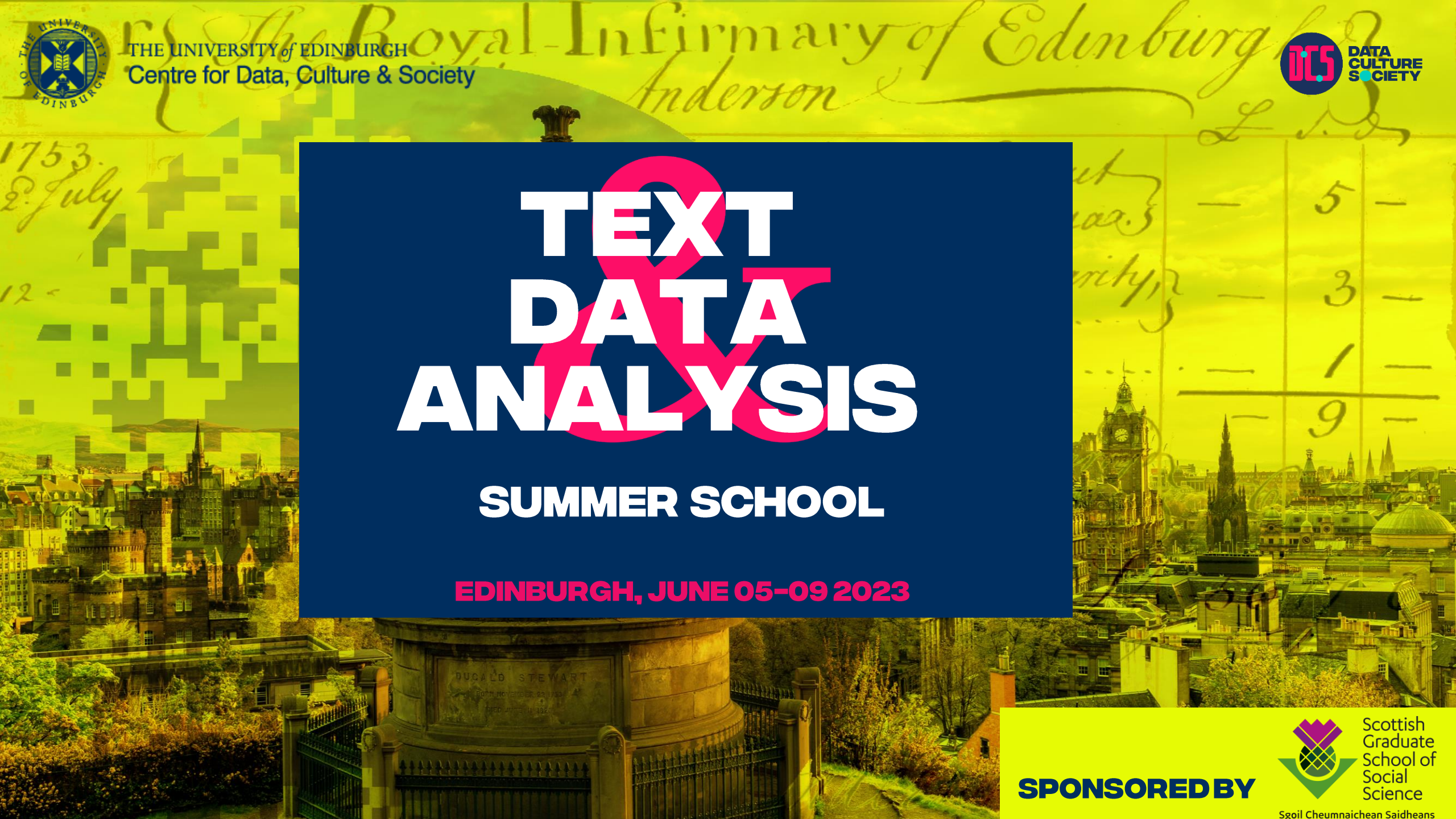
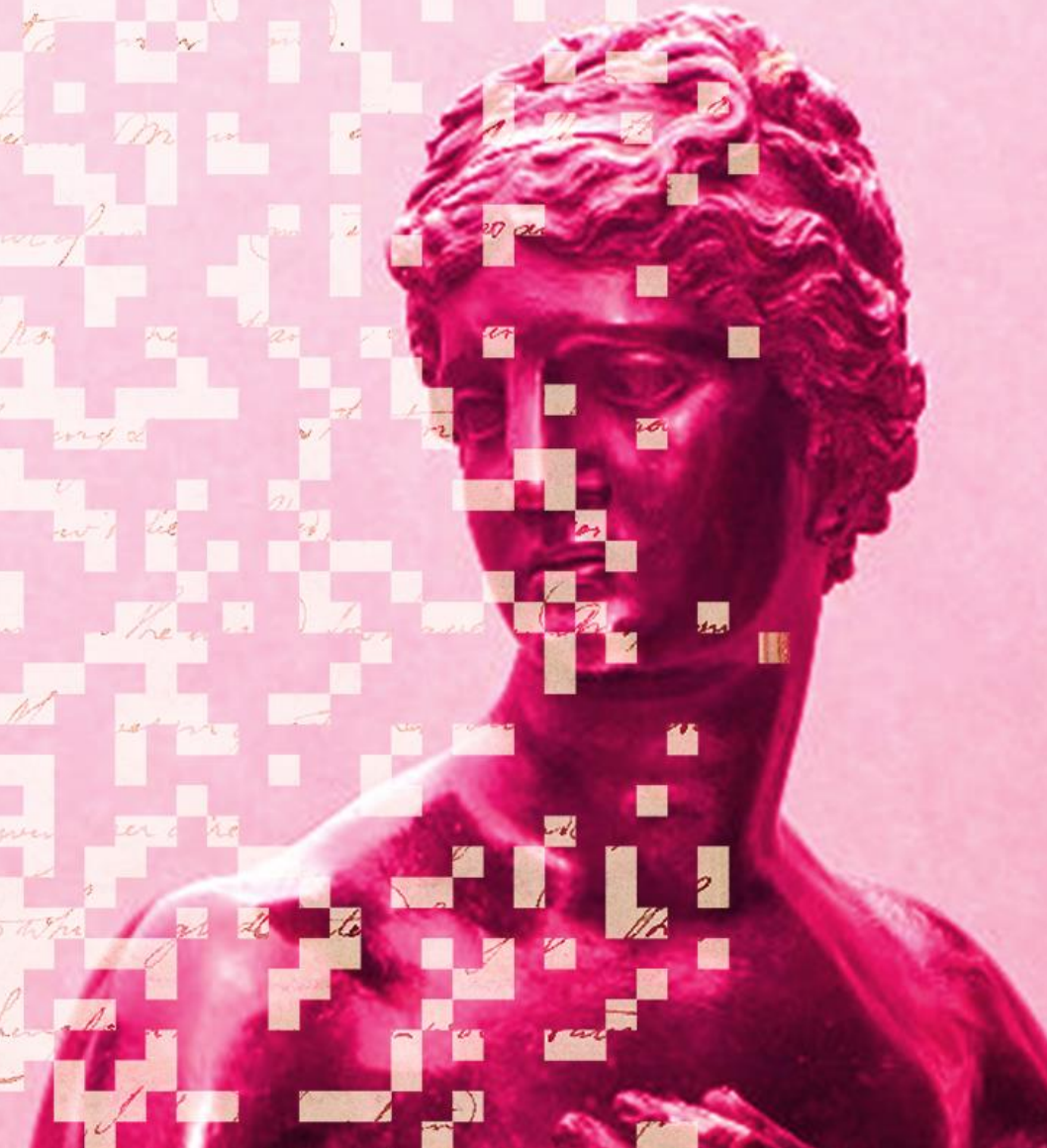

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TEXT & DATA ANALYSIS
SUMMER SCHOOL
EDINBURGH, JUNE 05-09 2023
SPONSORED BY
Scottish Graduate School of Social Science
Sgoil Cheumnaichean Saidheans

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

DATA CULTURE SOCIETY



@EDCDCS

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

HOUSE KEEPING

DATA CULTURE SOCIETY

- Toilets
- Food Consumption
- Water Fountains
- Fire Alarm
- Code of Conduct

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

TIMETABLE

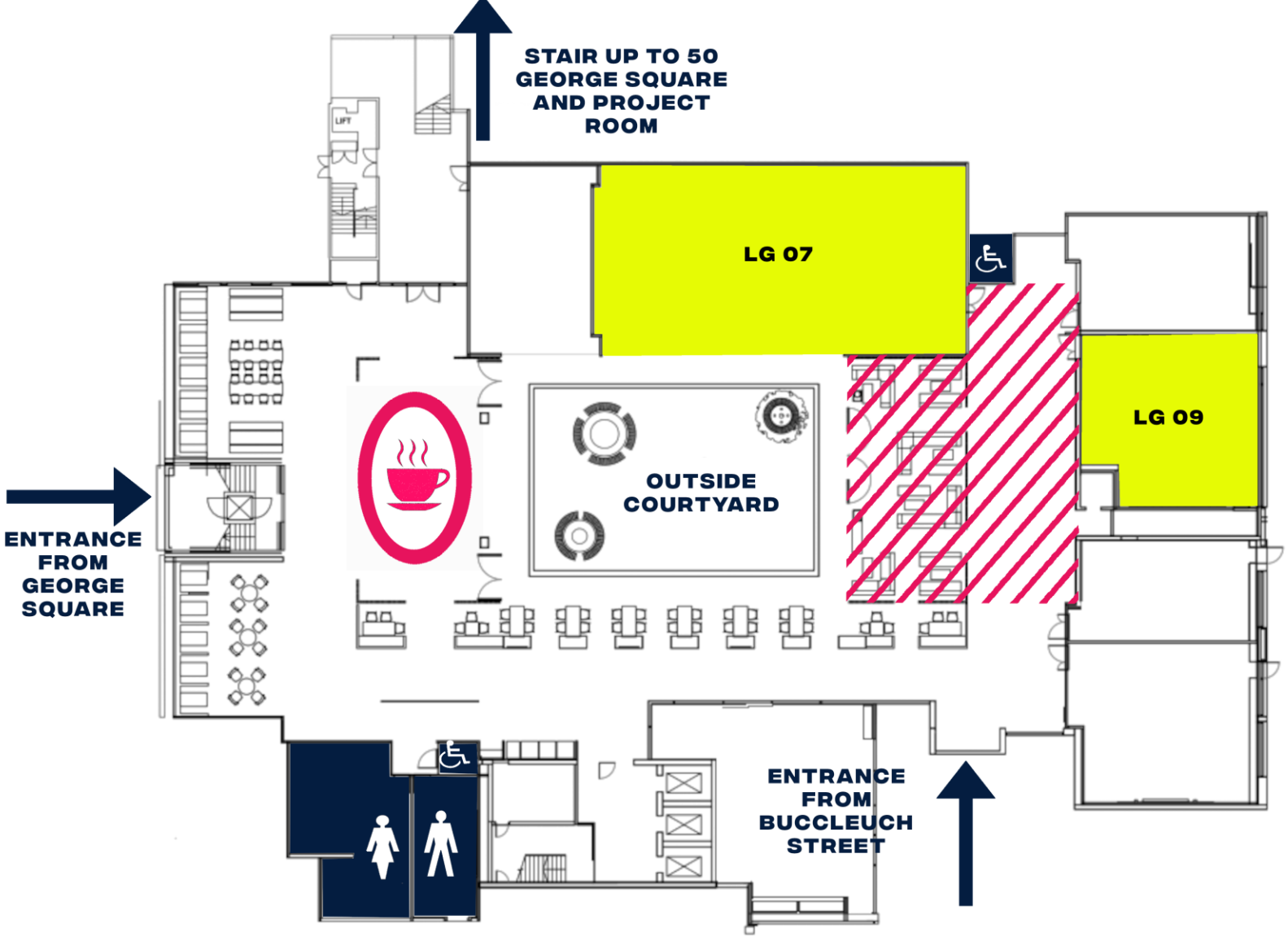
DATA CULTURE SOCIETY

	Monday	Tuesday	Wednesday	Thursday	Friday
09:00-09:30	Registration				
09:30-09:40	Welcome	Setting Up	Setting Up	Setting Up	Setting Up
09:40-10:40	Seminar	Seminar	Seminar	Seminar	Seminar
10:40-11:00	Coffee	Coffee	Coffee	Coffee	Coffee
11:00-12:30	Teaching Block	Teaching Block	Teaching Block	Teaching Block	Teaching Block
12:30-13:30	Lunch	Lunch	Lunch	Lunch	Lunch
13:30-15:00	Teaching Block	Teaching Block	Teaching Block	Teaching Block	Teaching Block
15:00-15:30	Coffee	Coffee	Coffee	Coffee	Coffee
15:30-17:00	Teaching Block	Teaching Block	Teaching Block	Keynote	Teaching Block
Evening	Pub Quiz	Pub Crawl	Ceilidh	Drinks Reception	Dinner

Grey: Events taking place in the **Teaching Rooms** (**LG07** and **LG09** Room, 40 George Square)
Yellow: Events taking place in the **Project Room, 50 George Square**
Pink: Refreshment breaks that will take place in the lounge area outside the teaching rooms
Teal: Events in the social programme of the summer school

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

BUILDING PLAN



UNDERGROUND LEVEL OF 40 GEORGE SQUARE

TEXT & DATA ANALYSIS SUMMER SCHOOL

Stream 1:
A Gentle Introduction to Coding for Data Analysis (LG09)

Stream 2:
Text & Data Analysis in the Wild (LG07)

WELCOME!

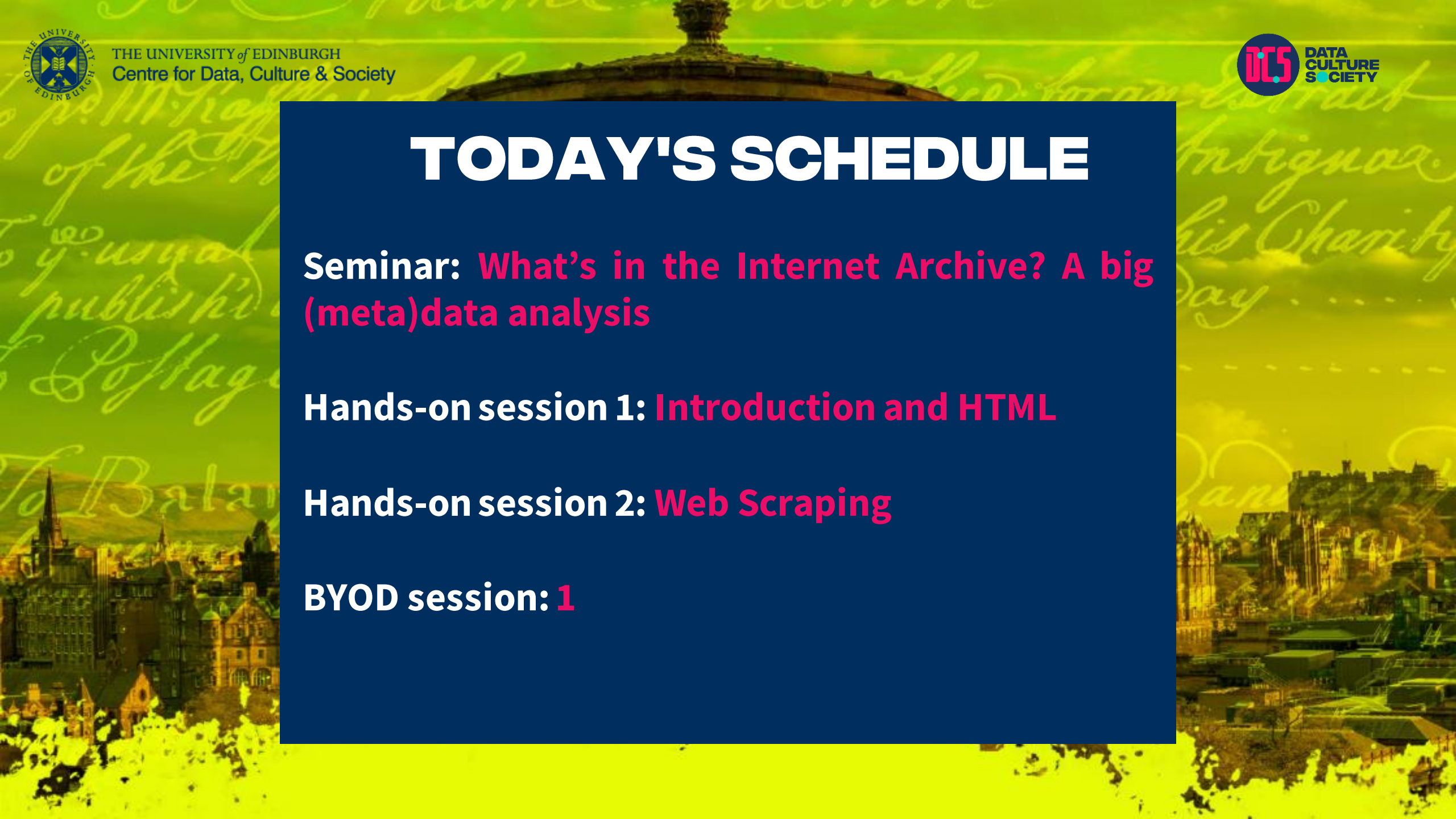
www.cdcs.ed.ac.uk

TEXT & DATA ANALYSIS IN THE WILD

The cost of Living in Scotland and the UK

WELCOME!

www.cdcs.ed.ac.uk


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
TODAY'S SCHEDULE
Seminar: What's in the Internet Archive? A big (meta)data analysis
Hands-on session 1: Introduction and HTML
Hands-on session 2: Web Scraping
BYOD session: 1

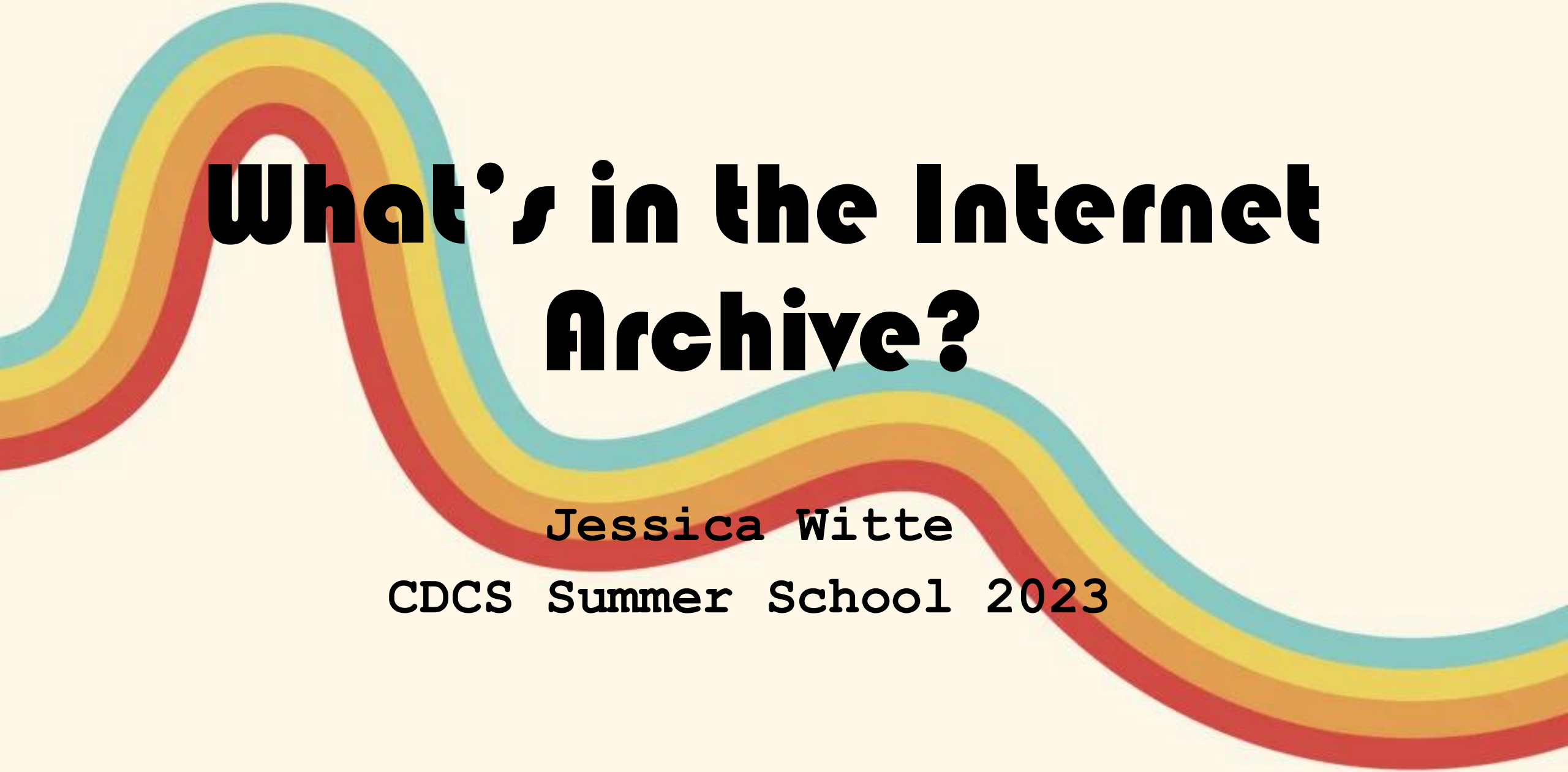
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

WHAT'S IN THE INTERNET ARCHIVE?
A BIG (META)DATA ANALYSIS

Dr Jessica Witte,

Post-doctoral fellow at the University of Edinburgh

*image borrowed from etsy



What's in the Internet Archive?

Jessica Witte

CDCS Summer School 2023

“I believe we have a massive project to do together, which is to put the best we have to offer within reach of our kids . . . we need to move all the best works online and then find mechanisms to serve these to anyone who wants them. We need to do this now . . . **I am afraid that our digitization will be selective,** that only dominant languages, dominant cultures, and dominant points of view will be represented in the digital future**. If we are biased in our selection of what we bring to the next generation, then we are committing a crime that will never be forgiven.**”

--Brewster Kahle, founder of the Internet Archive (interviewed by Ana Parejo Vadillo, 2015)

The Internet Archive (and its Text Archive)

- Founded in 1992 by Brewster Kahle, an American librarian with a background as a computer engineer
- Vision: a “Library of Alexandria” of the digital era
- Calls itself a library, but publishers disagree
- Lost a copyright lawsuit in March 2023
- Text Archive holds over 37 million items in nearly 27,000 collections
- Highly accessible: anyone with an account can contribute and read materials

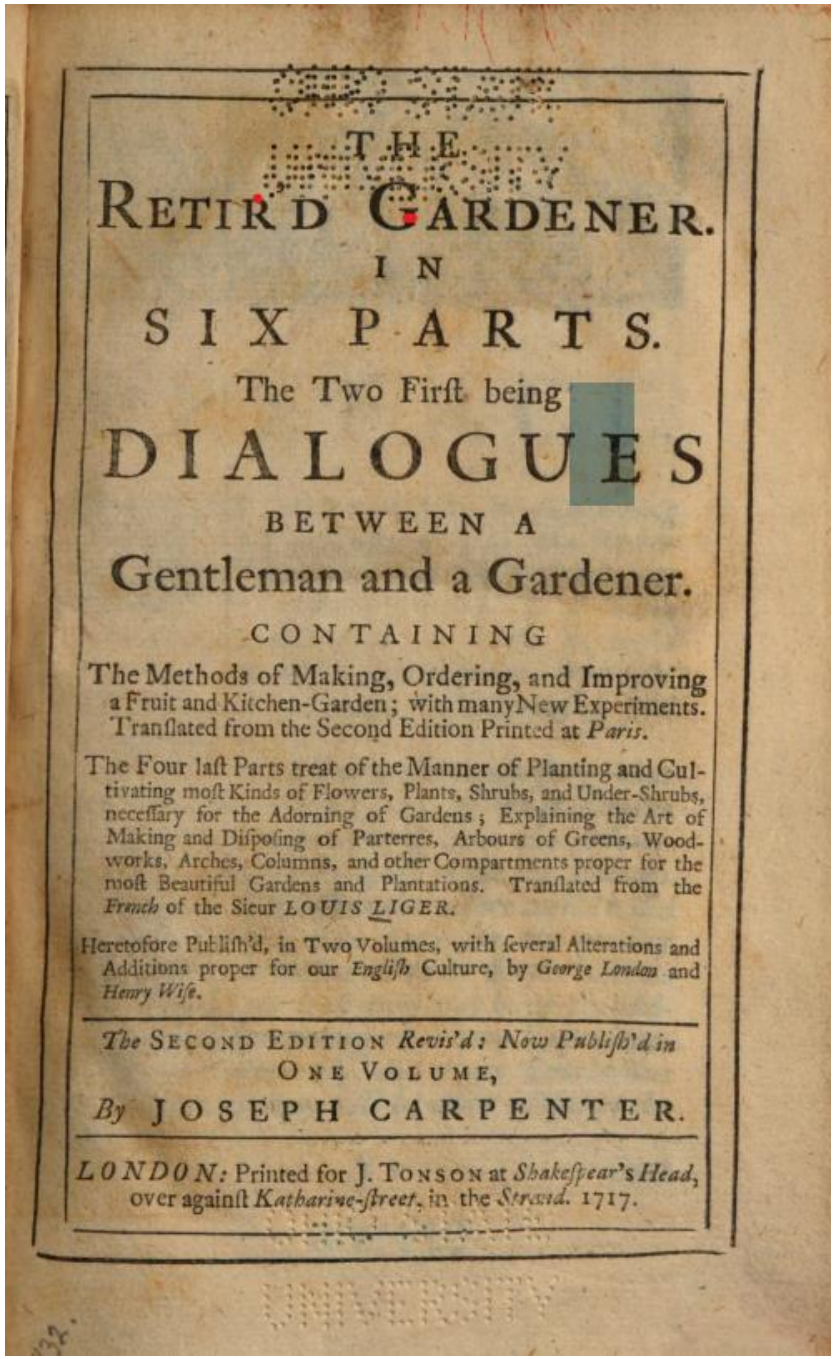
What's in the Text Archive?

- **The expected**: novels, cookbooks, drama, poetry, newspapers, magazines, nonfiction literature, self-help guides, spiritual & religious texts, children's books
- **The unexpected**: personal CVs, academic articles on microbiology, formerly classified government papers
- **The quirky**: user manuals for transistor radios, self-published poetry about turning 40, cartoons, maps, vintage primary school textbooks
- **The disturbing**: right-wing propaganda, conspiracy theories, pro-extremist content, etc.

Digital libraries: what we know

- Digital libraries have a problem with overrepresentation and bias
  - Multiple “filters” in the digitization pipeline
  - Globally, technological infrastructure is unequal
- Digitised texts are only as accurate as the OCR (book scanning) technology that produced them
- Resources, quality/scope of collections, and digital tools differ between institutions
- Not all physical objects are suitable for digitization (e.g. large folios, fragile texts)

Box of puns.com/book-puns



THE
RETIRD GARDENER.
IN
SIX PARTS.
The Two Firſt being
DIALOGUES
BETWEEN A
Gentleman and a Gardener.
CONTAINING
The Methods of Making, Ordering, and Improving a Fruit and Kitchen-Garden; with many New Experiments. Tranſlated from the Second Edition Printed at *Paris*.

The Four laſt Parts treat of the Manner of Planting and Cultivating moſt Kinds of Flowers, Plants, Shrubs, and Under-Shrubs, neceſſary for the Adorning of Gardens; Explaining the Art of Making and Diſpoſing of Parterres, Arbours of Greens, Wood-works, Arches, Columns, and other Compartments proper for the moſt Beautiful Gardens and Plantations. Tranſlated from the *French* of the Sieur *LOUIS LIGER*.

Heretofore Publiſh'd, in Two Volumes, with ſeveral Alterations and Additions proper for our *Engliſh* Culture, by *George London* and *Henry Wiſe*.

The SECOND EDITION *Revis'd: Now Publiſh'd in* ONE VOLUME,
By JOSEPH CARPENTER.

LONDON: Printed for J. TONSON at *Shakeſpear's Head*, over againſt *Katharine-ſtreet*, in the *Strand*. 1717.

#7
!
H.E.
RETIRD GARDENER .
I N
SIX PART S.
The Two Firſt being DIALOGUES
BETWEEN A Gentleman and a Gardener .
CONTAINING The Methods of Making , Ordering , and Improving
a Fruit and Kitchen - Garden ; with many New Experiments .
Tranflated from the Second Edition Printed at Paris . The Four laſt Parts tr
eat of the Manner of Planting and Cul
tivating moſt kinds of Flowers , Plants , Shrubs , and Under -
Shrubs , neceffary for the Adorning of Gardens ; Explaining the Art of Ma
king and Diſpoling of Parterres , Arbours of Greens , Wood works , Arches
, Columns , and other Compartments proper for the moſt Beautiful Gard
ens and Plantations . Tranſlated from the
French of the Sieur LOUIS LIGER . Heretofore Publiſh'd , in Two Volumes ,
with ſeveral Alterations and
Additions proper for our Engliſh Culture , by George London and Henry W
ije .The SECOND EDITION Revis'd : Now Publiſh'd in
ONE VOLUME , By JOSEPH CARPENTER .
LONDON : Printed for J. TONS ON at Shakeſpear's Head ,
over againſt Katharine - ſtreet , in the Str.3 : 1d . 1717 .
datharine

Research questions

1. Do user-driven repositories create more representative digital collections?
2. What authors, genres, languages, and time periods are represented in Text Archive?
3. Is the Internet Archive a library? What even *is* a library in the digital era?

INTERNET ARCHIVE

Workflow + process

- Scrape library metadata (using the Internet Archive's API, to follow best practice*)
- Clean metadata
  - Standardise column data
  - Check for null/empty rows
  - Identify and annotate errors
  - Verify information in a random sample
  - Drop unnecessary information
- Perform analysis

*…although unclear what the future holds

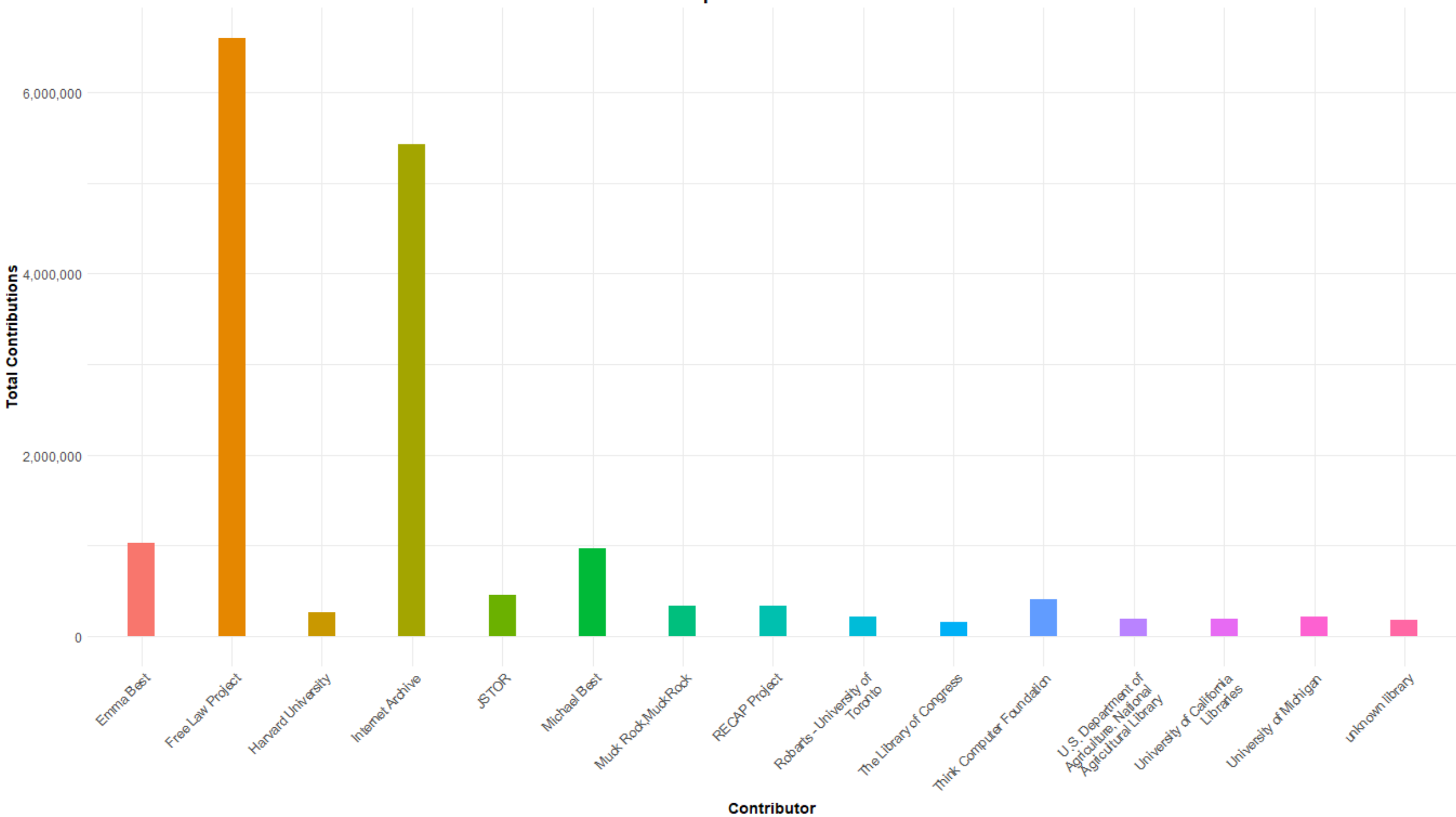
What is metadata?

- Literally “data about data”; information about objects in a collection (e.g. title, author, year of publication, language)
- Provides a large-scale snapshot of a collections’ contents
- Increasingly machine-generated

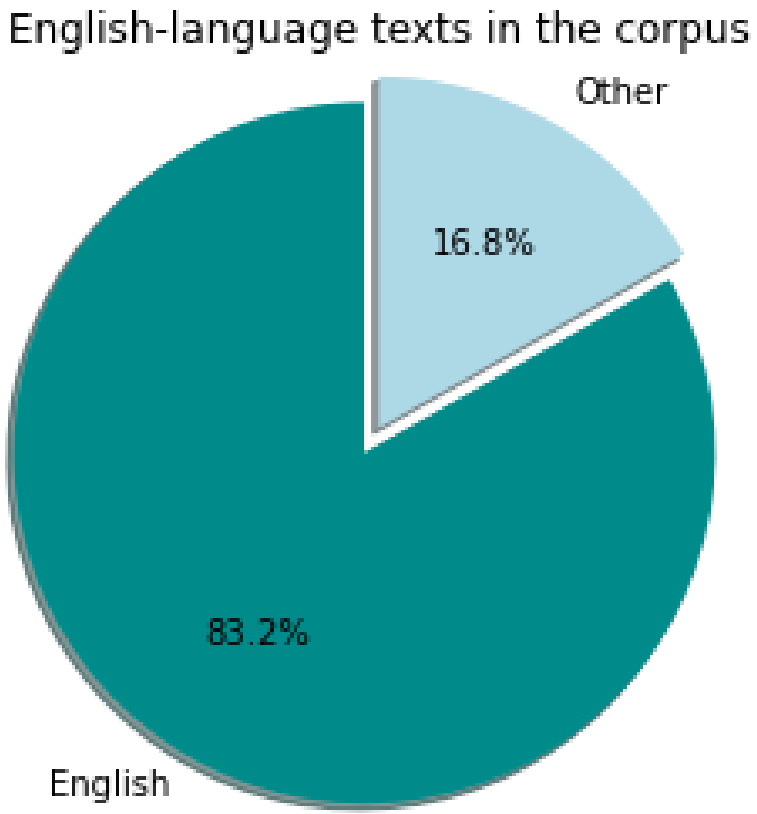
Results, problems, and more questions than conclusions...

Statistics + numbers

- Language metadata for 33,868,805 texts
- 13.09% contain limited metadata and need to be manually annotated
- 445 languages represented
- 49.07% of the texts were uploaded after 1 Jan. 2020
- Significant English-language bias: 83% of the metadata-containing corpus (compared to HathiTrust: 51%)
- Highest-frequency languages are all majority world languages
- Top contributors:primarily libraries and governmental organisations


Top 15 Contributors
Total Contributions
6,000,000
4,000,000
2,000,000
0
Emma Best
Free Law Project
Harvard University
Internet Archive
JSTOR
Michael Best
Muck Rock,MuckRock
RECAP Project
Robarts - University of Toronto
The Library of Congress
Think Computer Foundation
U.S. Department of Agriculture, National Agricultural Library
University of California Libraries
University of Michigan
unknown library
Contributor

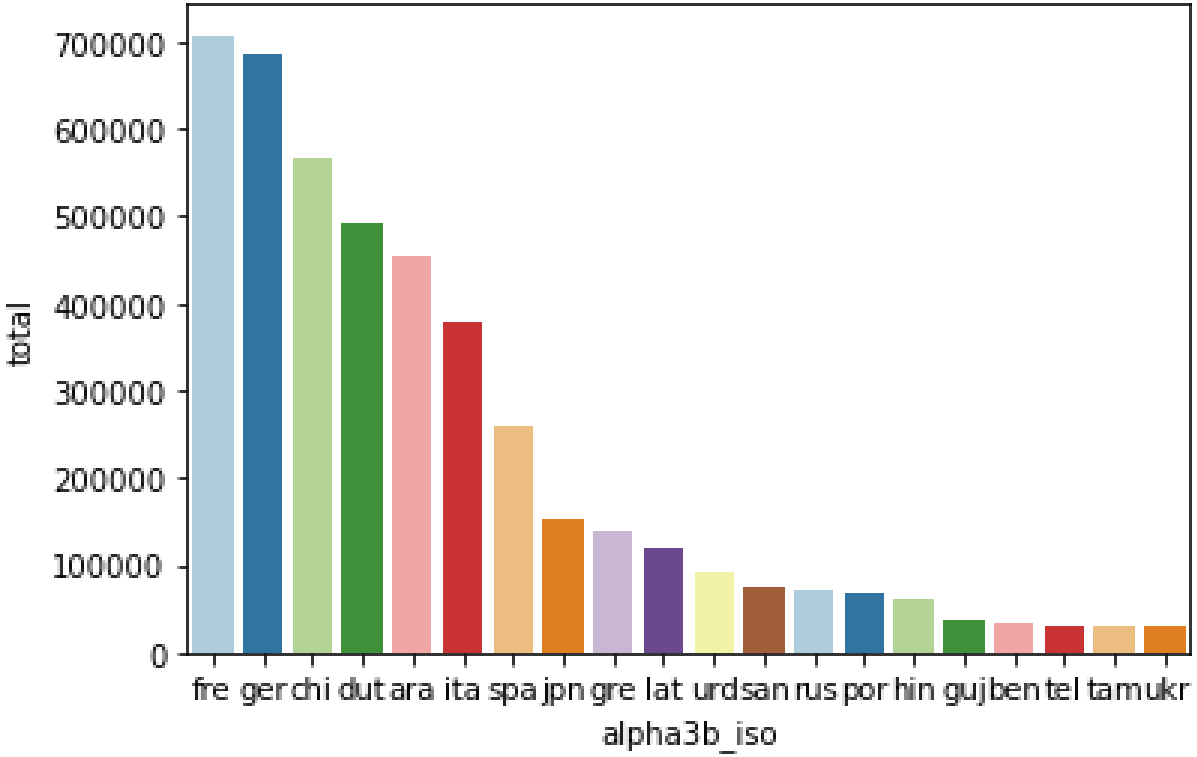
Language Balance Results



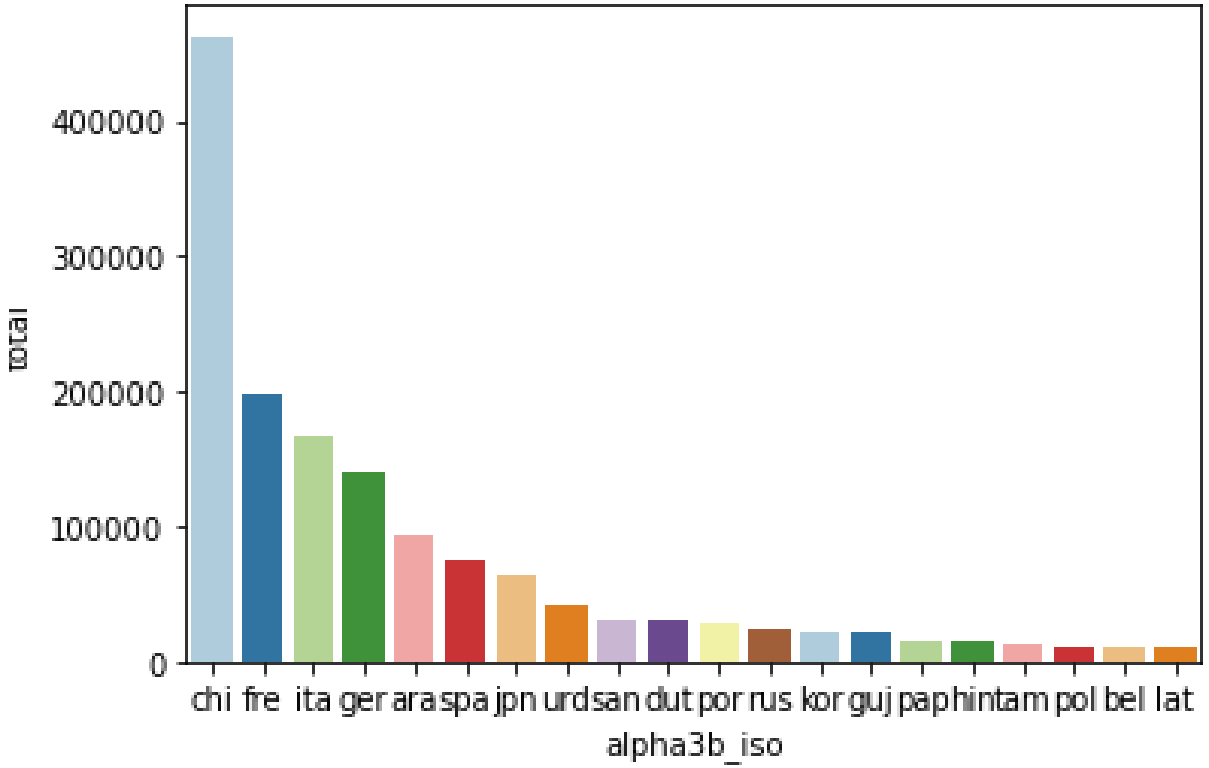
Non-English-Language Balance			
Language	Total	% of corpus	% of corpus (excl. English)
French	707195	2.40%	14.33%
German	685464	2.33%	13.89%
Chinese	566962	1.93%	11.49%
Dutch	491710	1.67%	9.96%
Arabic	452941	1.54%	9.18%
Italian	379768	1.29%	7.69%
Spanish	257678	0.88%	5.22%
Japanese	152463	0.52%	3.01%
Greek	141055	0.48%	2.86%
Latin	118392	0.40%	2.40%

language balance & the pandemic

1997-2022



COVID-19 Pandemic (2020-2021)



Problems

- Traumatic, violent content interspersed with books, magazines, and other expected library materials
- What is a library? What *isn't* a library?
- About 25% of the 446 languages are significantly underrepresented (appearing 3 or fewer times)
- Texts published in significantly underrepresented languages are almost entirely translations of the Bible facilitated by nineteenth- and twentieth-century colonisers
- Machine-created metadata with tesseract OCR erases records of non-majority languages

Sasak (2007) New Testament (Print) SABDA

by LAI

Publication date	2007
Topics	bible
Collection	folkscanomy_religion; folkscanomy; additional_collections
Language	sas
Addeddate	2022-04-23 07:47:22
External-identifier	urn:dbs:SASLAI
Identifier	SASLAI_DBS_HS
Identifier-ark	ark:/13960/s2q8f0drnkh
Ocr	tesseract 5.0.0-1-g862e
Ocr_autonomous	true
Ocr_detected_lang	jv
Ocr_detected_lang_conf	1.0000
Ocr_detected_script	Latin
Ocr_detected_script_conf	1.0000
Ocr_invalid_language	sas
Ocr_module_version	0.0.15

SHOW MORE

Rarotongan (1888) Genesis Portion

by Fiji: The Bible Society

Publication date	1888
Topics	bible
Collection	folkscanomy_religion; folkscanomy; additional_collections
Language	rar
Addeddate	2022-04-23 07:36:36
External-identifier	urn:dbs:RARMKA
Identifier	RARMKA_DBS_HS
Identifier-ark	ark:/13960/s2zzjb22z9z
Ocr	tesseract 5.0.0-1-g862e
Ocr_autonomous	true
Ocr_detected_lang	sw
Ocr_detected_lang_conf	1.0000
Ocr_detected_script	Cyrillic Latin
Ocr_detected_script_conf	0.9883 0.0117
Ocr_invalid_language	rar
Ocr_module_version	0.0.15

SHOW MORE

Challenges for future work

- Legal cases, copyright laws, API subscriptions, and "corporatisation" of data
- Examining the unlabelled objects in the metadata file
- Definition of a "library" in the past/present/future
- Preservation ethics: should we digitise everything just because we can? What should we do with extremist, counterfactual, and violent content?
- Text mining other columns of the metadata


Thank you ☺

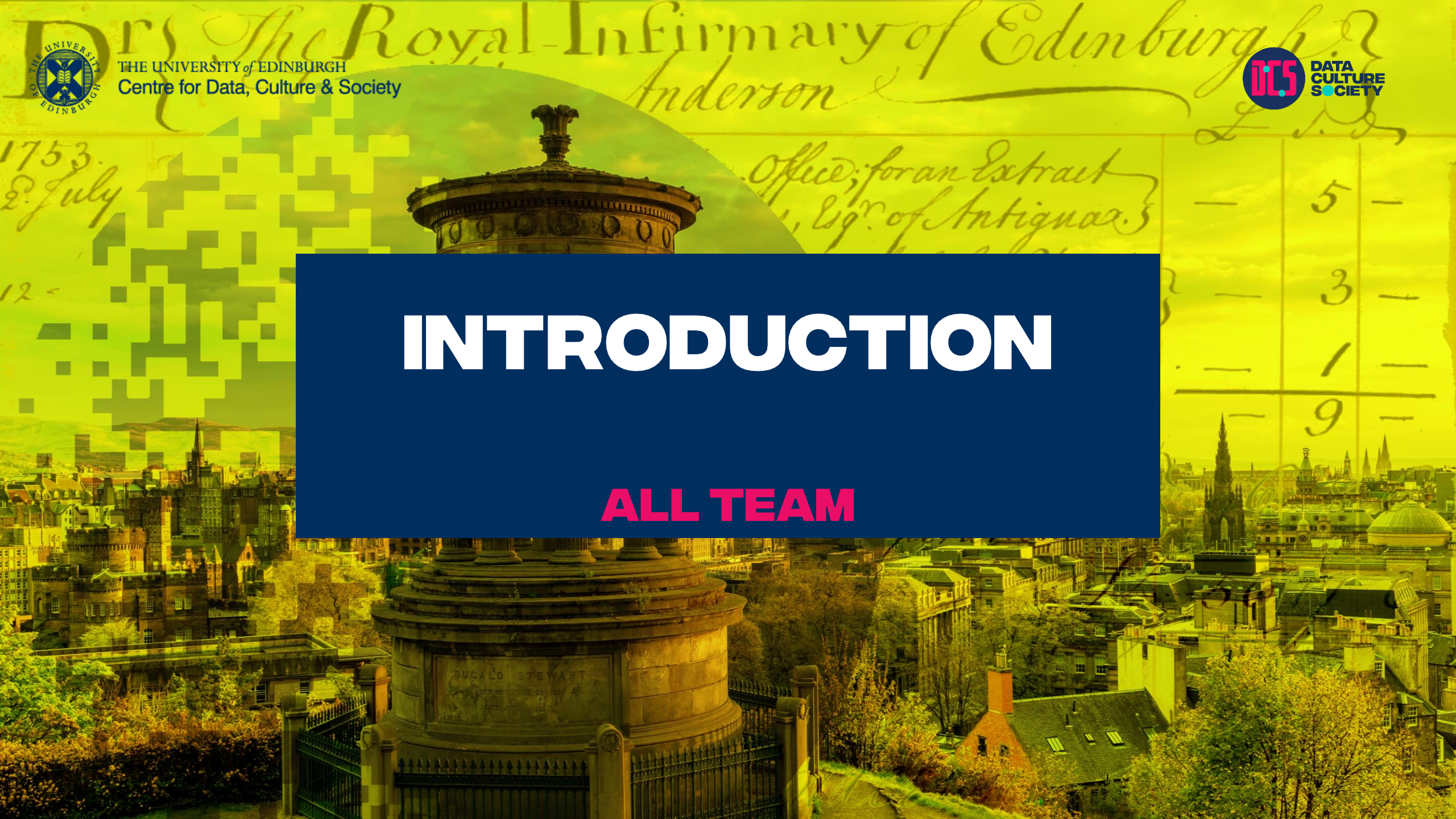

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
COFFEE BEAK
WE ARE GOING TO RESTART AT 11:00

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

INTRODUCTION

ALL TEAM



	Monday	Tuesday	Wednesday	Thursday	Friday
09:00-09:30	Registration				
09:30-09:40	Welcome	Setting Up	Setting Up	Setting Up	Setting Up
09:40-10:40	Seminar	Seminar	Seminar	Seminar	Seminar
10:40-11:00	Coffee	Coffee	Coffee	Coffee	Coffee
11:00-12:30	Introduction	Text Analysis	Sentiment Analysis	Data Analysis	Data Visualisation
12:30-13:30	Lunch	Lunch	Lunch	Lunch	Lunch
13:30-15:00	Webscraping	Text Analysis	Data Wrangling	Data Analysis	Data Visualisation
15:00-15:30	Coffee	Coffee	Coffee	Coffee	Coffee
15:30-17:00	BYOD	BYOD	BYOD	Keynote	Next Steps
Evening	Pub Quiz	Pub Crawl	Ceilidh	Drinks Reception	Dinner

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

GETTING SET FOR THE WEEK

- Sticky notes

- Getting Set on **Noteable**
  Go to
  https://noteable.edina.ac.uk/login

OUR TEAM

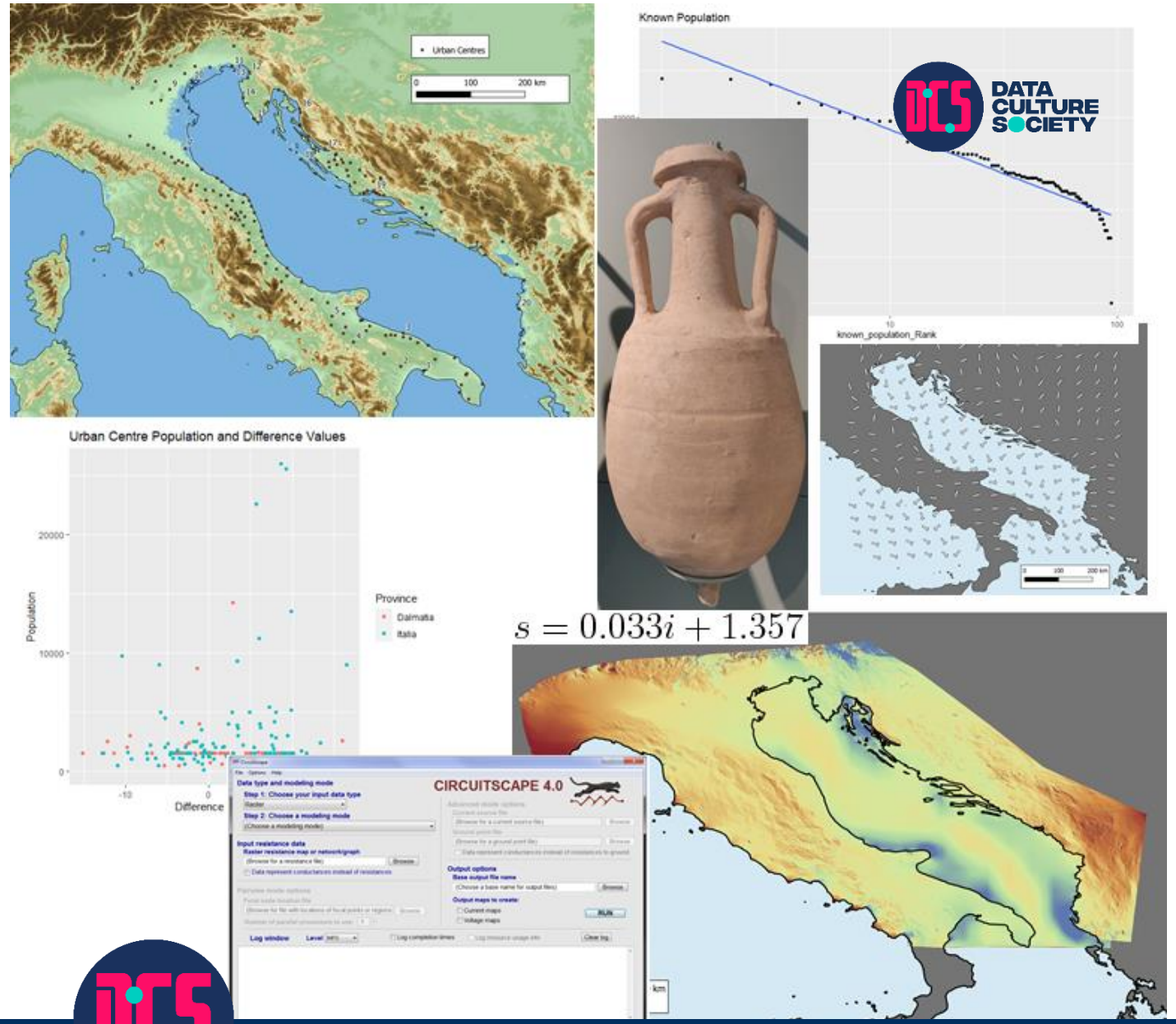
Instructors and Helpers

- "Language" barrier: The best people to teach digital research methods are researchers with a similar background
- They have been through the same steps and they will know where the "roadblocks" are

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

ANDREW MCLEAN

- PhD candidate in Classics/Archaeology at the University of Edinburgh
- He works on economy of the Roman Adriatic. He is expanding on traditional Least Cost Path (LCP) analysis by applying circuit theory to model maritime movement
- Training Fellow with DCS

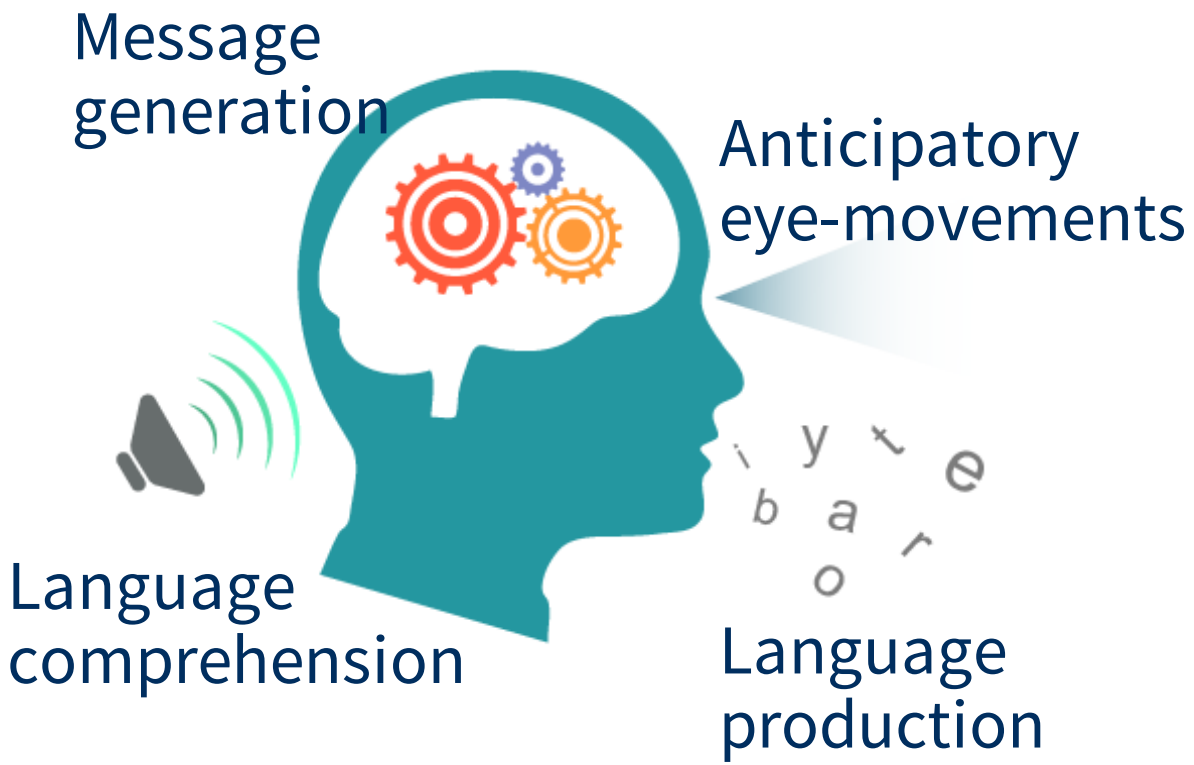


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

FANG JACKSON-YANG

- PhD candidate in the School of Philosophy, Psychology, and Language Sciences
- Her research investigates how speakers encode prominent information and how listeners predict upcoming utterances
- Training Fellow with DCS

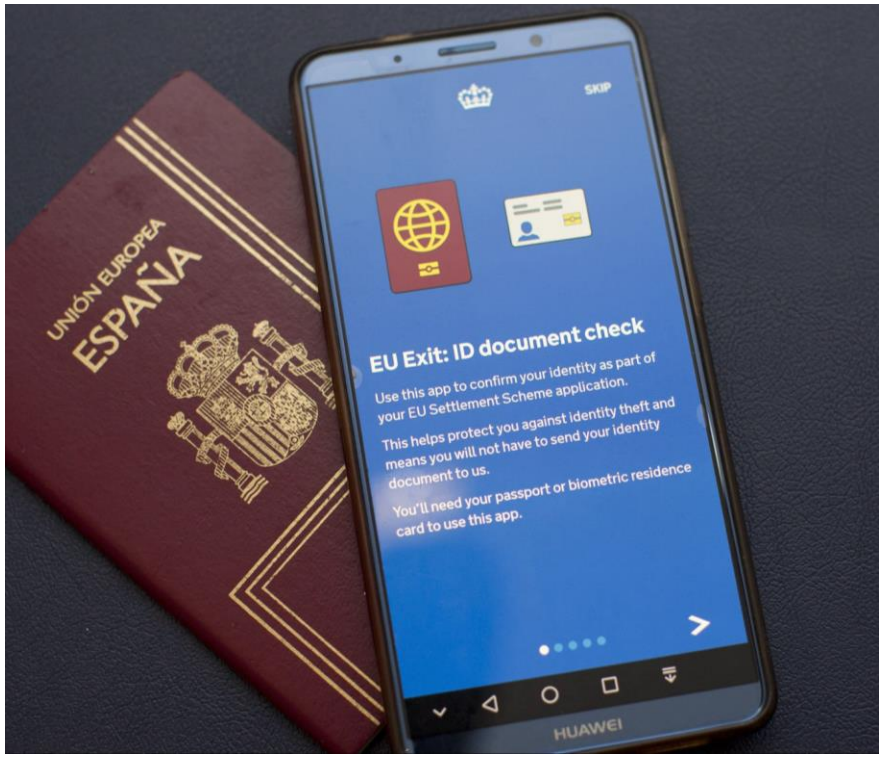


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

JAMES BESSE

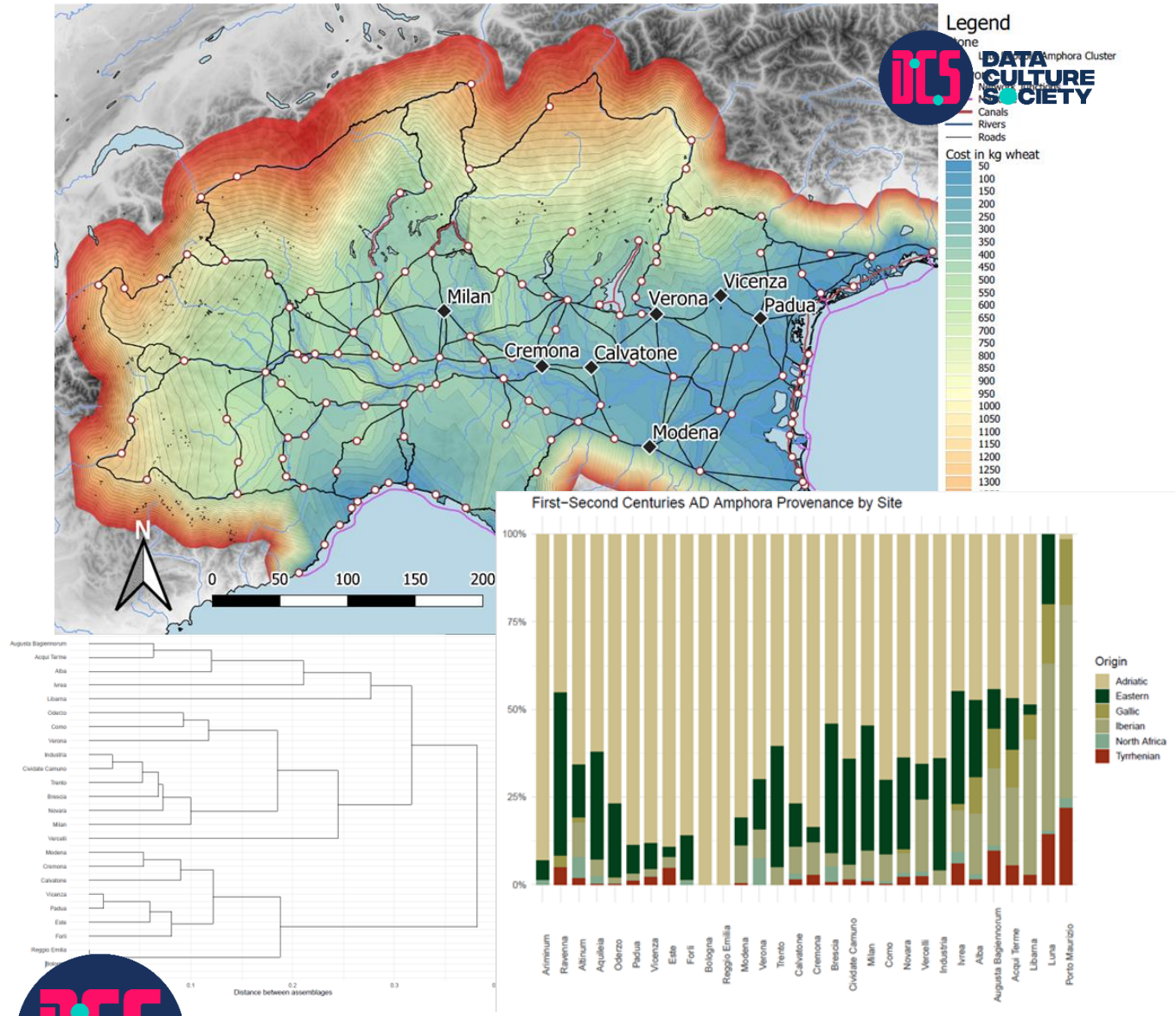
- PhD candidate in Science, Technology and Innovation Studies
- His research covers the implementation of identity and access management systems in the public sector. Primary focus is the EU Settlement Scheme.
- Training Fellow with DCS



THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

JAMES PAGE

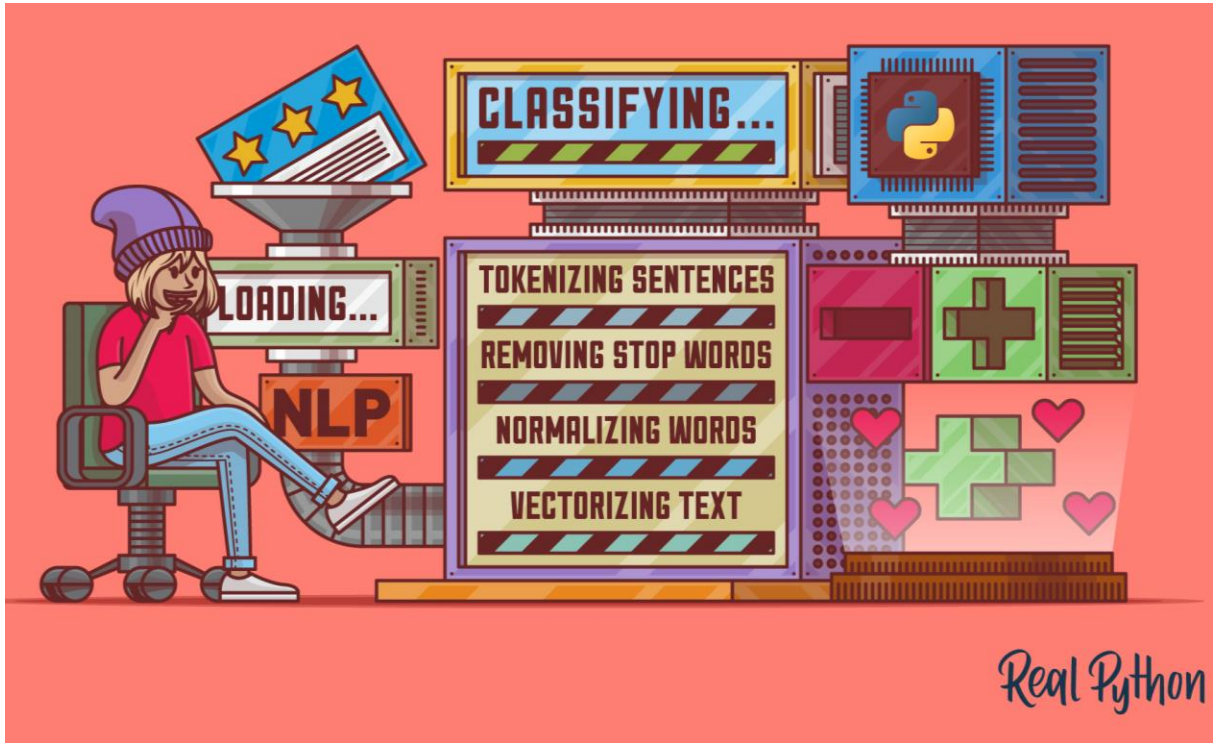
- Roman Archaeologist
- He works on trade and economy in Northern Italy during the Roman era, using network modelling and cluster analysis.
- Training Fellow with DCS



JESSICA WITTE

- Postdoctoral fellow in text & data mining (Centre for Data, Culture & Society/School of Languages, Literatures & Cultures, University of Edinburgh)

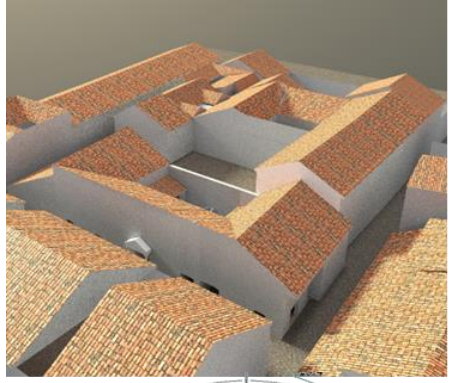
- Founder of the first official StackOverflow Fan Club
- Current research projects include:
  - NLP analysis of social media posts about Brexit
  - Deep dive into the Internet Archive's text collections
  - Data-driven medical humanities analysis of eating disorders as a disease category
  - Other assorted rabbit holes

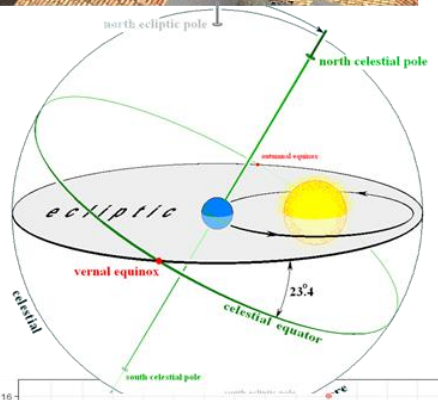


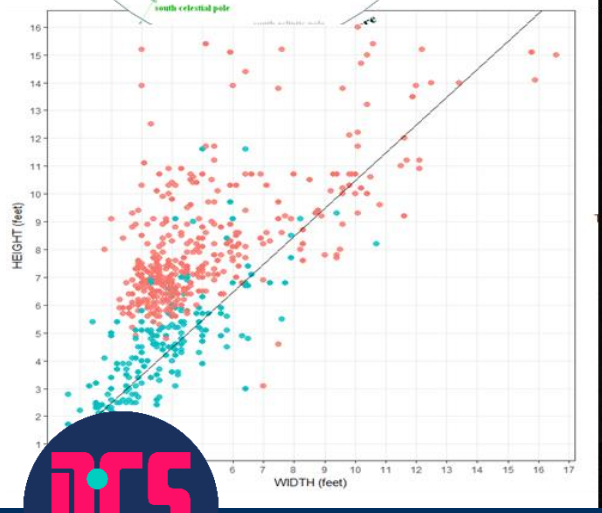
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

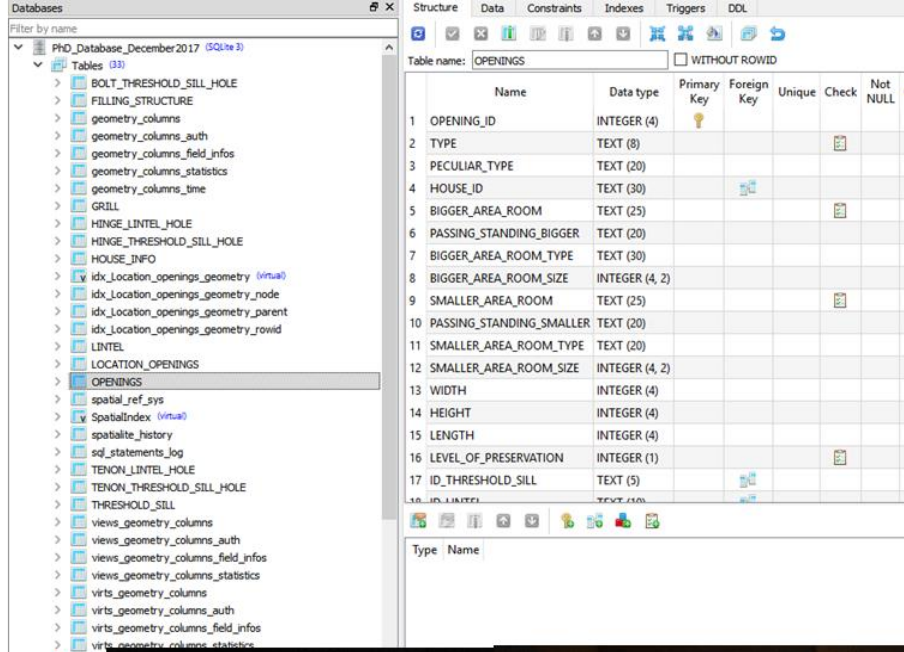
LUCIA MICHIELIN

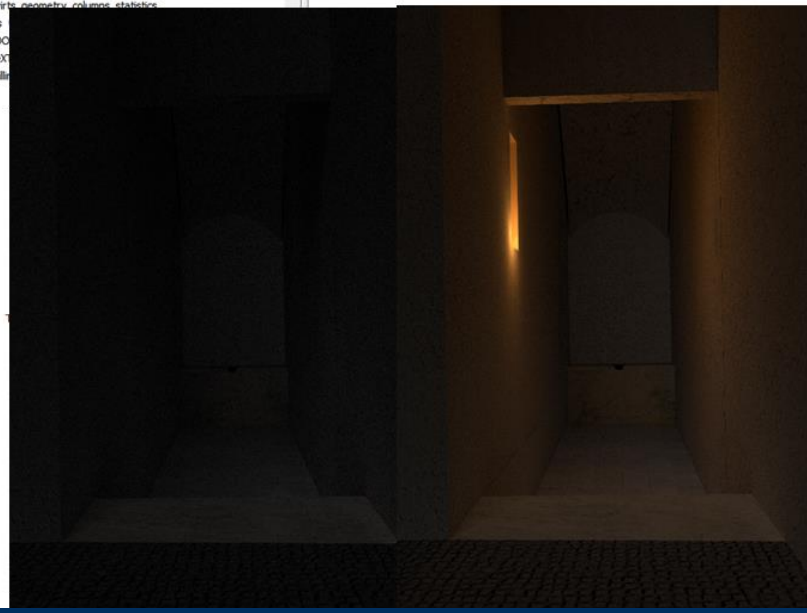
- CDCS Training Manager
- PhD in Classical Archaeology on computational analyses of doors and windows in Roman Private houses









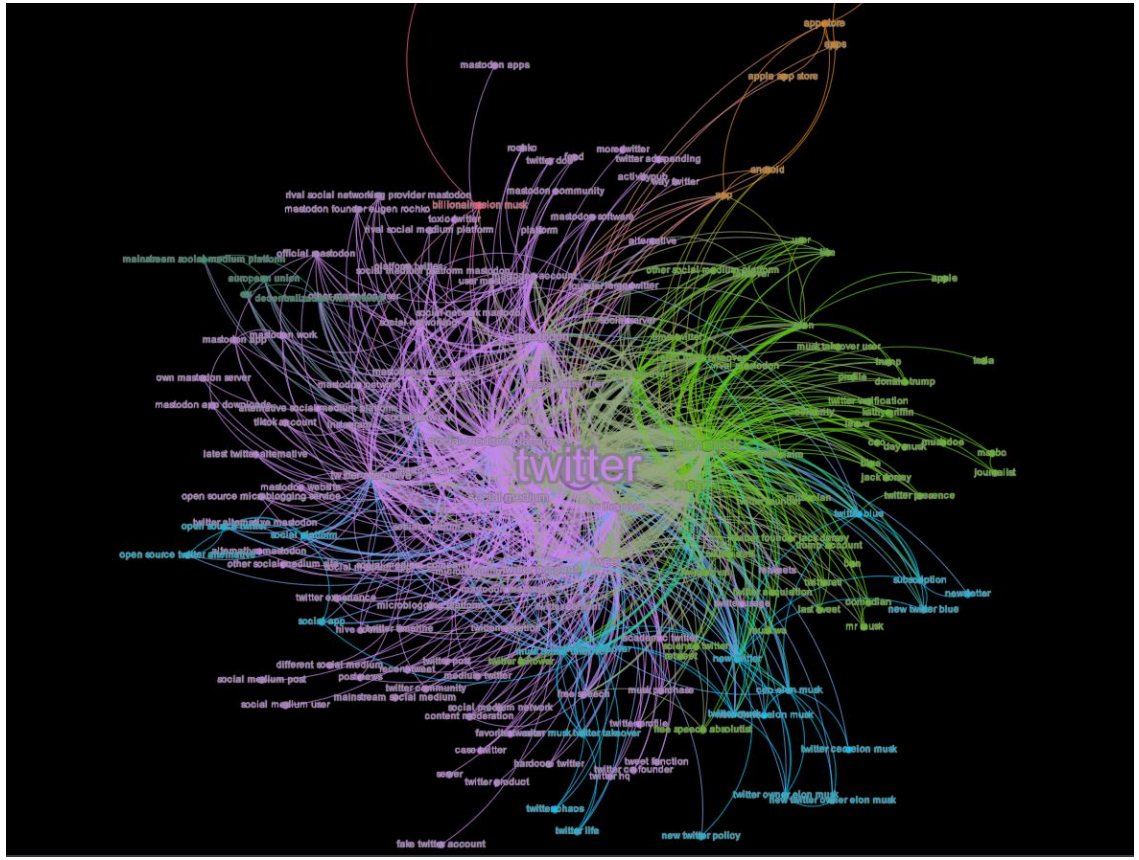


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

PEDRO JACOBETTY

- A sociologist whose research interests intersect include technology, digital culture, knowledge production and circulation, media and communication.
- Also interested in innovative ways of using digital methods for social sciences and art.



www.cdcs.ed.ac.uk

OUR DATASET

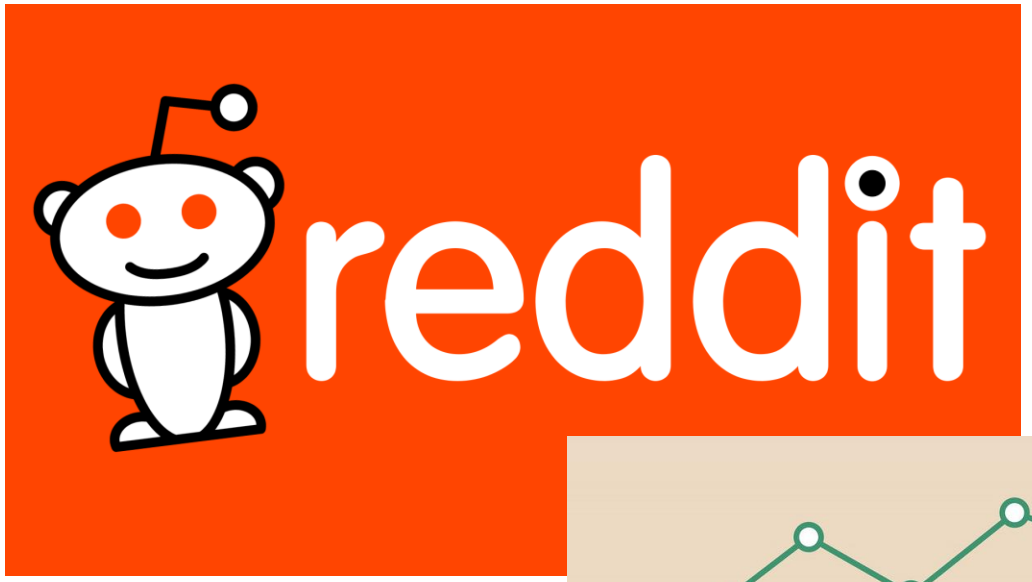
- Composite of multiple datasets.
- Real world data taken from Scottish and UK Government research, alongside research undertaken by charities.

GOV.UK

The Scottish Government
Riaghaltas na h-Alba

GOVERNMENT NEWS

- Scraped during Day 1
- News items from the UK GOV website and Scottish GOV containing the key word "cost of living"
- Subset for those published since 2020



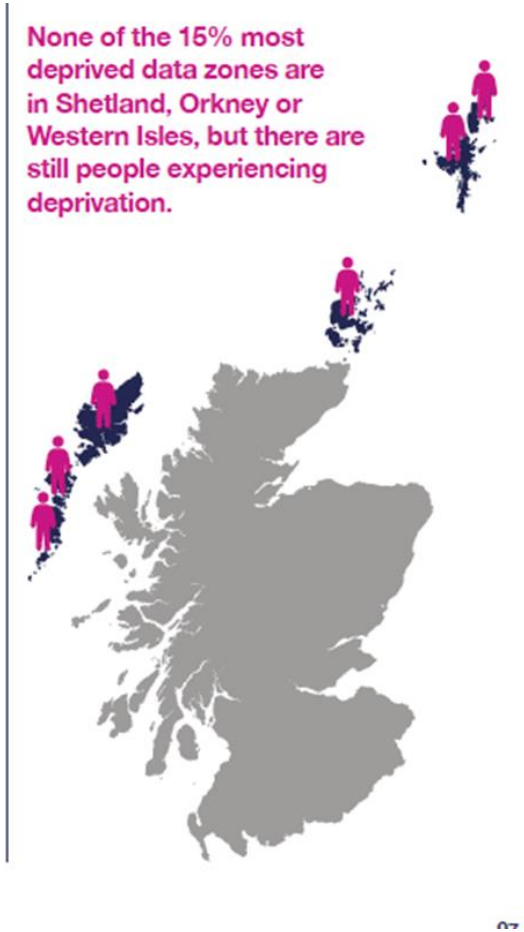
REDDIT DATA

- Scraped by Jessica Wittie
- Scraped from **r/AskUK** and **r/Scotland**
- Only those containing the key word "Cost of Living"

SCOT GOV STATS DATA

- Government statistical data collected by James Page
- From the Office of National Statistics
- Research undertaken by charities
- Indicators:
  - Homeless Applications
  - Gas Consumption
  - Energy Bills
  - Business
  - Foodbank Parcels
  - Rents
  - House Prices
  - Food Insecurity
  - Life Expectancy
  - Welfare Applications

SIMD DATA



- Scottish Index of Multiple Deprivation
- Scottish Government's official tool for identifying areas in Scotland of concentrations of deprivation by incorporating several different aspects of deprivation
- Last collection of data took place in 2020
- https://www.isdscotland.org/products-and-services/gpd-support/deprivation/simd/

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

OUR RESEARCH QUESTIONS

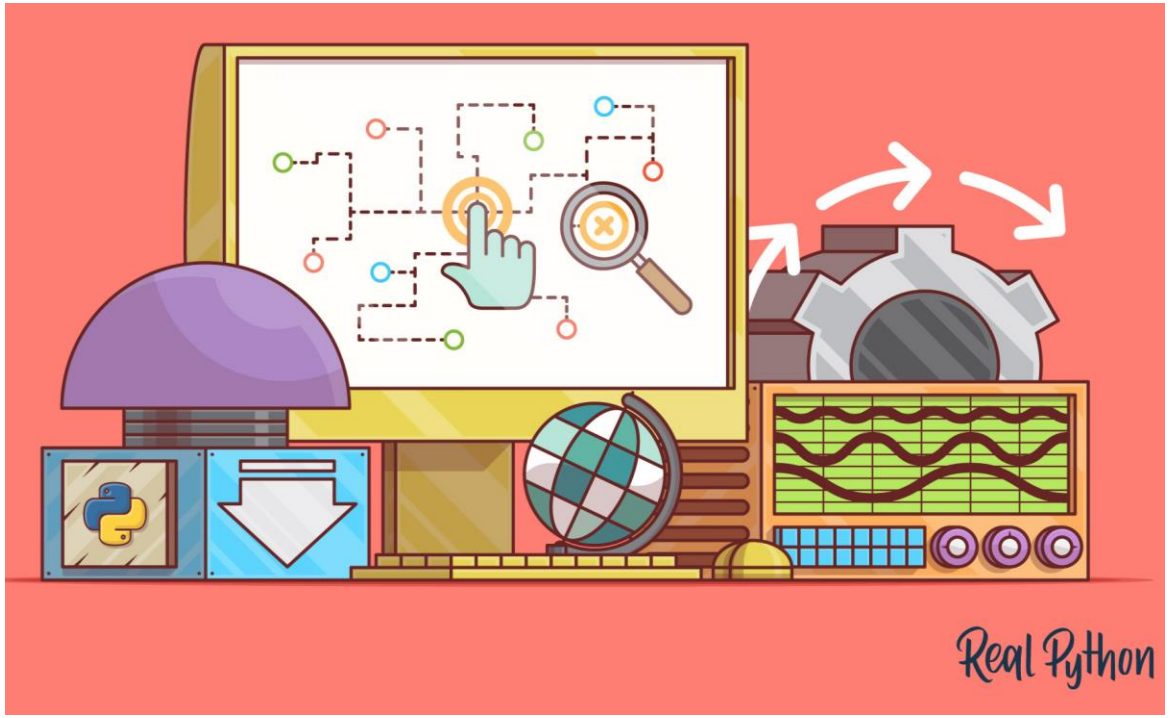
- How is the cost of living crisis portrayed by the Scottish and UK Governments?
- Is the cost of living crisis getting better in 2023?
- Are parts of Scotland affected differently by the cost of living crisis?

THE TECHNIQUES

- Web scraping
- Text Analysis
- Topic Modelling
- Sentiment Analysis
- Data Wrangling
- Linear Regression
- NHT
- Cluster Analysis
- PCA
- Data Visualisation

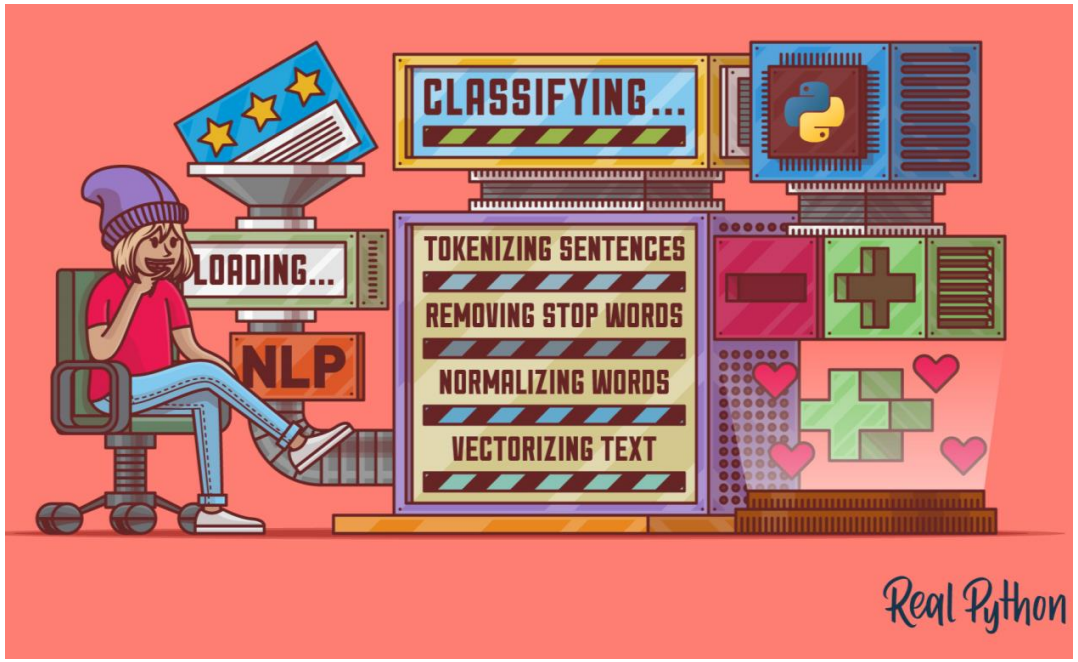
WEB SCRAPING

- Web scraping refers to the extraction of data from a website.
- This information is collected and then exported into a format that is more useful for the user.
- If you've ever copied and pasted information from a website, you've performed the same function as any web scraper, only you manually went through the data scraping process.
- Unlike the lengthy process of extracting data by yourself, web scraping uses intelligent automation to retrieve hundreds, millions, or even billions of data points from the internet.



TEXT ANALYSIS

- Computationally evaluating, investigating, and exploring textual (natural language) data
- Supervised and unsupervised methods
- Topic modelling, named entity recognition (NER), sentiment analysis, text classification



Some caveats:

- Training data matters
- Computers “read” differently than we do
- Potential for quantitative/qualitative mismatch in tools, results, statistical tests, etc.

TOPIC MODELLING

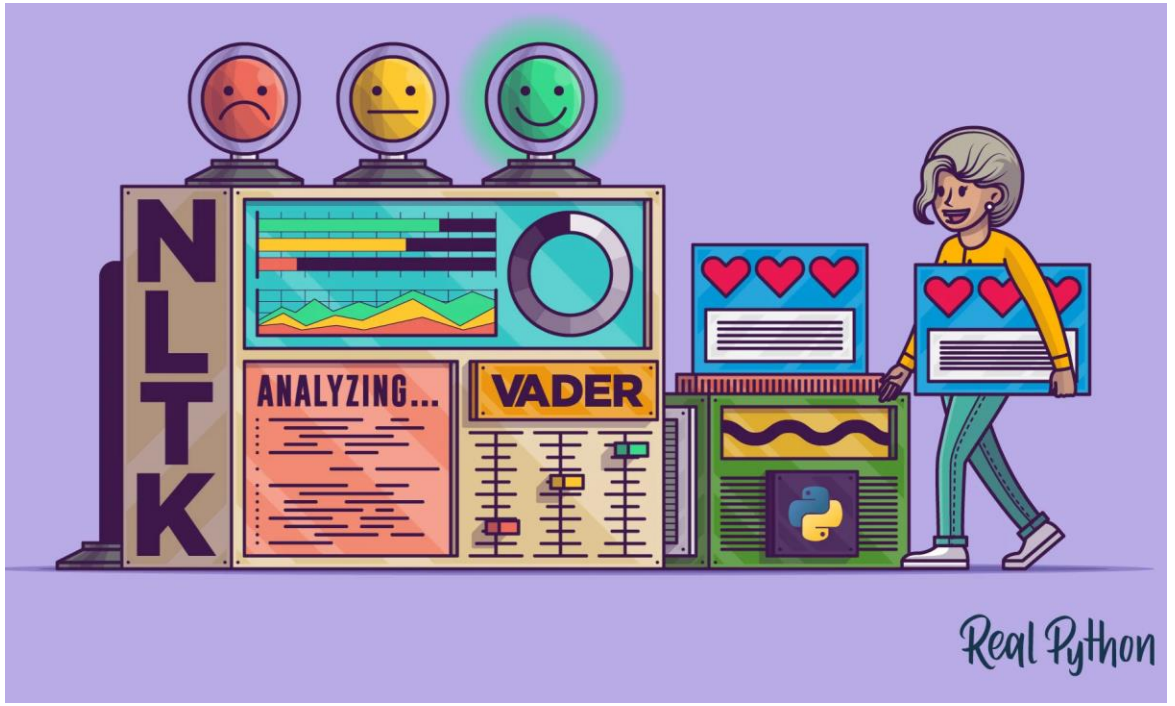
- Unsupervised method
- “Bag of words” (BoW) organised into a defined number of topics
- Latent Dirichlet analysis (LDA) and latent semantic analysis (LSA)
- Ideal for large, pre-processed datasets free from irony, sarcasm, humour, etc.
- Example use cases:
  - Identifying common themes or concepts Grouping documents based on similarity
  - Detecting changes in a dataset over time

Just bought a book from IKEA

https://jr.co.il/humor/humor001.htm

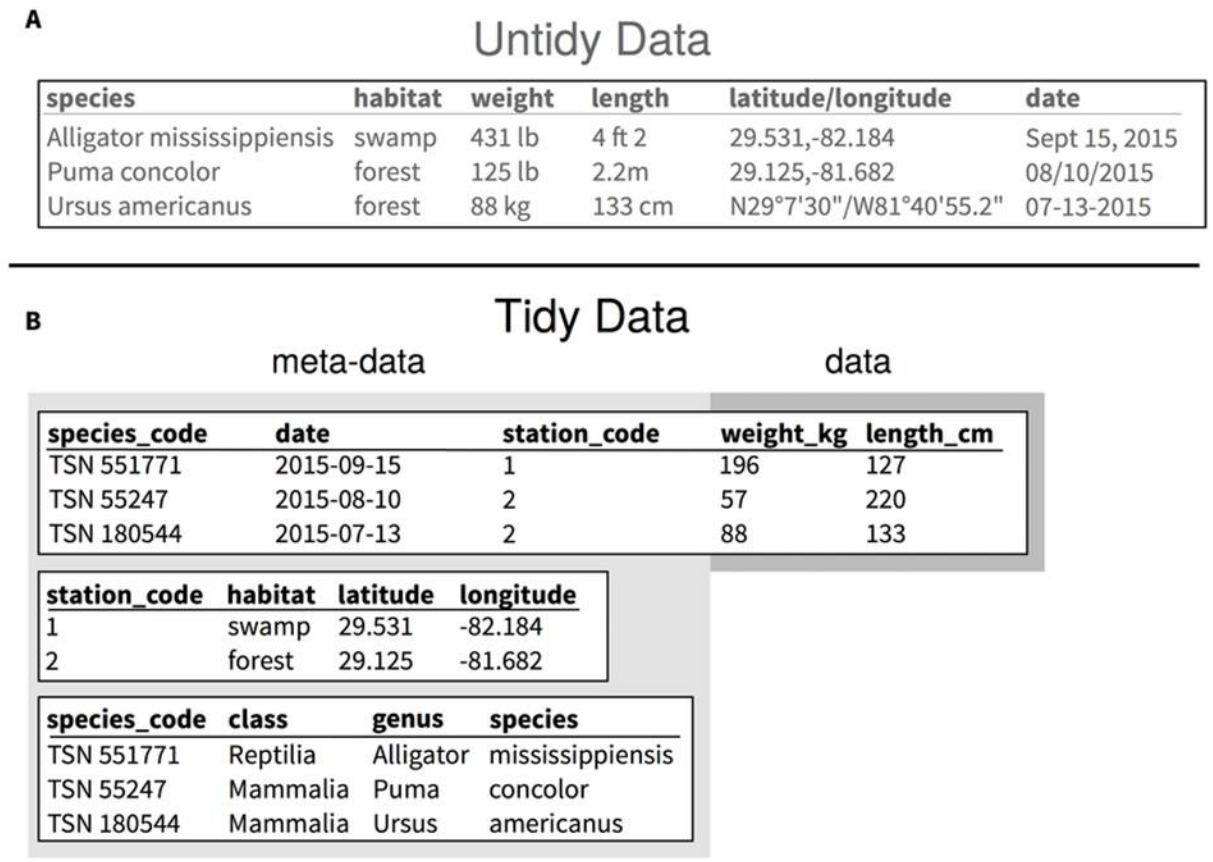
SENTIMENT ANALYSIS

- Computational detection of “emotion” defined as polarity, or relative positivity/negativity of textual data
- Rule-based or lexicon-based models
- Commonly applied at the document, sentence, or phrase level
- Performs well on opinionated text (e.g. consumer reviews, Tweets)
- Cannot detect irony, sarcasm, tone, or context
- Many existing models have only been trained on English-language text



DATA WRANGLING & TIDY DATA

- Data is often mislabelled or unorganised
- Data wrangling and tidying is simply the process of organizing this data
- The aim is to make it more computer friendly, which isn't always more human friendly



A

Untidy Data

species	habitat	weight	length	latitude/longitude	date
Alligator mississippiensis	swamp	431 lb	4 ft 2	29.531,-82.184	Sept 15, 2015
Puma concolor	forest	125 lb	2.2m	29.125,-81.682	08/10/2015
Ursus americanus	forest	88 kg	133 cm	N29°7'30"/W81°40'55.2"	07-13-2015

B

Tidy Data

meta-data

data

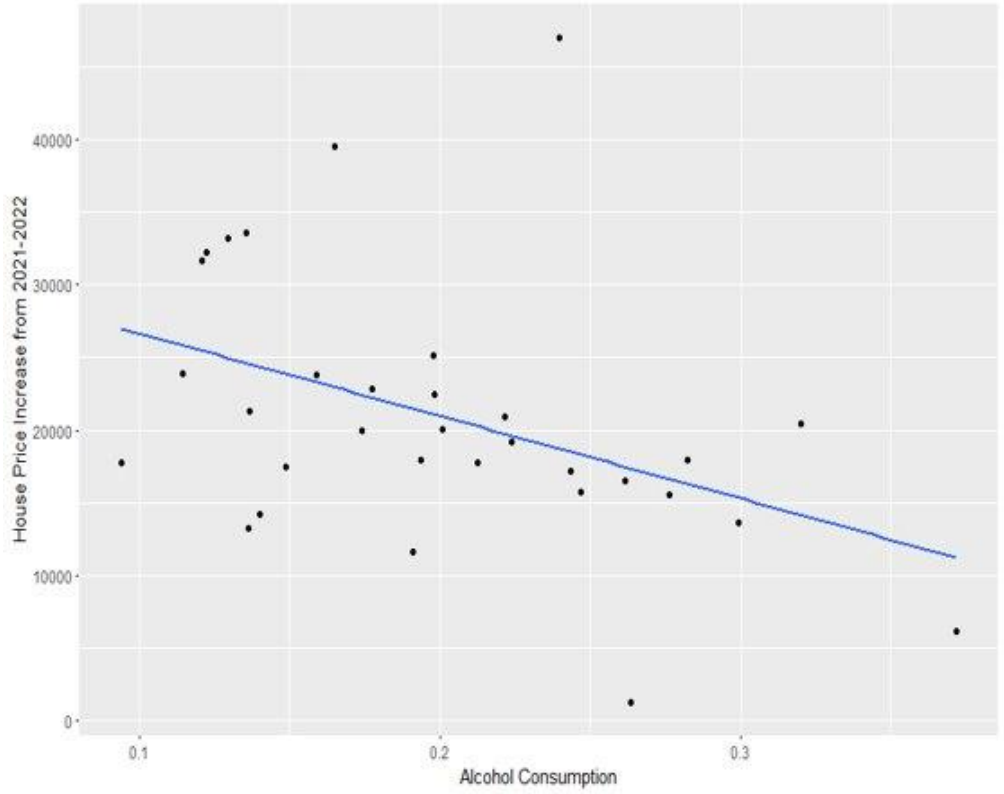
species_code	date	station_code	weight_kg	length_cm
TSN 551771	2015-09-15	1	196	127
TSN 55247	2015-08-10	2	57	220
TSN 180544	2015-07-13	2	88	133

station_code	habitat	latitude	longitude
1	swamp	29.531	-82.184
2	forest	29.125	-81.682

species_code	class	genus	species
TSN 551771	Reptilia	Alligator	mississippiensis
TSN 55247	Mammalia	Puma	concolor
TSN 180544	Mammalia	Ursus	americanus

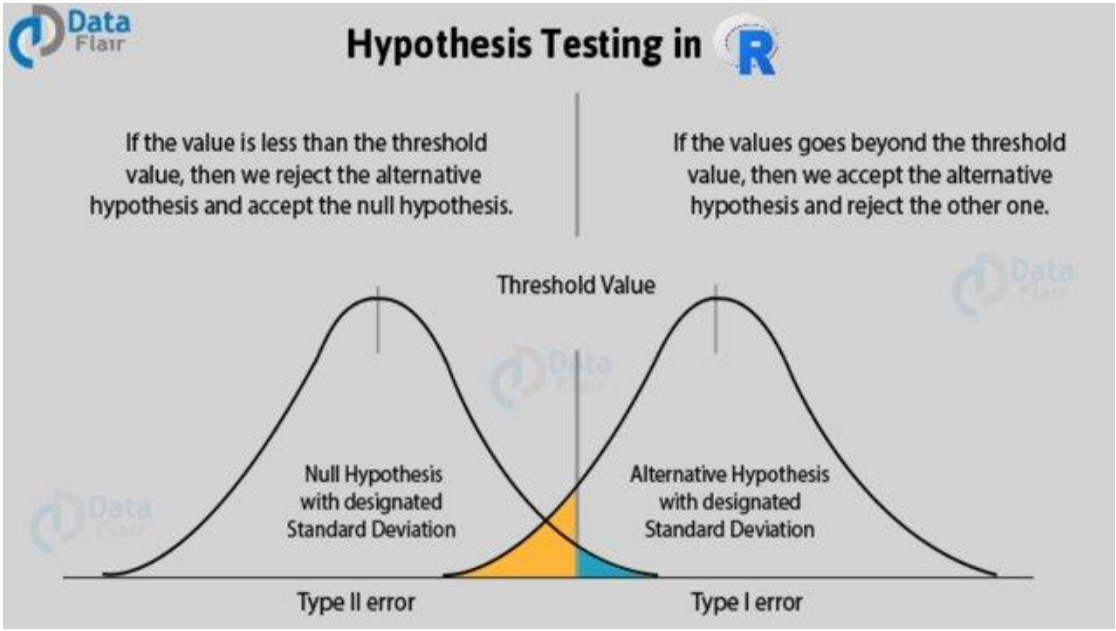
LINEAR REGRESSION

- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables.
- It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.



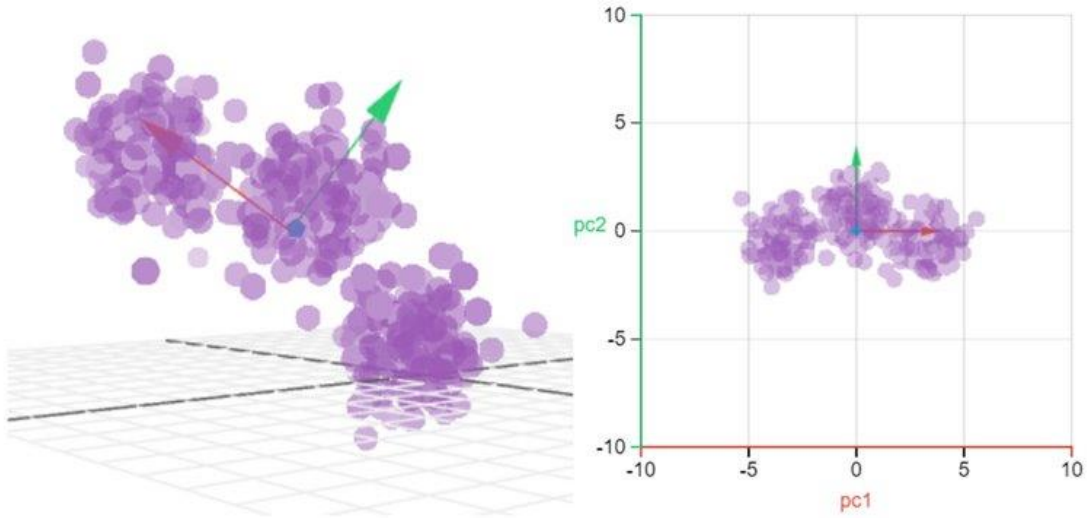
NHT

- Hypotheses are assumptions made by researchers about data they have collected or are researching.
- Hypothesis testing is a formal method for validating hypotheses- proving whether they are correct or not.
- Importantly, hypotheses do not need to be correct every time. There is merit in disproving hypotheses.



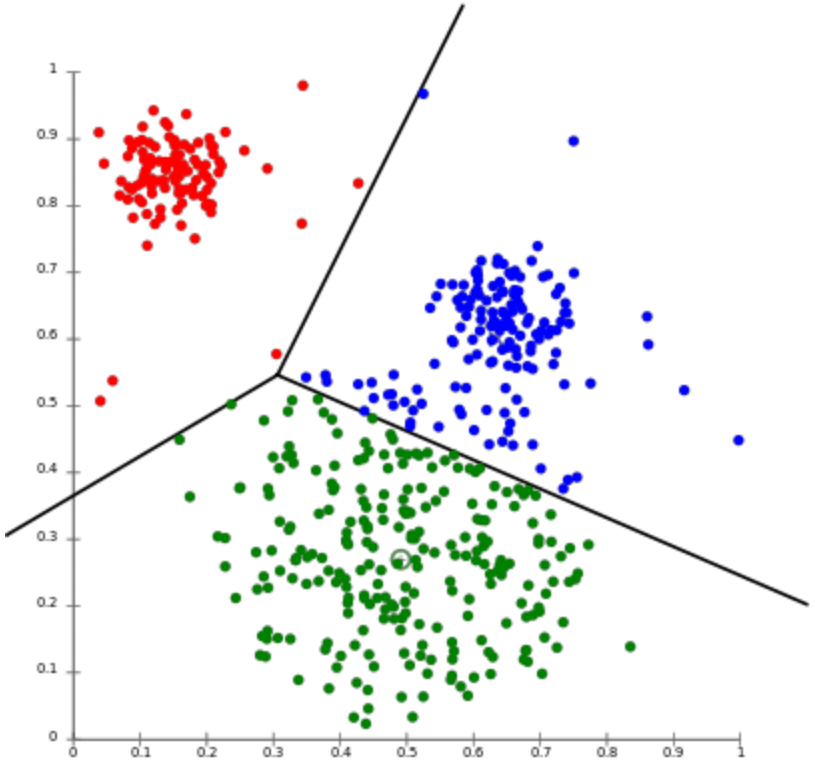
PCA

- Principal component analysis (PCA) is a technique used to emphasize variation and bring outline patterns in a dataset
- When there are more than two variables, it can be difficult to isolate underlying patterns in a dataset because it's hard to see through a cloud of data.
- PCA can be used to find out which of our variables cause the most variation in our dataset



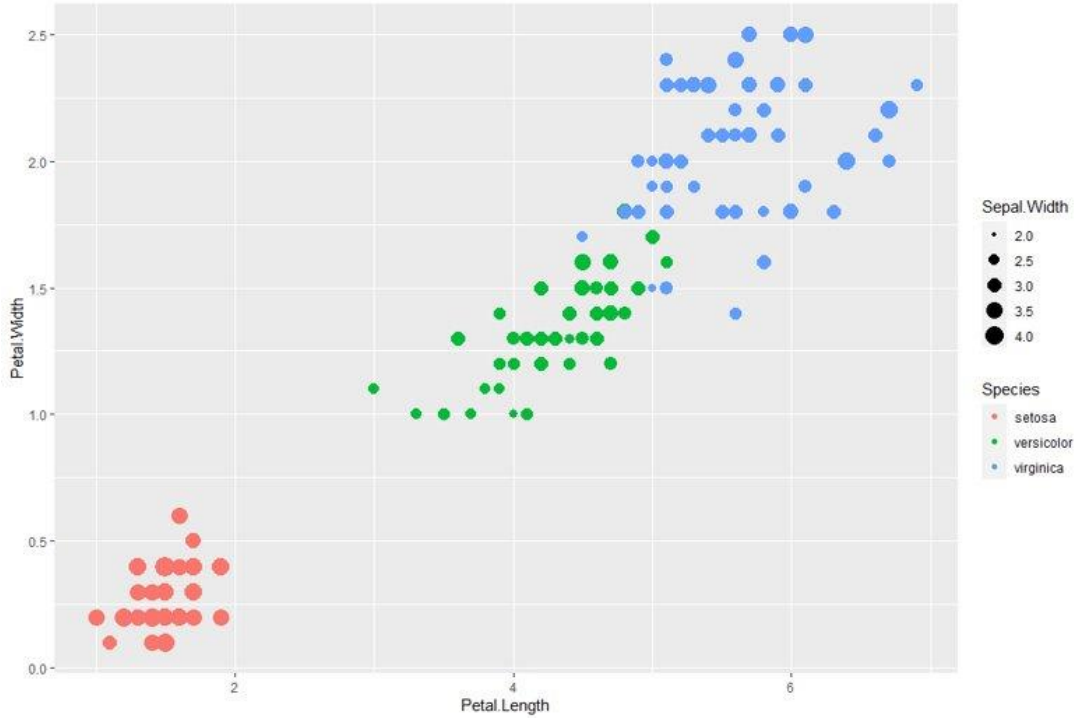
CLUSTER ANALYSIS

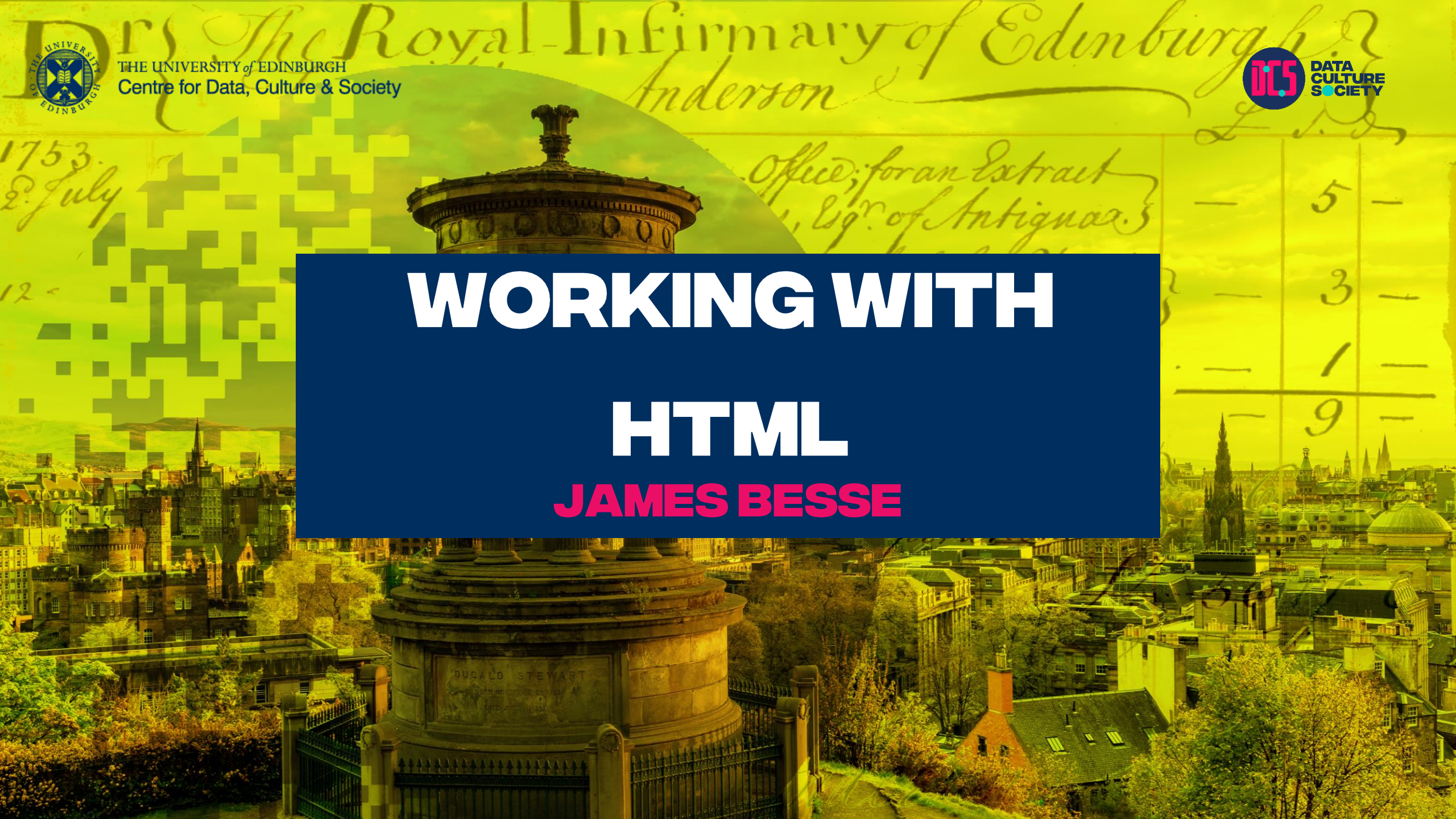
- Cluster analysis is a way of dividing objects in a dataset into groups made up of similar characteristics.
- Many different types of cluster analysis.



DATA VISUALISATION

- An important yet often overlooked part of data processing
- A good plot will elevate and enhance your results. A bad one will obscure patterns and data.



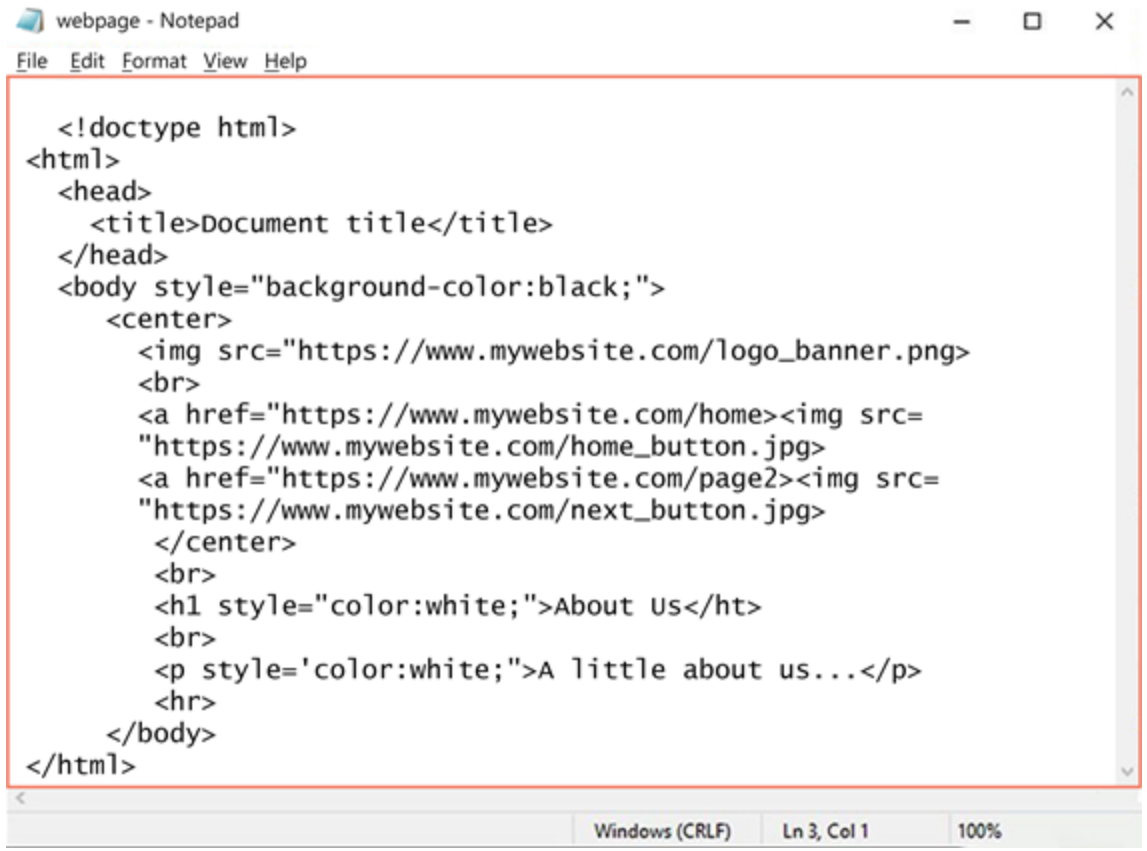

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
WORKING WITH
HTML
JAMES BESSE

HTML FOR WEB SCRAPING

- What is HTML? The standard markup language used for creating the structure and content of web pages. It defines the elements and their layout on a webpage.HTML provides a standardized way to structure and present information on the web, allowing for the creation of interactive and accessible websites that can be viewed by users across different devices and platforms.

- HTML tags are the building blocks of an HTML document. They are used to mark up and define the structure and content of different elements on a webpage. Tags are enclosed in angle brackets (< >), and most come in pairs, with an opening tag and a closing tag.

- HTML attributes provide additional information about HTML elements. They are used to modify the behavior or appearance of elements. Attributes are specified within the opening tag of an HTML element and consist of a name-value pair. For instance: The "src" attribute is used in the "img" tag to specify the source (URL or file path) of an image. The "href" attribute is used in the "a" tag to specify the target URL of a hyperlink.


webpage - Notepad
File Edit Format View Help
<!doctype html>
<html>
<head>
<title>Document title</title>
</head>
<body style="background-color:black;">

<img src="https://www.mywebsite.com/logo_banner.png>

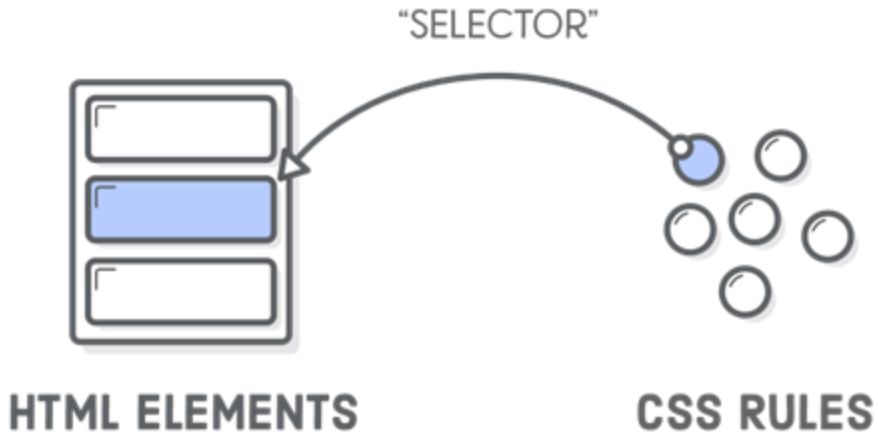
<a href="https://www.mywebsite.com/home><img src=
"https://www.mywebsite.com/home_button.jpg>
<a href="https://www.mywebsite.com/page2><img src=
"https://www.mywebsite.com/next_button.jpg>

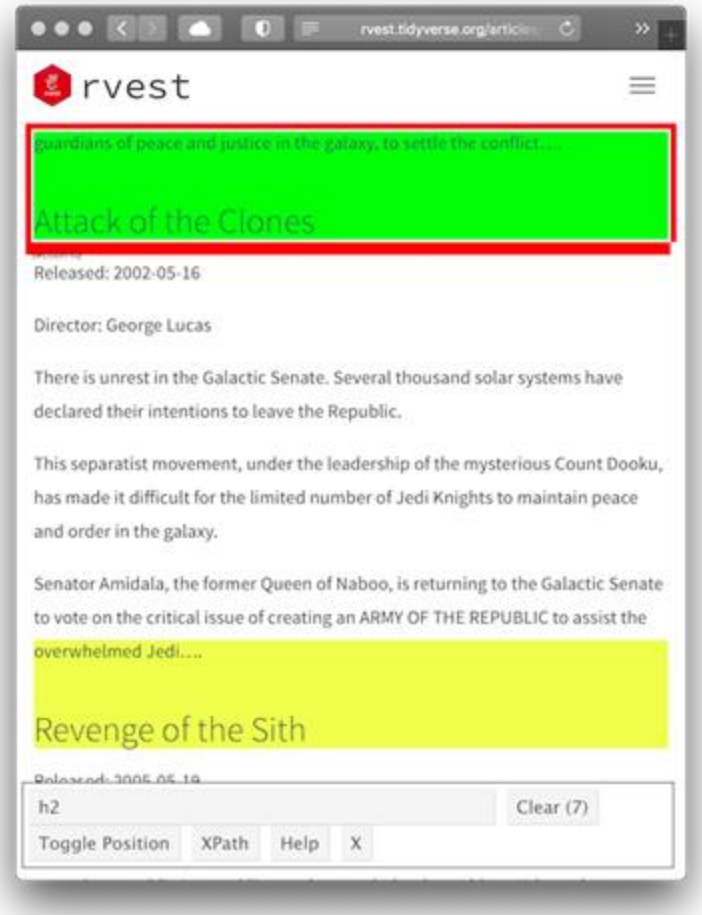
<h1 style="color:white;">About Us</ht>

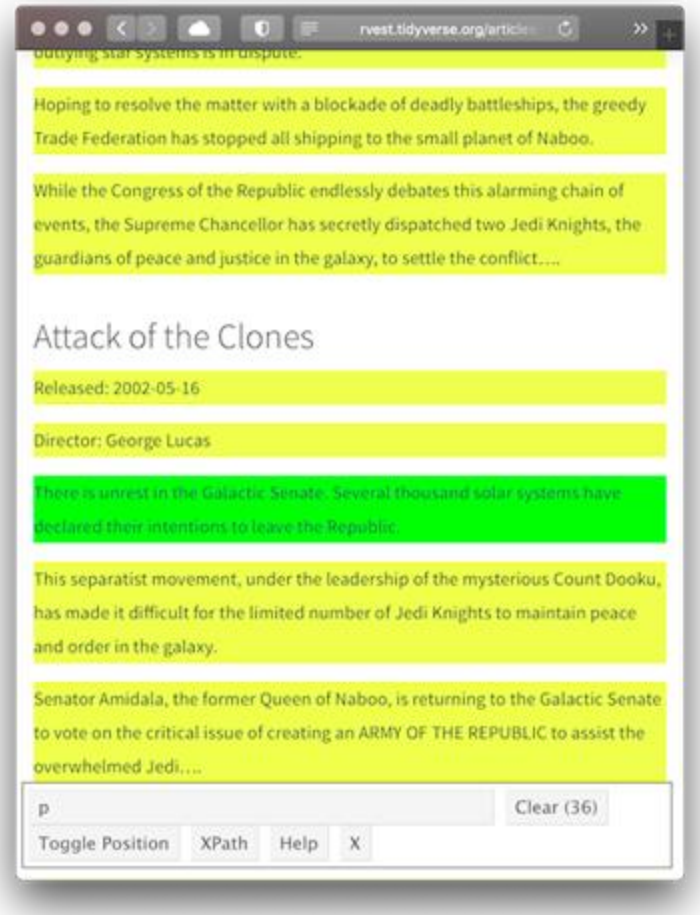
<p style='color:white;">A little about us...</p>
<hr>
</body>
</html>
Windows (CRLF)
Ln 3, Col 1
100%

CSS SELECTORS

- CSS selector gadgets, also known as CSS selector tools or generators, are web-based tools that assist in creating CSS selectors for targeting specific HTML elements.
- They offer an interactive interface where users can select elements visually, and the tool generates the appropriate CSS selector for the selected element(s).
- These gadgets analyze the structure and attributes of the HTML elements to create CSS selectors accurately, making it easier for developers to style or manipulate specific elements on a webpage.



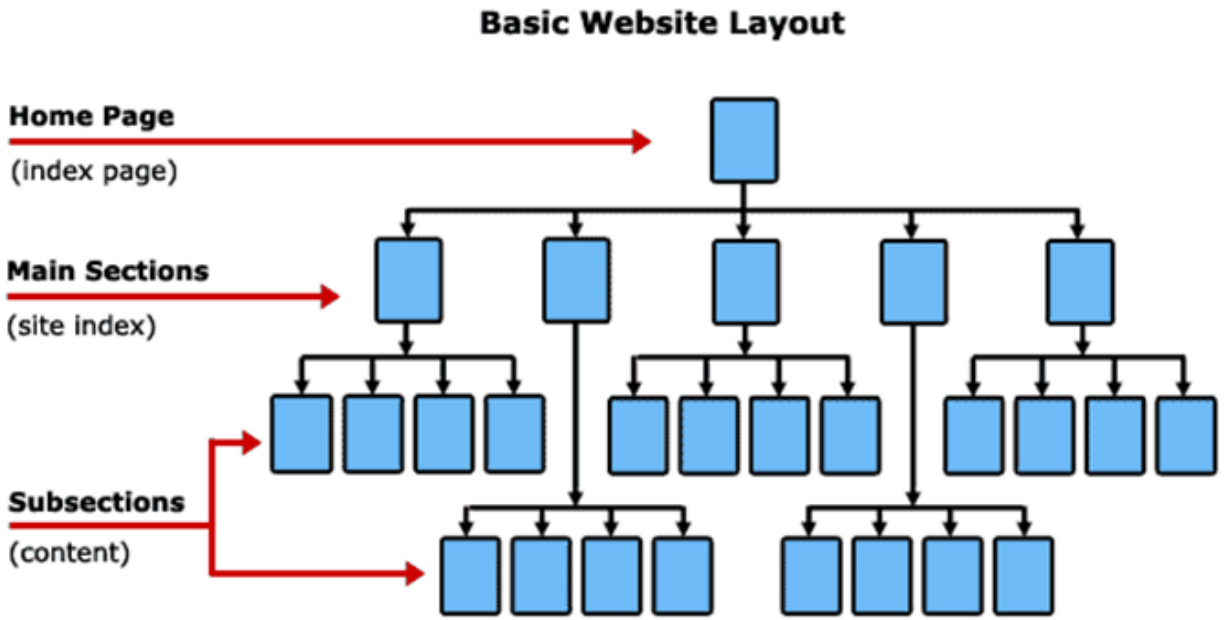

rvest.tidyverse.org/articles
rvest
guardians of peace and justice in the galaxy, to settle the conflict....
Attack of the Clones
Released: 2002-05-16
Director: George Lucas
There is unrest in the Galactic Senate. Several thousand solar systems have declared their intentions to leave the Republic.
This separatist movement, under the leadership of the mysterious Count Dooku, has made it difficult for the limited number of Jedi Knights to maintain peace and order in the galaxy.
Senator Amidala, the former Queen of Naboo, is returning to the Galactic Senate to vote on the critical issue of creating an ARMY OF THE REPUBLIC to assist the overwhelmed Jedi....
Revenge of the Sith
h2
Clear (7)
Toggle Position
XPath
Help
X


rvest.tidyverse.org/articles
Hoping to resolve the matter with a blockade of deadly battleships, the greedy Trade Federation has stopped all shipping to the small planet of Naboo.
While the Congress of the Republic endlessly debates this alarming chain of events, the Supreme Chancellor has secretly dispatched two Jedi Knights, the guardians of peace and justice in the galaxy, to settle the conflict....
Attack of the Clones
Released: 2002-05-16
Director: George Lucas
There is unrest in the Galactic Senate. Several thousand solar systems have declared their intentions to leave the Republic.
This separatist movement, under the leadership of the mysterious Count Dooku, has made it difficult for the limited number of Jedi Knights to maintain peace and order in the galaxy.
Senator Amidala, the former Queen of Naboo, is returning to the Galactic Senate to vote on the critical issue of creating an ARMY OF THE REPUBLIC to assist the overwhelmed Jedi....
p
Clear (36)
Toggle Position
XPath
Help
X

CSS SELECTORS

- A lot of traditional HTML websites are structured like trees, so we can follow a single URL down levels of sub-URLS to the specific features of the website we want

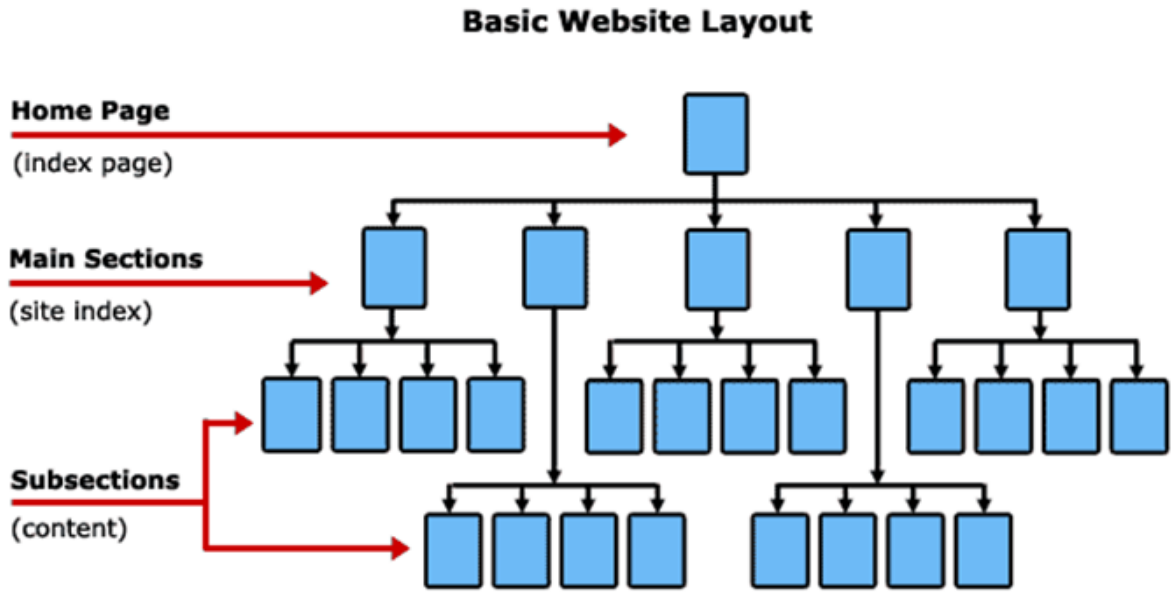
- News websites, web forums, usually follow this structure



CSS SELECTORS

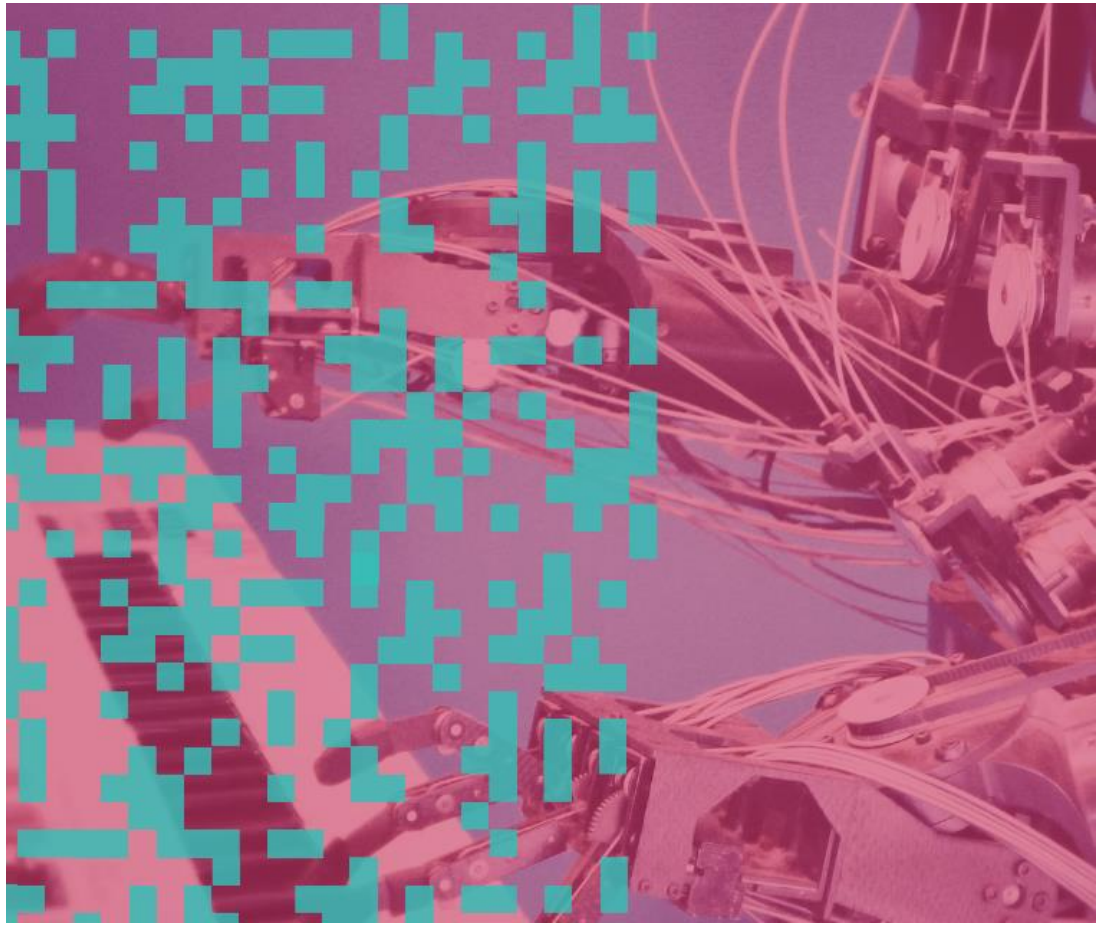
Process:

- Start with base URL
- Scrape URLs for sub-forums or news topic
- Generate list of URLS for pages
- Scrape URLs from each page
- Scrape content from each page
- Create dataframe with each article/post as a row, and each element from the article/post in its own column



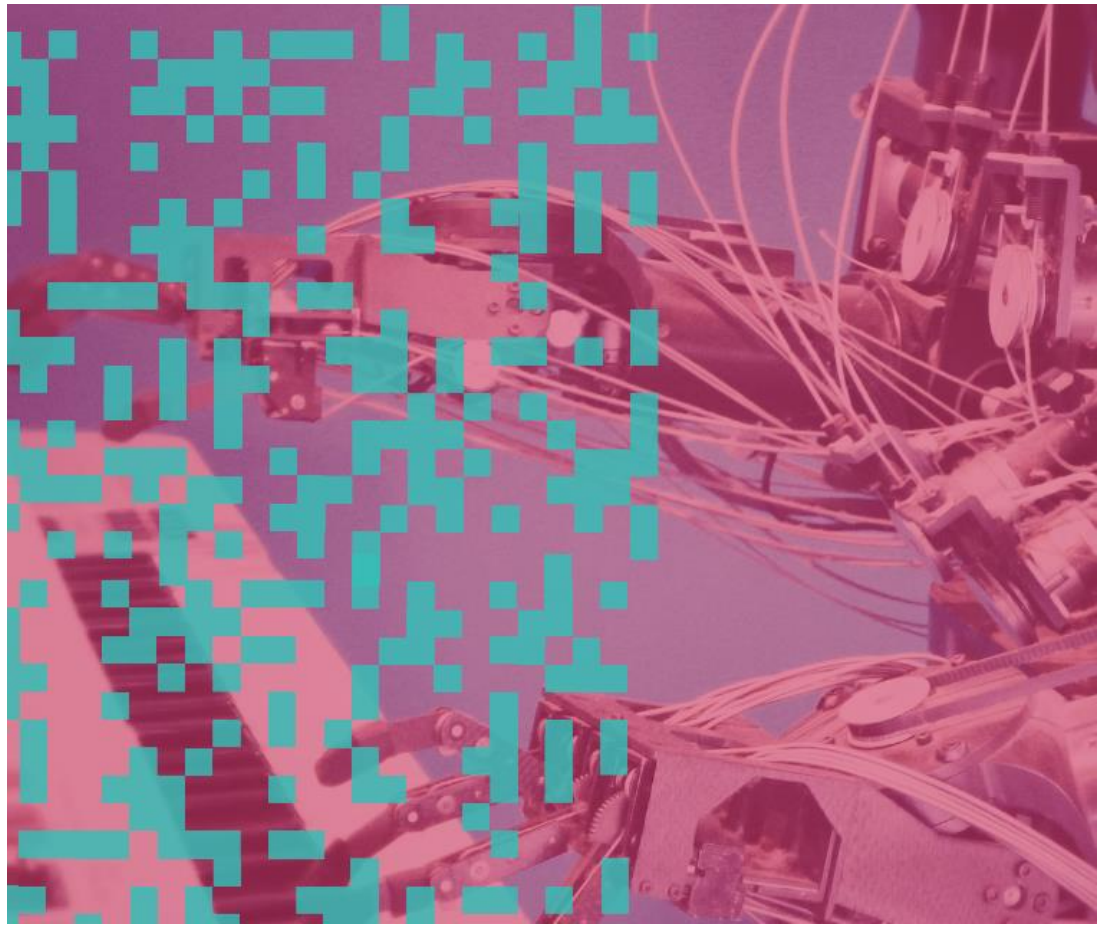
LAW AND ETHICS

- Is web scraping **legal**?
- Do you need to inform the website owners?
- Terms of service. Some websites explicitly forbid scraping.
    - Retaliation against scrapers. Facebook/ NYU.
- Public vs private websites/forums.
- Ethics guidelines typically distinguish between websites based on whether they are public/private, based on the expectations of their users about who will see the content.



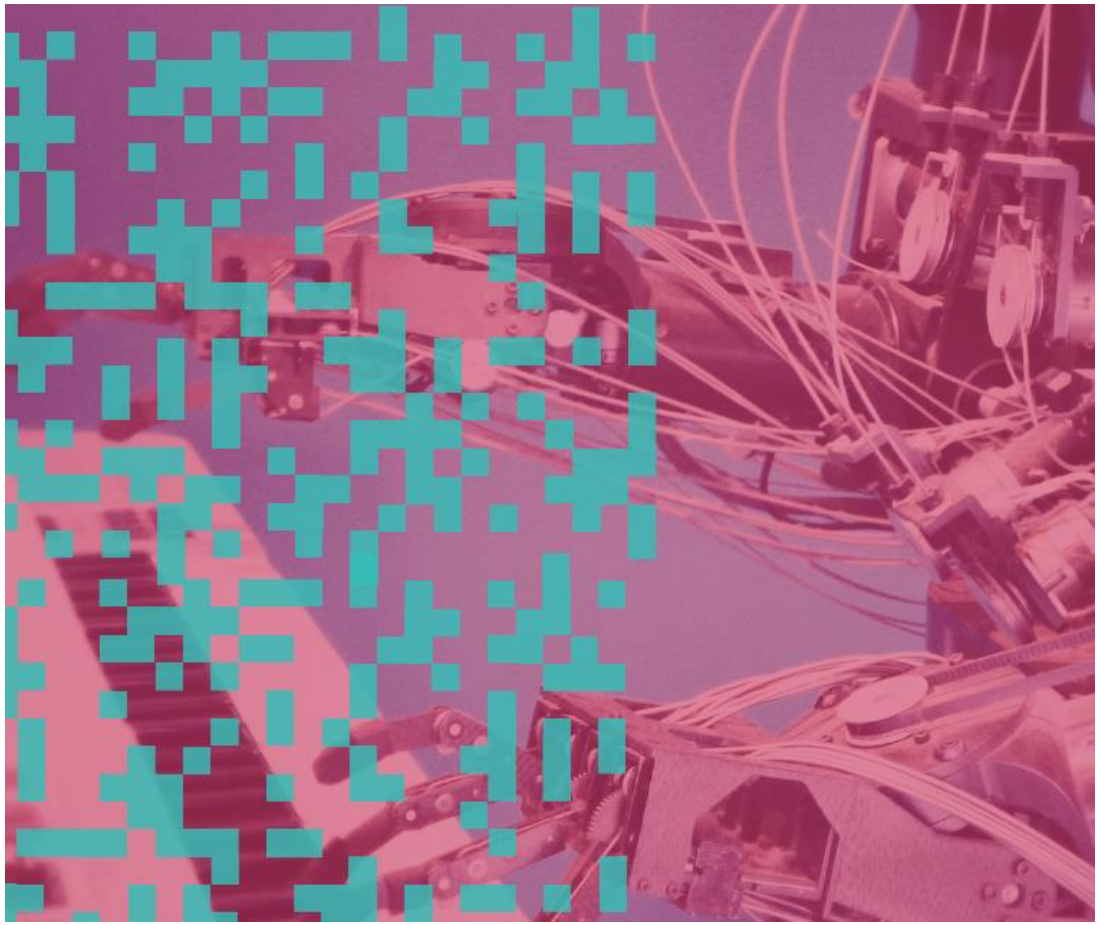
LAW AND ETHICS

- Are the following public or private?
  - Gov.uk news
  - Twitter posts by bereaved parents
  - A far-right subreddit
  - A private Facebook group for ERASMUS students
  - UoE staff intranet and Teams groups
  - Spanish in London WhatsApp group



LAW AND ETHICS

- **Anonymization**. If there is a reasonable chance the users or owners of a website see the content they are producing as private, it is important to prevent them from being identified.
    - Not publishing the names of websites
    - Removing the names of users/authors
    - Not reporting the texts of posts, since they might be searched for to identify

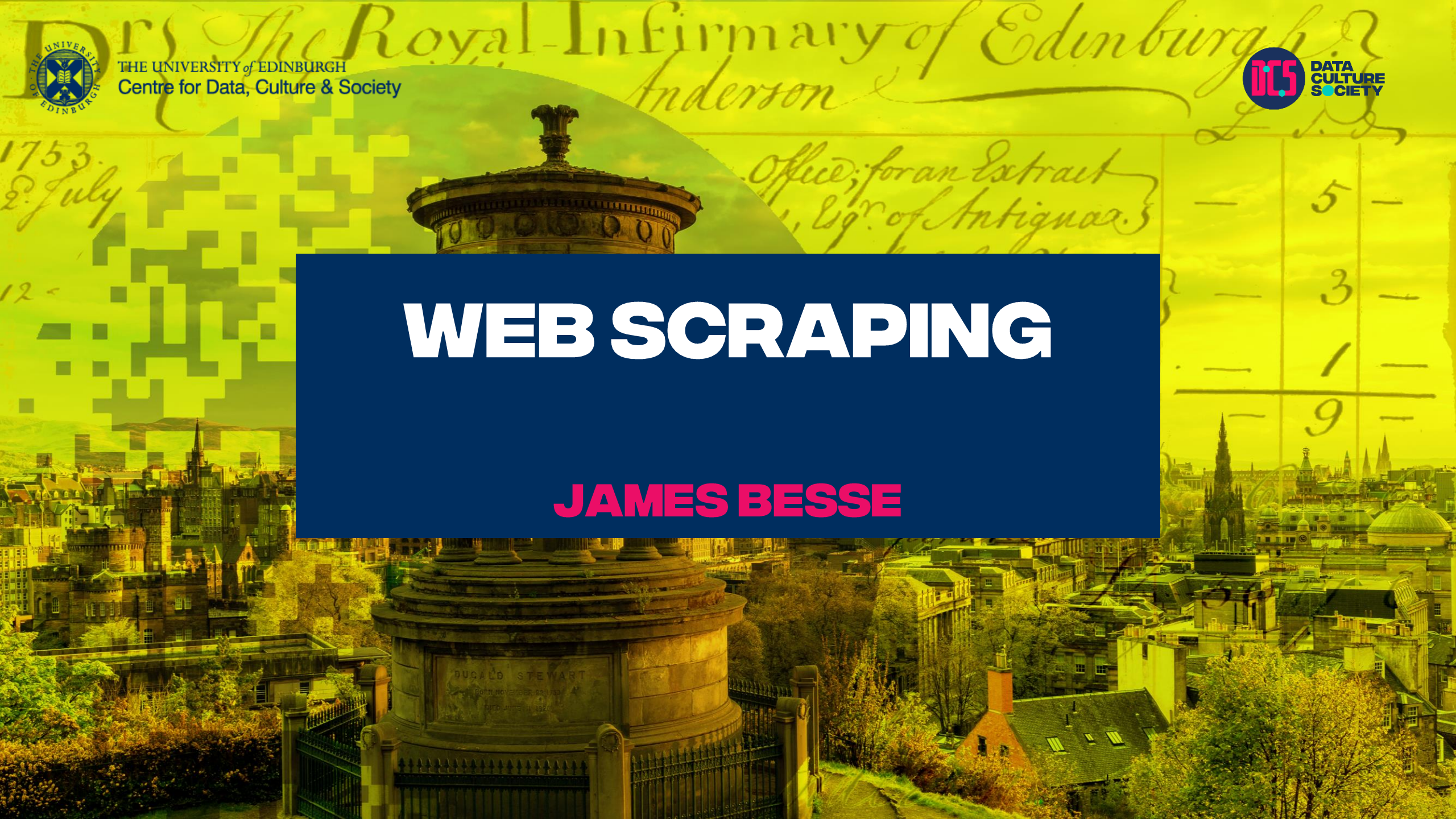


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

LUNCH BEAK

WE ARE GOING TO RESTART AT 13:30


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
WEB SCRAPING
JAMES BESSE

WHY SCRAPE THE WEB?

- Tons of interesting data about social interaction. Large datasets about granular interactions. Allows researchers to do 'big qualitative' data analysis.
- Content produced online has meaningful impacts on how people behave offline – ie. online activism, cybercrime, terrorism, government to citizen information, medical advice (websites and support groups), e-commerce, etc.

TERMINOLOGY

Web **crawling** refers to the automated process of systematically browsing the internet and indexing web pages for various purposes such as search engine indexing or data extraction.

Web **scraping**, on the other hand, involves **extracting specific data** from websites by programmatically accessing and parsing the HTML content of web pages, typically with the goal of retrieving structured information for analysis or reuse.

TERMINOLOGY

A **spider** is a program that systematically navigates the web by following links from one web page to another, often used in web crawling or web scraping tasks to discover and retrieve information from multiple pages.

A **scraper** is a script that is designed to extract specific data from web pages automatically. It typically utilizes various techniques to parse and extract desired information from the HTML structure of web pages.

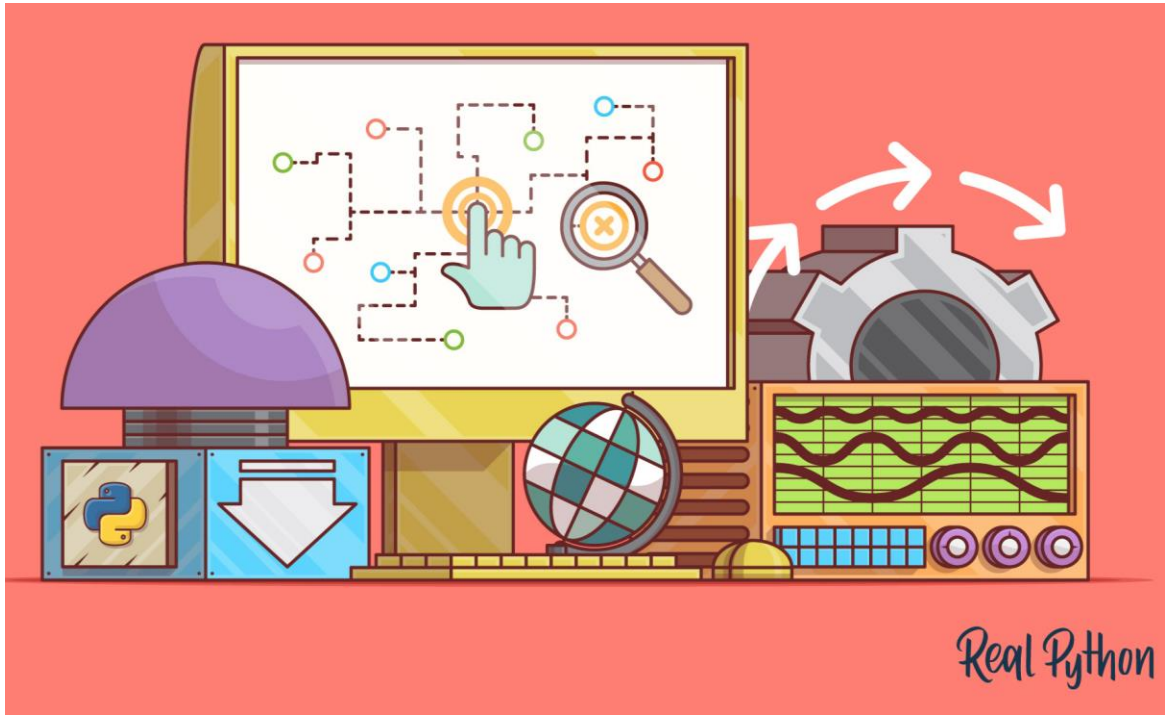
WEB SCRAPING OVERVIEW

Three main approaches:

- **Crawl** and scrape HTML/XML. Mostly used for the traditional internet (ie. institutional web pages, blogs, BBS). Generate lists of URLS and selectively download the HTML of web pages.
- Application programming interfaces (**APIs**). Request data from websites on their own terms. Many large websites, esp. platforms (ie. Google, Reddit, Twitter, Facebook, the Guardian, etc.) have these set up. This was (pre- Elon Musk) the mainstay of Twitter research, with most of the data used in these studies requested from the API.
- **Web browser automation** (ie. Selenium, Puppeteer). A technique for evading blocks on scrapers by automating a web browser to click through a website like a human user.

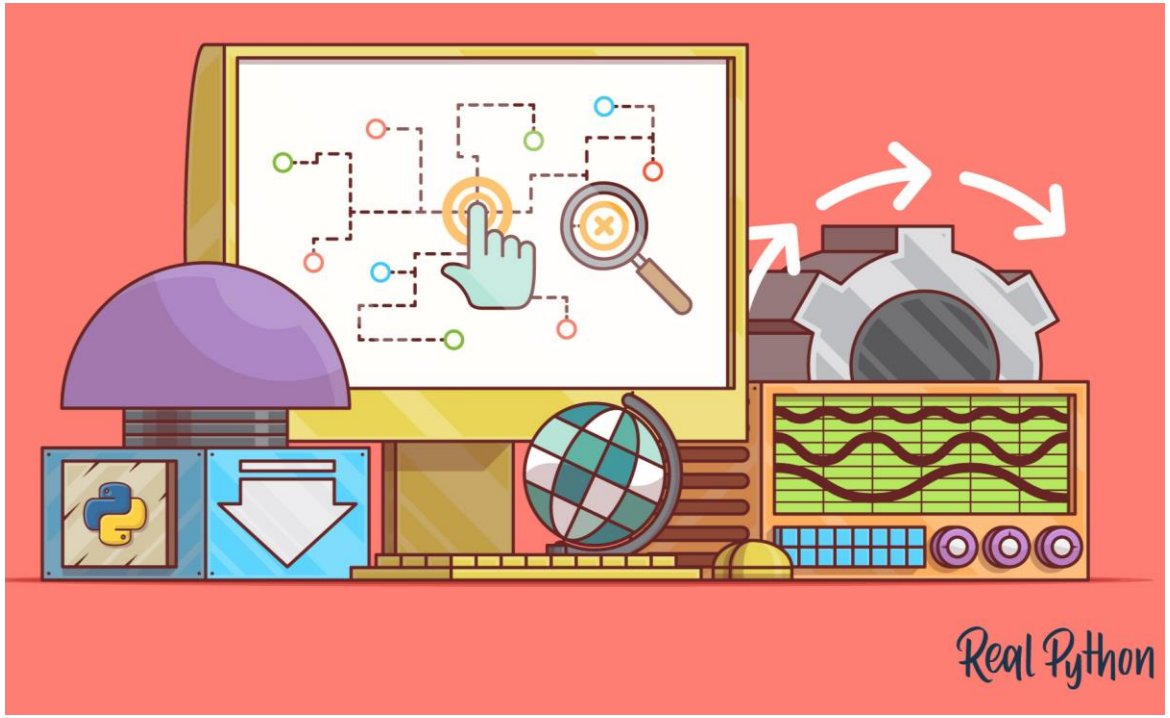
PRACTICALITIES

- Web scraping is to coding what hitchhiking is to business travel. It requires you to be much better at writing and editing code than statistical modelling requires.
- The data will be super messy, and you will probably need to do extensive cleaning after the fact.
- Websites may be internally and externally inconsistent in how they format HTML and text. What works once for a single website may only sometimes work, and may not work for other websites.



PRACTICALITIES

- Websites often change. If you scrape something once, chances are the website will be updated or modified the next time you try to scrape it.
- Some websites will try to block scrapers, so you may need to find ways of tricking them, or scrape extremely slowly to avoid rate limiting.
- People disagree about whether or not webscraping is ethical and legal. Make sure you're very careful about going through institutional review before you use web scraping for research.




THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY

TIME FOR R


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
COFFEE BEAK
WE ARE GOING TO RESTART AT 15:30


THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society
DATA CULTURE SOCIETY
BYOD
SESSION 1

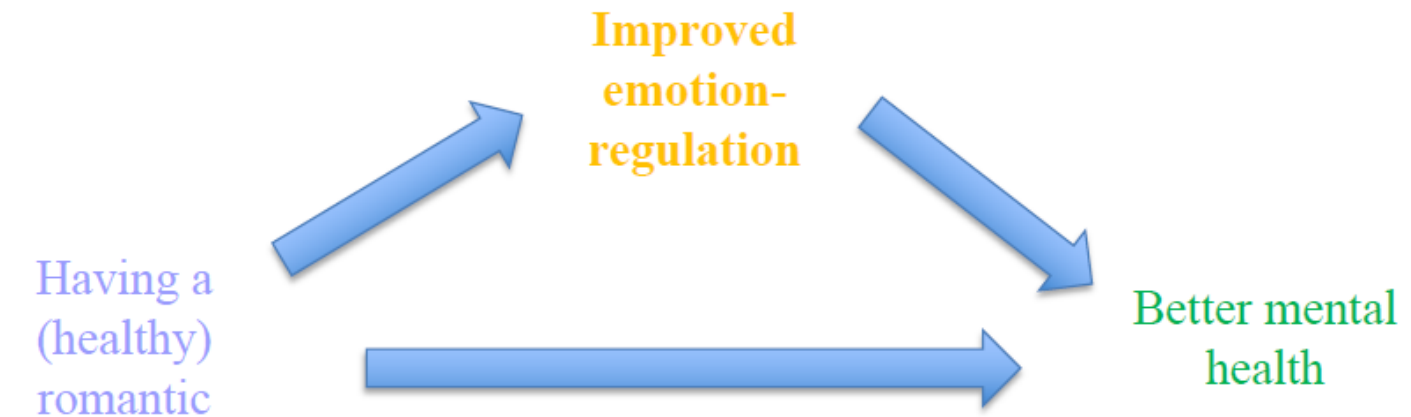
The University of Edinburgh

Love at first *lecture*?

The effect of romantic relationships on university students' mental health

Data: Online intake questionnaire and Ecological Momentary Assessment through app-based daily surveys (SEMA app)

Planned Analysis: Structural Equation Modeling, as outlined in the SEM map below



Clara Sanchez-Izquierdo Lozano
MSc Psychological Research Methods

What concerns do parents commonly report, and what aspects of their babies do they enjoy?

Who?

1,788 parents with a child under 27-months old

Text questions:

Do you have concerns about your baby's eating or sleeping behaviors?

Does anything about your baby worry you?

What do you enjoy about your baby?

Parent-reported measures:

Socio-emotional development of babies (continuous & categorical)

Parents’ depressive symptoms (continuous & categorical)

Demographics:

Parents’ age (continuous)

Parents’ gender (mother, father, would rather not say)

Babies’ age range (0-2, 3-8, 9-14, 15-20, 21-26 months)

Data Source:

Data for this study are from the posts from s r/askgaybros, r/gaybros, and r/gaybroscirclejerk subreddits in Reddit within the past year, including their titles, text and comments； excluded comments with less than 10 words, as insufficient information would make modeling and qualitative analysis impossible.

What is the data for:

Explore discourse from digital masculine gay communities to explore the underlying ideologies.

Data structure:

	A	B	C	D	E	F	G	H	I
1		title	text	date	comments	lean_tit	text_lis	ments_list	
2	0	Are my fa	I want to	2023-03-3	['No these are in		['i', 'wa	['no', 'these'	
3	1	Anal fiss	Hello!I h	2023-03-3	['Before you turn		['hello',	['before', 'yo	
4	2	[QUESTIO	Hello, I	2023-03-3	["As long as no or		['hello',	['as', 'long',	

Columns B, C, and E are the original title, text, and comments data; Columns F, G, and H are their respective data after being tidied.

The level of data cleaning:

I employed nltk library to clean the collected text data. I only tokenized the bodies and comments, as they did not require further processing for topic modelling. However, for the titles, I implemented additional steps such as stemming, removing stop words, and eliminating excessively frequent and rare words to lay the foundation for more accurate topic model results.

Research Background and Objective

The use of social media has been a recent breakthrough as it integrates subjective information with tangible formats, such as photos and images, therefore the results reflect public perception and preference (Liu et al., 2016; Tieskens et al., 2018). Despite these advances, investigating the correlation between perception of urban green space and subjective wellbeing is lacking, especially in Chinese cities with rapid urbanizing areas, as a large quantity of urban green space does not guarantee perceived beneficial effects of nature (Bell et al., 2014; Zhang et al., 2017). Utilizing visual social media with image recognition services provides opportunities to understand perceived nature and its influence on subjective wellbeing.

It is contended that subjective wellbeing is positively associated with perceived nature in urbanized areas. To test this hypothesis, we investigate how perception of urban green spaces relates to subjective wellbeing and life satisfaction in urbanized areas through harvesting user-generated photos from social media platform and analysing content of these photos with computer vision service.

Methods

Data collection and content analysis

Metadata of geo-tagged photographs taken from 2011 to 2015 are retrieved from Flickr within 5th ring road in Beijing. Geo-tagged photos are collected and labelled by Google Could Vision service. Up to 10 tags are generated to describe content of a photo (see data_photos). Perception of urban green space index (PGSI) can be calculated based on ratio of green dominated photos in each sampling unit.

Fixed effect linear regression analysis

Relying on Beijing Area Study (BAS) conducted in 2013 and 2015, wellbeing status (WS) of respondents are investigated by asking overall health status and mental health status and results are registered into five-level Likert scaling. Study site is divided based on subdistrict administrative units (Jiedao), which are linked to each participant in BAS (see data_survey). A fixed effect linear regression analysis can be conducted, where WS is entered as dependent variable, PGSI is entered as independent variable and sociodemographic characteristics (e.g., gender, age and education) are entered as dummy variables.

THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

DATA CULTURE SOCIETY

PUB QUIZ

SUMMER HALL

1, Summerhall, Newington,

Edinburgh EH9 1PL

www.cdcs.ed.ac.uk