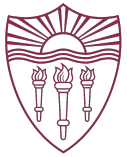


DSCI-531: FAIRNESS IN ARTIFICIAL INTELLIGENCE

BIAS IN DATA

Kristina Lerman

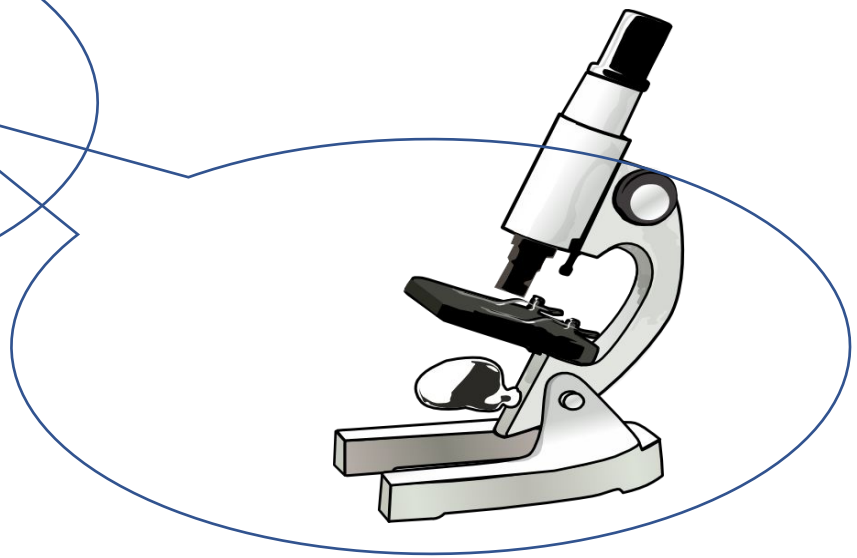
Spring 2025

Social data is heterogeneous

generated by individuals with different characteristics and behaviors

Social data is non-representative

individuals self-select on different platforms or for different behaviors



Social data is sampled

only a (possibly) biased subset of data is observed

Biases in data pose threats to the validity of learned models

Threats to Prediction

- Models are non-generalizable and non-reproducible
- Poor performance on held-out data

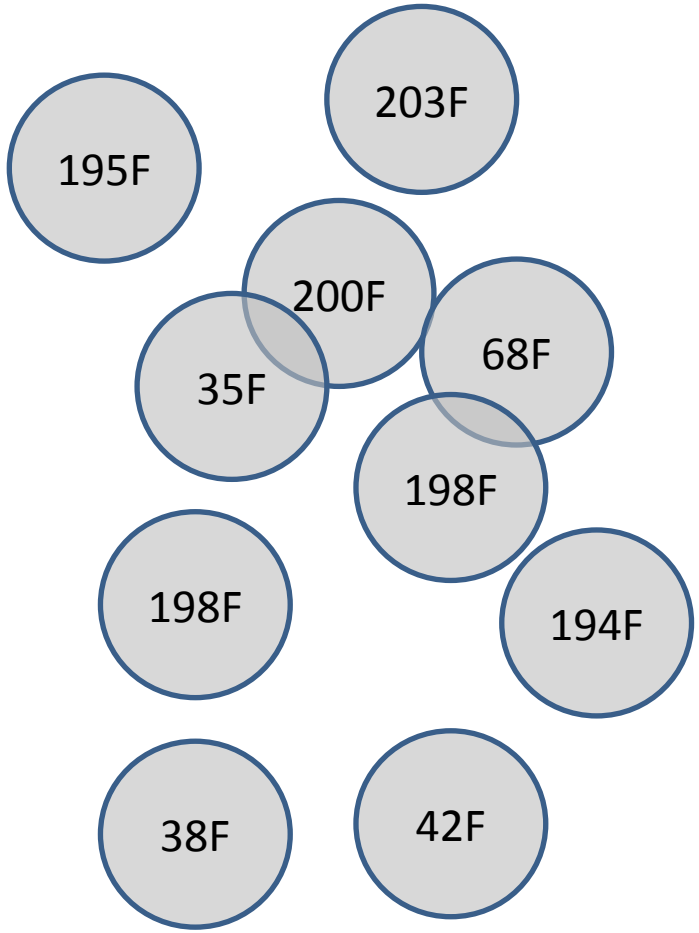
Threats to Explanation

- Ecological fallacy
- Misleading or wrong inferences about individuals
- Impact on interventions

Threats to Fairness

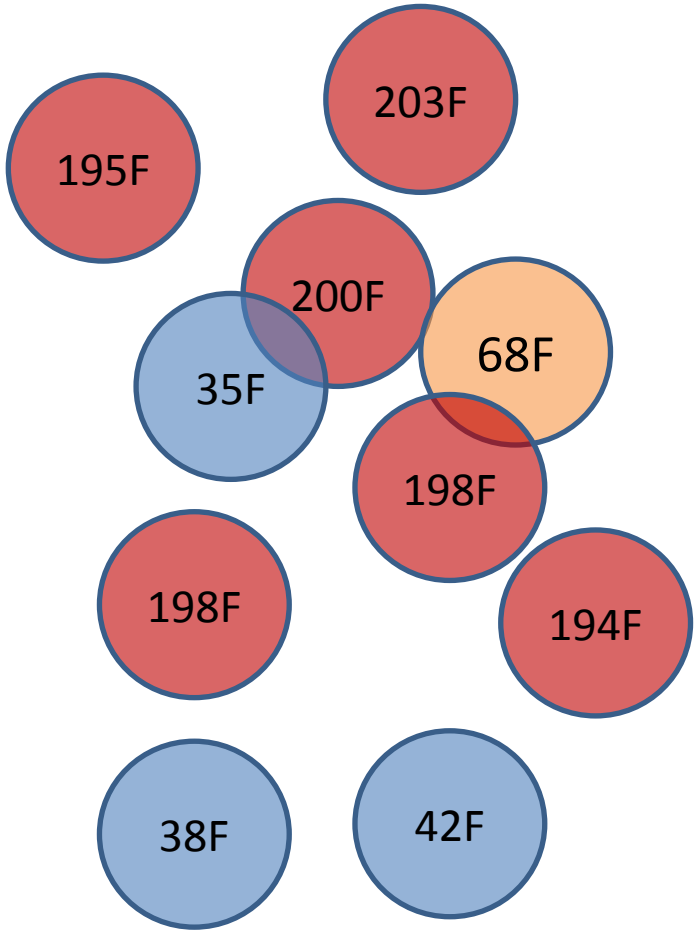
- Models learned on biased data may entrench and amplify discrimination

Thought Experiment: At what temperature do people like their coffee?

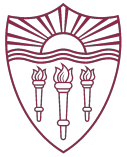


- Measure the temperature of coffee of people walking by
  - 137.1F after 10 measurements
- Ecological fallacy
  - A failure in reasoning that arises when an inference is made **about an individual** based on **aggregate** data for a group.

Thought Experiment: At what temperature do people like their coffee?



- Measure the temperature of coffee of people walking by
  - 137.1F after 10 measurements
- Ecological fallacy
- Population is heterogeneous, composed of different groups
  - Subgroups have different features
    - Hispanics vs Whites
  - Subgroups represent cohorts of different age
    - Veterans vs novices
  - Different waves of data collection
    - Twitter sample in 2011 vs 2016



Simpson's paradox

Subgroups with different behavior & population data

Selection, sampling bias

Subgroups not equally represented

Berkson's paradox

Selection induces a trend where there is none

Sources of bias in heterogeneous data

https://catalogofbias.org/biases/

Survivor bias

Subgroup dropout induces population differences

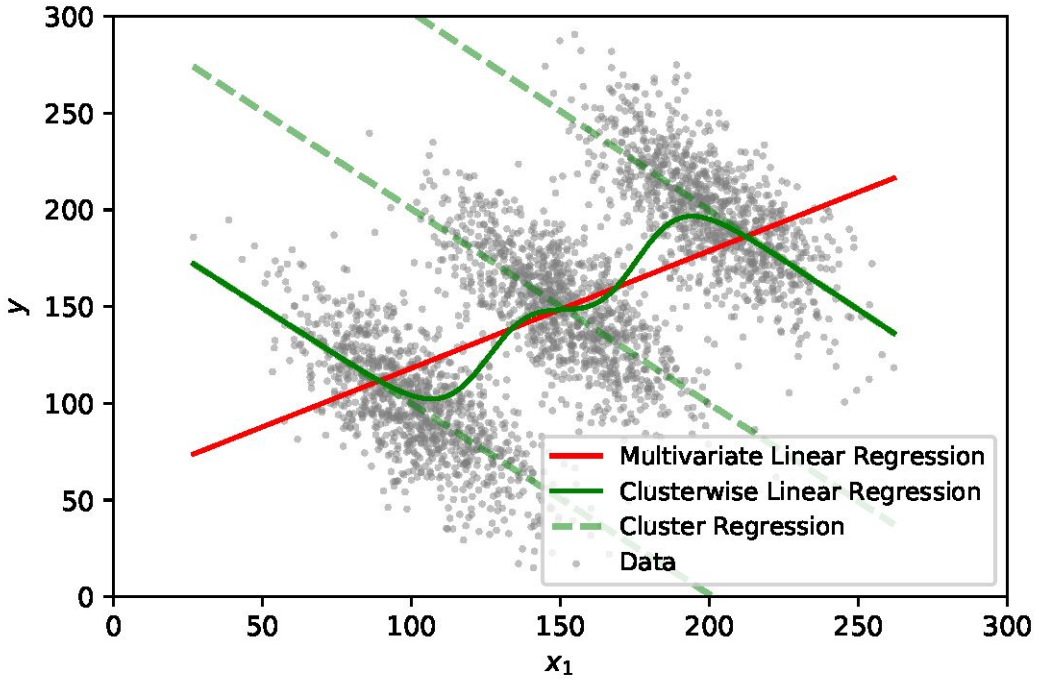
Aggregation bias

Different results and different temporal/spatial resolutions

Longitudinal fallacy

Different ages of cohorts distort cross-sectional analysis

Simpson's paradox



A **trend** appears in different sub-groups of data

but

disappears or reverses when these sub-groups are combined.*

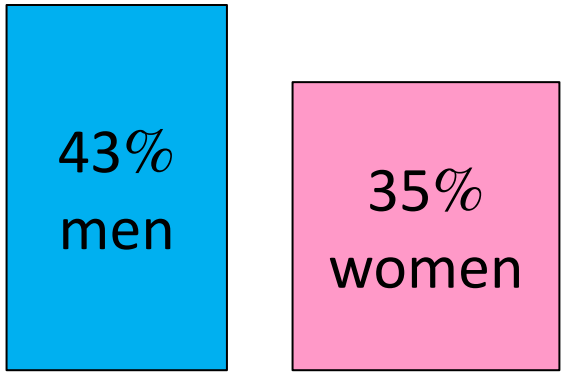
* Simpson (1951). "The Interpretation of Interaction in Contingency Tables" *JRSS*

Sex Bias in Graduate Admissions: Data from Berkeley

Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation.

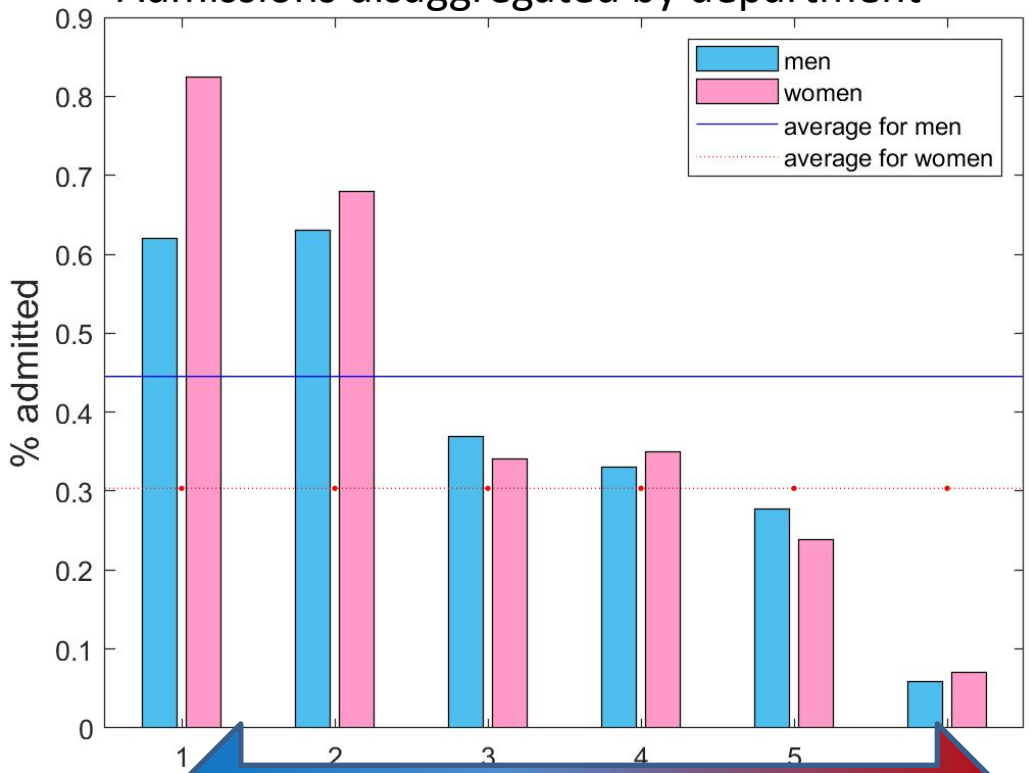
P. J. Bickel, E. A. Hammel, J. W. O'Connell

Source: Science, Vol. 187, No. 4175 (1975), pp. 398-404



Percent of applicants admitted for graduate study

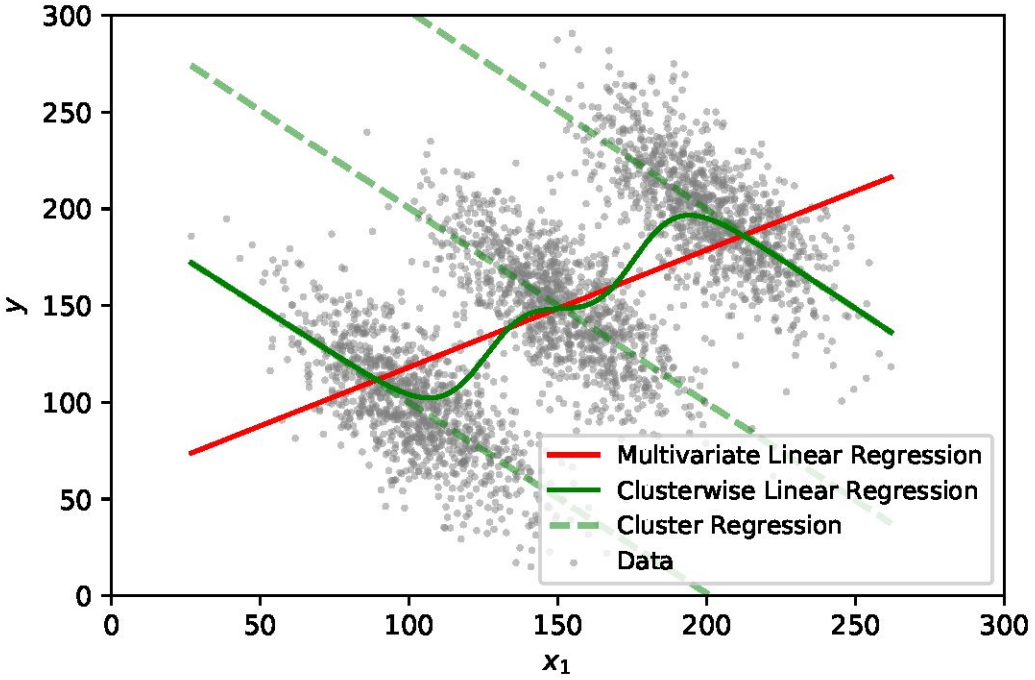
Admissions disaggregated by department



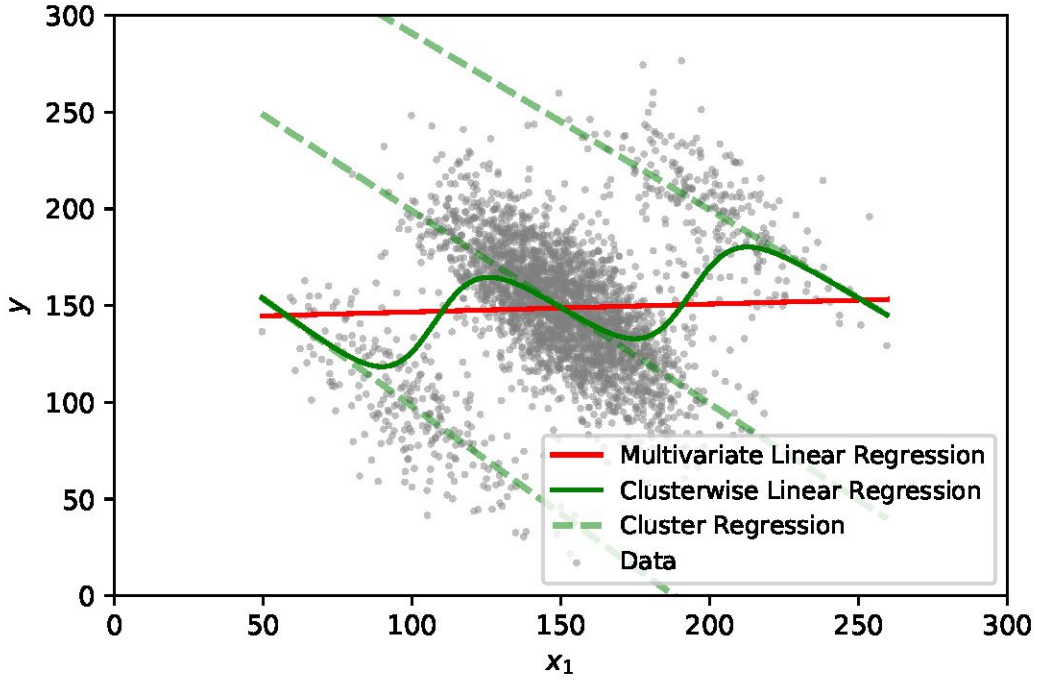
Different population of applicants: more women apply to selective depts

Sampling bias

Subgroups uniformly represented in population



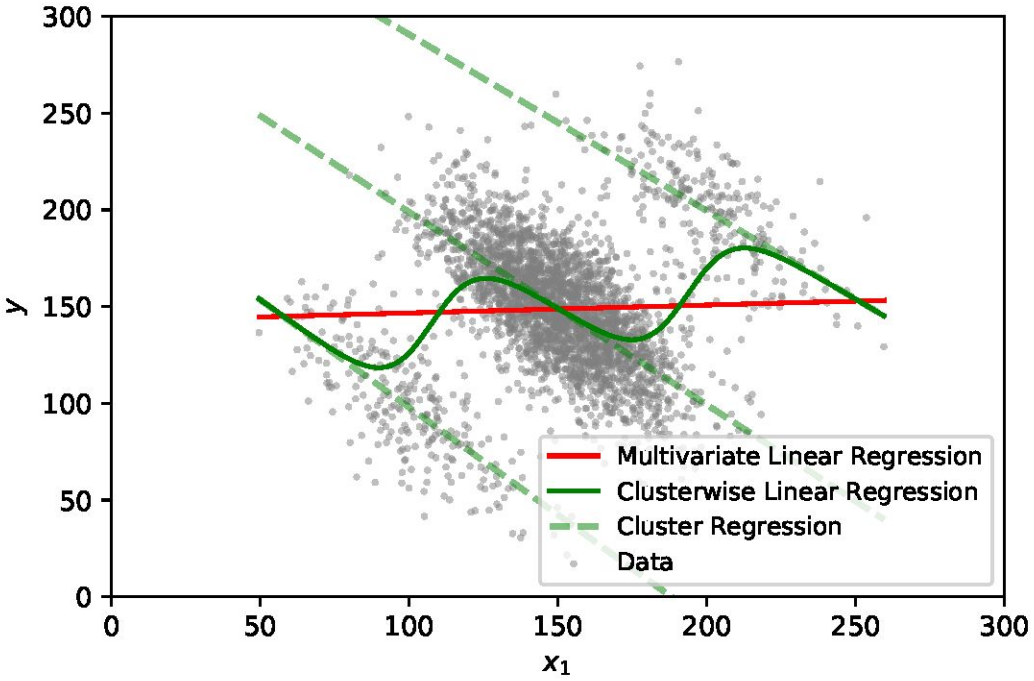
Subgroups overrepresented in the population



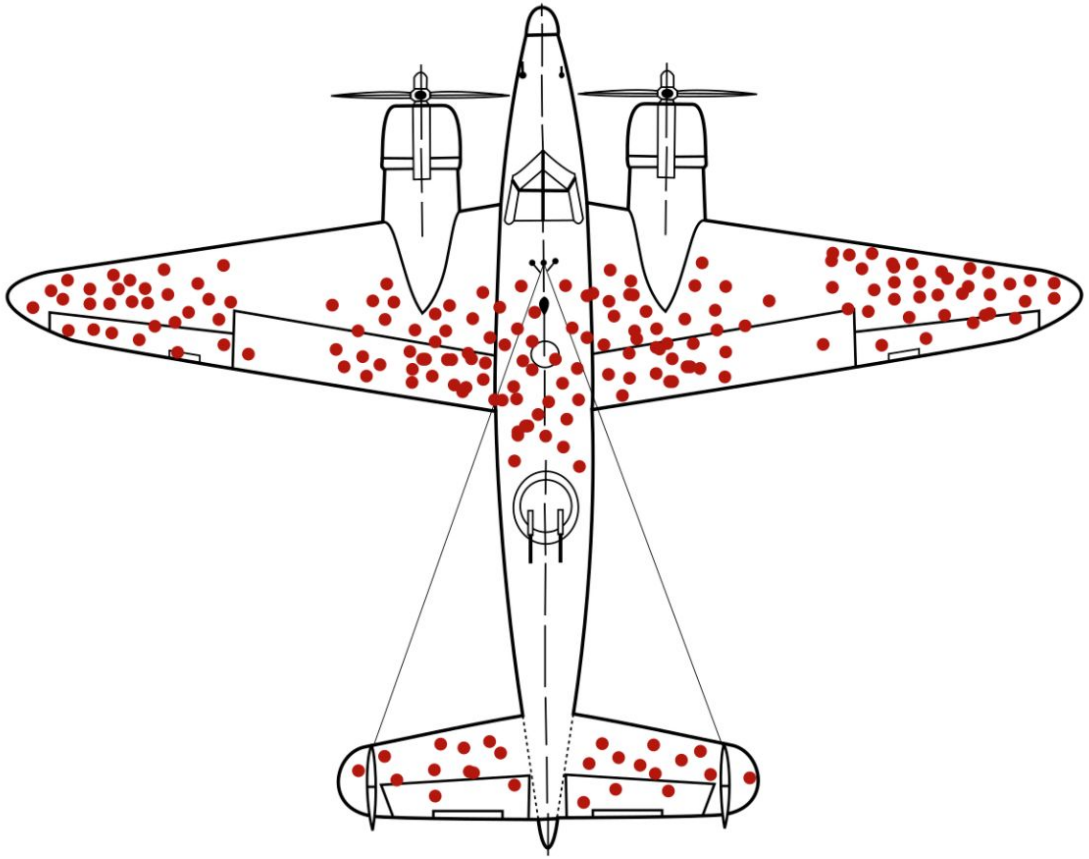
Selection bias

- **Selection bias –** data sample is not representative of the population, with groups or data points **preferentially included or excluded**. This leads to **systematic errors** in conclusions drawn from the data
- **Self-Selection bias** – individuals **opt into** a study based on their personal characteristics, leading to non-random participation.

Subgroups overrepresented in the population



Survivor bias



- WWII fighter planes coming in from battle had a distinctive pattern of damage from enemy guns
- Where should you reinforce the plane to ensure it would not be shot down?

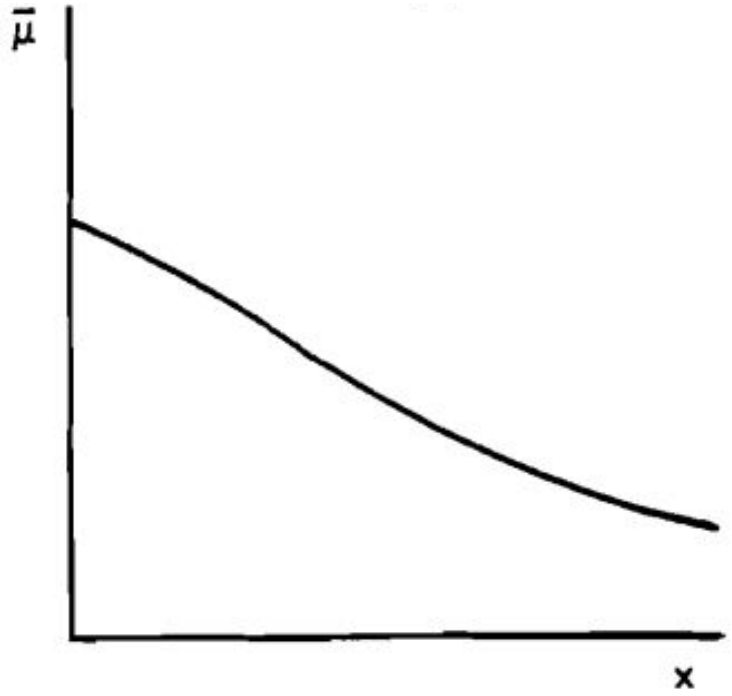
Survivor bias

[Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39(3):176-185.]

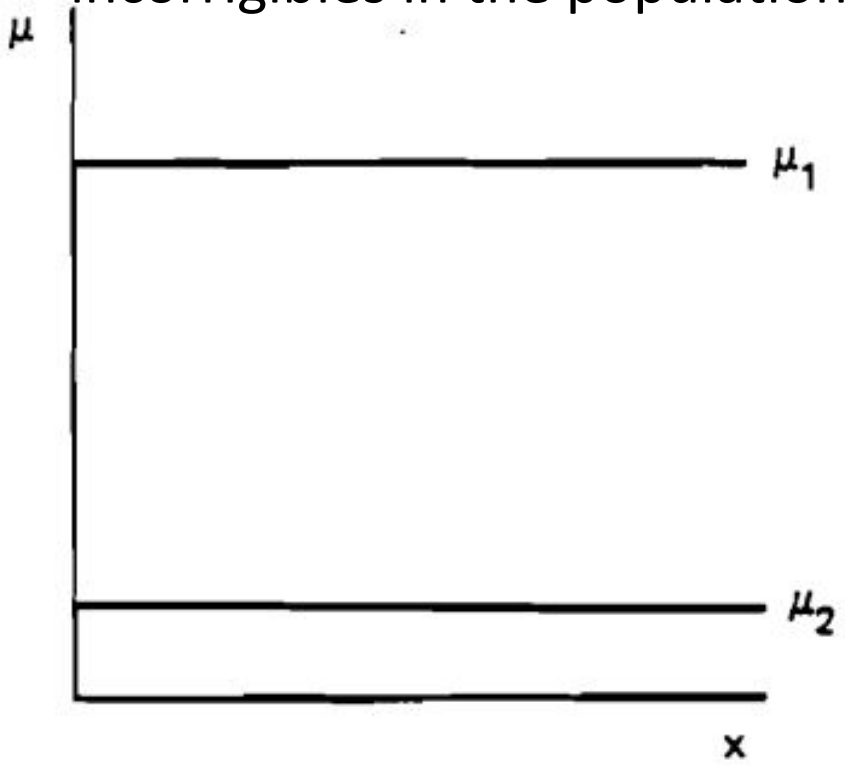
- Mortality rate among aging cohort
  - Age: time since release from prison
  - Death: recidivism (commit another crime)
  - Cohort mortality rate is the weighted sum of the subgroup mortality rates, where the weights are the relative size of the subgroups
- Cohort mortality rate does not tell you about the behavior of subgroups (or individuals)

Survivor bias

Recidivism rate of convicts released from prison declines with age

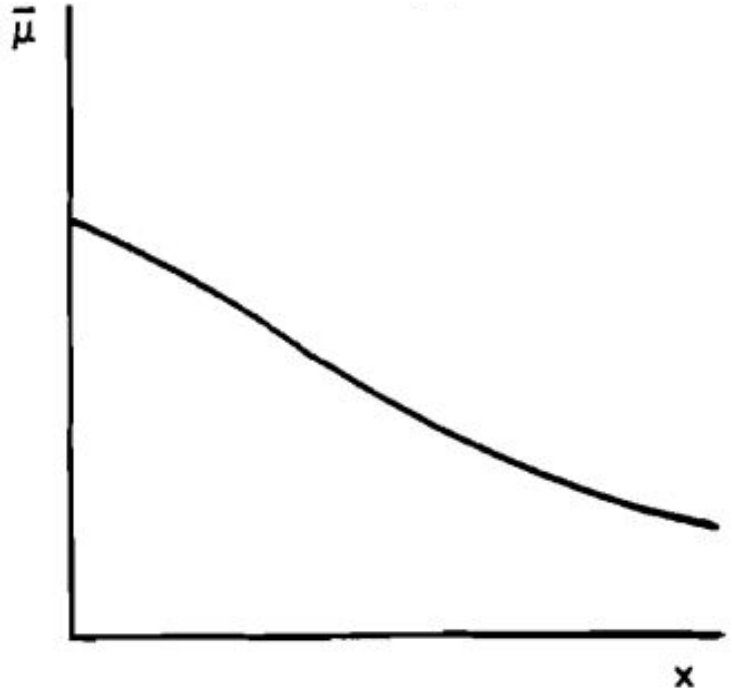


In reality, two subgroups: incorrigibles and reformed. Over time, fewer incorrigibles in the population

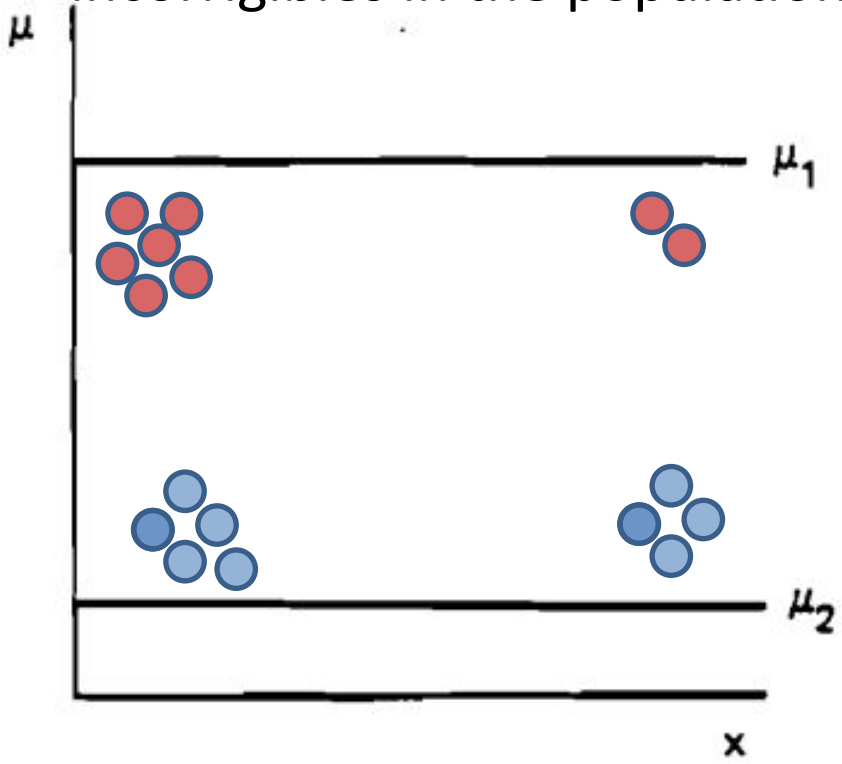


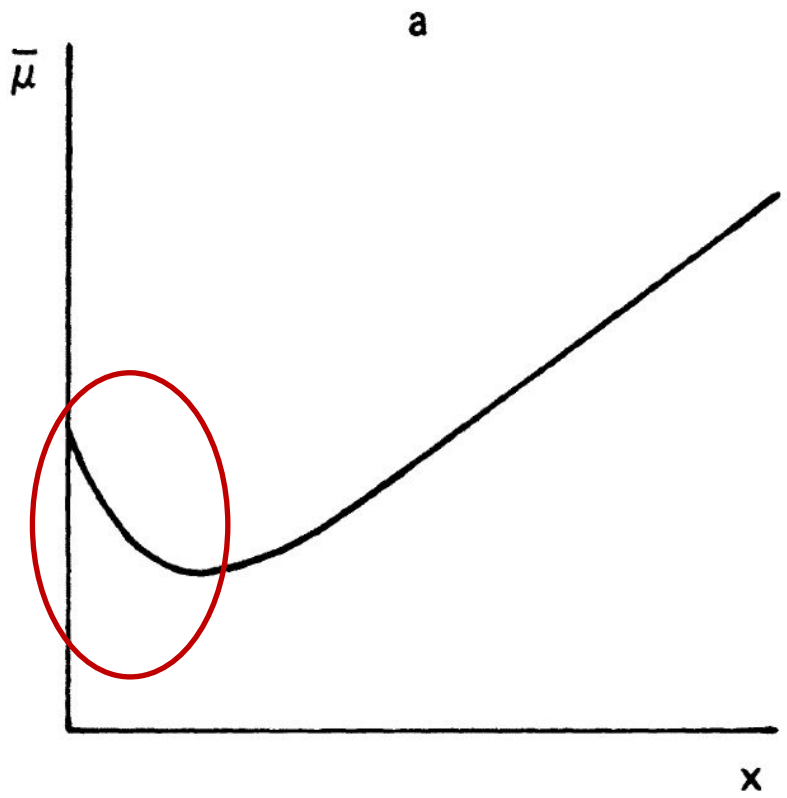
Survivor bias

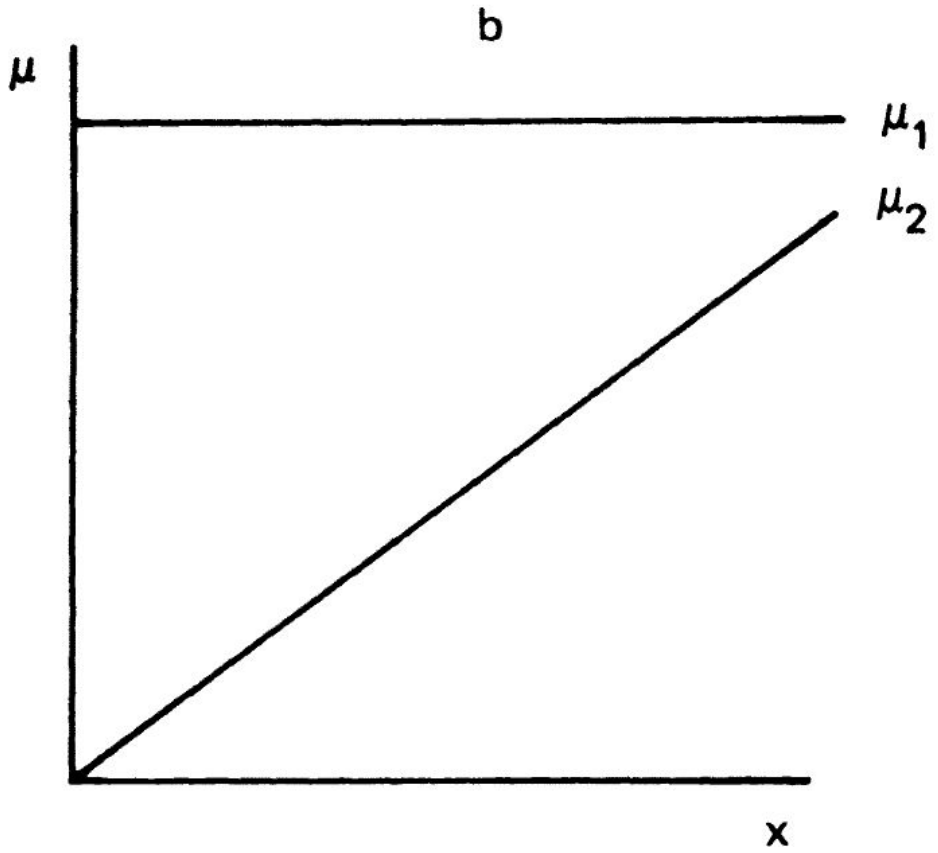
Recidivism rate of convicts released from prison declines with age



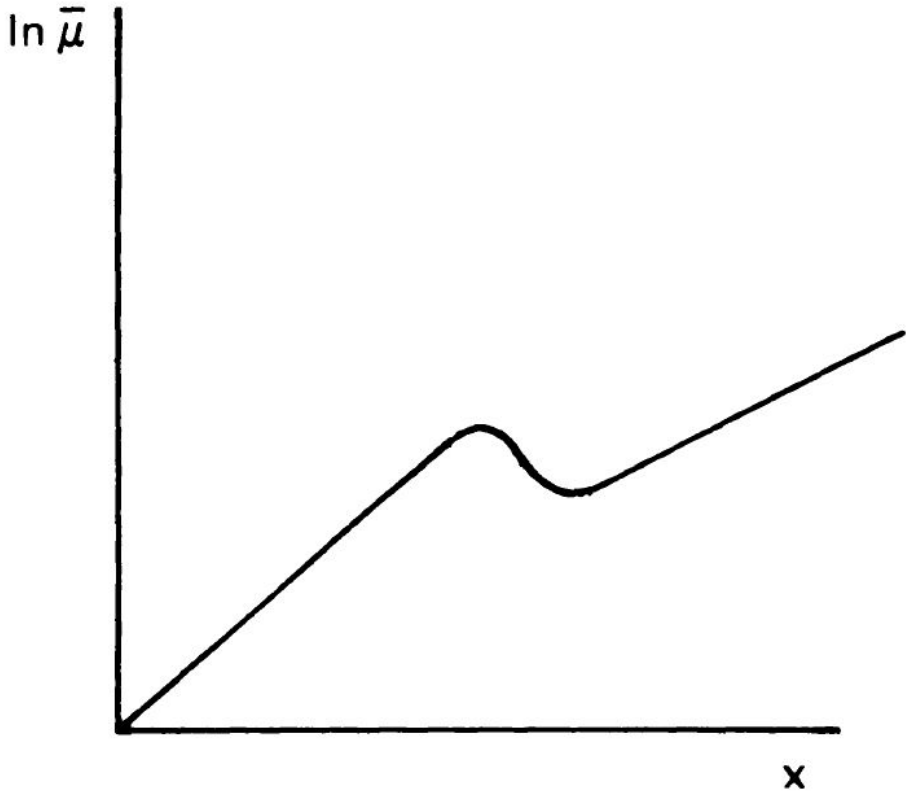
In reality, two subgroups: incorrigibles and reformed. Over time, fewer incorrigibles in the population

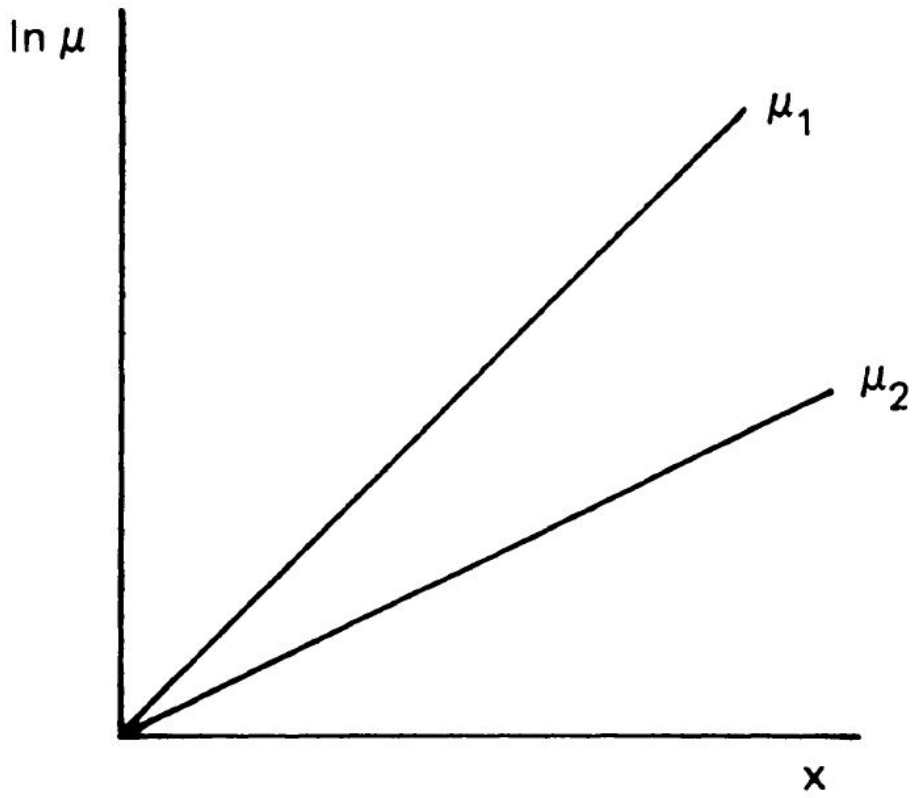




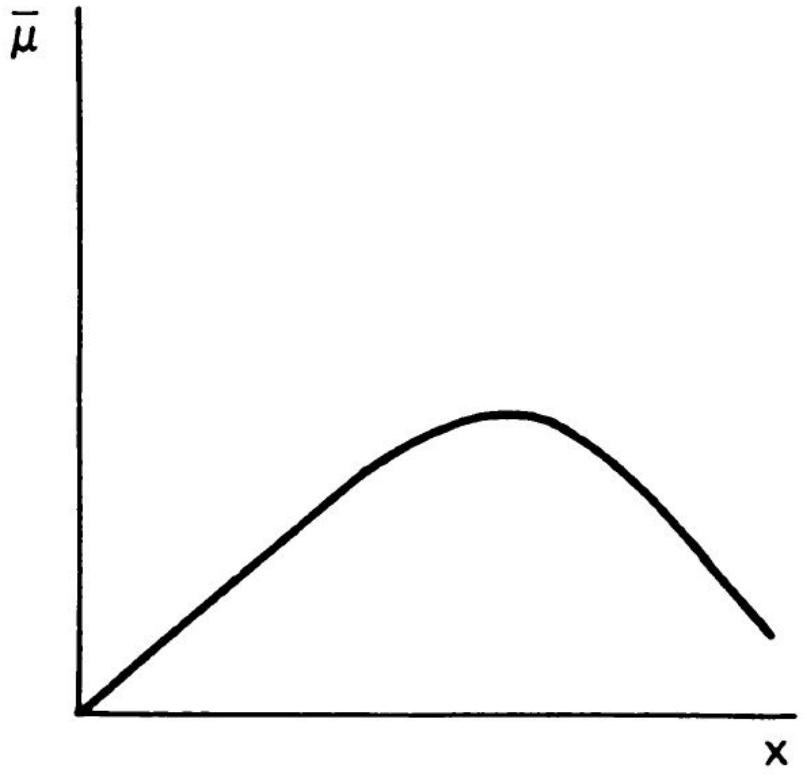


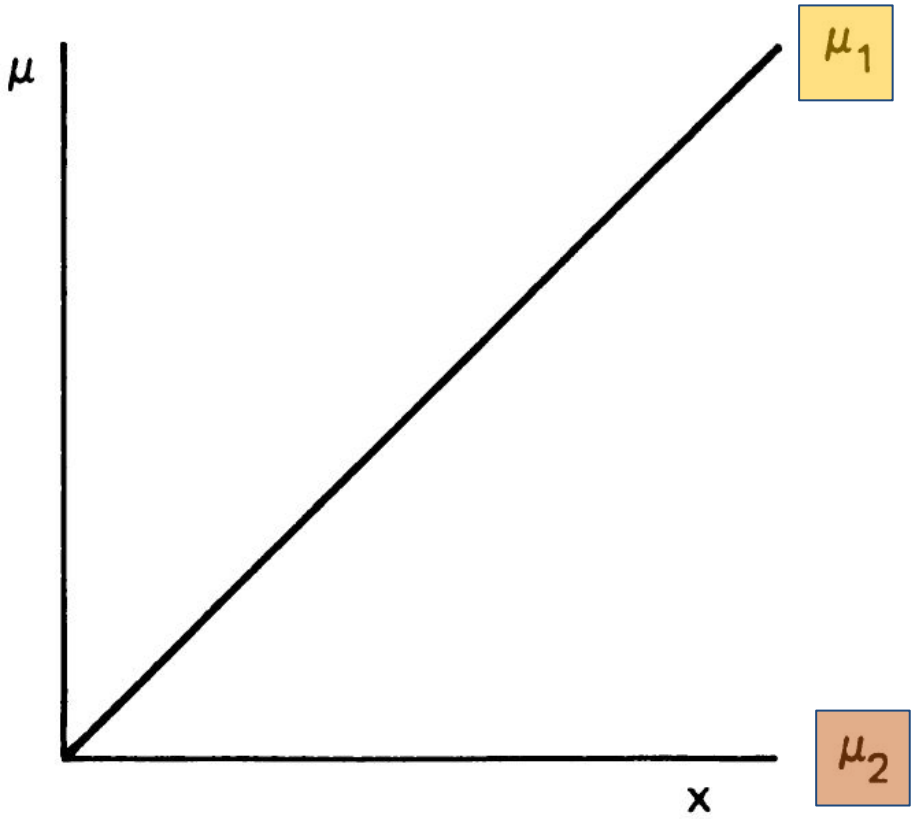
- Does this cohort curve imply that the failure rate for a specific device decreases during the infant mortality phase ?
- The high initial rate of breakdown could be due to a group of lemons (group 1).





- The sudden decline in the observed hazard rate is produced by the rapid extinction of the frailer subcohort ($\mu 1$).
- Then, due to the exponential increase in mortality, the death rates become sufficiently large that within a few years almost all of the frailer subcohort dies

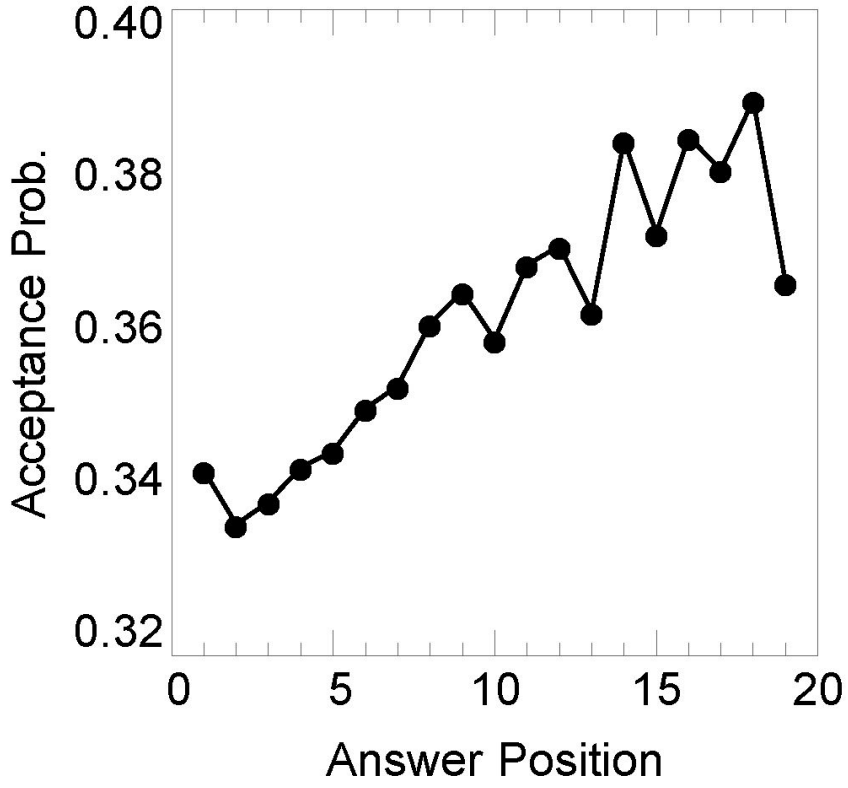




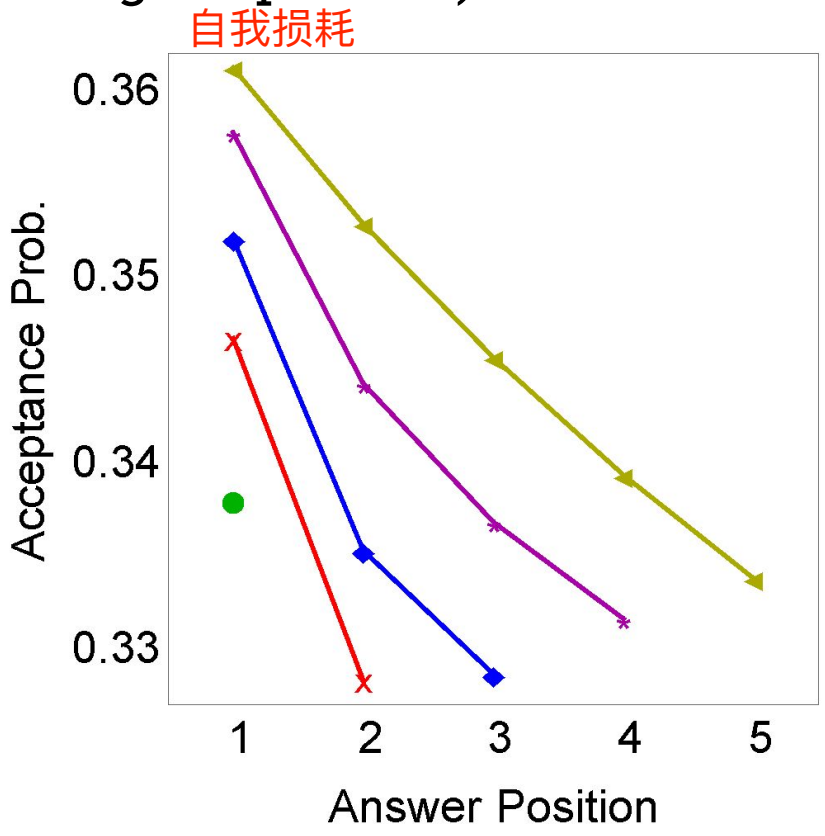
- Observed mortality for the entire population may rise and then fall
- E.g., divorce rates follow this pattern, but this does not imply that marriage is more likely to fail after the first few years. In reality, for one group marriage strengthens with duration, and for the other, it weakens

Survivor bias on stackoverflow

Every subsequent answer written by a user appears to be better (accepted as best answer) …



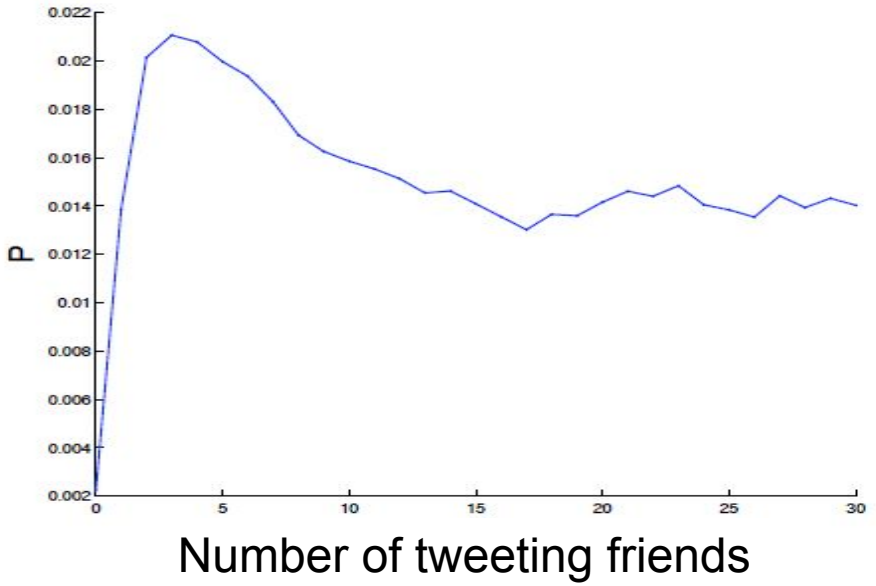
… when disaggregated by session length, every subsequent answer is worse (*cf* "ego depletion")



Alpourteifard et al. (2018) "Using Simpson's paradox to discover interesting behavioral patterns in data" in *ICWSM* 2018

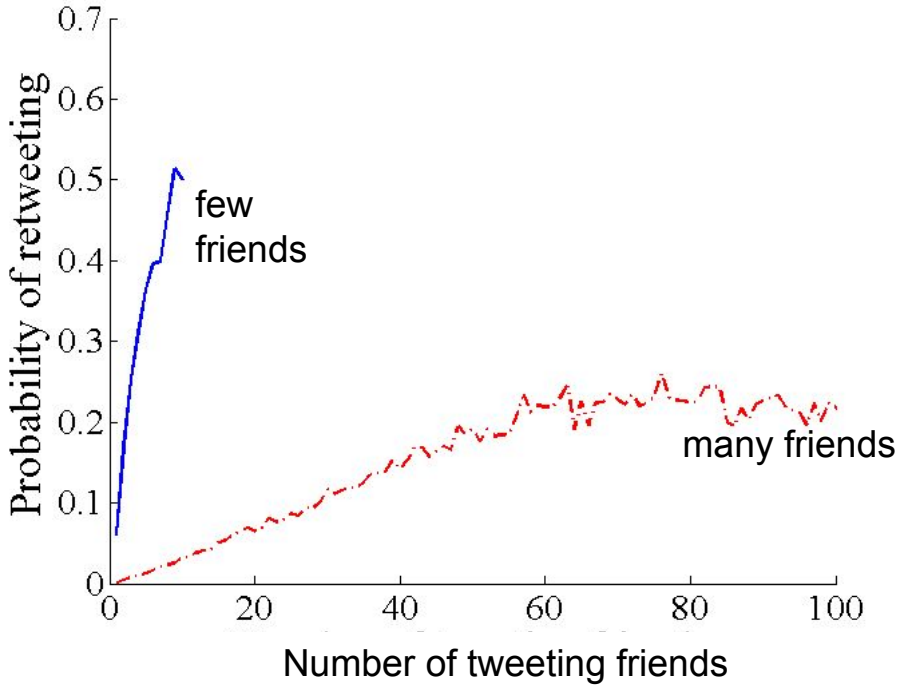
Survivor bias

Probability to retweet after exposure by *x* friends.
(peak is an artifact of survivor bias)



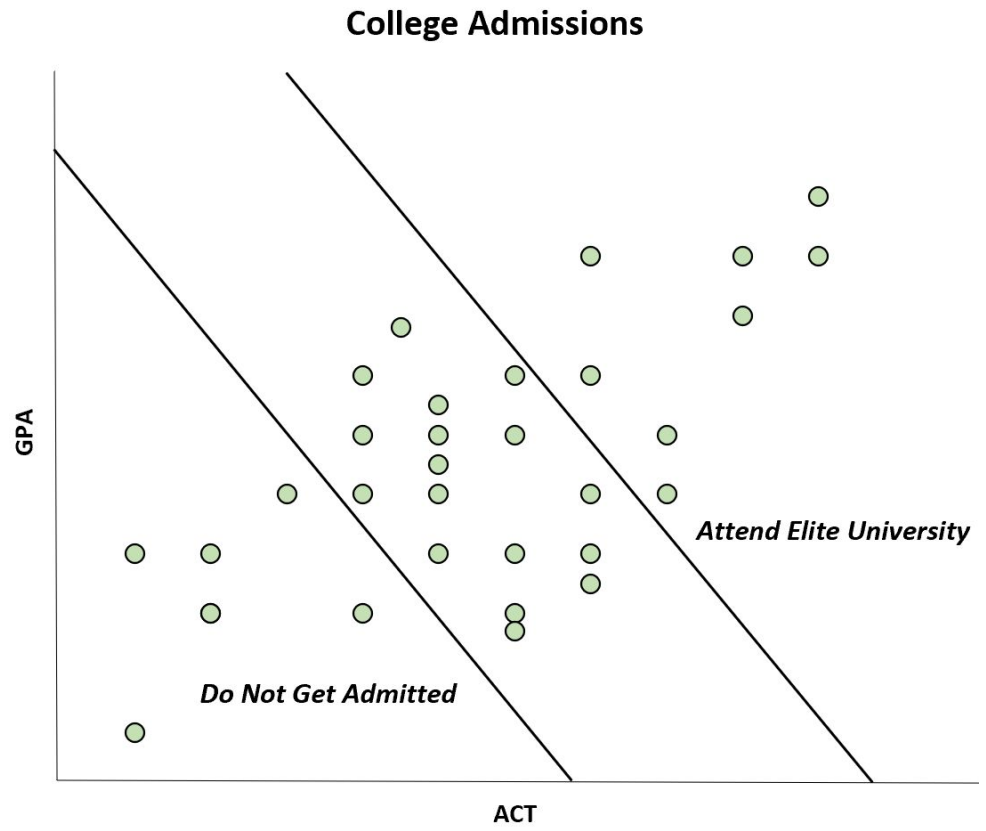
Romero et al. (2011) “Differences in the Mechanics of Information Diffusion Across Topics” in *WWW*.

Users with few friends drop out for larger *x* (exposures)
(users w/many friends are less susceptible)



Hodas & Lerman (2012) “How visibility and divided attention constrain social contagion”, in *SocialCom*.

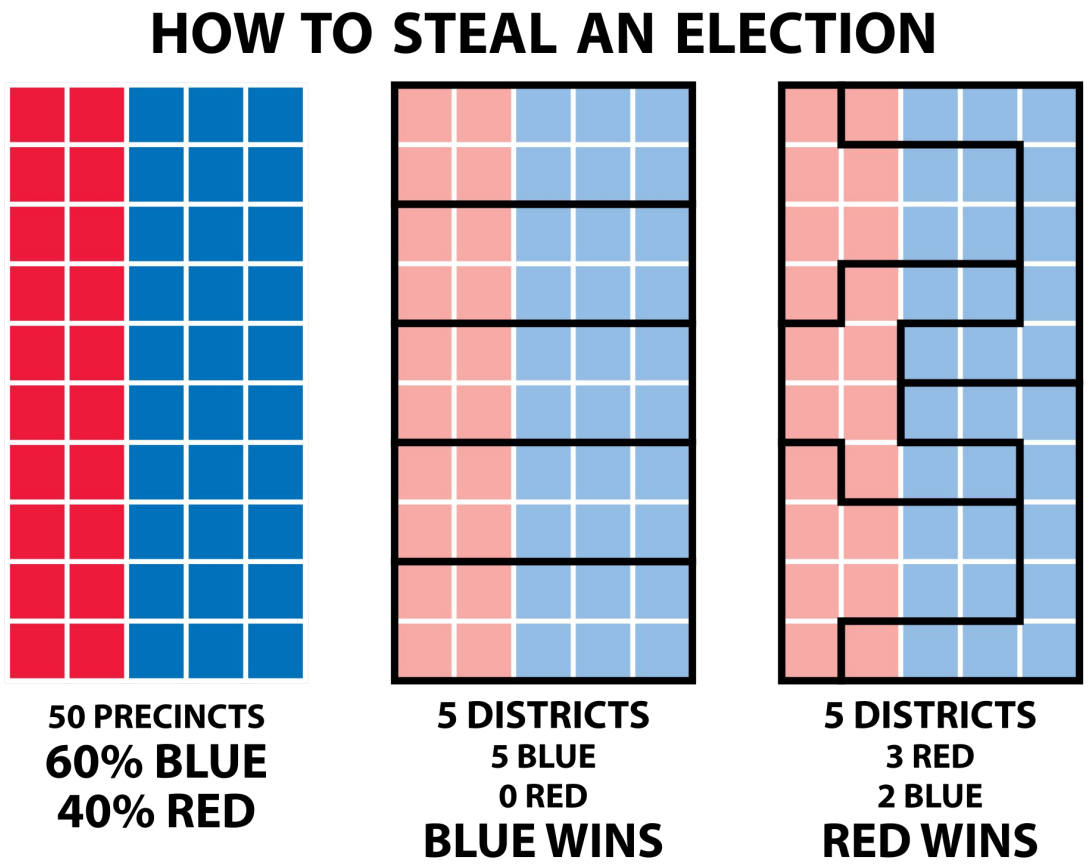
Berkson's paradox



- A positive correlation in a <u>population</u> can appear as negative correlation in a <u>sample</u>.
- Illustration: College only admits students with high enough GPA or high enough ACT score.
- GPA and ACT are correlated, but among the students who decide to go to a particular college, there appears to be a negative correlation between the two. Why?

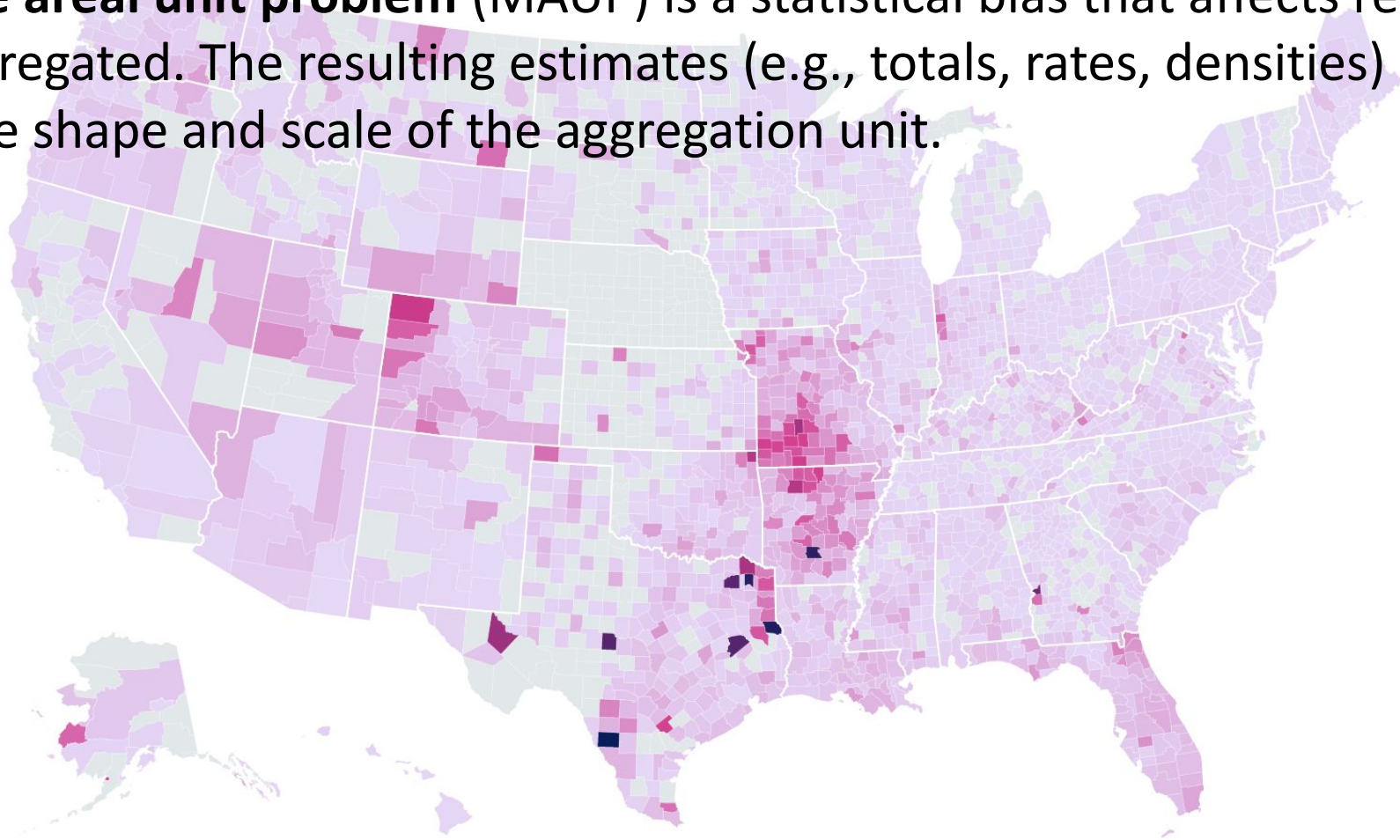
Aggregation bias

- **Modifiable areal unit problem** (MAUP) is a statistical bias that affects results when data is aggregated. The resulting estimates (e.g., totals, rates, densities) are affected by both the shape and scale of the aggregation unit.



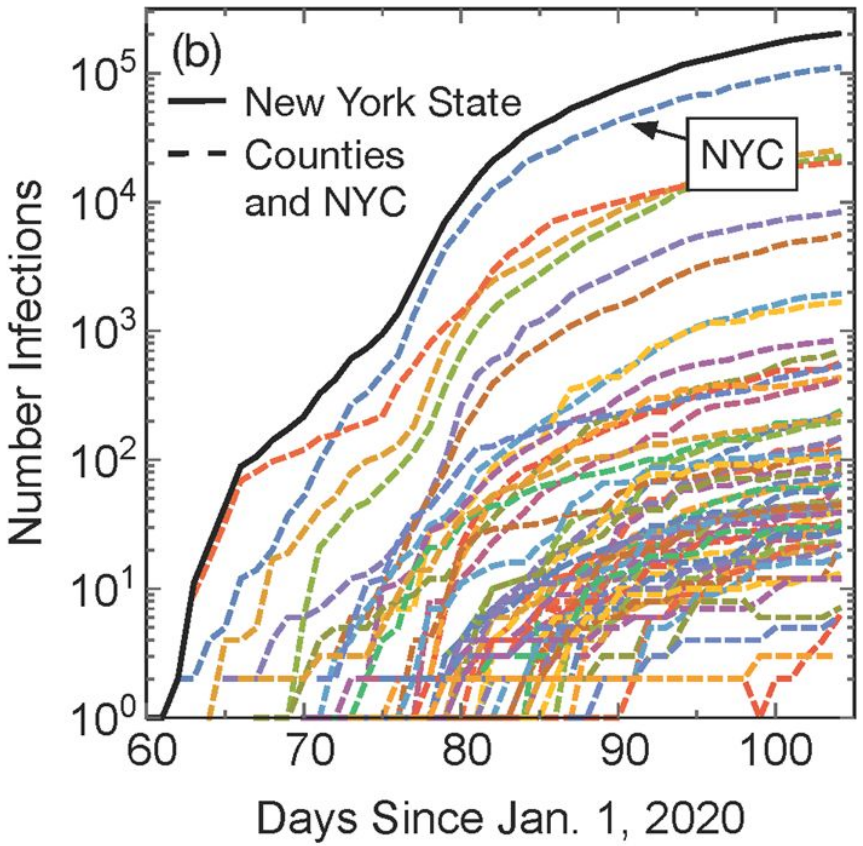
Aggregation bias

- **Modifiable areal unit problem** (MAUP) is a statistical bias that affects results when data is aggregated. The resulting estimates (e.g., totals, rates, densities) are affected by both the shape and scale of the aggregation unit.

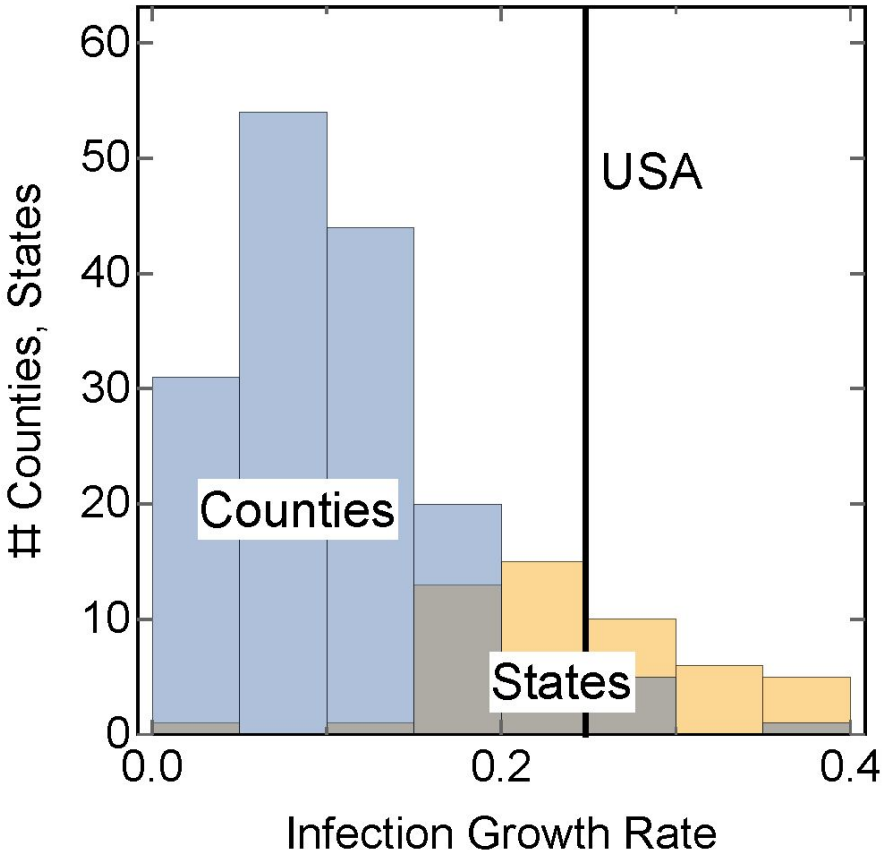


Aggregation bias

Growth rate of Covid-19 infections in US counties



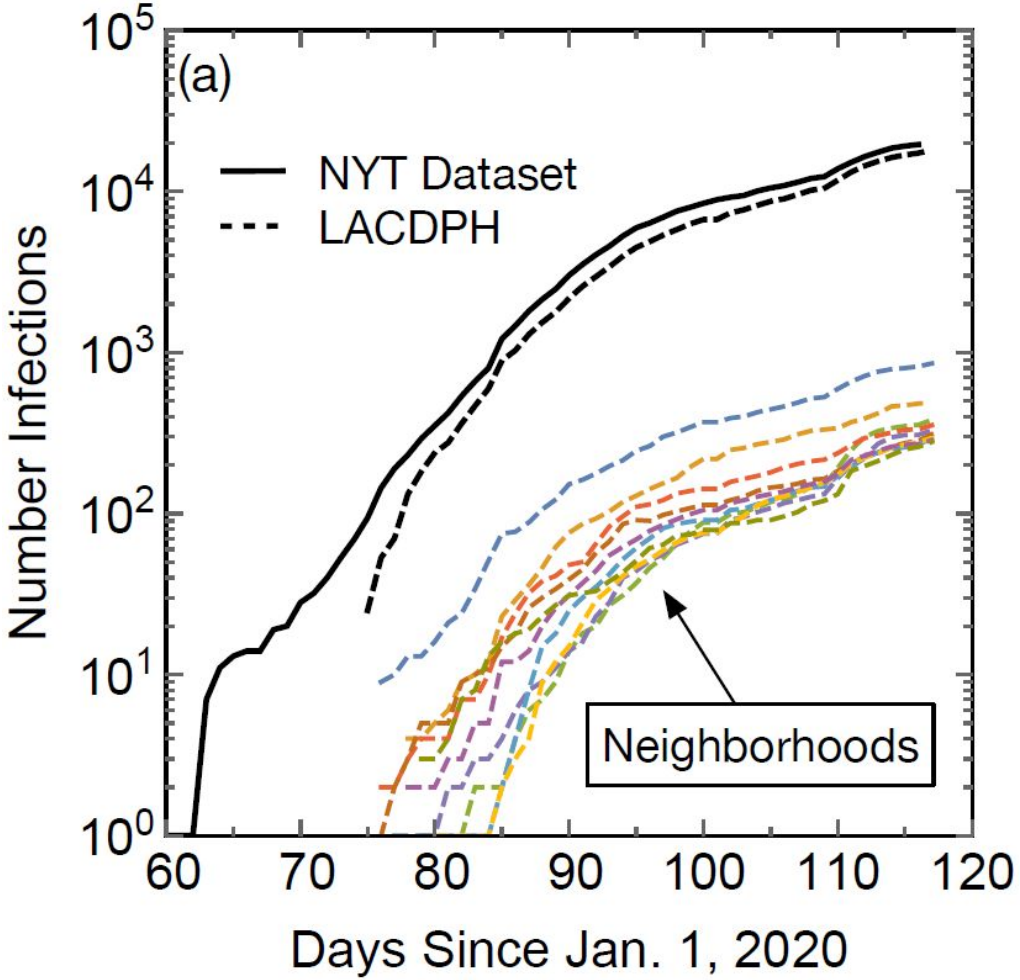
… appears slower than when the same data is aggregated by state, or country



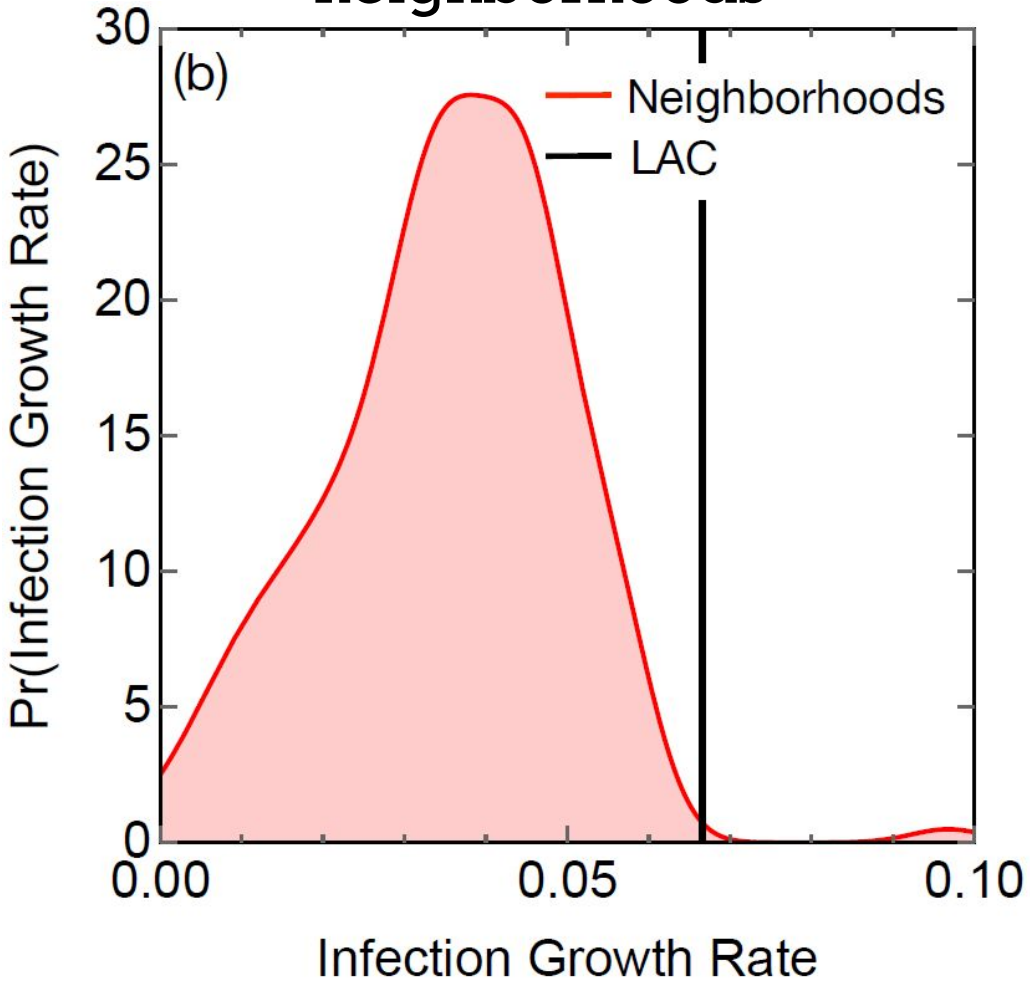
Burghardt, K., & Lerman, K. (2020). Systematic Biases in Aggregated COVID-19 Growth Rates. *arXiv preprint arXiv:2004.12994*.

… also at the metropolitan level

Infections in Los Angeles county



… appear to spread faster than in neighborhoods

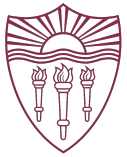


Analysis of temporal data

- Understanding data/behavior that evolves over time

- Cross-sectional analysis
  - Aggregate time-based analysis is a natural starting point
    - Approach: combine all available data about people (potentially over long periods of time)
  - This could be misleading and not describe any individual user
  - Why? Mixes old and new data
- Longitudinal analysis 纵向
  - Follow the same group over time

Accounting for time – cohort analysis



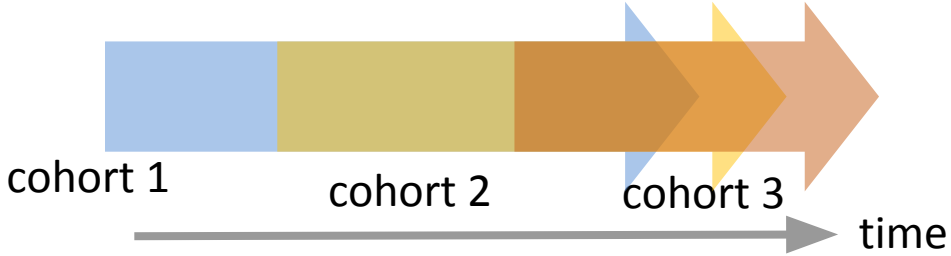
Online communities are evolving continuously.

- Tenure - How long a user has participated in the community
- Join date - When the user join the community

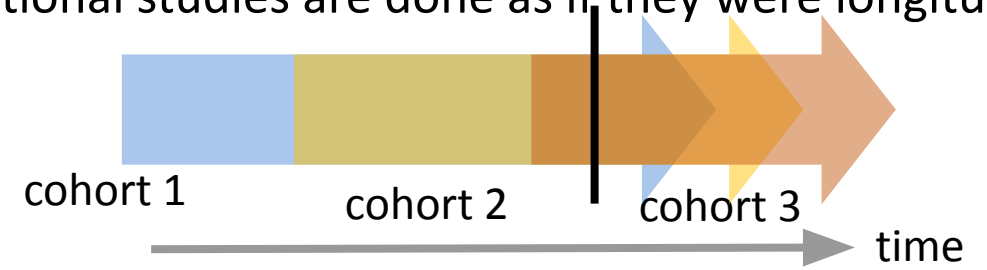
- Cohort analysis is a type of longitudinal analysis: A **cohort** is a group of people who share a characteristic, generally with respect to time
  - E.g., USC class of 2023, people born in 1990, etc.
- Cohorts in online communities
  - E.g., Wikipedia users who joined in 2005

Longitudinal vs Cross-sectional

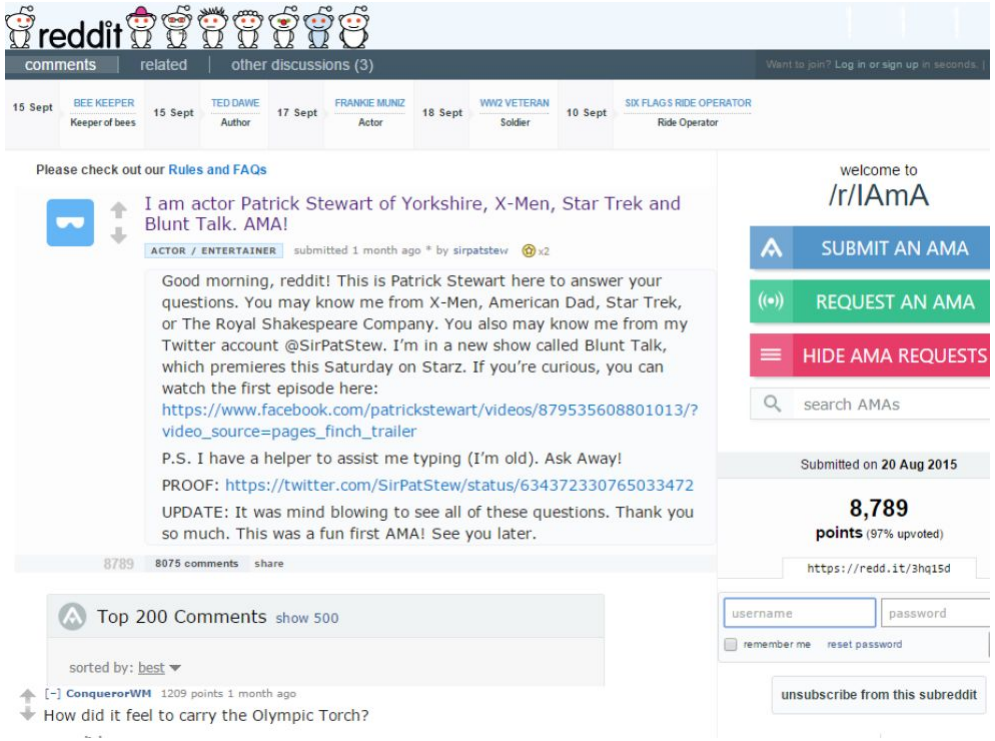
- **Longitudinal analysis** gathers data repeatedly from the same cohorts.
  - A **cohort** is a group of people who share a common characteristic, generally with respect to time: E.g., people born in 1990, Reddit users who joined in 2010, etc.



- **Cross-sectional analysis** aggregates data over all cohorts at a specific point in time.
  - Often cross-sectional studies are done as if they were longitudinal



Case study*: Is Reddit getting worse?



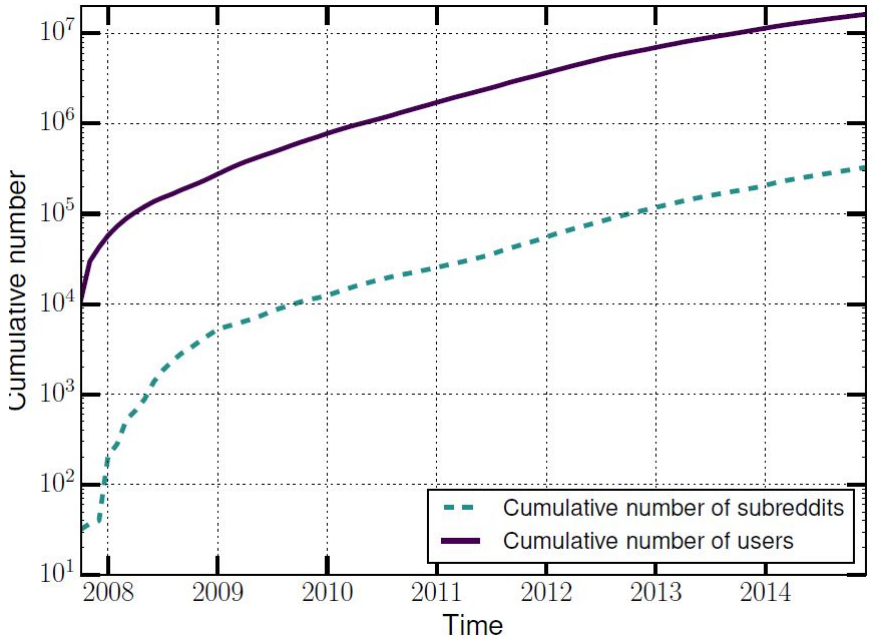
The data:

- 1.65 billion comments publicly available comments from October 2007 until May 2015
- 114 million submissions from October 2007 until December 2014
- UTC creation date, the username, the subreddit, and for comments, the comment text

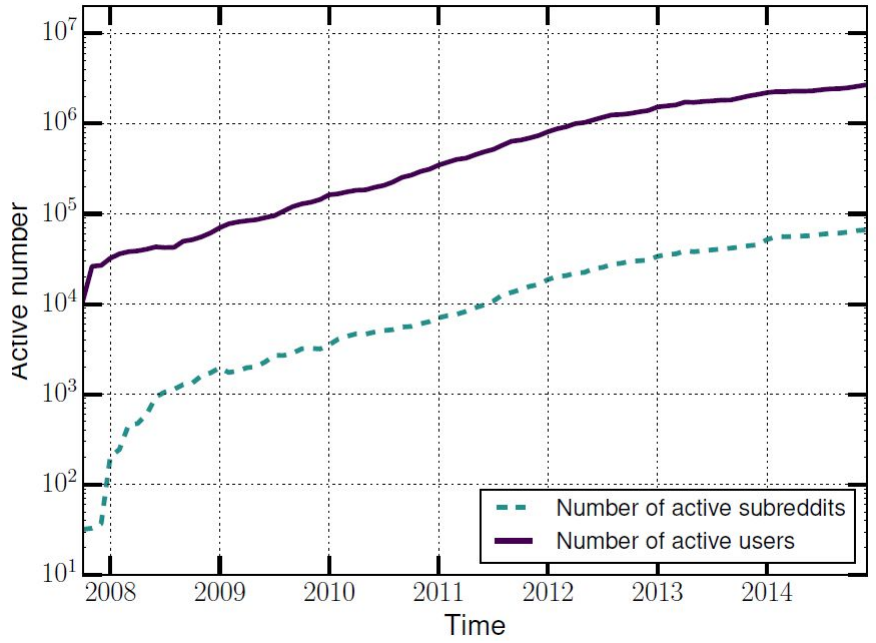
[Barbosa, S., Cosley, D., Sharma, A., & Cesar Jr, R. M. (2016, April). Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th International Conference on World Wide Web* (pp. 829-841). International World Wide Web Conferences Steering Committee.]

The size of Reddit

Cumulative number of users/subreddits



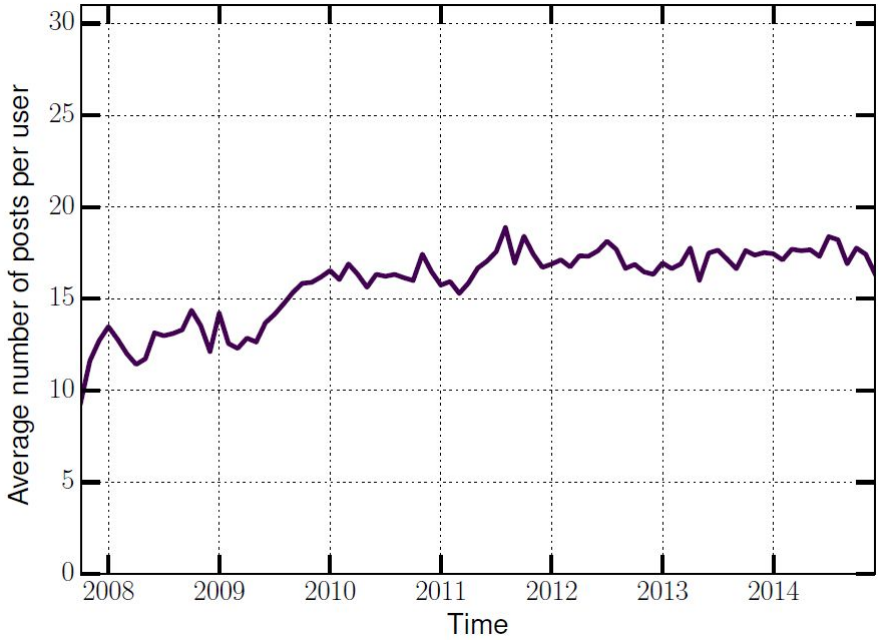
Number of active users/subreddits



Activity over time

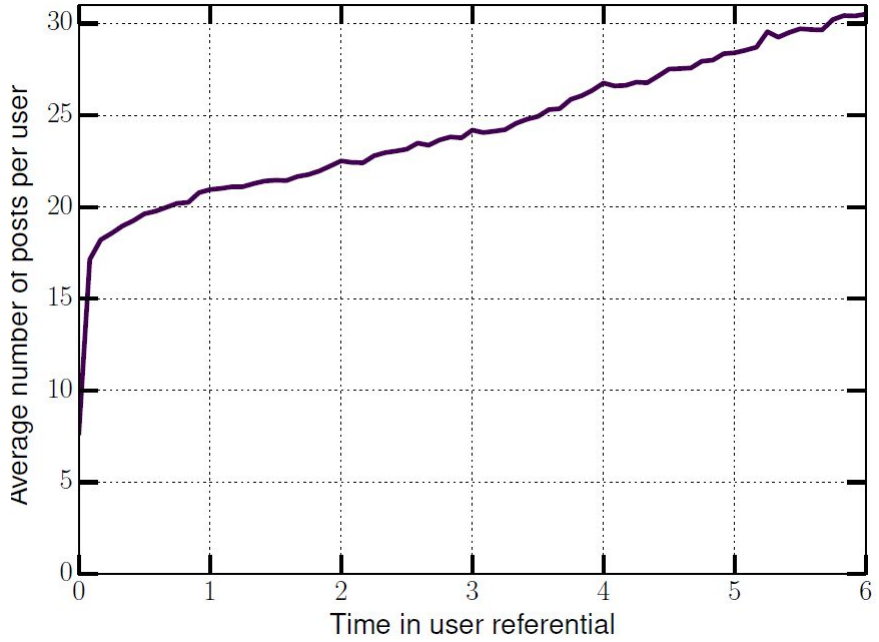
Activity over time:

Users may be becoming more active over time



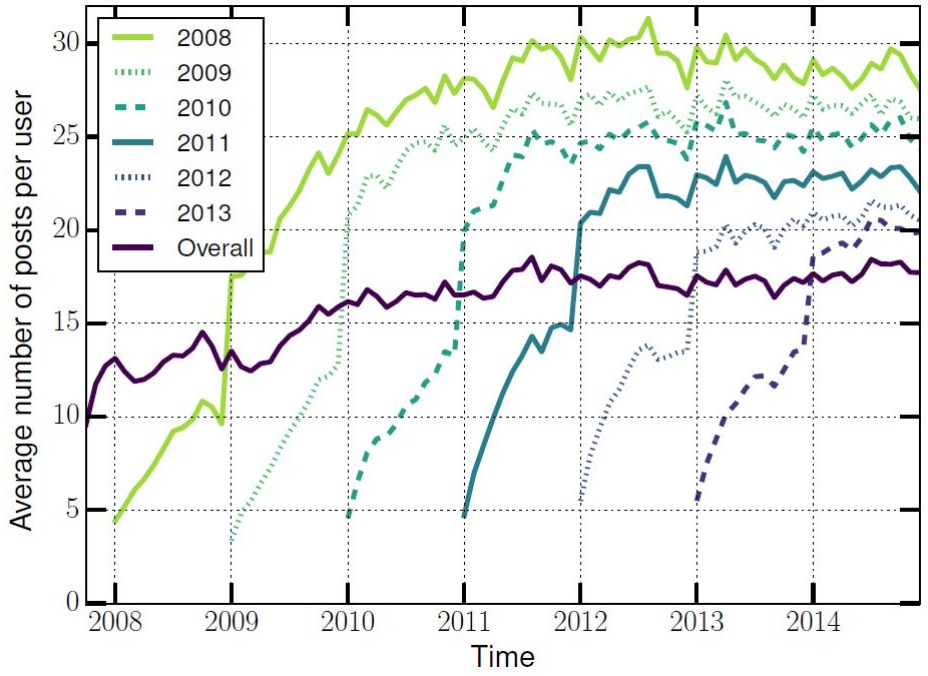
Activity with respect to user tenure:

The longer the user survives the more s/he posts. …?

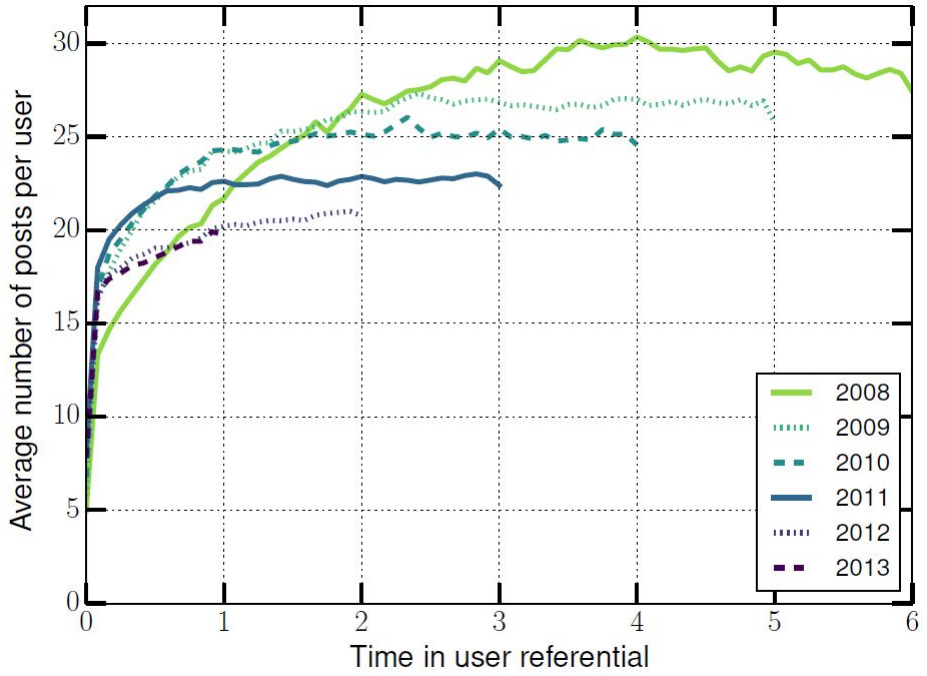


New cohorts do not catch up

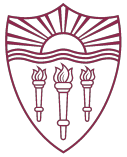
User activity split by join date



… and aligned



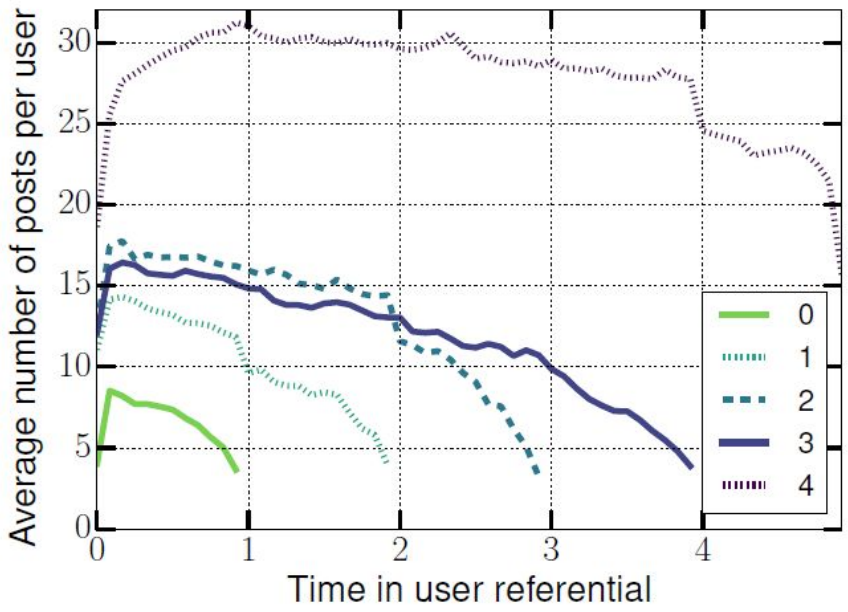
Does tenure predict activity or vice versa?



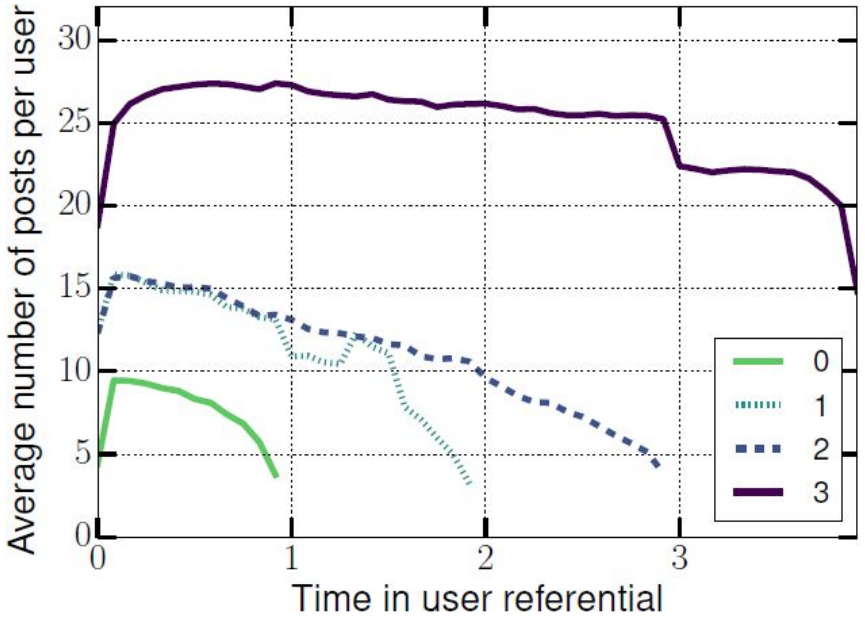
- Individual users come in with different motivations and posting propensities
- The rise over time in activity is not that individual users become more active but that low-activity users leave the system

A cohort becomes less active over time

2010 cohort – low activity users more likely to leave



2011 cohort

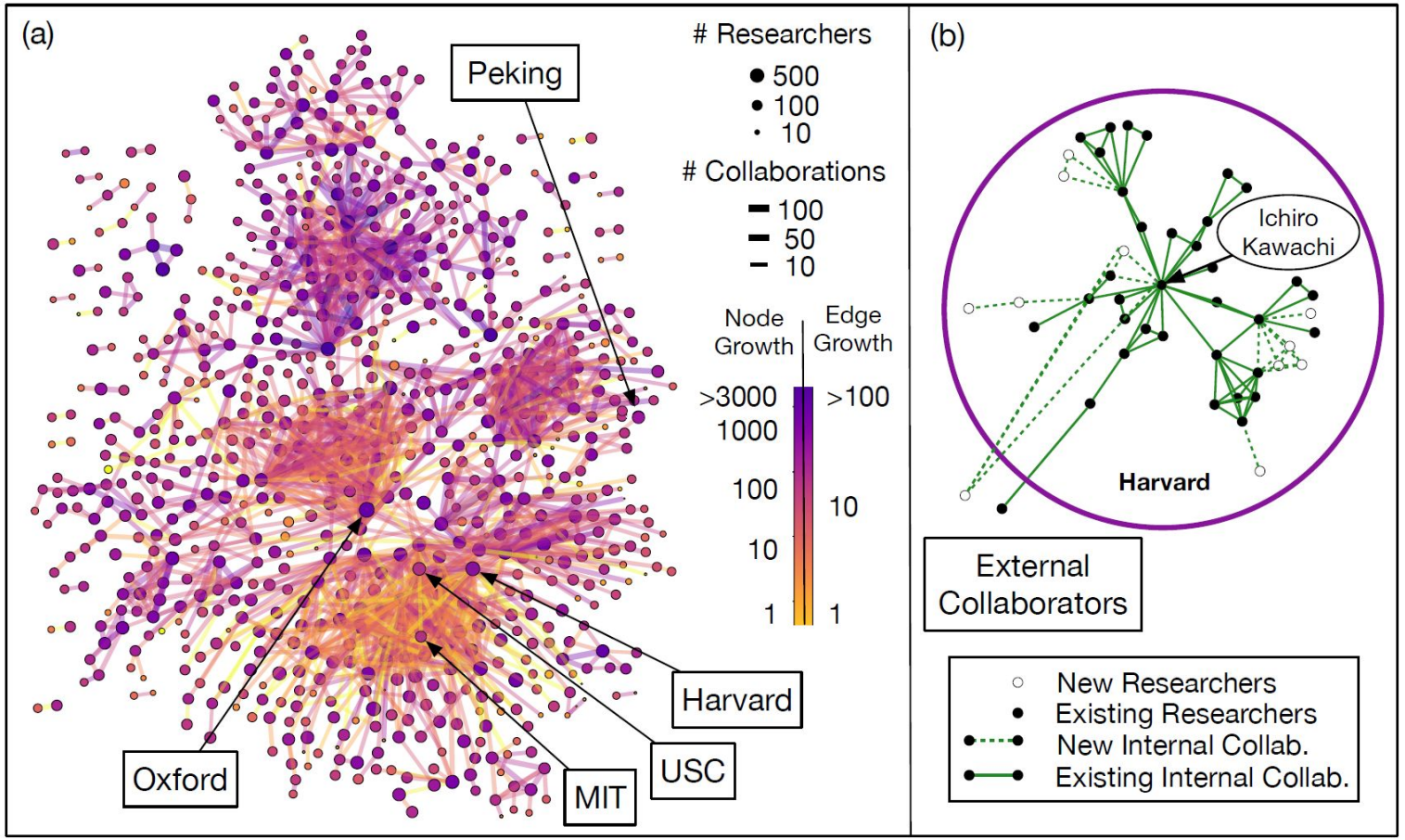


How do research institutions grow and facilitate scientific collaborations?

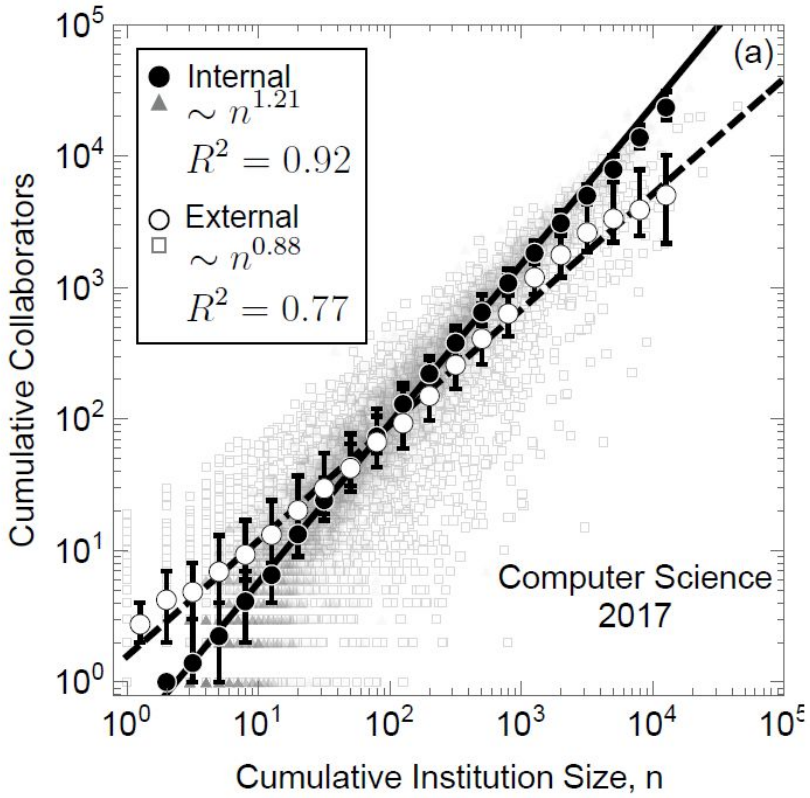
- How does the number of collaborations scale with institution size?
  - Do larger institutions facilitate disproportionately more collaborations than smaller ones?
  - Is this scaling universal, or does it vary across institutions?
- What are the statistical regularities in the growth of research institutions?
  - Do institutions follow known scaling laws (e.g., Zipf's law, Heaps' law)?
  - How does the number of institutions grow relative to the number of researchers?
- Is there a universal scaling law for collaborations across institutions?
  - Do all institutions exhibit the same growth exponent, or is there heterogeneous scaling?
  - How do internal and external collaborations differ in scaling behavior?
- Answer these questions with largescale bibliographic database

Scientific collaborations

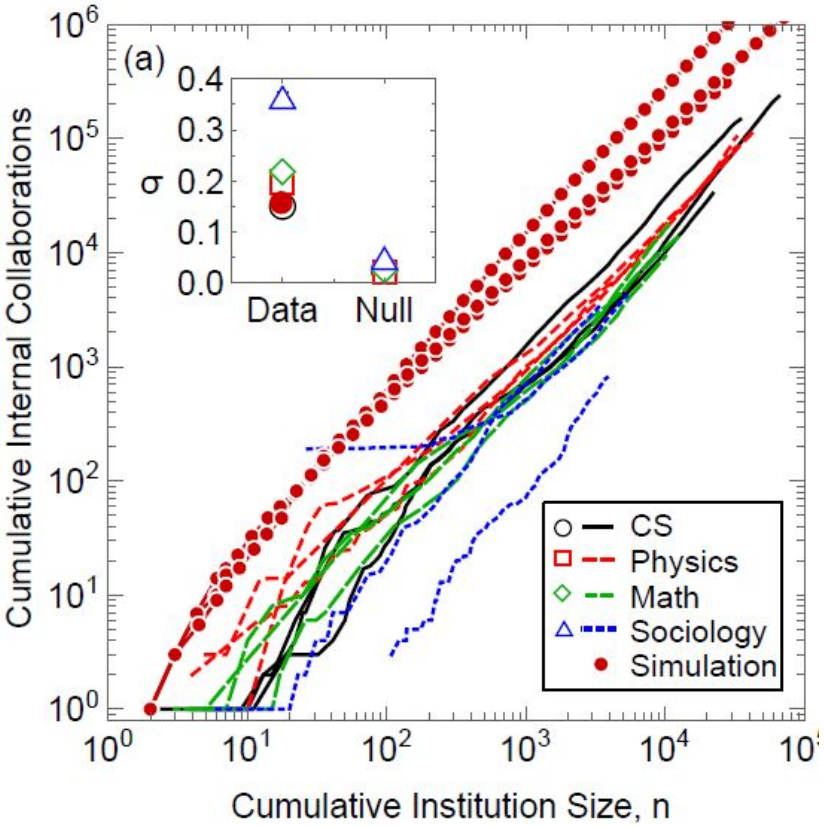
- Scientific collaborations as paper co-authorships
- Collaborations in sociology. (a) External collaborations and institution size. Each node represents a research institution. Institutions with more researchers are represented by larger nodes, and more collaborations are represented by thicker lines (edge weights).
- (b) The largest connected component of internal collaborations within Harvard University. Each node represents a researcher.



Cross-sectional vs longitudinal analysis = different results



Cumulative collaborators vs. institution size, n, for computer science (as of 2017). Shaded triangles (empty squares) are internal (external) cumulative collaborations for a single university. Black (white) circles are the mean internal (external) cumulative collaborations within institution of size n.



Number of collaborations as a function of institution size n for three different institutions in each field: CS, Physics, Math and Sociology.

Cross-sectional vs longitudinal analysis

Analysis Type	Collaboration Exponent	Stability Over Time	Observed Trend
Longitudinal (Tracks institution growth)	**≈ 1.2** (Superlinear, meaning institutions densify)	**Stable over time**	Institutions grow at different rates, creating heterogeneity in collaboration patterns.
Cross-Sectional (Single snapshot)	Varies over time but generally **< 1.2**	**Fluctuates over time**	Exponents change across years and tend to be lower than longitudinal values.

Debiasing heterogeneous data

debiasing data to create more fair and generalizable models

- disaggregate data into homogeneous groups
- disaggregate data into similarly-behaving groups
- remove correlations with a given feature

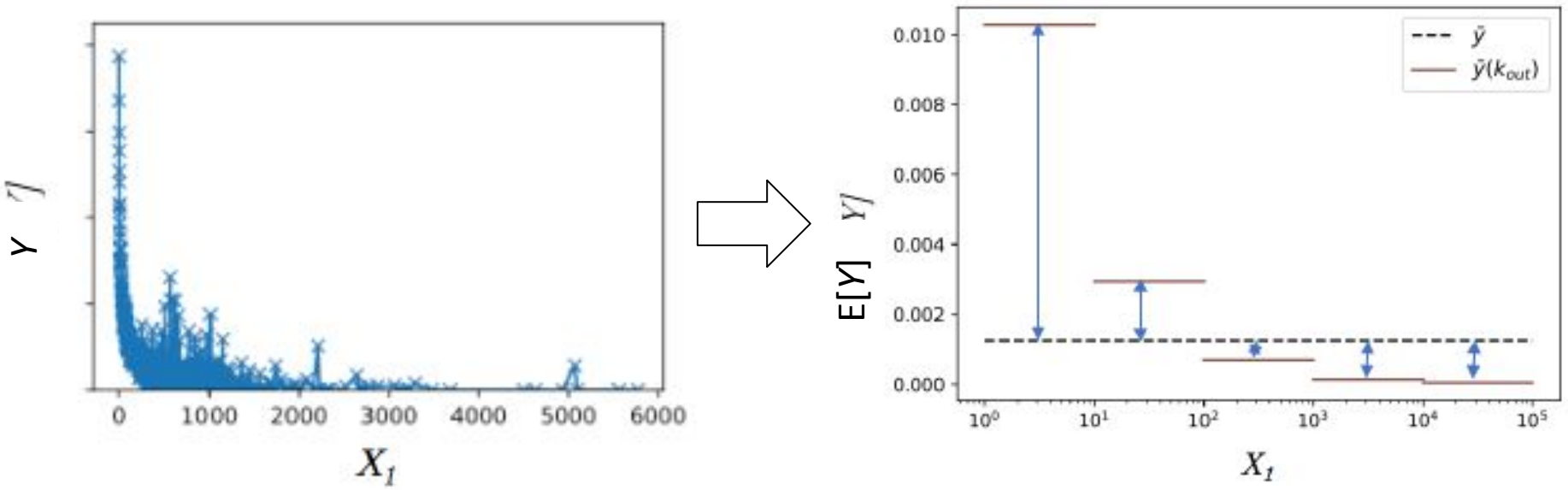
Central idea: Disaggregate data

- Reduce bias by disaggregating data into homogeneous subgroups
- Disaggregation tips
  - Ordinal variables: bin by value
    - E.g., disaggregate by department
    - E.g., disaggregate by year to create cohorts of users who joined in a given year
  - Continuous variables
    - Equal size bins? … some bins too sparse
    - Equal statistics bins? … some bins too heterogeneous
    - Data-driven binning

Data-driven binning

- Split the data so as to maximize the amount of variation of the outcome variable Y the bins explain

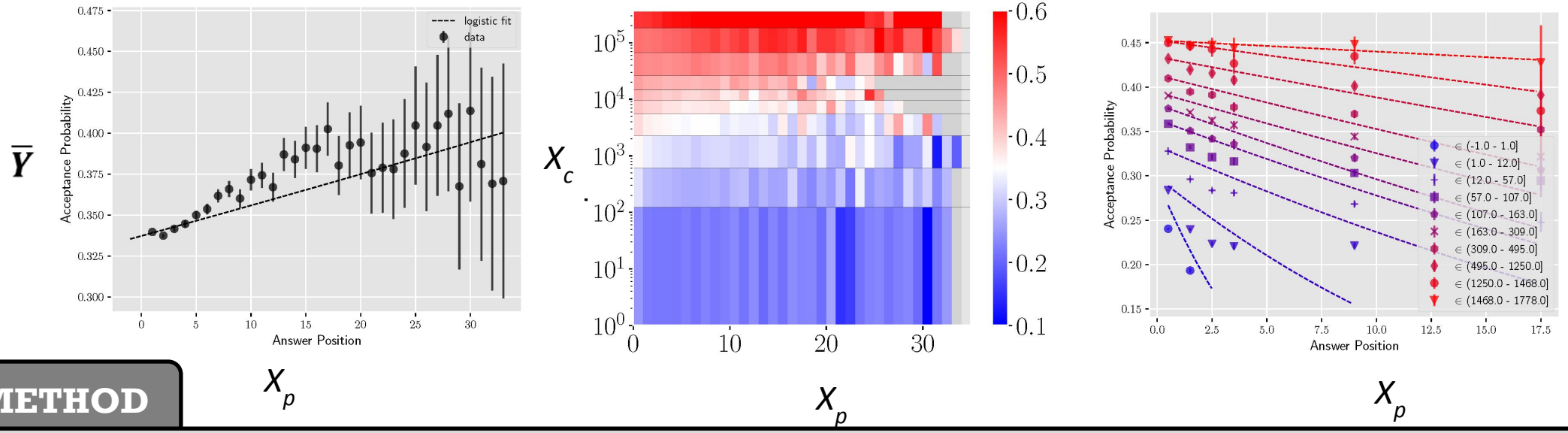
$$R^2 = \frac{\sum_{g=1}^{G} N_g (\bar{y}_g - \hat{y})^2}{\sum_{i=1}^{N} (y_i - \hat{y})^2}$$



[Fennell, Zuo & Lerman (2019), "Predicting and Explaining Behavioral Data with S3D", to appear in ***EPJ Data Science***]

Disaggregate data by observed features

- Which features create Simpson's paradox*
  1. **Estimate** regression trends within the overall (aggregated) data
  2. **Disaggregate** data by binning on an <u>observed</u> feature X_c
  3. **Compare** trends in disaggregated data to those for the aggregated data

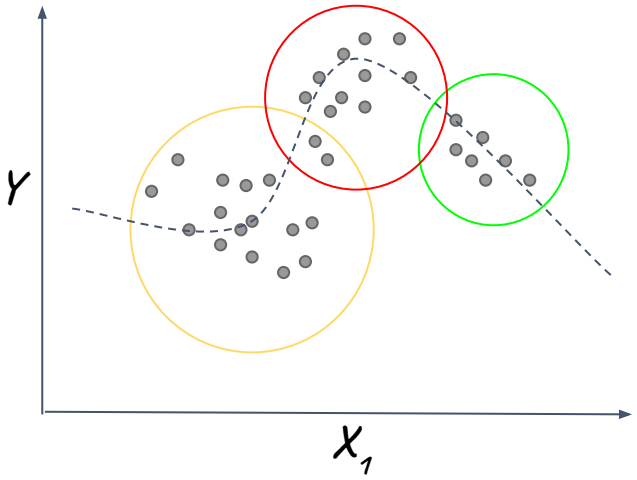


METHOD

Alipourfard, Fennell & Lerman (2018) *Using Simpson's paradox to discover interesting patterns in data.* ICWSM. *Code:* https://github.com/ninoch/Trend-Simpsons-Paradox

Debiasing data via disaggregation

- Disaggregate data on latent features to create more homogeneous groups
- Joint disaggregation + regression (DoGR)*
  - Disaggregate data along multiple dimensions
    - Represent population with (latent) subgroups as a Gaussian Mixture Model
    - Estimate parameters via EM
    - Soft clustering
  - Regression within subgroups
    - Separate regression coefficients within each subgroup



$$Y = a_1 X_1 + a_2 X_2 + a_3 X_3 + \ldots$$

	X_1	X_2	...
Green	-1.2	-0.6	...
Red	2.6	1.2	...
Orange	-0.1	3.01	...

METHOD

Alipourfard, Burghardt & Lerman (2021) *Disaggregation via Gaussian Regression for Robust Analysis of Heterogeneous Data.* Handbook on CSS. *Code:* https://github.com/ninoch/DoGR

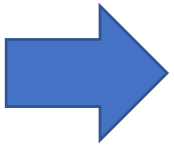
Illustration on Wine Data (UCI)

Groups
- Red vs White wine

Outcome
- Wine quality

Dimensions
- Citric acid,
- Chlorides,
- Free Sulfur Dioxide (SO2),
- Residual sugars,
- …

hide groups



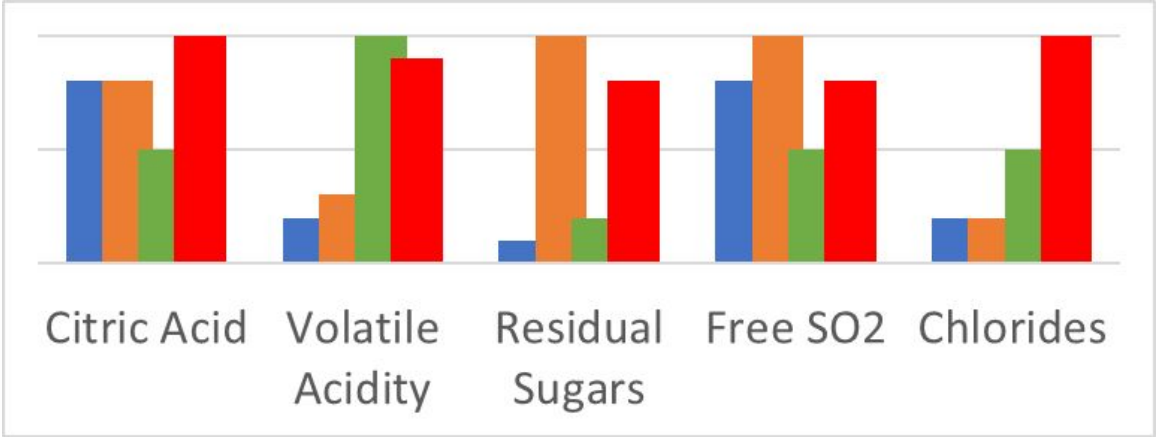
data

	outcome: quality	free SO2	citric acid	residual sugars	…
White wines	4898				
Red wines	1599				

Latent subgroups in wine data

`data`

	outcome: quality	free SO2	citric acid	residual sugars	…
White wines	4898				
Red wines	1599				

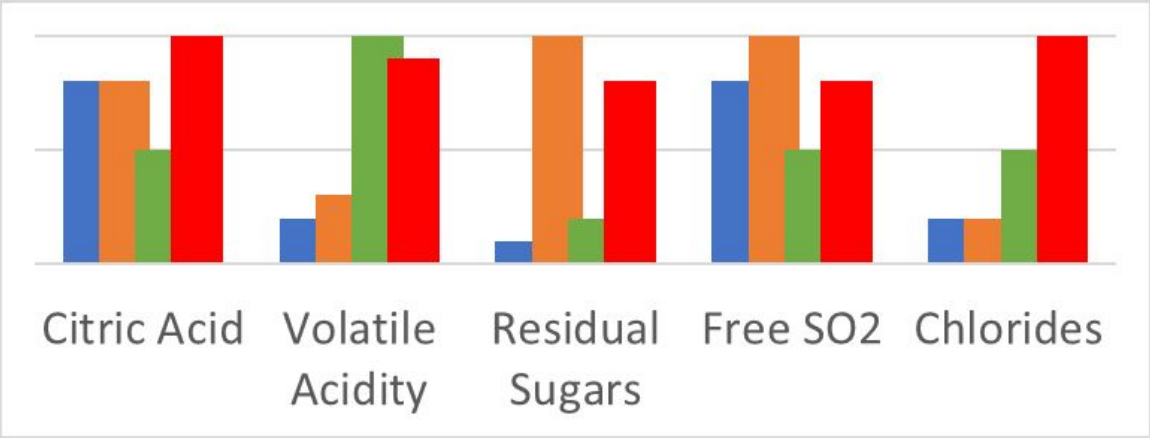


Subgroup	*Composition*	*Ave. quality*
Blue	98% white	6.02
Orange	100% white	5.91
Green	85% red	5.60
Red	57% red	5.36

METHOD

Alipourfard, Burghardt & Lerman (2021) *Disaggregation via Gaussian Regression for Robust Analysis of Heterogeneous Data.* Handbook on CSS. *Code:* https://github.com/ninoch/DoGR

Latent subgroups in wine data



Subgroup	*Composition*	*Ave. quality*
Blue	98% white	6.02
Orange	100% white	5.91
Green	85% red	5.60
Red	57% red	5.36

Subgroup regression: Simpson's reversal!

	Citric acid	*Free SO2*	*Residual sugar*
	β	β	β
All wines	*0.104*	*-0.001*	*-0.018*
Subgroup			
Blue	0.430	**0.017*** **	**0.427*** **
Orange	**-0.155****	-0.000	-0.016
Green	-0.097	-0.002	**0.323*** **
Red	0.304**	-0.011***	-0.004***

- Higher sugars and quality
- Higher citric acid and quality

METHOD

Alipourfard, Burghardt & Lerman (2021) *Disaggregation via Gaussian Regression for Robust Analysis of Heterogeneous Data.* Handbook on CSS. *Code:* https://github.com/ninoch/DoGR

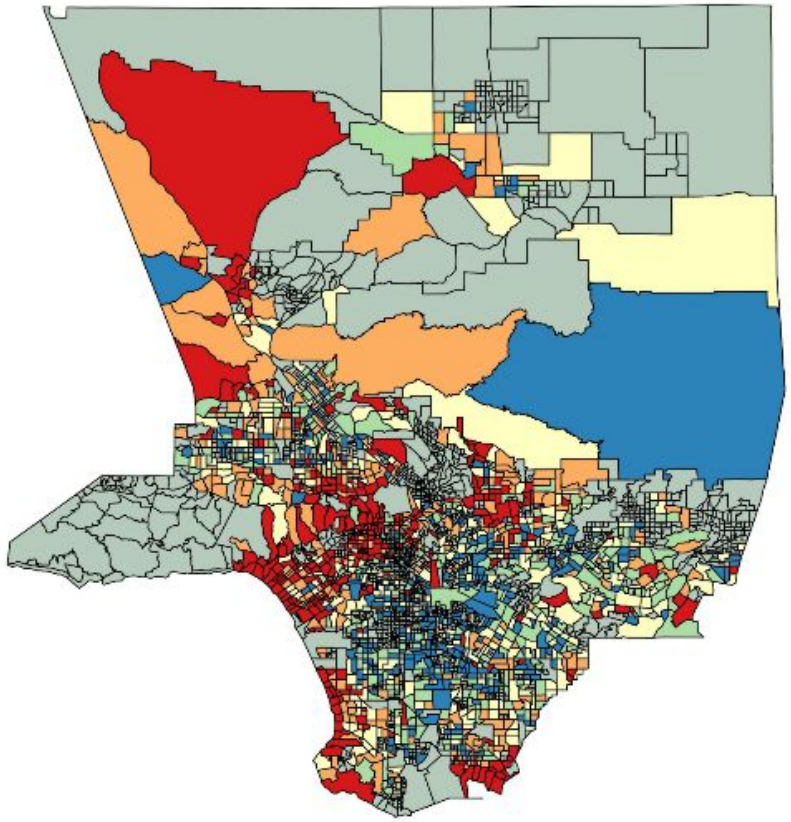
Evaluation

- **Prediction:** 5x5-fold nested cross validation to train model, test on out-of-sample data.
- **Evaluation**: Root Mean Square Error (RMSE) & Mean Absolute Error(MAE)
  - **Without disaggregation**
    - MLR: Multi-linear Regression
    - CART: Classification & Regression Trees
  - **With disaggregation (better)**
    - WCLR, FWCLR, GMR
    - DoGR

Method	RMSE ($\pm\sigma$)	MAE ($\pm\sigma$)
Synthetic		
MLR	294.88 ($\pm$ 1.236)*	288.35 ($\pm$ 0.903)*
CART	264.70 ($\pm$ 7.635)*	224.57 ($\pm$ 6.138)*
WCLR	261.14 ($\pm$ 3.370)*	232.76 ($\pm$ 2.682)*
FWCLR	261.27 ($\pm$ 4.729)*	233.05 ($\pm$ 3.772)*
GMR	257.36 ($\pm$ 4.334)	219.15 ($\pm$ 3.567)
DoGR	**257.32 ($\pm$ 3.871)**	**219.11 ($\pm$ 3.106)**
Metropolitan		
MLR	0.083 ($\pm$ 0.0061)	0.062 ($\pm$ 0.0033)
CART	0.086 ($\pm$ 0.0056)	0.064 ($\pm$ 0.0036)*
WCLR	0.083 ($\pm$ 0.0029)	0.062 ($\pm$ 0.0024)
FWCLR	**0.082 ($\pm$ 0.0044)**	**0.061 ($\pm$ 0.0021)**
GMR	0.083 ($\pm$ 0.0043)	0.061 ($\pm$ 0.0023)
DoGR	0.083 ($\pm$ 0.0052)	0.061 ($\pm$ 0.0031)
Wine Quality		
MLR	0.83 ($\pm$ 0.018)*	0.64 ($\pm$ 0.015)*
CART	0.79 ($\pm$ 0.015)	0.62 ($\pm$ 0.013)
WCLR	0.83 ($\pm$ 0.013)*	0.64 ($\pm$ 0.011)*
FWCLR	0.80 ($\pm$ 0.013)*	0.63 ($\pm$ 0.009)*
GMR	0.79 ($\pm$ 0.017)	0.62 ($\pm$ 0.014)
DoGR	**0.79 ($\pm$ 0.014)**	**0.62 ($\pm$ 0.011)**
NYC		
MLR	13.36 ($\pm$ 7.850)	2.20 ($\pm$ 0.064)*
CART	15.33 ($\pm$ 9.128)	**1.34 ($\pm$ 0.190)**
FWCLR	13.14 ($\pm$ 7.643)	1.76 ($\pm$ 0.321)*
DoGR	**11.88 ($\pm$ 9.109)**	1.40 ($\pm$ 0.222)

Happiness around LA County

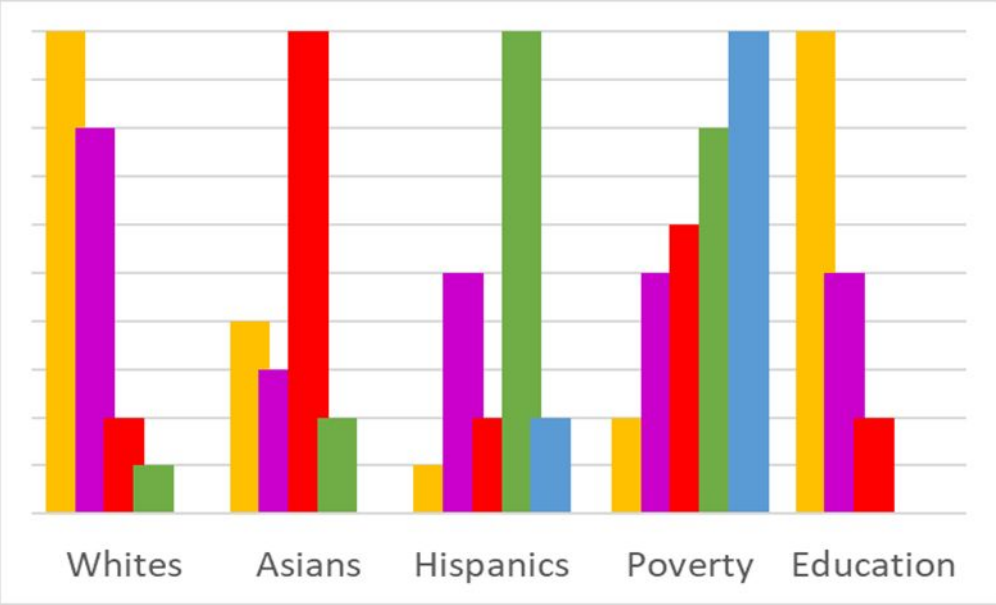
- Geolocated tweets linked to census tracts
- Measure outcome: sentiment of tweets



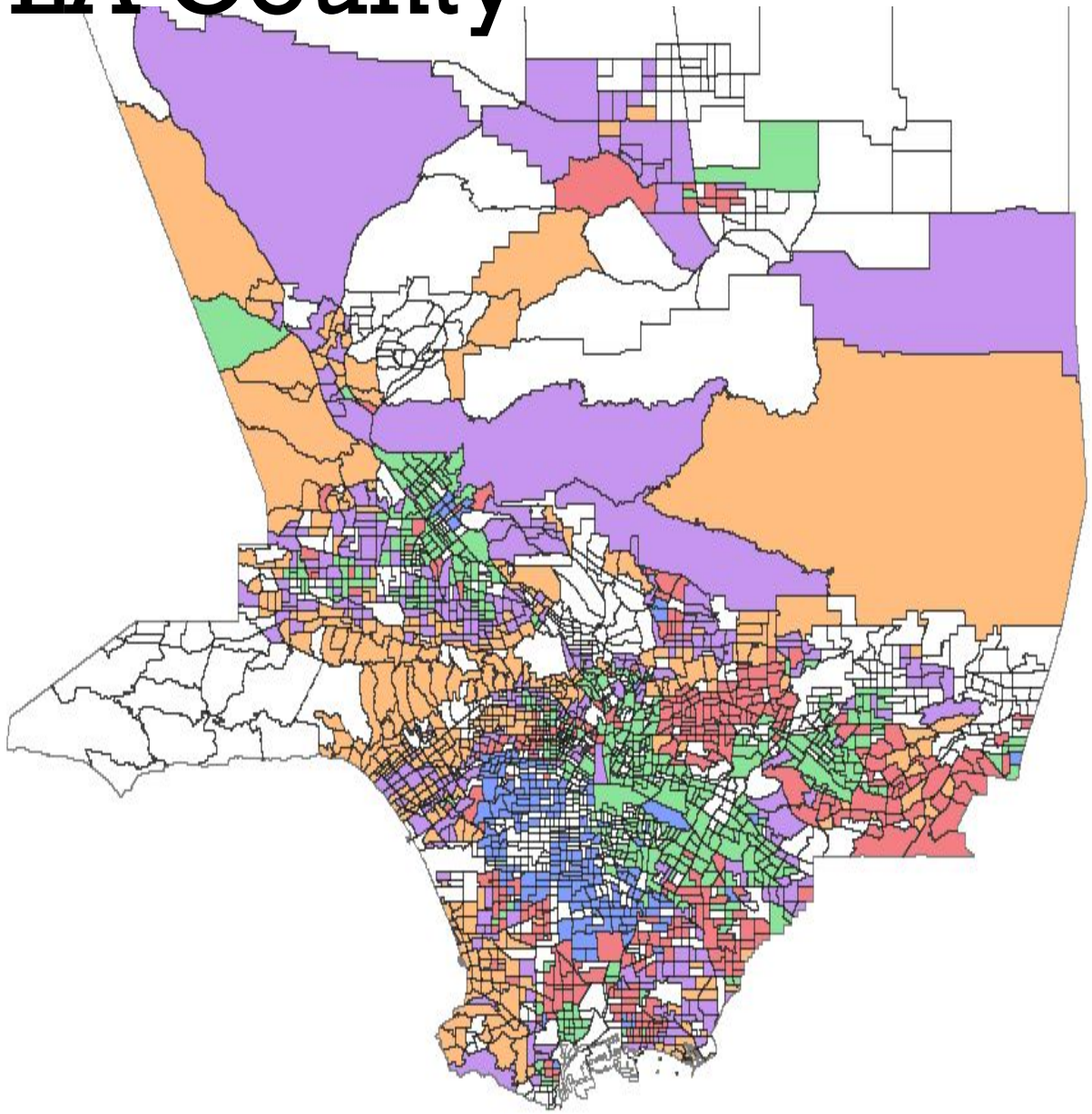
Features

- Whites,
- Blacks,
- Asians,
- Hispanics,
- Poverty (% below poverty)
- Education (% graduate degree)
- Sentiment (color: red=happy, blue=sad)

Latent subgroups of LA County



Subgroup	*Ave. valence*
Orange	5.86
Purple	5.80
Red	5.76
Green	5.74
Blue	5.72



Lessons learned

- Regardless of how many features are considered, individuals will differ along neglected dimensions. Some of these differences affect the individuals' chances of death, marriage, unemployment, or other transition.
- Because of this heterogeneity, selection will occur: the surviving population will differ from the original population.
- This means that observations of the surviving population cannot be directly translated into conclusions about the behavior of the individuals who made up the original population.
- The observed dynamics at the population level will deviate from the underlying dynamics at the individual level.

Lessons learned

- When is this not an issue? When the population is homogeneous.
- In other cases, patterns observed over the population may be surprisingly different from the underlying patterns on the individual level
- Researchers interested in uncovering these individual patterns – to help develop theories or to make predictions, might benefit from an understanding of heterogeneity.
- When should a researcher suspect substantial heterogeneity? Always.

"I can't understand why the whole audience hated my Simpson's Paradox joke. I tried it on the men and the women in the crowd separately and each group loved it!"