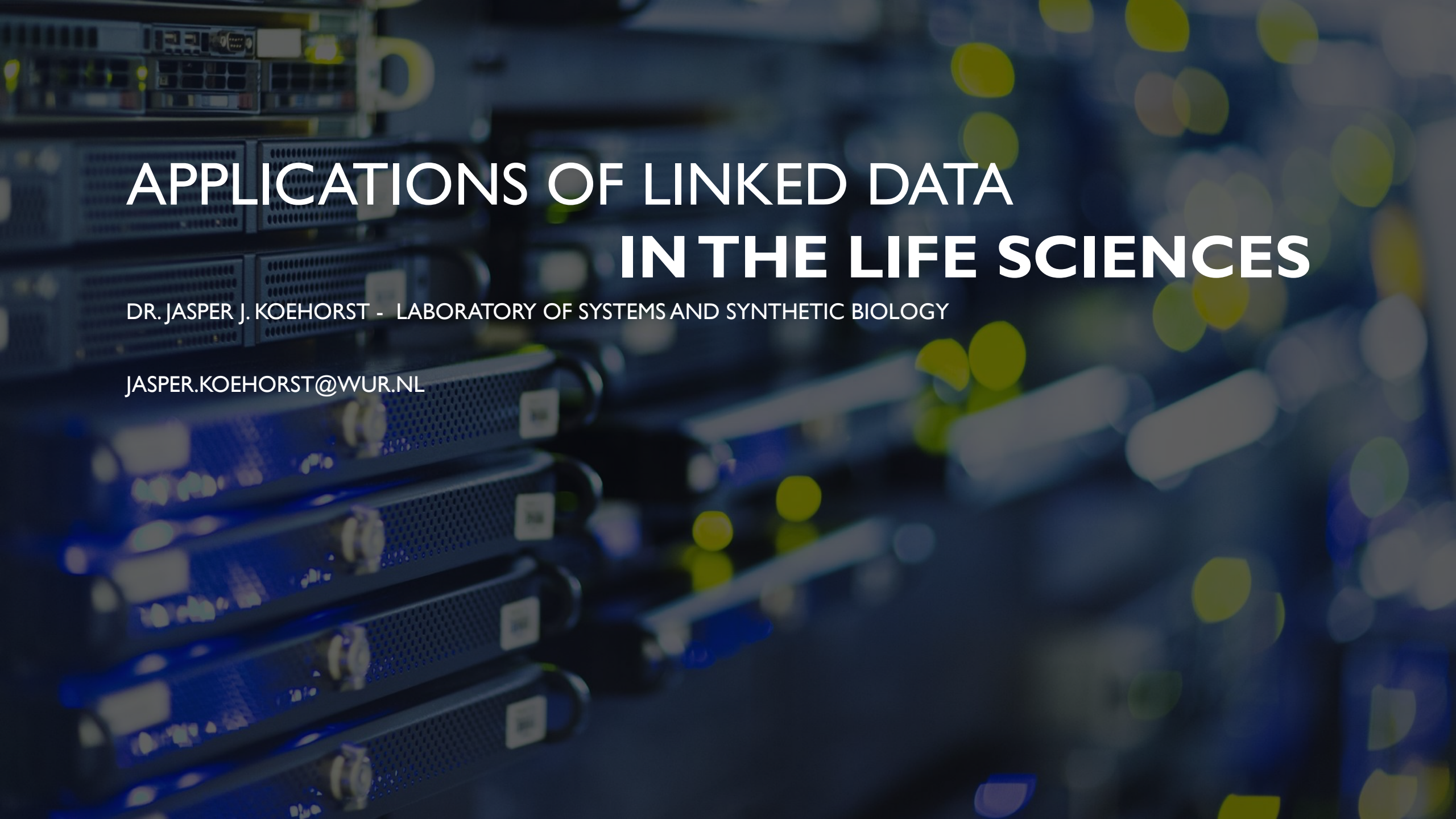
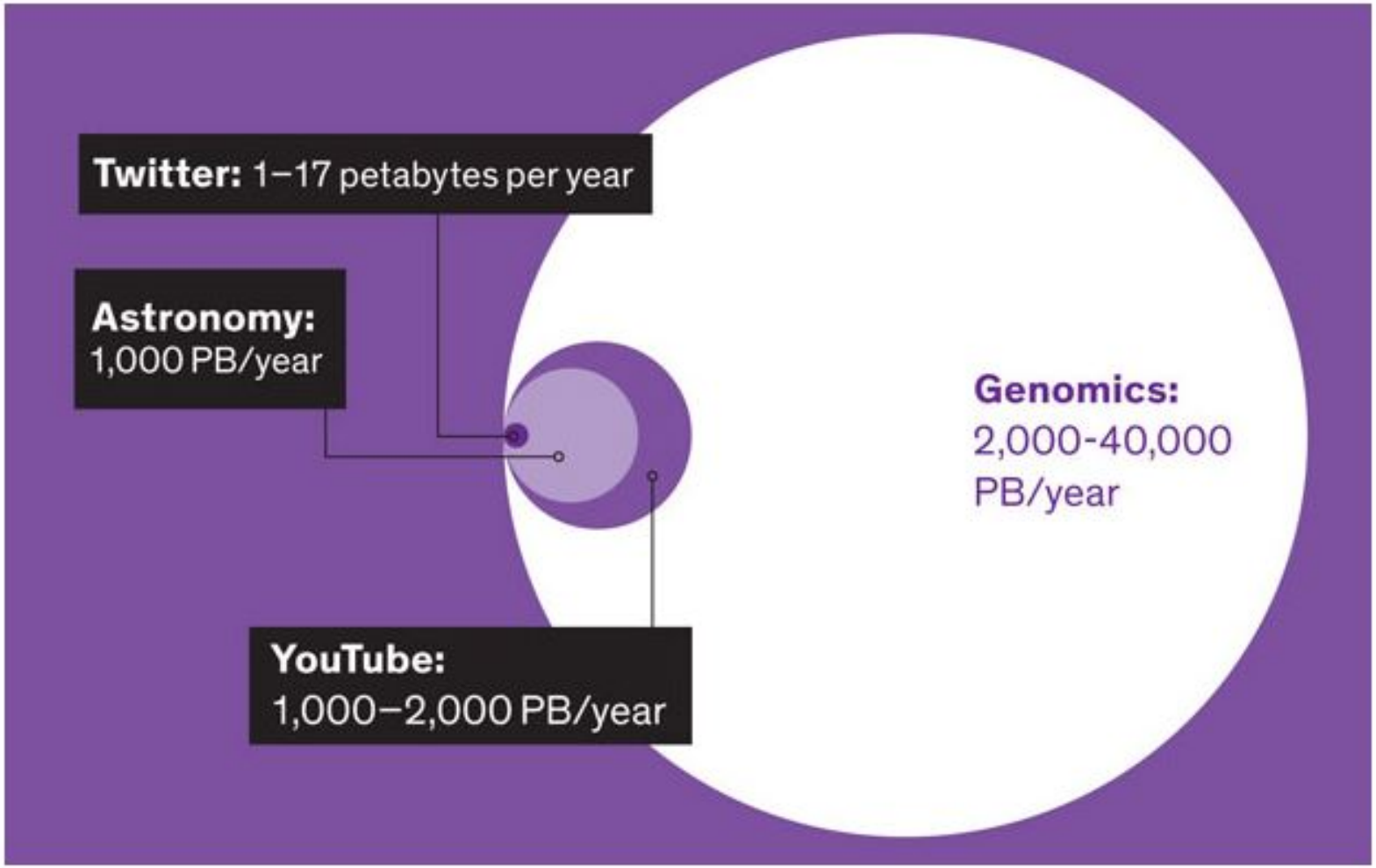


APPLICATIONS OF LINKED DATA IN THE LIFE SCIENCES

DR. JASPER J. KOEHORST - LABORATORY OF SYSTEMS AND SYNTHETIC BIOLOGY

JASPER.KOEHORST@WUR.NL

Projected annual storage in 2025

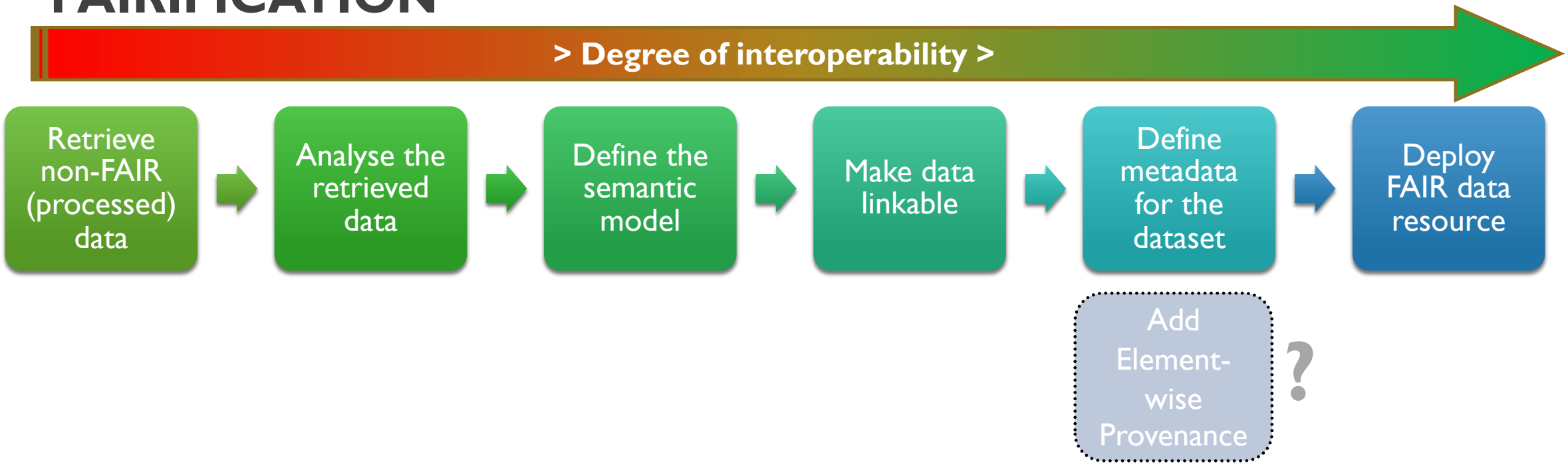


Source: "Big Data: Astronomical or Genomical?" PLoS Biology, 7 July 2015.

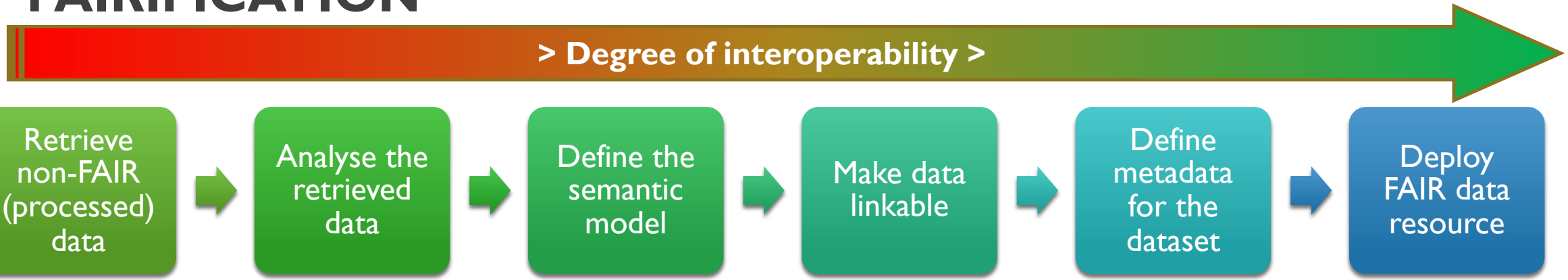
GENOMICAL AMOUNT OF DATA

- **Not only a lot of data**
- **Extremely diverse**
- **Interoperability problems: Inconsistencies in describing genomic information**
- **Current data standards are not sufficient.**
- **The process of FAIRification is complex and costly**
- **99.99% of information obtained is computationally predicted**

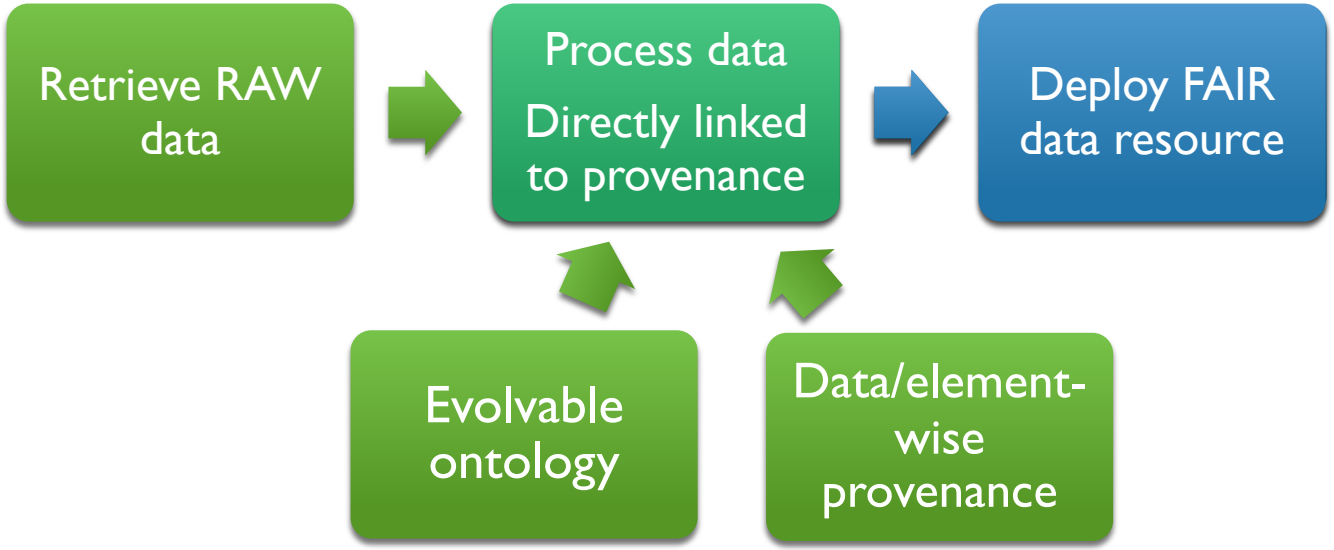
FAIRIFICATION



FAIRIFICATION



FAIR BY DESIGN



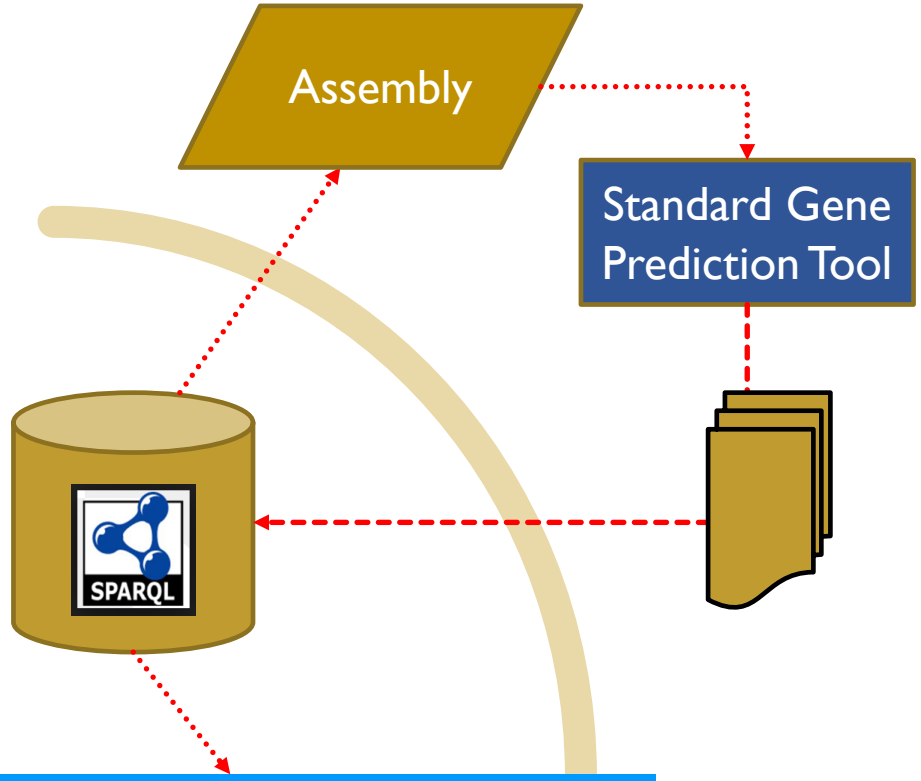
CLASSICAL ANALYSIS WORKFLOW: **DATA MEETS THE TOOL**

IN EACH STEP A DECISION IS MADE BASED ON A PRE-DEFINED THRESHOLD

THE ACTUAL ELEMENT-WISE PROVENANCE

- E.G CONFIDENCE SCORES - IS LOST
- NO STATISTICAL ANALYSIS ON METADATA POSSIBLE
- CHANGING THE THRESHOLD VALUE = **RECOMPUTE**

FAIR BY DESIGN SEMANTIC ANALYSIS WORKFLOW:
TOOL MEETS THE DATA



FAIR Data Sharing – Linked Data
Triple store can be queried with SPARQL

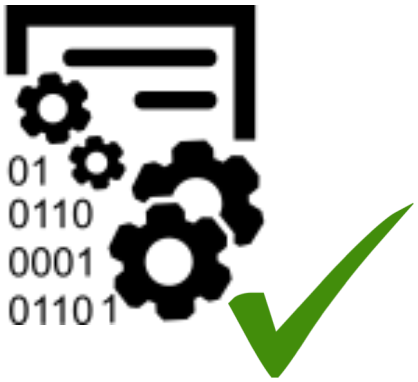
DATA/ELEMENT-WISE PROVENANCE AUTOMATICALLY AND DIRECTLY LINKED TO THE PREDICTION

Algorithm = ..
Version = ..
Run date = ..
……..
Genomic location
Confidence score
……..

GBOL ontology for Interoperable genome annotation

DYNAMIC THRESHOLD SELECTION (AFTERWARDS)
MULTIPLE (SIMILAIR) TOOLS CAN BE COMPARED

FROM **FAIRIFICATION** TO **FAIR BY DESIGN**



RAW DATA

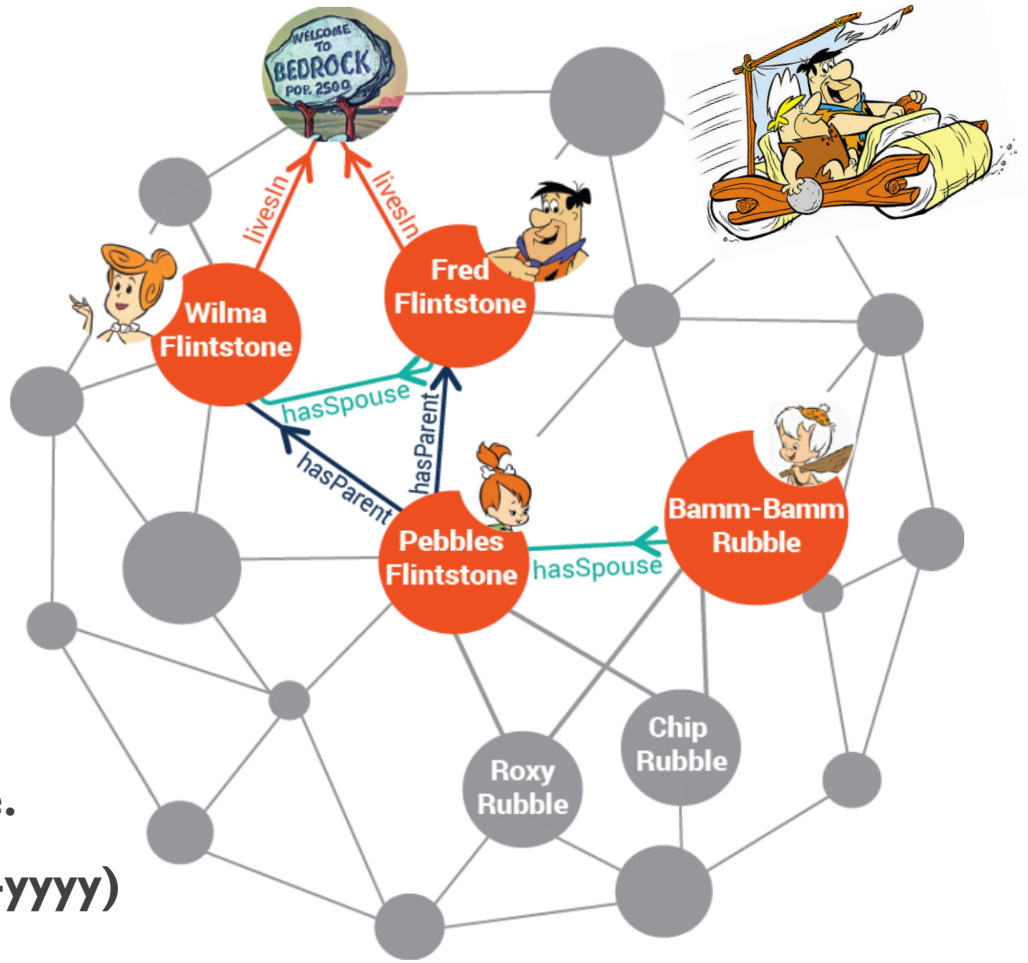
ONTOLOGIES

APPLICATION PROGRAMMING INTERFACE

SEMANTIC FRAMEWORK & LINKED (RDF) DATA MODEL

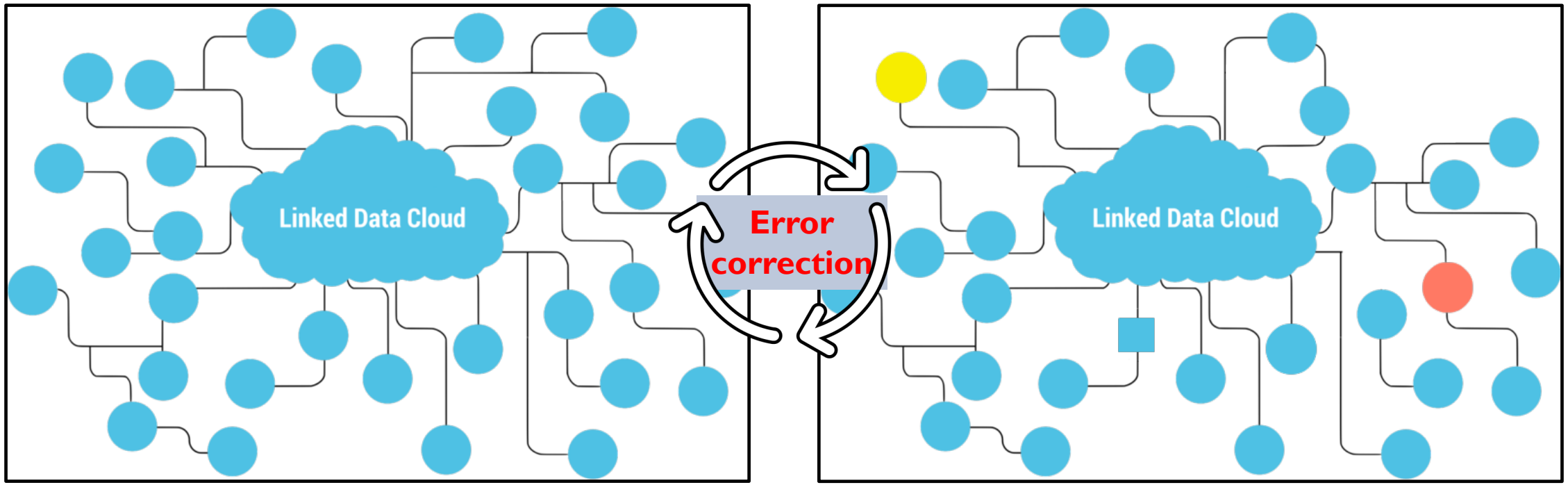
HOW TO ENSURE A HIGH DEGREE OF INTEROPERABILITY USING A GENOMICAL AMOUNTS OF DATA: **ONTOLOGIES**

- **DEFINITIONS, DEFINITIONS AND MOST IMPORTANTLY DEFINITIONS**
  - **MINIMAL INFORMATION MODEL REQUIRED FOR INTEROPERABILITY**
  - **ONTOLOGY AND CONSTRAINED RELATIONS**
  - **SHEX DEFINITION** (a language for validating and describing the RDF data according to the ontology)




```
<cartoonfigureShape> {      # A cartoonfigure (e.g Fred)  has:
 :name xsd:string+;         #   one or more names
 :birthdate xsd:date?       #   and optionally a  (single)  birthdate.
}                                    in a specific format (e.g. dd-mm-yyyy)
```

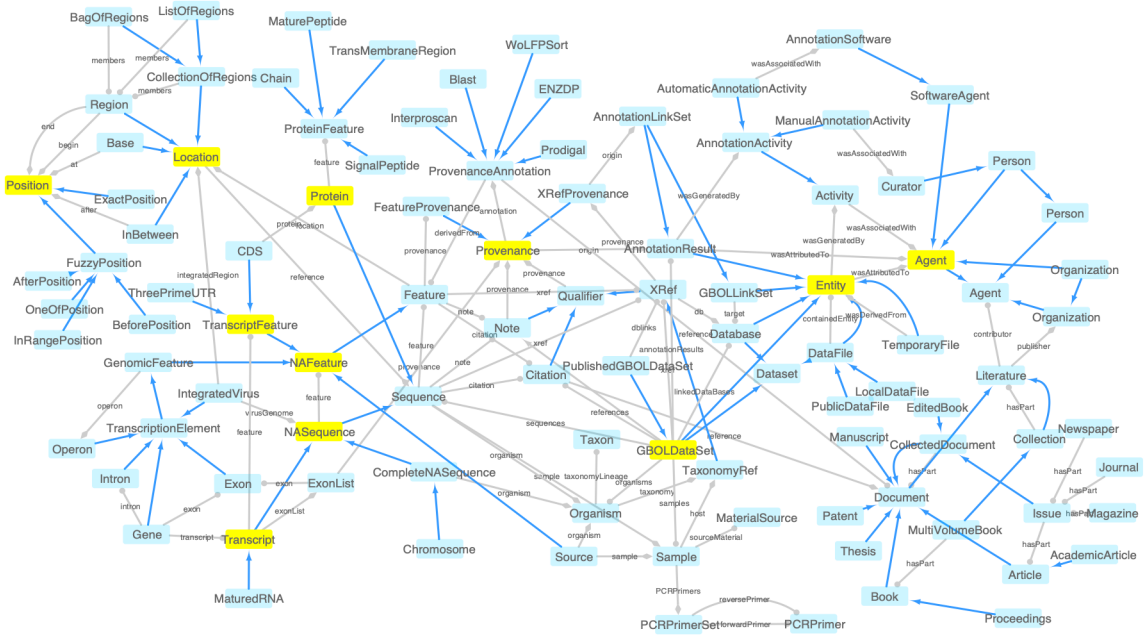
THE DYNAMIC NATURE OF SEMANTIC WEB

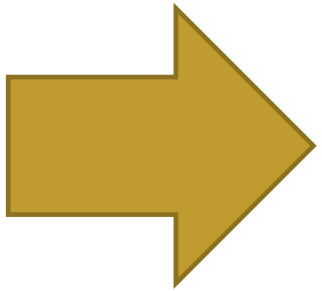


EXPECTED

OBSERVED

EMPUSA: ONTOLOGY VALIDATOR & CODE GENERATOR

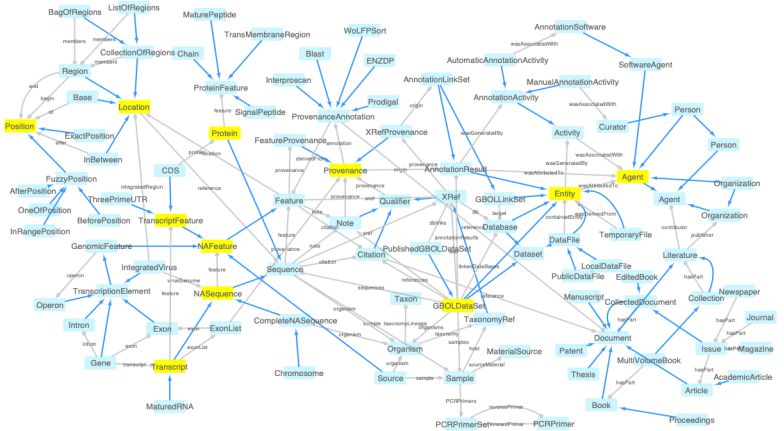


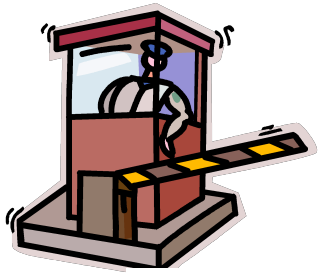


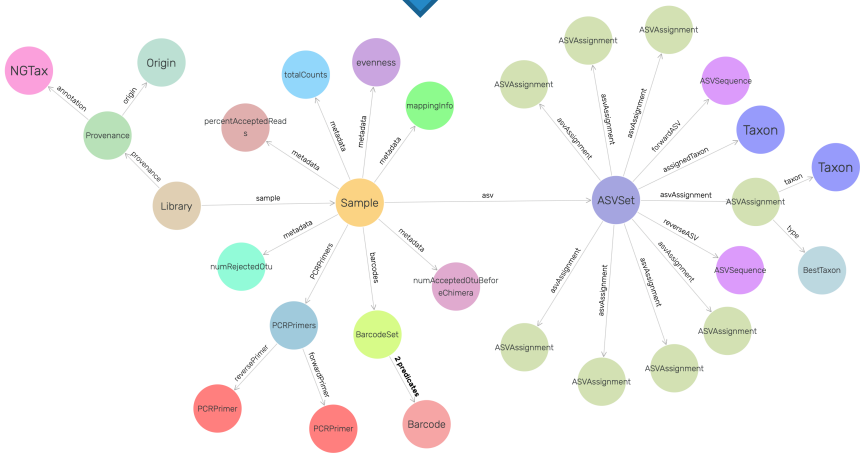
APPLICATION PROGRAMMING INTERFACE

EMPUSA GENERATES THE CODE THAT CAN ACT AS A GATEKEEPER FOR FAIR LINKED DATA GENERATION

FAIR BY DESIGN
FROM DATA TO LINKED DATA



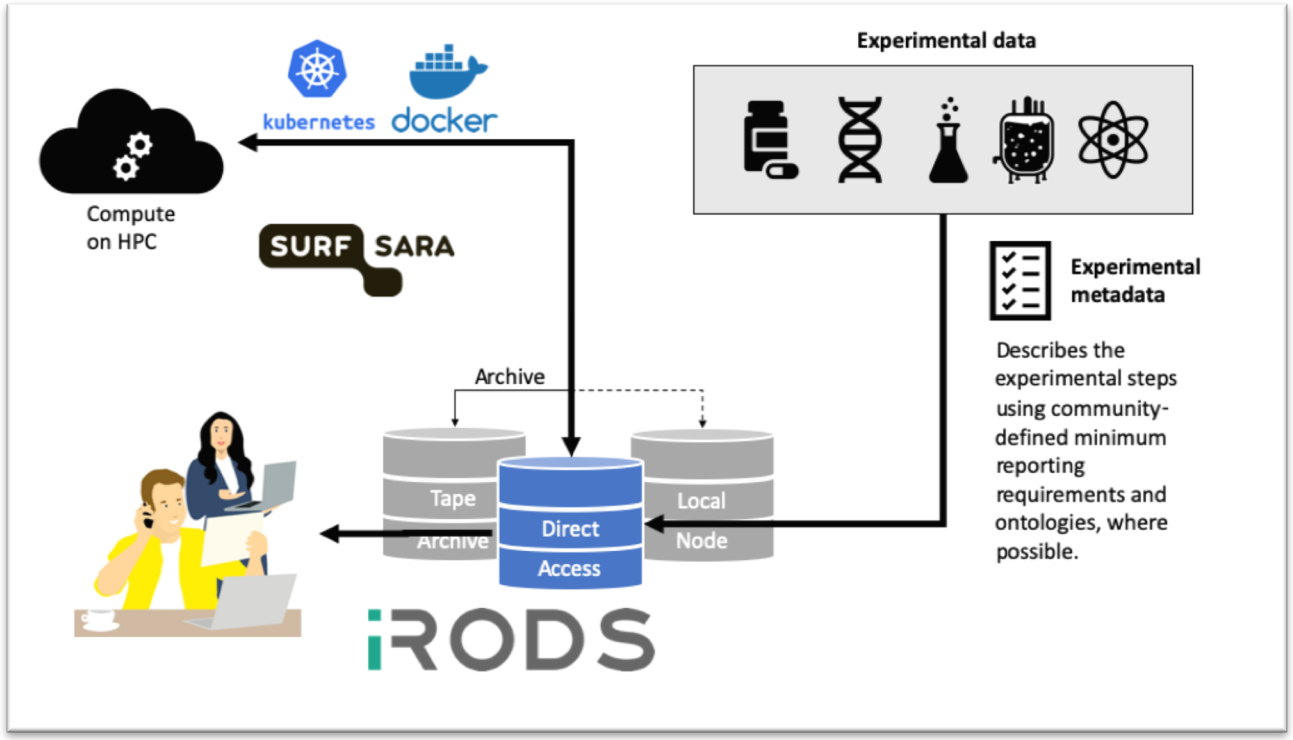




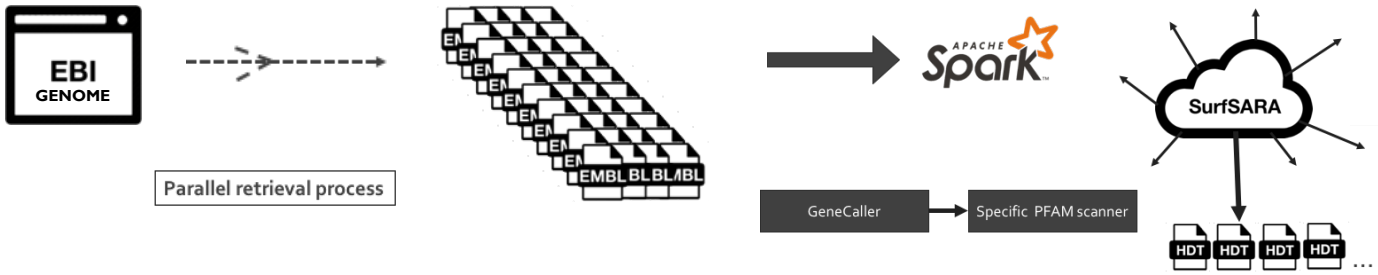
Validated linked (meta-)dataset

APPLICATION: PROJECT MANAGEMENT

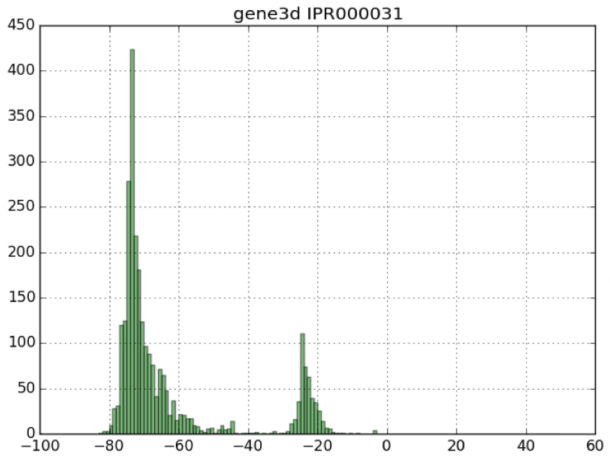
- According to the P-ISA principles
  - Project, Investigation, Study, Assay
  - All (meta-)data is available as linked data
  - Assay: compute results and provenance translated to RDF and embedded in the system

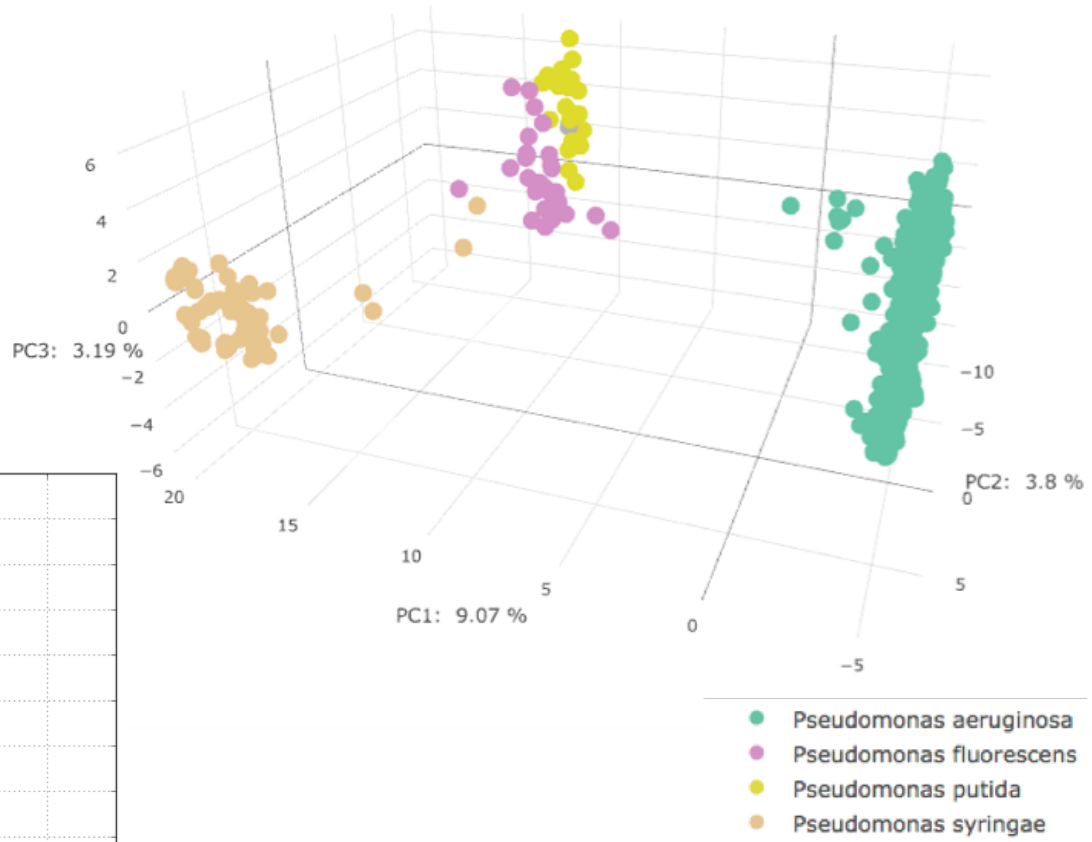


APPLICATIONS: SAPP, FUNCTIONAL GENOME ANNOTATION & COMPARATIVE FUNCTIONAL GENOMICS OF THOUSANDS OF SPECIES



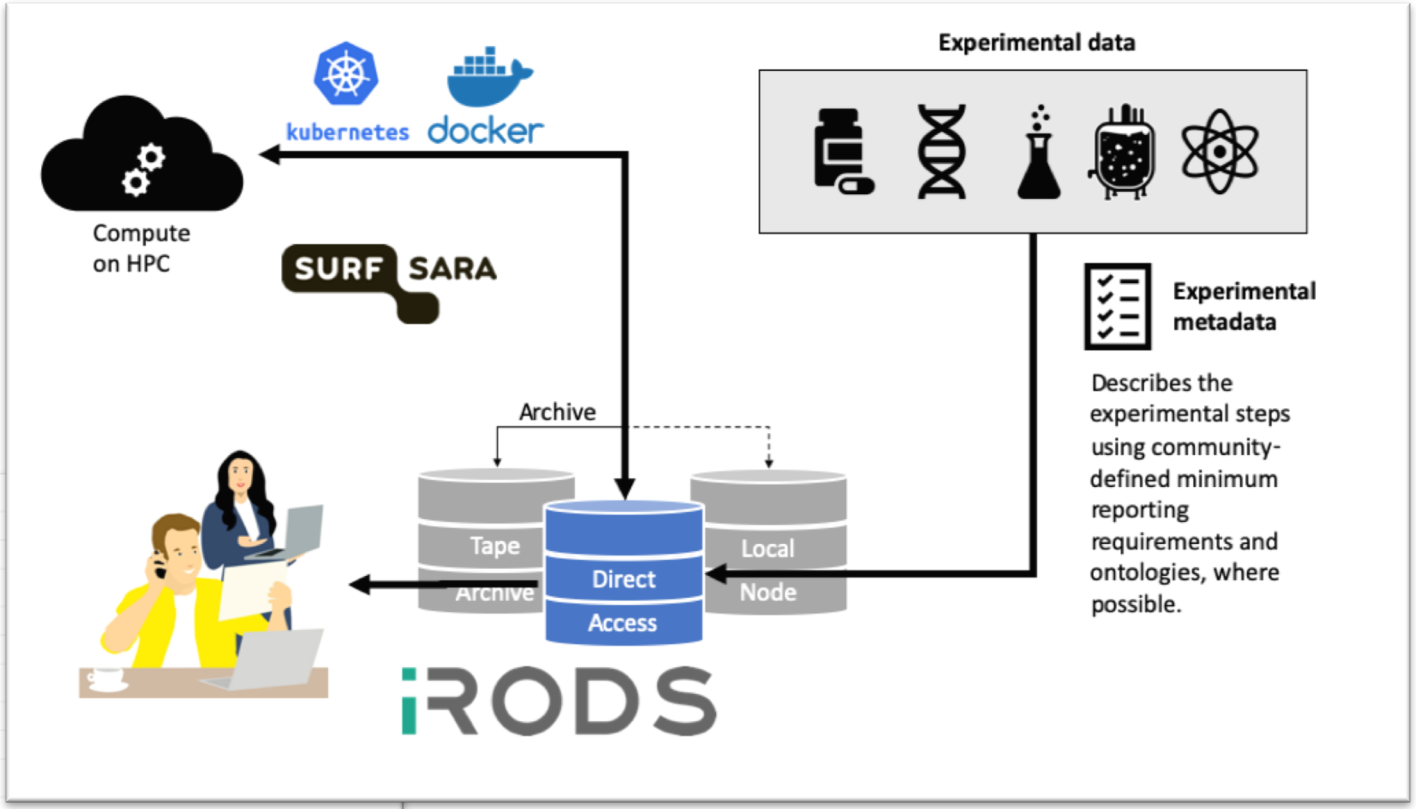

```
1  PREFIX gbol: <http://gbol.life/0.1/>
2  SELECT ?sample ?accession ?db
3  WHERE {
4      ?sample a gbol:Sample .
5      ?dnaobject gbol:sample ?sample .
6      ?dnaobject gbol:feature ?gene .
7      ?gene gbol:transcript ?transcript .
8      ?transcript gbol:feature ?cds .
9      ?cds gbol:protein ?protein .
10     ?protein gbol:feature ?domain .
11     ?domain gbol:xref ?xref .
12     ?xref gbol:db <http://gbol.life/0.1/db/pfam> .
13     ?xref gbol:accession ?accession .
14 } LIMIT 10
```

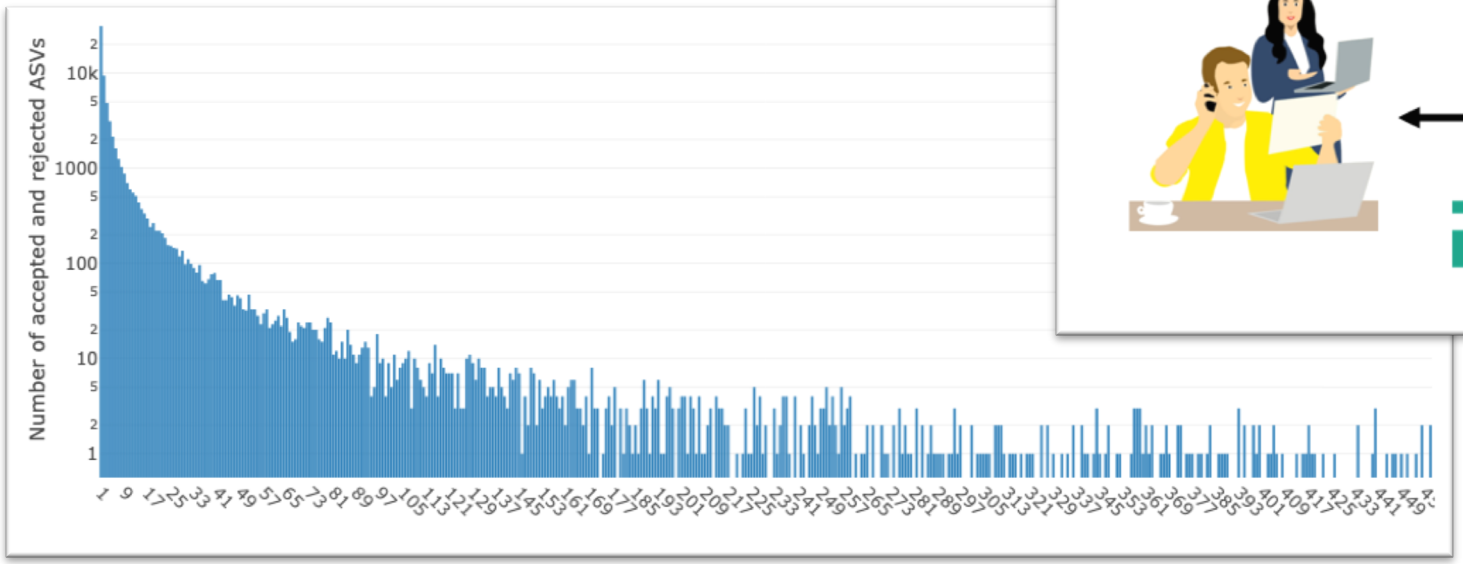




APPLICATIONS: NGTAX 2.0 HIGH THROUGHPUT AMPLICON ANALYSIS

- TRACK & TRACE EACH SPECIES
- SEQUENCE MAPPING ACROSS SAMPLES
- LINKED METADATA
  - AGE, GENDER, DIET, ...





NUMBER OF ASVs SHARED BETWEEN ~1800 SAMPLES

ANY OMICS CAN BE ANALYSED USING THE SAME PRINCIPLES

SOFTWARE - AVAILABILITY

- GENOME BIOLOGY ONTOLOGY LANGUAGE (GBOL)
  HTTP://GBOL.LIFE
- EMPUSA, THE CODE GENERATOR
  HTTP://EMPUSA.ORG
- SAPP, SEMANTIC GENOME ANNOTATION
  HTTP://SAPP.GITLAB.IO
- PROJECT MANAGEMENT
  HTTP://M-UNLOCK.NL
- NGTAX 2.0, AMPLICON ANALYSIS
  HTTP://WURSSB.GITLAB.IO/NGTAX/

WDCC

UNLOCK
MICROBIAL POTENTIAL