The name is Science!
d-Science!

Yannis Ioannidis

MaDgIK Lab

Dept. of Informatics & Telecom, Univ. of Athens

ATHENA Research Center

MADGIK

Ερευνητικό Κέντρο Αθηνά

Ερευνητικό Κέντρο Καινοτομίας στις Τεχνολογίες της Πληροφορίας, των Επικοινωνιών, της Γνώσης

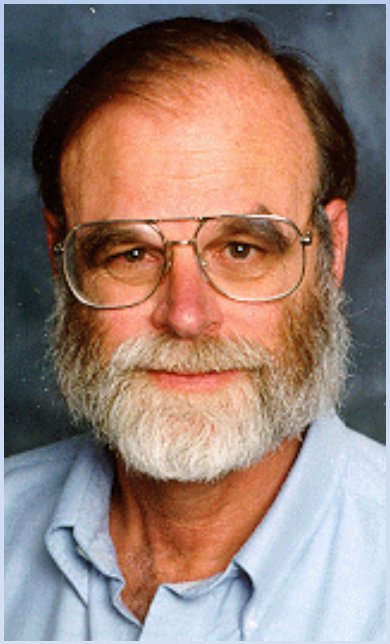
National and Kapodistrian UNIVERSITY OF ATHENS

DEPT. OF INFORMATICS & TELECOMMUNICATIONS

Dedication / Credits

Jim Gray (+)

Missing at sea since 28 January 2007



Turing Award 1999

Sigmod Innovations Award

US National Academy of Engineering

…

MADIK

Science Paradigms

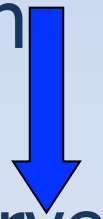
- 1st - Thousand years ago: science was empirical
  - describing natural phenomena w/ some models, generalizations

- 2nd - Last few hundred years: theoretical branch
  - using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

Really Early Times

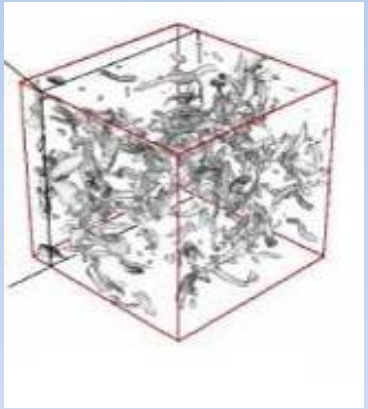
- One scientist
- One location
- One discipline
- One phenomenon



- One pencil (… carver …)
- One paper (… stone …)
- Street announcements, e.g., Εύρηκα!

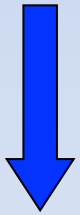
Science Paradigms

- 3rd - Last few decades:
  a computational branch
  simulating complex phenomena



Recent Times

- One small group of scientists
- One location
- One discipline
- One phenomenon



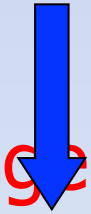
- One file system
- One local disk with custom files
- Publications at refereed forums

Science Paradigms

- 4th - Today:
  data exploration (eScience)
  unify theory, experiment, and simulation

Current Times

- Many/large teams of scientists
- Many locations
- Many disciplines
- Many phenomena



- Many data management systems
- Many data forms
- Web uploads for publications, data, processes, …

MAD IK

~~e~~-Science

*"… computationally-intensive science that is carried out in highly distributed **network** environments,*

or

*science that uses immense **data** sets that require **computing**."*

d-Science

As the **amount**, **variety**, and **complexity** of data increases, it becomes more and more apparent that the construction of well articulated information technology is needed for scientific discovery

nature

www.nature.com/nature

Vol 461 | Issue no. 7261 | 10 September 2009

Data's shameful neglect

Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.

More and more often these days, a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community. Pioneering archives such as GenBank have demonstrated just how powerful such legacy data sets can be for generating new discoveries — especially when data are combined from many laboratories and analysed in ways that the original researchers could not have anticipated.

All but a handful of disciplines still lack the technical, institutional and cultural frameworks required to support such open data access (see pages 168 and 171) — leading to a scandalous shortfall in the sharing of data by researchers (see page 160). This deficiency urgently needs to be addressed by funders, universities and the researchers themselves.

Research funding agencies need to recognize that preservation of and access to digital data are central to their mission, and need to be supported accordingly. Organizations in the United Kingdom, for instance, have made a good start. The Joint Information Systems also the software that will help investigators to do this. One important facet is metadata management software: tools that streamline the tedious process of annotating data with a description of what the bits mean, which instrument collected them, which algorithms have been used to process them and so on — information that is essential if other scientists are to reuse the data effectively.

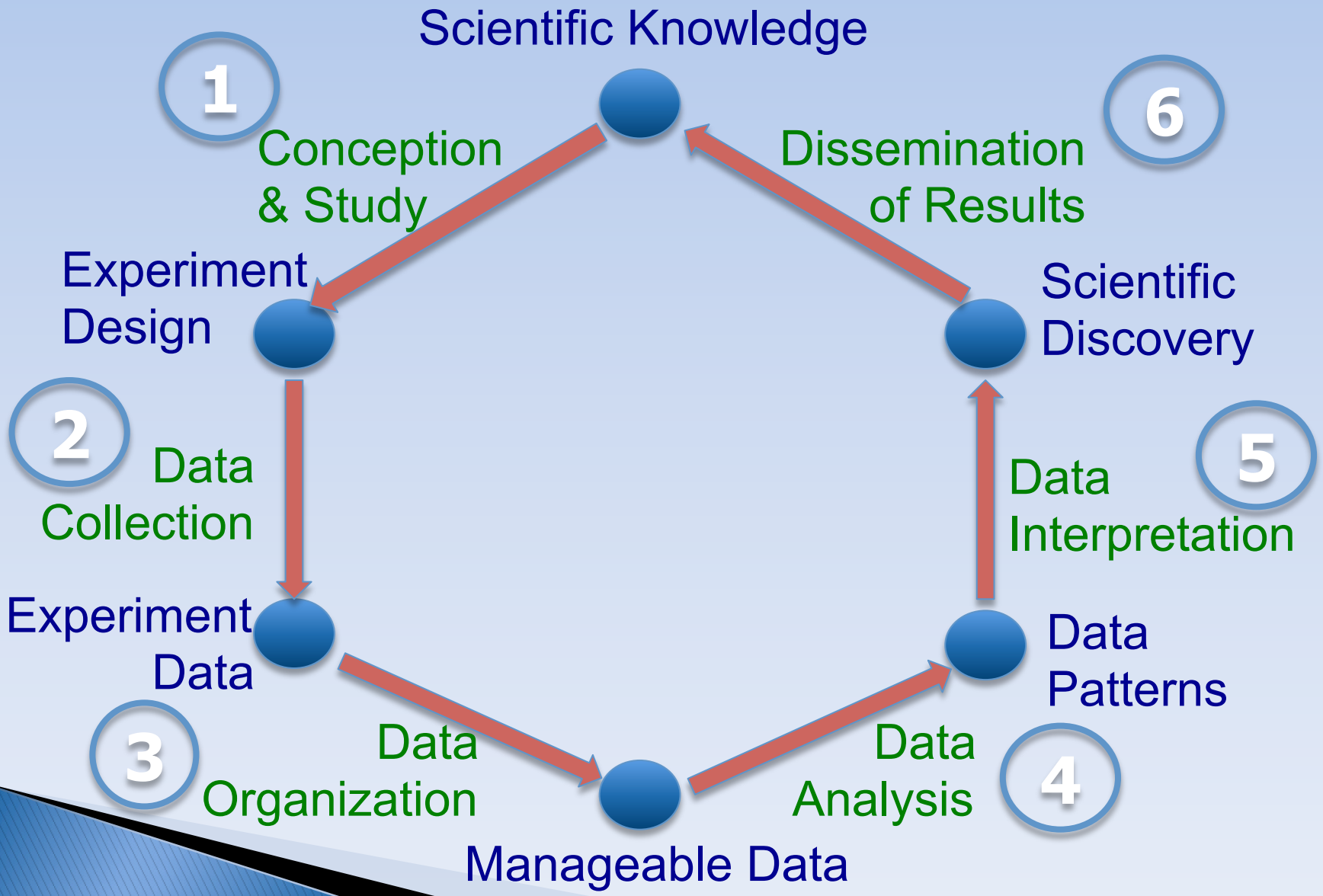
Also necessary, especially in an era when data can be mixed and combined in unanticipated ways, is software that can keep track of which pieces of data came from whom. Such systems are essential if tenure and promotion committees are ever to give credit — as they should — to candidates' track-record of data contribution.

"Data management should be woven into every course in science."

Who should host these data? Agencies and the research community together need to create the digital equivalent of libraries: institutions that can take responsibility for preserving digital data and making them accessible

More and more often these days, a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community. Pioneering archives such as GenBank have demonstrated just how powerful such legacy data sets can be for generating new discoveries — espe-

Research Life Cycle



Scientists' Role

- Early times: scientist at all stages
- Current times:

  Specializations at certain stages

  - Data collection (2) and Result dissemination (6)

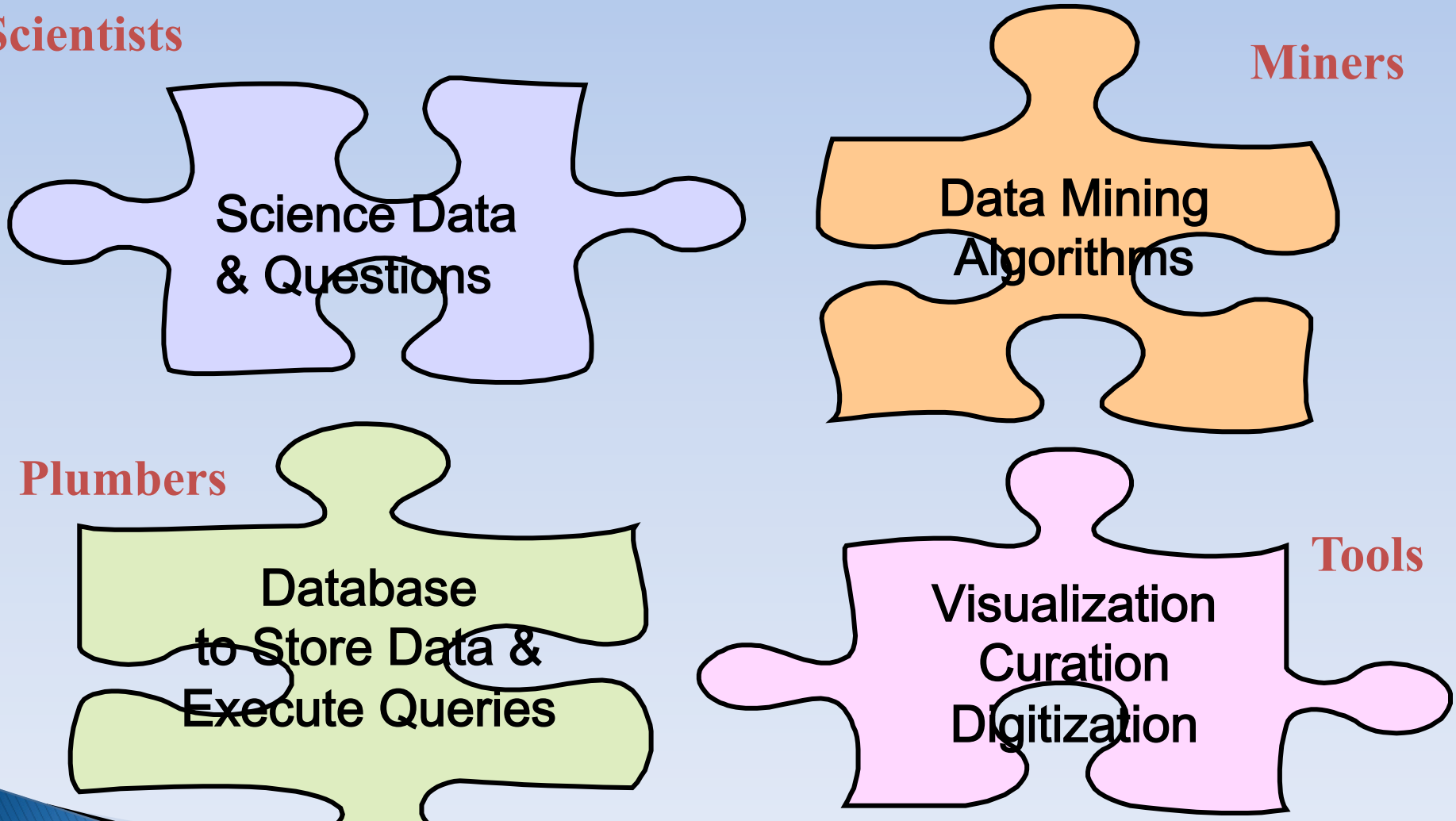
  Scientist at scientific stages

  - Conception and study (1) and Data interpretation (5)

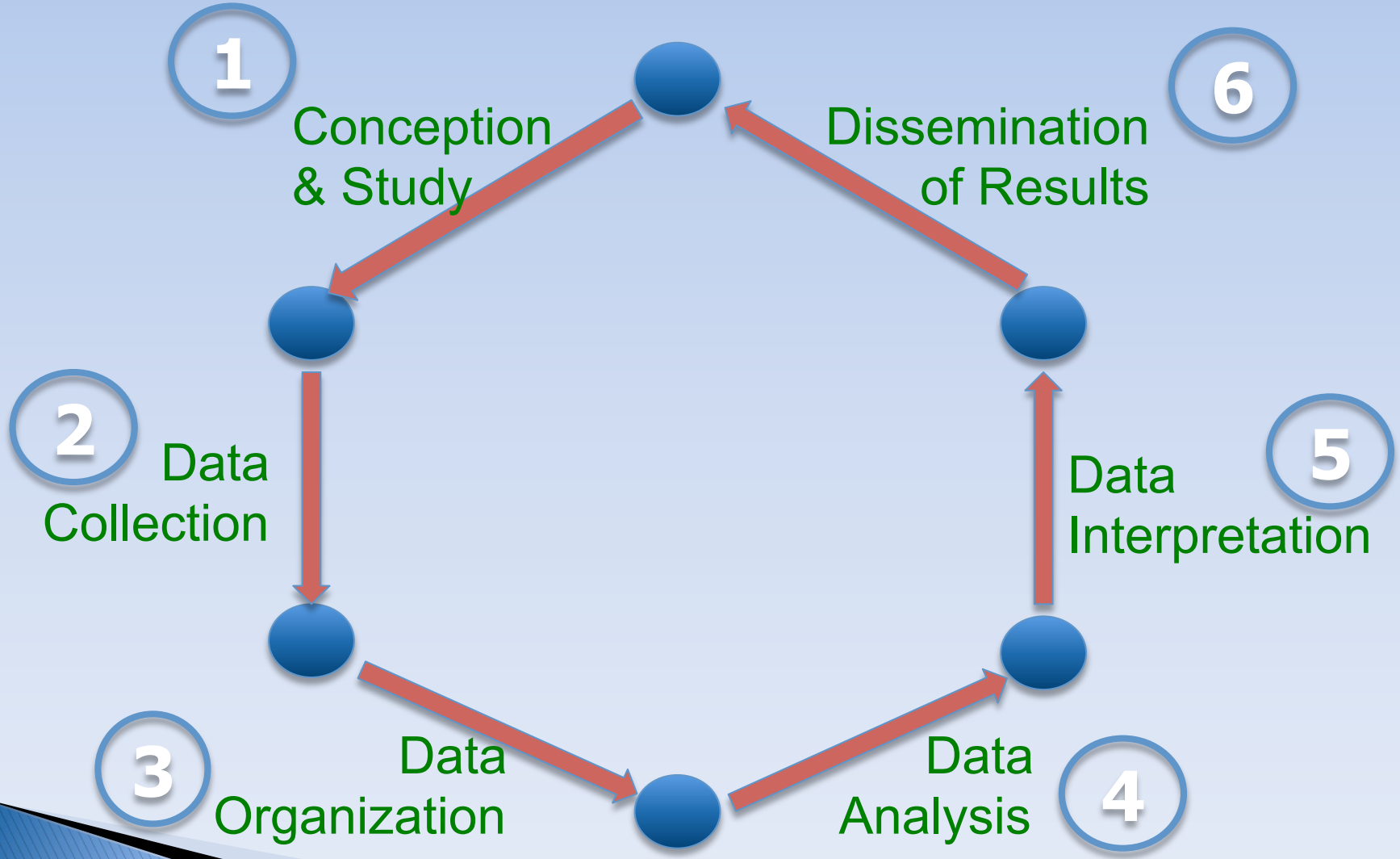
  Scientist at data management stages (often)

  - Data organization (3) and Data analysis (4)

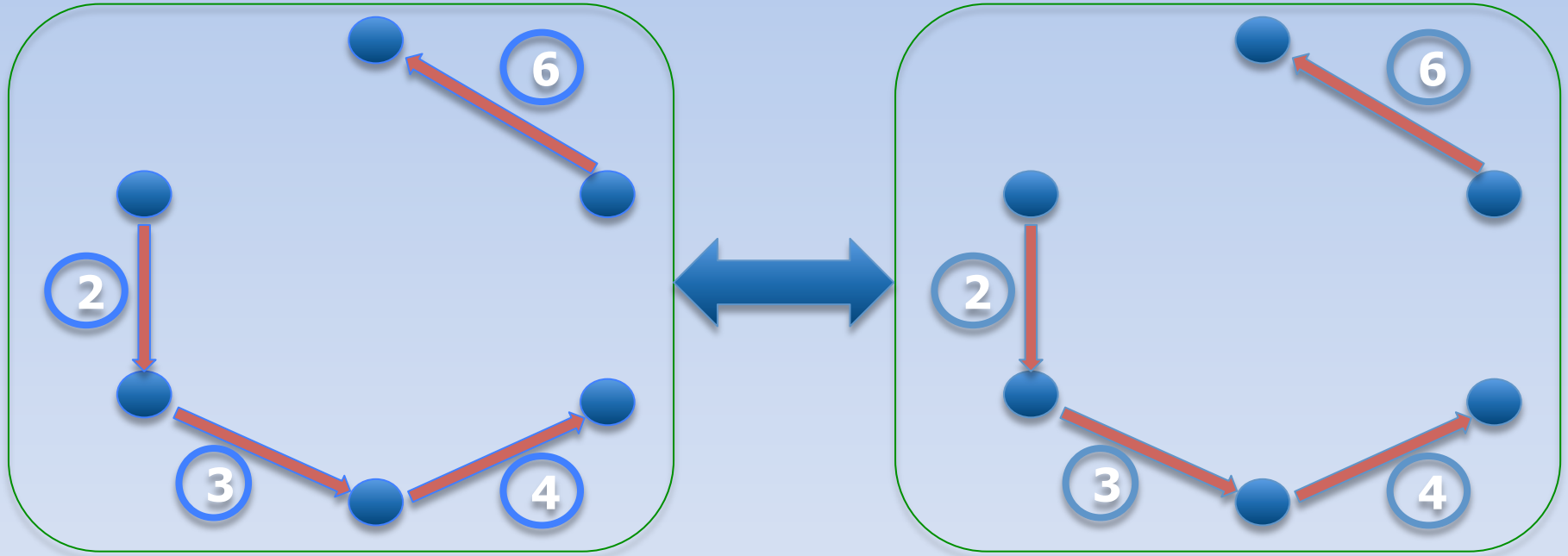
Roles wrt Informatics



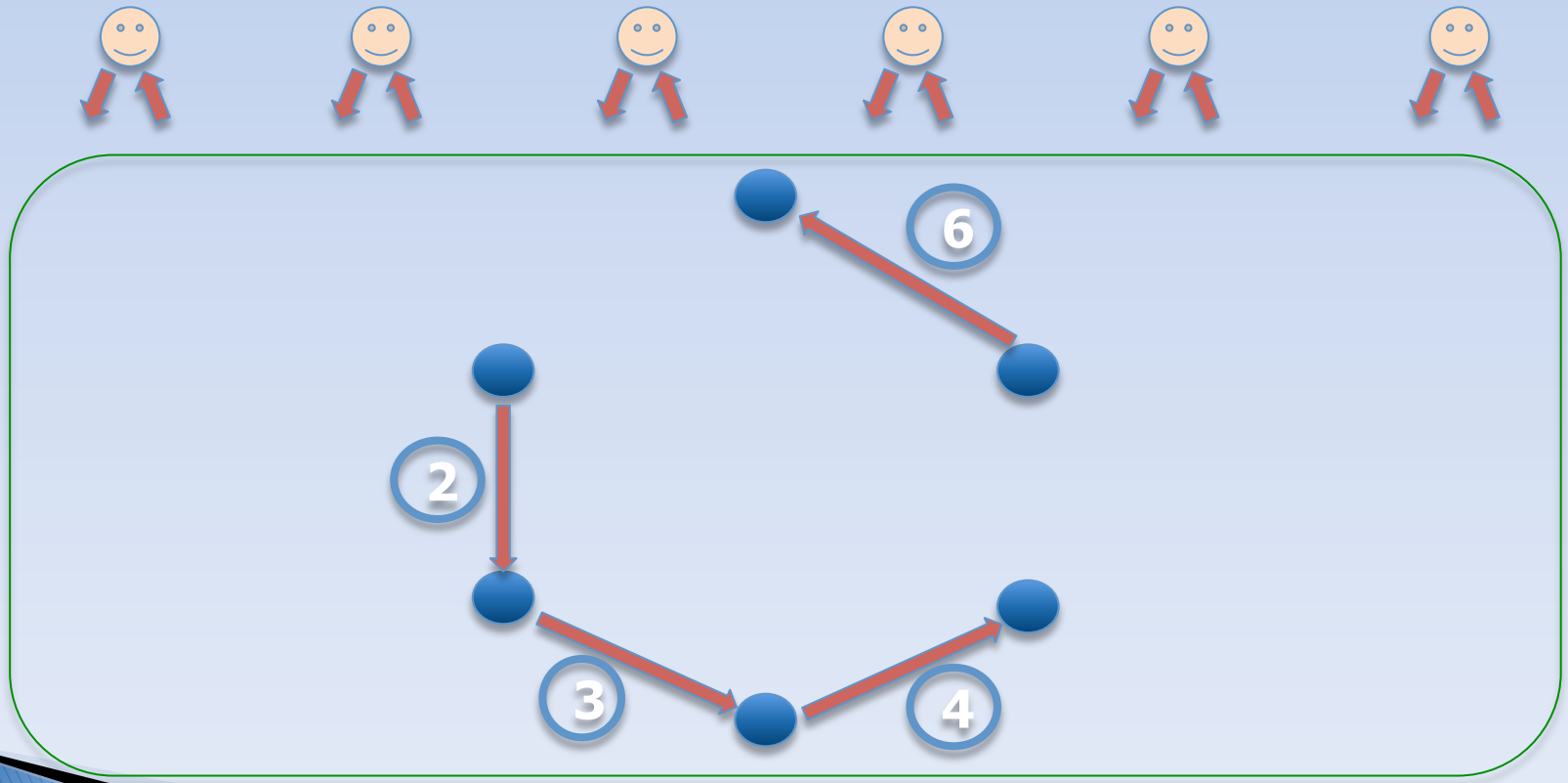
Research Life Cycle



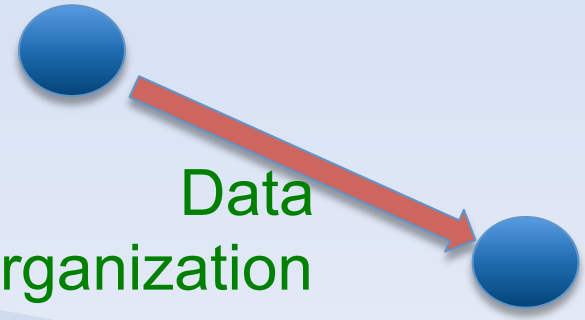
Agenda: Data Interoperability



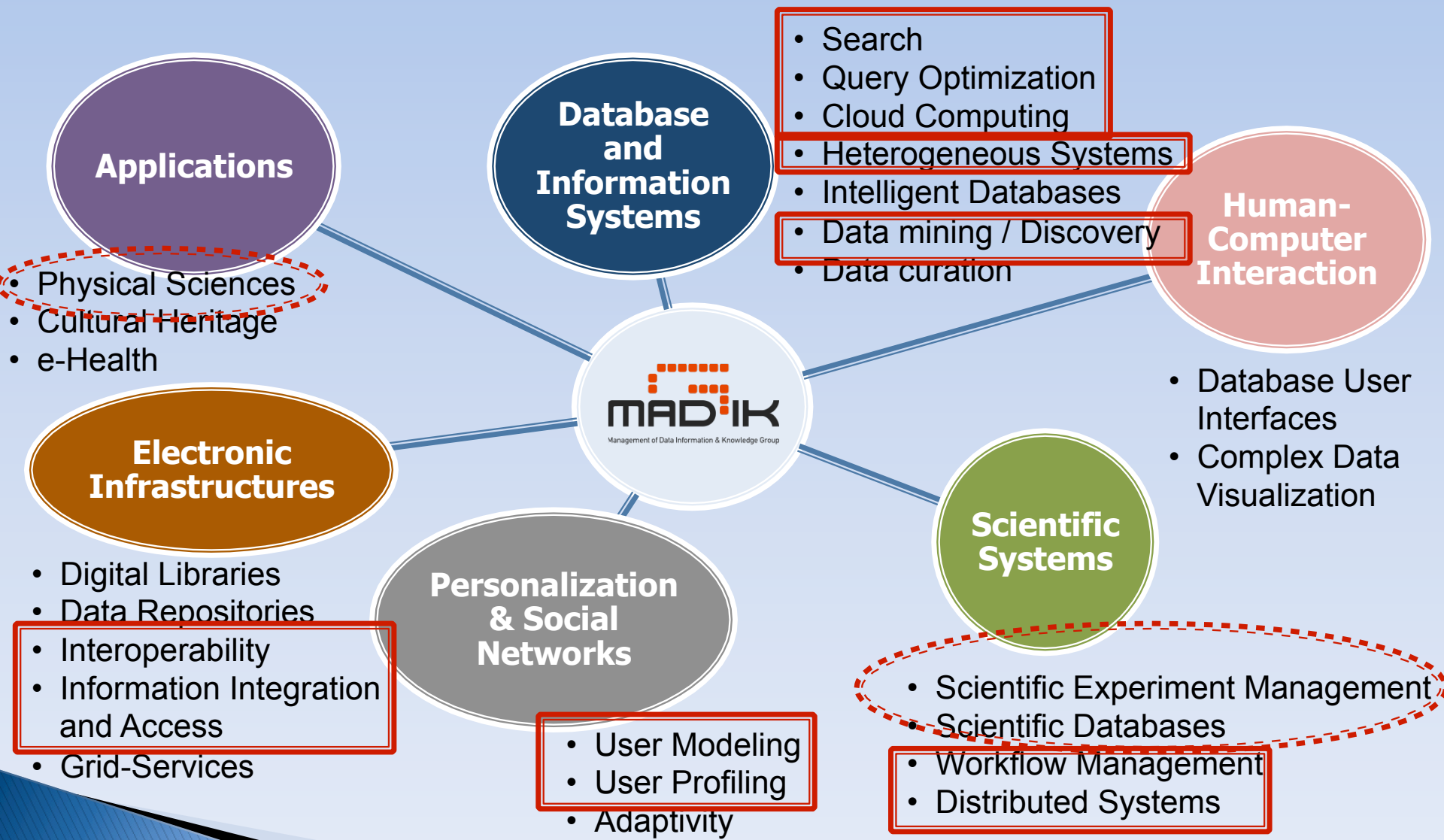
Agenda: Data Infrastructures



Agenda: Data Management



Research Areas



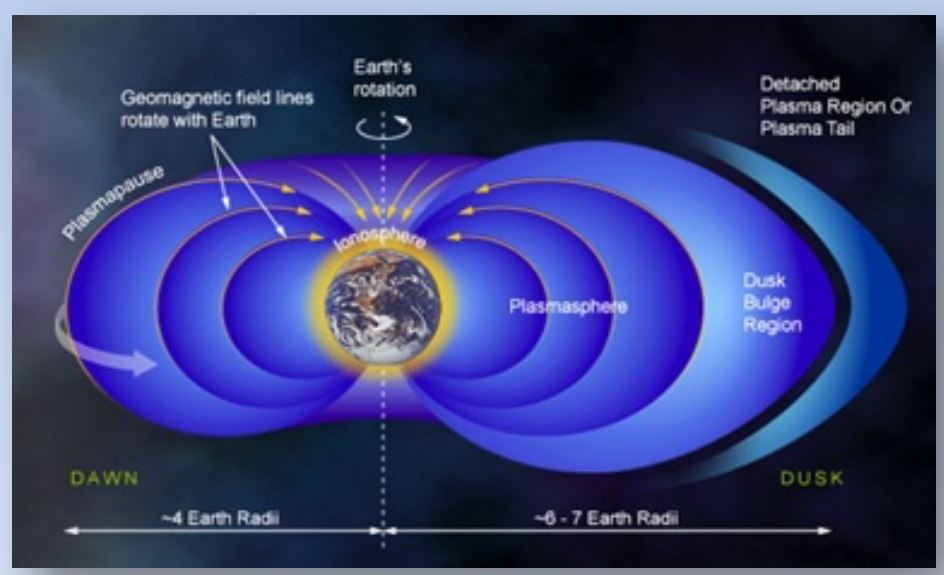
Spectrum of d-Science Projects

- ESPAS: Near-earth Space
- D4Science II: Environmental Monitoring
- iMarine: Fisheries and Marine Life
- EarthServer: Earth Sciences
- Interact: Arctic Environment
- EFG2: European Film Gateway
- CHESS: Museum Storytelling
- BMRB: Bio-magnetic Resonance (NMR)
- OpenAIRE/OpenAIREplus: Publication Monitoring

MAD-IK

Near-Earth Space

Disparate interrelated scientific communities



- Space weather
- Space climate
- Ionosphere
- Magnetosphere
- Lithosphere-Ionosphere Coupling
- Thermosphere

Industrial users and communities

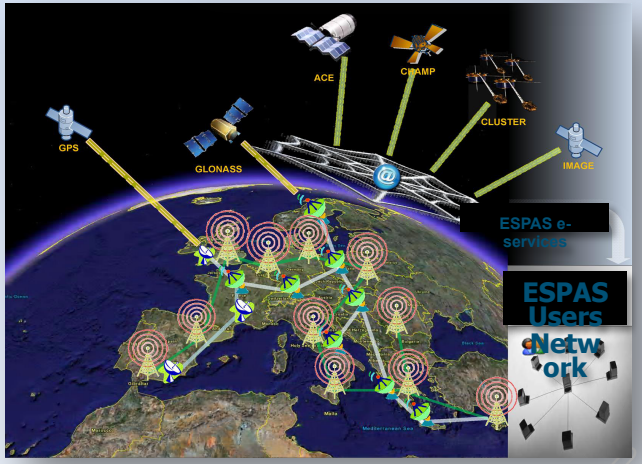
- Space communications
- Satellite Operation
- Navigation and Surveillance

Key issues and challenges

- Near-Earth space domain is yet to be established
  - Diverse organizations with different (or no) data policies
  - No formal data conceptualization
  - Different scopes, different use by researchers

22 partners

- Complexity of data/metadata
  - Rich set of distributed and moving sensors
  - In-situ & active/passive remote sensing

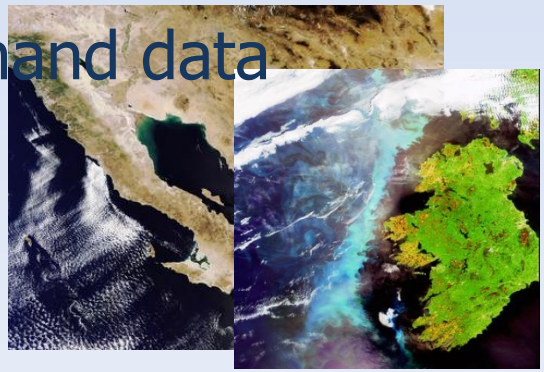


Key issues and challenges

- Data modeling
  - Diverse communities
- Data policies
  - Nothing exists
- Infrastructure to provide access to instrumentation
  - Different policies, different views
- Infrastructure experiment monitoring
  - Dynamic scientific workflows

D4Science: Environmental Monitoring and Fisheries

- European Space Agency, FAO, World Fish Center, …
- Global environmental issues: marine environment, forest ecosystem, air quality
- Sensor data analysis, integration and correlation of data sources; reasoning, information/knowledge mgnt
- Large amount of information (→ 1TB), added-value applications and services
- Seamless workflow definition & on-demand data processing



Fisheries & Marine Life

- Fishery@FAO, Unesco, …
- Worldwide researchers from many disciplines (biologists, climatologists, socio-economists, fishery managers, …)
- Sustainable environmental management
- Climate change mitigation, marine biodiversity loss containment, poverty alleviation, disaster risk reduction
- Use of global fishery ecosystem and aquaculture, e.g., species, aquatic resources, hydrological changes
- Extreme data diversity

EarthServer

Earth Sciences

- European Space Agency
- Atmospheric, oceanography, geology, and general earth observation communities
- Open access and ad-hoc analytics on Earth Science
- RASDAMAN raster data management system
- Queries on cross-domain data of 20+ Tbytes

MAD IK

Current limitations & gaps

- **Application-specific software** of limited long term value
- Absence of **consistent informatics perspective** for data management
- Discipline-specific solutions: science **re-builds** rather than **re-uses** software
- No set of **common requirements**

Common Challenges

- Managing and processing exponentially-growing volumes of non co-located data
- Dealing with time-sensitive stream data from arrays of sensors and instruments or simulations
- Significantly reducing data analysis cycles so that researchers can make timely decisions
- “Data cleaning” is much harder than data loading

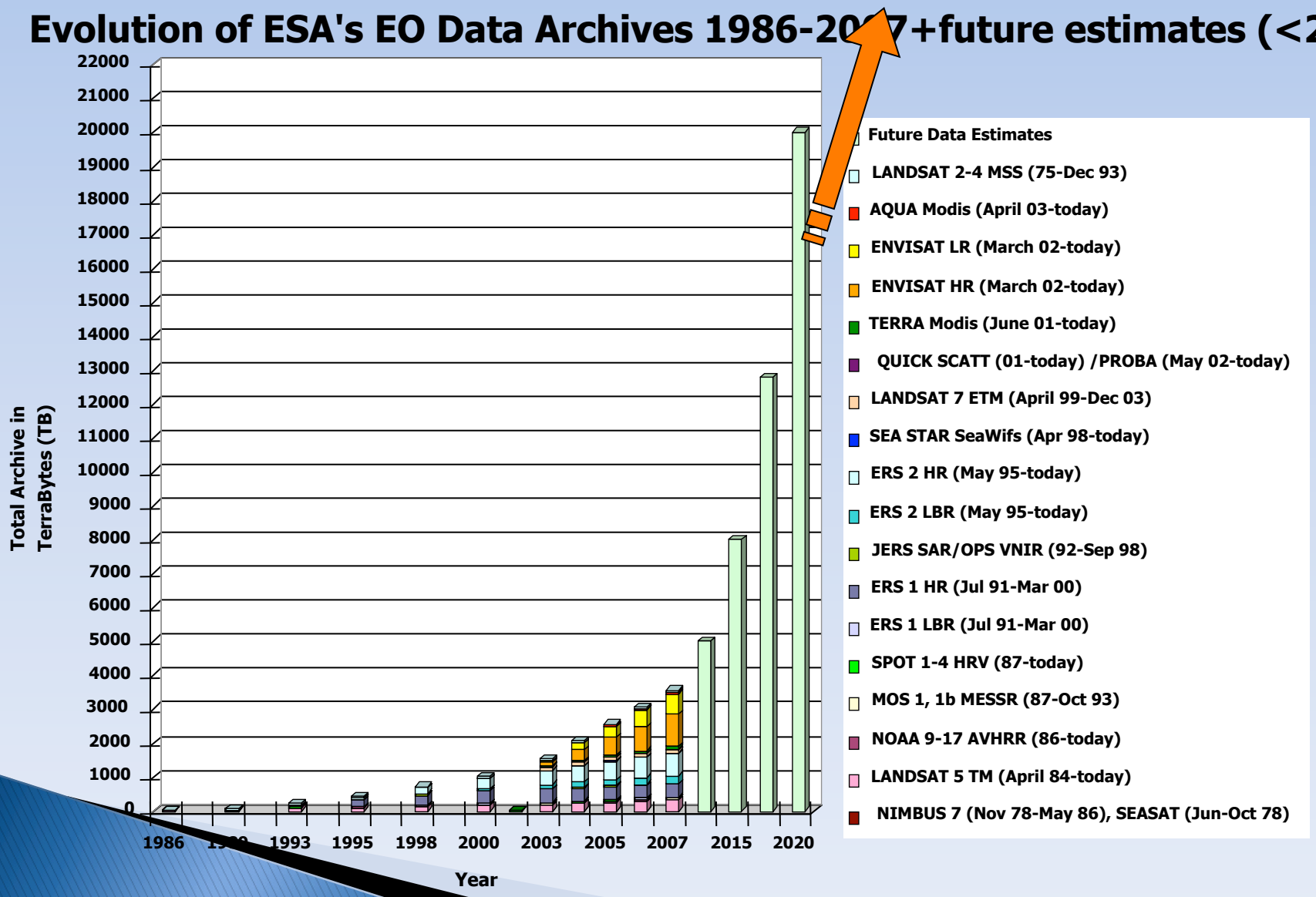
Common Research Issues

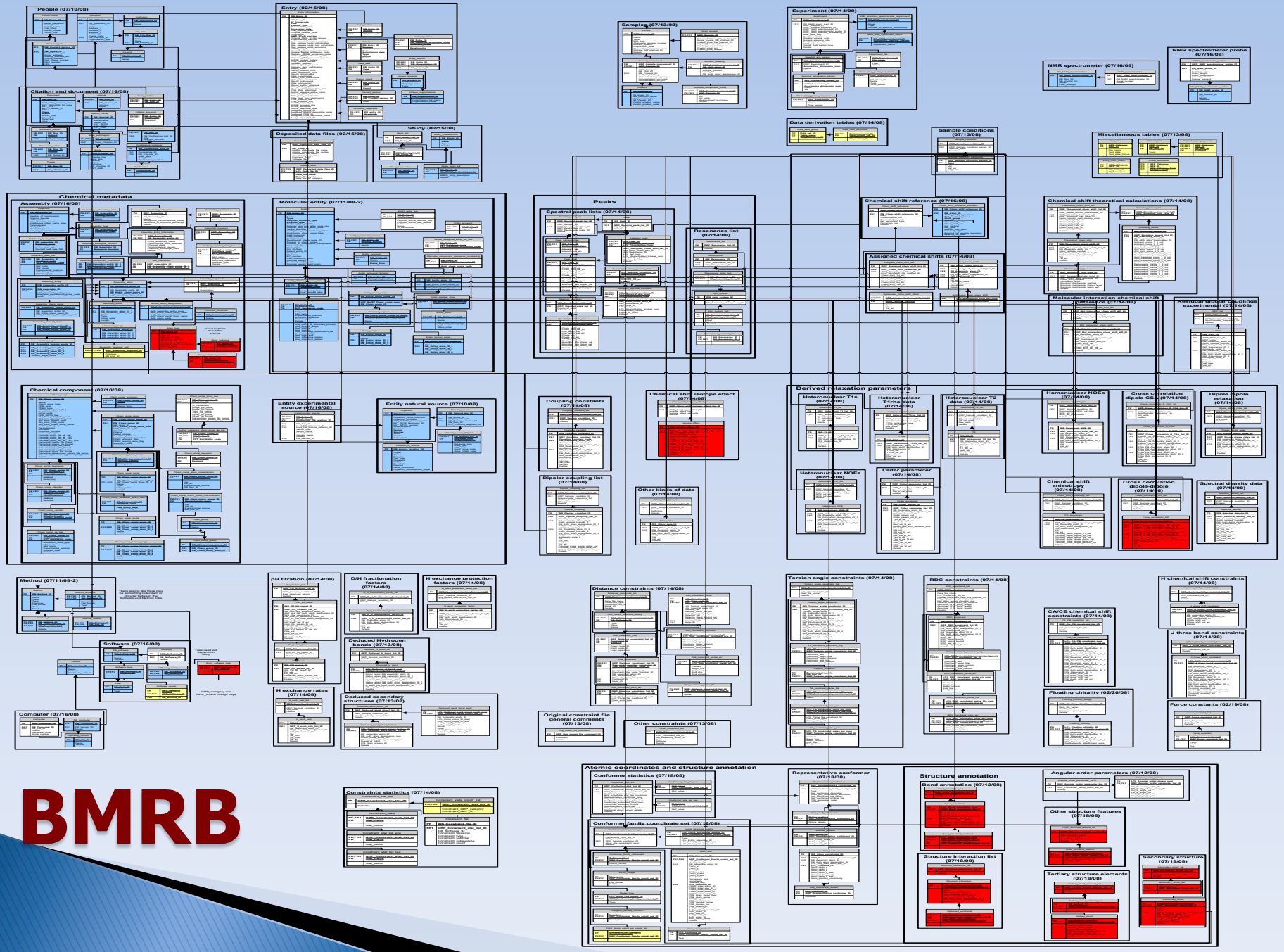
- Specialized storage and data structures
- Filtering and fusion
- Efficient querying and distribution
- Parallelism
- Very few predefined query patterns
  - Everything goes….
  - Rapidly extract small subsets of large data sets
  - Geospatial everywhere

Recent initiatives in informatics

- Open-source data management and analytics software specifically data-intensive scientific problems
- Main characteristics
  - scalability up to petabytes and beyond and down to megabytes (consistent interface)
  - built-in support for provenance (lineage), workflows, uncertainty
  - support for "external" data objects
  - optimizing query language
  - use of Open Standards
- **SciDB** (Stonebraker@MIT et al.), **Rasdaman** (Baumann@Jakobs U.), **MonetDB/SciLense** (Kersten@CWI)

Evolution of Space Info




People (07/10/08)
Entry (02/15/08)
Citation and document (07/16/08)
Deposited data files (02/15/08)
Study (02/15/08)
Samples (07/13/08)
Experiment (07/14/08)
NMR spectrometer (07/16/08)
NMR spectrometer probe (07/16/08)
Data derivation tables (07/14/08)
Sample conditions (07/13/08)
Miscellaneous tables (07/13/08)
Chemical metadata
Assembly (07/16/08)
Molecular entity (07/11/08-2)
Peaks
Spectral peak lists (07/14/08)
Resonance list (07/14/08)
Chemical shift reference (07/16/08)
Assigned chemical shifts (07/14/08)
Chemical shift theoretical calculations (07/14/08)
Molecular interaction chemical shift perturbation (07/14/08)
Residual dipolar couplings experimental (07/14/08)
Chemical component (07/10/08)
Entity experimental source (07/16/08)
Entity natural source (07/10/08)
Coupling constants (07/14/08)
Chemical shift isotope effect (07/14/08)
Dipolar coupling list (07/14/08)
Other kinds of data (07/14/08)
Derived relaxation parameters
Heteronuclear T1s (07/14/08)
Heteronuclear T1rho data (07/14/08)
Heteronuclear T2 data (07/14/08)
Heteronuclear NOEs (07/14/08)
Order parameter (07/13/08)
Homonuclear NOEs (07/14/08)
Cross correlation dipole CSA (07/14/08)
Dipole dipole relaxation (07/14/08)
Chemical shift anisotropy (07/14/08)
Cross correlation dipole-dipole (07/14/08)
Spectral density data (07/14/08)
Method (07/11/08-2)
Software (07/16/08)
Computer (07/16/08)
pH titration (07/14/08)
D/H fractionation factors (07/14/08)
H exchange protection factors (07/14/08)
Deduced Hydrogen bonds (07/13/08)
H exchange rates (07/14/08)
Deduced secondary structures (07/13/08)
Distance constraints (07/14/08)
Original constraint file general comments (07/13/08)
Other constraints (07/13/08)
Torsion angle constraints (07/14/08)
RDC constraints (07/14/08)
CA/CB chemical shift constraints (07/14/08)
Floating chirality (02/20/08)
H chemical shift constraints (07/14/08)
J three bond constraints (07/14/08)
Force constants (02/19/08)
Atomic coordinates and structure annotation
Conformer statistics (07/18/08)
Conformer family coordinate set (07/18/08)
Representative conformer (07/18/08)
Structure annotation
Bond annotation (07/12/08)
Angular order parameters (07/12/08)
Other structure features (07/18/08)
Structure interaction list (07/18/08)
Tertiary structure elements (07/18/08)
Secondary structure (07/18/08)
Constraints statistics (07/14/08)

BMRB

Complex Data Interdependencies

Problems

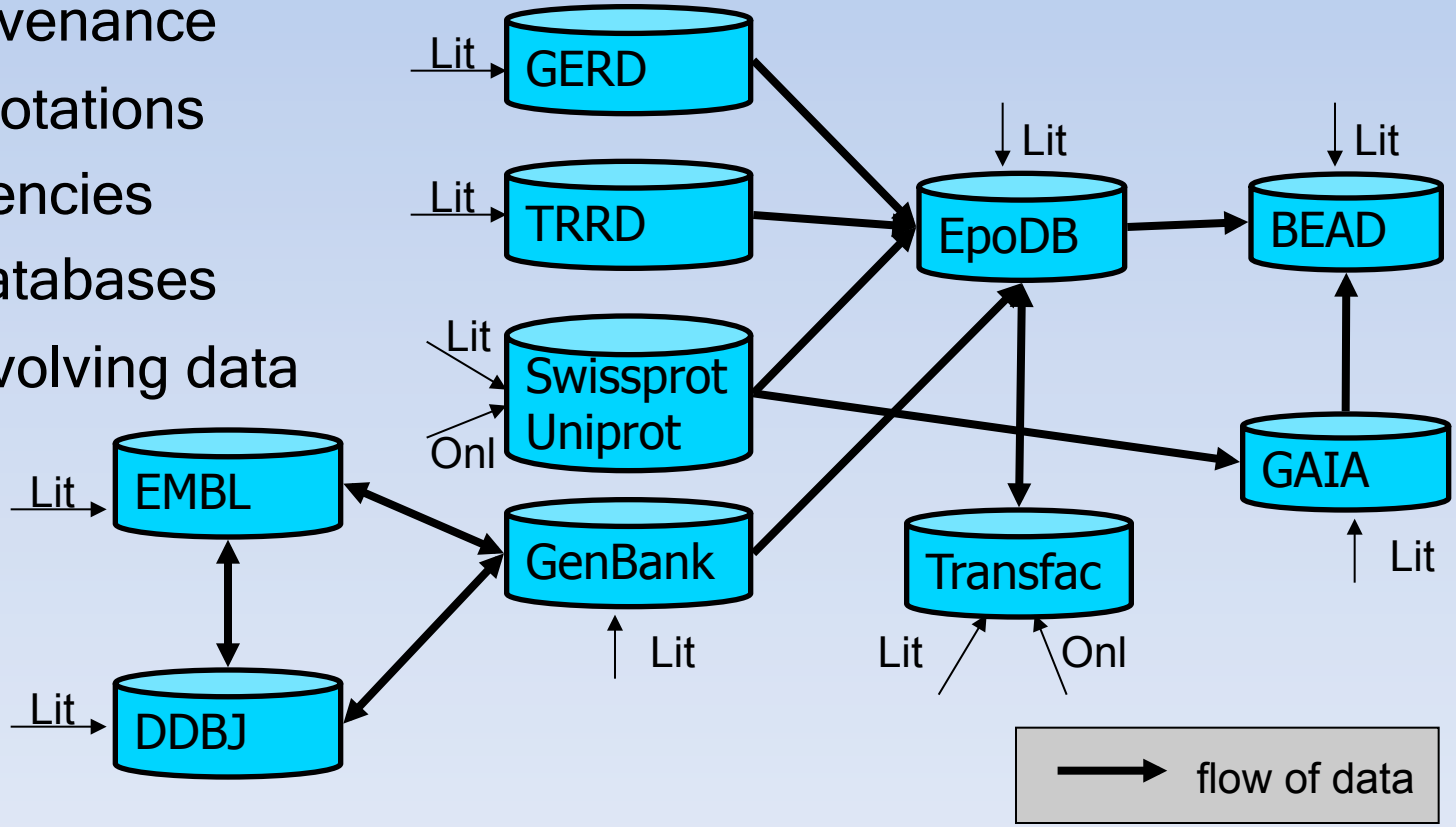
Trace data provenance

Propagate annotations

Cyclic dependencies

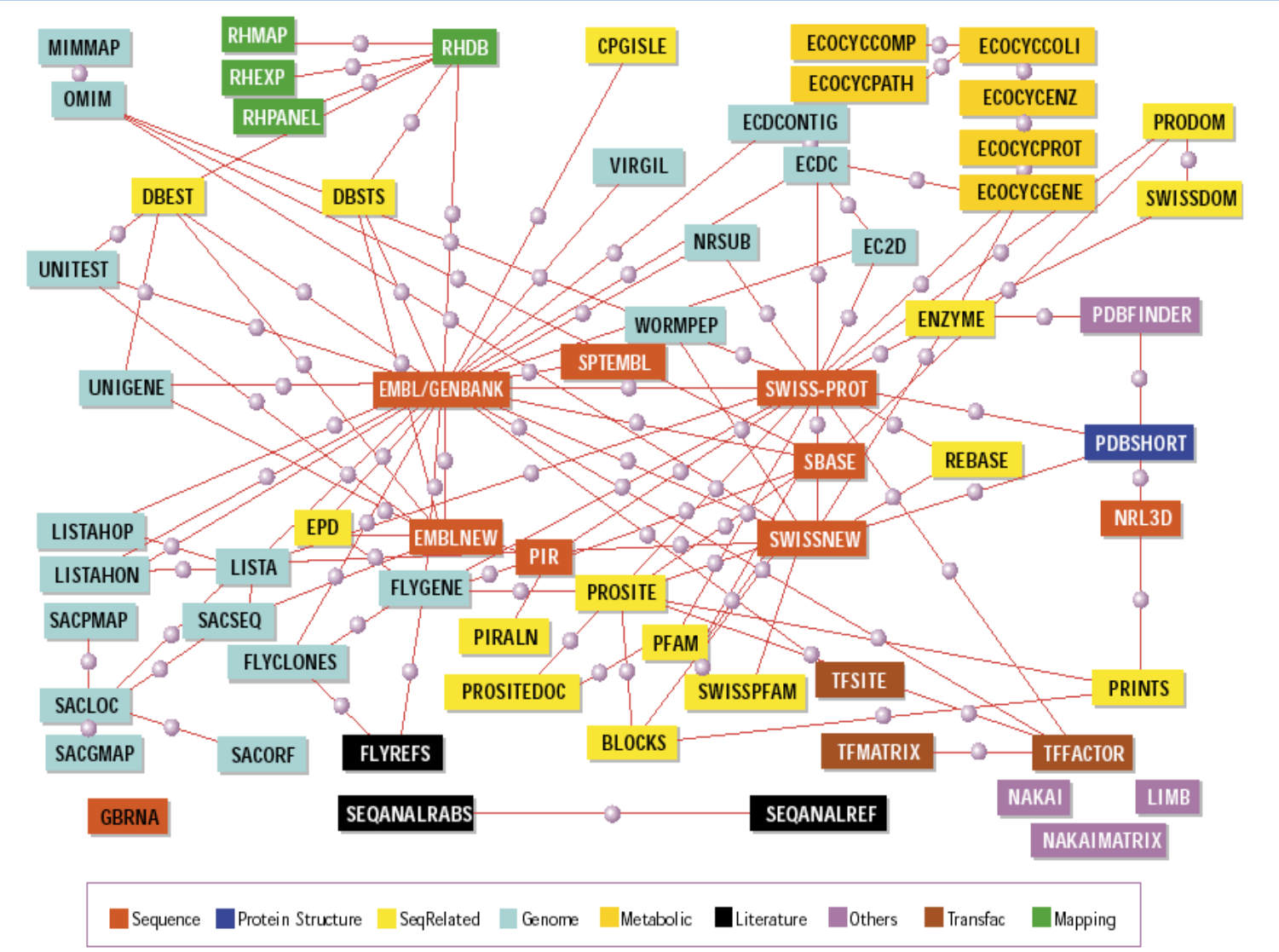
Cite curated databases

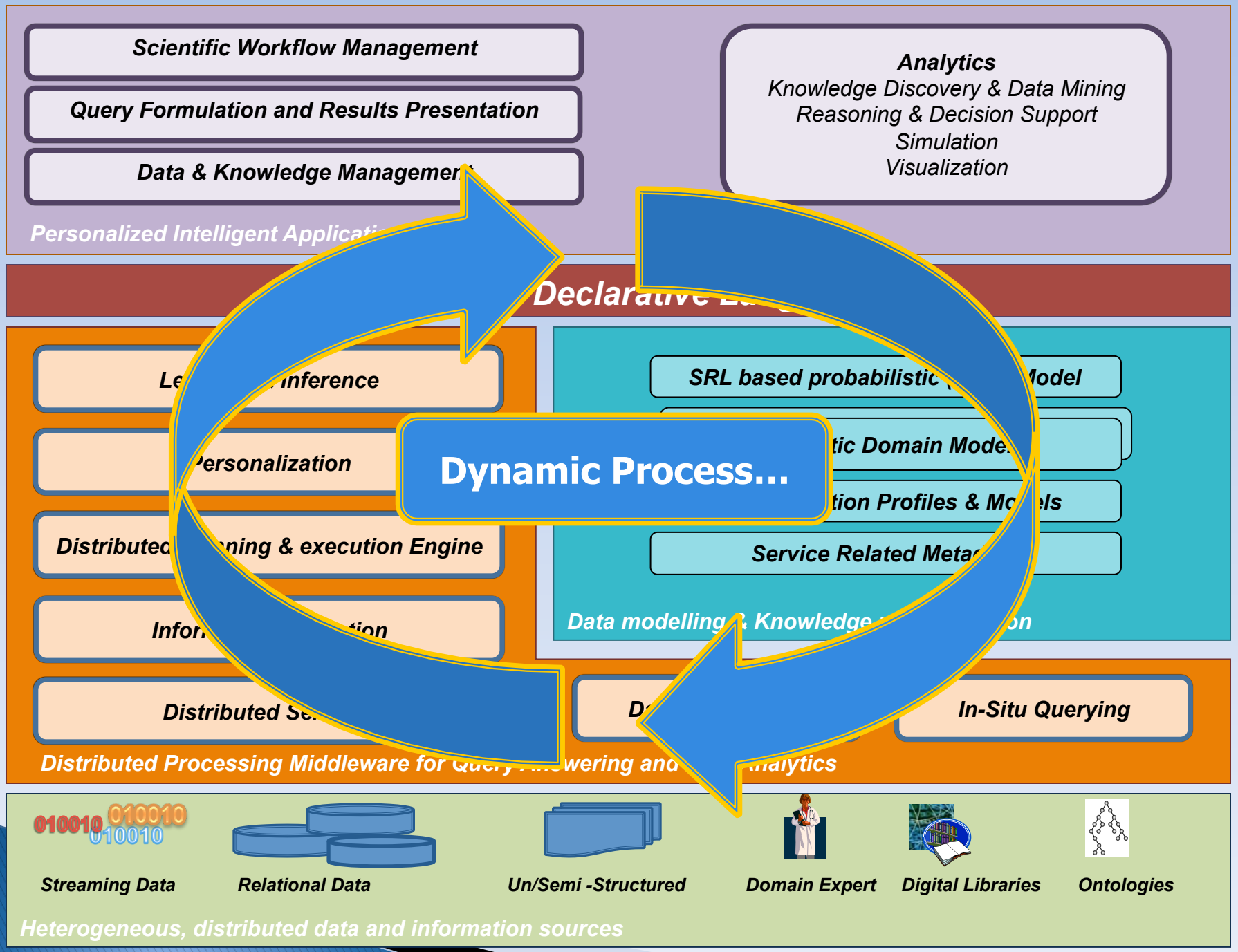
Synchronize evolving data



Lit: Data from Literature (human intervention)
Onl: Online Data Updates

Complex Data Interdependencies



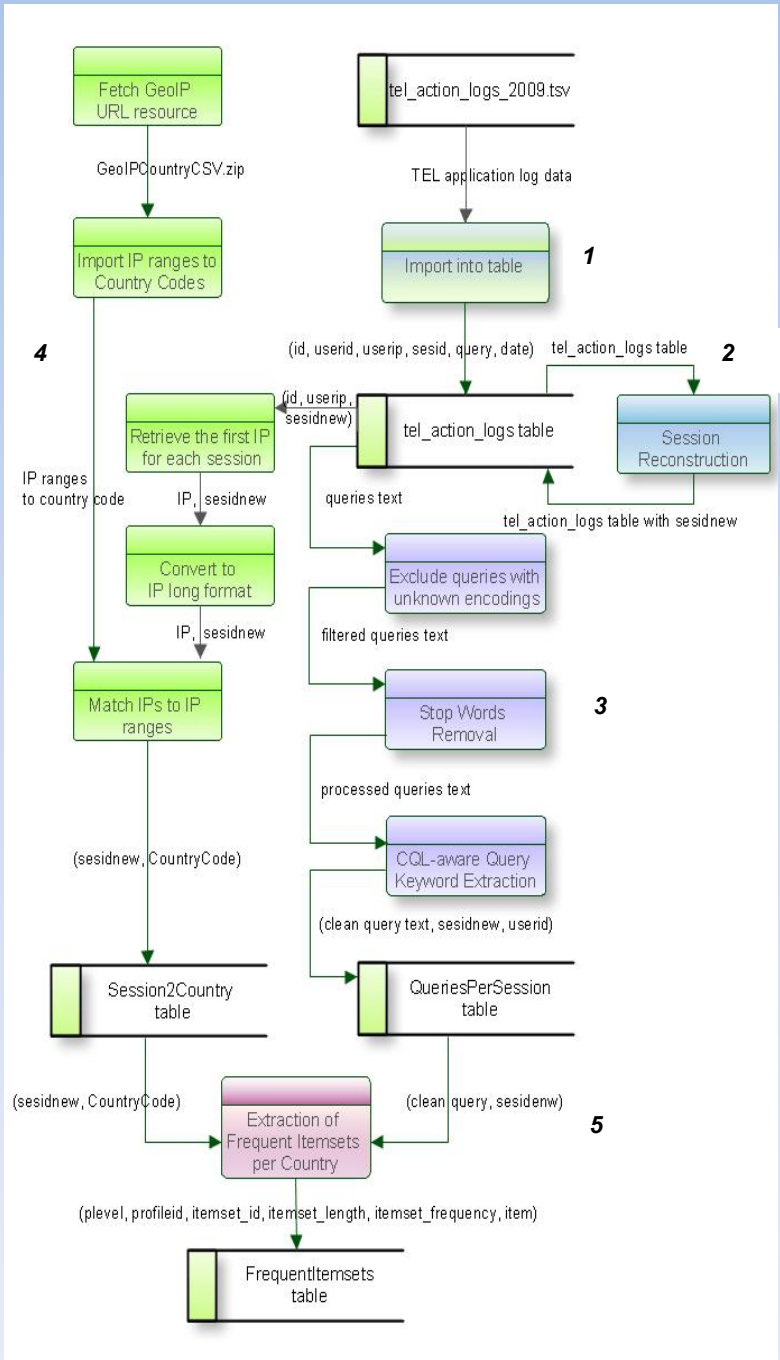

Scientific Workflow Management
Query Formulation and Results Presentation
Data & Knowledge Management
Analytics
Knowledge Discovery & Data Mining
Reasoning & Decision Support
Simulation
Visualization
Personalized Intelligent Applicati
Declarative La
Le inference
Personalization
Distributed ning & execution Engine
Infor tion
Distributed Se
SRL based probabilistic Model
tic Domain Mode
tion Profiles & Mo els
Service Related Meta
Data modelling & Knowledge on
Dynamic Process…
In-Situ Querying
Distributed Processing Middleware for Query nswering and nalytics
Streaming Data
Relational Data
Un/Semi -Structured
Domain Expert
Digital Libraries
Ontologies
Heterogeneous, distributed data and information sources

Emerging Query Optimization

- Query: graph of arbitrary operators
- Optimality: response time or completion time and money
- Environment: cloud of hosts (elasticity)

Motivation

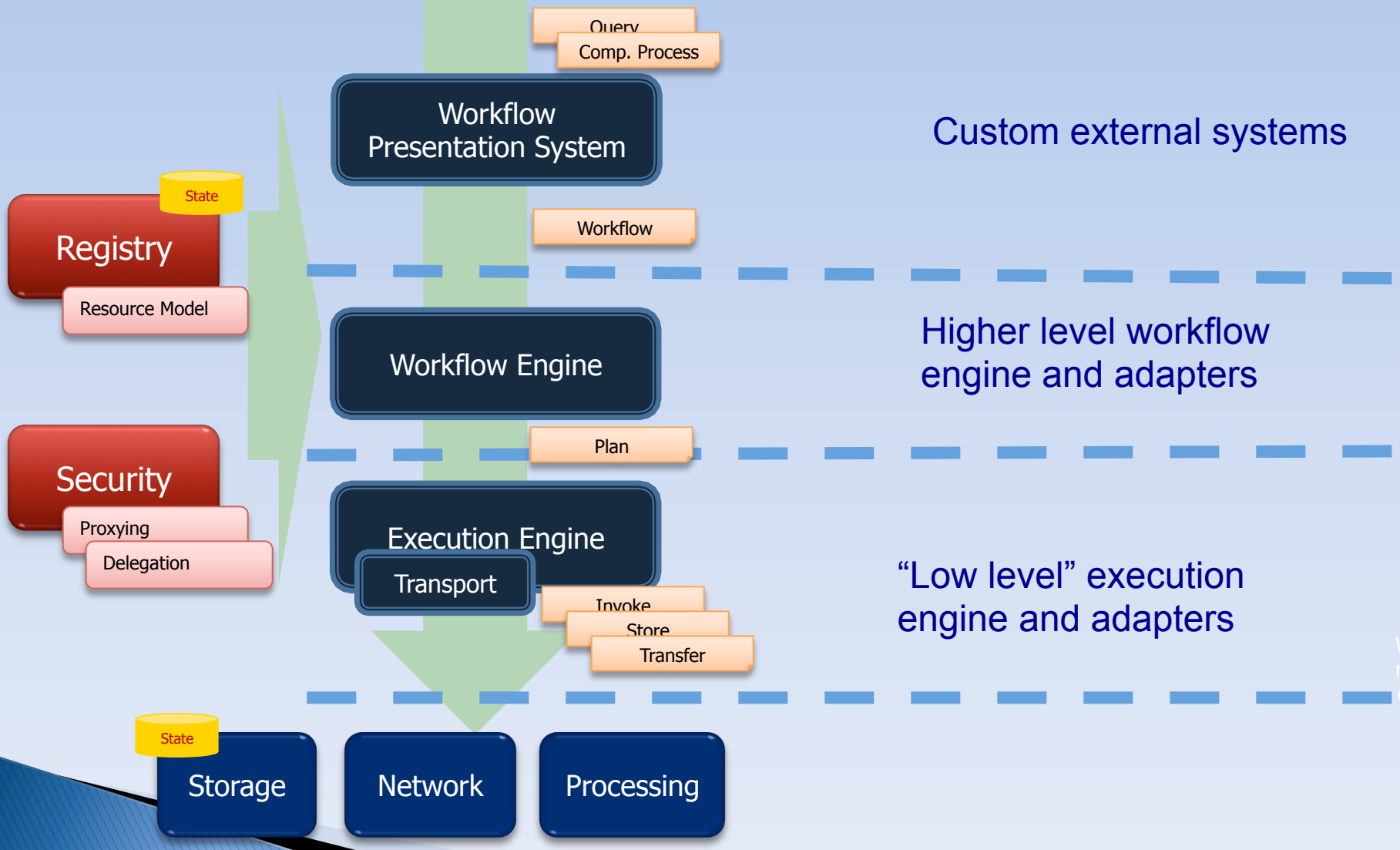
- Graph of arbitrary operators
- Non-relational data analytics
  - Query log analysis
  - Data mining
  - Simulation model composition
  - …
- User behavior analysis for European national libraries
  - One of sixteen flows



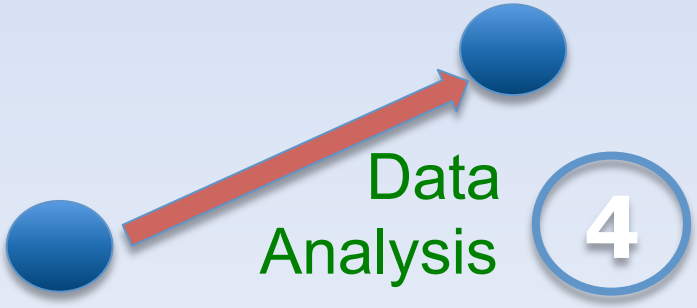
Objectives

- Execution of complex processing flow
- Staging of data among different storage providers
- Expressive and powerful execution plan language
- Technologies and disciplines abstractions - architecture of loosely coupled layers
- Unbound extensibility for integration with different environments (Storage, Security, Execution, …) – plugability
- Execution of … everything (SOAP/REST & WSRF web services, local executables and scripts, POJOs, native components) and in any manner (in-process, intra-process, intra-node)
- Align with cloud computing principles & data stream pipelining

Architecture Overview



Agenda: Data Management



Analysis and Databases

- Much statistical analysis deals with
  - Creating uniform samples
  - Data filtering
  - Assembling relevant subsets
  - Estimating completeness
  - Censoring bad data
  - Counting and building histograms
  - Generating Monte-Carlo subsets
  - Likelihood calculations
  - Hypothesis testing
- Traditionally these are performed on files
- Most tasks are much better done inside a database
  - Indices, parallel data search & analysis

MAD IK

AITION Data Mining Process

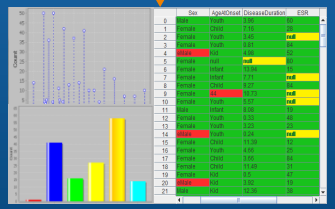
1. Qualitative dependency analysis: Learning the structure
2. Quantitative analysis: Learning the parameters (CPD)

Make inferences based on the model (Graph + CPD) using different known / unknown schemes

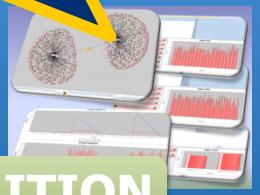
Raw Data

DCV

Data Curation & Validation

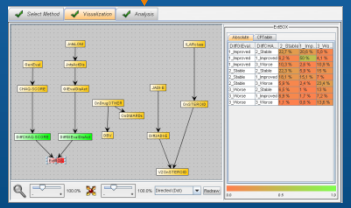


Curated Data

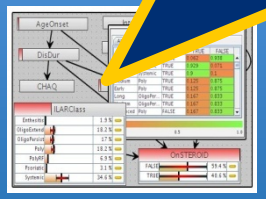


AITION DESK

Model Building

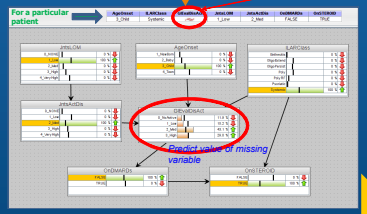


Model



AITION VIZ

Model Exploration & Reasoning

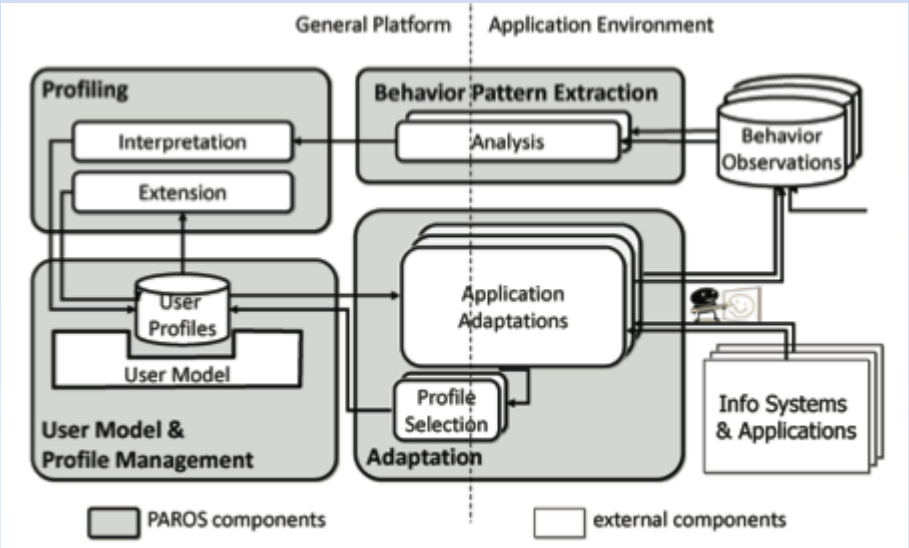


Decisions & Simulation

One way & static ....

PAROS Personalization Framework

- Provide:
  - Modelling user preferences & attitudes
  - Personalized querying, context & services
  - Graph mining for usage pattern analysis



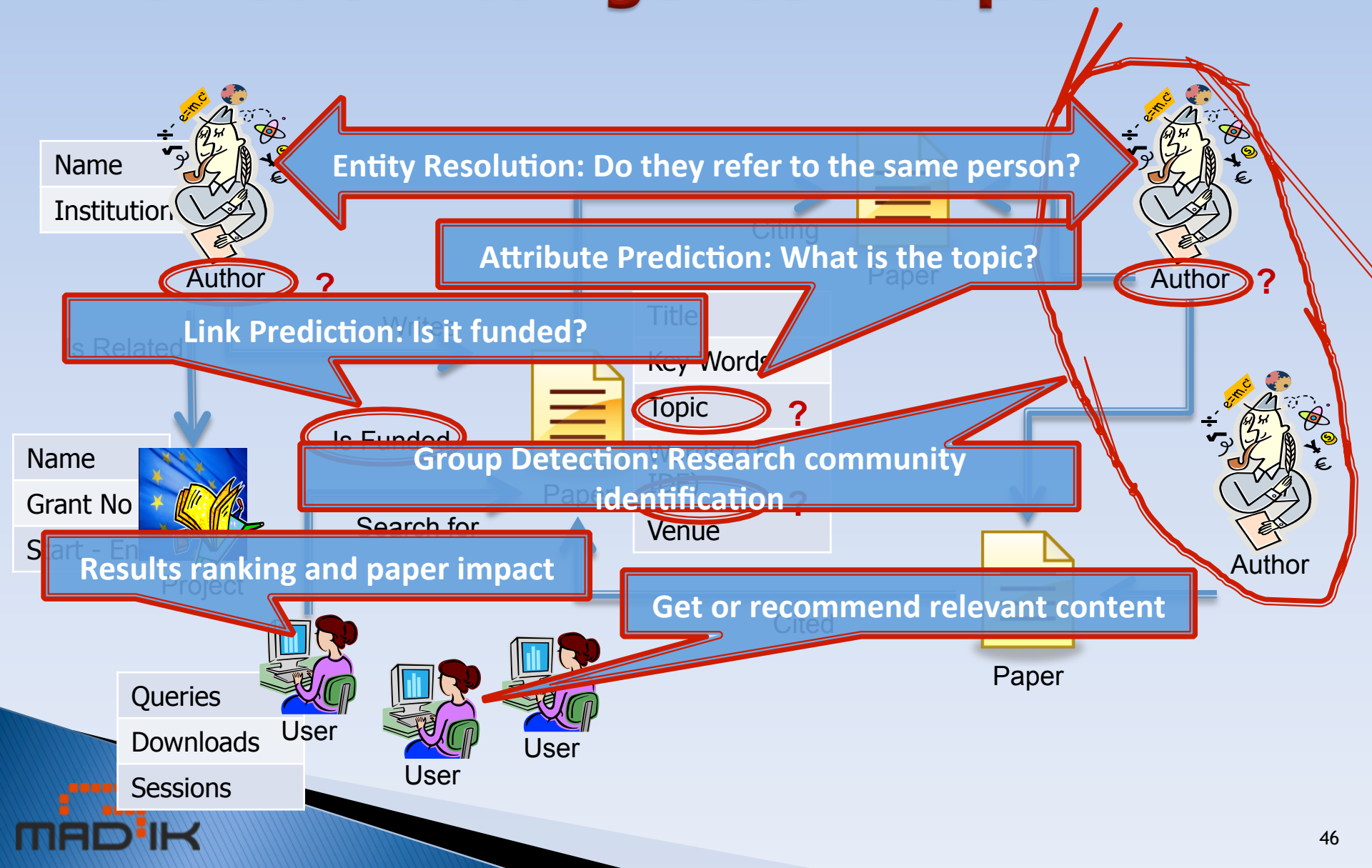
Phase II: Holistic approach…

- Evolutionary information processing and knowledge discovery framework able to provide highly accurate predictive and simulation models combining:
  - **Bottom-up data driven process:** analyzing huge volume of high distributed, federated, multi-type, vertically integrated data and streams,
  - **Top-down model driven process:** incorporating domain knowledge represented as relational & semantic theoretic models and First-Order Logic (FOL) rules & constraints.

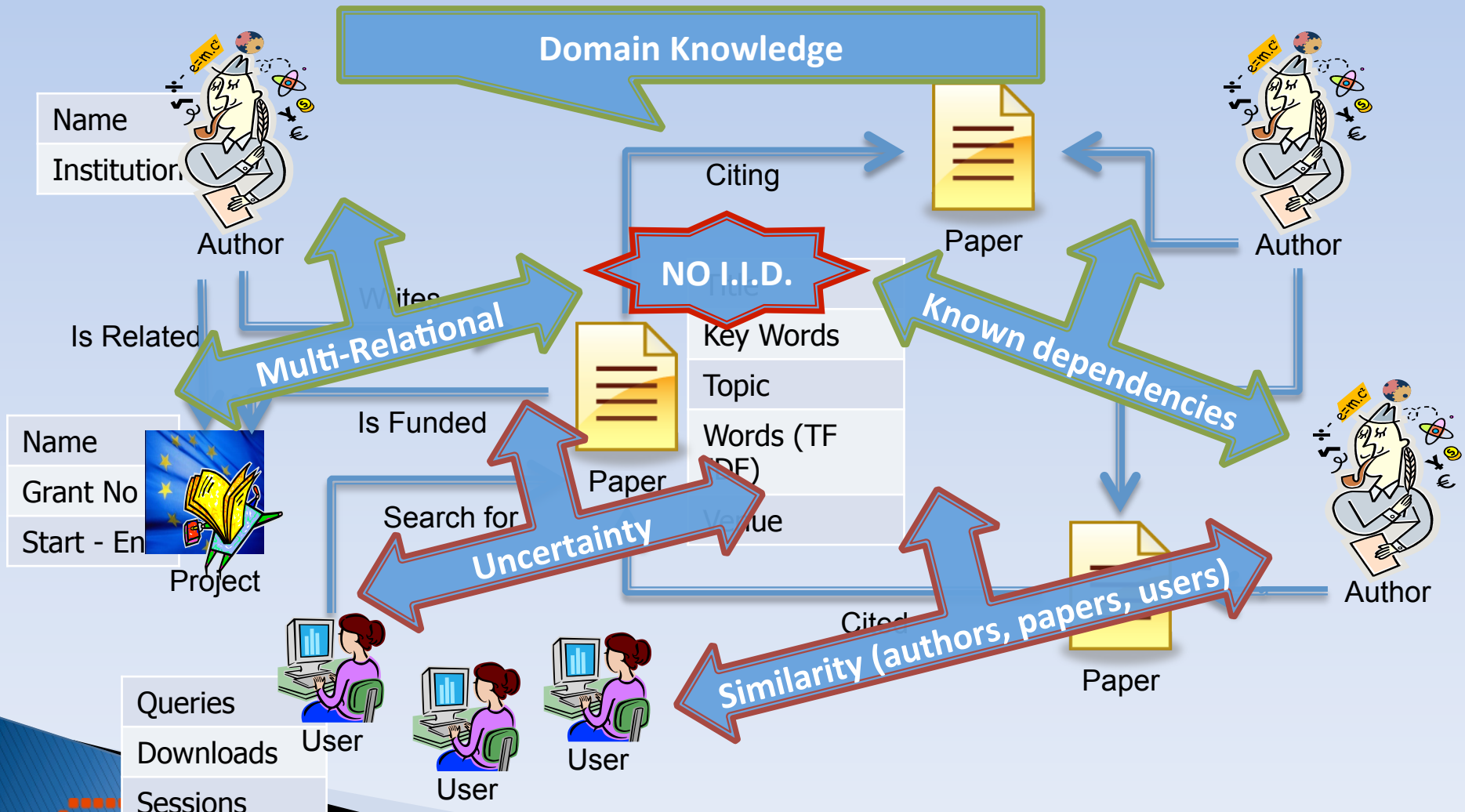
Basic Problem….

- Traditional Machine Learning (ML) approach based on single relation (propositional) and consider i.i.d. data
- Traditional Data Mining (DM) cannot handle uncertainty in attributes, entities, relations
- But:
  - Uncertainty is everywhere:
    - No complete knowledge
    - Noise/outliers in the data due to measurement procedures, human involvement in the loop, multi-cite, etc.
  - Real world has objects, properties, known relations
  - Instances are not always independent

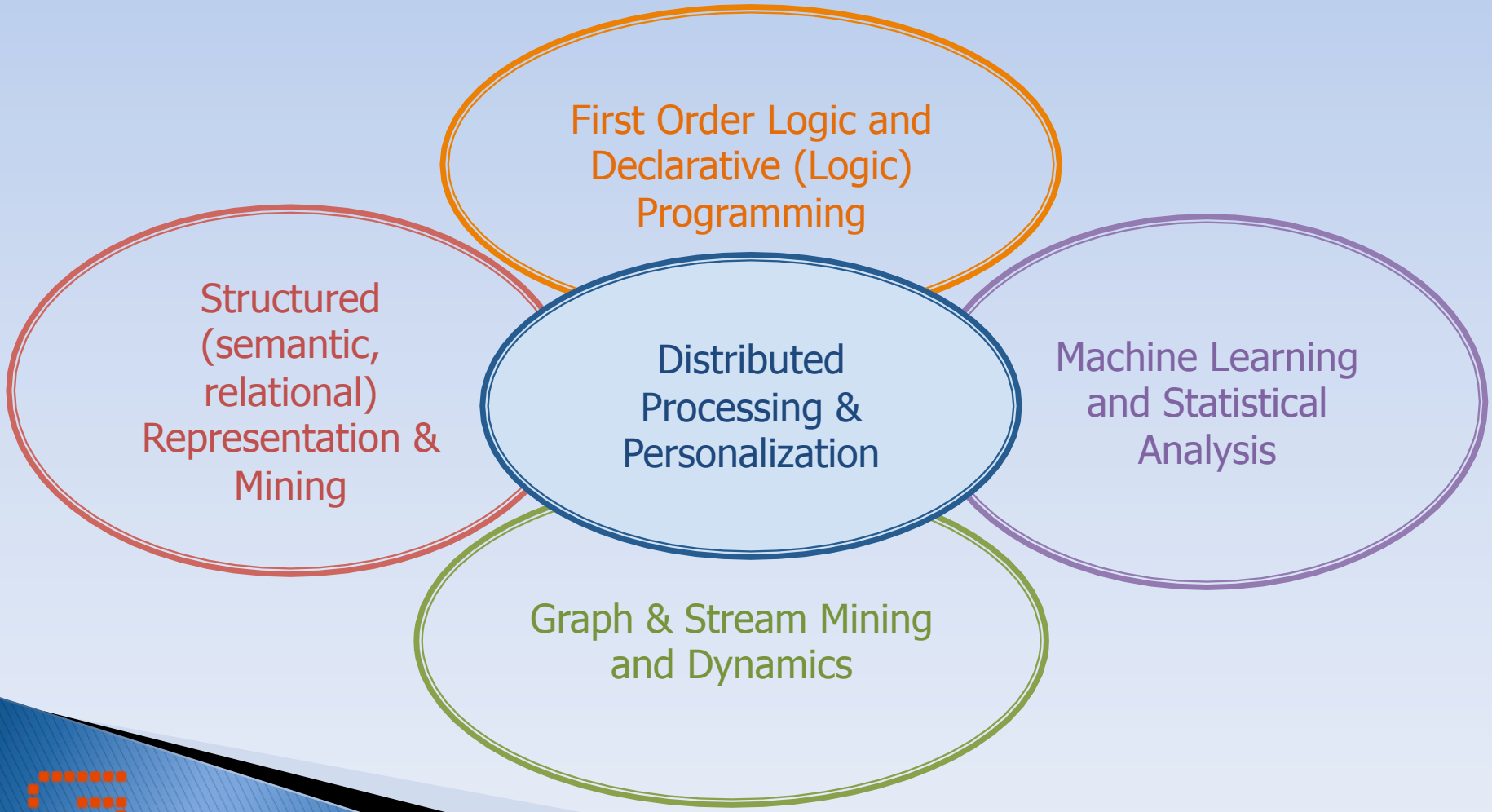
Hollistic Intelligence in OpenAIRE



Intelligence in Open AIRE



Holistic approach…



MAD IK

ICT Results ...Results that lead the way

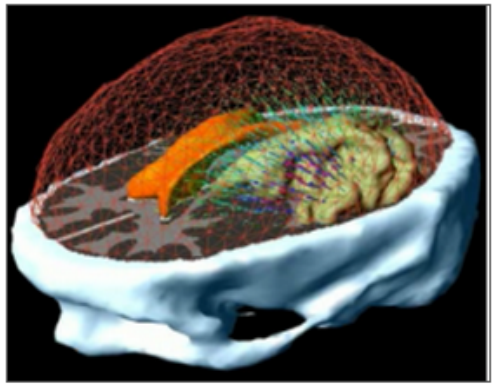
Relevant ICT Links Go

ICT Results
Features & Publications
Press Desk
Investors Room
Help & Links
ICT Results for You

Features

Print friendly Send this feature Send Feedback

Software for solving life-threatening medical puzzles



New software is under development that doctors hope will help them identify brain tumours in children that will grow aggressively.

Some brain tumours in children remain benign and doctors choose not to operate. But a small percentage of those will suddenly start to grow aggressively.

Doctors have not identified what triggers that aggressive tumour growth, despite the vast array of data they hold on their child patients – demographic, environmental, genetic and clinical data, as well as images such as MRI and CAT scans of the developing tumours.

But a new software tool called AITION can integrate all the medical data from a tumour patient and then analyse it to calculate the probable factors that are stimulating tumour development, combining up to 30 correlated variables. AITION provides an overview of the causal relationship across all factors.

Graphical network of causal relationships

AITION's conclusions are displayed as a 'knowledge model', a graphical network of medical factors with links that represent the correlations between them. Strongly interdependent concepts are directly connected, loosely dependent concepts are not connected at all. The patient's doctors can play around with the knowledge model. They can improve the model by adding information they know to be true about the patient. They can use the model to test the likely effects of different types of medication, surgery

Agenda: Data Management

Open Access

Open Access brings the innovation cycle to a new open world, contributing to the Digital Agenda for Europe

"Scientific data has the power to transform our lives for the better – it is too valuable to be locked away."

Neelie Kroes,

Vice-President of the EC, responsible for the Digital Agenda

Open Science

- **Open Science** enables researchers to conduct **efficiently** and **effectively** their research activities

- **Open Science** facilitate multidisciplinary **collaboration** and data **sharing** entailing **Open Access** to
  - research data
  - data services
  - tools
  - analyses
  - methods

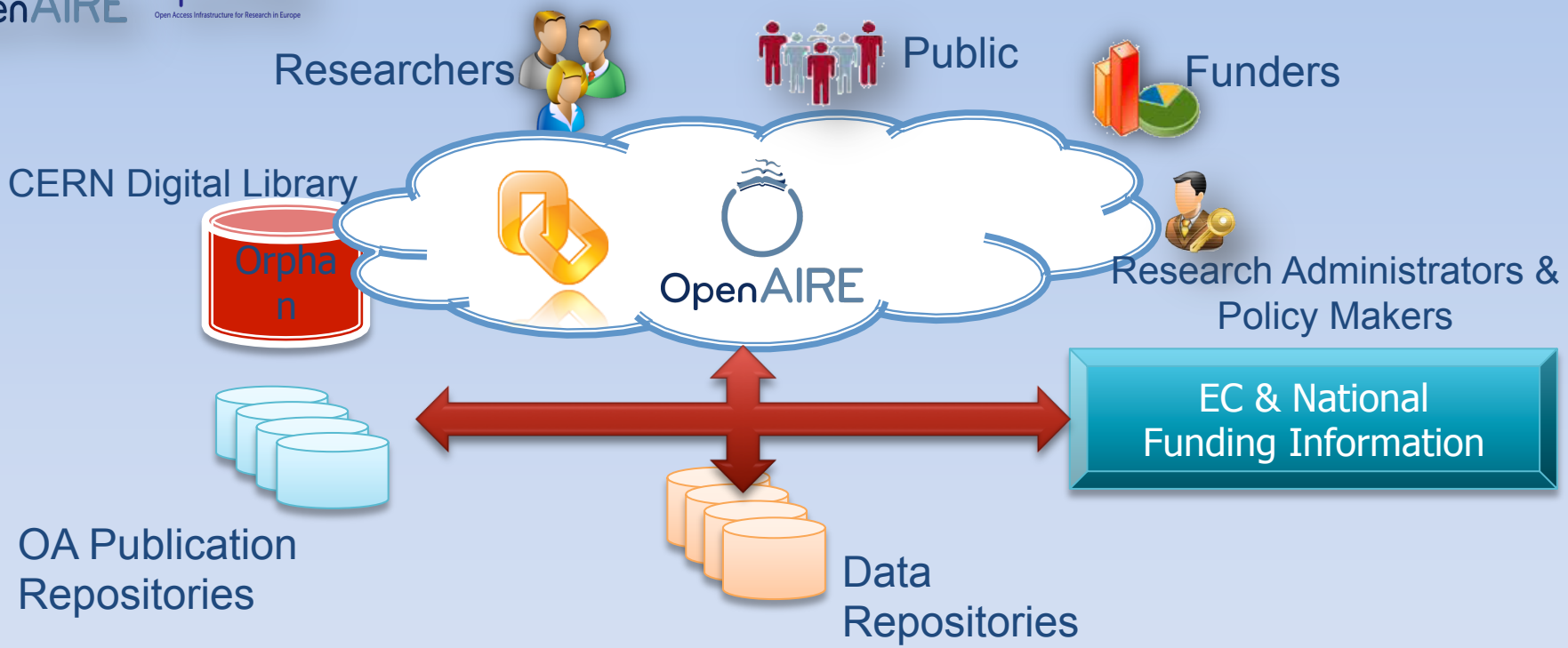
Open Access & Scientific Freedoms

- **Open Access** improves **information access, use and reuse -** easier & more reliable
- **Open Access** improves **free info movement** in ***Research*** and ***Public Sector***, benefiting
  - researchers, students, citizens
  - funding officers and industry

Investing in Open Data Infrastructures

- **Relax** and **bridge** geographic, organizational, disciplinary boundaries
- A domain with a **global** horizon

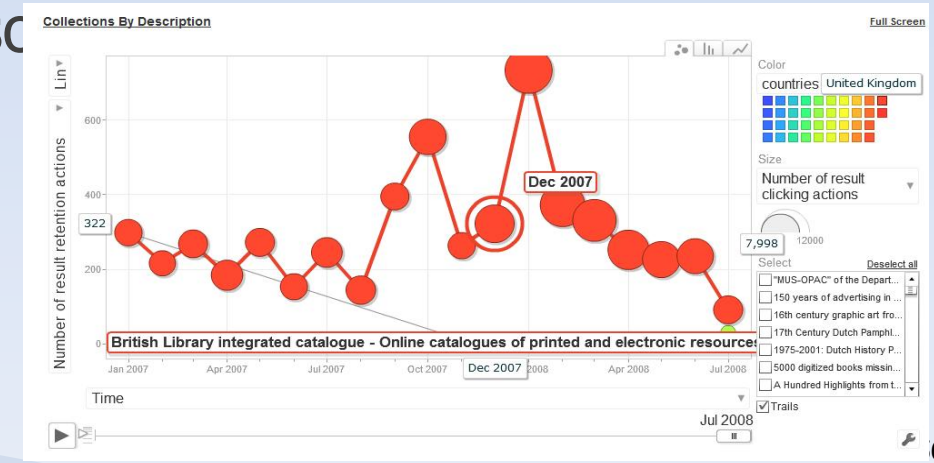
Unifying European Research



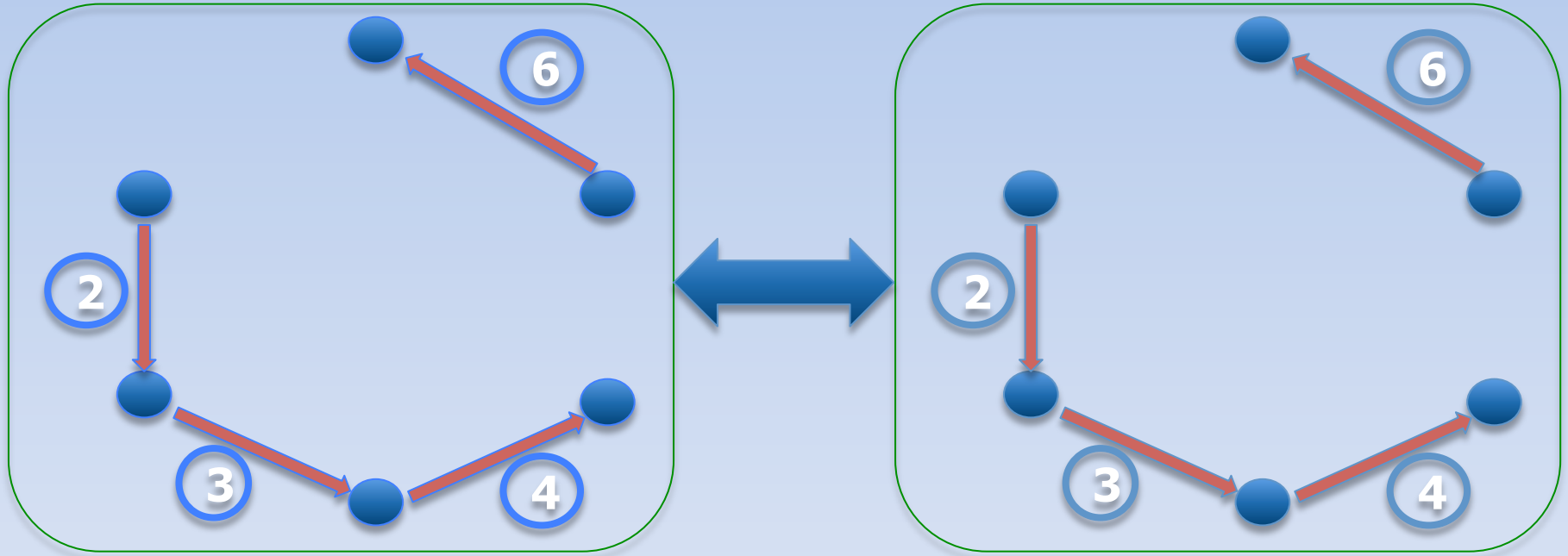
- Links OA repositories with FP7 / ERC /national funding
- Links publications with research data
- Provides repository for "orphan" publications & data
- Calculates statistics on OA/FP7/ERC /… deposition

Usage data

- Analysis to measure research impact
- Usage pattern analysis tools for content linking
- Challenges, research direction
  - Big volumes of data to process
  - Definition of metrics
  - Researcher behavior and sc
  - Organizational issues



Agenda: Data Interoperability



Interoperability

DEFINITION

- "*The ability of two or more systems or components to **exchange** information and to **use** the information that has been exchanged*" (IEEE definition)
- *"A property referring to the ability of diverse systems and organizations to work together (inter-operate)" (Wikipedia)*
- "*The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires minimal knowledge of the unique characteristics of those units" (ISO/IEC 2382-2001 Information Technology Vocabulary, Fundamental Terms*)

MADIK

Data interoperability

- The ability to exchange data between different systems and use them
- Indicative areas: information representation - profiling, discovery, security
  - Multidisciplinary information – Scientific heterogeneity
  - Inconsistency of intended use of information

Technical Challenges

- **Heterogeneity** of the exchanged information objects
  - Syntactic heterogeneity
  - Structural heterogeneity
  - Semantic heterogeneity
- **Inconsistency of the intended use of the information object**
  - by the producer entity and the intended exploitation of this object by the consumer entity
  - the exchanged information must be complemented with some **"descriptive"** information, such as contextual, provenance, quality, security, privacy, etc. information
  - The descriptive information should be modeled by **purpose-oriented descriptive data models /metadata models.**

MADIK

Organisational and policy challenges

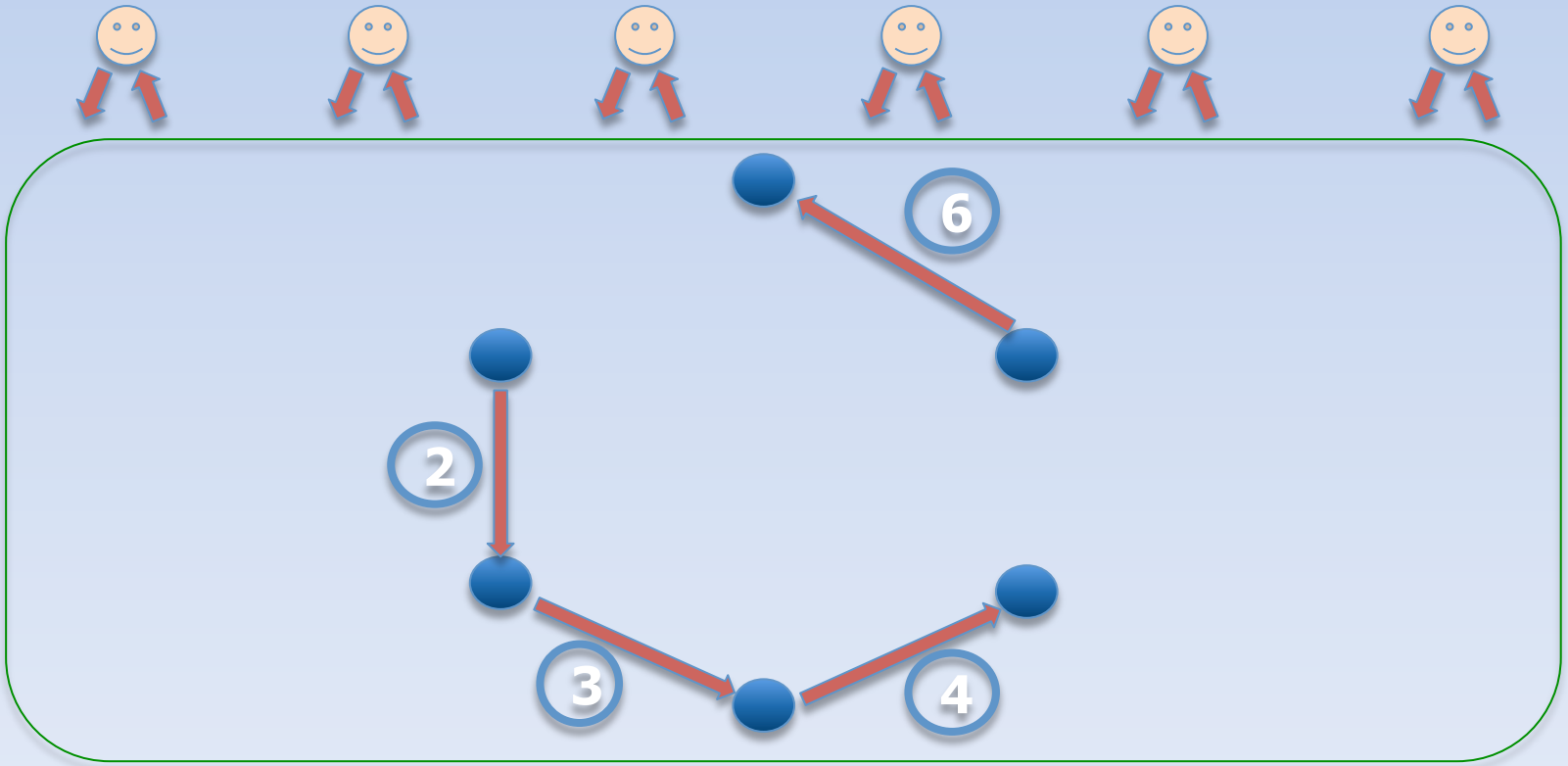
- Interoperability is not only about technical challenges
- Policy and organisational challenges exist:
  - Interoperability entails (global / con-federated) governance and trust
  - Interoperability entails certification and standardisation
  - Interoperability entails sustained funding and support!

Solution

- Quick and dirty: Mediation
- Slow and clean: Standardization

- Top-down, strong-armed will never be final
- Bottom-up, grass-roots will always exist
- Never-ending issue

Agenda: Data Infrastructures



What is “e-Infrastructure”?

“an environment where research resources (**hardware, software and content**) can be readily shared and accessed wherever this is necessary to promote better and more effective research”.

Networks, grids and middleware, comp. resources, exp workbenches, data reps, tools and instruments and the op support that enable global virtual research collabs.

Source: ICT infrastructures for e-Science

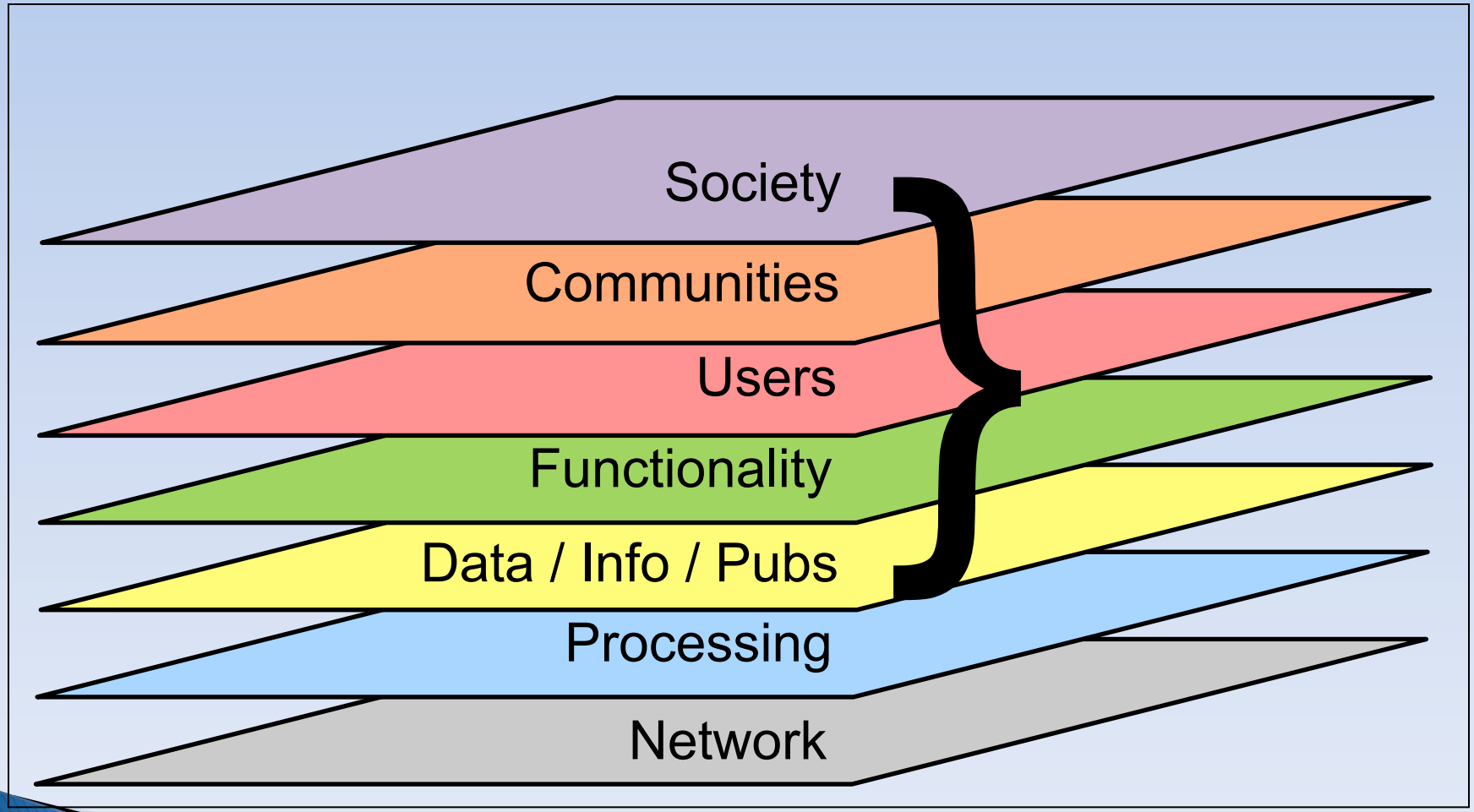
but is “probably” more than that

Wider interpretation: technologies of various kinds for **creating, collecting, annotating, manipulating, storing, finding and re-using information and services** such as those to provide **user support, training, preservation** …

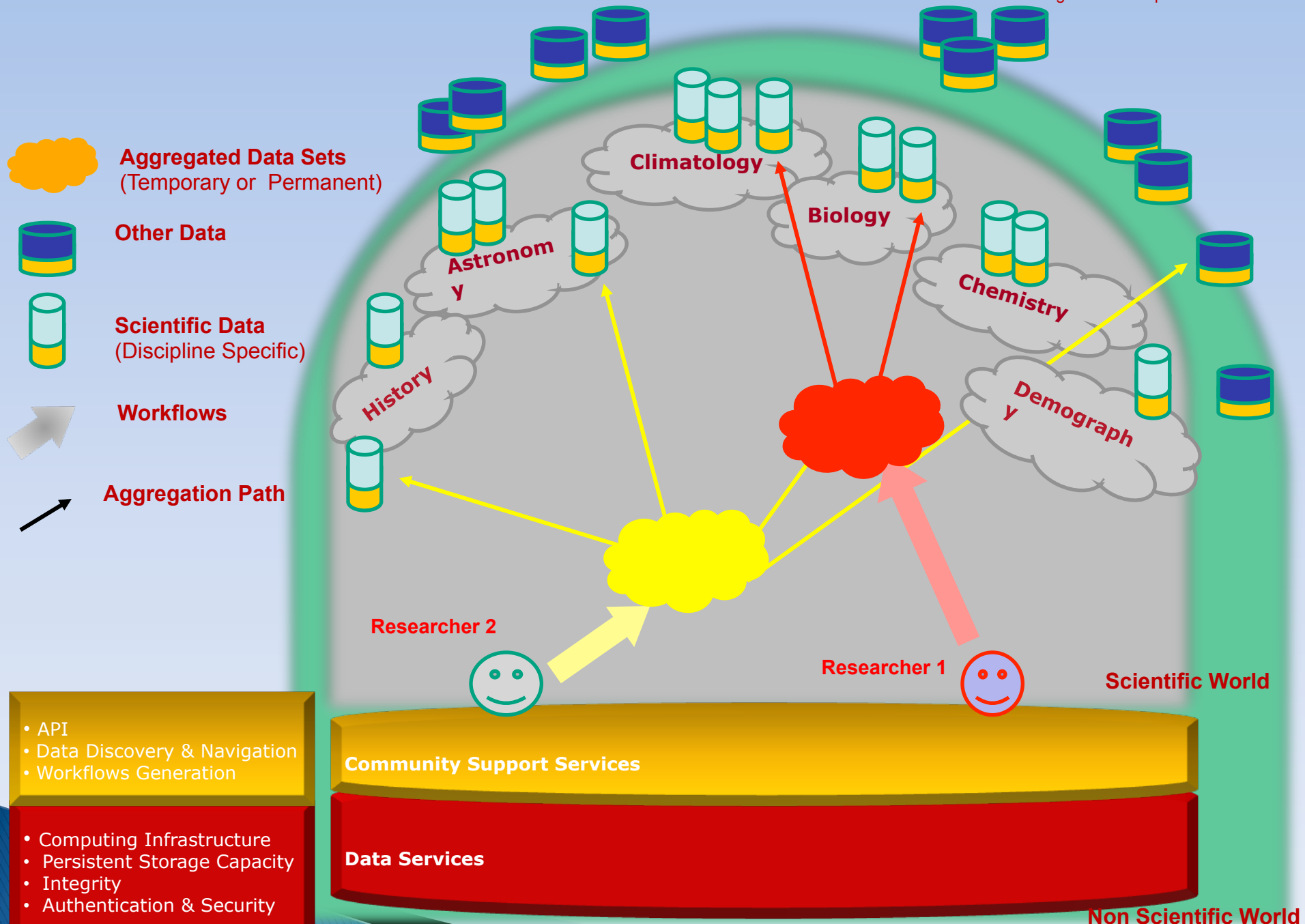
… information resources and tools such as **vocabs, ontologies, rights management and privacy protection systems, and curation**. Several of these resources depend upon manual **human** input.

Source: e-SciDR study

eInfrastructure Layers



Source: High-level Group on Scientific Data



Riding the wave

How Europe can gain from the rising tide of scientific data

Final report of the High Level Expert Group on Scientific Data
A submission to the European Commission

October 2010

raising tide of data…

"A fundamental characteristic of our age is the raising **tide of data** – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge."

Report of the High-Level Group on Scientific Data, October 2010
"Riding the Wave: how Europe can gain from the raising tide of scientific data"

Vision 2030

"Our vision is a scientific e-Infrastructure that supports **seamless access, use, re-use and trust of data**. In a sense, the physical and technical infrastructure becomes invisible and the **data themselves become the infrastructure** – a valuable asset, on which science, technology, the economy and society can advance."

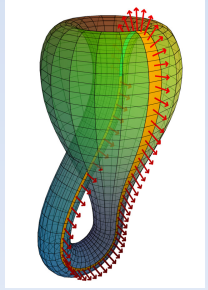
High-Level Group on Scientific Data

"Riding the Wave: how Europe can gain from the raising tide of scientific data"

scientific Information *con·tin·u·ums* allowed in a "new world"

BETWEEN

- experimental **data** and **publications** (new paradigm)
- different **scientific disciplines** (multidisciplinary)
- **past**, **present** and **future** (preservation)
- **different institutions** (organisation)
- **humans** and **computers** (e-Infrastructure)
- **research** and **education** (public mission)
- **current critical needs** and **visionary environments** (evolution)



Reference: Klein Bottle with Moebius Band. Reference to article "Imaging maths - Inside the Klein bottle" at http://plus.maths.org/issue26/index.html. The Klein bottle is a non-orientable surface found by Felix Klein in 1882 while working on a topological classification of surfaces.

Conclusions

- 4th Paradigm in Scientific Research
- Data size and complexity fundamental
- Data management and analysis research challenges
- Interoperability a huge challenge
  - Data, services, policies, communities, …
- Global Research Data Infrastructures

The name is Science!
d-Science!

Thank you!