INTEGRATING DATA MANAGEMENT INTO COMPUTE WORKFLOWS WITH IRODS

SURF SARA

Your instructors for today

Arthur Newton

Claudia Behnke

Claudio Cacciari

Agenda

09:30 – 10:00 Data Management, FAIR and iRODS

10:00 – 11:00 **Hands-on**: Data handling with the python API

11:00 – 11:15 **Coffee break**

11:15 – 12:00 **Hands-on**: Data handling with the python API continued

12:00 – 13:30 **Lunch**

13:30 – 14:00 The HPC system, compute workflows and iRODS

14:00 – 15:00 **Hands-on**: Two compute workflows

15:00 – 15:15 **Coffee break**

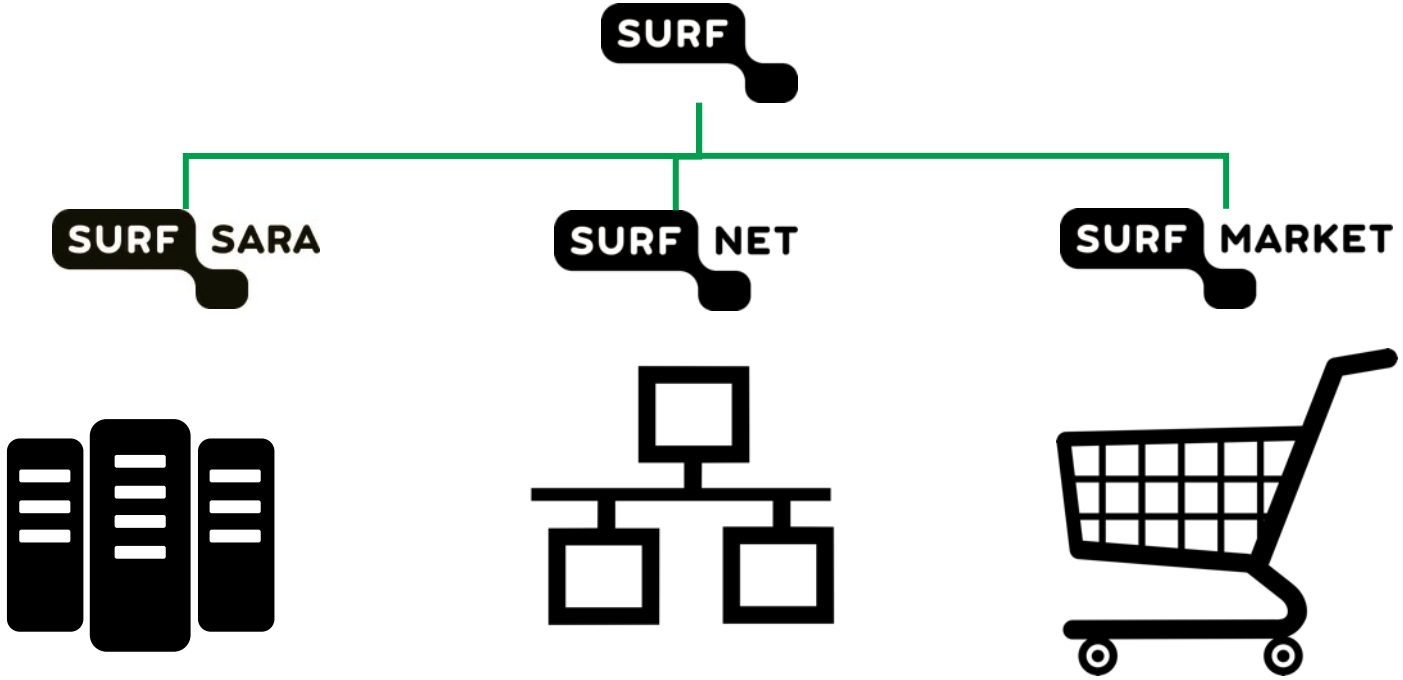
15:15 – 16:00 **Hands-on**: Two compute workflows / Open Questions

16:00 – 16:30 **Wrap up and evaluation**

SURF SARA

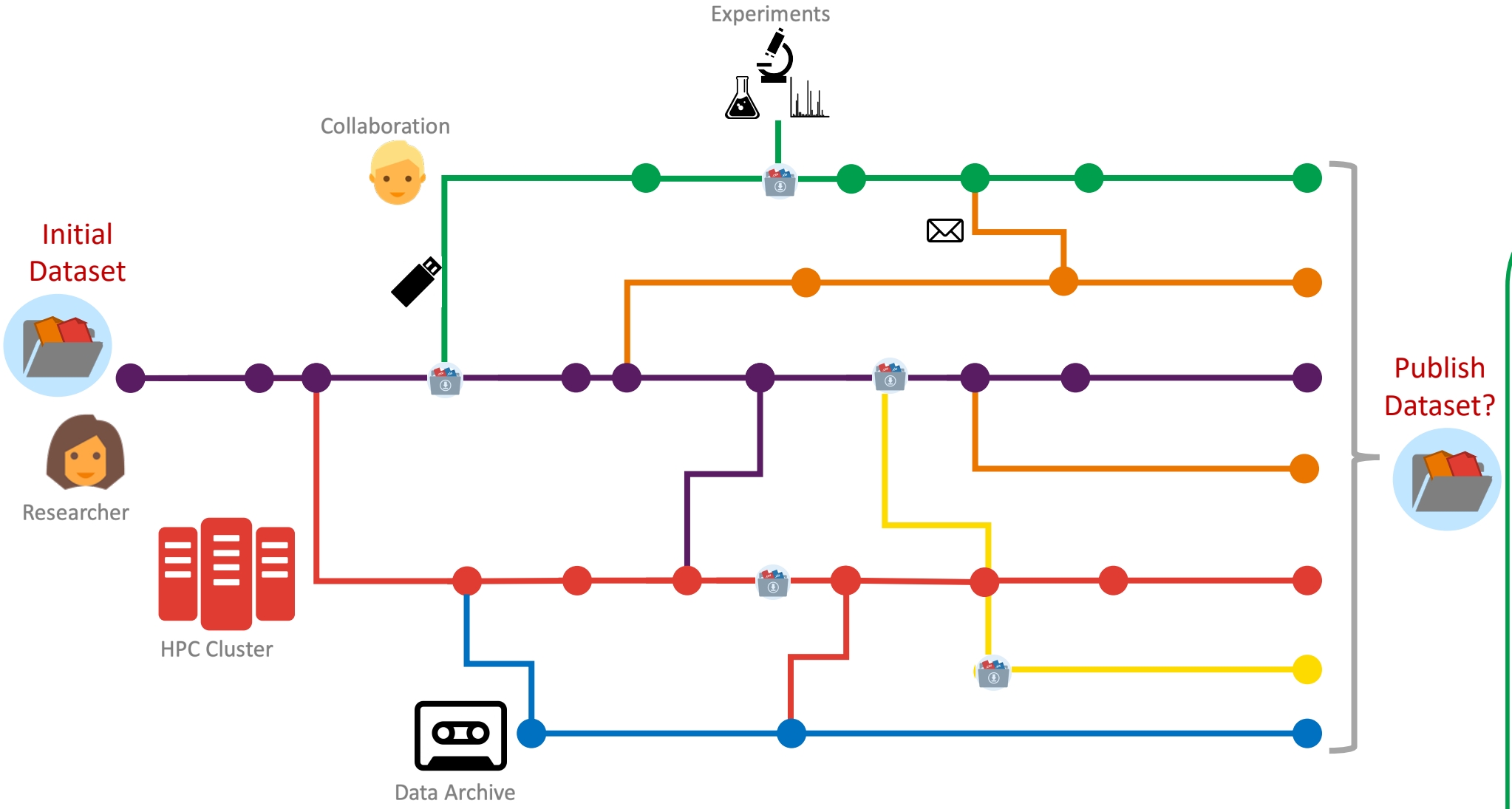
About SURF

Collaborative ICT* organization to support Dutch education and research

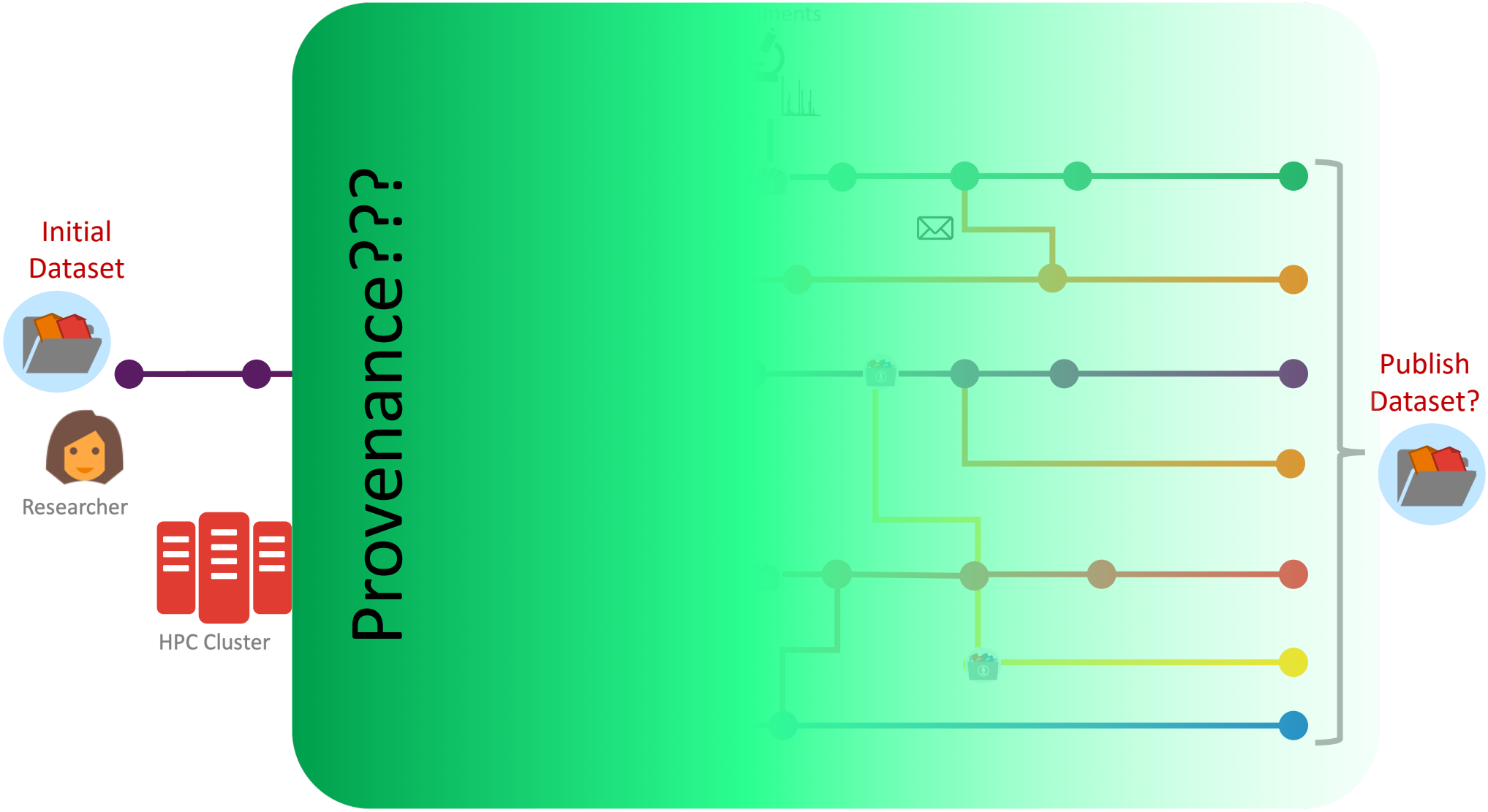


*Information and communications technology

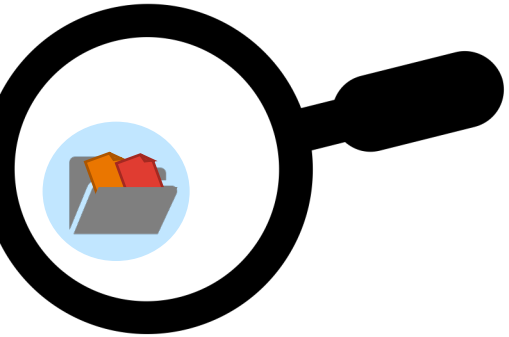
Data, what's the problem?

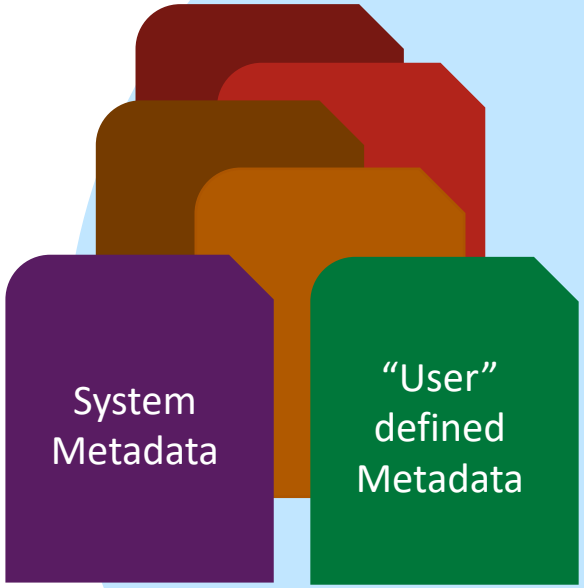


Data, what's the problem?

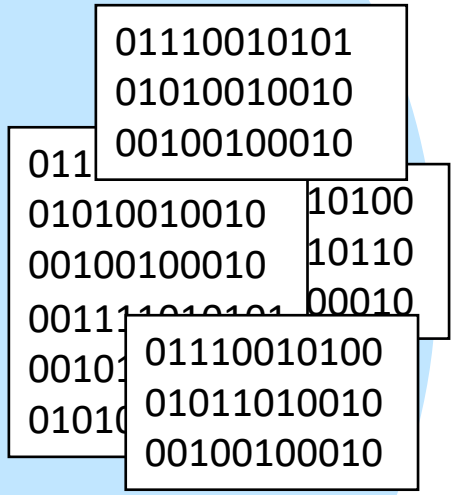


Data, what's the problem?





Referring



Metadata

Bitstreams (e.g. files)

11904/428b37b0-e42a-11e9-8c3a-38f9d326da8b

PID

Data Life Cycle

Experiments

Provenance

Descriptive

sensitive

CREATING

PROCESSING

ANALYSING

PRESERVING

SHARING

RE-USING

HPC Cluster

Operational Metadata

System Metadata

Access rights

Data Archive

SURF SARA

Data Life Cycle

https://www.nature.com/articles/sdata201618

Experiments

Provenance

Descriptive

sensitive

RE-USING

Findable

Accessible

Interoperable

Reusable

Operational Metadata

System Metadata

Access rights

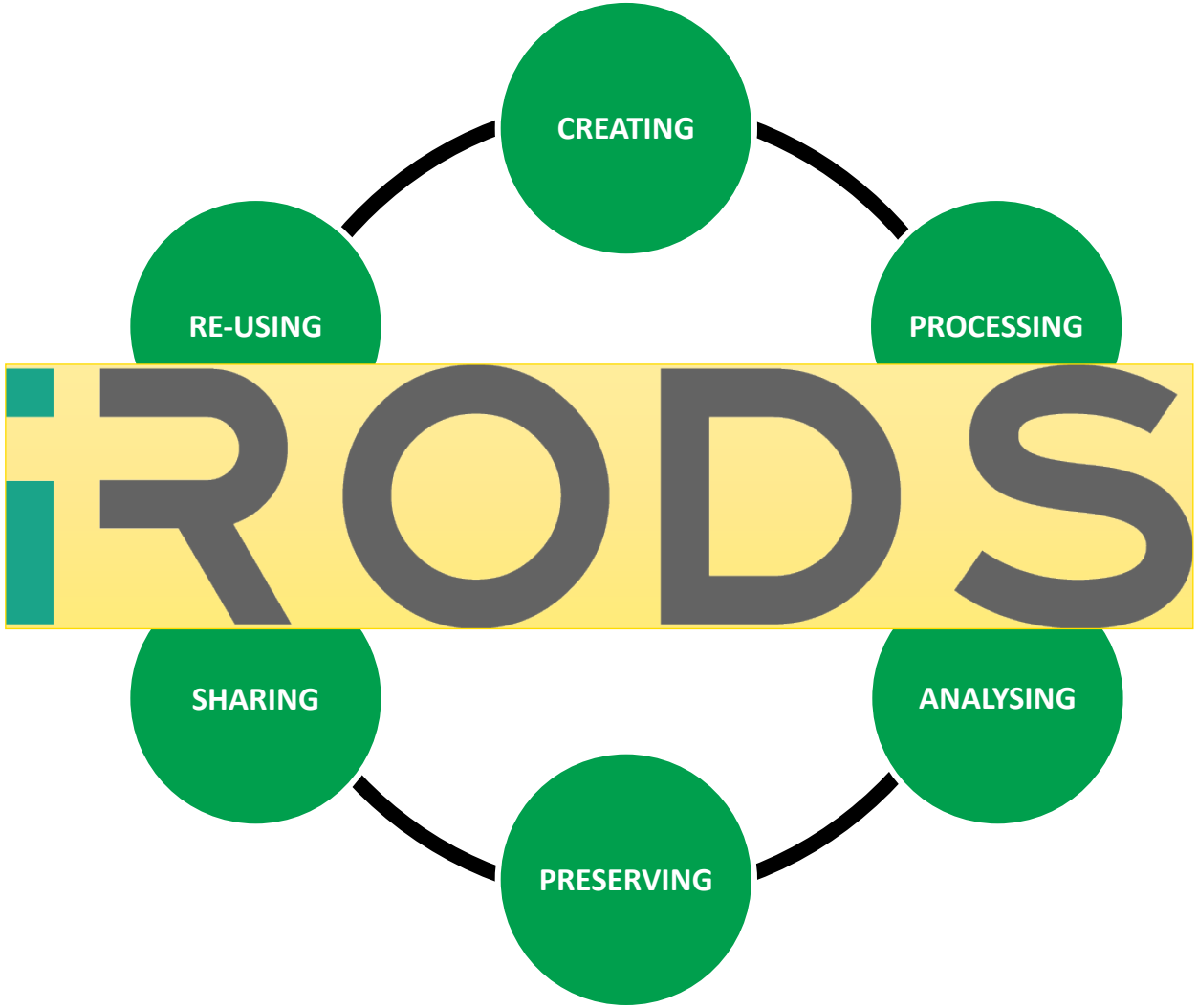
SHARING

PRESERVING

Data Archive

SURF SARA

Data Life Cycle



What is iRODS? - Core competencies

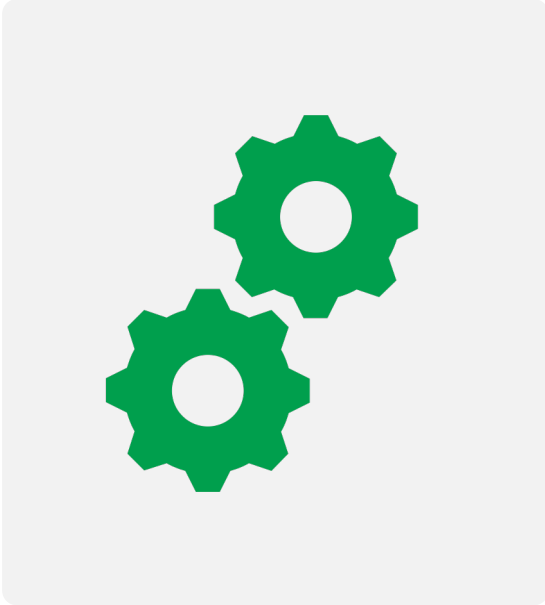
Unified namespace

Data Virtualization

Data Discovery

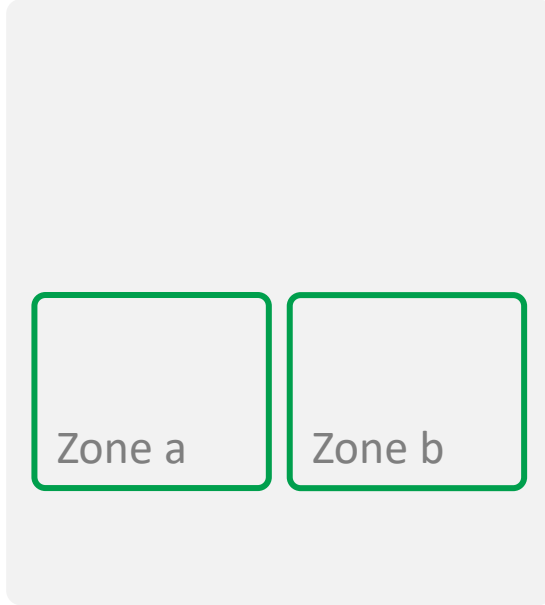
Metadata

Workflow Automation



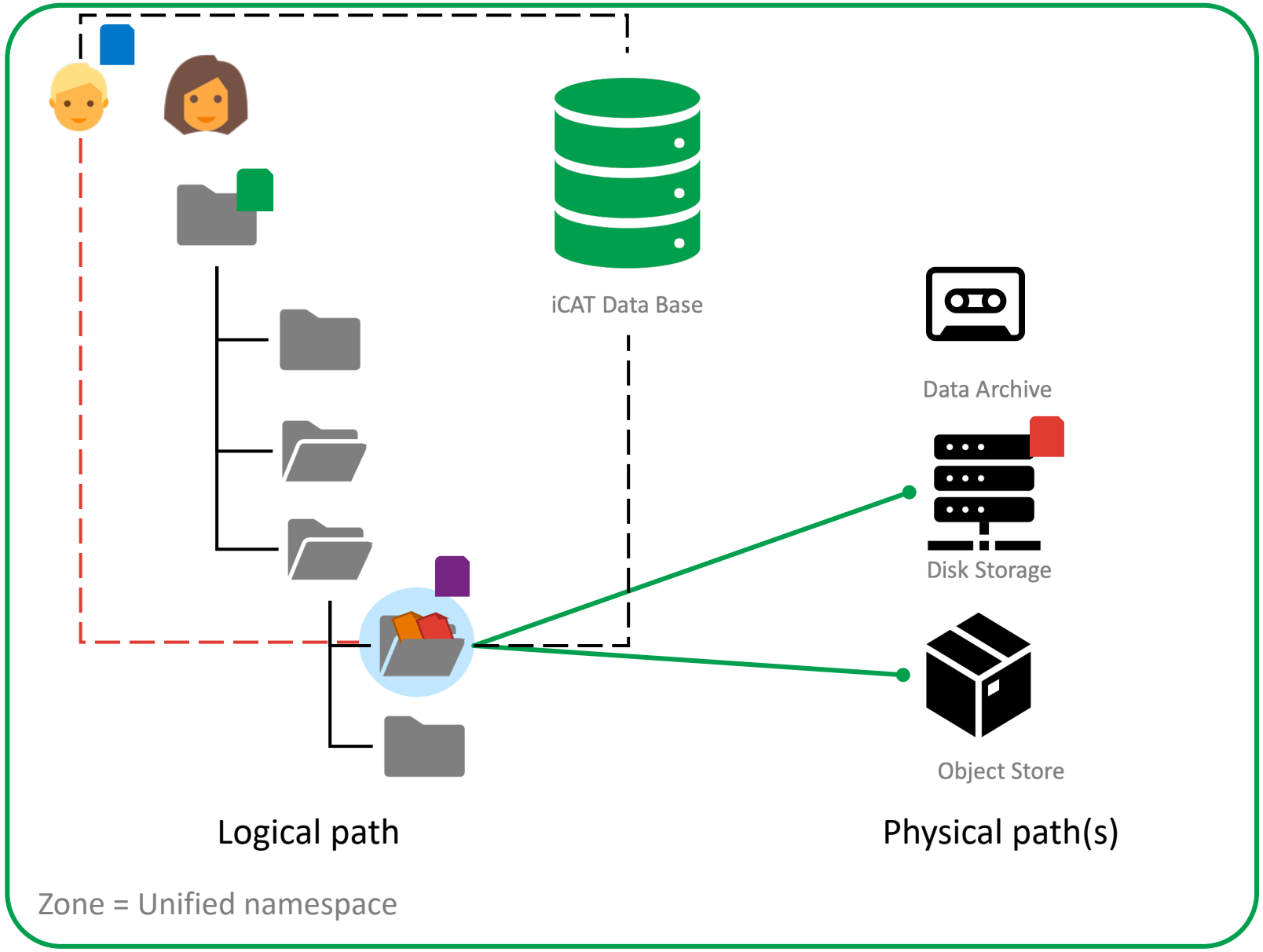
Rule Engine

Secure Collaboration

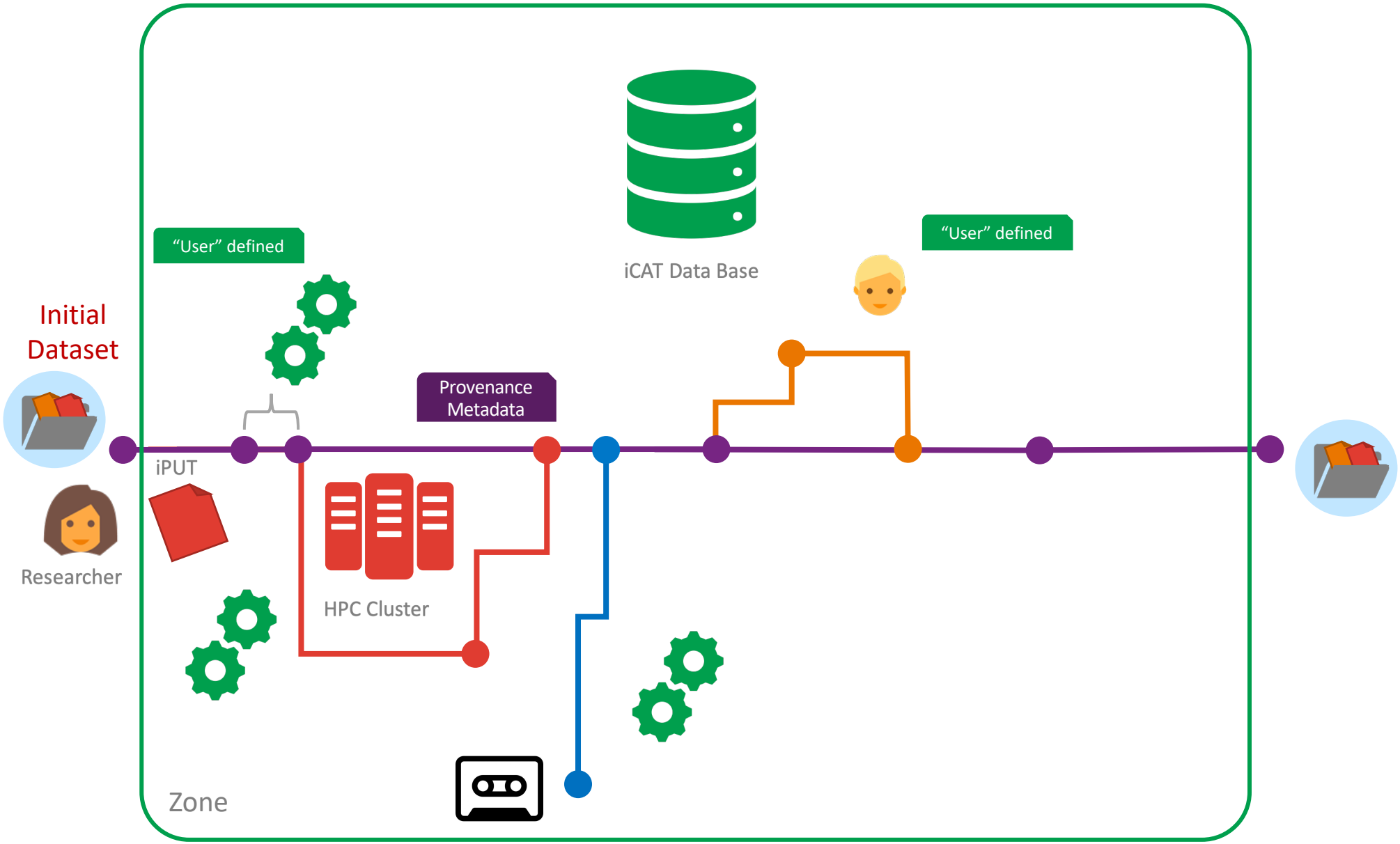


Federation

Data virtualisation & Meta data



Rule engine workflows



Data management platform based on iRODS

- One entrance point for the user to many storage services
- Good interface to compute services and other data applications
- Data policies: configure the behavior of data throughout the data life cycle
- Data sharing within research groups and with external collaborators

Findable: within the data management platform, external findability by e.g. persistent identifiers

Accessible: coupling with federated identities, accession control lists, well-defined transfer protocols

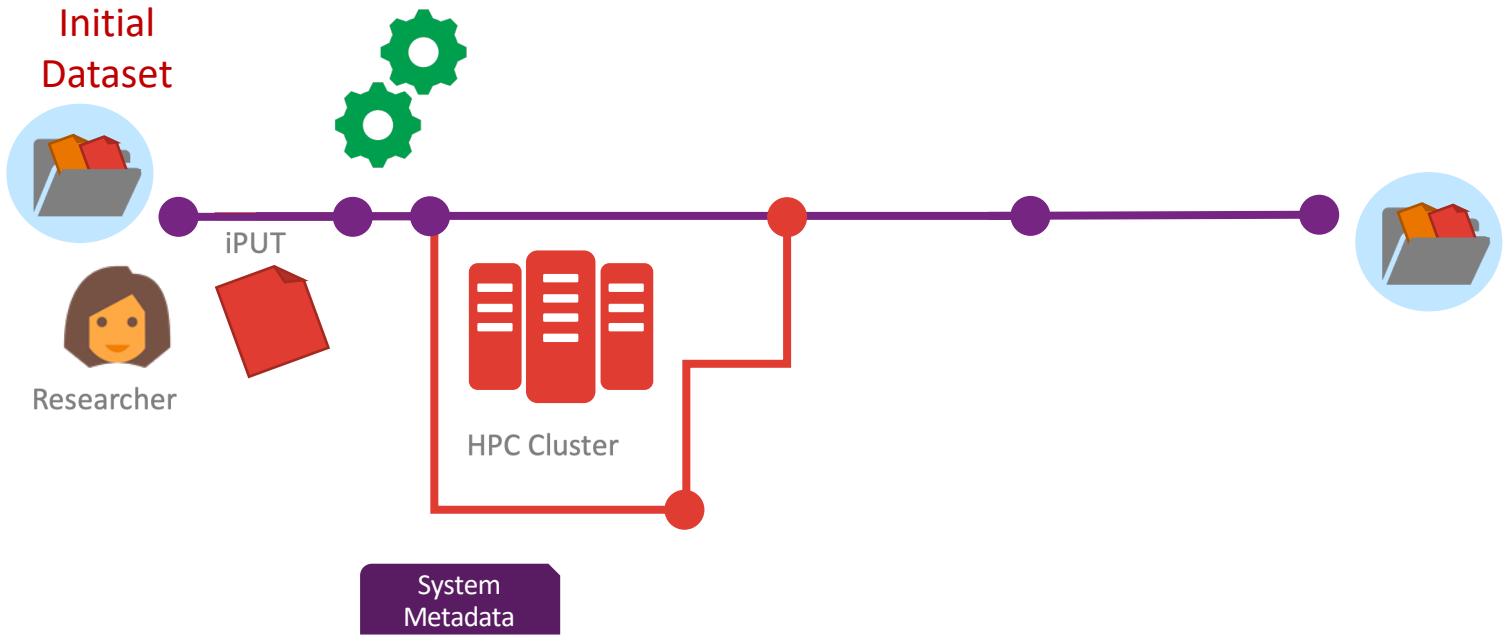
Interoperable: strongly dependent on data formats and employed metadata standards

Reusable: Depending on implemented data policies; within the limits of metadata annotation and standards for data formats

SURF SARA

Hands on for today

- Learn how to programmatically access data stored in iRODS via the iRODS python API
- Learn about iRODS concepts: data objects, collections, metadata handling querying
- Learn how to find data based on metadata (not on some knowledge of hand made directory trees) and use it in an HPC system



Some questions for the audience

- Who got a sdemo (or your own) account for Lisa?
- Who changed their password and was able to login to Lisa via a ssh client?
- Who knows the basics of Python?
- Who knows what iPython is?
- Who knows iRODS already?
- Who is doing data intensive compute jobs already on Lisa?
- Who is ready for the course!?

https://edu.nl/hgj9e

SURF SARA

Federated identities