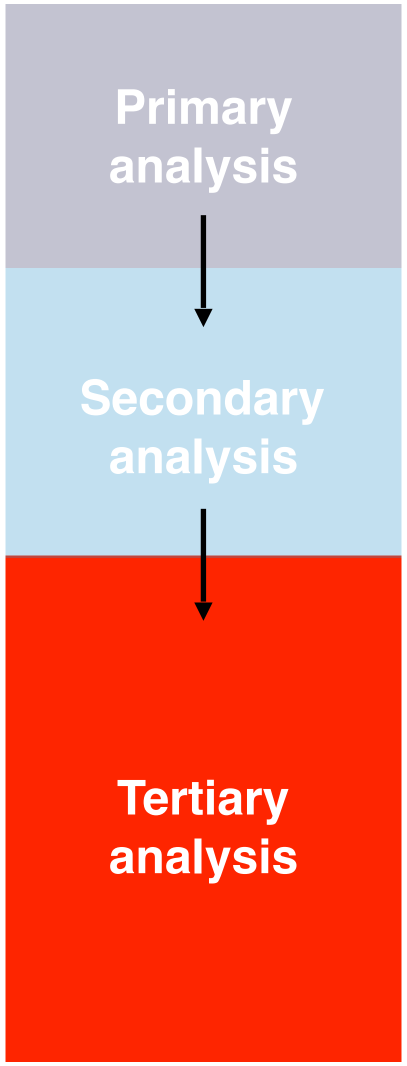
Integration of Genomic Big Data: Efficient Queries on ENCODE (Meta)data

Stefano Perna

DEIB | Dipartimento di Elettronica, Informazione e Bioingegneria

POLITECNICO
MILANO 1863

Big Data Analysis with NGS



- Hardware generated data, machine stats, etc.
- Production of sequence reads and quality scores

- QA filtering on raw reads
- Alignment/Assembly of reads
- QA and variant calling on aligned reads

- Multi-sample processing
- QA/QC of variant calls
- Annotation and filtering of variants
- Data aggregation
- Association analysis
- Population structure analysis
- Genome browser driven exploratory analysis

Source: http://blog.goldenhelix.com/grudy/a-hitchhiker%E2%80%99s-guide-to-next-generation-sequencing-part-2/

Main Questions

(from our interaction with **IEO - European Oncology Institute** and **IIT - Italian Institute of Technology**)

*“Can **interesting DNA regions and their relationships** be **discovered** using **genome-wide** queries?”*

*“Can **genomic data of patients** be **grouped** according to clinical phenotype and **compared**?”*

*“Can the **genomic features of all the genes** involved in the **same biological process** be **extracted and then analyzed**?”*

*“Can we r**etrieve portions of the genome of given patients**, extracting them fr**om remote servers** and comparing them?”*

Main Questions

(from our interaction with **IEO - European Oncology Institute** and **IIT - Italian Institute of Technology**)

*“Can **interesting DNA regions and their relationships** be **discovered** using **genome-wide** queries?”*

→ **GENOMETRIC QUERY LANGUAGE (GMQL)**

*“Can **genomic data of patients** be **grouped** according to clinical phenotype and **compared**?”*

→ GMQL
+
CLUSTERING

*“Can the **genomic features of all the genes** involved in the **same biological process** be **extracted and then analyzed**?”*

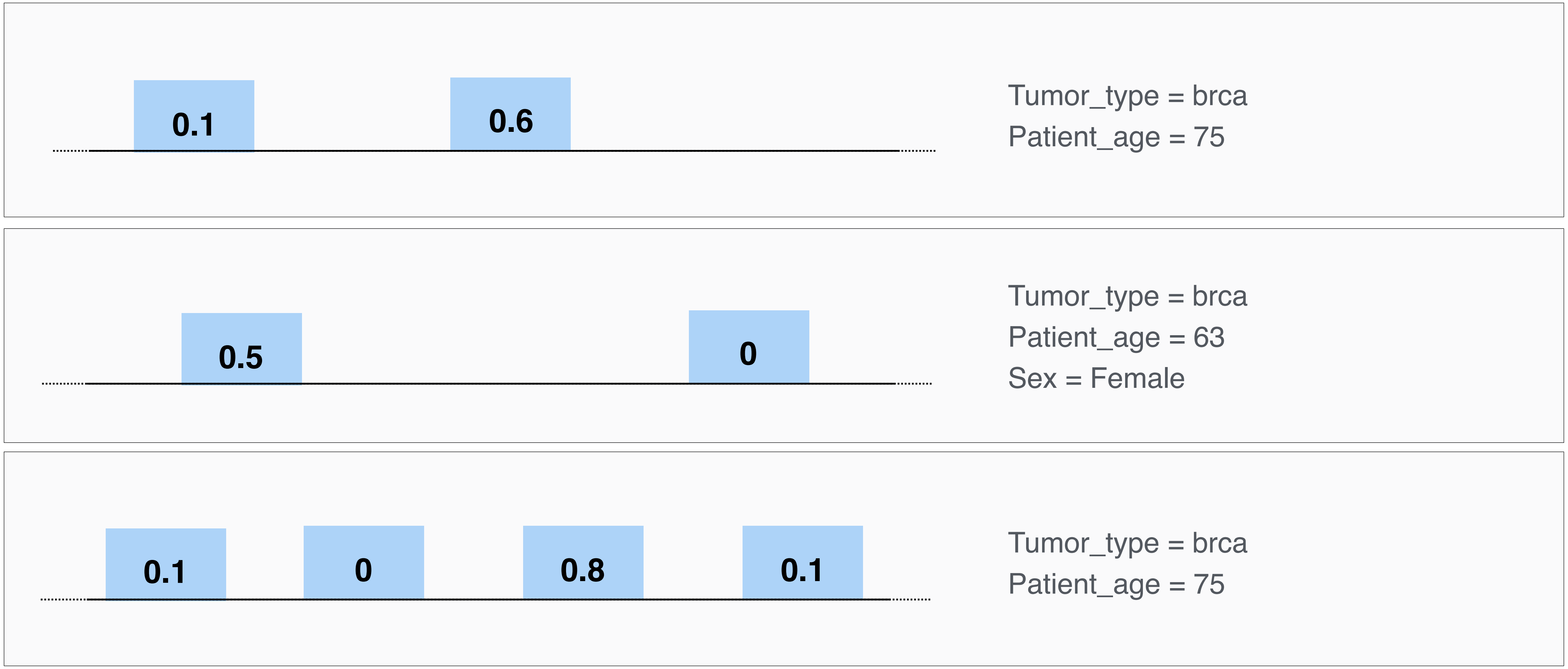
→ GMQL
+
DATA ANALYSIS

*“Can we r**etrieve portions of the genome of given patients**, extracting them fr**om remote servers** and comparing them?”*

→ GMQL
+
INDEXING & SEARCH

Genomic Data Model

REGIONS **METADATA**



Tumor_type = brca
Patient_age = 75

Tumor_type = brca
Patient_age = 63
Sex = Female

Tumor_type = brca
Patient_age = 75

QUERY LANGUAGE

SEQUENCE OF
ALGEBRAIC OPERATIONS

```
PROMS = SELECT(annotationType == 'promoter') ANNOTATIONS;
PEAKS = SELECT(dataType == 'ChipSeq') ENCODE;
RESULT = MAP(peak_count AS COUNT) PROMS PEAKS;
```

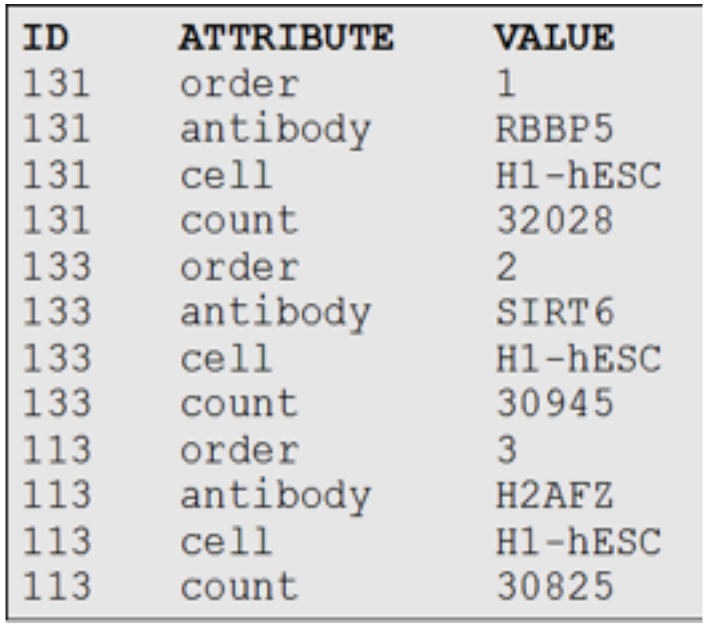
Masseroli et al. Bioinformatics 31(12):1881-8 (2015)

QUERY LANGUAGE

SEQUENCE OF ALGEBRAIC OPERATIONS

```
PROMS = SELECT(annotationType == 'promoter') ANNOTATIONS;
PEAKS = SELECT(dataType == 'ChipSeq') ENCODE;
RESULT = MAP(peak_count AS COUNT) PROMS PEAKS;
```

Executed over 2,423 ENCODE samples including a total of 83,899,526 peaks mapped to 131,780 promoters producing as result 29 GB of data



ID	ATTRIBUTE	VALUE
131	order	1
131	antibody	RBBP5
131	cell	H1-hESC
131	count	32028
133	order	2
133	antibody	SIRT6
133	cell	H1-hESC
133	count	30945
113	order	3
113	antibody	H2AFZ
113	cell	H1-hESC
113	count	30825

# Samples	# Regions	Join(dist <0)	Map(COUNT)	Cover
10	~1.9 M	14.66 sec.	20.29 sec.	19.25 sec.
50	~8.8 M	23.86 sec.	43.08 sec	46.34 sec.
100	~17.4 M	35.38 sec	74.43 sec.	79.02 sec.
1000	~60 M	120.98 sec	473.39 sec	235.22 sec.

Masseroli et al. Bioinformatics 31:12 (2015)

REPOSITORY

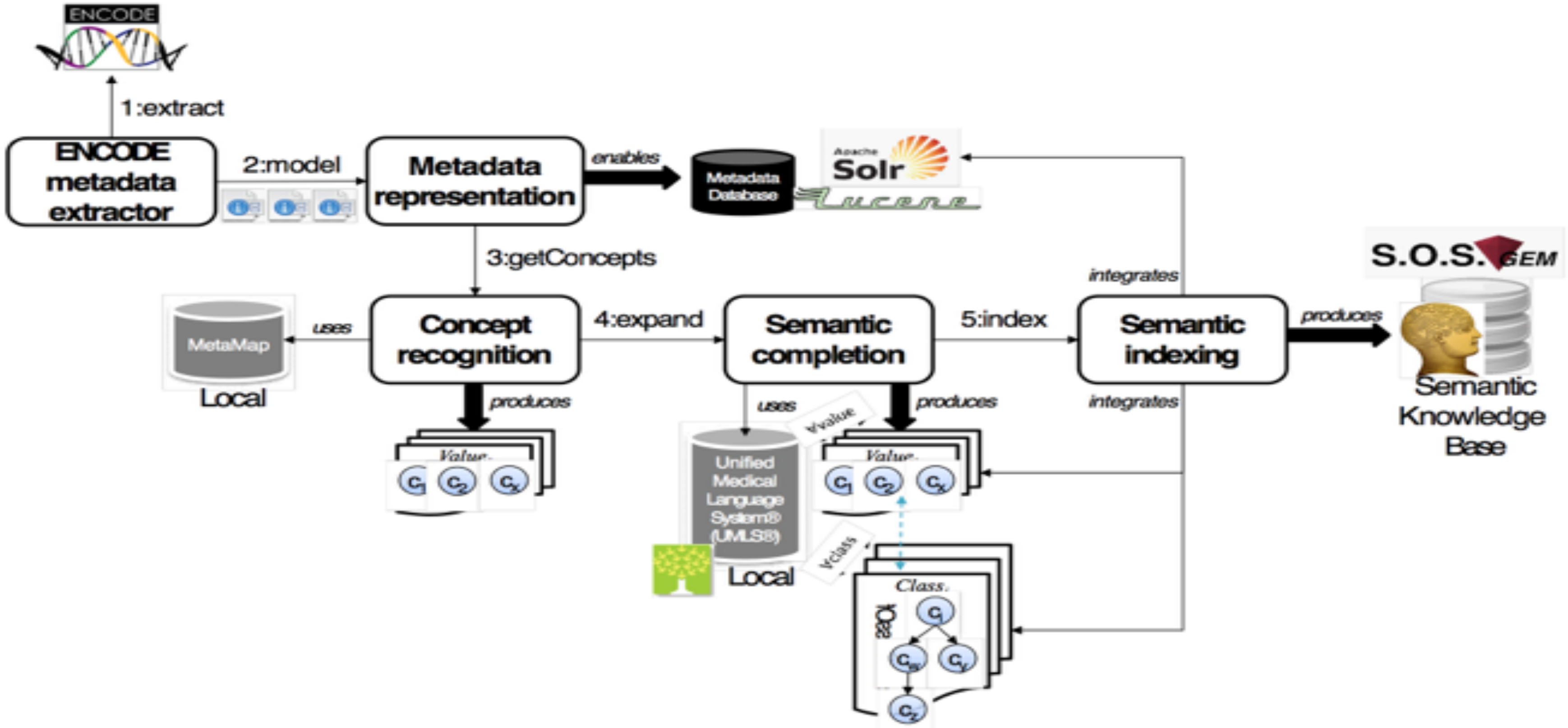
Consortium	Imported datasets	# of samples	File size (MB)
ENCODE	HG19_ENCODE_BED	1,933	34,201
	HG19_ENCODE_BROAD	1,970	23,552
	HG19_ENCODE_NARROW	1,999	7,168
	MM9_ENCODE_BROAD	441	2,355
	MM9_ENCODE_NARROW	277	1,162
EPIGENOMICS ROADMAP	HG19_EPIGENOMICS_ROADMAP_BED	78	595
	HG19_EPIGENOMICS_ROADMAP_BROAD	979	23,244
TCGA	HG19_TCGA_Cnv	2,623	117
	HG19_TCGA_DnaSeq	6,361	276
	HG19_TCGA_Dnamethylation	1,384	29,696
	HG19_TCGA_Mirna_Isoform	9,227	3,379
	HG19_TCGA_Mirna_Mirnaseq	9,227	569
	HG19_TCGA_RnaSeq_Exon	2,544	31,744
	HG19_TCGA_RnaSeq_Gene	2,544	3,584
	HG19_TCGA_RnaSeq_Spljxn	2,544	30,720
	HG19_TCGA_RnaSeqV2_Exon	9,217	114,688
	HG19_TCGA_RnaSeqV2_Gene	9,217	20,480
	HG19_TCGA_RnaSeqV2_Spljxn	9,217	105,472
	HG19_TCGA_RnaSeqV2_Isoform	9,217	49,152
Grand total	**19 datasets**	**81,012**	**412,835**

Semantic Understanding of ENCODE Metadata

- S.o.S.Gem searches for a**pproximate matching with Encode metadata** by **using the Unified Medical Language System** (UMLS, 173 vocabularies, 3M concepts, 12M atoms)

- More in detail, it builds the c**ompletion of the ontology w.r.t ENCODE metadata**, using **forward chaining**

- Leverages **MetaMap**, a tool for recognising UMLS atoms.



Semantic Understanding of ENCODE Metadata

- SoSGem is used in **pipeline with GMQL**, performing **information retrieval on Encode**
- Joint work with La Sapienza of Rome (Fernandez, Lenzerini),
- Published on **IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS**

Fernandez et al. IEEE/ACM Trans Comput Biol Bioinform 13(2):233-47 (2016)

OUR VISION

What comes next?

PRIMARY ANALYSIS

read counting
MOTIF finding
quality control MEME SNP BOWTIE
alignment peak calling GATK
ADAM variant calling
FASTA indel detection
SICER allele calling BWA
MACS HMMER

SECONDARY ANALYSIS

GMQL

GENOMETRIC QUERY LANGUAGE

TERTIARY ANALYSIS

Short-Term Goals

METADATA TRACING.

- **Support** users in **explaining** observed query **outputs**;
- **Study of data causality** based on determining **data lineage** (or provenance);
- Especially relevant with **queries over multiple sources**;

PATTERN-BASED REGION EXTRACTION.

- Define **complex patterns of genomic features**;
- Enable **the formulation of similarity queries** (e.g., **distal patterns**, or using the notions of **similar/dense/sparse genomic regions**)

DESCRIPTIVE STATISTICS.

- Provide **automatic summarisation of result samples**;
- Integrate **classic significance or regression tests** within the query capabilities;
- Adding innovative features (e.g. peak shapes).

Mid-Term Goals

INTERACTION NETWORKS

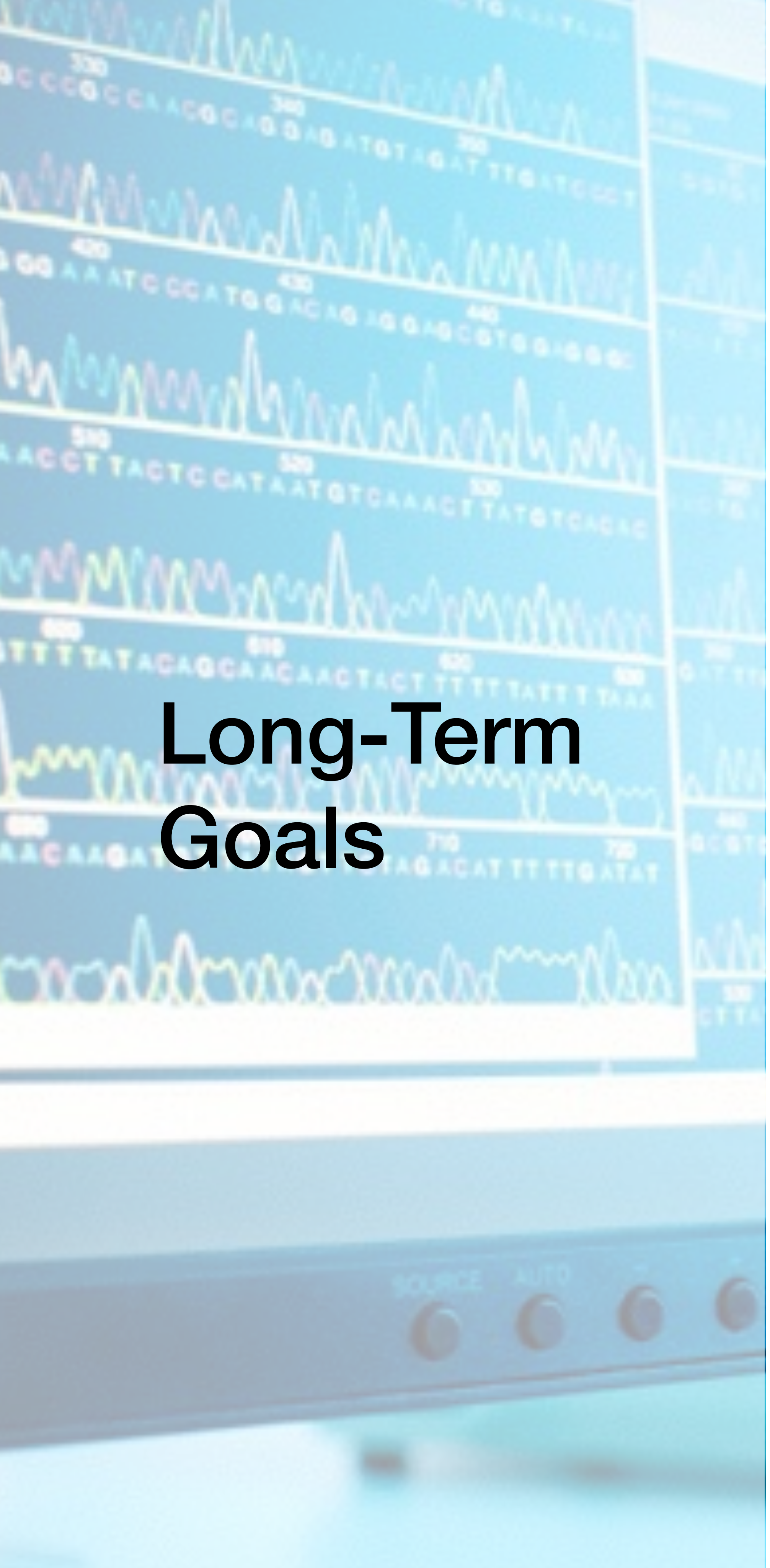
- Provide **automatic translation of query results as** interaction **networks**;
- Integrate **known and/or novel data analysis methods**, based upon **deep learning**, **topological data analysis** or others

INTEGRATED REPOSITORY

- Produce an **integrated repository** with **semantically well-defined and compatible metadata**;
- Include **data from ENCODE**, **TCGA**, **1000 Genomes**, **Roadmap Epigenomics** (and possibly **other** sources).

WEB SERVICES

- Use GMQL to build several **custom queries as public web services**, supporting powerful statistics to indicate the significance of query results



Long-Term Goals

INTERNET OF GENOMES

Use GMQL as a basis for simple interaction protocols for:

- **Requesting information** about remote datasets, using both metadata and region schemas;
- **Sending a query** and obtain result data about is compilation (including estimated data sizes);
- **Launching execution** and then controlling the staging resources and communication load

METADATA AND FEATURE-BASED SEARCH

- Develop **indexing** and **searching methods**, **supporting keyword-based search with semantic query expansion** (leveraging on available ontologies e.g., OBO, UMLS) and **feature-based search patterns**;
- Provide results in ranking order (as in classic search engines);
- Trace **query histories** and **build recommending systems**.

Collaborations

… with BIOLOGISTS

- IEO-IIT (Pier Giuseppe Pelicci, Giuseppe Testa, Stefano Campaner, Bruno Amati)
- University of Insubria (Giovanni Porta)
- NUS Singapore (Lamsoon Wong)
- Broad Institute (Noam Shorem)

Collaborations

… with BIOLOGISTS

- IEO-IIT (Pier Giuseppe Pelicci, Giuseppe Testa, Stefano Campaner, Bruno Amati)
- University of Insubria (Giovanni
- NUS Singapore (Lamsoon Won
- Broad Institute (Noam Shorem)

… with DATA SCIENTISTS

- Harvard University (Pavlos Protopapas)

POLITECNICO MILANO 1863

Collaborations

… with BIOLOGISTS

- IEO-IIT (Pier Giuseppe Pelicci, Giuseppe Testa, Stefano Campaner, Bruno Amati)
- University of Insubria (Giovanni
- NUS Singapore (Lamsoon Won
-

… with DATA SCIENTISTS

… with COMPUTER SCIENTISTS

- Roma1 University (Javier Fernandez, Maurizio Lenzerini): Ontology-based meta-data augmentation and query rewrite.
- Roma3 University (Emanuel Weitschek, Paolo Atzeni, Riccardo Torlone): Integration with TCGA.
- University Bologna (Paolo Ciaccia, Ilaria Bartolini, Piero Montanari): Supporting pattern-based queries from the genome browser.
- Flink Group (Volker Markl, Asterios Katsifodimos): Flink Implementation.
- Paradigm 4 (Marylin Matz, Mike Stonebraker): SciDB Implementation.

Collaborations

… with BIOLOGISTS

- IEO-IIT (Pier Giuseppe Pelicci, Giuseppe Testa, Stefano Campaner, Bruno Amati)
- University of Insubria (Giovanni
- NUS Singapore (Lamsoon Won
-

… with DATA SCIENTISTS

… with COMPUTER SCIENTISTS

- Roma1 University (Javier Fernandez, M based meta-data augmentation and query r
- Roma3 University (Emanuel Weitsch . Integration with TCGA.
- University Bologna (Paolo Ciaccia, Ilaria Barto ri): Supporting pattern-based queries from the genome browser.
- Flink Group (Volker Markl, Asterios Katsifodimos): Flink Implementation.
- Paradigm 4 (Marylin Matz, Mike Stonebraker): SciDB Implementation.

… maybe more! =)

Resources & Websites

http://www.bioinformatics.deib.polimi.it/genomic_computing/ Overview
http://www.bioinformatics.deib.polimi.it/genomic_computing/GMQL/

Includes:

- **Local mode** or **MapReduce mode** (over **Hadoop**, or Hadoop **YARN**) for GNU/Linux systems - Download (122 MB)
- **Web services** (over Hadoop YARN) - Download (60 MB)
- Quick start - Install GMQL and get started
- GMQL tutorial & Complete documentation
- Functional comparison with BEDTools & BEDOPS
- **Pointer to publication** on the Bioinformatics journal

http://www.bioinformatics.deib.polimi.it/GMQL/queries/

Includes:

- **User-friendly interface** to **creating/managing GMQL queries**
- **Custom queries** and **ENCODE / Roadmap Epigenomics datasets**

@ CINECA

We opened a link to CINECA, supporting:

- a **web interface,** where bioinformaticians can **browse the datasets of genomic features and biological/clinical metadata** and **build GMQL queries upon them;**
- **processed data** from **ENCODE** and **Roadmap Epigenomic public sources** (open and anonymised data for secondary use);
- future **availability of processed TCGA data**
- **user-friendly services** designed for biologists

http://www.bioinformatics.deib.polimi.it/GMQL/interfaces/

An Invitation to Join (1/2)

Conference:

BITS 2016 - 13th Annual Meeting of the Bioinformatics Italian Society

University of Salerno, June 15-17, 2016

Session:

Genomic Big Data Management, Modeling and Computing

Organizers: Marco Masseroli
Website: http://bits2016.bioinformatics.it/index.html

An Invitation to Join (2/2)

Conference:

ISMB 2016 - Intelligent Systems for Molecular Biology
Orlando, July 8-12, 2016

Session:

Genomic Big Data Management, Modeling and Computing

Organizers: Stefano Ceri, Marco Masseroli, Emanuel Weitschek
Website: http://www.iscb.org/cms_addon/conferences/ismb2016/specialsessions.php#SST03

... thank you for your attention!