BIAS IN DATA

Kristina Lerman
USC Information Sciences Institute
DSCI 552 – Spring 2021
March 1, 2021

Topics

- Bias in data
  - Sources of bias in data
  - Understand the impact of bias on data analysis
  - Learn how to evaluate bias in data
  - Computational strategies to mitigate bias in data
- Algorithmic fairness
  - What is fairness in AI?
  - Bias in data and algorithmic fairness
  - Measures of algorithmic fairness; the impossibility of total fairness
  - Methods: Improving fairness by debiasing data

bi·as

/ˈbīəs/

noun: bias

1. prejudice in favor of or against one thing, person, or group compared with another, usually in a way considered to be unfair

2. a concentration on or interest in one particular area or subject

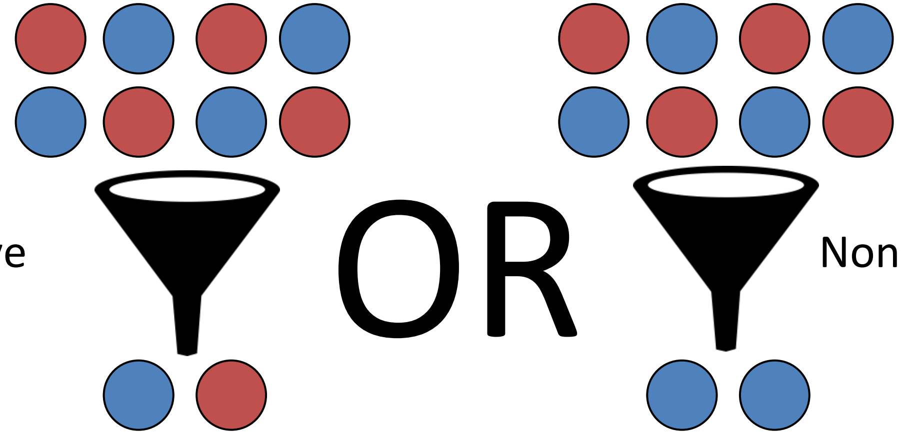
3. a systematic distortion of a statistical result due to a factor not allowed for in its derivation

Selection bias

- Operational Definition:

  Non-representativeness

Representative Sample



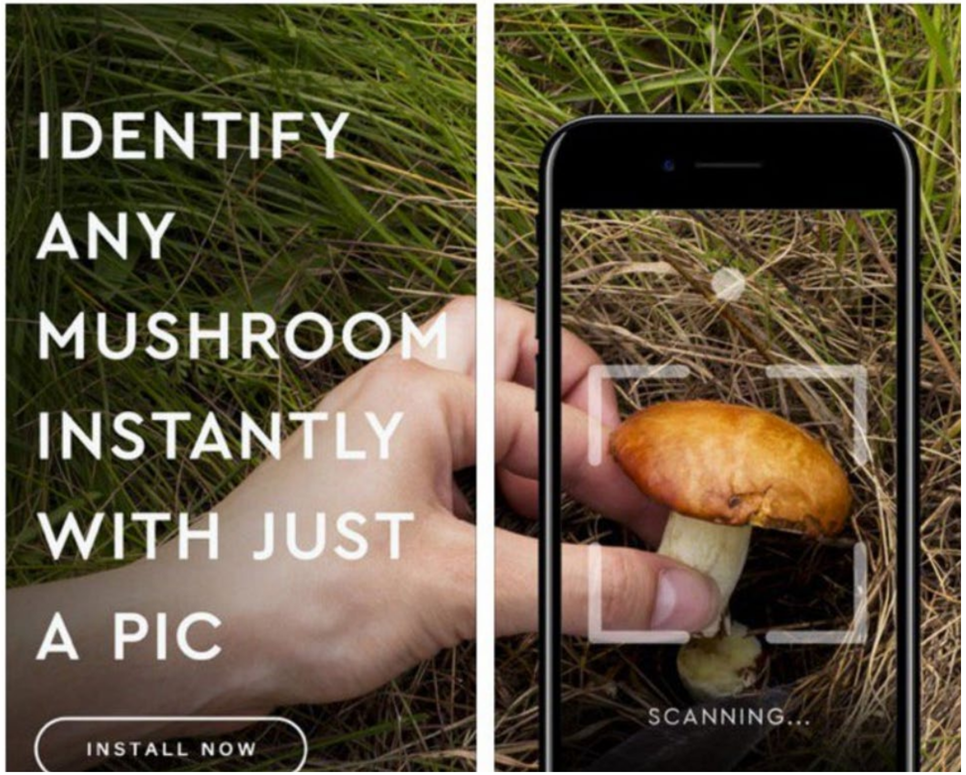
Non-Representative Sample

Selection bias is everywhere

Population of app reviewers with positive experience is different from the population of reviewers with negative experience

Darren Dahly
@statsepi

Survival bias means this this app will get fantastic reviews. #epidemiology ht @mathbabedotorg @EdwardBehan



Bias is a threat to the validity of models learned from data

Threats to Prediction

- Non-generalizable and non-reproducible models
- Poor performance on held-out data

Threats to Explanation

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

Threats to Fairness

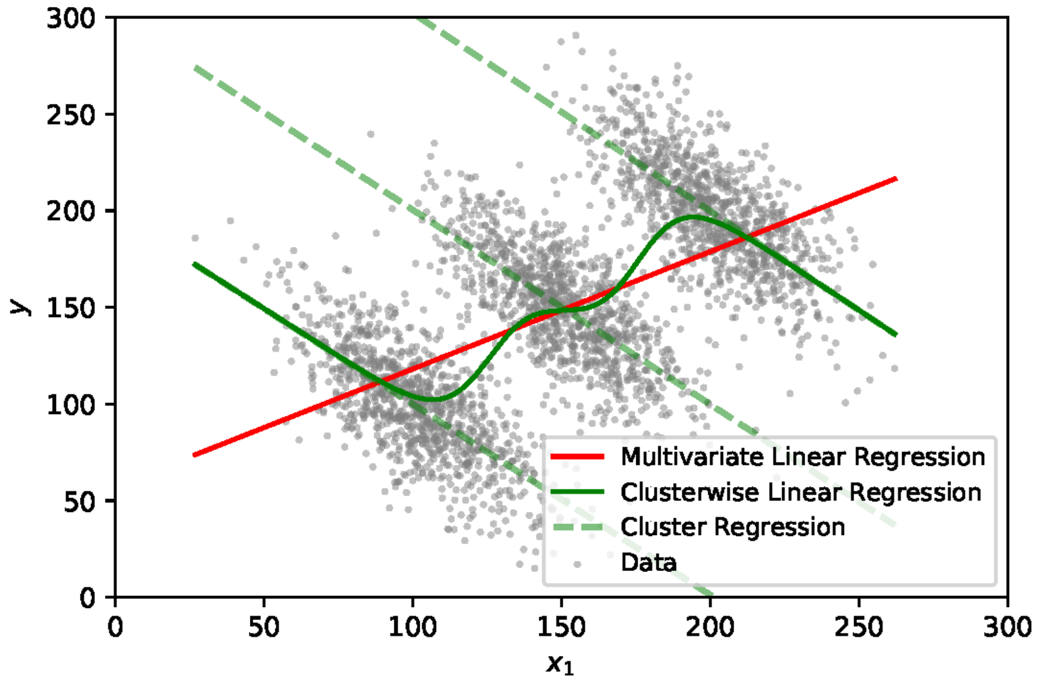
- Models learned on biased data may entrench and amplify discrimination

Simpson's paradox Subgroups with different behavior & population data	**Sampling bias** Subgroups not equally represented	**Filtering bias** Subsampling may distort data

Sources of bias in data

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



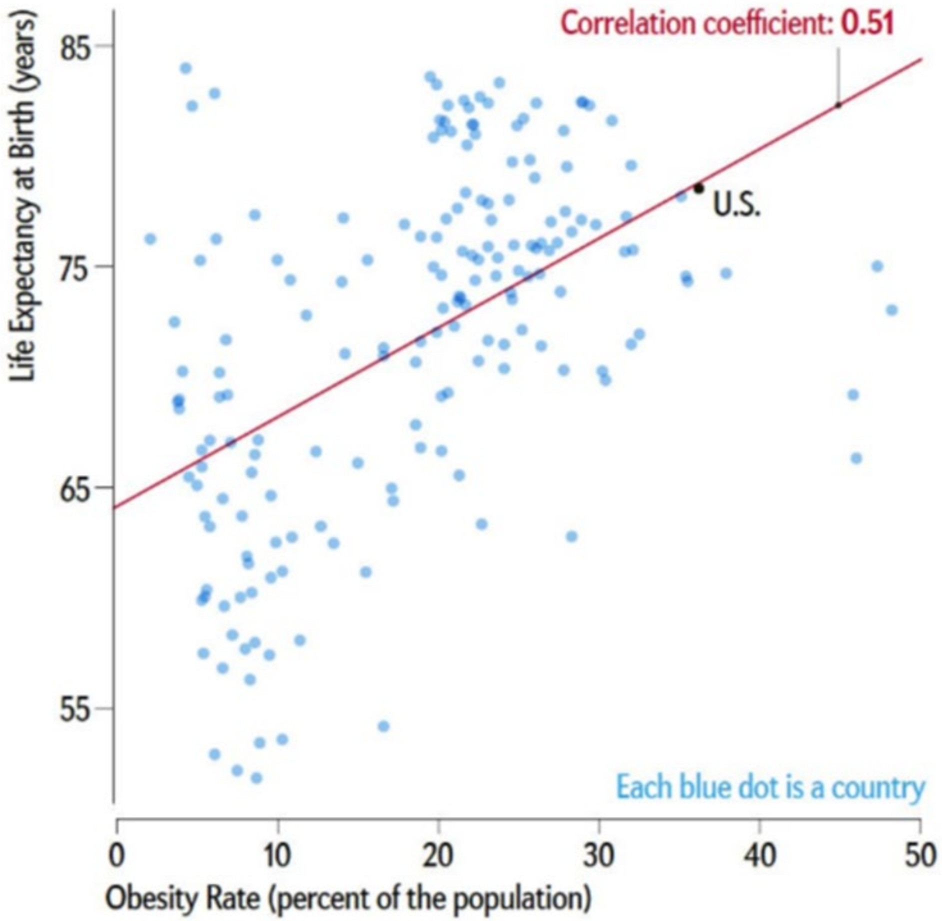
A **trend** appears in different sub-groups of data

but

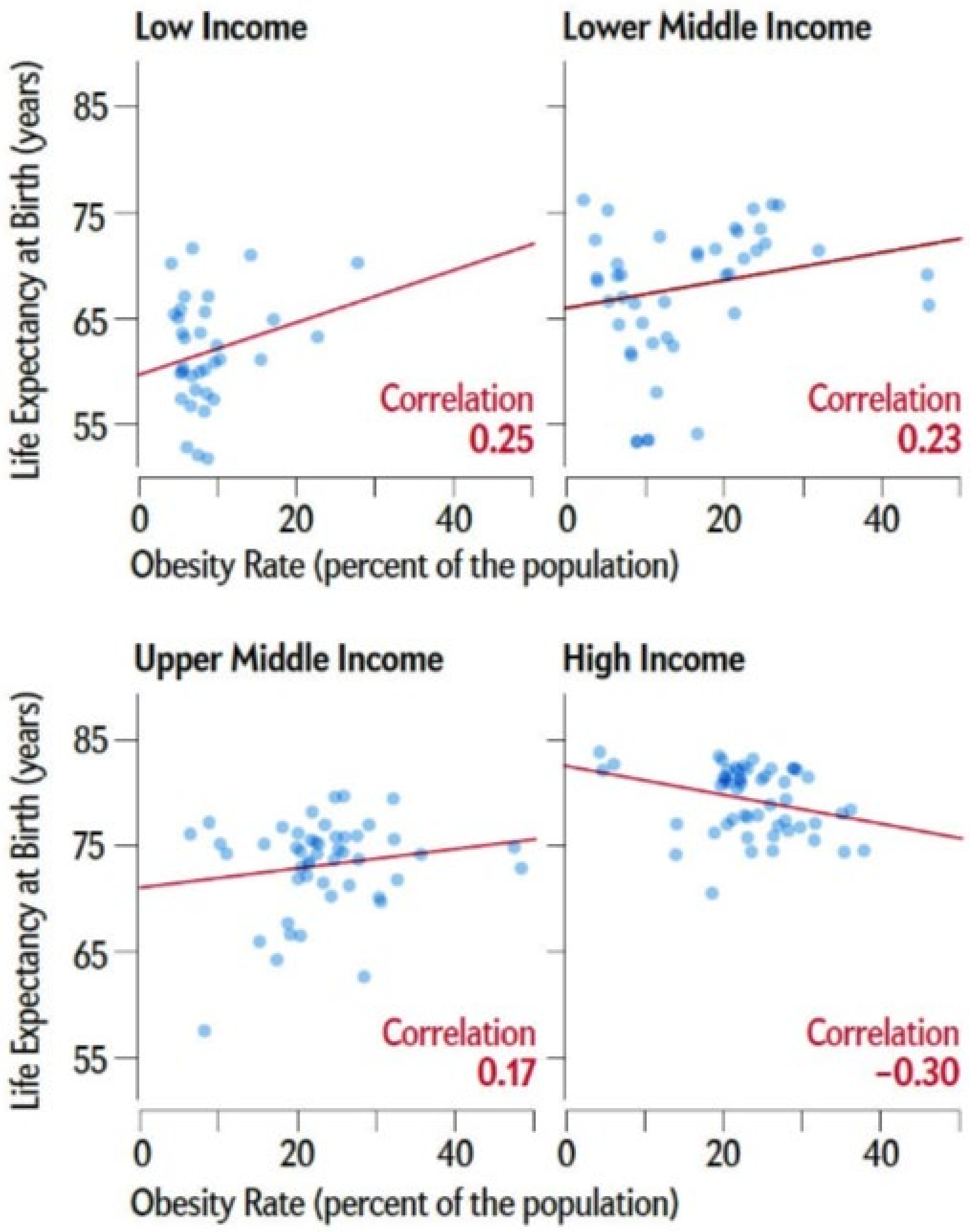
disappears or reverses when these sub-groups are combined.*

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

Does obesity shorten lives?



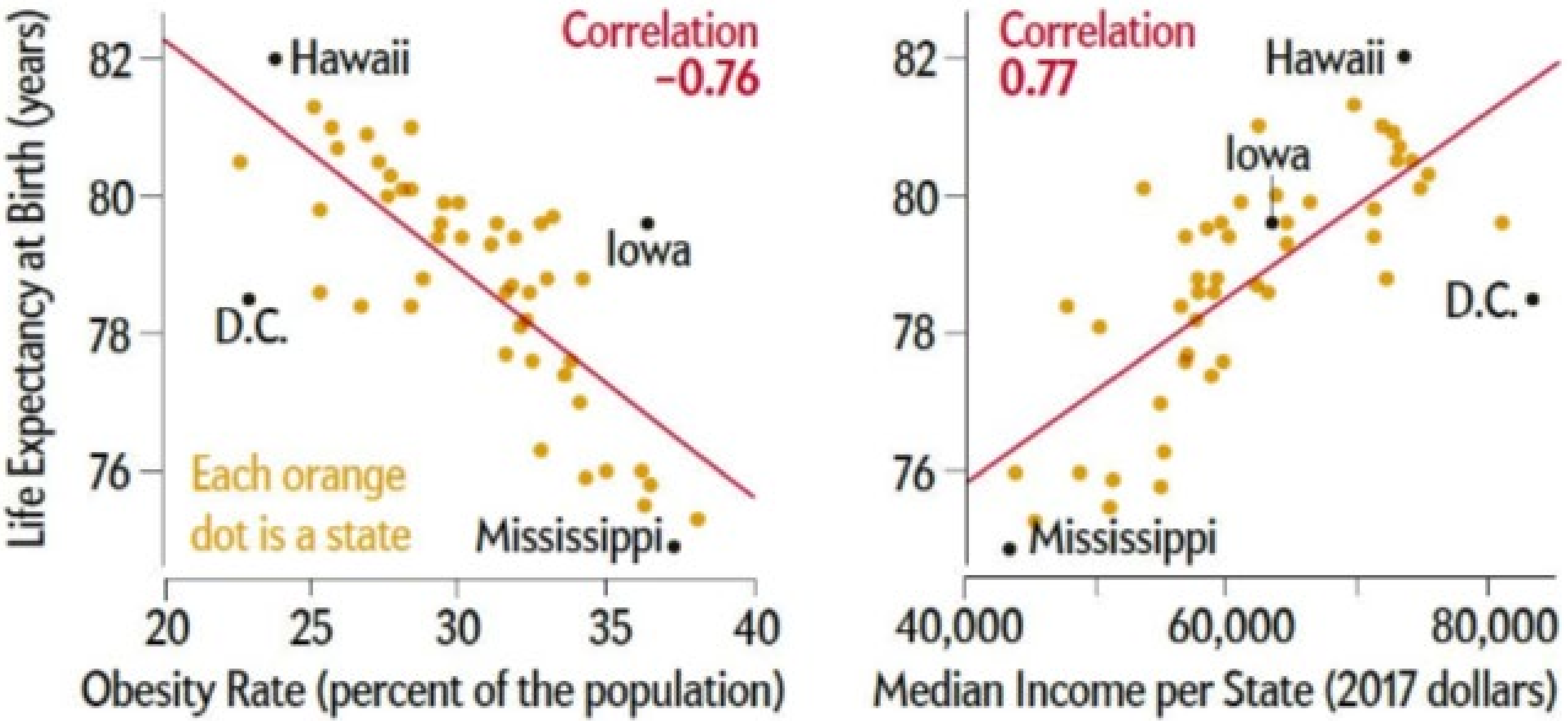
https://www.scientificamerican.com/article/graphics-that-seem-clear-can-easily-be-misread/



First, a pattern in aggregated data can disappear or even reverse once you explore the numbers at different levels of detail. If the countries are split by income levels, the strong positive correlation becomes much weaker as income rises. In the highest-income nations (*chart on bottom right*), the association is negative (higher obesity rates mean lower life expectancy).

https://www.scientificamerican.com/article/graphics-that-seem-clear-can-easily-be-misread/

The pattern remains negative when you look at the U.S., state by state: life expectancy at birth drops as obesity rises (*left*). Yet this hides the second fallacy: the negative association can be affected by many other factors. Exercise and access to health care, for example, are associated with life expectancy. So is income (*right*). The fallacy is trying to determine something about your individual risk by looking at aggregated data that do not reflect individual circumstances. If instead you saw data on individuals within a large sample of randomly selected people, you might discover that obesity may, or may not, relate to life expectancy for someone in your situation.



https://www.scientificamerican.com/article/graphics-that-seem-clear-can-easily-be-misread/

Sex Bias in Graduate Admissions: Data from Berkeley

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

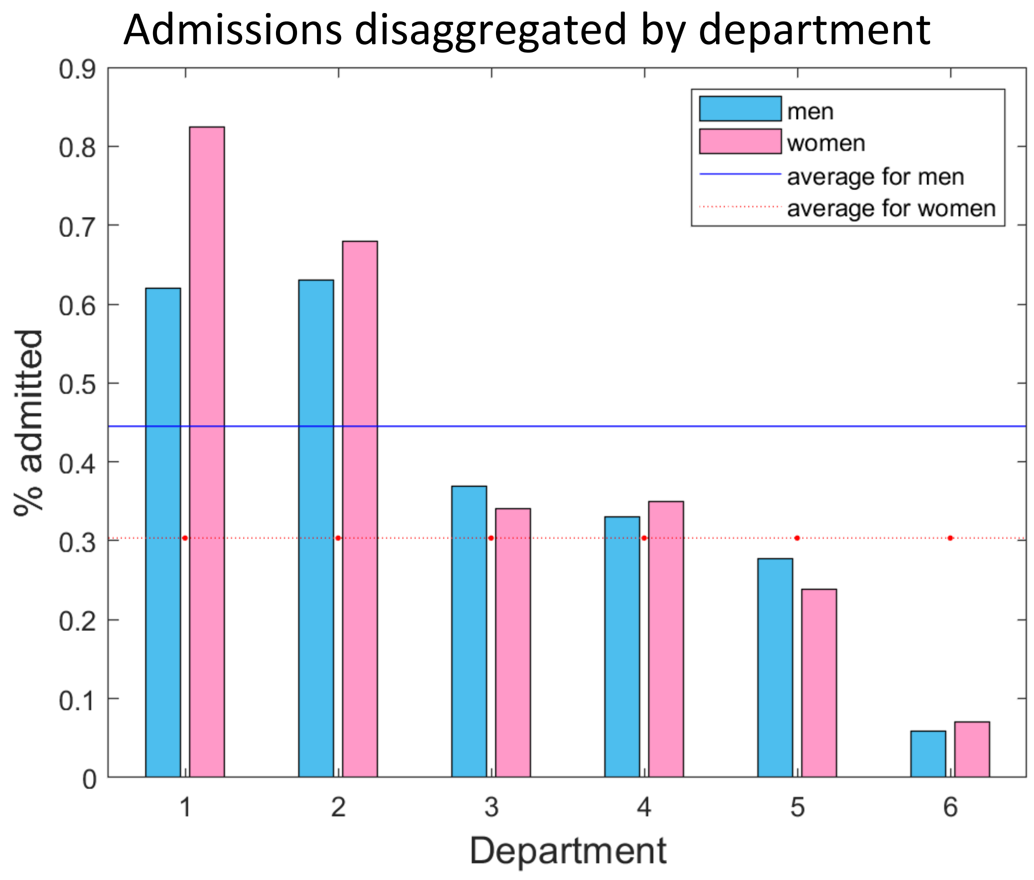
P. J. Bickel, E. A. Hammel, J. W. O'Connell

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

43% men

35% wome n

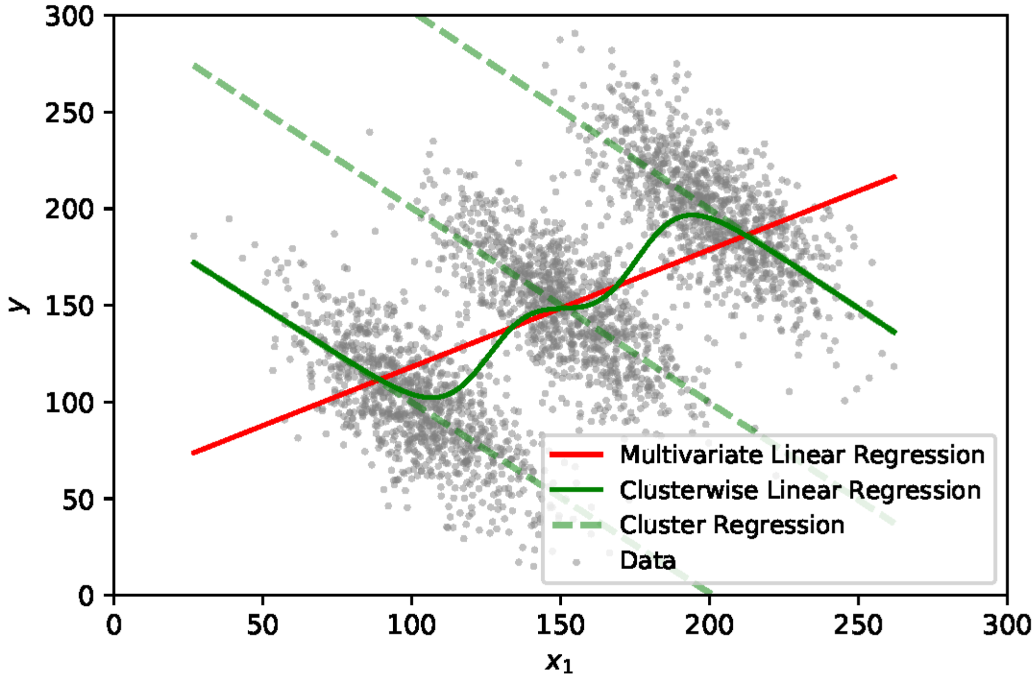
Percent of applicants admitted for graduate study



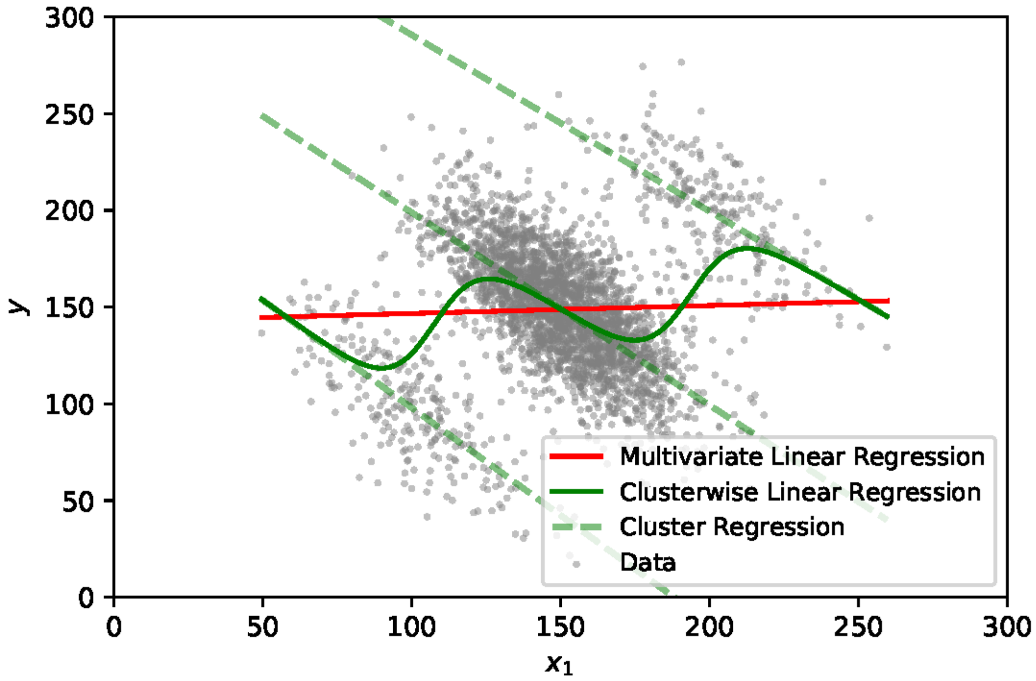
Why does the bias arise? More women apply to highly selective departments

Sampling bias

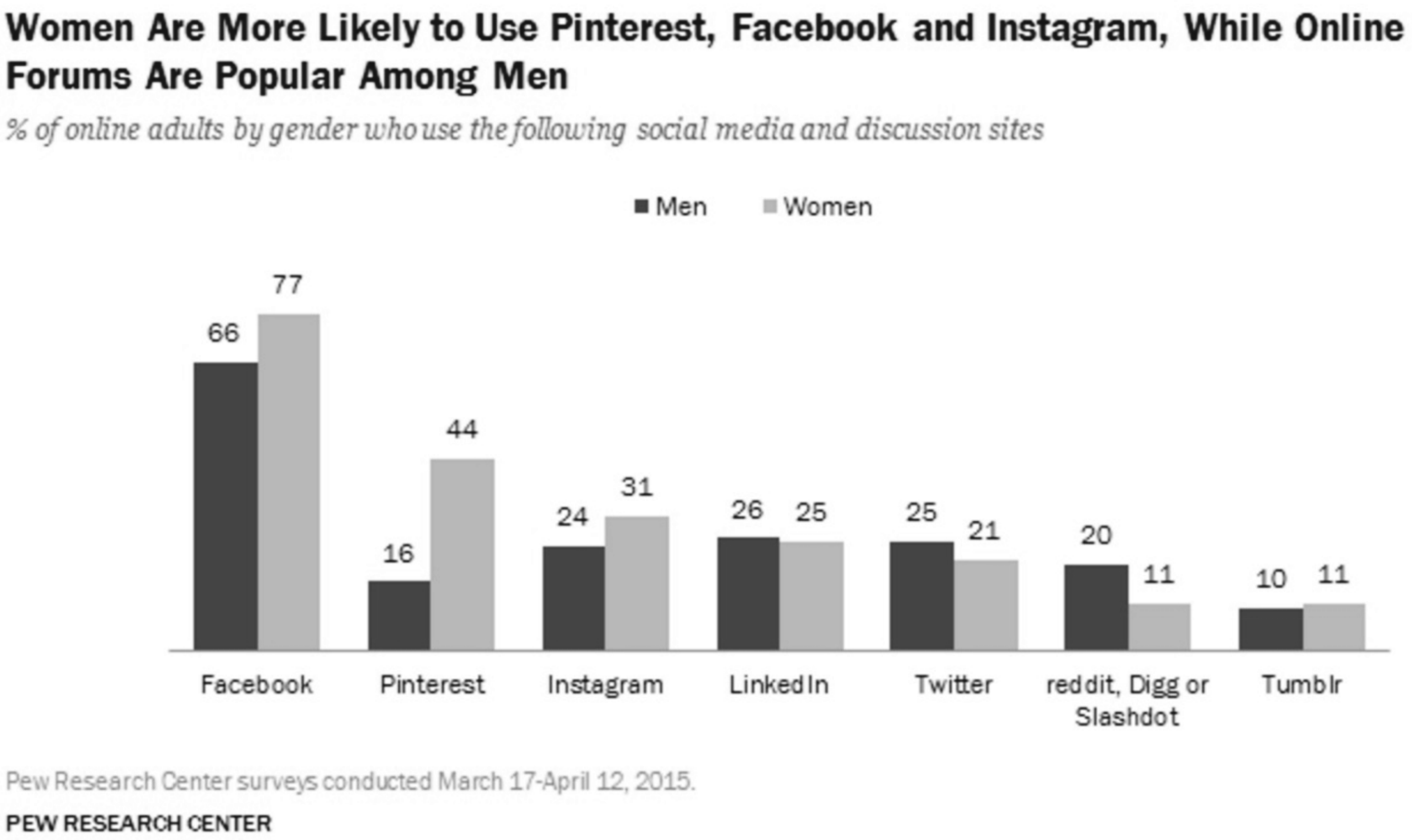
Subgroups uniformly represented in population



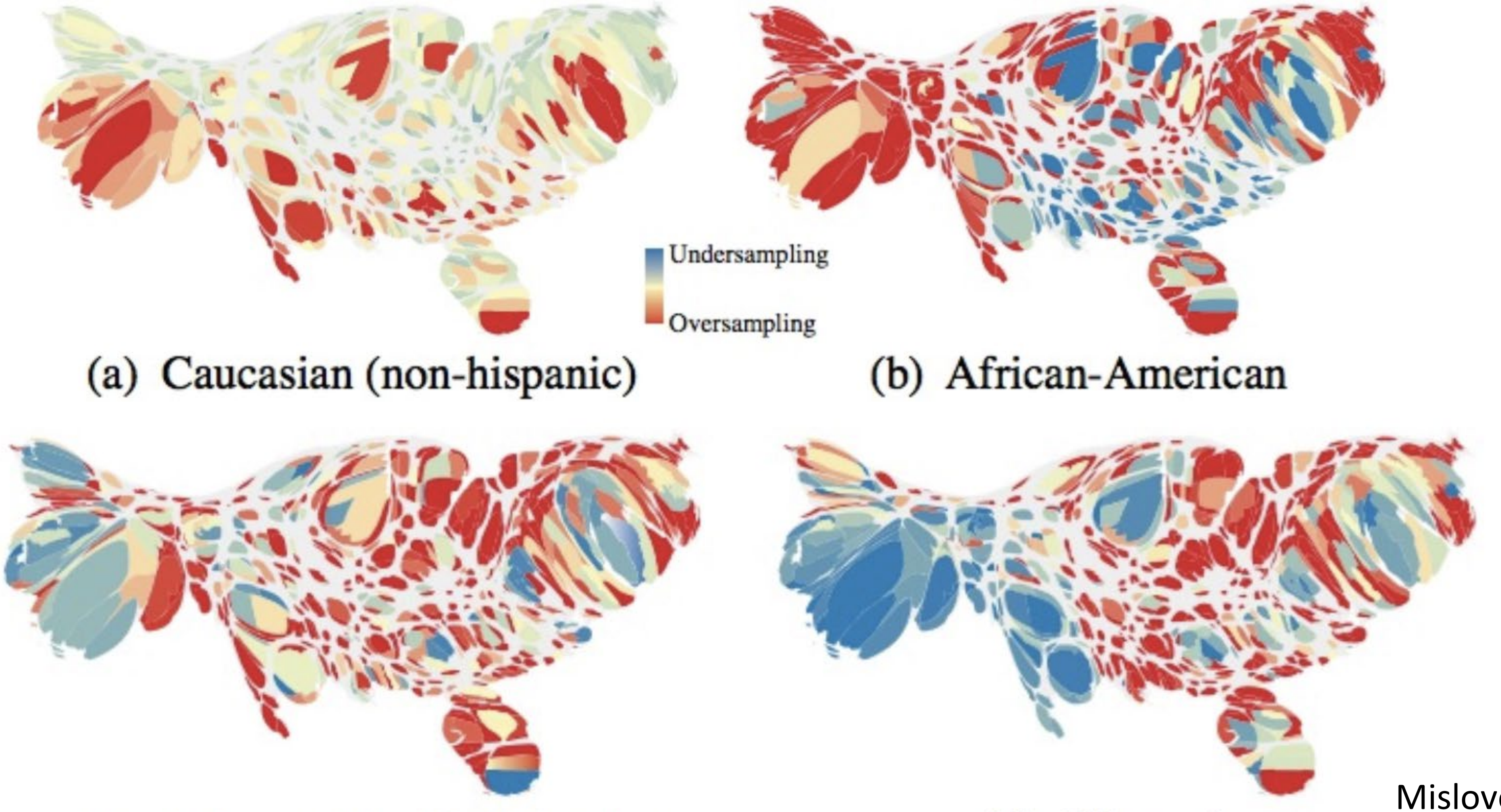
Subgroups overrepresented in the population



Sampling bias due to self-selection



Sampling bias: demographic variation of Twitter users

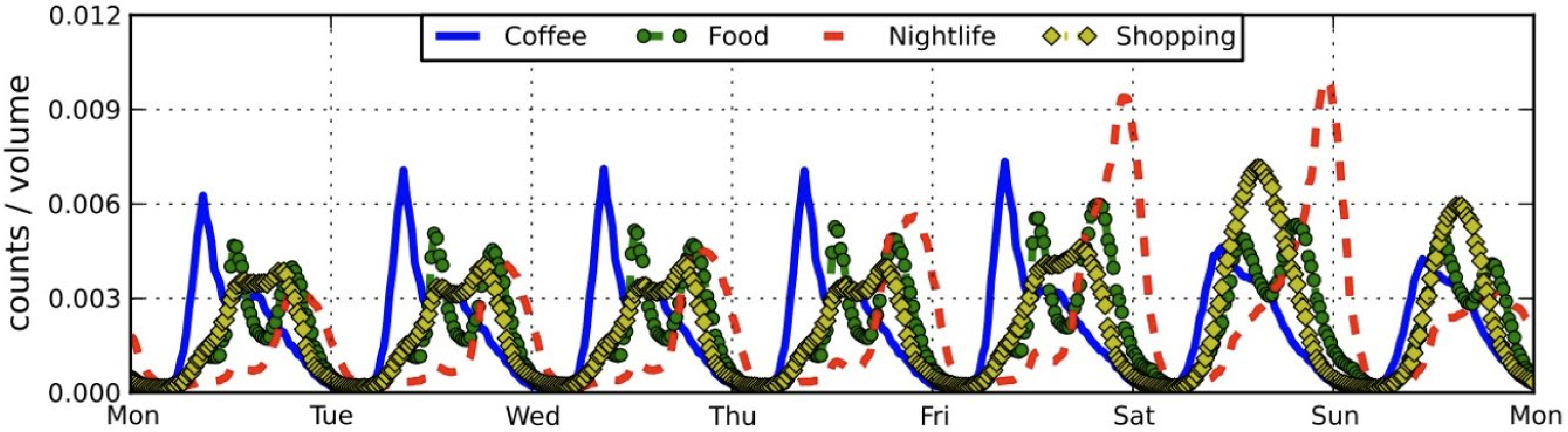


(a) Caucasian (non-hispanic)   (b) African-American

(c) Asian or Pacific Islander   (d) Hispanic

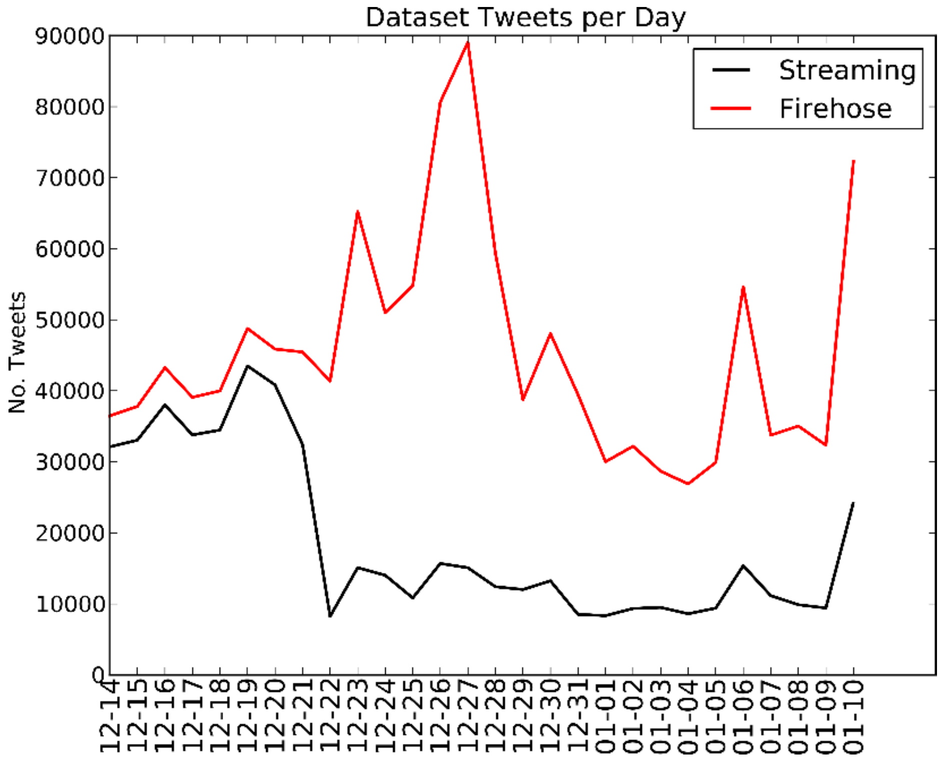
Mislove et al. 2011

Sampling bias: Seasonality affects temporal data



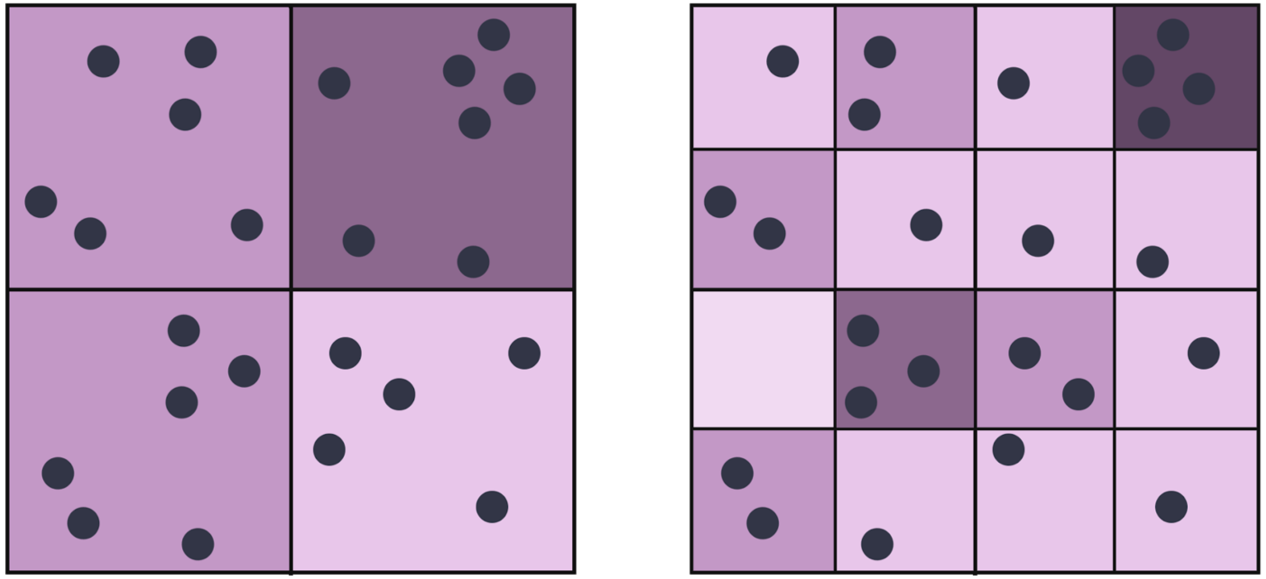
Filtering bias

- Platforms share their data through the API. But, API may not return a representative sample of data.
  - Twitter "Firehose" – all tweets, but costly.
  - "Gardenhose API" - 1% - free.
    - Takes no parameters from users.
    - Returns a random 1% sample.
  - "Streaming API" - 1% - free.
    - Takes query parameters from user.
    - Returns tweets matching query.
    - Samples data when volume reaches 1%.



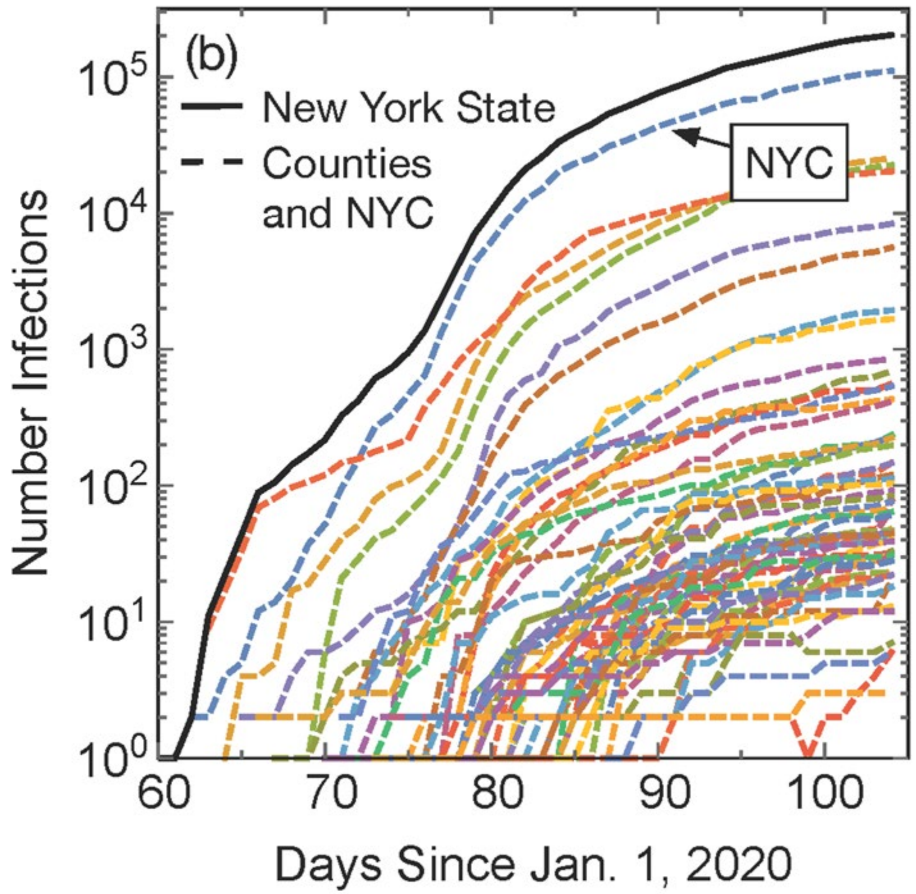
Aggregation bias

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

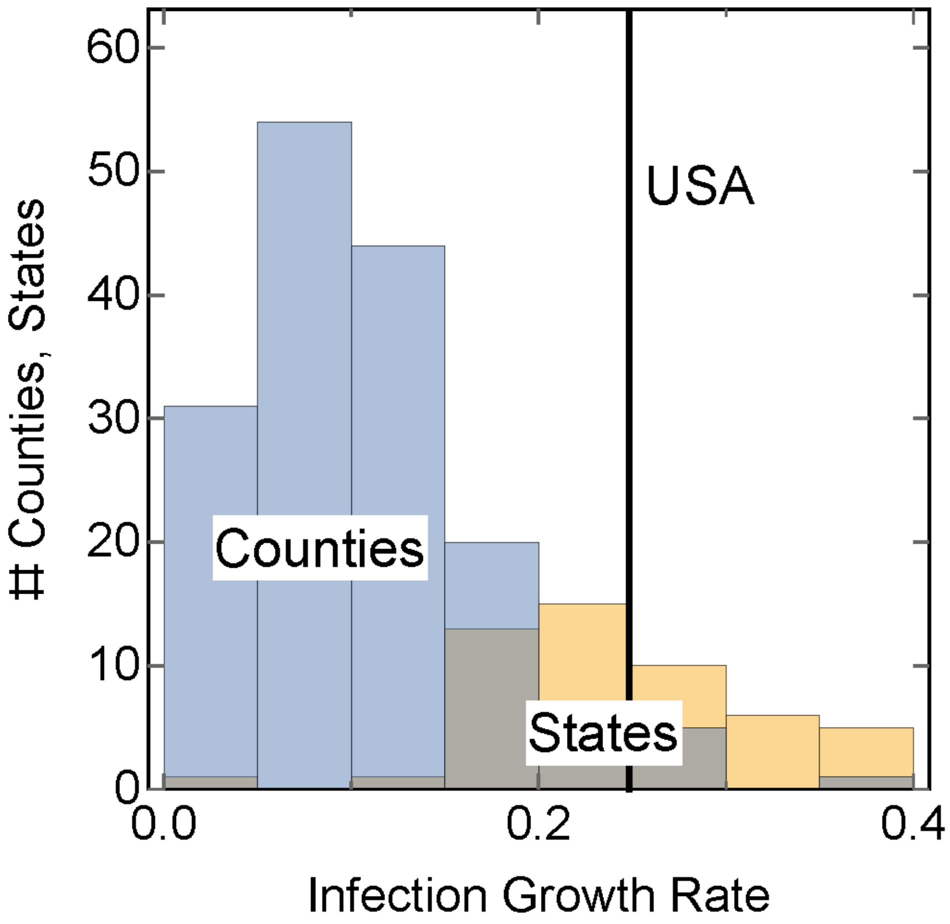


Aggregation bias

Growth of Covid-19 infections in US counties

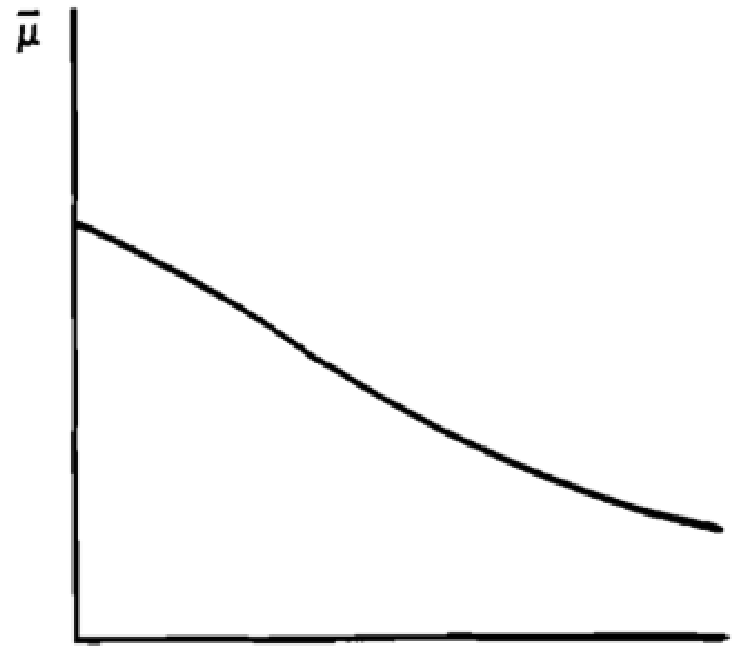


… appears slower than when the same data is aggregated by state



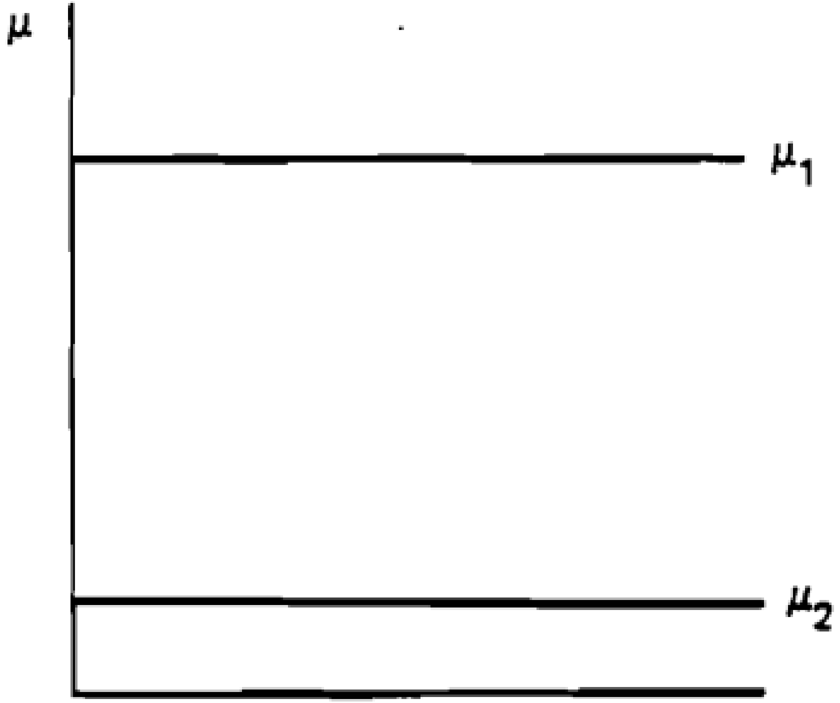
Survivor bias

Recidivism rate of convicts released from prison declines with time since release



Time

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

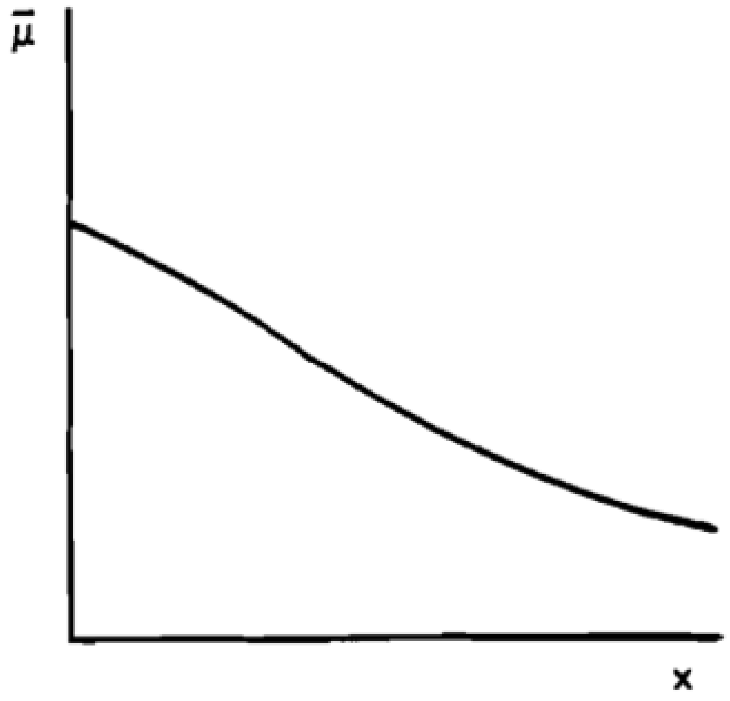


Time

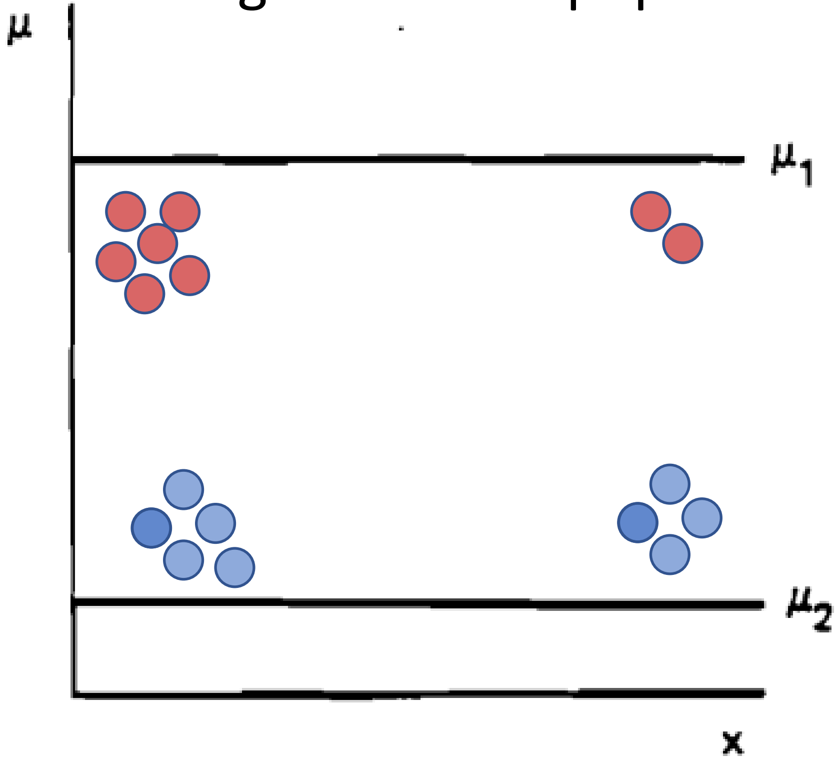
[Vaupel, J. W. and Yashin, A. I. (1985). Heterogeneity's ruses: some surprising effects of selection on population dynamics. *The American Statistician*, 39(3):176-185.]

Survivor bias

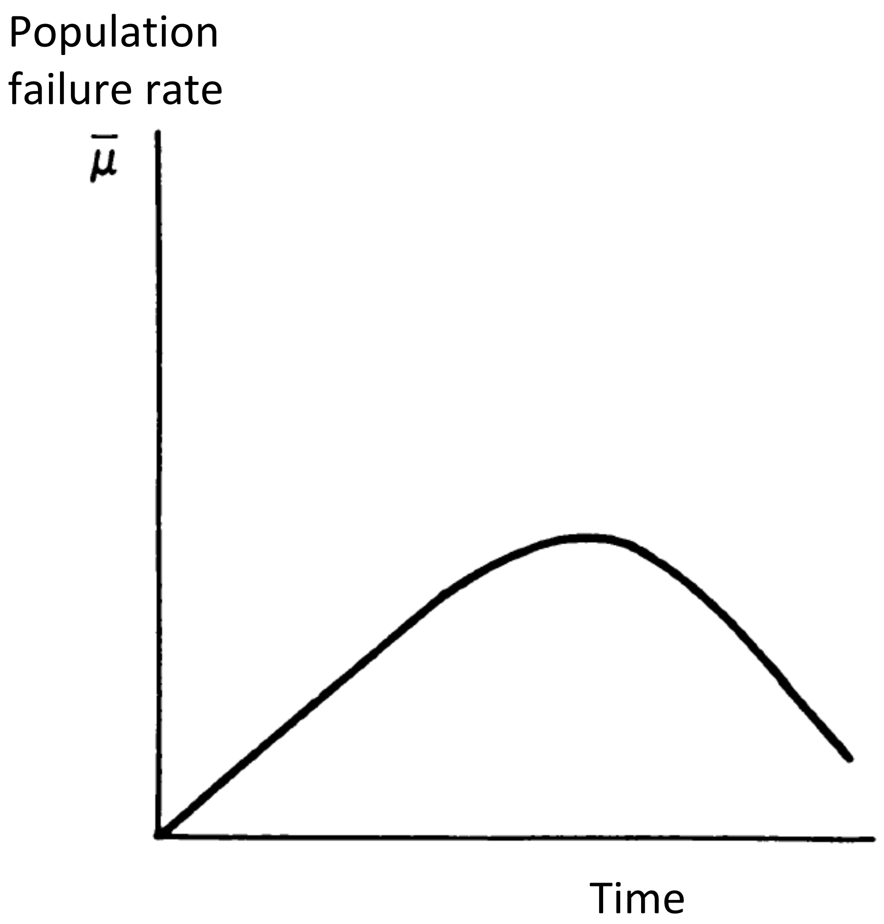
Recidivism rate of convicts released from prison declines with age

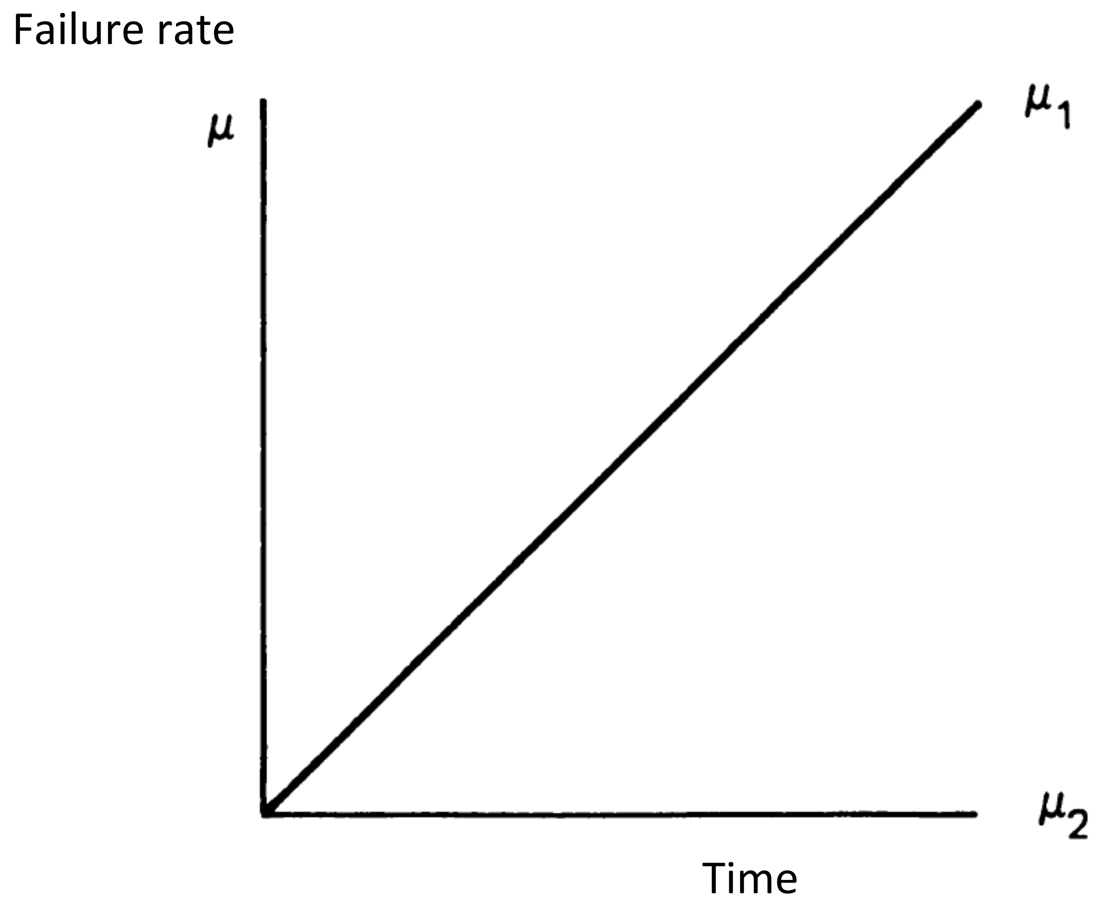


In reality, two subgroups: incorrigibles and reformed. Over time, fewer incorrigibles in the population



[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.]



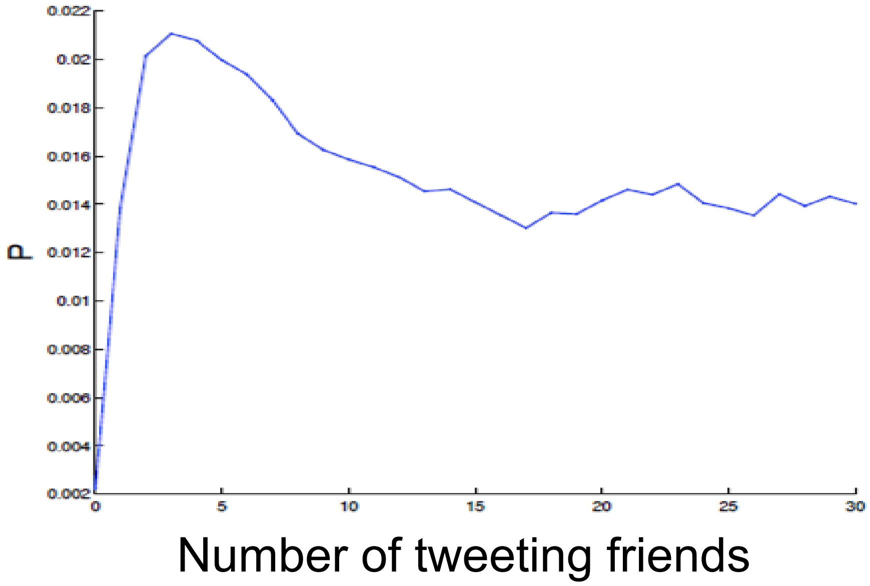


- Observed failure rate for the entire population may rise and then fall
- E.g., divorce rates follow this pattern, but this does not imply that marriage is more likely to fail after the first few years. In reality, for one group marriage strengthens with duration, and for the other, it weakens

[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.]

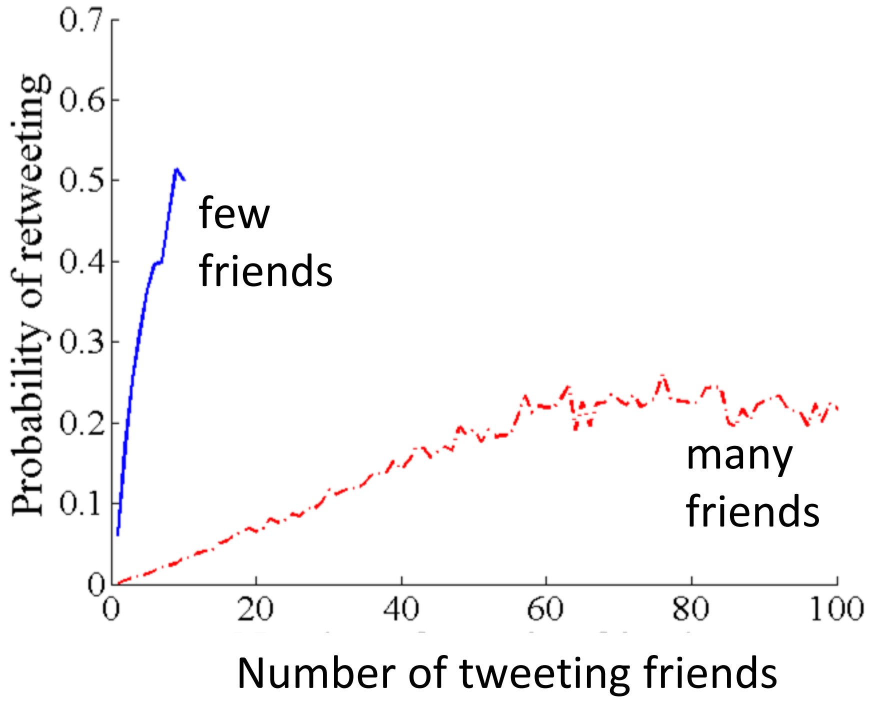
Survivor bias

Probability to retweet after x exposures by friends.
(peak is an artifact)



[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)
(high degree users less susceptible)

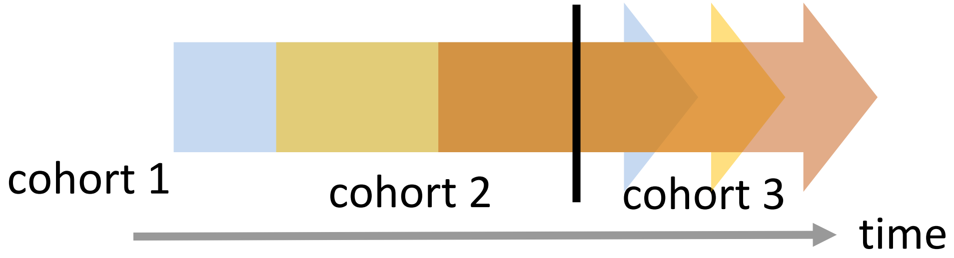


[Hodas & Lerman (2012) “How visibility and divided attention constrain social contagion”, in SocialCom.]

Longitudinal fallacy

- **Longitudinal analysis** gathers data for the same subjects repeatedly over a period of time.

- A **cross-sectional study** is a type of observational study that analyzes data from a population at a specific point in time.

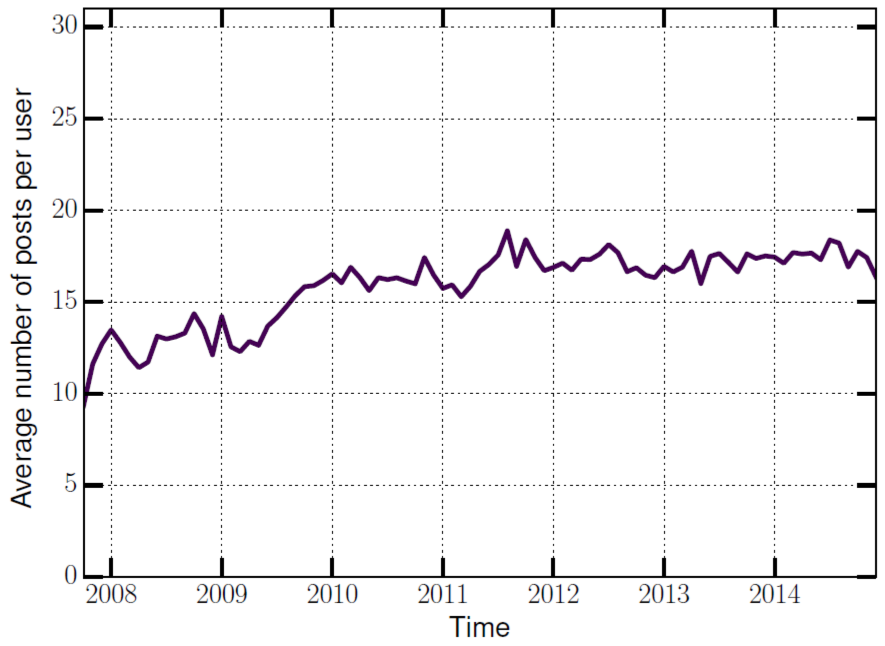


- A **cohort** is a group of people who share a common characteristic, generally with respect to time
  - E.g., USC class of 2023, people born in 1990, etc.

Reddit: Activity over time

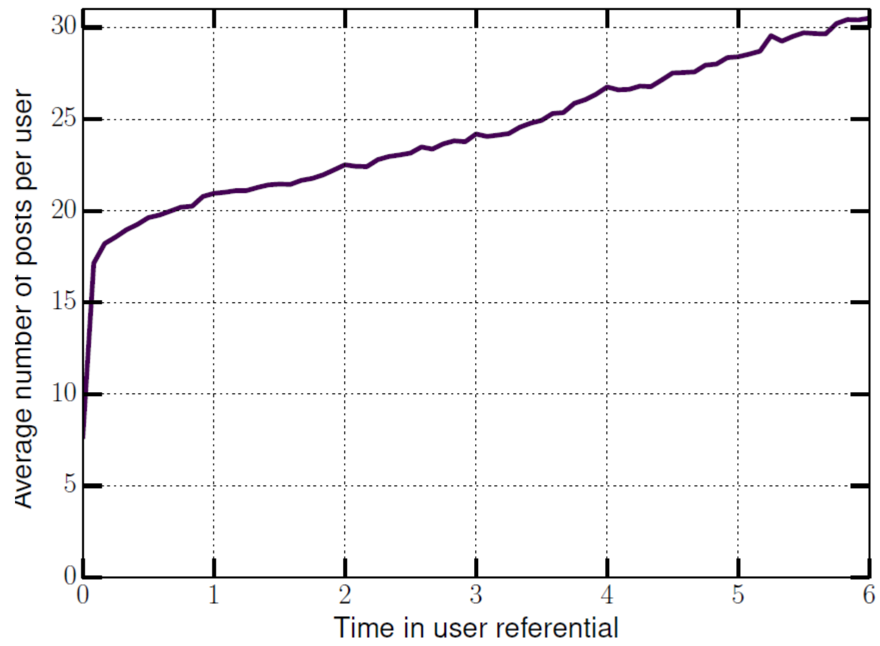
Activity over time:

Users may be becoming more active over time



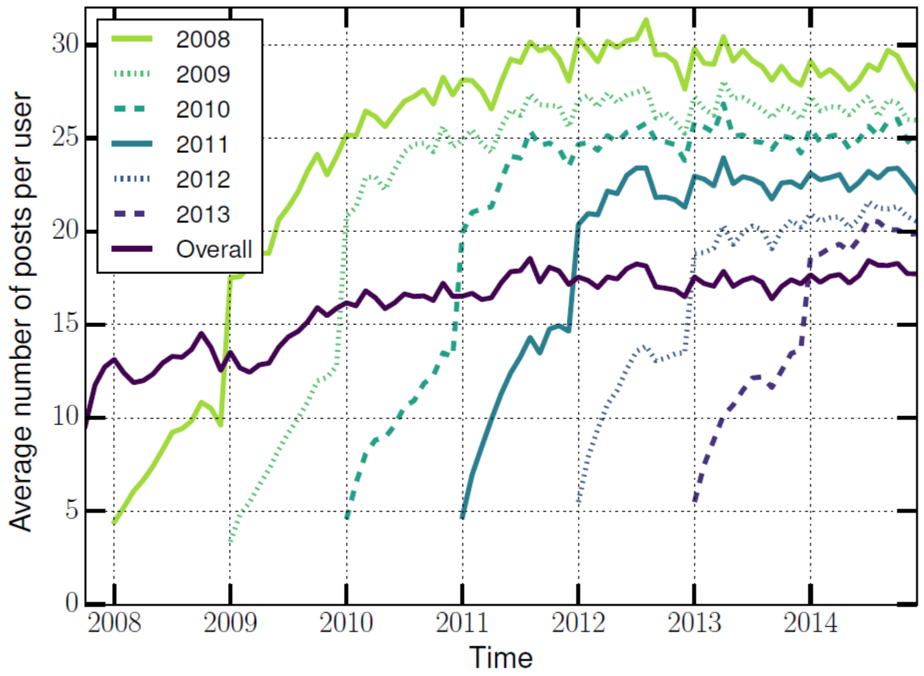
Activity with respect to user tenure:

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

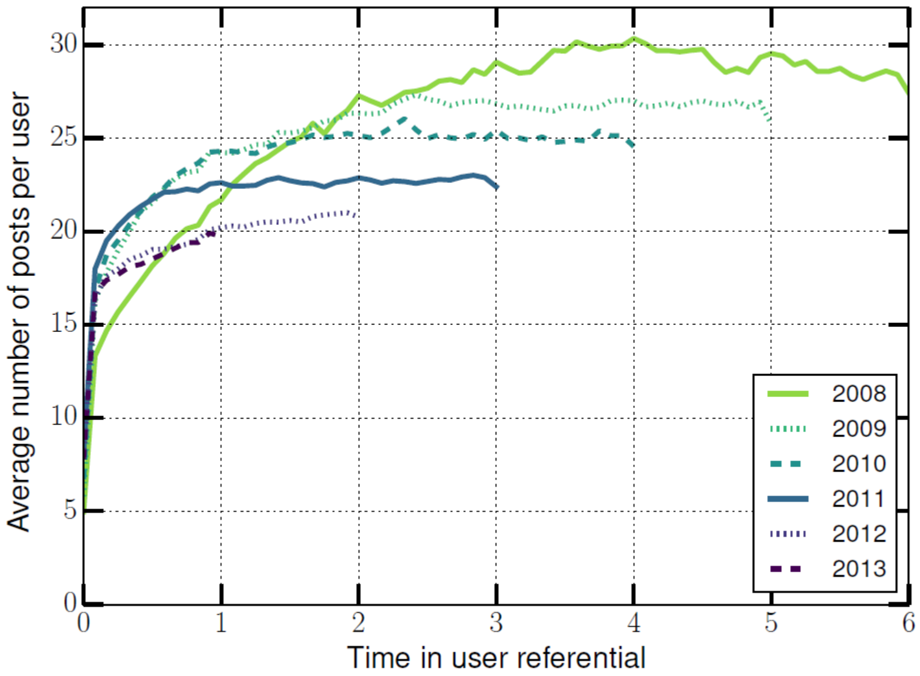


New cohorts do not catch up

User activity split by join date

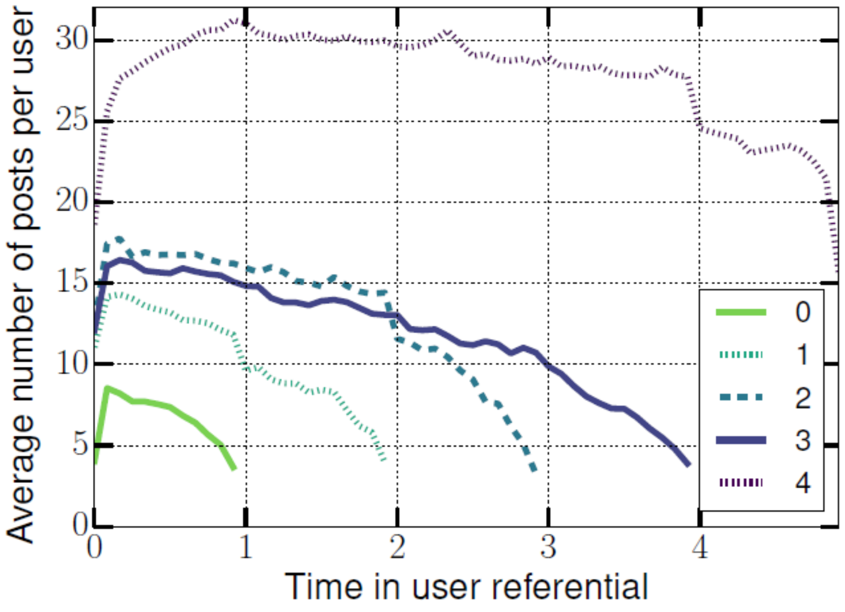


… and aligned

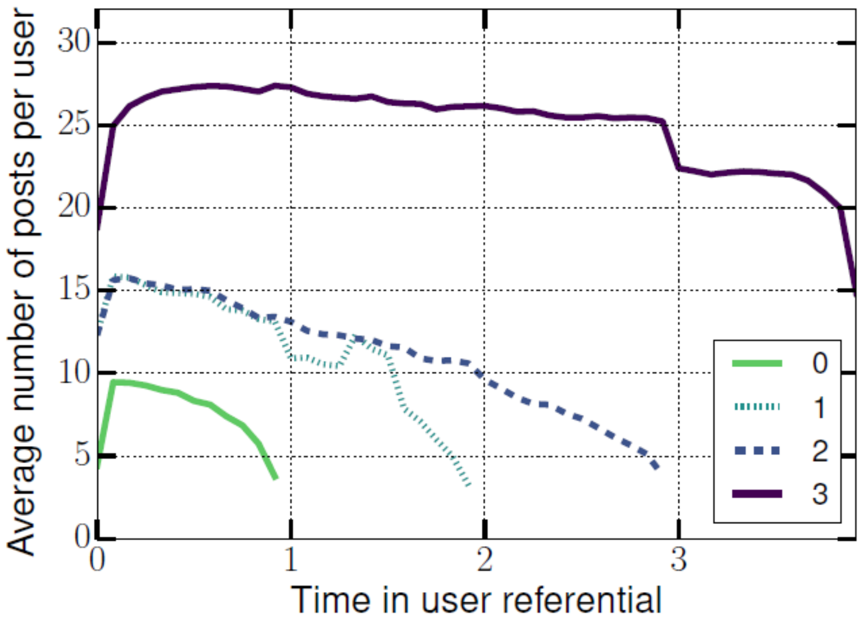


Does tenure predict activity or vice versa?

2010 cohort – low activity users more likely to leave



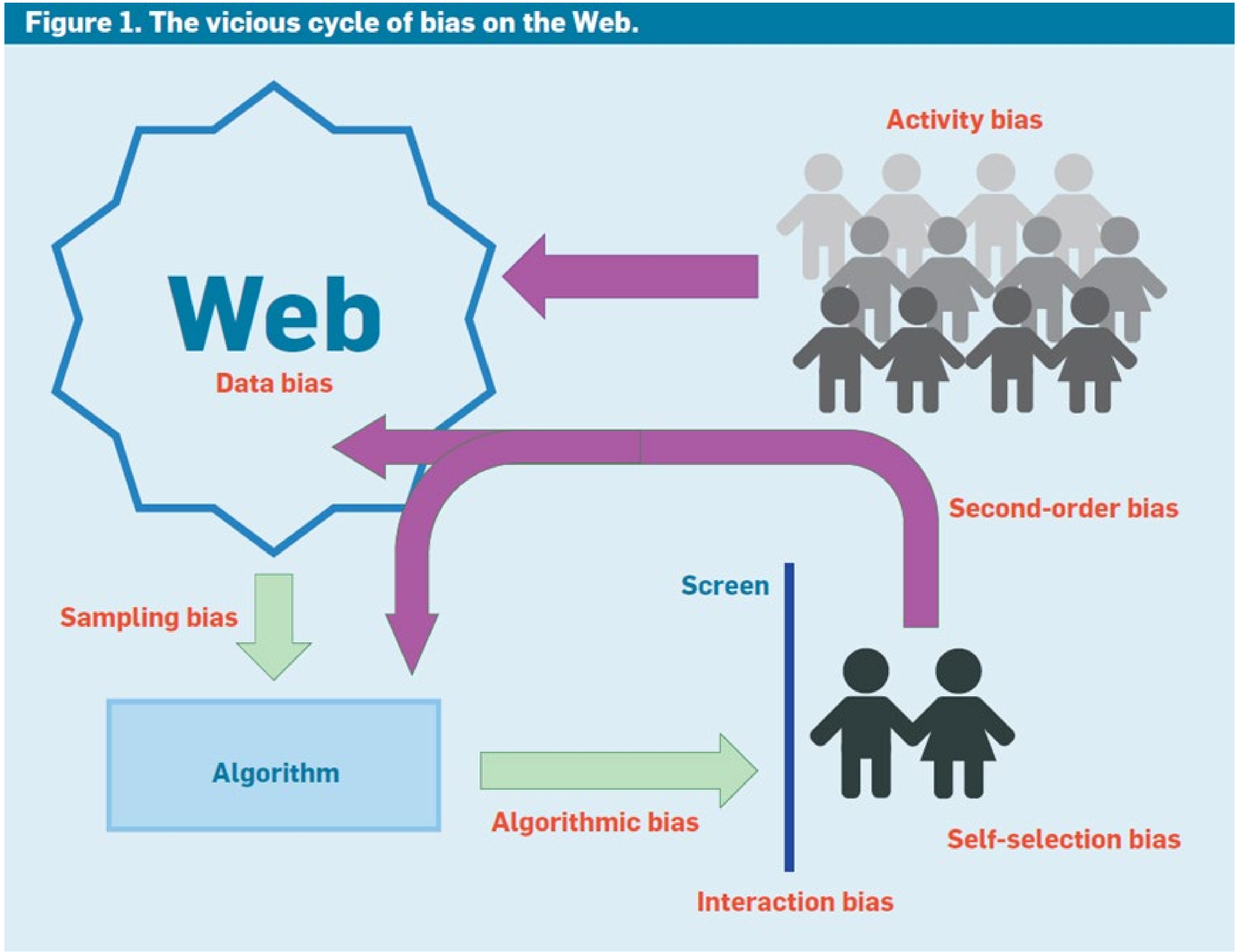
2011 cohort



Lessons learned

- Regardless of how many features are considered, individuals will differ along neglected dimensions. Some of these differences affect the individual outcomes (death, marriage, unemployment, etc).
- Because of this heterogeneity, selection will occur: the remaining (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 trends at the population level will deviate from the underlying trends at the individual level.

Interaction between biases amplifies them



[Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, *61*(6), 54-61.]

Biases in the user interface

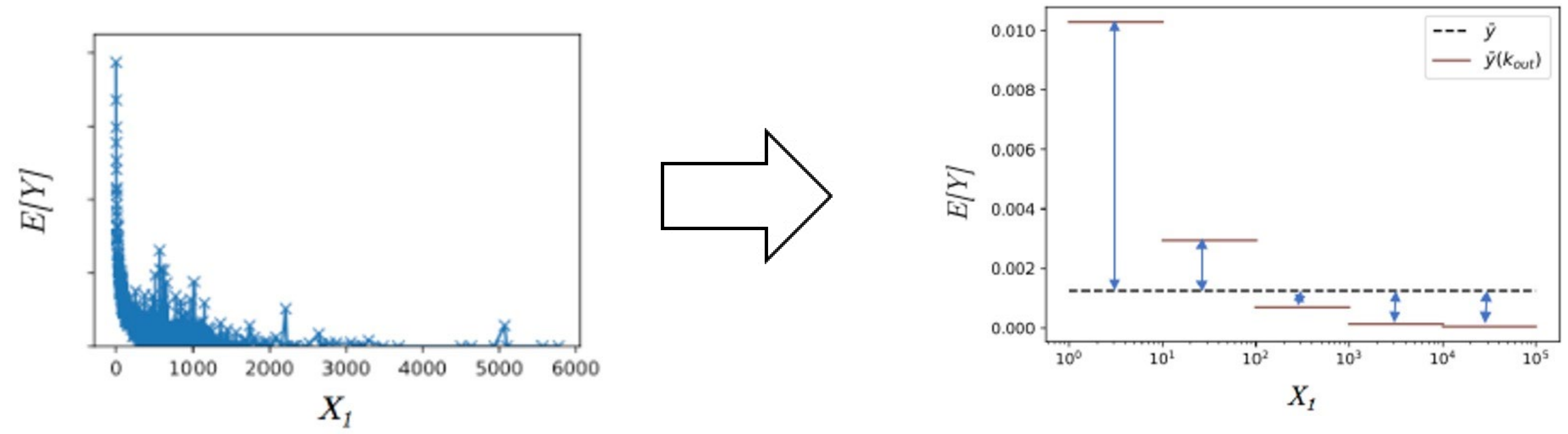
Reducing bias in data

- 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 b
- R2 measure

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



Automated discovery of Simpson's paradoxes

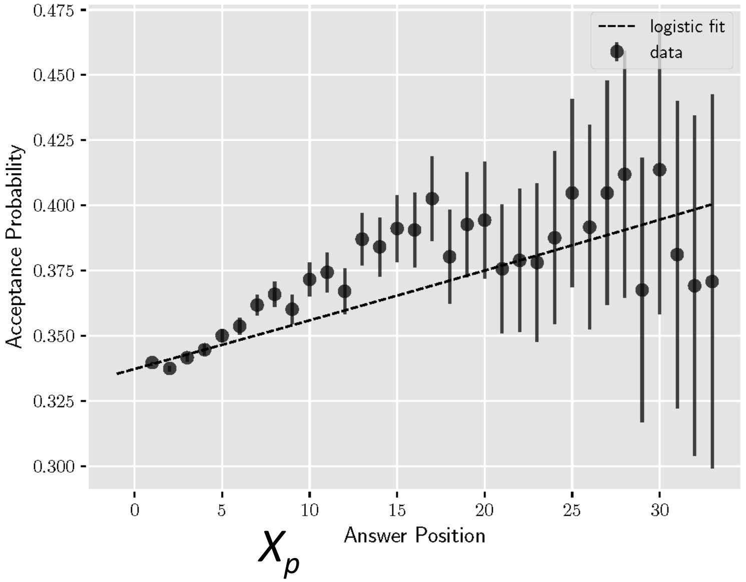
- Method to systematically disaggregate data into homogeneous subgroups
- Identify functional differences between subgroups and pooled data
  - "Using Simpson's paradox to discover interesting behavioral patterns in data" in *ICWSM* 2018
  - "Can you Trust the Trend? Discovering Simpson's Paradoxes in Social Data" in *WSDM* 2018

METHOD

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

Automated discovery of Simpson's paradoxes

1. **Estimate** trend of outcome *Y* with respect to a covariate X_p

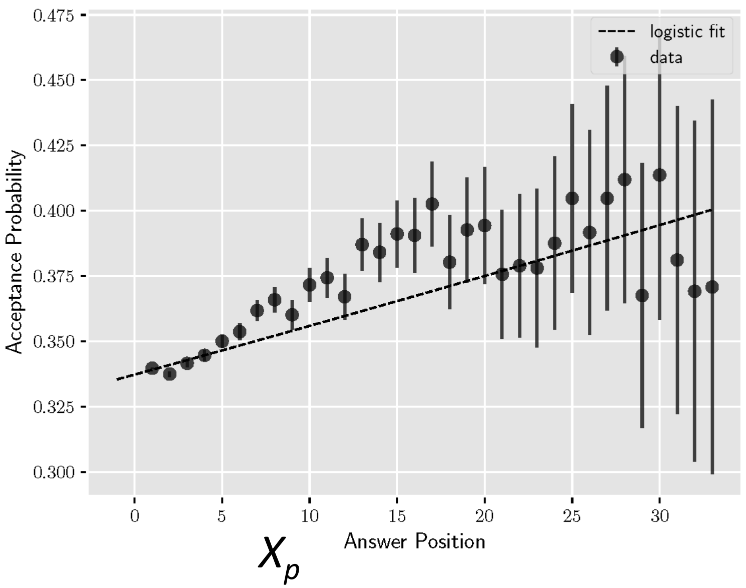


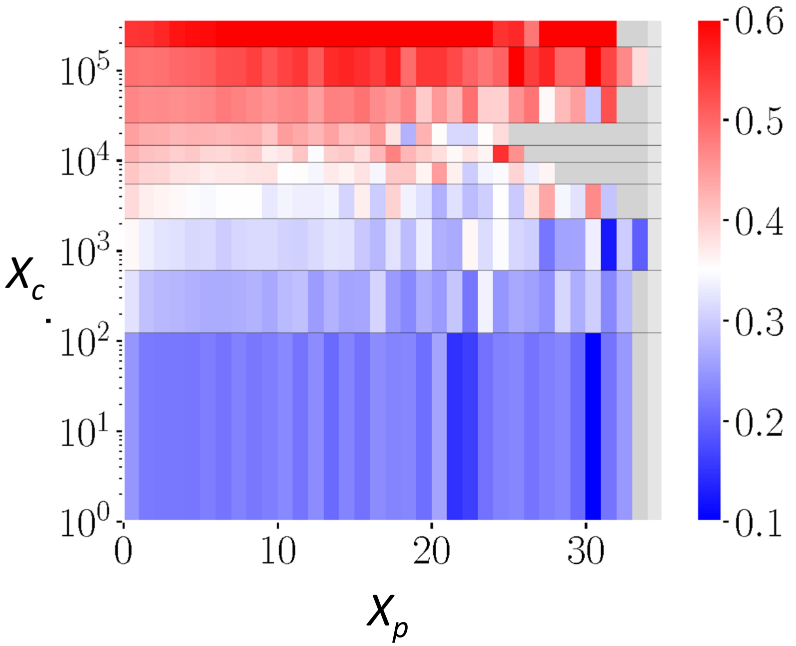
METHOD

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

Automated discovery of Simpson's paradoxes

1. **Estimate** trend of outcome Y with respect to a covariate X_p
2. **Disaggregate** data by conditioning on some other covariate X_c



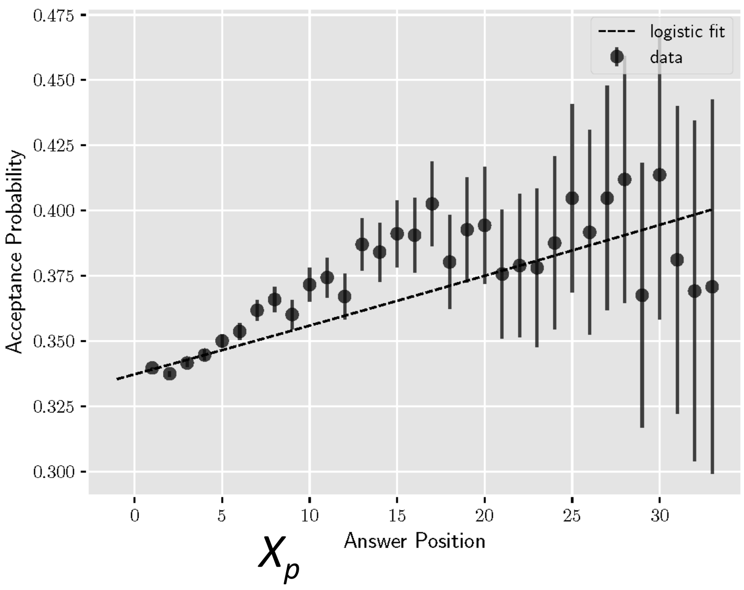


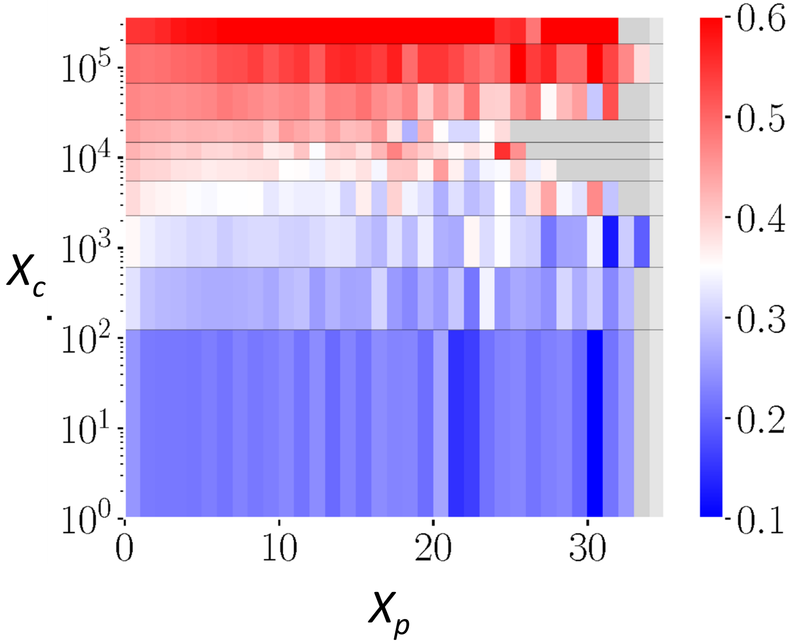
METHOD

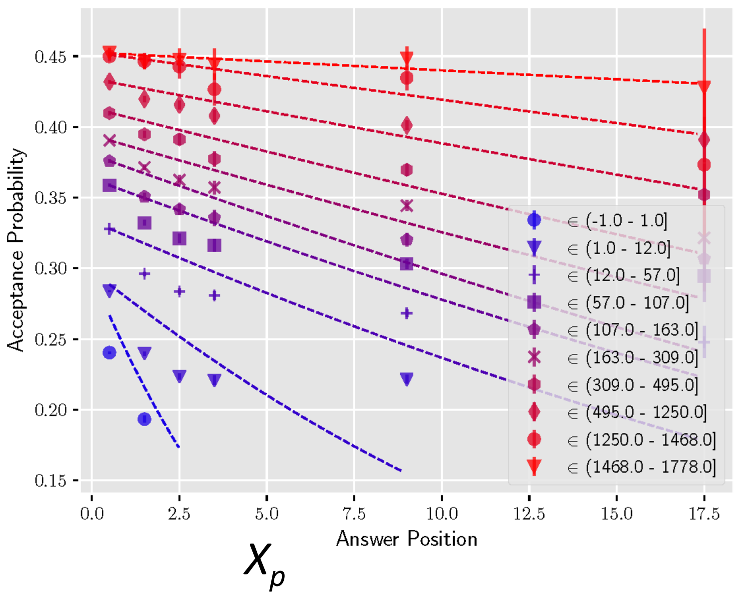
Alipourfard, Fennell & Lerman (2018) *Using Simpson's paradox to discover interesting patterns in data.* ICWSM. *Code:* https://github.com/ninoch/Trend-Simpsons-Paradox

Automated discovery of Simpson's paradoxes

1. **Estimate** trend of outcome *Y* with respect to a covariate X_p
2. **Disaggregate** data by conditioning on some other covariate 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

Simpson's paradox in real-world data

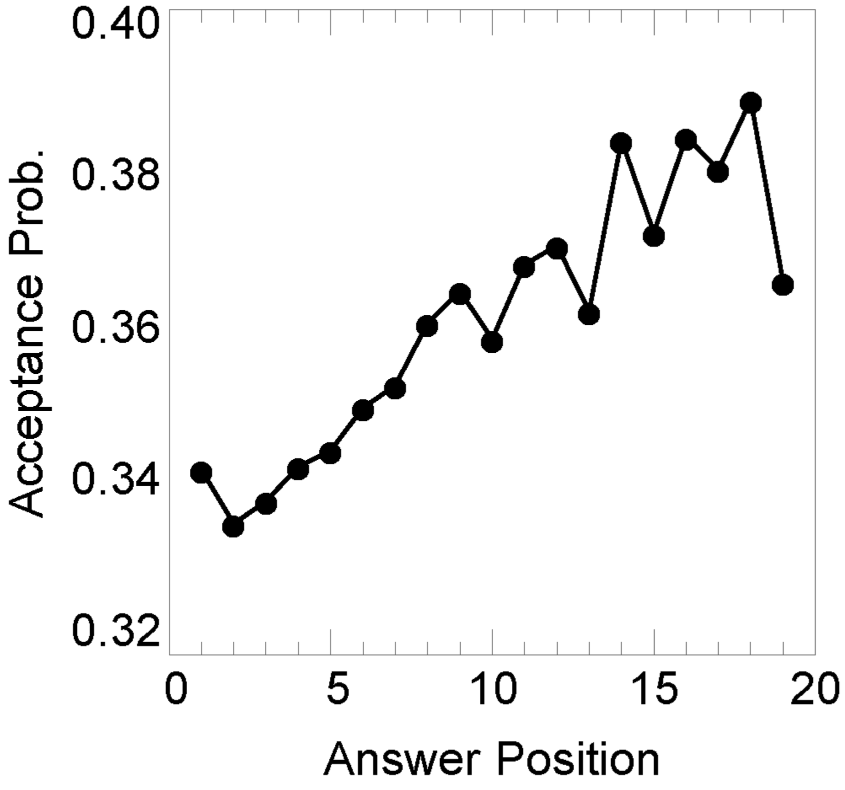
Simpson's reversal provides evidence for cognitive depletion □ The more time people spend online, the worse they perform

- Singer et al. (2016) Evidence of online performance deterioration in user sessions on Reddit, in PLoS One
- Kooti et al (2017) Understanding short-term changes in online activity sessions, in WWW Companion
- Ferrara et al (2017) Dynamics of content quality in collaborative knowledge production, in ICWSM
- Alipourfard et al (2018) Using Simpson's Paradox to Discover Interesting Patterns in Behavioral Data, in ICWSM
- Sapienza et al (2018) Individual performance in team-based online games, in Royal Society Interface
- Hodas et al (2018). Model of cognitive dynamics predicts performance on standardized tests. *JCSS*

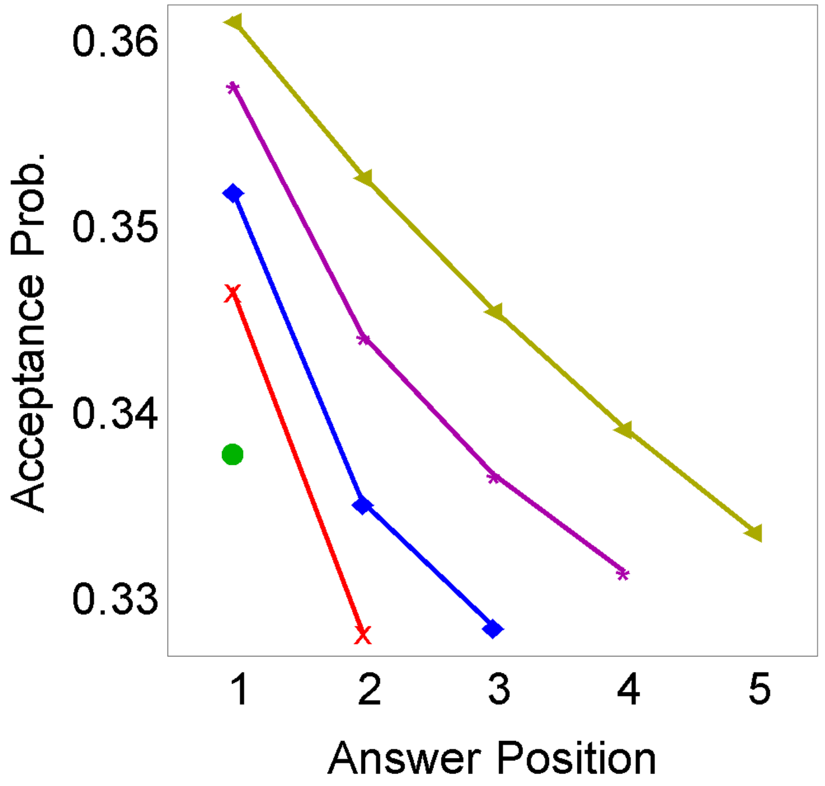
stackoverflow

Disaggregating by session length reveals later answers are worse

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 (less accepted)

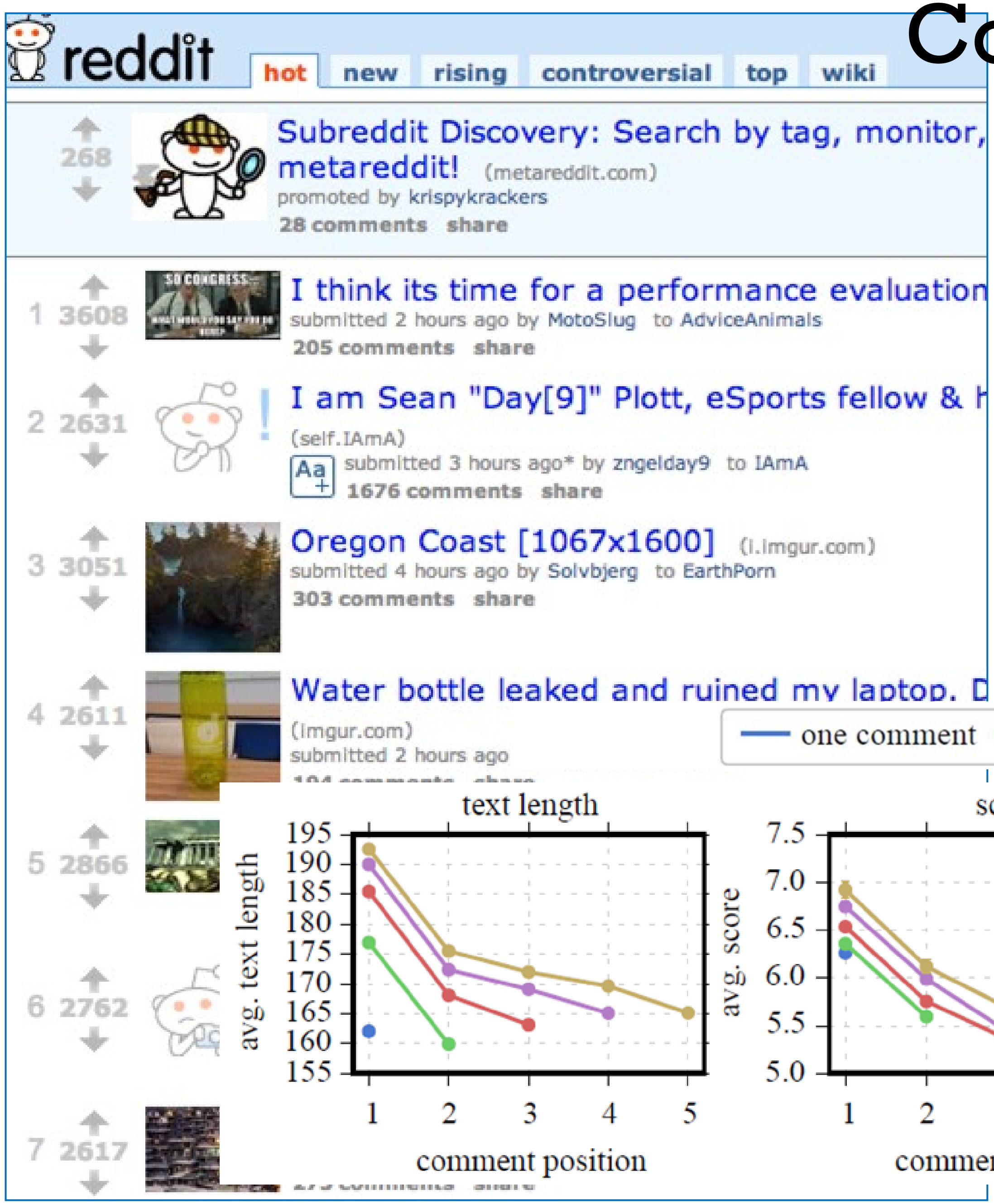


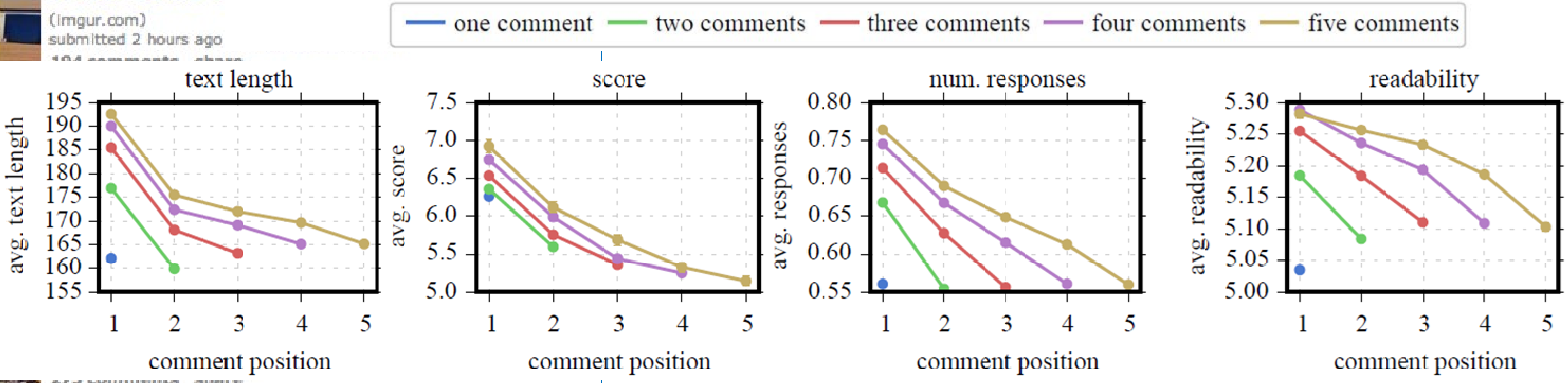
Alipourfard et al (2018) “Using Simpson’s paradox to discover interesting behavioral patterns in data” in *ICWSM* 2018

Cognitive depletion on Reddit

Over the course of a session, each successive comment is

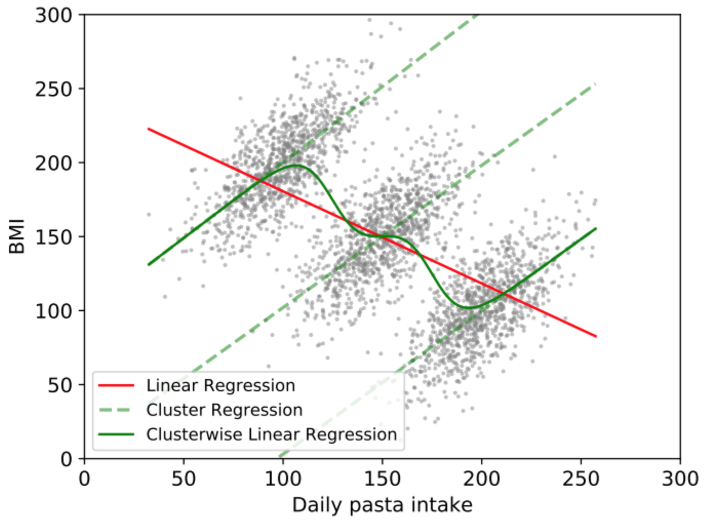
- Shorter
- Receives a lower score
- Receives fewer responses
- Textually less complex





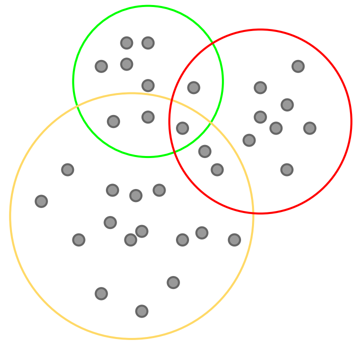
Discovering latent subgroups

- Disaggregate into subgroups to reduce bias
  - What if subgroups are not observed?
- Joint disaggregation + regression
  - Disaggregate data along multiple dimensions into latent subgroups
    - Soft clustering
  - Measure trends within subgroups
    - Regression coefficients within each subgroup



Mixture Models

- A probabilistic model for representing latent (unidentified) **subgroups** within the population.
- Use Gaussian Mixture Models, while also computing **regression coefficients** for subgroups
- When each subgroup has independent regression coefficients, we can study the effect of each independent variable on the outcome of interest



$$Y = a_1X_1 + a_2X_2 + a_3X_3 + \ldots$$

	X_1	X_2	...
Green	1.2	-0.6	...
Red	-2.6	1.2	...
Orange	0.01	3.01	...

Joint Density

- X is independent variable and Y is outcome.
- Each cluster k (f_X) is a Gaussian with mean μ_k and covariance Σ_k.
- Under the assumption of normality of residuals, $f_{Y|X}$ has normal distribution with mean $\hat{Y}^{(k)}$
- Then, the joint density is:

$$f_X^{(k)} \sim \mathcal{N}(\mu_k, \Sigma_k)$$

$$f_{Y|X}^{(k)} \sim \mathcal{N}(\hat{Y}^{(k)}, \sigma_k)$$

$$\Longrightarrow$$

$$\begin{aligned} f_{X,Y}^{(k)}(x, y) &= f_{Y|X}(y|x) f_X(x) \\ &= \varphi(y; \hat{y}^{(k)}, \sigma_k) \varphi(x; \mu_k, \Sigma_k) \end{aligned}$$

While: $\hat{Y}^{(k)} = \beta_{k,0} + \beta_{k,1}X_1 + \beta_{k,2}X_2 + ... + \beta_{k,p}X_p,$

Loss Function

Like GMM the universal joint density is **weighted average** of clusters:

$$f_{X,Y}(x,y) = \sum_{k=1}^{K} \omega_k \times f_{X,Y}^{(k)}(x,y)$$

Then, the **loss** function is:

$$\mathcal{L} = \sum_{i=1}^{N} log\Big(\sum_{k=1}^{K} \omega_k \times f_{X,Y}^{(k)}(x_i, y_i)\Big)$$

While we learn these **parameters** (for each cluster k):

$$\mu_k \quad \Sigma_k \quad \sigma_k \quad \omega_k \quad \beta_{k,0} \quad \beta_{k,1} \quad \dots \quad \beta_{k,p}$$

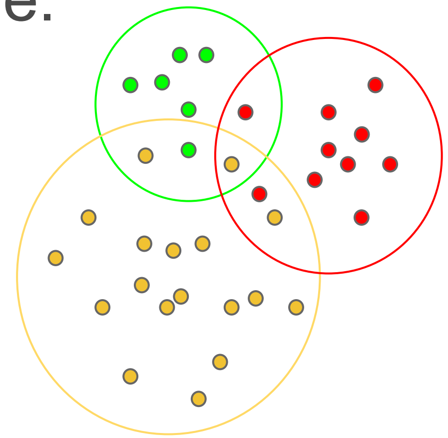
Using **EM-algorithm** and **Weighted Least Squares.**

Soft Clustering

- The membership parameters are the assignment of datapoint i to cluster k. This is a "**soft**" or "**fuzzy**" clustering.

$$\gamma_{i,k} = \frac{\omega_k \times f^{(k)}_{X,Y}(x_i, y_i)}{\sum_{k'} \omega'_k \times f^{(k')}_{X,Y}(x_i, y_i)}$$

- For the analytical purposes, we assign each datapoint to the cluster with the **highest** membership value.



Wine data

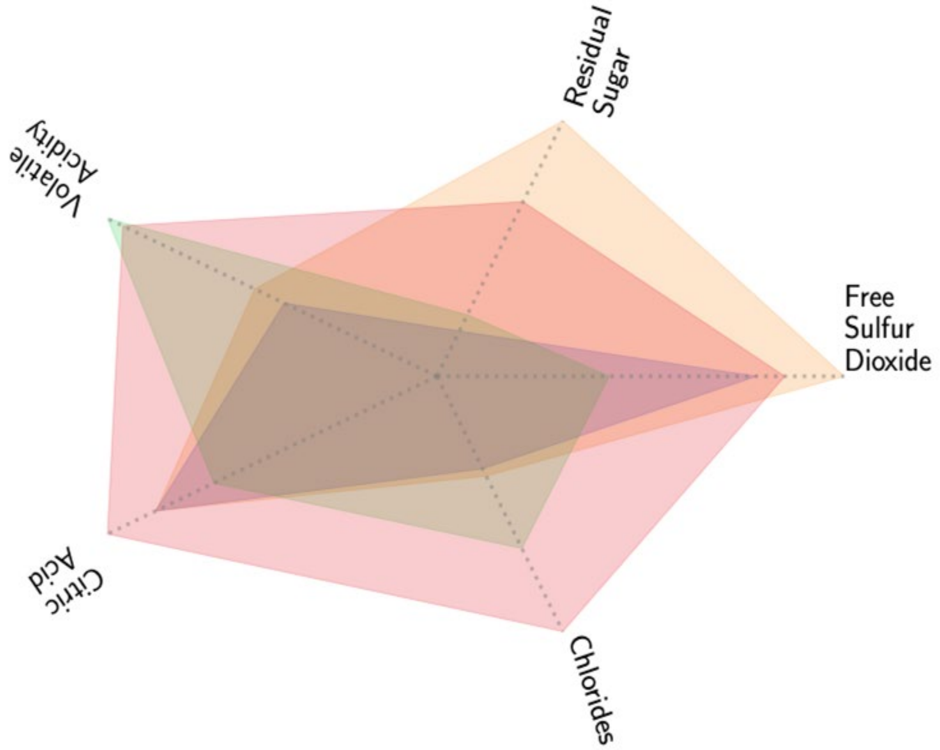
Outcome
- Wine quality

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

data

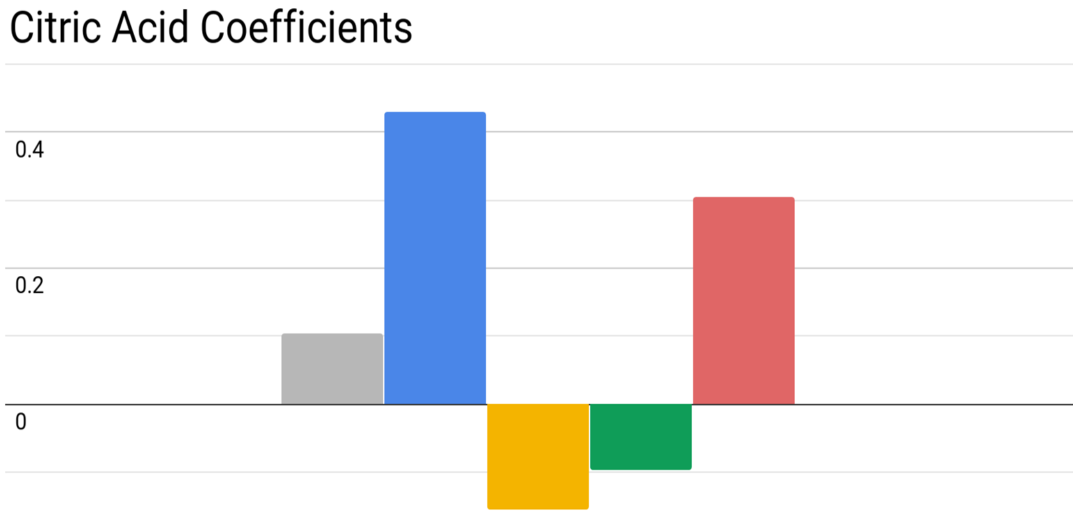
	outcome: quality	free SO2	citric acid	residual sugars	…
White wines	**4898**				
Red wines	**1599**				

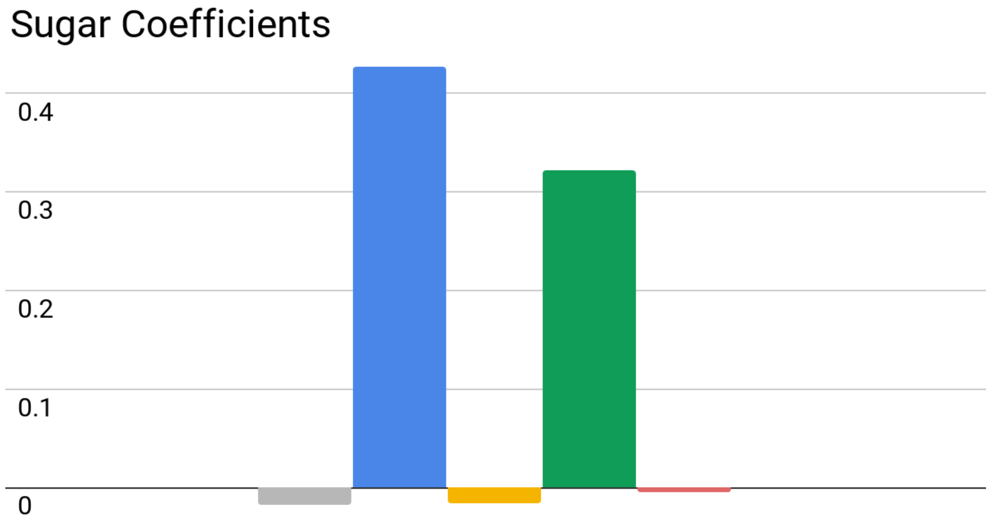
Disaggregation: subgroups in wine data



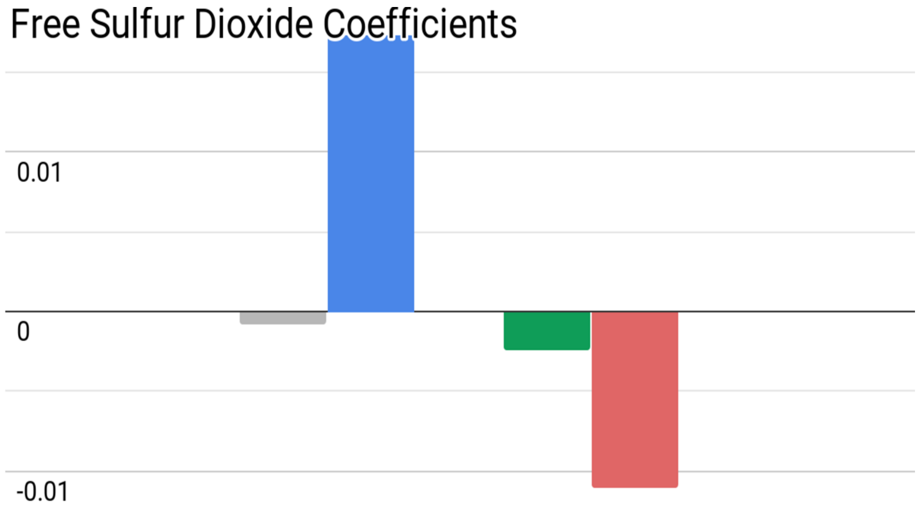
Subgroup	*Composition*	*Ave. quality*	*Comments*
Blue	98% white	6.02	High quality whites
Orange	100% white	5.91	
Green	85% red	5.60	High quality reds
Red	57% red	5.36	Low quality

Regression in subgroups shows trend reversal



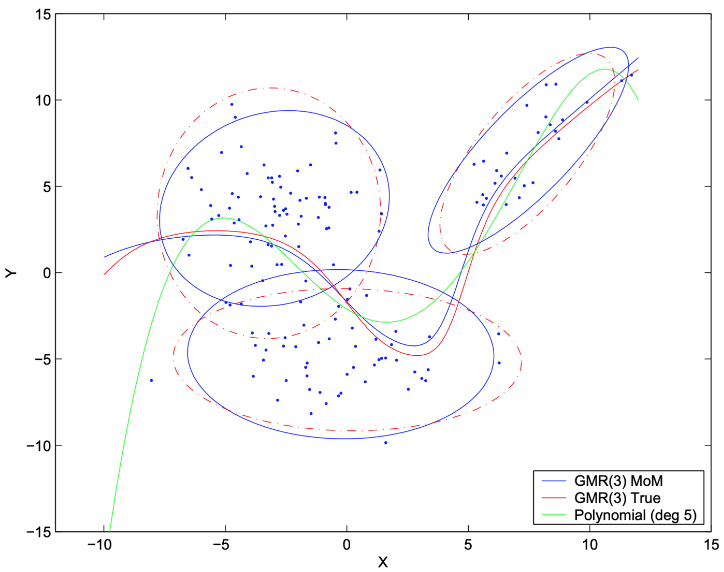


Subgroup	*Composition*	*Ave. quality*
Linear regression	All wines	
Blue	98% white	6.02
Orange	100% white	5.91
Green	85% red	5.60
Red	57% red	5.36



Baselines

- **MLR:** Multivariate Linear Regression
  - No clusters
- **WCLR:** Clusterwise Linear Regression
  - K-means as clustering algorithm
  - Coefficients as regression parameters
- **FWCLR:** Fuzzy Clusterwise Linear Regression
  - K-means as clustering algorithm
  - Coefficients as regression parameters
  - Like GMM, it has soft clustering
- **GMR:** Gaussian Mixture Regression
  - Gaussian Mixture for clustering
  - Regression slope is determined by covariance matrix:



$$\Sigma_j = \begin{bmatrix} \Sigma_{jX} & \Sigma_{jXY} \\ \Sigma_{jYX} & \Sigma_{jYY} \end{bmatrix}$$

Prediction Results

- **Prediction:** We use 5x5-fold nested cross validation to train the model on four folds and make predictions on the out-of-sample data in the fifth fold.
- **Evaluation**: Root Mean Square Error (RMSE) & Mean Absolute Error(MAE)

Method	RMSE ($\pm\sigma$)	MAE ($\pm\sigma$)
	Synthetic	
MLR	294.88 ($\pm$ 1.236)*	288.35 ($\pm$ 0.903)*
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)
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)*
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)*
FWCLR	13.14 ($\pm$ 7.643)	1.76 ($\pm$ 0.321)*
DoGR	**11.88 ($\pm$ 9.109)**	**1.40 ($\pm$ 0.222)**
	Stack Overflow	
MLR	60.69 ($\pm$ 1.118)	37.74 ($\pm$ 0.152)
FWCLR	60.47 ($\pm$ 0.960)	37.25 ($\pm$ 0.794)
DoGR	60.68 ($\pm$ 1.298)	37.62 ($\pm$ 0.314)

- Questions?

- Virtual office hour
- https://usc.zoom.us/j/95136500603?pwd=VEJhblhWK25IT2N3RC9FNWk3eTJKQT09