CSC2552
Topics in Computational Social Science: AI, Data, and Society

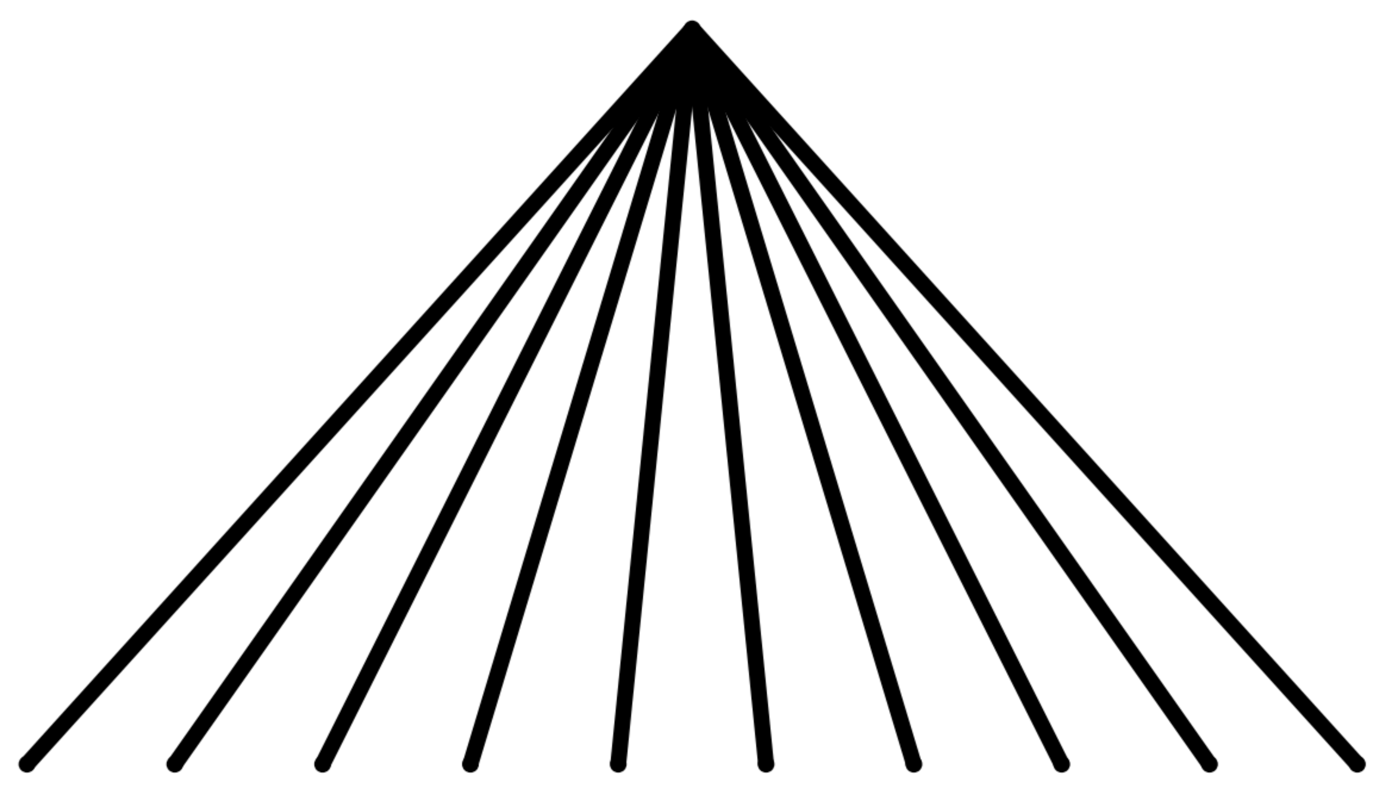
Lecture 1: Introduction to Computational Social Science

Prof. Ashton Anderson, Fall 2023

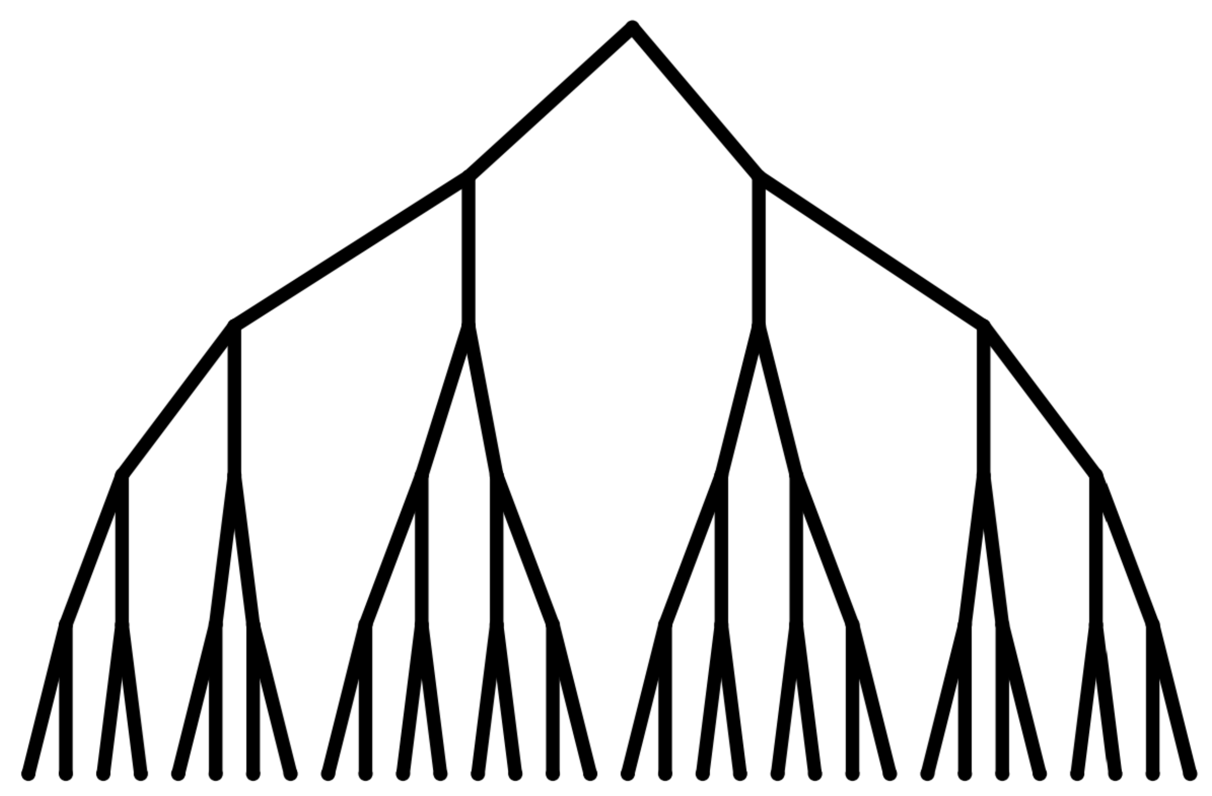
An example motivating question

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

Broadcast



Viral

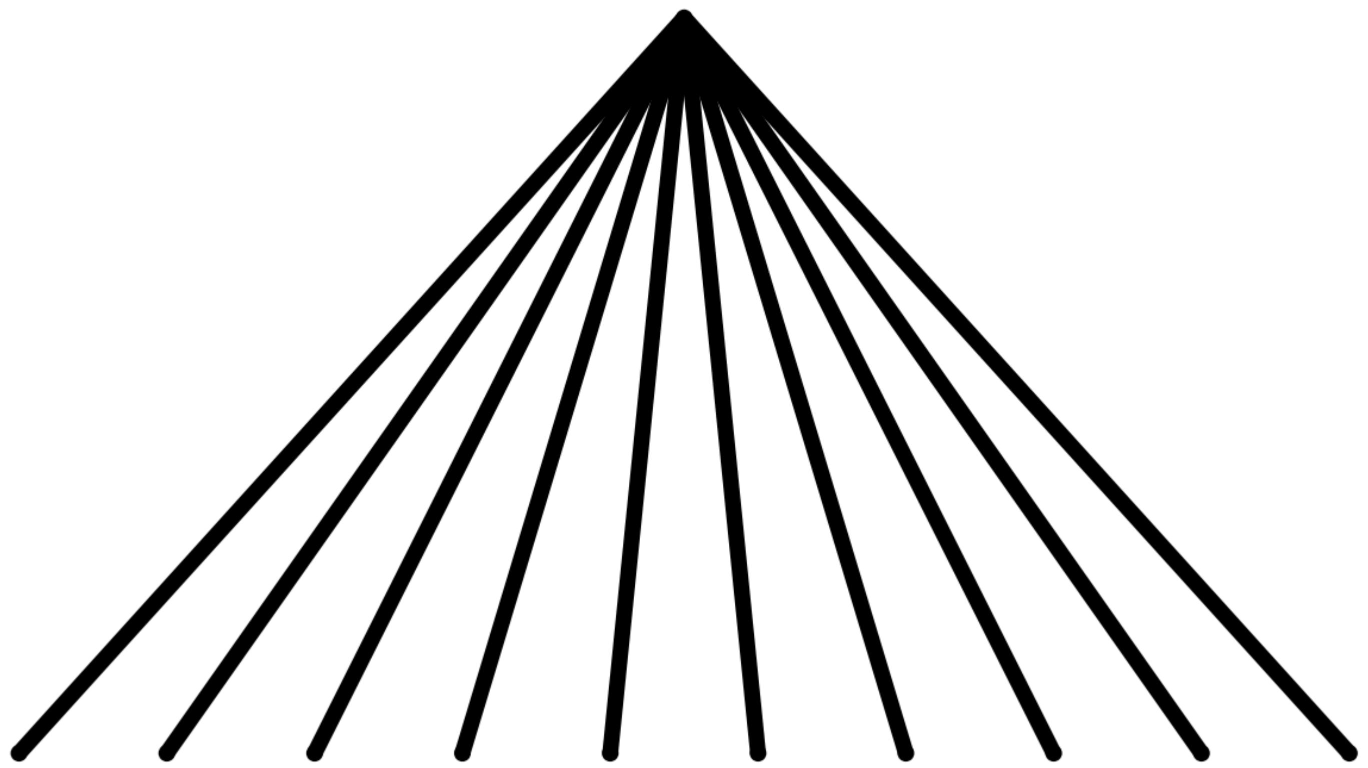


An example motivating question

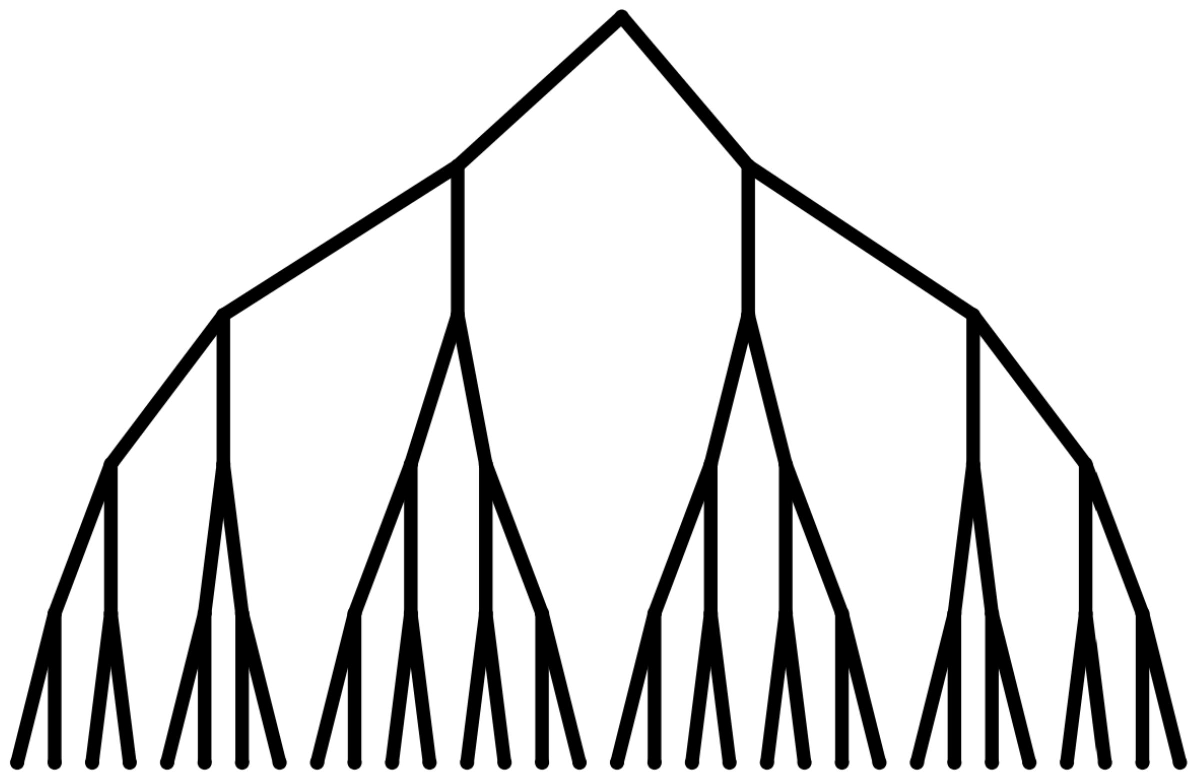
This is an important question:

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

Broadcast



Viral

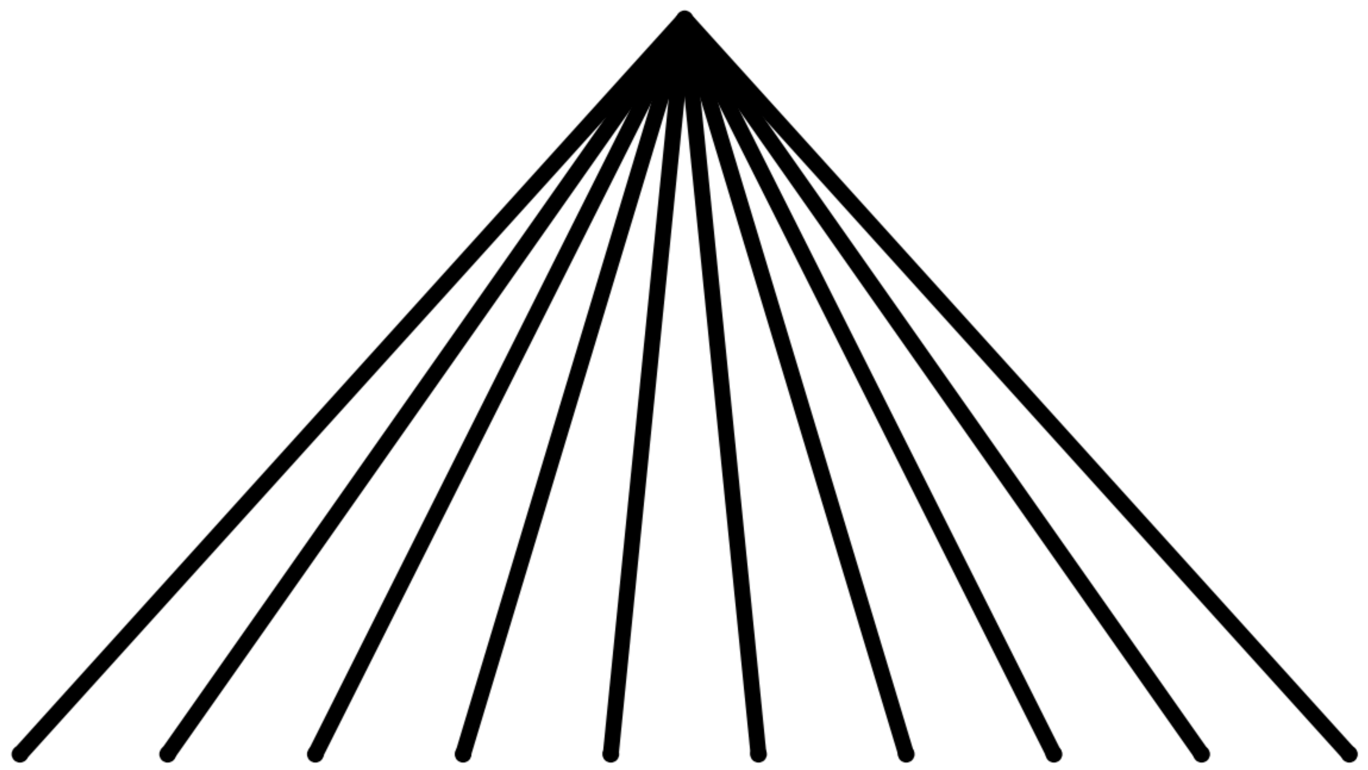


A motivating question

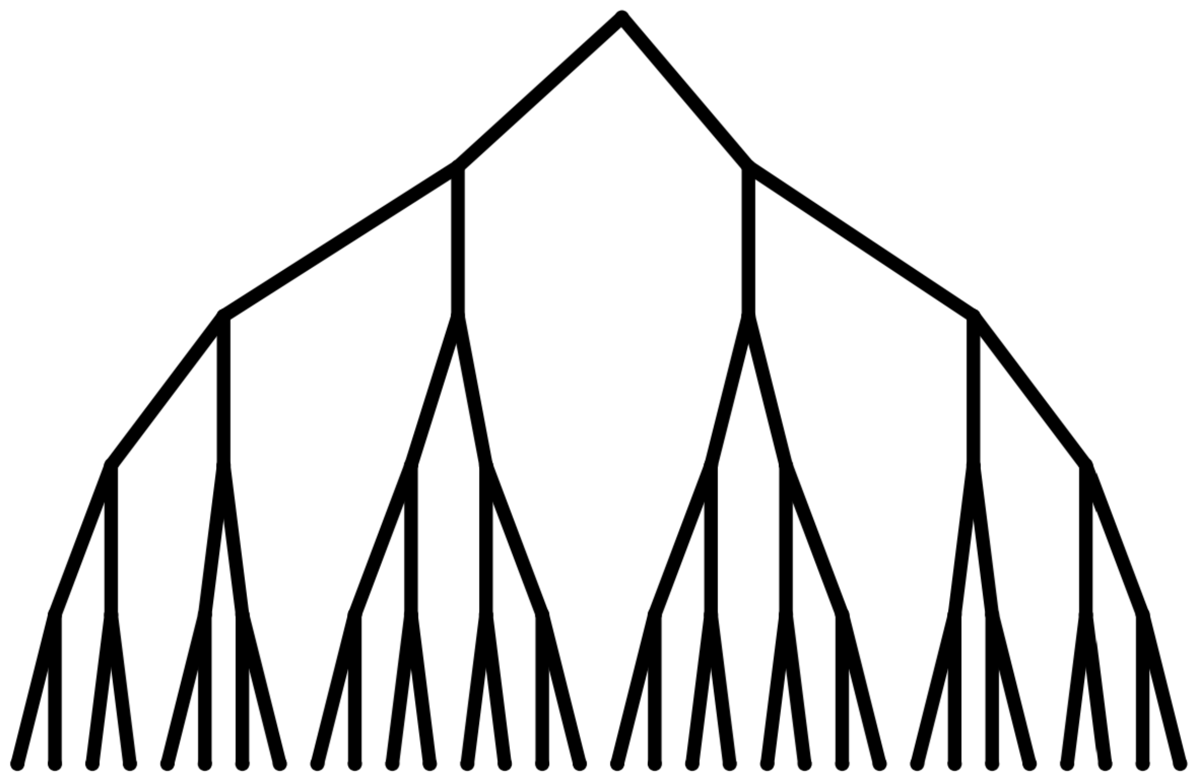
This is a difficult question:

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

Broadcast

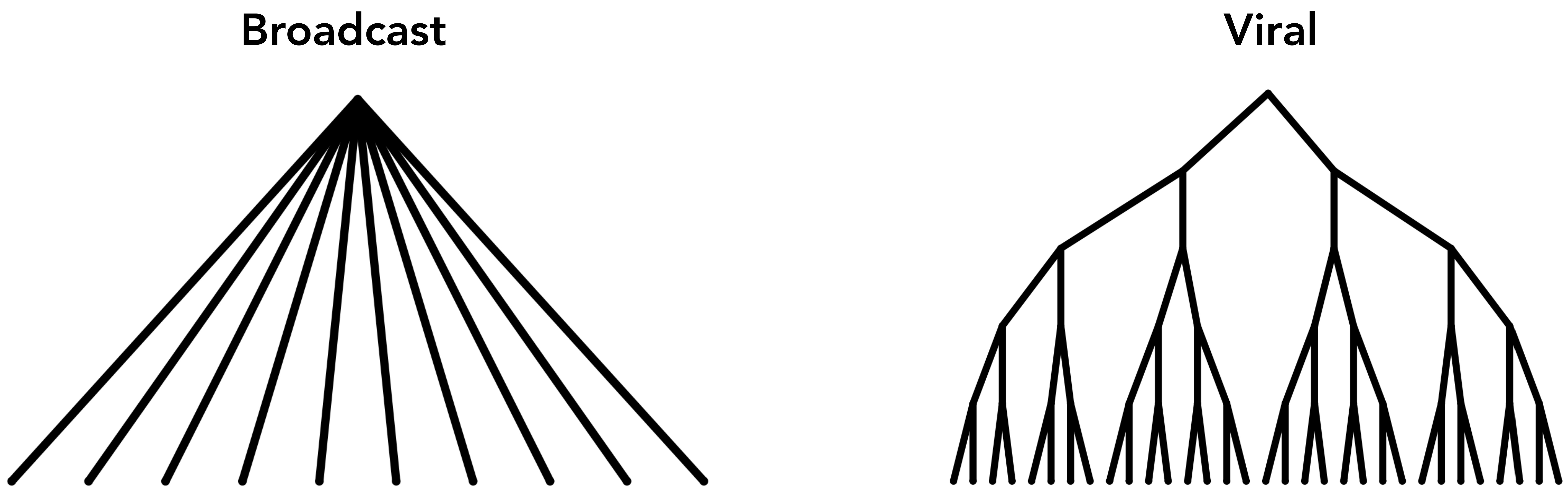


Viral



Traditional data & methods

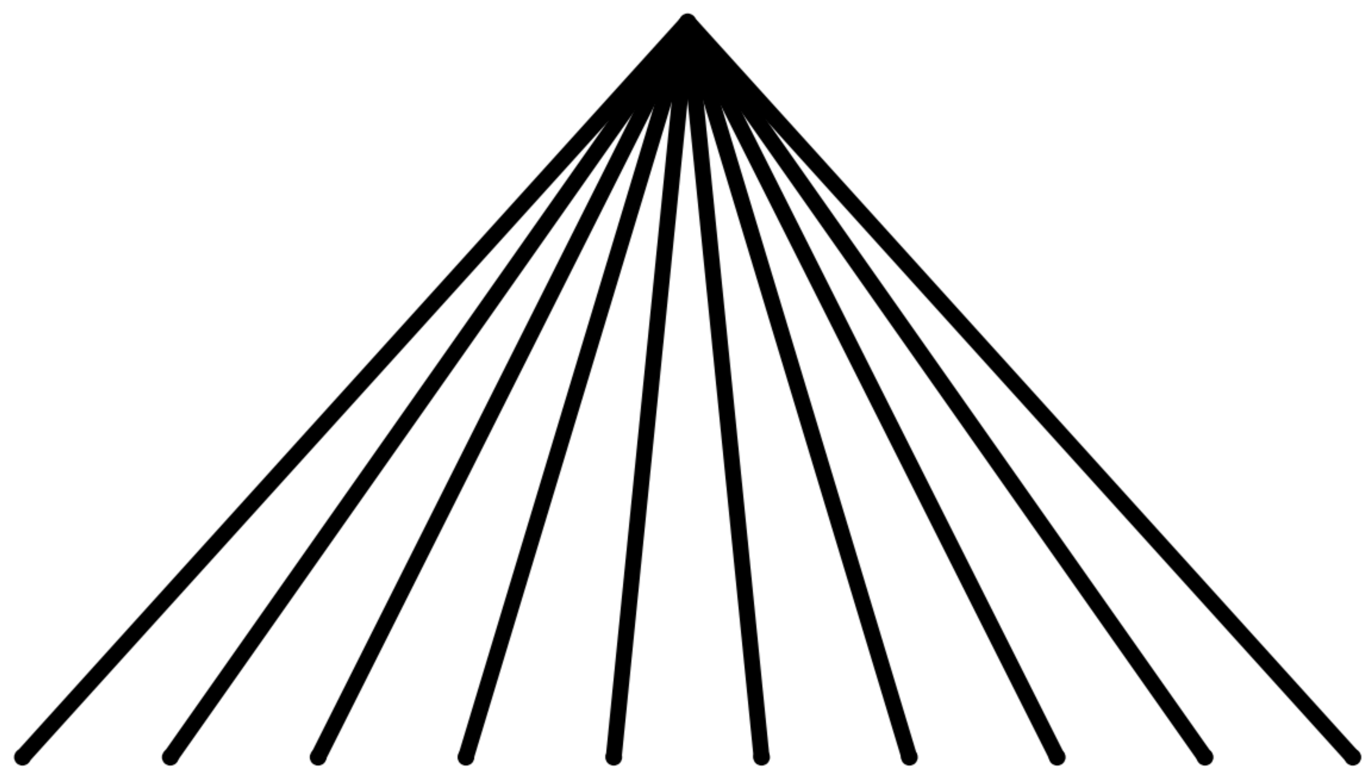
- Introspection
- Survey data
- Aggregate data
- Laboratory experiments
- Computer simulations



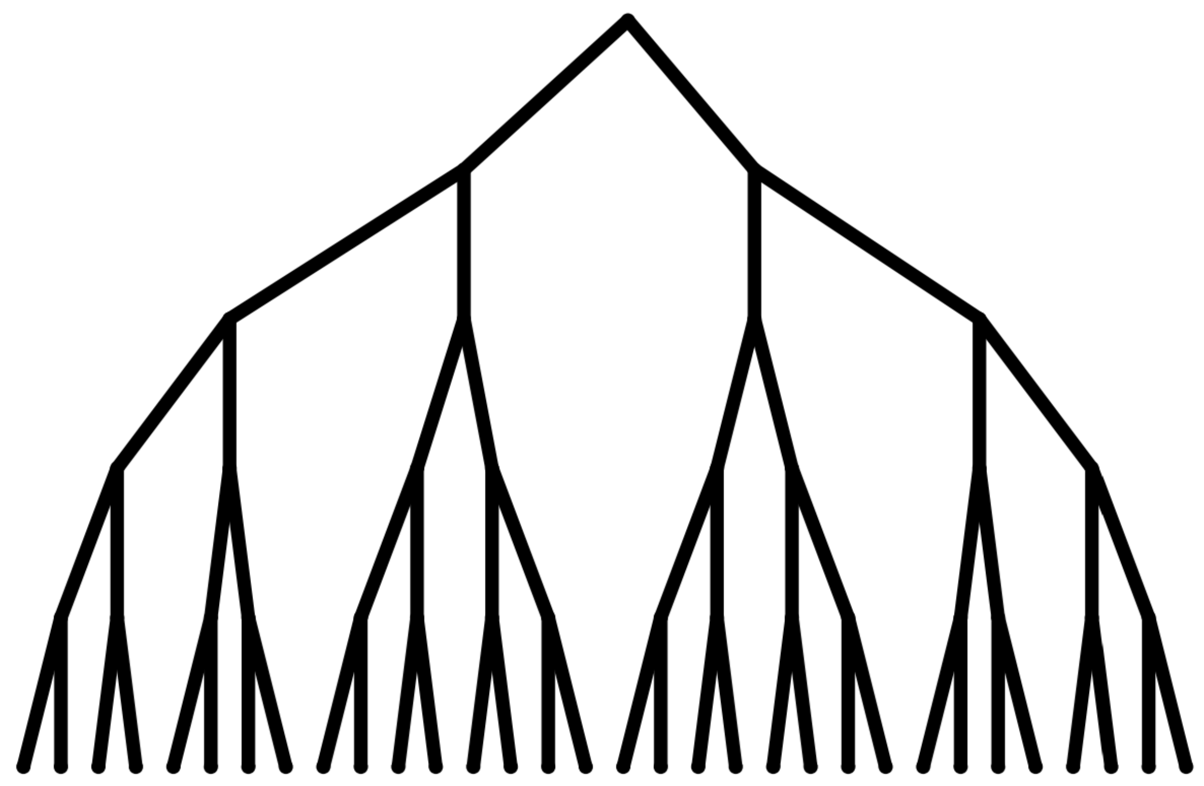
Problems?

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

Broadcast

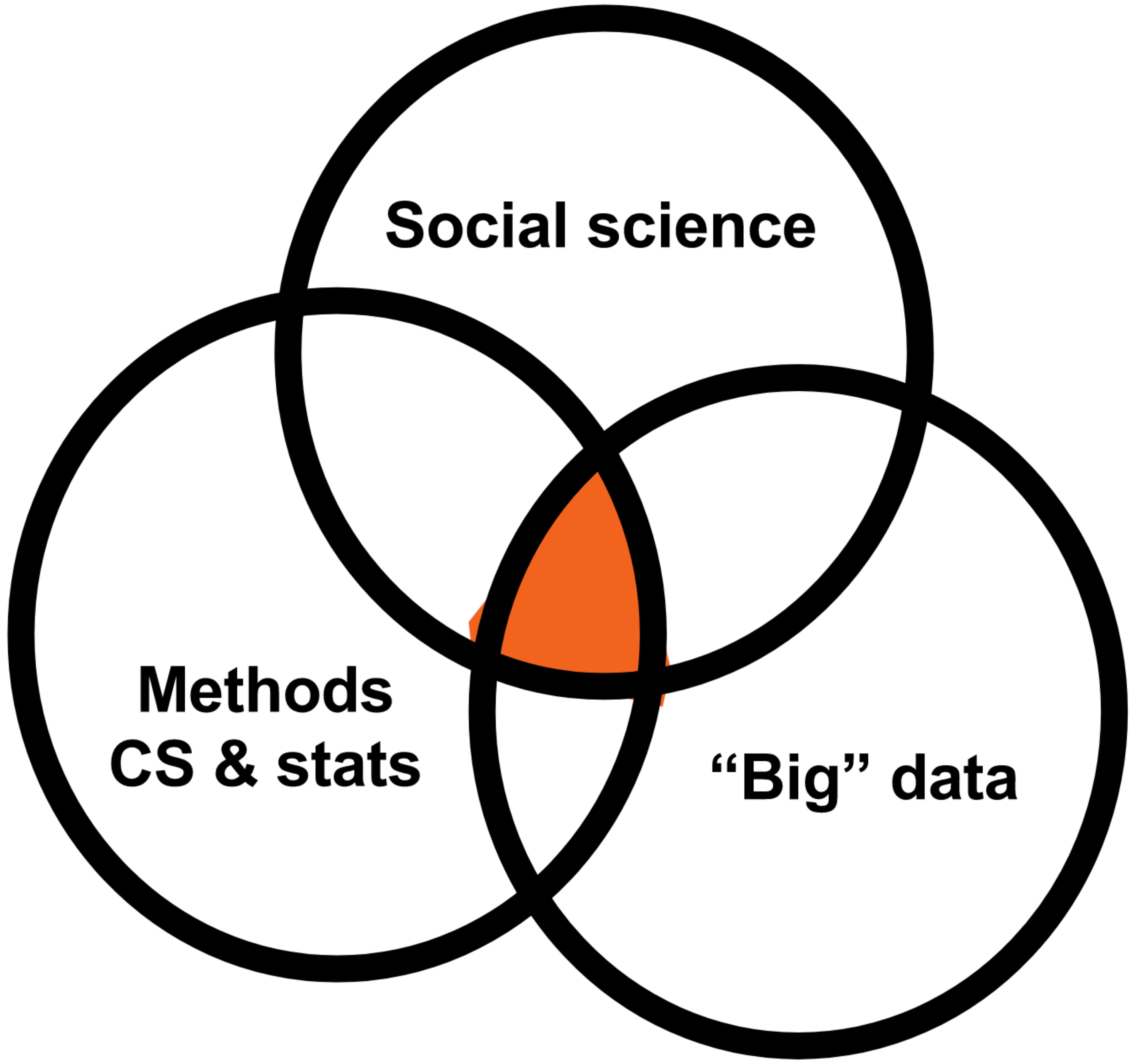


Viral



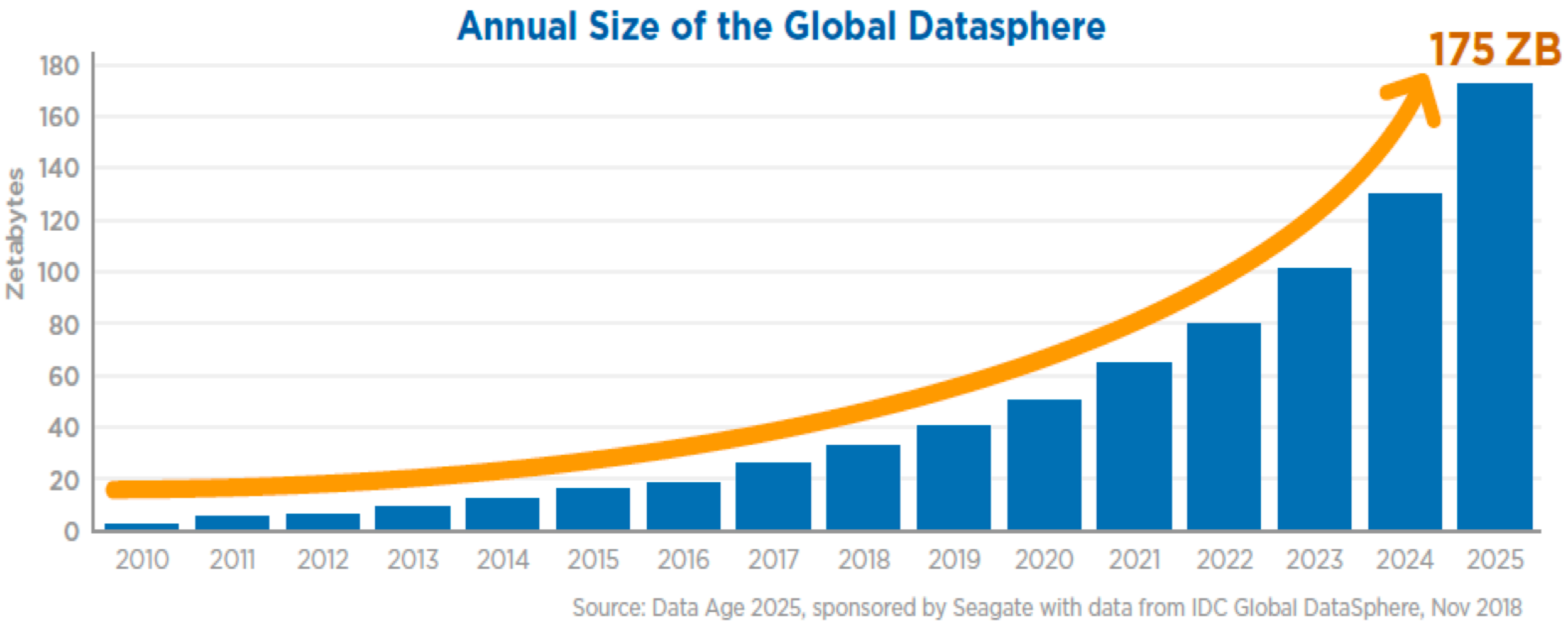
Computational social science

Social research in the digital age



The digital age is creating huge new opportunities for social research

Why now? Revolutions in data availability



Why now? Revolutions in computing

Massively distributed computing
- MapReduce, Spark, cloud computing

Big-memory machines
- Terabytes of RAM

Advances in machine learning
- Deep learning, transformers, large language models

Fast streaming algorithms
- Streaming aggregation, stochastic gradient descent

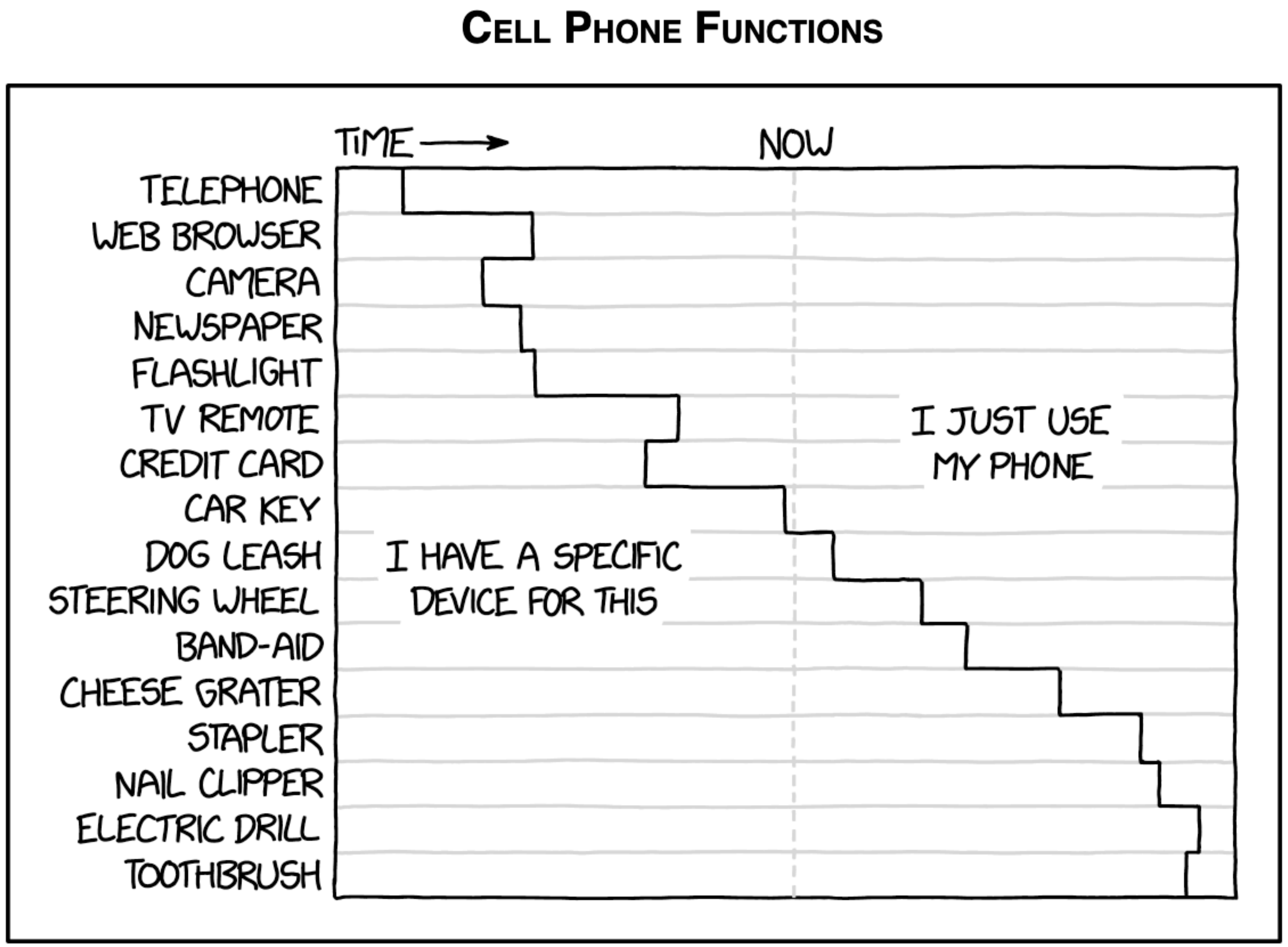
Human computation
- Crowdsourcing, Mechanical Turk

Why now? Revolutions in digitization

Everything online

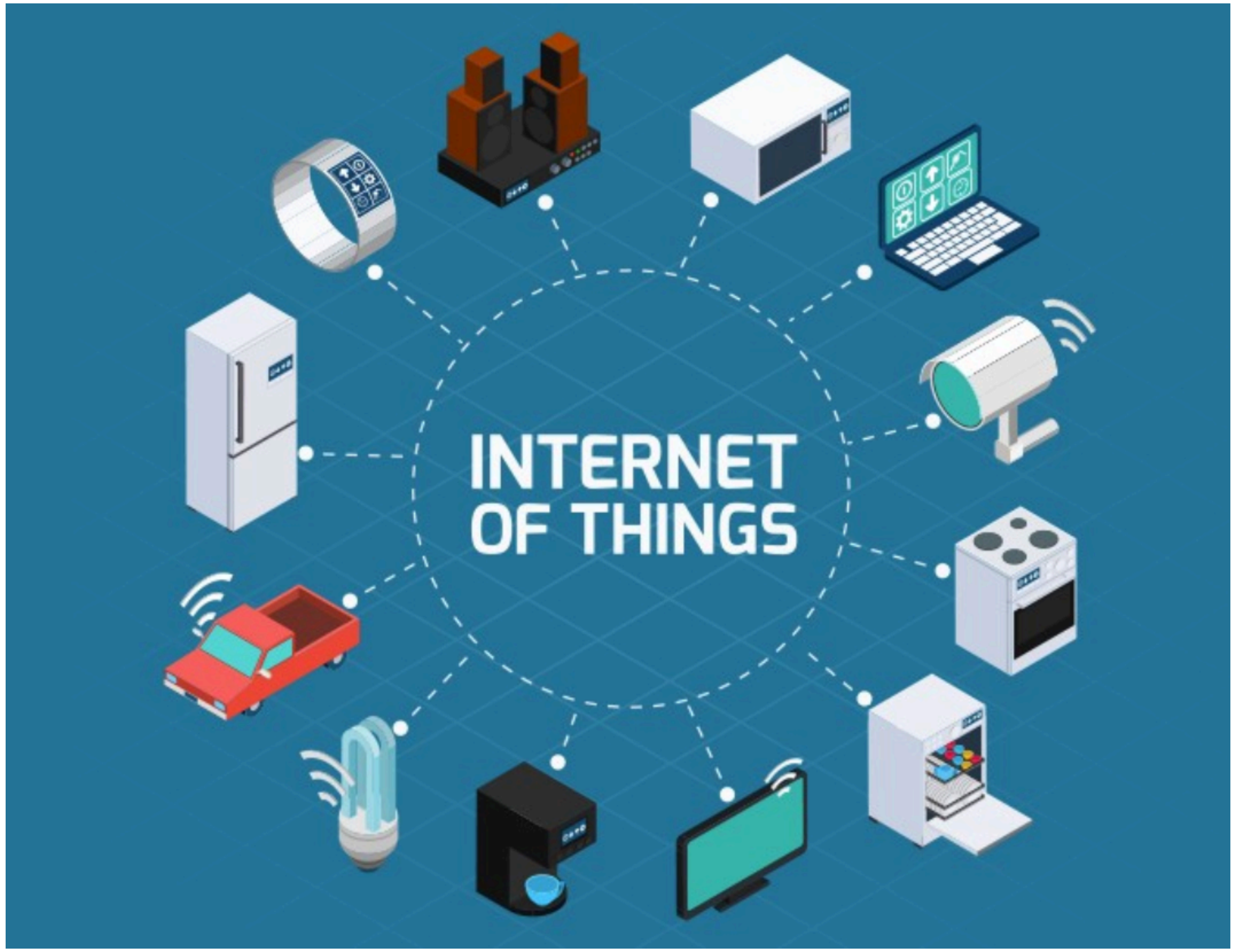
Why now? Revolutions in digitization

Computers everywhere



Why now? 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

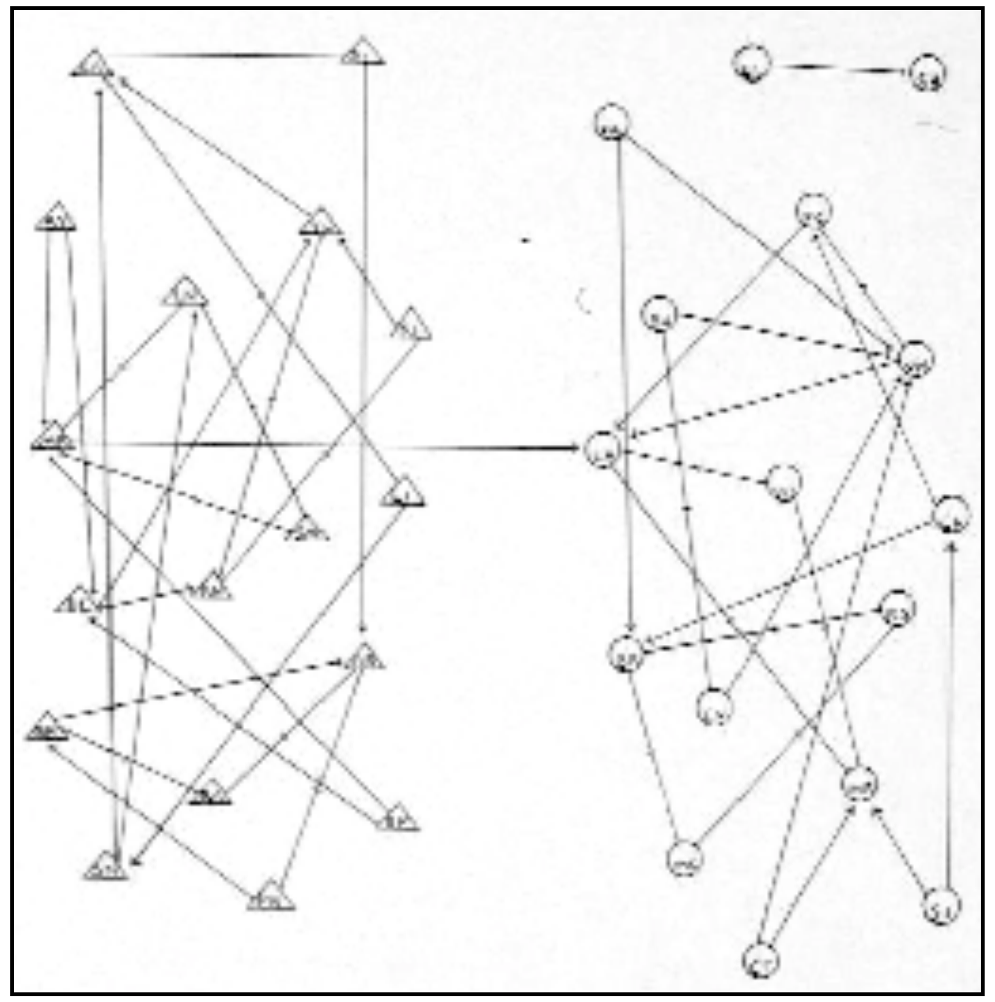
Revolutions in technology precipitate revolutions in science

Revolution in computational resources

+ Availability of large-scale human data

+ Developments in statistics

= Computational social science



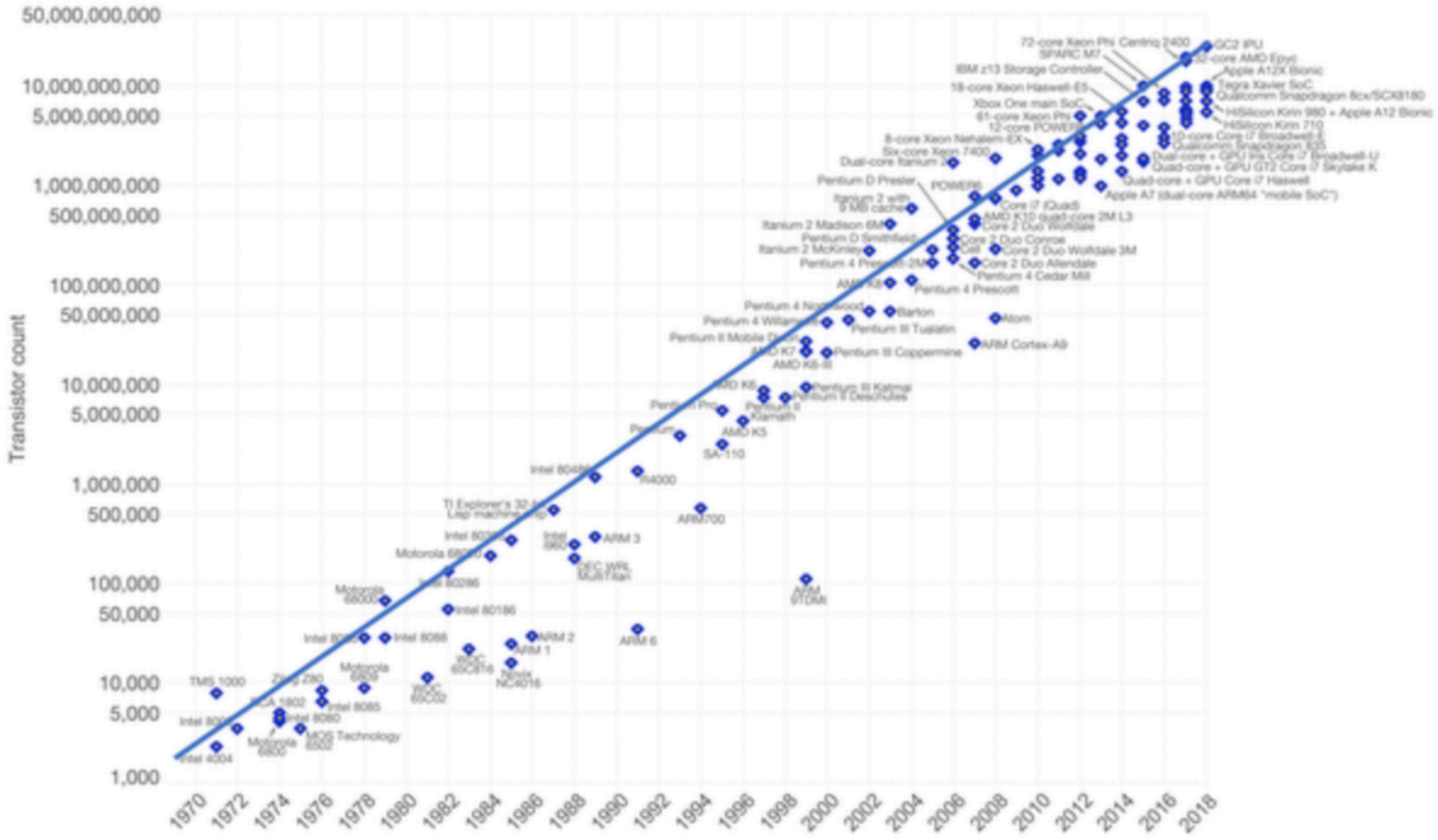
Computational Social Science

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

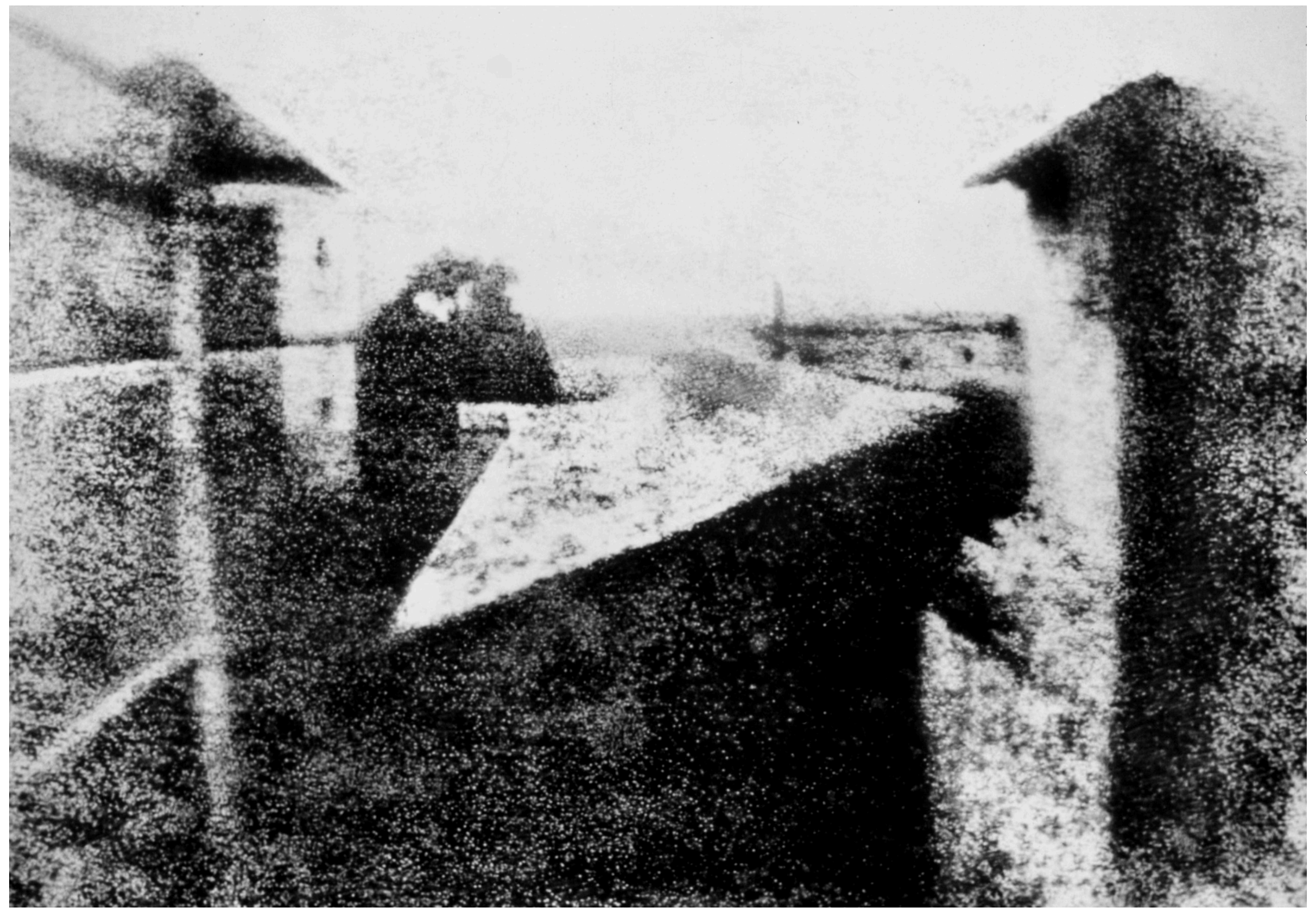
But wait… hasn’t this been happening for a long time?



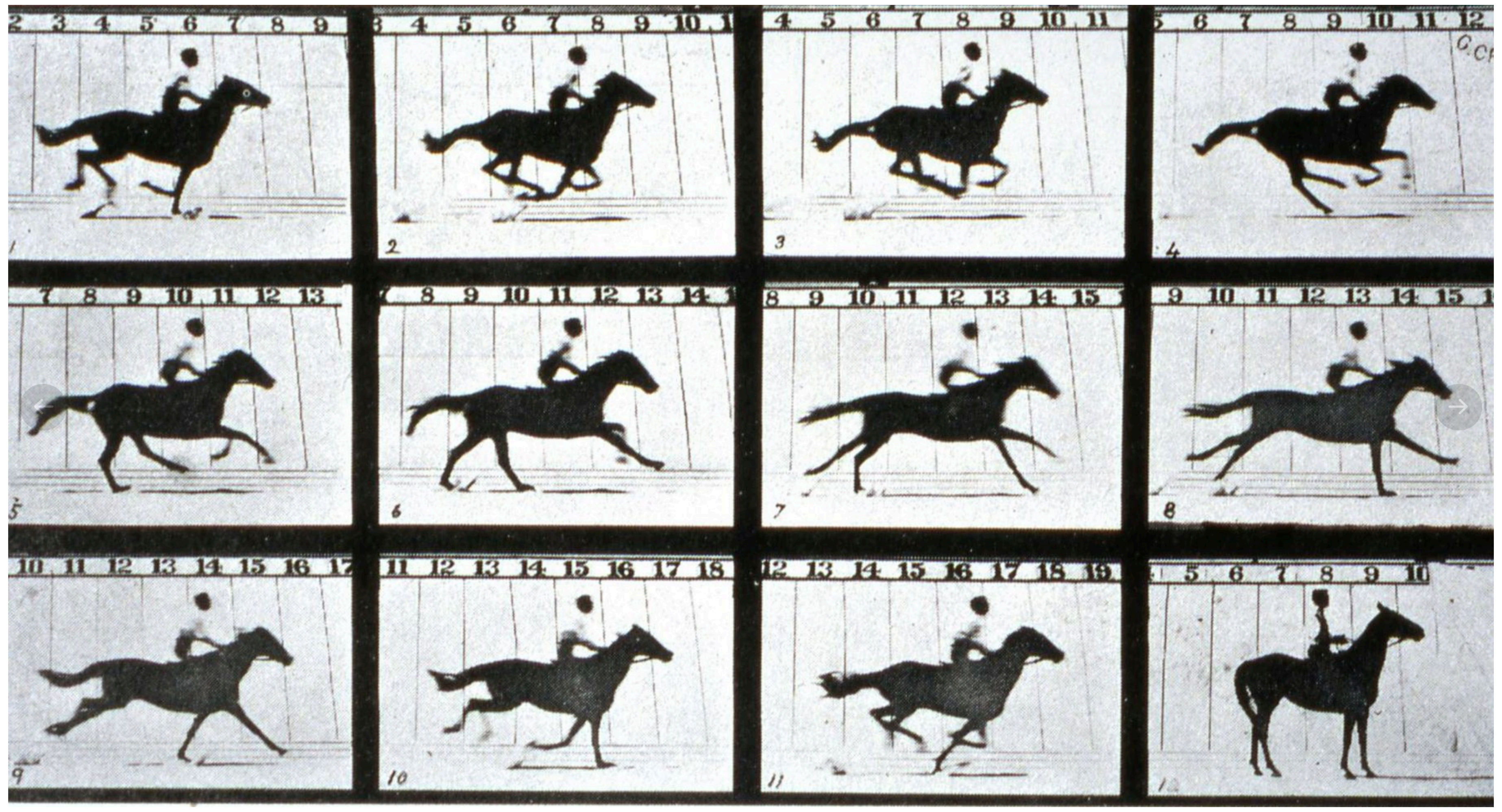
Moore’s law

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

About Me

0–18 → 18–22 → 22–29 → 29–31 → 31+

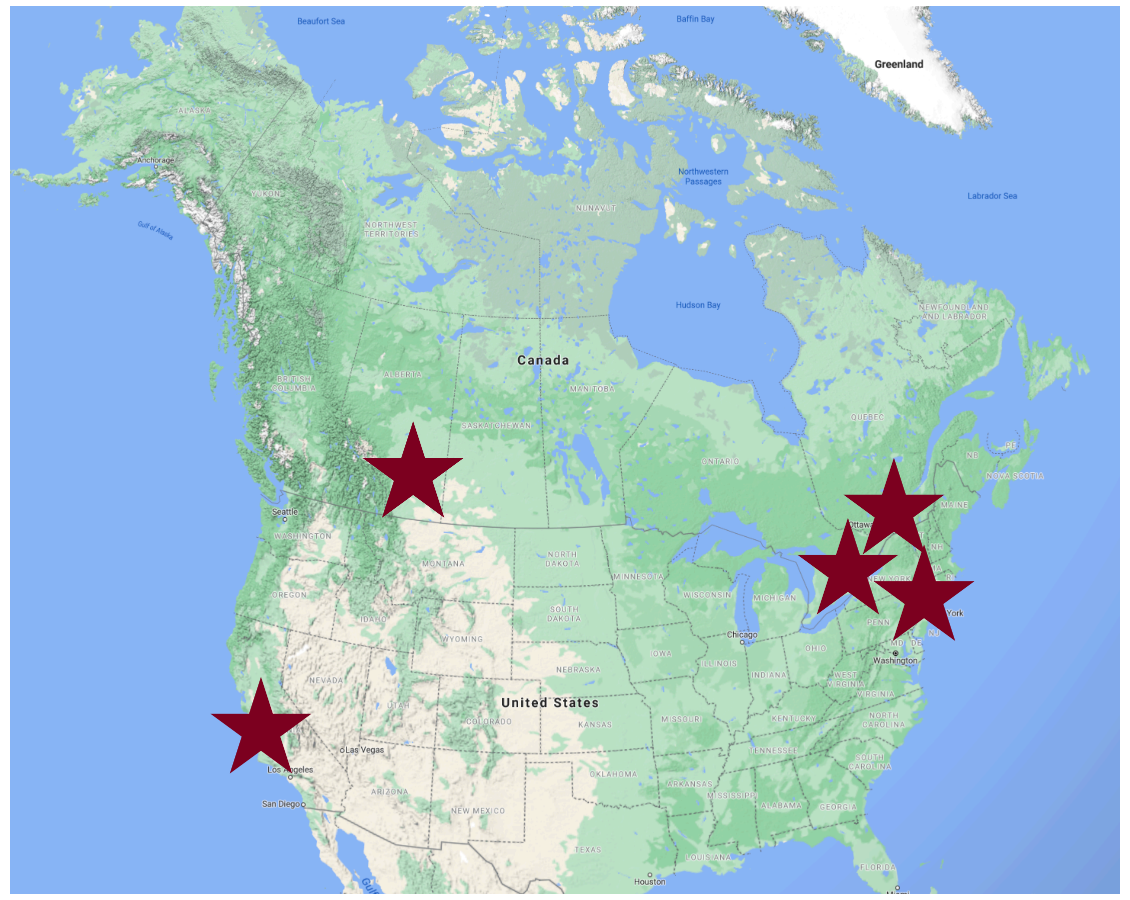
Calgary → Montreal → San Francisco → New York City → Toronto

1M → 4M → 7M → 20M → 6M

Now: Assistant Professor of Computer Science at U of T

Head of the Computational Social Science Lab (researching questions in AI, data, and society) 🤓

Computational Social Science Lab | Computer Science University of Toronto

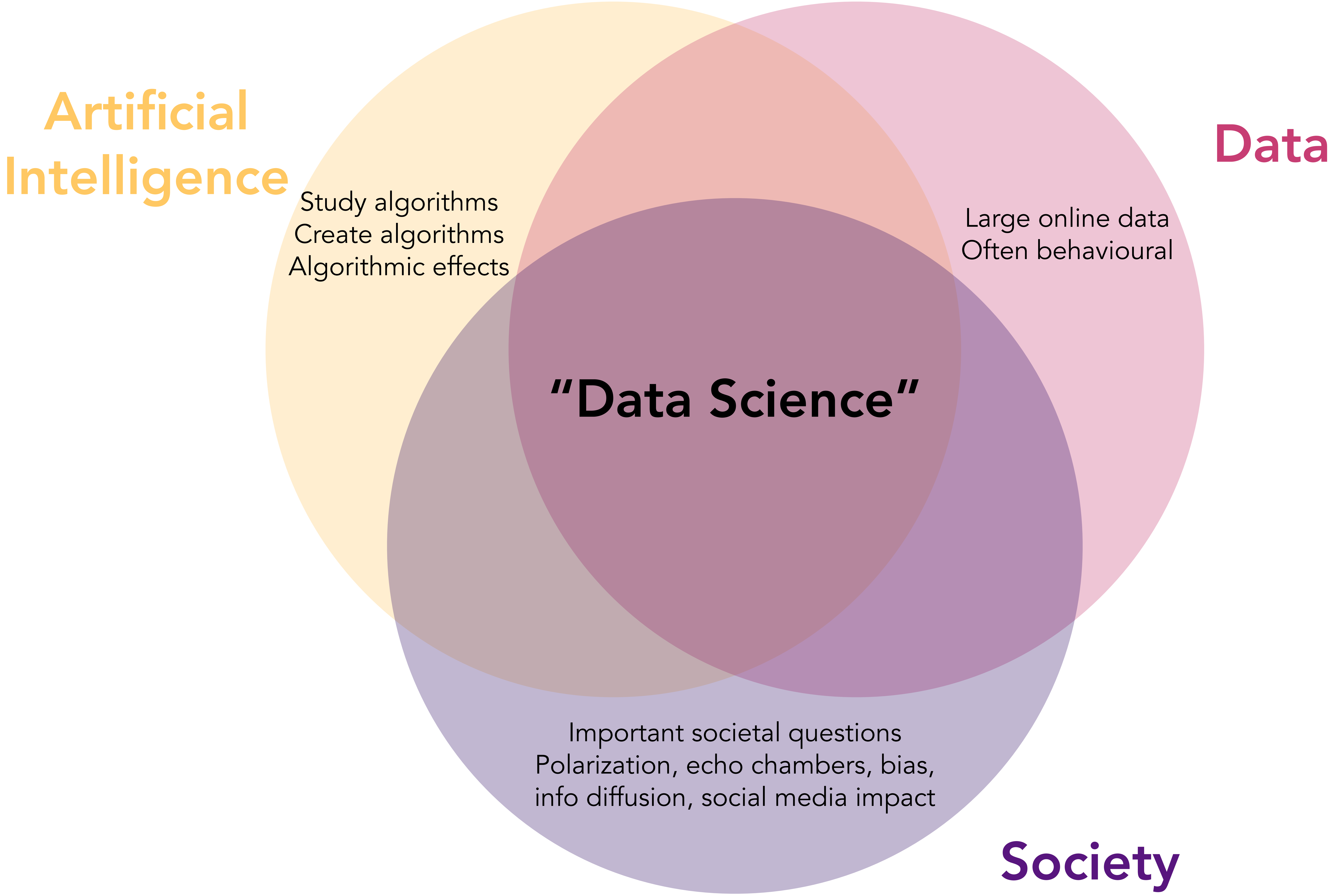


(Want to get involved? Email me after the course!)

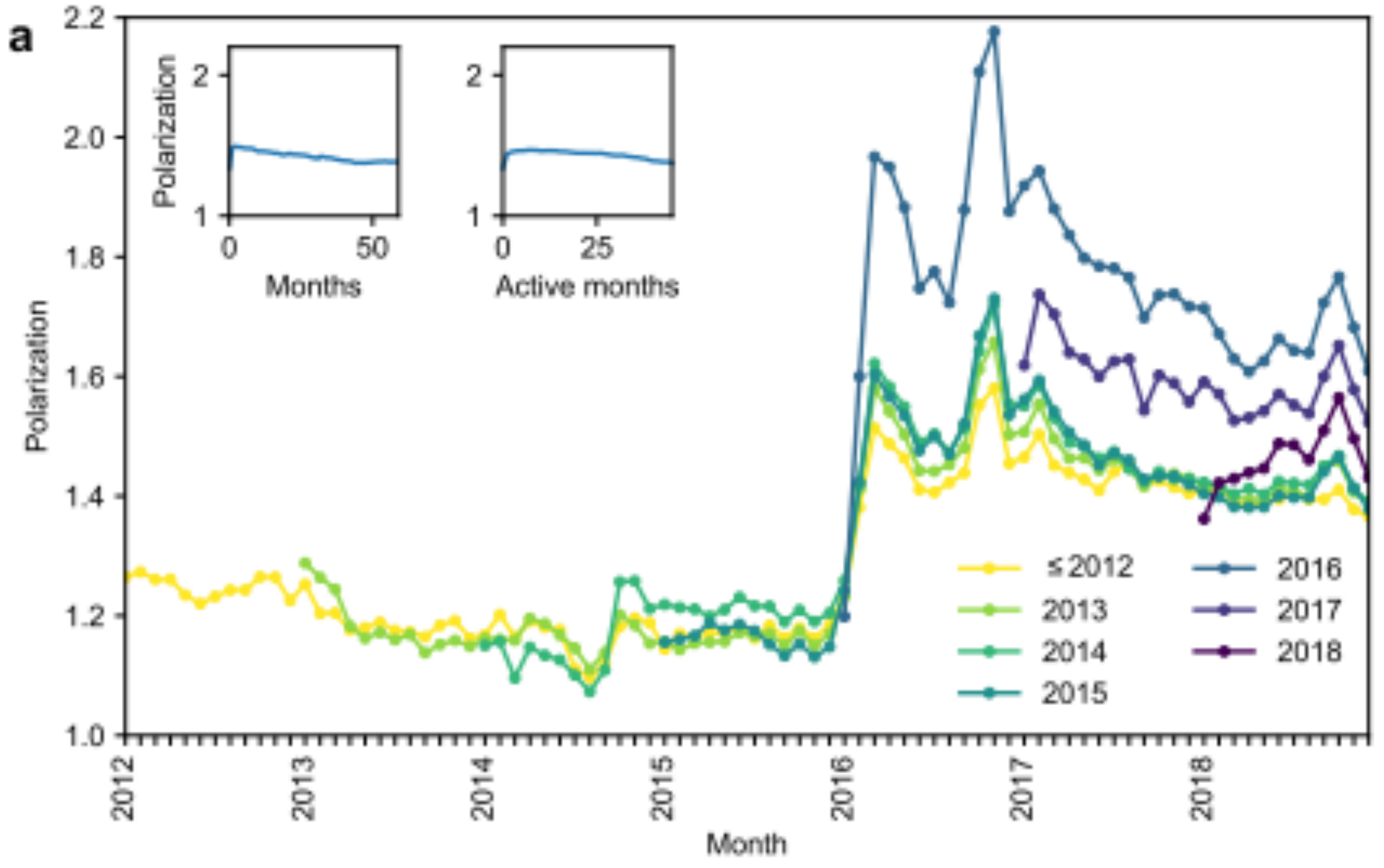
My path

Stage	Interests
McGill B.Soft.Eng ‘08	Theoretical Quantum algorithms and information Anything practical was impure
Stanford Master’s ‘10	“Hmm…would be nice to feel more connected to the world” Game theory: computational/economic lens on strategic interaction Mix of theoretical and applied
Stanford Ph.D. ‘15	Discovered the joy and power of large-scale empirical analysis Computational social science: social research in the digital age Mostly empirical analysis supplemented with theoretical modeling, experimentation, and surveys

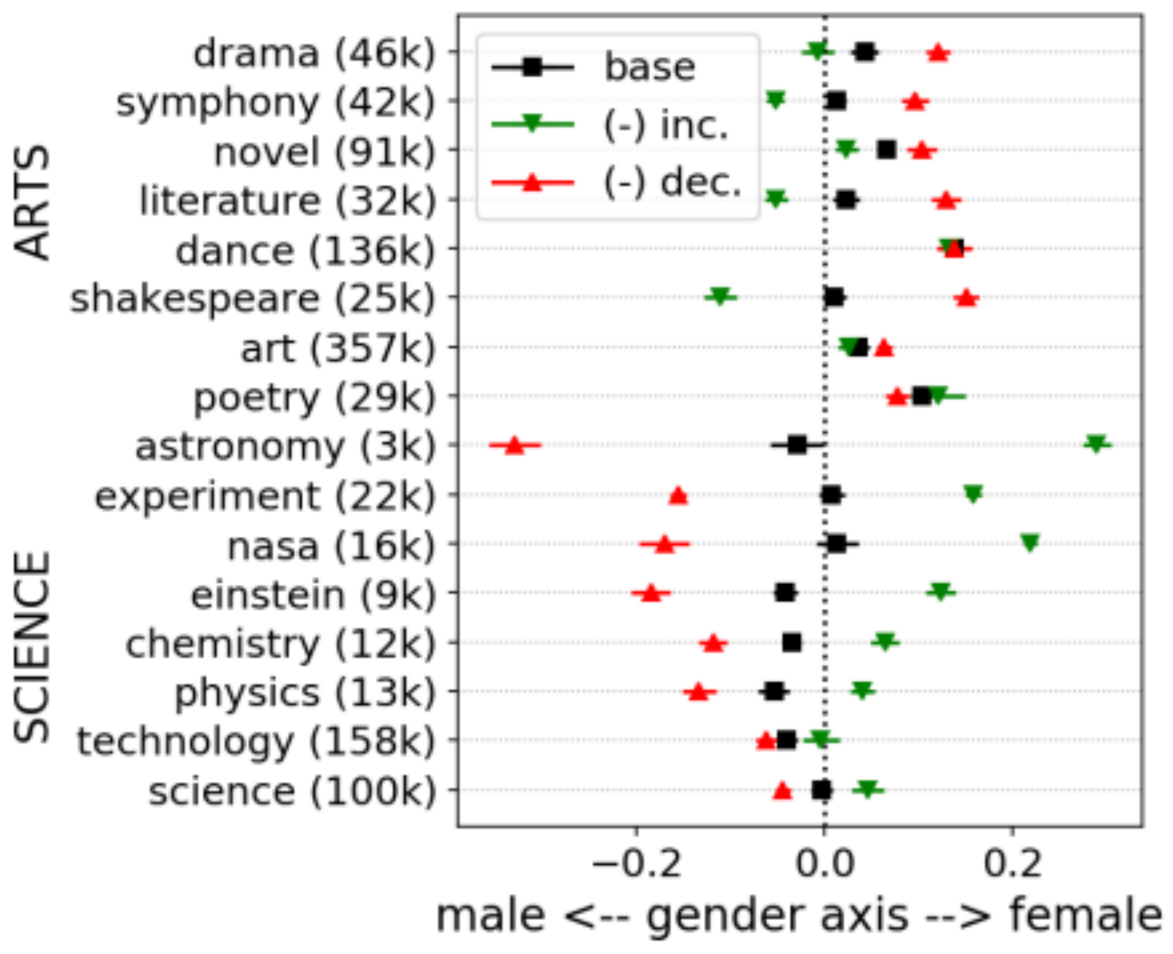
My research


Artificial Intelligence
Study algorithms
Create algorithms
Algorithmic effects
Data
Large online data
Often behavioural
"Data Science"
Important societal questions
Polarization, echo chambers, bias,
info diffusion, social media impact
Society

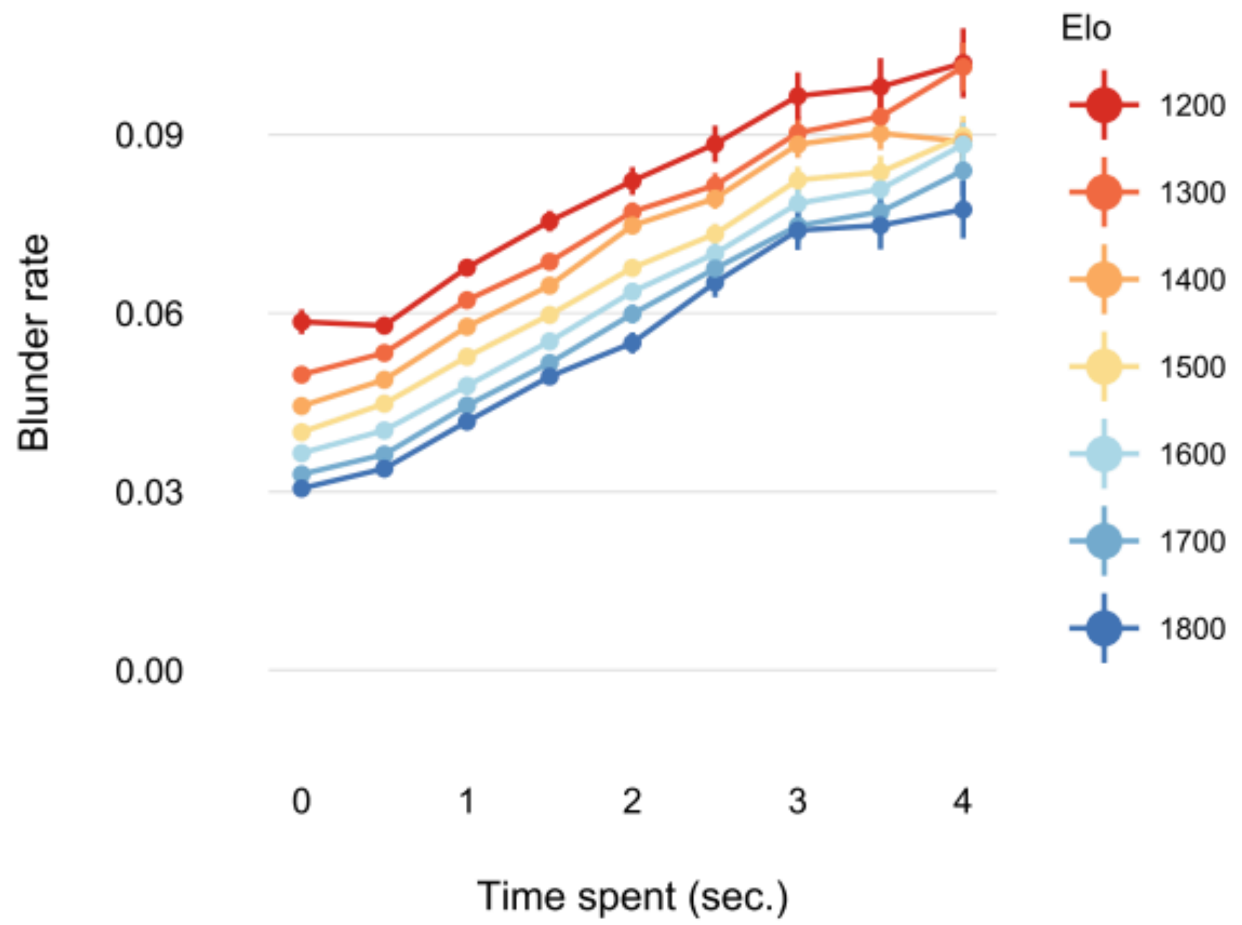
My research



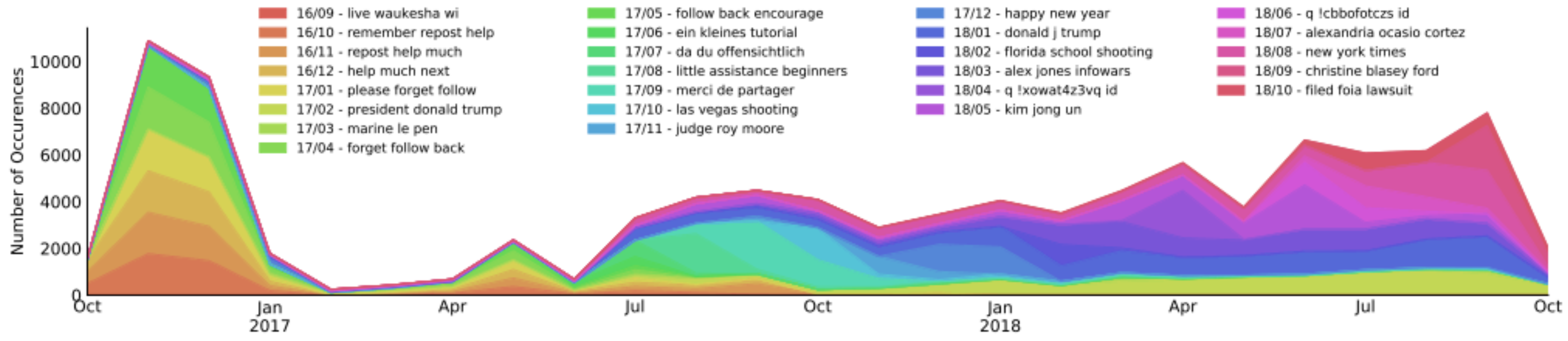
Political polarization on Reddit



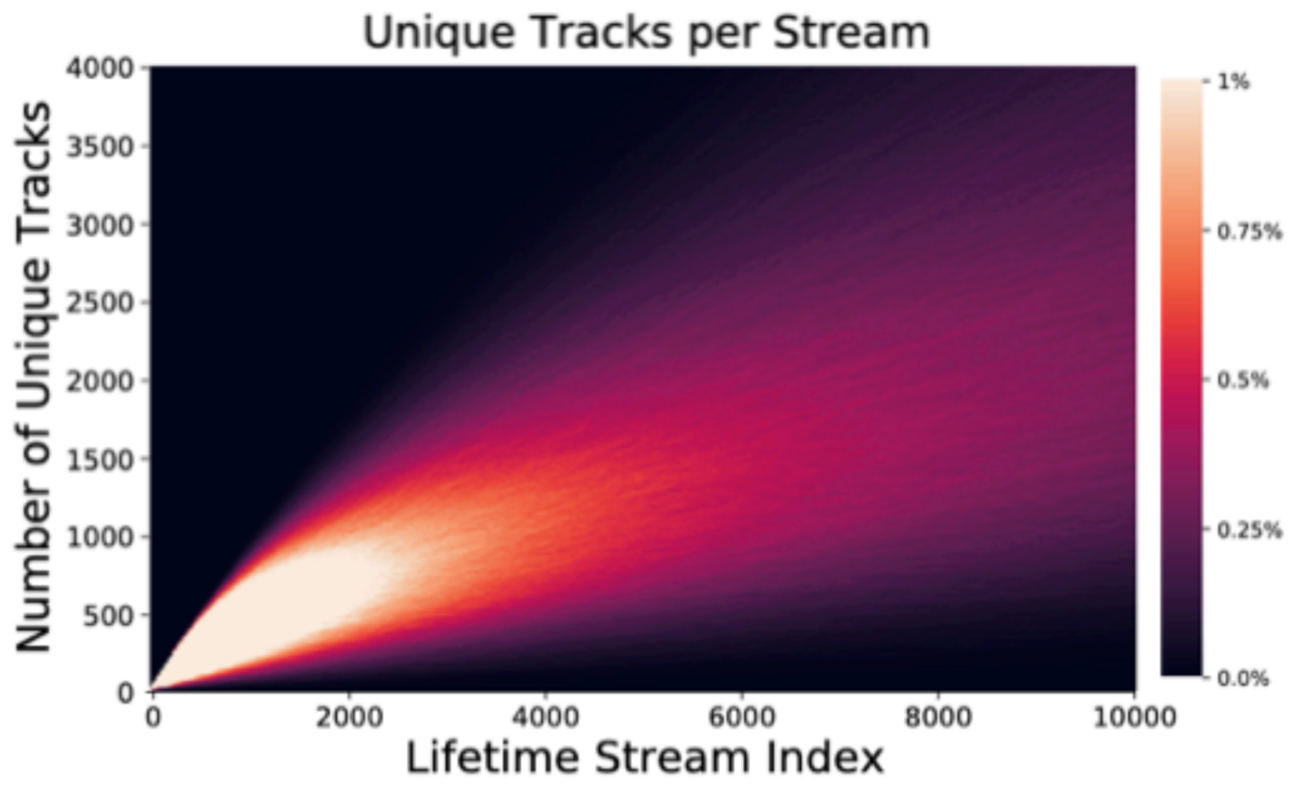
Gender bias in text algorithms



Nature of human error in chess



Discussion topics on Gab (alt-right platform)



Music exploration on Spotify

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

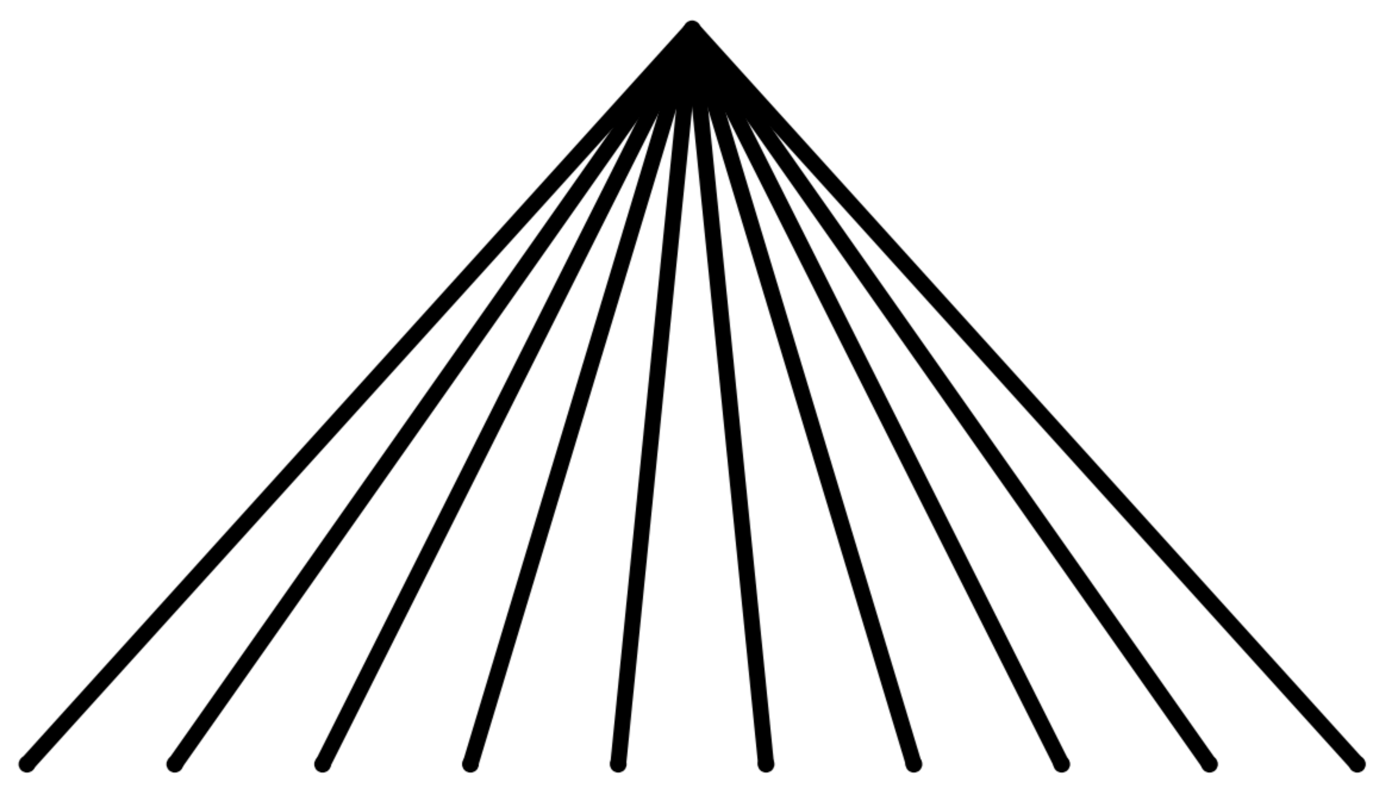
Final project

- 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

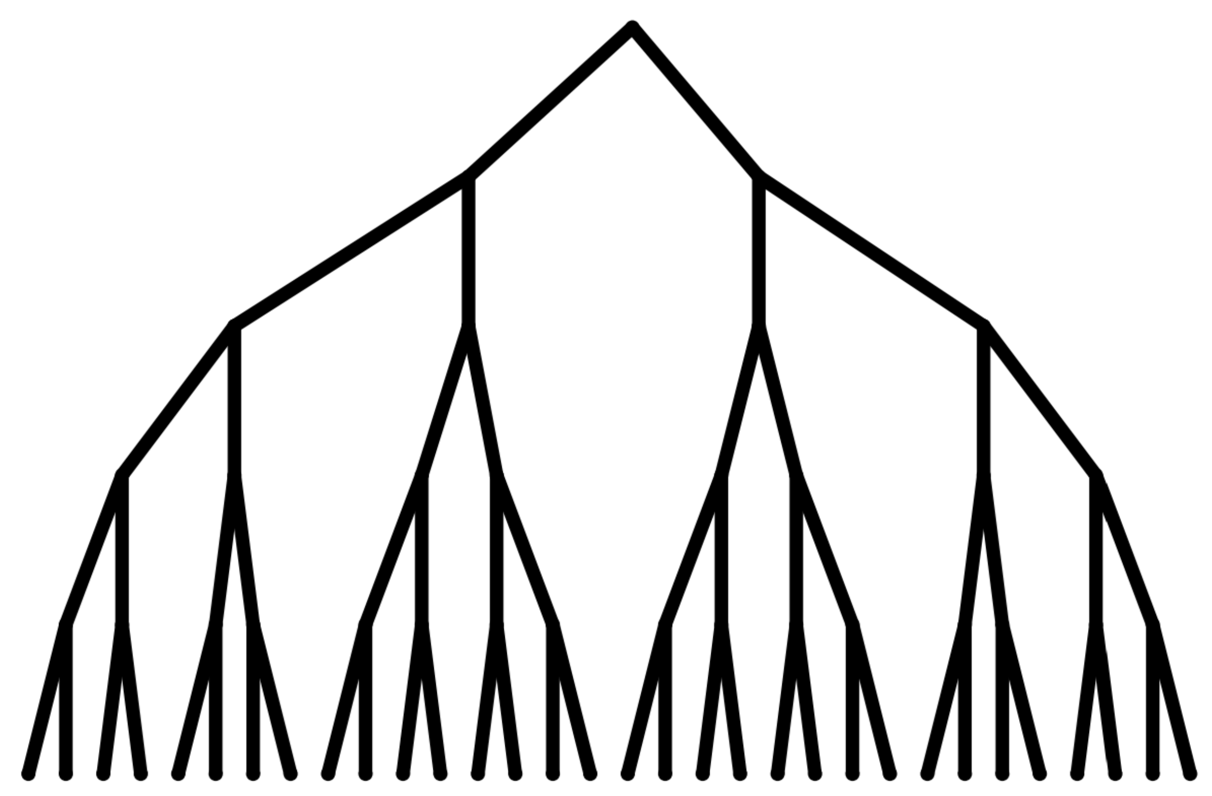
Back to the question

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

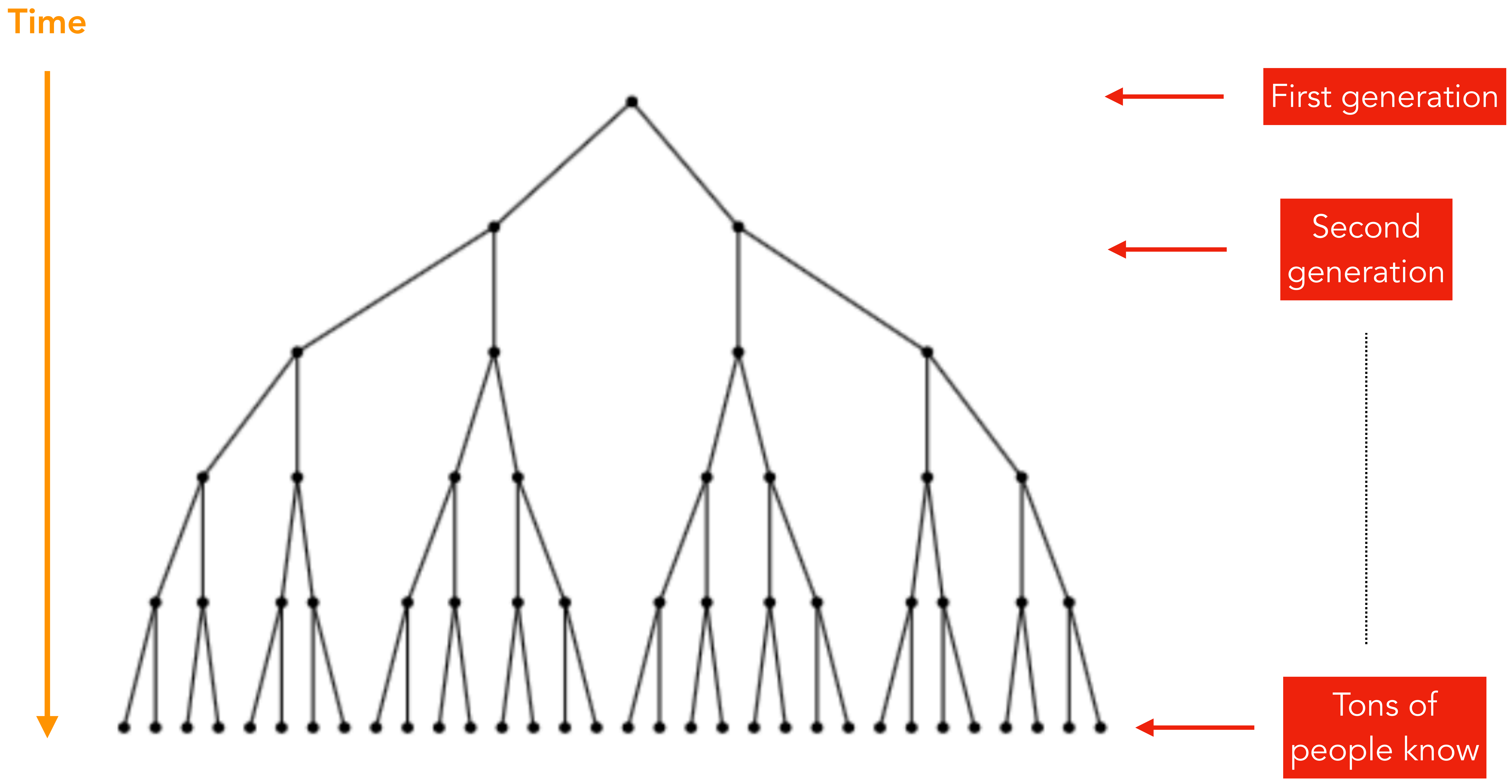
How to quantify structure?

What is "virality"?

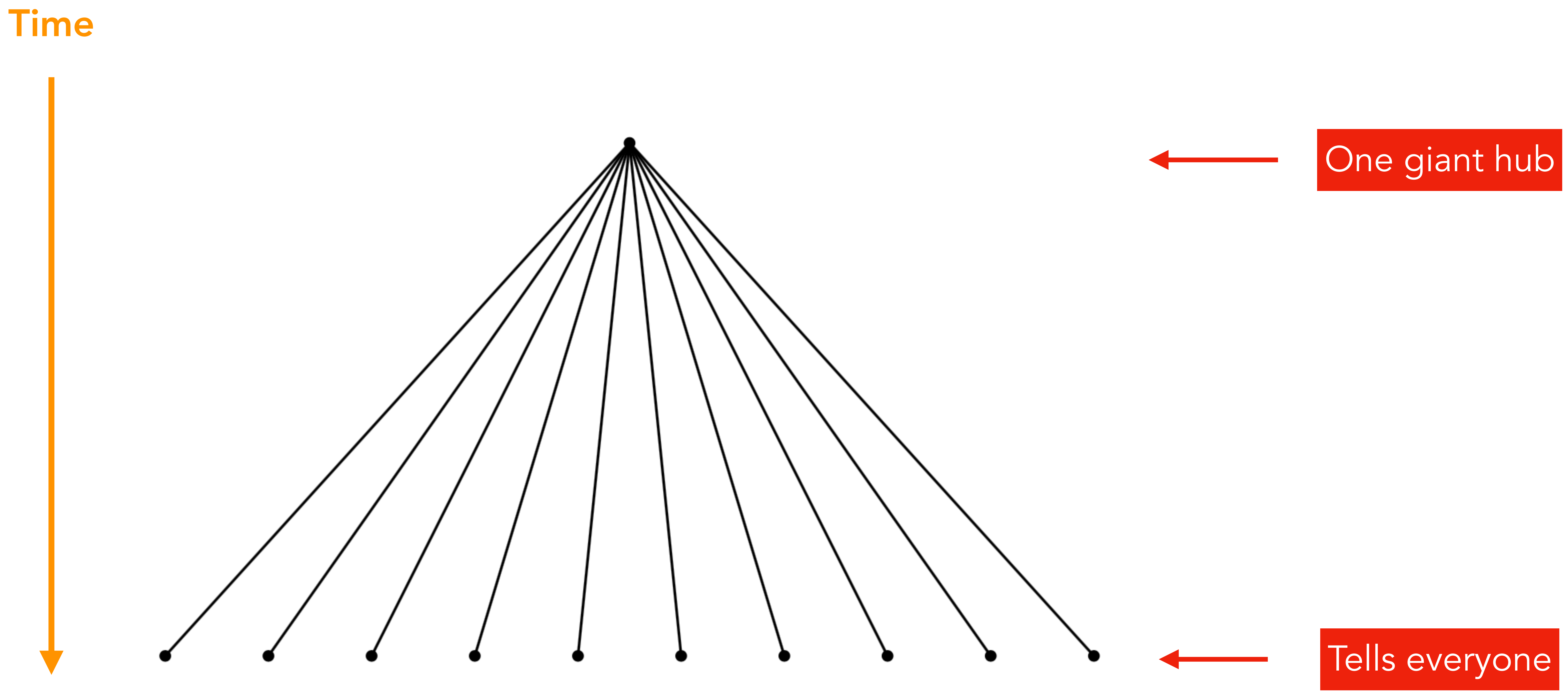
Methodological challenges

How do you analyze 1 billion cascades?

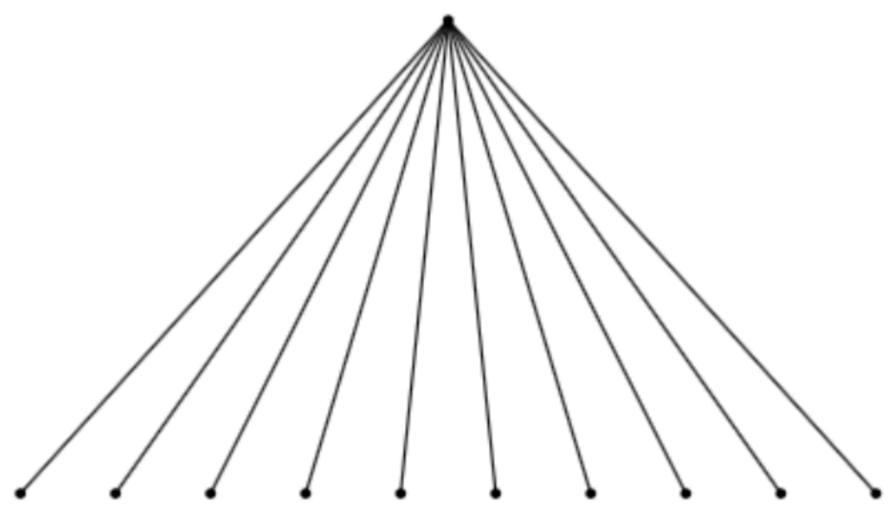
Viral diffusion



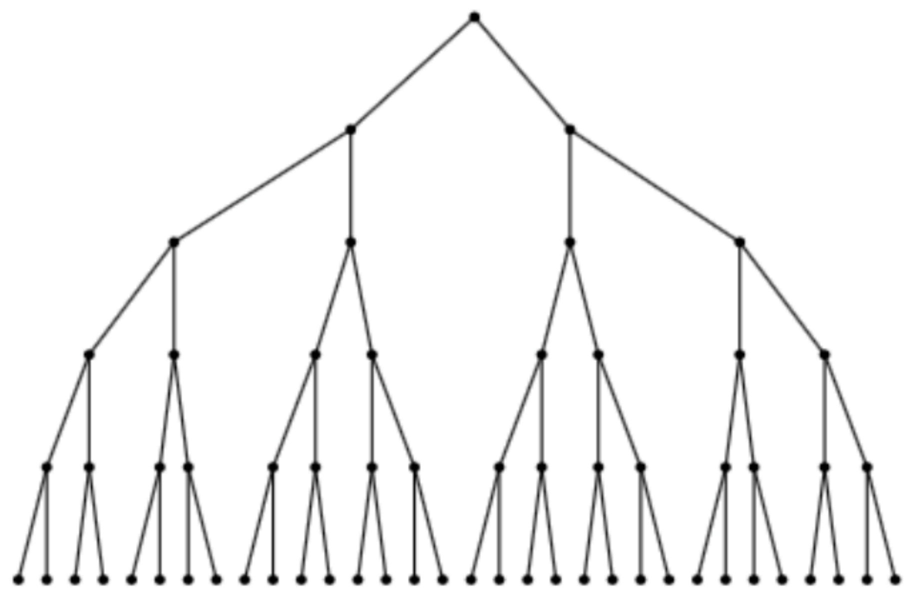
Broadcast diffusion



Which is it?



or



"Broadcast"

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

"Viral"

- Organically spreading content
- Chain letters

How to study information spread?

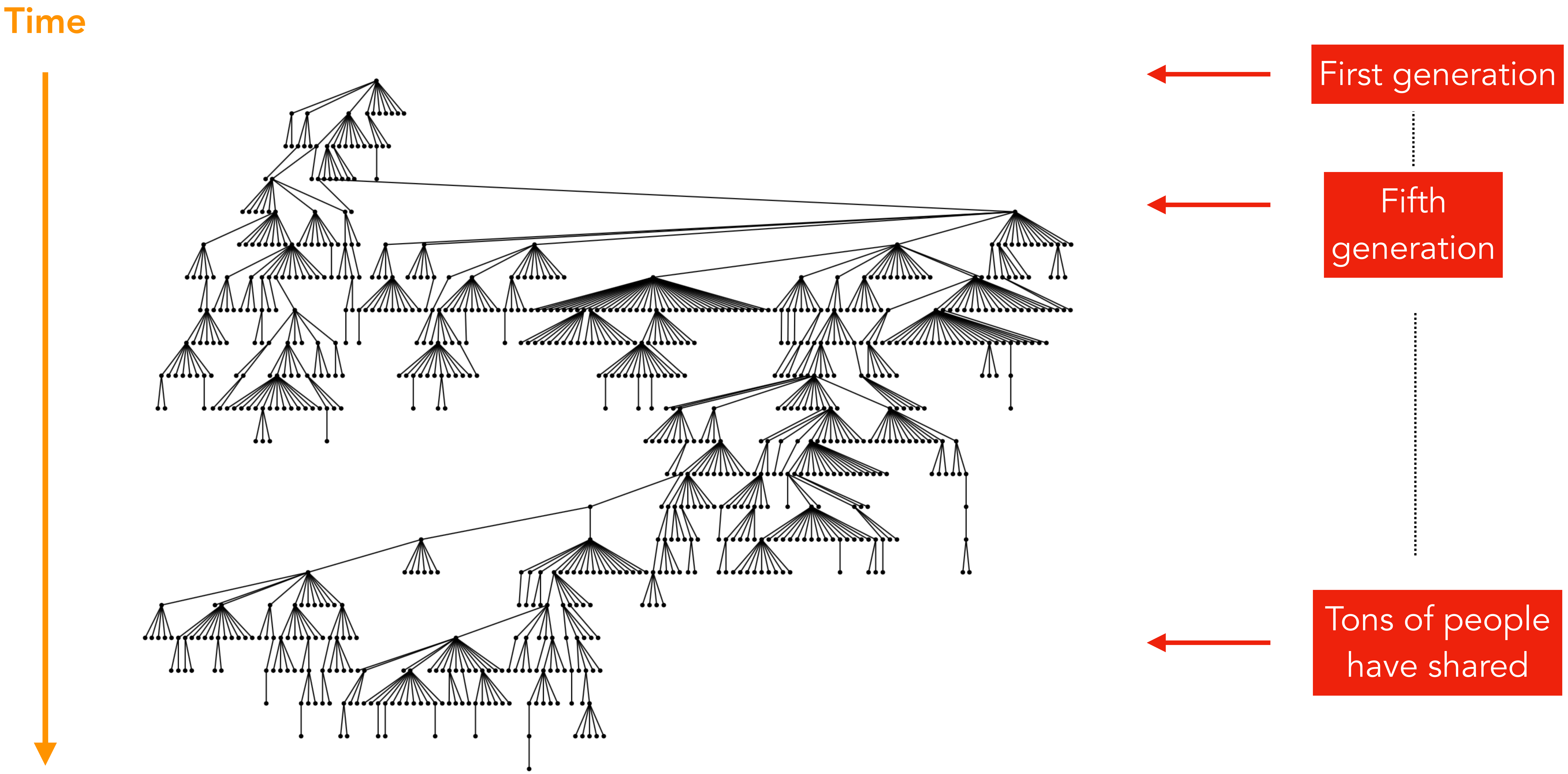
Hard to track “information” spreading from one mind to another

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

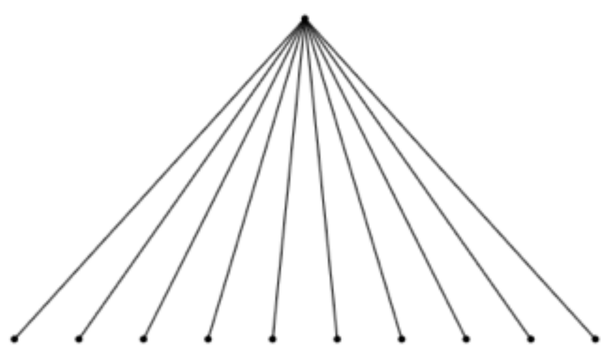
How to study information spread?

Connect these sharing edges into **trees**



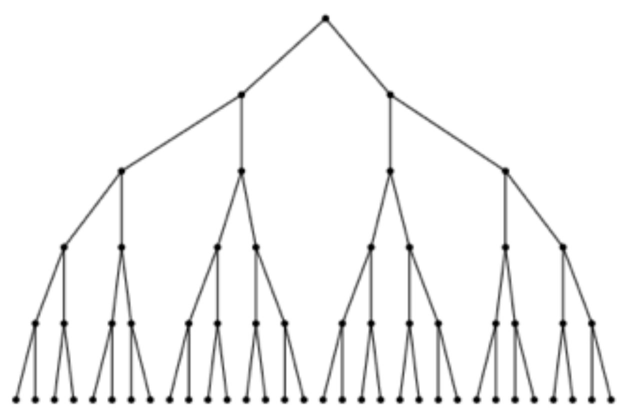
How to measure virality?

How **structurally viral** is a particular cascade?

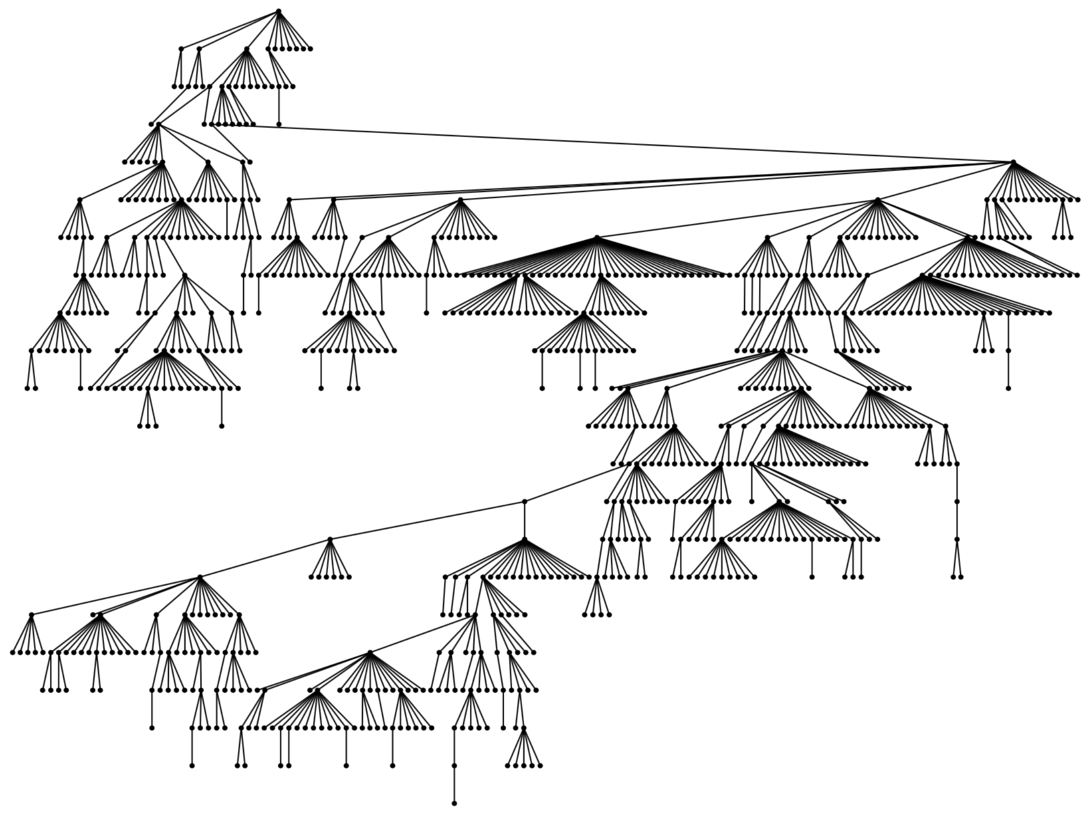


Not viral

?



Super viral



How to measure virality?

One idea: **depth of the cascade**

But this is **sensitive to a single long chain**

Not viral — Super viral

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!

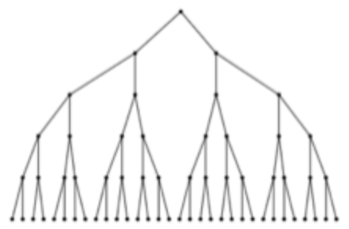
Now we have a way to construct information cascades on Twitter

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

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



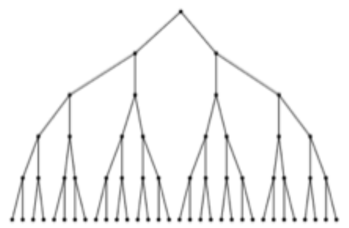
Super viral

Measure virality in data!

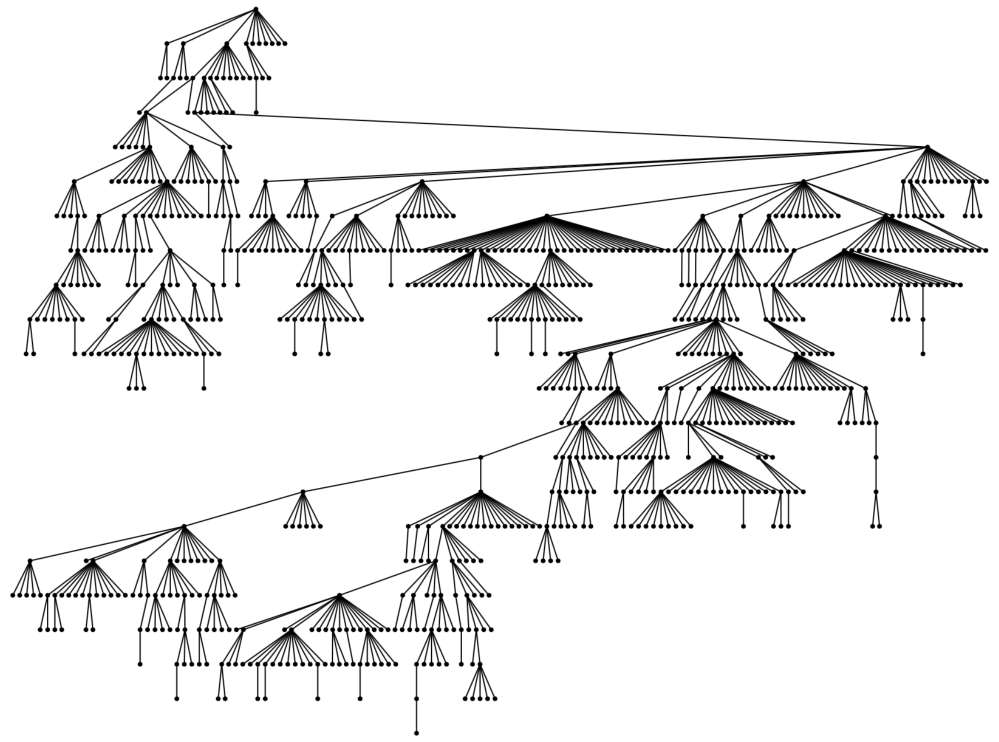
- Looked at an **entire year of Twitter data**
- 622 million unique URLs, 1.2 billion “adoptions” (tweets) of these URLs
- Every URL is associated with a forest of trees

Not viral

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



Super viral

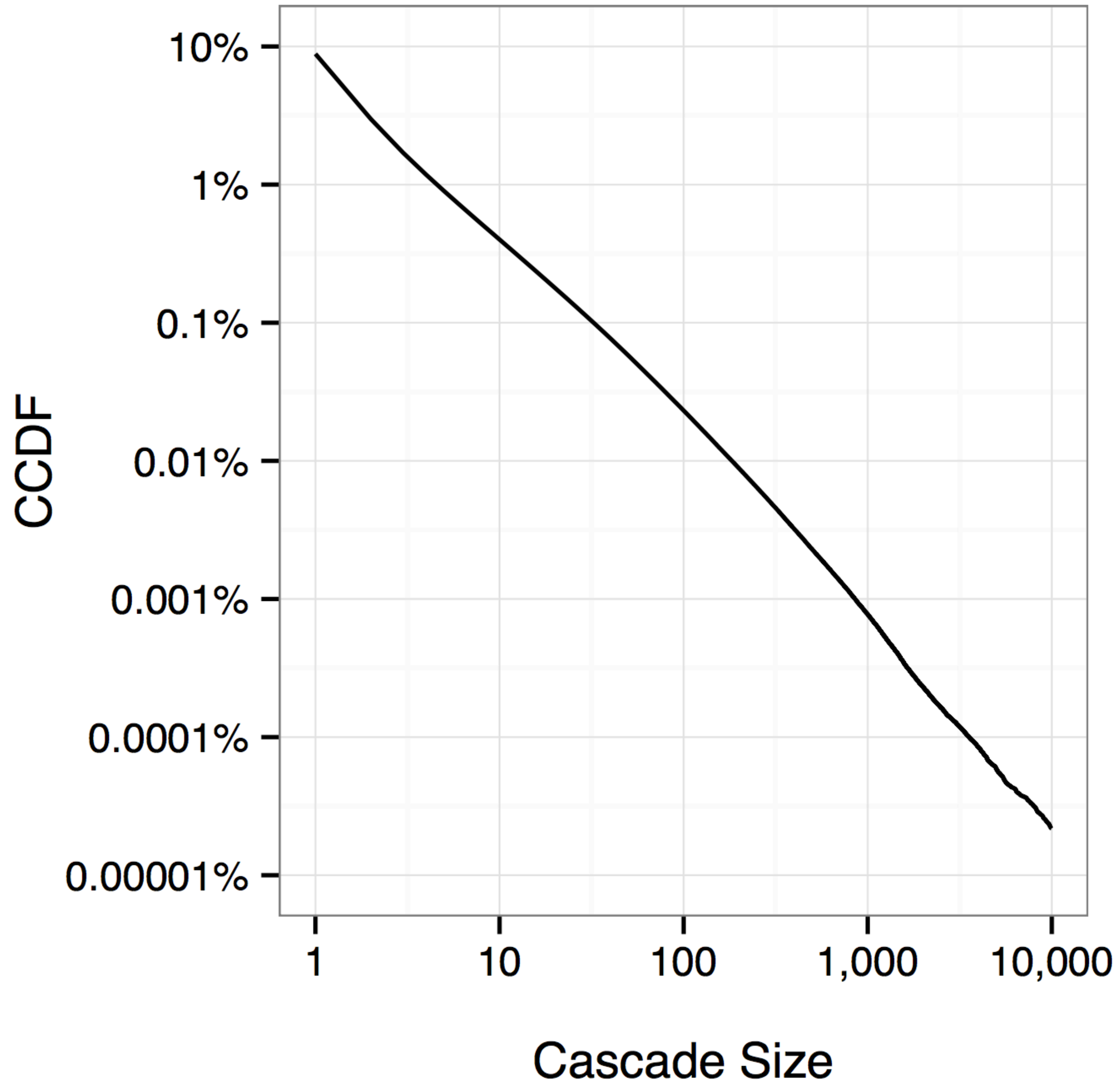


Measure virality in data!

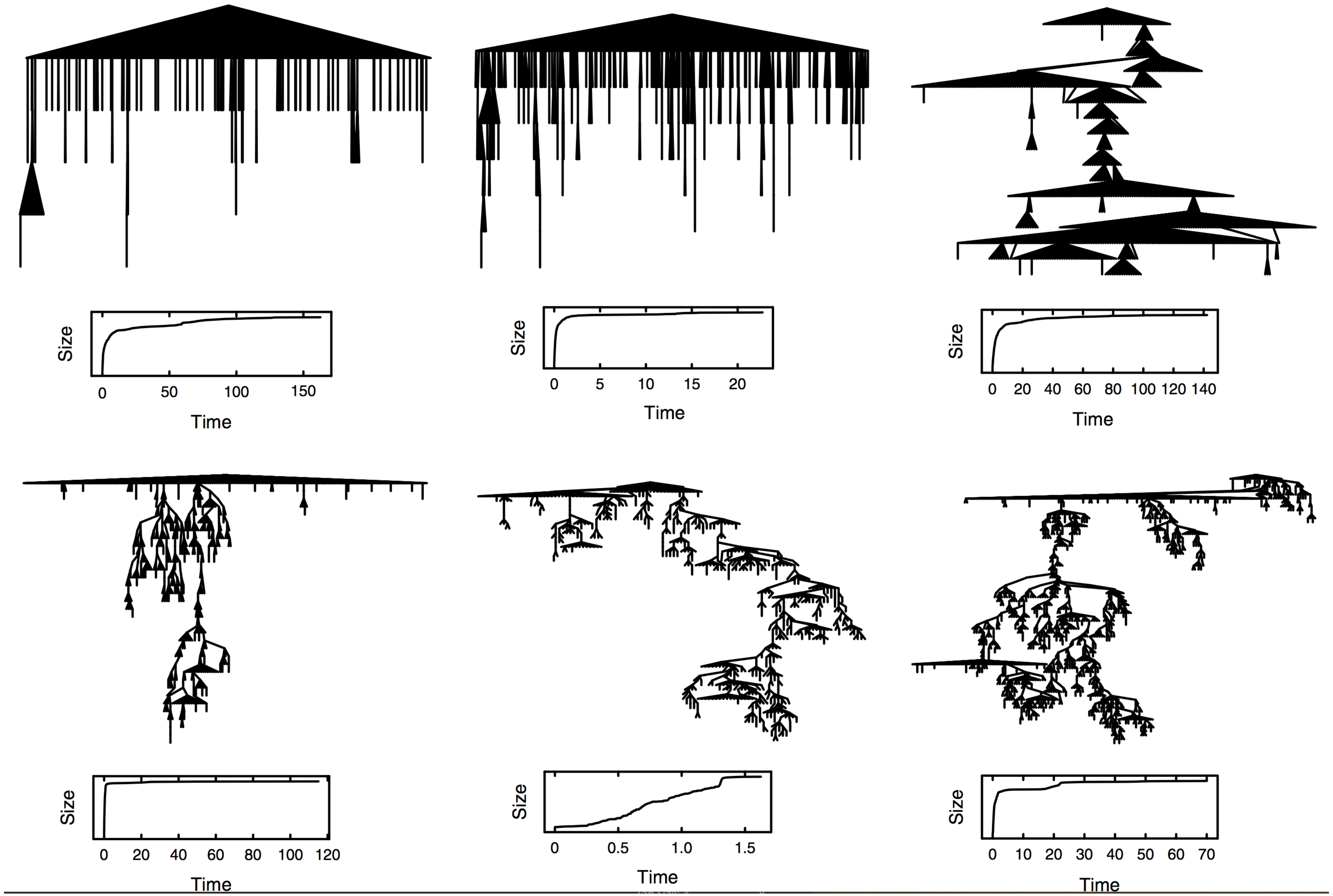
First conclusion: most stuff goes nowhere

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

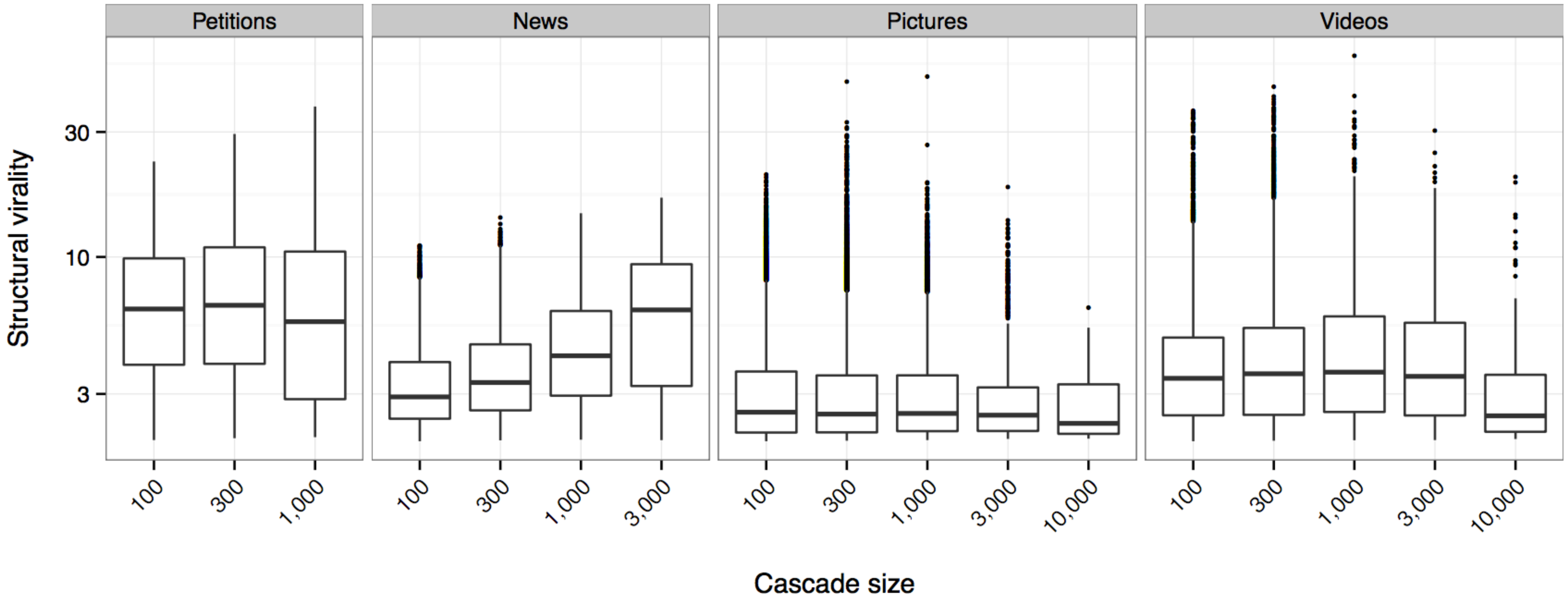


Surprising diversity at every scale

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



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)