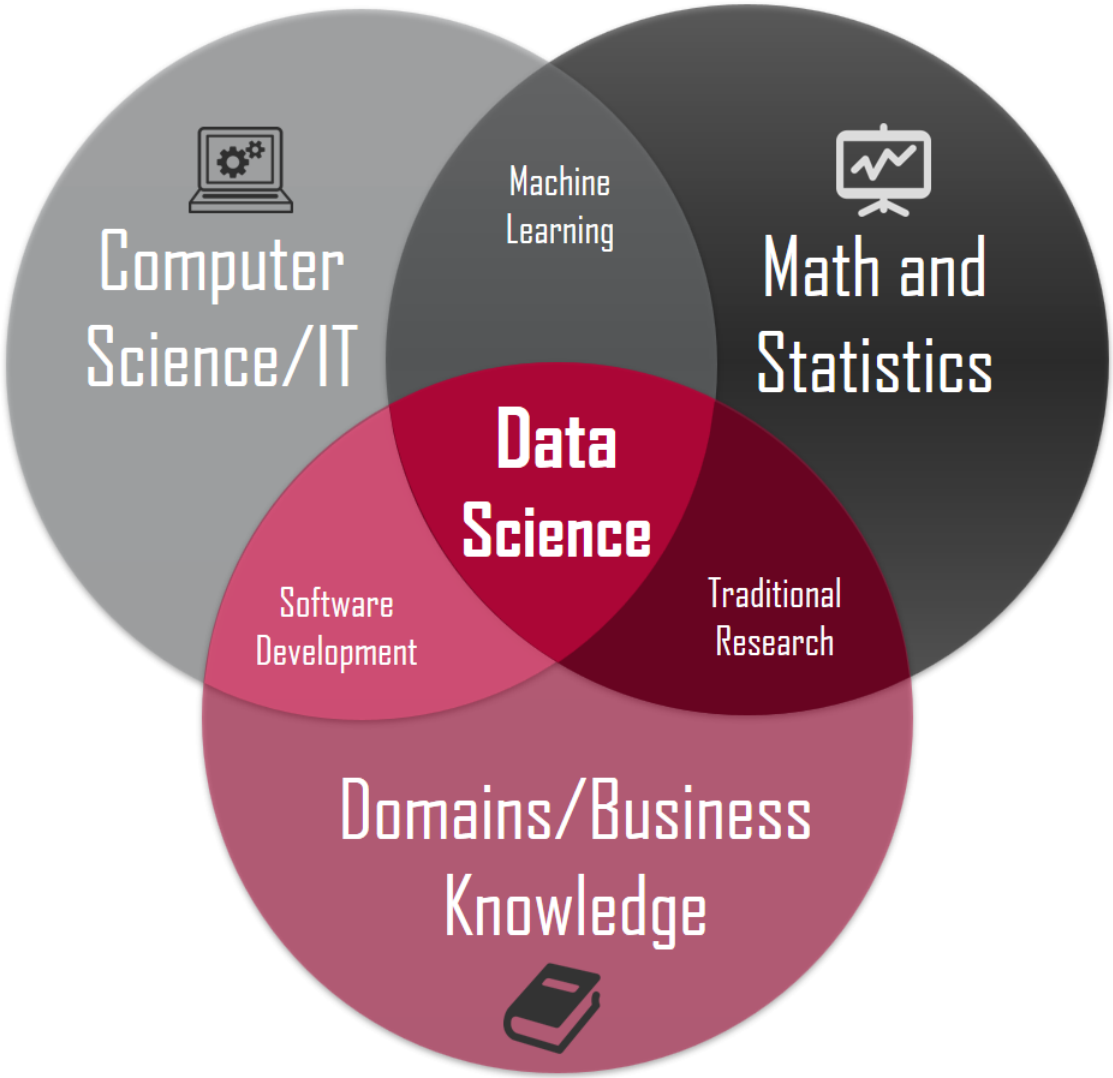
Data Science 101: an introduction

Data Science Team

Stanford University, Department of Statistics

One view


Computer Science/IT
Machine Learning
Math and Statistics
Data Science
Software Development
Traditional Research
Domains/Business Knowledge

A takehome message from this course:

Good Statistics and Data Science involves much more than running a Machine Learning algorithm on data.

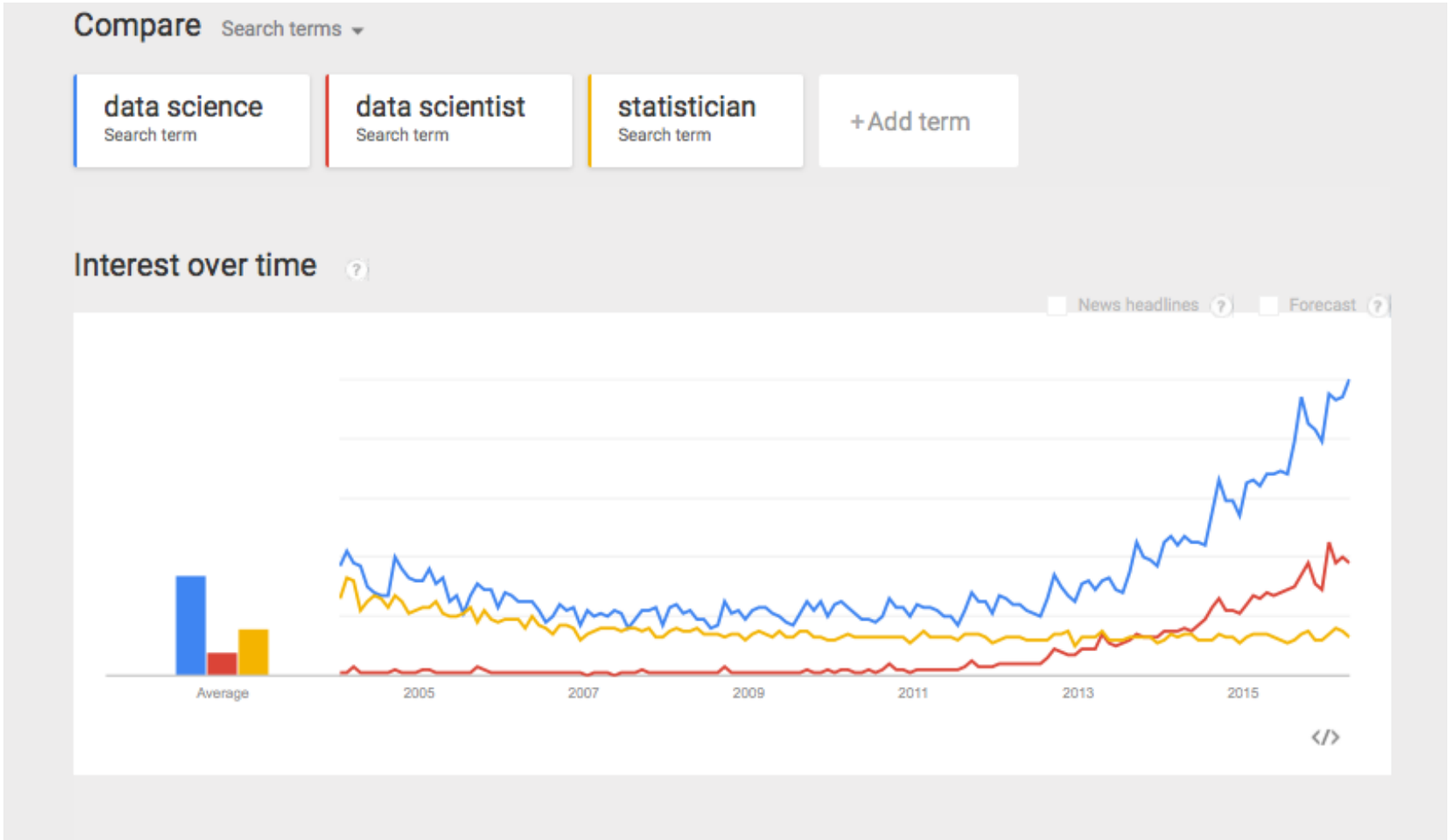
There are essential challenges like:

- How should I design my study to answer a question? Is it feasible and ethical to carry out such a study?
- What data should I collect? How much data?

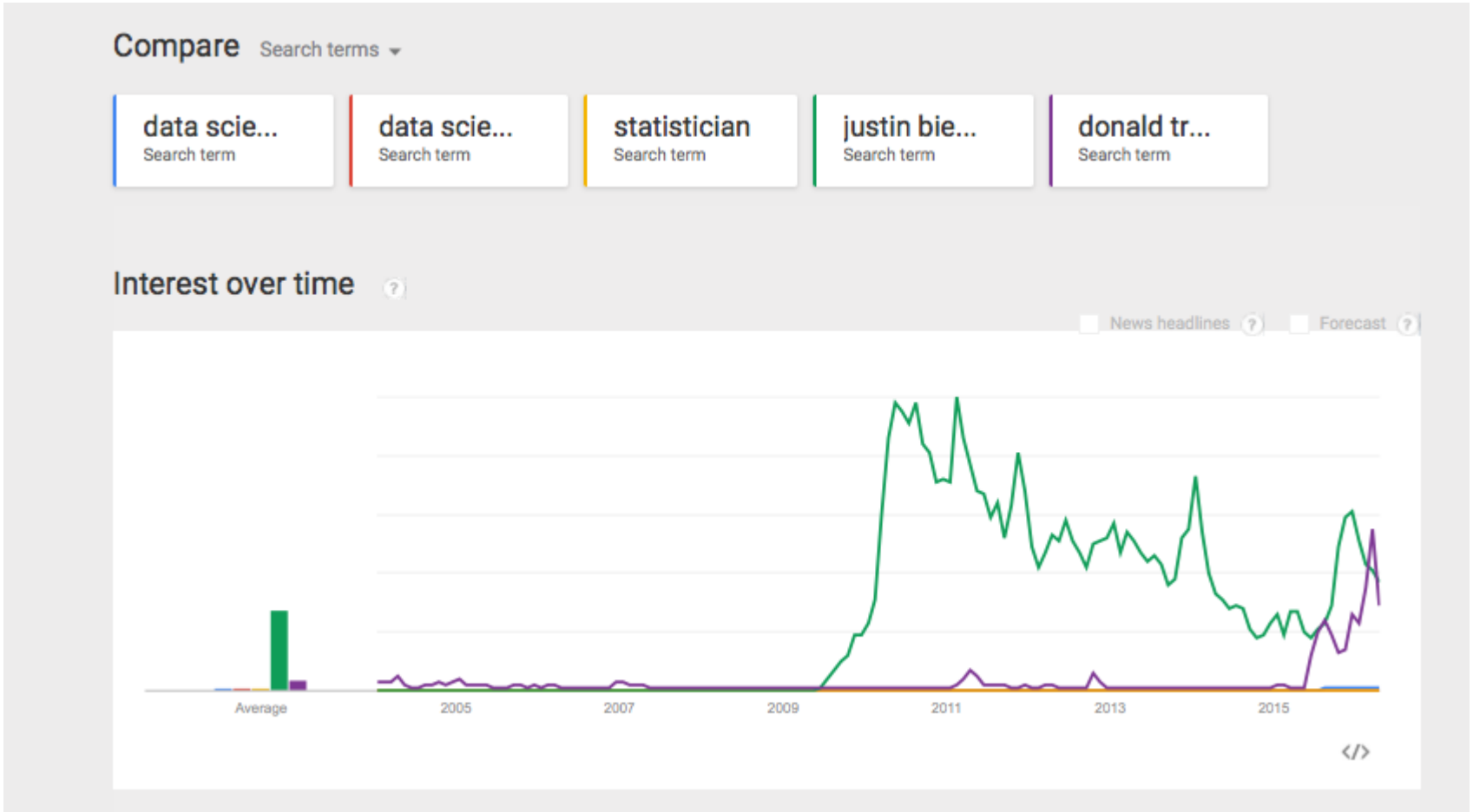
After collecting the data:

- What is the quality of my data?
- How should I model my data to try to answer the question at hand?
- What can I conclude from my study, and with how much certainty?

Data Scientist vs Statistician in Google search



Keeping things in perspective

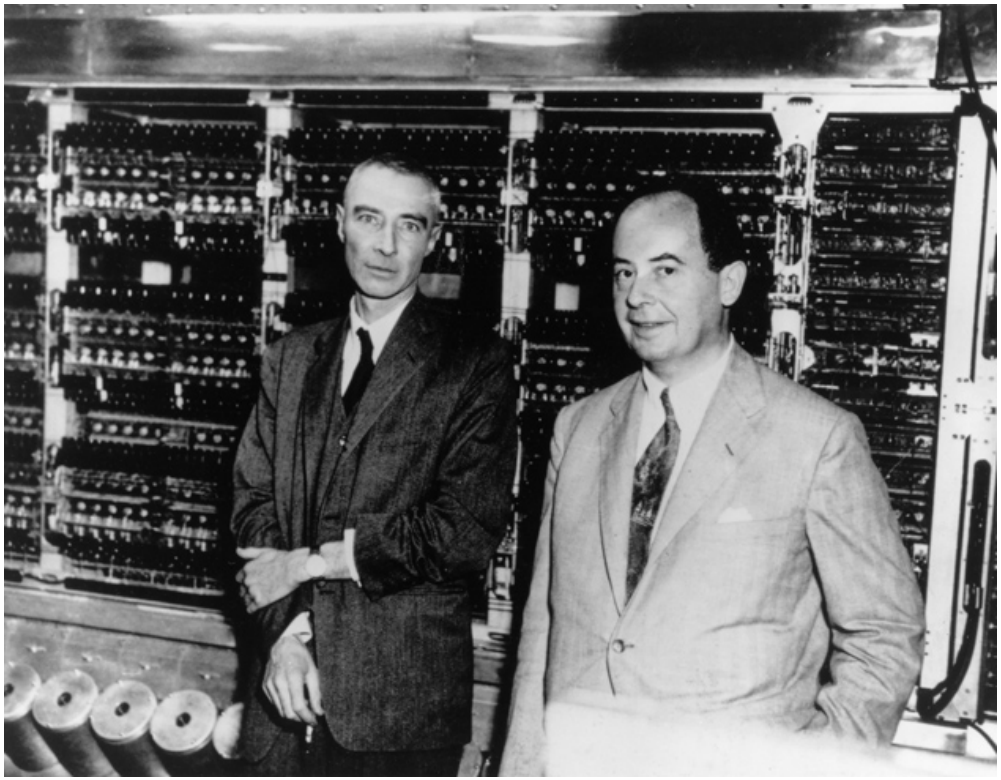


Data Science 101

- Literacy for the digital age
- Data science 101 = Self-defense 101 in the modern workplace
- Data
  - What is it?
  - How to explore it?
  - Where does it come from?
- Science
  - What does it mean to learn from data?
  - Prediction vs Inference
  - How do we know when we are right or wrong?

A lot of buzz about a lot of data

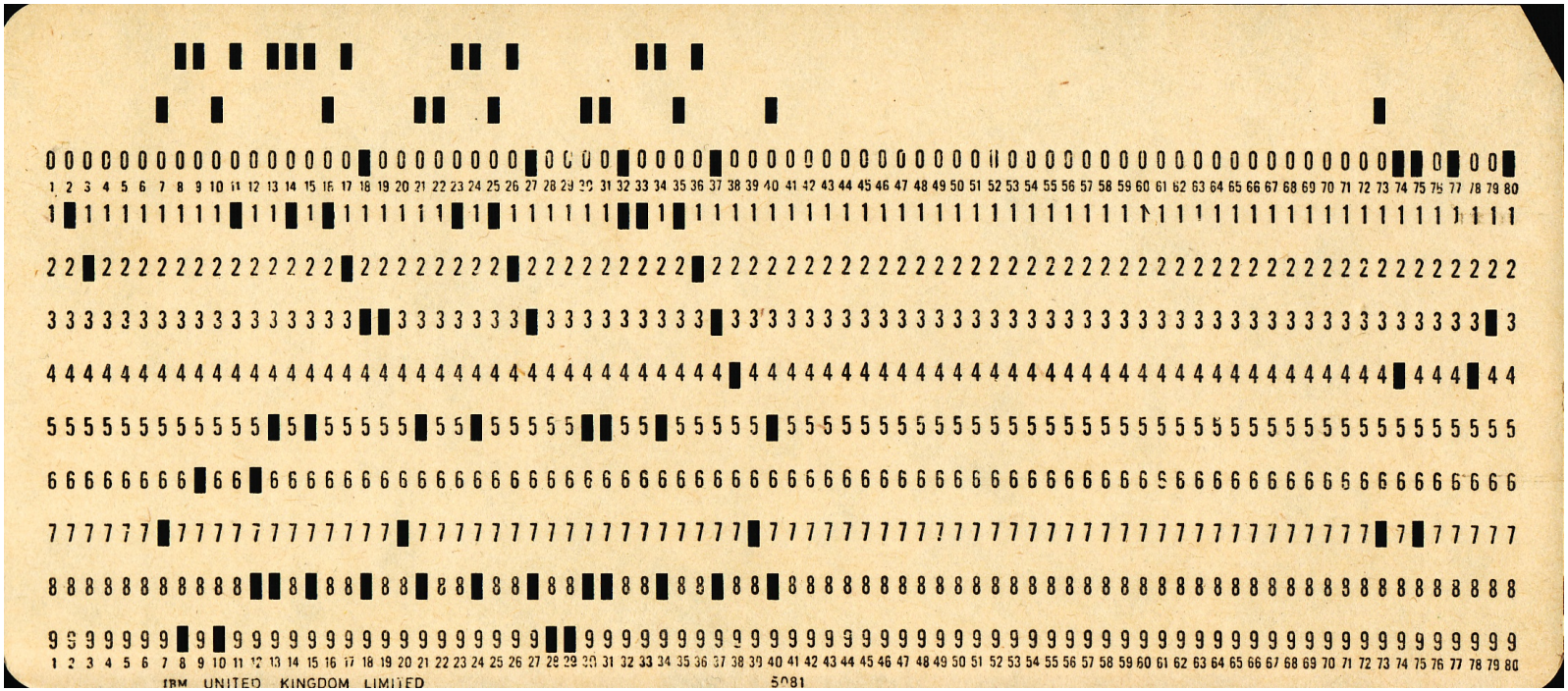
Things have changed a lot since the 1940s-1950s. . .



Robert Oppenhemier and John Von Neumann in front of the Electronic Numerical Integrator and Computer (ENIAC), source.

The ENIAC was the first programmable, electronic, general-purpose digital computer (1945).

And since the 1960s. . .

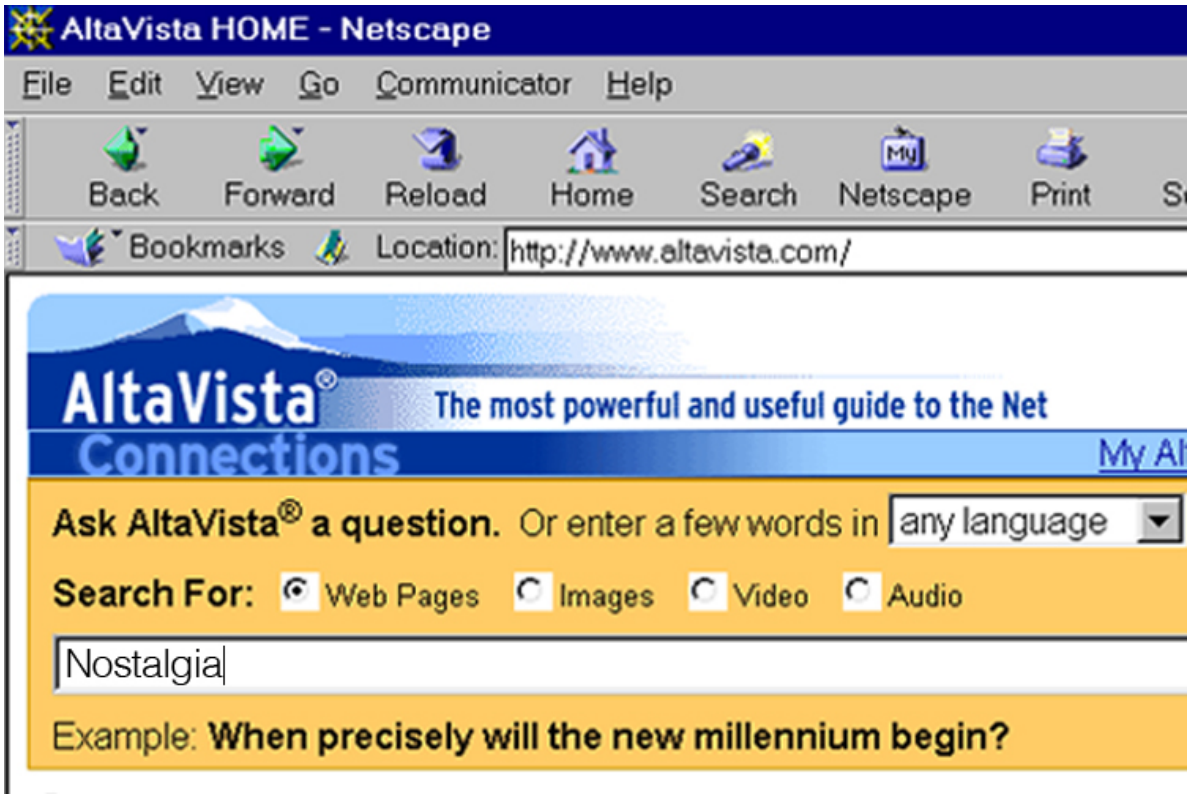


Punched card: Holds digital data represented by absence / presence of holes, source.

And since the 1980s. . .

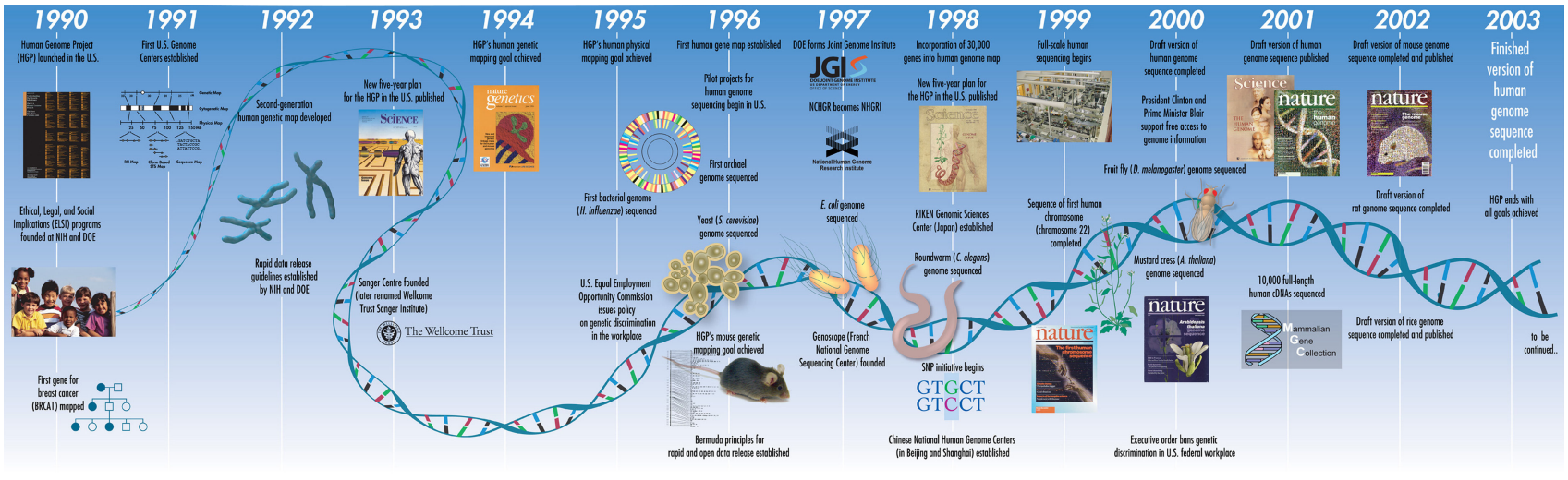
IBM PC 5150, source. The IBM 5150 was the first microcomputer released in the IBM model line and had a strong influence on the personal computer market.

And since the 1990s. . .



AltaVista was established in 1995 and was one of the most-used early search engines, image source.

And, yes, since the 2000s. . .



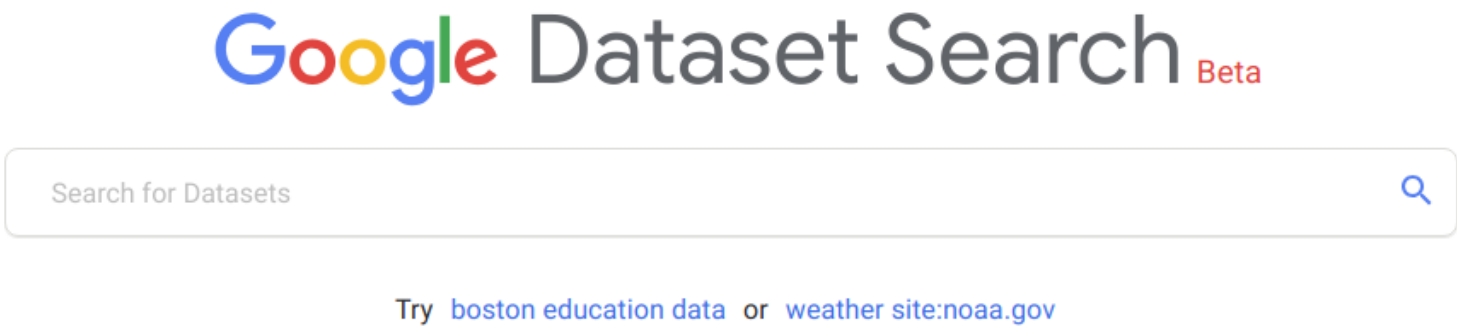
The Human Genome Project: 1990 - 2003, image source.

Some examples of Big Data

- **Genetics**: Now costs a few hundred dollars to sequence an entire genome. Typically there are many more measurements than subjects
- **Physics/CERN LHC**: LHC (Large Hadron Collider - particle accelerator) experiments produce about 90 petabytes (1 petabyte $= 2^{50}$ bytes) of data per year (another 25 petabytes for non-LHC experiments)
- **Astronomy/Sloan Digital Sky Survey**: about 25.5TB telescope imaging data
- **Passively gathered data**
  - Products we buy
  - Topics that engage us
  - Our levels of physical activity
  - Who we talk to

Some sources of Data

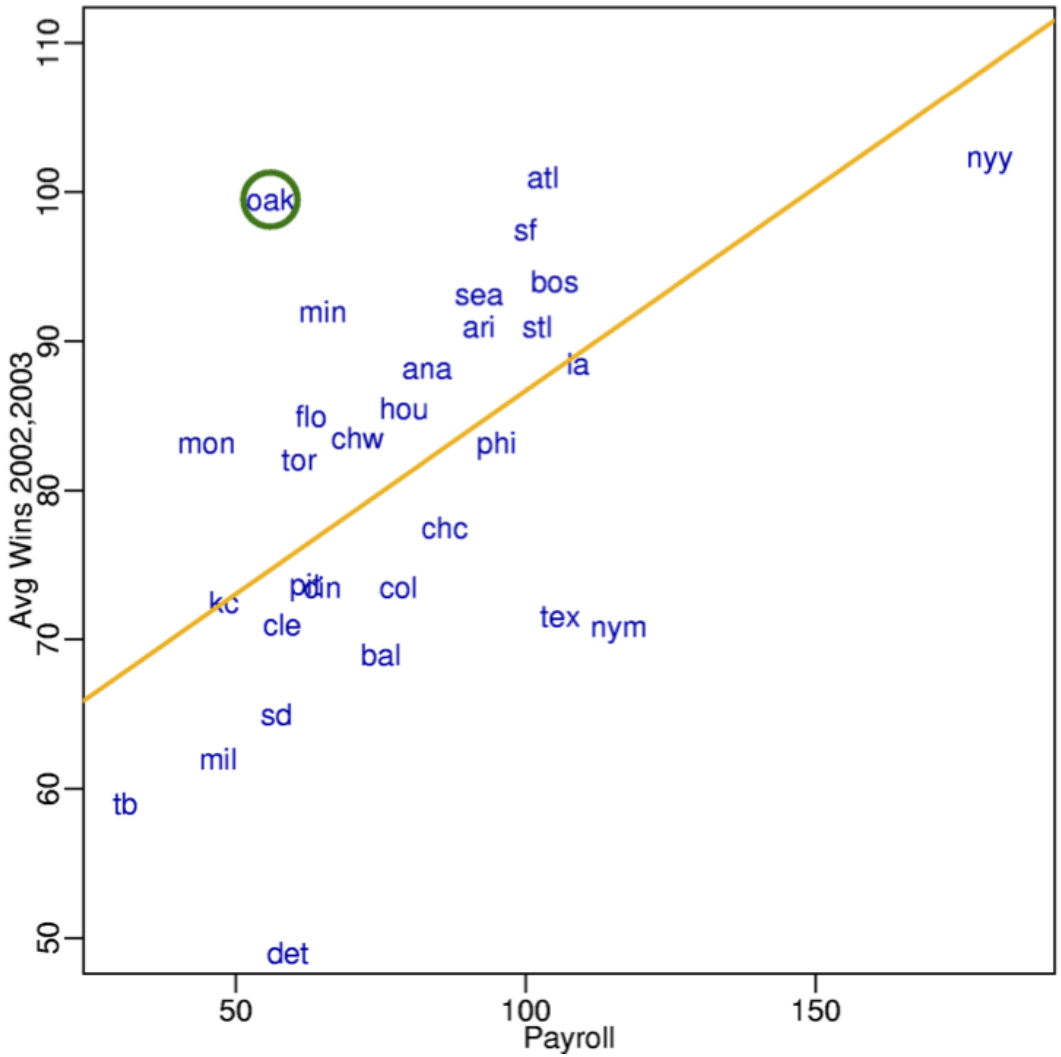
- Stats for change (public data sources related to social issues)
- Data.gov (U.S. Government open data)
- dbGap (database of Genotypes and Phenotypes)
- CERN (CERN particle physics data)
- Google Dataset Search
- Private data



Data science can do a lot: sports

Moneyball was a 2011 movie. Using "data science / statistics techniques", the Oakland Athletics baseball team build a team of "undervalued" talent, source.

Data science can do a lot: sports

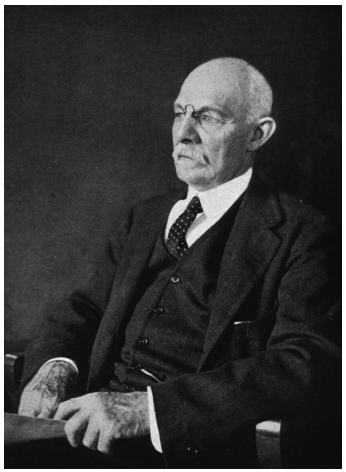


Starting around 2001, the Oakland A's picked players that scouts thought were no good but data said otherwise.

The problem with bad data

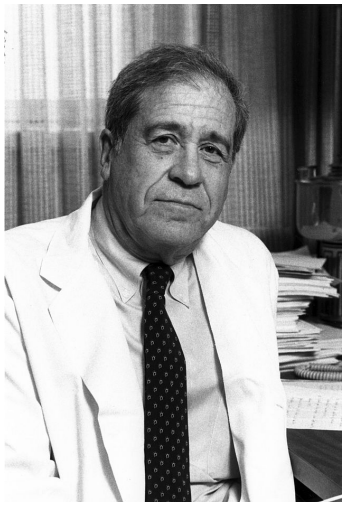
- Picture of Truman holding the Chicago Daily Tribune with an incorrect banner headline on front.
- Truman won over his opponent Dewey in the 1948 presidential election.
- The paper relied on their analyst who predicted the winner in previous presidential contests. The conventional wisdom (supported by various polls) was that Dewey would win the election by a "landslide", see here.

Data science can do a lot: medicine



- William Stewart Halsted of Johns Hopkins University introduced radical mastectomy for the treatment of Breast Cancer in the 1920s
- It became the standard treatment in the field, and became more radical (and disfiguring) as time went on
- Few doctors dared to challenge the conventional wisdom, partly because of Halsted's stature

Fast forward to 1958



- Finally in 1958 Dr. Bernard Fisher of the University of Pittsburgh managed to run the first clinical trial comparing radical mastectomy to less invasive treatments.
- His work showed that early-stage breast cancer could be more effectively treated by lumpectomy, in combination with radiation therapy, chemotherapy, and/or hormonal therapy

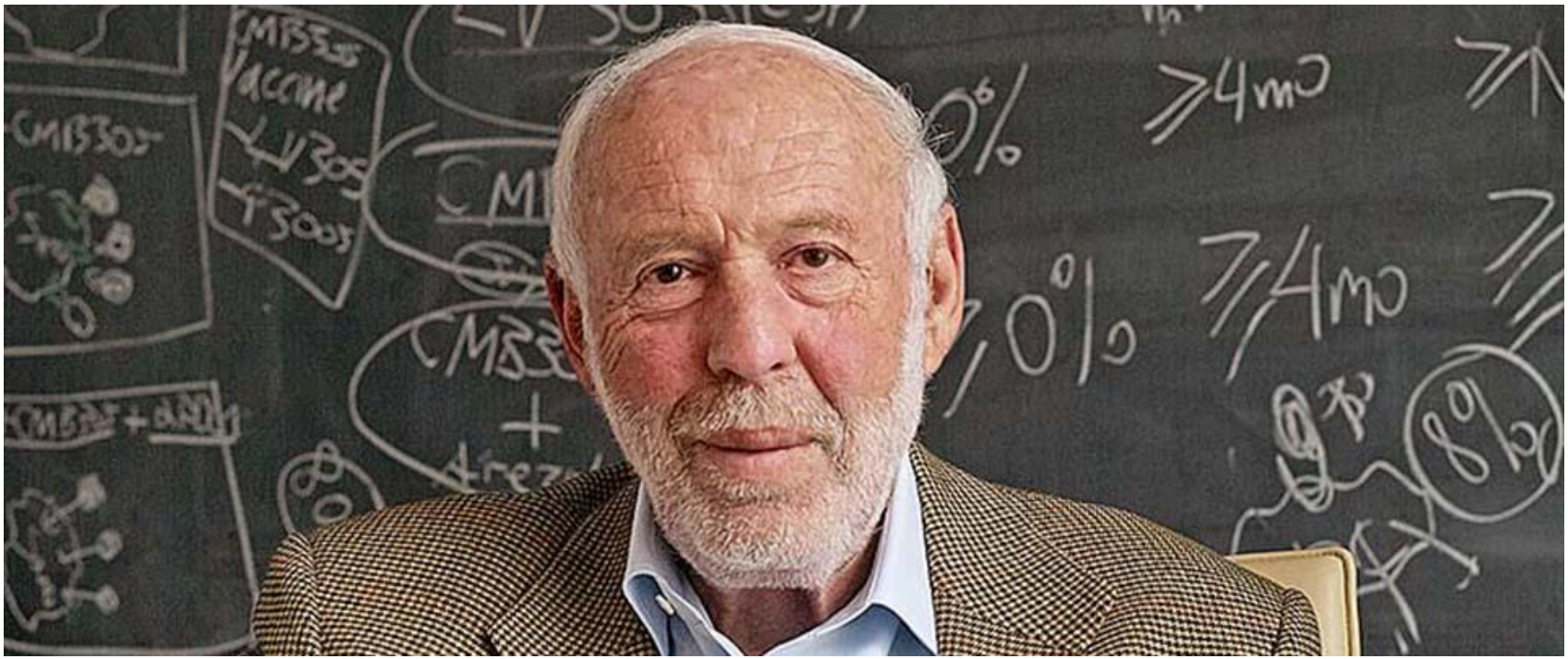
Data science can do a lot: medicine

The National Institutes of Health has announced a new opportunity for organizations interested in helping engage volunteers in the *All of Us* Research Program, part of the Precision Medicine Initiative. This funding opportunity, open to national and regional organizations, as well as local community groups, will support activities to promote enrollment and retention in the *All of Us* Research Program across diverse communities.

All of Us is an ambitious effort to gather data over time from 1 million or more people living in the United States, with the ultimate goal of accelerating research and improving health. Unlike research studies that are focused on a specific disease or population, *All of Us* will serve as a national research resource to inform thousands of studies, covering a wide variety of health conditions. Researchers will use data from the program to learn more about how individual differences in lifestyle, environment and biological make-up can influence health and disease. By taking part, people will be able to learn more about their own health and contribute to an effort that will advance the health of generations to come. NIH plans to launch the program later this year.

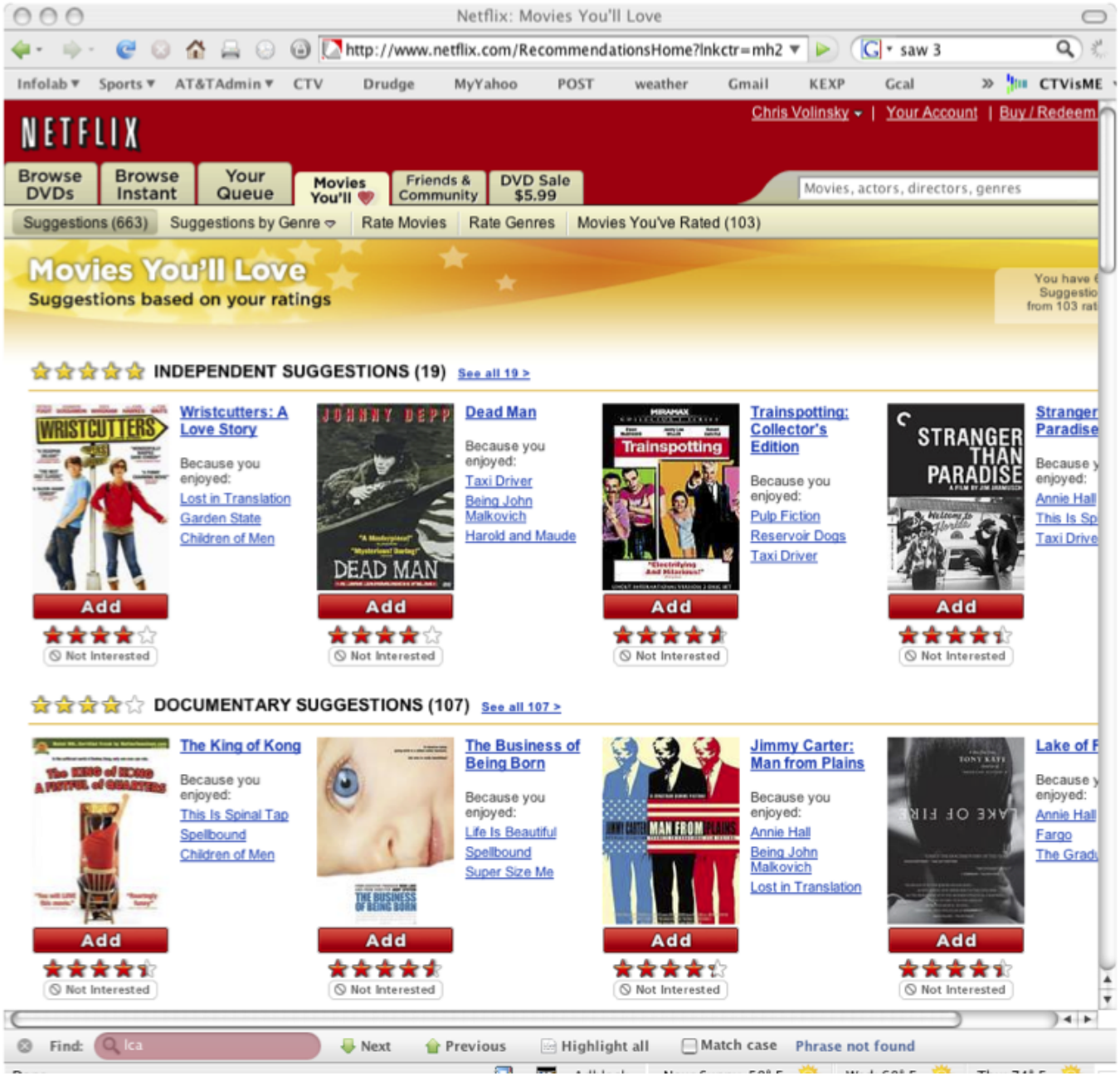
- The All of Us research program wants to build the most diverse health database in history.
- The goal is to investigate how biology, lifestyle and environment affect health in order to help find ways to treat and prevent disease.

Data science can do a lot: finance

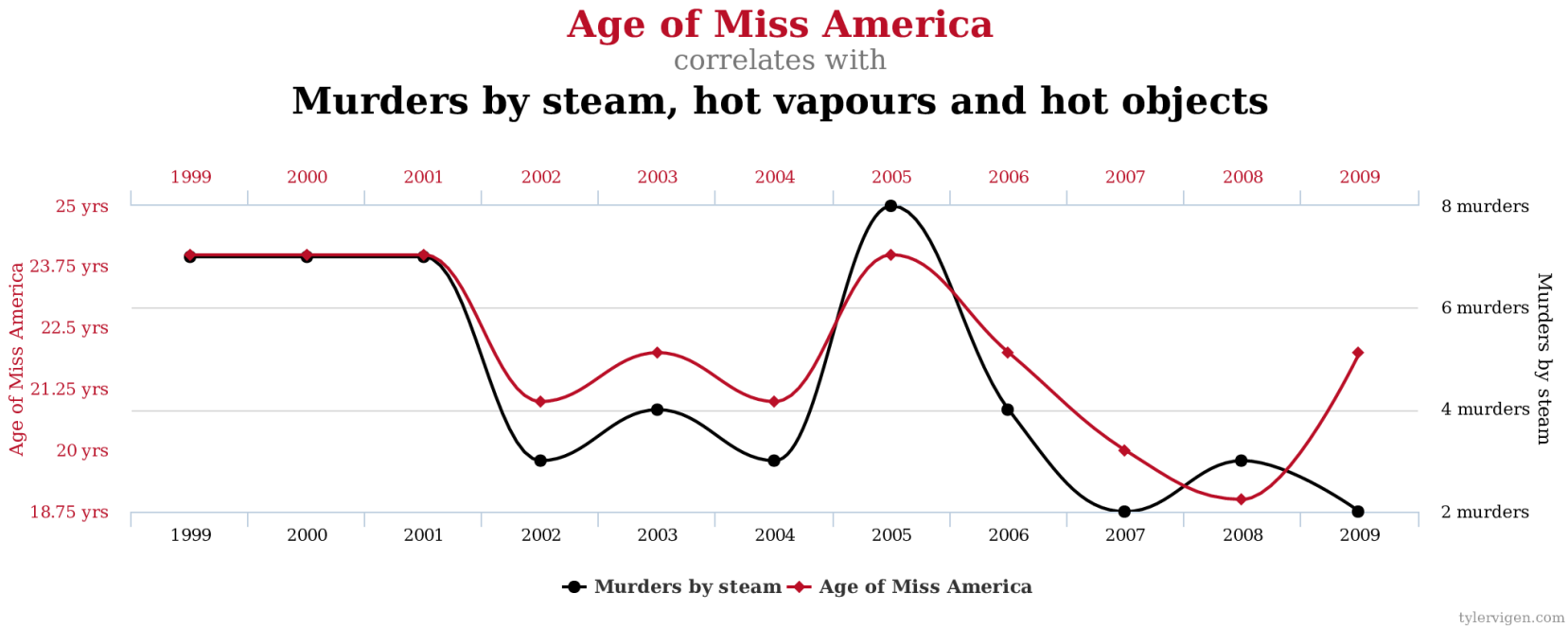


Jim Simons, Renaissance capital, see here. He is a mathematician and founder of a quantitative hedge fund. He is known for his studies on pattern recognition.

Data science can do a lot: recommendations



Is data all we need?



- It is easy to find "interesting" patterns where none exist! (image)
- How should we judge whether a "pattern" is interesting?
- When should we worry about falsely labeling patterns "interesting"? (E.g. Google mistranslates a sentence vs. incorrect cancer diagnosis. . .)

Is data all we need?

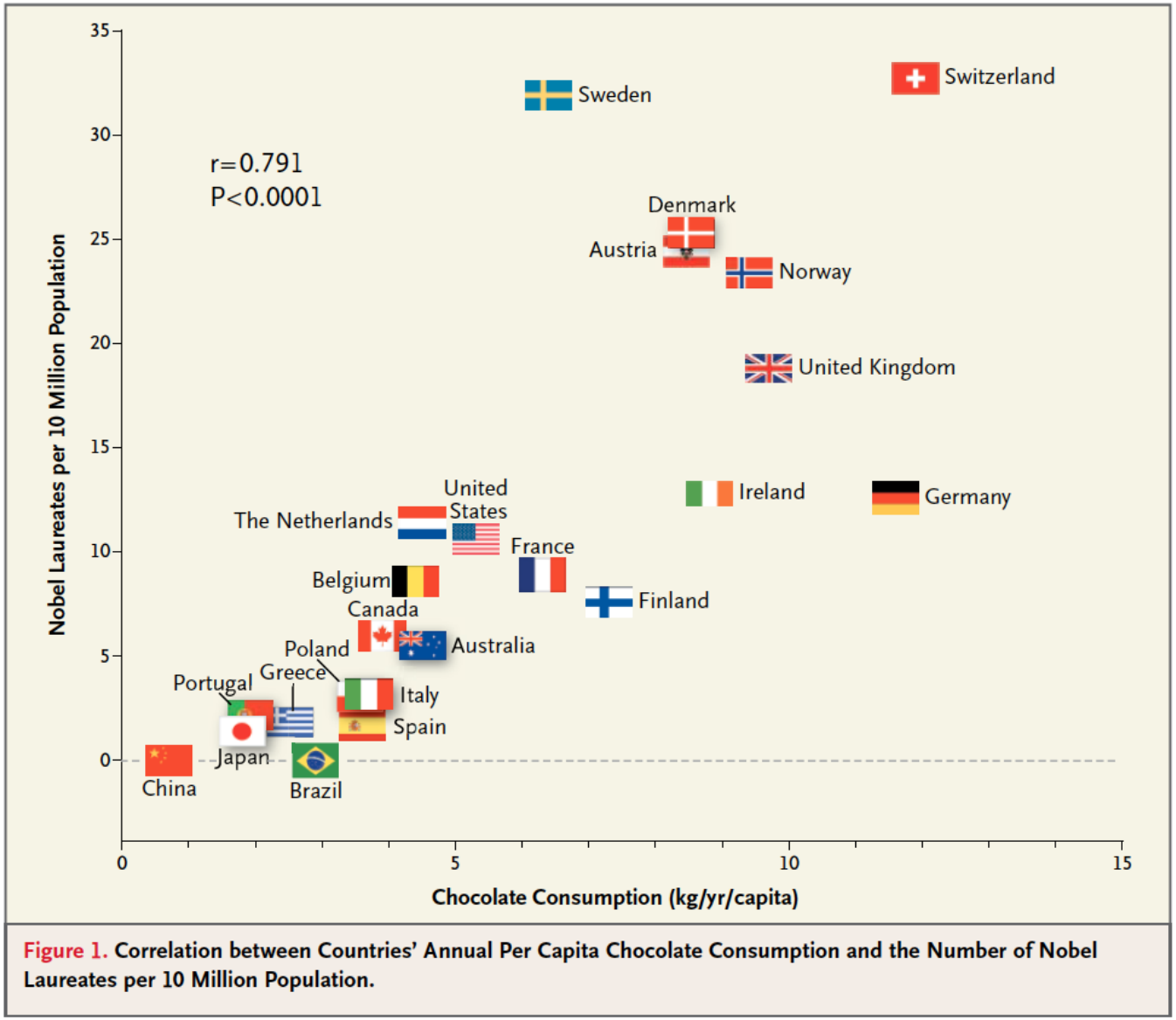


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Is data all we need?

Blindly used, machine learning algorithms can reinforce “hidden” biases.

Questions From the Pandemic

-To mask or not to mask?

-Should young children get vaccinated?

-Hydroxychloroquine? Ivermectin? Molnupiravir?

-Long covid?

Learn how to use data

- **Explore**: identify patterns in data
- **Predict**: make informed guesses about the future, or about subjects we haven't gathered data on yet
- **Infer**: quantify what you know/understand the limits of what you can know based on the data you have