

Enabling Data-Intensive Federations

Consultation on Cloud Computing Research Innovation Challenges for
WP 2018-2020

7 November 2016

Malcolm Atkinson

and

Alex Hardisty

Malcolm.Atkinson@ed.ac.uk

HardistyAR@cardiff.ac.uk



Environmental Research
Infrastructures Providing Shared
Solutions for Science and Society

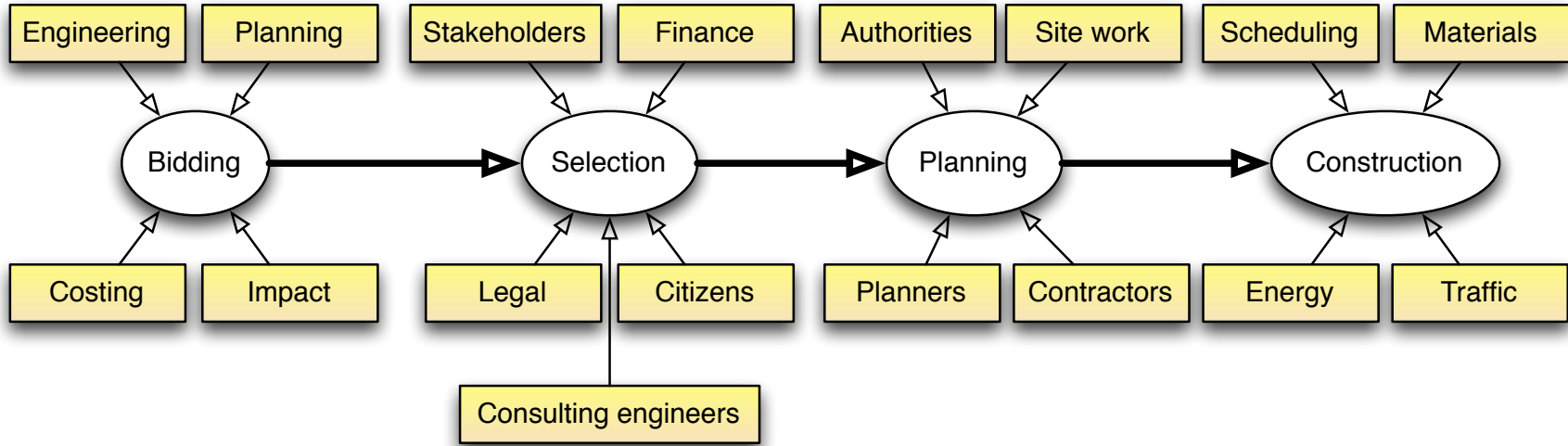
Additional Research Goal

- Improve our ability to **exploit diverse information resources**
 - Independently produced and managed
 - Active data
 - Pooling insights from experts in many fields and organisations
 - Innovation and productivity
 - Across Europe
- **Legal, ethical and commercial constraints visibly met**
 - Rules combined, mapped and enforced automatically
 - Validation of long-term compliance

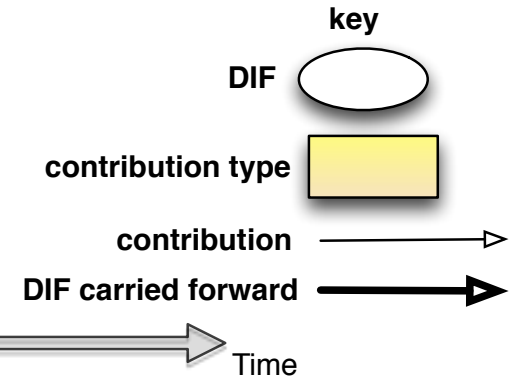


Data-Intensive Federation evolution

Civil engineering consortia

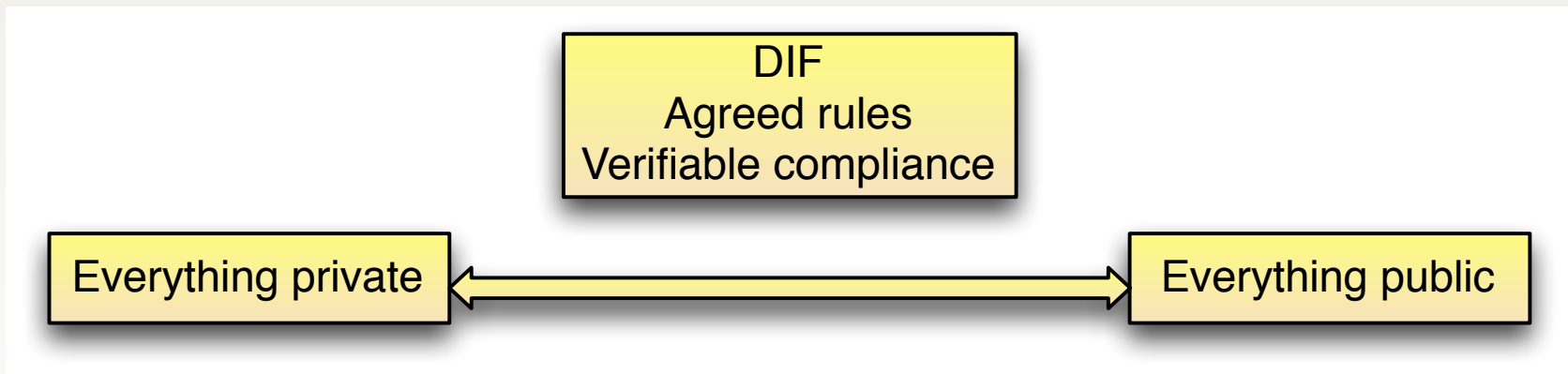


Shows how much can be achieved!

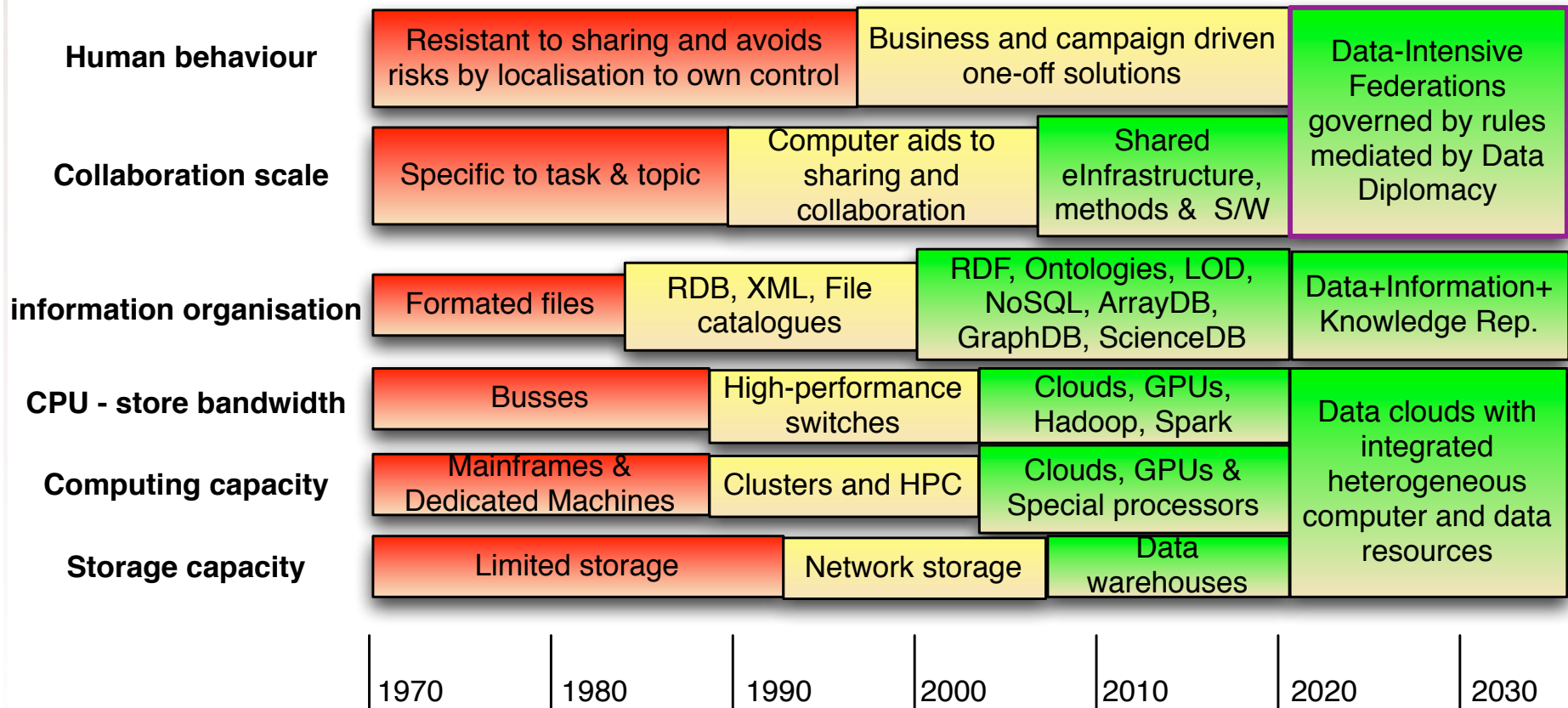


Data Intensive Federations

- Pooling effort to present data and capabilities from diverse sources
- A productive consistent work environment
 - An evolving cluster of data, analysis and simulation topics
- Respecting data owners' and consortium rules on data use*



Expanding our capacity to exploit diversity



Lowering barriers generated by group identity

DIF research is urgently needed

Today's DIF

- Demand exceeding capacity of experts

Unsustainable

- ★ Number of DIFs growing rapidly
- ★ Complexity of DIFs growing rapidly
- ★ Speed of creation critical
- ★ Acceptable rules requires **trust in their meaning and implementation**

We need

● Re-usable methods and architectures for DIF

- ★ **Rules** logically expressing intention
- ★ **Frameworks** automating distributed enforcement
- ★ Compliant across data cloud, integrated data services and specialist platforms



Example rules

- Every *contributing agency* receives *reports* of their *data use*
 - Including *indirectly* from caches and after *preserving transformations*
 - When *combined* with other inputs if its *contribution* > 5%
- Data produced using ESFRI research infrastructure *X*
 - remains *exclusive* to *originators* for *M* months after *timer starts*
 - then it is made FAIRly available
- Data from *consented study participant P*
 - must be *removed* throughout the system if *P withdraws consent*
- Data from *consented study CS*
 - may only be used for the *purposes covered* by its consent
- Only the *requestor* may know that *simulation S* was performed
 - with *parameters PAR*

Italic entities need agreed recognition methods



Data-Intensive Research Strategy

- **Case studies**
- **Rule notation development**
 - Human comprehension
 - Automation and reasoning
- **Automated *distributed* rule compliance**
 - Integrated with **platform mappings** and semantic translation
 - **Sustainable zones of trust**
 - Spanning independent heterogeneous platforms
- **Verification of compliance**
- **Data diplomacy rule manipulation tools**
- **Starter kits** and tailorable **frameworks** to ease path to adoption

*Environmental Research Infrastructures
a demanding example to pioneer
solutions for ERA competitive use*

*Input RDA, CERIF, CKAN, W3C standards, ... and rule concept ontologies
Link with data clouds, data curation and data services*

Data Cloud Requirements

- Engage with rules
 - Articulate their own
 - Develop trust in their handling of rules
 - Educate data diplomats
- Develop strategies for making rules pervasive
 - Consistent and persistent compliance
 - Data should not escape from protective wrappers
- Engage with DIF treaties and their amendment
 - Relationships preserved as platforms evolve
 - Correct identification of rule concepts
 - Inter-platform communication complies with rules



Thank you

Questions and Comments please

M Atkinson (University of Edinburgh)

Malcolm.Atkinson@ed.ac.uk

A Hardisty (Cardiff University)

HardistyAR@cardiff.ac.uk

Supported by:

Peter Wittenburg, RDA TAB Member and HLEG on Scientific Data ("Riding the Wave")

Keith Jeffery, PaaSage and HOLACloud projects and RDA (co-chair 4 groups)



H2020 Project



Project Number: 654182

Spare slides for Questions



H2020 Project



Project Number: 654182

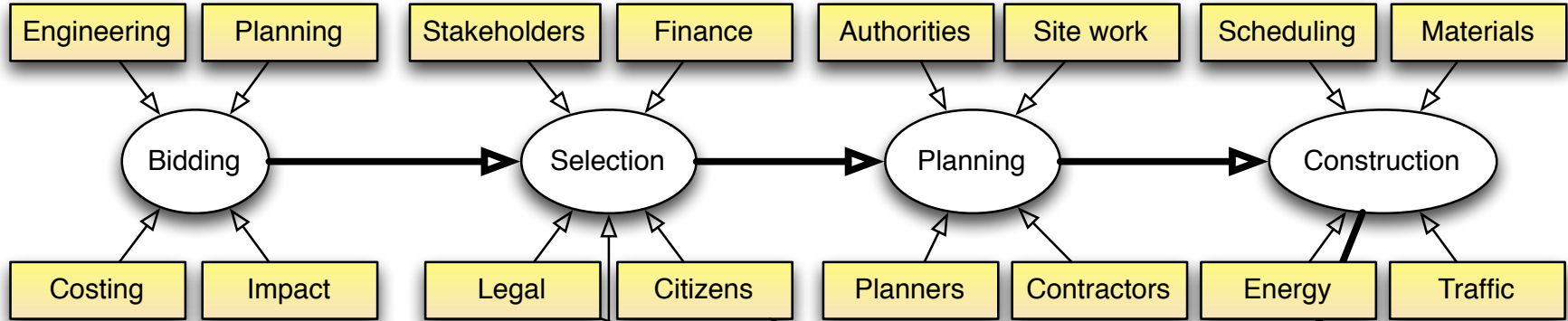
Data Intensive Federation elements

- Organisational patterns of data diplomacy
- Concepts, logic and rules for trusted use of data
- Automated rule implementation
- Independent validation of rule compliance
- Workbenches for data diplomacy
- Starter kits

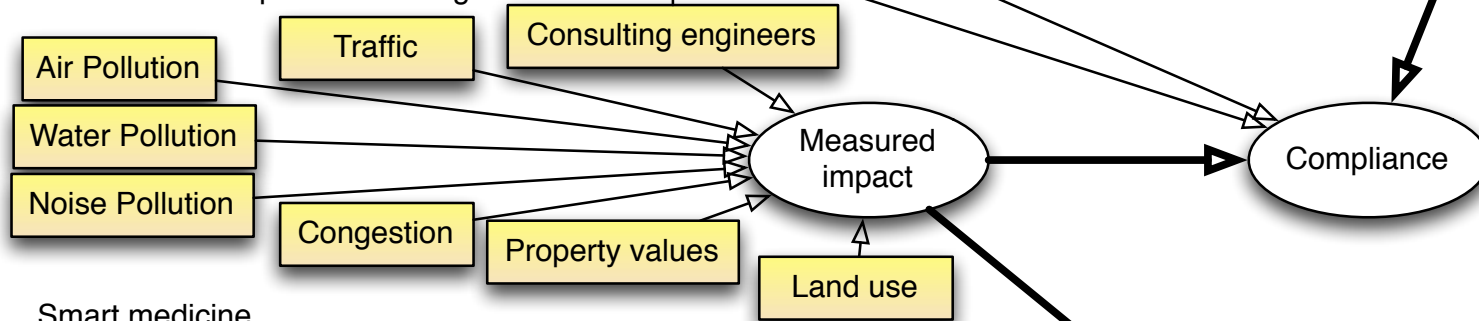


DIF evolution and mergers

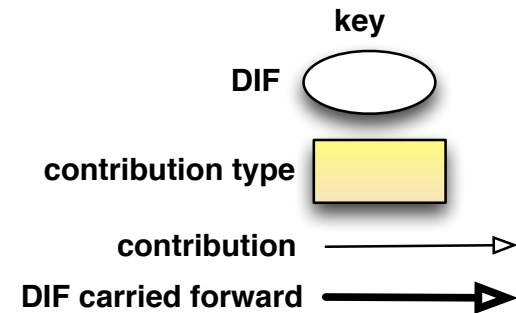
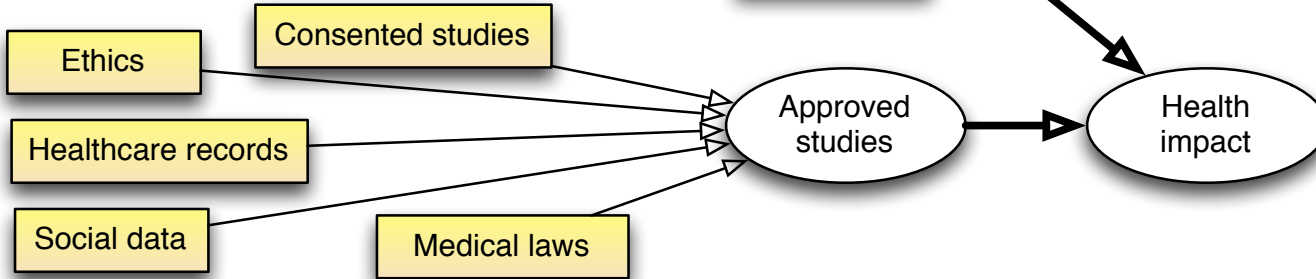
Civil engineering consortia



Environmental impact monitoring



Smart medicine

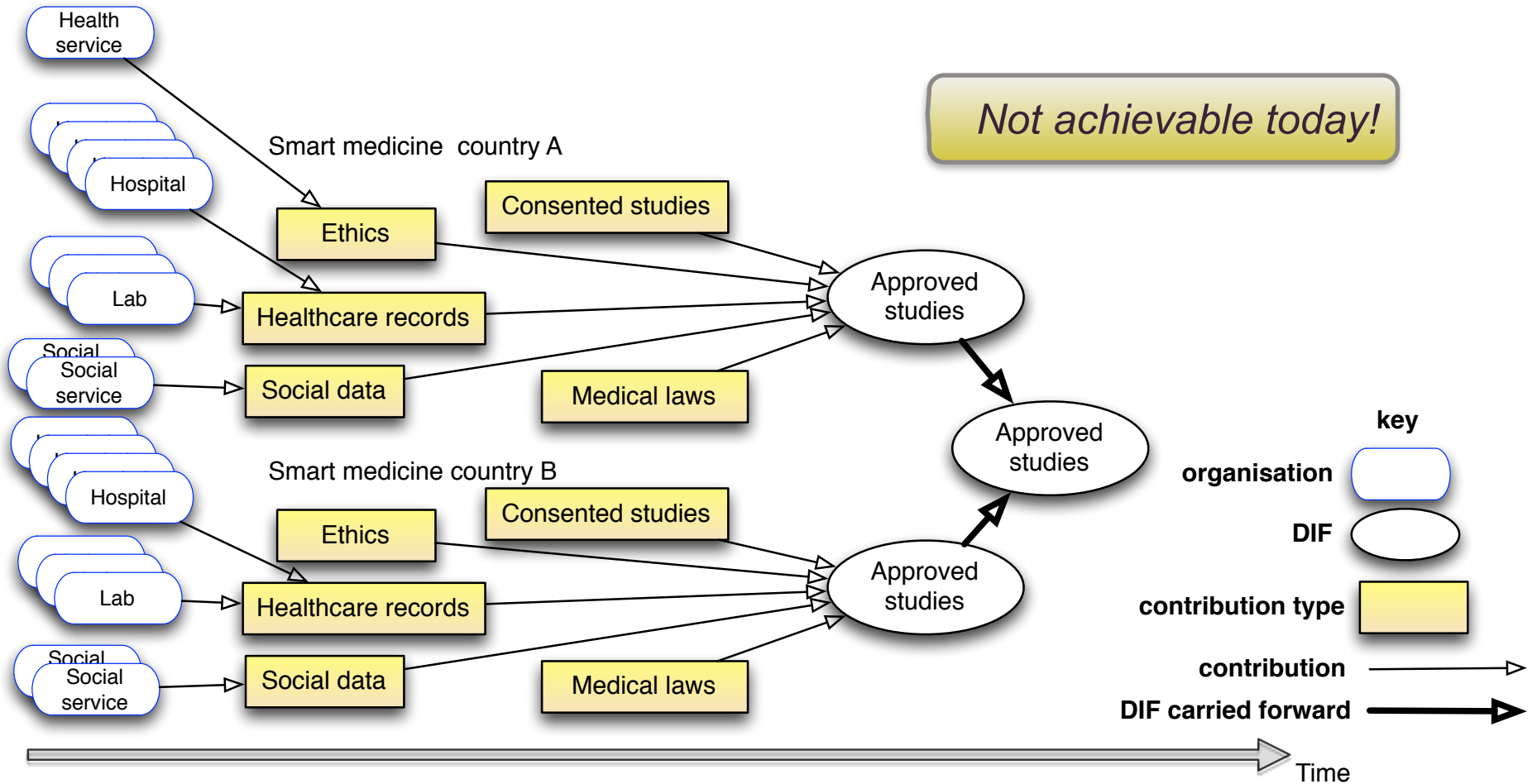


Example Data Intensive Federations

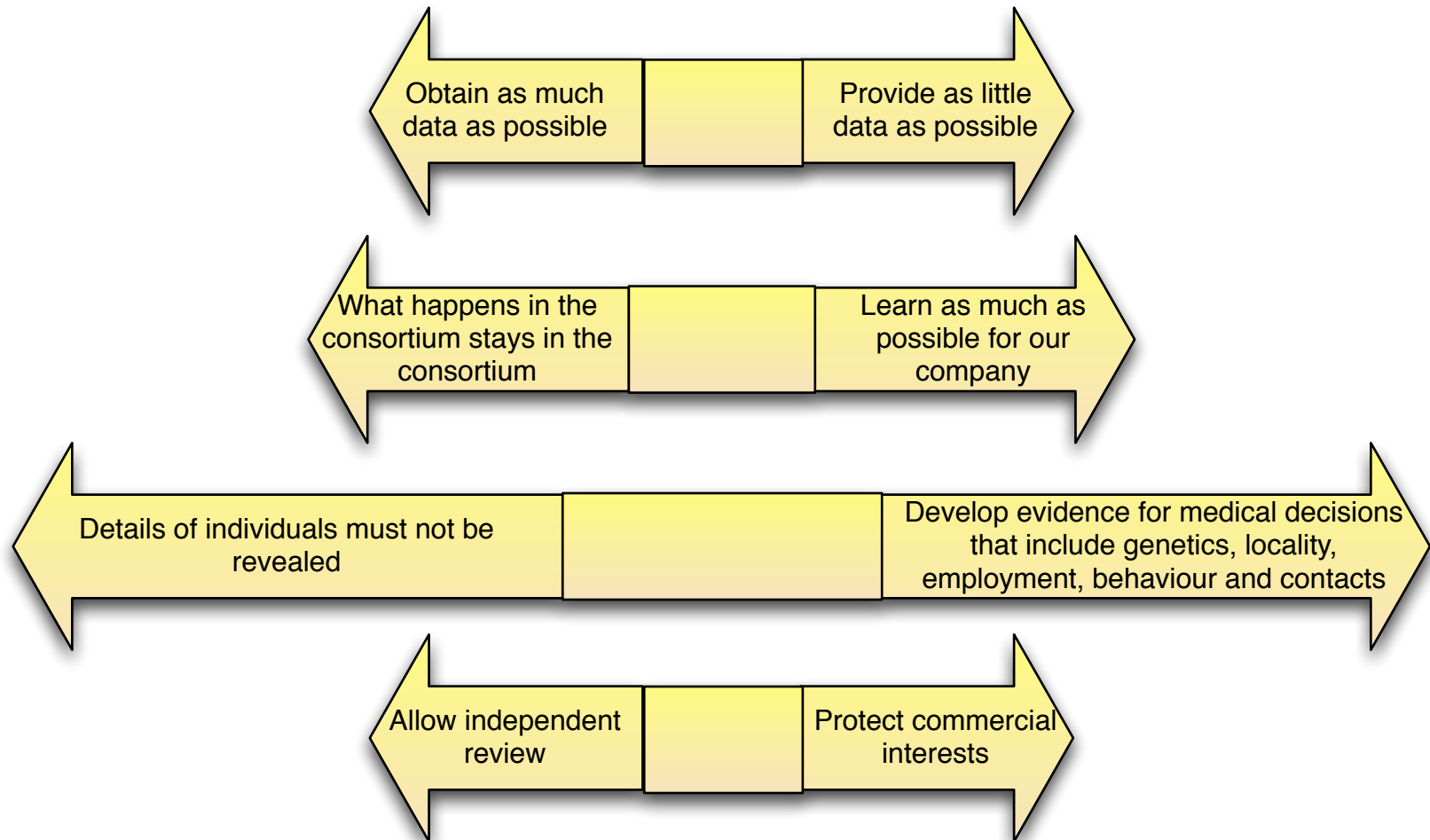
- **Pooling resources** to guide response to environmental disasters
- **Environmental and land-use** planning and management
- **Campaigns** to deliver engineering products
- **Research Infrastructures** harnessing many observational programs
- **Medical research** gaining breadth and statistical scale
- **Monitoring** environmental, health, economic and social impact



Data-Intensive Medical Research Federation



Data Diplomats resolve tensions



Diplomatically crafted rules

- Abstract expression of intent
 - Human comprehensible
 - Detailed logical reasoning
 - Easily composed, refined and extended
 - Socio-economic and technical concepts and verbs
- Automation wrapping users' requests in compliance clothing
 - Integrated with mapping to cloud and specialised platforms
 - Translations for semantic integration
- Effective and trusted verification

*Collected as
information sharing
treaties!*

