Minería de textos en Biomedicina

Alfonso Valencia
ICREA Professor
Director, Life Sciences Department BSC
Director, Spanish National Bioinformatics Institute (INB-ISCIII)
Head of the Spanish Node of ELIXIR (ELIXIR-ES)

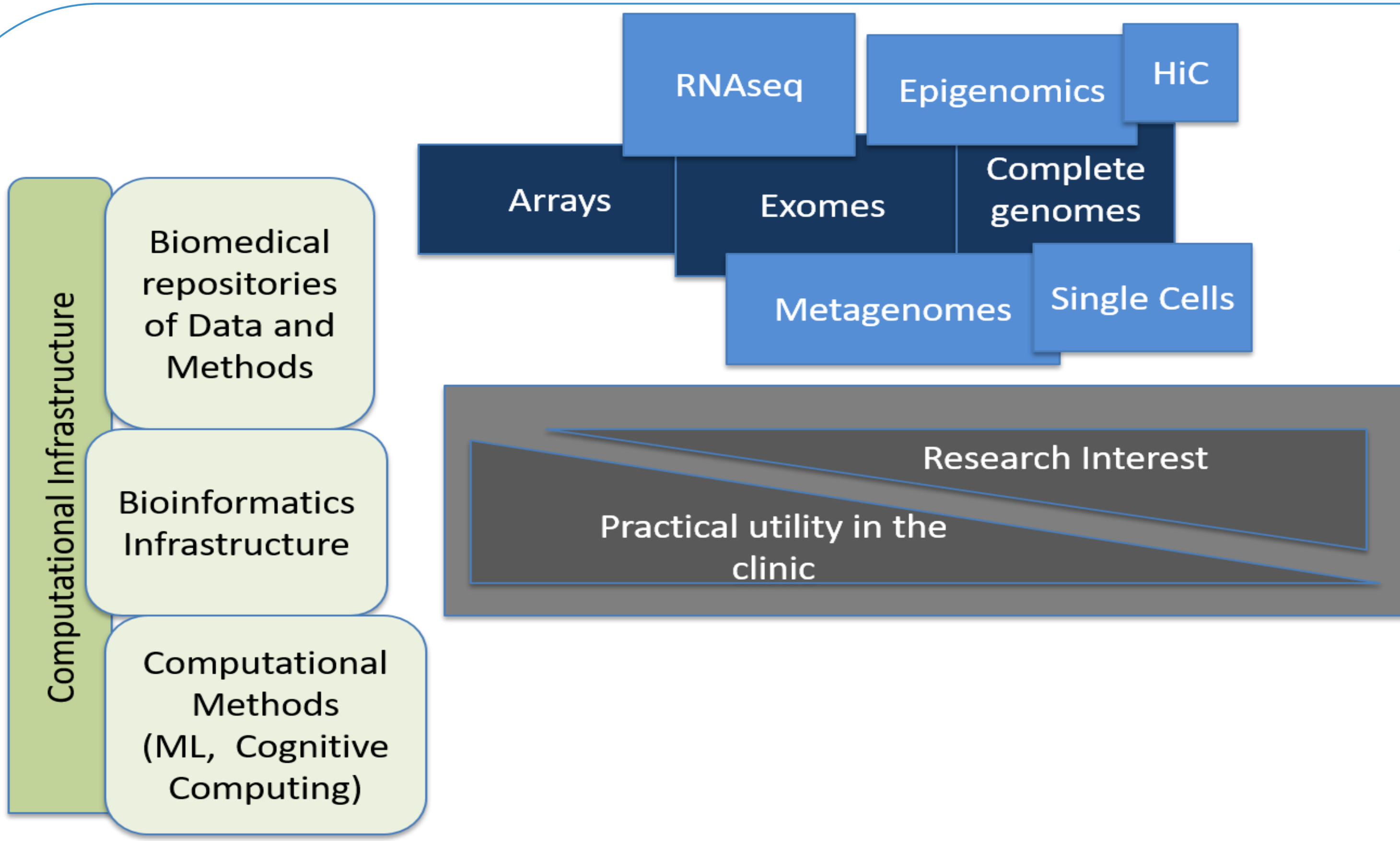
valencia@bsc.es
@alfons_valencia

Madrid April 2017

Plan de Impulso
de las Tecnologías del
Lenguaje

Personalized Medicine

Source New Yorker



Computational Infrastructure

Biomedical repositories of Data and Methods

Bioinformatics Infrastructure

Computational Methods (ML, Cognitive Computing)

RNAseq

Epigenomics

HiC

Arrays

Exomes

Complete genomes

Metagenomes

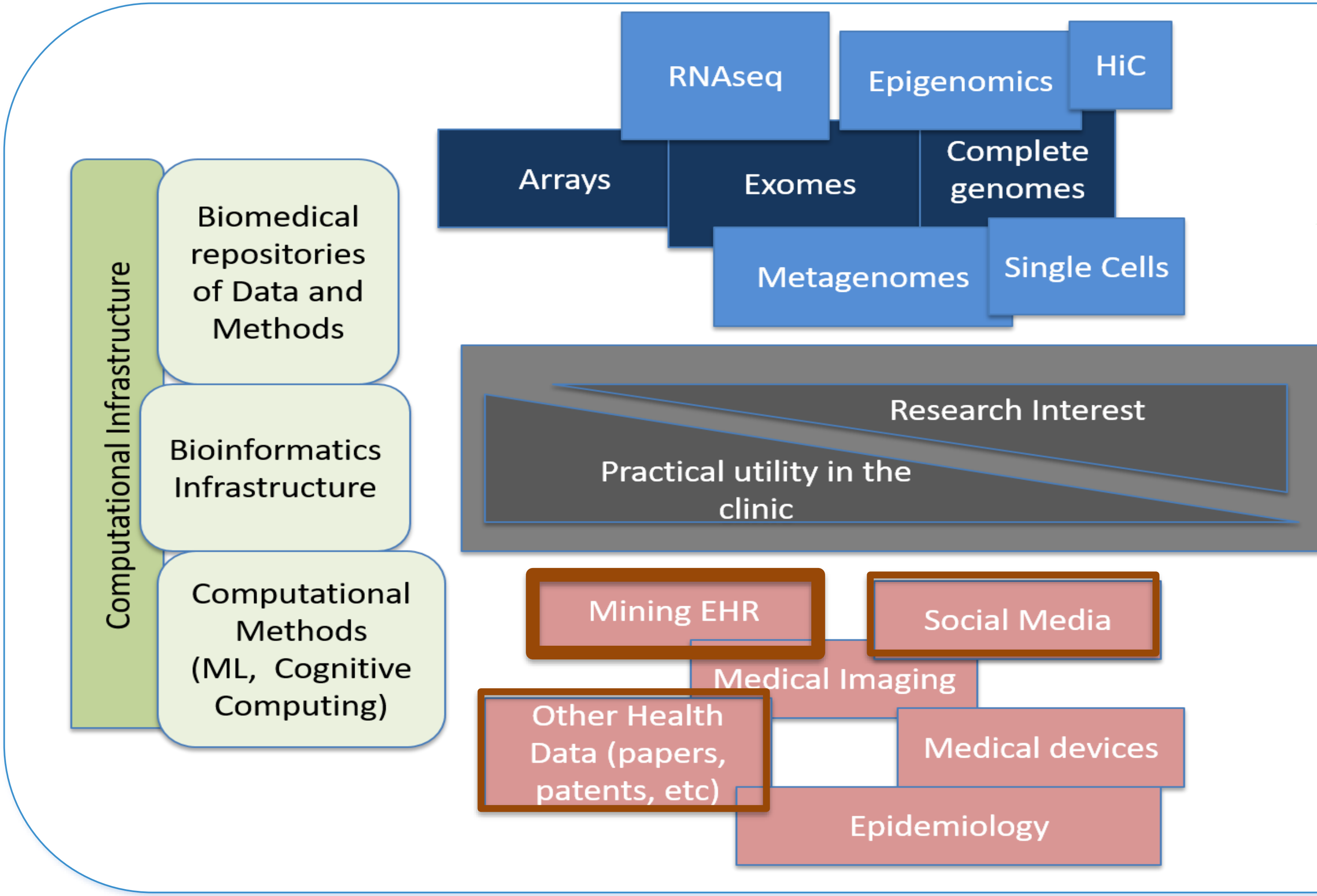
Single Cells

Research Interest

Practical utility in the clinic

Personalized Medicine

Source New Yorker



Computational Infrastructure

Biomedical repositories of Data and Methods

Bioinformatics Infrastructure

Computational Methods (ML, Cognitive Computing)

RNAseq

Epigenomics

HiC

Arrays

Exomes

Complete genomes

Metagenomes

Single Cells

Research Interest

Practical utility in the clinic

Mining EHR

Social Media

Medical Imaging

Other Health Data (papers, patents, etc)

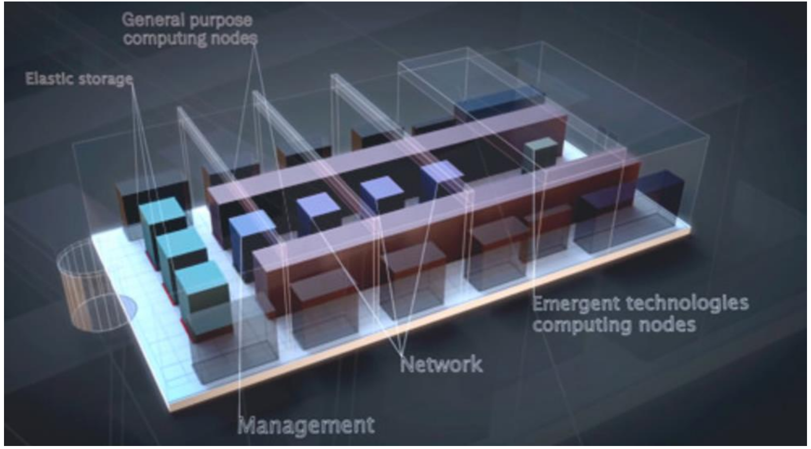
Medical devices

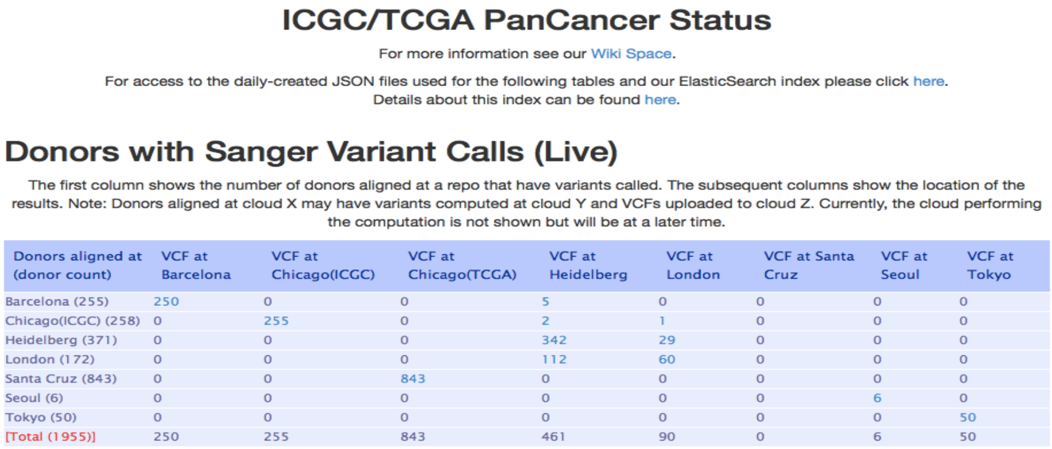
Epidemiology

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

13, 7 Petaflop/s
x12.4 times more powerful than MareNostrum 3.
IBM (+Lenovo, Intel, Fujitsu)
Aprox. €30 million.





ICGC/TCGA PanCancer Status

For more information see our Wiki Space.

For access to the daily-created JSON files used for the following tables and our ElasticSearch index please click here.
Details about this index can be found here.

Donors with Sanger Variant Calls (Live)

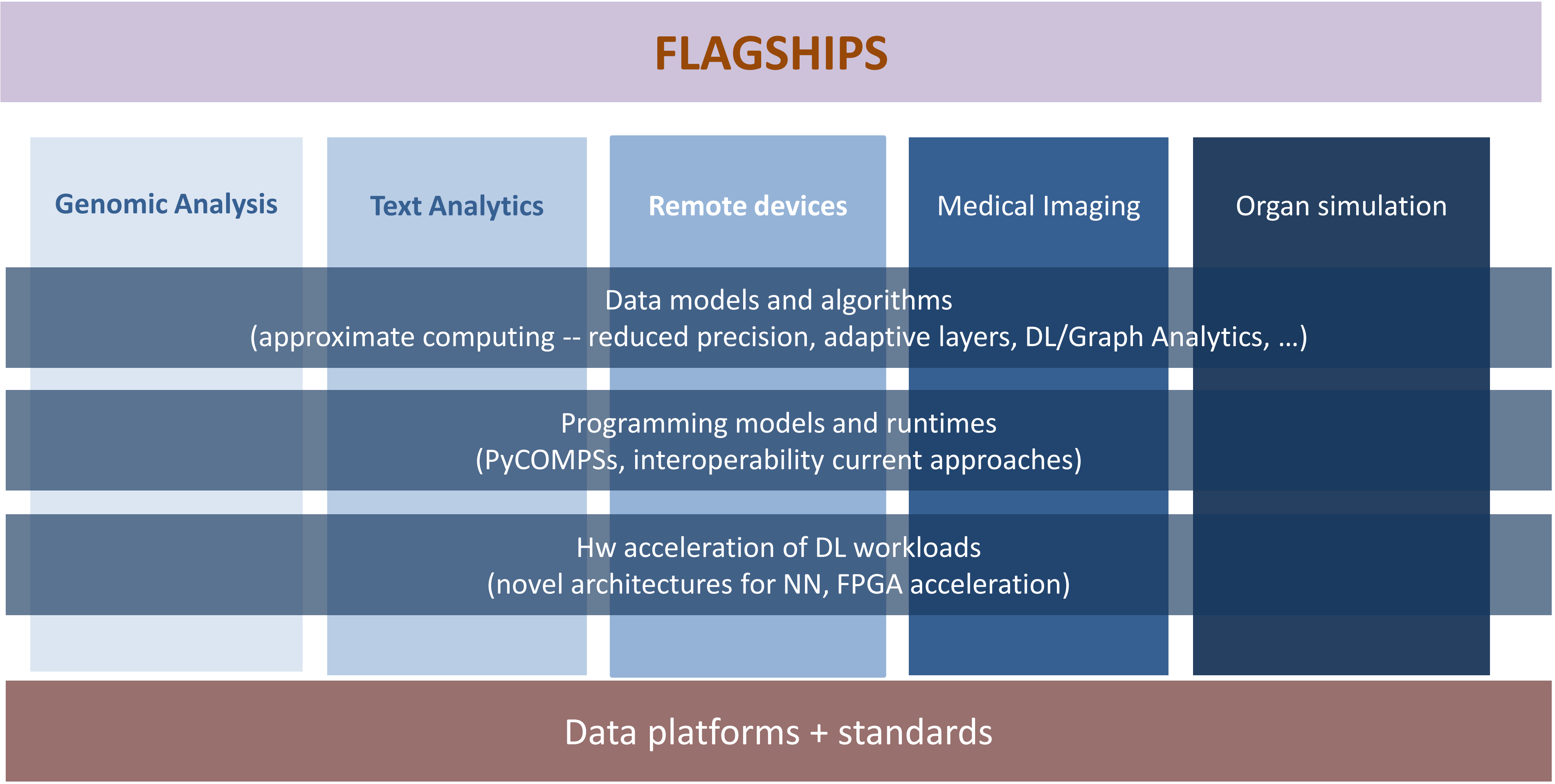
The first column shows the number of donors aligned at a repo that have variants called. The subsequent columns show the location of the results. Note: Donors aligned at cloud X may have variants computed at cloud Y and VCFs uploaded to cloud Z. Currently, the cloud performing the computation is not shown but will be at a later time.

Donors aligned at (donor count)	VCF at Barcelona	VCF at Chicago(ICGC)	VCF at Chicago(TCGA)	VCF at Heidelberg	VCF at London	VCF at Santa Cruz	VCF at Seoul	VCF at Tokyo
Barcelona (255)	250	0	0	5	0	0	0	0
Chicago(ICGC) (258)	0	255	0	2	1	0	0	0
Heidelberg (371)	0	0	0	342	29	0	0	0
London (172)	0	0	0	112	60	0	0	0
Santa Cruz (843)	0	0	843	0	0	0	0	0
Seoul (6)	0	0	0	0	0	0	6	0
Tokyo (50)	0	0	0	0	0	0	0	50
[Total (1955)]	250	255	843	461	90	0	6	50

INB
Instituto Nacional de Bioinformática

elixir SPAIN

La estrategia del BSC en Medicina Personalizada



FLAGSHIPS

Genomic Analysis

Text Analytics

Remote devices

Medical Imaging

Organ simulation

Data models and algorithms
(approximate computing -- reduced precision, adaptive layers, DL/Graph Analytics, …)

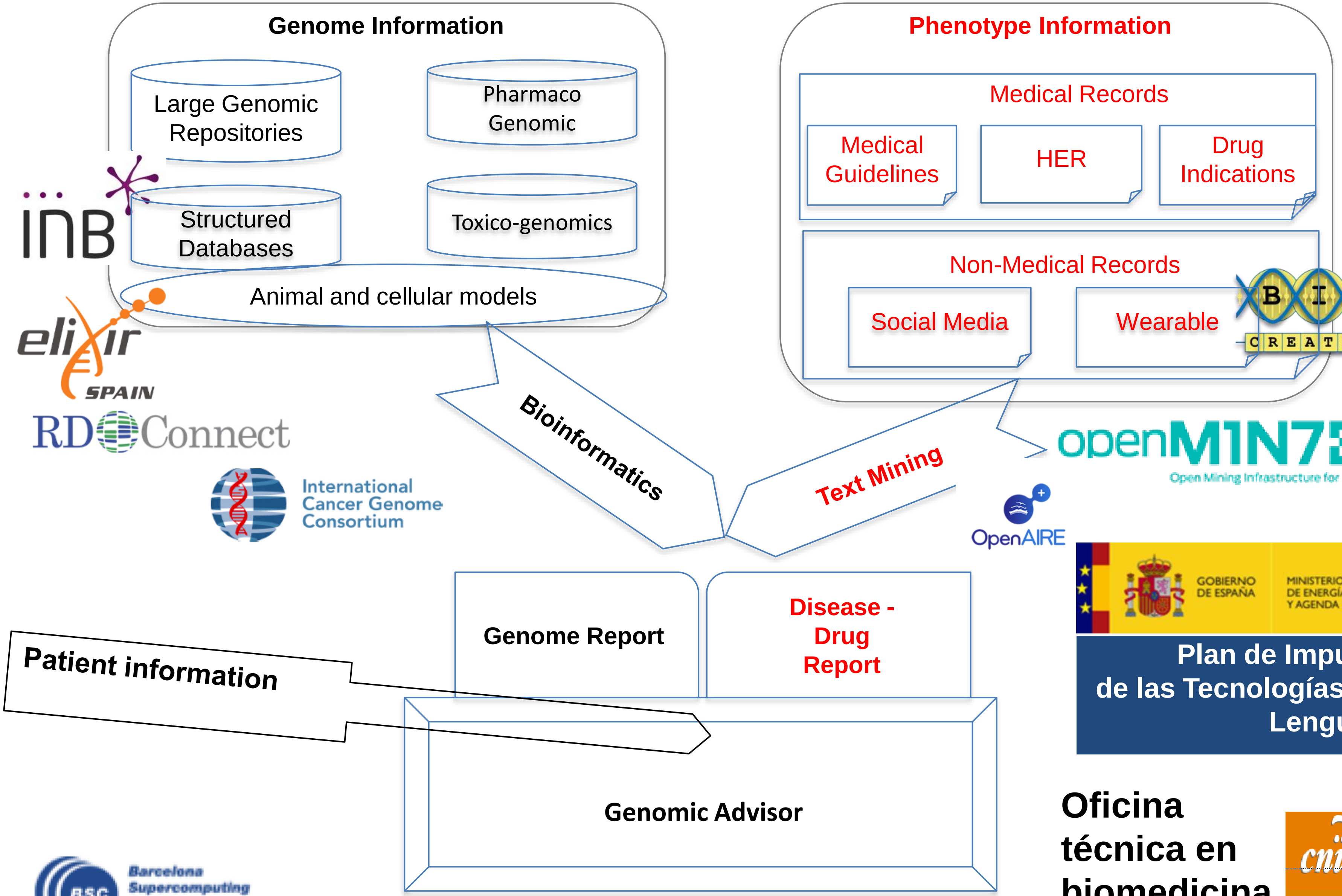
Programming models and runtimes
(PyCOMPSs, interoperability current approaches)

Hw acceleration of DL workloads
(novel architectures for NN, FPGA acceleration)

Data platforms + standards

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Precision Medicine



Genome Information

Large Genomic Repositories

Pharmaco Genomic

Structured Databases

Toxico-genomics

Animal and cellular models

Phenotype Information

Medical Records

Medical Guidelines

HER

Drug Indications

Non-Medical Records

Social Media

Wearable

Bioinformatics

Text Mining

Genome Report

Disease - Drug Report

Patient information

Genomic Advisor

INB

elixir SPAIN

RD Connect

International Cancer Genome Consortium

BIO CREATIVE

openMINTED

Open Mining Infrastructure for Text & Data

OpenAIRE

GOBIERNO DE ESPAÑA

MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL

Plan de Impulso de las Tecnologías del Lenguaje

Oficina técnica en biomedicina

cnio

BSC

Barcelona Supercomputing Center

Centro Nacional de Supercomputación

Actividades Oficina Técnica en Biomedicina

- Creación de corpora anotados y guidelines (abstracts de publicaciones científicas en castellano) > *RDA Iberia*
- Registro de sistemas y workflows (colaboracion con ELIXIR / OpenMinted infrastructures)
- Evaluación (automática) de métodos (basado en metaservidor *becalm) > IberEval*
- Integración en la infraestructura computacional (recursos BSC)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

GOBIERNO DE ESPAÑA

MINISTERIO DE ENERGÍA, TURISMO Y AGENDA DIGITAL

BioCreative V.5 Workshop Agenda [22/04/17]

Wednesday, April 26-27, 2017

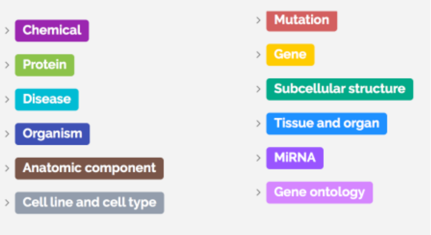
Universitat Politecnica de Catalunya (UPC)

text mining **evaluation of online systems**

- **TIPS (Technical interoperability and performance of annotation servers).**
- **CEMP (Chemical Entity Mention recognition).**
- **GPRO (Gene and Protein Related Object recognition).**

BeCalm

Biomedical Annotation Metaserver



Biomedical Interest

This platform interconnects multiple annotation servers with various recognition abilities. The aim is to offer users information of practical biomedical use.

Servers worldwide

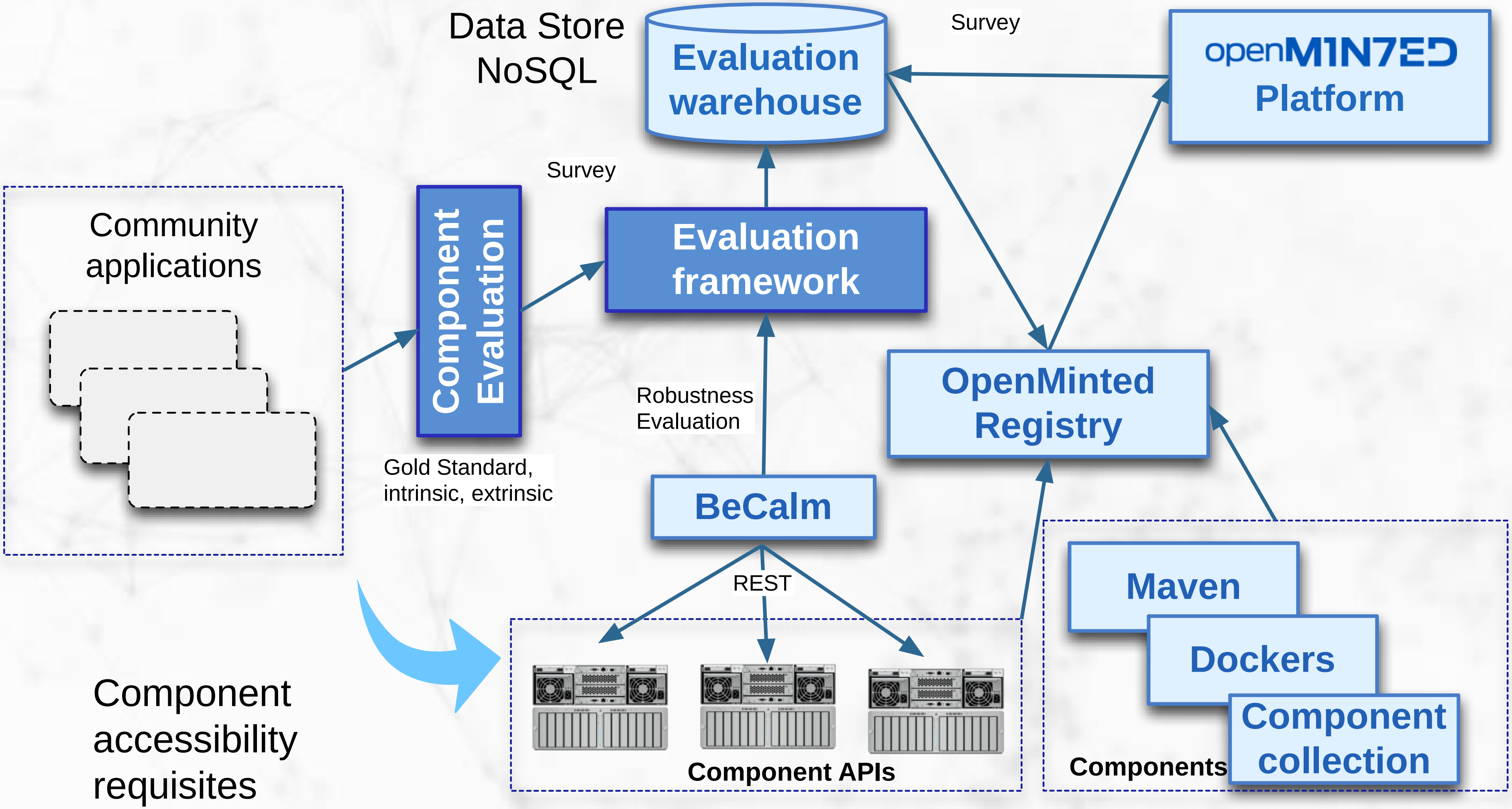
The platform unites and standardised access to textual information extracted by various automatic systems interconnected worldwide.

Benchmarking

Text miners may test the performance of their systems at the meta-server. The platform offers integrated access to high quality gold standards and state-of-the-art prediction systems.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

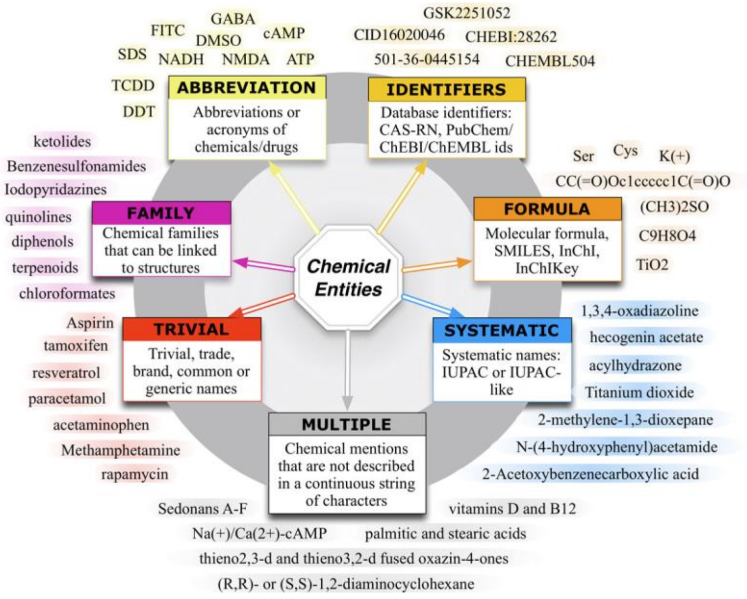
OMTD evaluation framework



Component accessibility requisites

Importancia de los **corpus**

Gold and silver standard



Chemical Entities
ABBREVIATION
Abbreviations or acronyms of chemicals/drugs
FITC
GABA
DMSO
cAMP
SDS
NADH
NMDA
ATP
TCDD
DDT
IDENTIFIERS
Database identifiers: CAS-RN, PubChem/ ChEBI/ChEMBL ids
GSK2251052
CID16020046
CHEBI:28262
501-36-0445154
CHEMBL504
FORMULA
Molecular formula, SMILES, InChI, InChIKey
Ser
Cys
K(+)
CC(=O)Oc1ccccc1C(=O)O
(CH3)2SO
C9H8O4
TiO2
FAMILY
Chemical families that can be linked to structures
ketolides
Benzenesulfonamides
Iodopyridazines
quinolines
diphenols
terpenoids
chloroformates
SYSTEMATIC
Systematic names: IUPAC or IUPAC-like
1,3,4-oxadiazoline
hecogenin acetate
acylhydrazone
Titanium dioxide
2-methylene-1,3-dioxepane
N-(4-hydroxyphenyl)acetamide
2-Acetoxybenzenecarboxylic acid
TRIVIAL
Trivial, trade, brand, common or generic names
Aspirin
tamoxifen
resveratrol
paracetamol
acetaminophen
Methamphetamine
rapamycin
MULTIPLE
Chemical mentions that are not described in a continuous string of characters
Sedonans A-F
vitamins D and B12
Na(+)/Ca(2+)-cAMP
palmitic and stearic acids
thieno2,3-d and thieno3,2-d fused oxazin-4-ones
(R,R)- or (S,S)-1,2-diaminocyclohexane

ChemdNER corpus

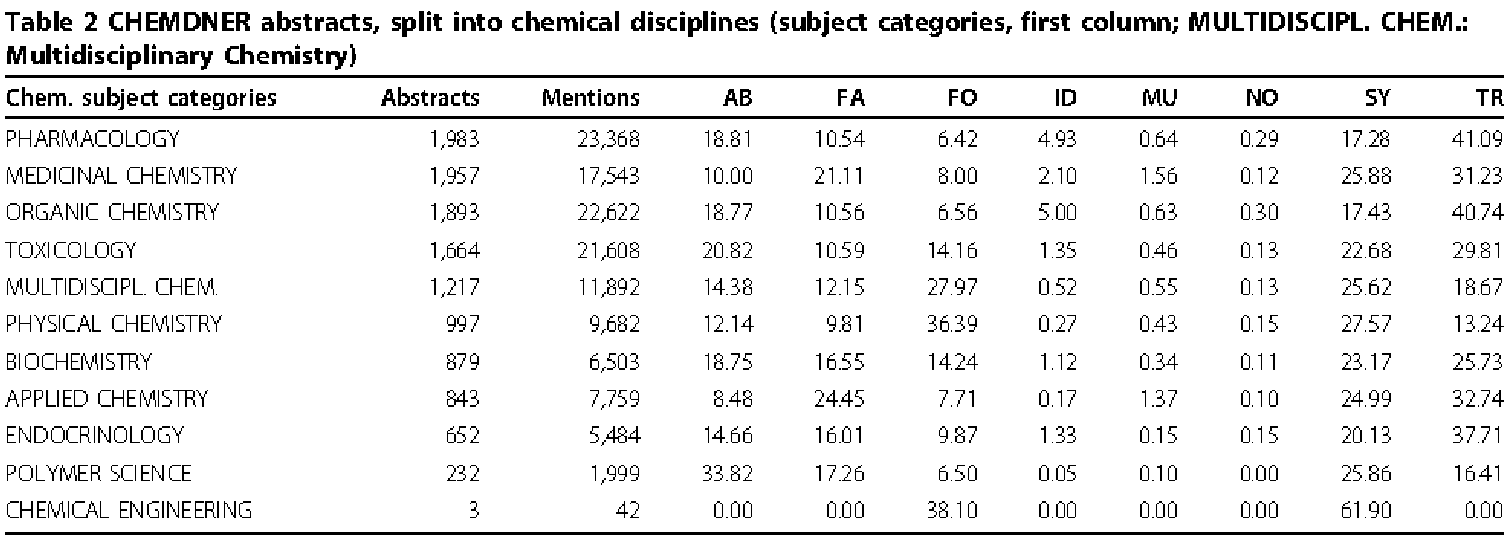
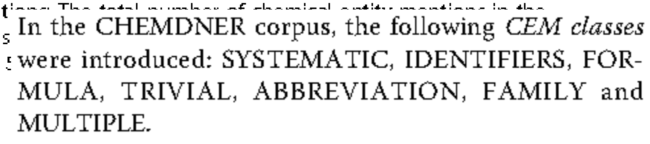


Table 2 CHEMDNER abstracts, split into chemical disciplines (subject categories, first column; MULTIDISCIPL. CHEM.: Multidisciplinary Chemistry)

Chem. subject categories	Abstracts	Mentions	AB	FA	FO	ID	MU	NO	SY	TR
PHARMACOLOGY	1,983	23,368	18.81	10.54	6.42	4.93	0.64	0.29	17.28	41.09
MEDICINAL CHEMISTRY	1,957	17,543	10.00	21.11	8.00	2.10	1.56	0.12	25.88	31.23
ORGANIC CHEMISTRY	1,893	22,622	18.77	10.56	6.56	5.00	0.63	0.30	17.43	40.74
TOXICOLOGY	1,664	21,608	20.82	10.59	14.16	1.35	0.46	0.13	22.68	29.81
MULTIDISCIPL. CHEM.	1,217	11,892	14.38	12.15	27.97	0.52	0.55	0.13	25.62	18.67
PHYSICAL CHEMISTRY	997	9,682	12.14	9.81	36.39	0.27	0.43	0.15	27.57	13.24
BIOCHEMISTRY	879	6,503	18.75	16.55	14.24	1.12	0.34	0.11	23.17	25.73
APPLIED CHEMISTRY	843	7,759	8.48	24.45	7.71	0.17	1.37	0.10	24.99	32.74
ENDOCRINOLOGY	652	5,484	14.66	16.01	9.87	1.33	0.15	0.15	20.13	37.71
POLYMER SCIENCE	232	1,999	33.82	17.26	6.50	0.05	0.10	0.00	25.86	16.41
CHEMICAL ENGINEERING	3	42	0.00	0.00	38.10	0.00	0.00	0.00	61.90	0.00

Abstracts: The number of abstracts associated with that category in the CHEMDNER corpus. Mentions: The total number of chemical entity mentions in the abstracts of that category. Remaining columns: The values provided for the different SACEM class AB: ABBREVIATION, FA: FAMILY, FO: FORMULA, ID: IDENTIFIER, MU: MULTIPLE, NO: NO CLASS, SY: S



In the CHEMDNER corpus, the following *CEM classes* were introduced: SYSTEMATIC, IDENTIFIERS, FORMULA, TRIVIAL, ABBREVIATION, FAMILY and MULTIPLE.