

Data Science

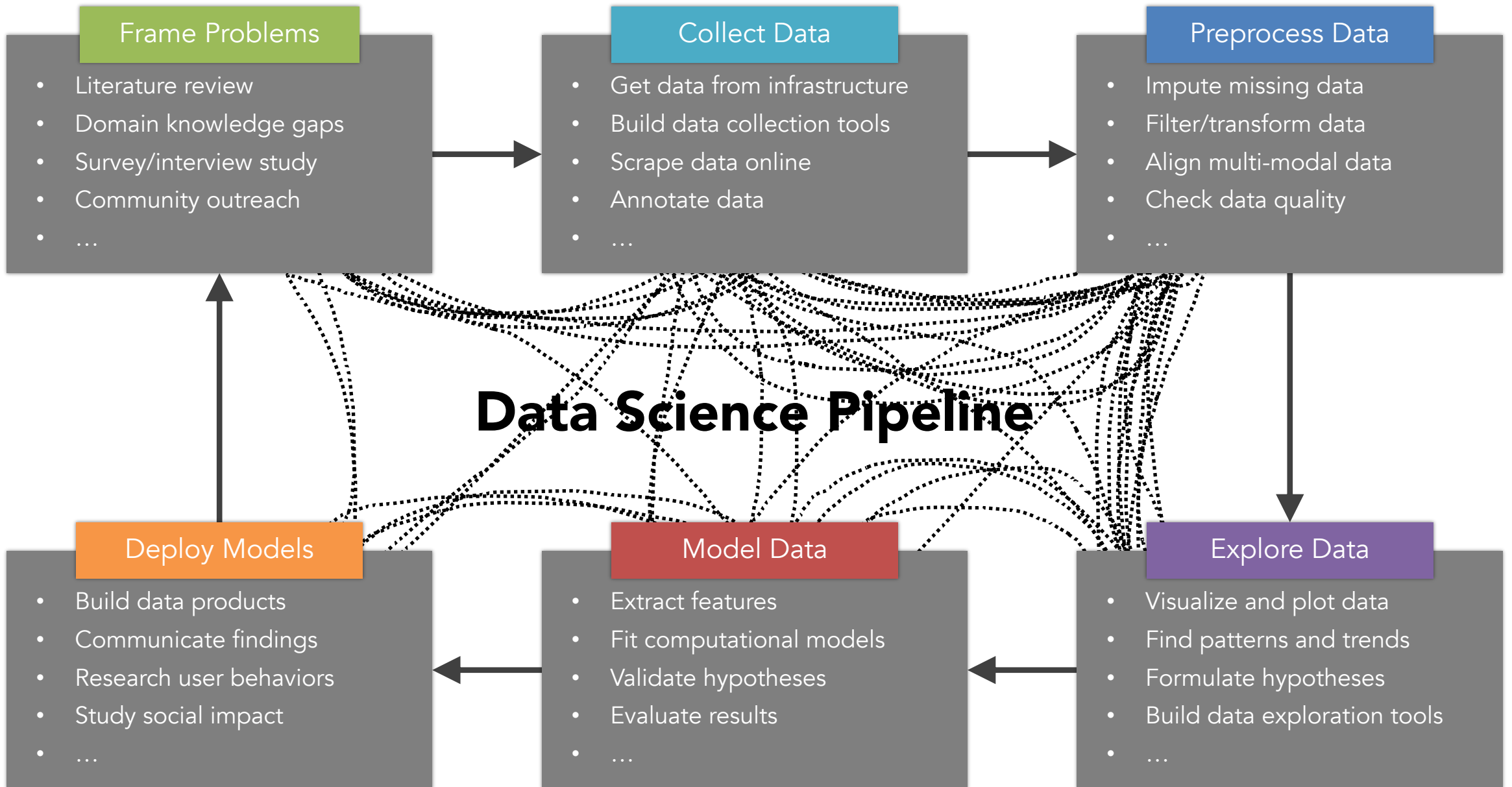
Lecture 2-1: Data Science Fundamentals (Pipeline)



Lecturer: Yen-Chia Hsu

Date: Feb 2024

This lecture shows a typical data science pipeline and recaps data cleaning techniques.



What people typically think : →

The reality of the data science pipeline: ·····

Frame Problems

This course will use existing scenarios and cases with well-defined problems. However, in the real world, we need to define and frame the problems first.

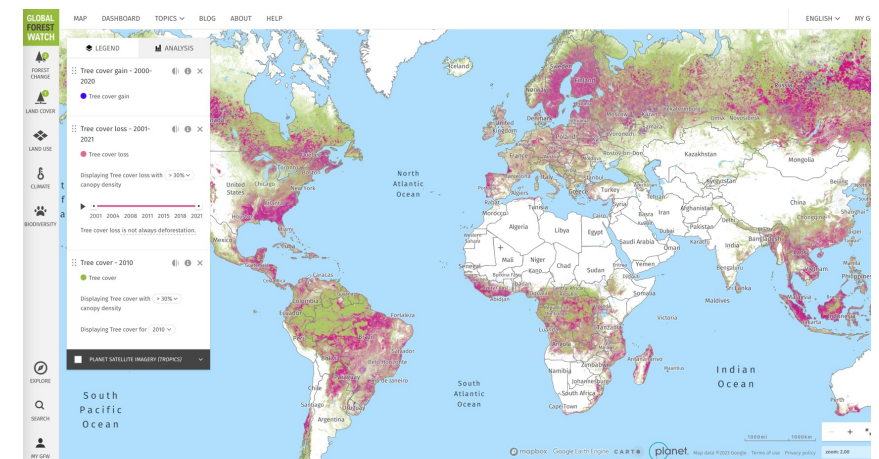
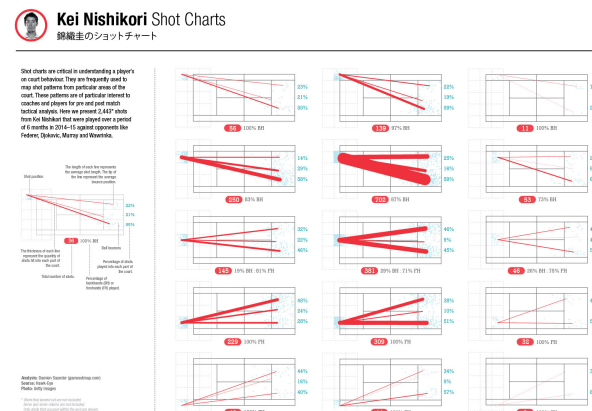
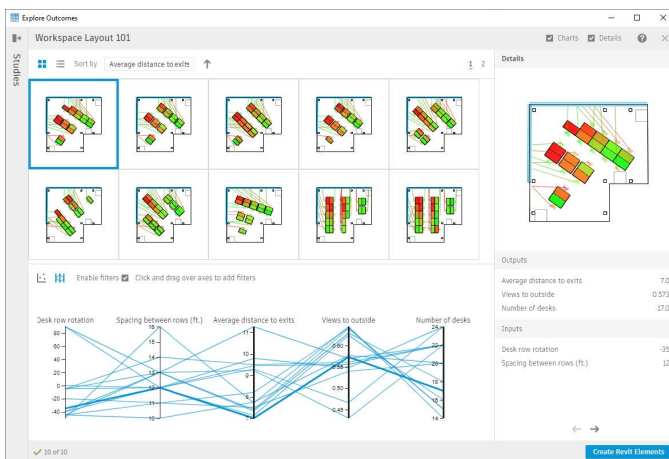
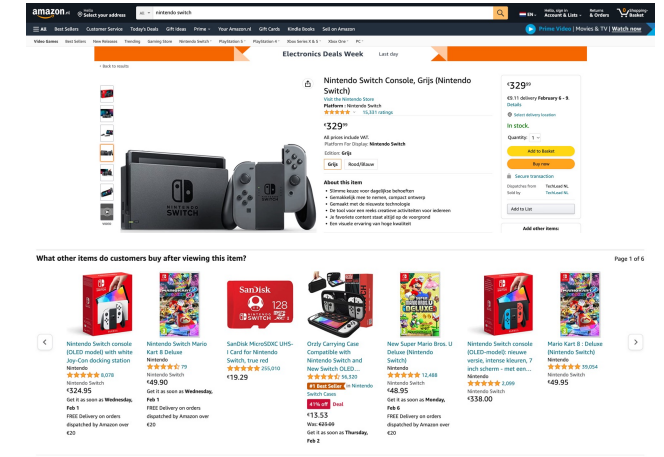
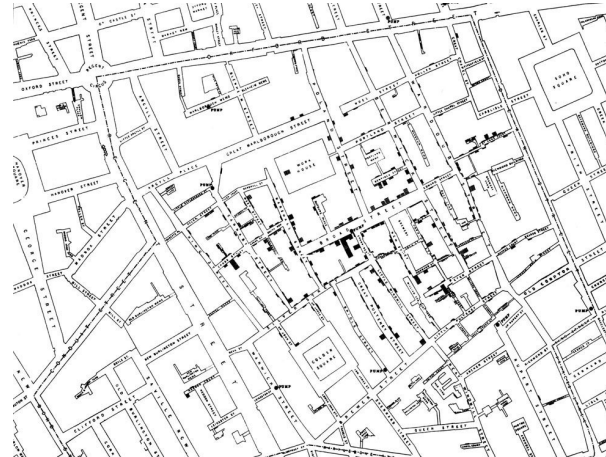


Figure sources are in the slides for the first lecture.

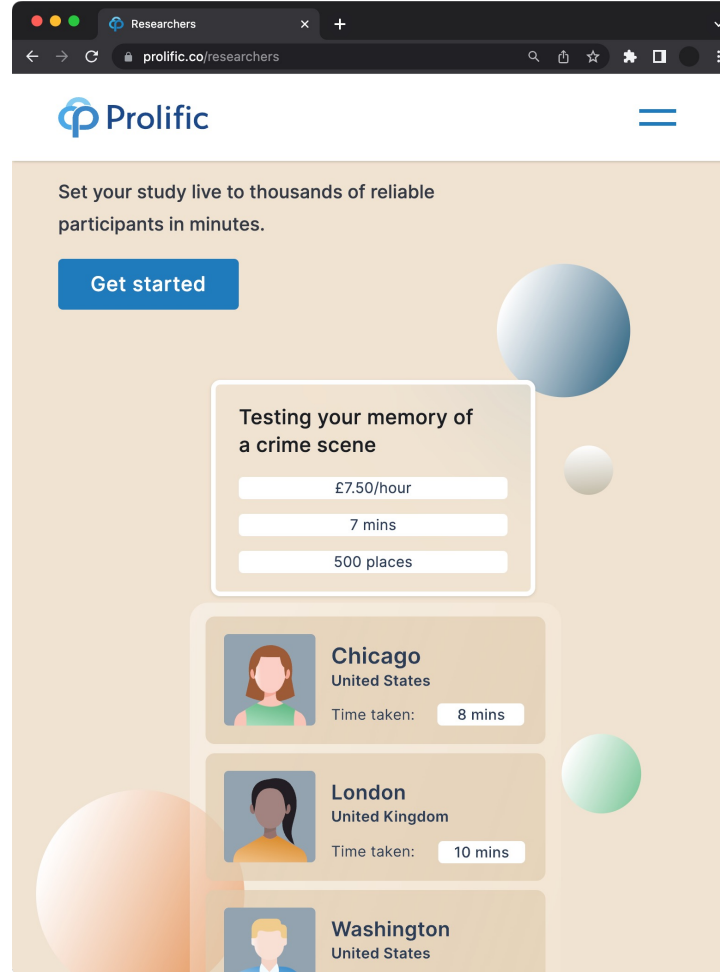
Collect Data

This course will assume that someone already collected the data for you. But we will briefly cover crowdsourcing and data annotation techniques.



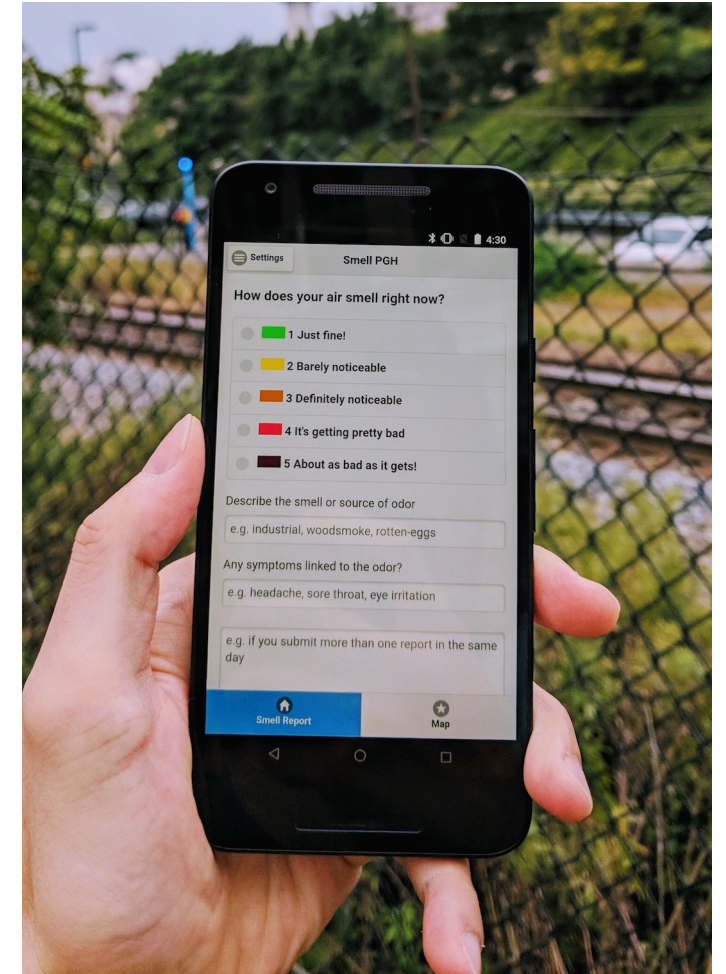
The screenshot shows the GGD Amsterdam Data Portal. The header includes the logo for Gemeente Amsterdam and navigation links for 'Inloggen' and 'Menu'. The main section is titled 'Populaire kaartlagen' (Popular map layers) and features three cards: 'Kadastrale perceelsgrenzen' (Cadastral parcel boundaries), 'Meetbouden - Zakingsnelheid' (Measuring buildings - Building speed), and 'Parkeren - Fiscale indelingen' (Parking - Fiscal divisions). Each card includes a brief description of the data and a 'Toon alle kaartlagen' (Show all map layers) link. The footer contains 'Vragen' (Questions), 'Colofon' (Credits), and 'Volg ons' (Follow us) with social media icons for Twitter, Facebook, LinkedIn, and YouTube.

[GGD Amsterdam Data Portal](https://data.amsterdam.nl)



The screenshot shows the Prolific website. The header includes the Prolific logo and a navigation menu. The main section is titled 'Set your study live to thousands of reliable participants in minutes.' and features a 'Get started' button. Below this, there is a card for 'Testing your memory of a crime scene' with a rate of £7.50/hour, a duration of 7 mins, and 500 places. There are also cards for 'Chicago United States' (8 mins), 'London United Kingdom' (10 mins), and 'Washington United States'.

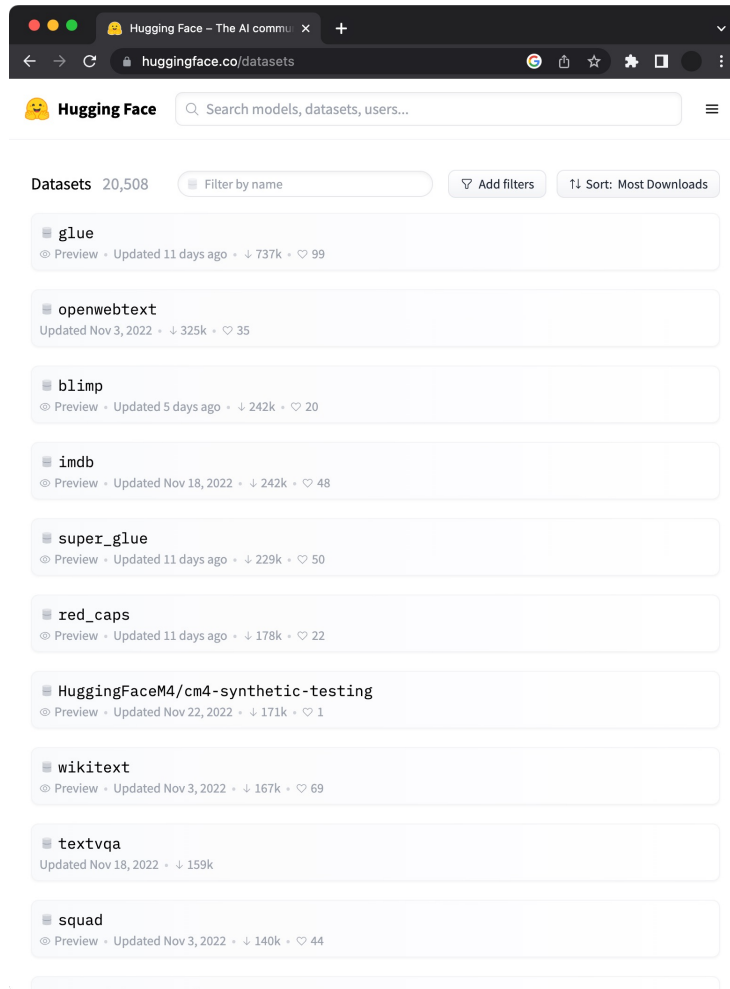
[Prolific Tool for Data Annotation](https://prolific.co)



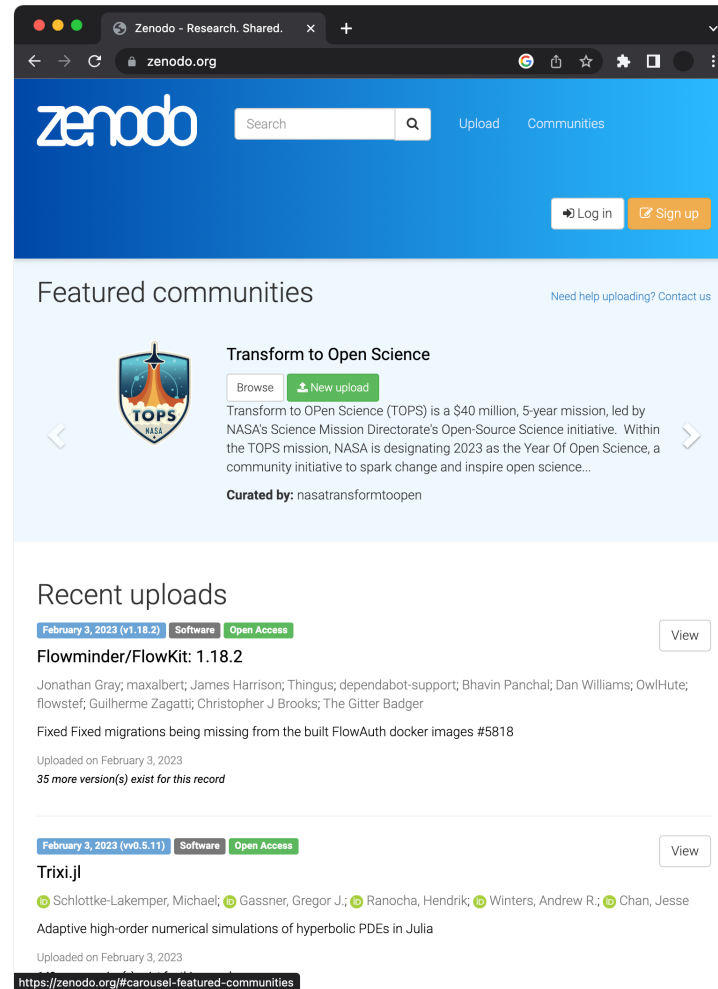
[Mobile App Data Collection](#)

Collect Data

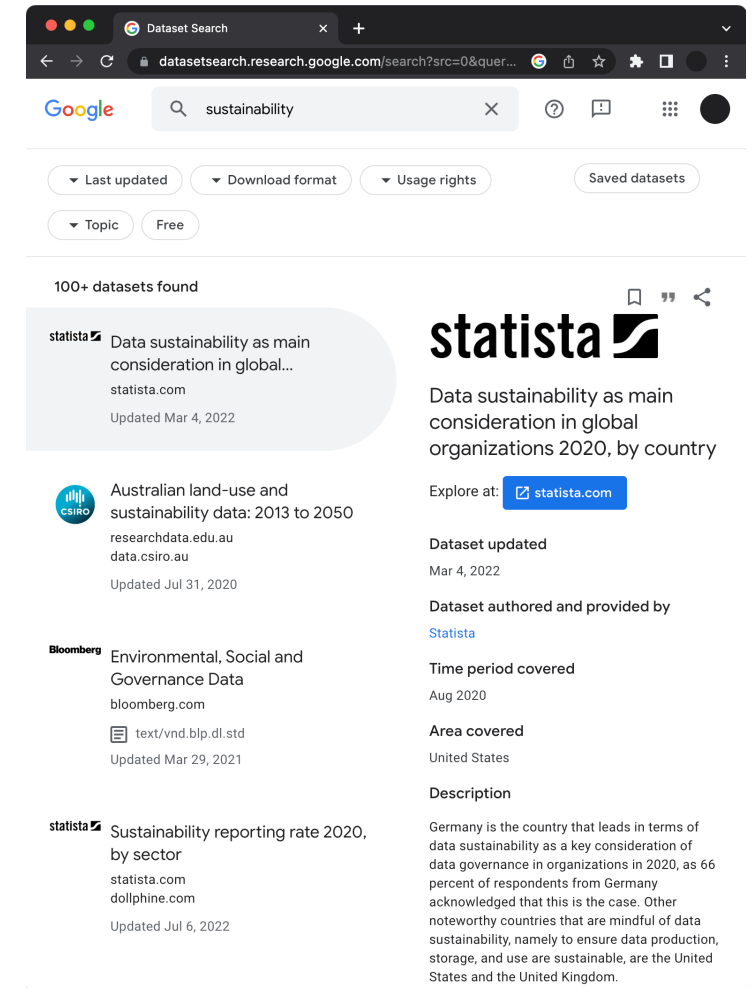
There are also other sources for getting public datasets, such as Hugging Face, Zenodo, Google Dataset Search, government websites, etc.



[Hugging Face](https://huggingface.co/datasets)



[Zenodo](https://zenodo.org)



[Google Dataset Search](https://datasetsearch.research.google.com/search?src=0&quer...)

This course will focus on [pandas](#), which is a very handy Python library for preprocessing structured data. We will cover the following techniques:

- | | |
|--|---|
| <ul style="list-style-type: none">• Filter unwanted data• Aggregate data (e.g., sum)• Group data based on a column• Sort rows based on a column• Concatenate data frames• Merge and join data frames• Quantize continuous values into bins | <ul style="list-style-type: none">• Scale column values• Resample time series data• Apply a transformation function• Use regular expressions• Drop rows or columns• Treat missing values |
|--|---|

Filtering can reduce a set of data based on specific criteria. For example, the left table can be reduced to the right table using a population threshold.

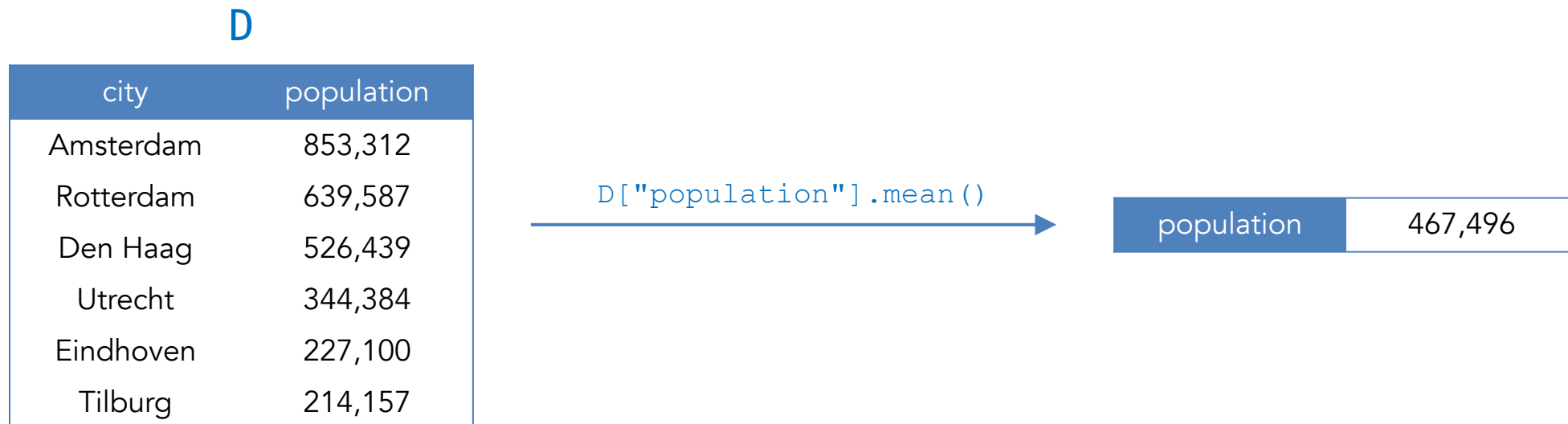
D

city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439
Utrecht	344,384
Eindhoven	227,100
Tilburg	214,157

`D[D["population"]>500000]`

city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439

Aggregation reduces a set of data to a descriptive statistic. For example, the left table is reduced to a single number by computing the mean value.



Grouping divides a table into groups by column values, which can be chained with data aggregation to produce descriptive statistics for each group.

D

city	province	population
Amsterdam	Noord-Holland	853,312
Rotterdam	Zuid-Holland	639,587
Den Haag	Zuid-Holland	526,439
Utrecht	Utrecht	344,384
Eindhoven	Noord-Brabant	227,100
Tilburg	Noord-Brabant	214,157

`D.groupby("province").sum()`



province	population
Noord-Holland	853,312
Zuid-Holland	1,166,026
Utrecht	344,384
Noord-Brabant	441,257

Sorting rearranges data based on values in a column, which can be useful for inspection. For example, the right table is sorted by population.

D

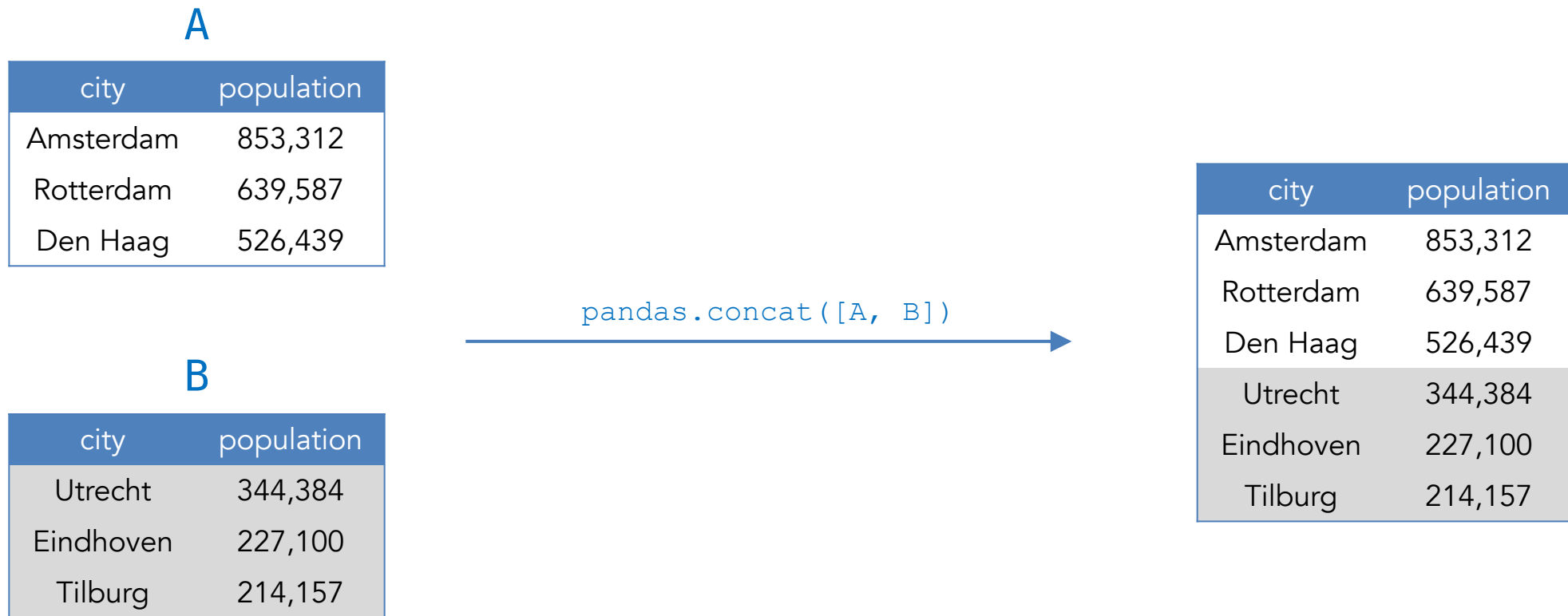
city	population
Eindhoven	227,100
Den Haag	526,439
Tilburg	214,157
Rotterdam	639,587
Amsterdam	853,312
Utrecht	344,384

`D.sort_values(by=["population"])`



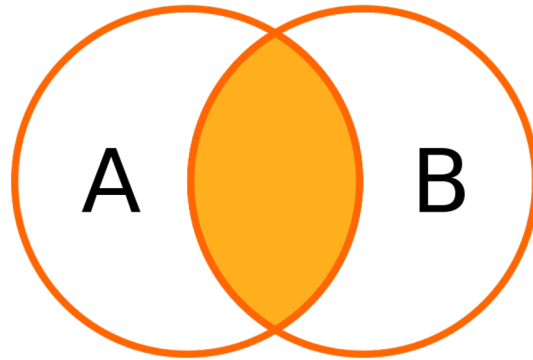
city	population
Tilburg	214,157
Eindhoven	227,100
Utrecht	344,384
Den Haag	526,439
Rotterdam	639,587
Amsterdam	853,312

Concatenation combines multiple datasets that have the same variables. For example, the two left tables can be concatenated into the right table.

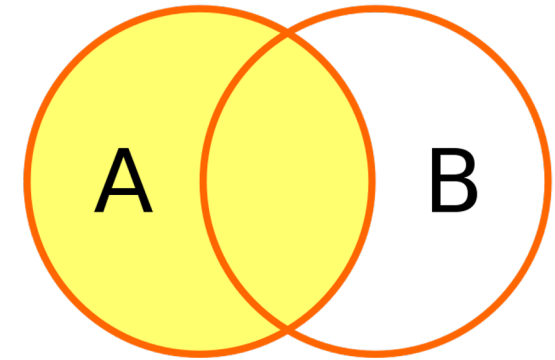


Merging and joining is a common method (in relational databases) to merge multiple data tables which have overlapping set of instances.

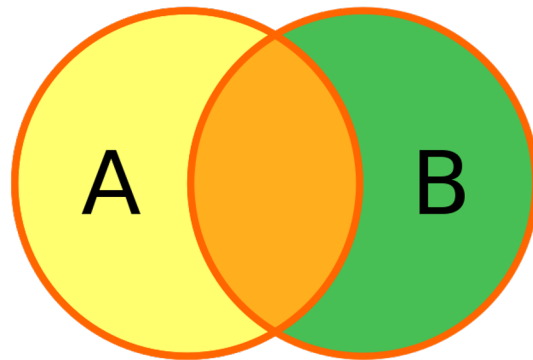
- Inner join



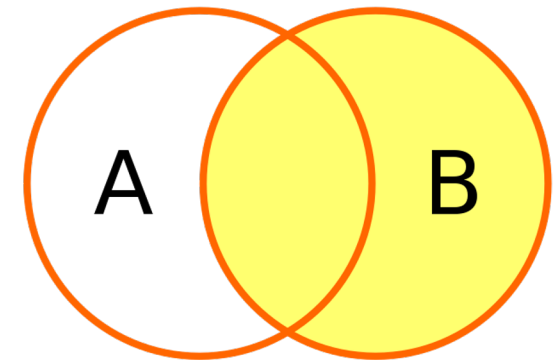
- Left (outer) join



- Outer join



- Right (outer) join



A

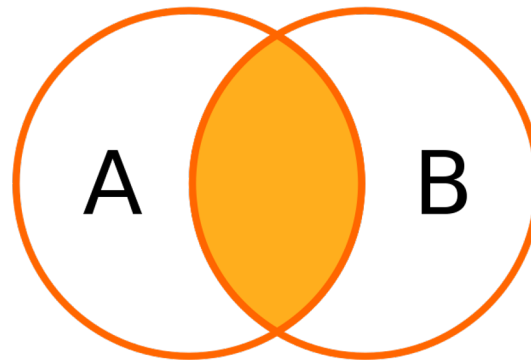
city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439
Utrecht	344,384
Eindhoven	227,100
Tilburg	214,157

B

city	air_quality
Amsterdam	42.4
Rotterdam	40.9
Den Haag	41.1
Utrecht	41.4
Eindhoven	43.8
Zwolle	40.9

Use "city" as the key to merge A and B

```
A.merge(B, how="inner", on="city")
```



- Inner join

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8

A

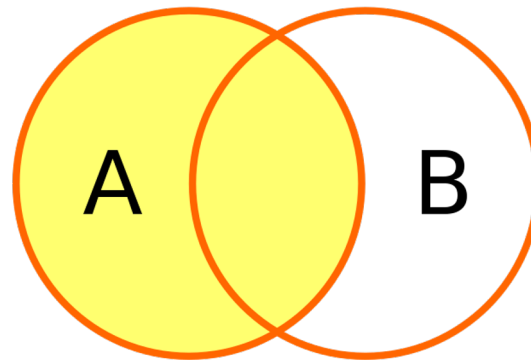
city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439
Utrecht	344,384
Eindhoven	227,100
Tilburg	214,157

B

city	air_quality
Amsterdam	42.4
Rotterdam	40.9
Den Haag	41.1
Utrecht	41.4
Eindhoven	43.8
Zwolle	40.9

Use "city" as the key to merge A and B

`A.merge(B, how="left", on="city")`



- Left join

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	NaN

A

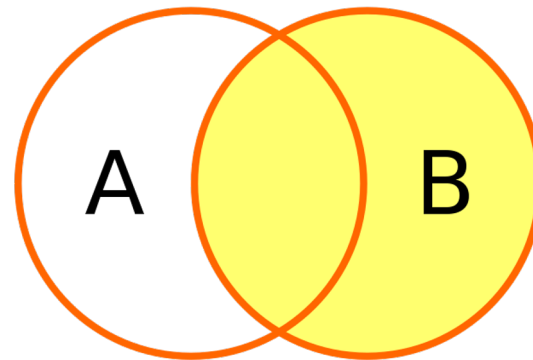
city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439
Utrecht	344,384
Eindhoven	227,100
Tilburg	214,157

B

city	air_quality
Amsterdam	42.4
Rotterdam	40.9
Den Haag	41.1
Utrecht	41.4
Eindhoven	43.8
Zwolle	40.9

Use "city" as the key to merge A and B

`A.merge(B, how="right", on="city")`



- Right join

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Zwolle	NaN	40.9

A

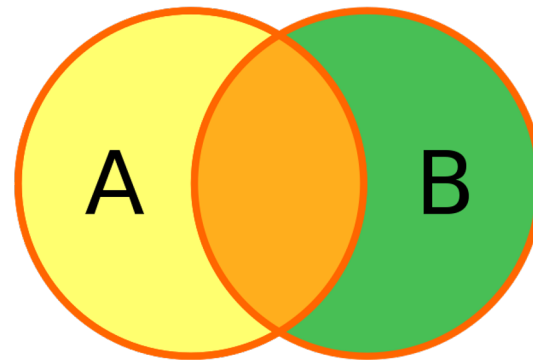
city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439
Utrecht	344,384
Eindhoven	227,100
Tilburg	214,157

B

city	air_quality
Amsterdam	42.4
Rotterdam	40.9
Den Haag	41.1
Utrecht	41.4
Eindhoven	43.8
Zwolle	40.9

Use "city" as the key to merge A and B

`A.merge(B, how="outer", on="city")`



- Outer join

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	NaN
Zwolle	NaN	40.9

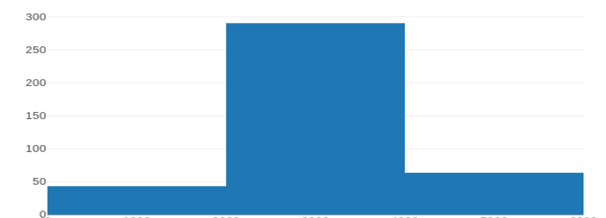
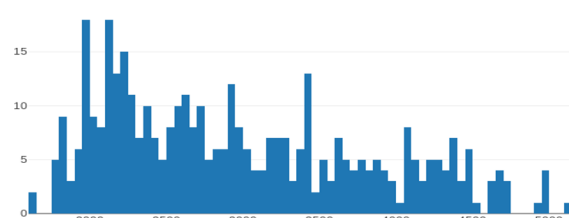
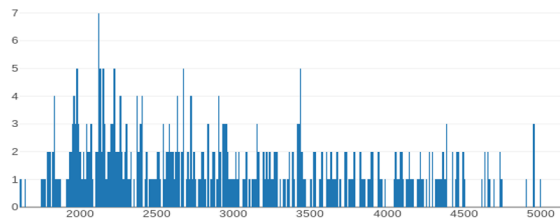
Quantization transforms a continuous set of values (e.g., integers) into a discrete set (e.g., categories). For example, age is quantized to age range.

D

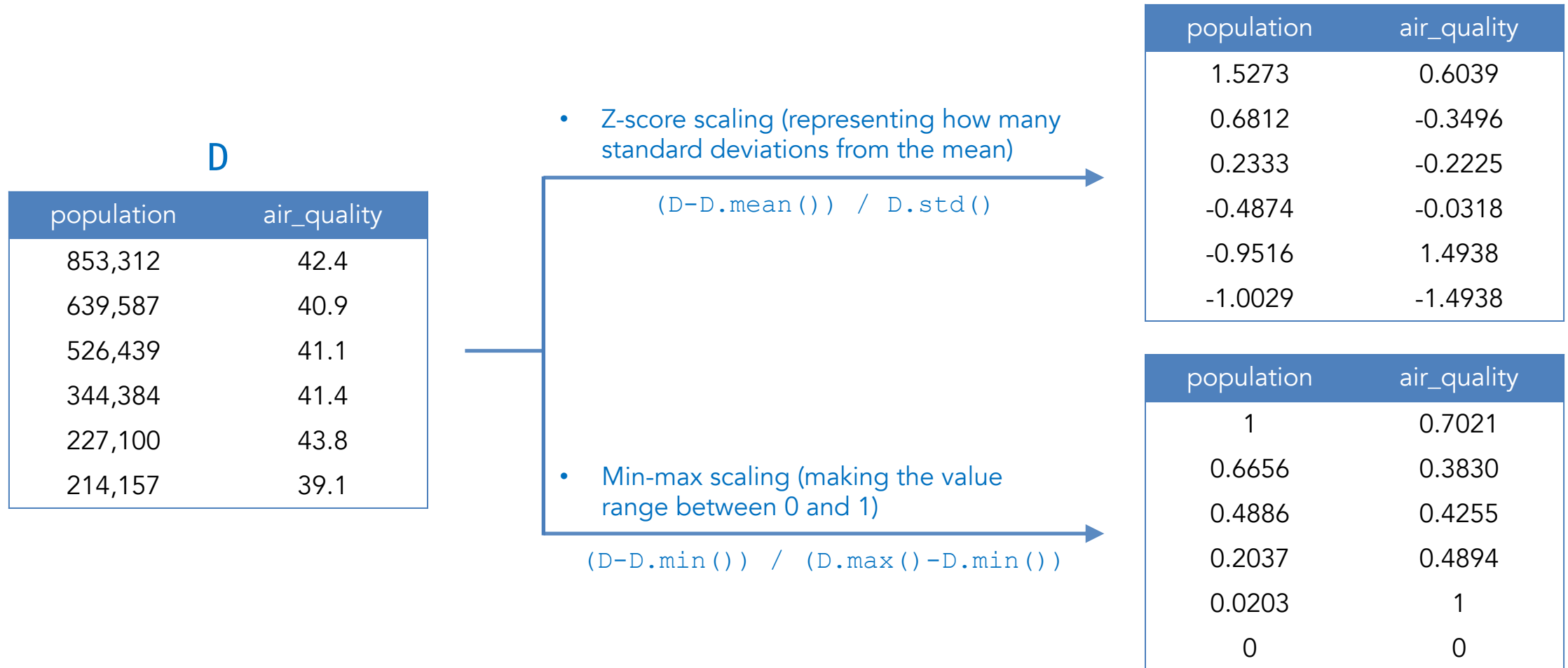
name	age
Jantje	8
Piet	16
Maria	22
Renske	34
Donald	65

```
pandas.cut(D["age"], [0,20,50,200],
           labels=["1-20", "21-50", "51+"])
```

name	age
Jantje	1-20
Piet	1-20
Maria	21-50
Renske	21-50
Donald	51+



Scaling transforms variables to have another distribution, which puts variables at the same scale and makes the data work better on many models.



Preprocess Data

You can **resample** time series data (i.e., the data with time stamps) to a different frequency (e.g., hourly) using different aggregation methods (e.g., mean).

D

timestamp	v1
2016-10-31 07:30:00	52.6
2016-10-31 08:30:00	48.3
2016-10-31 08:53:20	44.2
2016-10-31 09:30:00	31.1

`D.resample("60Min", label="right").mean()`

timestamp	v1
2016-10-31 08:00:00	52.60
2016-10-31 09:00:00	46.25
2016-10-31 10:00:00	31.10

You can apply a **transformation** to rows or columns in the data frame.

D

wind_mph
3.6
NaN
5.1

```
def f(x):
    if pd.isna(x): return None
    else: return x<5
D["is_calm"] = D["wind_mph"].apply(f)
```

wind_mph	is_calm
3.6	True
NaN	None
5.1	False

D

wind_deg
343
351
359
5
41
25
:



Very slow if you have a lot of rows!

```
def f(x):
    return numpy.sin(numpy.deg2rad(x))
D["wind_sine"] = D["wind_deg"].apply(f)
```



```
D["wind_sine"] = np.sin(np.deg2rad(D["wind_deg"]))
```

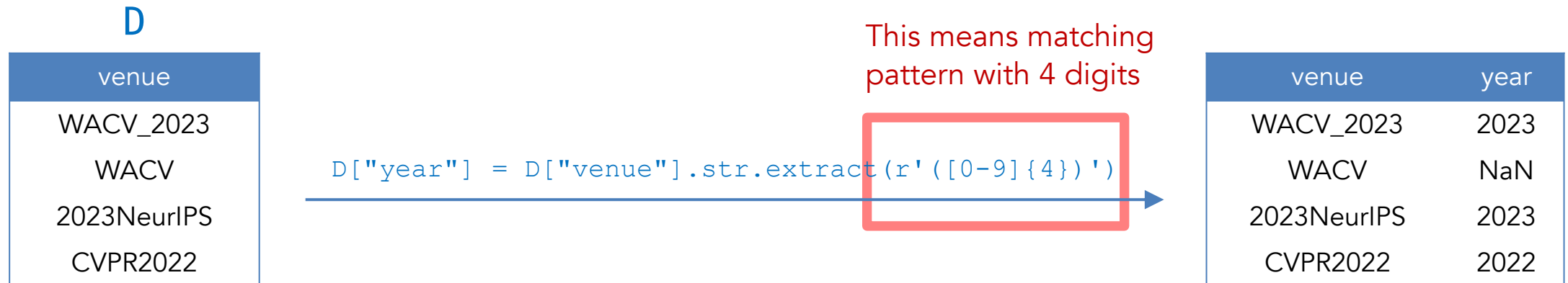


Better to transform the entire column directly!

wind_deg	wind_sine
343	-0.292372
351	-0.156434
359	-0.017452
5	0.087156
41	0.656059
25	0.422618
:	:

Preprocess Data

To extract data from text or match text patterns, you can use **regular expression**, which is a language to specify search patterns.



Preprocess Data

We can **drop** data that we do not need, such as duplicate data records or those that are irrelevant to our research question.

city	population	year
Amsterdam	853,312	2018
Rotterdam	639,587	2018
Den Haag	526,439	2018

`pandas.drop(columns=["year"])`

city	population
Amsterdam	853,312
Rotterdam	639,587
Den Haag	526,439

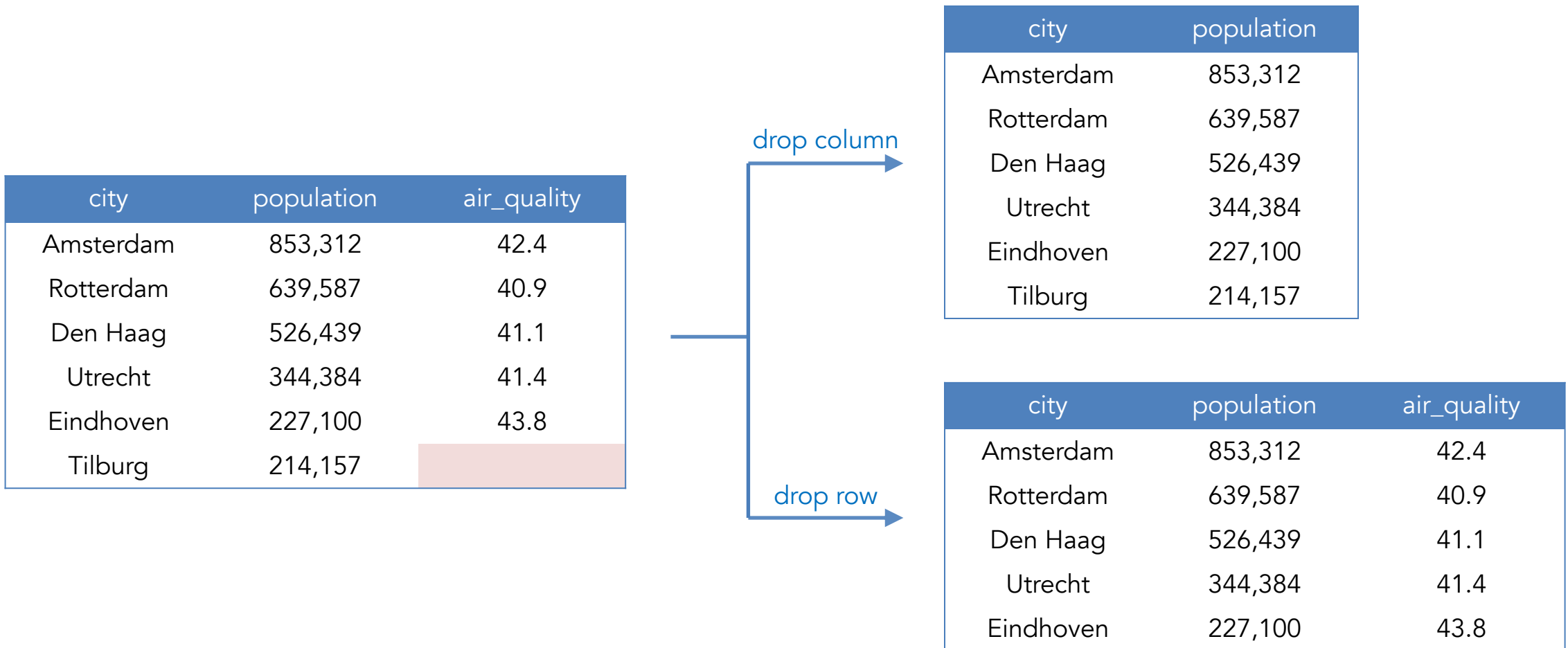
	city	population	year
0	Amsterdam	853,312	2018
1	Rotterdam	639,587	2018
2	Den Haag	526,439	2018
3	Utrecht	344,384	2018
4	Eindhoven	227,100	2018
5	Amsterdam	862,965	2019
6	Utrecht	344,384	2018

`pandas.drop([5, 6])`

	city	population	year
0	Amsterdam	853,312	2018
1	Rotterdam	639,587	2018
2	Den Haag	526,439	2018
3	Utrecht	344,384	2018
4	Eindhoven	227,100	2018

Preprocess Data

We can either **drop the rows** (i.e., the records/observations) or the **columns** (i.e., the variables/attributes) that contain the missing values.



Preprocess Data

We can **replace the missing values** (i.e., imputation) with a constant, mean, median, or the most frequent value along the same column.

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	

constant
imputation

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	-1

mean
imputation

city	population	air_quality
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	41.92

Preprocess Data

We can **model missing values**, where y is the variable/column that has the missing values, X means other variables, and F is a regression function.

city	population (X)	air_quality (y)
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	

$$y = F(X)$$

city	population (X)	air_quality (y)
Amsterdam	853,312	42.4
Rotterdam	639,587	40.9
Den Haag	526,439	41.1
Utrecht	344,384	41.4
Eindhoven	227,100	43.8
Tilburg	214,157	42.46

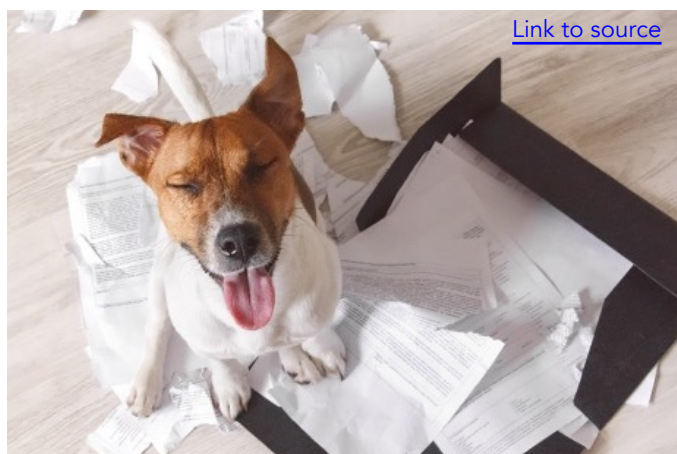
Different missing data may require different data cleaning methods.

Missing Not At Random is a big problem and cannot be solved simply with imputation.

MCAR

Missing Completely At Random:

- Missing data is a completely random subset (no relations) of the entire dataset.

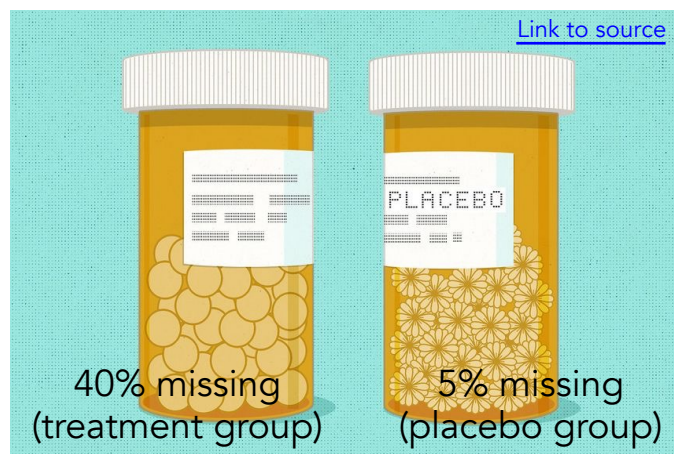


[Link to source](#)

MAR

Missing at Random:

- Missing data is only related to variables other than the one having missing data.

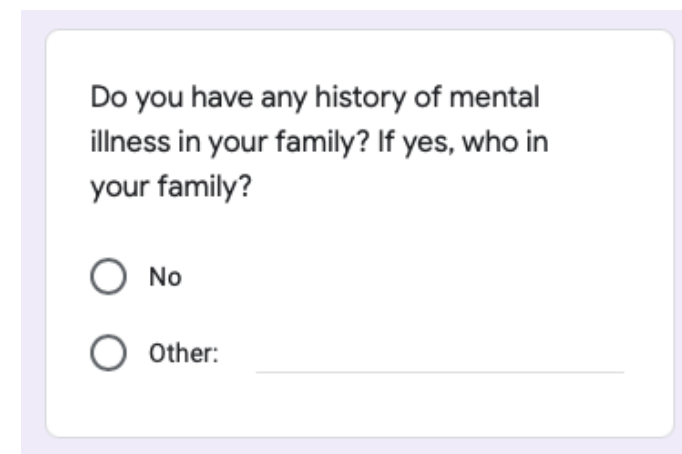


[Link to source](#)

MNAR

Missing Not At Random:

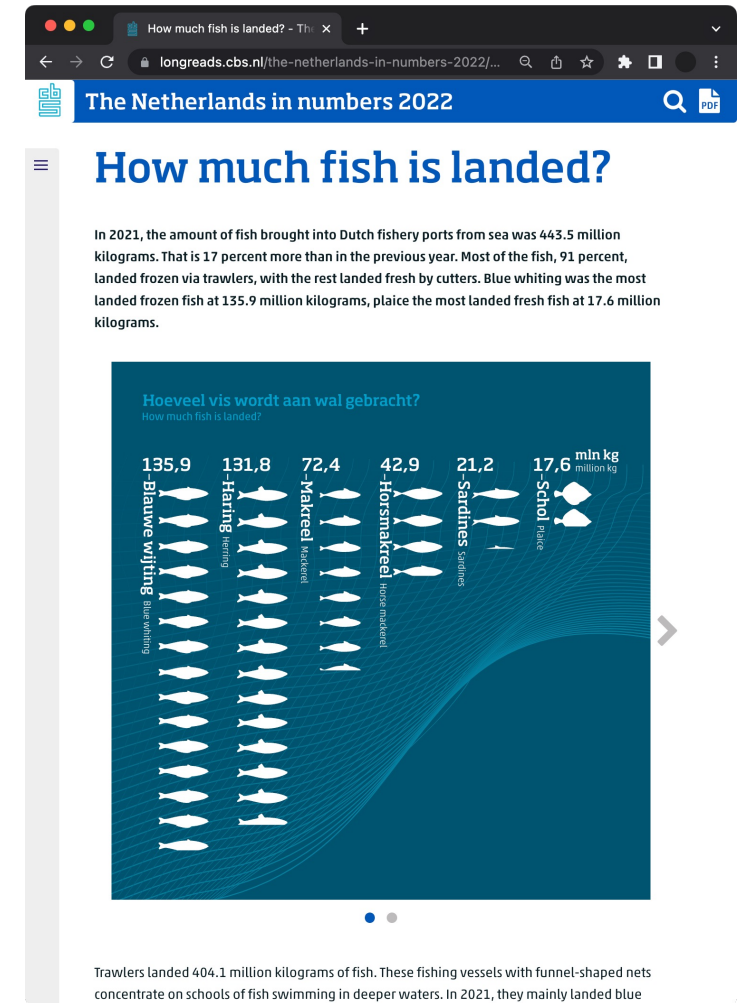
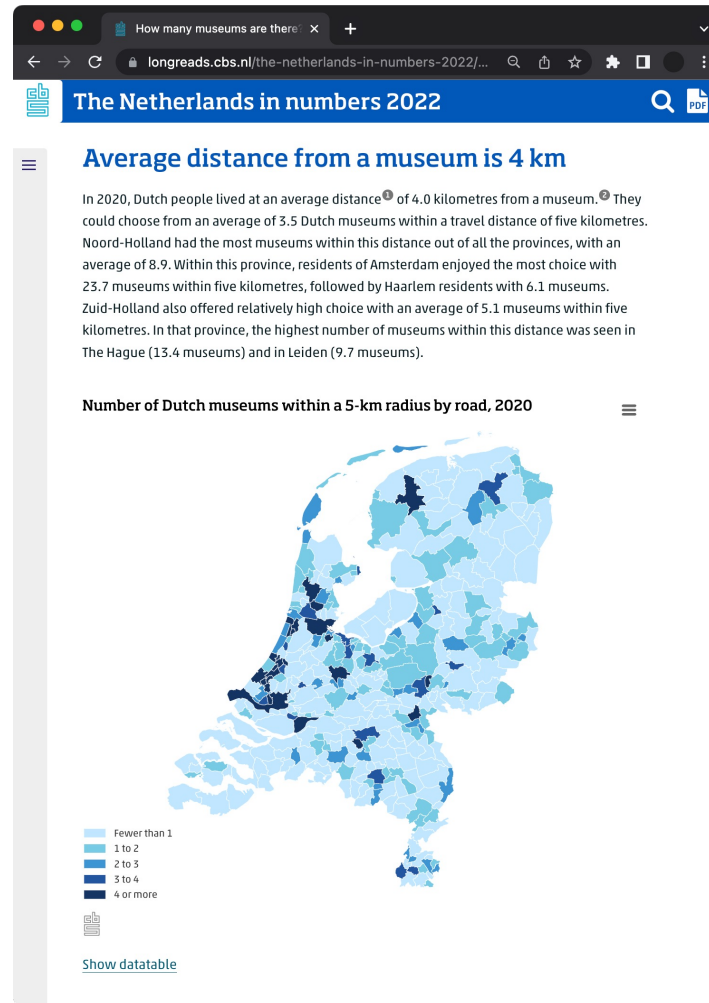
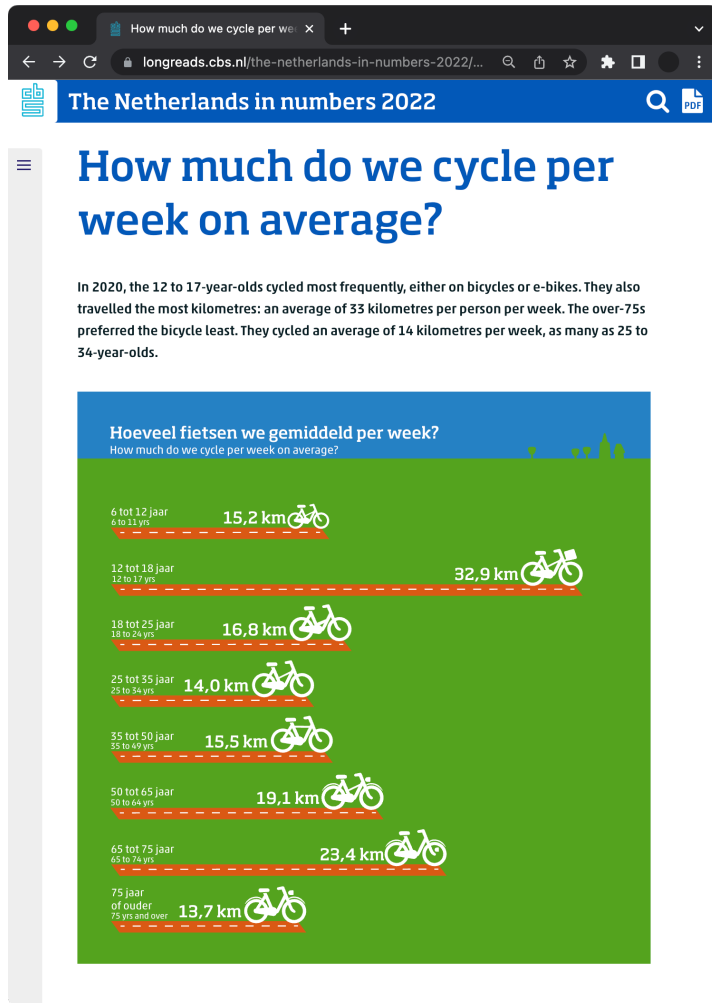
- Missing data is related to the variable that has the missing data. (e.g., sensitive questions)

A screenshot of a survey question. The question is: "Do you have any history of mental illness in your family? If yes, who in your family?". Below the question are two radio button options: "No" and "Other: _____". The "Other" option has a blank line for text entry.

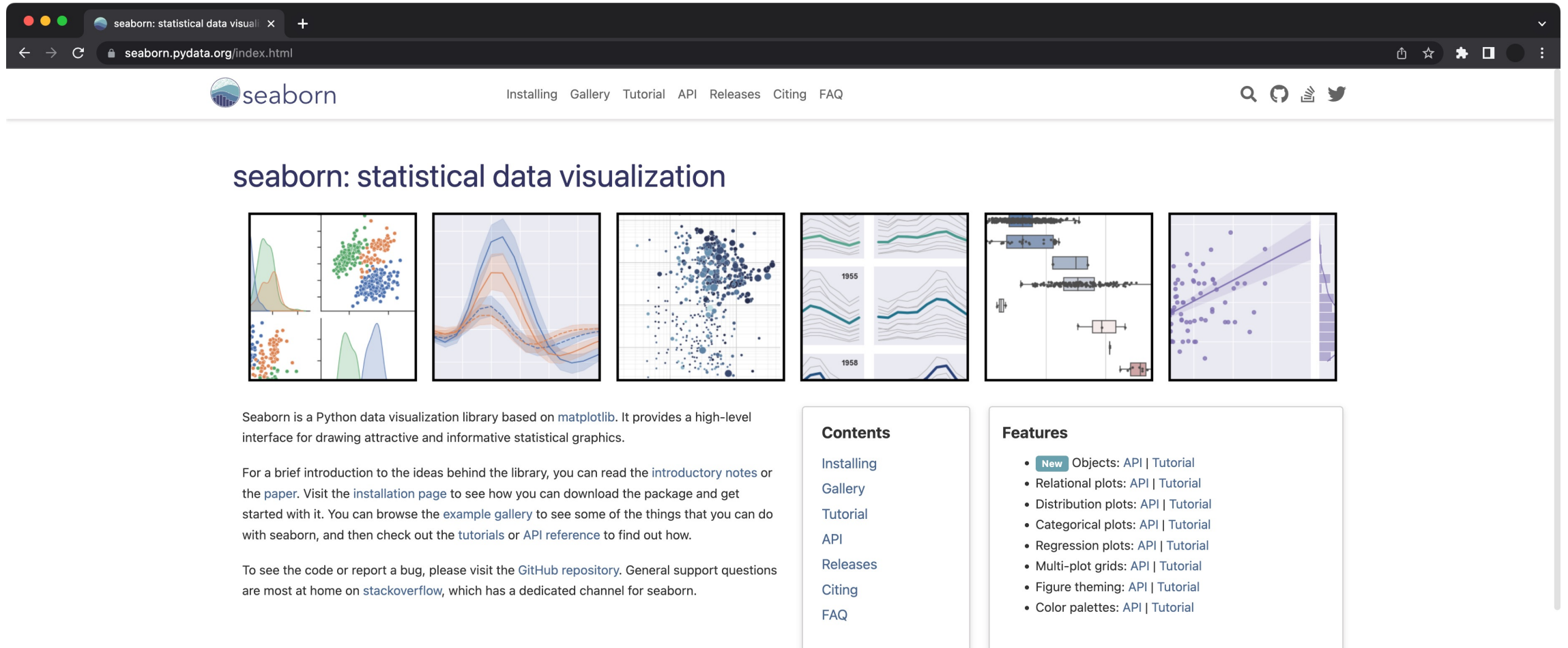
You really need to practice coding a lot to know and internalize how these things work!

- [Pandas exercises on GitHub](#)
- [Pandas exercises on Kaggle](#)
- [Pandas exercises on W3Schools](#)
- [Pandas exercises by UC Berkeley School of Information](#)
- [Pandas exercises on GeeksforGeeks](#)
- [Pandas exercises on w3resource](#)

The course will use examples (i.e., the modules) to teach you how to explore different types of data, either using existing tools or Python.



You can use the Python seaborn library (based on matplotlib) to quickly plot and explore structured data.



The screenshot shows the Seaborn website homepage. At the top, there's a navigation bar with links for "Installing", "Gallery", "Tutorial", "API", "Releases", "Citing", and "FAQ". The main heading is "seaborn: statistical data visualization". Below this, there are six thumbnail images showcasing different types of plots: a multi-panel plot with histograms and scatter plots, a density plot with multiple overlapping distributions, a scatter plot with a regression line, a faceted plot showing data for the years 1955 and 1958, a box plot with individual data points, and a scatter plot with a regression line and a shaded confidence interval.

seaborn: statistical data visualization

Seaborn is a Python data visualization library based on [matplotlib](#). It provides a high-level interface for drawing attractive and informative statistical graphics.

For a brief introduction to the ideas behind the library, you can read the [introductory notes](#) or the [paper](#). Visit the [installation page](#) to see how you can download the package and get started with it. You can browse the [example gallery](#) to see some of the things that you can do with seaborn, and then check out the [tutorials](#) or [API reference](#) to find out how.

To see the code or report a bug, please visit the [GitHub repository](#). General support questions are most at home on [stackoverflow](#), which has a dedicated channel for seaborn.

Contents

- [Installing](#)
- [Gallery](#)
- [Tutorial](#)
- [API](#)
- [Releases](#)
- [Citing](#)
- [FAQ](#)

Features

- New** Objects: [API](#) | [Tutorial](#)
- Relational plots: [API](#) | [Tutorial](#)
- Distribution plots: [API](#) | [Tutorial](#)
- Categorical plots: [API](#) | [Tutorial](#)
- Regression plots: [API](#) | [Tutorial](#)
- Multi-plot grids: [API](#) | [Tutorial](#)
- Figure theming: [API](#) | [Tutorial](#)
- Color palettes: [API](#) | [Tutorial](#)

The screenshot shows the Plotly Python Graphing Library website. On the left is a dark sidebar with the Plotly logo and navigation links: 'Quick Reference' (Getting Started, Is Plotly Free?, Figure Reference, API Reference, Dash, GitHub, community.plotly.com) and 'Examples' (Fundamentals, Basic Charts, Statistical Charts, Scientific Charts). At the bottom of the sidebar is a promotional banner for a webinar: 'Accelerating Energy Trading Analytics with Dash Enterprise' on February 15, 4pm CET, with a 'REGISTER' button.

The main content area features the title 'Plotly Open Source Graphing Library for Python' and a brief description: 'Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts. Plotly.py is free and open source and you can view the source, report issues or contribute on GitHub.' Below this is a link to 'Deploy Python AI Dash apps on private Kubernetes clusters: Pricing | Demo | Overview | AI App Services'.

Two main sections of examples are shown: 'Fundamentals' and 'Basic Charts'. The 'Fundamentals' section includes: 'The Figure Data Structure' (violin plots for Sunday and Saturday), 'Creating and Updating Figures' (polar plot with concentric circles), 'Displaying Figures' (bar chart titled 'US Export of Plastic Scrap'), 'Plotly Express' (a grid of scatter plots), and 'Analytical Apps with Dash' (a network graph). The 'Basic Charts' section includes: a scatter plot with colored points, a line plot with multiple series, a stacked bar chart, a pie chart with segments labeled 25%, 40%, and 18%, and a bubble chart with a red box labeled '(31,656k, 82,603) Japan'.

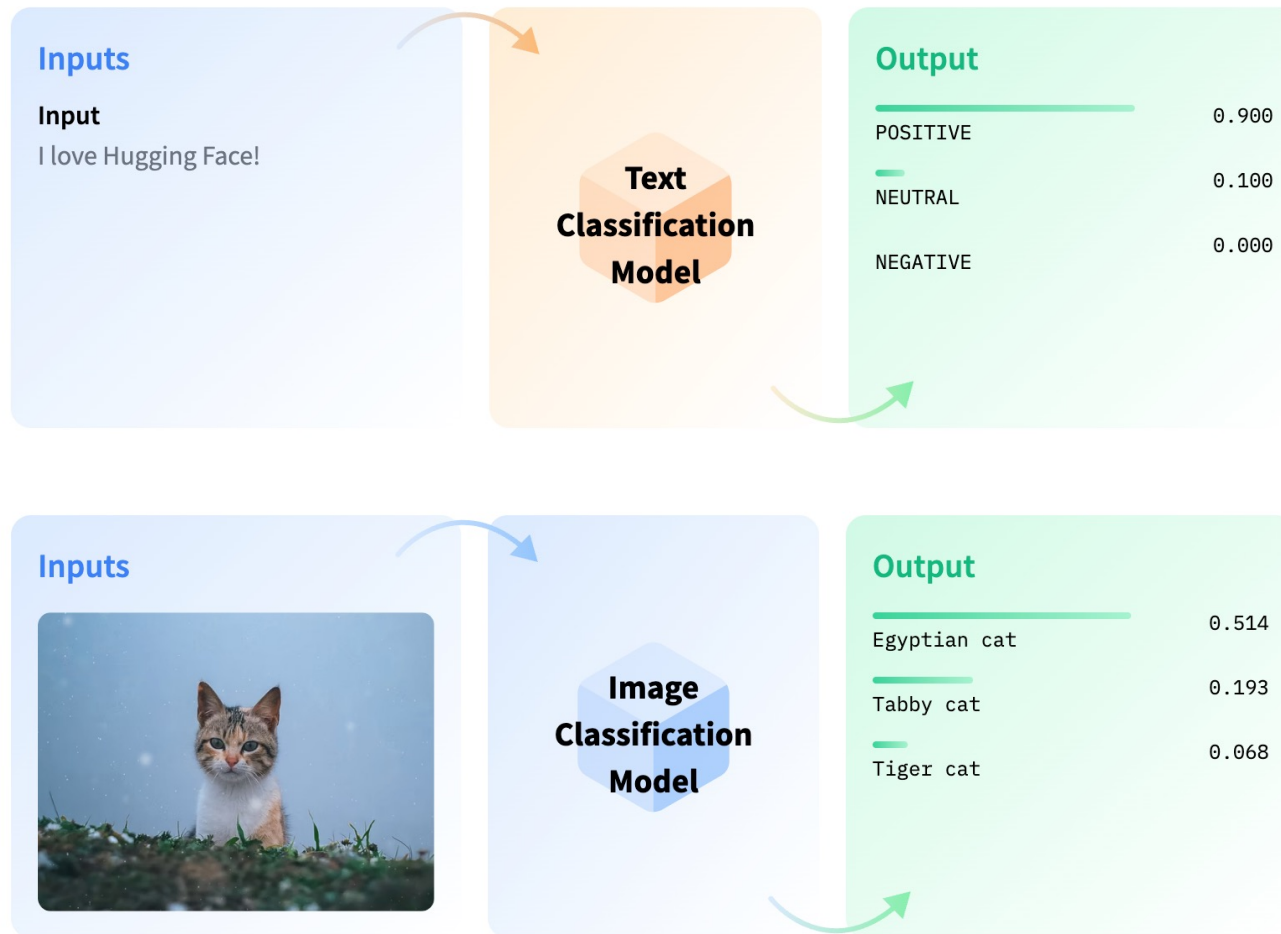
Introduction This content is written for the Du...

Introduction
 This content is written for the Dungeon Master. It contains a complete Dungeons & Dragons adventure, as well as descriptions for every creature and magic item that appears in the adventure. It also introduces the world of the Forgotten Realms, one of the game's most enduring settings, and it teaches you how to run a D&D game. The Basic Rules contain the rules you need to adjudicate situations that arise during the adventure. Running the Adventure Lost Mine of Phandelver is an adventure for four to five characters of 1st level.

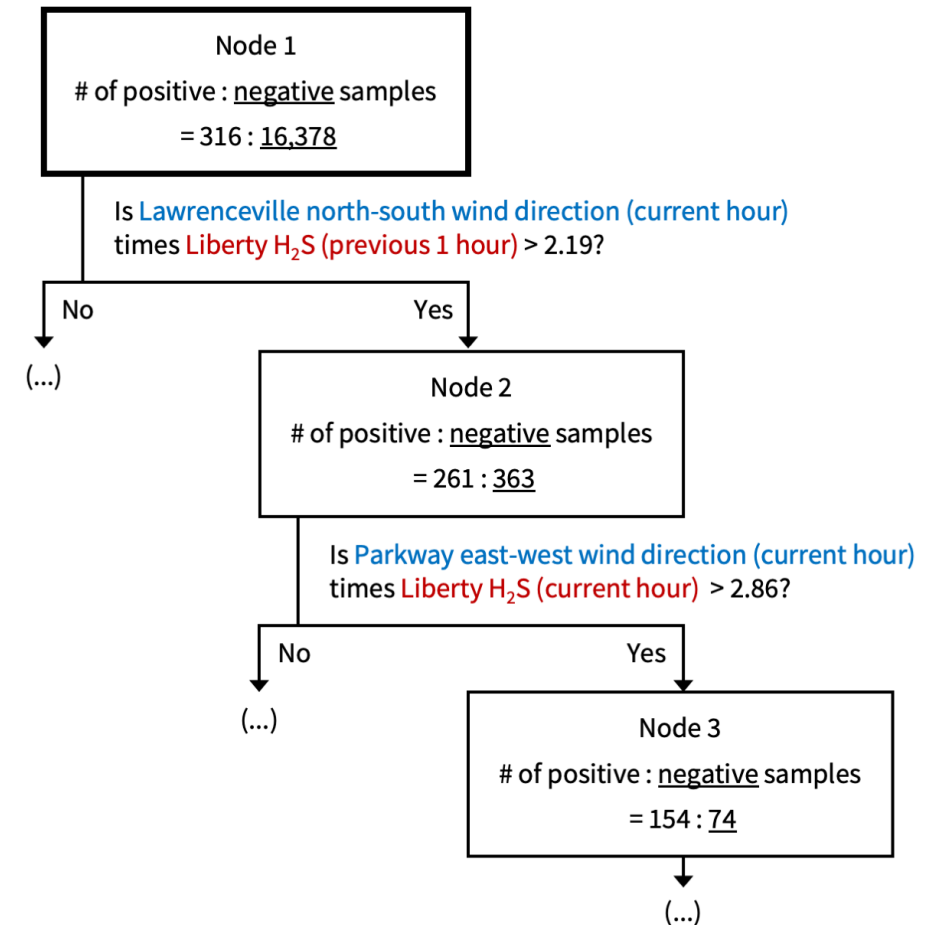
Document	Left	Term	Right
Introduct...	contains a complete Dungeons & Dragons	ad...	, as well as descriptions for
Introduct...	item that appears in the	ad...	. It also introduces the world
Introduct...	situations that arise during the	ad...	. Running the Adventure Lost Mine
Introduct...	during the adventure. Running the	ad...	Lost Mine of Phandelver is
Introduct...	Mine of Phandelver is an	ad...	for four to five characters

Model Data

This course will teach you techniques for modeling structured, text, and image data through three modules from a practical point of view.

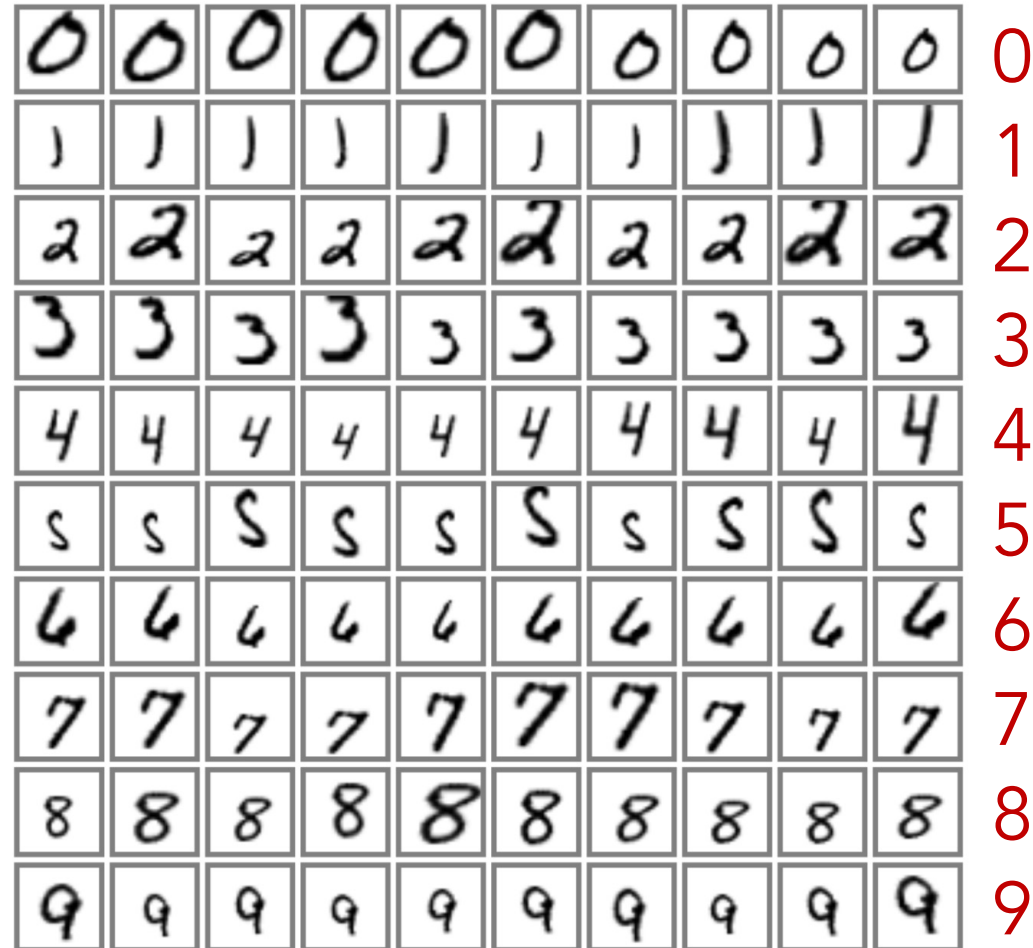


Source: <https://huggingface.co/tasks>



Source: <https://smellpgh.org/>

One example of image classification is **optical character recognition**, such as recognizing digits from hand-written images.



A more complicated image classification task is **fine-grained categorization**, such as categorizing the types of birds.



Barred Owl



American Robin



American Crow



Rufous Hummingbird



Rock Pigeon



Canada Goose











One example of text classification is **sentiment analysis**, such as identifying emotions from movie reviews.



STAR WARS: THE LAST JEDI REVIEWS

All Critics **Top Critics** All Audience

Page 1 of 4

	Matthew Rozsa <i>Salon.com</i>	★ Top Critic	✱ "Star Wars" is not "Breaking Bad," and the same narrative tricks that worked for the latter feel jarringly out of place in the former.	July 27, 2018	
	Leah Pickett <i>Chicago Reader</i>	★ Top Critic	🍅 What's most interesting to me about The Last Jedi is Luke's return as the mentor rather than the student, grappling with his failure in this new role, and later aspiring to be the wise and patient teacher.	December 26, 2017	
	Jake Wilson <i>The Age (Australia)</i>	★ Top Critic	🍅 While The Last Jedi may not receive top marks for originality, the eighth official entry in the Star Wars saga is still one of the most entertaining blockbusters of the year...	December 16, 2017	
	Peter Rainer <i>Christian Science Monitor</i>	★ Top Critic	🍅 Fanatics will love it; for the rest of us, it's a tolerably good time.	December 15, 2017	
	Matthew Lickona <i>San Diego Reader</i>	★ Top Critic	🍅 The devoted will no doubt be delighted; for the rest, a resigned acceptance may be the safest path to enjoyment.	December 15, 2017	

A more complex text classification task is annotating paragraphs, such as **categorizing the research aspect** for each fragment in the paper abstract.

For successful infection, viruses must recognize their respective host cells. A common mechanism of host recognition by viruses is to utilize a portion of the host cell as a receptor. Bacteriophage Sf6, which infects *Shigella flexneri*, uses lipopolysaccharide as a primary receptor and then requires interaction with a secondary receptor, a role that can be fulfilled by either outer membrane proteins (Omp) A or C. Our previous work showed that specific residues in the loops of OmpA mediate Sf6 infection. **To better understand Sf6 interactions with OmpA loop variants,** we determined the kinetics of these interactions through the use of biolayer interferometry, an optical biosensing technique that yields data similar to surface plasmon resonance. Here, we successfully tethered whole Sf6 virions, **determined the binding constant of Sf6 to OmpA to be 36 nM.** Additionally, we showed that Sf6 bound to five variant OmpAs and the resulting kinetic parameters varied only slightly. Based on these data, we propose a model in which Sf6: Omp receptor recognition is not solely based on kinetics, but likely also on the ability of an Omp to induce a conformational change that results in productive infection. All rights reserved. No reuse allowed without permission.

Background

Purpose

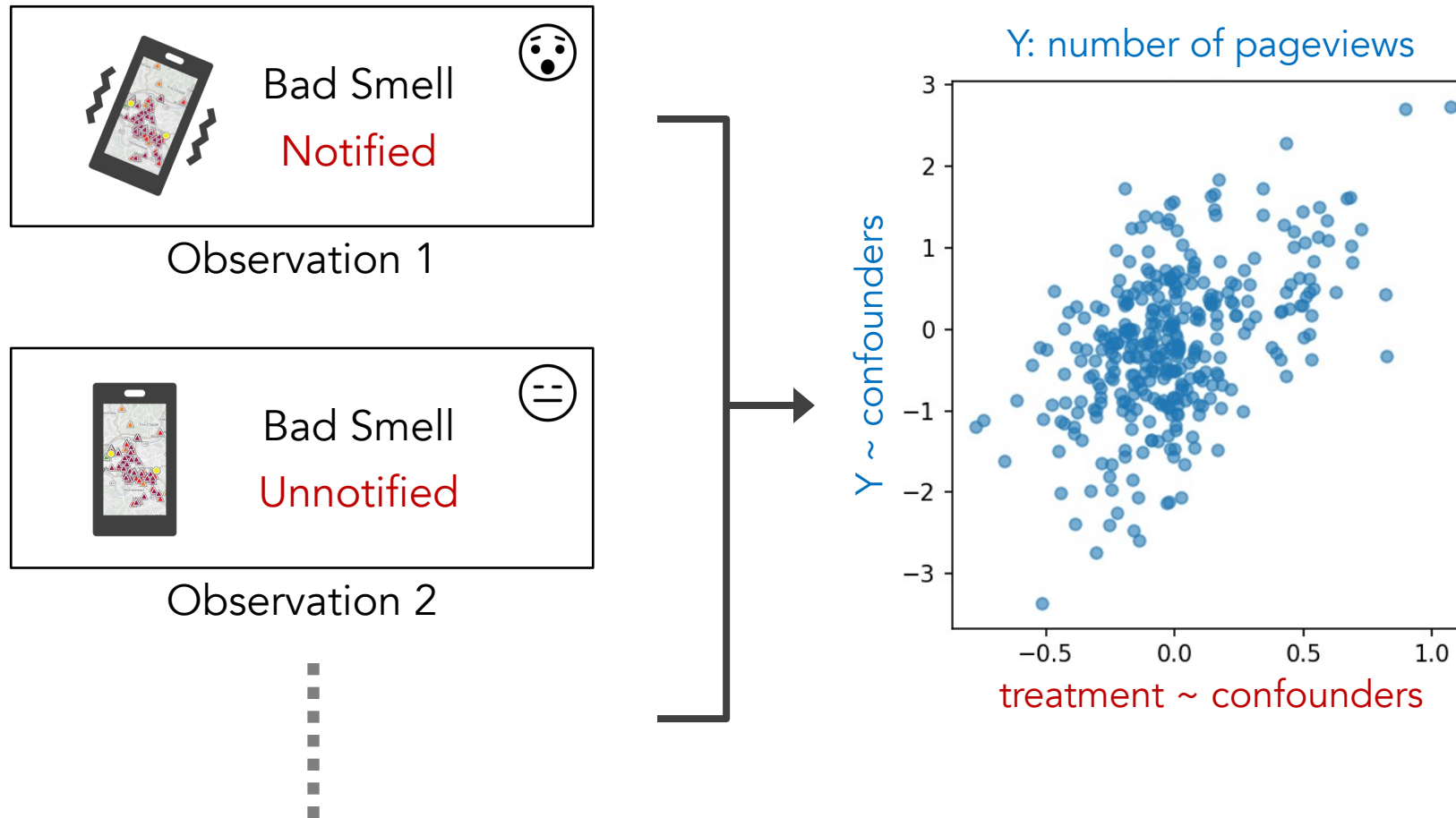
Method

Finding

Other

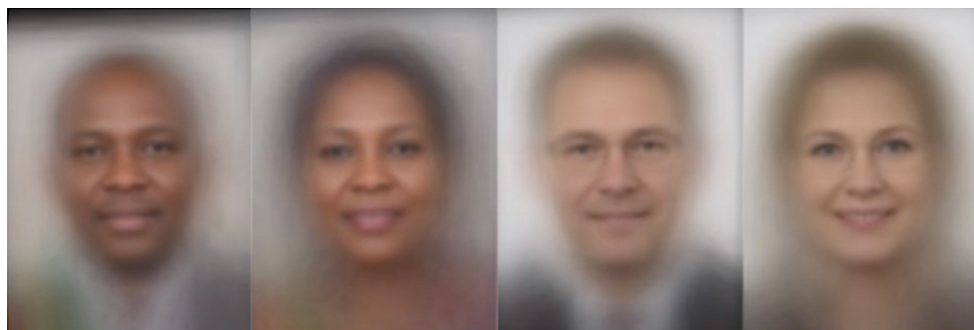
Deploy Models



















Deploying models in the wild can enable further quantitative or qualitative research with insights, such as the push notification study in Smell Pittsburgh.



Deploy Models

Another example is to study the behaviors of deployed systems to understand its social impact. For example, face recognition systems for recognizing gender did **worst on darker-skin female images**.



Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

Take-Away Messages

- The data science pipeline is not always linear. Be flexible and open-minded!
- Be aware of the step of framing problems. This course assumes that the problems are defined.
- Collecting data requires well-designed software/hardware infrastructure.
- Being familiar with pandas can speed up the data preprocessing step.
- Different types of missing data require different treatments.
- Besides using descriptive statistics, it is also a good idea to visualize and explore data.
- Different types of data need different modeling techniques. There is no “one solution for all”.
- It is important to study user behaviors and investigate the social impact of deployed models.



Questions?