Data Analytics

Unit 7 - Storytelling with Data, Web Scraping, APIs, AB Testing

NOV - DEC 2020 | BERLIN

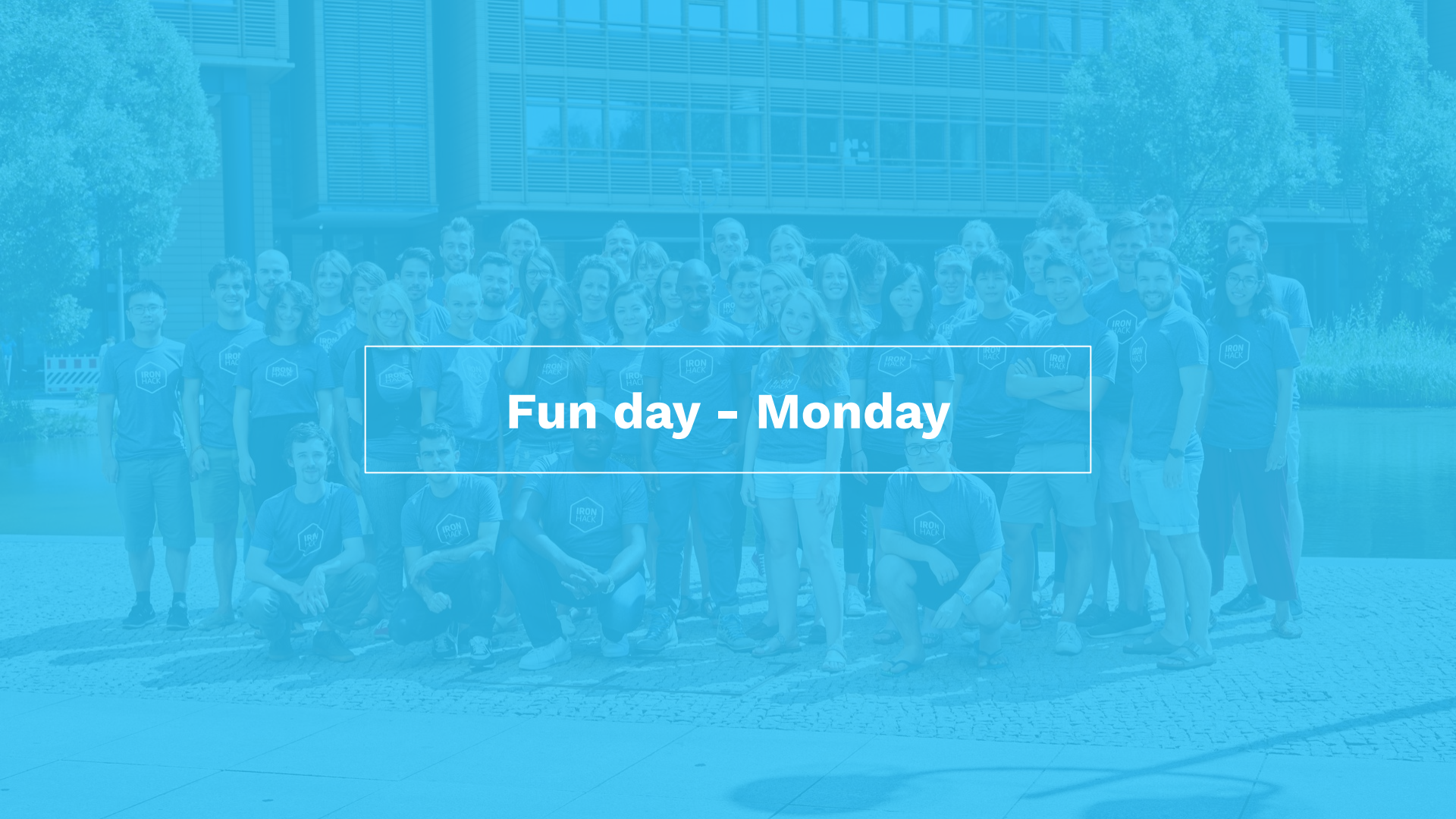
What will I learn in this unit?

Unit #8

Python HTML Tableau Presenting

The aim of this unit is to polish their data analytics and engineering skills by performing an end-to-end data product: we will create a program that takes an input from the user and automatically collects data from the internet through web scraping and APIs; then it goes through a clustering model and finally returns an output back to the user.

They will implement agile methodologies to develop the product and finally they will "sell it" with an engaging presentation



Fun day - Monday

Morning lecture

LFB Best of class dashboards

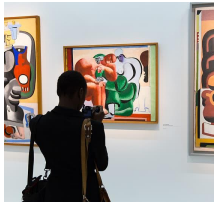
Why do we tell stories?
Zoom in Zoom out
Data storytelling
Narrative Arc
Tips on Tableau Story setup

--Project intro--
(split into groups)

Afternoon Session

Group Project

Covid-19 and Human Movement



Gallery

4.30-4.45 Break

4.45 Gallery of Data Stories

Why do we tell stories?

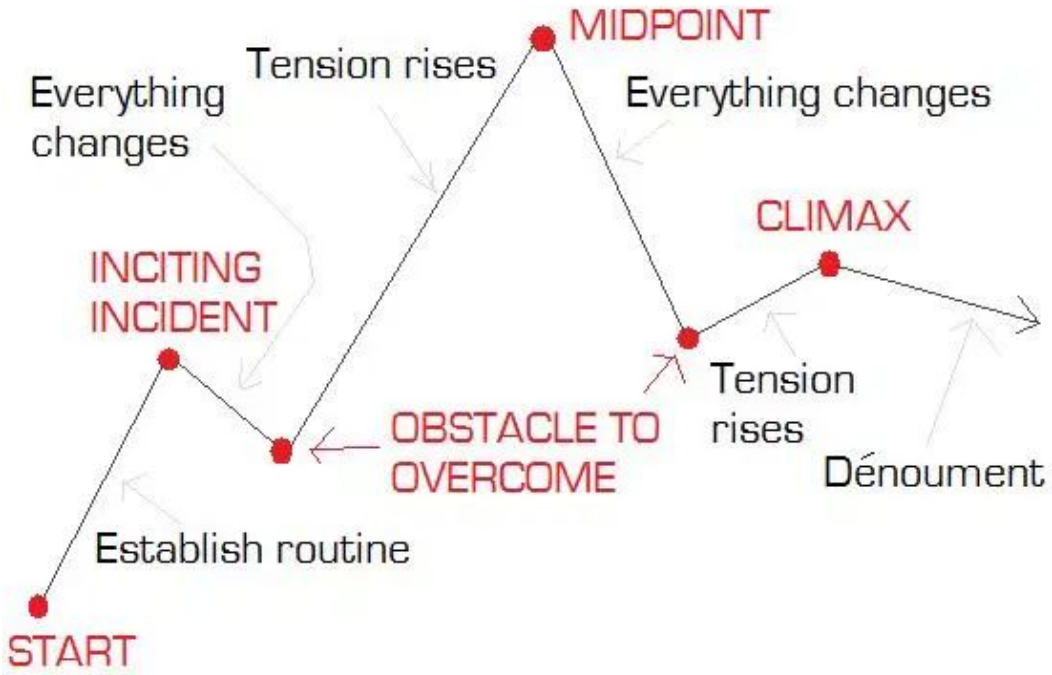
Value and meaning
Oldest tradition
Learn its important in childhood
To be remembered - emotion causes memory
Makes us human - relate the story of everything - how we relate to things- impact!
Connection - is a story - make something relevant - you feel involved
Communicate ideas - shared reality/ history
Tells you who you are
Explain our world
Distinctly human trait - religion, nationhood
Identifies us and other
Information - warnings - morality

What makes a good story?

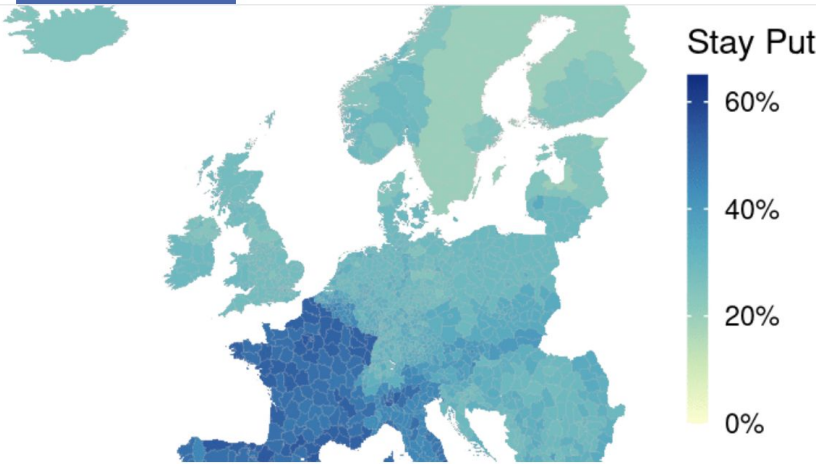
Clarity
Visuals are vivid
Tone
Humour
Strength in storyline - peak
Structure
Common thread
Relatability
Shock , Surprise, unexpected
Elicit an emotional reactions

Zoom in

Zoom out

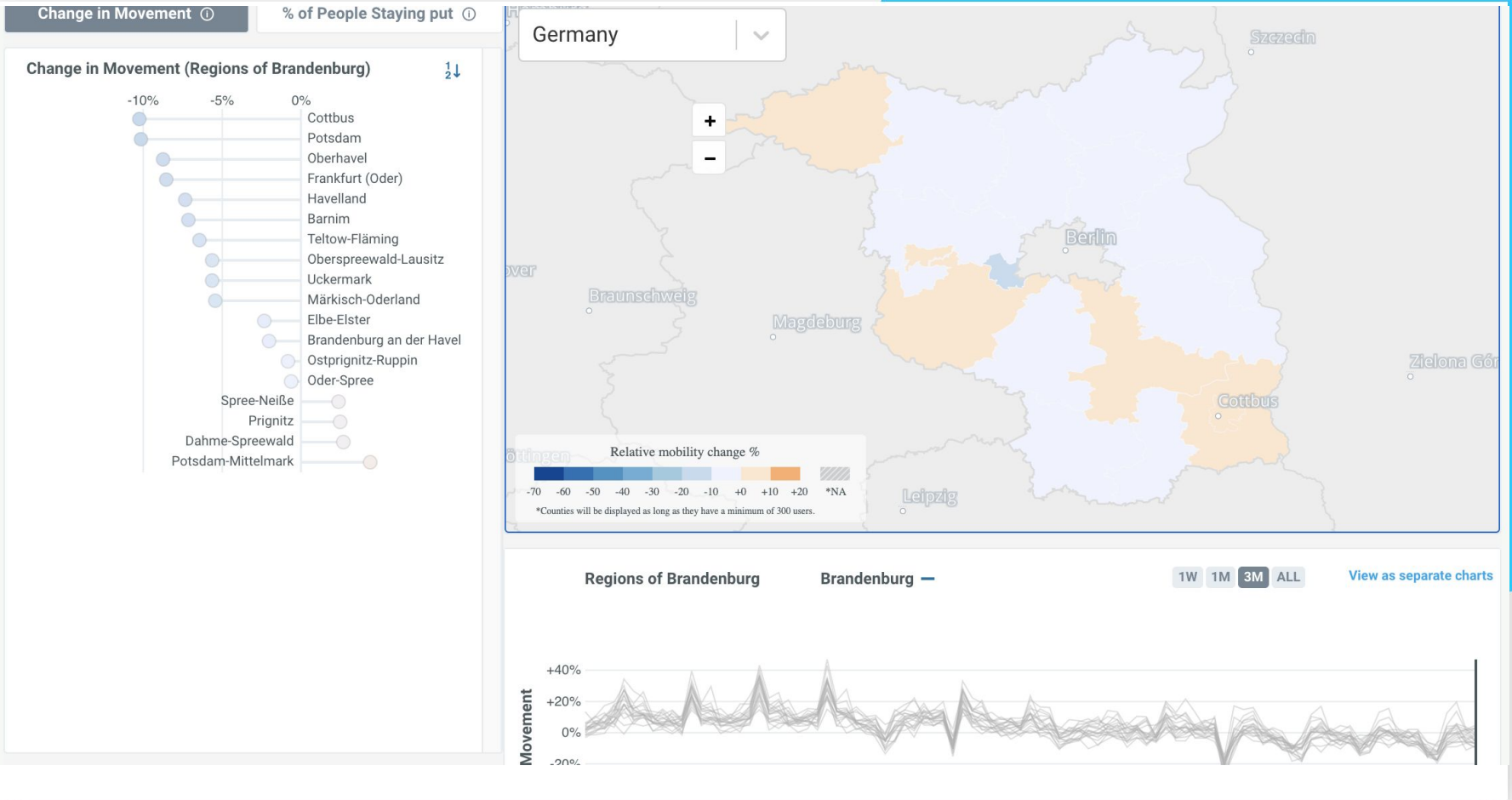


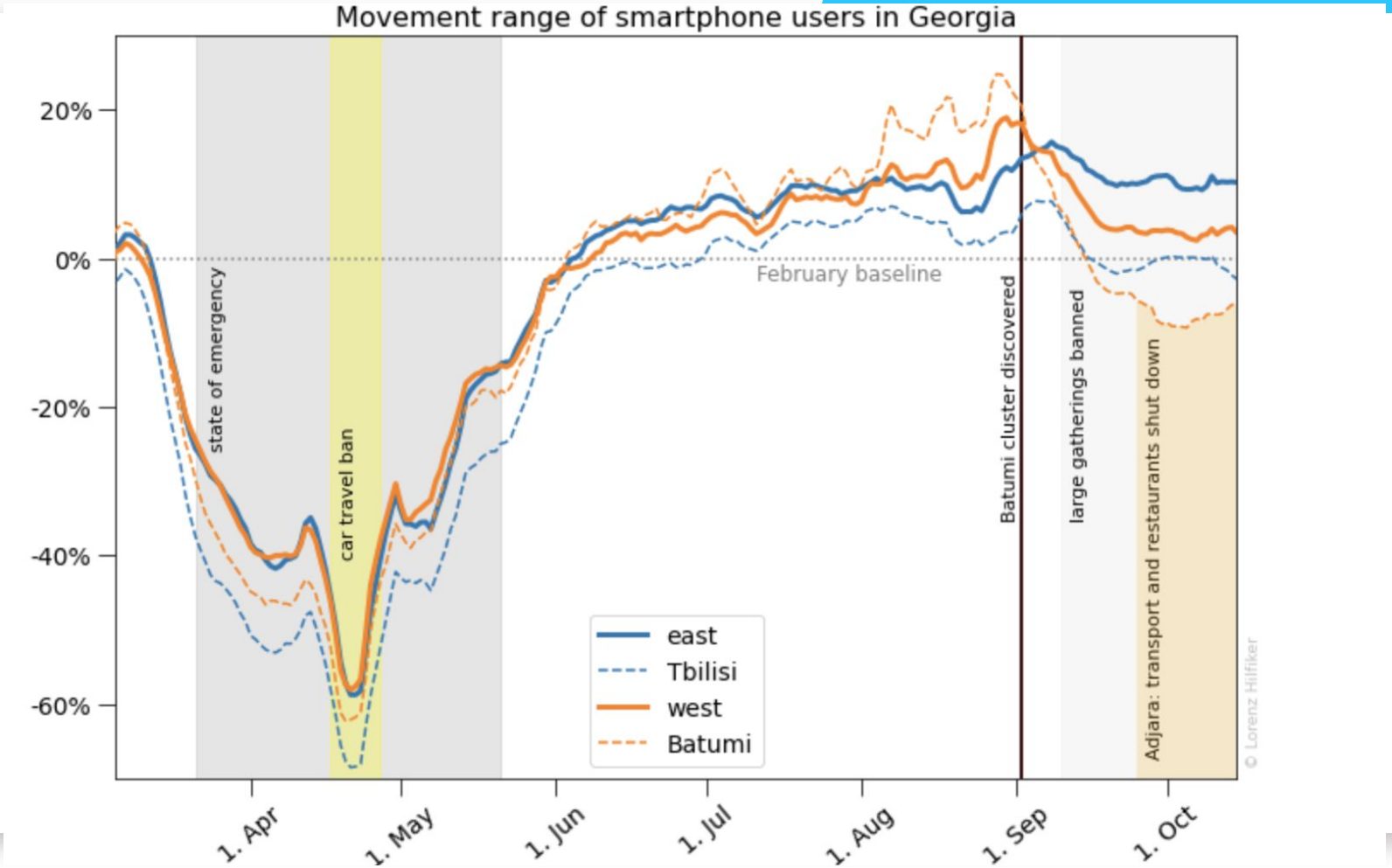
THE STORY ARC



Movement Range Maps

Movement Range Maps inform researchers and public health experts about how populations are responding to physical


Change in Movement
% of People Staying put
Change in Movement (Regions of Brandenburg)
-10%
-5%
0%
Cottbus
Potsdam
Oberhavel
Frankfurt (Oder)
Havelland
Barnim
Teltow-Fläming
Oberspreewald-Lausitz
Uckermark
Märkisch-Oderland
Elbe-Elster
Brandenburg an der Havel
Ostprignitz-Ruppin
Oder-Spree
Spree-Neiße
Prignitz
Dahme-Spreewald
Potsdam-Mittelmark
Germany
Szczecin
Berlin
Braunschweig
Magdeburg
Cottbus
Leipzig
Relative mobility change %
-70 -60 -50 -40 -30 -20 -10 +0 +10 +20 *NA
*Counties will be displayed as long as they have a minimum of 300 users.
Regions of Brandenburg
Brandenburg
1W 1M 3M ALL
View as separate charts
+40%
+20%
0%
Movement


Movement range of smartphone users in Georgia
20%
0%
-20%
-40%
-60%
state of emergency
car travel ban
February baseline
Batumi cluster discovered
large gatherings banned
Adjara: transport and restaurants shut down
east
Tbilisi
west
Batumi
1. Apr
1. May
1. Jun
1. Jul
1. Aug
1. Sep
1. Oct
© Lorenz Hilfiker

Two Day - Tuesday

Morning session

Introduction to WebScraping

Case Study explained

When do we need it? 8.01.1

Html basics

- **Tags, structure, inspect**
- **Next steps for newbies**

Beautiful Soup & Parsing 8.01.2

Scraped data & pandas

Andres presentation- final project - and deep learning

Afternoon Session

Lunch 12:30 - 1:30

Review of Pandas and Getting started with Web Scraping with Flo

Lab Session

->TA assisted Labs from 15:00 -

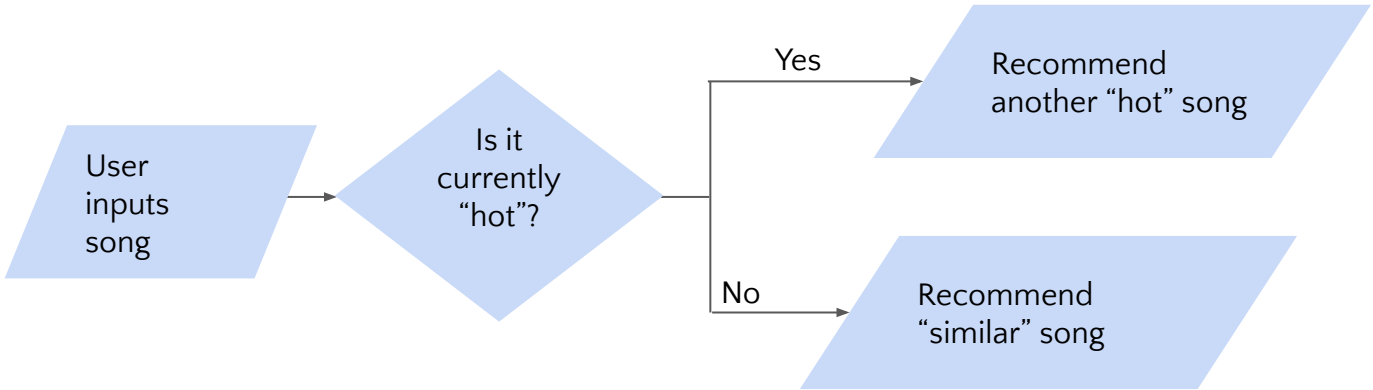
8.01 (inside lesson unit in Day 2 of student portal) **HTML Web Scraping**

Web Scraping -optional

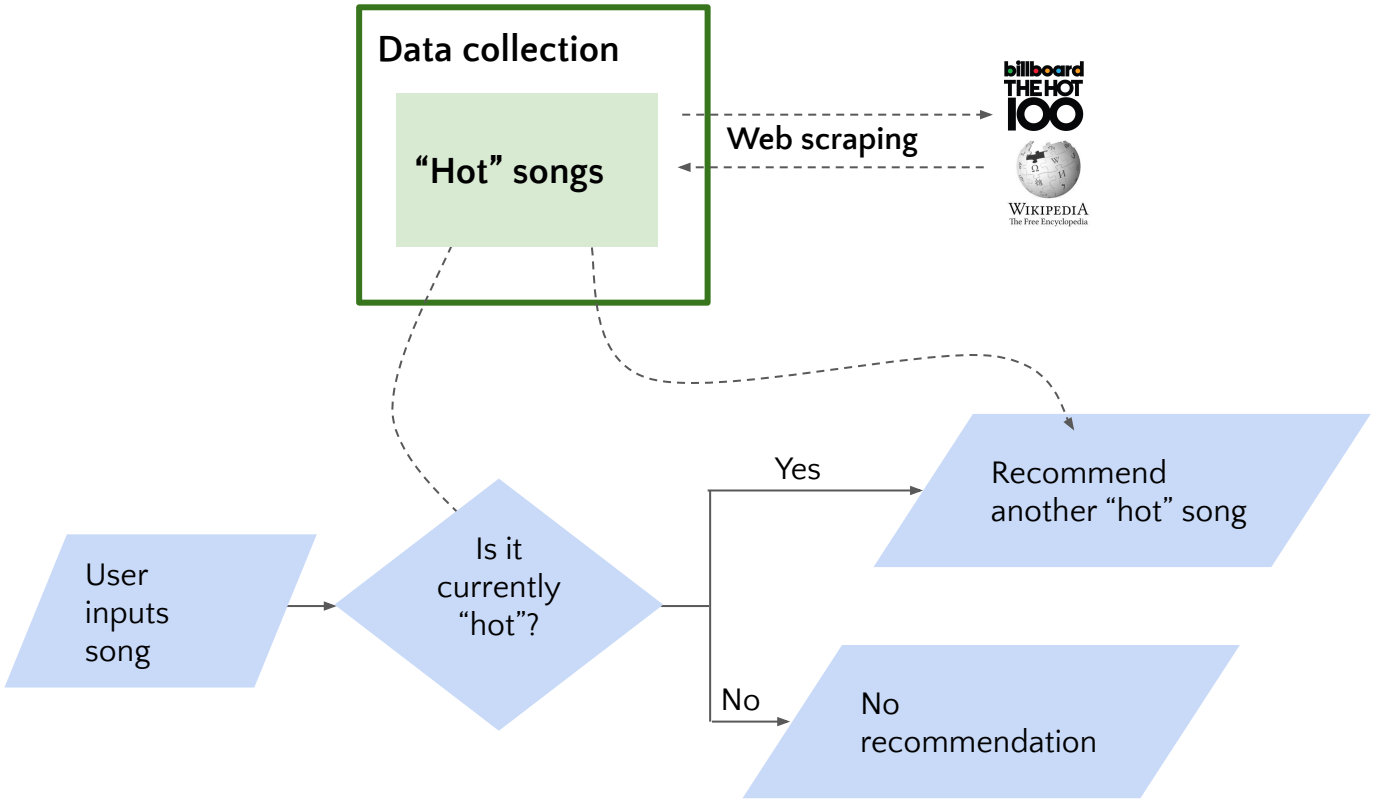
HTML Tutorial - optional

CSS Level 32 - optional

Project flowchart - GNOD Case Study



1st prototype



When to use web scraping

No API - if API is available, normally better to use it

Automation Needed - of course we can copy + paste but ... ugh

Less Restricted - eg no API account required, less rules to follow (eg limit on # requests)

ISSUES

- You depend on the structure of the site being scraped
  - Can be messy
  - Can change overnight
  - Website protections

When to use web scraping

Ideas for sites and use cases

Yellow pages - addresses of companies in a city
Reddit
Asos - images of menswear
Social networks
Amazon - prices of products
Bbc news - see how countries are described
Airbnb - apartments and room prices / sizes / locations - impact
Twitter - Bit coin all time high - look for acronyms
Skyscanner - demand forecasting - prices - best times to book
Linkedin - for filtering jobs

Web scraping slides

Basic html (tree) structure

```
<!DOCTYPE html>
<html>
    <head>
        <title>Page Title</title>
    </head>
    <body>
        <h1>My First Heading</h1>
        <p>My first paragraph.</p>
        <p>My second paragraph has a <b>bold<b> word!</p>
    </body>
</html>
```

- An **HTML element** is defined by a start tag, some content and an end tag. When web scraping we will mostly be interested in the content, but the tag will be crucial in locating the content.
- **Tags** are just keywords that encapsulate some content. They tell the web browser how to display the content. Some examples of common tags are:
- `<!DOCTYPE>` and <html> define the document type
- <head>, <title> and <body> define the main parts of the document
- <h1> to <h6> define headings
- <p> defines a paragraph
- <b> will make its contents bold

Tags , attributes and value pairs

```
<a href="https://www.ironhack.com/">a data bootcamp</a>
```

Attributes you need to know

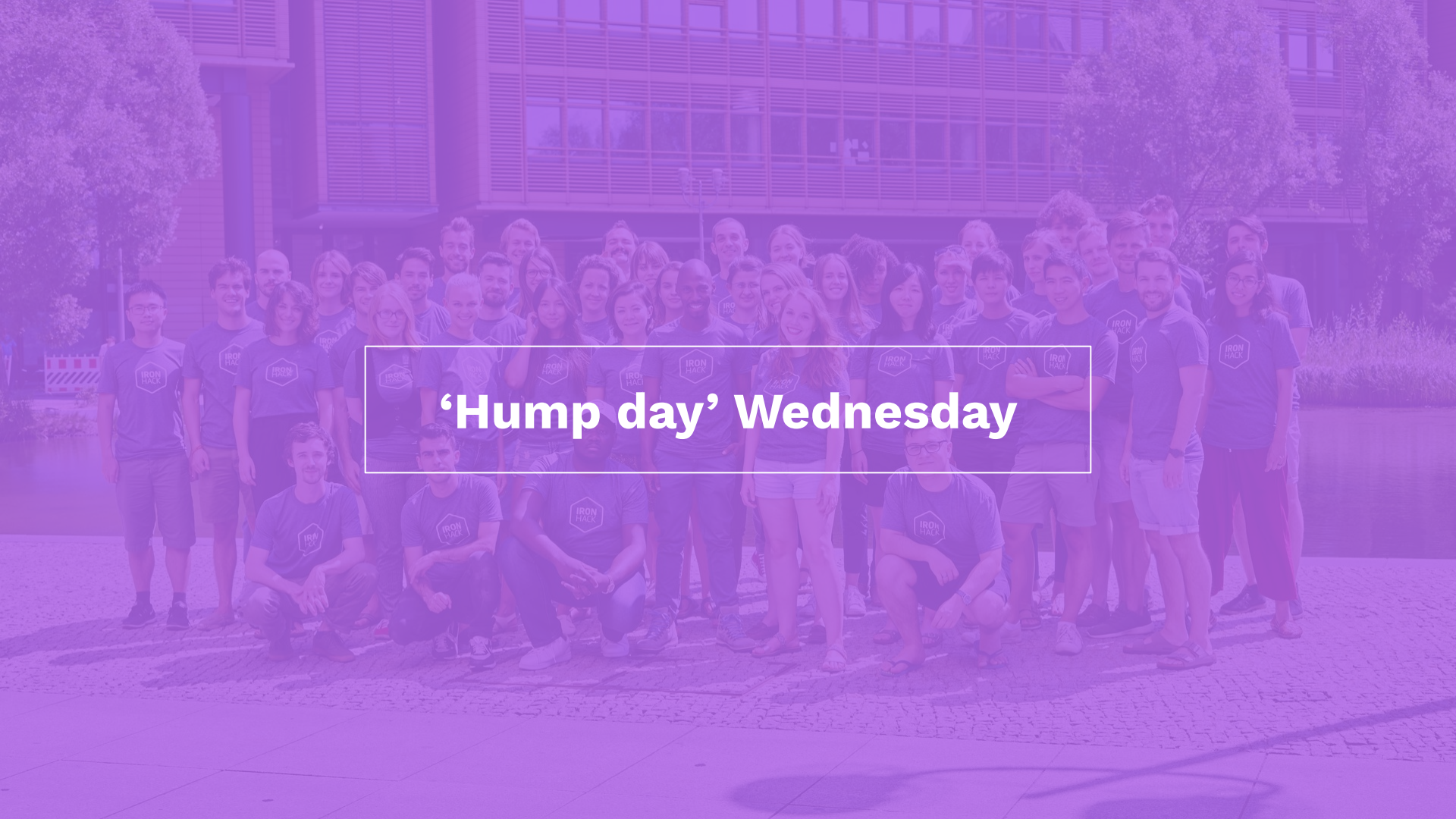
- The **id** attribute: unless the creator of the site has broken basic conventions, id's are unique. That makes them the best attributes for locating data in a site. If you discover that the piece of information you're trying to collect is an element that has an id, your job will be SO EASY. Bad news though: that doesn't happen often.

- The **class** attribute: it's often used to give style to multiple elements. For example, go to https://xkcd.com/. Notice how there are elements like "boxes" or "buttons" that are styled similarly in a site. Instead of defining the style for each one of these elements, the style for all the "boxes" might be defined in a different script (a CSS document), and it just points to all elements with `class = "box"`. This is often a useful way to locate content inside of an HTML script.


Found It!!!
Congratulations, it only took you 65298 seconds
www.jolyon.co.uk

The dormouse's story

Wikipedia - languages



‘Hump day’ Wednesday

Morning session

Lunch 13:00 -

Afternoon Session

1

Lab Session

->TA assisted Labs from 15:00-


Nearly there Thursday

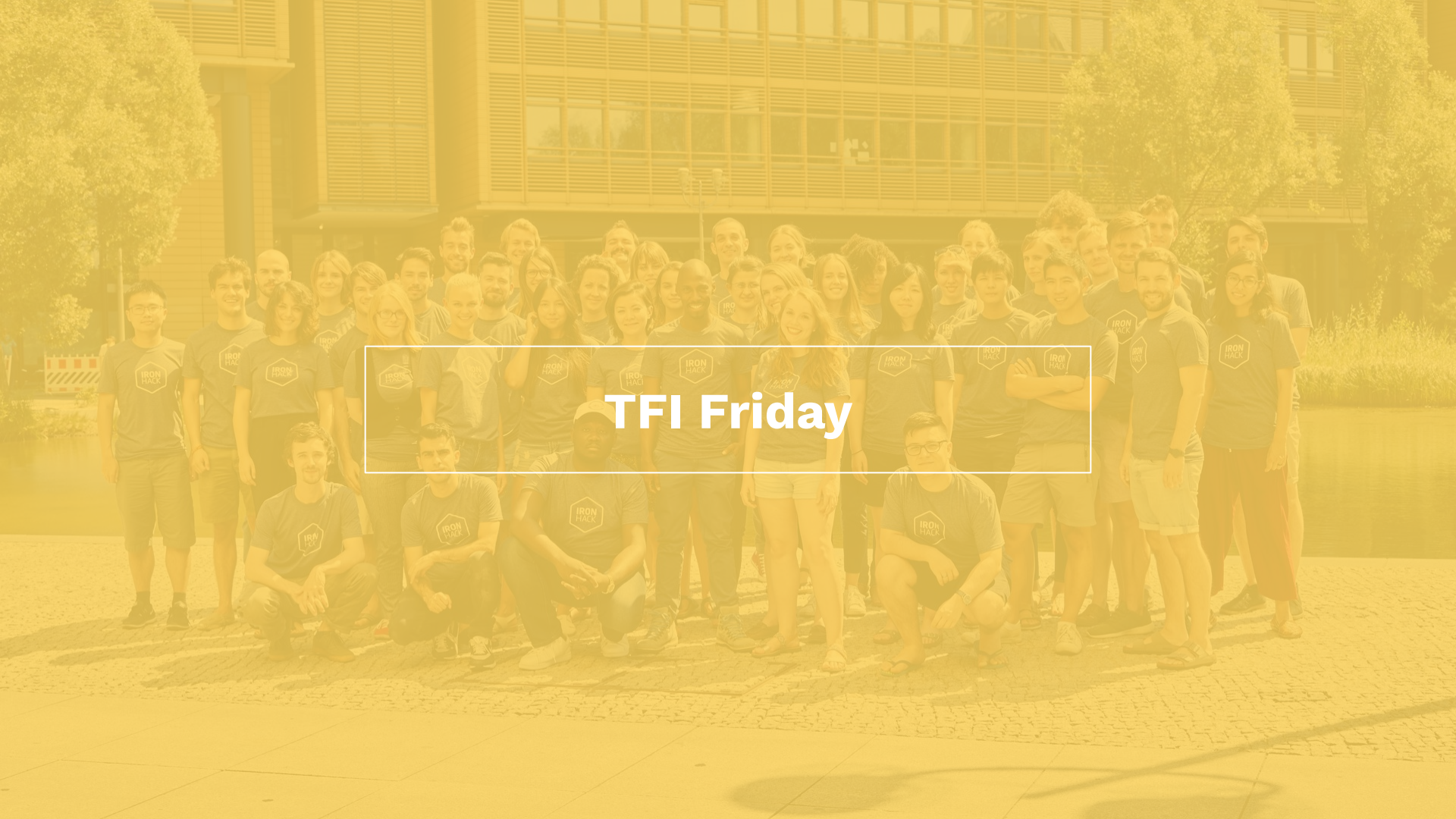
Morning session

Lunch 12:30 -

Afternoon Session

Lab Session

->TA assisted Labs from 16:00


TFI Friday

Morning session

Lunch 12:40 - 1:45

Afternoon Session