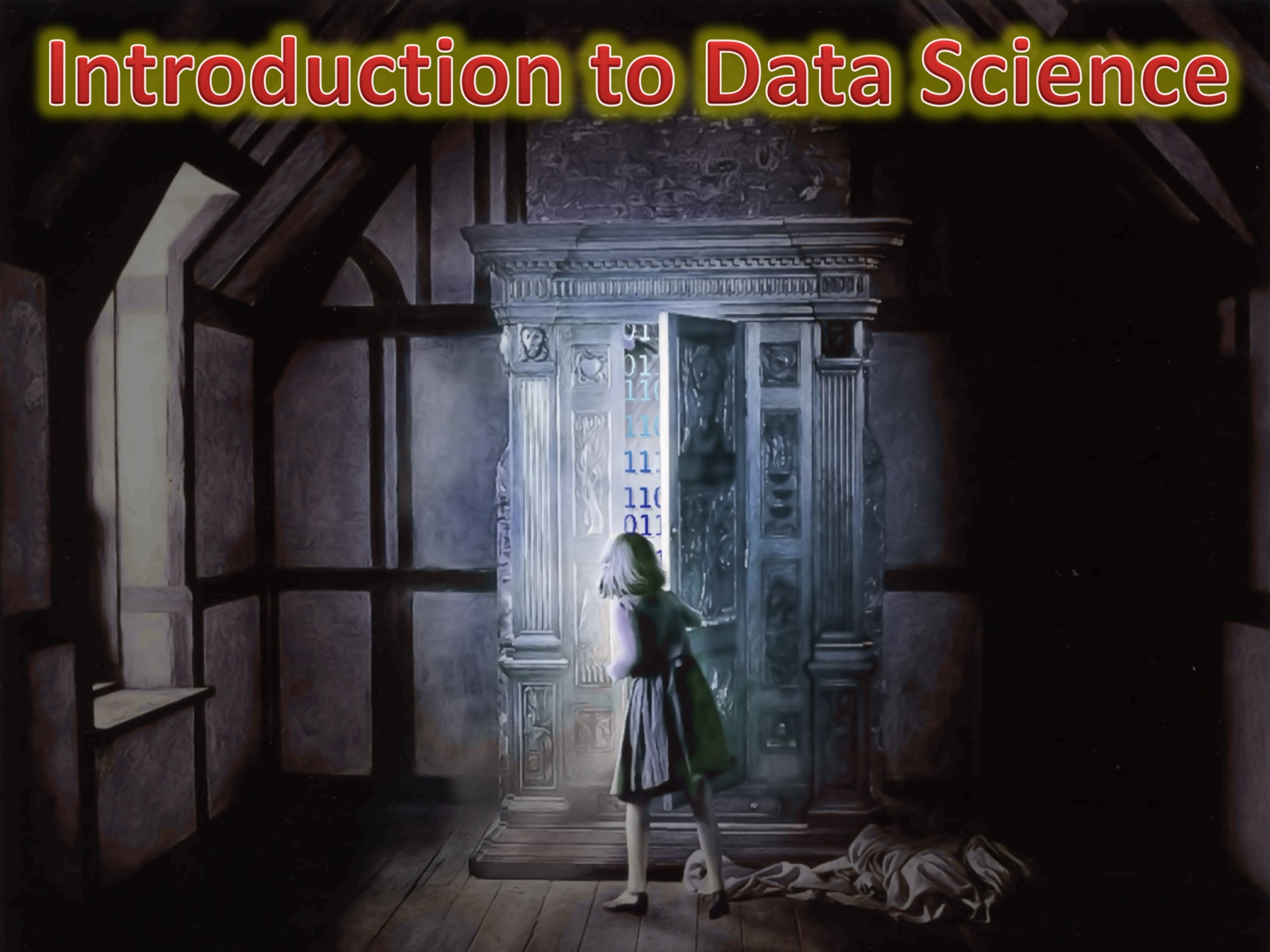
Introduction to Data Science



The program today

- Administrative
  - Course logistics
  - A reading and discussion of the sittyba

- Content
  - What is Data Science?
  - Why is it so special (particularly now)?

The teaching staff

- Instructors
  - Pascal Wallisch, PhD
  - Milan Bradonjic, PhD

- Teaching Assistants
  - Jianyu Zhang
  - Aditya Chawla
  - Anurag Rathore
  - Sumanyu Muku

- Graders
  - Aakash Bhattacharya
  - Avinav Goel
  - Anurag Rathore
  - Ayesha Ahmed

We will use the website on “Brightspace” as the LMS for this class

It features

- Announcements
- Lecture slides
- Datasets
- Code
- Assignments
- Assorted class materials (videos, etc.)

You can access it at

https://brightspace.nyu.edu/

The *sittyba*

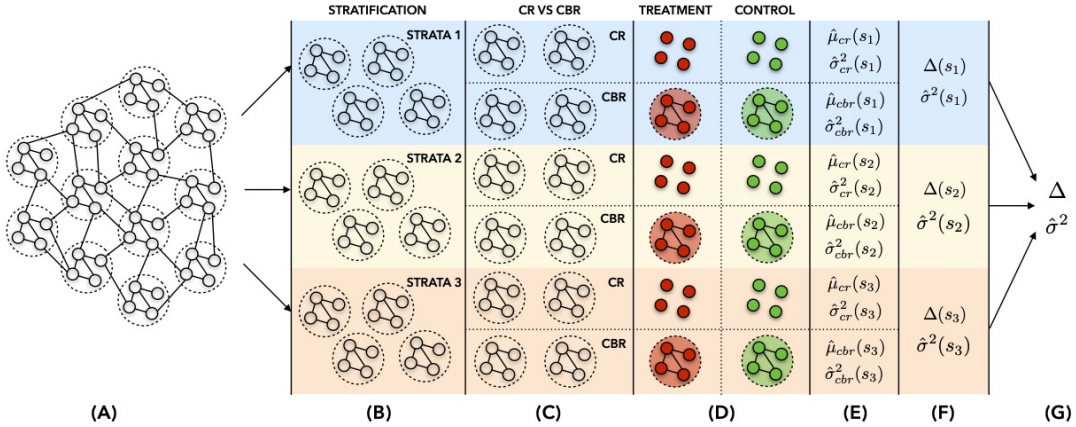
Anything else?

High level course goals

1. “Voltage regulator” – onboard people from many different backgrounds
2. Building strong foundations – planting seeds that will enable you to take more advanced classes
3. Kindling a burning passion for data
4. “Secret shelf”
5. [?]

The 3 principal parts of Data Science

Hypothesis Testing



Prediction

Machine Learning

What this class (and Data Science) is not

- A coding class, although we will be coding (a lot).
- A math class, although we will be deriving theorems.
- A statistics class, although we will do computations.

The difference?

In those fields, these activities are often an end in itself.

In data science, we will use concepts and techniques developed in those fields as a means to an end.

Towards what end?

Insights from *Data*

"The purpose of computation is insight, not numbers."

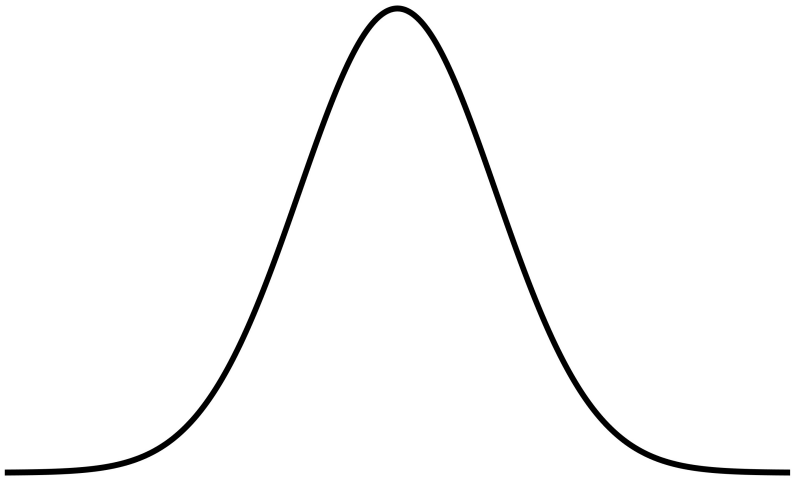
(Richard Hamming, 1962)

As a Data Scientist, the computer is your horse, you are the jockey

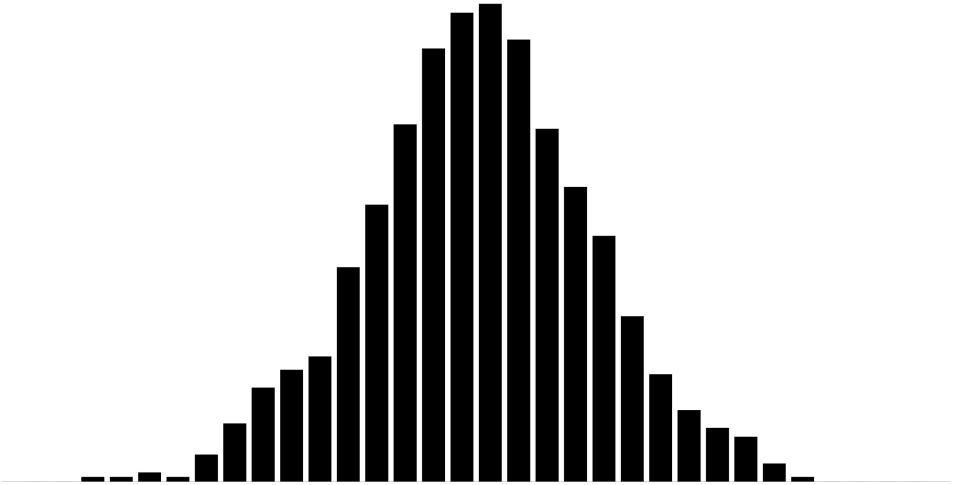
Make sure not to confuse the map with the territory, i.e. do not confuse your tools with what the job actually is (calling Python functions to process data is a means to an end, not an end in itself)

The fundamental difference between Mathematics and Data Science

Math is axiomatic
(Deductive)



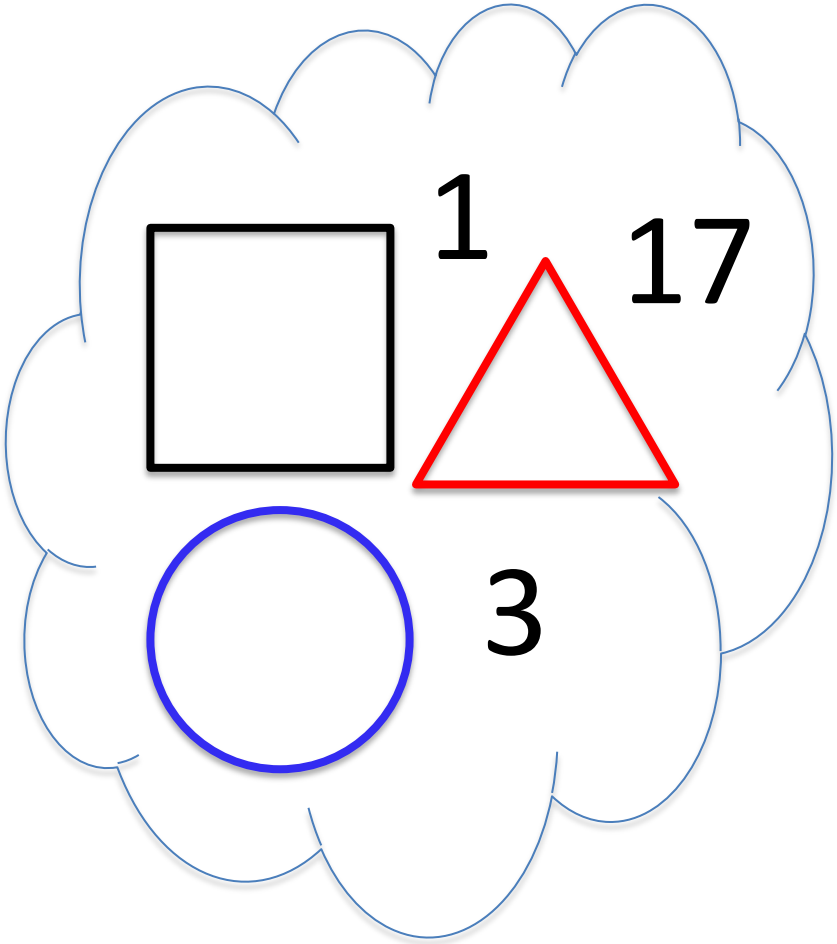
Science uses data
(Inductive)



What is data?

Data is very special. Coming up with the concept of data was a radical, paradoxical step

Mathematics



Simplicity, Beauty, Symmetry

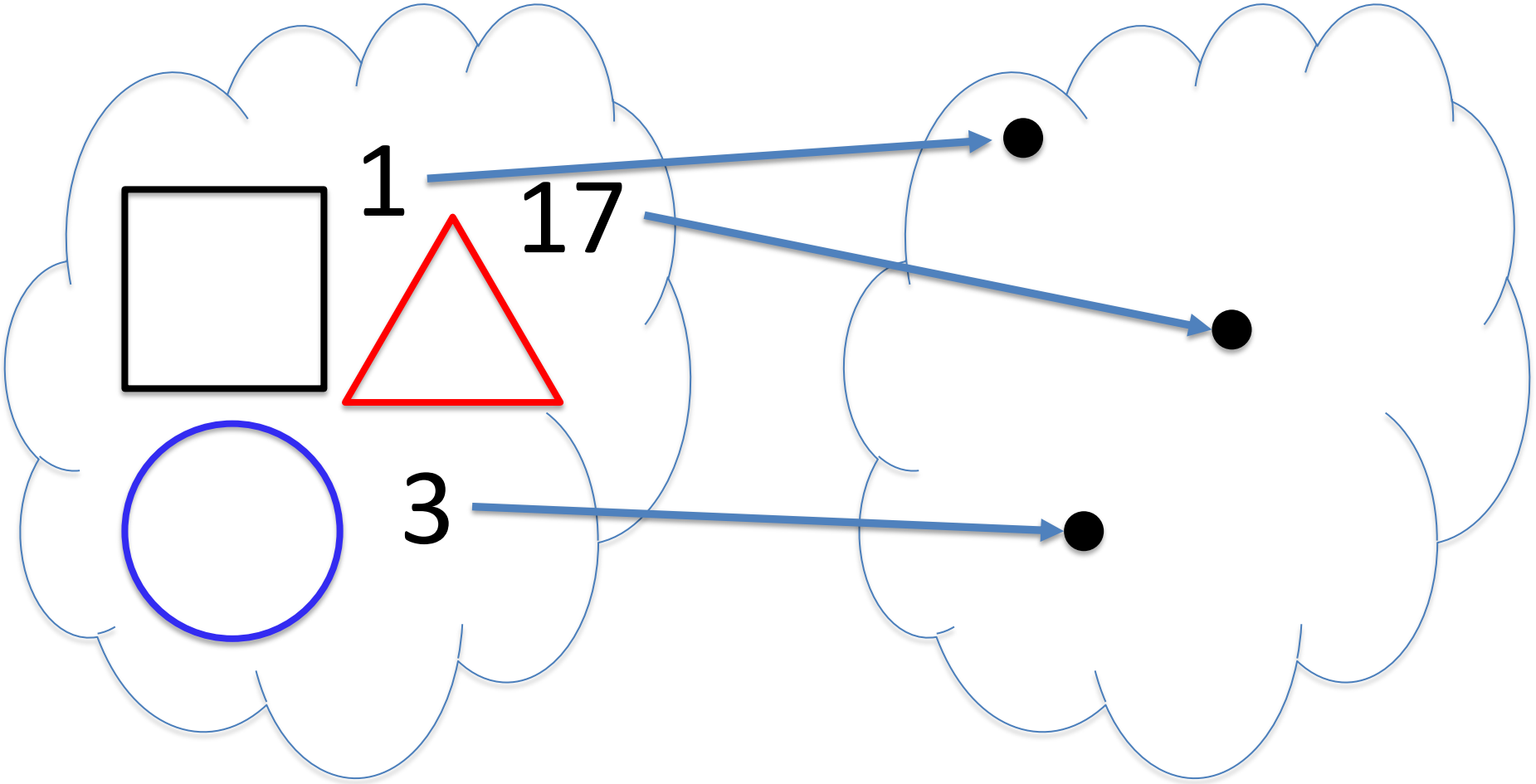
Qualitative descriptions of the natural world

Complex, messy, broken symmetry

Quantification = mapping between these realms

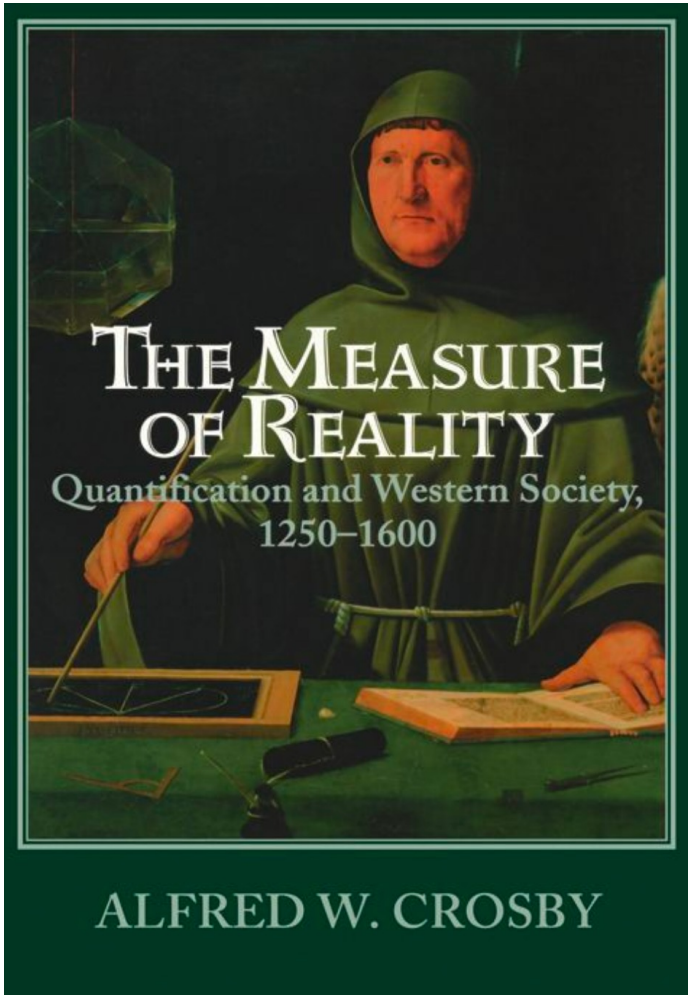
Mathematics

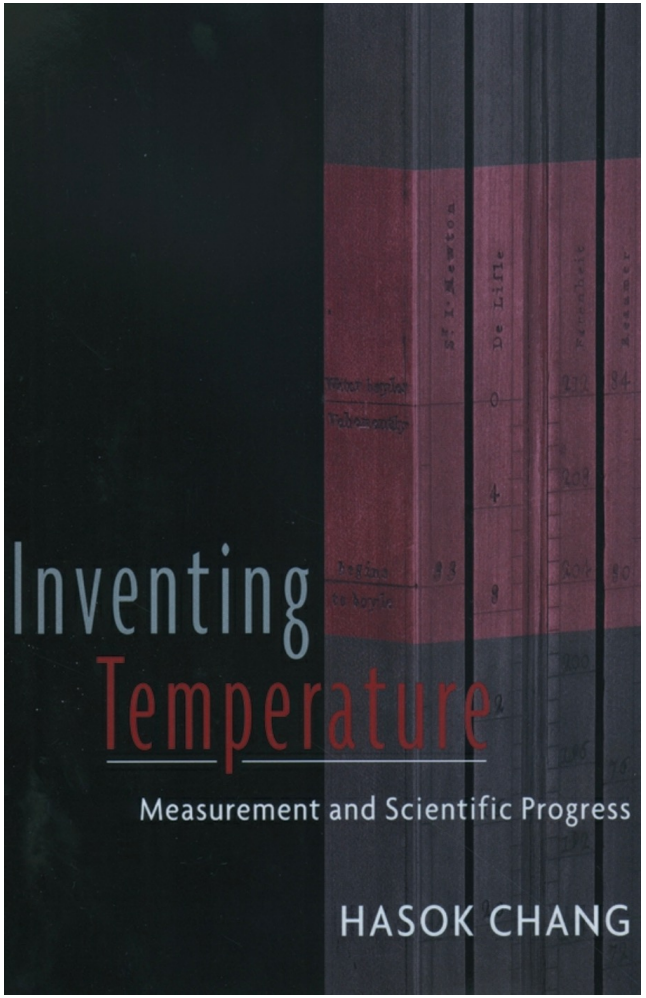
Descriptions of the natural world



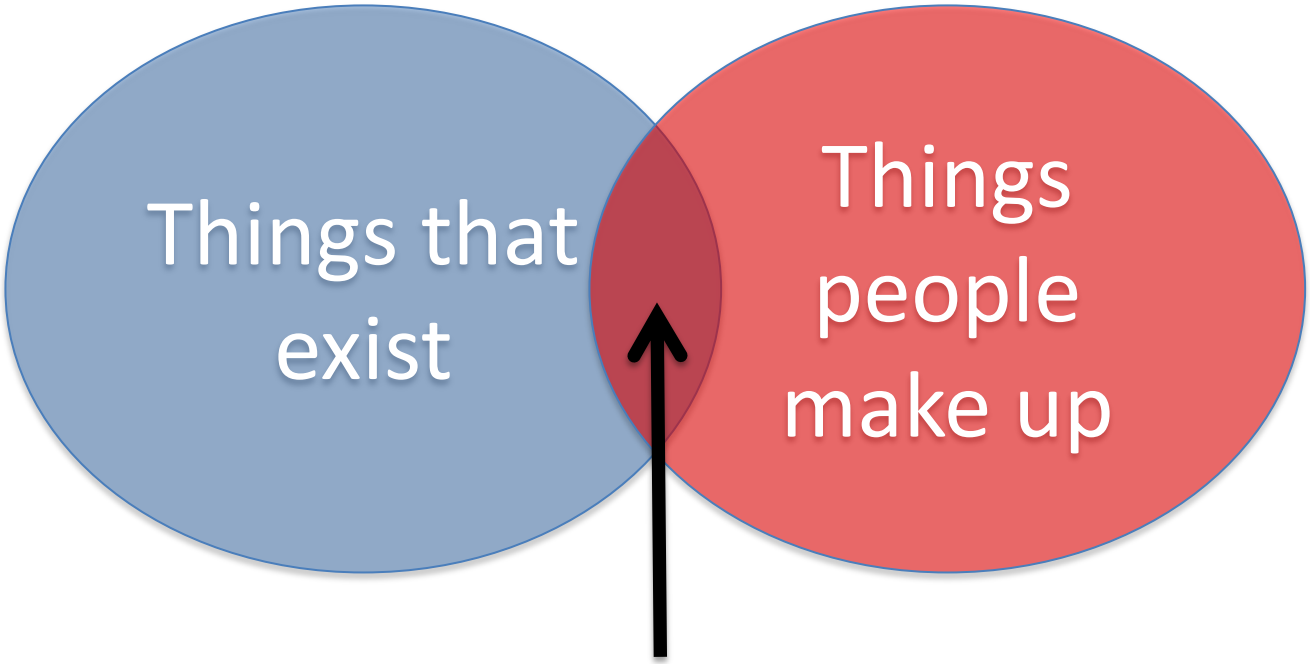
This was a radical step in the history of ideas

- Took until the 1250s to seriously consider the idea.
- Took another 500+ years to implement / realize.





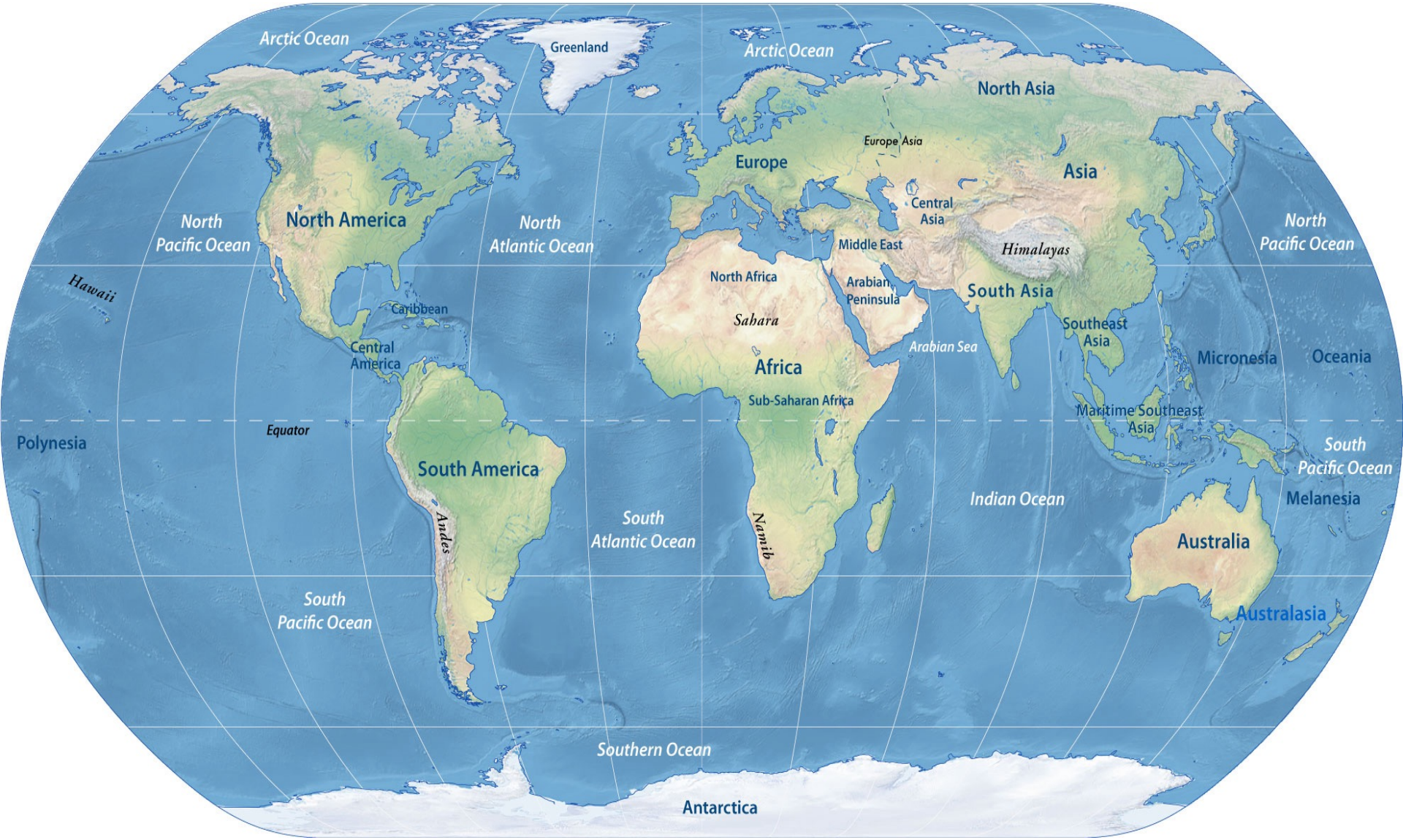
Data is truly special



Data is an epistemic interface on reality

- Data (Latin): “A given thing” (~1645) – intended to mean “quantitative facts”.
- There are no data in the ancient world. The concept of data is relatively modern.
- However, in science, it is *not* just given. Getting the data is usually the key part of science.
- However, in Data Science, we usually do presume that we already have the data.
- Data are an interface on reality that allows to represent its complexity, beyond preconceptions.

Data can transform our understanding, e.g. TO maps



Math vs. Machine Learning

$$
\begin{aligned}
Q = \; & 15 b_4 \beta^6 T_{cr} + 5 b_2 \beta^6 T_{cr} + 3 \alpha b_4 \beta^5 T_{cr} - 15 \alpha b_3 \beta^5 T_{cr} + \alpha b_2 \beta^5 T_{cr} \\
& - 15 \alpha b_1 \beta^5 T_{cr} - 22 {a_3}^2 \beta^5 T_{cr} - 7 a_2 a_3 \beta^5 T_{cr} - 14 a_1 a_3 \beta^5 T_{cr} - 3 {a_2}^2 \beta^5 T_{cr} \\
& - 7 a_1 a_2 \beta^5 T_{cr} - 4 {a_1}^2 \beta^5 T_{cr} - 12 \alpha^2 b_4 \beta^4 T_{cr} - 3 \alpha^2 b_3 \beta^4 T_{cr} + 6 \alpha^2 b_2 \beta^4 T_{cr} \\
& - 3 \alpha^2 b_1 \beta^4 T_{cr} + 12 {a_3}^2 \alpha \beta^4 T_{cr} + 37 a_2 a_3 \alpha \beta^4 T_{cr} + 30 a_1 a_3 \alpha \beta^4 T_{cr} \\
& + 7 {a_2}^2 \alpha \beta^4 T_{cr} + 19 a_1 a_2 \alpha \beta^4 T_{cr} + 18 {a_1}^2 \alpha \beta^4 T_{cr} + 12 \alpha^3 b_3 \beta^3 T_{cr} \\
& + 2 \alpha^3 b_2 \beta^3 T_{cr} + 12 \alpha^3 b_1 \beta^3 T_{cr} + 4 {a_3}^2 \alpha^2 \beta^3 T_{cr} - 20 a_2 a_3 \alpha^2 \beta^3 T_{cr} \\
& - 16 a_1 a_3 \alpha^2 \beta^3 T_{cr} - 12 {a_2}^2 \alpha^2 \beta^3 T_{cr} - 26 a_1 a_2 \alpha^2 \beta^3 T_{cr} - 8 {a_1}^2 \alpha^2 \beta^3 T_{cr} \\
& - 8 \alpha^4 b_2 \beta^2 T_{cr} - 4 a_2 a_3 \alpha^3 \beta^2 T_{cr} + 8 {a_2}^2 \alpha^3 \beta^2 T_{cr} + 8 a_1 a_2 \alpha^3 \beta^2 T_{cr} \\
& + 5 b_3 \beta^5 + 15 b_1 \beta^5 - 15 \alpha b_4 \beta^4 + \alpha b_3 \beta^4 - 15 \alpha b_2 \beta^4 + 3 \alpha b_1 \beta^4 - 4 {a_3}^2 \beta^4 \\
& - 9 a_2 a_3 \beta^4 - 18 a_1 a_3 \beta^4 - {a_2}^2 \beta^4 - 9 a_1 a_2 \beta^4 - 18 {a_1}^2 \beta^4 - 3 \alpha^2 b_4 \beta^3 \\
& + 6 \alpha^2 b_3 \beta^3 - 3 \alpha^2 b_2 \beta^3 - 12 \alpha^2 b_1 \beta^3 + 26 {a_3}^2 \alpha \beta^3 + 19 a_2 a_3 \alpha \beta^3 \\
& + 30 a_1 a_3 \alpha \beta^3 + 11 {a_2}^2 \alpha \beta^3 + 33 a_1 a_2 \alpha \beta^3 + 12 {a_1}^2 \alpha \beta^3 + 12 \alpha^3 b_4 \beta^2 \\
& + 2 \alpha^3 b_3 \beta^2 + 12 \alpha^3 b_2 \beta^2 - 8 {a_3}^2 \alpha^2 \beta^2 - 32 a_2 a_3 \alpha^2 \beta^2 - 12 a_1 a_3 \alpha^2 \beta^2 \\
& - 14 {a_2}^2 \alpha^2 \beta^2 - 18 a_1 a_2 \alpha^2 \beta^2 - 8 \alpha^4 b_3 \beta - 8 {a_3}^2 \alpha^3 \beta + 8 a_2 a_3 \alpha^3 \beta \\
& + 4 {a_2}^2 \alpha^3 \beta + 8 a_2 a_3 \alpha^4 \qquad (3.34)
\end{aligned}
$$

Simple learning rule

+

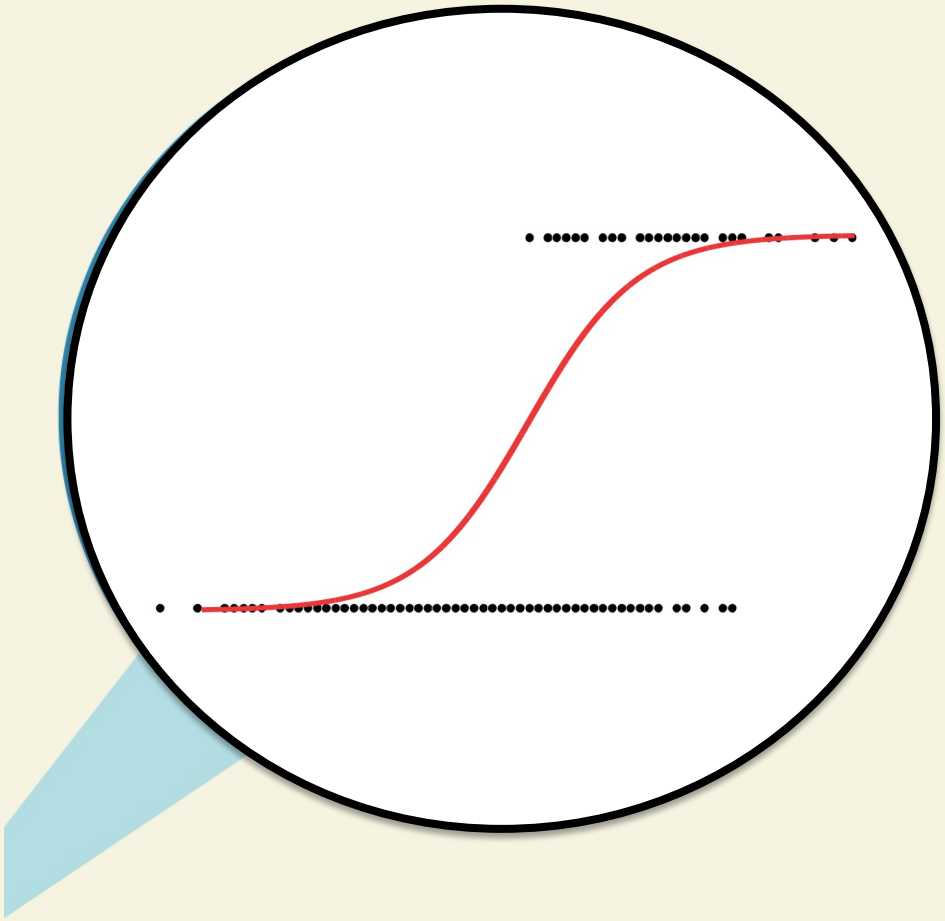
lots of data

+

fast computer

(lots of iterations)

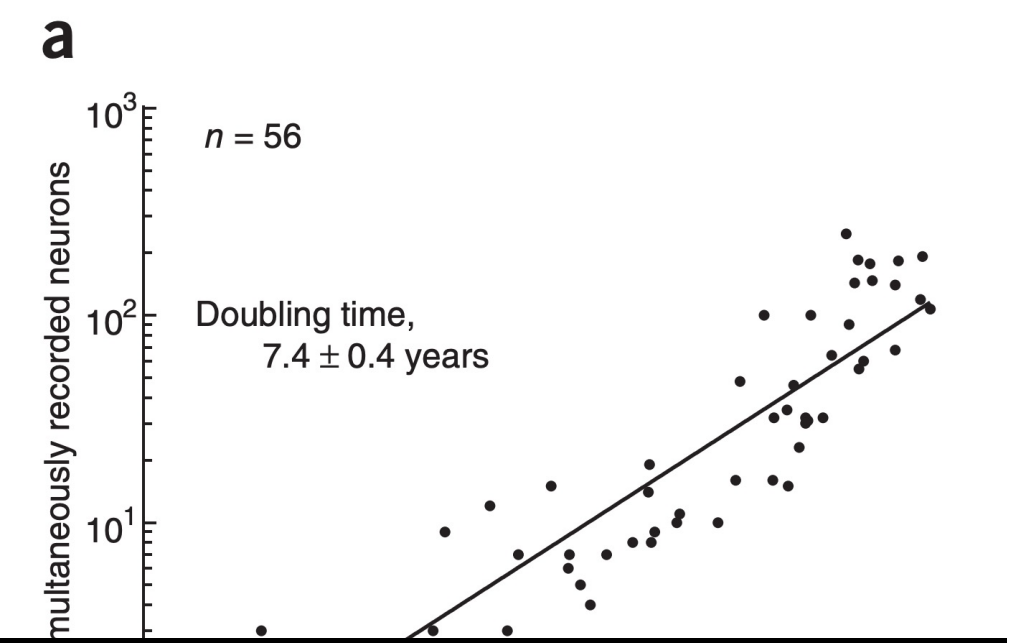
Rand (2011) – Differential Delay Equations

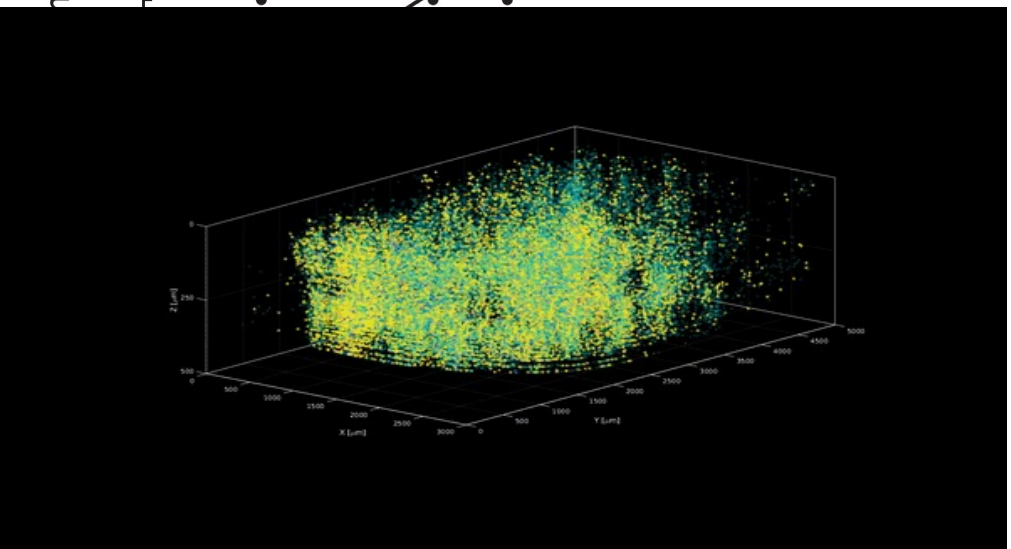


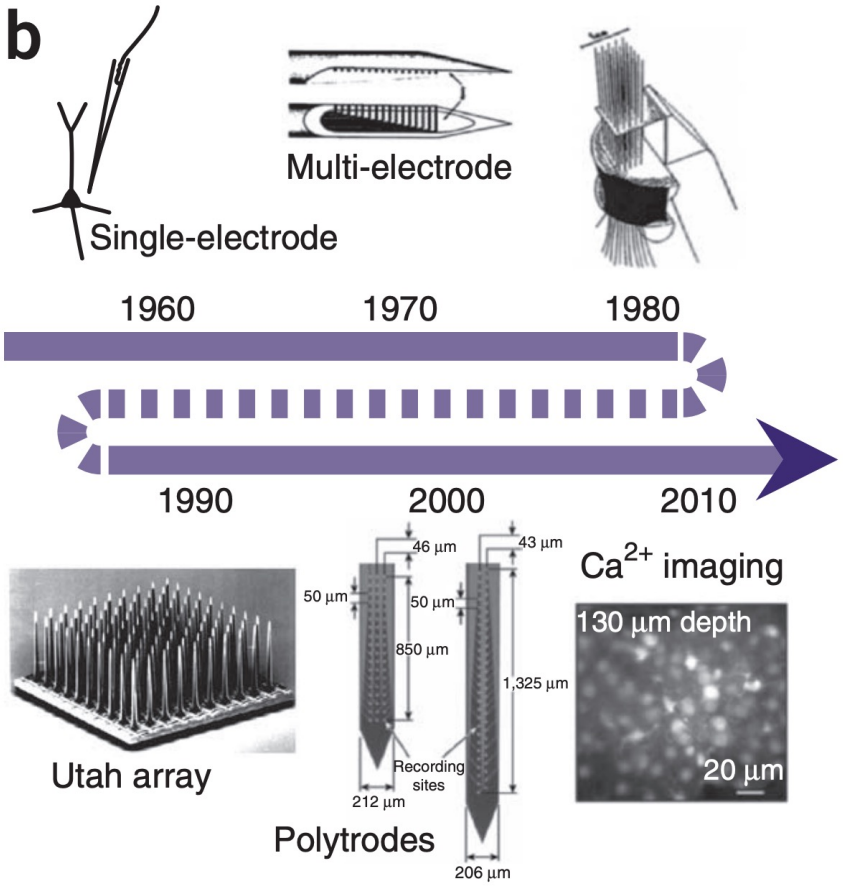
The computer is *our* microscope – on data

The age of Data (Science)

Data is increasing exponentially (due to exponentially increasing recording capabilities), as described by Stevenson's law







Vaziri et al. (2021)

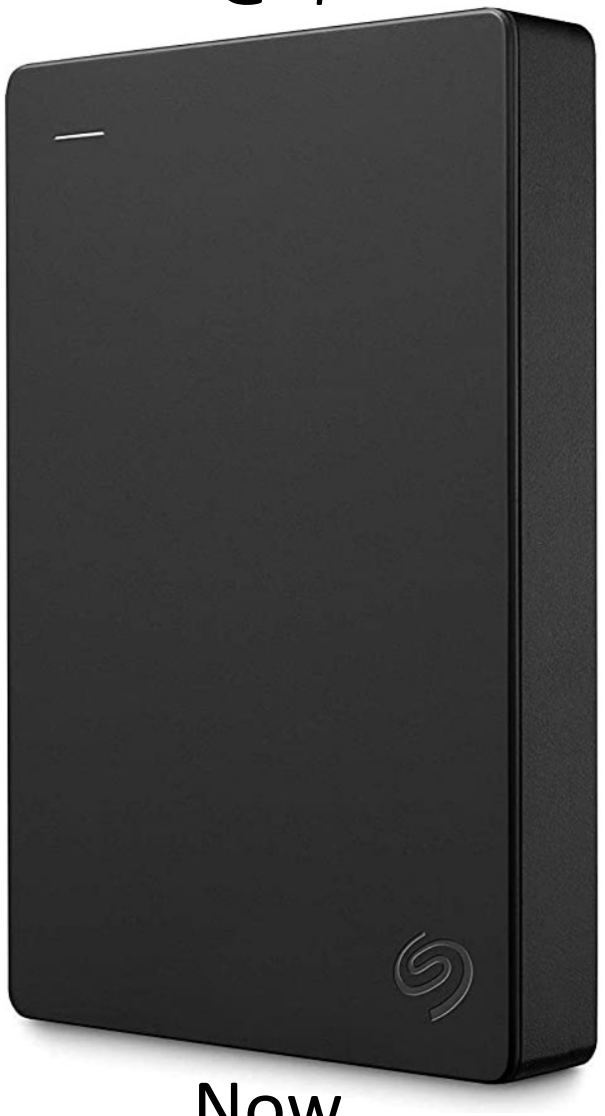
Stevenson & Kording (2011)

We are also able to store all this data

5 MB @ $30k/m

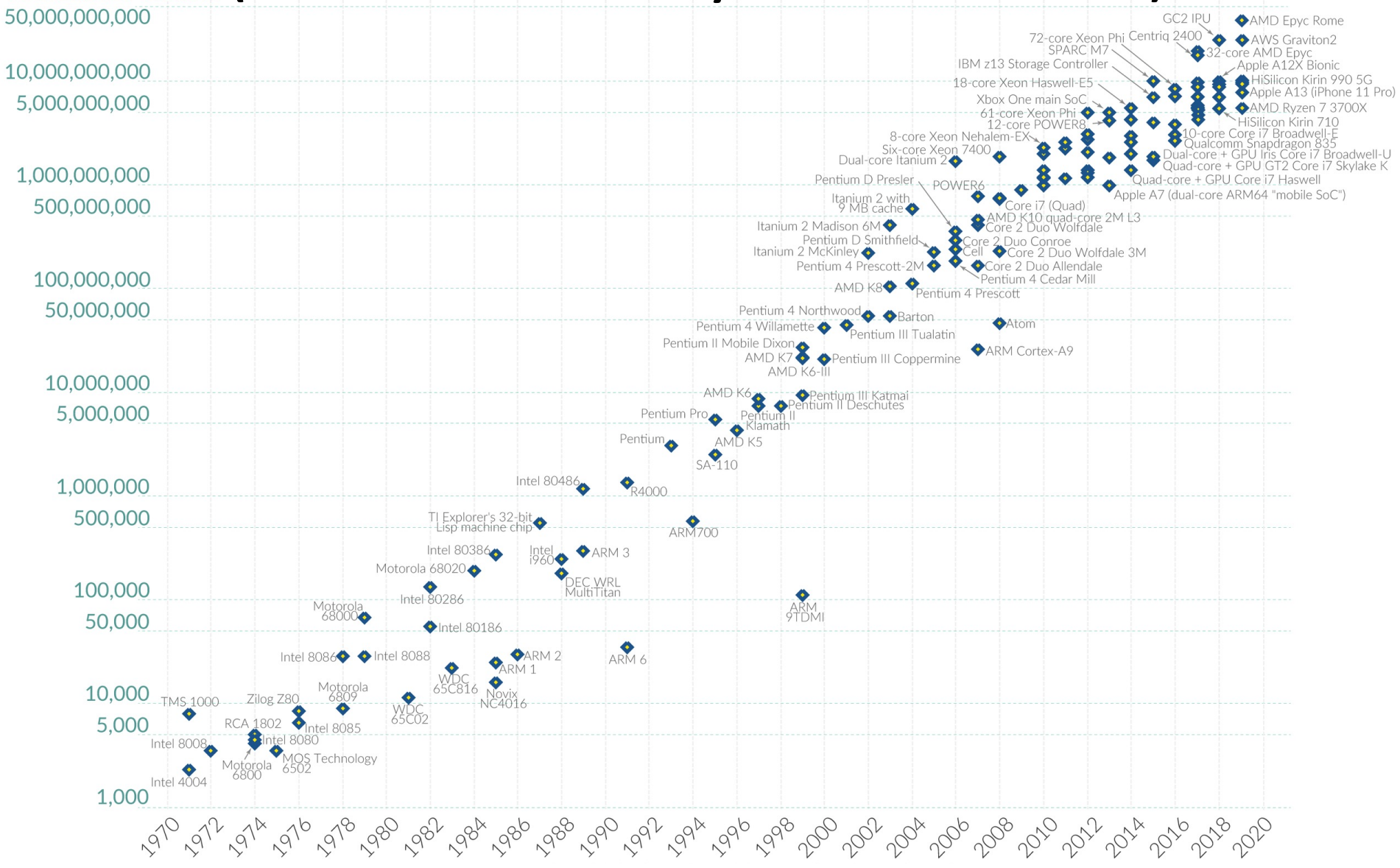
1956

5 TB @ $300



Now

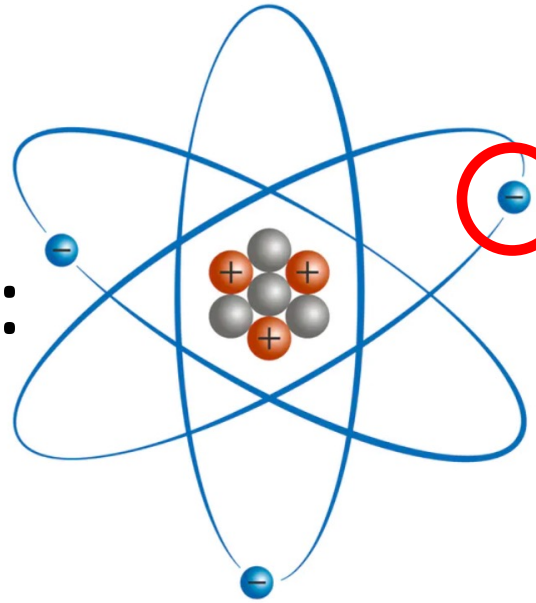
Data processing capabilities also increase exponentially (as described by Moore's law)



This is timely because a dire need has arisen

- Many scientific fields (economics, neuroscience, psychology, microbiome, nutrition, epigenetics, pharmacology, etc.) have run into a wall of diplexity.
- **Diplexity**: Fundamental, irreducible, inherently DIverse comPLEXITY.
- This renders most traditional data analysis approaches (importantly all that assume **ergodicity**) inappropriate or misleading.
- Impeding further progress in these fields.
- Luckily, there is salvation in novel (multivariate) big data analytics.

Physics:



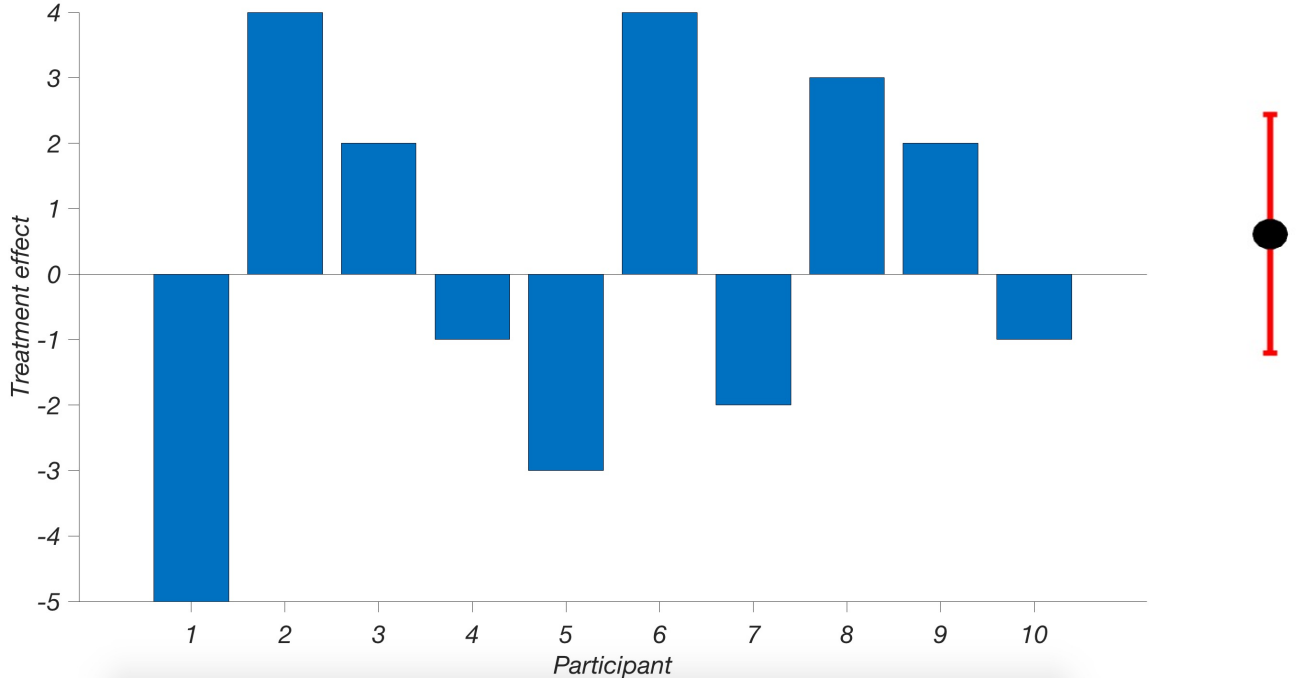
Electron(s)
Mass: 9.10938356 × 10-31 kg
Charge: -1.60217662 × 10-19 C
Spin: ½

Number: At least 10^{80}

They are all identical.

			Diplexity			

Microbiome:



The critical need for data-based decision making was highlighted by the novel Coronavirus pandemic

It also illustrated how narrative based approaches are insufficient and outright dangerous

BuzzFeed News

Bourbon Street Biden Endorsement US Airport

SCIENCE / CORONAVIRUS

Don't Worry About The Coronavirus. Worry About The Flu.

The first new disease outb
media era has been define
spread of panic and uncer
you should worry about —
shouldn't.

Dan Vergano
BuzzFeed News Reporter

Reporting From
Washington, D.C.

Updated on January 30, 2020, at
Posted on January 28, 2020, at 12

The Washington Post
@washingtonpost

Get a grippe, America. The flu i
much bigger threat than coron
for now.

Get a grippe, America. The flu is a much bigger threat than coronavirus, for now.
washingtonpost.com

13:35 · 2/1/20 · SocialFlow

Alison Buttenheim @abuttenheim · Feb 14

Novelty, availability, dread, ambiguity: #COVD19 has all the right ingredients for out-of-scale freakout. #BehavioralScience #behavioraleconomics

Coronavirus 'Hits All the Hot Buttons' for How We Misjudge Risk

Coronavirus 'Hits All the Hot Buttons' for How We Misjudge Risk
Psychologists say that differing responses to coronavirus and the flu illustrate our shortcomings when it comes to evaluating danger.
nytimes.com

Bloomberg Opinion

Technology & Ideas

The Cognitive Bias That Makes Us Panic About Coronavirus

Feeling anxious? Blame "probability neglect."

By Cass R. Sunstein
February 28, 2020, 1:09 PM EST

Why it matters whether you could read a graph in February 2020

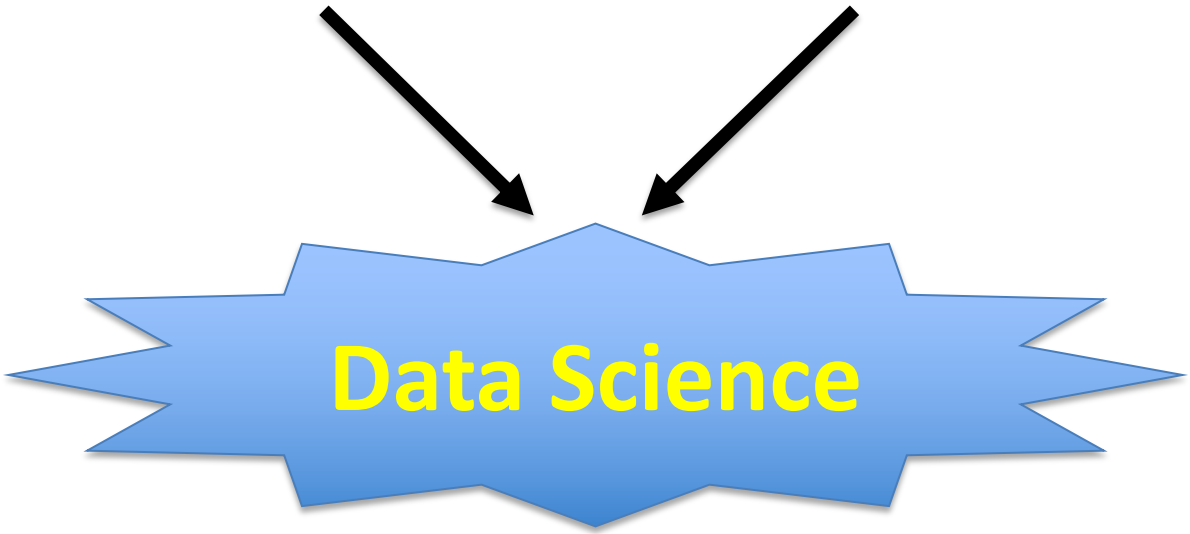
Data Science fills this niche

Affordances:

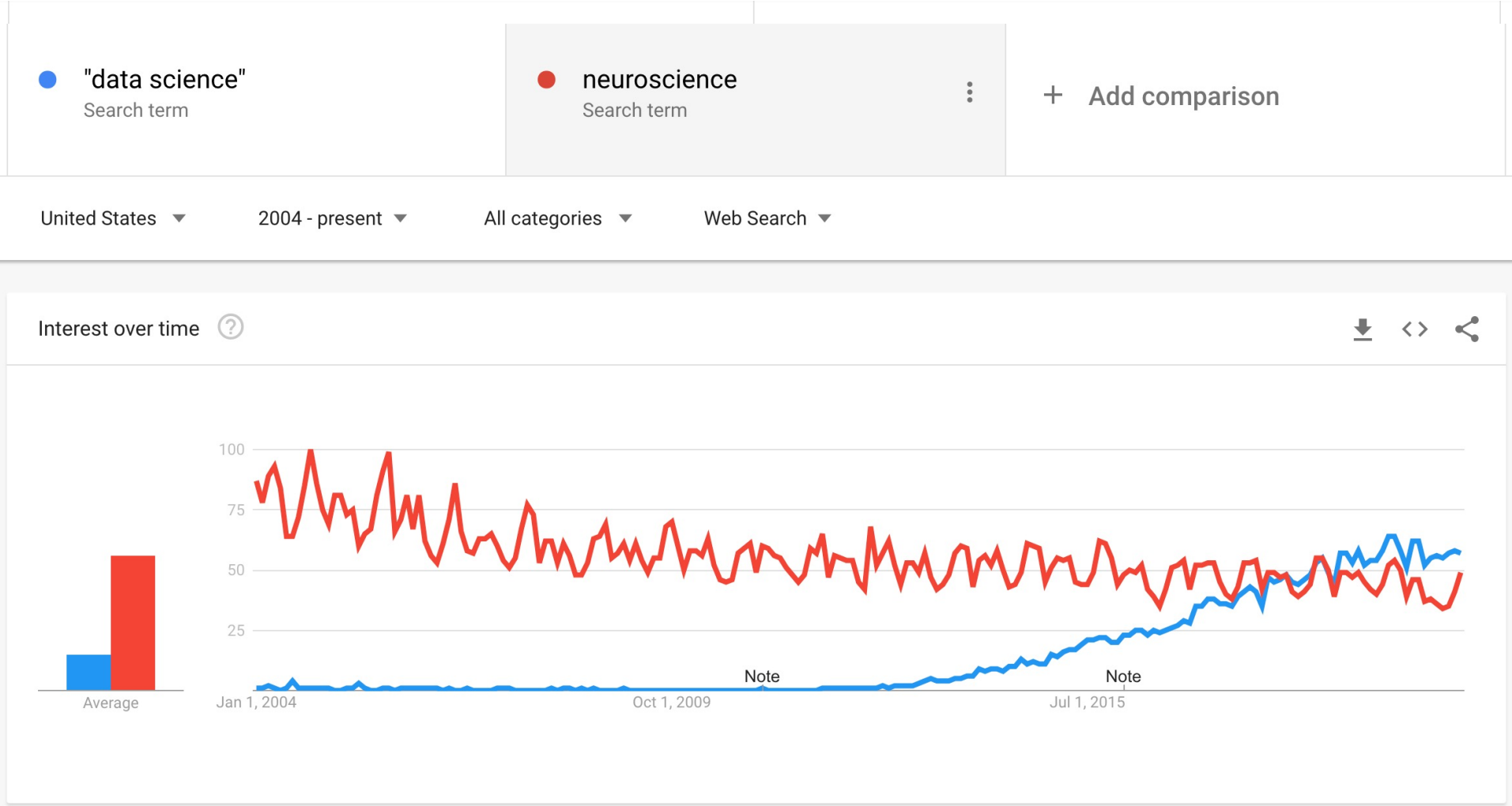
-Big Data
-Physical capabilities to store and process Big Data

Needs:

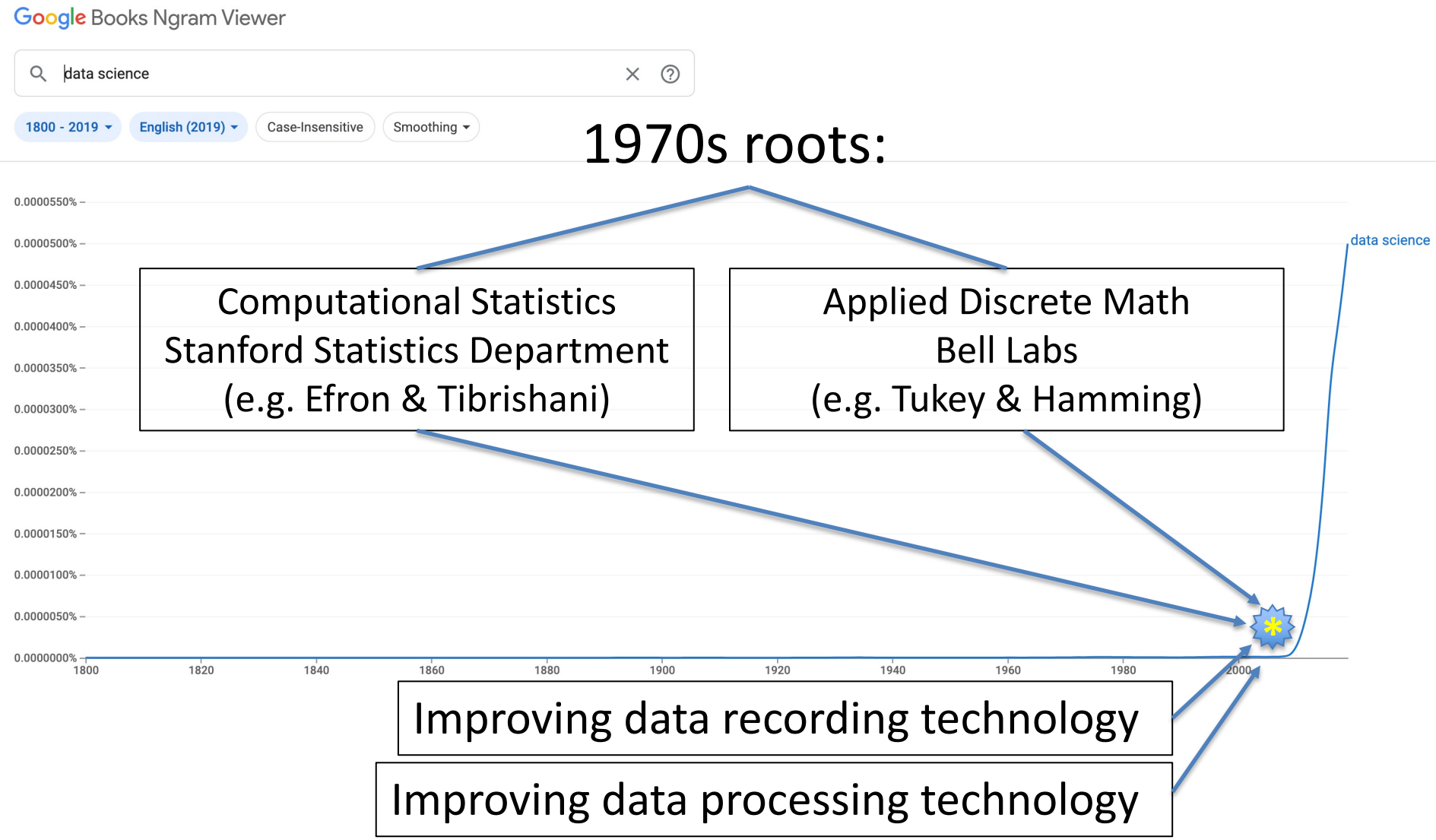
-Scientific need to handle diplexity
-Societal need to make Decisions based on data



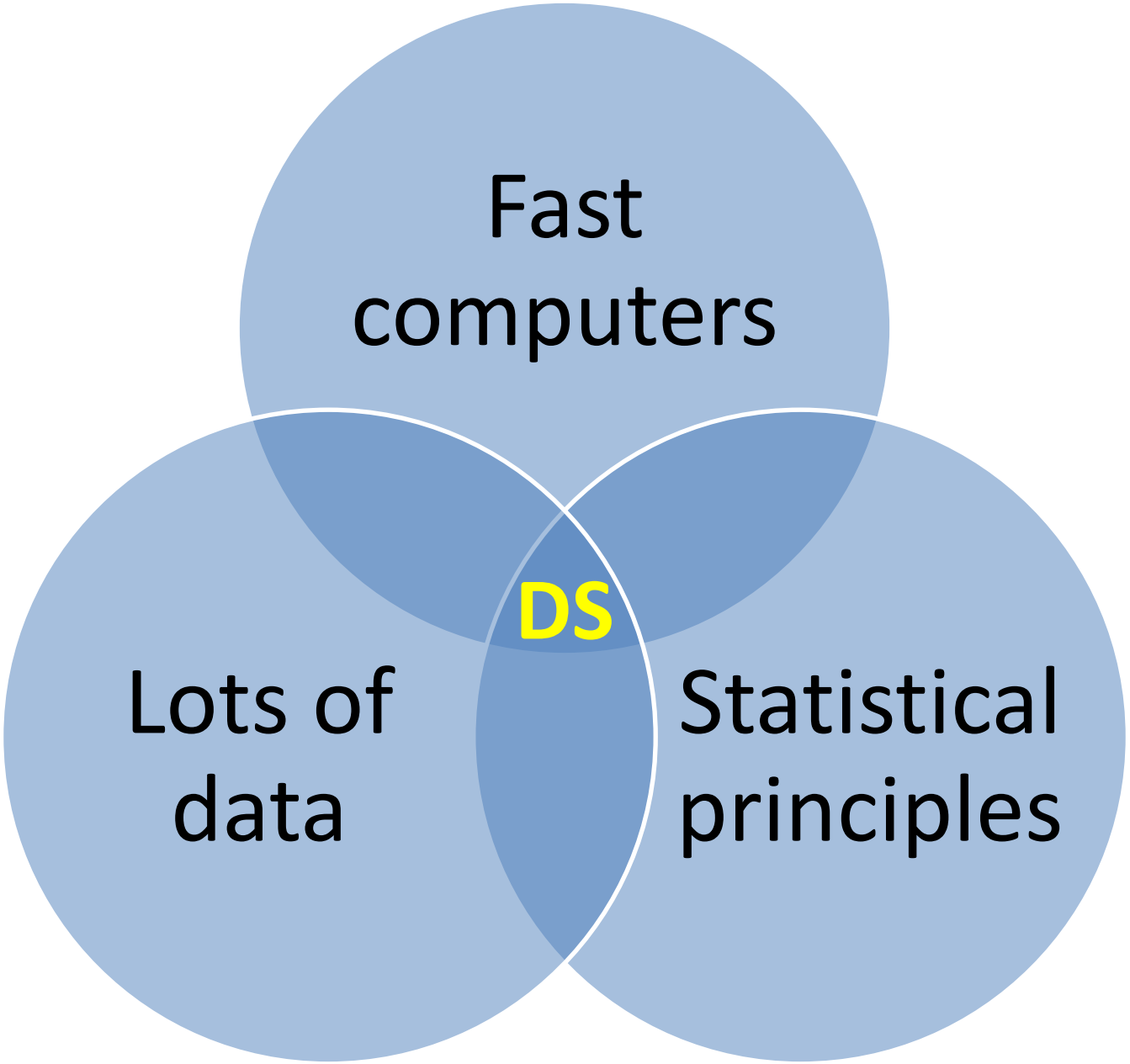
Data science as a relevant and independent field is less than a decade old



A genuinely 21st century enterprise



So what is Data Science?



The 10 most valuable companies in the world (by market capitalization)

Rank	2008
1	**Exxon Mobil**
2	**PetroChina**
3	**Gazprom**
4	General Electric
5	Microsoft
6	**Petrobras**
7	China Mobile
8	**Royal Dutch Shell**
9	ICBC
10	Walmart

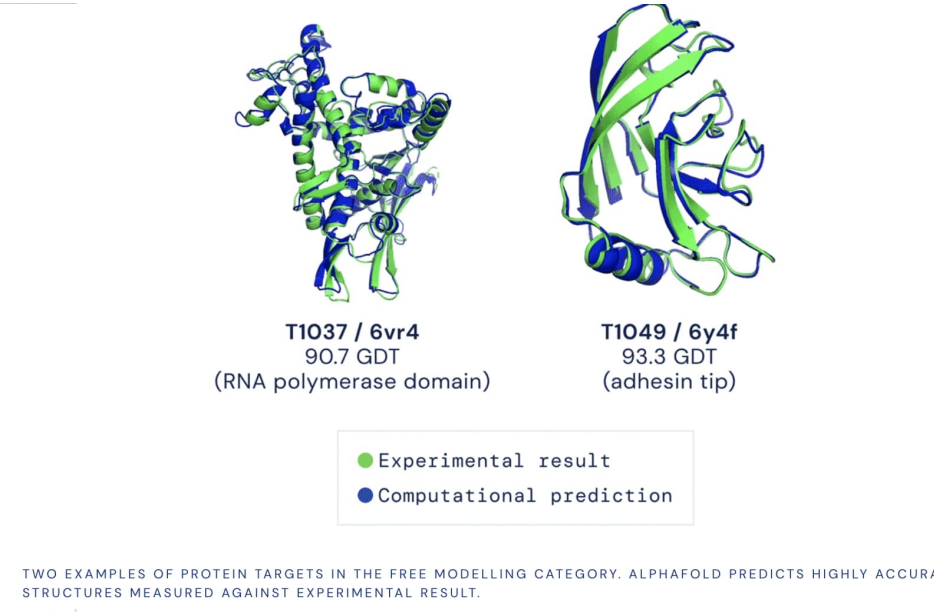
We are not the first to recognize this

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

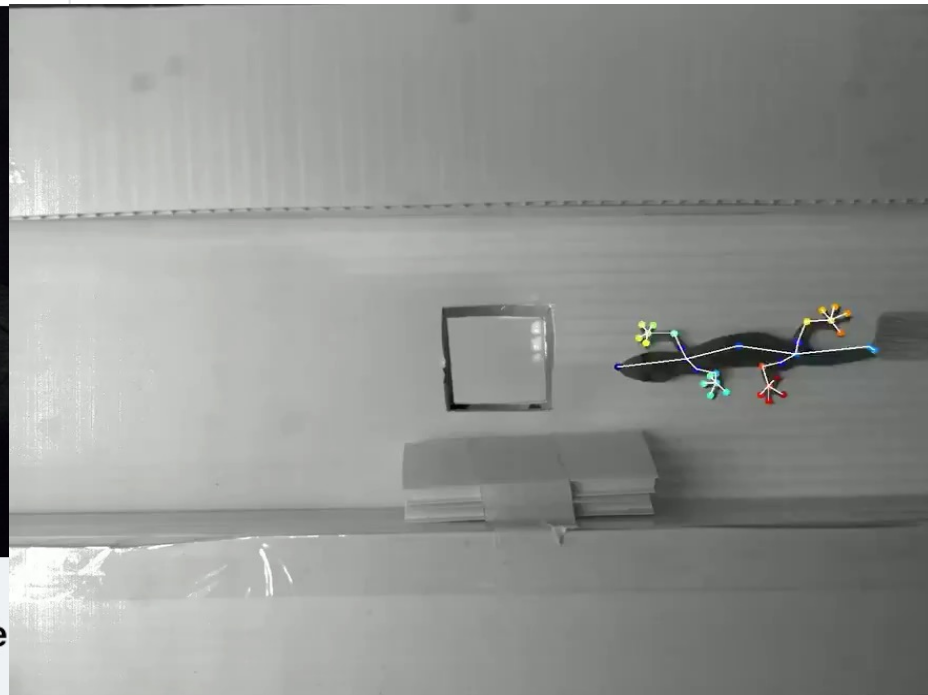
Data Science is already changing the world



TWO EXAMPLES OF PROTEIN TARGETS IN THE FREE MODELLING CATEGORY. ALPHAFOLD PREDICTS HIGHLY ACCURA
STRUCTURES MEASURED AGAINST EXPERIMENTAL RESULT.

EXPLICA.CO

Movies : Why do the specialized critics and the audience almost never coincide?




So… once again:
Welcome aboard

That’s it – for now

General questions?

Discussion?