DIGITAL AND COMPUTATIONAL METHODS FOR HERITAGE

Dr Lucia Michielin

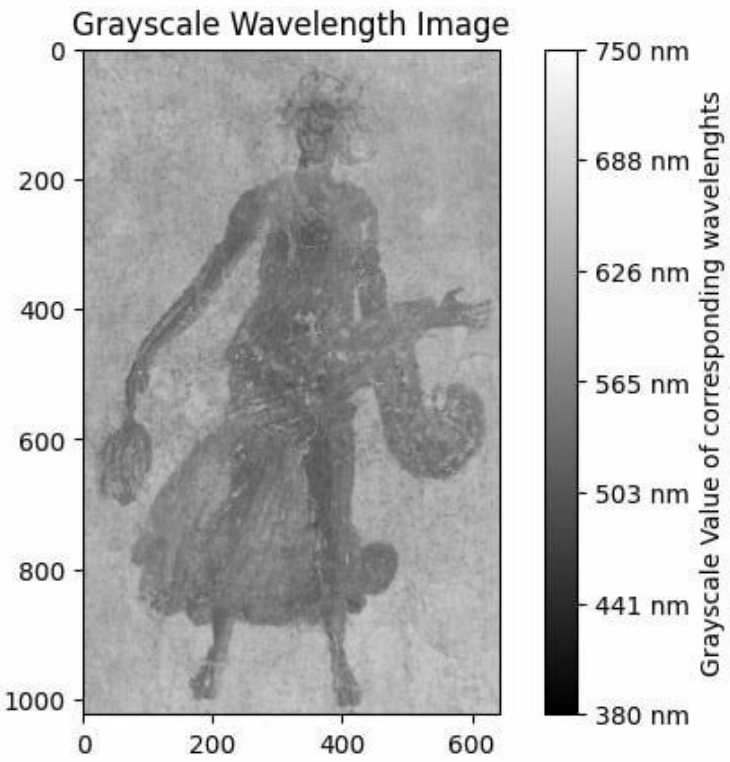
THE UNIVERSITY of EDINBURGH
Centre for Data, Culture & Society

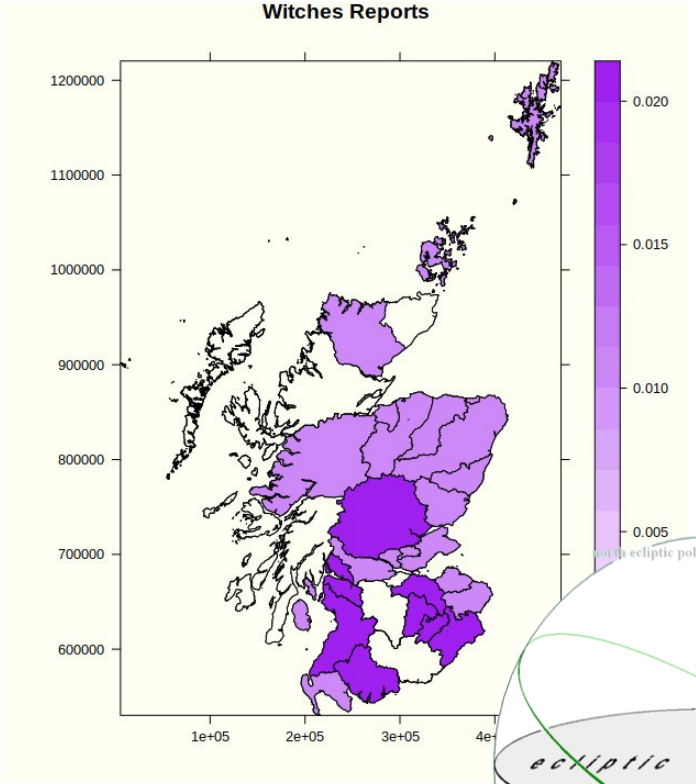
THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute

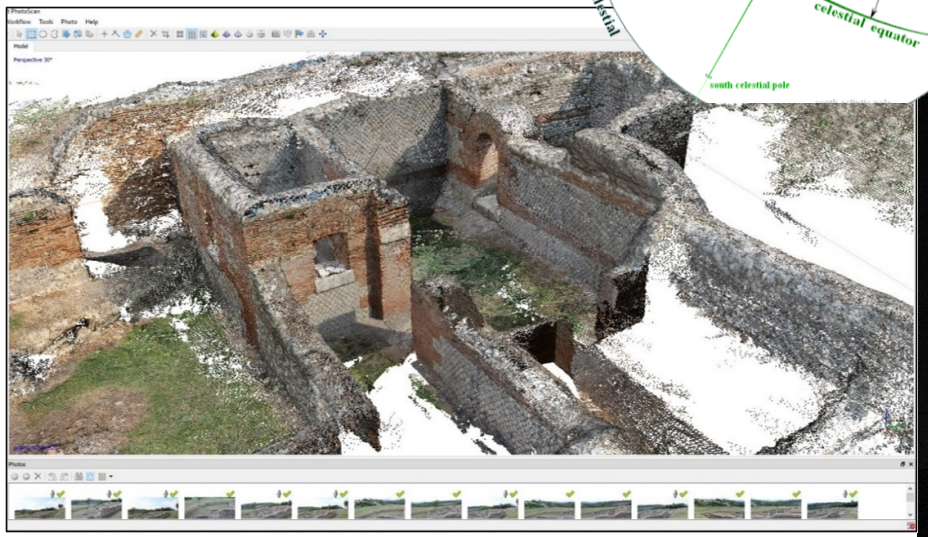
DATA CULTURE SOCIETY

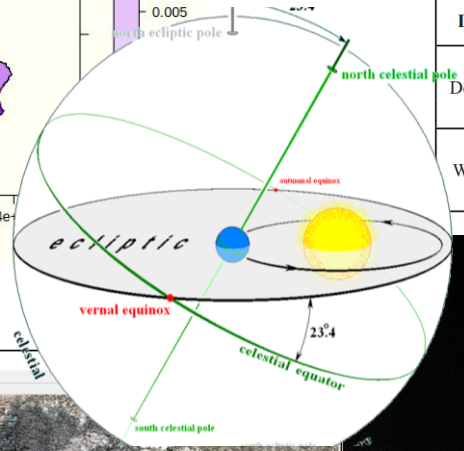
MY BACKGROUND

- PhD in Classics and Computational Archaeology
- Digital Skills Training Manager









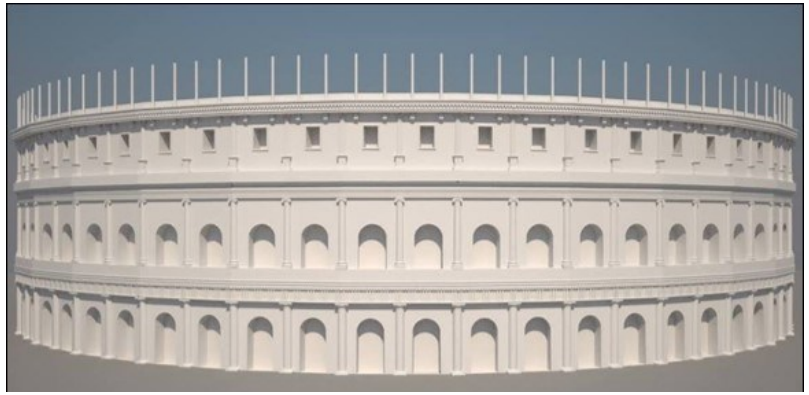
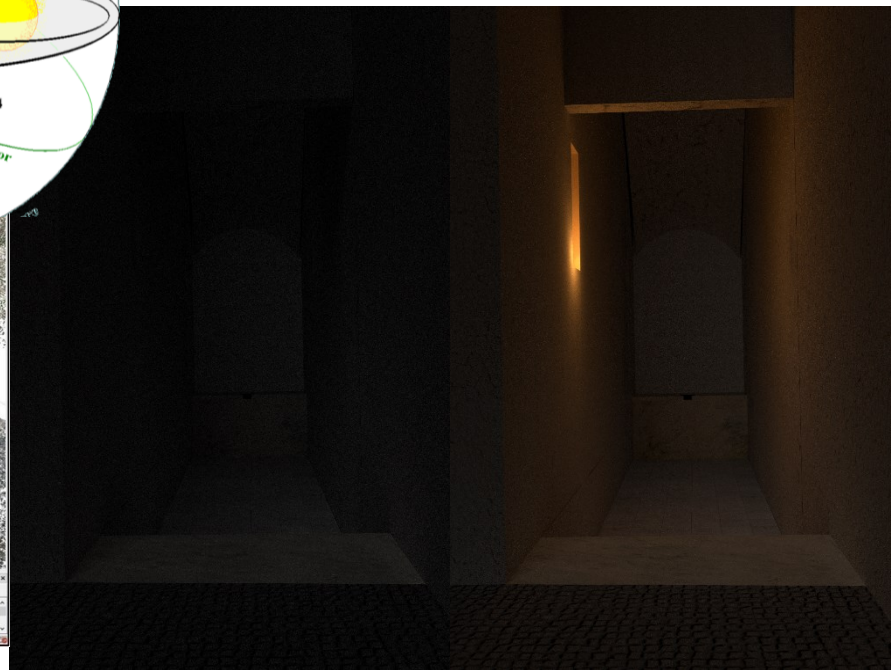


Table 6.6 – The width variation. The table collects the p-values resulting from the Null-Hypothesis tests performed between the opening facing outward/public areas and inward/private areas. To be considered statistically significant, the p-values need to be below the 0.05 mark. The table also represents which of the two subsets have the larger mean.

Dataset	Test	Variable	P-values after correction (outward vs inward-facing)	Larger mean
Doorways	T-test	Width	8.289e-06	Outward Facing
		Log-Width	2.943e-06	Outward Facing
	K.S. test	Width	5.1843-06	Outward Facing
Windows	T-test	Width	1	Outward Facing
		Log-Width	0.032	Outward Facing
	K.S. test	Width	0.014	Outward Facing



DATA CULTURE SOCIETY

THE UNIVERSITY of EDINBURGH
Edinburgh Futures Institute

SUPPORT FOR DATA-LED AND APPLIED DIGITAL RESEARCH ACROSS THE ARTS, HUMANITIES AND SOCIAL SCIENCES.

WHAT WE ARE GOING TO LOOK AT TODAY

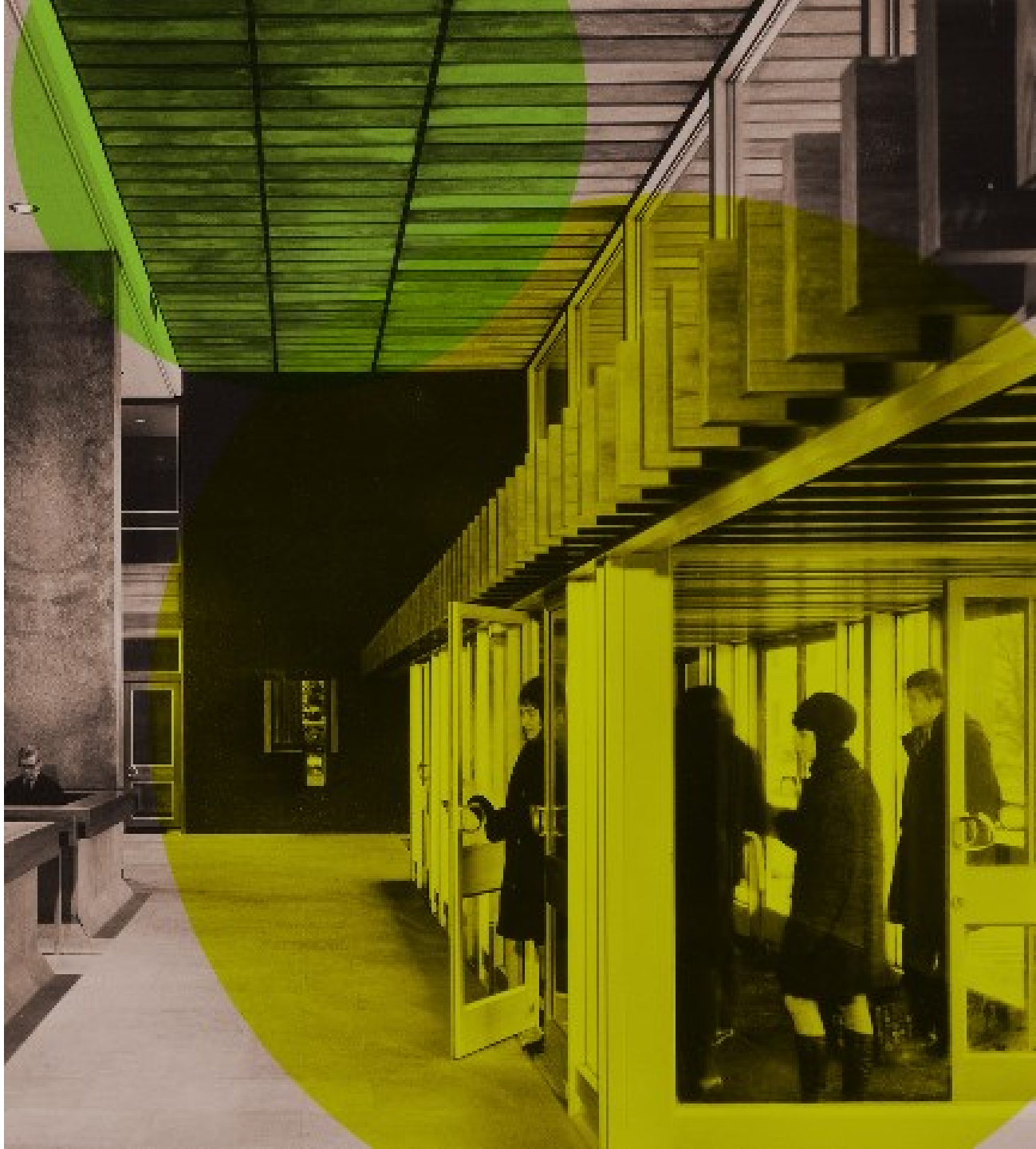
- Introduction on challenges and limitations
- Good practices of dealing with data
- Overview of main techniques and where to start with those
- Play around with the Statistical Accounts of Scotland

BREAKING THE "LANGUAGE" BARRIER

- Why is embedding digital method "Hard" for people in Heritage?
- "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

CHALLENGES

- Perceived as something separate from “standard” research
- Lack of a shared language / language barriers
- Many disciplines with no tradition in teaching digital skills
- Some techniques would have a fairly step learning curve



BUT

- Lot of similarities with mental processes you are familiar with (coding <-> working with dead languages)
- Can free you form repetitive tasks
- Make your research accessible and reproducible
- See data from a different perspective

DATA GOOD PRACTICES

- What kind of data do you work with?
- How do you store them?
- How do you edit them?
- Have you written a data management plan?

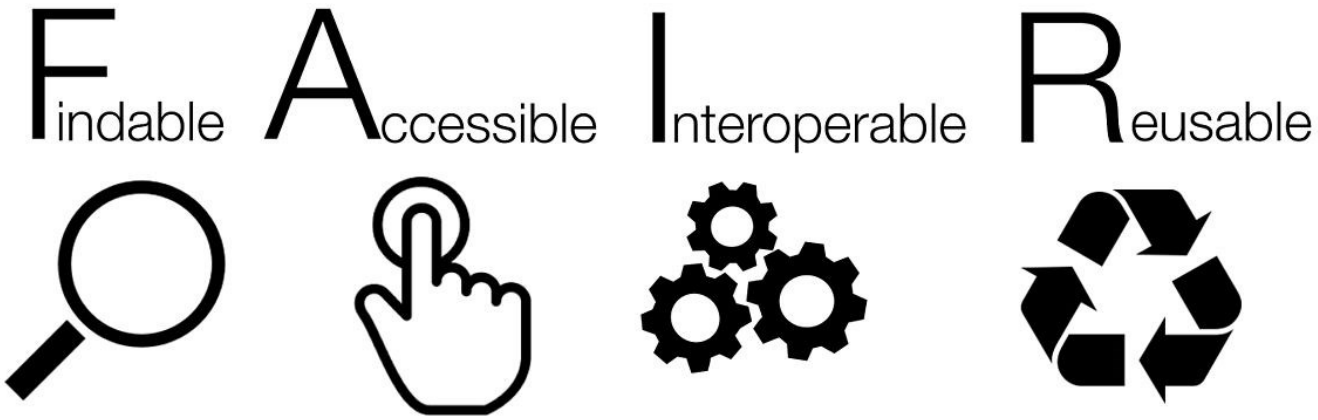
HOW TO ORGANISE YOUR DATA

- **Put all your variables in columns**
- **Put each observation in its own row**
- **Don’t combine multiple pieces of information in one cell** -
- **Leave the raw data raw - don’t change it!** Always work on a copy of your data
- **Export the cleaned data to a text-based format like CSV (comma-separated values)** - This ensures that anyone can use the data and is the format required by most data repositories

RESOURCES

- https://datacarpentry.org/spreadsheets-socialsci/.
- https://datacarpentry.org/spreadsheet-ecology-lesson/
- https://librarycarpentry.org/lc-spreadsheets/

FAIR DATA PRINCIPLES



RESOURCES

- https://www.ed.ac.uk/information-services/research-support/research-data-service/after/datavault/prepare-datavault/make-your-data-fair
- https://www.ed.ac.uk/sites/default/files/atoms/files/quick_guide_6_-_making_your_research_data_fair_v1.1_0.pdf

USEFUL TOOLS

- **Openrefine** to clean and wrangle your data code free
- **Version Control** and **GitHub Desktop** to avoid multiple versions of your files and share your analyses
- **Latex and Overleaf** to produce outputs and article draft (option to work collaboratively too)
- **Noteable** UoE service to access coding environments (Python and R) or Posit and Google Colab
- **Regex** method of using a sequence of characters to define a search to match strings. https://programminghistorian.org/en/lessons/understanding-regular-expressions

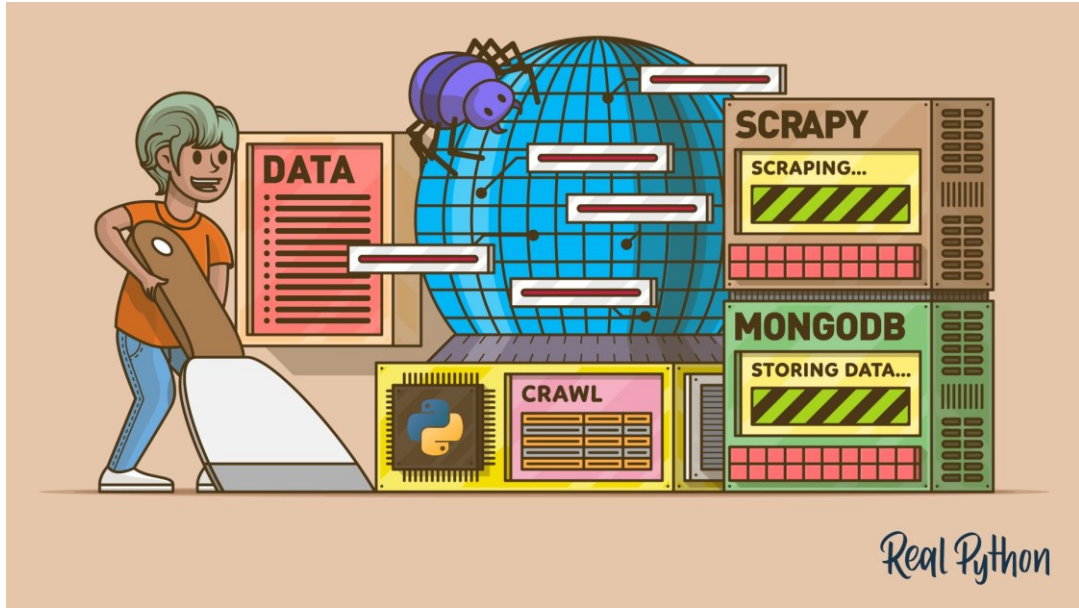
DIGITAL RESEARCH METHODS OVERVIEW

- Working with Textual Data
- Network Analysis
- Data Visualisation
- Data Analysis and Statistics
- Geographical Information Systems
- Working with 3D Data
- Image Processing
- Digital Surveys and Crowdsourcing

WORKING WITH TEXTUAL DATA: GETTING THE DATA

- **OCR** (optical character recognition) **and HTR** (hand text recognition)
  - Creates **machine-readable** documents that can be searched, edited, and analysed computationally both code and code free options
  - https://github.com/DCS-training/Image-to-Tech-Text-Extraction- (OCR)
  - https://github.com/DCS-training/Transkribus (HTR)

- **Webscraping** (web crawling and dynamic web pages scraping)
  - Two main techniques: web crawling (for static website e.g. forum or news sites) and “social media scraping” (API used to be the most common for scraping dynamic pages)
  - Intro to Beautiful Soup Python
  - Web scraping with R
  - Web Data Research Assistant (code free)



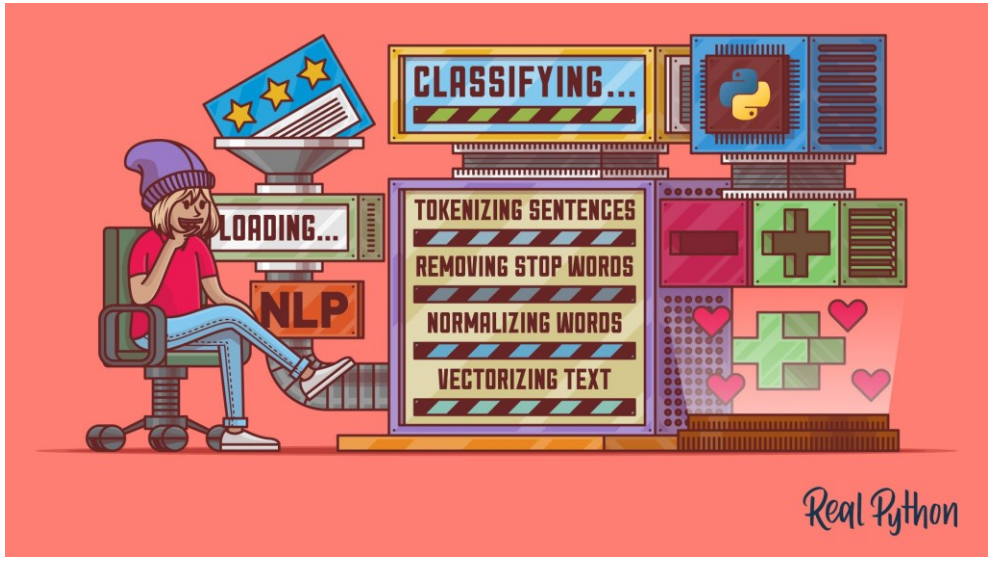
WORKING WITH TEXTUAL DATA: ANALYSE THE DATA

Text Analysis (Distant Reading) Computationally evaluating, investigating, and exploring textual (natural language) data. The first and most important step is the pre-processing (tokenisation, stopwords removal, text cleaning)

- *Sentiment Analysis*
- *Topic Modelling*
- *Named entity recognition* (NER)
- *Word Embeddings*

RESOURCES

- Gale Digital Scholar Lab. Explore UoE primary sources + analyse code-free textual data
- Corpus Analysis with Antconc (code free)
- https://lancsbox.lancs.ac.uk/
- Digitised Documents and Text Analysis Repositories on CDCS GitHub on different aspect of text analysis (code based)
- Text Analysis Pathway
- Sentiment Analysis Pathway



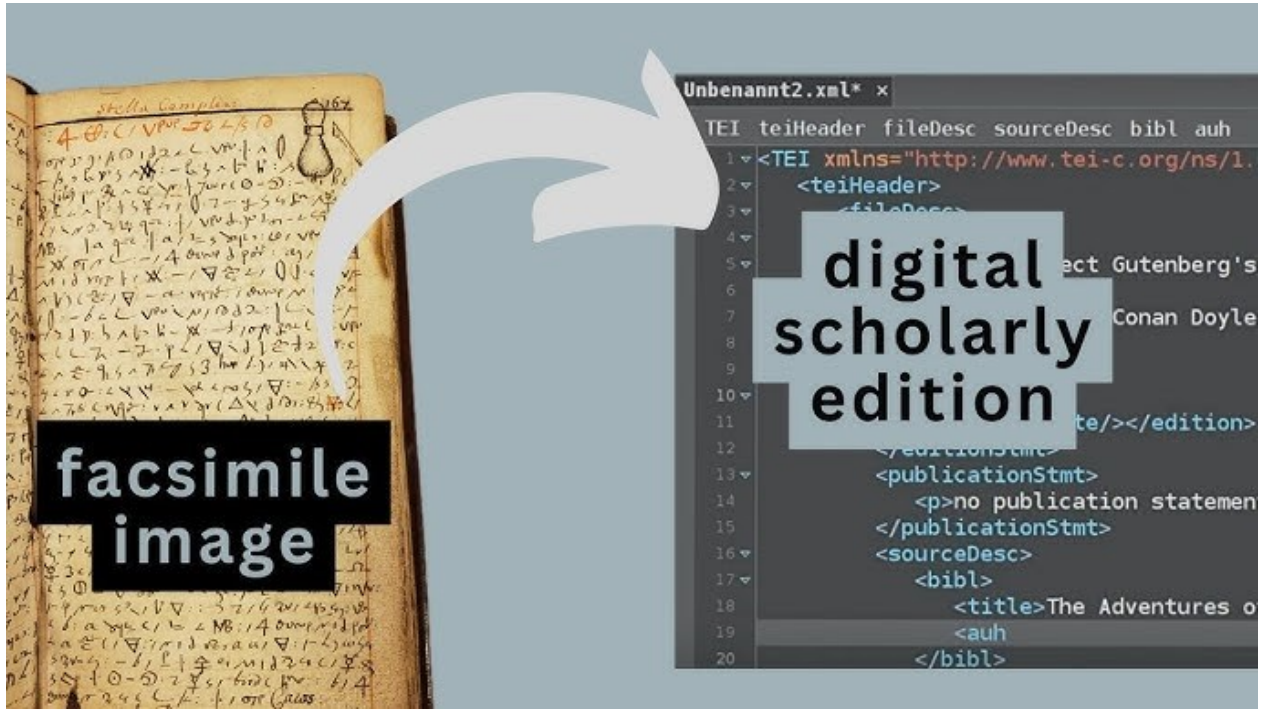
WORKING WITH TEXTUAL DATA: ANALYSE THE DATA

Text Encoding and **TEI Initiative**:

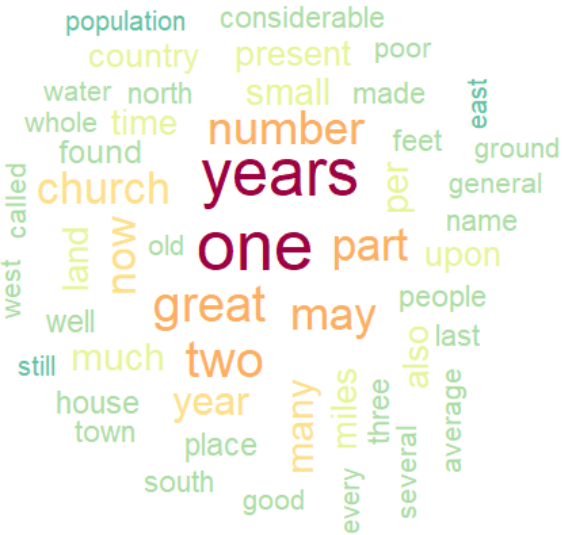
- Text encoding is a process whereby documents are transferred to an electronically searchable format for scholarly research
- The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form

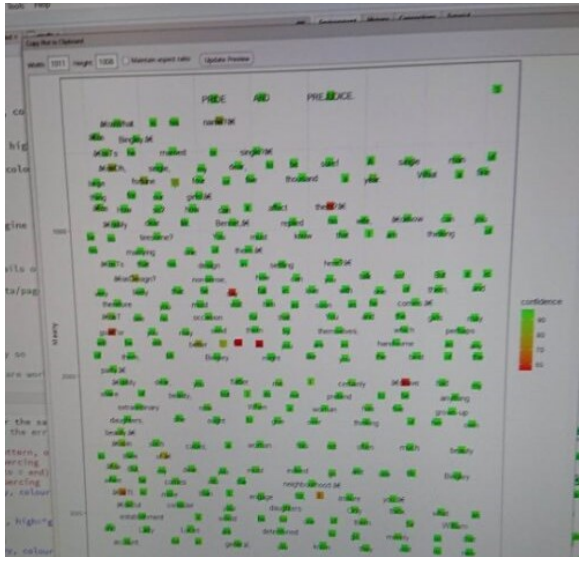
RESOURCES

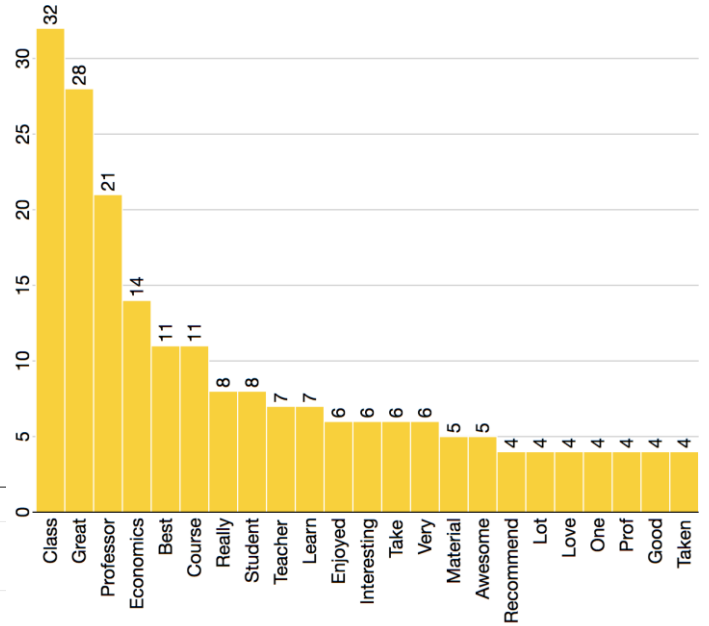
- TEI by Example

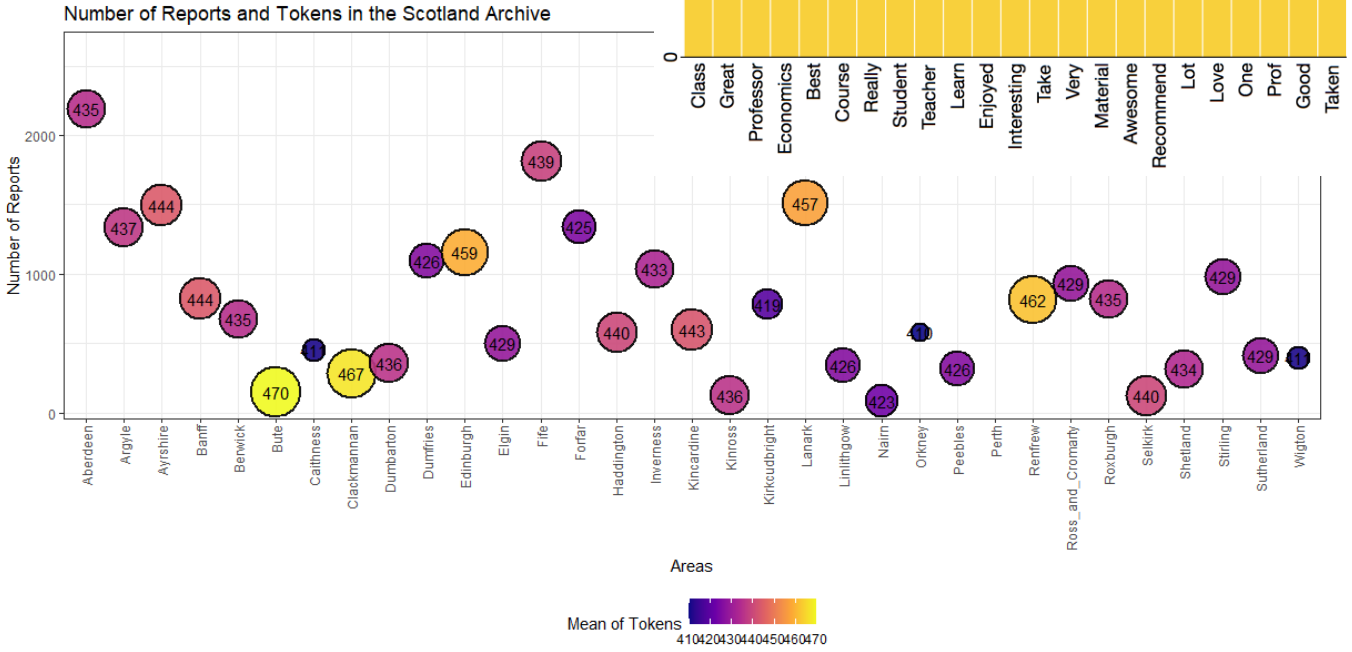


WORKING WITH TEXTUAL DATA









NETWORK ANALYSIS

- **Network analysis** can help to transform empirical complexity into a research object, allowing exploration and communication of complex relational phenomena.
- Networks are defined by **nodes (entities)** and **edges (relations)**. We often associate nodes with humans that are connected by all sorts of relations – the edges.
- Relations between characters in a plot can be traced by their **interaction patterns**, textual documents within a corpus by the **co-presence of keywords**, actors staring together in a film, etc
- Both code and non code approaches
- Hardest step is to **build the dataset** that you want to use because you need to define both notes and edges

RESOURCES

Gephi:

- https://jasonmkelly.com/jason-m-kelly/2021/1/21/gephi-and-historical-network-analysis-module
- https://www.youtube.com/watch?v=GXtbL8avpik
- https://www.youtube.com/watch?v=lPivwXdy9XY
- https://www.youtube.com/playlist?list=PLk_jmmkw5S2BqnYBqF2VNPcszY93-ze49

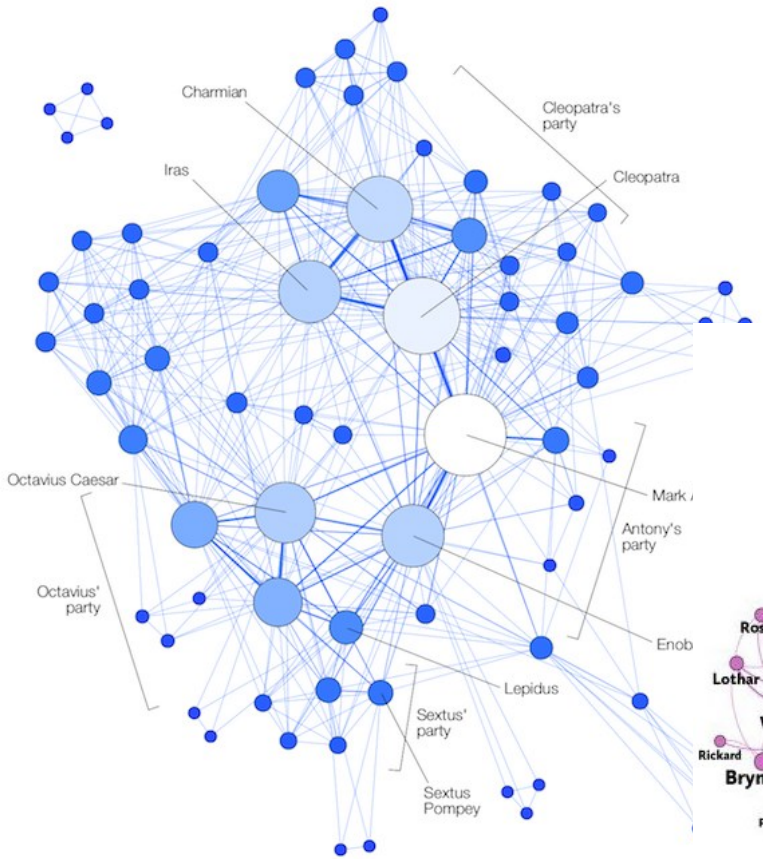
Python

- https://programminghistorian.org/en/lessons/exploring-and-analyzing-network-data-with-python
- https://github.com/DCS-training/Network-Analyisis-Python

R

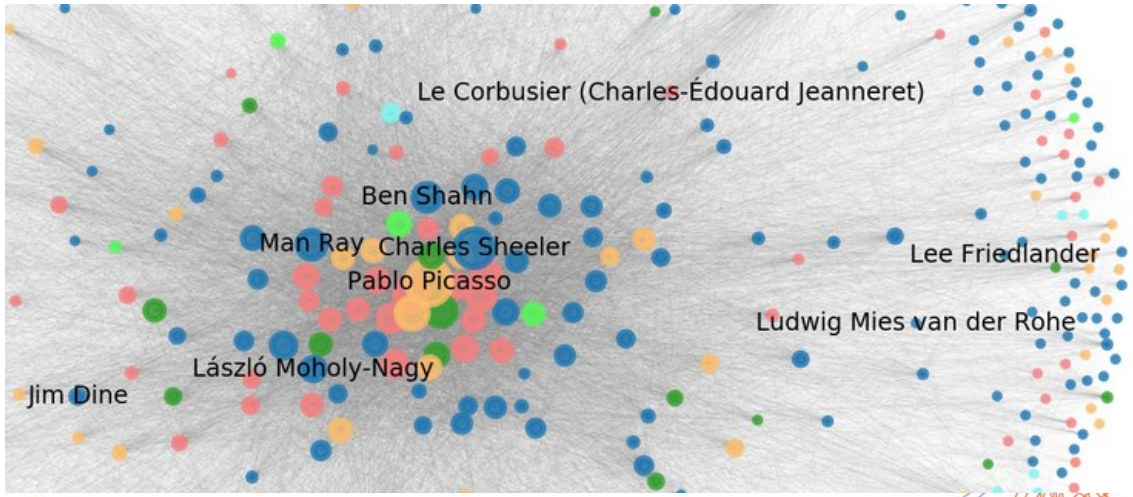
- https://programminghistorian.org/en/lessons/temporal-network-analysis-with-r

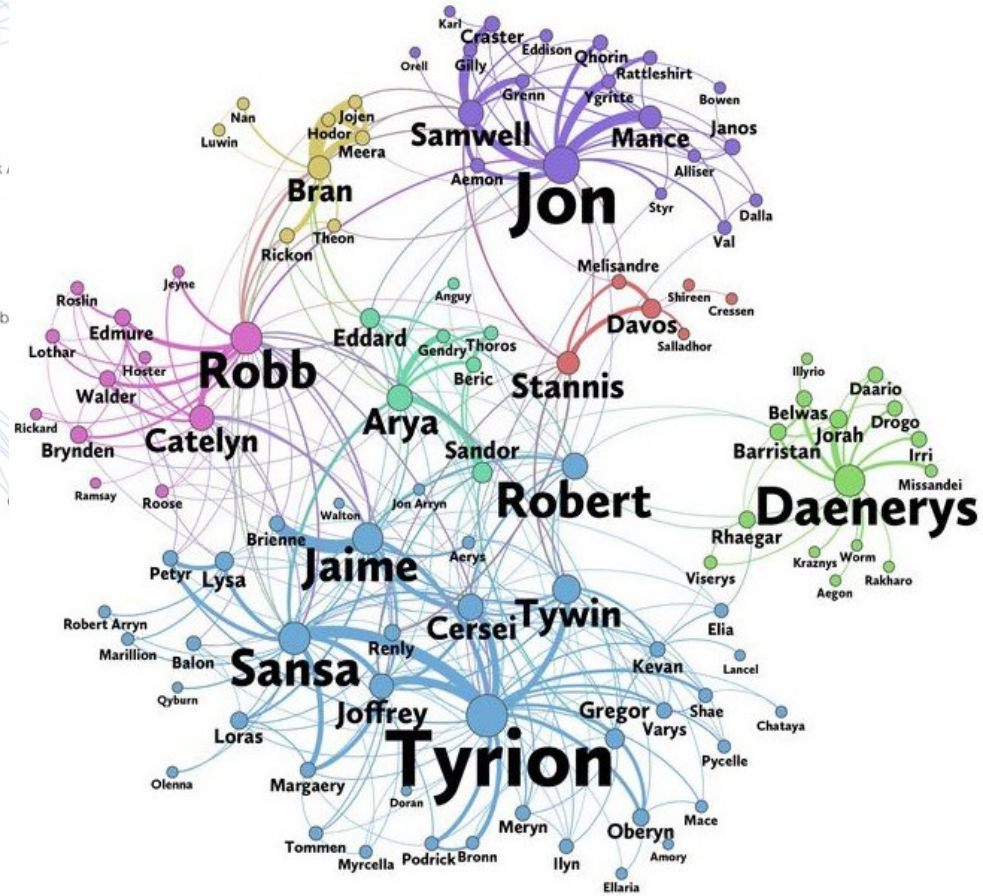
NETWORK ANALYSIS

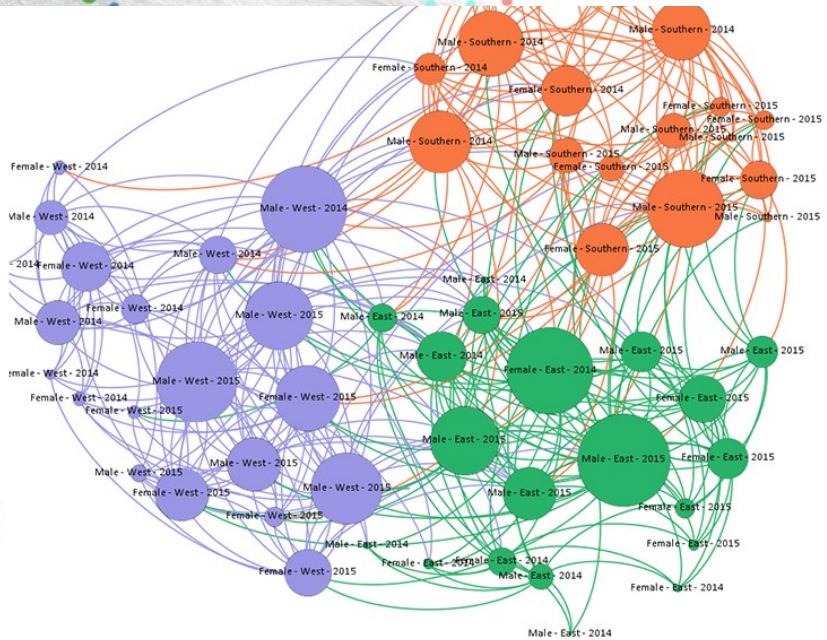


ANTONY AND CLEOPATRA

Number of characters 74 | 17% Network density







DATA VISUALISATION

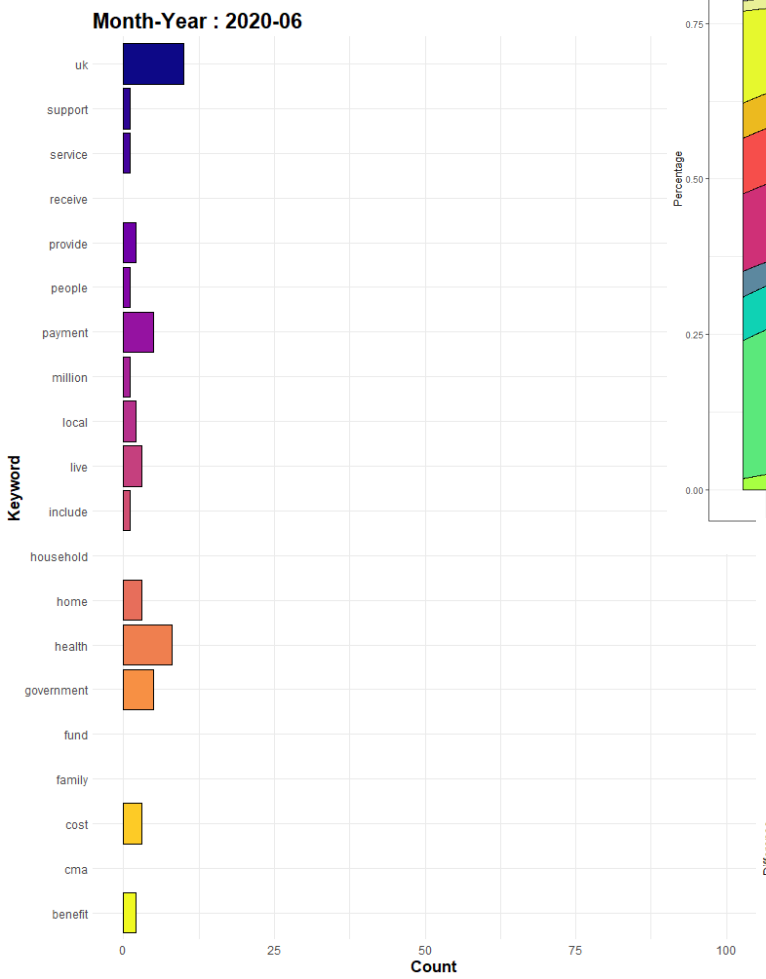
- **Be accurate and clear** e.g. make sure the unit of measurements are clearly expressed
- **Let the data speak**: show as much information as possible without hiding any aspects of the data
- **Avoid unnecessary frills** and showing off. Less is usually more in visualisations, making them clearer and better
- **Avoid pie charts**: they can be very misleading
- **Consider the medium** in which the visualisation will be presented: will the graph be printed or online? how big will it be?

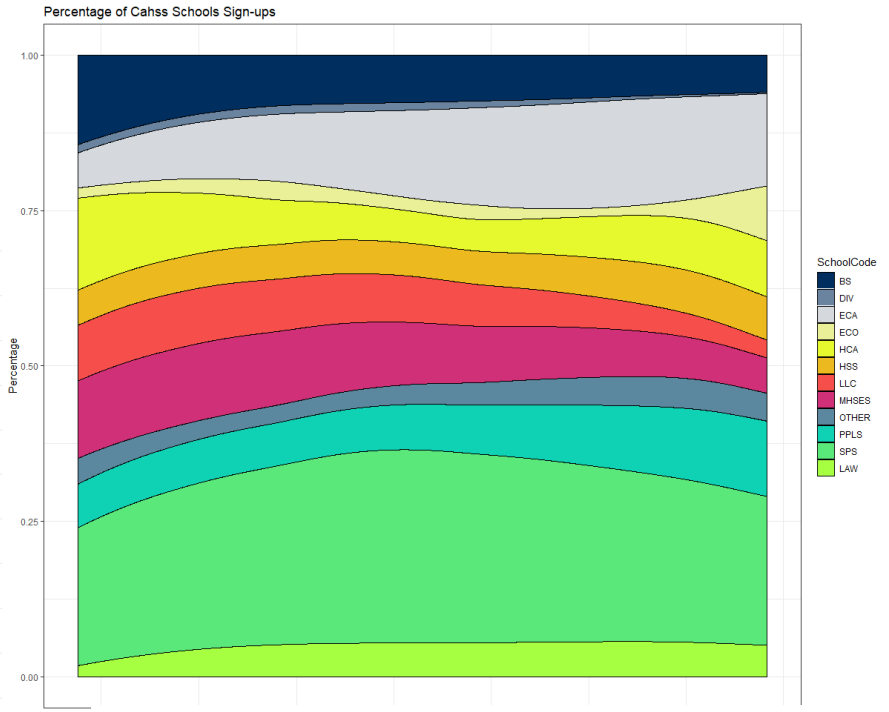
RESOURCES

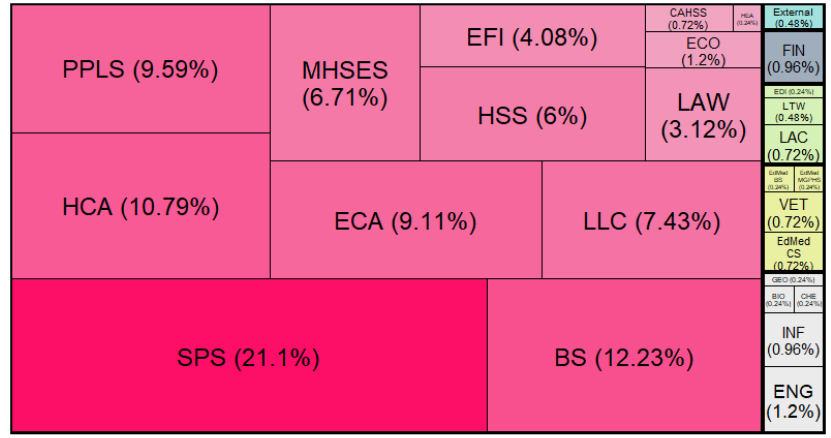
- Good place to start: Schwabish, Jonathan. Better Data Visualizations : A Guide for Scholars, Researchers, and Wonks / Jonathan Schwabish. New York, NY: Columbia University Press, 2021 (Full-Text available through the library)
- Overview of main data vis tools
- Repository on data vis

NB Excel is not a good tool to do data visualisation!

DATA VISUALISATION

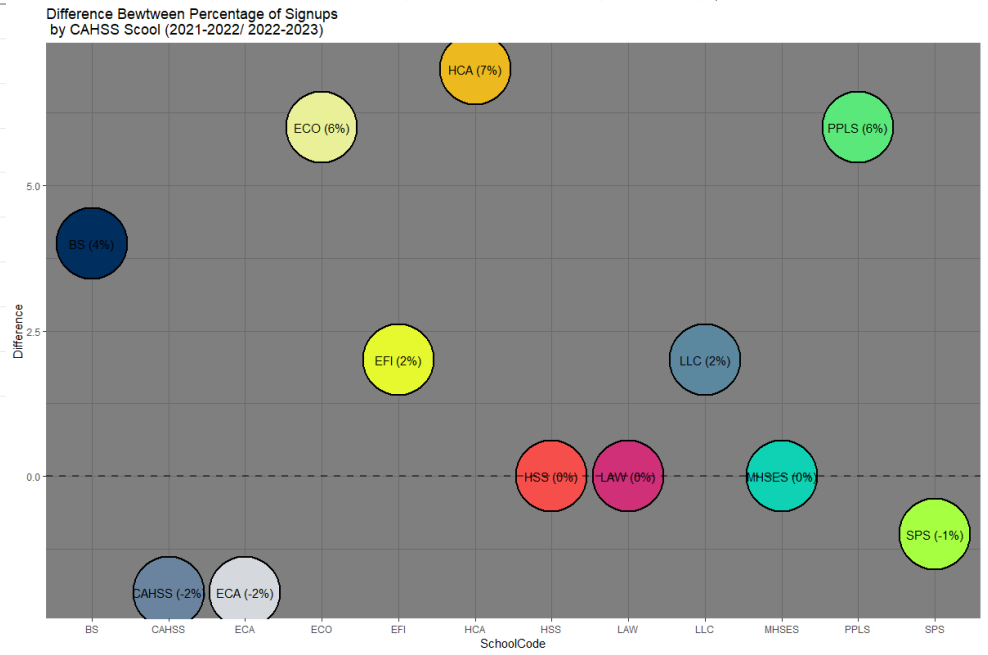


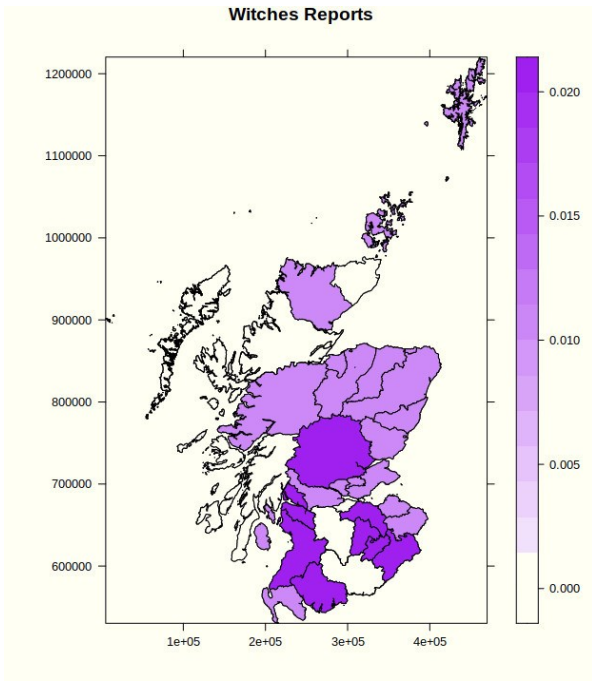




Colleges and other Clusters CAHSS CSG ISG CMVM External SCE

(N = 417)





DATA ANALYSIS AND STATISTICS

Too many techniques/methods to list but main concept to familiarise with

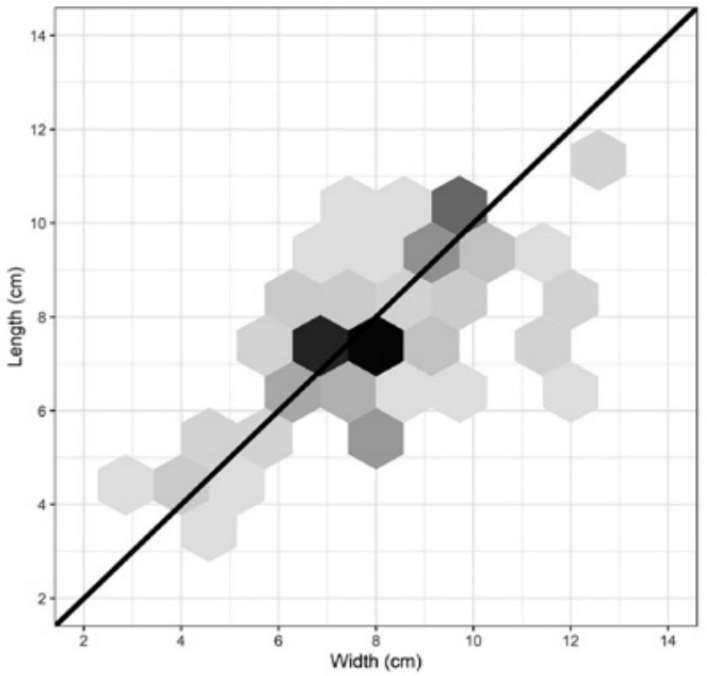
- Descriptive Statistics
- Inferential Statistics
- Measures of Central Tendency and dispersion
- Normal distribution and other distributions
- Null-Hypothesis Significance tests
- P-Values
- Population vs Sample
- Probability
- PCA
- Clustering Analysis
- Linear Regression

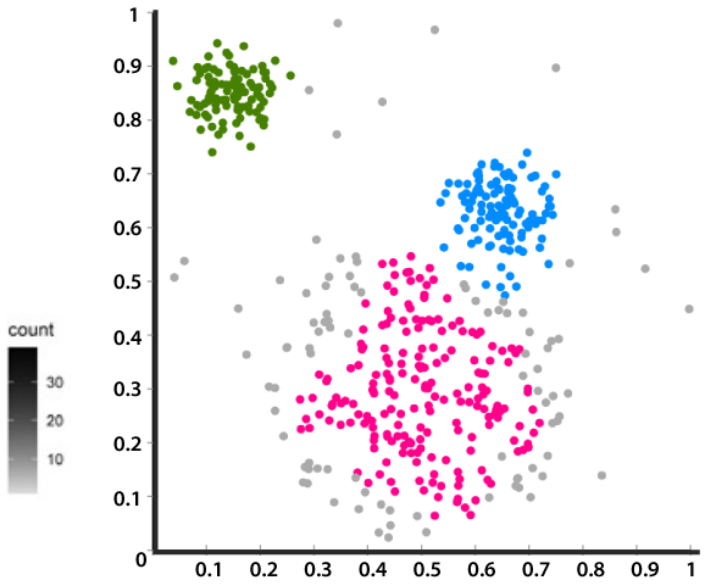
- **Pro of doing it via code is that makes your analysis fully reproducible**
- **NB Excel is not a good tool to do statistical analysis!**

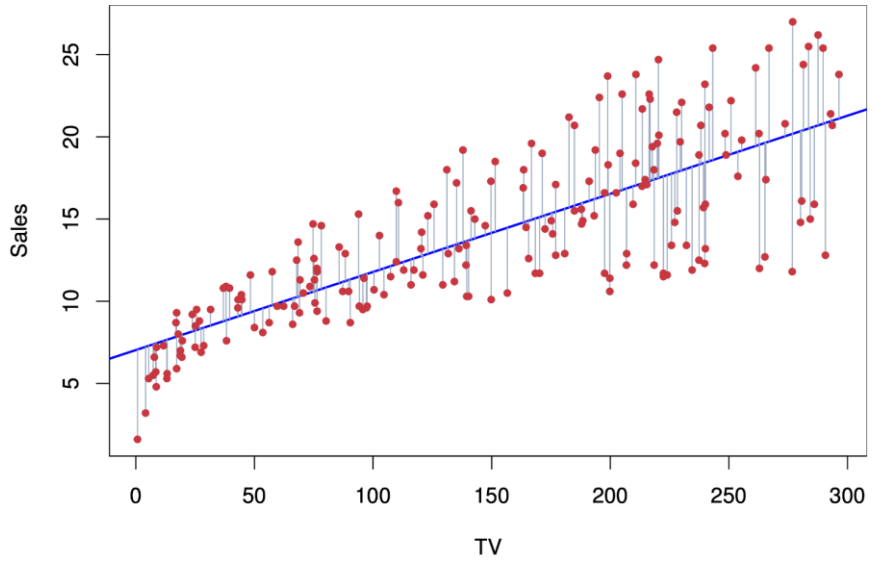
RESOURCES

- StatQuest Very clear collection of videos that explain statistical concepts in an easy way https://statquest.org/
- SPSS (Easy interface)
- Stata (Easy interface)
- R and R Studio (Language developed for stat)
- Python and SciPy
- Simulation in statistics
- Descriptive vs inferential stats
- Machine learning
- Linear Modelling

DATA ANALYSIS AND STATISTICS







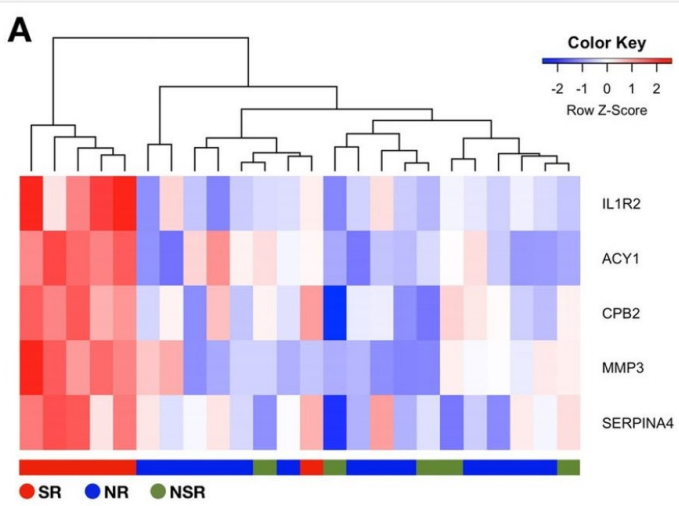


Table 6.6 – The width variation. The table collects the p-values resulting from the Null-Hypothesis tests performed between the opening facing outward/public areas and inward/private areas. To be considered statistically significant, the p-values need to be below the 0.05 mark. The table also represents which of the two subsets have the larger mean.

Dataset	Test	Variable	P-values after correction (outward vs inward-facing)	Larger mean
Doorways	T-test	Width	8.289e-06	Outward Facing
		Log-Width	2.943e-06	Outward Facing
	K.S. test	Width	5.1843-06	Outward Facing
Windows	T-test	Width	1	Outward Facing
		Log-Width	0.032	Outward Facing
	K.S. test	Width	0.014	Outward Facing

GEOGRAPHICAL INFORMATION SYSTEMS

- GIS can also help a researcher cross disciplinary boundaries when working on an interdisciplinary project or question. The use of GIS platform can be synthesised into 4 steps:
  - *Create* geographic data.
  - *Manage* it.
  - *Analyse* it and gather a better knowledge of the area as a whole system
  - *Display* it on a map.
- Work with both raster and vectors
- Important to familiarise with reference systems

(geographical vs cartographic SR)

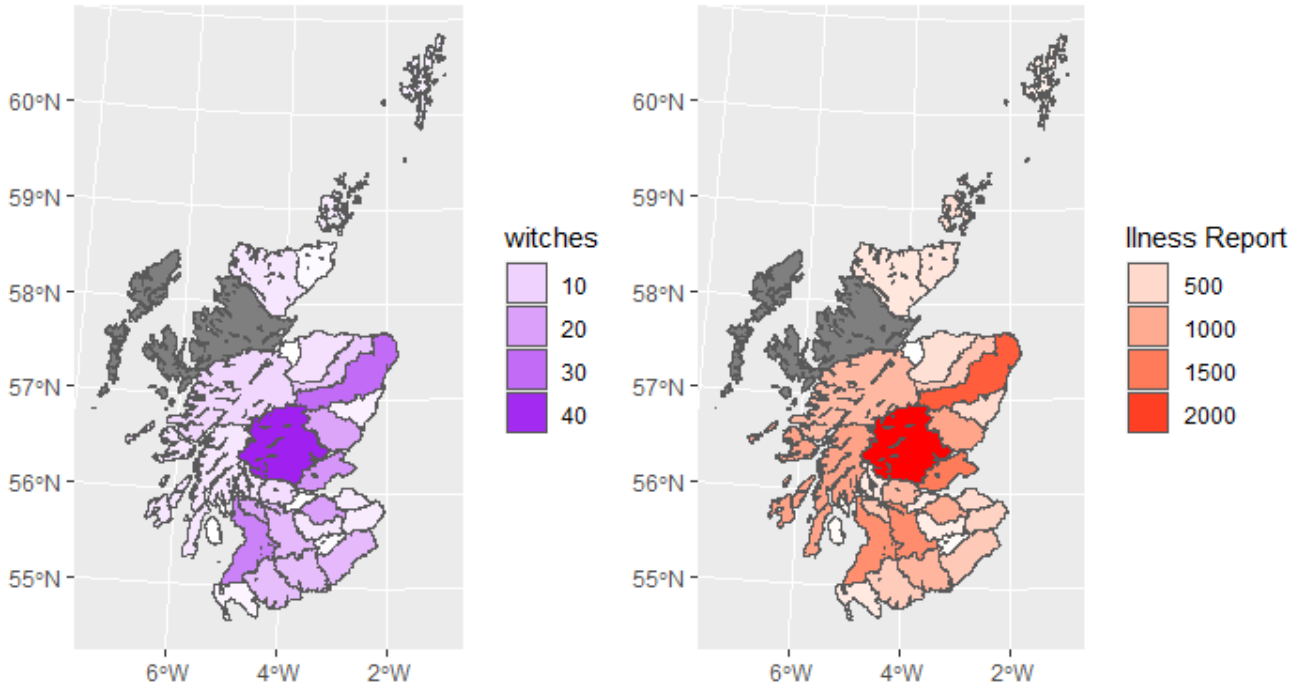
- Main software are QGIS and ArcGIS

https://gisgeography.com/qgis-arcgis-differences/

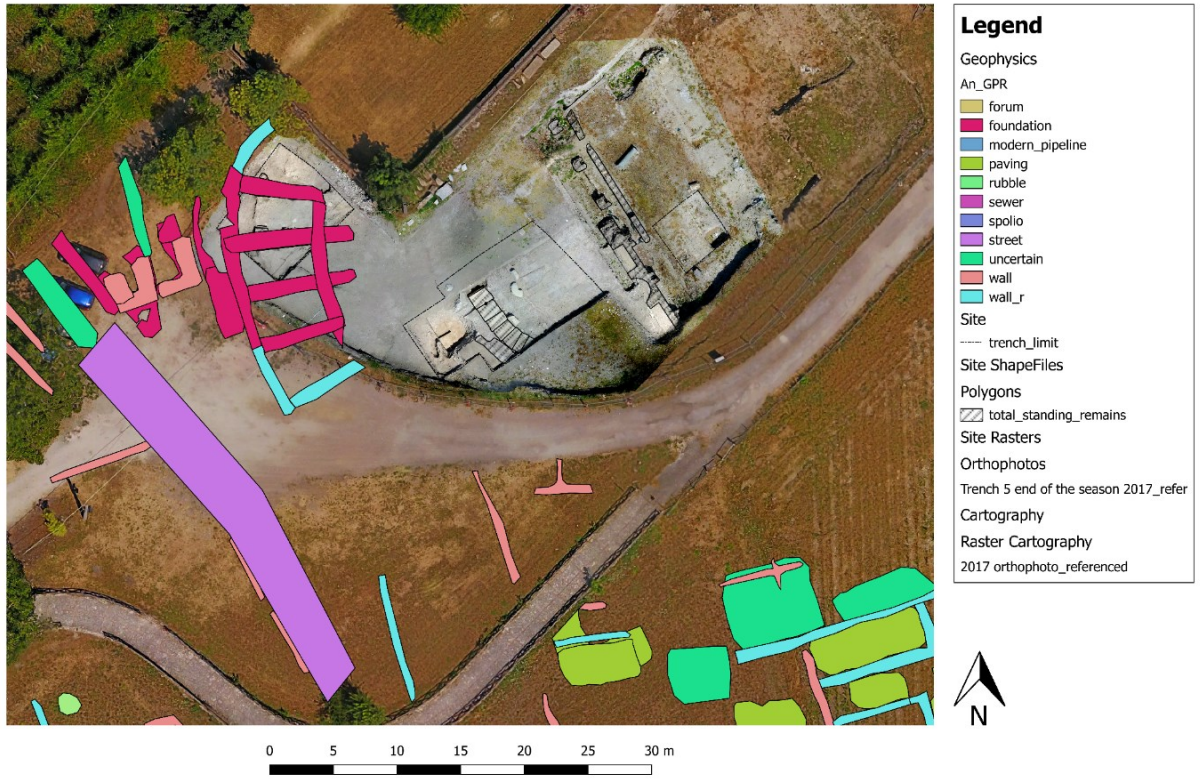
RESOURCES

- Good Introductory course
- If you are going to use QGIS its manual is available online and cover a lot
- And also a lot of embedded tutorials
- And a basic intro
- Another basic video that has a good intro on projections and RS In general Youtube is a good place to find a lot of video tutorial both on the general principles and on specific topics
- List of ESPG codes
- Google Maps Plotting Geographical data
- Repositories on Geographical data

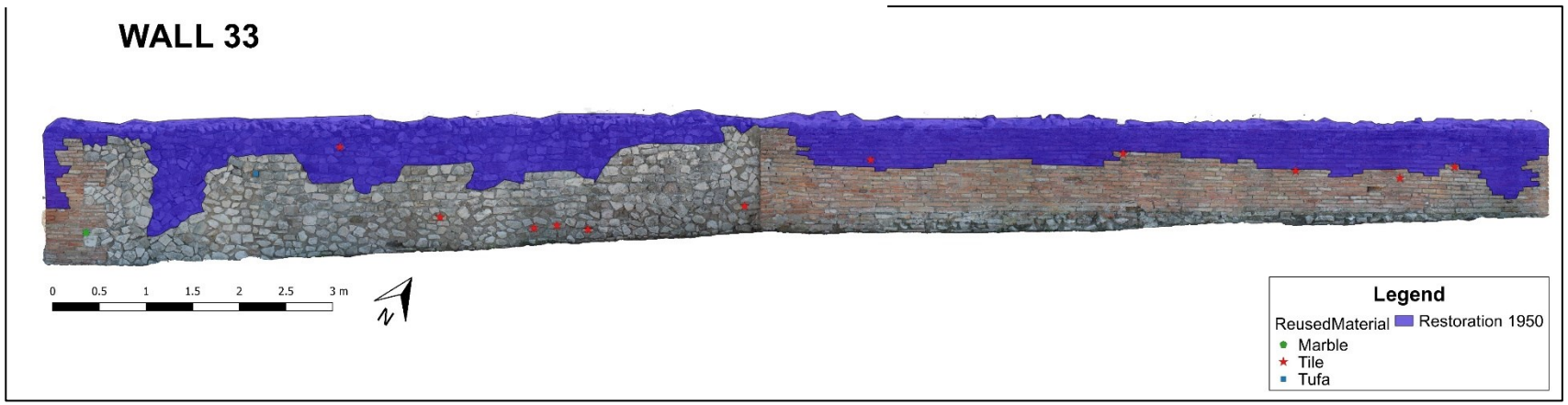
GEOGRAPHICAL INFORMATION SYSTEMS



Standing Remains and Geophisical data in the area of the theatre



WALL 33



WORKING WITH 3D DATA

- **Data Creation**
  - **Computer Drawing** Some software is extremely precise and suitable for complex engineering tasks, others can be more approximate emphasising creative free form modelling
  - **Active Scanners** The usual outcome is a point cloud that can then be further processed.
  - **Photogrammetry** and structure from motion Because they do not use active sensors they are called passive methods

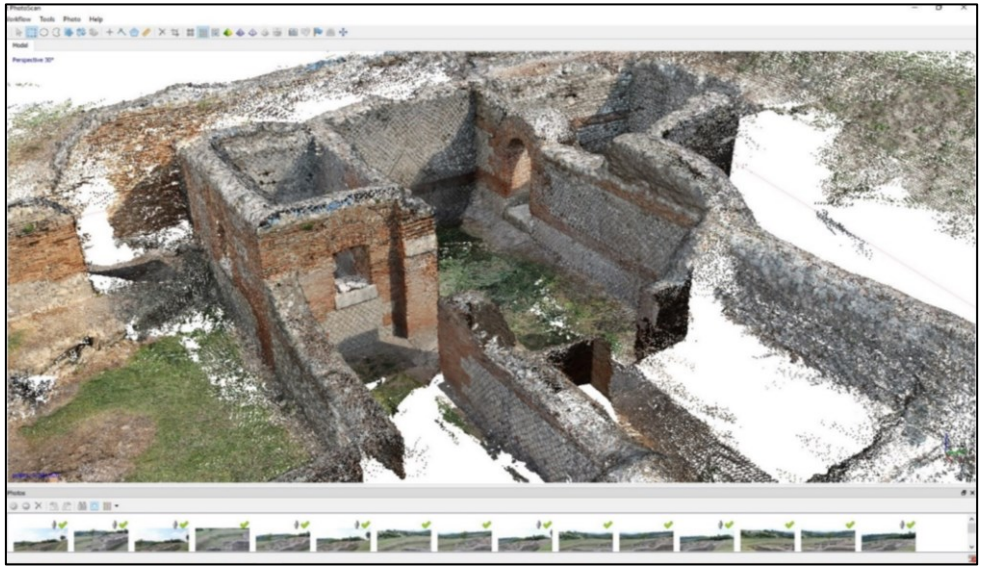
- **Data Wrangling** The first product of active and passive data acquisition is normally a point cloud. From the point cloud, different steps are needed to clean and refine the model

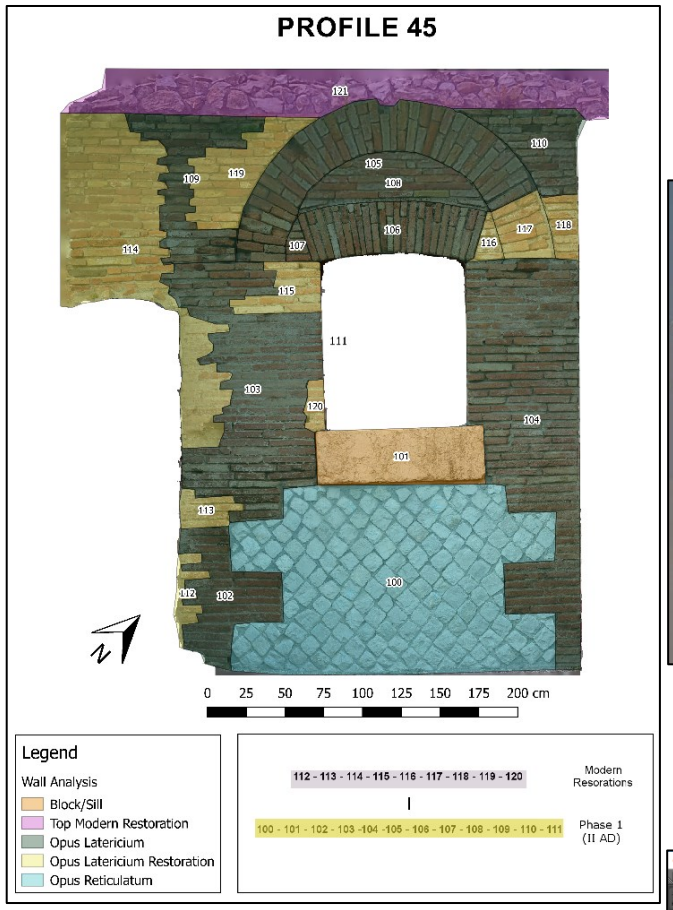
- **Outputs**
  - **3D Printing** 3D printing has been around for a long time but in the last 10-15 years it has become much more affordable and accessible
  - **Orthophotos Renders and Simulations**

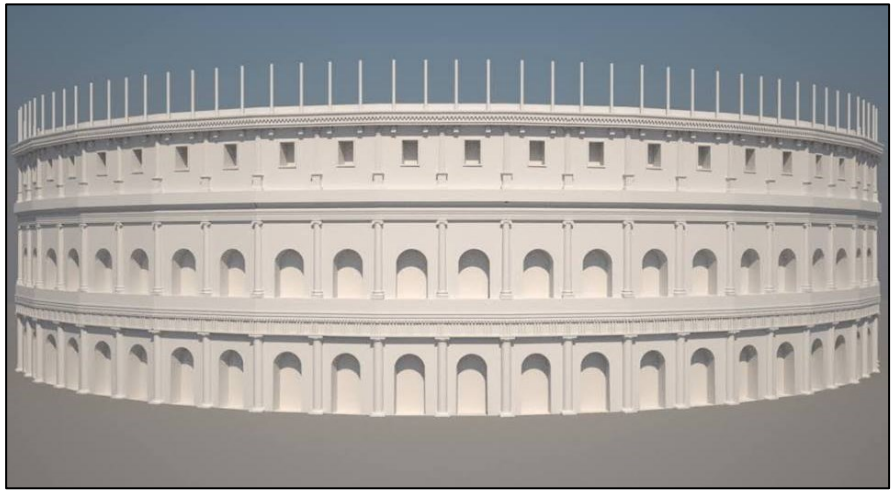
RESOURCES

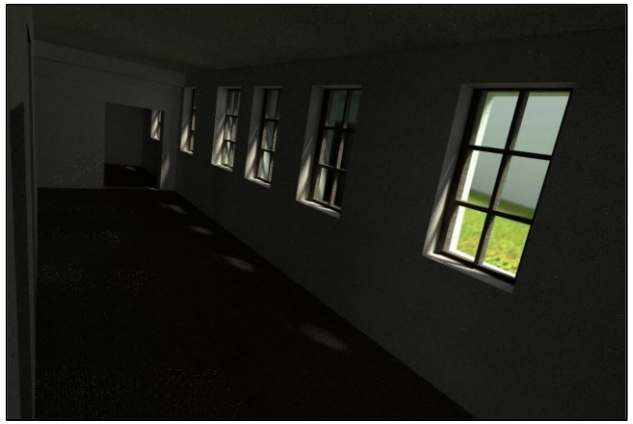
- 3D printing at UoE
- UCreate Bookable Equipment
- 3D Modelling and Cultural Heritage
- Metashape tutorials
- Working with 3D Data Pathway

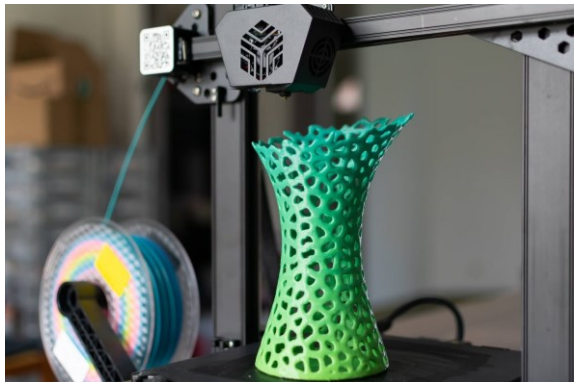
WORKING WITH 3D DATA











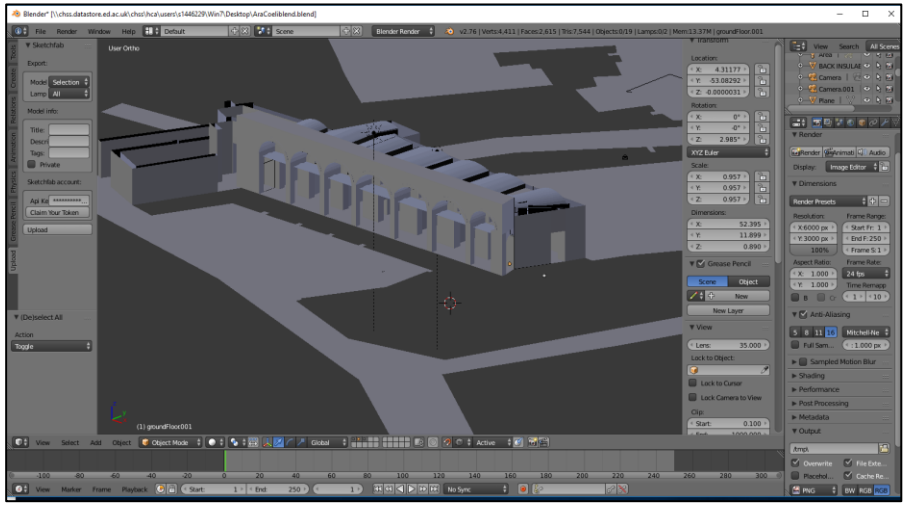


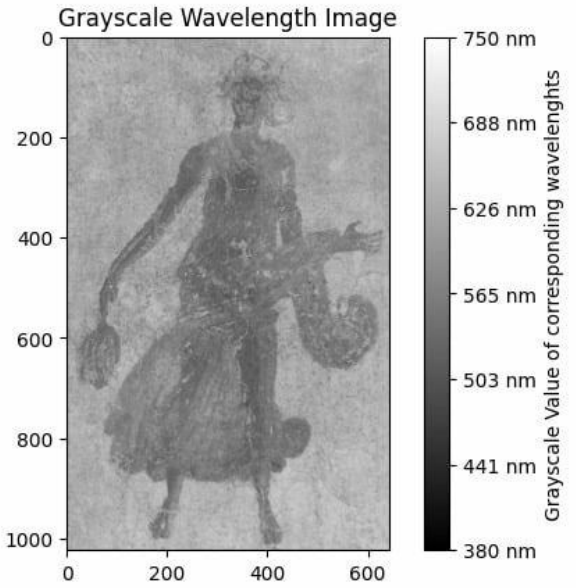
IMAGE PROCESSING

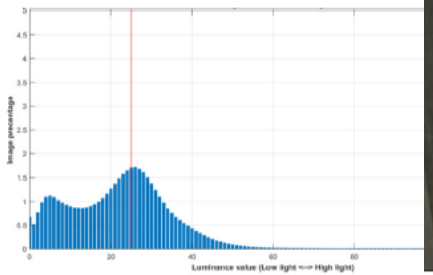
- **Digital Restauration:** Digitally restoring lost heritage (e.g. Buddha Project)
- **Image Segmentation:** Dividing an image into meaningful parts or regions based on certain characteristics, such as colour, intensity, or texture (used mostly in medicine diagnostic but many applications)
- **Image Analysis:** Extracting quantitative information from an image. This can involve measuring properties such as size, shape, or texture of objects within the image. Can be used across different colour bands eg. NDVI index
- **Pattern Recognition:** Identifying patterns within images based on predefined models or learned features. This is often used to identify features in landscapes
- **Machine Learning in Image Processing:** The integration of machine learning techniques, such as deep learning, to automatically learn and adapt to patterns and features in images. (e.g. automatically tag big datasets of images)

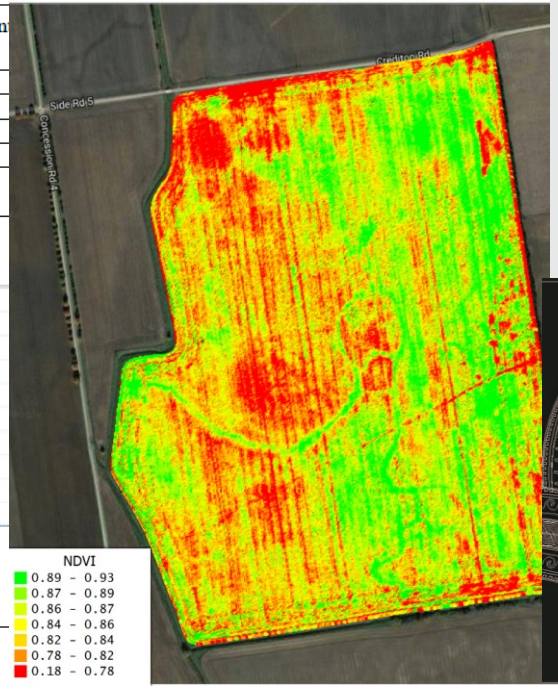
RESOURCES

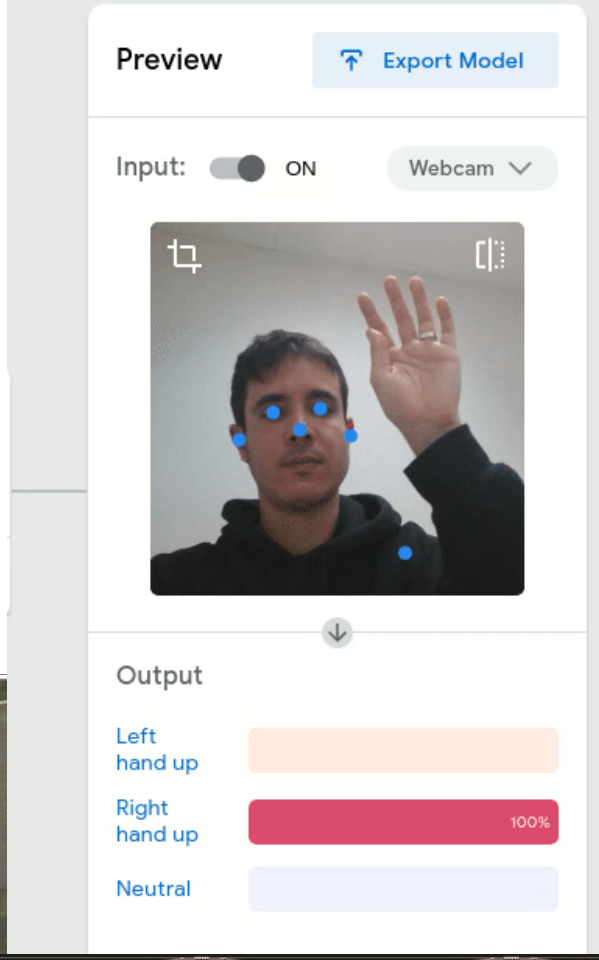
- Image Processing in the Digital Humanities
- Computer vision and Machine Learning
- Google teachable Machine
- Basic Raster Analysis (QGIS)
- Advanced Raster Analysis (QGIS)
- Image Processing with Python
- Computer Vision in DH

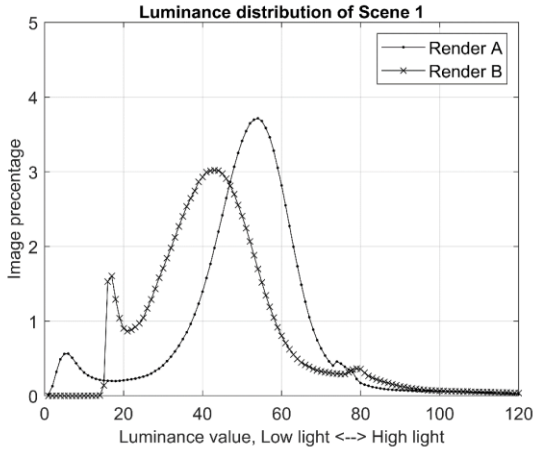
IMAGE PROCESSING

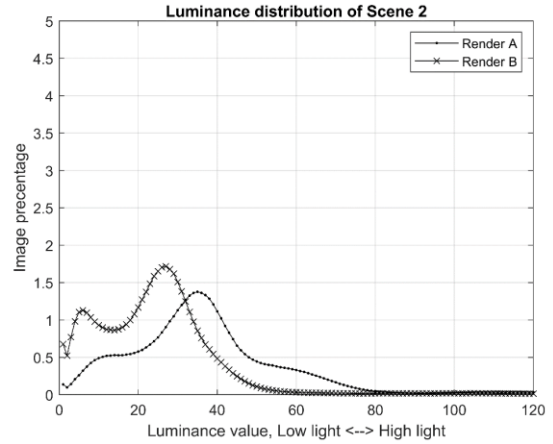


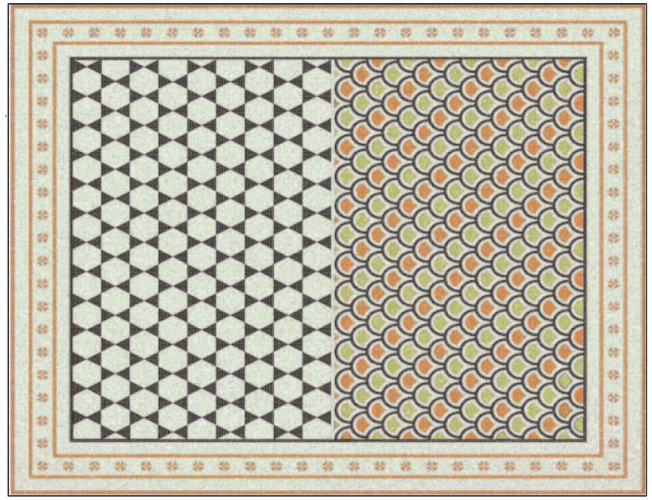
Title	*Medianum* 9 at the ground floor of the Apartmen Case a Giardino (glazed windows)
From	N
Time	14:30
Day	21/06
Sun Inclination	56.1°
Median Luminance	25/255
Histogram	











CROWDSOURCING SURVEY AND EXPERIMENT DATA

- Method of collecting data from human participants without having to meet them
- Can just be from links shared on (social) media, or specific crowdsourcing platforms (Amazon Mechanical Turk or Prolific)
- Usually much quicker than in-person data collection
- Can be more or less expensive than in-person data collection depending on methods
- Variety of tools available: point-and-click interfaces (e.g. Qualtrics, Testable, Gorilla) or code-based (e.g. jsPsych, PsyNet, from scratch in JavaScript + HTML)

Resources

- Blog post on advantages and disadvantages
- 2017 paper on crowdsourcing in cognitive science
- Two papers on data quality across different crowdsourcing platforms: 1 2

THE CDCS GITHUB PAGE

- Repositories with code and presentations developped by our instructors
- All material released under a CC BY-NC 4.0 license
- At the moment ~90 repositories and more than 30 collaborators

GITHUB

THE PEACEREP EXAMPLE

- Aims
  - Respond to a changing conflict context, to
  - Support peace and transition processes
  - Trends and shifts in the post-1990 peace and transition process landscape.

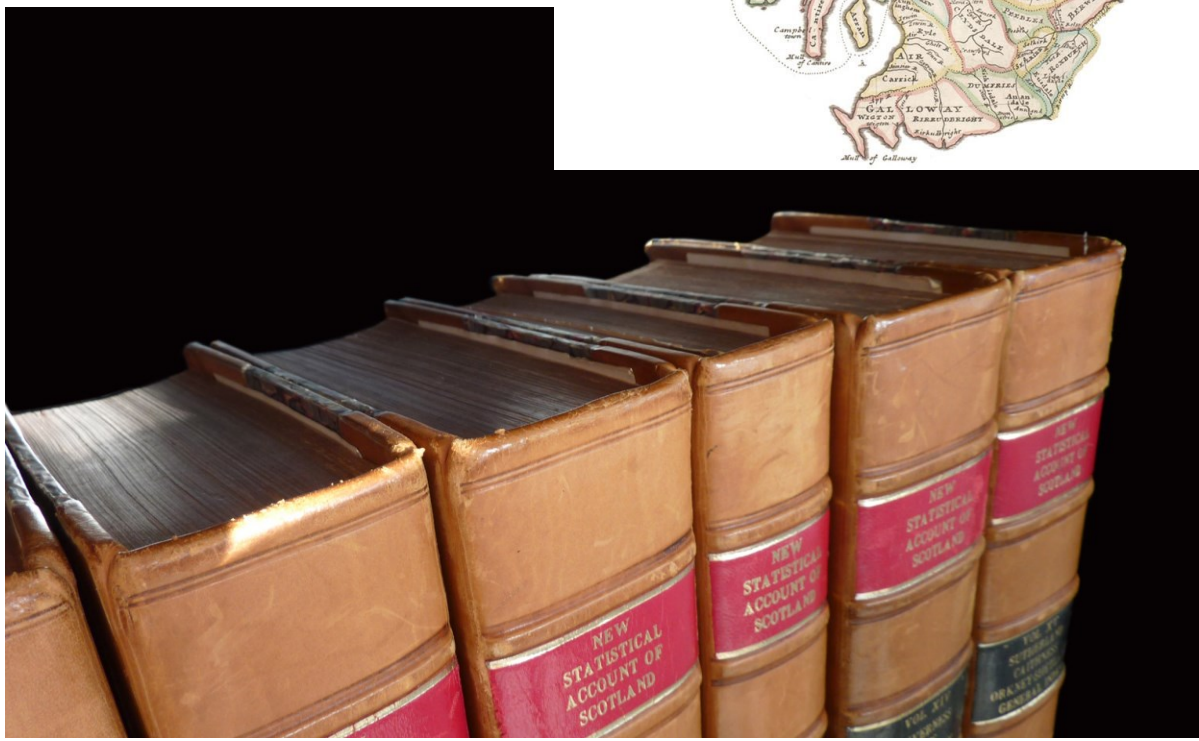
- Good practices of computational and interdisciplinary project on conflict
  - OCR
  - Text Analysis
  - Statistics
  - Interactive Data Visualisation
  - Crowdsource of Data

- Have a look at their website: https://peacerep.org/

PeaceRep
Peace and Conflict Resolution Evidence Platform

SCOTTISH ACCOUNT OF SCOTLAND

- The ‘Old’ *Statistical Account* (1791-99), under the direction of Sir John Sinclair of Ulbster, and the ‘New’ *Statistical Account* (1834-45) are reports of life In Scotland during the XVIII and XIX century.
- They offer uniquely rich and detailed parish reports for the whole of Scotland, covering a vast range of topics including agriculture, education, trades, religion and social customs.
- https://stataccscot.edina.ac.uk/static/statacc/dist/home
- Everything from changing fashions in dress to the different attitudes to smallpox inoculation and resulting high infant mortality between the north and south of Scotland
- Our datasets are **29,083 .txt files** corresponding to single reports from the statistical accounts



SCOTTISH ACCOUNT OF SCOTLAND

The websites you needs are

- Our Repository (where the code and data are)

https://github.com/Lucia-Michielin/HeritageDeepDive

- Posit (where you can run it if you do not have RStudio)

https://posit.cloud/

- Visualise the notebook

https://htmlpreview.github.io/? https://github.com/Lucia-Michielin/HeritageDeepDive/blob/main/DeepDivePosit.html

THANKS FOR YOUR ATTENTION

Q & A