KDDM1 - Data Collection

Roman Kern <rkern@tugraz.at>

Version 2.2.0

KDDM1 - Data Collection
Outline

Introduction

Why is data collection important?

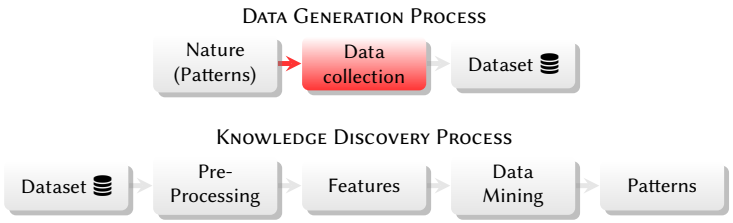
Introduction
Motivation

- To discovery patterns in data
  - ... one needs data
- ➔ Intuitively, the quality and quantity of the data will influence the usefulness

> **Motivation**:
Working with data is a key competence, which in today's work is of high importance.
> **Goal**:
This course aims at providing the key elements of working with data and extracting valuable information out of it.

Introduction

Data Processing Pipeline



> The upper row represents an exemplary data generation process.

> One can only loose information (more on this in an upcoming lecture covering the **data processing inequality**).
> Take away: Data (or information) relevant for the task, which is **not collected**, is lost and cannot be recover later in the process.

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

Introduction

Importance of Dataset Collection

- Depending on the way the data has been collected
  - ... varying data mining approaches are suitable
- Best known example
  - **Observational data** does NOT (in general) allow to derive causal statements

> i.e., **correlation does not imply causation**.

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

Introduction

Data Set Collection and Data Generation Process

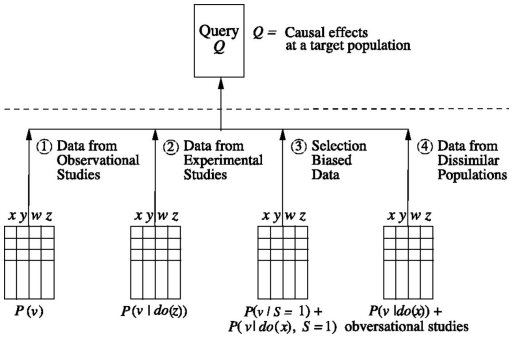
- Paper "Causal inference and the data-fusion problem" [1]
  - Consider causal inference i.e., estimate causal effects
  - Depending on the data generation process
    - The underlying causal relationships
    - ... and the way the data has been collected
  - This might be possible or not
    - In case it is possible, the procedure looks differently

> Paper is accessible online: https://www.pnas.org/doi/full/10.1073/pnas.1510507113

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

Introduction

Data Set Collection and Data Generation Process



> The image serves as illustration of four scenarios in which causal inference (= query) may look different (or might not be possible).

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

Introduction
Types of Data Set Collection I

Observational data

- Passively collected data
  - e.g., machine data, sensors, questionnaires
- The data collection process is not expected to interfere with the data generation process

Introduction
Types of Data Set Collection II

Interventional data

- Data collected while making changes
  - e.g., test, if a drug is working
- The data will reflect the inference in the data generation process

> Recall from the first lecture, where John Snow removed the handle to a water pump and then observed the change in death counts.

> Related, the changes may happen without us knowing.

Introduction
Types of Data Set Collection III

Biased data

- Some influencing factor on the data collection process
  - For example, **sampling bias**
    - Instead of a representative sample from the population
    - ... a biased sample is collected
- Problems will arise especially, once the sampling bias changes

Introduction
Types of Data Set Collection IV

Domain shift

- When combining multiple datasets
  - The conditions they were collected might differ
  - e.g., Demographic data from multiple countries
- Thus, we are not allowed to simply join these datasets together 💣

Types of Data Sets

What type of data are we collecting?

Types of Data Sets
Types of Data Sets

- The first stage is the understanding
  - What type of data are we about to collect
- Practical and theoretical considerations
  - First, we look at the level of structure
  - Next, we consider a couple of common dataset structures

> Practical considerations are e.g. available technologies (databases, file formats, etc.)

> Theoretical considerations are assumptions on the data (e.g., are there relations in the data)

Types of Data Sets
Structure of Data

- First we can distinguish between
  1. Structured data
     - Description (e.g., schema) available, e.g., numbers
     - e.g., Databases
  2. Semi-structured data
     - Mixture of well described data and unstructured data
     - e.g., JSON, spreadsheets, ...
  3. Unstructured data
     - e.g., Text (reports, posts), log files, slides

> Every dataset needs some form of preprocessing.
> Even structured datasets often require feature normalisation, etc.

> It is estimated that most data (in the real world) falls into the category of unstructured data.

Types of Data Sets
Tabular Data I

- The most common data structure in data science
  - ... is **tabular data**
- Organised like a matrix or spreadsheet

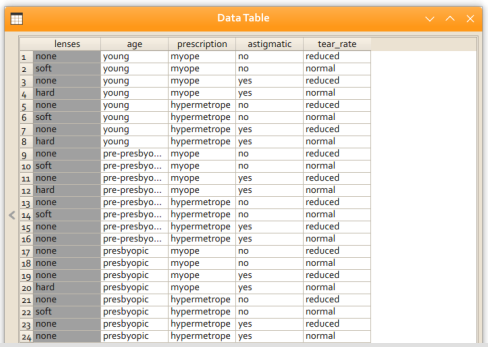
Types of Data Sets

Tabular Data II

- By convention
  - Rows reflect instances, e.g., people
  - Columns reflect features (attributes, variables), e.g., age, income, gender
- Often one column is “special”
  - Target for prediction
    - Target is called the **dependent variable**
    - To be predicted by the **independent variables** (the other columns)

Types of Data Sets

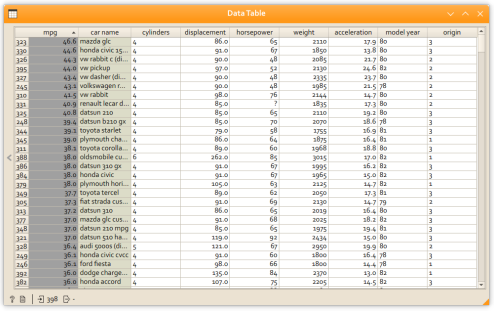
Tabular Data Examples



	lenses	age	prescription	astigmatic	tear_rate
1	none	young	myope	no	reduced
2	soft	young	myope	no	normal
3	none	young	myope	yes	reduced
4	hard	young	myope	yes	normal
5	none	young	hypermetrope	no	reduced
6	soft	young	hypermetrope	no	normal
7	none	young	hypermetrope	yes	reduced
8	hard	young	hypermetrope	yes	normal
9	none	pre-presbyo...	myope	no	reduced
10	soft	pre-presbyo...	myope	no	normal
11	none	pre-presbyo...	myope	yes	reduced
12	hard	pre-presbyo...	myope	yes	normal
13	none	pre-presbyo...	hypermetrope	no	reduced
14	soft	pre-presbyo...	hypermetrope	no	normal
15	none	pre-presbyo...	hypermetrope	yes	reduced
16	none	pre-presbyo...	hypermetrope	yes	normal
17	none	presbyopic	myope	no	reduced
18	none	presbyopic	myope	no	normal
19	none	presbyopic	myope	yes	reduced
20	hard	presbyopic	myope	yes	normal
21	none	presbyopic	hypermetrope	no	reduced
22	soft	presbyopic	hypermetrope	no	normal
23	none	presbyopic	hypermetrope	yes	reduced
24	none	presbyopic	hypermetrope	yes	normal

Types of Data Sets

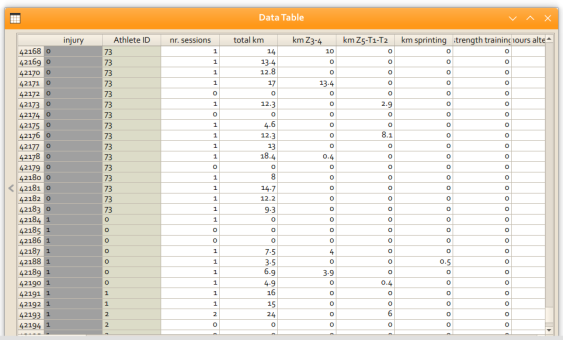
Tabular Data Examples



	mpg	car name	cylinders	displacement	horsepower	weight	acceleration	model year	origin
321	46.6	mazda glc	4	86.0	65	2110	17.9	80	3
330	44.6	honda civic 15...	4	91.0	67	1850	13.8	80	3
326	44.3	vw rabbit c (di...	4	90.0	48	2085	21.7	80	2
395	44.0	vw pickup	4	97.0	52	2130	24.6	82	2
327	43.4	vw dasher (di...	4	90.0	48	2335	23.7	80	2
245	43.1	volkswagen r...	4	90.0	48	1985	21.5	78	2
310	42.5	vw rabbit	4	98.0	76	2144	14.7	80	2
331	41.5	renault lecar d...	4	85.0	7	1835	17.3	80	2
325	40.8	datsun 210	4	85.0	65	2110	19.2	80	3
248	39.4	datsun b210 gx	4	85.0	70	2070	18.6	78	3
344	39.1	toyota starlet	4	79.0	58	1755	16.9	81	3
345	39.0	plymouth cha...	4	86.0	64	1875	16.4	81	1
311	38.1	toyota corolla...	4	89.0	60	1968	18.8	80	3
388	38.0	oldsmobile cu...	6	262.0	85	3015	17.0	82	1
386	38.0	datsun 310 gx	4	91.0	67	1995	16.2	82	3
384	38.0	honda civic	4	91.0	67	1965	15.0	82	3
379	38.0	plymouth hori...	4	105.0	63	2125	14.7	82	1
349	37.7	toyota tercel	4	89.0	62	2050	17.3	81	3
315	37.3	fiat strada cus...	4	91.0	69	2130	14.7	79	2
333	37.2	datsun 310	4	86.0	65	2019	16.4	80	3
377	37.0	mazda glc cus...	4	91.0	68	2025	18.2	82	3
348	37.0	datsun 210 mpg	4	85.0	65	1975	19.4	81	3
321	37.0	datsun 510 ha...	4	119.0	92	2434	15.0	80	3
328	36.4	audi 5000s (di...	5	121.0	67	2950	19.9	80	2
249	36.1	honda civic cvcc	4	91.0	60	1800	16.4	78	3
246	36.1	ford fiesta	4	98.0	66	1800	14.4	78	1
392	36.0	dodge charge...	4	135.0	84	2370	13.0	82	1
382	36.0	honda accord	4	107.0	75	2205	14.5	82	3

Types of Data Sets

Tabular Data Examples



	injury	Athlete ID	nr. sessions	total km	km Z3-4	km Z5-T1-T2	km sprinting	strength training	ours alte
42168	0	73	1	14	10	0	0	0	
42169	0	73	1	13.4	0	0	0	0	
42170	0	73	1	12.8	0	0	0	0	
42171	0	73	1	17	13.4	0	0	0	
42172	0	73	0	0	0	0	0	0	
42173	0	73	1	12.3	0	2.9	0	0	
42174	0	73	0	0	0	0	0	0	
42175	0	73	1	4.6	0	0	0	0	
42176	0	73	1	12.3	0	8.1	0	0	
42177	0	73	1	13	0	0	0	0	
42178	0	73	1	18.4	0.4	0	0	0	
42179	0	73	0	0	0	0	0	0	
42180	0	73	1	8	0	0	0	0	
42181	0	73	1	14.7	0	0	0	0	
42182	0	73	1	12.2	0	0	0	0	
42183	0	73	1	9.3	0	0	0	0	
42184	1	0	1	0	0	0	0	0	
42185	1	0	0	0	0	0	0	0	
42186	1	0	0	0	0	0	0	0	
42187	1	0	1	7.5	4	0	0	0	
42188	1	0	1	3.5	0	0	0.5	0	
42189	1	0	1	6.9	3.9	0	0	0	
42190	1	0	1	4.9	0	0.4	0	0	
42191	1	1	1	16	0	0	0	0	
42192	1	1	1	15	0	0	0	0	
42193	1	2	1	24	0	6	0	0	
42194	1	2	0	0	0	0	0	0	

> Here all variables are categorical.
> The first column (type of contact lens) is to be predicted.
> Making this a **classification problem**.

> Source of the dataset: https://archive.ics.uci.edu/ml/datasets/lenses

> Here the target is numeric, making this a **regression problem**.

> Source of the dataset: https://archive.ics.uci.edu/ml/datasets/auto+mpg

> Here “injury” is the target variable.
> And “Athlete Id” is a special variable, marked as meta (i.e., not helpful for prediction)

> Dataset taken from Kaggle: https://www.kaggle.com/code/mohamedbakrey/anticipate-injured-numbers-in-competitive-runners
> Proposed in paper [2].

> This dataset does not follow the I.I.D. assumption, as the same athlete is in the dataset in multiple rows (and since the past behaviour of an athlete may determine the future, there is a dependency).

Types of Data Sets
Properties of Tabular Data

- Many tabular datasets follow the **I.I.D. assumption**
  - Independence assumption
    - Instances are independent
    - i.e., no systematic dependency/correlation between rows
  - Identically distributed assumption
    - All rows are drawn from the same distribution
    - i.e., no change in distribution/behaviour between rows
- Many data science and machine learning method require IIDness

> Actually, there are two assumptions.

Types of Data Sets
Practical Aspects of Tabular Data

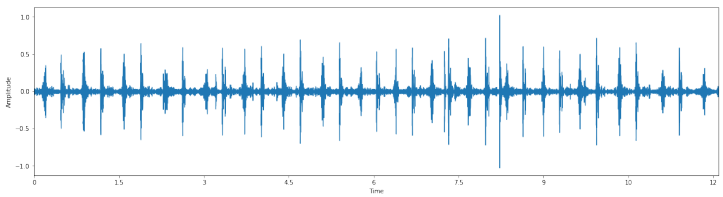
- Typically tabular data is stored as spreadsheets, or `.csv` files
  - The feature type (number, category, string, ...) is unknown
  - Encoding of the file might be unknown
  - Separator char, and escape characters might be unknown

> The CSV file is the lingua franca of data science.

Types of Data Sets
Time Series Data I

- Often we analyse change over time
  - e.g., stock market, health indicators,
- Time being an important variable
  - Often used as x-axis
- Time series data is in general not expected to follow the I.I.D assumption
  - Often considered autoregressive, i.e., the past helps to predict the future

Types of Data Sets
Time Series Data Examples



> Sound is a typical example of time series data.

> Source of the illustration: https://www.kaggle.com/code/osamaheikal/heartbeat-sound-lstm-classification-96

Types of Data Sets
Time Series Data Types

- **Univariate time series**
  - If just one variable changes over time
- **Multivariate time series**
  - Multiple variables change over time
  - ... and may even interact with each other

Types of Data Sets
Time Series Data Examples

> Multiple time series for temperature, humidity, wind speed, etc.
> Naturally, one would expect that there are dependencies between these variables.

> Source of the data from the OpenWeatherMap.

Types of Data Sets
Other Data Types
Common data types in data science

- Image data
  - Dependencies between close pixels
- Graph data
  - Network of nodes (vertices) and edges (connections)
  - Dependencies between connected nodes
    - e.g., Because of homophily
- Text data
  - Typical unstructured type of data
- Spatial data

> Next, we ask the question, where can be get some datasets.

Existing Datasets

Reuse what is there

Existing Datasets
Databases

- Often data is available in databases
  - e.g., company internal data
- Dataset collection
  - Extract (and transform) a subset of the data
  - Typically with structured query languages
    - e.g., SQL, Cypher, SPARQL, ...

> Later in the course we will go deeper on what the implications are.

Existing Datasets
External Datasets

- Adding external datasets
  - ... is common practice
  - In fact, for some it is the definition of **big data**
- Special care needs to be applied
  - i.e., simply adding the data is not advisable

> Also see dataset list of [3] for tabular data.

Existing Datasets
Scientific Datasets

- UCI Machine Learning Repository
  - https://archive.ics.uci.edu
- UEA & UCR Time Series Classification Repository
  - http://www.timeseriesclassification.com/index.php
- Zenodo
  - https://zenodo.org/
- Data (journal)
  - https://www.mdpi.com/journal/data

>

Existing Datasets
Public Datasets

- Awesome lists on Github
  - https://github.com/awesomedata/awesome-public-datasets
- Kaggle
  - https://www.kaggle.com/datasets
- Open governmental data
  - https://data.graz.gv.at/daten/liste
  - https://data.europa.eu/en (1,5m+ datasets)

Web Crawling

Collect data directly from the web

Motivation for Web crawling

- **Question**: How does a search engine know that all these pages contain the query terms?
- **Answer**: Because all of those pages have been crawled!

Motivation for Web crawling

Use Cases

- General web search engines (e.g. Google, Bing, ...)
- Vertical search engines (e.g. Yelp)
- Business Intelligence
- Online Reputation Management
- Data set generation

Names for Web crawling

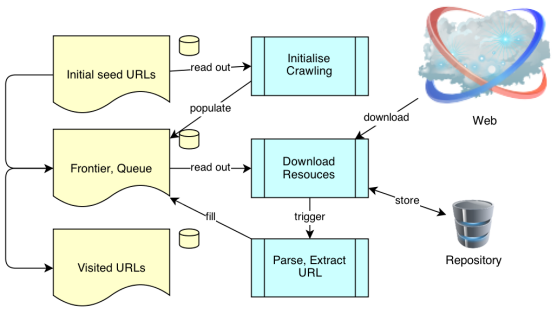
- A **web crawler** is a specialised Web client
- ... that uses the HTTP protocol
- There are many different names for Web crawlers:
  - Crawler, Spider, Robot (or bot), Web agent, Web scutter,
  - Wanderer, worm, ant, automatic indexer, scraper, ...
- Well known instances: googlebot, scooter, slurp, msnbot, ...
- Many libraries, e.g. Heretrix, scrapy
  - See also: `https://github.com/BruceDone/awesome-crawler`

Simple Web crawling schema

Basic Idea

- The crawler starts at an **initial web page**
- The web page is downloaded and its content gets analysed
- ... typically the web page will be HTML
- The **links** within the web page are being **extracted**
- All the links are candidates for the next web pages to be crawled

Simple Web crawling schema



Types of crawlers

- **Batch crawler** - snapshot of the current state, typically until a certain threshold is reached
- **Incremental crawler** - revisiting URLs to keep up to date
- **Specialised crawlers**, e.g focused crawler, topical crawler

Challenges of web crawling

- The **large volume** of the Web
- The **volatility** of the Web, e.g. many Web pages change frequently
- Dynamic Web pages, which are “rendered” in the client
- ... including **dynamically generated** URLs

Challenges of web crawling (cont.)

- Avoid crawling the **same resources** multiple times, e.g. normalise/canonicalise URLs
- Cope with **errors** in downloading, e.g. slow, unreliable connections
- Detect **redirect loops**
- Memory consumption, e.g. large frontier

Challenges of web crawling (cont.)

- Deal with many **content types**, e.g. HTML, PDF, ...
- **Gracefully parse** invalid content, e.g. missing closing tags in HTML
- Identify the **structure** of Web pages, e.g. main text, navigation, ...

Web crawling

Extract structured information

- Usually the information is embedded in HTML tailored towards being displayed
  - ... but crawlers would prefer to have the data already in a structured way
- → **Semantic Web**, highly structured, little uptake
- → **Microformats**, less structured, but more uptake

Web crawling & semantic web

- The "**Semantic Web**" should aid the process of Web crawling
- As it is targeted at making the Web **machine readable**
- Web pages expose their content typically as RDF (Resource Description Language)
  - ... instead of the human readable HTML, e.g. depending on the User Agent
- → specialised crawlers for the Semantic Web

Microformats

- **Microformats** as a lightweight alternative to the “Semantic Web”
- Embedded as HTML markup
- Supported by the major search engines
- http://microformats.org
- e.g. All 4.1+ billion OpenStreetMap nodes have a geo microformat

Microformats

Example: Taken from openstreetmap.org

```
<div class="geo">
    <a href="/?lat=47.0591997&amp;lon=15.4632963&amp;zoom=18">
        <span class="latitude">47.0591997</span>,
        <span class="longitude">15.4632963</span>
    </a>
</div>
```


Crawling strategies

Two main approaches

- **Breadth first search**
  - Data structure: Queue (FIFO)
  - Keeps shortest path to start
- **Depth first search**
  - Data structure: Stack (LIFO)
  - Quickly moves away from initial start node

Concurrent crawlers

- Run crawlers on **multiple machines** in parallel
- Even geographically dispersed
- → shared data structures need to be synchronised

Deep Web

- Also called **hidden Web** (in contrast to the surface Web)
- Consider a Web site that contains a form for the user to fill out
  - e.g. a search input box
- The task of the deep crawler is to fill out this box automatically and crawl the result

Topical Crawler

- Application: **On-the-fly crawling** of the Web
- Starting point: small set of seed pages
- Crawler tries to find similar pages
- Seed pages are used as reference

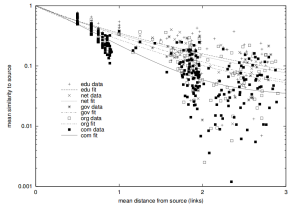
Focused Crawler

- Application: collect pages with **specific properties**, e.g. thematic, type
- For example: find all Blogs that talk about football
  - ... where Blog is the type and football is the topic
- Predict how well the pages in the frontier match the criteria
- Typically uses a manually assembled training data set
  - $\rightarrow$ classification

- Cues to predict relevant pages
  1. **Lexical**, e.g. the textual content of a page
  2. **Link topology**, e.g. the structure of the hyperlinks
- Cluster hypothesis: pages lexically (or topologically) close to a relevant page is also relevant with high probability.
- Need to address two issues:
  1. Link-content conjecture
  2. Link-cluster conjecture

Link-content conjecture

- Are two pages that link to each other more likely to be lexically similar?

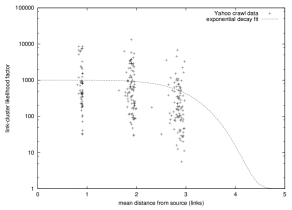


Decay of the cosine similarity as a function of their mean directed link distance

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

Link-cluster conjecture

- Are two pages that link to each other more likely to be semantically related?



Decay in mean likelihood ration as a function of mean directed link distance, starting from Yahoo! directory topics

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

- Examples to measure and compare the performance of crawlers
- **Harvest rate**
  - → Percentage of good pages
- **Search length**
  - → Number of pages to be crawled before a certain percentage of relevant pages are found

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

- **Web information extraction** is the problem of extracting target information item from Web pages
- → Two problems
  1. Extract information from natural language text
  2. Extract structured data from Web pages

Roman Kern <rkern@tugraz.at>, Institute for Interactive Systems and Data Science

Web information extraction

- Web information extraction **via structure**
- Motivation: Often pages on the Web are generated out of databases
- Data records are thereby transformed via **templates** into web pages
  - For example: Amazon product lists & product pages
- Task: Extract the original data record out of the Web page

This task is often called **wrapper generation**.

Wrapper generation

- Three basic approaches for wrapper generation:
  1. Manual - simple approach, but does not scale for many sites
  2. Wrapper induction - supervised approach
  3. Automatic extraction - unsupervised approach

We will have a look at the **wrapper induction**.

Wrapper generation

Wrapper induction

- Needs manually **labelled training examples**
- Learn a classification algorithm
- A simple approach:
  - Web page is represented by a sequence of tokens
  - Idea of landmarks: locate the beginning and end of a target item

Wrapper generation

- Manually labelling is tedious work
- Idea: reduce the amount of work by intelligently **selecting the training examples**
  - → Active Learning approach
- In the case of simple wrapper induction use **co-training**:
  - Search landmarks from the beginning and from the back at the same time
  - Use disagreement as indicator for a training example to annotate

Web crawler and ethics

- Web crawlers **may cause trouble**
  - If too many requests are sent to a single Web site
  - ... it might look like a denial of service (DoS) attack
  - → the source IP will be blacklisted
- Respect the `robots.txt` file (but it's not a legal requirement)
- Some bot disguise themselves and try to replicate a user's behaviour
- Some server disguise themselves, e.g., cloaking (various versions of the same Web page for different clients)

Web crawler and ethics

Example: orf.at/robots.txt

```
# do not index the light version
User-agent: *
Disallow: /l/stories/
Disallow: /full

# these robots have been bad once:

user-agent: stress-agent
Disallow: /

User-agent: fast
Disallow: /
```


Web crawler and legal aspects

- Before crawling a web page
  - Consult the terms and conditions
  - Check if data contains personal (or sensitive) information

Types of Studies

When to conduct which type of study? What data to collect?

Main Types of Studies

- **Interventional study**
  - Able to estimate causal effects
- **Observational study**
  - Estimate correlations (but not causal effects)
  - New insights e.g., candidates for further hypothesis

Controlled Experiment

- A precondition for conducting an experiment is a clear hypothesis (derived from theories)
- *Independent variables* are manipulated to measure their effect
  - ... on one or more *dependent variables*
- Each combination of values of the independent variables is a treatment
  - e.g., applying a method or not (= two groups from a single independent variable)
- We want to measure the effect of a treatment
  - i.e., the cause-effect relationships

Controlled Experiment

- Subjects should be drawn from a well-defined population
- ... with the idea that if it holds for the selected subject,
- ... it also holds for the whole population

Note

Students are not always a representative sample

Controlled Experiment

- Variables other than the chosen independent variables
- ... must not be allowed to affect the experiment
- Between subjects design
  - Split subjects according to variable
  - e.g., smokers vs. non-smokers
  - Assign randomly to treatment groups
- Within subjects design
  - Each subject uses all treatments
  - Downside: learning effect

Types of Studies
Controlled Experiment

Quasi-Experiments

- If a true experiment is not possible
  - Results need to be interpreted carefully
- Subject cannot be randomly assigned to the groups
- The effect can only be measured at discrete time stamps

> Noteworthy quasi experiments are Regression Discontinuity Design, and Difference in Differences.

Types of Studies
Survey Research

- Identify characteristics via a large sample of the population
- Typically questionnaires
- More recently, crowd sourcing

Types of Studies
Survey Research

- Selection of units of analysis is crucial
  - Does not need to be people
- Random sampling might introduce bias in certain populations
- ... stratification might be needed

Types of Studies
Survey Research

- **Cross-sectional survey**
  - Single snapshot at a given time
- **Case-controlled design**
  - Collect multiple variables
  - Study correlations between variables across population
- **Cohort study**
  - Changes over time in a sub-population (group)
  - Form of **longitudinal studies** (in contrast to cross-sectional studies)

Prospective studies are studies running over a long period of time

Types of Studies

Survey Research

- **Guideline for Questions**
  1. Avoid leading questions
  2. Avoid questions that invite the social desirability bias
  3. Avoid double-barreled questions
  4. Avoid long questions
  5. Avoid negations
  6. Avoid irrelevant questions
  7. Avoid poorly worded response options
  8. Avoid big words
  9. Avoid ambiguous words & phrases

> See more information: https://www.slideshare.net/rsmehta/3-types-of-research-study

Synthetic Datasets

Generate your own datasets

Synthetic Datasets

Why create own datasets?

- Real-world datasets may not be available
  - e.g., not ethical, or legal
- Test out data mining methods and approaches
  - Isolated changes
- Assess the impact of data set size on results

> For example, datasets that relate to private or sensitive data.

Synthetic Datasets

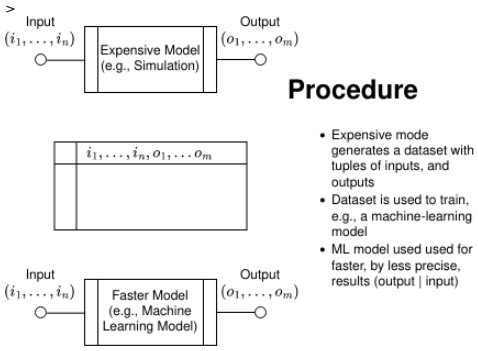
Dataset Augmentation

- Extend an existing dataset
  - By adding additional examples
  - Often by changing existing samples
    - Here the operation needs to be independent of any mechanism used for prediction
- The goal is to **improve the robustness** of methods trained on the augmented dataset

> For example, one wants to classify pictures of cats and dogs.
> Rotating, scaling, etc. will not affect the prediction mechanism (i.e., the way the model is support to predict the type of animal).

Synthetic Datasets
Surrogate Modelling

- Highly sophisticated simulation models produce precise results
  - e.g., finite element method, computational fluid dynamics
- Often these methods are computationally intense
  - i.e., slow
- Idea: Have these simulation models create a dataset
  - And train a machine learning model on it, the so call, **surrogate model**
  - Which is faster, but less precise

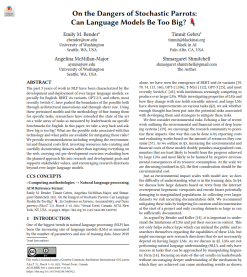
> 

- Expensive mode generates a dataset with tuples of inputs, and outputs
- Dataset is used to train, e.g., a machine-learning model
- ML model used used for faster, by less precise, results (output | input)

> [4]

Synthetic Datasets
Dataset Collection Documentation

- "we rely on ever larger datasets we risk incurring **documentation debt**"
- "putting ourselves in a situation where the datasets are both undocumented and too large to document post hoc"
- "undocumented training data perpetuates harm without recourse"



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Author

ABSTRACT

CCS CONCEPTS

1 INTRODUCTION

Synthetic Datasets
Domain Knowledge Documentation

- Existing expert knowledge
  - e.g., physical constraints
  - Should also be documented
    - e.g., via knowledge graphs
- Data knowledge
  - e.g., height can only be positive
  - e.g., missing values cannot occur

Thank You!

... for your attention

References I

[1] E. Bareinboim and J. Pearl, **Causal inference and the data-fusion problem**, Proceedings of the National Academy of Sciences, vol. 113, no. 27, pp. 7345–7352, 2016.

[2] S. S. Lövdal, R. J. Den Hartigh, and G. Azzopardi, **Injury prediction in competitive runners with machine learning**, International journal of sports physiology and performance, vol. 16, no. 10, pp. 1522–1531, 2021.

[3] R. Singh and S. Bedathur, **Embeddings for tabular data: A survey**, arXiv preprint arXiv:2302.11777, 2023.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, **On the dangers of stochastic parrots: Can language models be too big?** Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.