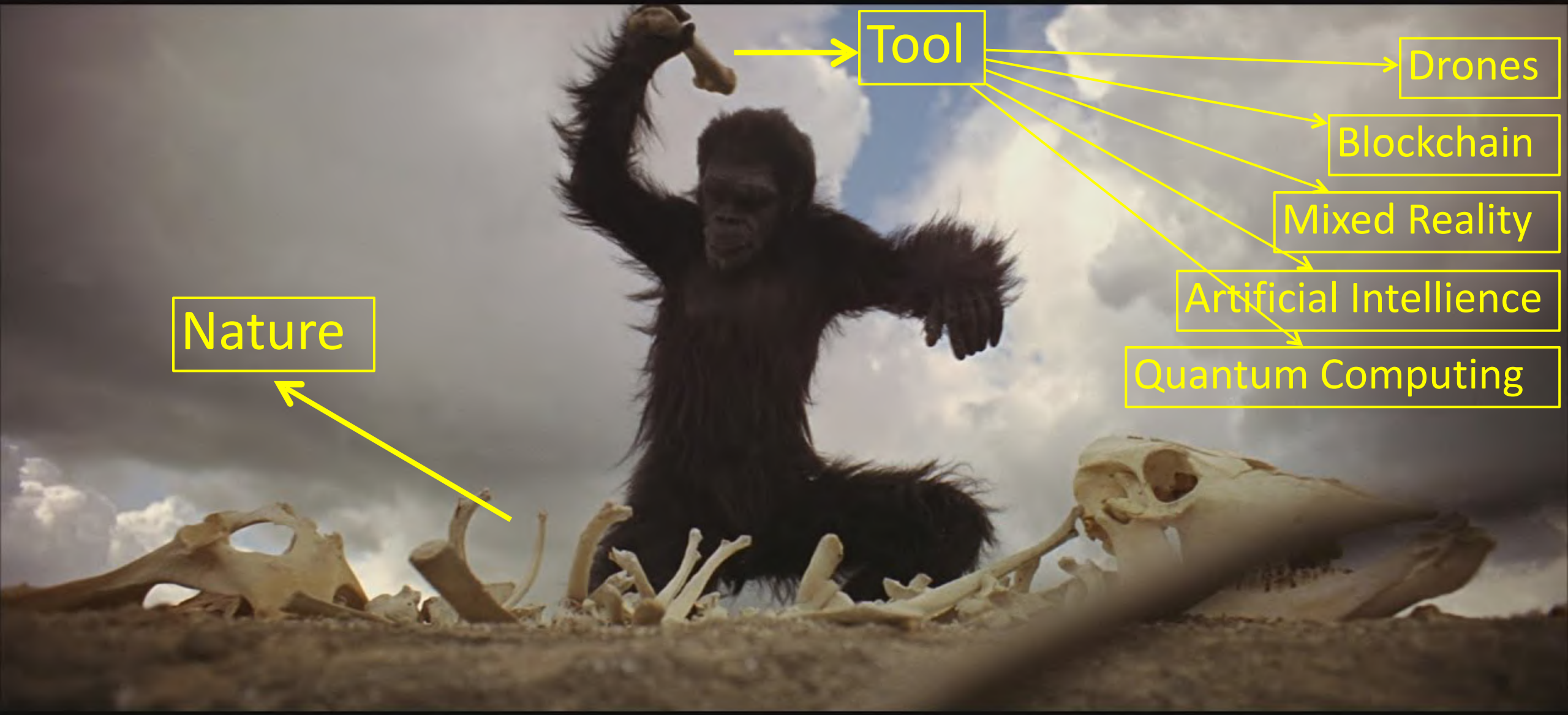
Boundaries of Data Science

A Critical Review

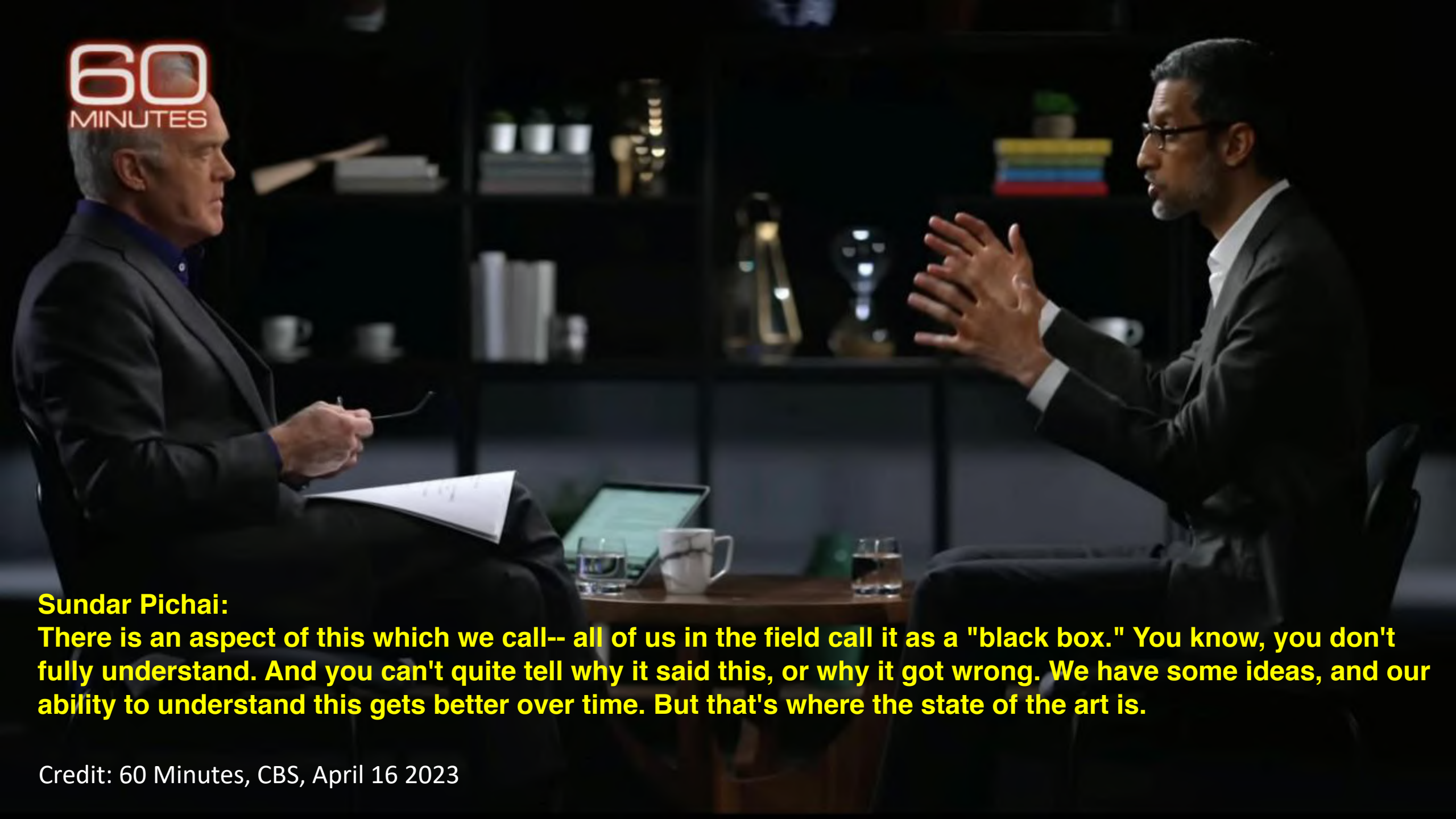
Prepared for Smartcon/Data Science Days 2023

Bora Üzüm

May 12, 2023/İstanbul



Credit: 2001: A Space Odyssey, Dawn of Man Scene



Sundar Pichai:
There is an aspect of this which we call-- all of us in the field call it as a "black box." You know, you don't fully understand. And you can't quite tell why it said this, or why it got wrong. We have some ideas, and our ability to understand this gets better over time. But that's where the state of the art is.

Credit: 60 Minutes, CBS, April 16 2023

Credit: The Lead, CNN, May 2 2023

The Battle for Data Science

1 Introduction

Through the years, the database community has periodically looked at developments in technology and engaged in hand-wringing over the idea that we are becoming irrelevant. The cry "have we missed the boat – again" is common; e.g., here is a panel I served on several years ago [8]. My goal in this essay is to argue that the database field and the techniques that have come from this research are still essential for "data science," that is, for the exploitation of data to solve problems of importance in application fields – science, commerce, medicine and such. I believe, as I assume most readers of this article believe, that the field of database systems has always had at its core the study of how to deal with the largest amounts of data possible at the time, whether that be megabytes of corporate payroll data, terabytes of genomic information, or petabytes of satellite output. Thus whatever study of data is necessary at the time – that's our job.

To advance this argument, I want to look at three issues:

1. Is the field of statistics really the essential ingredient in data science?
2. Is machine learning really what data science is all about?
3. Is data science a danger to decent societal norms?

Hint: my answer to all three is "no." I'll try to address each of these in turn.

Credit: Ullman, J.D. (2020). The Battle for Data Science. *IEEE Data Eng. Bull., 43*, 8-14.

The Battle Line: Venn Diagrams for Data Science

intersecting: "hacking skills," "math and statistics," and "substantive expertise." At the roundtable, this diagram was shown several times to illustrate the importance of statistics, and I have seen statisticians in several other contexts showing the same diagram to explain the importance of their field to data science. I reproduce the diagram in Fig. 1, but I have added my own edits and comments to explain what is misleading about the diagram.

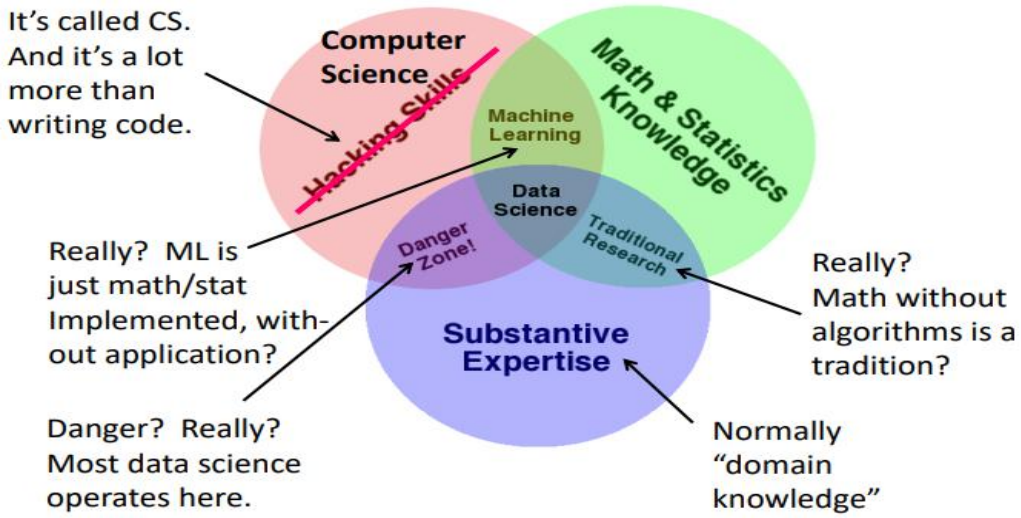


Figure 1: The Conway Venn diagram for data science

Credit: Ullman, J.D. (2020). The Battle for Data Science. *IEEE Data Eng. Bull., 43*, 8-14.

The Battle Line: Venn Diagrams for Data Science

I too offered a Venn diagram (Fig. 2) that I believe better represents the relationships between the fields. There is computer science and the various domain sciences, and somewhere in the intersection of these is data science. Machine learning is a branch of computer science – a very important subset these days. Some of machine learning is used for data science, although there are other uses of machine learning that are more internal to computing. Many of these applications are considered "artificial intelligence" these days, e.g., driverless cars or intrusion detection. Finally, I see both math and statistics as very important tools for all of computer science, and the small bubbles in my diagram do not do justice to their importance. However, I drew them as shown to emphasize that they do not really impact domain sciences directly. but rather they do so through the software that is developed, often with their important aid.

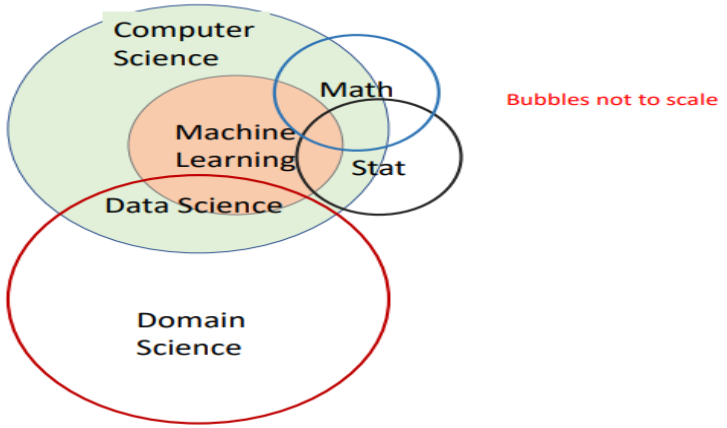


Figure 2: A personal view of the relationship between computer science, machine learning, and statistics

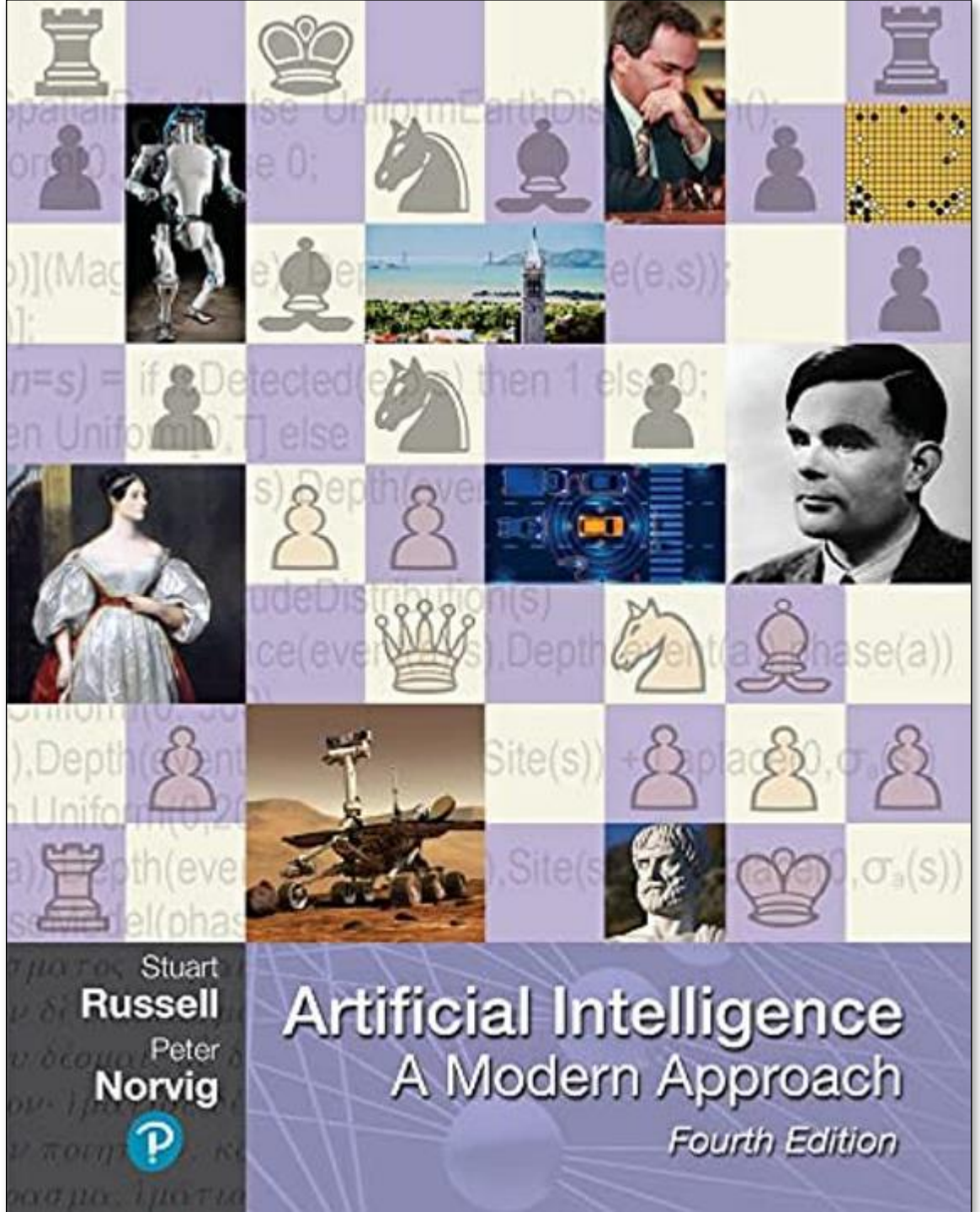
Credit: Ullman, J.D. (2020). The Battle for Data Science. *IEEE Data Eng. Bull., 43*, 8-14.

White House Meets With AI Leaders in Attempt to 'Protect Our Society‘

Jason Nelson, Yahoo Finance, May 5 2023

Credit: 2001: A Space Odyssey, Monolith Scene

Credit: The School of Athens, Raffaello



CHAPTER 28

PHILOSOPHY, ETHICS, AND SAFETY OF AI

In which we consider the big questions around the meaning of AI, how we can ethically develop and apply it, and how we can keep it safe.

Philosophers have been asking big questions for a long time: How do minds work? Is it possible for machines to act intelligently in the way that people do? Would such machines have real, conscious minds?

To these, we add new ones: What are the ethical implications of intelligent machines in day-to-day use? Should machines be allowed to decide to kill humans? Can algorithms be fair and unbiased? What will humans do if machines can do all kinds of work? And how do we control machines that may become more intelligent than us?

28.1 The Limits of AI

Weak AI
Strong AI

In 1980, philosopher John Searle introduced a distinction between **weak AI**—the idea that machines could act *as if* they were intelligent—and **strong AI**—the assertion that machines that do so are *actually* consciously thinking (not just *simulating* thinking). Over time the definition of strong AI shifted to refer to what is also called "human-level AI" or "general AI"—programs that can solve an arbitrarily wide variety of tasks, including novel ones, and do so as well as a human.

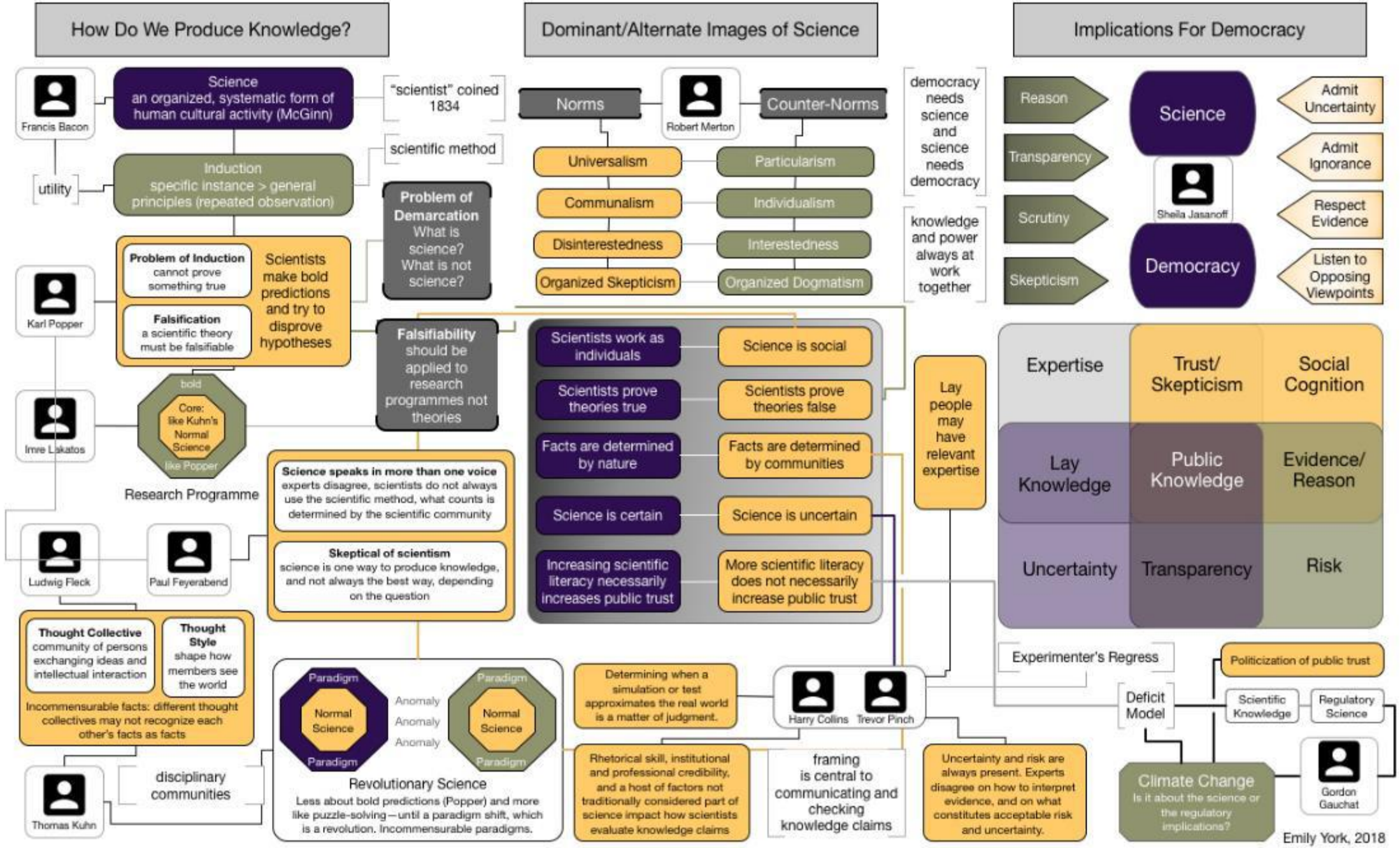
Critics of weak AI who objected to the very possibility of intelligent behavior in machines now appear as shortsighted as Simon Newcomb, who in October 1903 wrote "aerial flight is one of the great class of problems with which man can never cope"—just two months before the Wright brothers' flight at Kitty Hawk. The rapid progress of recent years does not, however, prove that there can be no limits to what AI can achieve. Alan Turing (1950), the first person to define AI, was also the first to raise possible objections to AI, foreseeing almost all the ones subsequently raised by others.

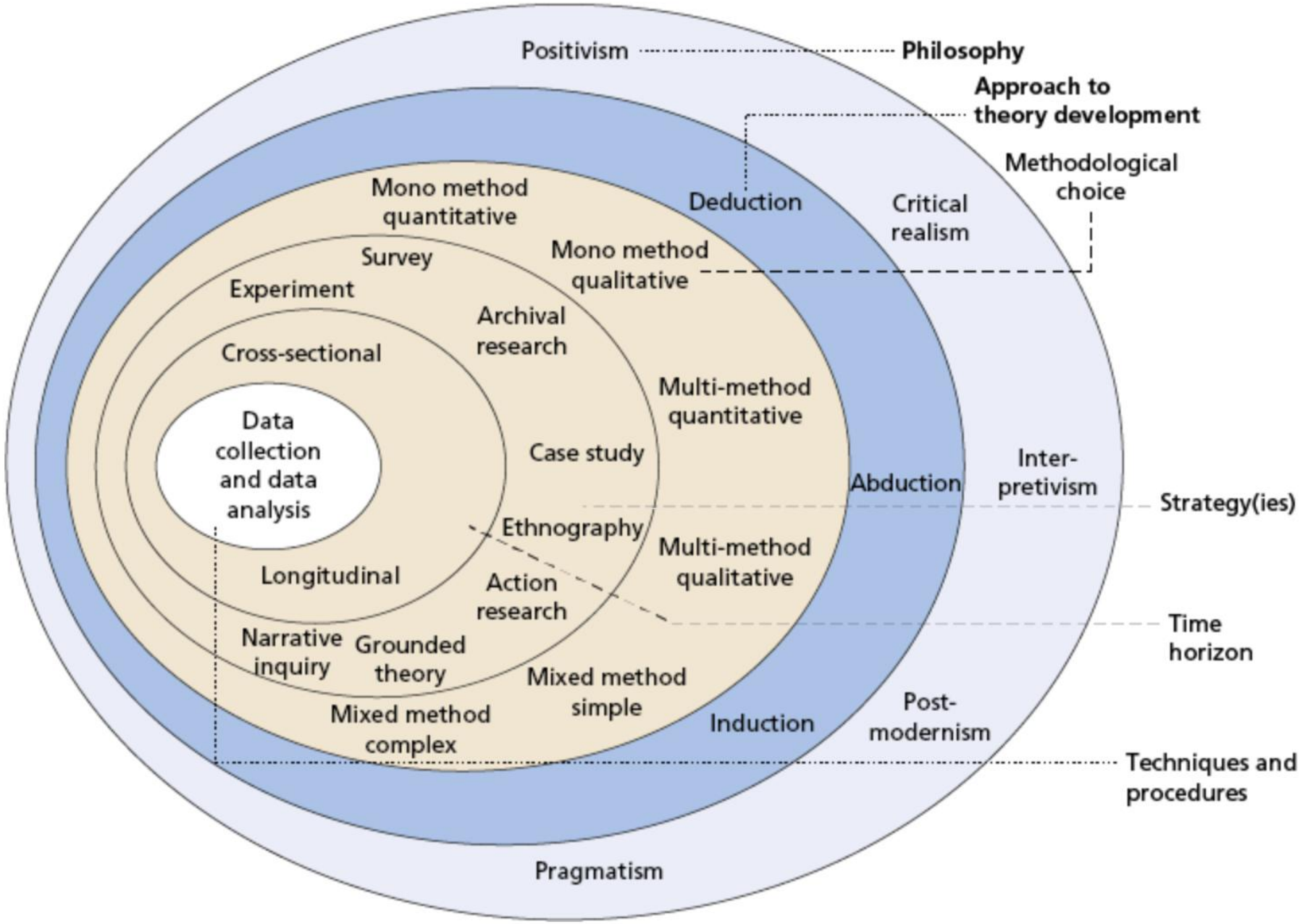
28.1.1 The argument from informality

Turing's "argument from informality of behavior" says that human behavior is far too complex to be captured by any formal set of rules—humans must be using some informal guidelines that (the argument claims) could never be captured in a formal set of rules and thus could never be codified in a computer program.

A key proponent of this view was Hubert Dreyfus, who produced a series of influential critiques of artificial intelligence: *What Computers Can't Do* (1972), the sequel *What*

Philosophy of Science: Key Concepts in ISAT 131 Technology, Science, and Society





Source: ©2018 Mark Saunders, Philip Lewis and Adrian Thornhill

The Paper by Floridi, Taddeo, Wang, Watson, Desai in 2022:

- Epistemological foundations of Data Science

Characterization	Enquiry	Generated Knowledge	Black-box Problems	Science in Data intensive paradigm
Minimalist & Maximalist	DS as Statistics	Inference Modes	Conceptual	Math application agnosticism
Descriptive	DS as Science	Epistemic Products	Non-Conceptual	Theory-free science
Normative	Else			

Minimalist & Maximalist Characterizations

- Minimalist:
  - DS is “the business of learning from data”
  - data scientist is someone who “uses data to solve problems”
- Maximalist:
  - DS concerns (a) Correlative/**predictive knowledge** & (b) **causal knowledge**
  - The **relation** of quantitative data **to a real-world** problem, often in the presence of **variability and uncertainty**
  - Data Scientist means a professional who uses scientific methods to **liberate** and **create meaning** from **raw data**
  - DS blends **statistical** and **computational** thinking… It connects statistical models and computational methods to **solve discipline-specific problems**.

Descriptive & Normative Taxonomies (1/3)

- 1962, Tukey
  - the first descriptive taxonomy of data analysis: “**procedures for analysing data** and techniques **for interpreting the results** of such procedures; ways of planning the **gathering of data** to make its analysis easier, more precise, or more accurate; all the **machinery and results of (mathematical) statistics** which apply when analysing data”
- 1997, Wu
  - **data collection** (experimental design, sample surveys); **data modelling and analysis**; **problem understanding/solving, and decision making**
- 2017, Donoho
  - **collection, management, processing, analysis, visualization, and interpretation** of vast amounts of **heterogeneous data** associated with a **diverse** array of scientific, translational, and **interdisciplinary** applications

Descriptive & Normative Taxonomies (2/3)

- 1993, Chambers
  1. Preparing data (planning, collection, organization, and validation);
  2. Analysing data (by models or other summaries);
  3. Presenting data (in written, graphical or other form).
- 2001, Cleveland
  1. Multidisciplinary investigations (data analysis in the context of different discipline specific areas)
  2. Models and methods for data (statistical models, model-building methods, estimation methods, etc.)
  3. Computing with data (hardware, software, algorithms)
  4. Pedagogy (curriculum planning, school/college/corporate training)
  5. Tool evaluation (descriptive and revisionary analysis of tools and their methods of development)
  6. Theory (foundational and theoretical problems in data science)

Descriptive & Normative Taxonomies (3/3)

- 2017, Donoho
  1. Data gathering, preparation and exploration
  2. Data representation and transformation
  3. Computing with data
  4. Data modelling
  5. Data visualisation and presentation
  6. Science about data science
- Floridi et al conclude as:

> «Data science is the **study of information systems** (natural or artificial), by **probabilistic reasoning** (e.g., inference and prediction) implemented with **computational** tools (e.g., databases and algorithms).»

The knowledge generated by data science

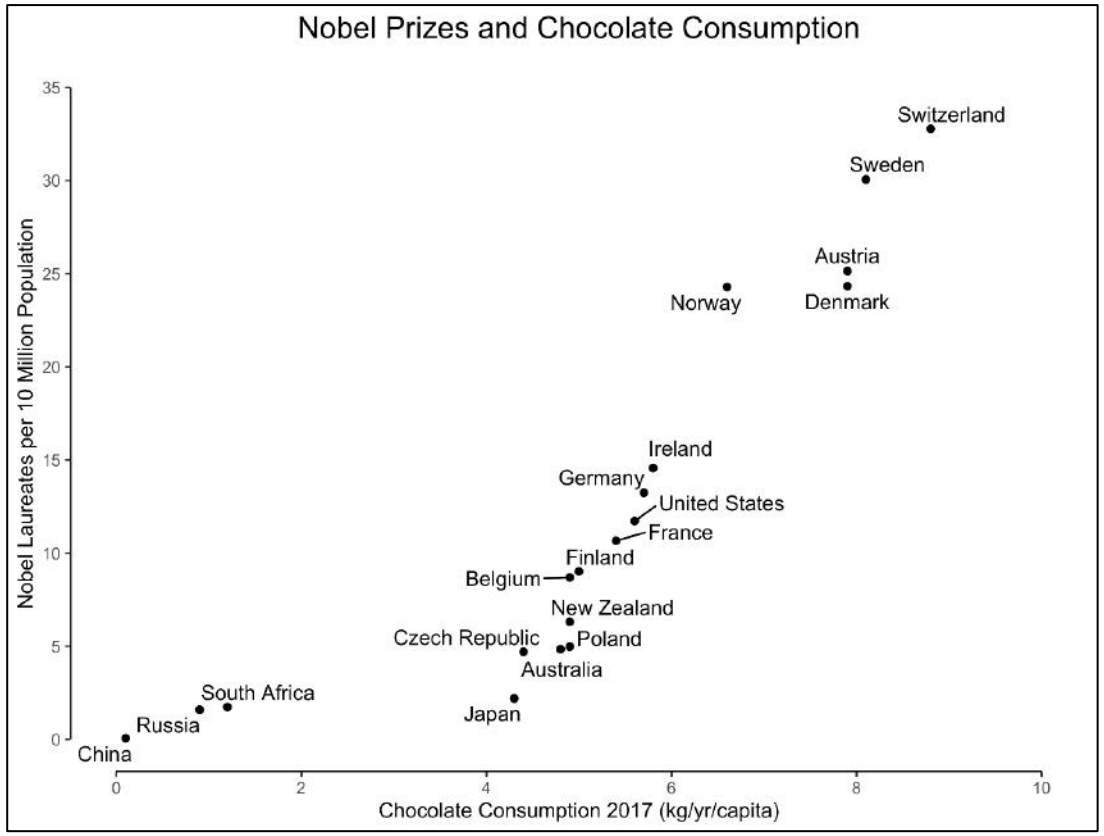
Modes of Inference (how)

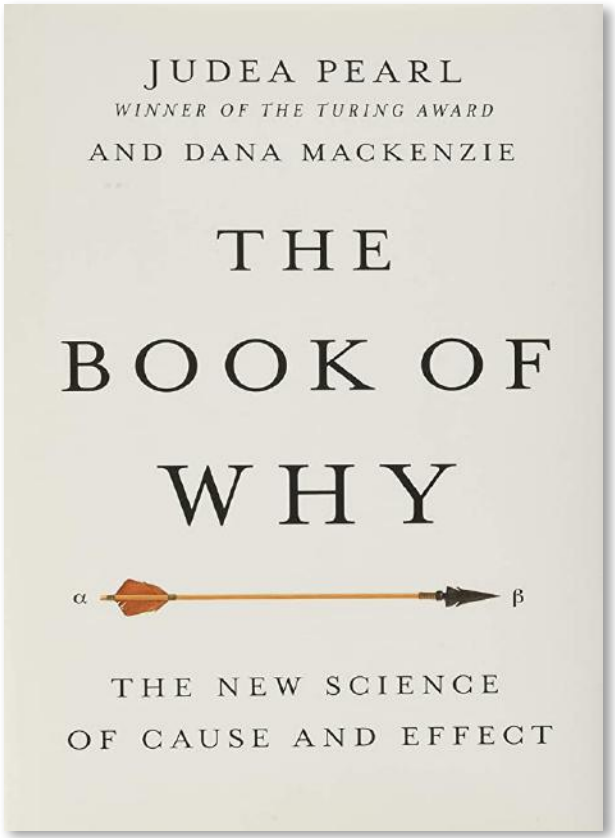
Deductive	Inductive			Abductive
Probablity Theory, Differential Calculus, Functional Analysis…	Solves Hume's problem by statistical testing?	Object Induction: next X is Y	Rule Induction: all Xs are Ys	Explore&Exploit

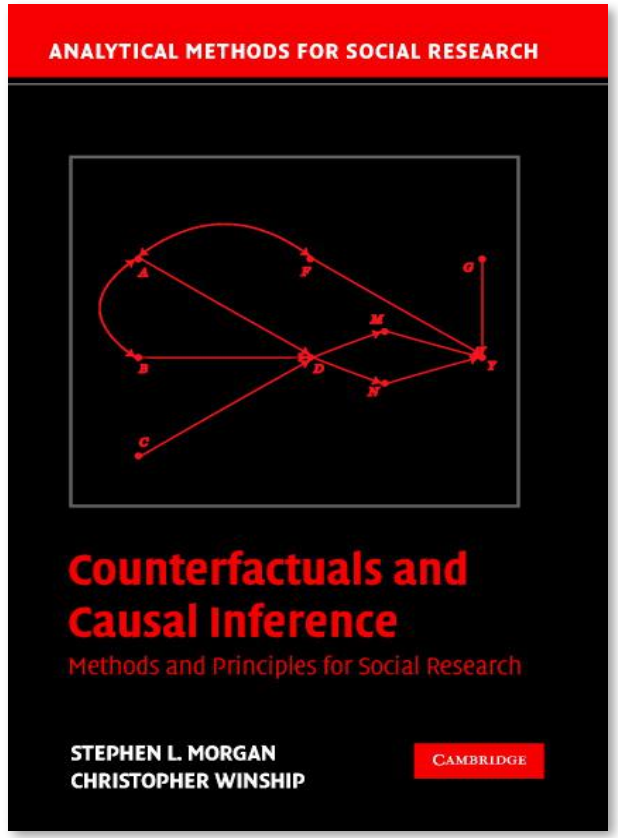
Epistemic Products (what)

Supervised	Unsupervised	Reinforcement
Context change problems, misclasification	Clustering, Generative Models, Autoencoders … Overfitting	Explore&Exploit AlphaGo…
Associations	Structures	Policies

Missing piece: Causal Knowledge







Reinforcement learning seems most promising way.

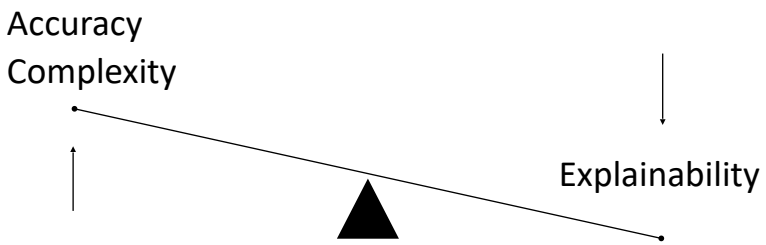
Structural Causal Model (association/intervention/counterfactuals)

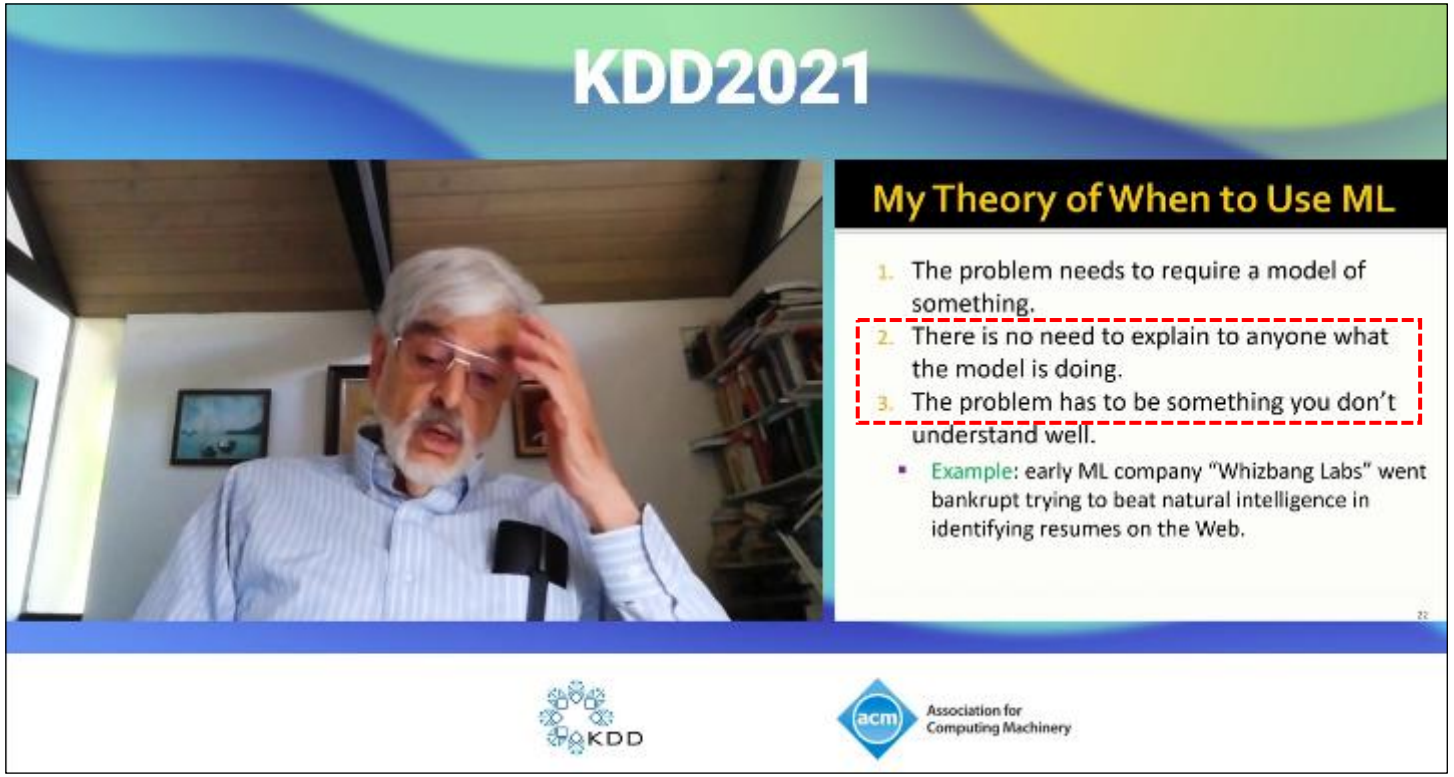
Black Box

Interpretable if experts can determine, by examining its inner workings, why it came to a conclusion.

Explainable if the system can give the reasons for its conclusion.

Auditable if we can tell how the system got to a state, produced an output, what was responsible for each step, and who is accountable.





- ❑ **Knowledge = Justified, True, Belief**
- ❑ Algorithmic **Gettier Case** occurs due to **Overfitting**

Watson, D. S., & Floridi, L. (2020). The explanation game: A formal framework for interpretable machine learning. Synthese. https://doi.org/10.1007/s11229-020-02629-9

Science in Data Driven Paradigm

Paradigm	Nature	Form	When
First	Experimental	Empiricism; describing natural phenomena	pre-Renaissance
Second	Theoretical	Modelling and generalization	pre-computers
Third	Computational	Simulation of complex phenomena	pre-Big Data
Fourth	Exploratory	Data-intensive; statistical exploration and data mining	Now

Scientific paradigms taken from Kitchin (2014, p. 3), compiled from Hey et al. (2009)

I. Wigner, E. P. (1960).The unreasonable effectiveness of mathematics in the natural sciences.RichardCourant lecture in mathematical sciences delivered at New York University, May 11, 1959. Communications on Pure and Applied Mathematics, 13(1), 1–14. https://doi.org/10.1002/cpa.3160130102

II. A. Halevy, P. Norvig and F. Pereira, "The Unreasonable Effectiveness of Data," in IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, March-April 2009, doi: 10.1109/MIS.2009.36.

Wegner: «Interactions cannot be reduced to algorithms»

Computer Science follows

MODELLED INPUT + **INTERACTION** + ALGORITHM = MODELLED OUTPUT

Data Science follows

FLAT INPUT + FLAT OUTPUT = ALGORITHM

Peter Wegner. 1997. Why interaction is more powerful than algorithms. Commun. ACM 40, 5 (May 1997), 80–91. https://doi.org/10.1145/253769.253801

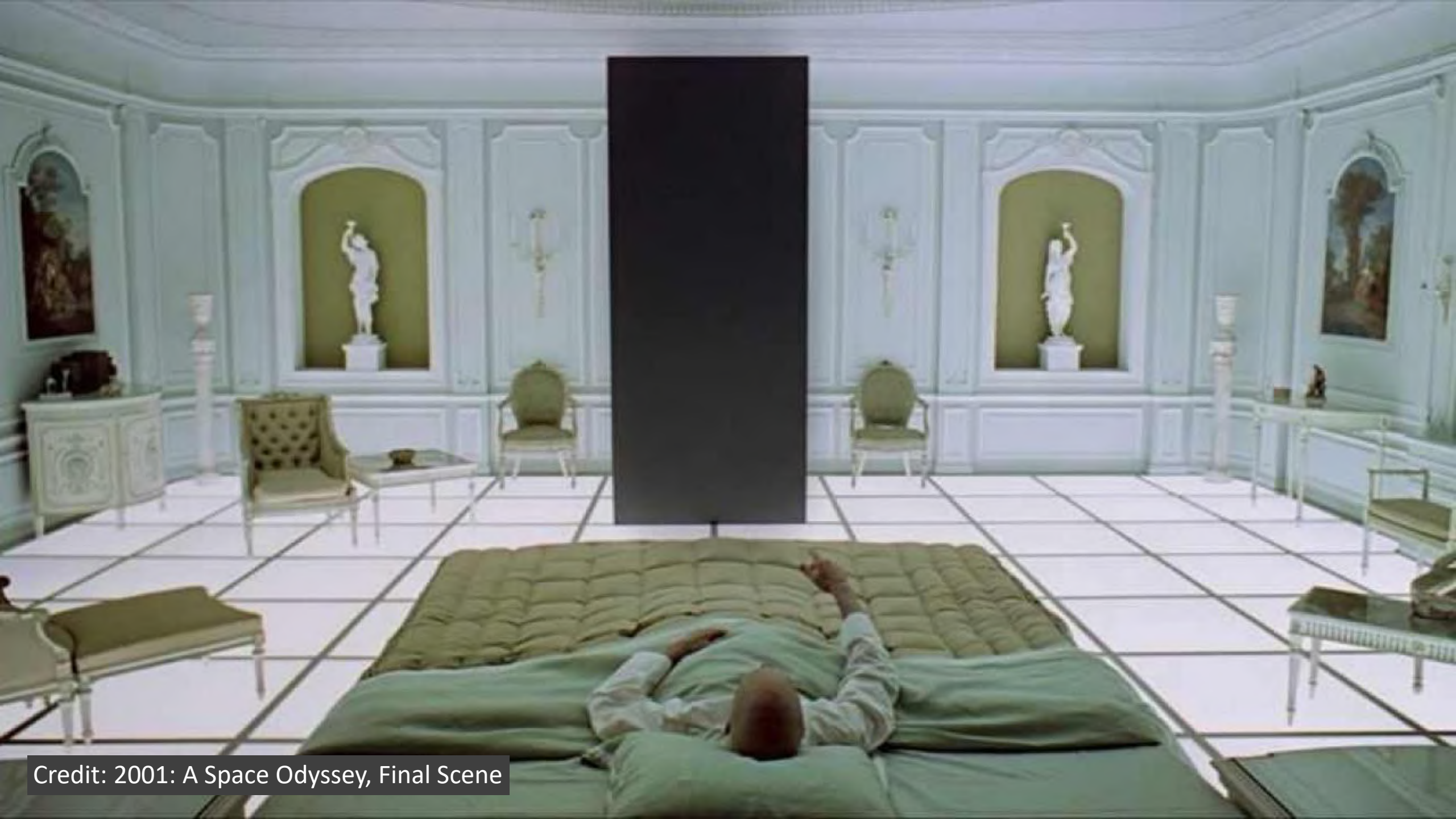
Computing Competencies for Undergraduate Data Science Curricula

ACM Data Science Task Force

January 2021

Andrea Danyluk, Co-chair
Paul Leidig, Co-chair

CONTENTS



Credit: 2001: A Space Odyssey, Final Scene