



«Human-Centered Data Science»

Examining the Role of Data in Your Data Science Practice

Prof. Dr. Claudia Müller-Birn

Human-Centered Computing, Institute of Computer Science

Freie Universität Berlin

May 19, 2022

Lecture Overview

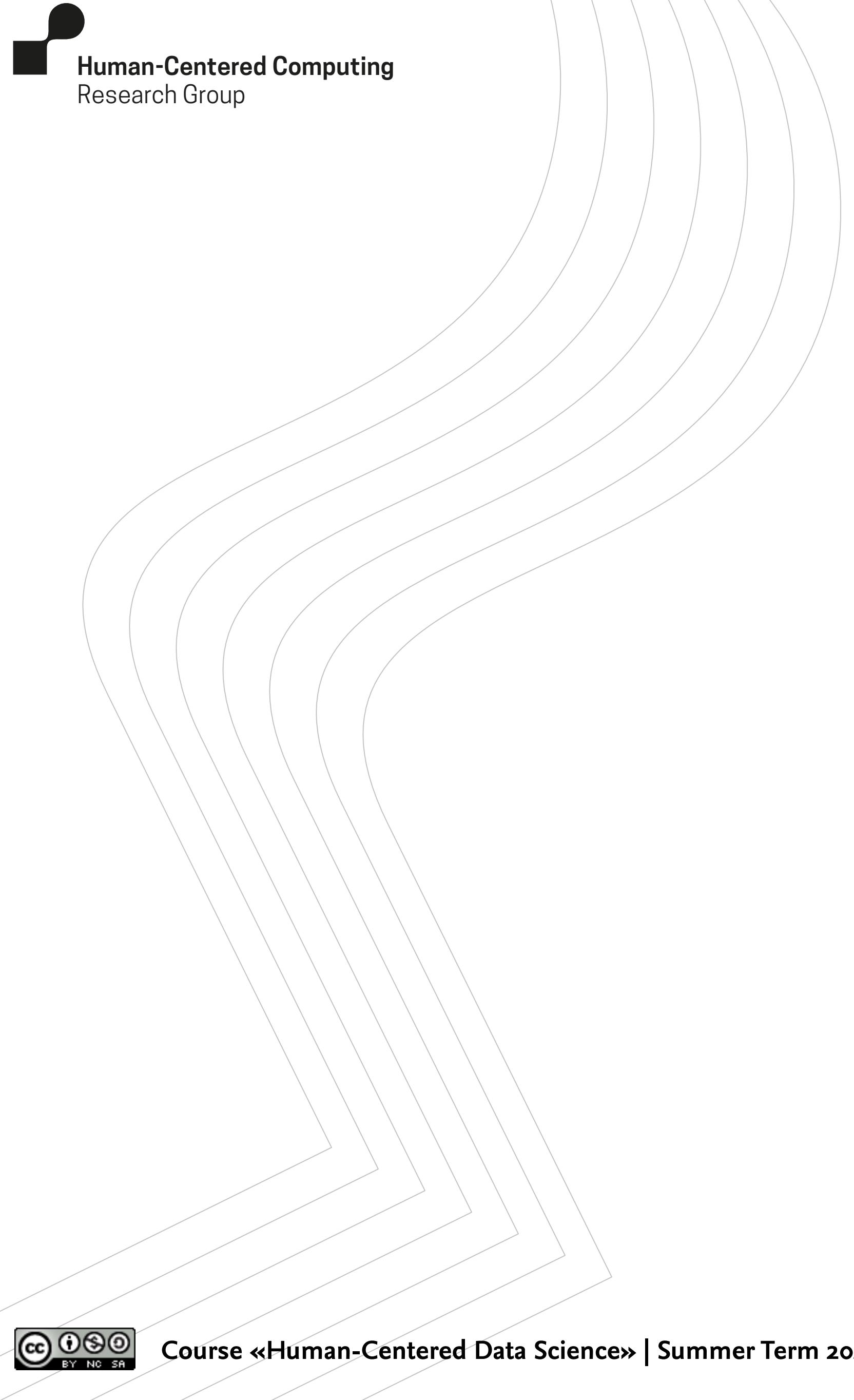
Recap & Motivation

Bias in Data Collection and Processing (Defining bias, overview on bias in the data science pipeline, discussing different types of bias and examples)

Break

Approaches to Mitigate Bias (datasheets for datasets, structure of the datasheets for datasets, other approaches and example), documentation as reflexive practice)





Recap



Incorporating Principles in Your Data Science Practice

Principles

Professional Responsibility

Promotion of Human Values

Fairness and Non-discrimination

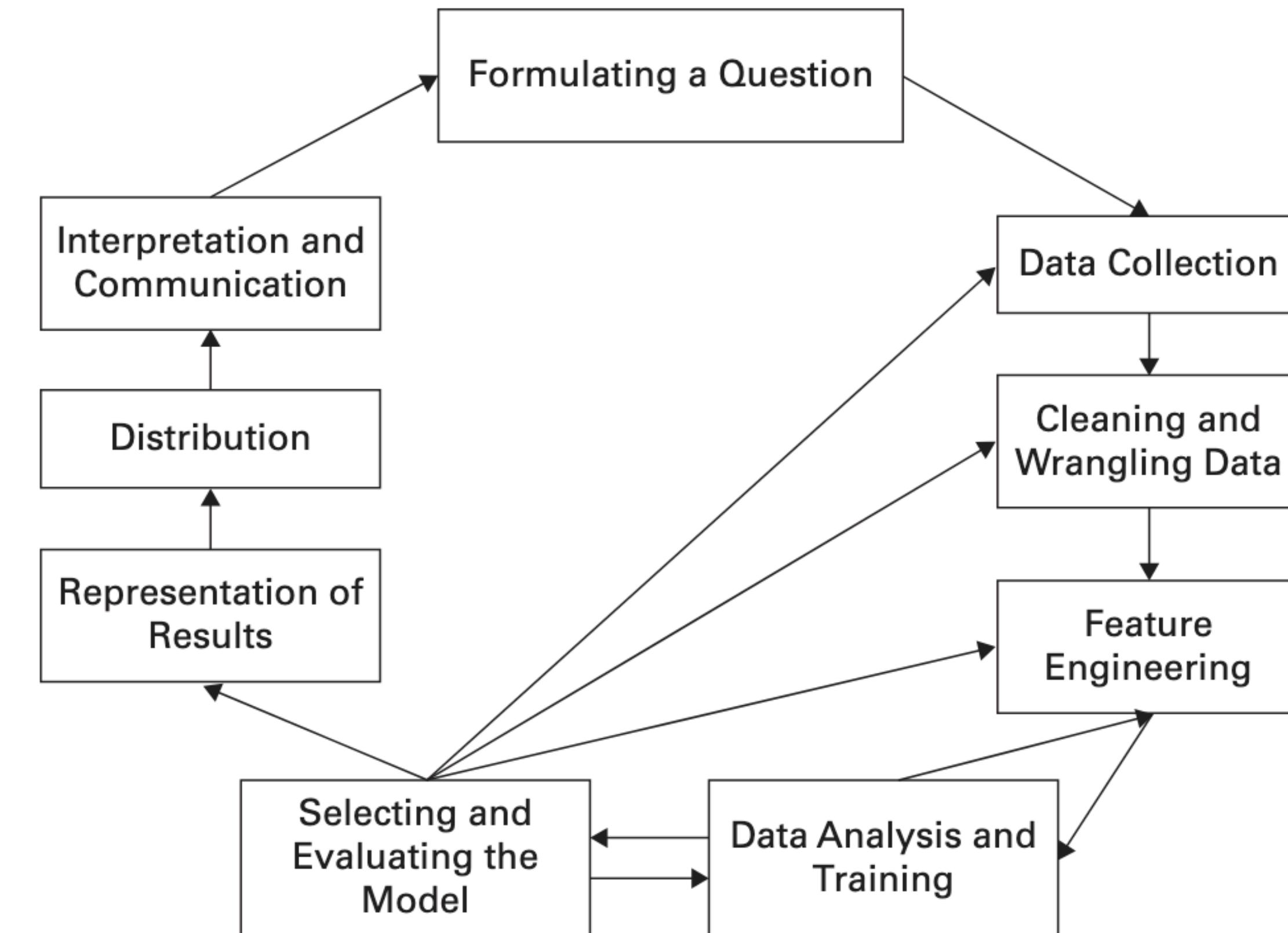
Human Control of Technology

Safety and Security

Transparency and Explainability

Accountability

Privacy





Motivation



Text Data

The screenshot shows a translation interface with two main sections. On the left, the input text is "She is a doctor.
He is a nurse." with a character count of 31/5000. On the right, the translated text is "O bir doktor.
O bir hemşire." with a character count of 31/5000. Both sections include icons for audio, microphone, keyboard, and sharing.

The screenshot shows a more detailed translation interface. The top navigation bar includes tabs for English, Turkish, Spanish, and "Turkish - c". The main area displays the input text "O bir doktor.
O bir hemşire" and its German translation "Ärztin (feminin)" and "Arzt (maskulin)". The interface includes tabs for English, German, French, and "DEUTSCH" (selected). It also shows a note about gender-specific translations and a character count of 6 / 5000.

Language Model

nature machine intelligence

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

[nature](#) > [nature machine intelligence](#) > [comment](#) > [article](#)

Comment | [Published: 17 June 2021](#)

Large language models associate Muslims with violence

[Abubakar Abid](#), [Maheen Farooqi](#) & [James Zou](#) 

[Nature Machine Intelligence](#) 3, 461–463 (2021) | [Cite this article](#)

1041 Accesses | 2 Citations | 72 Altmetric | [Metrics](#)

Large language models, which are increasingly used in AI applications, display undesirable stereotypes such as persistent associations between Muslims and violence. New approaches are needed to systematically reduce the harmful bias of language models in deployment.



Natural language processing (NLP) research has seen substantial progress on a variety of applications through the use of large pretrained language models^{1,2,3,4}. Although these increasingly sophisticated language models are capable of generating complex and cohesive

Quelle: <https://www.nature.com/articles/s42256-021-00359-2>

Image Data

 diri noir avec banan
@jackyalcine

Google Photos, y'all [REDACTED] up. My friend's not a gorilla.

Skyscrapers Airplanes Cars

Bikes Gorillas Graduation

© Twitter – @jackyalcine



Public Employment Service Austria



High chances

Opportunities for integration within the
next 7 months > 66%

Middle chances

Group on which resources are to be
concentrated in the future

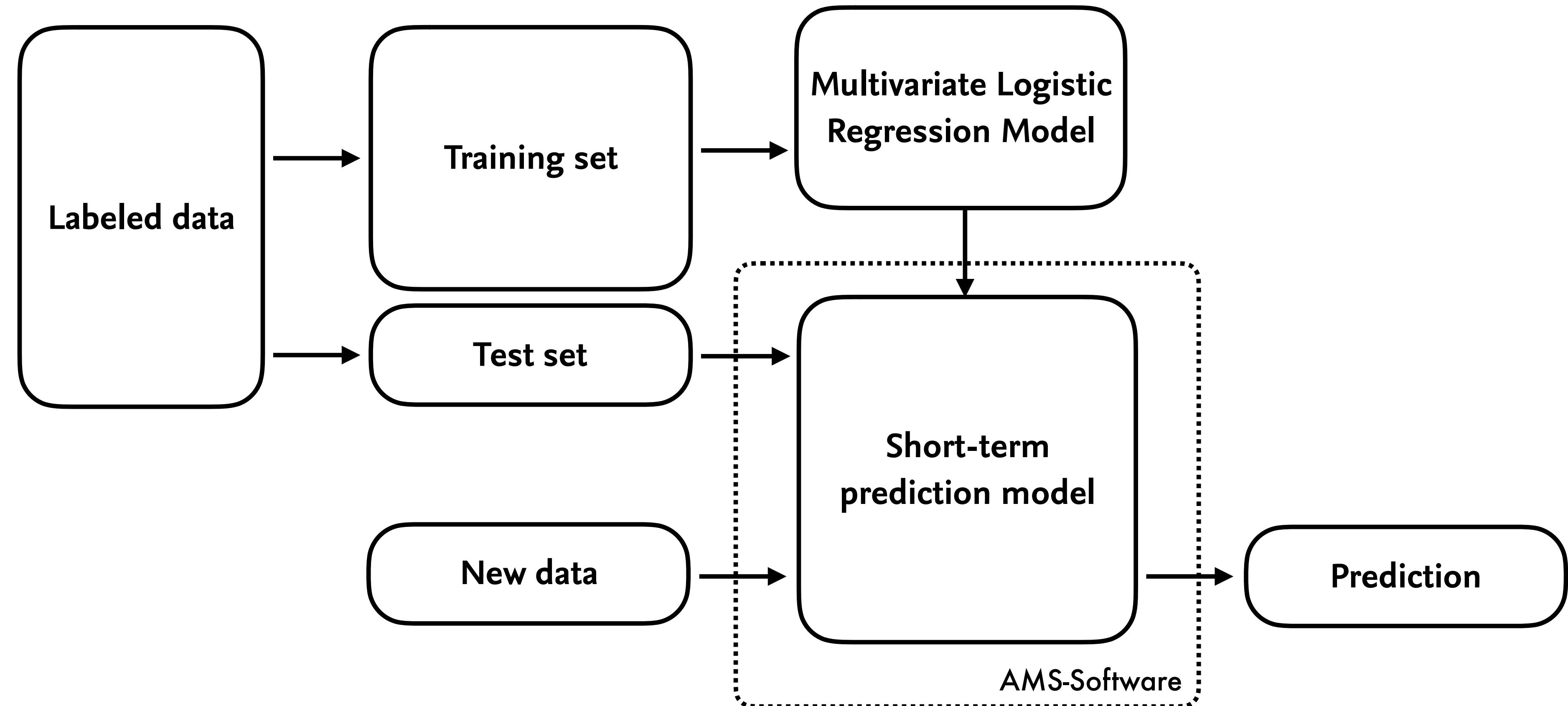
Low chances

Opportunities for integration within the
next 2 years < 25%

Possible Machine Learning (ML)-Pipeline

Repurposed data

- » Self-reported data of job seekers upon registration with the AMS
- » Social security data (e.g., gender)



Insights taken from

Allhutter, D., Cech, F., ooz7, F. F., Grill, G., & Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria - How Austerity Politics Are Made Effective. *Frontiers Big Data*, 3, 326. <http://doi.org/10.3389/fdata.2020.00005>





Used Features

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

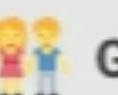
- 0,57 x MN_PARTICIPATION_1

- 0,21 x MN_PARTICIPATION_2

- 0,43 x MN_PARTICIPATION_3)

Kurzfristige Integrationschance

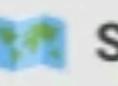
Dieser Rechner berechnet die kurzfristige Integrationschance für Arbeitssuchende wie [hier](#) beschrieben. Dieses Modell ist nur eines von 96 verschiedenen, die die Chancen von Arbeitssuchenden einschätzen sollen und bezieht sich nur auf die ersten sieben Monate nach Beginn der Arbeit des AMS.

 **Geschlecht**

männlich weiblich

 **Alter**

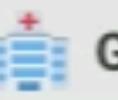
unter 30 30-49 Jahre 50 oder älter

 **Staatsangehörigkeit**

Österreich andere EU-Staaten Drittstaaten

 **Ausbildung**

höchstens Pflichtschule Lehre oder berufsbildende mittlere Schule (BMS)
 Matura oder höhere Ausbildung

 **Gesundheitliche Beeinträchtigung**

Nein Ja

 **Betreuungspflichten**

Nein Ja

Animation taken from <https://derstandard.at/2000089925698/Berechnen-Sie-Ihre-Jobchancen-so-wie-es-das-AMS-tun>





Der AMS-Algorithmus ist ein „Paradebeispiel für Diskriminierung“
The AMS algorithm is a prime example of discrimination



Der AMS-Algorithmus ist ein „Paradebeispiel für Diskriminierung“
Ein Computerprogramm soll ab 2019 die Arbeitsmarktchancen von Arbeitslosen berechnen. Experten von der TU und WU Wien schlagen Alarm.
futurezone.at

Austria's employment agency is rolling out a sorting algorithm that gives lower points to women and the disabled, in the name of efficiency.

A textbook example of automated – and possibly illegal – discrimination.

STORY

Austria's employment agency rolls out discriminatory algorithm, sees no problem

Austria's employment agency rolls out discriminatory algorithm, sees no pro...
AMS, Austria's employment agency, is about to roll out a sorting algorithm that gives lower scores to women and to the disabled. It is very likely illegal under...

What does discrimination mean?



Bias in Data Collection and Processing





Bias in Computer Systems

BATYA FRIEDMAN

Colby College and The Mina Institute
and

HELEN NISSENBAUM

Princeton University

From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

Categories and Subject Descriptors: D.2.0 [Software]: Software Engineering; H.1.2 [Information Systems]: User/Machine Systems; K.4.0 [Computers and Society]: General

General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values



Batya Friedman
University of Washington

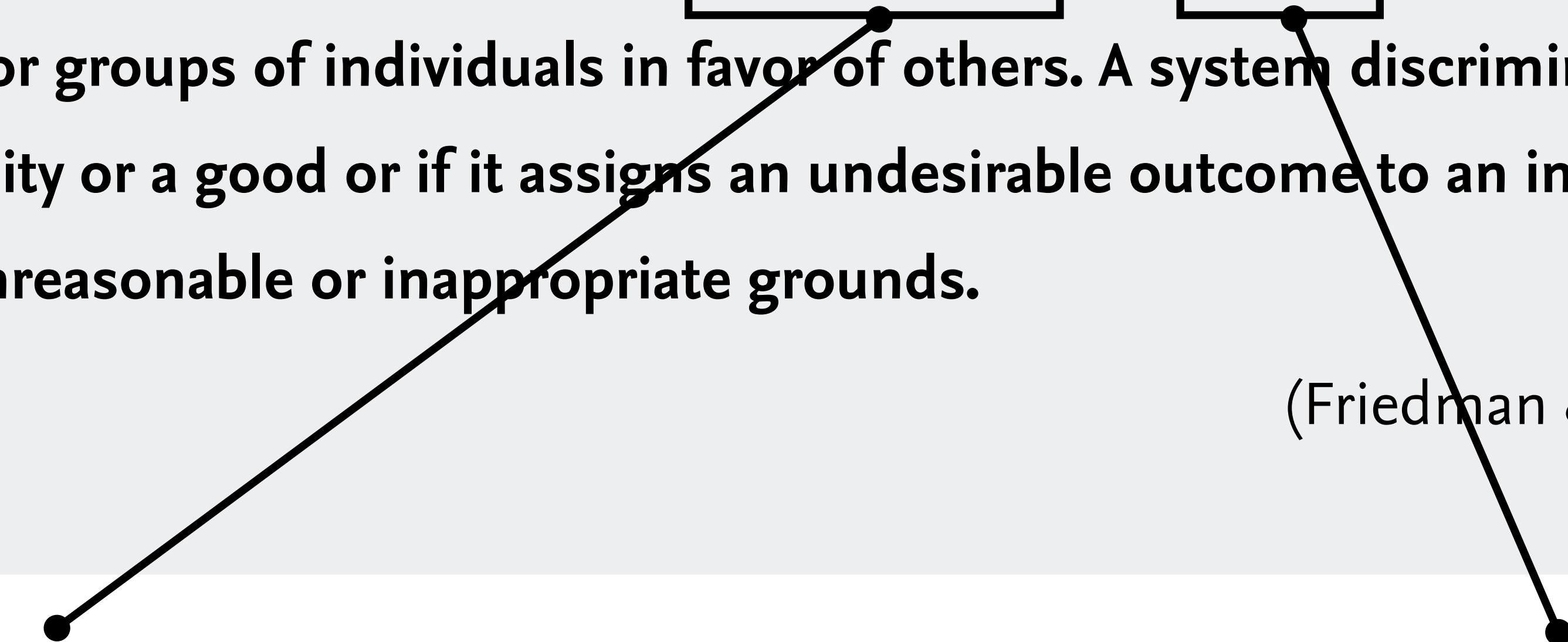


Helen Nissenbaum
Cornell Tech University

“

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)



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.

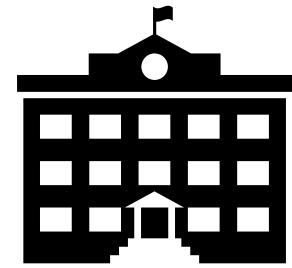


"Construction of distinctions of social groups and categories of persons, which are used for the creation, justification and legitimation of demarcations and hierarchies, especially of power asymmetries, socio-economic inequalities and unequal chances of recognition".

Scherr, Albert 2017. Soziologische Diskriminierungsforschung, in Scherr, Albert, El-Mafaalani, Aladin & Yüksel, Gökçen (Hg.): Handbuch Diskriminierung. Wiesbaden: Springer Fachmedien Wiesbaden, 39–58.



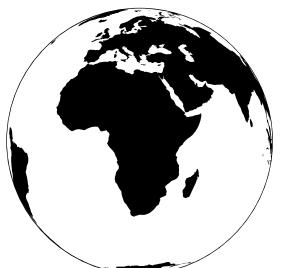
Types of Bias



Preexisting bias has its roots in social institutions, practices, and attitudes.

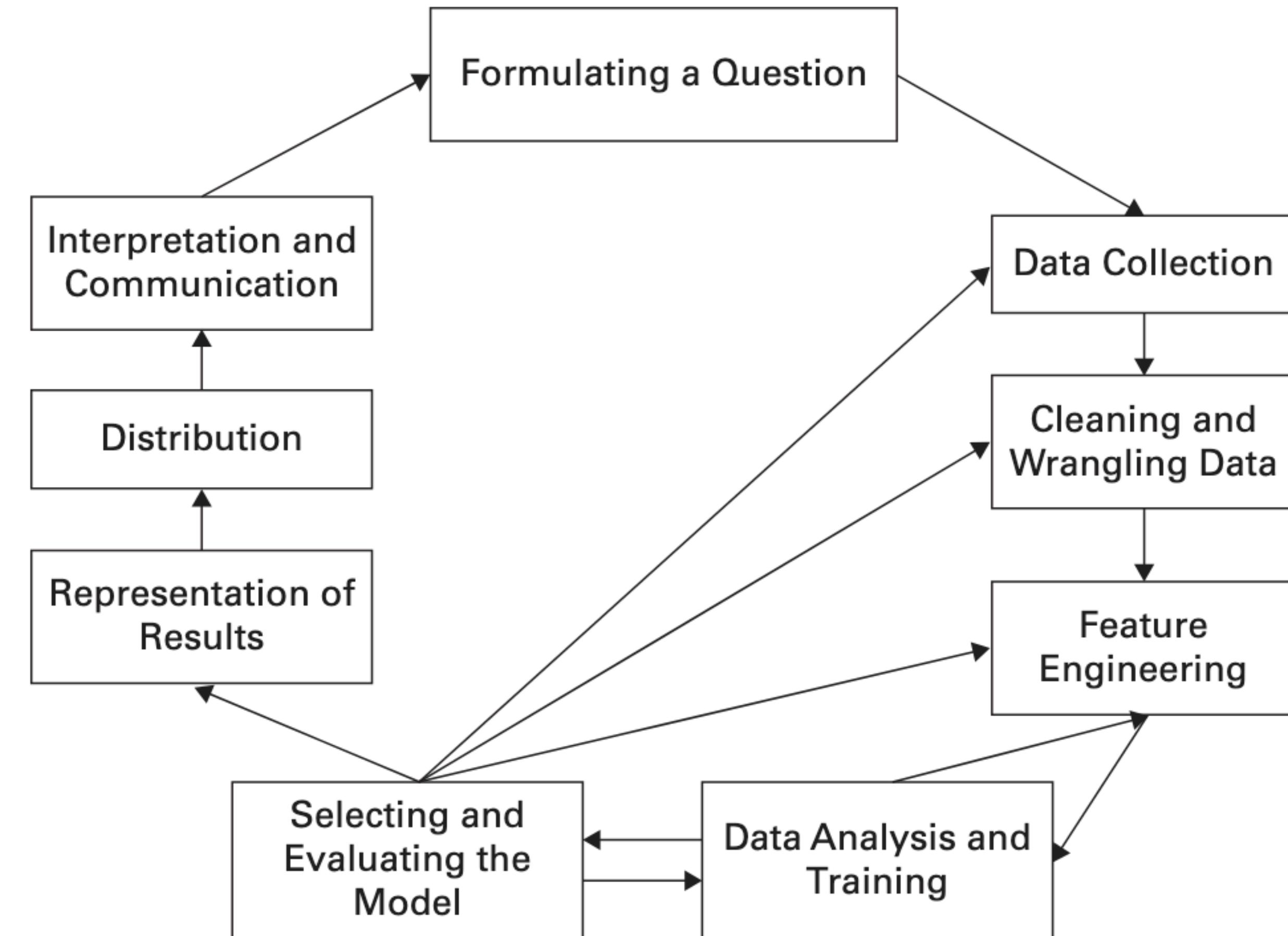


Technical bias arises from technical constraints or technical considerations.

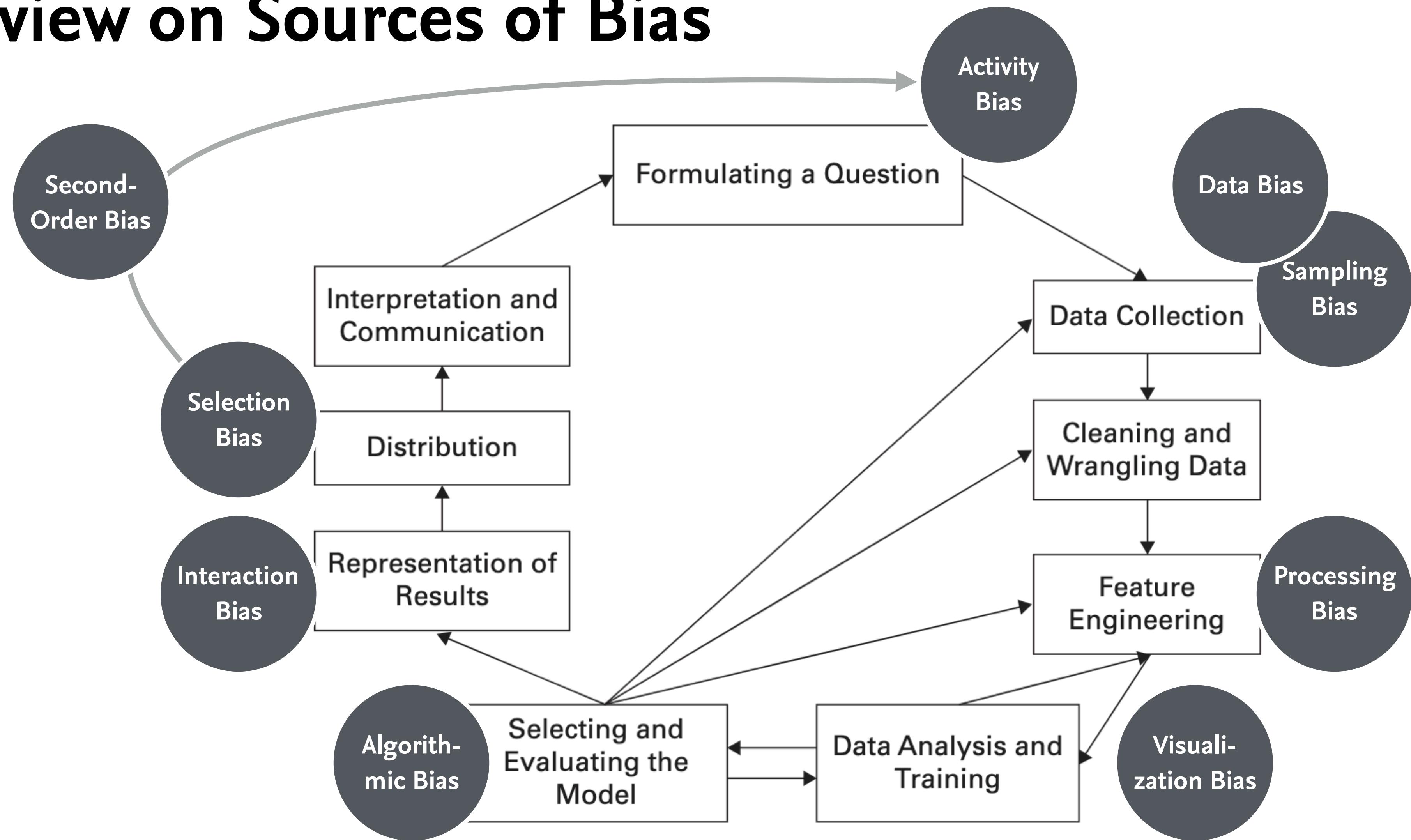


Emergent bias arises in a context of use.

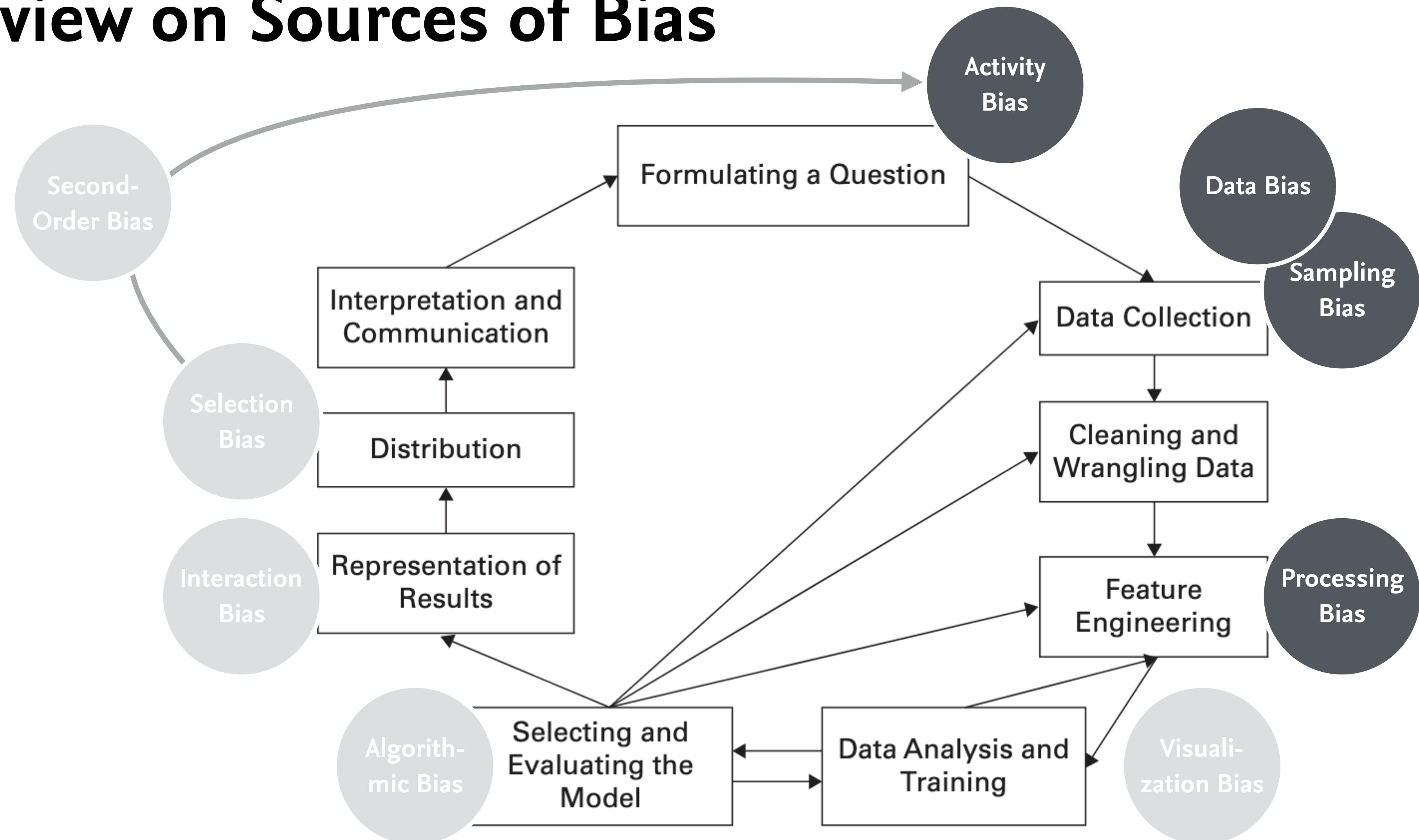
Data Science Pipeline



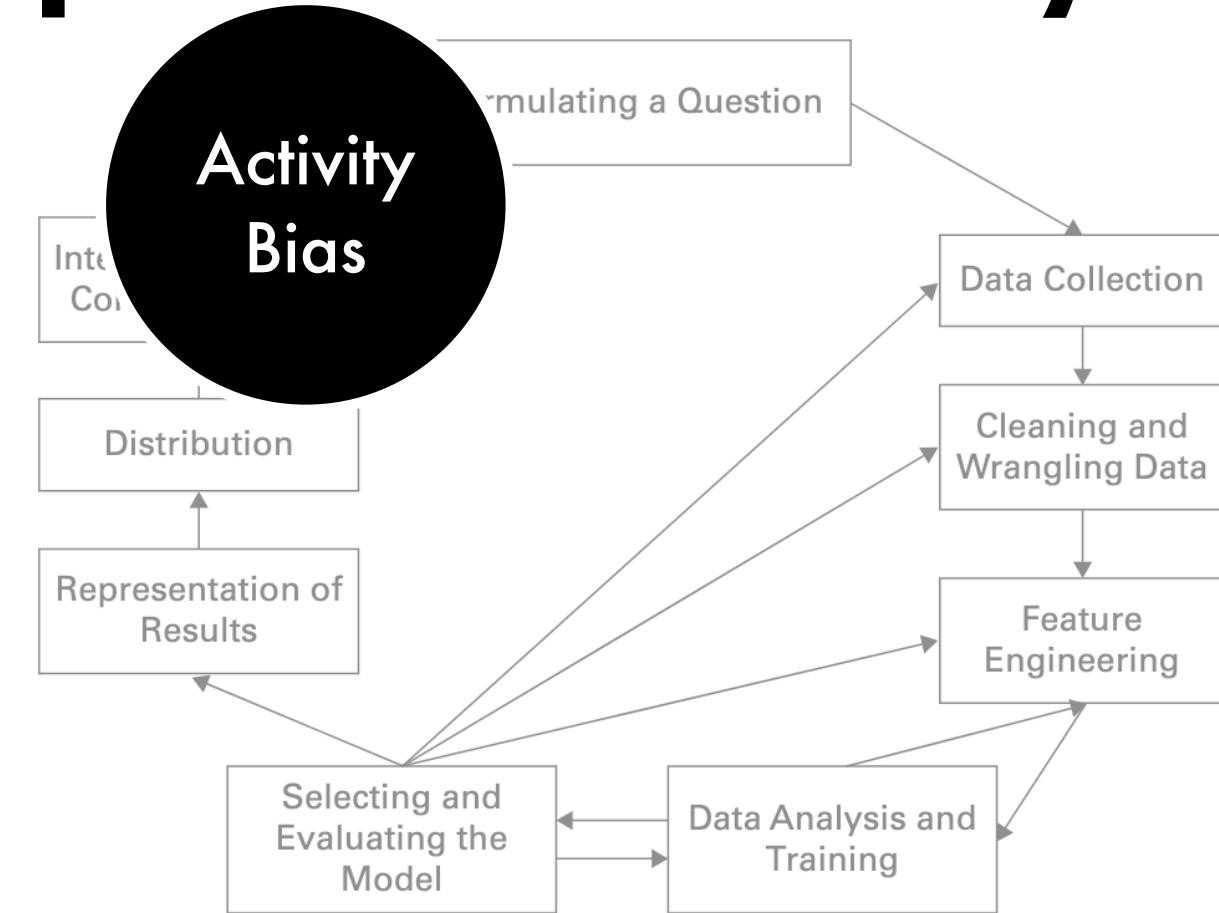
Overview on Sources of Bias



Overview on Sources of Bias



Types of Activity Bias



Population biases

Differences in demographics or other user characteristics between a population of users represented in a dataset or platform and a target population.

Behavioral biases

Differences in user behavior across platforms or contexts, or across users represented in different datasets.

Content production biases

Behavioral biases that are expressed as lexical, syntactic, semantic, and structural differences in the contents generated by users.

Temporal variations

Differences in populations or behaviors over time.

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: Population Bias

	Gesamt			Geschlecht		Alter			
	2018	2019	2020	Frauen	Männer	14-29 J.	30-49 J.	50-69 J.	ab 70 J.
WhatsApp*	65	63	68	73	63	92	79	62	32
Facebook	19	21	14	15	13	24	19	10	1
Instagram	9	13	15	16	14	53	13	1	1
Snapchat	6	5	6	5	7	27	1	-	0
Twitter	1	2	2	1	4	4	3	2	1
Xing	1	1	1	1	1	0	2	0	-
LinkedIn**	-	1	1	0	2	2	2	1	-
Twitch**	-	1	1	0	2	3	1	-	-
TikTok**	-	1	2	2	1	7	1	0	-

* WhatsApp 2018: Wert stammt aus dem Convergence Monitor.

** 2018 nicht erfasst.

Basis: Deutschspr. Bevölkerung ab 14 Jahren (2020: n=3 003; 2019: n=2 000; 2018: n=2 009).

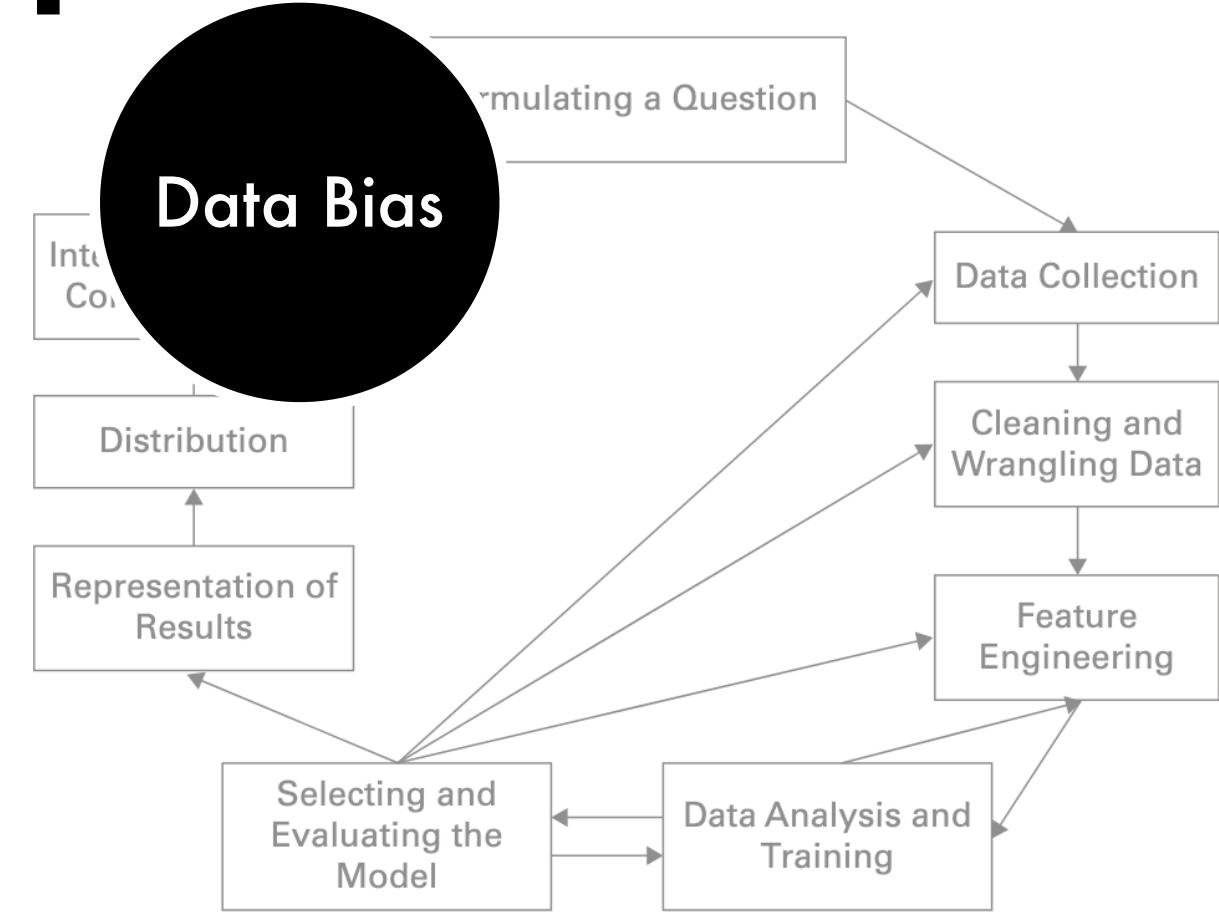
Quelle: ARD/ZDF-Onlinestudien 2018-2020.

Daily Use of Social Media in Germany

Natalie Beisch, Carmen Schäfer: Internetnutzung mit großer Dynamik: Medien, Kommunikation, Social Media. Ergebnisse der ARD/ZDF-Onlinestudie 2020 https://www.ard-zdf-onlinestudie.de/files/2020/0920_Beisch_Schaefer.pdf



Types of Data Bias



Functional biases

Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment.

Normative biases

Biases that are a result of written norms or expectations about unwritten norms describing acceptable patterns of behavior on a given platform.

External biases

Biases resulting from factors outside the social platform, including considerations of socioeconomic status, ideological/religious/political leaning, education, social pressure, privacy concerns, topical interests, language, personality, and culture.

Non-individual agents

Interactions on social platforms that are not produced by individuals, but by accounts representing various types of organizations, or by automated agents.

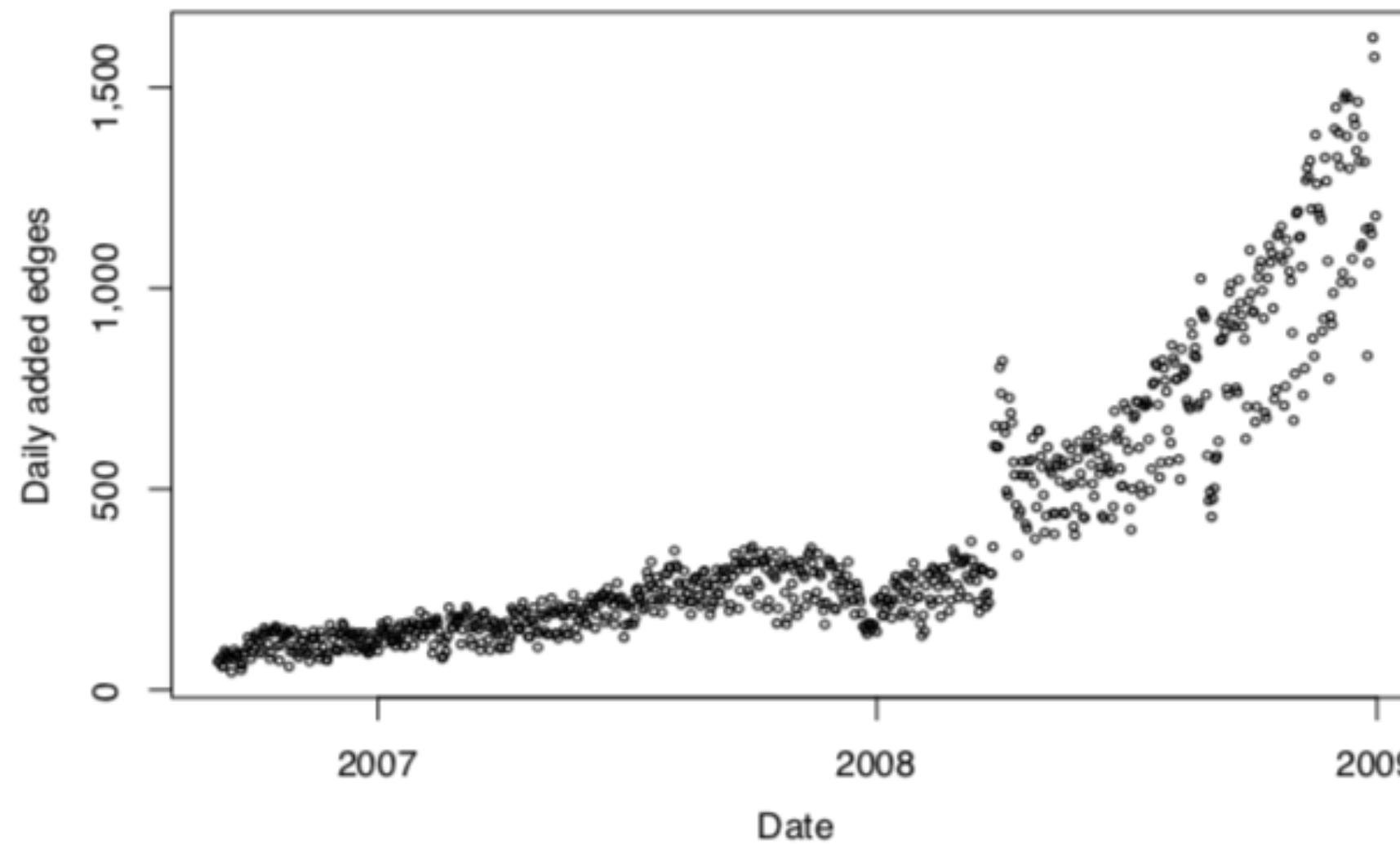
Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2, 13. <http://doi.org/10.3389/fdata.2019.00013>



Example: Functional Biases

Reused dataset from Facebook New Orleans data through a manual crawl starting from a single user and using breadth-first search.

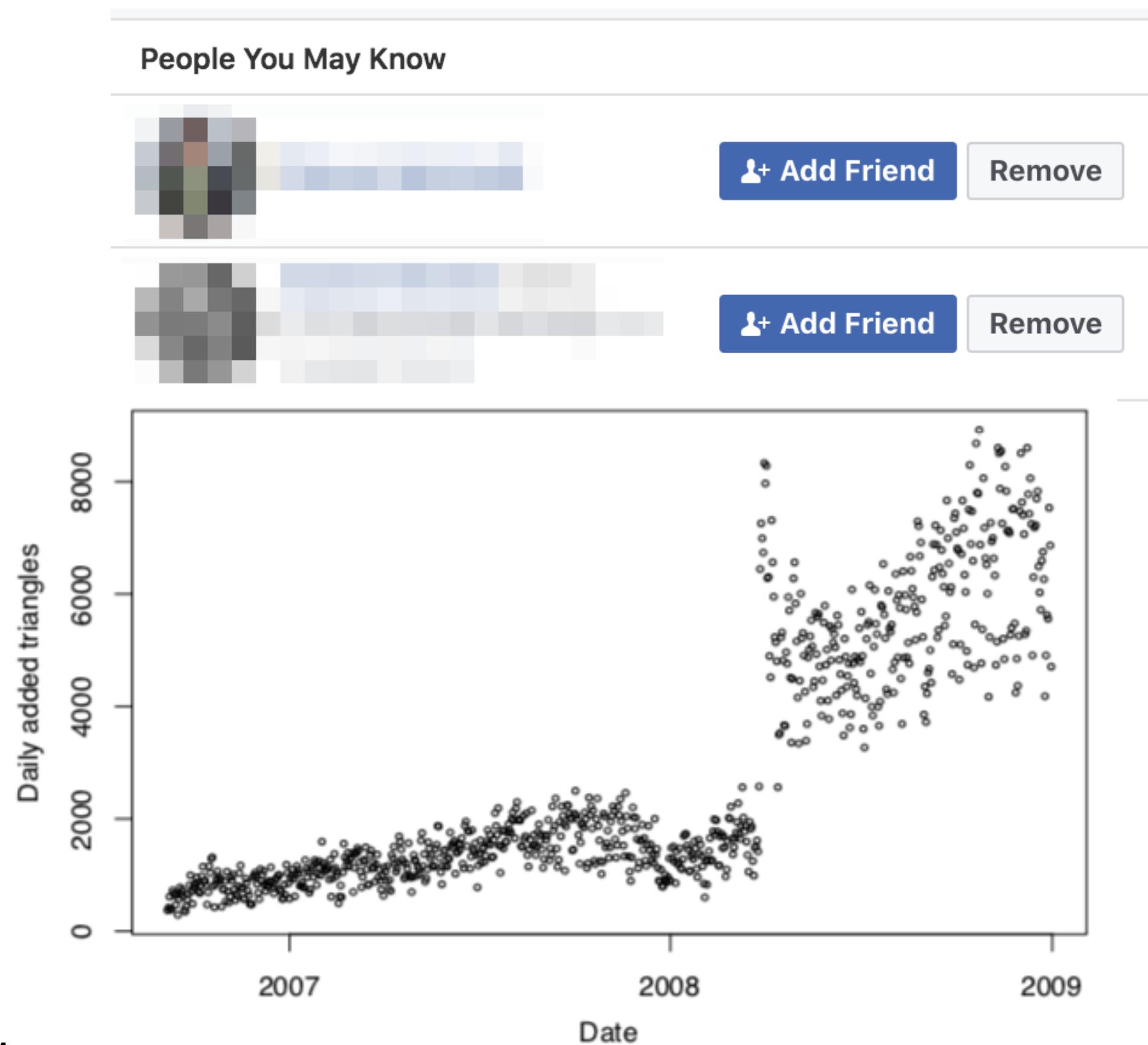
Facebook introduces “People You May Know” (PYMK) feature on March 26, 2008.



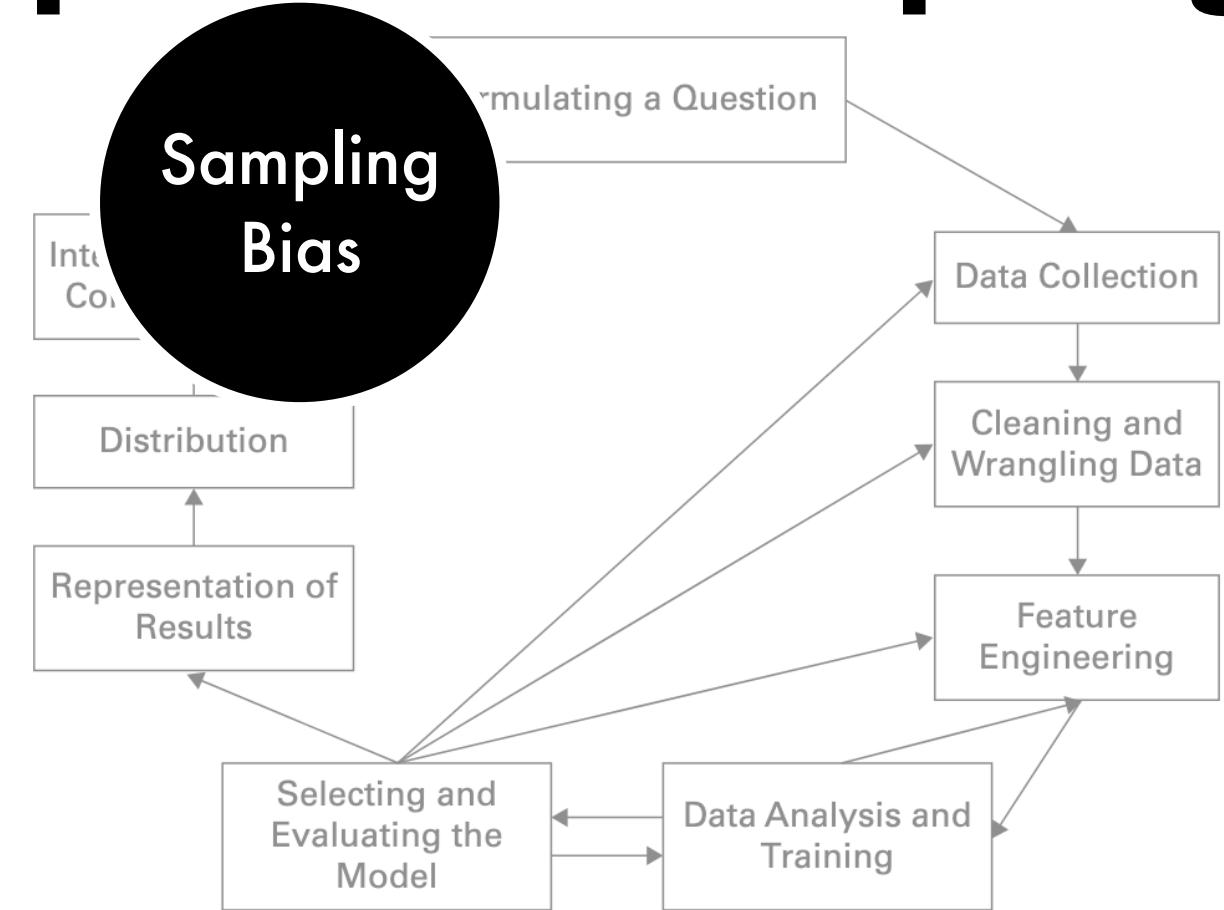
Malik, M. M., & Pfeffer, J. (2016, May). Identifying Platform Effects in Social Media Data. In ICWSM. pp. 241-249.



Course «Human-Centered Data Science» | Summer Term 2022 | Claudia Müller-Birn



Types of Sampling Bias



Data Acquisition

Data collection by third parties for example, by providing limited programmatic access. Their sampling strategies are often opaque.

Data Querying

APIs have limited expressiveness regarding information needs and the choice of keywords in keyword-based queries shapes the resulting datasets.

Data Filtering

Outliers are sometimes relevant for data analysis and text filtering operations may bound certain analyses.

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>

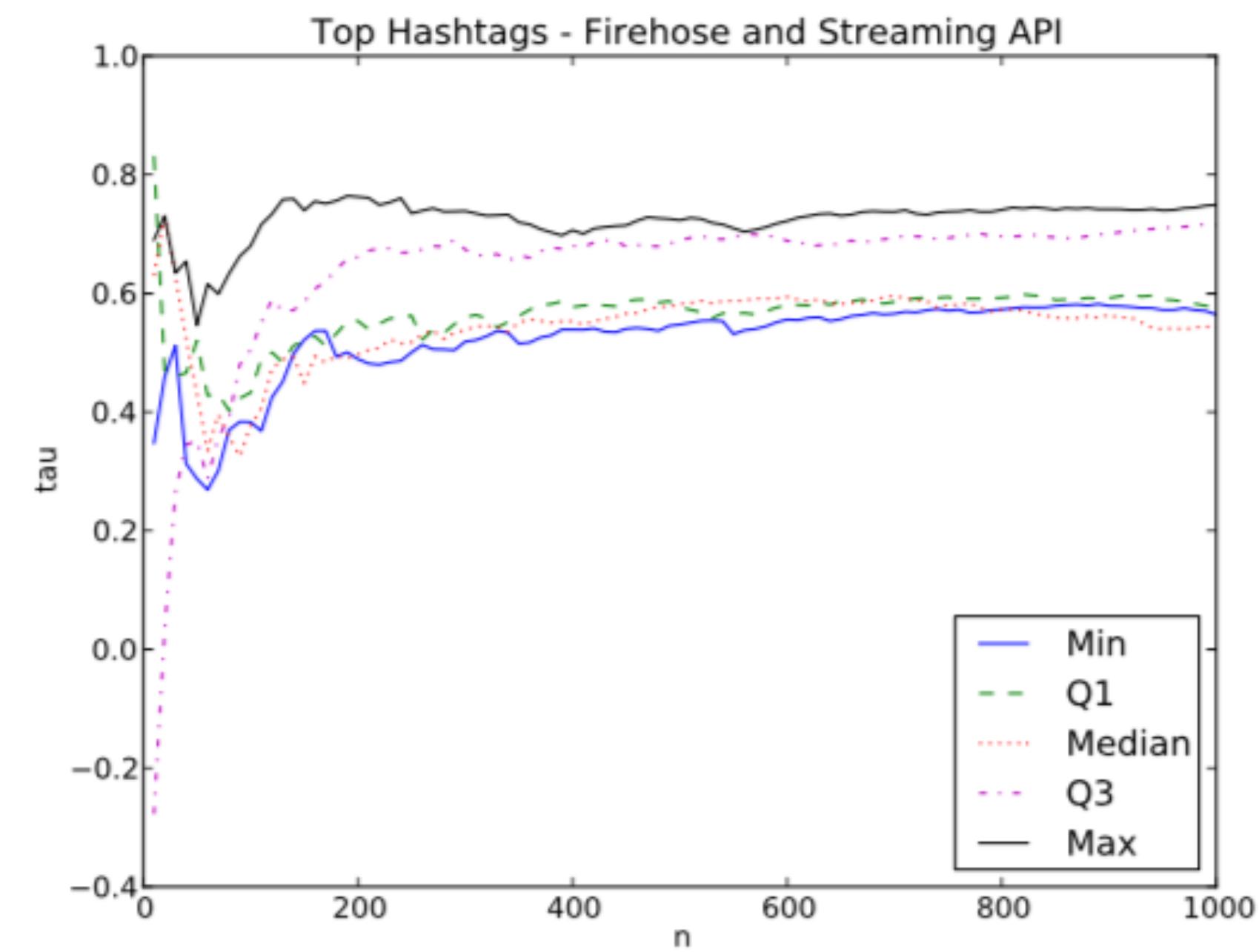
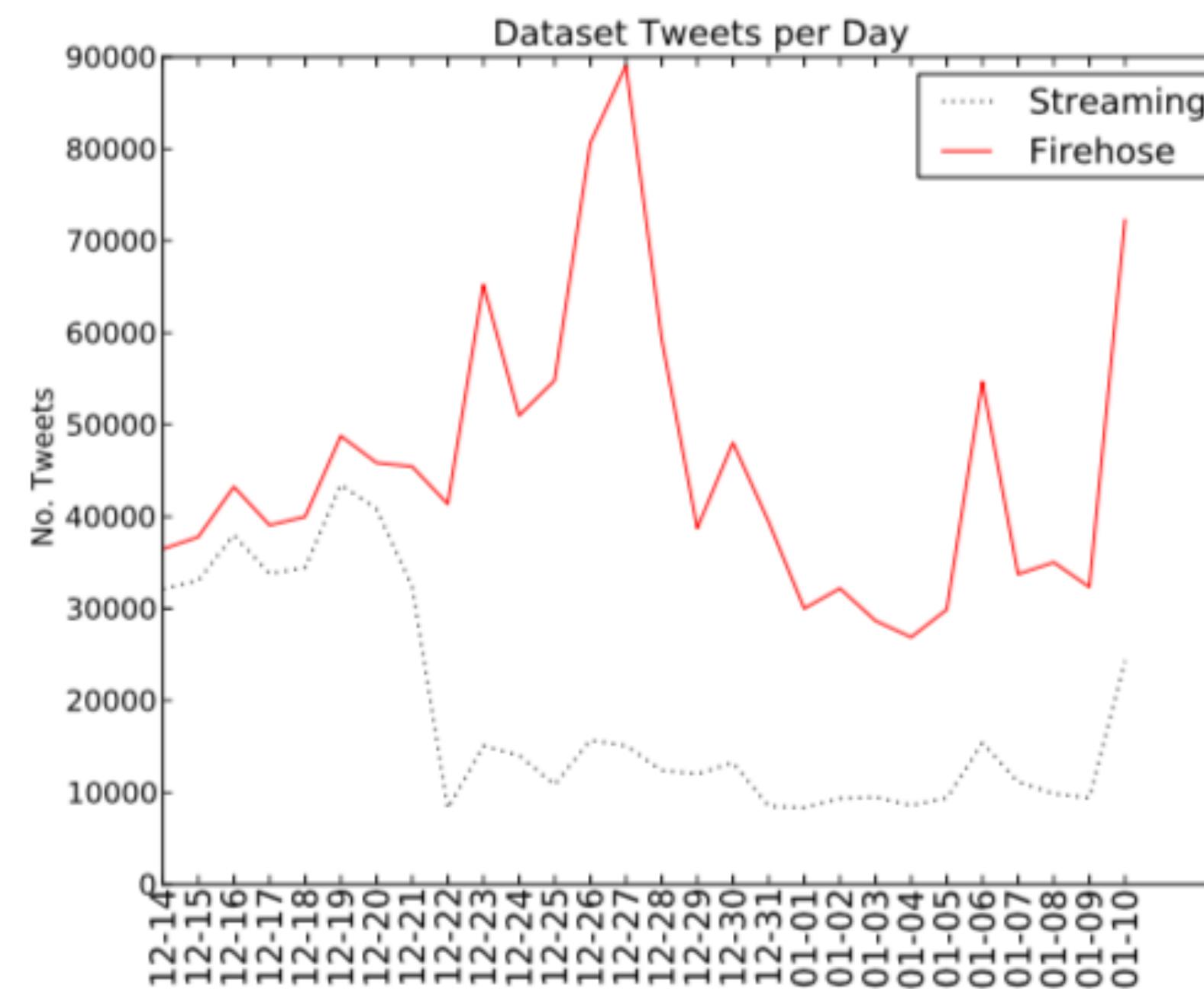




Example Data Acquisition: Is this sample good enough?

Keywords	Geoboxes	Users
#syria, #assad, #aleppovolcano, #alawite, #homs, #hama, #tartous, #idlib, #damascus, #daraa, #aleppo, #سوريا*, #houla	 (32.8, 35.9), (37.3, 42.3)	@SyrianRevo

* Arabic word for “Syria”



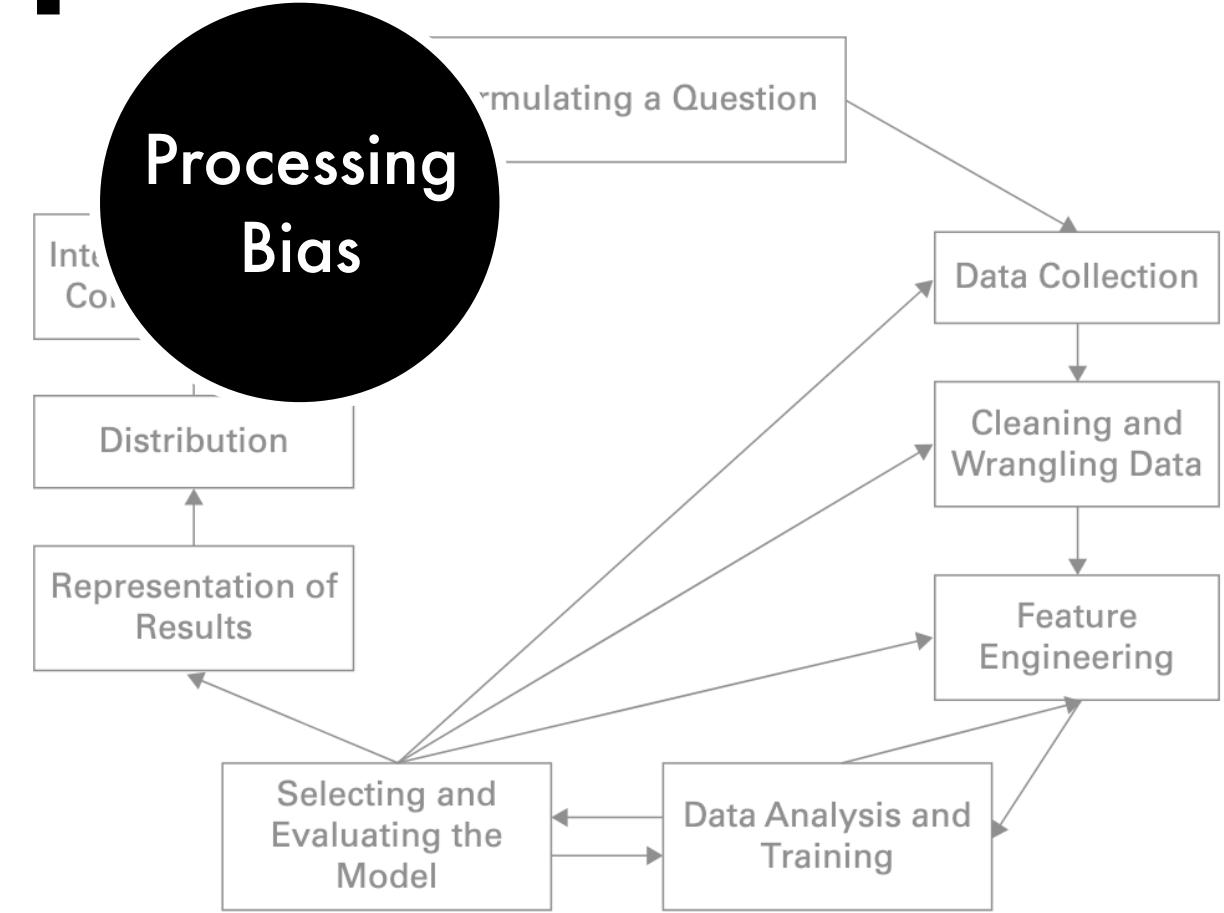
Kendall’s τ is a statistic which measures the correlation of two ordered lists.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? Comparing data from twitter's streaming API with Twitter's firehose. In In ICWSM 2013. pp. 400-408.



Course «Human-Centered Data Science» | Summer Term 2022 | Claudia Müller-Birn

Types of Data Processing Bias



Data Cleaning

Data representation choices and default values may introduce biases, for example the normalization of geographical references.

Data Enrichment

For example, manual annotation often yields subjective and noisy labels, or automatic annotation may introduce errors or bias.

Data Aggregation

For example, consider pre-processing heuristics that aggregate the data to make it more manageable at the cost of losing information.

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>



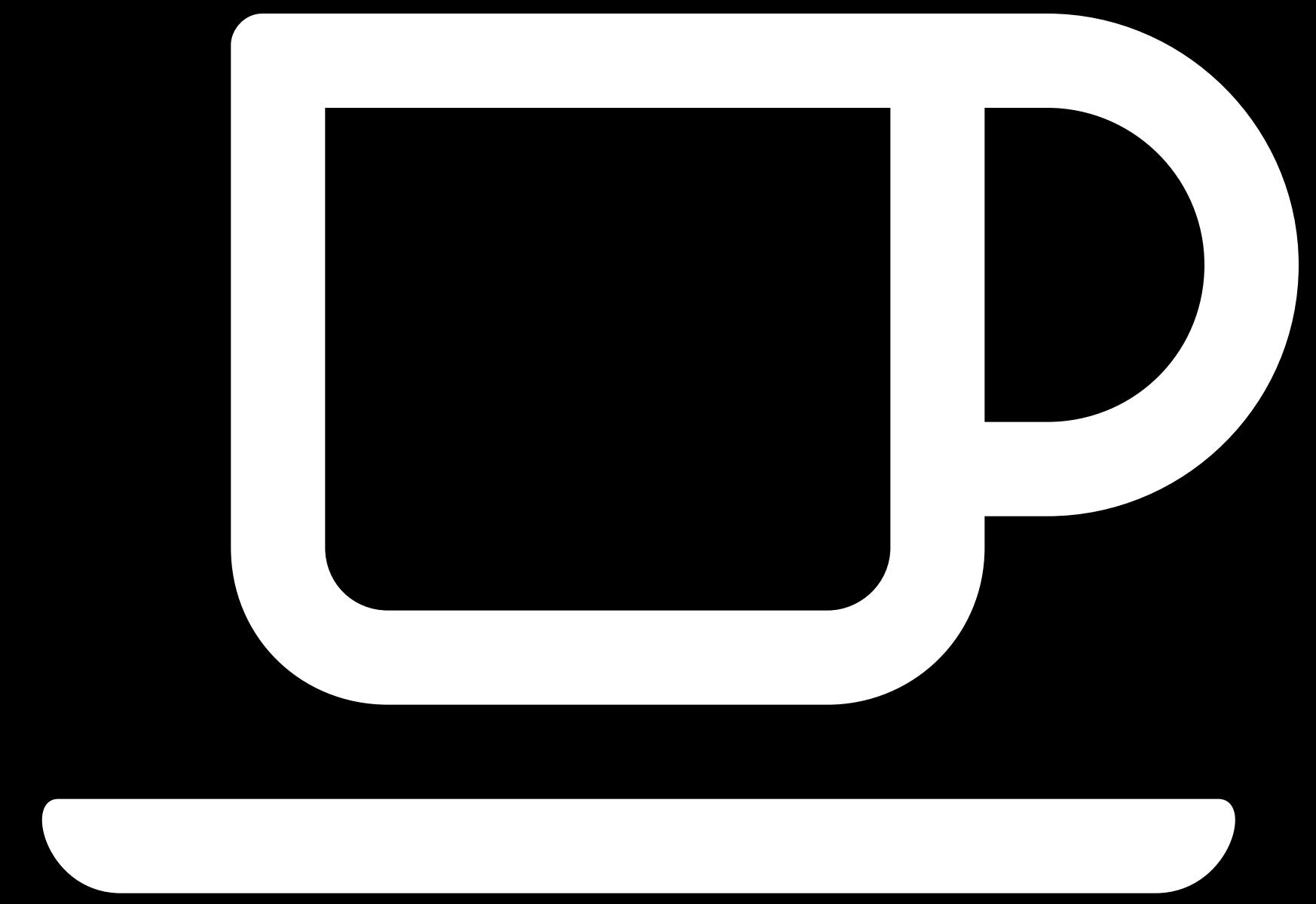
Bias in Word Embeddings

A word embedding is trained on word co-occurrence in text corpora. It represents each word (or common phrase) as a d-dimensional word vector. Words with similar semantic meanings tend to have vectors that are close together.

man:woman as king:queen
man:computer programmer as woman:homemaker
Father:doctor as woman:nurse

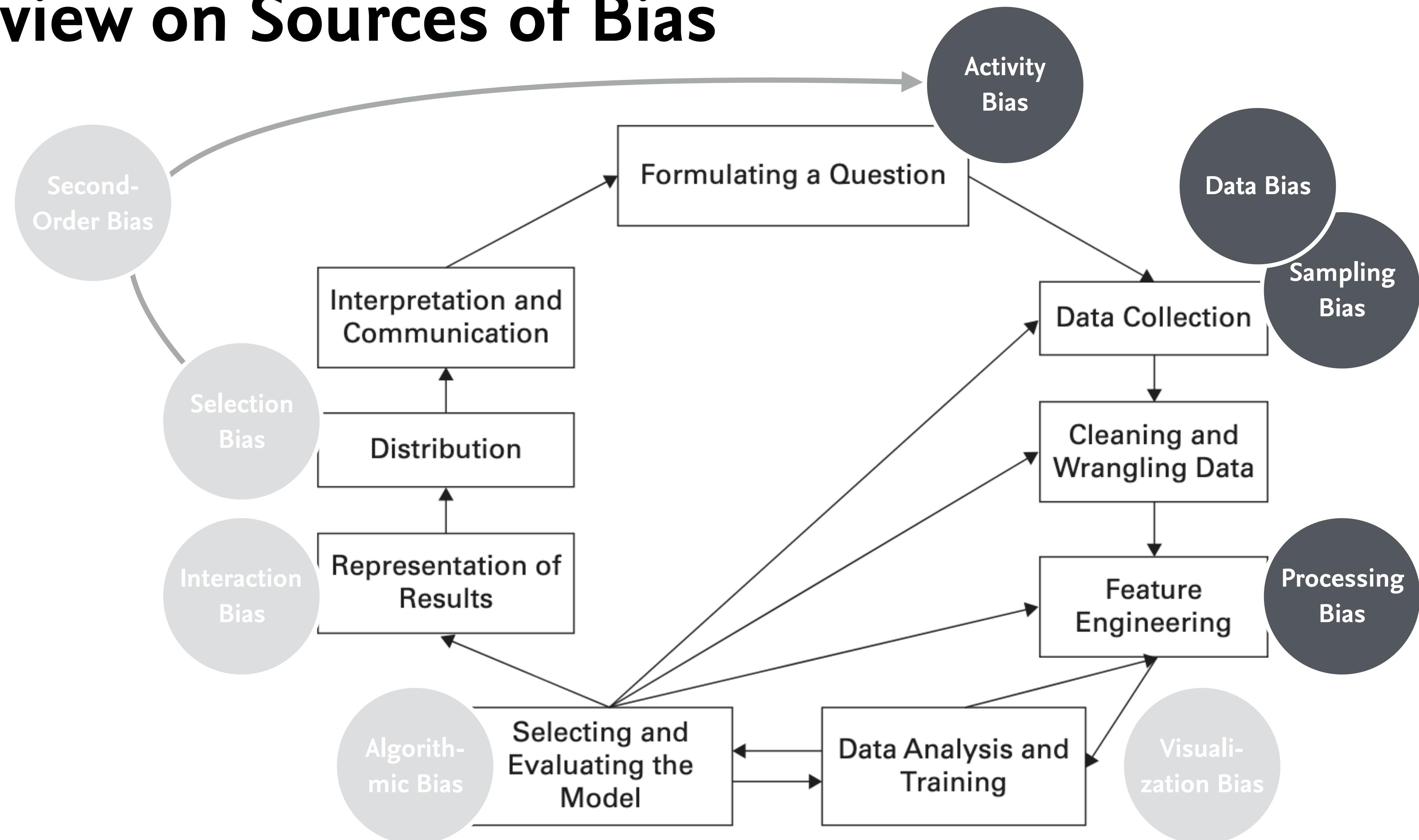
- | Extreme <i>she</i> | Extreme <i>he</i> |
|---------------------------|--------------------------|
| 1. homemaker | 1. maestro |
| 2. nurse | 2. skipper |
| 3. receptionist | 3. protege |
| 4. librarian | 4. philosopher |
| 5. socialite | 5. captain |
| 6. hairdresser | 6. architect |
| 7. nanny | 7. financier |
| 8. bookkeeper | 8. warrior |
| 9. stylist | 9. broadcaster |
| 10. housekeeper | 10. magician |

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 29, 4349–4357.



5 minutes break

Overview on Sources of Bias





How to Mitigate Bias



General Considerations to Mitigate Bias

Reflect on your motivation Which problem needs to be solved by your data-driven software or what questions do you want to answer?

Understand the data origin What are the “circumstances” under which the data were created? How well do you understand the context of data collection? How much could you control the nature of the data set? What are your assumptions? What are platform-specific characteristics of your data? Can you quantify platform population mismatches?

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?

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010.
Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063-1064.





Inspiration: Electronic Components

Sensirion Pressure sensor 1 pc(s) SDP610-025Pa -25 Pa up to 25 Pa (L x W x H) 29 x 18 x 27.05 mm



Item no.: 1313587

Manufacturer no.: 1-100759-02

EAN: 2050002956520



With the sensor out of the SDP600 series, launched one of the first digital dynamic Sensirion differential pressure sensor. The sensor has a digital I²C interface and measures even the smallest pressure differences (10 Pa) with highest sensitivity an...

[Full description ▾](#)



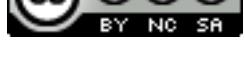
Documents & Downloads

Data sheets (1)

Datasheet 1313587 Sensirion Pressure sensor 1 pc(s) SDP610-025Pa -25 Pa up to 25 Pa (L x W x H) 29 x 18 x 27.05 mm

PDF

Screenshots taken from <https://www.conrad.com/p/sensirion-pressure-sensor-1-pcs-sdp610-025pa-25-pa-up-to-25-pa-l-x-w-x-h-29-x-18-x-2705-mm-1313587>



SDP600 Series (SDP6xx/5xx)
Low-cost Digital Differential Pressure Sensor

Product Summary

The SDP600 sensor family is Sensirion's series of digital differential pressure sensors designed for high-volume applications. They measure the pressure of air and non-aggressive gases with superb accuracy and no offset drift. The sensors cover a pressure range of up to ±500 Pa (±2 inch H₂O / ±5 mbar) and deliver outstanding accuracy even at the bottom end of the measuring range.

The SDP600 series operates from a 3.3 Vdc supply voltage and features a digital 2-wire interface, which makes it easy to connect directly to a microprocessor. The signal is internally linearized and temperature compensated.

Applications

- Medical
- HVAC
- Automotive
- Process automation
- Burner control

New versions

- Low pressure versions SDP600/610-125Pa and SDP600/610-25Pa are suited to measure very low and ultra low differential pressure.
- Low power versions (SDP606/SDP616) are developed especially for low power battery operation.
- Special calibration to measure a massflow in bypass configuration (SDP601/SDP611).

OEM options

A variety of custom options can be implemented for high-volume OEM applications. Ask us for more information.

www.sensirion.com Version 1.7 – September 2012 1/10

1. Sensor Performance

1.1 Physical specifications¹

Parameter	SDP600 SDP610	SDP600-500Pa SDP610-500Pa	SDP600-125Pa SDP610-125Pa	SDP600-25Pa SDP610-25Pa	SDP601 SDP611	SDP616
Short Description	Low cost	Standard	Low DP	Lowest DP	"Mass Flow"	Low Power ²
Calibrated range ³	0 Pa to +400 Pa (±2.0 in. H ₂ O)	-500 Pa to +500 Pa (±2.0 in. H ₂ O)	-125 Pa to +125 Pa (±0.5 in. H ₂ O)	-25 Pa to +25 Pa (±0.1 in. H ₂ O)	-500 to +500 Pa (±2.0 in. H ₂ O)	-500 to +500 Pa (±2.0 in. H ₂ O)
Measurement range	-500 to +500 Pa (±2.0 in. H ₂ O)	-125Pa to +125Pa (±0.5 in. H ₂ O)	-25Pa to +25Pa (±0.1 in. H ₂ O)	-500 to +500 Pa (±2.0 in. H ₂ O)	-500 to +500 Pa (±2.0 in. H ₂ O)	-500 to +500 Pa (±2.0 in. H ₂ O)
Temperature-compensation	yes	yes	yes	yes	mass flow ⁴	yes
Resolution	12 bits preset ⁵ (adjustable from 9 to 16 bit)	0 Pa	0 Pa	0 Pa	0 Pa	16 bit
Zero point accuracy ⁶	0.2 Pa	0.1 Pa	0.1 Pa	0.05 Pa	0.03 Pa	0.2 Pa
Span accuracy ⁷	4.5% of reading	3% of reading	3% of reading	3% of reading	3% of reading	3% of reading
Zero point repeatability ⁸	0.1 Pa	0.05 Pa	0.05 Pa	0.03 Pa	0.01 Pa	0.1 Pa
Span repeatability ⁹	0.5% of reading	0.5% of reading	0.5% of reading	0.5% of reading	0.5% of reading	0.5% of reading
Offset shift due to temperature variation	None	(less than resolution)	(less than resolution)	(less than resolution)	< 0.5% of reading for 10°C	< 0.1 Pa/year
Span shift due to temperature variation	None	(less than resolution)	(less than resolution)	(less than resolution)	< 0.5% of reading for 10°C	< 0.1 Pa/year
Offset stability	None	(less than resolution)	(less than resolution)	(less than resolution)	< 0.1 Pa/year	< 0.1 Pa/year
Response time ¹⁰	4.6 ms typical at 12-bit resolution	70 ms typical	70 ms typical	70 ms typical	70 ms typical	70 ms typical
Warm-up time for first reliable measurement	Typ. 50 ms (first measurement typically after 10 ms)	N/A	N/A	N/A	N/A	N/A

1 Unless otherwise noted, all sensor specifications are valid at 25°C with Vdd = 3.3 Vdc and absolute pressure = 966 mbar.
2 Low Power version are specified with Vdd = 3.0 Vdc and absolute pressure = 966 mbar.
3 The SDP600/SDP610 sensors do measure in the full range from -500 to +500 Pa. But in contrast to the SDP600/SDP610 we do not guarantee the long-term stability of the zero point and span shift over time with other resolutions, e.g. 1.3 mbar.
4 Please see chapter 5 for details.
5 One bit resolution is 1.3 mbar.
6 With 12-bit resolution, includes repeatability of hysteresis.
7 Total accuracy/repeatability is a sum of zero-point and span accuracy/repeatability.

www.sensirion.com Version 1.7 – September 2012 2/10

1.2 Ambient conditions

Parameter	SDP6xx / SDP6xx Series
Calibrated for ¹¹	0 Pa
Media compatibility	Air, N ₂ , O ₂
Calibrated temperature	-20 °C to +40 °C
Operating temperature ¹²	-20 °C to +40 °C
Storage temperature ¹³	-40 °C to +80 °C
Position sensitivity	Less than repeatability error

1.3 Materials

Parameter	SDP6xx / SDP6xx Series
Welded materials	PBT (polybutylene terephthalate), glass (silicon nitride, silicon oxide), silicon, gold, FR4 (ceramic as static sealing, epoxy, copper, tin, silver, gold, aluminum)
REACH, RoHS, WEEE	The SDP6xx/5xx series is REACH, RoHS and WEEE compliant

2. Electrical Specifications

2.1 Electrical characteristics

Parameter	SDP600 / SDP610 ¹⁴	SDP600 / SDP610 ¹⁵
Operating voltage	3.0 – 3.3 Vdc (A supply voltage of 3.3 V is recommended)	3.0 – 3.3 Vdc (A supply voltage of 3.3 V is recommended)
Current drain	< 400 µA	< 6 mA typical in operation
Interface	Digital 2-wire interface (I ² C)	Digital 2-wire interface (I ² C)
Default clock frequency	100 kHz typical, 400 kHz max.	100 kHz typical, 400 kHz max.
Default I ² C address	0x40 (binary: 1000 000)	0x40 (binary: 1000 000)
Scale factor ¹⁶	60 Pa ¹⁷	200 Pa ¹⁷
SDP600-500Pa	125Pa ¹⁷	125Pa ¹⁷
SDP600-125Pa	24100 mbar ¹⁷	24100 mbar ¹⁷
SDP600-25Pa	41388 mbar ¹⁷	41388 mbar ¹⁷
SDP601	149360 mbar ¹⁷	149360 mbar ¹⁷
SDP616	241000 mbar ¹⁷	241000 mbar ¹⁷

11 Contact Sensirion for information about other gases, wider calibrated temperature ranges and higher storage temperatures with 118 000 000. For older products, calibrated temperature range is 0°C to +40°C.
12 See section 5.1. The scale factor may vary with other configurations.
13 See section 5.1. The scale factor may vary with other configurations.
14 The SDP600/SDP610 sensors do measure in the full range from -500 to +500 Pa. But in contrast to the SDP600/SDP610 we do not guarantee the long-term stability of the zero point and span shift over time with other resolutions, e.g. 1.3 mbar.
15 Total accuracy/repeatability is a sum of zero-point and span accuracy/repeatability.
16 Contact Sensirion for information about other gases, wider calibrated temperature ranges and higher storage temperatures with 118 000 000. For older products, calibrated temperature range is 0°C to +40°C.
17 See section 5.1. The scale factor may vary with other configurations.
18 See section 5.1. The scale factor may vary with other configurations.
19 The sensor output may be divided by alternative scale factors to receive the physical value in another unit.

www.sensirion.com Version 1.7 – September 2012 3/10



BY TIMNIT GEBRU, JAMIE MORGENSTERN,
BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN,
HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD

Datasheets for Datasets



Objective of «Datasheets for Datasets»

For Dataset Creators Encouraging careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use.

For Dataset Consumers Ensuring that they (policy makers, consumer advocates, investigative journalists, individuals*) have the information they need to make informed decisions about using a dataset for their chosen tasks and avoid unintentional misuse.

() individuals whose data is included in datasets, and individuals who may be impacted by models*

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.



Sections of «Datasheets for Datasets»

motivation

composition

**collection
process**

**preprocessing/
cleaning/labeling**

uses

distribution

maintenance

Sections of «Datasheets for Datasets»

motivation	Describe the motivations for creating the dataset, including funding, any specific tasks the authors had in mind, and who the authors are.
composition	Describe the composition of the dataset, like what kinds of data are in it, how it was collected, whether labels are associated with the data, and whether the dataset contains sensitive information.
collection process	Describe the data collection process, like how the data was collected, where or who is was collected from, who was involved in the collection process, and, if people are involved, if consent was given for the data to be collected.
preprocessing/ cleaning/labeling	Whether the data was process or labelled and how it was done.
uses	The tasks the dataset is intended to be used for, how it has already been used, and limitations of use. Distribution: How the dataset will be distributed and to who, and any restrictions on distribution.
maintenance	Who and how the dataset will be maintained, and if and how others will be able to build on it.
distribution	Whether the dataset is distributed to third parties outside of the owner with what license by employing any restrictions.



Example

Movie Review Polarity	Thumbs Up? Sentiment Classification using Machine Learning Techniques
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.	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 ?
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	Figure 1. An example “negative polarity” instance, taken from the file neg/cv452_tok-18656.txt.
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.	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 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	What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.
Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?	Each instance consists of the text associated with the review, with obvious ratings information removed from that text (some errors were found and later fixed). The text was down-cased and HTML tags were removed. Boilerplate newsgroup header/footer text was removed. Some additional unspecified automatic filtering was done. Each instance also has an associated target value: a positive (+1) or negative (-1) sentiment polarity rating based on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in “Data Preprocessing”).
The dataset was created by Bo Pang and Lillian Lee at Cornell University.	Is there a label or target associated with each instance? If so, please provide a description.
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.	The label is the positive/negative sentiment polarity rating derived from the star rating, as described above.
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.	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.
Any other comments?	Everything is included. No data is missing.
None.	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
movie-review-data/r1-polaritydata README.r1.0.txt; http://www.cs.cornell.edu/people/pabo/movie-review-data/poldata README.2.0.txt	





Other Approaches

To appear in Transactions of the ACL

**Data Statements for Natural Language Processing:
Toward Mitigating System Bias and Enabling Better Science**

Emily M. Bender
Department of Linguistics
University of Washington
ebender@uw.edu

Batya Friedman
The Information School
University of Washington
batya@uw.edu

Abstract

In this paper, we propose *data statements* as a design solution and professional practice for natural language processing technologists, in both research and development—through the adoption and widespread use of data statements, the field can begin to address critical scientific and ethical issues that result from the use of data from certain populations in the development of technology for other populations. We present a form that data statements can take and explore the implications of adopting them as part of regular practice. We argue that data statements will help alleviate issues related to exclusion and bias in language technology; lead to better precision in claims about how NLP research can generalize and thus better engineering results; protect companies from public embarrassment; and ultimately lead to language technology that meets its users in their own preferred linguistic style and furthermore does not misrepresent them to others.

gramming jobs because of biases present in training text). There are both scientific and ethical reasons to be concerned. Scientifically, there is the issue of generalizability of results; ethically, the potential for significant real-world harms. While there is increasing interest in ethics in NLP,¹ there remains the open and urgent question of how we integrate ethical considerations into the everyday practice of our field. This question has no simple answer, but rather will require a constellation of multi-faceted solutions.

Toward that end, and drawing on value sensitive design (Friedman et al., 2006), this paper contributes one new professional practice—called *data statements*—which we argue will bring about improvements in engineering and scientific outcomes while also enabling more ethically responsive NLP technology. A data statement is a characterization of a dataset which provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software. In developing this

Dataset Nutrition Label
TaxBills NYC Dataset (joined.csv)

About

This dataset was created to make information about NYC's rent stabilized apartments more accessible to the public. Currently, information on rent stabilization in NYC is only published in aggregate by borough, leaving little information about specific buildings available. By parsing tax bill data about buildings contained in Notice of Property Value (NoPV) documents, which are available publicly in pdf format, this dataset is able to get a building-by-building count of rent stabilized units in NYC. See more about why this dataset was created here.

Data Creation Range: January, 2015 - Present
Created By: John Krauss
Content: Tabular (csv, JSON)
Source: <https://github.com/talos/nyc-stabilization-unit-counts>

Alert Count	
Completeness	0
Provenance	1
Misrepresentation	1
Collection	
Socioeconomic Bias	1
Inaccurate Prediction	1
Description	
Composition	1
Racial bias	1

* Please refer to the Objectives and Alerts section for more details

Use Cases
Potential real-world applications of the dataset

- How many rent stabilized units are in a particular building?
- Has a building lost rent stabilized units?
- Is there a pattern of deregulation in a building?
- Is there a pattern of deregulation for a specific landlord?
- Where might there be abuse of tax abatements? Is the landlord breaking the rules of the abatements?
- Where is gentrification happening?

Badges

Alert Count by Category

Category	Count
Collection	1
Composition	1
Provenance	1

Alert Count by Mitigation Potential

Potential	Count
Yes	1
Maybe	2

Alert Count by Potential Harm

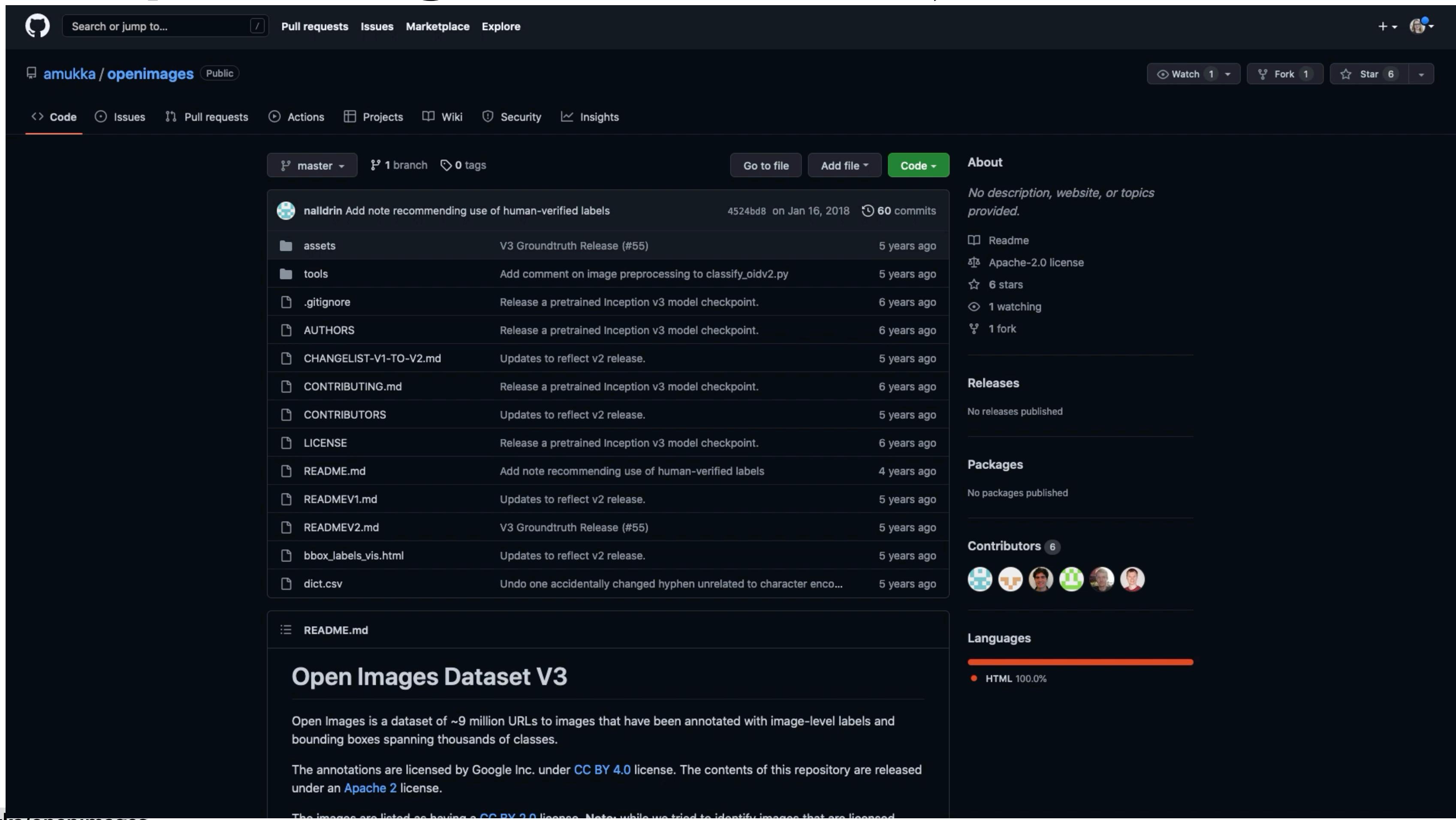
Harm	Count
Inaccurate Prediction	1
Misrepresentation	1
Racial bias	1
Socioeconomic Bias	1

Bender, Emily M, und Batya Friedman. „Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science“ 2018.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. “The Dataset Nutrition Label - a Framework to Drive Higher Data Quality Standards.” CoRR cs.DB (January).



Example: Open Images Dataset V3 (Data Fact Sheet)



The screenshot shows the GitHub repository page for `amukka/openimages`. The repository has 1 branch and 0 tags. The commit history shows 60 commits from `nalldrin`, with the most recent being a note recommending human-verified labels. The repository is licensed under Apache-2.0 and has 6 stars, 1 fork, and 1 watcher. The `About` section notes that there is no description, website, or topics provided. The `Releases` section indicates no releases have been published. The `Packages` section shows no packages published. The `Contributors` section lists 6 contributors with their profile icons. The `Languages` section shows that the code is written in HTML at 100%. The main content area displays the `README.md` file, which describes the Open Images Dataset V3 as a dataset of ~9 million URLs with annotated image-level labels and bounding boxes across thousands of classes. It mentions that annotations are licensed under CC BY 4.0 and the contents under Apache 2.

<https://github.com/amukka/openimages>

Krasin, I. et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification, 2017.





Documentation As Reflexive Practice





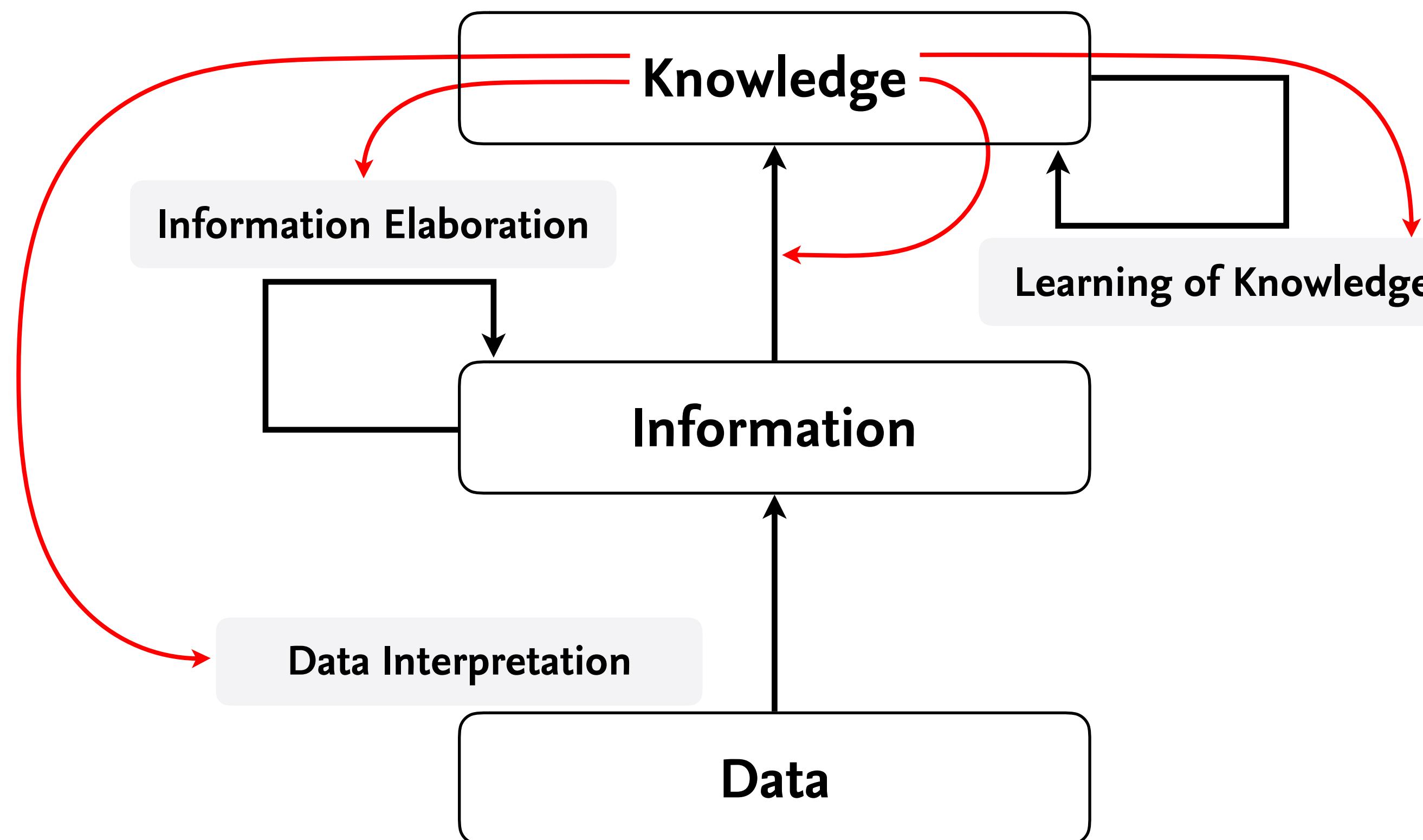
When Data Become Data

Rarely can a magic moment
be established when things
become data

Borgman, C. L. (2016). Big data, little data, no data: Scholarship in
the networked world. MIT press. (p. 62)



What is Data?



Knowledge Interpreted symbol structures

- » used to interpret data, elaborate on information, and learn
 - » used within decision steps

Information Interpretation

- » input to a decision step
 - » output from a decision step

Data Observed, uninterpreted symbols

- » signs, character sequences, patterns

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.



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.



Check your Insights

- » What is bias and what are the prerequisites that bias emerges in data science workflows?
- » Is it possible to completely avoid bias in data science?
- » What are the types of bias and where does bias can occur in the data science workflow?
- » What approaches do you know to mitigate bias? What is the basic idea that unites all these approaches?
- » Why is it necessary to consider the context of data when working with data?
- » What is meant by documentation as reflexive practice? What is the difference to the other approaches?





«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