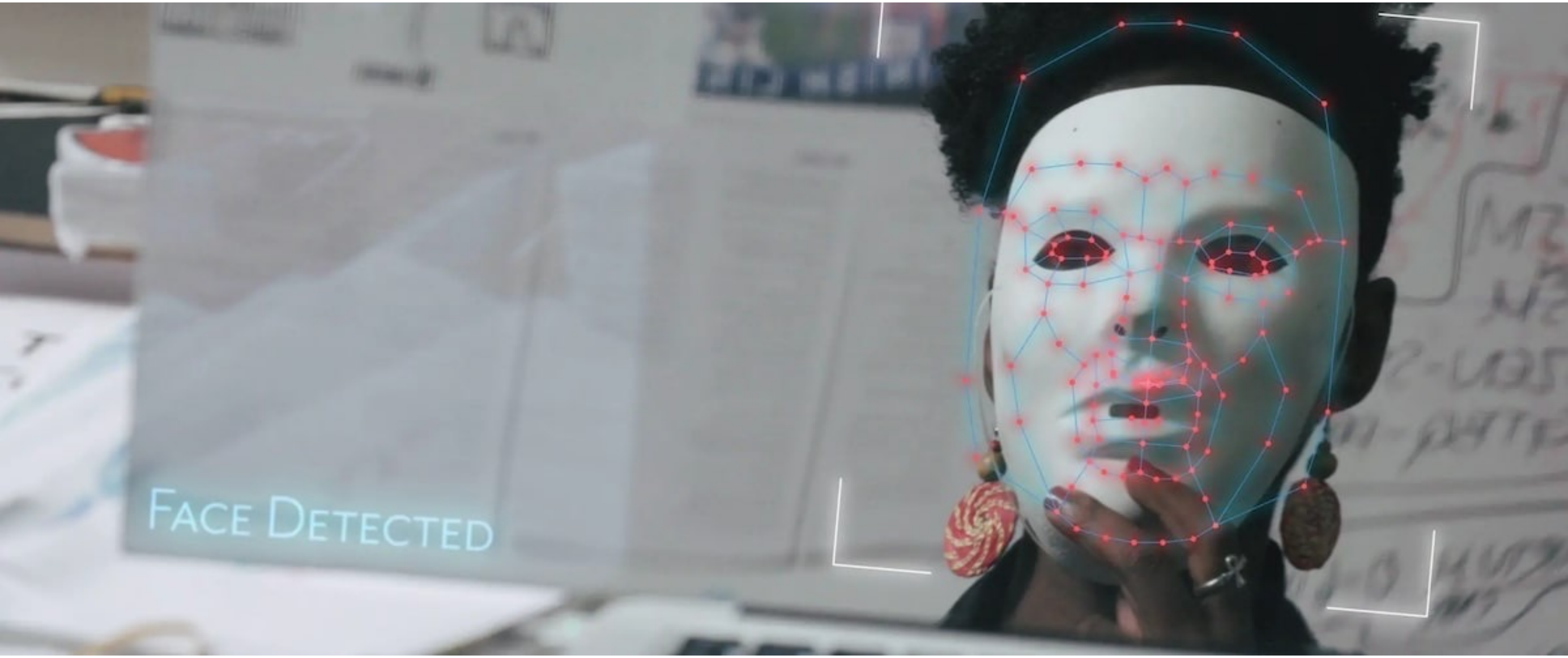
Data provenance

Sinead Williamson
Department of Statistics and Data Science

Coded bias



Kantayya, S. (2020). Coded bias. *7th Empire Media*.
Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAccT.*

Coded bias

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAccT.*

Coded bias

Caroline, C. P. (2019). Invisible Women: Data Bias in a World Designed for Men. *New York, NY: Harry N. Abrams*

Bennardo, M., *et al.* (2016). Day-night dependence of gene expression and inflammatory responses in the remodeling murine heart post-myocardial infarction. Am J Physiol Regul Integr Comp Physiol, *311*(6), R1243–R1254.

Coded bias

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14-19.

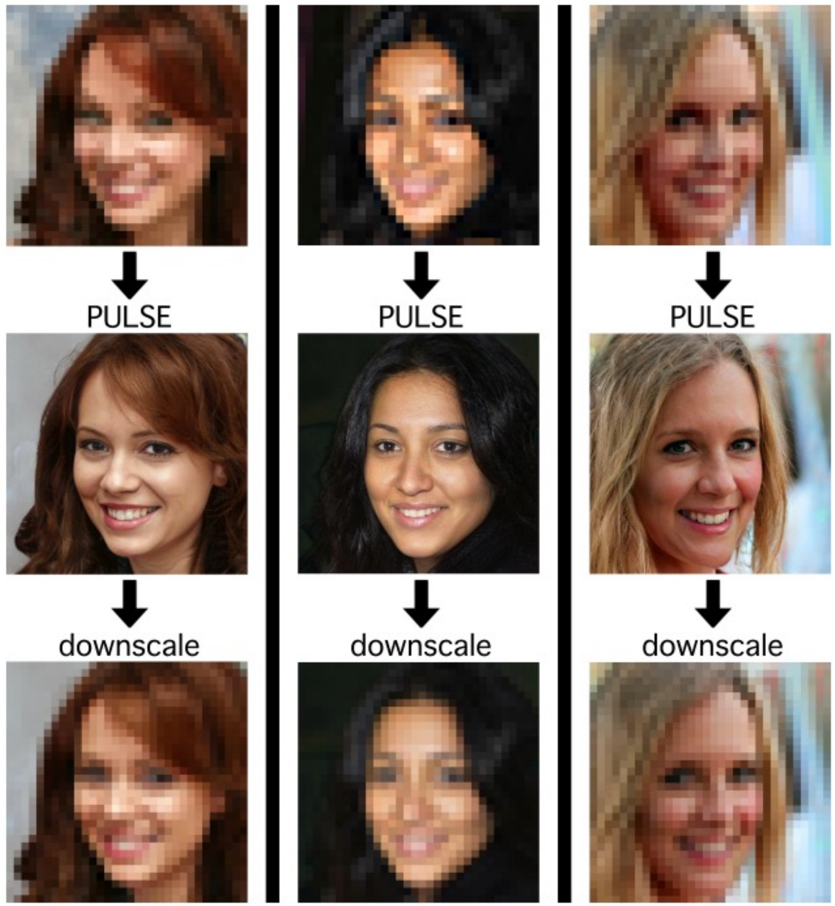
Ensign, D., *et al.* (2018). Runaway feedback loops in predictive policing. In *FAccT* (pp. 160-171).

Akpinar, N. J., *et al*. (2021). The effect of differential victim crime reporting on predictive policing systems. In *FAccT* (pp. 838-849).

Coded bias

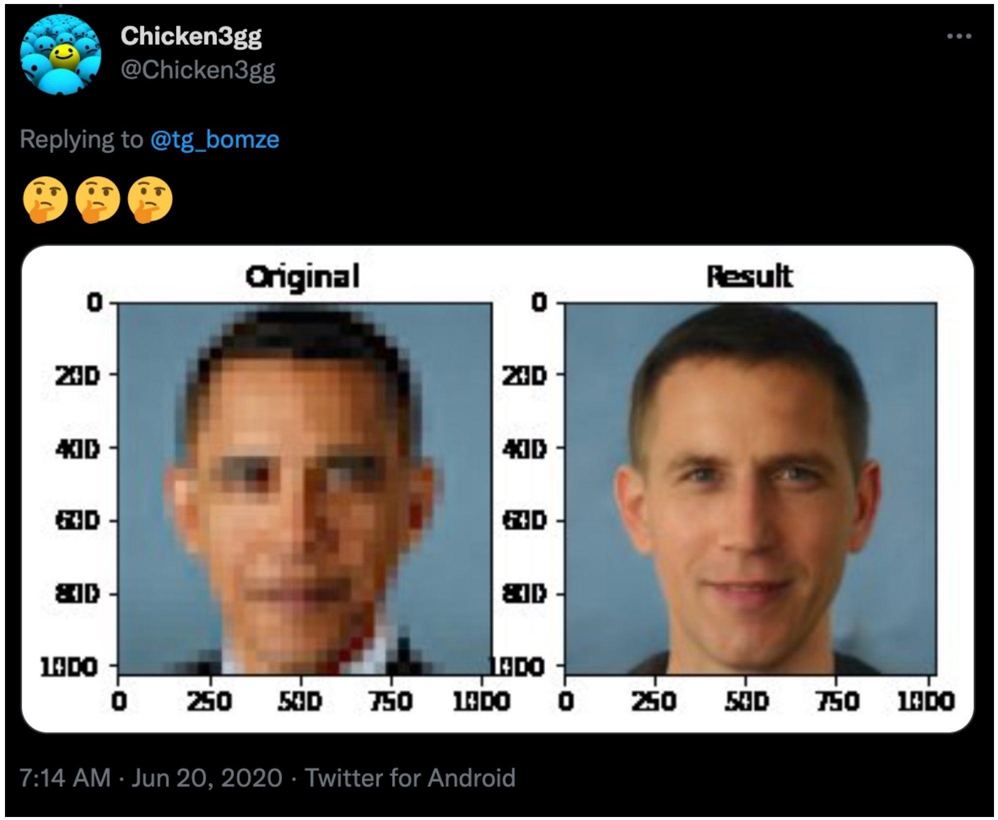
Dastin, J. (2018), Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*

Coded bias



Menon, S., *et al*. (2020). PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR* (pp. 2437-2445).

Coded bias

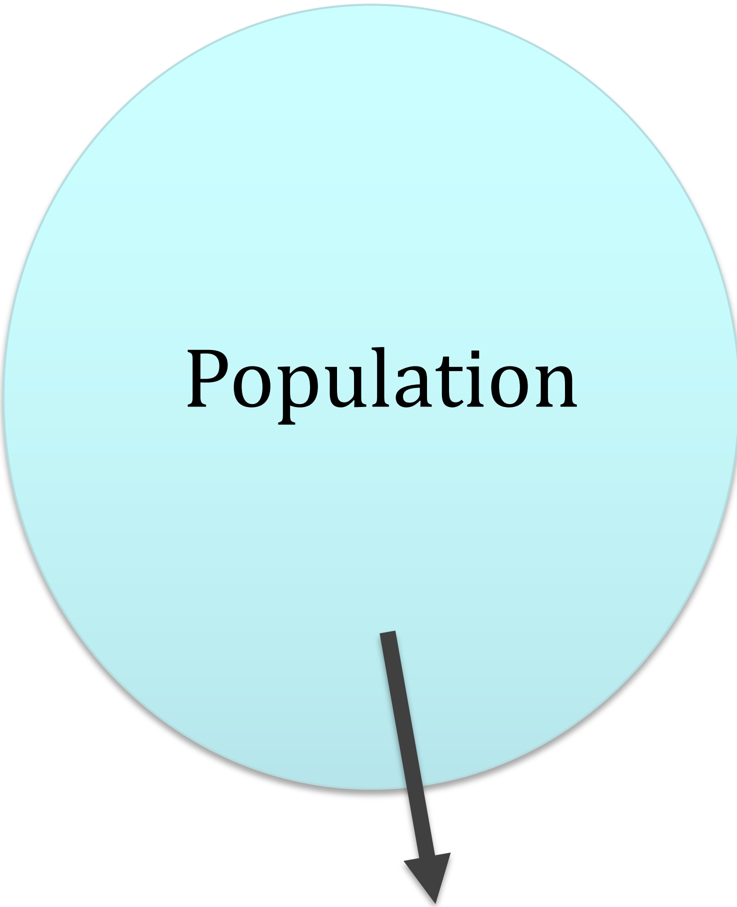


Menon, S., *et al*. (2020). PULSE: Self-supervised photo upsampling via latent space exploration of generative models. In *CVPR* (pp. 2437-2445).

The statistical pipeline

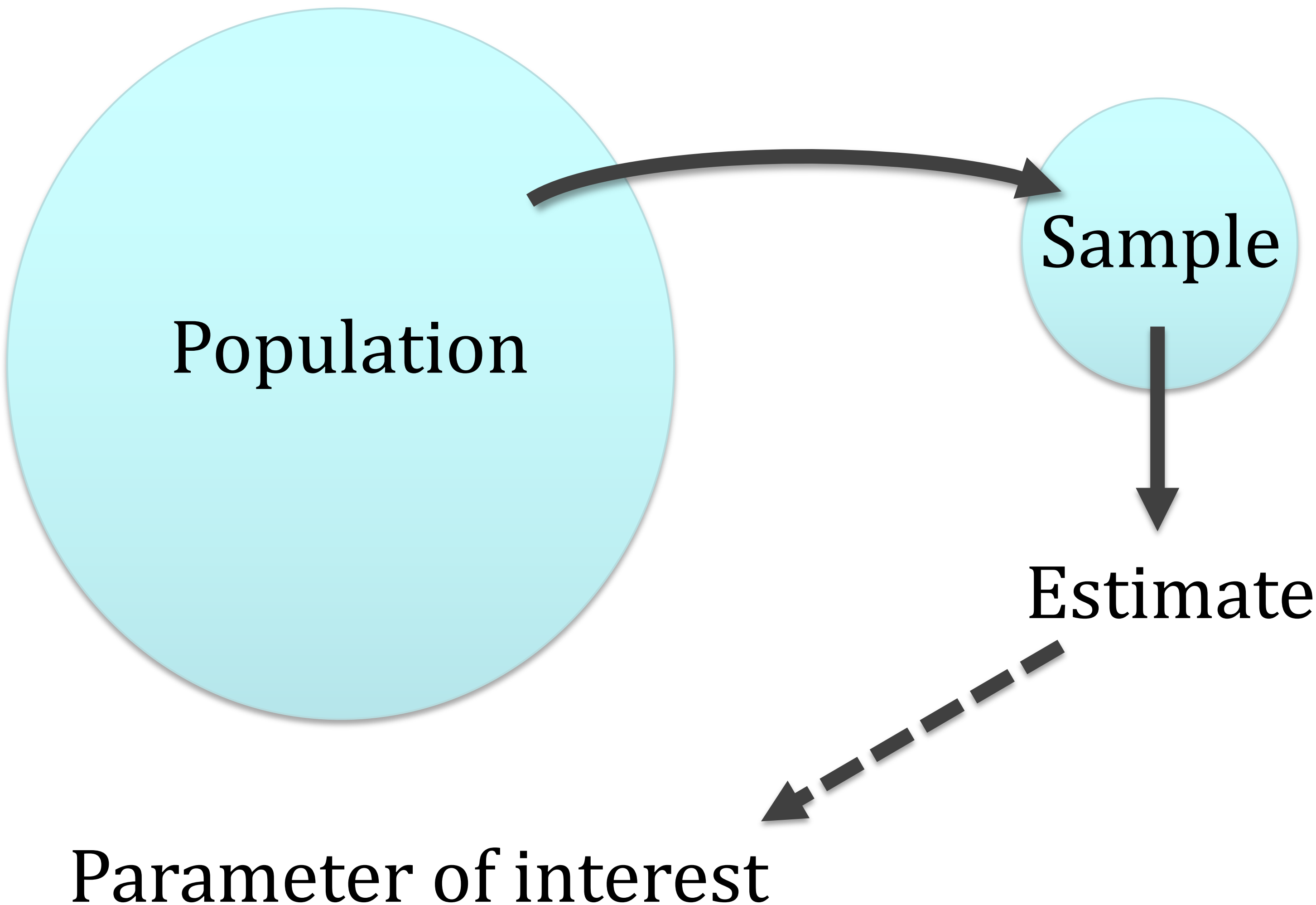
- Construct a question
- Gather data and perform any pre-processing
- Perform statistical analyses or modeling
- Make conclusions or predictions

The statistical pipeline

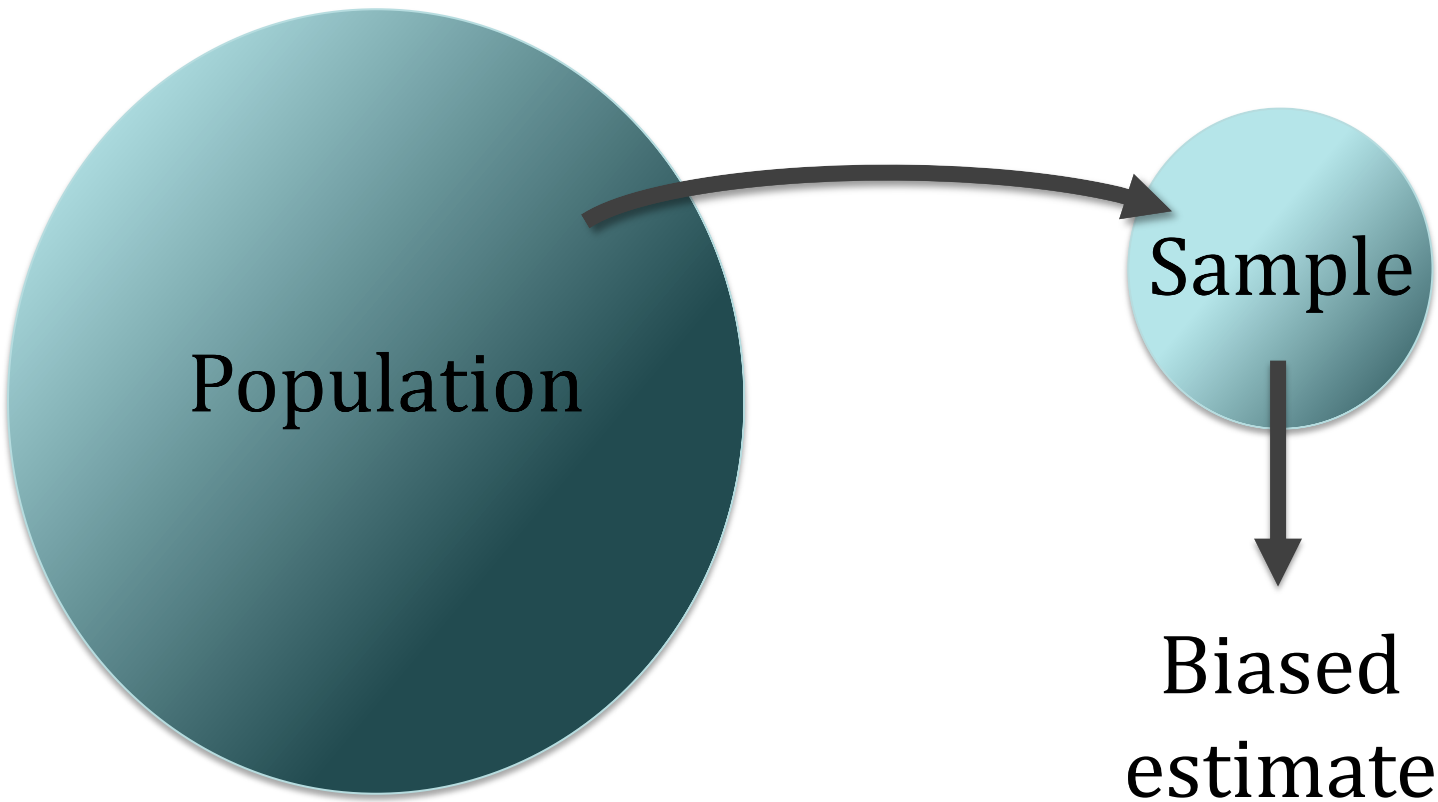


Parameter of interest

The statistical pipeline

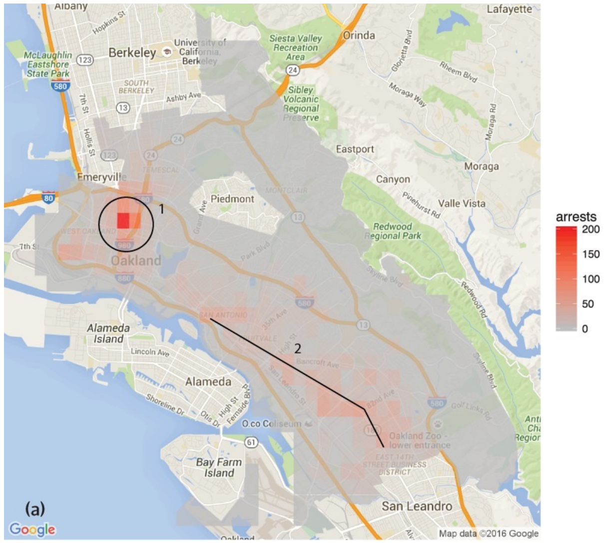


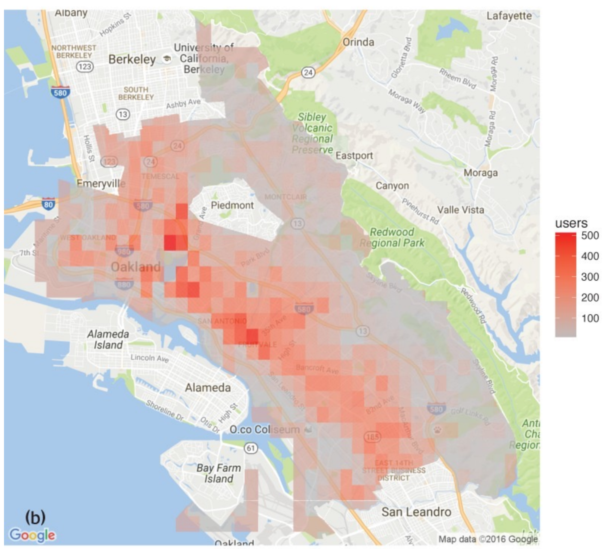
The statistical pipeline



Parameter of interest

The statistical pipeline





Top: Number of drug arrests made by Oakland police department, 2010. (1) West Oakland, (2) International Boulevard.

Bottom: Estimated number of drug users, based on 2011 National Survey on Drug Use and Health

Lum, K., & Isaac, W. (2016). To predict and serve?. *Significance*, *13*(5), 14-19.

All datasets are biased

All datasets are biased

But we can mitigate bias by thinking about our population and our data collection

Avoiding bias

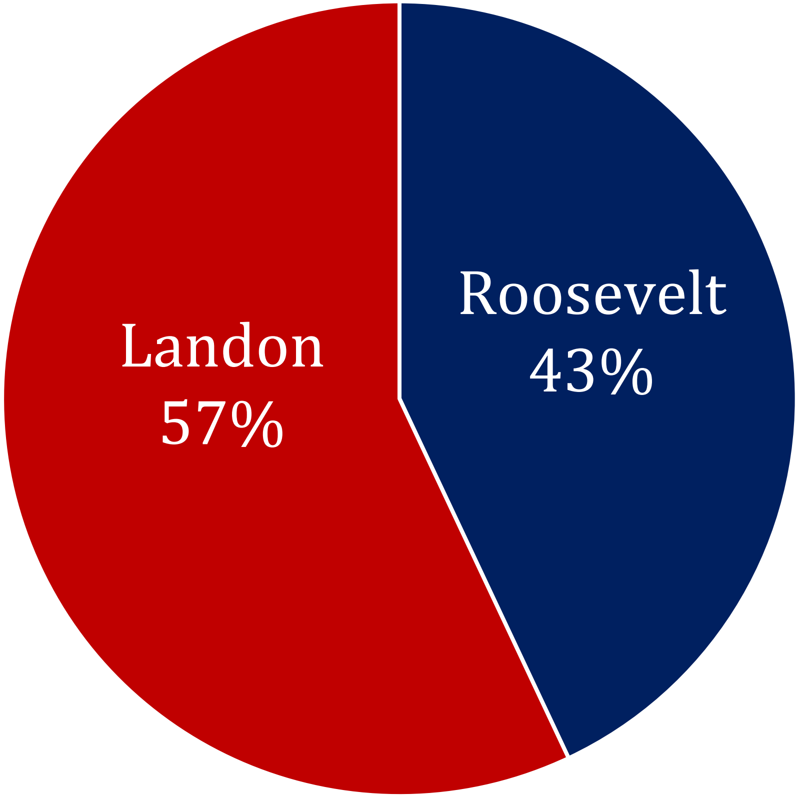
Once again, [we are] asking more than ten million voters -- one out of four, representing every county in the United States -- to settle November's election in October.

Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totaled.

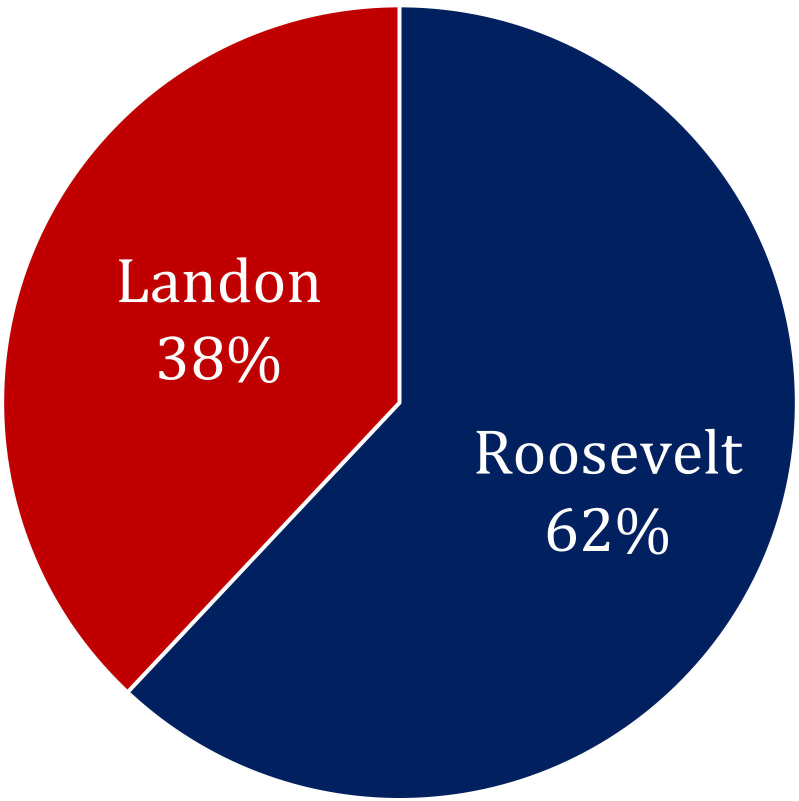
When the last figure has been totted and checked, if past experience is a criterion, the country will know to within a fraction of 1 percent the actual popular vote of forty million [voters].

Literary Digest, prior to the 1936 Alfred Landon (R) vs Franklin D. Roosevelt (D) election

Avoiding bias



Avoiding bias



Avoiding bias

Undercoverage bias: sample is not representative

- Literary digest: names taken from telephone lists, magazine subscription lists, club membership lists… highly biased towards upper/middle class.
- Mice: Only males included
- Facial recognition: datasets predominantly white

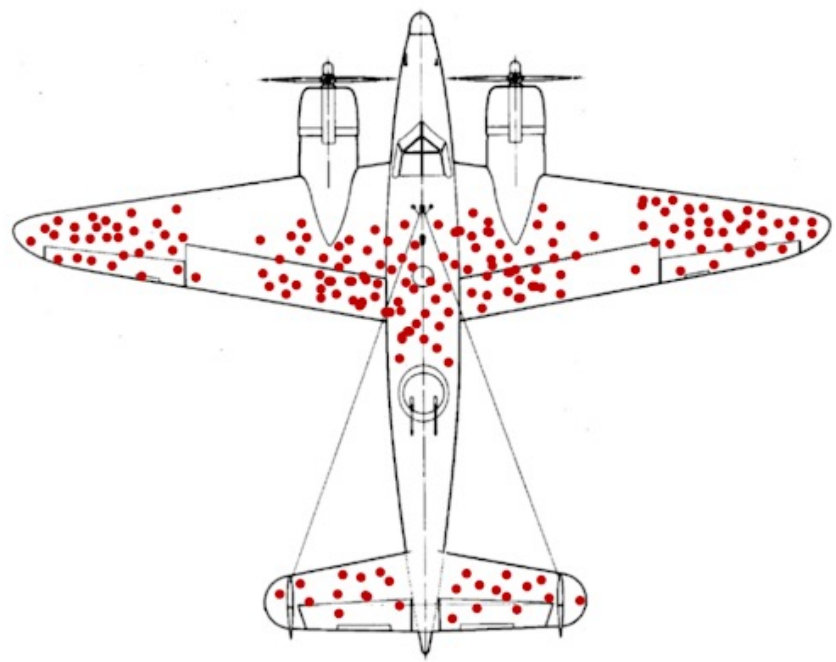
Avoiding bias

Volunteer bias: not everyone equally likely to respond

- Literary digest: only 2.4 million (out of 10 million) responded – are certain groups more likely to respond?
- Predictive policing: not all crimes equally likely to be reported or followed up

Avoiding bias

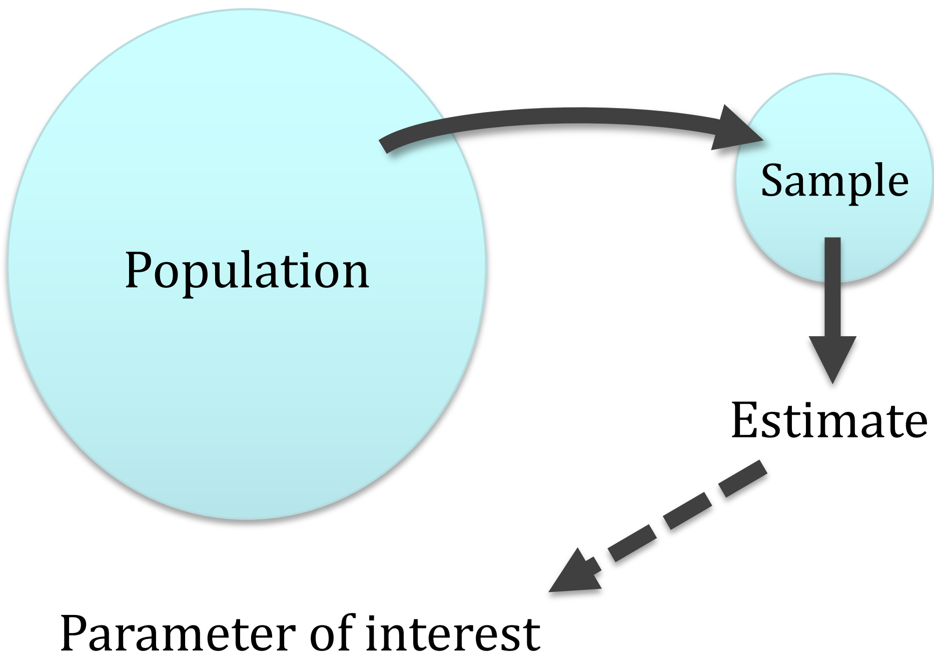
Survivorship bias: only looking at individuals who made it through some initial selection



Wald, Abraham. (1943). *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*. Statistical Research Group, Columbia University.

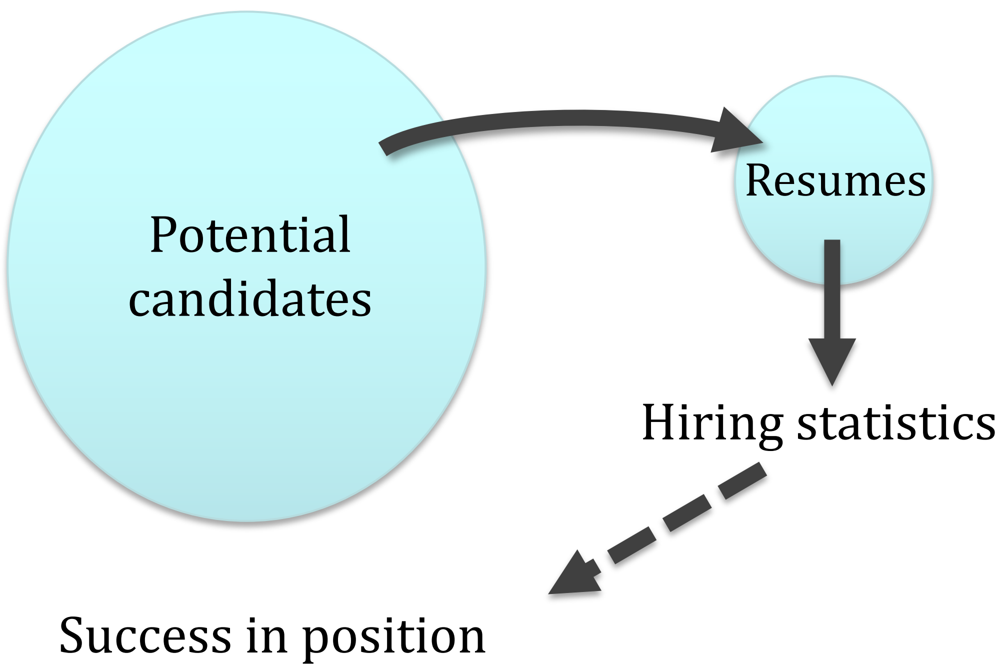
Are you using a biased proxy?

Is what you are predicting what you actually care about?



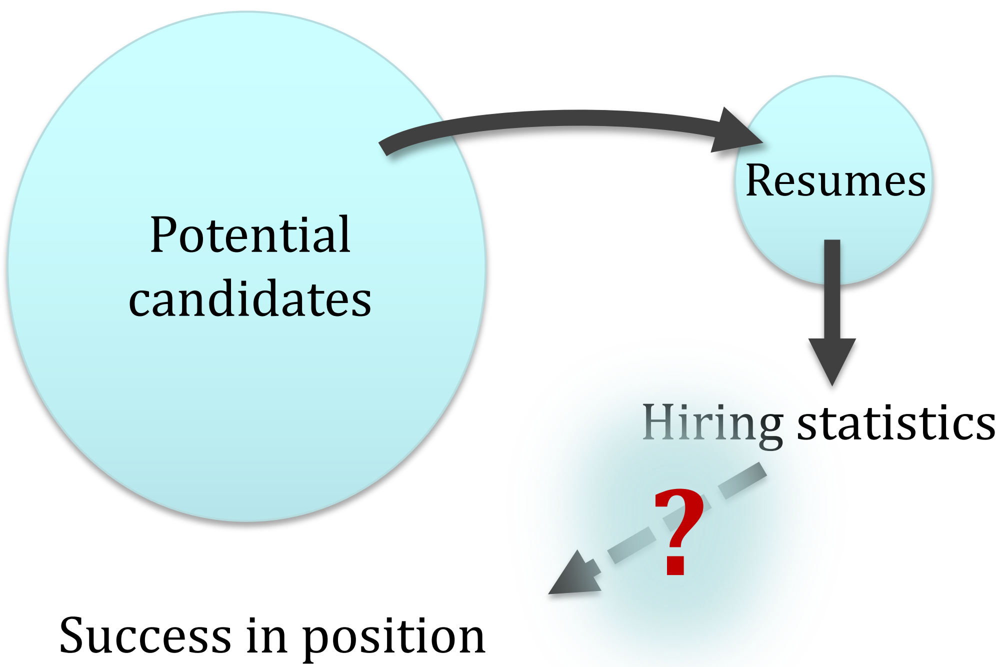
Are you using a biased proxy?

Is what you are predicting what you actually care about?



Are you using a biased proxy?

Is what you are predicting what you actually care about?



Are you using a biased proxy?

Is what you are predicting what you actually care about?

- What is predictive policing actually predicting?
- What is a hiring algorithm actually predicting?

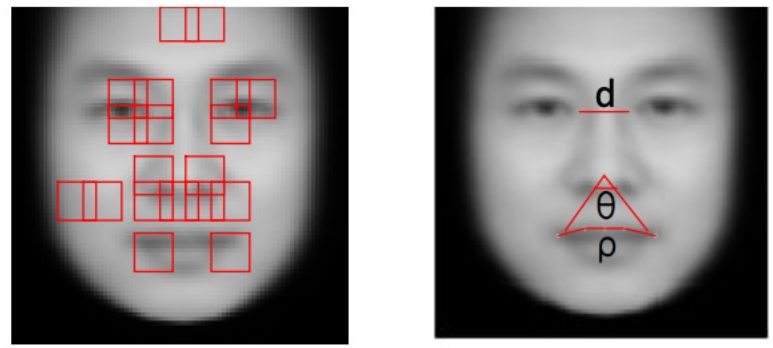
If you are using a proxy, you will propagate bias inherent in that proxy

Are you using a biased proxy?

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.



(a) (b)

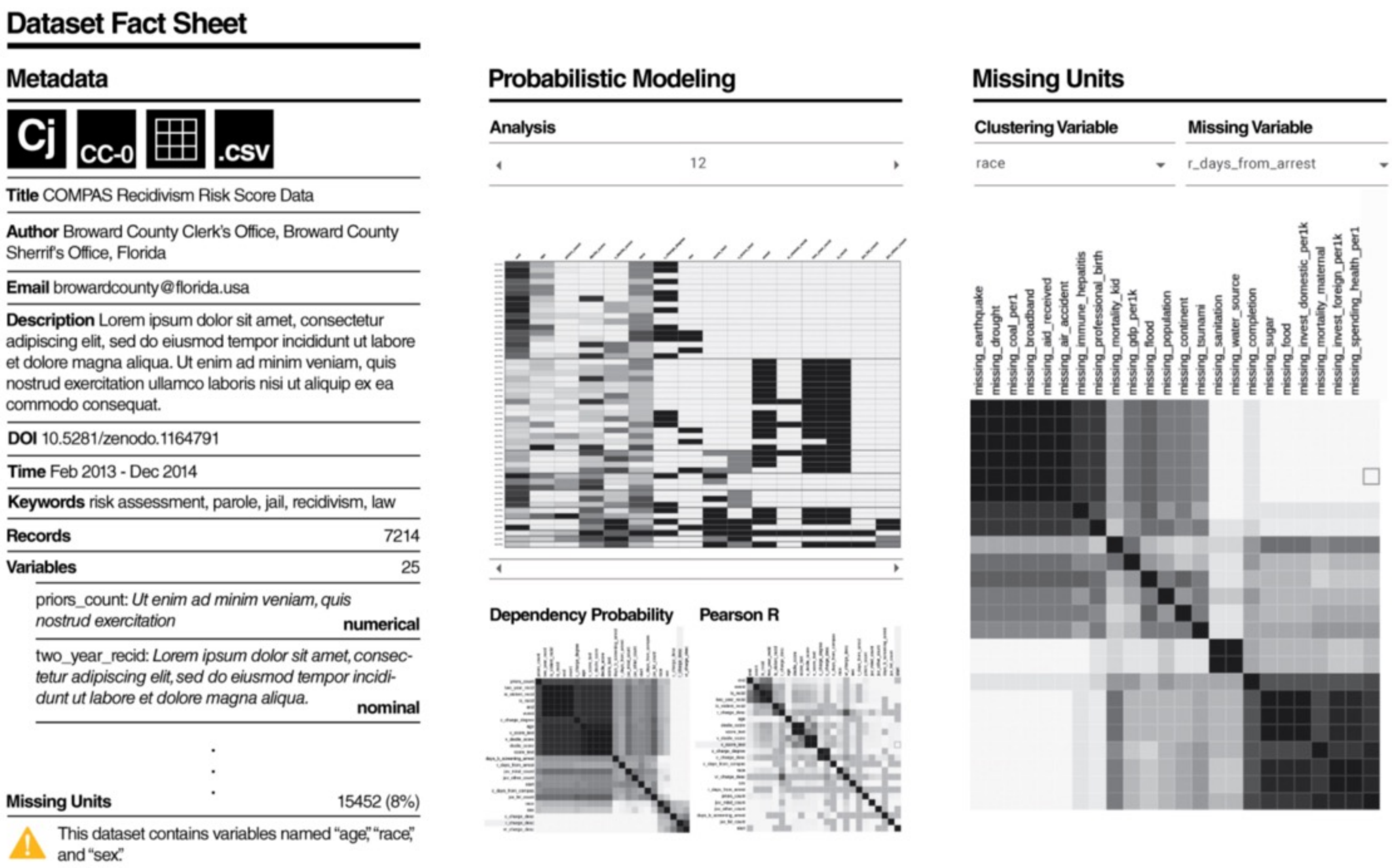
Figure 8. (a) FGM results; (b) Three discriminative features ρ, d and θ.

	Mean		Variance	
	criminal	non-criminal	criminal	non-criminal
ρ	0.5809	0.4855	0.0245	0.0187
d	0.3887	0.4118	0.0202	0.0144
θ	0.2955	0.3860	0.0185	0.0130

Table 4. The mean and variance for three normalized discriminative features ρ, d and θ.

Wu, X., & Zhang, X. (2016). Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*.

Documenting data



The Data Nutrition Project, https://datanutrition.org/

Why write data documentation?

Data creators… help process and streamline the process of data generation and encourage thoughtful data gathering

Data consumers… understand the data that they are using (and its limitations)

Data points… allows subjects to give informed consent and understand how their data is used

Stakeholders… allows people to understand and critique the analysis or predictions

The scientific community… allows reproducibility and fosters trust

Data Sheets for Data Sets

- **Motivation**
- **Composition**
- **Collection**
- **Uses**
- **Distribution and maintenance**

Gebru, T., *et al*. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Data Sheets for Data Sets

Motivation: why was the dataset created?

- Was there a task in mind?
- Who created the dataset?
- Who funded the dataset?

Gebru, T., *et al*. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Data Sheets for Data Sets

Composition:

- What does a data point represent?
- What does each data point consist of?
- Is it the entire dataset?
- If not, is it representative? How do you know?
- Is any information missing?

- Does the data set contain private or sensitive information?
- Are individuals identifiable?
- Does the dataset identify subpopulations? How?

Gebru, T., *et al*. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Data Sheets for Data Sets

Collection:

- How was the data collected?
- Over what timeframe?
- Was there any subsampling?
- Was any pre-processing done?

- Was an ethical review carried out?
- Was consent obtained from human subjects?
- Is there a mechanism for consent to be withdrawn?

Gebru, T., *et al*. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Data Sheets for Data Sets

Uses:

- How has this data been used?
- What might it be used for?
- What *shouldn't* it be used for

Distribution and maintenance

- Will the dataset be distributed? How/when?
- Are there copyright restrictions?
- Who is maintaining the dataset?
- Will it be updated? How?

Gebru, T., *et al*. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Data Sheets for Data Sets

Statlog (German Credit Data) Data Set

Download: Data Folder, Data Set Description

Abstract: This dataset classifies people described by a set of attributes as good or bad credit risks. Comes in two formats (one all numeric). Also comes with a cost matrix

Data Set Characteristics:	Multivariate	**Number of Instances:**	1000	**Area:**	Financial
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	20	**Date Donated**	1994-11-17
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	752214

Source:

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Data Sheets for Data Sets

Motivation: why was the dataset created?

- Was there a task in mind?
- Who created the dataset?
- Who funded the dataset?

Data Sheets for Data Sets

Motivation: why was the dataset created?

- Was there a task in mind?
- Who created the dataset?
- Who funded the dataset?

- Submitted to the UCI repository by Prof Dr Hans Hoffman

- Originally appears in a 1979 paper on credit scoring

Häußler, W. M. (1979). Empirische ergebnisse zu diskriminationsverfahren bei kreditscoringsystemen. *Zeitschrift für Operations Research*, *23*(8), B191-B210.

Data Sheets for Data Sets

Composition:

- What does a data point represent?
- What does each data point consist of?
- Is it the entire dataset?
- If not, is it representative? How do you know?
- Is any information missing?

- Does the data set contain private or sensitive information?
- Are individuals identifiable?
- Does the dataset identify subpopulations? How?

Data Sheets for Data Sets

Categorical variables

- Checking account status
- Credit history
- Purpose
- Amount in savings accounts
- Present employment duration
- Sex and relationship status
- Other debtors
- Property owned
- Other installment plans
- Housing status
- Job
- Has telephone?
- Foreign worker?
- Creditworthiness

Numeric variables

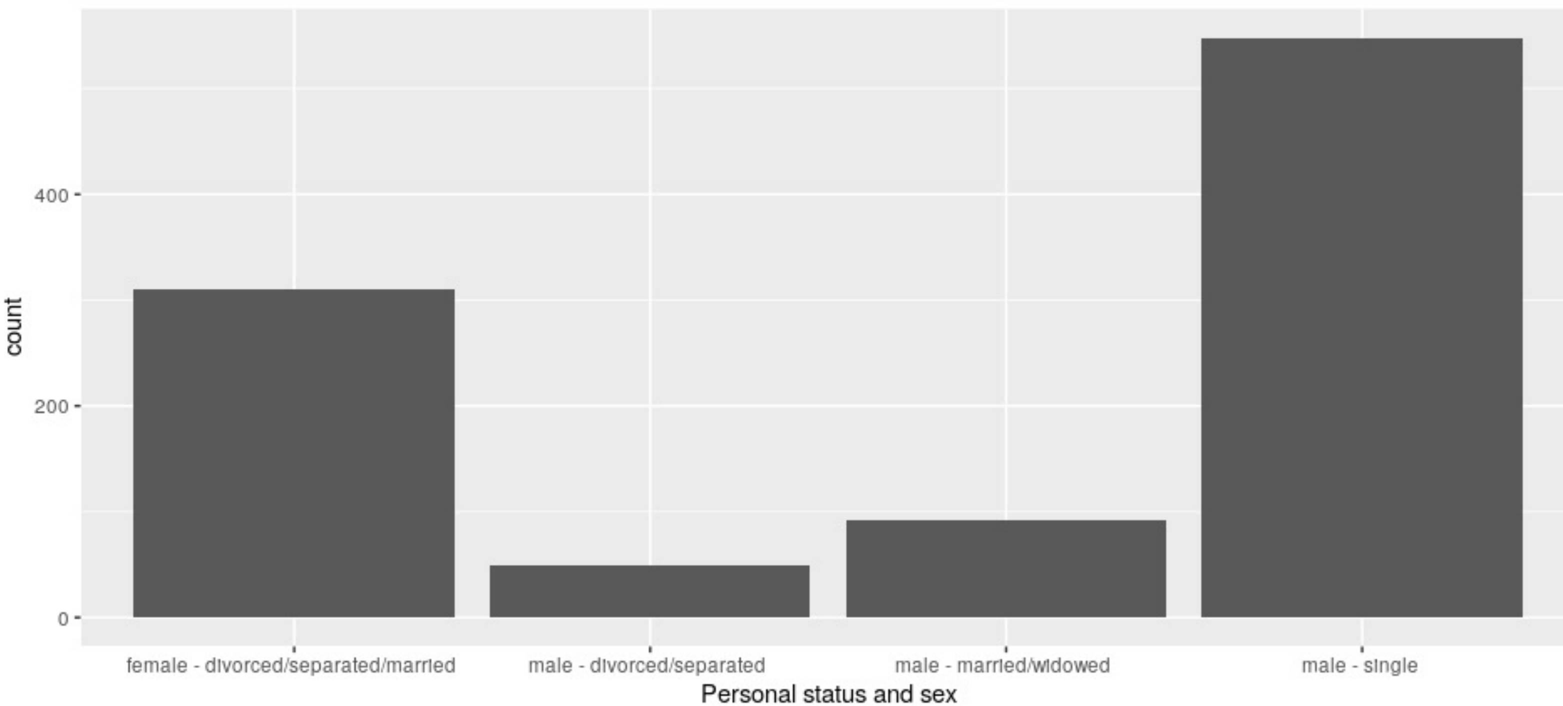
- Duration in months
- Credit amount
- Installment rate as % of disposable income (num)
- Time at present residence
- Age in years
- Number of existing credits
- Number of dependents

Data Sheets for Data Sets

- Is it the entire dataset?
- If not, is it representative? How do you know?

Data Sheets for Data Sets

- Is it the entire dataset?
- If not, is it representative? How do you know?



Data Sheets for Data Sets

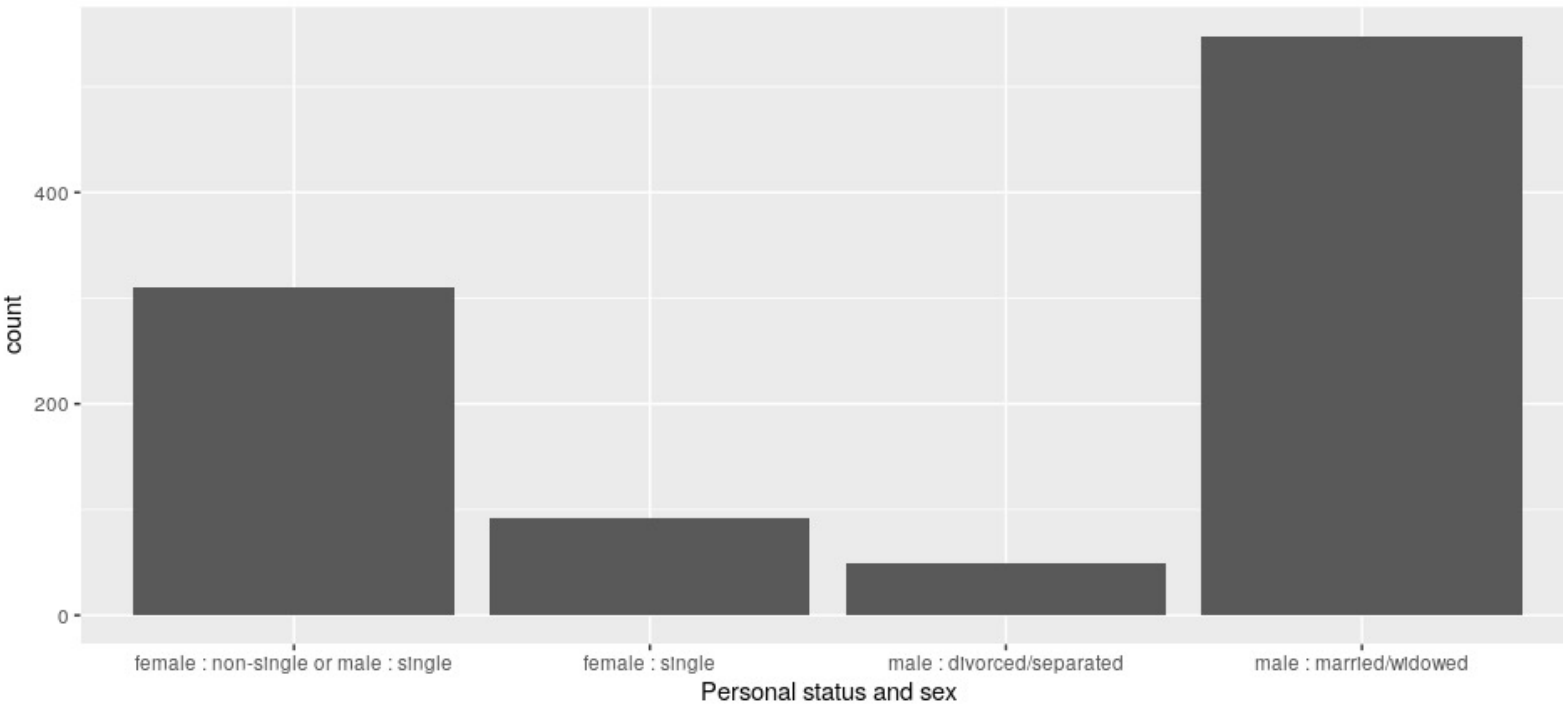
- Is it the entire dataset?
- If not, is it representative? How do you know?

- Corrected version created by Ulrike Grömping in 2019…

Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

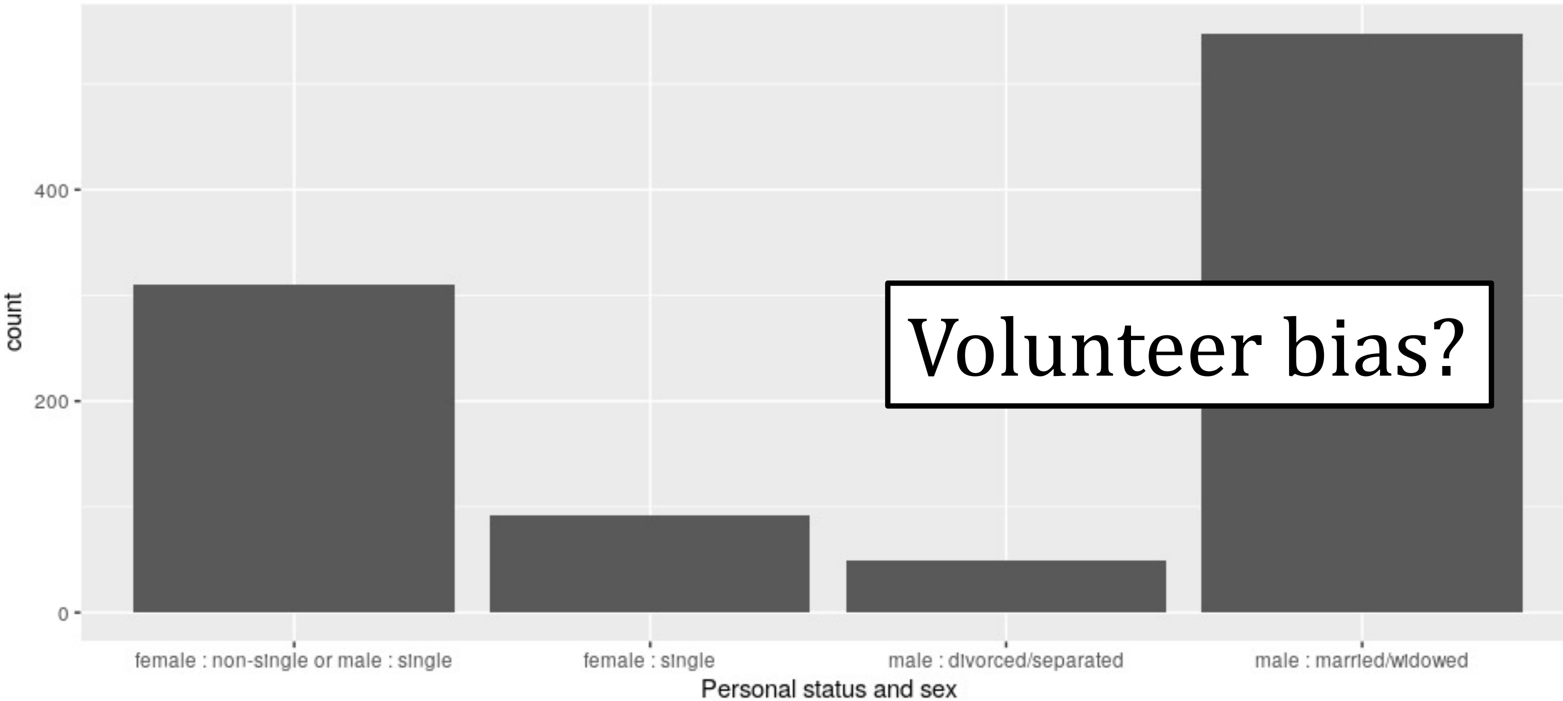
- Is it the entire dataset?
- If not, is it representative? How do you know?



Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

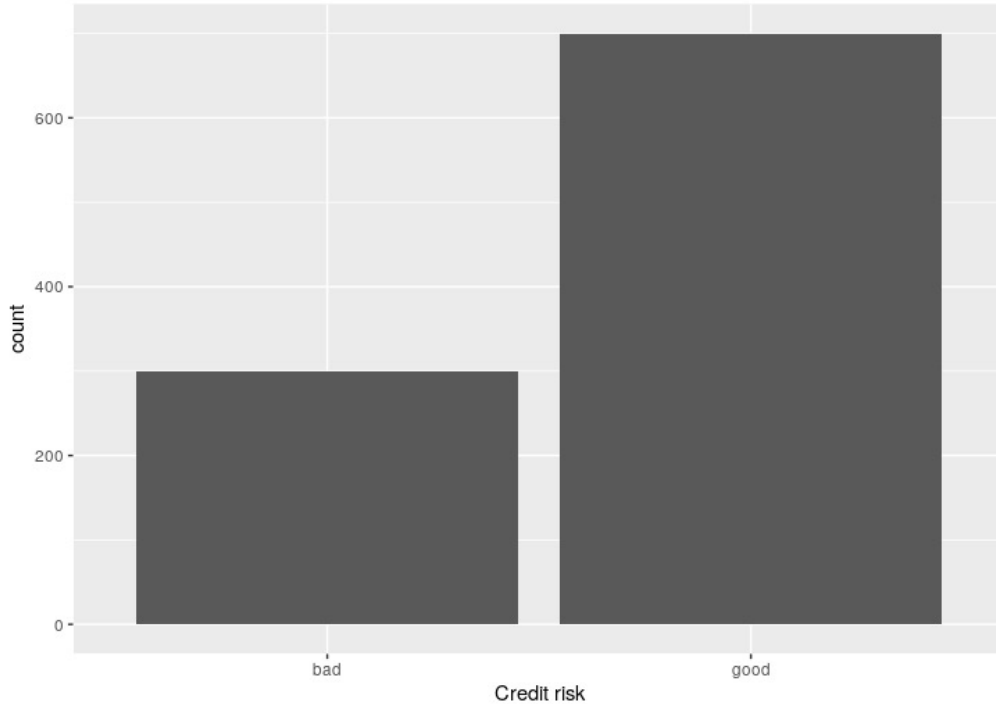
- Is it the entire dataset?
- If not, is it representative? How do you know?



Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

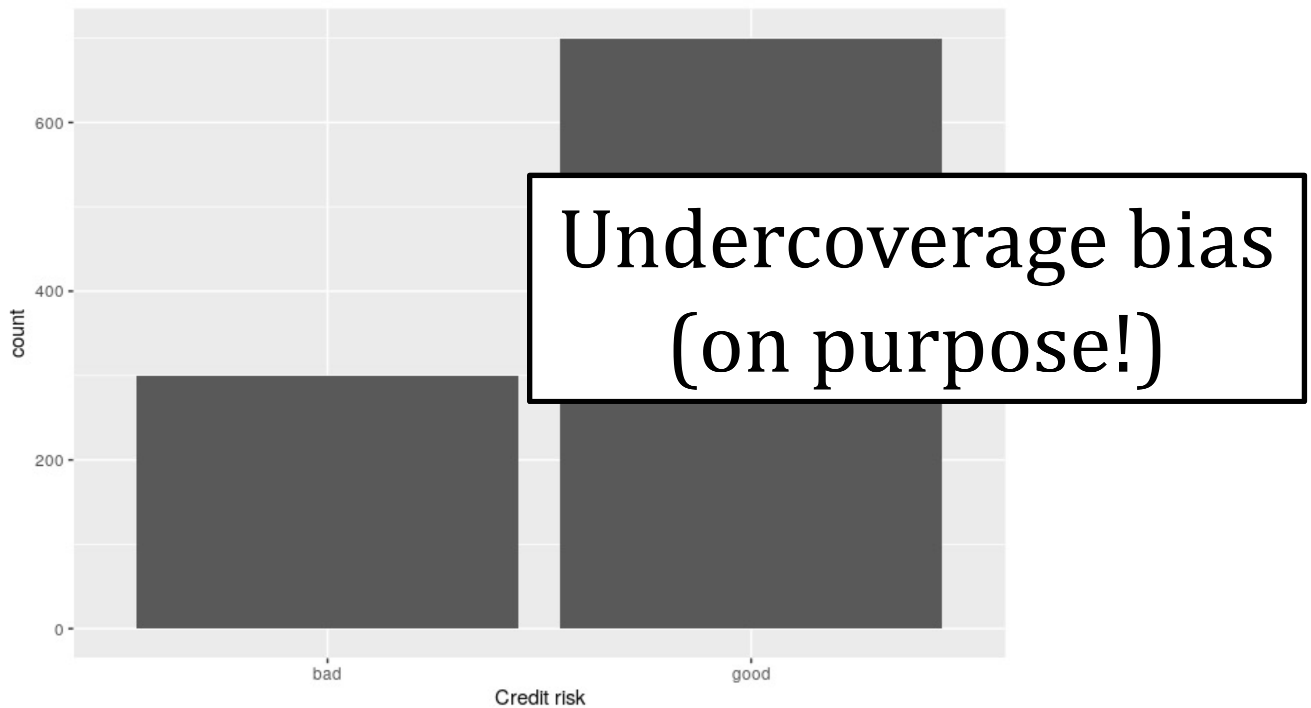
- Is it the entire dataset?
- If not, is it representative? How do you know?



Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

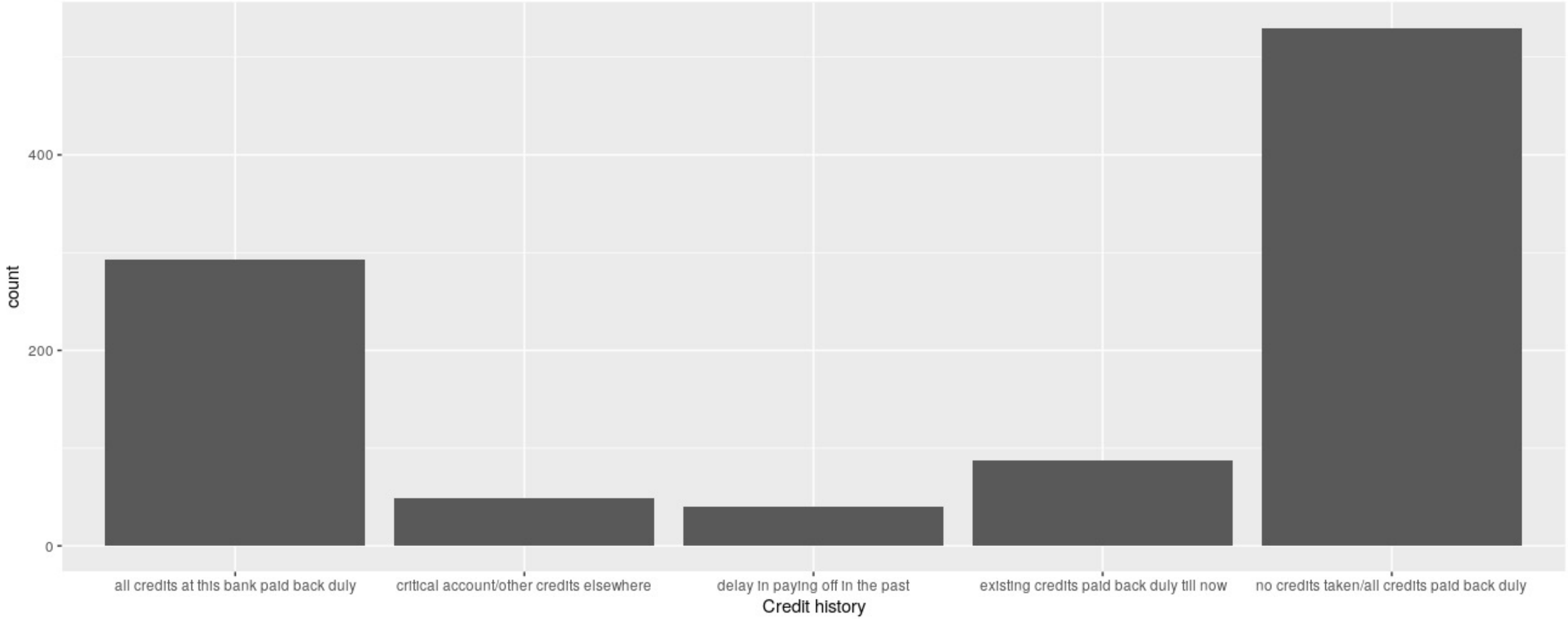
- Is it the entire dataset?
- If not, is it representative? How do you know?



Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

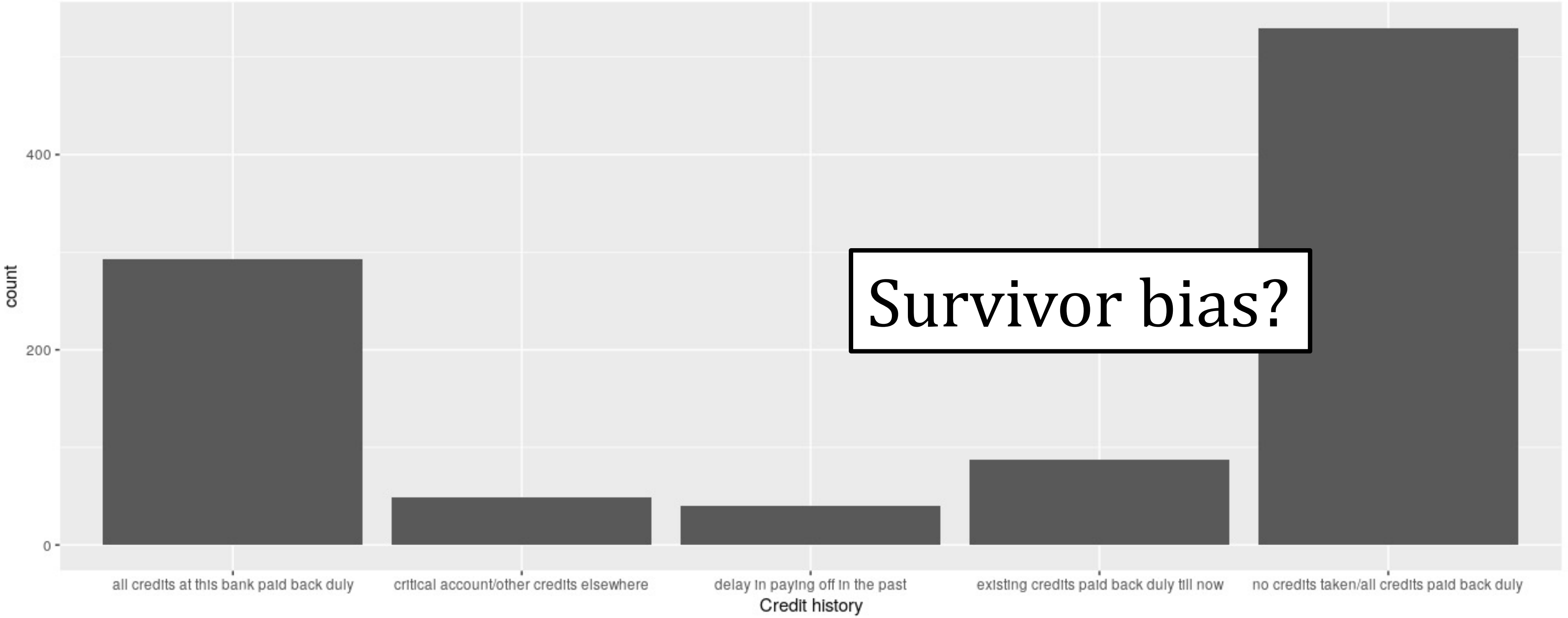
- Is it the entire dataset?
- If not, is it representative? How do you know?



Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

- Is it the entire dataset?
- If not, is it representative? How do you know?



Groemping, U. (2019). South German credit data: Correcting a widely used data set. *Rep. Math., Phys. Chem., Berlin, Germany, Tech. Rep*, *4*, 2019.

Data Sheets for Data Sets

Composition:

- Does the data set contain private or sensitive information?
- Are individuals identifiable?

Data Sheets for Data Sets

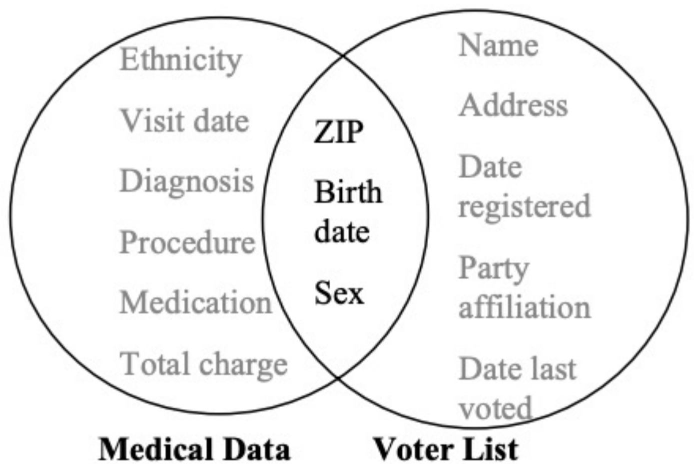
Composition:

- Does the data set contain private or sensitive information?
- Are individuals identifiable?

- Contains foreign worker status, sex, marital status
- Contains a large amount of financial information…

Data Sheets for Data Sets

For example, William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code.



Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

Data Sheets for Data Sets

We have…

- Only one person who is 74 and owns real estate
- Only one male foreign worker who is divorced/separated
- Only one person who owns their own home, but has been working under a year and doesn't have a telephone
- Only one single female home-owner

Data Sheets for Data Sets

Collection:

- How was the data collected?
- Over what timeframe?
- Was there any subsampling?
- Was any pre-processing done?

- Was an ethical review carried out?
- Was consent obtained from human subjects?
- Is there a mechanism for consent to be withdrawn?

Data Sheets for Data Sets

Collection:

- How was the data collected?
- Over what timeframe?
- Was there any subsampling?
- Was any pre-processing done?

- Was an ethical review carried out?
- Was consent obtained from human subjects?
- Is there a mechanism for consent to be withdrawn?

Data Sheets for Data Sets

Uses:

- How has this data been used?
- What might it be used for?
- What *shouldn't* it be used for

Data Sheets for Data Sets

Papers That Cite This Data Set[1]:

Rexa.info
Research • People • Connections

Jeroen Eggermont and Joost N. Kok and Walter A. Kosters. Genetic Programming for data classification: partitioning the search space. SAC. 2004. [View Context].

Ke Wang and Shiyu Zhou and Ada Wai-Chee Fu and Jeffrey Xu Yu. Mining Changes of Classification by Correspondence Tracing. SDM. 2003. [View Context].

Avelino J. Gonzalez and Lawrence B. Holder and Diane J. Cook. Graph-Based Concept Learning. FLAIRS Conference. 2001. [View Context].

Oya Ekin and Peter L. Hammer and Alexander Kogan and Pawel Winter. Distance-Based Classification Methods. e p o r t RUTCOR ffl Rutgers Center for Operations Research ffl Rutgers University. 1996. [View Context].

Chotirat Ann and Dimitrios Gunopulos. Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. Computer Science Department University of California. [View Context].

Paul O' Dea and David Griffith and Colm O' Riordan. DEPARTMENT OF INFORMATION TECHNOLOGY. P. O'Dea (NUI. [View Context].

Paul O' Dea and Josephine Griffith and Colm O' Riordan. Combining Feature Selection and Neural Networks for Solving Classification Problems. Information Technology Department, National University of Ireland. [View Context].

Citation Request:

Please refer to the Machine Learning Repository's citation policy

---

[1] Papers were automatically harvested and associated with this data set, in collaboration with Rexa.info

https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Data Sheets for Data Sets

Google Scholar

"german credit dataset"

Articles

About 1,010 results (**0.03** sec)

Any time
Since 2021
Since 2020
Since 2017
Custom range...

Sort by relevance
Sort by date

Any type
Review articles

include patents
include citations

Create alert

… of neural network, C5. 0, and classification and regression trees algorithms in the credit risk evaluation problem (case study: a standard **German credit dataset**)
MM Khoraskani, F Kheradmand… - … Journal of Knowledge …, 2017 - inderscienceonline.com
Due to the reducing global economic stability, the demand of banks for predicting their customer's credit risk has significantly increased and has become more critical, still challenging than ever. This paper addresses the problem of credit risk evaluation of bank's …
Save Cite Cited by 1 Related articles All 2 versions

Deep convolutional neural networks versus multilayer perceptron for financial prediction
VE Neagoe, AD Ciotec, GS Cucu - … International Conference on …, 2018 - ieeexplore.ieee.org
… The experiments have used the **German credit dataset** and the Australian credit dataset. The model performances are evaluated by the following indices: Overall Accuracy (OA); False Alarm Rate (FAR); Missed Alarm Rate (MAR). The experimental results have confirmed the …
Save Cite Cited by 29 Related articles All 2 versions

[HTML] Information gain directed genetic algorithm wrapper feature selection for credit rating
S Jadhav, H He, K Jenkins - Applied Soft Computing, 2018 - Elsevier

https://scholar.google.com/

Data Sheets for Data Sets

Uses:

- How has this data been used?
- What might it be used for?
- What *shouldn't* it be used for

- Has been used for credit score modeling, fairness and explainability analysis, and to showcase various prediction algorithms

- *Shouldn't* be used for predicting credit in a real-world setting
- *Shouldn't* be used in an interpretability setting (without corrections)

Data Sheets for Data Sets

Distribution and maintenance

- Will the dataset be distributed? How/when?
- Are there copyright restrictions?
- Who is maintaining the dataset?
- Will it be updated? How?

Data Sheets for Data Sets

Distribution and maintenance

- Will the dataset be distributed? How/when?
- Are there copyright restrictions?
- Who is maintaining the dataset?
- Will it be updated? How?

- Good: open access, clear citation policy, maintained by UCI

- Bad: no updates, no link to improved dataset, no mechanism for removing data

Data Sheets for Data Sets

Other resources for creating data documentation:

- Bender, E. M., & Friedman, B. (2018). **Data statements for natural language processing**: Toward mitigating system bias and enabling better science. In *ACL*, 587-604.

- Stoyanovich, J., & Howe, B. (2019). **Nutritional labels for data and models**. In *IEEE Technical Committee on Data Engineering*, *42*(3).

- Holland, S., *et al*. (2018). **The dataset nutrition label**: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*.