Data Science Process and Objectives

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Presentation of Project Plans

Use form for feedback

What is Data Science?

What is science?

- From the Latin word scientia, meaning **knowledge**
- A **systematic** enterprise that builds and organizes knowledge in the form of **testable explanations and predictions** about the universe

So what is data science?

- Data Science seeks to discover new knowledge by answering questions through data

What data science is **not**

https://xkcd.com/1838/

How to turn observational, biased, **scientifically "weak" data** into strong scientific results?

Fundamental Data Science Challenges

Scientific method in data science

1. Ask a **question**.
2. State a **hypothesis** about the answer to the question.
3. Make a **testable prediction** that would provide evidence in favor of the hypothesis if correct.
4. Test the prediction via an **experiment involving data.**
5. Draw the **appropriate conclusions** through analyses of experimental results.

[e.g. Roger Bacon, 1265]

Associated Challenges

Domain Knowledge & Theory

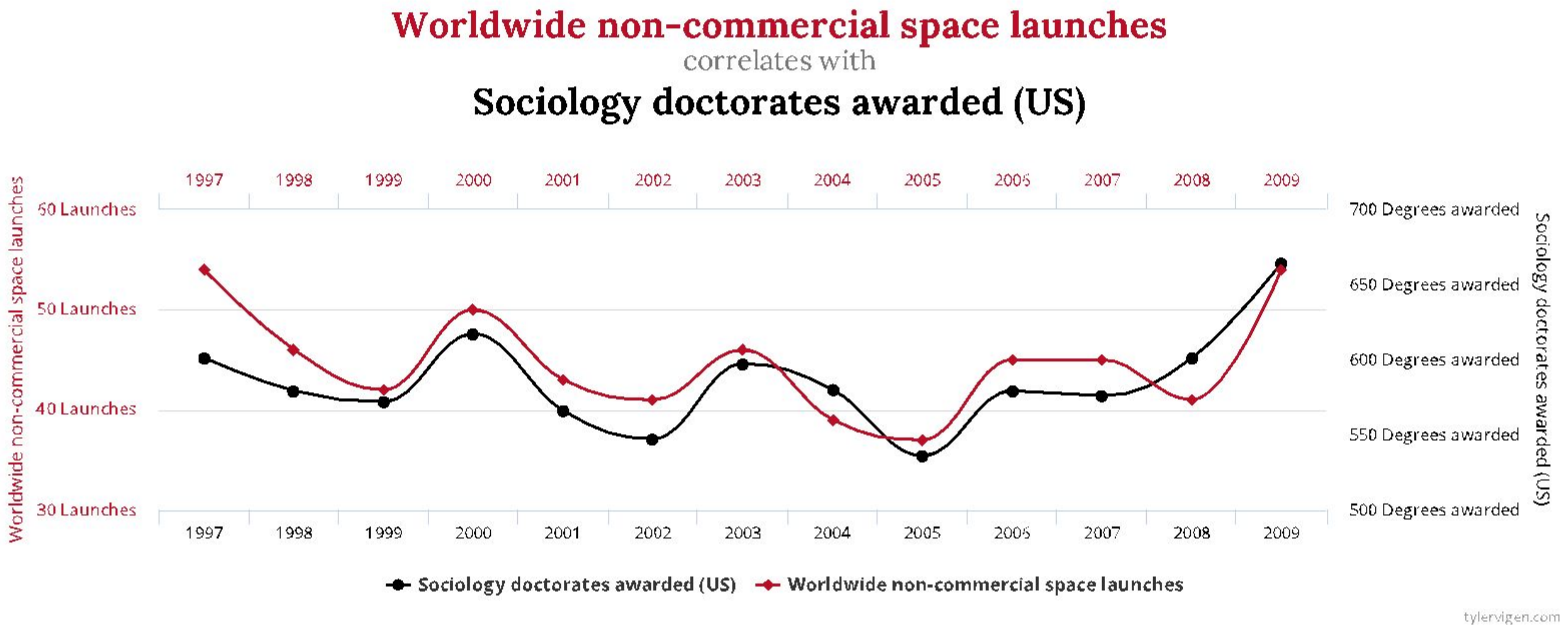
Construct Validity
Are you measuring what you think you are measuring?

Internal Validity
Confounding & Causal Inference
Robustness of findings

Model Intelligibility

External Validity
Incomplete picture of external world

Goal: Valid inferences from data



http://www.tylervigen.com/spurious-correlations

Prediction is not enough!

Causality

- We are typically interested in cause and effect
  - T causes Y if changing T leads to a change in Y *keeping everything else constant*
- Intervention: *What if we do X?*
- Counterfactual: *Was it X that caused Y? What if I had acted differently?*

We will learn about causality later in the course!

Interventionist definition (https://plato.stanford.edu/entries/causation-mani/)

Importance for Decision Making

- Which treatment should doctor recommend for kidney stones?
- **Simpson's paradox:** After accounting for the confounder (stone size) the best choice reverses.

	Treatment A	Treatment B
	78% (273/350)	**83%** (289/350)

Charig et al., BMJ, 1986

Why Observational (Data) Science is still critical

In many cases, we cannot randomize / intervene / A-B test (cf. offline evaluation).

- **Practicality:** Exposure to treatment may be hard to manipulate
  - Ex: Environmental effects (air pollution)
- **Ethical concerns:** Known negative effects
  - Ex: Is suicide contagious?
- **Efficiency:** Experimental science is expensive and takes time
  - Ex: Studying impact on mortality 10 years later
- …

What if I have a ton of data?

Big data to the rescue?

- “Look at how much data I had…”
- ”How could I be wrong? I used 3 billion data points!”
- “This is just noise. All the problems will cancel out…”

- Beware! You need to worry about bias and variance!
- **More data does not help you reduce bias!**
- **Today: Sources of bias, how to model it, & what to do about it**

The Reasonable Uneffectiveness of Big Data

- “The Unreasonable Effectiveness of Data”
  - By Alon Halevy, Peter Norvig, and Fernando Pereira at Google
  - Simple models + Lots of data work very well

- Now consider context of causal inference
  - Measurement error, confounding, and selection bias common threats to causal inference, are independent of sample size
  - When we can’t observe counterfactuals, observing more data will not help us!

We will learn about causality later in the course!

Big Data does not address…

…common threats to causal inference, including:

1. Construct validity
   - E.g. measurement error
2. Internal Validity
   - E.g. confounding
3. External Validity
   - E.g. selection effects

Challenge 1: Construct Validity

- Def: Are you measuring what you think you are measuring?
  - Especially important operationalization of theoretical construct / new “sensor”
    (e.g. social media, linguistic proxy)
- How to demonstrate?
  - Convergent validity: Simultaneous measures of same construct correlate
  - Discriminant validity: Doesn't measure what it shouldn’t

Big Data typically means little control over how anything was measured

Challenge 2: Internal Validity

- Def: Soundness of research design
- What potential selection effects / confounding are there?
  - Is data missing non-randomly?
  - Could measurement be biased across key groups?
  - Does population change across multiple analyses (complicating comparisons)?

Internal Validity (cont.)

- How robust are findings across different choices along the way?
  - How robust are results with respect to inclusion/exclusion of outliers?
- How many hypotheses are being tested?
  - May need to control false discovery rate
- Are distributional / parametric assumptions valid?
  - Consider non-parametric models and bootstrapping

Big Data typically means observational data, convenience samples, and no pre-registration

Challenge 3: External Validity

- Def: Can findings be generalized to other situations and to other people?
- How biased is the study population?
  - Ex: “Internet Explorer users”
  - Ex: “Chrome latest beta users”
  - Ex: “Smartphone owner + health app installed”
  - Convenience samples can be WEIRD, especially motivated, lack key groups of interest, …

Big Data typically means more data, but more of the same!

Summary: Data Science Objectives

1. Formulate a research question
2. Identify a dataset with which to answer the question
3. Design an analysis process (next)
4. Consider construct, internal and external validity
   - Remember that more data doesn’t necessarily help

Your turn…

- Divide into breakout rooms - for ca. 10 minutes
- Discuss: In your project…
  - What does construct, internal, external validity mean?
  - What are threats to validity in your project?
  - What can you do to address them?

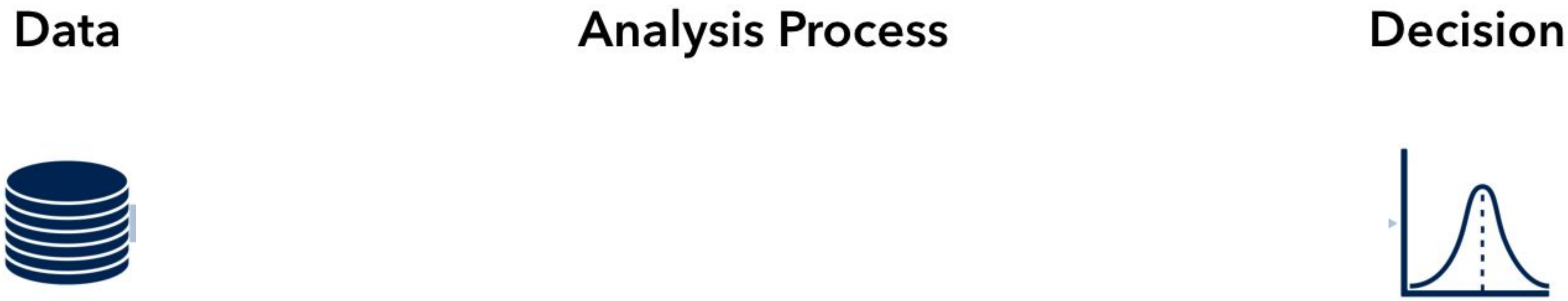
- Each breakout room will present their discussion & peer feedback
  - Help the other groups out with your feedback

- Note: You will do a validity reflection as a group on Oct 27. You will also do a personal validity reflection by next week.
  - It really pays off to take notes of your discussions and feedback 😊

What is the Data Science Process?

Data Science as a Process

- Separate iterative process into a sequence of activities with different points of failure
- **What does it take to get data science right?**



- Framework for your group projects and evaluating data science projects

Process Steps Explained

Define

- *Define the goal and type of the analysis.* Failure modes: Goal of analysis does not match scientific or business need.

Collect

- *Measure / collect data to analyze.* Failure modes: Selection bias (e.g., population mismatch, selective labeling…).

Process Steps Explained (2)

Annotate

- *Augment data with labels or other metadata.* Failure modes: Annotator disagreement; erroneous codes or labels.

Wrangle

- *Clean, filter, summarize, and/or integrate data.* Failure modes: Incorrect filtering, e.g., high-leverage outliers. Incorrect joins with other datasets.

Process Steps Explained (3)

Profile

- *Inspect shape and structure of data.* Failure modes: Overlook data quality issues or violations of distributional assumptions.

Operationalize

- *Define and validate central measures, which may be proxies.* Failure modes: Lack of construct validity (i.e., not measuring what you think you are measuring).

Process Steps Explained (4)

Explore

- *Interactively explore data and variable relationships.* Failure mode: Confirmation bias; unclean split between train/test data.

Model

- *Define and fit models of relationships in data.* Failure modes: Lack of internal validity. Failure to identify effect, e.g., due to confounding or violated assumptions.

Process Steps Explained (5)

Evaluate

- *Measure explanatory power or predictive accuracy of model using appropriate statistical techniques.* Failure modes: p-hacking, overuse of test set data.

Report

- *Report results and potential generalizations.* Failure modes: Misinterpretation (e.g., generalization, uncertainty), miscommunication via errors or omissions.

Process Steps Explained (6)

Deploy

- *Deploy model or enact decision.* Failure modes: Distribution drift, e.g., changes in data pipeline upstream, changing assumptions, adversarial input.

What does this mean for you?

- **Plan** your own project along these stages
- When learning about other projects **pay attention to potential pitfalls** across all phases
- When working on your own project, **explicitly address each step and failure modes**

Teaser for next week: Example Study

Brief Overview of Study Context & Main Results

How Physically Active Are We?

Physical activity is extremely important for health [Lee et al., 2012]. But we do not know how much physical activity people get!

According to WHO:

- 5-54% of Germans don't get enough activity
- No data for Switzerland and Israel

Health research limitations today:

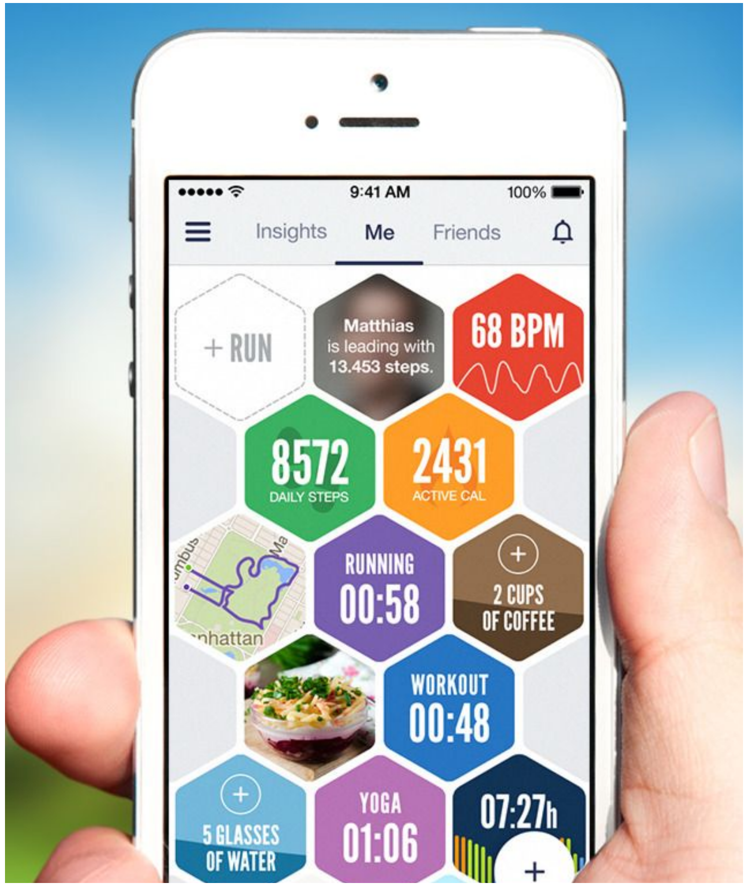
- High cost, short-term, limited scale
- Biases from self-reporting

Wearable and Mobile Devices

69% adults own smartphones in developed countries
46% in developing economies (rapidly growing)

Wearable and mobile devices generate massive digital traces of real-world behavior and health

Activity Tracking



azumio

Tracking actions

- Steps (automatic)
- Runs
- Walks
- Workouts
- Biking
- Weight
- Heart rate
- Food
- Drinks
- And many, many others

Dataset Statistics

- Data from 2011
- 717,527 anonymized users
- Users from 111 countries
- 68 million days of steps tracking
  - **100 billion** data points (2TB), Minute-by-minute
- Focus on 46 countries with >=1,000 users
  - 32 high-income, 14 middle-income countries

Today: 6M users, 160M days of activity, 800M actions tracked

Data in Context

- Our data: 68 million days of activity from over 700,000 individuals in 111 countries

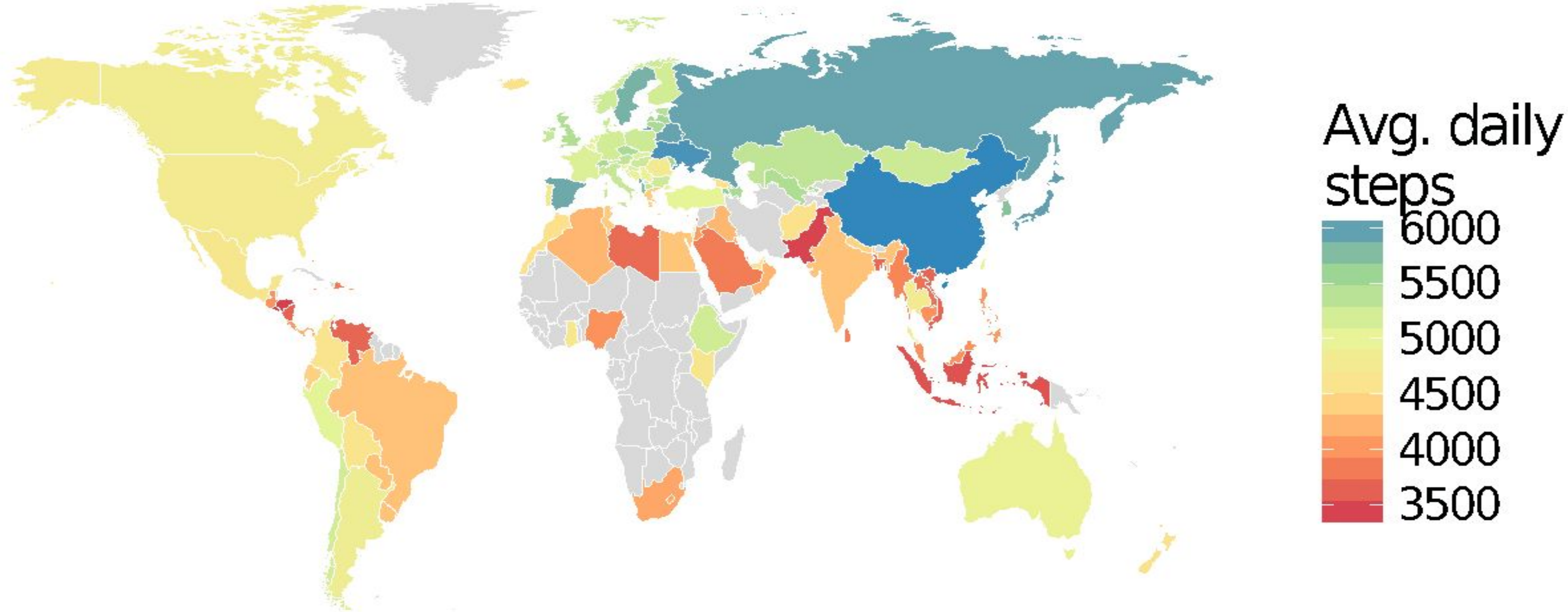
1400x larger than largest existing gold-standard datasets:

- NHANES [Troiano et al., MSSE 2008]
- IPEN [Van Dyck et al., Int. J. Obes. 2015]

Size of NHANES
relative to
full slide (Azumio)

Population data available at:
http://activityinequality.stanford.edu/

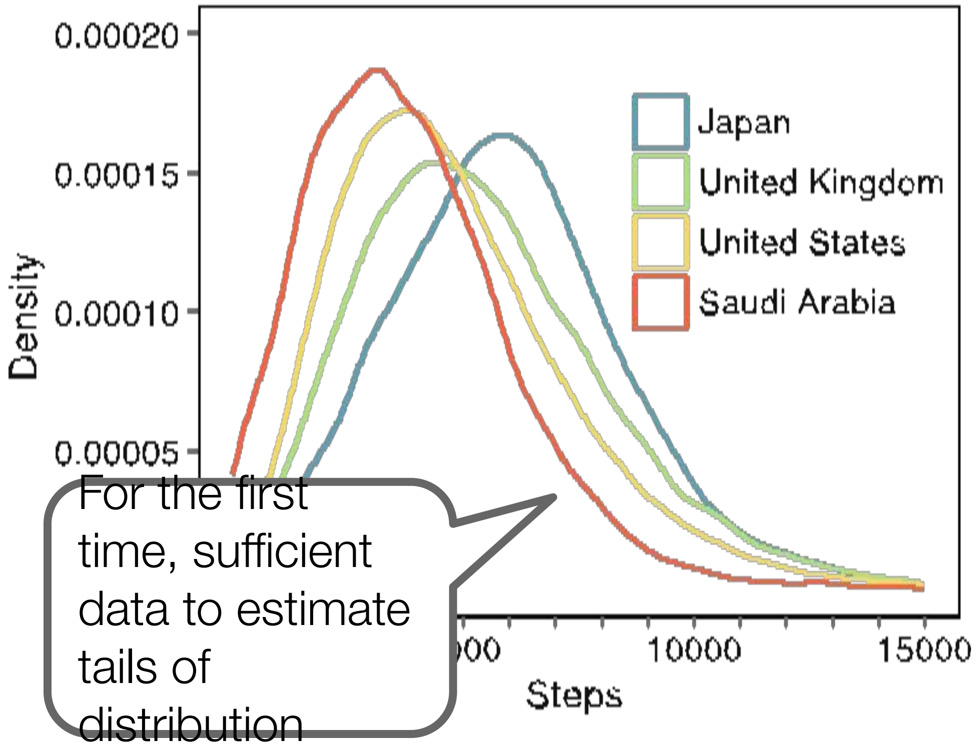
Worldwide Activity



But, how is activity distributed within the population?

Result 1: Inequality of Physical Activity

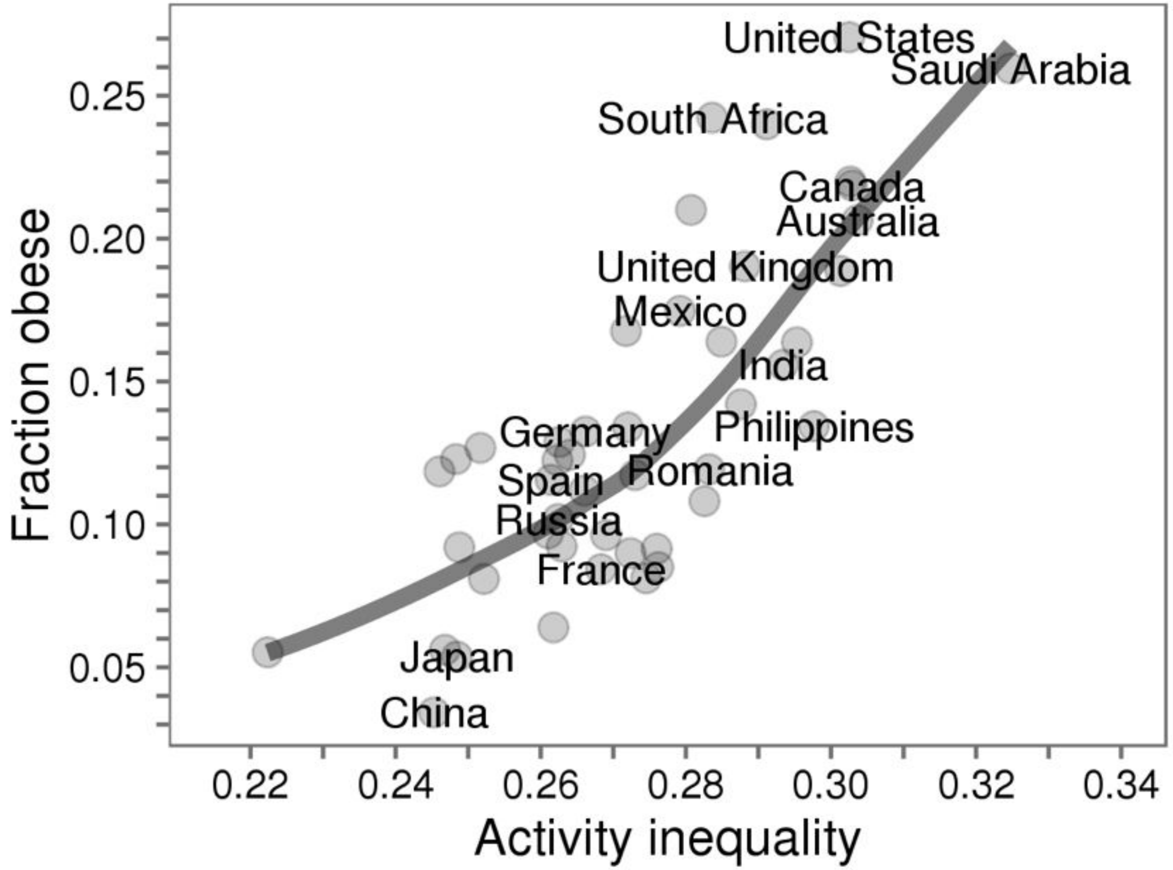
Difference in means



- **How (un)evenly is activity distributed?**
- Gini index of the activity distribution:
  - Activity rich vs. activity poor people

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j}$$

Result 2: Activity Inequality Predicts Obesity

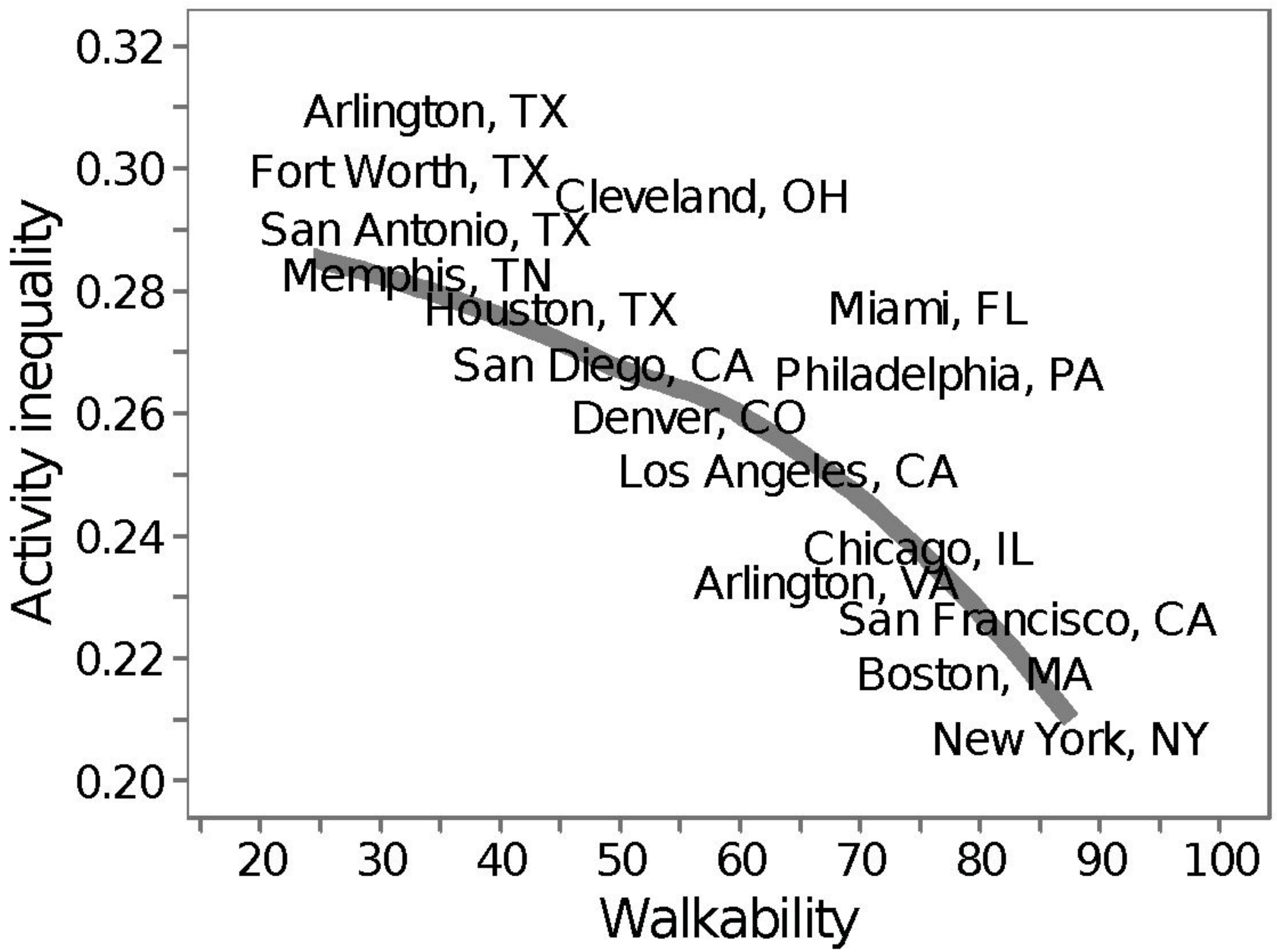


Tails/extremes matter more than the mean

R^2=0.64 (vs. 0.47 for avg. activity)

Massive digital traces uniquely enable studying tails!

Result 3: Walkability Reduces Inequality



Open Q&A

When is prediction / big data not enough?

Prediction is everywhere!

- Recommender Systems
- Social Networks
- …

- We have increasing amounts of data and highly accurate predictions! Why do we need causal inference?

1) Do prediction models guide decision-making?

From data to prediction

Can we predict a user's future activity based on exposure to their social feed?

Use the social feed to predict a user's future activity.

- Future Activity -> *f*(items in social feed) +

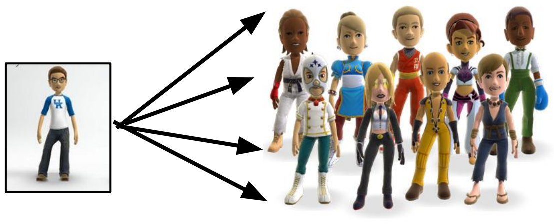
Highly predictive model.

Does it mean that feeds are influencing us significantly?

From prediction to decision-making

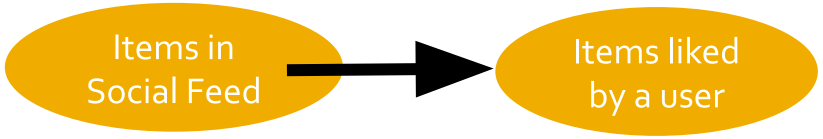
Would changing what people see in the feed affect what a user likes?

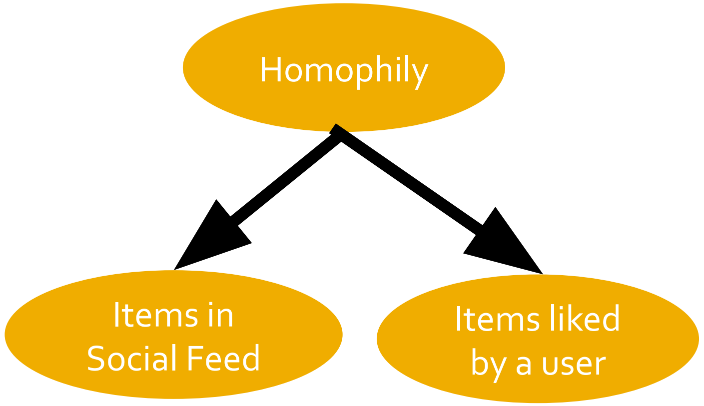
Maybe, maybe not (!)



Predictability due to
feed influence

Predictability due to
homophily





Friends' activity can predict a person's activity with high accuracy.
But that tells us *nothing* about the effect of the social feed.

A Motivating Example

- Which treatment should a doctor recommend for kidney stones?
- **Simpson's paradox:** After accounting for the confounder (stone size) the best choice reverses.
- Critical for decision making

Treatment A	Treatment B

Charig et al., BMJ 1986

Why Observational (Data) Science is still critical

In many cases, we cannot randomize / intervene / A-B test (cf. offline evaluation).

- **Practicality:** Exposure to treatment may be hard to manipulate
  - Ex: Environmental effects (air pollution)
- **Ethical concerns:** Known negative effe
  - Ex: Is suicide contagious?
- **Efficiency:** Experimental science is exp
  takes time
  - Ex: Studying impact on mortality 10 years later
- …

Big Data does not address…

…common threats to causal inference, including:

1. Construct validity
   - E.g. measurement error
2. Internal Validity
   - E.g. confounding
3. External Validity
   - E.g. selection effects

Challenge 1: Construct Validity

- Def: Are you measuring what you think you are measuring?
  - Especially important operationalization of theoretical construct / new “sensor”
    (e.g. social media, linguistic proxy)
- How to demonstrate?
  - Convergent validity: Simultaneous measures of same construct correlate
  - Discriminant validity: Doesn't measure what it shouldn’t

Big Data typically means little control over how anything was measured

Challenge 2: Internal Validity

- Def: Soundness of research design
- What potential selection effects / confounding are there?
  - Is data missing non-randomly?
  - Could measurement be biased across key groups?
  - Does population change across multiple analyses (complicating comparisons)?

Internal Validity (cont.)

- How robust are findings across different choices along the way?
  - How robust are results with respect to inclusion/exclusion of outliers?
- How many hypotheses are being tested?
  - May need to control false discovery rate
- Are distributional / parametric assumptions valid?
  - Consider non-parametric models and bootstrapping

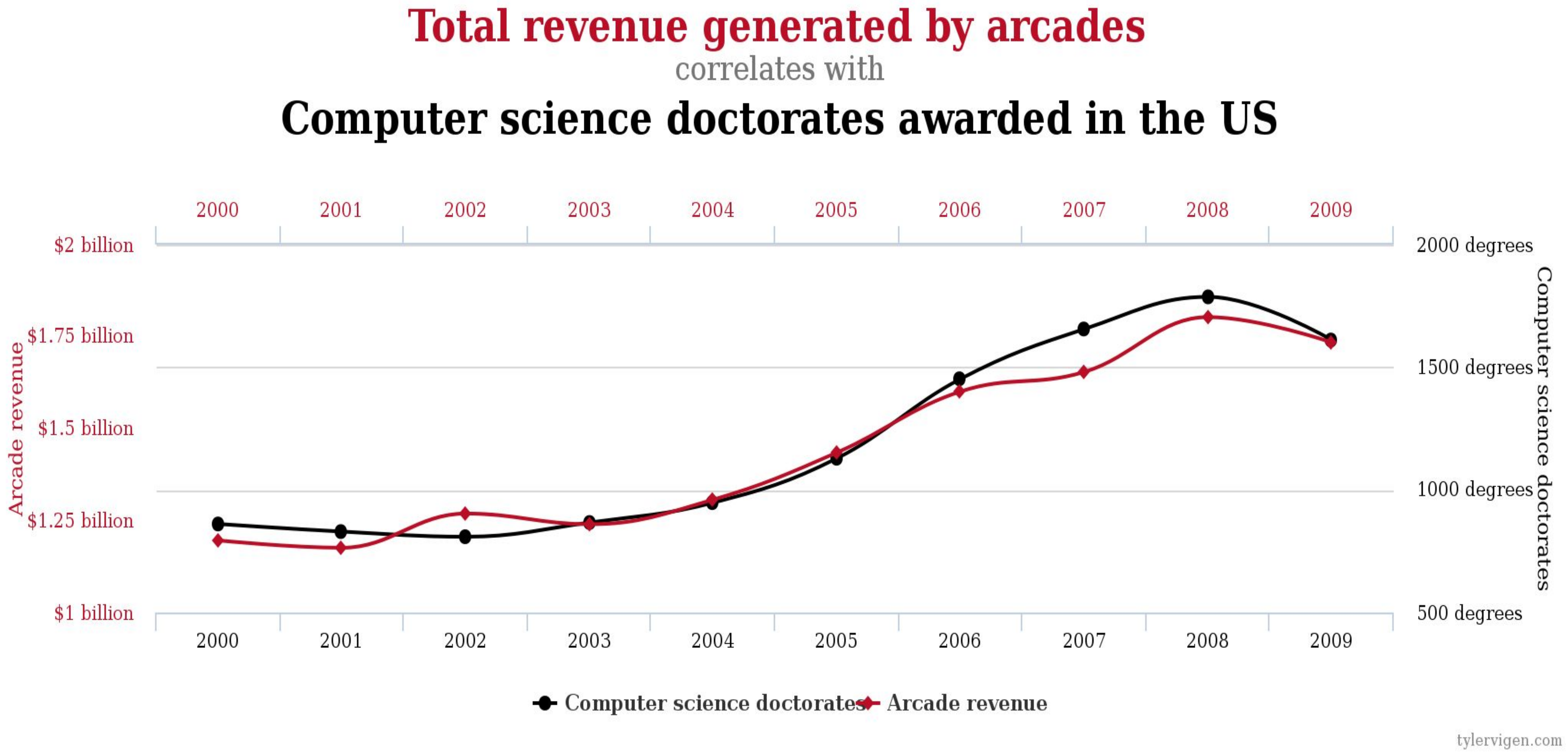
Big Data typically means observational data, convenience sample, and no pre-registration

Challenge 3: External Validity

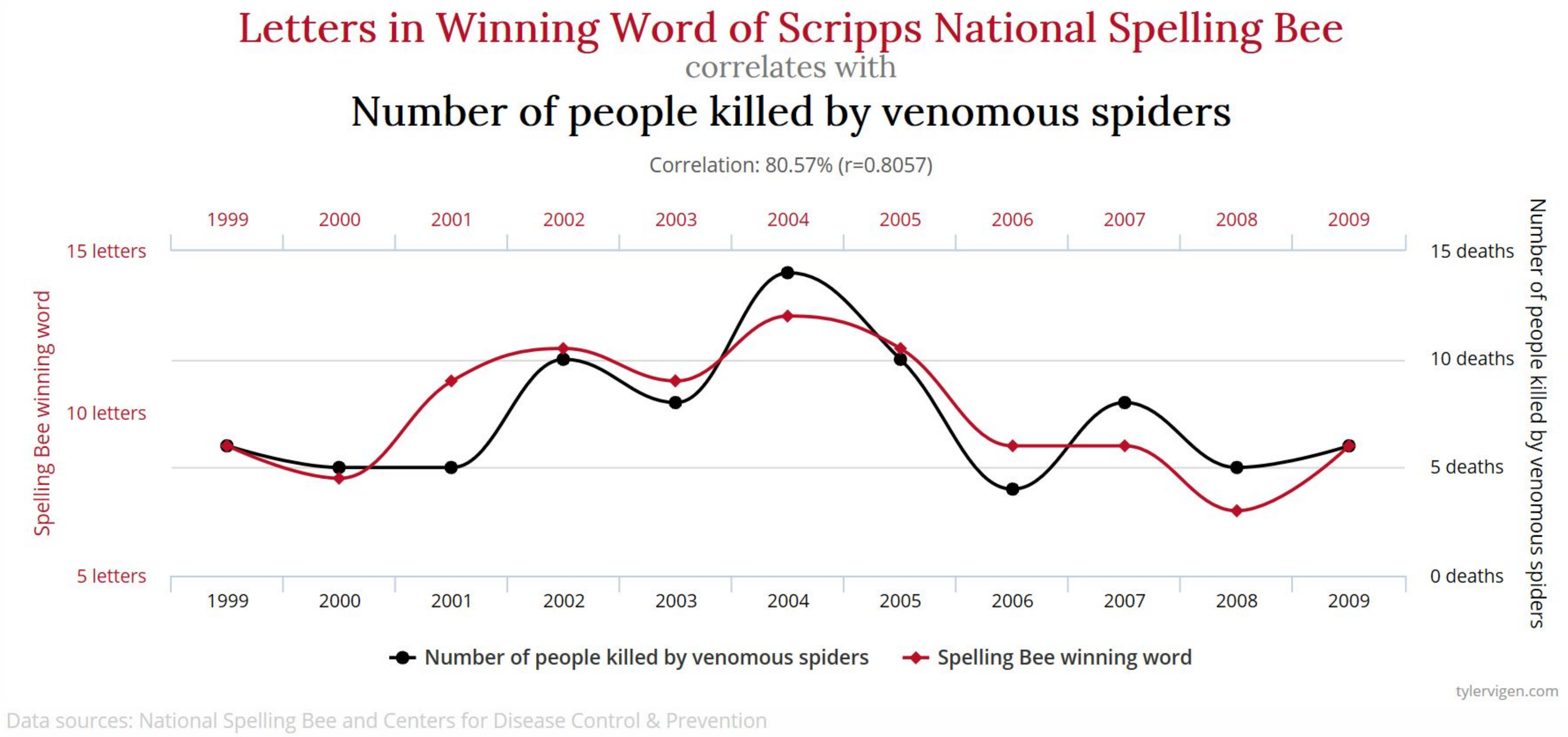
- Def: Can findings be generalized to other situations and to other people?
- How biased is the study population?
  - Ex: “Internet Explorer users”
  - Ex: “Chrome latest beta users”
  - Ex: “Smartphone owner + health app installed”
  - Convenience samples can be WEIRD, especially motivated, lack key groups of interest, …

Big Data typically means more data, but more of the same!

2) Will the predictions be robust tomorrow, or in new contexts?



http://www.tylervigen.com/spurious-correlations

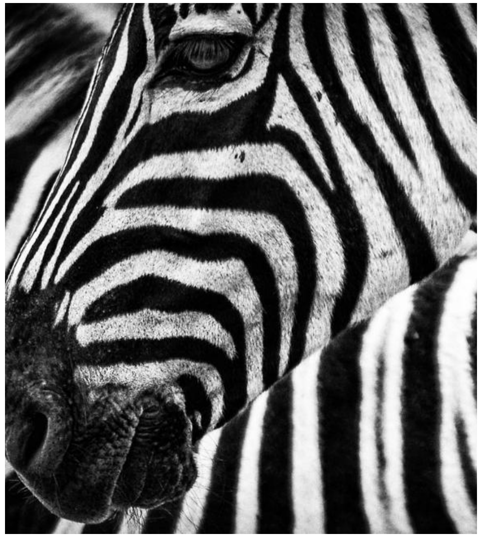

Letters in Winning Word of Scripps National Spelling Bee
correlates with
Number of people killed by venomous spiders
Correlation: 80.57% (r=0.8057)
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
15 letters
10 letters
5 letters
Spelling Bee winning word
15 deaths
10 deaths
5 deaths
0 deaths
Number of people killed by venomous spiders
Number of people killed by venomous spiders
Spelling Bee winning word
tylervigen.com
Data sources: National Spelling Bee and Centers for Disease Control & Prevention

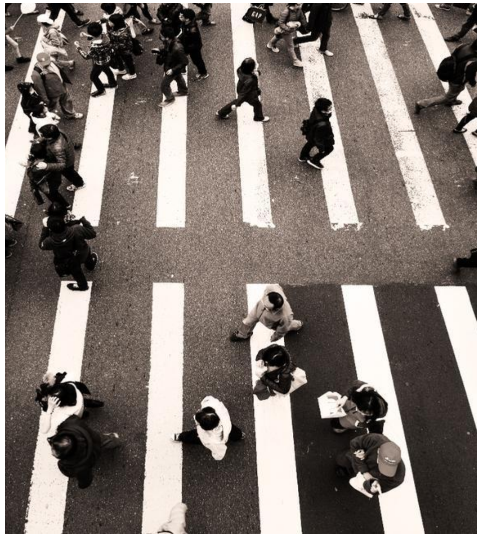
3) What if the prediction accuracy is really high?

Interventions change the environment

- Train/test from same distribution in supervised learning
- No such guarantee in real life!
- Problematic: Acting on a prediction changes distribution!
  - Incl. critical domains: healthcare or adversarial scenarios.

- Connections to covariate shift, domain adaptation [Mansour et al. 2009, Ben-David 2007].





Your turn… [10min in group, 10 min report]

- Divide into breakout rooms
- For your project…
  - Go through each of the steps for your group project
  - What activities will fall into each step?
  - How much time / risk do you anticipate in each step?
  - What questions & challenges come up?

- TODO think pair share?
- TODO highlight the deliverable two weeks from today
- TODO how does this interact with next week’s activities?

Presentations of project plans

- TODO group tentative Project plan (1-2 paragraphs, where data from, RQ, risks, infolab intro, make 6ish slide template for them)
  - 6ish min presentation + 8 min feedback + 1 min switch – 6 groups= 90 min!
- Very few slides
- How has your project idea evolved?
- Any key insights from activities today? (live)
- Allow me to give feedback on risks etc

- After this lecture today the group assignments are fixed. Ask them if anyone really wants to switch teams? Or tell them to check in with each other through Canvas?

Your turn [10 min in group, 5 min report]
students basically should have answered these in their proposal presentations last week

- Stage 1: Define
  - What is your exact research question?
  - Why does it matter?
  - What is unique about your planned approach?
  - Try to keep it short and very concrete

- Stage 2: Collect
  - Where do you get your data from?
    - You should have your dataset secure and accessible at this point. Otherwise this is a major risk you need to address immediately.
  - Does your dataset allow you to answer your research question?
  - What concerns should you consider (e.g. selection bias, population mismatch, known data quality issues)?

- Stage 3: Annotate
  - Could you combine your dataset with another dataset to create a unique and exciting opportunity?
  - Do you need to collect/annotate any additional data? How will you do this?
  - How will you ensure the data quality?