International Science Engagement Challenge

ISEC 2019 Project Booklet

El Solitario, Spain

Table of Contents

What is ISEC

International Science Engagement Challenge (ISEC) is a new educational experience designed for people between the ages of 16 and 24. The main goal of ISEC is to carry out scientific projects in an international, open minded, interdisciplinary, and collaborative environment. Each participant, according to his or her interests, will be assigned a scientific mentor. Our mentors are mainly final year Bachelor students, Master students, PhD students or young science teachers. Each mentor will propose several projects, adapting and tailoring them to the interests and capabilities of the participants. Each participant will complete one research project individually or in a small group, overseen by a mentor. All the participants will all be working together in the same room and each mentor will be accessible to every participant, further promoting interdisciplinary collaboration. Once each applicant have been assigned a mentor, the mentor will contact the applicant with several project suggestions, that can be modified to fit your skills and education. This international scope will require applicants to be conversationally fluent in English.

ISEC definitely is not all work and no fun! That is exactly why ISEC team consider ISEC to be a summer camp and not a summer school. We strongly believe that the best learning takes best when students are eager and motivated to learn, but also have time to relax and rest their brains. At ISEC you're going to front flip into the pool, sing songs around the campfire, and stargaze under the Milky Way. You will have the opportunity to be silly and laugh and remember that even as you're growing up and taking on new responsibilities and stresses, there is always room for a new friend and a good time. ISEC is organized by an organization called Inspiring Science Association (ISA)

What is ISA

The Inspiring Science Association (ISA) is a non profit organization based in Spain and formally registered in the Spanish National Registry of Associations. We are an international team of young people with very different backgrounds who want to do something beautiful. We all share a common love for Science and the will to spread it. Our driving principles and main objectives are:

1. To encourage, strengthen and inspire scientific vocations among people all over the world.
2. To present, develop and study new ways of learning Science.
3. To familiarize future scientists and science enthusiasts with the reality of scientific research.
4. To enhance critical thinking, skepticism, initiative, inquiry, curiosity, communication, open mindedness, and reflective thinking, qualities which we believe are essential characteristics of successful scientists and can improve society.
5. To create an international environment of cooperation and collaboration among people from very different places, backgrounds, ages and interests.

The Inspiring Science Association has a democratic, transparent structure, made up of:

- The General Assembly, containing all the Inspiring Science Association members, whose main faculties are to approve the Directive Board, examine and approve the yearly account balance, modify the Statutes of the association, discuss the admission of new associates, and select the mentors of ISEC following the open selection process.
- The Directive Board, which leads the main project and is in charge of the management of the association, as well as following and executing the resolutions of the General Assembly.
- External volunteers, who are not a part of the Inspiring Science Association but collaborate with the association as needed and to whom we are extremely grateful. We are always looking for more volunteers if you are interested!

The Board of ISA and the ISEC 2019 Mentors. From left to right are Radka, Carlos, and Oz.

Carlos Looks Back on ISEC 2019

I still remember as if it was yesterday the day ISEC was born. It was an unusual sunny day in Luxembourg. I was walking along my best friend while we were talking about science, education and teaching. We shared the same vision and we had lived experiences similar to what we were discussing. She saw my eyes vividly sparkling and then gave me the push that any sane person needs in order to endeavor such an adventure: "Carlos, if all you are telling me means so much for you, do not hesitate at all, and do it!"

Two years, several mistakes and infinite hours of work later, we made it! We compensated our inexperience with effort and courage, and the reward was to see your smile each one of the scarce seventeen days that ISEC lasted. Because despite my constant grunting, I could not have imagined a better start of what I hope it becomes an initiative that lasts many years. Because despite our insistence on reminding me how old I am, you made me stop feeling like that and made me laugh again as a child in every game, every working session and every evening at the pool.

I hope with all my heart that you have enjoyed these days at El Solitario as much as I did, that you liked the projects and the games and fun sessions. But what I desire the most is that you enjoyed so much that you are eager and expecting the beginning of ISEC 2020, since nothing would make me happier than sharing another summer with you.

It has been an honor being your mentor. Thank you so much to all and each one of you: Andrea, Anisia, Anthony, Eva, Katie, Lavinia, Maggie, Niu, Sofía, Tonya, Vivaan, Wytske and Yanni. None of this would have been possible without you, without the leap of faith you took when you trusted us. I hope I did not let you down. I wish you all the best – may we meet again in the near future.

ISEC 2019: Özgür's Insights

What a summer! I still think someone will wake me up from this dream. After long years of preparation and labour, we finally managed to organise the first ISEC! Here at ISEC 2019 saw many friendships established, memories generated, science learned, games played and projects made. It was surely an amazing experience from various perspectives and that fact alone makes ISEC 2019 a wonderful event for me.

I hope ISEC also managed to provide these feelings to all you participants as well! Because it was you who made this camp happen and give its final shape. You, the first participants of ISEC planted the tree of this camp's future traditions. It was exciting to observe how ISEC gained its final form with the help of your amazingness. Thanks to each one of you for being part of ISEC 2019 and I hope you all enjoyed the times we spent together at Spain's unique atmosphere at El Solitario (Baa!)

Finally, thank you for working with me as your project mentor. All of you have made tremendous effort and I am proud of spending an ISEC full of your projects. Also thank you all the participants: Andrea, Anisia, Anthony, Eva, Katie, Lavinia, Maggie, Niu, Sofía, Tonya, Vivaan, Wytske and Yanni.

So Long And Thanks For All The Fish! Let The Force Be With You.

Radka's ISEC 2019 Musings

Some things are easily forgotten. Some are not. ISEC 2019 was full of things that are not easily forgotten. I hope that you remember these last few weeks in Spain fondly. I hope you remember your (perhaps apprehensive) arrival in Baños de Montemayor, as well as your sleep deprived and bittersweet departure. I hope you learned from the successes in your projects, as well as the failures and challenges. But most importantly, I hope that you enjoyed each other's company and connected with peers you otherwise would not have had the opportunity to meet.

Personally, ISEC 2019 was everything that the mentors hoped for and more. This first iteration was multiple years in the making with anticipated and unanticipated challenges along the way. Seeing all of you come together and buy-in to our silly games and quirky personalities made it all worthwhile. There were certainly lessons to learn along the way. I have come away with invaluable knowledge about how to make projects run more smoothly in the future and I learned a lot of interesting details about Fred the Sticky Moose. I also promise you that the scavenger hunt will not be on the same day as the project deadline. And it will be easier. And I won't force Carlos to take selfies.

Thank you to Carlos and Oz for undertaking this journey with me and thank you to all the participants - Andrea, Anisia, Anthony, Eva, Katie, Lavinia, Maggie, Niu, Sofía, Tonya, Vivaan, Wytske, and Yanni. My last hope is that you will all be eager to come back to ISEC in the future, and to build on the wonderful friendships that began at El Solitario.

Top Row: Lavinia, Sofía, Yanni, Anisia, and Vivaan; Middle Row: Tonya, Katie, Maggie, Andrea, and Niu; Bottom Row: Carlos, Wytske, Oz, Eva, Anthony, and Radka

Maxwell's Equations

Andrea Majocchi; Muhammad Anthony Abadi
International Science Engagement Challenge 2019

This report shows that the natural phenomenon of electricity and of magnetism are closely intertwined with each other. Through the use of mathematical tools one can succinctly describe and couple said phenomena, which have successfully been proven time after time through empirical means. The mathematical description of electromagnetism culminates in the four Maxwell's equations, which have the capability of describing and predicting with high degree of accuracy all electromagnetic phenomena.

Introduction

In order to understand Maxwell's equations it is important for the reader to be familiar with several concepts regarding electromagnetism, namely the concept of electrical charges, magnets, electric and magnetic forces. By way of historical recounting shall these ideas be explained in this report.

Ancient civilizations were aware of electric and magnetic phenomena. Looking around them, people noticed that lodestones could attract iron and several writers attested to the numbing effect of electric shocks caused by some sea creatures. For millenia, electricity and magnetism remained merely strange and mystifying facts, without a theory to explain their enigmatic properties. Around 600 B.C. Greeks were aware of the curious behaviours possessed by two minerals, amber (elektron) and magnetic iron ore (the Magnesian stone). The presocratic philosopherTales of Miletus discovered that after he rubbed amber with wool, the amber could attract other objects [6]. In the 11th century, the Chinese scientist Shen Kuo was the first person to describe the magnetic needle compass, an invention that improved the navigation by employing the astronomical concept of true north.

In 1600 the English physician William Gilbert, in his book "De Magnete", showed a careful study of electricity and magnetism, distinguishing the lodestones effect from the static electricity produced by rubbering amber. After several experiments, he discovered that other objects had electrical properties and concluded that Earth itself was magnetic [7]. Gilbert's work was followed up by Robert Boyle, a natural philosopher and a member of the Royal Society who stated that electric attraction and repulsion can act across the vacuum without the dependence upon the air as a medium. Further work was conducted by Otto Von Guericke, Stephen Gray and John Ellicott in the early 18th century. Scientists started building the first electric machines and understanding the difference between conductors and non-conductors.

In the last part of the century, Benjamin Franklin promoted his investigations of electricity through the famous experiment of having a metal key attached to the bottom of a kite string in a storm-threatened sky [8]. The experiment showed that lightning was indeed electrical in nature. Franklin is also considered to have established the convention of positive and negative charge: two positive charges or two negative charges repel each other, a positive charge and a negative charge attract each other.

Before Charles-Augustine de Coulomb published his research concerning electricity [3], people had been suspecting that the electrostatic force, i.e. Coulomb force (F_C), is analogous to gravitational forces, in that the magnitude is directly proportional to the electrical charges ($Q_1; Q_2$) and inversely proportional to the square of the distance between them ($\vec{r}_{21}$). Very much similar to the experiment carried out by Henry Cavendish on gravitational force, a torsion balance was employed in order to measure this relationship. Coulomb reported in his paper that his experiment was able to successfully verify this proportionalities (1) and came up with the so called Coulomb's equation (2).

$$\vec{F_C} \propto \frac{Q_1 Q_2}{|\vec{r}_{21}|^2} \hat{r}_{21} \tag{1}$$

$$\vec{F_C} = \frac{1}{4\pi\epsilon_0} \frac{Q_1 Q_2}{|\vec{r}_{21}|^2} \hat{r}_{21} \tag{2}$$

In 1820 Hans Christian Ørsted found out that the direction of a compass needle can be altered by placing it near a straight long wire carrying an electric current (DC) [4]. He noticed that the compass needle aligned itself perpendicularly to the wire. Since it was widely known that the alignment of any compass needle corresponds to the direction of the magnetic field lines at the position of the needle, it could be said that there is magnetic field lines circling the wire. He therefore concluded that moving electric charges (I) induce magnetic field (H) in the area surrounding them as shown in figure 1, thus uniting electricity and magnetism for the first time. Furthermore, he discovered that the effect gets weaker as the shortest distance (r) between current carrying wire and the compass needle grows and so the following proportionality can be said.

$$H \propto \frac{I}{r} \tag{3}$$

This report is structured in two different section, namely the methods and the conclusion. In the methods two main ideas shall be elaborated. The first one con-

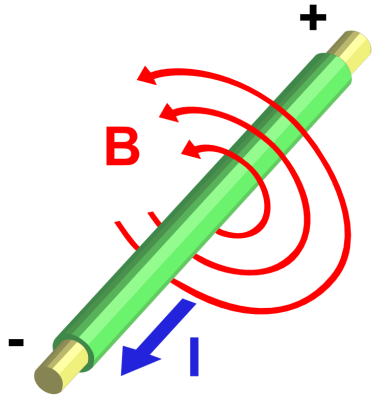


Figure 1: Ørsted's experiment (Taken from `https://en.wikipedia.org/wiki/Oersted\%27s_law`)

tains all the explanation about the mathematical tools necessary for understanding Maxwell's equations. Here we explain the concept of vector field, multiple integral, flux, divergence, curl and lastly a couple of theorems, which play an integral role to play for transforming the integration form of the Maxwell's equation to its respective differential form. This section mainly references two books: [1] for the mathematical structure of field and [2] for vector calculus.

In the second part of the section, the Maxwell's equations themselves are be explained in detail, by the end of which the reader should have a firm grasp on how these equations can describe electromagnetic phenomena. Moreover, the reader is expected to have some grounding on basic calculus.

The conclusion of this report lastly should give the reader a brief overview about Maxwell's equations and their implication.

Methods

A. Mathematical Tools

Maxwell's equations rely heavily on mathematical concepts, therefore it is indispensable that the reader has a grounding on these ideas. For example, the concept of flux and to a further extend the rate of change of flux is used in all the equations. Another example would be the application of line integral in the third and fourth equations. So, it has now been established that without these mathematical tools the reader could not grasp the meaning of these equations.

1. *Vector Field*

a. Group A set of M, on which the operation $\circ$ is defined, is said to be the group $(M;\circ)$ if all arbitrary elements $a;b;c \in M$ fulfill the following axioms:

G1) Closure $Cl_\circ$:

$$\forall a;b \in M : a \circ b = c \wedge c \in M$$

G2) Associativity $A_\circ$:

$$(a \circ b) \circ c = a \circ (b \circ c)$$

G3) Existence of Identity Element $N_\circ$:

$$\exists n \in M : a \circ n = a$$

G4) Existence of Inverse Element $I_\circ$:

$$\exists \bar{a} \in M : a \circ \bar{a} = n$$

A group is called an Abelian group or commutative group if in addition to the said axioms commutatvity $C_\circ$ holds:

$$a \circ b = b \circ a$$

b. Field A set F is called field $(F;\bigoplus;\bigodot)$ if on F the operations $\bigoplus$ and $\bigodot$ are defined and have the following characteristics:

F1) F is a Abelian group concerning the operation $\bigoplus$ with the identity element $n = 0$.

F2) $F\backslash\{0\}$ is a Abelian group concerning the operation $\bigodot$ with the identity element $e = 1$.

F3) F has distributive property:

$$a \bigodot (b \bigoplus c) = a \bigodot b \bigoplus a \bigodot c; \quad a,b,c \in K$$

c. Vector Field A vector field over a field F ("F"-Vector Field) $(V;\bigoplus;\bigodot)$ is a set V with the following characteristics:

V1) The operation $\bigoplus$ ("Vector Addition") is defined for $\vec{e}_1 \circ \vec{e}_2$ $(\vec{e_1},\vec{e_2} \in V)$ as the following:

$$\bigoplus : \vec{a} \bigoplus \vec{b} = \vec{c}; \quad \vec{a},\vec{b},\vec{c} \in V$$

With respect to the operation $\bigoplus$, V can build an Abelian group $(V;\bigoplus)$.

V2) The operation $\bigodot$ ("Scalar Multiplication") is defined for $e_1 \circ \vec{e}_2$ $(e_1 \in F \wedge \vec{e}_2 \in V)$ as the following:

$$\bigodot : \lambda \bigodot \vec{a} = \vec{b}; \quad \lambda \in F \wedge \vec{a},\vec{b} \in V$$

With the following properties:

$$A_\bigodot : \lambda \bigodot (\mu \bigodot \vec{a}) = (\lambda \cdot \mu) \bigodot \vec{a}$$
$$N_\bigodot : 1 \bigodot \vec{a} = \vec{a}$$

V3) V has distributive property:

$$D1) \quad \lambda \bigodot (\vec{a} \bigoplus \vec{b} = \lambda \bigodot \vec{a} \bigoplus \lambda \bigodot \vec{b}$$
$$D2) \quad (\lambda + \mu) \bigodot \vec{a} = \lambda \bigodot \vec{a} \bigoplus \mu \bigodot \vec{a}$$

d. n-Tupel as Vector Field A tuple can be simply defined as a finite ordered list of elements. This tuple T, in which all the elements are rational numbers, can be represented as follows:

$$T = \mathbb{R}^n = \{(a_1; a_2; a_3; ...; a_n) | a_i \in \mathbb{R}; i = 1...n; n \in \mathbb{N}\}$$

The reader can easily proof that the conditions of V1, V2 and V3 hold and thus $(T; \bigoplus; \bigodot)$ is valid as a vector field.

Two more operations are additionally required in order to describe the characteristics of electromagnetic forces using mathematical language. These operators are called the dot product "$\cdot$" and the cross product "$\times$". The reader might want to read the referenced materials in order to understand these operations.

e. Application of n-tupel $(n = 3)$ *Vector Field* Since the first element of the tuple can represent the magnitude of a geometrical structure in the x-axis, the second element in the y-axis and finally the third element in the z-axis, 3-tupel is able to sufficiently represent euclidean vectors in 3-D. Thus, all representations of physical phenomena which includes magnitude and direction, such as electromagnetic forces and fields, can be sufficiently described through 3-tuple vector fields. In virtue of this property, the only way to measure electrostatic force and to a further extend electric is through the placing of still test charges within a region influenced by the field, which depending on where it is placed will be accelerated with a certain magnitude and in a certain direction corresponding to the magnitude and direction of the electric field. The same can also be said to magnetic force, i.e. Lorentz force, and to magnetic field, which can be determined by measuring the magnitude and direction of the acceleration of moving electrical charges. Furthermore, the mapping of coordinates to vectors $\vec{E} : \mathbb{R}^3 \mapsto \mathbb{R}^3$ are entirely possible using this mathematical structure.

The property of $\bigoplus$ "Vector Addition" also plays a big role in the description of electromagnetic field due to the fact that when two or more electric charges are present, the electric field present at each position in space is the summation of all the electric fields from all the electric charges. Through the observation of coils carrying current the same property can also be found in the magnetic field. With each twist added into the coil the magnetic field generated by the coil gets stronger, because all the magnetic fields that are respectively induced by each twist are added up to determine the magnetic field strength of the whole coil. In addition to this, it is a well-known fact that from the experiment conducted by Coulomb regarding electric forces the property of $\vec{E} \propto Q_{present}$ was discovered and from Ørsted's experiment that $\vec{H} \propto I$. Hence, another property of vector field, namely $\bigodot$ "Scalar Multiplication", comes in handy, because as the strength of electric and magnetic field at almost every position grows as the magnitude of the charges involved or electric current grows.

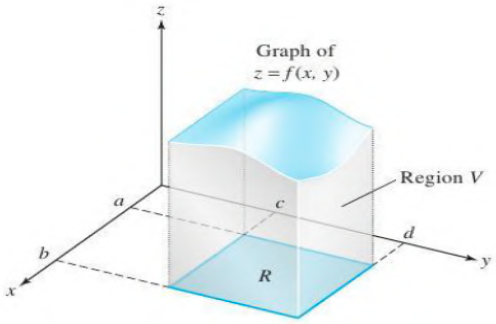


Figure 2: The region V bounded by $f(x, y)$ and R can be calculated by double integration. (Taken from p.264 of Vector Calculus [2])

2. Integration

a. Multiple Integral The multiple integral is a definite integral of a function $f : \mathbb{R}^n \to \mathbb{R}^m; n, m \in \mathbb{N}$ of more than one real variable, e.g. $f(x, y)$ or $f(x, y, z)$. Integrals of a function $f(x, y)$ over a region in $\mathbb{R}^2$, usually bounded by XY-plane, are called double integrals, and integrals of a function of three variables over a region of $\mathbb{R}^3$, typically made by three orthogonal vectors, are called triple integrals.

Since triple integration has more or less the same idea and mechanics as double integration, this section shall only discuss the double integration for brevity sake. Let a function with two real variables be $f(x, y)$ and the region over the function be R as demonstrated in figure 2. The integration of $f(x, y)$ over the region R shall be given below.

$$\iint_R f(x, y)\, dA \iff \iint_R f(x, y)\, dx\, dy \tag{4}$$

In geometry this translates to the calculation of the volume V bounded by the surface $f(x, y)$ and the XY-plane as depicted by the figure above. The limits that are being used to define the boundary of the integration should correspond to the region R which is a "shadow" of $f(x, y)$. The mechanism of double integration can be thought of finding the volume of a loaf of bread by way of summing up all the slices of bread.

b. Arc Length A path in a region $\mathbb{R}^2$ can be written as a function $y = f(x)$. To calculate the function's arc length with integration, one would need to find a way to divide the path into smaller components, which can be done through the process of parameterization. The main idea behind this is to use any one variable from the same or different coordinate system to map the whole function. The time variable t shall be chosen for this explanation, so through the use of 2-tuple the function can be parameterized as $\vec{l}(t) = (t, f(t))$. Since distance can be calculated by integrating a function of velocity with respect to time and $\vec{l}'(t) = (1, f'(t))$ can be thought of as a 2-tuple, in which the elements are two functions describing velocity in the x and y axis, the arc length can be found by integrating the magnitude of this function with respect to time. The concept can also be extended to a path in a region $\mathbb{R}^3$.

The length of any given path with the parameterized equation $\vec{c}(t) = (x(t), y(t), z(t))$ in the domain of $t_0 \leq t \leq t_1$ can be calculated with the following equation:

$$ArcLength = \int d(\vec{c}(t)) \tag{5}$$

$$= \int_{t_0}^{t_1} |\frac{d\vec{c}(t)}{dt}|\, dt \tag{6}$$

$$= \int \sqrt{[x'(t)]^2 + [y'(t)]^2 + [z'(t)]^2}\, dt \tag{7}$$

c. Path Integral A path is an oriented curve in the region of $\mathbb{R}^3$. The integral of a function $f(x, y, z) : \mathbb{R}^3 \to \mathbb{R}$ over a path $\vec{c} : P = [a, b] \in \mathbb{R} \to \mathbb{R}^3$ can be defined only under the condition that the function f is continuous along the path P. This can be written as the following:

$$\int_c f\, dP = \int_c f(x, y, z)\, dP \tag{8}$$

$$= \int_a^b f(x(t), y(t), z(t))|\vec{c}'(t)|\, dt \tag{9}$$

$$= \int_a^b f(\vec{c}(t))|\vec{c}'(t)|\, dt \tag{10}$$

Take the example of calculating the average temperature along the equator line E. The surface of the earth can be imagined as a scalar field of temperature and $T(x, y, z)$ as the function that maps the temperature at a given location. The mean temperature is given by dividing the sum of all measurements taken and the number of measurement taken. Since with the help of the function T there shall be infinitely many measurement, the average cannot be calculated by the usual statistical method, rather with the following formula.

$$\int_{Quito}^{nearQuito} T(x(t), y(n), z(n))\, dE \tag{11}$$

d. Line Integral The line integral on the other hand can only be done over a vector field. The idea is analogous to the path integral with only a slight difference. In the line integral the vector given by the field function in the direction of the path is the one that is taken into account rather that the vector itself, which may or may not contribute to the overall integral.

Let the vector field on $\mathbb{R}^3$ denoted by $\vec{F}$ and the path by $\vec{c} : P = [a, b] \in \mathbb{R} \to \mathbb{R}^3$. Under the assumption that $\vec{F}$ is continuous along the path P, the line integral is given by the equation below.

$$\int_c \vec{F}\, dP = \int_a^b \vec{F}(\vec{c}(t)) \cdot \vec{c}'(t)\, dt \tag{12}$$

e. Integral Over Surfaces Before diving into integral over surfaces in scalar and vector field, it is advisable to think about the way surfaces can be parameterized and calculated by integral. Unlike paths, surfaces given by $f(x, y, z)$ cannot be parameterized by only utilizing a single variable, therefore another one shall be added. The function $\Theta : D \subset \mathbb{R}^2 \to \mathbb{R}^3$ maps the surface from an elementary region in a plane. The function mapping the surface thus can be written as follows.

$$\Theta(u, v) = (x(u, v), y(u, v), z(u, v)) \tag{13}$$

It can be easily proven that by using the magnitude of the cross product between two vectors the area of a parallelogram made by vectors involved can be calculated. By means of partial derivation the function that maps all the vectors parallel to the surface from the uv-plane can determined. In one hand, if $\frac{\partial\Theta}{\partial u}$ is calculated, then all the mapped vectors shall be in the direction of the u-axis. On the other hand derivation by $\frac{\partial\Theta}{\partial u}$ gives all vectors in the direction of v-axis.

$$\vec{T}_u(u, v) = \frac{\partial\Theta}{\partial u} \tag{14}$$

$$= \frac{\partial x(u, v)}{\partial u} + \frac{\partial y(u, v)}{\partial u} + \frac{\partial z(u, v)}{\partial u} \tag{15}$$

$$\vec{T}_v(u, v) = \frac{\partial\Theta}{\partial v} \tag{16}$$

$$= \frac{\partial x(u, v)}{\partial v} + \frac{\partial y(u, v)}{\partial v} + \frac{\partial z(u, v)}{\partial v} \tag{17}$$

The surface area A can then be calculated by summing the magnitudes of all the cross product at every location on the surface, which is expressed with the following equation.

$$A = \iint_D |\vec{T}_u \times \vec{T}_v| \, du \, dv \tag{18}$$

Analogous to the case of path integral, when integrating a scalar function $f(x, y, z)$ over a surface A, the formula is as follows.

$$\iint f(x, y, z) \, dA = \iint f(\Theta(u, v))|\vec{T}_u \times \vec{T}_v| \, du \, dv \tag{19}$$

Integration of vector fields $\vec{F}$ over a surface however is more distinct than its counter part, namely line integral. Rater that taking the vector that is parallel to the surface into account, here only vectors in the direction perpendicular or normal to the surface contribute to the integral. Since $\vec{T}_u \times \vec{T}_v$ conveniently maps all the normal vectors of each point on the surface, the dot product $\vec{F}(\Theta(u, v)) \cdot (\vec{T}_u \times \vec{T}_v)$ projects $\vec{F}$ to the normal vectors of the surface. The equation to calculate is described below.

$$\iint \vec{F}(x, y, z) \, dA = \iint \vec{F}(\Theta(u, v)) \cdot (\vec{T}_u \times \vec{T}_v) \, du \, dv \tag{20}$$

f. Flux "In the case of fluxes, we have to take the integral, over a surface, of the flux through every element of the surface. The result of this operation is called the surface integral of the flux. It represents the quantity which passes through the surface." - James Clerk Maxwell [5]

Let $\vec{F}$ denote a vector field, A an oriented surface in three-dimension and Φ_F the flux of $\vec{F}$ on the surface A. The formula for finding the flux is the following:

$$\Phi_F = \oiint_A \vec{F} \cdot \, d\vec{A} \tag{21}$$

One may need to look into physics for further understanding, more specifically in the so called transport phenomena where the flux is the rate of flow of a property per unit area. Now, consider measuring the flow rate of a river with a net, out of which an imaginary permeable surface can be thought of. Moreover, the river itself can now be considered as a vector field of current density $\vec{J}$. The net is then thrown into the river, but it might not be oriented perpendicularly to the river current and takes a corrugated form, so the rate cannot be measured directly by dividing the current density with the total surface area formed by the net. In order to end this predicament, one just need to find the vector component of the current density in the direction normal to the area of the net $\vec{A}_{net}$, which can be conveniently done by using the dot product $\vec{J} \cdot \vec{A}_{net}$. If $\vec{A}_{net}$ is too big and the $\vec{J}$ is not uniform throughout, the measurement would not be accurate, rendering value of the measured flux false. Therefore, it is advisable not to utilize the total surface area of the net, but rather use the mesh of the net for a more accurate measurement. The finer the mesh the more precise the measured flux will become. If the mesh become so tiny that $\vec{A}_{mesh} \to d\vec{A}$, then the value of flux over the whole net can be calculated by summing up all the flux on each mesh ($d\Phi_J = d(\vec{J} \cdot \, d\vec{A})$). Therefore,

$$\iint d(\vec{J} \cdot \vec{A}) = \iint \vec{J} \cdot \, d\vec{A} \tag{22}$$

Interestingly enough the flux can be thought of as the more fundamental quantity than the vector field itself, which can be considered as flux density.

3. Differentiation

a. Nabla Operator The nabla operator is simply a mathematical operator that is rather versatile for the purposes of finding the gradient, divergence and curl of a field, all of which shall be discussed below. A mathematical operator can be thought of as a function that maps an element of a space to another element in another space. The operator is as follows:

$$\vec{\nabla} = \begin{pmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{pmatrix} \tag{23}$$

b. Gradient The gradient imparts the rate and direction of change in a scalar field. Given a function $f : \mathbb{R} \to \mathbb{R}$ that maps a scalar field, the gradient of the scalar field can be found using the following formula:

$$\vec{\nabla} \cdot f = \begin{pmatrix} \frac{\partial}{\partial x} f \\ \frac{\partial}{\partial y} f \\ \frac{\partial}{\partial z} f \end{pmatrix} \tag{24}$$

c. Divergence Let F be a vector field $\vec{F} : \mathbb{R}^3 \to \mathbb{R}^3$. The divergence at any given point in the vector field can be determined with the following equation.

$$div\vec{F} = \vec{\nabla} \cdot \vec{F} = \frac{\partial}{\partial x} F_1 + \frac{\partial}{\partial y} F_2 + \frac{\partial}{\partial z} F_3 \tag{25}$$

There are a couple of interpretations, that can be employed for intuiting the concept of divergence:

First interpretation (Rate of expansion per unit volume): Let's take into consideration the velocity field F of a nitrogen gas. At time t_0 a small region enclosing the point P_0 that contains a certain set of $N_{2(g)}$ molecules shall be defined as $V(t_0)$. If one were to go along the time axis and were to trace the movement of every $N_{2(g)}$ molecules in the set, one would observe that at time t_1 a

new region containing the same set of $N_{2(g)}$ can be designated and shall be denoted by $V(t_1)$. Then, in this case $\nabla\vec{F}$ at P_0 can be approximated as the rate of expansion per unit volume under the flow of $N_{2(g)}$ or the equation below:

$$\nabla\vec{F} \approx \frac{V(t_1) - V(t_0)}{t_1 - t_0}\frac{1}{V(t_0)} \tag{26}$$

The following applies:

$$divF < 0 \Leftrightarrow Compression$$
$$divF = 0 \Leftrightarrow No\ Change\ in\ Volume$$
$$divF > 0 \Leftrightarrow Expansion$$

As the volume of the region $V(t_0)$ and $\delta(t_1 - t_2)$ approaches to 0, the value of $\nabla\vec{F}(P_0)$ will become more accurate.

$$\frac{1}{V(t_0)}\frac{d}{dt}V(t)|_{t\to t_0} \to divF(P_0) \tag{27}$$

Second interpretation (Flux per unit volume): Another way to intuit the concept of divergence is to imagine the case of turning a tap on. A small imaginary permeable box B in $\mathbb{R}^3$, at which center lies the point P, is placed in a region influenced by the velocity field $\vec{F}$ of tap water. Since the volume V of B is constant, the value of the divergence only depends on the net flux on the closed surface of the box. The flux tells us whether there is more water flowing in or out of the box, i.e. the rate of flow across the box boundary. If B is placed at the point P_a in a region, near which the tap water lands, the flux can be said to be positive, hence $\nabla\vec{F}(P_a) > 0$ and P_a can be called a source. Conversely, if B is placed at the point P_b in a region, near which the tap water goes into the sink, there is more water flowing in into than out of O, so $\nabla\vec{F}(P_b) > 0$ and P_b can be designated as a sink.

As B shrinks to P, the value of flux per volume B at point P shall converge to the true value of $\nabla\vec{F}(P)$. Let Φ_x, Φ_y, Φ_z be the flux in the direction of flux in the x-axis, y-axis and z-axis respectively.

$$Divergence = lim_{Vol\to 0}\frac{Flux}{Vol} \tag{28}$$

$$Divergence = \frac{\partial\Phi_x}{\partial x} + \frac{\partial\Phi_y}{\partial y} + \frac{\partial\Phi_z}{\partial z} \tag{29}$$

d. Curl Let $\vec{F} = \begin{pmatrix} F_1 \\ F_2 \\ F_3 \end{pmatrix}$ be a vector field. It's curl is given by the following equation.

$$\vec{\nabla}\times\vec{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix} = \begin{pmatrix} \frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \\ \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \\ \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \end{pmatrix} \tag{30}$$

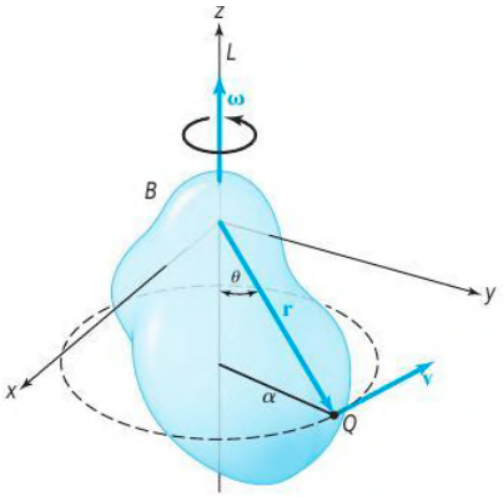


Figure 3: Rotating Rigid Body, a.k.a. Potato (Taken from p.250 Vector Calculus [2])

The concept of curl can be understood by considering the case of a rotating potato. A spinning potato, can be thought of as a rotating rigid body, which can be represented mathematically as a velocity vector field of all the component inside the body. Now, the curl of this body shall be calculated below.

Representing the vector field of a rotating potato with referencing the figure with reference to figure 3:

$$\vec{r} = x\mathbf{i} + y\mathbf{k} + z\mathbf{k}|\vec{v}| \qquad = \omega\alpha = |\vec{\omega}||\vec{r}|sin(\theta) \tag{31}$$
$$\vec{v} = \vec{\omega}\times\vec{r} \qquad = -\omega y\mathbf{i} + \omega x\mathbf{j} \tag{32}$$

Determining the curl of the potato's vector field:

$$\vec{\nabla}\times\vec{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ -\omega y & \omega x & 0 \end{vmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2\omega \end{pmatrix} = 2\omega\mathbf{k} \tag{33}$$

It has thus been proven that the curl of a rotating body's velocity field is equal to 2ω in the direction of the rotational axis given by the right hand rule. When the example of a flowing body of water is taken into account, it is implied that that the curl of any given point in the body of water can be determined by placing a rigid body at the said position and measuring its angular speed. As the size of the measuring object decreases the value of the curl at a given point will converge to the true value. This intuition can also be generalized to a different type of physical fields, e.g. electric field, in which the imagined "spin" of electron can be utilized for measuring curl.

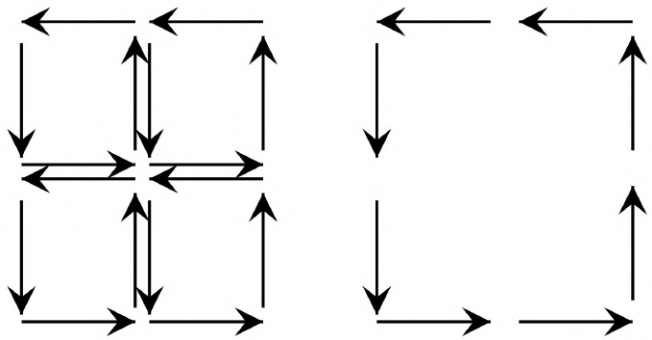


Figure 4: Visualization of Stokes' Theorem (Taken from `https://en.wikipedia.org/wiki/Stokes\%27_theorem`)

4. Theorems

a. Stokes' Theorem Stokes' theorem states that in any given oriented surface in space the sum of all the curl at each point on the surface tells us the closed path integral along the boundary of the said surface. This can be written as the following equation:

$$\iint_A \vec{\nabla} \times \vec{F}\, d\vec{A} = \oint_{\partial A} \vec{F}\, d\vec{l} \tag{34}$$

Where A is a surface inside a three-dimensional space given by the equation $z = f(x, y)$, ∂A be a closed path, which defines the border of the surface, and $\vec{F}$ be a vector field.

The underlying principle can be visualized in the figure 4. With its aid one might notice that the component of the curling paths or field lines that travel in opposite direction nullify each other effects and doesn't contribute to the line integral of the boundary. Since the curl at each point captures the rotating property of field lines satisfactorily, one can sum the curl at each point on the surface to cancel the effect of the field lines, which are anti-parallel to each other, and add up the effect of the parallel field lines, therefore leaving the field line components which is consequential to the line integral intact.

b. Gauß' Theorem Gauß's divergence theorem states that in any given symmetric elementary region enclosed by a surface the total flux can be determined by summing the divergence at all points inside the enclosed surface. This can be written as the following equation:

$$\iiint_V \vec{\nabla} \cdot \vec{F}\, d\vec{V} = \oiint_{\partial V} \vec{F}\, d\vec{A} \tag{35}$$

Where V is a symmetric elementary region in space, ∂V be the enclosing surface of the region and $\vec{F}$ be a vector field.

In order to develop a physical intuition for this theorem it is helpful to consider a flowing body of liquid. The flow of liquid can be modeled after a vector field of current density, thus the total flux of any enclosed surface tells us the net flow into or out of a volume's boundary, which implies "how much of a sink or a source" a given region is. According to the theorem one way of investigating the net of the rate of liquid flow across a given region, i.e. flux, is to sum up the sources and from the sum to subtract the sinks inside the enclosed region, which is intuitive bearing in mind that the whole, net flux, is the sum of its parts, the flux at each point inside a region.

B. Maxwell's Equations

In this section we are going to explain Maxwell's four equations using the mathematical tools described in the previous pages. In order to avoid misunderstanding, we want to mention that Maxwell did not discover all these equations single-handedly, but he did put them together and recognized their significance, particularly in predicting the existence of electromagnetic waves.

1. Gauß's Law

The first equation, or Gauß's law for the electric field, describes the relationship between the static electric field and the electric charges that cause it. The static electric field, which is the vector field that associates to each point in space the Coulomb's law, points away from positive charges and towards negative charges. Picturing the electric field by its field lines, the net outflow of the electric field through any closed surface is called electric flux.

The first equation states that the total electric flux through any closed surface is proportional to the total charge enclosed by the surface:

$$\oiint_A \vec{E} \cdot\, d\vec{A} = \frac{Q_{encl}}{\epsilon_0} \tag{36}$$

The symbol $\oiint_A$ indicates the closed surface integral over the surface A, Q_{encl} is the net charge enclosed by the surface, $\vec{E}$ is the vector of the electric field and ϵ_0 is the permittivity of free space.

Now, we are going to use the surface integral of a vector to find Maxwell's first equation:

$$\iint_A \vec{F} \cdot\, d\vec{A} = \iint_A F[z(x,y)] \cdot (\vec{T_x} \times \vec{T_y})\, dx\, dy \tag{37}$$

If we place a charge at the center of an imaginary spherical surface with radius r, we need to parametrize the surface of interest, by considering the latitude and

the longitude on the sphere. The coordinates of a sphere are:

$(x = r\sin\phi\cos\theta; y = r\sin\phi\sin\theta; z = r\cos\phi)$ where $\theta\epsilon(0, 2\pi)$ and $\phi\epsilon(0, \pi)$

The flux is:

$$\iint_A \vec{E} \cdot d\vec{A} = \vec{E} \iint_A (\vec{T_\theta} \times \vec{T_\phi})\, d\theta\, d\phi \tag{38}$$

Since $\vec{E}$ is constant in all the surface and doesn't depend on the two variables $d\theta$ and $d\phi$, it can be put outside the integral.

We know that the electric field at a point charge is:

$$E = \frac{Q}{4\pi r^2 \epsilon_0} \tag{39}$$

where $A = 4\pi r^2$ is the surface of the sphere.

The vectorial product between the longitude and the latitude is:

$$(\vec{T_\theta} \times \vec{T_\phi}) = r^2 \sin\phi \tag{40}$$

Now we can check the equation:

$$\Phi = \iint_A \vec{E} \cdot d\vec{A} = \frac{Q}{4\pi r^2 \epsilon_0} \iint_A r^2 \sin\phi\, d\theta\, d\phi \tag{41}$$

$$= \frac{Q}{4\pi\epsilon_0} \int_0^{2\pi} d\theta \int_0^{\pi} \sin\phi\, d\phi = \frac{1}{2\epsilon_0} 2Q \tag{42}$$

$$= \frac{Q}{\epsilon_0} \tag{43}$$

This is also true for the electric flux through any closed Gaussian surface in integral:

$$\Phi_E = \oiint_A \vec{E} \cdot d\vec{A} = \frac{Q_{encl}}{\epsilon_0} \tag{44}$$

Where $\vec{E}$ is the vector of the electric field, A is any closed surface and $d\vec{A}$ is the vector representing an infinitesimal element of the surface A. We can see that the flux is independent of the radius r and depends only on the charge enclosed by the surface.

The first Maxwell's equation can be expressed in differential form by using the divergence theorem.

$$\iiint_V \vec{\nabla} \cdot \vec{E}\, dV = \iint_A \vec{E} \cdot d\vec{A} \tag{45}$$

The total electric charge Q enclosed in A is the volume integral over of the charge density ρ.

$$Q = \iiint_V \rho\, dV \tag{46}$$

By the relation between equation (28) and equation (29) we have:

$$\iiint_V \vec{\nabla} \cdot \vec{E}\, dV = \iiint_V \frac{\rho}{\epsilon_0}\, dV$$

$$\vec{\nabla} \cdot \vec{E} = \frac{\rho}{\epsilon_0} \tag{47}$$

2. *Gauß's Law for magnetism*

The second equation, or Gauß' law for magnetism, states that there are no magnetic charges (also called magnetic monopoles [9]) analogous to electric charges. Instead the magnetic field is generated by a configuration called dipole, and the net outflow of the field through any closed surface is zero. We are able to see this phenomenon in magnets: their fields lines always move in a closed path from the north pole to the south pole. For this reason magnetic dipoles can be represented as loops where positive and negative charges are inseparably bound together, having no net magnetic charge. So the total magnetic flux through any closed surface is zero.

$$\Phi_B = \oiint_A \vec{B} \cdot d\vec{A} = 0 \tag{48}$$

where $\vec{B}$ is the vector of the magnetic field, A is any closed surface and $d\vec{A}$ is a vector, whose magnitude is the area of an infinitesimal piece of the surface A, and whose direction is the out-ward pointing surface normal.

The second Maxwell's equation in integral form states that the magnetic flux through any closed surface, like a sphere or a cube, is zero; but the magnetic flux through non-closed surfaces is not necessarily zero.

In differential form, the equation is:

$$\iiint_V \vec{\nabla} \cdot \vec{B}\, dV = \iint_A \vec{B} \cdot d\vec{A} = 0 \tag{49}$$

$$\iiint_V \vec{\nabla} \cdot \vec{B}\, dV = \iiint_V 0\, dV$$

$$\vec{\nabla} \cdot \vec{B} = 0 \tag{50}$$

3. Faraday's law

The third equation, or Faraday's law of induction, describes how a time varying magnetic field induces an electric field. During the 1830s, Micheal Faraday performed several experiments and made a great discovery: the variation of the magnetic field through the surface enclosed by a loop induces an electromotive force, defined as electromagnetic work done on a unit charge when it has traveled one round of a conductive loop:

$$\varepsilon = -\frac{d\Phi_B}{dt} \tag{51}$$

where ε is the induced electromotive force by the variation of the magnetic flux Φ_B. The most widespread version of Faraday's law states that the electromotive force around a closed path is equal to the negative of the time rate of change of the magnetic flux enclosed by the path. For a loop of wire in a magnetic field, the magnetic flux is defined for any surface A whose boundary is the given loop. The magnetic flux through the wire loop is the surface integral:

$$\Phi_B = \iint\limits_A \vec{B} \cdot d\vec{A} \tag{52}$$

where $\iint\limits_A$ is the surface integral over the surface A and $\vec{B}$ is the vector of the magnetic field.

To have the complete Maxwell-Faraday's equation, we have to find the connection between the work W required to move a charge around a closed loop and the rate of decrease of the magnetic flux through the enclosed surface:

$$W = \oint\limits_\delta \vec{F} \cdot d\vec{l} = q \oint\limits_\delta \vec{E} \cdot d\vec{l} \tag{53}$$

$$\frac{W}{q} = \oint\limits_\delta \vec{E} \cdot d\vec{l} \tag{54}$$

where $\oint\limits_\delta$ is the closed line integral around the boundary curve and $d\vec{l}$ is the infinitesimal vector element of the contour δ.

The electromotive force is equivalent to the voltage that would be measured by cutting the wire to create an open circuit, and attaching a voltometer to the leads:

$$\varepsilon = \frac{W}{q} = \oint\limits_\delta \vec{E} \cdot d\vec{l} \tag{55}$$

Finally, we have the Maxwell-Farady's equation, which states that a time-varying magnetic field always accompanies a spatially varying, non-conservative electric field, and vice-versa:

$$\oint\limits_\delta \vec{E} \cdot d\vec{l} = -\frac{d}{dt} \iint\limits_A \vec{B} \cdot d\vec{A} \tag{56}$$

where A is a surface bounded by the closed contour δ. This aspect of electromagnetic induction is the operating principle behind many electric generators: for example, a rotating bar magnet creates a changing magnetic field, which in turn generates an electric field in a nearby wire. William Gladstone, English finance minister, was sent by the government to see Faraday's discoveries. After the electric experiment, he asked Faraday what was the practical value of his work. The scientist replied: "One day, sir, the government will put taxes on it [10]."

We can use Lenz's law to determine the direction of an induced current or electromotive force. The law states that the direction of any magnetic induction effect is such as to oppose the cause of the effect. For example, if we move a magnet towards a loop, the motion of the magnet causes an increasing flux through the loop and induces an electromotive force. The induced current in the loop generates a magnetic field with direction opposite to that of the magnet. Lenz's law is directly related to energy conservation. If the induced current were in the direction opposite to that given by Lenz's law, the magnetic force on the rod would accelerate it with no external source and the electric energy would increase, thus violating the principle of conservation.

The third Maxwell's equation can be written in differential form by using the Stokes' theorem:

$$\oint\limits_\delta \vec{E} \cdot d\vec{l} = \iint\limits_A (\vec{\nabla} \times \vec{E})\, d\vec{A} \tag{57}$$

$$\iint\limits_A (\vec{\nabla} \times \vec{E})\, d\vec{A} = -\frac{d}{dt} \iint\limits_A \vec{B} \cdot d\vec{A}$$

$$\vec{\nabla} \times \vec{E} = -\frac{dB}{dt} \tag{58}$$

4. Ampère's law

The fourth equation, or Ampère-Maxwell's law, states that magnetic fields can be generated in two ways: by electric current and by changing electric fields. In integral form, the magnetic field induced around any closed loop is proportional to the electric current and to the rate of change of electric flux.

Originally, Ampère's circuital law related the integrated magnetic field around a closed loop to the electric current passing through the loop:

$$\oint_{\delta} \vec{B} \cdot \, d\vec{l} = \mu_0 \iint_{A} \vec{J} \cdot \, d\vec{A} = \mu_0 I_{encl} \tag{59}$$

The symbol $\oint_{\delta}$ is the closed line integral of the magnetic field $\vec{B}$ around a current enclosed by the curve δ, μ_0 is the magnetic permeability of free space,$\vec{J}$ is the total current density in the surface A enclosed by the curve and $d\vec{l}$ is the infinitesimal element of the curve.

The original circuital law is correct only in a magnetostatic situation, where the system is static; but for systems with electric fields that change over time, the original law must be modified to include a term known as Maxwell's correction [11]. Maxwell conceived, next to the conduction current I_c through a wire, a displacement current I_d that is related to the change of electric field in free space.

$$I_d = \epsilon_0 \frac{d\Phi_E}{dt} = \epsilon_0 \frac{d}{dt} \iint_{A} \vec{E} \cdot \, d\vec{A} \tag{60}$$

In this way we can have Ampère's law in complete form:

$$\oint_{\delta} \vec{B} \cdot \, d\vec{l} = \mu_0 (\iint_{A} \vec{J} \cdot \, d\vec{A} + \epsilon_0 \frac{d}{dt} \iint_{A} \vec{E} \cdot \, d\vec{A}) \tag{61}$$

In differential form, the equation is

$$\oint_{\delta} \vec{B} \cdot \, d\vec{l} = \iint_{A} \vec{\nabla} \times \vec{B} \, d\vec{A} \tag{62}$$

$$\iint_{A} \vec{\nabla} \times \vec{B} \, d\vec{A} = \mu_0 (\iint_{A} \vec{J} \cdot \, d\vec{A} + \epsilon_0 \frac{d}{dt} \iint_{A} \vec{E} \cdot \, d\vec{A})$$

$$\vec{\nabla} \times \vec{B} = \mu_0 (J + \epsilon_0 \frac{dE}{dt}) \tag{63}$$

Discussion and Conclusion

There is a remarkable symmetry in Maxwell's four equations. In empty space, where there is no charge, the two Gauß' equations are identical in form, one containing the electric field and the other containing the magnetic field.

$$\oiint_{A} \vec{E} \cdot \, d\vec{A} = 0 \tag{64}$$

$$\oiint_{A} \vec{B} \cdot \, d\vec{A} = 0 \tag{65}$$

When we compare the other two equations, Faraday's law says that a changing magnetic flux creates an electric field, and Ampère's law says that a changing electric flux creates a magnetic field. In empty space, where there is no conduction current, the two equations have the same form, apart from numerical constants and a negative sign.

$$\oint_{\delta} \vec{E} \cdot \, d\vec{l} = -\frac{d}{dt} \iint_{A} \vec{B} \cdot \, d\vec{A} \tag{66}$$

$$\oint_{\delta} \vec{B} \cdot \, d\vec{l} = \mu_0 \epsilon_0 \frac{d}{dt} \iint_{A} \vec{E} \cdot \, d\vec{A} \tag{67}$$

An important consequence of the equations is that they demonstrate how electromagnetic disturbances consist of time-varying electric and magnetic fields that travel or propagate from one region of space to another, even though there is no matter in the intervening space. Such disturbances, called electromagnetic waves, provide the physical basis for the electromagnetic spectrum.

Although Maxwell's equations have stood the test of time remarkably well, some physicists wish that magnetic monopoles existed, for the sake of the mathematical poetry of the equations. Indeed, magnetic charges have never been observed, despite extensive searches. If they did exist, both Gauß' law for magnetism and Faraday's law would need to be modified, and the resulting four equations would be fully symmetric under the interchange of electric and magnetic fields.

Maxwell's equations are extraordinarily successful at explaining and predicting a variety of phenomena, however they are not exact and have several limits. For example, they can't explain some observed electromagnetic phenomena such as the quantum entanglement of electromagnetic fields, the photoelectric effect and Planck's law.

Acknowledgements

To our lovely mentor and kindhearted "Sklaventreiber" Carlos we dedicate this pièce de résistance.

[1] Arens, T., Busam, R., Hettlich, F., Karpfinger, C., Stachel, H. *Grundwissen Mathematikstudium*. Spektrum Akademischer Verlag, 2013.
[2] Jerrold E. Marsden, Anthony Tromba *Vector Calculus*. W. H. Freeman, Reading, 2012.
[3] C.A. Coulomb. *Premier Mémoire sur l'Électricité et le Magnétisme. Construction & usage d'une Balance électrique, fondée sur la propriété qu'ont les Fils de métal, d'avoir une force de réaction de Torsion proportionnelle à l'angle de Torsion*. (French) Mémoires de l'Académie Royale des Sciences, 1788.
[4] Oersted, H. C. *Experiments on the effect of a current of electricity on the magnetic needles*. Annals of Philosophy, 1820.
[5] Maxwell, James Clark *Treatise on Electricity and Magnetism*. Oxford University Press, 1892.
[6] Edward Tatnall Canby. *A History of Electricity*. Hawthorn books, NY, 1963.
[7] David P Stern. *A Millennium of Geomagnetism*. Reviews of Geophysics, 2002.
[8] Charles Byrne. *A Brief History of Electromagnetism*. Lowell, MA 01854 USA, 2015.
[9] John David Jackson. *Classical Electrodynamics*. University of California, Berkeley, 1999.
[10] Michio Kaku. *Hyperspace*. Oxford University Press, 1994.
[11] George E. Owen. *Electromagnetic Theory*. Mineola (NY), Dover Publications, 2003.

Introduction to Quantum Mechanics

María Magdalena Castro Sam
International Science Engagement Challenge 2019

Quantum Mechanics underlies a lot of aspects of physics and it can be found in numerous subjects. The mathematics and the theoretical concepts behind Quantum Mechanics can be too difficult to understand since they require a high level of abstraction. This work aims to revise the firsts steps to understand Quantum Mechanics including some of the History behind it and a brief introduction to the mathematical formalism and concepts.

Introduction

At the end of the 19th century scientists believed that physics was almost complete, and that there was nothing else to study but a few mysteries. However, some people were introducing concepts that slowly began to shape what would be a new area of physics. In this report we will try to address some of the aspects of Quantum Mechanics such as the history, a revision of some mathematical tools and concepts that are useful.

In Section II we talk about some experiments that paved the way to the birth of the theory. In Section III we will discuss some concepts and history of Heisenberg's interpretation of Quantum Mechanics. In Section IV we introduce some definitions about notation in order to address the postulates of Quantum mechanics in Section V. Finally, in Section VI we discuss the birth of Schrödinger's wave equation theory and the probabilistic interpretation of Quantum Mechanics. To properly understand Quantum Mechanics it is necessary to have mastered some of the mathematical tools. We include further references in the Appendix A to deepen into Quantum Mechanics.

Founding ideas

To this day there have been numerous experiments that show the quantum nature of the world. Here we try to discuss the most important experiments and ideas that led to the birth of Quantum Mechanics as we know it nowadays.

One of the unsolved mysteries at that time was the physics behind the black body. A black body is defined to be such that absorbs all the radiation. However, it doesn't mean that the object we talk about is necessarily black. When receiving and absorbing radiation, it can change colors by emitting its own radiation, like the metals that shine when they are heated.

There were numerous attempts to plot the intensity of the emitted radiation as a function of the wavelength. Nevertheless, classical laws couldn't describe accurately what was happening. Rayleigh and Jeans developed a formula that fit the experimental data at really high temperatures, but didn't described accurately low temperatures. This was named "The Ultraviolet Catastrophe", because the Rayleigh-Jeans' law does not predict correctly the behavior of the frequencies of the ultraviolet region of the electromagnetic spectrum.

Max Planck came up with an idea in order to solve the black body radiation problem. He supposed that energy wasn't a continuous, but rather came in small bits, named *quanta*, which had very specific values for energy, and there could not be intermediate values. This simple assumption perfectly corresponded the experimental data and paved the way to a new formulation to describe our world.

Another phenomenon that couldn't be described was the photoelectric effect. Certain metals emit electron when light shines on them. Classical intuition would tell us that the number of electrons would be proportional to the intensity of light. However, the experiments showed that it was not the intensity but the frequency of the radiation what matters. The higher the frequency, the more energetic the electrons are.

Albert Einstein used Planck's hypothesis in order to come up with a formula that described and explained the photoelectric effect. At the time, the accepted conception of light was that it was a wave. Following Planck's hypothesis, Einstein proposed that light was quantized too. This quanta of energy were later called 'photons': the particles that compose light.

Besides these two experiments, there were a couple of ideas that also contributed to the formulation of Quantum Mechanics. The first is Bohr's atom model. The most known idea of the structure of an atom is Rutherford's model, a nucleus in the center and electrons orbiting around it. However, if atoms were this way, and followed Rutherford's model, we would not exist: electrons wouldn't be able to orbit and they would quickly fall towards the nucleus. Nowadays we know that Bohr's atomic model is the correct one: Bohr explained that just as energy had discrete values, electrons that moved around the atom are confined to specific orbits in discrete levels of energy.

The second important idea towards Quantum Mechanics was De Broglie's hypothesis. Louis De Broglie started from Einstein's hypothesis of the photoelectric effect. He tried to come up with a formalism that showed that if waves could be particles, particles could be waves too. "De Broglie's intuition that solid matter might be wave-like revealed a unity in Nature that played a central role in the formulation of Quantum Mechanics"[1].

De Broglie quantized the electron's angular momentum

in the hydrogen atom using Bohr's ideas about Quantum Mechanics, which required to consider the electron in the hydrogen atom a wave and not a particle. Numerous experiments have confirmed that matter has wave-like nature, not only in electrically charged particles but in all particles. "All physical objects, small or large, have an associated matter wave as long as they remain in motion. The universal character of De Broglie matter waves is firmly established" [2]. This would represent another step towards the conception of the wave nature of particles. Wave like behavior, which happens at extremely small scales. This has been already proven in experiments, such as the double-slit experiment.

Heisenberg's Uncertainty Principle

The solutions to the problems discussed above were not a complete theory yet. Niels Bohr considered that physics was meant to be only for what could be observed and measured, and that everything else had no reason to be studied. Werner Heisenberg, influenced by Bohr's idea, developed a theory that focused on what was measured rather than what it was.

Alongside Born and Jordan they developed the matrix formalism for Quantum Mechanics. Although the formulation fit perfectly the experimental data, Heisenberg was well aware that he was using very abstract concepts. Despite the fact that Heisenberg struggled to understand the physical significance of the mathematical formalism, after making the calculations the final value did have a physical meaning and it matched very well the experimental measurements. His mathematics described what happened only when the measurement is done and there's no point on asking what's happening before or after.

But one of the his most important contributions was what would be later called the Heisenberg's Uncertainty Principle. He discovered that some of the quantities that could be measured were non commutative, that is

$$ab \neq ba \tag{1}$$

This quantities were linked in pairs. For example, momentum and position, or, as Dirac would discover later, energy and time. When he tried to comprehend what meant that the observables weren't commutative, Heisenberg realized that the uncertainty principle was a consequence of the non-commutativeness.

This means that if a and b are observable quantities, and $ab = ba$ then we can measure them with arbitrary precision. A common misconception of Heisenberg's Uncertainty principle is that you can't measure, for example, position and momentum at the same time. However, you can. But the more precise is the measurement of position, the more imprecise the one of momentum will be.

Dirac notation

The notation introduced by Paul Dirac is useful as it is very convenient for physicists. For a reader without these knowledge, it will be hard to grasp the explanation of the notation, so we recommend to have prior knowledge of linear algebra. Now we will explain some of the basics of the notation in order to move to the next section. Further references can be found in [3]. Let us consider a complex vector space which we will call Ket space. The dimension of the ket space is specified according to the nature of the physical system that is being studied. If it is infinite, it's called a Hilbert space. Here we will call vectors 'kets' and we'll represent them by $|\alpha\rangle$.

Kets fulfill the sum and complex scalar product:

$$|\alpha\rangle + |\beta\rangle = |\gamma\rangle \tag{2}$$

$$c|\alpha\rangle = |\alpha\rangle c \tag{3}$$

We define an observable as anything that can be measured and we represent it with an operator which in linear algebra is a linear function that maps vectors from Hilbert space to the same space. Operators act on a ket from the left.

$$A(|\alpha\rangle) = A|\alpha\rangle \tag{4}$$

The eigenvectors of the operators are called eigenkets. A given set of eigenkets $\{|a'\rangle, |a''\rangle, |a'''\rangle ...\}$ have the following property:

$$A|a'\rangle = a'|a'\rangle, A|a''\rangle = a''|a''\rangle, ... \tag{5}$$

where a', a'', a''',... are numbers. The set of numbers $\{a', a'', a''' \ldots\}$ will be summarized as $\{a'\}$ and are called the eigenvalues. The physical state of the system is linked to an eigenket and is called an eigenstate.

Now, let us define the Bra space, which will be dual to the ket space, which means that for every bra there will be a corresponding ket. Bras can be thought as functions that map kets to the set of complex numbers. Bras will be denoted as $\langle\alpha|$. There's also a set of eigenbras corresponding to the eigenkets $\langle a'|$. It's important to say that the bra dual to $c|\alpha\rangle$ is not $c\langle\alpha|$ but $c^*\langle\alpha|$. We define the inner product as

$$\langle\alpha|\beta\rangle = (\langle\alpha|)(|\beta\rangle) \tag{6}$$

The inner product is a complex valued function $\omega : V \times V \to \mathbb{C}$. For making an inner product we take a ket and a bra from the corresponding vectorial spaces. The inner product $\langle\alpha|\beta\rangle$ is a number. Note that

$$\langle\alpha|\beta\rangle = \langle\beta|\alpha\rangle^* \tag{7}$$

This means that $\langle\alpha|\beta\rangle$ and $\langle\beta|\alpha\rangle$ are complex conjugates of each other and that $\langle\alpha|\beta\rangle \neq \langle\beta|\alpha\rangle$ as it happens

in the real product. It's also easy to see that $\langle\alpha|\alpha\rangle$ is a real number.

$|\alpha\rangle$ and $|\beta\rangle$ are said to be orthogonal if $\langle\alpha|\beta\rangle = 0$. We can also normalize a ket by using the "Gram-Schmidt orthogonalization process":

$$|\tilde{\alpha}\rangle = \left(\frac{1}{\sqrt{\langle\alpha|\alpha\rangle}}\right)|\alpha\rangle \tag{8}$$

with the property $\langle\tilde{\alpha}|\tilde{\alpha}\rangle = 1$ in order to make the vectors orthonormal, which is useful for the probabilistic interpretation of Quantum Mechanics.

The postulates of Quantum Mechanics

The postulates of Quantum Mechanics are a set of agreements for how Quantum Mechanics work. If the reader wishes to dive in more in the technical aspects of the postulates, we recommend to go directly to [4].

A. First Postulate

The state of a physical system is represented by a ket $|\psi\rangle$ in the space of states, which is a complex separable Hilbert space. Since the space of states is a vector state it means that a linear combination of states is another states. The superposition of states indicates that if $|\Psi_1\rangle$ and $|\Psi_2\rangle$ are states then

$$|\Psi\rangle = a_1|\Psi_2\rangle + a_2|\Psi_2\rangle \tag{9}$$

is also a state where a_1 and a_2 are complex numbers.

B. Second Postulate

Each observable of a physical system is represented by an operator that acts on the kets that describe the system.

C. Third Postulate

The only possible result of the measurement of an observable A is one of its corresponding eigenvalues.

D. Fourth Postulate (Born's Rule)

The measurement of an observable A on a state $|\psi\rangle$ is always one of its eigenvalues and the probability of obtaining an eigenvalue a' when measuring A on a state $|\psi\rangle$ is given by

$$|\langle a'|\psi\rangle|^2 \tag{10}$$

which is the square of the inner product and $|a'\rangle$ is the corresponding eigenket of the eigenvalue a'.

E. Fifth Postulate

When an observable A has been measured and yielded a value a', the state of the system is the normalized eigenstate $|a'\rangle$. This means that the state is not the same after the measurement. As we saw from the first postulate, the particle can be in a superposition of states (because the Hilbert space is linear and accepts linear combinations). After the measurement, the wave function collapses into one of the states, $|a'\rangle$, which is an eigenstate of the observable A. This is not the only interpretation that exists, but it belongs to one of the most orthodox interpretations of Quantum Mechanics called the Copenhagen Interpretation.

F. Sixth Postulate

The time evolution of a quantum system preserves the normalization of the associated ket. The time evolution of the state of a quantum system is described by $|\psi(t)\rangle = U(t, t_0)|\psi(t_0)\rangle$, for some unitary operator U.

Schrödinger's Wave Equation

It is well known that Schrödinger didn't like the Heisenberg's matrix formulation, so he tried to come up with another formulation of Quantum Mechanics. Wave theory was well known, and waves had physical meaning, while Heisenberg's theory had mathematics that were far beyond intuition.

At the time, Maxwell's laws were already consolidated and it's description for light waves was very well known, so Schrödinger tried to find an expression that could describe the wave-like nature of particles and that also described Newton's laws. He tried with different wave expressions until he discovered one that fit in his new theory, but he encountered that his new wave was a complex one, and it raised some questions.

Schrödinger's theory was well received by the scientific community as it was simpler than Heisenberg's, but it had conceptual problems. Since the wave was complex, it had no physical meaning.

However, this mathematical artifact described amazingly well the experiments and the quantum behavior of particles. A derivation and a deeper analysis of this equation can be found on [5]. If a particle with mass m is subjected to a potential $V(\mathbf{r}, t)$, the Schrödinger equation is:

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{r}, t) = -\frac{\hbar^2}{2m}\Delta\psi(\mathbf{r}, t) + V(\mathbf{r}, t)\psi(\mathbf{r}, t) \tag{11}$$

where Δ is the Laplacian operator $\partial^2/\partial x^2 + \partial^2/\partial y^2 + \partial^2/\partial z^2$, $\hbar$ is the reduced Planck constant $\frac{h}{2\pi}$, and $\psi(\mathbf{r}, t)$ is the wave function. Notice that the equation is linear and homogeneous in ψ. Therefore superposition states are allowed, as we previously saw from the postulates. "Consequently, for material particles, there exists a superposition principle which combined with the interpretation of ψ as a probability of amplitude, is the source

of wave-like effects" [5]. When $V(\mathbf{r},t)=0$, the equation (11) becomes:

$$i\hbar\frac{\partial}{\partial t}\psi(\mathbf{r},t)=-\frac{\hbar^2}{2m}\Delta\psi(\mathbf{r},t) \tag{12}$$

Whose solution is:

$$\psi(\mathbf{r},t)=Ae^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \tag{13}$$

where ω is the angular frequency of the wave and k is the wave number. A is the normalization constant thus we make

$$\int_{-\infty}^{\infty}|\psi(\mathbf{r},t)|^2dr=\int_a^b|A|^2dr=1 \tag{14}$$

to find the value of A that will normalize the wave function. The limits of the integral of A depend on the system that is being analyzed. The relation between ω and k is given by

$$\omega=\frac{\hbar k^2}{2m} \tag{15}$$

Schrödinger thought that his formulation helped him escape the problem of interpretation that Heisenberg's theory had, but the complex numbers situation entailed too much trouble. He tried to solve the fact that the wave was complex by interpreting its intensity, as the intensity of a wave is always a real number. But the intensity gives you a distribution, how can we interpret that? Schrödinger said that if we were studying an electron, the intensity of the wave meant that there was a distribution of mass and charge of the electron, it would be distributed all over space depending on the experiment.

But experimental physicists showed that this could not be true, for some experiments didn't show up these distributions but it was seen as a punctual particle, in a specific part of space, and not as a wave [6].

Due to this problem, Born decided to propose an interpretation that is largely used nowadays: the probabilistic interpretation of Quantum Mechanics. He said that what was really happening was that the distribution wasn't a distribution of the electron's mass and charge but a probability of finding the particle in a determined state according to what was being measured, hence the need of normalizing the wave functions. For Born the wave function's intensity was the probability of finding the particle in a certain state.

This interpretation of Born's probability matched the experiments extremely well and it's similar to Heisenberg's interpretation in the sense that the wave function and Heisenberg's matrices don't describe a physical entity. They are a mathematical tool that allows us to know the prediction of what we're going to observe after the measurement.

Even if Schrödinger didn't like Heisenberg's theory, he would later prove that both theories were equivalent [7].

Appendix: Further readings

The reason why the Blackbody radiation and the Photoelectric experiments are mentioned in this reports is because they played a major role in the development of Quantum Mechanics: they hold historical importance. However, these are not the only experiments that were made at the time. If the reader wishes to read more another useful experiments, we recommend to read about the Compton effect here [8].

Another useful experiment to understand Quantum Mechanics is the Stern-Gerlach with you can find here [3]. These are not the only books that contains them and you can find them in almost any modern physics books. If you are interested in the biography of Heisenberg and Einstein (and other interesting stories involving scientists and science) you can visit this website [9]. Last but not least, if you are a Spanish speaker you can find here a good explanation of some concepts about Quantum Mechanics here [10].

Conclusion

Quantum Mechanics is an important part of physics as its a theory that helps to explain a lot of the phenomena of our world. In this work we have addressed some of the basic ideas that are introduced at college level when being first exposed to Quantum Mechanics.

Quantum Mechanics understanding (if such thing exists), or the understanding of why we can't understand it needs a formal mathematical knowledge and training. The discussion about the interpretations of Quantum Mechanics is still a relevant issue and it's addressed in numerous articles to the day, so it's understandable that these concepts may seem at first hard to get. However, whether today's scientists discuss the present theories or try to come up with their own, it's in no doubt that Quantum Mechanics is an important area of physics.

A lot of people still don't like the interpretations despite Heisenberg's and Schrödinger's theories and they disagree. However, these theories match perfectly with the experiments and they are a really accurate description of reality and that makes them essential for physics. We could even think that if a new Quantum theory emerges, it will be equivalent to these two formulations or include them, in the same way General Relativity includes Newton's laws.

Acknowledgements

The author of this report wishes to thank Carlos Morales Lóbez for his explanations about Quantum Mechanics and comments that greatly improved the manuscript.

[1] Daniel Shanahan. *The de Broglie Wave as Evidence of a Deeper Wave Structure.* [physics.hist-ph] arXiv:1503.02534v2

[2] Paul Peter Urone, Kim Dirks and Manjula Sharma. *De Broglie's Matter Waves*
`http://bit.ly/2NK3cjE`

[3] Jun John Sakurai. *Modern Quantum Mechanics.* Adison-Wesley Publishing Company, 1994

[4] R. Jaffe. *Suplementary notes on Dirac notation, Quantum states, etc.*
`http://web.mit.edu/8.05/handouts/jaffe1.pdf`

[5] Claude Cohen-Tannoudji, Bernard Die and Franck Laloë. *Quantum Mechanics Vol. 1.* Wiley, 1991

[6] Falkenburg B., Mittelstaedt P. *Probabilistic Interpretation of Quantum Mechanics.* In: Greenberger D., Hentschel K., Weinert F. (eds) Compendium of Quantum Physics (2009). Springer, Berlin, Heidelberg.

[7] Carlos M. Madrid Casado. *A brief history of the mathematical equivalence between the two quantum mechanics.* Latin Amercan Journal of Physics Education, Vol. 2 No. May 2008 ISSN 1870-9095

[8] Stephen Thorthon and Andrew Rex. *Modern physics for scientists and engineers.* Cengage Learning, 2013

[9] American Institute of Physics. *History of Science Web Exhibits*
`https://history.aip.org/web-exhibits/`

[10] Pedro Gómez-Esteban. *Cuántica sin fórmulas.*
`https://eltamiz.com/cuantica-sin-formulas/`

How To Get To Mars: A Review of Past, Present, and Future Missions

by Anisia, Yannis and Katie
Project Mentor: Özgür Can Ozudogru

This report presents an ideal trip to Mars. First, the project considers why humans should travel to Mars, then presents a detailed review of past, present, and future missions to Mars. It focuses on specific details of the mission including rocket and fuel type, the landing and habitat locations, and the crew on the spacecraft. It also touches upon the medical problems and solutions faced by humans during the mission.

Introduction

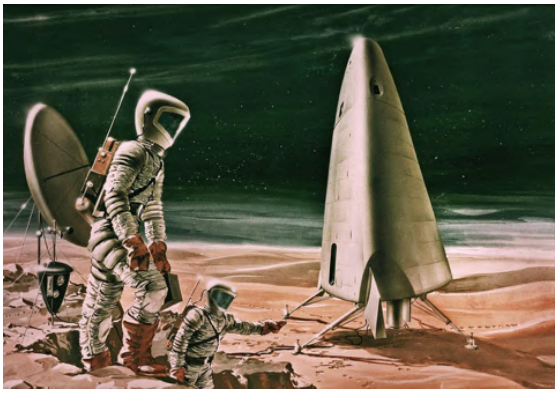


Figure 1: Astronauts on Mars depicted by NASA before the Apollo Missions

Mars is the fourth planet from the Sun and the second-smallest planet in the solar system. It is known as the Red Planet because of the Iron Oxide on its surface. The red planet was one of the 5 planets known to ancient civilizations, dating back to the Roman empire, where it represented the god of war and state. Mars has always sparked great interest in numerous fields of science. The creation of more advanced telescopes allowed astronomers and paleontologists to see Mars in greater detail; astronomer Giovanni Schiaparelli discovered the presence of linear formations on the surface, and again due to the possibly similar conditions on Mars, it was predicted to be water channels. Very little was known about the planet until *Mariner 4*, the first satellite ever to reach the Martian Sphere of Influence, sent back photos to Earth. We now know more about Mars and have discovered that it is a barren planet with a very thin atmosphere. The exploration and study of Mars can provide useful information, and so scientists have been working towards a mission to Mars for decades, and are getting closer every day.

Why Mars?

A. Why Humans Should Travel to Mars

Mars is the closest planet to the Earth both in the physical distance that needs to be traveled to reach it and in similarities with habitability. There are several reasons why humans must travel to Mars.

Firstly, we must ensure the survival of our species. As temperatures rise and resources on our Earth dwindle, humans will soon need to find a new habitat to occupy. It is the best available option, as Venus and Mercury are too hot and the Moon has no atmosphere to protect us from dangerous meteors and radiation. Mars has ice on the surface, meaning that it has the potential to support life.

Secondly, it is time to search for the presence of life on Mars. The discovery of extraterrestrial life would be a huge step forward in science. This discovery would enable humans to study organisms from other planets, and we may find new cell structures and ecosystems. We may even discover different genetics that we may learn from or use for our advancement. It gives us a comparison of our lives and may alter how we choose to live our lives.

We may also use the scientific and technological discoveries from the exploration of Mars to improve our quality of life on Earth. It is common for innovation in one field to inspire innovation in another. For example, in 1993, astronomers created a new algorithm to extract clearer images from the Hubble telescope. Later, a medical doctor discovered that this same algorithm could be used to detect the early stages of breast cancer. The trip to Mars will test our resourcefulness, our creativity, our knowledge, and our abilities in every way.

B. Dust Storms

Dust is a critical component in the Martian atmosphere. It alters the atmosphere's circulation by heating or cooling it and is then redistributed around Mars by atmospheric winds. In this dust cycle, dust storms play a particularly important role. These storms are problematic, as they endangers not only the astronauts but also their equipment and energy sources. Celestial dust exposure is dangerous for crew members, especially after extravehicular missions. The spaceship's design uses

thin solar panels, which are light and can be easily transported, to bring in energy to power the spaceship. The dust can decrease the electrical power from the spaceship's solar panels. Since the dust particles are tiny and electrostatic, they stick to the solar panels. The dust lodges behind the panels and clogs up filters, blocking the solar cells from receiving light from the sun unless they are regularly cleaned.

Measures can be taken to avoid the failure of solar panels during and after dust storms. During the duration of dust storms, operations are severely limited while the astronauts wait for the storm to subside. While the power is significantly decreased during dust storms, there is still enough to keep the crucial systems functioning. To conserve energy, non-critical systems are shut off. Water and Oxygen production is turned off, and the spaceship will use water and Oxygen from storage tanks, greenhouse lighting is dimmed, and rover operation is limited. If necessary, the rovers can be used to remove dust from the solar panels. The effects of the dust can also be limited through electromagnetic screens. These screens have high-voltage, multi-phase electrodes that produce an electromagnetic field which decrease the electrostatic properties of the dust. This makes it harder for the dust to stick to the astronaut's suits and the surfaces of the spacecraft. If the solar panels fail, the spaceship would have alternative energy production systems.

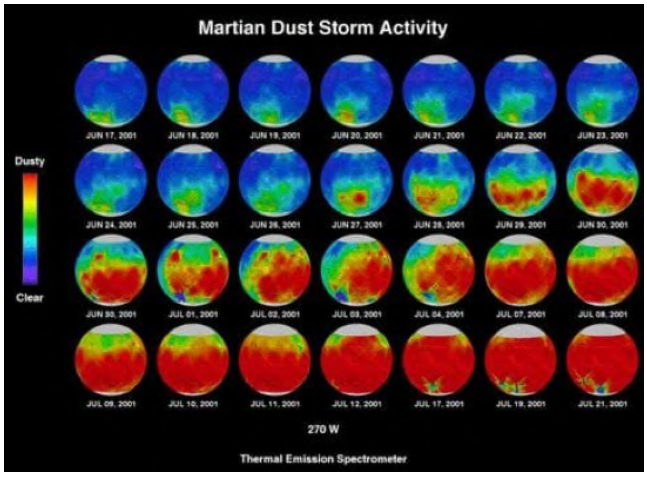


Figure 2: Martian dust storm. Image taken from TES team, MGS, JPL, NASA

To protect the crew inside the spaceship, air filters remove dust from the air to prevent it from damaging both the astronauts and the internal systems. Astronauts will be trained to cope with the dust, such as using algorithms to predict dust storms to avoid the outside atmosphere during a dust storm warning. Furthermore, to avoid skin contact with the spacesuits, the suits should remain outside, hence they would be attached to the habitat's wall, and the colonists would be able to climb directly into the habitat.

C. Geology and Climate of Mars

1. Geology on Mars

Mars is a terrestrial planet that consists of materials of Silicon and Oxygen, as well as metals that make up the rock. The Southern hemisphere of Mars consists of cratered highland terrain, while the Northern section of Mars contains lighter cratered, volcanic plains. The surface is mostly composed of extrusive, low-silicon igneous rock, although some parts of Mars have higher quantities of silicon. There are large regions with low albedo, which contain silicon sheets and high-silicon glass. Much of the surface is also covered in finely grained Iron Oxide dust. Mars also has a ton of elements with low boiling points, such as chlorine, phosphorus, and sulfur. These elements are much more common on Mars than the Earth, probably because Earth's proximity to the Sun has boiled away these elements.

The rocks on Mars also give clues to the series of events that took place on Mars. The most important information is about the composition of rocks that can only be created in the presence of water, meaning that Mars once contained liquid water. The presence of liquid water indicates that there was once life on Mars.

Mars is home to many huge volcanoes, which can be 10-100 times larger than those on Earth. This is because the crust on Mars does not move the way it does on Earth, so all of the lava piles up into one big volcano. Many of these volcanoes became inactive a billion years ago. The largest volcano on Mars is known as Olympus Mons, which is 100 times larger than the largest volcano on Earth. This volcano is much younger than most others, becoming inactive only hundreds of years ago, rather than a billion. Olympus Mons may be intermittently active even today, which is a possibility that future Mars exploration missions might want to consider. Most of these volcanoes are in the Tharsis Bulge, a highland region in Southern Mars.

Apart from the high presence of volcanoes in the Tharsis Bulge, there are many other interesting geological features. The surface of the planet has bulged upwards as a result of the high pressure below. This pressure creates large cracks in the surface of the planet. The largest canyon is known as Valles Marnier is, which is 5000 kilometers long, almost a quarter of the way around the planet. The previous presence of water on Mars has helped shape Valles Marnier is by seeping from deep springs and undercutting the cliffs. This undercutting led to large landslides on Mars. Today, the primary erosion on the surface of Mars is caused by wind.

2. Climate on Mars

The weather on Mars can vary from -153 degrees C (at the poles) to 20 degrees C. Son l temperatures at the *viking* lander site range from -17.2 degrees C to -107 de-

gree C. The highest soil temperature approximated by the *viking* Orbiter was 27 degrees C. In the shady parts of the planet, the Spirit rover recorded a maximum daytime air temperature of 35 degrees C, and often recorded temperatures well above 0 degrees C, except in winter. Because of Mars' lack of atmosphere, the friction between atoms is lower, resulting in slower heat exchange, meaning that it feels less cold on Mars than on Earth at the same temperature. Because of the low concentration of water vapor, rain on Mars is not likely to happen. Snow, on the other hand, is possible. Mars has clouds that sometimes produce dry ice snow. When the temperature increases, the ice sublimes, transforming into gas. This repeated process forms creases in the Martian surface.

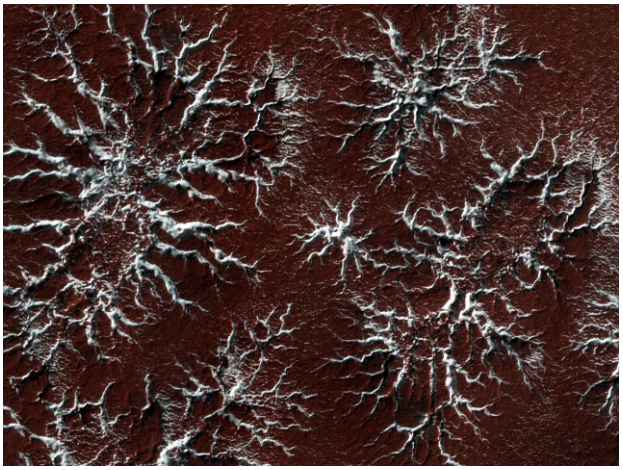


Figure 3: Creases on Mars. Image taken from Mars Reconnaissance Orbiter

One of the reasons for the lower temperature on Mars is its increased distance from the sun compared to Earth. At the poles, during winter, the temperature decreases enough to turn Carbon Dioxide from the atmosphere into dry ice.

Mars is a spinning, windy planet. The wind speed can reach up to 100 km/h, producing great dust storms. Although the wind is strong enough to produce dust storms, the atmospheric pressure of Mars is about 1 percent of Earth's sea level pressure. For a human, experiencing a 100 km/h Martian storm feels like a light terrestrial breeze.

A recent discovery by the *Mars Global Surveyor* has shown large areas of magnetic material on the surface of Mars. Although Mars does not currently have a magnetic field, this means that there was once a magnetic field around Mars. This is probably because the core of Mars is pure solid metal, and there are currently no liquid elements in Mar's core to conduct electricity. Magnetic fields are created by the movement of liquid metals in the core.

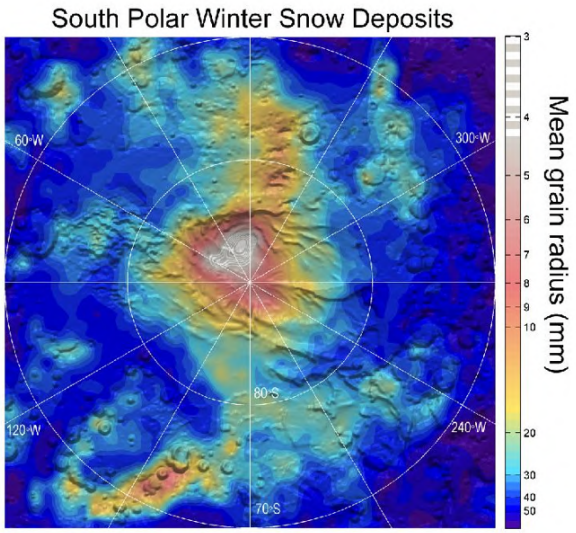


Figure 4: South polar winter snow deposits. Image taken from NASA

D. Colonizing Mars

Colonizing Mars is possible for several reasons, including the similarities between the Earth and Mars and the prospect for generating food, Oxygen, and building materials. In all, Mars can be nearly self-sufficient, once the population is big enough (eg. in the millions) to supply the required division of labor. According to Elon Musk, there would be an economy based on real estate. With the growing population and shrinking available land on Earth, on a new planet with plenty of room. There is also likely to be a market for research of Mars, which could help fund colonization.

One possibility is to build connecting habitats underground. According to data collected by the *Mars Reconnaissance Orbiter*, there are habitable environments below the surface of Mars. This solution would eliminate the need for Oxygen tanks and pressure suits when colonists leave their homes, as well as protecting the colonists from dangerous radiation exposure. The soil on Mars could also be baked into bricks to create solid structures, helping to build homes underground.

It may also be possible to grow food on Mars. The atmosphere on Mars is thick enough to protect plants from a solar flare, meaning that thin-walled plastic greenhouses are enough to protect and nurture crops. The greenhouse effect of the greenhouses would be perfect for creating a habitable environment for the plants to grow. The length of Martian days is almost the same as terrestrial days, which is familiar to our plants. Colonists could potentially gather water from Mars as well. There is a large quantity of water in the form of ice, and studies show that if all the permafrost and ice on Mars was melted, there would be enough water to create an ocean 100 meters deep. Most landers have found that there

may be water in the form of ice beneath the surface of Mars as well. If found, this ice could be extracted and purified for human consumption.

Mars has higher quantities of Deuterium than the Earth has. Deuterium is the key fuel for fusion reactors, meaning that Mars has the potential to create large amounts of nuclear energy.

Previous Missions to Study Mars

There are 3 main types of unmanned vehicles that have been sent to Mars: **Landers**, **Rovers** and **Orbiters**; each of the 3 types of vehicles is advantageous in different aspects of studying Mars.

A. Orbiters

The first vehicle to ever take a picture of Mars was an orbiter called *Mariner 4* during a mission in 1965. *Mariner 6, Mariner 7 and Mariner 9* were the first vehicles to take significant scientific data about Mars, including information about craters, canyons, and the presence of dust storms and volcanoes. *Mariner 9* photo-mapped the entire surface of the planet and revealed relics of ancient river beds covering the surface.

Mars Global Surveyor, *Mars Odyssey*, *Mars Express*, *Mars reconnaissance orbiter*, *Maven* and the *ExoMars program's orbiter (ExoMars 2016)* are all successful missions that further expanded our understanding of Mars' atmosphere and compositions. These missions continue to bring new data about the surface and atmosphere of Mars to this day.

B. Landers

One of the many programs by NASA to explore Mars, the *discovery program*, launched the *InSight* Lander in 2018. Its purpose was to study the crust, mantle, core and tectonic activity of Mars, as well as further studying the atmosphere and surface by measuring temperature, pressure, and wind at different times throughout the year. *InSight* was launched along with 2 small satellites called Cubesats. These satellites, known as *Mars Cube One (MarCO)*, tested a new kind of deep space communication. Both successfully transmitted data from the *Insight* lander during its entry and landing on the Martian surface. The success of this technology provided a cheaper and easier alternative to direct communication from the Martian surface to Earth.

Only 3 other landers successfully landed and gathered data from the Martian surface. The first two were the *Viking 1 And 2* landers that were launched in 1976. Their main mission was to take the first-ever pictures from the Martian surface, as well as scan it to find the presence of life on Mars. They discovered that the Martian soil presents chemical activity, even in the absence of any microorganisms near the landing sites. The ultraviolet radiation and the oxidizing nature of the soil result in a self-sterilizing property. The only other successful lander before insight was *Mars Phoenix* which landed in the northern ice cap in May 2008. It drilled into the ice to discover if Mars was capable of sustaining life. It also took soil samples and found their chemical compositions. This information was used to understand how soil composition varied depending on its location on Mars.

C. Rovers

Rovers have great advantages compared to orbiters and landers. Since they can move on the surface, it can take a greater number of significant measurements whereas landers are more specialized in their instrumentation. However, with the development of energy production for the rovers and the efficiency of rockets, rovers can conduct a wider variety of experiments in a larger area than a lander.

Mars Pathfinder, launched in December of 1996, was the first rover to land on the Martian surface. Similarly to the *Marco satellites*, the purpose of the rover was to demonstrate and test new technological development. The rover surpassed expectations and lasted a greater time than its expected mission life, managing to send data about the Martian atmosphere, climate, geology, and composition of various rocks through 15 chemical composition tests during its journey.

In January of 2004, the twin geologist rovers *Opportunity* and *Spirit* were launched. They both achieved more than could ever be expected, as they both had an expected mission time of 90 days, but *Spirit* lasted 6 years before being abandoned when it got stuck in sand, and Opportunity continued to explore the surface of Mars for longer and further distances than any previous extraterrestrial rovers before losing communication. Thanks to these two rovers, proof that there was once proper lakes, rivers and possibly seas on Mars was discovered. Using this information, among more data transmitted, we have been able to deduce that conditions on Mars were once suitable enough to sustain microbial life.

The latest rover to be sent to the Martian surface, the *Mars Science Laboratory (Curiosity Rover)*, was launched in 2011 and, as of August 17, 2019, has completed its 2500th sol (a day on Mars). *Curiosity* was the largest unmanned vehicle to land on an extraterrestrial surface and being so heavy, it could not land with the same method as all previous rovers. Instead, an entirely new method was conceptualized, designed and successfully tested (see section 8 "From Earth to Mars"). The project was initially launched by NASA, and in collaboration with numerous international and national partners such as ESA, Aerojet, JPL, etc. Curiosity had many objectives of varying importance and difficulty:

1. To test the new Martian landing system and prove

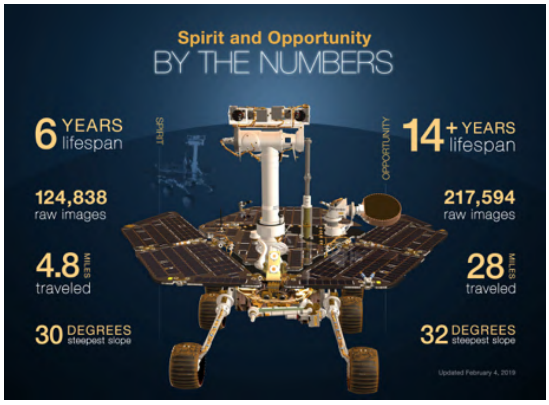


Figure 5: Facts about the twin rovers 'Spirit' and 'Opportunity'. Image taken from NASA

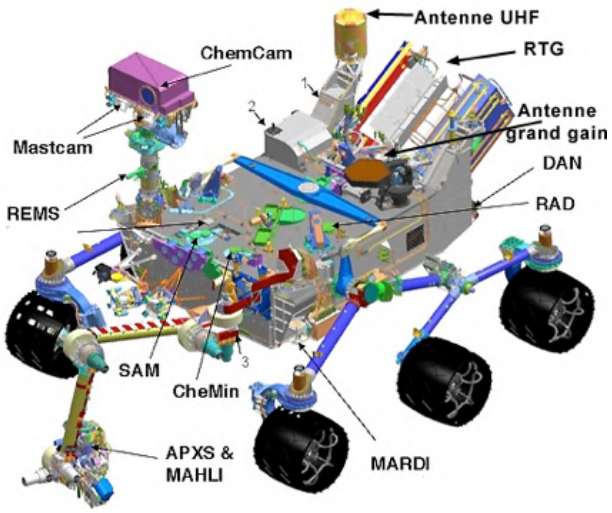


Figure 6: Details about the instruments on the Curiosity Rover. Image taken from NASA/JPL

its effectiveness and accuracy.

2. Determine the habitability of the area it landed to prepare for future manned missions. It is the only rover to take radiation readings from within itself as a means of determining how dangerous solar (and other) radiation could be even with protection.

3. Further study the climate and geology of Mars through investigation of:

1. Organic compound and chemical building blocks such as Carbon, Hydrogen, Nitrogen, Oxygen, Phosphorus, and Sulphur

2. Biosignatures

3. The cycle of water and Carbon Dioxide within the atmosphere Characterization of the spectrum of surface radiation

D. Failures

All these vehicles had purposes that each expanded our knowledge of Mars by practicing autonomous and different methods of landing, finding proof of liquid water and habitability, exploring the possibility of the previous life, determining the chemical composition of the air and soil, and learning about the planet's geology. However, in order to achieve so many successes in Martian Exploration, many failed attempts were especially useful to the learning process. With each failed mission, engineers have been able to learn more about the difficulty and tactics required to successfully land on the surface.

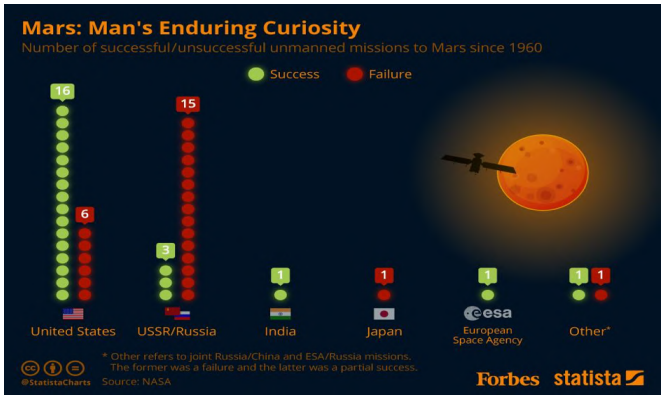


Figure 7: Number of (un)successful missions to Mars since 1960. Image taken from STATISTA

Future Missions and programs

There are 7 planned upcoming missions, 4 of which will launch in 2020 with a planned arrival of late 2020 or early 2021. 2 of these missions will be a first for the space program; the United Arab Emirates (UAE) and China, will be launching the *Hope Mars* mission (orbiter) and the *Mars Global Remote Sensing Orbiter and small rover* respectively. NASA will be launching its *Mars 2020 rover and helicopter*, and ESA with support of the Russin Administration of Science will be launching the *ExoMars 2020 rover and lander* as a continuation to the already in-orbit *ExoMars orbiter*, as part of the *ExoMars* program. Japan is planning to launch *Mars Terahertz microsatellites* in 2022 and 2024, an orbiter and lander to one of Mars' moons, Phobos, called the *Martian Moons Exploration*.Moreover, India is planning on launching its second orbiter as part of its *Mars Orbiter* program, called the *Mars Orbiter Mission 2* in 2022.

A. Mars2020

The *Mars 2020* rover is planned to be launched mid-July of 2020, to land in the Jezero crater of Mars. It is

similar to the Mars Curiosity Rover in size, mass, capability and hence, its main objectives will be to assess the possibility of past and future habitability of the planet, as well as the potential to preserve biosignatures within geological materials. This mission will have the very distinct new features; the rover will be able to take small samples of martian soil or rocks that are processed, and then stored in small containers that it can either hold on to or drop along its route. These containers will then be able to be recovered by a future mission and ideally returned to Earth for examination before human exploration. This would be especially useful in learning about the fertility of this soil and how it reacts to the presence of various elements, rare in the martian atmosphere.

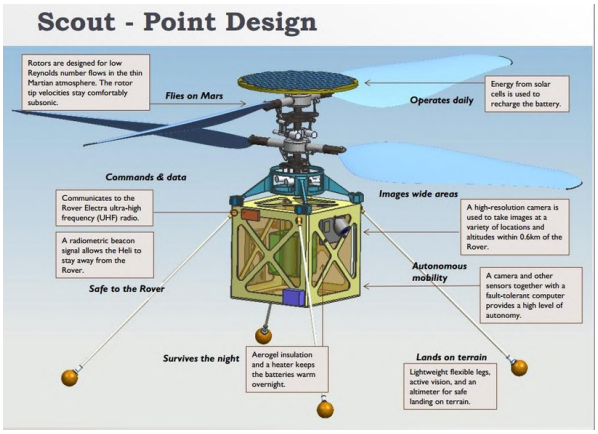


Figure 8: Diagram of MHS. Image taken from NASA

The second new addition to this is the *Mars Helicopter Scout (MHS)* which is a 1.8 KG (4.0 lb) vehicle with blades spinning 3 times faster than a multi-tonne helicopter on Earth as the atmosphere is so much thinner, much more lift is required to be produced by the blades regardless of the severe decrease in mass. The goal of this helicopter is to be able to map out the route the Rover should take, and find certain interesting rocks and patterns that the Mars 2020 rover should investigate.

B. ExoMars2020

ExoMars was a long term project between the European Space Agency (ESA) and ROSCOSMOS. The first part was the ExoMars Trace Gas Orbiter which was part of the 2016 mission and will be responsible for supporting communication between the 2020 rover, and mission control at the Rover Operations Control Centre (ROCC) in Italy. The rover will be sent to a site on the Martian surface, with a high potential to find well preserved organic materials. Due to the lack of a thick atmosphere on Mars, solar radiation is a large issue in regards to preservation of organics, hence, the *ExoMars 2020* rover is equipped with a drill capable of going as deep as 2 meters, and is also equipped with an infrared spectrometer to determine the type of minerals within the holes it digs.

C. *Artmeis* program

One of the current programs operated by NASA, the *Artmeis* program, is one of the largest steps towards human exploration of Mars. The program's first step is to land the next man and the very first woman, on the surface of the Moon, by launching them aboard the new powerful rocket created by NASA, the Space Launch System (SLS). The crew, aboard the *Orion* spacecraft, will dock to yet another project conducted by NASA, *the Gateway. The gateway* is a new international space station that will be set in an orbit around the Moon, it will consist of parts assembled and launched by various research companies and space programs; the goal of this space station, is to allow astronauts to remain near the Moon for a much longer amount of time as all the Apollo Missions, and more importantly, it can allow crew to go down to the lunar surface multiple times. However, Lunar exploration is not its only goal as it will allow for testing of new landing technologies on extraterrestrial surfaces and deep space docking, crucial towards the ambitious 2028 manned Martian missions by NASA.

From Earth to Mars

Traveling to Mars is an extremely difficult process in which there are many aspects to consider before leaving. The first biggest question that needs to be answered is whether the mission would be a one-way mission, or have a return trip. Depending on the type of mission, it is then possible to decide which landing method is best suited, and more importantly how much propellant is required to be able to return if need be.

A. Energy Needed

To this day, the largest payload ever sent had a mass of 1 metric tonne (the *Curiosity Rover*). The heavier the payload, the more energy would need to be expended in changing the velocity of the payload which requires more propellant which in turn would further increase the mass. Thankfully, the spacecraft does not need to burn propellant until it reaches Mars as there is no air resistance in the vacuum of space, which means that if it accelerates up to a precise velocity, within 4-6 months it will eventually arrive within the Martian sphere of influence. In which it can then either orbit Mars or one of its 2 moons or enter the atmosphere and attempt to land.

In order to have a better understanding of 3 dimensional maneuvering, and relative speed in an orbit, we look at the spacecraft's velocity, and depending on the change in velocity, represented by delta v, we can determine information about the spacecraft's orbit, such as

whether it is a closed orbit, or if it will have enough velocity to escape the sphere of influence of the body it is orbiting.

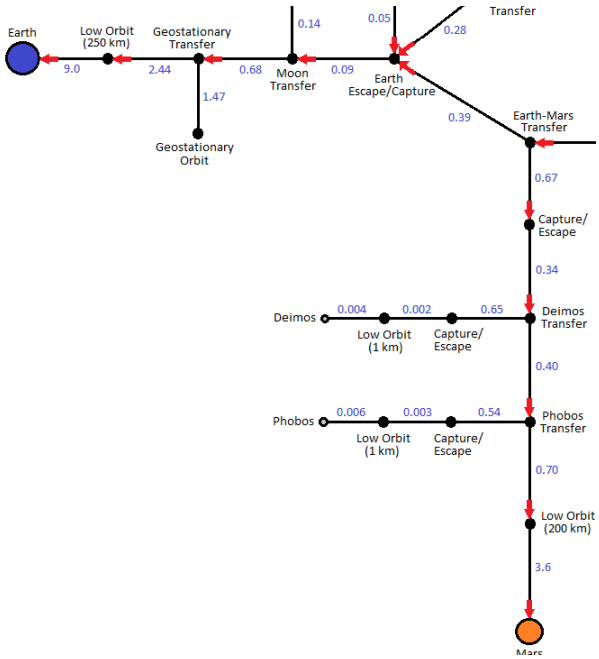


Figure 9: A Delta-V map of the journey to Mars. Image taken from NASA

According to the above diagram, a journey to the surface of Mars, from Earth, would take an approximate 12.6 Kilometers/second change in velocity (delta v); around 3/4 of all this increase in velocity is actually spent escaping Earth's atmosphere and forming a Low Earth Orbit (LEO), once in orbit however, another 2.44 Km/s of delta v will be needed to increase the apogee (the highest point of a non-circular orbit around Earth) up to the point known as the geostationary transfer point. After spending so much energy increasing the height of the orbit around Earth, a relatively small ΔV of less than 1 km/s is required to adjust the trajectory of the spacecraft for it to increase the apogee far enough for it to reach the escape velocity of Earth, and then further increase the aphelion (highest point of an orbit around the sun) in order for it to be at the same or similar altitude as Mars would be at. With enough precision, the spacecraft can arrive directly into the Martian atmosphere without hitting the planet, which can then be used, along with the assistance of one or multiple other means, to slow down the vehicle enough to land on the surface. Once the trajectory has been set, the vehicle then just needs to coast for 6-8 months until it reaches the Martian sphere of influence, the time greatly depends on the location of the 2 planets in relation to each other, however the ideal position of the planets would be at a 44 degree angle in relation to the sun.

The most commonly used method of interplanetary transfer and the most efficient and simple is called the

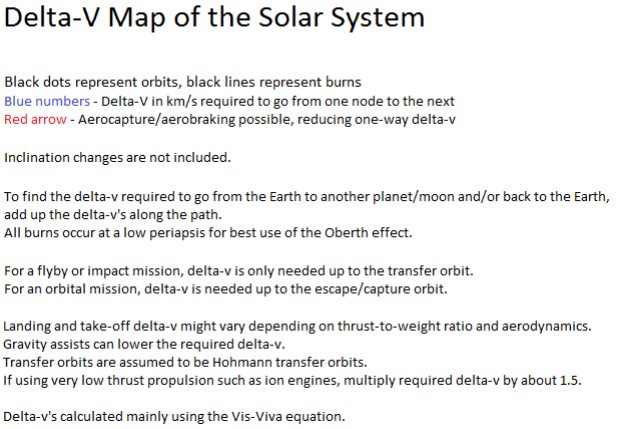


Figure 10: Instruction on how to use the above Delta-V map

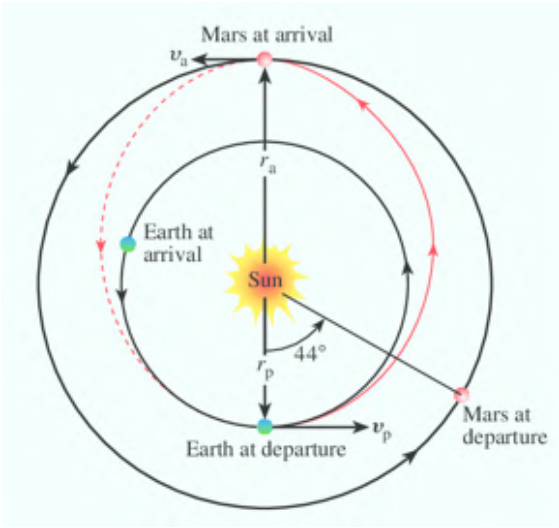


Figure 11: An example of the Hohmann orbit transfer trajectory. Image taken from ResearchGate

Hohmann transfer orbit. This is illustrated by the red orbit path in the image on the right (Note: Earth's and Mars' orbits are not perfectly circular as depicted). This method involves increasing the aphelion up to the altitude of Mars' orbital path but keeping the perihelion (the lowest point of the orbit around the sun) at the same altitude as that of Earth's orbital path.

B. One vs Two Way Missions

In a one-way trip to the surface of Mars, a lot less propellant would need to be carried which would leave a lot more space for supplies and people. This means that each astronaut would have to choose between a new life on Mars, or keeping their old life Earth. A one-way trip

would also require a lot more preparations in order to ensure the safety and comfort of the colonists once on the planet and hence, in addition to other reasons including resource management, supply missions, and design projects, the price would increase significantly. The advantages however of the first humans to step on Mars to go there on a one-way trip, is that it would allow for a much greater number of people to be present and at once on the planet, hence increasing the amount of research that can be done, and the number of fields that can be researched, as well as the safety of the colonists through means such as bringing actual trained medical personnel as opposed to astronauts with a brief medical training for emergencies.

In a two way mission, all the crew will have the opportunity to return to Earth, either in the case of an emergency or when it is time to return as the return vehicle will be ready at all times. Furthermore, supplies would need to last the crew 2 years as 1 year would be spent traveling to Mars and back and the other would be spent staying on Mars. Any mission to Mars is optimum to launch within a small window once every 2 years, hence, it is only optimum for a vehicle to travel back to Earth from Mars within a specific window once every 2 years. Luckily, the delta v required to get off the surface of Mars and on a trajectory back to Earth is about 1/2 that of the Earth-Mars journey; it would take about 6.1 km/s of delta v and after that the six month coast would lead the spacecraft back into the dense atmosphere of Earth, which would be capable of, similarly to the Apollo Missions, slow down the vehicle and allow for a safe landing.

C. Landing Methods

One of the most complex matters in traveling to Mars is the approach and landing. 3 main ways of landing that have been used before on the Martian surface. Each works best for different types of vehicles, and sometimes a combination of 2, or even 3, methods have been used. Vehicles entering the Martian atmosphere directly from a transfer orbit from Earth, will be entering with a velocity of approximately 7.3 km/s, and depending on the size, mass and angle of attack of the vehicle, it would take roughly 3 minutes of burning in the atmosphere to reduce its speed to its terminal velocity of usually around 400m/s. Even though the Martian atmosphere is much thinner than that of the Earth, the heat produced from the friction between the Carbon Dioxide, Nitrogen, and dust, acting on the rapidly moving spacecraft, is enough to produce temperatures as high as 2100 degrees Celsius, hence, it is necessary to have some sort of heat shield to protect the spacecraft from the extreme temperatures.

1. Bouncing Airbags

This method has been used to deliver the *Opportunity* and *Spirit* rovers as it allows for an accurate and effective way of placing the rover on the surface. The method includes an aeroshell burning through the atmosphere until the vehicle reaches its terminal velocity, and then the aeroshell separates and the vehicle descends towards the surface in 2 parts, the crane and the rover. When close to the surface, retro thrusters are activated slowing down the vehicle to almost 0 m/s just above the surface, and the rover detaches and immediately inflates large airbags all around itself in a triangular based pyramid. The decrease in mass means that the crane will accelerate upwards rapidly and eventually crash further away, barely contaminating and disturbing the landing zone, mean whilst the rover keeps bouncing and rolling until it stops, deflates and opens up to let the rover out. This method, unfortunately, would not be able to be used for a manned spacecraft, for reasons ranging from the sheer mass of the vehicle(the 1-tonne *Curiosity* rover was too heavy for this system), to the dangers and effects that bouncing and rolling on the ground would have on the crew.

2. Parachutes

Almost every single astronaut to have ever flown into outer space has always landed back on Earth by parachute. The very first man in space, the famous Russian Yuri Gagarin, completed a suborbital flight, and after re-entry, was ejected from the capsule and landed with a parachute on foot. Apart from these exceptions and the Space Shuttle's landing, every crewed capsule to re-enter the atmosphere (Eg. the *Gemini* capsules and the Soyuz crew capsule) have deployed parachutes to land. One issue with this method is that its extremely inaccurate, which would be an issue for a one-way trip to Mars, as the astronauts would be transporting a considerable amount of material with them, that would need to be brought to a pre-built habitat. Secondly, the landings are quite rough, which is why most capsules land in the ocean and get recovered as the water can soften the landing, however, Russians have integrated a solid fuel burning system that ignites just before touchdown to have the same slowing down effect as water; the reason for the rough landing is because the atmosphere can only slow down so much, as the capsule has quite a great mass. On Mars, however, the atmosphere is approximately 1 percent that of the Earth so a parachute would only be able to reduce the speed after or towards the end of the entry. However, they would not be sufficient to land the vehicle even remotely safely on the hard surface of Mars.

3. *Propulsive Landing*

Propulsive landing is the ideal way of landing any vehicle on the Martian surface. It has been used to land all the immobile landers on the surface of Mars and has also been used to land the latest MSL (*Cuiosity Rover*). The same method will also be used for the 2 newest large rovers launching summer 2020. All these rovers have and will use a system called a Skycrane. Skycranes are modules containing both propellant and thrusters that are used to slow down the entire vehicle, whilst the rover is attached by cables underneath. The Skycrane slows down the whole vehicle to hover over the surface, and the rover is lowered by a winch until it touches down on the surface and detaches the cables, allowing the Skycrane to accelerate upwards and crash outside the landing zone. The issues associated with using this for a manned mission is that it is not reusable and would require a lot more fuel and much more advanced engines, which could potentially raise the center of mass of the whole system to above the midpoint of the system, resulting in a top-heavy craft that would be much less stable in reentry. However, the best alternative would be to propulsively land with engines below the vehicle. It is the hardest method as, unlike the Apollo Missions where the crew manually controlled the LEM, With the time delay of between 5-30 minutes with Mars, the increase in gravity, and complexion and size of the rocket, it would need to land automatically. However, with the rate at which technology is advancing, and with enough practice on Earth, it should be easily achievable. A clear example of this is the Falcon 9 booster made by SpaceX, which with enough practice, they are now able to automatically land an orbital class booster within a 1-meter range of the desired spot, SpaceX CEO Elon Musk hopes to launch a new rocket shortly with the capability of landing propulsively on Mars.

Preparations Needed

A. One and Two Way Mission?

Many steps need to be taken in the planning and testing phases before sending humans to step on the Martian surface. However the very first decision that needs to be made, is whether the mission should be a planned return mission, or a one-way/long term stay; the advantages and disadvantages of the 2 types of missions have already been discussed, but the habitation that the crew would live in has not been discussed yet. There are 3 possible solutions:

1. The crew would live within the interplanetary transport vehicle, which would be useful as only what is necessary would be brought on the journey, however, this would greatly decrease the number of crew members, and their comfort, as the single-vehicle would need to carry supplies for approximately 2 years.

2. Multiple vehicles could be sent to arrive at the same location, the distribution could be split with the crew in one and cargo in the others, or crew and cargo distributed among all the vehicles. This would pose some issues however as the vehicles would need to land very close to each other to minimize travel distance between them and facilitate cargo transfers, which means that they will either reenter at very similar times which could be very dangerous if there were to be a malfunction with one of them and make debris that could interfere with the landing zone or entry of another of the vehicles. Furthermore, the landing of the vehicles in such close proximity and the departing of them too, could damage the other rockets or create large dust clouds that due to the low gravity, may not settle fast enough before the next vehicle was to land.

3. Unmanned missions could be sent ahead of time, to automatically construct a sturdy habitat, that would be suitable for long term stay, or even indefinite stay. Many Martian habitat concepts have been designed by both space programs, and competitions, all with the necessary equipment and rooms, but often additional parts and equipment that could prove extremely helpful. Afterward, once the habitation has been constructed, a crew can be sent there, and, granted the landing could damage or affect the area, the habitation will be prepared to combat dust storms worse than the landing vehicle, and will have the equipment necessary to clean the harmful dust and potential debris from the area.

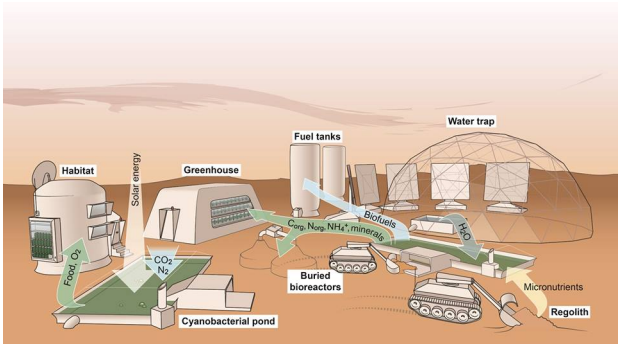


Figure 12: A habitat design. Image taken from NASA illustrators

B. Habitat Location

Once the type of mission has been decided and what kind of habitation the crew will live in for a given period,

the next stage is to be able to decide where the habitat should be constructed (or where the landing site should be). Thankfully, with all the orbiters currently still functional, detailed maps of the Martian surface have been made, and with the 4 previous rover missions, and the 2 largest upcoming rover missions, more and more ideal landing and habitation areas will be discovered and proposed during mission planning. The location of this habitat is crucial, as it needs to ensure an easier landing and departure, hence meaning the equator would be ideal, such as when we launch satellites and missions on Earth. Theoretically, as long as the mission has a long enough duration, a rough radius of approximately 100Km around the habitat could be explored by crew using manned rovers, similarly to the last few Apollo Missions on the Moon. There have already been multiple suggested ideas for the landing site of the future Mars2020 and ExoMars rovers, however, these rovers can only land in and explore one of these landing sites each; the other landing sites are just as important and possibly the perfect place for a habitat.

1. Northeast Syrtis

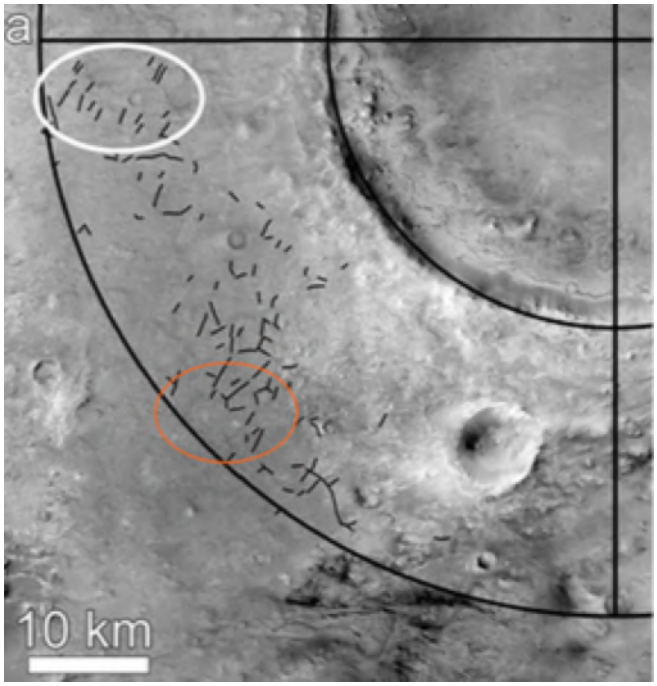


Figure 13: Image of the northeast Syrtis, mapped within 1 crater radius of the south west rim of Jezero crater. Image taken from the Planetary Science Centre

Located in the Northeastern section of Syrtis Major, the area is quite rich in mineral diversity. There is access to geological environments such as Sulphate bearing terrains and clay mineral deposits and Carbonated olivine minerals, all of which hint at the once habitability of Mars, but would greatly improve our knowledge of Mars' history. The landing zone is ideal both for the rover and for a habitation to be constructed as its relatively flat, low and has plenty areas of interest to examine. The entire Syrtis Major area is also rich in silicon, which can be used for making glass, circuits, 3D printer resource.

2. Jezero Crater

The Jezero crater is also located in the northeastern section of Syrtis major and may look close on a map. However, it is still very far, and the steepness of the crater's hills would make it extremely difficult for a rover or manned rover to traverse. Considering that Mars 2020 rover could travel a maximum of approximately 100 meters per day (with minimal research done), it would not be a great idea to attempt to examine both zones with the rover as it could take multiple years (increasing risk of communication lost) for it to reach the other area. The crater was once home to a River Delta, and it was probably capable of supporting life in the form of microorganisms. "To the extent that ancient lakes and deltas were habitable environments, and that we believe them to now preserve traces of ancient life, Jezero is the best choice among the remaining sites," says John Mustard, a planetary geologist. Furthermore, as the crater is located in the Syrtis area, it is also very rich in multi-use silicon.

3. Colombia Hills

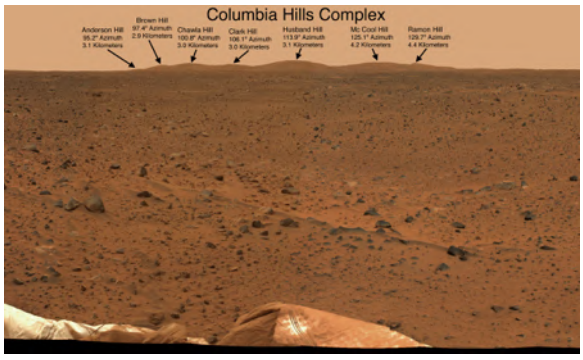


Figure 14: Image of the Columbia Hills taken by Spirit rover

The Columbia Hills are located in the Gusev Crater, almost on the opposite side of the planet to the Syrtis major area. The crater was the landing zone of the Spirit rover which traveled 3 km to the Columbia hills where it took pictures and examined the hills in the crater, of which each of the 7 peaks were named after the crew lost in the Space Shuttle Columbia disaster. The area is super rich in Sulphur minerals, and it also has a considerable amount of silicon, however not as much as in the Syrtis area.

4. Oxia Planum

Oxia Planum is a very low altitude plain, located near the equator in the northern hemisphere. It was chosen as the landing place for the *ExoMars2020* rover because of its low altitude, as it makes it ideal for parachuting landing (as the atmosphere is slightly thicker 3000m below average level). The area is also rich in iron-magnesium and has a valley system that proves that water was once present there, allowing for an increased chance in biosignature preservation. It is, however, much less Silicon rich.

Another issue that needs to be looked at is the communication network. Just like Earth, Mars rotates about an axis, meaning that half the sol, a habitat on the surface would have direct access to communicate with Earth, and the other half of the sol, the habitat would not be able to communicate at all with Earth. The best solution would be to rely on communication satellites as we do on Earth. To be efficient, these satellites should be organized in a triangle formation. However, if one satellite fails, the network could fail and communication with Earth would be lost as the habitat would likely lack a sufficiently powerful antenna to communicate directly with Earth. As previously mentioned, NASA tested, during the insight lading, 2 small cube stats called *MarCO*, which both successfully proved the capability of the smallest industry in deep space communication.

C. Vehicle Testing

Once the type of mission, habitat, and location have all been chosen, the difficult task of designing and testing a new rocket capable of completing the various tasks it needs to. The process of designing and constructing the prototype of a new rocket already takes a substantial amount of time and money. When the vehicle is planned to be able to transport crew, it requires a lot more time as various tests need to be conducted for the Federal Aviation Administration (FAA) to approve the rockets' safety. These tests often include multiple proofs of soft landing, stress tests and of course abort tests that should be able to take place at any point during the vehicle's journey (eg. on the launchpad, during liftoff or even in orbit). This vehicle will need to coast for six months twice, before performing a re-entry and propulsive landing both times, and it will also need to stand on a foreign planet for at least 1 year and still be fully functional and safe. Therefore, the amount of testing this vehicle will have to go will take years.

D. Crew Members

In Apollo 17, there were 3 crew members: Lunar Module Pilot, Commander, Command Module Pilot. The lunar module pilot is responsible for landing the spaceship on the surface of Mars. The commander was in charge of the coordination of the crew and the payload. The command module pilot was in charge of navigation and guidance. The Apollo 17 mission was very successful. It had the largest assortment of lunar samples which led to a fresh understanding of the Lunar Surface, the longest period spent in lunar orbit and on the Lunar Surface, and the largest amount of extravehicular activities. The *STS-1* mission was the first the *Space Shuttle Program*. The purpose of this mission was to demonstrate a safe launch, orbit, and return of the reusable orbiter vehicle.

This mission had a crew of two: a Mission Commander and a Pilot. While there were a few anomalies in the mission, in the end, it was very successful without any major issues and with a safe and successful landing. Future Space Shuttle missions also included a Payload Commander, Mission Specialist, Flight Engineer, International Mission Specialist, Educator Mission Specialists, Payload Specialist, and a Manned Space-Flight Engineer.

Impact on Humans

A. Social

Social isolation is defined by the absence of social contact or a state in which the number of social encounters can be counted. It can impair health, raise levels of stress hormones and diminish the ability to do basic tasks such as preparing meals and taking baths. Although not everybody necessarily feels lonely in a state of social isolation, studies show (*like the Senni and Laures experiment*) that long term periods of isolation affect mental health and environment perception. In group isolation, cultural language, religious backgrounds, language, and emotional baggage every member has impacts the dynamic of the group and their ability to keep control of the situation when they encounter moments of tension. The effects of social isolation are lowered if the overall levels of stress are lowered and one of the most effective ways of doing so is meeting all the needs a human has equally.

Maslow's Pyramid separated the human needs into categories and presented a hierarchy based on the importance of each level. Following his theory, human needs are motivated by this hierarchy, but the order is not necessarily rigid. The first 4 levels are shown to decrease motivation as they are met whereas the last one increases motivation to keep accomplishing goals.

The first level of the Pyramid represents biological and psychological needs like water, air, shelter, warmth and human contact. With the information gathered in earlier Mars missions we can meet most of them. A day on Mars is 24 hours and 37 minutes, so the normal sleeping schedule will not change. In conditions where there is no light, it is believed that people turn to 48 hours sleeping sessions (Human Physiology edited by Robert F. Schmidt and Gerhard Thews). Even though this could be avoided

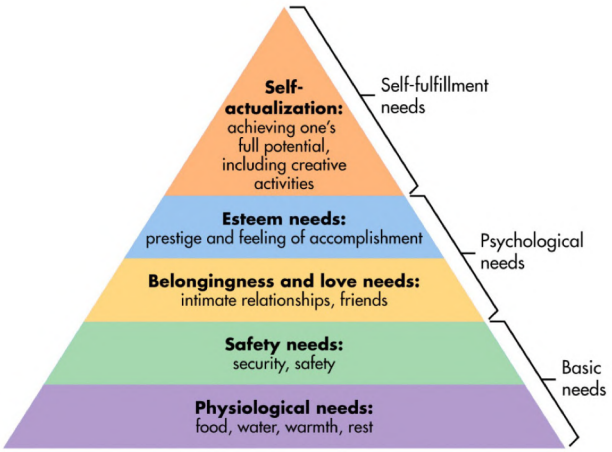


Figure 15: Maslow's hierarchy of needs. Image taken from Simply Psychology

by using artificial lighting programs, by relying on sunlight you save electricity. Earth's atmosphere has various components (Nitrogen, Oxygen, Argon, etc), but it was not like that since the beginning. Long ago, it was mostly made of Carbon Dioxide and water vapor, just as Mars' atmosphere is now. Scientists have already found ways of transforming Carbon Dioxide into Oxygen and Carbon in nonbiological ways. (For example, the instrument that the *2020 Mars* rover has, MOXIE, that breaks CO_2 in Carbon Monoxide and O_2.) Mars' atmosphere is thick enough to offer protection for potential agriculture results from solar flare. Therefore, inside a greenhouse with UV protective layers, it's possible to produce food on Mars. The abundant resources of ice in the permafrost make it possible to be exploited for water resources.

The second level of the Pyramid represents safety needs like protection from the elements, security, order, and stability. The tilt of Mars' axis is 25 degrees, close to Earth's 23.5 degrees tilt, meaning that the seasonal changes and the weather phenomenons are similar. As discovered by the *Vicking vehicles*, Martian summer is repetitive, but the rest of the seasons have various weather phenomenons, like cyclone variations, differences between temperature limits and the existence of ice in the winter. A controlled global warming operation can be attempted, in order to raise the temperatures to one more close to Earth's. In order to maintain order and stability, support and psychological training could be attempted from inside the mission as well as from the outside. In stressful isolated situations, the emerging stress and tension among the astronauts can be taken out on people on Earth. This situation can be controlled by strong leadership and psychological expertise, as well as good communication skills and effective problem solving.

The first two levels are basic human needs and need to be prioritized; the third level is love and belongingness needs such as friendship, intimacy, trust, acceptance, and receiving/giving love. As humans are social creatures, they usually create strong bonds between each other in nearly any conditions. Belonging needs do not emerge until basic human needs, as biological, psychological and security, are met. Humans need frequent, positive interactions with a few people, which form and maintain relationships. The need is not fully met if the interaction is not frequent nor if it's with different groups of people all the time.

The fourth level is esteem needs. This means self-esteem, born through achievement, independence, and respect from others. Crew members must be open to new experiences, conscious, extraverted, and agreeable.

Research shows (*Anderson, Marc H. "The Role of Group Personality Composition in the Emergence of Task and Relationship Conflict within Groups." ResearchGate, Mar. 2009*) that a team made of different types of personalities perform better than homogeneous groups of people. Group personality is important in conflict management and coordination of tasks, mostly because these aspects are a group level phenomenon and require efficient personality assemblage. This does not mean that there is a universal perfect team based on personality types, but choosing a member of a team based on personality traits is supported by science and can help in establishing a better task management and avoiding conflict [13]. If the team is composed of a heterogeneous group of personalities and skills, they will have different ways of feeling accomplished, they will have different needs and ways to feel independent and different anger outlets.

These are necessary human needs that need to be met for a group to function. Although there are other needs that appear once the first 4 levels are achieved, but those can't be influenced and created artificially.

B. Medical

Humans will be exposed to different conditions than those on Earth for long periods of time. There are a few measures that need to be taken into consideration when traveling to Mars. One of the main changes is a long period of lack of gravity, followed by a quick change of gravitational force. On Mars, the gravity is 0.334 percent of Earth's gravity. The approximate minimum period of time we need to get to Mars in 6 months.

1. Dust

Celestial dust is highly abrasive and can cause ocular and pulmonary injuries if it gets inside the spacesuit. Martian dust is highly oxidative, and when in contact with skin or eyes it can irritate and burn the area, which might result in dangerous conditions if not irrigated and treated properly. The pressure on Mars is also much lower, so dust is more likely to be deposited in the lung periphery than on Earth conditions, reducing the pulmonary volume and breath capacity.

2. *Lowered bone density*

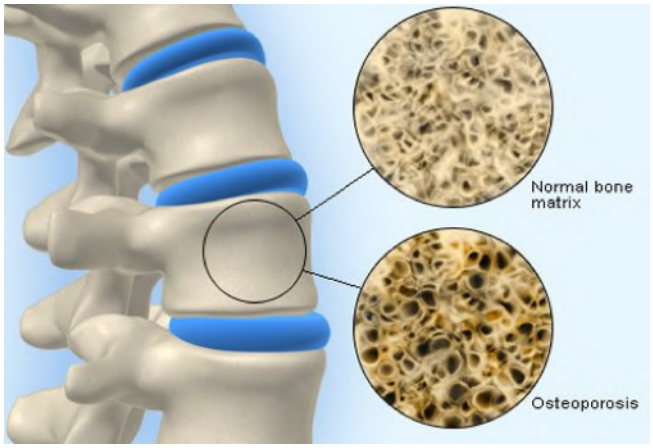


Figure 16: Difference in bone density. Image taken from MedicineNet

Bone density is the mineral mass of the bone, which is tissue and minerals, divided at the volume of the bone. It varies with race, gender, age and even the region of the body that is measured. You are more likely to break a bone with low bone density. From previous space explorations, we found out that humans lose 1 percent of their bone density each month. On Earth, elder humans lose about 1 percent per year.

After landing, the eye-hand coordination will be affected, blood will flow more easily, bones will need to regain density and humans will have to adapt to these new and untested conditions. Osteopenia, which decreases bone density, can be controlled by an adequate alimentary intake of vitamin D and Calcium, and Hormone-based medication, (Calcitonin, Progesterone, and Estrogen) or bisphosphonates. With the proper treatment, the bone density can stabilize and the risk for a bone fracture decreases.

3. *Muscle atrophy*

Another effect of the lack of gravity will be muscle atrophy. Since astronauts work in a weightless environment, their muscles need to do very little work to support different parts of the body. The only way to grow muscles is to break muscle cells (by exercising, for example) because when the body repairs them it also adds additional ones. If a muscle is not used for a long period of time, the cells in the muscles will die. Loss of muscle mass can be problematic if the astronauts need to perform emergency procedures that require strength. Astronauts can counter muscle atrophy through a combination of intensive exercise, particularly weight training, and a proper diet. We are also working on medication to inhibit muscle atrophy. There are different drugs still

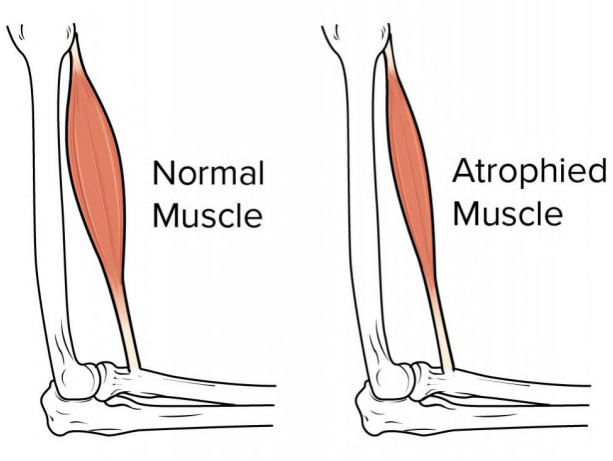


Figure 17: Muscle atrophy. Image taken from Jamie Eska

in trial that raise the level of Growth Hormone (responsible with muscle growth) or lower the levels of growth hormone inhibitors. These drugs will help regulate the human body while it adapts to new conditions, especially in cases in which exercise and a balanced food diet are not possible.

4. *Space Radiation Exposure*

Space radiation represents another dangerous risk during a mission to Mars, because of the extended period of time humans will be exposed to it. Inside the Earth's atmosphere and magnetic field, the radiation we are exposed to is less than 0,005 sieverts a year. From previous missions (like the Curiosity One) it was revealed that a trip from Earth to Mars would subject humans to approximately 0,66 sieverts. As 1 sievert increases the risk of cancer with about 5 percent, a trip to Mars and back might be too perilous.

The effects of radiation depend on the amount of radiation you are exposed to and the length of the period. In space, we are exposed to 3 types of ionizing radiation: particles trapped in the Earth's magnetic field, particles shot into space during solar flares, and galactic cosmic rays (high-energy protons and heavy ions from outside our solar system). Ionizing radiation has a substantial quantity of energy, removing all electrons from an atom they strike. Their energy is absorbed as they travel through materials. As they ionize atoms, they also produce changes in the structure of cells that affect their metabolism and functionality. Space radiation has different effects on DNA, tissues and cells. One of the consequences of radiation on DNA is that it breaks the DNA strings causing aggregated clusters of breaks. These DNA ruptures as well as the deterring of cell replication do not happen if exposed to Earth's radiation. Ionizing radiation also alters water structure, creating harmful free radicals, altering the structure of the cell.

Long term effects of repeated exposure to radiation reduce white cell percentage, weakening the immune system them, reduction of thrombocytes, cells responsible for stopping hemorrhages, fertility problems and changes in kidney function. Judging by the approximated quantity of radiation Mars explorers will be exposed to, besides the increase in cancer risk, degradation of the blood marrow(causing blood-forming deficiency) as well as degradation of the eyes' lenses(causing milky cataract).

Furthermore, a problem that might impale Mars colonization is the effect space radiation have on stem cells and fetuses since the rapidity of their division involves the rapidity of radiation spread. This means that the radiation not only makes it harder to procreate on Mars because of the effects it has on the reproductive system, but it also makes it impossible to bring kids and/or embryos safely on Mars.

5. Renal Stone Formation

Previous space missions unveiled that in space, humans are more exposed to renal stone formations. Different types of renal stones are caused by different alterations of the body. Renal stone formation is caused by lowered water intake causing lower urine output and increased Calcium excretion caused by how the lack of gravity affects bone metabolism.

Calcium Oxalate is the most common stone, caused by increased Calcium levels in urine (provoked by changes in bone metabolism, for example). Their formation can be suppressed by raising the urine pH (with oral intake of potassium citrate or other alkali). Uric acid stones are similar, but they occur more rarely and they can be prevented by reducing meat intake.

Struvite stones are generated by infections, caused by microorganisms that chemically change the urea in ammonia and Carbon Dioxide. They are formed when the pH raises over 7.2, and they usually need to be extracted by surgery. Most of them can be avoided by tracking the cause and check the alimentary intake of different elements.

6. Decompression Sickness

Decompression sickness is an effect of rapid changes in pressure, and in space, it occurs during extravehicular missions. The Nitrogen inside the tissues transforms into bubbles and rises after sudden pressure changes. These tiny bubbles cause different consequences depending on what part of the body they are in. Also known as aerobiosis, this affection is influenced by different factors, such as dehydration, age (as it's more likely to affect older people), previous affections, temperature, body types (persons with higher fat content are more likely to be affected). It can affect joints, mostly bigger ones (like elbows, shoulders, and knees) causing localized deep pain. It can cause chest pain and dry consistent cough, loss of balance and hearing loss, dizziness, headaches, paralysis, itching and a lot more. The effects of DCS can be lowered by lowering Nitrogen intake before a mission. Astronauts avoid the effects of DCS by breathing 100 percent Oxygen for hours before a mission.

7. Sensorimotor Alterations

Even though astronauts make efforts to maintain their bones and muscles in shape during missions, after long periods spent in microgravity the oculomoteur abilities, standing position, and lower limb ataxia are seriously affected. This is problematic because once they enter the Mars atmosphere, they might have to perform different tasks for the landing. Also, although Mars's gravity is much lower than Earth's, after at least 6 months spent in microgravity, they will still have to readjust. It was concluded after other space missions that the brain takes some time to adjust to space conditions, long-duration space flights affecting sensorimotor functions. To combat motion sickness, the drug promethazine can be used, as it is believed it does not affect sensorimotor adaptation. Because gravity has an important role in orientation, the lack of it forces astronauts to rely on visual perception for adaptation.

It was observed that the ability to adapt to microgravity conditions varies in different individuals and it can be predicted by exposing the individual to altered gravity conditions. It is also known that pre-adaptation training in said altered gravity conditions improves the individual's ability to adapt in another altered gravity environment.

EEG (electroencephalogram) is a test that measures brain wave patterns. Mostly, EEGs are used to diagnose and monitor seizures, but studies are made to understand how they can be used in motor learning and control. EEG was used in motor deficiency disjunctions, deepening our understanding of how the brain works. By monitoring the brain of the crew members constantly during the missions, we might be able to predict what motor skills are affected and we might be able to speed up the adaptation process and even correct the effects of space travel/ lower gravity in real-time by giving real-time feedback about brain and neurons activity.

8. Cardiac Rhythm Alteration

Cardiac rhythm refers to a sequential beating of the heart generated by electrical impulses. The heart has 3 nodes that produce the electrical impulse, but in normal conditions, the sinoatrial node is the one controlling the frequency of the beating. A normal sinus rhythm is between 60 and 100 bpm.

It was reported that after prolonged periods in space there were cases of cardiac rhythm alterations Since 1959, 11 cases of atrial fibrillation (AF), atrial flutter, or supraventricular tachycardia have been recorded among active corps crewmembers, and six additional cases have been identified among retired astronauts, but no episode was observed during space flight.

Atrial fibrillation happens when ventricles do not receive regular impulses and contract out of rhythm, and the heartbeat becomes uncontrolled and irregular. It is the most common arrhythmia, and 85 percent of people who experience it are older than 65 years. When the sinoatrial node fails to generate an impulse, atrial tissues or internodal pathways may initiate an impulse. Atrial flutter is a coordinated rapid beating of the atria.

All of these conditions are treatable by modern medication and it is not demonstrated that they are a direct result of microgravity, being associated with pre-existing, undiagnosed conditions.

Concepts

A. Manned VS Unmanned

There are both benefits and downsides of having a manned trip to Mars. Unmanned spacecrafts tend to be cheaper, safer, and able to provide new scientific data in ways that astronauts cannot. However, robotic equipment is unable to think creatively and look at new scientific discoveries from a unique perspective. Robots also cannot adapt to unforeseen problems the way that humans can.

The cost of a manned spacecraft, in both risk and price, is much higher than an unmanned spacecraft. Humans are incredibly expensive to launch past our atmosphere. People are heavy and require heavy life support equipment (such as water, air, and some means of recycling solid waste), meaning that it will take much more fuel to move a manned spacecraft to Mars. There is also the cost of taking care of a human living on Mars. NASA estimates that each astronaut on Mars will need to bring at least 1 ton of food/year. Each ton of food costs about 50 million, so each additional person on the spacecraft costs NASA an extra 100 million for food alone on a return trip. It is also becoming increasingly difficult to justify the cost of putting humans in space because of the rising quality of robots. The vast majority of discoveries made in the solar system have been carried out by robot probes, such as the discovery of the great hydroCarbon lakes on Saturn's moon, Titan. The risk is also much greater. If a robotic spacecraft explodes in space, the only loss is the cost of its materials. If a manned spacecraft explodes in space, it is disastrous not only for the world but for the future of the project and space program, not to mention the emotional damage to the astronaut's friends and family.

Robots are more advantageous at scientific exploration as they are less fragile and can exist in more extreme environments. They can be mounted with more advanced technology than humans can carry, which allows them to analyze a much wider variety of data. Unmanned missions can also stay out in space for much longer. The Hubble Telescope, one of the most celebrated observatories run by NASA, has been gathering information for 25 years. The longest time a human has ever stayed in space is 437 days and the physical and mental impacts that this long of an exposure to zero-G and small confined spaces, can be quite severe.

One benefit of a manned mission is that humans are much more capable of fixing unexpected problems. When the Hubble telescope required repairs, it was a human that fixed it, not a machine. Machines are incapable of fixing themselves or collecting information that they were not programmed for. When new information raises more questions, it is much easier for a human to conduct further studies to answer these questions than it is for a robot.

B. Types of Rockets

There are various types of space vehicles and methods that can be used to get a large range of people and cargo to the surface of Mars. Each has its advantages and disadvantages in fields such as comfort, safety, and price.

1. Light Landers

A Light lander is advantageous for short journeys, unlike that to Mars. It was used in the Apollo Missions, for instance, where there was a crew capsule, that was responsible for the reentry and safety of the crew. Also, there was a Lunar Excursion Module (LEM) docked to the top of the capsule. The journey to the Moon was just under 5 days and the crew had very limited space (6.2 m3)to be able to move around in. This was sufficient, as they had enough room to store the food and water needed for about 2 weeks for each astronaut, as well as space for all the life support systems.

If this method was used in a mission to Mars, a larger room would be needed, as spending 6 months in such a confined space would take a huge mental and physical toll on the crew. Furthermore, the storage of the capsule would need to be much larger. Approximately 26 times more rations would need to be brought on the trip than what was on the Apollo Missions. The Apollo command module also used a Carbon Dioxide absorbing system, which means the air was not recycled in any way. Instead, Oxygen was pumped into the capsule from storage, and Carbon Dioxide was removed. The amount of Oxygen, and Carbon Dioxide absorbers, that would be needed to support life for a total of 1 year traveling would be massive. Considering that 1 year of traveling in this module would be relatively impossible, it should be con-

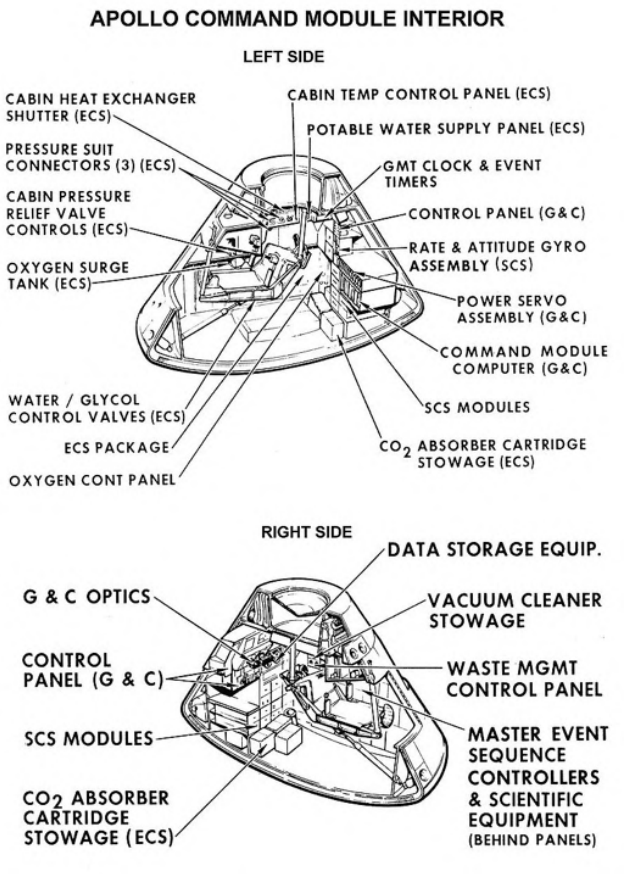


Figure 18: Interior of Apollo command module. Image taken from NASA

sidered as only possible for a one-way trip, arriving at a pre built habitat for the crew. However, it is not the most efficient way since only a small crew could be brought at first, meaning multiple rockets would be needed. This increases both cost and risk, and the capsule would not be able to control its landing, thus it would only be able to reenter and land in a rough area around the habitat with no vehicle to transport the crew. Furthermore, as previously discussed, to land any kind of vehicle on Mars, propulsion is usually needed as there is too strong a gravitational force for its weak atmosphere; ideally the capsule would travel alone, not attached to a fuel tank (in order to avoid any possibility of an issue such as that of the Apollo 13 mission).

2. Heavy Landers

Contrary to the light landers, a heavy lander is advantageous for long journeys; heavy landers have never properly been used before, other than the Space Shuttle which in this case would not work as the Martian atmosphere is too thin. The goal of a heavy lander is to transport large quantities of either crew, cargo, or both, which is ideal for 6 or 12 months of traveling. It would be used both in a one-way mission, just as much as it could be used in a return mission as a vehicle of that mass would only be able to land propulsively. With the right calculations, it would have sufficient fuel to set a return trajectory to Earth's atmosphere. A large vehicle like this would not even use parachutes to slow down as they would be almost insignificant, hence the entire re-entry process will be propulsive, making it much more accurate allowing for landing on a flat area near the pre-built habitat.

An example of such a vehicle is SpaceX's long-awaited "Starship" (a.k.a. BFR). The entire rocket is a 2 part combination that consists of the bottom section, "Super Heavy", which will be capable of lifting the entire top section, "Starship", into the upper atmosphere, before decoupling and returning to the surface landing propulsively, whilst Starship continues to form an orbit and eventually a transfer orbit to the desired planet or Moon. The advantage of using a large ship like Starship to do a 6-month journey is that its very large, spacious, and can carry enormous cargo. Hypothetically (as the designs change regularly) the vehicle should be able to hold a maximum of 100 passengers on a short journey, such as around the Moon and back, comfortably. On a journey to Mars, to increase comfort, no more than 50 passengers should be on the ship, as all the supplies for these passengers' survival will be on board. A vehicle of this size and mass however requires an extensive amount of fuel to move, which means that once it reaches Earth's orbit, another ship dedicated to this job, would need to rendezvous with the passenger ship and refuel it to ensure sufficient fuel is on board to complete the full journey.

Such a vehicle, of course, can be used for other purposes than transporting crew to the Martian surface. With the immense power and quantity of fuel in these vehicles, very large multi-tonne cargo missions can be done, both to support Humans on Mars and prepare the habitation that will support astronauts.

Another advantage of such large ships is they can be designed in such a way that allows for artificial gravity to be created. This could for instance be an internal section that spins, or the whole vehicle rotating at a steady speed. The consequences of the lack of gravity for extended periods of time are discussed in section XX B

3. Space Stations

The use of a Space station is the most expensive method of space travel, complicated and resource (time and money) consuming, however, it is the best solution for a two way mission. The method has never been tried before, but back before the Apollo Missions, Wernher von Braun conceptualized such a station that could be constructed in space and used to "go to the Moon and beyond". The station/large interplanetary ship would be constructed in Earth's orbit piece by piece like the

International Space Station (ISS). However, this version requires fuel storage and engines. Since escaping the atmosphere and forming an orbit is the most energy consuming part, the station could use slow but efficient engines to make a transfer orbit to Mars. Once the station reaches Mars, it would even enter the atmosphere. Hence, the station would enter a Martian orbit and a light lander such as a Martian adapted version of the LEM that will bring some crew down to the surface to take samples that can be examined in the orbital laboratory.

Although this won't allow for much surface exploration, it is the easiest way to keep the crew safe and comfortable as the station can be spacious and relatively light as there is no need for thick, high melting point metal, and for a large heavy heat shield and other required heavy landing equipment. One of the greatest physical issues on astronauts is the lack of gravity, which causes a wide range of medical conditions. The space station, as it's not entering an atmosphere, could theoretically have an artificial gravity ring or be built in such a way that allows for artificial gravity to be created.

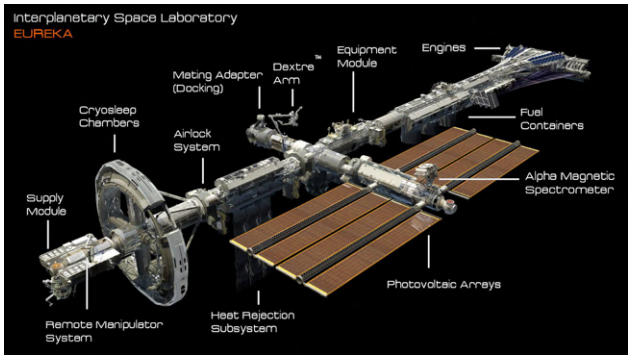


Figure 19: Concept of an interplanetary space laboratory. Image taken from Eureka

Another problem is that there would be no space for self sufficiency, so 2 years worth of food will need to be brought for each crew member. To put this into scale, the ISS receives new supplies around every 2 months.

C. Fuels and Engines

1. *Nuclear Thermal Engine*

One of the most effective ways to create enough energy to accelerate the spacecraft up to the desired velocity is through the use of nuclear thermal engines. The engines can use a light gas, such as Hydrogen, and heat it to at least 3,000 Kelvin in the chamber. During nuclear fusion, the incredibly dense, extremely heated Hydrogen particles clash together in the engine, releasing enormous amounts of energy to power the spacecraft. The gas is then shot out of the spacecraft in order to propel it forward. This type of engine is effective because not only is it fuel-efficient, but it has a relatively high thrust to weight ratio. To accelerate the craft sufficiently to reach Mars, the nuclear engine would need 230 grams (a half-pound) of uranium fuel. While uranium is radioactive, the actual fuel is not dangerous. However, once the engine starts to run, through nuclear fusion the uranium is split into other elements that are highly radioactive and can be deadly. The extreme heat produced by the engines can also be very dangerous to both the spacecraft and the crew on board, as not only is it in the direct solar radiation, but the engines produce easily double the thermal energy as a conventional chemical engine. Nuclear engines have also never been used as they are very controversial. If there was an explosion on the launch pad or low enough in the atmosphere, radioactivity could easily be released into the vicinity. Even though the crew could escape the explosion, the launchpad would need to be abandoned and in an area like Kennedy Space Centre, it could be dangerous for nearby large cities.

2. *Liquid Fuel Engines*

Liquid fuel engines, also known as chemical engines, are the most commonly used method of propelling a spacecraft. A liquid fuel engine combusts liquid propellants to create thrust so it has many benefits over all other types of fuels. Its biggest disadvantage is its need for liquid Oxygen onboard as it relies on the combustion of fuel to release extremely hot gasses to propel it. Since there is no Oxygen in space, a liquid fuel propelled spacecraft would need to bring its Oxygen to allow for complete combustion. These engines have the greatest thrust to weight ratio of any engine, which is why the first stage of a rocket almost always uses these engines, to be able to lift the mass of the whole rocket through the thick atmosphere. Nonetheless, these engines consume an extensive amount of fuel, which means for it to be useful, fuel needs to be hauled in the rocket. One more benefit of using liquid fuel engines, both in space and in the atmosphere, is that the thrust can be controlled. This is especially useful for very small precise movements in space and for controlling the acceleration during launch to avoid putting too much stress on the rocket. It also means that they can be switched on and off a lot. However, it is one of the hardest types of engines to reignite in space as a separate fuel tank of hypergolic fuels needs to be hauled up. The most common types of fuels used are Hydrogen and RP-1, as they are light and relatively easy to be kept cold, which is ideal as most rockets use liquid versions of the fuel and oxidizer. For example, the Falcon 9 rocket uses liquid RP-1 fuel and liquid Oxygen as its propellant in both the first and second stages.

3. *Solid Fuel Rockets*

These rockets work by burning a solid substance that burns rapidly but does not explode. There are some ad-

vantages to these engines which include simplicity and price of manufacture. Granted, they are a lot simpler to build as they work based on setting fire to large quantities of highly flammable material and don't need to mix the exact amounts of Oxygen and liquid fuel heated to the correct temperature as in chemical engines. Nevertheless, they are a lot more dangerous, since the thrust in a solid-fuel rocket cannot be controlled, and once ignited, the engine cannot stop, decrease in thrust or even detach from the main rocket. In the case of the Space Shuttle Columbia (Columbia Disaster), there was a fault in one of the solid fuel boosters strapped to the side, and it began leaking extremely hot gas which eventually led to an explosion killing all on board.

4. Ion Engines

More efficient than a conventional rocket, this type of engine uses electrical energy from solar cells and Xenon gas to propel itself forward. The ratio of thrust to the rate of propellant consumption is high, meaning that an ion rocket requires significantly less fuel than a rocket-fueled by chemical propulsion, making it the most efficient engine ever made. Unfortunately, the engine is only effective on very small spacecraft such as communication satellites as the thrust to weight ratio is smaller than any engine. In this engine, atoms are pushed from their storage container into a chamber where a high potential difference of electricity (provided by large solar panels or small reactors) causes the ionization of the atoms. This causes them to be shot out the back of the rocket. According to the third law of Newtonian mechanics, the almost insignificant force of the ion being pushed out is applied to the spacecraft, pushing it forward. Therefore this method of propulsion only works on very light vehicles with no resistance of any kind.

Conclusion and Opinion

Mars is the fourth and smallest planet in our Solar System. It is necessary to explore Mars because it can ensure the survival of our species even through an extinction level event on Earth, provide information on past or present life forms, broaden our understanding of how life started, and improve the quality of our own life (see section XV and XV A).

The ideal mission would travel to Mars with no planned return because it would provide more time to study Mars and would have the potential to save the human race from extinction. It would also be much simpler than a mission with a planned return trip. A one-way mission also requires less fuel, thus offering more room for supplies and colonists per trip (see sections XXI A and XVIII B).

Regarding vehicles, this mission relies on a Heavy Lander over the light lander and space stations. The light lander is significantly smaller, meaning that there is limited space for supplies and colonists (see section XXI B 1). The Space Station, while viable for two-way missions, is quite expensive and as it can not land on the surface of Mars, it would not be used as a Martian habitat. Furthermore, the only way to get to and from the station would be with the assistance of a light lander, which from Martian orbit, has a lower potential for an accurate landing (see section XXI B 3). However, Heavy landers are much larger and could bring up to 50 passengers, significantly reducing the effects of social isolation. The Heavy Lander, due to its large size, provides sufficient room for all life support systems and rations for the journey whilst still fitting the crew comfortably, which would decrease stress among the people on the spaceship. This spacecraft, because of its size, can be designed in such a way to be able to rotate in space during the 6-month journey and create an artificial gravity effect (see section XXI B 2). The artificial gravity could prevent medical problems like muscle atrophy, lowered bone density, renal stone formation, decompression sickness, sensorimotor alterations, and cardiac rhythm alterations (see section XX B). The Heavy Lander could also hypothetically allow an emergency return from Mars back to Earth, in cases such as the destruction of the habitat.

However, Heavy landers are much larger and could bring up to 50 passengers, significantly reducing the effects of social isolation. The Heavy Lander, due to its large size, provides sufficient room for all life support systems and rations for the journey whilst still fitting the crew comfortably, which would decrease stress among the people on the spaceship. This spacecraft, because of its size, can be designed in such a way to be able to rotate in space during the 6 month journey and create an artificial gravity effect (see section XXI B 2). The artificial gravity could prevent medical problems like muscle atrophy, lowered bone density, renal stone formation, decompression sickness, sensorimotor alterations, and cardiac rhythm alterations (see section XX B). The Heavy Lander could also hypothetically allow an emergency return from Mars back to Earth, in cases such as the destruction of the habitat.

The mission would use propulsive landing, as it is much more accurate, safe and feasible than parachutes and bouncing airbags. The vehicle is too heavy for parachutes to effectively slow down in the martian atmosphere and certainly not for a safe landing, and bouncing airbags would pop under the heavy weight of the ship as well as damage the ship and the crew. Therefore, the most practical option is propulsive landing, allowing for a soft, accurate and safe landing (see section XVIII C).

Because no such Heavy Lander has ever been tested, a long and rigorous testing process would be required, especially because it can carry many more lives than any previously flown rocket. This means that a failure in space would put more passengers at risk. Due to the communication delay with Earth and the sheer complexity of a manual landing, the Heavy Lander must prove that it is

capable of landing autonomously in various conditions, meaning it could take years of practicing on Earth and unmanned missions to Mars before it's ready (see section XIX C).

The Heavy lander would work best with chemical engines, rather than an ion or a nuclear engine. A nuclear engine, along with being expensive and unsafe for the crew and passengers, is only efficient in a vacuum and is therefore not suited for a landing on Mars, which has an atmosphere and relatively strong gravity compared to the thrust produced. Ion engines are too feeble; they would hardly move a spacecraft the size of a heavy lander, even in a vacuum. However, chemical engines are extremely high powered, which is favorable during a propulsive landing in the Martian atmosphere. They are well tested and less likely to fail during the mission. Furthermore, they require Oxygen to work, which could also be used for life support during the journey (see section XXI C).

The location of the habitat is also crucial to the success of the mission, and after comparing several sites, the Jezero Crater has been determined to be the ideal place to construct the habitat and send the crew (as long as the Mars2020 rover does not land there). Other options include Oxia Planum, Columbia Hills, and Northeast Syrtis. Oxia Planum is not suitable because there is not enough silicon, a crucial recourse to self-sufficiency. Both Columbia Hills and Northeast Syrtis provide a much lower chance of encountering biosignature presence than the other options. However, the Jezero Crater not only has a high chance of biosignature presence, but it is also high in silicon and is easy to land on due to its flat terrain. The Jezero Crater can also better protect human colonists from dangerous dust storms as it has steeps sides. Lastly, the Jezero crater, although not as much as the Northeast Syrtis, is rich in minerals and old rocks from deep in the Martian ground, which were brought out by the impact that created the crater (see section XIX B).

The process of colonizing Mars requires three unmanned missions and one final manned mission. The first unmanned mission brings construction supplies necessary for the development of the habitat and landing pad. Depending on how many supplies each rocket can hold, several rockets may be necessary for this mission. These supplies would include materials for the construction of: landing pads, the habitats, power production systems and more. The second unmanned mission delivers robots whose purpose is to use the resources brought in the first mission, to prepare everything for habitability. The third unmanned mission carries supplies and equipment which can be transported into the habitat by the robots from the second mission. In the final manned mission, the first group of crew and colonists would land on the landing pad and begin to organize the first Martian colony.

The crew on the manned mission would consist of a commanding officer, engineers who specialize in different fields, as well as scientists and researchers. All the colonists will also be trained in various nonscientific fields such as medicine or psychology. The commanding officer is in charge of the crew and the payload of the ship. He is responsible for the safety of the crew and takes final decisions in times of time pressure or lack of communication with Earth to avoid a possible conflict. Types of engineers can include flight, agricultural, electrical, civil, communications engineers, all specialized in different aspects of engineering that can solve unexpected issues. The researchers would help gather and analyze data to study the planet and possibility of life. (see section XIX D)

Theoretically, the rocket would be capable of escaping Mars and returning to the surface to bring back the crew in the case of an emergency, or to bring back the rockets (with Waste materials) as a means of reusing them for another cargo shipment and to avoid polluting the Martian surface. In order for this to be possible, the vehicle must have sufficient fuel to complete all the burns (see section XVIII A).

The humans will never cease to learn, and space exploration is one of the many fields, that we know so little about. The colonization of Mars would not only teach us so much, but it will allow the human race to become a space bearing civilization.

Appendix

1. Celestial: Pertaining to the universe beyond the Earth's atmosphere.
2. Cube sat: A small satellite usually no bigger than 10cm x 10cm x 11.5cm.
3. Extravehicular: Relating to work done in space, outside a spacecraft.
4. Oxidise: To combine chemically with Oxygen.
5. Bio signature: A chemical or physical marker indicating the presence of life.
6. Terrestrial: On or relating to the Earth.
7. Albedo: The proportion of the incident light or radiation that is reflected by a surface, typically that of a planet or moon.
8. Dry ice: Solid Carbon Dioxide.
9. Magnetic field: A region around a magnetic material or moving charge within which the force of magnetism acts.
10. Propellant: A substance used to create thrust, typically in rockets.
11. Elliptical orbit: An orbit that is not a perfect circle around a body, it has a highest point and a lowest point.

12. Apoapsis: The highest point in an elliptical orbit around a body. Around the Earth: Apogee, around the sun: aphelion.

13. Periapsis: The lowest point in an elliptical orbit around a body. Around the Earth: perigee, around the sun: perihelion.

14. Aeroshell: A casing which protects a spacecraft during re-entry

15. Geostationary: Having a circular equatorial orbit at an altitude 35,900km, so that the satellite always travels directly above the same position on the surface of the body it orbits.

16. Olivine: An olive green mineral that occurs naturally in Basalt, peridotite and other basic igneous rocks

17. Sol: A Martian day, approximately 24 hours and 37 minutes.

18. Permafrost: A thick subsurface of soil that remains at below freezing point throughout the year, occurring chiefly in polar regions.

19. Ataxia: The loss of full control of bodily movements.

20. Ionize: Convert an atom, molecule or substance into an ion or ions typically by removing one or more electrons.

---

[1] Adam Mann, *"The Problem With Dust On The Moon And Mars"*.

[2] Klein,Harold P.t *"Mariner 8 and Mariner 9"* NASA.

[3] *"Viking Mission to Mars"*.

[4] NASA Facts, NASA 20 July 2016

[5] NASA, NASA 26 June 2019 *"NASA's Insight Mars Lander"*.

[6] NASA, NASA,30 Nov 2018, *"InSight Mission Overview"*.

[7] Mcleod, Saul. Simply Psychology, 21 May 2018, *"Maslow's Hierarchy of Needs"*.

[8] Wall, Mike. Space.com, Space,1 August 2014, *"Oxygen-Generating Mars Rover to Bring Colonization Closer"*

[9] NASA, NASA, *"Overview"*, Mars.nasa.gov/mer/mission/overview/

[10] Zurbin, Robert. National Space Society, *"The Case for Colonizing Mars, by Robert Zubrin"*

[11] Google, Google *"Missions To Mars Have Had A High Failure Rate Historically [Infographic]"*

[12] American Psychological Institution 1995 *Baumeister, Roy F. "The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Desire"*

[13] Anderson, Marc H. ResearchGate, March 2009, *"The Role of Group Personality Composition in the Emergence of Task and Relationship Conflict within Groups"*

[14] Bone Densitometry *"Bone Density"*

[15] Driver, Catherine Burt. "Osteopenia Diet, Causes, Treatment & Medications." MedicineNet

[16] Guest. "Muscle Atrophy- NASA." Mafiadoc.com, MAFIADOC.COM

[17] "Mars Exploration Program Archives." Universe Today, 30 March 2019,

[18] Eske, Jamie. "Muscle Atrophy: Causes, Symptoms, and Treatments." Medical News Today, MediLexicon International

[19] "ExoMars Mission (2020)." ESA

[20] Tate, Karl. "Space Radiation Threat to Astronauts Explained (Infographic)." Space.com, Space, 30 May 2013

[21] "How Long Does It Take to Travel to Mars? - A Mission to Mars." Mars One

[22] "1990 Recommendations of the International Commission on Radiological Protection." ICRP, www.icrp.org/publication.asp?id=ICRP Publication 60

[23] "How Does Ionizing Radiation Affect Our Body." How Does Ionizing Radiation Affect Our Body?, Hong Kong Observatory

[24] "Understanding Space Radiation." NASA Facts, National Aeronautics and Space Administration, Oct. 2002

[25] McKie, Robin. "Astronauts Lift Our Spirits. But Can We Afford to Send Humans into Space?" The Guardian, Guardian News and Media, 7 Dec. 2014,

[26] Filmer, Joshua. "What Is the Value of Manned Space Exploration?" Futurism, Futurism, 18 July 2014

[27] "Renal Stone Formation in Space." Wikipedia, Wikimedia Foundation, 17 Apr. 2019 en.wikipedia.org/wiki/Renal_stone_formation_in_space

[28] Sibonga, Jean D. "Risk of Renal Stone Formation." Risk of Renal Stone Formation, Human Research Program Human Health Countermeasures Element , Mar. 2008, humanresearchroadmap.nasa.gov/evidence/reports/RenalStone.pdf

[29] *"Rocket Propellants."* Basics of Space Flight: Rocket Propellants

[30] Dunbar, Brian. *"Ion Propulsion."* NASA, NASA, 18 Aug. 2015

[31] Verhovek, Sam Howe. "The 123,000 MPH Plasma Engine That Could Finally Take Astronauts To Mars." Popular Science, Popular Science, 18 Mar. 2019

[32] Government of Canada, Canadian Space Agency. "What Is Decompression Sickness?" Canadian Space Agency Website, 18 Aug. 2006

[33] Stratford Monday, Frank, and Frank Stratford. "Why Should Humans Go to Mars?" The Space Review: Why Should Humans Go to Mars?

[34] "Dust in the Atmosphere of Mars and its Impact on Human Exploration" edited by Joel S. Levine, Daniel Winterhalter, Russell L. Kerschmann - chapter ten

[35] "Solution To Clean Space Dust From Mars Exploration Vehicles." Mars Exploration News

[36] Hille, Karl. "The Fact and Fiction of Martian Dust Storms." NASA, NASA, 18 Sept. 2015

[37] Williams, Matt. "How Do We Colonize Mars?" Universe Today, 21 Nov. 2016

[38] Clark, Torin K. "Development of a Countermeasure to

Enhance Sensorimotor Adaptation to Altered Gravity Levels." NASA
[39] Whittier, Tyler T. "ELECTROENCEPHALOGRAPHY (EEG) AND ITS USE IN MOTOR LEARNING AND CONTROL." Master's Thesis Tyler Whittier, July 2017
[40] Caiani, et al. "Weightlessness and Cardiac Rhythm Disorders: Current Knowledge from Space Flight and Bed-Rest Studies." Frontiers, Frontiers, 9 Aug. 2016
[41] "Geology." NASA, NASA, Mars.nasa.gov/programmissions/science/goal3/
[42] "Surface Geology of Mars." The Center for Planetary Science
[43] OpenStax. "Astronomy." Lumen
[44] "What Are the Duties of an Astronaut?" Career Trend
[45] Gonzalez, Robbie. "The Best Map Yet of What Could Be NASA's Next Mars Landing Site." Wired, Conde Nast, 3 June 2017.
[46] Bramble, M S. "Stratigraphy of the Northeast Syrtis Major Mars 2020 Landing Site and the Ejecta of Jezero Crater, Mars." Lunar and Planetary Science Conference 2018
[47] "Columbia Hills (Mars)." National Aeronautics and Space Administration Wiki
[48] "Shuttle Mission: STS-1." History of the US Shuttle Program, US Space Shuttle.com
[49] https://en.wikipedia.org/wiki/OxiaPlanum

Searching for exoplanets around EPIC 206103150 with the transit method

Lavinia Finalde Delfini
Project Mentor: Ozgur Can Ozudogru

This paper presents the analysis of the lightcurve of the star EPIC 206103150 (also known as WASP-47). The aim of the research project is to find the exoplanets orbiting the star and to determine their properties through the transit method. The observation was acquired by Kepler's Space Observatory during its third K2 campaign. This paper also contains background research about exoplanets, such as their classification, the methods of detection and Kepler's mission.

Introduction

A. Exoplanets

Exoplanets are planets orbiting stars that are not the Sun. The first exoplanet was discovered in 1992. Since then, the number of known exoplanets has greatly increased: as of today (29 August 2019), 4043 exoplanets have been detected and there are thousands of possible exoplanets being analysed.

Discovering and studying exoplanets is particularly important for two reasons:

1. To understand the formation of our solar system by developing models of star and planet formation based on the observation of other planetary systems.

2. To find Earth-like planets that could possibly host life.

B. Classification

Exoplanets can be classified on the basis of their mass, orbit or composition. Using mass as the basis for the classification, they can be divided in Terrestrial Planets and Gas Giant Planets.

Terrestrial Planets

Terrestrial planets have a rocky composition, and they can have deserts, oceans, ice and atmospheres. There are 3 types of terrestrial planets:

- Subterran (0.1-0.5 Earth masses)
- Terran (0.5-2 Earth masses)
- Superterran (2-10 Earth masses)

Gas Giant Planets

Gas giants are large planets composed mostly of hydrogen and helium. There are 2 types of gas giants:

- Neptunian (10-50 Earth masses)
- Jovian (50-5000 Earth masses)

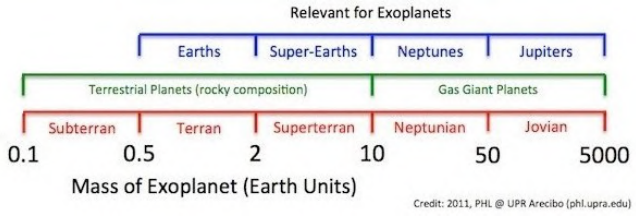


Figure 1: Exoplanets Mass Classification (EMC)
Image taken from UPR Arecibo

C. Methods of detection

1. *Direct detection*

Imaging: Exoplanets are hard to detected directly: they are small, distant, and much dimmer than the star they orbit. To be able to see them, the light of the star has to be dimmed or blocked out. This can be done by using infrared radiation or using a coronagraph.

2. *Indirect detection*

Finding planets through the effect they have on the star they orbit.

Radial velocity tracking: As a planet orbits a star, it pulls on it and causes it to move in a small orbit, getting closer and then farther from Earth. The changes in the radial velocity of the star (the velocity with respect to the observer on Earth), make the star's spectrum shift because of the Doppler effect (as can be observed in FIG. 2). The spectrum is detected using high precision spectrographs and the changes in radial velocity can be used to calculate the minimum mass of the exoplanet, which depends on the inclination of its orbit.

Astrometry: Astrometry is the most sensitive method of detection. It consists in analysing the changes in position of the host star. Unlike radial velocity tracking, astrometry gives us an accurate estimate of the mass of the exoplanet.

Gravitational microlensing: According to Einstein's General Theory of Relativity, when light em-

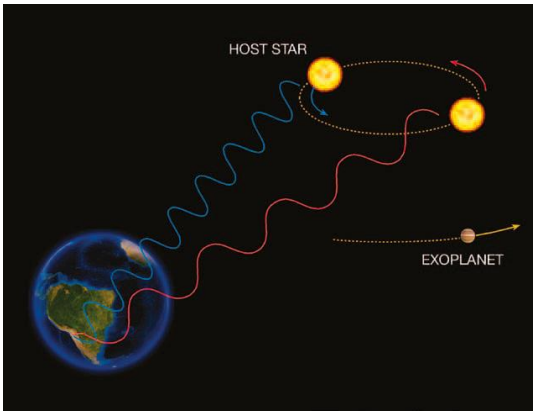


Figure 2: Radial Velocity method
Image taken from ESO

anated from a star passes next to another star the gravitational pull of the second star will make the light bend. If the two stars are aligned when seen from Earth, the light will pass on both sides of the intermediary star, "lensing" it. This causes the star to appear much brighter. If the star has a planet orbiting around it, its gravitational pull will make the light ray on one side of the star bend and it will cause the brightness to increase further, depending on its mass and distance from the star.

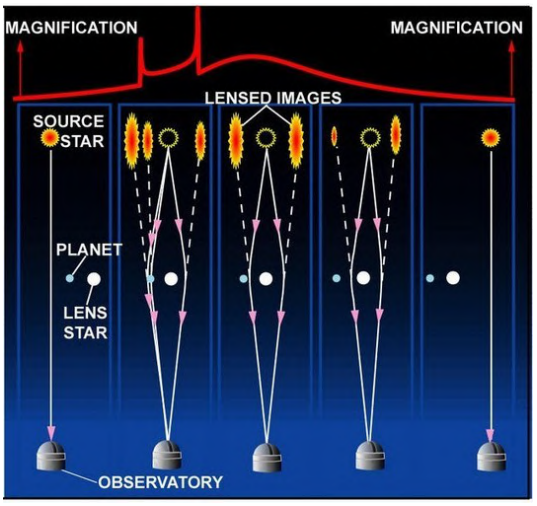


Figure 3: Gravitational Microlensing
Image taken from The Planetary Society

Transits: During a transit (when a planet passes between Earth and its host star) the planet blocks some of the light going from the star to Earth, causing a decrease in the observed brightness.

In detail, the light curve during a transit will behave in the following way: during ingress, the area of the star that is covered by the planet increases more and more, so the brightness decreases gradually. During the transit, the planet covers parts of the star with different brightnesses. Since the star is brightest at its centre, the lowest brightness is observed when the planet is aligned with the centre of the star.

The transit method can be used to identify exoplanets and to calculate their period, radius and (if combined with the radial velocity method) their mass.

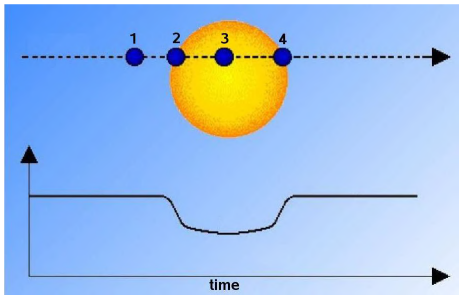


Figure 4: Light curve of a star during a planet transit
Image taken from Optical, Infrared, and Millimeter Space Telescopes, SPIE Conference 5487

D. Kepler's mission

NASA's Kepler Space Observatory operated from March 2009 to October 2018. The aim of the mission was to look for exoplanets, and in particular for Earth-like ones, using the transit method of detection.

In order to increase the probability of detecting transits, Kepler's photometer was designed to scan more than 100,000 stars at the same time.

During the first part of the mission, Kepler always scanned the same patch of the sky (the Cygnus-Lyra region). In 2014, the mission had to be changed after two of the four reaction wheels, that kept the telescope stable, broke. In June, the K2 mission started.

1. K2 mission

The K2 missions consisted of a series of campaigns of the duration of approximately 80 days. After that, the spacecraft had to turn to avoid the light of the sun.

The K2 mission was entirely community-driven, since the targets were proposed by the community.

2. Discoveries

In more than 9 years in space, Kepler observed over 500.000 stars. Its findings changed our understanding of the universe: we learned that there are more planets than stars, and that 20-50% of solar systems are likely to have planets which may be habitable. The mission also showed us the great diversity of planets and solar systems.

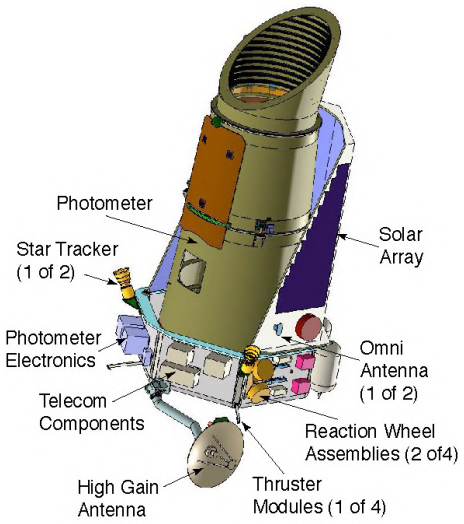


Figure 5: Kepler's spacecraft
Image taken from Johannes Koppenhofer

3. EPIC 206103150

EPIC 206103150, also known as WASP-47, is a star studied by Kepler's telescope during its third K2 campaign. The telescope was positioned at the following celestial coordinates:

right ascension 331.203045; declination -12.018893.

The telescope registered the change in the flux of the star from 17 November 2014 at 22:22:05.054 to 23 January 2015 at 18:20:15.127. The data acquired by the telescope can be found on the website of Harvard University's CfA. The telescope also provided other information on the star, which can be found on "K2 Ecliptic Plane Input Catalogue" of the Mikulski Archive for Space Telescopes:

mass: $m_{star} = 1.003$ solar masses

radius: $R_{star} = 1.104$ solar radius

Methods

A. Graph of the transit

A *corrected flux* vs. *time* graph was plotted using the information provided by the observation.

In the graph, time is expressed in Barycentric Julian Date (BJD). This is an astronomical unit of time that takes into consideration the time difference between the emission of light and when the observer sees it, based on its position with respect to the barycentre of the Solar System.

The flux is the number of photons that go through a cm^2 in a second. The corrected flux is different from the flux in that it doesn't show the background noise from the sky. However, noise can still be observed on the graph, represented by rising peaks. The noise is caused by solar flares from the star. These explosive events emit lots of photons, causing a sudden increase in the flux that can last from minutes to hours.

After removing the rising peaks, from the clean graph it was possible to observe that:

- The brightness od the star changes over time, increasing gradually at first and then becoming more stable.
- There are two planets orbiting the star, causing the brightness to decrease periodically as they transit it. Planet A is blocking more light than the other. Planet B has a larger orbital period.

In the clean graph, some irregularities can still be observed.

The first irregularity happens around 2179.0 BJD (12:00 of December 20, 2014). There can be several possible causesm, such as an error of the telescope, a non-bright object positioned between the telescope and the star or a big solar flare that happened while planet A was transiting the star.

The second irregularity happens at 2210.649 BJD (03:34 of January 21, 2014). It might have been caused by a third planet orbiting the star, or any other non-bright object positioned between the telescope and the star.

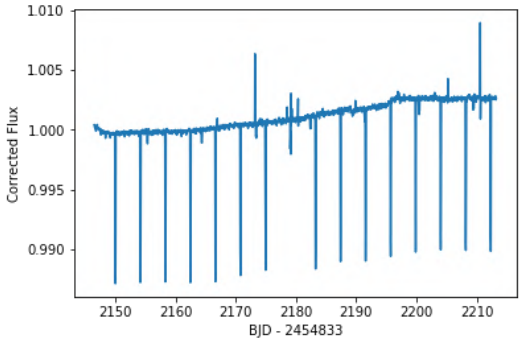


Figure 6: Original graph

B. Calculating the orbital period

On the graph, the orbital period P of a planet is the time between one minimum point and the next one. It was possible to make an accurate estimate of the orbital period of the two planets by calculating the average of 5 of these.

The following table shows the time (in days) between the planets' transits, and their average.

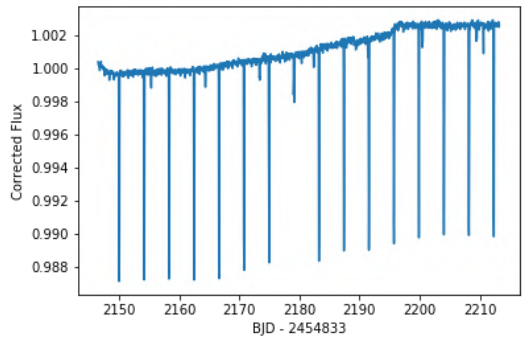


Figure 7: Clean graph

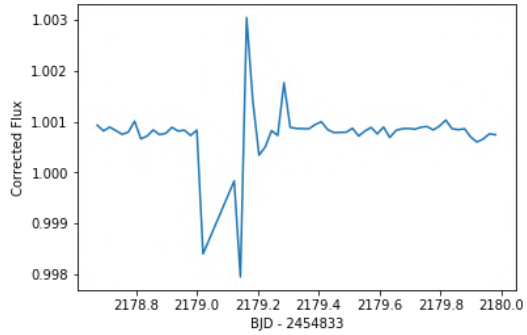


Figure 8: Graph from 2178.673426 to 2179.960614 BJD: second irregularity

Planet A:	Planet B
$P_{transit_{1,2}} = 4.15$	$P_{transit_{1,2}} = 9.05$
$P_{transit_{5,6}} = 4.19$	$P_{transit_{2,3}} = 9.01$
$P_{transit_{9,10}} = 4.15$	$P_{transit_{3,4}} = 8.99$
$P_{transit_{11,12}} = 4.17$	$P_{transit_{4,5}} = 9.07$
$P_{transit_{15,16}} = 4.17$	$P_{transit_{6,7}} = 8.99$
$P_{average} = 4.16$	$P_{average} = 9.02$

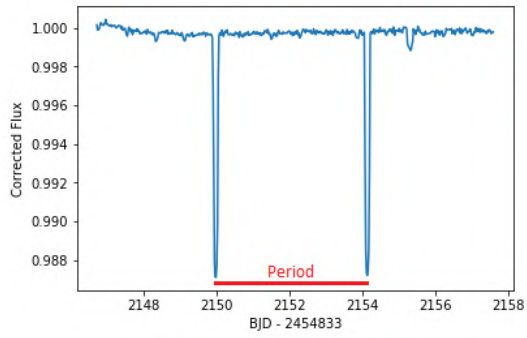


Figure 9: Orbital period on the graph

C. Calculating the radius

The drop in brightness of a star during the transit of a planet depends on the radius of the star and that of the planet:

$$depth = (\frac{R_{planet}}{R_{star}})^2 \tag{1}$$

Therefore, if the radius of the star is known, the depth can be used to calculate the radius of the planet:

$$R_{planet} = R_{star}\sqrt{depth} \tag{2}$$

The depth was calculated by making the average of the flux at the point where the drop starts and the flux at the point where it ends, and then subtracting the flux at the minimum point. The average depth of the planet is the average of the depth of different drops.

From the depth and the radius of the star, it was possible to calculate the radius of the two planets using the formula (2).

D. Calculating the semi-major axis

The semi major axis a is the distance between the centre of the orbit of a planet and the perimeter. It can be calculated using Kepler's third law:

$$P^2 = \frac{a^3}{M_{star}} \tag{3}$$

Therefore:

$$a = \sqrt[3]{M_{star}P^2} \tag{4}$$

where P is in years, M is in solar masses and a is in AU.

E. Results

The following table summarizes the properties of the exoplanets, calculated as explained at point B, C and D.

	Planet A	Planet B
Orbital period (days)	4.16	9.02
Radius (solar radius)	0.13	0.04
Radius (Earth radius)	13.54	4.01
Radius (Jupyter radius)	1.23	0.37
Semi-major axis (AU)	0.05	0.09

Discussion and Conclusion

The first part of the project consisted mostly of researching information about the topic. This allowed me to learn more about the current situation of the search for exoplanets, before moving on to my own project.

The project I worked on required to write many codes on Python, to first plot and then study some graphs. This was complicated but at the same time it forced me to learn more about a field that I didn't know much about.

Moreover, working on the project I learned how scientific research is conducted.

The final goal of the project was to find the exoplanets orbiting around the star and to calculate their properties. I found and analysed 2 of the 4 planets that orbit the star. Their official names are WASP-47 b for what I called "planet A" and WASP-47 d for what I called "planet B". Comparing my results to the official data published on the Open Exoplanet Catalogue, I found them to be extremely similar and therefore, I think we can say that the research project was successful.

---

[1] NASA Exoplanet Archive, ipac, Caltech, accesed in: 2019
[2] Searching for Extra-solar Planets with the Transit Method, Johannes Koppenhofer, Ludwig–Maximilians–University (LMU), Munich, 2009
[3] Exoplanets presskit, ESO, accessed in: 2019
[4] A Mass Classification for both Solar and Extrasolar Planets, Planetary Habitability Laboratory, UPR Arecibo, accessed in: 2019
[5] Astrometry: The Past and Future of Planet Hunting, The Planetary Society, accessed in: 2019
[6] Microlensing: Beyond our Cosmic Neighborhood, The Planetary Society, accessed in: 2019
[7] Transit Method, Las Cumbres Observatory, accessed in: 2019
[8] Calculating Exoplanet Properties, Simon Fraser University, accessed in: 2019
[9] Kepler and K2: Mission overview, NASA, accessed in: 2019
[10] Mission objectives, Kepler & K2 Science Center, NASA, 2019
[11] Overview and status of the Kepler Mission, Optical, Infrared, and Millimeter Space Telescopes, SPIE Conference 5487, Glasgow, 2004
[12] NASA's First Planet Hunter, the Kepler Space Telescope: 2009-2018, NASA, accessed in: 2019
[13] Famed planet-hunting spacecraft is dead. Now what?, Nadia Drake, National Geographic, accessed in: 2019
[14] Using Commercial Amateur Astronomical Spectographers, Appendix A: Astronomical Time, Jeffrey L. Hopkins, 2013
[15] Barycentric Julian Date, Jason Eastman, Ohio State University, accessed in: 2019
[16] WASP-47, Open Exoplanet Catalogue, accessed in: 2019
[17] K2 Ecliptic Plane Input Catalogue, Mikulski Archive for Space Telescopes, accessed in: 2019

Galaxy Image Partial Reconstruction With Machine Learning

Kunwanhui Niu
International Science Engagement Challenge 2019

The aim of this report is to attempt to reconstruct half of the image of a galaxy from the other half using machine learning. Pix2Pix, a powerful image to image transformation algorithm, was modified and two approaches were used, one rotating the input and the other one not rotating them. The code was depicted, and the results were analyzed, which showed the power of Pix2Pix and lead to further studies.

Introduction

A. Theory

1. Galaxy

A galaxy is a cluster of stellar objects, planetary objects, gas, dust, and dark matter. Within a galaxy, matter is bound by gravity, while spaces outside galaxies are usually voids. A galaxy can be classified into three categories: spiral, elliptical, and irregular. Most galaxies have a super-massive black hole (SMBH) at the core and any other objects orbit the black hole. The galaxy we live in, the Milky Way, is a spiral galaxy. See FIG. 1 for a diagram of Hubble sequence, one of the most accepted classification of galaxies.

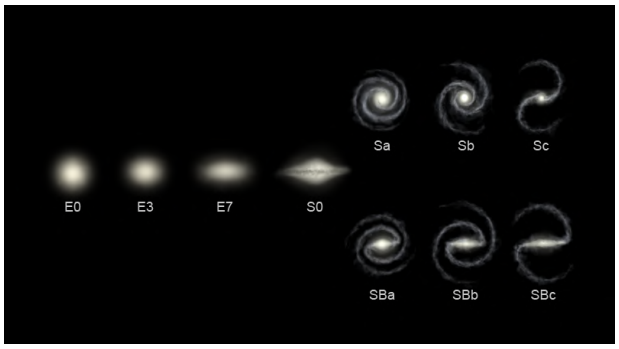


Figure 1: Hubble Sequence Diagram from Wikimedia commons

In spiral galaxies, there are often a dense central bulge consisting of mainly older stars and a supermassive black hole at the center. Around the bulge, most matter resides in a rotating flat disk and form spirals, in which many young bright stars are present. Other matter forms a spherical halo surrounding the galaxy, in which stars often congregate in globular clusters. Sometimes a bar structure may be present extending into two opposite directions from the central bulge, and spiral arms emerge from the end of the bar.

In elliptical galaxies, there is also a denser central region consisting of a supermassive black hole. However, the stellar density is lower compared to that of spiral galaxies, and most stars are old orange or red stars. As a result, elliptical galaxies usually appear yellow and featureless, with little dust or gas visible. Their shape varies between spherical to highly elongated, and they often contain many globular star clusters.

Other galaxies that could fit neither classifications are considered irregular galaxies. Most of them have been distorted by gravitational influence of other galaxies or formed through galactic collision, so they have various irregular shapes and features.

2. Artificial Neural Network

An artificial neural network is a computing system that simulates biological neural networks. The network consists of many nodes, which could take in input from connecting nodes, apply weight to each input, calculate an output via a function, and deliver the output to the next layer (or if the node is at the last layer, generate an output). It is frequently used in various data analysis tasks, such as image recognition, speech recognition, machine translation, content filtering, and medical analysis. See FIG. 2 for diagram of a simple neural network.

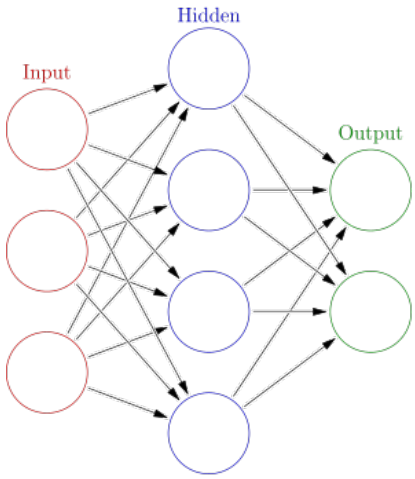


Figure 2: Simple ANN diagram from Wikimedia commons

A network is often initiated with random weight at every node, then trained for multiple epochs. In each epoch, the network takes in a series of inputs from a dataset and generates corresponding outputs. Then, the outputs by the network are compared with the desired

output from the dataset. The network then uses the difference in output to modify the weight at each node. After training, the network could be applied on a new input, and it would attempt to generate an output despite having never encountered the input.

3. *Pix2Pix*

The algorithm used in this program is Pix2Pix [1]. It was developed by researchers from UC Berkeley, and it could be used on a variety of image to image transformation tasks without modifying the loss function. It includes two parts: the generator that generates an output image from an input image, and the discriminator that determines whether one image is a valid transformation from the other. Noise is introduced in the form of dropout on various layers in order to randomize the output.

For the generator, an algorithm called U-net [2] is used. It utilizes encoder and decoder blocks as well as skip connections: the encoders downsample the input image until one layer called the bottleneck layer, after that the decoders upsample the input to the final output. However, skip connections deliver information from an encoder directly to the decoder with the same size of input without passing through the bottleneck layer. Thus, some details in the input could be conserved.

Pix2Pix has been used for many community programs and succeeded in most of the tasks [1, 4.7]. Therefore, it was selected as the algorithm for this task.

B. Programming

1. *Python*

The programming language used was Python 3.7.4 [3]. Python is an interpreted, high-level, general-purpose programming language. It was invented by Dutch programmer Guido van Rossum and first released in 1991. In 2008, Python 3 was released, with significant modification to the grammar from Python 2. Python featured dynamic typing, dynamic name resolution and a garbage collector for memory management. Python is also highly extensible with a very active programmer community.

2. *TensorFlow and Keras*

TensorFlow [4] is an end-to-end open source platform for machine learning developed by Google. It stores the model as tensors, which could be processed effectively on GPUs or TPUs (tensor processing units). Keras [5] is a high-level API for TensorFlow that allows users to easily build, train, evaluate, and deploy their neural networks. Together, they provide both beginners and experts a powerful tool in machine learning.

3. *Environment*

The machine used for this task was a Clevo P750TM1-G with Z370 chipset, i9-9900K CPU, RTX 2070 (mobile 115W) GPU and 2*16GB of DDR4 RAM. Geforce driver 436.02, CUDA 10.0 and CUDnn 7.6.2.24 from Nvidia was installed to enable TensorFlow to use GPU in training. Anaconda was used to set up a separate environment for TensorFlow. The version of TensorFlow used was 1.14.0 and other important packages such as PIL, numpy and sklearn were installed as well. The coding was done in Spyder 3.

The program

A. Training and testing Data

The data used for the program was images of galaxies from Hubble Space Telescope. For every image, there should be a galaxy at the center and the size is always $424 * 424$ pixels.

For training, 1000 images without obvious artifacts were manually selected, because the original dataset was both too big for the machine to handle (Memory was exhausted during processing of the full dataset), and it contained some faulty images in which either the galaxy was blurry, there was colored lines across the image, or the image was in shades of other color. See FIG. 3 for first 25 pairs of training data after shuffling (the two parts of images had been merged together in the plot).

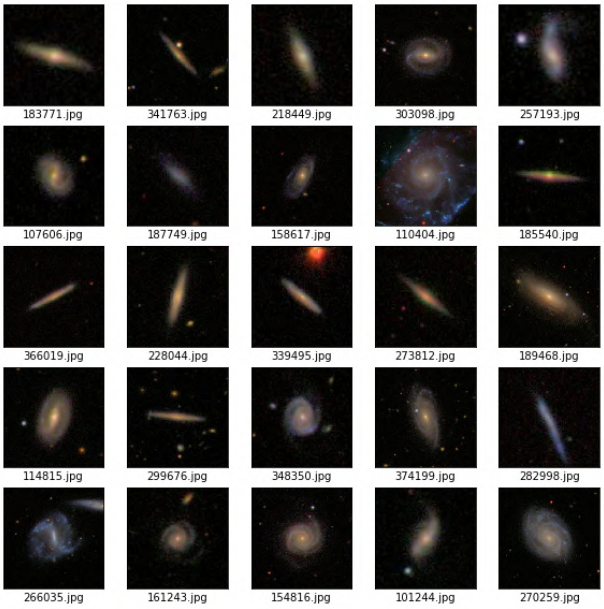


Figure 3: Examples of the training data

For testing, 12 images were selected to evaluate the program, and none of them were part of training data.

All of them were selected to pose challenges for the program. See FIG.4 for the test images in order (compressed into $3 * 4$ image array).

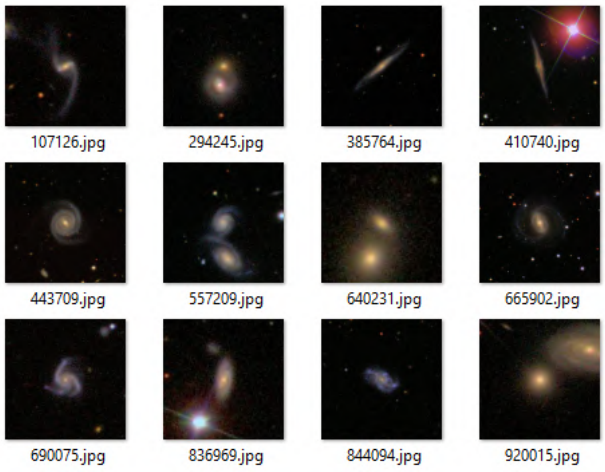


Figure 4: The test data

The 1st, 5th, 8th, 9th, and 11th image had special features such as spiral arms and rings. They were used to test the program's ability to reconstruct these features. It is worth noticing that the 8th image had a very faint ring, the 9th image had 3 spiral arms instead of 2, and the 11th image was rather irregular. Those images could be challenging for the program because a rotated image would be inaccurate. The 3rd and 4th image were galaxies pictured edge-on, testing the program's ability to reconstruct edge-on galaxies. Other images were elliptical galaxies which should be easy for the program to reconstruct. Of all images, the 4th and 12th had large objects on the top half, while the 2nd, 6th, 7th and 10th had objects obstructing the lower part. This is to test both the ability to filter out other objects in the input, and the effectiveness in removing the obstruction, which was the task of this paper.

B. Data preparation

In order to convert the raw images to training data, the images are cropped down to $256 * 256$ pixels in size while preserving the galaxy at the center. This is because Pix2Pix is most compatible for images with side lengths in powers of 2. After that, the image was split in half: the top half would be used as input while the bottom half would be the desired output. The input, output, and name of the image file (for demonstration purpose) was stored in numpy arrays. The three arrays were then shuffled randomly, saved to files for quick retrieval, and returned to main program. The code used to rotate the input image for better understanding by the program has been commented. Listing 1 was used to prepare and display the data:

Listing 1: Source code for data preparation

```
from __future__ import print_function
import numpy as np
from PIL import Image
from os import listdir
from sklearn.utils import shuffle

orig=424
iSize=256
iHalf=int(iSize/2)
crop=84
img_shape=(iHalf,iSize,3)
trainP='D:\\Files\\ISEC\\Datasets\\train\\'

def convimg(path):
    img=Image.open(path)
    img=img.crop((crop,crop,orig-crop,orig-crop))
    img=img.resize((iSize,iSize))
    arr=np.array(img.getdata(),np.float32)
    arr=arr.reshape(iSize,iSize,3)
    arr/=255
    out=[]
    for i in range(2):
        out.append(np.split(arr,[iHalf])[i])
#   out[0]=np.flip(np.flip(out[0],0),1)
    return out

def prep_data(path):
    imgsList=listdir(path)
    l=len(imgsList)
    x1=np.empty([l,iHalf,iSize,3],
                dtype="float32")
    x2=np.empty([l,iHalf,iSize,3],
                dtype="float32")
    iList=np.empty([l],dtype='object_')
    i=0
    for x in imgsList:
        arr=convimg(path + x)
        x1[i]=arr[0]
        x2[i]=arr[1]
        iList[i]=imgsList[i]
        i+=1
    x1,x2,iList=shuffle(x1,x2,iList,
                        random_state=114514)

    np.save("x1",x1)
    np.save("x2",x2)
    np.save("iList",iList)
    return x1,x2,iList

def syn_img(a1,a2):
    a1*=255
    a2*=255
#   a1=np.flip(np.flip(a1,0),1)
    syn=np.empty([iSize,iSize,3],dtype="uint8")
    for i in range(iHalf):
        for j in range(iSize):
            for k in range(3):
                syn[i][j][k]=int(a1[i][j][k])
    for i in range(iHalf,iSize):
        for j in range(0,iSize):
            for k in range(0,3):
                syn[i][j][k]=int(a2[i-iHalf][j][k])
    imgA=Image.fromarray(syn)
    return imgA,syn

x1,x2,iList=prep_data(trainP)
```

C. Model construction

The Pix2Pix code was slightly modified to fit this task, because the input and output were rectangular 1:2 images instead of square images. As a result, the size of input was modified and the last (7th) pair of encoder-decoder was removed to fit the rectangular shape of the data. Moreover, instead of completely removing dropout at one certain decoder block, the dropout was gradually reduced to 0. From the third decoder, dropout decreased by 0.1 every decoder until it reached 0.

The epoch definition was also modified: originally, one epoch passes when the program has been trained with the number of inputs equal to the size of training data. However, in this program, one epoch passed when the program took in an input. Hence, with a training data size of 1000, our epoch 1000 would be the original epoch 1.

The optimizer used was Adam and the loss function used was mean squared logarithmic error (it was combined with mae in the final model with default weight [1, 3.1], 1 to 100).

The following code was used to construct the model: See Listing 2 for the discriminator, Listing 3 for the generator, and Listing 4 for the full GAN model.

Listing 2: Source code for the discriminator

```
from keras.optimizers import Adam
from keras.initializers import RandomNormal
from keras.models import Model
from keras.models import Input
from keras.layers import Conv2D
from keras.layers import LeakyReLU
from keras.layers import Activation
from keras.layers import Concatenate
from keras.layers import BatchNormalization
from keras.layers import Conv2DTranspose
from keras.layers import Dropout

lf='mean_squared_logarithmic_error'
opt=Adam(lr=0.0002,beta_1=0.5)

def discriminator(img_shape):
        init=RandomNormal(stddev=0.02)
        in_src=Input(shape=img_shape)
        in_target=Input(shape=img_shape)
        merged=Concatenate()([in_src,in_target])
        d=Conv2D(64,(4,4),strides=(2,2),
              padding='same',
              kernel_initializer=init)(merged)
        d=LeakyReLU(alpha=0.2)(d)
        d=Conv2D(128,(4,4),strides=(2,2),
              padding='same',
              kernel_initializer=init)(d)
        d=BatchNormalization()(d)
        d=LeakyReLU(alpha=0.2)(d)
        d=Conv2D(256,(4,4),strides=(2,2),
              padding='same',
              kernel_initializer=init)(d)
        d=BatchNormalization()(d)
        d=LeakyReLU(alpha=0.2)(d)
        d=Conv2D(512,(4,4),strides=(2,2),
              padding='same',
              kernel_initializer=init)(d)
        d=BatchNormalization()(d)
        d=LeakyReLU(alpha=0.2)(d)
        d=Conv2D(1,(4,4),padding='same',
              kernel_initializer=init)(d)
        patch_out=Activation('sigmoid')(d)

        model=Model([in_src,in_target],patch_out)
        model.compile(loss=lf,optimizer=opt
                      ,loss_weights=[0.5])
        return model

d=discriminator(img_shape)
```

Listing 3: Source code for the generator

```
def encoder(layer_in,n_filters,batchnorm=True):
        init=RandomNormal(stddev=0.02)
        g=Conv2D(n_filters,(4,4),strides=(2,2),
             padding='same',
             kernel_initializer=init)(layer_in)
        if batchnorm:
                g=BatchNormalization()(g,
                                training=True)
        g=LeakyReLU(alpha=0.2)(g)
        return g

def decoder(in1,in2,n_filters,d=0.5):
        init=RandomNormal(stddev=0.02)
        g=Conv2DTranspose(n_filters,(4,4),
                        strides=(2,2),
                        padding='same',
                        kernel_initializer=init)(in1)
        g=BatchNormalization()(g,training=True)
        if d>0:
                g=Dropout(d)(g,training=True)
        g=Concatenate()([g,in2])
        g=Activation('relu')(g)
        return g

def generator(img_shape=(iHalf,iSize,3)):
    init=RandomNormal(stddev=0.02)
    in_img=Input(shape=img_shape)
    e1=encoder(in_img,64,batchnorm=False)
    e2=encoder(e1,128)
    e3=encoder(e2,256)
    e4=encoder(e3,512)
    e5=encoder(e4,512)
    e6=encoder(e5,512)
    b=Conv2D(512,(4,4),strides=(2,2),
             padding='same',
             kernel_initializer=init)(e6)
    b=Activation('relu')(b)
    d1=decoder(b,e6,512)
    d2=decoder(d1,e5,512,d=0.4)
    d3=decoder(d2,e4,512,d=0.3)
    d4=decoder(d3,e3,256,d=0.2)
    d5=decoder(d4,e2,128,d=0.1)
    d6=decoder(d5,e1,64,d=-1)
    g=Conv2DTranspose(3,(4,4),strides=(2,2),
                      padding='same',
                      kernel_initializer=init)(d6)
    out_img=Activation('tanh')(g)
    model=Model(in_img,out_img)
    return model

g=generator(img_shape)
```

Listing 4: Source code for the GAN

```
def gan(g,d,img_shape):
        d.trainable=False
        in_src=Input(shape=img_shape)
        gen_out=g(in_src)
        dis_out=d([in_src,gen_out])
        model=Model(in_src,[dis_out,gen_out])
        model.compile(loss=[lf,'mae'],
                      optimizer=opt,
                      loss_weights=[1,100])
        return model

gan=gan(g,d,img_shape)
```

Now a GAN model has been constructed. See FIG.5 for the diagram of the model.

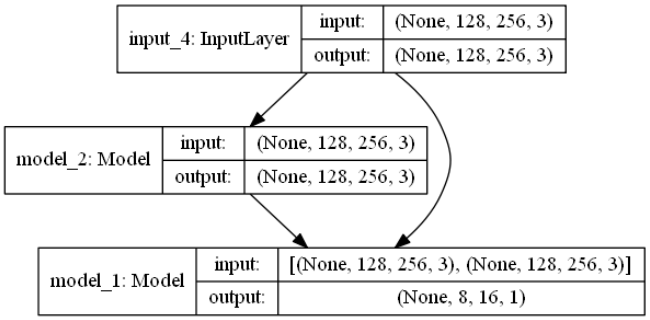


Figure 5: Simple diagram for the GAN

D. Training

The training and evaluation was carried out in one function. At every epoch, one random pair of data would be selected from the training data. The desired output from the real data would be labeled "1", and the generator will generate an output labeled "0". The two sets of data would then be passed to the discriminator, and its feedback would be returned to the generator.

At user defined epochs (from a tuple), the program would perform an evaluation with the test data provided, and store the results in lists to be returned after training.

See Listing 5 for the training section of the code.

Listing 5: Source code for training

```
testP='D:\\Files\\ISEC\\Datasets\\test\\'
testEpochs=(1,10,100)
E=100

def genR(x1,x2,n,shapeP):
        ix=np.random.randint(0,x1.shape[0],n)
        X1,X2=x1[ix],x2[ix]
        y=np.ones((n,int(0.5*shapeP),
                shapeP,1),dtype='float32')
        return [X1,X2],y

def genF(g,samples,shapeP):
        X=g.predict(samples)
        y=np.zeros((len(X),int(0.5*shapeP),
                shapeP,1),dtype='float32')
        return X,y

def test_model(path):
    imgsList=listdir(path)
    l=len(imgsList)
    Iin=np.empty([l,iHalf,iSize,3],dtype="float32")
    Id=np.empty([l,iHalf,iSize,3],dtype="float32")
    Itest=np.empty([l,iHalf,iSize,3],dtype="float32")
    temp=np.empty([1,iHalf,iSize,3],dtype="float32")
    syn1=[]
    syn2=[]

    i=0
    for x in imgsList:
        arr=convimg(path + x)
        temp[0]=arr[0]
        Iin[i]=arr[0]
        Id[i]=arr[1]
        Itest[i]=gan.predict(temp)[1][0]
        syn1.append(syn_img(Iin[i],Id[i])[1])
        Iin[i]/=255
        syn2.append(syn_img(Iin[i],Itest[i])[1])
        i+=1

    return syn1,syn2,imgsList

def train(d,g,gan,x1,x2,
          testEpochs,n_epochs,n_batch=1,n_patch=16):
    testResults=[]
    j=0
    for i in range(n_epochs):
        [x1R,x2R],yR=genR(x1,x2,n_batch,n_patch)
        x2F,yF=genF(g,x1R,n_patch)
        d_loss1=d.train_on_batch([x1R,x2R],yR)
        d_loss2=d.train_on_batch([x1R,x2F],yF)
        g_loss,_,_=gan.train_on_batch(x1R,[yR,x2R])
        print('>Epoch:[%d] d1[%.3f] d2[%.3f] g[%.3f]'
              %(i+1,d_loss1,d_loss2,g_loss))
        if i+1 in testEpochs:
            tR,temp,tN=test_model(testP)
            testResults.append(temp)
            j+=1
    return tR,testResults,tN

tr,trs,tn=train(d,g,gan,x1,x2,testEpochs,E)
```

During training, indication of epoch number and losses would be in the form of FIG.6 .

```
>Epoch:[1] d1[0.046] d2[0.116] g[18.233]
```

Figure 6: Example of output during training

E. Output

After the training, an image would be constructed from the lists returned by the train function. It would include both the real images and the images constructed at different epochs. Then, it would be saved and displayed to the user. Finally, a beeping sound would be played to indicate the user that the training was complete.

See Listing 6 for the code for output.

Listing 6: Source code for output

```
import winsound

x=(len(testEpochs)+1)
y=(len(tn))
for i in range(y):
    for j in range(x-1):
        if j==0:
            temp=np.concatenate((tr[i]
                                 ,trs[j][i])
                                ,axis=1)
        else:
            temp=np.concatenate((temp,
                                 trs[j][i])
                                ,axis=1)
    if i==0:
        fa=temp
    else:
        fa=np.r_[fa,temp]
imgf=Image.fromarray(fa)
imgf.save('out.png')
imgf.show()

frequency=2500
duration=1500
winsound.Beep(frequency,duration)
```

Connecting all the Listings together, a fully functional program could be constructed.

Analysis

The model was trained for 50000 epochs with testing at epoch 1, 1000, 10000, 20000, 30000, and 50000 twice, one with the input rotated and one without.

A. Rotated input

At first, the input was rotated by 180° by uncommenting the two lines in Listing 1. In this way, the input would be very similar to the desired output, and the program would not need to learn to rotate the input.

See FIG.7 for the results.

At epoch 1, the output consisted of mostly random noise. However, if examined closely, the shape of the flipped input could be seen as shades of green in the output, indicating the input was rotated.

At epoch 1000, although most noise was removed, the color of the output was not ideal, lacking contrast and appearing white for most images. Also, major obstructions on the top half of the input were either colored like the galaxy and preserved or converted to noise. Still, for image 1, 5, 6, 8, and 9, the spiral or ring structures were rather successfully reconstructed.

At epoch 10000, the coloring was greatly improved, being much closer to the real color of the galaxies. The obstructions on the top part were further weakened in the output, and noise was also reduced. However, spiral structures were reconstructed shorter and more blurry.

At epoch 20000, the improvements became more evident, but generally the spirals are slightly more distorted.

Later at epoch 30000 and 50000, the noise appeared again. The reason behind could be that the program was attempting to remove the shades cast by the obstructions in the input, but failed to do so because such problem was not present in the training data. Still, at epoch 50000, spiral structures were somehow generated again, and the colors were more accurate.

Over the whole training process, the influence from the big object in the 4^{th} and 12^{th} image could not be removed. Understandably, the program failed to generate 2 more spirals for the 9^{th} image. In all, this program performed considerably good in reconstructing the obstructed lower part of the 2^{nd}, 6^{th}, 7^{th} and 10^{th} image, and was in some degrees accurate for the 1^{st}, 3^{rd}, 8^{th} and 11^{th} image.

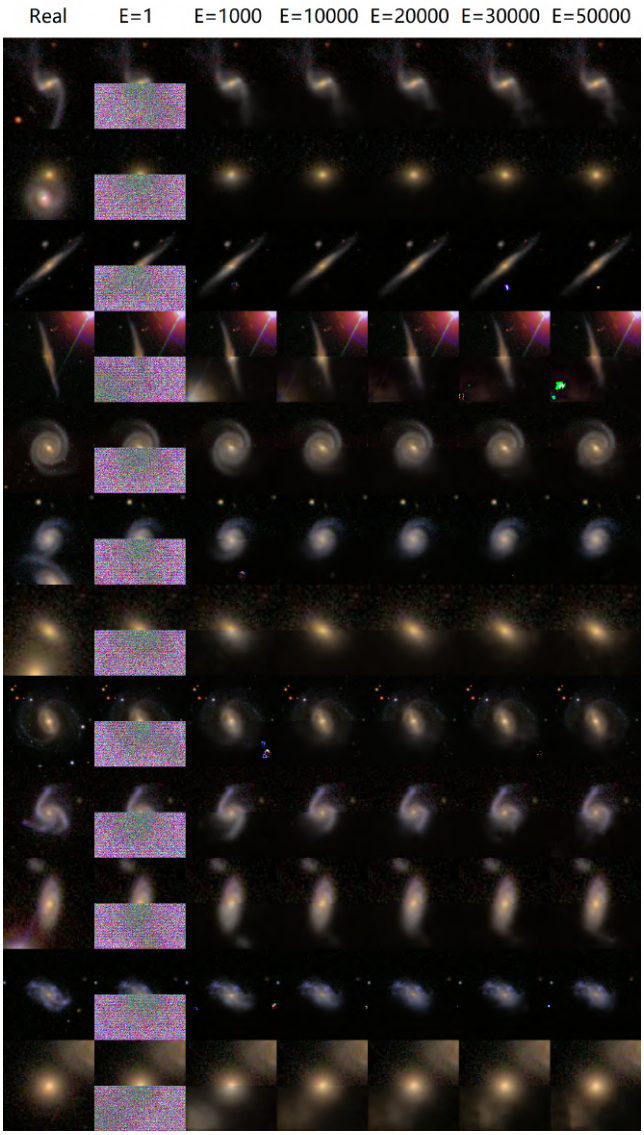


Figure 7: Results of training with rotated input

B. Non-rotated Input

When the input was not rotated, the difficulty for the program was much higher. See FIG.8 for the results.

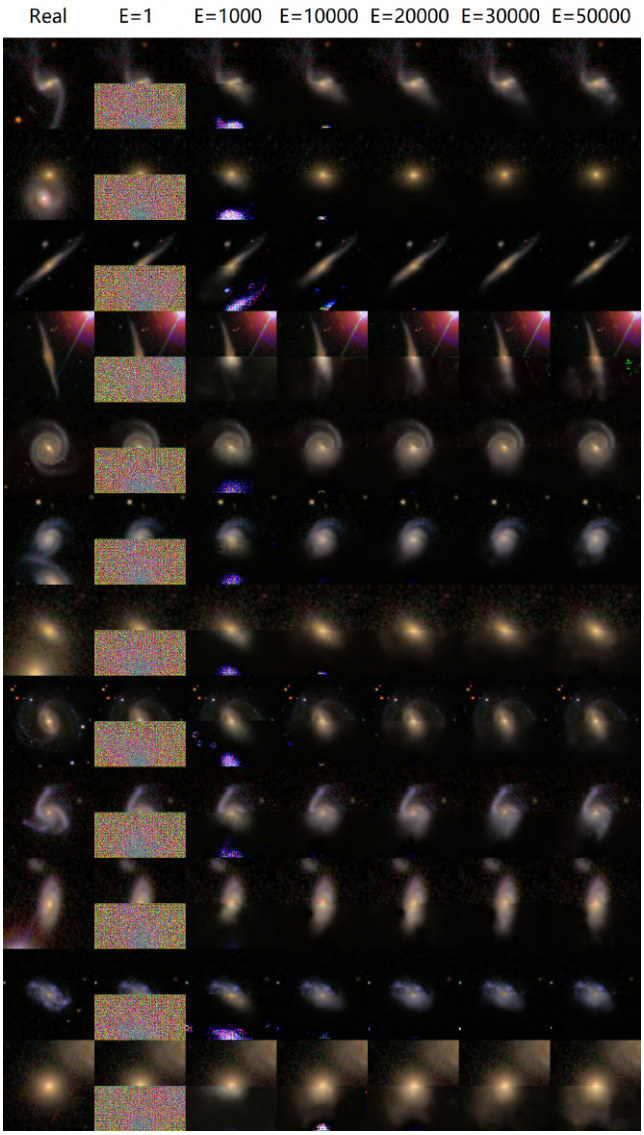


Figure 8: Results of training without rotating input

At epoch 1, the output was mostly noise, which is consistent from the previous result, and shades of non-rotated input could still be observed. It is worth noticing that the main color of the noise was different because the model was initiated with different random parameters every time.

At epoch 1000, blurry shapes of galaxies emerged at the top of the output, though they could not match the shape of the galaxies properly yet. Noise was reduced to a region corresponding to the bright parts in the input. The color of all output galaxy halves was light yellow, between that of elliptical galaxies and that of spiral galaxies. That meant the program could not color the galaxies properly, which was also consistent with the previous run.

At epoch 10000, the noise was further reduced, and in some images the noise was almost eliminated. The output could roughly emulate the input in shape, but could not generate any special features such as spirals. Also, the coloring for galaxies improved, beginning to differentiate between yellow elliptical galaxies and white-blue spiral galaxies.

At epoch 20000, all noise was almost removed except at the bottom of the output. The coloring continued to improve but the output was still blurry. The shape for some output could not fit the input. At epoch 30000, the improvement continued as at epoch 20000, except the noise at the 4^{th} appeared again, just as in the previous run.

At epoch 50000, some structures that could resemble a spiral structure finally appeared in the 1^{st} and 9^{th} image. The noise in the 4^{th} image was worse, but for most inputs, satisfactory output could be obtained.

Overall, the level of detail was lower compared to the run with rotated input, but the adaptation of the program could be observed easily. For the 2^{nd}, 10^{th} and 12^{th} image, the obstruction was removed, including the big galaxy in the top part of the 12^{th}.

Though the output was less satisfactory, this program showed better the capacity of the algorithm. The output was completely generated by the program, instead of filtered and slightly transformed in section XXIX A.

Conclusion

To conclude, Pix2Pix could be modified to accomplish galaxy image partial reconstruction. In the rotation approach, better quality output could be obtained, while the non-rotation approach fitted the concept of machine learning better, and could solve some problems in the rotation approach. Further study could be done on further fine-tuning the model, enlarging the variety of the training dataset, training the model for more epochs, and train with higher resolution images. Generally, a more powerful machine would be required.

Acknowledgments

I would thank the developers of Python, Anaconda, Spyder, TensorFlow, Keras, Pix2Pix, and the communities for those programs. They provide wonderful tools and solutions for me to complete this study. Also, I thank Oz, Radka, Carlos and all participants at ISEC 2019. The two weeks we had was extraordinary and I learned a lot from you.

[1] Phillip Isola,Jun-Yan Zhu,Tinghui Zhou,Alexei A. Efros. *Image-to-Image Translation with Conditional Adversarial Networks*. arXiv:1611.07004 [cs.CV].

[2] Olaf Ronneberger, Philipp Fischer, Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv:1505.04597 [cs.CV].

[3] Python Software Foundation. *python.org*. 2019.

[4] Google. *tensorflow.org*. 2019.

[5] François Chollet. *keras.io*. 2019.

Analysis Of The Spectrum Of a Spinning Black Hole's Accretion Disk

Vivaan Parvinder Singh, Efanova Antonina
International Science Engagement Challenge 2019

This study presents the analysis of the spectrum of a rotating black hole's accretion disk obtained by the NICER telescope using the example of MAXI J1535-571. The introduction discusses the theoretical component of the given topic, touching upon topics such as spectrum and spectral lines, the Doppler shift, X-ray binary stars and their constituents: Ergosphere and Roche Lobe. The main part describes in detail the method for analysing the spectrum of the accretion disk including data cleaning, extraction of the spectrum and analysis.

Introduction

Figure 1: The NICER Telescope sourced from NASA

NICER Telescope

The Neutron Star Interior Composition Explorer (NICER) is stationed in the international space station controlled by NASA. It was launched in 2017. It contains one of the highest X-Ray sensitivity sensors that detect radiation ever built. Although its main objective was to find out what was at the core of neutron stars, it enabled and helped us find many properties about compact objects such as spinning (Kerr) Black Holes. NICER provides scientists an opportunity to observe the extraordinary gravitational, electromagnetic, and nuclear-physics environments emitted by neutron stars [1].

Theory

Flux and Luminosity

What is Flux?

Flux is the amount of particles that is received by a detector per unit time. In this case it is the number of photons that reach the detector per second.

$$L = wr^2F \tag{1}$$

where w (solid angle), F (flux density observed at a distance r), L (total flux of a star or luminosity).

Flux density is the number of particles that is received by a detector per unit time per unit distance. In this case it is the amount of photons that reaches the detector per second per cm squared.

What is Luminosity?

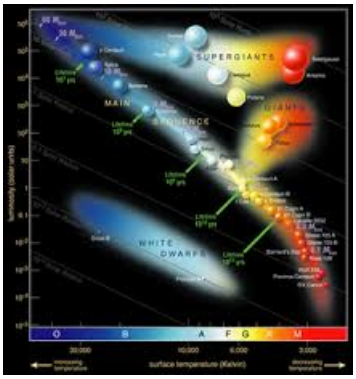


Figure 2: The Hertzsprung-Russell Diagram image taken from Wikipedia

Luminosity is the quantity of light energy emitted by a star in a given amount of time. It is measured in joules per second.

The luminosity of a star is dependent on two factors:

Surface Temperature - Hotter stars are more luminous than cooler stars. The relationship between luminosity and surface temperature is shown through the Hertzsprung-Russell Diagram.

Radius - The larger the radius of the star, the brighter it will be as bigger stars are more luminous.

Luminosity is determined in two ways:

- Apparent Magnitude - The apparent magnitude of the star is its luminosity or brightness of the star as it appears from Earth affected by several factors including its actual/natural brightness, its distance from the Earth and any celestial body in the line of sight of that star.

- Absolute Magnitude - The absolute magnitude of a star is its luminosity from a standard distance of 10 parsecs or 32.6 light years from Earth [2].

Black Holes

Black holes are formed from the implosion of a massive star at the end of its lifetime. This is followed by a supernova explosion and another implosion of material moving towards this dense core of the collapsed star to make an ultradense point. Depending upon the mass of the progenitor, this ultradense core either becomes a neutron star or if the mass is enough, then it becomes a black hole.

Characteristics of a black hole:

Black holes are characterised by two aspects only and can only be differentiated using these two aspects: angular momentum and mass.

Black holes without (zero) angular momentum (and therefore no spin) but with mass are known as Schwarzschild Black Holes. The Schwarzschild Radius is the radius of a black hole; the region in which the black hole's gravitational force acts on light. The outline of this region is defined as the Event Horizon.

Characteristics of a spinning black hole:

A black hole with spin is a black hole with angular momentum. Black holes which have not only mass but also angular momentum are often called Kerr Black Holes. However, a black hole cannot be seen due to the fact that light cannot escape from a Black Hole. Scientists observe the paths and orbits of the stars in the X-Ray Binaries and analysis from these observations allow scientists to assess whether it is a black hole, as black holes would cause these orbits to change as a result of their gravitational effect/influence. Matter that is not yet consumed by the black hole and taken inside the event horizon, lies in a region just outside of the event horizon called an Accretion Disk. This is the region where the matter rotates with angular momentum as a consequence of the black holes axis of rotation. The matter contained within the accretion disk is subject to the black hole's gravitational force and is what keeps it around the black hole.

Ergosphere:

The region outside of the outer event horizon is known as the Ergosphere. In this region, everything (all spacetime and matter) is forced to rotate as a result of the black hole's rotation through an effect called frame-dragging. Because of the fact that everything is forced to rotate within the ergosphere, the edge of this region is known as Static Limit. So, due to frame-dragging, nothing can be stationary in the ergosphere with respect to an observer outside of the ergosphere, hence the static limit resides on the edge of the ergosphere. The size of the ergosphere is determined by the gravity and angular momentum of the black hole. Due to there being a difference in angular momentum in the poles (where there is no angular momentum) and the equator (where there is most angular momentum) it causes the ergosphere to have an oblate shape [3].

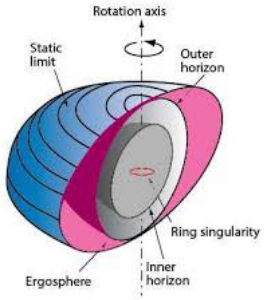


Figure 3: Representation of a Kerr Black Hole sourced from Semantic Scholar

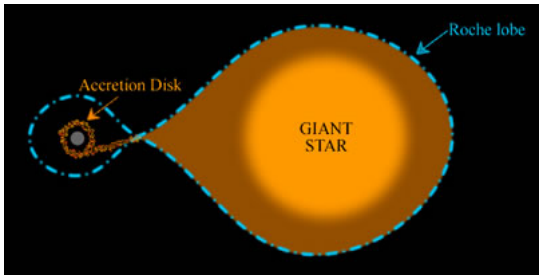


Figure 4: A picture of a Roche Lobe sourced from Centre for Astrophysics and Supercomputing

Roche Lobes

The Roche Lobe is the area around a star in a binary system within which orbital material is gravitationally bound to this star. This is an approximately drop-shaped region bounded by the critical gravitational equipotential, with the top of the drop directed to another star. In a non accelerating, rotating binary system with two objects, at some point the centrifugal and gravitational forces equal out and create a Lagrangian Point [4].

X-Ray Binaries (XRB)

X-Ray Binaries are binary systems with a compact star and a non-compact star. In order for it to be classified as an X-Ray binary system it must emit x-ray radiation and this is only possible through mass transfer amongst the compact and non-compact stars. Mass transfers allow the accretion disk to become more dense and permit high energy collisions between particles, which is what gives this x-ray radiation.

There are two types of XRBs:

- High Mass X-Ray Binaries are binary systems with compact stars and giant stars. The mass transfers prevail in the form of stellar wind and they are usually form near star forming galaxies.
- Low Mass X-Ray Binaries are binary systems with compact stars and inter-medially massed stars

with mass transfers through Roche Lobes overflow. However, unlike High mass X-ray binary systems, the mass transfers between stars is continuous. They are situated far from the star forming regions as they are older binary systems [5].

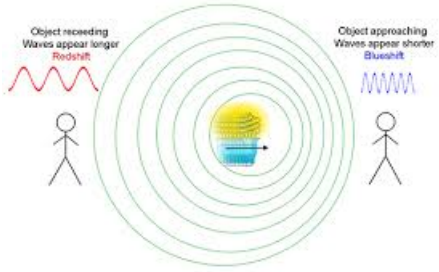


Figure 5: Doppler Effect sourced from NASA

Doppler Shift

The Doppler effect is a change in the frequency and wavelength (it is recorded by the receiver) caused by the movements of both the wave source and the receiver. Moreover, the motion of the medium in which the waves move is not related to this motion, and the speed of the wave depends on the characteristics of this medium. The wave source itself can no longer influence the further behaviour of the waves.

Gravitational Redshift The concept of gravitational redshift is based on the Doppler Effect. Gravitational redshift is the Doppler Effect in practice in a gravitational influence (in interstellar space). It is the increase in observed wavelength and a decrease in observed frequency from the perspective of the observer or the receiver. It is through this concept scientists are made aware if, for instance, galaxies are moving closer or further away from us. In essence, it is the wavelength and frequency of the photon of light emitted by that galaxy or celestial body that helps scientists decipher whether it (the celestial object or body) is indeed moving away from us. A photon is subject to gravity, so it is affected by the gravitational force of very big and dense celestial objects such as black holes and neutron stars. As a result, the photon needs to escape the gravitational pull from these objects and it is redshifted in the process [6].

Spectrum

If a light coming from a source is detected without any interference, we would receive a continuous spectrum. However, if the incoming light is affected by an outside influence such as a cold gas cloud, then the elements in that gas cloud will absorb some of the energy of the light

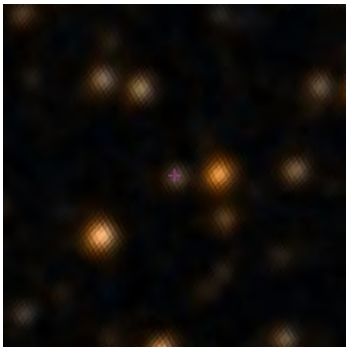


Figure 6: Native Optical Image of MAXI J1535-571 sourced from ResearchGate

and the remainder of the energy of the light will be detected by a detector and it will produce an absorption line.

However, if a light coming from a source is affected by an outside influence such as a hot gas cloud, then some of the energy of the incoming photon will be absorbed by the elements which will emit this energy in the form of radiation and the spectrum of this emitted radiation will be received by the detector. The pattern formed i.e. the spectrum, is called emission lines [7].

Spectral Line Broadening Line broadening is absorption lines or emission lines spreading across a greater wavelength, or frequency range. If an object which emits radiation is revolving around a compact star its absorption or emission lines will be redshifted or blue shifted. Since observation times are very long, it is not possible to observe spectral lines revolving around a compact source as a single line. Therefore, superposition of the lines in the spectrum and this forms spectral line broadening in Gaussian shape [8].

MAXI J1535-571

MAXI J1535 - 571 is a part of a Low-Mass X-Ray Binary system and is a Kerr Black Hole candidate. Its spin parameter is measured as 0.994 by Miller et al. 2018 and scientists have detected a line broadening emission of Iron from this black hole candidate. It also contains an accretion disk [9].

Method

Observation

The observation ID is 10560360106. The date this observation was September 13, 2017. The exposure time of the uncleaned data was 9628 seconds, however after this data was (refined)/cleaned it reduced the exposure time to 7413 seconds.

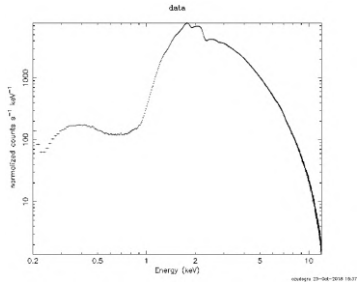


Figure 7: Representation of the spectrum from 0-12 keV

Data Cleaning

In order to clean and compile the data, three commands from the Nicerdas software were used.

- **NICERCAL** - The purpose of using this command was to calibrate the data and check for obvious errors such as checking for angles and if the image is titled.
- **NIMAKETIME** - This command checks the auxiliary files (file containing time and orbit information) as well as any errors in the log file.
- **NIMERGETIME** - This command is performed to obtain a clean observation by compiling seven different observations (mpu's) and merging them.

After we run the operations the exposure time reduced to 7413 seconds from 9628 seconds.

Extraction of the Spectrum

XSELECT is a software designed by NASA that enables one to extract the spectrum of MAXI J1535-571 through observing and logging the frequency of light and number of photons. With the help of the Planck-Einstein equation, i.e.

$$E = hf \tag{2}$$

the frequency observed by NICER is converted to the respective energy values.

Analysis

XSPEC is a software that was developed by NASA and is used to plot a graph of these energy values against the number of photons. However, as a consequence of errors in the instruments of NICER, we omit the spectrum below 1 keV. The graph produces a mathematical shape that clearly resembles the power law (after the omission of the incorrect data below 1 keV).

As a result of the errors due to the NICER instruments, we set boundaries to our spectrum from 1-10 keV.

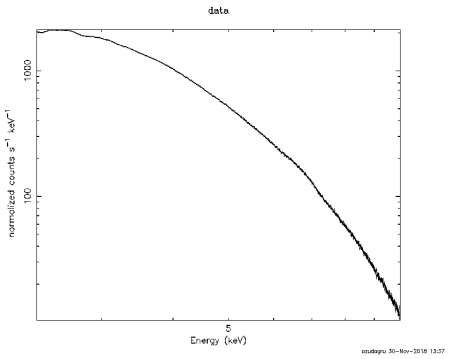


Figure 8: The spectrum from 1-10 (Power Law)

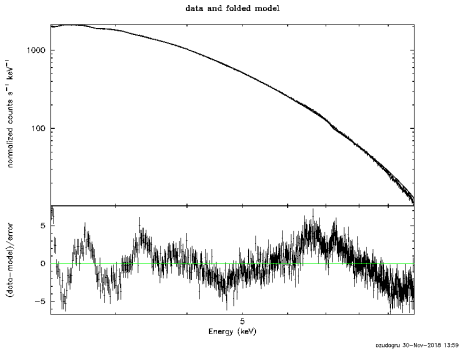


Figure 9

Although, the spectrum of the NICER extends from 0-12 keV, we do not use the results from the ranges 0-1 and 10-12 keV as it produces unreliable results.

In order to better analyse the spectrum, several models were applied:

- phabs (Photon absorption) - It describes the photons absorbed by the accretion disk
- diskbb (Accretion disk) - It describes that we have an accretion in our source
- gauss (Gaussian shape) - The presence of an iron emission in our source
- relline (Relativistic line) - It is used to take the relativistic effects into account
- pow (Power law) - The spectrum fits the shape of the power law shape

$$phabs(diskbb + gauss + relline + pow) \tag{3}$$

We observed the emission line and acquired detailed information about the emission line by using the models. Through measuring the width of the emission line of the spectrum we acquired the spin parameter of MAXI J1535-571.

Results and Conclusion

This is a table of results acquired after administrating NICERDAS commands:

Iron emission line (keV)	Spin parameter (Speed of light)
6.40	0.998
6.37	0.998

Our initial task was to familiarise ourselves with some convoluted concepts relating to black holes, most of which we had not come across prior to the initiation of this project. Although comprehending such complex ideas was difficult, it was rewarding and captivating. The next part of our assignment required knowledge in the language of the Ubuntu Operating System. This was stupefying and made us develop and acquire new skills that we did not have prior to the initiation of the project. Our ultimate objective was to analyse the spectrum of the accretion disk of a black hole using the example of MAXI J1535-571 by calculating the spin parameter. This allowed us to learn more about spinning black holes and X-ray binaries. Firstly, we had to acquire data from the NICER telescope, however, some of its data may be mistaken or have a large error arising from its movement or the penetration of light particles into the intermolecular cloud, these errors are displayed on the graphs above. To avoid them, we used several operations, which are described in detail in this work. Besides, in order for us to measure the spin parameter, we had to obtain the width of an iron emission line. Therefore, the width of the iron emission line is produced due to gravitational line broadening; the emission line is blue-shifted and red-shifted, therefore, it produces this wide shape called the Gaussian shape. But, the width of the iron emission line is directly related to the general relativistic effects of the spin of the black hole, which is directly related to the rotational spin of the black hole. Thus, we used the obtained data to determine the rotation parameter of the black hole.

After the administration of the NICERDAS programs, we calculated that the spin of MAXI J1535-571 which was **0.998** times the speed of light. However, this was contrary to the results attained by Miller et at. and the spin of this black hole was **0.994** times the speed of light. A plausible reason for a difference in results could be the consequence of the quantity of exposure time. Finally, we can conclude that even though there was a slight difference in the results that the results we have obtained are reliable nonetheless as the methods we used were identical, but our increased exposure time allowed us to gain a more reliable result due to a greater number of results to calculate an average from.

[1] Özgür Can Ozudogru. *"Timing and Spectral Analysis of Black Hole Candidate X-Ray Binary MAXI J1535-571 from NICER Observation"* Middle East Technical University, April 5, 2019.

[2] James E. Heath *"Jim Heath's Starry Site"* Austin Community College, 1999.

[3] Flint Wild *"NASA Knows!"* June 4, 2014.

[4] Francesca Valsecchi, Frederic A. Rasio, Jason H. Steffen *"Earth and Planetary Astrophysics"* Cornell University, 15 August 2014.

[5] – *"X-ray Binaries"* The University of Oxford Department of Physics, 2019.

[6] – *"Imagine the Universe!"* the High Energy Astrophysics Science Archive Research Center (HEASARC), 05 May 2016.

[7] – *"The SAO Encyclopedia of Astronomy"* Swinburne University of Technology.

[8] – *Physics Open Lab*, September 7, 2017.

[9] H. Stiele and A. K. H. Kong *A Spectral and Timing Study of MAXI J1535–571, Based on Swift/XRT, XMM-Newton, and NICER Observations Obtained in Fall 2017* The American Astronomical Society, November 21, 2018.

Creating a Python Script for a Timelapse

Eva and Wytske
International Science Engagement Challenge 2019

By learning the basics of Python we wrote a script to create a timelapse from a set of images. We had an interest in furthering our knowledge of programming, and this project allowed us to learn a lot about Python within a short amount of time while also having a specific goal to work towards. Most of the time spent was not on developing the timelapse script, but rather becoming comfortable using Python to write simple scripts. The timelapse code was created relatively fast, giving us time to take enough pictures of a working session to create our timelapses and learn more about camera optics. The process was an overall positive experience as we met our goal and received a lot of knowledge in limited time.

Introduction

Our project aimed to create a Python script which would form a timelapse from a set of pictures. Our knowledge of Python was very limited. Therefore we first had to learn the basics of writing Python scripts. We used the Jupyter notebook to write and execute the script, and used the os and OpenCV libraries to write the final code. The goal was to create the script with as few libraries as possible, so we could learn more of Python and not let pre-made functions do most of the work for us. This report will briefly cover the steps taken to learn the basics of Python, explain the final script used for the timelapse and explain the optics of the camera used to create the timelapse photos.

Python Process

In order to start learning Python we learned basic plotting using the pyplot package within the matplotlib library. We learned how to import, open, read, transform, and plot data from a .txt file. Creating plots from the files became increasingly more complicated as headers and data from different sources were introduced. This side project was time consuming, but taught us many skills which made the main project easier to complete. Once we were more comfortable with Python, we installed and used OpenCV to create the timelapse script.

To learn about creating videos with OpenCV we followed a tutorial from Medium (Enrique, A. 2018) showing how to make a 10 second video of colour noise. Understanding this code was very useful for creating our own Python script, and we implemented some similar code in our project.

After some trial and error and finalizing our project we created the code seen in Figure 1.

First the two required libraries were imported; os and OpenCV. The os module was required in order to read a directory, or the contents of a folder. The OpenCV library allowed us to work with video and image formats in Python. The Frames Per Second (FPS), folder name of the pictures, and video name of the final timelapse were not specified, but instead became an interactive element

```python
import cv2
import os
from cv2 import VideoWriter, VideoWriter_fourcc

Folder = input('What is the folder name? ')
FPS = input('What is the fps? ')
File_Name = input('What is the name of the video? ')

lst = os.listdir("C:/Users/Wytske/Documents/ISEC/"+Folder)
image = cv2.imread("C:/Users/Wytske/Documents/ISEC/"+Folder+"/"+lst[0])
height, width = image.shape[:2]

fourcc = VideoWriter_fourcc(*'MJPG')
video = VideoWriter("./"+File_Name+".avi", fourcc, float(FPS), (width, height))


y = 0
for _ in range(len(lst)):
    x = lst[y]
    img = cv2.imread("C:/Users/Wytske/Documents/ISEC/"+Folder+"/"+x)
    video.write(img)
    y = y + 1

video.release()
```

Figure 1: The final Python script that creates timelapses from a set of images

as seen in Figure 2.

```python
Folder = input('What is the folder name? ')
FPS = input('What is the fps? ')
File_Name = input('What is the name of the video? ')
```

Figure 2: To make the code more user friendly, certain variables were defined through a user input function

This was in order to make the script easier to use, as different timelapses have different values and names for these three variables. The "input()" function allows the user to write in the answers to the question in a separate box outside of the code itself when the code is executed. As the user inputs the values, they are assigned to the variables "Folder", "FPS" and "File Name". This is more efficient as the user is not required to change the code itself for every different timelapse. An example of how this interaction looks like in the Jupyter notebook can be seen in Figure 3.

To simplify the input for the "Folder" variable, it was assumed that all of the timelapse picture folders were within the same larger folder named "ISEC" on the computer we used to execute the script, as shown in the path to the files in Figure 4. On other computers this path would be different. Alternatively, the input question could ask the full path to the folder, which would be

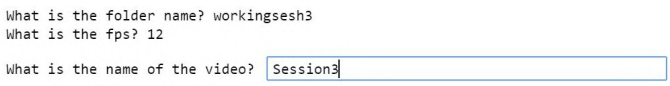


Figure 3: When the code is executed, the user is able to type in the folder name which contains the images, the frames per second (FPS) and the desired name of the output video

more time consuming but also be more general.

```
lst = os.listdir("C:/Users/Wytske/Documents/ISEC/"+Folder)
image = cv2.imread("C:/Users/Wytske/Documents/ISEC/"+Folder+"/"+lst[0])
height, width = image.shape[:2]
```

Figure 4: All the timelapse picture folders were within the same larger folder named “ISEC”

To make a video, the video’s dimensions need to be specified. As the video will be made up of already taken images, the dimensions of the video must be the same as that of the images. In order to find out what the dimensions of the images are, the contents of the specified “Folder” were listed. This was done using the “os.listdir()” function as shown in Figure 4. This function comes from the os library and reads the name of each file in a specified folder, and makes it into a list which we named “lst”.

The first image is opened with the function “cv2.imread()” from the library OpenCV. The height and width is found using the “.shape” function . These values are then assigned to the variables “height” and “width” as shown in Figure 4.

The details of the output video were defined through OpenCV. FourCC refers to “Four Character Code” which specifies what format the input material is in, which in this case is jpg. The FourCC for jpg is MJPG. The output is defined as “video” using the function “VideoWriter”. In order, it specifies the output video name (File Name), the FourCC, the specified FPS, and the defined dimensions of the images. When the input values are given, they are automatically strings, however in the “VideoWriter” function the FPS value is required to be an integer or float, therefore there is a “float()” function to transform the input value into a float. This section is shown in Figure 5.

```
fourcc = VideoWriter_fourcc(*'MJPG')
video = VideoWriter("./"+File_Name+".avi", fourcc, float(FPS), (width, height))
```

Figure 5: The FPS value is required in the “VideoWriter” function to be an integer or float

In order to compile the images into a timelapse a “for loop” function was used. To loop through every image in the file, the loop continued for a specific “range”. This range was defined as the length of the list of items in the specified folder, meaning the loop will repeat for every item in the folder. The value of “y” represents the different items in the list, and is increased by one at the end of the loop so every image will be read and opened. Once an image was opened, it was immediately written into the already defined video using the “.write” function. When all images have been appended to the video, the loop ends and the timelapse is saved using the “.release” function as shown in Figure 6.

```
y = 0
for _ in range(len(lst)):
    x = lst[y]
    img = cv2.imread("C:/Users/Wytske/Documents/ISEC/"+Folder+"/"+x)
    video.write(img)
    y = y + 1

video.release()
```

Figure 6: To read and open an image, the items in the list are represented as the value of “y” which is increased by one, the “.write” function is used to define the video and the “.release” function is used to save the timelapse

Camera and Optics

In order to test and use our code, we decided to take the “Working Session” as our scene for the timelapse project.

A Nikon D90 camera was used to take the photos needed to make the video. In order to take a well exposed photo for time-lapse photography, it is necessary to establish some manual settings instead of automatic settings in the digital camera due to the manual mode is going to give much more control over the look of the photos, so the settings of the digital camera will not change from shot to shot unless these would be changed. The four important settings in the camera are aperture, ISO, shutter speed and white balance.

The first setting, the aperture, is the opening in the lens. When the shutter release button of the camera is pressed, a hole opens up that allows the camera’s image sensor to catch a brief sight of the scene that it is capturing. The aperture setting impacts the size of that hole. The camera captures more light when the hole is larger and less light when the hole is smaller. The aperture is measured in f/numbers. For example f/2.8, f/4, f/5.6, f/8, f/22 etc. Moving from one f-stop to the next doubles or halves the size of opening of the lens (and the amount of light getting through). So f/2.8 is in fact a much larger aperture than f/22.

The second setting, the ISO, measures the sensitivity of the image sensor. A higher ISO setting means that camera’s sensor is more responsive to light, so it needs less light to reach the sensor to create a well-exposed photograph. However, the quality of the images may be affected by digital noise, colors may be less vibrant and overall image contrast is flatter. When the ISO setting is low, the sensor is less responsıve to light, so, therefore, it requires more light to create a well-exposed photograph. Using a low ISO setting is preferable as it will result in

better quality photos. There will be little or no digital noise, and the colors and contrast in images will be better. Although 100 ISO is generally accepted as a 'normal' or 'standard' ISO in sunlight conditions, we used 800 ISO due to darker lightning conditions.

The shutter speed is the amount of time that light hits the image sensor. This is generally measured in fractions of seconds. The bigger the denominator the faster the speed. For example, 1/1000 is much faster than 1/30. The amount of time the shutter speed is opened directly affects the exposure of the photo. If a photo is too dark, it is underexposed, so details will be lost in the shadows and the darkest areas of the image. If a photo is too light, it is overexposed, so details will be lost in the highlights and the brightest parts of the image. Figure 7 shows how different shutter speed settings affect the exposure of a photo. All three photos had the same aperture and ISO setting of f/3.5 and 400 respectively. However, the shutter speeds were 1/13, 1/5 and 1/1.6, resulting in differently exposed images.

Figure 7: Example of an underexposed, correctly exposed and overexposed photo due to different shutter speeds but with the same aperture, ISO and WB settings for all three images

Finally, the white balance (WB) setting determines how the camera displays color under different lightning conditions, so that objects which appear white in person are rendered white in the photo. Moreover, it takes into account the "color temperature" of a light source, which refers to the relative warmth or coolness of white light. It is measured in kelvin (K), so the correct camera setting depends on the color temperature of the light source. The color temperature of 10000 - 15000 K corresponds to the clear blue sky light source, the color temperature of 5500 - 6500 K belongs to the average daylight light source and the color temperature of 4000 - 5000 K corresponds to the fluorescent light source.

As an example, Figure 8 shows a picture taken with a white balance setting of 3030 K and 8330 K respectively, but the same aperture, shutter speed and ISO.

Once we learned about and understood the various

Figure 8: Example of an image with different color temperature, an aperture of f/3.5, a shutter speed of 1/8 and an ISO of 200

manual camera settings, we decided to place the camera in the corner of the working room so we could capture a view of everyone working. We were able to start taking photos for our timelapse video with an aperture of f/4.2, an ISO of 800, an exposure of 1/40 and a color temperature of 3300 K. This white balance temperature has chosen due to the fluorescent light in the room.

We wanted to shoot as many photos as possible to make our video longer, so an intervalometer was needed to take the images automaticaly during the entire "Working Session". This is an external device that attaches to the camera, so it is useful due to its automatic shooting function. Multiple timing options can be configured so that it is not necessary to press the shutter button to take every single image. We specified that the images captured on the intervalometer of the self-timer were taken at 30 second intervals.

Conclusion

We were able to use the Python code to make multiple timelapses, meaning we met our goal and completed the project. We aimed to write as much of the code ourselves as possible, and not rely on libraries which have functions that complete most of the work for us. We achieved this to an extent, but it may have been that there is a more raw method to creating the timelapse script. While going through Python tutorials, we learned that there are multiple methods of approaching a problem, and to create a timelapse we only considered one method. In future projects we may consider trying multiple different approaches to learn more methods and techniques. Additionally we considered trying other programming languages also, but ended up focusing on learning only Python and about camera optics. With more time and resources it could be interesting to attempt the same project using other programs.

[1] Enrique, A. (2018) *How to create a video animation using Python and OpenCV*. [Online] Available from: https://medium.com/@enriqueav/how-to-create-video-animations-using-python-and-opencv-881b18e41397 [Accessed 25th of August 2019]

[2] Mansurov, N. (2019a) *Understanding Aperture in Photography*. [Online] Available from: https://photographylife.com/what-is-aperture-in-photography [Accessed 25th of August 2019]

[3] Mansurov, N. (2010) *What is ISO? The complete guide for beginners*. [Online] Available from:https://photographylife.com/what-is-iso-in-photography [Accessed 25th of August 2019]

[4] Mansurov, N. (2019b) *What is white balance and why is it important in photography*. [Online] Available from:https://photographylife.com/what-is-white-balance [Accessed 25th of August 2019]

[5] Rowse, D. (2019) *Introduction to shutter speed in digital photography*. [Online] Available from:https://digital-photography-school.com/shutter-speed/ [Accessed 25th of August 2019]

Analyzing Properties of Stars Through Nuclear Reactions

Sofía Marin-Quiros
International Science Engagement Challenge 2019

The purpose of this project is to analyse general trends in mass, luminosity, and temperature throughout the different types of nuclear fusion in main sequence stars. To make some commentary on this, the stellar modeling software MESA was used to model the main sequence lifetime over 7 different stars masses. One of the main trends found was that the hydrostatic equilibrium does not completely keep the star from gravitationally collapsing. Another important point of discussion found is that the CNO cycle requires higher temperature to produce more energy because temperature is actually just a "symptom" of a higher density. This density is the true condition that is required to create a substantial amount of energy.

Introduction

Summary

By definition, a star is a body of mass that is able to go through fusion. This report is focusing on the base of stars and looking at the two main types of fusion in a star's lifetime: the CNO cycle and the PP-chain. The project is focusing on the general trends of the two process in temperature, mass, and luminosity.

A star's lifetime starts with a molecular cloud in which gravity forces it to collapse into small particles. The particles formed will start spinning and create a protostellar disk orbiting a protostar. At this time, the mass will increase gravity and the radius will decrease, therefore, temperature increases along with density. The particles in the protostellar disk will start to stick together and form planetoids. If these planetoids gain enough mass, they will become round and the rest of the dust will scatter away. Throughout this star formation process, the protostar is still not a star since it does not have any fusion.

Hydrogen Fusion

As the protostar gains more and more mass, the temperature will keep increasing until it gets to 10 million Kelvin at which hydrogen fusion can begin. Once hydrogen fusion starts, the star has entered the Main Sequence. The Main Sequence (when a star fuses hydrogen at its core) is the stage in which stars spend most of their lifetime in. This is a process in which the star turns hydrogen fusion into helium and the reaction loses energy through neutrinos and photons. The force of the energy being released by nuclear fusion counteract the force of gravity, which prevents the star from collapsing during the main sequence. This concept is known as hydrostatic equilibrium. As the temperature rises the CNO cycle will be the main source of fusion. The CNO cycle is able to fuse carbon through reactions with catalysts.

The process of hydrogen fusion can happen through what is known as the proton-proton chain (PP chain). This chain has three different "paths" it may take to fuse helium known as PP-I, PP-II, and PP-III chains.

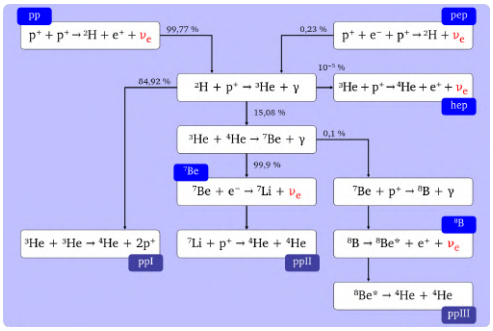


Figure 1: A graph shows the various paths included in the proton-proton chain. Image by Dorottya Szam, taken from `https://en.wikipedia.org/wiki/Proton-proton_chain_reaction`

The PP-I chain dominates the hydrogen fusion process. As temperature increases, PP-II and PP-III start competing. The PP-I chain is the fastest process to produce helium atoms. The PP-I chain starts with two hydrogen nuclei (protons) being combined to make a deuterium atom. This will also form a positron (an electron of positive charge) and a neutrino (a form in which energy is released in fusion). Afterward, the deuterium atom will react with another hydrogen nucleus to make hydrogen 3. This reaction loses mass through the release of a photon. Lastly two helium 3 react together to create a helium 4 and two protons.

PP-II and III chains are not as prominent at lower temperatures. The PP-II chain uses an electron capture of Be 7 from the PP-I chain (this is when the Be atom is made more stable by adding an electron). This turns into a Lithium atom, a neutrino, and a photon. The Li 7 atom then reacts with a proton to form two He 4 atoms. Lastly, the PP-III chain happens at extremely high temperatures and is very rare among stars in the main sequence. The last reaction of the PP-I chain changes: Be 7 atom and a proton which combines to turn into a photon and an extremely unstable atom (a B 8 atom). Since the B 8 atom is so unstable (has a half-life of 0.8 seconds) it decays quickly into a Be 8 atom, a positron, and a neutrino. The Be 8 atom is also very unstable so it decays into two He 4 atoms.

CNO Cycle

When the stars are F stars or larger (1.5 solar masses or more), the CNO cycle will rule the fusion holding the star in equilibrium. This is due to the higher temperature. The star produces an environment in which the CNO cycle happens faster at 18 million Kelvin or higher. The CNO cycle uses a set of catalysts to create He 4. Like the PP-chain, this cycle has three different paths to reach the same end product of He 4 and C 12. The first path's energy generation will depend on the concentration of hydrogen nuclei when fusion starts.

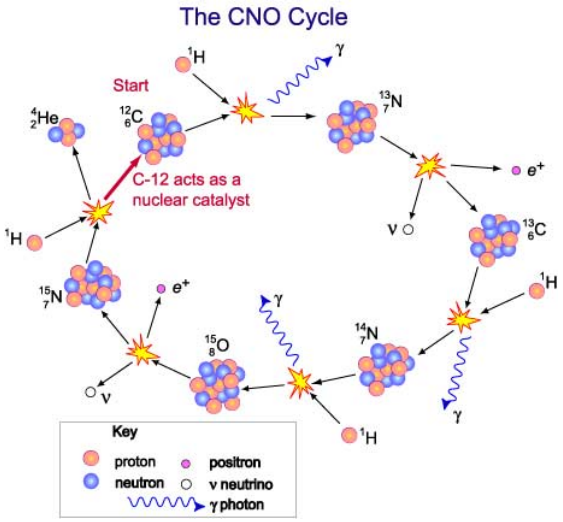


Figure 2: The reactions involved in the CNO Cycle

Methods

The purpose of this project was to find trends in temperature, luminosity, and mass and relating them to the CNO cycle and the PP-chain. The stellar model used took characteristics including: temperature of the center of the star, temperature of the outside layer of the star, the age, the luminosity of the star produced by the PP-chain, and the luminosity of the star produced by the CNO cycle. The stellar model output has a change in mass for each data set. The project looks at different masses since mass changes the temperature of the star and therefore the amount of energy produced by the CNO cycle and hydrogen fusion. Since the CNO cycle and the PP-chain happen in the core of stars on the main sequence, the output has each of the center temperature of the main sequence stars. The main sequence star types that output was taken about is an 0.2 solar masses star, 0.7 solar masses star, 1 solar mass star, 1.5 solar masses star, 2 solar masses star, 10 solar masses star, and a 50 solar masses star.

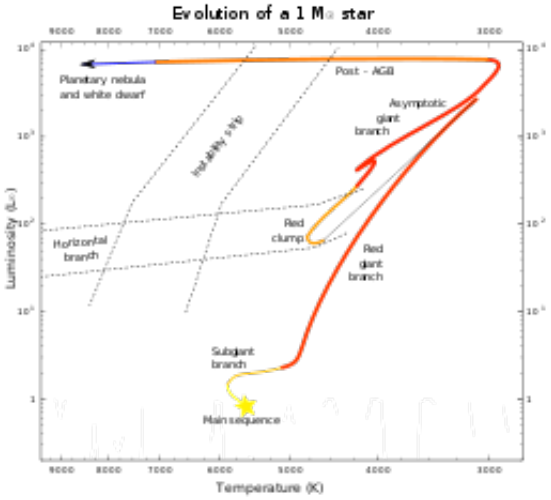


Figure 3: Complete evolutionary track of a 1 solar mass star.

To obtain this high variety of data, the stellar evolution model program MESA was used (Modules for Experiments in Stellar Astrophysics). This program can be used for modeling multiple stellar processes including nuclear astrophysics, galactic chemical evolution supernovae, stellar populations, stellar hydrodynamics, and stellar activity. The program is able to model the internal and outer layer characteristics of stars throughout the different stages before, during, and after the main sequence.

One of the features of the program is the ability to stop the data output when the star reaches a certain stage, temperature, mass, radius etc. The process of the modeling in this project is to stop output once the star reaches the horizontal branch. This is the stellar evolution stage after the main sequence in which the star's hydrogen core is exhausted and a small gravitational collapse happens, resulting in a temperature increase which is high enough to fuse helium. Because of this, the radius decreases, which maintains the luminosity constant in a flat line on the HR diagram after the main sequence (Luminosity vs. Temperature), hence called the horizontal branch.

While the output in this project contains a limit of the end of the output, it does not have a starting condition. Because of this, the output is given from the protostellar stage. However, during this stellar evolution stage, there is no hydrogen fusion or CNO cycle at the core. Since the focus of this project is on energy generation and there is none during this stage, this output is not included for analysis.

Data

After the data output was given by MESA, four graphs were made for each mass stated in the methodology section.

With all of the modeling output, three graphs were created in order to notice and understand trends within the main sequence.

First, the luminosities due to the CNO cycle and PP-chain were plotted versus the age of the star. This can be useful when looking at the changes in luminosities. By looking at the luminosities, trends on the amount of each of the energy generation processes can be seen. When reaching conclusions, luminosity graphs can be compared to temperature in order to analyze the role of temperature in energy generation.

Secondly, temperature of the center of the star over age was plotted. This helps gain a deeper understanding of how the temperature changes even within the main sequence. By comparing this graph and the luminosity, correlations can be drawn from the fluctuation of both luminosity and density.

Lastly the luminosity of the CNO cycle and PP-chain (separately) over the center temperature was plotted on the same graph. With this in mind, the project can directly show the mathematical relationship between the luminosity and temperature. Some assumptions can be made to compare how the abundance of each fusion process affects the temperature.

Log(Luminosity) vs. Age

The smallest mass of main sequence star (0.2 solar masses), shows a trend that is seen in all: the CNO line seems linear, however it is important to take into account that the y axis is a log of 10. Therefore, the data says that the luminosity given off by CNO vs age is an exponential equation. This is important that even despite the CNO cycle having a larger slope, the PP-chain sill dominates in luminosity.

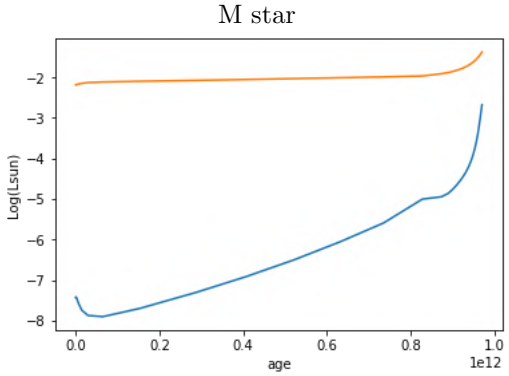


Figure 4: A graph of Luminosity vs Age of the pp-chain (orange) and CNO cycle (blue) for a 0.2 solar mass star

This F star has the mass of 1.5 solar masses at which the CNO cycle luminosity starts to become more prominent than the PP-chain does at the age of approximately 1.25 billion years. This is where luminosity of the CNO cycle becomes larger than the luminosity of the PP-chain (where the blue line y axis value is larger than the orange line). The important phenomenon to focus on is that the fusion process ruling the star can change within its life on the main sequence. Furthermore, when a 1.5 solar mass star is 2 billion years old, the luminosity from the PP chain increases and the luminosity from the CNO cycle decreases.

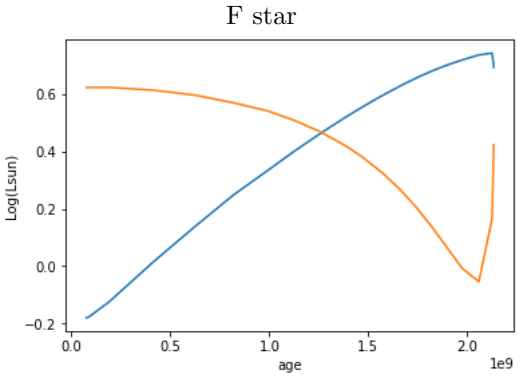


Figure 5: A graph of Luminosity vs Age of the pp-chain (orange) and CNO cycle (blue) for a 1.5 solar mass star

The 50 solar mass data graph shows two important trends seen in higher mass stars. In the CNO cycle luminosity does not increase by a significant amount. In fact, it appears that the blue line (representing CNO) has a slope close to 0.

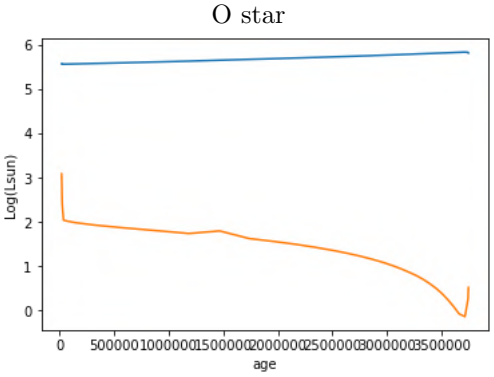


Figure 6: A graph of Luminosity vs Age of the pp-chain (orange) and CNO cycle (blue) for a 50 solar mass star

Log(Center Temperature) vs. Age

The 0.2 and 50 solar mass stars show the general trend of the temperature curves. The temperature is increasing at a constant rate until it approaches the end of the main sequence where there is a sharp increase. As the mass increases so does the speed at which the temperature rises.

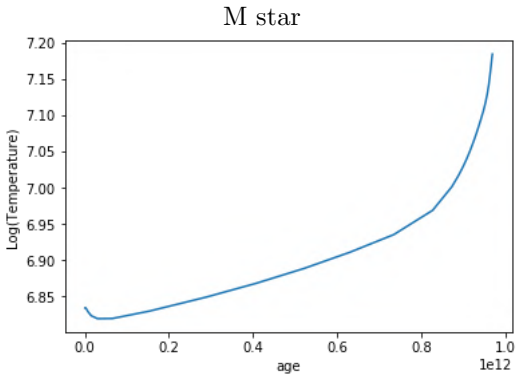


Figure 7: A graph of Temperature vs Age of the pp-chain (orange) and CNO cycle (blue) for a 0.2 solar mass star

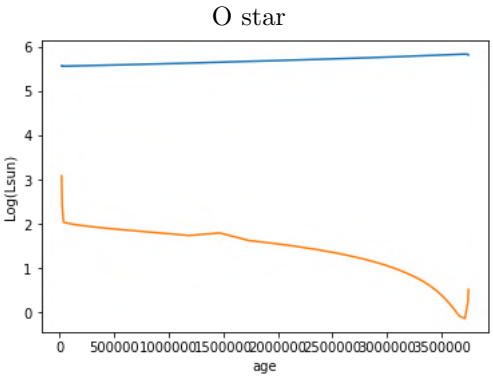


Figure 8: A graph of Temperature vs Age of the pp-chain (orange) and CNO cycle (blue) for a 50 solar mass star

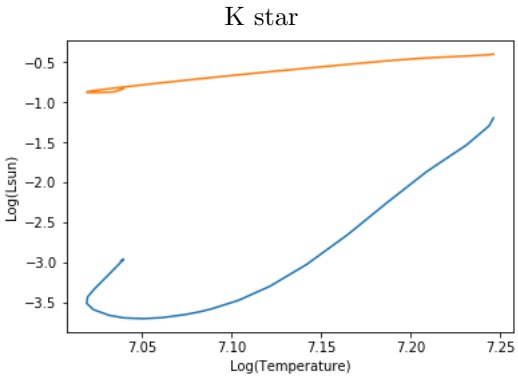


Figure 9: A graph of Luminosity vs Temperature of the pp-chain (orange) and CNO cycle (blue) for a 0.7 solar mass star

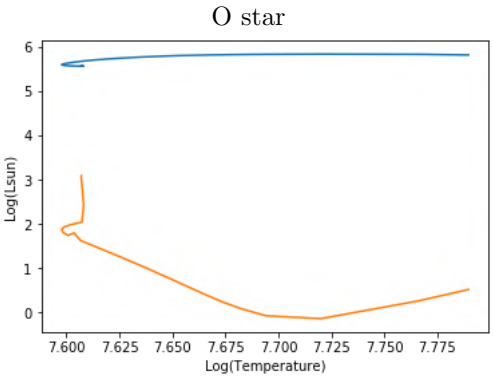


Figure 10: A graph of Luminosity vs Temperature of the pp-chain (orange) and CNO cycle (blue) for a 50 solar mass star

Log(Luminosity) vs. Log(Temperature)

The graph of the 0.7 solar mass star shows how an increase in temperature will increase the luminosity when at lower temperatures. However, as seen in the 50 solar mass star, when the temperature keeps increasing and gets substantially hot, the luminosity will decrease as temperature increases. This is clearly shown with the PP-chain (orange line) which is positive at the lower mass star (experiences lower temperatures) and then switches to a negative slope at the higher temperature and higher mass star.

Discussion

Through the data, it is possible to draw some conclusions about the stellar fusion behind the graphs.

By correlating the graphs of temperature over time and luminosity over time, a vital characteristic can be determined. This is that the luminosity is related to the speed at which the reactions of the CNO cycle or the PP-chain occur. This is because as the temperature increases at a fast rate, the CNO cycle luminosity rises at a fast rate too. At the beginning of this project, it was known that CNO cycle becomes the main source of energy production as the temperature rises. This correlation is important because it will be used for further analytics in the discussion.

The most valuable graph that is vital to understanding the other graphs is the core temperature vs. age. Since the temperature rapidly increases in all 7 masses of the stars, an assumption can be made that temperature fluctuates within the main sequence. Based on background information, an increase in temperature means the radius is smaller, and higher density. From this information, an assumption can be made that there is an increase in density within the main sequence. From this, the conclusion is that the hydrostatic equilibrium does have some fluctuation within the main sequence. The rapid increase seen in these graphs is the small gravitational collapse right before the helium flash.

When looking at the luminosity vs time graphs, an explanation for the need of a higher temperature for the CNO cycle can be reached. Based on previous assumptions, the increase in temperature means that the increase in density is the principle behind the CNO cycle

functionality. Based on this, it can be said that the CNO cycle requires a high density environment (in the core) in order to produce a significant amount of energy.

Lastly there is a theory that can be made about the decrease in the production of hydrogen fusion when at substantially high temperatures. This theory is that since the density keeps getting exponentially high this will somehow affect the reactions so they slow down and produce less energy. A possibility could be that there is some type of degeneration of one of the reactants so they are less abundant.

This fluctuation in the hydrostatic equilibrium explains the luminosity graph. The sudden drop in temperature and luminosity can't be scientifically explained by the stages of the main sequence. When analyzing the possible explanations, the abundance of hydrogen was considered. One possibility was that before the hydrogen fusion starts hydrogen is abundant and therefore high amounts of energy are made. However, once the hydrogen fusion begins, the hydrogen amount will decrease and energy production will suddenly decrease. However this theory fails to recognize that hydrogen fusion does not consume a high enough amount of hydrogen for it to affect the fusion in a fast manner.

Conclusion

This project looked at the general trends that happen with the energy generation process throughout the varied masses. It looked at all 7 types of stars on the main sequence. The general trends includes that the temperature rises as the star ages, the CNO becomes more prominent the higher the mass is, and lastly that the PP-chain energy generation will eventually start to decrease when the temperature gets past a certain temperature.

Some improvements could include gaining more data on qualities. This includes the size of the convection zone, or the actual amount of energy being created by each process instead of the luminosity. Some next steps would be to look at masses that are lower than main sequence stars or larger and investigate the energy used to maintain equilibrium.

[1] Carl Hansen, *"Stellar interiors"*, 1963
[2] O.R. Pols, *"Stellar Structure and evolution"*, 2009
[3] Belén López Martí, *"Stellar evolution Cesar's booklet"*
[4] Astro.ru , *"Stellar models and stellar stability"*, 2013

Apendix A. Additional Graphs

*Blue= CNO cycle

*Orange= PP-chain

M star

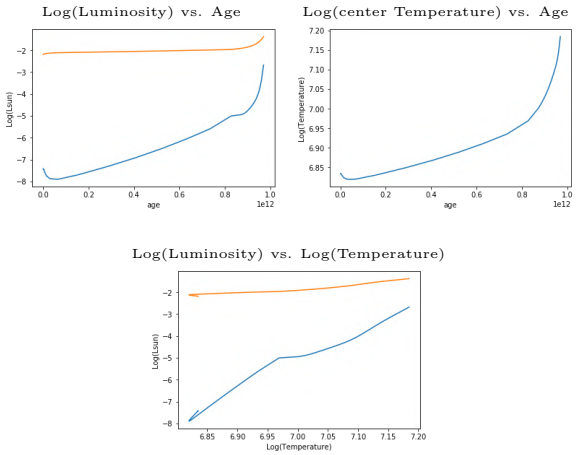


Figure 11: The graphs generated for a 0.2 solar mass star

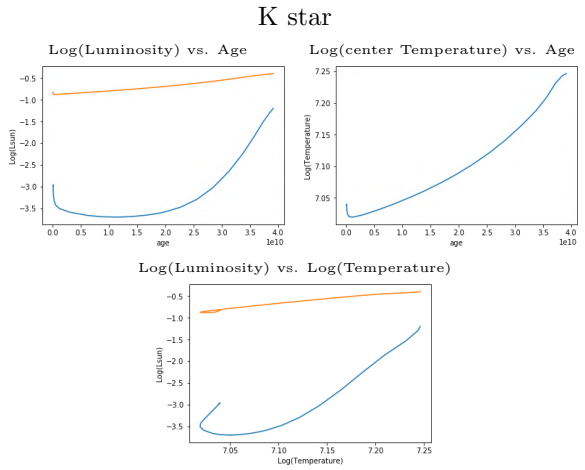


Figure 12: The graphs generated for a 0.7 solar mass star

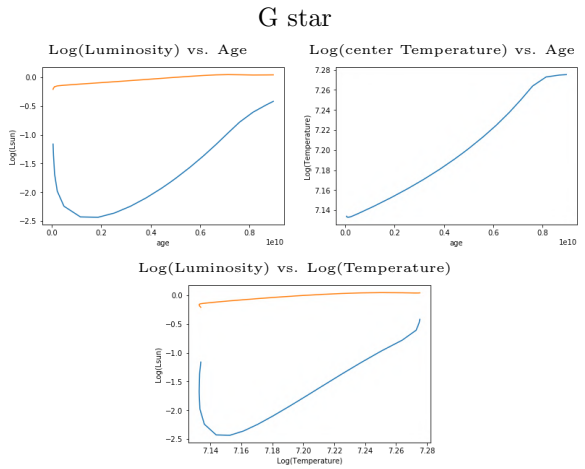


Figure 13: The graphs generated for a 1 solar mass star

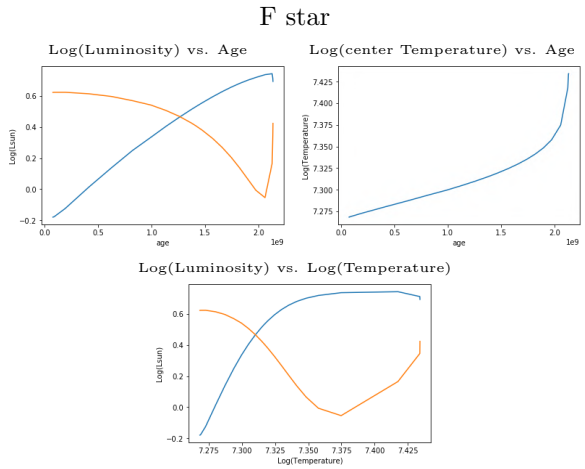


Figure 14: The graphs generated for a 1.5 solar mass star

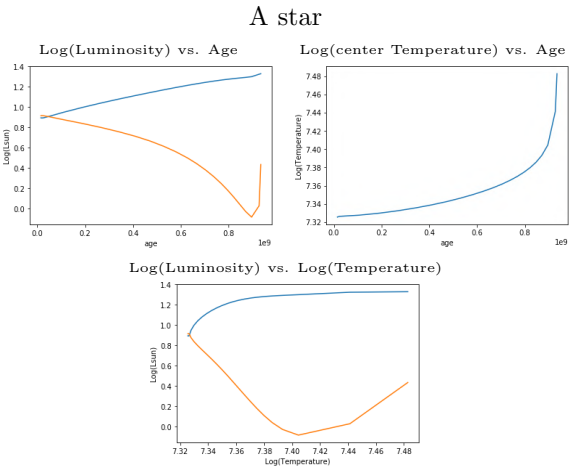


Figure 15: The graphs generated for a 2 solar mass star

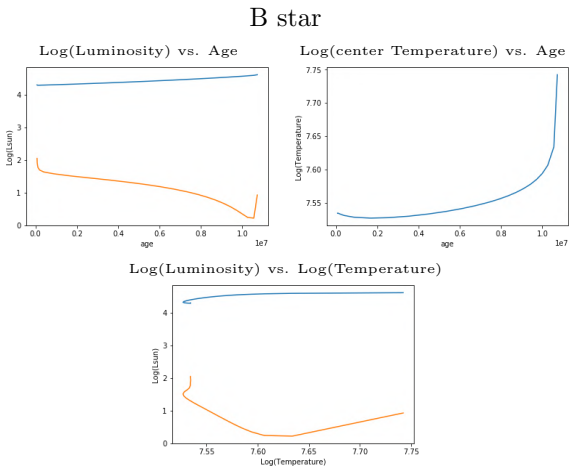


Figure 16: The graphs generated for a 10 solar mass star

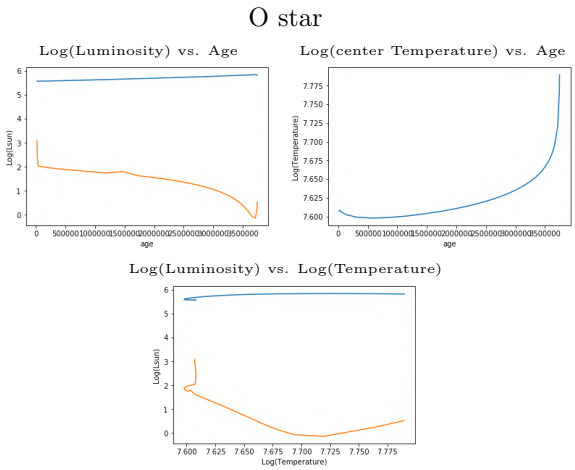


Figure 17: The graphs generated for a 50 solar mass star

ISEC 2019

DIRECTED BY: Carlos, Oz, Radka

PHOTOGRAPHY: Radka

SOUNDTRACK: Oz

LOGISTICS: Carlos

COVER DESIGN: Sılanur Sevgen

HOUSING & FOOD: CTR El Solitario

SPECIAL THANKS TO: Bea, the manager of El Solitario

STARRING

Andrea
Anisia
Anthony
Eva
Katie
Lavinia
Maggie
Niu
Sofía
Tonya
Vivaan
Wytske
Yanni

Wytske, Andrea, Anthony,
Tonya, Niu, Goats
Anisia, Yanni, Sofía

Carlos, Oz, Eva,
Maggie, Vivaan,
Lavinia, Katie, Radka

At ISEC, we learned about new cultures

and we sang songs together

Ground Control to Major Tom... Ground Control to Major Tom...
Take your protein pills and put your helmet on!

Games and Fun revealed both our collaborative and competitive spirits

it may have broken us a little bit

but the laughter made it all worthwhile

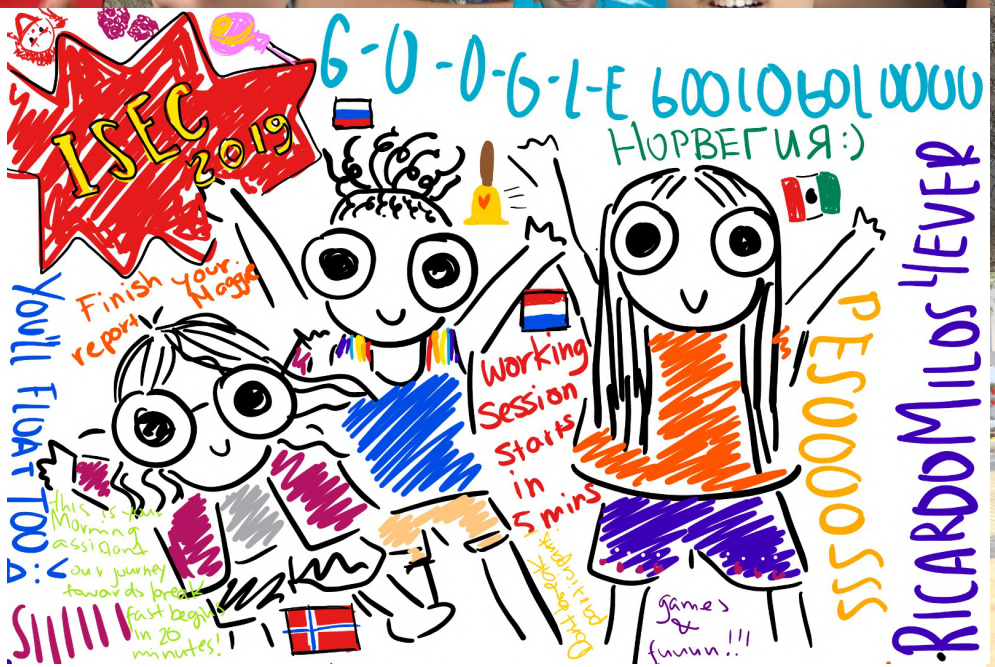
Skopa
Skopa Skopa
Skopa Skopa Skopa
Skopa Skopa Skopa Skopa

and for one day, the mentees even became the mentors

Let's leave the balloons here, Radka, maybe they'll fill them up for us

however,
it wasn't
all fun
and
no work...

we left with new friends, new memories,
and new knowledge



and when the time came, we didn't say 'goodbye', but rather 'take care until we see each other again'