Diagnostic hypothesis refinement in reproducible workflows for advanced medical data analysis

Cezary Mazurek, Raul Palma, Juliusz Pukacki
Poznań Supercomputing and Networking Center

Scientific workshop. Big Data: processing and exploration, 22.04.2016, Poznań,

Workflows

- The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules.
  (From The Workflow Management Coalition Specification)

- Workflows serve a dual function *):
  - first as detailed documentation of the method (i. e. the input sources and processing steps taken for the derivation of a certain data item)
  - second as re-usable, executable artifacts for data-intensive analysis.

- Workflows stitch together a variety of data manipulation activities such as data movement, data transformation or data visualization to serve the goals of the scientific study*).

*) D.Garijo,P.Alper,K.Belhajjame,O.Corcho,Y.Gil,C.Goble,Common motifs in scientific workflows: an empirical analysis, Future Gener. Comput. Syst.(2014) http://dx.doi.org/10.1016/j.future.2013.09.018.

Scientific workflows

Becoming widely used in many fields

- Coordinate **execution** of *services* and linked *resources*
- **Dataflow** between services
  - Web services (SOAP, REST)
  - Command line tools
  - Scripts
  - User interactions
  - Components (*nested workflows*)
- **Method** becomes:
  - **Documented** visually
  - **Shareable** as single definition
  - **Reusable** with new inputs
  - **Repurposable** other services
  - **Reproducible**?

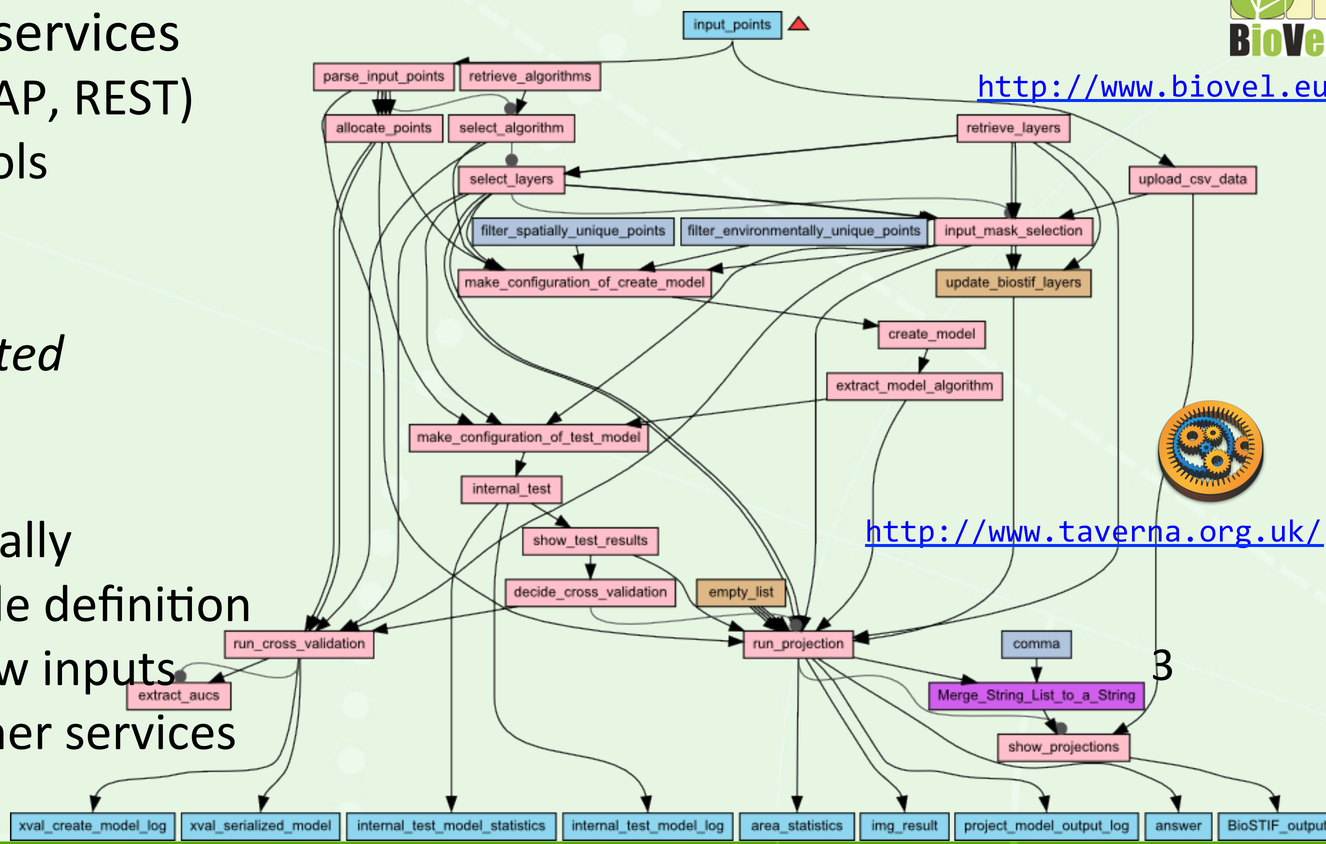
myexperiment

http://www.myexperiment.org/workflows/3355

BioVeL

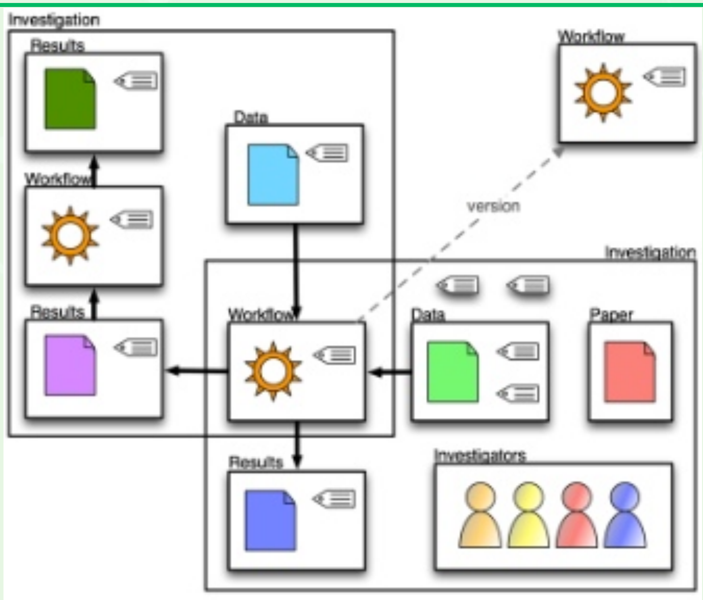
http://www.biovel.eu/

http://www.taverna.org.uk/



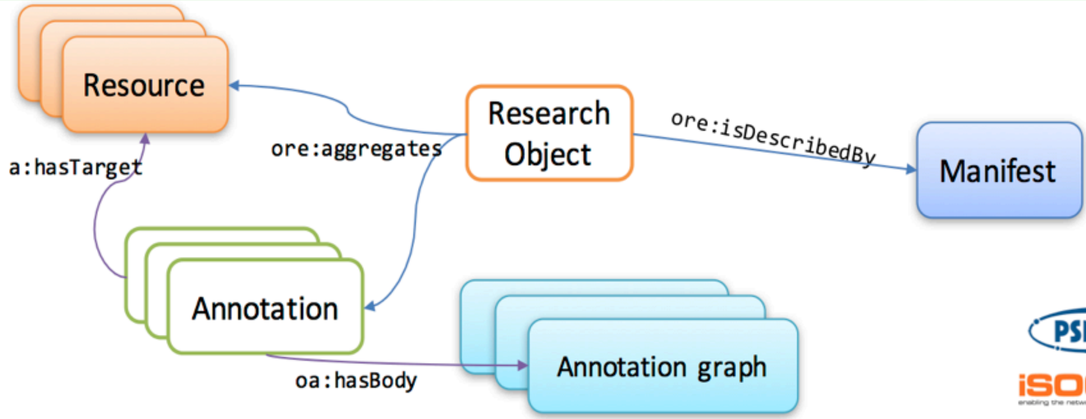
Research objects

- Semantic **aggregations** of related scientific resources, their **annotations** and research **context**.
- Enable referring a bundle of research artifacts supporting an **investigation**
- Provide mechanisms to **associate** human and machine-readable **metadata** to these artifacts.
- **RO model** enables to capture and describe these objects, their provenance and lifecycle
  - Ontology network (based on OAI-ORE, OA, PROV-O)



ro (aggregation and annotation)
wfdesc (workflow description)
wfprov (workflow provenance)
roevo (evolution model)
minim (minimum information model)

RO primer: http://wf4ever.github.com/ro-primer
RO specification: http://wf4ever.github.com/ro



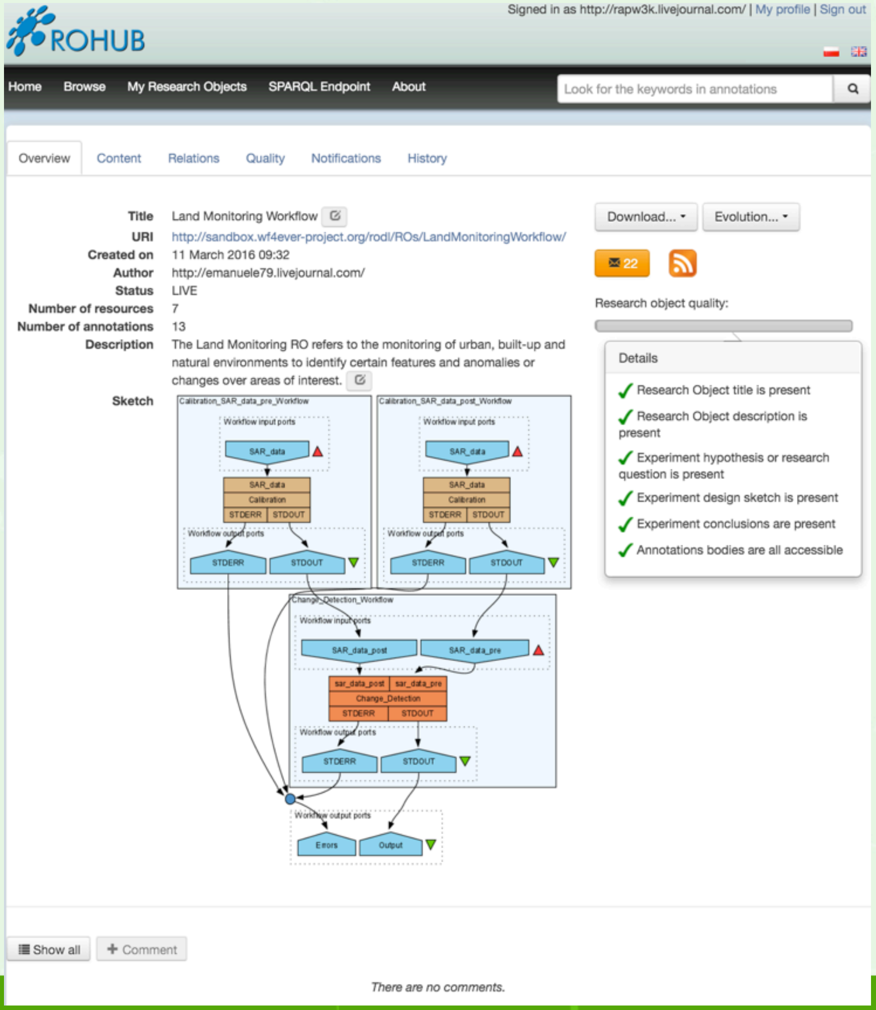
ROHub (http://www.rohub.org)

RO storage, lifecycle management and preservation

- Enables the **sharing** of scientific findings

- Support scientists throughout the **research lifecycle** to create and maintain high-quality ROs that can be interpreted and reproduced in the future.

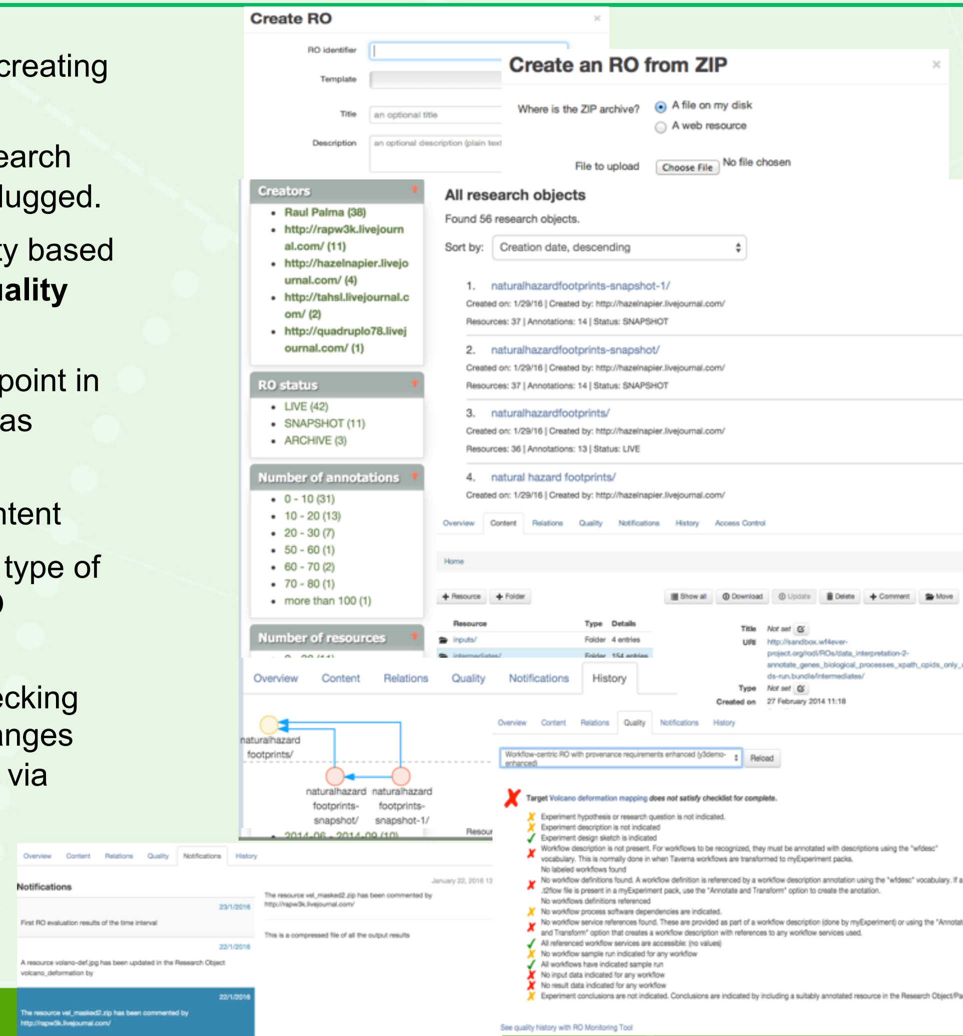
- Combination of digital libraries, long term-preservation and semantic technologies.



ROHub (http://www.rohub.org)

RO storage, lifecycle management and preservation

- **Create, manage and share ROs**: different methods for creating ROs and different **access modes** to share them
- **Finding ROs**: a **faceted search** interface, a **keyword** search box, and other interfaces as the collab spheres can be plugged.
- **Assessing RO quality**: a **progress bar** of the RO quality based on set of predefined basic RO requirements. Detailed **quality information**
- **Managing RO evolution**: create **RO snapshots** at any point in time, release and **preserve** the RO when the research has concluded. Visualize the **evolution of the RO**
- **RO Inspection:** Navigation panel to traverse the RO content
- **External resources and workflow run**: **aggregate** any type of resource, including **links to external resources** and **RO bundles** (ZIP serialization)
- **Monitoring ROs**: monitoring features, such as fixity checking and **RO quality**, which generate **notifications** when changes are detected. Visualize those notifications and subscribe via **atom feed**.



Reproducibility

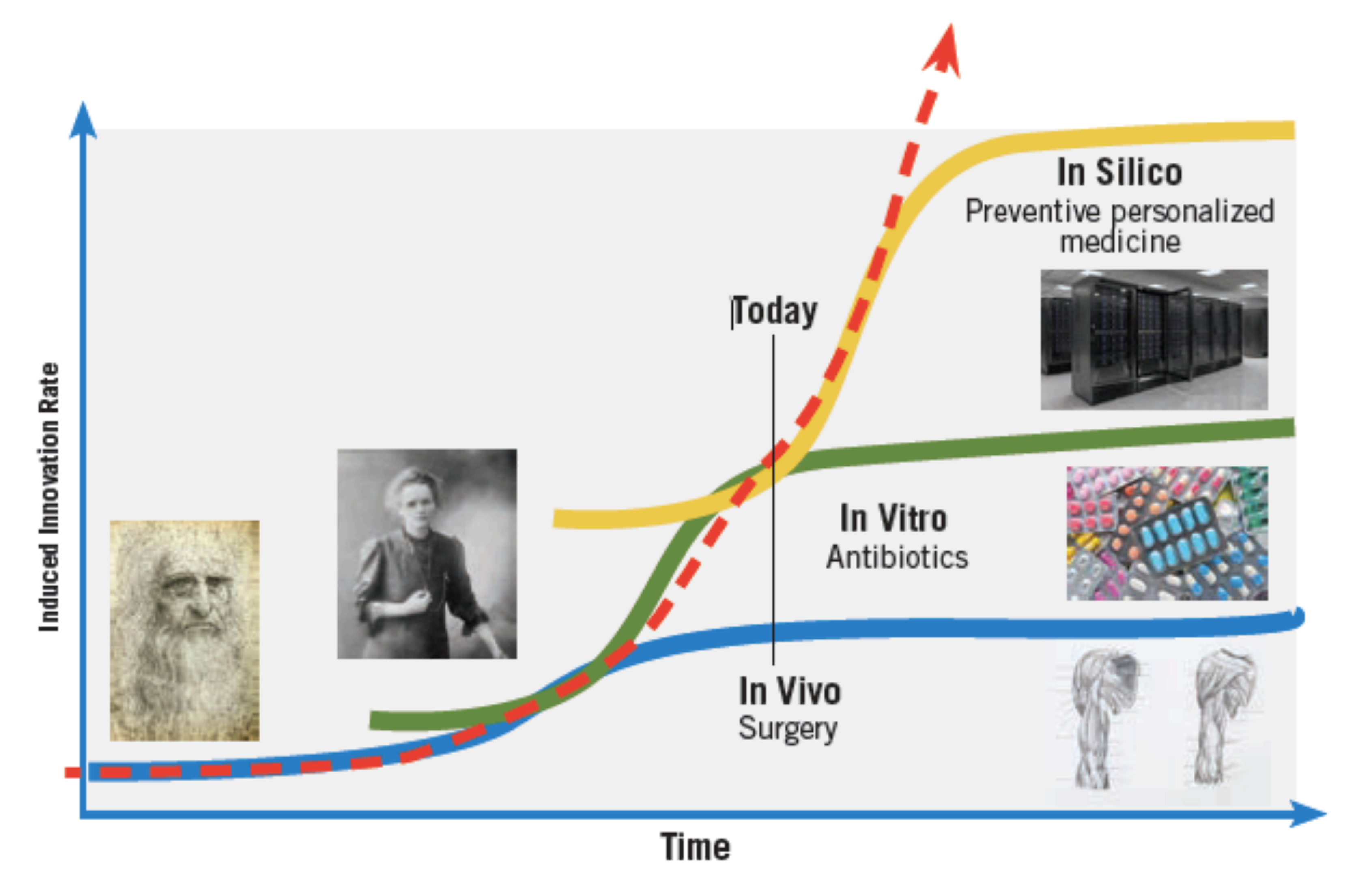
Challenge

Reproducibility for computational experiments is challenging.

It is hard both for authors to derive a compendium that encapsulates all the components (e.g., data, code, parameter settings, environment) needed to reproduce a result, and for reviewers to verify the results.

There are also other barriers, from practical issues – including the use of proprietary data, software and specialized hardware, to social – for example, the lack of incentives for authors to spend the extra time making their experiments reproducible.

Big Data *Surfing*



T.Marschal: In Vivo, in vitro, in Silico!. ANSYS Advantage, vol. IX, Issue 1, 2015

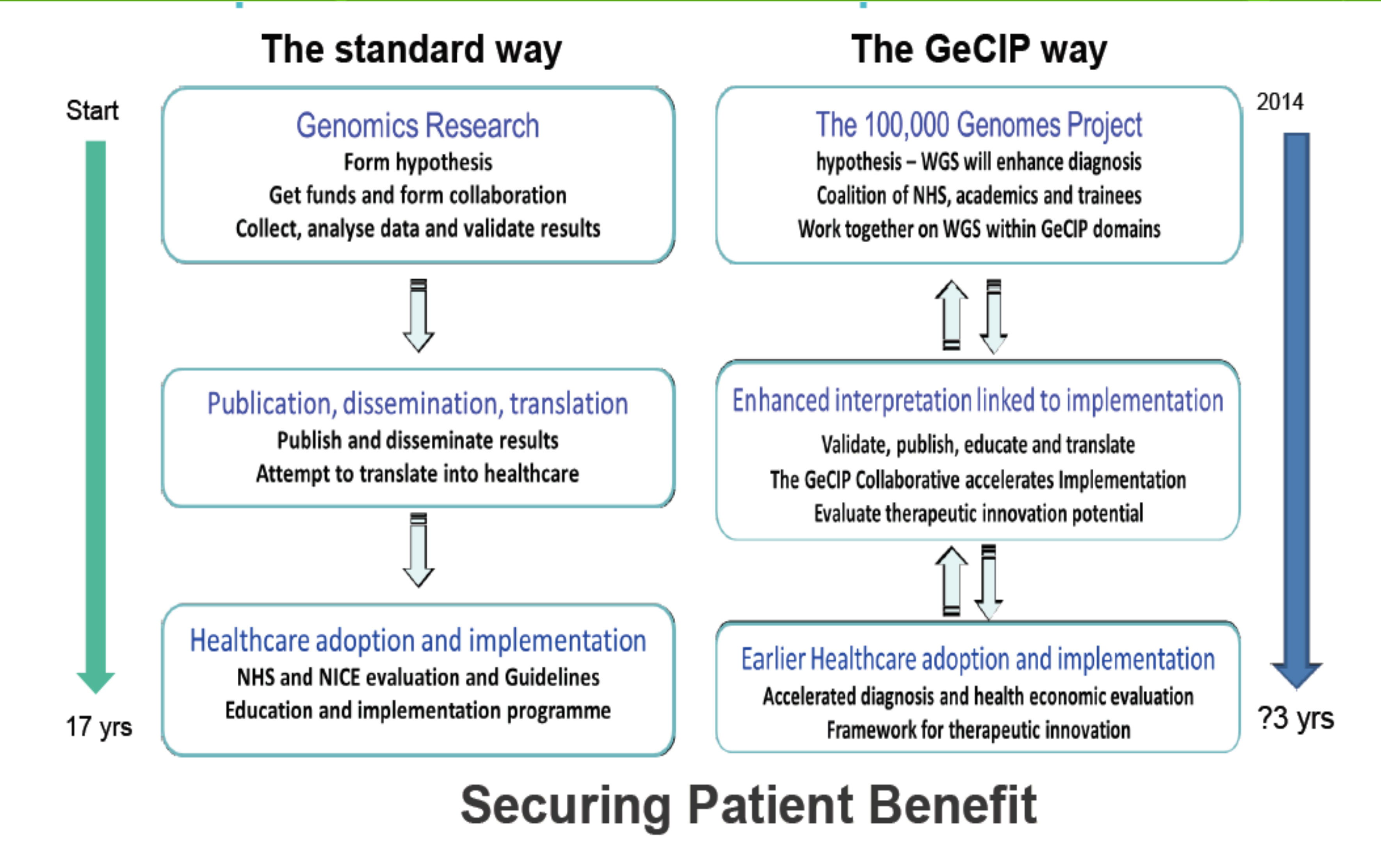
Problem/Challenge

- Historically, the scientific method is well known and was introduced by Louis Pasteur in XIX century.
- This method is in fact a cycle of following steps:
  - Observations->Questions->Hypotheses-> Predictions>Experiment (incl. refinement) -> Discussion.
- These steps allowed for many years to report scientific experiments conducted In-Vivo and In-Vitro.
- However we think that even if steps are still the same while performing in-Silico experiments, the way of reporting them need to be changed, especially in fields where part of experiment is creation of software tools

What it means?

- Smart data processing and experients but….
- What data means for doctors?
  - They need treatment instructions and its expected results
- We need new environment for *in-silico* disease hypothesis refinement and building decision support systems

This is a challenge for researchers in interdisciplinary teams



Prof Mark Caulfield FMedSci, Genomics England Clinical Interpretation Partnership

Are the answers obvious?

Are **the questions** obvious?

Towards precision (personal) medicine

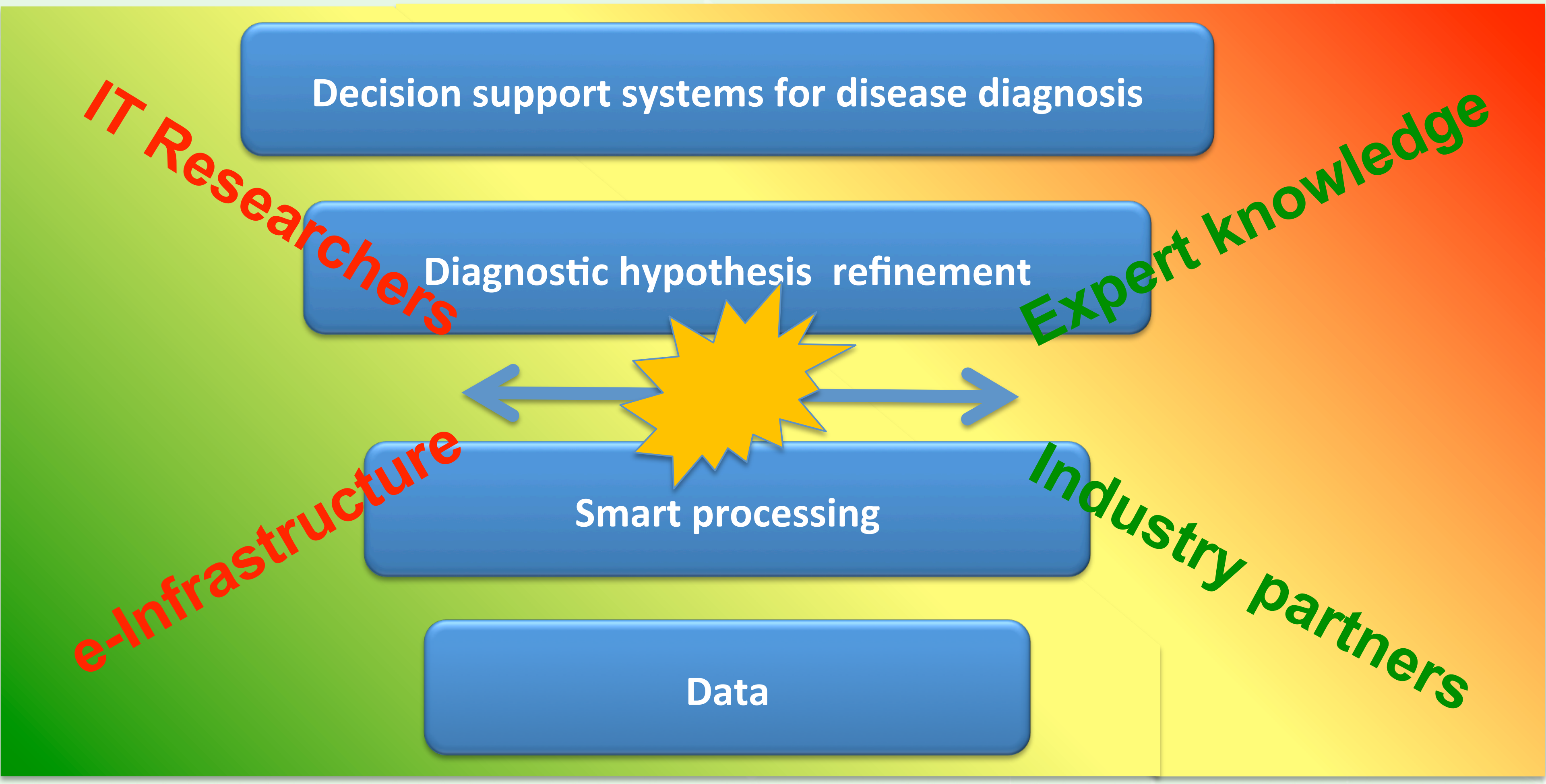
- Questions-driven (smart) data experiments
- If failed – lead to other questions and experiments

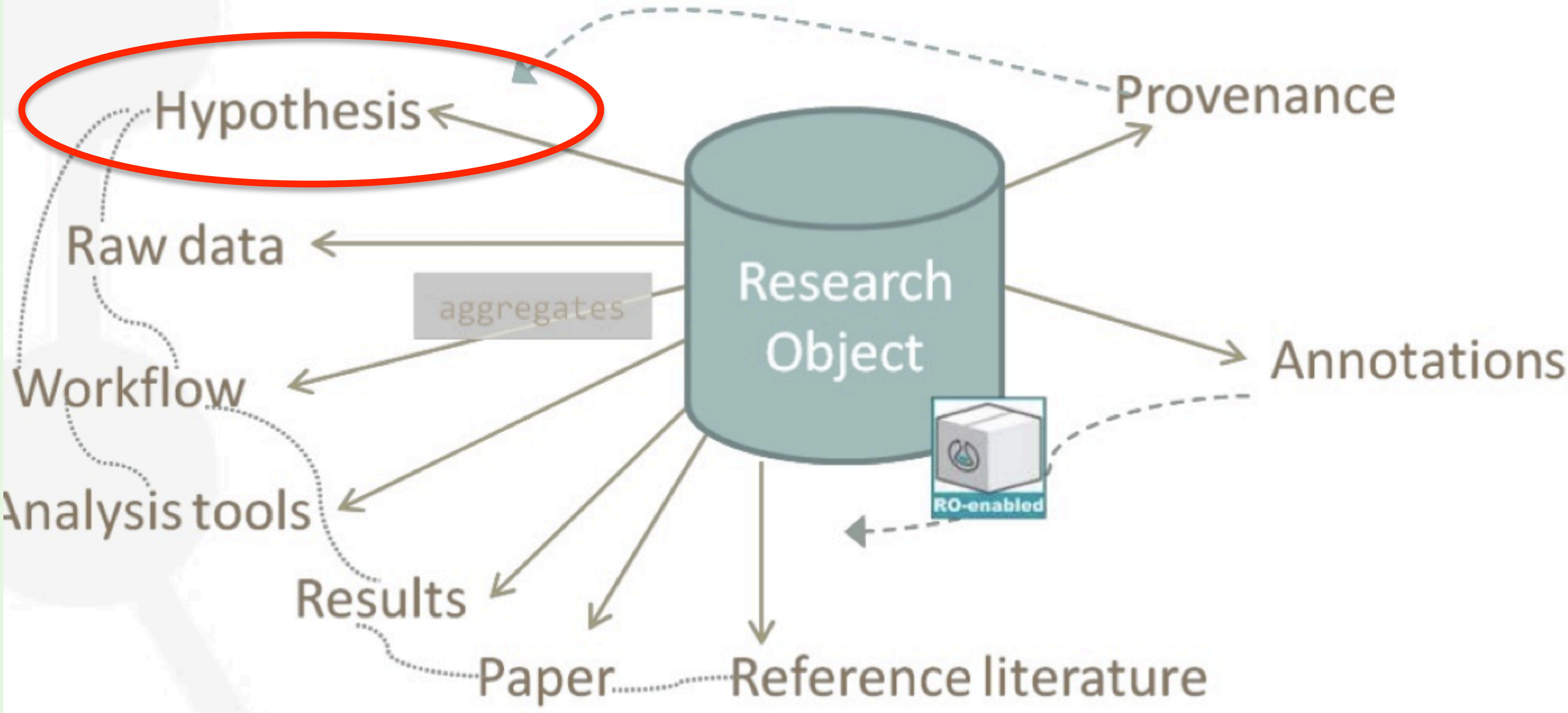
- So…do not start from transfering existing knowledge and statistical approach to data space

- We need to start thinking from like being lived in data space and create experiments to quickly verify hypothesis (diagnostic hypothesis refinement)

- Precision medicine makes it even more challenging!!!
  - Data experiments are being defined for individual patient and route to personal treatment

Disruptive Innovation in Interdisciplinary Teams





ROs as web pages http://rohub.linkeddata.es/

ROs as part of a Linked Data Platform (alpha): http://purl.org/net/ldp4ro

Hypothesis refinement

Practical cases

- In-Silico experiments, especially in their **refinement cycle**, lead to creation of new software tools, algorithms and even computer science challenges. To make this experiment valuable such a process needs to be **controlled and recorded** while achieving milestone stages;
- Scientific experiments are performed in cycles, when each cycle is a refinement of the **hypothesis**. Continuing research starting from any cycle and branching this process further on, require that each cycle is checkpointed and stored as a scientific procedure step;
- Medical research reliant on data analysis, focused on early disease diagnosis or stopping the disease progress, very often results in providing **software tools** helping in data analysis and created during the experimentation cycles.
- To treat the process of knowledge discovery based on data analysis and development of processing tools, as a research method, we need to provide the way of formal description of **stages of such a process**, be paired with hypothesis refinement stages.

PSNC POZNAŃ SUPERCO

Domain examples

- Bioinformatics
  - *omics research
- Earth Science (EVEREST)
  - European Virtual Environment for Research - Earth Science Themes: a solution
- Cardiac rehabilitation and early risk identification of cardiovasular diseases
  - Personal prevention plan
- Glaucoma diagnosis and early prevention

Glaucoma research experiment

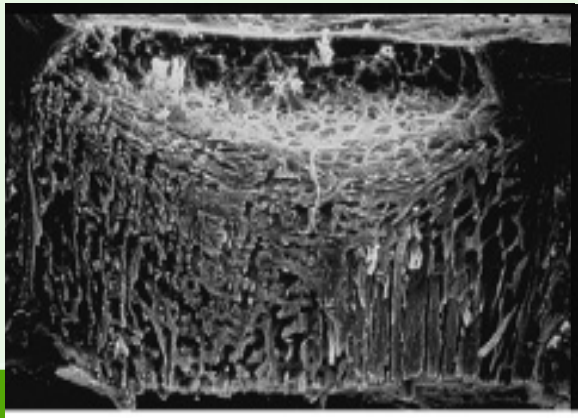
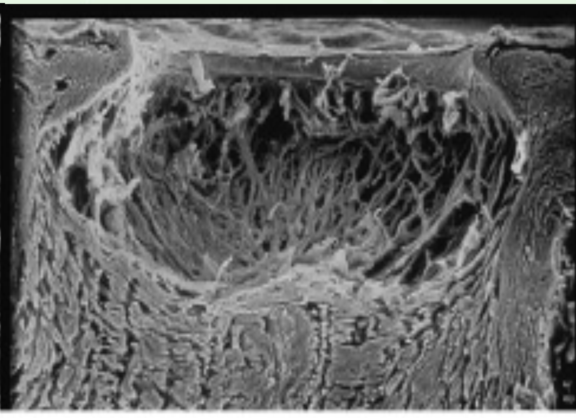
Glaucoma - group of progressive optic nerve neuropaties releted with:

a) accelerated apoptosis of Retinal Ganglion Cells due to neurotrophic deprivation

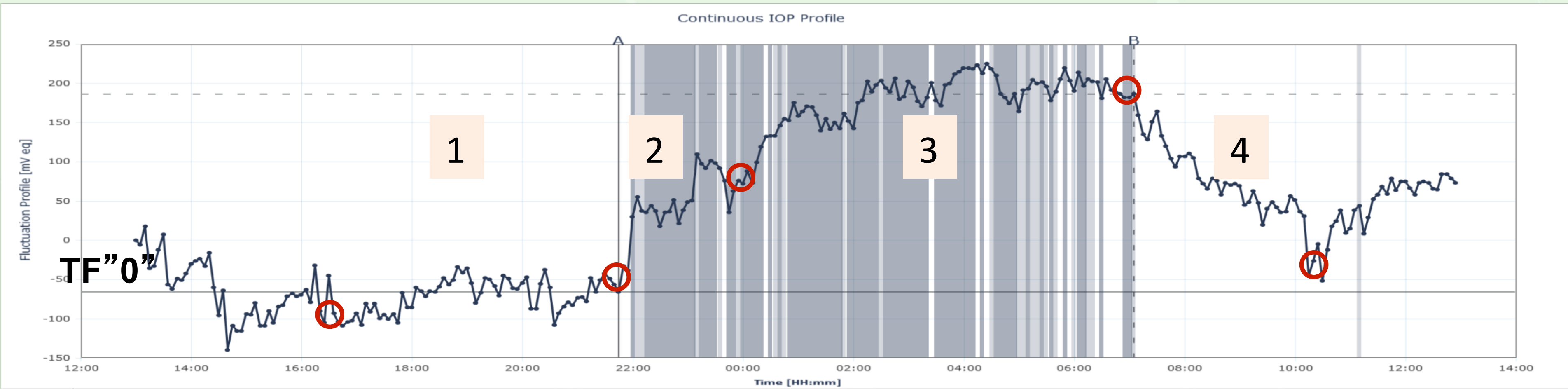
[Band L.R., 2009; Balaratnasingam C., 2008; Fechtner R.D., Weinreb R.N., 1994; Garcia-Valenzuela E., 1995; Quigley H.A., 1976, 1995, 2000; Yablonski M., Asamoto A., 1993]

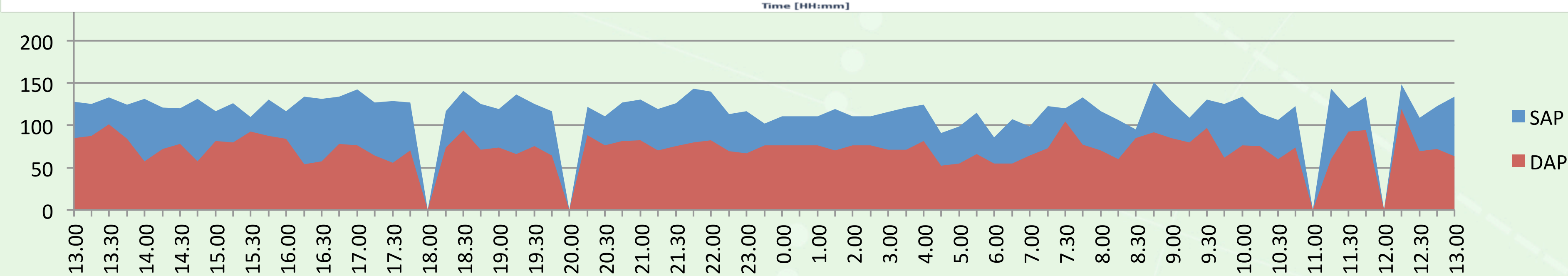
b) lamina cribrosa sclerae pathognomonic phenotype changes

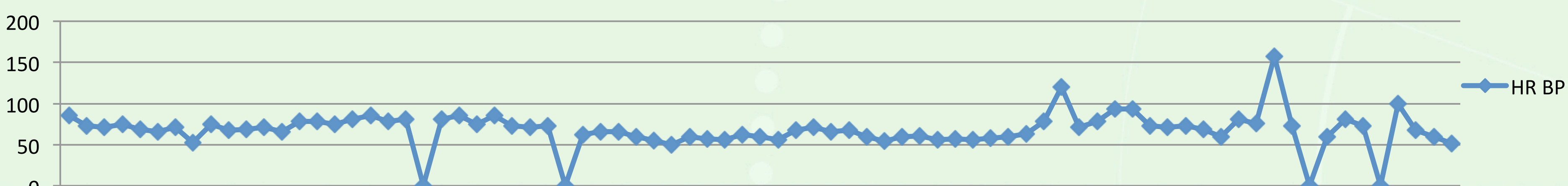
[Ernest J.T. and Potts A.M., 1968; Quigley H.A., 1983; Roberts M.D., 2009].

1. GENERAL ANALYSIS - AREA UNDER CURVE (AUC) 24h
2. TIME-INTERVAL DEPENDENT ANALYSIS (Linear Model α & β)







Checkpoint

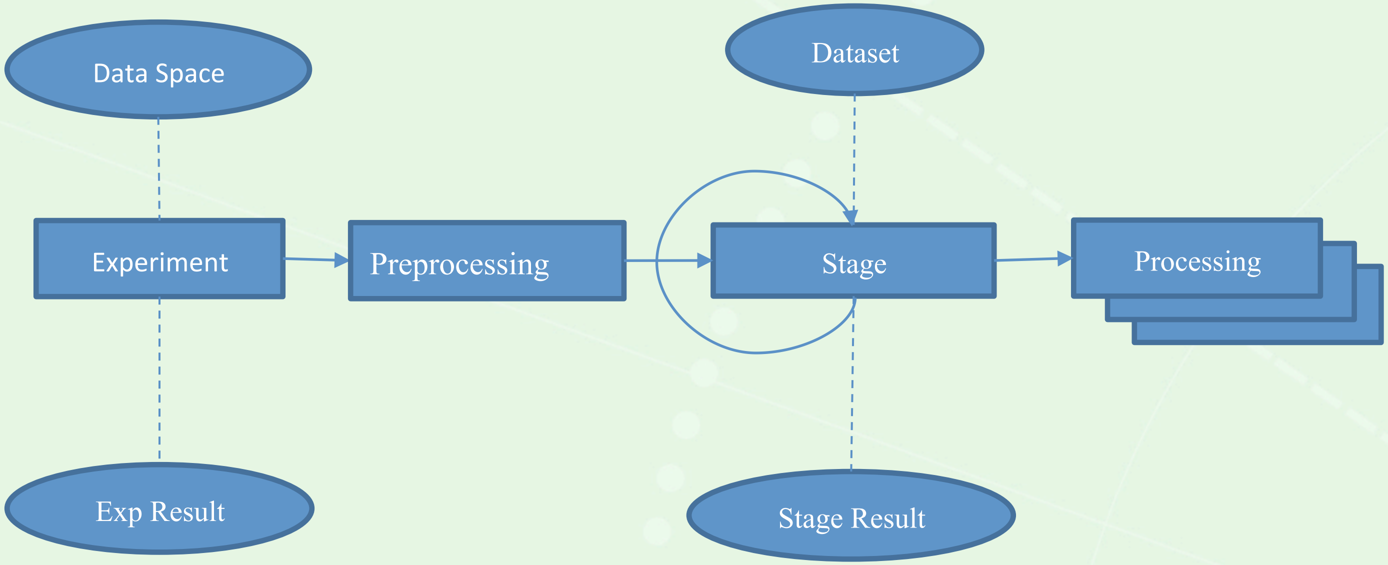
VC-DomLEM is a sequential covering algorithm inducing strong decision rules satisfying constraints on consistency.

J. Błaszczyński, R. Słowiński, M. Szeląg, *Sequential Covering Rule Induction Algorithm for Variable Consistency Rough Set Approaches*, Information Sciences (2011), 181, pp. 987-1002

- 280 rules assigned into 50 classifiers (role of Experts)
- Classifiers Voting (round table) decide of diagnosis
- Rules indicated by algorithm in diagnosis pointed at specific place of pathology in checked system?

Decision Rule Models in Differentiation of Healthy and Glaucomatous Patients" R. Wasilewicz; P. Wasilewicz; A. Radziemski, J. Błaszczyński, C. Mazurek; R. Słowinski, Cardiovascular Mobile Health Conference 2015, Tabarz, Germany

Hypothesis


Data Space
Dataset
Experiment
Preprocessing
Stage
Processing
Exp Result
Stage Result

International Consortium for Technology in Biomedicine

ICTBioMed

International Consortium

Open Health System Laboratory, USA

Centre for Development of Advanced Computing, India

University of Notre Dame, USA

Chalmers Unviersity of Technology, Sweden

Internet2, USA

Poznań Supercomputing and Networking Center, Poland

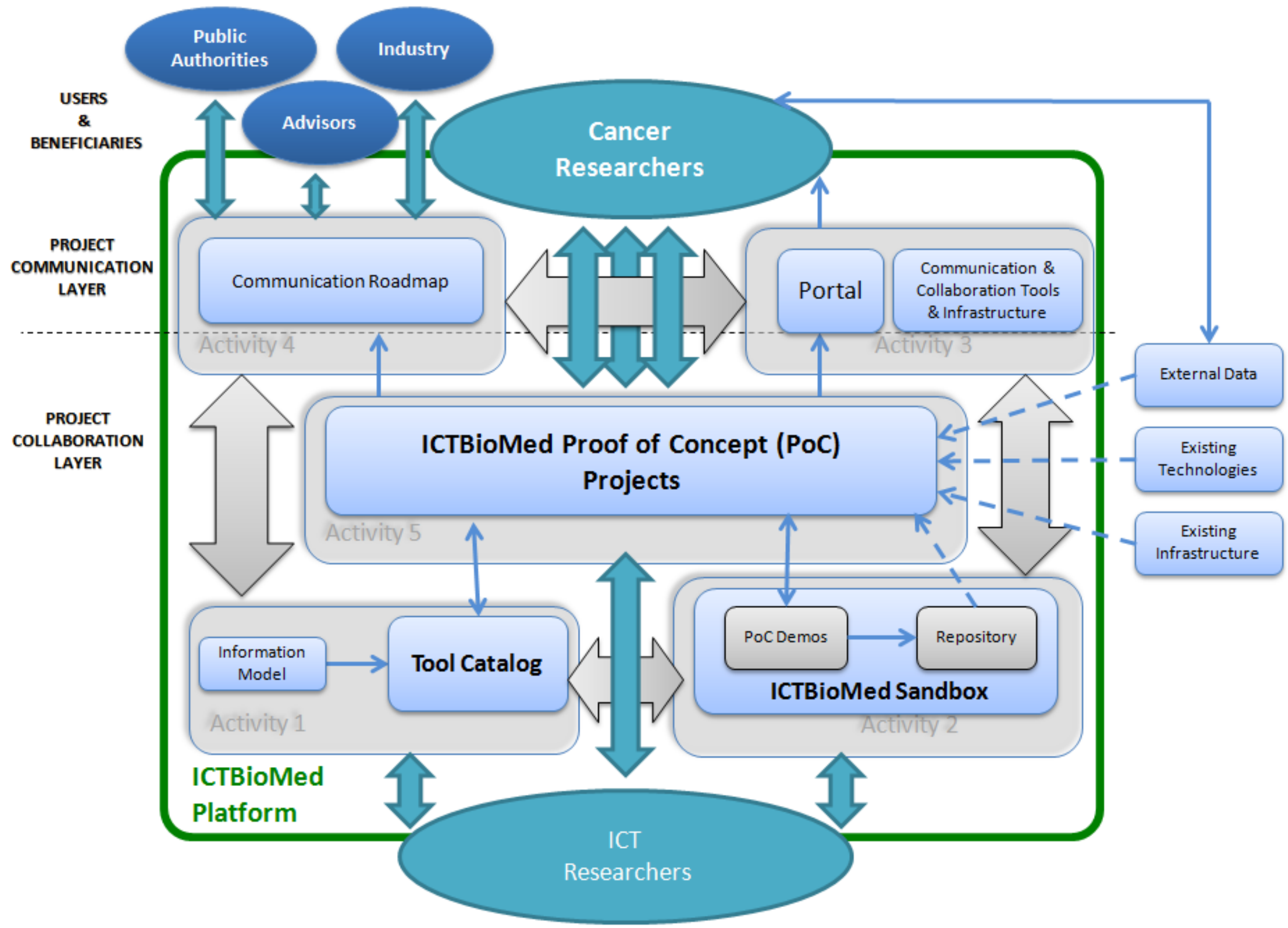
In collaboration with:

Duke University — Applied Therapeutics Section, USA

Indian Institute of Technology, Dehli, India

International Consortium for Technology in Biomedicine
ICTBioMed

International collaboration for biomedicine



Applications (some examples)

CDAC

Biomolecular Simulations and molecular docking: Research on cancer proteins, antisense molecules, GPCRs

Next Generation Sequencing Data Analysis: Applications in cancer genomics (Breast Cancer transcriptome)

High throughput comparative genomics studies on salmonella and mycobacterium

Chalmers

Chalmers Life Science and Engineering: Europe's leading center for Metabolic Engineering and Systems Biology (Jens Nielsen Lab)

Gothenburg University (Molecular Biology, Europe's leading Center for Systems Biology,NGS)

Sahlgrenska University Hospital and Academy (Centers for Cancer and Cardiovascular and Metabolic Diseases)

Biotech Industries: AstraZeneca worldwide research and innovation hub.

PSNC

Support for complex eScience research tasks in the area of post-genomic clinical trials and virtual physical human modeling for clinical purposes: ACGT and P-Medicine projects

RNASeq analysis (role of proteins and retroelements in induced pluripotent stem cells)

Breast cancer therapy (novel biomarkers) and diagnostics (applying TCGA data)

Interactive visualization of correlations between genomic analysis observations

Pilot workflow integration with UT MD Anderson Cancer Center

GEN Exclusive

GEN Exclusive

More

Sep 15, 2015 (Vol. 35, No. 16)

Industry, Academia Reconfigure Ties

Time for a Radical Change in Collaborative Approaches to Healthcare Research

Stephen K. Klasko, M.D., MBA

It's time to break the mold. We need new models for collaboration between the health research industry and academia.

In the face of unprecedented generation of knowledge, we need to rethink how we speed discovery to patients. But we have an equally great mandate—to rethink how we select, train, and grow the next generation of health professionals and researchers who will create meaning from all this data.

When we hosted BIO2015 in Philadelphia this summer, it became even clearer to me that we need transformation of not just American healthcare delivery, but also of fundamental relationships between scientists and educators as they work to improve health.

Click Image To Enlarge +

Stephen K. Klasko, M.D.

We need new models for collaboration between the health research industry and academia.

The only way that will happen is if we can reduce some of the local competition and fragmentation and create super-centers of innovation for:

- *regional consortia for clinical research,*
- *experimental therapeutics centers,*
- *advanced biomanufacturing centers,*
- *centralized repositories for patient data.*

http://leadership.jefferson.edu/blog/

DIAGNOSTIC HYPOTHESIS REFINEMENT IN DATA SPACE

Glaucoma

PERSONALIZED MEDICINE

PERSONALIZED DIAGNOSIS & TREATMENT

- Glaucoma
- Nephroblastoma
- Breast Cancer
- Cardiovascular Diseases (CVD)
- ENT Disorders

TOOLS EXTRACTION

DIAGNOSIS REFINEMENT

DATA ANALYSIS

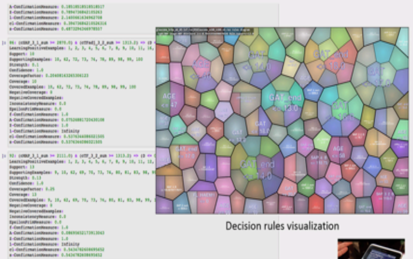
- Semantic Data Integration
- Data Warehousing
- Modeling and Simulations (VPH)
- Clinical Decision Support Systems
- NGS Data Processing Pipelines

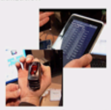
KNOWLEDGE EXTRACTION

DATA SPACE

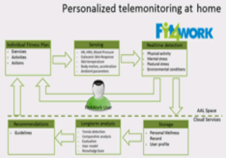
HETEROGENOUS DATA SOURCES

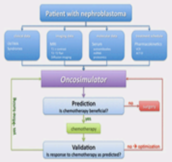
- Imaging Data
- Genomic Data
- Phenotypic Data
- Environmental Data
- Lifestyle Data

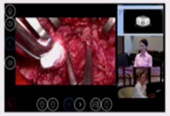

Decision rules visualization


Personalized telemonitoring at home

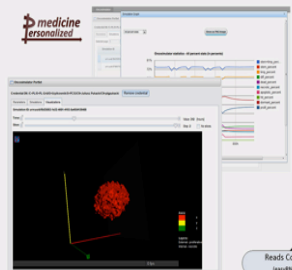

Personalized prevention for older adults

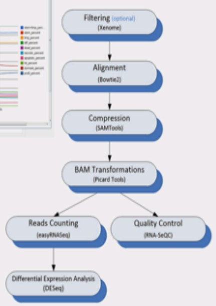

Personalized prevention for ageing workers


VPH model as CDSS tool

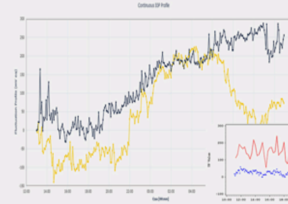

Intraoperative surgery teleconsultation

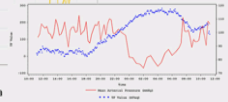

Notebook based working space


Cancer model simulation visualization

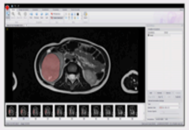


RNAseq computation workflow


Eye volumetry data


Cardiovascular factors


Registry of major salivary glands benign tumours


Naproblastoma images repository


Registry of trauma cases in the Wielkoposka Center of Telemedicine

PARTNERS

Poznan University of Medical Sciences
prof. Krzysztof Słowiński
prof. Witold Szyfter

Poznan University of Technology
prof. Roman Słowiński
dr. Jerzy Błaszczyński

[W] eye clinic

W-Eyeclinic, Poznan
dr. Robert Wasilewicz

UNIVERSITÄT DES SAARLANDES

Saarland University Hospital
prof. Norbert Graf

Institute of Communications and Computer Systems, Athens
dr. Georgios Stamatakos

CONTACT >>>>>>>>>>

POZNAN SUPERCOMPUTING AND NETW

Mazurek Cezary phone: (+48 61) 858 2030, +48

Publications

- R.Wasilewicz, P.Wasilewicz, E.Czaplicka, J.Kocieckil, J.Blaszczynski, C.Mazurek and R.Slowinski: 24 hour continuous ocular tonography Triggerfish and biorhythms of the cardiovascular system functional parameters in healthy and glaucoma populations. Acta Ophthalmologica, 91: 0. doi: 10.1111/j.1755-3768.2013.2721.x
- Palma R., Corcho O., Hołubowicz P., Pérez S., Page K., Mazurek C., Digital libraries for the preservation of research methods and associated artefacts. Proc. 1st International Workshop on the Digital Preservation of Research Methods and Artefacts (DPRMA 2013) at Joint Conference on Digital Libraries (JCDL 2013). pp. 8-15. Indianapolis, Indiana, USA, July 2013
- Mazurek, C., Pukacki, J., Kosiedowski, M., Trocha, S., Darbari, H., Saxena, A., Joshi, R., Brenner, P., Gesing, S., Nabrzyski, J., Sullivan, M., Dubhashi, D., Thankaswamy, S., and Srivastava, A. (2014) Federated Clouds for Biomedical Research: Integrating OpenStack for ICTBioMed. Cloud Networking (CloudNet), 2014 IEEE 3rd International Conference on, pp.294-299, 8-10 Oct. 2014, doi: 10.1109/CloudNet.2014.6969011
- Palma R., Corcho O., Gómez-Pérez J.M., Mazurek, C., “ROHub A Digital Library of Research Objects Supporting Scientists Towards Reproducible Science”. In Semantic Publishing Challenge of Proc. Extended Semantic Web Conference (ESWC), Crete, Greece, May 25-29, 2014.
- M.Krysinski, M.Krystek, C.Mazurek, J.Pukacki, P.Spychala, M.Stroinski, J.Weglarz. Semantic Data Sharing and Presentation in Integrated Knowledge System. [In:] R. Bembenik, Ł. Skonieczny, H. Rybiński, M. Kryszkiewicz, & M. Niezgódka (Eds.), Intelligent Tools for Building a Scientific Information Platform: Advanced Architectures and Solutions, pp. 67–83. Springer International Publishing 2013
- J.Andersen, P.Shah, K.Korski, M.Ibbs, V.Filas, M.Kosiedowski, J.Pukacki, C.Mazurek, Y.Wu, E.Chang, C.Toniatti, G.Draetta, M.Wiznerowicz: Applying TCGA data for breast cancer diagnostics and pathway analysis, Cancer Research 10/2014; 74(19 Supplement):4272-4272
- J.Pukacki, H.Świerczyński, C.Mazurek, M.Kosiedowski "RNA-Seq data analysis pipeline in Poznan Supercomputing and Networking Center", 1st Congress of the Polish Biochemistry, Cell Biology, Biophysics and Bioinformatics, September 2014, Warsaw, Poland
- M. Kosiedowski, C. Mazurek, K. Słowiński, M. Stroiński, K. Szymański, J. Węglarz: „Telemedical systems for the support of regional Healthcare In the area of trauma” , Global Telemedicine and Health Updates: Knowledge Resources, vol. 3 str. 592 – 596, 2010

PSNC

Poznań Supercomputing and Networking Center

affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences,

ul. Noskowskiego 12/14, 61-704 Poznań, POLAND,
Office: phone center: (+48 61) 858-20-00, fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.psnc.pl