Karlsruhe Institute of Technology

SCC Steinbuch Centre for Computing

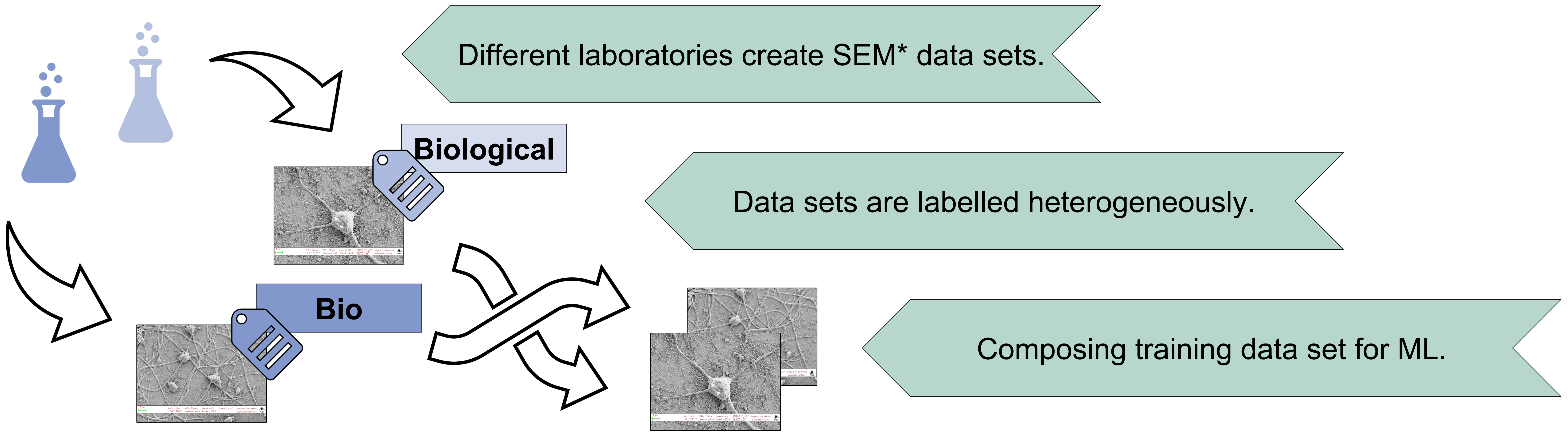
HMC HELMHOLTZ METADATA COLLABORATION

FAIR Digital Object for Accessing Label Information of ML Training Data Stored in a Metadata Schema

Nicolas Blumenröhr, Thomas Jejkal, Andreas Pfeil, Rainer Stotzka

Composing Machine Learning (ML) training data sets from heterogeneous sources is laborious due to their relabelling into uniform categories.

For automation, the FAIR Digital Object (FAIR DO) concept can be used in conjunction with a metadata schema.



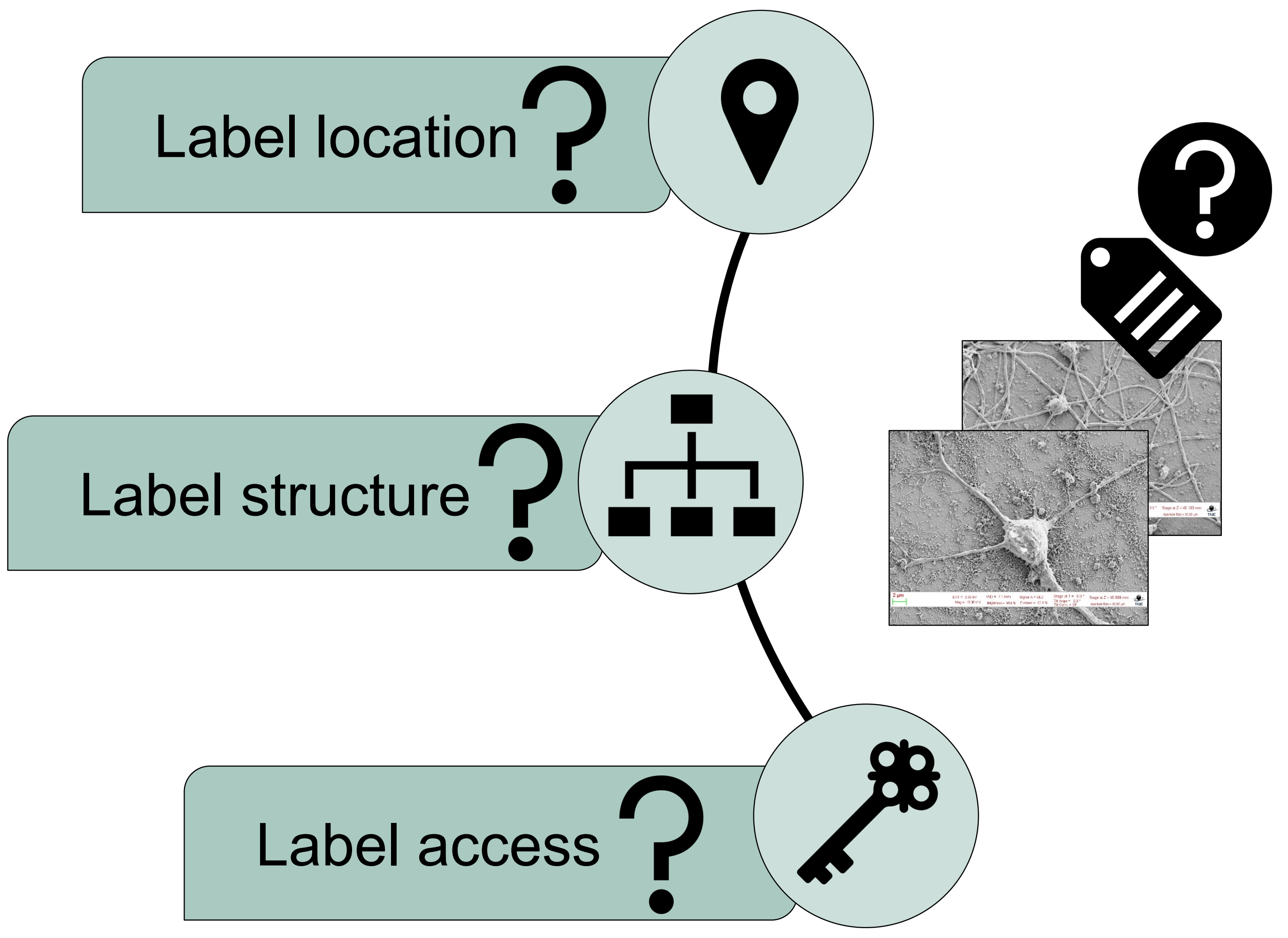
Different laboratories create SEM* data sets.

Data sets are labelled heterogeneously.

Composing training data set for ML.

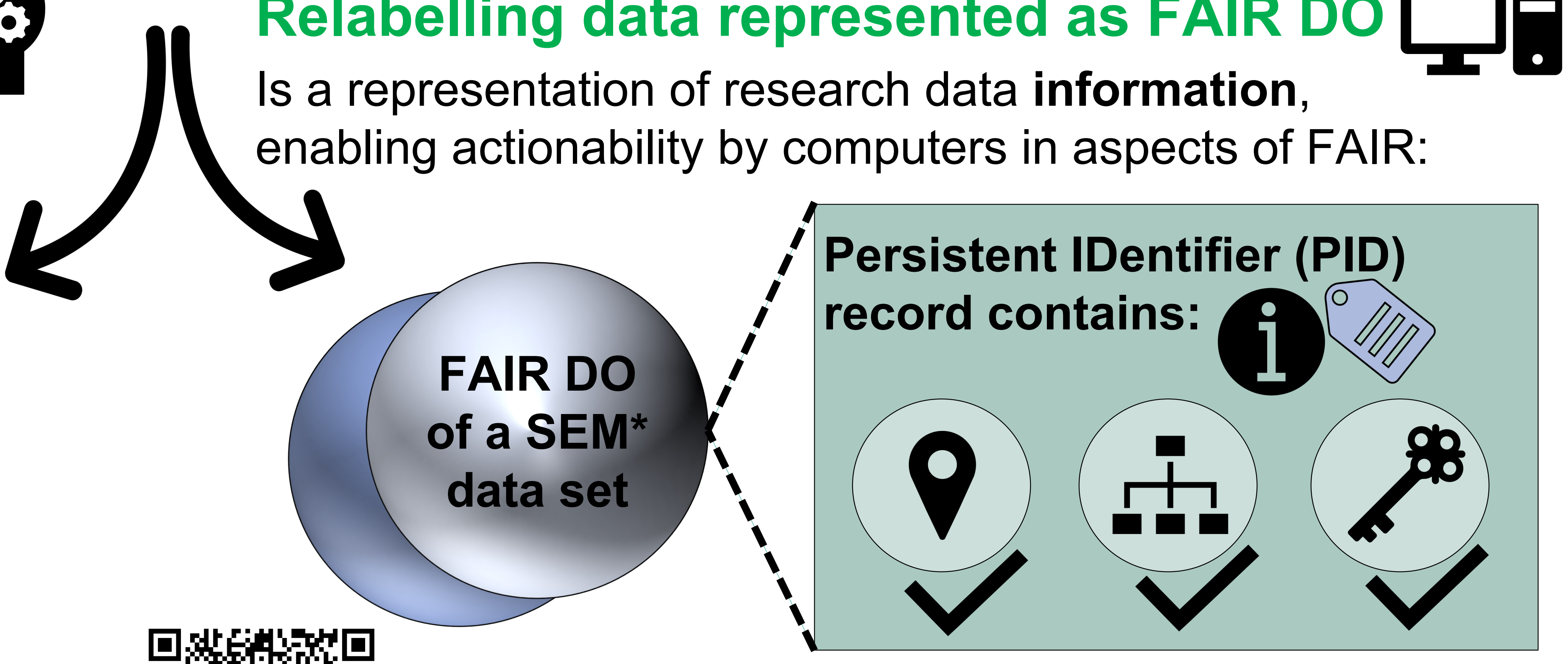
Relabelling data the classic way

Very time consuming, because research data **information** needs to be searched manually:



Relabelling data represented as FAIR DO

Is a representation of research data **information**, enabling actionability by computers in aspects of FAIR:



FAIR DO of a SEM* data set
Persistent IDentifier (PID) record contains:

Resolve the FAIR DO at: https://hdl.handle.net/21.11152/37833c54-1d36-42e4-858d-831447122863

Labels described with a metadata schema

Provides a standard that is referenced in the FAIR DO

```
{
    "labelProperties": {
        "levelOfLabel" : "image",
        "typeOfLabel": "string",
        "descriptionOfLabel": "SEM images labelled with
        10 different terms.",
        "labelTerms-DataObjectsAssignment": [{
            "labelTerm": "Biological",
            "descriptionOfLabelTerm": "Characteristics of
            cells and tissue.",
            "dataObjects": [
              "L7_0a800855e3b88fd72a83fe7dd8257f88",...]
            },
            ...
        ]
    }
}
```

Example structure of a custom JSON metadata schema-based document, where the label information for a ***Scanning Electron Microscopy (SEM)** training data set is described. (Provided by R. Aversa et. al. http://doi.org/10.23728/b2share.19cc2afd23e34b92b36a1dfd0113a89f)

Conclusions

- If laboratories represented their data as FAIR DOs, associated label information could be located and accessed easier.
- Additional description of the labels using a metadata schema provides a standardized structure of the label information.
- Clients and additional tools that are compatible with FAIR DOs and schemas can be used to enable partially, or fully, automated relabelling and other data preparation steps.
- This saves a lot of time for the ML user.