Data Lakehouse to support the development of AI models for predicting patient response to anti-tumor therapies

Elodine COQUELET[1], Marta SILVA[2], Laura BALBI[2], Jofre RIBA[3], Javier ALFARO[4], Fabio Massimo ZANZOTTO[5], Catia PESQUITA[2], Rohit KUMAR[3] and Christophe BATTAIL[1]

[1] Université Grenoble Alpes, IRIG, Laboratoire Biosciences et Bioingénierie pour la Santé, UA 13 INSERM-CEA-UGA, 38000 Grenoble, France.
[2] LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal.
[3] Fundacio Eurecat, Spain.
[4] International Center for Cancer Vaccine Science, University of Gdansk, Poland.
[5] University of Rome Tor Vergata, Italy.

Horizon 2020 KATY project (grant No 101017453)
Horizon Europe CANVAS project (grant No 101079510)

1. European project KATY

European project KATY

- The KATY project (4 years, 2021 - 2024) aims:
  - to develop an AI-empowered personalized medicine system to improve cancer treatments ;
  - to prototype it for the prediction of the response of patients with metastatic kidney cancer to targeted and immuno-therapies.

- Multi-disciplinary consortium spread over 20 institutions and 12 counstries.

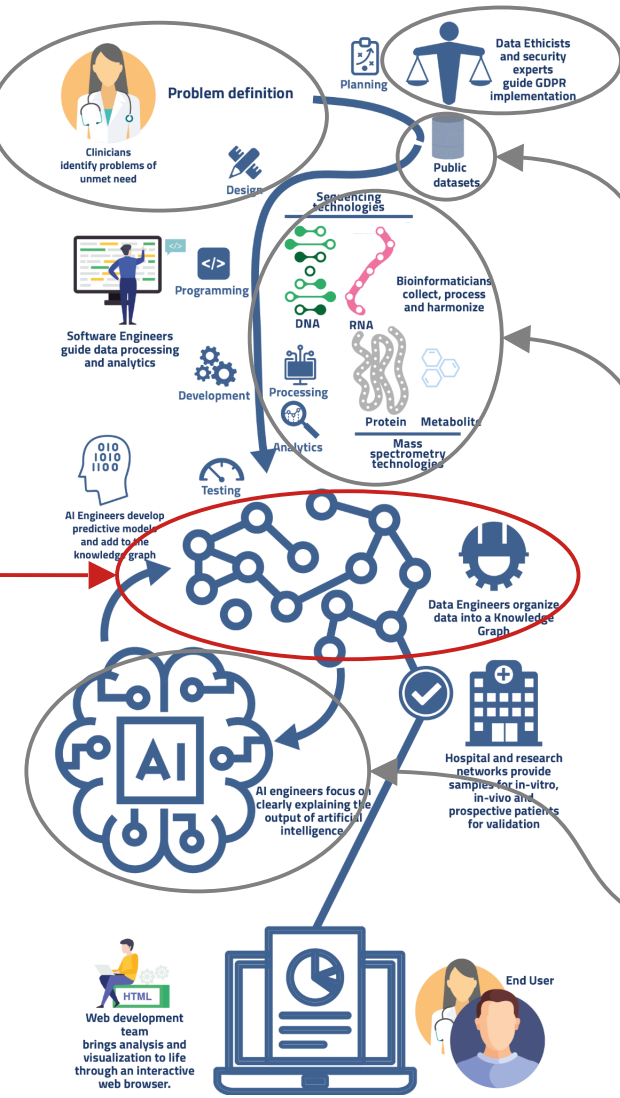
- Collection and generation of large scale “omics” molecular data for kidney cancer:
  - from public databases (processed and raw)
  - from cohorts of patient tumors collected by KATY clinical partners.

KATCHY
Clinical Knowledge for Humanity

European project KATY - Workflow

Definition of the clinical problem

Data management in a **Data Lake** & a Knowledge Graph hosted on a European computing infrastructure



Secure data sharing (GDPR compliance)

Identification and homogeneous processing of relevant public datasets

Production of new omics data in accordance with patient consent

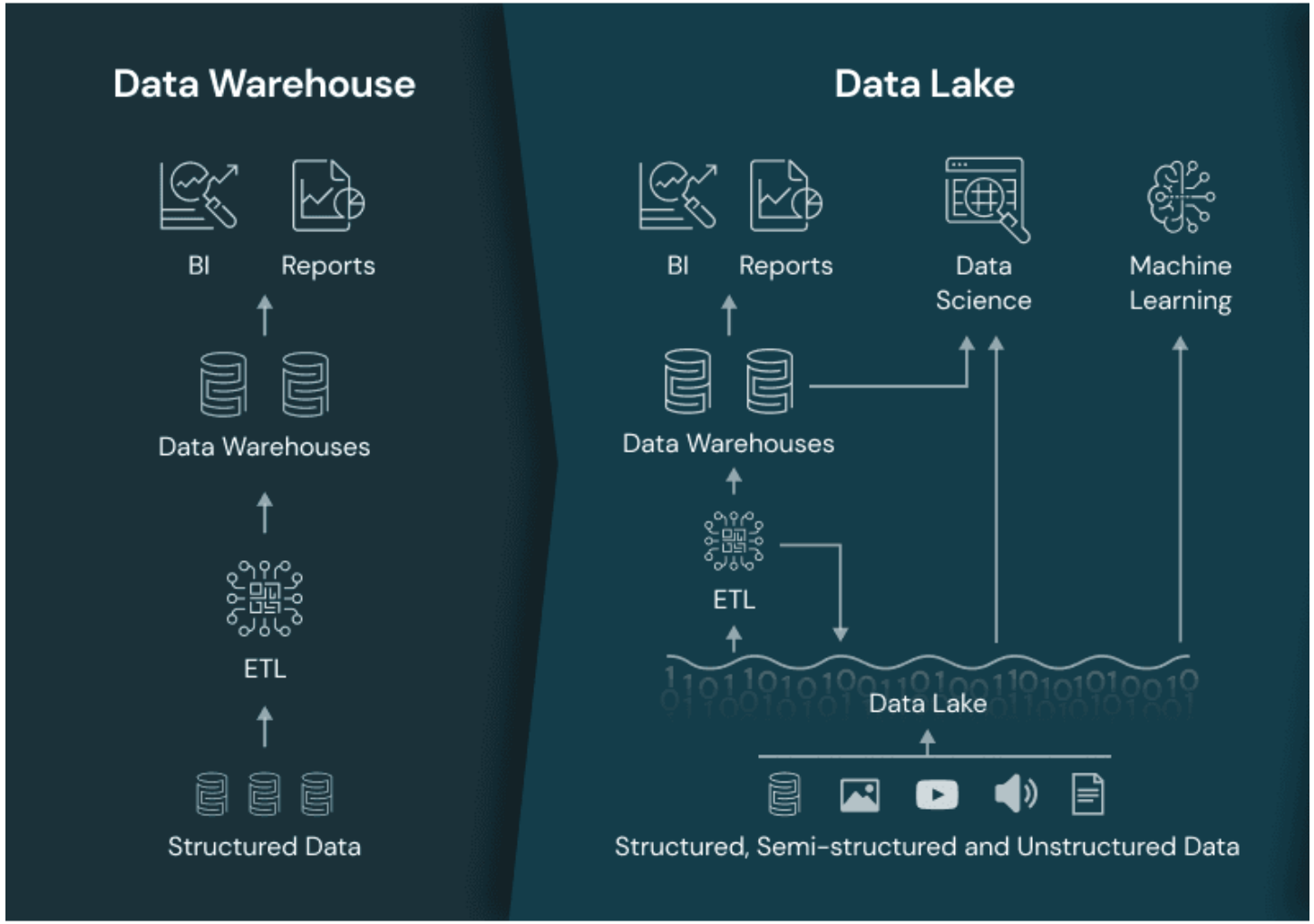
Computational modelling using explainable ML and AI algorithms

2. Big data management

Big data technologies

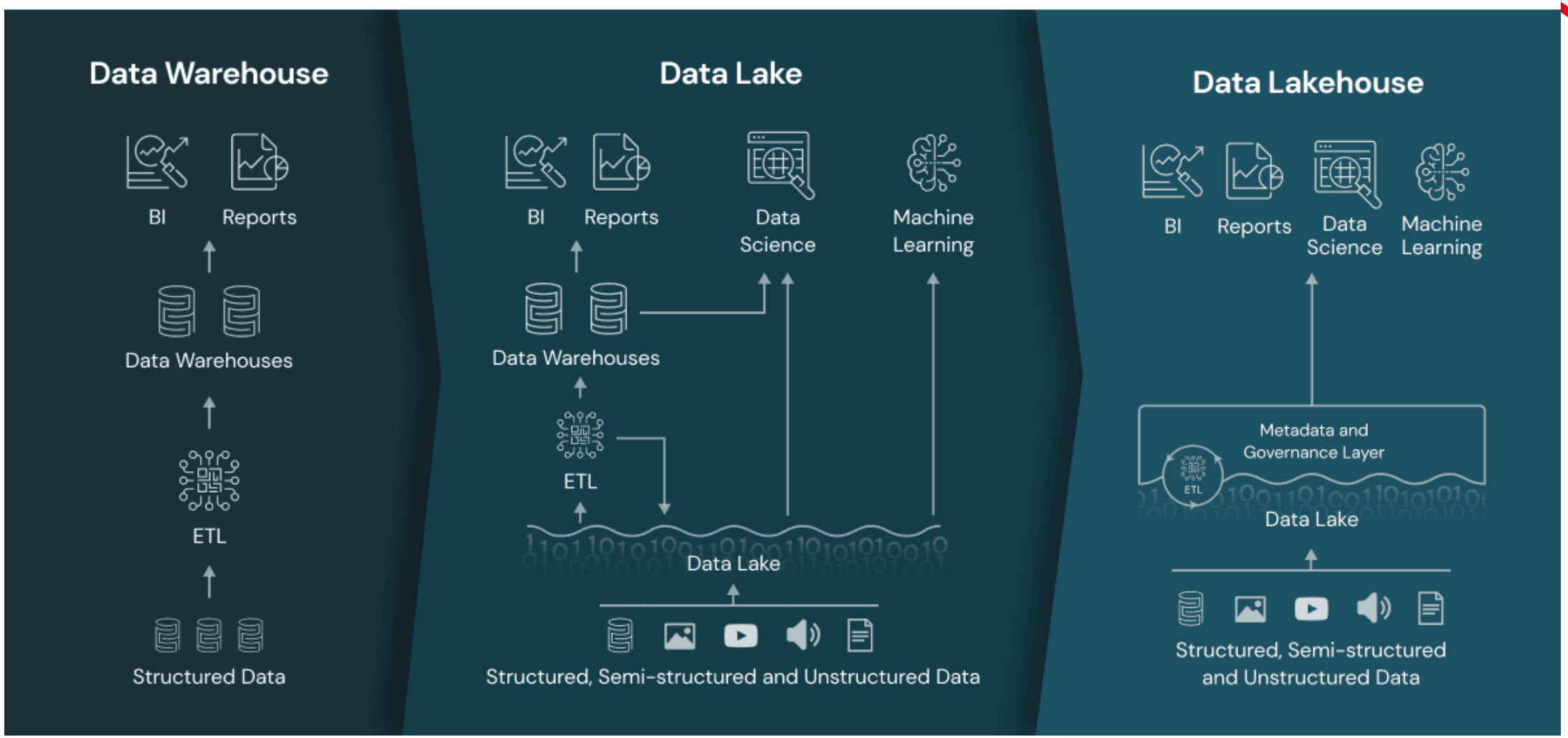
https://www.databricks.com/glossary/data-lakehouse

Big data technologies



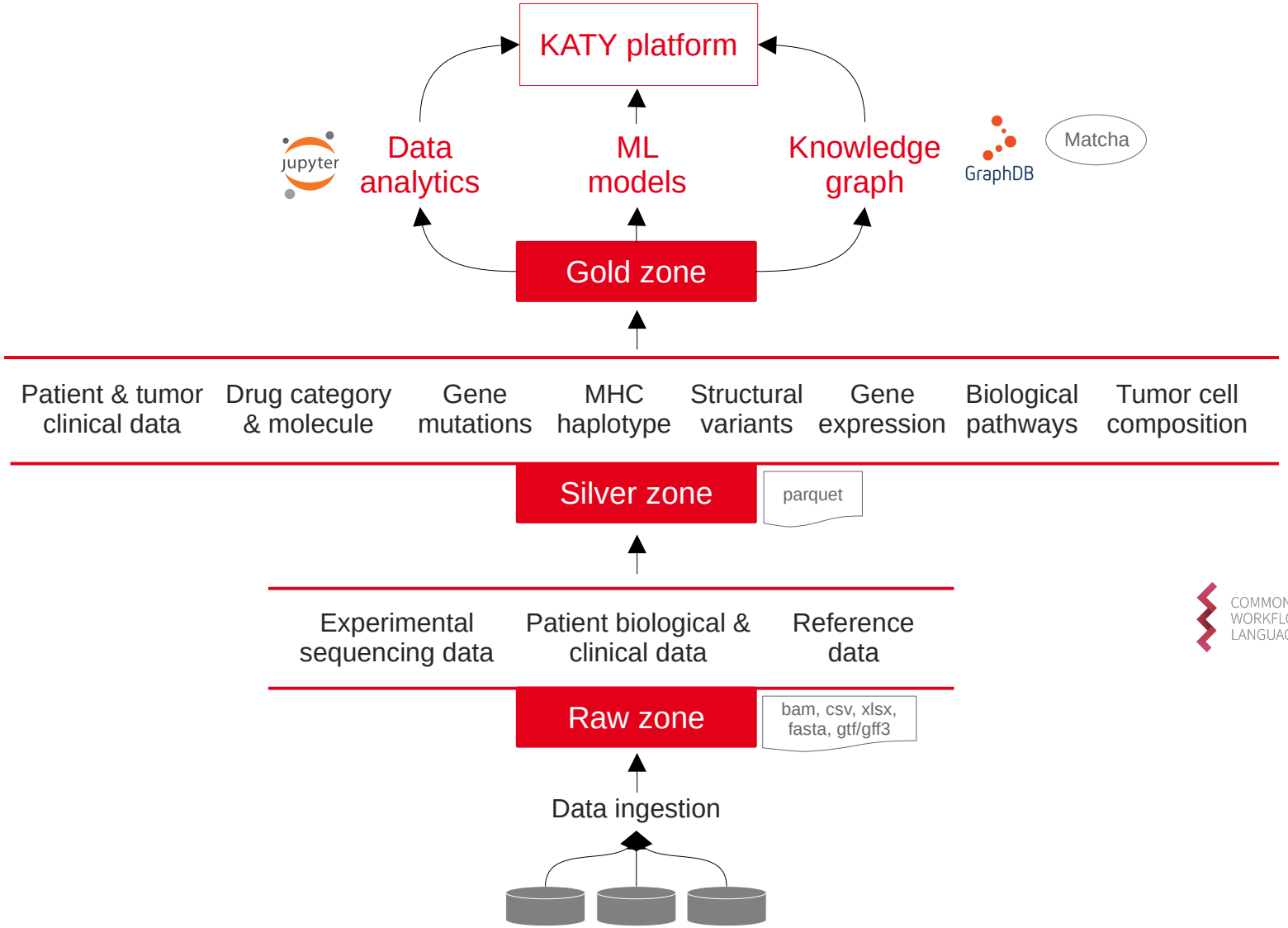
https://www.databricks.com/glossary/data-lakehouse

Big data technologies



https://www.databricks.com/glossary/data-lakehouse

Data Lakehouse



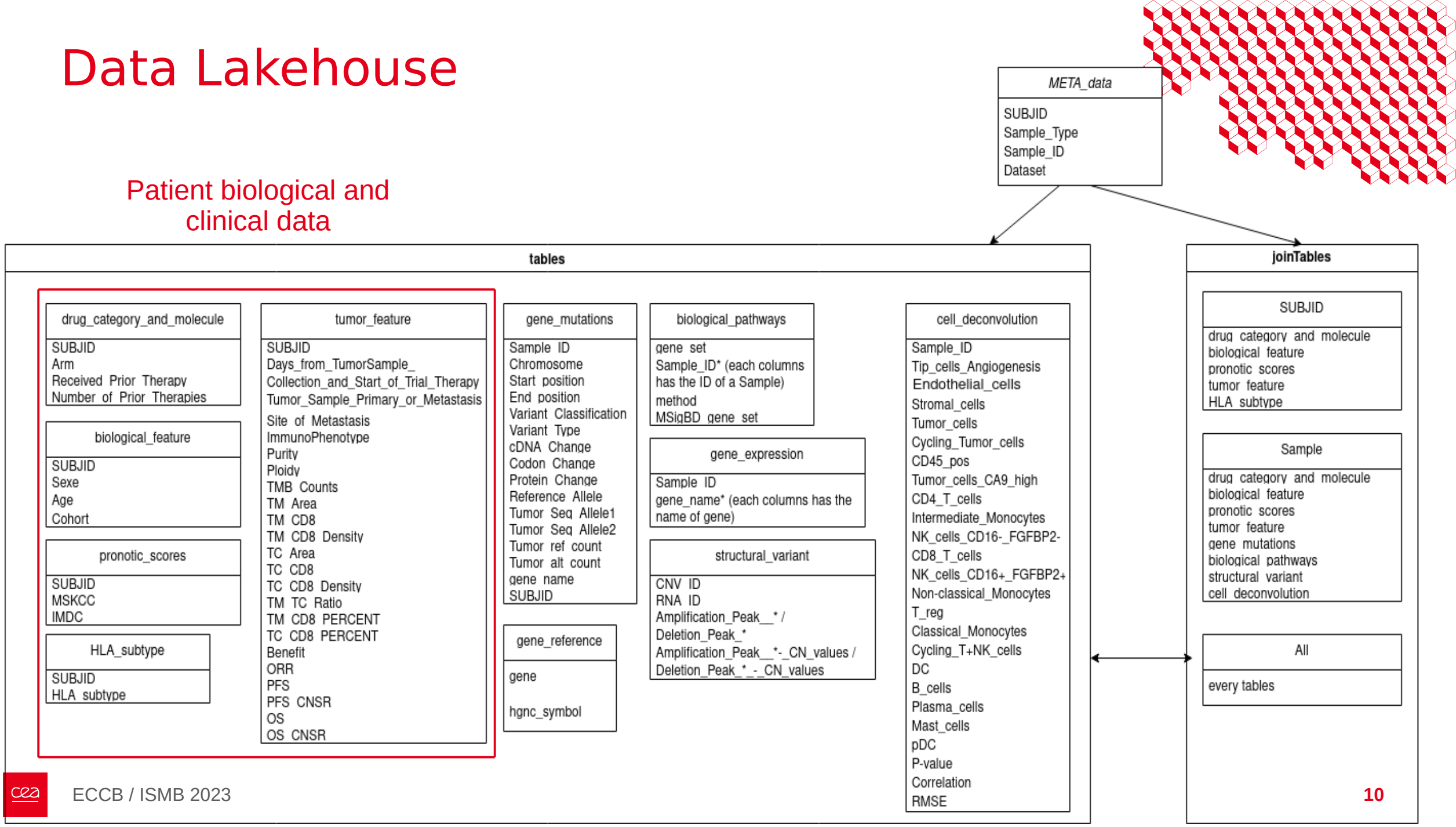
jupyter Apache Airflow

DELTA LAKE Apache Ranger Apache Atlas

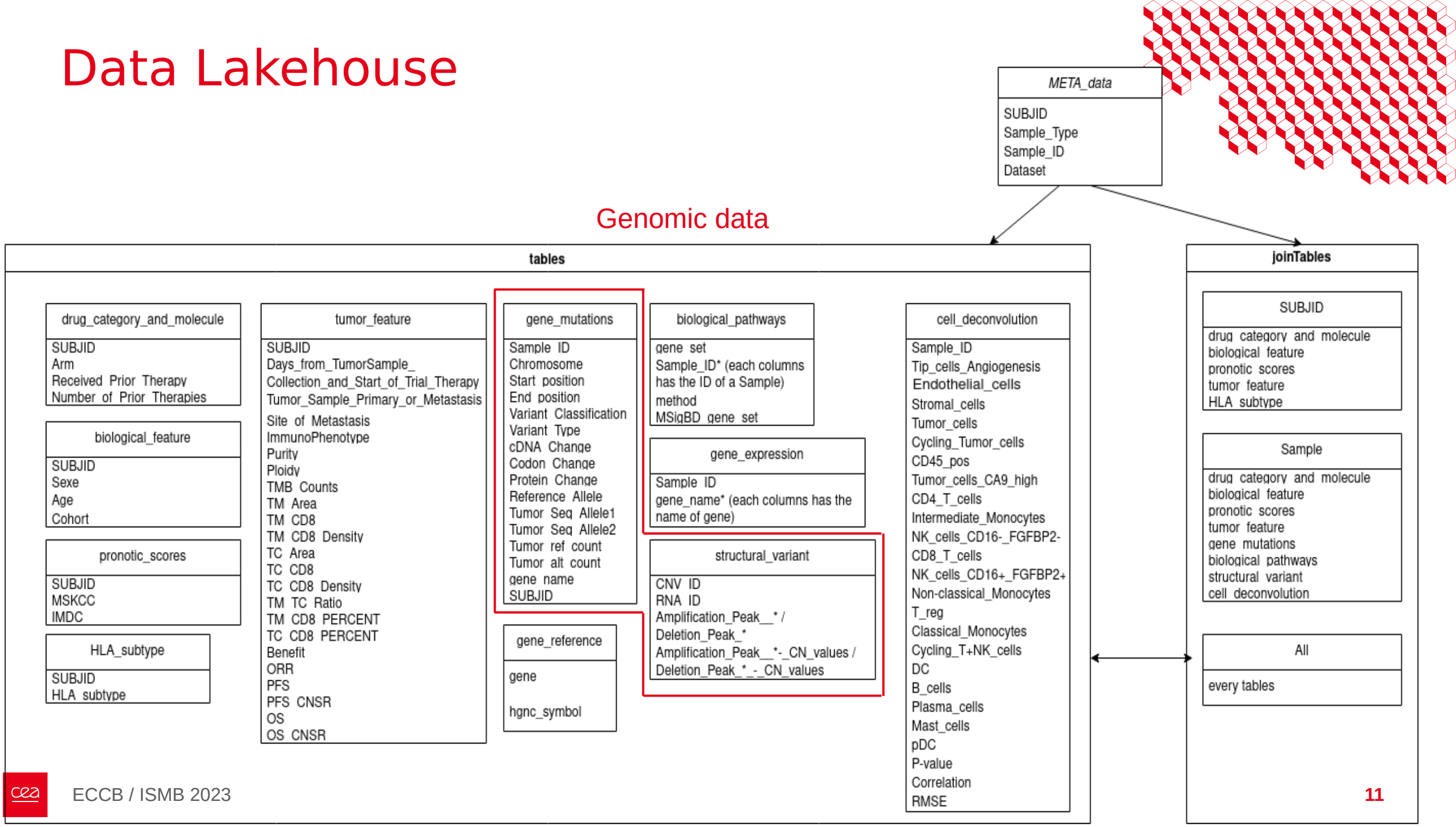
COMMON WORKFLOW LANGUAGE openstack. HDFS Apache Spark

Data Lakehouse

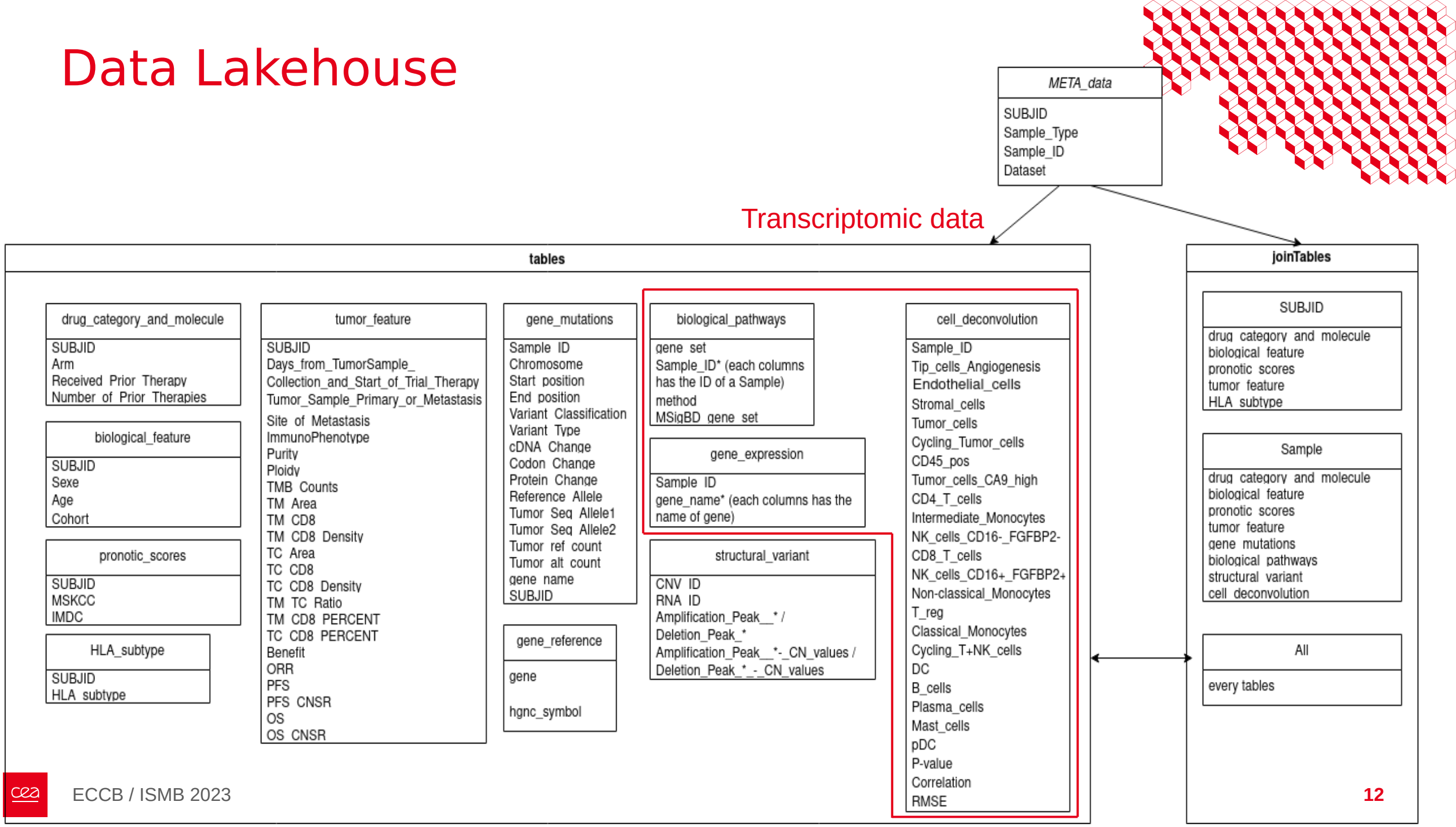
Patient biological and clinical data



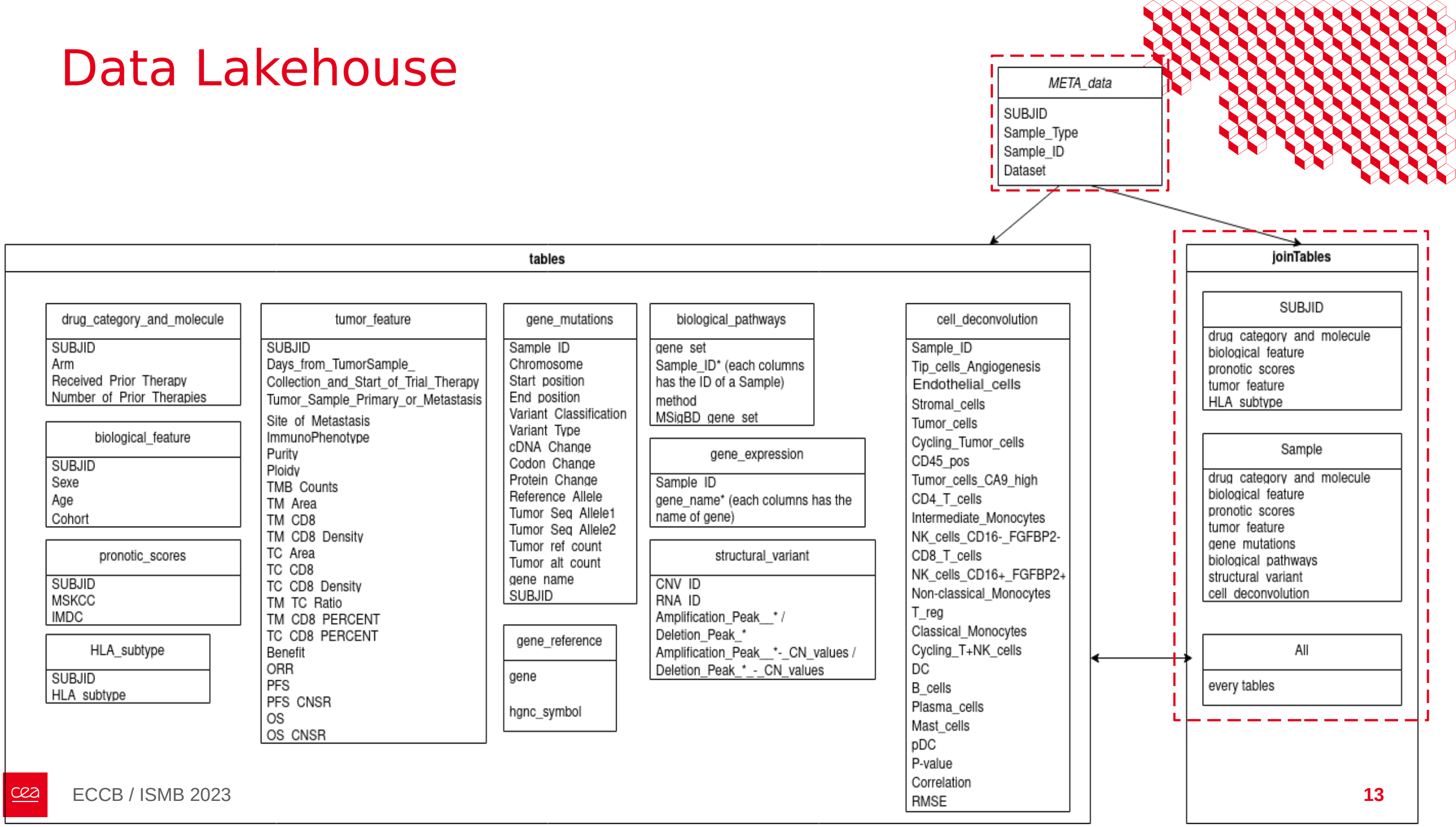
Data Lakehouse



Data Lakehouse



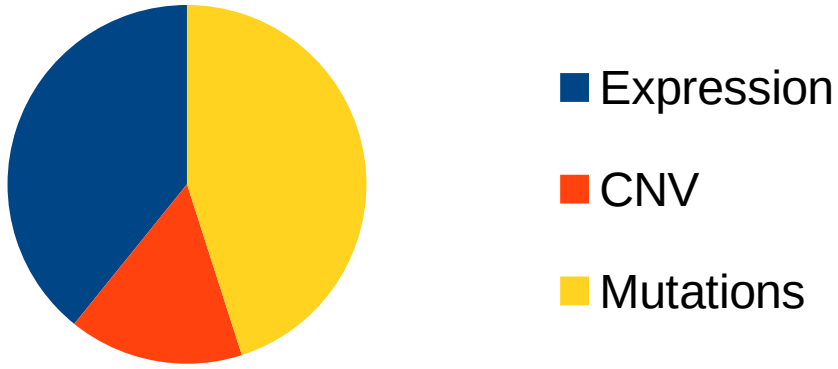
Data Lakehouse



Data Lakehouse

Repartition of experiments

2287 Samples



Repartition of Treatments

1478 Subjects

Prototyping from two data sets combining molecular profiling of kidney tumor tissues with clinical drug trials:

- Braun, D. A. *et al.* (2020). Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. Nature medicine, 26(6), 909–918.
- Motzer, R. J. *et al.* (2020). Avelumab plus axitinib versus sunitinib in advanced renal cell carcinoma: biomarker analysis of the phase 3 JAVELIN Renal 101 trial. Nature medicine, 26(11), 1733–1741.

3. Data queries

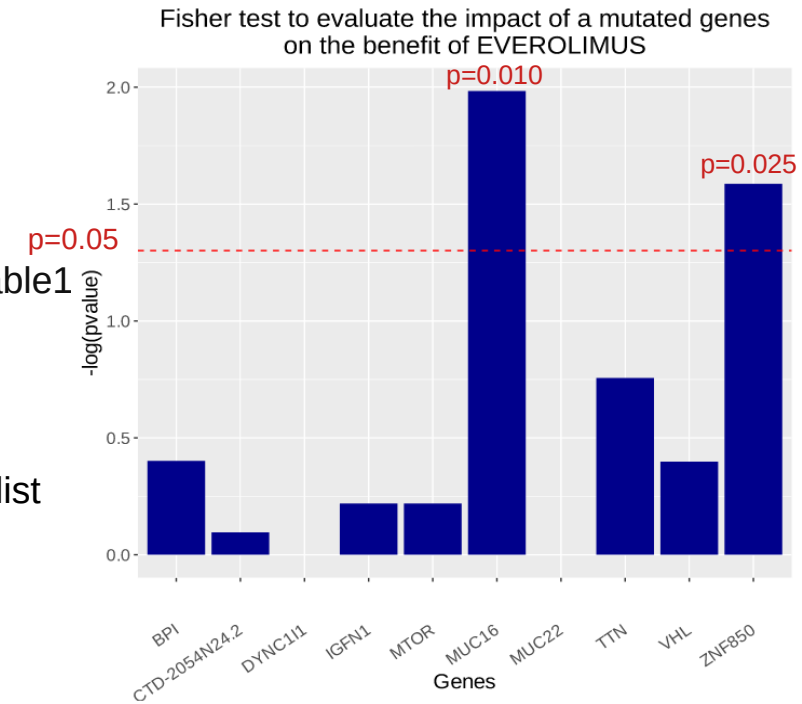
Data query – Use case 1

Investigate whether **genes frequently mutated in kidney cancer** are associated with **patient response to treatment**

1) **SELECT** gene_name,count(**gene_name**) **FROM DELTA**.`**gene_mutations**`
WHERE (Variant_Type == '**SNP**' **AND** Variant_Classification == '**Missense_Mutation**')
GROUP BY gene_name **ORDER BY** count(gene_name) **DESC LIMIT** **10**

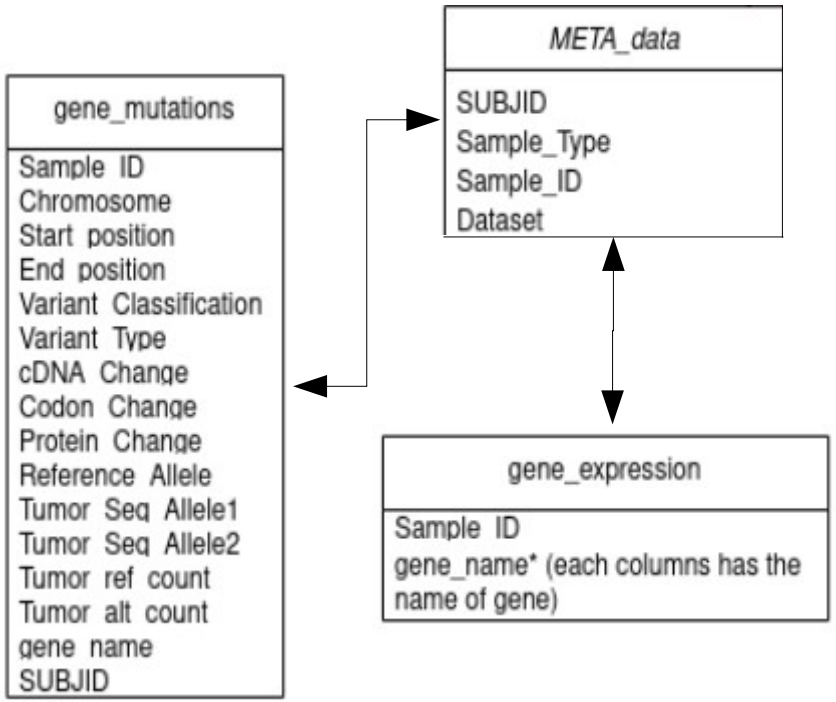
2) **SELECT** table1.SUBJID, table1.Sample_ID, table2.benefit **FROM DELTA**.`META_data` table1
JOIN DELTA.`joinTables/SUBJID` table2 **ON** table1.SUBJID == table2.SUBJID
WHERE table1.Sample_Type == '**MAF_Tumor_ID**' **AND** table1.Sample_ID != 'None'
AND table2.Arm == '**EVEROLIMUS**' **AND** table1.Dataset == '**BRAUN_2020**'

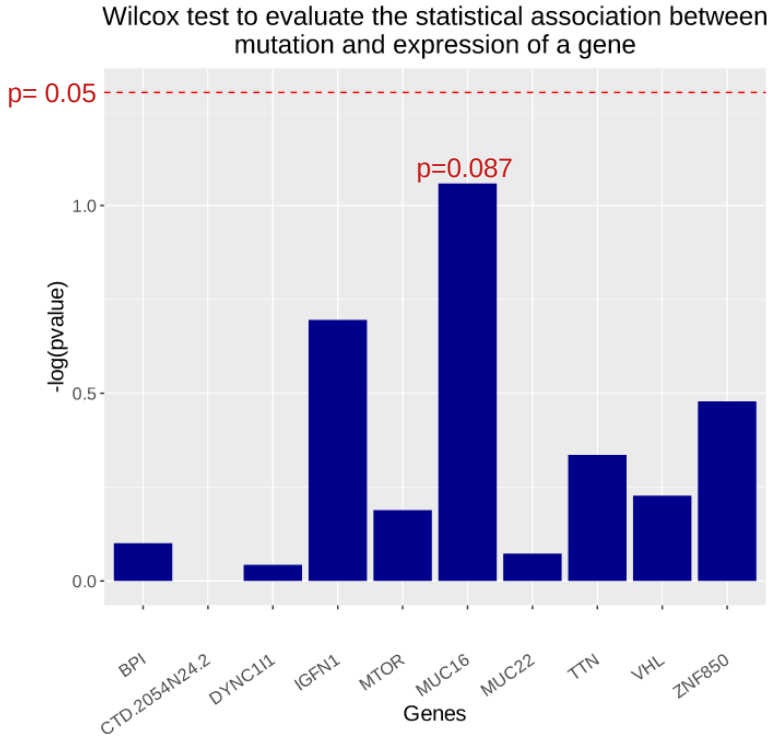
3) Sort for each mutation whether or not the Sample of the sub-population is in the mutation list



Data query – Use case 2

Investigate whether, for frequently mutated genes in kidney cancer, **the mutation status** is associated with a **change in gene expression**





Conclusions

- Prototyping of a Data Lakehouse integrating patient molecular and clinical data, to support the development of AI models for predicting response to targeted and immuno-therapies for patients with kidney cancer.

- Validation of the data structure with first data queries.

- Work in progress on semantic ontologies to harmonize and control the vocabulary used.

- On going deployment on a cloud infrastructure.

- Establishment of a data security and governance strategy for the integration of restricted access datasets.

Poster 1309

Thank you

Contact:
elodine.coquelet@gmail.com
christophe.battail@cea.fr

CEA GRENOBLE

17 avenue des Martyrs

38000 Grenoble, France

Horizon 2020 KATY project (grant No 101017453)
Horizon Europe CANVAS project (grant No 101079510)