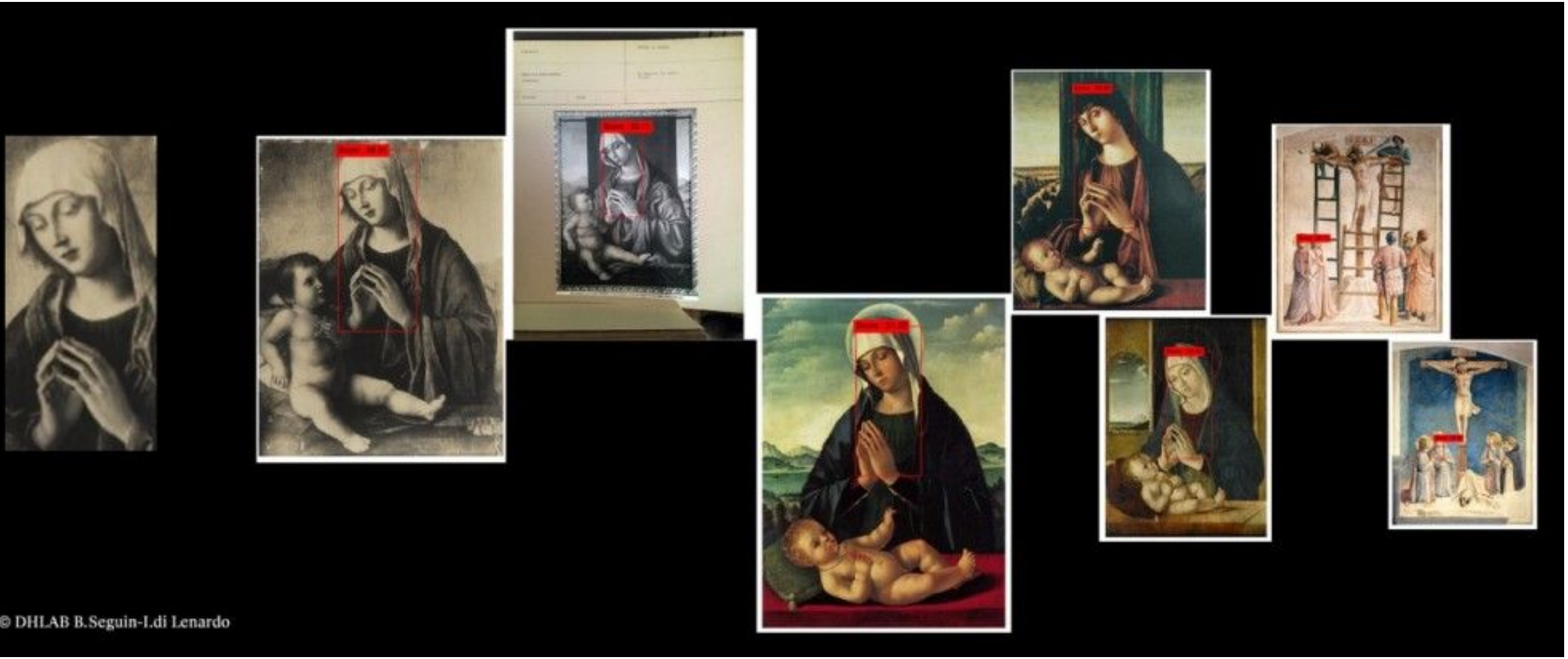
1.1 Introduction

Applied Data Analysis (ADA)

Oxford DH Summer School - 2022

Reminder!

Pre-school survey:
https://forms.gle/6yfTZPxnRQxcUGnx6

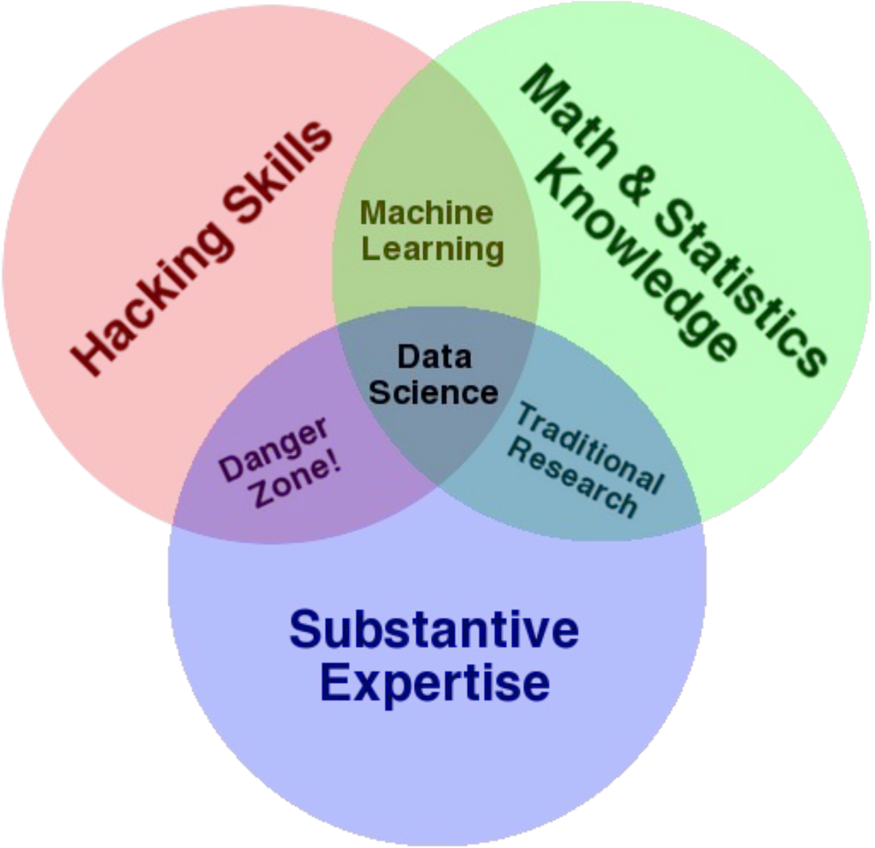


© DHLAB B.Seguin-I.di Lenardo

Google
Books

INTERNET
ARCHIVE

FAKE
NEWS


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

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.

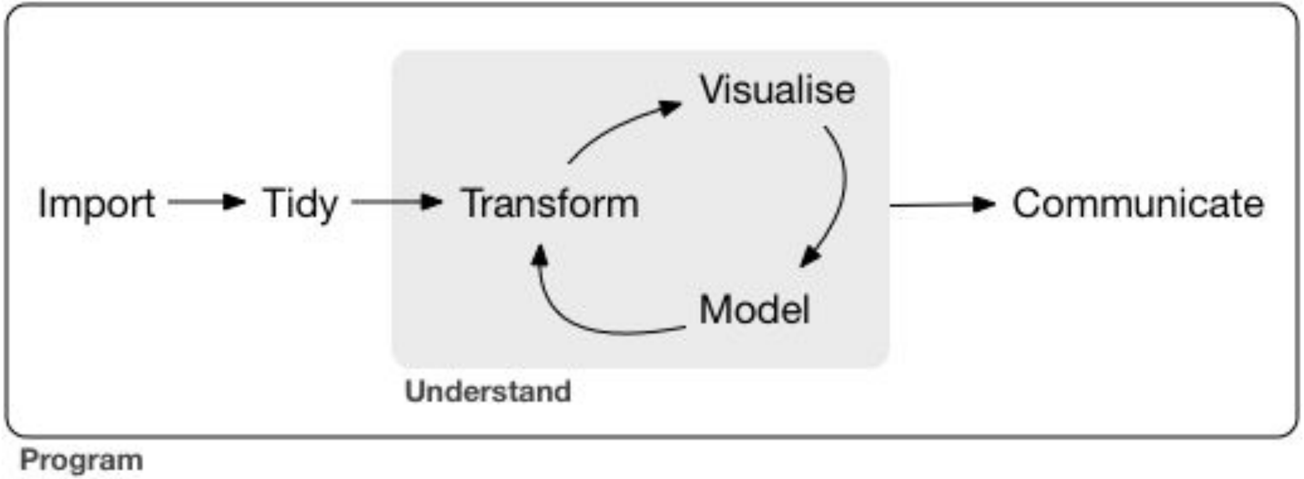
What’s also there but not in the title

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



See https://r4ds.had.co.nz/introduction.html

How we think about ADA

* Observational rather than causal evidence.

* Complementary and enriching rather than exclusive.

Topics

1. Python ADA skills
2. Import data
3. Tidy data
4. Wrangling
5. Visualizing
6. Modelling
7. Communicating

Schedule

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Introduction	Skills 🐼	Exploratory DA	Joker session	Social Network Analysis
🐍 Refresher Data carpentry Intro to 🐼	Tidy data	Exploratory DA	Data visualization	Dataset construction
Afternoon session	*Afternoon session*	*Afternoon session*	*Afternoon session*	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)*

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

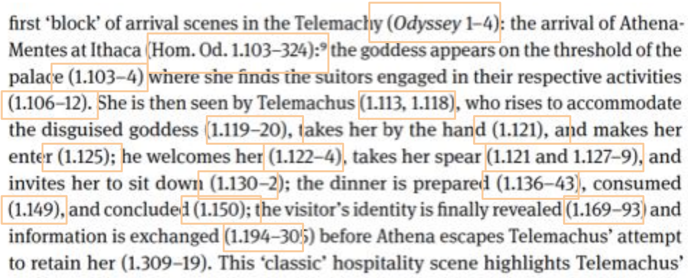
Example of ADA: Mining Classics Citations from JSTOR

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.

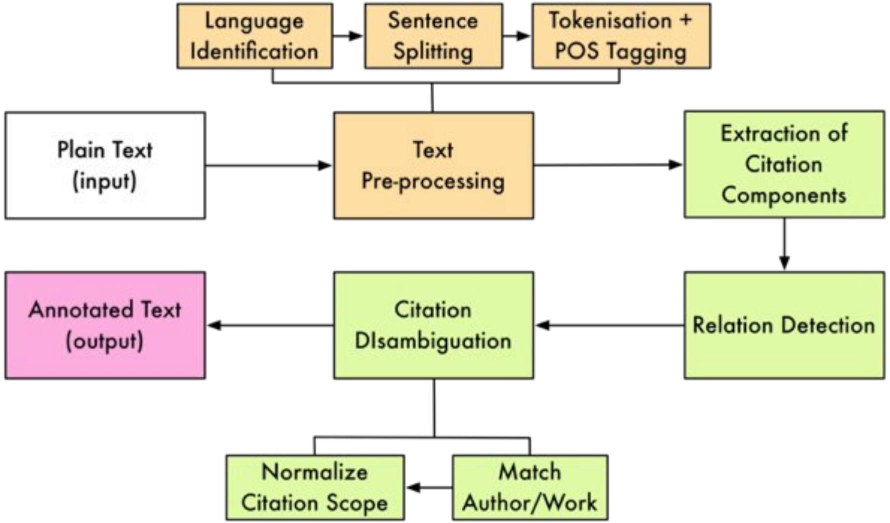
Pre-print: https://doi.org/10.5281/zenodo.3736455

Data processing: mining canonical citations

Example of canonical references

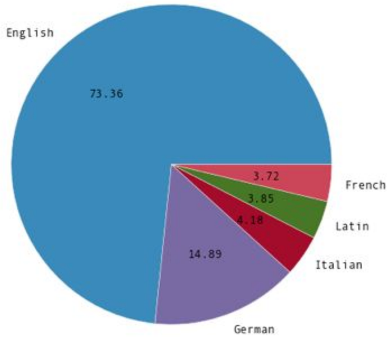


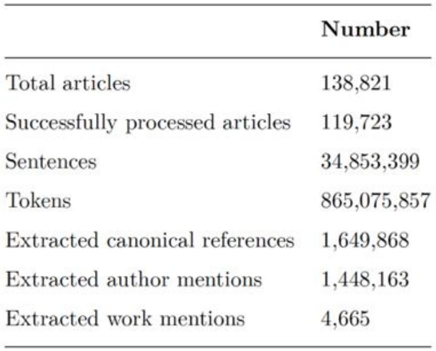
first 'block' of arrival scenes in the Telemachy (*Odyssey* 1–4): the arrival of Athena-Mentes at Ithaca (Hom. Od. 1.103–324):[9] the goddess appears on the threshold of the palace (1.103–4) where she finds the suitors engaged in their respective activities (1.106–12). She is then seen by Telemachus (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/

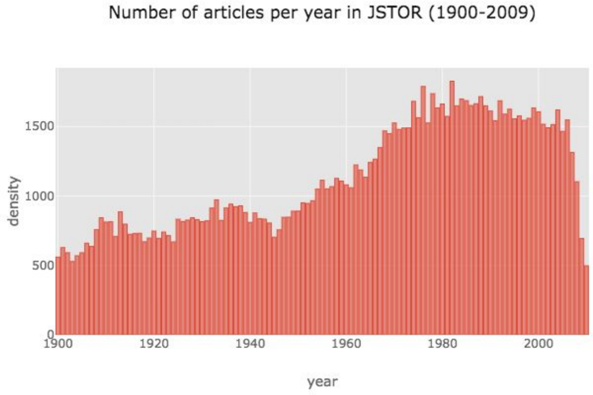
Data overview: Classics articles in JSTOR

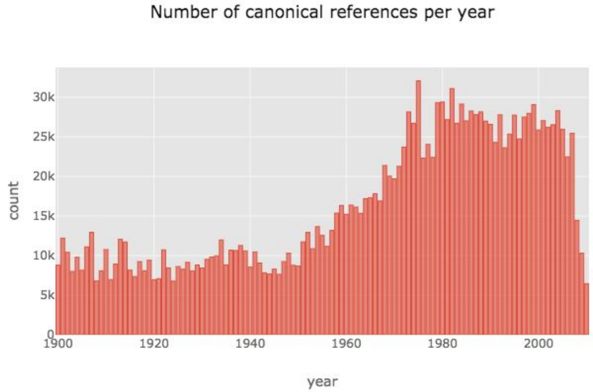




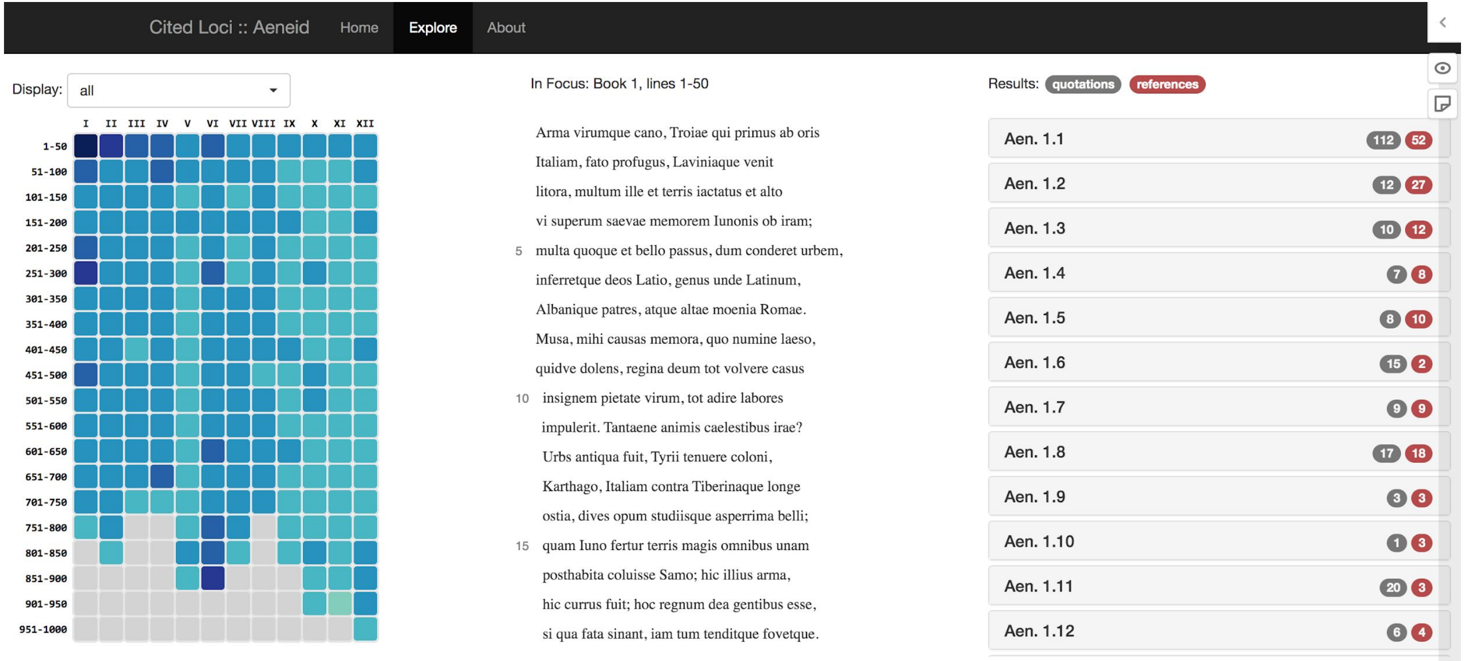
	Number
Total articles	138,821
Successfully processed articles	119,723
Sentences	34,853,399
Tokens	865,075,857
Extracted canonical references	1,649,868
Extracted author mentions	1,448,163
Extracted work mentions	4,665

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



Display: all

In Focus: Book 1, lines 1-50

Arma virumque cano, Troiae qui primus ab oris
Italiam, fato profugus, Laviniaque venit
litora, multum ille et terris iactatus et alto
vi superum saevae memorem Iunonis ob iram;
multa quoque et bello passus, dum conderet urbem,
inferretque deos Latio, genus unde Latinum,
Albanique patres, atque altae moenia Romae.
Musa, mihi causas memora, quo numine laeso,
quidve dolens, regina deum tot volvere casus
insignem pietate virum, tot adire labores
impulerit. Tantaene animis caelestibus irae?
Urbs antiqua fuit, Tyrii tenuere coloni,
Karthago, Italiam contra Tiberinaque longe
ostia, dives opum studiisque asperrima belli;
quam Iuno fertur terris magis omnibus unam
posthabita coluisse Samo; hic illius arma,
hic currus fuit; hoc regnum dea gentibus esse,
si qua fata sinant, iam tum tenditque fovetque.

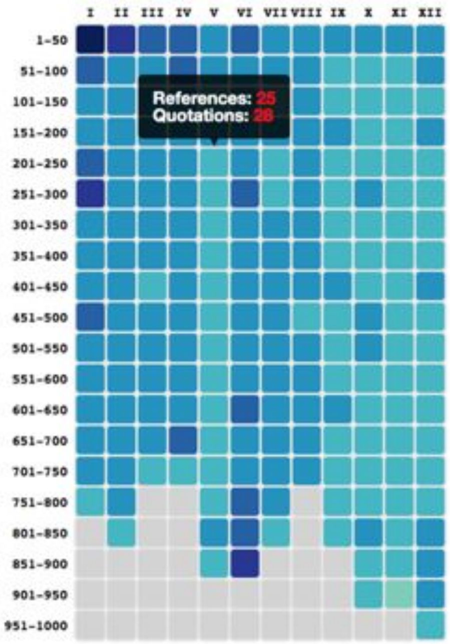
Results: quotations references

	quotations	references
Aen. 1.1	112	52
Aen. 1.2	12	27
Aen. 1.3	10	12
Aen. 1.4	7	8
Aen. 1.5	8	10
Aen. 1.6	15	2
Aen. 1.7	9	9
Aen. 1.8	17	18
Aen. 1.9	3	3
Aen. 1.10	1	3
Aen. 1.11	20	3
Aen. 1.12	6	4

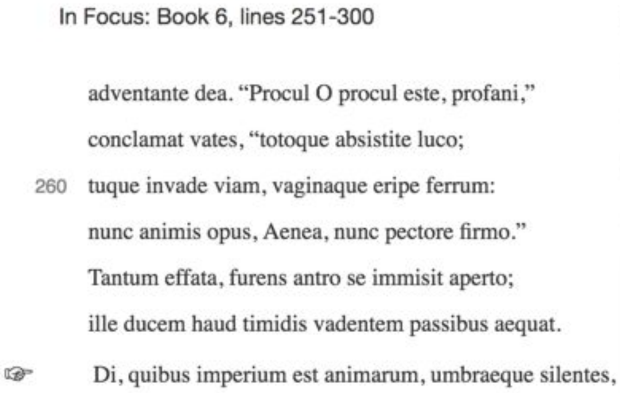
https://mromanello.github.io/Aeneid_in_JSTOR/

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:

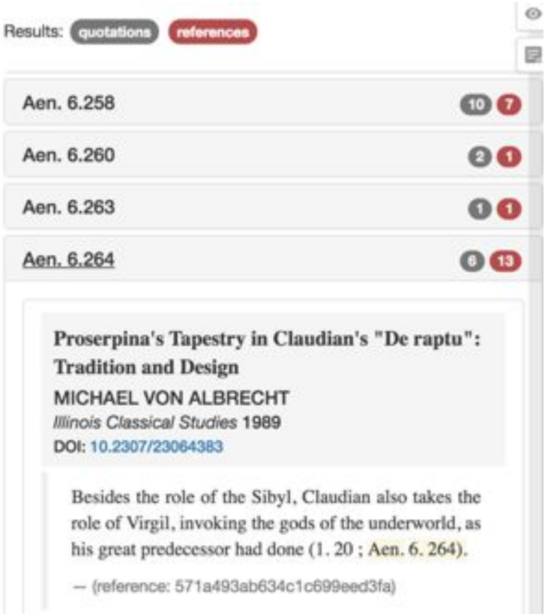
nunc animis opus, Aenea, nunc pectore firmo."

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

Aen. 6.258	10	7
Aen. 6.260	2	1
Aen. 6.263	1	1
Aen. 6.264	6	13

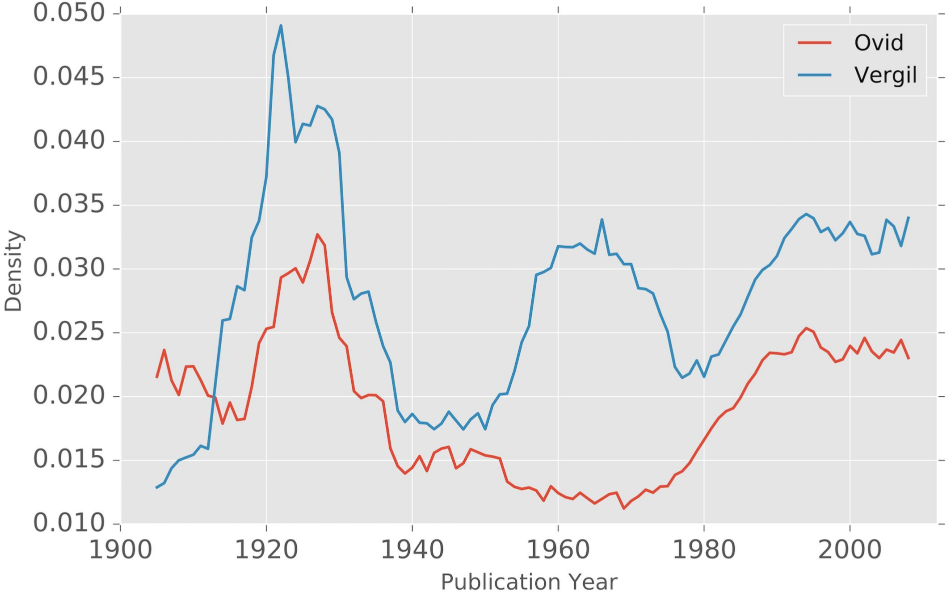
Proserpina's Tapestry in Claudian's "De raptu": Tradition and Design
MICHAEL VON ALBRECHT
Illinois Classical Studies 1989
DOI: 10.2307/23064383

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.*

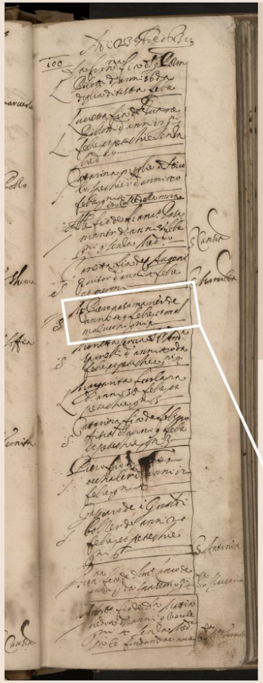
Example of ADA: Epidemics in Venice

Analysis of death records in Venice during the 1630-31 plague outbreak.

Paper:
https://www.nature.com/articles/s41598-020-74775-6

Der Doctor Schna- -bel von Rom
Vos Creditis, als eine fabel,
quod ſcribitur vom Doctor Schnabel,
der fugit die Contagion
et autert ſeinen Lohn darvon.
Cadavera ſucht er zu friſten,
gleich wie der Corvus auf der Miſten.
Ah Credite, ſihet nicht dort hin,
dann Romæ regnat die Peſtin.
Quis non deberet ſehr erſchrecken,
für ſeiner Virgul oder ſtecken,
qua loquitur, als wär er ſtumm.
und deütet ſein Conſilium.
Wie mancher Credit ohne zweiffel
das ihn tentir ein ſchwartzen Teüffl,
Marſupium heiſt ſeine Höll,
und aurum die geholte ſeel.
I. Columbina, ad vivum delineavit.
Paulus Fürſt. Excudit
Kleidüng wider den Tod zü Rom. Anno 1656.
Alſo gehen die Doctores Medici daher zü Rom, wann ſie die, an der Peſt erkranckte Per-
ſonen beſüchen, ſie zü curiren ünd fragen, ſich vor dem Gifft zü ſichern, ein langes Kleid von ge-
wärtem Tüch ihr Angeſicht iſt verlarvt, für den Augen haben ſie groſſe Crÿſtalline Brillen, vor der
Naſen einen langen Schnabel voll wolriechender Specereÿ, in der Hände, welche mit Handſchühen
wol verſehen iſt, eine lange Rüthe ünd darmit deüten ſie, was man thün, ünd gebraüche ſoll.

Data collection: Death records



Patriarchal Archive of Venice

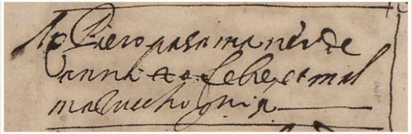
N° of registers: **89**
Records extracted*: **43,088**
Sant'Eufemia case study: **1785**
*(deaths counted)

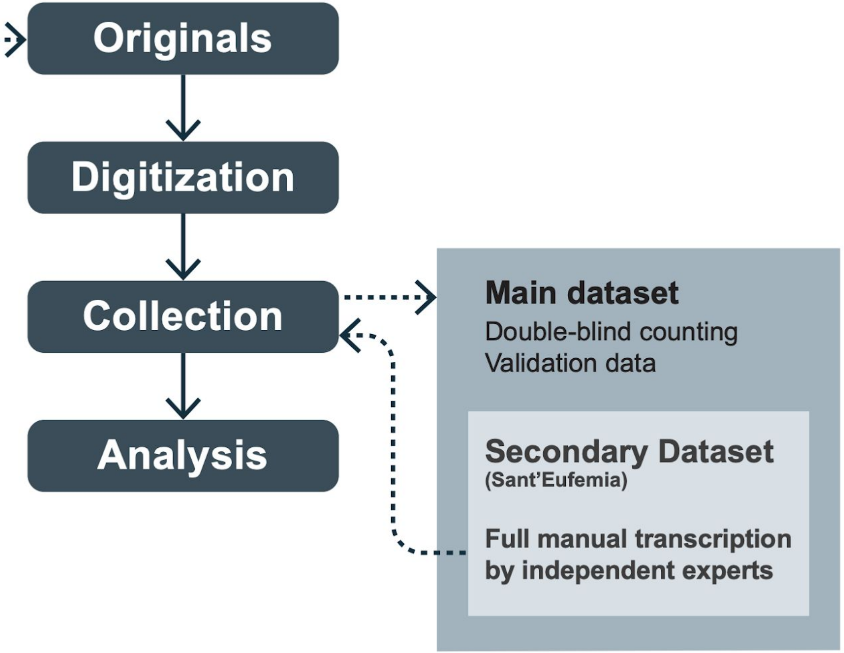
State Archive of Venice

N° of registers: **11**

Record structure

1 book per parish/year
1 death per line
Ordered chronologically





Originals

Digitization

Collection

Analysis

Main dataset

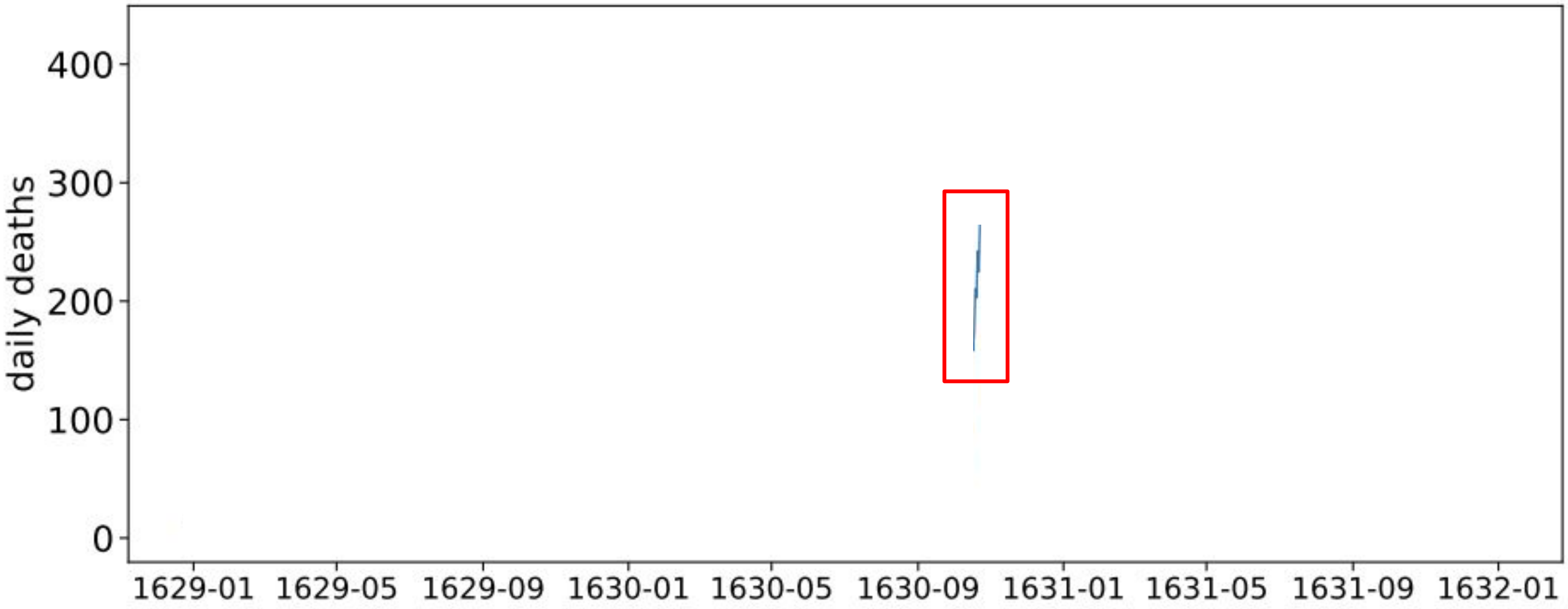
Double-blind counting

Validation data

Secondary Dataset
(Sant'Eufemia)

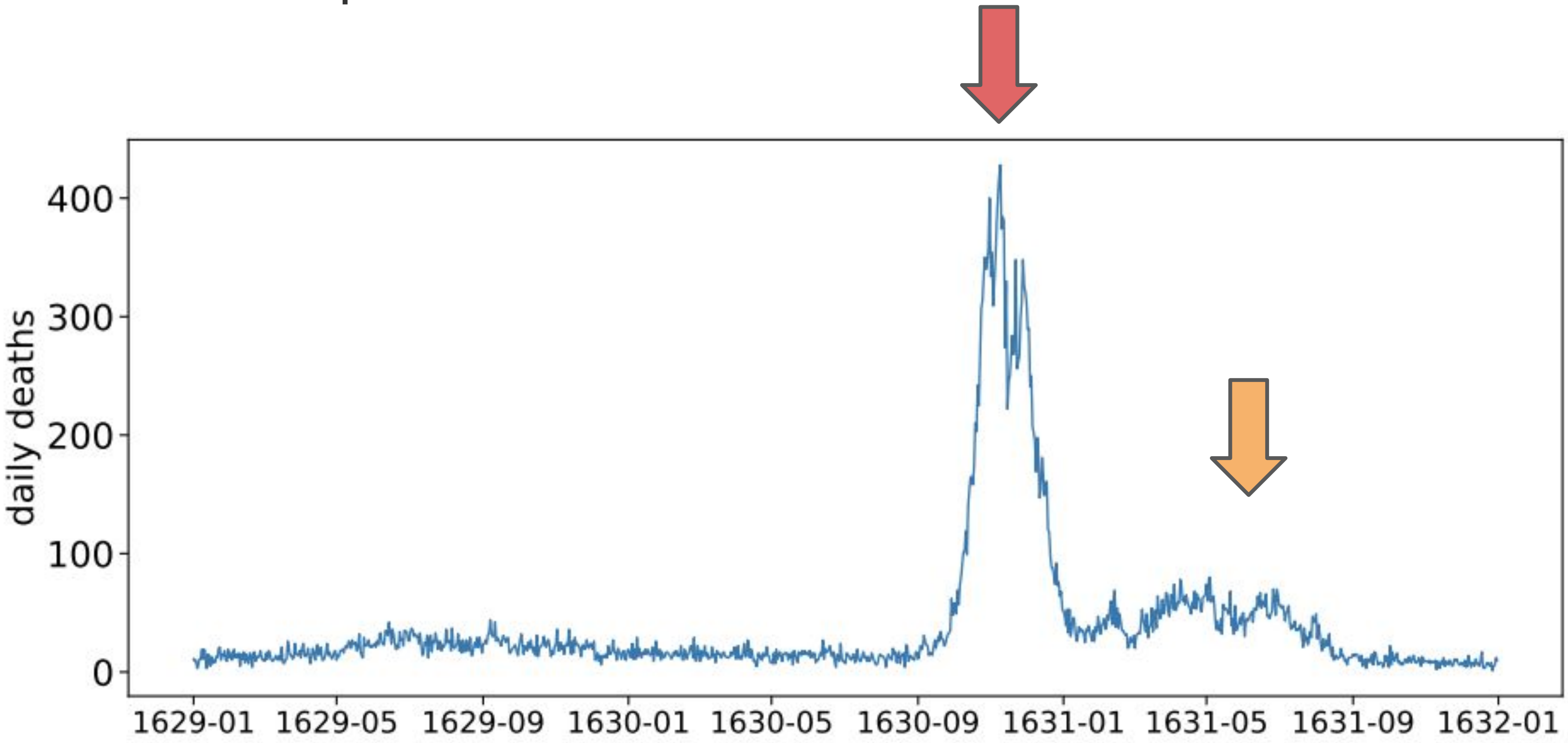
Full manual transcription by independent experts

Previous art

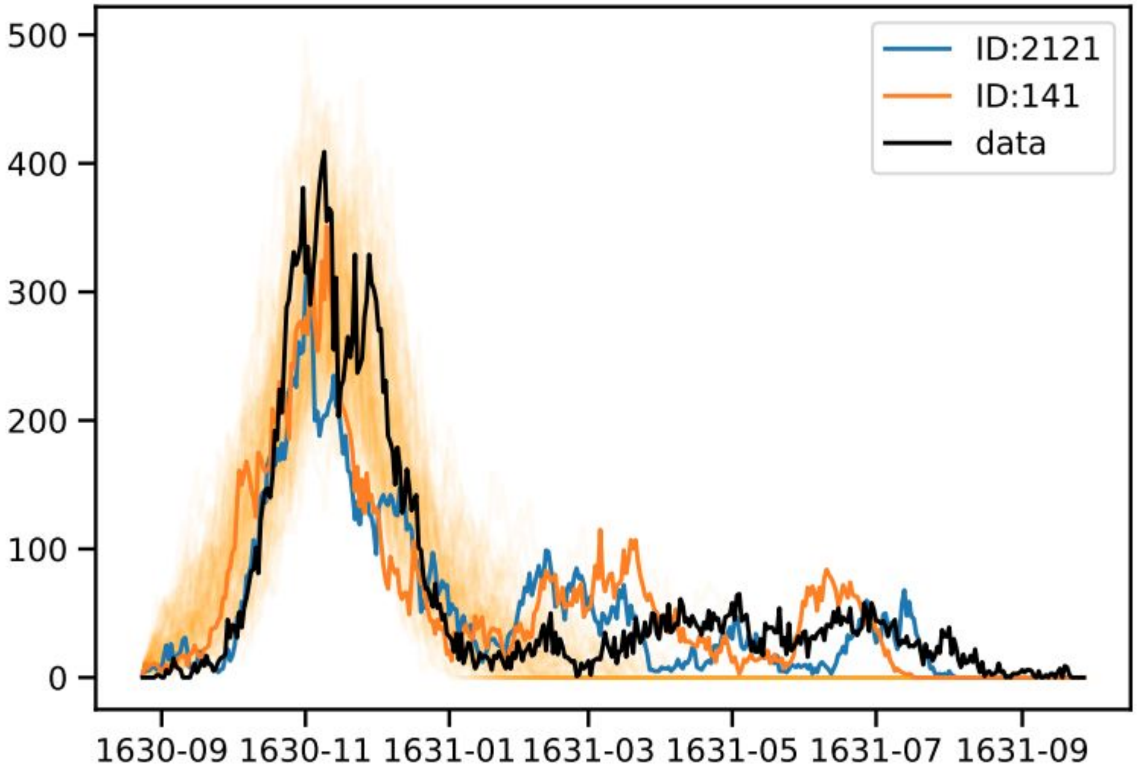


Ell, S. R. Three days in October of 1630: detailed examination of mortality during an early modern plague epidemic in Venice. *Rev. Infect. Dis.* 11(1), 128–139 (1989).

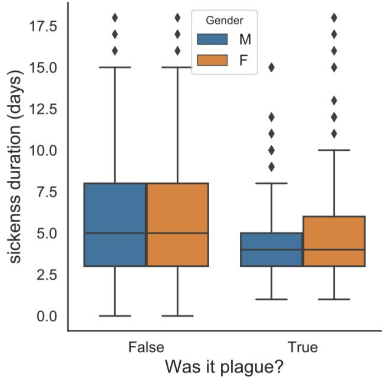
Our results: A double peak



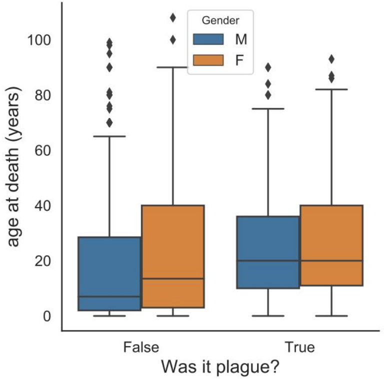
Asking new questions: outbreak dynamics



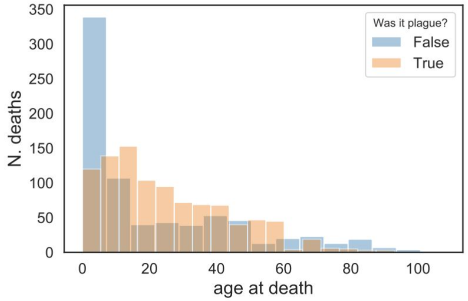
Data analysis: Age and gender



(a)



(b)

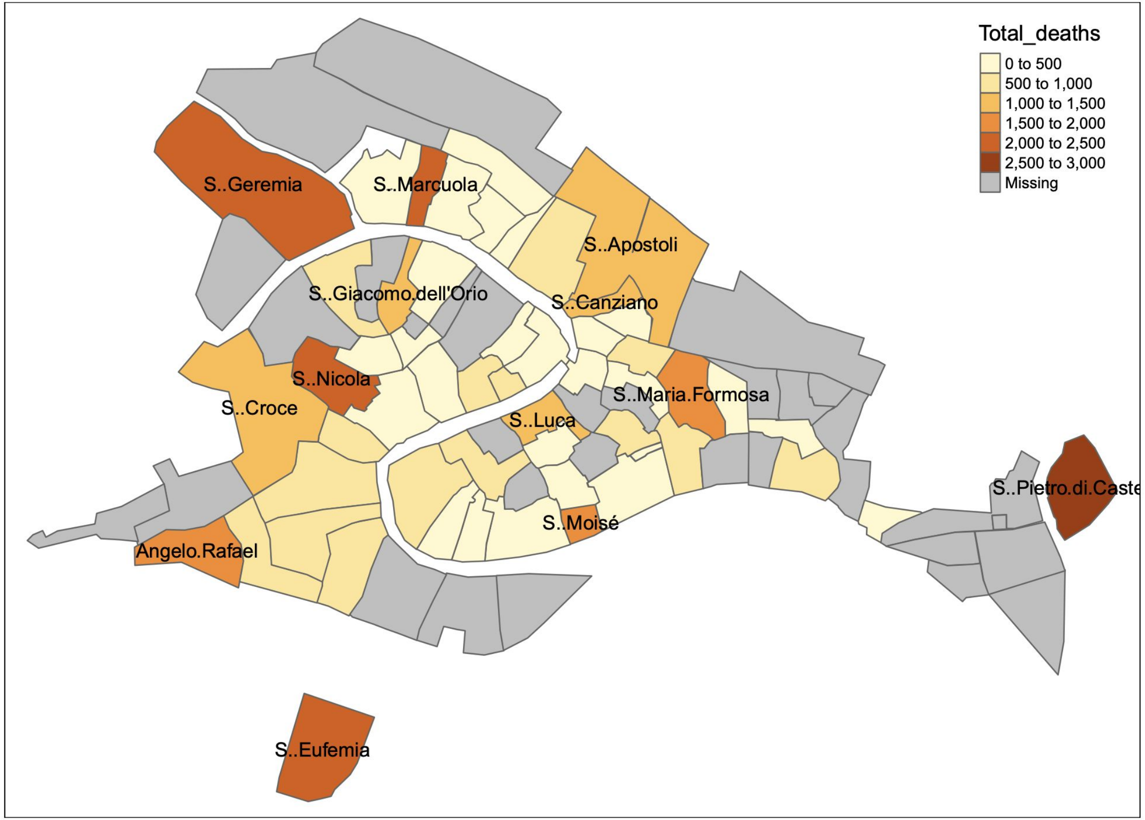


(c)

Age cohort (years old)	Plague	Not plague	Aggregated
Infant (0-2)	3.38%	23.52%	12.16%
Child (3-13)	32.01%	29.2%	30.42%
Young (14-23)	20.95%	8.74%	15.63%
Adult (24-44)	29.2%	16.45%	23.64%
Old (45+)	17.28%	19.28%	18.15%

(d)

Data analysis: Location (parishes)

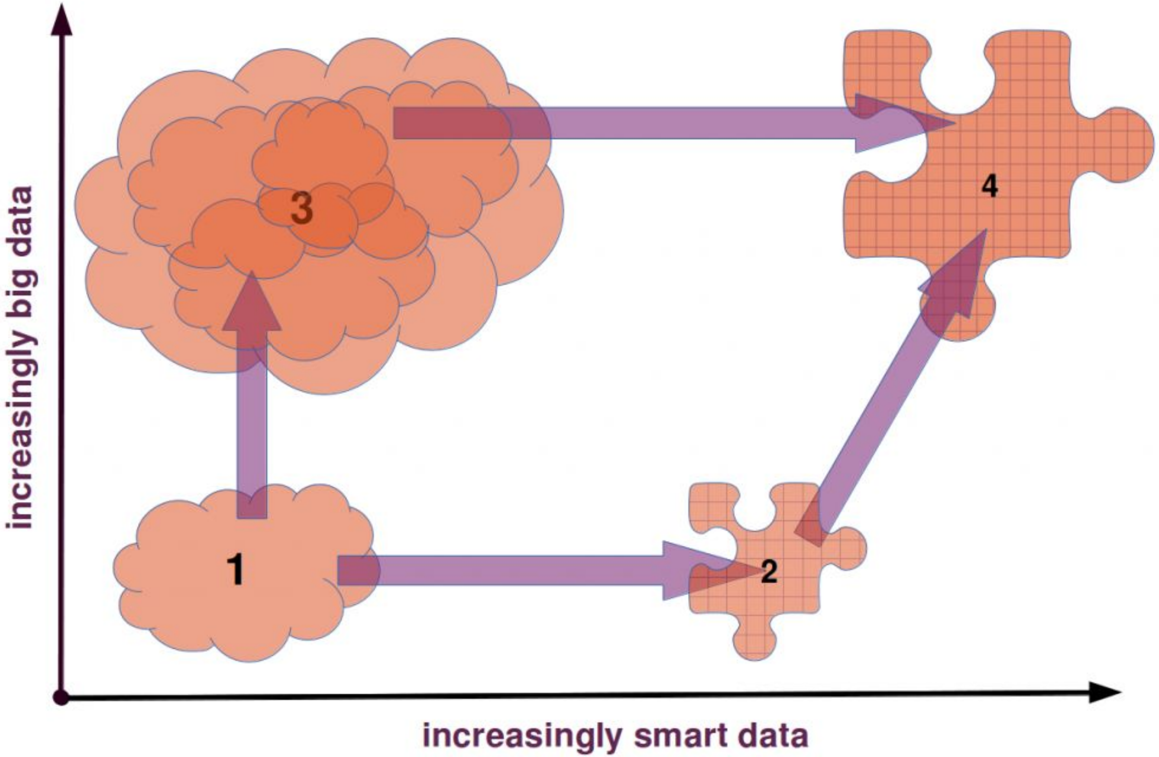


Take homes

- Huge death toll is confirmed.
- No gender or location effects, but significant age effects.
- Epidemic-wise, the double peak is the most interesting finding and it it not to be found in previous literature: how to explain it is still an open question.

We rarely can just use one source. No matter how rich a source is, there is always more to be considered (public officers, personal recounts, lived environment such as burials, etc.).

Conclusion



http://journalofdigitalhumanities.org/wp-content/uploads/2013/11/story-of-smart-big-data-1024x673.png

Tools and materials

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

Slides are on Google Drive: shared GDrive folder

Collaborative question area: shared GDoc

Pre-school survey: https://forms.gle/RQjt5VssHabQKh2U7

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

Twitter: #DHOxSS2022 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.