DATA MANAGEMENT

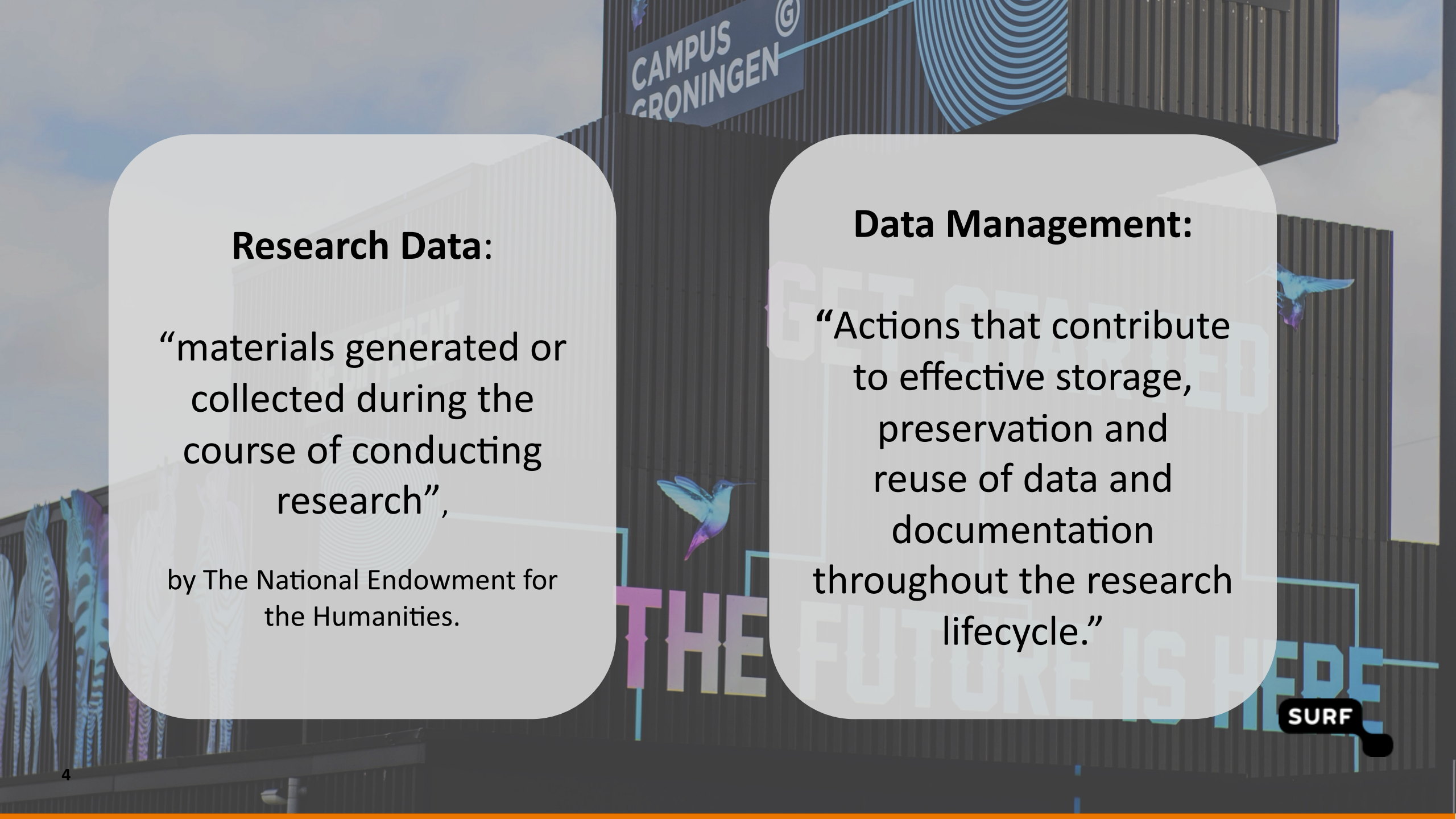
(WITH SURF)

What are we covering

- 09:00 – 09:45: Data management with SURF
- 10:00 – 11:00: Hands-on: Data processing with Lisa and Research Drive
- 11:00 – 12:30: Hands-on: Data management with Yoda

What are we covering

- What is Research Data Management
- The data lifecycle
- The data management landscape, according to SURF



Research Data:

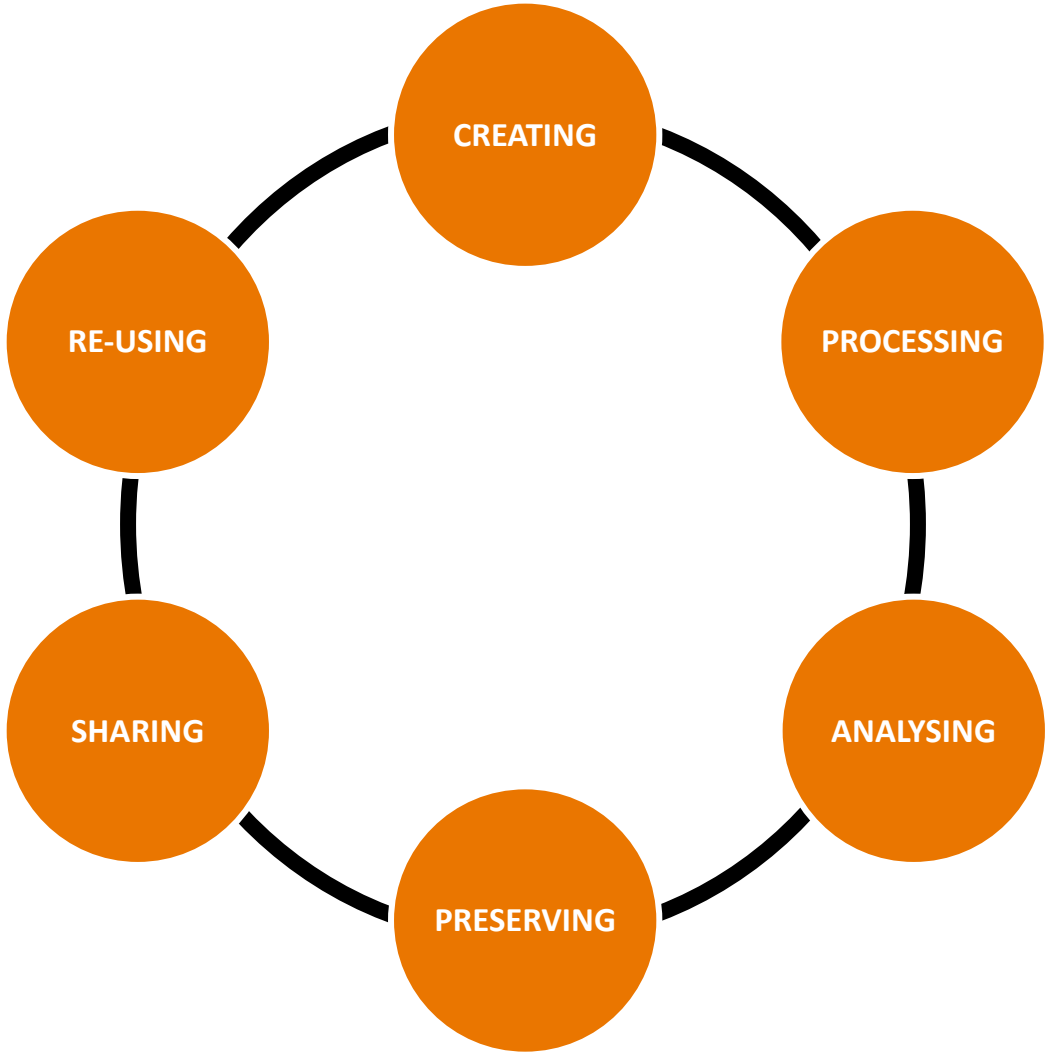
“materials generated or collected during the course of conducting research”,

by The National Endowment for the Humanities.

Data Management:

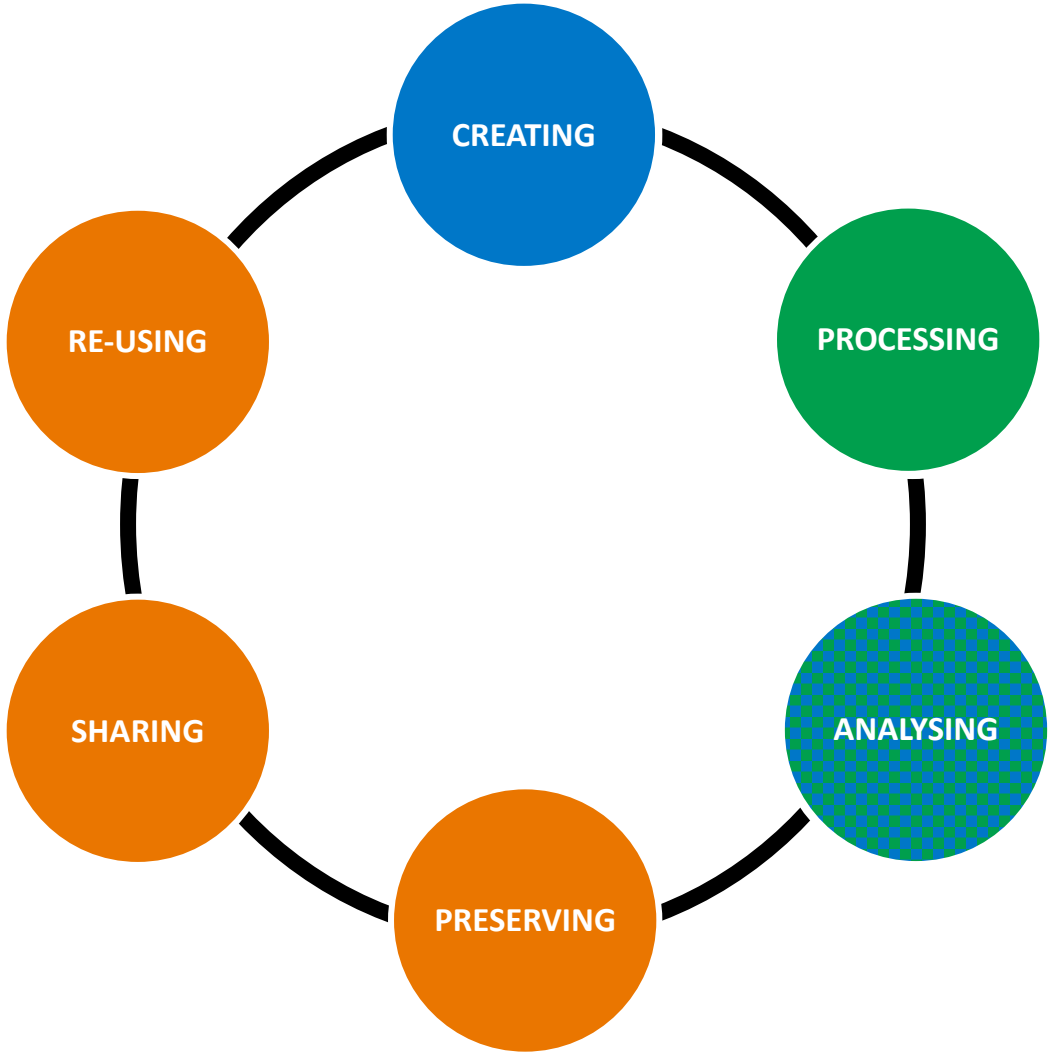
“Actions that contribute to effective storage, preservation and reuse of data and documentation throughout the research lifecycle.”

Data Life Cycle – the holy grail of Research Data Management



Data Life Cycle – the holy grail of Research Data Management

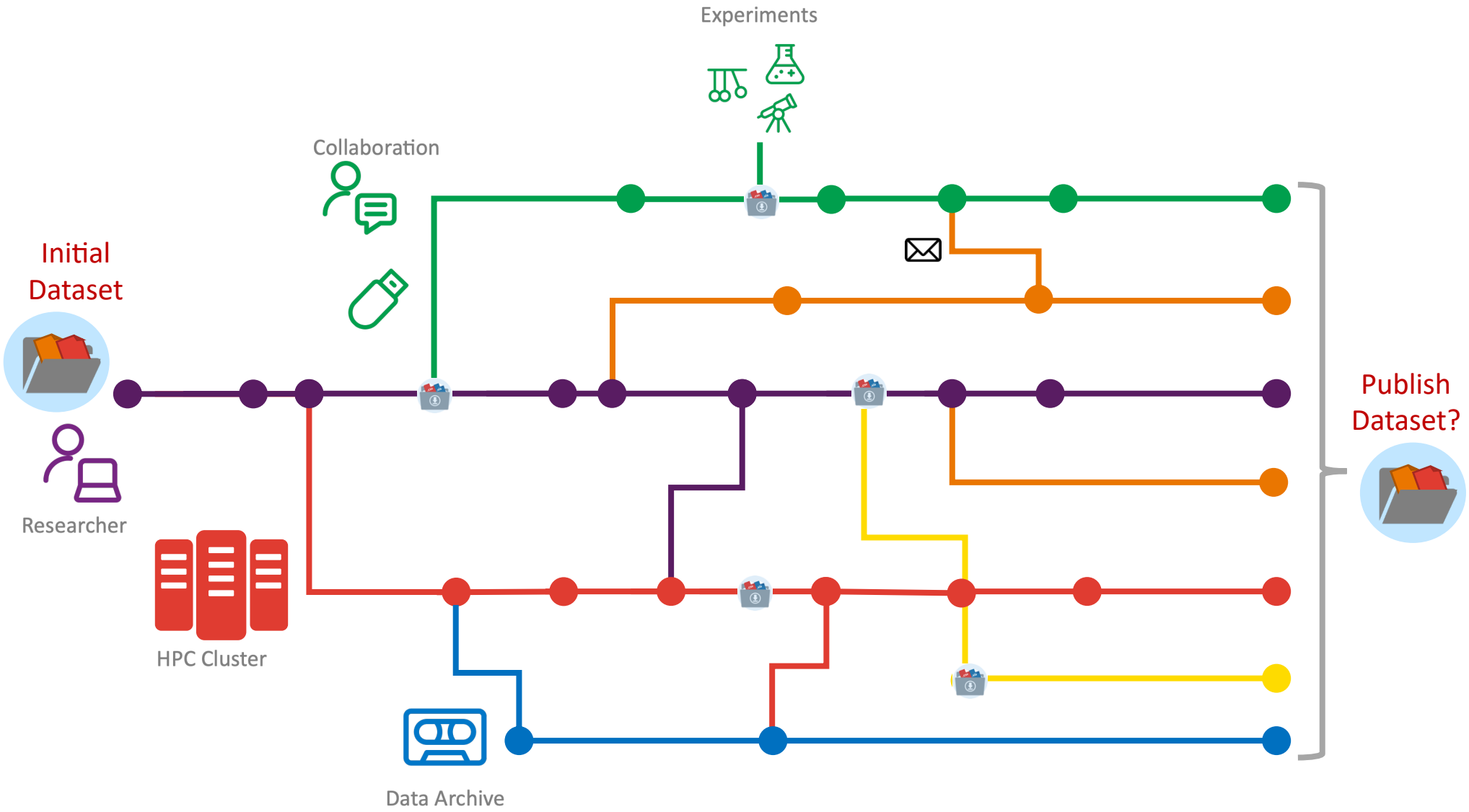
Data Services

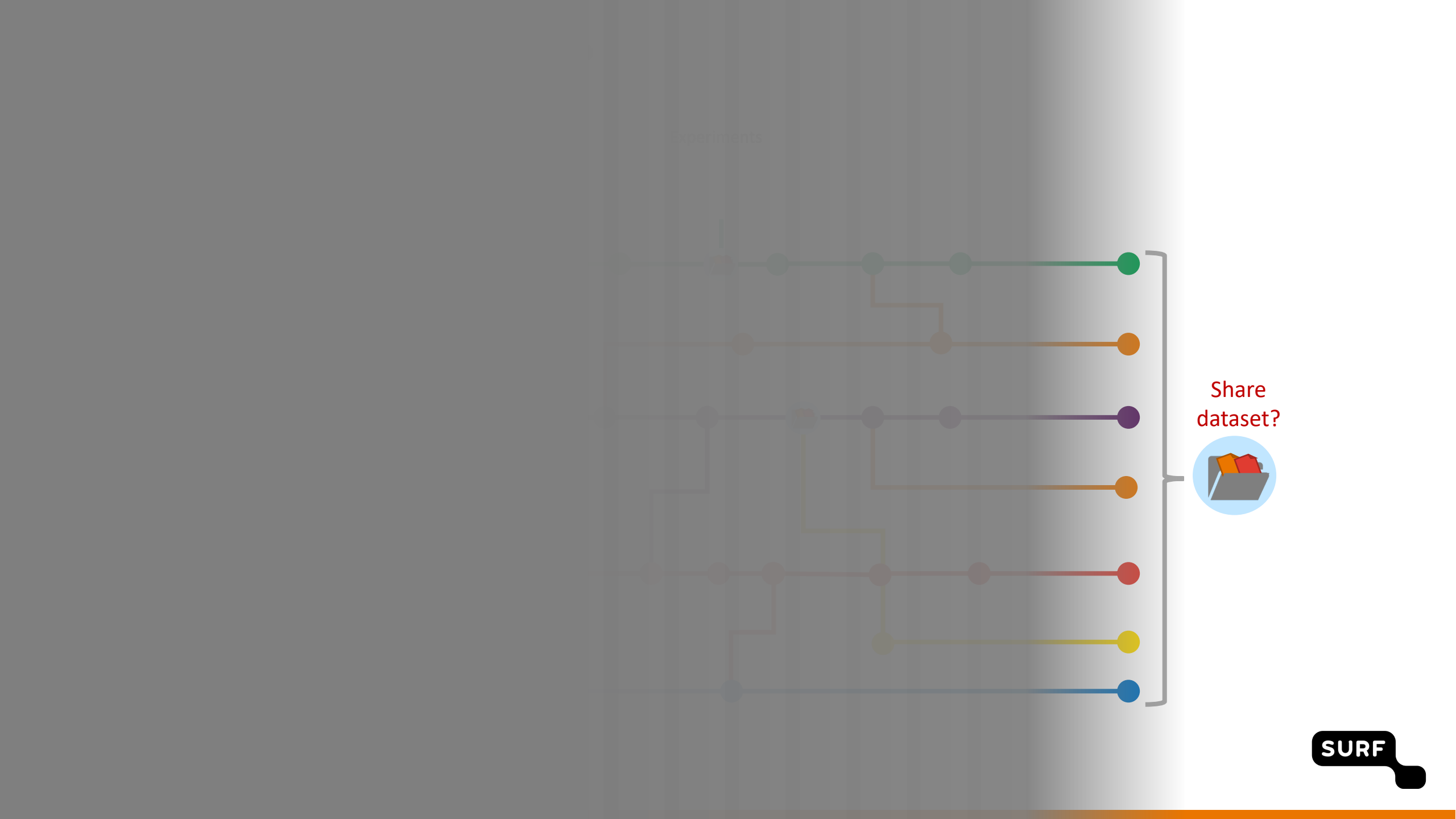


Compute

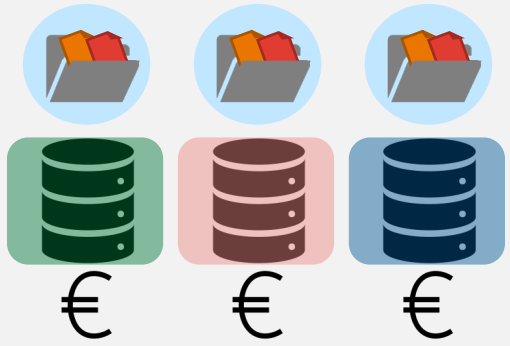
Data Processing

Data, what's the problem?




Share
dataset?

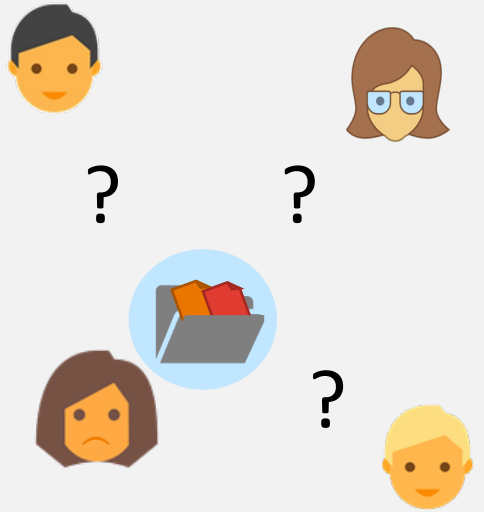
Consequences of bad research data management



No cost effective data storage

Data gets lost by disaster or loss of context

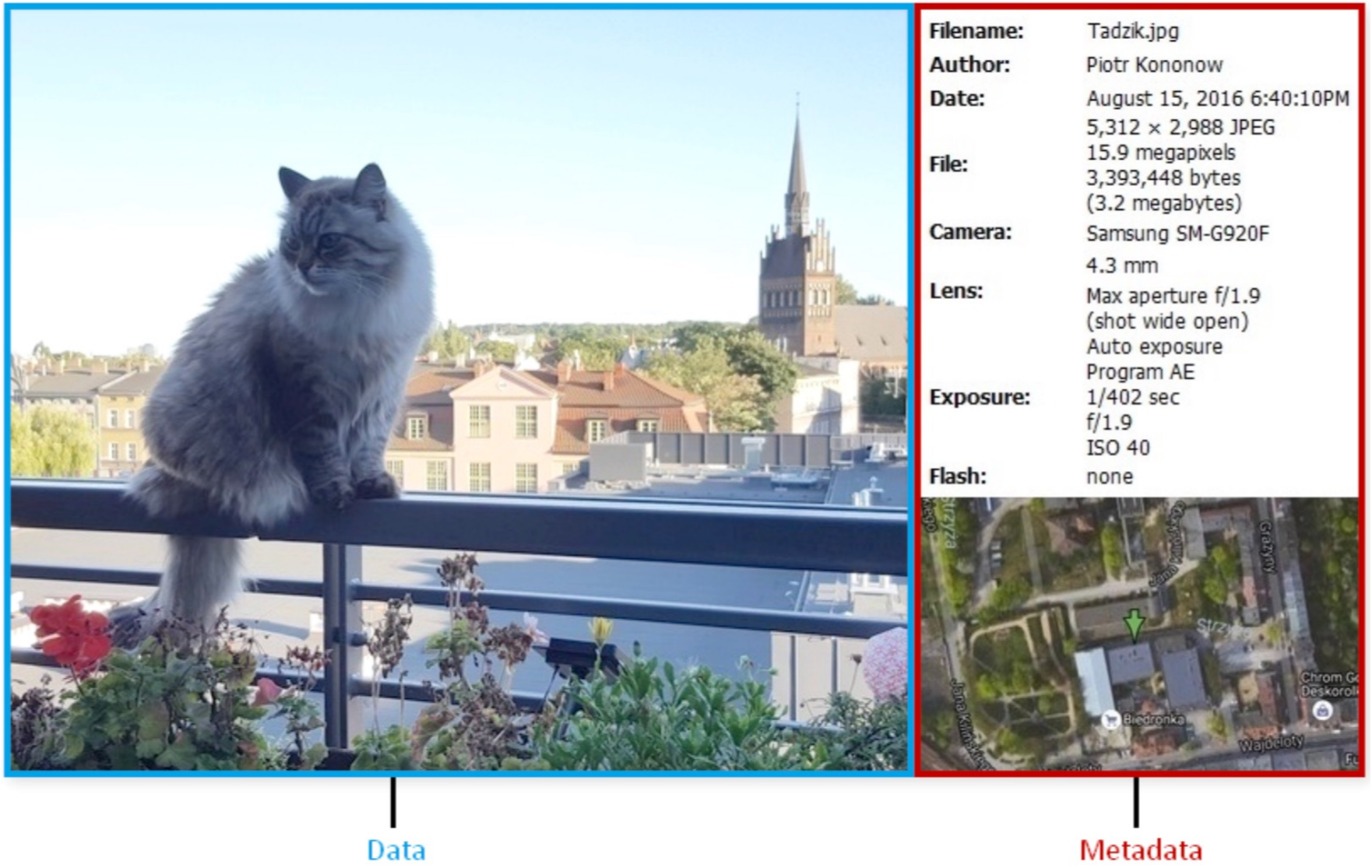
Redoing experiments and no interoperability



Not able or fear to share / publish data

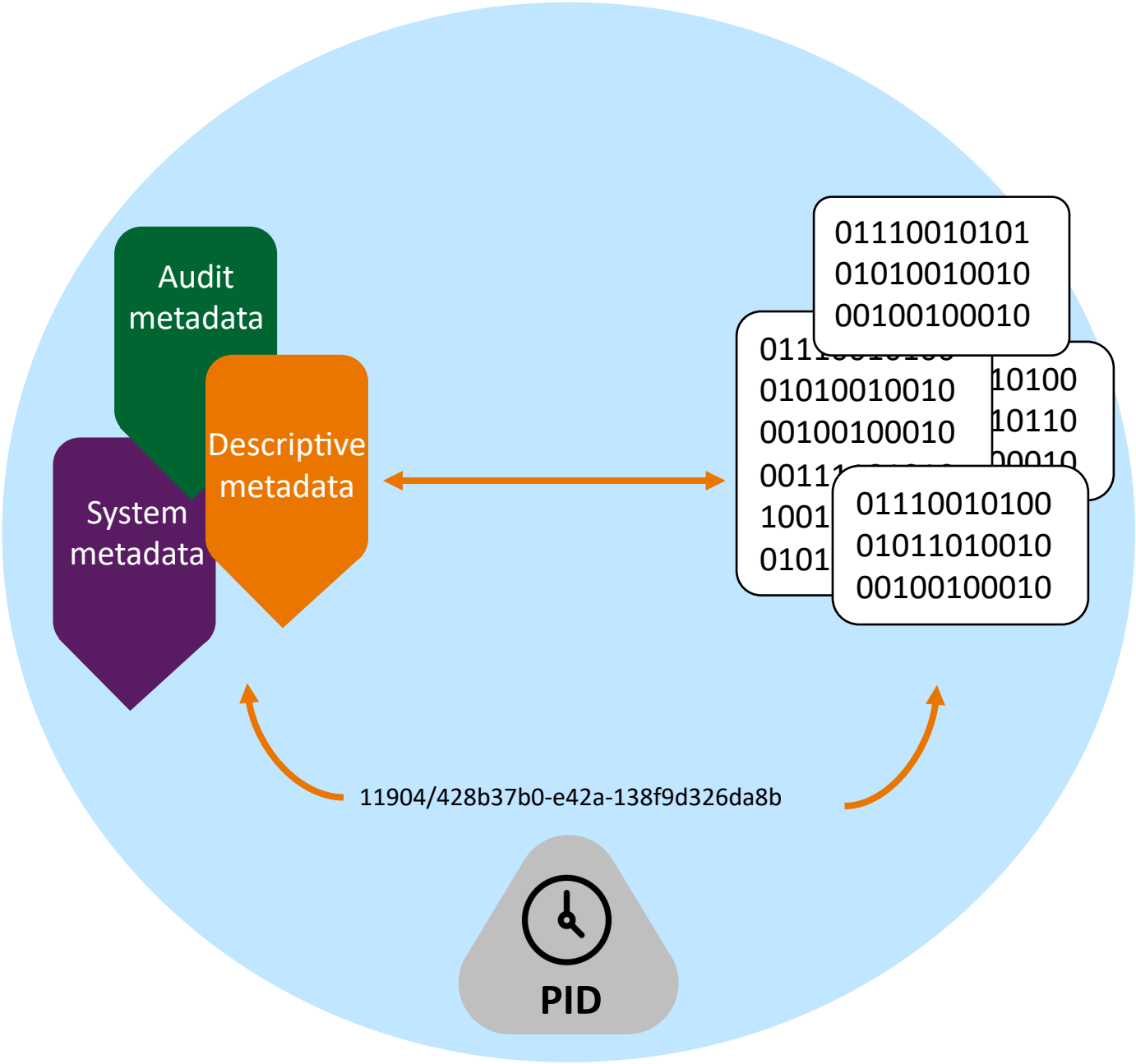
Why using metadata?

- Facilitate data discovery
- Help users determine the applicability of the data
- Enable interpretation and reuse of data
- Allow any limitations to be understood
- Clarify ownership and restrictions on reuse

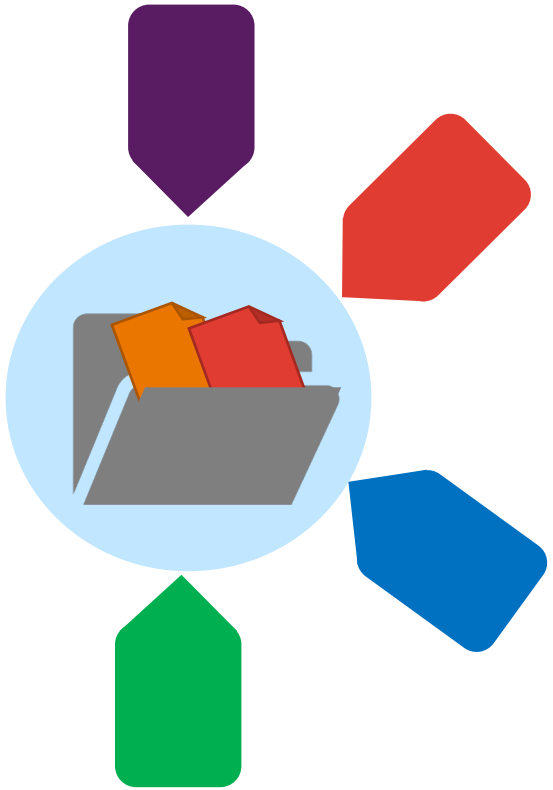


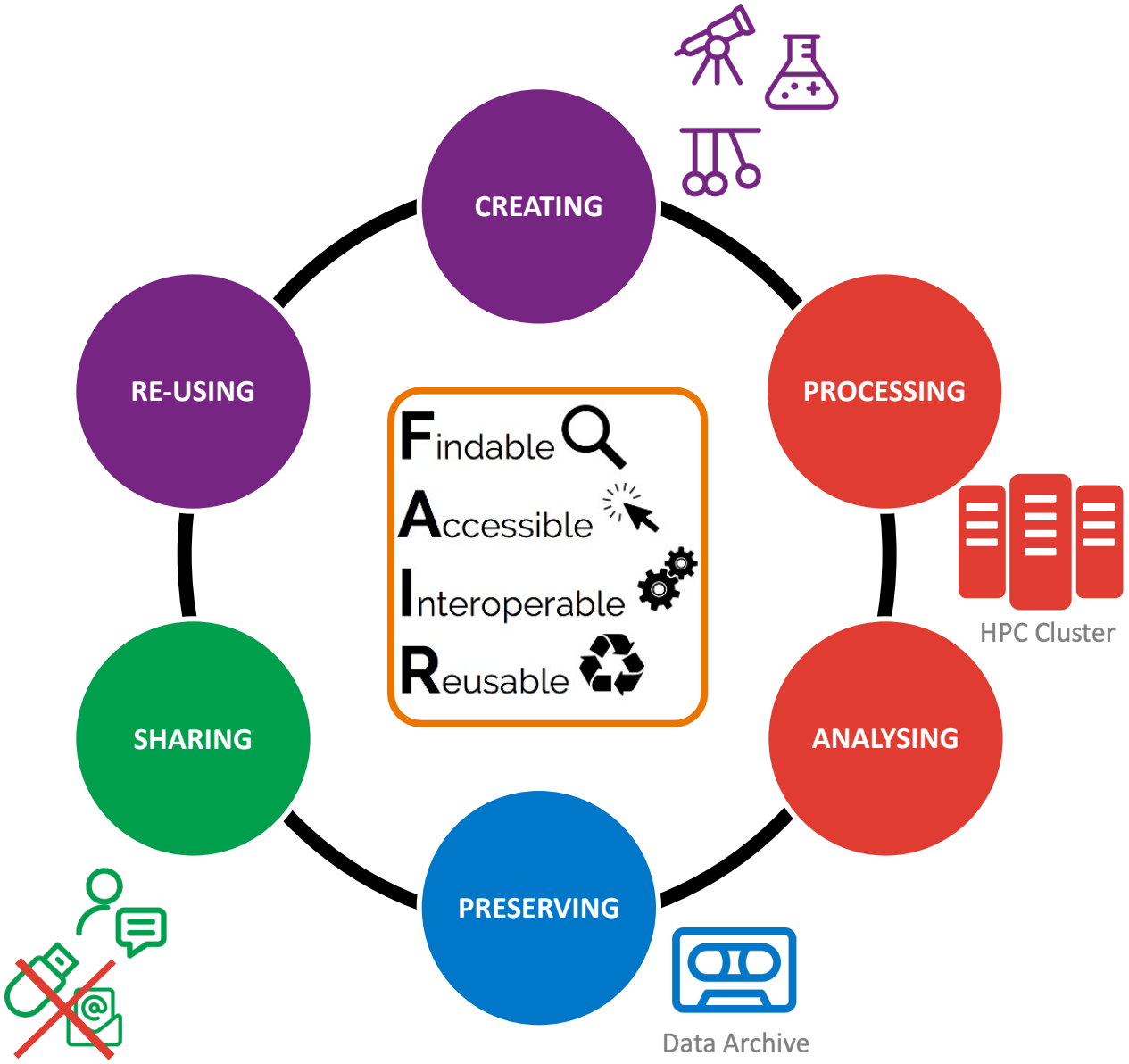
Ref: https://dataedo.com/kb/data-glossary/what-is-metadata

Data, what is it?

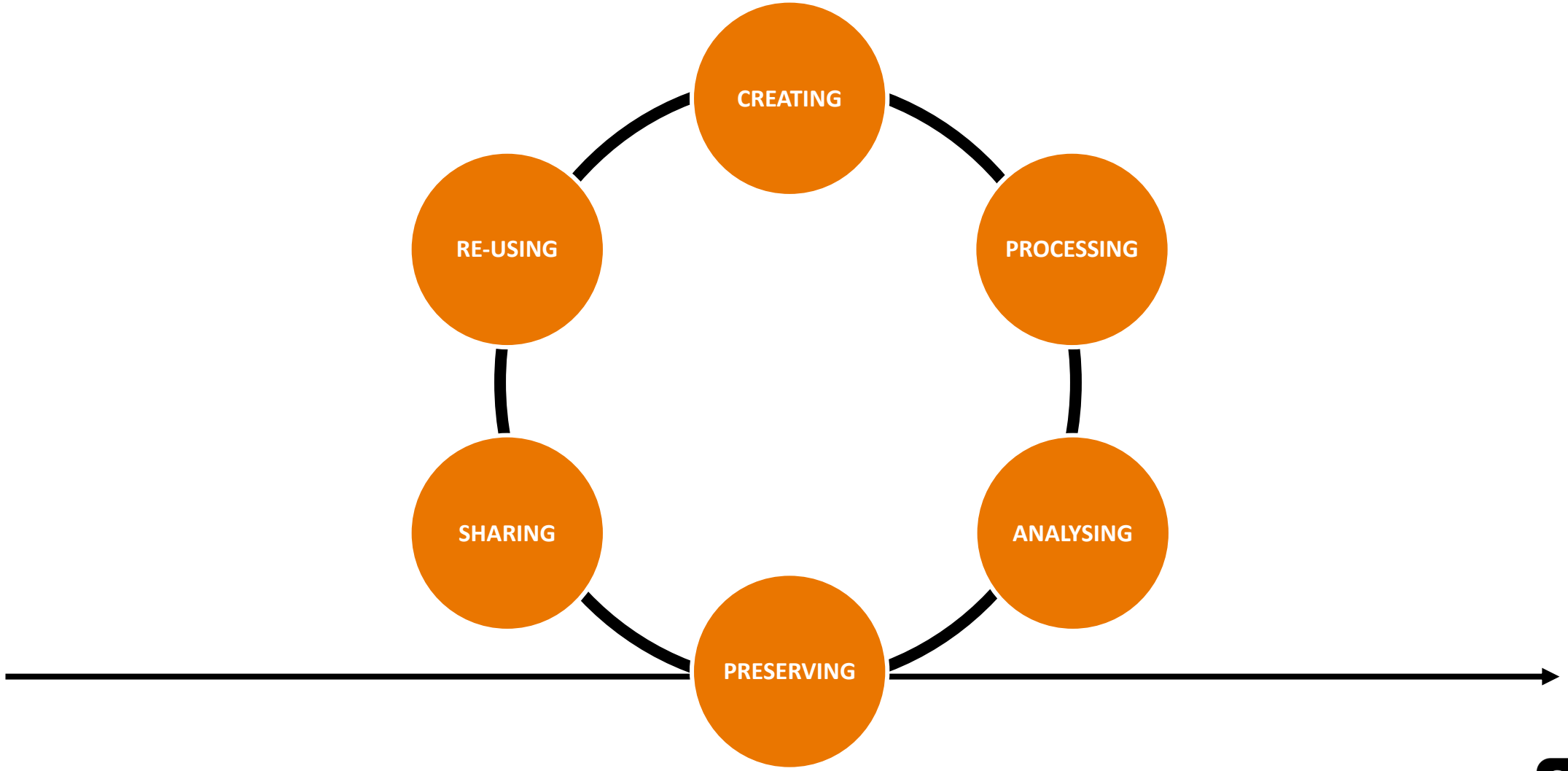


Data Life Cycle is described by metadata

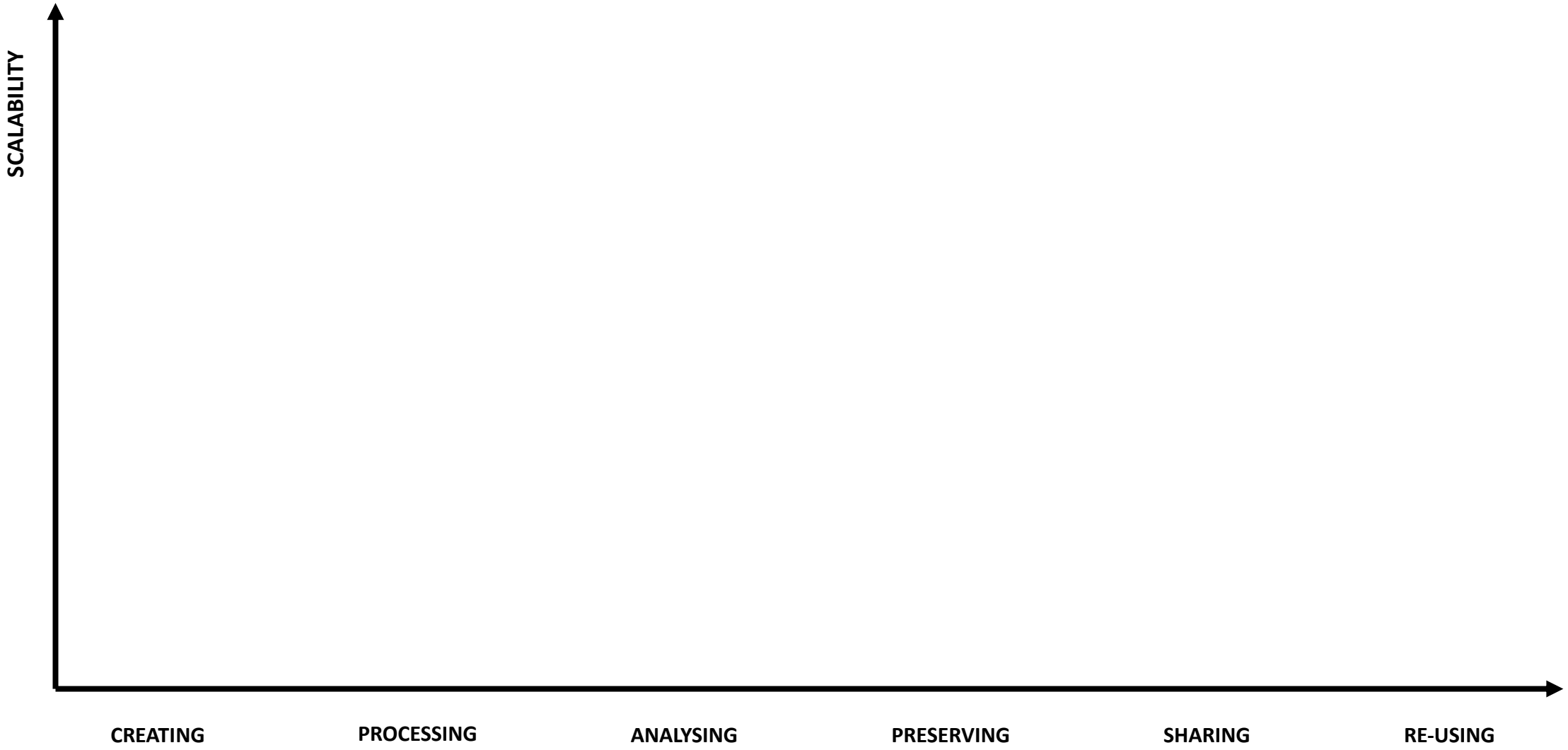




Data Life Cycle – the holy grail of Research Data Management

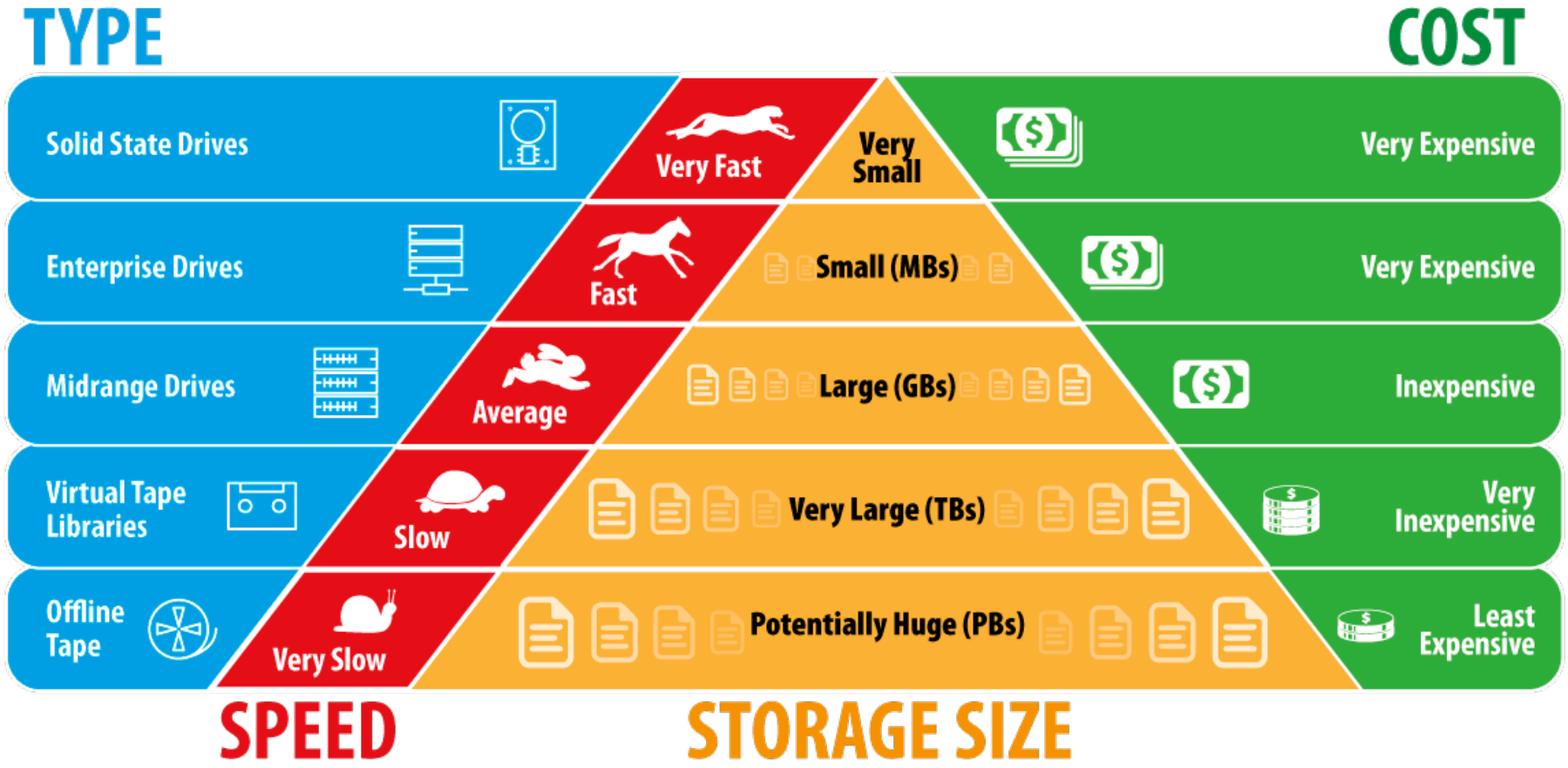


Services at SURF Data Services



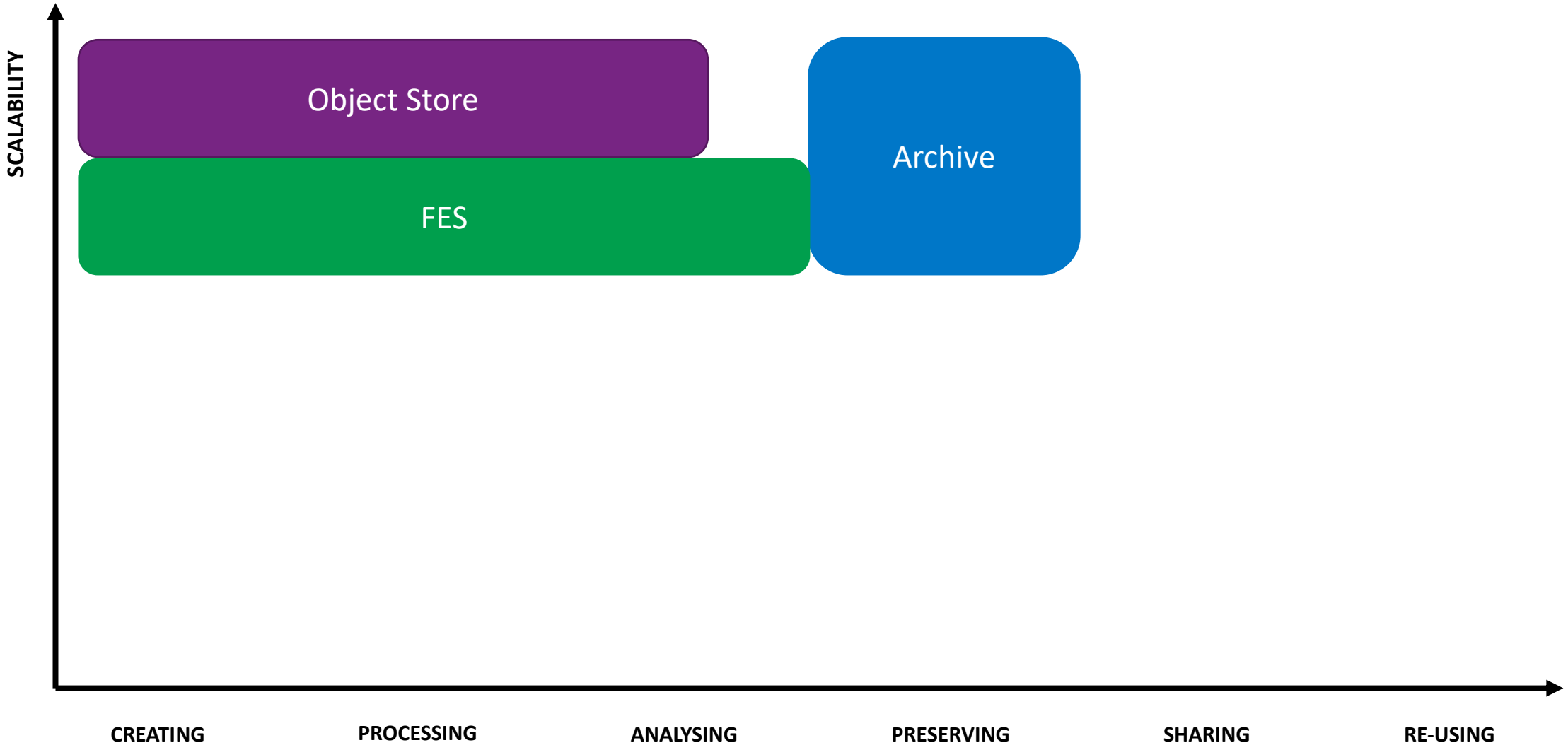
Data storage

- How to determine cost-effective (long-term) storage for your data?



Services at SURF Data Services

Services at SURF Data Services



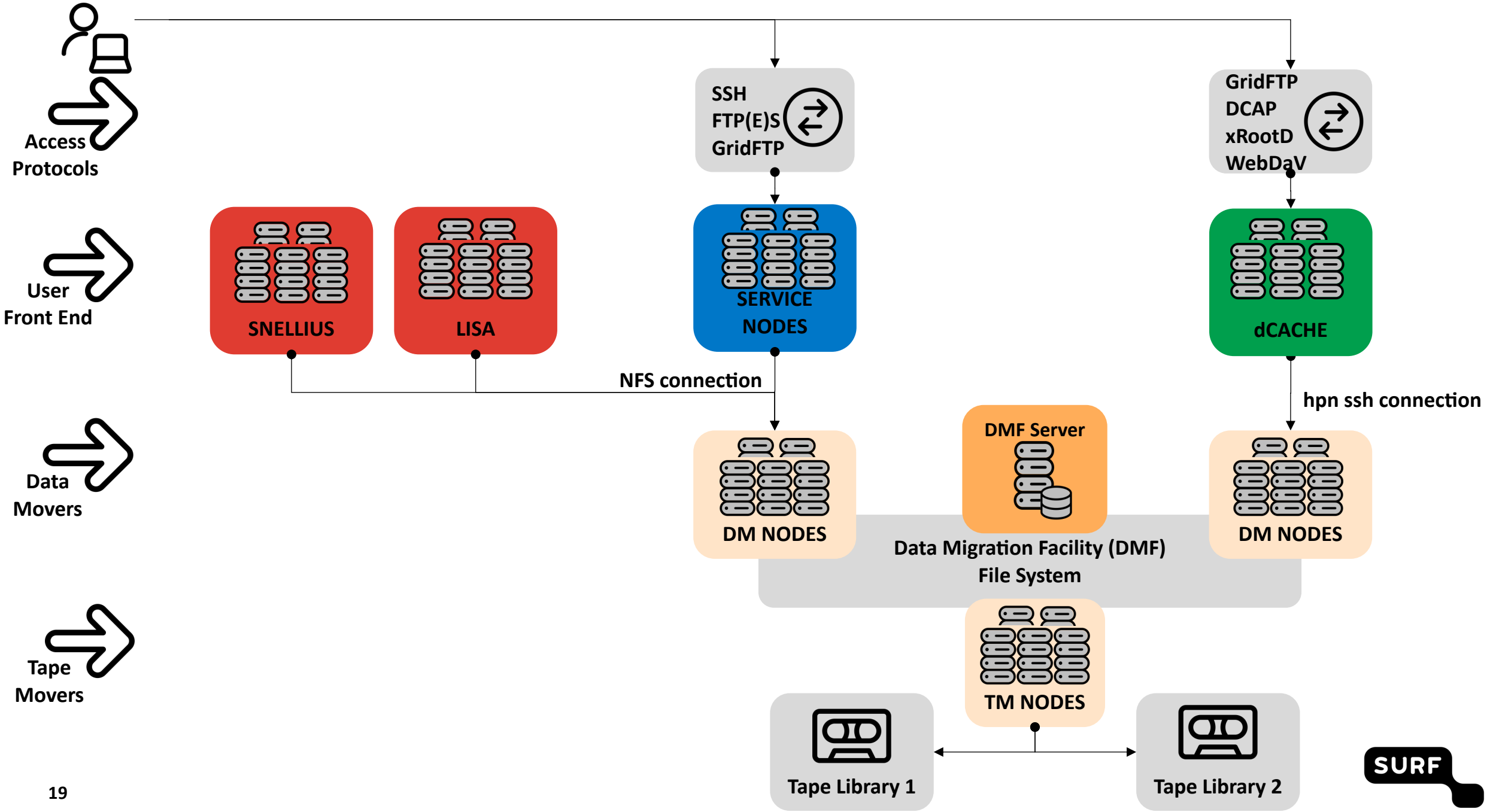
Data Archive: typical user stories

"I need to **cheaply** store my dataset of over 100 TB."

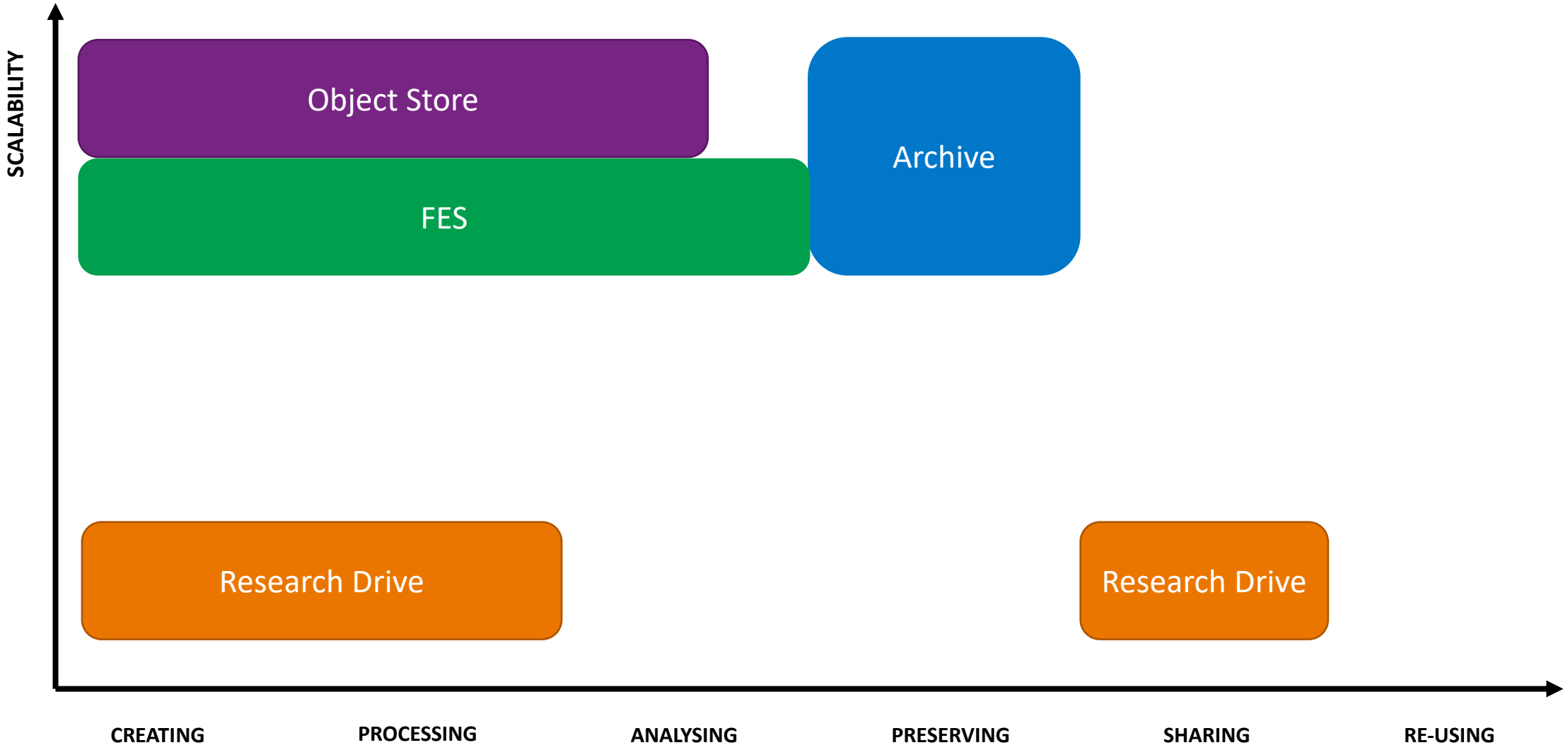
"I need a **temporary storage scale-out** for my compute jobs."

"We want to enable **distributed tiered storage** for delayed processing of data."

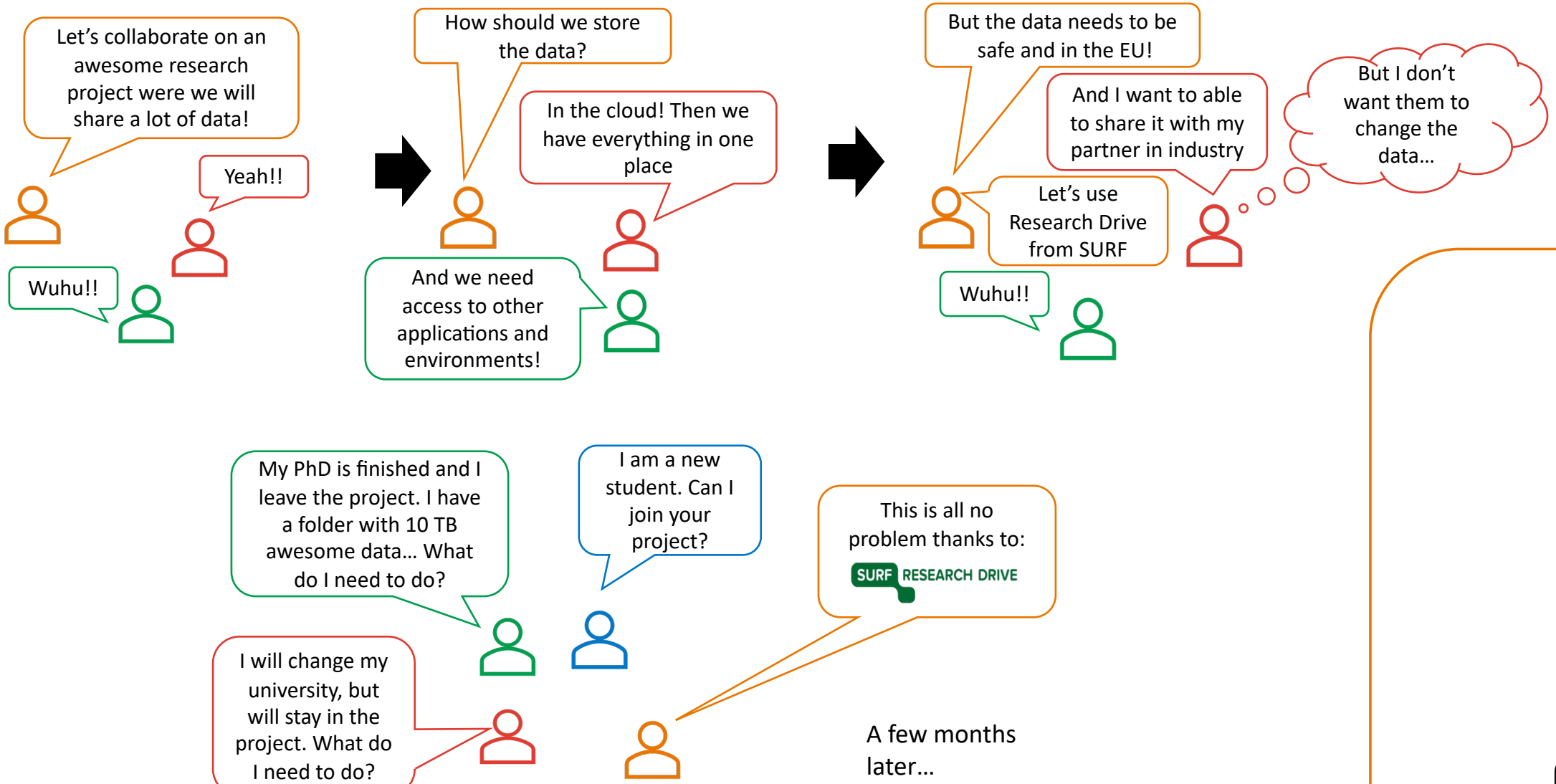
"Our own storage infrastructure will be jettisoned, can you store our **20 PB of data**?"


Access Protocols
User Front End
Data Movers
Tape Movers
SSH
FTP(E)S
GridFTP
GridFTP
DCAP
xRootD
WebDaV
SNELLIUS
LISA
SERVICE NODES
dCACHE
NFS connection
hpn ssh connection
DMF Server
DM NODES
DM NODES
Data Migration Facility (DMF) File System
TM NODES
Tape Library 1
Tape Library 2

Services at SURF Data Services



Why do we need Research Drive?



Research Drive: typical user stories

"I want to **manage our team's research data** during my research project."

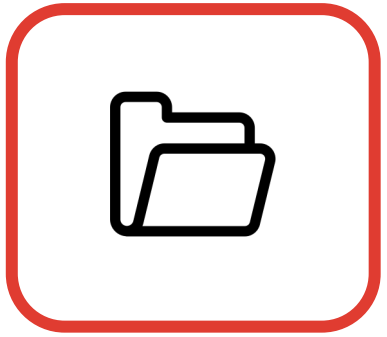
"I want to **share** my research project's data with my colleagues and **external** co-workers."

"I want to **attach external storage and remote services** to manage stored data."

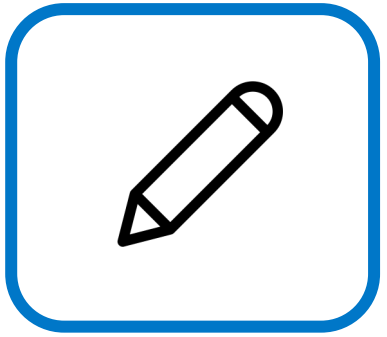
"I need to **manage access and permissions** on a per-file or folder level."

What is reasearch drive?

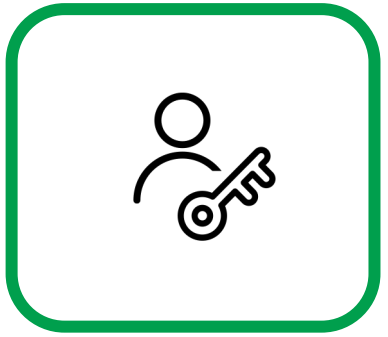
Cloud environment



Sharing files & folders



Collaborative editing

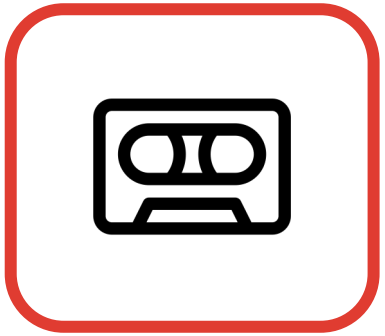


Data Stewardship

Access for Guests

Unlimited storage

What is not research drive?



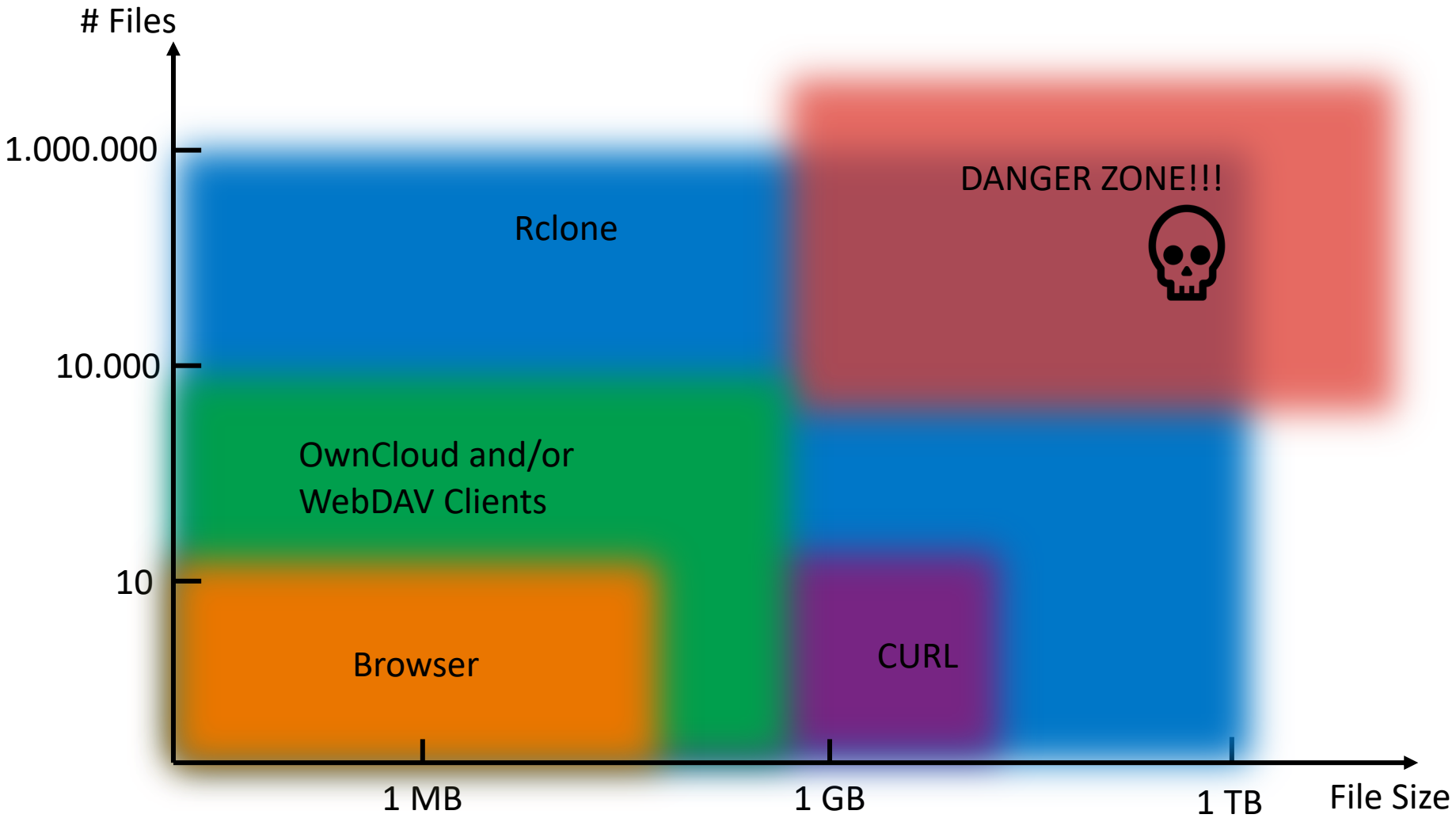
Long-term archiving

Data publishing

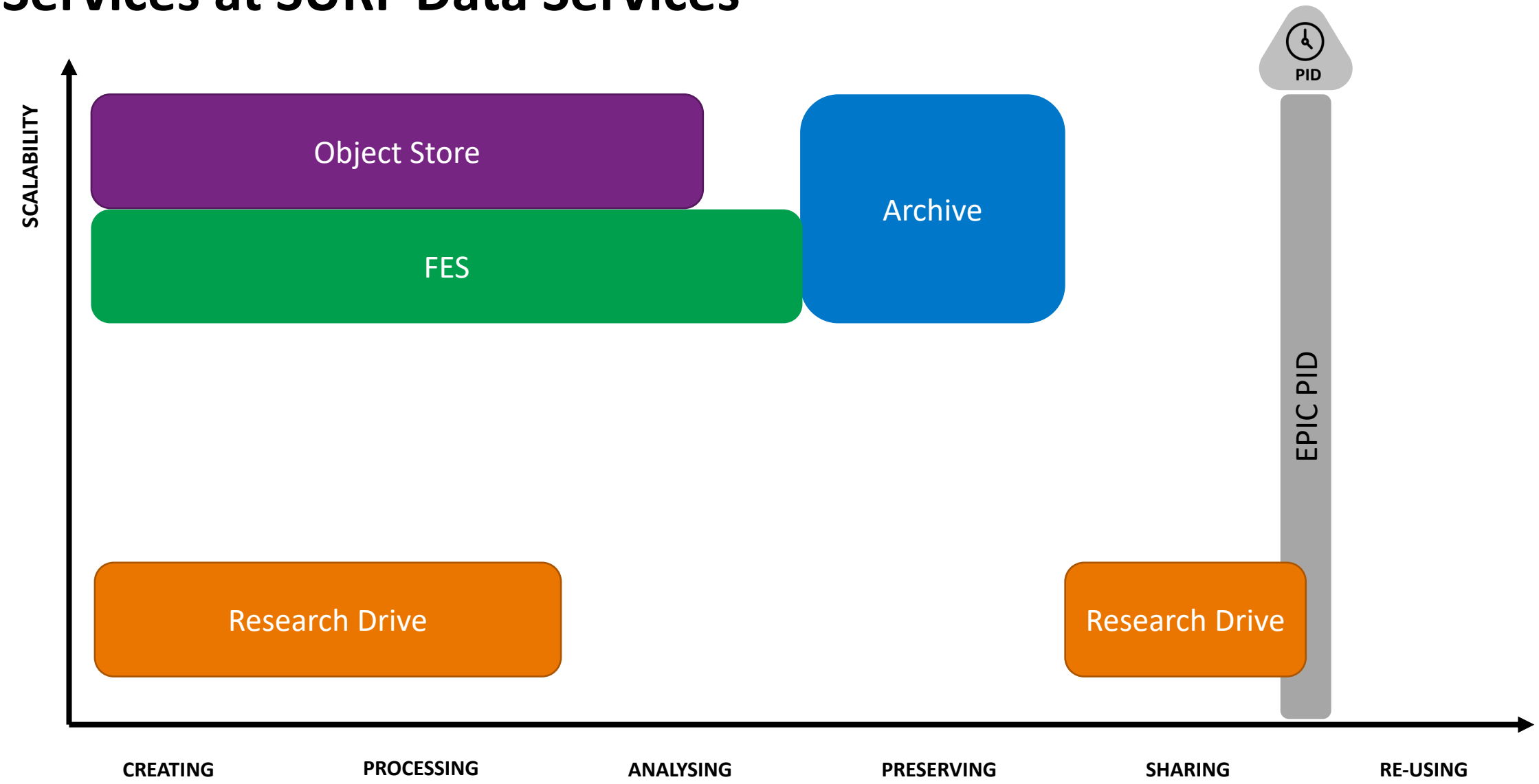
Store and access data

Web Brower Interface	OwnCloud desktop client	WebDaV client
• Data stays in cloud • Access and modification of data via build-in applications or download	• Local copy of data is created on machine • Access and modification of data offline possible • Automatic resyncing	• Data stays in cloud but is visible as file system in computer • Access and modification of data via local applications
• No access for offline users • Limited amount of build in applications • Modified data needs upload again	• ”Out-of sync” data is possible • Large amount of data should not be synchronized • Local copy of data is created on machine	• No access for offline users • Limited capacity when it comes to many files

Getting your data into Research Drive



Services at SURF Data Services



EPIC PID: typical user stories

"I want to give my dataset and files a **persistent identifier** for use in publications."

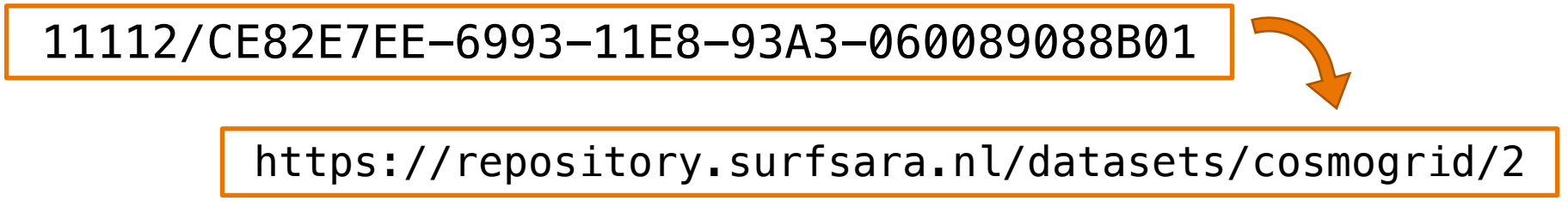
"My PID hosting does not **scale**, can EPIC PID handle all my PIDs?"

"I want SURF to **host** all my **existing** PIDs."

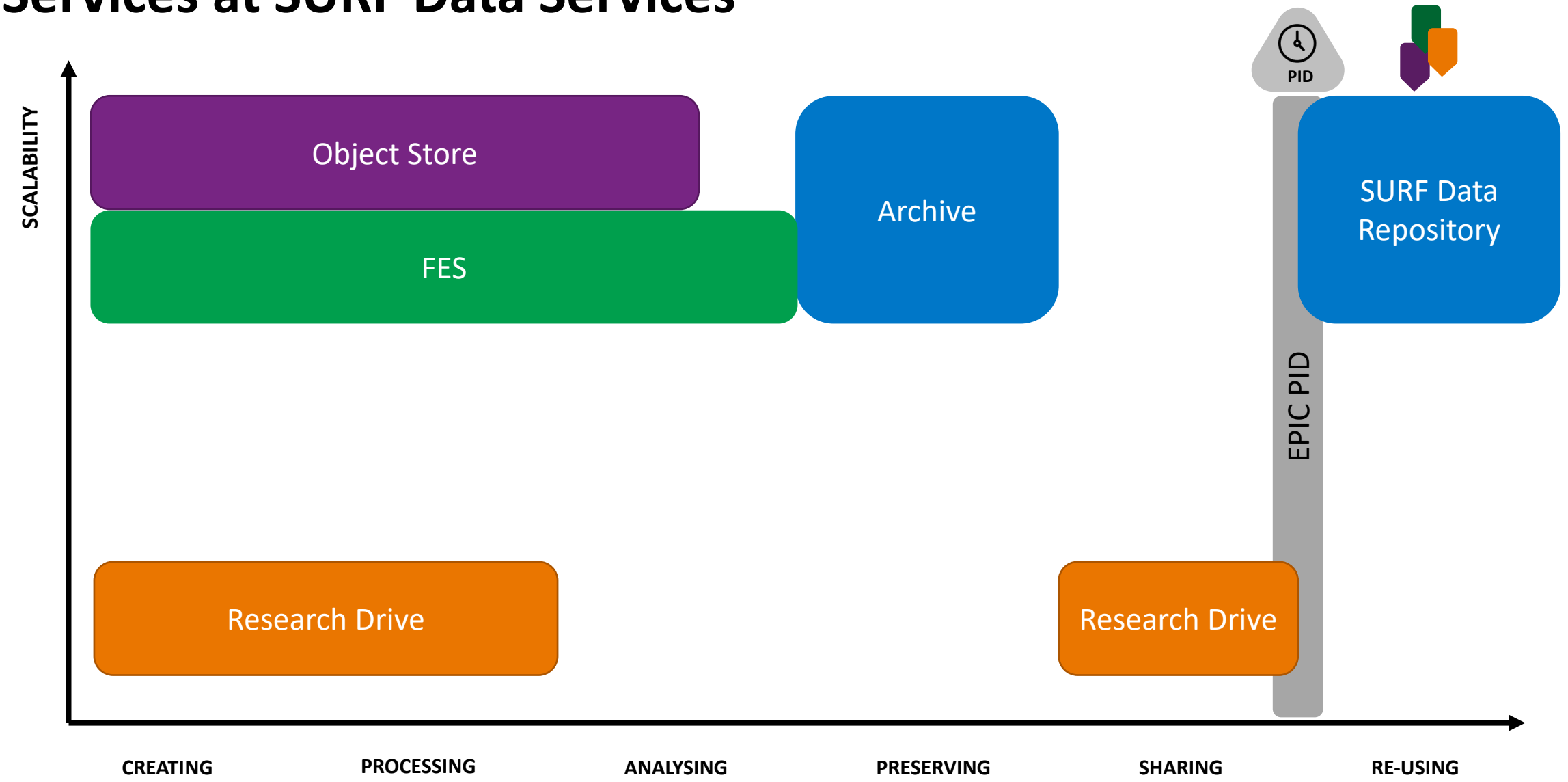
"We want all the **objects** in our museum to have a PID."

EPIC PID

- High-availability service for **persistent identifiers**
- Resolve **unique string of characters** into a URL or similar location together with additional metadata through the global Handle service
- Administer your PIDs using your own **prefix**



Services at SURF Data Services



Data Repository: typical user stories

"I want to **publish** my very large dataset in a **trustworthy** repository**.**"

"I need **automated publishing** of my datasets but only with **specific** people."

"My dataset should have a **PID** and **metadata** attached, and should be publicly **available** for 10 years."

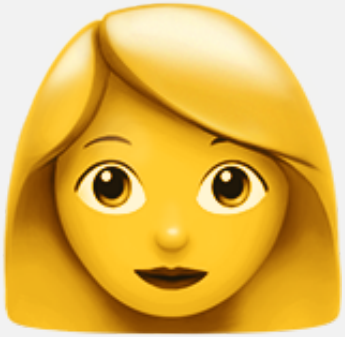
"I want to manage my own **collections, schemas, and publications.**"

Data Repository: why?

Cost-effective long-term preservation solution - Publish datasets of **any size**, even up to PBs!

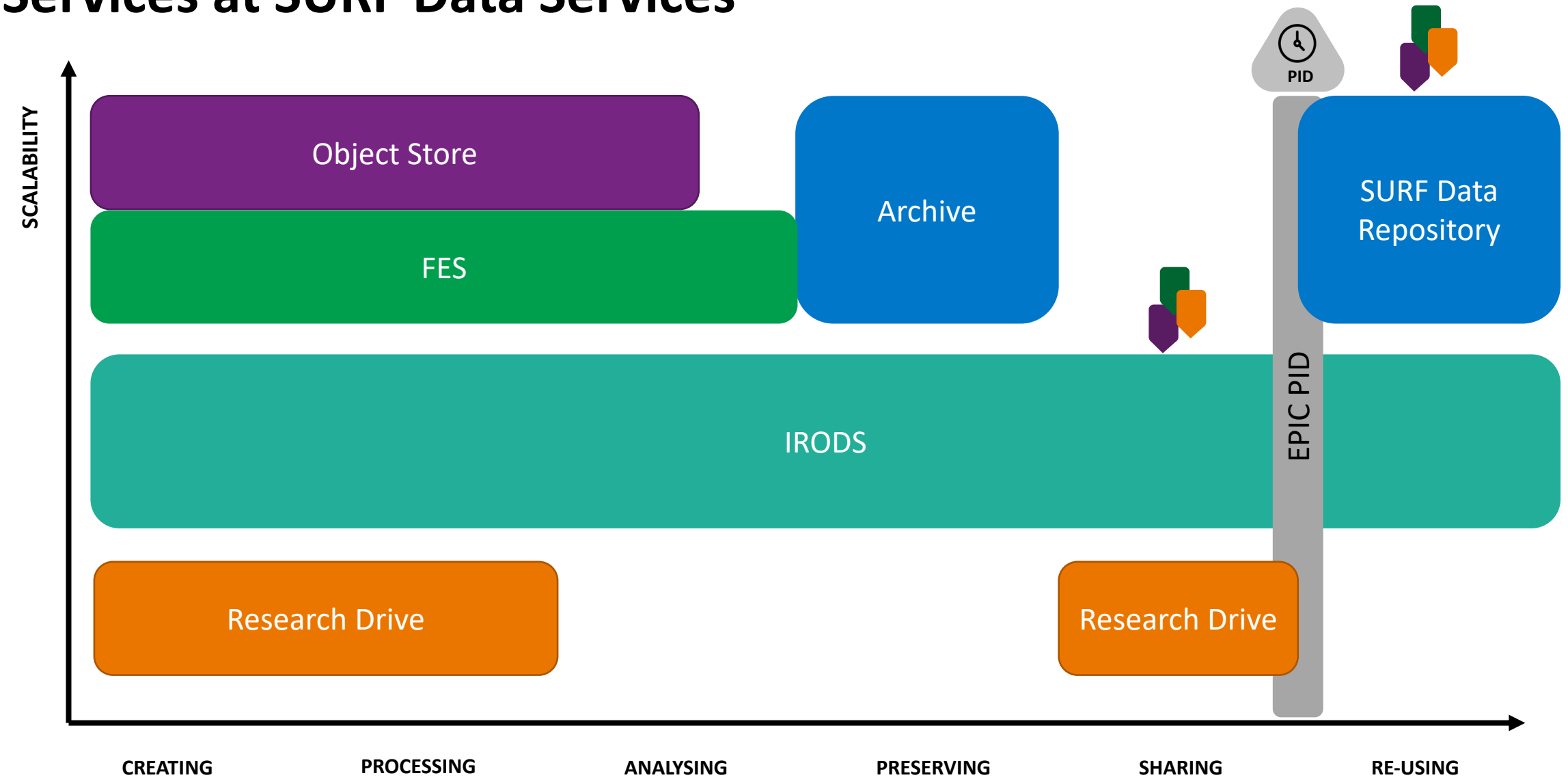
Automated **massive download** of datasets via a **REST API.**

Self-service platform for deposits, collections and metadata schemas.



Close proximity to (compute) infrastructures.

Services at SURF Data Services



iRODS: typical user stories

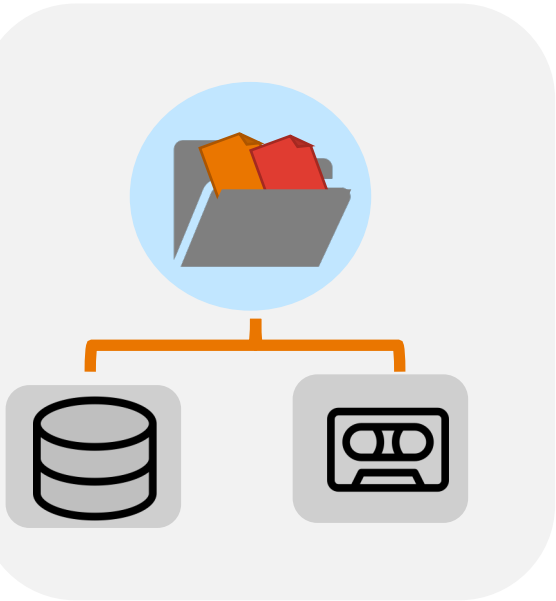
"I want to **automate the management** of research data within my institute."

"I need to manage my research data stored **at different locations**."

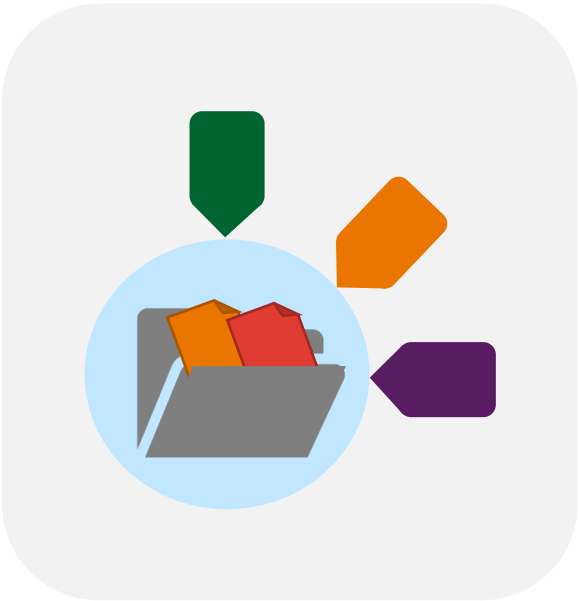
"I want SURF to **host my data management** solution."

"I need a **secure synchronized remote copy** of all our institute's research data."

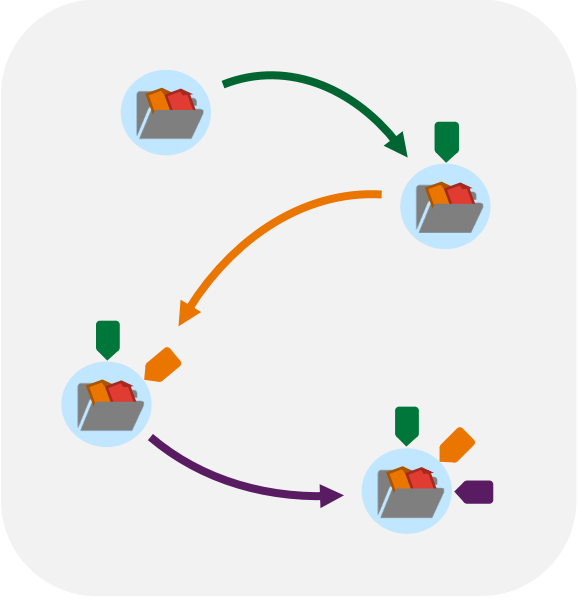
What is iRODS?



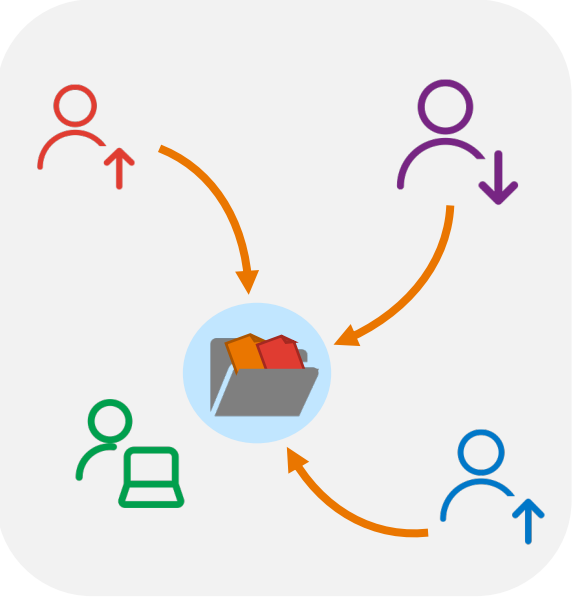
Unified storage of disk and tape



Metadata for data discovery

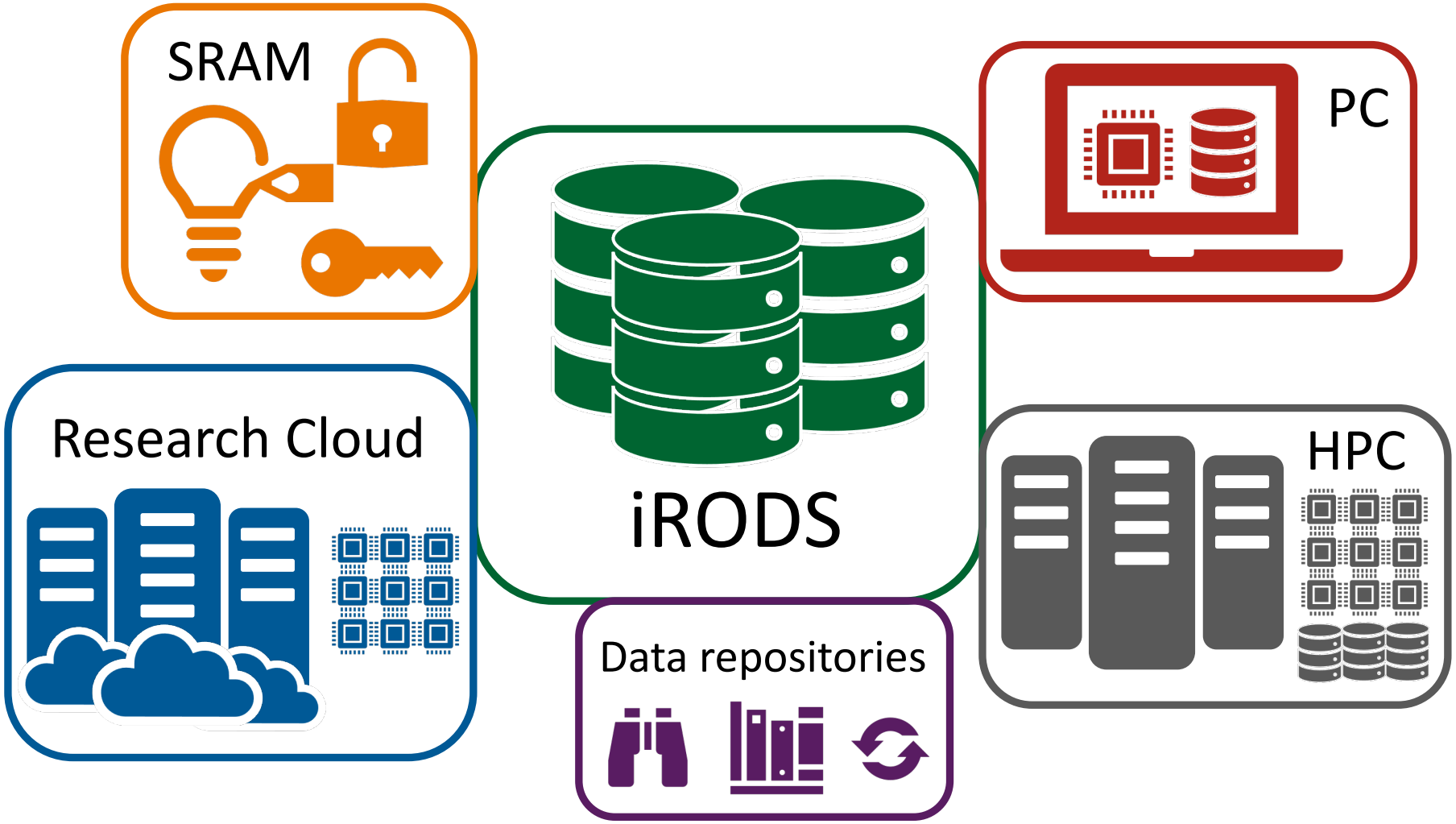


Secure collaboration and auditing



Rule engine to automate policies

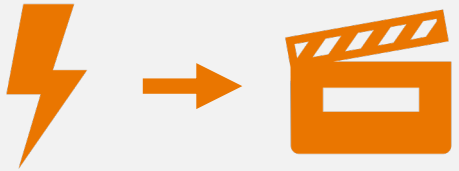
iRODS is designed to connect to research



iRODS rule engine automates workflows and enforces policy

User invocation

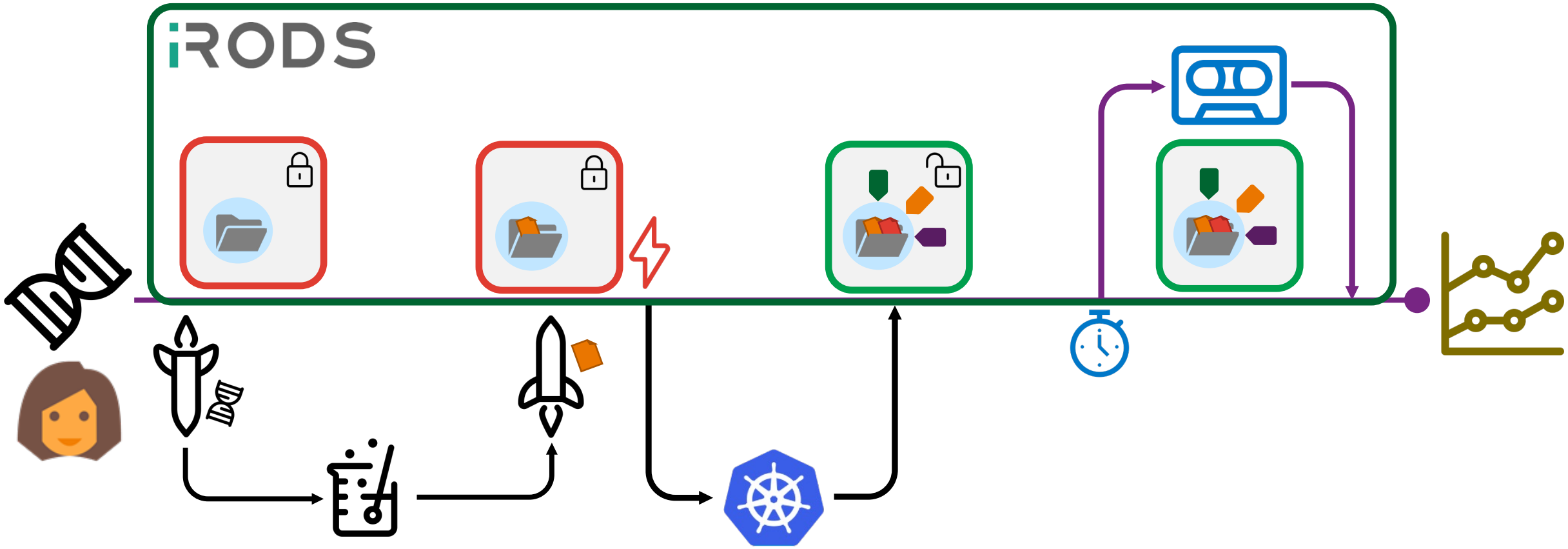
Time based



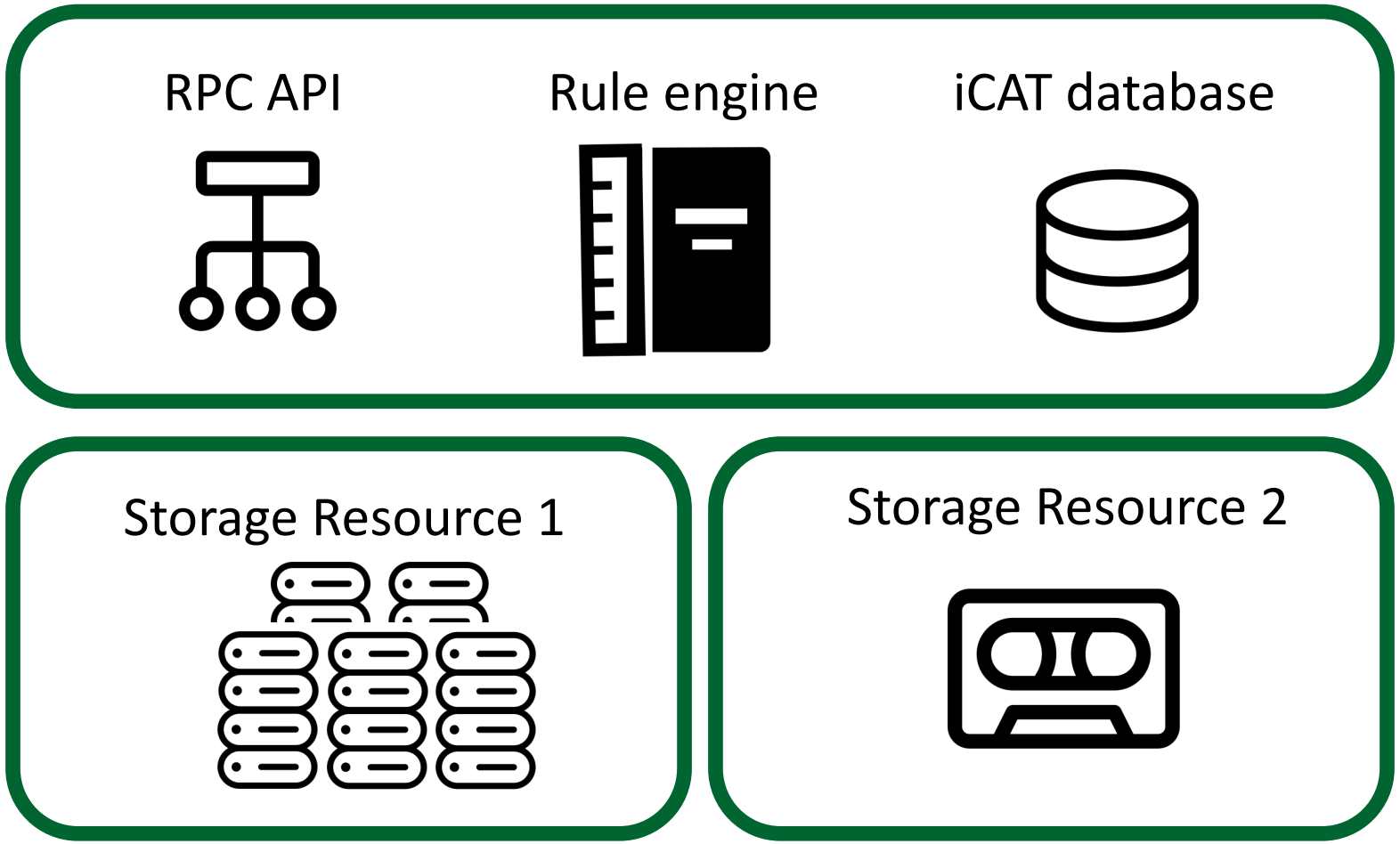
Event based

Different languages

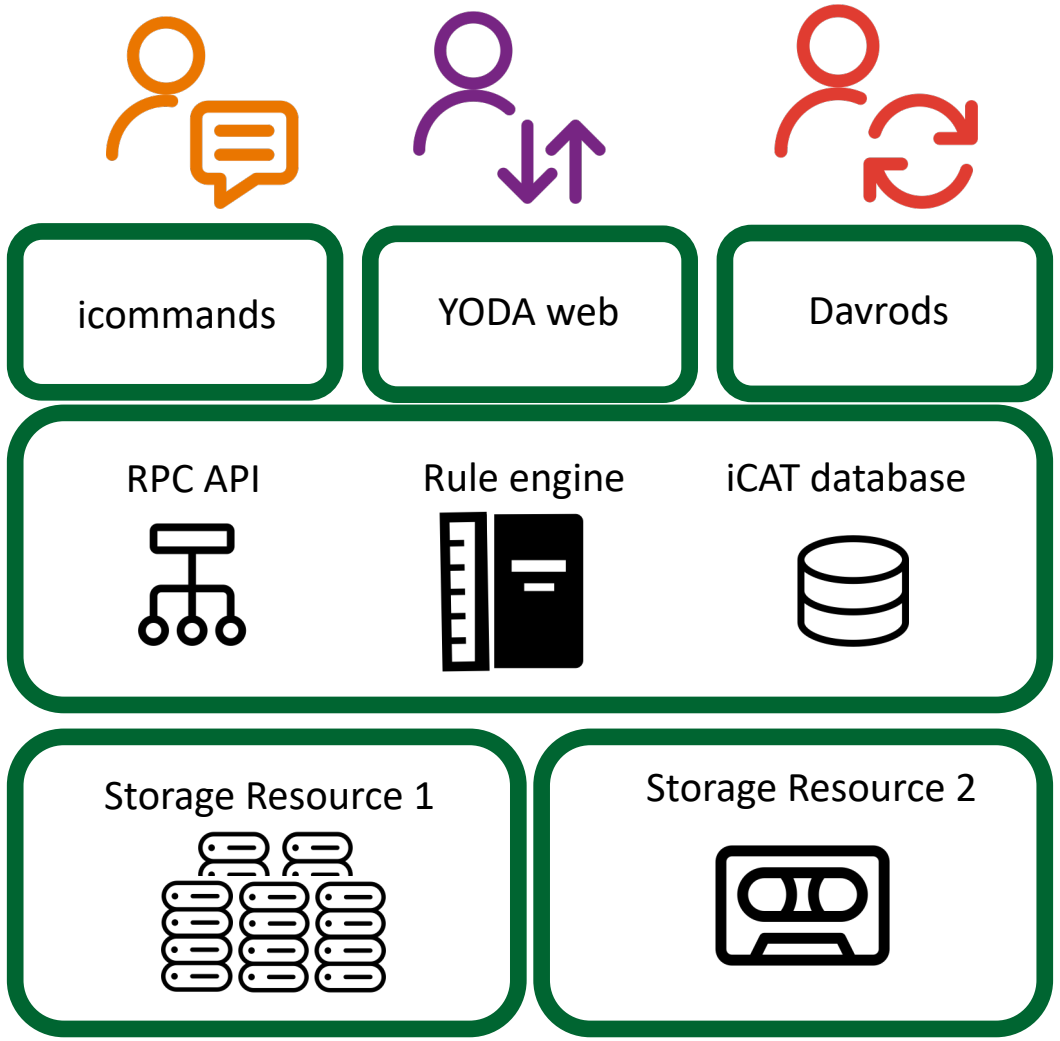
iRODS User journey: DNA sequencing (real use case)



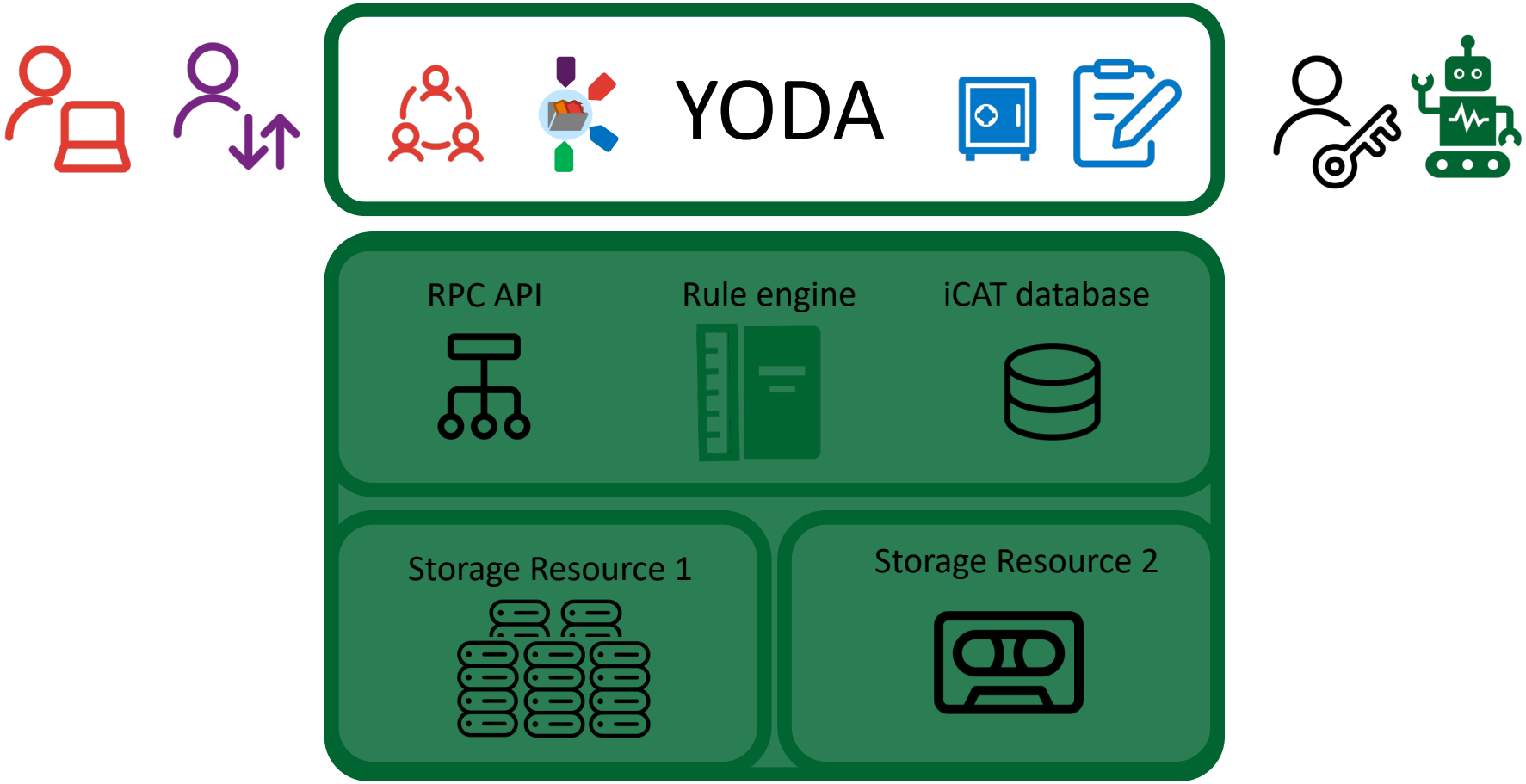
iRODS components in one administrative domain



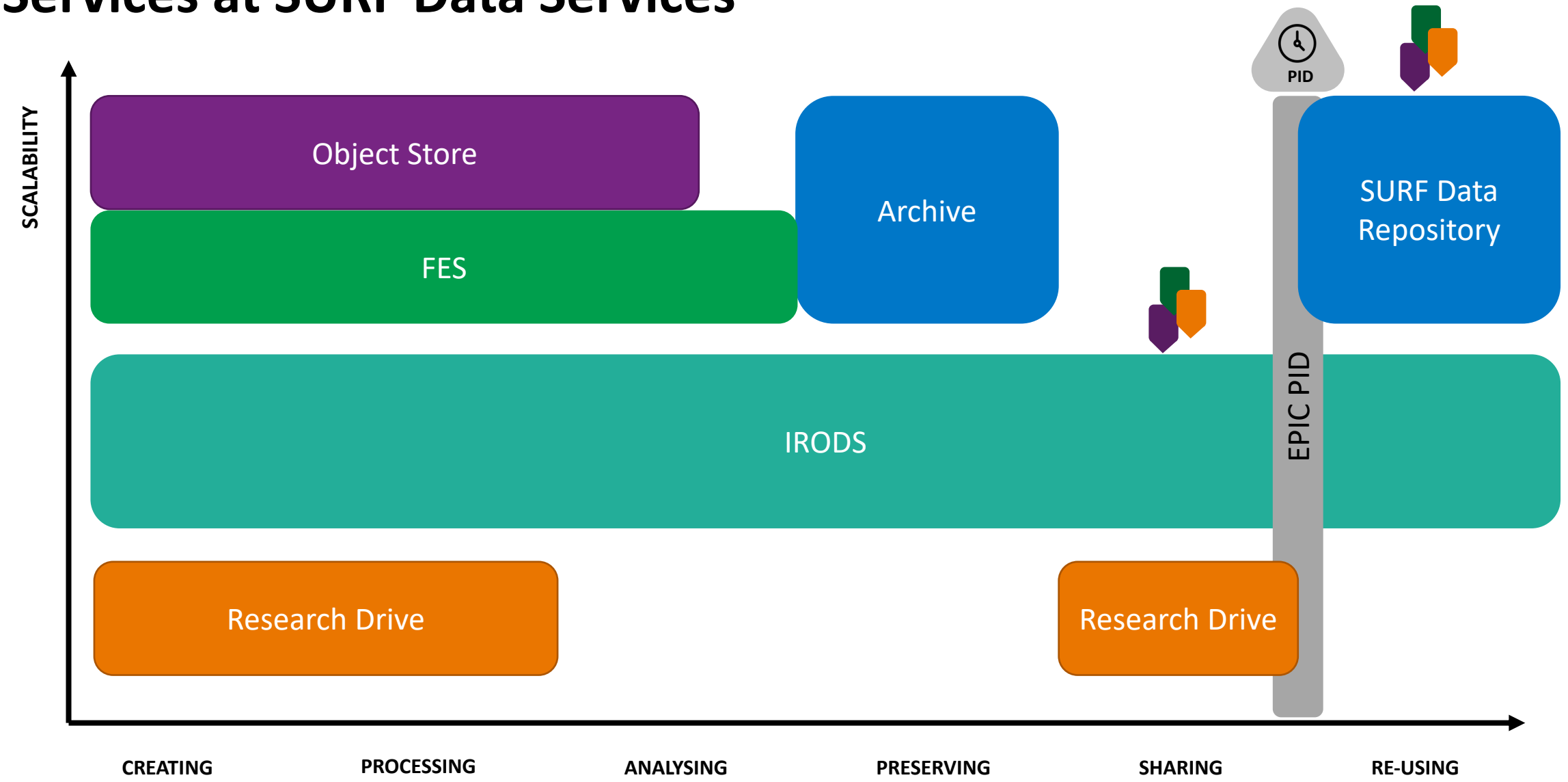
iRODS user interfaces development within SURF



iRODS with YODA web portal and YODA rules installed



Services at SURF Data Services




QUESTIONS?
SURF

What are we covering

- 09:00 – 09:45: Data management with SURF
- 10:00 – 11:00: Hands-on: Data processing with Lisa and Research Drive
- 11:00 – 12:30: Hands-on: Data management with Yoda

edu.nl/xgvtg

HAPPY SHARING!

SURF Data Services

E-mail: info@surf.nl

www.surf.nl

Social media: @surf_nl

Driving innovation together

Driving innovation together