# Pushing the Limits of Data Powered Research

Malcolm Atkinson

Malcolm.Atkinson@ed.ac.uk

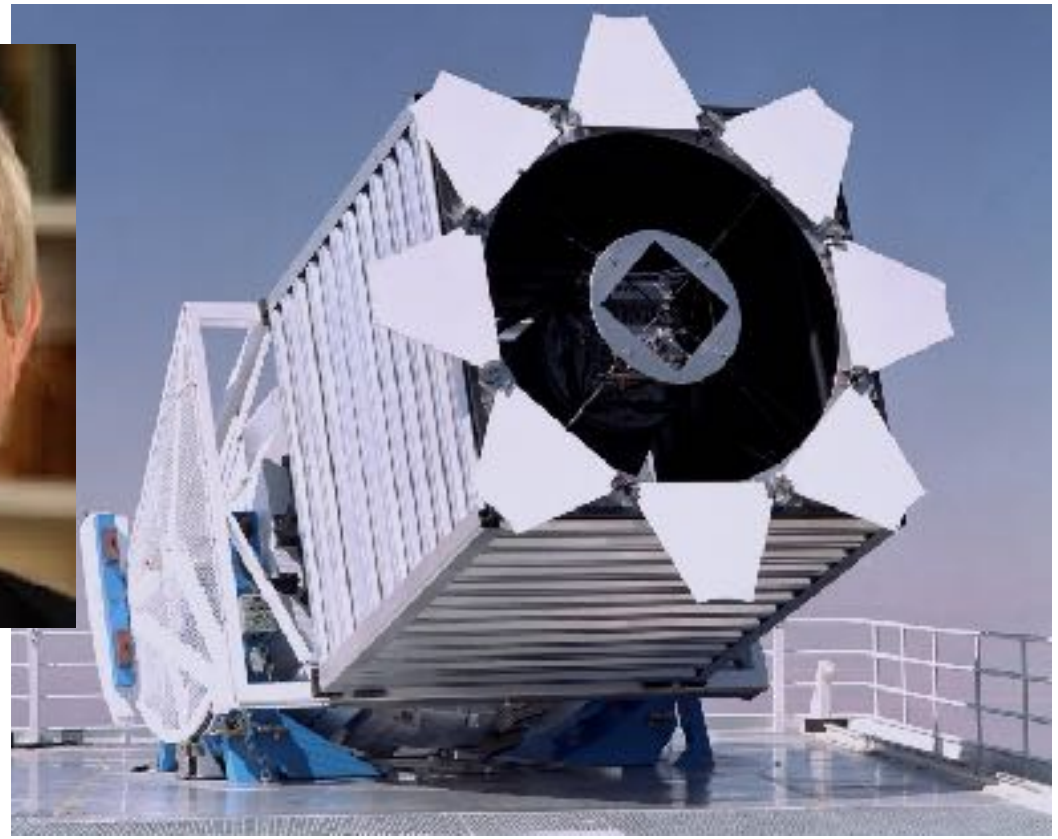
School of Informatics, University of Edinburgh

THE UNIVERSITY of EDINBURGH informatics | epcc |

No. 777413

DARE

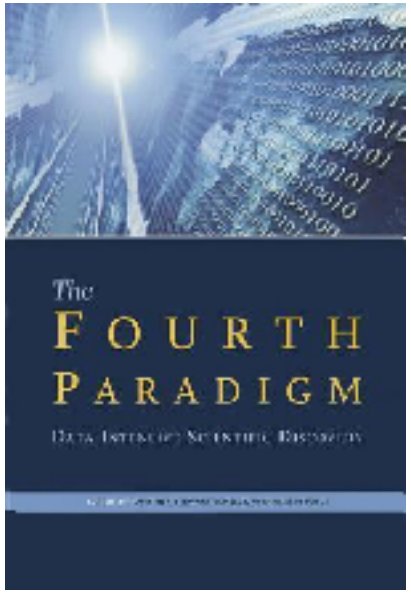
# The 4th Paradigm

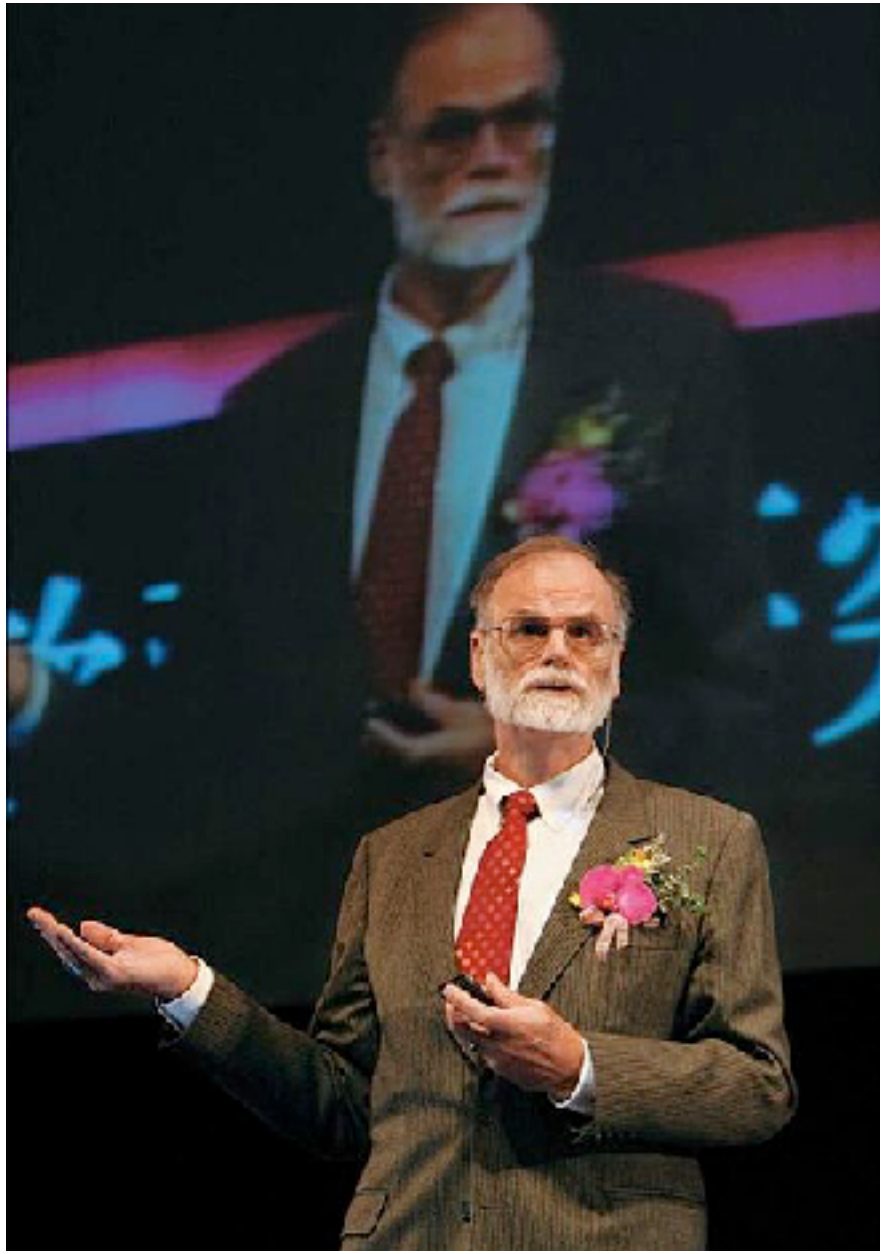


- Experimental Science
- Theoretical Science
- Computational Science
- **Data-intensive Science**
  - Digital technology yields data bonanza
  - Laden with *latent* information
  - Challenges learning
    - to handle volume
    - to discover the knowledge
    - to share the opportunities openly
  - Outruns Moore's law
  - Sociological and technological limits

# The 4th Paradigm

- Experimental Science
- Theoretical Science
- Computational Science
- **Data-intensive Science**
  - Digital technology yields data bonanza
  - Laden with *latent* information
  - Challenges learning
    - to handle volume
    - to discover the knowledge
    - to share the opportunities openly
  - Outruns Moore's law
  - Sociological and technological limits

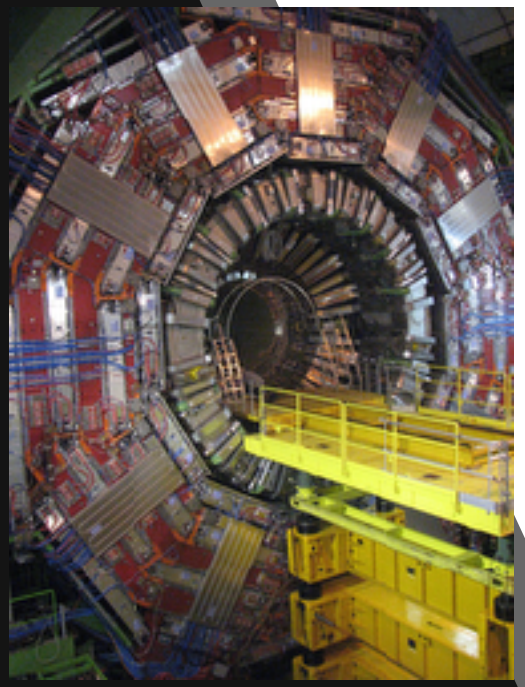


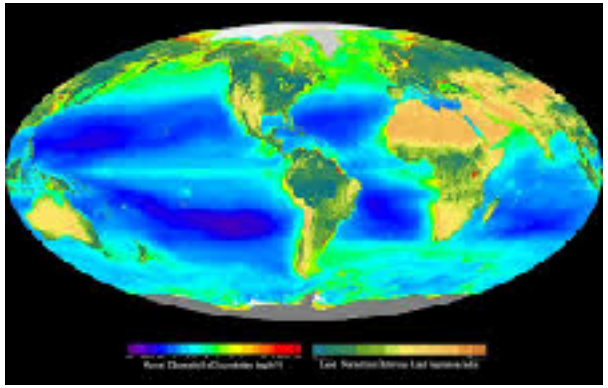



Jim Grey, Microsoft Research

# Examples

- Astronomy
- Environmental sciences
- Climate and weather
- High-energy physics
- Life sciences
- Social sciences
- Humanities

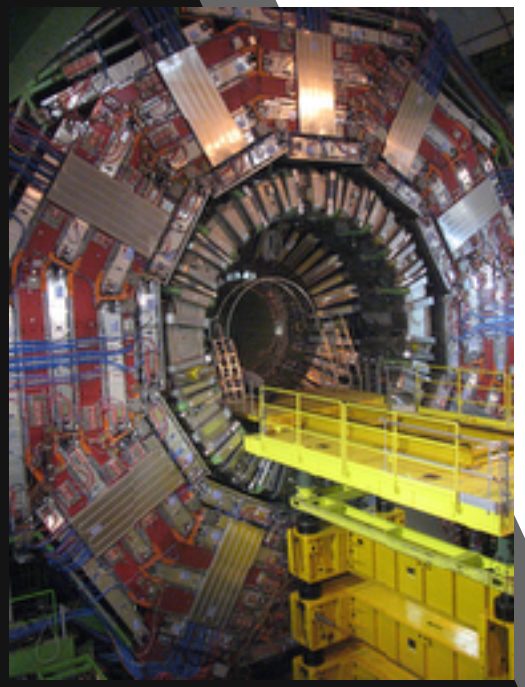


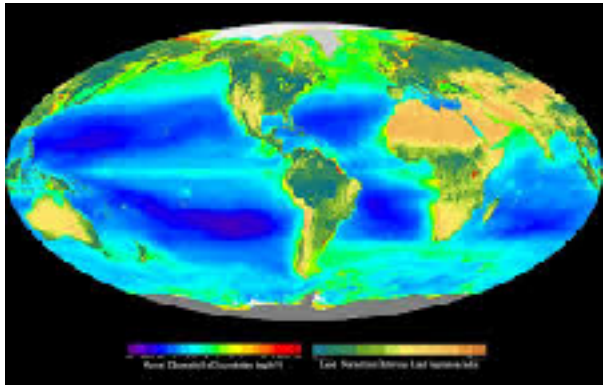


twitter

# Examples

- Astronomy
- **Environmental sciences**
- **Climate and weather**
- High-energy physics
- Life sciences
- Social sciences
- Humanities

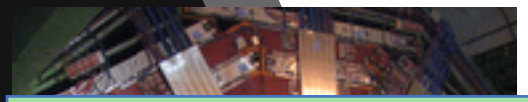


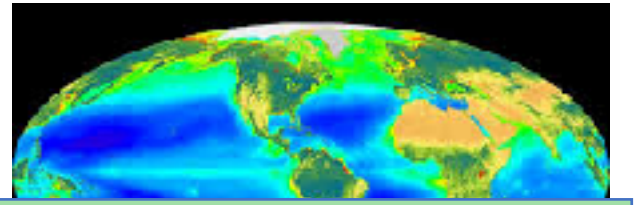


twitter

# Examples

- Astronomy
- **Environmental sciences**
- **Climate and weather**
- High-energy physics
- Life sciences
- Social sciences
- Humanities





**Pressing challenges**

**with**

**global and local pressures**

# 50 years ago

Edsger Dijkstra

## Letters to the Editor

### Go To Statement Considered Harmful

Key Words and Phrases: go to statement, jump instruction, branch instruction, conditional clause, alternative clause, repetitive clause, program intelligibility, program sequencing
*CR* Categories: 4.22, 5.23, 5.24

EDITOR:

For a number of years I have been familiar with the observation that the quality of programmers is a decreasing function of the density of **go to** statements in the programs they produce. More recently I discovered why the use of the **go to** statement has such disastrous effects, and I became convinced that the **go to** statement should be abolished from all "higher level" programming languages (i.e. everything except, perhaps, plain machine code). At that time I did not attach too much importance to this discovery; I now submit my considerations for publication because in very recent discussions in which the subject turned up, I have been urged to do so.

My first remark is that, although the programmer's activity ends when he has constructed a correct program, the process taking place under control of his program is the true subject matter of his activity, for it is this process that has to accomplish the desired effect; it is this process that in its dynamic behavior has to satisfy the desired specifications. Yet, once the program has been made, the "making" of the corresponding process is delegated to the machine.

dynamic progress is only characterized when we also g
call of the procedure we refer. With the inclusion of
we can characterize the progress of the process via a
textual indices, the length of this sequence being e
dynamic depth of procedure calling.

Let us now consider repetition clauses (like, **while**
or **repeat** *A* **until** *B*). Logically speaking, such clau
superfluous, because we can express repetition with
recursive procedures. For reasons of realism I don't
clude them: on the one hand, repetition clauses ca
mented quite comfortably with present day finite eq
the other hand, the reasoning pattern known as
makes us well equipped to retain our intellectual g
processes generated by repetition clauses. With the
the repetition clauses textual indices are no longer
describe the dynamic progress of the process. With eac
a repetition clause, however, we can associate a so-
namic index," inexorably counting the ordinal nu
corresponding current repetition. As repetition clau
procedure calls) may be applied nestedly, we find t
progress of the process can always be uniquely charac
(mixed) sequence of textual and/or dynamic indices.

The main point is that the values of these indices
programmer's control; they are generated (either by
of his program or by the dynamic evolution of the proc
he wishes or not. They provide independent coordina

# Download considered harmful

- Prevalent today
- FAIR pushes download
- Evidence of value?
- Complexity for researchers
- Provenance unsupported

# Download considered harmful

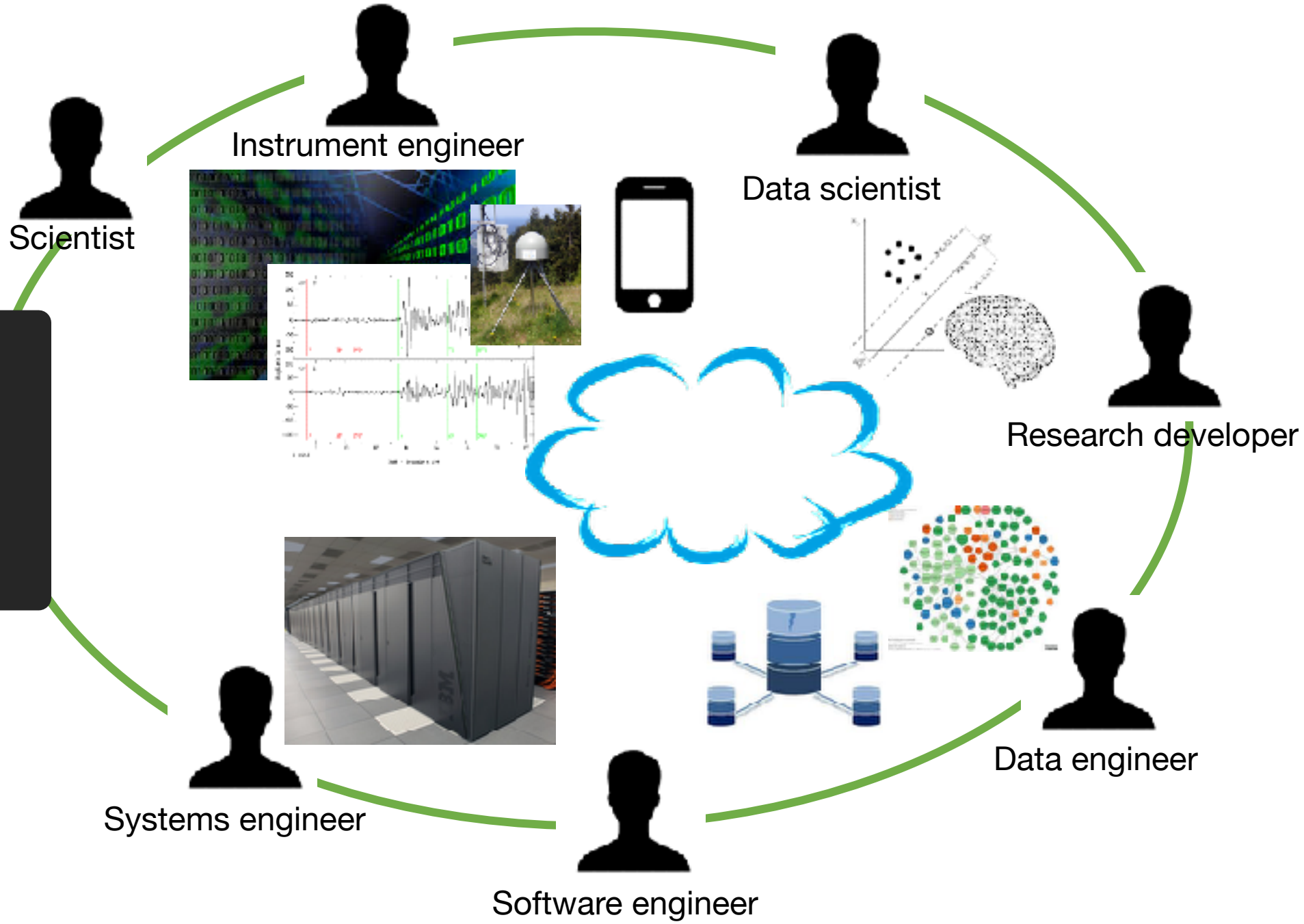
- Prevalent today
- FAIR pushes download
- Evidence of value?
- Complexity for researchers
- Provenance unsupported

**Evidence led policy**

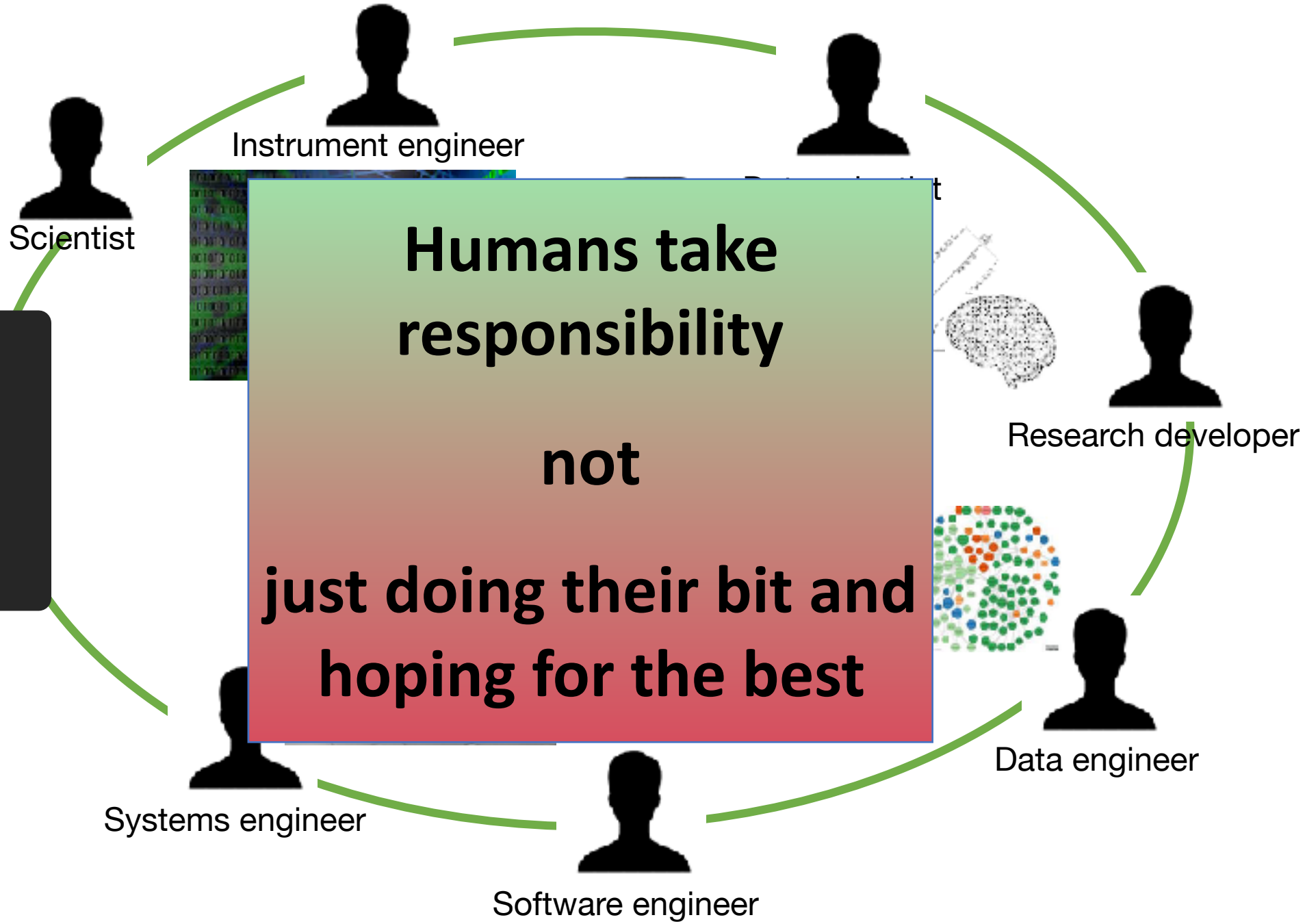
**not**

**policy led evidence**

Pooling expertise across disciplines



Pooling expertise across disciplines

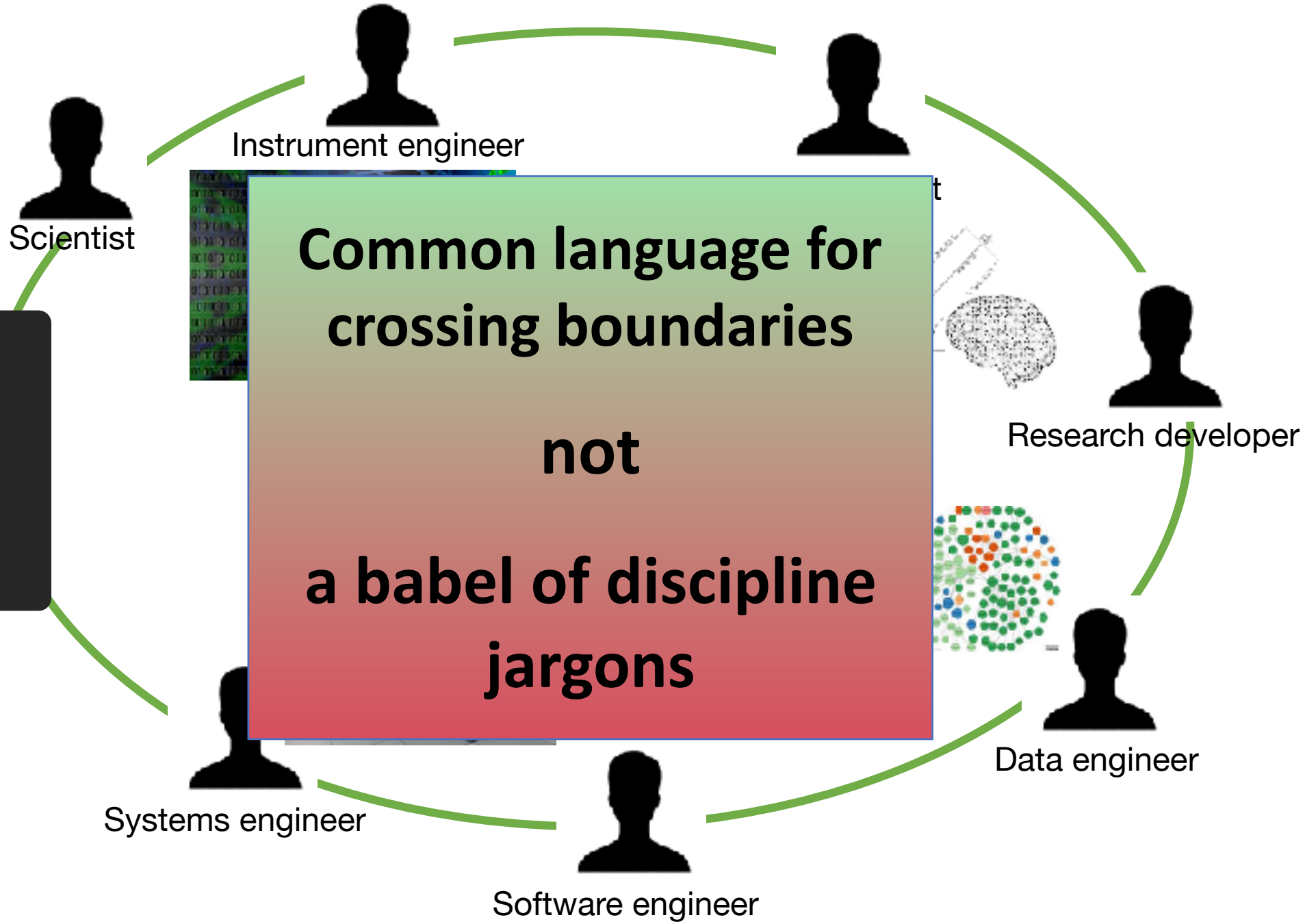



**Humans take responsibility**

**not**

**just doing their bit and hoping for the best**

Pooling expertise across disciplines

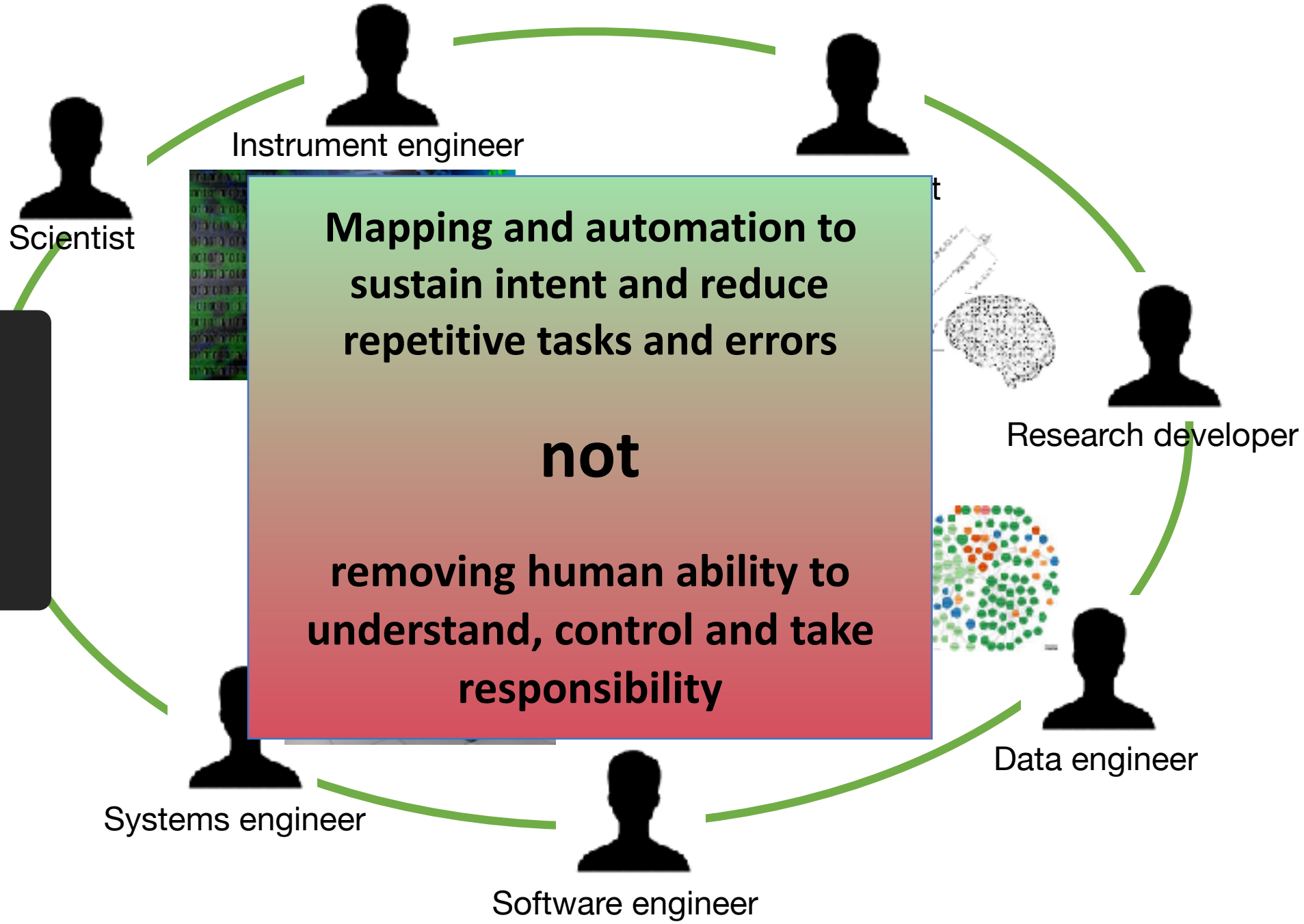



**Common language for crossing boundaries**

**not**

**a babel of discipline jargons**

Pooling expertise across disciplines




**Mapping and automation to sustain intent and reduce repetitive tasks and errors**

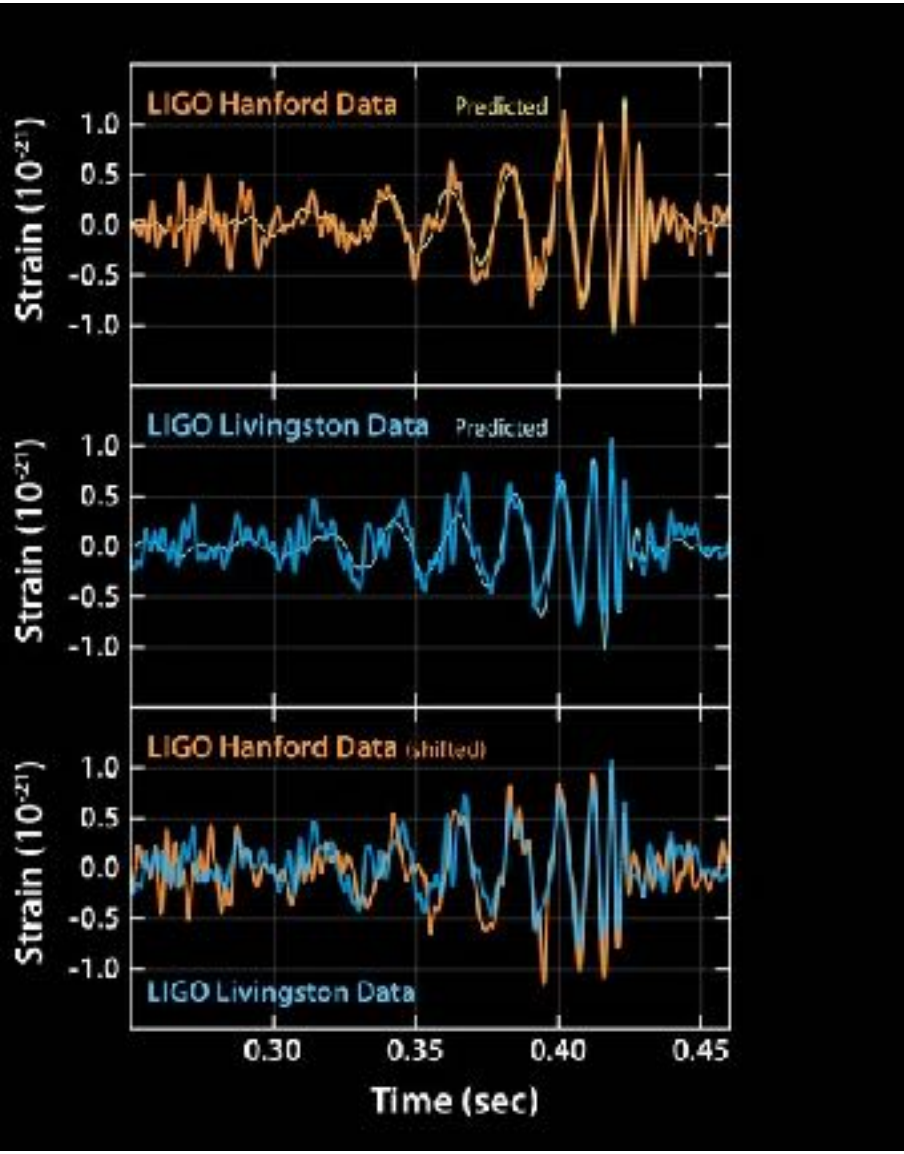
**not**

**removing human ability to understand, control and take responsibility**

# Gravitational wave detection

## LIGO: (Laser Interferometer Gravitational-Wave Observatory)

- Aims to detect gravitational waves predicted by Einstein's theory of relativity.
- Can be used to detect
  - binary pulsars
  - mergers of black holes
  - "starquakes" in neutron stars
- Two installations: in Louisiana (Livingston) and Washington State
  - Other projects: Virgo (Italy), GEO (Germany), Tama (Japan)
- Instruments are designed to measure the effect of gravitational waves on test masses suspended in vacuum.
- Data collected during experiments is a collection of time series (multi-channel)



# Effective sharing embracing diversity

- Many **autonomous** organisations
- Cluster of **very long** research campaigns
- Scale and scope determined by **challenges**
- Immense diversity plus **global**

# Effective sharing embracing diversity

- Many **autonomous** organisations
- Cluster of **very long** research campaigns
- Scale and scope determined by **challenges**
- Immense diversity plus **global**

**Reaches scale and breadth for *long-term* clusters of campaigns tackling global and societal challenges**

**business drives *rapid* change in heterogeneous digital environment**

# Effective sharing embracing diversity

- Many **autonomous** organisations
- Cluster of [illegible] research camp[illegible]
- Scale and scope determined by **challenges**
- Immense diversity plus **global**

**Darwinian challenge to research campaigns**

**Reaches scale and breadth for *long-term* clusters of campaigns tackling global and societal challenges**

**business drives *rapid* change in heterogeneous digital environment**

# Gaining allegiance incrementally

- Many individuals, roles and priorities
- Established practices plus innovation
- Fair as well as FAIR
- Data, Information, Knowledge, Visualisations, Judgement, Methods, Cultures, Professional practices, QA, Digital platforms, Observational systems, Archives, Computational models, Networks, VREs, Stores, Scientific databases, Ethical rules, Collaboration agreements, DMPs, FAIR, …

# Gaining allegiance incrementally

- Many individuals, roles and priorities
- Established practices plus innovation
- Fair as well as FAIR
- Data, Information, Knowledge, Visualisations, Judgement, Methods, Cultures, Professional practices, QA, Digital platforms, Observational systems, Archives, Computational models, Networks, VREs, Stores, Scientific databases, Ethical rules, Collaboration agreements, DMPs, FAIR, …

**Building collaboration**

**without**

**losing human talent**

# Gaining allegiance incrementally

- Many individuals, ro

- Established practices

- Fair as well as FAIR

- Data, Information, Knowledge, Visualisations, Judgement, Methods, Cultures, Professional practices, QA, Digital platforms, Observational systems, Archives, Computational models, Networks, VREs, Stores, Scientific databases, Ethical rules, Collaboration agreements, DMPs, FAIR, …

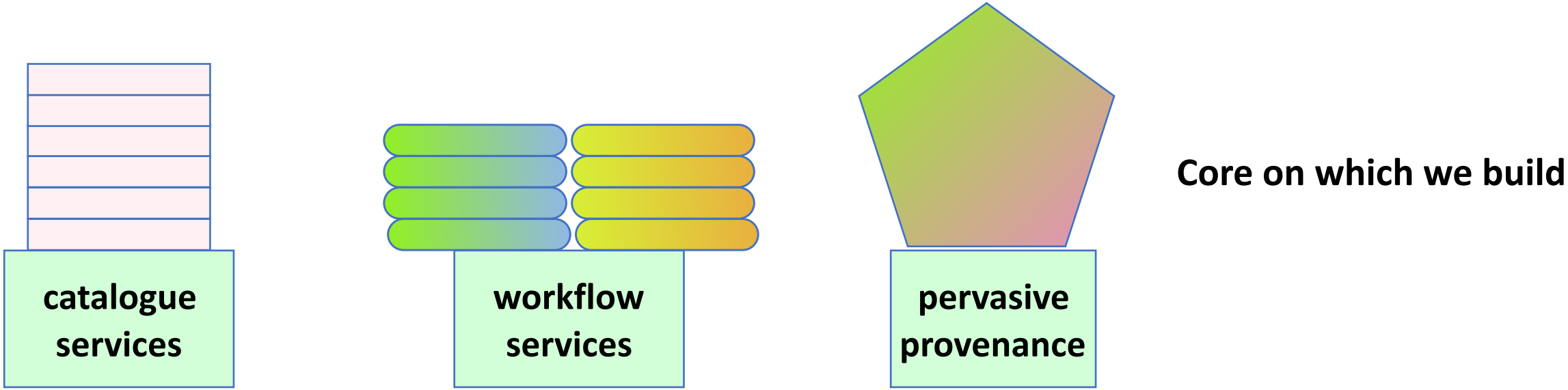
**Complexity harnessed** via **Integrated understandable controllable abstractions**

**Building collaboration**

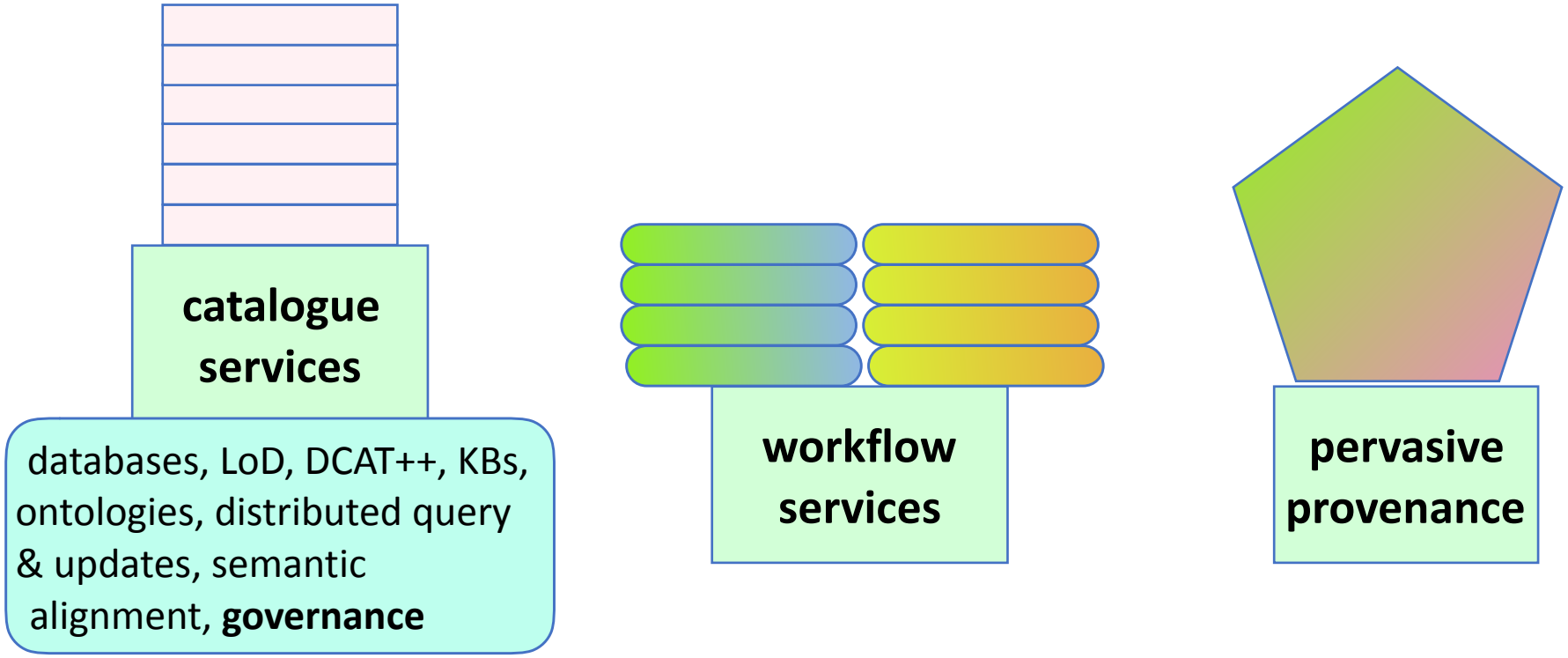
**without**

**losing human talent**

# Flexible Federation Framework



# Flexible Federation Framework

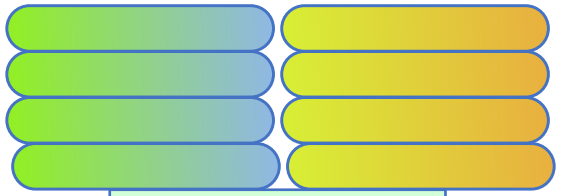


# Flexible Federation Framework

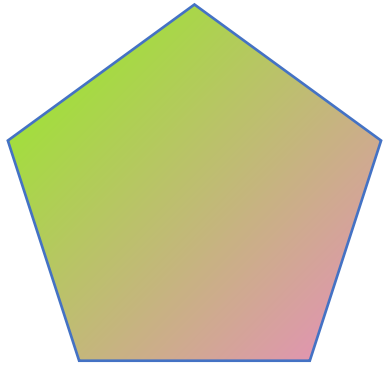
**Growing inventory of everything**
**for**
***human* understanding and communication**
***machine* automation, interpretation and optimisation**

**catalogue services**

databases, LoD, DCAT++, KBs, ontologies, distributed query & updates, semantic alignment, **governance**

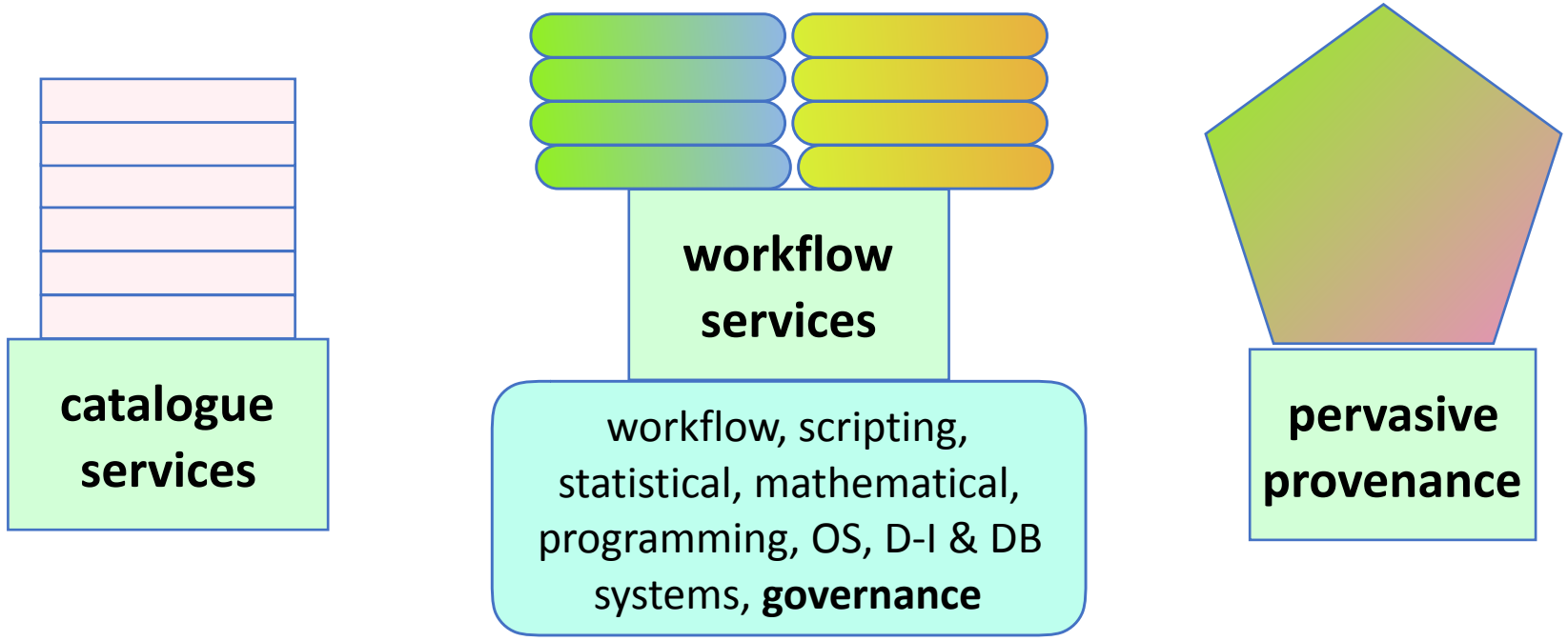


**workflow services**



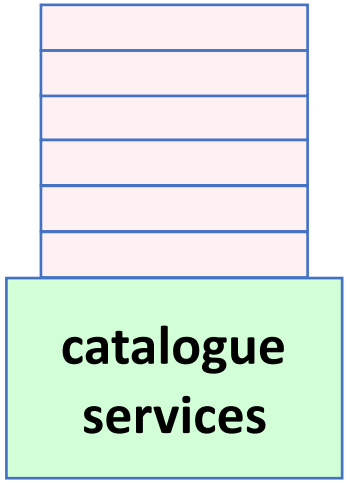
**pervasive provenance**

# Flexible Federation Framework

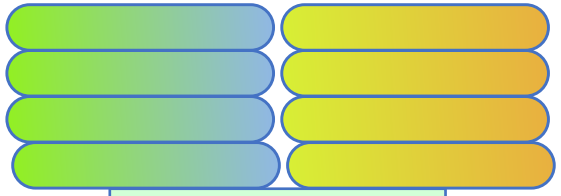


# Flexible Federation Framework

**Actions encoded and executed by actors**
***human* specifying and refining intent + judging results**
***machine* automation, interpretation and optimisation, choosing and mapping to *platforms*, *modes* and *contexts***

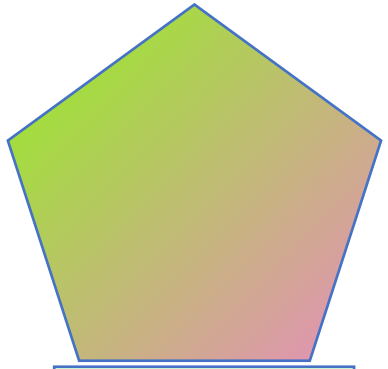


**catalogue services**

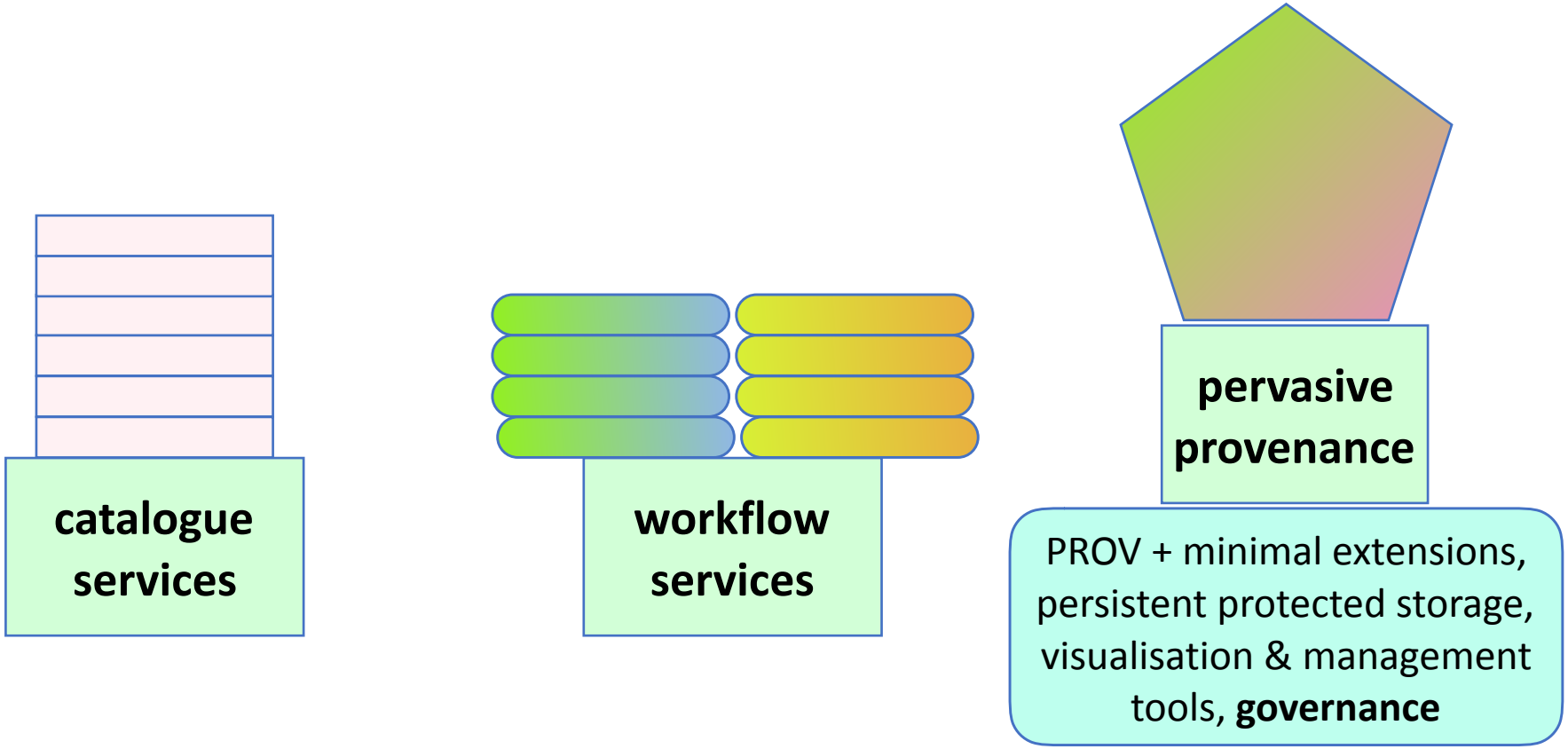


**workflow services**

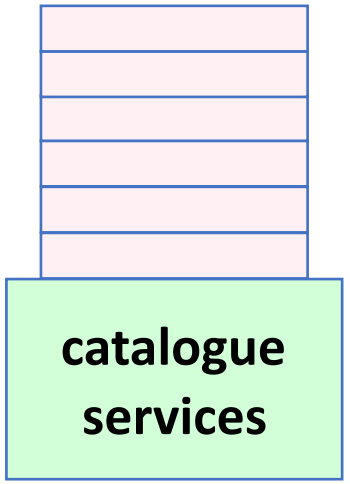
workflow, scripting, statistical, mathematical, programming, OS, D-I & DB systems, **governance**



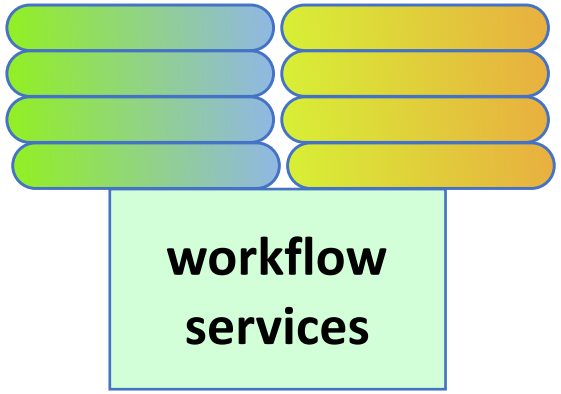
**pervasive provenance**


catalogue services
workflow services
pervasive provenance
PROV + minimal extensions, persistent protected storage, visualisation & management tools, **governance**

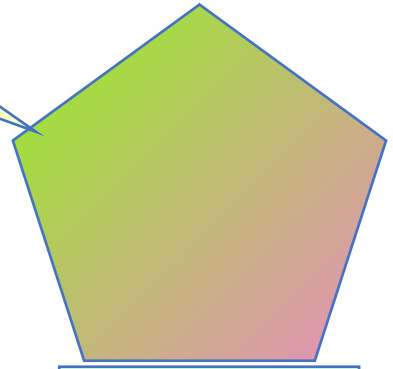
**Reliable consistent history**
***foundation* for**
***human* understanding, annotating, validating and organising**
**+ judging results + investigations**
***machine* automation, diagnostics, optimisation, recovery,**
**avoiding redundant work, planning and**
**provisioning**



**catalogue services**



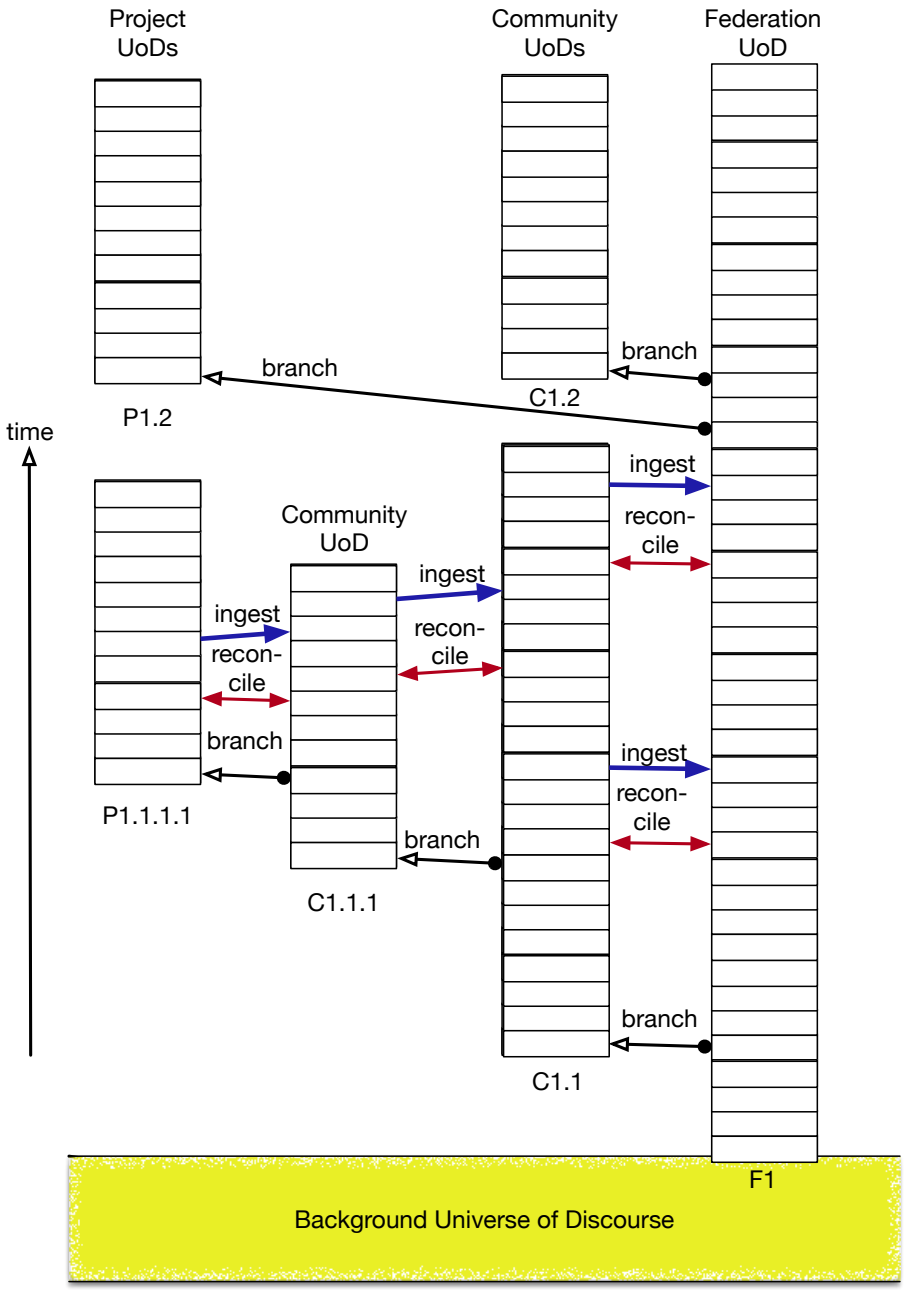
**workflow services**



**pervasive provenance**

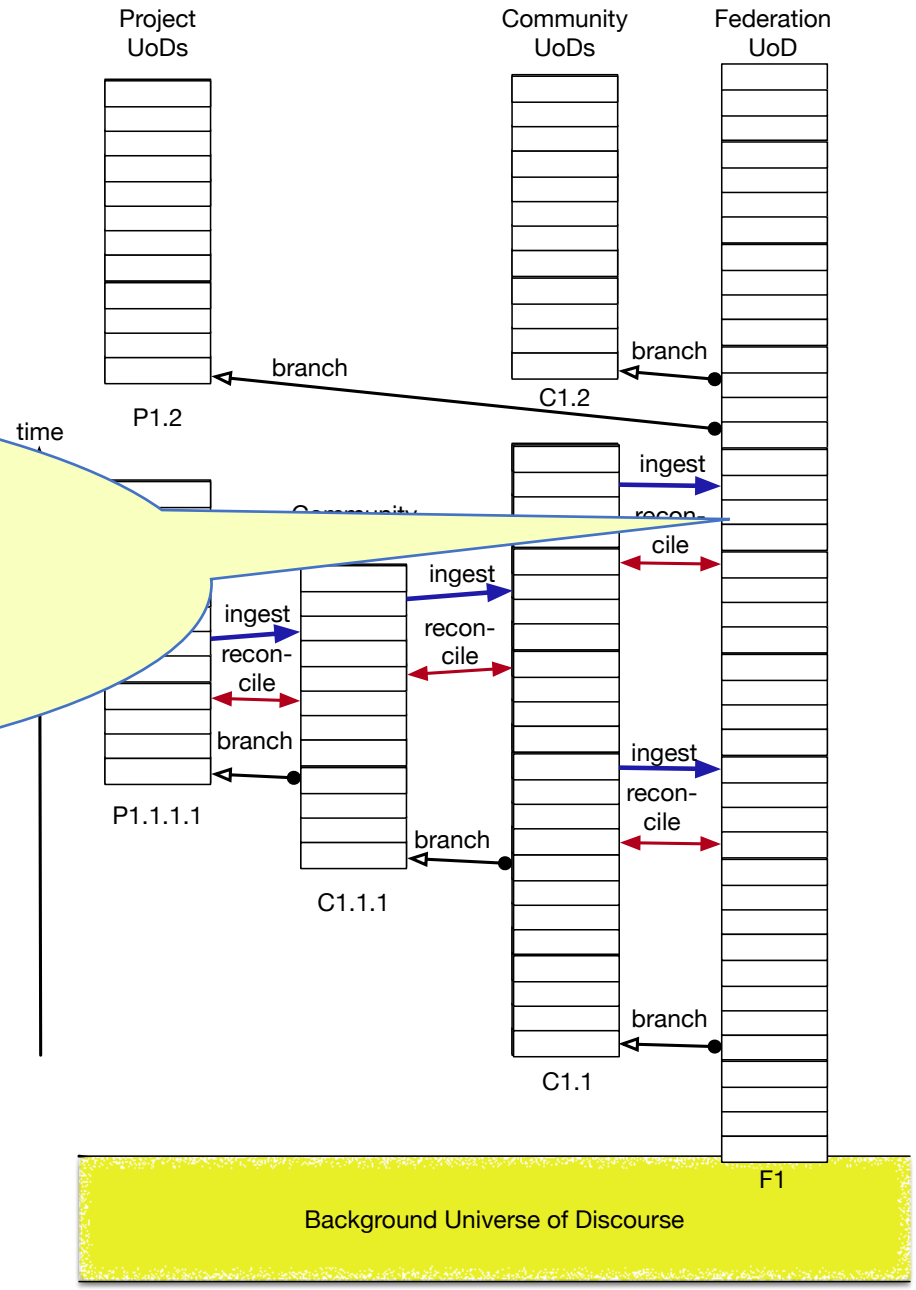
PROV + minimal extensions, persistent protected storage, visualisation & management tools, **governance**

# Universes of Discourse



# Universes of Discourse

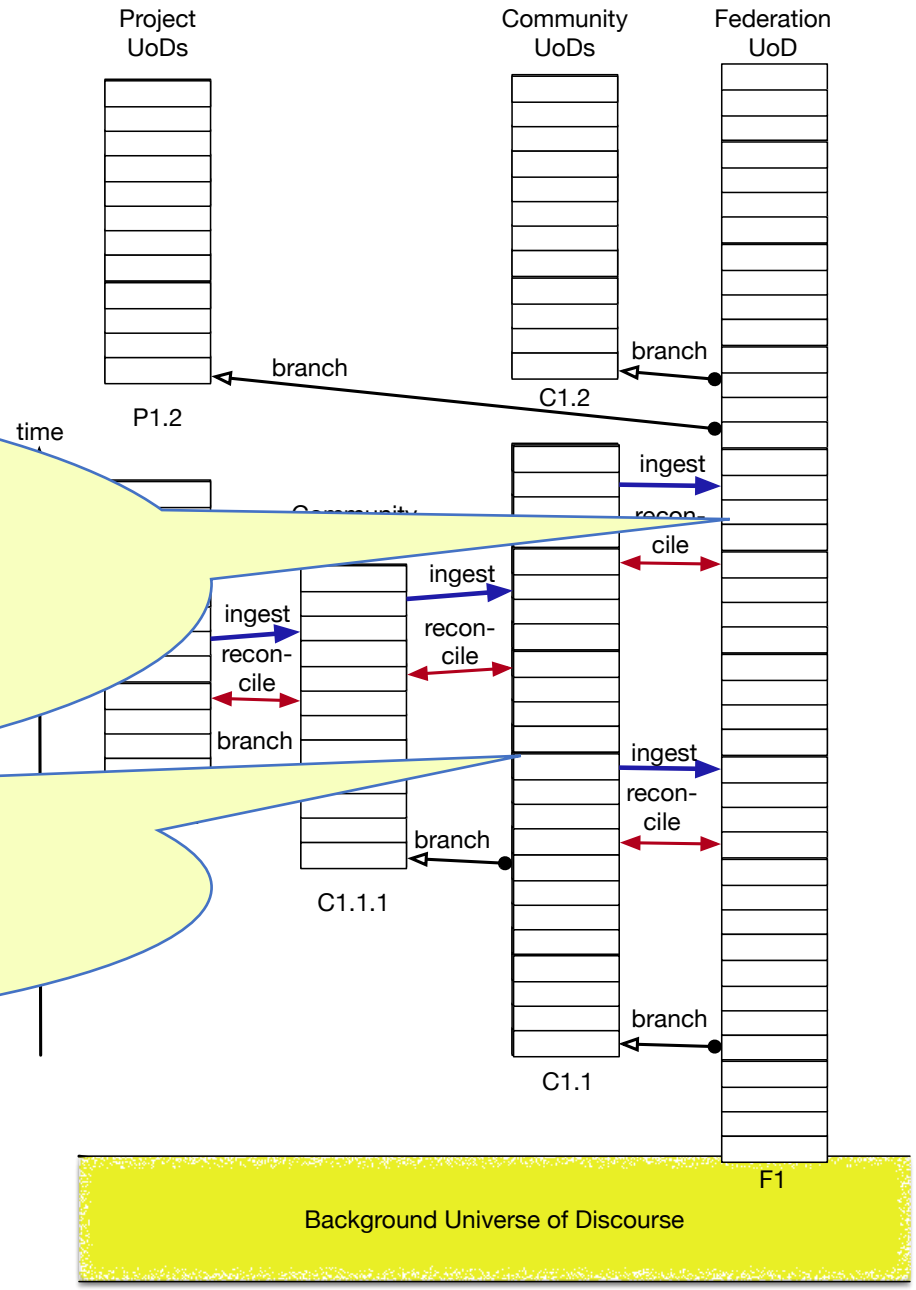
**Common *stabilised* concepts, procedures and resources**
**for**
***entire* federation *relative* to background**



# Universes of Discourse

**Common *stabilised* concepts, procedures and resources**
**for**
***entire* federation *relative* to background**

***UoD* for a significant sustainable sub-community**
***relative* to common core**

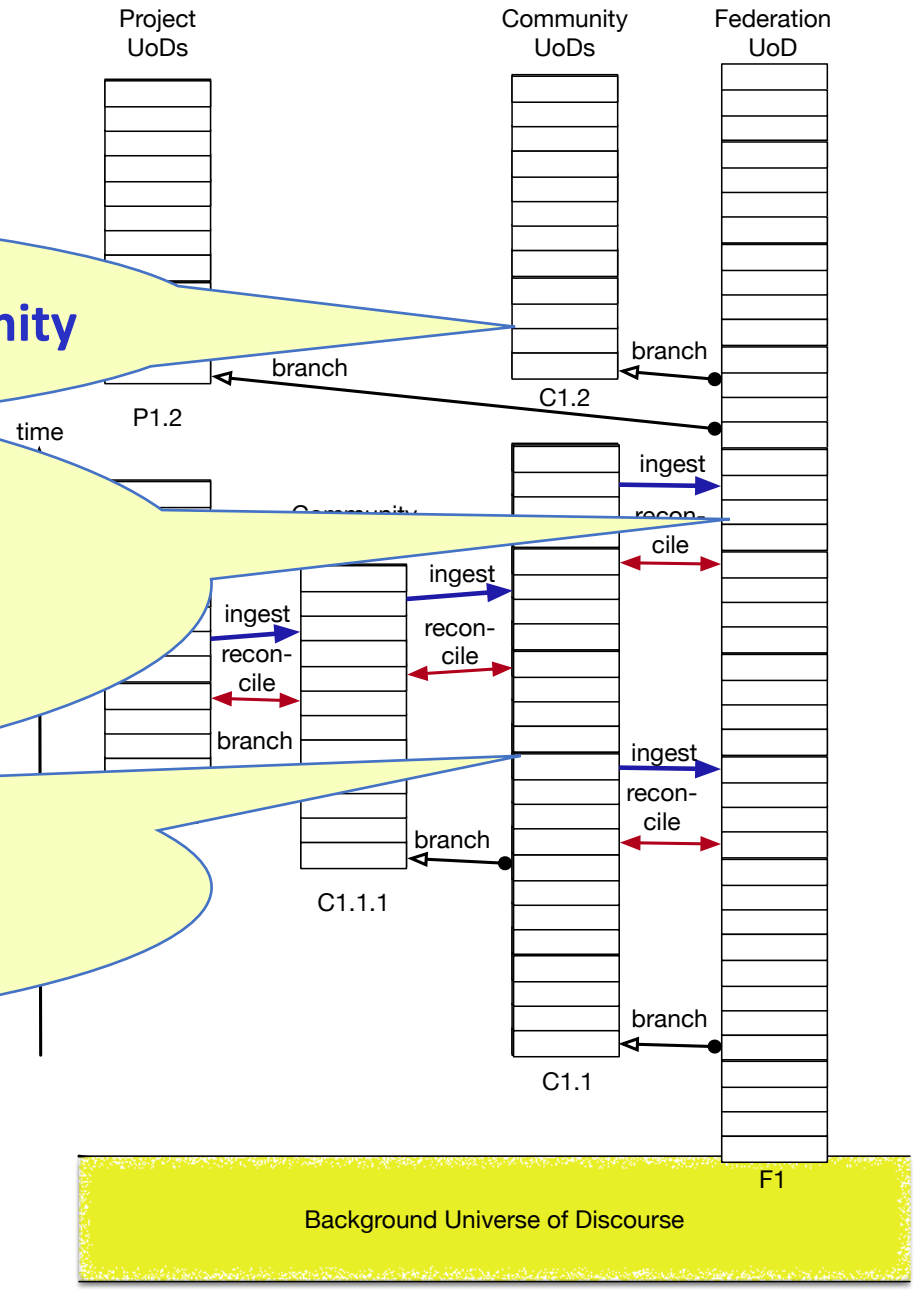


# Universes of Discourse

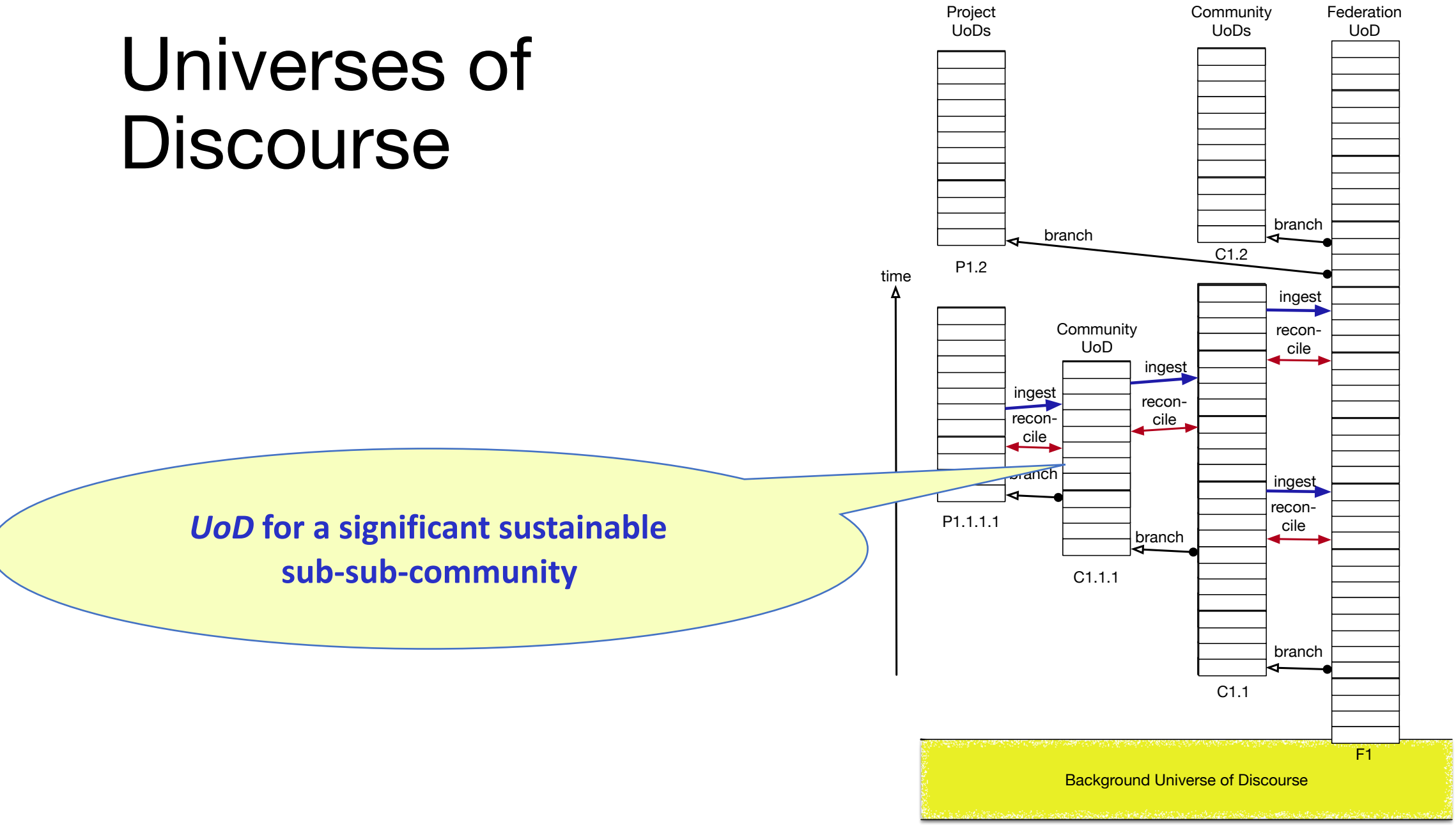
***UoD* for another significant sustainable sub-community**

**Common *stabilised* concepts, procedures and resources for *entire* federation *relative* to background**

***UoD* for a significant sustainable sub-community *relative* to common core**



# Universes of Discourse

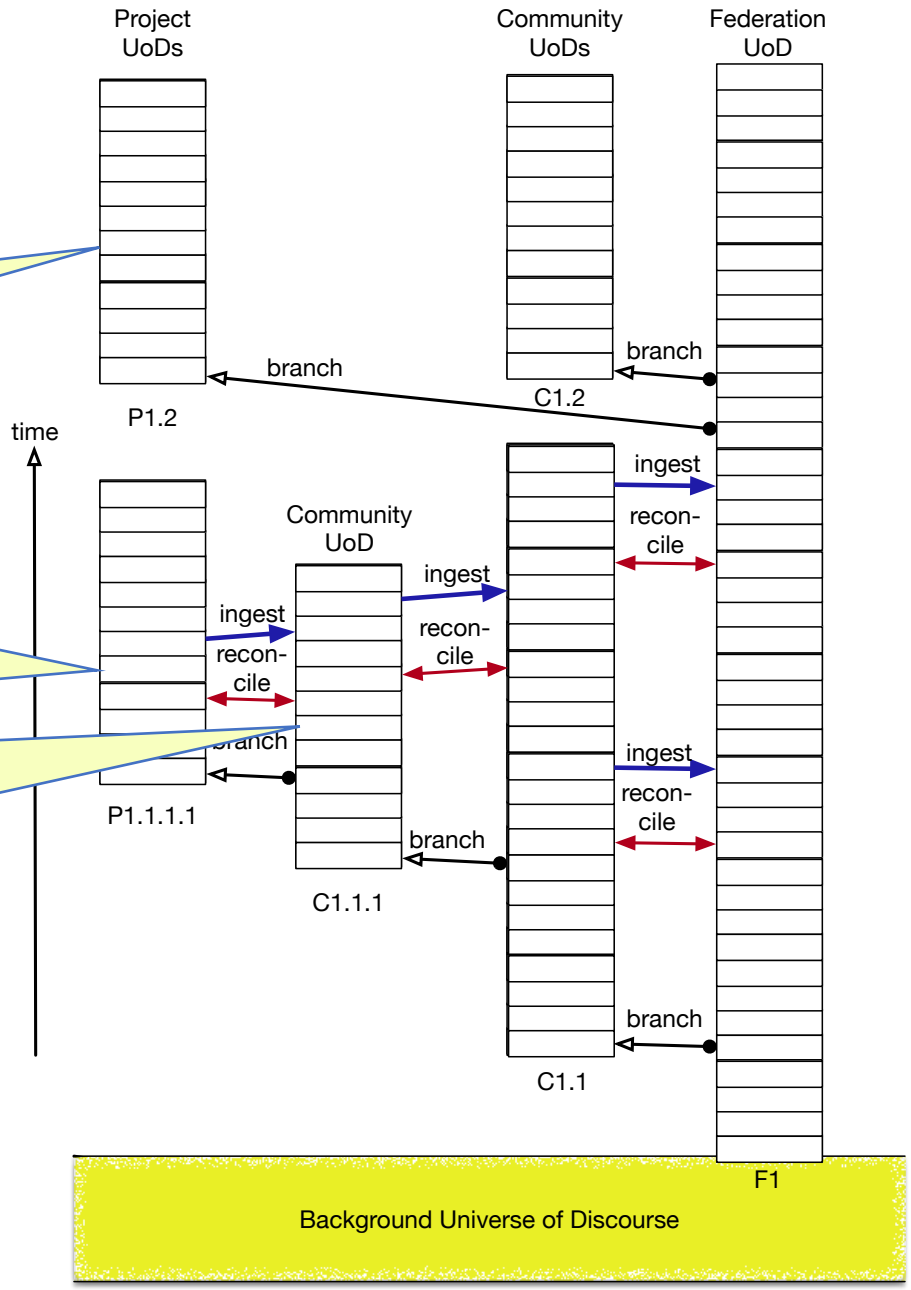


# Universes of Discourse

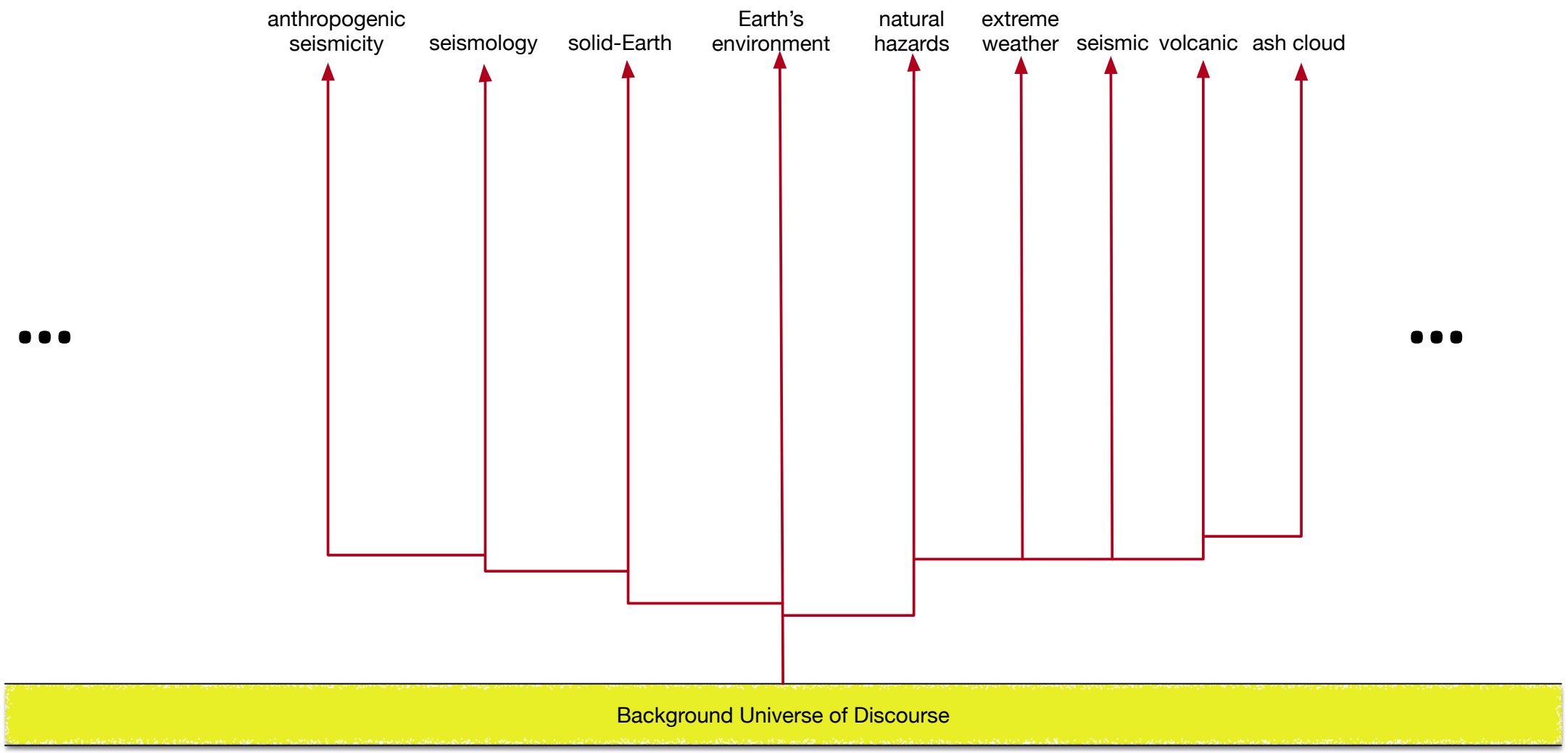
***Short*** **project with unfettered innovation**

***Short*** **project with unfettered innovation still with supporting work environment**

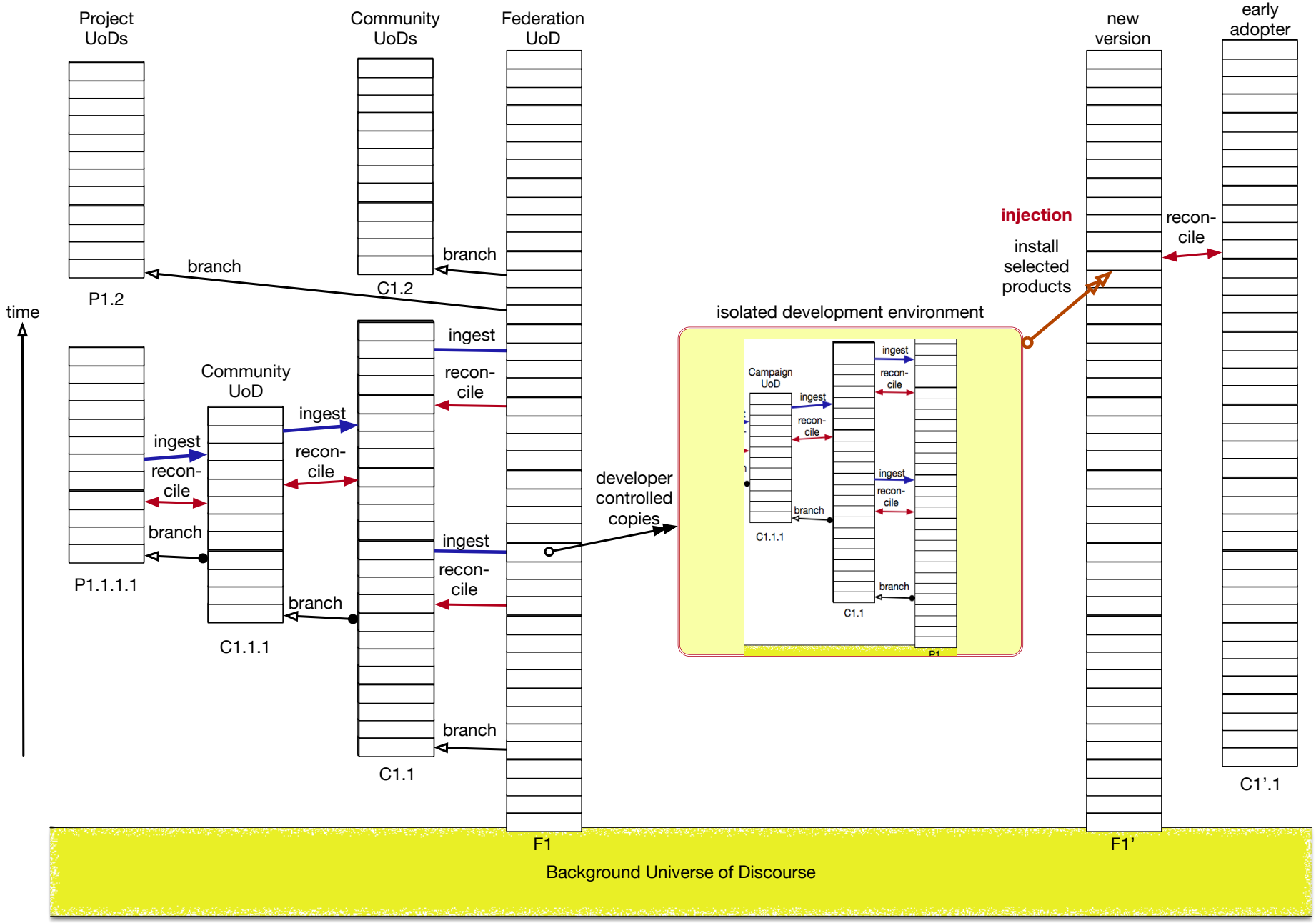
***UoD*** **for a significant sustainable sub-sub-community**



# Federation structure



# Managing change to limit disruption with protection



# Take home messages

- Ensure human responsibility and control
- Value diversity and build for it
- Mirror human structures
- Accommodate organisational autonomy
- Nurture innovation while retaining established practices
- Invest in sustaining collaborations and methods
- You can’t dodge complexity our world is complex
- But sustain manageable niches and eliminate technology intrusion

# Take home messages

- Ensure human responsibility and control
- Value diversity and build for it
- Mirror human structures
- Accommodate organisational autonomy
- Nurture innovation while retaining established practices
- Invest in sustaining collaborations and methods
- You can’t dodge complexity our world is complex
- But sustain manageable niches and eliminate technology intrusion

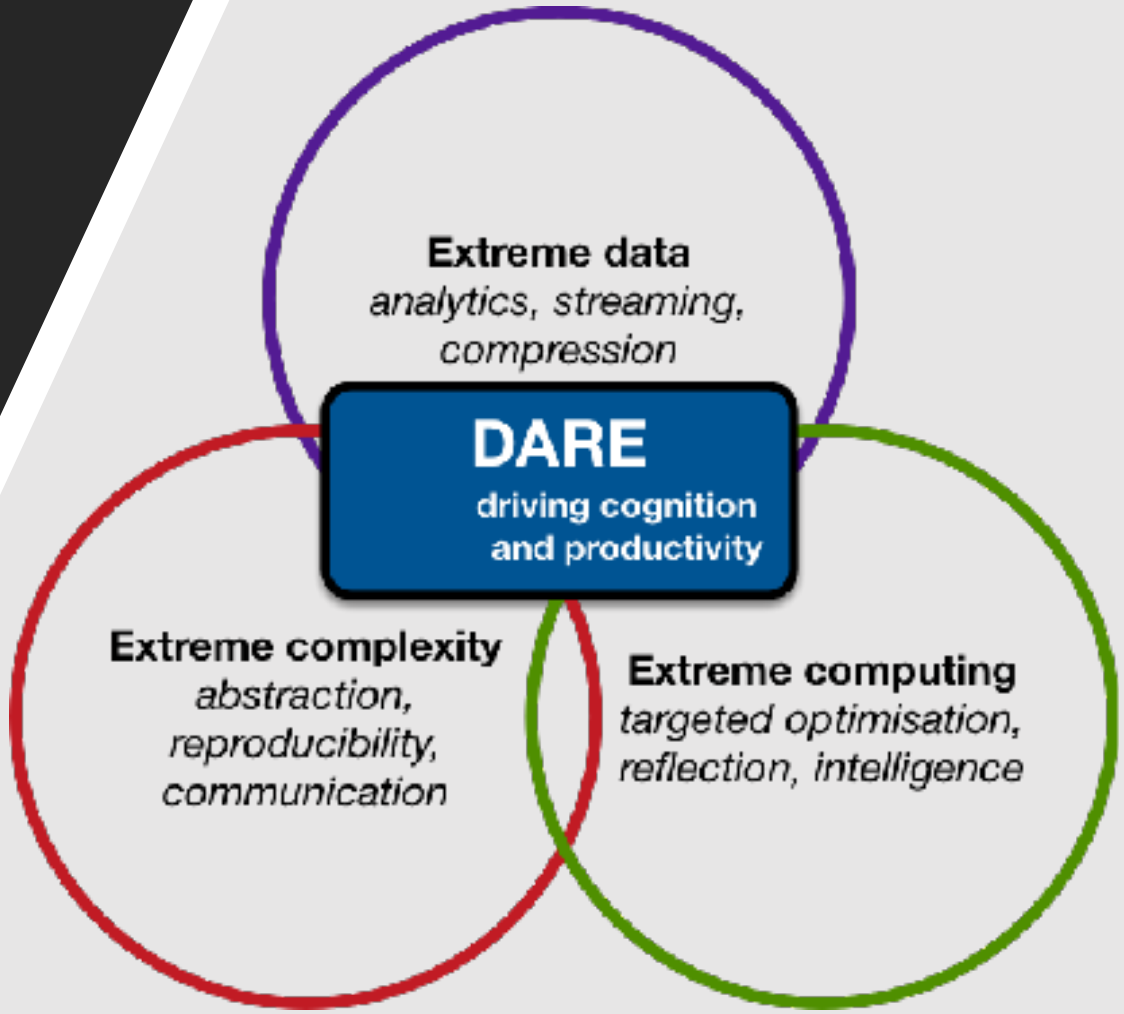
***Encourage*** **innovation and adoption of advances**

**do not**

***impose*** **change or uniformity**

# Project DARE: Taming the Extremes

- Started **January 2018**
- *Agile* Research Excellence
- Abstraction is key
- User groups
- Hyper-services layer
  - Conceptual integration
  - Provenance-powered reasoning
  - Mappings sustaining intent
  - Exploiting distributed diverse advanced platforms



# Thank you!

*Your* questions please

## Visit Edinburgh

**IWSG 2018**

**13-15 June 2018**


## Acknowledgements

**Colleagues**

Rosa Filgueira, Iraklis Klampanos, Alessandro Spinuso, Luca Trani

**Projects**

UK e-Science, ADMIRE, VERCE, ENVRI, ENVRIplus, SKA-link, DARE


# Bibliography


50 years observing Pulsars Jocelyn Bell-Burnell https://www.rse.org.uk/event/fifty-years-of-pulsars-pulasting-radio-stars/

50 years avoiding unnecessary complexity Edsger Dijkstra "*Go To Statement Considered Harmful*", CACM, **11** (3), 147-148, 1968. doi:10.1145/362929.362947

Pegasus' role in gravitational wave detection https://pegasus.isi.edu/tag/ligo/

M.P. Atkinson, S. Gesing, J. Montagnat and I. Taylor, *Scientific Workflows: Past, Present and Future*, FGCS, 75, 216–227, 2017.

R. Filgueira, A. Krause, M.P. Atkinson, I. Klampanos and A. Moreno, *dispel4py: A Python Framework for Data-Intensive Scientific Computing*. In Int. J. of HPC Apps, vol. 31, no. 4, pp. 316-334, 2016

C.S. Liew, M. P. Atkinson, M. Galea, T. F. Ang, P. Martin, J. I. van Hemert, *Scientific workflows: Moving across paradigms*, ACM Comput. Surv. 49 (4) (2017) 66:1–66:39. doi:10.1145/3012429.

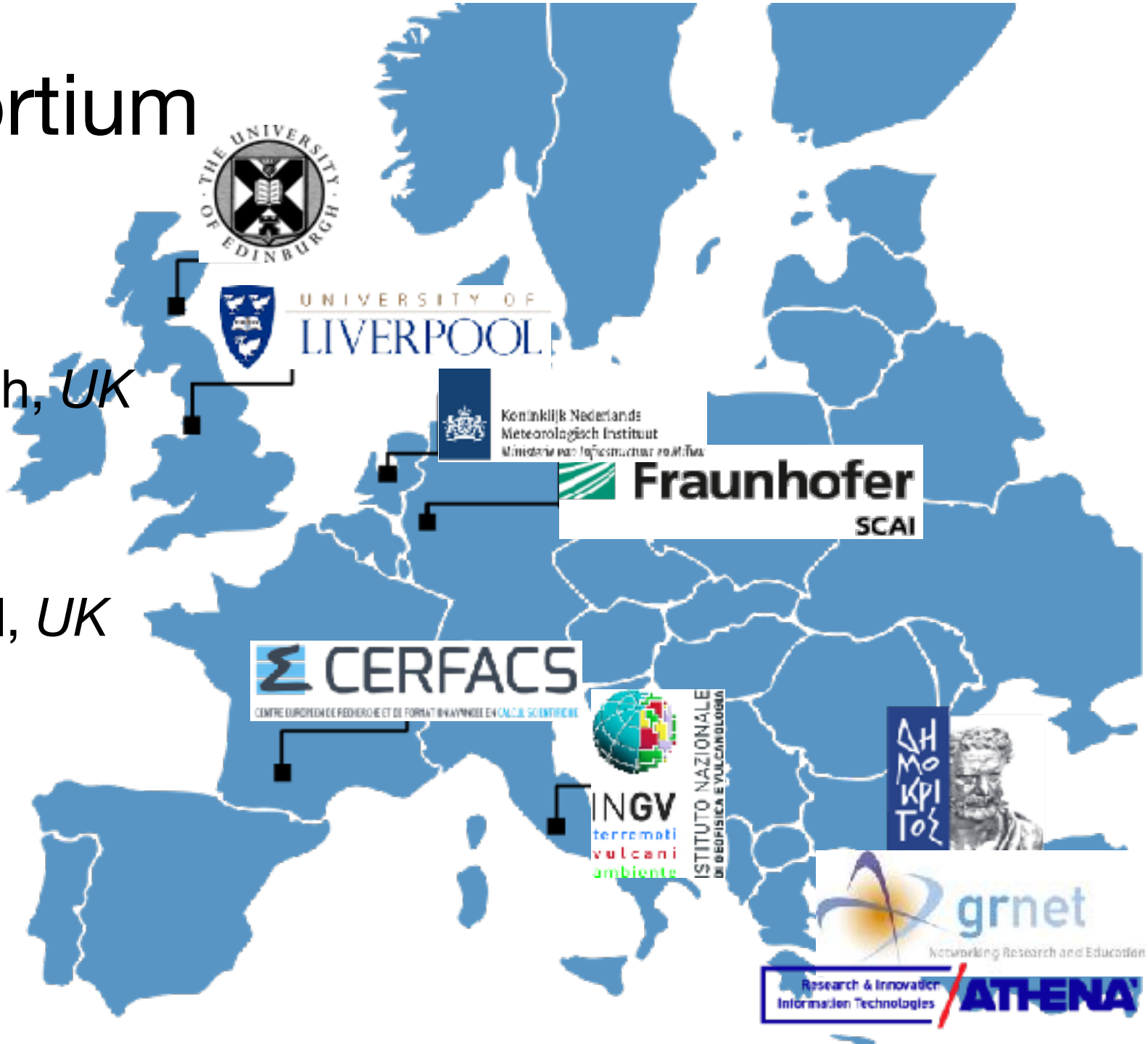
A. Spinuso. *Active Provenance for Data-Intensive Research*, PhD thesis, School of Informatics, University of Edinburgh, to be submitted, 2018.

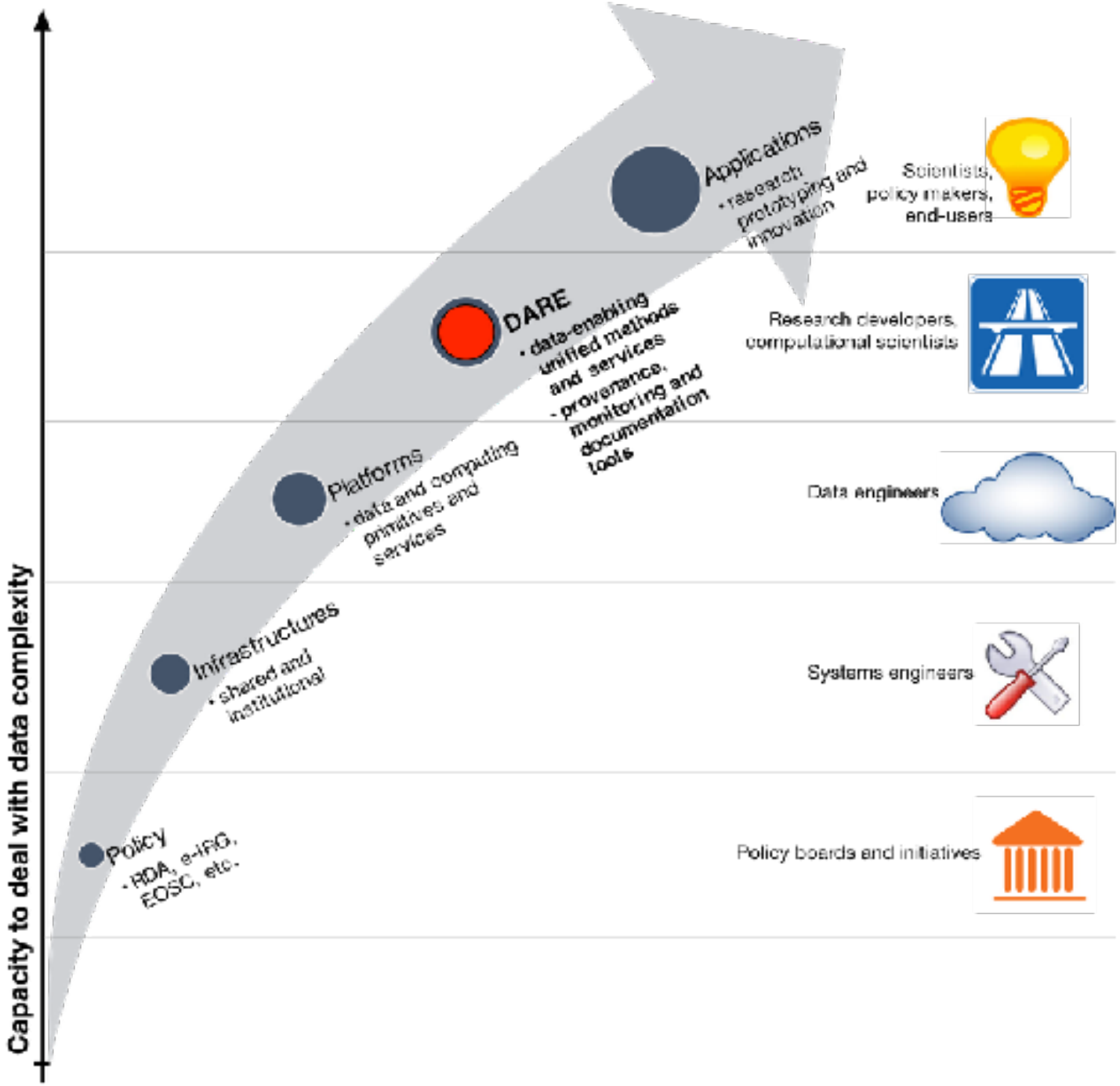
L. Trani, M. Koymans, M.P. Atkinson, R. Sleeman and R. Filgueira. *WFCatalog: A catalogue for seismological waveform data*, Computers Geosciences 106; 101 – 108, 2017.

L. Trani, M.P. Atkinson, D. Bailo, R. Paciello and R. Filgueira. *Establishing Core Concepts for Information-Powered Collaborations – pioneered by solid-Earth sciences*, submitted 2017.

# The DARE Consortium

- NCSR-”Demokritos”, *EL* (Coordinator)
- The University of Edinburgh, *UK*
- INGV, *IT*
- CERFACS, *FR*
- The University of Liverpool, *UK*
- KNMI, *NL*
- GRNET S.A., *EL*
- Fraunhofer SCAI, *DE*
- “ATHENA” RIC, *EL*



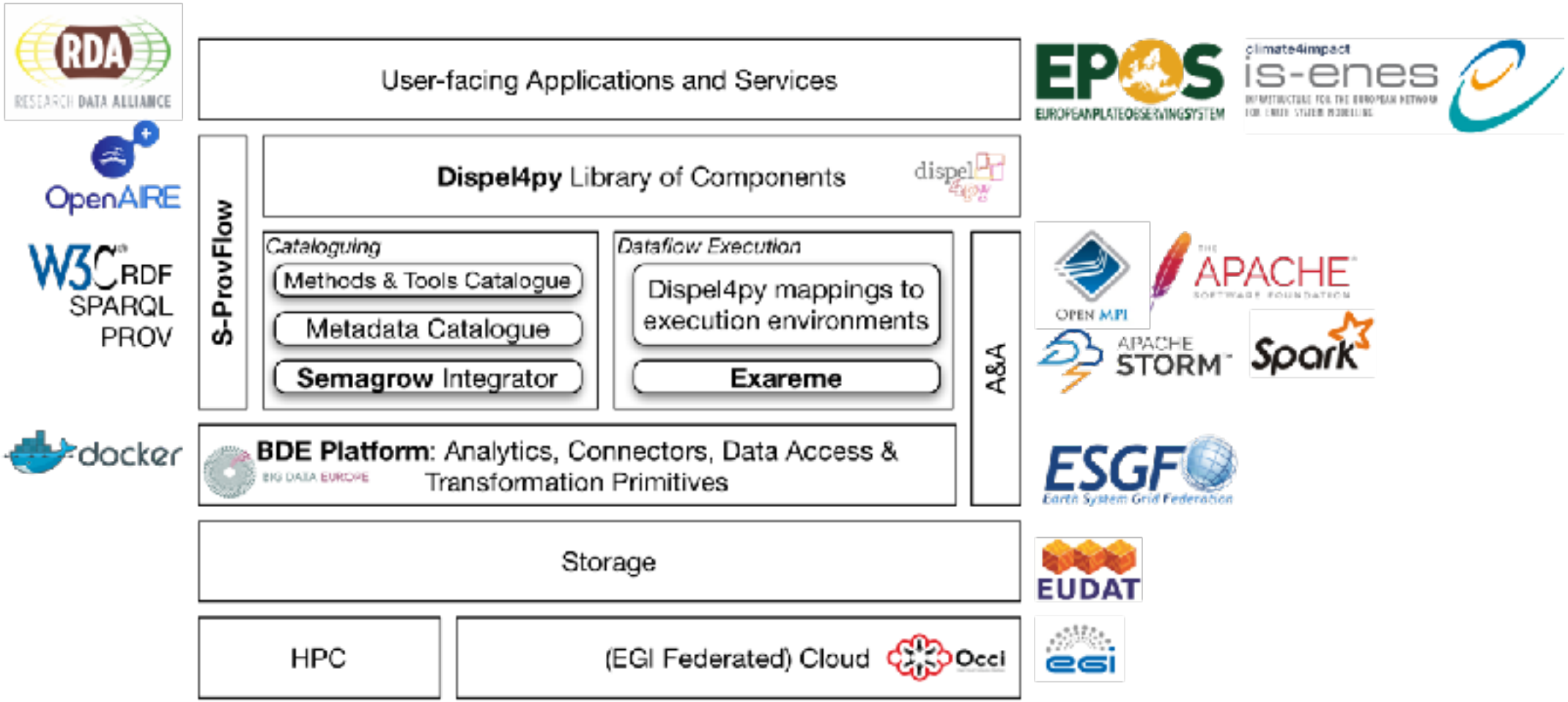



# Users and Requirements

- Key to productivity
- Raising and widening abstraction
  - Communicability of results
  - Reusability
  - Transparency
  - Attribution
  - Reproducibility
- Sustainability

DARE

# Components




- **Minimum TRL: 6**
- **Target TRL by the end of the project: 8**

DARE