

Welcome to Data Science

IFT6758

Fall 2019

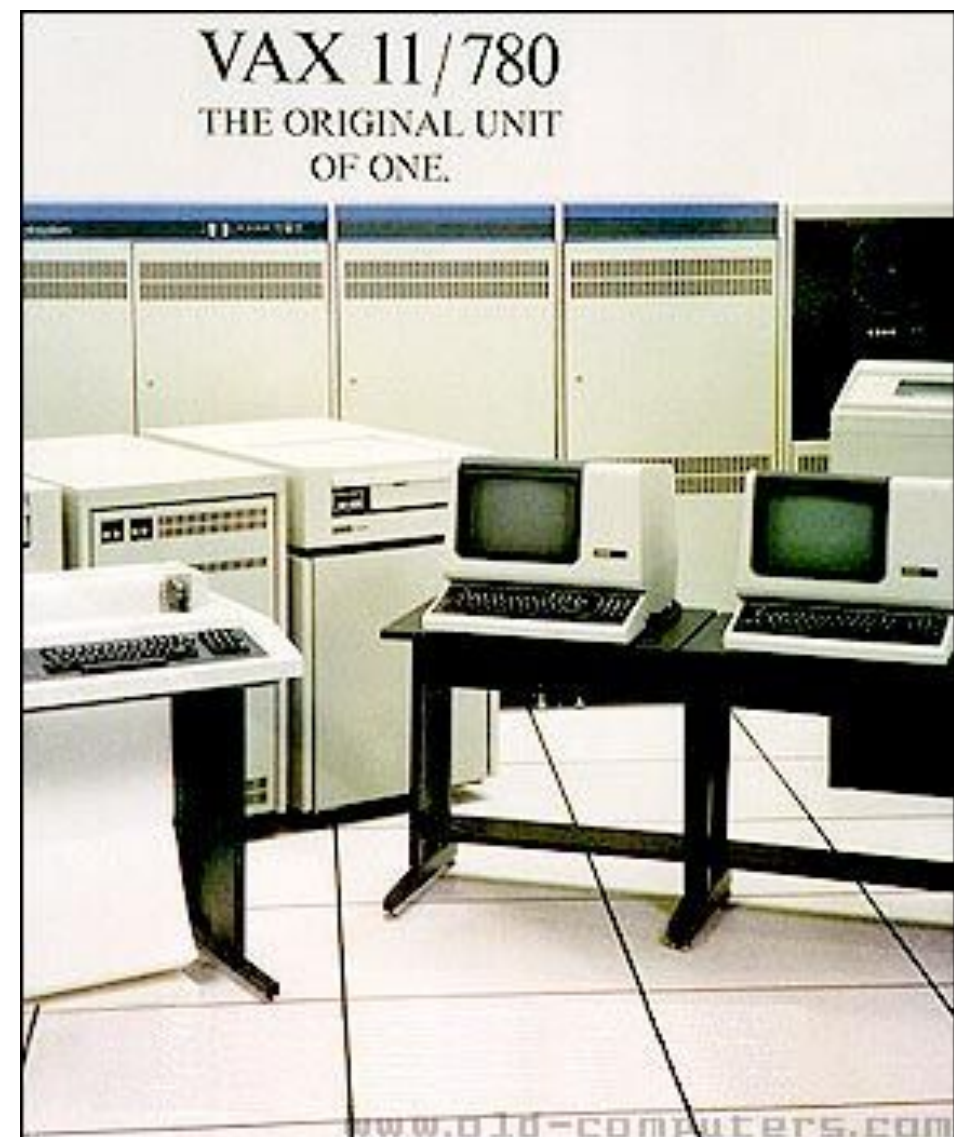
Survey: <https://forms.gle/bEjKNMXzuzBeBMxc9>

Teams: <https://forms.gle/793jBEcBh9U57Qp99>

Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Private organizations
 - Civil society

A lot can change in
30 years...



Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - **Scientific discovery**
 - Private organizations
 - Civil society



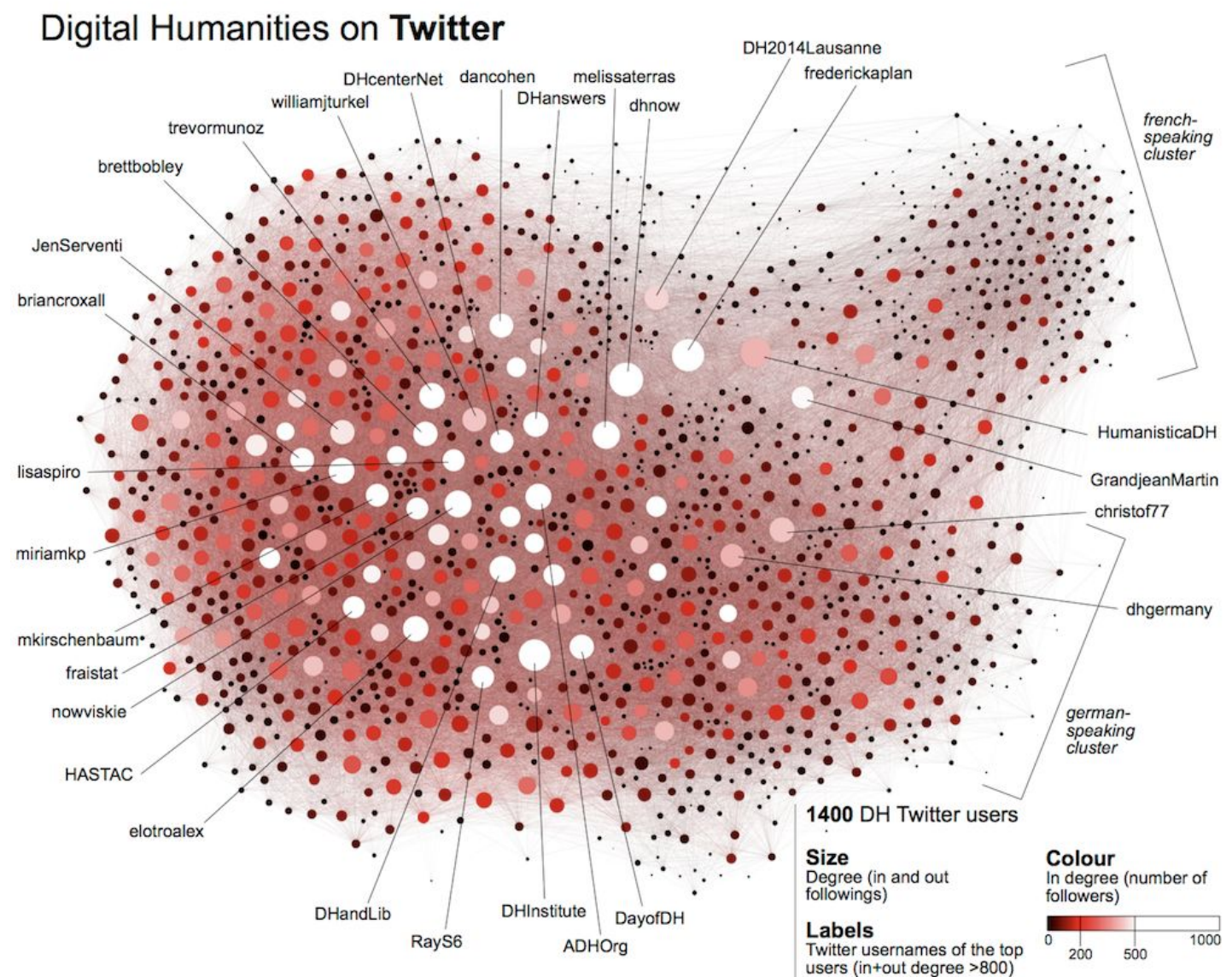
Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - **Scientific discovery**
 - Private organizations
 - Civil society



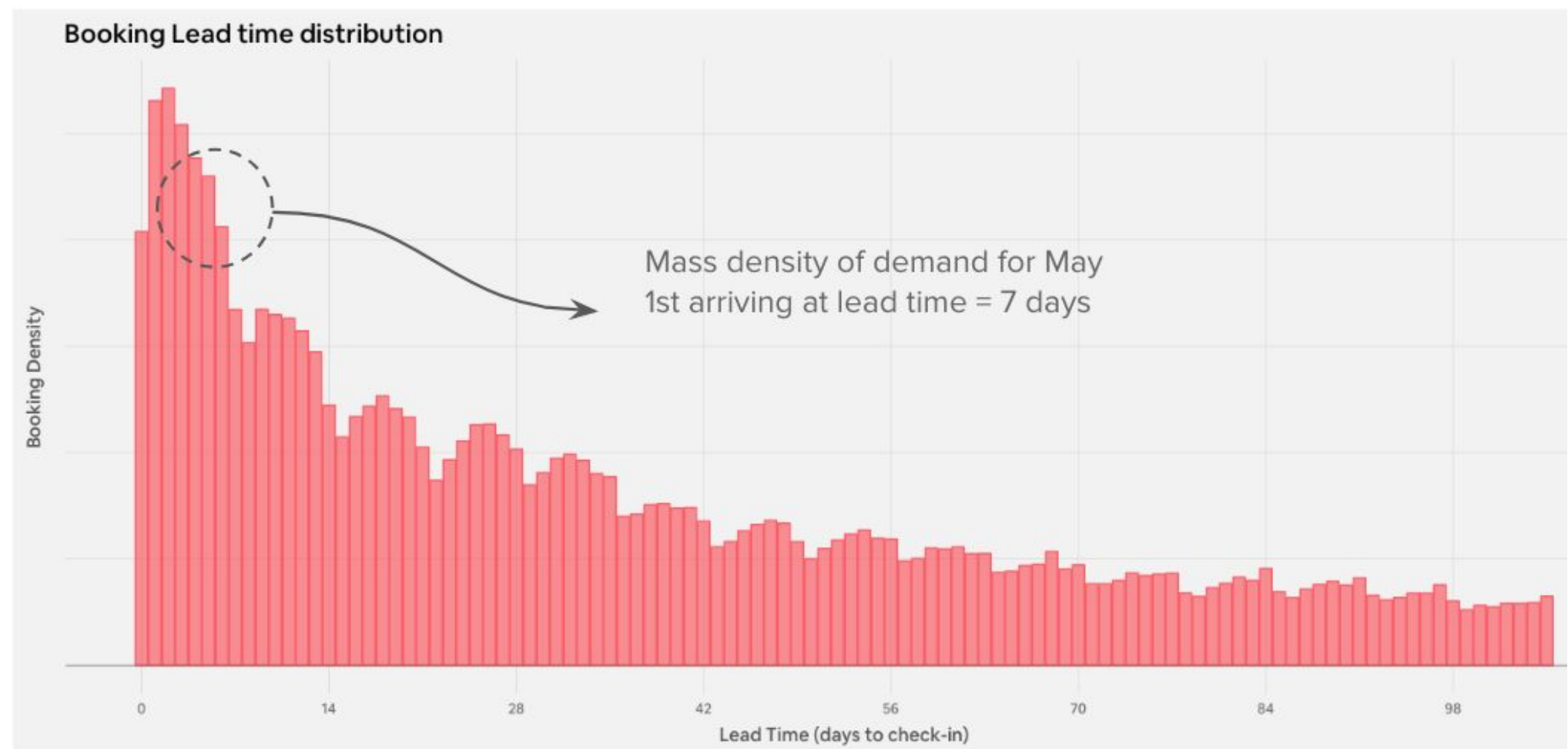
Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - **Scientific discovery**
 - Private organizations
 - Civil society



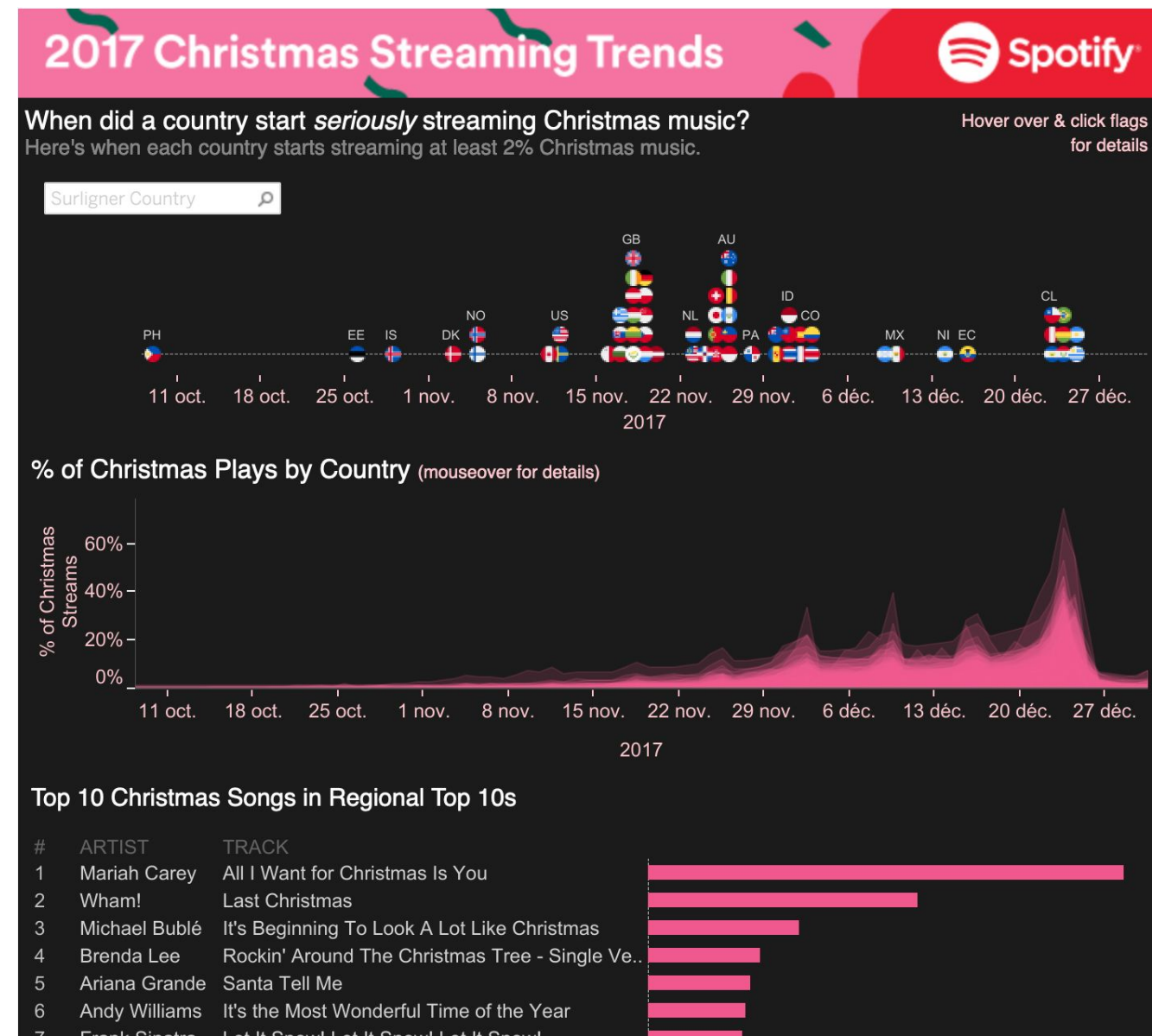
Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - **Private organizations**
 - Civil society



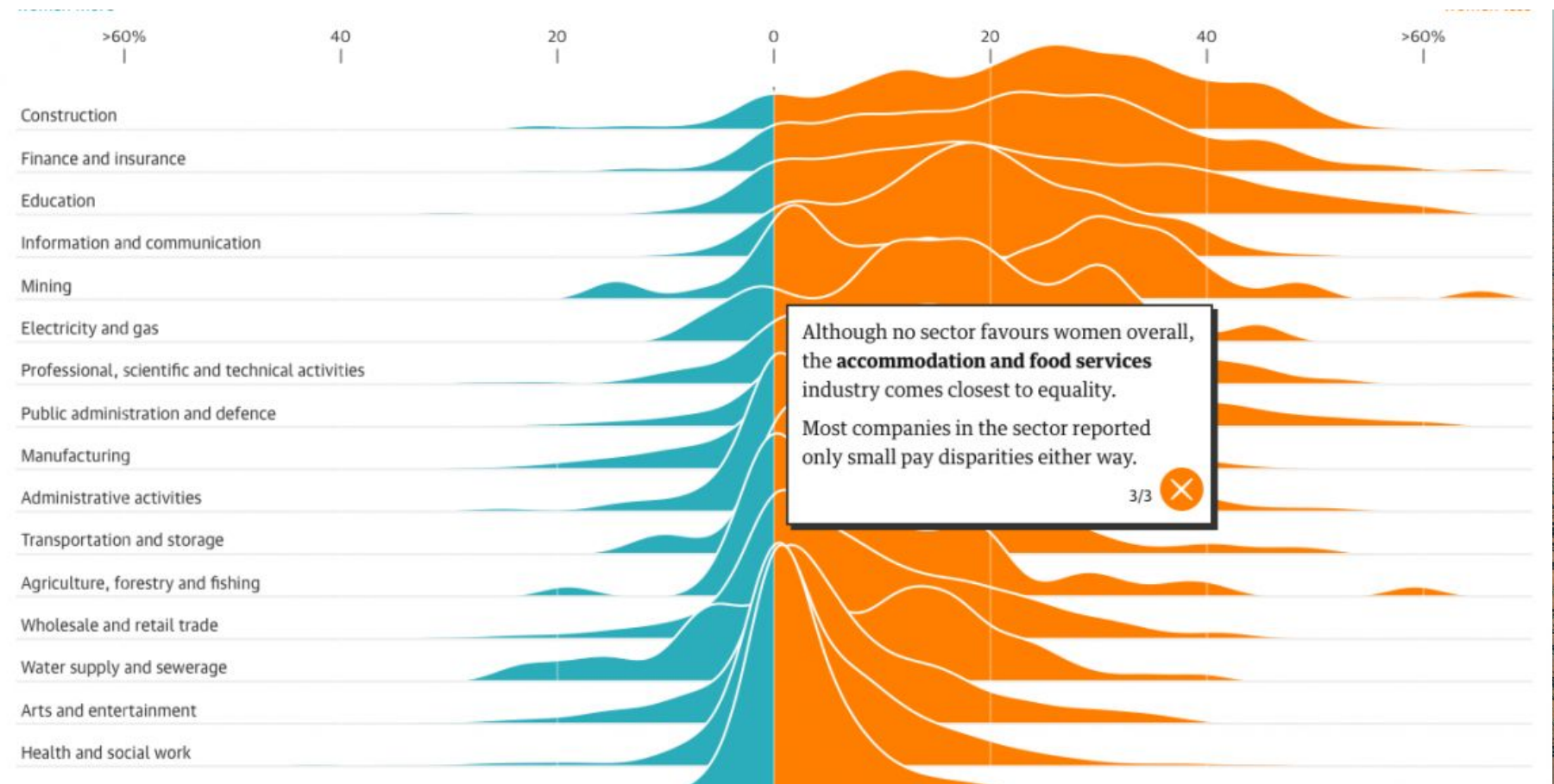
Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - **Private organizations**
 - Civil society



Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Private organizations
 - **Civil society**



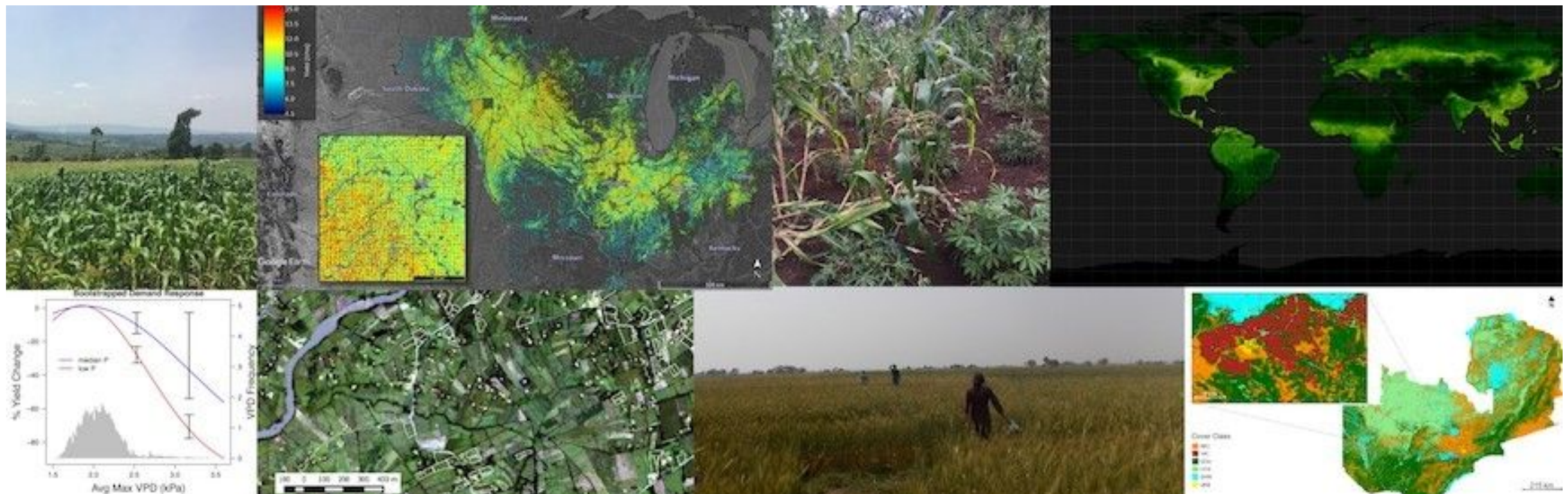
Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Private organizations
 - **Civil society**



Data Science: Year 2050

- Billions of sensors have been collecting data for decades
- Data is routinely collected and employed for
 - Scientific discovery
 - Private organizations
 - **Civil society**



Data Science: Year 2050

- I could go on and on...everything is already in motion
 - <https://github.com/awesomedata/awesome-public-datasets>
- This isn't even the end of your career!
- With so much driven by data, it's important that data scientists work **responsibly** and for the **greater good**

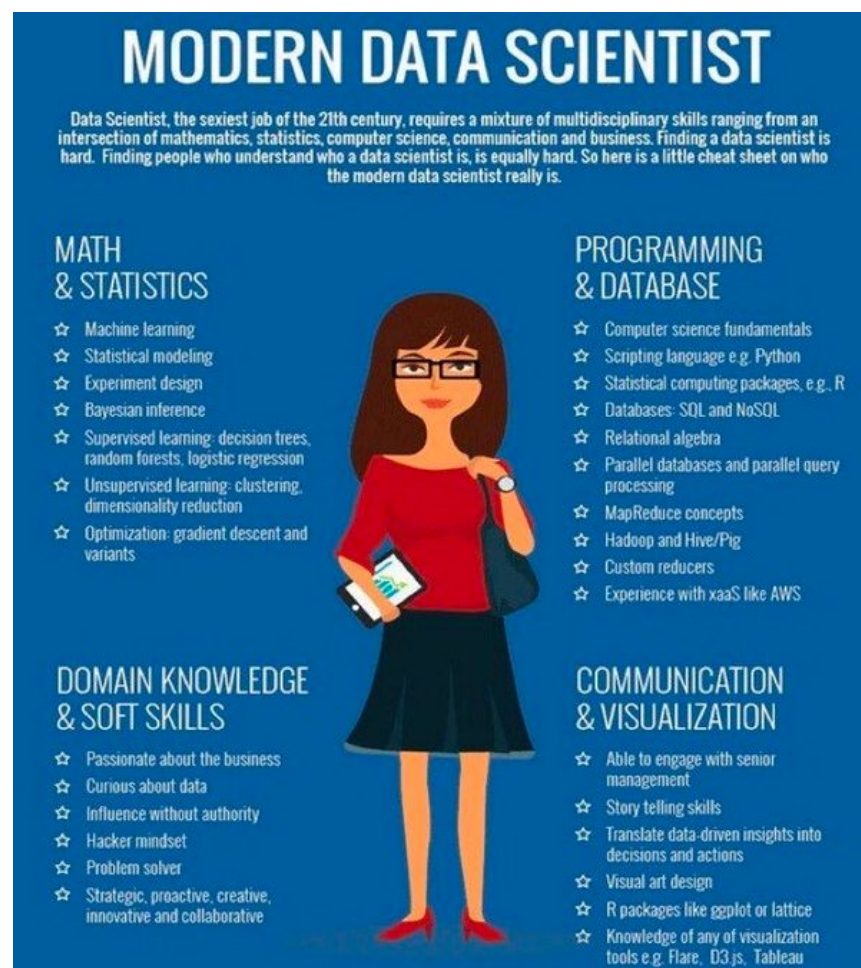
Learning Objectives

- **Data fluency**
 - Build competence working with multimodal data sets
 - Exposure to the full data science workflow
- **Become a data detective**
 - Learn to ask good questions
 - Reason about uncertainty, think critically
- **Learn responsible data science**
 - Understand risks at all stages of data science workflow



Data Scientists

- You want to...
 - Apply data science in your own field
 - Work in industry or research
 - Understand data's role in society



Course Outline

- **Part 1: Summaries and Inference**
 - Data transformations and visualization
 - Supervised and unsupervised summaries
 - Inference and model comparison
- **Part 2: Nontabular Data**
 - Text and image data
 - Graph Mining
- **Part 3: Frontiers**
 - Advanced Inference
 - Ensembling
 - Privacy and explainability

Logistics

- Website: <https://ift6758.github.io>
- Fill [survey](#) for access to discussion forum
- Grading: 35% project, 25% Final, 25% HW, 15% Midterm
- Professor & TA Office Hours TBA

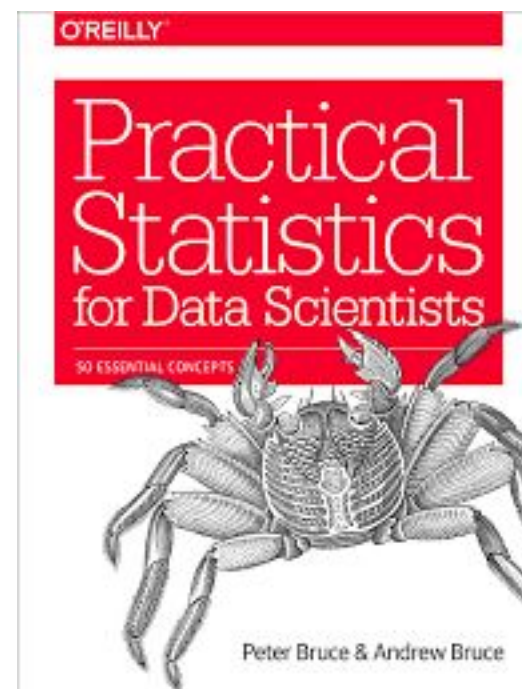
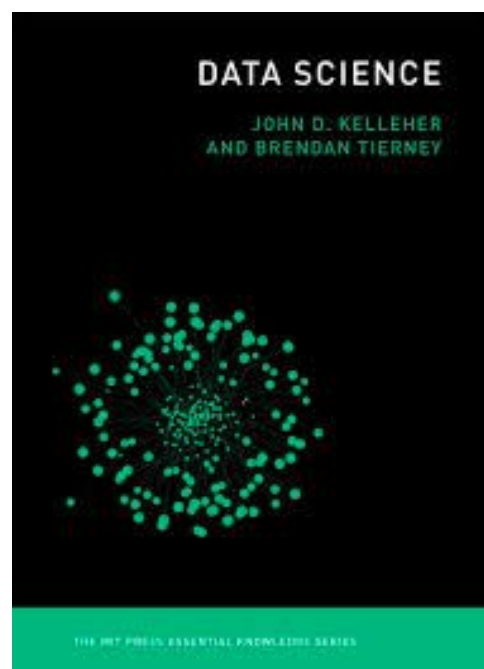
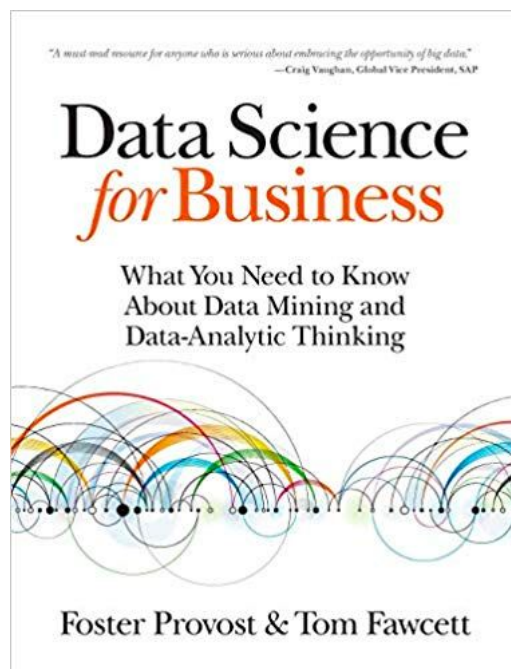
Contact (but use Forum when possible!)

Kris -- kris.sankaran@umontreal.ca

Golnoosh -- farnadig@mila.quebec

Resources

- [Online courses](#)
- Other sources of inspiration...
 - Leo Breiman's [commencement address](#)
 - Harvard [Data Science Review](#)
 - [Data is Plural](#)
 - [Data Humanism](#)
 - A new [Elements of Style](#)



Project

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.

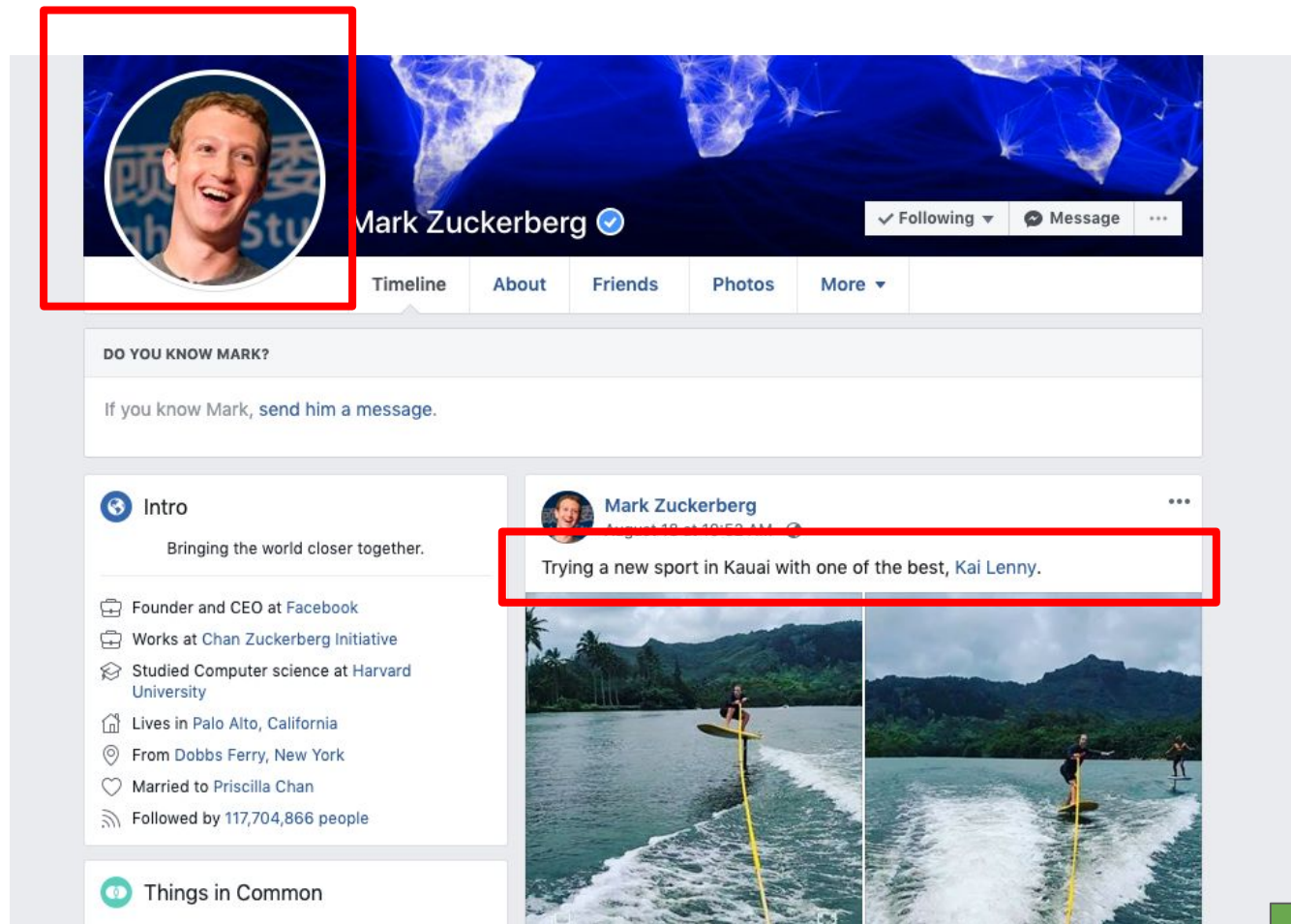
User Profiling in Social Media

Task: Infer users' gender, age, and personality traits

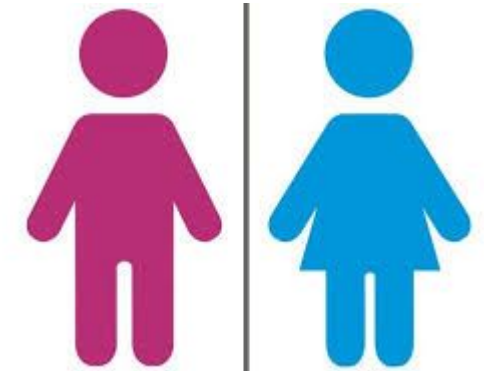
Data: Profile picture (image), status updates (text),
page likes (relation)

Grade: 35%

Deliverables: 2 presentations, 1 group report, 1
individual report, and couple of weekly evaluations on
the software performance



Gender:



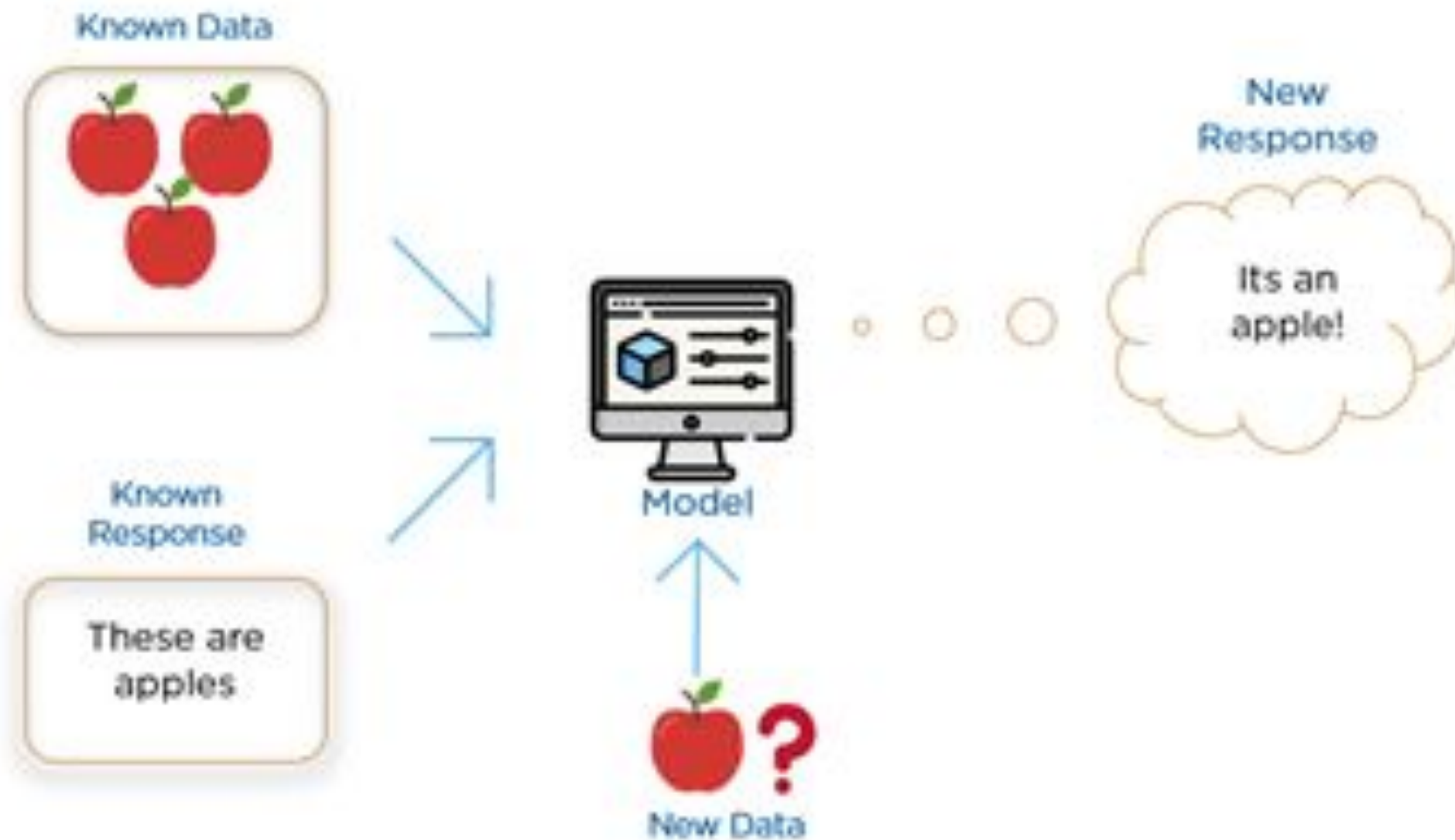
Age:



Personality traits:



Supervised learning

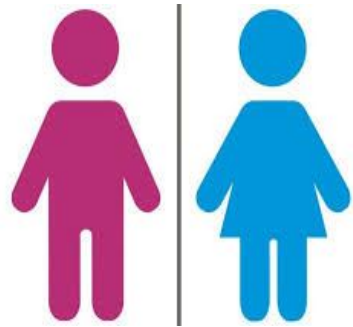


Training data: 9500 facebook users with labels

Public test dataset with data of 334 Facebook users (no labels!)

Hidden test data: 1334 users (1000 new users + 334 public test users)

Supervised learning Tasks



Binary Classification
Female vs. Male



Multi-class Classification
4 classes: “xx-24”, “25-34”,
“35-49”, or “50-xx”



Regression
Score between [1,5]

Evaluation

Starting date: TBA

Your software will be tested every week.

Your solutions for age and gender will be assessed based on **accuracy**.

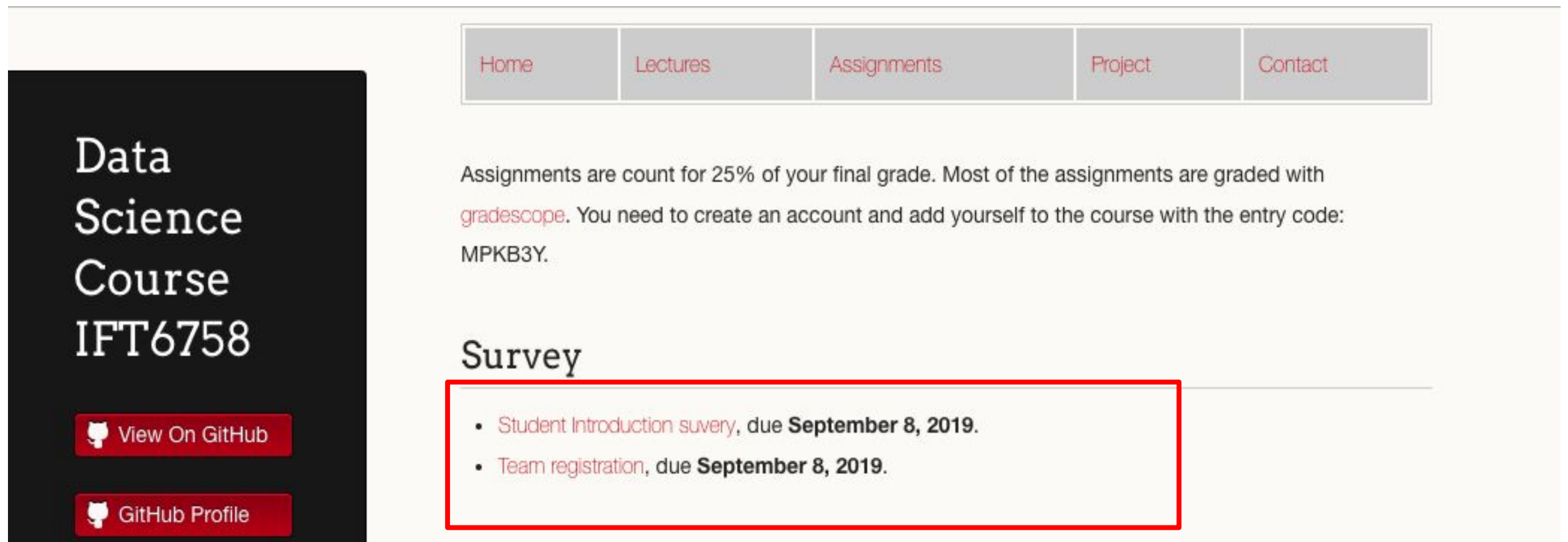
For personality identification, we will use the average Root Mean Squared Error (**RMSE**).

The score of all teams will be posted on the course website.

Team

Go to the course webpage and register your team (3-5 members):

<https://ift6758.github.io/project.html>



Data Science Course IFT6758

[View On GitHub](#)

[GitHub Profile](#)

Home Lectures Assignments Project Contact

Assignments are count for 25% of your final grade. Most of the assignments are graded with [gradescope](#). You need to create an account and add yourself to the course with the entry code: MPKB3Y.

Survey

- [Student Introduction suvery](#), due **September 8, 2019**.
- [Team registration](#), due **September 8, 2019**.



Data access

Step 1: Register your team



Step 2: Receive data agreement form and project description



Step 3: All team members need to sign the user agreement form to access the data



Step 4: Each team will receive an account to access the server

Prizes



You have a chance to win a prize! We will give prizes to the teams:

- With the best score on the last evaluation of the course
- With the most innovative approach

Lab

Lab Assignments: * optional, will not be graded

Practical Labs: Intro to Python, and different packages

Where/When: Tuesdays 12:30PM-14:30PM at B3250

Pavillon 3200 J-Brilliant, Université de Montréal (subject to change)