

# Data Science

Lecture 1-1: Course Introduction and Organization

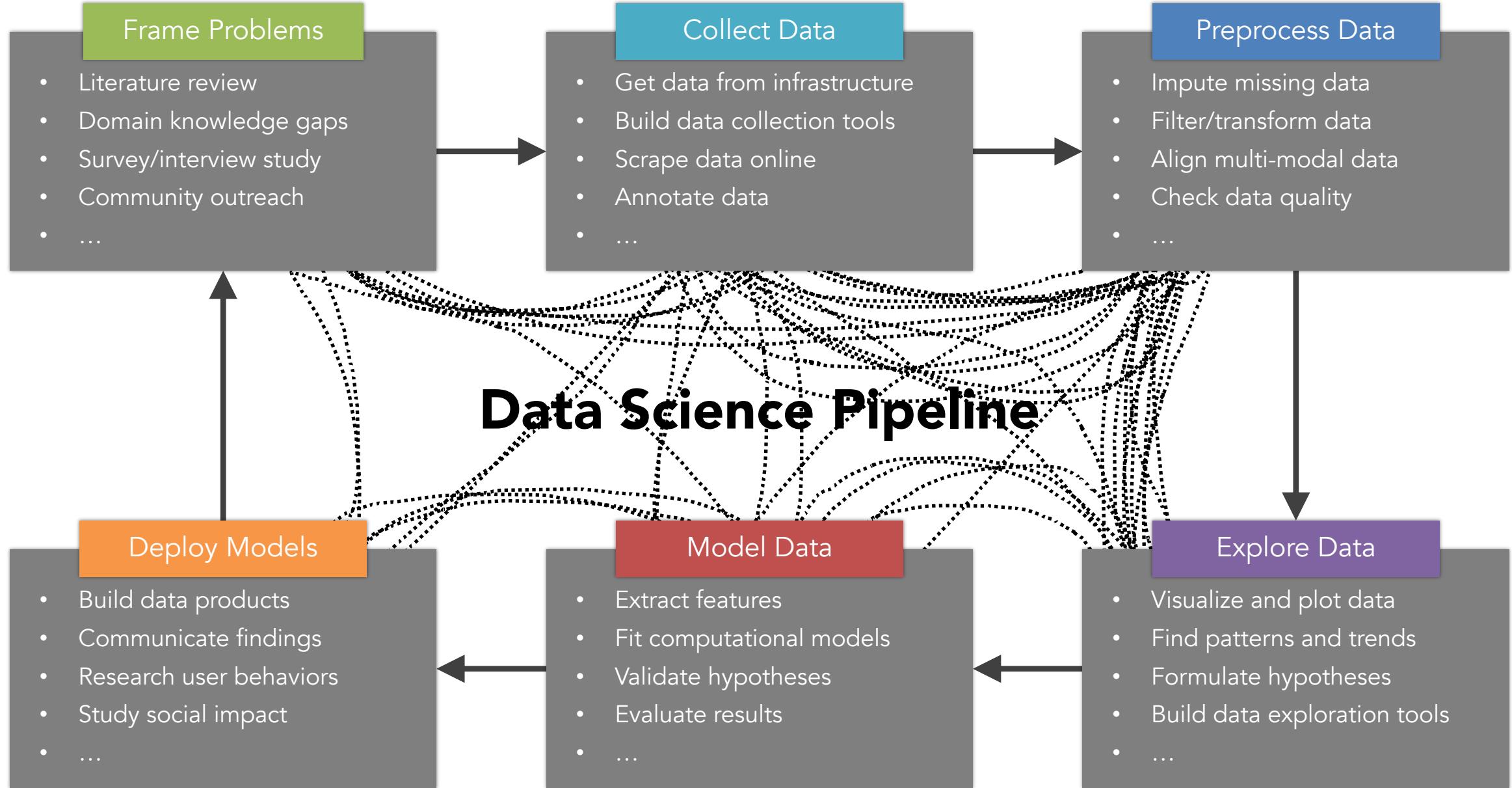


UNIVERSITY  
OF AMSTERDAM

Lecturer: Yen-Chia Hsu

Date: Feb 2026

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



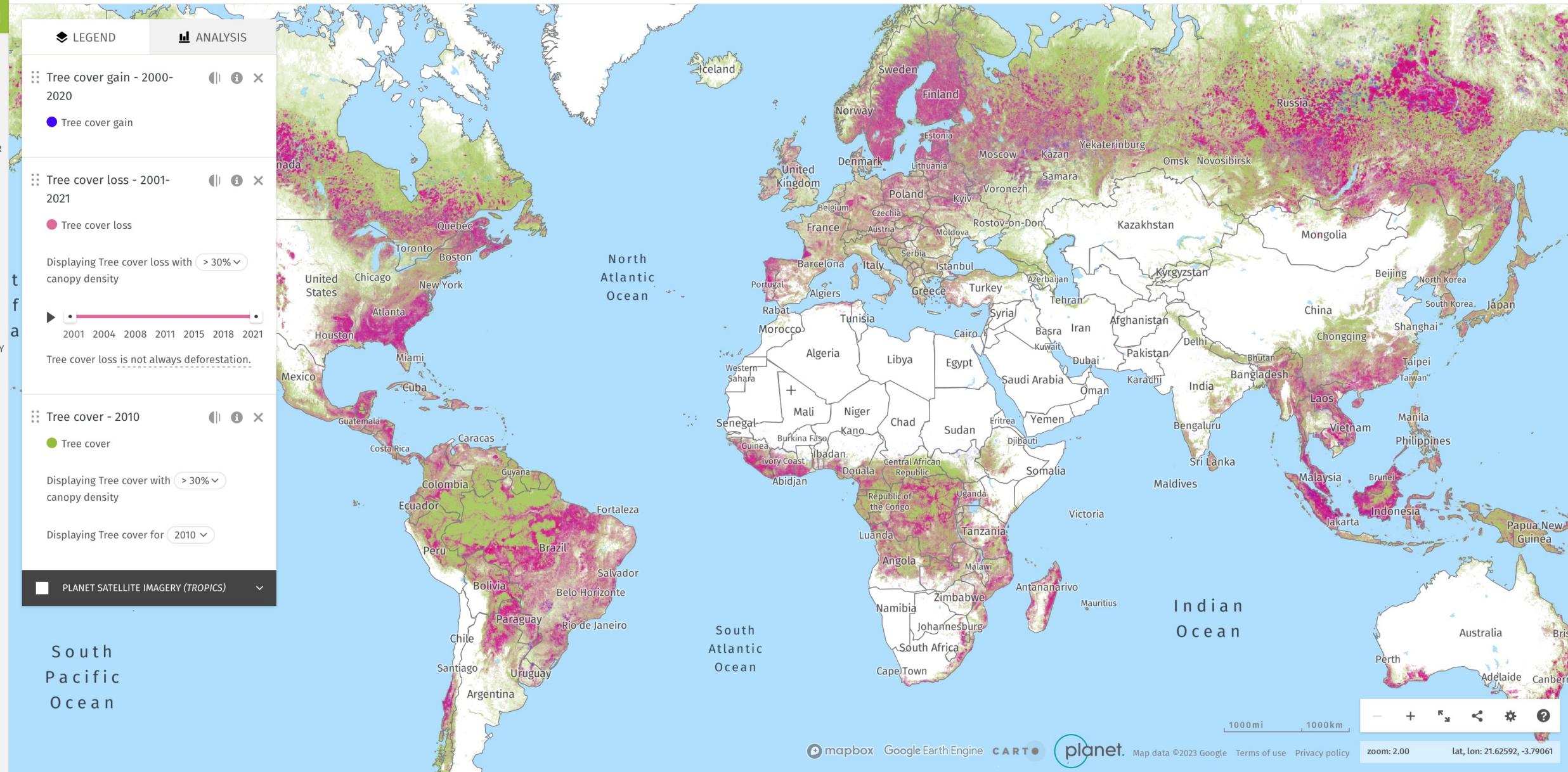
What people typically think : →

The reality of the data science pipeline: ..... 3

## 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: analyzing the situation of global forest situation -- <https://www.globalforestwatch.org/map/>

Electronics Deals Week Last day

< Back to results



## Nintendo Switch Console, Grijs (Nintendo Switch)

Visit the Nintendo Store

Platform: Nintendo Switch

4.5 stars 15,331 ratings

€329<sup>99</sup>

All prices include VAT.

Platform For Display: Nintendo Switch

Edition: Grijs

Grijs Rood/Blauw

### About this item

- Slimme keuze voor dagelijkse behoeften
- Gemakkelijk mee te nemen, compact ontwerp
- Gemaakt met de nieuwste technologie
- De tool voor een reeks creatieve activiteiten voor iedereen
- Je favoriete content staat altijd op de voorgrond
- Een visuele ervaring van hoge kwaliteit

€329<sup>99</sup>

€9.11 delivery February 6 - 9.

Details

Select delivery location

In stock.

Quantity: 1

Add to Basket

Buy now

Secure transaction

Dispatches from TechLead NL  
Sold by TechLead NL

Add to List

Add other items:

### What other items do customers buy after viewing this item?



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



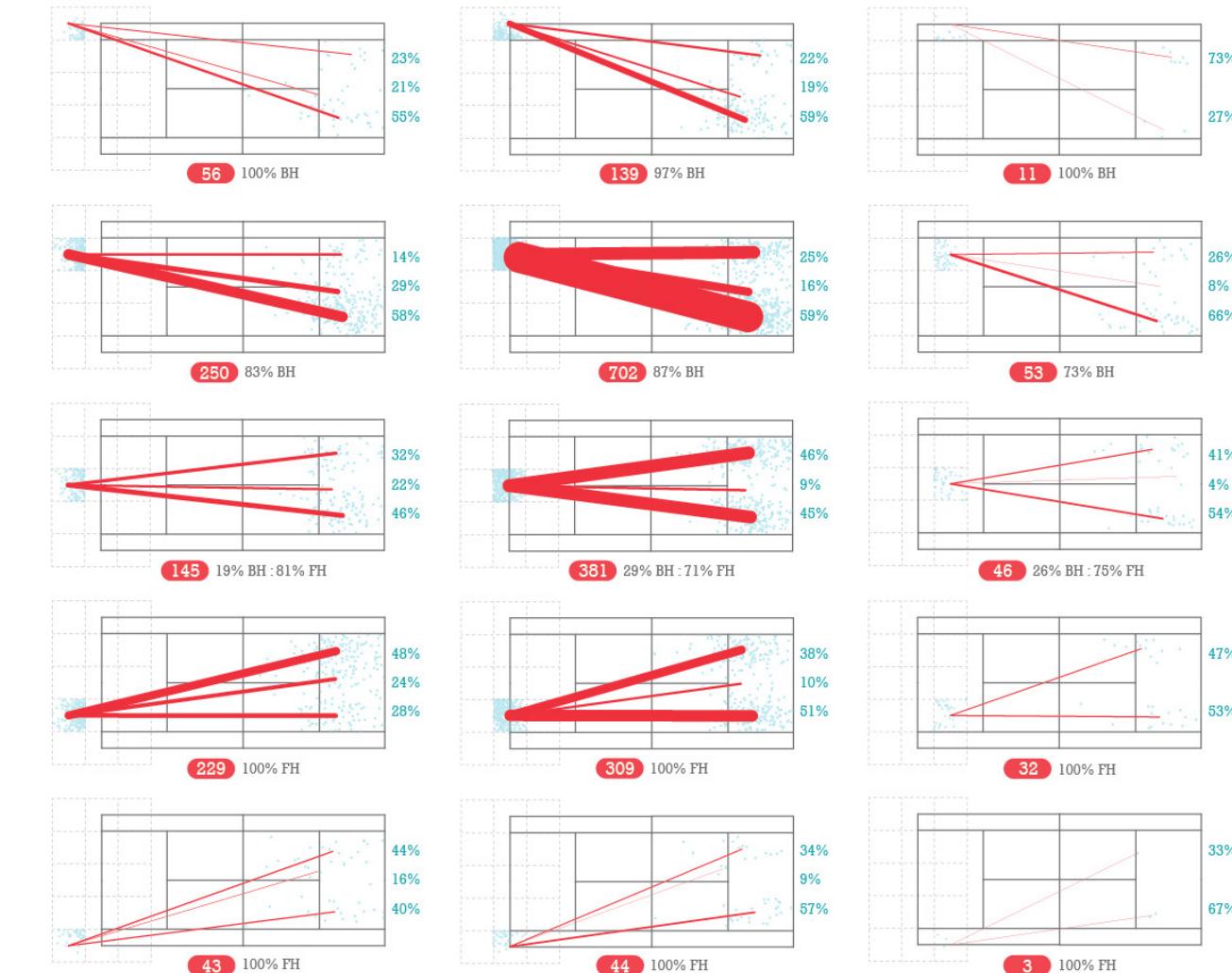
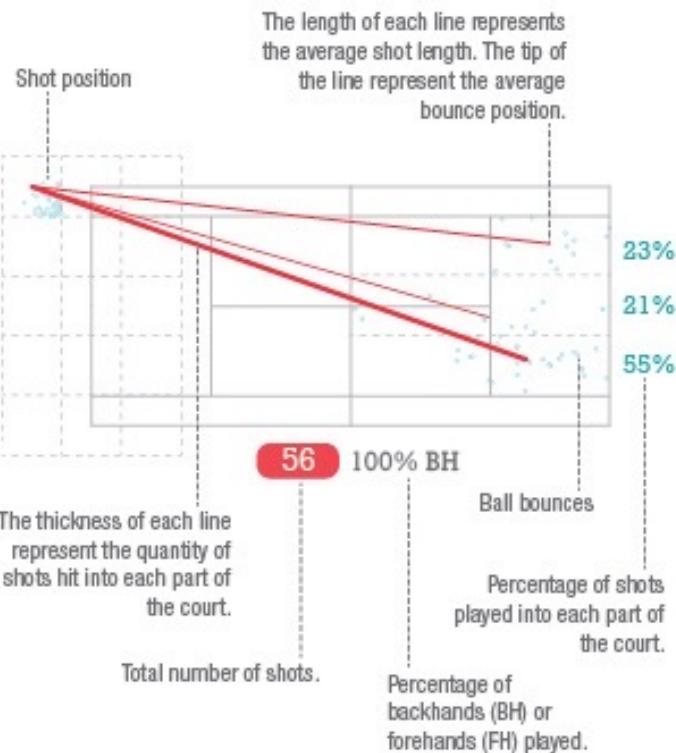
Page 1 of 6



# 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.

The screenshot shows the 'Populaire kaartlagen' (Popular map layers) section of the GGD Amsterdam Data Portal. It features three thumbnail images: a panoramic view of Amsterdam, a canal scene with buildings, and a street view. Below each image is a title and a brief description:

- Kadastrale perceelsgrenzen**: De Basisregistratie kadaster (BRK) bevat informatie over kadastrale objecten (percelen en appartementsrechten). De kaartlaag kadastrale perceelsgrenzen toont de kadastrale grenzen en de grenzen van de gemeente Amsterdam.
- Meetbouten - Zakkingsnelheid**: De registratie meetbouten bevat de meetgegevens van boutingen in de Amsterdamse panden. Het doel van de meetbouten is het monitoren van de deformatie (zakkingen of stijgingen). Dit wordt inzichtelijk gemaakt op deze kaartlaag.
- Parkeren - Fiscale indelingen**: De kaartlaag parkeren - fiscaal geeft visueel informatie over de fiscale indeling in Amsterdam. Krijg via de kaart inzicht in waar er fiscaal parkeren en waar er niet-fiscaal parkeren van toepassing is.

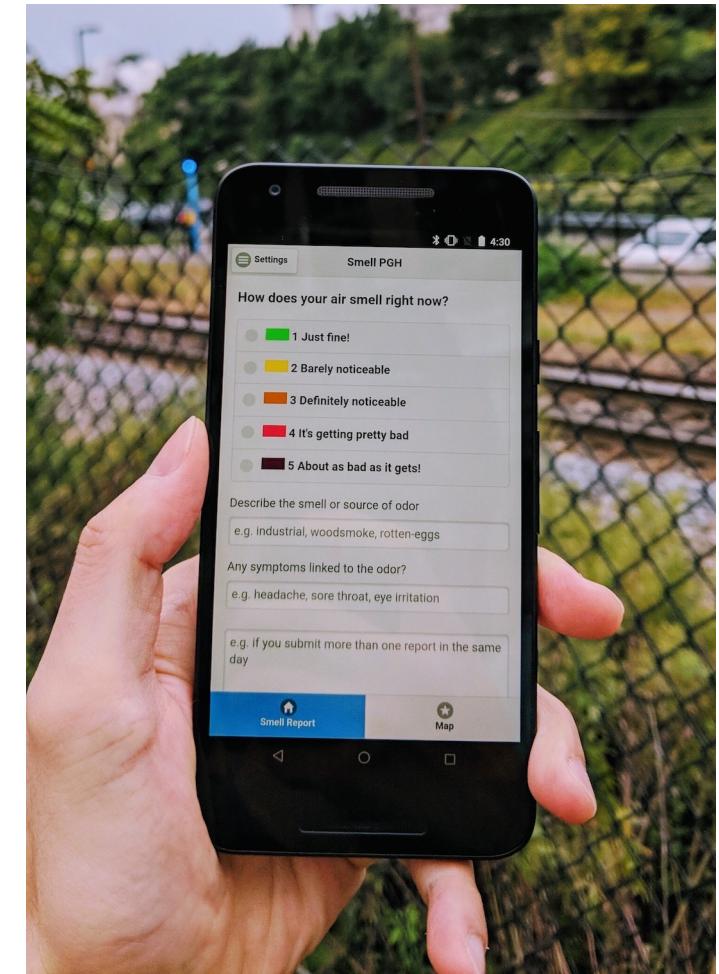
At the bottom, there are links for 'Vragen' (Questions), 'Colofon' (Colophon), and 'Volg ons' (Follow us) with icons for Twitter, Facebook, LinkedIn, and YouTube.

The screenshot shows the Prolific platform interface. The top banner reads 'Set your study live to thousands of reliable participants in minutes.' with a 'Get started' button. Below, a card displays a study titled 'Testing your memory of a crime scene' with the following details:

- £7.50/hour
- 7 mins
- 500 places

Three participant profiles are listed below:

- Chicago, United States**: Time taken: 8 mins
- London, United Kingdom**: Time taken: 10 mins
- Washington, United States**



[GGD Amsterdam Data Portal](#)

[Prolific Tool for Data Annotation](#)

[Mobile App Data Collection](#)

## Collect Data

Hugging Face, Zenodo, Google Dataset Search, government websites, etc.

The Hugging Face dataset search interface shows a list of datasets. Key datasets listed include:

- glue
- openwebtext
- blimp
- imdb
- super\_glue
- red\_caps
- HuggingFaceM4/cm4-synthetic-testing
- wikitext
- textvqa
- squad

[Hugging Face](#)

The Zenodo research shared interface features a "Featured communities" section with a NASA Transform to Open Science badge. It also displays "Recent uploads" and a "Curated by: nasatransformtoopen" section. Recent uploads include:

- Flowminder/FlowKit: 1.18.2
- Trixi.jl

[Zenodo](#)

The Google Dataset Search interface shows results for "sustainability". Key findings include:

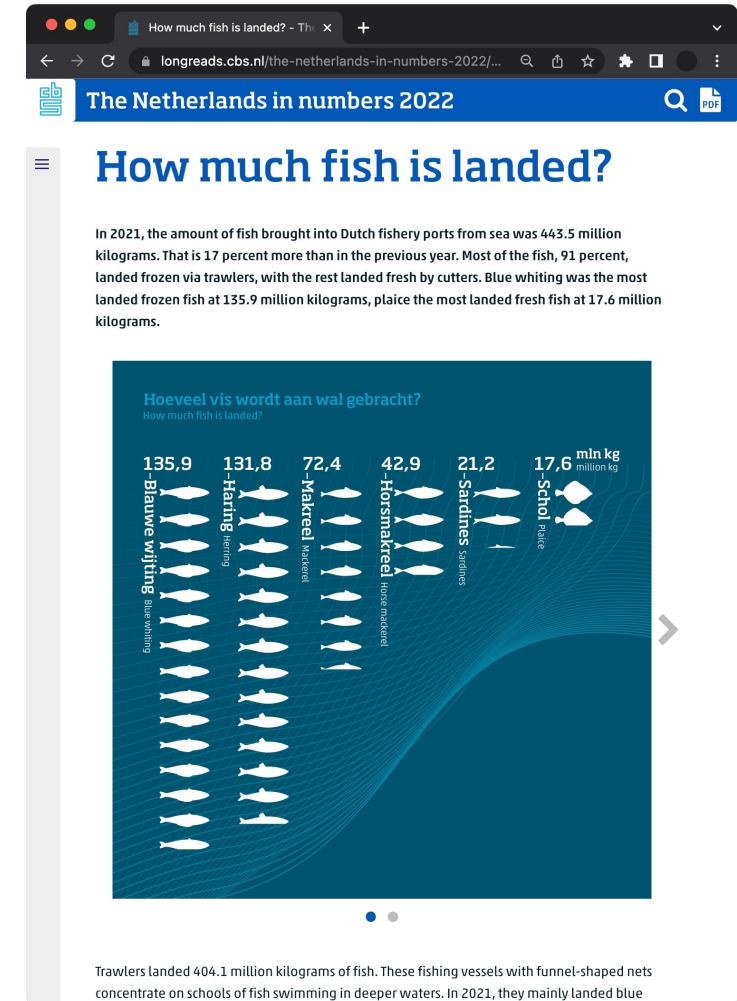
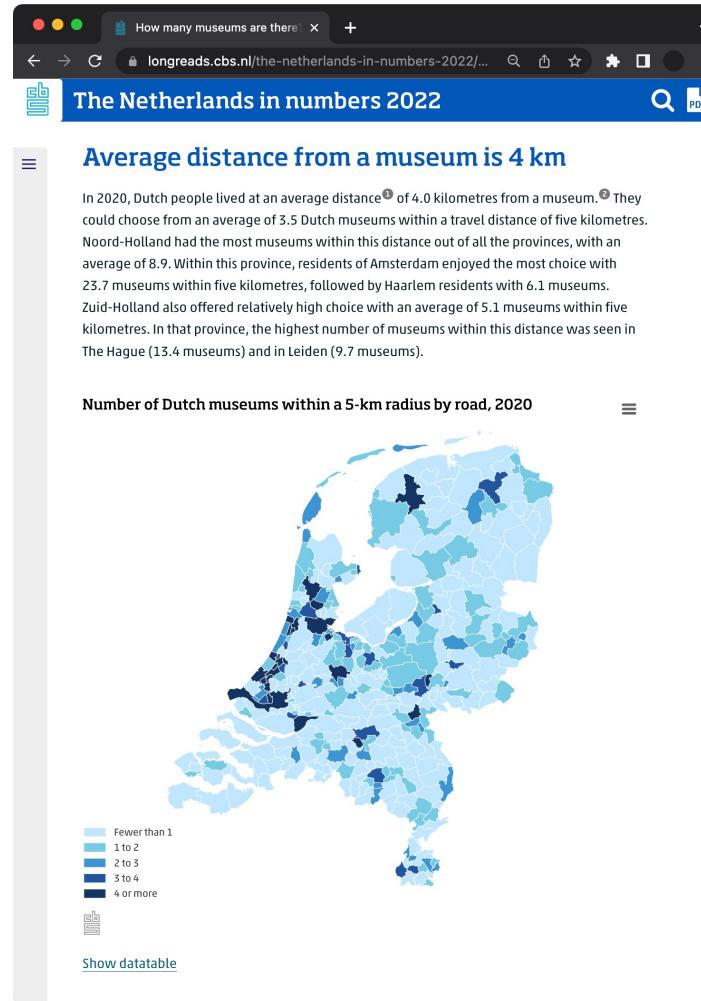
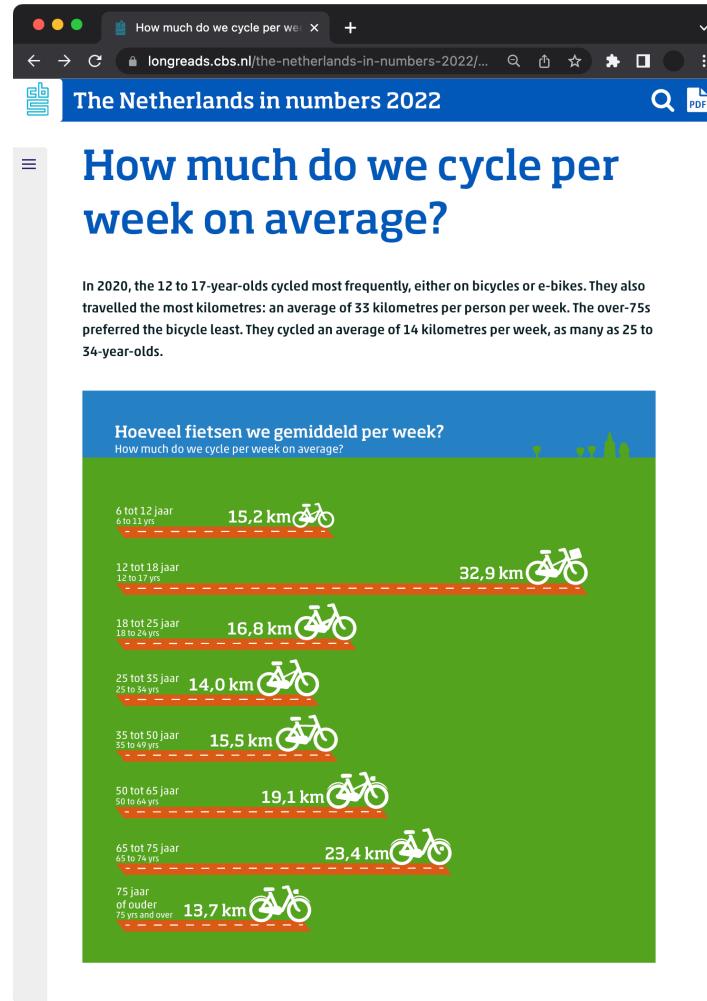
- statista: Data sustainability as main consideration in global organizations 2020, by country
- csiro: Australian land-use and sustainability data: 2013 to 2050
- bloomberg: Environmental, Social and Governance Data
- statista: Sustainability reporting rate 2020, by sector

[Google Dataset Search](#)

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

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

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



## Explore Data

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

The screenshot shows the official Seaborn website at [seaborn.pydata.org/index.html](https://seaborn.pydata.org/index.html). The page features a header with the Seaborn logo, navigation links for Installing, Gallery, Tutorial, API, Releases, Citing, and FAQ, and social media icons for search, GitHub, YouTube, and Twitter. Below the header is a section titled "seaborn: statistical data visualization" displaying six examples of Seaborn plots: a joint plot with marginal distributions, a density plot with multiple layers, a scatter plot with a regression line, contour plots for two years (1955 and 1958), a box plot with violin plots, and a scatter plot with a linear regression model. The main content area includes a brief introduction, links to introductory notes and a paper, installation instructions, and links to the example gallery, tutorials, API reference, releases, citations, and frequently asked questions. A sidebar on the right lists "Contents" (Installing, Gallery, Tutorial, API, Releases, Citing, FAQ) and "Features" (Objects, Relational plots, Distribution plots, Categorical plots, Regression plots, Multi-plot grids, Figure theming, Color palettes).

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.

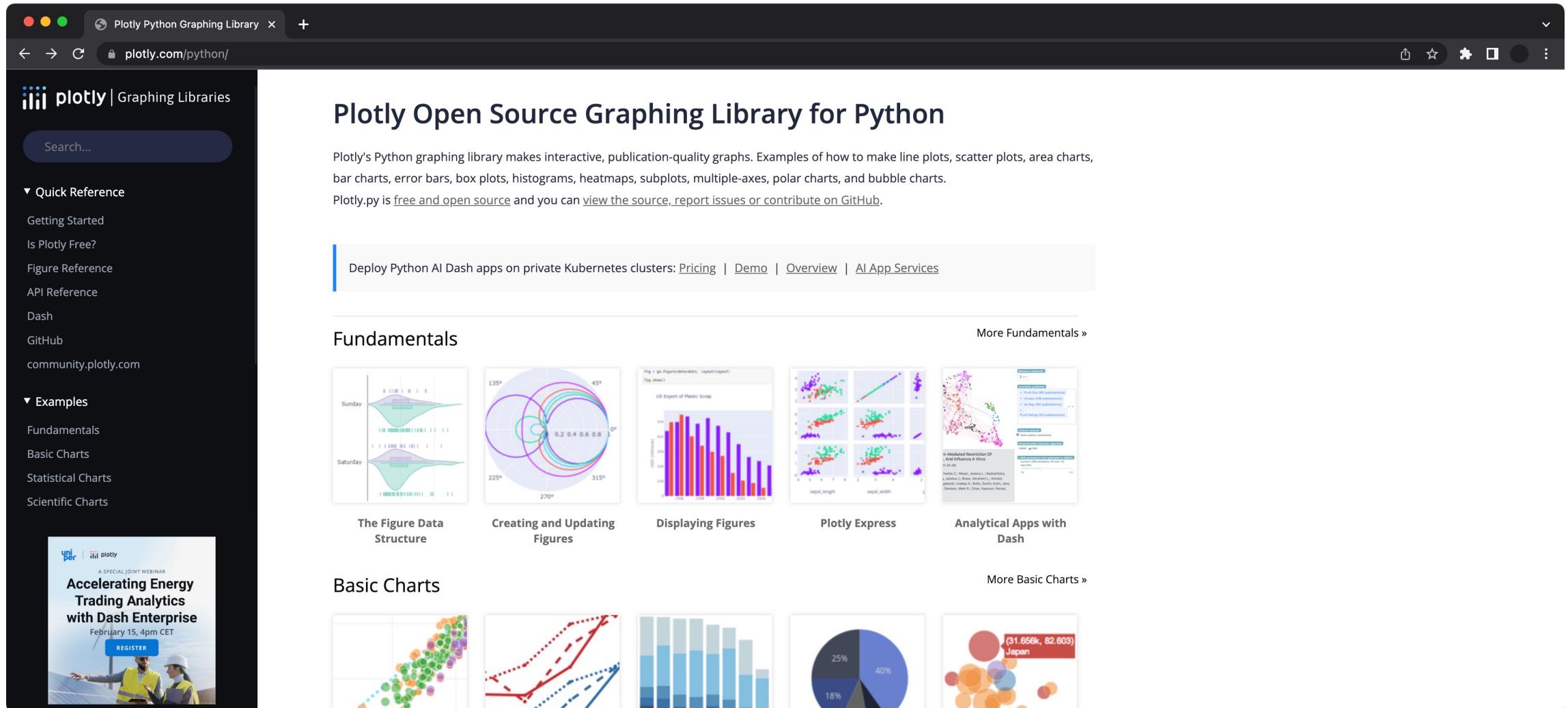
To see the code or report a bug, please visit the [GitHub repository](#). General support questions are most at home on [stackoverflow](#), which has a dedicated channel for seaborn.

**Contents**

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

**Features**

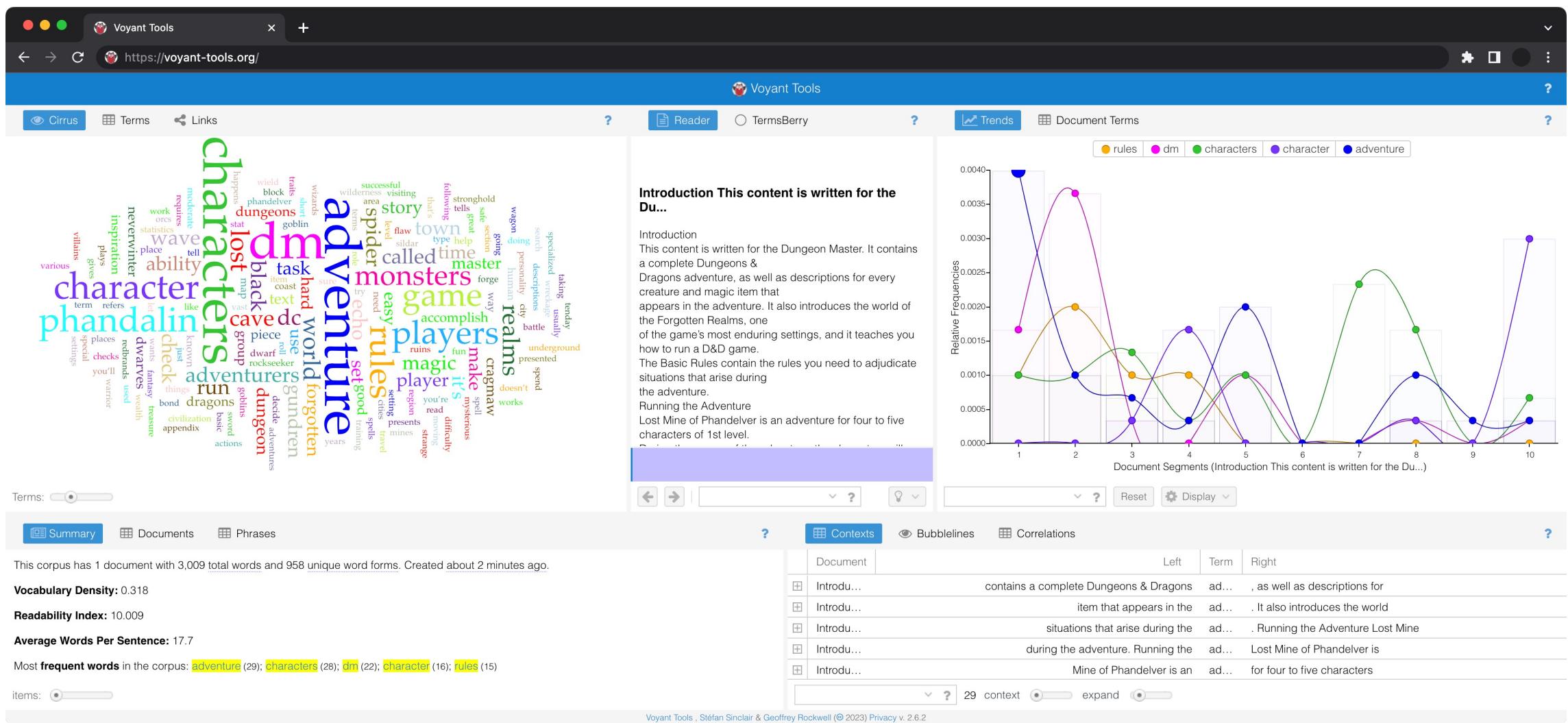
- **New** Objects: [API](#) | [Tutorial](#)
- Relational plots: [API](#) | [Tutorial](#)
- Distribution plots: [API](#) | [Tutorial](#)
- Categorical plots: [API](#) | [Tutorial](#)
- Regression plots: [API](#) | [Tutorial](#)
- Multi-plot grids: [API](#) | [Tutorial](#)
- Figure theming: [API](#) |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



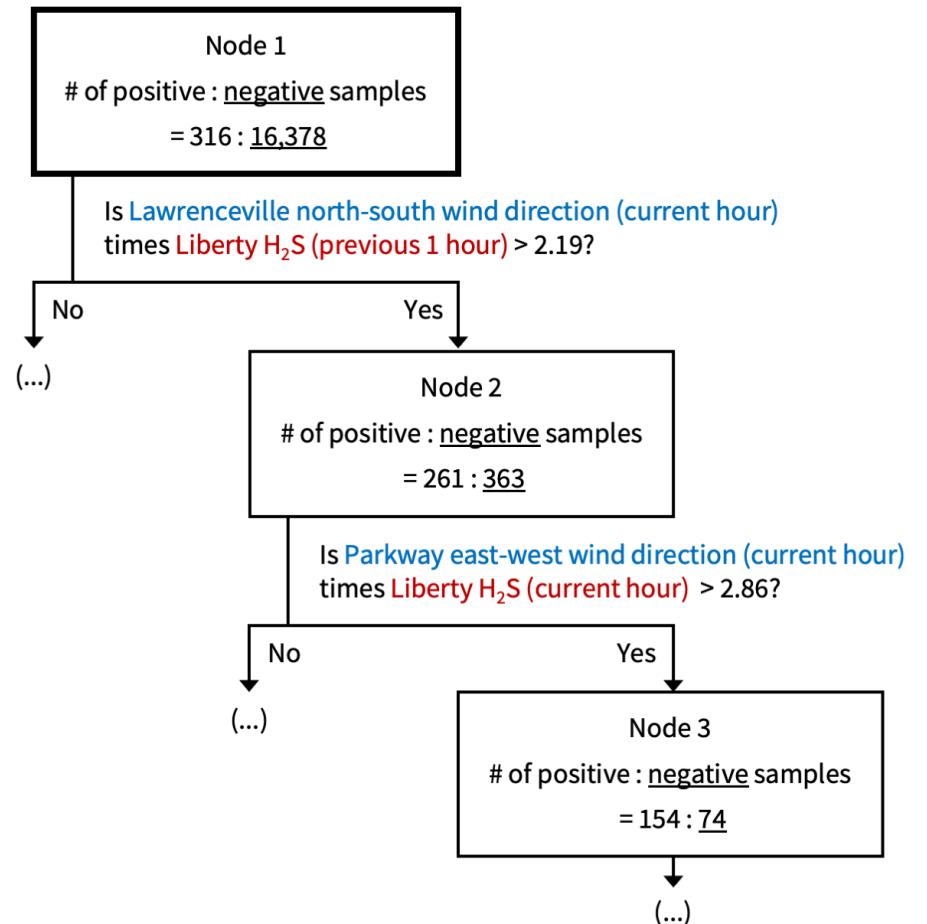
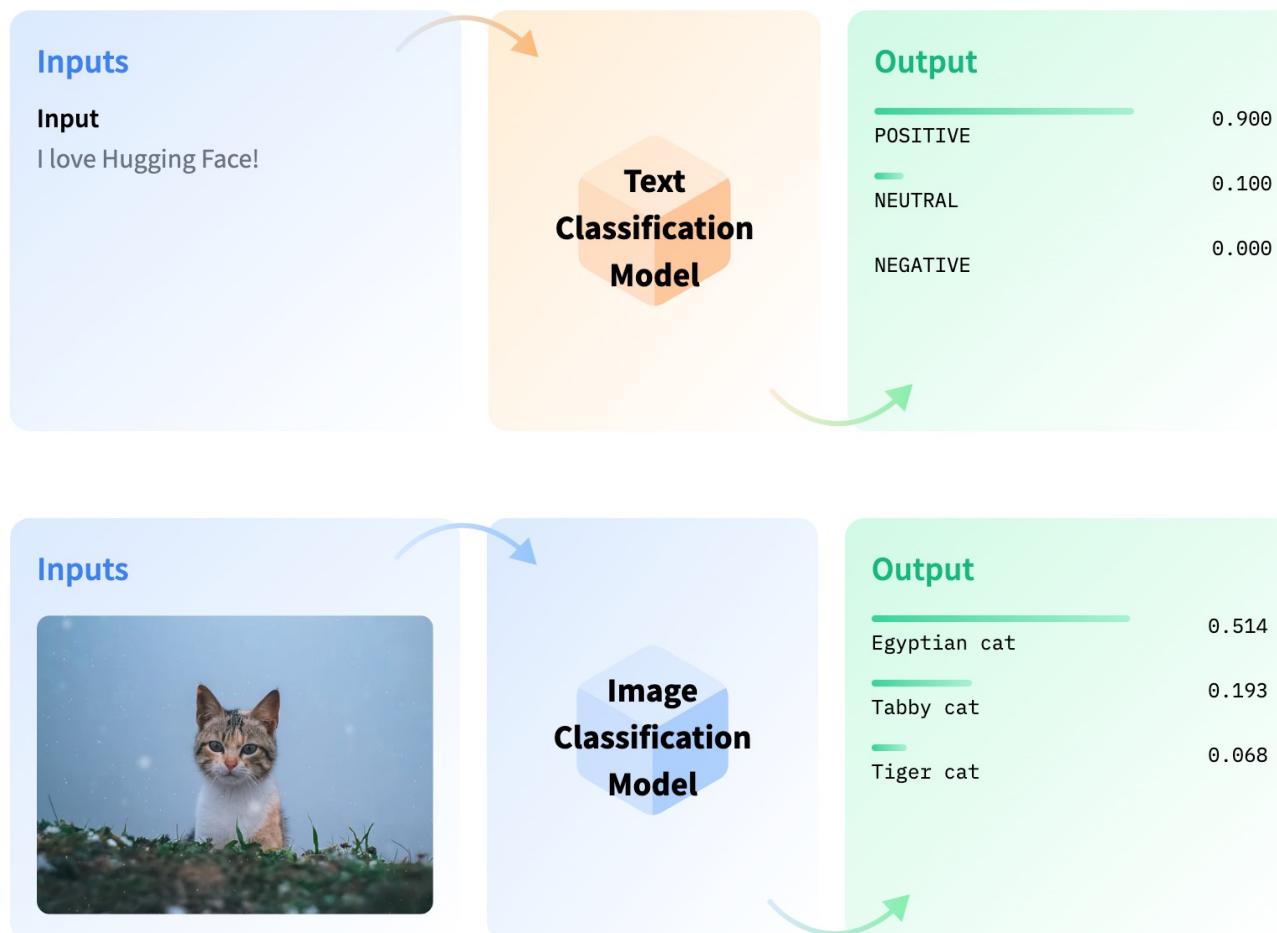
The screenshot shows the official Plotly Python Graphing Library website at <https://plotly.com/python/>. The page has a dark-themed header with the Plotly logo and navigation links for "Search...", "Quick Reference", "Examples", and "Fundamentals". A sidebar on the left lists "Getting Started", "Is Plotly Free?", "Figure Reference", "API Reference", "Dash", "GitHub", and "community.plotly.com". A promotional banner for a joint webinar with UniPer is visible. The main content area features sections for "Fundamentals" (with examples like a sunburst chart, a Venn diagram, a histogram, and a scatter plot grid), "Basic Charts" (with examples like a bubble chart, a line chart with multiple series, a bar chart, and a pie chart), and "More Fundamentals" and "More Basic Charts" sections. A callout box in the center promotes deploying AI Dash apps on private Kubernetes clusters with links to Pricing, Demo, Overview, and AI App Services.

## Explore Data

You can use the Voyant Tools to explore text 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.



## Model Data

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

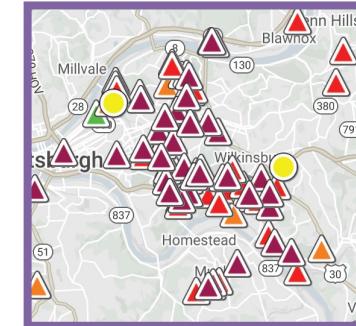
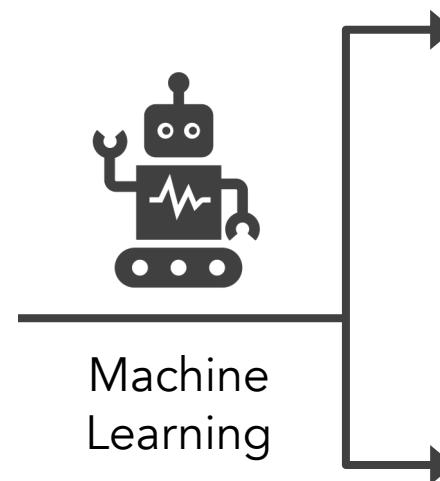
O <sub>3</sub> : 26 ppb	CO: 127 ppb
H <sub>2</sub> S: 0 ppb	PM <sub>2.5</sub> : 9 µg/m <sup>3</sup>
Wind: 17 deg	...

Observation 1

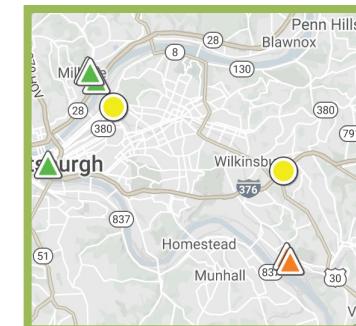
O <sub>3</sub> : 1 ppb	CO: 1038 ppb
H <sub>2</sub> S: 9 ppb	PM <sub>2.5</sub> : 23 µg/m <sup>3</sup>
Wind: 213 deg	...

Observation 2

⋮



😢 Has Event



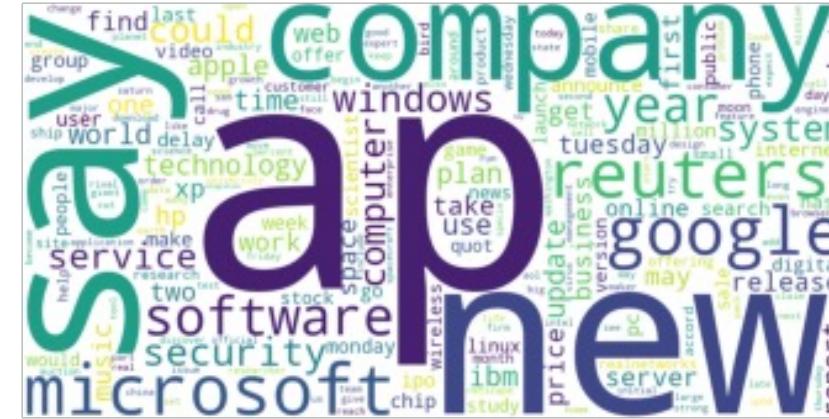
😊 No Event

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

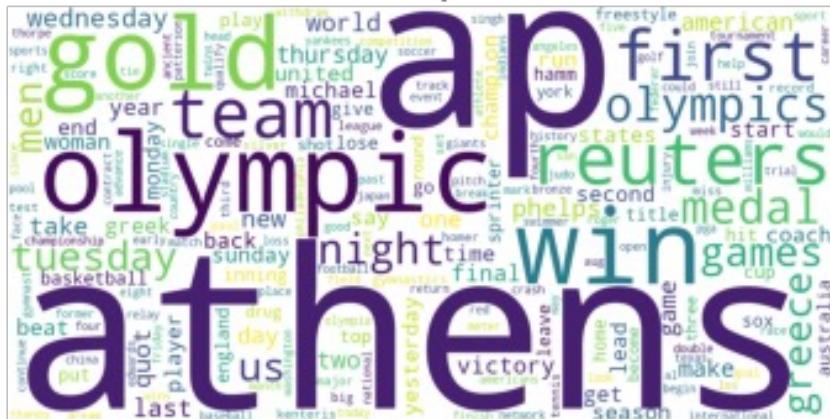
Class: Business



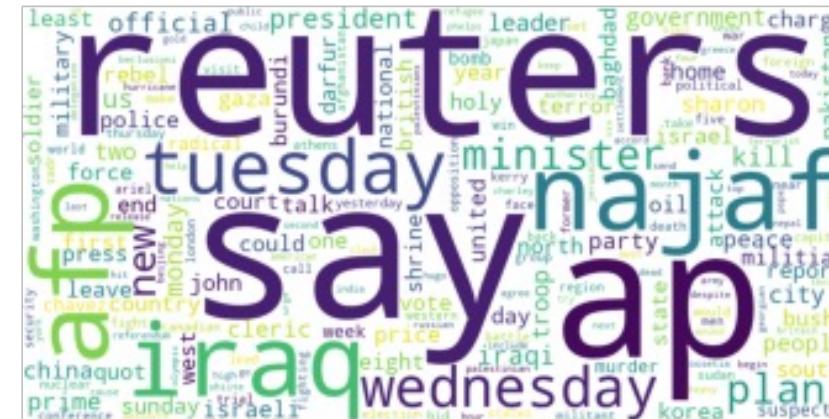
Class: Sci/Tech



## Class: Sports

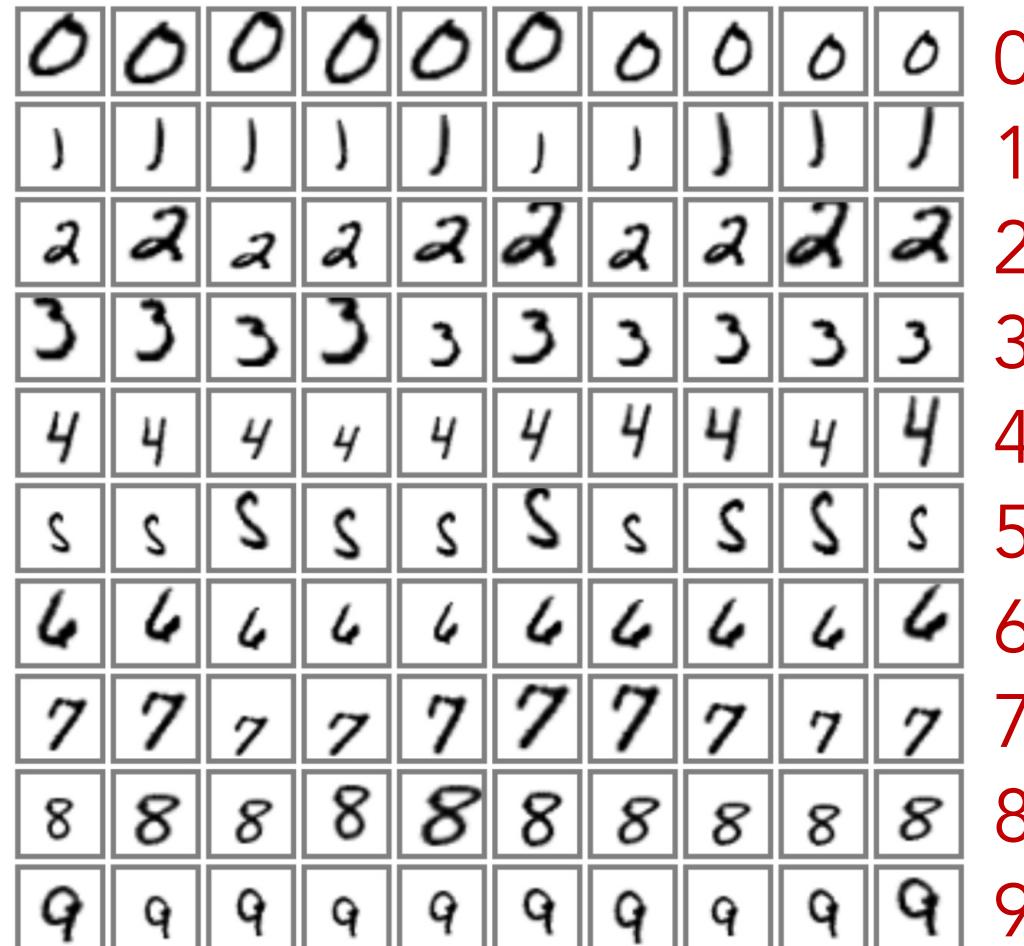


Class: World



## Model Data

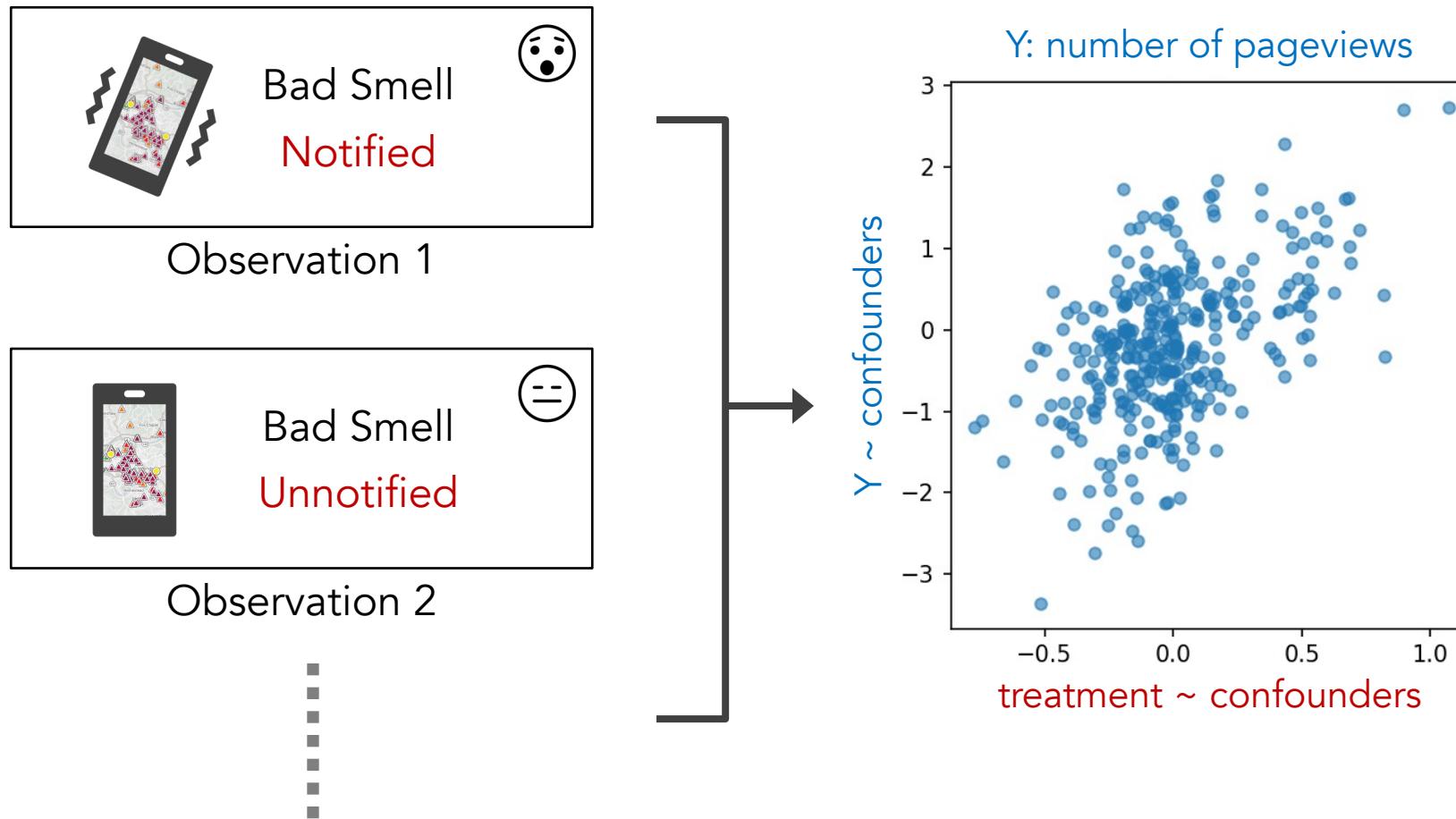
The image data processing module uses deep neural networks to recognize digits from hand-written images.



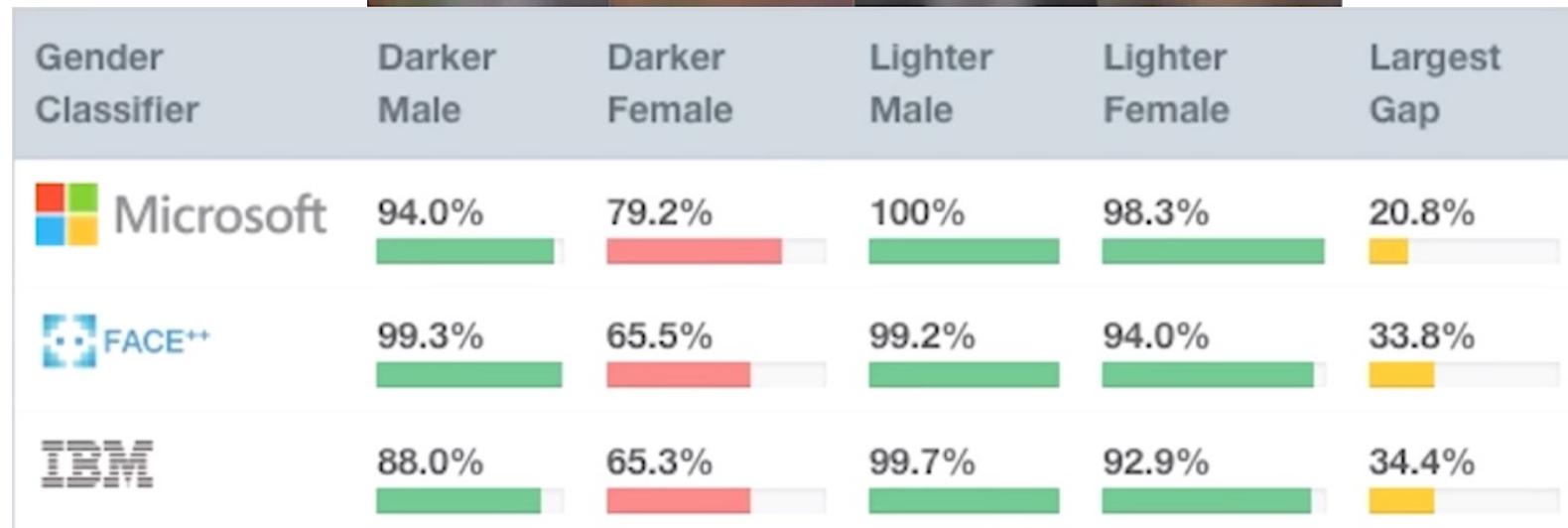
A more complicated image data processing task is **fine-grained categorization**, and we use garbage classification as an example.



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



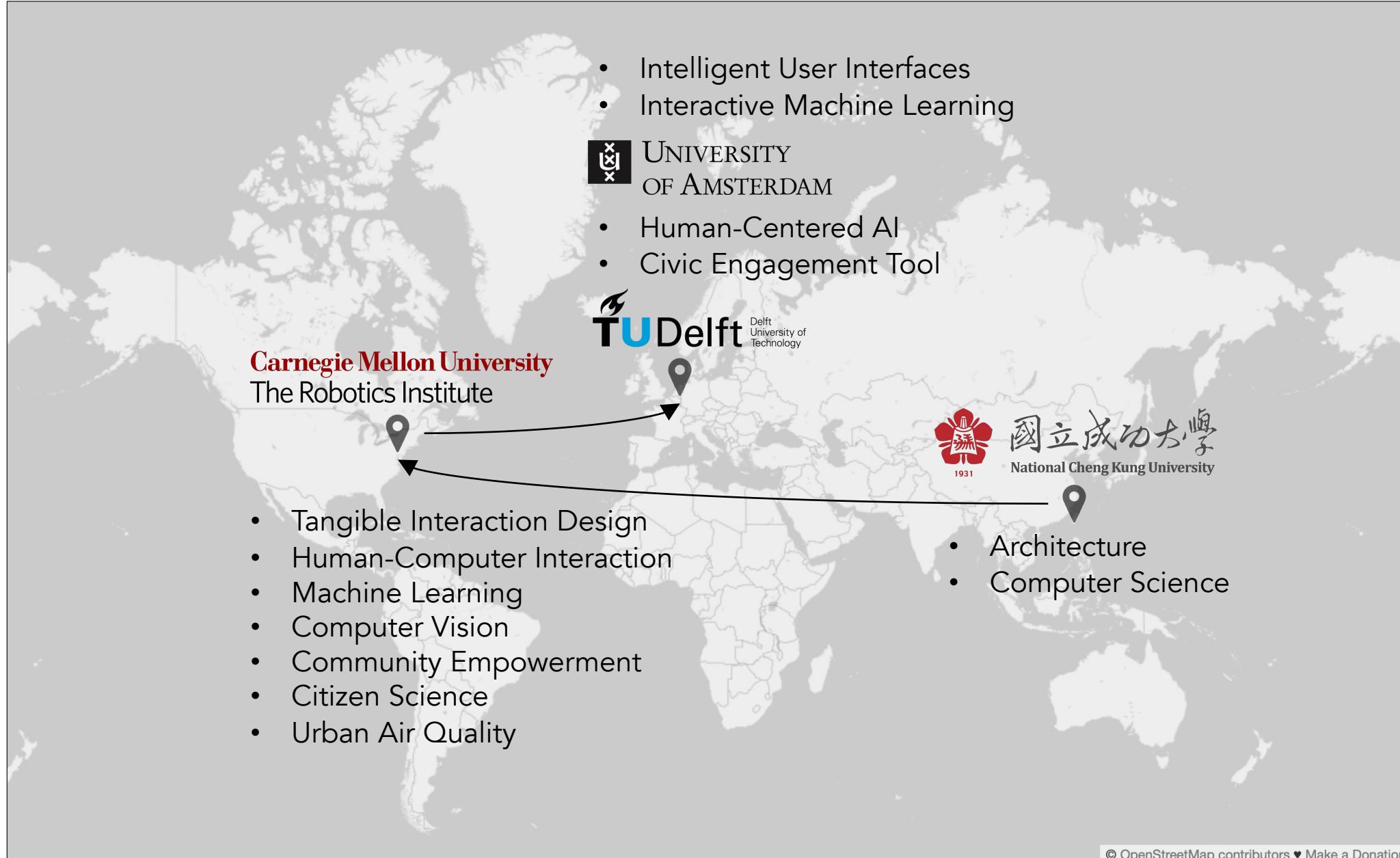
Another example is to study the behaviors of deployed systems to understand its social impact. For example, face recognition systems for recognizing gender did **worst on darker-skin female images**.



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)



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?