FRED GUTH

THE EMERGENCE OF AN INFORMATION BOTTLENECK THEORY OF DEEP LEARNING

UNIVERSIDADE DE BRASÍLIA

Contents

Welcome

AS FAR AS THE LAWS OF MATHEMATICS REFER TO REALITY, THEY ARE NOT CERTAIN;
AND AS FAR AS THEY ARE CERTAIN, THEY DO NOT REFER TO REALITY.
— ALBERT EINSTEIN

Introduction

In his acceptance speech for the Test-of-Time award in NeurIPS 2017,[1] Ali Rahimi[2] started a controversy by frankly declaring (Rahimi 2018, 12'10"). His concerns on the lack of theoretical understanding of machine learning for critical decision-making are rightful:

> *'We are building systems that govern healthcare and mediate our civic dialogue. We would influence elections. I would like to live in a society whose systems are built on top of verifiable, rigorous, thorough knowledge and not on alchemy.'*

The next day, Yann LeCun[3] responded:

> *'Criticising an entire community (…) for practising "alchemy", simply because our current theoretical tools have not caught up with our practice is dangerous.'*

Both researchers, at least, agree upon one thing: *the practice of machine learning has outpaced its theoretical development*. That is certainly a research opportunity.

A Tale of Babylonians and Greeks

Richard Feynman (Figure 1) used to lecture this story (Feynman 1994): Babylonians were pioneers in mathematics; Yet, the Greeks took the credit. We are used to the Greek way of doing Math: start from the most basic axioms and build up a knowledge system. Babylonians were quite the opposite; they were pragmatic. No knowledge was considered more fundamental than others, and there was no urge to derive proofs in a particular order. Babylonians were concerned with the phenomena, Greeks with the

[1] Conference on Neural Information Processing.

[2] Research Scientist, Google.

Rahimi, Ali. 2018. "Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech." https://youtu.be/x7psGHgatGM.

[3] Deep Learning pioneer and 2018 Turing award winner. https://bit.ly/3CQNwTU

Figure 1: Richard Feynman, Nobel laureate physicist.

Feynman, Richard. 1994. *The Character of Physical Law*. Modern Library.

ordinance. In Feynman's view, science is constructed in the Babylonian way. There is no fundamental truth. Theories try to connect dots from different pieces of knowledge. Only as science advances, one can worry about reformulation, simplification and ordering. Scientists are Babylonians; mathematicians are Greeks.

Mathematics and science are both tools for knowledge acquisition. They are also social constructs that rely on peer-reviewing. They are somewhat different, however.

Science is empiric, based on facts collected from **experience**. When physicists around the world measured events that corroborated Newton's *"Law of Universal Gravitation"*, they did not prove it correct; they just made his theory more and more plausible. Still, only one experiment was needed to show that Einstein's *Relativity Theory* was even more believable. In contrast, we can and do prove things in mathematics.

In mathematics, knowledge is absolute truth, and the way one builds new knowledge with it, its *inference method*, is deduction. **Mathematics is a language**, a formal one, a tool to precisely communicate some kinds of thoughts. As it happens with natural languages, there is beauty in it. The mathematician expands the boundaries of expression in this language.

In science, there are no axioms: a falsifiable hypothesis/theory is proposed, and logical conclusions (predictions) from the theory are empirically tested. Despite inferring hypotheses by induction, there is no influence of psychology in the process. A tested hypothesis is not absolute truth. A hypothesis is never verified, only falsified by experiments (Popper 2004, 31–50). Scientific knowledge is belief justified by experience; there are degrees of plausibility.

Popper, Karl. 2004. *A Lógica Da Pesquisa Científica*. Translated by Leonidas Hegenberg and Octanny Silveira. São Paulo: Cultrix.

Understanding the epistemic contrast between mathematics and science will help us understand the past of AI and avoid some perils in its future.

The importance of theoretical narratives

Science is a narrative of how we understand Nature (Gleiser and Sowinski 2018). In science, we collect facts, but they need interpretation. The logical conclusion from the hypothesis that predicts some behaviour in nature gives a plausible *meaning* to what we observed.

Gleiser, Marcelo, and Damian Sowinski. 2018. "The Map and the Territory." In *The Frontiers Collection*, edited by Shyam Wuppuluri and Francisco Antonio Doria. Springer International Publishing. https://doi.org/10.1007/978-3-319-72478-2.

To illustrate, take the ancient human desire of flying. There have always been stories of men strapping wings to themselves and attempting to fly by jumping from a tower and flapping those wings like birds (see Farrington

2016). While concepts like lift, stability, and control were poorly understood, most human flight attempts ended in severe injury or even death. It did not matter how much evidence, how many hours of seeing different animals flying, those ludicrous brave men experienced; the *meaning* they took from what they saw was wrong, and their predictions incorrect.

Farrington, Karen. 2016. *The Blitzed City: The Destruction of Coventry, 1940*. London: Aurum Press.

Figure 2: 'A way of flying', Francisco Goya, 1815–1820, Amsterdam, Rijksmuseum.

They did not die in vain[4]; Science advances when scientists are wrong. Theories must be falsifiable, and scientists cheer for their failure. When it fails, there is room for new approaches. Only when we understood the observations in animal flight from the aerodynamics perspective, we learned to fly better than any other animal before. Science works by a "natural selection" of ideas, where only the fittest ones survive until a better one is born. Chaitin also points out that an idea has "fertility" to the extent to which it *"illuminates us, inspires us with other ideas, and suggests unsuspected connections and new viewpoints"* (Chaitin 2006, 9).

[4] Those "researchers" deserved, at least, a Darwin Award of Science. The Darwin Award is satirical honours that recognise individuals who have unwillingly contributed to human evolution by selecting themselves out of the gene pool.

Chaitin, Gregory. 2006. *Meta Math! The Quest for Omega*. Vintage Books.

Being a Babylonian enterprise, science has no clear path. One of the exciting facts one can learn by studying its history is that robust discoveries have arisen through the study of phenomena in human-made devices (Pierce, n.d.). For instance, Carnot's first and only scientific work (M. J. Klein 1974) gave birth to thermodynamics: the study of energy, the conversion between its different forms, and the ability of energy to do work; the science that explains how steam engines work. However, steam engines came before Carnot's work and were studied by him. Such human-made devices may present a simplified instance of more complex natural phenomena.

Pierce, John R. n.d. *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications.

Klein, Martin J. 1974. "Carnot's Contribution to Thermodynamics." *Physics Today* 27 (8): 23–28. https://doi.org/10.1063/1.3128802.

Another example is Information Theory. Several insights of Shannon's

theory of communication were generalisations of ideas already present in Telegraphy (Shannon 1948). New theories in artificial intelligence can, therefore, be developed from insights in the study of deep learning phenomena.[5]

Shannon, Claude E. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27 (3): 379–423.

[5] Understanding human intelligence using artificial intelligence is a field of study called Computational Neuroscience.

Bringing science to Computer Science

Despite the name, Computer Science has been more mathematics than science. We, computer scientists, are very comfortable with theorems and proofs, not much with theories.

Nevertheless, AI has essentially become a Babylonian enterprise, a scientific endeavour. Thus, there is no surprise when some computer scientists still see AI with some distrust and even disdain, despite its undeniable usefulness:

- Even among AI researchers, there is a trend of "mathiness" and speculation disguised as explanations in conference papers (Lipton and Steinhardt 2018).
- There are few venues for papers that describe surprising phenomena without trying to come up with an explanation. As if the mere inconsistency of the current theoretical framework was unworthy of publication.

Lipton, Zachary C., and Jacob Steinhardt. 2018. "Troubling Trends in Machine Learning Scholarship." https://arxiv.org/abs/1807.03341.

While physicists rejoice in finding phenomena that contradict current theories, computer scientists get baffled. In Natural Sciences, unexplained phenomena lead to theoretical development. Some believe they bring *winters*, periods of progress stagnation and lack of funding in AI. This seems to be LeCun's opinion.[6]

[6] Due to all possible alternative explanations (lack of computational power, no availability of massive datasets), it seems harsh or simply wrong to blame theorists.

Artificial Intelligence has been through several of the aforementioned "winters". In 1957, Herbert Simon[7] famously predicted that within ten years, a computer would be a chess champion (Russell, Norvig, and Davis 2010, sec. 1.3). It took around 40 years, in any case. Computer scientists lacked understanding of the exponential nature of the problems they were trying to solve: Computational Complexity Theory had yet to be invented.

[7] Herbert Simon (1916–2001) received the Turing Award in 1975, and the Nobel Prize in Economics in 1978.

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

Machine Learning Theory (computational and statistical) tries to avoid a similar trap by analysing and classifying learning problems according to the number of samples required to learn them (besides the number of steps). The matter of concern is that it currently predicts that generalisation requires simpler models in terms of parameters. In total disregard to

the theory, deep learning models have shown spectacular generalisation power with hundreds of millions of parameters (and even more impressive overfitting capacity).

Problem

In the last decade, we have witnessed a myriad of astonishing successes in Deep Learning. Despite those many successes in research and industry applications, we may again be climbing a peak of inflated expectations. If in the past, the false solution was to "add computation power" on problems, today we try to solve them by "piling data" (Figure 3). Such behaviour has triggered a winner-takes-all competition for who collects more data (our data) amidst a handful of large corporations, raising ethical concerns about privacy and concentration of power (O'Neil 2016).

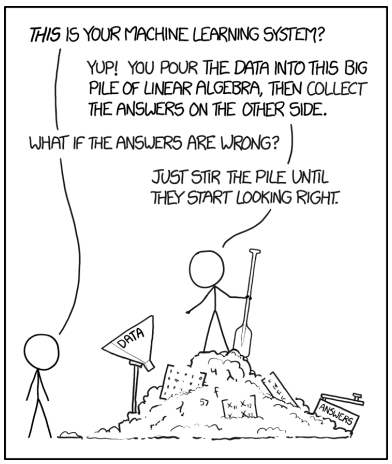


Figure 3: Source: https://xkcd.com/1838/. Reprinted with permission.

O'Neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. USA: Crown Publishing Group.

Nevertheless, we know that learning from way fewer samples is possible: humans show a much better generalisation ability than our current state-of-the-art artificial intelligence. To achieve such needed generalisation power, we may need to understand better how learning happens in deep learning. Rethinking generalisation might reshape the foundations of machine learning theory (Zhang et al. 2016).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. "Understanding Deep Learning Requires Rethinking Generalization." https://arxiv.org/abs/1611.03530.

Possible new explanation in the horizon

In 2015, Tishby and Zaslavsky (2015) proposed a theory of deep learning (Tishby and Zaslavsky 2015) based on the information-theoretical concept of the bottleneck principle, of which Tishby is one of the authors. Later, in 2017, Shwartz-Ziv and Tishby (2017) followed up on the IBT with the paper , which was presented in a well-attended workshop[8], with appealing visuals that clearly showed a *"phase transition"* happening during training. The video posted on Youtube (Tishby 2017) became a "sensation"[9], and received a wealth of publicity when well-known researchers like Geoffrey Hinton[10], Samy Bengio (Apple) and Alex Alemi (Google Research) have expressed interest in Tishby's ideas (Wolchover 2017). they are called formal languages.

Tishby, Naftali, and Noga Zaslavsky. 2015. "Deep Learning and the Information Bottleneck Principle." In *2015 IEEE Information Theory Workshop (ITW)*, 1–5. IEEE.

> *'I believe that the information bottleneck idea could be very important in future deep neural network research.' — Alex Alemi*

Shwartz-Ziv, Ravid, and Naftali Tishby. 2017. "Opening the Black Box of Deep Neural Networks via Information." https://arxiv.org/abs/1703.00810.

Andrew Saxe (Harvard University) rebutted Shwartz-Ziv and Tishby (2017) claims in and was followed by other critics. According to Saxe, it was impossible to reproduce (Shwartz-Ziv and Tishby 2017)'s experiments with different parameters.

[8] Deep Learning: Theory, Algorithms, and Applications. Berlin, June 2017 http://doc.ml.tu-berlin.de/dlworkshop2017

Tishby, Naftali. 2017. "Information Theory of Deep Learning." https://youtu.be/bLqJHjXihK8. https://youtu.be/bLqJHjXihK8.

[9] By the time of this writing, this video as more than 84,000 views, which is remarkable for an hour-long workshop presentation in an academic niche. https://youtu.be/bLqJHjXihK8

[10] Another Deep Learning Pioneer and Turing award winner (2018).

Wolchover, Natalie. 2017. "New

Has the initial enthusiasm on the IBT been unfounded? Have we let us "fool ourselves" by beautiful charts and a good story?

Problem statement

The practice of modern machine learning has outpaced its theoretical development. In particular, deep learning models present generalisation capabilities unpredicted by the current machine learning theory. There is yet no established new general theory of learning which handles this problem.

IBT was proposed as a possible new theory with the **potential** of filling the theory-practice gap. Unfortunately, to the extent of our knowledge, **there is still no comprehensive digest of IBT nor an analysis of how it relates to current MLT**.

Objective

This dissertation aims to investigate *to what extent* can the emergent Information Bottleneck Theory help us better understand Deep Learning and its phenomena, especially generalisation, presenting its strengths, weaknesses and research opportunities.

Research Questions

1. What are the fundamentals of IBT? How do they differ from the ones from MLT?
2. What is the relationship between IBT and current MLT? How different or similar they are?
3. Is IBT capable of explaining the phenomena MLT already explains?
4. Does IBT invalidate results in MLT?
5. Is IBT capable of explaining phenomena still not well understood by MLT?
6. What are Information Bottleneck Theory's (IBT) strengths?
7. What are Information Bottleneck Theory's (IBT) weaknesses?
8. What has been already developed in IBT?
9. What are Information Bottleneck Theory's (IBT) research opportunities?

Methodology

1. Given that IBT is yet not a well-established learning theory, there were two difficulties that the research had to address:

1. There is a growing interest in the subject, and new papers are published every day. It was essential to select literature and restrain the analysis.
2. Early on, the marks of an emergent theory in its infancy manifested in the form of missing assumptions, inconsistent notation, borrowed jargon, and seeming missing steps. Foremost, it was unclear what was missing from the theory and what was missing in our understanding.

An initial literature review on IBT was conducted to define the scope.[11] We then chose to narrow the research to **theoretical perspective on generalisation**, where we considered that it could bring fundamental advances. We made the deliberate choice of going deeper in a limited area of IBT and not broad, leaving out a deeper experimental and application analysis, all the work on ITL[12] (Principe 2010) and statistical-mechanics-based analysis of SGD (P. Chaudhari and Soatto 2018; Pratik Chaudhari et al. 2019). From this set of constraints, we chose a list of pieces of IBT literature to go deeper.

[11] Not even the term IBT is universally adopted.

[12] ITL makes the opposite path we are taking, bringing concepts of machine learning to information theory problems.

Principe, Jose C. 2010. *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Science & Business Media.

Chaudhari, P., and S. Soatto. 2018. "Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks." In *2018 Information Theory and Applications Workshop (ITA)*, 1–10. https://doi.org/10.1109/ITA.2018.8503224.

Chaudhari, Pratik, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. 2019. "Entropy-Sgd: Biasing Gradient Descent into Wide Valleys." *Journal of Statistical Mechanics: Theory and Experiment* 2019 (12).

2. In order to answer , we discuss the epistemology of AI to choose fundamental axioms (definition of intelligence and the definition of knowledge) with which we deduced from the ground up MLT, IT and IBT, revealing hidden assumptions, pointing out similarities and differences. By doing that, we built a "genealogy" of these research fields. This comparative study was essential for identifying missing gaps and research opportunities.
3. In order to answer , we first dissected the selected literature ([[ch:literature]][3]) and organised scattered topics in a comprehensive sequence of subjects.
4. In the process of the literature digest, we identified results, strengths, weaknesses and research opportunities.

Contributions

In the research conducted, we produced three main results that, to the extent of our knowledge, are original:

1. The dissertation itself is the main expected result: a comprehensive digest of the IBT literature and a snapshot analysis of the field in its current form, focusing on its theoretical implications for generalisation.
2. We propose an Information-Theoretical learning problem different from MDL proposed by (Hinton and Van Camp 1993) for which we

Hinton, Geoffrey E, and Drew Van Camp. 1993. "Keeping the Neural Networks Simple by Minimizing the Description Length of the Weights." In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, 5–13.

derived bounds using Shannon's . These results, however, are only indicative as they lack peer review to be validated.

3. We present a critique on Achille (2019)'s explanation (Achille 2019; Achille and Soatto 2018) for the role of layers in Deep Representation in the IBT perspective (**?@sec-achille_proof_critique**), pointing out a weakness in the argument that, as far as we know, has not yet been presented. We then propose a counter-intuitive *hypothesis* that layers reduce the model's "effective" hypothesis space. This *hypothesis* is not formally proven in the present work, but we try to give the intuition behind it (**?@sec-proposed_hypothesis**). This result has not yet been validated as well.

Achille, Alessandro. 2019. "Emergent Properties of Deep Neural Networks." PhD thesis, UCLA. https://escholarship.org/uc/item/8gb8x6w9.

Achille, Alessandro. 2019. "Emergent Properties of Deep Neural Networks." PhD thesis, UCLA. https://escholarship.org/uc/item/8gb8x6w9.

Achille, Alessandro, and Stefano Soatto. 2018. "Emergence of Invariance and Disentangling in Deep Representations." *J. Mach. Learn. Res.* 19 (1): 1947–80.

Dissertation preview and outline

The dissertation is divided into two main parts (`Background` and `The emergence of a theory`), with a break in the middle (`Intermezzo`).

1. Background
    - Chapter 2 — Artificial Intelligence: The chapter defines what artificial intelligence is, presents the epistemological differences of intelligent agents in history, and discusses their consequences to machine learning theory.
    - Chapter 3 — Probability Theory: The chapter derives propositional calculus and probability theory from a list of desired characteristics for epistemic agents. It also presents basic Probability Theory concepts.
    - Chapter 4 — Machine Learning Theory: The chapter presents the theoretical framework of Machine Learning, the PAC model, theoretical guarantees for generalisation, and expose its weaknesses concerning Deep Learning phenomena.
    - Chapter 5 — Information Theory: The chapter derives Shannon Information from Probability Theory, explicates some implicit assumptions, and explains basic Information Theory concepts.
2. Intermezzo
    - Chapter 6 — Information-Theoretical Epistemology: This chapter closes the background part and opens the IBT part of the dissertation. It shows the connection of IT and MLT in the learning problem, proves that Shannon theorems can be used to prove PAC bounds and present the MDL Principle, an earlier example of this kind of connection.

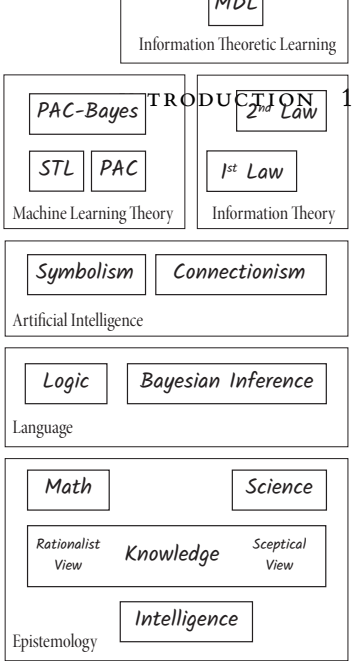


Figure 4: IBT "genealogy" tree.

3. The emergence of a theory

   - Chapter 7 — IB Principle: Explains the IB method and its tools: KL as a natural distortion (loss) measure, the IB Lagrangian and the Information Plane.
   - Chapter 8 — IB and Representation Learning: Presents the learning problem in the IBT perspective (not specific to DL). It shows how some usual choices of the practice of DL emerge naturally from a list of desired properties of representations. It also shows that the information in the weights bounds the information in the activations.
   - Chapter 9 — IB and Deep Learning: This chapter presents the IBT perspective specific to Deep Learning. It presents IBT analysis of Deep Learning training, some examples of applications of IBT to improve or create algorithms; and the IBT learning theory of Deep Learning. We also explain Deep Learning phenomena in the IBT perspective.
   - Chapter 10 — Conclusion: In this chapter, we present a summary of the findings, answer the research questions, and present suggestions for future work.

We found out that IBT does not invalidate MLT; it just interprets complexity not as a function of the data (number of parameters) but as a function of the information contained in the data. With this interpretation, there is no paradox in improving generalisation by adding layers.

Furthermore, they both share more or less the same "genealogy" of assumptions. IBT can be seen as particular case of MLT. Nevertheless, IBT allows us to better understand the training process and provide a different narrative that helps us comprehend Deep Learning phenomena in a more general way.

Background

I VISUALISE A TIME WHEN WE WILL BE TO ROBOTS WHAT DOGS ARE TO HUMANS

...AND I AM ROOTING FOR THE MACHINES.

— CLAUDE SHANNON

Artificial Intelligence

This chapter defines artificial intelligence, presents the epistemological differences of intelligent agents in history, and discusses their consequences to machine learning theory.

What is Artificial Intelligence?

Definition 0.1. **AI** is the branch of Computer Science that studies general principles of intelligent agents and how to construct them (Russell, Norvig, and Davis 2010).

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

This definition uses the terms *intelligence* and *intelligent agents*, so let us start from them.

What is intelligence?

Despite a long history of research, there is still no consensual definition of intelligence.[13] Whatever it is, though, humans are particularly proud of it. We even call our species *homo sapiens*, as intelligence was an intrinsic human characteristic.

In this dissertation:

[13] For a list with 70 definitions of intelligence, see Legg and Hutter (2007).

Legg, Shane, and Marcus Hutter. 2007. "A Collection of Definitions of Intelligence." https://arxiv.org/abs/0706.3639.

Definition 0.2. **Intelligence** is the ability to predict a course of action to achieve success in specific goals.

Intelligent Agents

Under our generous definition, intelligence is not limited to humans. It applies to any agent[14]: animal or machine. For example, a bacteria can perceive its environment through chemical signals, process them, and then produce chemicals to signal other bacteria. An air-conditioning can observe temperature changes, know its state, and adapt its functioning, turning off if it is cold or on if it is hot — *intelligence exempts understanding*. The air-conditioning does not comprehend what it is doing. The same way a calculator does not know arithmetics.

[14] An agent is anything that perceives its environment and acts on it.

A strange inversion of reasoning

This competence without comprehension is what the philosopher Daniel Dennett calls *Turing's strange inversion of reasoning*[15]. The idea of a *strange inversion* comes from one of Darwin's 19th-century critics Dennett (2009):

> *'In the theory with which we have to deal, Absolute Ignorance is the artificer; so that we may enunciate as the fundamental principle of the whole system, that,* ***in order to make a perfect and beautiful machine, it is not requisite to know how to make it****. This proposition will be found, on careful examination, to express, in condensed form, the essential purport of the [Evolution] Theory, and to express in a few words all Mr Darwin's meaning; who, by* ***a strange inversion of reasoning****, seems to think Absolute Ignorance fully qualified to take the place of Absolute Wisdom in all of the achievements of creative skill.'*
>
> — *Robert MacKenzie*

[15] In his work, Turing discusses if computers can "think", meaning to examine if they can perform indistinguishably from the way thinkers do.

MacKenzie, Robert Beverley. 1868. *The Darwinian Theory of the Transmutation of Species Examined*. J. Nisbet.

Counterintuitively to MacKenzie (1868) and many others to this date, intelligence can emerge from absolute ignorance. Turing's strange inversion of reasoning comes from the realisation that his automata can perform calculations by symbol manipulation, proving that it is possible to build agents that behave intelligently, even if they are entirely ignorant of the meaning of what they are doing (Turing 2007).

MacKenzie, Robert Beverley. 1868. *The Darwinian Theory of the Transmutation of Species Examined*. J. Nisbet.

Turing, Alan M. 2007. "Computing Machinery and Intelligence." In *Parsing the Turing Test*, 23–65. Springer Netherlands. https://doi.org/10.1007/978-1-4020-6710-5_3.

Dreaming of robots

From mythology to Logic

The idea of creating an intelligent agent is perhaps as old as humans. There are accounts of artificial intelligence in almost any ancient mythology: Greek, Etruscan, Egyptian, Hindu, Chinese (Mayor 2018). For example, in Greek mythology, the story of the bronze automaton of Talos built

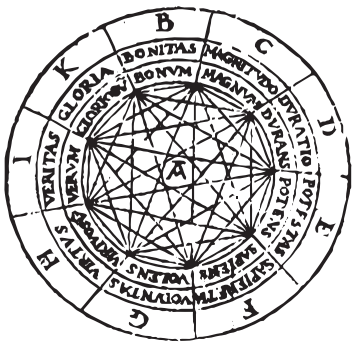
Mayor, Adrienne. 2018. *Gods and Robots: Myths, Machines, and Ancient Dreams of Technology*. Princeton University Press.

by Hephaestus, the god of invention and blacksmithing, first mentioned around 700 BC.

This interest may explain why, since ancient times, philosophers have looked for *mechanical* methods of reasoning. Chinese, Indian and Greek philosophers all developed formal deduction in the first millennium BC.In particular, Aristotelian syllogism, *laws of thought*, provided patterns for argument structures to yield irrefutable conclusions, given correct premises. These ancient developments were the beginning of the field we now call *Logic*.

Rationalism: The Cartesian view of Nature

Example of Lull's Ars Magna's paper discs.



Gardner, Martin. 1959. *Logic Machines and Diagrams*. McGraw-Hill Book Company.

In the 13th century, the Catalan philosopher Ramon Lull wanted to produce all statements the human mind can think. For this task, he developed *logic paper machines*, discs of paper filled with esoteric coloured diagrams that connected symbols representing statements. Unfortunately, according to Gardner (1959), in a modern reassessment of his work, *"it is impossible, perhaps, to avoid a strong sense of anticlimax"*. With megalomaniac self-esteem that suggests psychosis, his delusional sense of importance is more characteristic of cult founders. On the bright side, his ideas and books exerted some magic appeal that helped them be rapidly disseminated through all Europe.

Lull's work greatly influenced Leibniz and Descartes, who, in the 17thcentury, believed that all rational thought could be mechanised. This belief was the basis of **rationalism**, the epistemic view of the *Enlightenment* that regarded reason as the sole source of knowledge. In other words, they believed that reality has a logical structure and that certain truths are *self-evident*, and all truths can be derived from them.

There was considerable interest in developing artificial languages during this period. Nowadays, they are called formal languages.

> *'If controversies were to arise, there would be no more need for disputation between two philosophers than between two accountants. For it would suffice to take their pencils in their hands, to sit down to their slates, and to say to each other:* Let us calculate.*'*
>
> *— Gottfried Leibniz*

The rationalist view of the world has had an enduring impact on society

until today. In the 19thcentury, George Boole and others developed a precise notation for statements about all kinds of objects in Nature and their relations. Before them, Logic was philosophical rather than mathematical. The name of Boole's masterpiece, *"The Laws of Thought"*, is an excellent indicator of his Cartesian worldview.

At the beginning of the 20th century, some of the most famous mathematicians, David Hilbert, Bertrand Russel, Alfred Whitehead, were still interested in formalism: they wanted mathematics to be formulated on a solid and complete logical foundation. In particular, Hilbert's *Entscheidungs Problem* (decision problem) asked if there were limits to mechanical Logic proofs (Chaitin 2006).

Chaitin, Gregory. 2006. *Meta Math! The Quest for Omega*. Vintage Books.

Kurt Gödel's incompleteness theorem (1931) proved that any language expressive enough to describe arithmetics of the natural numbers is either incomplete or inconsistent. This theorem imposes a limit on logic systems. There will always be truths that will not be provable from within such languages: there are "true" statements that are undecidable.

Alan Turing brought a new perspective to the *Entscheidungs Problem*: a function on natural numbers that an algorithm in a formal language cannot represent cannot be computable (Chaitin 2006). Gödel's limit appears in this context as functions that are not computable, no algorithm can decide whether another algorithm will stop or not (the halting problem). To prove that, Turing developed a whole new general theory of computation: what is computable and how to compute it, laying out a blueprint to build computers, and making possible Artificial Intelligence research as we know it. An area in which Turing himself was very much invested.

Chaitin, Gregory. 2006. *Meta Math! The Quest for Omega*. Vintage Books.

Empiricism: The sceptical view of Nature

David Hume, Scottish Enlightenment philosopher, historian, economist, librarian and essayist.

The response to **rationalism** was **empiricism**, the epistemological view that knowledge comes from sensory experience, our perceptions of the world. Locke explains this with the peripatetic axiom[16]: *"there is nothing in the intellect that was not previously in the senses"* (Uzgalis 2020). Bacon, Locke and Hume were great exponents of this movement, which established the grounds of the scientific method.

[16] This citation is the principle from the Peripatetic school of Greek philosophy and is found in Thomas Aquinas' work cited by Locke.

Uzgalis, William. 2020. "John Locke." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2020. https://plato.stanford.edu/archives/spr2020/entries/locke/; Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2020/entries/locke/.

David Hume, in particular, presented in the 18th century a radical empiricist view: reason only does not lead to knowledge. In (Hume 2009), Hume distinguishes *relations of ideas*, propositions that derive from deduction and

Hume, David. 2009. *Tratado Da Natureza Humana*. Editora UNESP.

matters of facts, which rely on the connection of cause and effect through experience (induction). Hume's critiques, known as *the Problem of Induction*, added a new slant on the debate of the emerging scientific method.

From Hume's own words:

> *'The bread, which I formerly eat, nourished me; that is, a body of such sensible qualities was, at that time, endued with such secret powers: but does it follow, that other bread must also nourish me at another time, and that like sensible qualities must always be attended with like secret powers? The consequence seems nowise necessary.'*
>
> *— David Hume*

There is no logic to deduce that the future will resemble the past. Still, we expect uniformity in Nature. As we see more examples of something happening, it is *wise* to expect that it will happen in the future just as it did in the past. There is, however, no *rationality*[17] in this expectation.

[17] In the philosophical sense.

Hume explains that we see conjunction repeatedly, "bread" and "nourish", and we expect *uniformity in Nature*; we hope that "nourish" will always follow "eating bread"; When we fulfil this expectancy, we misinterpret it as causation. In other words, we *project* causation into phenomena. Hume explained that this connection does not exist in Nature. We do not "see causation"; we create it.

This projection is *Hume's strange inversion of reasoning* (Huebner 2017): We do not like sugar because it is sweet; sweetness exists because we like (or need) it. There is no sweetness in honey. We wire our brain so that glucose triggers a labelled desire we call sweetness. As we will see later, sweetness is *information*. This insight shows the pattern matching nature of humans. Musicians have relied on this for centuries. Music is a sequence of sounds in which we expect a pattern. The expectancy is the tension we feel while the chords progress. When the progression finally *resolves*, forming a pattern, we release the tension. We feel pattern matching in our core. It is very human, it can be beneficial and wise, but it is, *stricto sensu*, *irrational*.

Huebner, Bryce. 2017. *The Philosophy of Daniel Dennett*. Oxford University Press.

The epistemology of the sceptical view of Nature is science: to weigh one's beliefs to the evidence. Knowledge is not absolute truth but justified belief. It is a Babylonian epistemology.

In rationalism, Logic connects knowledge and good actions. In empiricism, the connection between knowledge and justifiable actions is determined by

probability. More specifically, Bayes' theorem. As Jaynes puts it, probability theory is the "Logic of Science" . [18]

[18] The Bayes' theorem is attributed to the Reverend Thomas Bayes after the posthumous publication of his work. By the publication time, it was an already known theorem, derived by Laplace.

The birth of AI as a research field

In 1943, McCulloch and Pitts, a neurophysiologist and a logician, demonstrated that neuron-like electronic units could be wired together, act and interact by physiologically plausible principles and perform complex logical calculations (Russell, Norvig, and Davis 2010). Moreover, they showed that any computable function could be computed by some network of connected neurons (McCulloch and Pitts 1943). Their work marks the birth of ANNs, even before the field of AI had this name. It was also the birth of **Connectionism**, using artificial neural networks, loosely inspired by biology, to explain mental phenomena and imitate intelligence.

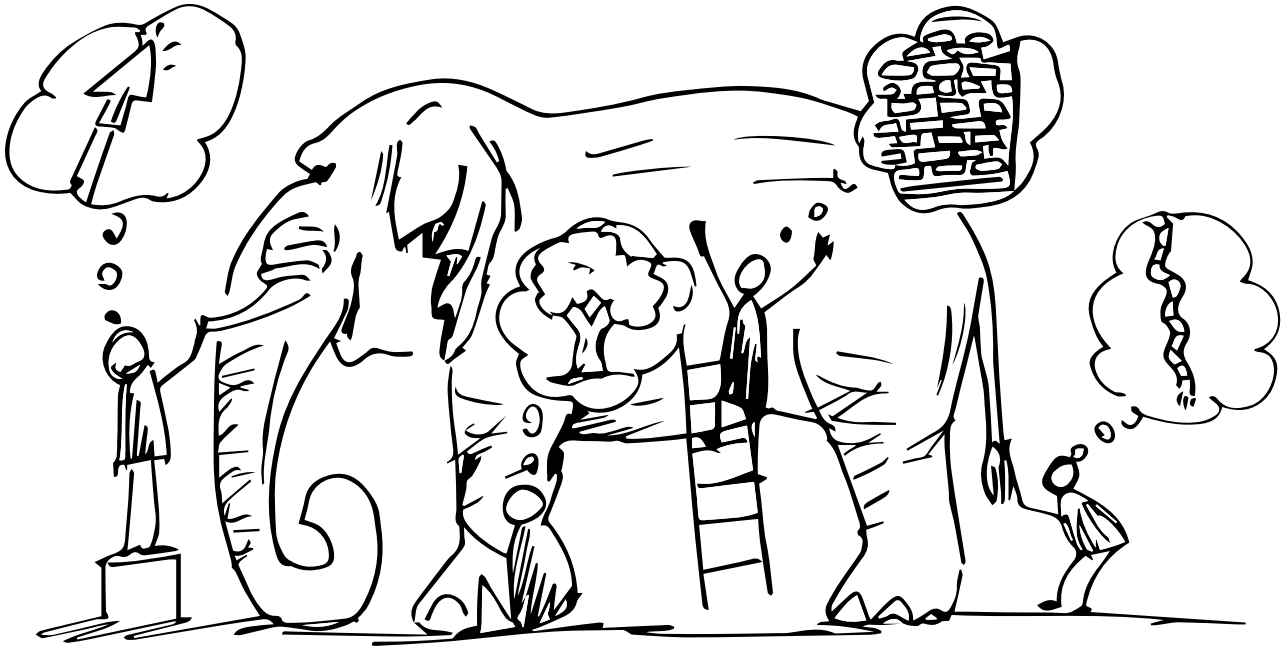
Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5 (4): 115–33.

Their work inspired John von Neumann's demonstration of how to create a universal Turing machine out of electronic components, which lead to the advent of computers and programming languages. Ironically, these advents hastened the ascent of the formal logicist approach called **Symbolism**, disregarding Connectionism.

In 1956, John McCarthy, Claude Shannon (who invented Information Theory, Figure ??), Marvin Minsky and Nathaniel Rochester organised a 2-month summer workshop in Dartmouth College to bring researchers of different fields concerned with *"thinking machines"* (cybernetics, information theory, automata theory). The workshop attendees became a community of researchers and chose the term *"artificial intelligence"* for the field.

Claude Shannon, father of "information theory".



The Blind Men and the Elephant.

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind

—John Godfrey Saxe,
The Blind Men and the Elephant

Building Intelligent Agents

Anatomy of intelligent agents

Like the blind men in the parable, an intelligent agent shall model her understanding of Nature from limited sensory data.

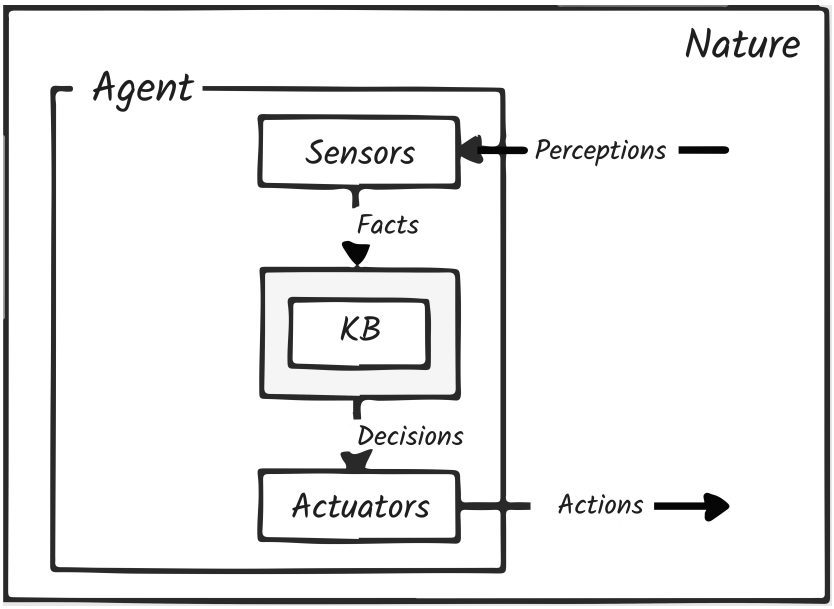


Figure 5: Anatomy of an Intelligent Agent.

The expected result of this conversation is a change in the agent's KB, therefore in her model and, more importantly, her future decisions. The model is an abstraction of how the agent "thinks" the world is (her "mental picture" of the environment). Therefore, it should be consistent with it: if something is true in Nature, it is equally valid, *mutatis mutandis*, in the model. A Model should also be as simple as possible so that the agent can make decisions that maximise a chosen performance measure, but not simpler. As the agent knows more about Nature, less it gets surprised by it.

This rudimentary anatomy is flexible enough to entail different epistemic views, like the rationalist (mathematical) and the empiricist (scientific); different approaches to how to implement the knowledge base (it can be learned, therefore updatable, or it can be set in stone from an expert prior knowledge); and also from how to implement it (a robot or software).

Noteworthy, though, is that the model that transforms input data into decisions should be the target of our focus.

Symbolism

Symbolism is the pinnacle of rationalism. In the words of Thomas Hobbes, one of the forerunners of rationalism, *"thinking is the manipulation of symbols and reasoning is computation".* Symbolism is the approach to building intelligent agents that does just that. It attempts to represent knowledge with a formal language and explicitly connects the knowledge with actions. It is *competence from comprehension*. In other words, it is *programmed*.

Even though McCulloch and Pitts work on artificial neural networks predates Von Neumann's computers, Symbolism dominated AI until the 1980s. It was so ubiquitous that symbolic AI is even called "good old fashioned AI" (Russell, Norvig, and Davis 2010).

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

The symbolic approach can be traced back to Nichomachean Ethics (Aristotle 2000):

> *'We deliberate not about ends but means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade, nor a statesman whether he shall produce law and order, nor does anyone else deliberate about his end. They assume the end and consider how and by what means it is to be attained; and if it seems to be produced by several means, they consider by which it is most easily and best produced, while if it is achieved by one only they consider how it will be achieved by this and by what means this will be achieved, till they come to the first cause, which in the order of discovery is last.'*
>
> *— Aristotle*

Aristotle. 2000. *Aristotle: Nicomachean Ethics*. Cambridge Texts in the History of Philosophy. Cambridge University Press. https://doi.org/10.1017/CBO9780511802058.

This perspective is so entrenched that Russell, Norvig, and Davis (2010, 7) still says: *"(...) Only by understanding how actions can be justified can we understand how to build an agent whose actions are justifiable"*; even though, in the same book, they cover machine learning (which we will address later in this chapter) without noticing it is proof that there are other ways to build intelligent agents. Moreover, it is also a negation of competence without comprehension. It seems that even for AI researchers, the strange inversion of reasoning is uncomfortable ([[ch:introduction]][3]).

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

All humans, even those in prisons and under mental health care, think their actions are justifiable. Is that not an indication that we rationalise our actions *ex post facto*? We humans tend to think our rational assessments lead to actions, but it is also likely possible that we act and then rationalise afterwards to justify what we have done, fullheartedly believing that the rationalisation came first.

Claude Shannon's Theseus

After writing what is probably the most important master's dissertation of the 20th century and "inventing" IT, what made possible the Information Age we live in today, Claude Shannon enjoyed the freedom to pursue any interest to which his curious mind led him (Soni and Goodman 2017). In the 1950s, his interest shifted to building artificial intelligence. He was not a typical academic, in any case. A lifelong tinkerer, he liked to "think" with his hand as much as with his mind. Besides developing an algorithm to play chess (when he even did not have a computer to run it), one of his most outstanding achievements in AI was Theseus, a robotic maze-solving mouse.[19]

Soni, Jimmy, and Rob Goodman. 2017. *A Mind at Play: How Claude Shannon Invented the Information Age*. Simon; Schuster.

[19] Many AI students will recognise in Theseus the inspiration to Russel and Norvig's Wumpus World.

To be more accurate, Theseus was just a bar magnet covered with a sculpted wooden mouse with copper whiskers; the maze was the "brain" that solved itself (D. Klein 2018).

Klein, Daniel. 2018. "Mighty Mouse." Technology Review. https://www.technologyreview.com/s/612529/mighty-mouse/.

> *'Under the maze, an electromagnet mounted on a motor--powered carriage can move north, south, east, and west; as it moves, so does Theseus. Each time its copper whiskers touch one of the metal walls and complete the electric circuit, two things happen. First, the corresponding relay circuit's switch flips from "on" to "off," recording that space as having a wall on that side. Then Theseus rotates 90° clockwise and moves forward. In this way, it systematically moves through the maze until it reaches the target, recording the exits and walls for each square it passes through.'*
>
> *— Martin Klein*

Symbolic AI problems

Several symbolic AI projects sought to hard-code knowledge about domains in formal languages, but it has always been a costly, slow process that could not scale.

Anyhow, by 1965, there were already programs that could solve any solvable problem described in logical notation (Russell, Norvig, and Davis 2010, 4). However, hubris and lack of philosophical perspective made computer scientists believe that "intelligence was a problem about to be solved[20]."

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

[20] Marvin Minsky, head of the artificial intelligence laboratory at MIT (1967)

Those inflated expectations lead to disillusionment and funding cuts[21] (Russell, Norvig, and Davis 2010). They failed to estimate the inherent difficulty in slating informal knowledge in formal terms: the world has many shades of grey. Besides, complexity theory had yet to

[21] Sometimes called *winters*.

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

be developed: they did not count on the exponential explosion of their problems.

Connectionism: a different approach

The fundamental idea in Connectionism is that *intelligent behaviour emerges from a large number of simple computational units when networked together* (Goodfellow, Bengio, and Courville 2016).

It was pioneered by McCulloch and Pitts in 1943 (McCulloch and Pitts 1943). One of Connectionism's first wave developments was Frank Rosenblatt's Perceptron, an algorithm for learning binary classifiers, or more specifically threshold functions:

$$y = \begin{cases} 1 \text{ if } Wx + b > 0 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

where W is the vector of weights, x is the input vector, b is a bias, and y is the classification. In neural networks, a perceptron is an artificial neuron using a step function as the activation function.

See Figure 7, termites self-cooling mounds keep the temperature inside at exactly $31°C$, ideal for their fungus-farming; while the temperatures outside range from 2 to $40°C$ throughout the day. Such building techniques inspired architect Mike Pearce to design a shopping mall that uses a tenth of the energy used by a conventional building of the same size.

From where does termites intelligence come?

> *'Individual termites react rather than think, but at a group level, they exhibit a kind of cognition and awareness of their surroundings. Similarly, in the brain, individual neurons do not think, but thinking arises in their connections.'*
>
> — *Radhika Nagpal, Harvard University (Margonelli 2016).*

Such collective intelligence happens in groups of just a couple of million termites. There are around 80 to 90 billion neurons in the human brain, each less capable than a termite, but collectively they show incomparable intelligence capabilities.

In contrast with the symbolic approach, in neural networks, the knowledge is not explicit in symbols but implicit in the strength of the connections be-

Figure 6: Building in Harare, Zimbabwe, is mod- elled after termite mounds. Photo by Mike Pearce.

Figure 7: Cathedral termite mound, Australia. Photo by Awoisoak Kaosiowa, 2008.

Goodfellow, Ian J., Yoshua Bengio, and Aaron C. Courville. 2016. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press.

McCulloch, Warren S., and Walter Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *The Bulletin of Mathematical Biophysics* 5 (4): 115–33.

Margonelli, Lisa. 2016. "Collective Mind in the Mound: How Do Termites Build Their Huge Structures?" https://www.nationalgeographic.com/news/2014/8/140731-termites-mounds-insects-entomology-science/.

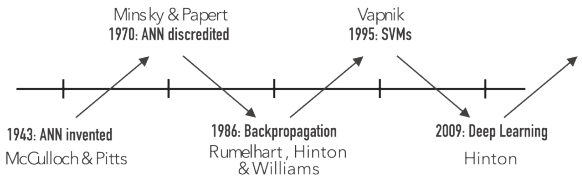


Figure 8: A brief history of connectionism.

tween the neurons. Besides, it is a very general and flexible approach since these connections can be updated algorithmically: they are algorithms that *learn*: the connectionist approach is an example of what we now call Machine Learning.

Machine Learning

Look at Figure 9. Is this a picture of a cat? How to write a program to do such a simple classification task (cat/no cat)? One could develop clever ways to use *features* from the input picture and process them to guess. Though, it is not an easy program to design. Worse, even if one manages to program such a task, how much would it worth to accomplish a related *task*, to recognise a dog, for example? For long, this was the problem of researchers in many areas of interest of AI:CV, NLP, Speech Recognition SR; much mental effort was put, with inferior results, in problems that we humans solve with apparent ease.

Figure 9: Is this a cat?

The solution is an entirely different approach for building artificial intelligence: instead of making the program do the *task*, build the program that outputs the program that does the *task*. In other words, learning algorithms use "training data" to infer the transformations to the input that generates the desired output.

Types of learning

Machine Learning can happen in different scenarios, which differ in the availability of training data, how training data is received, and how the test data is used to evaluate the learning. Here, we describe the most typical of them (Mohri, Rostamizadeh, and Talwalkar 2012):

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. The MIT Press.

- **Supervised learning**: The most successful scenario. The learner receives a set of labelled examples as training data and makes predictions for unseen data.
- **Unsupervised learning**: The learner receives unlabelled training data and makes predictions for unseen instances.
- **Semi-supervised learning**: The learner receives a training sample con-

sisting of labelled and unlabelled data and makes predictions for unseen examples. Semi-supervised learning is usual in settings where unlabelled data is easily accessible, but labelling is too costly.

- **Reinforcement learning**: The learner actively interacts with the environment and receives an immediate reward for her actions. The training and testing phases are intermixed.

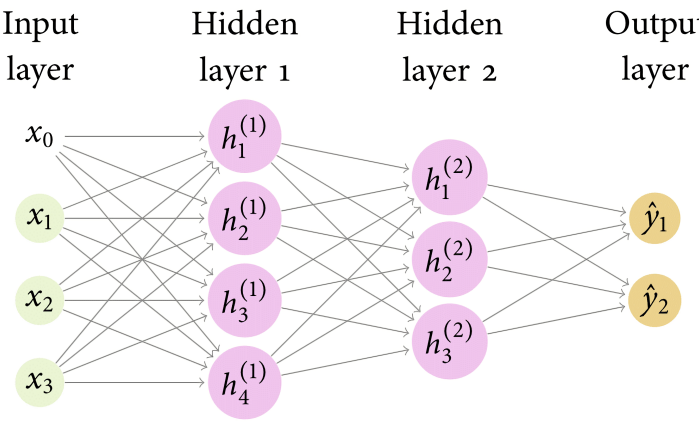
Deep Learning

The 2010s have been an AI Renaissance not only in academia but also in the industry. Such successes are mostly due to DL, in particular, supervised deep learning with vast amounts of data trained in GPUs. It was the decade of DL.

> *'Deep learning algorithms seek to exploit the unknown structure in the input distribution to discover good representations, often at multiple levels, with higher-level learned features defined in terms of lower-level features.'*
>
> *— Joshua Bengio (Bengio 2012)**

The name is explained by Goodfellow, Bengio, and Courville (2016): "*A graph showing the concepts being built on top of each other is a deep graph. Therefore the name, deep learning*" (Goodfellow, Bengio, and Courville 2016). Although it is a direct descendant of the connectionist movement, it goes beyond the neuroscientific perspective in its modern form. It is more a general principle of learning multiple levels of compositions.

The quintessential example of a deep learning model is the deep feedforward network or MLP (Russell, Norvig, and Davis 2010).



Bengio, Yoshua. 2012. "Deep Learning of Representations for Unsupervised and Transfer Learning." In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 17–36.

Goodfellow, Ian J., Yoshua Bengio, and Aaron C. Courville. 2016. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press.

Goodfellow, Ian J., Yoshua Bengio, and Aaron C. Courville. 2016. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press.

Russell, Stuart J., Peter Norvig, and Ernest Davis. 2010. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall Series in Artificial Intelligence. Prentice Hall.

Definition 0.3. Let,

- x be the input vector $\{x_1, \ldots, x_m\}$
- k be the layer index, such that $k \in [1, l]$,
- $W_{i,j}^{(k)}$ be the matrix of weights in the k-th layer, where $i \in [0, d_{k-1}], j \in [1, d_k]$ and $W_{0,:}^{(k)}$ are the biases
- σ be a nonlinear function,

a **MLPs** is a neural network where the input is defined by:

$$h^{(0)} = 1 \frown x$$

a hidden layer is defined by:

$$h^{(k)} = \sigma^{(k)}(W^{(k)\top} h^{(k-1)}).$$

The output is defined by:

$$\hat{y} = h^{(l)}.$$

Deep Learning is usually associated with DNNs, but the network architecture is only one of its components:

1. DNN architecture
2. SGD — the optimiser
3. Dataset
4. Loss function

The architecture is not the sole component essential to current Deep Learning success. The SGD plays a crucial role, and so does the usage of large datasets.

A known problem, though, is that DNNs are prone to overfitting ([[sec:bias-variance]][6]). Zhang et al. (2016) show state-of-the-art convolutional deep neural networks can easily fit a random labelling of training data (Zhang et al. 2016).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. “Understanding Deep Learning Requires Rethinking Generalization.” https://arxiv.org/abs/1611.03530.

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. “Understanding Deep Learning Requires Rethinking Generalization.” https://arxiv.org/abs/1611.03530.

Concluding Remarks

This chapter derived the need for a *language* from the definitions of *intelligence* and *intelligent agents*. An intelligent agent needs *language* to store her knowledge (what she has learned) and with that to communicate/share this knowledge with its future self and with other agents.

We claim (without proving) that a language can be derived from a definition of knowledge: an epistemic choice. We claim that mathematics and science can be seen as languages that differ in consequence of different views on what knowledge is and gave historical background on two epistemic views, Rationalism and Empiricism (Section 16,Section 16).

We gave historical background on AI and showed that different epistemic views relate to AI movements: Symbolism and Connectionism. We gave some background on basic AI concepts: intelligent agents, machine learning, types of learning, neural networks and deep learning, showing that DL relates to Connectionism and, hence, to science and an empiricist epistemology. Previously (**?@sec-bringing_science**), we have discussed that Computer Science generally relates to the rationalist epistemology. We hope this can help us better understand our research community.

Assumptions

1. A definition of intelligence Definition 0.2
2. An epistemic choice on the definition of Knowledge Section 16

A WISE MAN PROPORTIONS HIS BELIEF TO THE EVIDENCE.

— DAVID HUME

Probability Theory

In this chapter, propositional calculus and probability theory are derived from a list of desired characteristics for sceptical agents.

From Language to Probability

Formal Languages

We, as intelligent agents, do not know how Nature is; we only know how we perceive it. Our ideas are mental pictures of how we imagine Nature. Like in the story of the blind men and the elephant ([[blind_men]][1]), how do we know that our model is the same as someone else's? *Communicating*. We need to communicate with each other to check if our mental picture of Nature, our model, is consistent with the experience of others.[22]

[22] We can take this idea further and think that at any moment, we need to communicate with our past selves to check if new evidence is consistent with our prior model.

We use language to describe Nature. However, natural languages, like English, German, Portuguese, are ambiguous, and we need contextual clues and other information to more clearly communicate meaning. To avoid this, an intelligent agent uses formal language.

A *formal language* is a mathematical tool created for precise communication about a specific subject. For example, arithmetic is a language for calculations. Chemists have a language that represents the chemical structures of molecules. Programming languages are formal languages that express computations. In a nutshell, a formal language is a set of words (strings) whose letters (symbols) are taken from an alphabet and are well-formed according to a specific set of rules, grammar. Let $\mathrm{L} =< \Sigma, \Phi >$ be a formal language:

$$\Sigma = \{S_1, S_2, \cdots, S_n\} \text{ is an alphabet,} \tag{2}$$

$$\Phi = \Phi_1 \cup \Phi_2 \cup \cdots \cup \Phi_k \text{ is the grammar, where:} \tag{3}$$

Φ_1 is the set of unary operations,

Φ_2 is the set of binary operations,

...

Φ_k is the set of k-ary operations.

A formal language allows a quantitative description of a state of knowledge and defines how this state can be updated on new evidence.[23]

[23] An inference method defines the rules for updating knowledge.

With this definition, we can also think that a formal language is what Sowinski (2016) calls a *realm of discourse*, all the valid formed *strings*[24] that one can derive; everything one can *say* about Nature.

Sowinski, Damian Radoslaw. 2016. "Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy." PhD thesis.

[24] Strings, words, sentences, propositions, formulae are names used interchangeably through the literature.

Interestingly, formal languages allow us to manipulate representations of the environment without dealing with their semantics. They are the basis of *"Turing's strange inversion"*, (see [[turing_strange_inversion]][2]) by doing allowed operations on strings, computers can compute at a superhuman speed and accuracy without ever comprehending what they are doing.

From Rationalism to Propositional Calculus

Rational Agents can form representations of a complex world, use deduction as the inference process to derive updated representations, and use these new representations to decide what to do. In other words, rational agents are the consequence of the epistemological view of *rationalism*.

When a rational agent establishes a particular statement's truth value, all statements formed in her knowledge base from that statement instantly feel that update. Therefore, a rational agent cannot hold contradictions.

Desiderata for a rational language

We want to build a language for rational agents with the following desired characteristics:

i. **knowledge is absolute**; a sentence[25] can be either true or false;

ii. **unambiguous**, a constructed sentence can only have one meaning;

iii. **consistent**; a language without paradoxes, whatever path chosen to derive a sentence truth value will lead to the same assignment;

[25] A sentence can be either a single symbol or a string formed with several symbols according to the grammar.

iv. **minimal**; uses the most reduced set of symbols possible.

We call $L_R =< \Sigma_R, \Phi_R >$ the formal language built from these constraints, where sentences are either axiom symbols or compounded sentences formed using special symbols called operators, each operator denoting one operation, $\phi \in \Phi_R$.

It is possible to prove that L_R only needs one operator (Sowinski 2016; Jaynes 2003): NAND (or XOR), and it is also equivalent to Propositional Calculus.[26] In other words, Logic is the language that emerges from our desiderata, from rationalism. **Logic is the language of mathematics.**

Sowinski, Damian Radoslaw. 2016. "Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy." PhD thesis.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

A point worth mentioning is that using Logic as an agent formal language means the **implicit acceptance** of the constraints above.

From Empiricism to Probability Theory

The constraints that lead to Logic are very restrictive to use in the real-world; rational language has a comparatively small *realm of discourse*. Hume would say that it is only helpful for *relations of ideas*, talking in the abstract, and not for *matters of facts*, talking about reality.

A realm of discourse to talk about reality needs at least the empiricist perspective where knowledge is justified belief, and that one should *weigh her beliefs to the evidence*. The quantity that specifies to what degree we believe a proposition is true is constrained by other beliefs, i.e., previous experience and evidence gathered.

[26] Proposition is synonym to sentence and Propositional Calculus is also known as Sentential Calculus.

Sceptical Agents {#sec:sceptical_agents}

In the sceptical agent, the one derived from the empiricist epistemology (authors have called these agents epistemic agents (Caticha 2008), idealised epistemic agents (Sowinski 2016) or robots (Jaynes 2003)), beliefs are not independent of each other (Caticha 2008), they form an interconnected web that is the agent's knowledge base. The update mechanism, its inference method, follows the principle of minimality, i.e. it tries to minimise the change in the knowledge base.

Caticha, Ariel. 2008. "Lectures on Probability, Entropy, and Statistical Physics." https://arxiv.org/abs/0808.0012.

Sowinski, Damian Radoslaw. 2016. "Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy." PhD thesis.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Desiderata for a sceptical language {#sec-desiderata_language_sceptical}

As we did for rational agents, let us state a set of desired characteristics for the language of science, $L_S =< \Sigma_S, \Phi_S >$ (Sowinski 2016; Caticha 2008; Jaynes 2003):

Caticha, Ariel. 2008. "Lectures on Probability, Entropy, and Statistical Physics." https://arxiv.org/abs/0808.0012.

Sowinski, Damian Radoslaw. 2016. "Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy." PhD thesis.

Caticha, Ariel. 2008. "Lectures on Probability, Entropy, and Statistical Physics."

i. **Knowledge is a set of beliefs, quantifiable by real numbers and dependent on prior evidence**: Let $S_i \in \Sigma_S$ be sentences about the world. Given any two statements S_1, S_2, the agent must be able to say that S_1 is more plausible than S_2, or that S_2 is more plausible than S_1 or that S_1 and S_2 are equally plausible. Thus we can list statements in an increasing plausibility order. Real numbers can represent this transitive ordering.[27] Let b be a measure of degrees of belief in S given some previous knowledge K: [28]

$$b : \Sigma_S \to \mathbb{R} \tag{4}$$
$$b : S \mapsto b(S|K) \tag{5}$$

Here we capture that plausibility (degrees of belief) is not a function of a sentence, but a relation between a sentence and a given assumed prior knowledge K.

ii. **"Common sense:"**

The plausibility of compound sentences should be related by some logical function to the plausibility of the sentences that form them. We already showed that a minimal rational language has only one operator. Here, instead of using the NAND operator, for a matter of familiarity, let us use the almost minimal language with the operators NOT ($\neg$) and AND ($\wedge$). In this setting, we are saying there are such functions f and g that (Sowinski 2016):

$$b(\neg S|K) = f[b(S|K)] \tag{NOT}$$
$$b(S-1 \wedge S_2|K) = g[b(S_1|K), b(S_1|S_2), b(S_2|K), b(S_2|S_1)] \tag{AND}$$

iii. **Consistency**: The functions f and g must be consistent with the grammar Φ (production rules). Consistency guarantees that whatever path used to compute the plausibility of a statement in the context of the same knowledge web (the same set of constraints) must lead to the same degree of belief.

(a) Beliefs that depend on multiple propositions cannot depend on the order in which they are presented.

(b) No proposition can be arbitrarily ignored.

(c) Propositions that are identical must be assigned the same degree of belief.

Such desiderata have a name; it is known as Cox's axioms, and one can derive the Sum Rule and the Product Rule (see [1.4]) from them,

[27] We are implicitly assuming that the language we are building has infinite statements. A further discussion on this continuity assumption can be found in .

[28] Using $(S|K)$ in a function is a notation abuse that we accept to explain the idea better.

Sowinski, Damian Radoslaw. 2016. "Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy." PhD thesis.

therefore, also the Bayes' Theorem ([1.9]), and reverse-engineer Kolmogorov's Axioms of Probability Theory (that will be seen in [[sec:kolmogorov_axioms]][3], [[fig:kolmogorov]][4]) (Sowinski 2016; Jaynes 2003; Caticha 2008; Terenin and Draper 2015).

In other words, Probability Theory is the language that emerges from our desiderata, from empiricism. *'Probability theory is the Logic of Science'* (Jaynes 2003), and our measure b is usually called probability P.

Again, here we explicit that by using Bayesian inference to build and communicate concepts of the world (models), we are assuming Cox's axioms above.

Assumptions and their consequences

Let us take this opportunity to explore what some assumptions mean to human intelligence in particular. It is indisputable[29] that humans are not rational, neither sceptical agents. The whole idea of imagining an epistemic agent is a consequence of addressing intelligence without human complexities.

However, are humans irrational because of biology or psychology? Are we irrational for lack of will, or could it be that Nature wires the human brain in a way that pr*events* us from following these axioms? Here we argue that biology has an important role. Researchers have found, for instance, that visual acuity can be permanently impaired if there is a sensory deficit during early post-natal development (Wiesel 1982). Futhermore, if the human brain is not exposed to some samples in its infancy, it will never achieve the accuracy level if it had experienced them, regardless of experiencing those examples later. In other words, *human beliefs depend on the order in which pieces of evidence are presented*, contradicting Cox's axiom [[axiom:order]][5].

Formalizing Probability Theory

We derived Cox's axioms from a list of desired properties of the language for sceptical agents. We also know that it is possible to derive Kolmogorov's Axioms (which will be defined soon in [[sec:kolmogorov_axioms]][3]) from those axioms. In the next sections, we will use the Kolmogorov Axioms to formalise Probability theory.

Several concepts in the following sections are *relations of ideas*, not *matters of fact*. For example, the probability of an *event* E, P(E), can be computed by marginalisation (as we will show in [1.8]), but as discussed before, there are

Sowinski, Damian Radoslaw. 2016. "Complexity and Stability for Epistemic Agents: The Foundations and Phenomenology of Configurational Entropy." PhD thesis.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Caticha, Ariel. 2008. "Lectures on Probability, Entropy, and Statistical Physics." https://arxiv.org/abs/0808.0012.

Terenin, Alexander, and David Draper. 2015. "Cox's Theorem and the Jaynesian Interpretation of Probability." https://arxiv.org/abs/1507.06597.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

[29] Unless you are an economist.

Wiesel, Torsten N. 1982. "Postnatal Development of the Visual Cortex and the Influence of Environment." *Nature* 299 (5884): 583–91. https://doi.org/10.1038/299583a0.

no beliefs in a vacuum. In reality, there is only the probability of an *event* E given some background knowledge *K*. This change of epistemological perspective is essential to be remembered now that we will expose the idealised development of Probability Theory.

Experiments, Sample Spaces and Events

The set of possible outcomes of an *experiment* is the *sample space* Ω. Let us use the canonical *experiment* of rolling a dice. In this experiment, the sample space is:

$$\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\} \tag{6}$$

An **outcome** or **realisation** is a point $\omega \in \Omega$:

$$\omega_3 = \boxdot \tag{7}$$

$$\Omega = \{\omega_1 = \boxdot, \cdots, \omega_6 = \boxdot\}\,. \tag{8}$$

An **Event** is something that can be said about the *experiment*, "The dice rolled to an odd number". It is a true proposition. Nevertheless, easier than writing so much, we denote *events* with letters. **Events are subsets of** Ω (see [[fig:event_A]][6]).

$$A = \{a_1 = \boxdot, a_2 = \boxdot, a_3 = \boxdot\} \tag{9}$$

$$A \subset \Omega \tag{10}$$

We say that $A_1, A_2, \cdots$ are **mutually exclusive** or **disjoint** *events* if $A_i \cap A_j = \emptyset, \forall i \neq j$. For example, A is the *event* "the dice rolled to the value 5" and B is the *event* "the dice rolled to an even number". In this case, A and B are disjoint (see [[fig:disjoint_events]][7]).

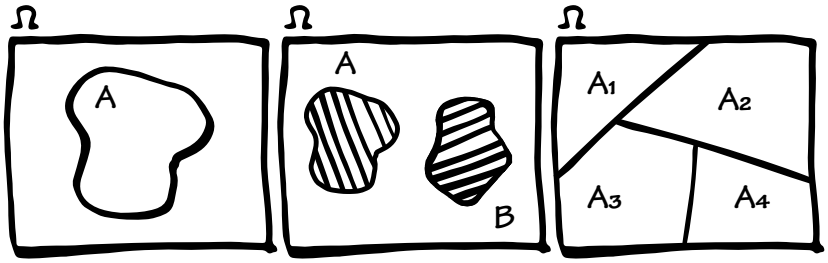


Figure 10: A partition of Ω.

Visually, we can represent the probability of an *event* A, $P(A)$, as the proportion of the sample space the *event* occupies. To differentiate *events* from

their probabilities, we will shade the area of the *event*.

Kolmogorov's Axioms and their direct consequences.

Directly from the Kolmogorov Axioms, one can derive (Jaynes 2003) other properties (see [[fig:axiom1, fig:axiom2, fig:axiom3]][9]):

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

<— [1]: #blind_men {reference-type="ref" reference="blind_men"}
[2]: #turing_strange_inversion {reference-type="ref" reference="turing_strange_inversion"} [1.4]: #sec:probability {reference-type="ref" reference="sec:probability"} [1.9]: #sec:bayes_theorem {reference-type="ref" reference="sec:bayes_theorem"} [3]: #sec:kolmogorov_axioms {reference-type="ref" reference="sec:kolmogorov_axioms"}
[4]: #fig:kolmogorov {reference-type="ref" reference="fig:kolmogorov"}
[5]: #axiom:order {reference-type="ref" reference="axiom:order"} [1.8]: #marginalisation {reference-type="ref" reference="marginalisation"}
[6]: #fig:event_A {reference-type="ref" reference="fig:event_A"} [7]: #fig:disjoint_events {reference-type="ref" reference="fig:disjoint_events"}
[8]: #fig:partition {reference-type="ref" reference="fig:partition"} [9]: #fig:axiom1, fig:axiom2, fig:axiom3 {reference-type="ref" reference="fig:axiom1, fig:axiom2, fig:axiom3"} [1.10]: #sec:random_variables {reference-type="ref" reference="sec:random_variables"} [10]: #eq:P(A, B)>0 {reference-type="eqref" reference="eq:P(A, B)>0"}
[1.1.3.0.1]: #sec:sceptical_agents {reference-type="ref" reference="sec:sceptical_agents"} [11]: #eq:conditional_probability {reference-type="ref" reference="eq:conditional_probability"} [12]: #eq:joint_probability {reference-type="ref" reference="eq:joint_probability"}
[13]: #eq:law_of_total_probabilities {reference-type="ref" reference="eq:law_of_total_probabilities"} [14]: #eq:P(X=x) {reference-type="eqref" reference="eq:P(X=x)"} [15]: #fig:sampling {reference-type="ref" reference="fig:sampling"} [16]: #sec:from_rationalism, sec:from_empiricism {reference-type="ref" reference="sec:from_rationalism, sec:from_empiricism"} [17]: #ch:artificial_intelligence {reference-type="ref" reference="ch:artificial_intelligence"} [1.1.1]: #sec:formal_language {reference-type="ref" reference="sec:formal_language"} [18]: #sec:desiderata_language,sec:desiderata_language_sceptical {reference-type="ref" reference="sec:desiderata_language,sec:desiderata_language_sceptical"}
[19]: #cox {reference-type="ref" reference="cox"} [20]: #def:intelligence {reference-type="ref" reference="def:intelligence"} [21]: #sec:rationalism, sec:empiricism {reference-type="ref" reference="sec:rationalism, sec:empiricism"} [1.1.3.0.2]: #sec:desiderata_language_sceptical {reference-type="ref" reference="sec:desiderata_language_sceptical"}
[22]: #consistency {reference-type="ref" reference="consistency"}
[1.1.2.0.2]: #sec:desiderata_language {reference-type="ref" ref-

erence="sec:desiderata_language"} [23]: #rational_consistency
{reference-type="ref" reference="rational_consistency"} [24]: #ratio-
nal_minimality {reference-type="ref" reference="rational_minimality"}
[25]: #absolute_truth {reference-type="ref" reference="absolute_truth"}
[26]: #unambiguous {reference-type="ref" reference="unambiguous"}
[27]: #beliefs {reference-type="ref" reference="beliefs"} [28]: #com-
mon_sense {reference-type="ref" reference="common_sense"} –>
–>

Tufte-Quarto

Questions

List of questions

- Why `Welcome` from `/index.qmd` is appearing in pdf?
- ~~`Part /background.qmd is not appearing in TOC (pdf). Is this a Tufte-book.cls problem?`~~
- pdf figures in website are being framed. Why?

  – can I remove file extension and site get svg or png and pdf get .pdf?

- Header configuration is set to include short-tile instead of title. Couldn't insert just a space, though. Why blank space without mathmode didn't work?
- How to include math macros in the html?
- What is the page full of definitions that appear while routing in the website? `math_definitions.tex`?
- Too much info in the margin citation. Create custom bib style?
- How to show title-block for each chapter in website?

Known-issues

1. Tufte-book can't handle label inside caption.

   Tufte-book.cls breaks when processing the line bellow: `markdown ![A way of flying](/Images/goya.jpg){.column-body #fig-goya}` which becomes:

```
\begin{figure}

{\centering \includegraphics{Images/goya.jpg}

}

\caption{\label{fig-goya}A way of flying}

\end{figure}
```

2. Can't render svg image in pdf and can't render pdf image in html.

   - current solution is quite ugly:

```
::: {.content-hidden unless-format="pdf" }

![IBT "genealogy" tree.](/Images/dissertation-map.pdf){.column-margin width=90%}

:::

::: {.content-hidden unless-format="html"}

![IBT "genealogy" tree.](/Images/dissertation-map.svg){.column-margin width=90%}

:::
```

3. Tufte-class works only until subsection -> `### subsection`; `#### subsubsection`-> returns error

4. sidecite is duplicating citations in the same margin. Solved this same problem in my dissertation and in the kaobook class. Only problem is that Tufte-book class is a little too cryptic for me.

To-dos

- [ ] create documentation showcasing project type
  - [ ] show jupyter/matplotlib examples:
    - https://matplotlib.org/2.0.2/examples/pylab_examples/polar_legend.html
    - https://www.ajnisbet.com/blog/tufte-in-matplotlib
    - https://www.andrewheiss.com/blog/2017/08/10/exploring-minards-1812-plot-with-ggplot2/
    - https://deepnote.com/workspace/fred-guth-c6b0391e-7e85-43ad-ba15-bd505c864b75/project/tufte-5ba3f476-1ce2-4d99-b9be-d045148bbc77/notebook/tufte-in-python-58e3589d00c34ac699f715ad69c4bfd3
    - https://matplotlib.org/2.0.2/examples/pylab_examples/ellipse_collection.html
    - https://matplotlib.org/2.0.2/examples/lines_bars_and_markers/fill_demo_features.html
    - https://matplotlib.org/2.0.2/examples/shapes_and_collections/scatter_demo.html
- [ ] move my dissertation to an examples folder (maybe another repo? .quartoignore my dissertation?)
- [ ] fix known bugs