GeCo 5.0

Shoot the piano players:

Anna Bernasconi, Arif Canakoglu, Stefano Ceri, Pietro Pinoli

POLITECNICO
MILANO 1863

Present – Aug 2021
New glue of group's activity…
…GeCo 5.0!

Sep 2016 – Present
advanced grant ERC «Data-driven Genomic Computing»

Dec 2019
GMQL federated

Jul 2019
GenoSurf

Nov 2019
pyGMQL

Nov 2017
Genomic Conceptual Model

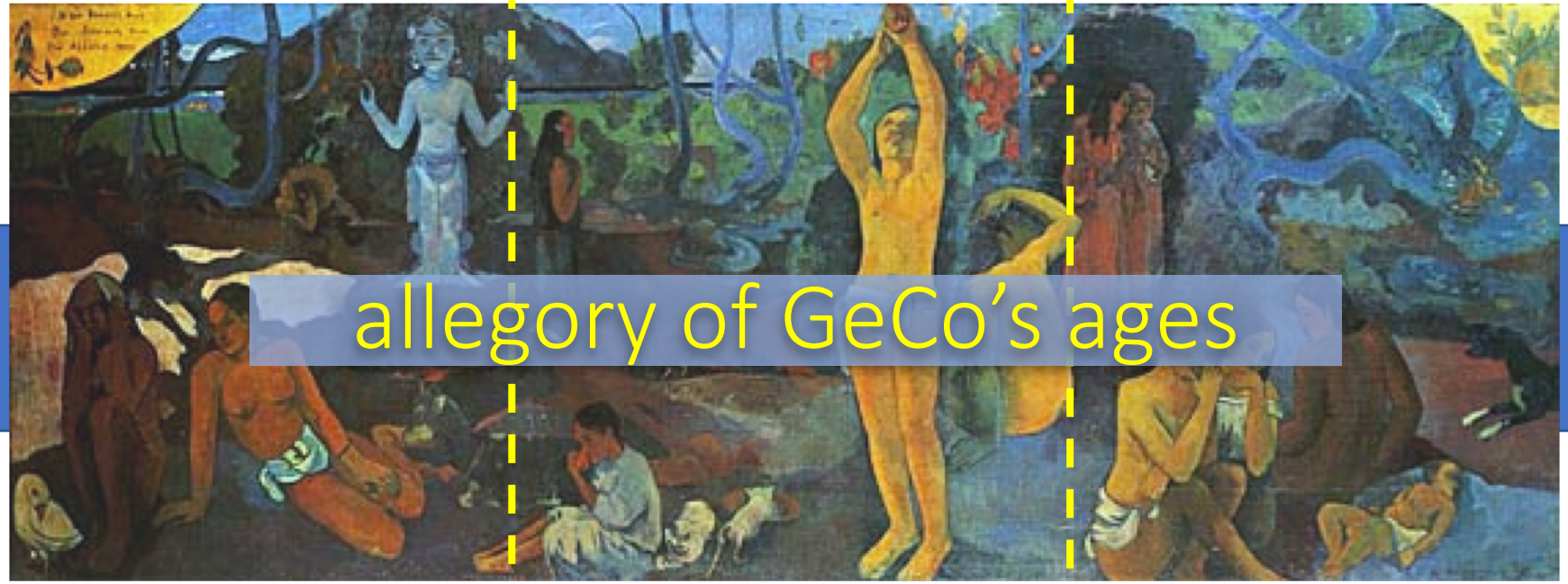
Jan 2015
first release of GMQL V1 (at Polimi and IEO-IIT)

Sep 2012: Start of collaboration with IEO-IIT

Mar 2015
first accepted paper on BioInformatics

Mar 2013 – Feb 2016
PRIN Project Gendata 2020 (with Sapienza, Roma3, Unibo, PoliTo,…)

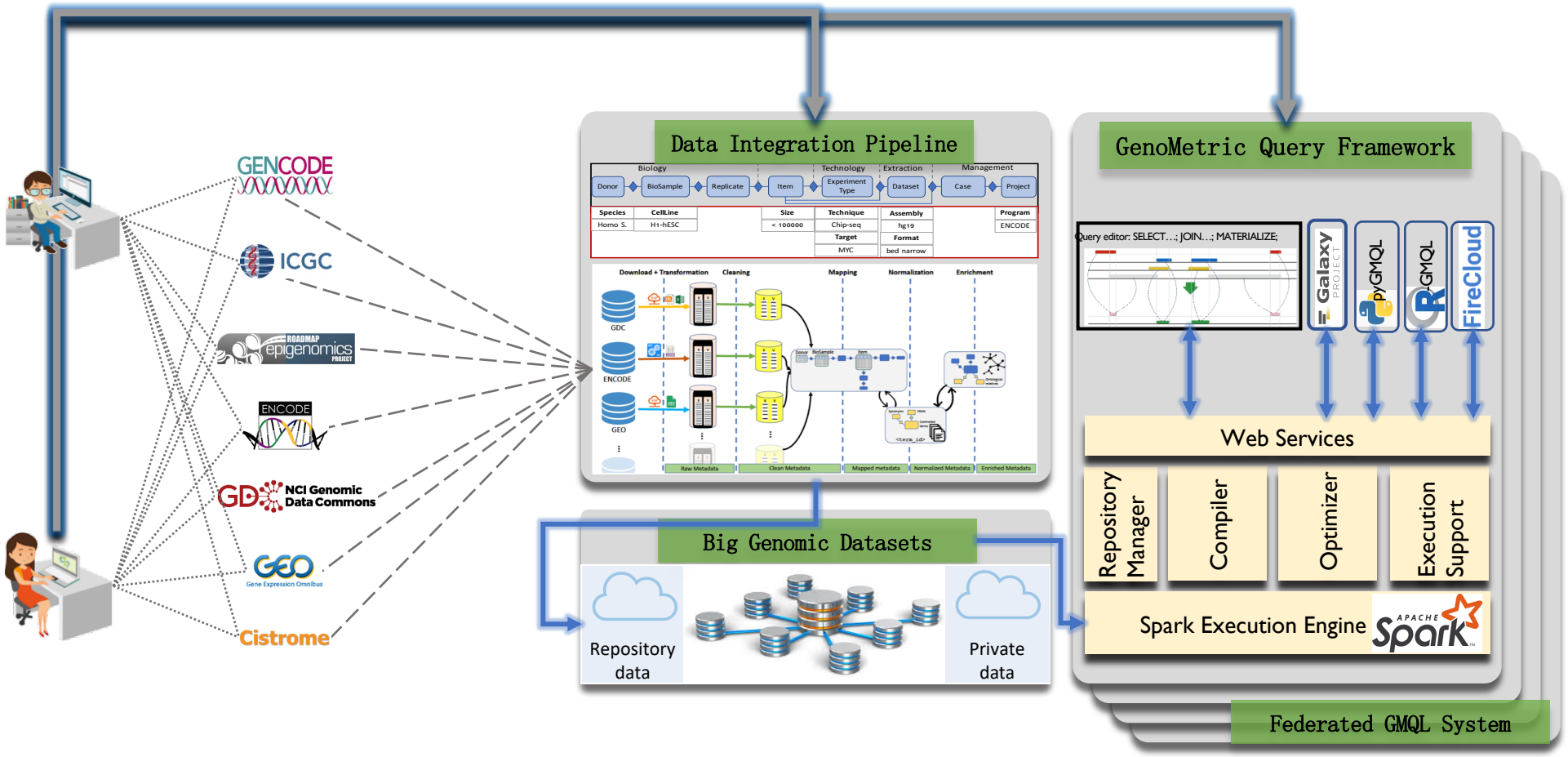
Where are we going? *Where are we?* *Where do we come from?*



Paul Gauguin, 1897. Museum of Fine Arts, Boston, USA.

POLITECNICO MILANO 1863

Bird eye's view of GeCo



POLITECNICO
MILANO 1863

GeCo abstractions

Data Model (GDM)

- GDM Objects = [Regions + Metadata]
- Regions = [Coordinates + Feature Vector]
  - Key-value in a rich space
  - Embedding most data types representing processed data (Expression, Mutations, Peaks, Annotations)
- Metadata = [Attribute-Value] pairs
  - Mediation/simplification due to the complexity of sources

Query Language (GMQL)

- Algebra of GDM objects
- Based on classical relational abstractions
  - Unary: SELECT, PROJECT, EXTEND, ORDER, GROUP, MERGE, COVER
  - Binary: UNION, DIFFERENCE, MAP, JOIN
- Domain specific qualities
  - Support of genomic distances and rich predicate calculus
  - Support of intersections, histograms, overlap counts
  - Support of grouping, metadata enrichment on-the-fly

Conceptual Model (GCM)

- Centered on the notion of sample=file=Item
- Subschemas:
  - BIOLOGY: Donor, BioSample, Replicate
  - MANAGEMENT: Process, CaseStudy
  - TECHNOLOGY: ExperimentType
  - EXTRACTION: Dataset
- Ten attributes are normalized and ontologically enriched

GenoMetric Query Framework

Galaxy PROJECT | pyGMQL | R-GMQL | FireCloud

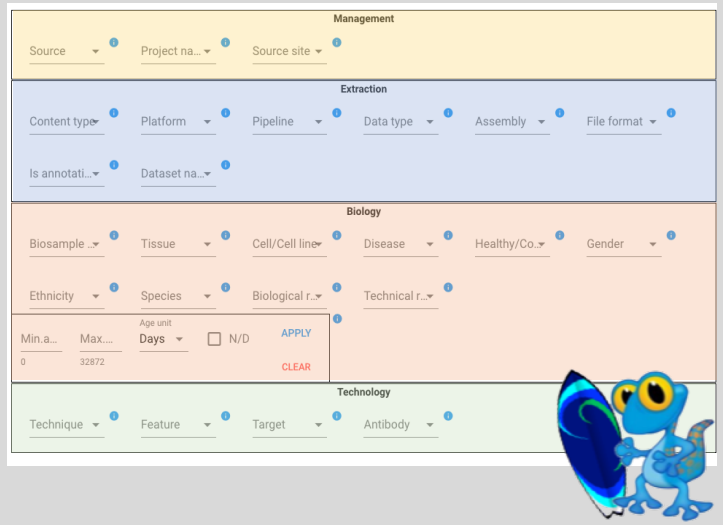
Web Services

Repository Manager | Compiler | Optimizer | Execution Support

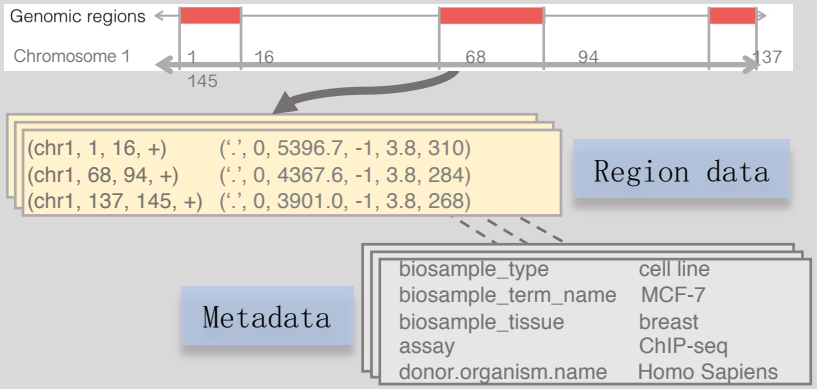
Spark Execution Engine — APACHE Spark

Federated GMQL System

GenoSurf Interface

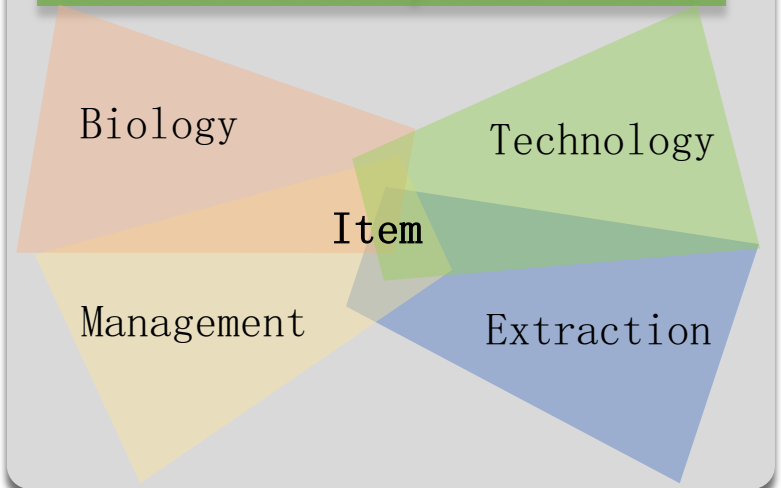


Genomic Data Model



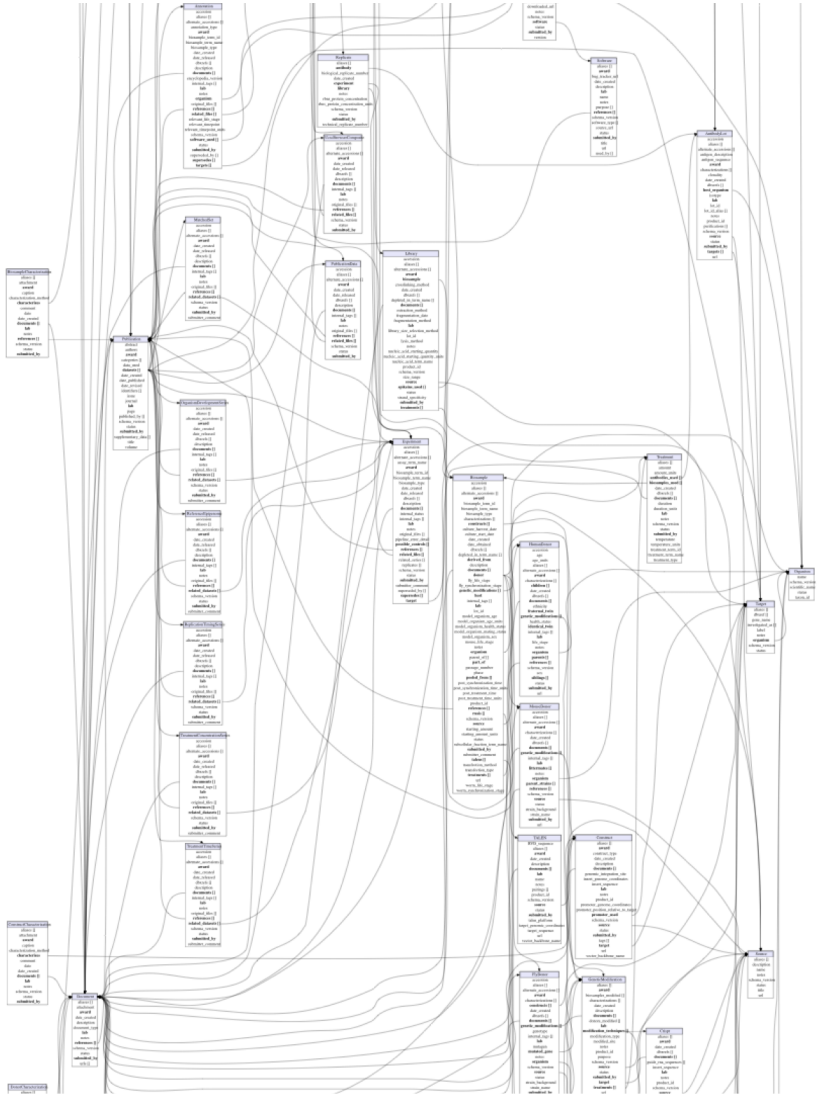
GenoMetric Query Language

Genomic Conceptual Model

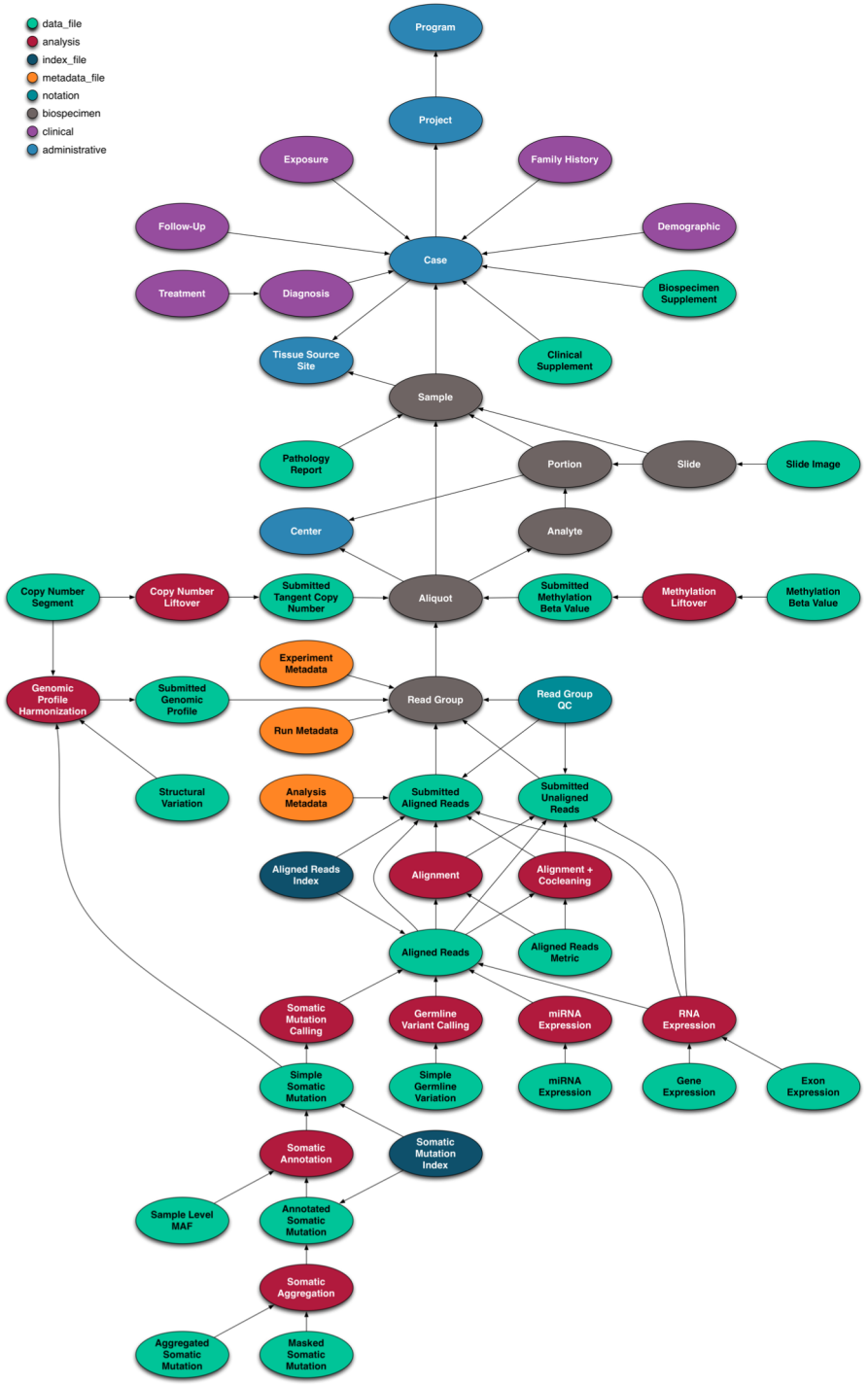


Today’s data models’ mess

Encode



Genomic Data Commons (TCGA + TARGET)



GeCo results so far

Model, Language and System

1. **Modeling and interoperability of heterogeneous genomic big data for integrative processing and querying.** Methods, 2016
2. **GenoMetric Query Language: A novel approach to large-scale genomic data management.** Bioinformatics 2015
3. **Multi-Dimensional Genomic Data Management for Region-Preserving Operations.** IEEE ICDE, 2019.
4. **Metadata Management for Scientific Databases.** Information Systems Elsevier, 2018.
5. **Processing of big heterogeneous genomic datasets for tertiary analysis of NGS data.** Bioinformatics, 2018.
6. **Optimal Binning for Genomics.** IEEE TC, 2018.
7. **Framework for Supporting Genomic Operations.** IEEE TC, 2016.
8. **Data management for heterogeneous genomic datasets.** IEEE/ACM TCBB, 2016.
9. **Evaluating Cloud Frameworks on Genomic Applications.** IEEE Big Data, 2015.

Metadata and Repository

1. **GenoSurf: Metadata driven semantic search server for integrated genomic datasets.** Database, 2019.
2. **From a Conceptual Model to a Knowledge Graph for Genomic Datasets**. ER, 2019.
3. **Conceptual Modeling for Genomics: Building an Integrated Repository of Open Data.** ER, 2017.
4. **TCGA2BED: extracting, extending, integrating, and querying The Cancer Genome Atlas.** BMC Bioinformatics, 2017.
5. **Ontology-based search of genomic metadata.** IEEE/ACM TCBB, 2016.
6. **Exploiting Conceptual Modeling for Searching Genomic Metadata: A Quantitative and Qualitative Empirical Study.** ER Workshops, 2019.
7. **Ontology-Driven Metadata Enrichment for Genomic Datasets.** SWAT4HCLS, 2018.
8. **Using Metadata for Locating Genomic Datasets on a Global Scale.** CIKM Workshops, 2018.

Interfaces and Tools

1. **PyGMQL: scalable data extraction and analysis for heterogeneous genomic datasets.** BMC Bioinformatics, 2019.
2. **Demonstration of GenoMetric Query Language.** CIKM, 2018.
3. **Exploring Genomic Datasets: from Batch to Interactive and Back.** ExploreDB'18, 2018.
4. **Next Generation Indexing for Genomic Intervals.** IEEE TKDE, 2018.
5. **Explorative visual analytics on interval-based genomic data and their metadata.** BMC Bioinformatics, 2017.
6. **Indexing Next-Generation Sequencing data.** Information Sciences Elsevier, 2016.
7. **Pattern Similarity Search in Genomic Sequences.** IEEE TKDE, 2016.
8. **Analysis and Visualization of Mutation Enrichments for Selected Genomic Regions and Cancer Types.** IEEE BIBM, 2019.

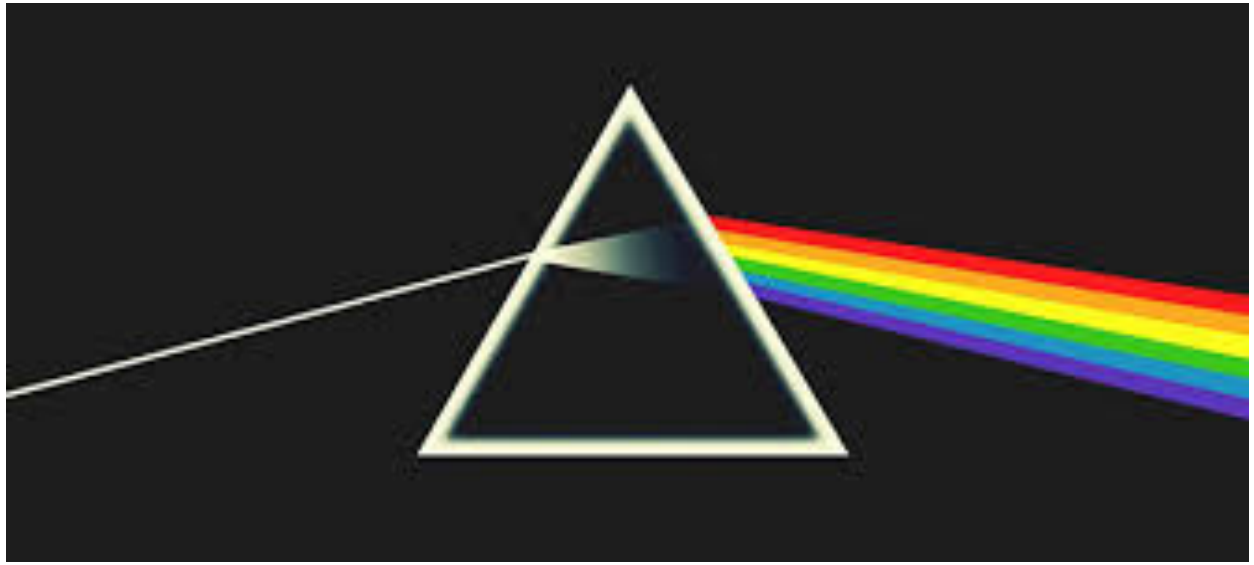
Applications

1. **Non-negative Matrix Tri-Factorization for Data Integration and Network-based Drug Repositioning.** IEEE CIBCB, 2019
2. **Drug Repositioning Predictions by Non-Negative Matrix Tri-Factorization of Integrated Association Data.** ACM BCB, 2019
3. **Association rule mining to identify transcription factor interactions in genomic regions.** Bioinformatics, 2019.
4. **Extensive epigenomic integration of the glucocorticoid response in primary human monocytes and in vitro derived macrophages.** Scientific Reports, 2019.
5. **Combining DNA methylation and RNA sequencing data of cancer for supervised knowledge extraction.** BioData Mining, 2018.
6. **Identifying Collateral and Synthetic Lethal Vulnerabilities within the DNA-damage Response.** Making it Personal: Cancer Precision Medicine, 2018.
7. **Designing and Evaluating Deep Learning Methods for Cancer Classification on Gene Expression Data.** CIBB, 2018.
8. **TICA: Transcriptional Interaction and Coregulation Analyser.** Genomics, Proteomics and Bioinformatics, 2018.
9. **Implementing a Transcription Factor Interaction Prediction System Using the GenoMetric Query Language.** Data Mining for Systems Biology, 2018.
10. **Modeling gene transcriptional regulation by means of hyperplanes genetic clustering.** IEEE IJCNN, 2018.
11. **Exploiting Ladder Networks for Gene Expression Classification.** IWBBIO, 2018.
12. **Impact of mutational signatures on microRNA and their response elements.** PSB, 2020.
13. **Deleterious Impact of Mutational Processes on Transcription Factor Binding Sites in Human Cancer.** IEEE BIBE, 2019.
14. **Analysis of Gene Regulatory Networks Inferred from ChIP-seq Data.** IWBBIO, 2019

GeCo 5.0 objective

WHAT

A high-level, user-friendly collection of methods, concepts, interactions, that use our background (and not only!) for doing concrete biological tasks



framework

interaction paradigm

applications

emphasis on statistics and machine learning

TARGET

the *PhD biological researcher* who does not want to learn how to query/program

POLITECNICO MILANO 1863

GeCo 5.0 change of paradigm

BEFORE

- Top-down
- Applications used to verify our model/system

NOW

- Bottom-up, Application-driven
  - Applications as first-class citizens to feed and consolidate the framework
- Workflow-driven approach, two stages:
  - Define data
  - Choose patterns

Pattern: Typical biological question a genomic researcher tries to answer:

- Identify characteristics of the genomic regions;
- Select features and find classifiers for different populations;
- Enrich genomic regions;

…

POLITECNICO
MILANO 1863

Novelties

Focus on **interpretability** of results

Re-use past pattern searches with different set of samples/ binning/ indexing

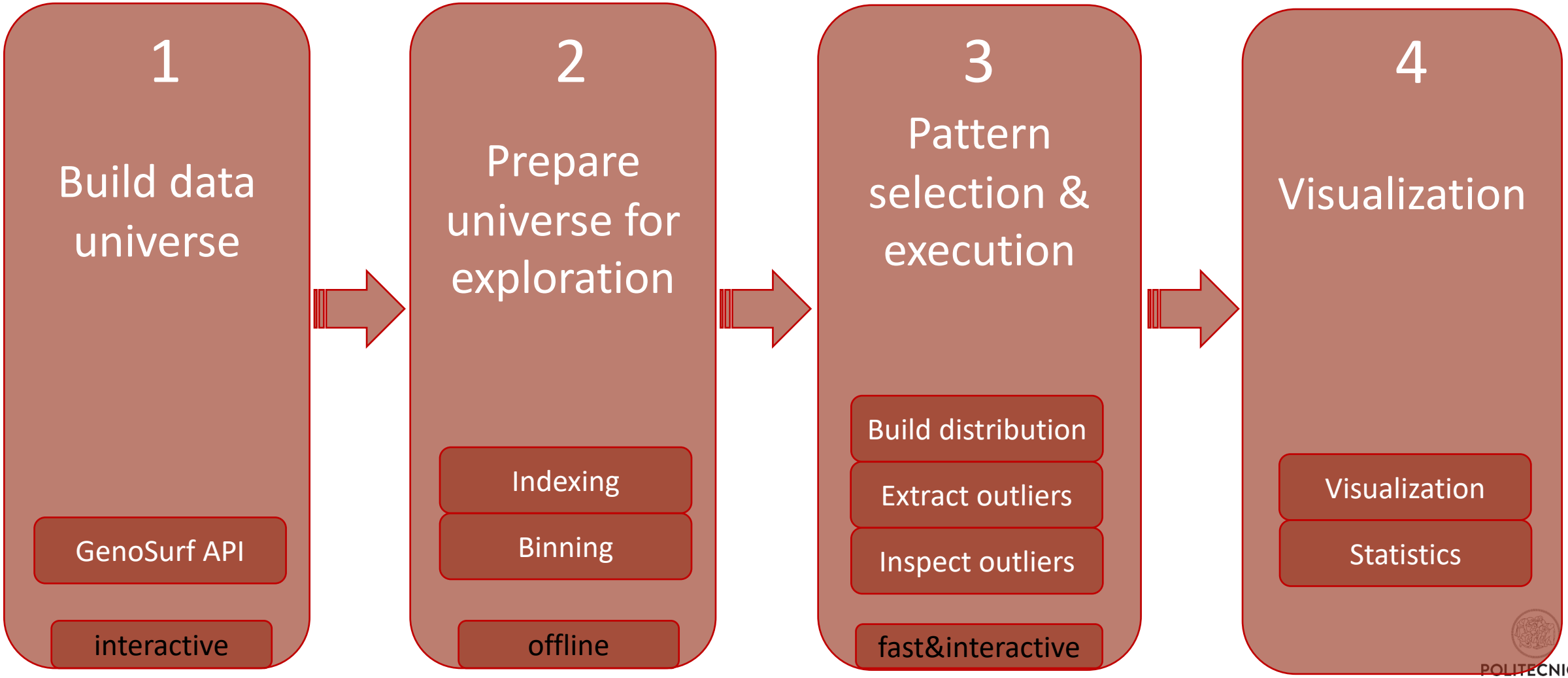
Fast pattern-based **data exploration**

Create a **direct link** between the extraction and the analysis process

Help researchers to create **novel hypotheses**

POLITECNICO MILANO 1863

Workflow



1 Defining objectives

2 Inspecting universe

3 Defining procedures

4 Inspecting results

Understanding data / **WHAT:** Conceptual model-driven **FOR WHAT:** Support of statistical abstraction

Organizing data / **"TABLING" INDEXING BINNING**

BACKGROUND

Repository of data

GenoSurf

DOMAIN UNIVERSE

Data visualization and exploration

Understanding solutions / **WHICH:** Procedure

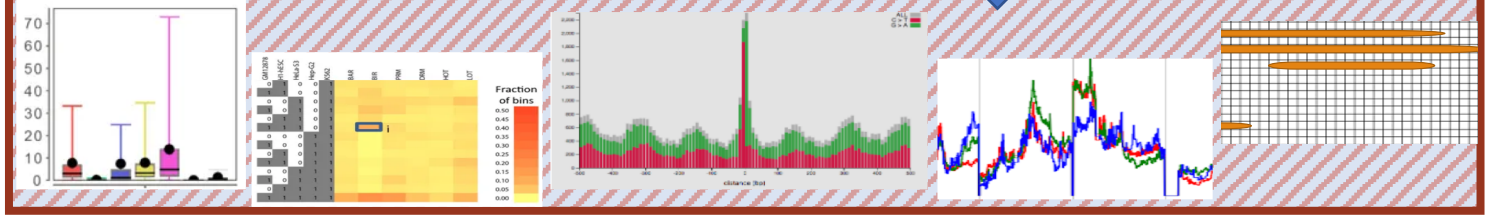
Formalizing solutions / **WHICH:** Parameter setting

BACKGROUND

Repository of solutions

Stats + ML

GMQL ABSTRACTIONS



1 Defining objectives

Repository of data

GenoSurf

Summary of interaction

Extraction parameters

3 Defining procedures

Repository of solutions

Stats + ML

Summary of procedures

Computation parameters

GenoSurf_pyGMQL_demo.

Code

jupyter

GeCo 5.0 example notebook

API GenoSurf

pyGMQL

[ ]:

matplotlib

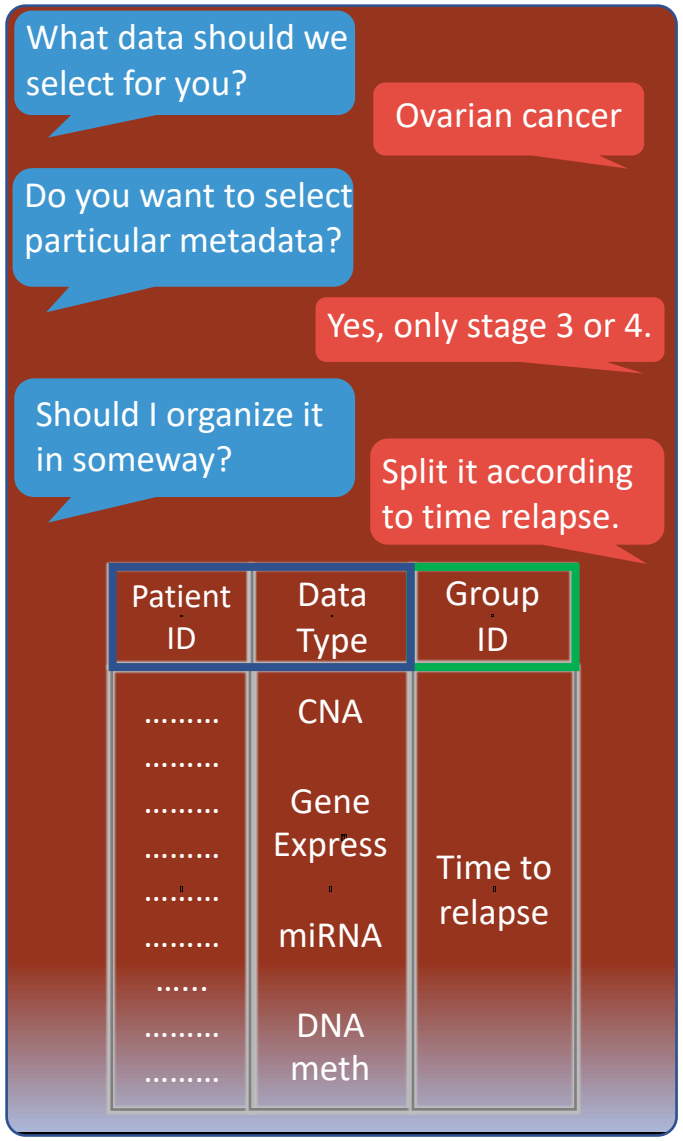
scikit learn

TensorFlow

[ ]:

MILANO 1863

Worked-out example: find outliers for HGS-OC



TCGA_ovarian_cancer_hg19 | **3rd and 4th stage patients** | **grouped by time relapse**

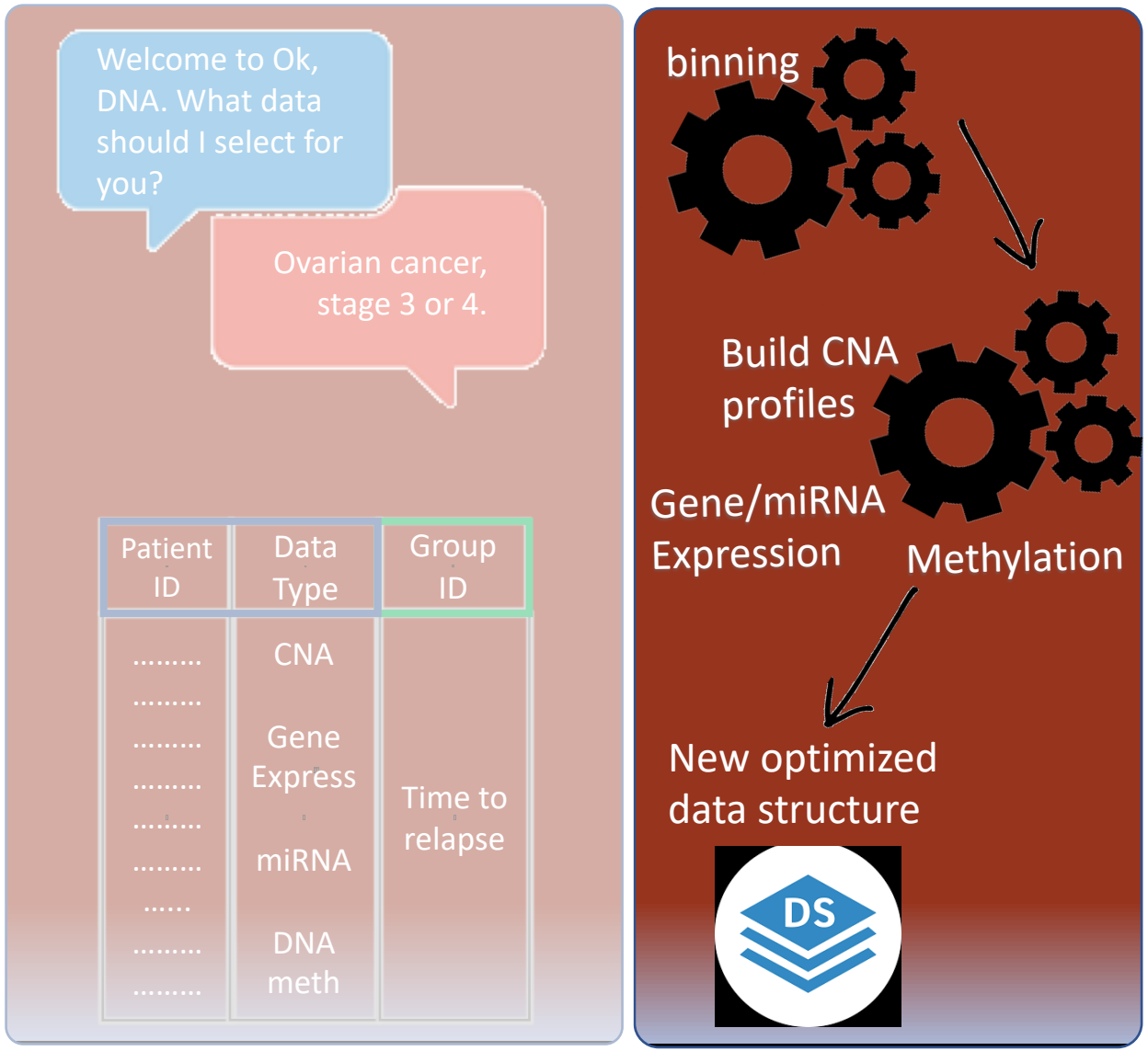
Group by time relapse

Group by health status

Conversational interface idea by Pietro Crovari & Co.

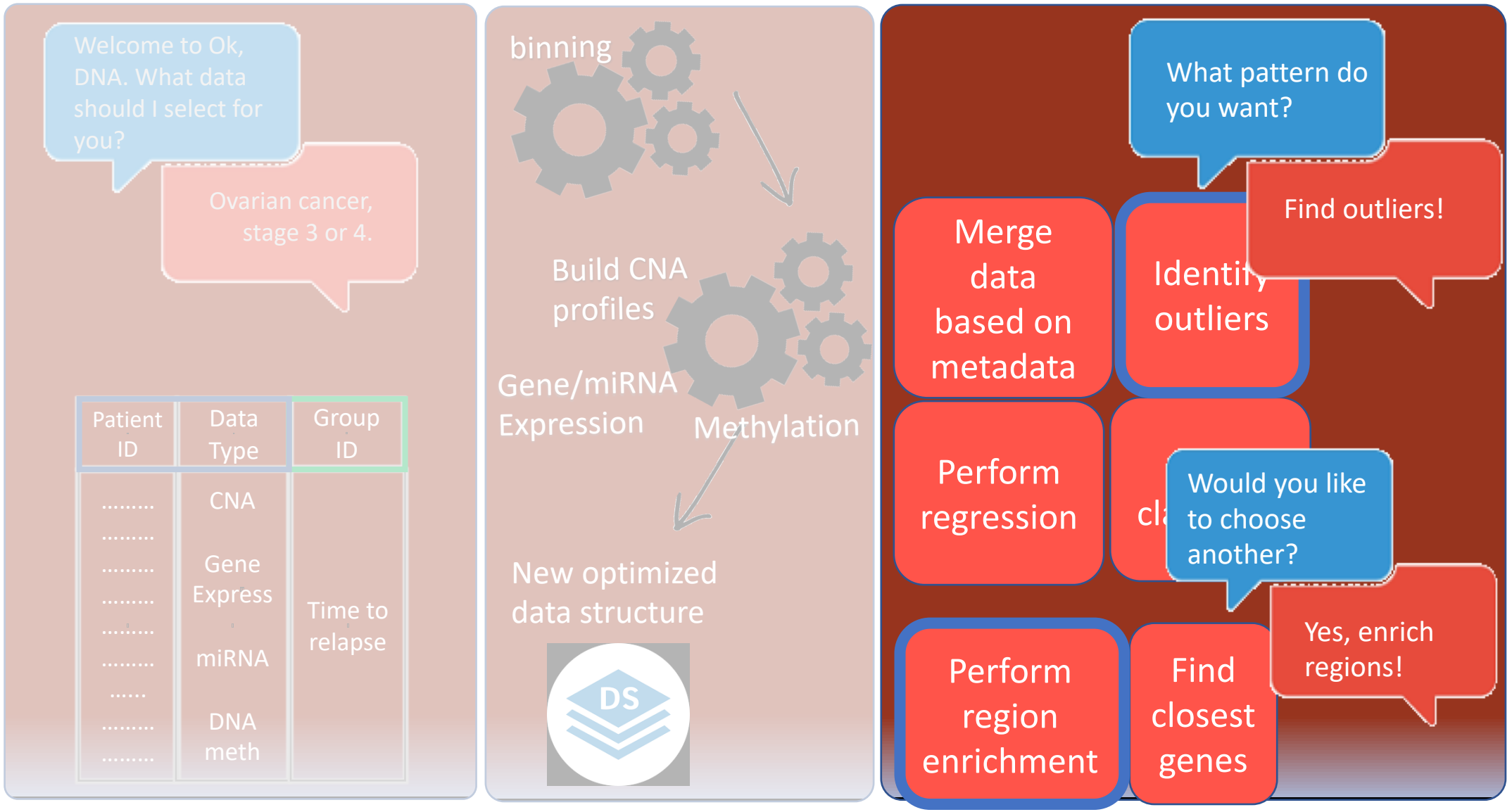
POLITECNICO MILANO 1863

Worked-out example: find outliers for HGS-OC



POLITECNICO MILANO 1863

Worked-out example: find outliers for HGS-OC



POLITECNICO MILANO 1863

Worked-out example: find outliers for HGS-OC

Welcome to Ok, DNA. What data should I select for you?

Ovarian cancer, stage 3 or 4.

Patient ID	Data Type	Group ID
.....	CNA	
.....		
.....	Gene Express	
.....		
.....		Time to relapse
.....	miRNA	
.....		
.....	DNA meth	
.....		

binning

Build CNA profiles

Gene/miRNA Expression

Methylation

New optimized data structure

What pattern do you want?

Merge data based on metadata

Identify outliers

Find outliers!

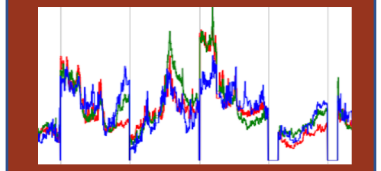
Perform regression

Would you like to chose another?

Perform region enrichment

Find closest genes

Yes, enrich regions!



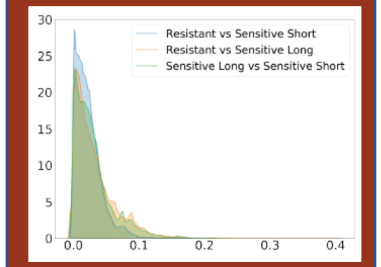


TABLE I
BEST PERFORMANCES FOR RESISTANT VS SENSITIVE USING CNA, GENE EXPRESSION, MIRNA AND DNA METHYLATION, SEPARATELY, AND THEN MERGING THEM

Type of data	N° features	Precision	Recall	Accuracy	AUC
CNA	225	0.51 ± 0.10	0.61 ± 0.19	0.68 ± 0.07	0.72 ± 0.11
Gene expression	20	0.71 ± 0.20	0.37 ± 0.10	0.77 ± 0.10	0.79 ± 0.11
miRNA	11	0.77 ± 0.30	0.37 ± 0.20	0.75 ± 0.10	0.72 ± 0.15
Methylation	65	0.79 ± 0.30	0.35 ± 0.10	0.78 ± 0.10	0.78 ± 0.09
Merge	311	0.68 ± 0.18	0.74 ± 0.11	0.80 ± 0.10	0.82 ± 0.09

TABLE II
BEST PERFORMANCES FOR RESISTANT VS SENSITIVE SHORT USING CNA, GENE EXPRESSION, MIRNA AND DNA METHYLATION, SEPARATELY, AND THEN MERGING THEM

Type of data	N° features	Precision	Recall	Accuracy	AUC
CNA	236	0.53 ± 0.11	0.57 ± 0.11	0.64 ± 0.09	0.69 ± 0.09
Gene expression	40	0.72 ± 0.20	0.55 ± 0.20	0.76 ± 0.10	0.79 ± 0.15
miRNA	12	0.70 ± 0.30	0.50 ± 0.30	0.74 ± 0.10	0.76 ± 0.16
Methylation	31	0.65 ± 0.30	0.32 ± 0.20	0.69 ± 0.10	0.80 ± 0.09
Merge	310	0.80 ± 0.16	0.65 ± 0.14	0.82 ± 0.09	0.83 ± 0.10

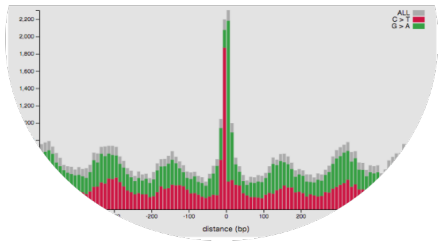
TABLE III
BEST PERFORMANCES FOR RESISTANT VS SENSITIVE LONG USING CNA, GENE EXPRESSION, MIRNA AND DNA METHYLATION, SEPARATELY, AND THEN MERGING THEM

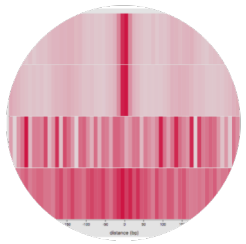
Type of data	N° features	Precision	Recall	Accuracy	AUC
CNA	153	0.89 ± 0.07	0.88 ± 0.10	0.83 ± 0.08	0.90 ± 0.08
Gene expression	26	0.86 ± 0.20	0.93 ± 0.10	0.85 ± 0.20	0.86 ± 0.15
miRNA	21	0.77 ± 0.10	0.90 ± 0.10	0.75 ± 0.10	0.84 ± 0.17
Methylation	18	0.79 ± 0.10	0.90 ± 0.10	0.77 ± 0.10	0.84 ± 0.17
Merge	213	0.86 ± 0.09	0.95 ± 0.08	0.84 ± 0.08	0.91 ± 0.10

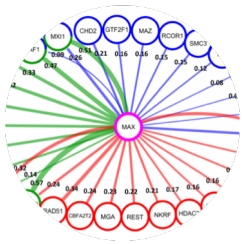
Example patterns

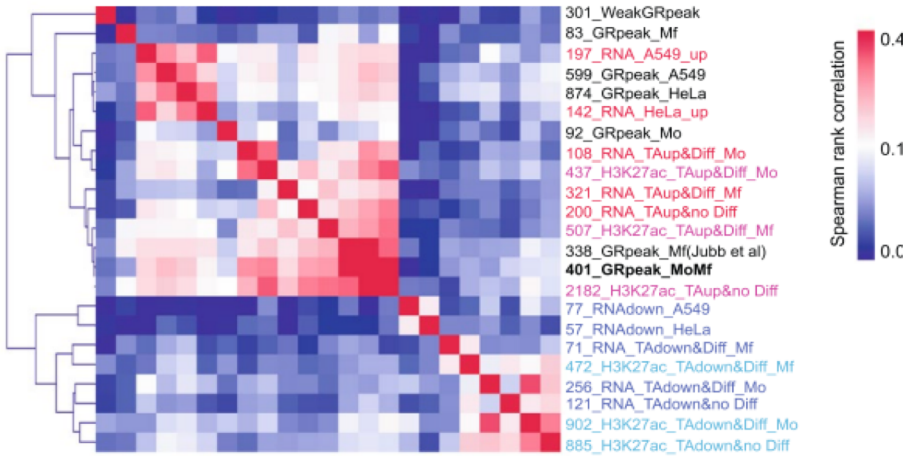
1. Provide general statistics
2. Identify and inspect outliers
3. Merge data based on metadata (e.g., patientID, tissue, …)
4. Find distribution of region-distance-based measures (e.g., distance oncogenes-tads boundary)
5. Find classifiers to distinguish populations (e.g., healthy vs tumor, cancer subtyping, …)
6. Classify regions predicting their function (e.g., how likely a region is to be a gene)
7. Perform regression (e.g., survival analysis)
8. Provide concise view of all characteristics of specific region/bin
9. Find co-occurrence / mutual exclusivity of events (e.g., regions, genes that are always co-expressed)
10. Find the closest/overlapping region given type constraints
11. Find areas that are dense/ sparse of a region type
12. Perform region enrichment (e.g., mutation, gene, transcription factors)
13. …

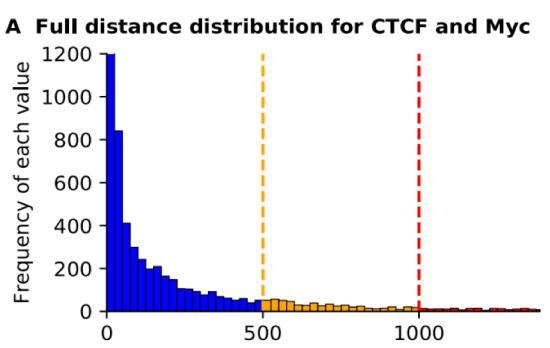
Result visualization



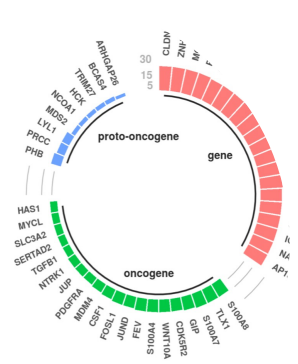


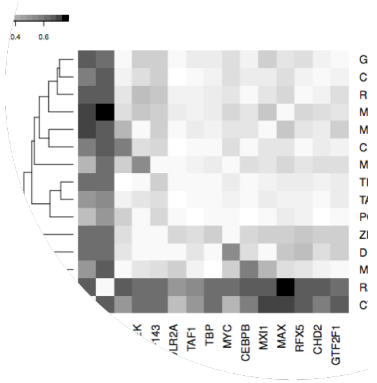


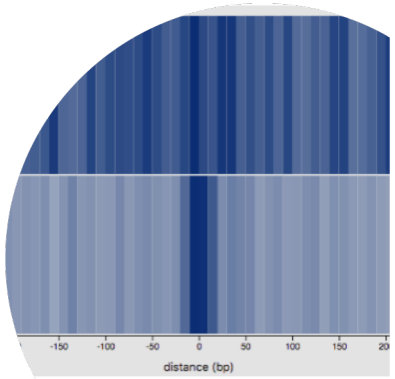




Name tf1	Name tf2	Couples	Couples Tss	Average	Average Passed	Median	Median Passed	Mad	Mad Passed	Tail 1000	Tail 1000 Passed
SIN3A	TAF1	104	0.654	111.644	Passed	44.0	Passed	35.0	Passed	0.019	Failed
SIN3A	TBP	878	0.61	105.196	Passed	47.0	Passed	39.0	Passed	0.003	Passed
TAF1	SIN3A	104	0.654	111.644	Passed	44.0	Passed	35.0	Passed	0.019	Failed
TAF1	TBP	1235	0.593	98.548	Passed	41.0	Passed	34.0	Passed	0.002	Passed
TBP	SIN3A	878	0.61	105.196	Passed	47.0	Passed	39.0	Passed	0.003	Passed
TBP	TAF1	1235	0.593	98.548	Passed	41.0	Passed	34.0	Passed	0.002	Passed

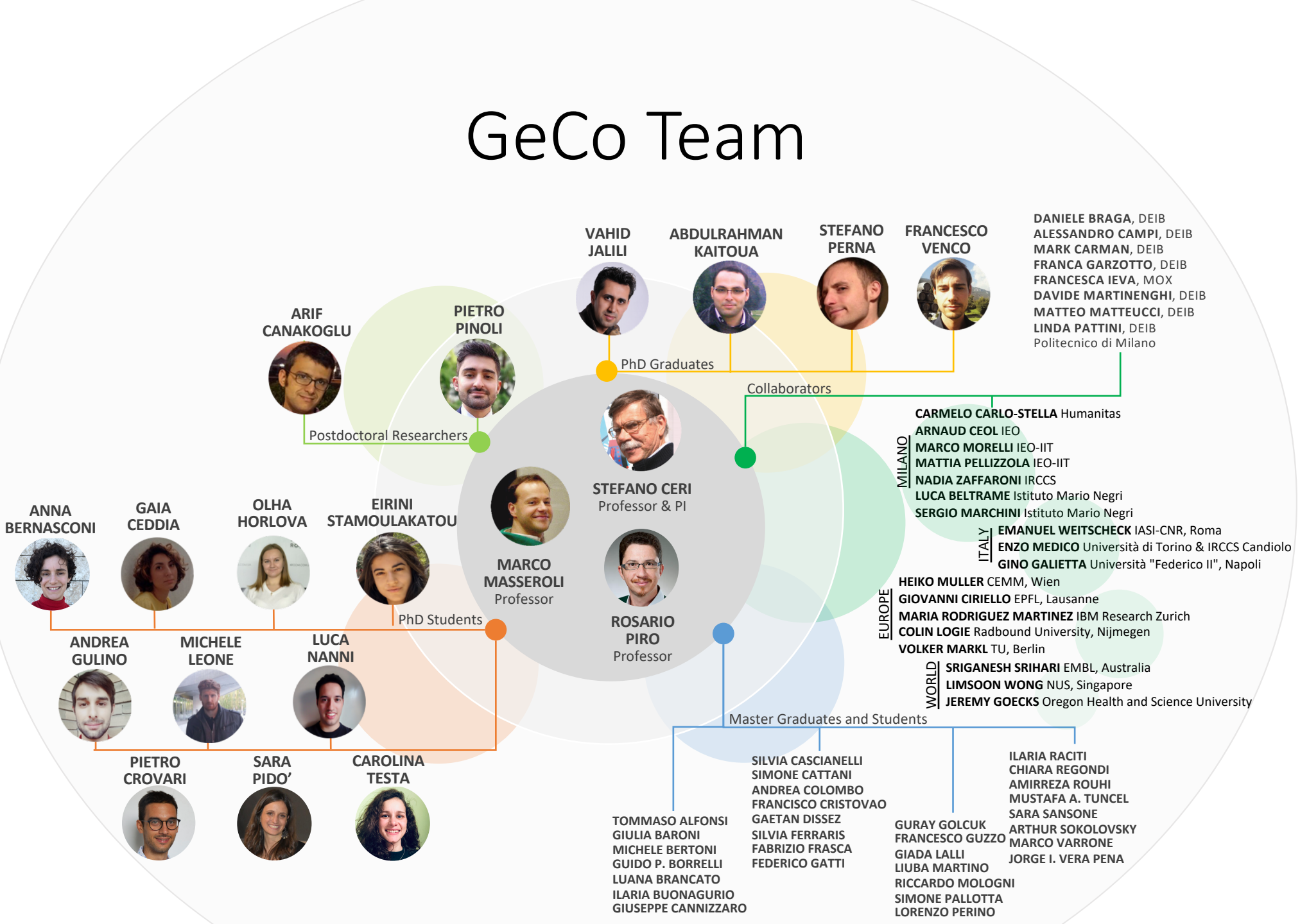






POLITECNICO
MILANO 1863

GeCo Team



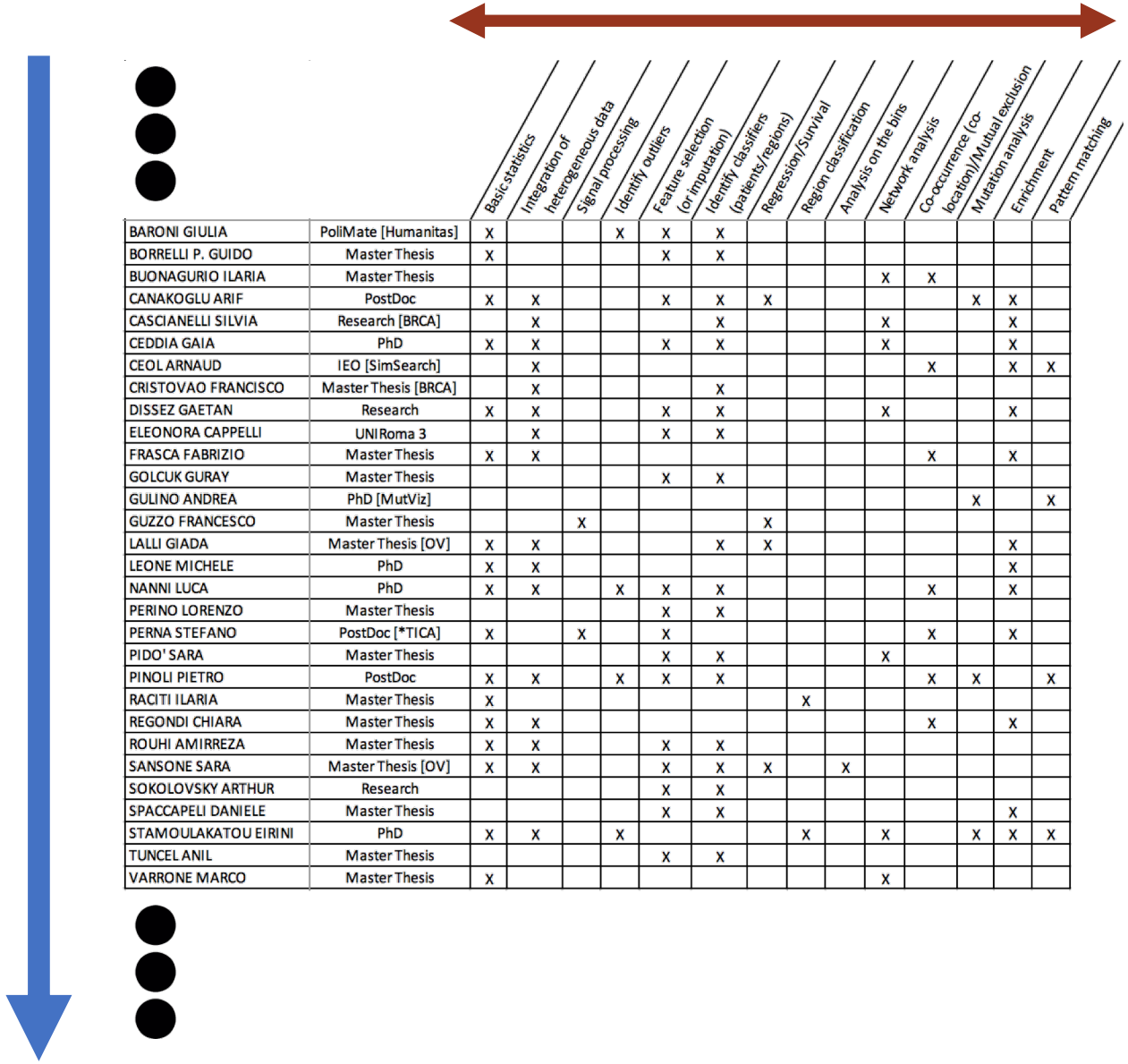
POLITECNICO
MILANO 1863

		Basic statistics	Integration of heterogeneous data	Signal processing	Identify outliers	Feature selection (or imputation)	Identify classifiers (patients/regions)	Regression/Survival	Region classification	Analysis on the bins	Network analysis	Co-occurrence (co-location)/Mutual exclusion	Mutation analysis	Enrichment	Pattern matching
BARONI GIULIA	PoliMate [Humanitas]	X			X	X	X								
BORRELLI P. GUIDO	Master Thesis	X				X	X								
BUONAGURIO ILARIA	Master Thesis										X	X			
CANAKOGLU ARIF	PostDoc	X	X			X	X	X					X	X	
CASCIANELLI SILVIA	Research [BRCA]		X				X				X			X	
CEDDIA GAIA	PhD	X	X			X	X				X			X	
CEOL ARNAUD	IEO [SimSearch]		X									X		X	X
CRISTOVAO FRANCISCO	Master Thesis [BRCA]		X				X								
DISSEZ GAETAN	Research	X	X			X	X				X			X	
ELEONORA CAPPELLI	UNIRoma 3		X			X	X								
FRASCA FABRIZIO	Master Thesis	X	X									X		X	
GOLCUK GURAY	Master Thesis					X	X								
GULINO ANDREA	PhD [MutViz]												X		X
GUZZO FRANCESCO	Master Thesis			X				X							
LALLI GIADA	Master Thesis [OV]	X	X				X	X						X	
LEONE MICHELE	PhD	X	X											X	
NANNI LUCA	PhD	X	X		X	X	X					X		X	
PERINO LORENZO	Master Thesis					X	X								
PERNA STEFANO	PostDoc [*TICA]	X		X		X						X		X	
PIDO' SARA	Master Thesis					X	X				X				
PINOLI PIETRO	PostDoc	X	X		X	X	X					X	X		X
RACITI ILARIA	Master Thesis	X							X						
REGONDI CHIARA	Master Thesis	X	X									X		X	
ROUHI AMIRREZA	Master Thesis	X	X			X	X								
SANSONE SARA	Master Thesis [OV]	X	X			X	X	X		X					
SOKOLOVSKY ARTHUR	Research					X	X								
SPACCAPELI DANIELE	Master Thesis					X	X							X	
STAMOULAKATOU EIRINI	PhD	X	X		X				X		X		X	X	X
TUNCEL ANIL	Master Thesis					X	X								
VARRONE MARCO	Master Thesis	X									X				

Results reached so far by members of the group (Master Thesis students + PhDs + PostDocs)

GeCo 5.0 builds upon diffuse research experience

POLITECNICO
MILANO 1863



		Basic statistics	Integration of heterogeneous data	Signal processing	Identify outliers	Feature selection (or imputation)	Identify classifiers (patients/regions)	Regression/Survival	Region classification	Analysis on the bins	Network analysis	Co-occurrence (co-location)/Mutation exclusion	Mutation analysis	Enrichment	Pattern matching
BARONI GIULIA	PoliMate [Humanitas]	X			X	X	X								
BORRELLI P. GUIDO	Master Thesis	X				X	X								
BUONAGURIO ILARIA	Master Thesis										X	X			
CANAKOGLU ARIF	PostDoc	X	X			X	X	X					X	X	
CASCIANELLI SILVIA	Research [BRCA]		X				X				X			X	
CEDDIA GAIA	PhD	X	X			X	X				X			X	
CEOL ARNAUD	IEO [SimSearch]		X									X		X	X
CRISTOVAO FRANCISCO	Master Thesis [BRCA]		X				X								
DISSEZ GAETAN	Research	X	X			X	X				X			X	
ELEONORA CAPPELLI	UNIRoma 3		X			X	X								
FRASCA FABRIZIO	Master Thesis	X	X									X		X	
GOLCUK GURAY	Master Thesis					X	X								
GULINO ANDREA	PhD [MutViz]												X		X
GUZZO FRANCESCO	Master Thesis			X				X							
LALLI GIADA	Master Thesis [OV]	X	X				X	X						X	
LEONE MICHELE	PhD	X	X											X	
NANNI LUCA	PhD	X	X		X	X	X					X		X	
PERINO LORENZO	Master Thesis					X	X								
PERNA STEFANO	PostDoc [*TICA]	X		X		X						X		X	
PIDO' SARA	Master Thesis					X	X				X				
PINOLI PIETRO	PostDoc	X	X		X	X	X					X	X		X
RACITI ILARIA	Master Thesis	X							X						
REGONDI CHIARA	Master Thesis	X	X									X		X	
ROUHI AMIRREZA	Master Thesis	X	X			X	X								
SANSONE SARA	Master Thesis [OV]	X	X			X	X	X		X					
SOKOLOVSKY ARTHUR	Research					X	X								
SPACCAPELI DANIELE	Master Thesis					X	X							X	
STAMOULAKATOU EIRINI	PhD	X	X		X				X		X		X	X	X
TUNCEL ANIL	Master Thesis					X	X								
VARRONE MARCO	Master Thesis	X									X				

Many other users
Many case studies

Relatively few
new patterns!

Driving Principles for Building GeCo 5.0

- **Enable Effective Composition of Patterns**
  - Bottom up, supported by effective infrastructure and HCI
- **Support Extensibility**
  - High-level, orthogonal, complete, correct
- **Support Reproducibility**
  - Not only dialogue, but also summarization and formalization

Thank you for your attention

Questions?

POLITECNICO
MILANO 1863