

Data Provenance and Reproducibility

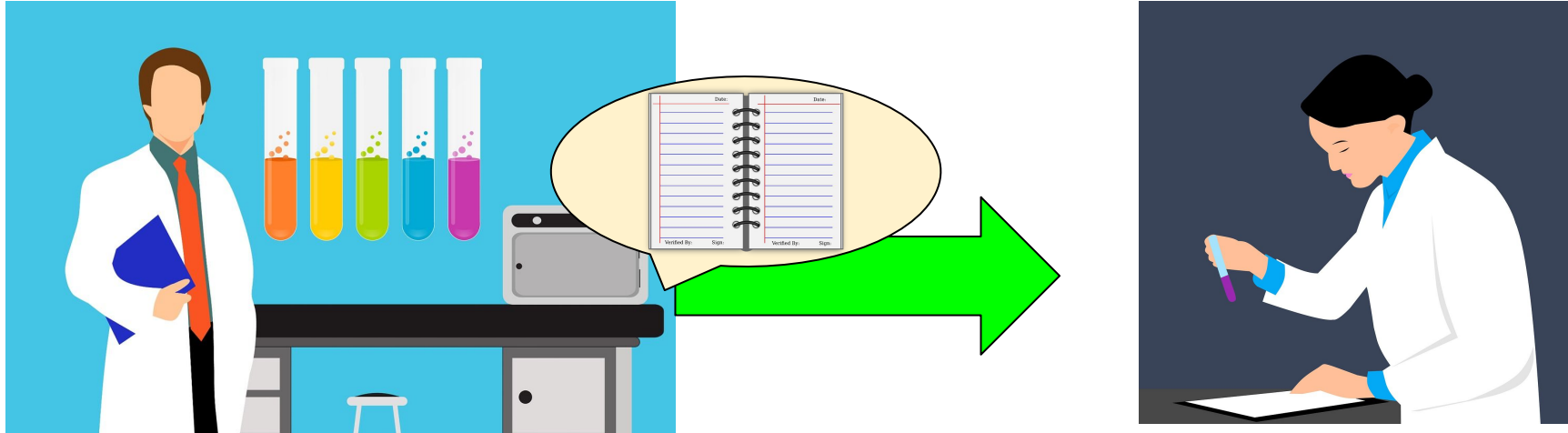
Tim Schmidt

Syeda Hiba Ahmad



- Reproducibility
 - Importance
 - Crisis
 - In ML
 - In Companies
 - What to do about it?
- Data Provenance
 - Importance
 - Challenges
 - Current Standards
 - Different Approaches
 - Provenance Taxonomy
 - Examples
- Future Work

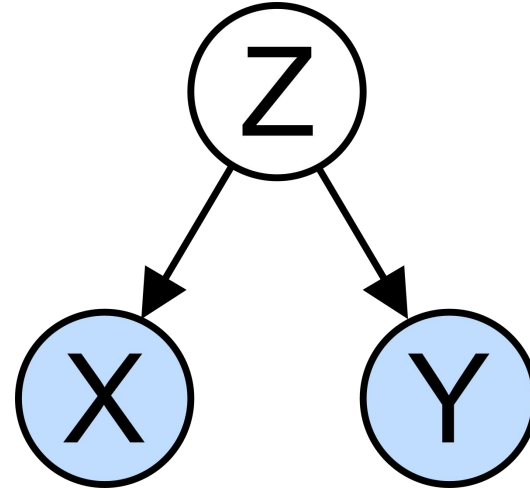
Reproducibility



“A measure of whether results can be attained by a different research team, using the same methods.”

Importance of Reproducibility

- shows that there are no confounding variables
 - protects against fraud
 - human error



Reproducibility Crisis

A crisis of repeatability:

“Of these 100 studies, just 68 reproductions provided [...] results that matched the original findings.”

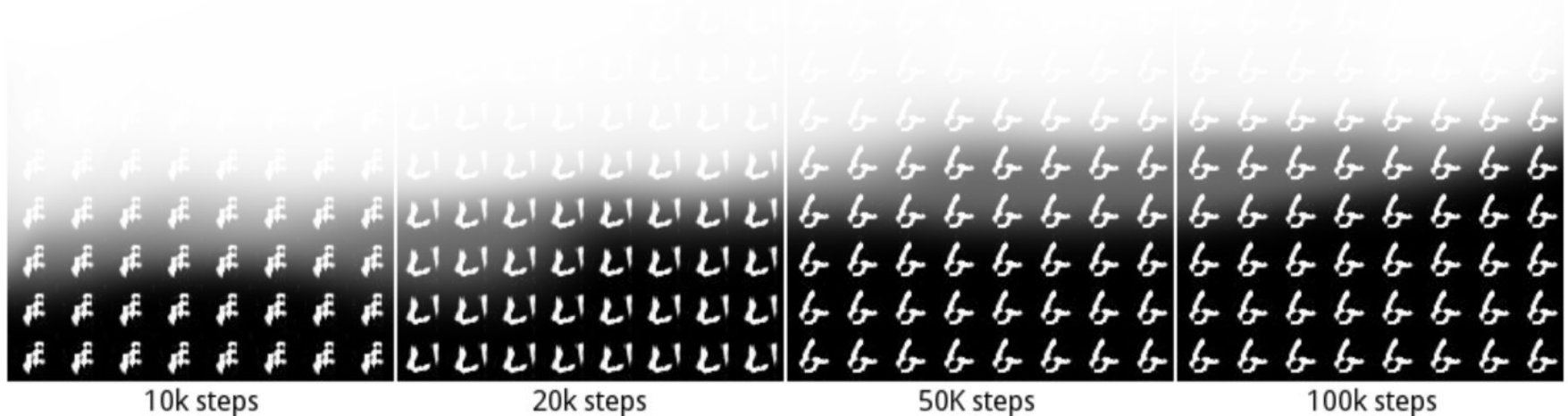
A crisis of description:

Of 400 algorithms [...] He found that only 6% [...] shared the algorithm's code. Only a third shared the data they tested their algorithms on, and just half shared “pseudocode”.



Reproducibility in ML

- ML swaps heuristics for blackbox for better results
- Randomness between runs (need to fix meta parameters)
- Need to store data



Reproducibility in Companies

- If a researcher drops out, somebody else should be able to step in
- Cope with changed requirements or platforms
 - time saver in the long run
- It's a lot more risky to try different variation

**CAN'T STOP.
TOO BUSY!!**



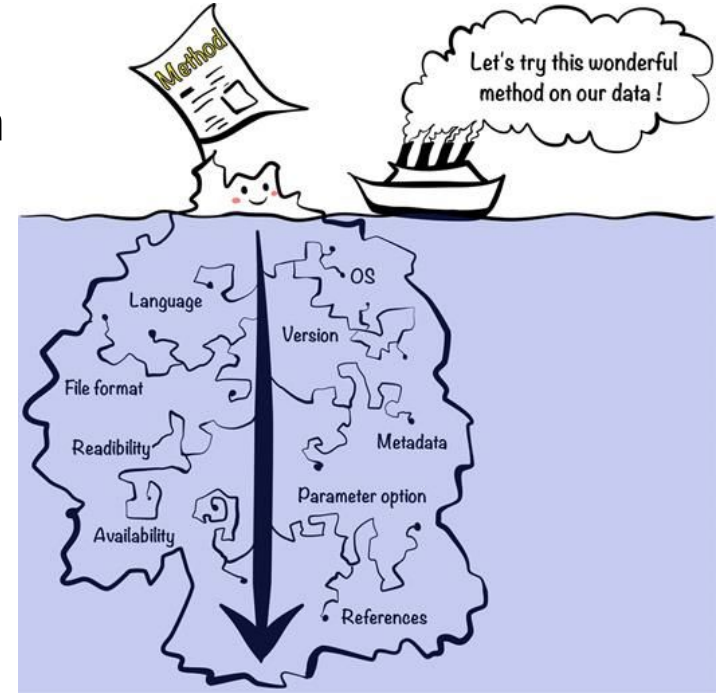
What to do about it?

Science:

- Provide all the info: code, data, description

Companies:

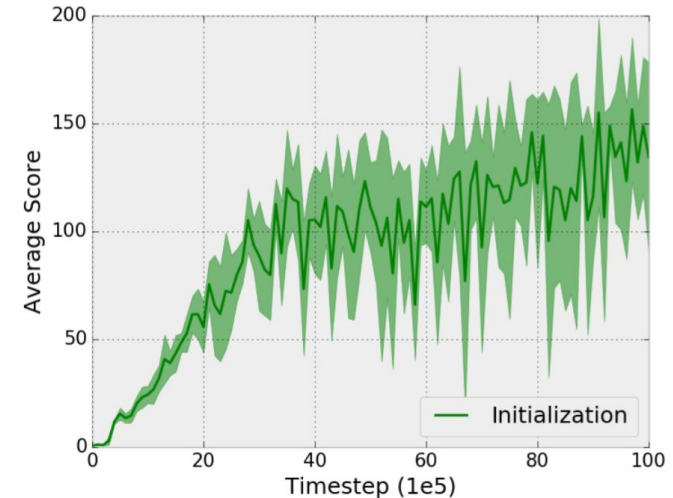
- No wrong incentives
- Repetition of experiments
- Keep the team educated



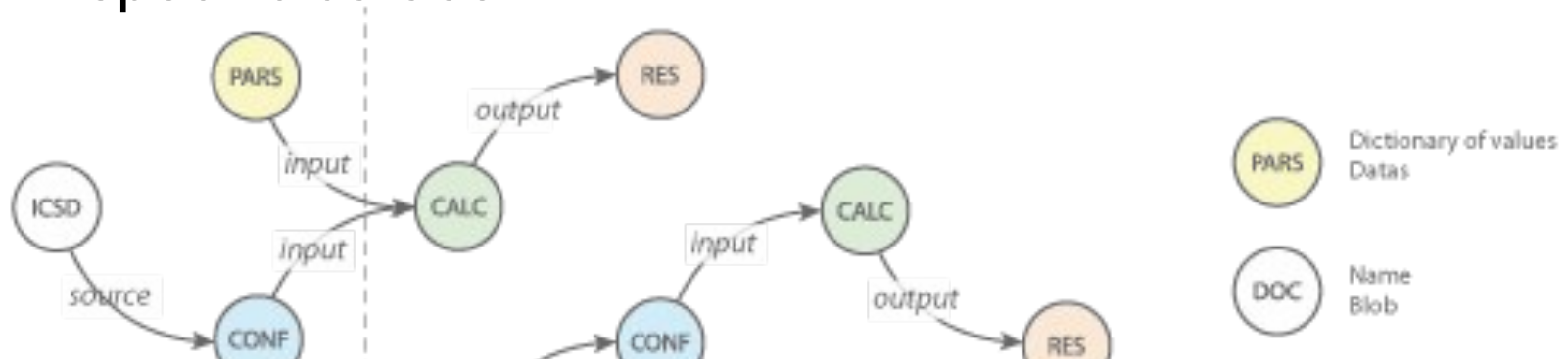
What to do about it?

ML:

- Features causing non-deterministic results are disabled
 - Versioning of models
- (Jupyter Notebooks are a nightmare)
- **Data versioning**



“Data Provenance is the documentation of data in sufficient detail to allow reproducibility of a specific dataset.”

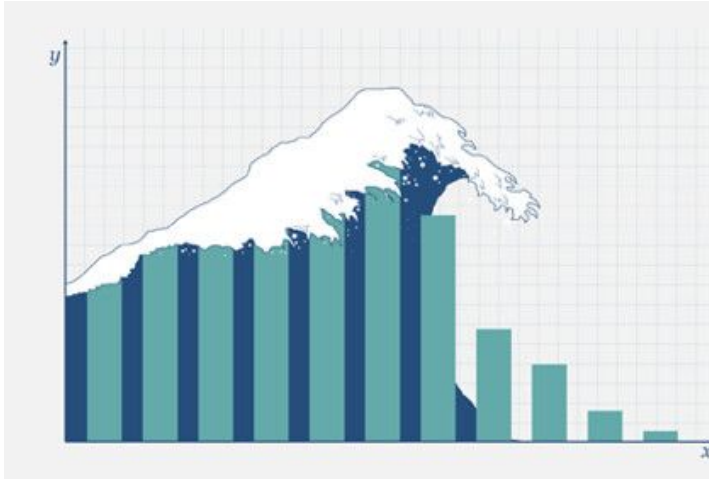


Importance

- Compliance (e.g. DSGVO)
- Necessary if accused of fraud
- Prevents manual errors
- Changes in underlying database
- Data needs to be trustworthy
- Root case analysis



Challenges



- Large data sizes
- Provenance overhead
- Archiving (vs changes)



Archiving (vs changes)

Key

Added
Deleted
Modified

Versions of data

Day 1

name	gender	age	primary interest
Anna	F	35	swimming
Bob	M	28	weight lifting

Day 24

name	gender	age	primary interest
Anna	F	35	swimming
Bob	M	28	weight lifting
Carol	F	20	kickboxing

Day 40

name	gender	age	primary interest
Anna	F	35	swimming
Carol	F	20	kickboxing

Day 44

name	gender	age	primary interest
Anna	F	35	swimming
Carol	F	20	weight lifting
Dan	M	38	cardio

Storing only the changes

Day 1

name	gender	age	primary interest
Anna	F	35	swimming
Bob	M	28	weight lifting

Day 24

name	gender	age	primary interest
Carol	F	20	kickboxing

Day 40

name	gender	age	primary interest
Bob	M	28	weight lifting

Day 45

name	gender	age	primary interest
Carol	F	20	weight lifting
Dan	M	38	cardio

Archiving the data

Day 1

[1-today]	name	gender	age	primary interest
	Anna	F	35	swimming
	Bob	M	28	weight lifting

Day 24

[1-today]	name	gender	age	primary interest
	Anna	F	35	swimming
	Bob	M	28	weight lifting
[24-today]	Carol	F	20	kickboxing

Day 40

[1-today]	name	gender	age	primary interest
	Anna	F	35	swimming
[1-39]	Bob	M	28	weight lifting
[24-today]	Carol	F	20	kickboxing

Day 45

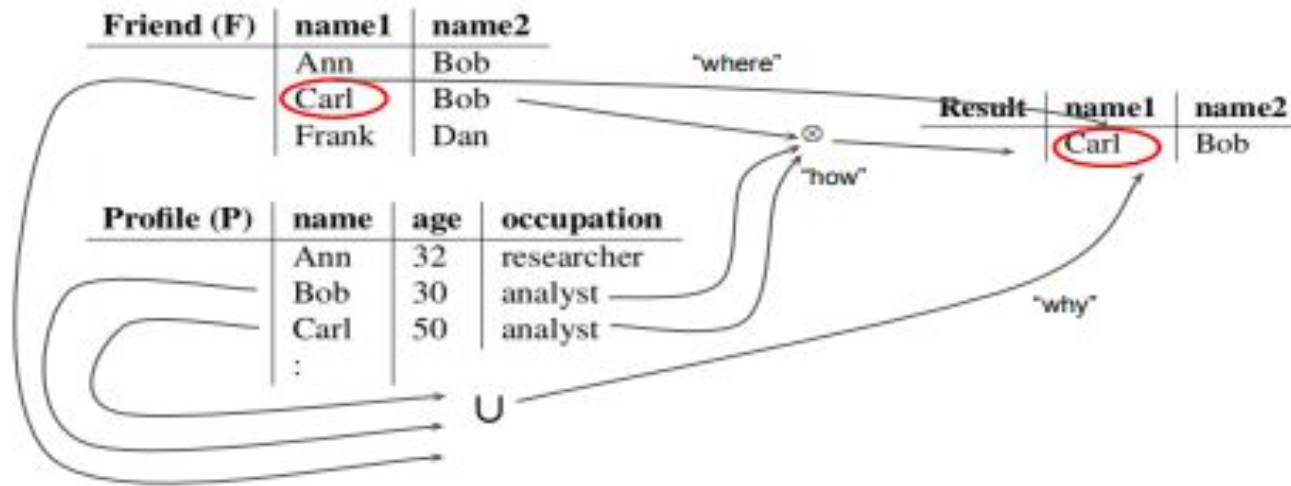
[1-today]	name	gender	age	primary interest
	Anna	F	35	swimming
[1-39]	Bob	M	28	weight lifting
[24-today]	Carol	F	20	[24-44] Kickboxing [45-today] weight lifting
[45-today]	Dan	M	38	cardio

“The PROV Family of Documents defines a **model**,
corresponding serializations and other **supporting definitions** to enable the inter-operable interchange of provenance information in heterogeneous environments such as the Web.”

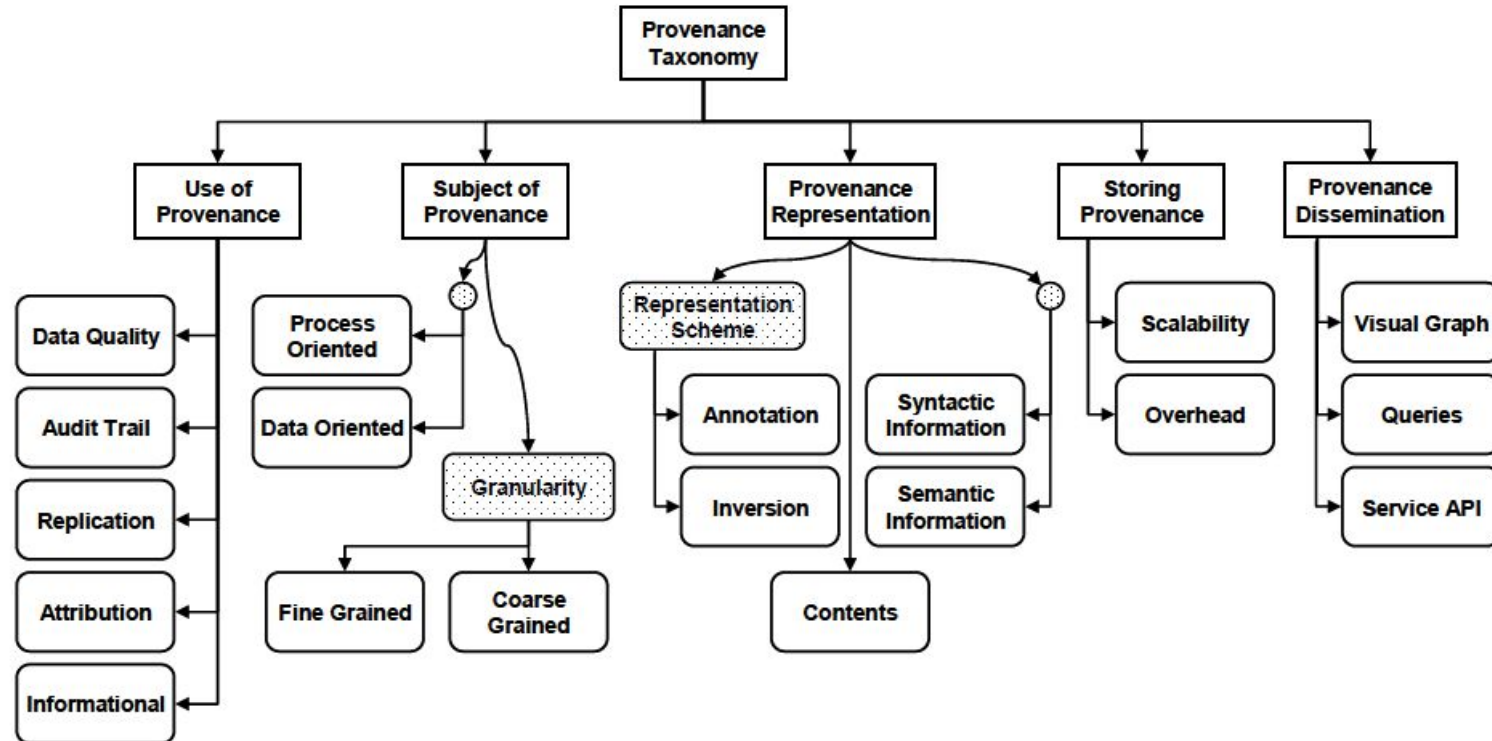
<https://www.w3.org/TR/prov-overview/>

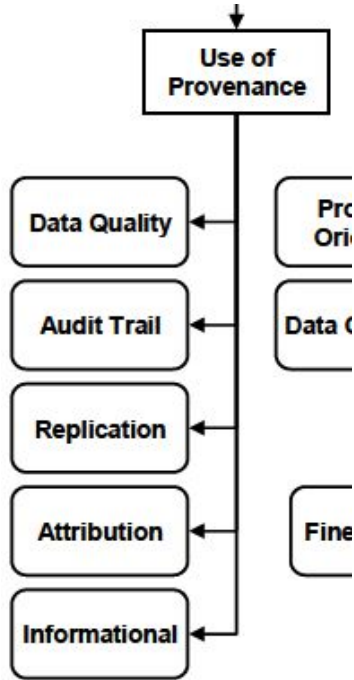
Different Approaches

- Where-Provenance (Original Source), Why-Provenance (Contributing Source) and How-Provenance (Transformation)



Provenance Taxonomy

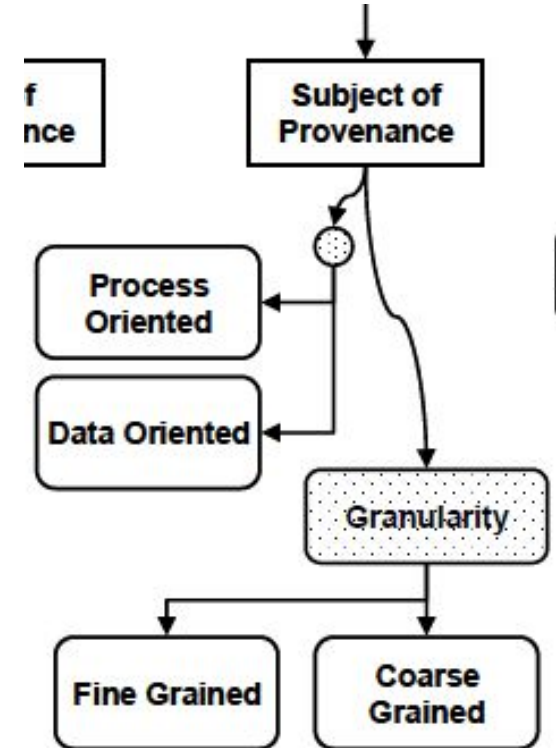


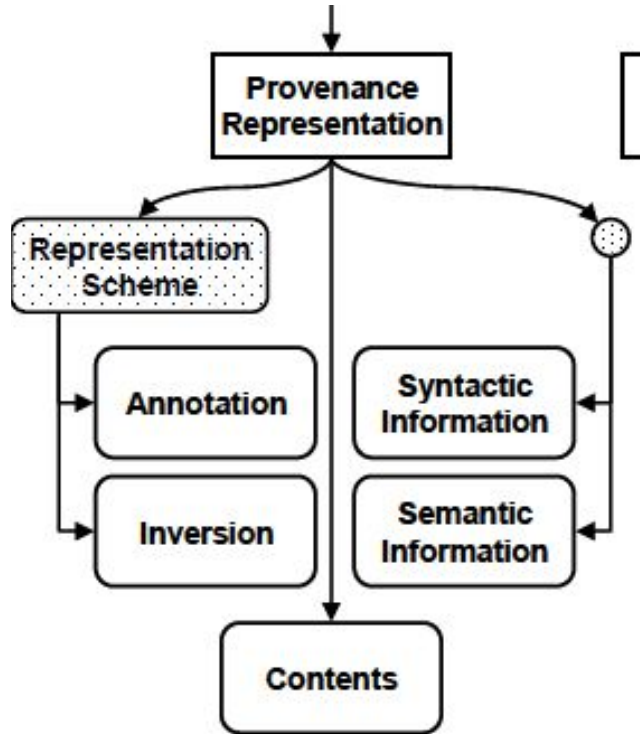


- **Data Quality:** Level of detail and error
- **Audit Trail:** Process with which data is produced
- **Replication:** Availability of similar sources
- **Attribution:** Pedigree (ownership)
- **Informational:** Metadata (descriptive)

Subject of Provenance

- **Data Oriented (Explicit) Model:** Metadata from source data
- **Process Oriented (Indirect):** Metadata from process inputs and outputs
- **Granularity:** Level of detail

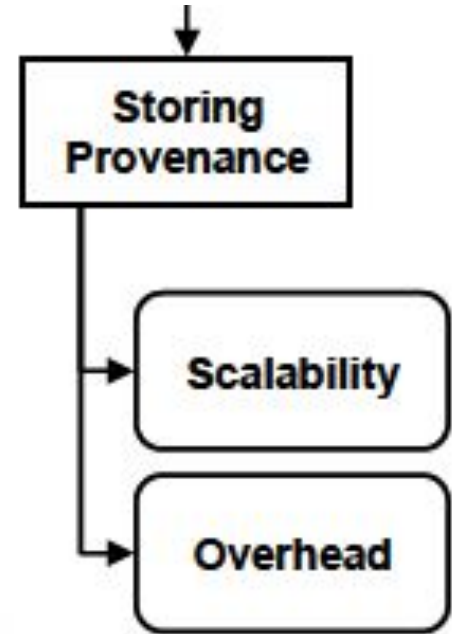




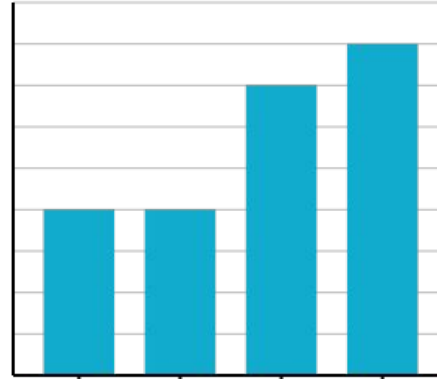
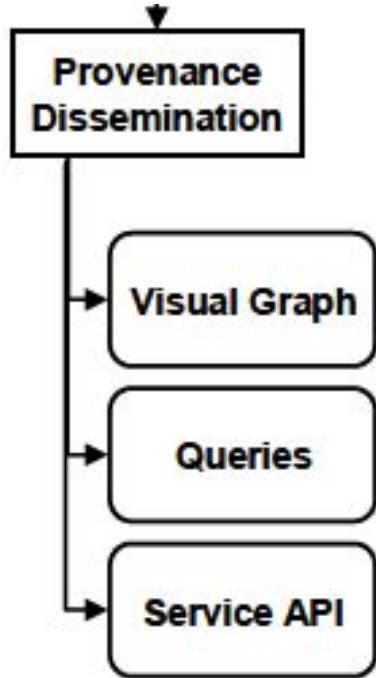
- **Annotation:** Descriptions about source data and processes
- **Inversion:** Reverse-engineering queries
- **Contents:** Of annotation and inversion methods
- **Syntactic Information:** The form in which data is stored
- **Semantic Information:** The meaning given to the data

Storing Provenance

- **Tightly Coupled:** Close relation with data
- **Loosely Coupled:** Slight relation with data
- **Scalability:** Growth of system
- **Overhead:** Management costs



Provenance Dissemination



Resultset 1

id	propname	dtype	intval	floatvalue	strvalue
1	eventValue	0	2	NULL	NULL
1	CustomEmailMessage	2	NULL	NULL	NULL
1	CustomEmailSubject	2	NULL	NULL	NULL
2	eventValue	0	1	NULL	NULL
2	CustomEmailMessage	2	NULL	NULL	NULL
2	CustomEmailSubject	2	NULL	NULL	NULL
3	eventValue	0	2	NULL	NULL
3	CustomEmailMessage	2	NULL	NULL	NULL
3	CustomEmailSubject	2	NULL	NULL	NULL
4	setpointA	1	NULL	50	NULL
4	eventValue	1	NULL	50184	NULL
4	CustomEmailMessage	2	NULL	NULL	NULL
4	CustomEmailSubject	2	NULL	NULL	NULL
5	eventValue	0	0	NULL	NULL
5	CustomEmailMessage	2	NULL	NULL	NULL
5	CustomEmailSubject	2	NULL	NULL	NULL
6	eventValue	0	2	NULL	NULL
6	CustomEmailMessage	2	NULL	NULL	NULL
6	CustomEmailSubject	2	NULL	NULL	NULL

Result Data

1000 rows fetched in 0.037s

Auto Refresh Edit Apply Discard



Practical Examples

- DVC (for small projects)
 - Git expansion for data
- Pachyderm (for bigger projects)
 - Runs on Kubernetes



[More Informations](#)

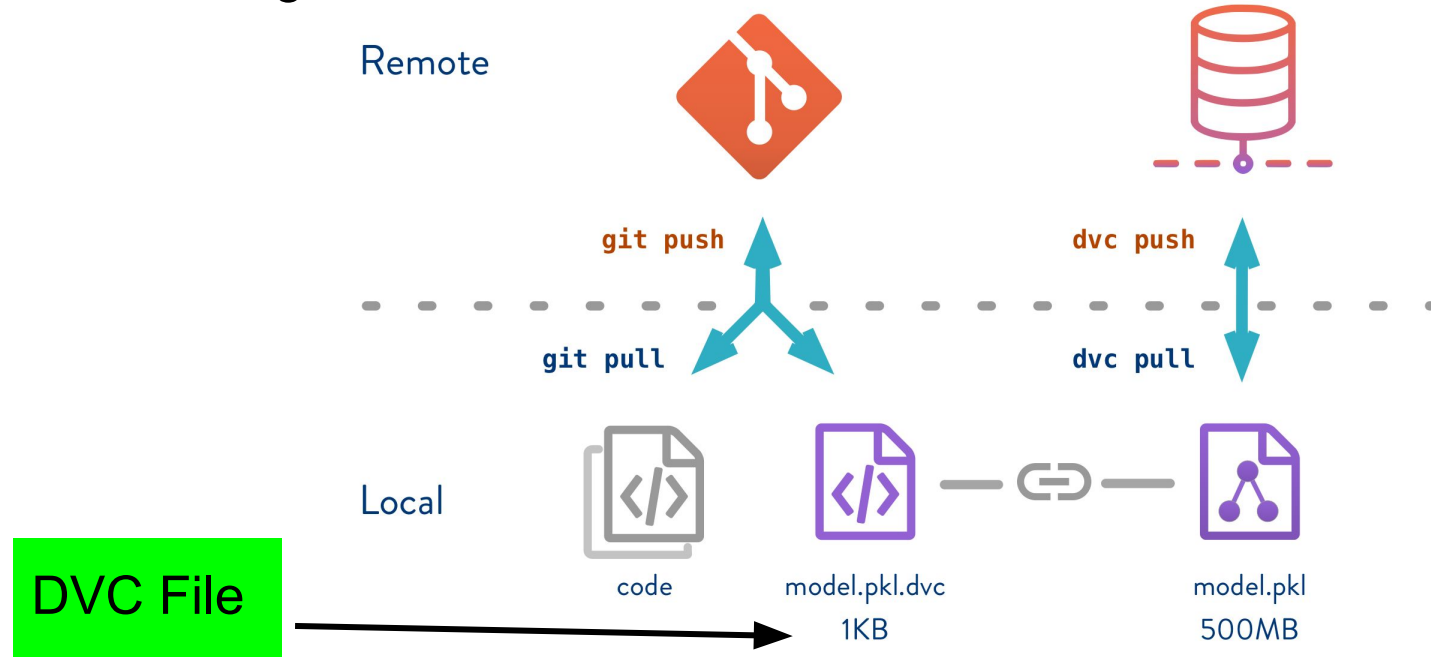
Data Versioning Control



- Data Versioning:

Github, Gitlab, any Git Server

S3, Azure, Google Cloud, SSH

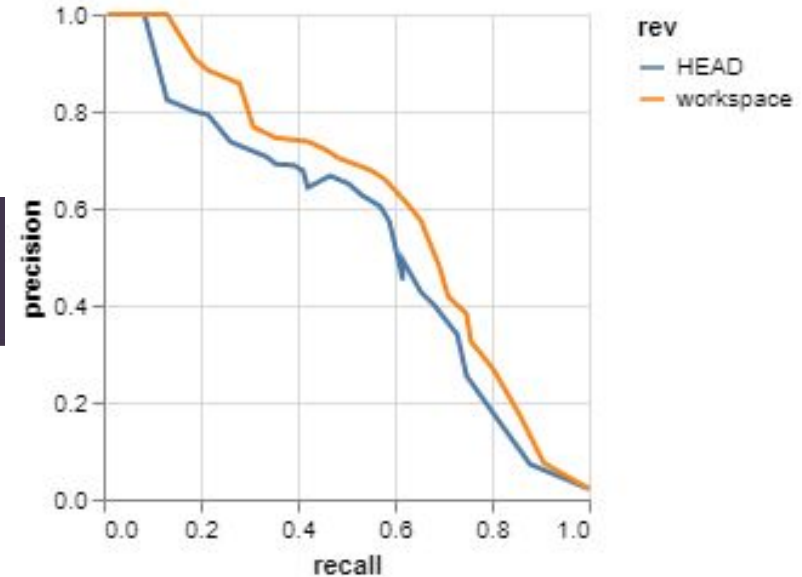


Data Pipelines & Experiments



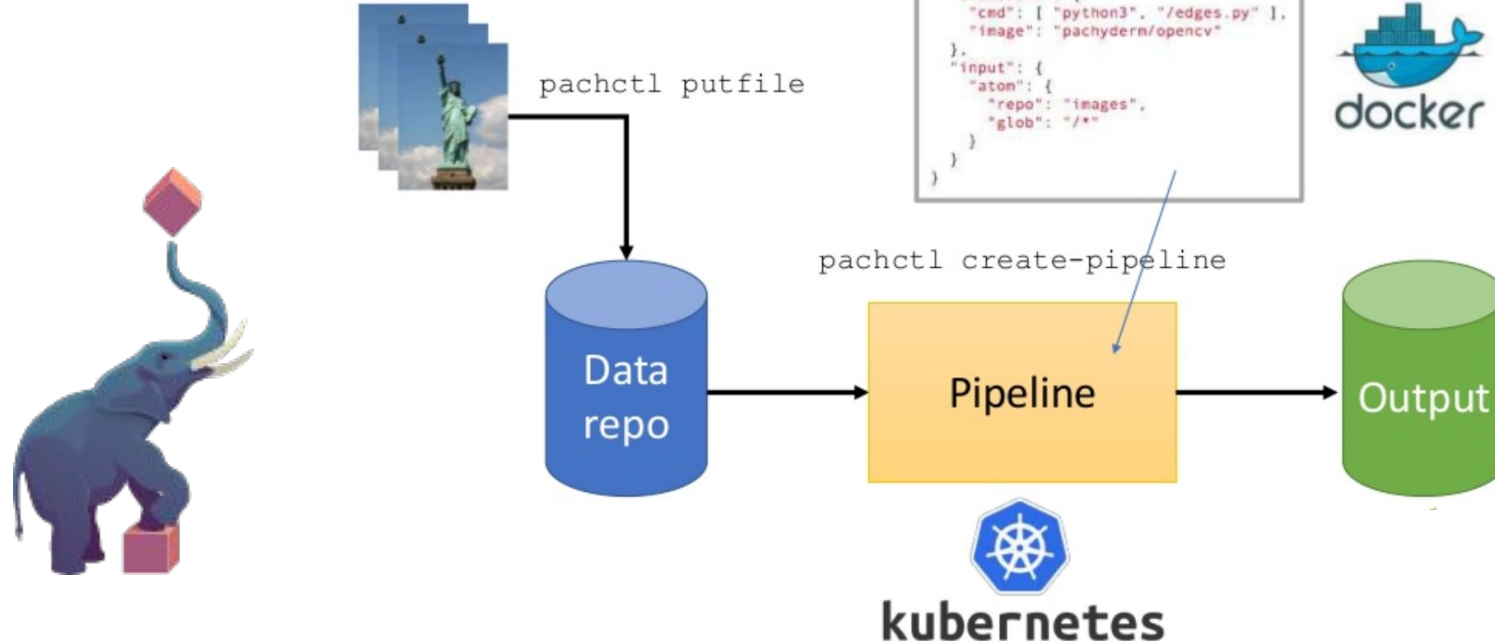
- Link processing steps together
- Store versions and parameters
- Compare versions

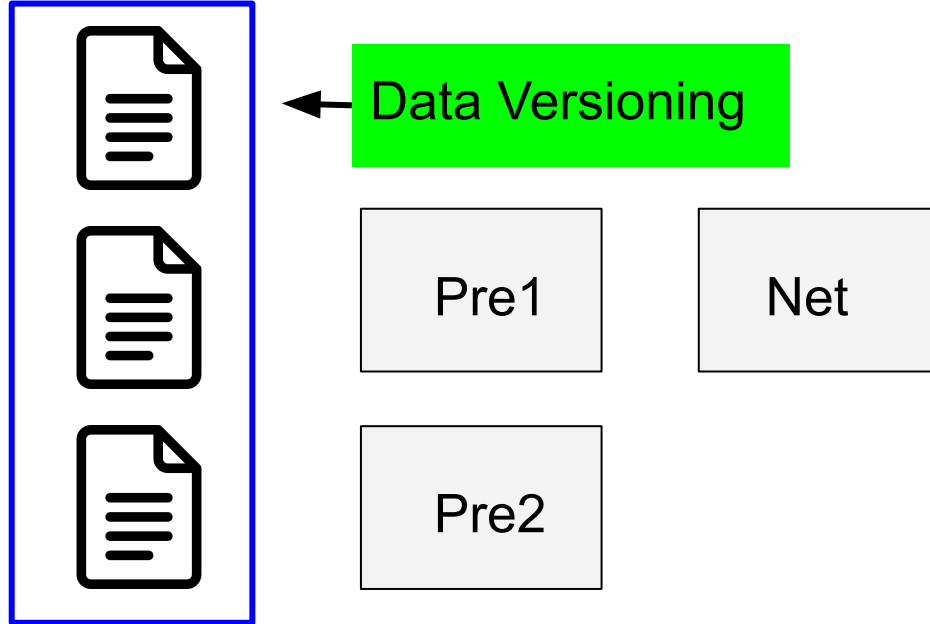
```
$ git add *.dvc auc.metric data/.gitignore  
$ git commit -m "Add tag prediction pipeline (6 stages)"
```

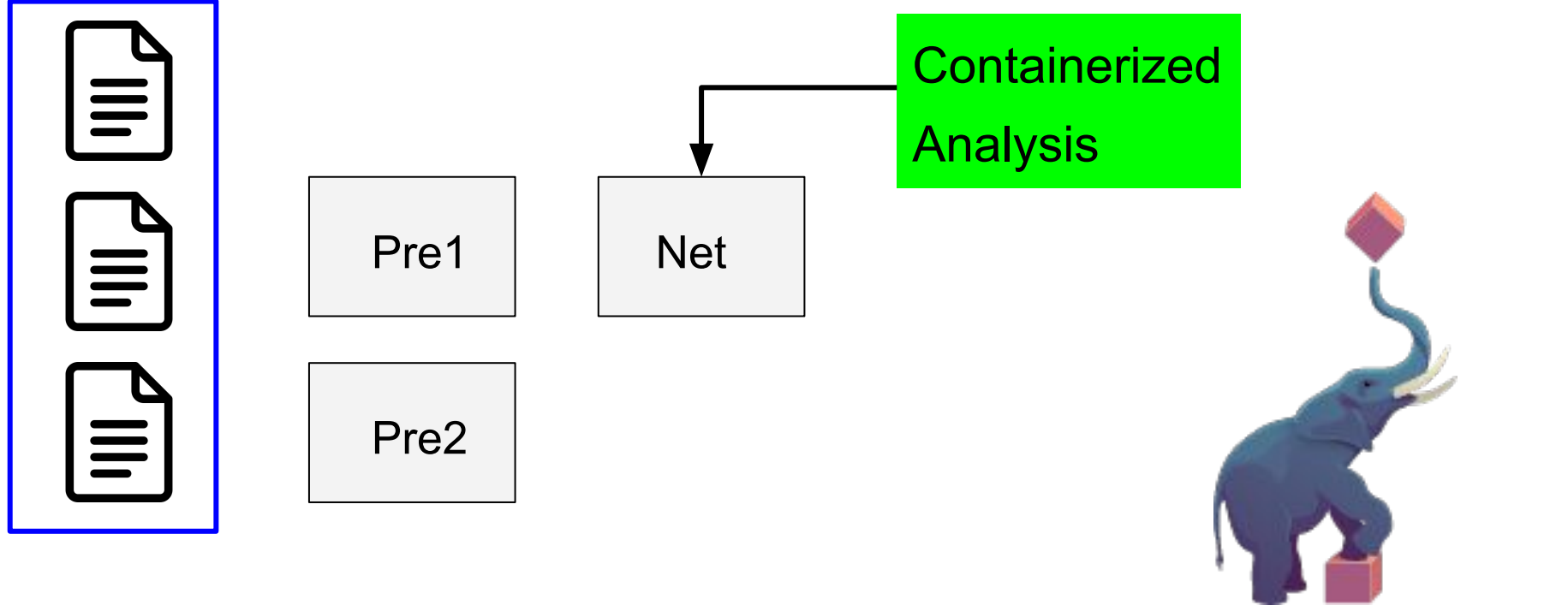


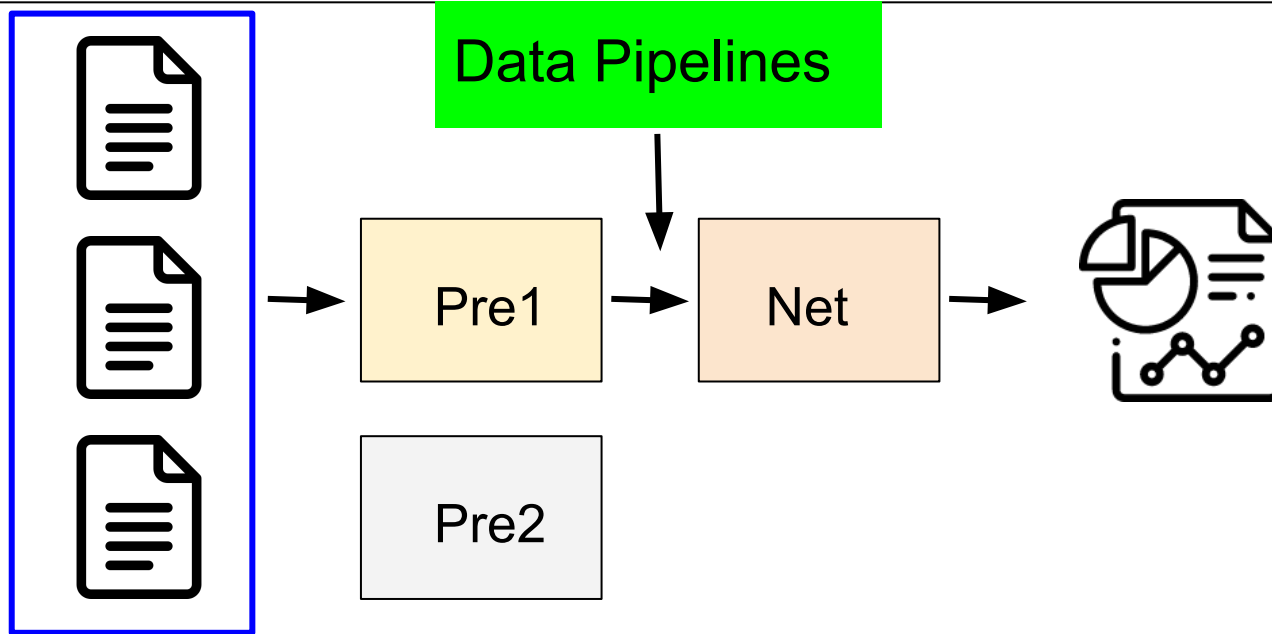
Pachyderm

- Based on Kubernetes and Docker

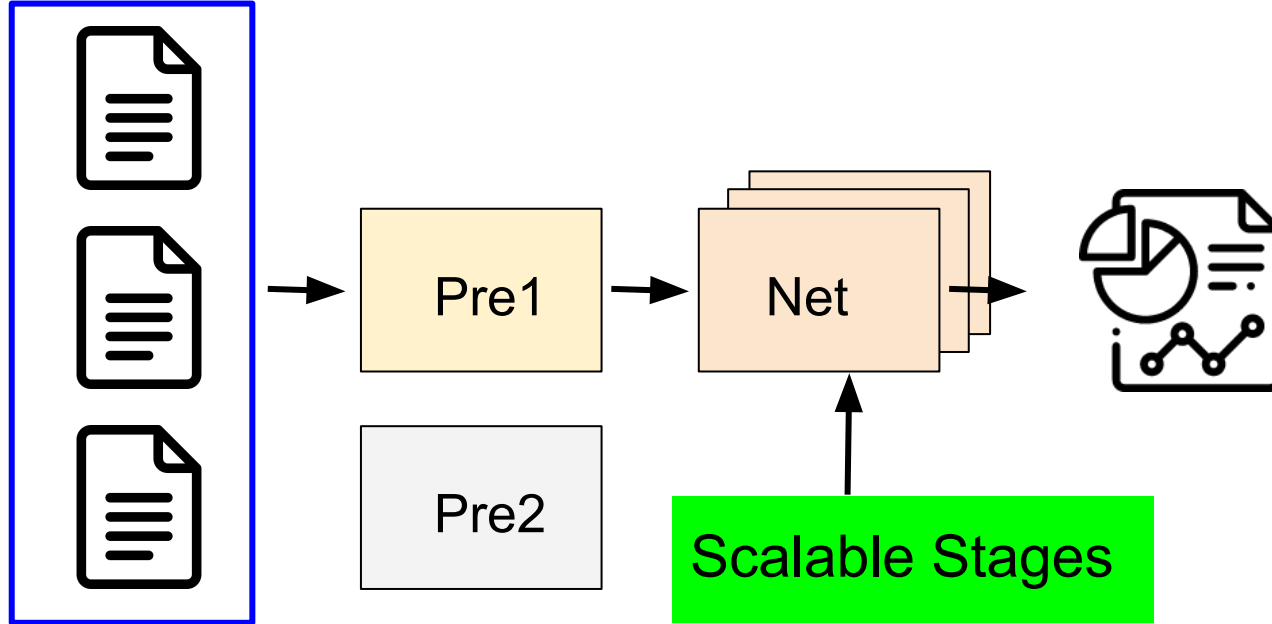


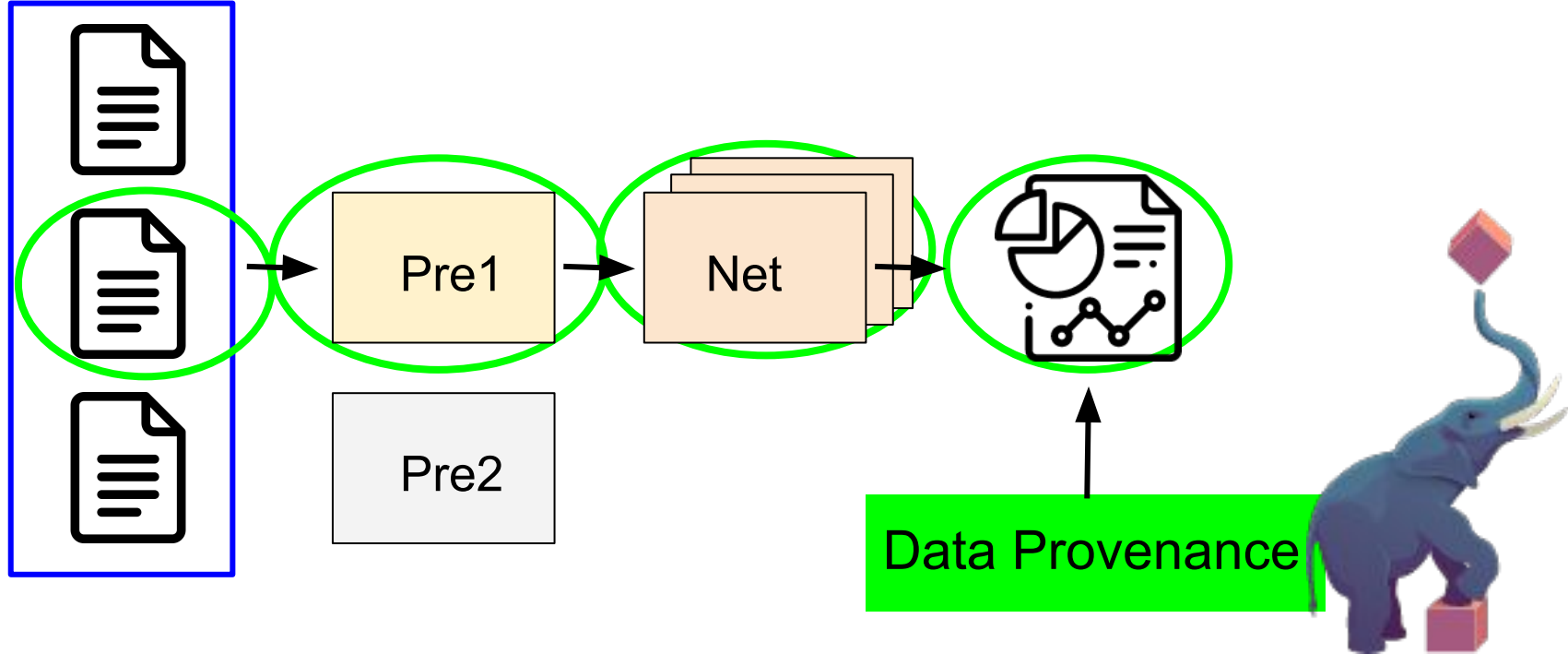




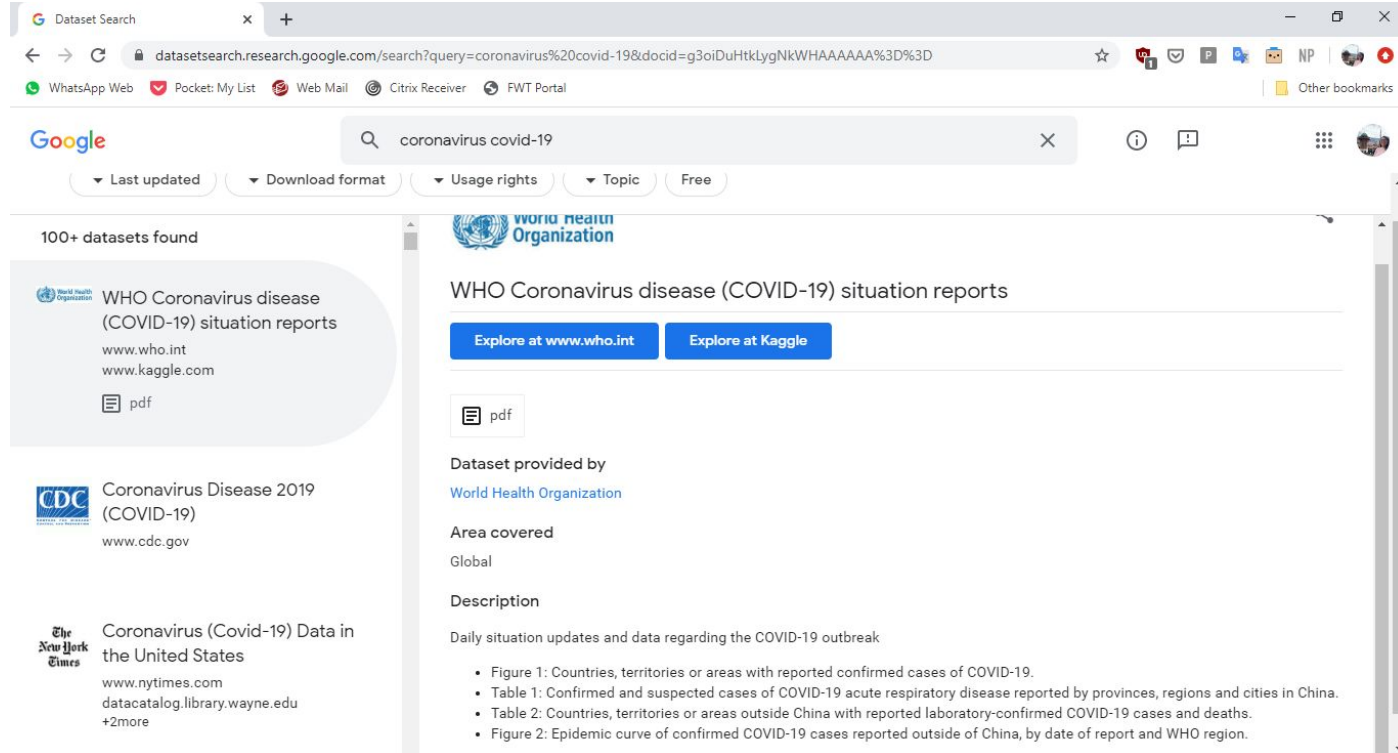


Pachyderm





Google Dataset Search (GOODS)



The screenshot shows a web browser window with the Google Dataset Search interface. The search query is 'coronavirus covid-19'. The results page displays '100+ datasets found'. On the left sidebar, three datasets are listed: 'WHO Coronavirus disease (COVID-19) situation reports' from the World Health Organization, 'Coronavirus Disease 2019 (COVID-19)' from the CDC, and 'Coronavirus (Covid-19) Data in the United States' from The New York Times. The main content area shows the details for the WHO dataset, including a PDF icon, a link to 'Explore at www.who.int', and a description of daily situation updates and data regarding the COVID-19 outbreak. The description includes a list of four figures: Figure 1 (countries with confirmed cases), Table 1 (confirmed and suspected cases in China), Table 2 (cases and deaths outside China), and Figure 2 (epidemic curve outside China).

Google Dataset Search

datasetsearch.research.google.com/search?query=coronavirus%20covid-19&docid=g3oiDuHtkLygNkWHAAAAA%3D%3D

Google coronavirus covid-19

Last updated Download format Usage rights Topic Free

100+ datasets found

WHO Coronavirus disease (COVID-19) situation reports
www.who.int
www.kaggle.com
pdf

CDC Coronavirus Disease 2019 (COVID-19)
www.cdc.gov

The New York Times Coronavirus (Covid-19) Data in the United States
www.nytimes.com
datacatalog.library.wayne.edu
+2more

WHO Coronavirus disease (COVID-19) situation reports

Explore at www.who.int Explore at Kaggle

pdf

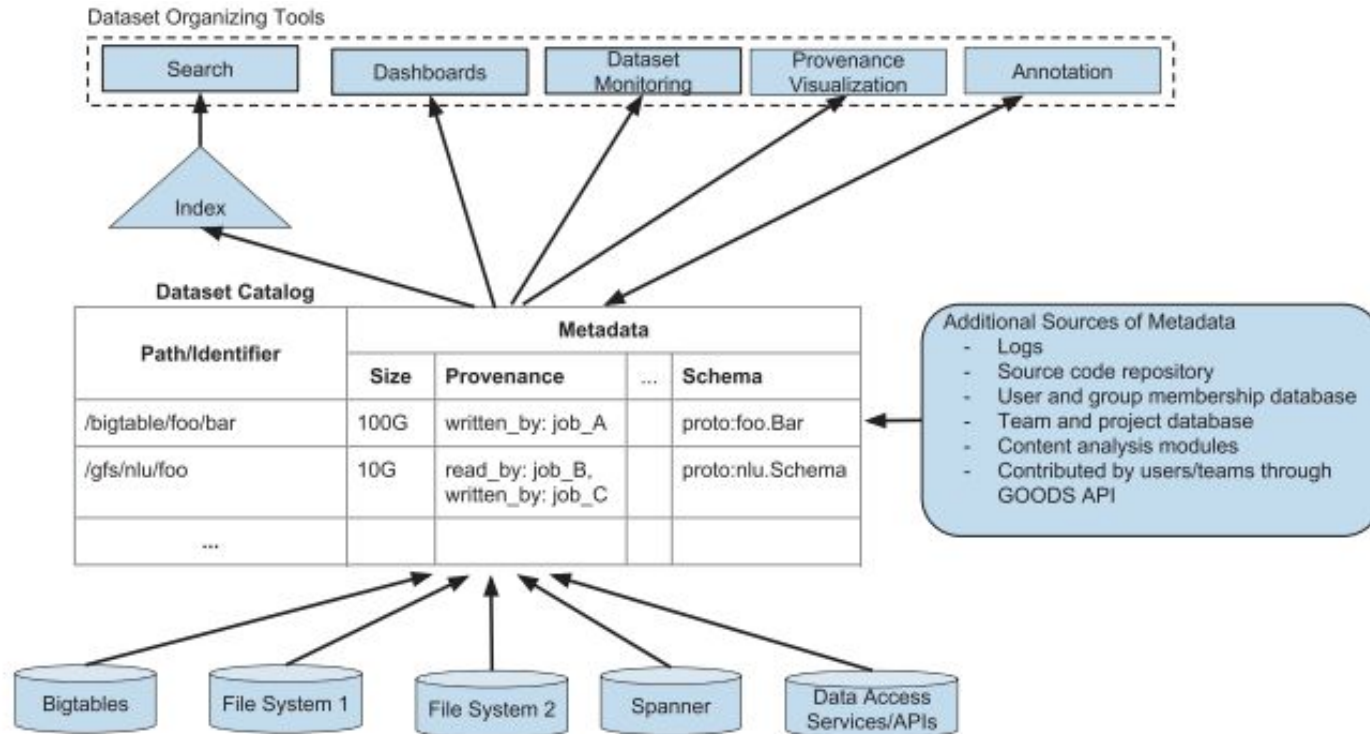
Dataset provided by
World Health Organization

Area covered
Global

Description
Daily situation updates and data regarding the COVID-19 outbreak

- Figure 1: Countries, territories or areas with reported confirmed cases of COVID-19.
- Table 1: Confirmed and suspected cases of COVID-19 acute respiratory disease reported by provinces, regions and cities in China.
- Table 2: Countries, territories or areas outside China with reported laboratory-confirmed COVID-19 cases and deaths.
- Figure 2: Epidemic curve of confirmed COVID-19 cases reported outside of China, by date of report and WHO region.

Google Dataset Search (GOODS)



What is Next?

- How to rank datasets
- How to identify important datasets
- Handling missing metadata
- More work on data semantics
- Data citation
- Environment information
- Applications in ML, social media, block chain, cybersecurity



Questions?

Thank you! :)

References

Reproducibility

- Jennifer Villa and Yoav Zimmerman. "Reproducibility in ML: why it matters and how to achieve it". May 25, 2018. Determined AI.
<https://determined.ai/blog/reproducibility-in-ml/>
- Ruairi J Mackenzie. "Repeatability vs. Reproducibility". Mar 25, 2019. Technology Networks.
<https://www.technologynetworks.com/informatics/articles/repeatability-vs-reproducibility-317157>
- Hutson, Matthew. "Artificial intelligence faces reproducibility crisis." Science. (2018): 725-726.
<https://science.sciencemag.org/content/359/6377/725.summary>
- Pascal Fecht. "Reproducibility in Machine Learning". Feb 26, 2019. Computer Science Blog.
<https://blog.mi.hdm-stuttgart.de/index.php/2019/02/26/reproducibility-in-ml/>
- Pete Warden. "The Machine Learning Reproducibility Crisis". Mar 19, 2018. Pete Warden's blog.
<https://petewarden.com/2018/03/19/the-machine-learning-reproducibility-crisis/>
- Stuart Buck. "Why Your Company Needs Reproducible Research". Apr 10, 2019. towards data science.
<https://towardsdatascience.com/why-your-company-needs-reproducible-research-d4a08f978d39>

References

Data Provenance

- Mike Brody. "Do You Know Where Your Data Came From?". Oct 30, 2018. DATAVERSITY.
<https://www.dataversity.net/know-data-came/#>
- Matthias Parbel. "Carsten Lux: "Data Lineage dreht den ETL-Prozess um"". Dec 28, 2018. heise Developer.
<https://m.heise.de/developer/artikel/Carsten-Lux-Data-Lineage-dreht-den-ETL-Prozess-um-4257607.html?seite=all>

Importance

- Cassie Kozyrkov. "All about data provenance". April 3, 2020. towards data science.
<https://towardsdatascience.com/how-to-work-with-someone-elses-data-6c45d467d7a2>
- Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. "Data provenance: Some basic issues." International Conference on Foundations of Software Technology and Theoretical Computer Science. Springer, Berlin, Heidelberg, 2000.
https://link.springer.com/chapter/10.1007/3-540-44450-5_6

References

Challenges

- Wang, Jianwu, et al. "Big data provenance: Challenges, state of the art and opportunities." 2015 IEEE International Conference on Big Data (Big Data). IEEE, 2015
<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7364047&tag=1>
- Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. "Data provenance: Some basic issues." International Conference on Foundations of Software Technology and Theoretical Computer Science. Springer, Berlin, Heidelberg, 2000
<http://db.cis.upenn.edu/DL/fsttcs.pdf>
- Buneman, Peter, and Wang-Chiew Tan. "Data provenance: What next?." ACM SIGMOD Record 47.3 (2019): 5-16
https://sigmodrecord.org/publications/sigmodRecord/1809/pdfs/03_Principles_Buneman.pdf

Different Approaches

- Cheney, James, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. Now Publishers Inc, 2009.
<http://homepages.inf.ed.ac.uk/jcheney/publications/provdbsurvey.pdf>

References

Provenance Taxonomy

- Simmhan, Yogesh L., Beth Plale, and Dennis Gannon. "A survey of data provenance techniques." Computer Science Department, Indiana University, Bloomington IN 47405 (2005): 69
<ftp://ftp.cs.indiana.edu/pub/techreports/TR618.pdf>

Google Dataset Search (GOODS)

- Halevy, Alon, et al. "Goods: Organizing google's datasets." Proceedings of the 2016 International Conference on Management of Data. 2016
<http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/45390.pdf>

Image References

Slide 4:

https://en.wikipedia.org/wiki/Confounding#/media/File:Simple_Confounding_Case.svg

Slide 6:

https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b

Slide 7:

<https://www.cristiesoftware.de/news/empirical-technical-debt/>

Slide 8:

<https://www.pinterest.de/pin/381891243402527019/>

Slide 9:

<https://openreview.net/pdf?id=S1e-OsZ4e7>

Slide 10:

<http://www.aiida.net/feature/data-provenance/>

Slide 11:

<https://www.earth.com/news/plant-roots-find-water/>

Image References

Slide 21:

<https://icons8.com/icons/set/>

Slide 22:

<https://dvc.org/>

<https://www.pachyderm.com/>

Slide 23:

<https://dvc.org/doc/use-cases/versioning-data-and-model-files>

Slide 24:

<https://dvc.org/doc/start/experiments#experiments>

Slide 25:

<https://www.slideshare.net/joshlk100/reproducible-data-science-review-of-pachyderm-data-version-control-and-git-lfs-tools>

Slide 26:

<https://fontawesome.com/icons/file?style=solid>

Slide 28:

https://www.flaticon.com/de/kostenloses-icon/analyse_944053?term=results&page=1&position=7