Data Science

Lecture 1-1: Course Introduction and Organization

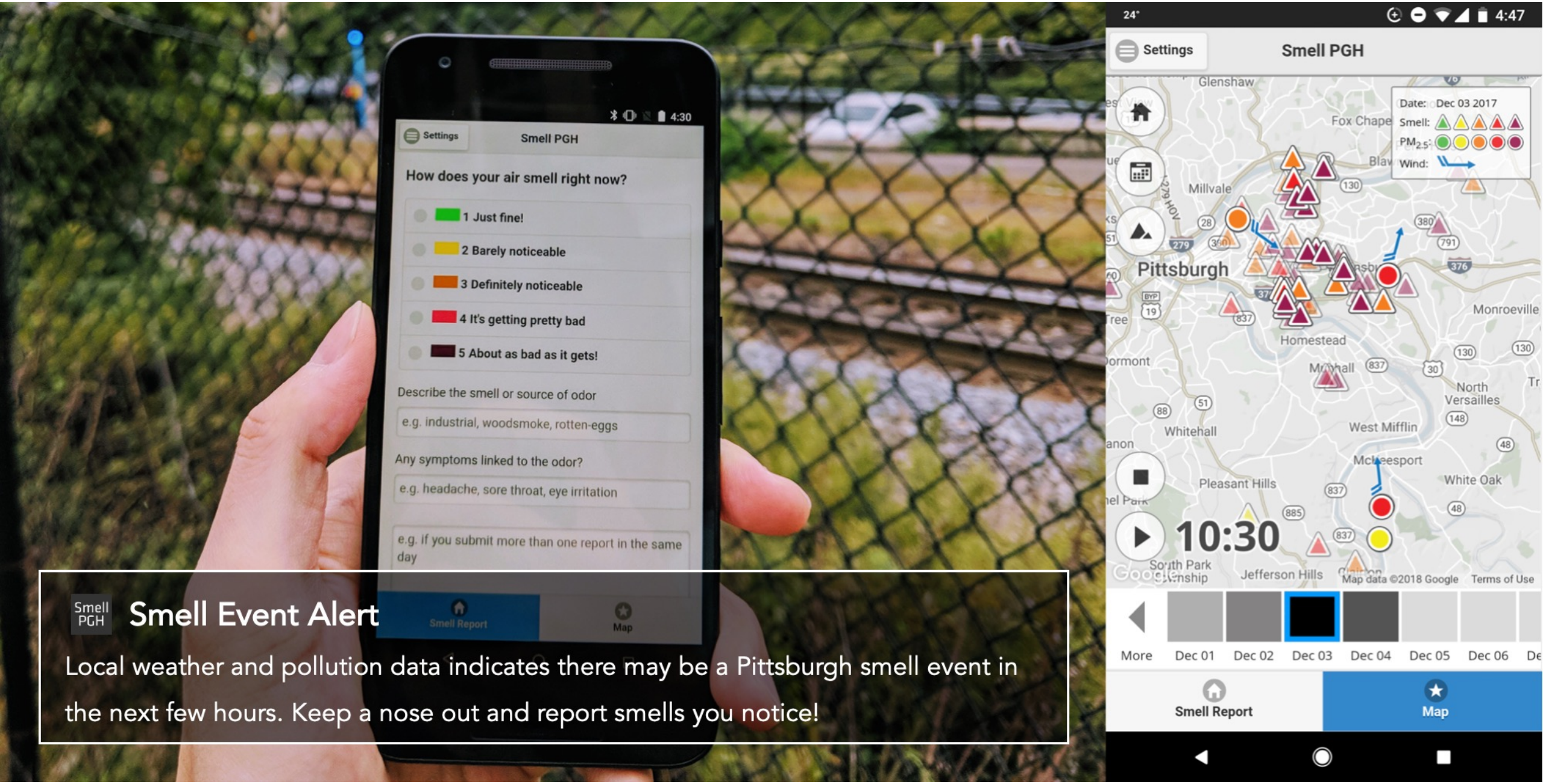
University of Amsterdam

Lecturer: Yen-Chia Hsu

Date: Feb 2025

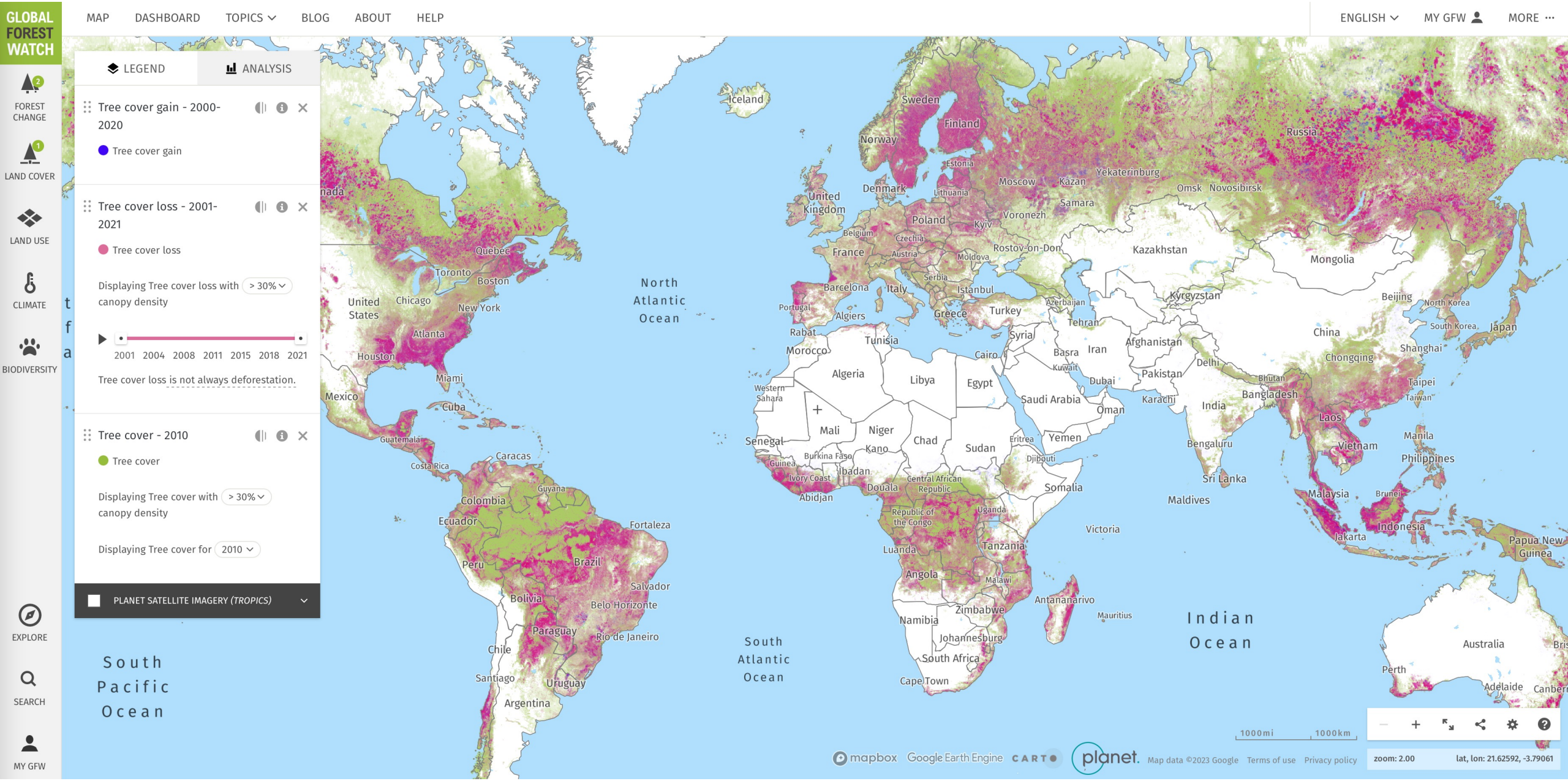
Data science is about turning rich data into actionable insight and making data impactful!

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.

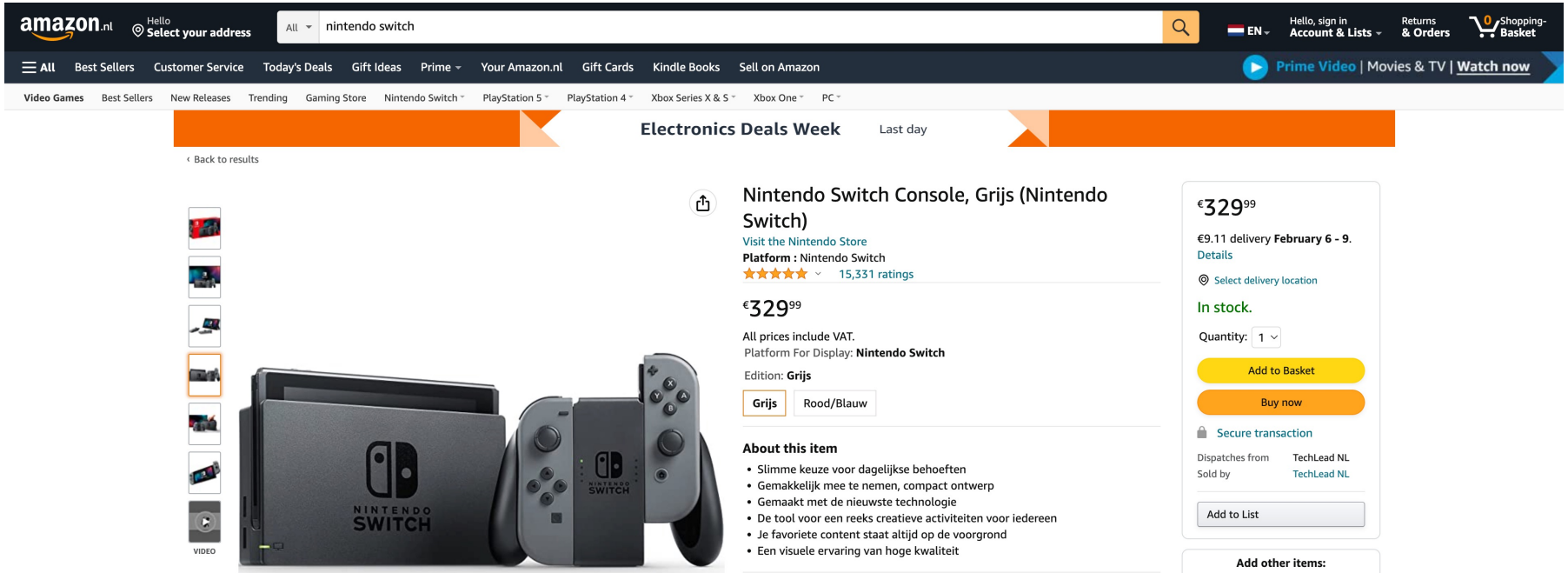


Example: making sense of air pollution data in Pittsburgh -- https://smellpgh.org/

Example: understanding parking patterns and law enforcement -- https://algoritmeregister.amsterdam.nl/en/automated-parking-control/



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

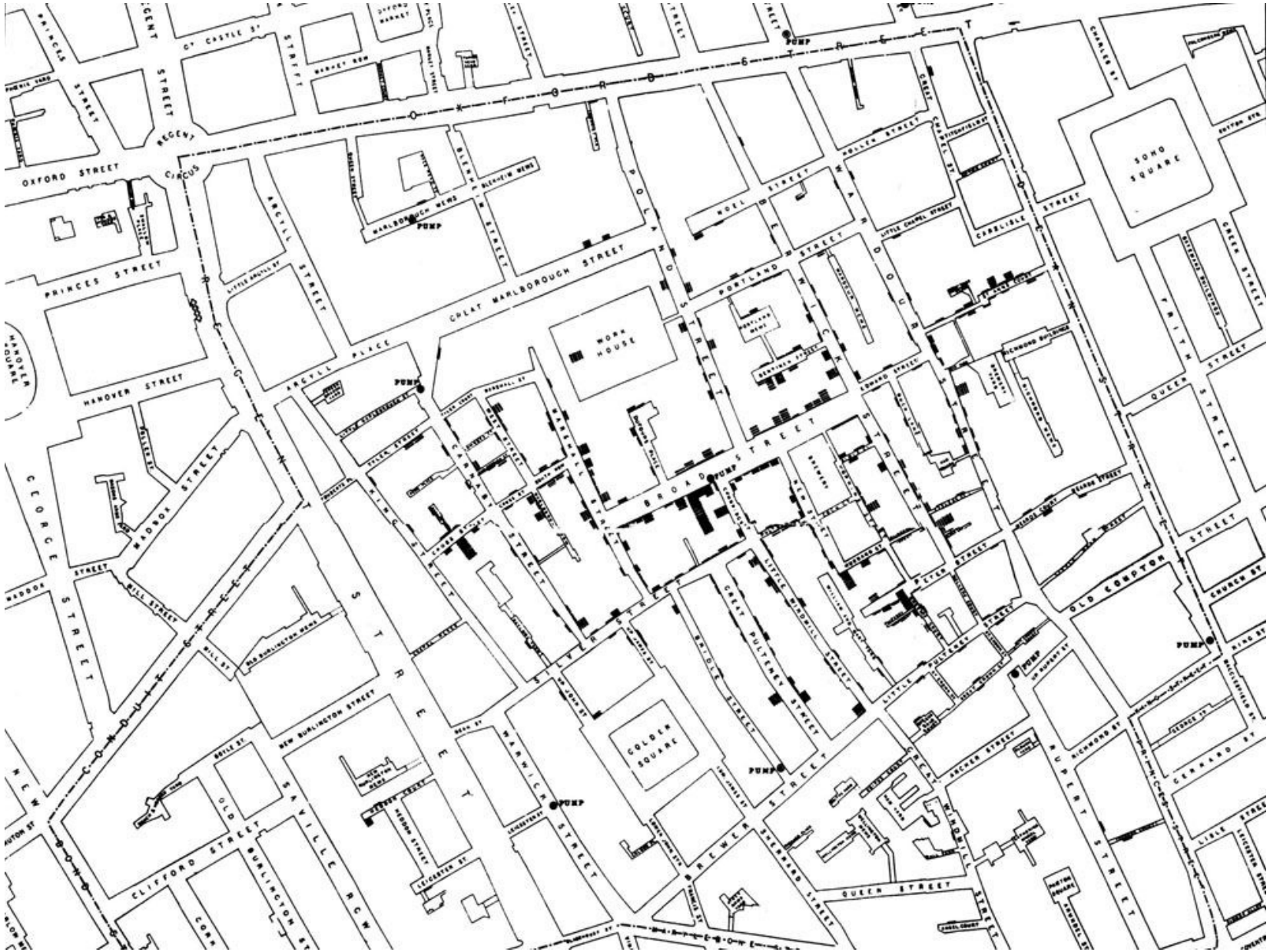


Example: recommending products to customers -- https://www.amazon.nl/

Cholera Map

This map of London was created by John Snow in 1854. London was experiencing a deadly cholera epidemic, when Snow tracked the cases on this map. The cholera cases are highlighted in black. Using this map, Snow and other scientists were able to trace the cholera outbreak to a single infected water pump.

ILLUSTRATION BY JOHN SNOW, IMAGE COURTESY FINAVON, WIKIPEDIA

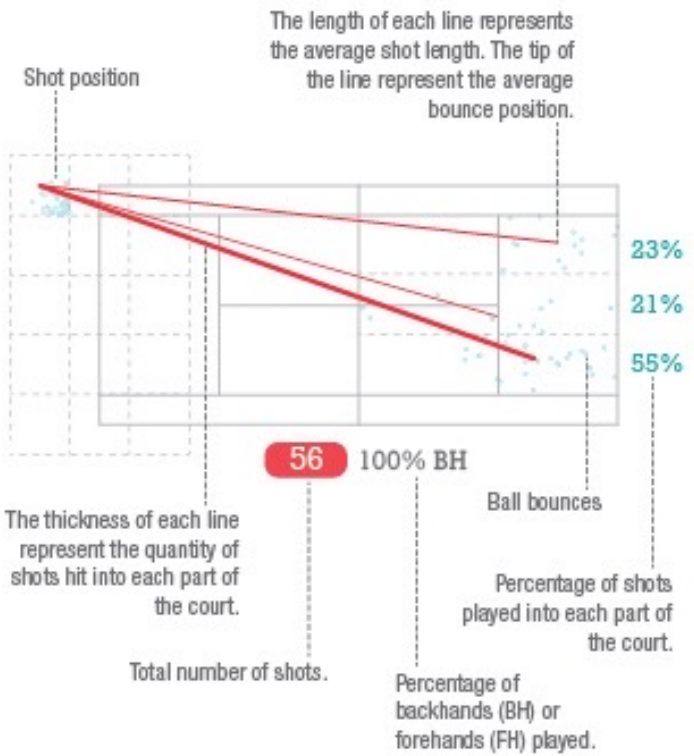


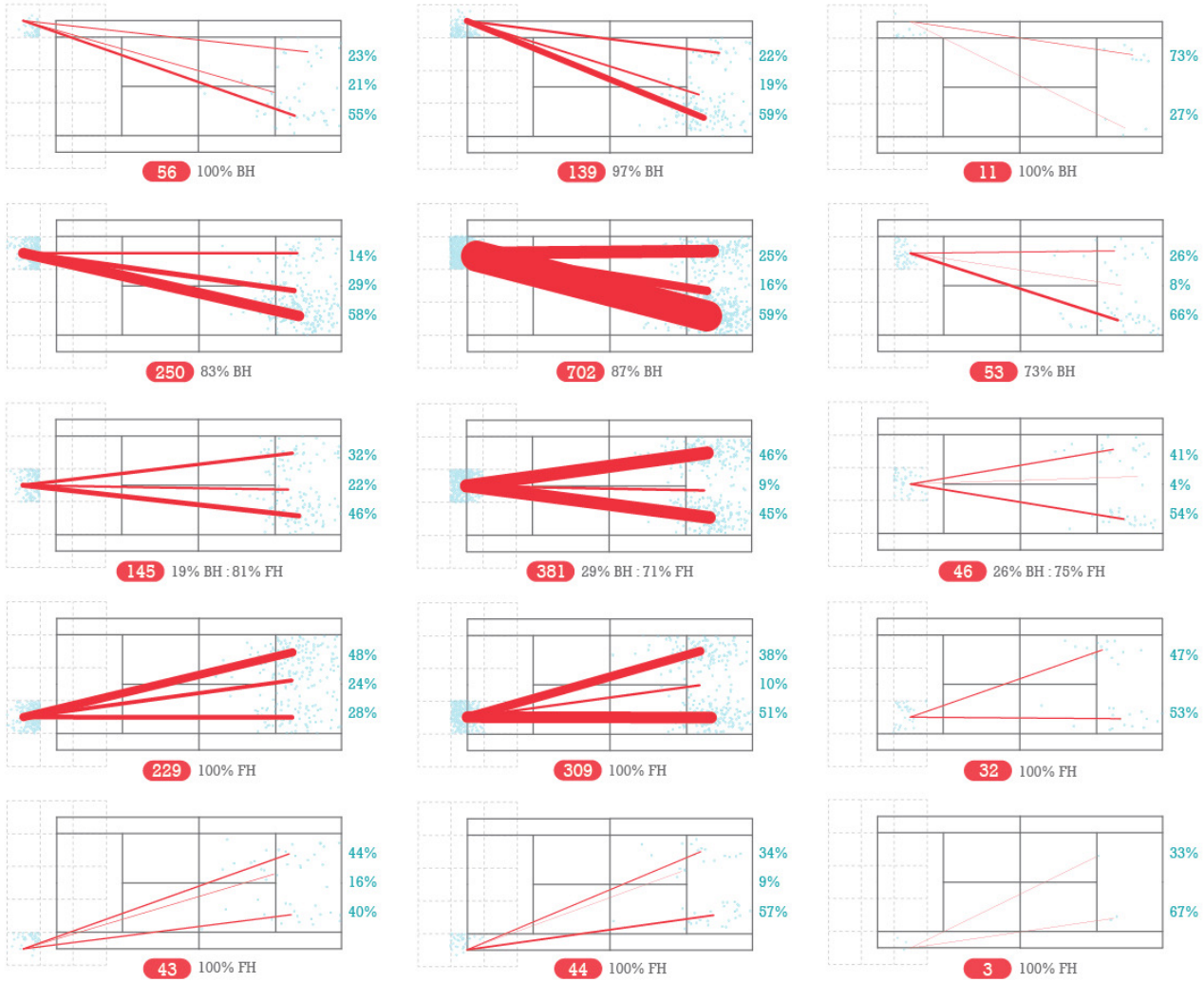
Example: investigating the source of epidemic -- https://www.nationalgeographic.org/activity/mapping-london-epidemic/

Kei Nishikori Shot Charts

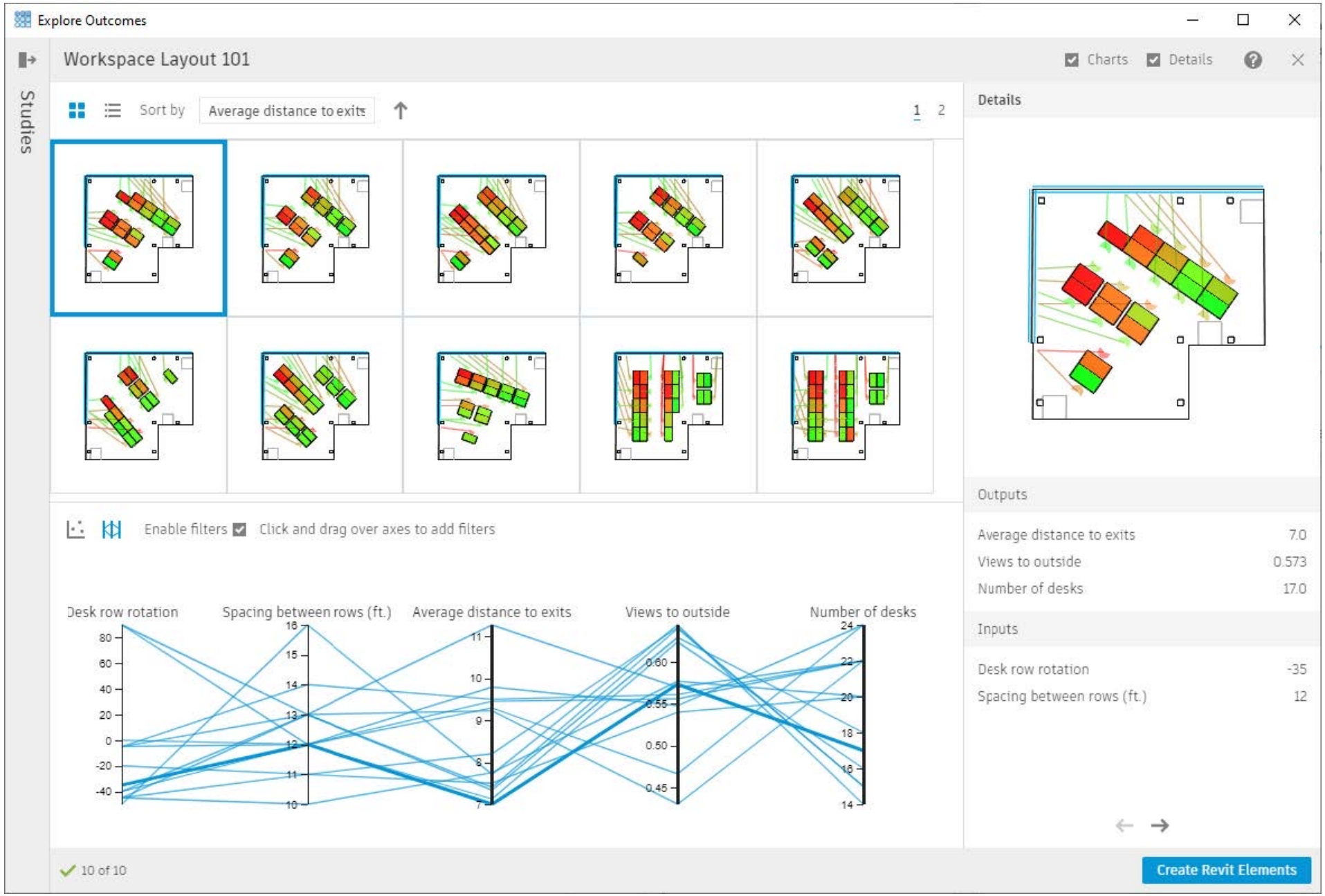
錦織圭のショットチャート

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/

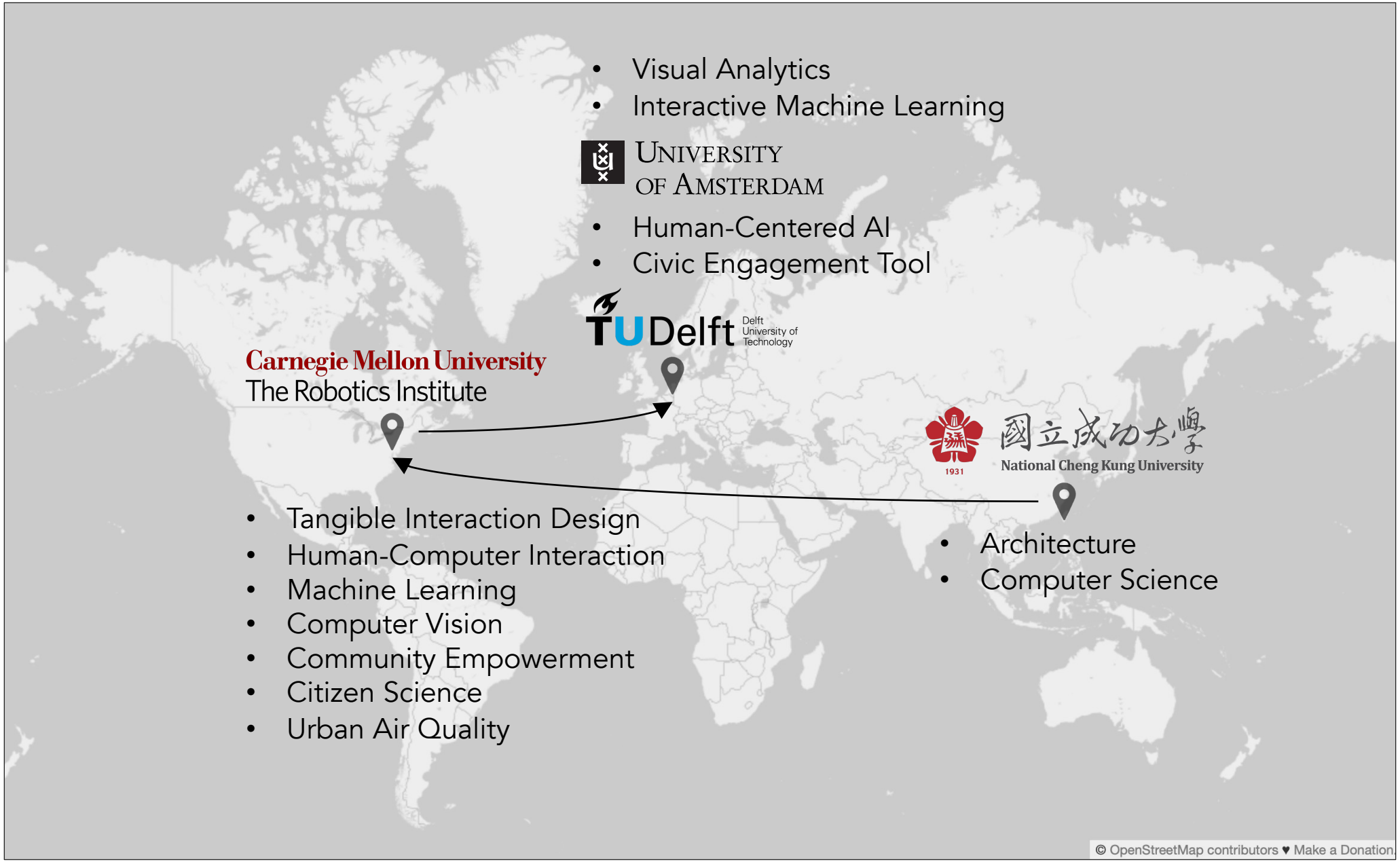


Example: exploring automatically generated design options -- https://www.autodesk.com/autodesk-university/de/node/132606

This course has four teaching team members.

Check Canvas for their contact information.

- Yen-Chia Hsu, course coordinator
- Yahia Dalbah, head of teaching assistants
- Tycho Stam, teaching assistant
- Alejandro Monroy, teaching assistant



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

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, only one lecture)	
5	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 and no lecture)	

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 will 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. But you still need to get at least 5.5 in the final score. There is a resit, which is 90% weight (will override your original weighted sum of exam scores).

Reflective Writing Submissions for Assignments

- We only grade your reflective writing submissions but not the assignments.
- Use the reflective writing template that we provide for submissions.
- We grade the reflective writing with pass/fail.
- 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.
- Assignment materials can appear in the mid-term exam, final exam, and resit.

Changes from Last Year

- Based on students' feedback, we improved the slides by adding clarification information (e.g., validation error, SVM, Logistic Regression, regularization, class activation mapping, conditional probability), materials (e.g., feature importance, PCA, weight initialization, positional encoding), and exercises (e.g., linear classifier, convolution operation).
- Tutorial notebooks are improved based on students' feedback (e.g., move some functions and the answers out of the main notebook, improve the assignment for PyTorch practice and structured data processing). We also added another tutorial for the image data processing module.
- TAs will give recitations during the seminars to repeat or clarify concepts that are covered in lectures.
- Late submission policy is different (automatic deduction of 10% max points of the assignment per day). We also added more mock exams for preparing mid-term and final exams.

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.

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?