

# 1.1 Introduction

Applied Data Analysis (ADA)

Oxford DH Summer School - 2023

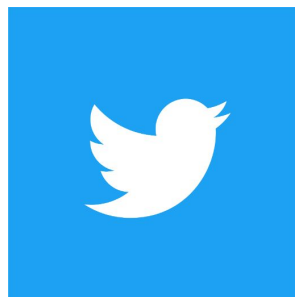
## Reminder!

Pre-school survey:

<https://forms.gle/9NfFeK9U7tNGw8YY8>

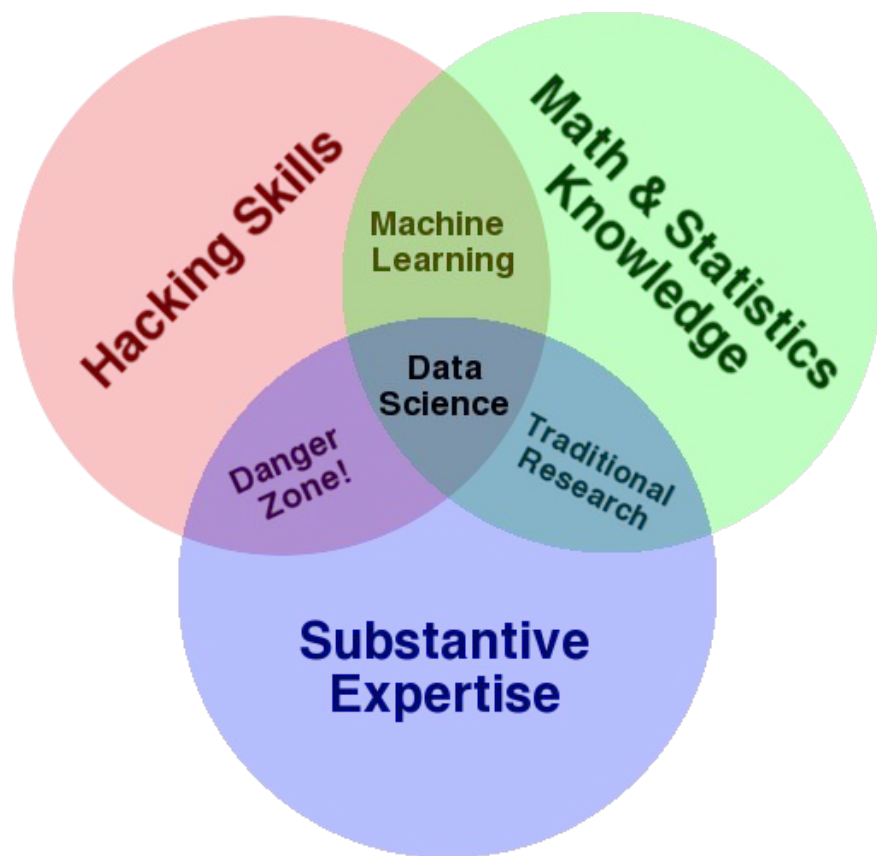












## What's in a title

**Applied:** we foreground techniques and methods in real-world scenarios, over implementations (but not theory!). I.e., we focus on what is done by a technique or method, rather than how [1].

**Data:** we use datasets which are too large to process (i.e., read) manually. We also consider datasets which are too large to be perused manually in their entirety without high risk of bias.

**Analysis:** we strive for insight. Data and tools serve little purpose without a motivation, question or information need, which should in turn help in creating new insight or knowledge.

## What's also there but not in the title

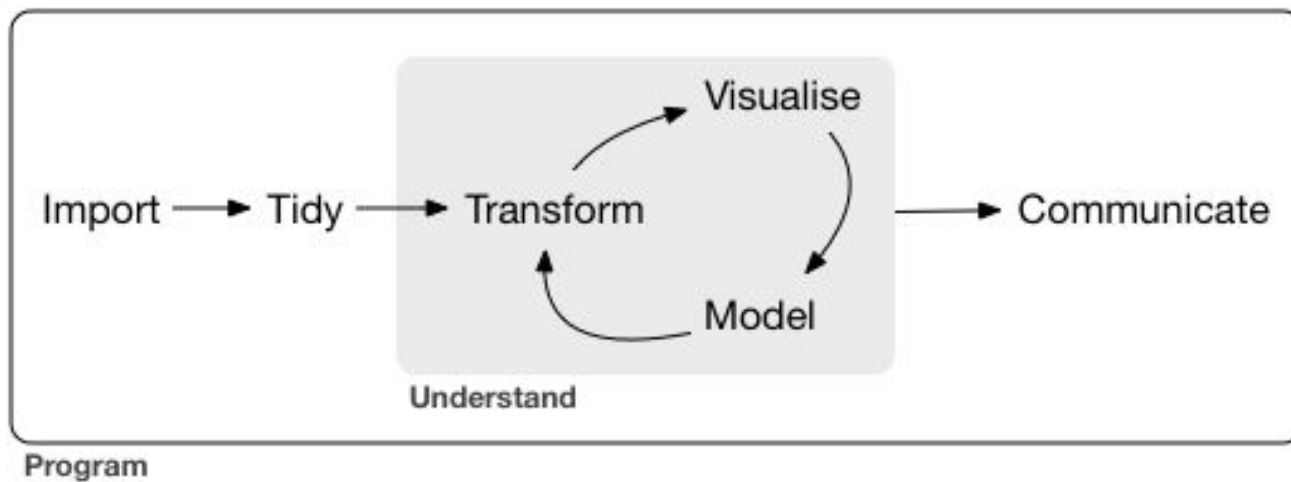
**Humanities:** we focus on data of interest to humanities scholars, professionals, practitioners.

**Advanced:** we assume some previous coding grounding on your side.

[1] <http://dhdebates.gc.cuny.edu/debates/text/99>.




## How we think about ADA



## **How we think about ADA**

- \* Observational rather than causal evidence.
- \* Complementary and enriching rather than exclusive.

## Schedule

<b>Monday</b>	<b>Tuesday</b>	<b>Wednesday</b>	<b>Thursday</b>	<b>Friday</b>
Introduction	Intro to Tidy data	Exploratory DA	Data visualization	Social Network Analysis
Data carpentry Intro to 🐼	Tidy data + 🐼	Exploratory DA	Geo-mapping	
<i>Afternoon session</i>	<i>Afternoon session</i>	<i>Afternoon session</i>	<i>Afternoon session</i>	Communicating and wrap-up

## Afternoon sessions

Options (attendees chose from the below):

- \* **Catching-up:** assistance is provided to clarify any issue from the previous classes or in setting-up your Python environment.
- \* **Exercises/project:** exercises or mini-projects will be provided for practice. Alternatively, attendees can bring their own mini-project to the class and work on it, individually or with others.
- \* **Lectures at Text to Tech:** attend the invited lectures given as part of the Text to Tech strand <https://www.dhoxss.net/from-text-to-tech>.

## Datasets

- \* Tweets from Elon Musk (21st century, text as a time series)
- \* 19th-century books from the British Library (metadata and text)
- \* Contracts of apprenticeship from Venice (16-17th centuries, numerical data)
- \* *Early African-American film database (20th century, metadata)*
- \* *Network of crypto art transactions (21st century, network data)*
- \* *Sample of historical British newspapers (19th century, metadata and text)*

*for afternoon sessions*

**Want more?** Have a look at [Humanities Datasets in Context](#) by Humanities Computing at Princeton.

## Teaching methods

Most classes are using **Jupyter notebooks**: interactive snippets of code and comments. Several short or not-so-short **assignments** are there for you to engage with. You should try to **play with code and data** as we go along, don't just execute our code.

Some classes use slideshows or the board.

**Questions and comments are encouraged.**

One of us will always move around: use **post-its** to signal if you have an issue (orange) or not (green, or something).

**Afternoon sessions** are for you to decide what to do, with us moving around to help.

*Please come to us with comments and feedback at any time during the week: we can always improve as we go.*



## Round of introductions

## **Results of the pre-school survey**

# Examples of data analysis applications

## Example of ADA: Mining Classics Citations from JSTOR

Mining of canonical references from Classics articles in JSTOR, to study scholarly reception of classical texts via citations as a proxy for scholar's attention.

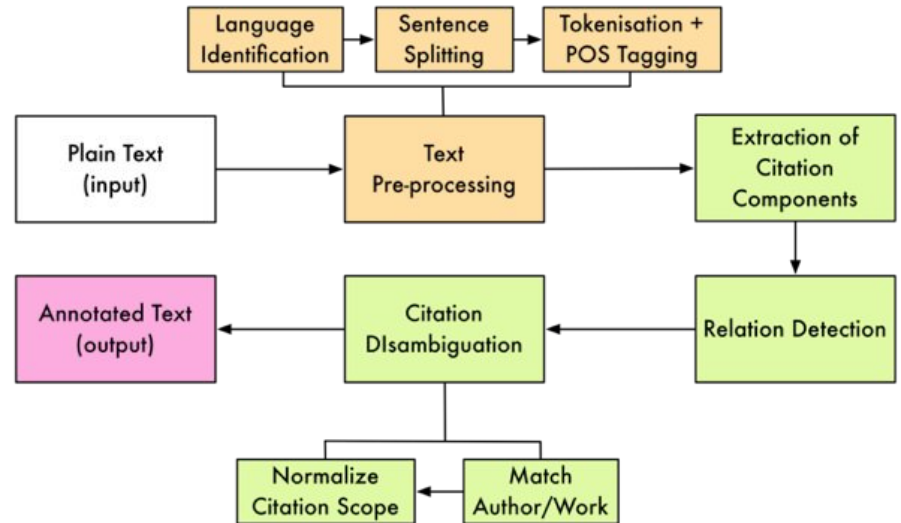
Pre-print: <https://doi.org/10.5281/zenodo.3736455>



## Data processing: mining canonical citations

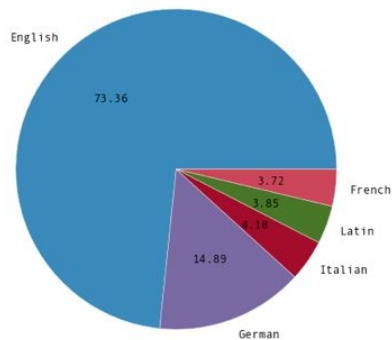
### Example of canonical references

first 'block' of arrival scenes in the Telemachy (*Odyssey* 1–4): the arrival of Athena-Mentes at Ithaca (Hom. Od. 1.103–324):<sup>9</sup> the goddess appears on the threshold of the palace (1.103–4) where she finds the suitors engaged in their respective activities (1.106–12). She is then seen by Telemachus (1.113, 1.118), who rises to accommodate the disguised goddess (1.119–20), takes her by the hand (1.121), and makes her enter (1.125); he welcomes her (1.122–4), takes her spear (1.121 and 1.127–9), and invites her to sit down (1.130–2); the dinner is prepared (1.136–43), consumed (1.149), and concluded (1.150); the visitor's identity is finally revealed (1.169–93) and information is exchanged (1.194–305) before Athena escapes Telemachus' attempt to retain her (1.309–19). This 'classic' hospitality scene highlights Telemachus'



<https://citedloci.org/>

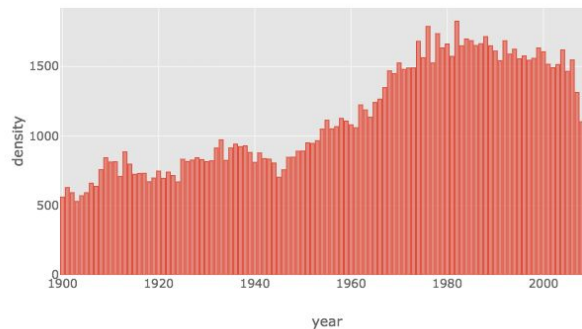
## Data overview: Classics articles in JSTOR



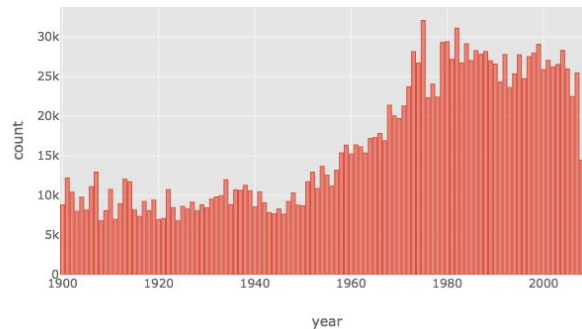
	Number
Total articles	138,821
Successfully processed articles	119,723
Sentences	34,853,399
Tokens	865,075,857
Extracted canonical references	1,649,868
Extracted author mentions	1,448,163
Extracted work mentions	4,665

Table 2: Basic statistics about the JSTOR data.

Number of articles per year in JSTOR (1900-2009)

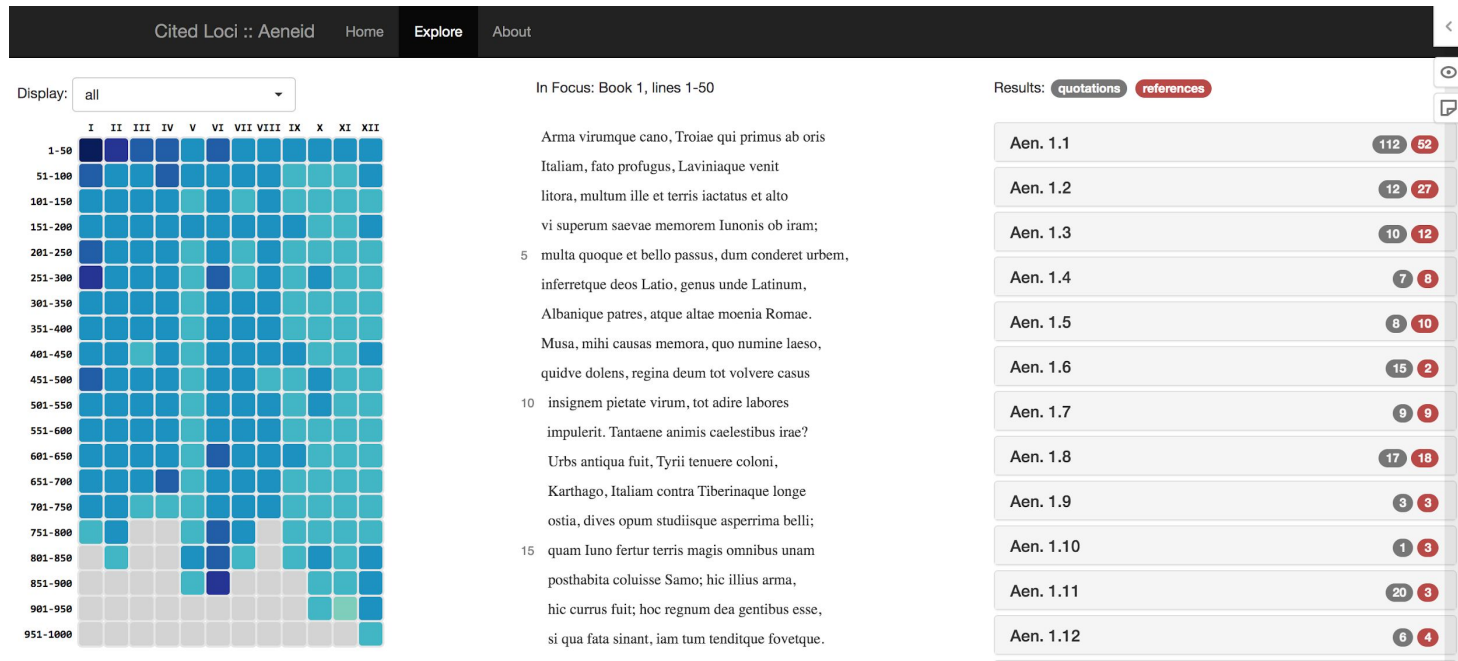


Number of canonical references per year





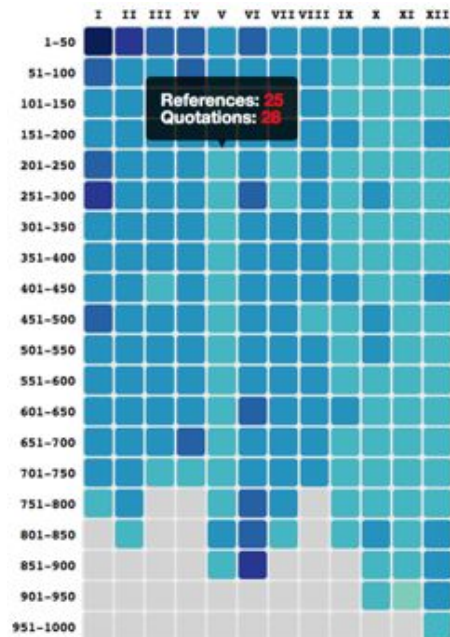
# Interactive data visualization: *Aeneid* in JSTOR



[https://mromanello.github.io/Aeneid\\_in\\_JSTOR/](https://mromanello.github.io/Aeneid_in_JSTOR/)

<http://labs.jstor.org/blog/cited-loci-of-the-aeneid/>

## Interactive data visualization: *Aeneid* in JSTOR



Distant reading

In Focus: Book 6, lines 251-300

adventante dea. "Procul O procul este, profani,"  
conclamat vates, "totoque absistite luco;  
260 tuque invade viam, vaginaque eripe ferrum:  
nunc animis opus, Aenea, nunc pectore firmo."  
Tantum effata, furens antro se immisit aperto;  
ille ducem haud timidis vadentem passibus aequat.  
Di, quibus imperium est animarum, umbraeque silentes,

Close reading  
(primary literature)

Results: quotations references

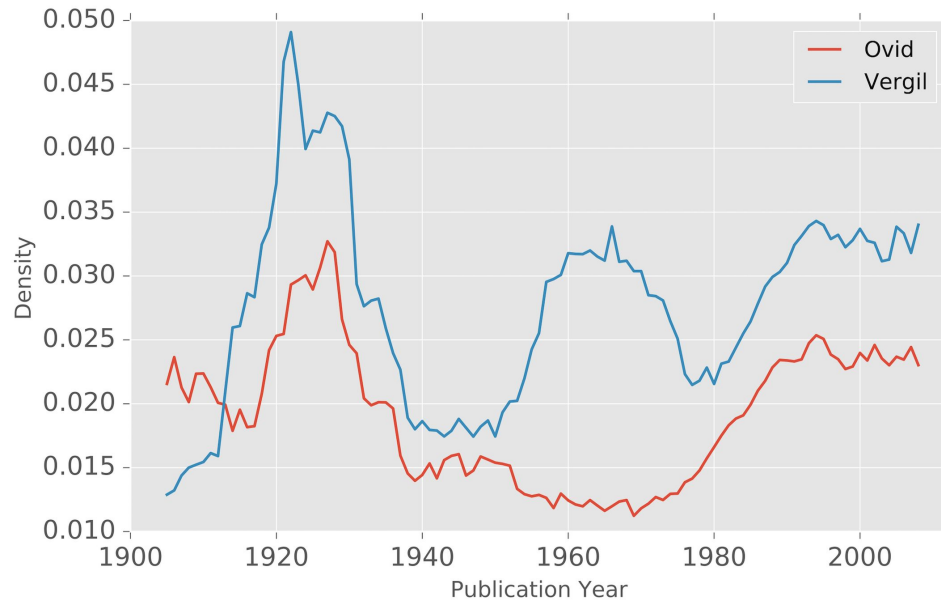
Line	Quotations	References
Aen. 6.258	10	7
Aen. 6.260	2	1
Aen. 6.263	1	1
<u>Aen. 6.264</u>	6	13

**Proserpina's Tapestry in Claudian's "De raptu": Tradition and Design**  
MICHAEL VON ALBRECHT  
*Illinois Classical Studies* 1989  
DOI: [10.2307/23064383](https://doi.org/10.2307/23064383)

Besides the role of the Sibyl, Claudian also takes the role of Virgil, invoking the gods of the underworld, as his great predecessor had done (1. 20 ; Aen. 6. 264).  
— (reference: 571a493ab634c1c699eed3fa)

Close reading  
(secondary literature)

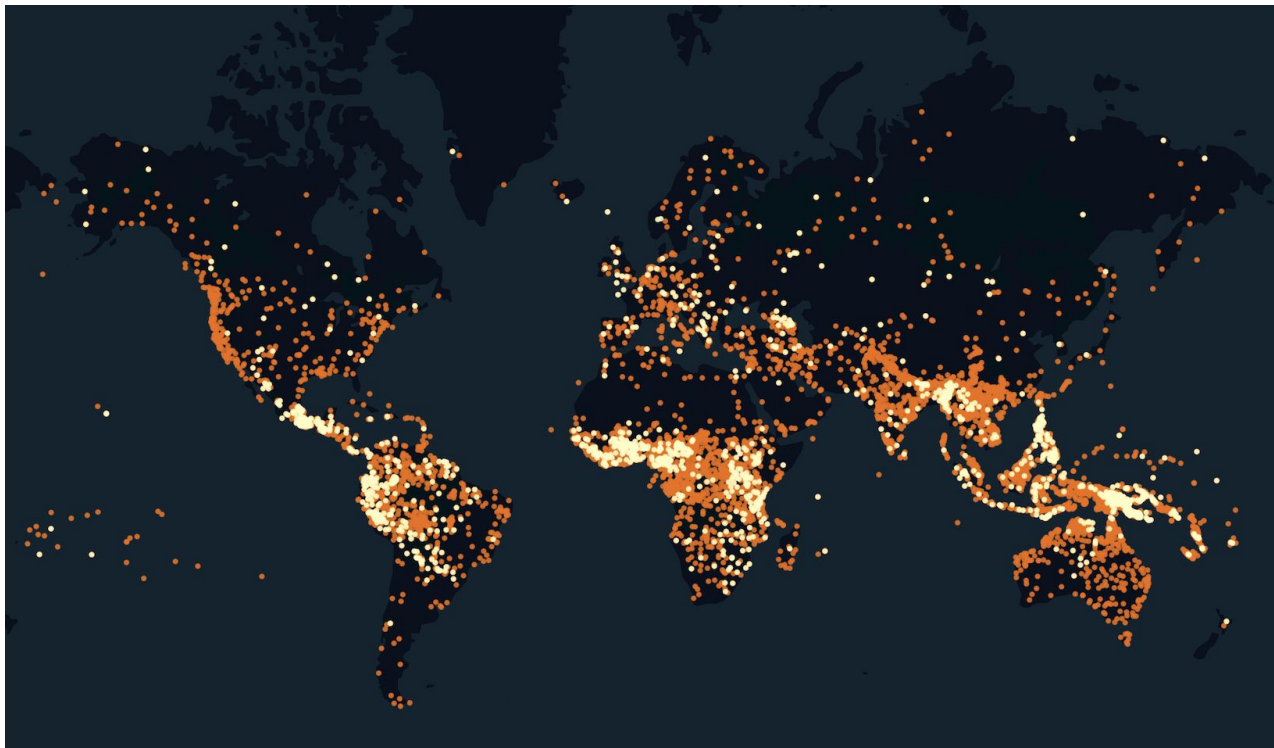
## Longitudinal analysis: Waves of reception in Ovid and Vergil



T. Ziolkowski 2009, *Ovid in the Twentieth Century*.



## Example of ADA: the typology of *when*-clauses across the world's languages



*How do different languages express temporal relationship between events?*

- Alignment of 1400+ Bible translations at token level
- Application of data-analysis techniques from outside the humanities applied to the humanities

**Figure 1**

Languages in the data (yellow) among the languages of the world (orange). Map generated using Kepler and geopandas in Python.

## Source data wrangling example: Maori (Austronesian)

42005038, but new wine must be put into  
fresh wineskins, engari me riringi te waina  
hou ki nga ipu hou

but new wine must be put into fresh  
wineskins {##} engari me riringi te waina  
hou ki nga ipu hou {##} 0-0 1-5 2-4 3-1  
5-2 6-6 7-9 8-8

42005038, but new wine must be put into  
fresh wineskins, but (engari) new (hou)  
wine (waina) must (me) be (NOMATCH) put  
(riringi) into (ki) fresh (hou) wineskins  
(ipu)

...

[Repeat for all languages]

Regex to find all occurrences of  
connective *when* and its parallels




	eng	mri	por	fin	...	kaz	kor
1	when	no	quando	kun		қашан	때 올
2	when	ka	quando	jolloin		кейін	때 올
$n$	...	...	...	...	...	...	...



Given word  $w$  at row  $n$ , we can deduct how  
different or similar its usage (e.g. meaning,  
sense) is from word  $w+1$  at row  $n+1$  by counting  
how many languages use the same word for both  
rows versus how many use two different words →  
*Hamming Distance*

## Similarity matrices and dimensionality reduction



	x	y
0	0.153564	-0.077936
1	-0.180358	-0.079010
2	0.294392	-0.123286
3	-0.208843	-0.074415
4	0.052290	0.275351
...	...	...
471	-0.116570	-0.113176
472	-0.093750	-0.159694
473	0.177257	-0.140021
474	-0.170266	0.051007
475	-0.090192	0.092315

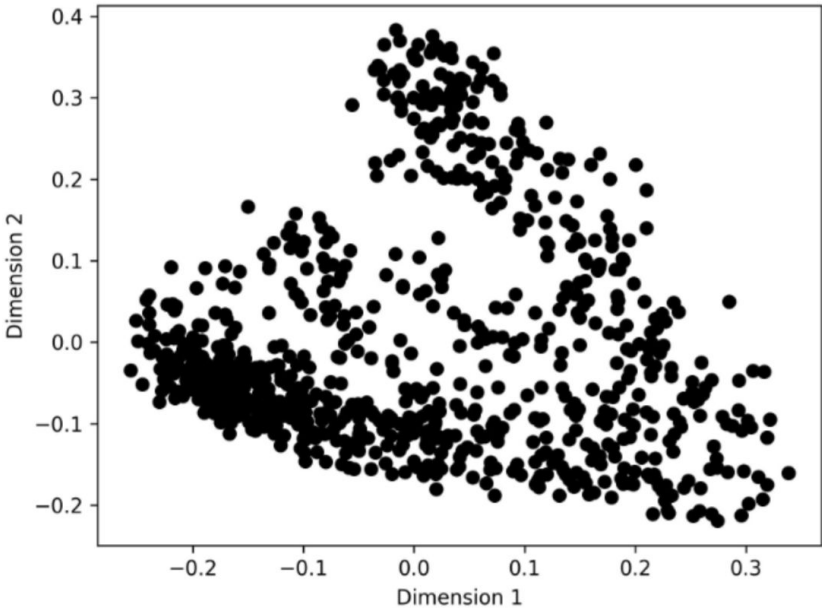
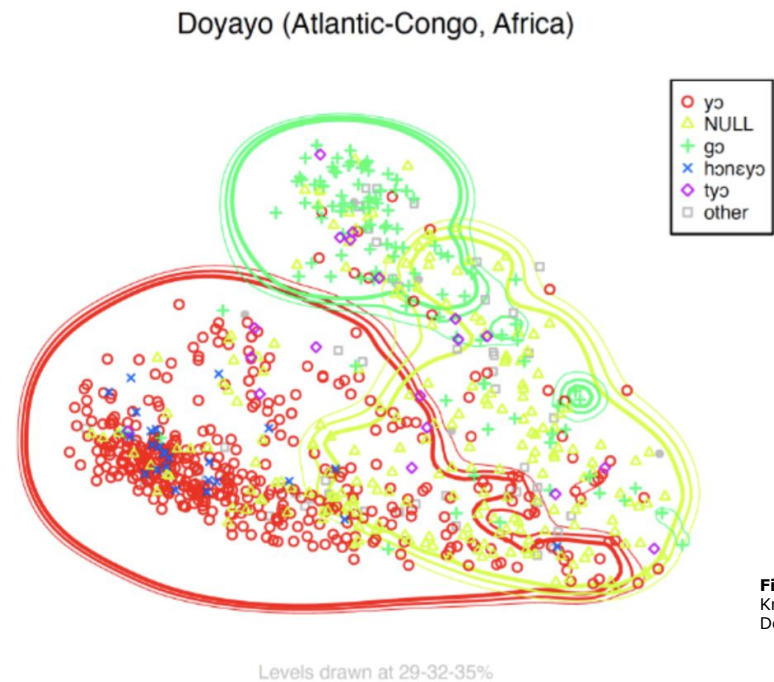
	0	1	2	...	473	474	475
0	0.000000	0.717813	0.826331	...	0.697500	0.774194	0.839394
1	0.717813	0.000000	0.831435	...	0.681905	0.733471	0.786543
2	0.826331	0.831435	0.000000	...	0.797583	0.828767	0.889262
3	0.651341	0.459270	0.770732	...	0.598077	0.676275	0.791045
4	0.802083	0.762994	0.575499	...	0.731507	0.783951	0.839344
...	...	...	...	...	...	...	...
471	0.812709	0.793103	0.843373	...	0.715254	0.747368	0.810277
472	0.786372	0.739071	0.855172	...	0.654851	0.740000	0.832962
473	0.697500	0.681905	0.797583	...	0.000000	0.657382	0.792023
474	0.774194	0.733471	0.828767	...	0.657382	0.000000	0.823151
475	0.839394	0.786543	0.889262	...	0.792023	0.823151	0.000000

**Figure 2**  
Hamming-distance matrix

**Figure 3**  
2-dimensional matrix obtained after dimensionality reduction of the Hamming distance matrix using multi-dimensional scaling (MDS).



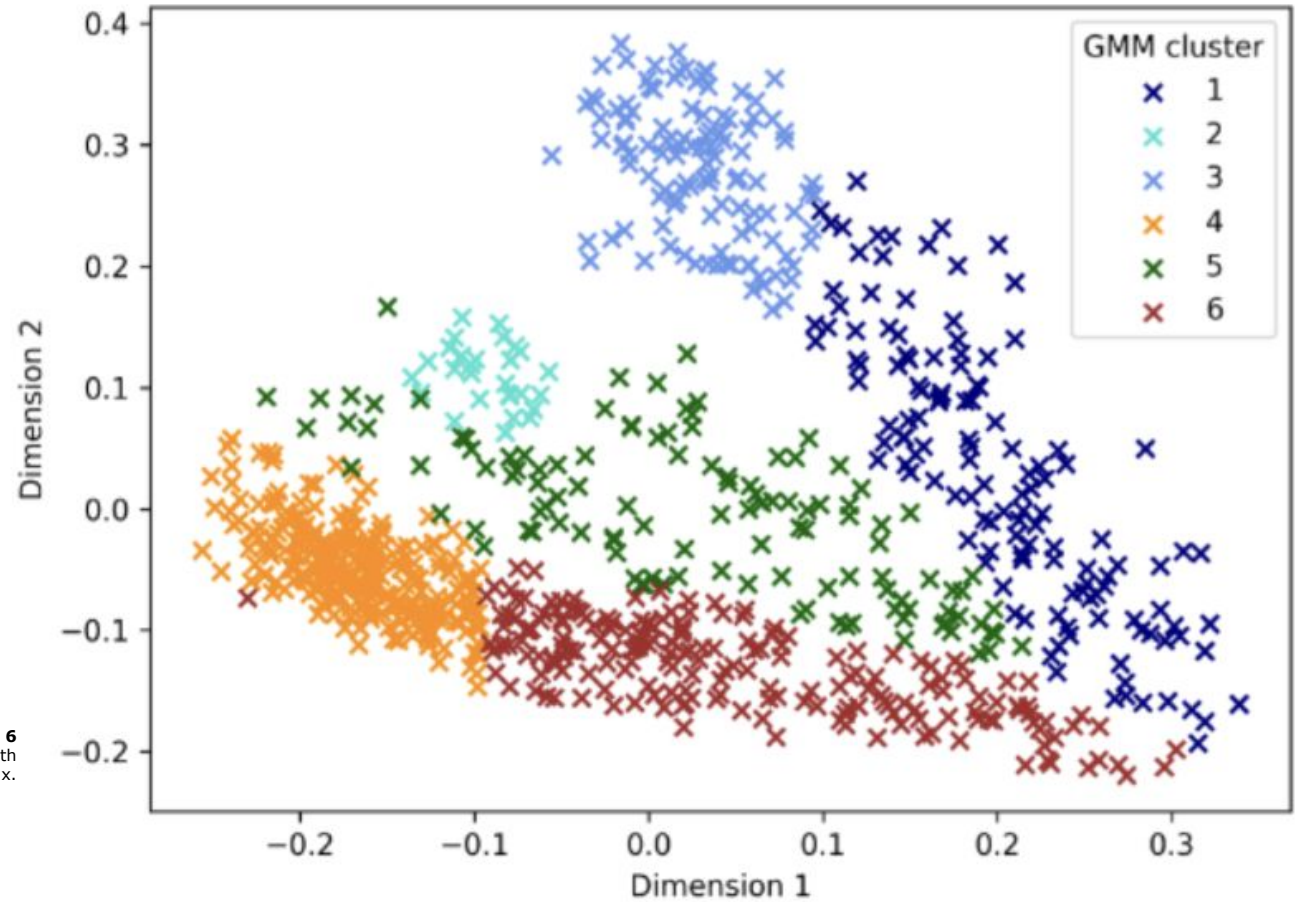
# Visualization of MDS matrix + Kriging



**Figure 4**  
Simple MDS scatter plot.

**Figure 5**  
Kriging map for *when*-parallels in  
Doyayo (Atlantic-Congo, Africa).

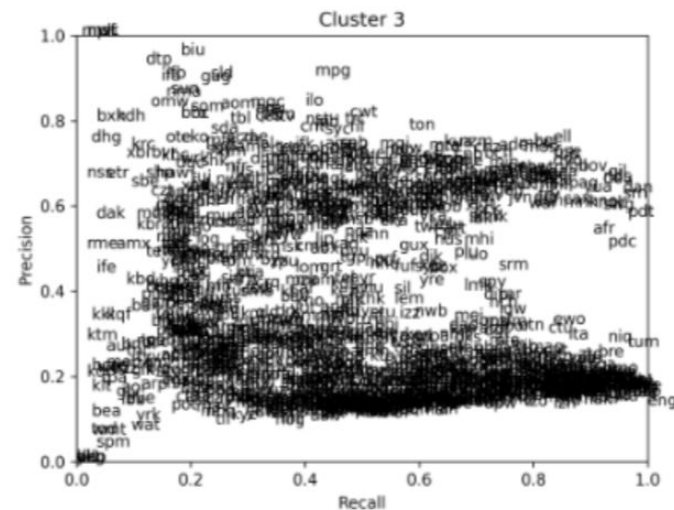
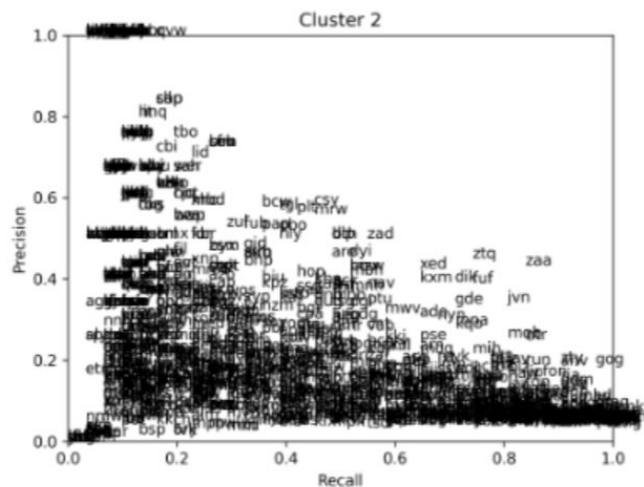
## Clustering data points



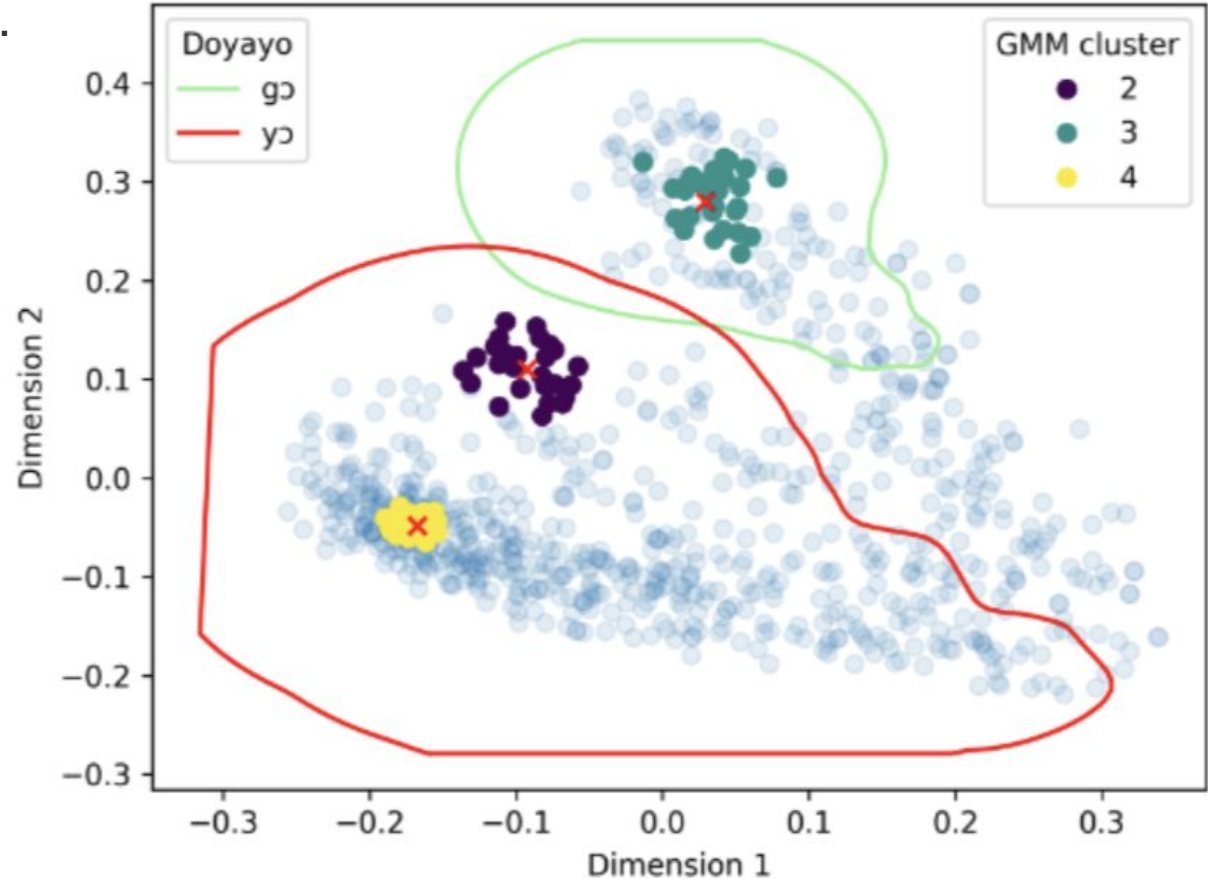
**Figure 6**  
Plot for a Gaussian Mixture Model with  
6 clusters fitted on the MDS matrix.

## Precision, recall and F1 score

$$pr = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$



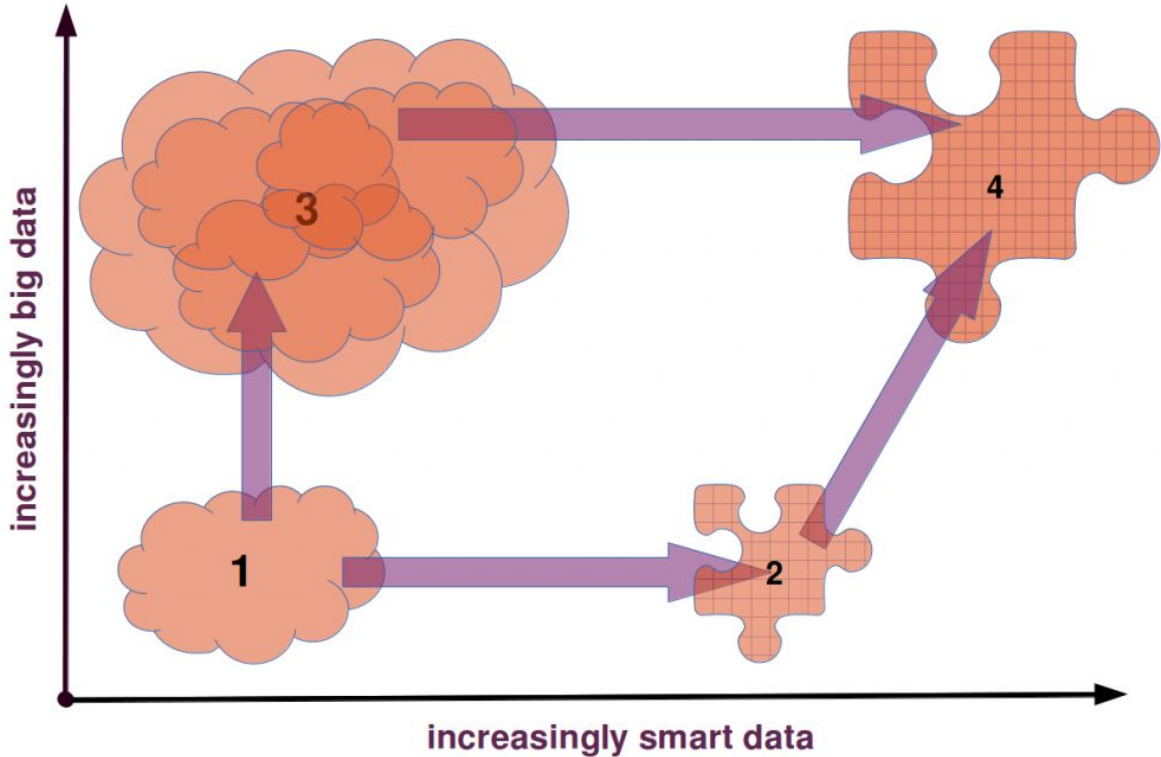
## Combining all of the above...



**Figure 6**

Result of a 30-nearest-neighbour search using the 'balltree method', with an example of its application to Doyayo (Atlantic-Congo, Africa). The red marks are the centroid of the respective GMM clusters (as represented in Figure 6, previous slide), while the points in which they are embedded are their 30 nearest neighbours. The contour lines in green and red correspond to the Kriging areas for Doyayo gO and yO at 29% probability.

# Conclusion



## Tools and materials

Canvas: <https://bit.ly/dhoxss-ada-2023>

Code and data are on GitHub: <https://github.com/mromanello/ADA-DHOxSS>

Slack channel: see invitation in [Canvas – Applied Data Analysis](#)

*Let's try these out!* <https://mybinder.org/v2/gh/mromanello/ADA-DHOxSS/master>

*Twitter/Mastodon: #DHOxSS2023 and #ADA*



## Setup and warm-up

- Launch Binder from the repo (it might take a little while)
- Go to /notebooks
- Open the *HelloWorld* notebook
- Play along and see if it's all sound and clear, use post-its to signal if there is any issue

**If you want to work locally, you are welcome to fork the repo and use your own copy.**

**We can help setting you up during lunch time and during the last afternoon session.**