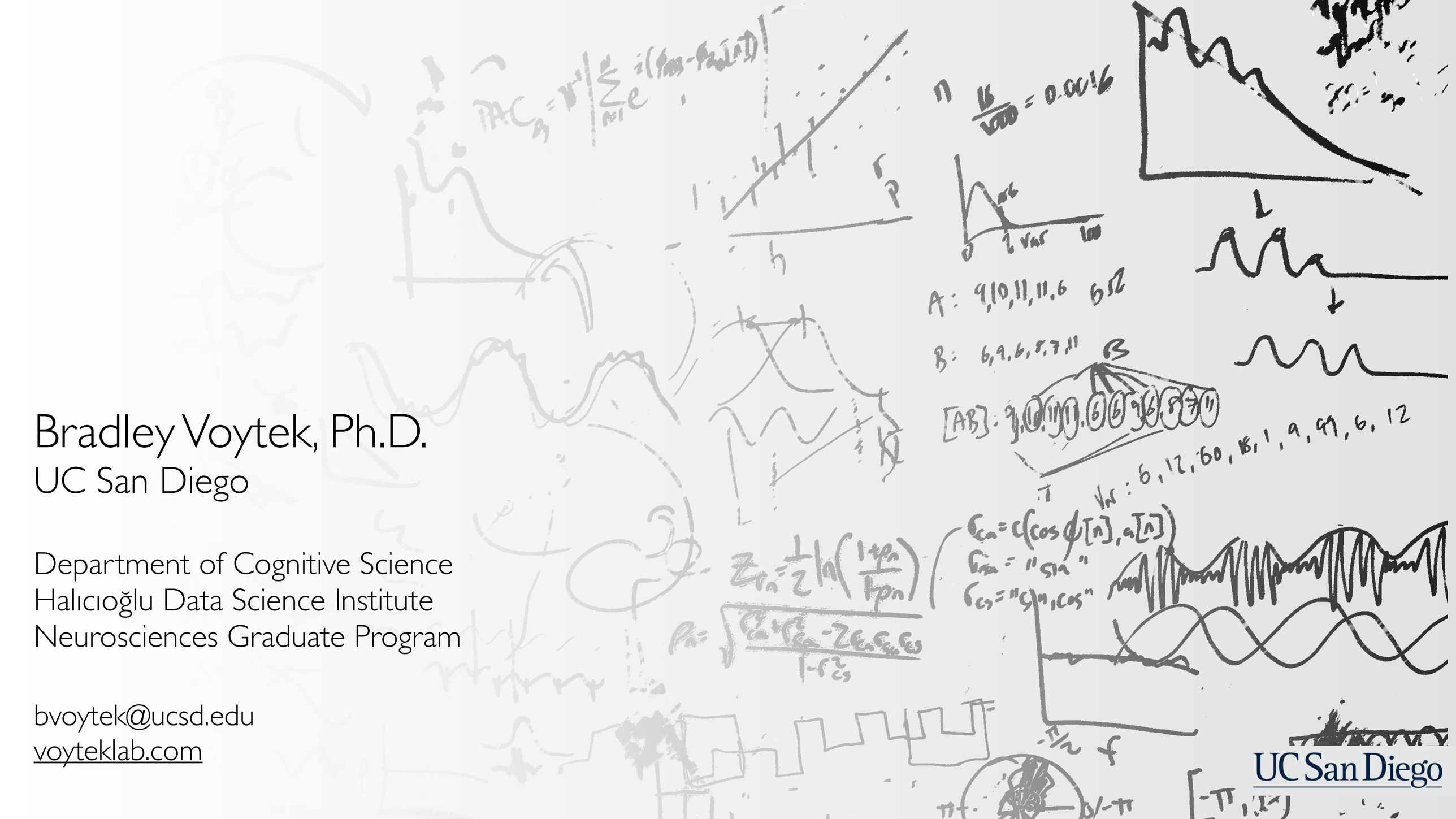
Bradley Voytek, Ph.D.

UC San Diego

Department of Cognitive Science
Halıcıoğlu Data Science Institute
Neurosciences Graduate Program

bvoytek@ucsd.edu
voyteklab.com



UC San Diego

Office hour updates

Date & Time	Location	Instructional Staff	email
*** Tuesdays 11:00a-12:00p	CSB 169	Prof. Bradley Voytek	bvoytek@ucsd.edu
* Thursdays 8:15 - 9:15 am	Zoom (link)	TA: Eena Kosik	ekosik@ucsd.edu
** Mondays 9:00-10:00a	Mandeville Coffee Cart	TA: Yueying Dong	yud070@ucsd.edu
* Thursdays 4:00-5:00p	Zoom (link)	TA: Kaushik Madhavan	kmadhavan@ucsd.edu
Tuesdays 3:00-4:00p	Mandeville Coffee Cart	IA: Hannah Song	hsong@ucsd.edu
Tuesdays 2:00-3:00p	Mom's Café	IA: Jingyi Wang	jiw109@ucsd.edu
Fridays 11:15a-12:15p	TBD	IA: Abhay Anand	a7anand@ucsd.edu
Tuesdays 1:00-2:00p	TBD	IA: Shivani Kedila	skedila@ucsd.edu

COGS 9
Introduction to Data Science

UC San Diego

Where is “cognition” in data science?

Cognition in Data Science

By turning many observations into numbers we can fit equations to them, or learn relationships between them, which lets us understand how observations relate and sometimes guess what might happen in the future.

These are all *cognitive phenomena!*

Data Science at UC San Diego

UC San Diego Data Science

UC San Diego

Data Science at UCSD

Facebook pioneer donates $75 million to UCSD for data science

Taner Halicioglu is donating $75 million to UC San Diego to make his alma mater a national leader in data science. (Erik Jepsen / UC San Diego)

Final Project examples: UCSD grade optimizer

Out[37]:

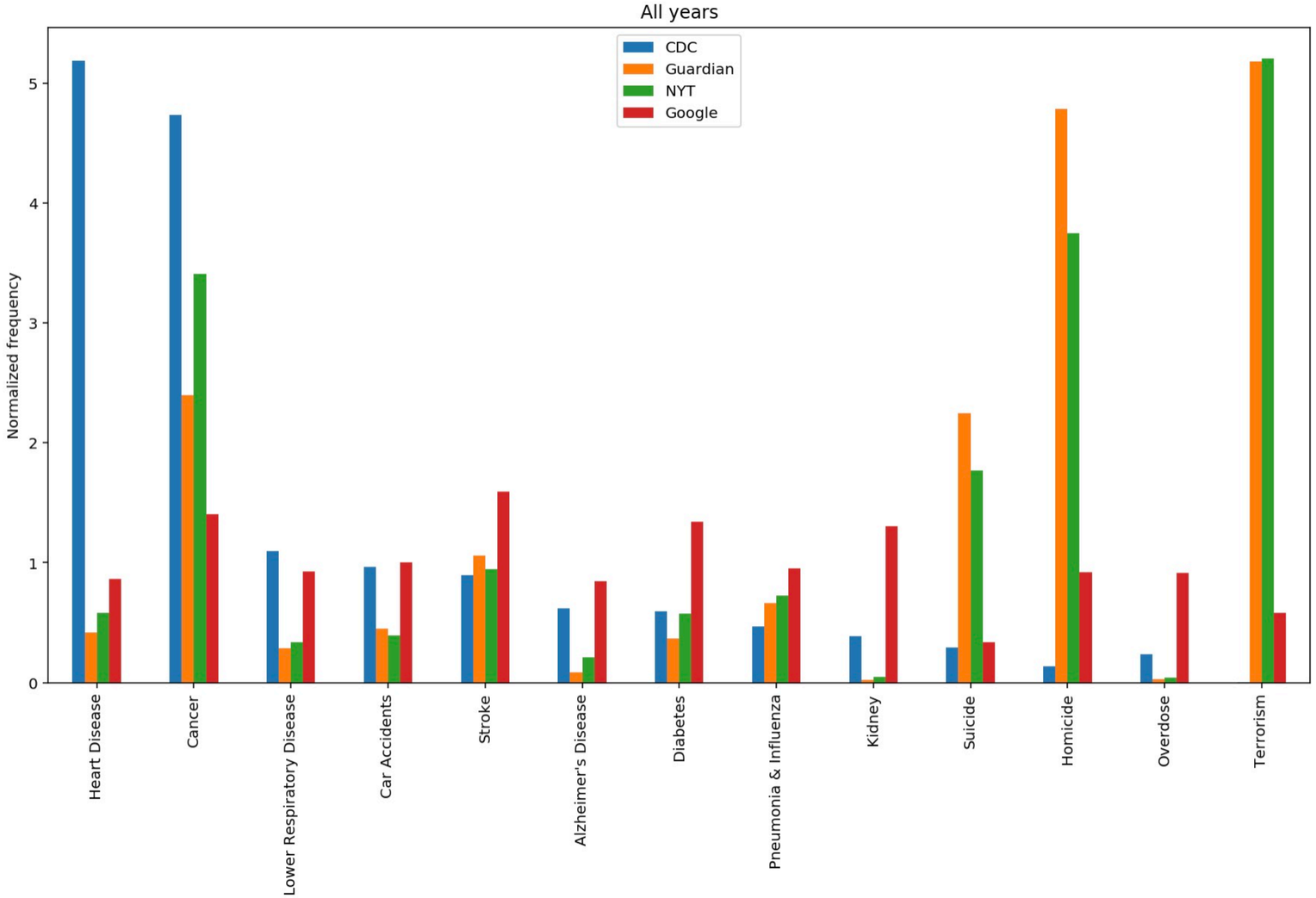
	course	0	1	2	3	4	5	6	7	8	...	990	991	992	993	994	995	996	997	998	999
0	MATH+20A	2.38	3.21	2.96	2.93	2.62	3.05	2.58	3.02	2.77	...	2.53	3.22	2.91	2.60	2.88	2.62	2.38	2.76	2.80	2.69
1	MATH+20B	2.89	2.21	3.04	3.22	3.34	3.02	2.82	2.53	3.21	...	3.05	2.53	3.39	3.19	2.95	3.04	2.86	3.39	2.77	3.35
2	MATH+20C	2.76	2.69	2.62	2.35	3.21	2.90	2.50	2.76	3.34	...	2.82	2.83	3.33	2.50	3.45	2.62	3.01	2.79	3.02	3.18
3	MATH+20F	3.22	2.63	3.03	3.02	3.31	2.70	3.22	3.17	2.50	...	2.74	2.82	3.33	2.58	2.62	3.17	2.77	2.75	2.95	2.68
4	COGS+1	3.34	2.40	3.19	2.58	3.45	2.56	2.80	3.08	2.76	...	2.66	2.53	2.53	2.71	3.47	3.33	3.52	2.79	2.63	2.90
5	COGS+14A	2.62	3.01	3.25	3.05	3.22	2.93	2.83	2.77	2.81	...	2.83	2.39	2.89	3.08	2.77	3.20	2.35	2.77	3.31	2.21
6	COGS+101A	2.50	2.77	2.90	3.22	2.74	3.19	3.19	2.76	2.95	...	3.21	3.45	3.38	2.52	2.41	2.81	2.53	2.90	2.50	2.80
7	COGS+102A	2.76	2.75	2.89	2.56	2.56	3.15	2.58	2.75	2.70	...	2.52	3.01	3.06	2.80	3.04	2.63	3.45	2.50	3.39	3.16
8	COGS+107A	2.77	3.20	2.76	3.02	2.46	2.38	2.44	2.53	3.33	...	2.80	2.58	2.33	2.87	2.96	2.26	2.83	2.68	3.05	3.02
9	CSE+7	2.90	3.04	2.74	2.58	2.90	2.52	3.39	2.41	3.47	...	3.24	2.62	2.71	3.02	3.22	2.57	2.50	2.95	2.53	2.46

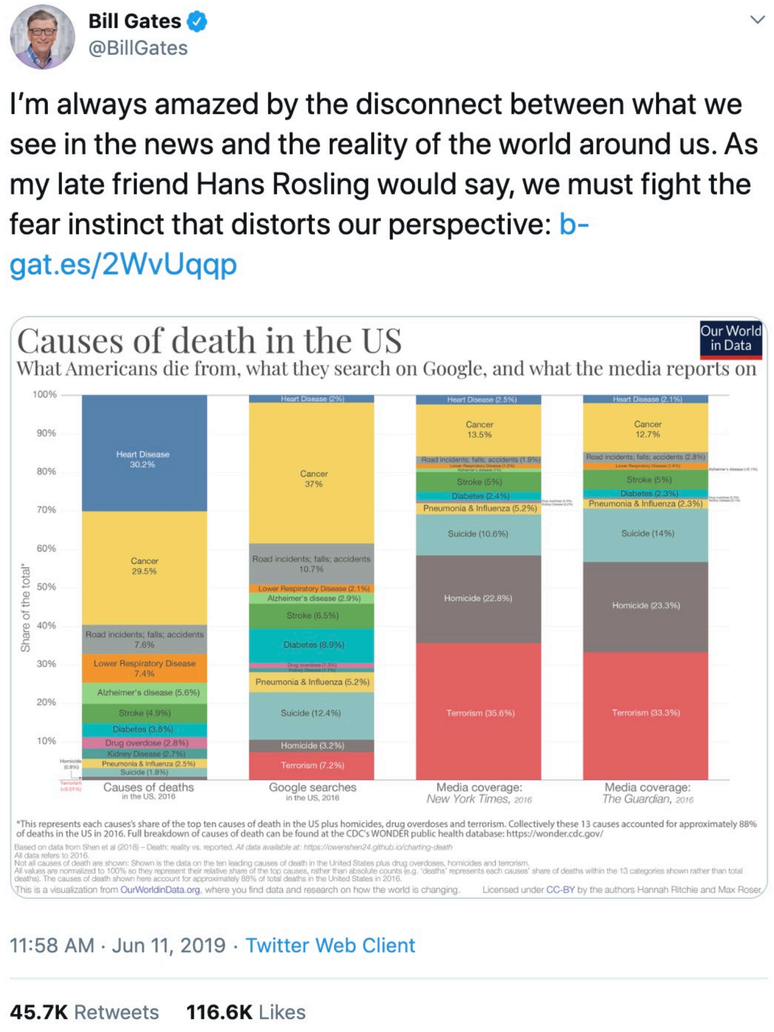
10 rows × 1001 columns

Create a dataframe filled with 1000 random student GPAs from students who have taken a bad professor from 10 different courses.

UC San Diego

Final Project examples: Death



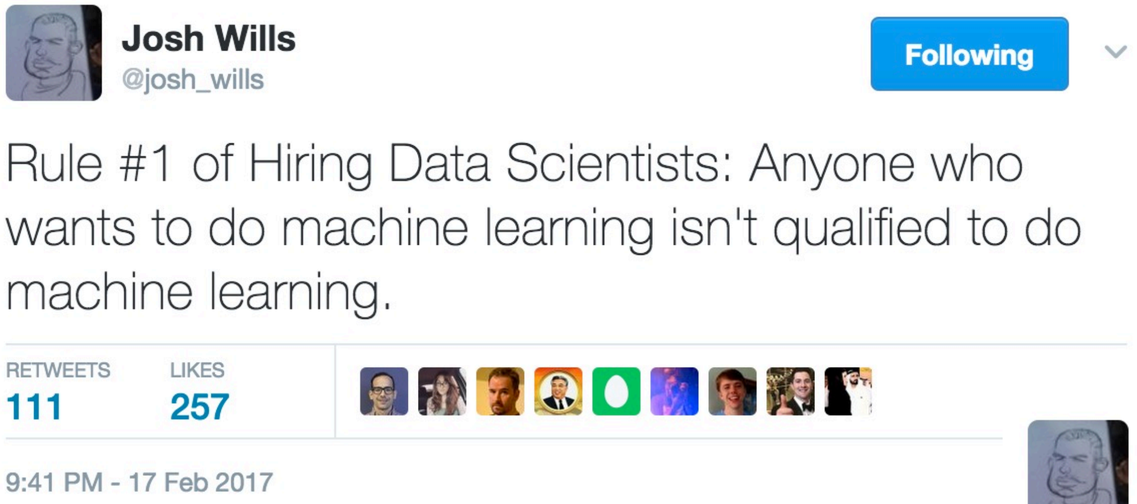


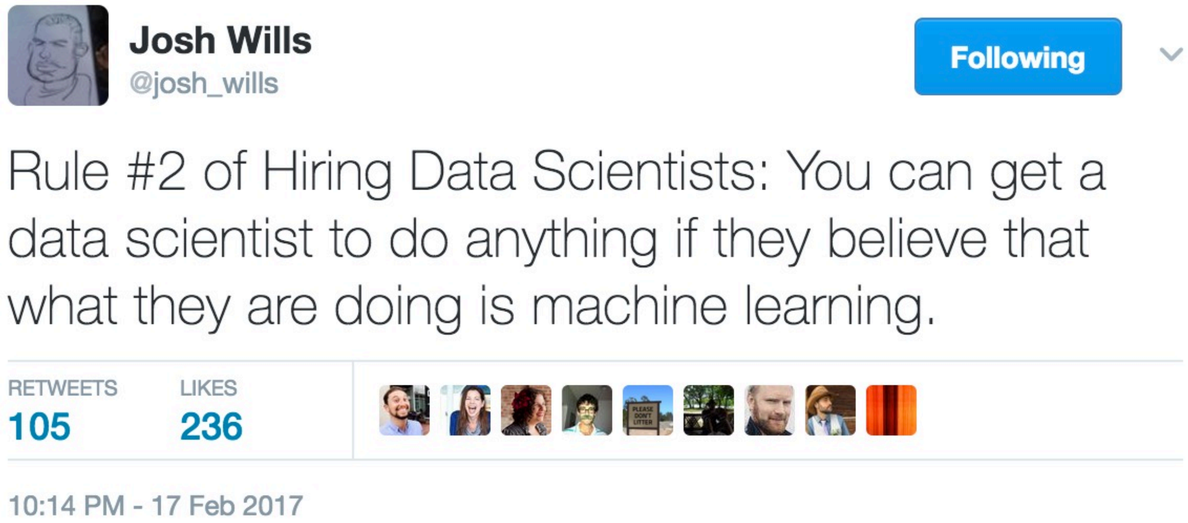
Source: Owen Shen, Hasan Al-Jamaly, Maximillian Siemers, and Nicole Stone.

UC San Diego

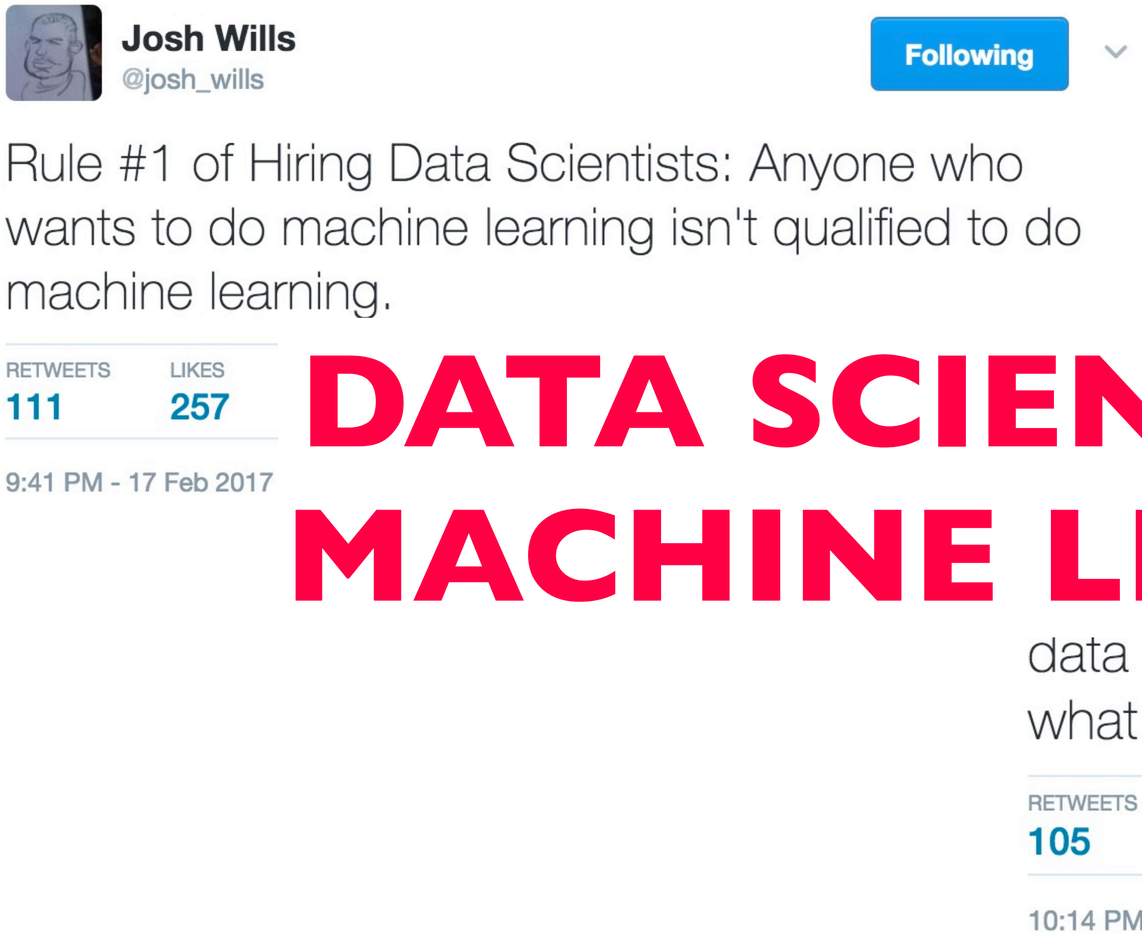
Okay data science is cool, but...?

Data Science





Data Science



Data science practicum

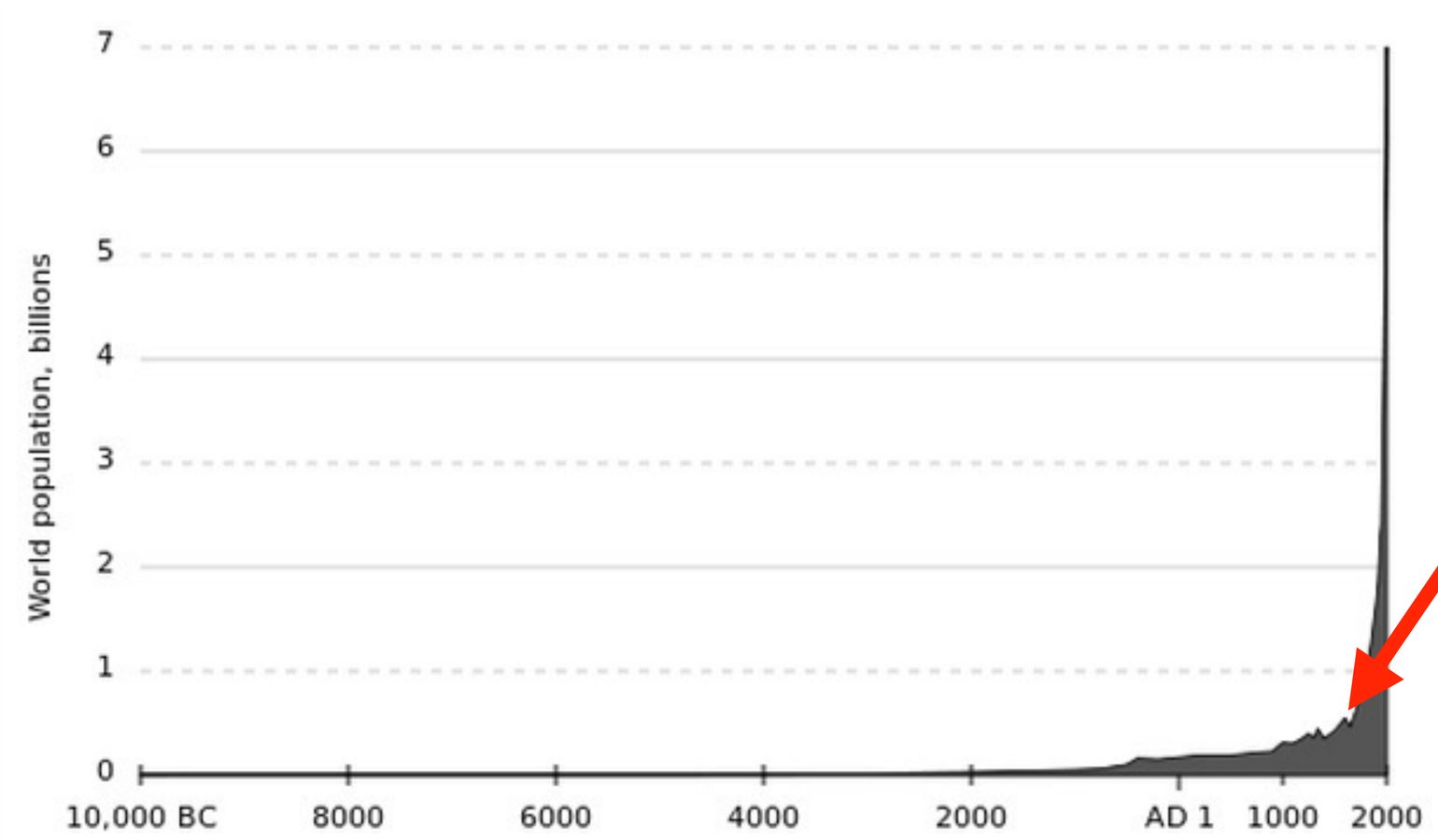
How many people were born in Great Britain on September 10, 1752?

TABLE 1. WORLD POPULATION, YEAR 0 TO NEAR STABILIZATION

Year	*Population (in billions)*
0	0.30
1000	0.31
1250	0.40
1500	0.50
1750	0.79
1800	0.98
1850	1.26
1900	1.65
1910	1.75
1920	1.86
1930	2.07
1940	2.30
1950	2.52
1960	3.02
1970	3.70
1980	4.44
1990	5.27
1999	**5.98**
2000	6.06
2010	6.79
2020	7.50
2030	8.11
2040	8.58
2050	8.91
2100	9.46
2150	9.75
Near stabilization (after 2200)	Just above 10 billion

Source: United Nations Population Division.

Data science practicum



$\Delta P = B - D$

divided by the proportion of world population living in Great Britain...

Data science practicum

::drumroll::

UC San Diego

Data science practicum

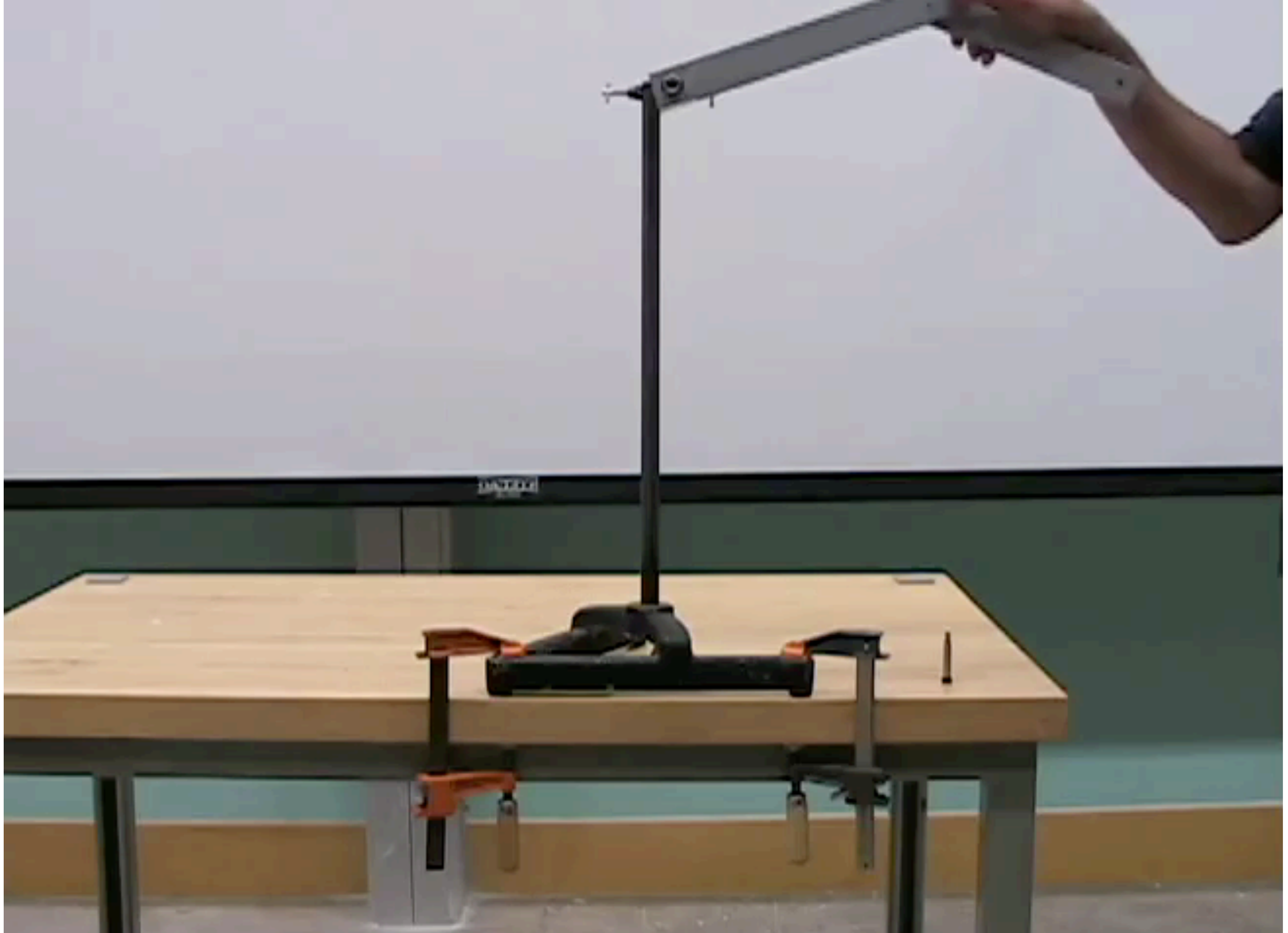
ZERO!

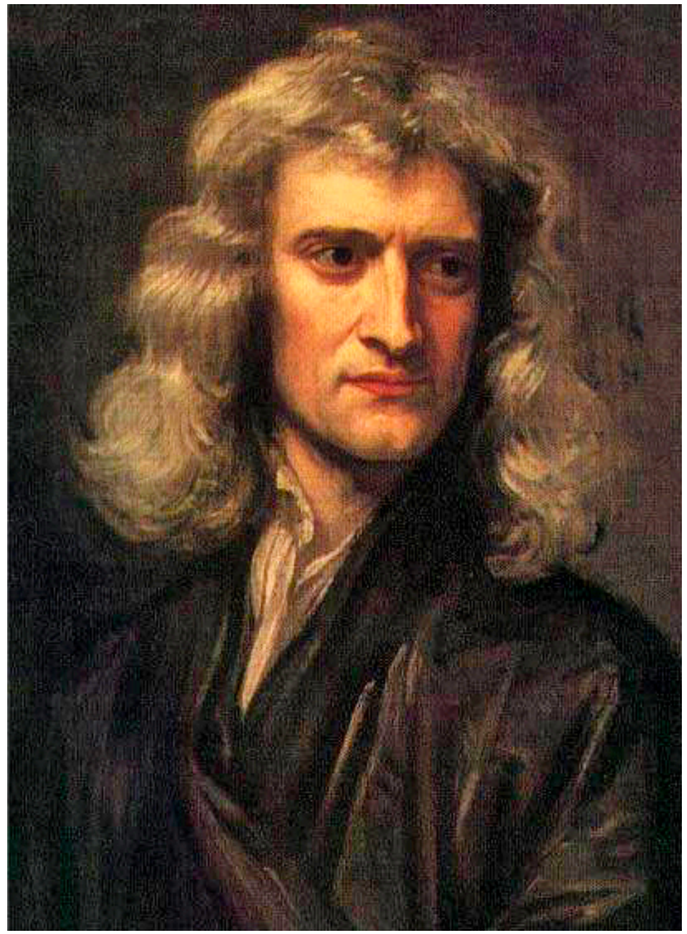
It was impossible for any citizen of the British Empire to have been born between September 3 and September 13 (inclusive) 1752.

This is because the Empire adopted the Gregorian Calendar after having been on the Julian.

By decree of the Calendar (New Style) Act 1750, September 2, 1752 was followed by September 14, 1752.

Data *cannot* be interpreted outside of their *context*





Sir Isaac Newton

$$\mathbf{F} = m\mathbf{a}$$

force = mass*acceleration

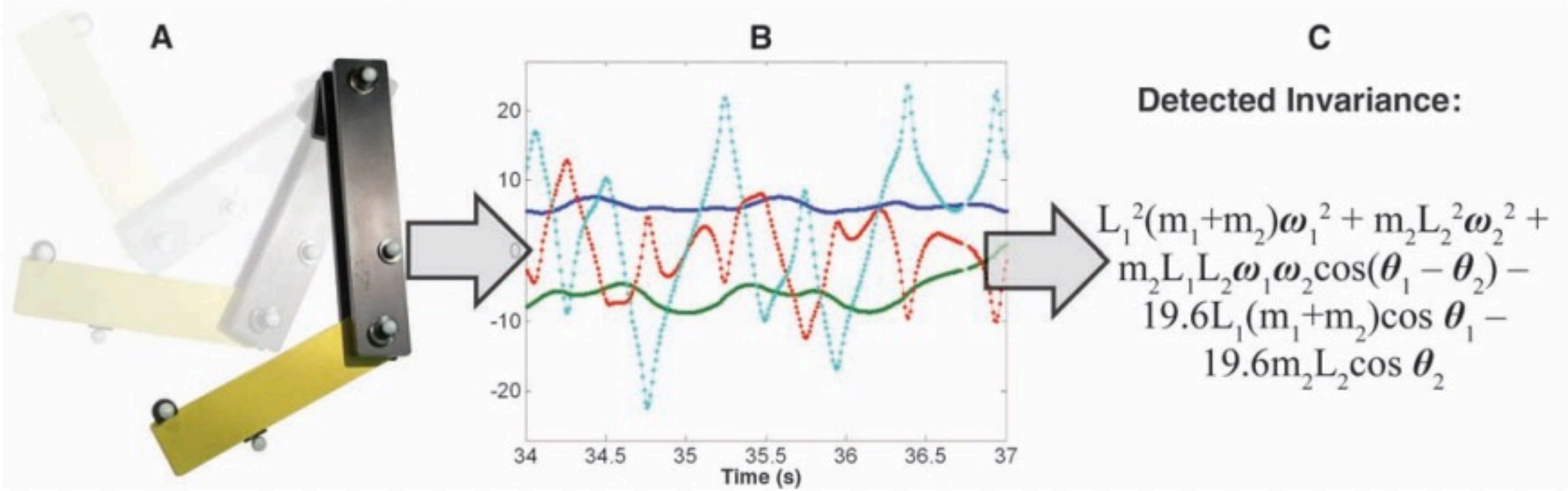


Fig. 1. Mining physical systems. We captured the angles and angular velocities of a chaotic double-pendulum (**A**) over time using motion tracking (**B**), then we automatically searched for equations that describe a single natural law relating these variables. Without any prior knowledge about physics or geometry, the algorithm found the conservation law (**C**), which turns out to be the double pendulum's Hamiltonian. Actual pendulum, data, and results are shown.

Source: Schmidt & Lipson, *Science* 2009

"We have demonstrated the discovery of physical laws, from scratch, directly from experimentally captured data with the use of a computational search… without prior knowledge about physics, kinematics, or geometry"

Source: Schmidt & Lipson, *Science* 2009

"The concise analytical expressions that we found are amenable to human interpretation and help to reveal the physics underlying the observed phenomenon. Many applications exist for this approach, in fields ranging from systems biology to cosmology, where theoretical gaps exist despite abundance in data."

Source: Schmidt & Lipson, *Science* 2009

"Might this process diminish the role of future scientists? Quite the contrary: Scientists may use processes such as this to help focus on interesting phenomena more rapidly and to interpret their meaning."

Source: Schmidt & Lipson, *Science* 2009

Where is "cognition" in Data Science?

Data Science is different from "using data to come to conclusions in science".

Where is "cognition" in Data Science?

The study of data: how data are collected, what can be learned from data, and how can data be biased.

Where is "cognition" in Data Science?

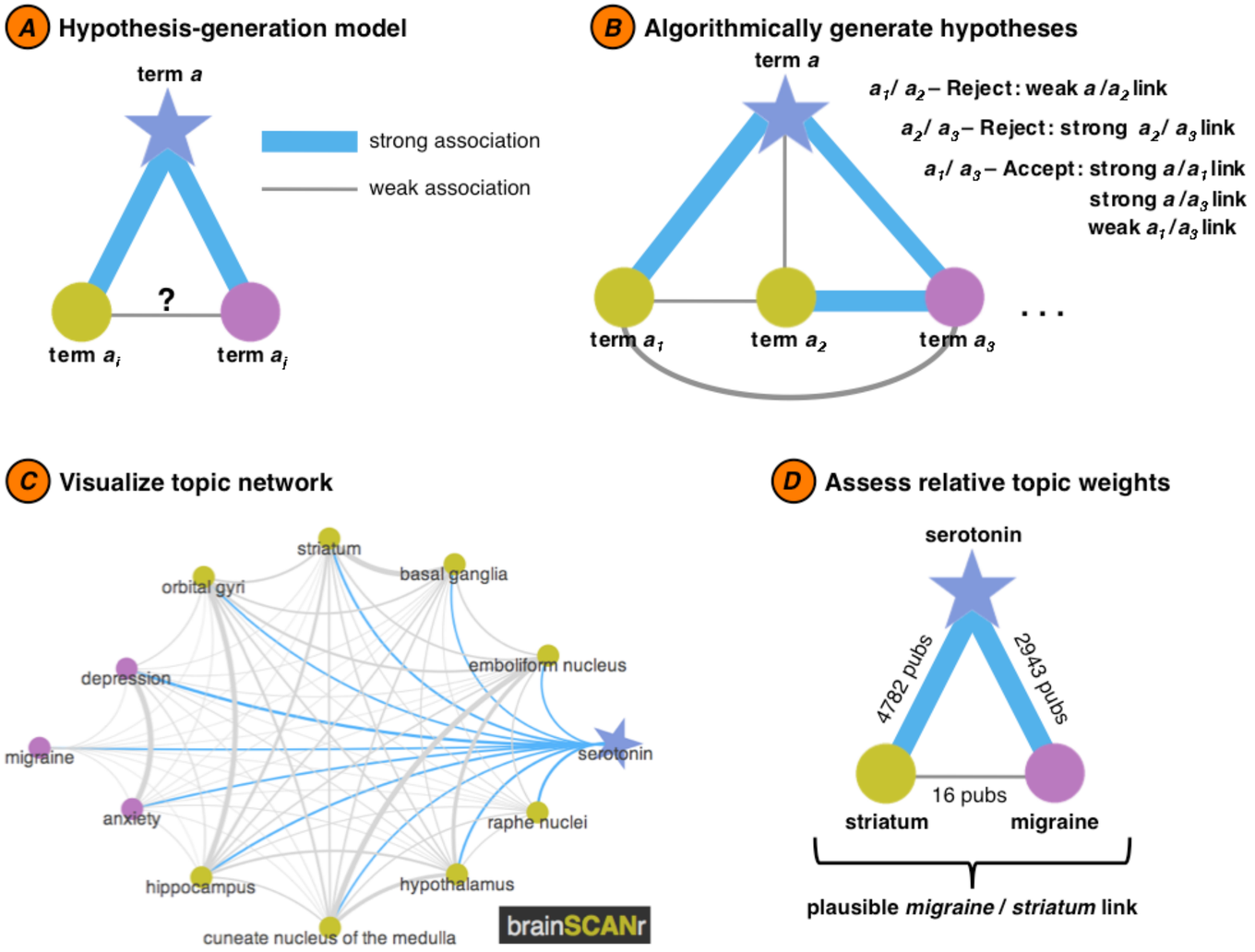
The study of data: how data are collected, what can be learned from data, and how data can bias.

These are all *cognitive phenomena!*

Let's step back for a second and talk about data-driven approaches to **knowledge discovery**: using data to guide us to new hypotheses and ideas.

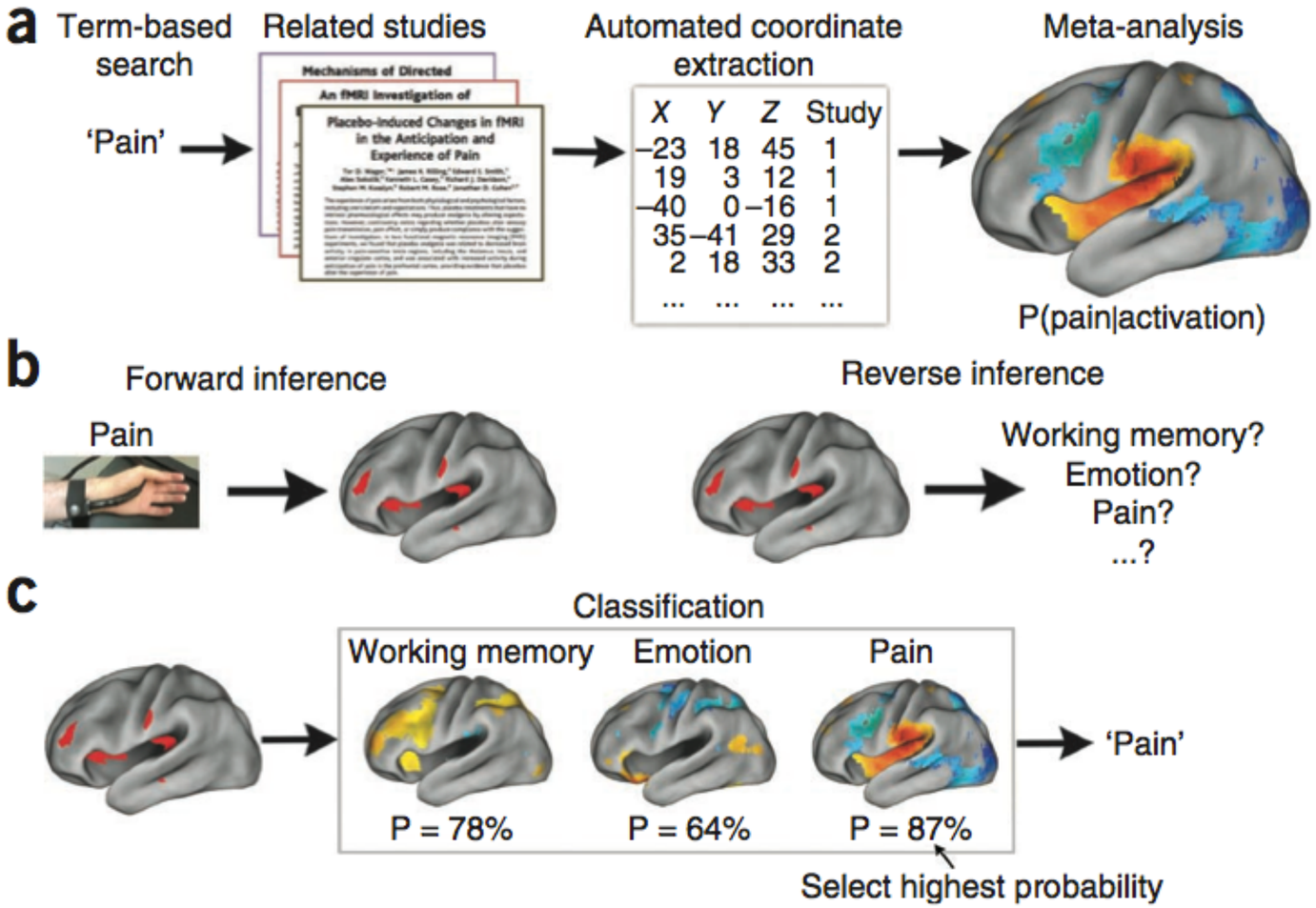
Note that data-driven approaches to **knowledge discovery** are analogous to *the human process of hypothesis generation*, but automated.

brainSCANr



Source: Voytek & Voytek, *J Neurosci Methods* 2012

NeuroSynth



Source: Yarkoni et al., *Nat Methods* 2011

These kinds of approaches can be **powerful**.

These kinds of approaches can be **powerful**.

But they can also *recapitulate biases*.

Biased data

brainSCANr: brain regions and neurotransmitters

NeuroSynth: brain regions and expected results

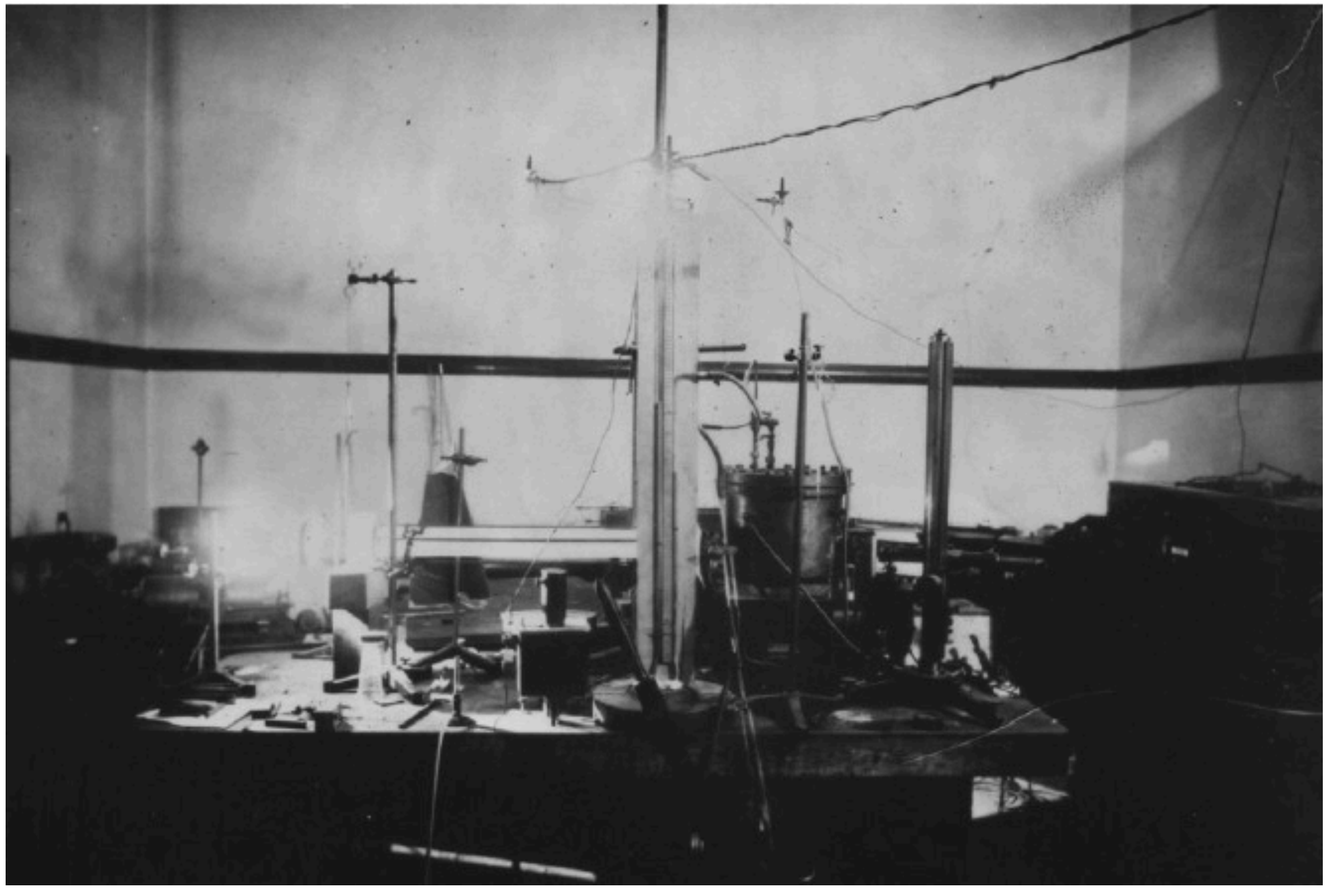
Biased data

Essentially "the rich get richer," or:

ideas that are over-represented due to biases get algorithmically surfaced as being more important.

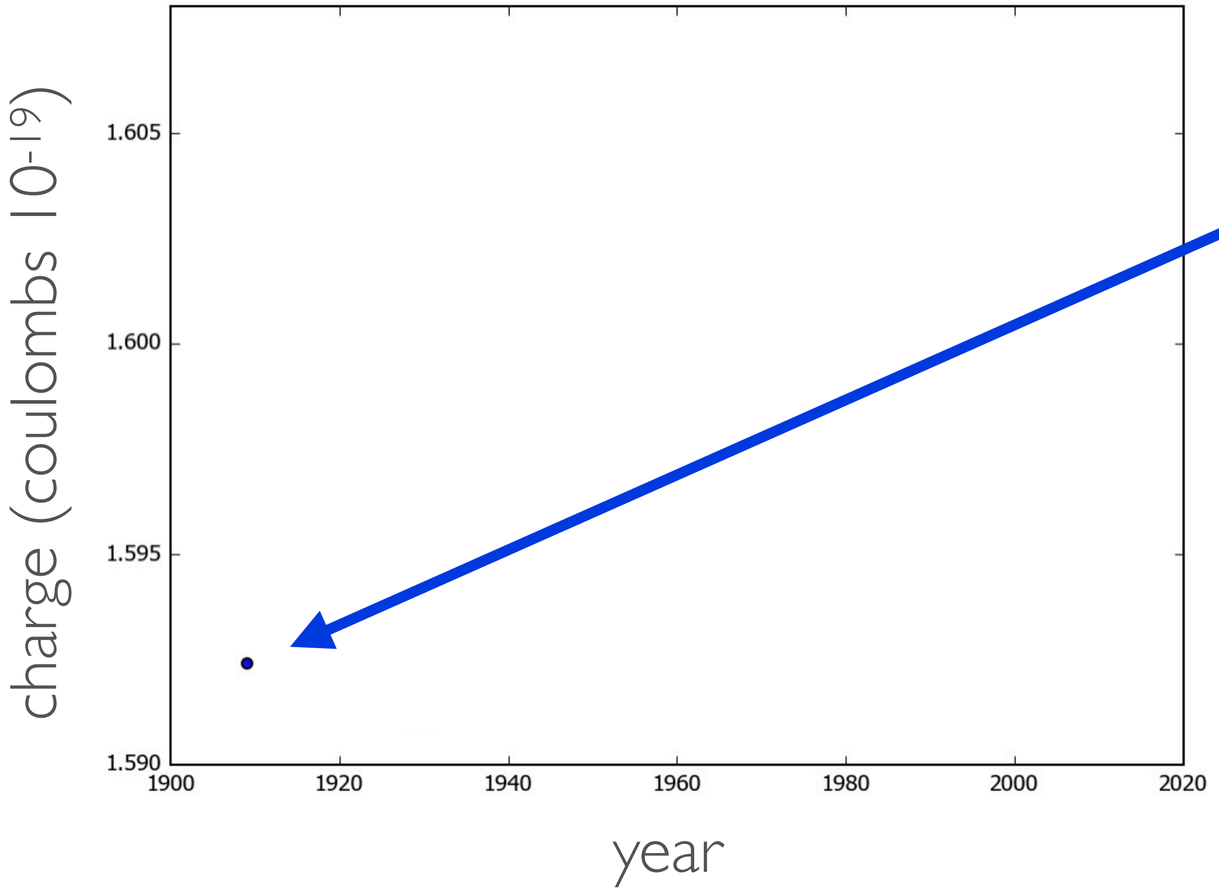
Millikan Oil Drop Experiment

Clever experiment to measure the charge of an electron



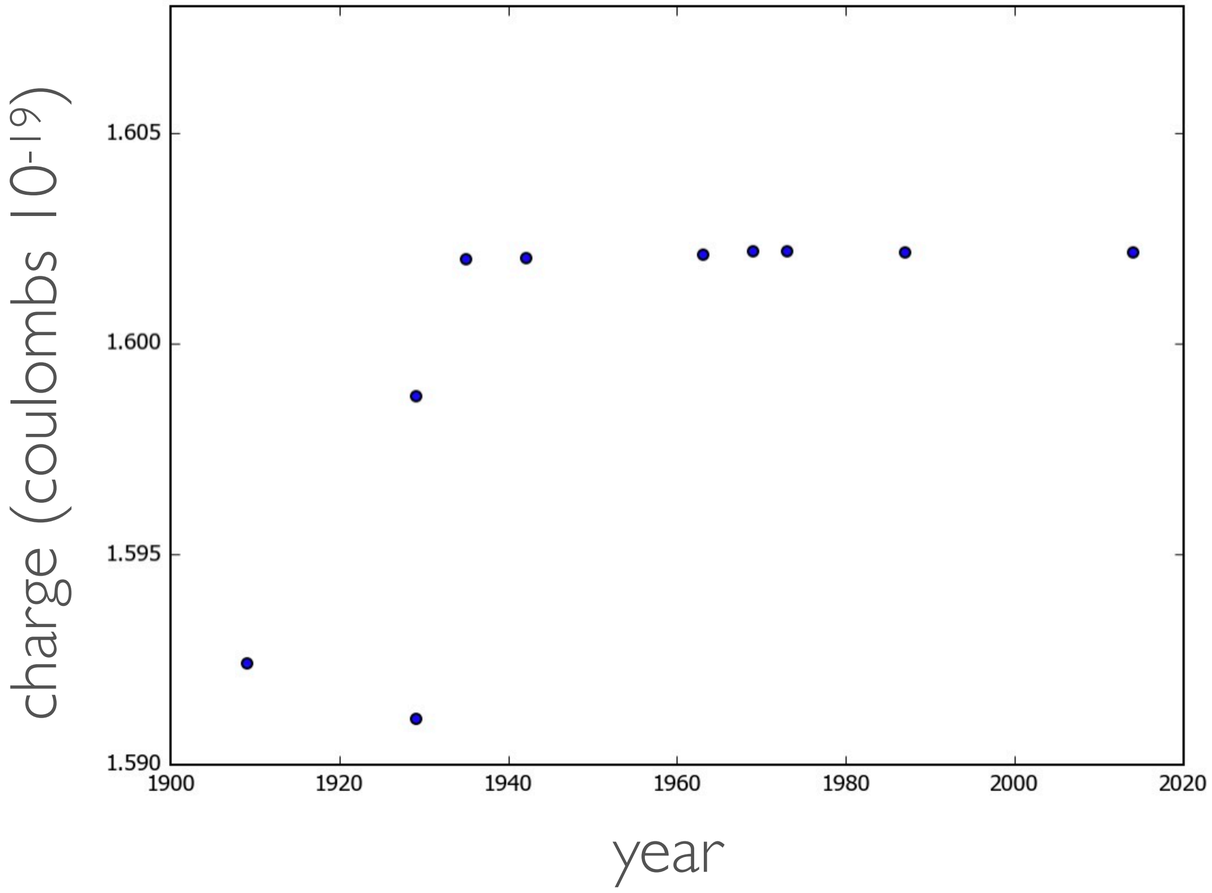
Source: *Wikimedia*

Millikan Oil Drop Experiment



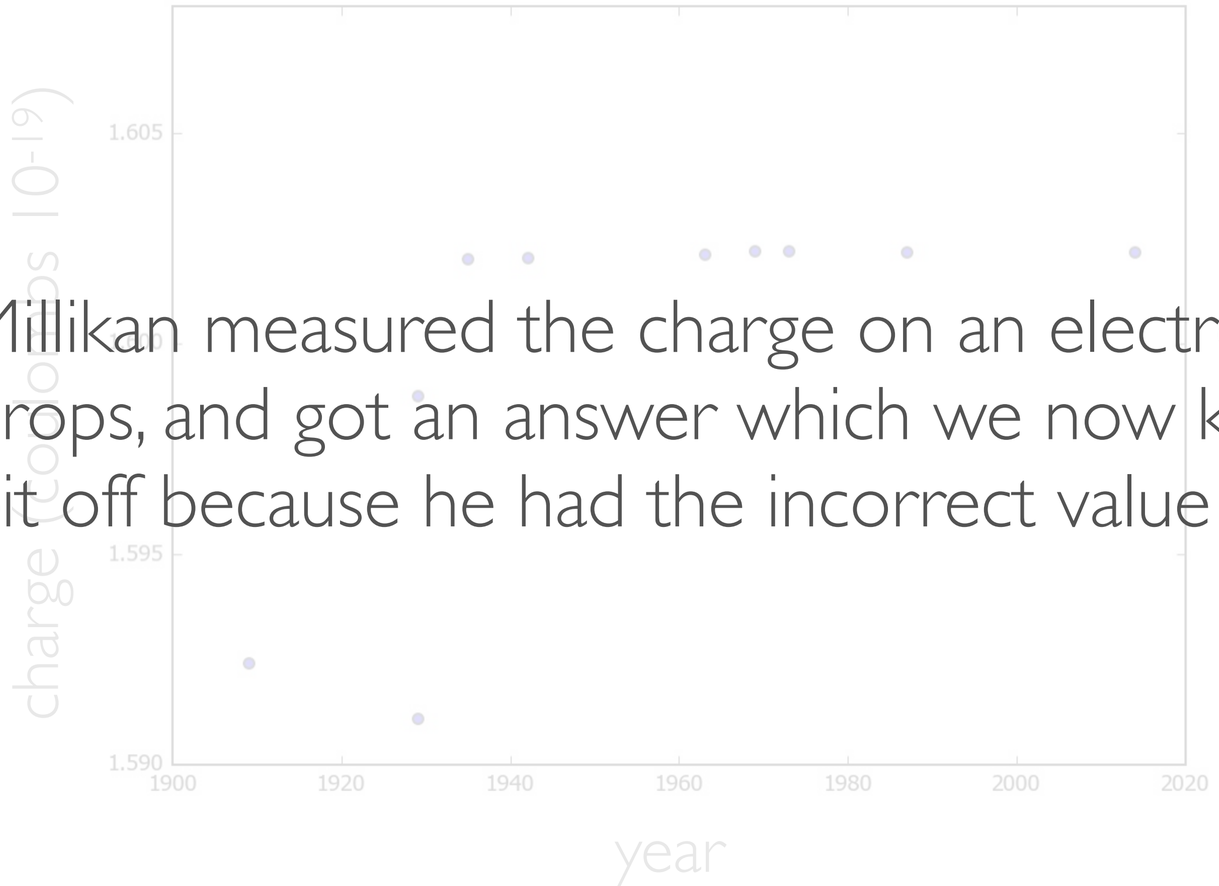
1909 Millikan value:
$1.5924(17) \times 10^{-19}$ C

Millikan Oil Drop Experiment



Source: *StackExchange*

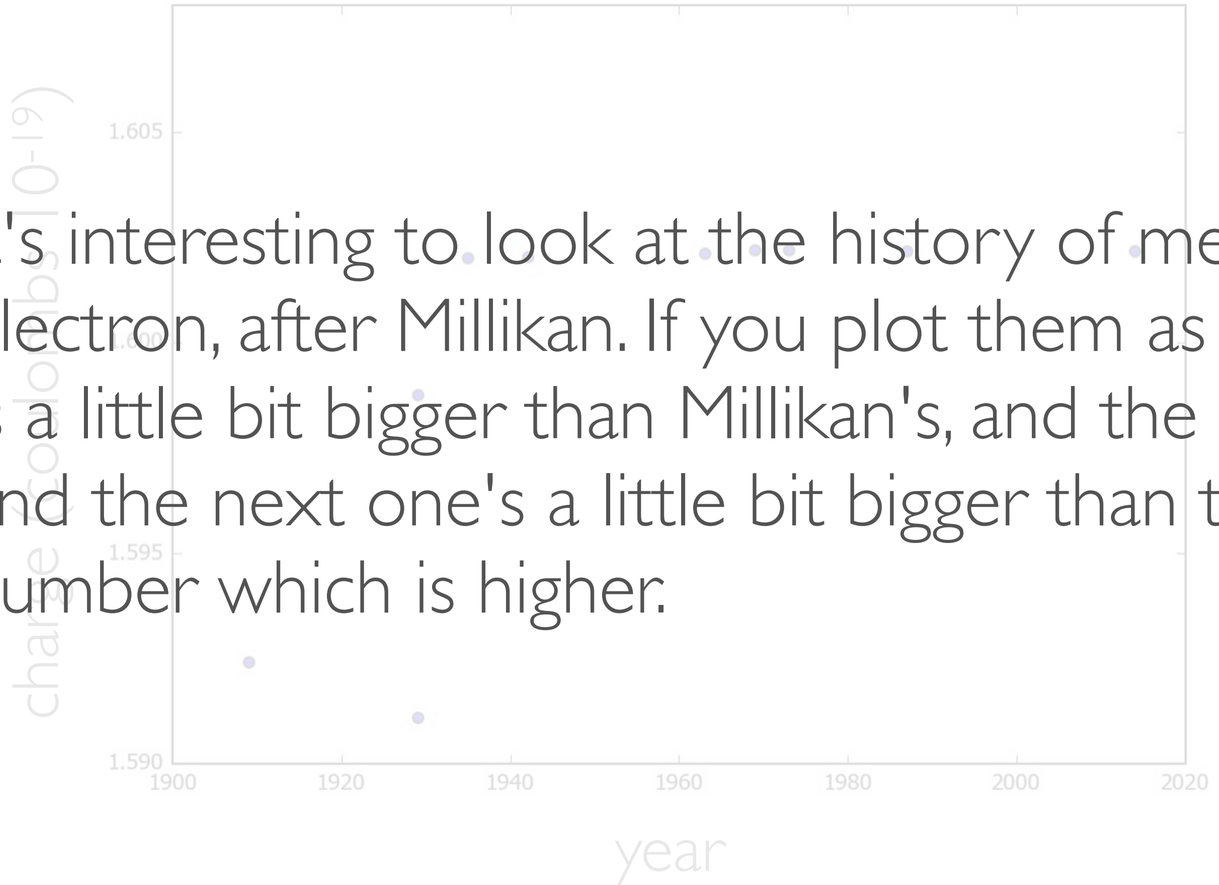
Millikan Oil Drop Experiment



Millikan measured the charge on an electron by an experiment with falling oil drops, and got an answer which we now know not to be quite right. It's a little bit off because he had the incorrect value for the viscosity of air.

Source: Richard Feynman, *Surely You're Joking, Mr. Feynman!*

Millikan Oil Drop Experiment



It's interesting to look at the history of measurements of the charge of an electron, after Millikan. If you plot them as a function of time, you find that one is a little bit bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher.

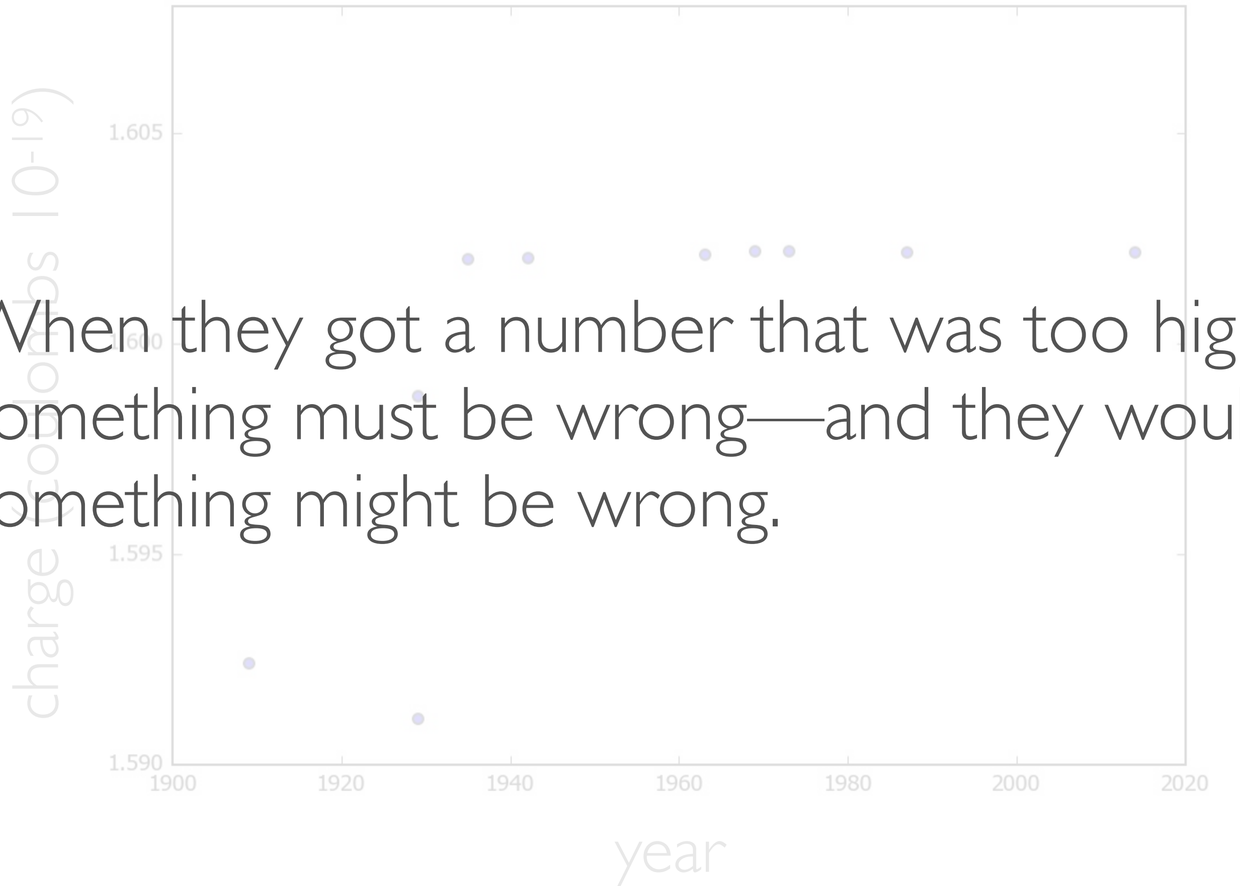
Source: Richard Feynman, *Surely You're Joking, Mr. Feynman!*

Millikan Oil Drop Experiment

Why didn't they discover the new number was higher right away? It's a thing that scientists are ashamed of—this history—because it's apparent that people did things like this:

Source: Richard Feynman, *Surely You're Joking, Mr. Feynman!*

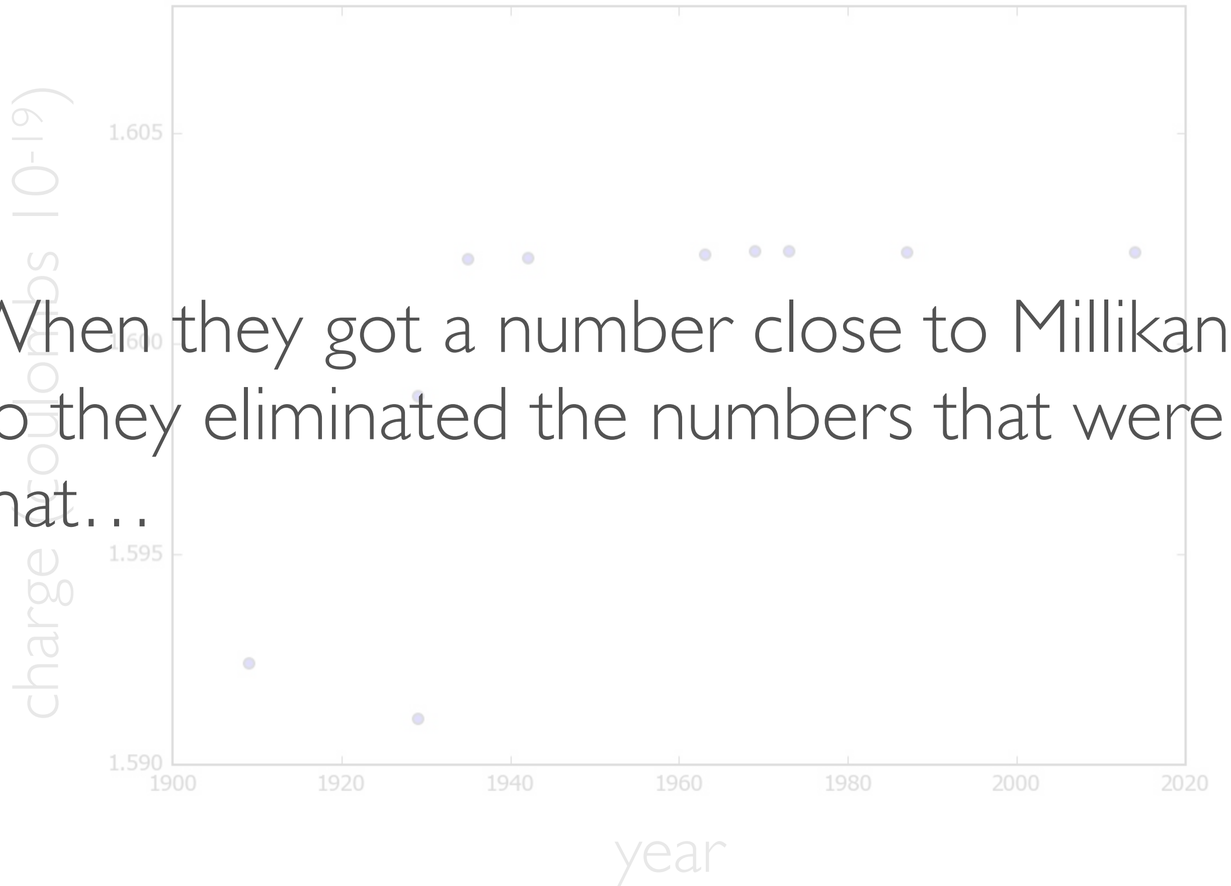
Millikan Oil Drop Experiment



When they got a number that was too high above Millikan's, they thought something must be wrong—and they would look for and find a reason why something might be wrong.

Source: Richard Feynman, *Surely You're Joking, Mr. Feynman!*

Millikan Oil Drop Experiment



When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that…

Source: Richard Feynman, *Surely You're Joking, Mr. Feynman!*

Confirmation bias

A cognitive bias wherein humans have a tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses.