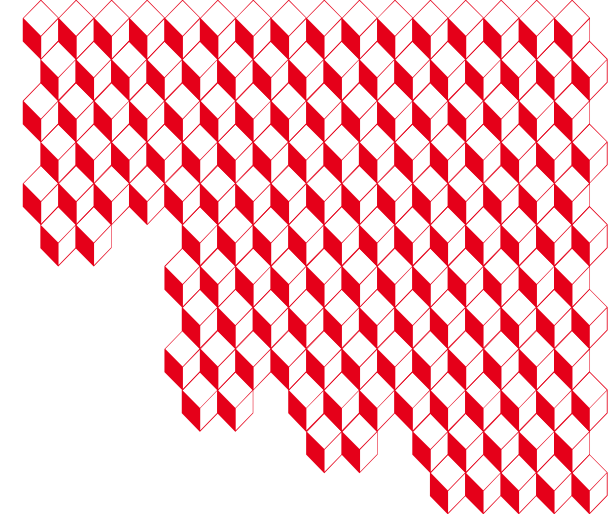




irig



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

Elodine COQUELET¹, Marta SILVA², Laura BALBI², Jofre RIBA³, Javier ALFARO⁴, Fabio Massimo ZANZOTTO⁵, Catia PESQUITA², Rohit KUMAR³ and Christophe BATTAIL¹

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.



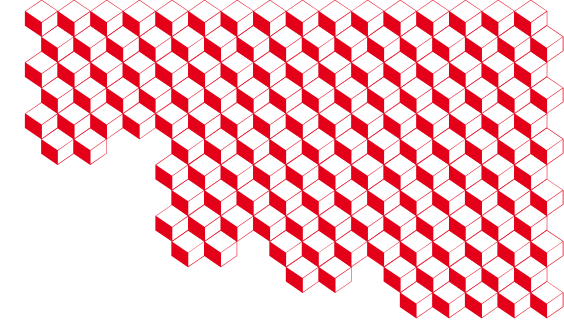
Horizon 2020 KATY project (grant No 101017453)
Horizon Europe CANVAS project (grant No 101079510)





1. European project KATY

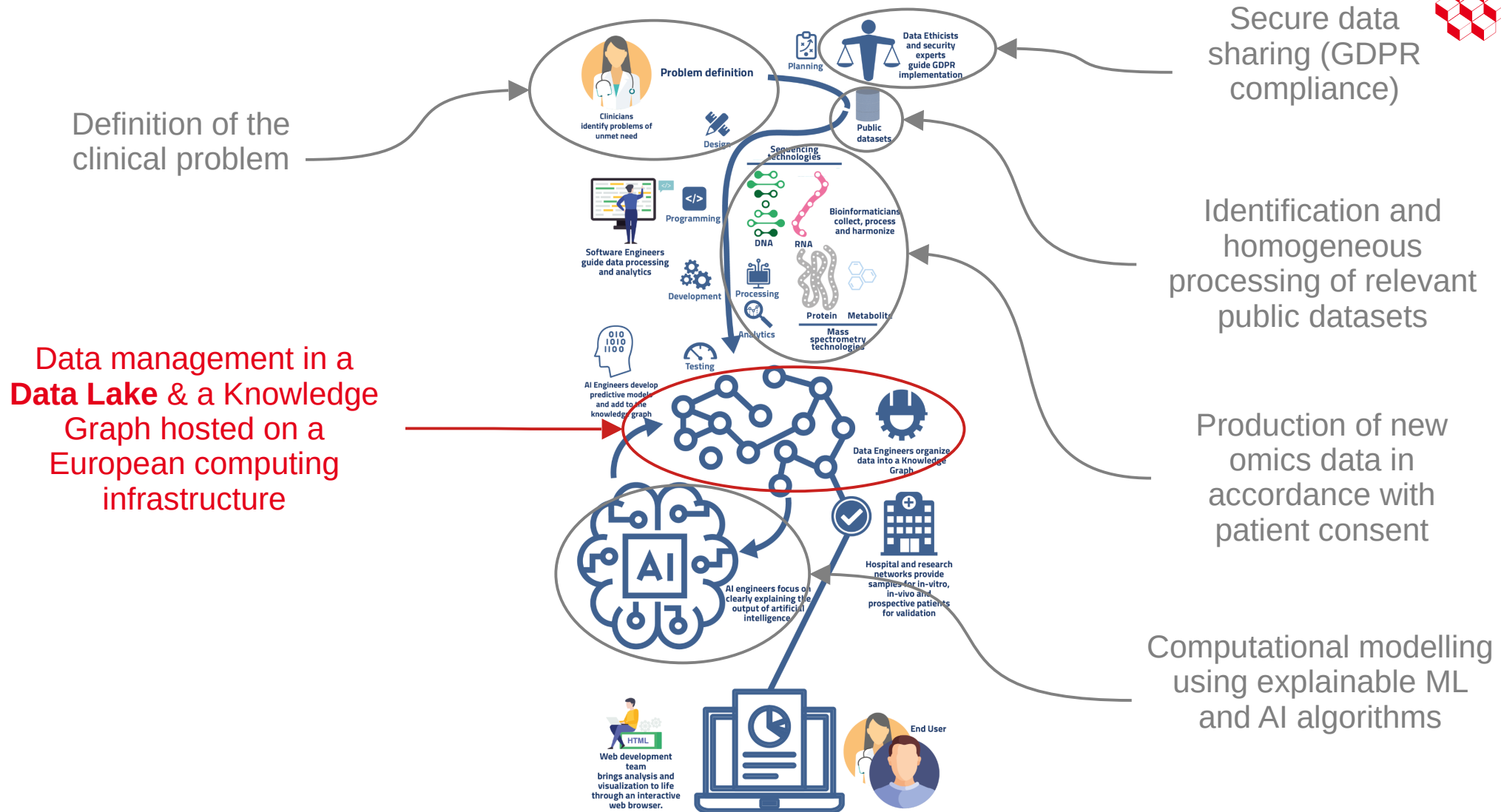
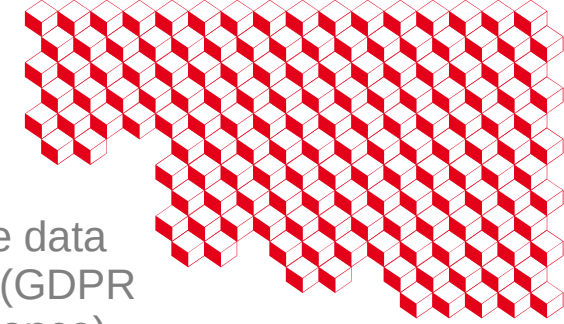
European project KATY



- 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 countries.
- **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.



European project KATY - Workflow



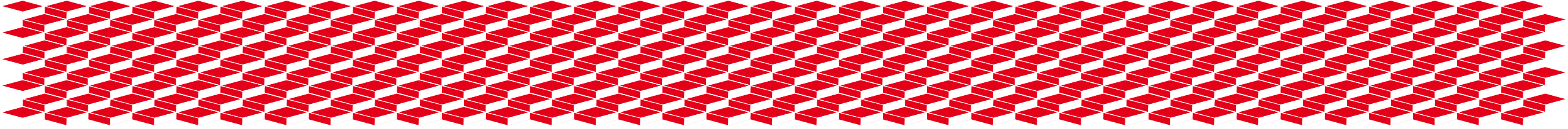
Definition of the clinical problem

Secure data sharing (GDPR compliance)

Identification and homogeneous processing of relevant public datasets

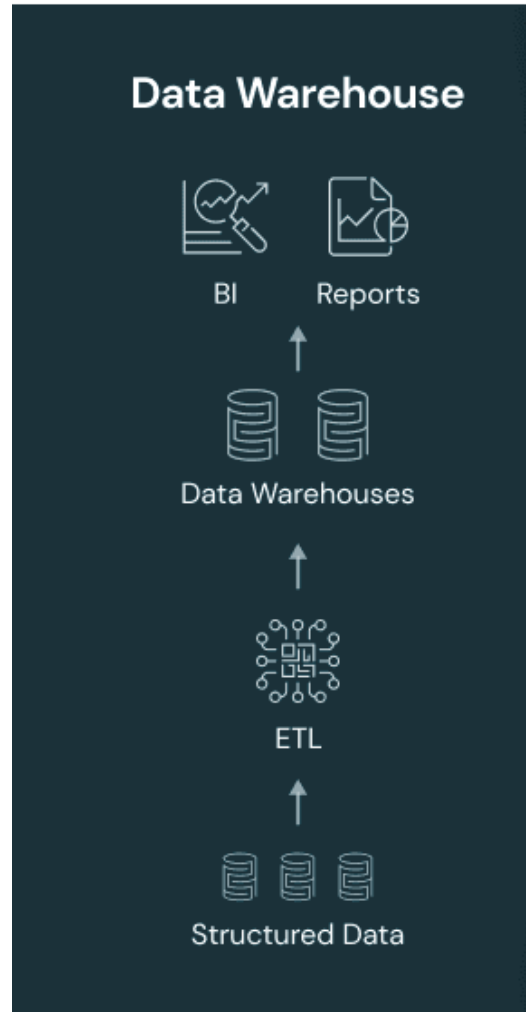
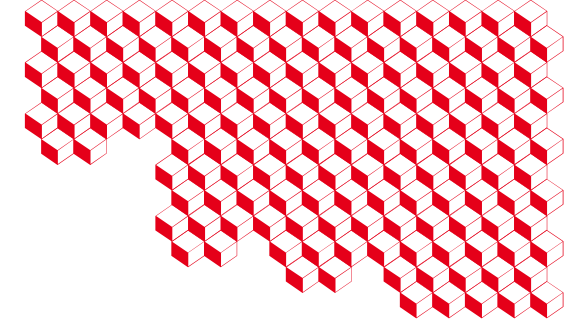
Production of new omics data in accordance with patient consent

Computational modelling using explainable ML and AI algorithms



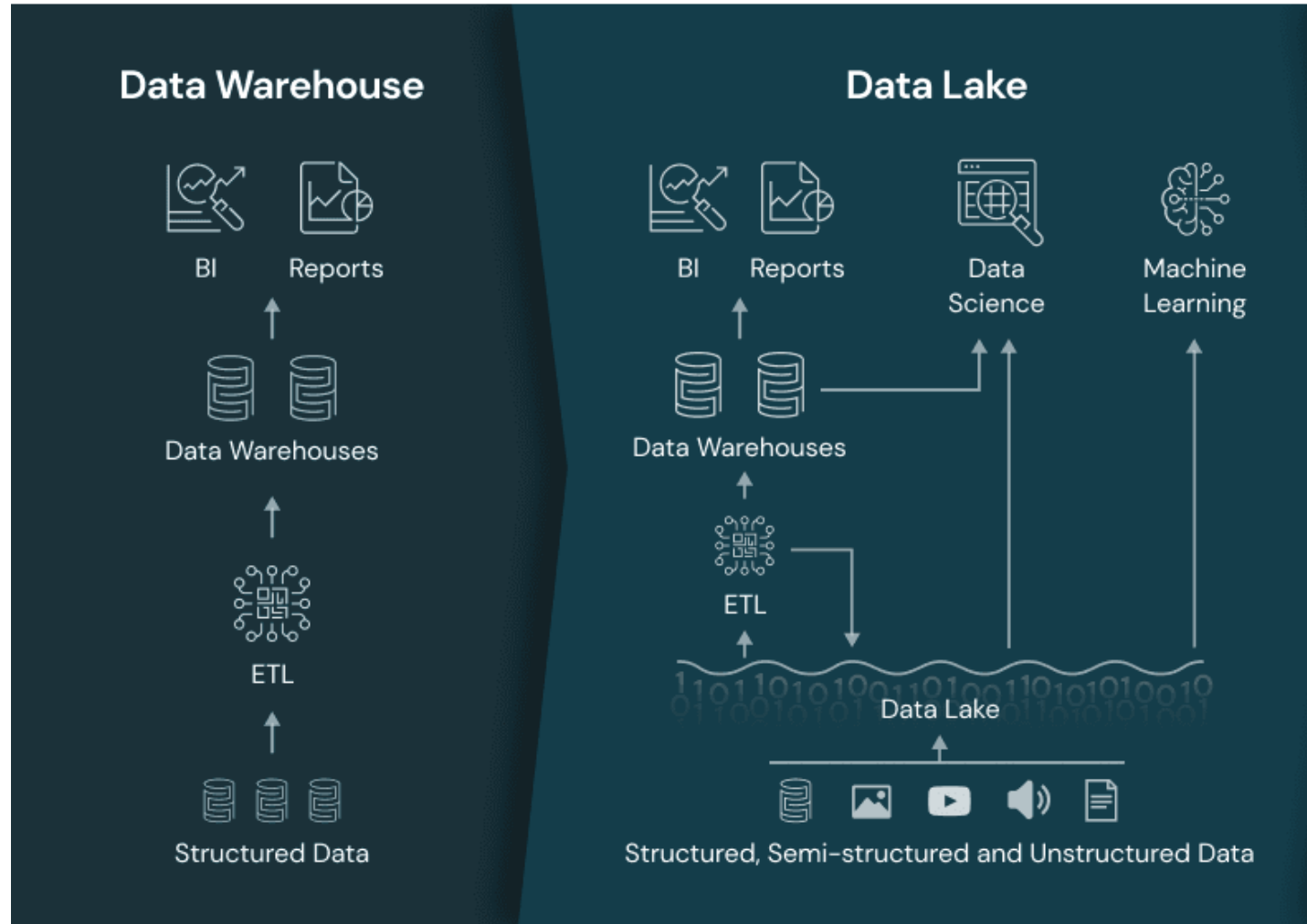
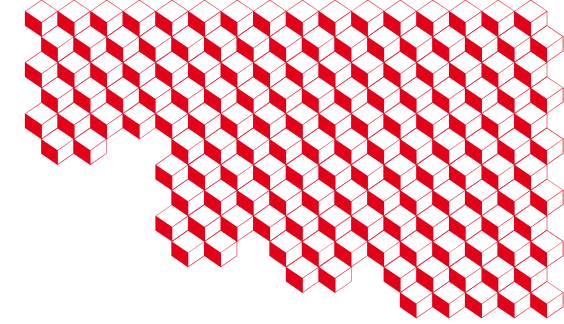
2 ■ Big data management

Big data technologies



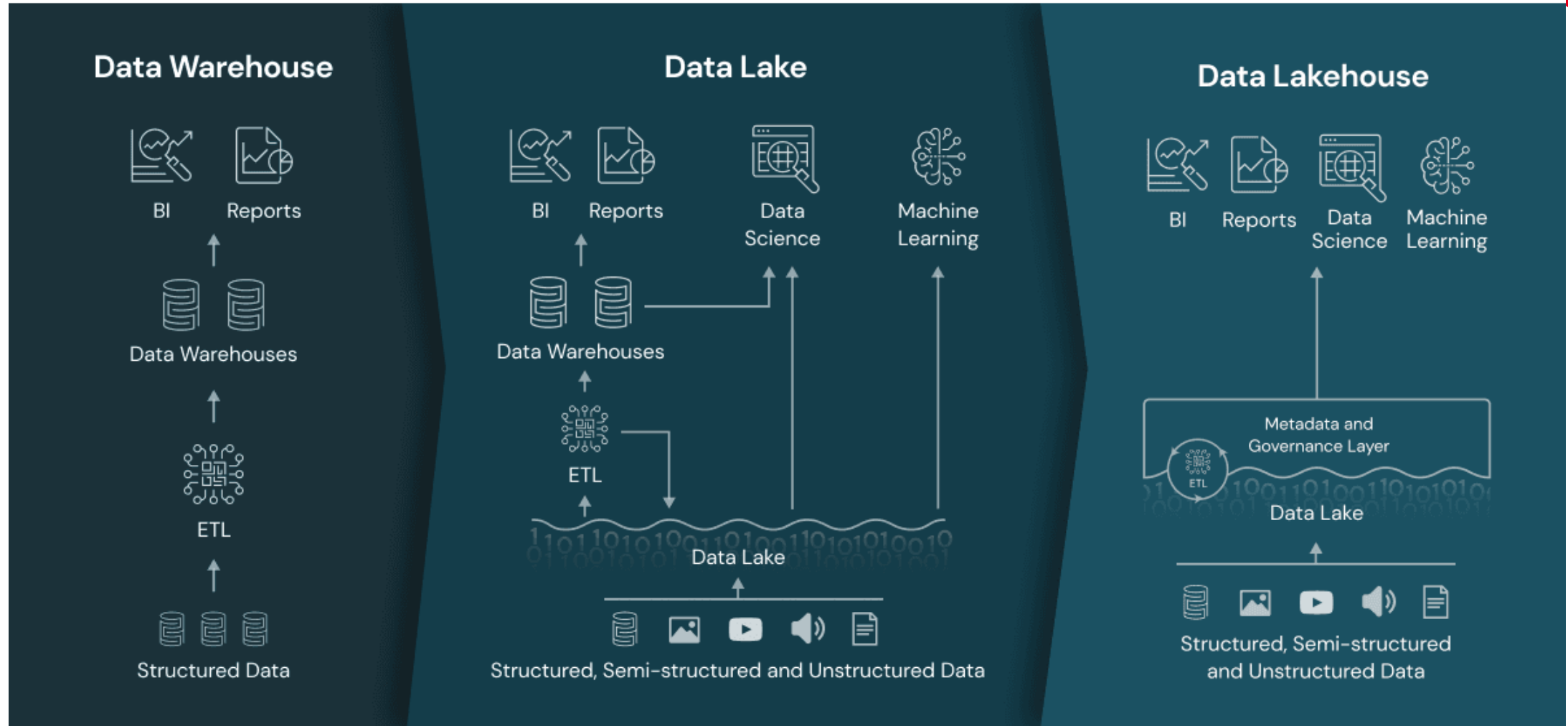
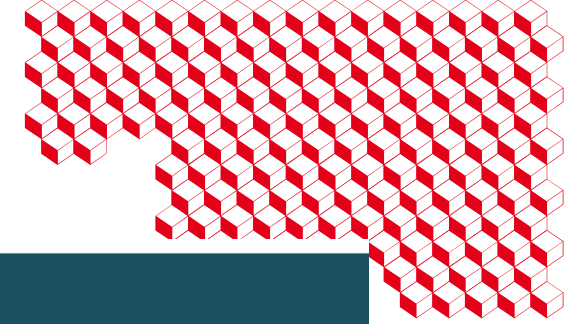
<https://www.databricks.com/glossary/data-lakehouse>

Big data technologies



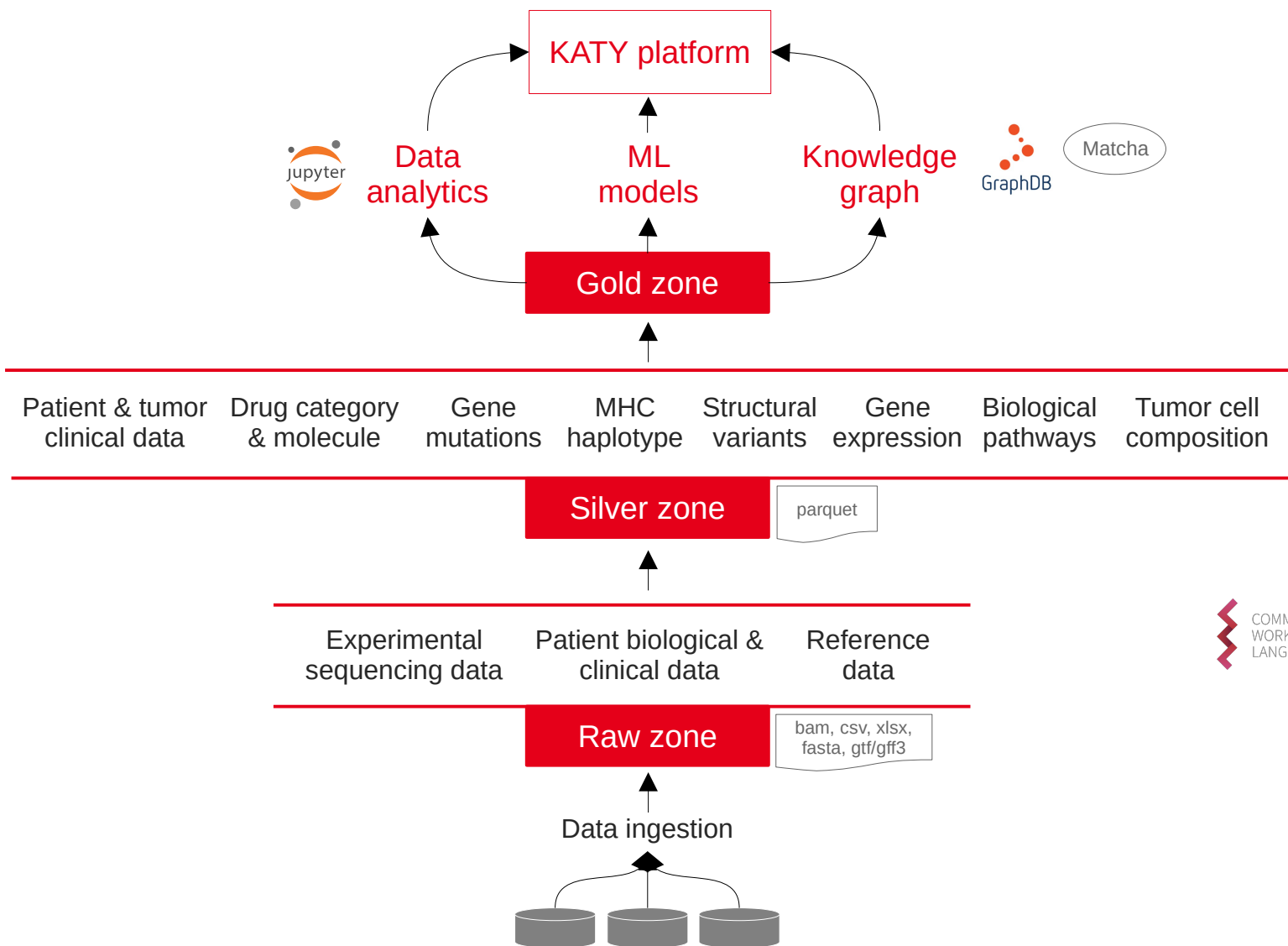
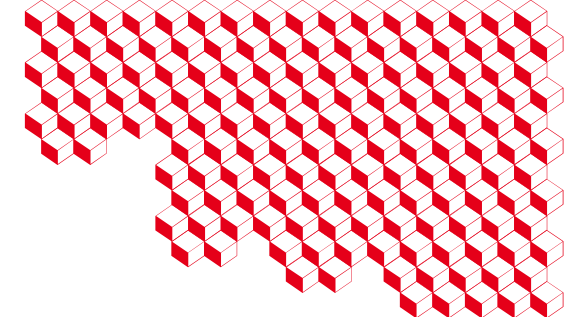
<https://www.databricks.com/glossary/data-lakehouse>

Big data technologies

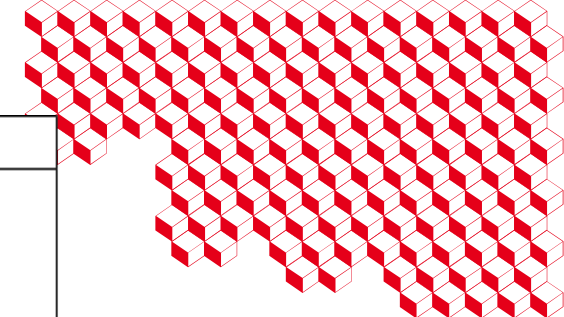


<https://www.databricks.com/glossary/data-lakehouse>

Data Lakehouse

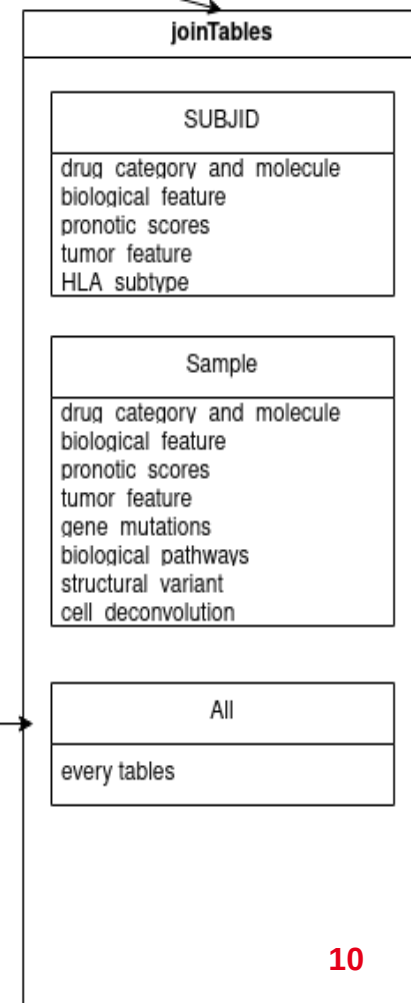
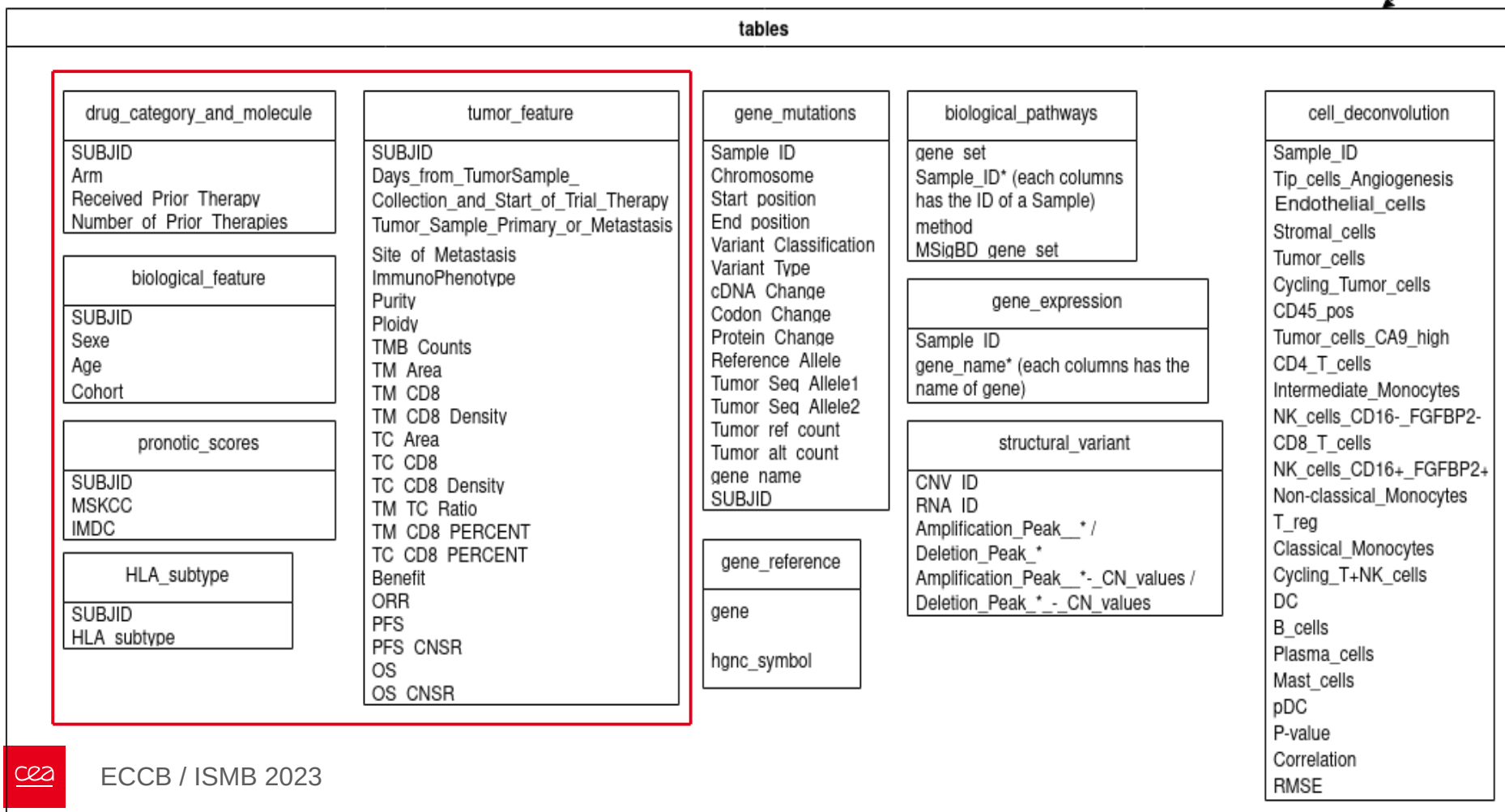


Data Lakehouse

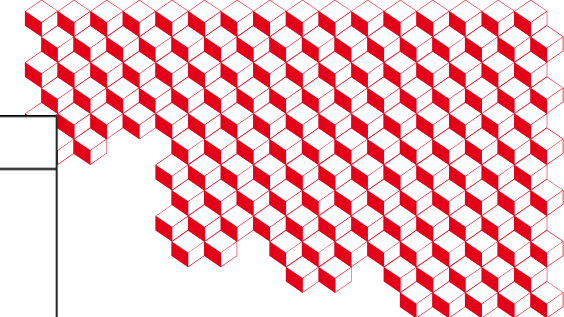


Patient biological and clinical data

META_data
SUBJID
Sample_Type
Sample_ID
Dataset

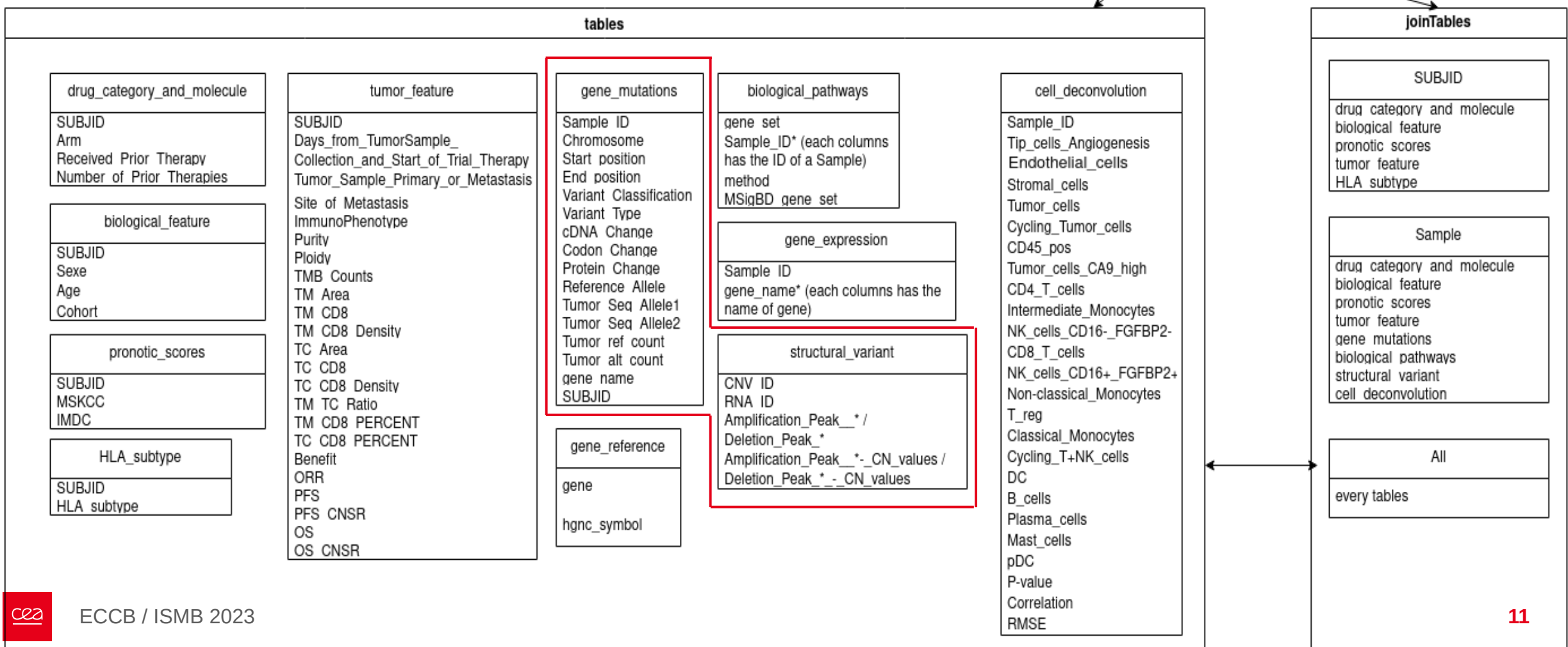


Data Lakehouse

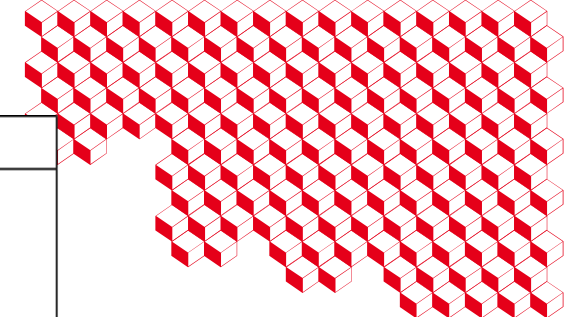


META_data
SUBJID
Sample_Type
Sample_ID
Dataset

Genomic data

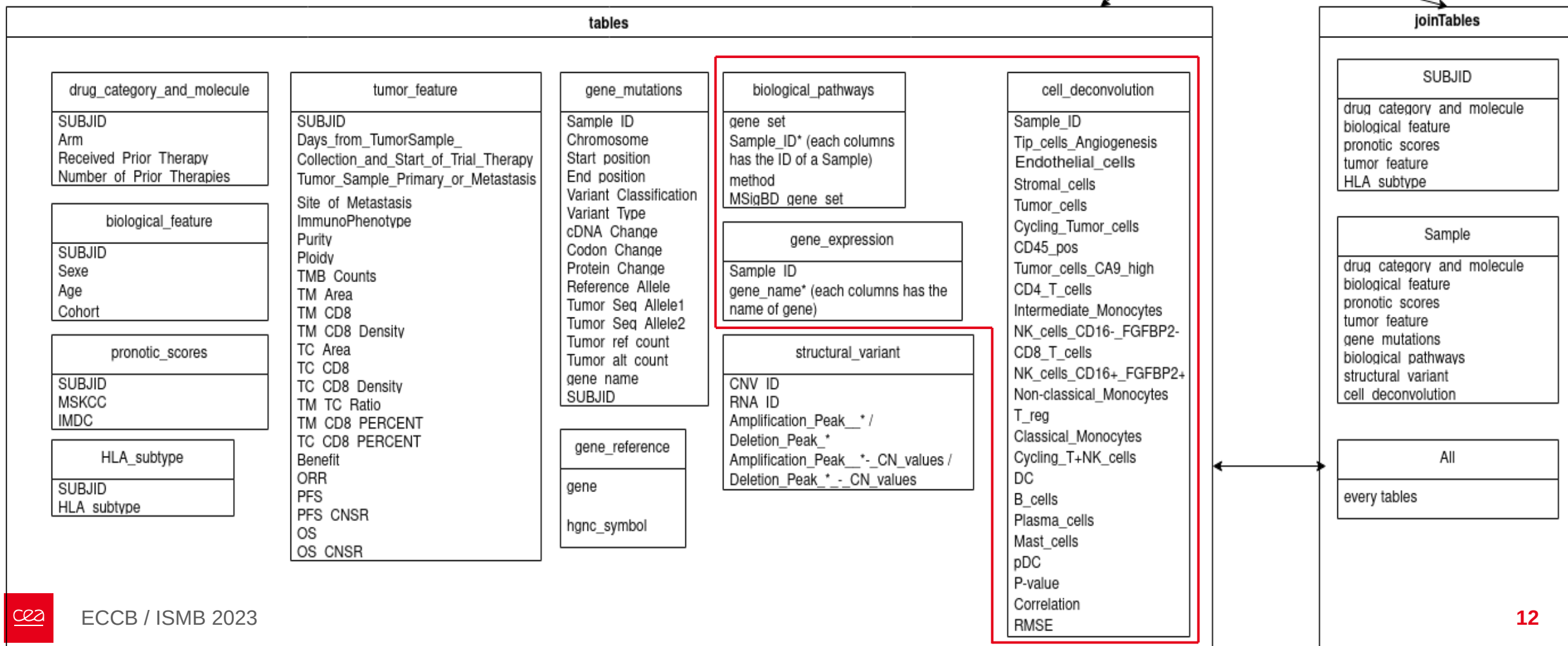


Data Lakehouse

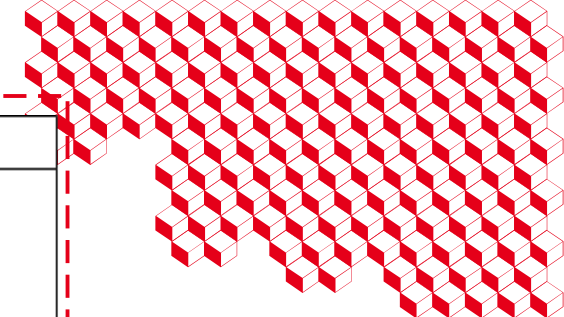


META_data
SUBJID
Sample_Type
Sample_ID
Dataset

Transcriptomic data



Data Lakehouse

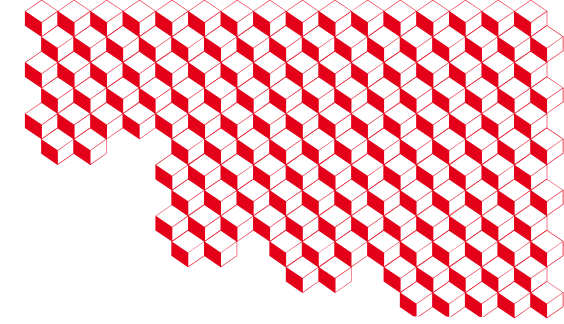


META_data
SUBJID
Sample_Type
Sample_ID
Dataset

tables																																																																																	
<table border="1"> <thead> <tr> <th>drug_category_and_molecule</th> </tr> </thead> <tbody> <tr> <td>SUBJID</td> </tr> <tr> <td>Arm</td> </tr> <tr> <td>Received Prior Therapy</td> </tr> <tr> <td>Number of Prior Therapies</td> </tr> </tbody> </table>	drug_category_and_molecule	SUBJID	Arm	Received Prior Therapy	Number of Prior Therapies	<table border="1"> <thead> <tr> <th>tumor_feature</th> </tr> </thead> <tbody> <tr> <td>SUBJID</td> </tr> <tr> <td>Days_from_TumorSample_Collection_and_Start_of_Trial_Therapy</td> </tr> <tr> <td>Tumor_Sample_Primary_or_Metastasis</td> </tr> <tr> <td>Site of Metastasis</td> </tr> <tr> <td>ImmunoPhenotype</td> </tr> <tr> <td>Purity</td> </tr> <tr> <td>Ploidy</td> </tr> <tr> <td>TMB Counts</td> </tr> <tr> <td>TM Area</td> </tr> <tr> <td>TM CD8</td> </tr> <tr> <td>TM CD8 Density</td> </tr> <tr> <td>TC Area</td> </tr> <tr> <td>TC CD8</td> </tr> <tr> <td>TC CD8 Density</td> </tr> <tr> <td>TM TC Ratio</td> </tr> <tr> <td>TM CD8 PERCENT</td> </tr> <tr> <td>TC CD8 PERCENT</td> </tr> <tr> <td>Benefit</td> </tr> <tr> <td>ORR</td> </tr> <tr> <td>PFS</td> </tr> <tr> <td>PFS CNSR</td> </tr> <tr> <td>OS</td> </tr> <tr> <td>OS CNSR</td> </tr> </tbody> </table>	tumor_feature	SUBJID	Days_from_TumorSample_Collection_and_Start_of_Trial_Therapy	Tumor_Sample_Primary_or_Metastasis	Site of Metastasis	ImmunoPhenotype	Purity	Ploidy	TMB Counts	TM Area	TM CD8	TM CD8 Density	TC Area	TC CD8	TC CD8 Density	TM TC Ratio	TM CD8 PERCENT	TC CD8 PERCENT	Benefit	ORR	PFS	PFS CNSR	OS	OS CNSR	<table border="1"> <thead> <tr> <th>gene_mutations</th> </tr> </thead> <tbody> <tr> <td>Sample ID</td> </tr> <tr> <td>Chromosome</td> </tr> <tr> <td>Start position</td> </tr> <tr> <td>End position</td> </tr> <tr> <td>Variant Classification</td> </tr> <tr> <td>Variant Type</td> </tr> <tr> <td>cDNA Change</td> </tr> <tr> <td>Codon Change</td> </tr> <tr> <td>Protein Change</td> </tr> <tr> <td>Reference Allele</td> </tr> <tr> <td>Tumor Seq Allele1</td> </tr> <tr> <td>Tumor Seq Allele2</td> </tr> <tr> <td>Tumor ref count</td> </tr> <tr> <td>Tumor alt count</td> </tr> <tr> <td>gene name</td> </tr> <tr> <td>SUBJID</td> </tr> </tbody> </table>	gene_mutations	Sample ID	Chromosome	Start position	End position	Variant Classification	Variant Type	cDNA Change	Codon Change	Protein Change	Reference Allele	Tumor Seq Allele1	Tumor Seq Allele2	Tumor ref count	Tumor alt count	gene name	SUBJID	<table border="1"> <thead> <tr> <th>biological_pathways</th> </tr> </thead> <tbody> <tr> <td>gene set</td> </tr> <tr> <td>Sample_ID* (each columns has the ID of a Sample)</td> </tr> <tr> <td>method</td> </tr> <tr> <td>MSigBD gene set</td> </tr> </tbody> </table>	biological_pathways	gene set	Sample_ID* (each columns has the ID of a Sample)	method	MSigBD gene set	<table border="1"> <thead> <tr> <th>cell_deconvolution</th> </tr> </thead> <tbody> <tr> <td>Sample_ID</td> </tr> <tr> <td>Tip_cells_Angiogenesis</td> </tr> <tr> <td>Endothelial_cells</td> </tr> <tr> <td>Stromal_cells</td> </tr> <tr> <td>Tumor_cells</td> </tr> <tr> <td>Cycling_Tumor_cells</td> </tr> <tr> <td>CD45_pos</td> </tr> <tr> <td>Tumor_cells_CA9_high</td> </tr> <tr> <td>CD4_T_cells</td> </tr> <tr> <td>Intermediate_Monocytes</td> </tr> <tr> <td>NK_cells_CD16-_FGFBP2-</td> </tr> <tr> <td>CD8_T_cells</td> </tr> <tr> <td>NK_cells_CD16+_FGFBP2+</td> </tr> <tr> <td>Non-classical_Monocytes</td> </tr> <tr> <td>T_reg</td> </tr> <tr> <td>Classical_Monocytes</td> </tr> <tr> <td>Cycling_T+NK_cells</td> </tr> <tr> <td>DC</td> </tr> <tr> <td>B_cells</td> </tr> <tr> <td>Plasma_cells</td> </tr> <tr> <td>Mast_cells</td> </tr> <tr> <td>pDC</td> </tr> <tr> <td>P-value</td> </tr> <tr> <td>Correlation</td> </tr> <tr> <td>RMSE</td> </tr> </tbody> </table>	cell_deconvolution	Sample_ID	Tip_cells_Angiogenesis	Endothelial_cells	Stromal_cells	Tumor_cells	Cycling_Tumor_cells	CD45_pos	Tumor_cells_CA9_high	CD4_T_cells	Intermediate_Monocytes	NK_cells_CD16-_FGFBP2-	CD8_T_cells	NK_cells_CD16+_FGFBP2+	Non-classical_Monocytes	T_reg	Classical_Monocytes	Cycling_T+NK_cells	DC	B_cells	Plasma_cells	Mast_cells	pDC	P-value	Correlation	RMSE
drug_category_and_molecule																																																																																	
SUBJID																																																																																	
Arm																																																																																	
Received Prior Therapy																																																																																	
Number of Prior Therapies																																																																																	
tumor_feature																																																																																	
SUBJID																																																																																	
Days_from_TumorSample_Collection_and_Start_of_Trial_Therapy																																																																																	
Tumor_Sample_Primary_or_Metastasis																																																																																	
Site of Metastasis																																																																																	
ImmunoPhenotype																																																																																	
Purity																																																																																	
Ploidy																																																																																	
TMB Counts																																																																																	
TM Area																																																																																	
TM CD8																																																																																	
TM CD8 Density																																																																																	
TC Area																																																																																	
TC CD8																																																																																	
TC CD8 Density																																																																																	
TM TC Ratio																																																																																	
TM CD8 PERCENT																																																																																	
TC CD8 PERCENT																																																																																	
Benefit																																																																																	
ORR																																																																																	
PFS																																																																																	
PFS CNSR																																																																																	
OS																																																																																	
OS CNSR																																																																																	
gene_mutations																																																																																	
Sample ID																																																																																	
Chromosome																																																																																	
Start position																																																																																	
End position																																																																																	
Variant Classification																																																																																	
Variant Type																																																																																	
cDNA Change																																																																																	
Codon Change																																																																																	
Protein Change																																																																																	
Reference Allele																																																																																	
Tumor Seq Allele1																																																																																	
Tumor Seq Allele2																																																																																	
Tumor ref count																																																																																	
Tumor alt count																																																																																	
gene name																																																																																	
SUBJID																																																																																	
biological_pathways																																																																																	
gene set																																																																																	
Sample_ID* (each columns has the ID of a Sample)																																																																																	
method																																																																																	
MSigBD gene set																																																																																	
cell_deconvolution																																																																																	
Sample_ID																																																																																	
Tip_cells_Angiogenesis																																																																																	
Endothelial_cells																																																																																	
Stromal_cells																																																																																	
Tumor_cells																																																																																	
Cycling_Tumor_cells																																																																																	
CD45_pos																																																																																	
Tumor_cells_CA9_high																																																																																	
CD4_T_cells																																																																																	
Intermediate_Monocytes																																																																																	
NK_cells_CD16-_FGFBP2-																																																																																	
CD8_T_cells																																																																																	
NK_cells_CD16+_FGFBP2+																																																																																	
Non-classical_Monocytes																																																																																	
T_reg																																																																																	
Classical_Monocytes																																																																																	
Cycling_T+NK_cells																																																																																	
DC																																																																																	
B_cells																																																																																	
Plasma_cells																																																																																	
Mast_cells																																																																																	
pDC																																																																																	
P-value																																																																																	
Correlation																																																																																	
RMSE																																																																																	
<table border="1"> <thead> <tr> <th>biological_feature</th> </tr> </thead> <tbody> <tr> <td>SUBJID</td> </tr> <tr> <td>Sexe</td> </tr> <tr> <td>Age</td> </tr> <tr> <td>Cohort</td> </tr> </tbody> </table>	biological_feature	SUBJID	Sexe	Age	Cohort			<table border="1"> <thead> <tr> <th>gene_expression</th> </tr> </thead> <tbody> <tr> <td>Sample ID</td> </tr> <tr> <td>gene_name* (each columns has the name of gene)</td> </tr> </tbody> </table>	gene_expression	Sample ID	gene_name* (each columns has the name of gene)	<table border="1"> <thead> <tr> <th>structural_variant</th> </tr> </thead> <tbody> <tr> <td>CNV ID</td> </tr> <tr> <td>RNA ID</td> </tr> <tr> <td>Amplification_Peak_* /</td> </tr> <tr> <td>Deletion_Peak_*</td> </tr> <tr> <td>Amplification_Peak_*-CN_values /</td> </tr> <tr> <td>Deletion_Peak_*-CN_values</td> </tr> </tbody> </table>	structural_variant	CNV ID	RNA ID	Amplification_Peak_* /	Deletion_Peak_*	Amplification_Peak_*-CN_values /	Deletion_Peak_*-CN_values																																																														
biological_feature																																																																																	
SUBJID																																																																																	
Sexe																																																																																	
Age																																																																																	
Cohort																																																																																	
gene_expression																																																																																	
Sample ID																																																																																	
gene_name* (each columns has the name of gene)																																																																																	
structural_variant																																																																																	
CNV ID																																																																																	
RNA ID																																																																																	
Amplification_Peak_* /																																																																																	
Deletion_Peak_*																																																																																	
Amplification_Peak_*-CN_values /																																																																																	
Deletion_Peak_*-CN_values																																																																																	
<table border="1"> <thead> <tr> <th>pronotic_scores</th> </tr> </thead> <tbody> <tr> <td>SUBJID</td> </tr> <tr> <td>MSKCC</td> </tr> <tr> <td>IMDC</td> </tr> </tbody> </table>	pronotic_scores	SUBJID	MSKCC	IMDC		<table border="1"> <thead> <tr> <th>gene_reference</th> </tr> </thead> <tbody> <tr> <td>gene</td> </tr> <tr> <td>hgnc_symbol</td> </tr> </tbody> </table>	gene_reference	gene	hgnc_symbol																																																																								
pronotic_scores																																																																																	
SUBJID																																																																																	
MSKCC																																																																																	
IMDC																																																																																	
gene_reference																																																																																	
gene																																																																																	
hgnc_symbol																																																																																	
<table border="1"> <thead> <tr> <th>HLA_subtype</th> </tr> </thead> <tbody> <tr> <td>SUBJID</td> </tr> <tr> <td>HLA subtype</td> </tr> </tbody> </table>	HLA_subtype	SUBJID	HLA subtype																																																																														
HLA_subtype																																																																																	
SUBJID																																																																																	
HLA subtype																																																																																	

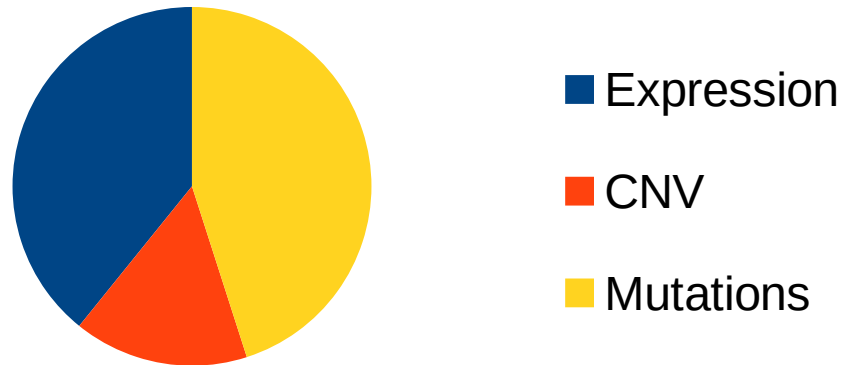
joinTables									
<table border="1"> <thead> <tr> <th>SUBJID</th> </tr> </thead> <tbody> <tr> <td>drug category and molecule</td> </tr> <tr> <td>biological feature</td> </tr> <tr> <td>pronotic scores</td> </tr> <tr> <td>tumor feature</td> </tr> <tr> <td>HLA subtype</td> </tr> </tbody> </table>	SUBJID	drug category and molecule	biological feature	pronotic scores	tumor feature	HLA subtype			
SUBJID									
drug category and molecule									
biological feature									
pronotic scores									
tumor feature									
HLA subtype									
<table border="1"> <thead> <tr> <th>Sample</th> </tr> </thead> <tbody> <tr> <td>drug category and molecule</td> </tr> <tr> <td>biological feature</td> </tr> <tr> <td>pronotic scores</td> </tr> <tr> <td>tumor feature</td> </tr> <tr> <td>gene mutations</td> </tr> <tr> <td>biological pathways</td> </tr> <tr> <td>structural variant</td> </tr> <tr> <td>cell deconvolution</td> </tr> </tbody> </table>	Sample	drug category and molecule	biological feature	pronotic scores	tumor feature	gene mutations	biological pathways	structural variant	cell deconvolution
Sample									
drug category and molecule									
biological feature									
pronotic scores									
tumor feature									
gene mutations									
biological pathways									
structural variant									
cell deconvolution									
<table border="1"> <thead> <tr> <th>All</th> </tr> </thead> <tbody> <tr> <td>every tables</td> </tr> </tbody> </table>	All	every tables							
All									
every tables									

Data Lakehouse



Repartition of experiments

2287 Samples



Repartition of Treatments

1478 Subjects



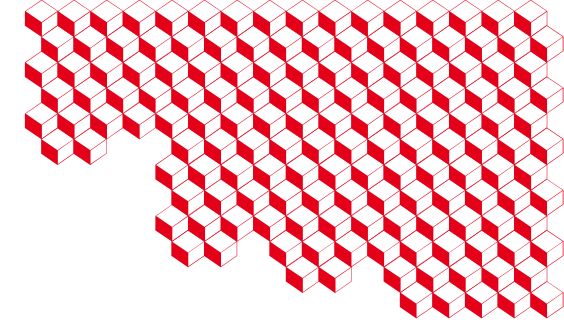
Prototyping from two data sets combining molecular profiling of kidney tumor tissues with clinical drug trials:

- 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.



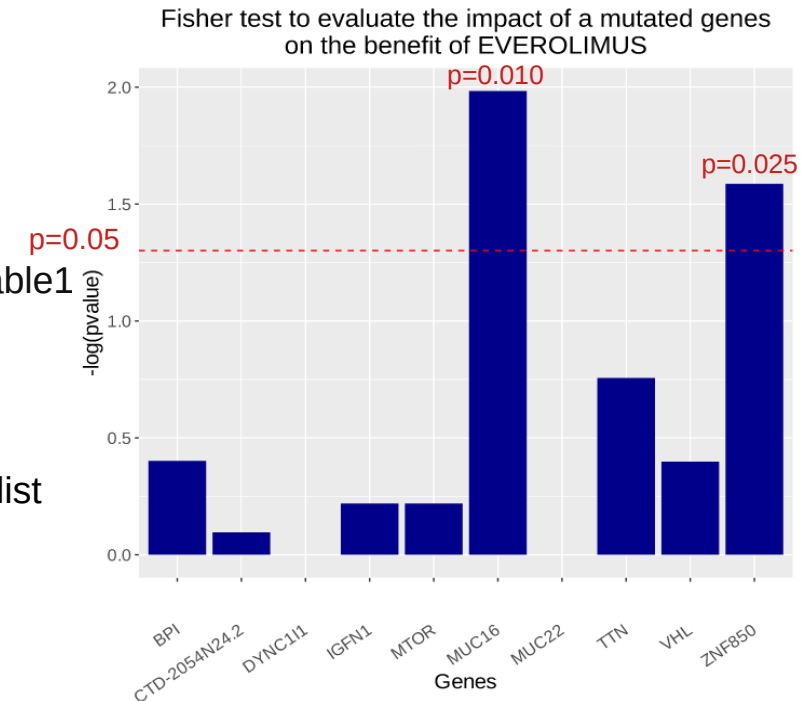
3 ■ Data queries

Data query - Use case 1

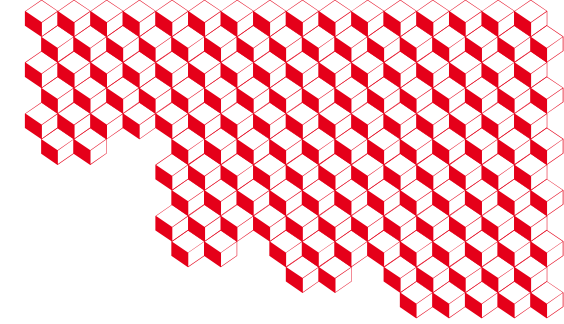


Investigate whether **genes frequently mutated in kidney cancer** are associated with **patient response to treatment**

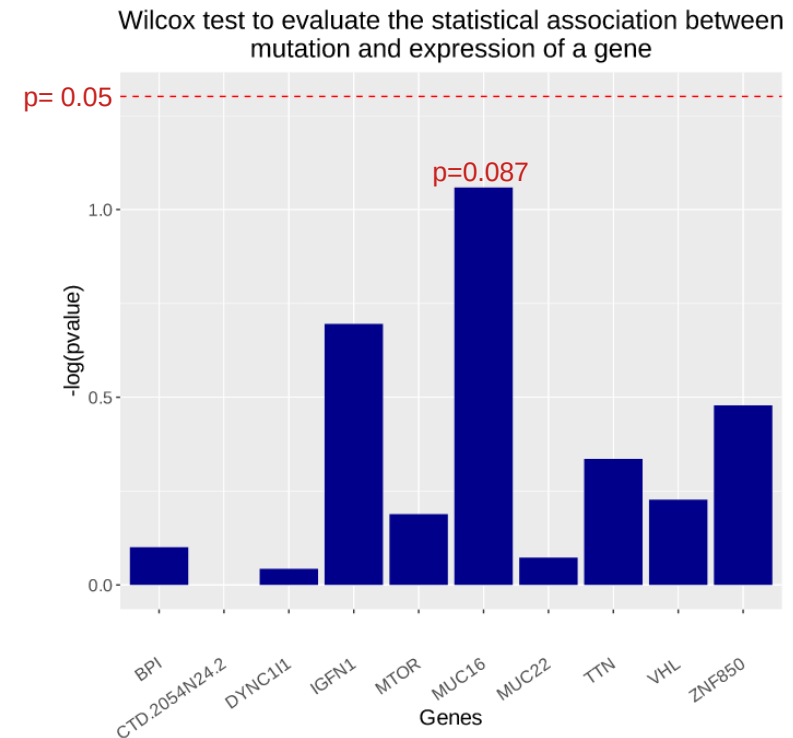
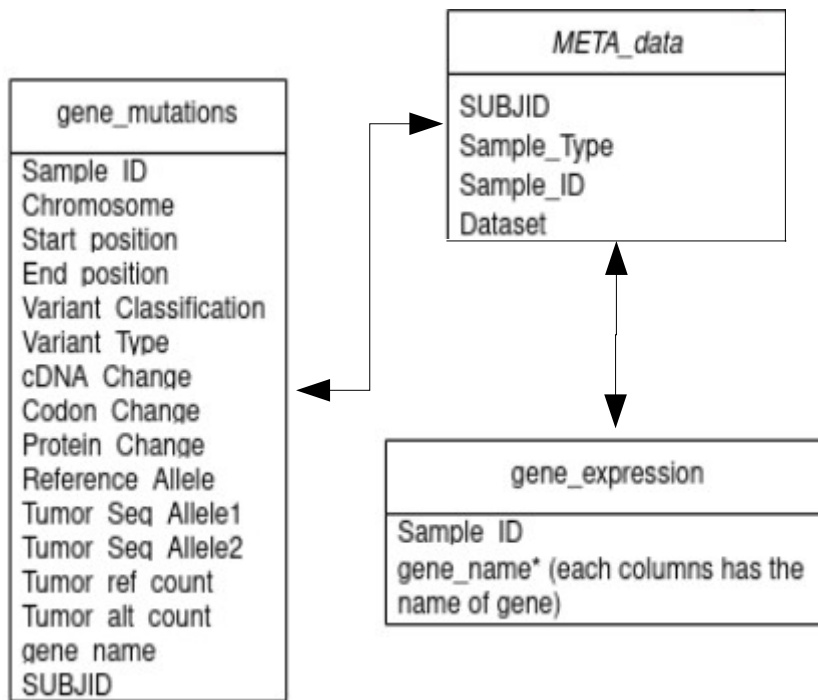
- 1) **SELECT** gene_name,count(**gene_name**) **FROM** DELTA.`gene_mutations`
WHERE (Variant_Type == '**SNP**' **AND** Variant_Classification == '**Missense_Mutation**')
GROUP BY gene_name **ORDER BY** count(gene_name) **DESC LIMIT 10**
- 2) **SELECT** table1.SUBJID, table1.Sample_ID, table2.benefit **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



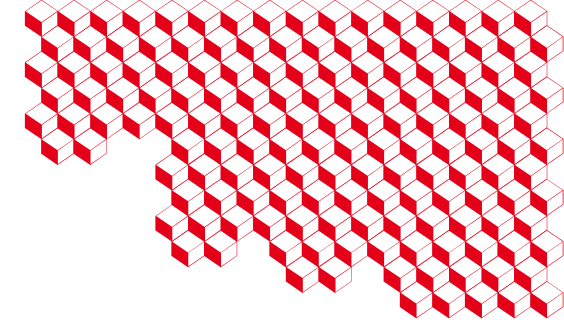
Data query - Use case 2



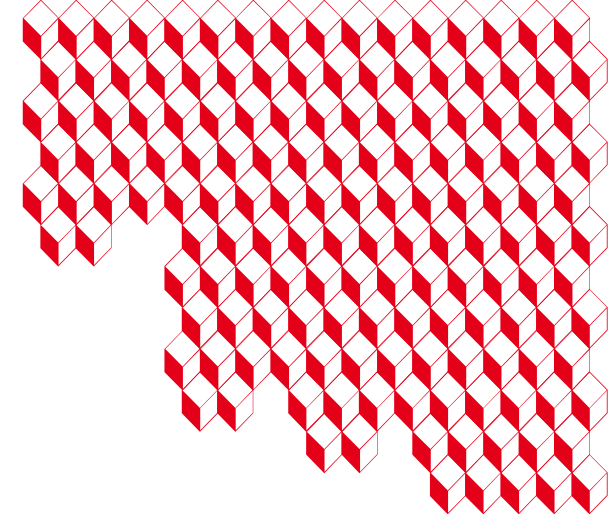
Investigate whether, for frequently mutated genes in kidney cancer, **the mutation status** is associated with a **change in gene expression**



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 immunotherapies 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



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

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

CEA GRENOBLE
17 avenue des Martyrs
38000 Grenoble, France