THE APPLICATION OF INFORMATION THEORY TO SCIENTIFIC REVOLUTIONS

DAVIN MARRO*

1. Abstract. For decades, the structure of scientific revolutions has been discussed as a complex dynamical system, driven by observations of natural phenomena that are further susceptible to the hypothetico-deductive process. Thus, the structure and nature of scientific change on a global scale has been controversial. Often, it is thought that the adoption of new scientific methodology, and technological advances work in tandem; as it is equally often that technological advances limit observational ability. Hence, the consensus of global scientific revolutions or paradigm shifts being unpredictable or chaotic in nature is made. However, this paper seeks to apply mathematical and probability statistical concepts and functions (within the context of systems theory), to observe their potential in understanding the nature of these models; as well as look to existing datasets which may be useful in decoding the quantitative attributes of said pattern.

Two datasets will be utilized in an attempt to generate probabilities which explain the potential for geopolitical regions or economies to implement and adopt new methodologies, as well as utilize these potentials to quantify the probability of a global consensus on new methodologies derived from the constituent systems, and adopt them. These datasets being the Human Development Index (HDI), provided by the Human Development Report Office- a sub-sect of the United Nations Development Programme, and the Global Innovation Index (GII) under the World Intellectual Property Organization.

The benefit to better understanding the nature of these systems is to give light to patterns in the adoption and implementation of new methodologies and ideologies in science, but could potentially be used more broadly in sociological contexts, as the scientific advancements undergone are ubiquitous in their influence on general perception. Understanding potential patterns in this mechanism should allow for a better observation of the nature of large-scale change.

2. Theoretical Basis. In order to model large-scale methodological shifts or scientific revolutions, it is first important to highlight important variables and datasets which must be accessed in order to derive insight. To provide a basis for these variables- The Kuhn Cycle, alongside The Hypothetico-Deductive (Scientific) method are referenced. The main distinction between the two processes is their span of influence. Whereas, the hypothetico-deductive method pertains to a more personal-localized approach to scientific methodology, and Kuhn's cycle - which addresses an en-masse, global approach to scientific methodology. Both of these processes are critical to formulating the introduced models, as they serve as the backbone for general scientific methodological systems. In addition, it is noted that there must be a distinction between localized, and globalized developments.

*Thomas More University (dmarro58@thomasmore.edu, https://www.thomasmore.edu).

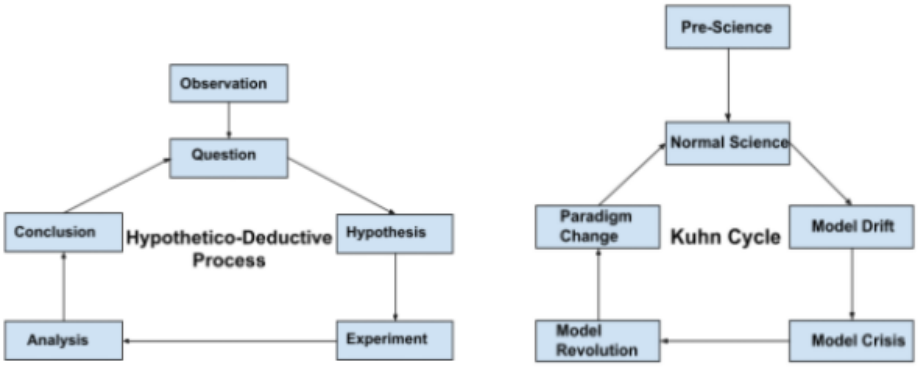

Figure 1. Hypothetico-Deductive Process v. Kuhn Cycle
Adapted From: (Kuhn 2015)

As seen in the above figures, there are similarities between the two cycles- inevitably due to the main difference involving their scale. In one, the relationship between individual observers and a natural phenomenon is expressed- leading one to generate and test hypotheses. And on the other hand, the relationship between a global-observer and a natural phenomenon is expressed- leading to hypothesis generation and experimentation. Both situations contain valuable data- however, the issue with collecting data per individual hypothesis and implementing it into its own model would be computationally and logistically costly.

It seems as individual entities within the system are highlighted - variability is also increased. In other words, comparing one entity to another will likely generate more difference than if millions of entities were compared to other millions. Thus, as sample size increases standard deviation and variance decrease. It is likely that, in the context of finding patterns to methodological shifts and advancements, too much attention is brought to per-hypothetico-deductive processes which are bound to have greater variability between each other, making the structure of the overall process seem too unique and chaotic. However, one could argue that the paradox introduced by a smaller scale for observation makes the system seemingly chaotic. An alternative explanation to a *seemingly chaotic*[4] system is the limitations of technology and data that is sufficient in quantifying credible probabilities for the system. For example, if the phenomenon has yet to be observed or quantified - it may be beyond the limits of instrumental observation- given the current state of technological advances.

Therefore, the model which equates observers to individual people is too reliant on arbitrary variables (contributing to a seemingly chaotic system), whereas one applying the scale of a *globalized* observer is preferred - as it relates to more quantifiable variables while minimizing the variability through maximizing the number of entities per observation. Thus, while the Kuhn cycle and hypothetico-deductive method are similar in process, the larger scale of the former implies reference to institutional progress that can be compared between geopolitical regions. This application breaks a globalized approach to scientific methodology down into two parts being:

1. The ability of individual geopolitical areas to innovate and implement new methods as they become more credible.
2. The ability of the globalized approach to adopt credible methods.

This approach distinguishes two larger-scale systems that work alongside each other within a global society such as one witnessed in the modern day, given the technological innovations. These observations and relationships will serve as the foundation for the proposed model in this paper, which utilizes a combined-approach given input quantities derived from the HDI and GII datasets.

The Human Development Index serves as a scale for geopolitical development with reference to metrics such as life expectancy at birth, expected years of schooling, average years of schooling, and gross national income (GNI) per capita. These metrics assess three primary areas of importance being: life expectancy, quality of life, and knowledge. This assessment includes data for approximately 189 economies over approximately 30 years from 1990-2020.[8]

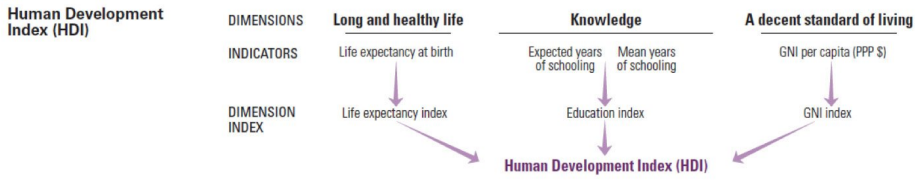

Figure 2. HDI Architecture from:
(https://ourworldindata.org/human-development-index)

The Global Innovation Index is representative of a focused assessment of output. Output which highlights both academic and technical contexts, as well as creative output. The measurements of GII include references to educational, infrastructural, and economic standards. GII defines two sub-indexes being the innovation input, and output systems. These work together to calculate GII values per economic structure or country. GII's assessment contains data for approximately 132 economies over approximately 13 years from 2007-2020.[10]

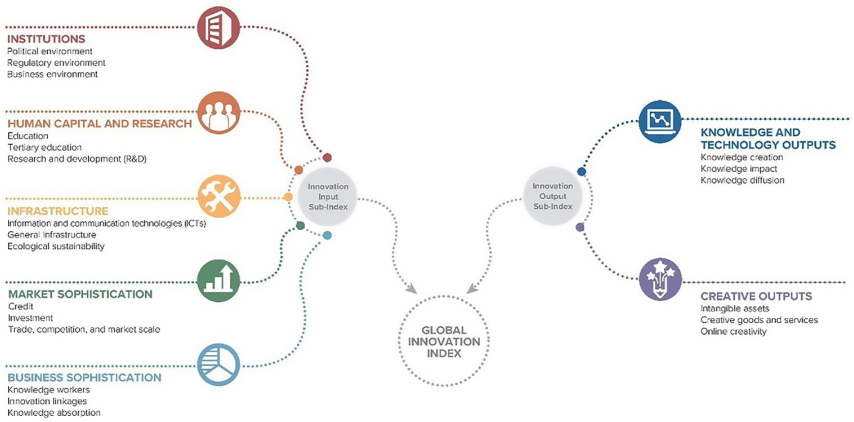

Figure 3. GII Architecture from:
(https://www.globalinnovationindex.org/about-gii)

HDI is useful as the indexes within its broader calculation are relevant when addressing geo-political or economic ability. In essence, the developmental system architecture

relies on the interconnectedness of elements which influence each other while defining the system. In the case of HDI, these variables interconnect in their reliance on one another. For example, with reference to the figure above, lower gross national income will yield a lower standard of living, thus, life expectancy at birth would decrease as well - and furthermore, the expected years of schooling. This reliance between variables provides credibility in quantifying the ability of a large scale structure such as a country to innovate and thrive. However, the variability of this ability can be focused more so with comparing elements under the context of scientific discovery and the development of new methods and technologies. GII behaves in a similar manner, but with more specificity in its main constituents, however, the similarities between the two data sets are found at this architectural level- as they address a majority of the same variables; whereas HDI constituents fit easily in the innovation input sub-index of GII.

With these data-sets, along with the foundational understanding of scientific progression in micro-localized and macro-economical scale there can be a beginning to quantifying and generating mathematical models for methodological shifts in science, as well as broader sociological contexts. In addition, a minimization of arbitrary and seemingly chaotic variables can be achieved with the utilization of data that accesses a broader survey of entities which include many entities per observation in the assessment of geopolitical economies, rather than individuals.

3. Mathematical Basis. The mathematical basis of this paper relies on a combination of functions. To begin, the data for both GII and HDI consist of measurements per year for each available geopolitical economy. For both datasets, the constituents relate to one-another in ways which allow for the values of each to act as indicators for future measurement. In other words, the prior year's data often determines the next year's data. Although, it is important to note that while yearly progression shows trends per economy, the trends per economy are not dependent on which year it is- rather, the measurements are derived from an interaction between the many elements of the systems involved when calculating the indexes. Therefore, in order to observe these progressions on a per-economy scale, a Markov chain could potentially be applied.

A markov chain is defined as a process which changes its state depending on the current state. For example, applying this to the per-economy HDI and GII data, each year has a value which is inevitably referencing prior values. Whereas, the overall process of an economy's development, with reference to their HDI and GII values, differs between economies often due to the initial-state differences.

Hidden markov models (HMM) have been utilized to understand the nature of hidden states that exist in a system which has only some observable variables. They have applications from protein folding pattern recognition, and cancer signaling pathways to speech and gesture recognition, and can be defined as modeling of a Markov process which is not directly observable (X_n).[3][2][1]

$$P(Y_n \epsilon A | X_1 = x_1, ..., X_n = x_n) = P(Y_n \epsilon A | X_n = x_n)$$

Definition 1. Hidden Markov Model

Whereas the Markov definition can be applied to this method-shift mechanism as emission sequences which represent methodological shifts resulting from transition probabilities derived from hidden-state determinants. Thus, when we apply this thinking to the HDI and GII datasets, the language is translated - and inclusive of implications of how HDI and GII are related to method-shifts. This relationship, along with how HDI and GII data relate to one another will be observed.

As mentioned prior, there is inevitable entropy in the systems described. This needs to be addressed with reference to Shannon Entropy, which is closely related to entropy within the Markov model. Perhaps as the entropy of an observable emission becomes complexified by ambiguity, there is an effect on the adoption of a method-shift. This potential must be addressed mathematically. Overall, when observing global-scale scientific methodological advancement- the information gained through the progression of these methodologies is determined by the weight of the transition probability to the nature of the transitions within the progression itself.[7] The entropy in a Markov chain is defined as:

$$H(T_i) = -\sum_{k=1}^{n} P_(i,k) * \log P_(i,k)$$

Definition 2. Shannon Entropy fit to Markov Chain

Where an entropic variable T_i is the transition from state x_i to $x_k$influenced by the probability distribution of the markov chain with state probabilities $(Pi_1, Pi_2 \cdots Pi_n)$. [5][11]

These definitions will be applied to the HDI and GII datasets in order to provide a transition matrix per economy and then enlarge the scale to the collective, global model which can assess the ability of both a localized geopolitical economy and collective global effort to innovate and implement new methods as they become more credible. For the application of the Markov model - three states are identified. State A is a state of increasing HDI/GII, State B is a state of decreasing HDI/GII and State C is a state of constant HDI/GII from the previous.

4. Data. An example of the Markov model is found below, which references HDI data from 1990 to 2020, depicting probabilities for global HDI to increase (A) decrease (B) and stay the same (C) given the prior measure.

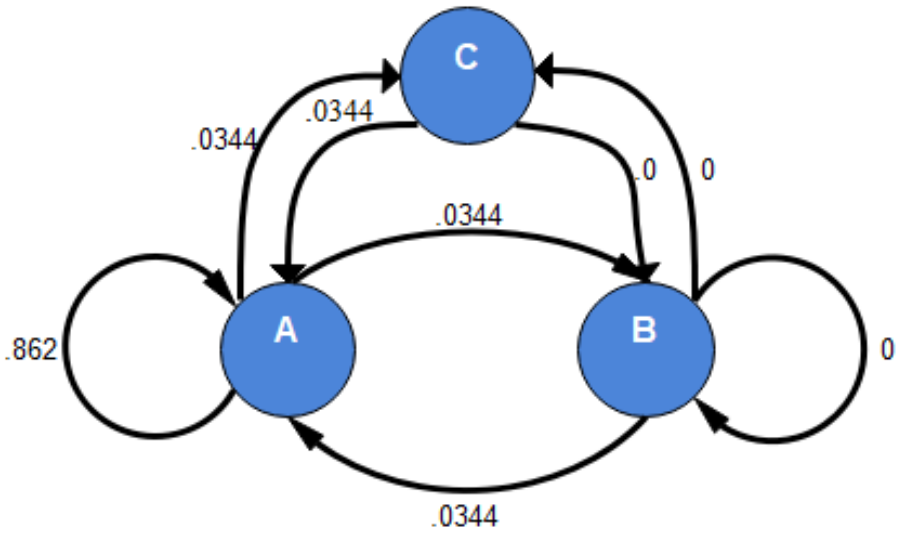

Figure 5. Global HDI Markov Chain

This Markov model is derived of overall transition probabilities from the provided HDI dataset, represented by the following transition matrix:

State Transitions	Overall Probability
B : A	0.0345
A : B	0.0345
B : C	0
A : C	0.0345
C : B	0
C : A	0.0345
B : B	0
A : A	0.862
Sum:	1

Table 1. Global HDI Transition Matrix

Noticeably, when minimizing the scale of the Markov chain, some states may not be present at all in the data. This would suggest that over time, most geopolitical entities trend towards the positive. These diagrams would translate to a transition matrix which lists probability of transitioning from one state (A-C) to another. These two examples of Markov chain diagrams show a case where there are constants present, and a case where a constant is absent. While they both address global probabilities, it should be noted that the likelihood of state complexity increases when smaller data sets are involved. For example, using 10 entities' data will yield more sporadic transition probabilities.

State Transitions	2013:2014	2014:2015	2015:2016	2016:2017	2017:2018	Overall
B : A	0.368	0.03	0.635	0.087	0.524	0.329
A : B	0.104	0.39	0.032	0.603	0.079	0.241
B : C	0.032	0.00	0.048	0.008	0.040	0.025
A : C	0.056	0.02	0.000	0.016	0.008	0.019
C: B	0.056	0.09	0.016	0.048	0.016	0.045
C : A	0.024	0.01	0.000	0.000	0.008	0.008
B : B	0.328	0.45	0.246	0.198	0.286	0.302
A : A	0.032	0.02	0.024	0.040	0.040	0.030
Sum	1.000	1.000	1.000	1.000	1.000	1.000

Table 2. Global GII Transition Matrix

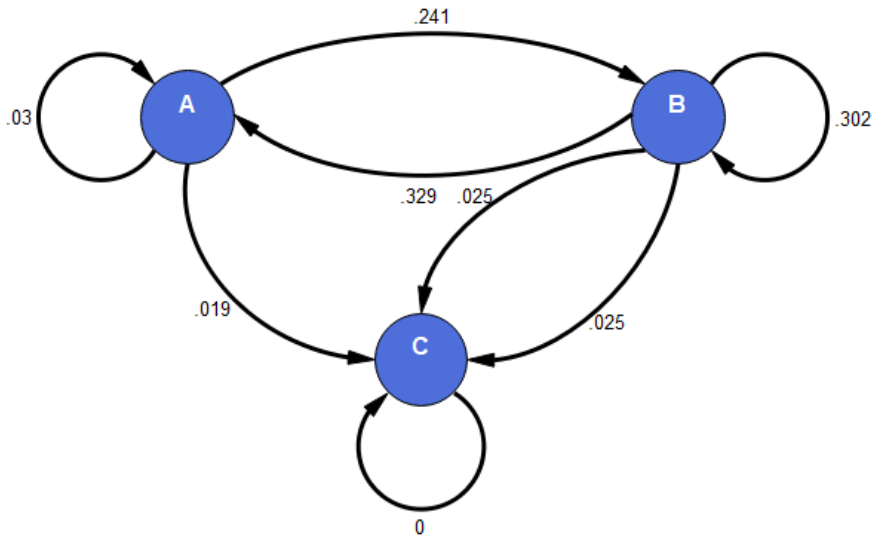

Figure 6. Global GII Markov Chain

In near opposition, the GII data trends towards the negative over time. With less than 1 percent frequency of increase. Further observation of GII and HDI relationship can be observed as follows:

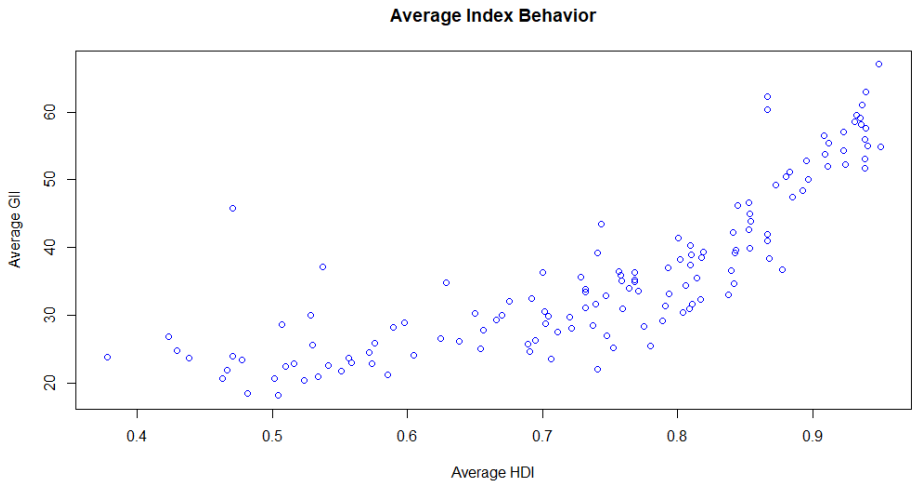

Figure 7. HDI data v. GII data

In combination, HDI and GII trend towards positive as seen in Figure 7. Now, the GII-HDI Markov chain can be generated in order to observe how an increase or decrease in one can effect the other. This relationship can be further witnessed observing the individual Markov models for HDI and GII, and seeing how in tandem the probability of moving from state B (decrease) to state A (increase) is .6, suggesting that negative values have a relatively high probability of transitioning to a state of increase.

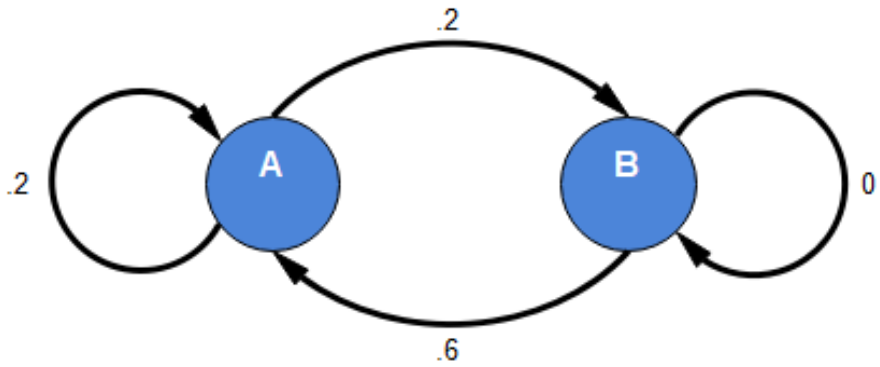

Figure 8. Combined Markov Chain (HDI and GII)

In Figure 7, it can be observed that the likelihood of trending increase is evident, as it is most likely that decreasing GII values can be overwhelmed by increasing HDI numbers.

5. Analysis and Conclusion. With the prior addressed generated Markov chains, probabilities for transitioning from states of decreasing, increasing, and constant HDI and GII values are quantified. In addition, the relationship between the two values is interpreted both graphically and via Markov chain as an overall positive-trending process.

In order to assess this, reiteration of what the data represents is necessary. HDI represents the standing of geopolitical economy in terms of life expectancy, education, and economic stability. Whereas GII represents a geopolitical economy's ability to contribute to innovation through measures of output, efficiency, and their standing. Thus, these interconnected datasets can provide insight as to how the input to innovation can be determined by the dynamic between their underlying elements. The Shannon entropy of these Markov chains can be assessed now that the transition states are identified. As the asymptotic progression of the graphical representation is asymptotic. Therefore the use of these transition probabilities is maintained in their ability to both assess information entropy, and explain how these datasets yield insight to innovation output.[7]

In conclusion, the HDI and GII datasets were observed to identify potential quantification to elements which would be of interest when developing a model for pattern-recognition in global scientific change in terms of output. Thus, the output of these geopolitical constituents feeds back into the system by allowing for technological innovations. Therefore, this model of scientific change could be broadened to represent more sociological observations as they inevitably coalesce.

These Markov chains provide transition states which could be utilized in algorithmic observation and simulation of the progression of the two datasets as representative indicators for scientific method-shifts. This research serves as foundation for a new, applied method of thinking about global scientific change, which improves upon prior works which highlighted the chaotic nature of scientific revolutions whereas the individual cases per hypothetico-deductive method. However, the main proposition this research provides is whether this chaotic nature of scientific change is due to a paradoxical notion of individual weight in the scientific process, and if it can be overwhelmed by a larger-scale economic assessment and combined approach towards answering the question: "What do methodological shifts produce?"

Shortcomings with this research begin with the limitations of the data. While HDI has approximately 30 years of data - accessible GII data is limited to 7 years. Therefore, the assessment of GII and HDI data in tandem is limited to the 7 years where both data have values. However, it is possible that a retroactive simulation of GII data may be created, as similar efforts are seen with the HDI in the Historical Index of Human Development.

In addition, while these Markov chains do in fact quantify the patterns within the two data-sets, the heavy implication is that the elements of these datasets quantify a significant amount of variables which cause the global development and adoption of scientific methodologies. Therefore, it is correlating previous-state input for economic, education, and innovation-input systems to propensity to develop *and* adopt methodologies. It may be possible that these quantification only assess propensity to develop- rather than adopt, or vice-versa.

Therefore, future research should observe the behavior of adoption of new methodologies. At the moment, the data utilized in this research would serve a good foundation for an input/output mechanism for innovation which takes the form of technological innovations and advancements given scientific progress. The key to understanding the pattern of scientific change begins with these elements, however, it ends with its

inclusion and effect on the individuals within its influence. Societies will progress in response to progression, whether the rate is less than or equal to before. In addition, many questions involving the portrayal of these innovations, in sociological contexts, are raised. As this is the area of scientific progress that is most susceptible to chaotic variables.

REFERENCES

[1] Alquraishi, M. (2019). AlphaFold at CASP13. Bioinformatics, 35(22), 4862-4865. doi:10.1093/bioinformatics/btz422

[2] Gagniuc, P. A. (2017). Markov chains from theory to implementation and experimentation. Hoboken: John Wiley amp; Sons.

[3] Haddock, S. H., amp; Dunn, C. W. (2018). Practical computing for biologists. Sunderland, MA: Sinauer.

[4] Kuhn, T. S. (2015). The structure of scientific revolutions. Chicago, IL: The University of Chicago Press.

[5] Lorenz, E. N. (2008). The Essence of Chaos. Seattle: Univ. of Washington Press.

[6] Oppenheim, A. V., Willsky, A. S., amp; Nawab, S. H. (1997). Signals amp; systems (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

[7] Shannon, C. (2013, July 29). A Mathematical Theory of Communication. Retrieved January 18, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x

[8] Roser, M. (2014, July 25). Human development Index (HDI). https://ourworldindata.org/human-development-index

[9] Who will finance innovation? (n.d.). https://www.globalinnovationindex.org/Home

[10] About the gii. (n.d.). https://www.globalinnovationindex.org/about-gii

[11] LeBlanc, P. (n.d.). Information Theory: Entropy, Markov Chains, and Huffman Coding. University Of Notre Dame.