Data Science

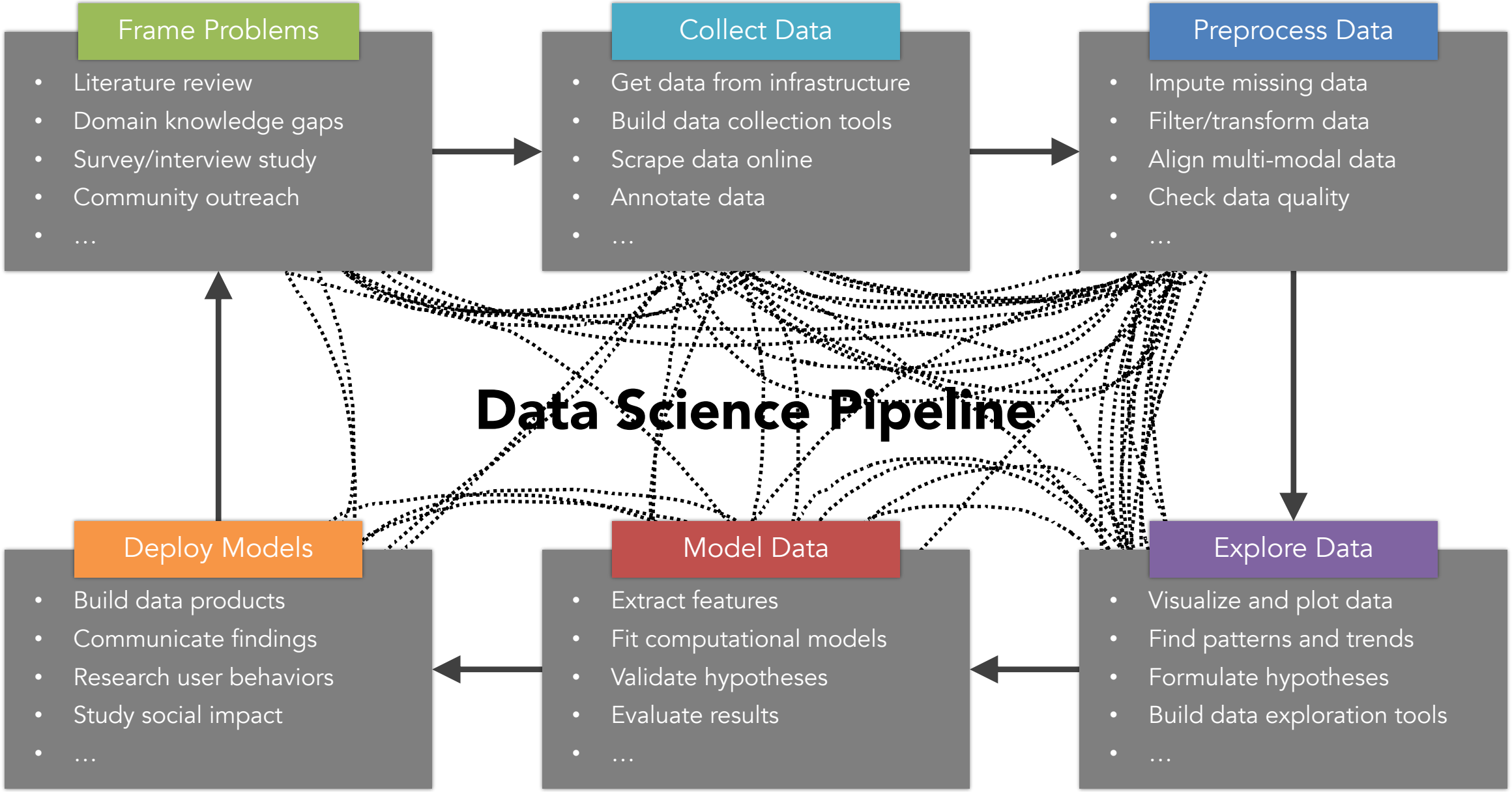
Lecture 1-1: Course Introduction and Organization

University of Amsterdam

Lecturer: Yen-Chia Hsu

Date: Feb 2026

Data science is about turning rich data into actionable insight and making data impactful!



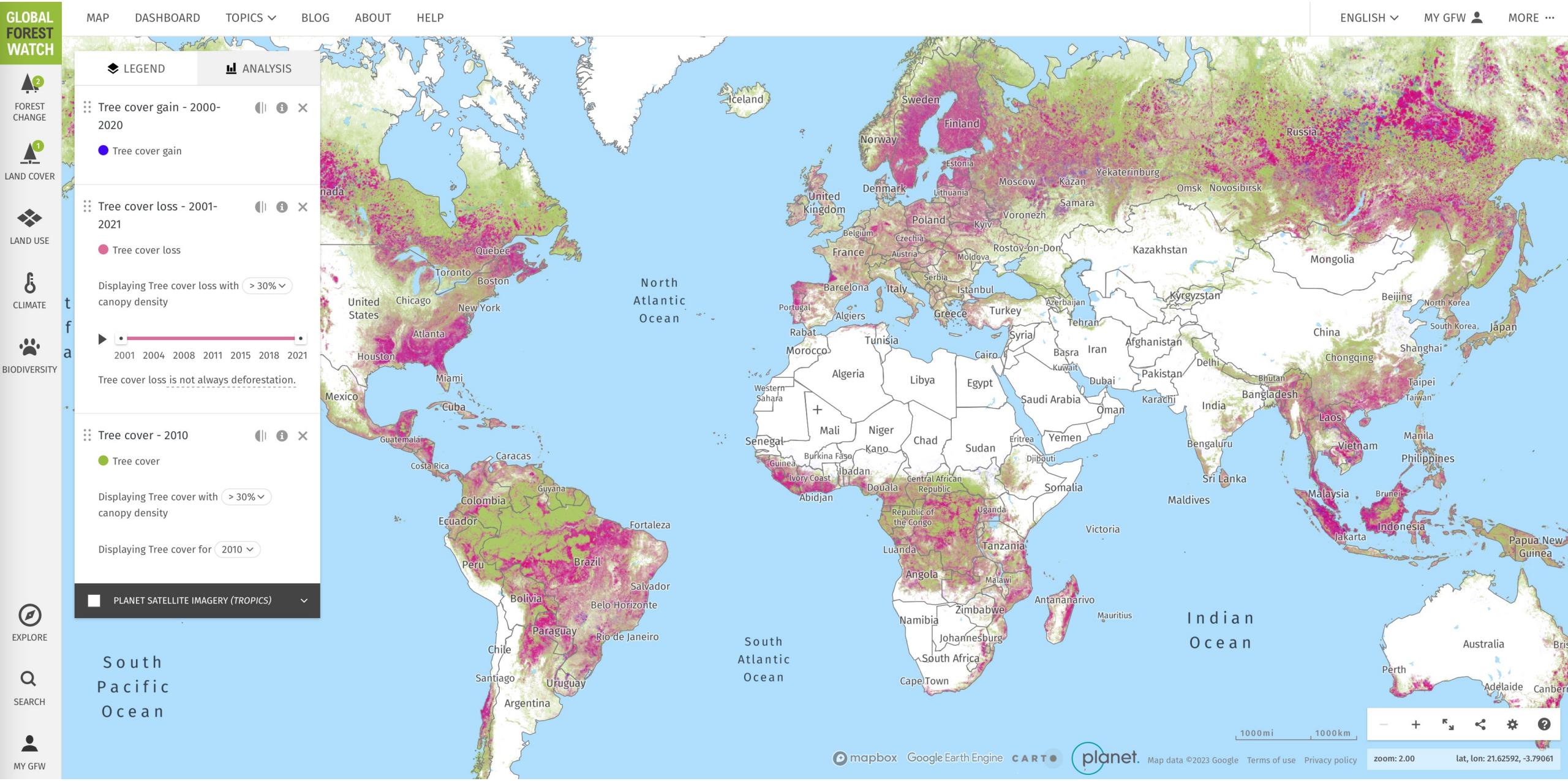
What people typically think : →

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

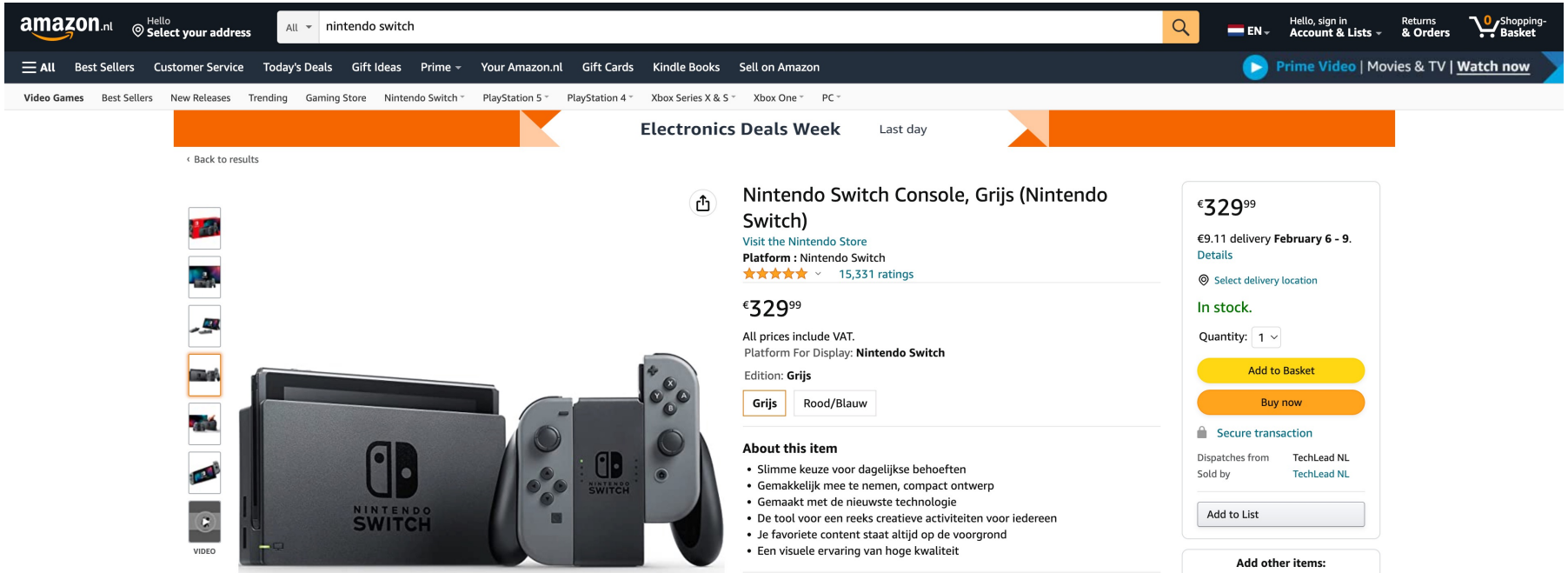
Frame Problems

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

Example: understanding parking patterns and law enforcement -- https://responsiblesensinglab.org/projects/scan-cars



Example: analyzing the situation of global forest situation -- https://www.globalforestwatch.org/map/



What other items do customers buy after viewing this item?

Page 1 of 6

Nintendo Switch console (OLED model) with white Joy-Con docking station
Nintendo
8,078
Nintendo Switch
€324.95
Get it as soon as **Wednesday, Feb 1**
FREE Delivery on orders dispatched by Amazon over €20

Nintendo Switch Mario Kart 8 Deluxe
Nintendo
79
Nintendo Switch
€49.90
Get it as soon as **Wednesday, Feb 1**
FREE Delivery on orders dispatched by Amazon over €20

SanDisk MicroSDXC UHS-I Card for Nintendo Switch, true red
255,010
€19.29

Orzly Carrying Case Compatible with Nintendo Switch and New Switch OLED…
56,320
#1 Best Seller in Nintendo Switch Cases
41% off Deal
€13.53
Was: ~~€23.09~~
Get it as soon as **Thursday, Feb 2**

New Super Mario Bros. U Deluxe (Nintendo Switch)
Nintendo
12,488
Nintendo Switch
€48.95
Get it as soon as **Monday, Feb 6**
FREE Delivery on orders dispatched by Amazon over €20

Nintendo Switch console (OLED-model): nieuwe versie, intense kleuren, 7 inch scherm - met een…
Nintendo
2,099
Nintendo Switch
€338.00

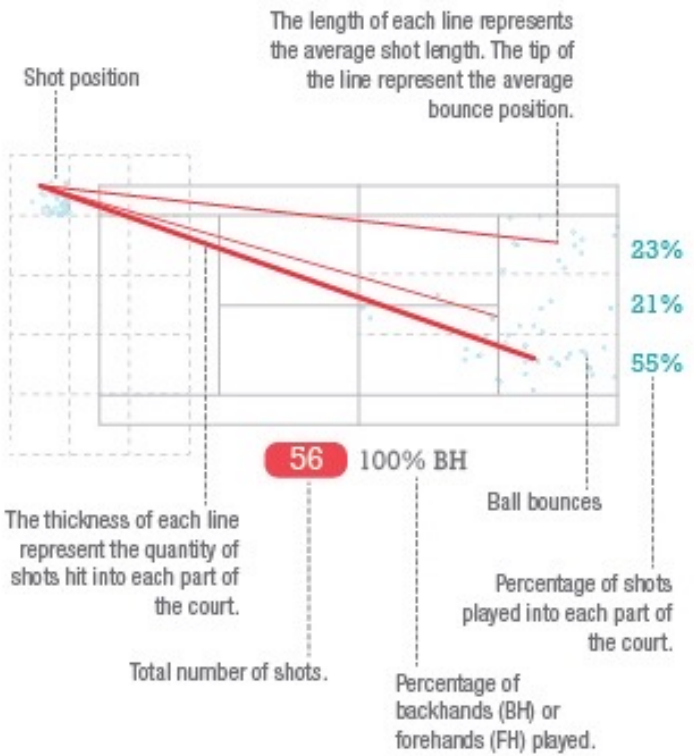
Mario Kart 8 : Deluxe (Nintendo Switch)
Nintendo
39,054
Nintendo Switch
€49.95

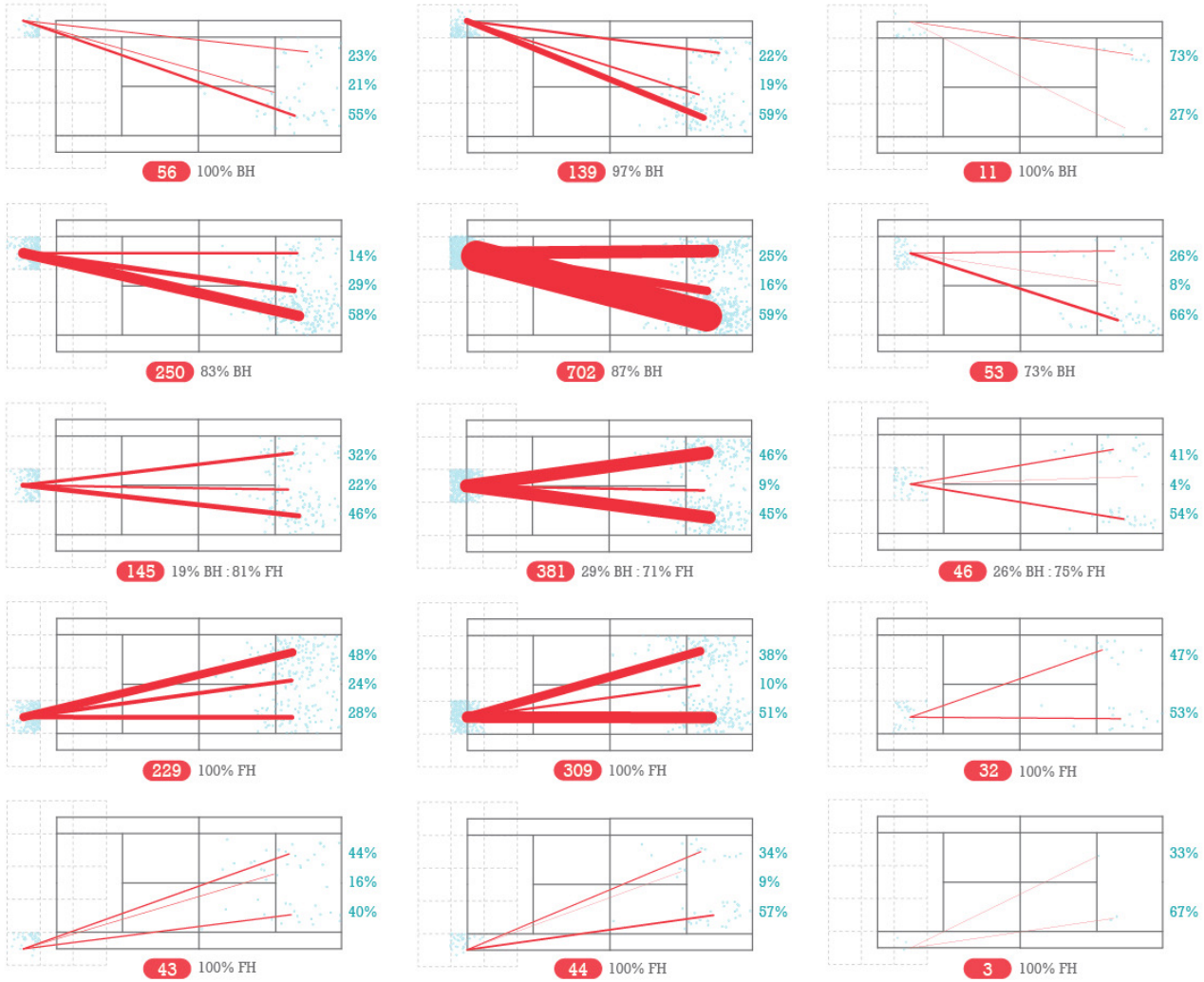
Example: recommending products to customers -- https://www.amazon.nl/

Kei Nishikori Shot Charts

錦織圭のショットチャート

Shot charts are critical in understanding a player's on court behaviour. They are frequently used to map shot patterns from particular areas of the court. These patterns are of particular interest to coaches and players for pre and post match tactical analysis. Here we present 2,443* shots from Kei Nishikori that were played over a period of 6 months in 2014–15 against opponents like Federer, Djokovic, Murray and Wawrinka.



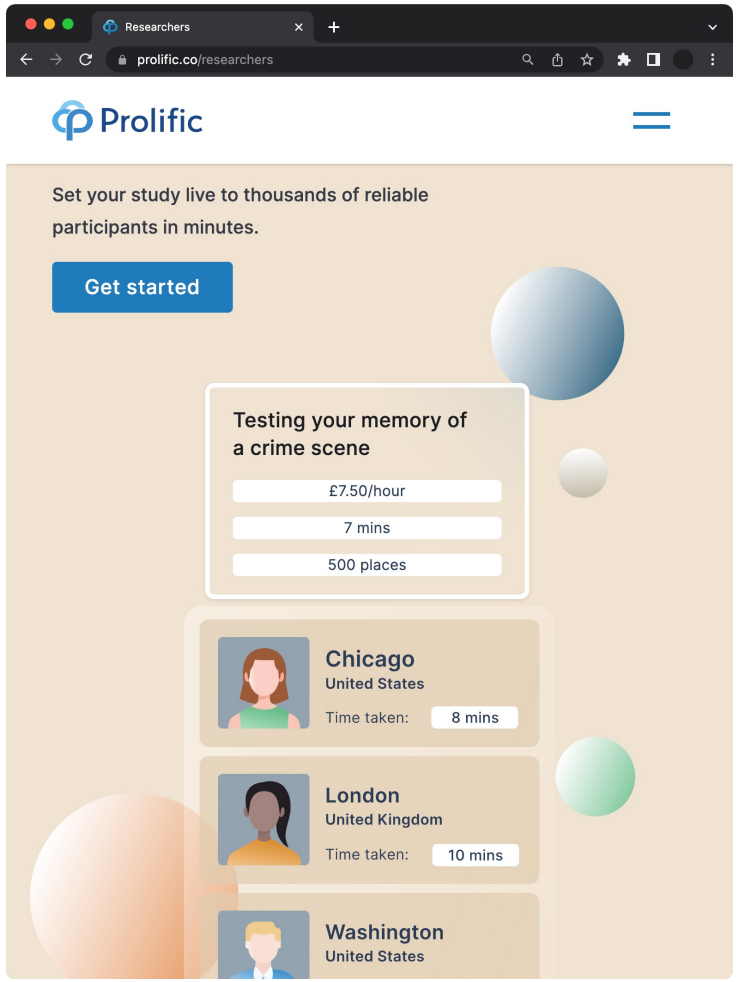


Example: analyzing tennis player behavior -- https://tennismash.com/2016/01/18/kei-nishikori-shot-charts/

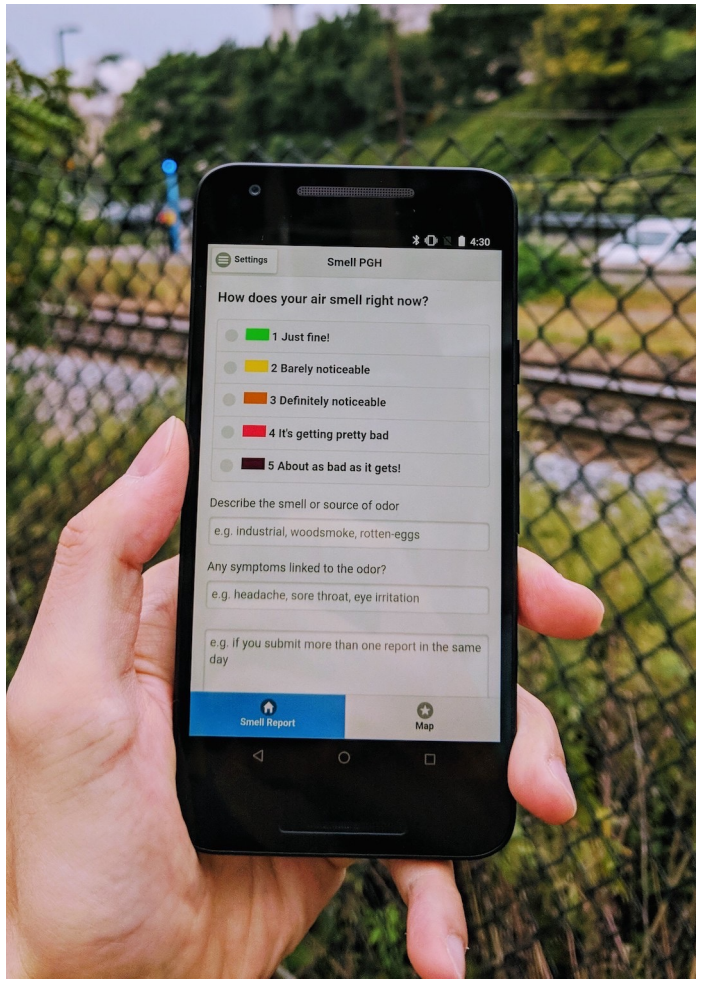
Collect Data

This course assumes that someone has collected the data for you. In reality, you may need to collect data using sensors, crowdsourcing, mobile apps, etc.

GGD Amsterdam Data Portal



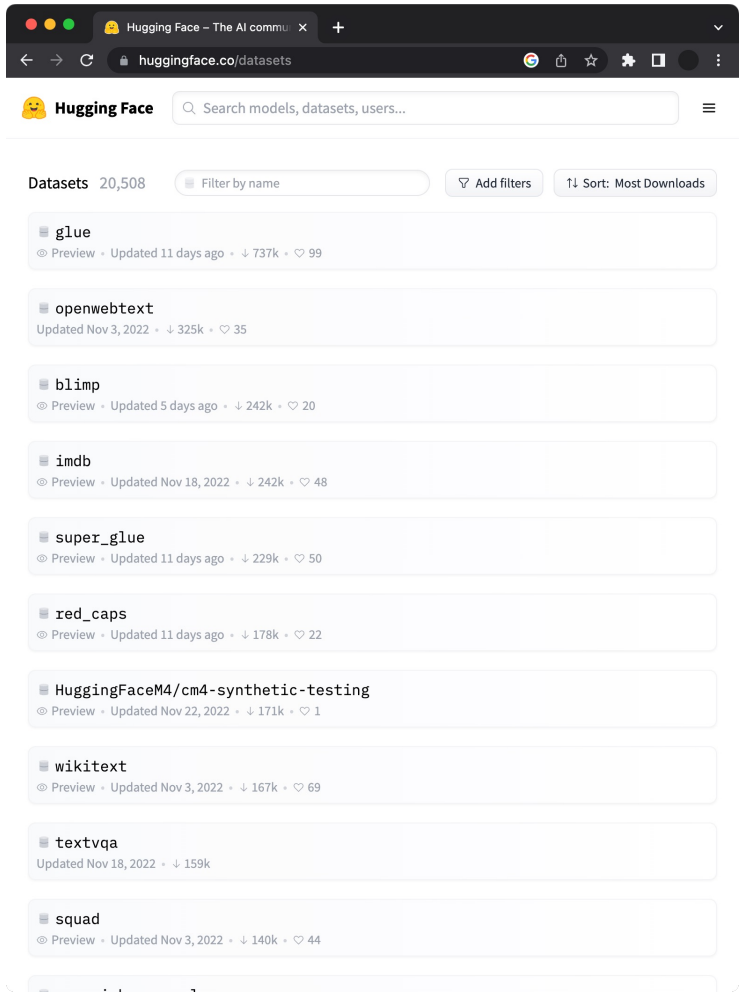
Prolific Tool for Data Annotation

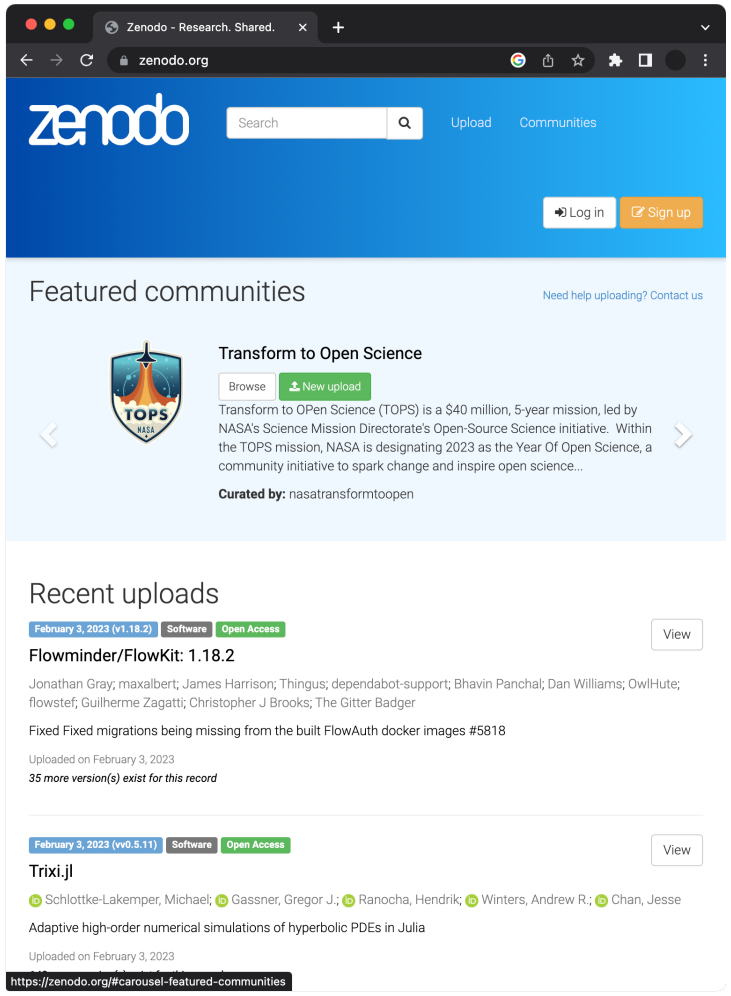


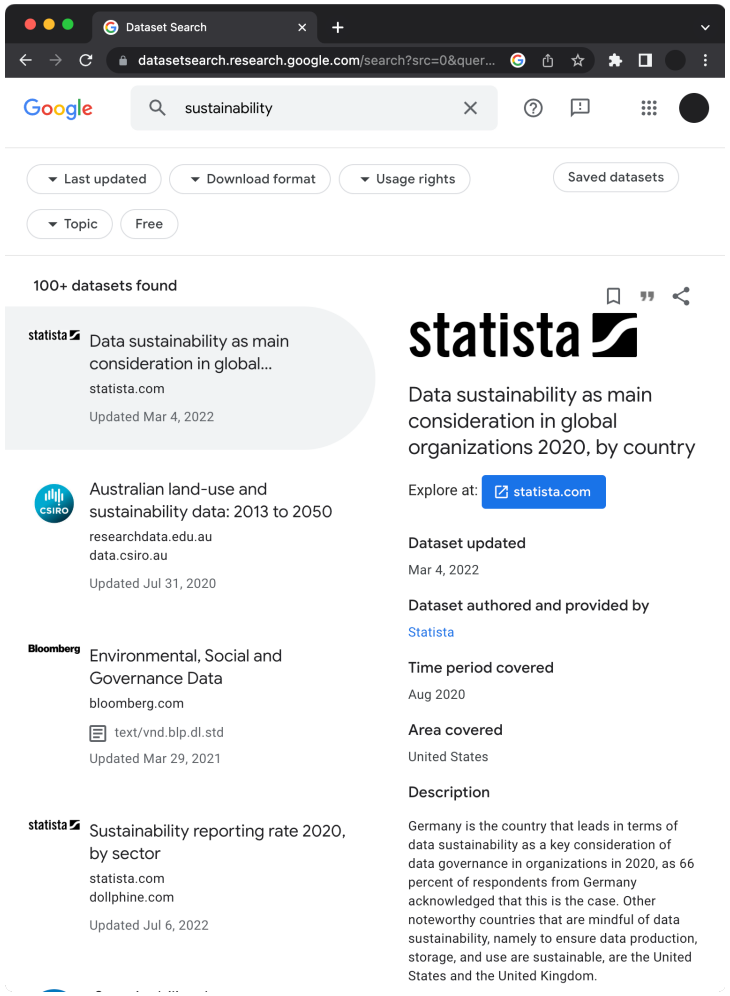
Mobile App Data Collection

Collect Data

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







Hugging Face

Zenodo

Google Dataset Search

Preprocess Data

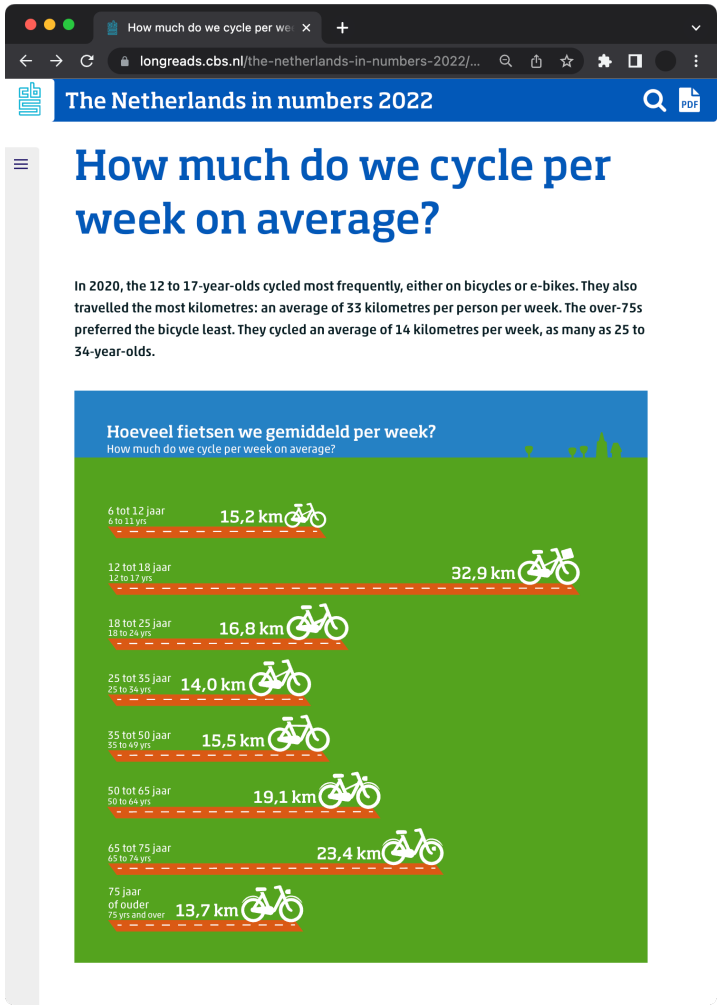
This course will use pandas, which is a very handy Python library for preprocessing structured data. We will cover the following techniques:

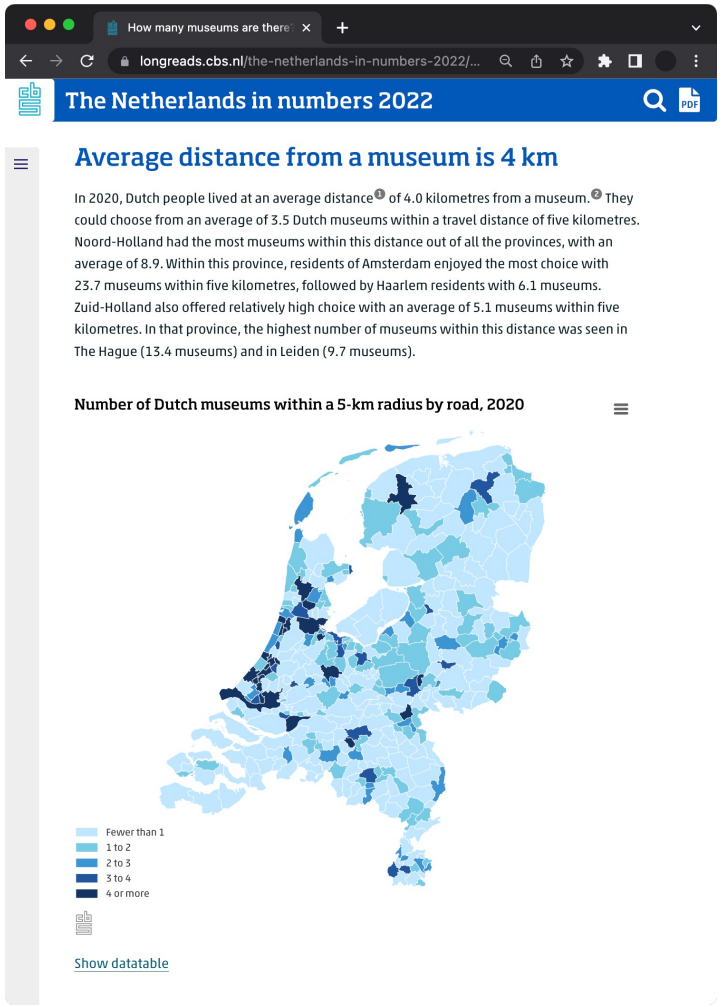
- Filter unwanted data
- Aggregate data (e.g., sum)
- Group data based on a column
- Sort rows based on a column
- Concatenate data frames
- Merge and join data frames
- Quantize continuous values into bins
- Scale column values
- Resample time series data
- Roll time series data in a window
- Apply a transformation function
- Use regular expressions
- Drop rows or columns
- Treat missing values

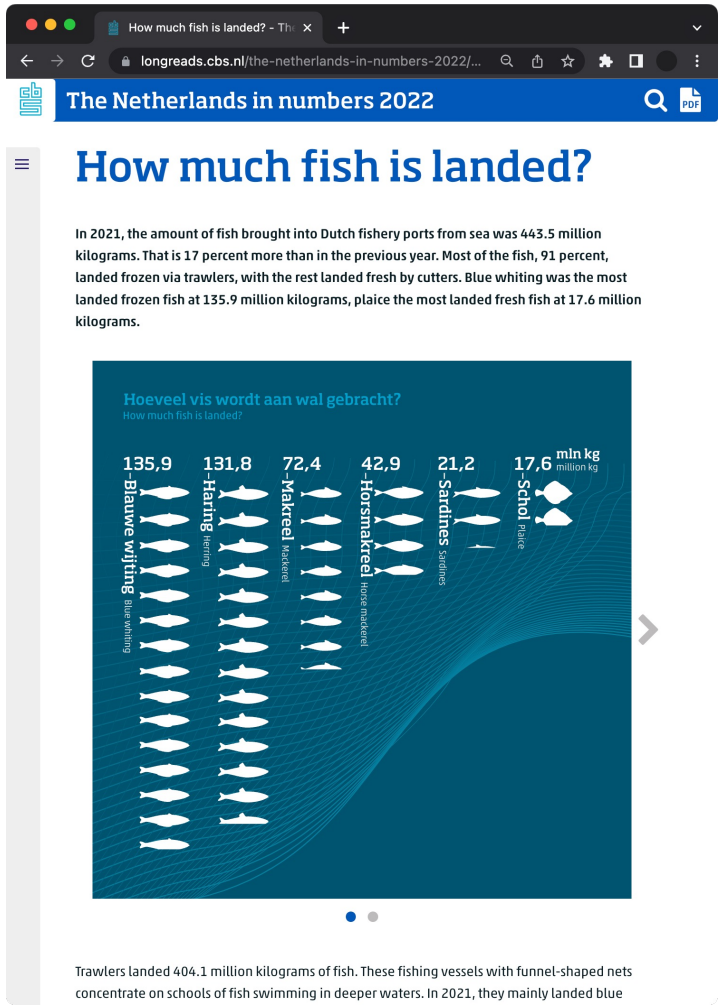
10 minutes to pandas -- https://pandas.pydata.org/docs/user_guide/10min.html

Explore Data

Information visualization is a good way for both experts and laypeople to explore data and gain insights.



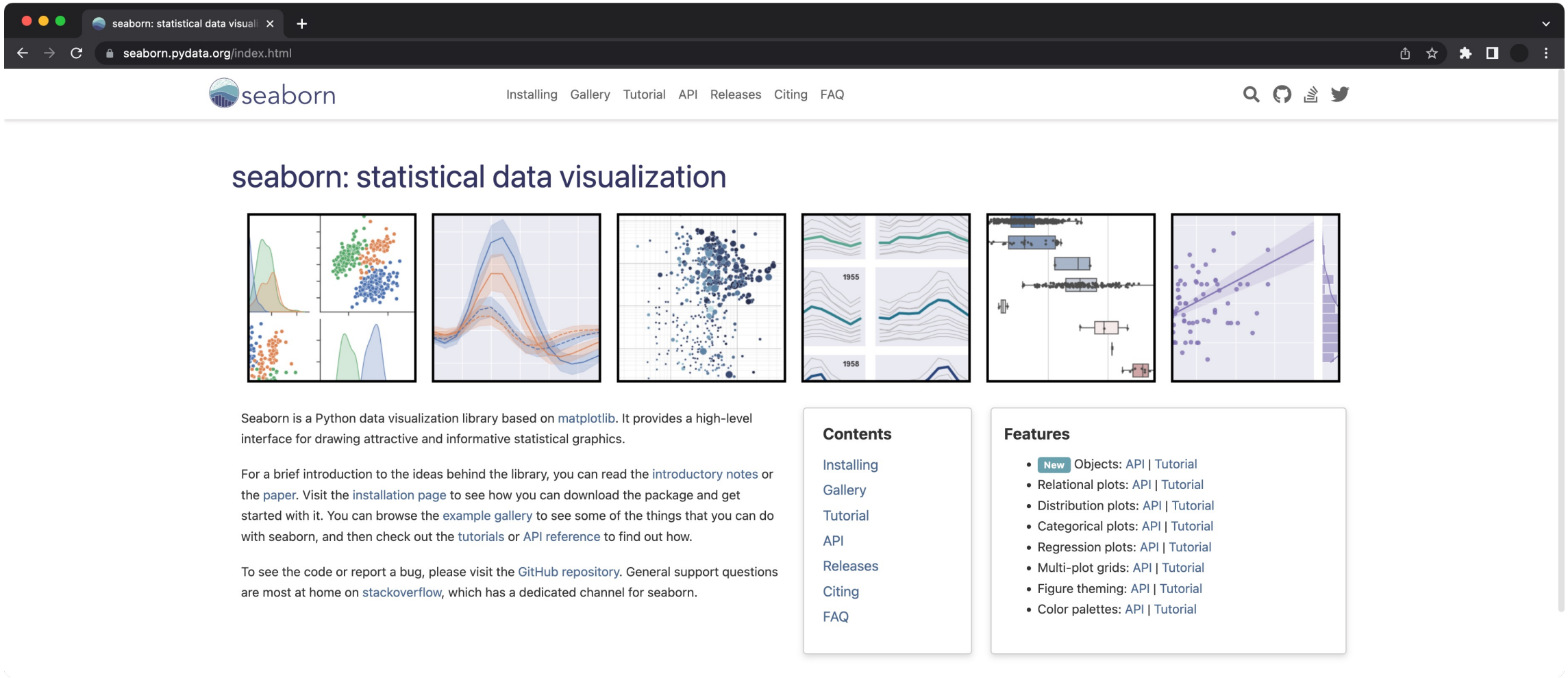




The Netherlands in Numbers -- https://longreads.cbs.nl/the-netherlands-in-numbers-2022/

Explore Data

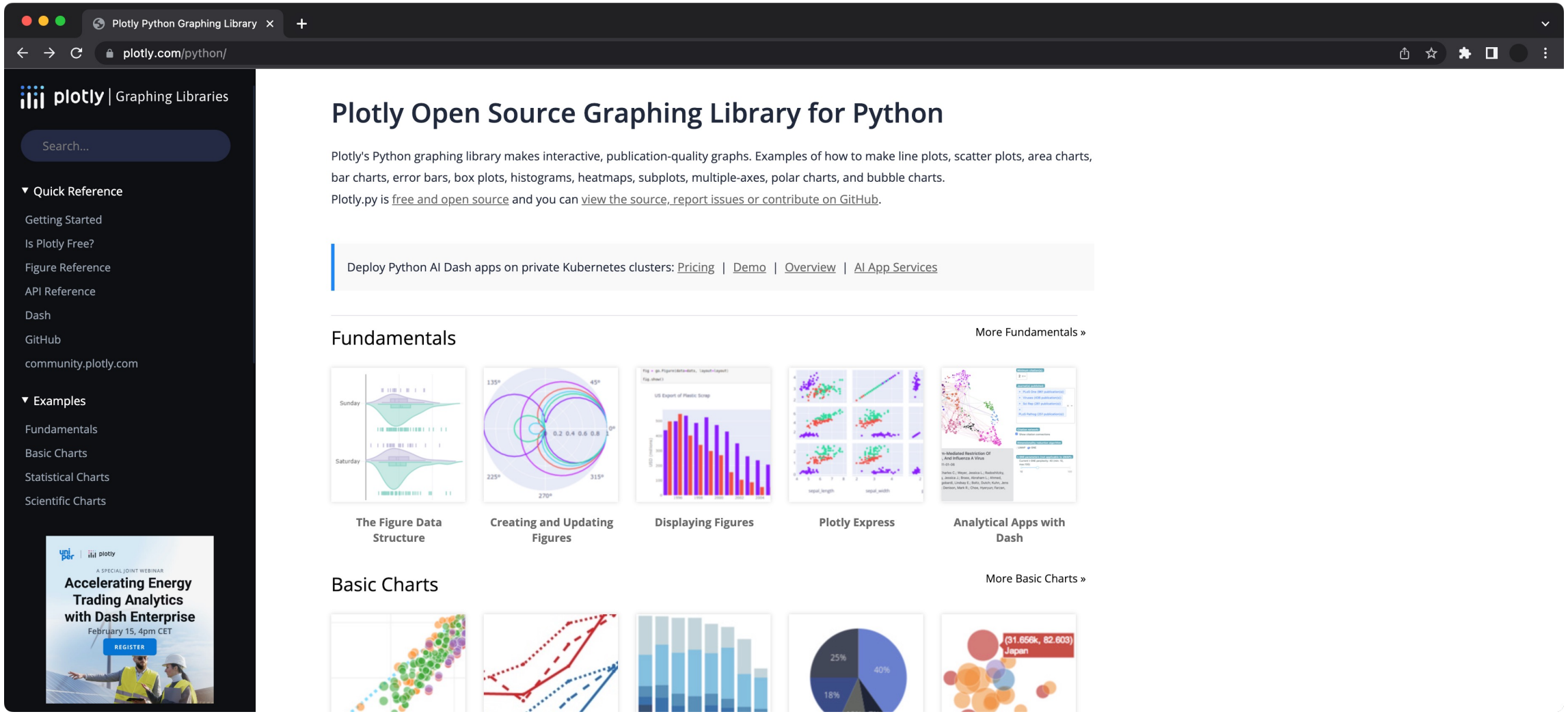
You can use the Python seaborn library (based on matplotlib) to quickly plot and explore structured data.



The Seaborn Library for Statistical Data Visualization -- https://seaborn.pydata.org/

Explore Data

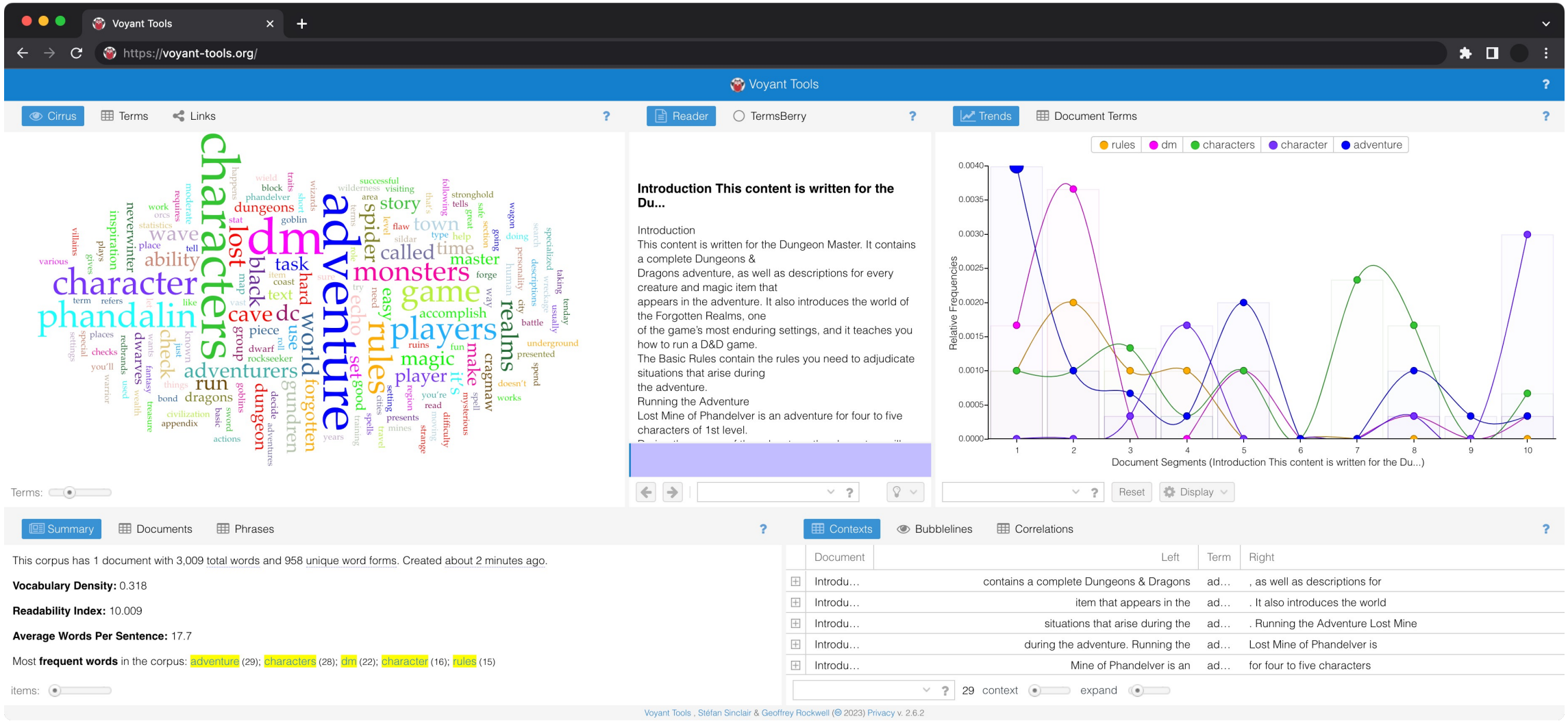
You can use the Python Plotly library to build interactive visualizations.



Plotly Open Source Graphing Library for Python -- https://plotly.com/python/

Explore Data

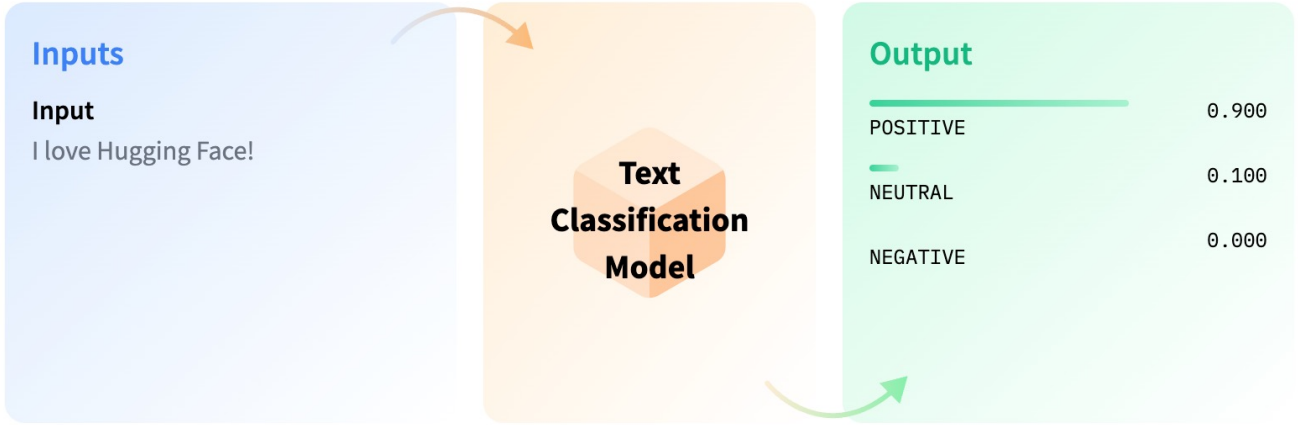
You can use the Voyant Tools to explore text data.

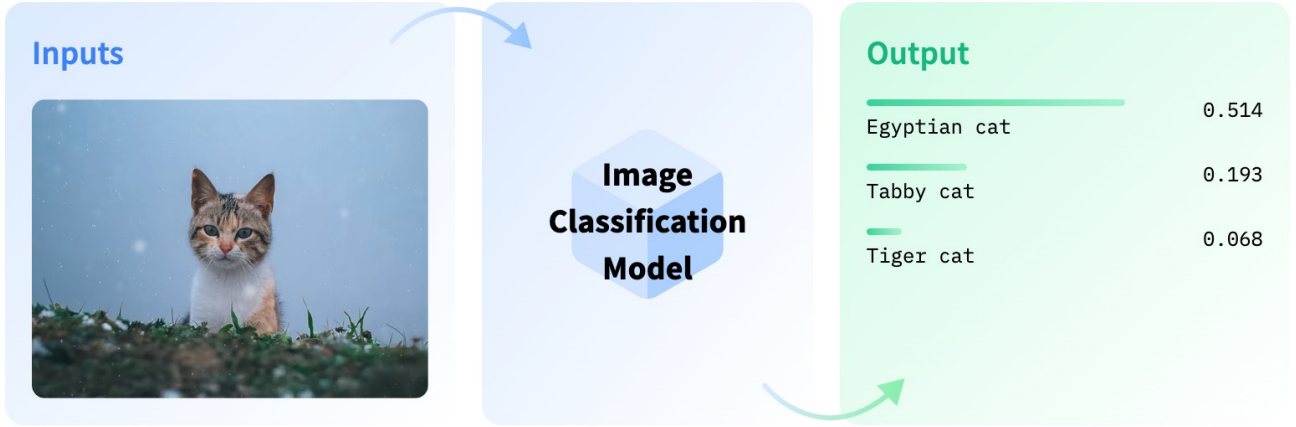


Voyant Tools -- https://voyant-tools.org/

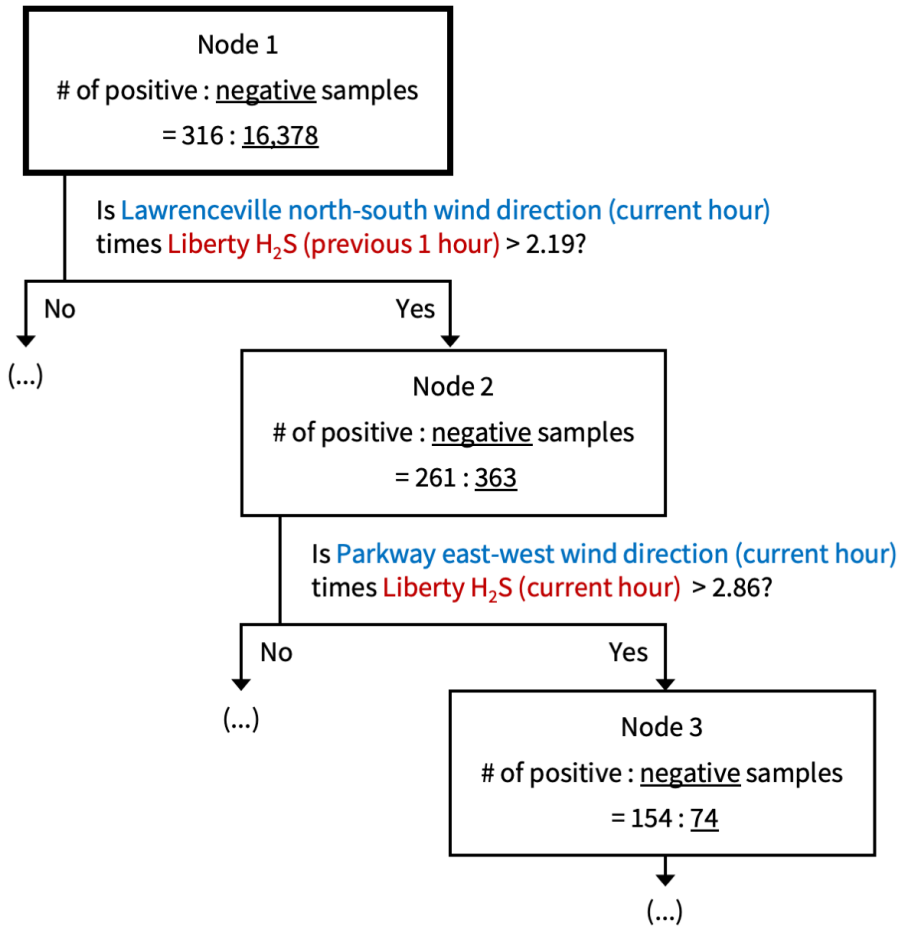
Model Data

This course will teach you machine learning 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/

Model Data

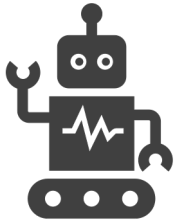
The structured data processing module uses air quality sensing and weather data to predict smell events.

O_3: 26 ppb	CO: 127 ppb
H_2S: 0 ppb	$PM_{2.5}$: 9 µg/m^3
Wind: 17 deg	…

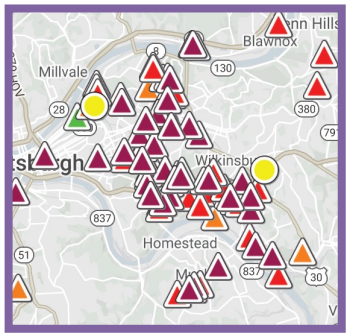
Observation 1

O_3: 1 ppb	CO: 1038 ppb
H_2S: 9 ppb	$PM_{2.5}$: 23 µg/m^3
Wind: 213 deg	…

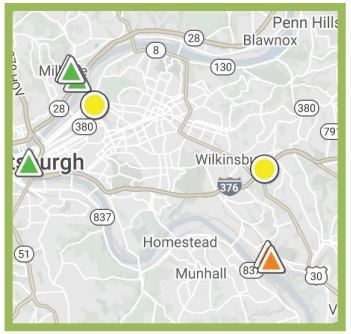
Observation 2



Machine Learning



Has Event



No Event

Source -- https://smellpgh.org/

Model Data

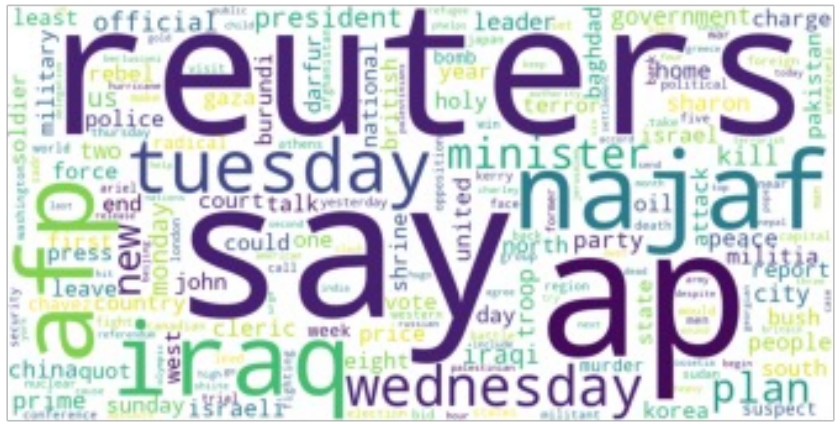
The text data processing module is about topic classification and topic modeling of online news items.

Class: Business

Class: Sci/Tech

Class: Sports

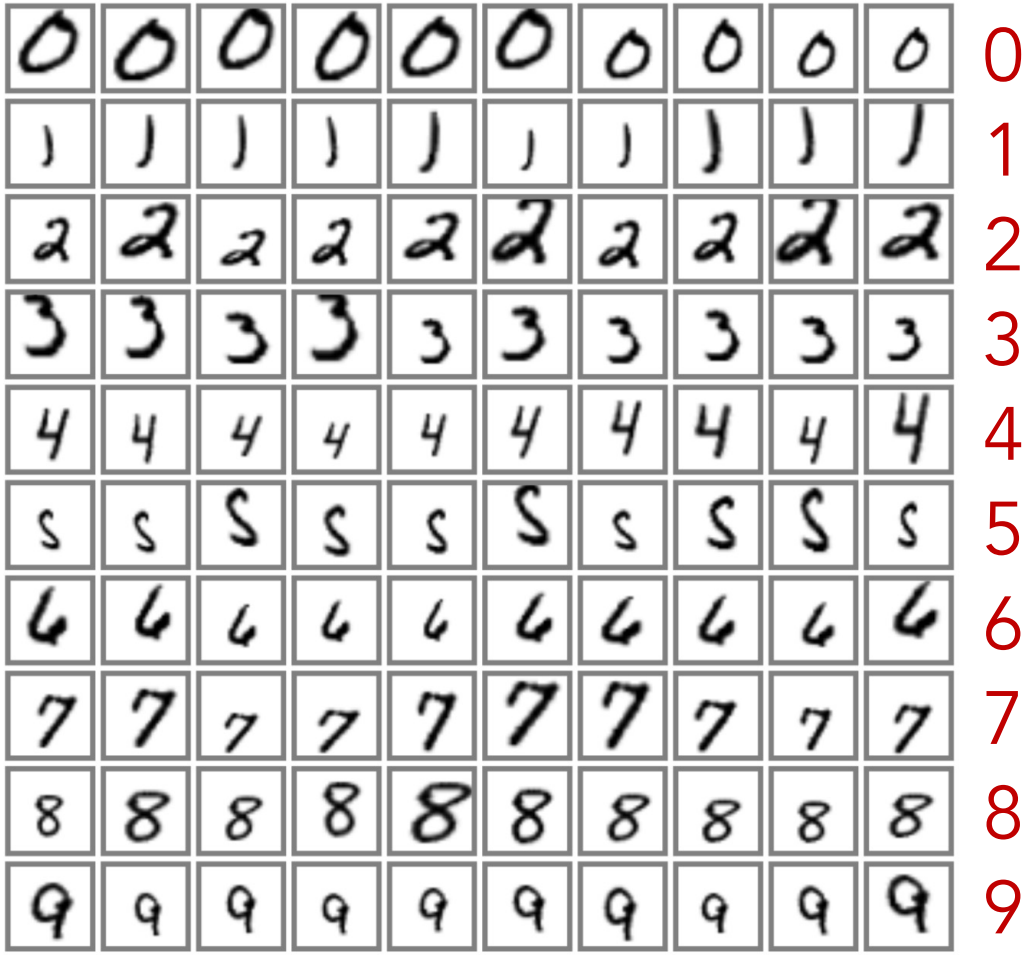
Class: World



Source -- https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset

Model Data

The image data processing module uses deep neural networks to recognize digits from hand-written images.



LeCun et al. 1998. Gradient-based learning applied to document recognition. IEEE.

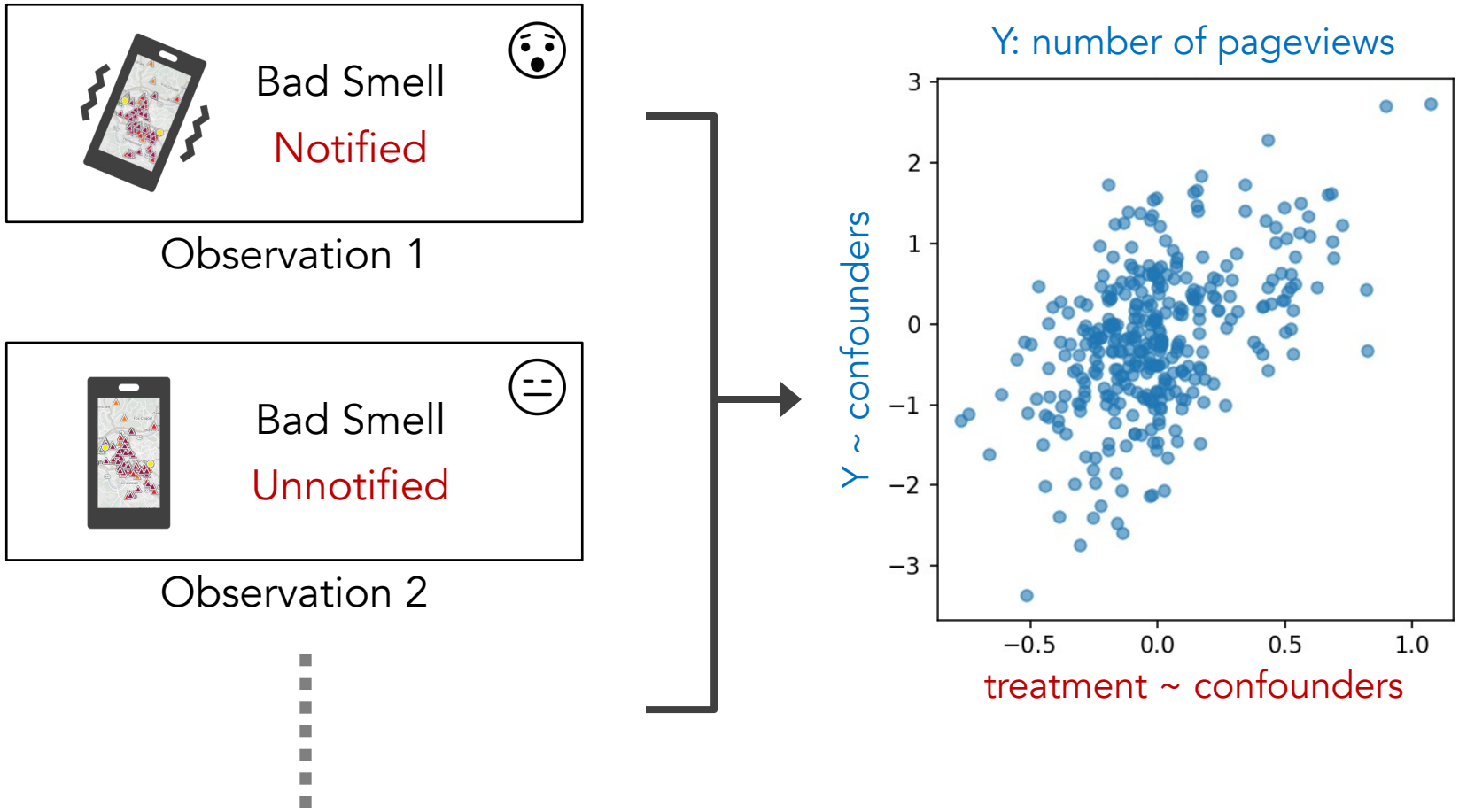
Model Data

A more complicated image data processing task is fine-grained categorization, and we use garbage classification as an example.

Source -- https://www.kaggle.com/datasets/asdasdasasdas/garbage-classification

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.



Hsu, Y. C., et al. (2020). Smell Pittsburgh: Engaging Community Citizen Science for Air Quality. ACM TiiS.

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.

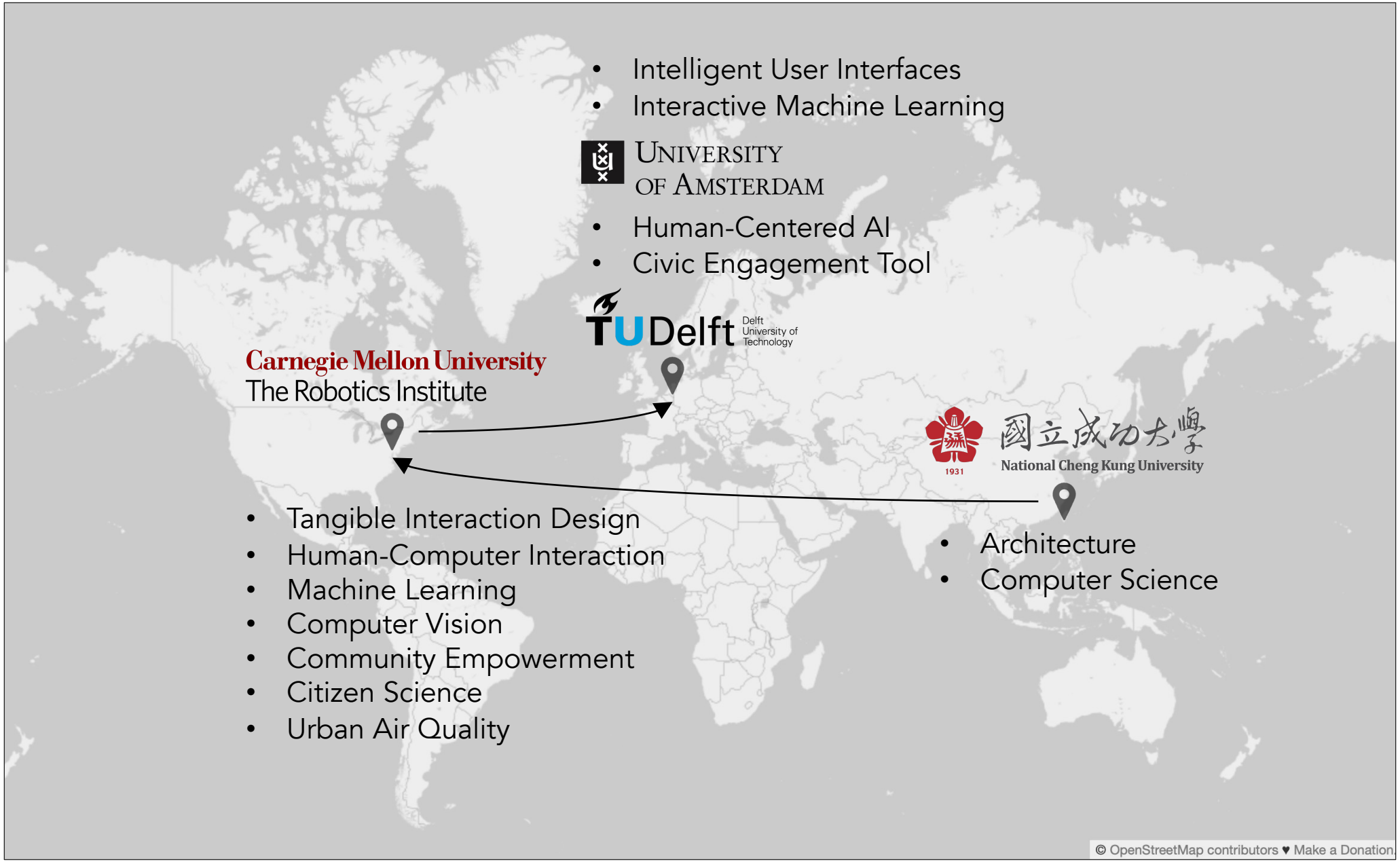
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Retrieved from the Gender Shades project on 2018 -- http://gendershades.org

This course has three teaching team members.

Check Canvas for their contact information.

- Yen-Chia Hsu, course coordinator
- Yahia Dalbah, head of teaching assistants (except the first two weeks)
- Basten Leeftink, teaching assistant (only the first two weeks)



Information about the course instructor -- http://yenchiah.me

This course aims to familiarize you with various data science pipelines (processing structured data, text data, and image data) by alternating between theories and practices.

Schedule Outline

Week	Content	Deadline
1	Introduction + Data Science Fundamentals	
2	Structured Data Processing	
3	Deep Learning Overview + PyTorch	Reflective writing due, Tuesday 23:59
4	Mid-term Exam (no seminars, no lectures)	
5	Mid-term Exam Review (online) + Text Data Processing	
6	Image Data Processing	Reflective writing due, Tuesday 23:59
7	Multimodal Data Processing + Guest Lecture	Reflective writing due, Tuesday 23:59
8	Final Exam (no seminars, no lectures)	

Administration

- Announcements will be made on Canvas.
- Lectures will be streamed and recorded with links on the Canvas home page (quality not guaranteed).
- Lectures may be given virtually if unexpected situations happen, same as seminars.
- Use TicketVise (preferred) or email to ask questions.
- There is no attendance requirement.
- You are expected to treat others with mutual respect and appreciation regardless of any differences.
- It is strongly recommended to stay home if you have symptoms associated with respiratory infections.

Assessment includes two exams (midterm 40%, final 50%) and three reflective writing submissions of assignments (10% total).

Exams

- Check https://multix.io/data-science-book-uva/#schedule-outline about the coverage of exams.
- Exams are based on multiple choice questions. We provide mock exams for you to practice.
- You may bring an A4-size cheat sheet with two sides of content (written or printed) to the exams. No other materials are allowed during the exams (except the cheat sheet).
- Check the date and time of the exams carefully on UvA DataNose.
- No minimum grade requirement for each exam to pass the course. There is a resit, which is 90% weight (will override your original weighted sum of exam scores).
- We will use guess correction for the exams (check the syllabus for details).

Exam Coverage Range

Lecture	Topic	Mid-term	Final	Resit
1	Introduction	yes	yes	yes
2	Data Science Fundamentals	yes	yes	yes
3	Structured Data Processing (Part I)	yes	yes	yes
4	Structured Data Processing (Part II)	yes	yes	yes
5	Deep Learning Overview	yes	yes	yes
6	PyTorch Basics	yes	yes	yes
7	Text Data Processing (Part I)	no	yes	yes
8	Text Data Processing (Part II)	no	yes	yes
9	Image Data Processing (Part I)	no	yes	yes
10	Image Data Processing (Part II)	no	yes	yes
11	Multimodal Data Processing	no	yes	yes

Reflective Writing Submissions for Assignments

- We only grade your reflective writing submissions (pass/fail) but not the assignments.
- Use the reflective writing template that we provide for submissions.
- You need to show that you have done the assignment by explaining what you have learned.
- We only accept submissions on Canvas (no email submissions).
- Check the syllabus for the late submission policy (automatic deduction of 10% max points per day).
- Assignment materials can appear in the mid-term exam, final exam, and resit.
- Share your GenAI usage experiences for this course in the reflective writing.

Important Notes on Expectations

- We do not teach programming. Instead, we teach how to do data science using programming and machine learning techniques. Basic concepts of machine learning will be covered.

- We do not aim to cover everything in data science. Instead, we introduce ways of doing data science that enable students to study further in relevant topics.

- We do not teach you data collection. Instead, we assume that someone has collected the datasets.

- We do not teach data science in production. Most of the techniques that are covered in this course are for the development environment.

- We expect students to have good Python programming skills already. The Python Coding Warm-Up practice reflects our expectations.

GenAI Usage and Policy

- We allow students to use GenAI tools, such as ChatGPT, Gemini, Claude, UvA AI Chat, etc.
- GenAI tools can hallucinate information, generate wrong content, or provide fake sources/citations. You are responsible for fact-checking and verifying all AI-generated content.
- The official lecture slides and course readings remain the single source of truth. If the AI generates content that contradicts the course materials, you must follow the course materials.
- The teaching team is neither responsible nor liable for your exam errors resulting from reliance on unverified AI-generated content.

Math, Coding, and Course Readings in the Exam

- If a math equation is on the slide, and we explained it in the lecture, it means that the equation may appear in the exam (in different ways). You also need to know what they are doing at a high level.
- Multiple-choice questions in the exam can contain coding-related questions. Only the code in the "Task Answers" pages for all modules will be covered in the exam (check the syllabus for details).
- This course has many readings. The exams will be based on the lecture slides (primarily) and the required course readings (secondly). Optional course readings are used for GenAI fact-checking.
- If a concept appears on the lecture slides, it may appear on the exam, and the course readings are used to support the concept with more explanations and details.
- If a concept appears in the course readings but not on the slides, it will not appear on the exam.

For more details, check the course website and the syllabus page below:

- https://multix.io/data-science-book-uva/
- https://multix.io/data-science-book-uva/syllabus.html

Questions?