Action Theme 3 Scoping Workshop

Belmont Forum
Data Management and e-Infrastructure CRA
Paris, November 28-29, 2016

Contacts:
Jean-Pierre Vilotte vilotte@ipgp.fr
Mark Asch mark.asch@u-picardie.fr
Tsair-Fuh Lin (MOST) tflin@mail.ncku.edu.tw
Michael Vogelsanger (JST) belmont@jst.go.jp

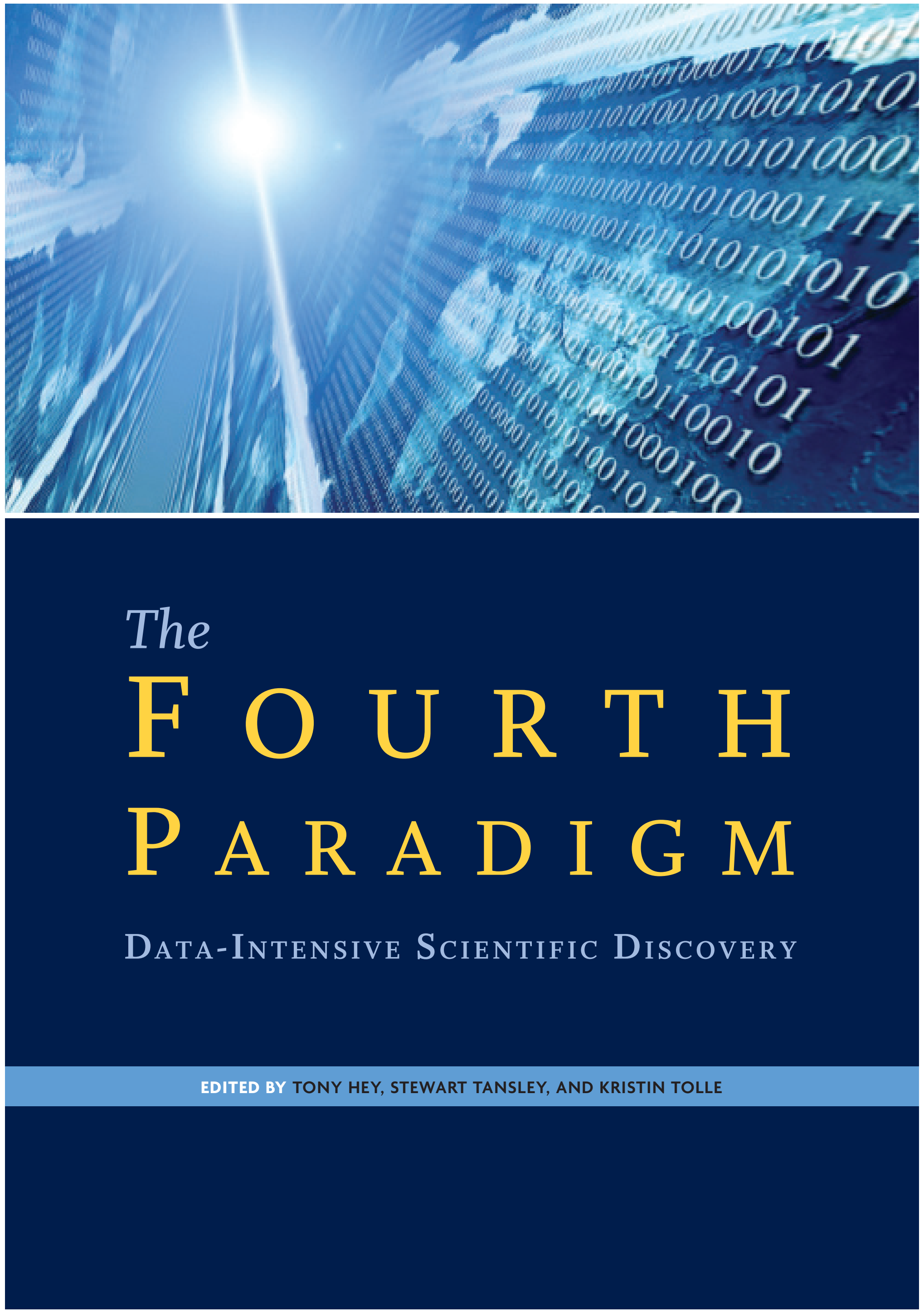
Agenda

DAY 1	Monday 28th November		2016
Session I	**9:00-12:45 Opening and Keynotes**		
9h00-9h30	Opening: Welcome, Objectives, Agenda	R. Samors, M. Asch, J-P. Vilotte	CRA, Belmont Forum
9h30-10h15	ESGF: a federation for data analysis	S. Denvil	ESGF
10h15-10h45	**Coffee Break**		
10h45-11h30	DIAS: Interoperable and interdisciplinary data	A. Kawasaki, E. Ikoma	U. Tokyo, Japan
11h30-12h15	CMIP: data and model inter comparison	S. Joussaume	IPSL, France
12:15-12:45	Open Discussion		
12h45-14:00	**Lunch Break**		
Session II	**14:00-18:00 Projects and Discussion**		
14:00-15h30	Project presentations identifying gaps and barriers (see next page)		
15h30-16h00	**Coffee Break**		
16h00-18h00	Project presentations identifying gaps and barriers (see next page)		
18h10-20h00	**Cocktail Dinner**		

Context

- Data Explosion - in volume and complexity (large instruments, monitoring networks, large simulations)
- Open data / Open science context (toward reproducible science).
- Inter- and Trans-disciplinary data use for environmental change
- Critical need to bring together application domain scientists and "digital" scientists (computer scientists, data analysts, statisticians, mathematicians).
- Need for reliable decision-making tools and decision-supporting predictions (see DMIPs), particularly for risk/hazard policies for prevention and mitigation.

Context: “big-data paradigm”



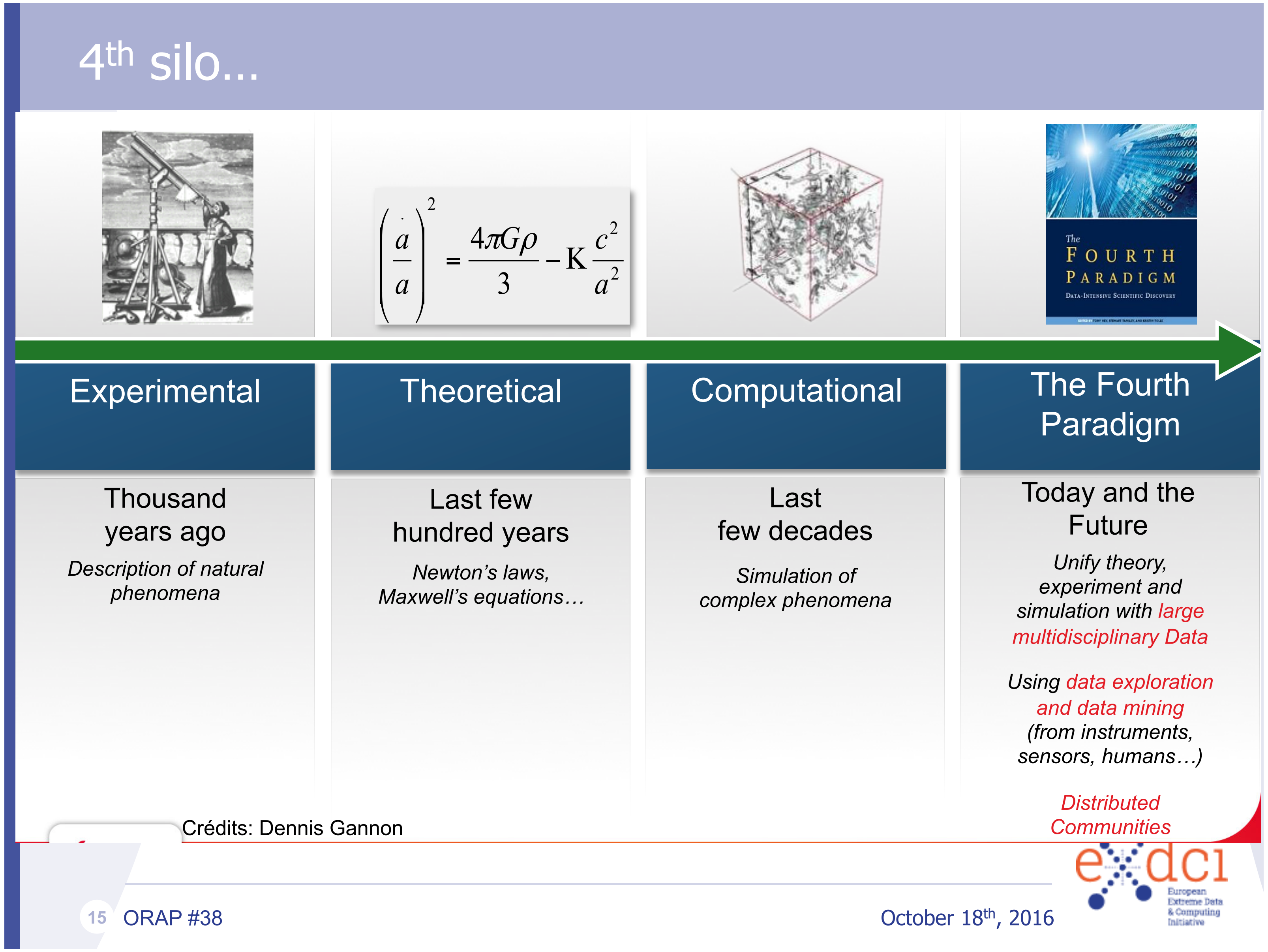
The

FOURTH

PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



4th silo…

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

The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY

Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago *Description of natural phenomena*	Last few hundred years *Newton’s laws, Maxwell’s equations…*	Last few decades *Simulation of complex phenomena*	Today and the Future *Unify theory, experiment and simulation with large multidisciplinary Data* *Using data exploration and data mining (from instruments, sensors, humans…)* *Distributed Communities*

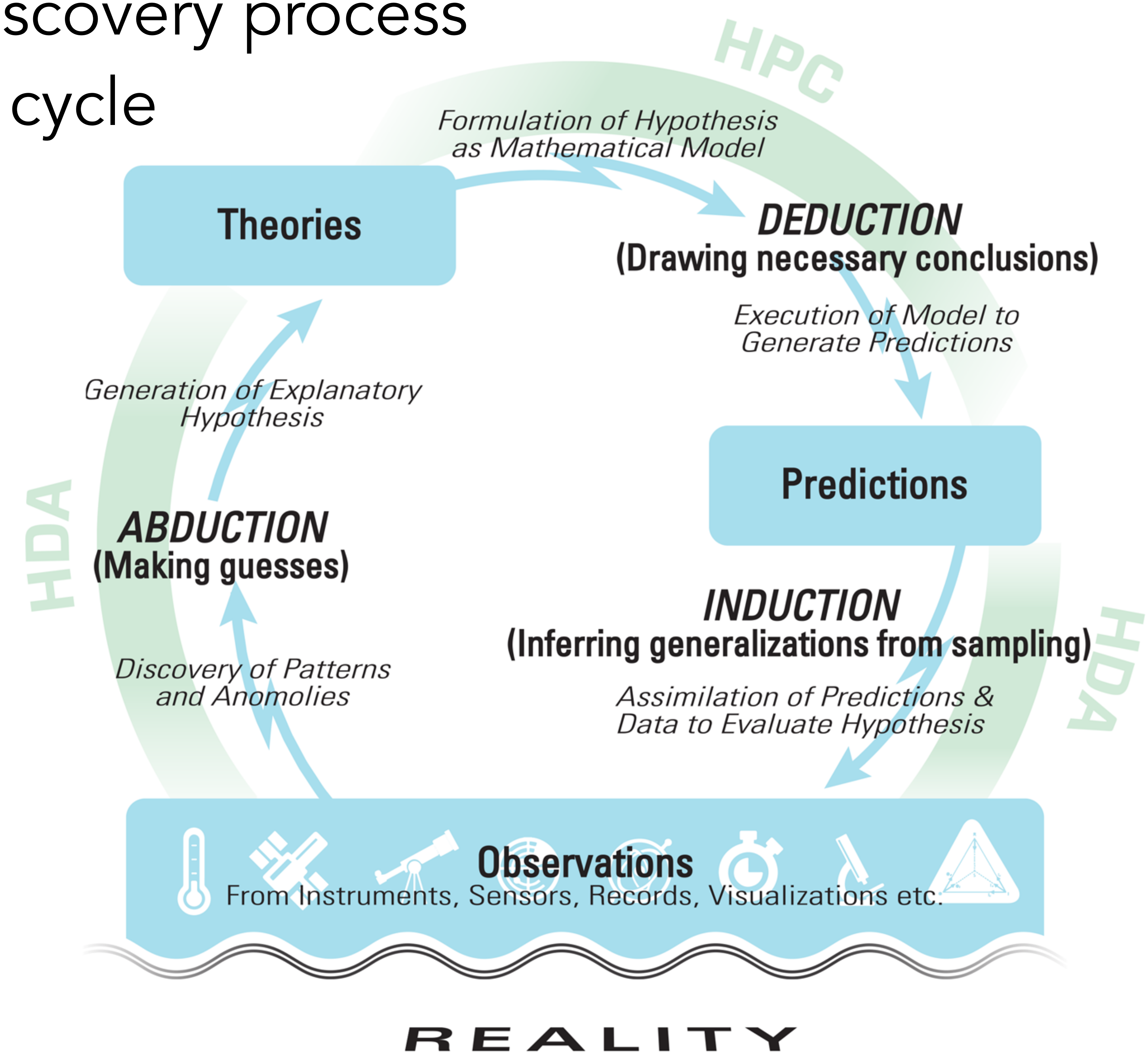
Crédits: Dennis Gannon

15 ORAP #38 October 18th, 2016

exdci European Extreme Data & Computing Initiative

Context: a new paradigm

The scientific discovery process
= the inference cycle



HPC

Formulation of Hypothesis as Mathematical Model

Theories

DEDUCTION
(Drawing necessary conclusions)

Execution of Model to Generate Predictions

Predictions

INDUCTION
(Inferring generalizations from sampling)

Assimilation of Predictions & Data to Evaluate Hypothesis

HDA

Generation of Explanatory Hypothesis

ABDUCTION
(Making guesses)

Discovery of Patterns and Anomolies

HDA

Observations
From Instruments, Sensors, Records, Visualizations etc.

REALITY

Context: convergence

•BDEC	see exascale.org
•EOSC - see http://ec.europa.eu/research/openscience/	

European Commission

Realising
the European
Open Science Cloud

First report and recommendations
of the Commission High Level Expert Group
on the European Open Science Cloud

Research and Innovation

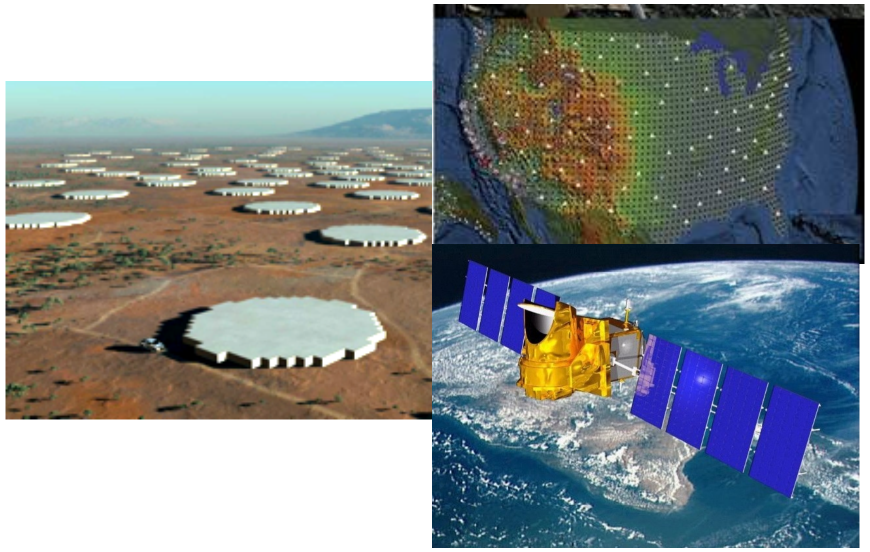
Workshop outcomes

- Input for a draft call text, based on our discussions, covering:
  - Common data and e-infrastructure gaps and barriers where collaboration between existing projects, sharing common research practice, would be beneficial and extensible to other projects.
  - Data and Model Intercomparison Projects (DMIPs).
  - An initial evaluation framework.

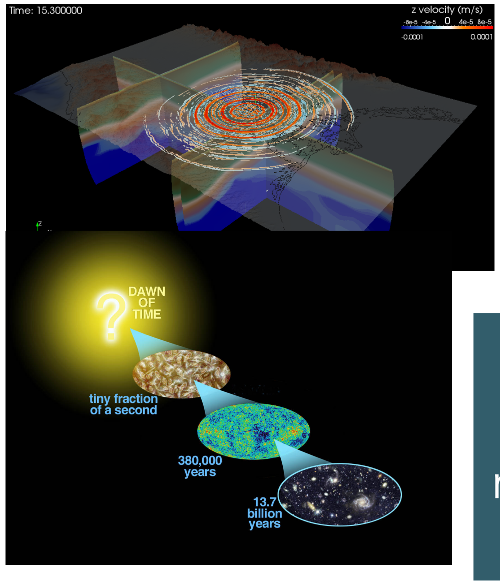
AT3 Timeline

- Early November 2016: Feedback from Belmont Forum in Doha.
- Late November 2016: Scoping workshop in Paris.
- **March 2017:** Publication of call.
- June. 2017: Proposal submission deadline.
- Sept./Oct. 2017: Announcement of call awardees.
- **Late 2017/Early 2018**: Launch of the funded projects.

Not a single dimensional challenge



Data Generation



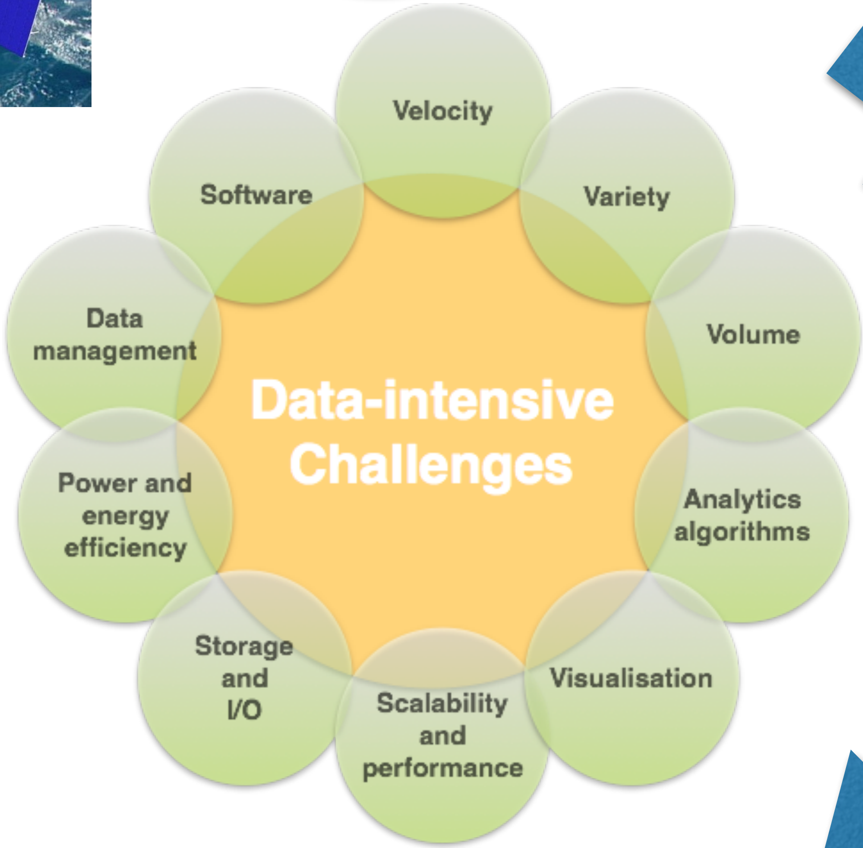
Data management

Data reduction query

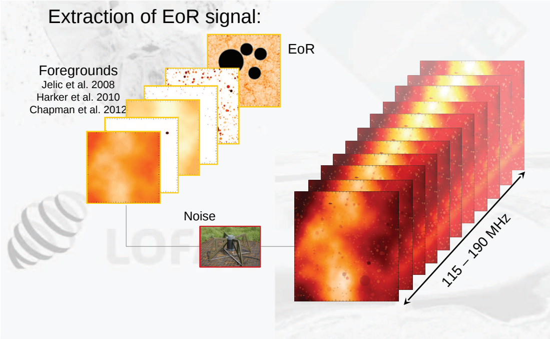
Data visualisation

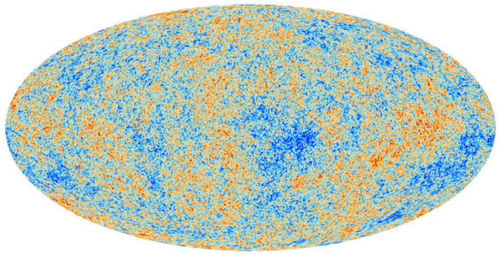
Data and method sharing

data processing, transformation

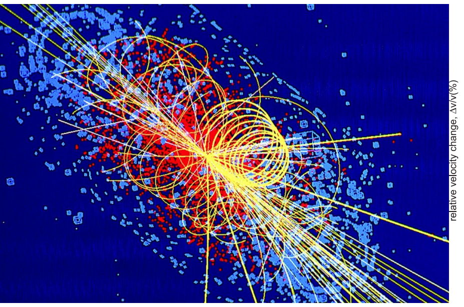


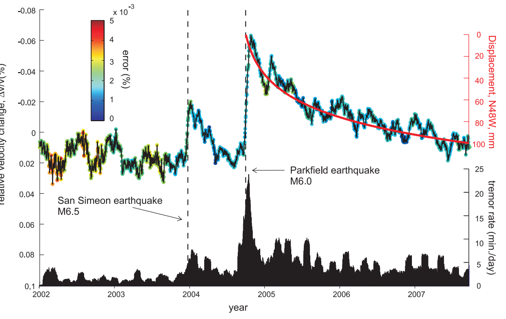
Discovery, Insights, Prediction

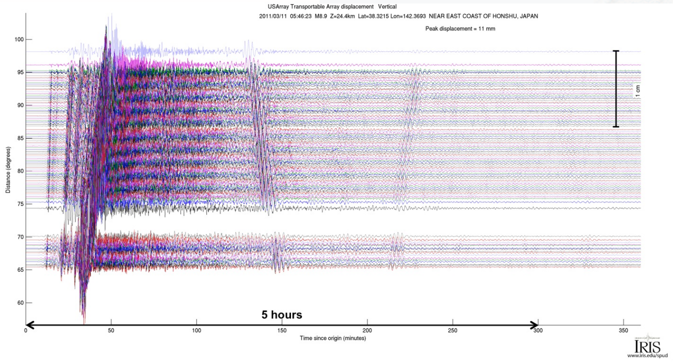




Data analytics, Mining, unsupervised learning




relative velocity change, Δv/v (%)
error (%)
Displacement, N48W, mm
tremor rate (min/day)
San Simeon earthquake M6.5
Parkfield earthquake M6.0
year
2002
2003
2004
2005
2006
2007



Adapted from Choudhary

Data -> Extraction/cleaning -> Integration/aggregation -> Learning models -> trigger / question -> predict

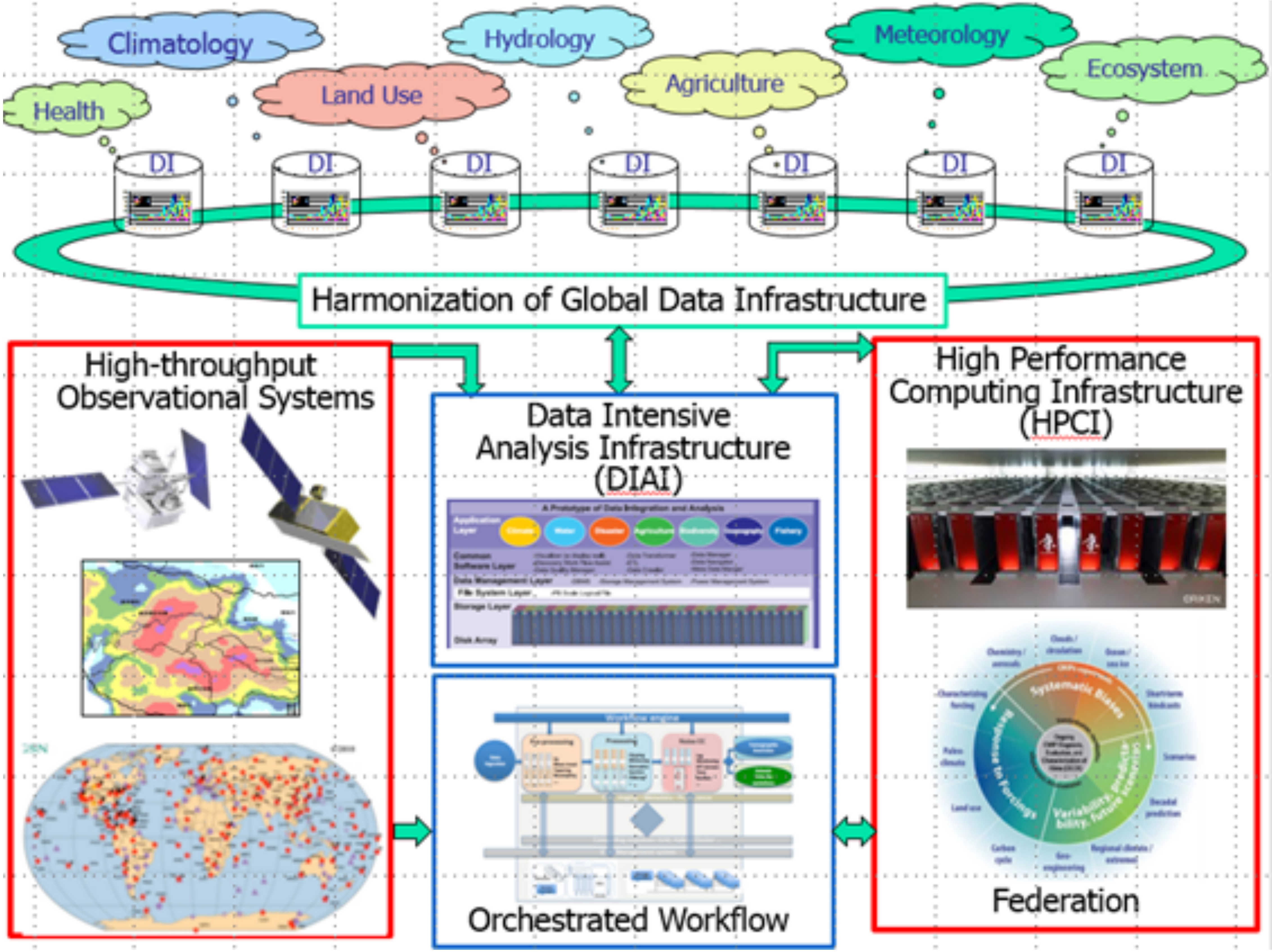
A research-driven variety of infrastructures

Ashby's Law of
Requisite Variety

Only variety absorbs variety

Federating autonomous infrastructures

A research-driven strategy .....



Climatology
Hydrology
Meteorology
Ecosystem
Agriculture
Land Use
Health
DI
DI
DI
DI
DI
DI
DI
Harmonization of Global Data Infrastructure
High-throughput Observational Systems
Data Intensive Analysis Infrastructure (DIAI)
High Performance Computing Infrastructure (HPCI)
Orchestrated Workflow
Federation

after Toshio Koike

E-infrastructure enabling interdisciplinary studies

- Interdisciplinary and trans-disciplinary research conducted around problems rather than in silos.
- Drivers for the co-evolution of data-intensive e-infrastructure system,
- Agreement around interchange and inter-operability of data and metadata.

Action articulated around a step-by-step and multi-level strategy:

1. **Survey, analyse and promote collaboration between cross-disciplinary case studies**:

   *Accelerate **inter-disciplinary and trans-disciplinary** – natural, social and economic - global change research and improve the quality of decision by **enabling effective use and valorisation of data** (international monitoring and observation systems, large-scale earth systems simulation);*

   *Emphasise "**going the last mile**", i.e., transforming scientific knowledge into actionable information for society and achieving influence;*

   *Increase **quality of science and decision making** through relevant and standardised framework for collaborative interdisciplinary data-and-models inter-comparison.*

2. **Distill and collate findings to inform the BF strategy**

   toward a holistic cross-disciplinary data infrastructure, training and "intellectual ramps" in harmony with interdisciplinary and trans-disciplinary research practice.

A multi-level approach: research domain specialists, data scientists, IT researchers, data-aware engineers, and critical stakeholders, i.e., including infrastructure providers.

Cycle-up series: **scoping workshops, competitive call for interdisciplinary case-study collaboration**

Start with existing supported projects by Belmont Forum and other international initiatives.

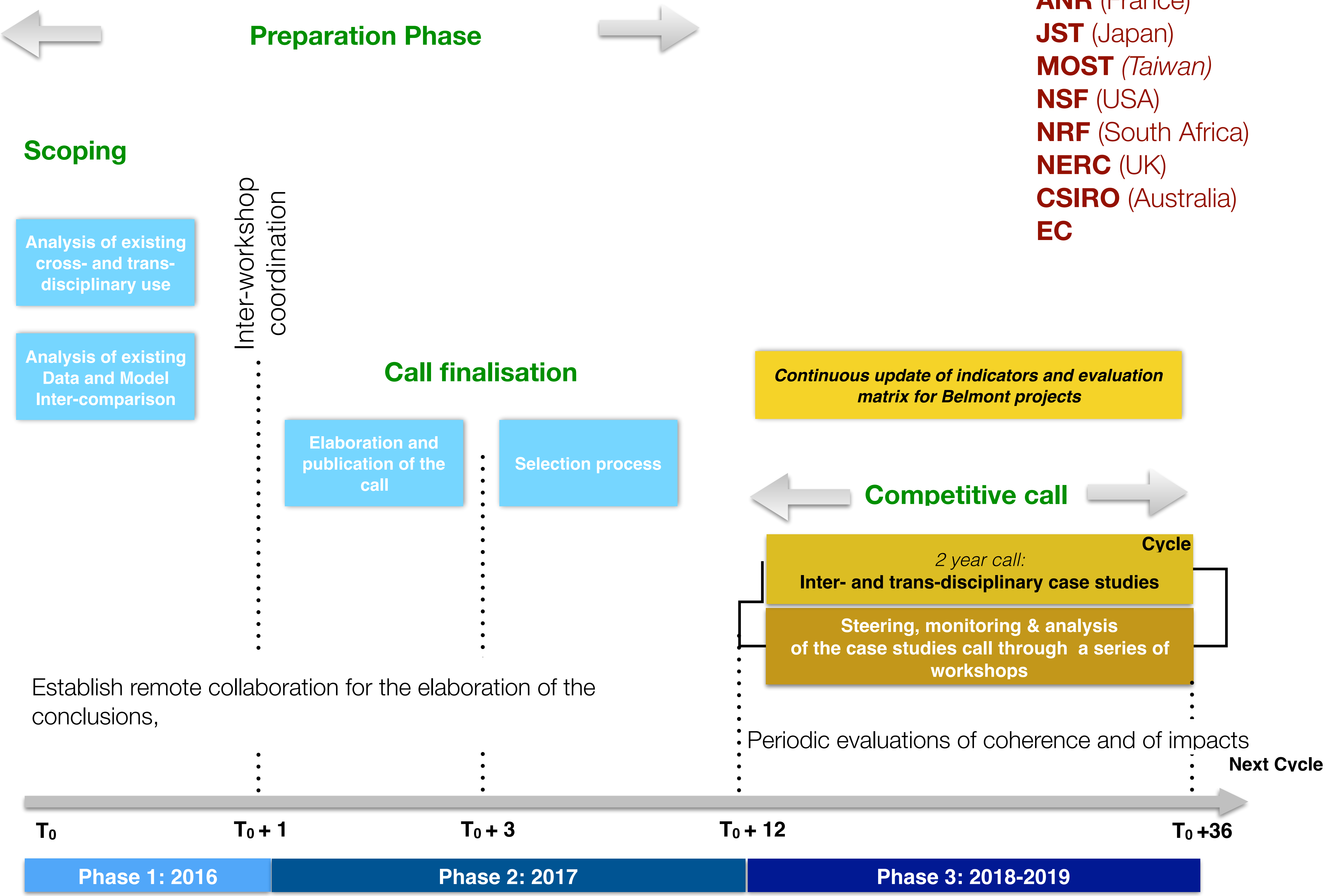
Action Theme 3: Objectives

1. **Identify** a first set of active **interdisciplinary data-use projects and use-cases** federating data- and e-infrastructure for environmental and global change problems and **foster coordination/collaboration between some of them** around common research practices with the aim to develop **mutual understanding and address collaboratively well-identified gaps and barriers.**

2. **Identify** large-scale Data and Model Inter-comparison Projects that are relevant for global change and natural risk research, and **foster coordination and collaboration between some of them** with the aim to develop **mutual understanding and address collaboratively well-identified theoretical and practical issues.**

3. Through the above two, **inform the data- and e-infrastructure policy** with bacon of best practices responding to concrete issues, and the **human capacity action theme** with well identified needs in **training and "intellectual ramps" to be developed collaboratively**.

ANR (France)
JST (Japan)
MOST *(Taiwan)*
NSF (USA)
NRF (South Africa)
NERC (UK)
CSIRO (Australia)
EC



Preparation Phase

Scoping

Analysis of existing cross- and trans-disciplinary use

Analysis of existing Data and Model Inter-comparison

Inter-workshop coordination

Call finalisation

Elaboration and publication of the call

Selection process

Continuous update of indicators and evaluation matrix for Belmont projects

Competitive call

Cycle

2 year call:
Inter- and trans-disciplinary case studies

Steering, monitoring & analysis of the case studies call through a series of workshops

Establish remote collaboration for the elaboration of the conclusions,

Periodic evaluations of coherence and of impacts

Next Cycle

T_0 **T_0 + 1** **T_0 + 3** **T_0 + 12** **T_0 +36**

Phase 1: 2016 **Phase 2: 2017** **Phase 3: 2018-2019**

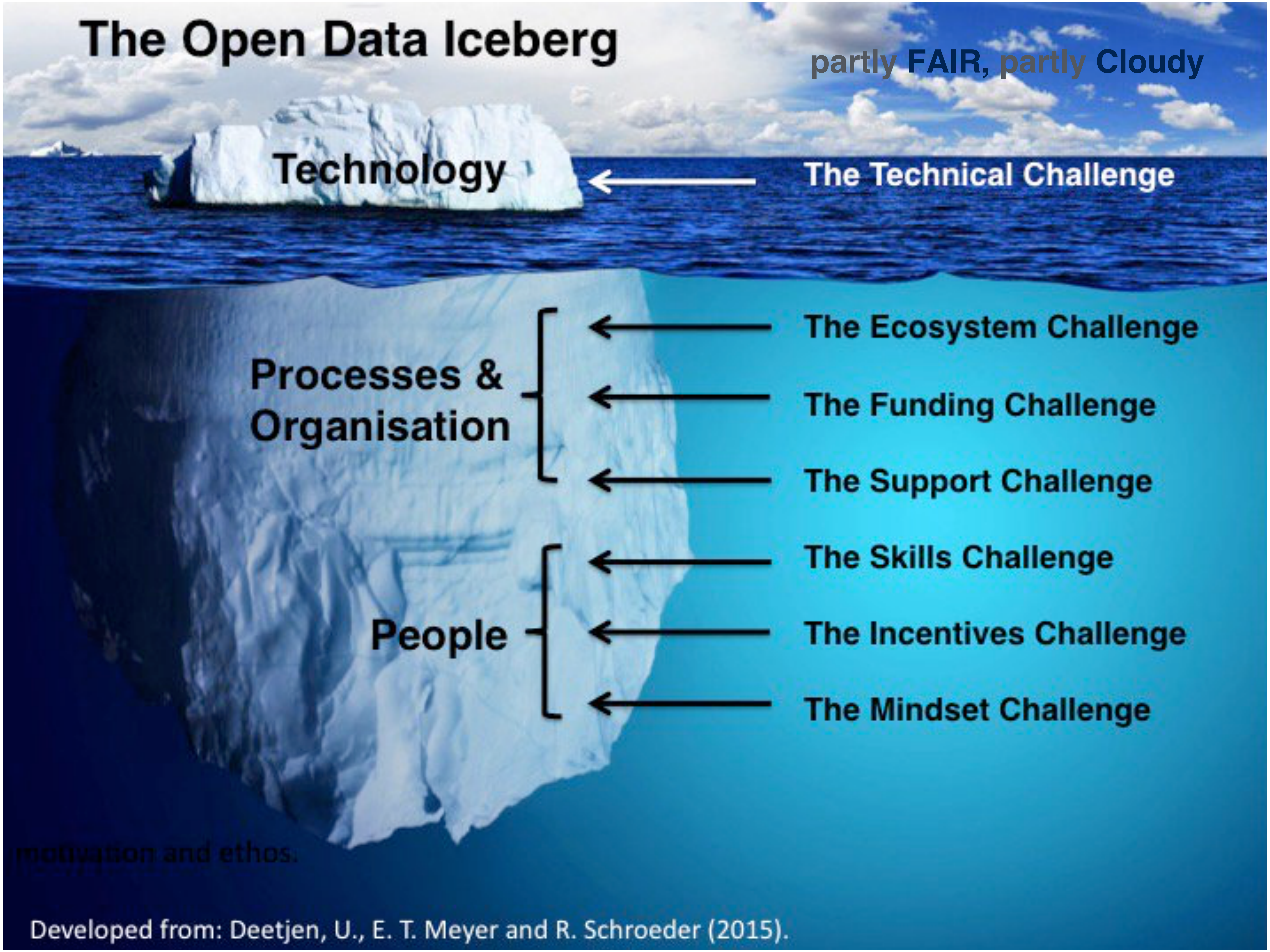
Targeted Projects of the Belmont call

- Any project with **well-identified** e-infrastructure, data analysis workflows and data management related problems.

- **Interdisciplinary** and trans-disciplinary projects where:
  - ***Big Data and 4Vs*** issues with ***multi-type and multi-disciplinary data*** are present;
  - ***Findable, Accessible, Interoperable, Reusable Data*** concepts are present;
  - ***Data management and Data stewardship*** are present;
  - ***Environmental, Social and Economical challenges*** are present;
  - Needs to ***federate data and compute infrastructures*** to address the above are well identified and timely to address collaboratively.

The Open Data Iceberg

partly FAIR, partly Cloudy



Technology ← The Technical Challenge

Processes & Organisation
← The Ecosystem Challenge
← The Funding Challenge
← The Support Challenge

People
← The Skills Challenge
← The Incentives Challenge
← The Mindset Challenge

motivation and ethos.

Developed from: Deetjen, U., E. T. Meyer and R. Schroeder (2015).

Technical challenges

Challenges for federated data-analysis platform

Foster international collaboration and community building toward know-how exchange for

- Storage and computing architecture in support of massive and complex interdisciplinary data
- Streaming data analysis workflow orchestrating analysis of distributed data sources with pervasive provenance systems
- Network-based and provenance-based data movement between different and distributed data and computing sources honouring data and AAI policies
- Concurrent data access and data representation for data-intensive analysis
- Adding access, data analysis and visualisation services on top of the data
- Energy and Green technology challenges
- Collaboration with private providers: public cloud and others

Bridging the gap between multi-type and multi-disciplinary data

- Data stewardship, data and metadata formats, data exchange protocols
- Credential and interoperability at the data level
- Implementing FAIR data principles
- Structured/unstructured data
- Dealing with and assimilating different data spatial and temporal scales
- Strengthen the use of data by and from other communities especially socio-economical communities

Data Model Inter-comparison - validation - prediction

- Identify trans-national expertise and beacons of good practices
- From model to coupled models framework
- Bridging scientific-driven and policy-driven concerns into a framework
- Extension to other socio-economic and health issues
- Foster standardisation of protocols and methods across disciplinary

The ecosystem challenges

Federating autonomous data and compute infrastructures ecosystems

- Research-driven strategy
- Hourglass architecture individualising and isolating layers supporting different concerns
- Data policies: across different disciplines, data providers and countries
- Involving data and compute resources providers, national science agencies

Federating data policies across domains and national boundaries

- AAI and data licensing

The funding and support challenges

The incentive challenges

Data publication and citation
Data plan for cross national boundary projects
Intellectual ramps

Skills and mindset challenges

Data literacy
Data analytics literacy including statistics and machine learning