

BIG DATA?

and small data



Concept and Works
related to Big Data
in our Research Group

TIC i Big Data en Asma i MPOC

Miguel Angel Mayer

Research Programme on Biomedical Informatics (GRIB)

Institut de Recerca de l'Hospital del Mar (IMIM) i

Universitat Pompeu Fabra (IMIM-UPF)

Coordinador Grup TIC CAMFiC



Jornada Debat en Gestió Clínica Asma i MPOC
Grup de Gestió CAMFiC,
Barcelona 3 febrer 2015



Parc de Recerca Biomèdica
de Barcelona (PRBB)



Hospital del Mar – Parc de Salut Mar

Facultat Ciències de la Salut i de la Vida.

Campus Universitari Mar
Universitat Pompeu Fabra



**Universitat
Pompeu Fabra**
Barcelona





“**Big Data** are data whose scale, diversity and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it”

IMIA working group on “Data Mining and Big Data Analytics”
From R. Bellazi, IMIA Yearbook of Medical Informatics 2014

40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005



Volume SCALE OF DATA

It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day

Most companies in the
U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored

The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015
4.4 MILLION IT JOBS
will be created globally to support big data,
with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be

150 EXABYTES
[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**
are shared on Facebook
every month



Variety DIFFERENT FORMS OF DATA

By 2014, it's anticipated there will be

**420 MILLION
WEARABLE, WIRELESS
HEALTH MONITORS**



**4 BILLION+
HOURS OF VIDEO**
are watched on
YouTube each month



400 MILLION TWEETS
are sent per day by about 200
million monthly active users



The New York Stock Exchange captures

**1 TB OF TRADE
INFORMATION**

during each trading session



By 2016, it is projected there will be

**18.9 BILLION
NETWORK
CONNECTIONS**

— almost 2.5 connections
per person on earth



Velocity ANALYSIS OF STREAMING DATA

Modern cars have close to
100 SENSORS
that monitor items such as
fuel level and tire pressure



**1 IN 3 BUSINESS
LEADERS**

don't trust the information
they use to make decisions



**27% OF
RESPONDENTS**

in one survey were unsure of
how much of their data was
inaccurate

Veracity UNCERTAINTY OF DATA

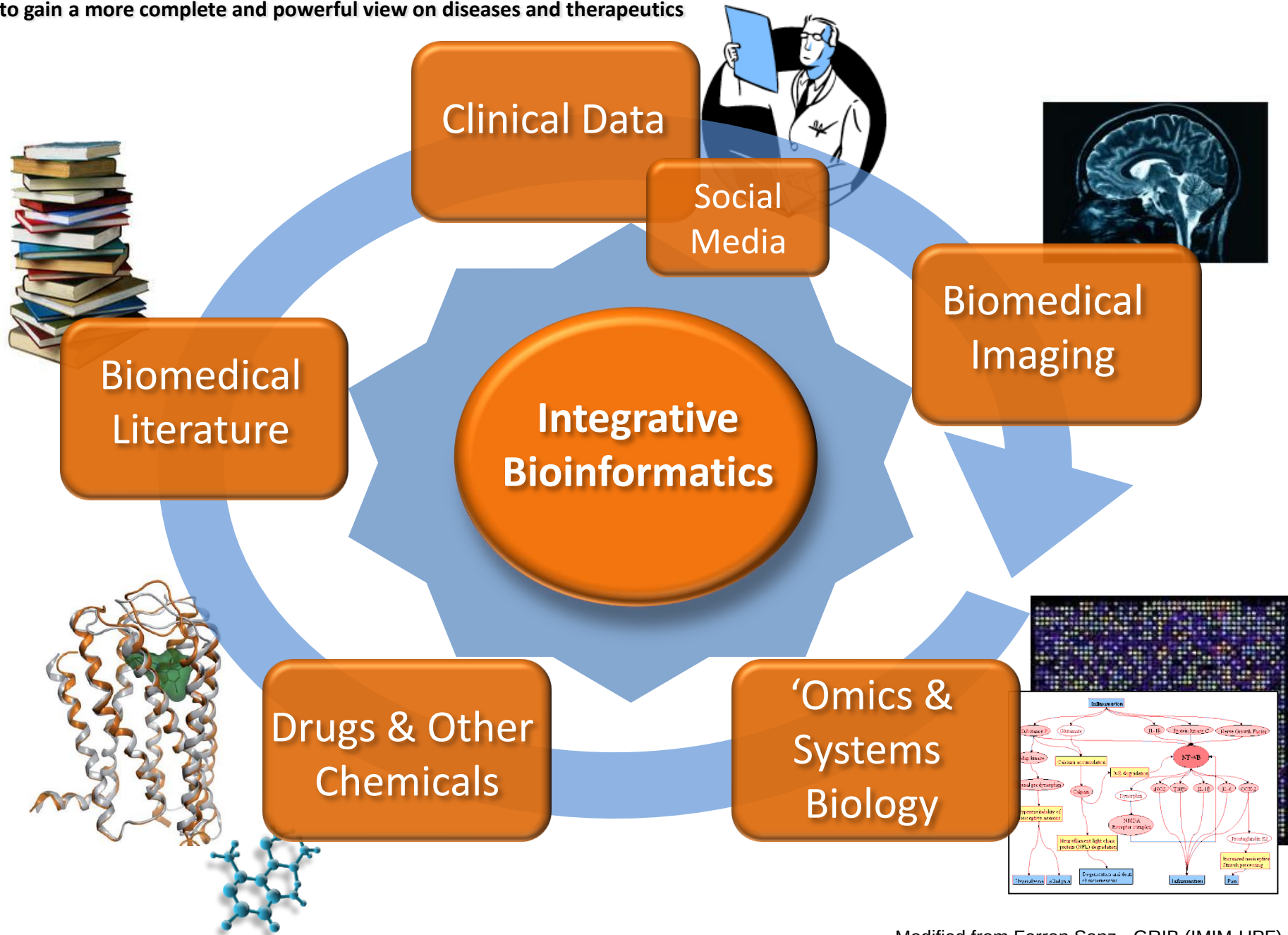
Poor data quality costs the US
economy around

\$3.1 TRILLION A YEAR



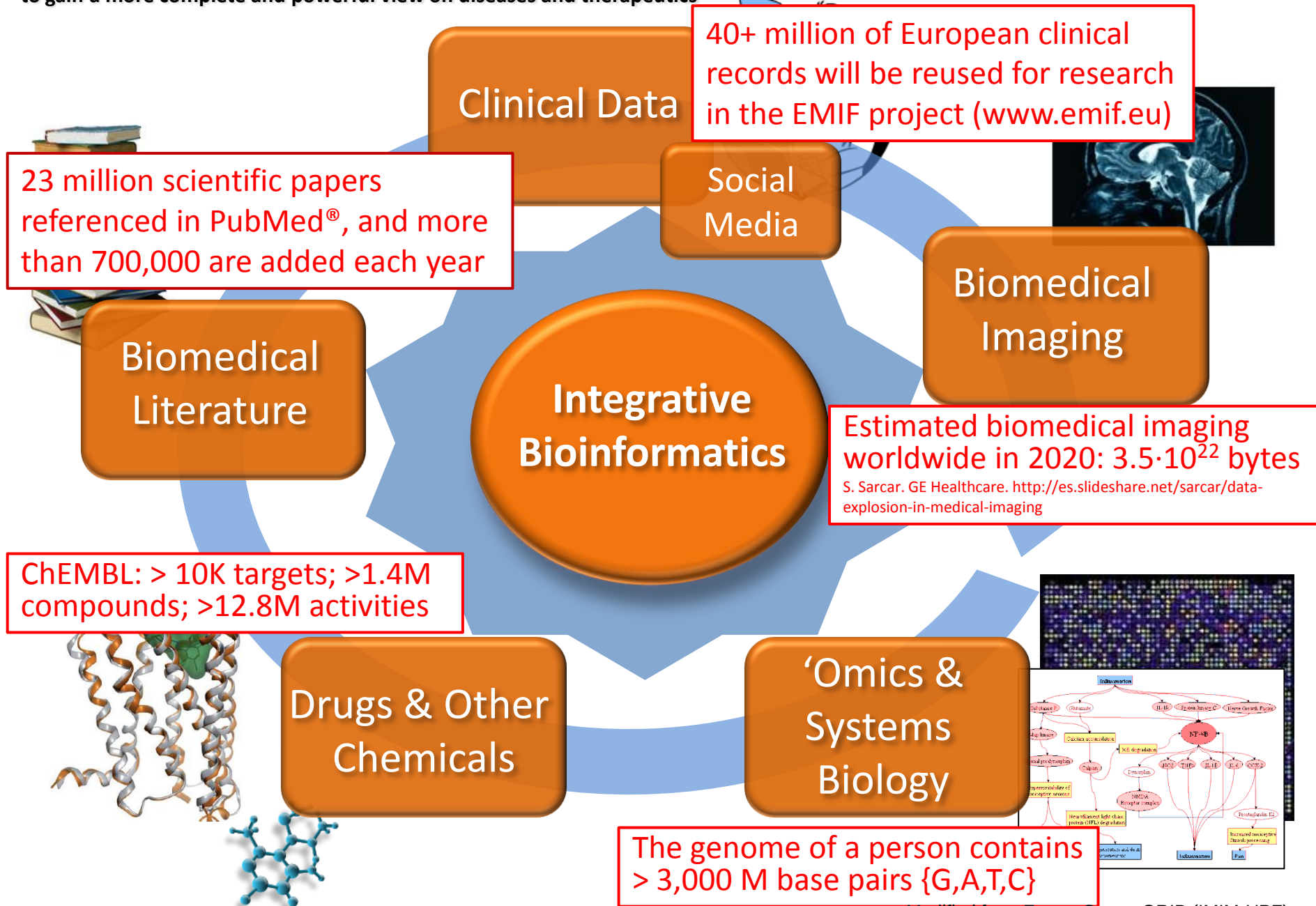
Integration of heterogeneous biomedical information

to gain a more complete and powerful view on diseases and therapeutics



Integration of heterogeneous biomedical information

to gain a more complete and powerful view on diseases and therapeutics



Social Media



Grupos sobre **alimentación saludable**
en Facebook: contenido y características

Leis A, Mayer MA et al. Gaceta Sanitaria 2013

Information about **Vaccines** in English
and Spanish on Facebook.

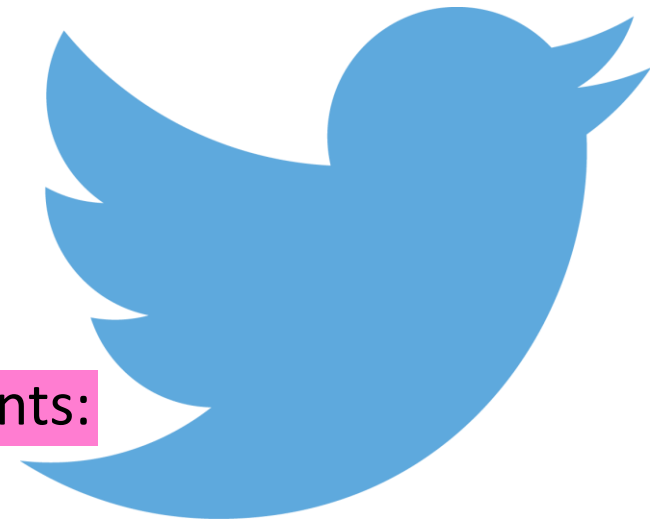
Mayer MA et al. iProceedings Medicine 2.0 2013

Exploring **Brand-Name Drug** Mentions
on Twitter for Pharmacovigilance

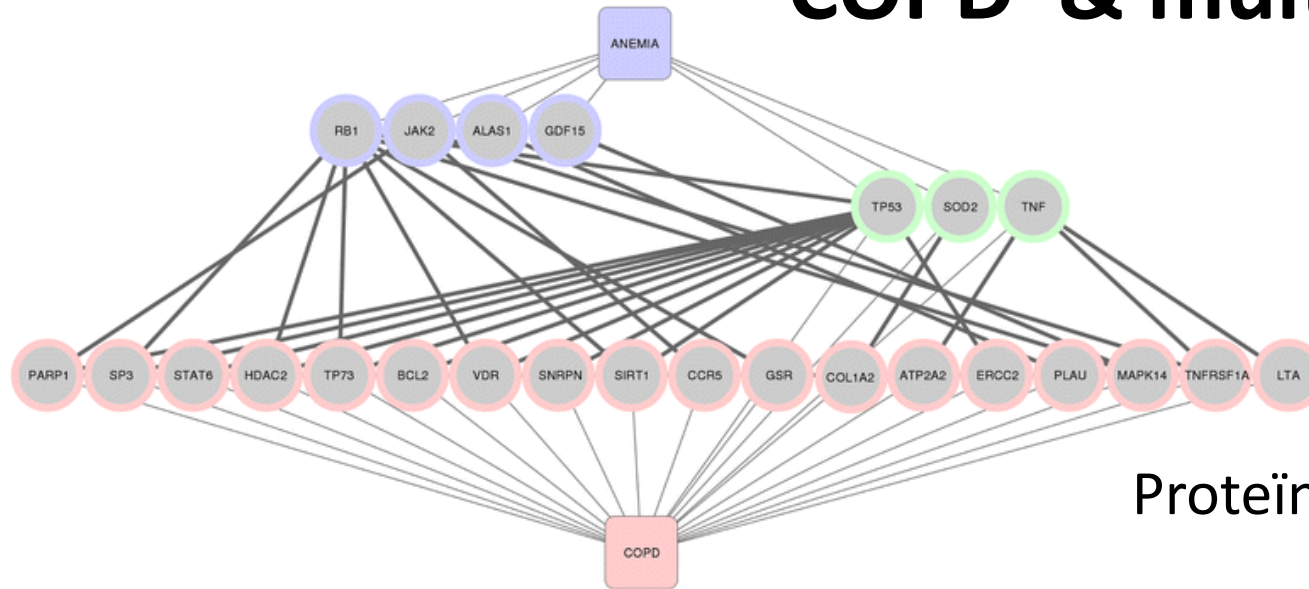
Carbonell P, Mayer MA, Bravo A. Proceedings MIE 2015

How Twitter Is Used in **International Health Events:**
World Aids Day Case Study.

Leis A, Mayer MA. iProceedings Medicine 2.0 London 2013

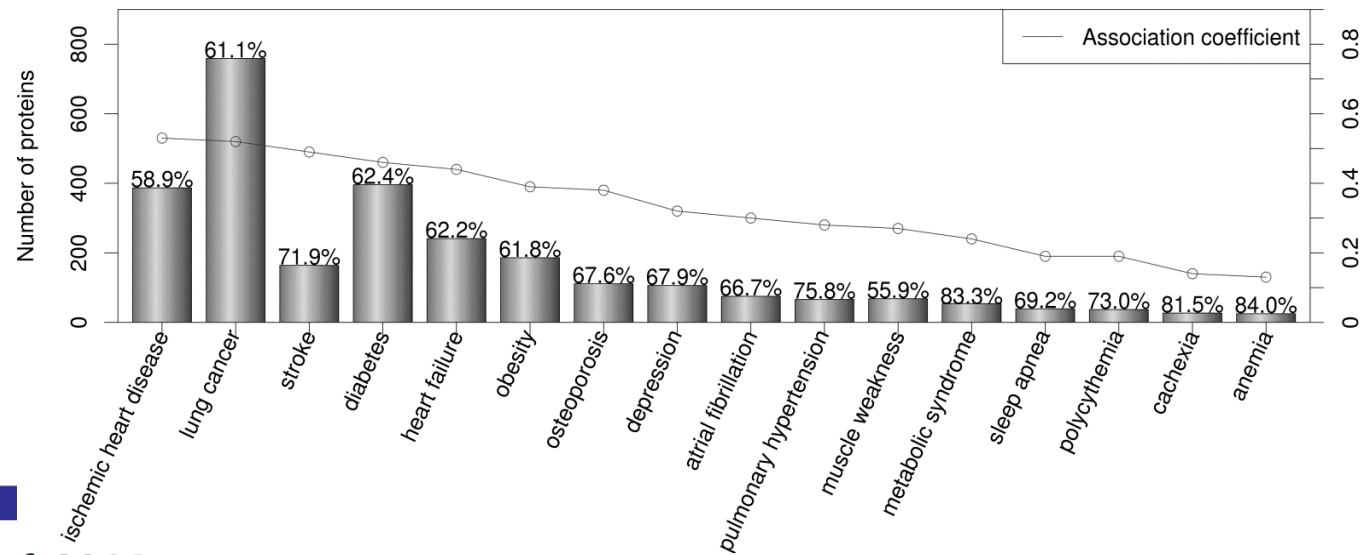


COPD & multimorbidities



Proteïnes, anèmia & MPOC

Tabac i nombre de
proteïnes compartides
MPOC i altres malalties



Grosdidier et al. *Respiratory Research* 2014, **15**:111
<http://respiratory-research.com/content/15/1/111>

RESEARCH

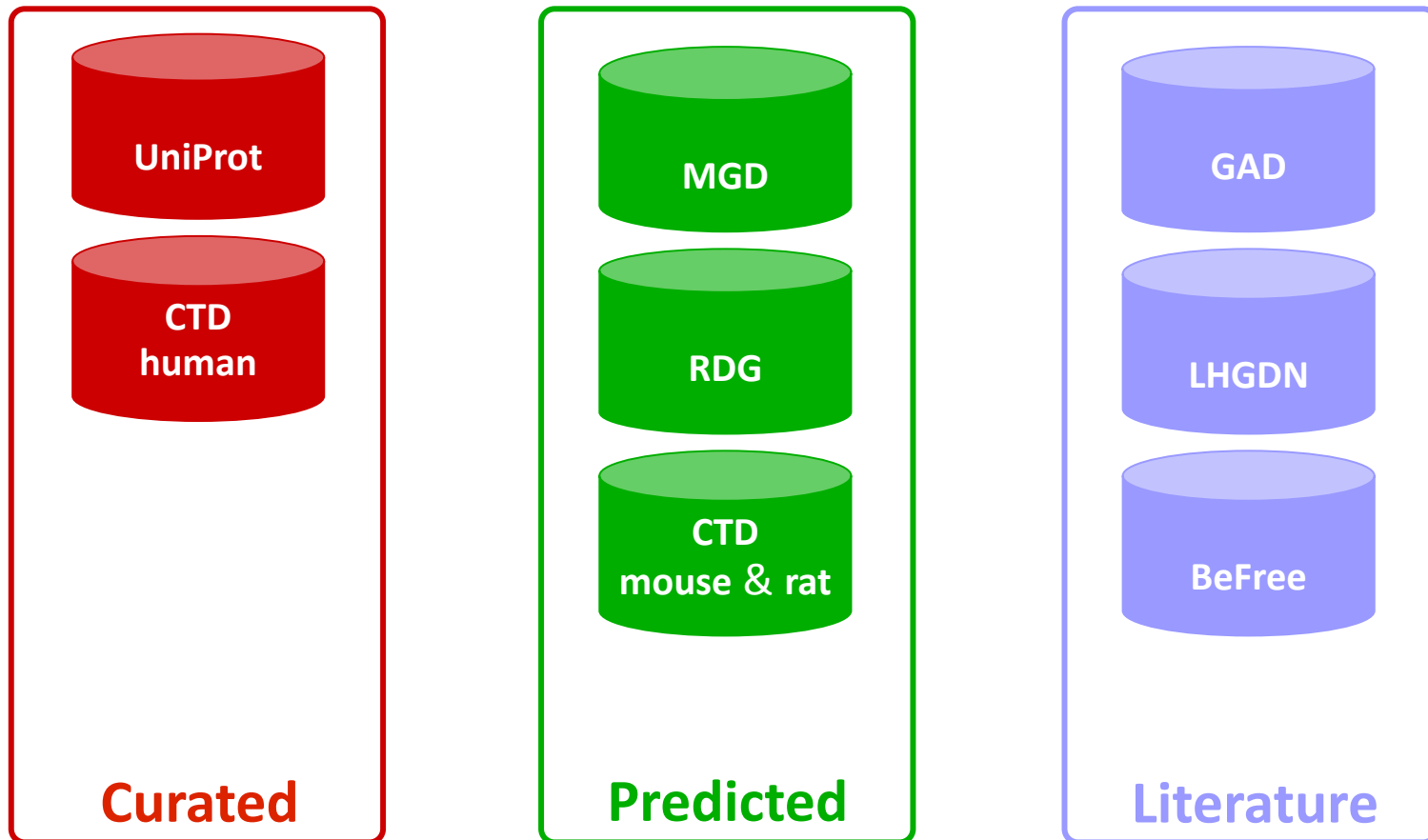
Network medicine analysis of COPD multimorbidities

Solène Grosdidier¹, Antoni Ferrer², Rosa Faner³, Janet Piñero¹, Josep Roca⁴, Borja Cosío⁵, Alvar Agustí^{3,4}, Joaquim Gea², Ferran Sanz¹ and Laura I Furlong^{1*}

DisGeNET

<http://ibi.imim.es/DisGeNET>

- A comprehensive resource on gene-disease associations
- Integrates information from publicly available databases and the literature (text mining)



EHR

Electronic Health Records
Special role in Big Data

Projectes que participa el GRIB amb l'IMI



- **Exploitation of millions of electronic patient records for contributing to the advancement of biomedical research.**

- Budget: 56.4 M€ (24.1 M€ of industrial contribution).
- Duration: January 2013 – December 2017.
- Partners: 9 pharma companies, 7 SMEs, 36 academic institutions and 3 patients organisations.



- **Computational extraction, integration and exploitation of the open source knowledge in order to facilitate pharmaceutical R&D.**

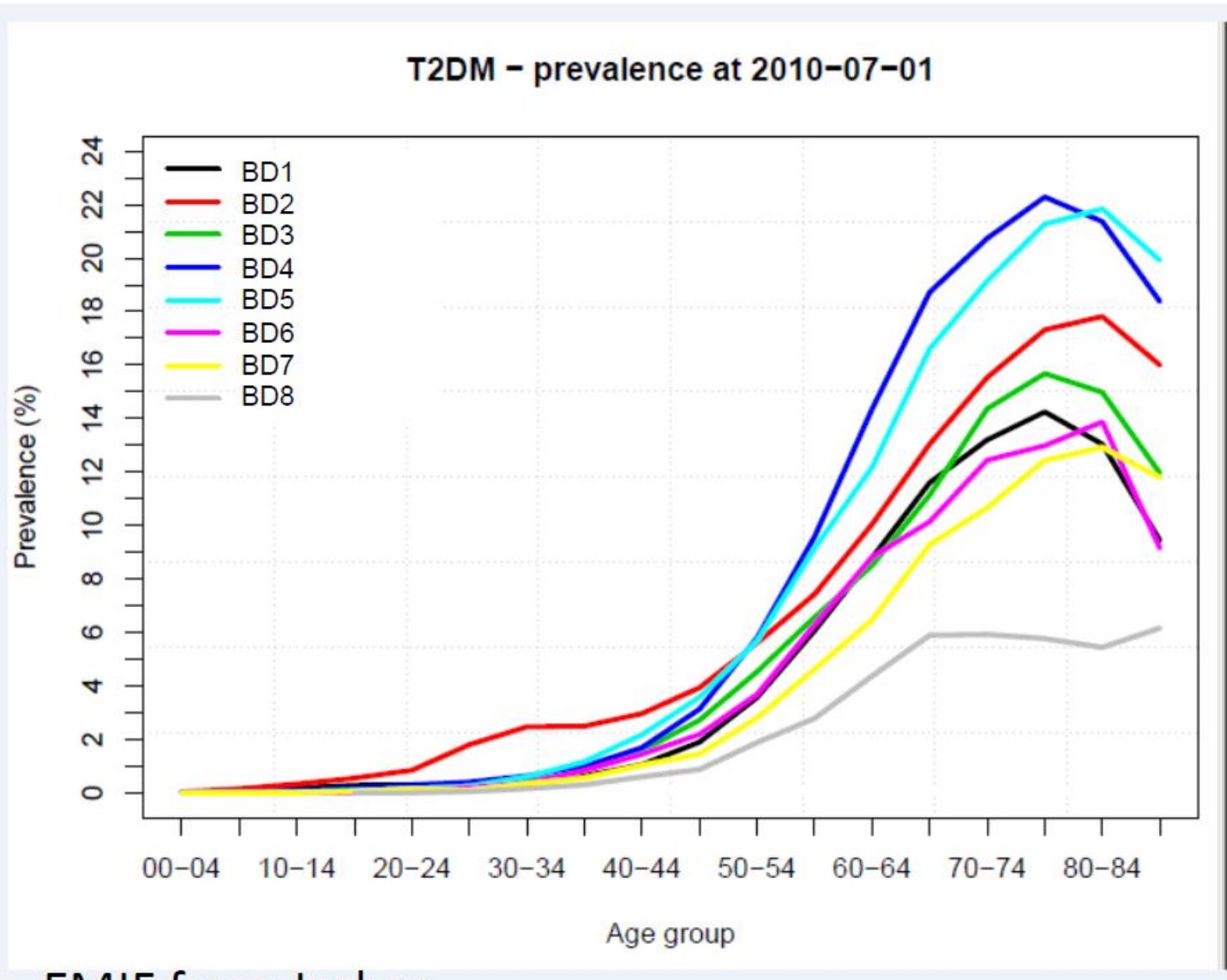
- Budget: 20.9 M€ (6.5 M€ of industrial contribution).
- Duration: February 2011 – February 2016 (3 years plus 2 years through an ENSO extension).
- Partners: 10 pharma companies, 6 SMEs and 15 academic institutions.



- **Information sharing and integration for the development of advanced predictive models of drug toxicity.**

- Budget: 18.7 M€ (10 M€ of industrial contribution).
- Duration: January 2010 – December 2016 (5 years plus 2 years through an ENSO extension).
- Partners: 13 pharma companies, 6 SMEs and 11 academic institutions.
- GRIB: academic coordinator.

Diabetes Mellitus type 2 in different countries (EMIF)



EMIF from Jerboa

Data **anonymisation**
and pseudonymisation
techniques and access
control

To assure that each
participating
database retains **full
control** over their
data

To manage **different
languages** for narrative
information and
numerous coding
schemes

To assure that all
data sources share a
**common
understanding** of the
required data

To **implement
dedicated extraction
software** and
controlled processing of
data by a shared
software environment
with novel analytical
methods and tools

To overcome the
risk related to
**evolving ethical and
legal regulations** in
Europe

To manage
different **coding
systems and
versions** (ICD9-CM,
ICD10, ICPC, READ,
etc.)

Challenges and critical issues

Reflexions finals

Big data ofereix **moltes expectatives**

Ciutadans participació més activa
sobre la seva salut

Els sistemes sanitaris cap una medicina més
personalitzada (patient-centric)

Especial rellevància dels aspectes
ètics i legals

Sostenibilitat dels sistemes sanitaris, utilització més
eficient informació i superar fragmentació

Formació específica en Big Data

Promoure el **Public and Private Partnership (PPP)**

Integrative Biomedical Informatics Group (GRIB)

<http://grib.upf.edu>



M.A. Mayer



A. Gutiérrez



A. Bauer-Mehren

DisGeNET



Open PHACTS



eTOX



Universitat
Pompeu Fabra
Barcelona

Contact:

Miguel Angel Mayer

 @mmayerp

mmayer@imim.es

GRIB (IMIM-UPF)

<http://grib.upf.edu>

CYBER THREAT MAP



Subscribe to FireEye Alerts

[X] NEW ATTACK: FROM [UNITED STATES] TO [KOREA, REPUBLIC OF]
[X] NEW ATTACK: FROM [JORDAN] TO [COLOMBIA]
[X] NEW ATTACK: FROM [UNITED STATES] TO [SAUDI ARABIA]

LOCAL TIME

9:53:17

ATTACKS TODAY

19,668

CYBER THREAT MAP



ATTACKERS

TOP COUNTRIES
(PAST 30 DAYS)

Powered by FireEye Labs

TOP 5 REPORTED INDUSTRIES [PAST 30 DAYS]

EDUCATION
ENERGY/UTILITIES
HIGH-TECH
GOVERNMENT: FEDERAL
GOVERNMENT: STATE & LOCAL

VIEW FULL SCREEN

The "Cyber Threat Map" is based on a subset of real attack data, which is optimized for better visual presentation. Customer information has been removed for privacy.