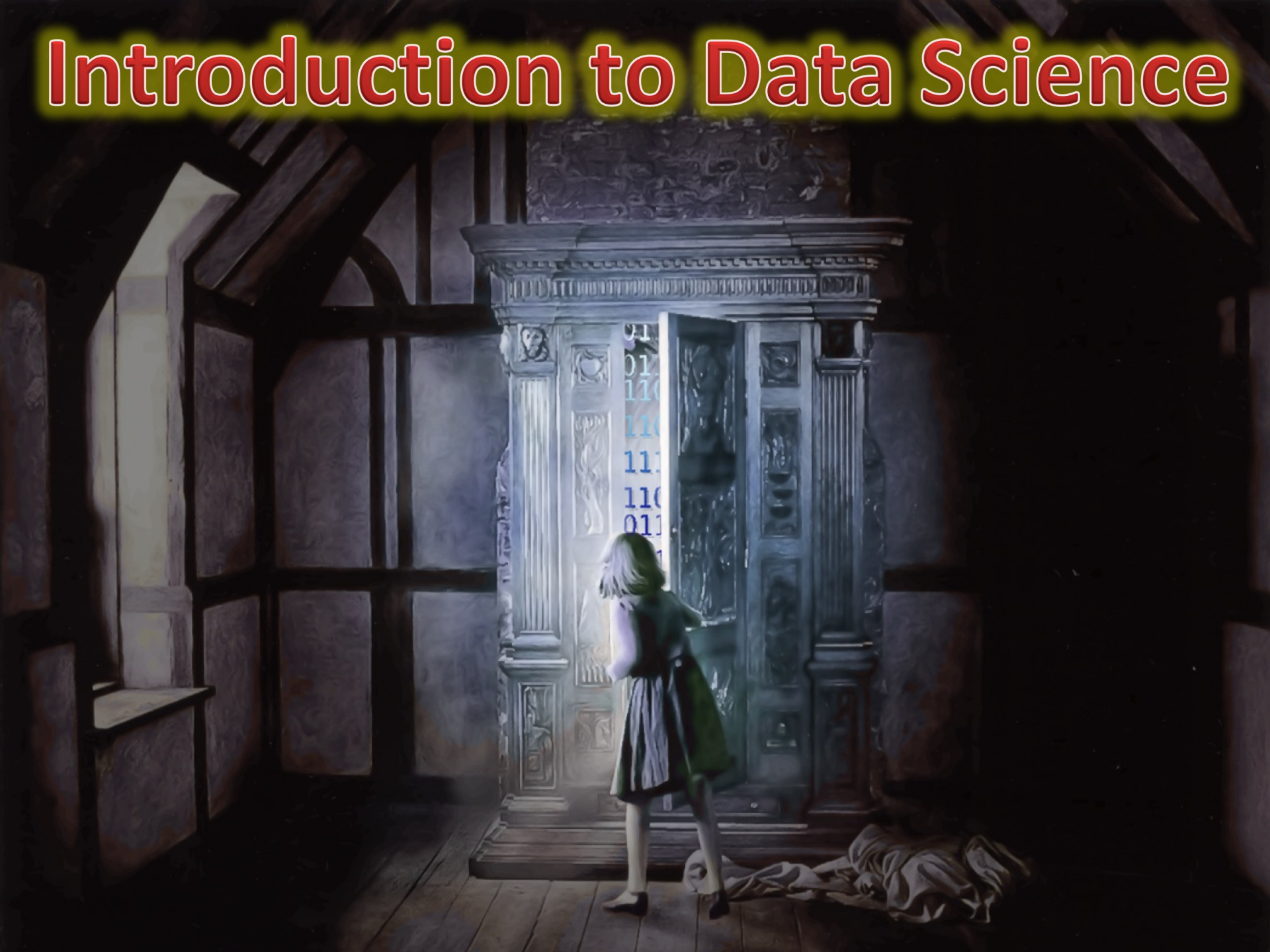
Smallest font

Please turn off and put away your cell phone

Calibration slide

Smallest font

Introduction to Data Science



The program today

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

- Content
  - Why do we need a class like this?
  - What is Data Science?

The teaching staff

Instructor

Pascal Wallisch, PhD

Section leaders

Ansh Riyal

Aman Singhal

Stephen Spivack

Tutor

Avinav Goel

Graders

Anonymous 1

Anonymous 2

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*

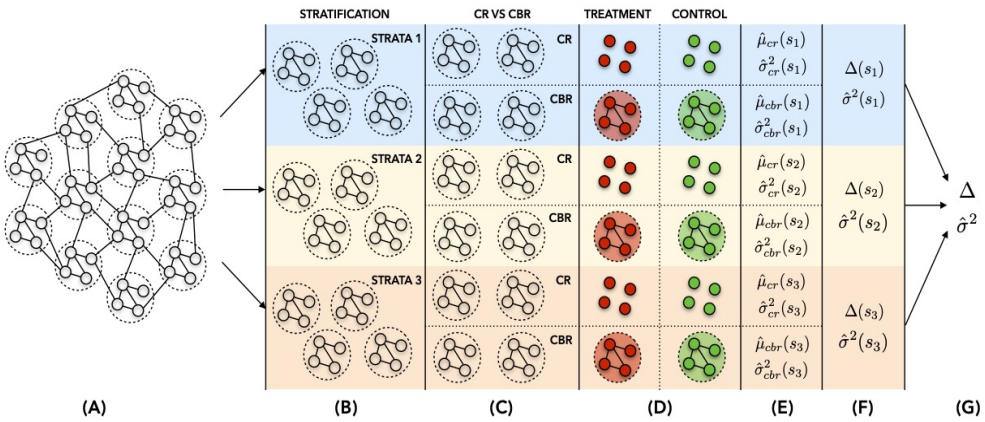
Anything else?

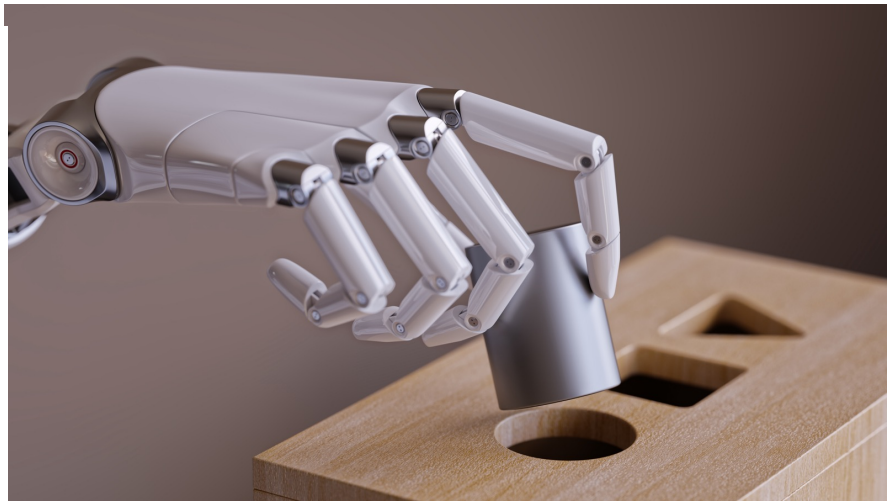
Why these two theme blocks?

The 2 principal domains of Data Science:

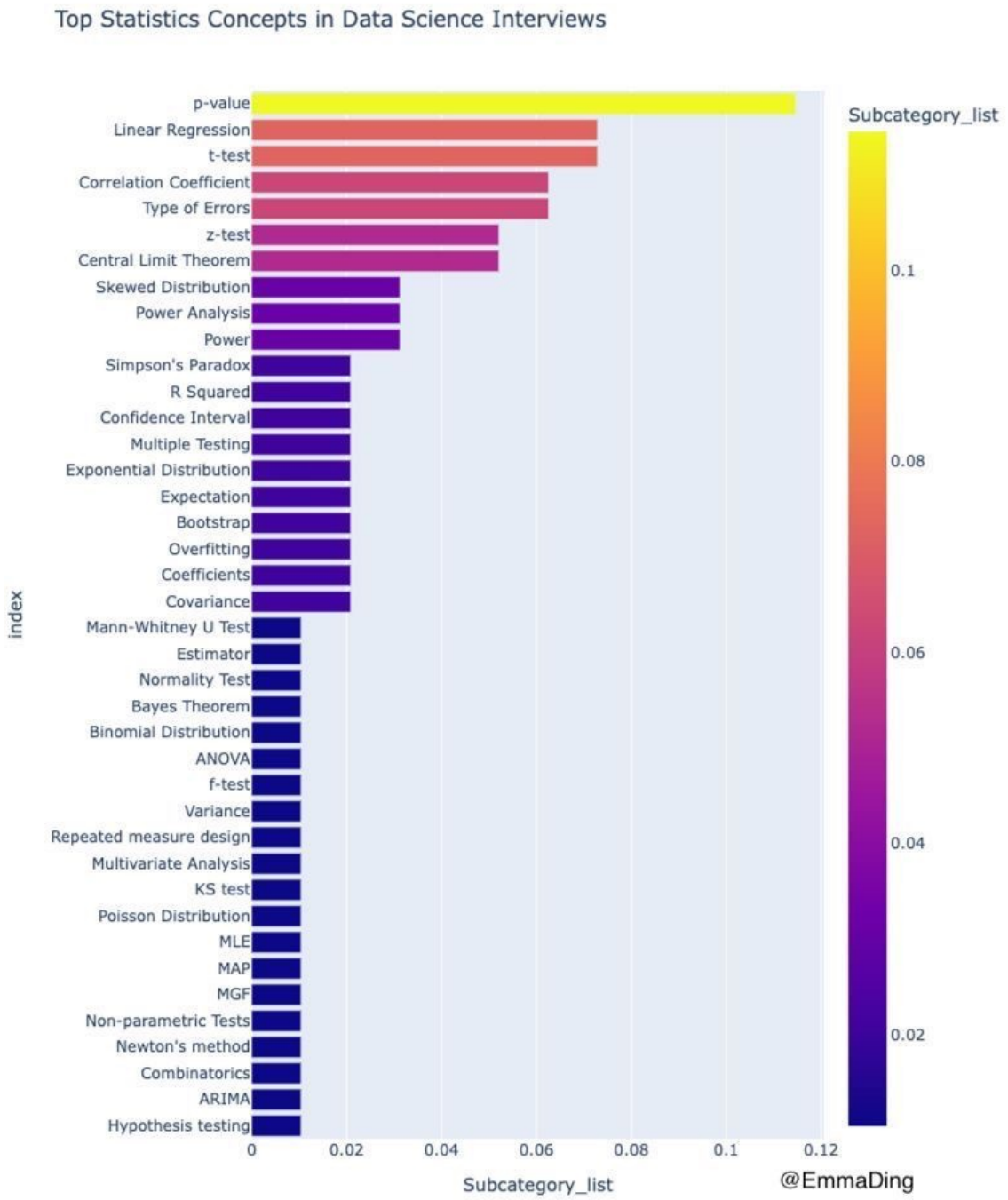
Inference

Machine Learning



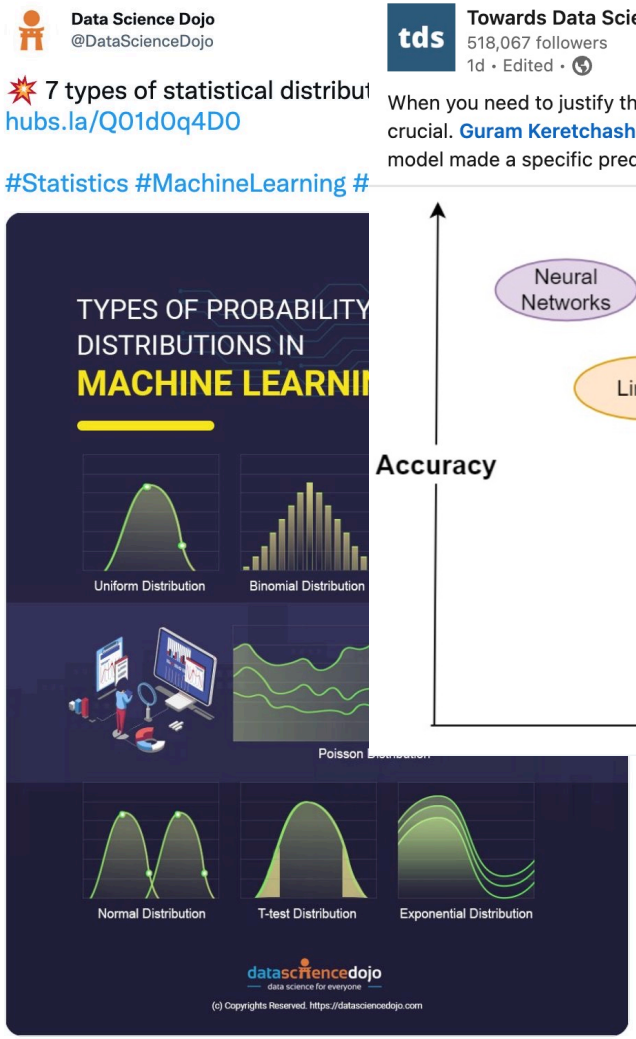


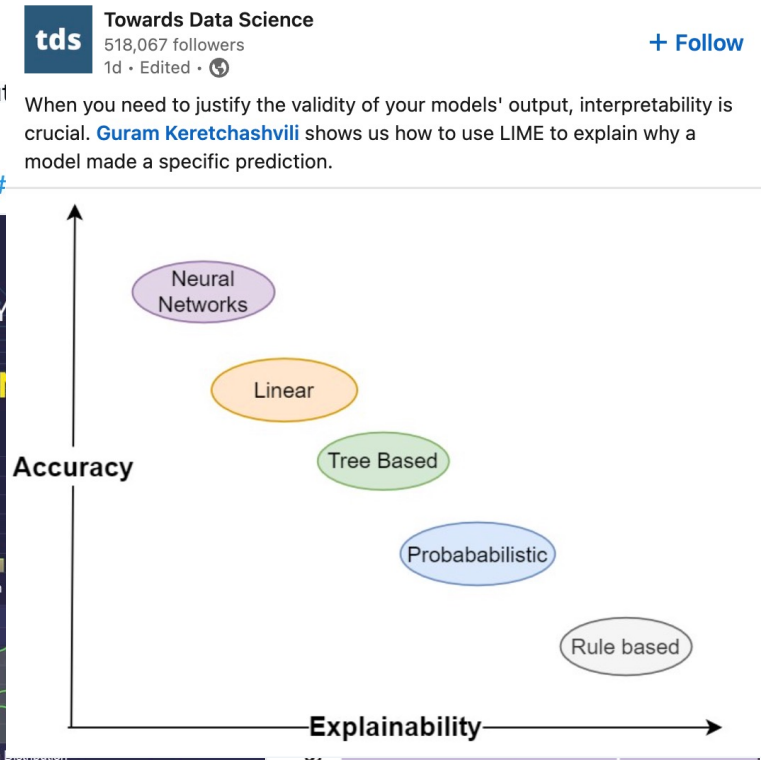
Why these topics in particular?

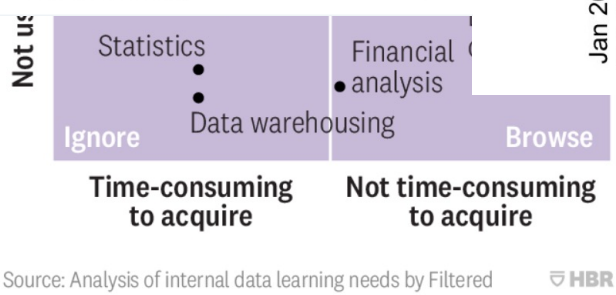


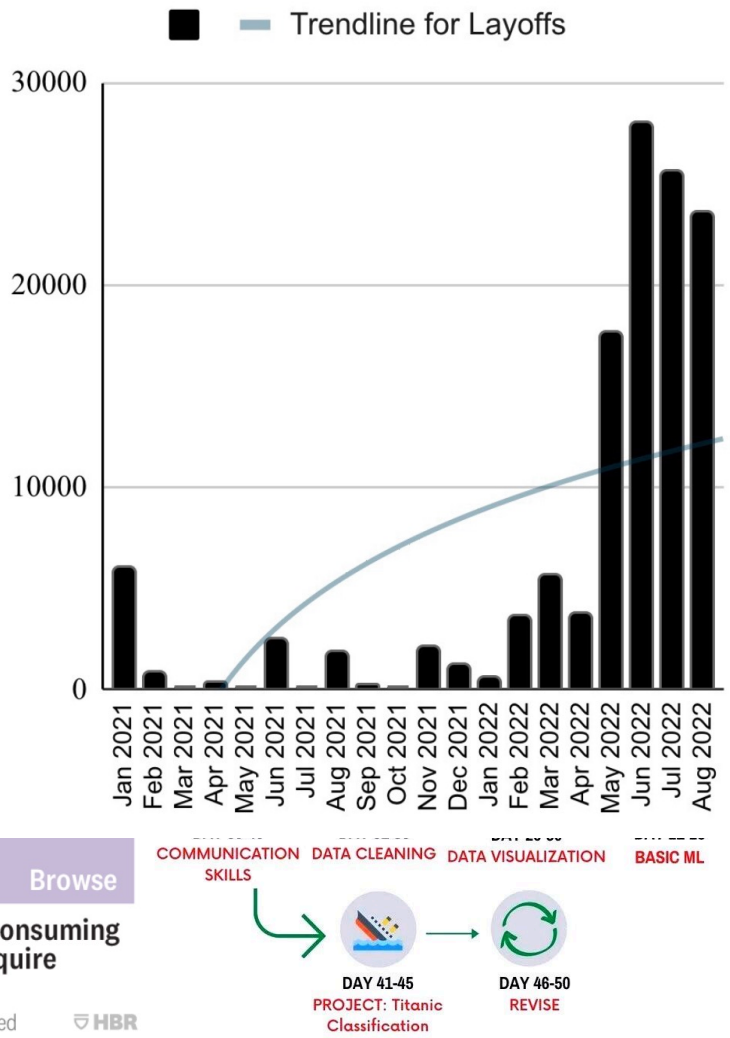
Why do we even need this class, in the age of abundant and freely available information?

Because “Bad Data Science” exists – it’s out there, lots of it:









Knowing things that are not true is only part of the problem with "Bad Data Science"

- Arguably, there is an even more insidious issue:
- Just having shards of specialized knowledge (even if they are accurate and valid) is not sufficient.
- You also need to be able to put them together in a meaningful way without too many unintended/unexpected consequences.
- There are many such cases in industry.
- Here, we will illustrate this with the cautionary tale of the Xerox automatic scanners.
  (courtesy David Kriesel)

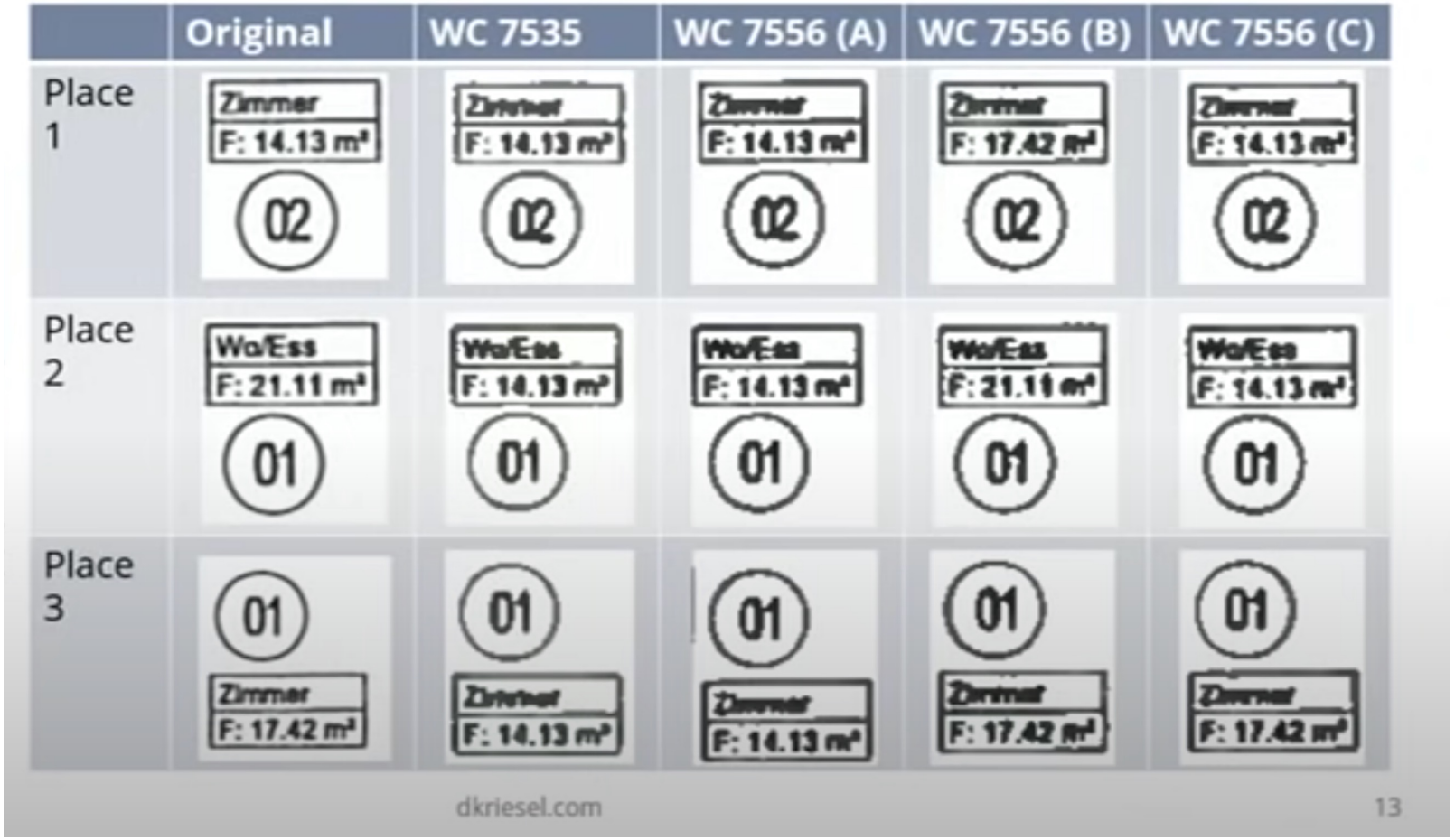
What can happen to you if you are missing the big picture: The case of the Xerox autoscanners

Xerox
Workstation 7535

Used by large organizations – corporations, national archives, the military, research institutions, courts, etc.

The issue first came to light in July 2013:

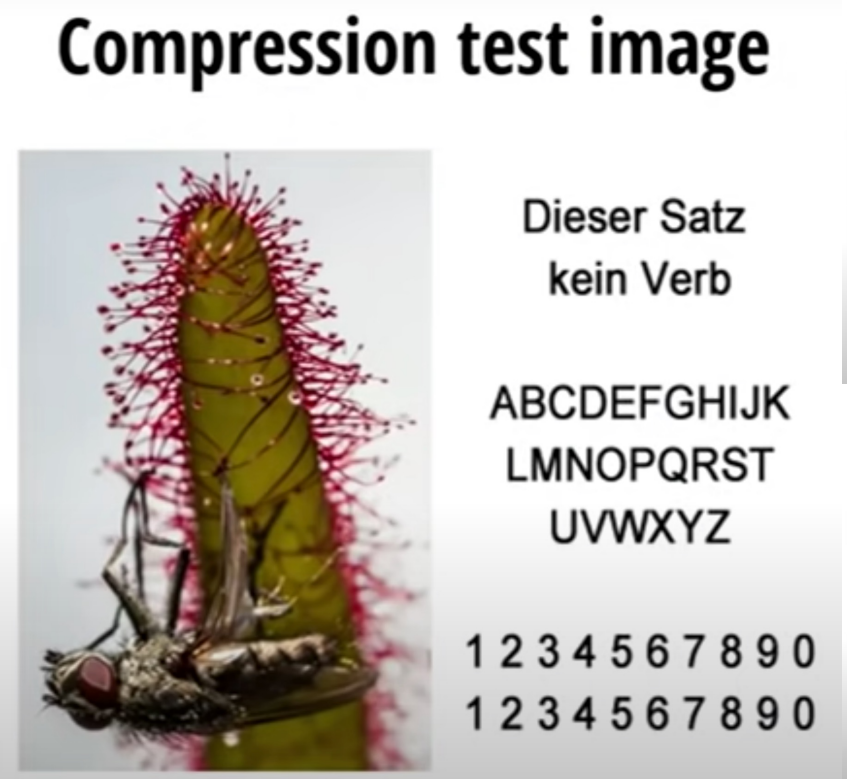
July 24th, 2013: Construction plan scan

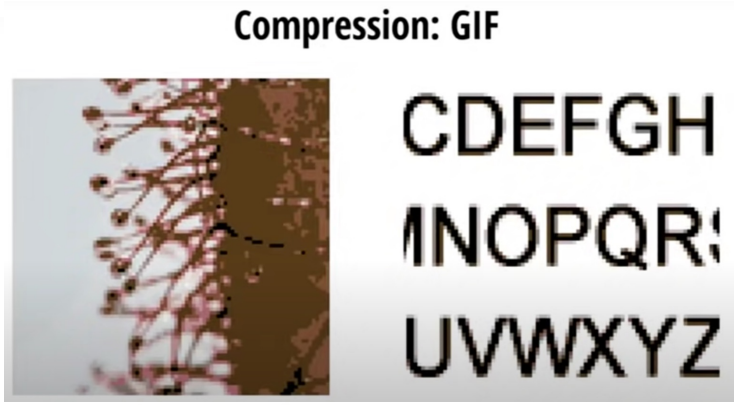


How could this have happened?

Insight 1: Images take up a lot of space, and pixel values are highly correlated, so it makes sense to compress images

Insight 2: Different compression schemes have different pros and cons



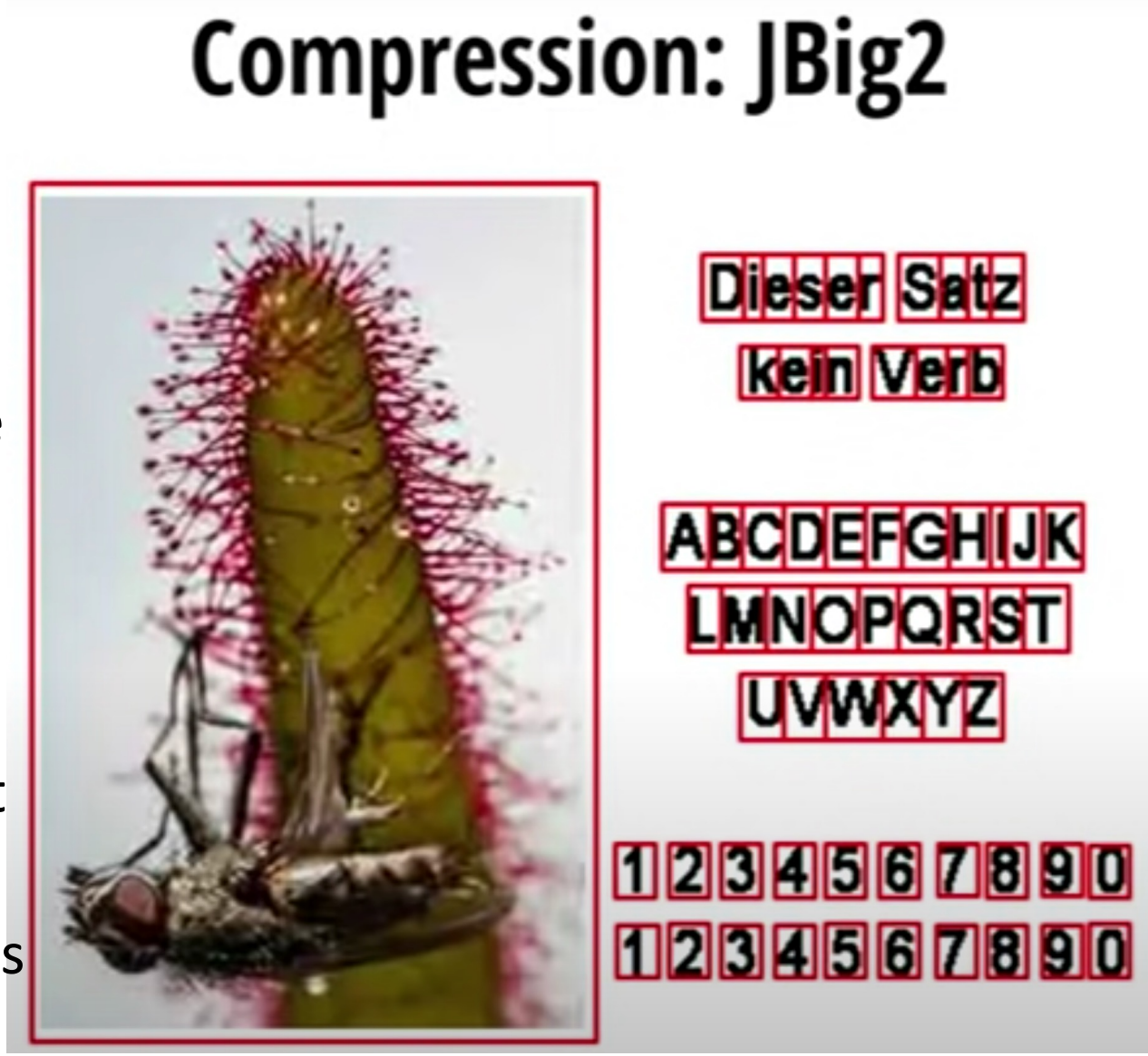


Lossless
Limited colors
Good for text
Bad for images

Lossy
Edge issues
Bad for text
Good for images

Insight 3: Text consists of many repeating elements –images often as well

- So Xerox decided to use JBig2 for image compression in their scanners.
- JBig2 works by partitioning the image into regions, then storing the cleanest instance of each region, and then replacing every repeat with that one.
- Dramatic space savings
- Technically lossless



Sounds great – what could go wrong?

Compression: JBig2

Dieser Satz
kein Verb

ABCDEFGHIJK
LMNOPQRST
UVWXYZ

1 2 3 4 5 6 7 8 9 0
1 2 3 4 5 8 7 8 9 0

JBig2 presumes perfect classification accuracy

A cautionary tale

- To create this problem, you had to have a fairly deep understanding and knowledge of image compression, computer vision and classification.
- And yet, you caused incalculable damage to your company (and those who relied on the veracity of the scans)
- There are hundreds of thousands of these machines in use around the world, each with many users.
- The “bug” was demonstrably in the wild and undocumented for at least 8 years (2005-2013).
- Users of the machines could not escape or avoid this issue, as it happened – undisclosed – on all compression settings.

Not great, as this happened at the height of digitization of paper documents

Bloomberg

Live Now | Markets | Economics | Industries | Tech | AI | Politics | Wealth | Pursuits | Opinion | Businessweek | Equalit

Business

Xerox Can Fix Number-Switching Scanners, but Not Altered Docs

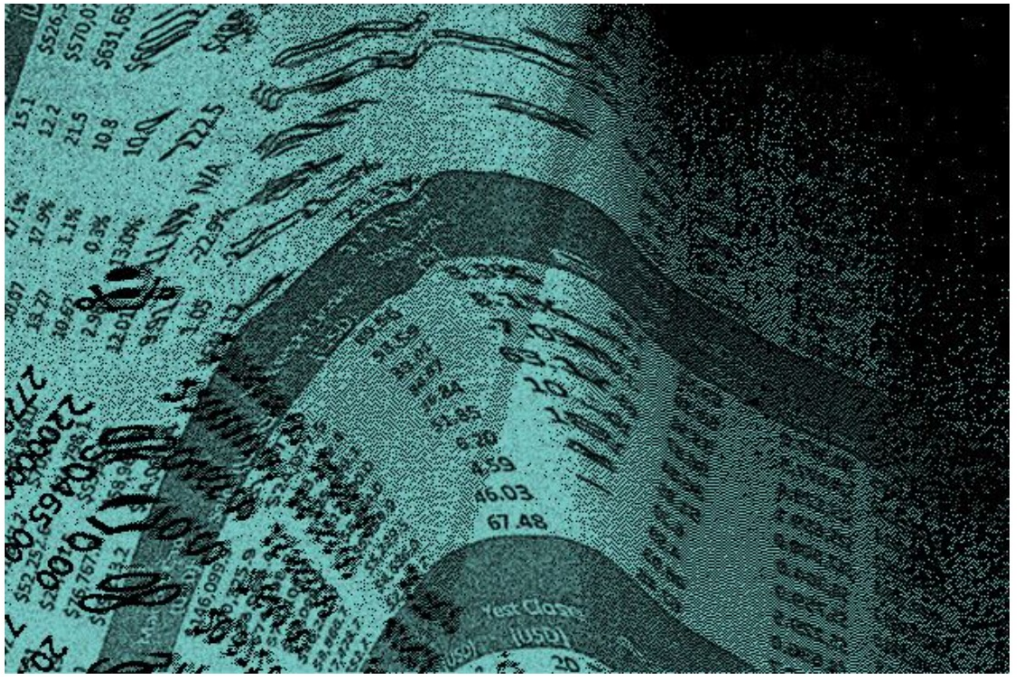


Illustration by 731

By Peter Coy
August 23, 2013 at 12:26 PM EDT

This is not an isolated case, it is fairly common

- Imagine that your institution offers a substantial childcare scholarship – but you have to apply for it by the due date.
- Imagine further that you have been at your institution for a long time, so you have a legacy email account that is auto-forwarded to your new institutional address (the institution switched systems at some point, but the old address is how you are registered with the school, in terms of netID, etc.).
- You missed out on the scholarship because you never saw any of the solicitations to apply.
- What happened?

How could this happen?

- The messages were not in the spam folder of your email account.
- They *were* in the spam folder of your original account – that you haven't checked in many years.

> The way it was configured before, by forwarding everything , it was expected that spam emails from NYU (pw44@nyu.edu) would not be rerouted to your new account (pascal.wallisch@nyu.edu), since spam is really never that important. So you never noticed anything important missing.

- What else have you been missing over the years?

Even academia is not spared

Bad Data Science is everywhere

Article | Open Access | Published: 29 August 2022

Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated

Eran Elhaik

Scientific Reports **12**, Article number: 14683 (2022) | Cite this article

38k Accesses | **619** Altmetric | Metrics

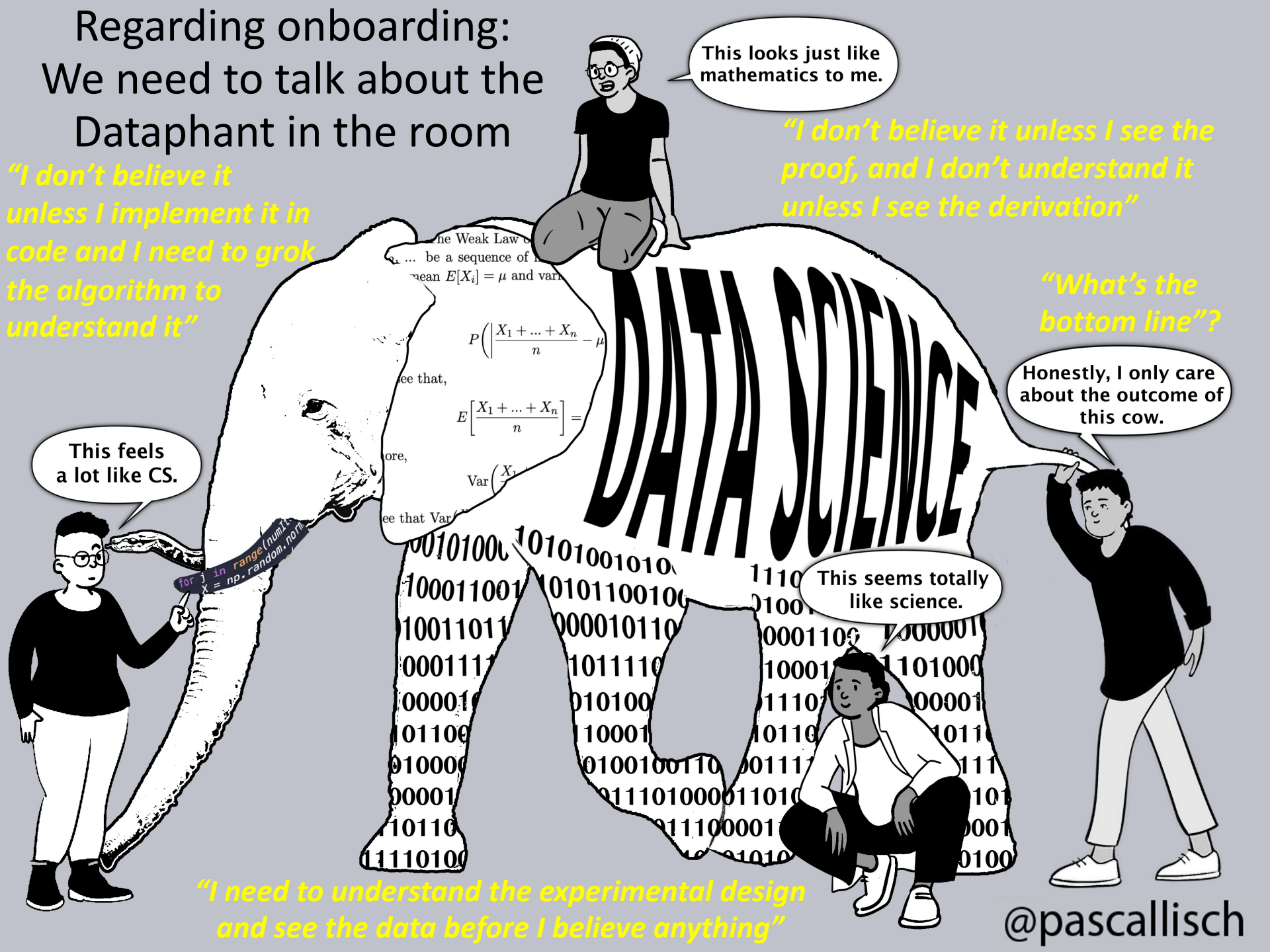
Abstract

Principal Component Analysis (PCA) is a multivariate analysis that reduces the complexity of datasets while preserving data covariance. The outcome can be visualized on colorful scatterplots, ideally with only a minimal loss of information. PCA applications, implemented in well-cited packages like EIGENSOFT and PLINK, are extensively used as the foremost analyses in population genetics and related fields (e.g., animal and plant or medical genetics). PCA outcomes are used to shape study design, identify, and characterize individuals and populations, and draw historical and ethnobiological conclusions on origins, evolution, dispersion, and relatedness. The replicability crisis in science has prompted us to evaluate whether PCA results are reliable, robust, and replicable. We analyzed twelve common test cases using an intuitive color-based model alongside human population data. We demonstrate that PCA results can be artifacts of the data and can be easily manipulated to generate desired outcomes. PCA adjustment also yielded unfavorable outcomes in association studies. PCA results may not be reliable, robust, or replicable as the field assumes. Our findings raise concerns about the validity of results reported in the population genetics literature and related fields that place a disproportionate reliance upon PCA outcomes and the insights derived from them. We conclude that PCA may have a biasing role in genetic investigations and that 32,000–216,000 genetic studies should be reevaluated. An alternative mixed-admixture population genetic model is discussed.

High level course goals

1. “Voltage regulator” – onboard people from many different backgrounds
2. Getting used to an integrated, “big picture” perspective on problem solving in Data Science (trying to prevent “Bad Data Science”)
3. Building strong foundations – planting seeds that will enable you to take more advanced classes
4. “Secret shelf”
5. [?]

Regarding onboarding: We need to talk about the Dataphant in the room



@pascallisch

Data Science is different from and more than the sum of its parts

- **Mathematics**: Rigor & Logic → *Clarity*
- **CS**: Algorithms & Computers → *Power*
- **Science**: Epistemology & Measurements → *Data*
- **Econ/Engineering**: Problem optimization → *Solutions*

Implications

- Recognize (and work on overcoming?) your priors
- Developing and adopting a genuine Data Science mindset requires a broad-minded person (as a decent understanding of all 3 foundations is necessary and an eye on problem solving is helpful).
- Therefore, intellectual humility should be the default.
- A commitment to life-long learning is essential.

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)

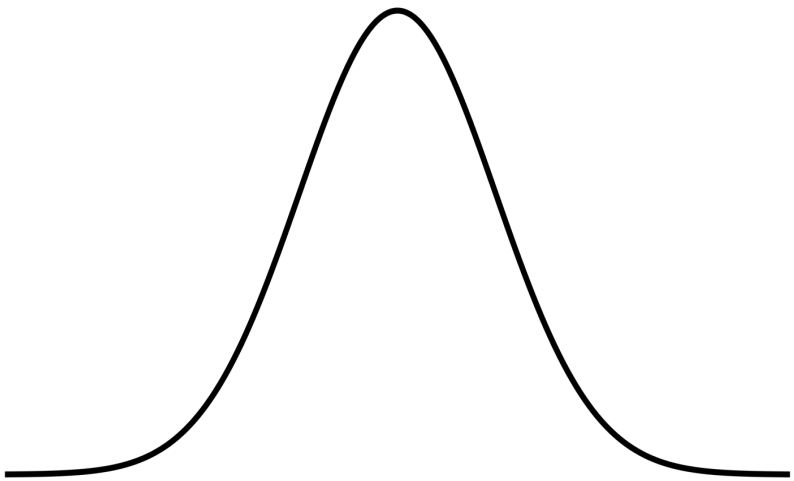
As a Data Scientist, the computer is your horse, you are the jockey

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)

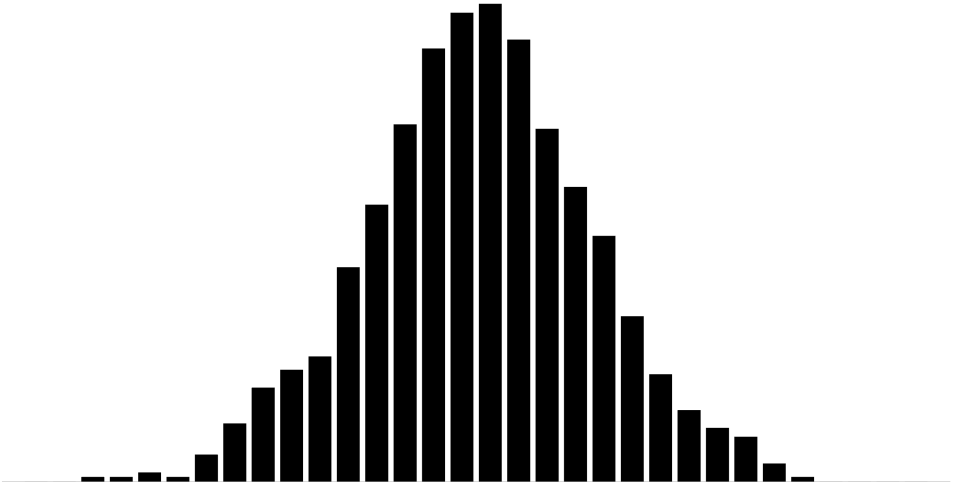
DATA SCIENCE

So much for CS, what about math?

Math is axiomatic
(Deductive)



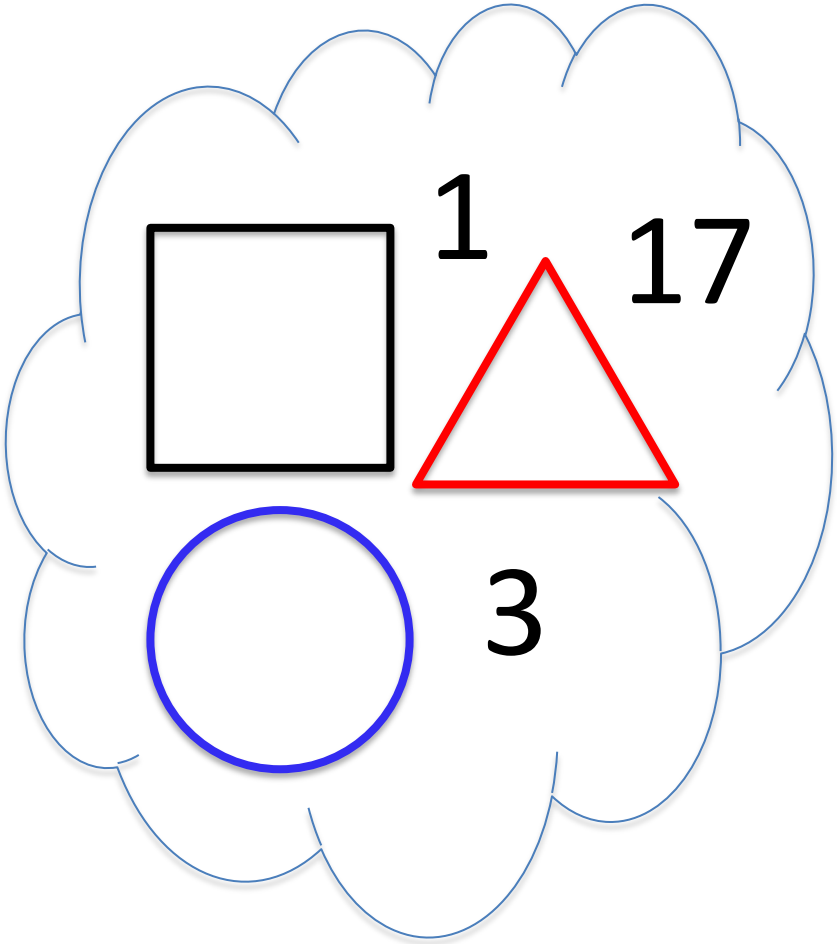
Science uses data
(Inductive)



What is data?
(From a DS perspective)

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

Mathematics



Simplicity, Beauty, Symmetry

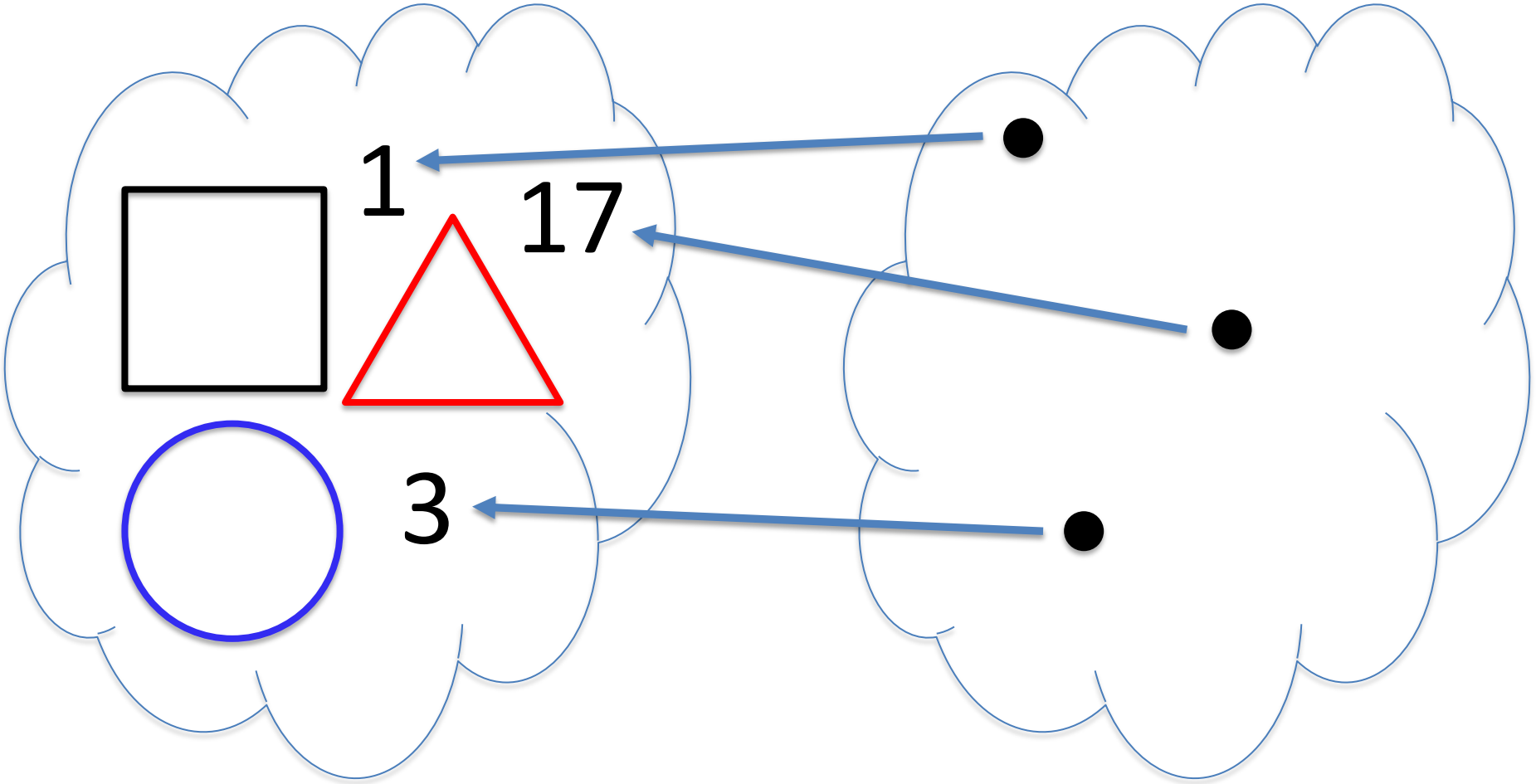
Qualitative descriptions of the natural world

Complex, messy, broken symmetry

Quantification = mapping between these realms

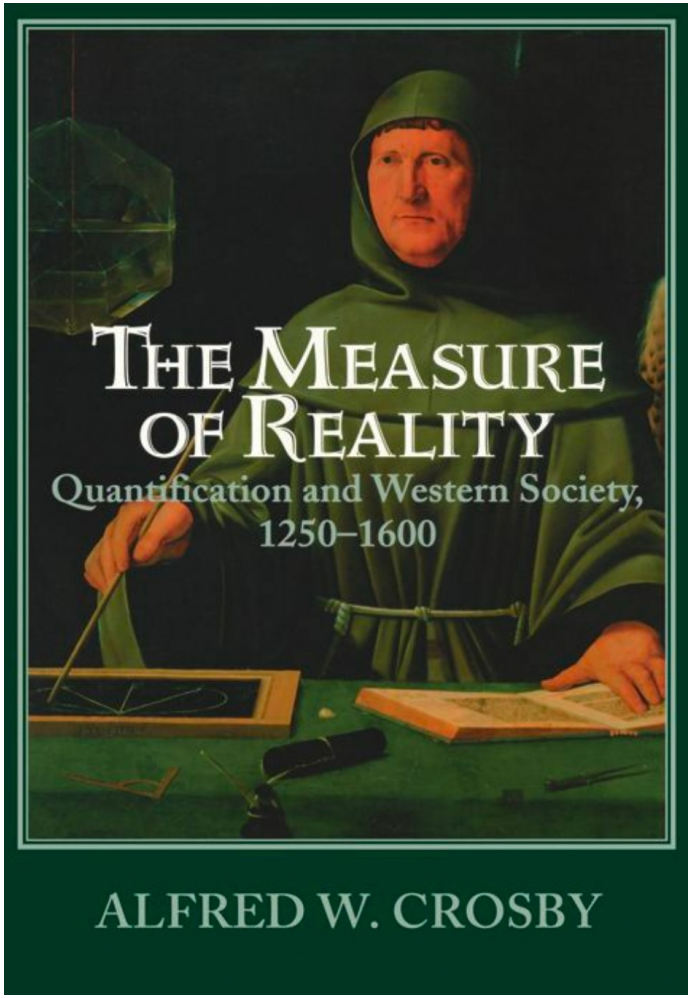
Mathematics

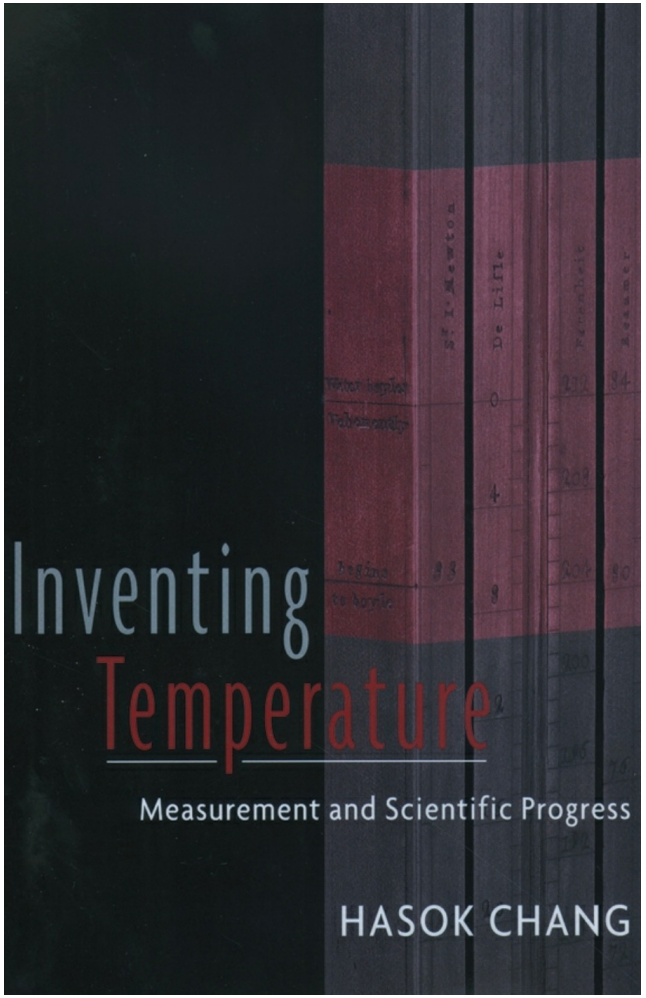
Descriptions of the natural world



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.

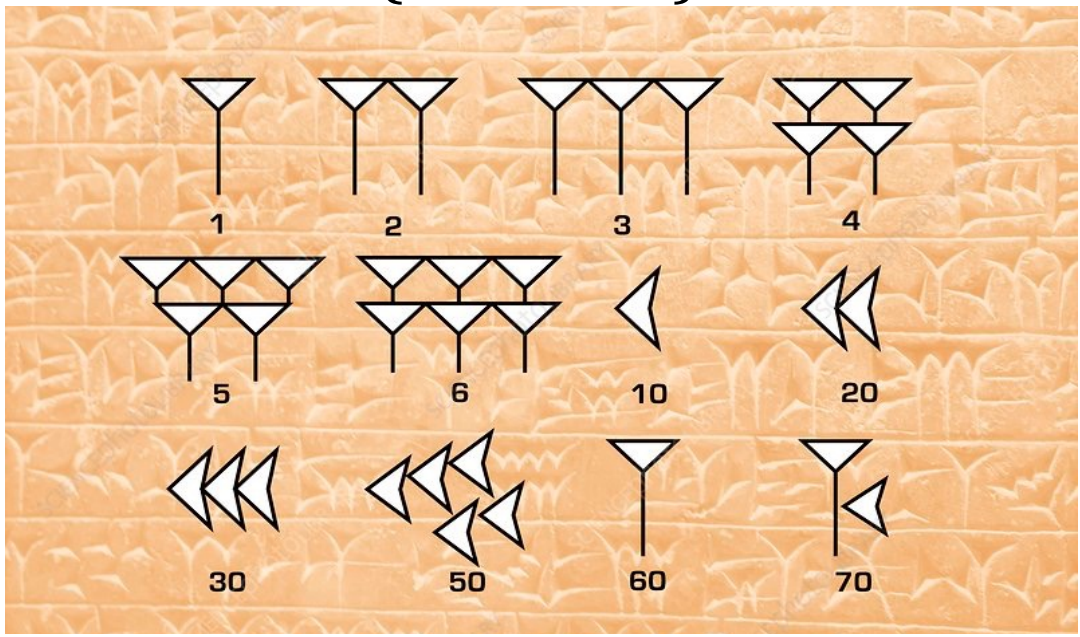




To be clear:
The idea to use numbers to tally specific concrete objects is more than ancient

Tally sticks like the "Wolf Bone" found in Vestonice ~30,000 BC

Babylonian cuneiform numerals (2000 BC)



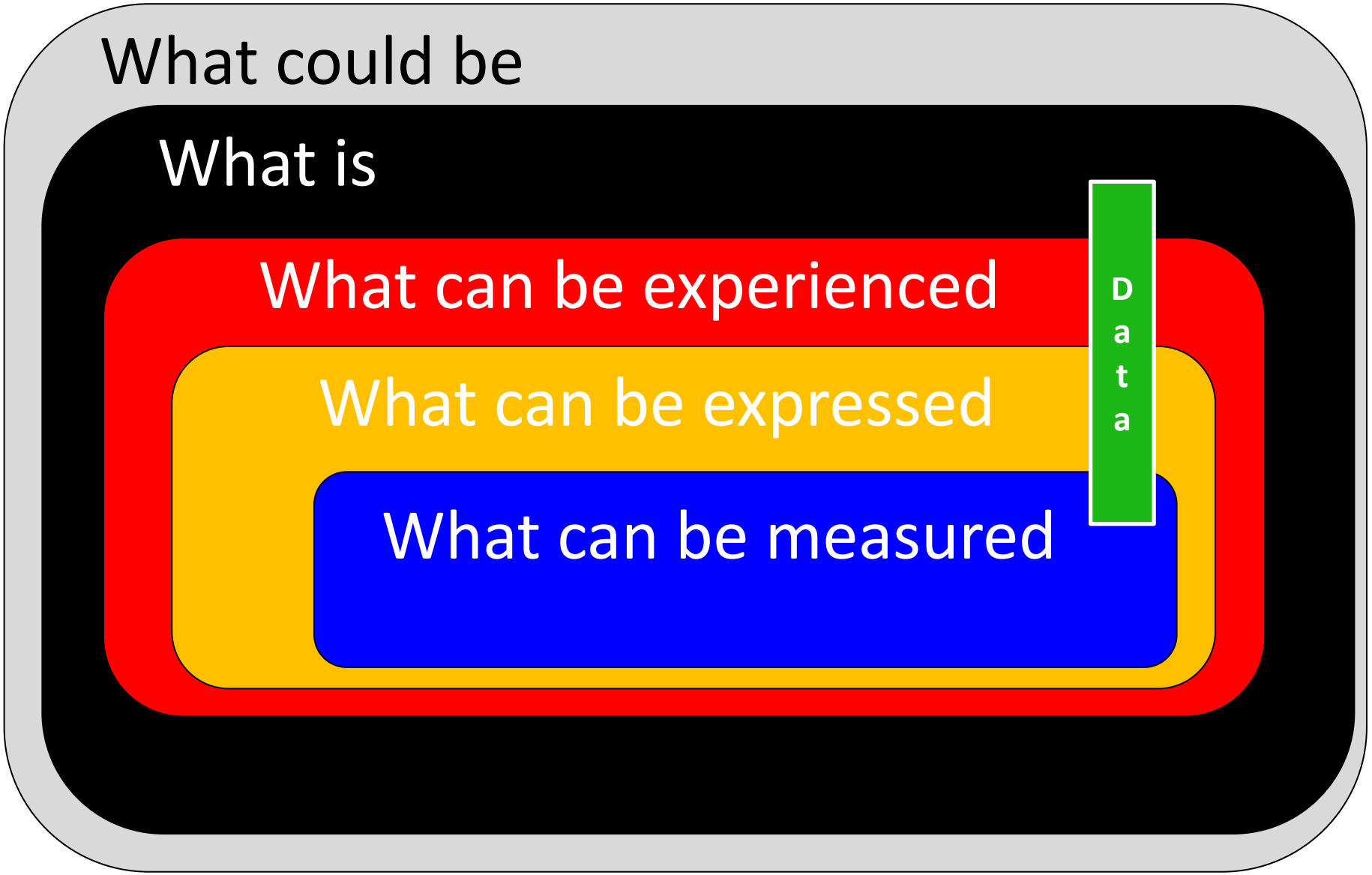
Flegg (2002). Numbers: Their History and Meaning

But the idea to use numbers to represent abstract qualities (and digitize them) is not

What can't be

The epistemic landscape

(not to scale)



This is still an ongoing development

THE AMERICAN
JOURNAL OF SOCIOLOGY

VOLUME XXXIII JANUARY 1928 NUMBER 4

ATTITUDES CAN BE MEASURED[1]

L. L. THURSTONE
University of Chicago

ABSTRACT

The object of this study is to devise a method whereby the distribution of attitude of a group on a specified issue may be represented in the form of a frequency distribution. The base line represents ideally the whole range of opinions from those at one end who are most strongly in favor of the issue to those at the other end of the scale who are as strongly against it. Somewhere between the two extremes on the base line will be a neutral zone representing indifferent attitudes on the issue in

RESEARCH ARTICLE | ECONOMIC SCIENCES |

The scientific value of numerical measures of human feelings

Caspar Kaiser and Andrew J. Oswald Authors Info & Affiliations

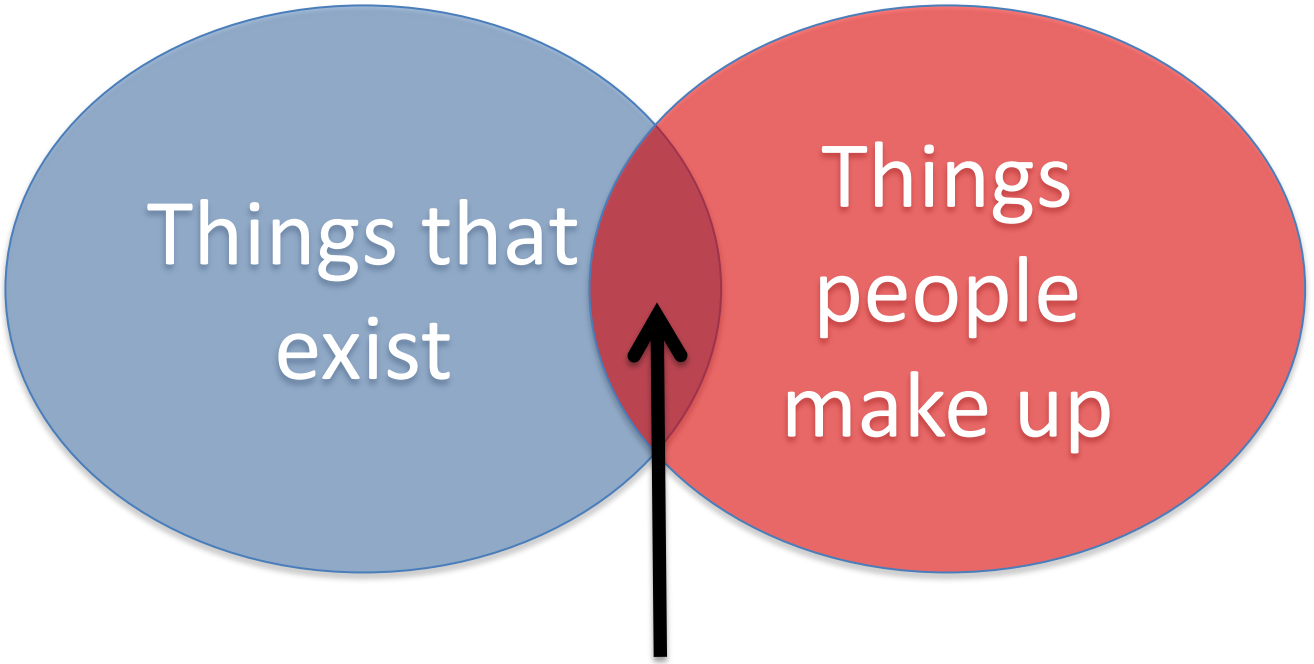
Edited by Richard Easterlin, University of Southern California, Los Angeles, CA; received June 30, 2022; accepted August 29, 2022

October 3, 2022 | 119 (42) e2210412119 | https://doi.org/10.1073/pnas.2210412119

Significance

Human feelings cannot be expressed on a numerical scale. There are no units of measurement for feelings. However, such data are extensively collected in the modern world—by governments, corporations, and international organizations. Why? Our study finds that a feelings integer (like *my happiness is X out of 10*) has more predictive power than a collection of socioeconomic influences. Moreover, there is a clear link between those feelings numbers and later get-me-out-of-here actions. Finally, the feelings-to-actions relationship appears replicable and not too far from linear. Remarkably, therefore, humans somehow manage to choose their numerical answers in a systematic way as though they sense within themselves—and can communicate—a reliable numerical scale for their feelings. How remains an unsolved puzzle.

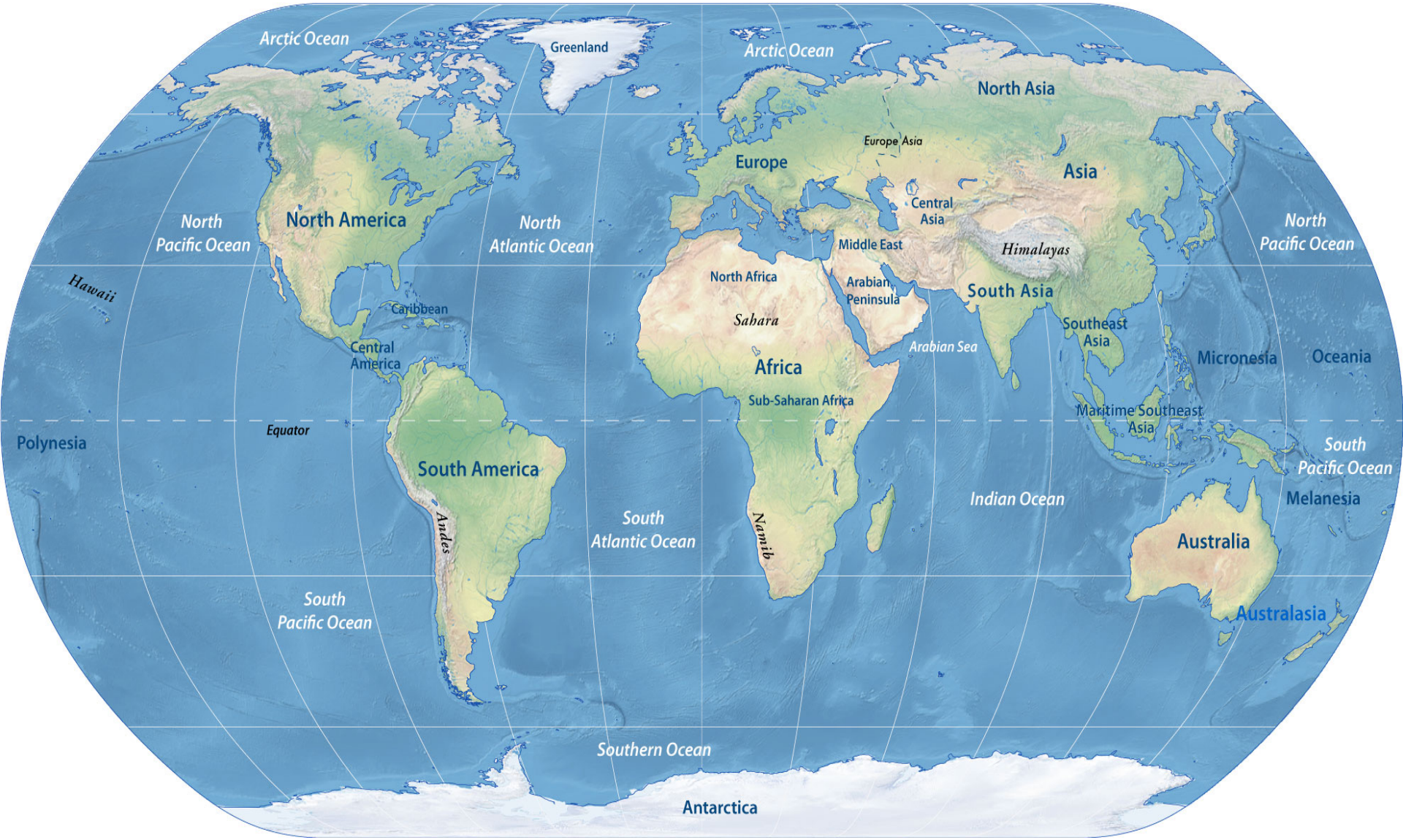
Data is truly special



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



(Lots of) data and the power to process them opens up a new approach:

The Lagrangian of the standard model of physics:

$$\mathcal{L}_{SM} = -\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc}f^{ade} g^b_\mu g^c_\nu g^d_\mu g^e_\nu - \partial_\nu W^+_\mu \partial_\nu W^-_\mu - M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu \partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu \partial_\mu A_\nu - igc_w(\partial_\nu Z^0_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - Z^0_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + Z^0_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - igs_w(\partial_\nu A_\mu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu (W^+_\mu \partial_\nu W^-_\mu - W^-_\mu \partial_\nu W^+_\mu) + A_\mu (W^+_\nu \partial_\nu W^-_\mu - W^-_\nu \partial_\nu W^+_\mu)) - \tfrac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \tfrac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2 (Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + g^2 s_w^2 (A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w (A_\mu Z^0_\nu (W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu) - \tfrac{1}{2}\partial_\mu H \partial_\mu H - 2M^2 \alpha_h H^2 - \partial_\mu \phi^+ \partial_\mu \phi^- - \tfrac{1}{2}\partial_\mu \phi^0 \partial_\mu \phi^0 - \beta_h \left(\tfrac{2M^2}{g^2} + \tfrac{2M}{g}H + \tfrac{1}{2}(H^2 + \phi^0\phi^0 + 2\phi^+\phi^-)\right) + \tfrac{2M^4}{g^2}\alpha_h - g\alpha_h M (H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-) - \tfrac{1}{8}g^2\alpha_h (H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2 H^2) - gMW^+_\mu W^-_\mu H - \tfrac{1}{2}g\tfrac{M}{c_w^2} Z^0_\mu Z^0_\mu H - \tfrac{1}{2}ig\left(W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)\right) + \tfrac{1}{2}g\left(W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) + W^-_\mu(H\partial_\mu\phi^+ - \phi^+\partial_\mu H)\right) + \tfrac{1}{2}g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) + M\left(\tfrac{1}{c_w}Z^0_\mu\partial_\mu\phi^0 + W^+_\mu\partial_\mu\phi^- + W^-_\mu\partial_\mu\phi^+\right) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + igs_w MA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + igs_w A_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac{1}{4}g^2 W^+_\mu W^-_\mu (H^2 + (\phi^0)^2 + 2\phi^+\phi^-) - \tfrac{1}{8}g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu (H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)^2\phi^+\phi^-) - \tfrac{1}{2}g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) - \tfrac{1}{2}ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac{1}{2}g^2 s_w A_\mu\phi^0(W^+_\mu\phi^- + W^-_\mu\phi^+) + \tfrac{1}{2}ig^2 s_w A_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2 - 1)Z^0_\mu A_\mu\phi^+\phi^- - g^2 s_w^2 A_\mu A_\mu\phi^+\phi^- + \tfrac{1}{2}ig_s\lambda^a_{ij}(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu - \bar{e}^\lambda(\gamma\partial + m^\lambda_e)e^\lambda - \bar{\nu}^\lambda(\gamma\partial + m^\lambda_\nu)\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial + m^\lambda_u)u^\lambda_j - \bar{d}^\lambda_j(\gamma\partial + m^\lambda_d)d^\lambda_j + igs_w A_\mu\left(-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j\gamma^\mu d^\lambda_j)\right) + \tfrac{ig}{4c_w}Z^0_\mu\{(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{e}^\lambda\gamma^\mu(4s_w^2 - 1 - \gamma^5)e^\lambda) + (\bar{d}^\lambda_j\gamma^\mu(\tfrac{4}{3}s_w^2 - 1 - \gamma^5)d^\lambda_j) + (\bar{u}^\lambda_j\gamma^\mu(1 - \tfrac{8}{3}s_w^2 + \gamma^5)u^\lambda_j)\} + \tfrac{ig}{2\sqrt{2}}W^+_\mu\left((\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)U^{lep}{}_{\lambda\kappa}e^\kappa) + (\bar{u}^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)\right) + \tfrac{ig}{2\sqrt{2}}W^-_\mu\left((\bar{e}^\kappa U^{lep\dagger}_{\kappa\lambda}\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\kappa\lambda}\gamma^\mu(1+\gamma^5)u^\lambda_j)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_e(\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa}(1-\gamma^5)e^\kappa) + m^\lambda_\nu(\bar{\nu}^\lambda U^{lep}{}_{\lambda\kappa}(1+\gamma^5)e^\kappa)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_e(\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa}(1+\gamma^5)\nu^\kappa) - m^\kappa_\nu(\bar{e}^\lambda U^{lep\dagger}_{\lambda\kappa}(1-\gamma^5)\nu^\kappa)\right) - \tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{\nu}^\lambda\nu^\lambda) - \tfrac{g}{2}\tfrac{m^\lambda_e}{M}H(\bar{e}^\lambda e^\lambda) + \tfrac{ig}{2}\tfrac{m^\lambda_\nu}{M}\phi^0(\bar{\nu}^\lambda\gamma^5\nu^\lambda) - \tfrac{ig}{2}\tfrac{m^\lambda_e}{M}\phi^0(\bar{e}^\lambda\gamma^5 e^\lambda) - \tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\hat{\nu}_\kappa - \tfrac{1}{4}\bar{\nu}_\lambda M^R_{\lambda\kappa}(1-\gamma_5)\hat{\nu}_\kappa + \tfrac{ig}{2M\sqrt{2}}\phi^+\left(-m^\kappa_d(\bar{u}^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + m^\lambda_u(\bar{u}^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j)\right) + \tfrac{ig}{2M\sqrt{2}}\phi^-\left(m^\lambda_d(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m^\kappa_u(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1-\gamma^5)u^\kappa_j)\right) - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{u}^\lambda_j u^\lambda_j) - \tfrac{g}{2}\tfrac{m^\lambda_u}{M}H(\bar{d}^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m^\lambda_u}{M}\phi^0(\bar{u}^\lambda_j\gamma^5 u^\lambda_j) - \tfrac{ig}{2}\tfrac{m^\lambda_d}{M}\phi^0(\bar{d}^\lambda_j\gamma^5 d^\lambda_j) + \bar{G}^a\partial^2 G^a + g_s f^{abc}\partial_\mu\bar{G}^a G^b g^c_\mu + \bar{X}^+(\partial^2 - M^2)X^+ + \bar{X}^-(\partial^2 - M^2)X^- + \bar{X}^0(\partial^2 - \tfrac{M^2}{c_w^2})X^0 + \bar{Y}\partial^2 Y + igc_w W^+_\mu(\partial_\mu\bar{X}^0 X^- - \partial_\mu\bar{X}^+ X^0) + igs_w W^+_\mu(\partial_\mu\bar{Y}X^- - \partial_\mu\bar{X}^+ Y) + igc_w W^-_\mu(\partial_\mu\bar{X}^- X^0 - \partial_\mu\bar{X}^0 X^+) + igs_w W^-_\mu(\partial_\mu\bar{X}^- Y - \partial_\mu\bar{Y}X^+) + igc_w Z^0_\mu(\partial_\mu\bar{X}^+ X^+ - \partial_\mu\bar{X}^- X^-) + igs_w A_\mu(\partial_\mu\bar{X}^+ X^+ - \partial_\mu\bar{X}^- X^-) - \tfrac{1}{2}gM\left(\bar{X}^+X^+H + \bar{X}^-X^-H + \tfrac{1}{c_w^2}\bar{X}^0X^0H\right) + \tfrac{1-2c_w^2}{2c_w}igM\left(\bar{X}^+X^0\phi^+ - \bar{X}^-X^0\phi^-\right) + \tfrac{1}{2c_w}igM\left(\bar{X}^0X^-\phi^+ - \bar{X}^0X^+\phi^-\right) + igMs_w\left(\bar{X}^0X^-\phi^+ - \bar{X}^0X^+\phi^-\right) + \tfrac{1}{2}igM\left(\bar{X}^+X^+\phi^0 - \bar{X}^-X^-\phi^0\right).$$

vs.:

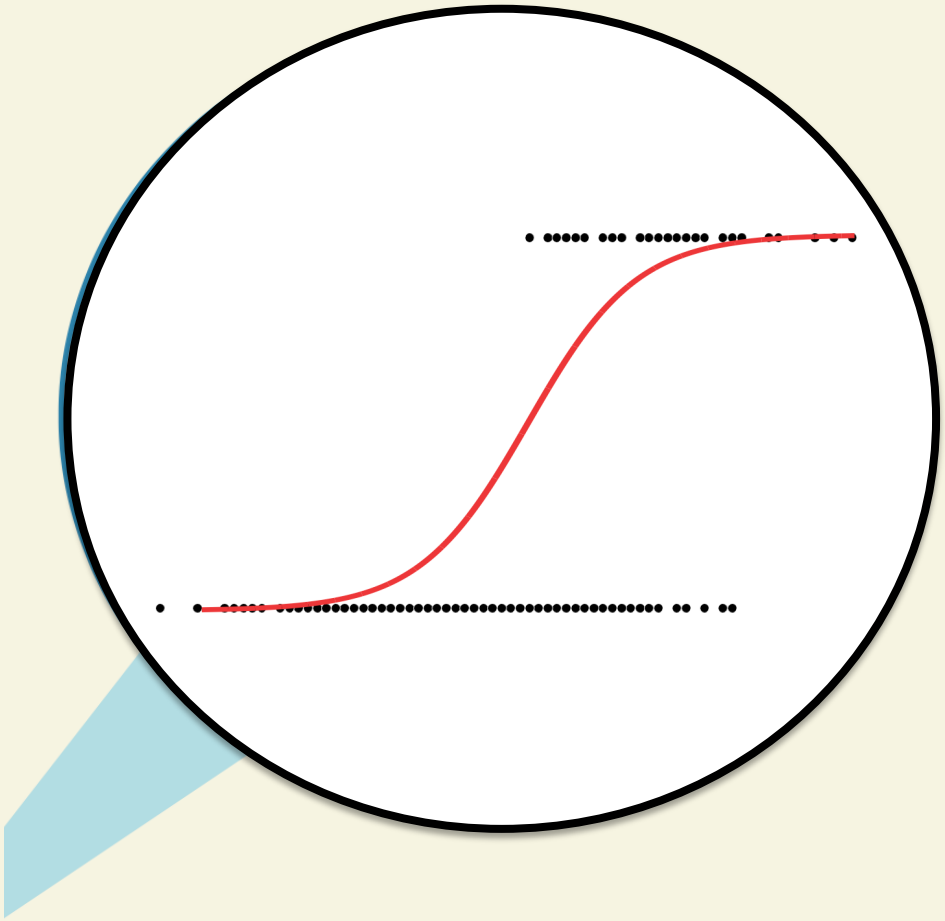
Simple learning rule (e.g. backprop)

+

lots of data

+

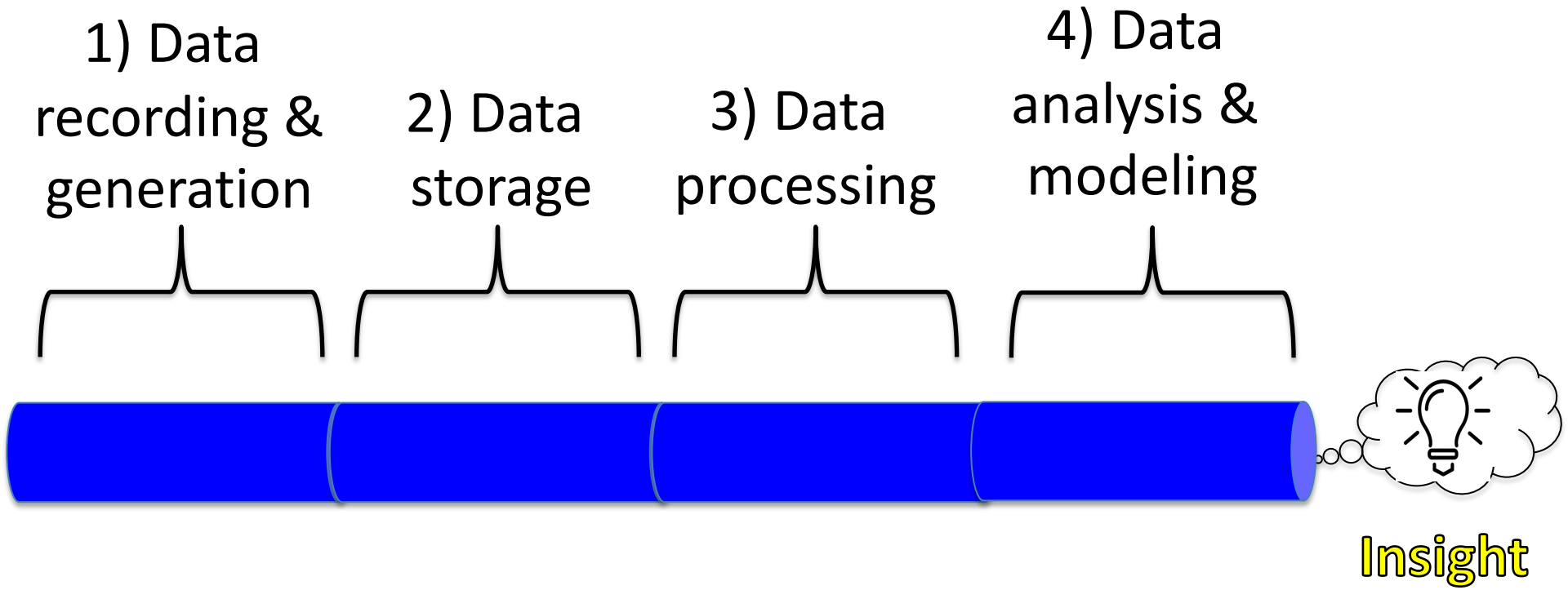
fast computer (lots of iterations)



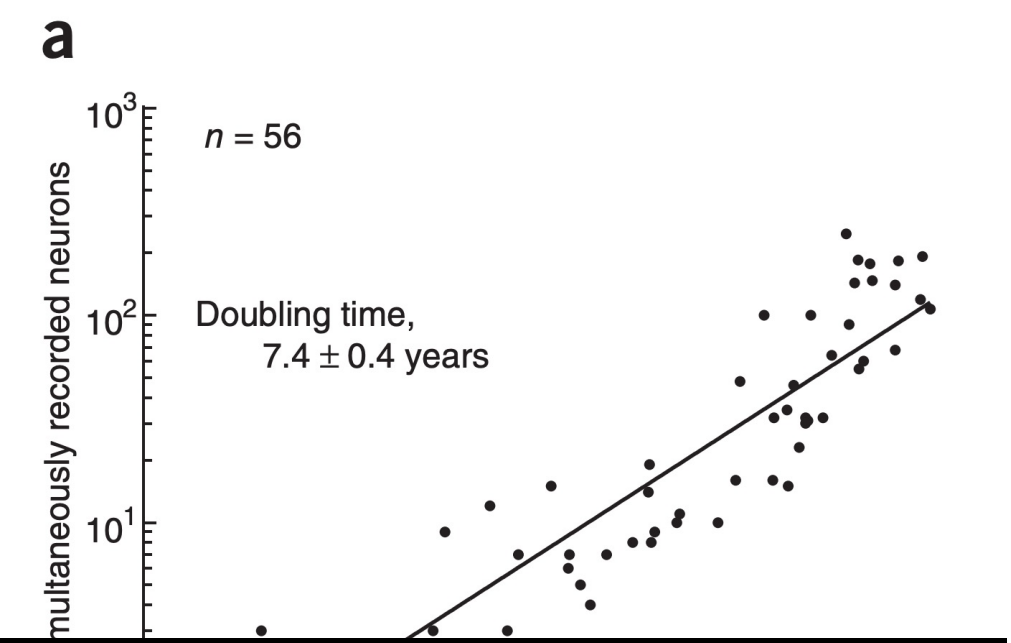
The computer is *our* microscope – on data

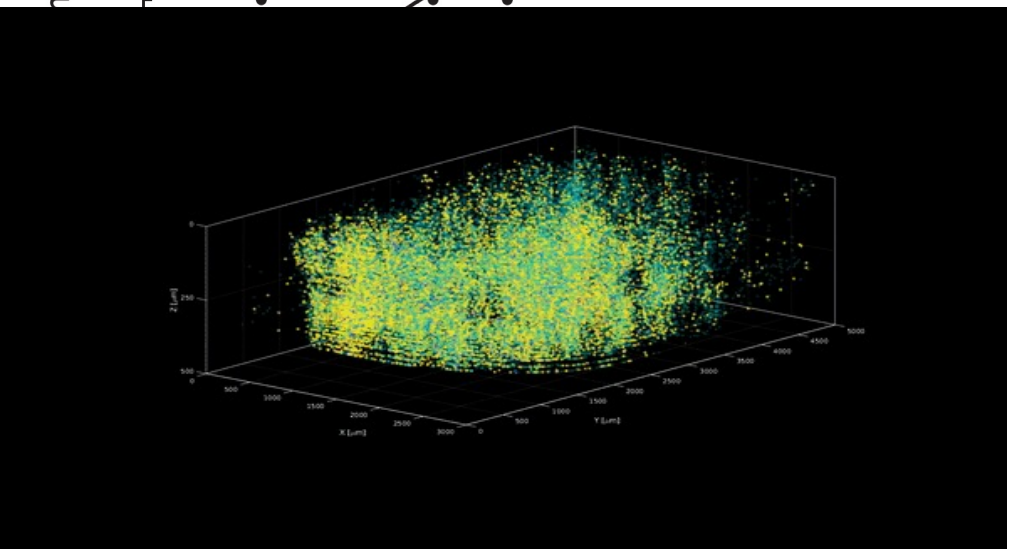
The age of Data (Science)

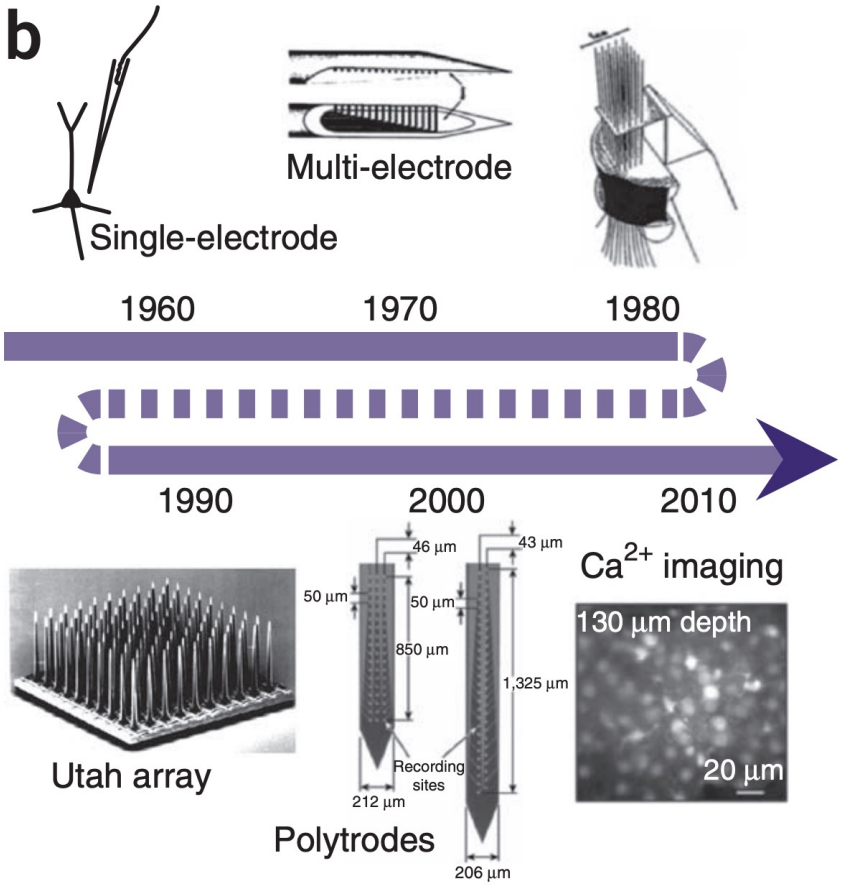
The data pipeline in the middle of the 20th century:



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







Vaziri et al. (2021)

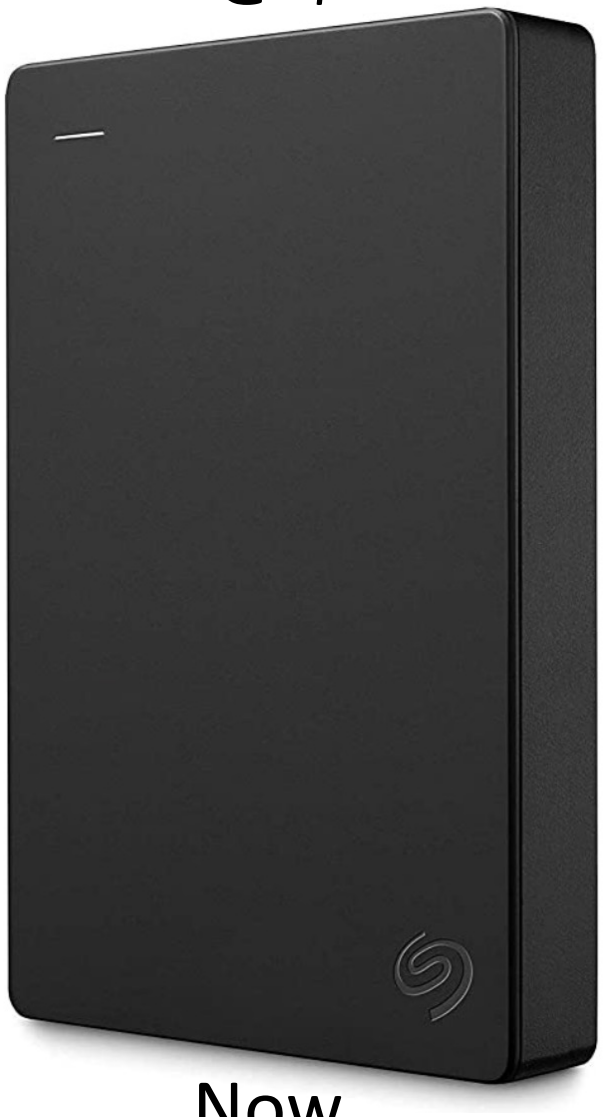
Stevenson & Kording (2011)

2) We are also able to store all this data

5 MB @ $30k/m

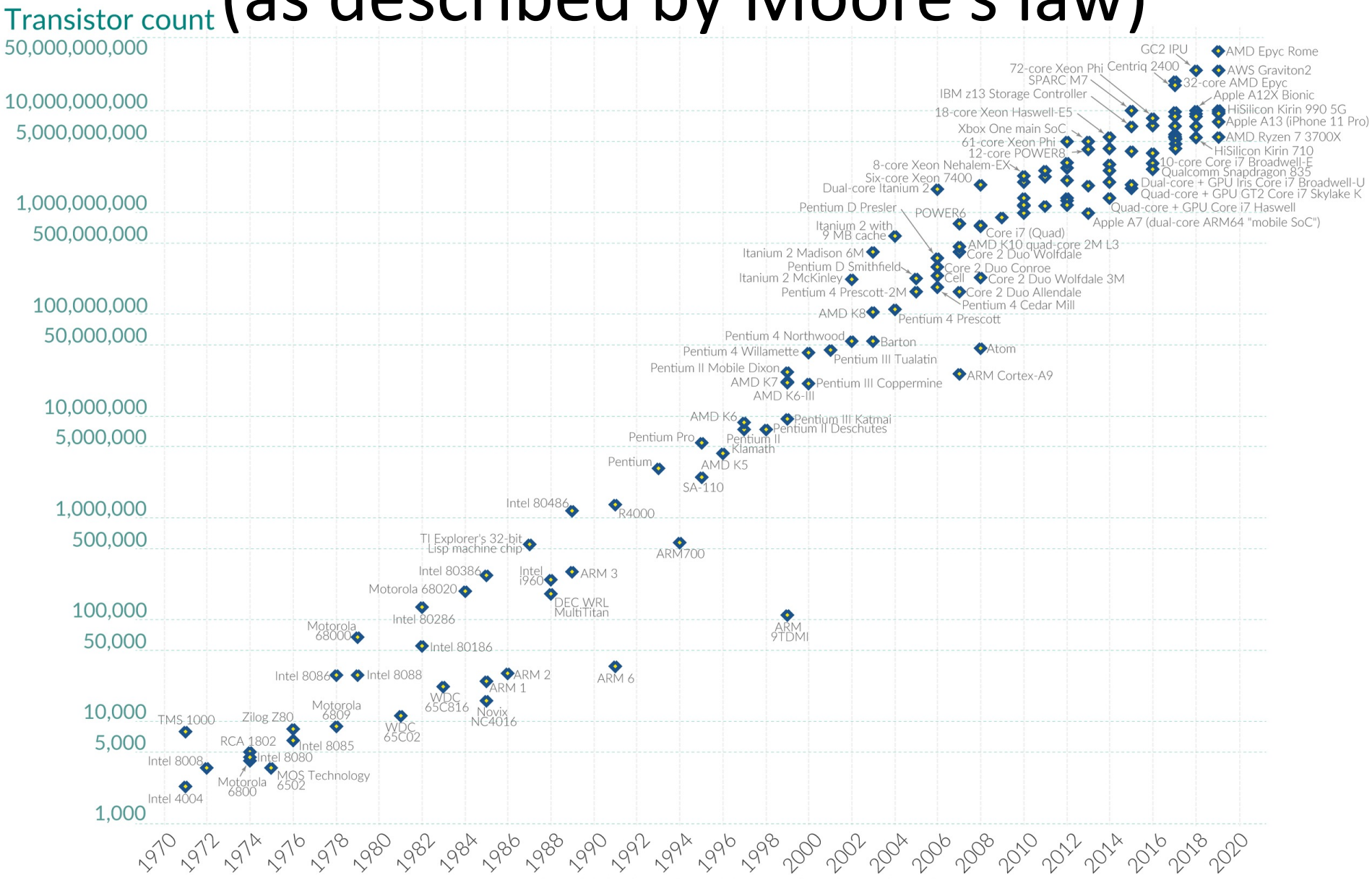
1956

5 TB @ $300

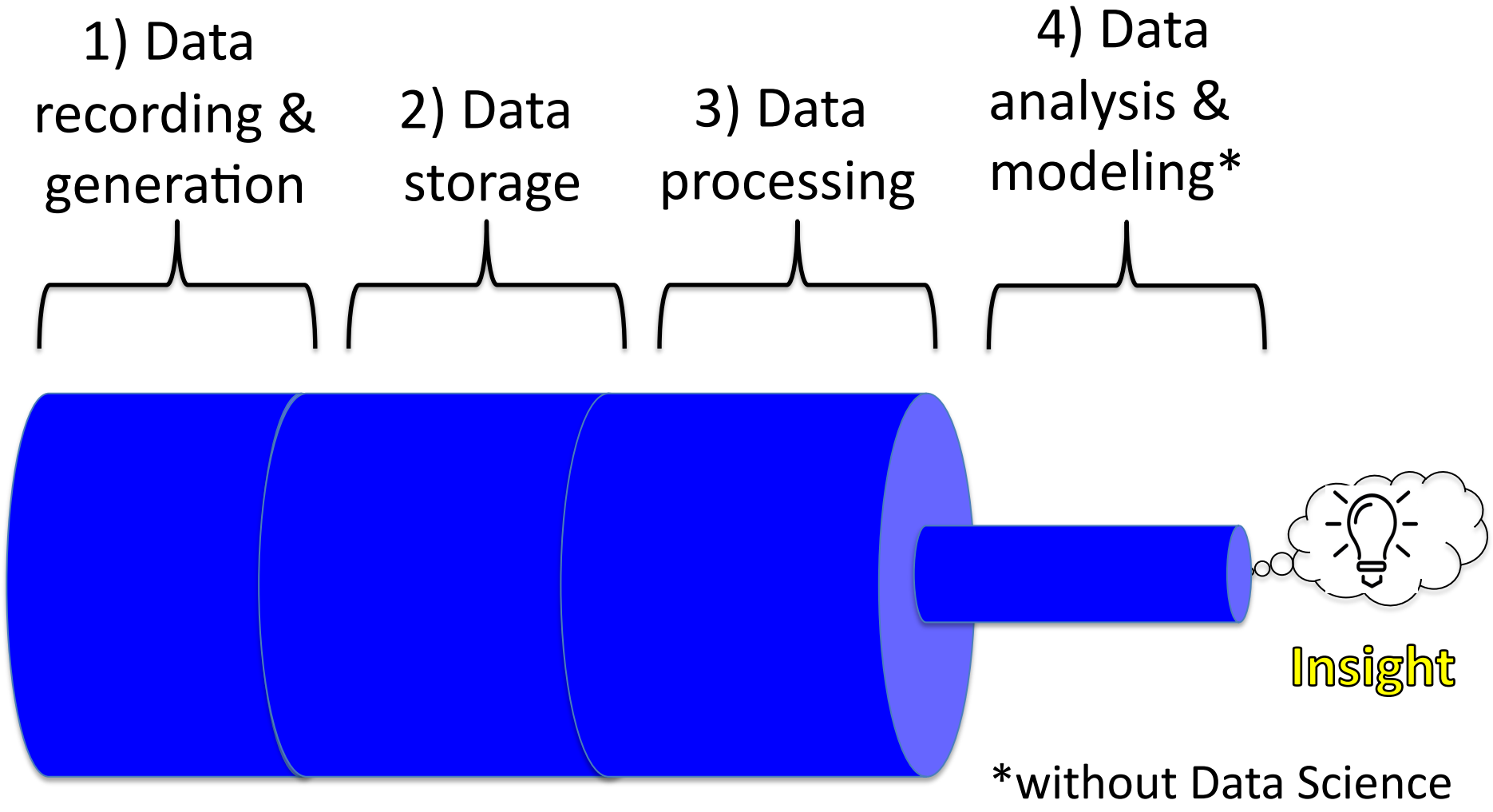


Now

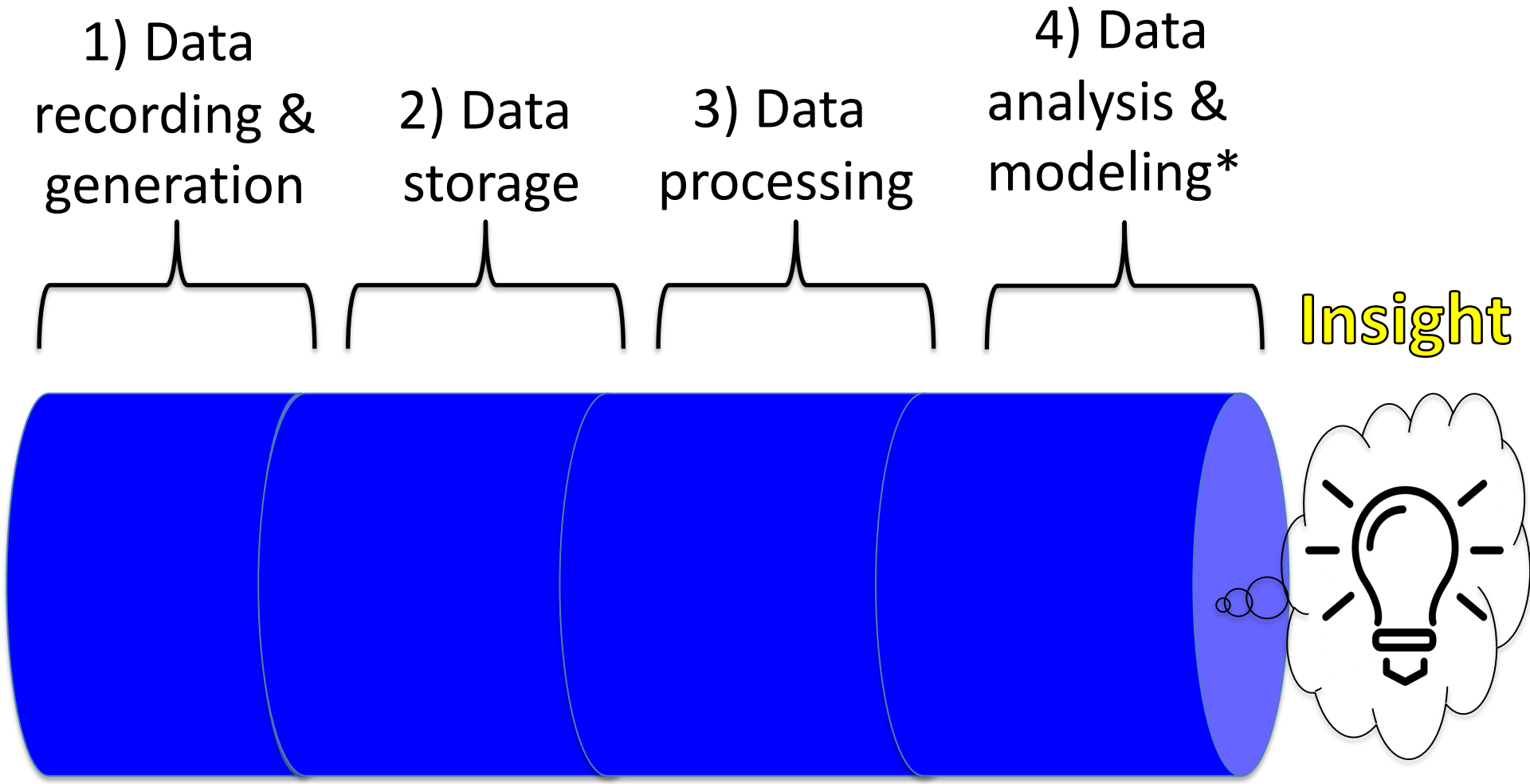
3) Data processing capabilities also increase exponentially (as described by Moore's law)



The modern data pipeline



The modern data pipeline

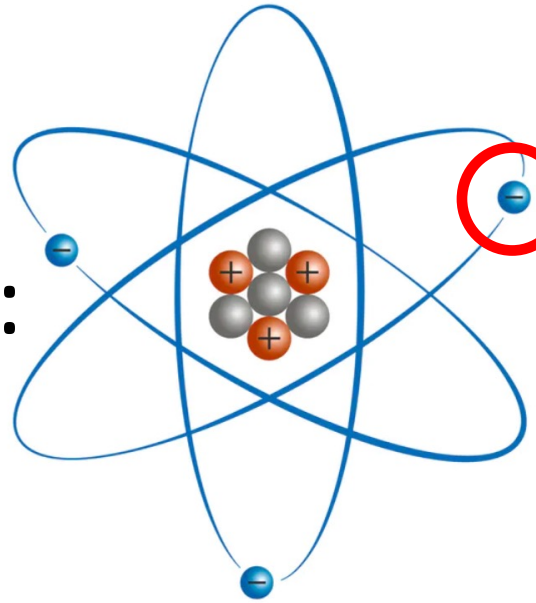


*with Data Science

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:



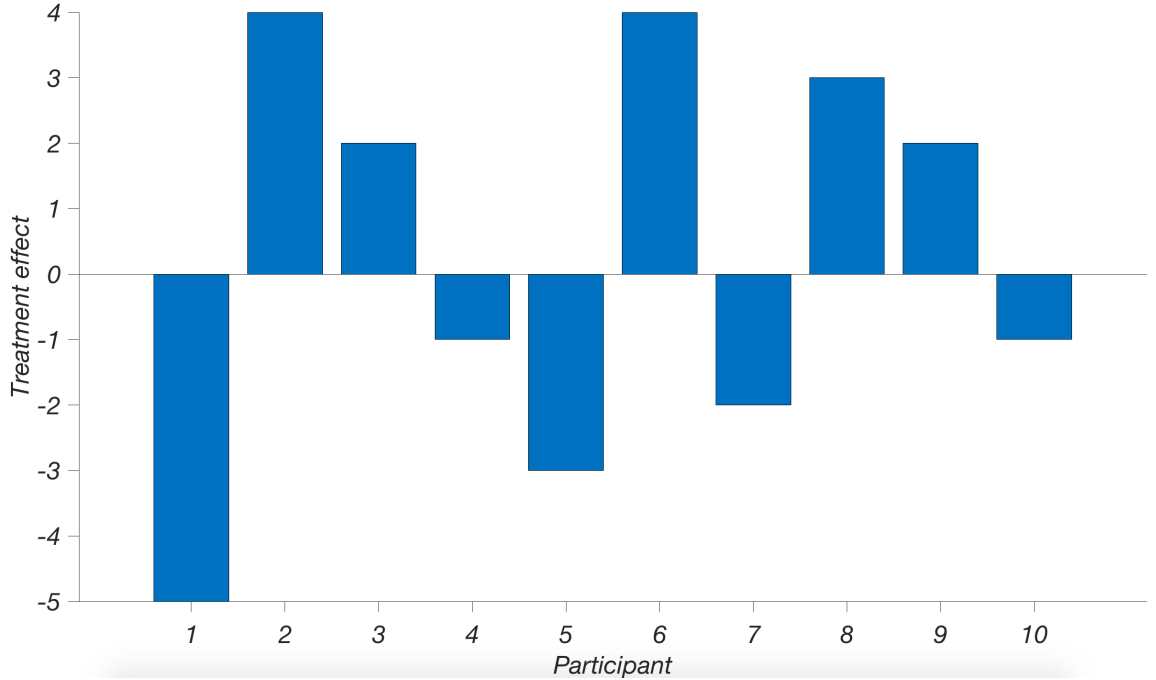
Electron(s)
Mass: 9.10938356 × 10-31 kg
Charge: -1.60217662 × 10-19 C
Spin: ½

Number: At least 10^{80}

They are all identical.

			Diplexity			

Microbiome:



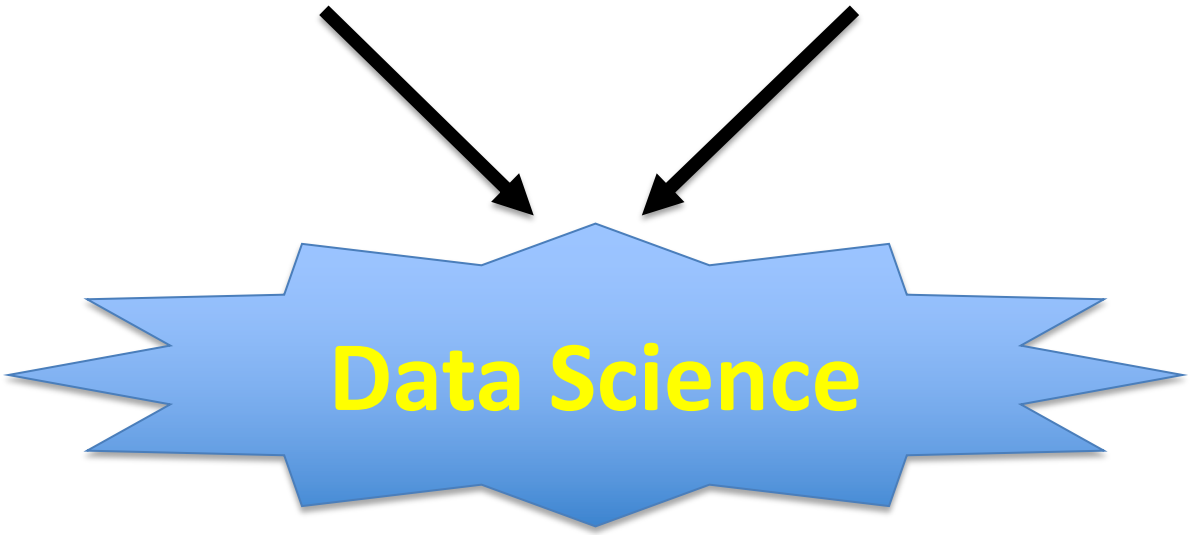
Data Science fills this niche

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 exactly a decade old

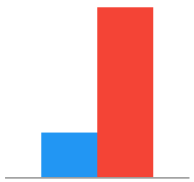
"data science"
Search term

neuroscience
Search term

Primus et optimus:

United States

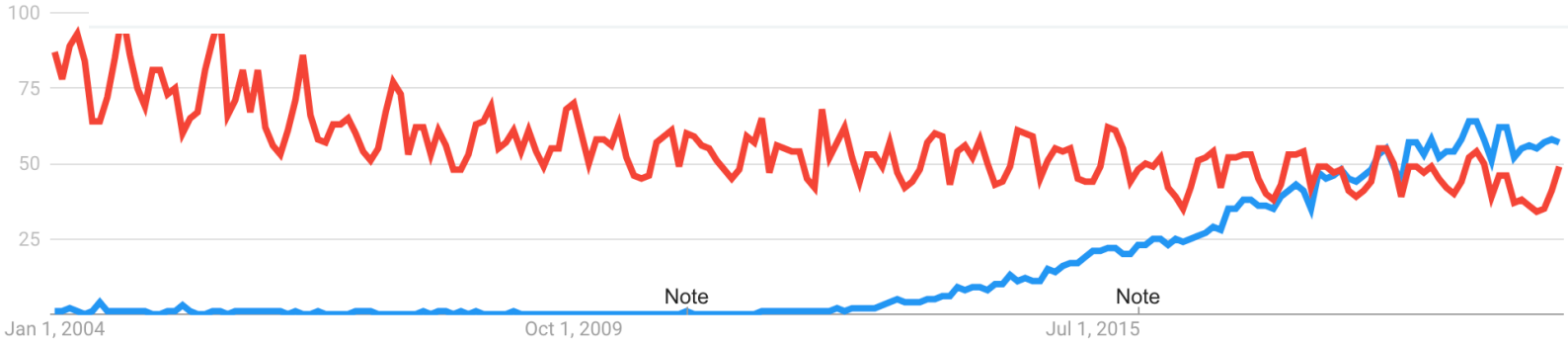
Interest over time



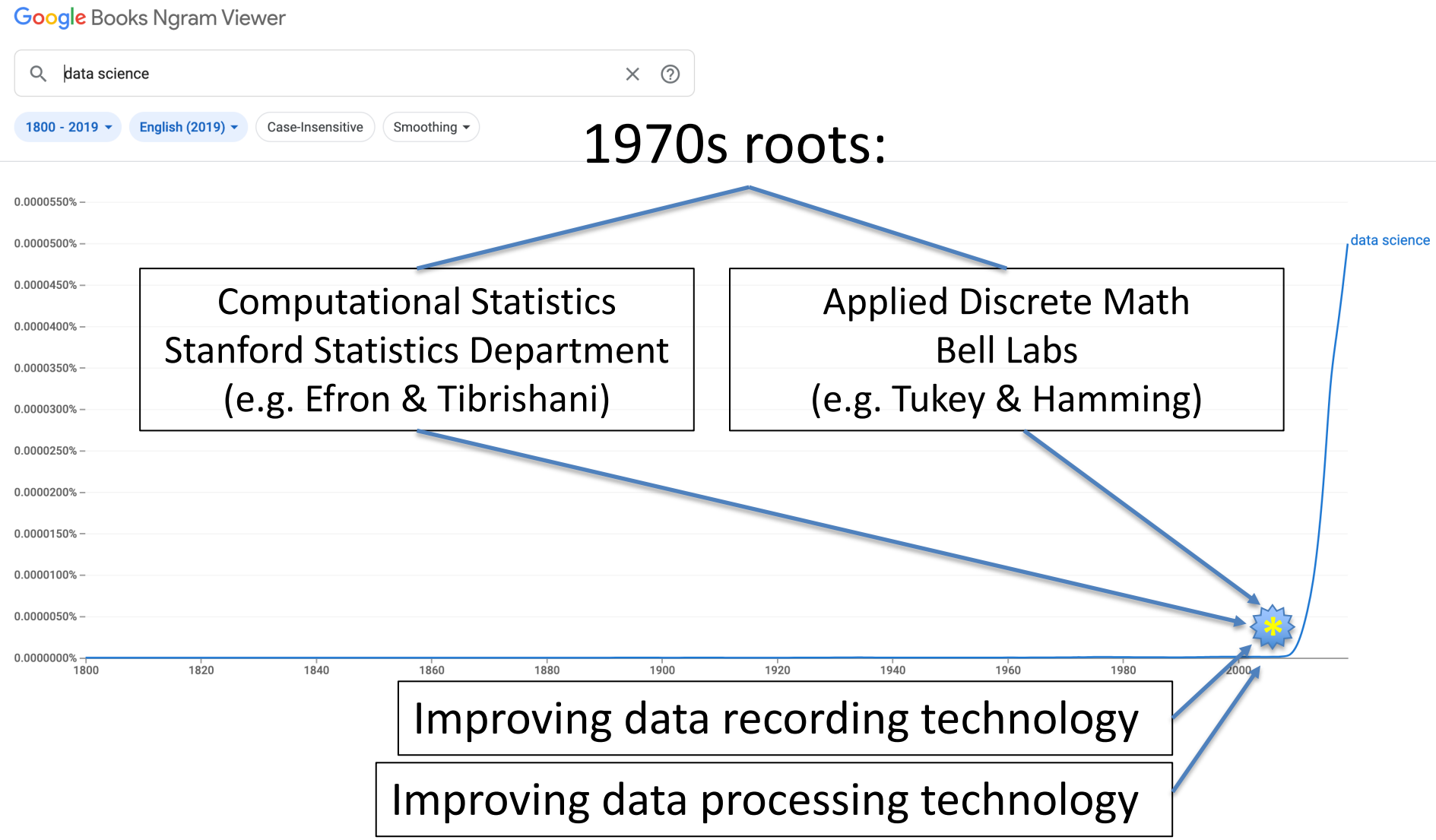
Pascal Wallisch @Pascallisch · Sep 1

Today marks the 10 year anniversary of the Center for Data Science (@NYUDataScience) at @nyuniversity, the first operational academic unit dedicated to data science in the world.

1 1 12 1,188



A genuinely 21st century enterprise



Data Science has
a long past,
a short history
and – likely - a
bright future

It's a different world now:
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

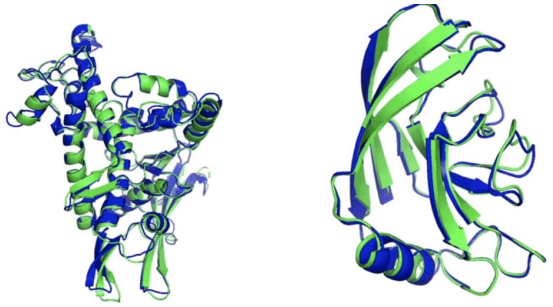
The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules

Data Science is already changing the world

EXPLICA.CO

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

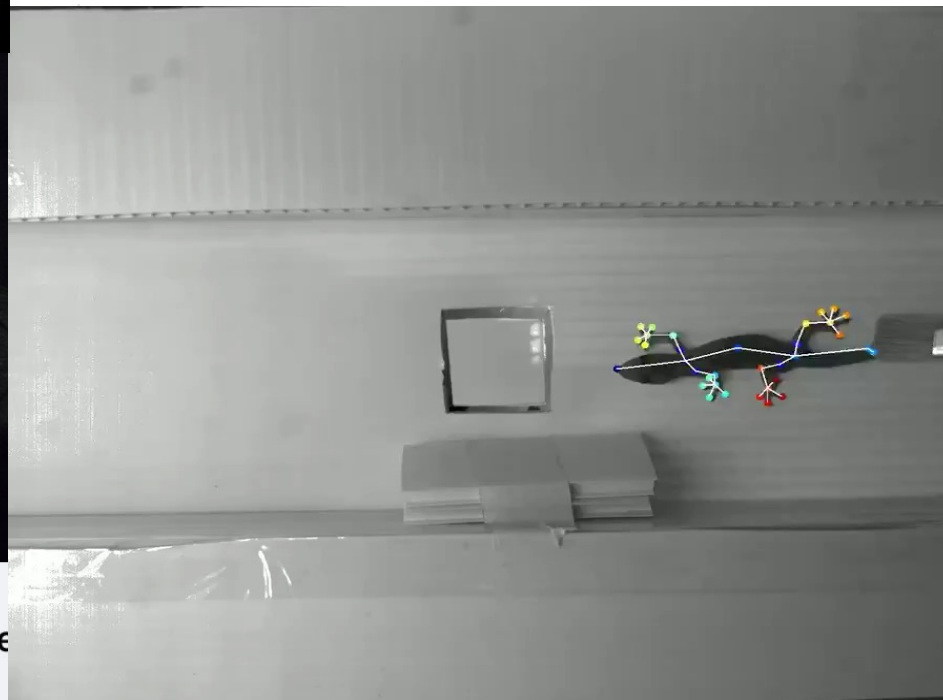


T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

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




So… once again:
Welcome aboard

That's it – for now

General questions?

Discussion?