CSC2552

Topics in Computational Social Science: AI, Data, and Society

Winter 2021

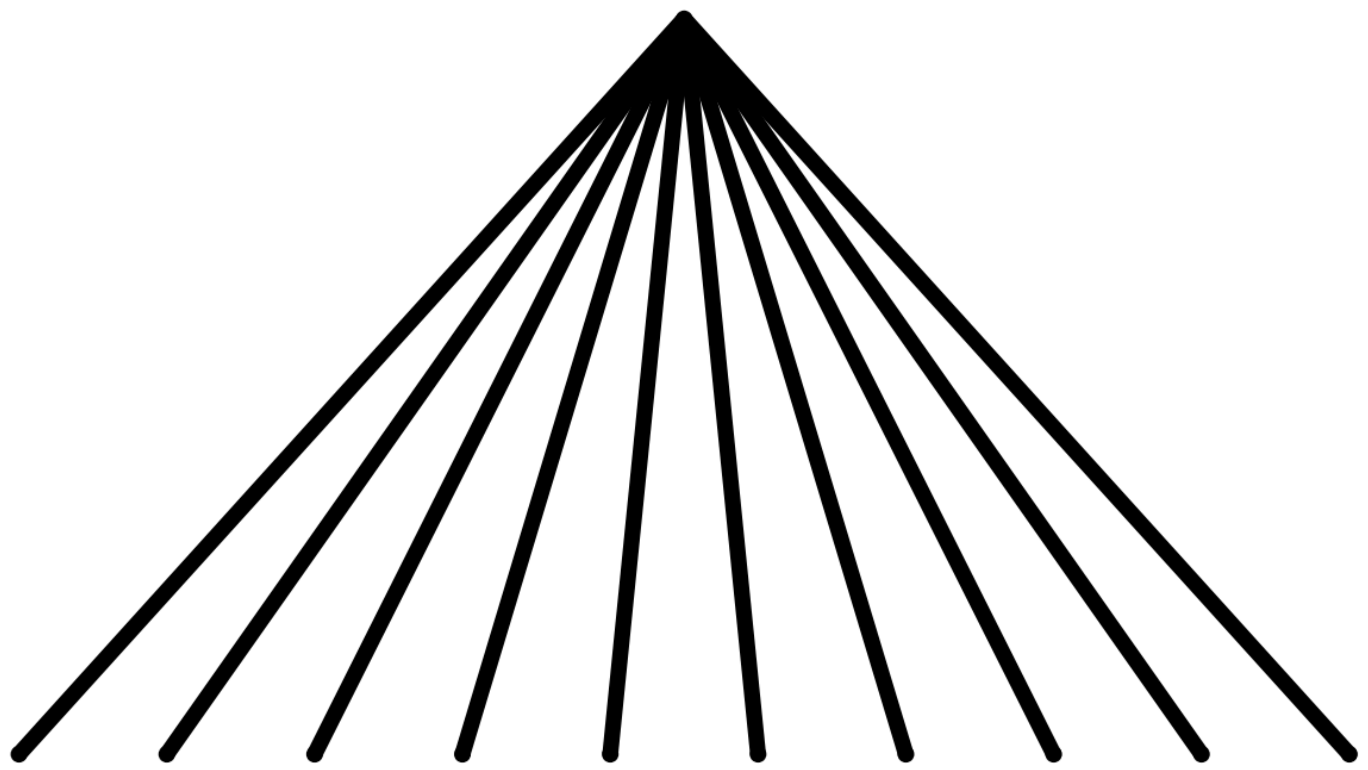
Lecture 1: Introduction to Computational Social Science

Ashton Anderson
University of Toronto

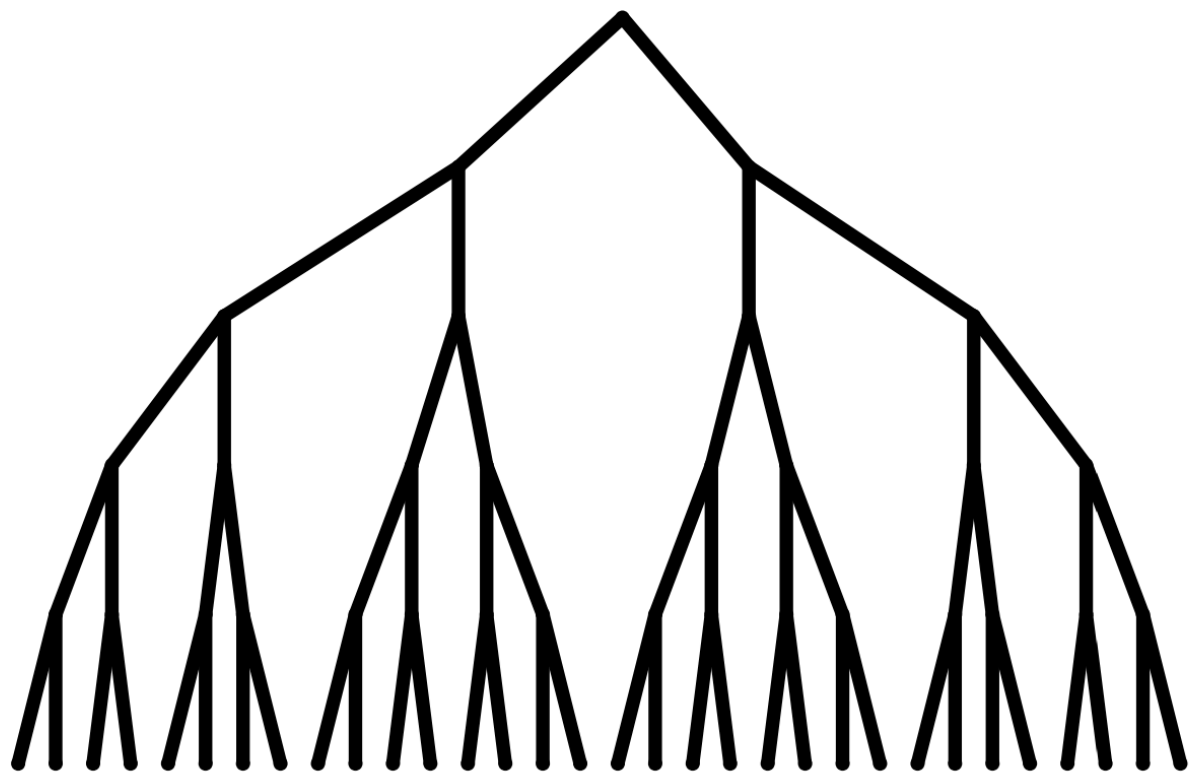
A motivating question

How do people in connected societies learn about new ideas, products, opinions, and beliefs?

Broadcast



Viral

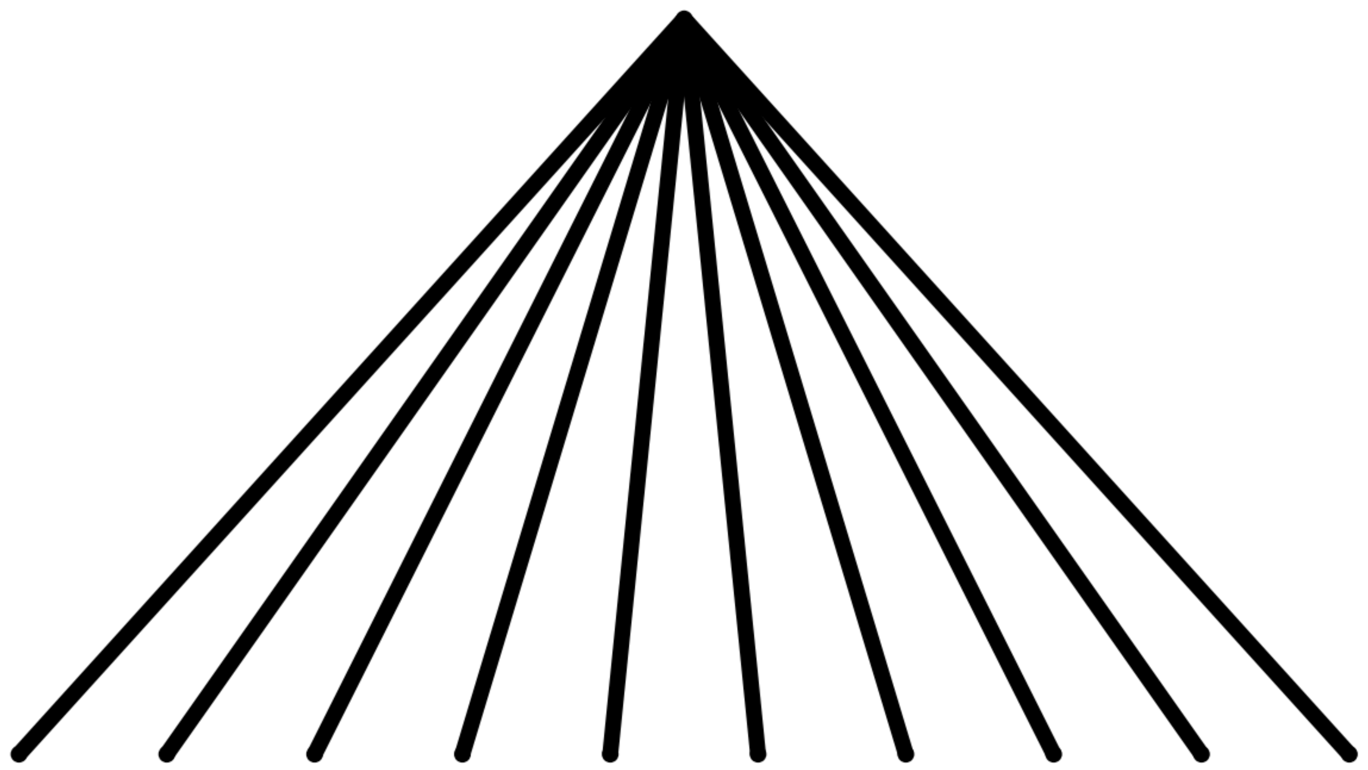


A motivating question

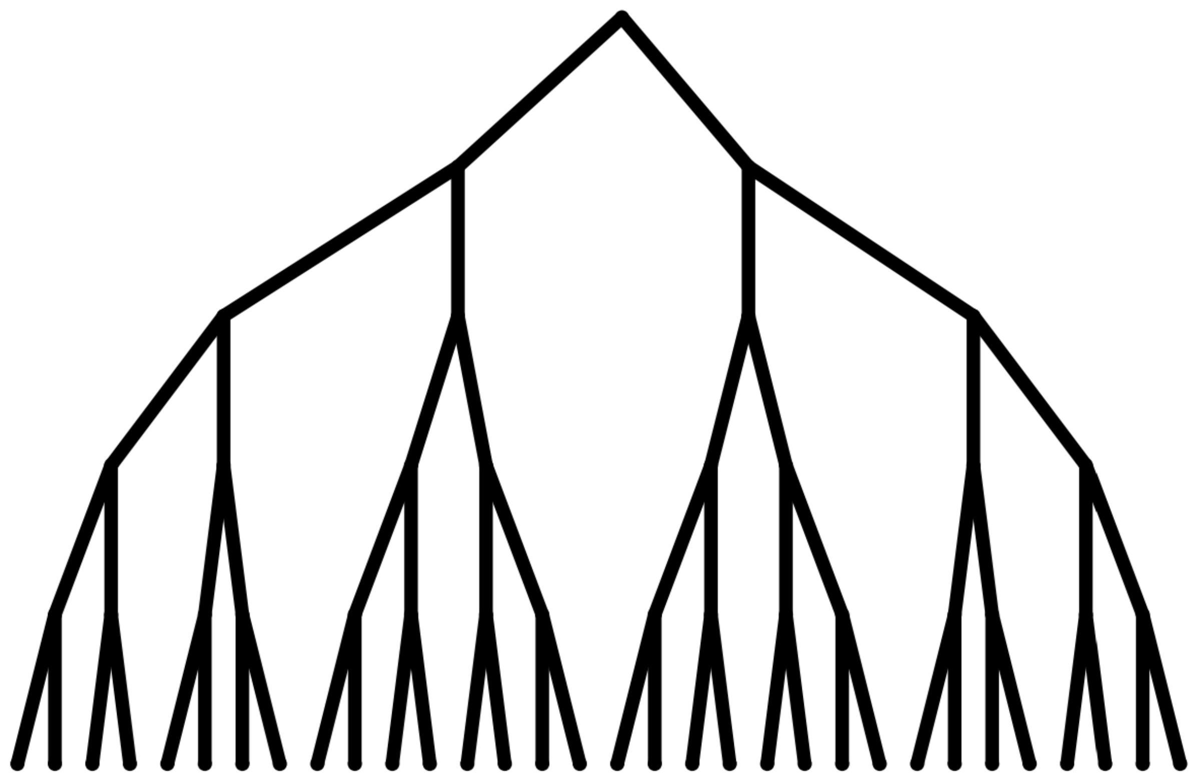
This is an important question:

How people receive information influences
what information they are exposed to,
when they are exposed to it, and
who controls information flow

Broadcast



Viral

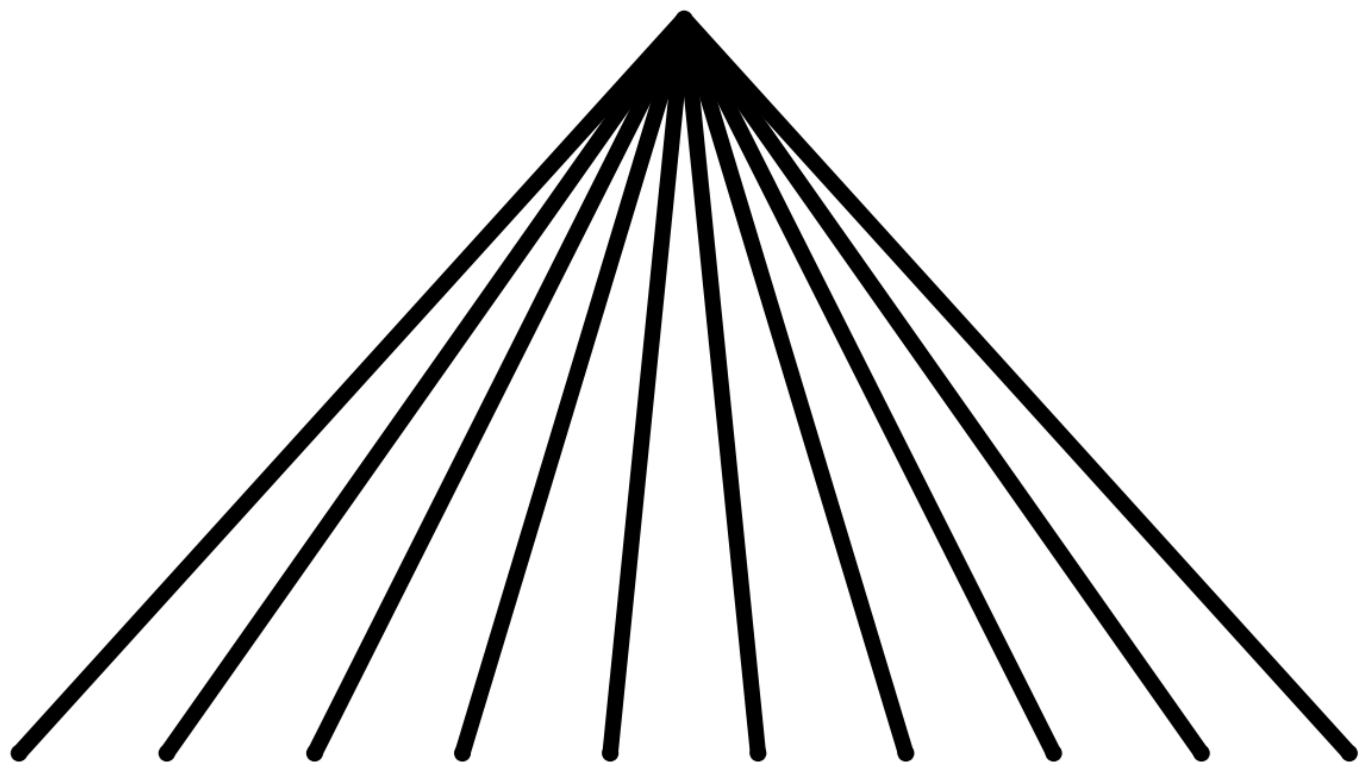


A motivating question

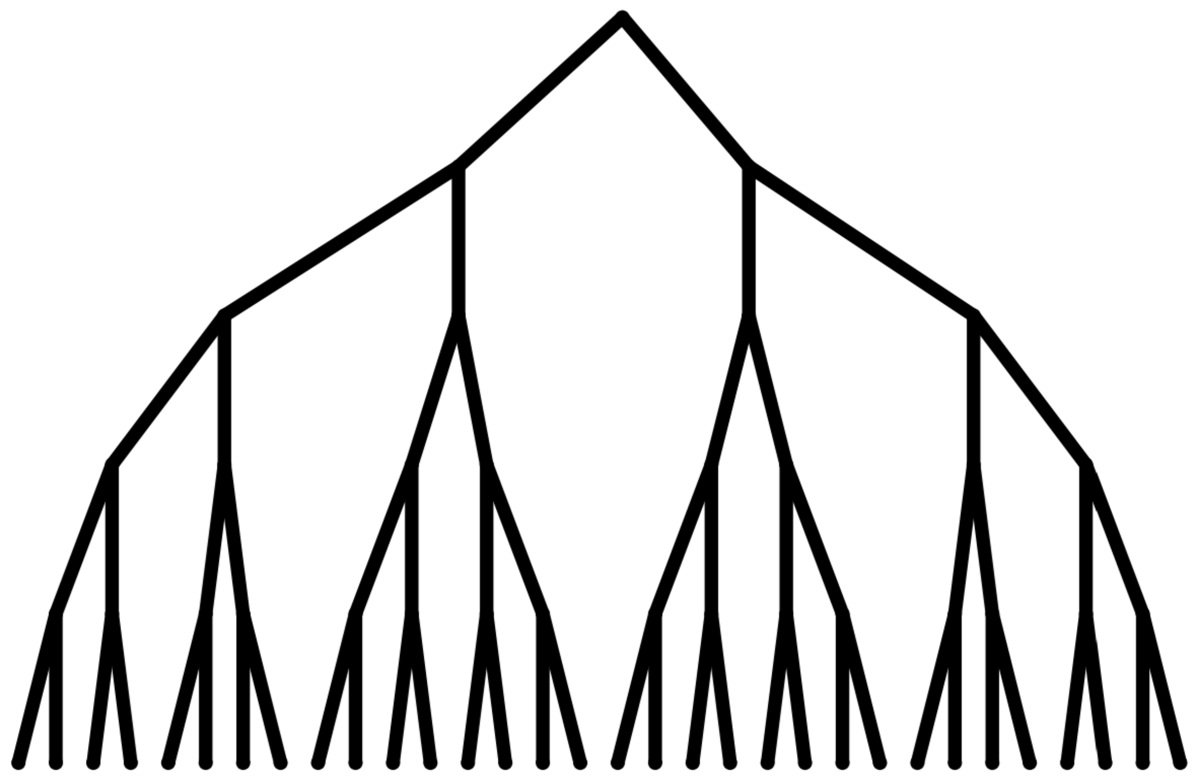
This is a difficult question:

How can we find out how information flows among billions of people?

Broadcast

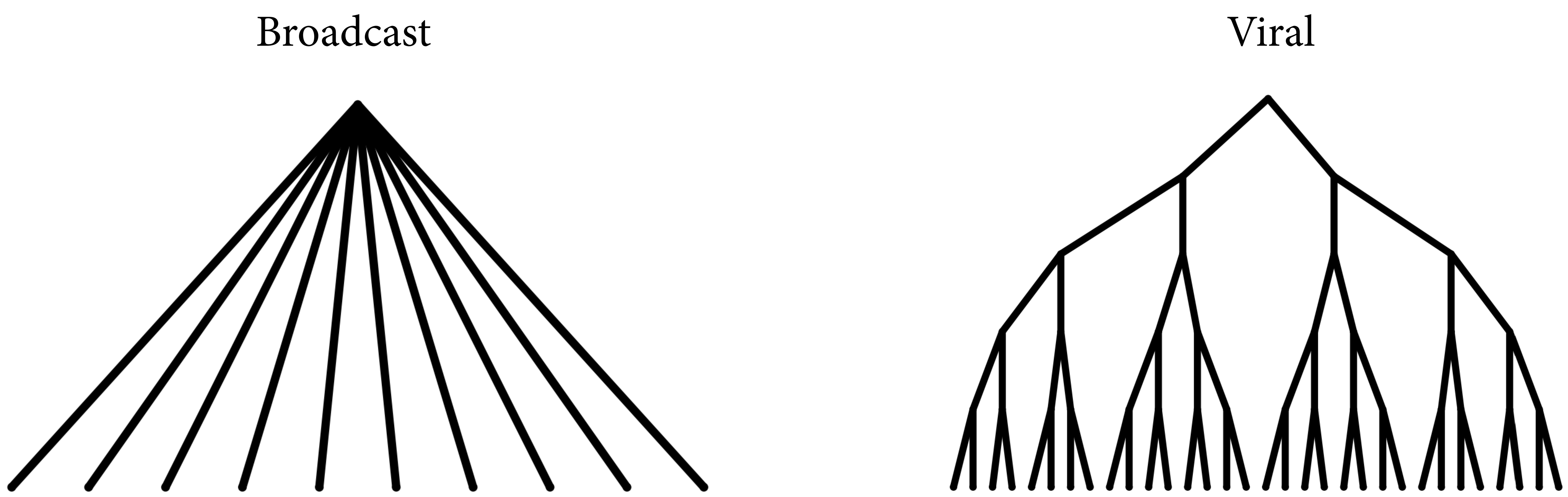


Viral



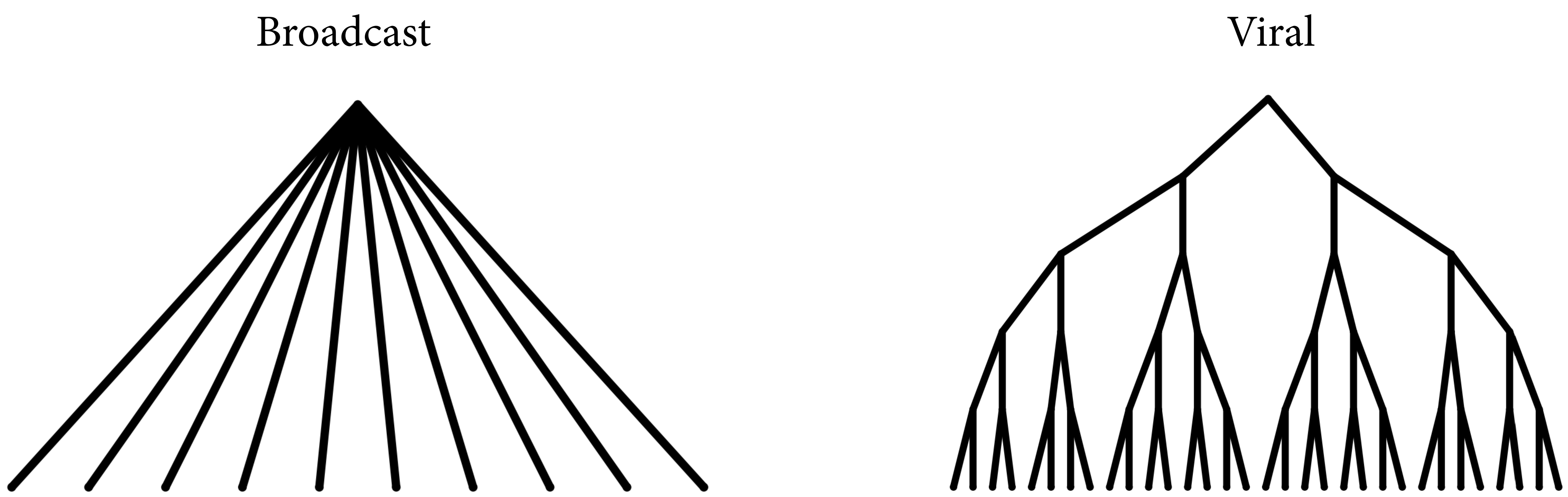
Traditional data & methods

- Introspection
- Survey data
- Aggregate data
- Laboratory experiments
- Computer simulations



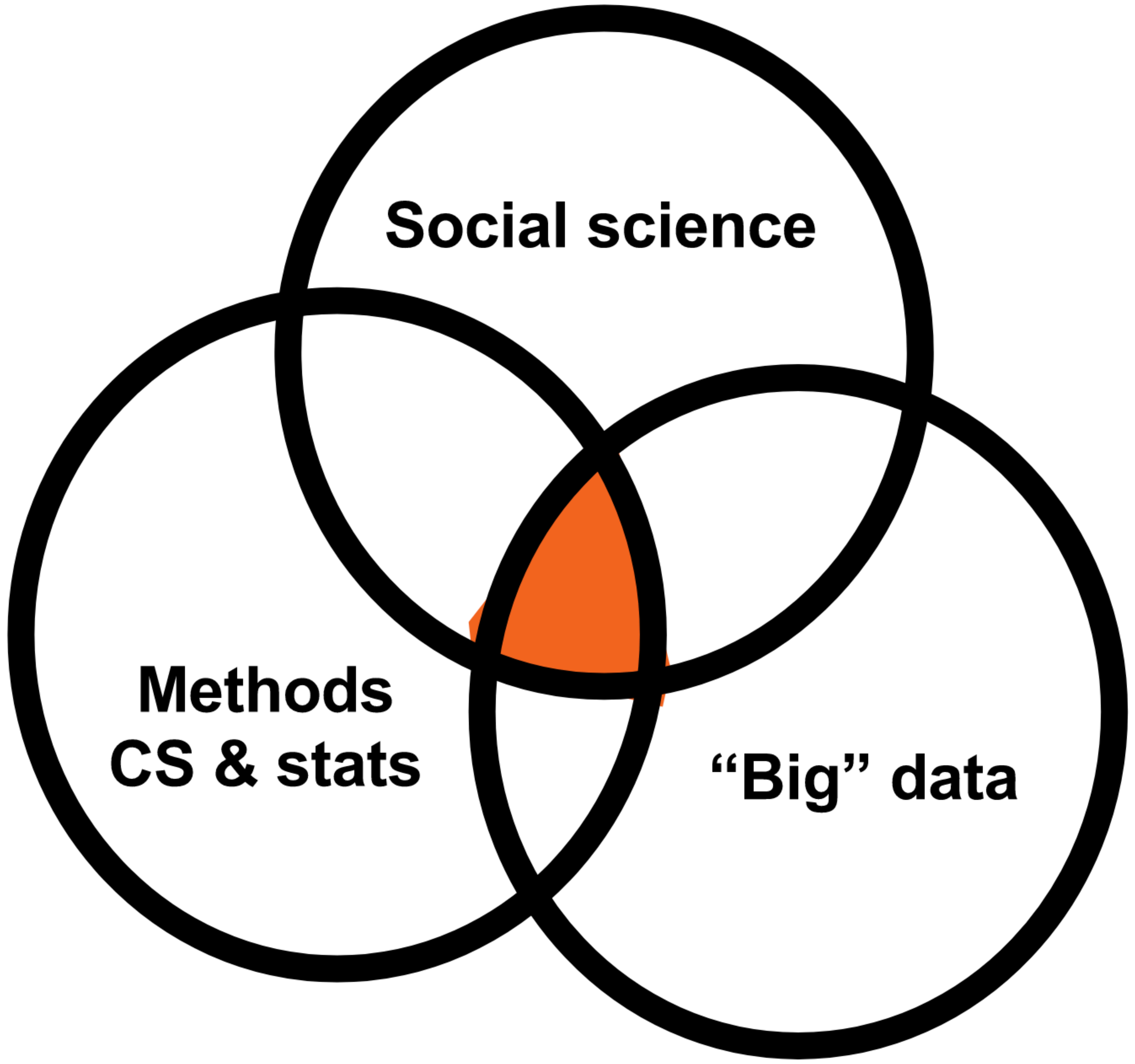
Problems?

- Introspection: biased
- Survey data: incomplete, small
- Aggregate data: insufficiently informative
- Laboratory experiments: generalizable?
- Computer simulations: real?



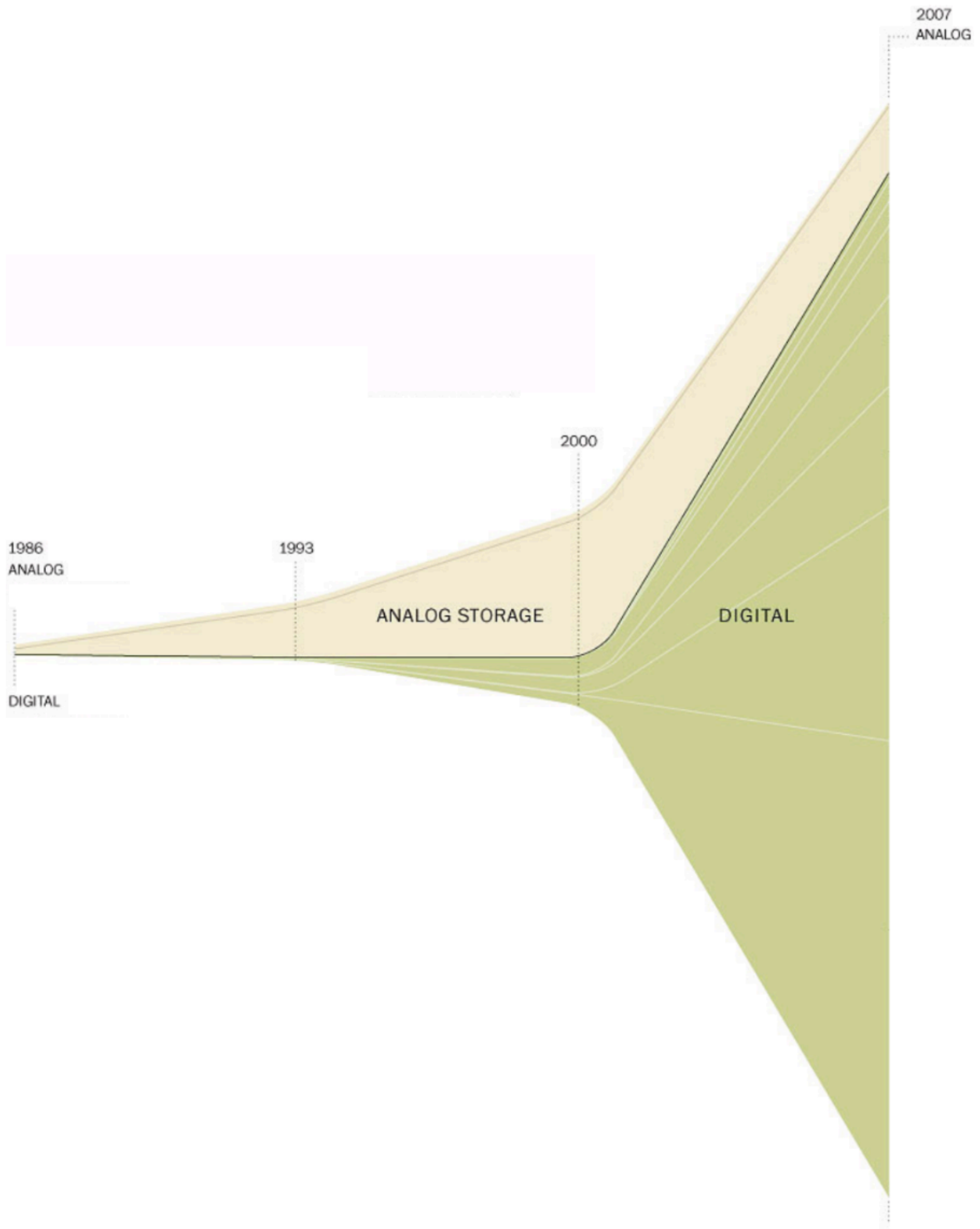
Computational social science

Social research in the digital age



The digital age is creating huge new opportunities for social research

Revolutions in data availability



……..

Revolutions in computing

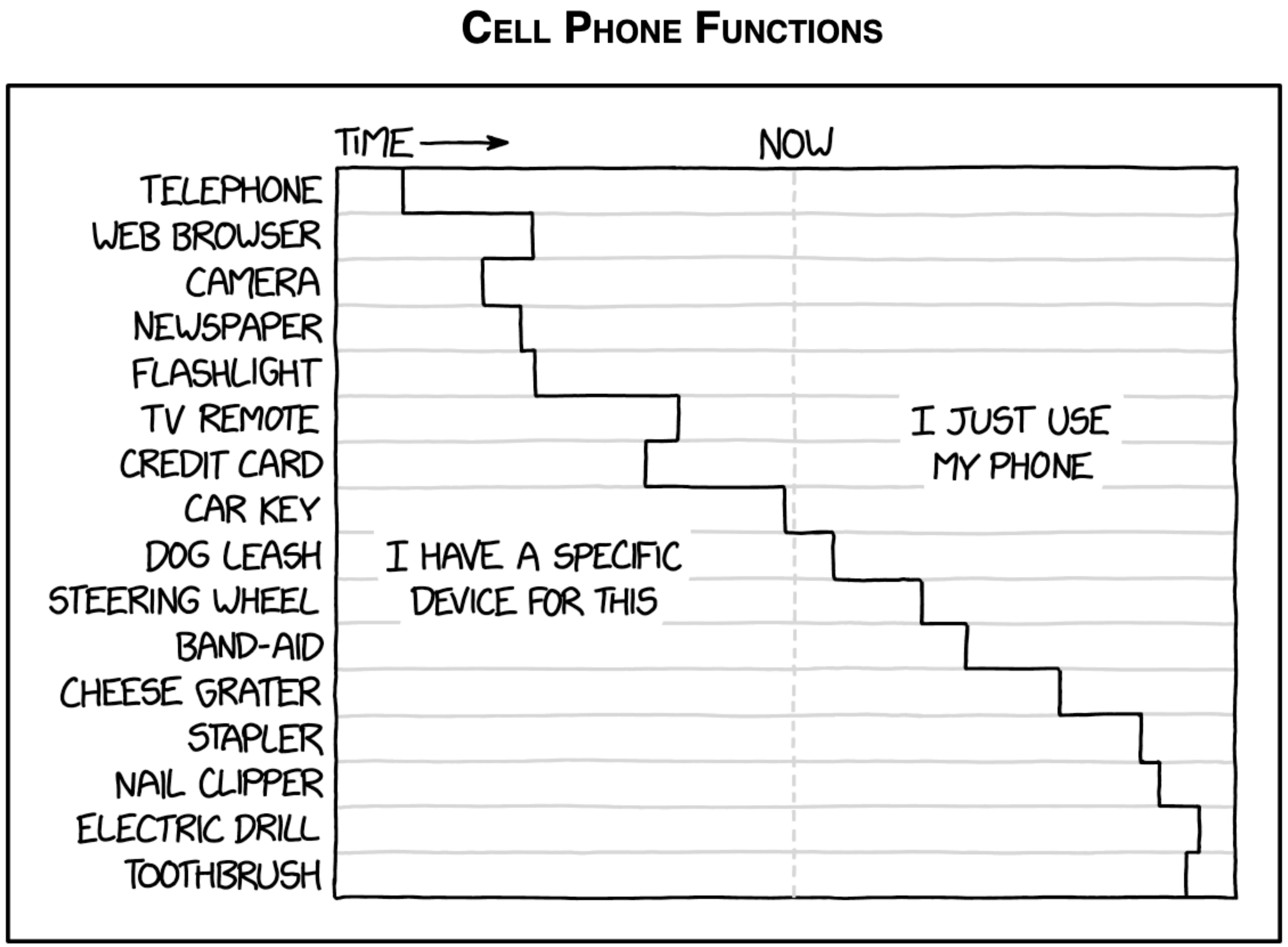
- Massively distributed computing
  - MapReduce, Spark, cloud computing
- Big-memory machines
  - Terabytes of RAM
- Fast streaming algorithms
  - Streaming aggregation, stochastic gradient descent
- Human computation
  - Crowdsourcing, Mechanical Turk

Revolutions in digitization

Everything online

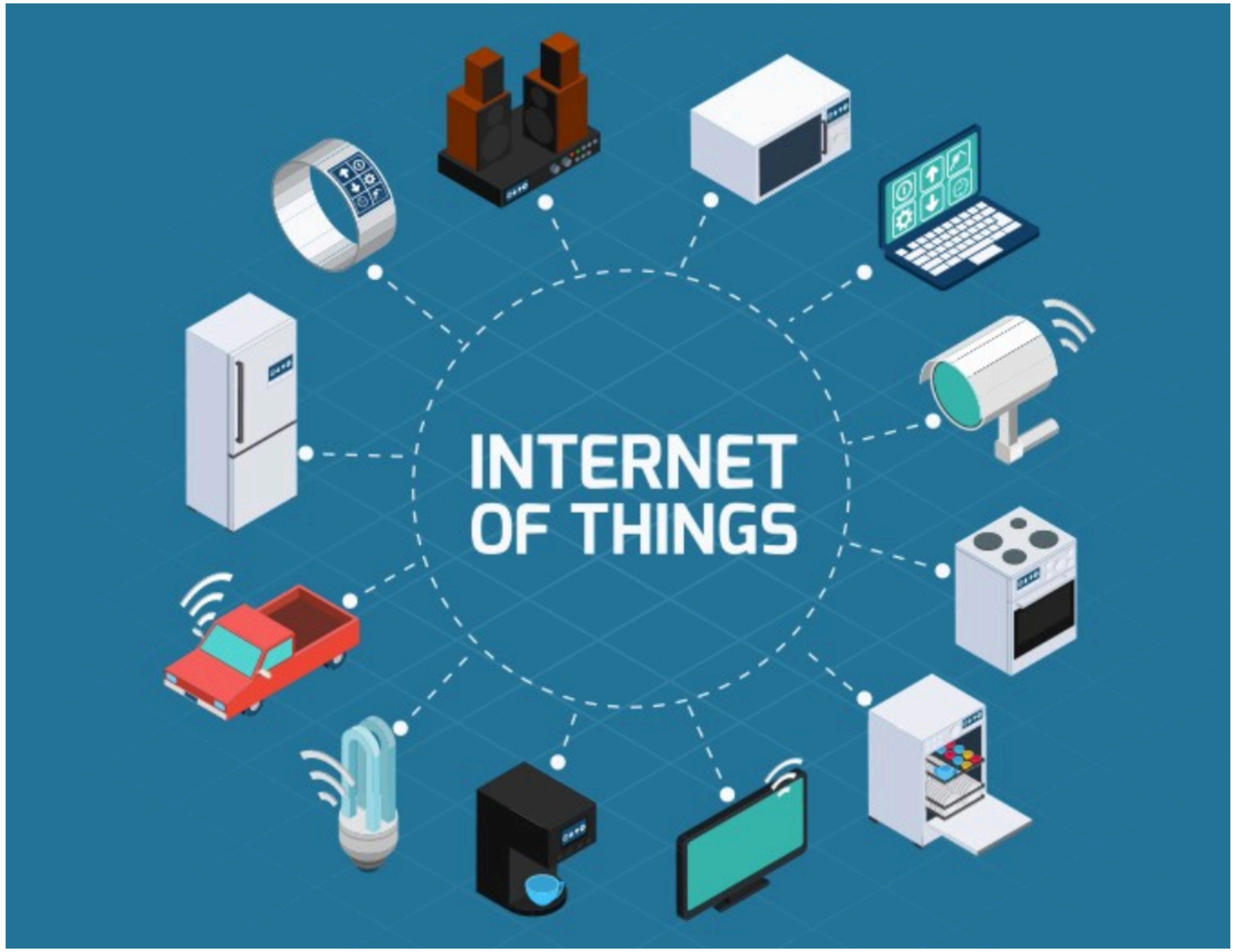
Revolutions in digitization

Computers everywhere



Revolutions in digitization

Computers everywhere



Computers Everywhere

Analog → Digital:

Online:

- Fully measured environments
- Massive, tightly controlled randomised experiments

Offline:

- Similar to online platforms now too
- Physical stores collect data and run experiments

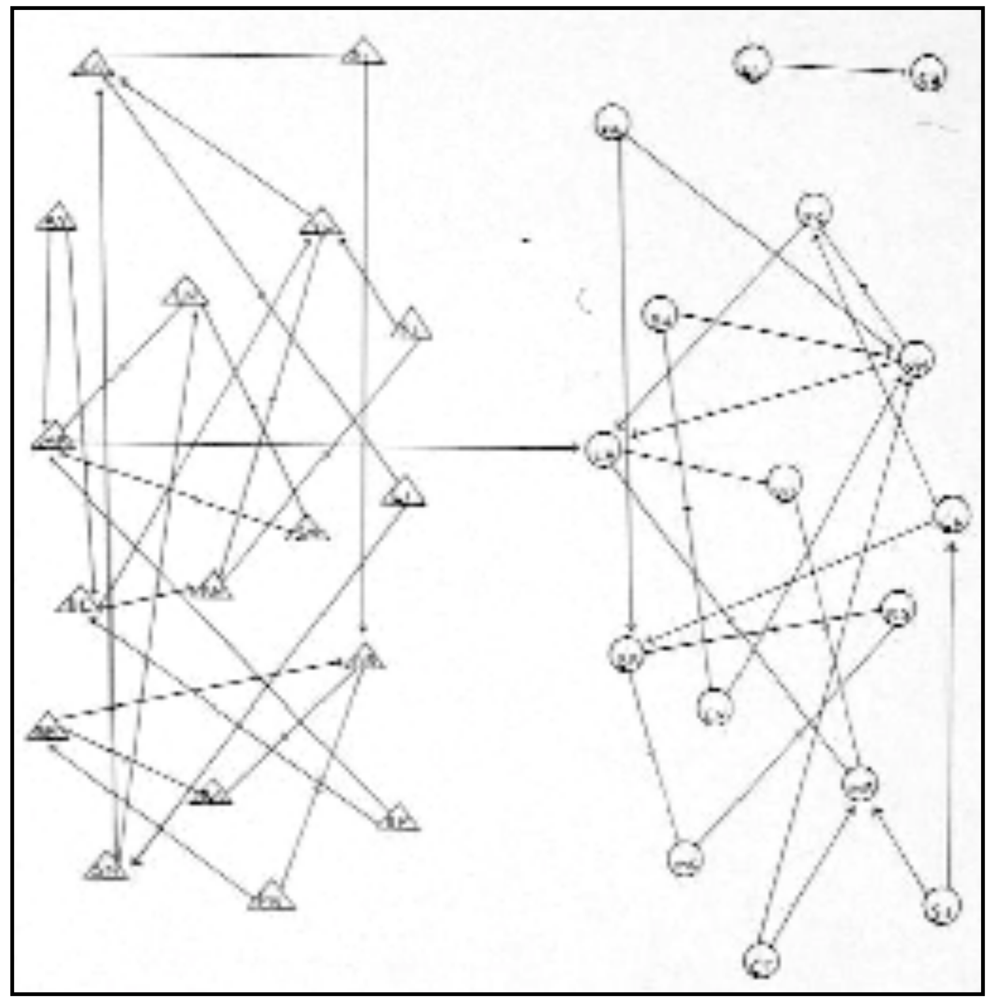
Computational Social Science

Revolutions in technology precipitate revolutions in science

Computational Social Science

Revolutions in technology precipitate revolutions in science

Revolution in computational resources
+ Availability of large-scale human data
+ Developments in statistics
= Computational social science



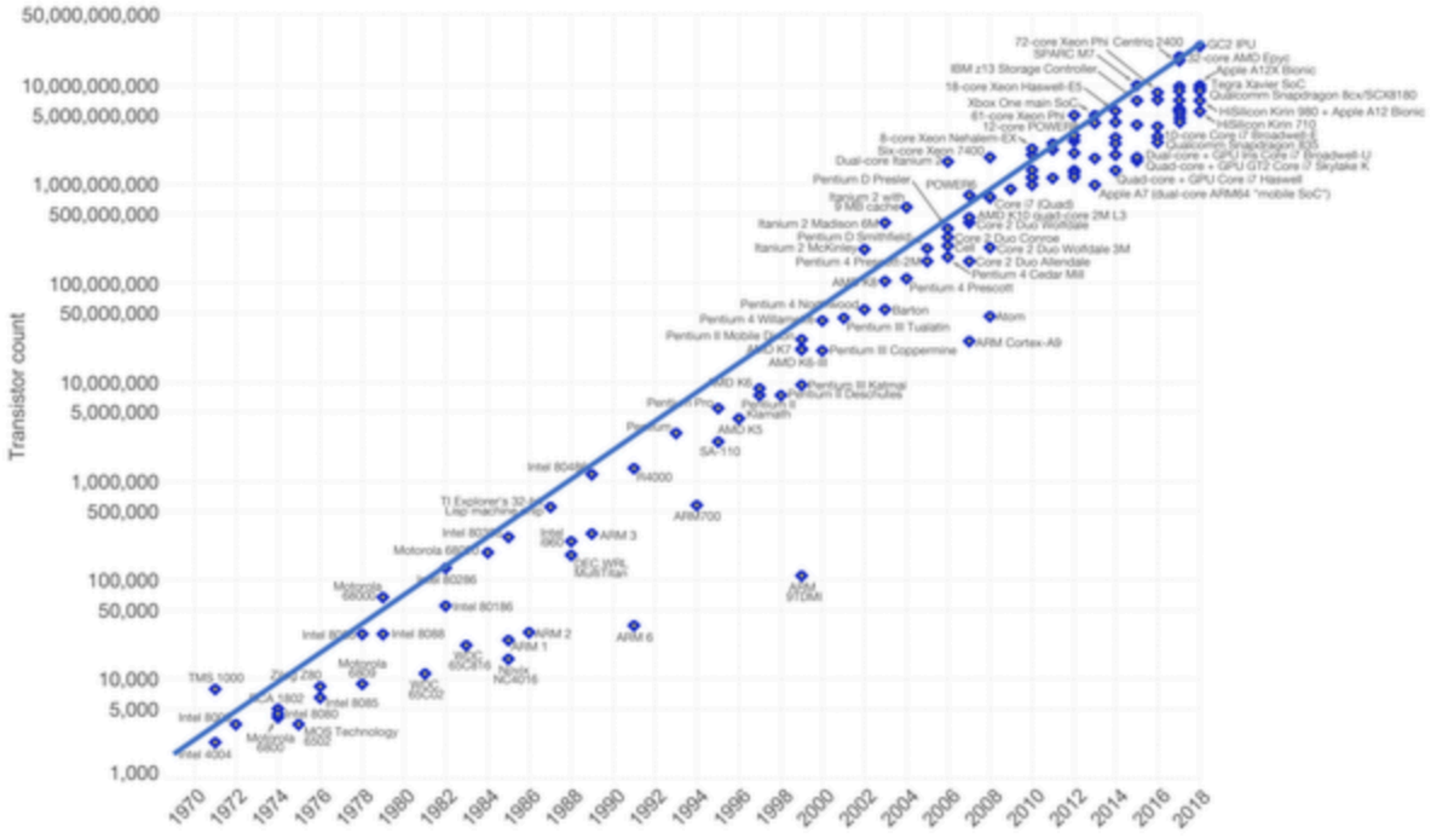
Computational Social Science

Revolutionary advances in computing power and data availability let us observe social phenomena in ways we couldn't before

CSS in a phrase:

peering through the socioscope

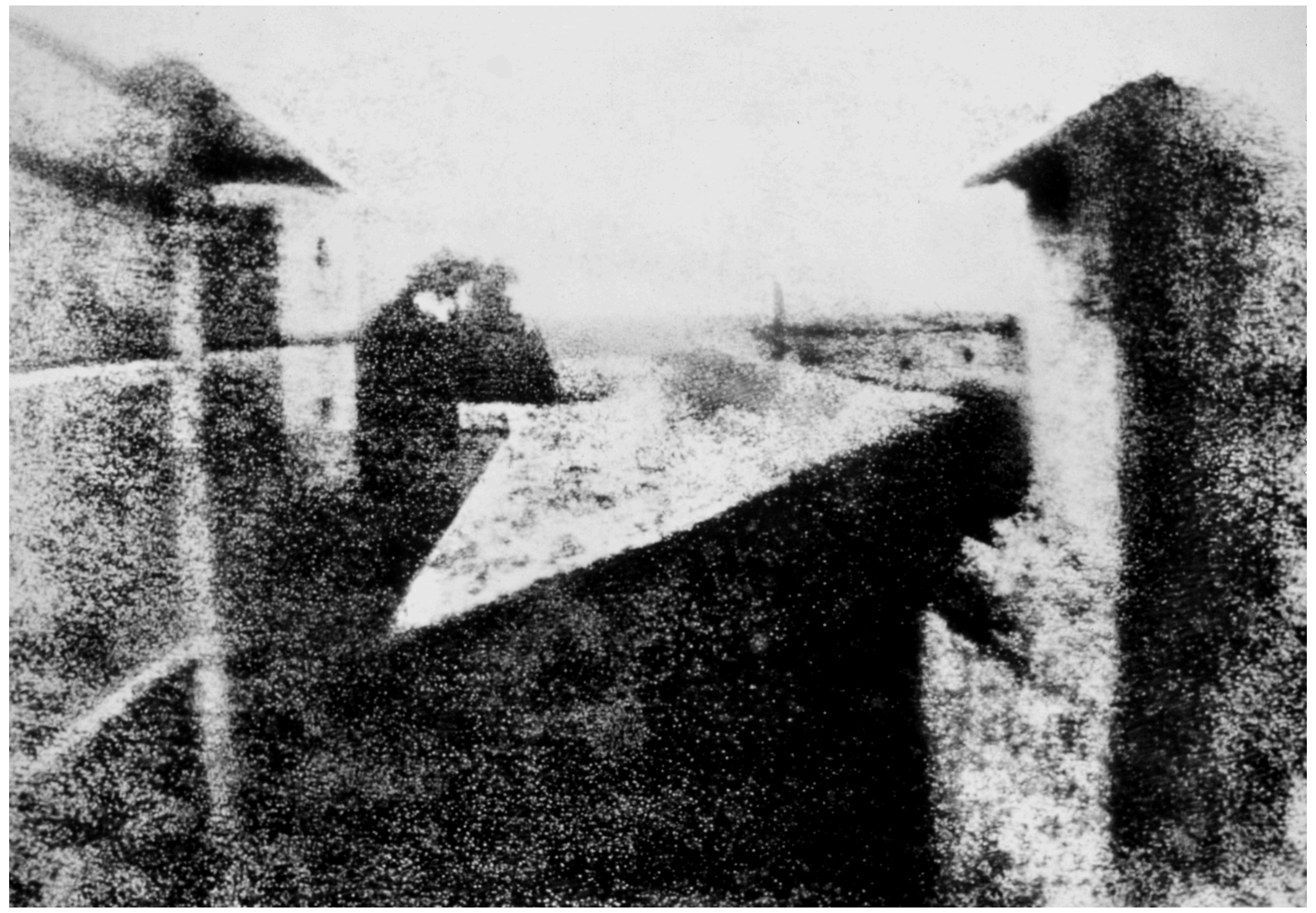
But wait… hasn’t this been happening for a long time?



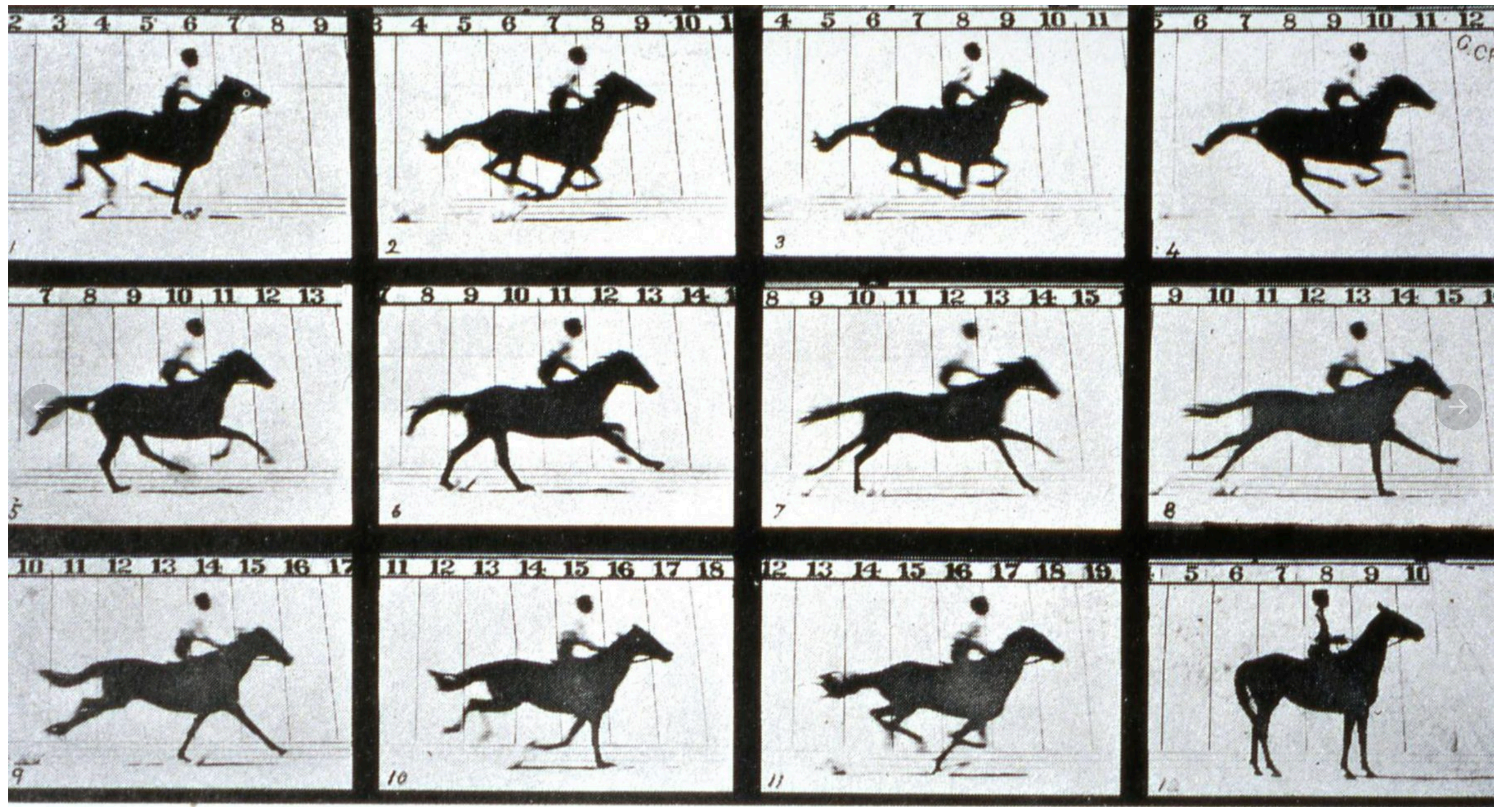
Moore’s law

A revolution in progress; a difference in kind

First photograph



First "moving pictures"



A movie is "just" a bunch of photos, but there is a qualitative difference

Similarly, social research has qualitatively changed

Course goals

- Learn the modern methods used to do social research in the digital age
- Develop research skills: reading papers, reviewing papers, presenting research, discussing research problems, doing a research project
- Emphasis on AI & Society

Course logistics

- 2 intro lectures by instructor
- 7 classes of student-led discussions of research papers
- 3 classes of student project presentations (1 proposal and 2 final)

Student responsibilities

- Write reviews of the main papers of the week before each class
- Lead a group discussion of a paper
- Do a final project on a topic related to the course
- 1–2 assignments to supplement class material

Reviews

- Not just a summary of the paper
- Briefly distill the paper, then summarize the paper's strengths and weaknesses
- How could it be extended?
- What is missing?
- What were the tradeoffs involved, and did the authors make the right compromises? Why or why not?

Group discussions

- Most of the class will be discussion-based group learning
- CSS is so new that the frontier is still very accessible!
- Everyone will get a chance to lead a discussion of a paper
- Come to class ready to discuss

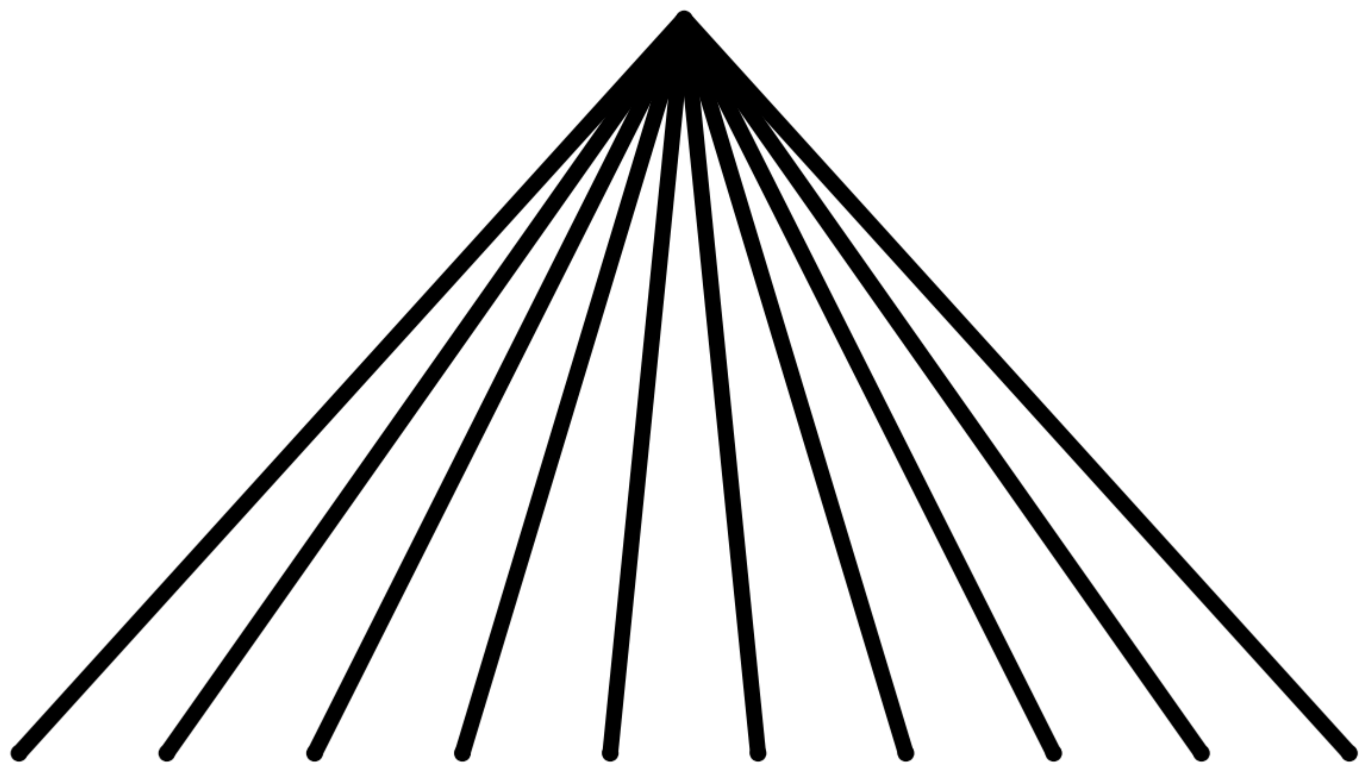
Final project

- Computational social science, like most computer science, is best learned by getting your hands dirty!
- Opportunity to do something tangible
- Example form of good project: implement a paper's analysis (new dataset?), extend in a non-trivial and interesting way, find something new
- Other project types too
- Lightning proposal presentations class; project presentation; project report

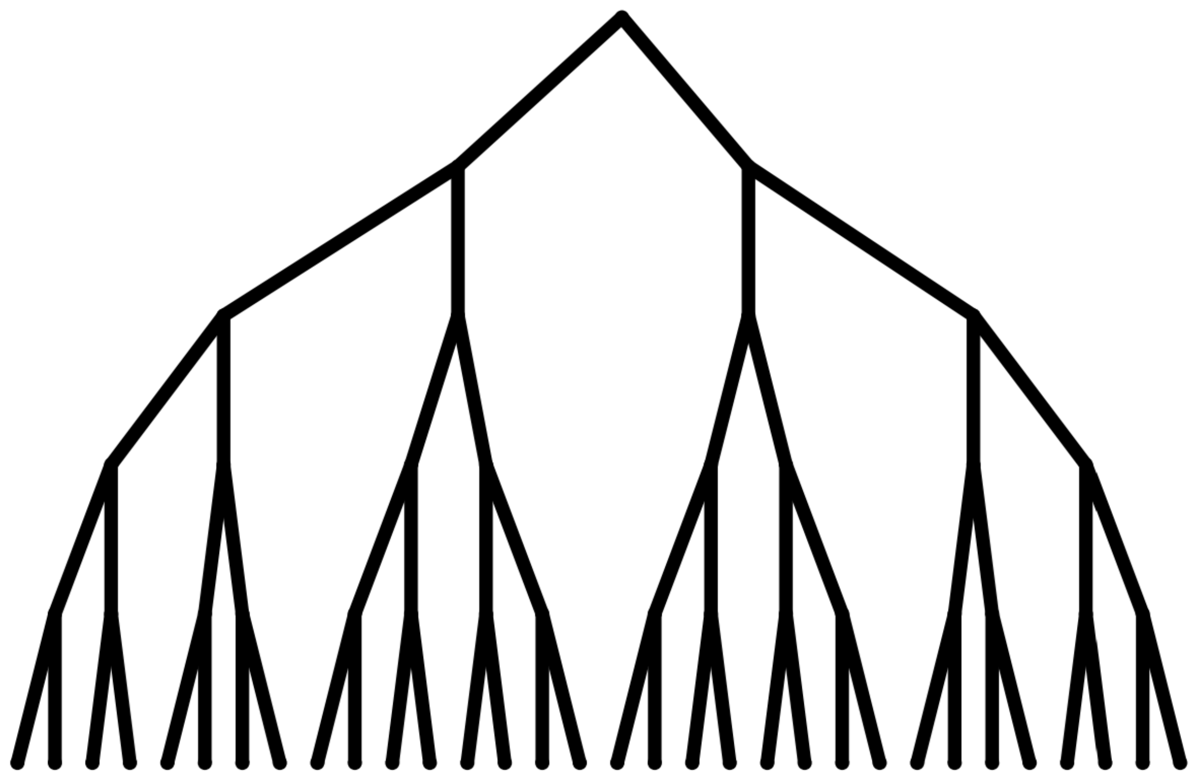
Back to the question

How do people in connected societies learn about new ideas, products, opinions, and beliefs?

Broadcast



Viral



Data

What data could we use to answer this question?

- Voting choices
- Reading habits
- Browsing histories
- Music preferences
- Purchasing behaviour
- …

The structural virality of online diffusion

[Goel, Anderson, Hofman, Watts 2015]

Question: how do links spread through online social networks?

Data: 1 billion links to videos, news stories, images, and petitions on Twitter

Methodological challenges

What is “influence”?

How to infer influence?

Methodological challenges

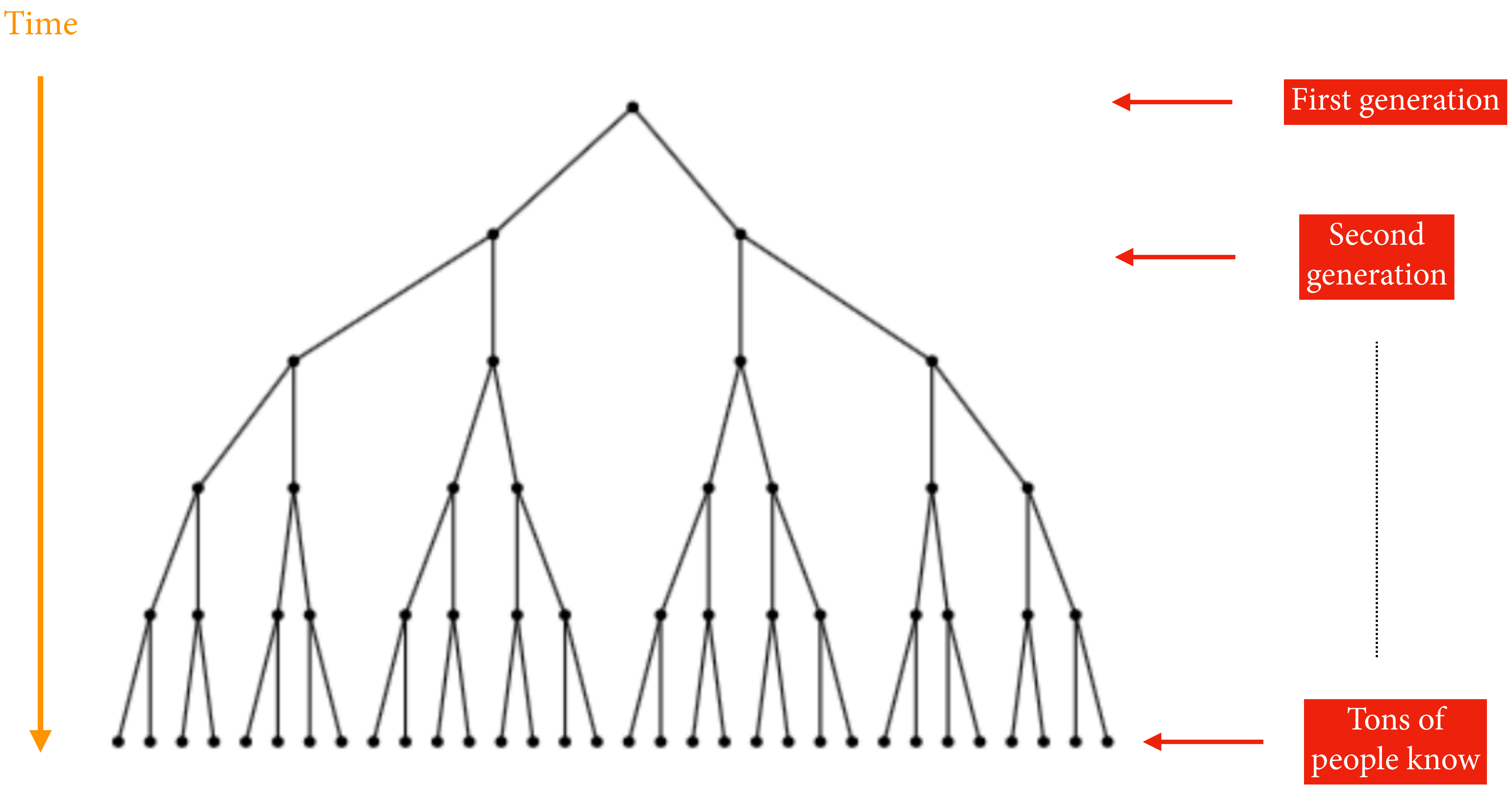
How to quantify structure?

What is “virality”?

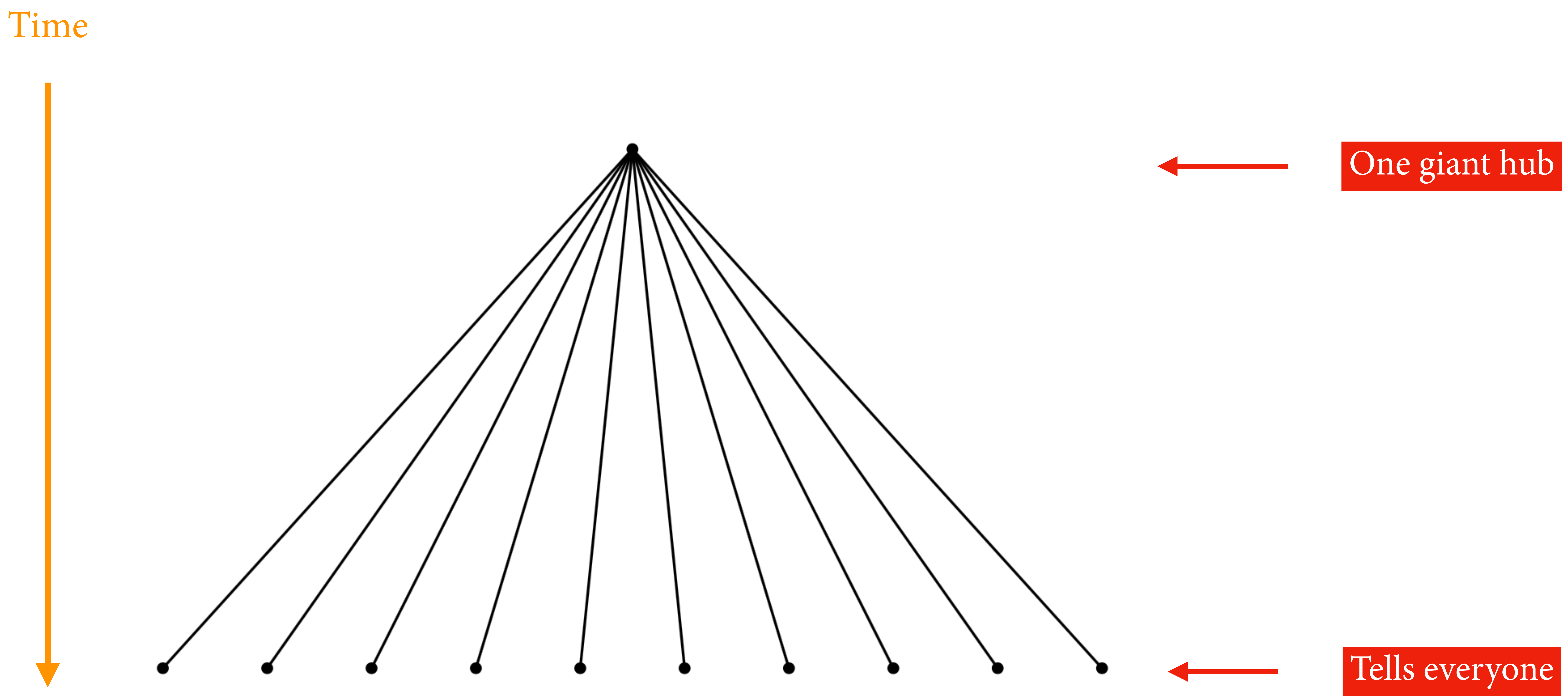
Methodological challenges

How do you analyze 1 billion cascades?

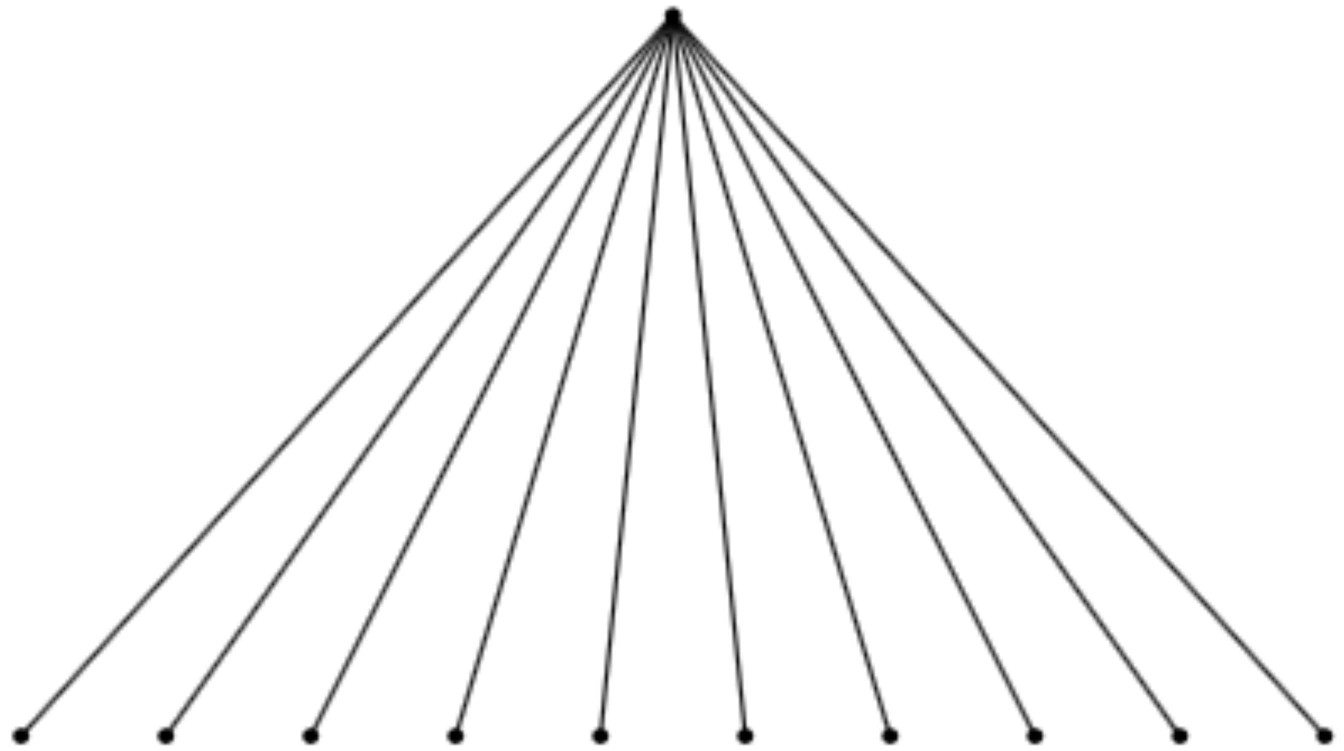
Viral diffusion



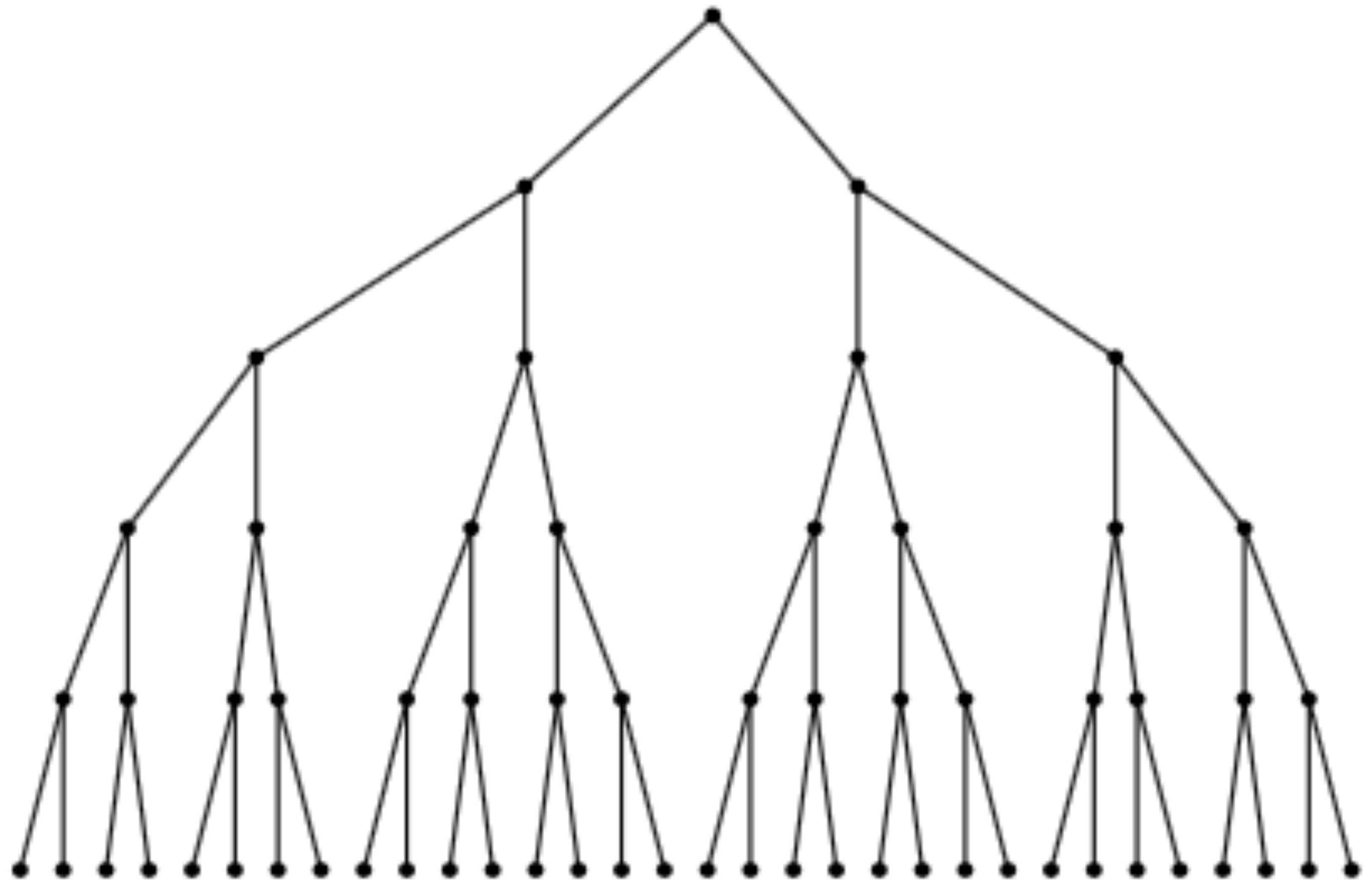
Broadcast diffusion



Which is it?



or



"Broadcast"

"Viral"

- Big media (CNN, BBC, NYT, Fox)
- Celebrities (Biebs, Taylor Swift)

- Organically spreading content
- Chain letters

How to study information spread?

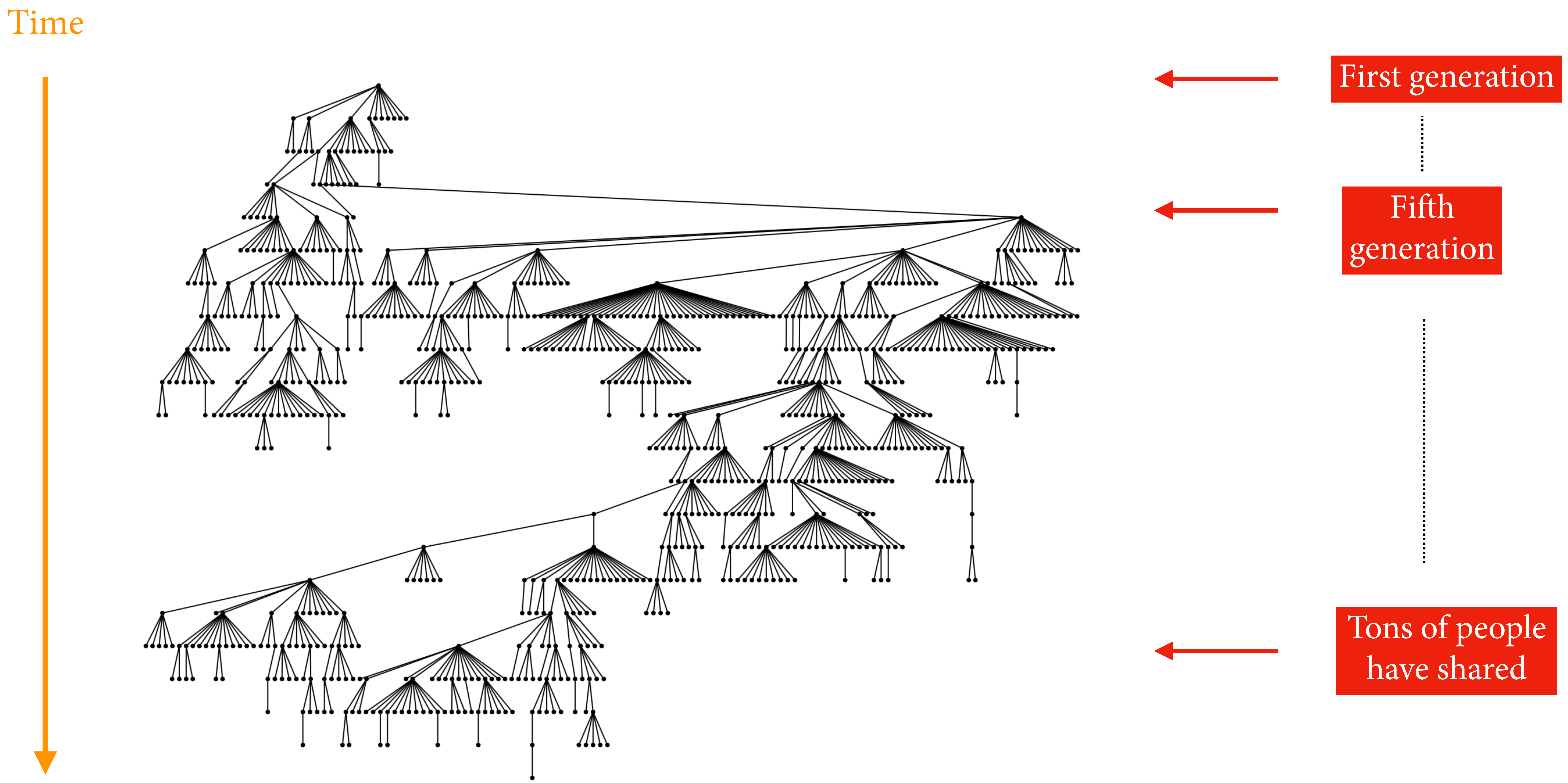
Hard to track “information” spreading from one mind to another

Online proxy: people sharing URLs

Twitter: person A tweets a URL, then a friend B tweets it (or directly retweets)
We say the URL passed from A to B

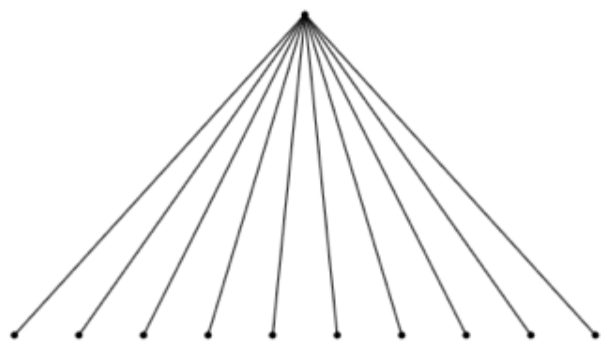
How to study information spread?

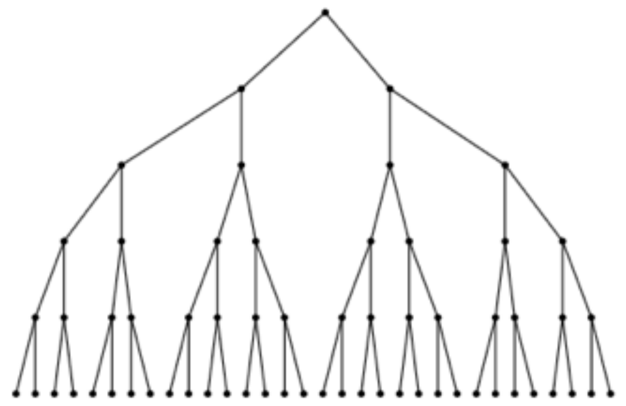
Connect these sharing edges into trees



How to measure virality?

How structurally viral is a particular cascade?

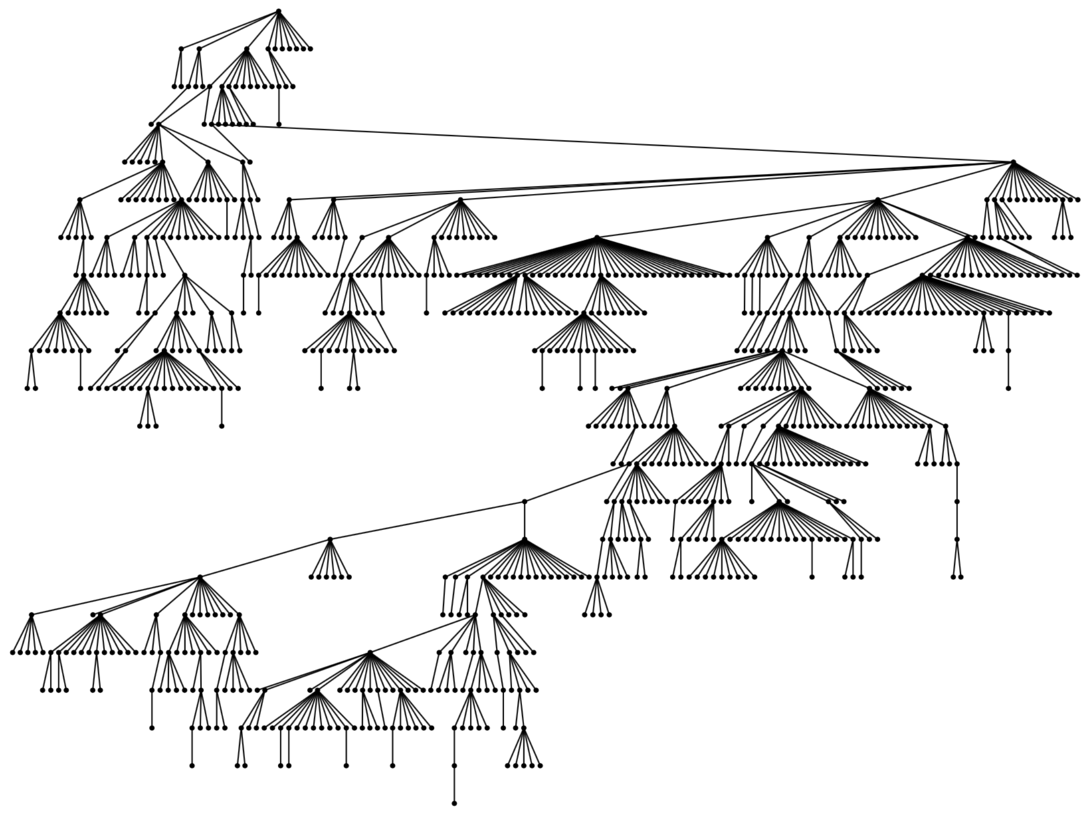




Not viral

?

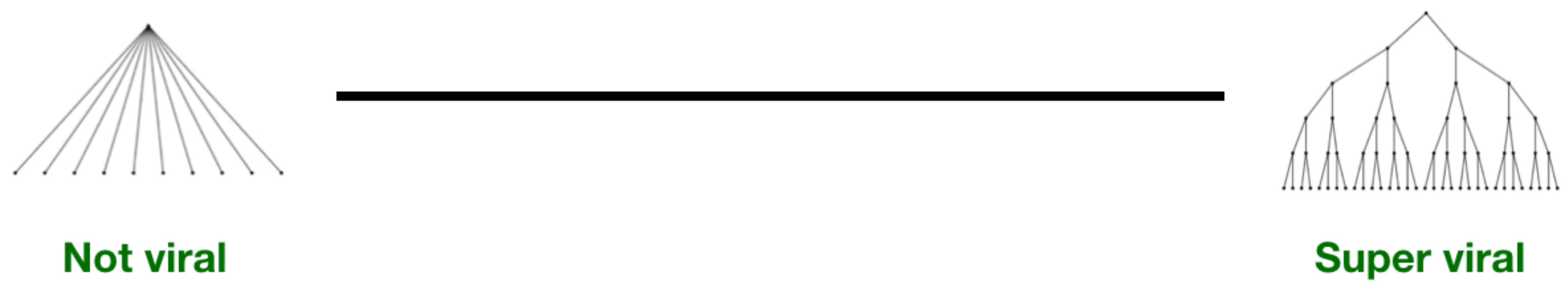
Super viral



How to measure virality?

One idea: depth of the cascade

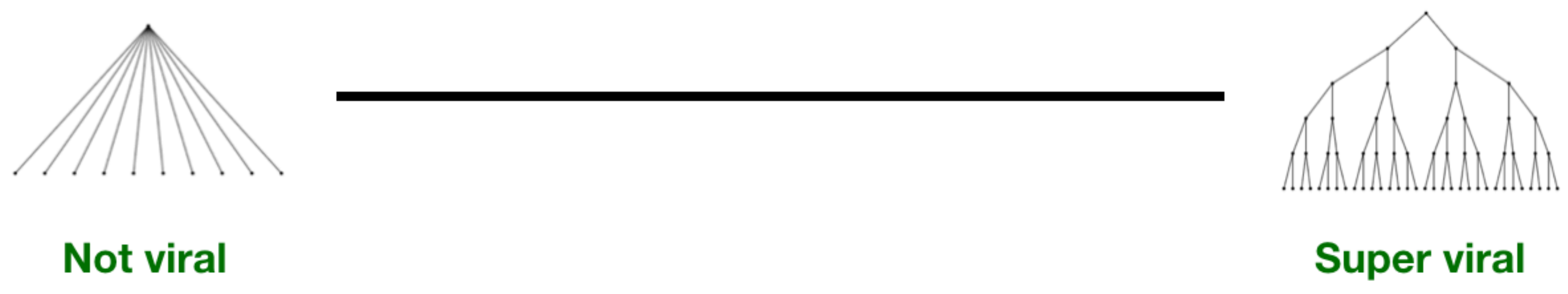
But this is sensitive to a single long chain



How to measure virality?

Another idea: average depth of the cascade

But even this sometimes fails: long chain then a big broadcast



How to measure virality?

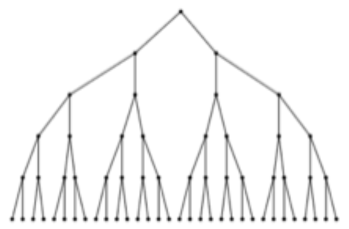
Solution: average path length between nodes

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$

Simple average!

Originally studied in mathematical chemistry [Wiener 1947] → “Wiener index”

Not viral



Super viral

Measure virality in data!

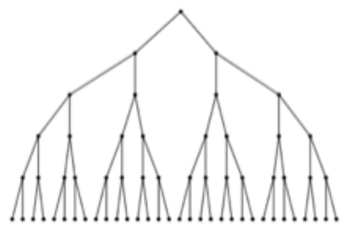
Now we have a way to construct information cascades on Twitter

And for each cascade we can compute a number that determines how "structurally viral" it is

So how often does stuff go viral?

Not viral

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$



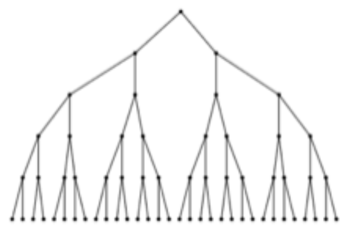
Super viral

Measure virality in data!

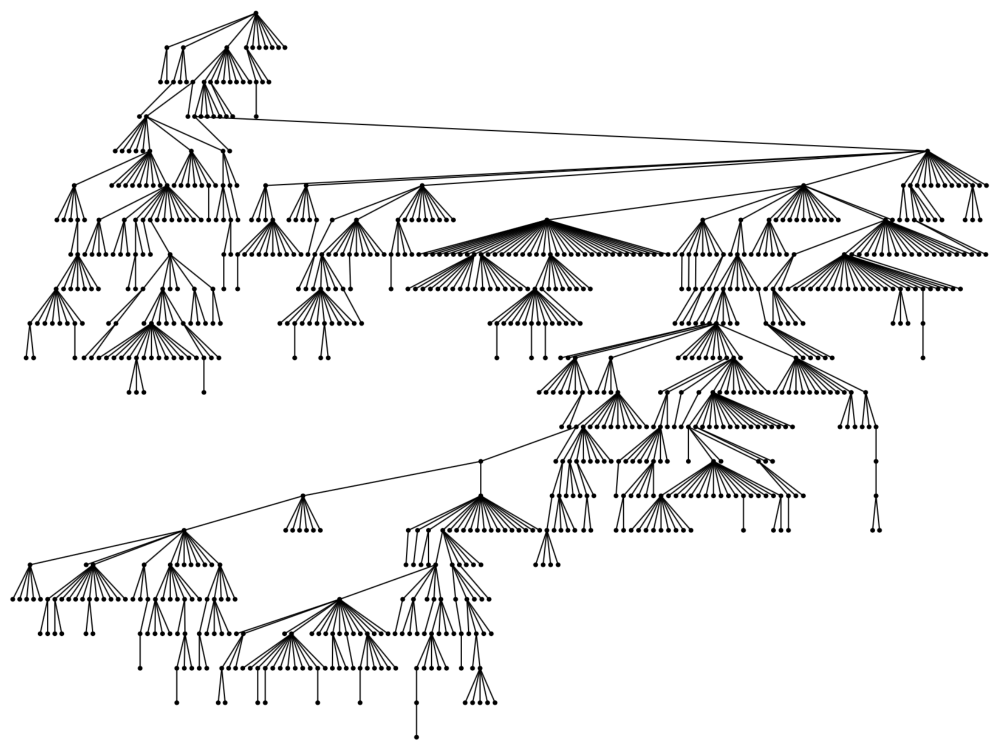
- Looked at an entire year of Twitter data
- 622 million unique URLs, 1.2 billion “adoptions” (tweets) of these URLs
- Every URL is associated with a forest of trees

Not viral

$$\nu(T) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}$$



Super viral

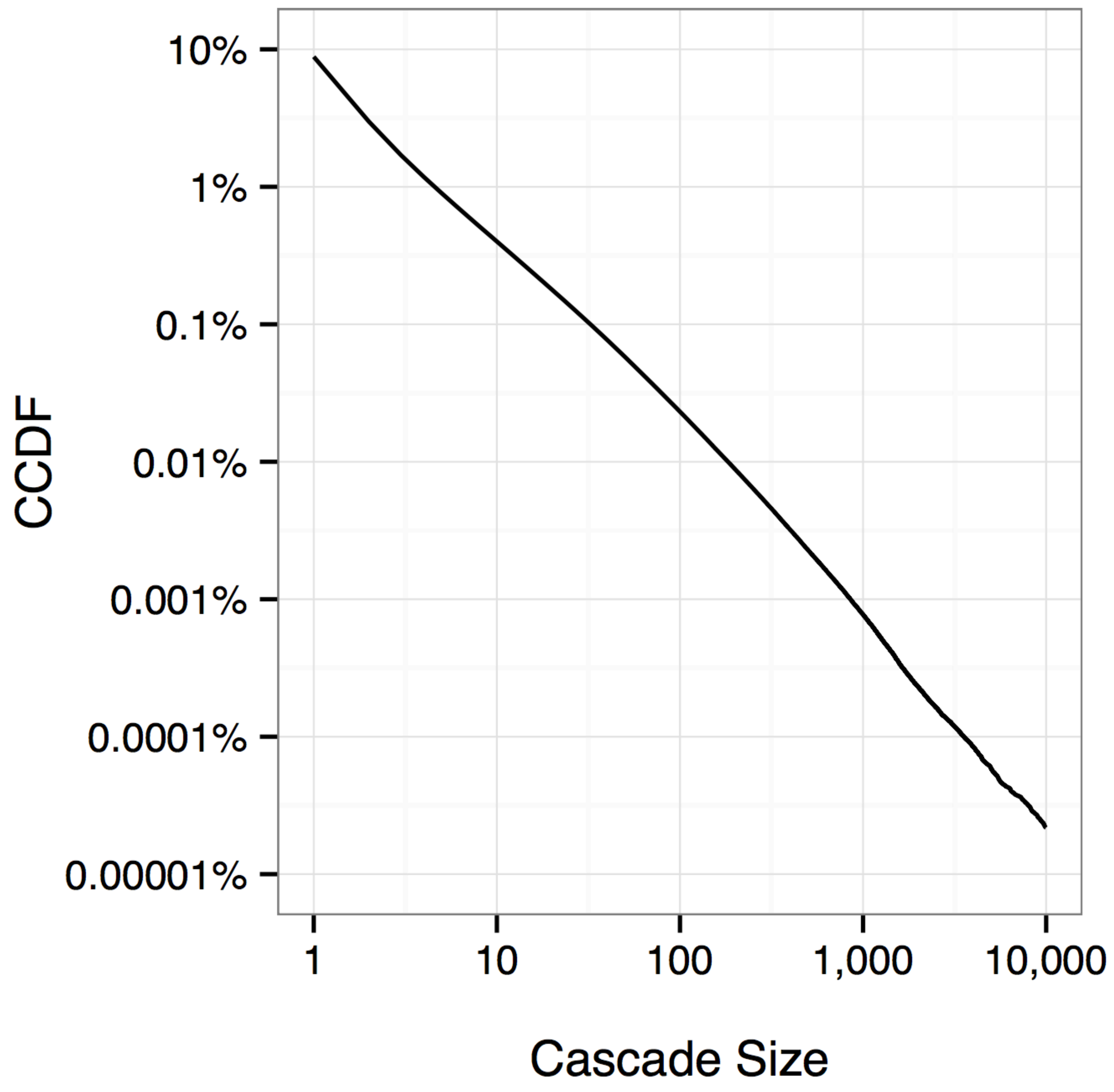


Measure virality in data!

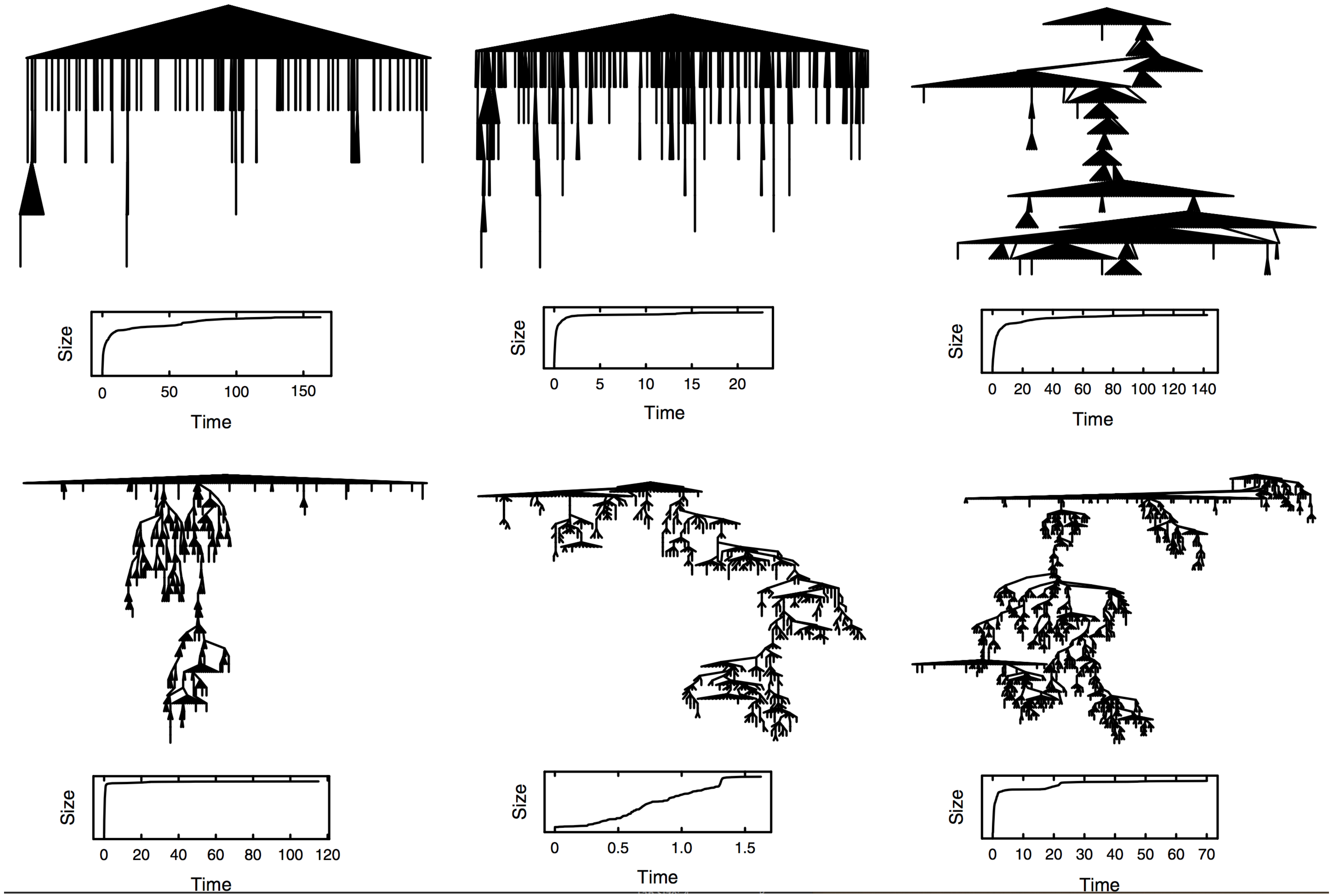
First conclusion: most stuff goes nowhere

Average cascade size: 1.3

Not very interesting cascades: focus on trees of size at least 100 (empirically 1/4000)



A new look into how ideas travel

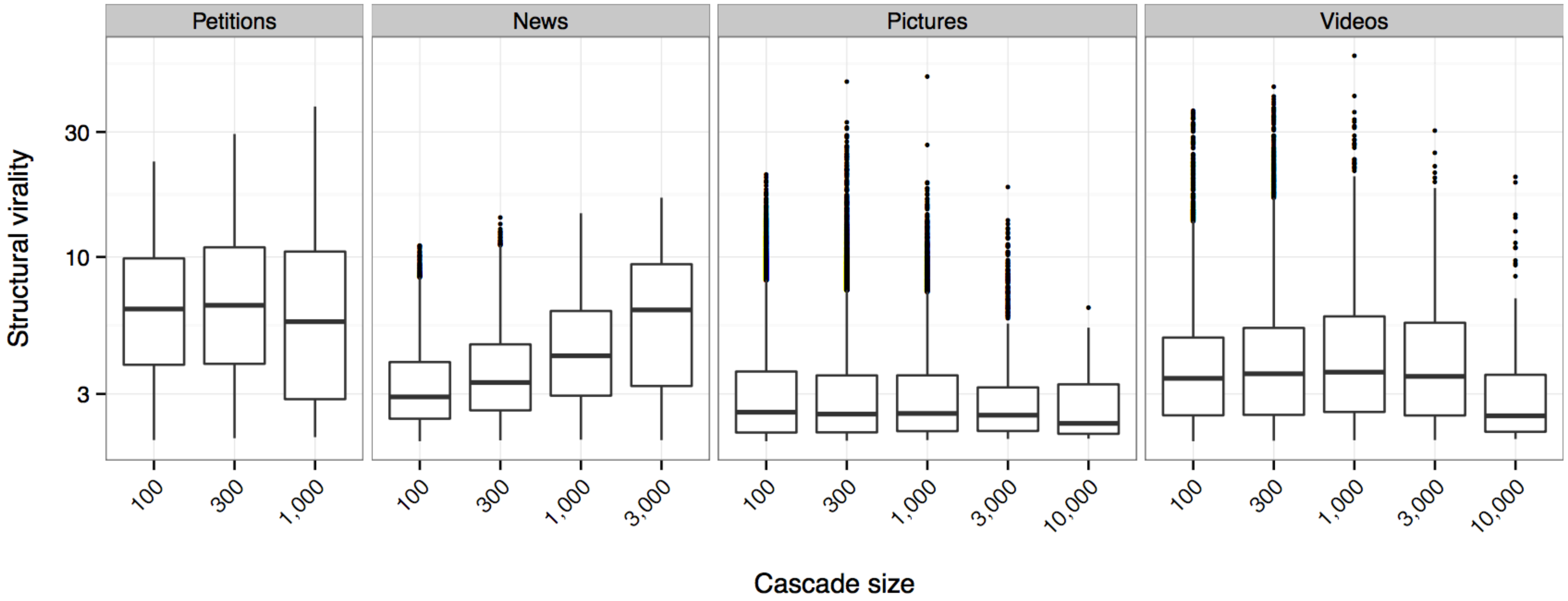


Surprising diversity at every scale

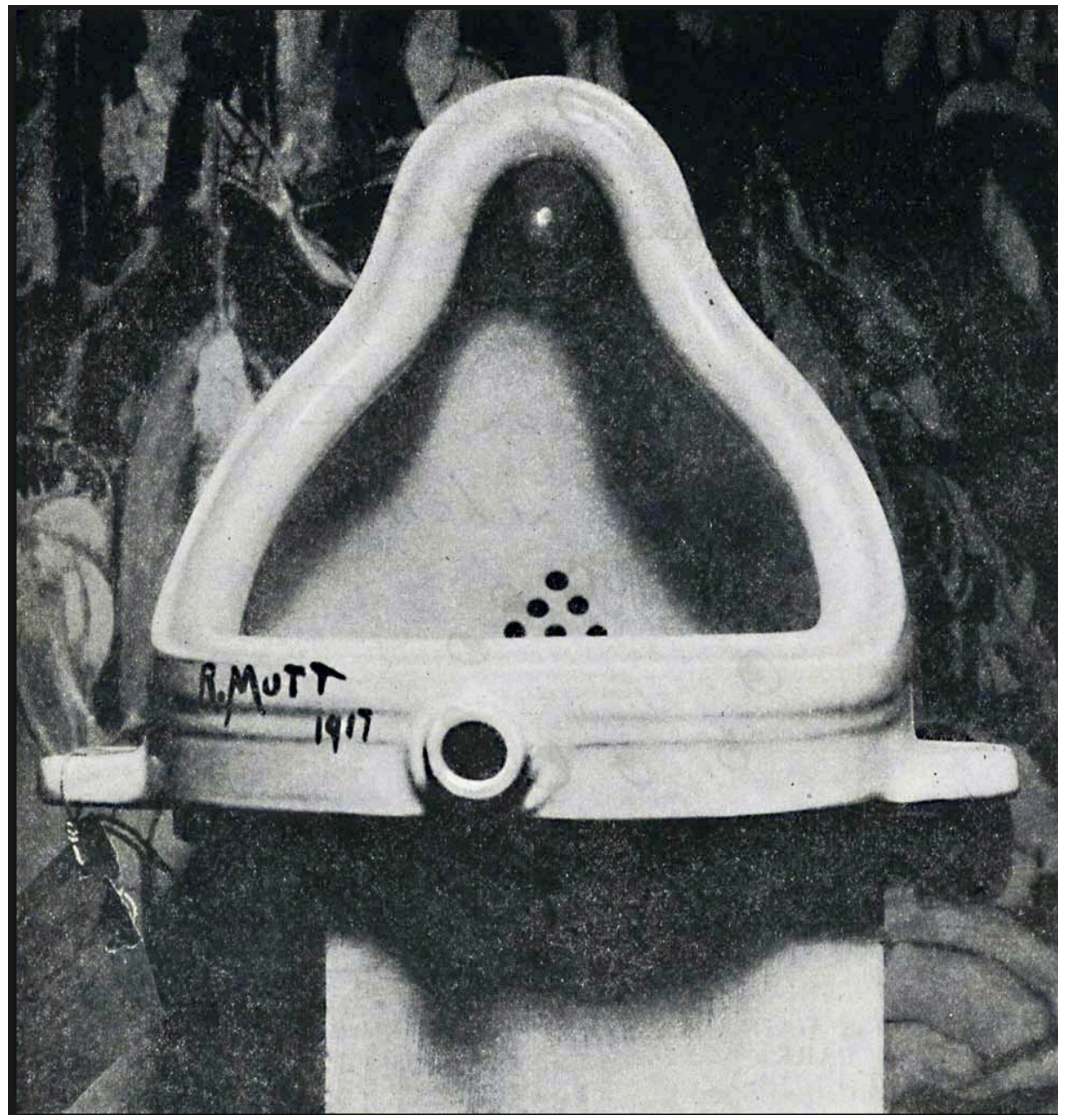
Across domains and across sizes, we see lots of different types of structures from broadcast to viral

Very low correlation between size and virality!

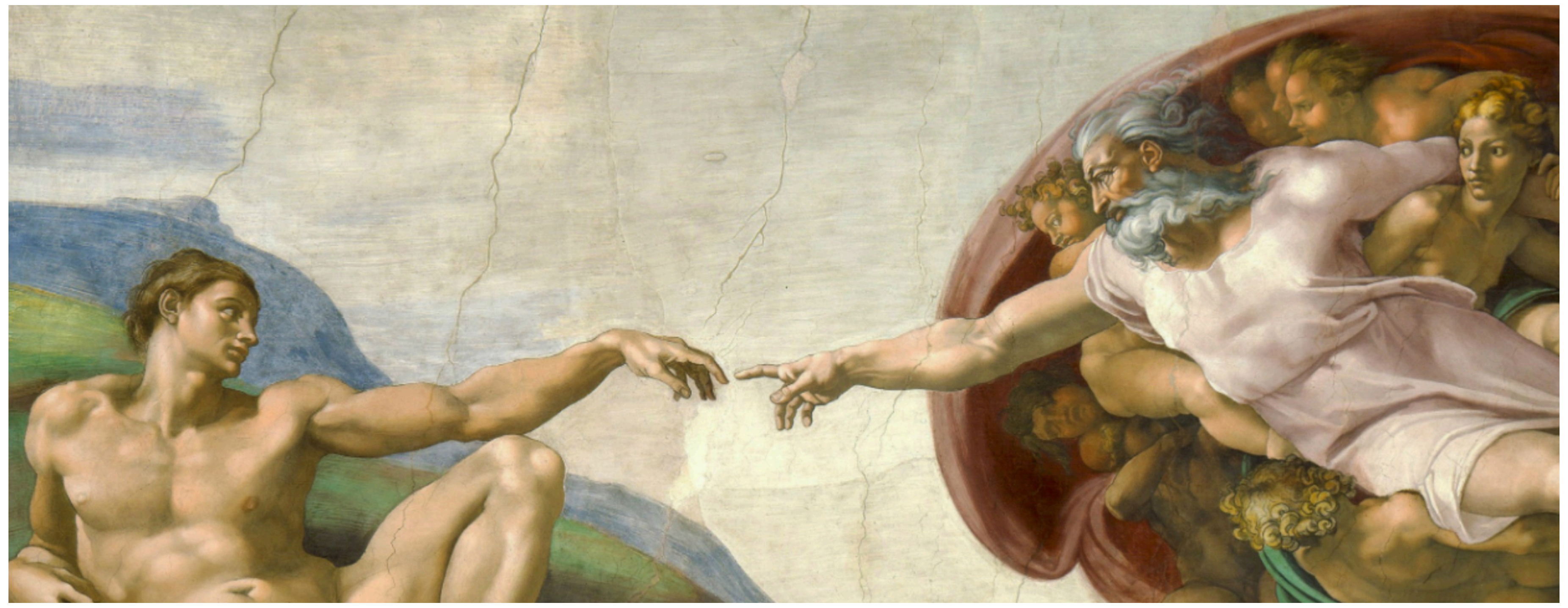
This means something about the world: big things aren't always viral OR broadcast



Ways of doing computational social science

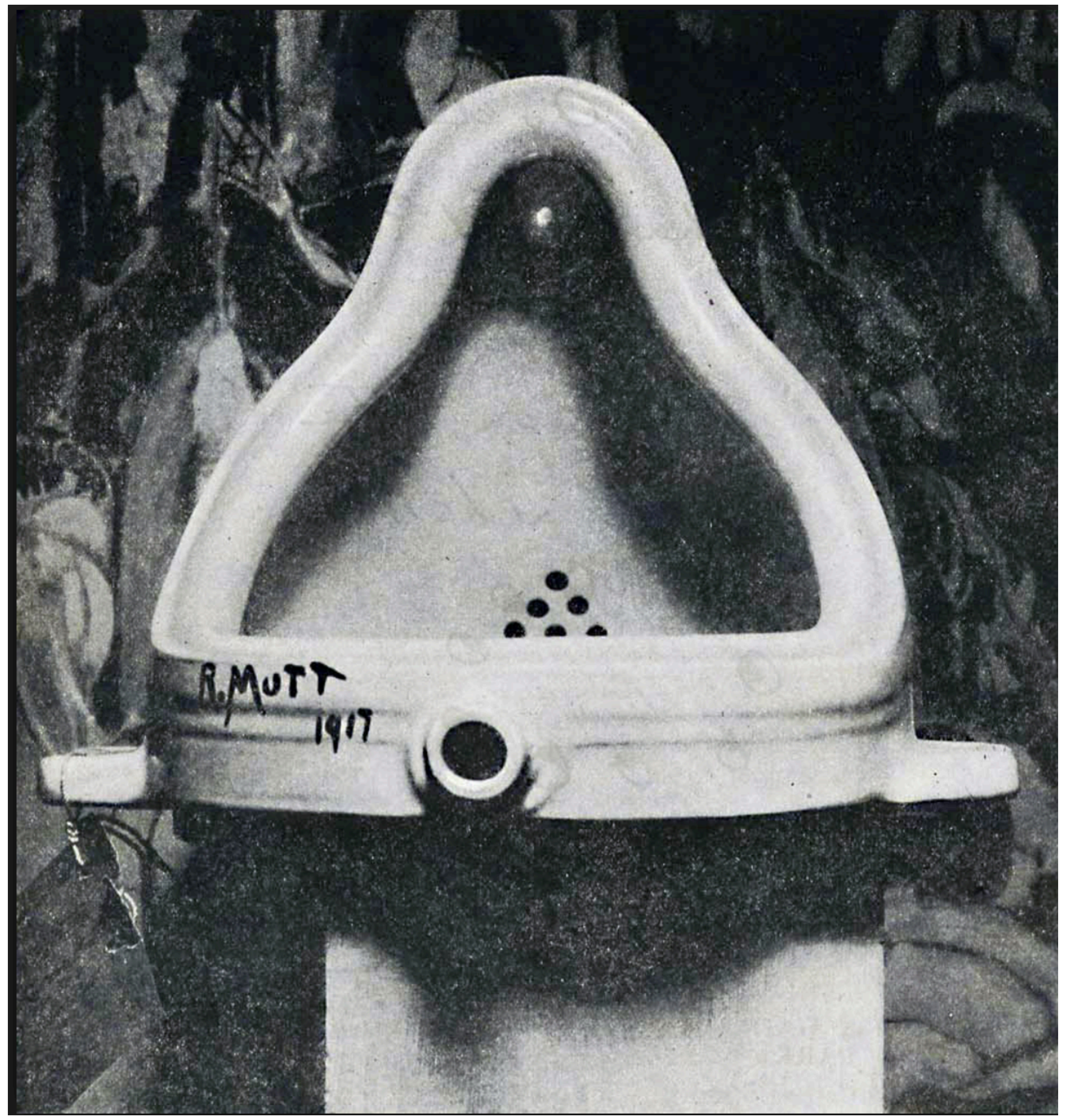


Readymades

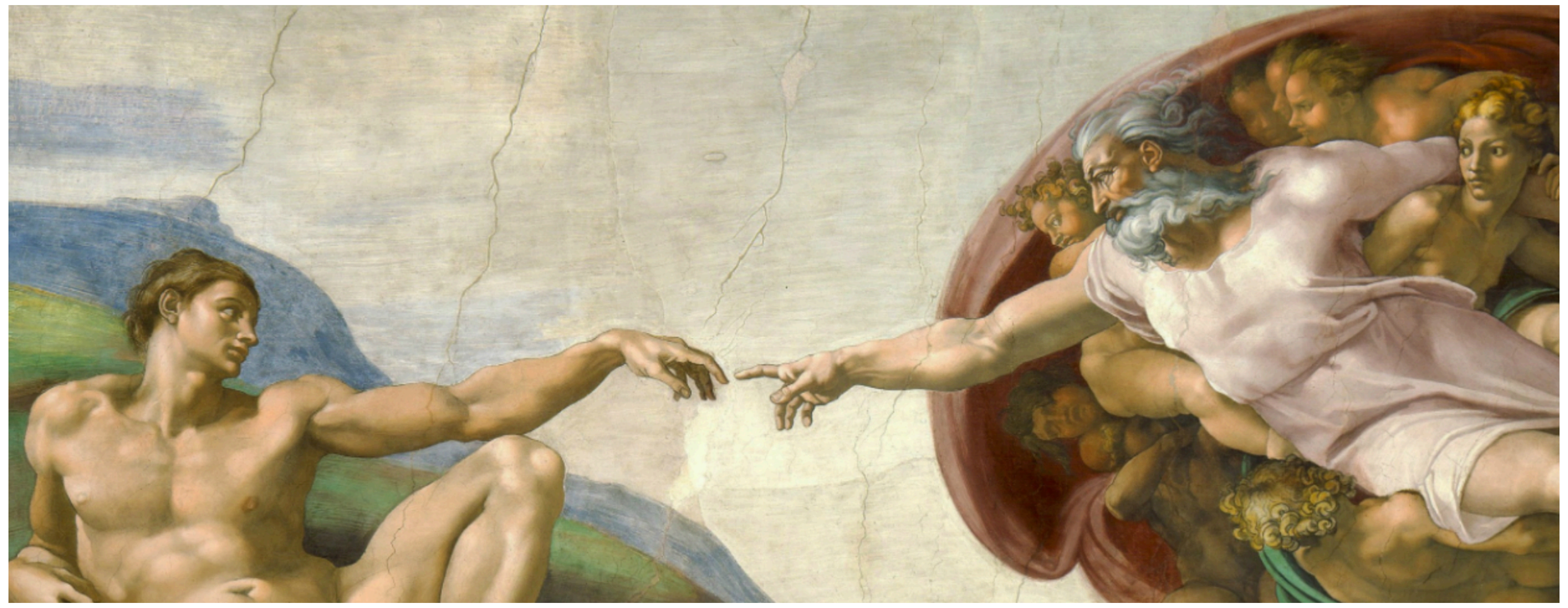


Custommades

Ways of doing computational social science



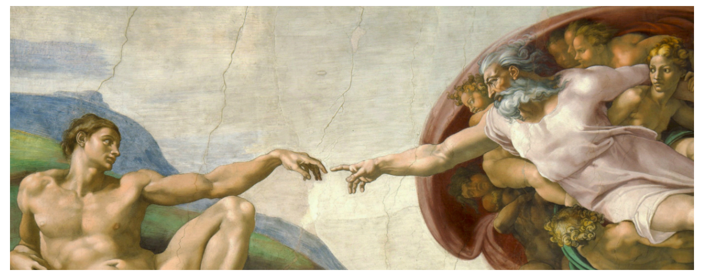
"Found" data



Experiments

A spectrum between the two

Ways of doing computational social science



Observational analyses

Human computation

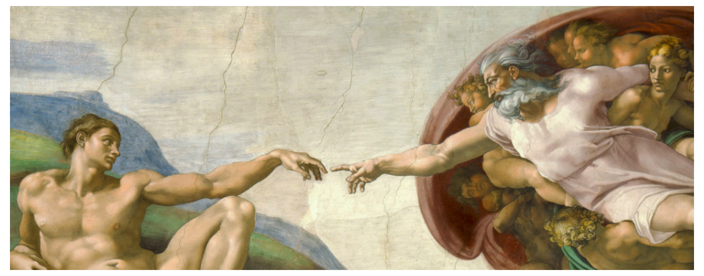
Natural experiments

Surveys

Field experiments

Lab studies

Ways of doing computational social science

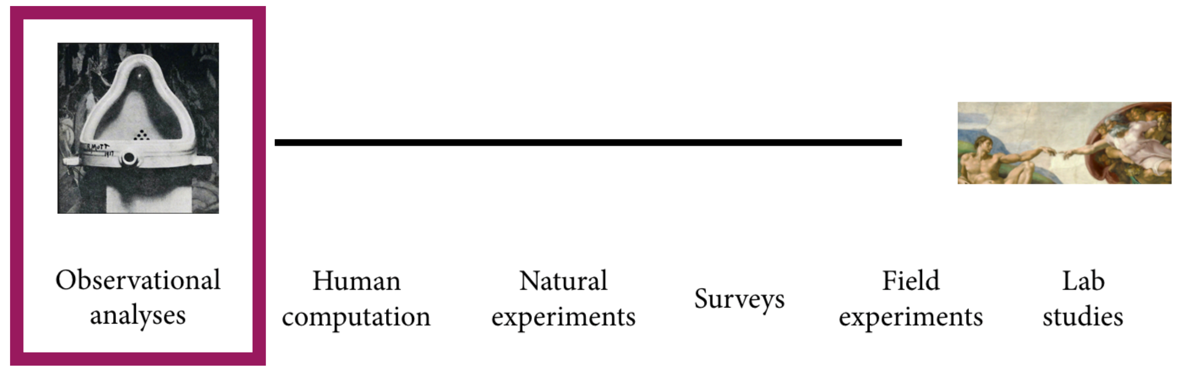


Observational analyses | Human computation | Natural experiments | Surveys | Field experiments | Lab studies

Observational analyses of existing data

- Massive datasets of all kinds of human behaviour are now available for study
  - Wikipedia, GPS traces, health databases, Facebook, Twitter, Reddit, reviews, purchases, dating, invitations, exercise apps, etc., etc…
- Key part of the “socioscope”: huge traces of things that we couldn’t see before
- Lack of detail/fidelity in individual records is hopefully made up for by large numbers of records (small noisy errors cancel out, big patterns are signal)

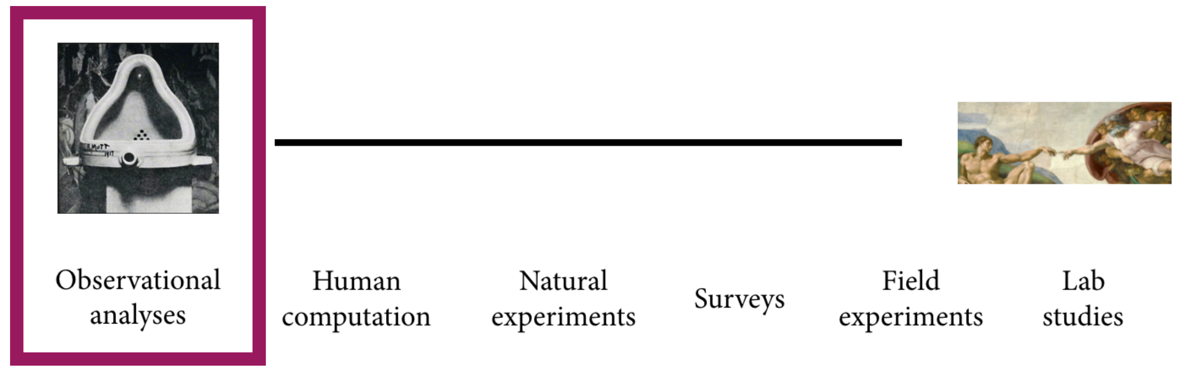
“Big data” / “Found data”



Ten common characteristics of big data

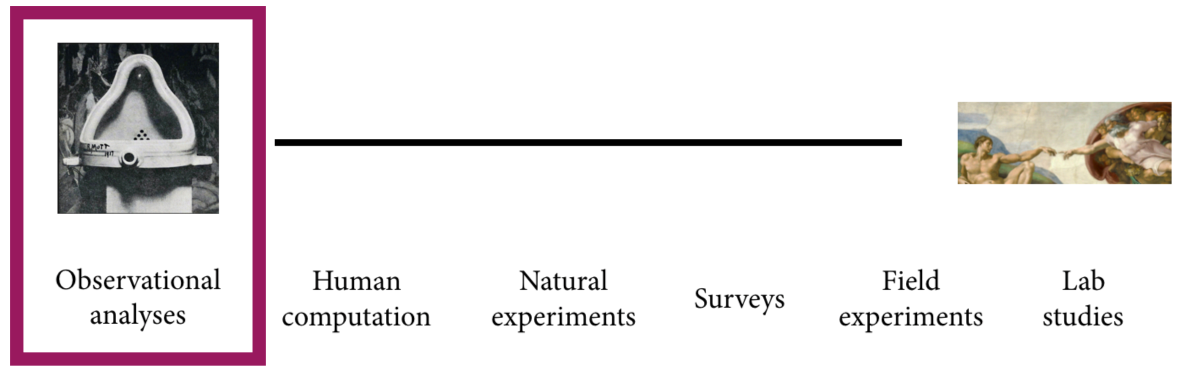
- Big: statistical power, rare events, fine resolution
- Always-on: unexpected events, real-time measurement
- Nonreactive: measurement probably won't change behaviour

- Incomplete: probably won't have the ideal information you want
- Inaccessible: difficult to access (gov't, companies)
- Nonrepresentative: bad out-of-sample generalization (good in-sample)
- Drifting: Population drift, usage drift, system drift
- Algorithmically confounded: want to study behaviour, not an algorithm
- Dirty: Junk, spam
- Sensitive: Private, hard to tell what's sensitive



Observing Behaviour: Three research strategies

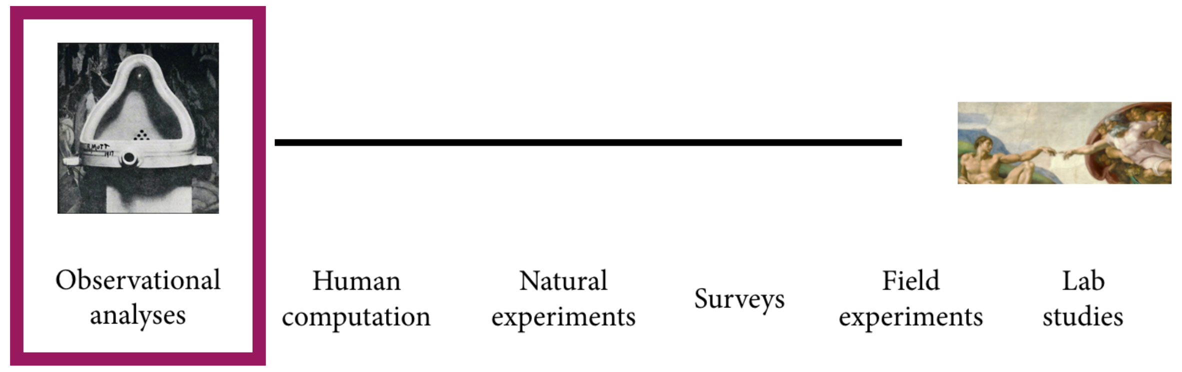
1. Counting things
2. Forecasting/nowcasting
3. Approximating experiments



Observing Behaviour: 1. Counting Things

Example: Measuring viral vs. broadcast diffusion on Twitter

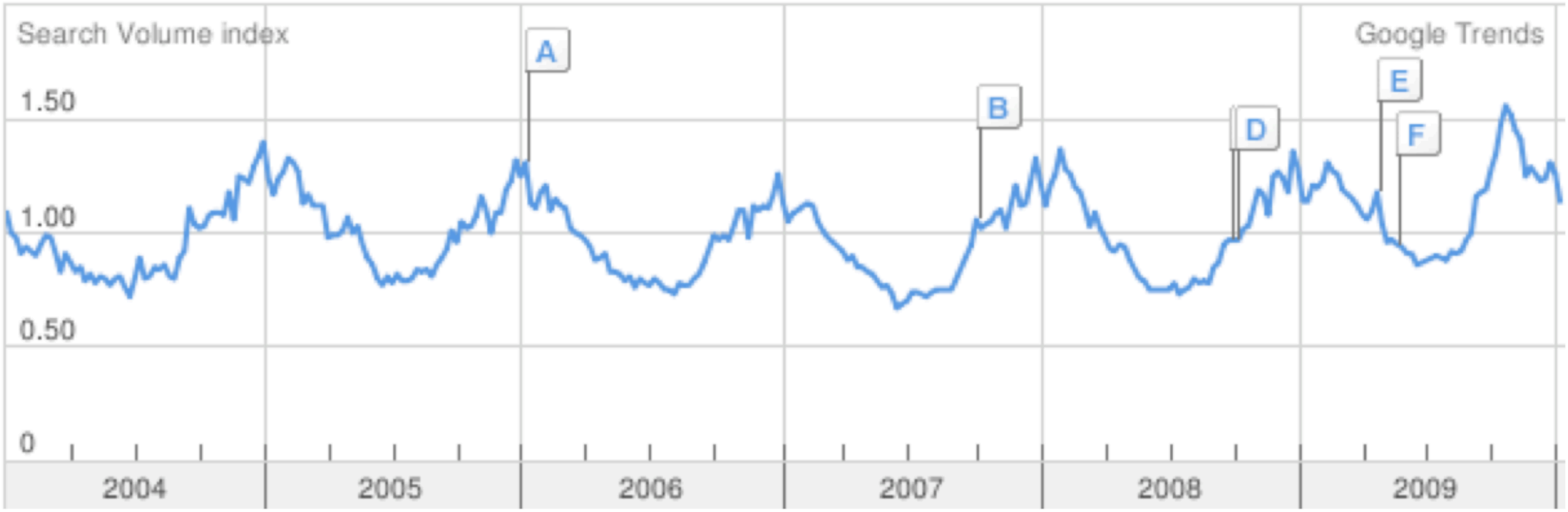
With newfound datasets and computational resources, many valuable initial contributions are measurements of quantities we couldn't measure before → counting at scale



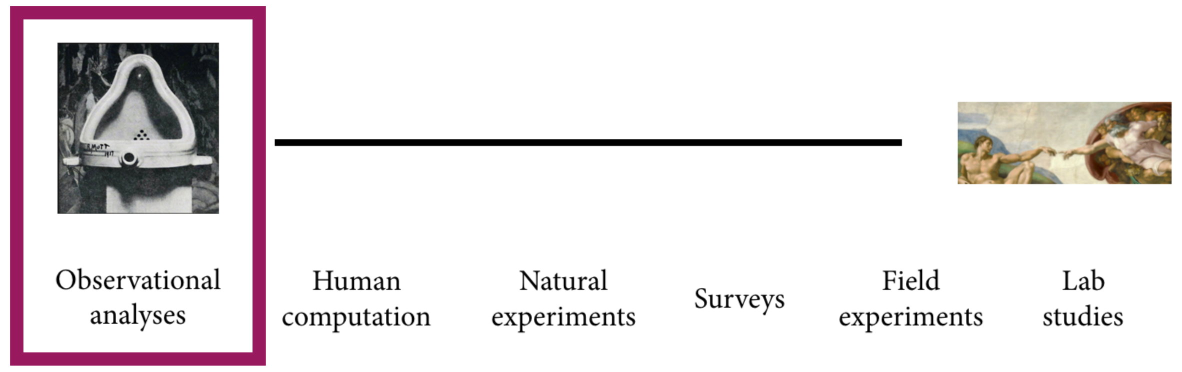
Observing Behaviour: 2. Nowcasting

Google Flu Trends

Idea: find 50 most correlated search query volume trends with flu data

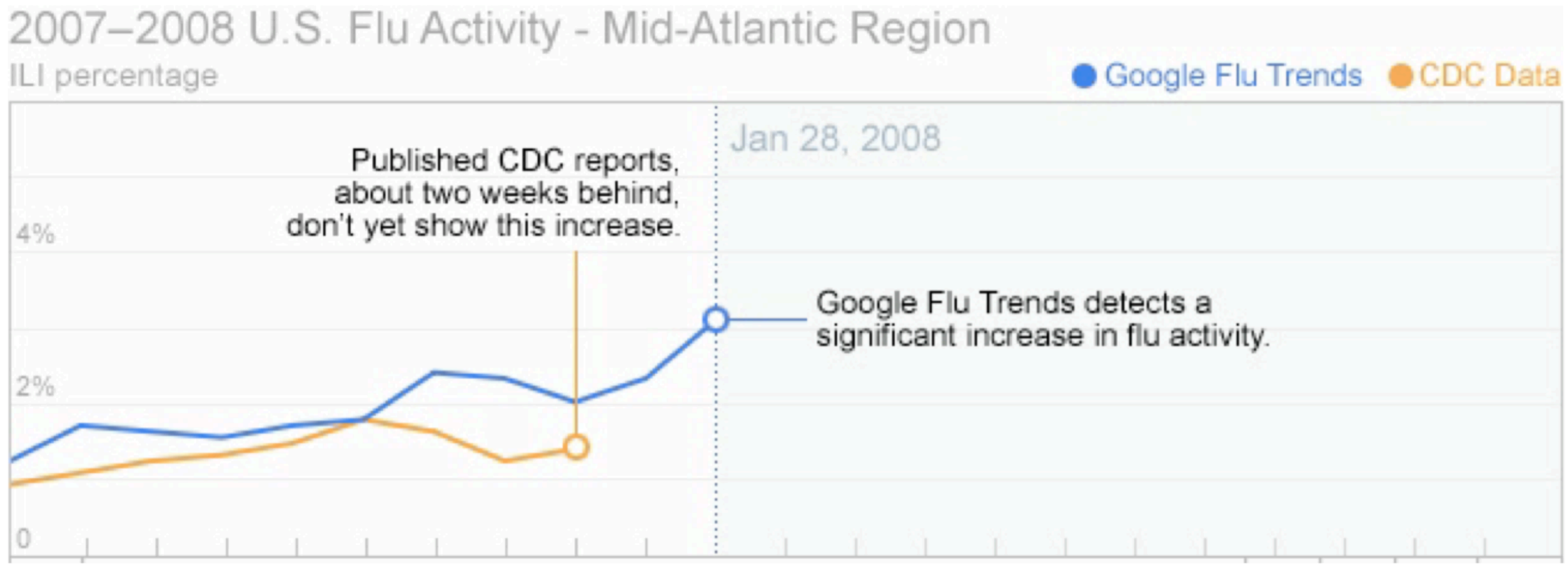


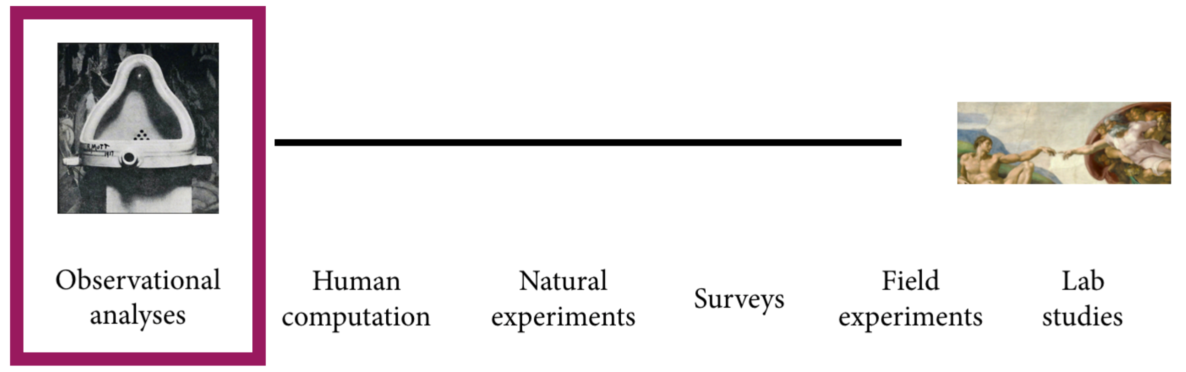
Search volume for the term “cough”



Observing Behaviour: 2. Nowcasting

The flu has a 1-2 week lag from when cases are reported to when the CDC releases official stats





Observing Behaviour: 2. Nowcasting

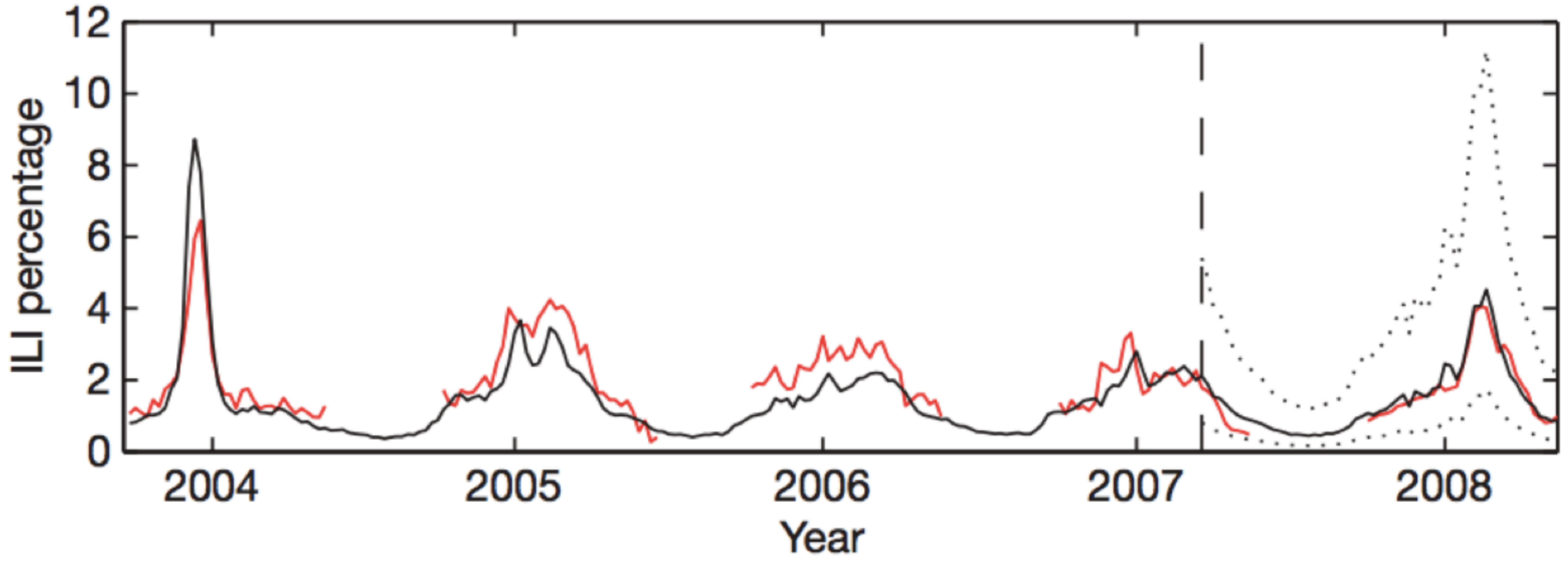
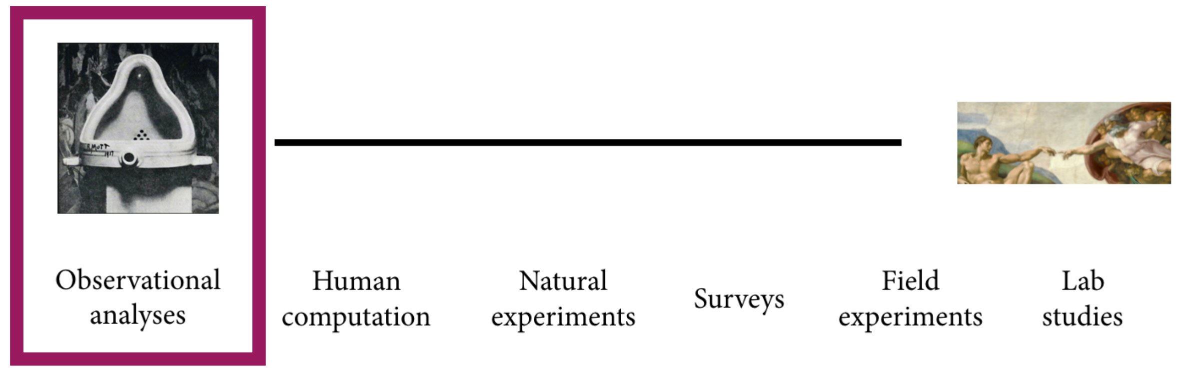
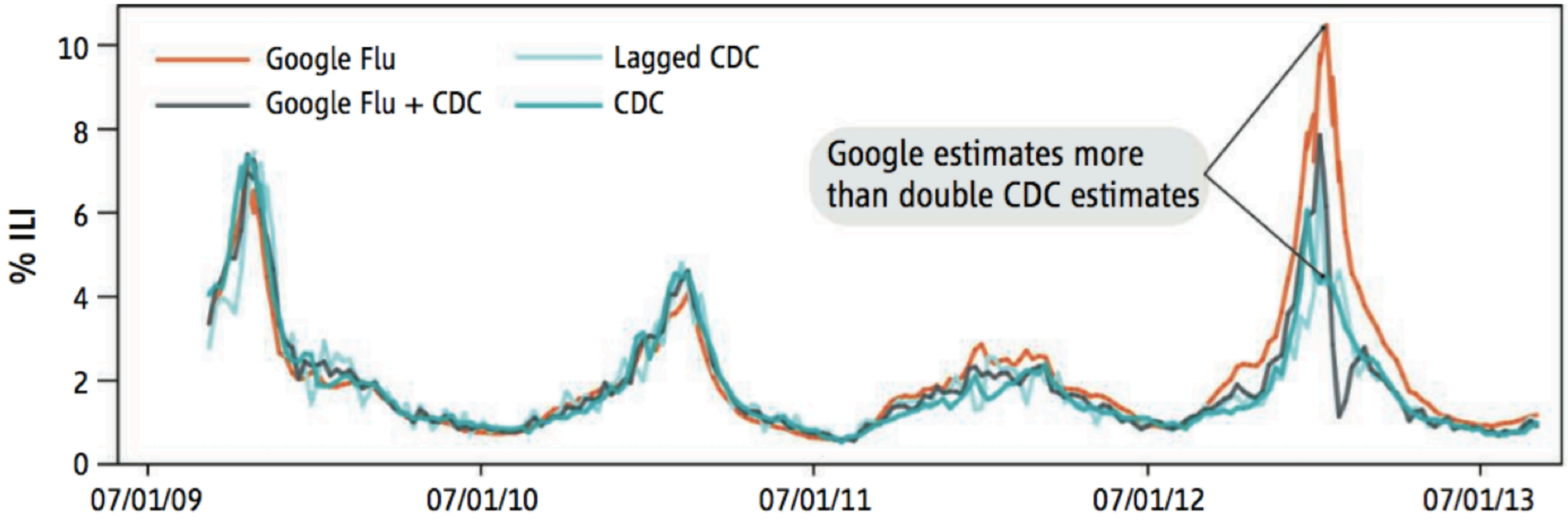


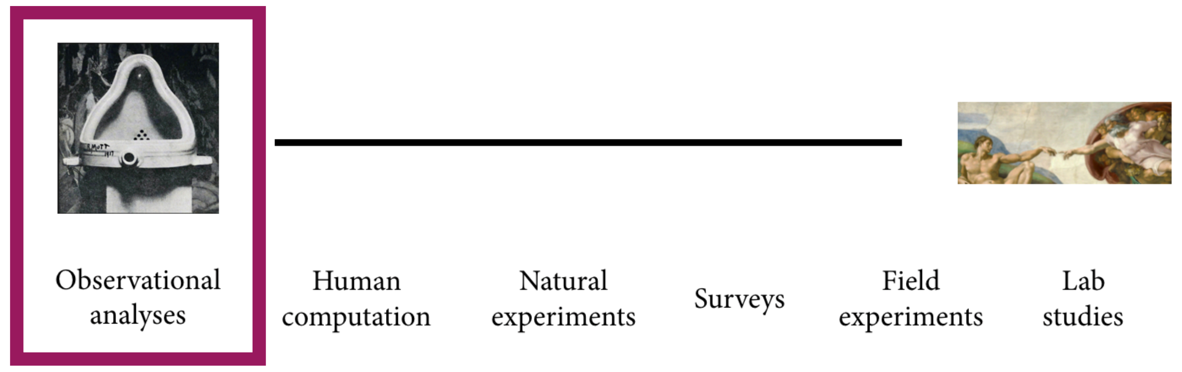
Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.



Observing Behaviour: 2. Nowcasting



Soon after Google Flu Trends launched, it was drastically off



Observing Behaviour: 2. Nowcasting

Media attention

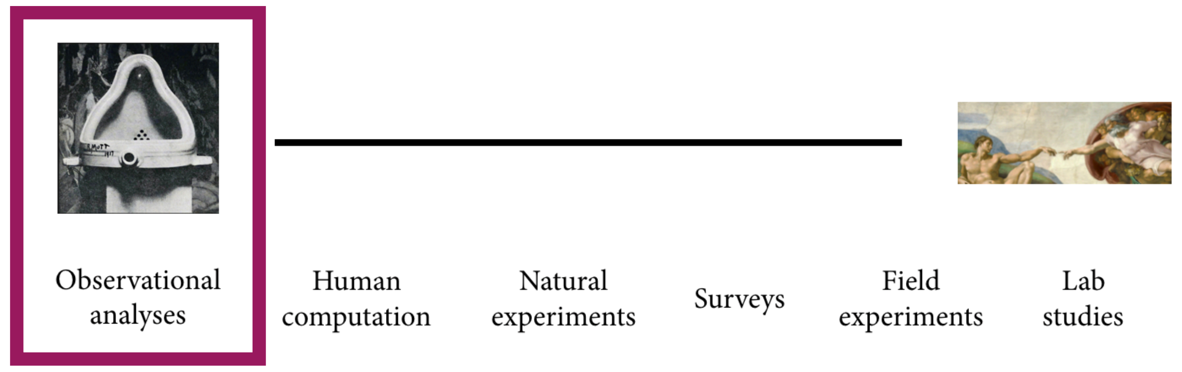
- “Bird flu”, “swine flu”

Algorithm changes

- Starting suggesting search terms

“Social hacking”

- Hey look we can screw up Google’s flu predictions

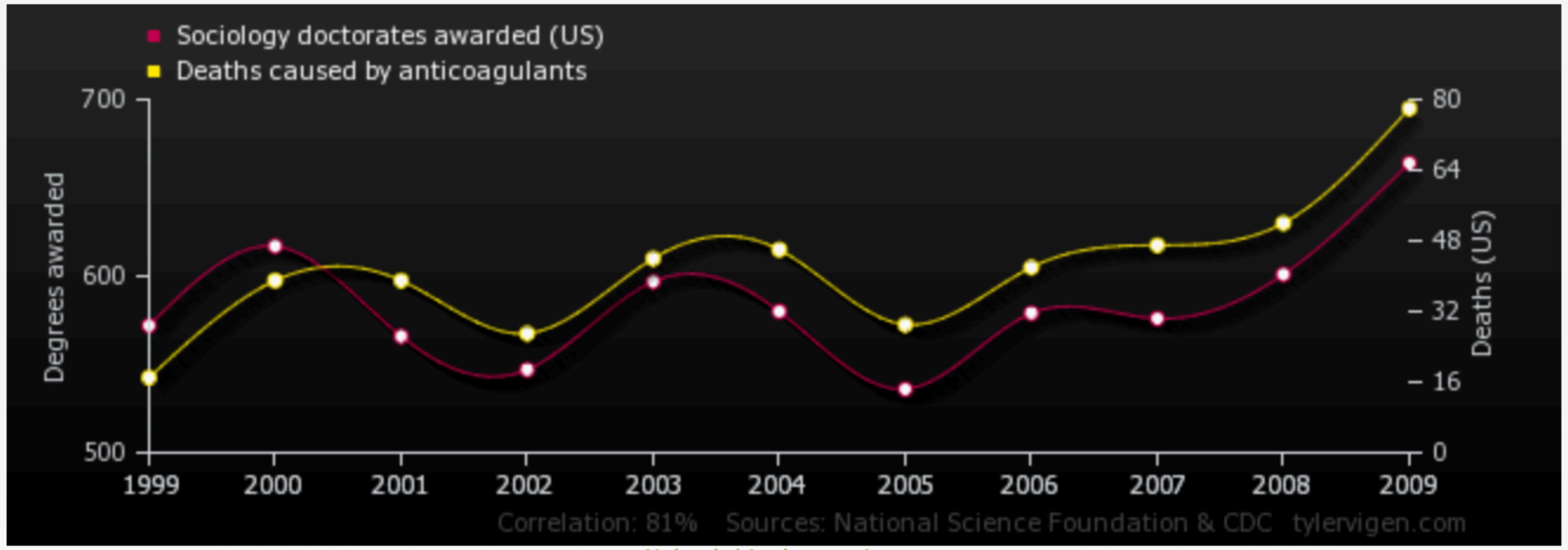


Correlation and causation

Sociology doctorates awarded (US)

correlates with

Deaths caused by anticoagulants



Upload this chart to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Sociology doctorates awarded (US) *Degrees awarded (National Science Foundation)*	572	617	566	547	597	580	536	579	576	601	664
Deaths caused by anticoagulants *Deaths (US) (CDC)*	17	39	39	27	44	46	29	42	47	52	78

Correlation: 0.811086

Observational analyses

Human computation

Natural experiments

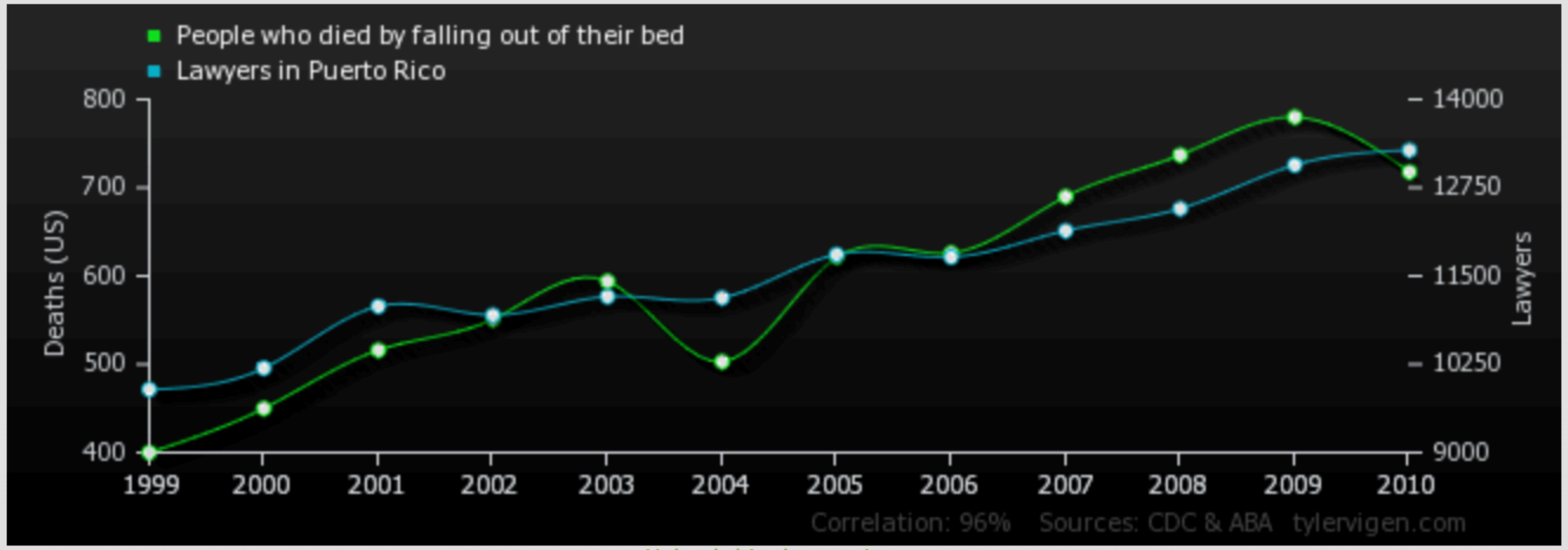
Surveys

Field experiments

Lab studies

Correlation and causation

People who died by falling out of their bed
correlates with
Lawyers in Puerto Rico



Upload this chart to imgur

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
People who died by falling out of their bed *Deaths (US) (CDC)*	400	450	516	551	594	503	621	626	690	737	780	718
Lawyers in Puerto Rico *Lawyers (ABA)*	9,892	10,195	11,071	10,947	11,209	11,191	11,805	11,767	12,142	12,454	13,071	13,282

Correlation: 0.957087

Observational analyses

Human computation

Natural experiments

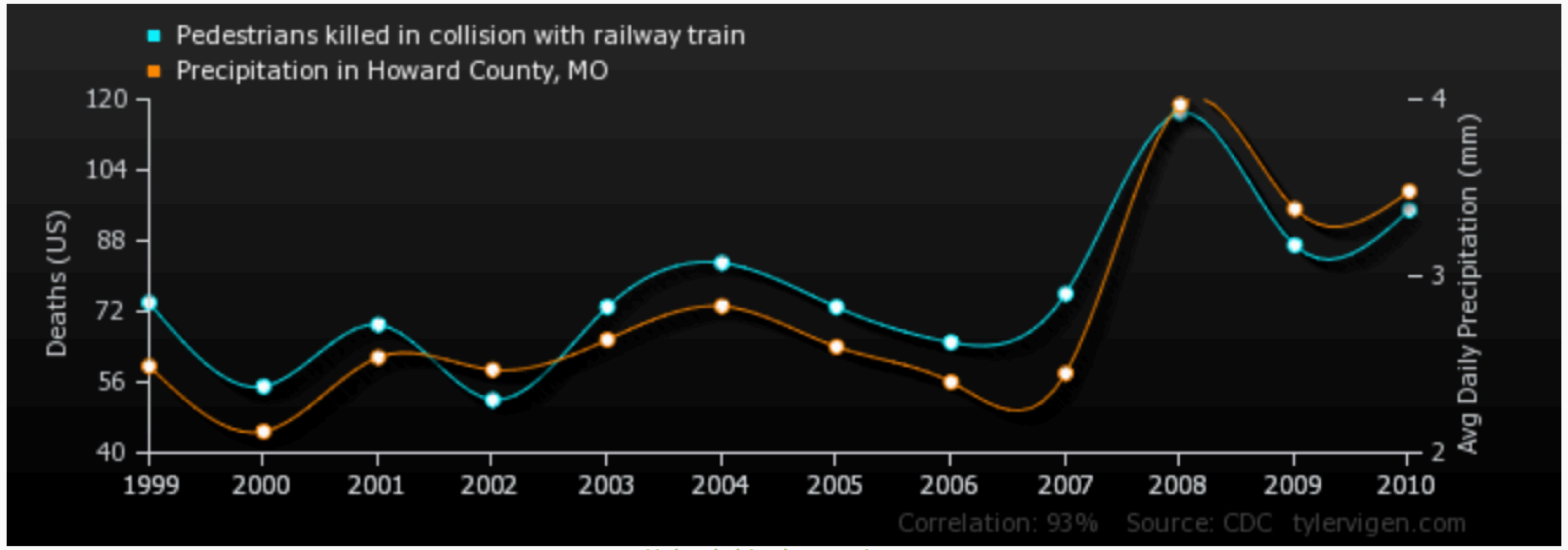
Surveys

Field experiments

Lab studies

Correlation and causation

Pedestrians killed in collision with railway train
correlates with
Precipitation in Howard County, MO



Upload this chart to imgur

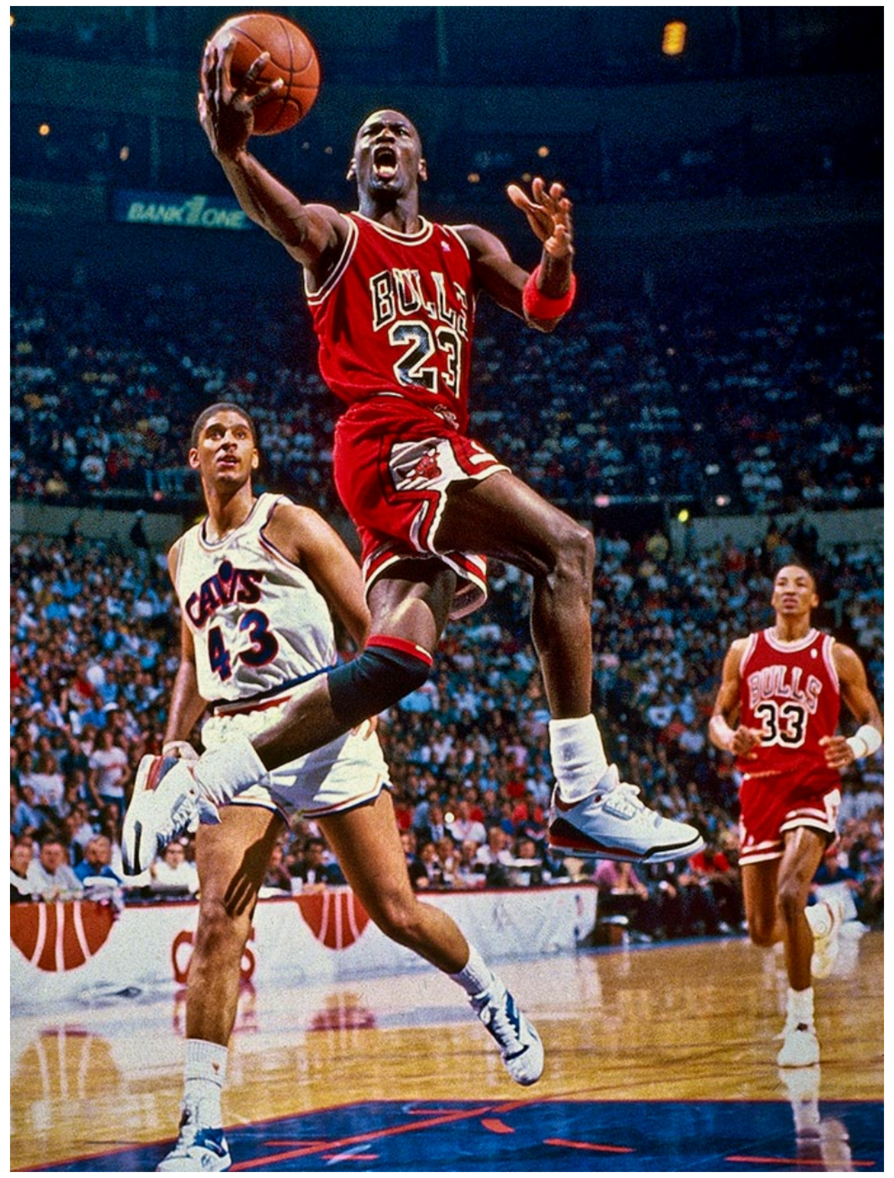
	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Pedestrians killed in collision with railway train Deaths (US) (CDC)	74	55	69	52	73	83	73	65	76	117	87	95
Precipitation in Howard County, MO Avg Daily Precipitation (mm) (CDC)	2.49	2.12	2.54	2.47	2.64	2.83	2.6	2.4	2.45	3.97	3.38	3.48

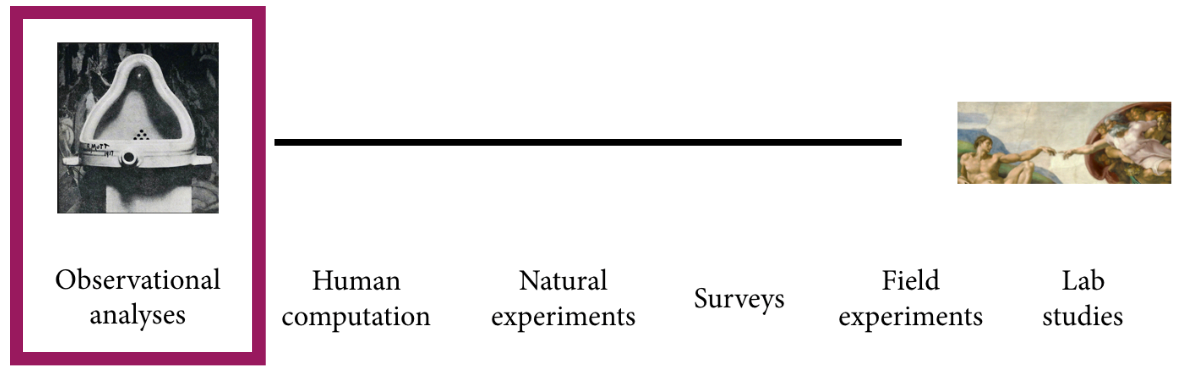
Correlation: 0.92783

Observational analyses
Human computation
Natural experiments
Surveys
Field experiments
Lab studies

Perils of big data

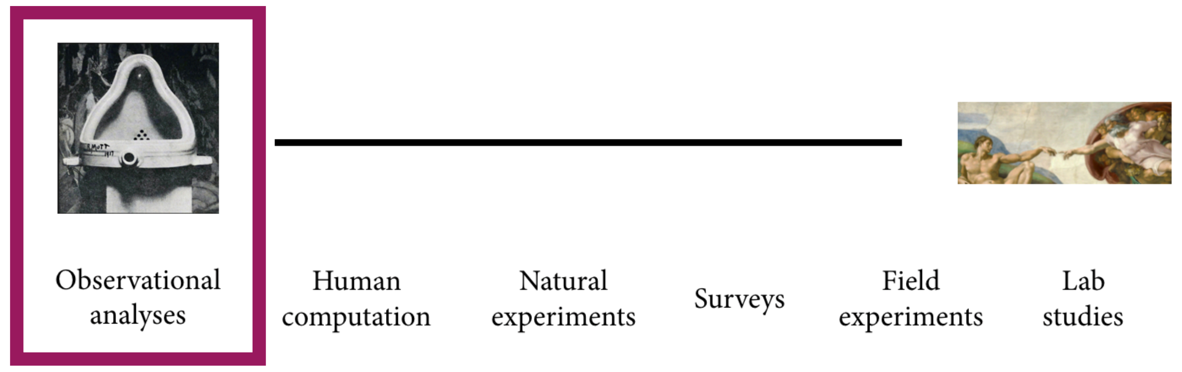
“When you have large amounts of data, your appetite for hypotheses tends to get even larger. And if it’s growing faster than the statistical strength of the data, then many of your inferences are likely to be false. They are likely to be white noise.” — Michael Jordan





Perils of big data

“When you have large amounts of data, your appetite for hypotheses tends to get even larger. And if it’s growing faster than the statistical strength of the data, then many of your inferences are likely to be false. They are likely to be white noise.” — Michael Jordan



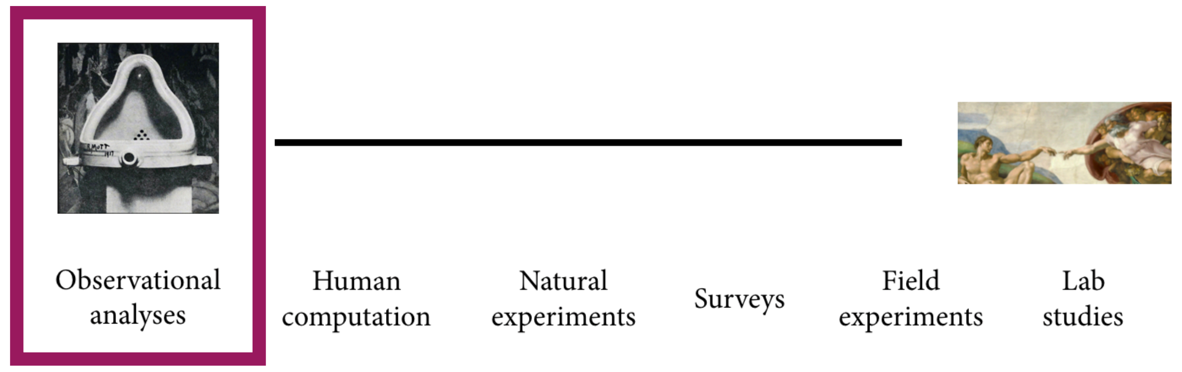
Observing Behaviour: 3. Approximating Experiments

Some clever strategies allow us to do "causal inference": make causal claims from observational data (i.e. arrive at experiment-like conclusions without actually running an experiment)

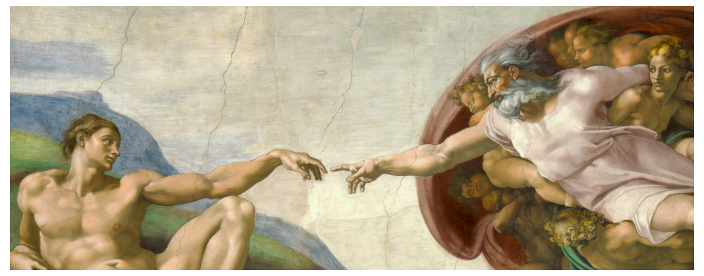
One well-known technique is instrumental variables: exploit natural variation in something to make a causal claim

Rain → Exercise

Friends exercising → You exercise?



Ways of doing computational social science



Observational analyses — Human computation — Natural experiments — Surveys — Field experiments — Experiments

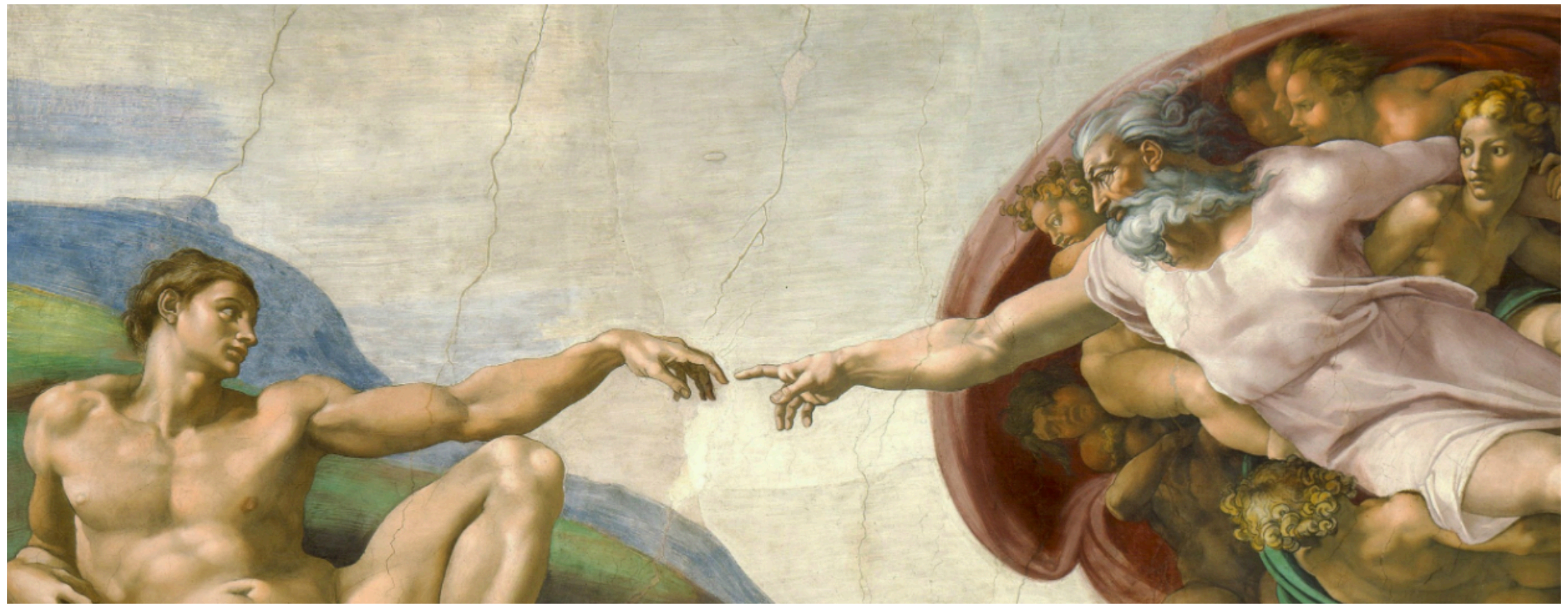
Experiments

On the other end of the spectrum is experimentation

The goal is to learn about causal relationships (cause-and-effect questions)

The strategy is to directly manipulate the environment and observe the consequences

Design the ideal scenario that will create just the data you need to answer your question



Experiments

Here, researchers intervene in the world to isolate and study a specific question

<u>Nomenclature:</u>

  “Experiment”: perturb and observe

  “Randomized controlled experiment”: Intervene for one group, don’t for another (randomly)

Correlation is not causation

  Observational data often riddled by unknown or hard-to-control confounding variables

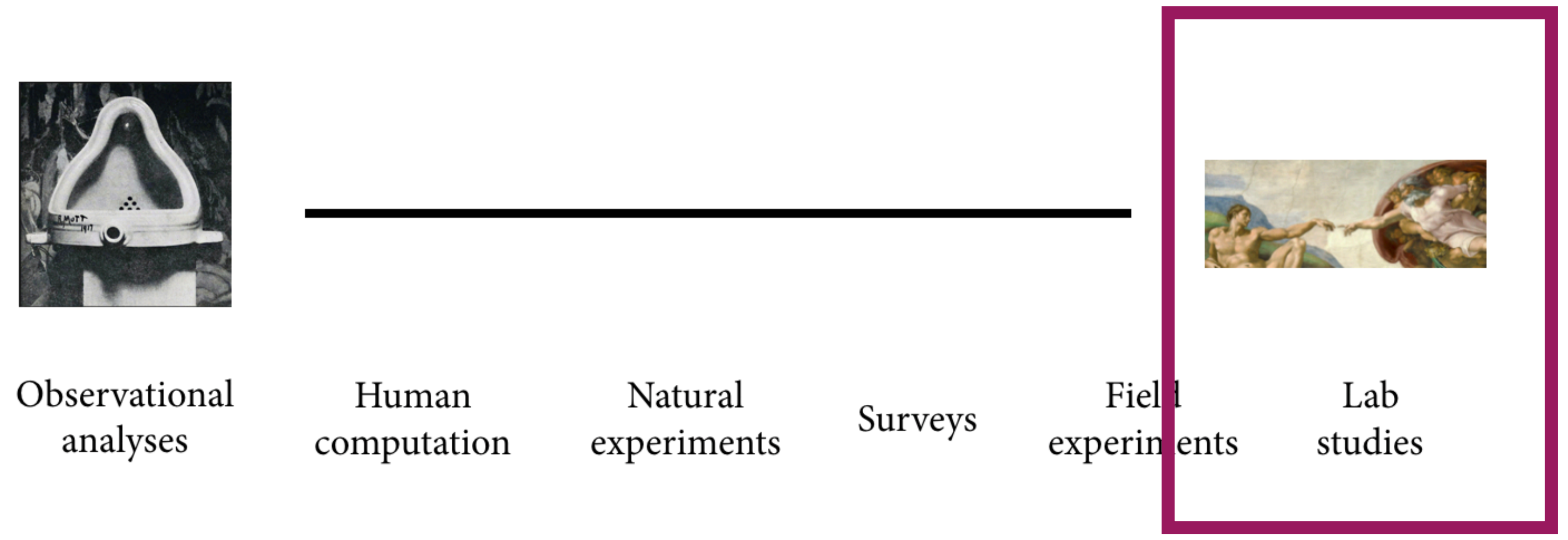
E.g. Do students learn more in schools that offer high teacher salaries?

  What’s an observational way to study this question?

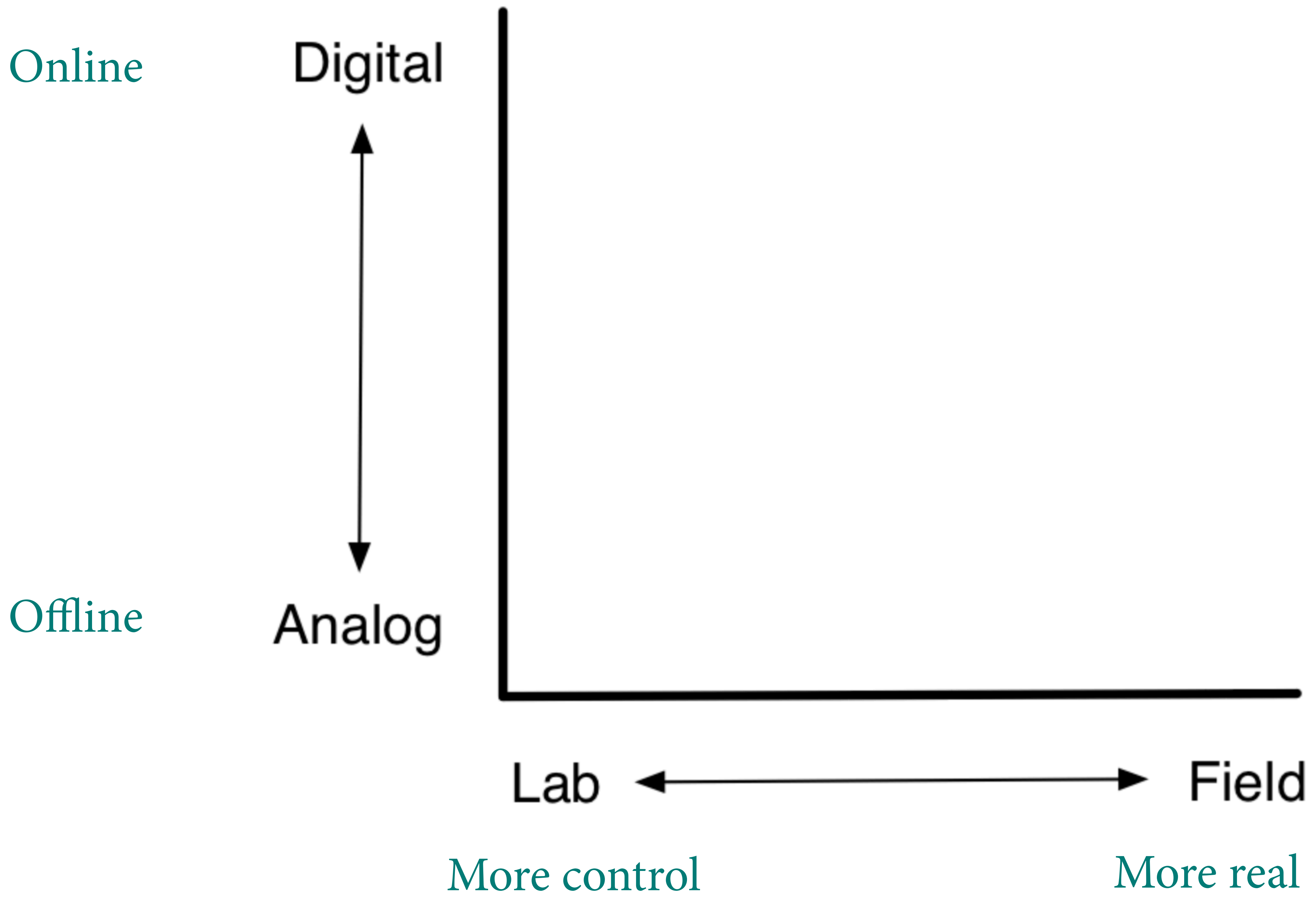
  What’s wrong with it?

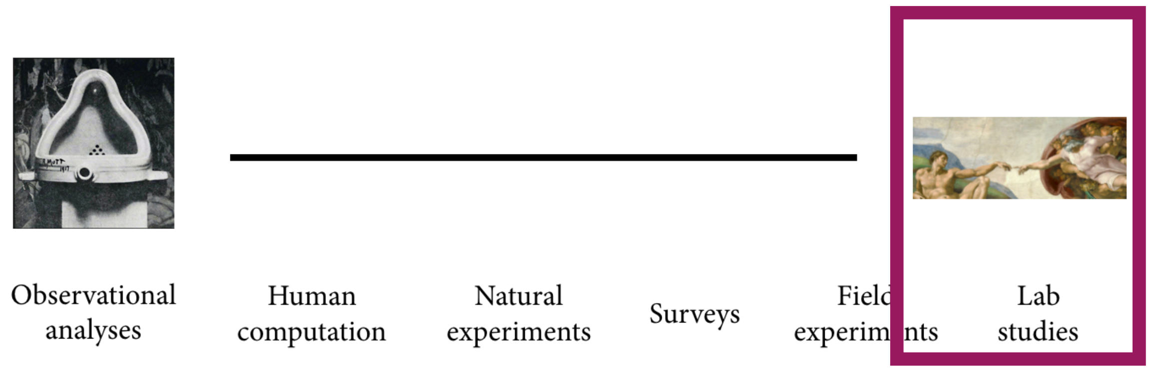
  What’s an experimental way to study this question?

  What’s wrong with it?

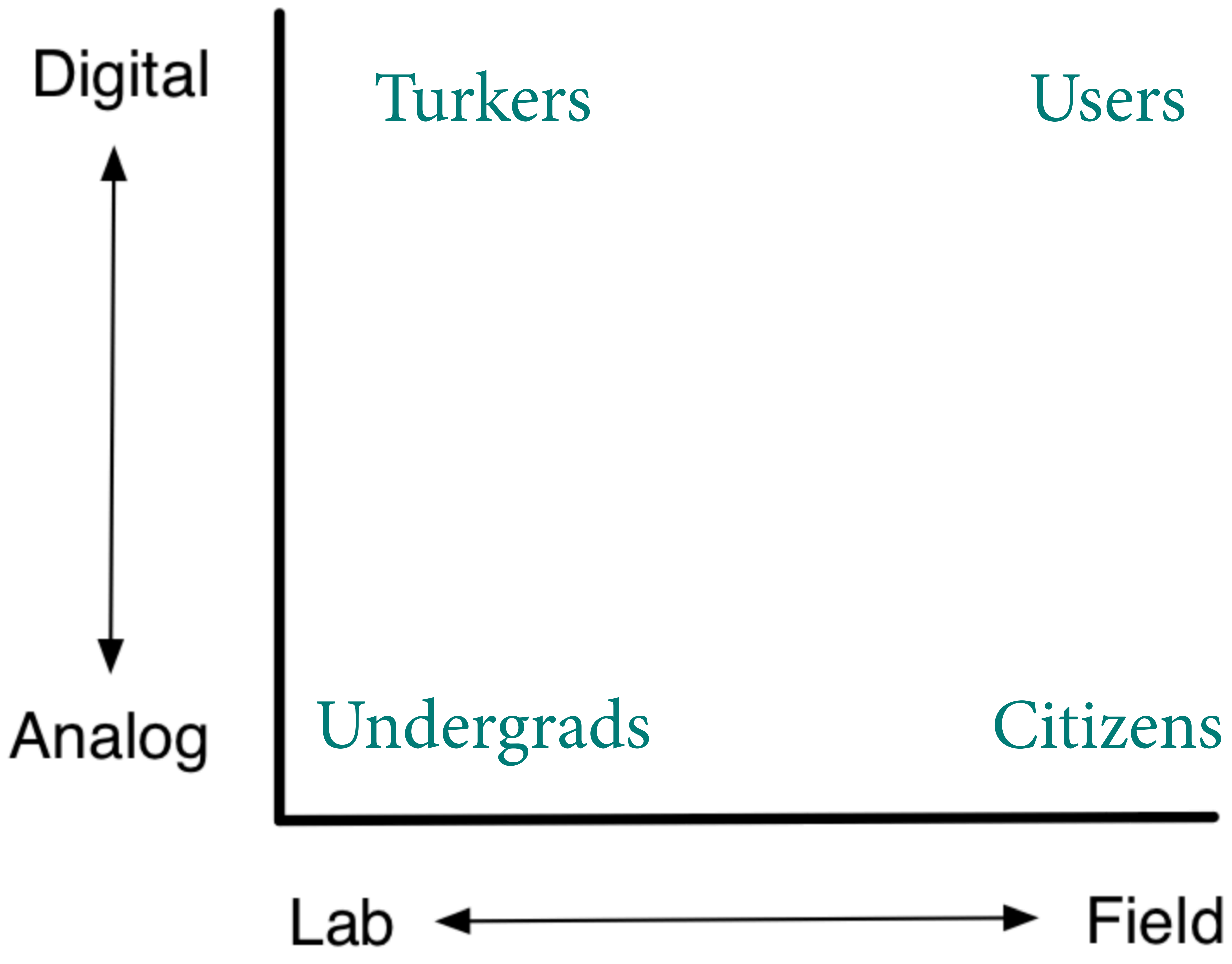


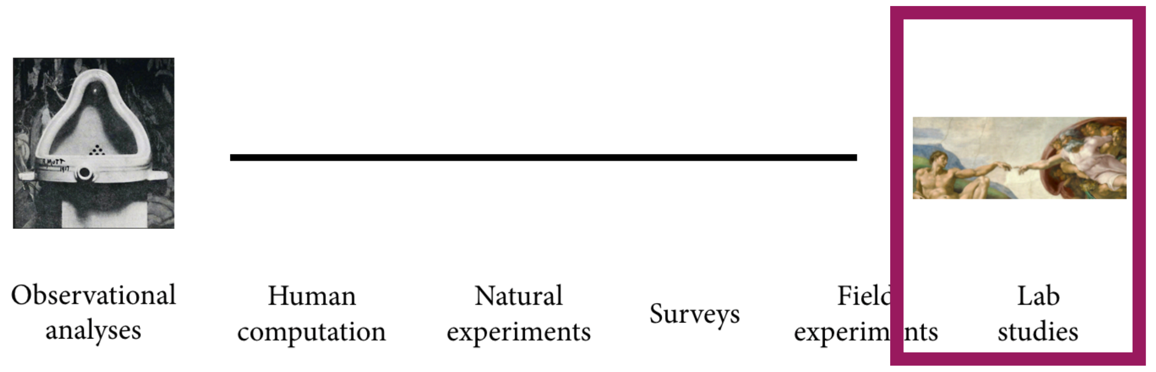
Experiments





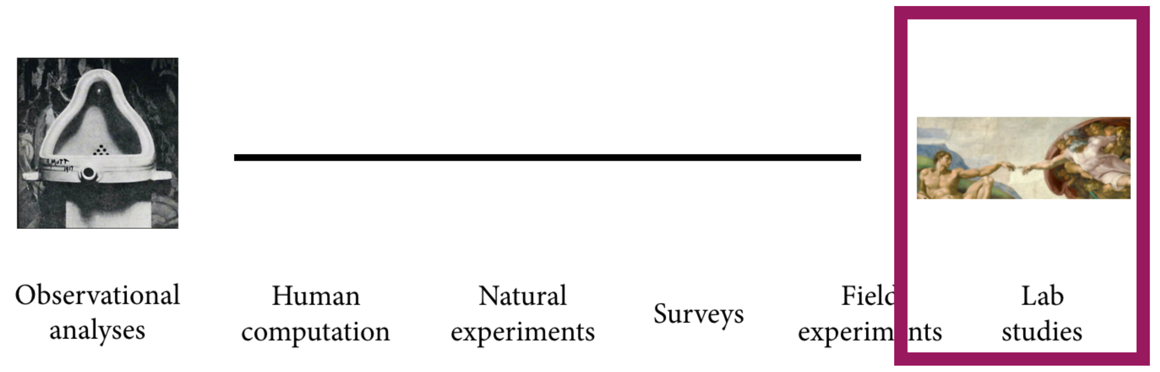
Experiments





Three major components of rich experiments

1. Validity
2. Heterogeneity
3. Mechanisms

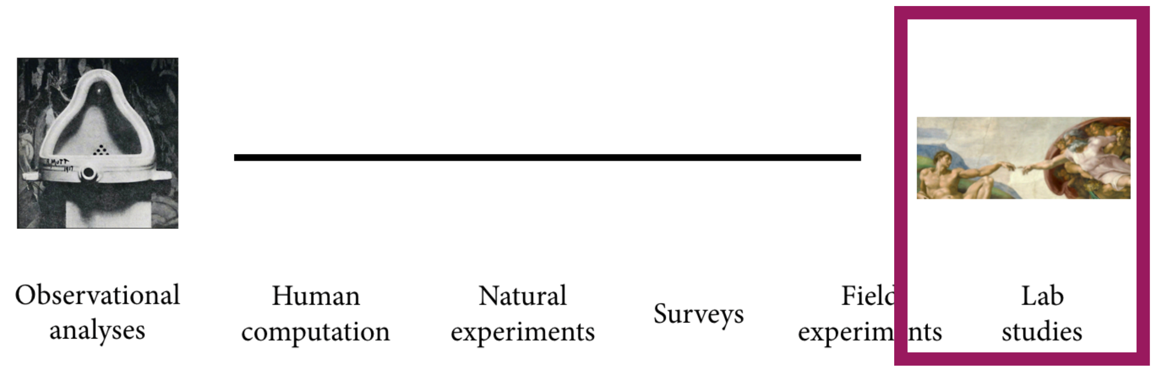


Three major components of rich experiments: 1. Validity

Validity: How general are the results?

Types of validity:

1. Statistical conclusion validity: were the stats done right?
2. Internal validity: was the experiment done right?
3. Construct validity: are we measuring the right thing?
4. External validity: is this applicable in other settings?

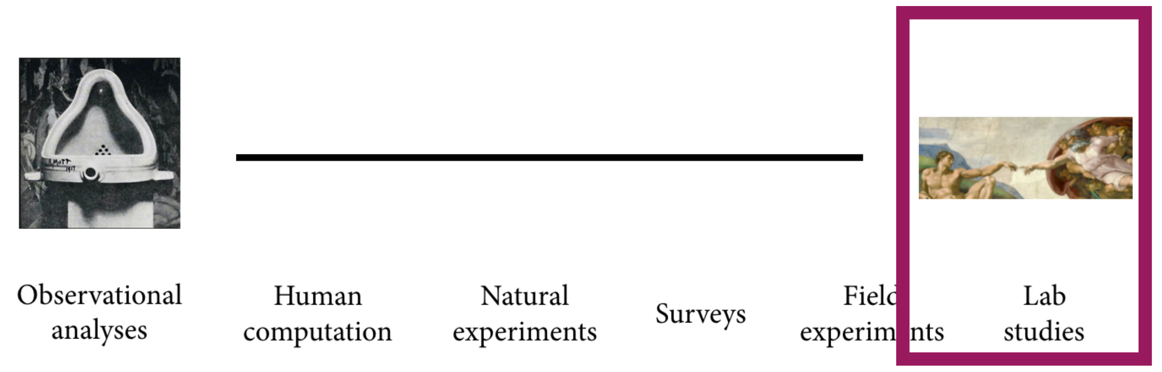


Three major components of rich experiments: 2. Heterogeneity

Barebones experiment: measure the average treatment effect (ATE)

But in social research, people almost always vary.

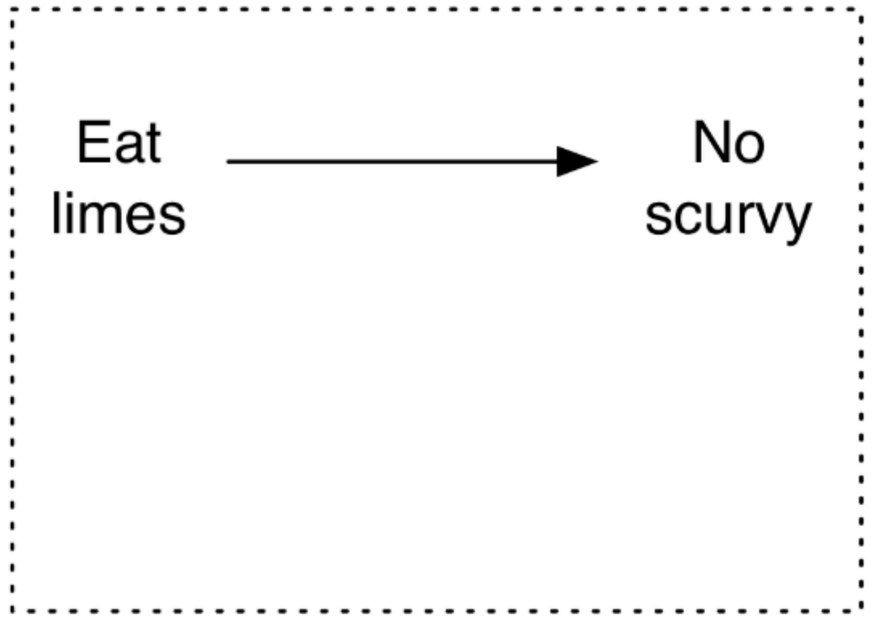
Digital research presents many more opportunities to measure how causes affect people differently



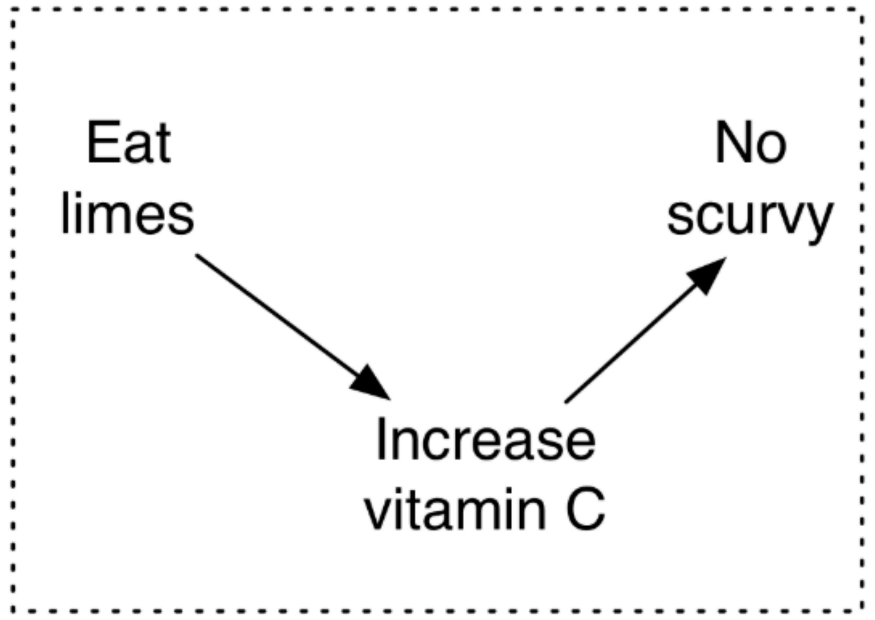
Three major components of rich experiments: 3. Mechanisms

Barebones experiment: measure what happened.

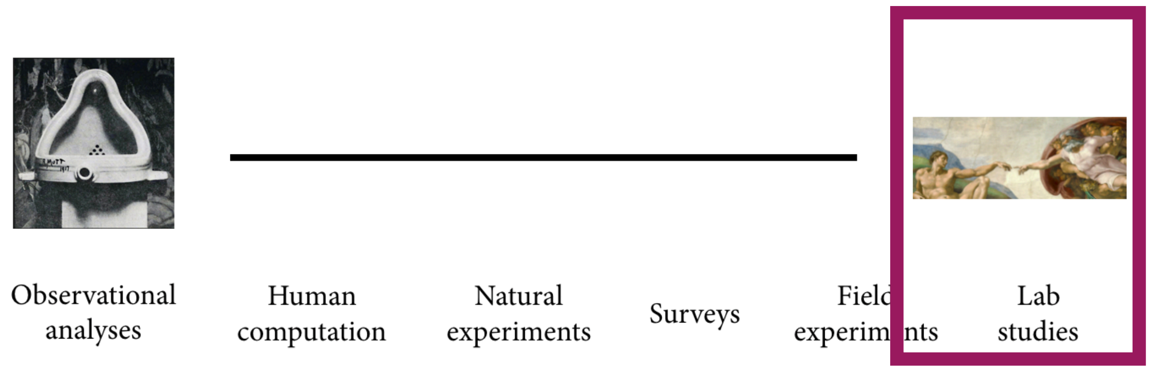
Mechanisms: why and how did it happen?



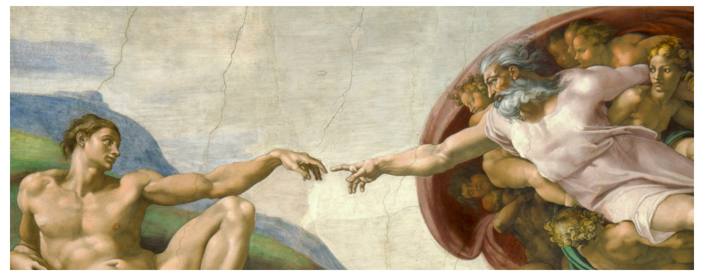
Causal effect
without mechanism



Causal effect
with mechanism



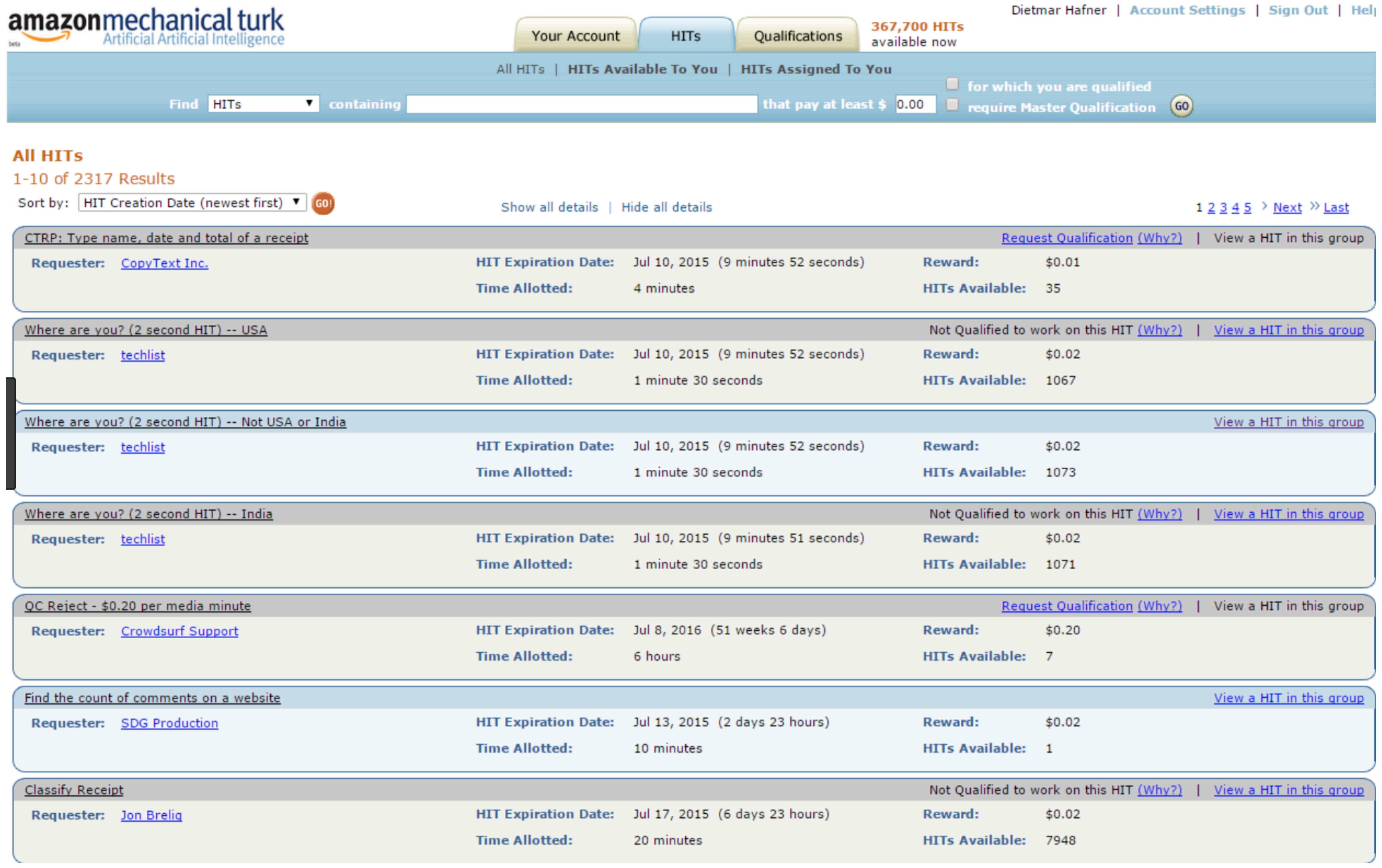
Ways of doing computational social science



Observational analyses | Human computation | Natural experiments | Surveys | Field experiments | Experiments

Human computation

- Online crowdsourcing platforms allow dividing work into microtasks
- Human-in-the-loop computing, modern-day lab studies, mass collaboration to build big resources (Wikipedia etc.)



Ways of doing computational social science

Observational analyses

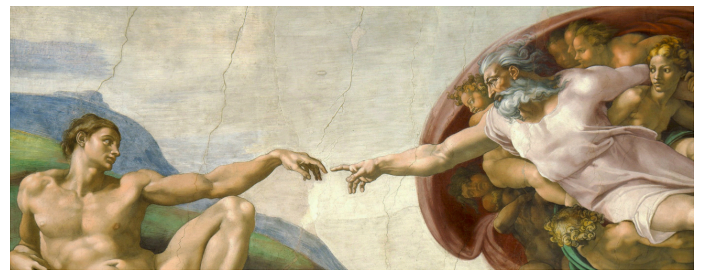
Human computation

Natural experiments

Surveys

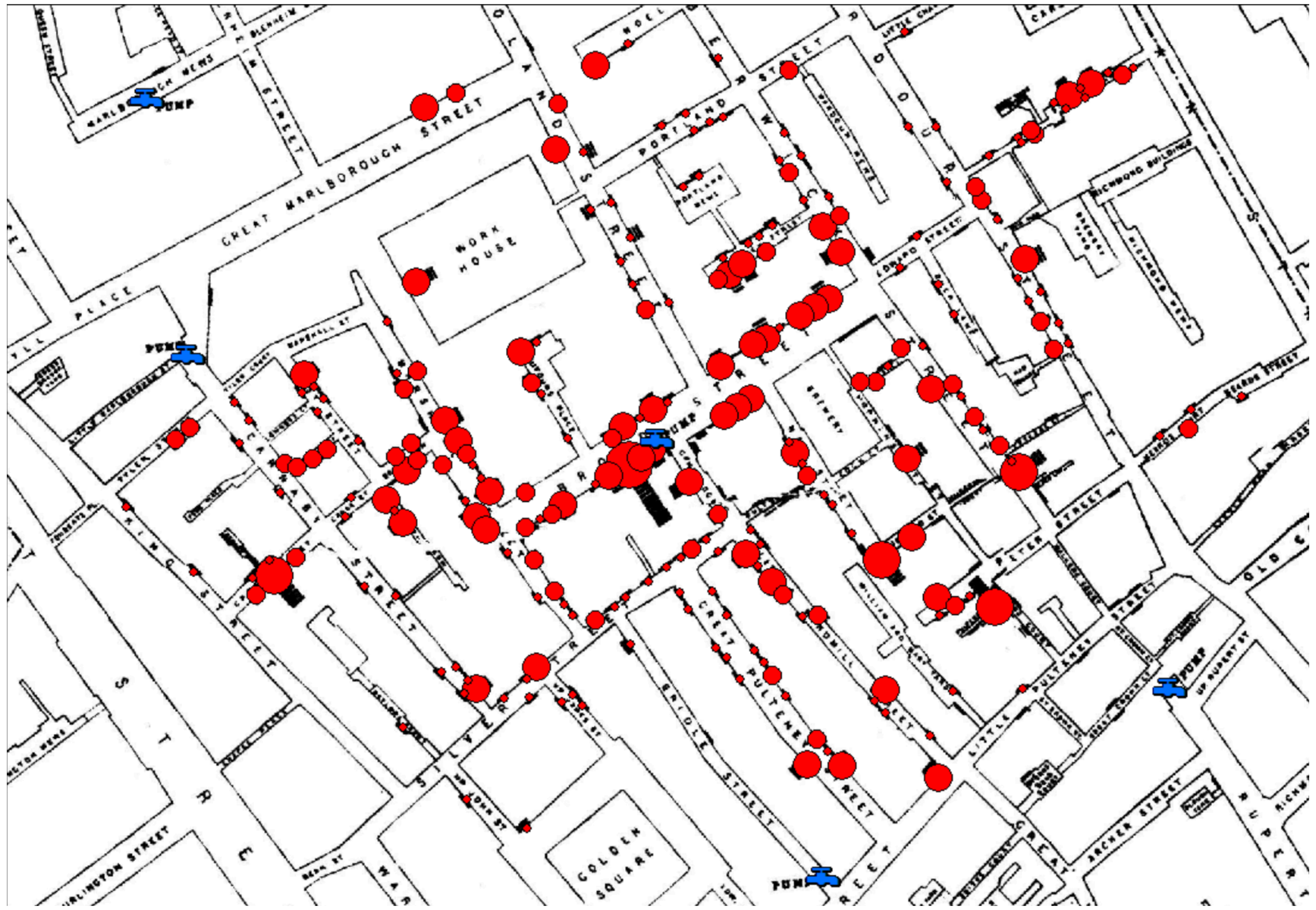
Field experiments

Experiments



Natural experiments

Sometimes observational data has some random component you can exploit, and analyze as a “natural” experiment

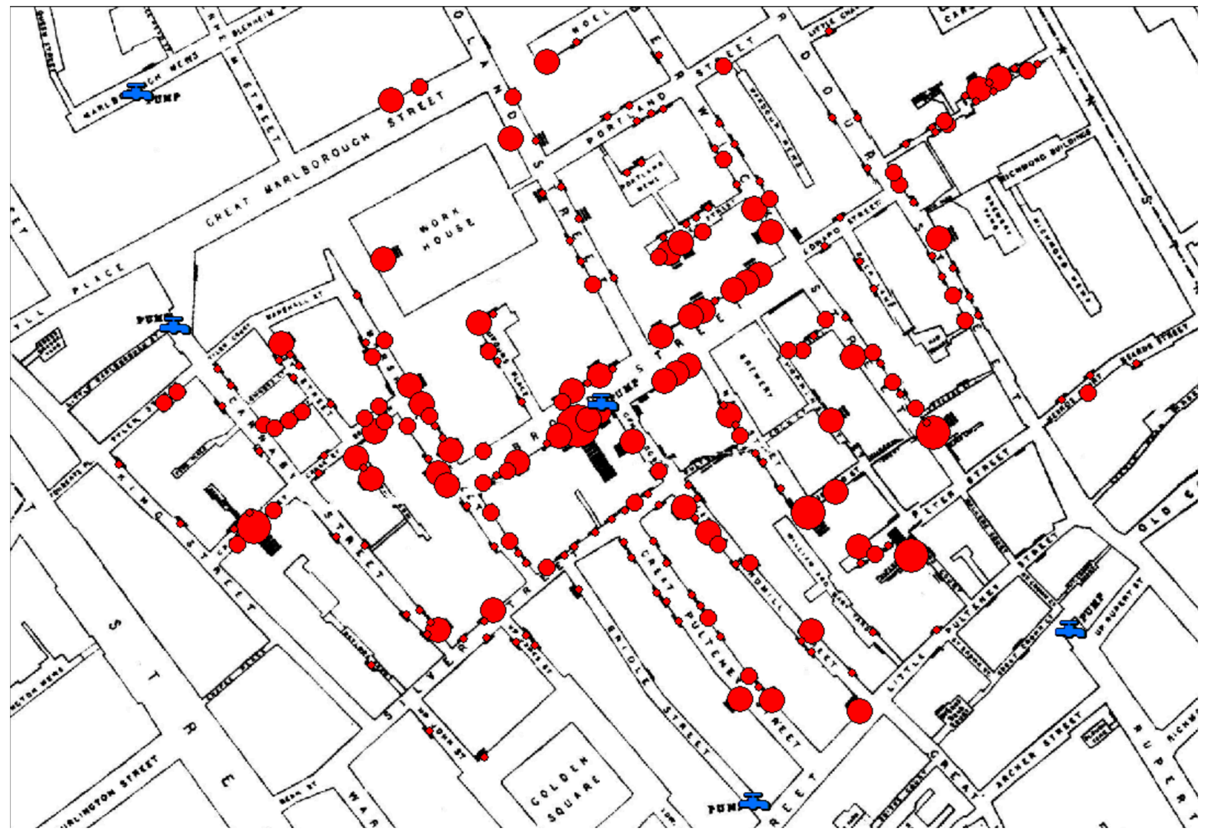


Cholera outbreak in London in 1850s

Natural experiments

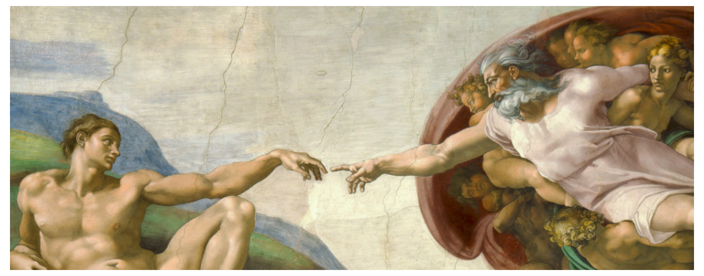
- Physician John Snow produced a map suggesting particular water was the culprit
- Two main water suppliers: one from downstream Thames where raw sewage was dumped in the water (high attack rates), and one from upstream (low attack rates)
- Which supplier you had was arbitrary (varied even within same house, same neighbourhood, etc.)
- Exposure to polluted water was as-if random

Now: in large datasets, more opportunities to identify and argue for as-if random assignment



Cholera outbreak in London in 1850s

Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

Surveys

Field experiments

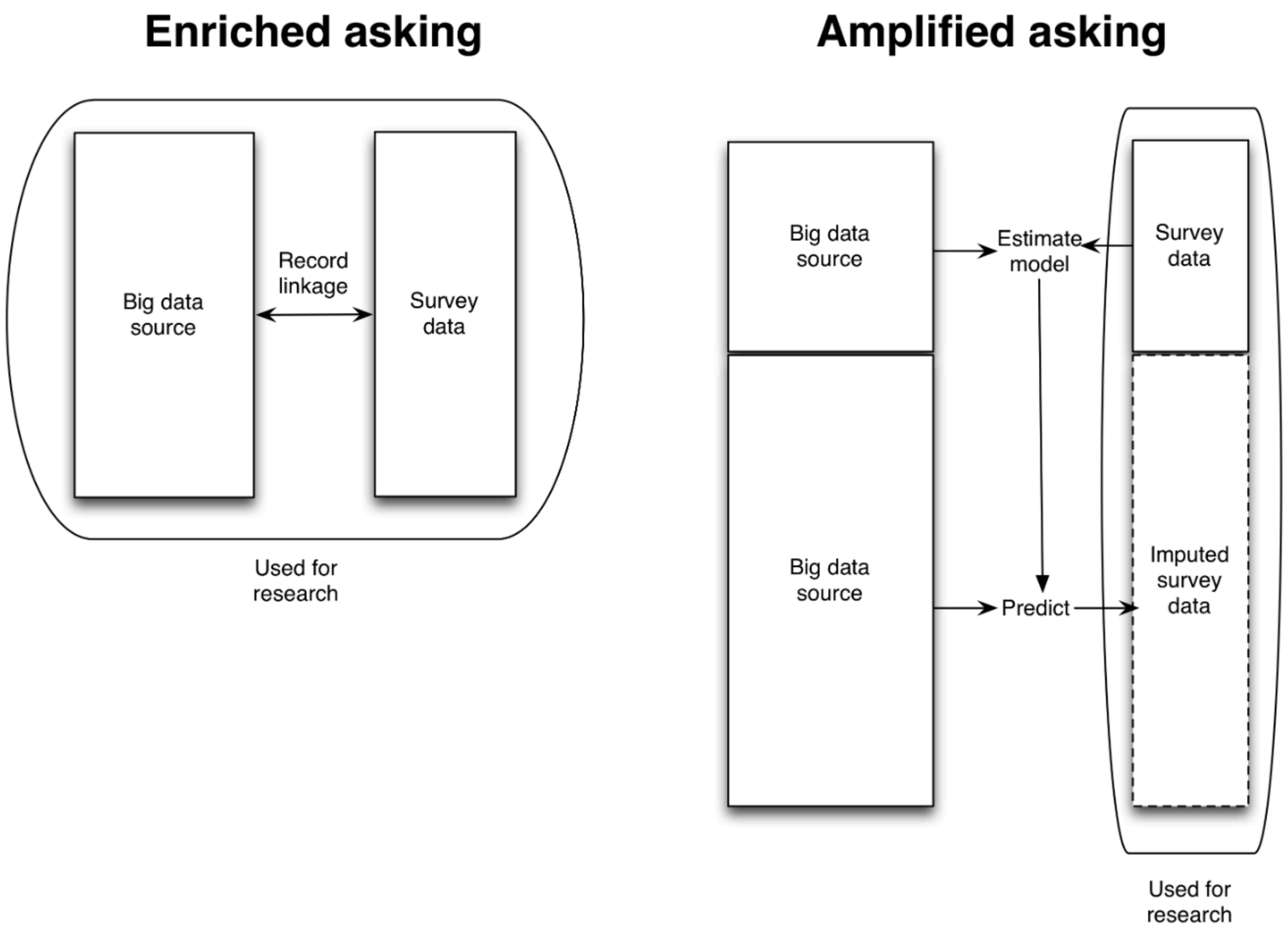
Experiments

Surveys: asking questions

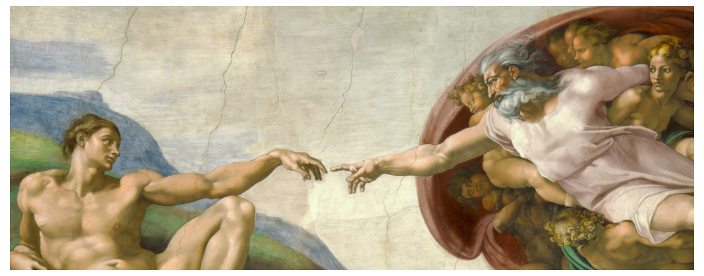
Social research has a unique advantage: we can ask our subjects what they're thinking!

Still the best way to learn the answer to many questions

In the digital era, there are new ways of asking questions



Ways of doing computational social science



Observational analyses

Human computation

Natural experiments

Surveys

Field experiments

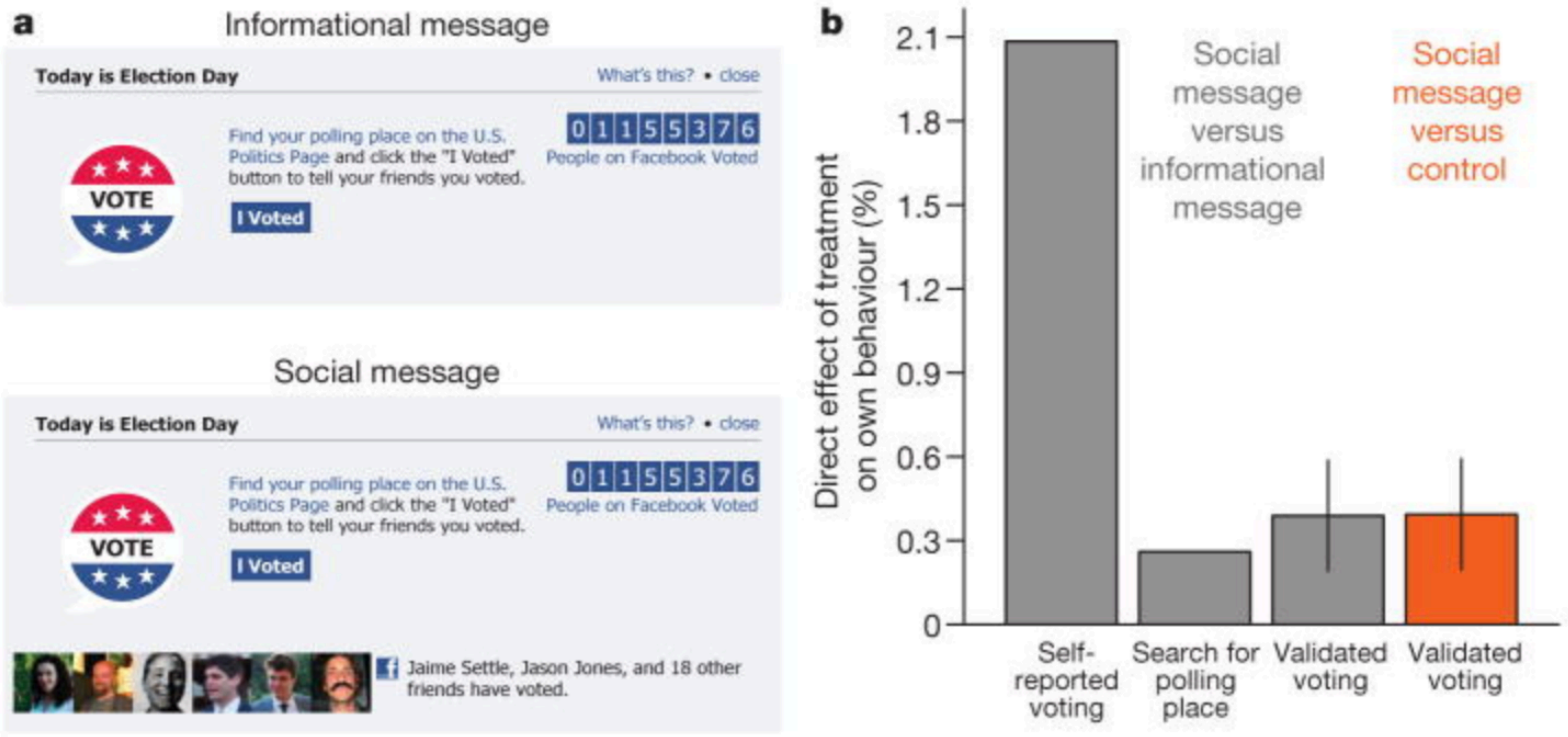
Experiments

Field experiments

- Introducing a treatment into a real system
- Much more possible now with algorithmic systems

Voting experiment on Facebook

Figure 1



The experiment and direct effects

a, **b**, Examples of the informational message and social message Facebook treatments (**a**) and their direct effect on voting behaviour (**b**). Vertical lines indicate s.e.m. (they are too small to be seen for the first two bars).

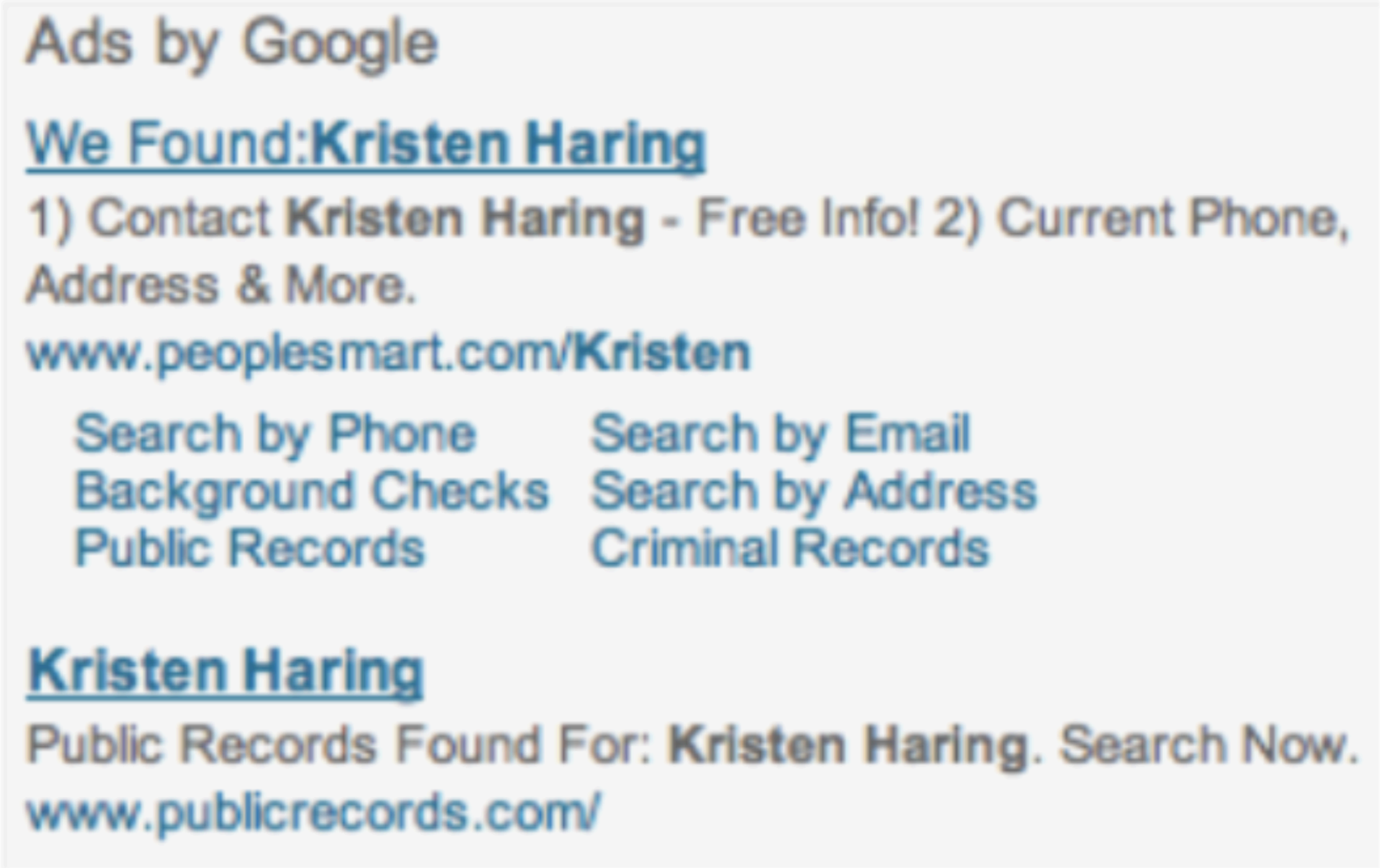
~300,000 more validated votes

AI & Society: Algorithmic decision-making

St. George's Hospital in the UK developed an algorithm to sort medical school applicants. Algorithm trained to mimic past admissions decisions made by humans.

But past decisions were biased against women and minorities. It codified discrimination.

Web search ads for “Kristen Haring”



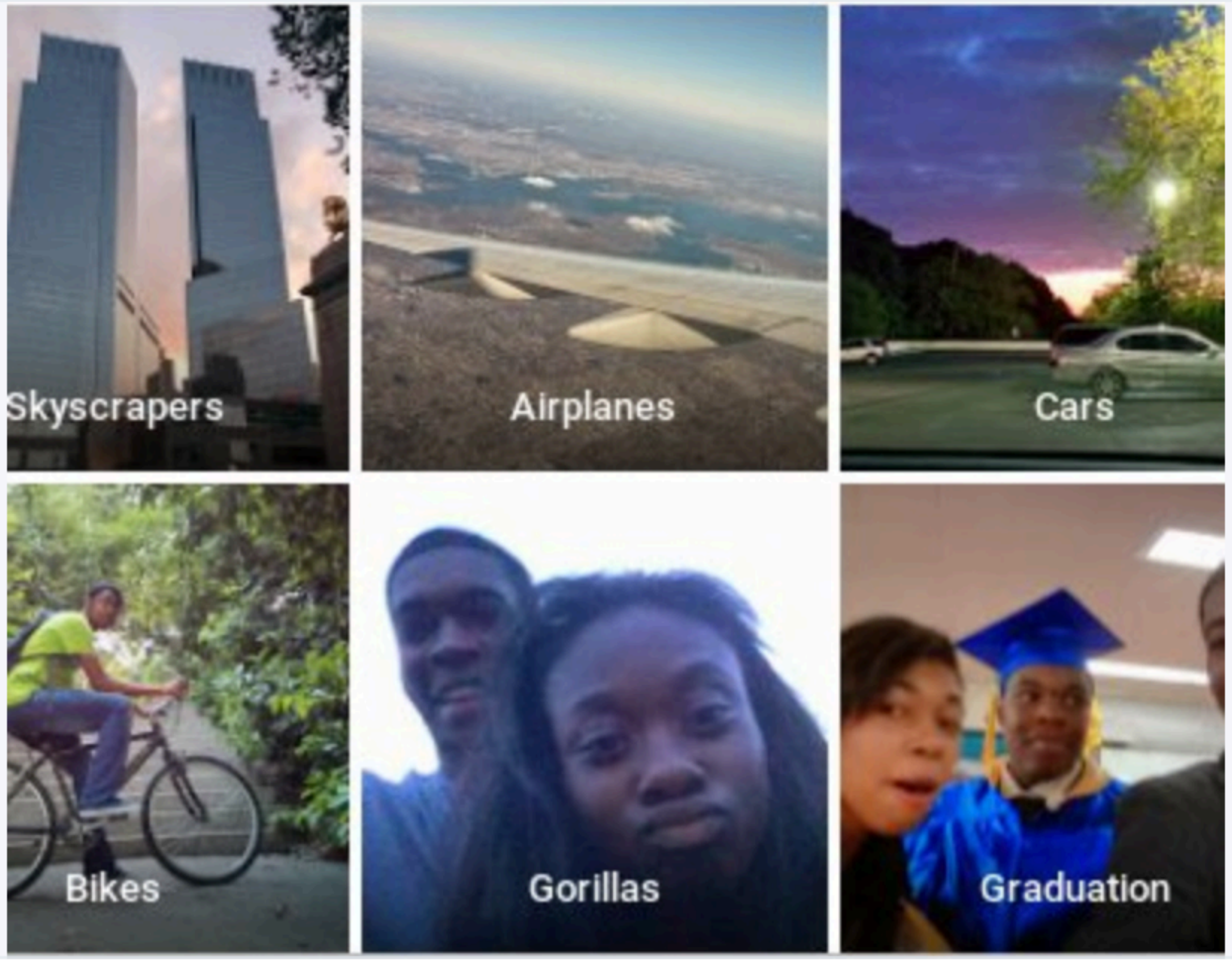
Web search ads for "Latanya Farrell"

Ads related to **latanya farrell**

Latanya Farrell, Arrested?
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

Latanya Farrell
www.publicrecords.com/
Public Records Found For: **Latanya Farrell**. View Now.

Image labeling gone wrong



Jacky Alciné
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.

8:22 PM - Jun 28, 2015

226 3,214 2,067

Image searching for "CEO"

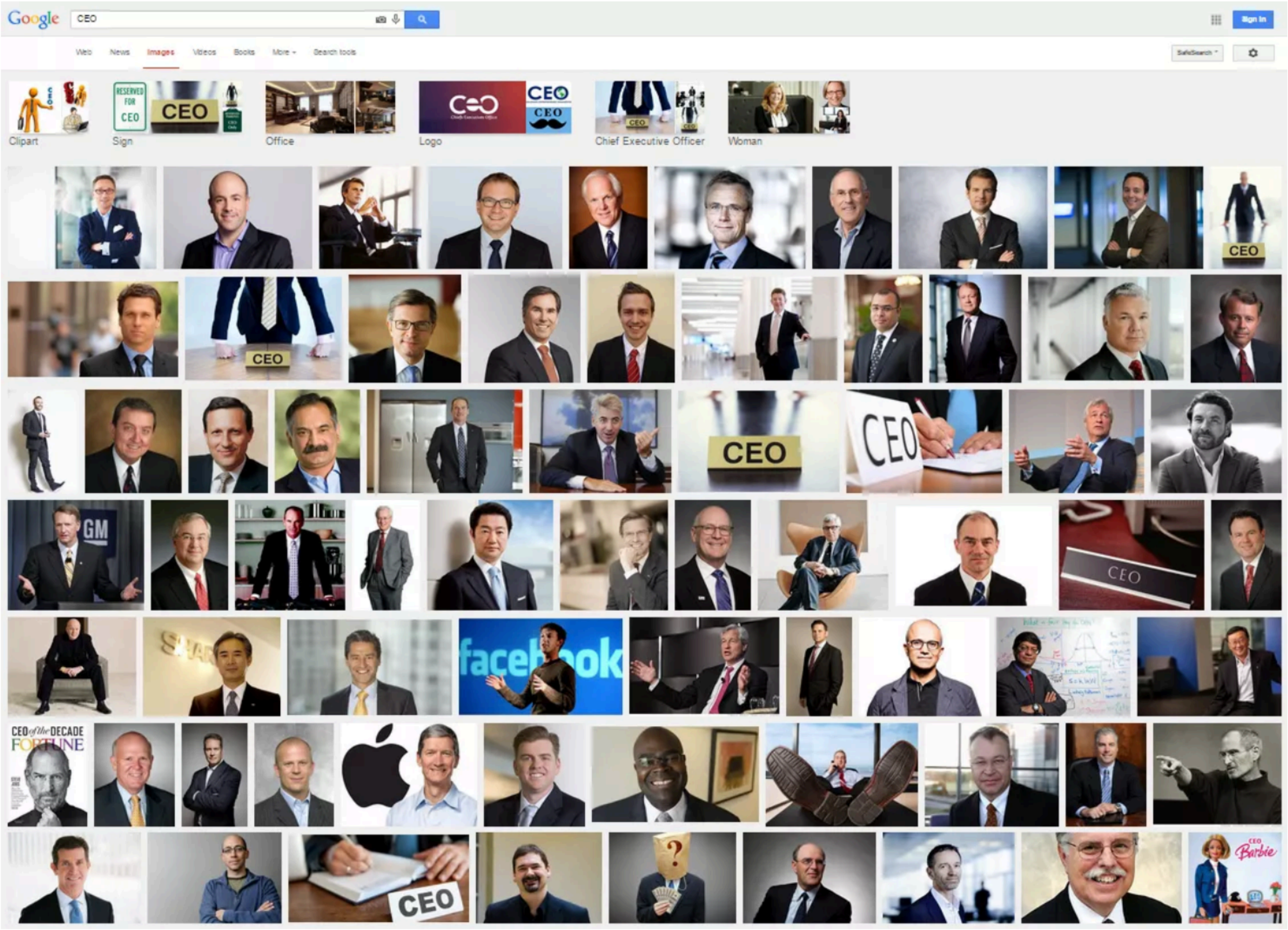


Image searching for “CEO”

Last nail in the coffin: this picture is from an Onion article.

Ethics and privacy

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

Facebook's Users Outraged Over Emotion Experiment

Facebook reveals news feed experiment to control emotions

Facebook emotion experiment sparks criticism

Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry

Facebook conducted secret psychology experiment on users' emotions

Everything We Know About Facebook's Secret Mood Manipulation Experiment

Computational social science

Game-changing opportunity to improve our understanding of human behaviour and have positive societal impact.

Doing so requires addressing serious technical, scientific, and ethical challenges.

Logistics

- http://www.cs.toronto.edu/~ashton/csc2552/
- Office hours by appointment
- Lectures Thursday 3–5pm
- Textbook: Bit by Bit by Matthew Salganik
- Read Chapter 1 (short)