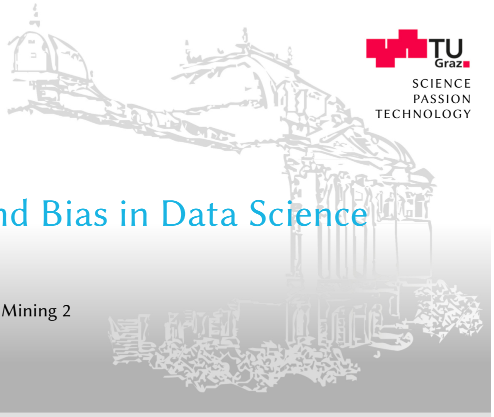


SCIENCE
PASSION
TECHNOLOGY

Assumptions and Bias in Data Science

Roman Kern

Knowledge Discovery and Data Mining 2

> **Motivation**: When generating insights from data, we have to make many (explicit, or accidental) assumptions, and there might be many biases - both will negatively influence the validity of our results.

> **Goal**: This lecture aims to make the implicit assumption explicit, understand their implications, be aware of the biases and their countermeasures. Finally, one should be able to apply the matching methods to a given problem setting and achieving valid results.

Assumptions and Bias in Data Science

Outline

1 Introduction

2 Assumptions

3 Bias

4 Fairness

5 Shifts

Introduction

Motivation & Basics

Introduction

Example “Apple Heart Study” I

- Study with 419,297 participants (volunteers, who responded) to diagnose *atrial fibrillation*
- Notification from the watch
  - Based on pulse measurements ❤
  - ... should seek advice & follow-up analysis
- 2,161 got notification, 658 got further analysed (ECG), only 450 usable
- Authors report 84% who got notification, actually were (later) diagnosed with atrial fibrillation

> Taken from `https://catalogofbias.org/2019/11/14/big-is-not-always-beautiful-the-apple-heart-study/`.

> The study: Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., ... & Hung, G. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. New England Journal of Medicine, 381(20), 1909-1917.

> Not blinded design study (i.e., open-label).

Introduction

Example “Apple Heart Study” II

- **Ascertainment bias**
  - Some participants are less likely to be included in the results
- **Compliance bias**
  - Not all participants adhere to same protocol
- **Detection bias**
  - Different ways to measure the outcome

> And potentially **spin bias**, **hot stuff bias**, and **confirmation bias**.
> Although this study deals with patients/participants, this also applies to other entities, e.g., machines.

Introduction

Motivation

- In data science we are given some (observational) data, and we need to gain **(correct) insights** from it!
- For this to work, we have to make a number of **assumptions**
  - About the data generation process, about the nature of the data, ...
  - ... since also our algorithms make (implicit) assumptions
- → A mis-match of these assumption will render the result invalid 🐞!

> We always make assumptions, be it explicit, or **accidental** (implicit, intrinsic)!
> In short, we always get a (numeric) result, but is it correct?
> Recall the danger zone in the data science Venn diagram.

Introduction

Motivation

- Also, the **data** might have a (systematic) bias
  - ... and our analysis will “inherit” this bias
  - → invalid (biased) results 💣!
- Also, **algorithms** might have a bias
  - ... and our results will reflect this
  - → invalid (biased) results 💣!

> In any case, we need to be aware of the assumptions and the bias.

Assumptions

What we implicitly/explicitly take for given

Assumptions

Smoothness Assumption

- Basic assumption in statistics: **smoothness assumption**
  - e.g., Needed to estimate $P(X)$ given a sample of data
  - Also applies to machine learning & data science
  - Assumption about a process or a data set
- Intuition
  - The close vicinity of data point is “similar” to the data point
  - e.g., $P(Y = y|X = x) \approx P(Y = y|X = x + \varepsilon)$

> For example, a *Kernel Density Estimator* makes this assumption.
> Without this assumption, ML might fail (see also universal approximation theorem).
> Assumption about the data generating process.
> Whatever similar is in the actual context.
> For more detail please also see *Lipschitz continuity* for a more rigorous definition.
> There is also a connection to *Differential Privacy*.

Assumptions

Smoothness Assumption

The smoothness assumption does not hold for

- Chaotic systems
  - e.g., Fractals
- Stochastic systems
  - e.g., Random data

☞ Need to understand the generative process (domain)

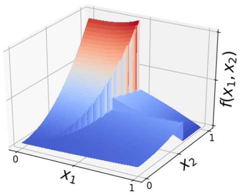
> Both may result in randomness, while the chaotic system follows a deterministic generation process.

Assumptions

Smoothness Assumption

The smoothness assumption does not hold for

- Non-continuous functions
  - e.g., Dirac δ function, step functions, ...
  - Less severe case
    - With *local smoothness*
    - e.g., A decision tree can model such behaviour
- Categorical variables (incl. binary)



☞ Need to understand the datset

> The data generation process needs to be continuous (in practical terms: watch out of **“if-clauses”** in the data generation process).
> Deep neural networks are capable to approximate a **piecewise smooth function**, e.g., a step function (https://arxiv.org/pdf/1802.04474.pdf).

Assumptions

Smoothness Assumption - Counter Measures I

- **More data**
  - Increase *instances*
  - ... will yield a better estimate (of local smoothness)
- **More data**
  - Increase *features*
  - ... may improve to explain the “unsmoothness”

> See also https://www.linkedin.com/pulse/smoothness-underlying-assumption-machine-learning-ravi-kothari/

Assumptions

Smoothness Assumption - Counter Measures II

- **Pre-Processing**
  - Transform feature (smooth)
    - e.g., splitting into multiple features
- **Fitting models**
  - Avoid models that (heavily) rely on smoothness
    - e.g., linear models

> Generally, it is a good idea at the beginning to look into the features, e.g. with **visual tools**.

> As hinted, the model differ on their dependence on the smoothness assumption (deep neural networks with non-linear activation function should be efficient in such cases).
> ... but watch out for **model bias**!

Assumptions

Unconfoundedness

- Assumption: There are no **hidden confounders** (in our data)
  - Hidden confounders may systematically introduce variance
    - e.g., partial correlation between variables
    - e.g., visible as multi-collinearity in the data

Worst case scenario

Hidden confounder is constant throughout our (training) data set, but changes to different value(s) in the (test) set

> Features may appear redundant, since they correlate to a high degree.
> Due to some lurking variables, as **common cause**.
> If we want to understand the model (e.g., check if it is biased), some proxy variables get (mistakenly) identified as most important features (but not the true cause).

Assumptions

Identically and Independently Distributed - IID

- The **IID assumption** is well known in machine learning
- The distribution does not change
  - ... between the training and the test set → **identical**
- The instances are **independent** from each other
  - You could not predict the next data point
  - ☞ sequence does not play a role

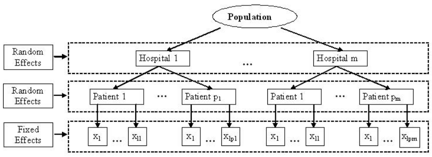
> Not the case, if there are some **confounders** as stated in the previous slide.
> The data follows a **stationary distribution**.
> i.e., $P(X_{train}) = P(X_{test})$
> Especially important, if our training dataset precedes the test dataset, as typical in real-world scenarios (i.e., use historic data to predict the present/future).
> Simple example, if we train a system to predict ice cream sales with data from the summer month, but then apply them on winter month, the results might be sub-optimal.

> Every row in training & test set must be identically distributed

> Another approach is non-random sampling, e.g., Kennard Stone Algorithm
> R. W. Kennard & L. A. Stone (1969): Computer Aided Design of Experiments, Technometrics, 11:1, 137-148.

Assumptions

Non-IID - Countermeasures



- **Add features**
  - Make non-independence explicit
  - ... and the machine learning algorithms can make use of it
  - ☞ requires domain knowledge

> Dundar, M., Krishnapuram, B., Bi, J. and Rao, R.B. 2007. Learning classifiers when the training data is not IID. IJCAI International Joint Conference on Artificial Intelligence (2007), 756–761.
> Introduce **synthetic features** to capture the random effects, e.g., hospital.
> For example, a decision tree can use this feature as splitting criterion.
> ... so at least within this “bucket”, the IID assumption should hold.

> Possible way to detect via statistical tests (compare distributions).

Assumptions

Assumption in Time Series I

Stationarity (strict, weak)

- Intuition
  - Mean and variance constant over time
  - The probability of observing a sub-sequence does not depend on time
- *Weak stationarity*
  - Mean is constant, variance is finite
  - Autocovariance does not depend on time

> See the lecture on time series in KDDM2 (especially for countermeasures).
> IID implies strict stationarity, but strict stationarity does not imply IID.
> We already know that time series data is typically not expected to be IID (e.g., we expect autocorrelation).
> Stationary is hard to assess, typically requires domain knowledge (of the data generating process).

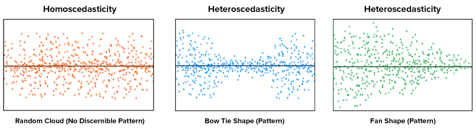
> Strict & weak stationarity imply **homoscedasticity**.

Assumptions

Assumption in Time Series II

Homoscedasticity (not limited to time series)

- We assume that certain “statistics” will not change over time
  - e.g., variance, noise
- If violated, it is called *Heteroscedasticity*



> Not exclusive to time series, also an assumption in linear regression. https://towardsdatascience.com/assumptions-of-linear-regression-algorithm-ed9ea32224e1.
> For example, if the noise is not IID.

Assumptions

Cluster Hypothesis

Search Results

- Closely associated documents tend to be relevant to the same requests
- 👉 justifies pseudo-relevant feedback

Machine Learning

- Algorithms like *k-means clustering* or *k-NN classification*
  - ... assuming close document to be similar

> Documents that form clusters are expected to be similarly relevant to queries. https://ie.technion.ac.il/~kurland/clustHypothesisTutorial.pdf.

> Where closeness is computed via a distance measure, which sometimes is also called a similarity measure (i.e., the assumption is also contained in the measure).
> In data science we encode our (domain) knowledge of similarity within the distance/similarity function.

> Does other apply to other areas, e.g., in network analysis, the assumption is that communities (densely-connected regions) also share similarities, which is consecutively used for, e.g., recommender systems.

Assumptions

Independent Variable Assumption

- **Bag-of-Words**
  - Split a sentence/document into words, ignoring their sequence
    - 💬 Words are clearly not independent from each other!
- **Naive Bayes**
  - Each variable (feature) is assumed to be independent from all others
  - ... and provides (often) good performance
    - e.g., Text classification

> Why can it be that Naive Bayes works so (unreasonably) well, if the assumption is obviously violated?
> Also called **maximum conditional independence** (inductive bias).
> Should we even care about assumptions?

Assumptions

Multicollinearity

- A dataset might contain **multicollinearity**
  - If more then one variable are *linear related*
  - **Perfect multicollinearity** for identical variables
- e.g., linear regression may behave erratically
  - Small changes in the data may have big changes in the output

> Multiple variables encode the same phenomena.
> Often results of a confounder, yielding partial correlation.

> Most often the problem, if only a few data points (small dataset) and increased risk with more correlating variables, see **spurious correlation**.
> Here the problem might be that the **difference between two (highly similar) variables** (e.g., due to randomness) appears to have predictive power on the target variable.

Assumptions

Multicollinearity - Countermeasures I

- **Less data**
  - Remove “redundant” variables
- **More data**
  - Additional observations (rows)

Assumptions

Multicollinearity - Countermeasures II

- **Preprocessing**
  - e.g., Principal Component Analysis
    - Multiple variables will be then represented by a single principal component
- **Regularise**
  - e.g., Ridge regression

> Since PCA is unsupervised it is agnostic to the spurious relations.
> ... but PCA is not robust.

Assumptions

Assumptions Regarding Regularisation

- Often we regularise our model
  - ... to include **fewer features**, e.g., L1 norm
  - Assuming less is “better”
- Or, we make explicit *feature selection*
- → assuming fewer features yield a better model

> There are many types of regularisation, some do affect the size of the coefficients.
> See also parsimonious models.
> The complexity of the model (here: amount of features) should match the complexity of the problem.
> Also applies to Occam’s razor (see in a few slides).

> Hope that a sparser model also **generalises better** as is should reduce the risk of overfitting.

> Use techniques from **interpretability** to manually check model for plausibility.

Assumptions in Causality

- *Causal inference* and *causal discovery*
  - … do not work without making (strong) assumptions
- For example
  - *Faithfulness assumption*: to see a correlation where there is causality
  - *SUTVA*: no interference, no hidden variance in treatment levels

> Please see slides on causal data science for details.

Gaussian Assumption I

- Often we assume that our data (or our noise) follows the **Gaussian distribution**
- Justified by the **central limit theorem** ❗
  - Given multiple random variables
    - … even if they are non-normally distributed
  - Their normalized sum will tends to be **normally distributed**

> And since we assume that many natural processes are actually a mixture of many underlying phenomena, we have a tendency to attribute a Gaussian behaviour to many observations.

Gaussian Assumption II

- However, this is based on the **assumptions**
  - The random variables are *independently and identically distributed*
  - … and have *finite variance*
- If the later is violated, but has power-law tails
  - … the sum will tend to a stable distribution
- Some statistical tests, e.g., **t-test**, requires the data to be Gaussian

> If the noise follows a Cauchy distribution, the mean and the variance will not be defined (i.e., not finite).
> One can (technically) standardize a Cauchy to 0-mean and 1-variance (z-normalisation), because empirical mean and variance exist. However, this does not change population mean and variance, all it might do is lead the analyst to spurious conclusions.
> See also generalised central limit theorem.

> The estimated mean of a Cauchy distribution is also Cauchy distributed across multiple datasets with the same Cauchy distribution (the empirical mean is the same as a random point drawn from the Cauchy distribution).

> For example, the X^2 test can be used to assess, if a dataset follows a given distribution (but there are many more).

Homogeneity of Variances

- Multiple sub-groups have the same **variance**
  - … required by some statistical tests, e.g., **ANOVA**
- Check of this assumption
  - Visualisation of the data to identify
  - Special tests to assess if this is the case
    - Levene's test, Fligner Killeen test, Bartlett's test

> See also `https://www.real-statistics.com/one-way-analysis-of-variance-anova/homogeneity-variances/`.

Assumptions

Occam’s Razor

- **Occam’s razor** dictates
  - Given two options
  - Pick the one with fewer assumptions
- For machine learning
  - ... this would relate to preference of simpler models
- But, a too simple model may just have bad *predictive performance*

> Similar to statistical tests, if there is no (statistical) evidence, make fewer assumptions, i.e., assume the observed difference is due to randomness (**null hypothesis**).
> Intrinsic methods to follow this guidance: regularisation, pruning (of decision trees).
> It can also be seen as an **inductive bias** of some algorithms.

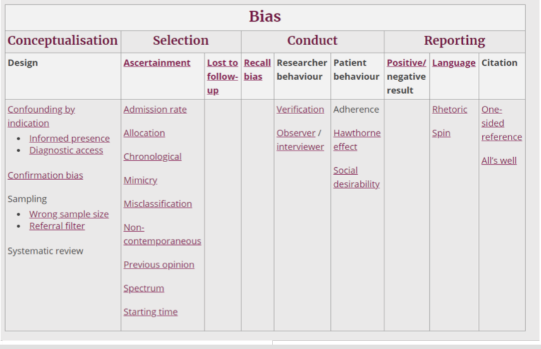
> See also *Principle of Maximum Entropy* (among multiple distributions, select the distribution which leaves you the largest remaining uncertainty).

Bias

What may influence our results?

Bias

Many!



Bias								
Conceptualisation	Selection		Conduct			Reporting		
Design	Ascertainment	Lost to follow-up	Recall bias	Researcher behaviour	Patient behaviour	Positive/ negative result	Language	Citation
Confounding by indication • Informed presence • Diagnostic access Confirmation bias Sampling • Wrong sample size • Referral filter Systematic review	Admission rate Allocation Chronological Mimicry Misclassification Non-contemporaneous Previous opinion Spectrum Starting time			Verification Observer / interviewer	Adherence Hawthorne effect Social desirability		Rhetoric Spin	One-sided reference All’s well

> Image taken from https://catalogofbias.org/2020/02/11/a-taxonomy-of-biases-progress-report/.
> The Catalogue of Bias list of **40 different types** of bias!
> Furthermore, this list is **not specific to data science**.

Bias

Bias in Machine Learning

- Bias is often an **additional parameter**
  - In neural networks, many regression methods, ...
  - $\mathit{Output} = \sum \mathit{weights} * \mathit{input} + \textbf{bias}$
  - ☞ unrelated to statistical bias

> This dummy parameter only shares the name with the concepts discussed in here.
> See Bishop, C. M. (2006). Pattern recognition and machine learning. for a more in depth analysis on the bias term (page 142f).

Bias in Data - Simple Example

ID	Weight	Name	Gender	Age	Label
e1	1.0	John	M	20	1
e2	1.0	Joe	M	20	0
e3	1.0	Joseph	M	20	0
e4	1.0	Sally	F	30	1
e5	1.0	Sally	F	40	0
e6	1.0	Sally	F	300	1

- **Quality impairments**
  - Duplicates e2 & e3
  - Outlier age in e6
- ☞ Will introduce bias e.g., arithmetic mean of age will be biased

> Example taken from: [1] Tae, K.H., Roh, Y., Oh, Y.H., Kim, H. and Whang, S.E. 2019. Data cleaning for accurate, fair, and robust models: A big data - AI integration approach. Proceedings of the ACM SIGMOD International Conference on Management of Data (2019).
> Task predict high income (label).
> Joe (e2, e3) has bigger influence on the results.
> Distribution of gender might be biased.

Bias of an Estimator

- Difference between the **expected value** of an estimator and the **true value**
- Unbiased estimator
  - ... with zero bias
    - ... does not exist without further assumptions
  - e.g., sample mean

> Assumptions about a population.
> Mean-unbiasedness vs median-unbiasedness - it need to be specified in relation to what an estimator is unbiased.
> In German called *Erwartungstreue*.
> e.g., assume the data does not follow a Cauchy distribution.
> No notion about **what** the bias actually is and **where** it comes from.
> A good estimator should also have a **low variance**.
> So it appears, an unbiased estimator is **always preferred**?

Stein Paradox

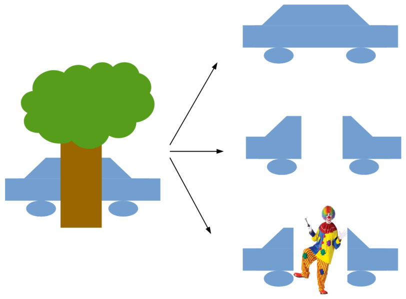
- **Biased** combined estimator for multiple parameters
  - Is **on average** better than individual (unbiased) estimators
  - ... even if the parameters are unrelated ↯
- **James–Stein estimator**
  - Biased estimator of the mean
  - $\widehat{\theta}_{JS} = \left(1 - \frac{(m-2)\sigma^2}{\|\mathbf{y}\|^2}\right)\mathbf{y}.$

> $Y = \{Y_1, Y_2, ..., Y_m\}$ with unknown means, with m the number of parameters to estimate, and assumed to be Gaussian with a known covariate matrix $\sigma^2 I$, and $\mathbf{y}$ are single observations.

> Stein's Paradox is caused by using MSE, i.e. the problem is with MSE and assumed Gaussianity.

> Example from baseball
» We want to estimate the batting score of all players, e.g., of a team
» Each player is a parameter, and the player's history
» ... can be used to compute the respective sample mean and variance
» The JS estimator gives on average better estimates than the "average of averages"
» Sideeffects: exceptionally good players will be over-corrected to the average players (and vice versa)

Inductive Bias



> Recall Occam's razor - guidance to make as few assumptions as needed.
> Simple question: **what is behind the tree**?
> Image credits: https://www.datascienceafrica.org/dsa2016/downloads/model_selection.pdf

Bias

Inductive Bias

Assumption on unseen data given observed data

- Examples
  - linear → linear relationships
  - k-NN → neighbours are similar
  - Maximum margin → boundary b/w classes, generalisation via distance to boundary
- In a Bayesian setting via the prior

> In machine learning bias is often seen as inductive bias or statistical bias (bias and variance).
> The inductive bias is part of the model, the main assumption.
> Also known as learning bias, variance hint.
> Declarative bias of the learner to choose a hypothesis/model.

> Bias is needed to obtain good generalisation:
> Mitchell, T. M. (1980). The need for biases in learning generalizations (pp. 184-191). New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers University.

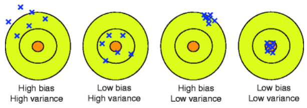
> Priors are almost always chosen because they are mathematically convenient, not because they have any connection to reality.

Bias

Bias and Variance

Trade-off between **bias and variance** ⚖

- A high bias represent models that make strong assumptions
  - e.g., a linear model assumes the response to be linear
- A high variance represents models that can adapt well
  - ... they can represent many hypothesis



> For more detail look into VC-dimension and PAC learning.
> The dartboard analogy was put forward by Moore & McCabe (2002)

> Please see here: (2011) Bias Variance Decomposition. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_74

> More recently, with the progress in deep learning, some authors argue that the trade-off does not apply to deep neural networks, see https://www.bradyneal.com/bias-variance-tradeoff-textbooks-update

Bias

Bias in Evaluation Measures

- **Accuracy** is biased towards the majority class
  - In skewed data sets
  - e.g., 90/10 split between majority and minority class
    - Always returning the majority will yield a accuracy of 90%
- **Cross-validation** (CV)
  - Try out multiple models
  - ... pick to one with the best CV performance
  - No free lunch → biased

> Using accuracy is therefore less useful in such settings, prefer F_1, or Matthews Correlation Coefficient.

> No free lunch theorem tells us that there is no best bias, there are **equally many problems** where one model is better than the other.

> → Select the bias/model/algorithm **according** to the problem (data), often domain knowledge is the key here ❗

Bias

Confounding Bias

- Bias due to **common cause**
- Preferred solution: randomised controlled studies
- Alternatively, condition on the common cause
  - i.e., render its effect out

> Example: https://medium.com/causal-data-science/understanding-bias-a-pre-requisite-for-trustworthy-results-ee5
> More active users may skew the distribution of searching for a product or seen an ad.
> “Activity bias”

Bias

Funding Bias I

- Relationship between **eggs** and (serum) **cholesterol level**
- Analysis of studies
  - 0% industry funded studies in the 1950s
  - 60% in the timespan 2010 - 2019

> Barnard, N. D., Long, M. B., Ferguson, J. M., Flores, R., & Kahleova, H. (2019). Industry Funding and Cholesterol Research: A Systematic Review. American Journal of Lifestyle Medicine, 1559827619892198.

Bias

Funding Bias II

- Results of studies, in the results section
  - Non-industry: 93% report increase[1]
  - Industry: 83% report increase
- Interestingly, 49% of industry funded studies found discordant
  - Statements in conclusion did not match results ⚡

[1] not necessarily statistically significant

> It might be that the funding bias had an effect on how the authors presented the results of their work in the conclusions of their studies.
> Readers should be aware of the implications.

Bias

Selection Bias

Goal: hide the relationship between egg consumption and serum cholesterol

- **Selection bias**
  - Patients with already high levels
  - ... are not sensitive to additional dietary intake
  - Select them as study participants → observe no treatment effect (e.g., one extra egg per day)

> We have discussed selection bias already previously (yielding the **Berkson's paradox**).
> Recall the Apple heart study - are the participants a **unbiased sample** from the population?
> Another idea would be to conduct a series of small studies (few participants), and then only select those, where the results were (due to randomless) not significant, and the publish a meta-study summarising the small studies.
> Selection (or sampling) bias is not limited to scenarios as described, unfortunately in many data collection processes there is an **intrinsic sampling bias**.

Bias

Publication Bias

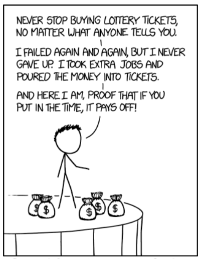
- The **outcome** of a study has an effect on the **publication** of results
- 👉 Only **good results** are published, bias towards positive results
- Creates incentive to *p-value hacking* and *HARKing*

> Another selection on publication is novelty, papers that show known relationships (e.g., egg → cholesterol), will be less likely written/published.
> Similar to the combination bad film/bad book is rarely observed, negative results in publication are rarely observed.
> p-value hacking - e.g., via many hypothesis (include many parameters) and due to randomness some results will yield false positives (Type I error).
> Hypothesizing After the Results are Known - once a signification relationship has been found, a suitable hypothesis is build to support the findings.

Bias

Survival Bias

- Special form of selection bias
  - “*Men get tough fighting in the coliseum*” vs “*only tough men survive the coliseum.*”
- Due to randomness (spurious) or systematic
- If systematic, the selection process suppresses to observe a fair, unbiased sample


NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
$
$

EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT SURVIVORSHIP BIAS.

> If the same experiment is conducted by multiple teams, even if there is no effect, one of the teams might be “lucky” to observe an effect due to randomness - i.e., this team was “selected for survival”.
> Image credits: https://xkcd.com/1827/
> See also: https://dataschool.com/misrepresenting-data/survivorship-bias/
> Another example: **stone age** (a only stone tools survived the times).

Bias

Cognitive Bias

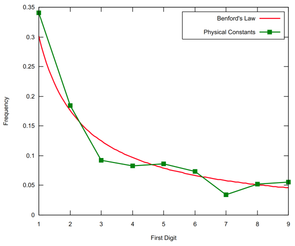
- Even (domain) **experts** are not immune to bias
- **Confirmation bias**
  - Prefer a hypothesis in line with previous assumption
- **Anchoring effect**
  - Judgement changes depending on the sequence
  - Problematic, if dataset contains subjective “ground truth”

> Since domain expertise plays an important role, and my biases can only be detected via expert knowledge, it is important to understand that there is bias as well.
> There are **much more** cognitive bias than listed here!

Bias

Bias Detection

- **Benford’s Law**
  - ... for numerical data with multiple orders of magnitude
  - Expect a power-law distribution
    - Of the first digit
    - Deviations indicate some **non-random process**



> By Drnathanfurious at English Wikipedia - Transferred from en.wikipedia to Commons by Tam0031 using CommonsHelper., Public Domain, https://commons.wikimedia.org/w/index.php?curid=6948975.

Bias

Bias Countermeasures I

- **Scientific rigour** (aka More Knowledge)
  - Understand the problem setting
  - Expectations on the data generating & collecting process
- **Repetition** (aka More Data)
  - Bias is most harmful, if not in the data
  - Collect same data, but different context
    - Where one would expect the same outcome
    - e.g., multiple domain experts, different sequence, ...

Bias

Bias Countermeasures II

- **Algorithmically** (statistically) remove bias
  - Over-correcting for bias may introduce unwanted bias
    - ... or moved true relationships

> Correct for a mediator might render a relationship invisible.
> e.g., Correct for cholesterol level when measuring the impact of a healthy diet on heart attack rates will make the influence of the diet disappear, or a least appear less prominent.

Fairness

When do we consider an algorithm to be fair?

> Please also have a look at the privacy-preserving lecture, where the notion of fairness has been introduced in a systematic way.

Fairness

Introduction to Algorithmic Fairness

- In a pre-algorithmic world
  - **Humans** made decisions
  - ... taking the **law** and other **constraints** into account
- Now, algorithms are expected to provide decisions
  - We can observe an **algorithmic bias**

> https://fairmlbook.org/
> There are two types of scenarios:
> 1. **Decision support systems** - where the algorithms only make suggestions, but the final decision it taken by a human.
> 2. **Decision making systems** - where the algorithms take the final decision (without manual intervention).
> Suggested tutorial: https://mrtz.org/nips17

Fairness

Examples of Algorithmic Bias

- **Recruitment**
  - Algorithms prefers gender/age/...
- **Face detection**
  - Error rate vary w.r.t to ethnicity/smiling/...
- **Credit assessment**
  - Based on location...

> https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-po

> Cases of unwanted bias in algorithmic outcomes, but also example of wanted bias, e.g., Apple users get targeted ads (for more expensive products).

> Example from natural language processing:
> **Gender bias** in word embedding: “Man is to Computer Programmer as Woman is to Homemaker”
> Inherited from the **underlying datasets**.
> 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.

Fairness

Algorithmic Fairness

Cause of algorithmic bias

- Based on datasets capturing (human) decisions
  - ... faithfully replicate
  - ... including **historical human biases**
- Obvious limitation
  - The algorithm does **not** directly “observe” the law & constraints
    - i.e., the machine only observes a subset of “features”

> The supervised machine learning algorithm are trained on human decisions.
> Since the machine has no **world knowledge**, there is no way to correct for biases in the data.

Fairness

Algorithmic Fairness

Other reasons for bias

- Too small dataset
  - e.g., influence of randomness, sampling bias
- Measurement errors
  - e.g., how to measure subjective judgements?
  - Watch out for automatically generated assessments
    - e.g., sentiment detection, spam, ...

> If a **minority class** has just a few samples, the randomness (or systematic differences) play a bigger role.
> For example, the quality of products may be judged by a person; What happens when this **person leaves** the organisation (see annotator shift)?

Fairness

Algorithmic Fairness

Implications

- Biased decisions
- **Feedback loops**
  - Bias will be include in succeeding training data
  - ☞ reinforcement

Fairness

Algorithmic Fairness

Detecting bias

- Validation datasets / scenarios
  - e.g., comparison of expected value and output
  - Might be a dedicated (data science) project
- Interpretability
  - Explainable AI (XAI)

> For example, in an abduction study one tries to find out, how certain results have been achieved.
> The field of explainable AI is large, many different approaches (e.g., Shaply values), include methods for interpretability.

Fairness

Algorithmic Fairness - Countermeasures

- **More Data**
  - e.g., remove influence of randomness, & sampling bias
- **More Knowledge**
  - Capture domain knowledge
    - Introduce via constraints
- **Algorithmic correction**
  - e.g., via modelling causalities

>

Fairness

Algorithmic Fairness

Prevention

- Active inclusion
- Fairness
- Right of understanding
- Access to remedy

> Forum, W. E. (2018) How to Prevent Discriminatory Outcomes in Machine Learning.

> *Active Inclusion*: The development and design of ML applications must actively seek a **diversity of input**, especially of the norms and values of specific populations affected by the output of AI systems.

> *Fairness*: People involved in conceptualizing, developing, and implementing machine learning systems should consider which **definition of fairness** best applies to their context and application, and prioritize it in the architecture of the machine learning system and its evaluation metrics.

> Right to Understanding: Involvement of ML systems in decision-making that affects individual rights must be disclosed, and the systems must be able to **provide an explanation** of their decision-making that is understandable to end users and reviewable by a competent human authority. Where this is impossible and rights are at stake, leaders in the design, deployment and regulation of ML technology must question whether or not it should be used.

> Access to Redress: Leaders, designers and developers of ML systems are responsible for identifying the potential negative human rights impacts of their systems. They must make visible avenues for redress for those affected by disparate impacts, and establish **processes for the timely redress** of any discriminatory outputs.

Fairness

Types of Fairness

- **Individual fairness** 👤
  - Similar individuals treated similarly
- **Group fairness** 👥
  - Similar classifier statistics across groups
    - e.g., statistical parity

> Often assumed to be mutually exclusive.

> Again one has to clarify what similar means, and how to measure.

> Lipschitz continuity implies individual fairness (see smoothness).

> See also: Cisse, M. and Koyejo, S. Fairness and Representation Learning. Tutorial at NeurIPS 2019.

> https://pair-code.github.io/what-if-tool/ai-fairness.html

Fairness

Private and Fair Presentations

Statistical parity For any value that the sensitive attribute takes we will have the same amount predictions for each class.

Error parity For any value that the sensitive attribute takes we will have the same error rates.

Sufficiency For any value that the sensitive attribute takes we will have the same prediction probability.

Connection between fairness ⇄ privacy

It turns out, that the problem of removing confidential information from a dataset or model is equivalent to ensuring statistical parity in algorithmic fairness.

> Recap from the lecture on privacy-preserving!

> Let X be a dataset, Y is the true label, $\hat{Y}$ is our prediction (the representation $Z = \hat{Y}$) and $S \in \{0, 1\}$ is the sensitive (binary) attribute we want to protect.

$$\begin{array}{l|l|r}
\textbf{Statistical parity} & \hat{Y} \perp S & P(\hat{Y} \mid S = 0) = P(\hat{Y} \mid S = 1) \\
\textbf{Error parity} & \hat{Y} \perp S \mid Y & P(\hat{Y} \mid Y = y, S = 0) = P(\hat{Y} \mid Y = y, S = 1) \\
\textbf{Sufficiency} & Y \perp S \mid \hat{Y} & P(Y \mid \hat{Y} = \hat{y}, S = 0) = P(Y \mid \hat{Y} = \hat{y}, S = 1)
\end{array}$$

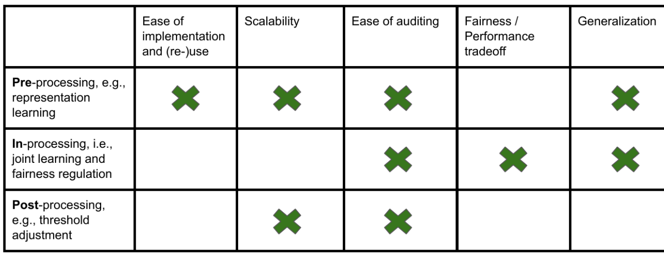
> In statistical parity we want the **same distribution** for different values of a sensitive attribute. This is equivalent to a independence constraint between the algorithm output and the sensitive attribute

Fairness

Achieve Algorithmic Fairness

- **Pre-Processing**
  - e.g., find a metric to measure similarity, representation learing
- **In-Processing**
  - e.g., constrains during learning
- **Post-Processing**
  - e.g., adjustment of models

> From NeurIPS 2019 tutorial:



	Ease of implementation and (re-)use	Scalability	Ease of auditing	Fairness / Performance tradeoff	Generalization
Pre-processing, e.g., representation learning	✖	✖	✖		✖
In-processing, i.e., joint learning and fairness regulation			✖	✖	✖
Post-processing, e.g., threshold adjustment		✖	✖		

Shifts

When does the data/process/... change?

Shifts

Stationary Distribution

- **Common assumption** in machine learning
  - i.e., the data remains stable between training and test
- Might not be realistic **in real-life**
  - Slight changes in the data generation (or collection) process, e.g., change in product portfolio of a retailer

☞ we need to **understand** types of shift, & their causes

> Often the **default route** in data science project:

1. Acquire a (large enough) dataset
1a. Optionally annotated by domain experts (painful exercise)
2. Split the dataset (80/20, or Cross-Validation - optionally also split for validation/hyperparameter optimisation)
3. Train the model (on the training split)
4. Get the evaluation results
4a. Assume this results will hold in the future ϟ

> We still get results from our ML models, even if there is a shift in the data.
> But our results get increasingly “more invalid” 💣.

Shifts

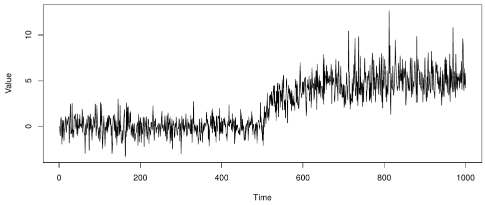
Overview of Shifts

1. Concept drift
2. Covariate shift
3. Prior distribution shift
4. Domain shift

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). Dataset shift in machine learning. The MIT Press.

Concept Drift

- As seen in **time series data**, e.g., $P_{train}(Y \mid X) \neq P_{test}(Y \mid X)$
  - ☞ requires to update the model



> One example from the **time series lecture**.
> It might be that the underlying **generative process changes**, but we do not know from the data alone.
> We just know that the data is not stationary and therefore methods that require stationary data are **inappropriate**.
> There are many approaches for **drift detection** (in time series).

Covariate Shift

- Change in distribution in (one or more) **independent** variables
  - $P_{train}(Y \mid X) = P_{test}(Y \mid X)$
  - $P_{train}(X) \neq P_{test}(X)$
- Problematic, if the training dataset does **not cover** the full $P(Y \mid X)$
  - Consider the relationship b/w X and Y as ReLU
  - ... and during training we only observed $x < 0$
  - → we could false assume $y = 0$, and $X \perp\!\!\!\perp Y$

> Let X be the independent variables and Y the dependent (target) variable to predict.
> See also `https://gsarantitis.wordpress.com/2020/04/16/data-shift-in-machine-learning-what-is-it-and-how-to-detect-it`
> Recall the "car" behind the tree - this was the training data, in testing we only get "samples" from behind the tree - only, when we correctly guessed what is behind the tree, we will be correct.
> Causal interpretation: $X \rightarrow Y$
> For **imbalanced data**, we may introduce a selector V that depends on the target:
> $X \rightarrow Y, Y \rightarrow V$
> If a **sample selection bias** causes the covariate shift, with V being the selector:
> $X \rightarrow Y, X \rightarrow V, Y \rightarrow V$ (without the last causal connection there would be no bias).

Covariate Shift - Countermeasure

- Learn a model
  - To distinguish **between training and test split**
- If the model works sufficiently well
  - i.e., can discriminate samples from training and test split
  - ☞ assume there is a covariate shift

Prior Probability Shift

- Change in distribution in (one or more) **dependent** variables
  - $P_{train}(Y \mid X) = P_{test}(Y \mid X)$
  - $P_{train}(Y) \neq P_{test}(Y)$
- Problematic, if the model uses $P(Y)$
  - e.g., Bayes rule to get $P(Y \mid X)$ via $P(X \mid Y)P(Y)$
  - An assumption made by Naive Bayes

> Can be detection as shift in $P(y)$ via statistical tests, parametric if we know/assume the distribution or non-parametric.
> Causal interpretation: $Y \rightarrow X$ (anticausal)

Shifts

Domain Shift

- (Systematic) change in the distribution
  - Intuition: independent variables depend on (latent) confounder
  - ... change in confounder changes variables
- Example from *Natural Language Processing*
  - Model learnt on text from **newspapers**
  - Model applied on text from **social media**
  - 👉 performance drop

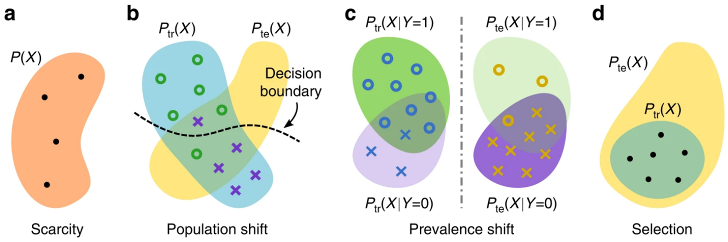
> Causal interpretation: We cannot observe the true cause, x_0, just a transformation $x_{newspaper}$, or $x_{socialmedia}$, ...
> $X_0 \rightarrow Y$, $X_0 \rightarrow X_{dataset}$, the latter relationship may change (e.g., due to some other variables like a confounder).
> The relationship $P(y \mid x_0)$ remains the same.

> The given example could also be seen as shift sample bias.

> Another example from **industrial application**:
> Change in sensors (in slightly different behaviour), typical we observe a slow degradation of sensor (drift, e.g., temperature sensors), which are then rectified via a maintenance (either via calibration, or swap in sensors).

Shifts

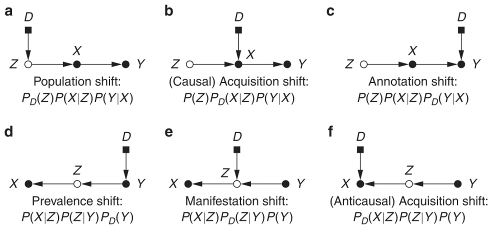
Same Concepts - Different Names



Castro, D.C., Walker, I. and Glocker, B. 2020. Causality matters in medical imaging. Nature Communications. 11, 1 (2020), 1–10. DOI:https://doi.org/10.1038/s41467-020-17478-w.

> In different domains (**medical image analysis**), the same concepts are known under different names.

> *'population shift', 'annotation shift', 'prevalence shift', 'manifestation shift' and 'acquisition shift'*
... correspond to
> *'covariate shift', 'concept shift', 'target shift', 'conditional shift' and 'domain shift'*



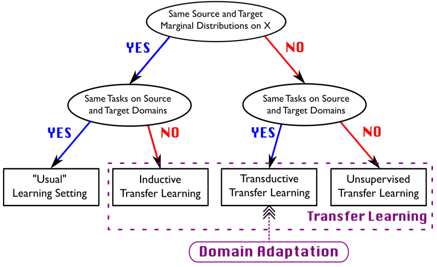
Note: the labels (a-f) do not correspond to the slide

Shifts

Dataset Shift - Countermeasure

- **Transfer learning**
  - Learn on a (large) dataset
  - Adapt to the (potentially small) target dataset
    - Which might be much smaller
  - The original and the target dataset should be related
- **Multitask learning**
  - Machine learning model designed for change in task

> Taken from https://en.wikipedia.org/wiki/Domain_adaptation


Same Source and Target Marginal Distributions on X
YES
NO
Same Tasks on Source and Target Domains
Same Tasks on Source and Target Domains
YES
NO
YES
NO
"Usual" Learning Setting
Inductive Transfer Learning
Transductive Transfer Learning
Unsupervised Transfer Learning
Transfer Learning
Domain Adaptation

> Multitask learning can be seen as special form of transfer learning.

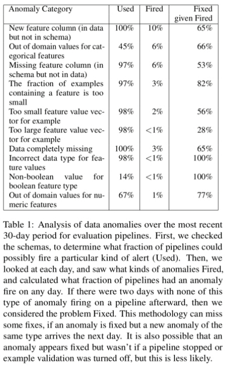
> Additional approach:
> Taking the **causal perspective**, if shift are invariant w.r.t. the

Shifts

Model Unit Testing

- Check if the model **can cope** with the expected nature of the data
  - e.g., via fuzz testing
- Integrate into **continuous integration**

> Taken from: Breck, E., Polyzotis, N., Roy, S., Whang, S.E. and Zinkevich, M. 2019. Data Validation for Machine Learning. SysML. (2019), 1–14.



Anomaly Category	Used	Fired	Fixed given Fired
New feature column (in data but not in schema)	100%	10%	65%
Out of domain values for categorical features	45%	6%	66%
Missing feature column (in schema but not in data)	97%	6%	53%
The fraction of examples containing a feature is too small	97%	3%	82%
Too small feature value vector for example	98%	2%	56%
Too large feature value vector for example	98%	<1%	28%
Data completely missing	100%	3%	65%
Incorrect data type for feature values	98%	<1%	100%
Non-boolean value for boolean feature type	14%	<1%	100%
Out of domain values for numeric features	67%	1%	77%

Table 1: Analysis of data anomalies over the most recent 30-day period for evaluation pipelines. First, we checked the schemas, to determine what fraction of pipelines could possibly fire a particular kind of alert (Used). Then, we looked at each day, and saw what kinds of anomalies Fired, and calculated what fraction of pipelines had an anomaly fire on any day. If there were two days with none of this type of anomaly firing on a pipeline afterward, then we considered the problem Fixed. This methodology can miss some fixes, if an anomaly is fixed but a new anomaly of the same type arrives the next day. It is also possible that an anomaly appears fixed but wasn't if a pipeline stopped or example validation was turned off, but this is less likely.

Shifts

Summary

1. **Understand** the data generation/collection/... process
2. Check for bias in the **data**
3. Make reasonable **assumptions**
4. Select & apply **matching** algorithms
5. Check for bias in the **results**
6. **Celebrate** ☑ (or goto 1 ↻)

> If there is a bias in the data, one would need to recollect (more) data.
> The assumptions of the algorithm should match the problem setting.
> The algorithms introduce new bias, which need to be carefully analysed.

Shifts

The End

Thank you for your attention!

> Hope, that this course (and lecture) provided sufficient insights to allow for elevation outside the danger zone into...