Data Provenance

Melanie Ganz
Cyril Pernet
Robert Oostenveld

What is Data Provenance?

1 - what do you think it means?
2 - what is it useful for?
3 - can you think of different applications?

Go to https://tinyurl.com/bdfcujbv and spend 5 minutes to think about this

Definition of the word “Provenance”?

According to the Oxford English Dictionary:

“The fact of coming from some particular source or quarter; origin, derivation.”

Or

“The history of the ownership of a work of art or an antique, used as a guide to authenticity or quality; a documented record of this.”

[https://www.oed.com/view/Entry/153408?redirectedFrom=provenance#eid]

What is Data Provenance?

Data provenance is a record of the origins and handling of data.

It typically includes information about who created the data, when it was created, how it was created, and how it has been modified.

If you get a directory listing with `ls -al`, you are seeing some metadata.

Data provenance can be used to help track down errors and to verify the results of data analysis.

It can also be used to comply with regulations and standards for data management.

Potential Uses for Data Provenance?

For research:

- Resolving discrepancies in data sets
- Can improve both the accuracy (i.e., precision) and completeness (i.e., recall) of information
- Helps tracking down errors and to verify the results of data analysis.
- It can also be used to comply with regulations and standards for data management.

But also:

- Tracking down the source of a cyberattack
- Verifying business transactions

W3C PROV specifications (more usage)

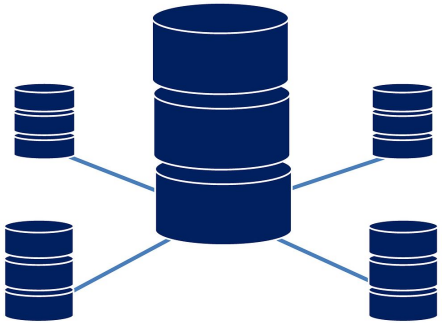
→ The World Wide Web Consortium is the main international standards organization for the World Wide Web.

Provenance, a form of structured metadata designed to record the origin or source of information

- allows checking where data/code/information is coming from (trust issue),
- allows integration of multiple sources
- tracks contributors, authors, curators (attribution)

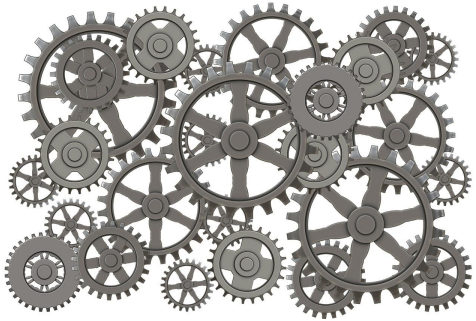
Provenance in Machine Learning

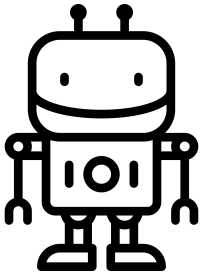
Data collection



Data curation and assemblage

Data preprocessing and feature selection





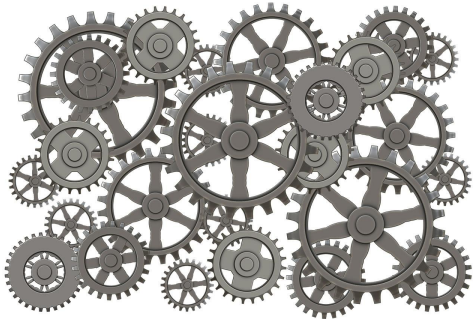
Model ready
→ prediction, ranking, etc …

Provenance in Machine Learning

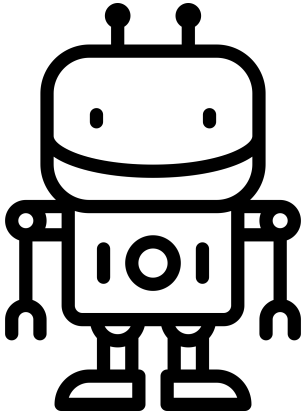
Who collected the data?
For which purpose?
Do you have the rights to use the data for the machine model usage?

Who curated the data and assembled them?
Is there any bias in the data?

Provenance in Machine Learning



How was preprocessing performed
Balance between sources? bias ?
How were missing values handled?
Which features were used?



Who is generating prediction with model?
Who is using the prediction?

Example you could face in your work

You want to replicate and extend some work (1) reusing experimental data (2) reusing existing code

- does not have to be from the web, could be the student who was in the lab last year …

Lucky for you, that student attended a lecture on data management it's all organized :-) … unfortunately he/she did not attend the lecture on provenance

A case study

- Provenance in practice! Go to GitHub and locate the Jupyter notebook: https://github.com/CPernet/ReproducibleQuantitativeDataScience/blob/main/provenance/ProvenanceInPractice.ipynb

- Try and redo the “scientific” analysis presented to you there!

Provenance summary

Provenance: all that is needed is a little documentation

→ metadata file (.md .json .txt)

→ Provenance graphs are rarely complete

→ Some provenance is better than no provenance

→ Management of base data is out of scope

https://www.mitre.org/sites/default/files/publications/practical-provenance-guide-MP100128.pdf