GOLAM: A framework for Analyzing Genomic Data

Lorenzo Baldacci, Matteo Golfarelli,
Simone Graziani, and Stefano Rizzi

DISI – University of Bologna, Italy

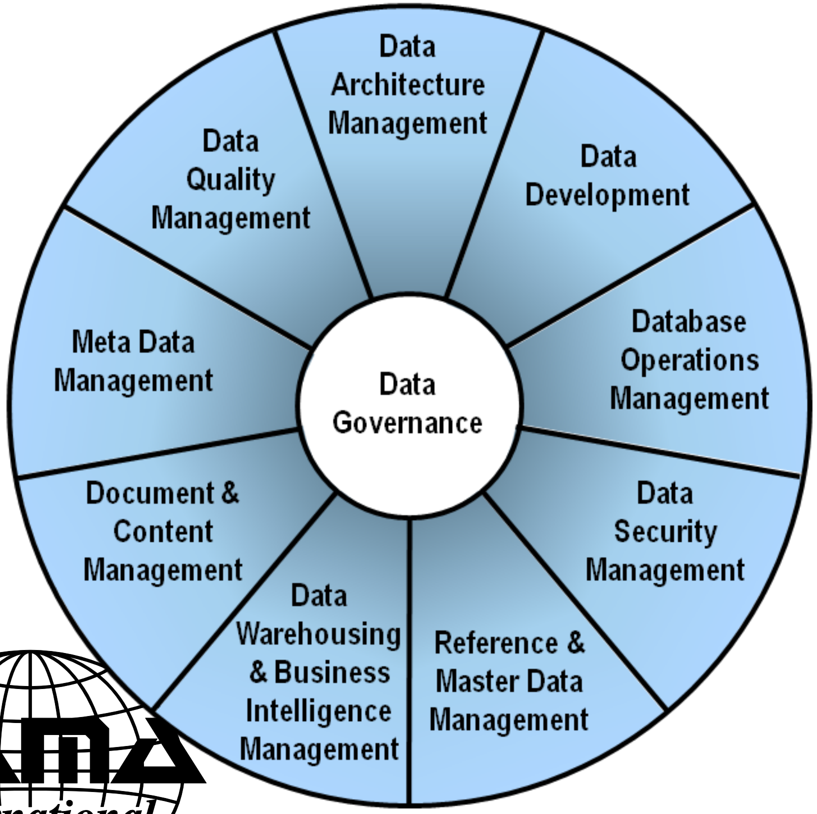
Business
Intelligence
Group

Contents

- Genomic Data
- Multidimensional Analytics applied to Genomic Data
- Discussion and Current Research Directions

Dealing with Genomic Data

- Genomic challenges: from sequencing to data management
  - ✓ Bioinformatics challenges are moving towards storage, retrieval, security, and presentation of genomic information



Data Governance

Data Architecture Management

Data Development

Database Operations Management

Data Security Management

Reference & Master Data Management

Data Warehousing & Business Intelligence Management

Document & Content Management

Meta Data Management

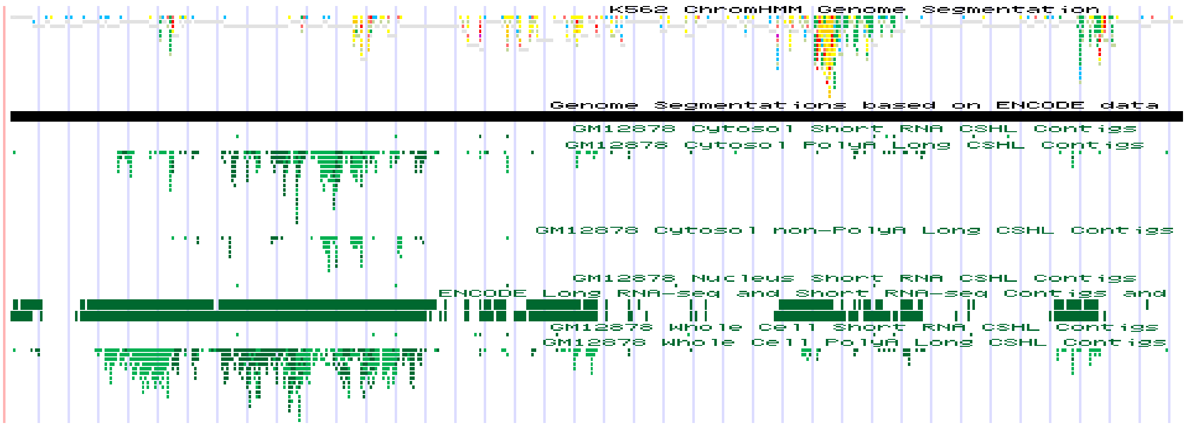
Data Quality Management

Sequencing is not enough!

DAMA International

Genomic Analytics Challenges

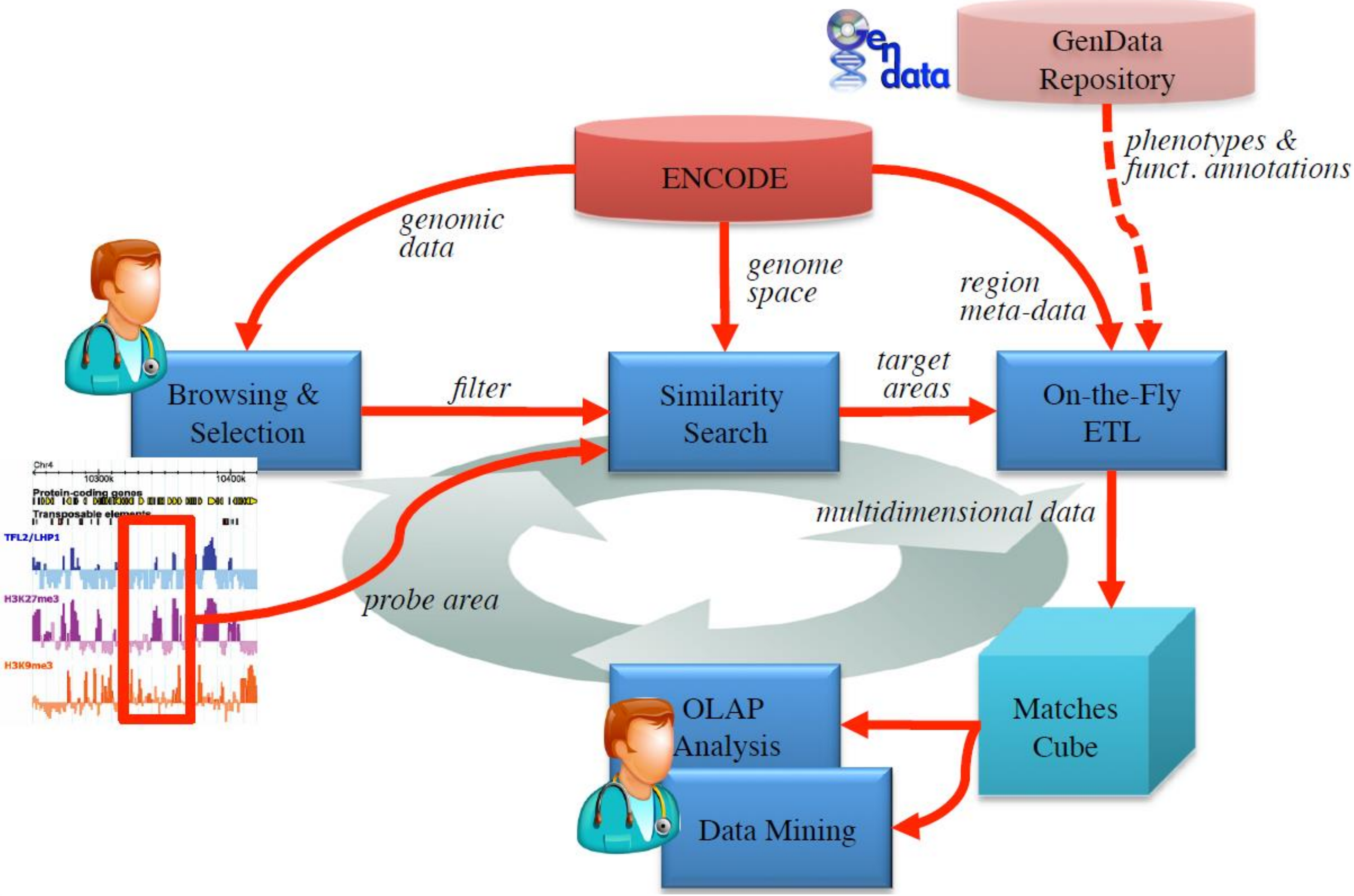
- Ad hoc (vertical) applications supporting specific biological questions
- Laboratories and Biologists frequently undertake analytical tools development in-house
- Genomic browsers are effective when analyzing detailed data but fall short when analyzing aggregated data


K562 ChromHMM Genome Segmentation
Genome Segmentations based on ENCODE data
GM12878 Cytosol Short RNA CSHL Contigs
GM12878 Cytosol PolyA Long CSHL Contigs
GM12878 Cytosol non-PolyA Long CSHL Contigs
GM12878 Nucleus Short RNA CSHL Contigs
ENCODE Long RNA-seq and Short RNA-seq Contigs and
GM12878 Whole Cell Short RNA CSHL Contigs
GM12878 Whole Cell PolyA Long CSHL Contigs

GOLAM: Genomic OLAM

- is a framework for OLAP analysis and mining
- helps biologists in overcoming the rigidity of genome analysis methods
- automates and speeds up analysis sessions and introduces a multi-resolution view of the data
- addresses the issue of loading data to the cube “On-the-Fly”

The GOLAM Framework



GenData Repository

ENCODE

phenotypes & funct. annotations

genomic data

genome space

region meta-data

Browsing & Selection

filter

Similarity Search

target areas

On-the-Fly ETL

multidimensional data

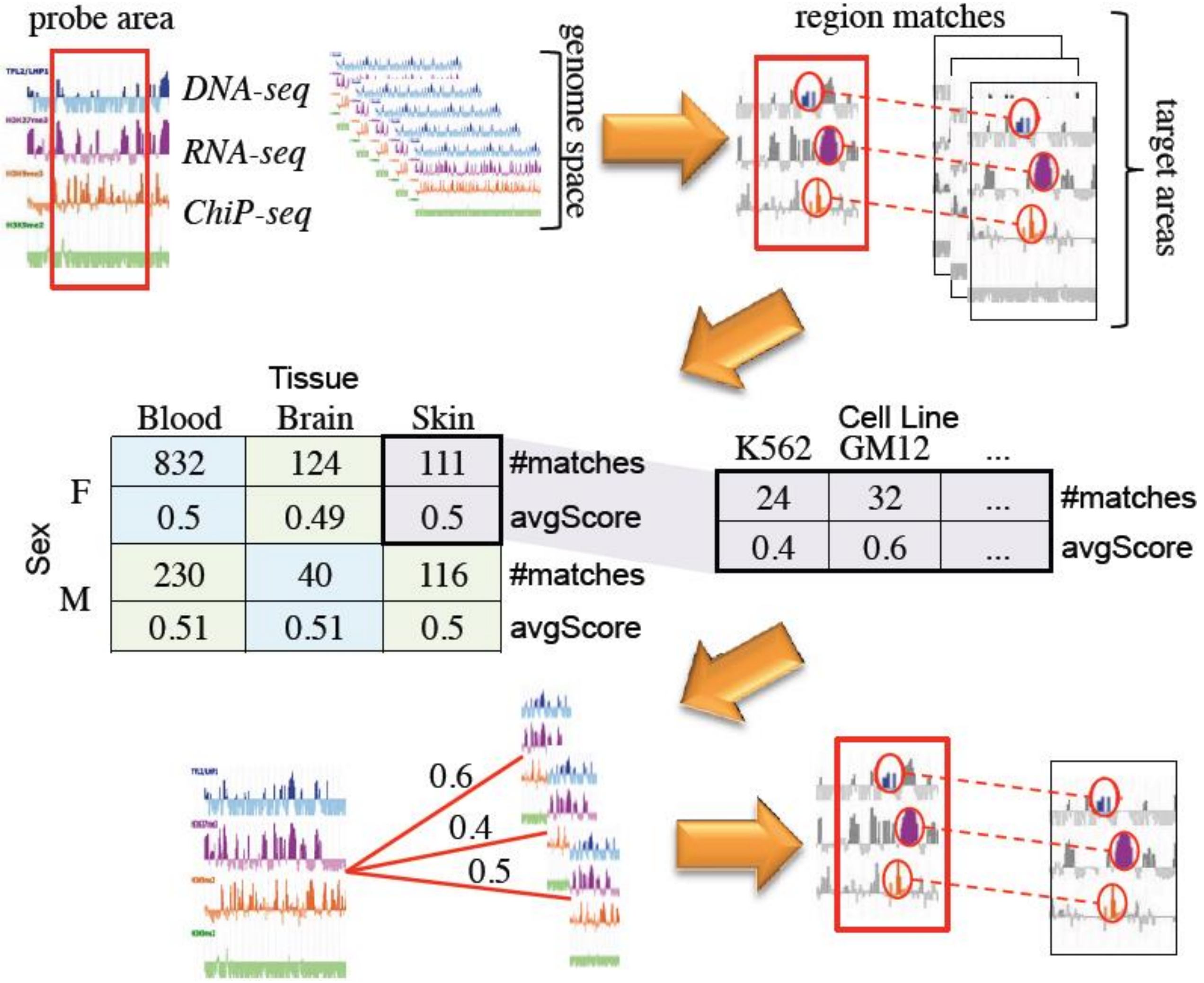
probe area

Matches Cube

OLAP Analysis

Data Mining

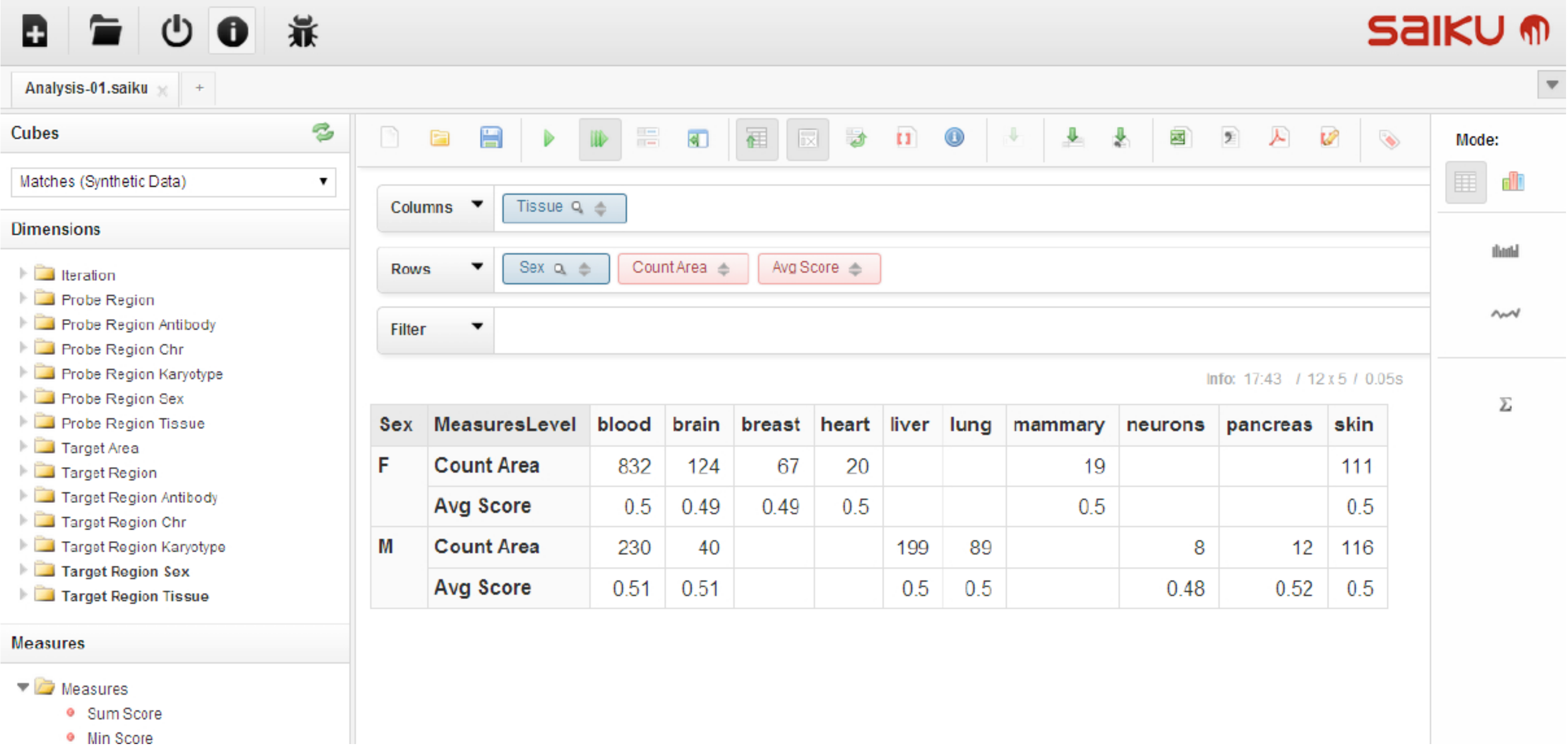
Genomic OLAP



		Tissue			
		Blood	Brain	Skin	
Sex	F	832	124	111	#matches
		0.5	0.49	0.5	avgScore
	M	230	40	116	#matches
		0.51	0.51	0.5	avgScore

Cell Line			
K562	GM12	...	
24	32	...	#matches
0.4	0.6	...	avgScore

Genomic OLAP



Analysis-01.saiku

Cubes

Matches (Synthetic Data)

Dimensions

- Iteration
- Probe Region
- Probe Region Antibody
- Probe Region Chr
- Probe Region Karyotype
- Probe Region Sex
- Probe Region Tissue
- Target Area
- Target Region
- Target Region Antibody
- Target Region Chr
- Target Region Karyotype
- Target Region Sex
- Target Region Tissue

Measures

- Measures
  - Sum Score
  - Min Score

Columns: Tissue

Rows: Sex, Count Area, Avg Score

Filter

Mode:

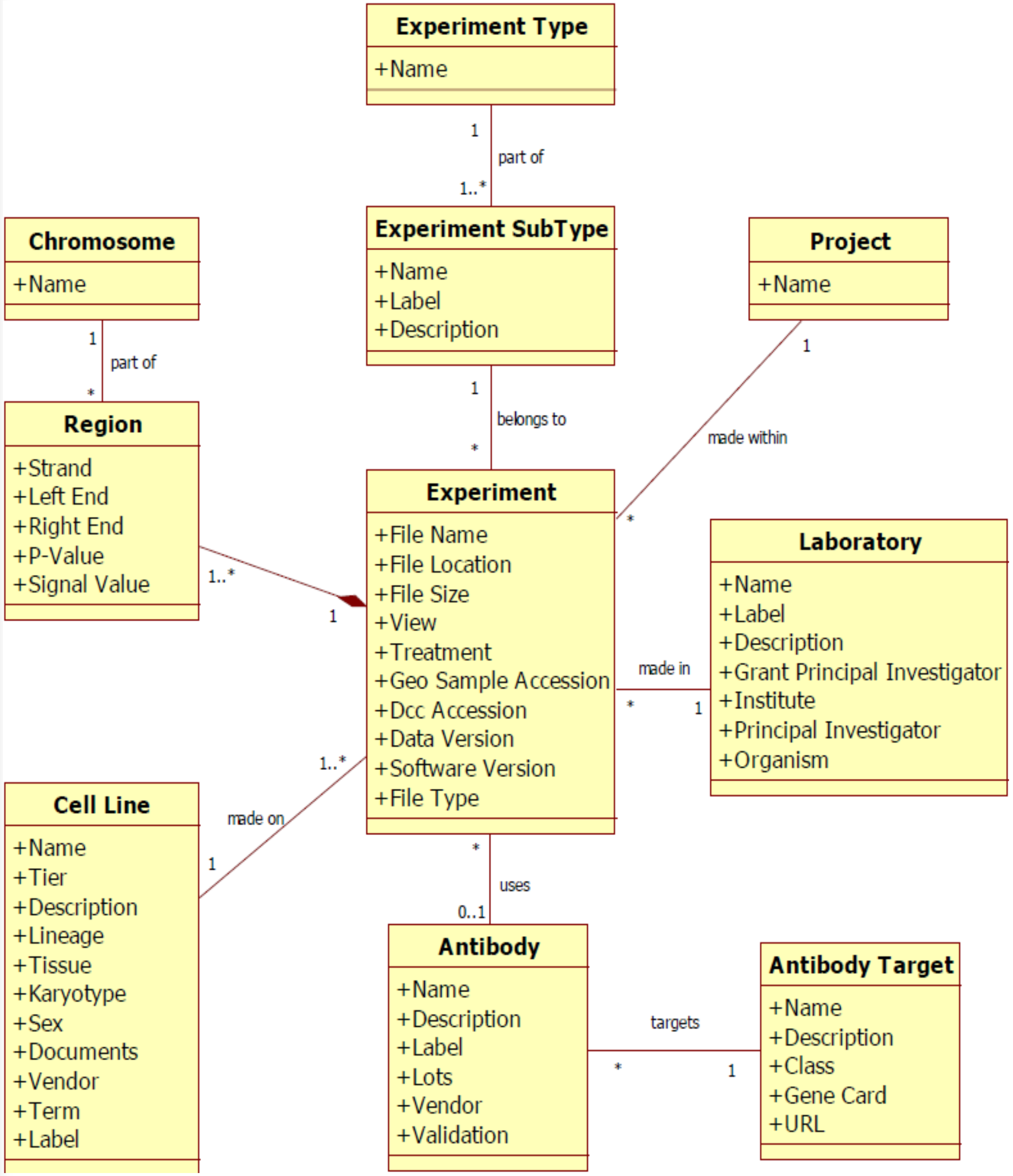
Info: 17:43 / 12 x 5 / 0.05s

Sex	MeasuresLevel	blood	brain	breast	heart	liver	lung	mammary	neurons	pancreas	skin
F	Count Area	832	124	67	20			19			111
	Avg Score	0.5	0.49	0.49	0.5			0.5			0.5
M	Count Area	230	40			199	89		8	12	116
	Avg Score	0.51	0.51			0.5	0.5		0.48	0.52	0.5

Research Challenges

- Genomic data modeling
- Schemaless data structure to deal with
  - Metastars
- Non-traditional data sources
  - Files
  - Ontologies
  - Big Data

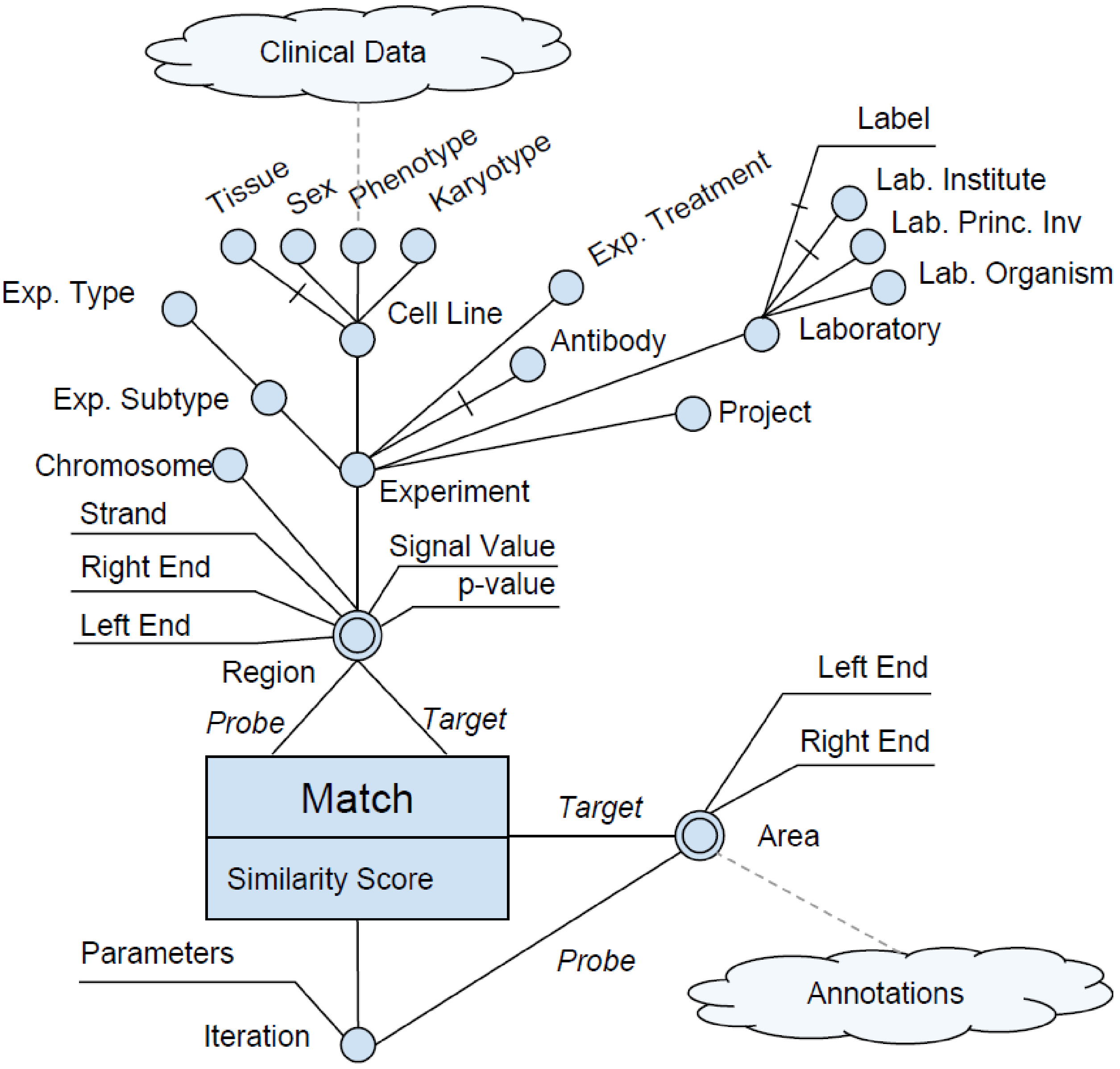
ENCODE data model



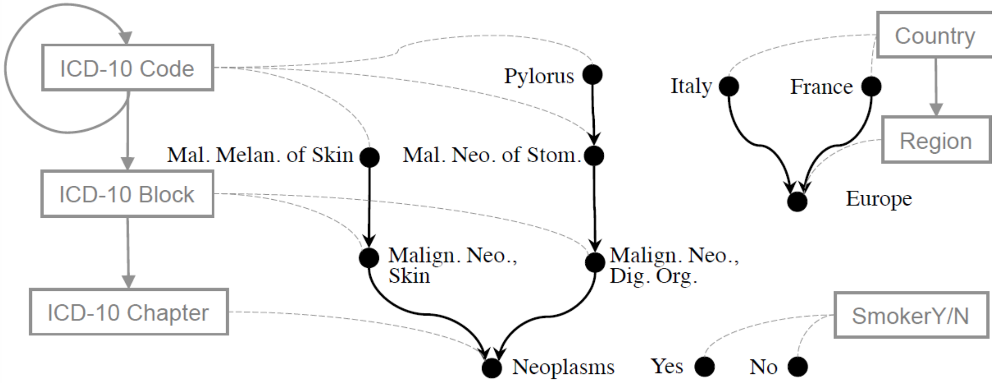
ENCODE is the *Encyclopedia of DNA Elements* publicly available and recognized to be the standard repository for genomic data and functional elements.

- **Cell Line**: is the biological sample on which an experiment is carried out.
- **Experiment**: is a single analysis on a cell in a laboratory.
- **Region:** is a segment of an experiment. It holds biological functional information.

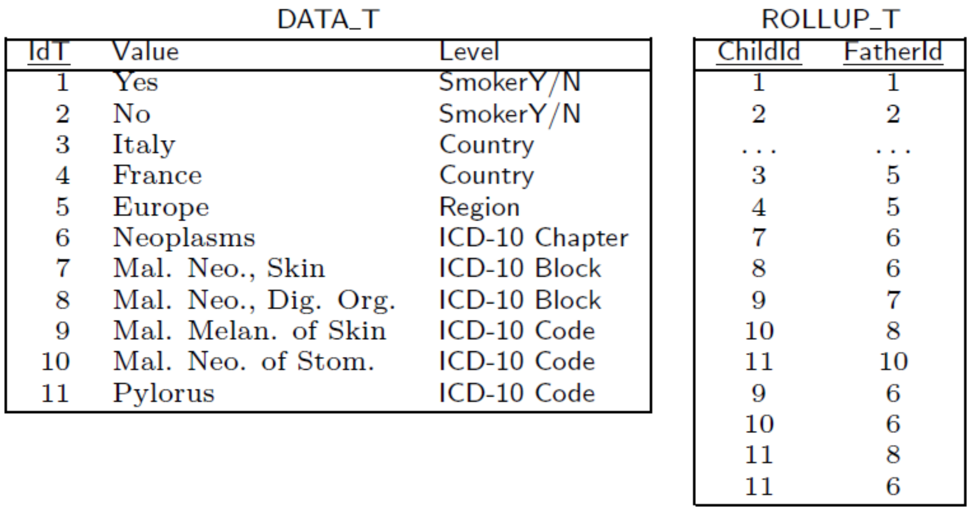
Multidimensional Schema



Metastars



Metastars use meta-modelling coupled with traditional dimension tables to support non-onto, non-covering, and non-strict hierarchies.



DATA_T

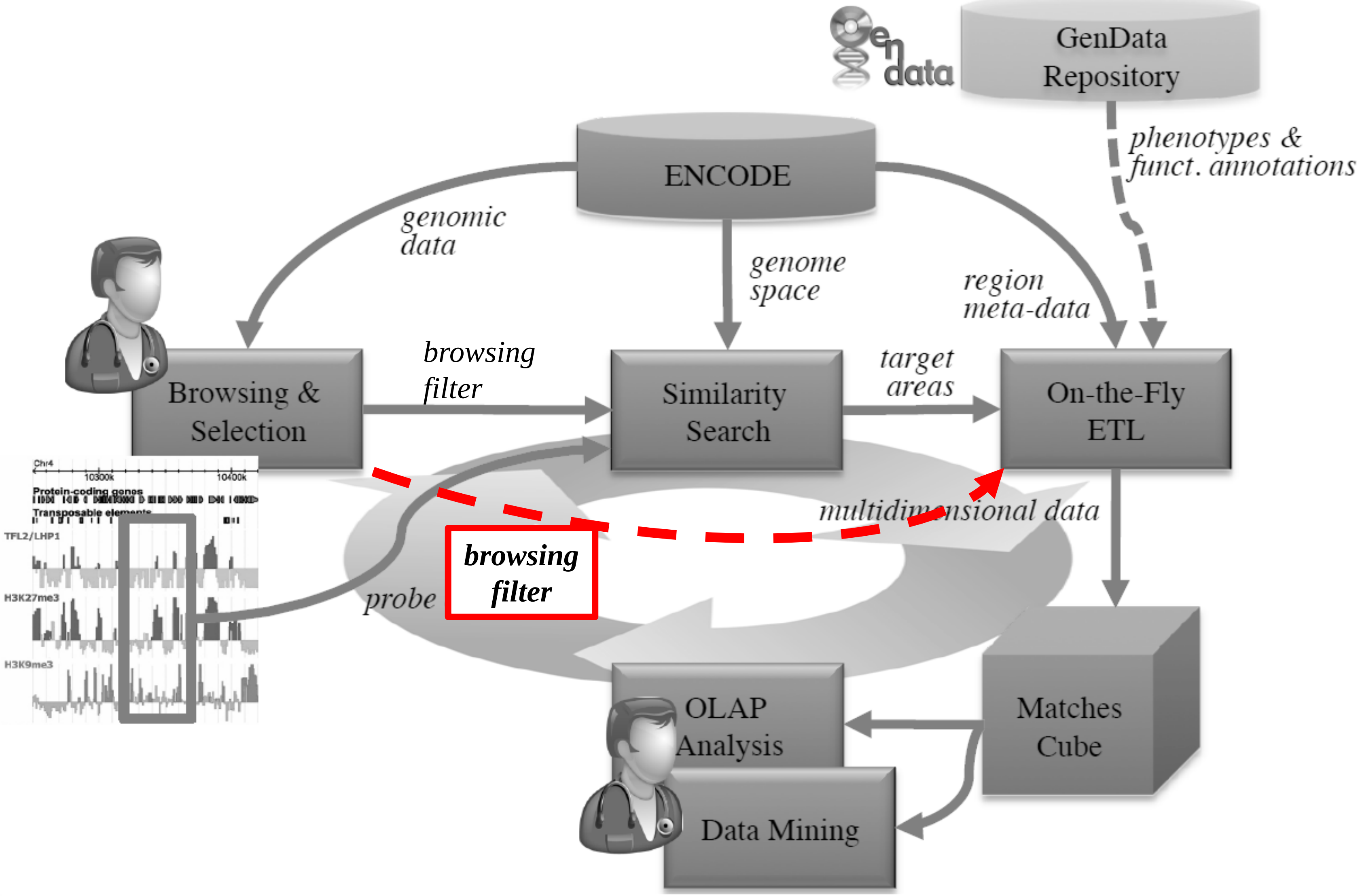
IdT	Value	Level
1	Yes	SmokerY/N
2	No	SmokerY/N
3	Italy	Country
4	France	Country
5	Europe	Region
6	Neoplasms	ICD-10 Chapter
7	Mal. Neo., Skin	ICD-10 Block
8	Mal. Neo., Dig. Org.	ICD-10 Block
9	Mal. Melan. of Skin	ICD-10 Code
10	Mal. Neo. of Stom.	ICD-10 Code
11	Pylorus	ICD-10 Code

ROLLUP_T

ChildId	FatherId
1	1
2	2
...	...
3	5
4	5
7	6
8	6
9	7
10	8
11	10
9	6
10	6
11	8
11	6

By employing Metastars we are able to **model hierarchies** whenever the schema is **dynamic or missing**.

On-The-Fly ETL



GenData Repository

phenotypes & funct. annotations

ENCODE

genomic data

genome space

region meta-data

Browsing & Selection

browsing filter

Similarity Search

target areas

On-the-Fly ETL

multidimensional data

browsing filter

probe

Matches Cube

OLAP Analysis

Data Mining

ETL Tests

- ENCODE counts a number of experiment files in the range of 25K, leading to **over a billion instances in the region dimension**
- The number of matches might count **over 300K events for each analysis session**
- Our preliminary tests aim at evaluating the On-the-Fly ETL from the efficiency point of view
  - ✓ The “eager” approach is compared vs. the approach that loads the matching genome space only
  - ✓ Tractability threshold is set afterwards

ETL Tests

- A probe area composed of 50 regions and 3 different genome spaces
  - ✓ Test 1 (*T1*) consisting of 740K regions
  - ✓ Test 2 (*T2*) consisting of 4.42M regions
  - ✓ Test 3 (*T3*) consisting of 54.5M regions

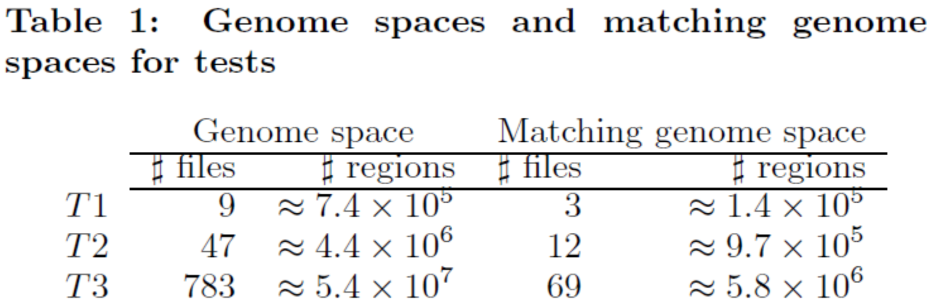


Table 1: Genome spaces and matching genome spaces for tests

	Genome space		Matching genome space	
	♯ files	♯ regions	♯ files	♯ regions
$T1$	9	$\approx 7.4 \times 10^5$	3	$\approx 1.4 \times 10^5$
$T2$	47	$\approx 4.4 \times 10^6$	12	$\approx 9.7 \times 10^5$
$T3$	783	$\approx 5.4 \times 10^7$	69	$\approx 5.8 \times 10^6$

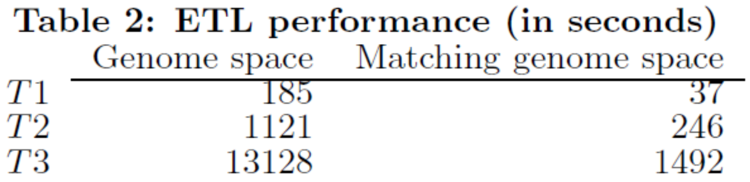


Table 2: ETL performance (in seconds)

	Genome space	Matching genome space
$T1$	185	37
$T2$	1121	246
$T3$	13128	1492

- The tractability threshold can be reasonably set to 50 ENCODE files

Discussion

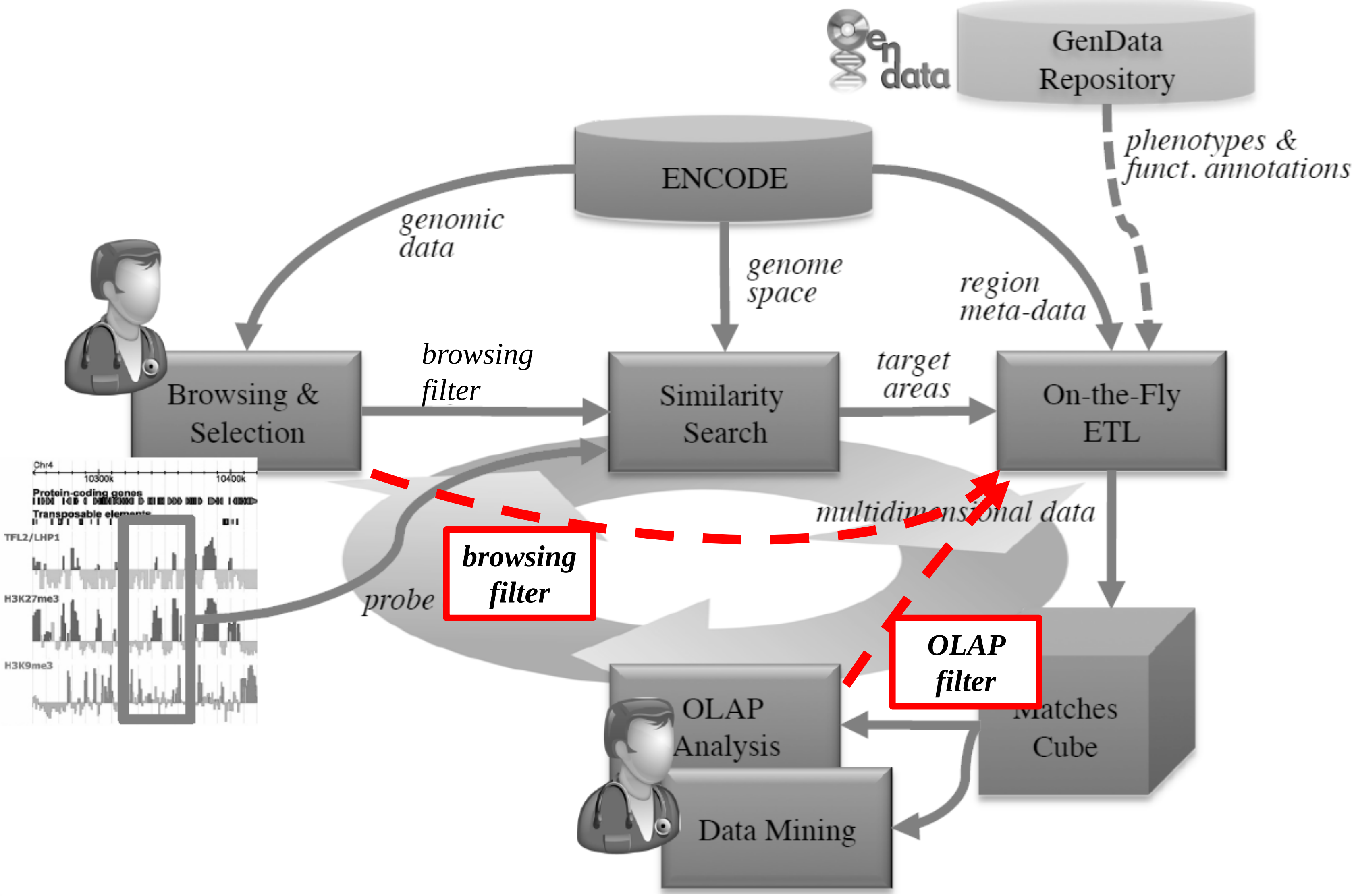
- With GOLAM we took a first step towards overcoming current limitations of genome analysis methods
  - ✓ Analysis session has been automated and speeded up
  - ✓ More analysis flexibility has been introduced

- We proved that ETL processes can be integrated within the analysis session in order to improve efficiency in those DW applications employed in non-traditional domains
  - ✓ big data
  - ✓ scientific data storage
  - ✓ open linked data
  - ✓ etc.

(Not so) Future work

- On-the-Fly ETL can be further pushed so to be framed within the OLAP session itself. In this manner interesting data (according to the user) might be loaded into the cube
  - ✓ Multidimensional indexing and new dice operators must be employed by the ETL processes
  - ✓ Source data extraction must be done according to a cost function that considers many facets of optimization (e.g. time, cost, etc.)
- Approximate results could be provided in order to improve the overall session’s responsiveness
  - ✓ ETL should be integrated and designed so that it gather data as a stream and exposes partial results to the user

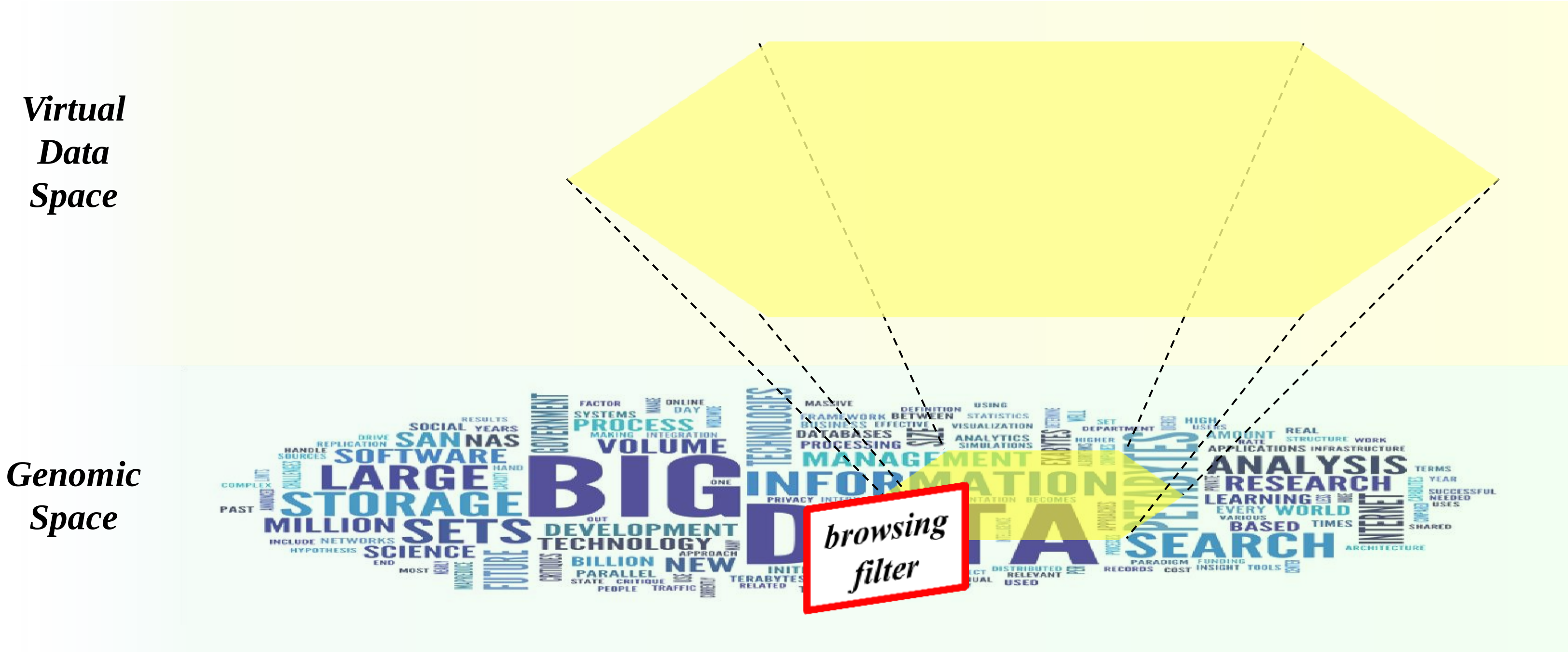
(Not so) Future work



(Not so) Future work

Genomic Space

(Not so) Future work



Virtual Data Space

Genomic Space

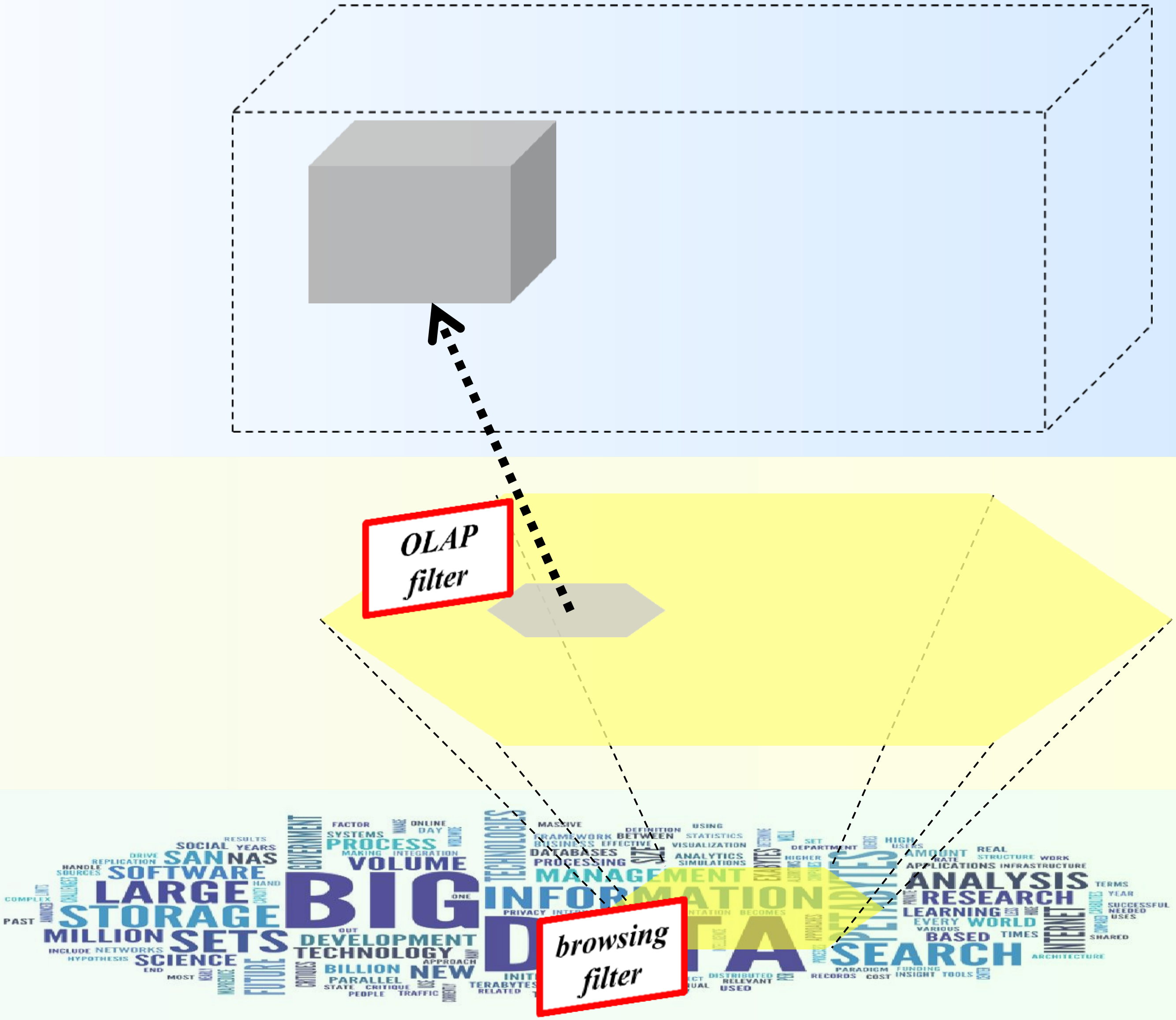
browsing filter

(Not so) Future work

Multidimensional Space

Virtual Data Space

Genomic Space

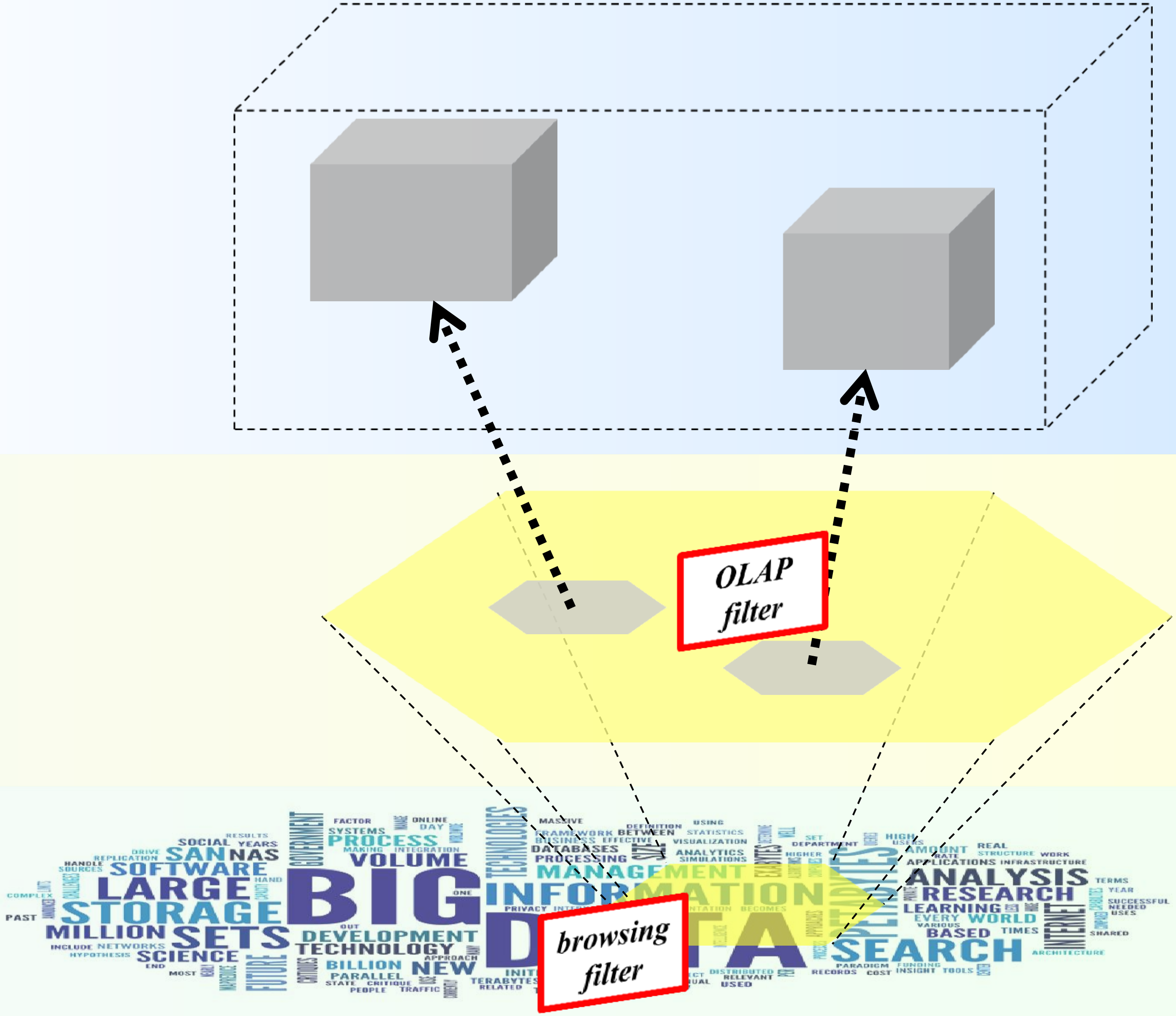


(Not so) Future work

Multidimensional Space

Virtual Data Space

Genomic Space

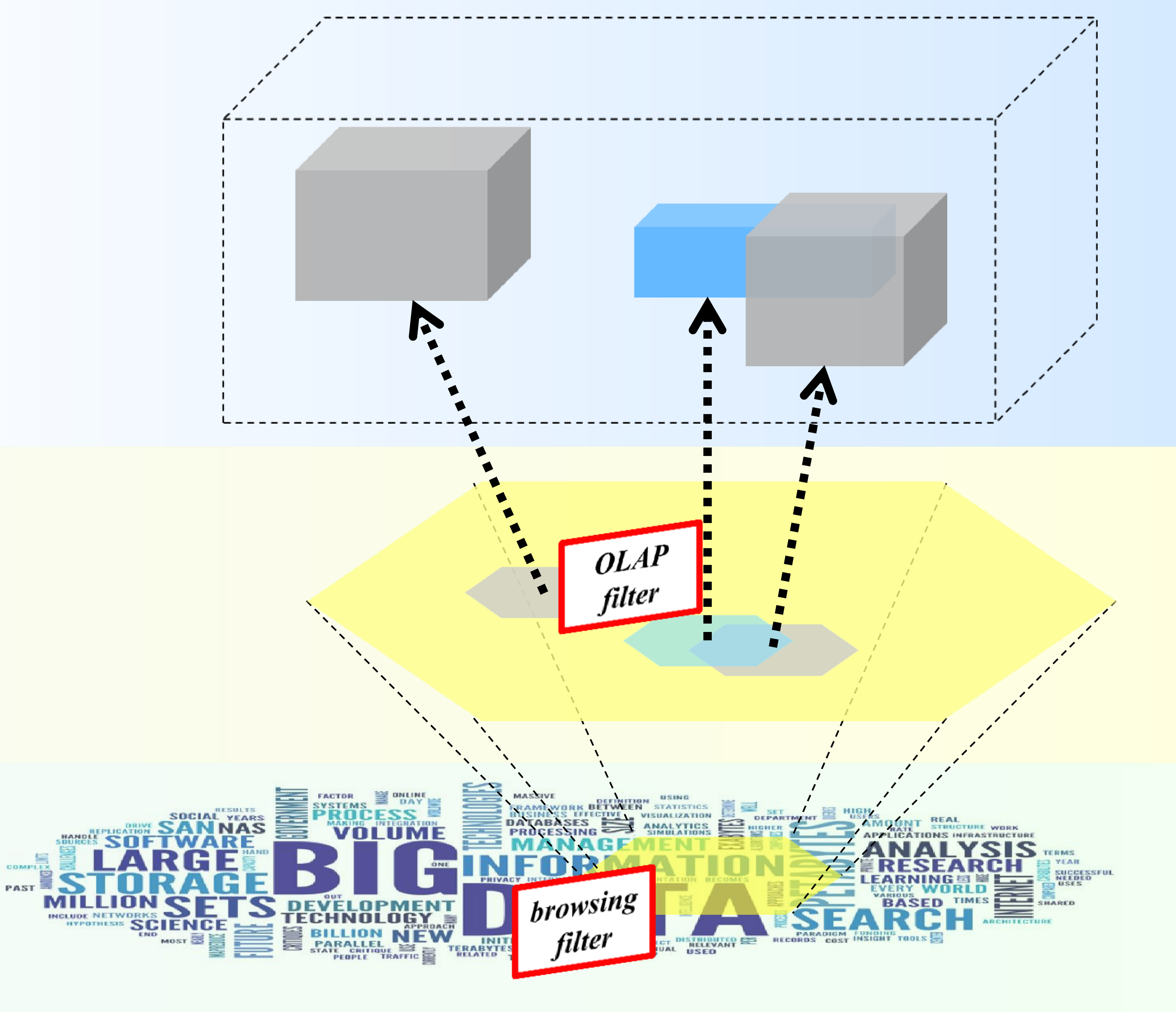


(Not so) Future work

Multidimensional Space

Virtual Data Space

Genomic Space



OLAP filter

browsing filter

…questions time…