BIOMEDICAL INFRASTRUCTURES IN THE BIG DATA ERA

HARIDIMOS KONDYLAKIS

COMPUTATIONAL BIOMEDICINE LABORATORY,

FOUNDATION FOR RESEARCH & TECHNOLOGY – HELLAS

TOKYO, MARCH 2016
Develop novel ICT technologies in the wider context of predictive, individualized, preventive and participatory (the P4) medicine aiming at:

- the semantic interoperability of biomedical data tools and models for enhancing biomedical knowledge discovery
- the optimal management of chronic diseases (such as diabetes, cardiovascular disease)
- the optimization of diagnosis and treatment through the development of novel predictive models, medical imaging analysis and clinical decision support tools
- the implementation of well-established in silico methods and tools towards novel approaches that could be incorporated in the medical clinical research

Coupled with the Center for eHealth and Applications and Services for evolving R&D results into commercial products focused on Integrated Care Solutions
OUTLINE

1. BIG DATA
   1. BIG DATA INTEGRATION
   2. BIOMEDICAL INFRASTRUCTURES IN THE BIG DATA ERA

2. EXPERIENCES

3. LESSONS
BIG DATA
DO THEY REALLY EXIST?

16,000 hospitals worldwide collect data on patients

4.9 million patients worldwide will use remote monitoring devices by 2016

An 18% annual compound growth rate is anticipated between 2010 and 2016 for patients that will use remote monitoring devices

80% of health data is unstructured and stored in hundreds of forms such as lab results, images, and medical transcripts

Patient monitoring equipment pumps out an average of 1,000 readings per second or 86,400 readings in a day

The Body as a Source of Big Data

Today, data usage is essential for healthcare providers to base a patient's complete care strategy, make more informed decisions and enhance treatment and outcomes.

Access to electronic health records beyond the hospital

665TB in data

It is estimated that by 2020, the average hospital will generate over 750MB of data per patient per day.

20-40% savings

36.6M in savings

20-40% increase in revenue

80% of all data is never used.
SCIENTIFIC DATA GROWTH

[Howe & Halperin, 2012]
Value

Volume

Variety

Veracity

Velocity

TA

- genotypes
- drug profiles
- text sources
- molecular profiles
- electronic health records
- wearable sensor data
- environmental data

- genotypes
- molecular profiles
- images
WHY DO WE CARE?

• **Google Flu Trends:**
  - advance indication in the 2007-08 flu season (Nature, 2009)
  - the 2009 H1N1 outbreak

• **IBM: Predict Heart Disease Through Big Data Analytics**
  - traditional: EKGs, heart rate, blood pressure
  - big data analysis: connecting
    - exercise and fitness tests:
    - diet
    - fat and muscle composition
    - genetics and environment
    - social media and wellness: share information
WHY DO WE CARE?
The quest for knowledge used to begin with grand theories. Now it begins with massive amounts of data. Welcome to the Petabyte Age.
DATA INTEGRATION
INTEROPERABILITY & DATA INTEGRATION

• Often although people build databases in isolation, they want to share their data.

• Interoperability means the ability of two or more information systems to accept data from each other \[\text{[eJHI2014]}\]

• Data integration is the problem of providing unified and transparent access to a collection of data stored in multiple, autonomous, and heterogeneous data sources \[\text{[BIT2006]}\]

• Integration generally goes beyond mere interoperability to involve some degree of functional dependency also.
DATA INTEGRATION APPROACHES

Virtual Warehouses
Ontologies: “Formal models about how we perceive a domain of interest and provide a precise, logical account of the intended meaning of terms, data structures and other elements modeling the real world”

Can Serve as the Global Schema either in Virtual Integration or in DWHs
Big Data + Data Integration
BIG DATA INTEGRATION

Virtual Integration

Traditional Data Warehouses

Big Data Warehouses

Query reformulation
Query over materialized data

OLTP

ETL

DWH

OLTP

ETL

Haddop

DWH

Wrapper Extractor

Wrapper Extractor

Wrapper Extractor

Wrapper Extractor

RDBMS

HTML/XML

RDBMS

Big Data DB
(HDFS/Cassandra etc.)
FP6 - THE ACGT PROJECT (2006-2010)

• Developing open-source, semantic and grid-based technologies in support of post genomic clinical trials in cancer research

• Issues identified
  • Tied to grid architecture (tight)
  • Performance issues
  • One ontology not enough
  • The interaction with data sources is not easy
  • Clinical Trial domain is evolving at high rate
FP7 - THE P-MEDICINE PROJECT (2011-2015)

• From data sharing and integration via VPH models to personalized medicine -

• www.p-medicine.eu

• Goals: to develop new tools, IT infrastructure and VPH models that help accelerating personalized medicine for the benefit of the patient
FP7 - THE P-MEDICINE PROJECT

• Project Outcomes (wrt data management)
  • HDOT Ontology this time trying to integrate existing ontologies
  • Ontology Annotator/Data Translator
  • Data Warehouse

• Issues identified
  • Interfaces should conform to highly restrictive legal policies
  • Late access to real data
  • There are many textual data

Dispersed operational systems
- Trial management systems
- Clinical record storage
- Biobanking databases
- Clinical imaging stores
- NGS data

ETL (Extract, Transform, Load)

Data Warehouse

Analysis tools

- Clinical decision making
- Data mining
- Modelling

- Tools and environments enabling the re-use of electronic health records
- EURECA aimed at establishing a semantic link between EHRs and CTs
  - Building a pragmatic scalable secure semantic interoperability solution
  - Define a modular Standards-based semantic core dataset to cover our clinical domain
  - Define information models of the sources and their mappings
- Manage the various sources of heterogeneity in existing systems
  - Many clinical research and care systems, often home-grown
  - Structured, semi-structured and free-text documents
  - Many standards and terminologies, but with low adoption
  - Different laws and policies
- Build advanced applications to support clinical research and to speed up knowledge transfer into practice

• Driving Excellence in Integrative Cancer Research through Innovative Biomedical Infrastructures
Multiple DWHs at clinical sites
Textual sources under consideration as well
Legal implications avoided by local installations
Although HL7 RIM is extensively used, data type’s definitions are not trivially translated into a relational model
FP7 MYHEALTHAVATAR & H2020 IMANAGECANCER (2013-2018)

- MyHealthAvatar (http://www.myhealthavatar.eu/)
  - an attempt at a proof of concept for the digital representation of patient health status
  - a lifetime companion for individual citizens that will facilitate the collection of, and access to, long-term health-status information

- iManageCancer (http://imanagecancer.eu/)
  - provide a cancer specific self-management platform
  - focusing on the wellbeing of the cancer patient
  - special emphasis on avoiding, early detecting and managing adverse events of cancer therapy but also, importantly, on the psycho-emotional evaluation and self-motivated goals
SO MANY ONTOLOGIES, WHAT IF SOMETHING CHANGES?

Approach
1. Reuse Past Mappings
2. Accept queries in new ontology version
3. Identify Ontology Changes
4. Rewrite queries from new ontology version to the past ontology version

SIGMOD2011
SWJ2013
ER2014
PROPOSALS UNDER DEVELOPMENT

• Responsible clinical trial data sharing is in the public interest to advance science
  • Many CTs now not analyzed and published in a timely manner
    • 1/3 of trials, results not published after 4 years
• Already a momentum for data sharing
• Question is not whether to share, but what types of clinical trial data, when, & how to share

Sharing Clinical Trial Data
MAXIMIZING BENEFITS, MINIMIZING RISK
LESSONS LEARNT
FALLACIES

• Your data are “big data”
• Big data introduces no fundamental problems
• Big Data = Map Reduce (Hadoop)
• Big Data = Data Quantity (Scalability)
LESSONS - ARCHITECTURE

• Large technical collaborations can only succeed when sufficient attention is given to using as much as possible industry standards.
LESSONS – ONTOLOGIES/STANDARDS

• Lack of coordination between standards/terminologies/ontologies.
  • Semantic inconsistencies between them.

• Multiple Ontologies/Terminologies for Clinical Data Management are needed.
  • No single ontology to rule them all.
  • The community can benefit from guidance on vocabularies to represent data and an integrated library with the recommended ontologies.

• Challenging to fill the gaps between generality of the ontologies versus the specificity of a (legacy) database that we need to integrate
LESSONS – DATA INTEGRATION

• Data Integration is hard.

• Big Data Integration is harder
  • No single solution exists.
  • We like experimenting (in many cases “try and cry”)

• High value in defining common methodologies concerning
  • data management
  • build and share comprehensive datasets
MORE LESSONS

• 80% of data in textual format

• The study of data quality is as important as data quantity
  • (10%-75% error rates in telecommunication)
  • “information perceived as being needed for clinical decisions was unavailable 13.6%--81% of the time” (2005)

• Tools should be available enabling data source exploration

• Methods for analyzing and summarizing data are equally important

• Data and their schemas are not static artifacts but subject to evolution
EVEN MORE LESSONS

• GUI and usability aspects equally important as functionality

• In many cases technical problems are less important than the legal and economic issues.
  • Most of the data are proprietary. Getting approval from the legal department can be challenging.
  • The right of the patient to own his data is crucial.
  • Compliance to trial related legislation, especially to the data protection laws, is a critical success factor for any research-network.
REFERENCES


The ACGT project in retrospect: Lessons learned and future outlook. ICCS 2011


REFERENCES


ARE YOU SURE THIS IS HOW WE GET DATA INTO THE CLOUD?

LET’S SOLVE THIS PROBLEM BY USING THE BIG DATA NONE OF US HAVE THE SLIGHTEST IDEA WHAT TO DO WITH

WE’VE DECIDED TO TAKE BIG DATA TO THE NEXT LEVEL...

HUMONGOUS DATA

© D.Fletcher for CloudTweaks.com
LINKED DATA BENCHMARK COUNCIL (LDBC)

- Objectives:
  - Produce an initial set of *industrial benchmarks* for *Graph* and *RDF* management technologies
  - Establish the LDBC Non Profit organization to survive the project that fosters the *creation* and *maintenance* of industrial-strength benchmarks
HOBBIT: HOLISTIC BENCHMARKING OF BIG LINKED DATA

• **Scope:**
  • Focus on *Big Linked Data*
  • Cover the *business-critical steps of the Linked Data lifecycle*
HOBBIT: HOLISTIC BENCHMARKING OF BIG LINKED DATA

• **Scope:**
  - Focus on *Big Linked Data*
  - Cover the *business-critical steps of the* *Linked Data lifecycle*

• **Objectives:**
  - Gather *real requirements from industry*
    - *Performance indicators and thresholds*
  - Develop *benchmarks based on real data*
  - Provide *universal benchmarking platform*
    - *Standardized hardware*
    - *Comparable results*
VRE to Empower multi-disciplinary research communities and accelerate Innovation and Collaboration.