Bias Types

Mittwoch, 9. Juli 2025 21:36

Introduction

Bias
L> preference in favour of or against a person, a group of people, or thing
L> can be conscious or subconscious

Data Bias
L> type of error, systematically skews results in a certain directions

Unbiased Sampling
L> when a sample is representative of the population being measured
L> class of 50 people - you ask 10 randomly - if 10 only women - gender bias

Types Of Data Bias

Sampling Bias
L> when a sample isn't representative of the population as a whole

Observer Bias
L> also called experimenter bias, research bias
L> tendency for different people to observe things differently
L> two scientists looking through a microscope observe different things
L> multiple nurses rounding up BP measures generate inaccurate data

Interpretation Bias
L> tendency to always interpret ambiguous situations in a positive or negative way
L> how two people rate the talking of another person *(calm, angry)*
L> maybe because they have different backgrounds and experiences

Confirmation Bias
L> "*people see what they want to see"*
L> tendency to search for or interpret information in a way that confirms pre-existing beliefs

End

Datei auswählen Keine ausgewählt

Good Data, Bad Data

Mittwoch, 9. Juli 2025 21:36

ROCCC - How to identify good data
R - Reliable *(accurate, complete, unbiased)*
O - Original *(second-, third-party source - make sure it's validated)*
C - Comprehensive *(contain all critical information needed)*
C - Current *(usefulness of data decreases as time passes)*
C - Cited *(Citing makes the information we provide more credible)*

Best data sets to go with:
+ governmental agency data
+ academic papers
+ public data sets
+ financial data

Bad data sources do not ROCCC
R - inaccurate, incomplete unbiased
O - just relying on second- or third-party data without original
C - missing important information to answer question or human error
C - out of date and irrelevant
C - no citing or vetting - that's a no go

Every good solution is found by avoiding bad data.

Datei auswählen Keine ausgewählt

Data Ethics

Mittwoch, 9. Juli 2025 21:36

Ethics
+ well-founded standards of right and wrong
+ prescribe what humans ought to do
+ usually in terms of right, obligations, benefits to society, fairness or specific virtues

Data Ethics
+ well-founded standards of right and wrong
+ dictate how data is collected, shared, and used

GDPR
L> General Data Protection Regulation of the European Union

Aspects of data ethics

Ownership
L> individuals who own the raw data they provide
L> have primary control over it's usage & how processed & shared

Transaction transparency
L> all data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data

Consent
L> individual's right to know explicit details
L> about how and why their data will be used
L> before agreeing and providing it

Currency
L> individuals should be aware of financial transactions
L> resulting from the use of their personal data
L> and the scale of these transactions

Privacy
L> preserving a data subject's information and activity
L> at any time a data transaction occurs
L> sometimes called data privacy or data protection
L> it's all about access, use and collection of data
L> also covers the person's legal right to their data

Openness
L> free access, usage and sharing of data
L> we will discuss details in *"open data"*

Privacy includes..
..protection from unauthorised access to our private data
..freedom from inappropriate use of our data
..the right to inspect, update or correct our data
..ability to give consent to use our data

Datei auswählen Keine ausgewählt

Data anonymisation

Mittwoch, 9. Juli 2025 21:36

Data anonymisation
L> process of protecting people's private or sensitive data
L> by eliminating that kind of information
L> blanking, hashing, masking personal information
L> using fixed-length codes to represent data columns
L> hiding data with altered values

PII = Personally identifiable information
L> information that can be used by itself
L> or with other data to track down a person's identity

Data Analysts are rarely responsible for the anonymisation of data

2 most sensitive types of data
L> healthcare data
L> financial data

De-identification
L> a process used to wipe data clean of all personally identifying infos

A list of data that is often anonymised
+ Telephone numbers
+ Names
+ License plates and license numbers
+ Social security numbers
+ IP addresses
+ Medical records
+ Email addresses
+ Photograph
+ Account numbers

Datei auswählen Keine ausgewählt

Open Data

Mittwoch, 9. Juli 2025 21:36

Features of Open Data
L> easy accessible format
L> must be available as a whole
L> re-use and re-distribution
L> universal participation

A whole lot of resources are needed
to make the technological shift to open data

Data Interoperability
L> ability of data systems & services to openly connect and share data
L> that's why doctor

Biggest Benefit - credible databases can be used more widely

Resources for open data

U.S. government data site: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.

U.S. Census Bureau: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.

Open Data Network: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.

Google Cloud Public Datasets: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into Big Query.

Dataset Search: The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.

Datei auswählen Keine ausgewählt

Glossary

Mittwoch, 9. Juli 2025 21:36

Terms and definitions for Course 3, Module 2

Bad data source: A data source that is not reliable, original, comprehensive, current, and cited (ROCCC)

Bias: A conscious or subconscious preference in favour of or against a person, group of people, or thing

Confirmation bias: The tendency to search for or interpret information in a way that confirms pre-existing beliefs

Consent: The aspect of data ethics that presumes an individual's right to know how and why their personal data will be used before agreeing to provide it

Cookie: A small file stored on a computer that contains information about its users

Currency: The aspect of data ethics that presumes individuals should be aware of financial transactions resulting from the use of their personal data and the scale of those transactions

Data anonymization: The process of protecting people's private or sensitive data by eliminating identifying information

Data bias: When a preference in favour of or against a person, group of people, or thing systematically skews data analysis results in a certain direction

Data ethics: Well-founded standards of right and wrong that dictate how data is collected, shared, and used

Data interoperability: A key factor leading to the successful use of open data among companies and governments

Data privacy: Preserving a data subject's information any time a data transaction occurs

Ethics: Well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues

Experimenter bias: The tendency for different people to observe things differently (also called observer bias)

Fairness: A quality of data analysis that does not create or reinforce bias

First-party data: Data collected by an individual or group using their own resources

General Data Protection Regulation of the European Union (GDPR): Policy-making body in the European Union created to help protect people and their data

Good data source: A data source that is reliable, original, comprehensive, current, and cited (ROCCC)

Interpretation bias: The tendency to interpret ambiguous situations in a positive or negative way

Observer bias: The tendency for different people to observe things differently (also called experimenter bias)

Open data: Data that is available to the public

Openness: The aspect of data ethics that promotes the free access, usage, and sharing of data

Sampling bias: Overrepresenting or underrepresenting certain members of a population as a result of working with a sample that is not representative of the population as a whole

Transaction transparency: The aspect of data ethics that presumes all data-processing activities and algorithms should be explainable and understood by the individual who provides the data

Unbiased sampling: When the sample of the population being measured is representative of the population as a whole

Datei auswählen Keine ausgewählt