«Human-Centered Data Science»

Next Session: Digging Deeper into Approaches to Identify, Mitigate and Avoid Bias

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 2, 2022

Lecture Overview

Recap

Discrimination (social science, legal perspective, computer science perspective)

Sources of Bias (cleaning and wrangling data; feature engineering)

Break

Sources of Bias (data analysis and training; selecting and evaluating the model)

Types of Emergent Bias

Recap

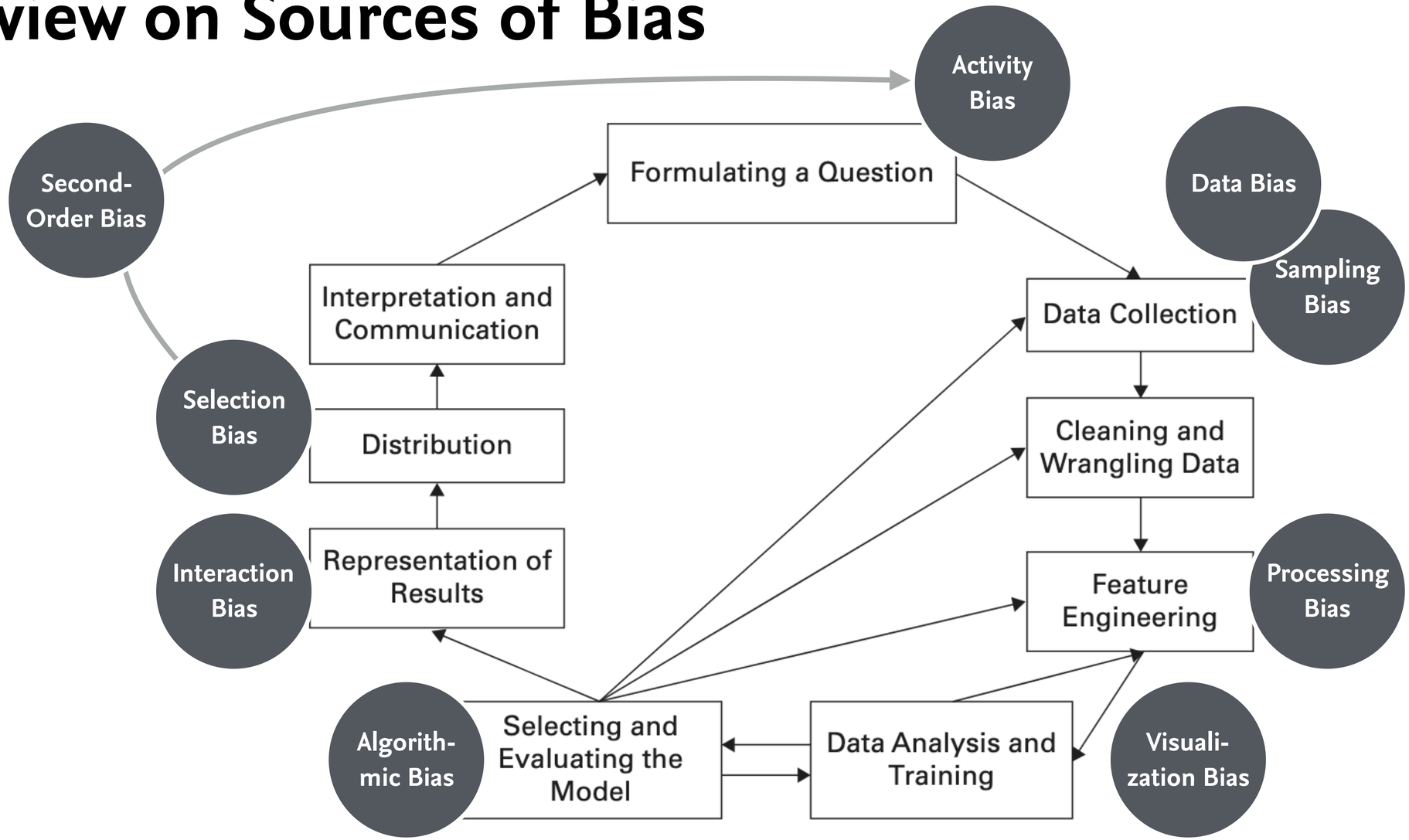
> **The term bias refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on unreasonable or inappropriate grounds.**
>
> (Friedman & Nissenbaum 1996)

Systematic discrimination does not establish bias unless it is joined with an unfair outcome.

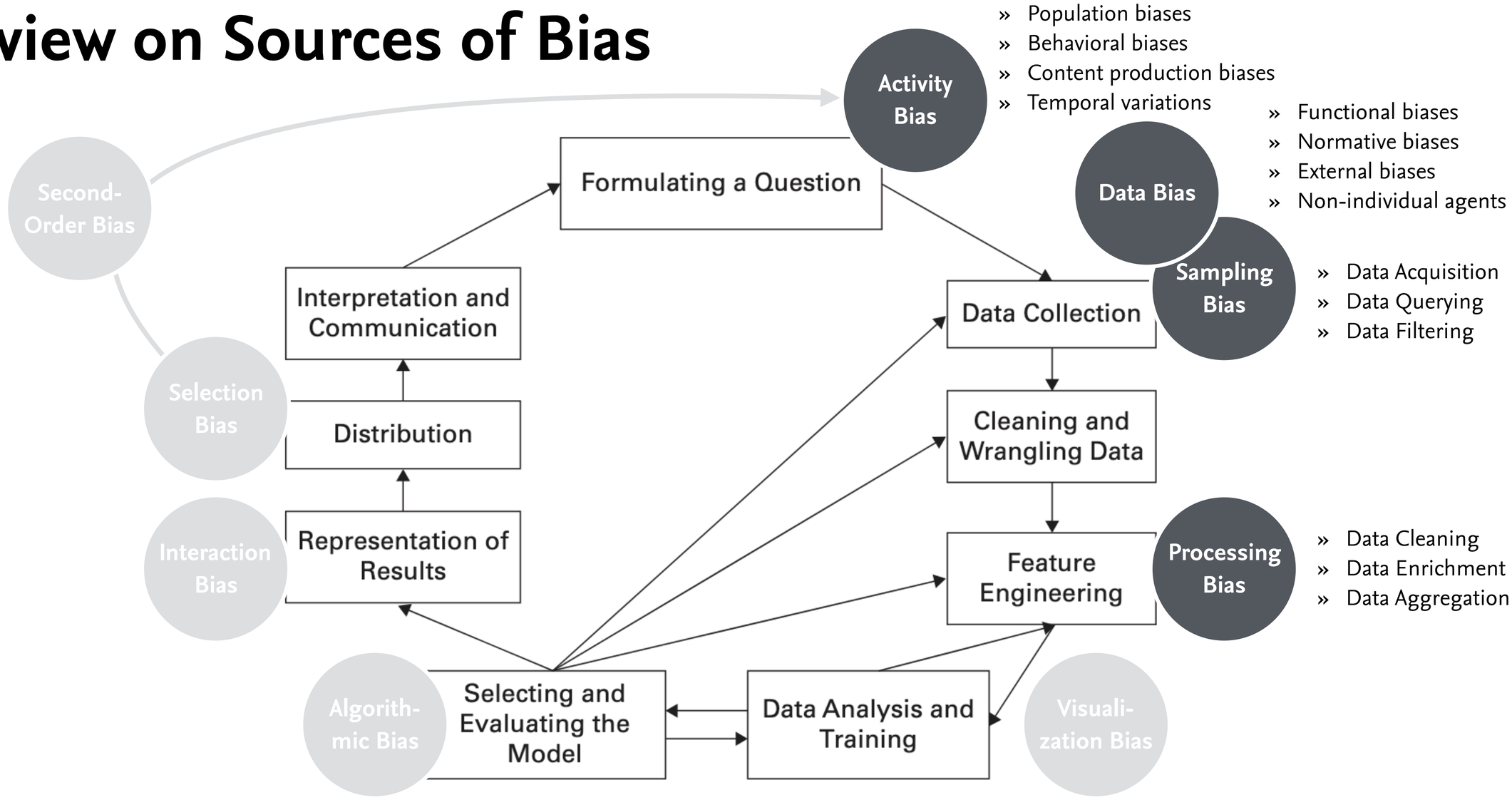
Unfair discrimination alone does not give rise to bias unless it occurs systematically.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. ACM Trans. Inf. Syst., 14(3), 330–347. http://doi.org/10.1145/230538.230561

Overview on Sources of Bias



Overview on Sources of Bias



Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting sentiment polarity—i.e., given a piece of English text, predict whether it has a positive or negative affect—or stance—toward its topic. The dataset was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.[1]

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

these are words that could be used to describe the emotions of john sayles' characters in his latest , limbo . but no , i use them to describe myself after sitting through his latest little exercise in indie egomania . i can forgive many things . but using some hackneyed , whacked-out , screwed-up * non * - ending on a movie is unforgivable . i walked a half-mile in the rain and sat through two hours of typical , plodding sayles melodrama to get cheated by a complete and total copout finale . does sayles think he's roger corman ?

Figure 1. An example "negative polarity" instance, taken from the file `neg/cv452_tok-18656.txt`.

exception that no more than 40 posts by a single author were included (see "Collection Process" below). No tests were run to determine representativeness.

The instances are movie reviews extracted from newsgroup postings, together with a sentiment polarity rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly negative (low number of stars). The sentiment polarity rating is binary {positive, negative}. An example instance is shown in figure 1.

How many instances are there in total (of each type, if appropriate)?

There are 1,400 instances in total in the original (v1.x versions) and 2,000 instances in total in v2.0 (from 2014).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset is a sample of instances. It is intended to be a random sample of movie reviews from newsgroup postings, with the

[1]All information in this datasheet is taken from one of the following five sources; any errors that were introduced are the fault of the authors of the datasheet: http://www.cs.cornell.edu/people/pabo/movie-review-data/; http://xxx.lanl.gov/pdf/cs/0409058v1; http://www.cs.cornell.edu/people/pabo/movie-review-data/rt-polaritydata.README.1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata.README.2.0.txt.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included. No data is missing.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

None explicitly, though the original newsgroup postings include poster name and email address, so some information (such as threads, replies, or posts by the same author) could be extracted if needed.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in classification accuracy.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

See preprocessing below.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset was created to enable research on predicting senti-

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

Funding was provided from five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Sloan Foundation.

Any other comments?

None.

Sections of «Datasheets for Datasets»

motivation | composition | collection process | preprocessing/ cleaning/labeling | uses | distribution | maintenance

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. Commun. ACM 64, 12 (December 2021), 86–92.

Documentation As Reflexive Practice

Documentation as reflexive practices should be seen as a **constitutive part** of data work.

Reflexive documentation could

- »**make** praxis-based and **situated decision-making explicit** and help preserve it in documentation
- »be especially useful to **improve traceability**
- »provide the context of dataset production could constitute a useful tool for **auditability**
- »become an additional supportive tool for data workers to contribute towards the **compliance with** existing **legal frameworks**

Miceli, Milagros, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, und Alex Hanna. „Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices“. In _Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency_, 161–72. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021.

Discrimination

Discrimination from a Social Science Perspective

Discrimination is not an isolated, individual action, but a structural, social phenomenon. There are very heterogeneous forms of discrimination: e.g. racism, sexism, antisemitism.

> **"Construction of distinctions of social groups and categories of persons, which are used for the creation, justification and legitimation of demarcations and hierarchies, especially of power asymmetries, socio-economic inequalities and unequal chances of recognition".**
>
> Scherr, Albert 2017. Soziologische Diskriminierungsforschung, in Scherr, Albert, El-Mafaalani, Aladin & Yüksel, Gökçen (Hg.): Handbuch Diskriminierung. Wiesbaden: Springer Fachmedien Wiesbaden, 39–58.

Discrimination from a Legal Perspective

> „Not all discriminations can be prohibited; the word "to discriminate," once divested of its emotional connotation, simply means to distinguish or to draw a line“
>
> Owen M. Fiss

Article 3 Basic Law for the Federal Republic of Germany

(1) All persons shall be equal before the law.

(2) Men and women shall have equal rights. The state shall promote the actual implementation of equal rights for women and men and take steps to eliminate disadvantages that now exist.

protected variables

(3) No person shall be favoured or disfavoured because of sex, parentage, race, language, homeland and origin, faith or religious or political opinions. No person shall be disfavoured because of disability.

Discrimination: Treatment vs Impact

Modern legal frameworks offer various levels of protection for being discriminated by belonging to a particular class of: gender, age, ethnicity, nationality, disability, religious beliefs, and/or sexual orientation.

Often two concepts are differentiated:

- **Disparate treatment:** Treatment depends on class membership
- **Disparate impact:** Outcome depends on class membership

Example: Austrian AMS System

BE_INT

= f (0,10

– 0,14 x GENDER_FEMALE

– 0,13 x AGE-GROUP_30_49

– 0,70 x AGE-GROUP_50_PLUS

+ 0,16 x STATE_GROUP_EU

– 0,05 x STATE_GROUP_THIRD

+ 0,28 x EDUCATION_APPRENTICESHIP

+ 0,01 x EDUCATION_MATURA_PLUS

– 0,15 x CARE_TAKING

– 0,34 x LIVING_TYP_2

– 0,18 x LIVING_TYP_3

– 0,83 x LIVING_TYP_4

– 0,82 x LIVING_TYP_5

...

– 0,67 x IMPAIRED

+ 0,17 x OCCUPATION_PRODUCTION

– 0,74 x OCCUPATION_DAYS_LITTLE

+ 0,65 x FREQUENCY_CASE_1

+ 1,19 x FREQUENCY_CASE_2

+ 1,98 x FREQUENCY_CASE_3_PLUS

– 0,80 x CASE_LONG

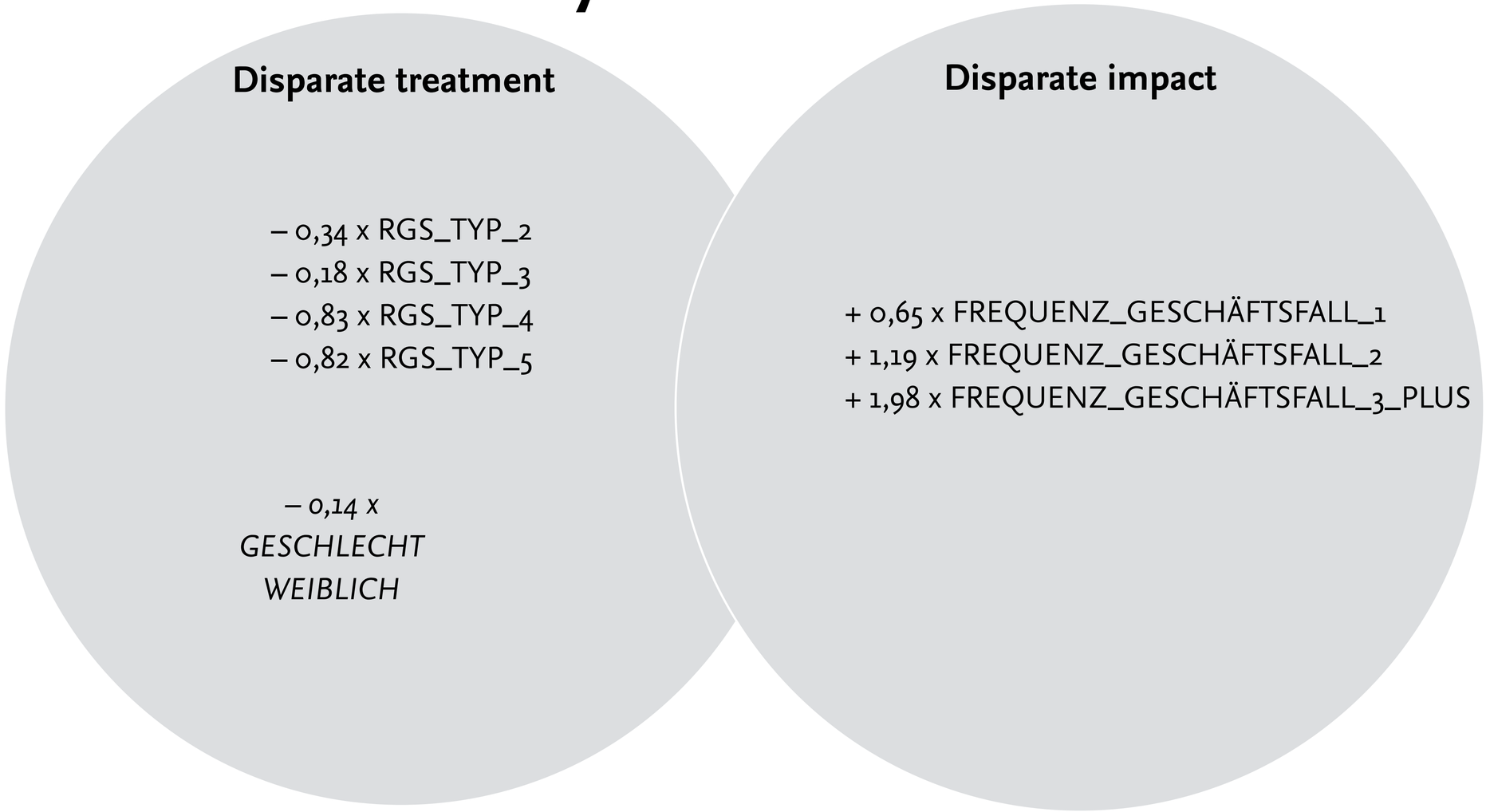
– 0,57 x MN_PARTICIPATION_1

– 0,21 x MN_PARTICIPATION_2

– 0,43 x MN_PARTICIPATION_3)

Translated features from http://www.forschungsnetzwerk.at/downloadpub/arbeitsmarktchancen_methode_%20dokumentation.pdf

Example: Austrian AMS System



Discrimination from a Computer Science Perspective

The focus is on quantification. In the context of an algorithm generating a prediction:

- Predictions for people with similar non-protected attributes should be similar
- Differences should be mostly explainable by non-protected attributes

Two basic frameworks for measuring discrimination:

- **Discrimination at the individual level:** consistency or individual fairness
- **Discrimination at the group level:** statistical parity

Hajian, Sara, Francesco Bonchi, und Carlos Castillo. 2016. Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. KDD '16. the 22nd ACM SIGKDD International Conference. New York, New York, USA: SIGMOD, ACM Special Interest Group on Management of Data. https://doi.org/10.1145/2939672.2945386.
Žliobaitė I. 2015. A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

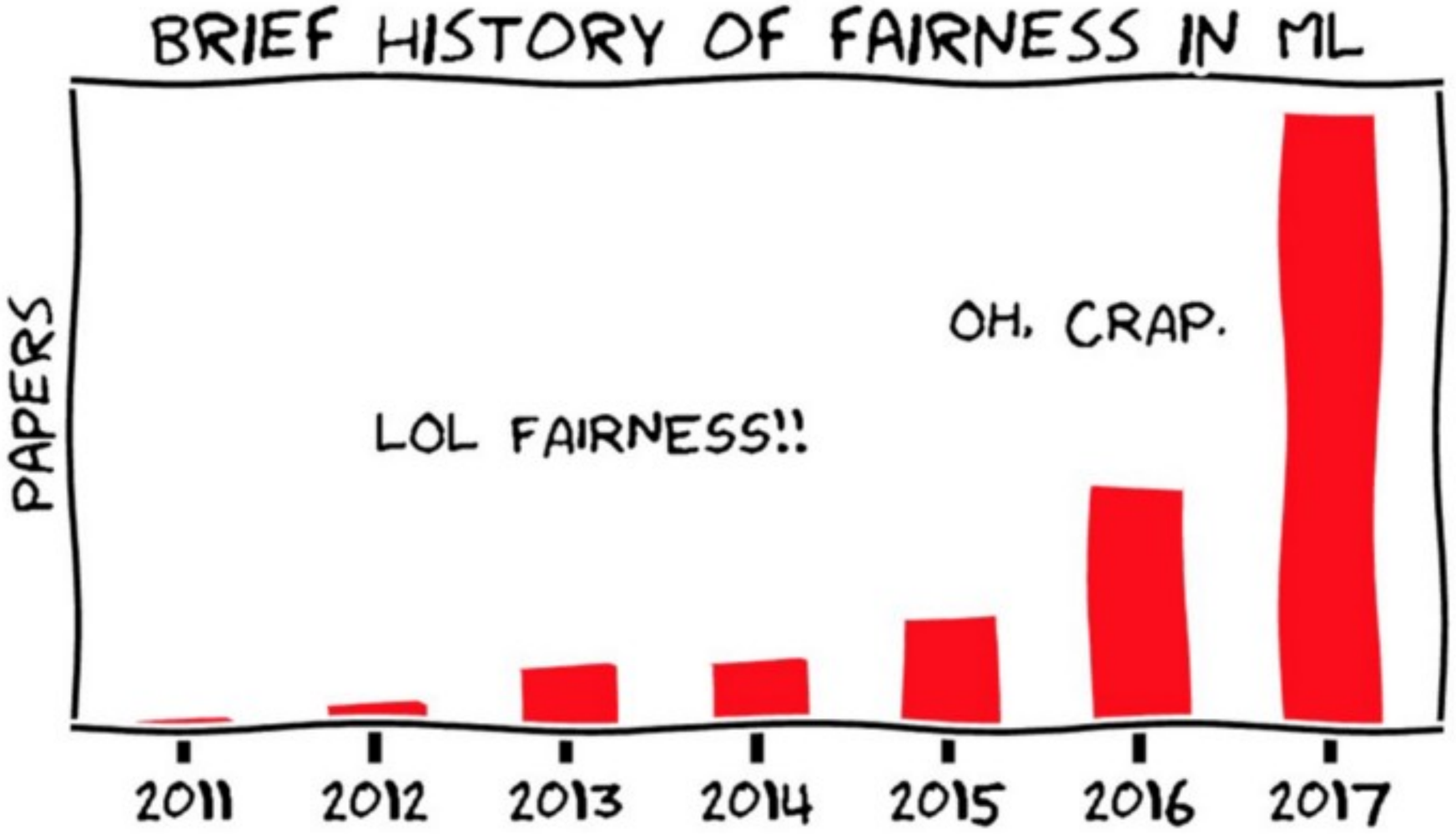
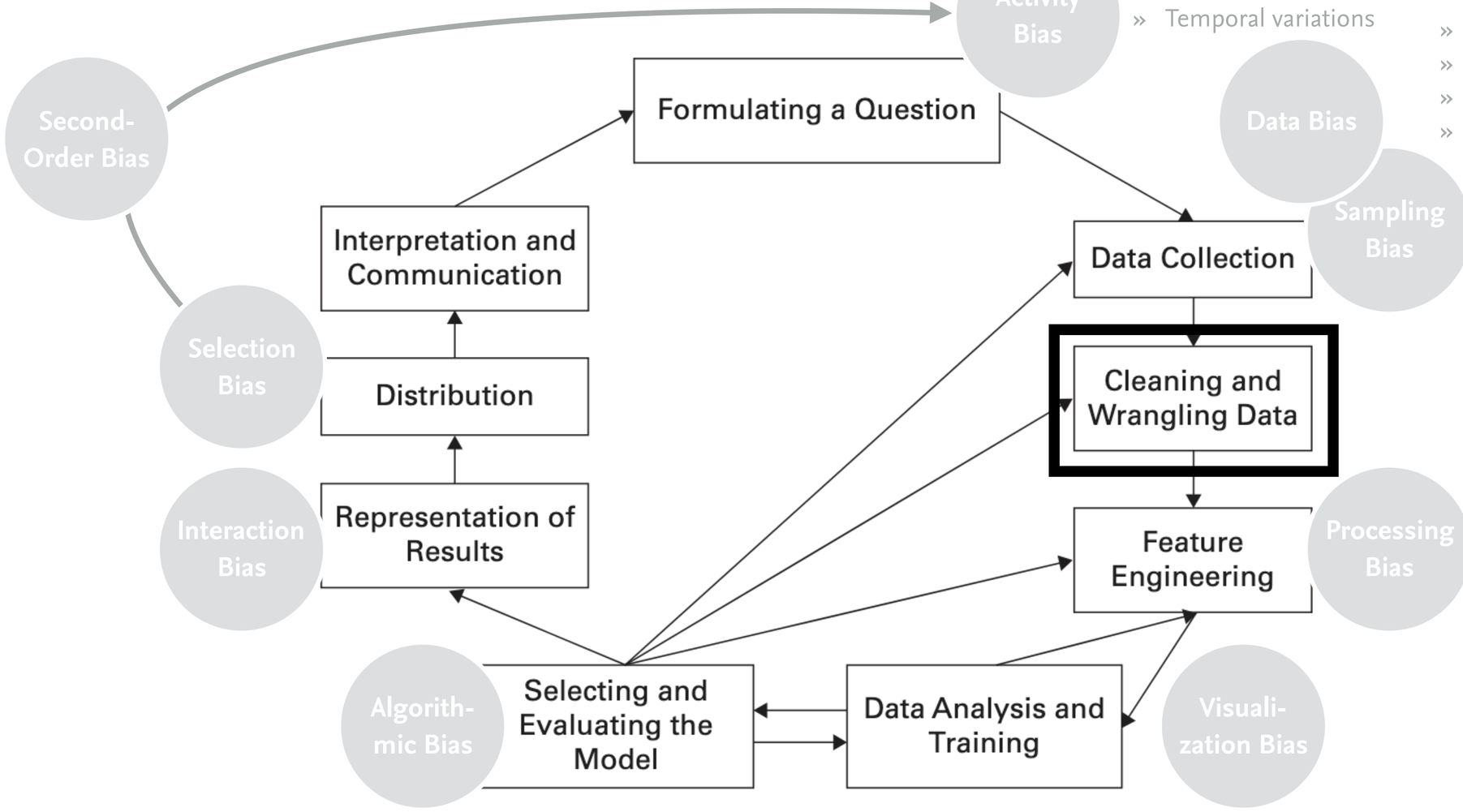


Diagram created by Moritz Hardt (unknown first time use)

Overview on Sources of Bias



Cleaning and Wrangling Data

Data wrangling refers to the often-messy work of getting data ready for analysis which might account for 80 to 90 percent of the time and effort of a data science project.

Main steps comprise the understanding, structuring, cleaning, enriching, and validating the data.

Starting Data

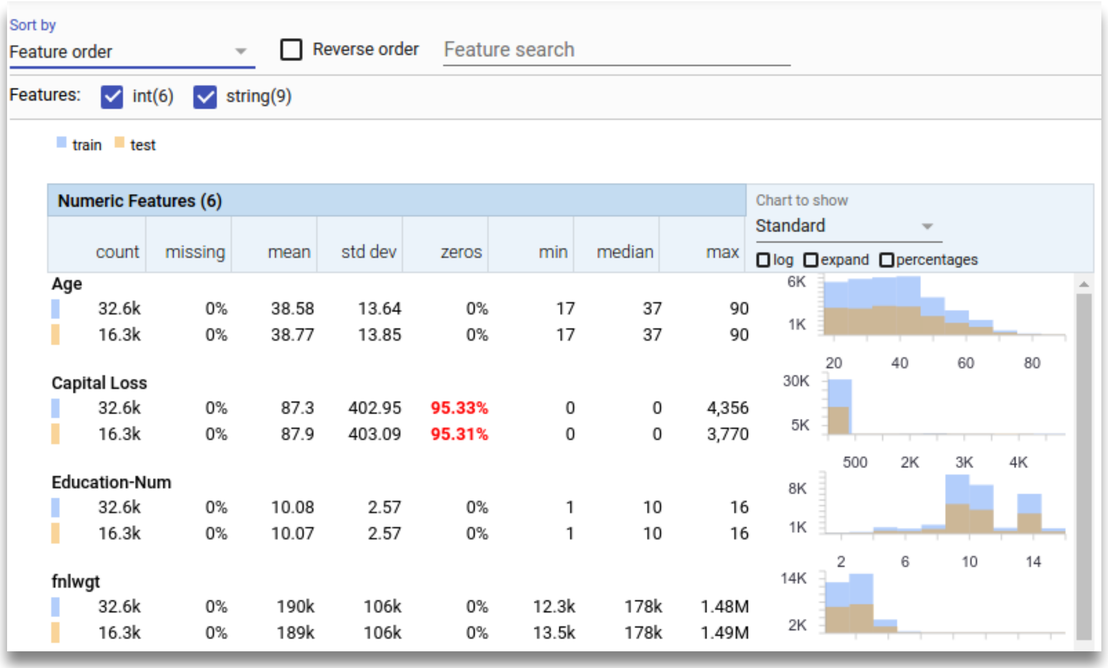
Name	Phone	Birth Date	State
John, Smith	445-881-4478	August 12, 1989	Maine
Jennifer Tal	+1-189-456-4513	11/12/1965	Tx
Gates, Bill	(876)546-8165	June 15, 72	Kansas
Alan Fitch	5493156648	2-6-1985	Oh
Jacob Alan	156-4896	January 3	Alabama

Result

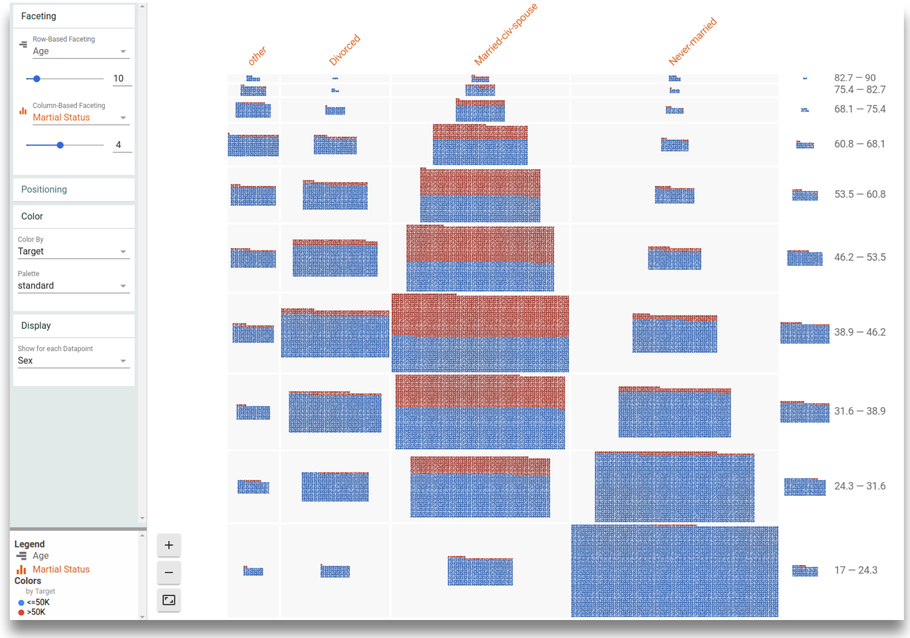
Name	Phone	Birth Date	State
John Smith	445-881-4478	1989-08-12	Maine
Jennifer Tal	189-456-4513	1965-11-12	Texas
Bill Gates	876-546-8165	1972-06-15	Kansas
Alan Fitch	549-315-6648	1985-02-06	Ohio

Aragon, Cecilia, Shion Guha, Marina Kogan, Michael Muller, und Gina Neff. Human-Centered Data Science: An Introduction. Cambridge, MA, USA: MIT Press, 2022.
Example from https://en.wikipedia.org/wiki/Data_wrangling

Understanding and Analyzing Datasets



Facets provides two visualizations for understanding and analyzing machine learning datasets consisting of **Facets Overview** and **Facets Dive**



It can be easily embedded into Jupyter notebooks or webpages.

The source code and WebApp is available on github.

Bias Mitigation Strategy

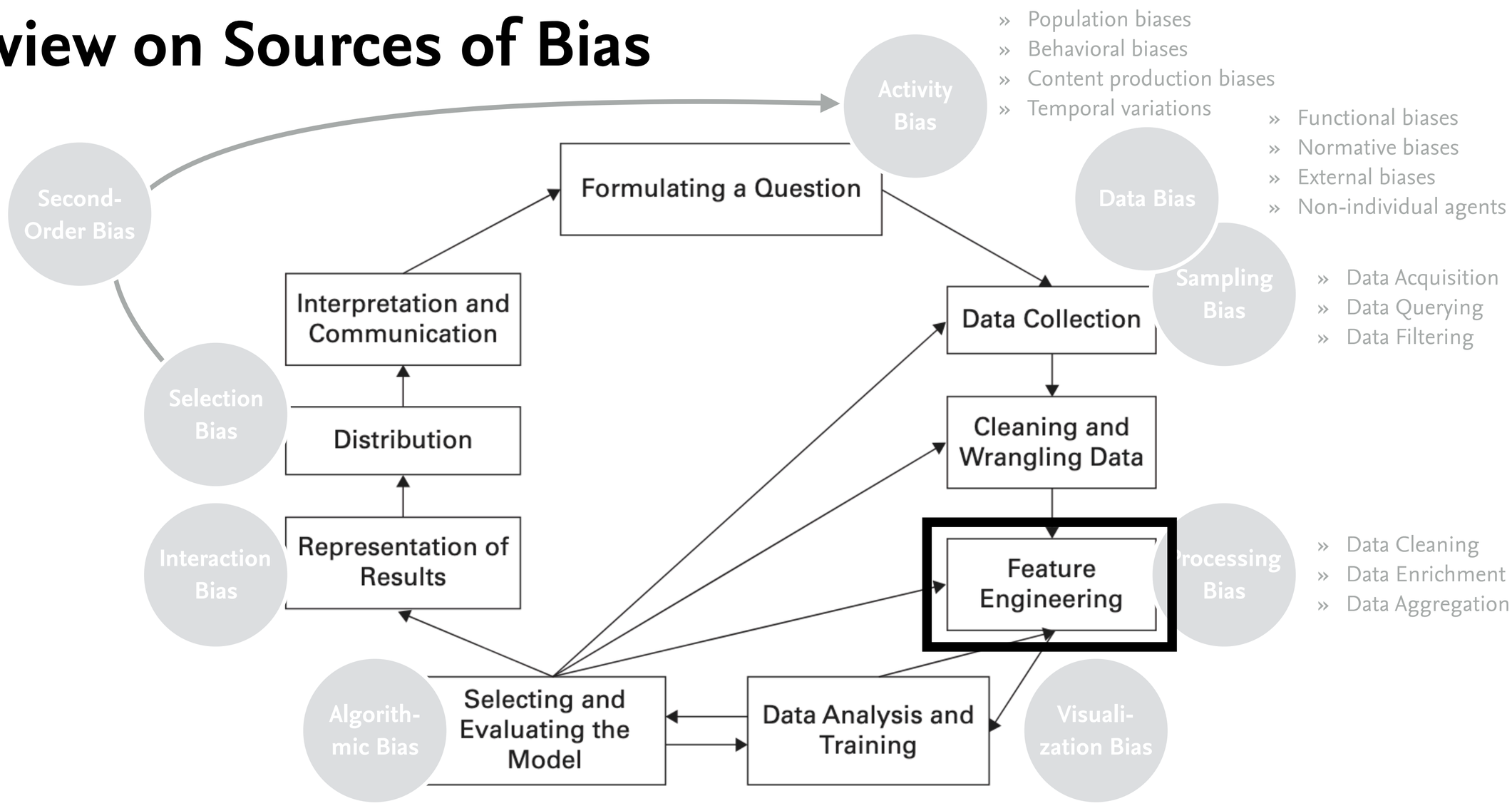
The aim is to produce a “balanced” dataset. The intuition behind these approaches is that the fairer the training data is, the less discriminative the resulting model will be.

Possible methods

- modify the original data distribution by altering class labels of carefully selected instances
- assign different weights to instances based on their group membership or
- carefully sample data from each group.

Ntoutsi, Eirini, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria Esther Vidal, Salvatore Ruggieri, u. a. 2020. „Bias in data−driven artificial intelligence systems—An introductory survey“. WIREs Data Mining and Knowledge Discovery 10 (3): 60–14. https://doi.org/10.1002/widm.1356.

Overview on Sources of Bias



Feature Engineering

> **Feature engineering is the practice of selecting existing variables (or features) in the dataset for inclusion in the model, as well as producing new combined features that capture additional aspects of the dataset.**
>
> (Aragon et al., 2022)

Feature engineering is often performed for dimensionality reduction for splitting split multi-class variables (one-hot encoding).

Aragon, Cecilia, Shion Guha, Marina Kogan, Michael Muller, und Gina Neff. Human-Centered Data Science: An Introduction. Cambridge, MA, USA: MIT Press, 2022.

Data Needs Context

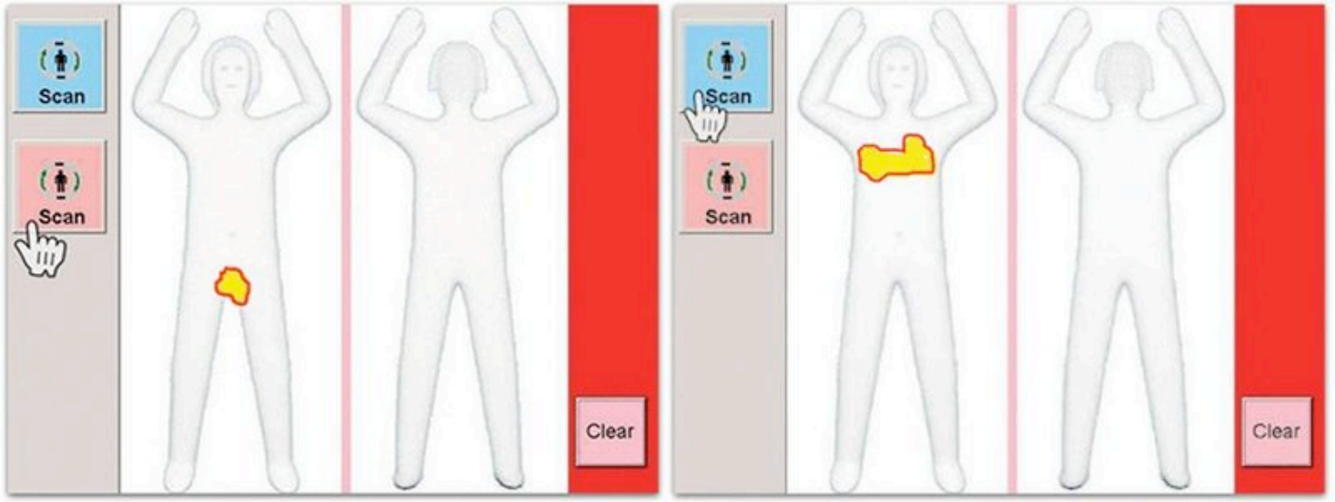
Your task as a data scientist is to craft an interpretation that will be useful for analysis and to record that interpretation as a class or subclass of data.

The act of crafting implies human intention, human action, and in many cases human creativity.

These acts of crafting have consequences .

Aragon, Cecilia, Shion Guha, Marina Kogan, Michael Muller, und Gina Neff. Human-Centered Data Science: An Introduction. Cambridge, MA, USA: MIT Press, 2022.

Example: Millimeter Wave Scanner



Costanza-Chock, S. (2018). Design justice: Towards an intersectional feminist framework for design theory and practice. Proceedings of the Design Research Society.

Recap: What is the Origin of Data?

Situated knowledges emphasis on **disclosing the mechanisms for the production of data.** These mechanisms for data production include social, cultural, historical and material conditions.

Additionally, a reflection on your **own perspective** is necessary but also **on existing values**.

Data need Context

Reflexivity is a precondition for restoring context in data creation.

D'Ignazio, C., & Klein, L. F. (2020). Data feminism. MIT Press.
Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. Feminist studies, 14(3), 575-599.

Thin (Big) Data Needs Thick Data

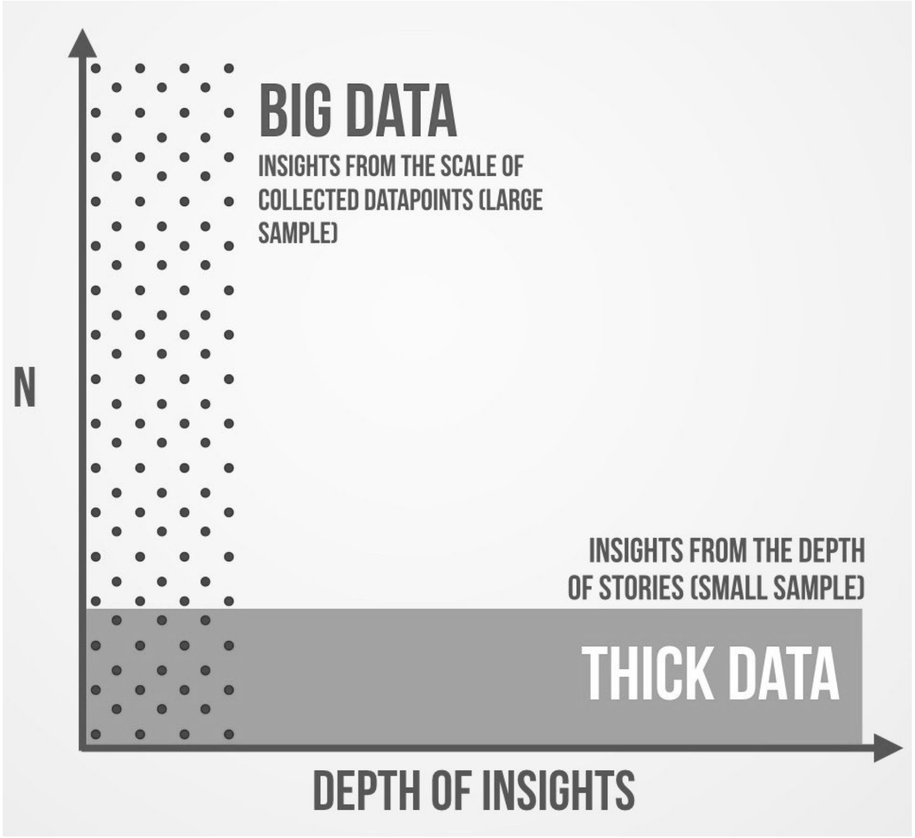
Thick Data

» Relies on Human Learning

» Reveals the social context of connections between data points

» Accepts irreducible complexity

» Loses scales



Thin (big) Data

» Relies on Machine Learning

» Reveals insights with a particular range of quantified data points

» Isolates variables to identify patterns

» Loses resolution

Tricia Wang (2016). Why Big Data Needs Thick Data. https://medium.com/ethnography-matters/why-big-data-needs-thick-data-b4b3e75e3d7

Combining Quantitative and Qualitative Methods

> Mixed methods research is [...] [a] class of research where the researcher mixes or combines quantitative and qualitative research techniques, methods, approaches, concepts or language into a single study.

Johnson, R. B., & Onwuegbuzie, A. J. (2016). Mixed Methods Research: A Research Paradigm Whose Time Has Come:. Educational Researcher, 33(7), 14–26. http://doi.org/10.3102/0013189X033007014

Strengths and Weaknesses of Mixed Research

Strengths

» Words, pictures, and narrative can be used to add meaning to numbers.

» Numbers can be used to add precision to words, pictures, and narrative.

» The strengths of an additional method can overcome the weaknesses in another method.

» Qualitative and quantitative research used together produce more complete knowledge necessary to inform theory and practice.

Weaknesses

» Can be difficult for a single researcher to carry out both qualitative and quantitative research.

» Researcher has to learn about multiple methods and approaches and understand how to mix them appropriately.

» More expensive and time consuming.

» A clear methodological guidance is still missing.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. Educational researcher, 33(7), 14-26.

Example:
Replication Study by Geiger & Halfaker

RESEARCH-ARTICLE OPEN ACCESS

Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'

ACM DL DIGITAL LIBRARY

Authors: R. Stuart Geiger, Aaron Halfaker Authors Info & Affiliations

Publication: Proceedings of the ACM on Human-Computer Interaction • December 2017 • Article No.: 49 • https://doi.org/10.1145/3134684

5 424

eReader PDF

R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 49 (November 2017), 33 pages. DOI:https://doi.org/10.1145/3134684

Recap: Results of Geiger & Halfaker's Replication Study

They could show that the overwhelming majority of bot-bot reverts constitute routine, productive, and even collaborative work between bots.

They define bots as assemblages of "code and a human developer" which are responsible for operating the bot in alignment with Wikipedia's complex policy environment.

Why these researchers arrived at this very different perspective.

R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 49 (November 2017), 33 pages. DOI:https://doi.org/10.1145/3134684

Methodological Approach: Trace Ethnography

Trace ethnography involves an ethnographer's socialization into a group **prior** to the ability to decode and interpret trace data.

Researchers need to learn how to follow and interpret log data as part of the lived and learned the experience of a community.

The understanding is based on documentary approaches by following documents as they "travel" across the platform by asking why, how, where, and by whom they are produced, edited and revised.

Geiger, R. S., & Ribes, D. (2011, January). Trace ethnography: Following coordination through documentary practices. In 2011 44th hawaii international conference on system sciences (pp. 1-10). IEEE.

Employed Approach by Geiger & Halfaker

> *"Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care."*

Geiger & Halfaker used the same database dumps of Wikipedia edit activity. Additionally, they "*interrogate, refine, supplement, and extend this dataset in an iterative manner,* ***integrating computational and ethnographic expertise*** *to distinguish cases of conflict from non-conflict at various scales*".

Tim Harford. 2014. Big data: A big mistake? Significance 11, 5 (2014), 14–19. https://doi.org/10.1111/j.1740-9713.2014. 00778.x
R. Stuart Geiger and Aaron Halfaker. 2017. Operationalizing Conflict and Cooperation between Automated Software Agents in Wikipedia: A Replication and Expansion of 'Even Good Bots Fight'. Proc. ACM Hum.-Comput. Interact. 1, CSCW, Article 49 (November 2017)

Bias Mitigation Strategy

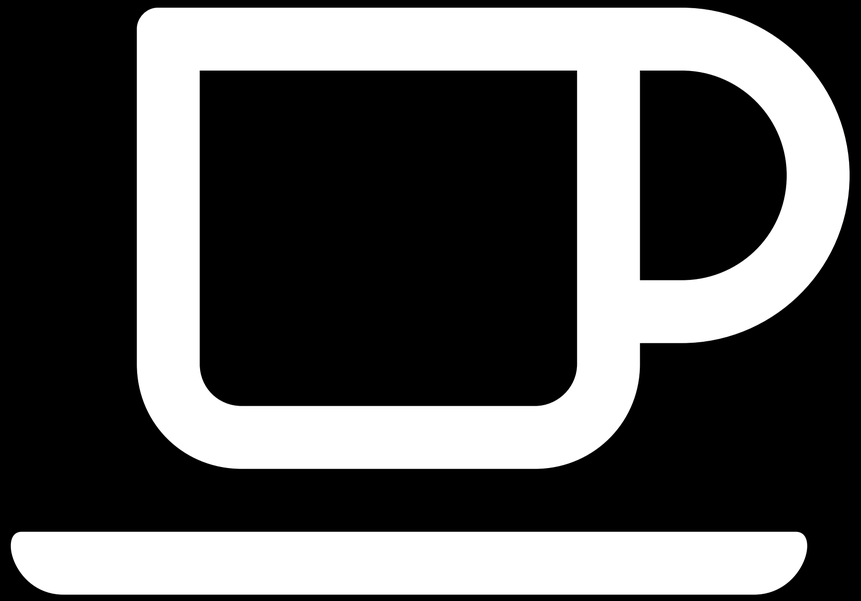
Ask not only "How does context give meaning to our data/design” but also "How does our data/design fit the context?”.

Connect data to viewpoints, interactions, histories, and local resources.

People's understanding of the world, of themselves and of interaction is strongly influenced by their different physical, historical, social and cultural situations. Thus, try to understand these situations.

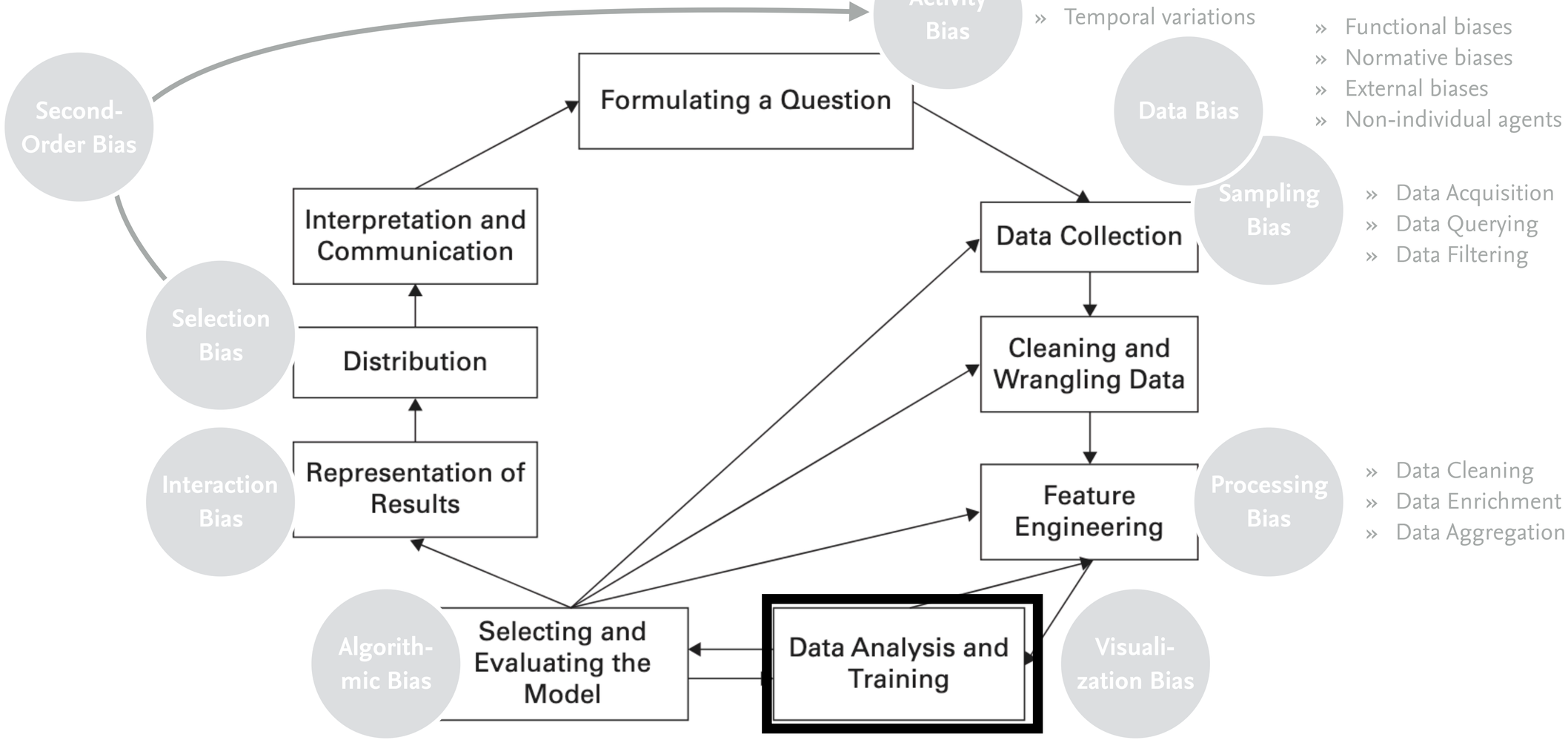
Evaluation is nothing that is universally valid (e.g. efficiency, effectivity), but something that depends on the context.

Harrison, Steve, Deborah Tatar, and Phoebe Sengers. "The three paradigms of HCI." Alt. Chi. Session at the SIGCHI Conference on human factors in computing systems San Jose, California, USA. 2007.



5 minutes break

Overview on Sources of Bias



Data Analysis and Training

In some data science techniques (supervised learning), labeled data is needed, i.e., data needs to be annotated.

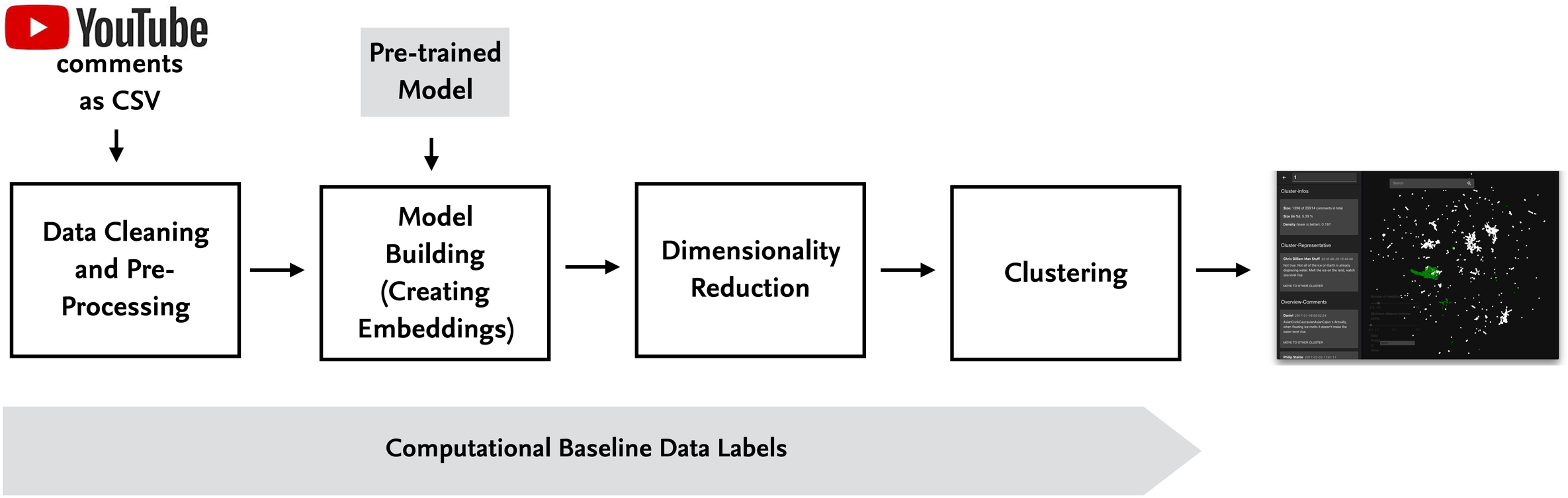
The labels are used as the ground truth —an accurate representation of the world with a high construct validity.

The choices made during labeling the data, for example, how the training is conducted, have implications for the results.

Aragon, Cecilia, Shion Guha, Marina Kogan, Michael Muller, und Gina Neff. Human-Centered Data Science: An Introduction. Cambridge, MA, USA: MIT Press, 2022.

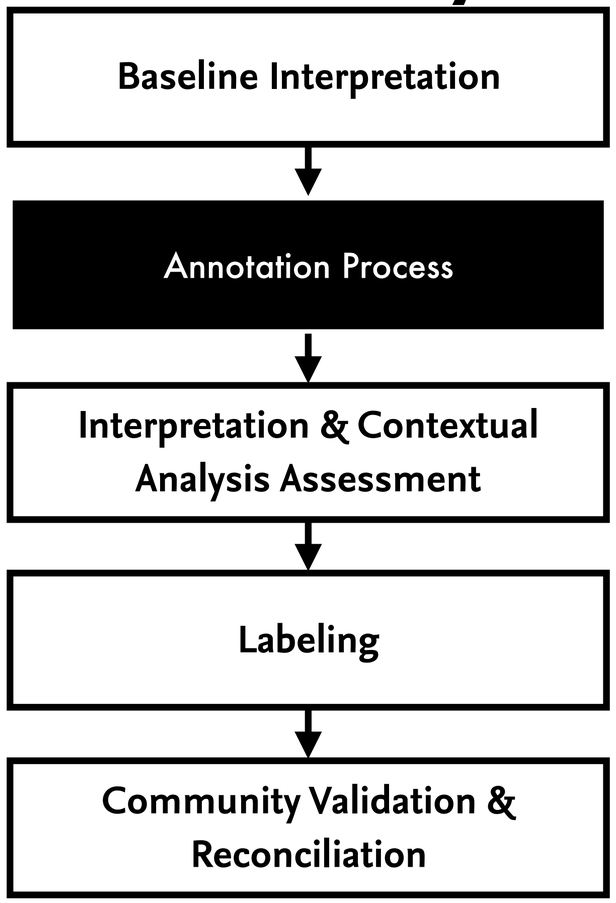
Example:
(Human) Data Annotation

Text Analysis of YouTube Comments



Hirsbrunner, S.; Tebbe, M. Müller-Birn, C. (2022): From critical technical practice to reflexive data science. Submission for Special Issue on "Critical Technical Practice(s) in Digital Research" in Convergence

Contextual Analysis of Social Media



→ **Contextualization** →

- » Original Social Media Post
- » Utilize Web-Based Resources
- » Original Post's Author
- » Peer Network
- » Offline Events
- » Virality
- » Engagement

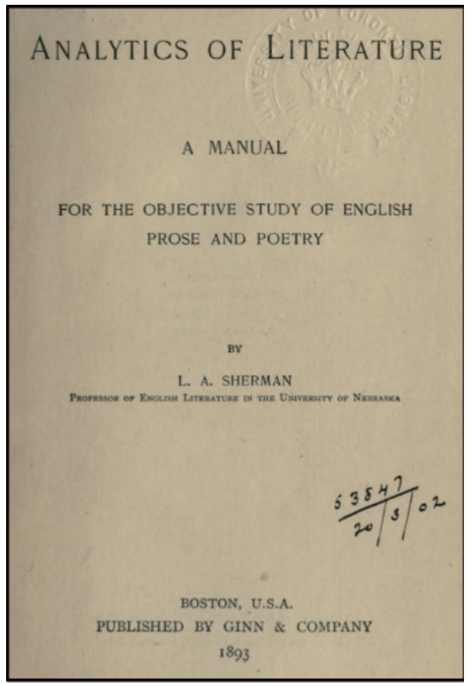
CASM utilizes a team-based approach to annotating and qualitatively analyzing social media data, explicitly grounded by community expertise and understanding.

Desmond U. Patton, William R. Frey, Kyle A. McGregor, Fei-Tzin Lee, Kathleen McKeown, and Emanuel Moss. 2020. Contextual Analysis of Social Media: The Promise and Challenge of Eliciting Context in Social Media Posts with Natural Language Processing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20). Association for Computing Machinery, New York, NY, USA, 337–342.

Structured Content Analysis (or closed coding)

An established method in the humanities and social sciences for generations.

> **...a systematic and replicable method...**
>
> (Riff, Lacy, and Frederick 2013)



ANALYTICS OF LITERATURE

A MANUAL

FOR THE OBJECTIVE STUDY OF ENGLISH PROSE AND POETRY

BY

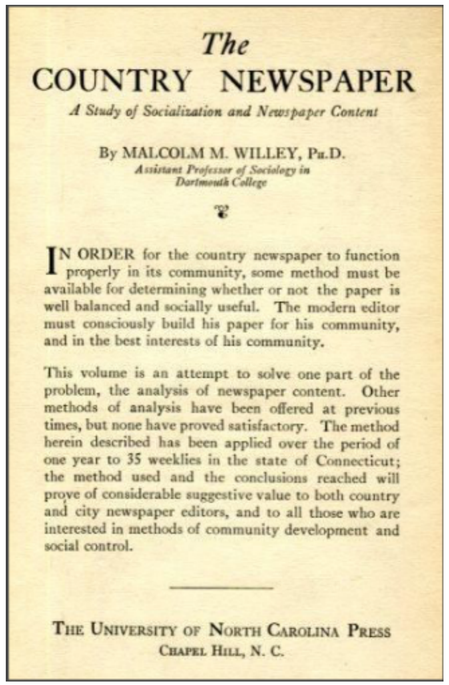
L. A. SHERMAN

PROFESSOR OF ENGLISH LITERATURE IN THE UNIVERSITY OF NEBRASKA

53547
20/3/02

BOSTON, U.S.A.
PUBLISHED BY GINN & COMPANY
1893

Sherman (1893)



The
COUNTRY NEWSPAPER
A Study of Socialization and Newspaper Content

By MALCOLM M. WILLEY, Ph.D.
Assistant Professor of Sociology in Dartmouth College

IN ORDER for the country newspaper to function properly in its community, some method must be available for determining whether or not the paper is well balanced and socially useful. The modern editor must consciously build his paper for his community, and in the best interests of his community.

This volume is an attempt to solve one part of the problem, the analysis of newspaper content. Other methods of analysis have been offered at previous times, but none have proved satisfactory. The method herein described has been applied over the period of one year to 35 weeklies in the state of Connecticut; the method used and the conclusions reached will prove of considerable suggestive value to both country and city newspaper editors, and to all those who are interested in methods of community development and social control.

THE UNIVERSITY OF NORTH CAROLINA PRESS
CHAPEL HILL, N. C.

Willey (1923)

Slide adapted from Stuart Geiger (2020) https://stuartgeiger.com/assets/gigofat2020_slides.pdf

Best Practices of Structured Content Analysis

1. **Define a "coding scheme"** with procedures, definitions, and examples.
2. **Recruit and train multiple "coders"** (or "annotators", "labelers", or "reviewers") with the coding scheme.
3. Have **coders independently code at least a portion of the same items**, then calculate "inter-annotator agreement" or "inter-rater reliability."
4. Define and **follow a process of "reconciliation" for disagreements**, e.g. majority rule, talk to consensus, expert/leader decides.
5. **Modify coding scheme**, training, and/or reconciliation as needed.

Slide adapted from Stuart Geiger (2020) https://stuartgeiger.com/assets/gigofat2020_slides.pdf

Human-Labeled Training Data: Garbage in, garbage out?

Motivation: Many of the ethical issues that arise in data-driven applications can be traced back to the quality of training data.

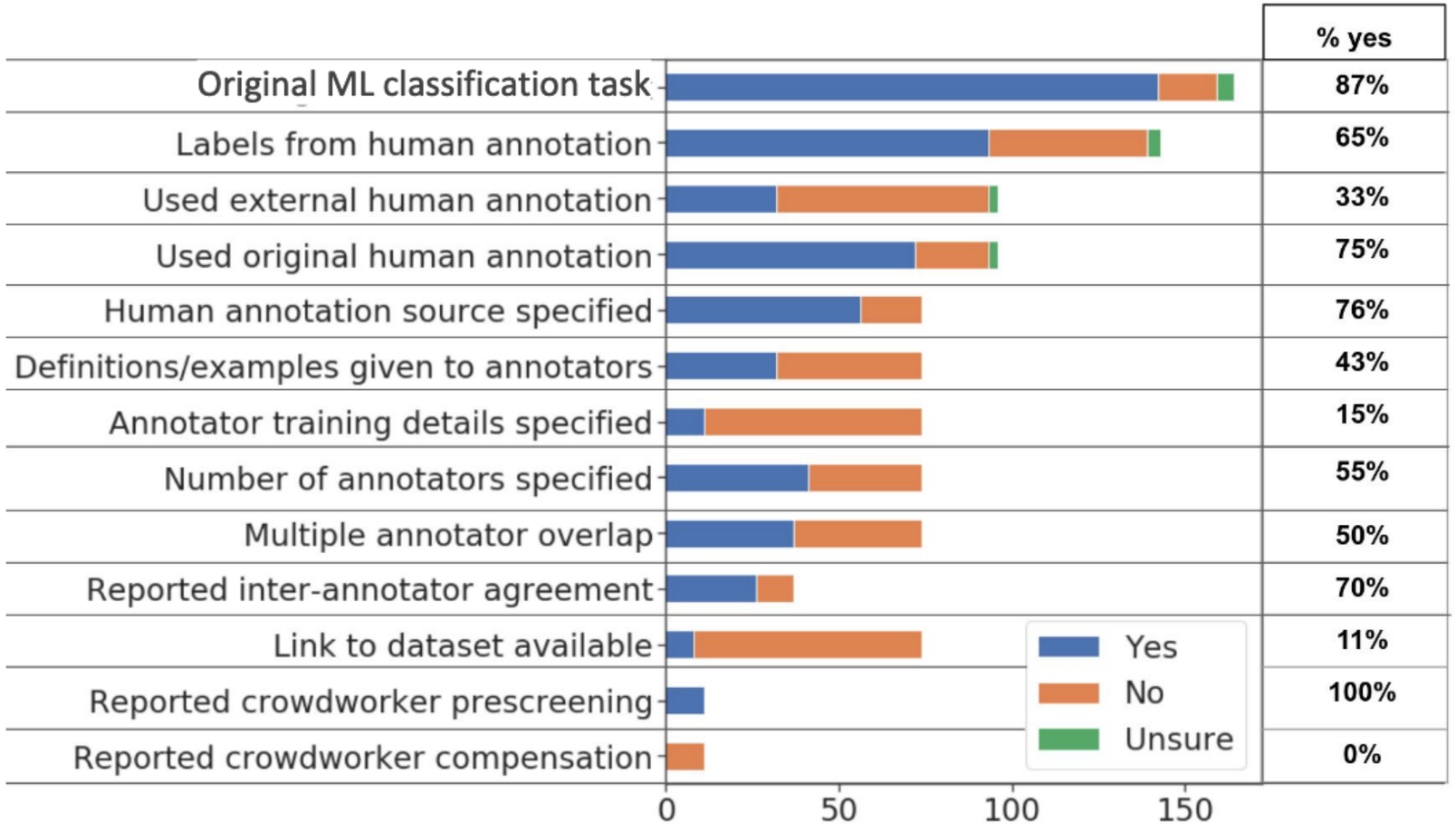
Approach: The way training data is labeled by humans is often a form of structured content analysis, which has established best practices. Sampling of 164 papers published from 2010 to 2018.

Research Question: How many papers in an application domain of ML — classifiers trained on tweets — report following these practices?

Results: Existing practices vary substantially, showing need for more focus on data labeling practices in ML education, evaluation, and regulation.

Slide adapted from Stuart Geiger (2020) R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 325–336.

Study Results



Slide adapted from Stuart Geiger (2020) R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 325–336.

Bias Mitigation Strategy for Data Labeling

Human annotation and labeling is as important task in data science but it is challenging.

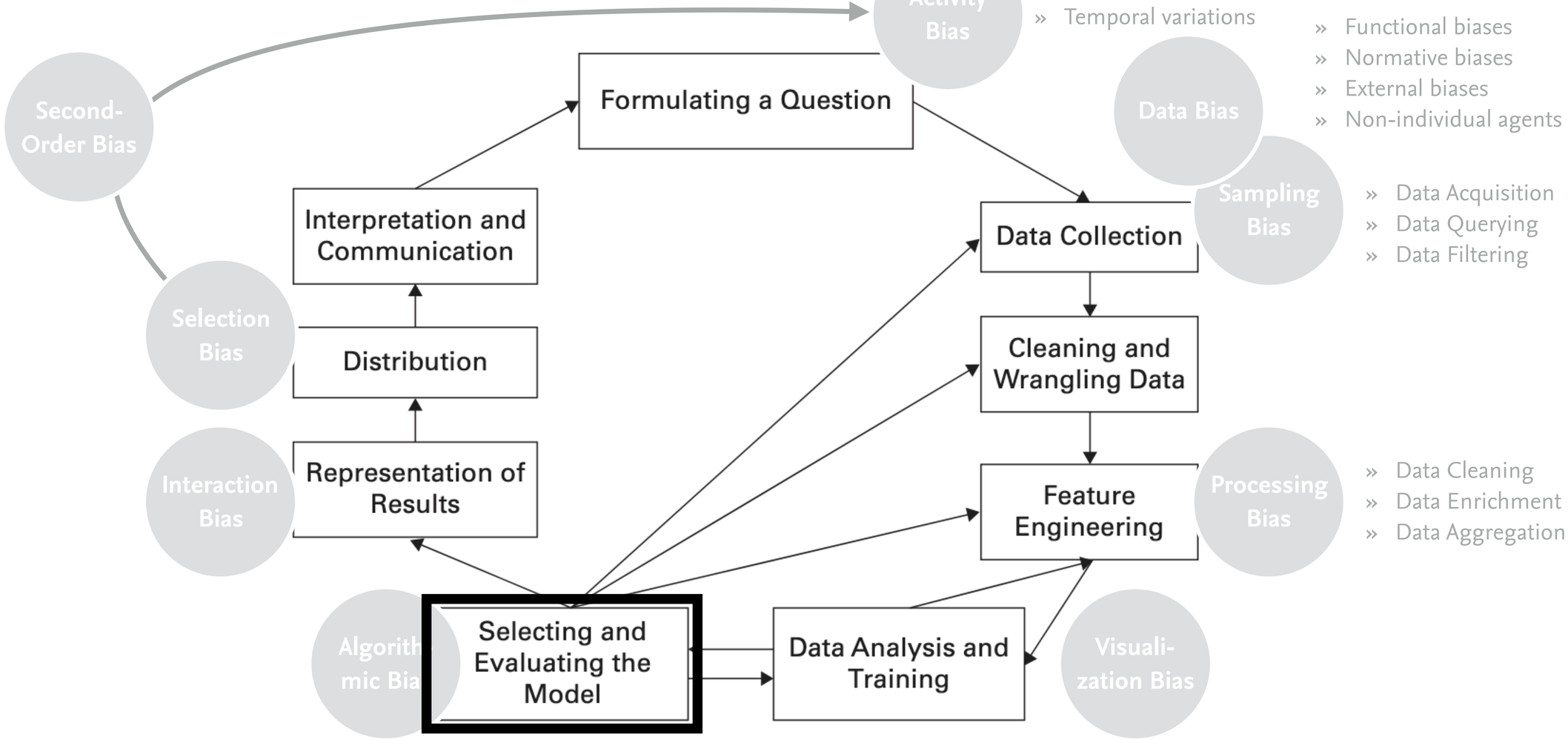
Operationalization & construct validity decisions play out in the design of human annotation processes. These decisions should be made explicit!

Consider best practice from *structured content analysis*.

For data science projects consider: Is the labeling process described enough so any reader can, with sufficient resources, independently produce a substantively similar dataset?

Slide adapted from Stuart Geiger (2020) R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, 325–336.

Overview on Sources of Bias



One Dataset - Many Choices

29 different data teams

One data set

One research question

Are soccer referees more likely to give red cards to dark skin toned players than light skin toned players?

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F.,... Nosek, B. A. (2017, September 21). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. http://doi.org/10.17605/OSF.IO/QKWST

One Dataset - Many Choices

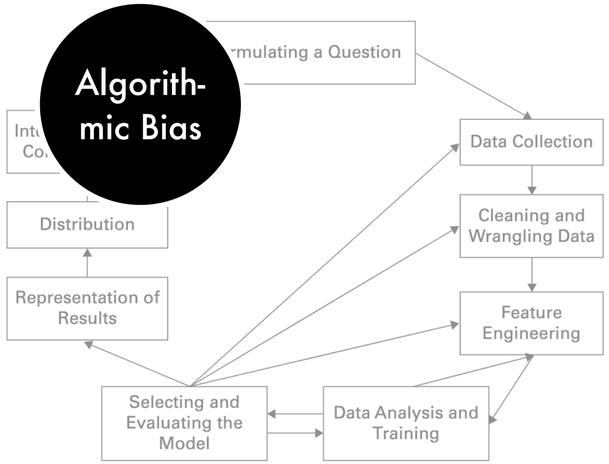
Team	Analytic Approach
10	Multilevel regression and logistic regression
1	Ordinary least squares with robust standard errors, logistic regression
4	Spearman correlation
14	Weighted least squares regression
11	Multiple linear regression
6	Linear Probability Model
17	Bayesian logistic regression
15	Hierarchical log-linear modeling
31	Logistic regression
30	Clustered robust binomial logistic
3	Multilevel Binomial Logistic Regr
23	Mixed model logistic regression
2	Linear probability model, logistic
5	Generalized linear mixed models
24	Multilevel logistic regression
28	Mixed effects logistic regression
32	Generalized linear models for bina
8	Negative binomial regression with
25	Multilevel logistic binomial regres
9	Generalized linear mixed effects m
7	Dirichlet process Bayesian clusteri
21	Tobit regression
12	Zero-inflated Poisson regression
26	Three-level hierarchical generalize
16	Hierarchical Poisson Regression
20	Cross-classified multilevel negativ
13	Poisson Multi-level modeling
27	Poisson regression
32	Generalized linear models for binary data

Multiple linear regression
Linear Probability Model
Bayesian logistic regression
Hierarchical log-linear modeling
Logistic regression
Clustered robust binomial logistic regressio
Multilevel Binomial Logistic Regression us
Mixed model logistic regression
Linear probability model, logistic regressio
Generalized linear mixed models
Multilevel logistic regression
Mixed effects logistic regression

Twenty teams (69%) found a statistically significant positive effect and nine teams (31%) observed an insignificant relationship.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F.,... Nosek, B. A. (2017, September 21). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. http://doi.org/10.17605/OSF.IO/QKWST

Challenges Modeling Data



» Using the right method, in the right place, and at the right time.

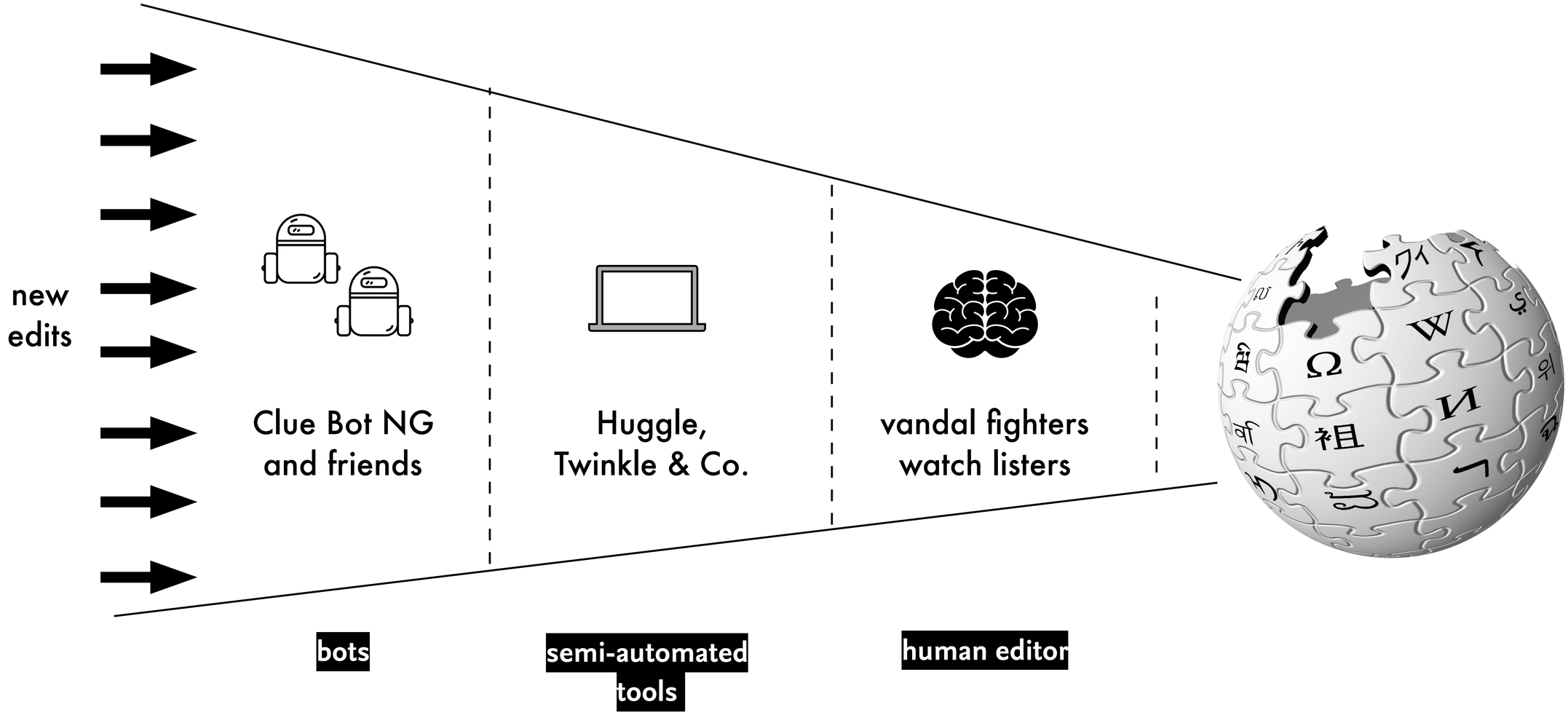
» There are performance variations across and within datasets.

» The composition of test and training data samples impacts the results.

» Distinct target variables, class labels, or data representations may lead to different results.

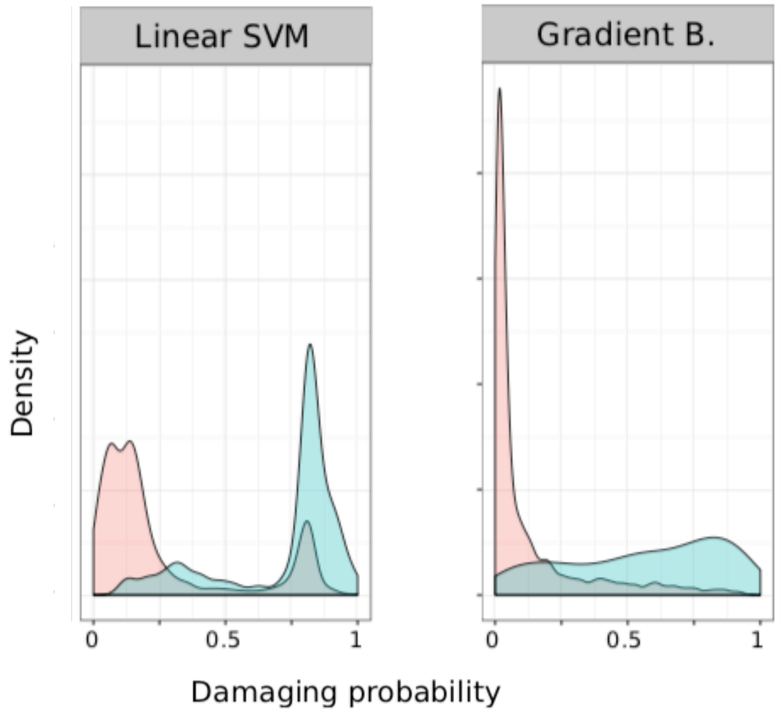
Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 13. http://doi.org/10.3389/fdata.2019.00013

Example (1): ML-driven Quality Assurance System

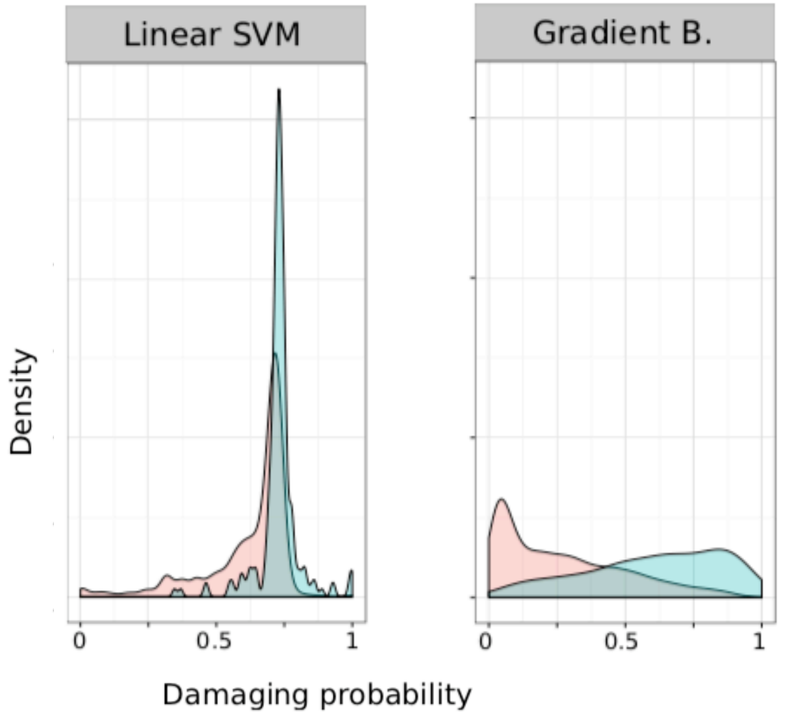


Example (2): Algorithmic Bias

No user-related features



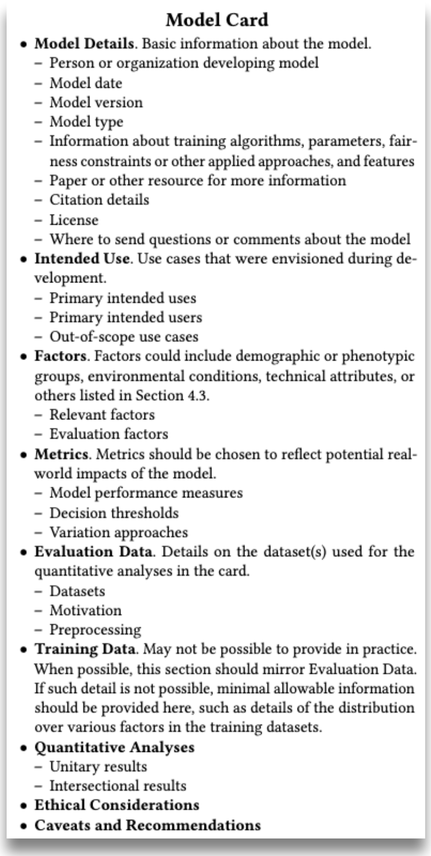
Everyone is newly registered



Halfaker, A., & Geiger, R. S. (2020). ORES: Lowering Barriers with Participatory Machine Learning in Wikipedia. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2), 1-37.

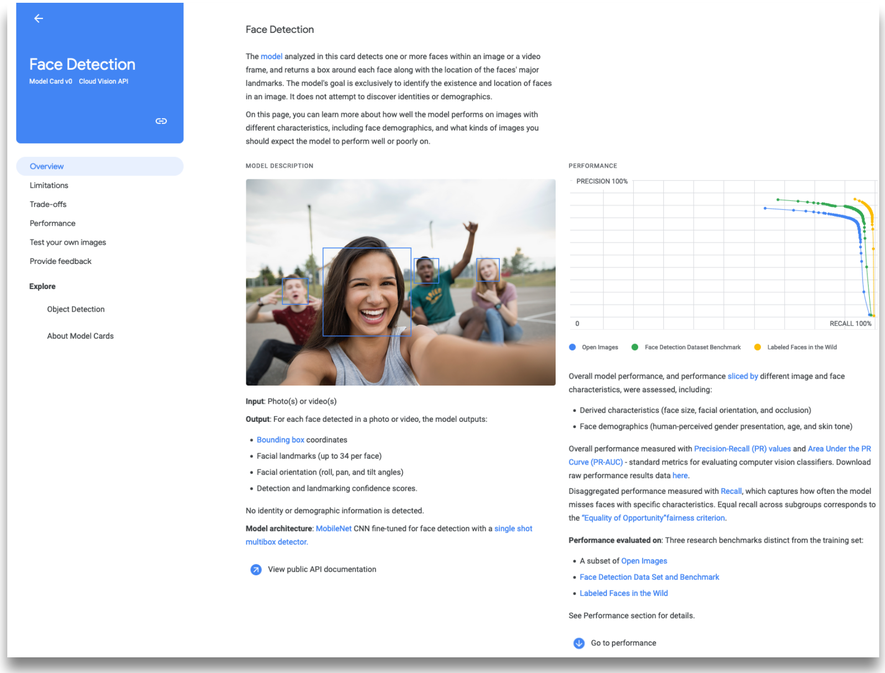
Mitigation Strategy: Model Cards for Model Reporting

Model cards is a framework that encourages transparent model reporting. It is designed as a short documentation that accompany trained machine learning models.



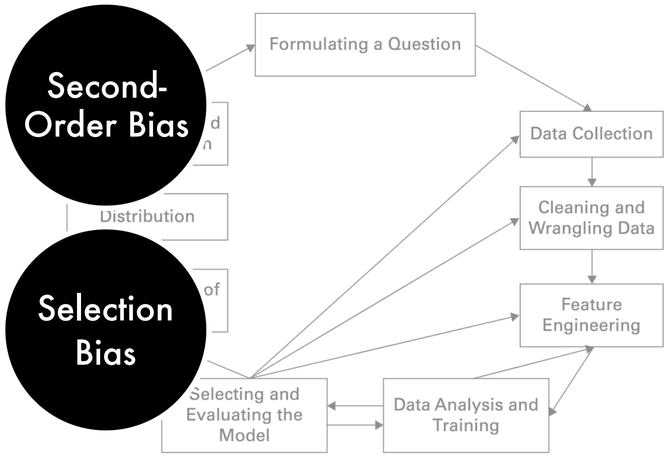
Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**



Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, und Timnit Gebru. 2019. „Model Cards for Model Reporting.“ FAT, Januar, 220–29. https://doi.org/10.1145/3287560.3287596. Image taken from https://modelcards.withgoogle.com/face-detection#overview
Further information: https://github.com/tensorflow/model-card-toolkit

Types of Emergent Bias

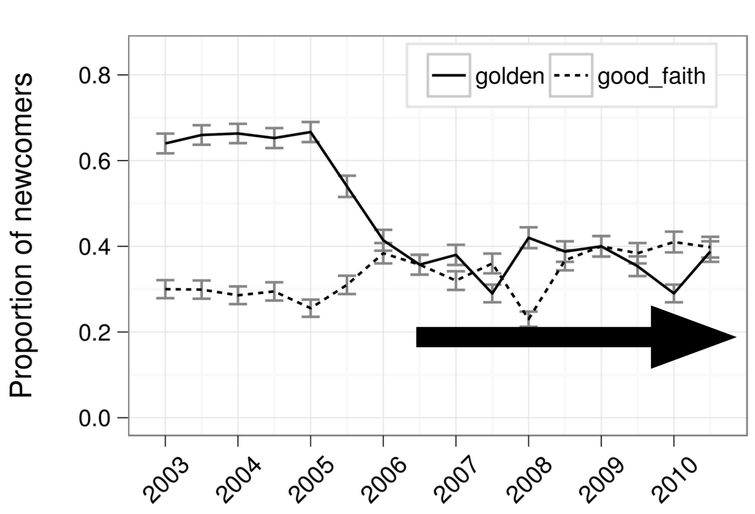


Self-selection bias	Exists if a user selects themselves to a certain group, for example, in the form of a social bias, such as “social conformity” or certain tech usages.
Second-Order Bias	Bias creates bias.

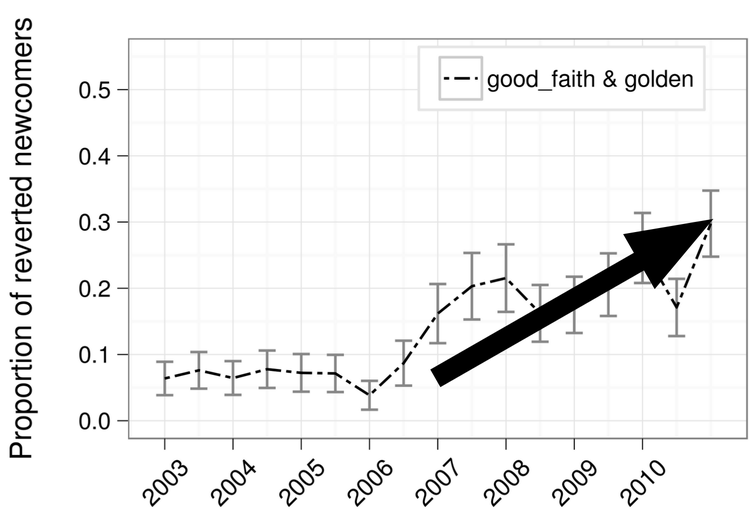
Baeza-Yates, R. (2018). Bias on the web. Commun. ACM, 61(6), 54–61.

Example (2): Effects of Automating “Spam” Detection

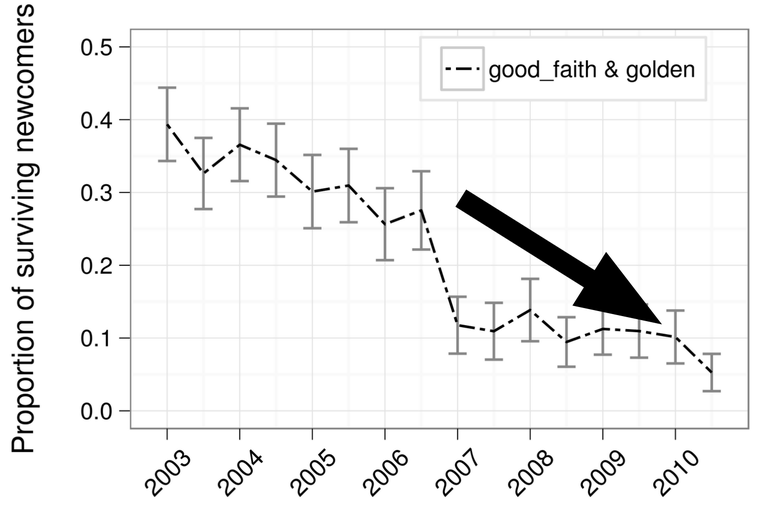
Stable quality of first contributions from new editors



Increasing rejection rate of these first contributions

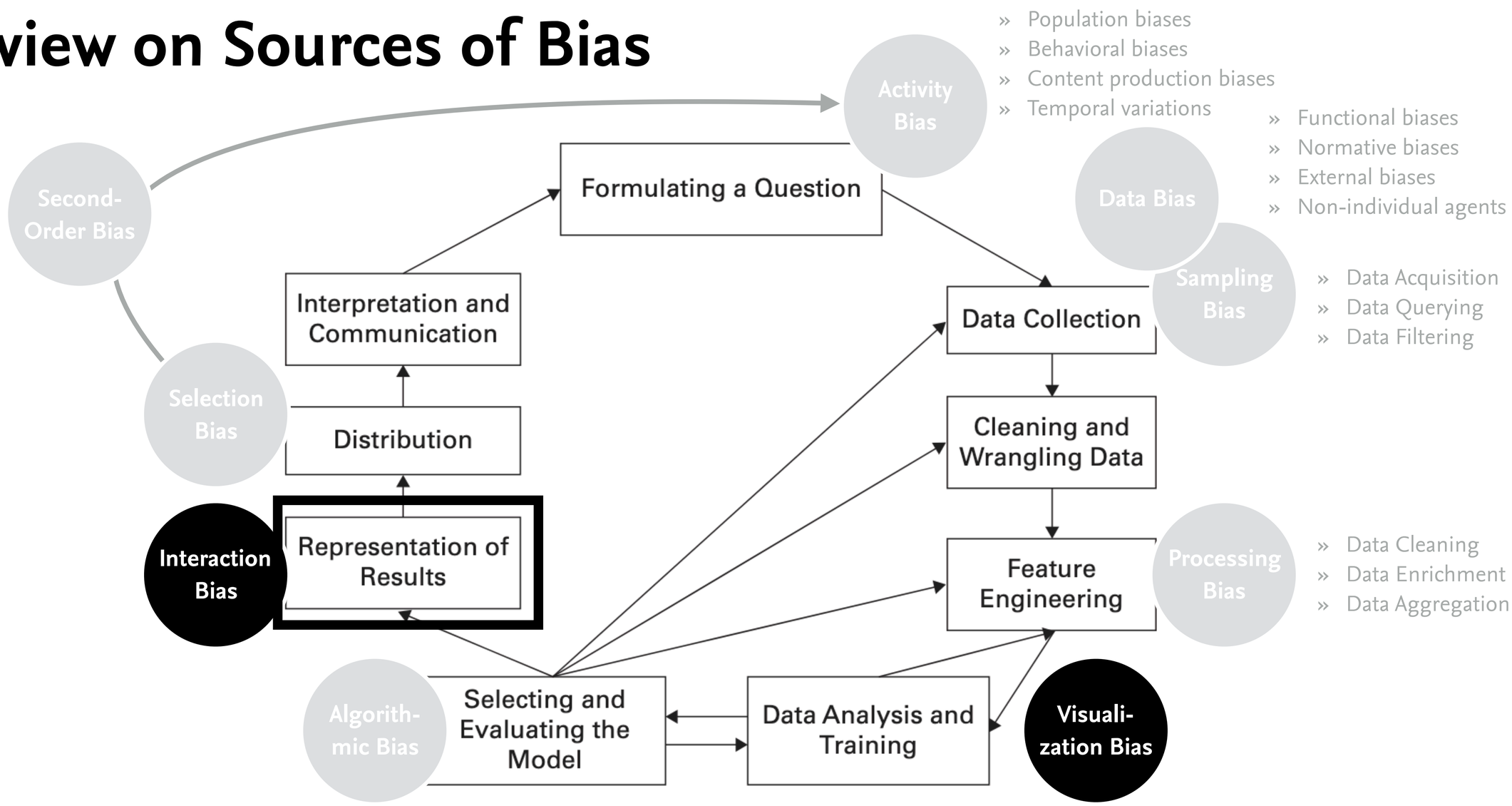


Decreasing binding rate of new editors

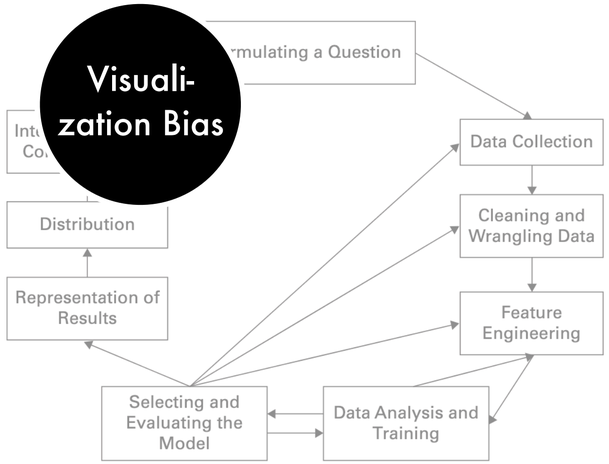


Aaron Halfaker, R. Stuart Gieger, Jonathan Morgan & John Riedl. (2013). The Rise and Decline of an Open Collaboration System: How Wikipedia's reaction to sudden popularity is causing its decline. American Behavioral Scientist, 57(5), 664-688. https://www-users.cs.umn.edu/~halfaker/publications/The_Rise_and_Decline/

Overview on Sources of Bias



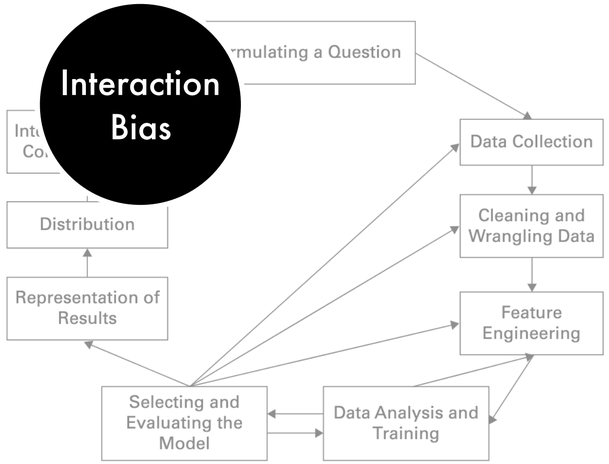
Types of Visualization (Perceptual) Bias



Clustering Illusion	A bias that explains why people see patterns in small sets of random data. People underestimate the consequence of variance and how even little sets of random data might have clustered data.
Weber-Fechner Law	The Weber-Fechner Law indicates that differences between stimuli are detected on a logarithmic scale.
Priming Biases	Priming relates to findings from theories of associative memory. It refers to the idea that concepts are more quickly activated after a similar concept has been activated.

Baeza-Yates, R. (2018). Bias on the web. Commun. ACM, 61(6), 54–61.

Types of Interaction Bias



Presentation biases	Can be caused by everything seen by the user can get clicks while everything else gets no clicks.
Position biases	Can be caused by the position of the item on the user interface. An example is the ranking bias.
User interaction biases	The provided interaction possibilities on the user interface provide the way people can approach the items. For example, content near images has a greater probability of being clicked.

Baeza-Yates, R. (2018). Bias on the web. Commun. ACM, 61(6), 54–61.

Take Aways for Your Critical-Reflexive Practice

Reconsider the collection process: What mechanisms or procedures were used to collect the data? What was the sampling strategy? Who was involved in the data collection process? What was the timeframe?

Reflect on pre-processing: Was any preprocessing/cleaning/labeling of the data done? Is the raw data saved in addition? Is the software used available? What approaches that aim to produce a "balanced" dataset that represented the target population (→ *fairness*)

Question the algorithmic choice: How robust is your model? Did you employed different models (and data) and compared their outcome? Do you understand your model? (→ *transparency*)

Check for result presentation: Do you know your users? Do you know the context of use? Did you evaluate different user interfaces? Can the results be verified? (→ *explainability, interpretability*)

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. Science, 346(6213), 1063-1064.

Check your Insights

What is the difference between bias and discrimination? How does it relate to fairness?

What are the three general types of biases and how does technology design and emergent bias relate?

How can you mitigate bias when cleaning data or during feature engineering?

In what context the use of qualitative data might be valuable for your data science practice?

Can you avoid bias in result presentation?

«Human-Centered Data Science»

Next week: Approaching Fairness in Data Science Workflows

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

June 9, 2022