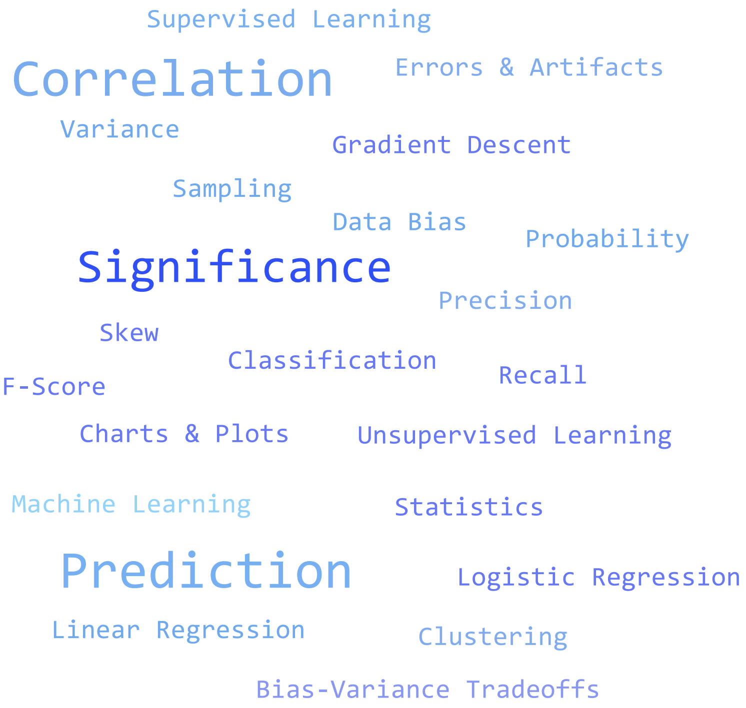


Data Science 1:
Introduction to Data Science

Data Munging, Data Cleaning & Data Bias

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Lecture slides based on content from “The Data Science Design Manual” (Steven Skiena, 2017) and associated course materials generously made available online by the author at https://www3.cs.stonybrook.edu/~skiena/data-manual/.

Special thanks to Professor Skiena for sharing these valuable teaching resources!

Supervised Learning
Correlation
Errors & Artifacts
Variance
Gradient Descent
Sampling
Data Bias
Probability
Significance
Precision
Skew
Classification
Recall
F-Score
Charts & Plots
Unsupervised Learning
Machine Learning
Statistics
Prediction
Logistic Regression
Linear Regression
Clustering
Bias-Variance Tradeoffs

Data Science 1:
Introduction to Data Science

Data Munging, Data Cleaning & Data Bias

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Semester Schedule

CW 42	14. Oct	Lecture	1	Orga & Intro	1-26
CW 43	21. / 23. Oct	Lecture + Exercises	2	Probability, Statistics & Correlation	27-56
CW 44	28. Oct	Lecture	3	Data Munging, Cleaning & Bias	57-94 / “Invisible Women”
CW 45	04. / 06. Nov	Lecture + Exercises	4	Scores & Rankings	95-120
CW 46	11. Nov	Lecture	5	Statistical Distributions & Significance	121-154
CW 47	18. / 20. Nov	Lecture + Exercises	6	Building & Evaluating Models	201-236
CW 48	25. Nov	Guest Lecture	7	Data Visualization	155-200
CW 49	02. / 04. Dec	Lecture + Exercises	8	Intro to Machine Learning	351-390
CW 50	09. Dec	Lecture	9	Linear Algebra	237-266
CW 51	16. / 18. Dec	Lecture + Exercises	10	Linear Regression & Gradient Descent	267-288
CW 02	06. Jan	Lecture	11	Logistic Regression & Classification	289-302
CW 03	13. / 15. Jan	Lecture + Exercises	12	Nearest Neighbor Methods & Clustering	303-350
CW 04	20. Jan	Lecture	13	Data Science in the Wild	391-426
CW 05	27. / 29. Jan	Lecture + Exercises	14	Q&A / Feedback	
CW 06	03. / 04. Feb	Oral Exams (Block 1)		Preparation in our last session („Oral Exam Briefing“)	
CW 13	24. / 25. Mar	Oral Exams (Block 2)			

Data Munging

Good data scientists spend most of their time cleaning and formatting data.

The rest spend most of their time complaining there is no data available.

Data munging or *data wrangling* is the art of acquiring data and preparing it for analysis.

Languages for Data Science

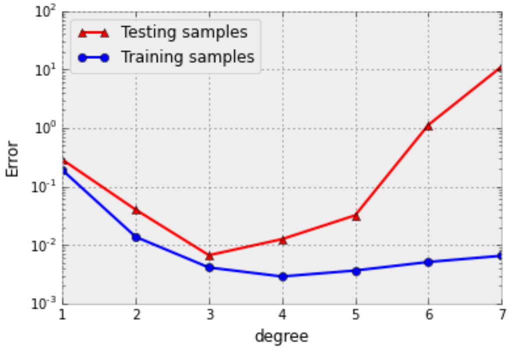
- *Python*: contains libraries and features (e.g regular expressions) for easier munging.
- *R*: programming language of statisticians.
- *Matlab*: fast and efficient matrix operations.
- *Java/C*: language for Big Data systems.
- *Mathematica/Wolfram Alpha*: symbolic math.
- *Excel*: bread and butter tool for exploration.

Notebook Environments

Mixing code, data, computational results, and text are essential for projects to be:

- reproducible
- tweakable
- documented.

In [40]:
```
degrees = range(1, 8)
errors = np.array([regressor3(d) for d in degrees])
plt.plot(degrees, errors[:, 0], marker='^', c='r', label='Testing samples
plt.plot(degrees, errors[:, 1], marker='o', c='b', label='Training sample
plt.yscale('log')
plt.xlabel("degree"); plt.ylabel("Error")
_ = plt.legend(loc='best')
```



By sweeping the degree we discover two regions of model performance:

- **Underfitting** (degree < 3): Characterized by the fact that the testing error will get lower if we increase the model capacity.
- **Overfitting** (degree > 3): Characterized by the fact the testing will get higher if we increase the model capacity. Note, that the training error is getting lower or just staying the same!.

Data Pipelines

Notebooks make it easier to maintain data pipelines, the sequence of processing steps from start to finish.

Expect to have to redo your analysis from scratch, so build your code to make it possible.

Standard Data Formats

Historically, computer scientists would rather share a toothbrush than a data format.

But accepted standards are now available, e.g.:

- *CSV files*: for tables like spreadsheets
- *XML*: for structured but non-tabular data.
- *JSON*: Javascript Object Notation for APIs.
- *SQL databases*: for multiple related tables.

Where Does Data Come From?

The critical issue in any modelling project is finding the right data set.

Large data sets often come with valuable metadata: e.g. book titles, image captions, Wikipedia edit history...

Repurposing metadata requires imagination.

Sources of Data

- Proprietary data sources
- Government data sets
- Academic data sets
- Web search
- Sensor data
- Crowdsourcing
- Sweat equity

Useful (but uncurated) dataset search engine: toolbox.google.com/datasetsearch

Proprietary Data Sources

Facebook, Google, Amazon, Blue Cross, etc. have exciting user/transaction/log data sets.

Most organizations have/should have internal data sets of interest to their business.

Getting outside access is usually impossible.

Companies sometimes release rate-limited APIs, including Twitter and Google.

Government Data Sources

- City, State, and Federal governments are increasingly committed to open data.
- Data.gov has over 100,000 open data sets!
- The Freedom of Information Act (FOI) enables US citizens to ask if something is not open.
- Preserving privacy is often the big issue in whether a data set can be released.

Academic Data Sets

- Making data available is often a requirement for publication in many fields.
- Expect to be able to find economic, medical, demographic, and meteorological data if you look hard enough.
- Track down from relevant papers, and ask.
- Google topic and “Open Science” or “data”

Web Search/Scraping

Scraping is the fine art of stripping text/data from a webpage.

Libraries exist in Python to help parse/scrape the web, but first look for existing solutions:

- Are APIs available from the source?
- Did someone previously write a scraper?

Terms of service limit what you can legally do.

Available Data Sources

- Bulk Downloads: e.g. Wikipedia, IMDB, Million Song Database.
- API access: e.g. New York Times, Twitter, Facebook, Google.

Be aware of limits and terms of use.

Sensor Data Logging

The “Internet of Things” can do amazing things:

- Image/video data can do many things: e.g. measuring the weather using Flicker images.
- Measure earthquakes using accelerometers in cell phones.
- Identify traffic flows through GPS on taxis.

Build logging systems: storage is cheap!

Crowdsourcing

Many amazing open data resources have been built up by teams of contributors:

- Wikipedia/Freebase
- IMDB

Crowdsourcing platforms like Amazon Turk enable you to pay for armies of people to help you gather data, like human annotation.

Sweat Equity

But sometimes you must work for your data instead of stealing it.

Much historical data still exists only on paper or PDF, requiring manual entry/curation.

At one record per minute, you can enter 1,000 records in only two work days.

Often projects require sweat equity.

FAIR Data Principles

- Guidelines for data sharing (esp. publishing)
  - **F**indable – data can be located and navigated easily, e.g. through unique identifiers and rich metadata
  - **A**ccessible – data is available through standard protocols, meta data remains available permanently
  - **I**nteroperable – data can be integrated with other data, tools & systems seamlessly, e.g. via standard formats
  - **R**eusable – data is clearly documented and licensed for use by others, ensuring transparency & reproducibility

Supervised Learning
Correlation
Errors & Artifacts
Variance
Gradient Descent
Sampling
Data Bias
Probability
Significance
Precision
Skew
Classification
F-Score
Recall
Charts & Plots
Unsupervised Learning
Machine Learning
Statistics
Prediction
Logistic Regression
Linear Regression
Clustering
Bias-Variance Tradeoffs

Data Science 1:
Introduction to Data Science

Data Munging,

Data Cleaning & Data Bias

Winter 2025

Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Cleaning Data: Garbage In, Garbage Out

Many issues arise in ensuring the sensible analysis of data from the field, including:

- Distinguishing errors from artifacts.
- Data compatibility / unification.
- Imputation of missing values.
- Estimating unobserved (zero) counts.
- Outlier detection.

Errors vs. Artifacts

- Data errors represent information that is fundamentally lost in acquisition.
- Artifacts are systematic problems arising from processing done to data.

The key to detecting artifacts is the sniff test, examining the product closely enough to get a whiff of something bad.

First-time Scientific Authors by Year?

In a bibliographic study, Skiena et al. analyzed PubMed data to identify the year of first publication for the 100k most frequently cited authors.

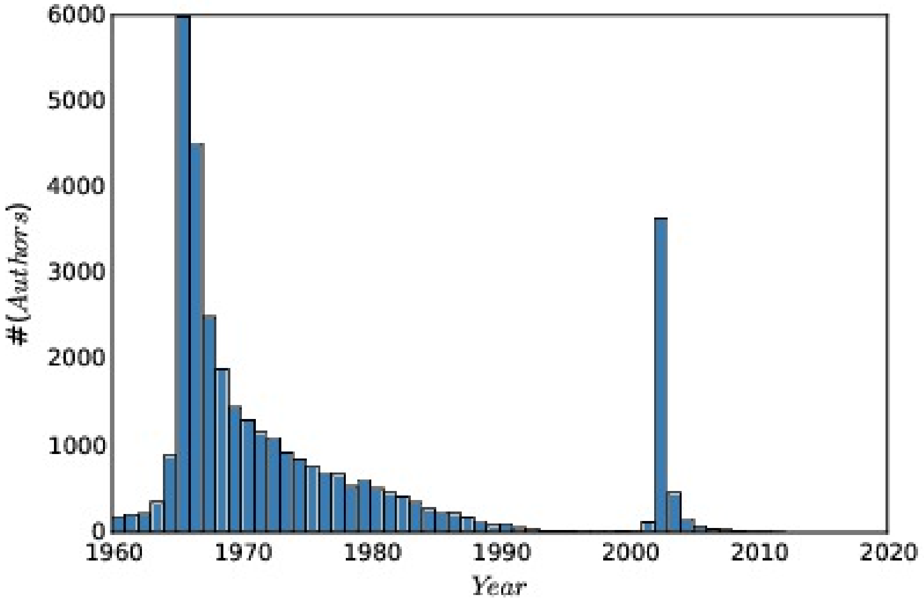
What *should* the distribution of new top authors by year look like?

It is important to have a preconception of any result to help detect anomalies.

Might this be Right?

What artifacts do you see?

What possible explanations could cause them?

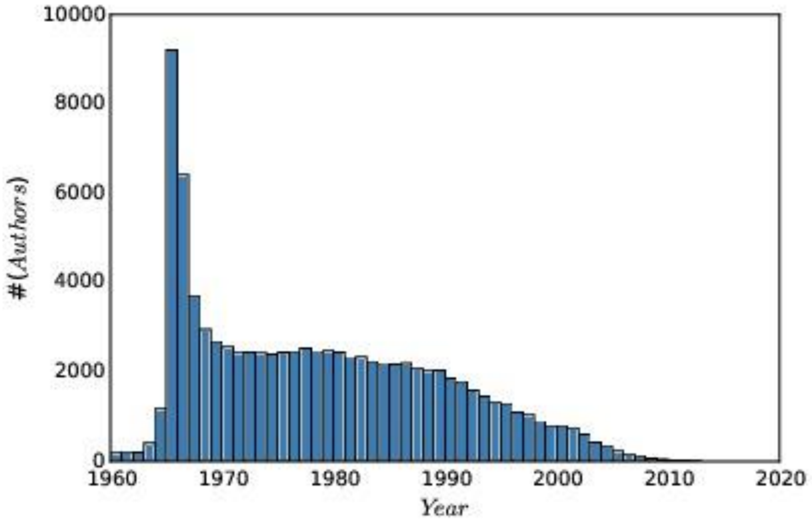


Mystery Solved!

Pubmed used author first names starting in 2002.

SS Skiena became Steven S Skiena

Data cleaning gets rid of such artifacts.



Data Compatibility

Data needs to be carefully massaged to make “apple to apple” comparisons:

- Unit conversions
- Number / character code representations
- Name unification
- Time/date unification
- Financial unification

Unit Conversions

NASA's Mars Climate Orbiter exploded in 1999 due to a metric-to-English conversion issue.

- Even sticking to the metric system has potential inconsistencies: cm, m, km?
- Bimodal distributions can indicate trouble
- Z-scores are dimensionless quantities.

Vigilance in data integration is essential.

Number / Character Representations

The Ariane 5 rocket exploded in 1996 due to a bad 64-bit float to 16-bit integer conversion.

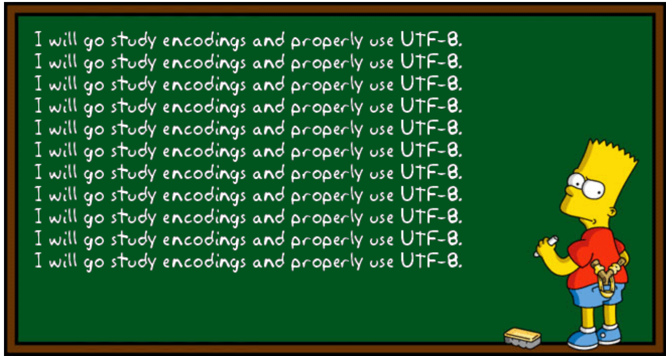
- Measurements should generally be decimal numbers
- Counts should be integers
- Fractional quantities should be decimal, not (q,r) pairs like (pounds, oz) or (feet, inches)

Character Representations

A particularly nasty cleaning issue in textual data is unifying character code representations:

- ISO 8859-1 is a single byte code for ASCII
- UTF-8 is a multibyte encoding for all Unicode characters.

Unicode font, UTF8 format	Unicode font, XXX... format
搜索简体中文网页	????????
Recherche avancée	Recherche avancée
網路畫廊，含中、港、澳參展作品	??????????????
โทรตˆดง่ายกว่า	?????????????
ウエブ全体から	???????
kehren Sie zur Suche zurück	kehren Sie zur Suche zurück
Сделайте Google стартово	???????? Google ????????
إخدوقت بحث أقل ووقت مطالعة أطول	?????? ??? ??? ???? ?????? ????



Name Unification

I appear on the web as:

(Wolle|Wolfram|W.|Wolfgang) (Wingerath|Winegrat)

- Use simple transformations to unify names, like lower case, removing middle names, etc.
- Consider phonetic hashing methods like Soundex and Metaphone.

Tradeoff between false positives and negatives.

Time / Date Unification

September 1752

Su	M	Tu	W	Th	F	Sa
-	-	1	2	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Aligning temporal events from different datasets/systems can be problematic.

- Use Coordinated Universal Time (UTC), a modern standard subsuming GMT.
- Financial time series are tricky because of weekends and holidays: how do you correlate stock prices and temperatures?

Time / Date Unification

Aligning temporal events from different datasets/systems can be problematic.

- Use Coordinated Universal Time (UTC), a modern standard subsuming GMT.
- Financial time series are tricky because of weekends and holidays: how do you correlate stock prices and temperatures?

Tom Scott: The Problem with Time & Timezones, Computerphile (2014).

Financial Unification

- Currency conversion uses exchange rates.
- Correct stock prices for splits and dividends.
- Use returns / percentage change instead of absolute price changes.
- The time value of money needs correction for inflation for fair long-term comparisons.

Why do stock/oil prices correlate over 30 years?

Dealing with Missing Data

An important aspect of data cleaning is properly representing missing data:

- What is the year of death of a living person?
- What about a field left blank or filled with an obviously outlandish value?
- The frequency of events too rare to see?

Setting such values to zero is generally wrong

Imputing Missing Values

With enough training data, one might drop all records with missing values, but we may want to use the model on records with missing fields

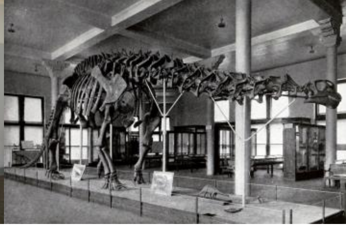
Often it is better to estimate or impute missing values instead of leaving them blank.

Example:
A good guess for your death year is birth + 80.

Imputation Methods

- *Mean value imputation* - leaves mean same.
- *Random value imputation* - repeatedly selecting random values permits statistical evaluation of the impact of imputation.
- *Imputation by interpolation* - using linear regression to predict missing values works well if few fields are missing per record.

Outlier Detection

The largest reported dinosaur vertebra is 50% larger than all others: presumably a data error.

- Look critically at the maximum and minimum values for all variables.
- Normally distributed data should not have large outliers, *k sigma* from the mean.

Fix why you have an outlier. Don’t just delete.

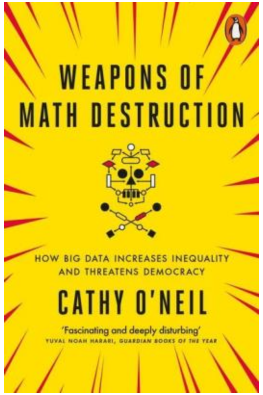
Detecting Outliers

- Visually, it is easy to detect outliers, but only in low dimensional spaces.
- It can be thought of as an unsupervised learning problem, like clustering.
- Points which are far from their cluster center are good candidates for outliers

Delete Outliers Prior to Fitting?

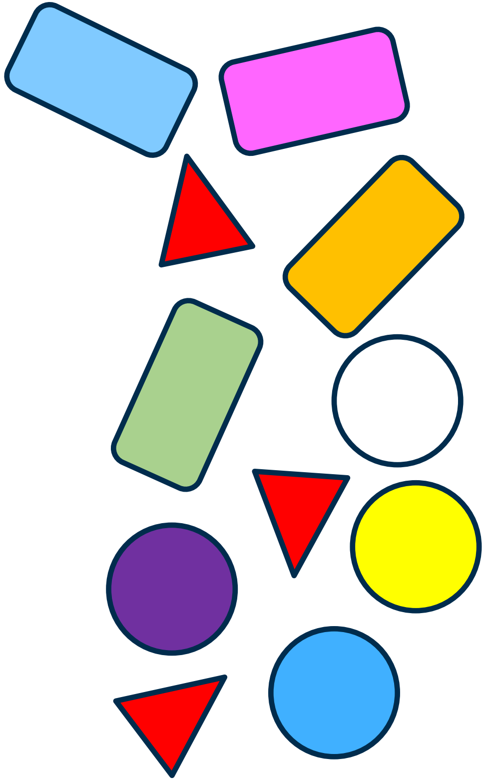
- Deleting outliers prior to fitting can yield better models, e.g. if these points correspond to measurement error.
- Deleting outliers prior to fitting can yield worse models, e.g. if you are simply deleting points which are not explained by your simple model.

Data Bias



„we have this belief – which is just wrong – that data itself is inherently objective"

—Cathy O'Neil



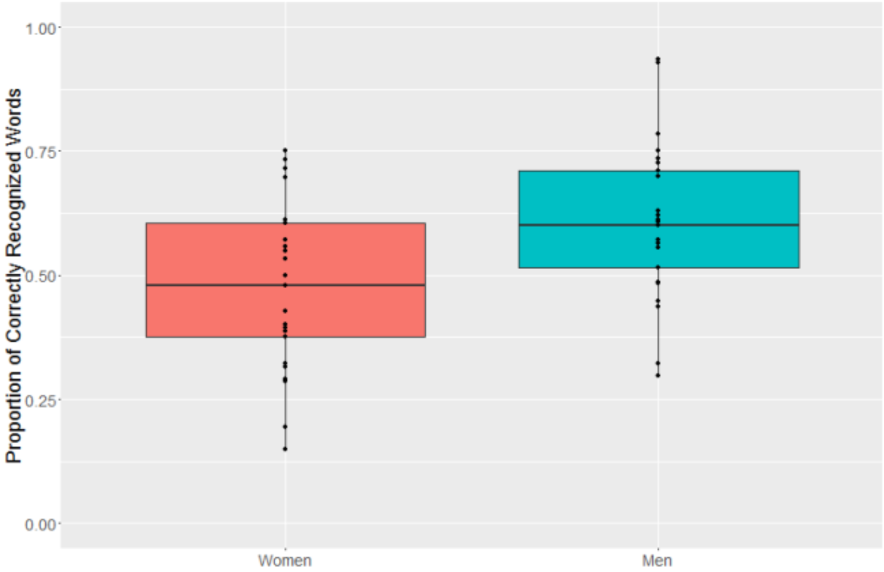
Data bias is a systematic error or distortion in the data pipeline (collection, analysis, interpretation, presentation) that results in a skewed or inaccurate representation of the underlying population.

Example: Gender-Data Gap in ASR

Automatic Speech Recognition (ASR) training data is skewed and thus does not work equally well for everybody.

„women have a tougher time using speech-recognition technology than men because the systems have a hard time deciphering what was actually said."

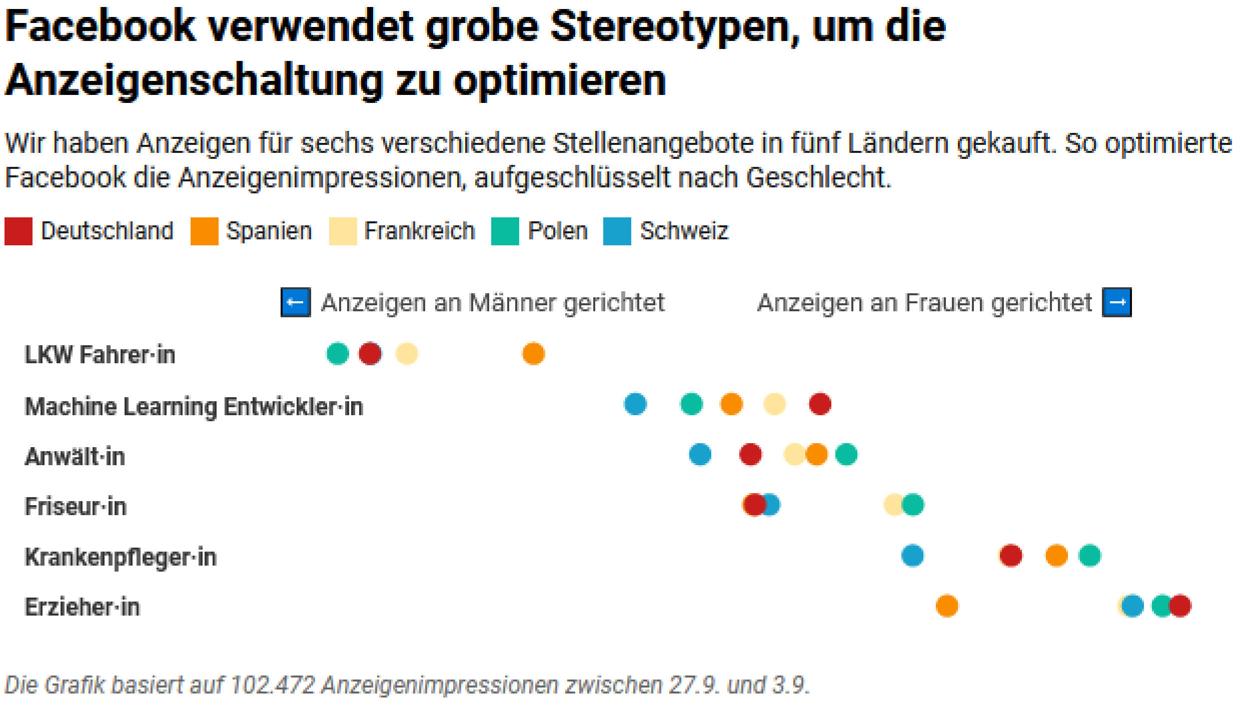
Graeme McMillan: It's Not You, It's It: Voice Recognition Doesn't Recognize Women, Time (2011).



„On average, for each female speaker less than half (47%) her words were captioned correctly. The average male speaker, on the other hand, was captioned correctly 60% of the time."

Rachael Tatman: Google's Speech Recognition Has a Gender Bias, Making Noise & Hearing Things (2016).

Example: Gender-Biased Job Ads

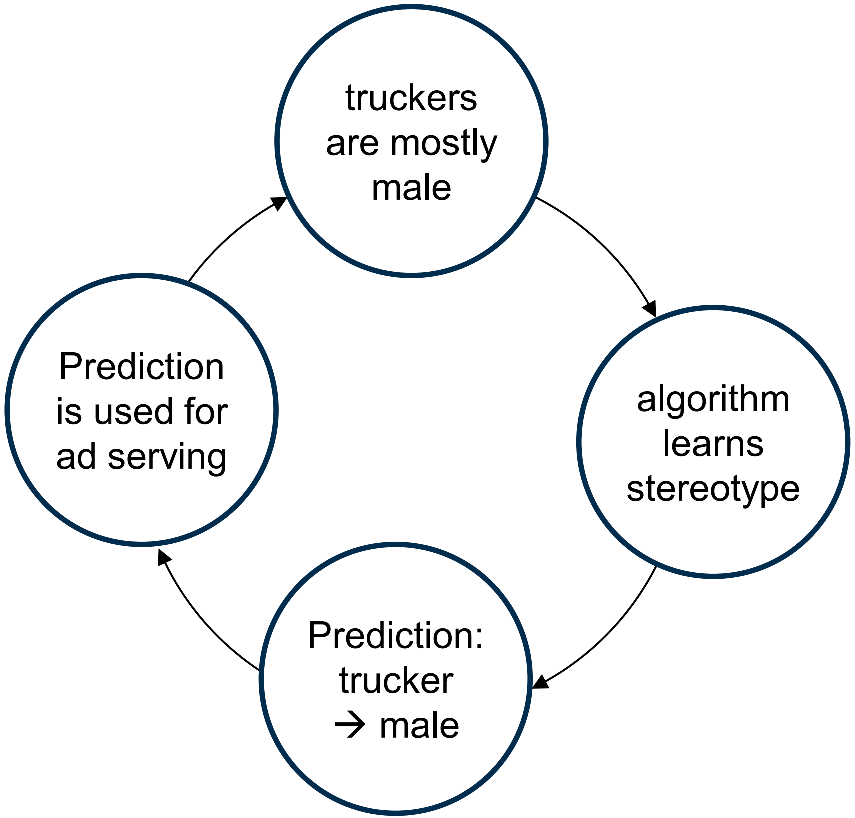


Nicolas Kayser-Bril: Automatisierte Diskriminierung: Facebook verwendet grobe Stereotypen, um die Anzeigenschaltung zu optimieren, AlgorithmWatch (2020).

Bias Amplification

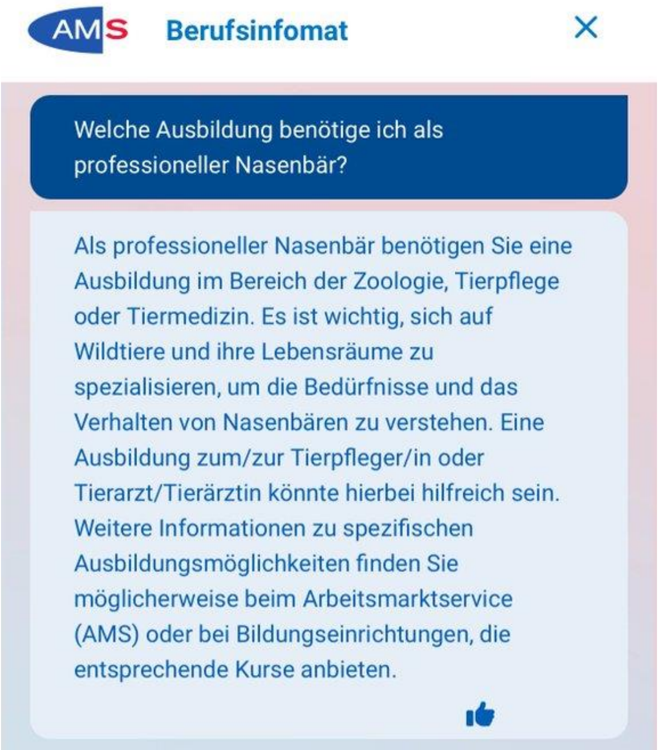
Models trained on biased data produce biased predictions.

Unless mitigated, this can reproduce or even amplify existing stereotypes.



Prominent Example: Berufsinformat

- AMS (Austrian Public Employment Service) released a ChatGPT-based career advisor

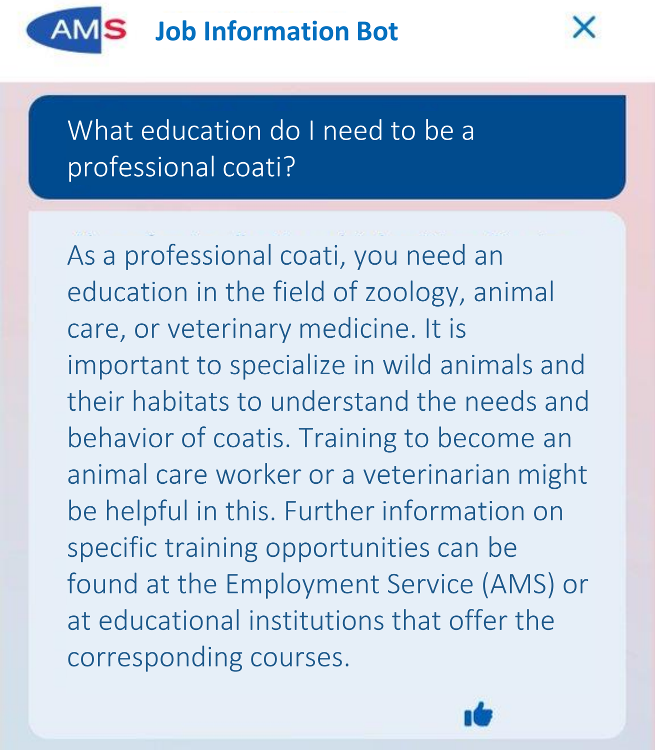


Chris Köver: Diskriminierung: AMS erntet Hohn mit neuem KI-Chatbot, Netzpolitik.org (2024).

Michael Staudacher: Ich sehe nichts, was mich zum jetzigen Zeitpunkt am Einsatz von KI in der Berufsberatung zweifeln ließe., Twitter (2024).

Prominent Example: Berufsinformat

- AMS (Austrian Public Employment Service) released a ChatGPT-based career advisor
- The system reflected heavy gender bias as it routinely recommended IT and commerce careers for men while recommending gender studies or philosophy for women
- Lessons learned (among others):
  - Biases inherent to training data are difficult to remove from a trained model
  - Rigorous testing is not optional, even with a sophisticated base model like ChatGPT



Chris Köver: Diskriminierung: AMS erntet Hohn mit neuem KI-Chatbot, Netzpolitik.org (2024).

Michael Staudacher: Ich sehe nichts, was mich zum jetzigen Zeitpunkt am Einsatz von KI in der Berufsberatung zweifeln ließe., Twitter (2024).

Wrapup:

Data Munging, Cleaning & Bias

- **Data Munging**: acquiring and preparing data for analysis
- **Data Cleaning**: identifying and removing errors, e.g. via unification, imputation, outlier handling
- **Data Bias** occurs when the data does not accurately represent the studied population, leading to distorted or inaccurate conclusions