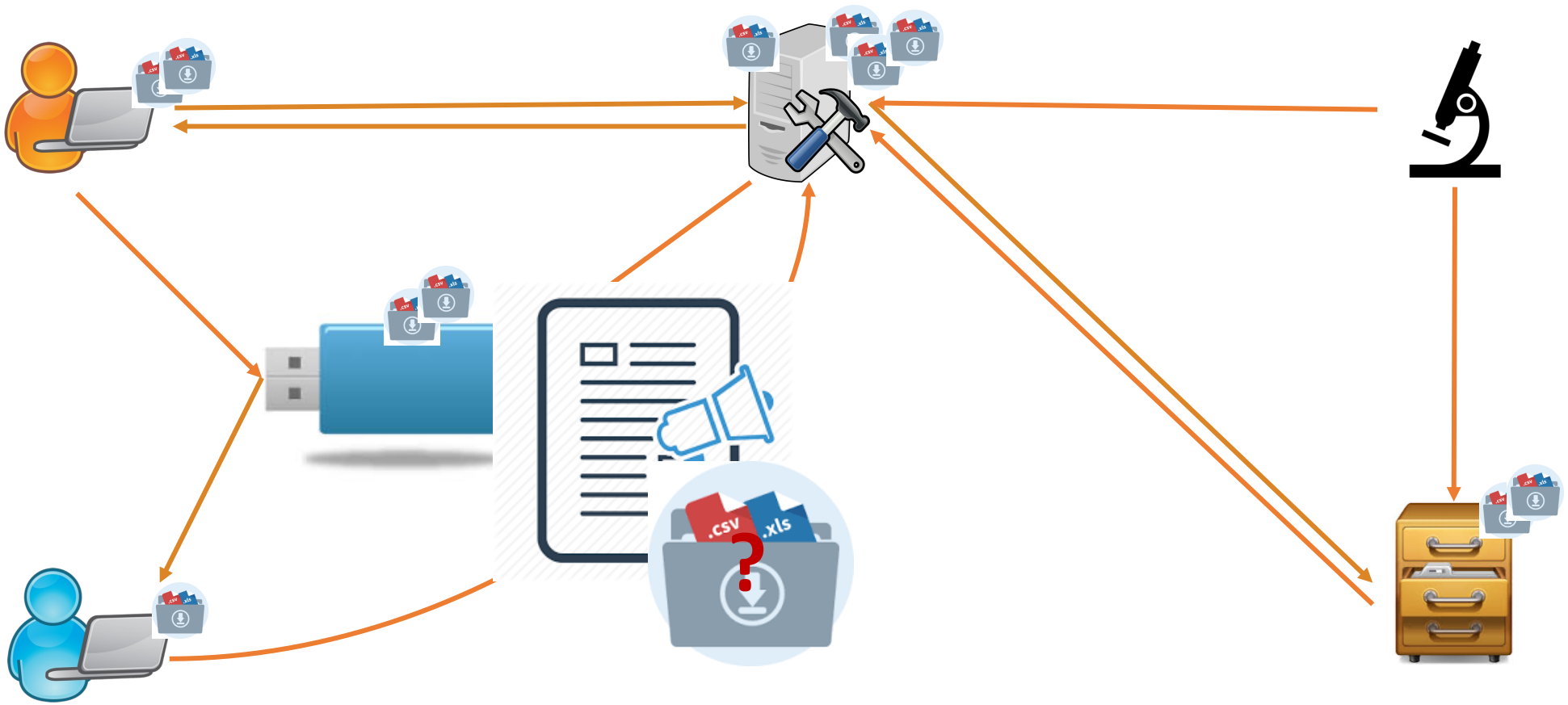
Using annotated data objects and annotating data objects in workflows

Data – Where is the problem?

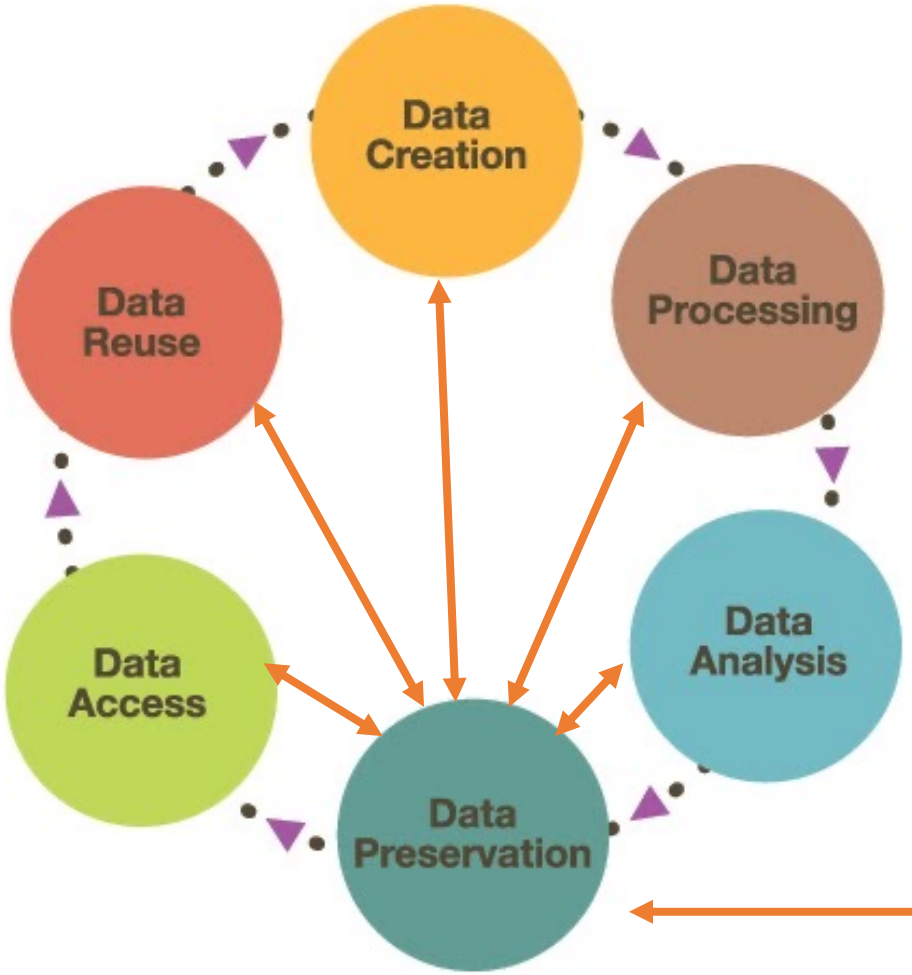


What would we need to master the chaos?

- What if data objects
  - Could tell us where their copies are stored?
  - Could tell us which other data is derived from them?
  - Would carry metadata describing how they were generated?
  - Could tell us in which state they are (version, published, volatile)?

- Other metadata than publishing metadata
- Metadata that cannot be separated from the data object
- A service that shows us the data object and not only the files

From data creation to archiving: Yoda



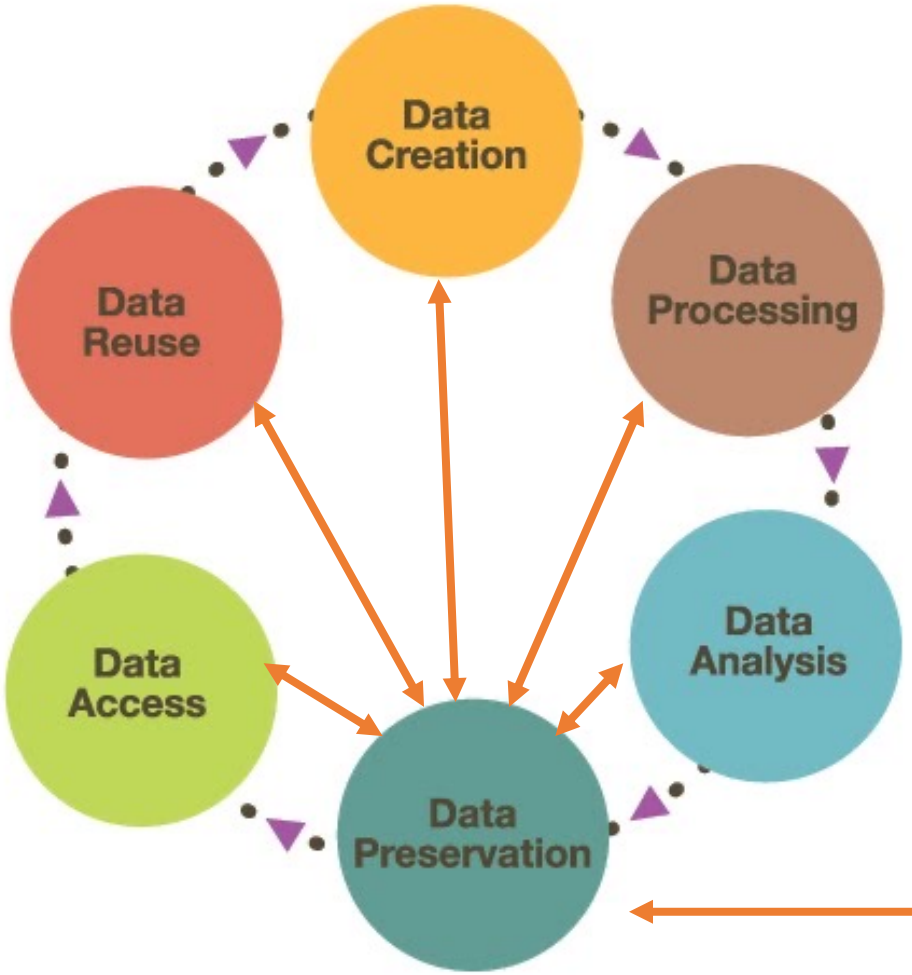
- Provides data steward workflow to
  - Ensure quality of archived data, data curation
  - Transfer responsibility from researcher to data manager

- No matter in which state your data is, Yoda is the route and guide to safe archiving

- Check out archived data to any stage in which the project is

Yoda: Preparing data for archiving and properly archive data

From data creation to archiving: Yoda



Side Note:

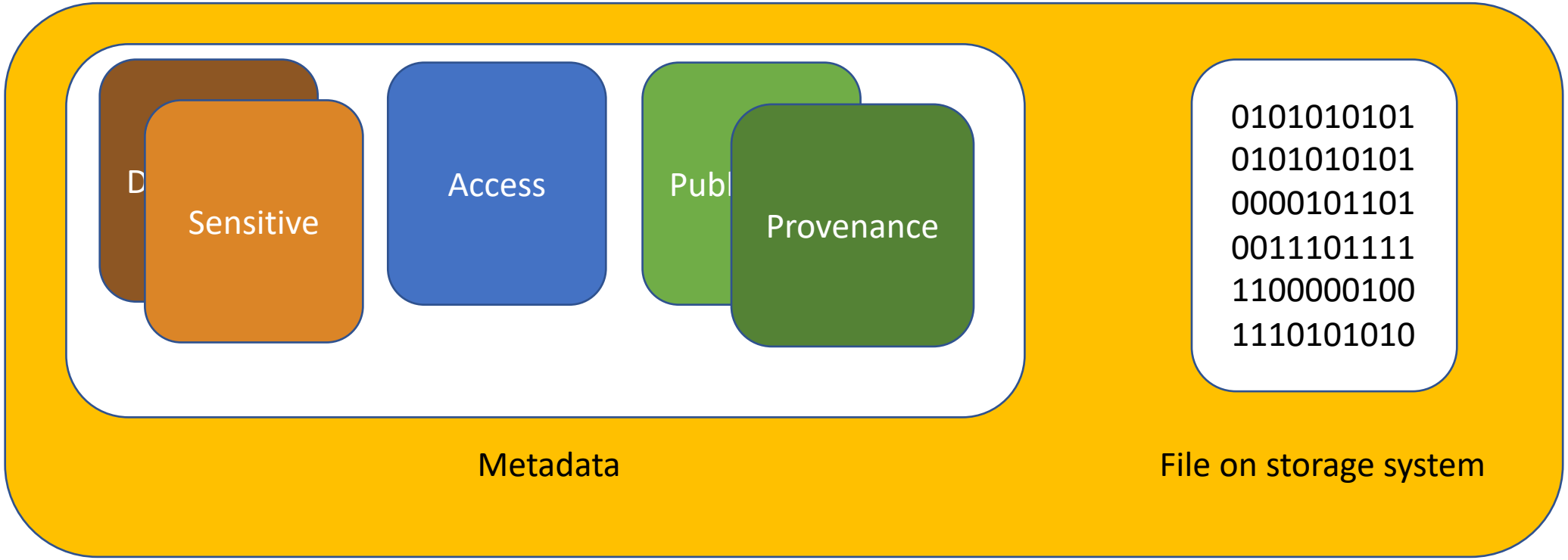
In practice we are using YODA research mainly for dumping data, Dropbox-like usage.

But there is more to it!!!!

Yoda: Preparing data for archiving and properly archive data

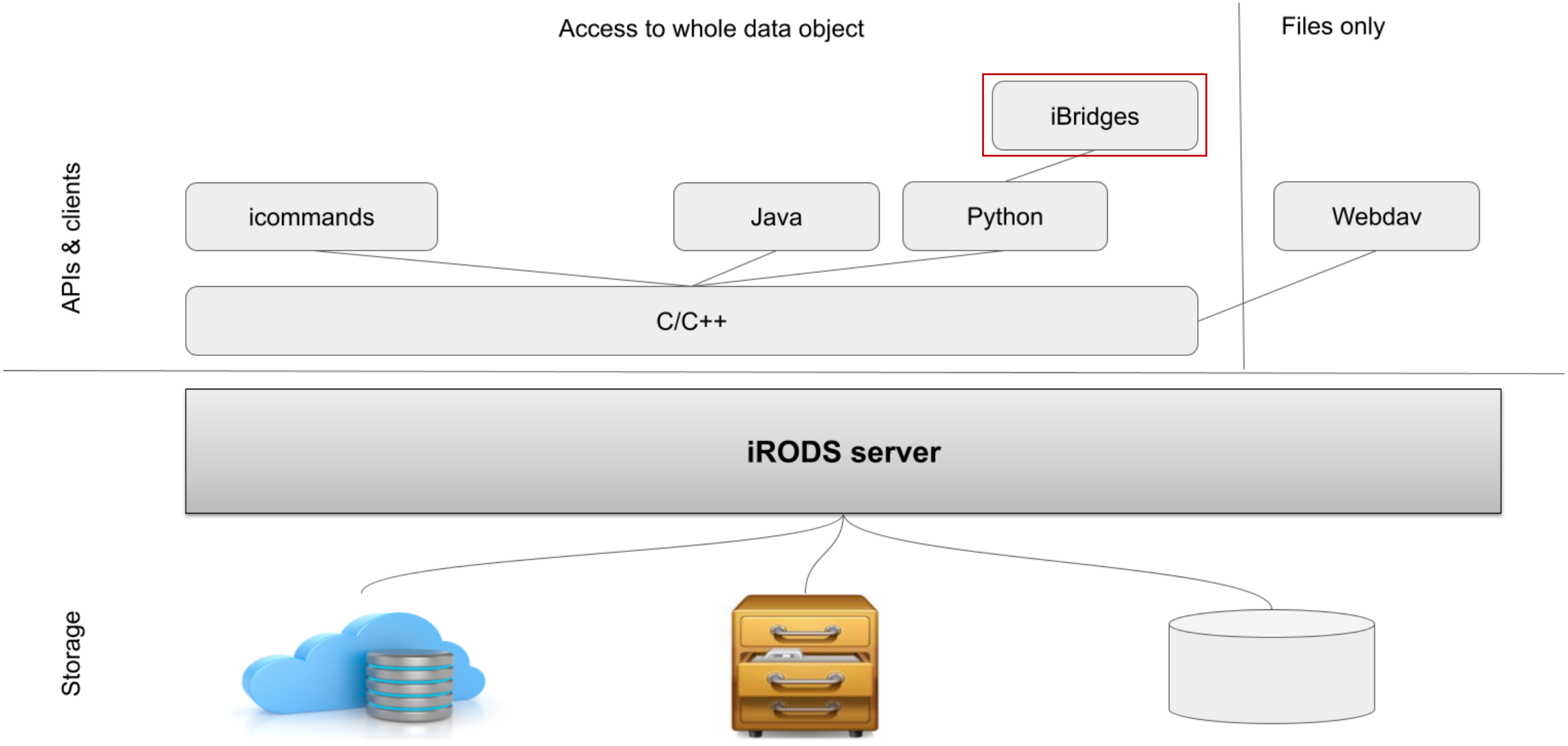
Data object

Data object = File + metadata + identifier

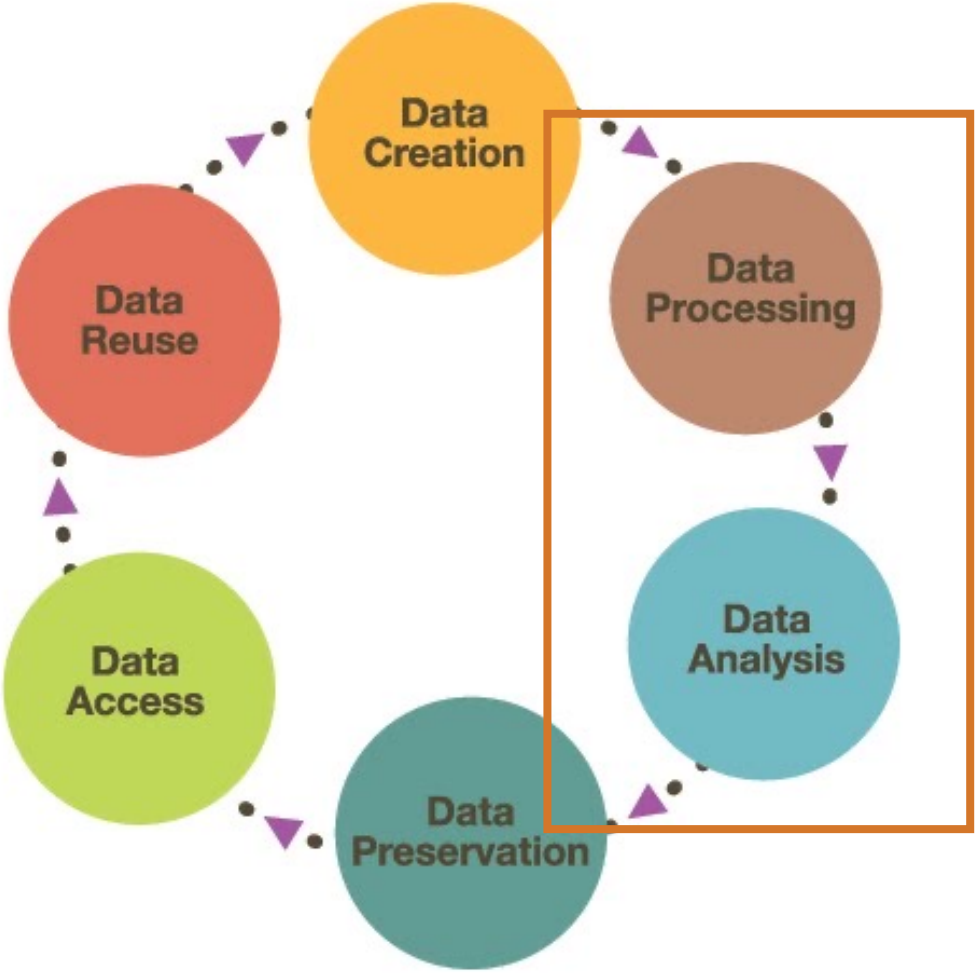


Keeping data and metadata tied together: iRODS → YODA extends to manged data packages

YODA: Accessing data objects



Simple workflow: Wordcount

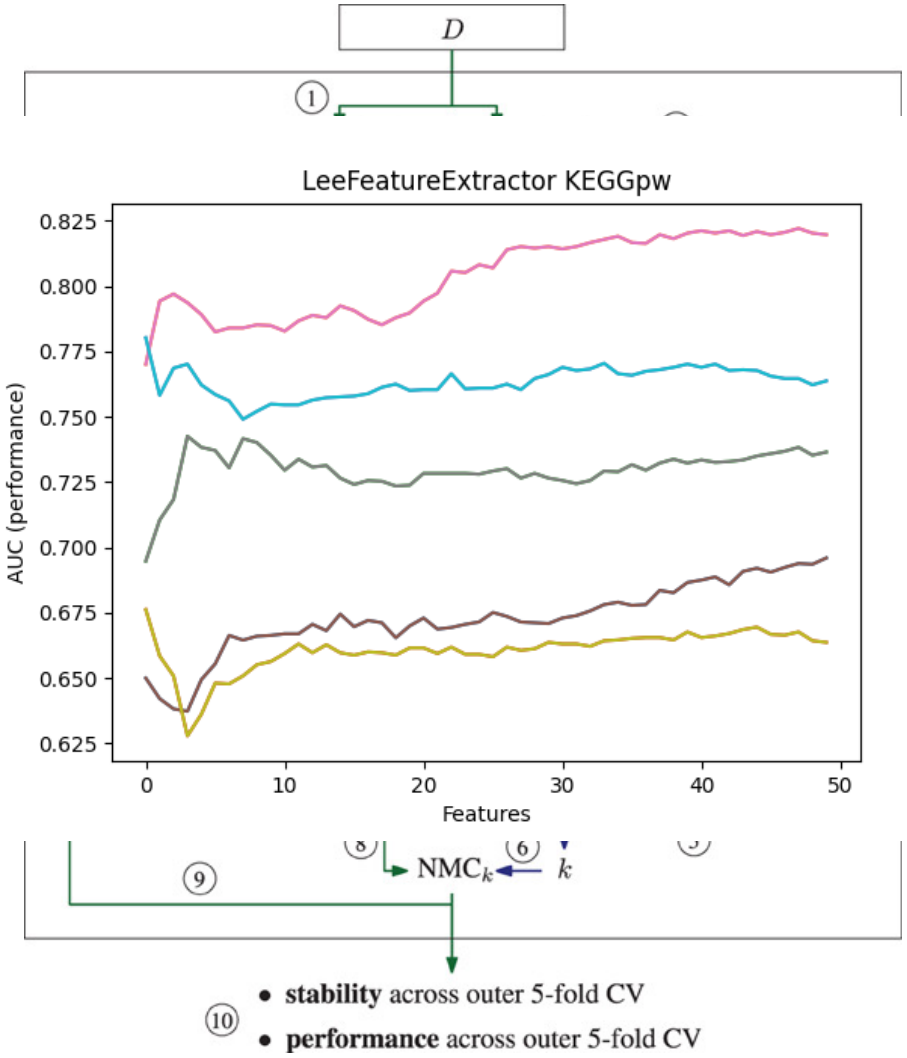


- Assume a colleague has gathered some (raw) data which you would like to analyse
- Assume the data is tagged with metadata (AUTHOR)

Workflow:

- Search for data in YODA
- Stream data into the analysis workflow
- Analyse data
- Upload results to YODA and tag with metadata from the PROV-O ontology

Realworld workflow: ACES



- **Combining published data from an external source with curated data stored in iRODS**

- RNA expression data from cancer patients (published data)
- Gene networks and pathway data (curated data)
- ACES: software to train classifiers to distinguish between patients

- **Research question which algorithm trained with which network data performs best, what are the selected features?**

Realworld workflow: ACES

- General iRODStraining material:
  - https://github.com/chStaiger/irods-training-compendium
- iRODS python API:
  - https://github.com/chStaiger/irods-training-compendium/blob/master/07-python-API-for-users.ipynb
  - https://github.com/chStaiger/irods-training-compendium/tree/master/13-HPC-RDM-Training