The Role of Virtual Observatories and Data Frameworks in an Era of Big Data

NIST bIG dATA

June 14, 2012, Gaithersburg, MD

Peter Fox (RPI and WHOI) pfox@cs.rpi.edu, @taswegian, http://tw.rpi.edu/web/person/PeterFox
Tetherless World Constellation http://tw.rpi.edu and AOP&E

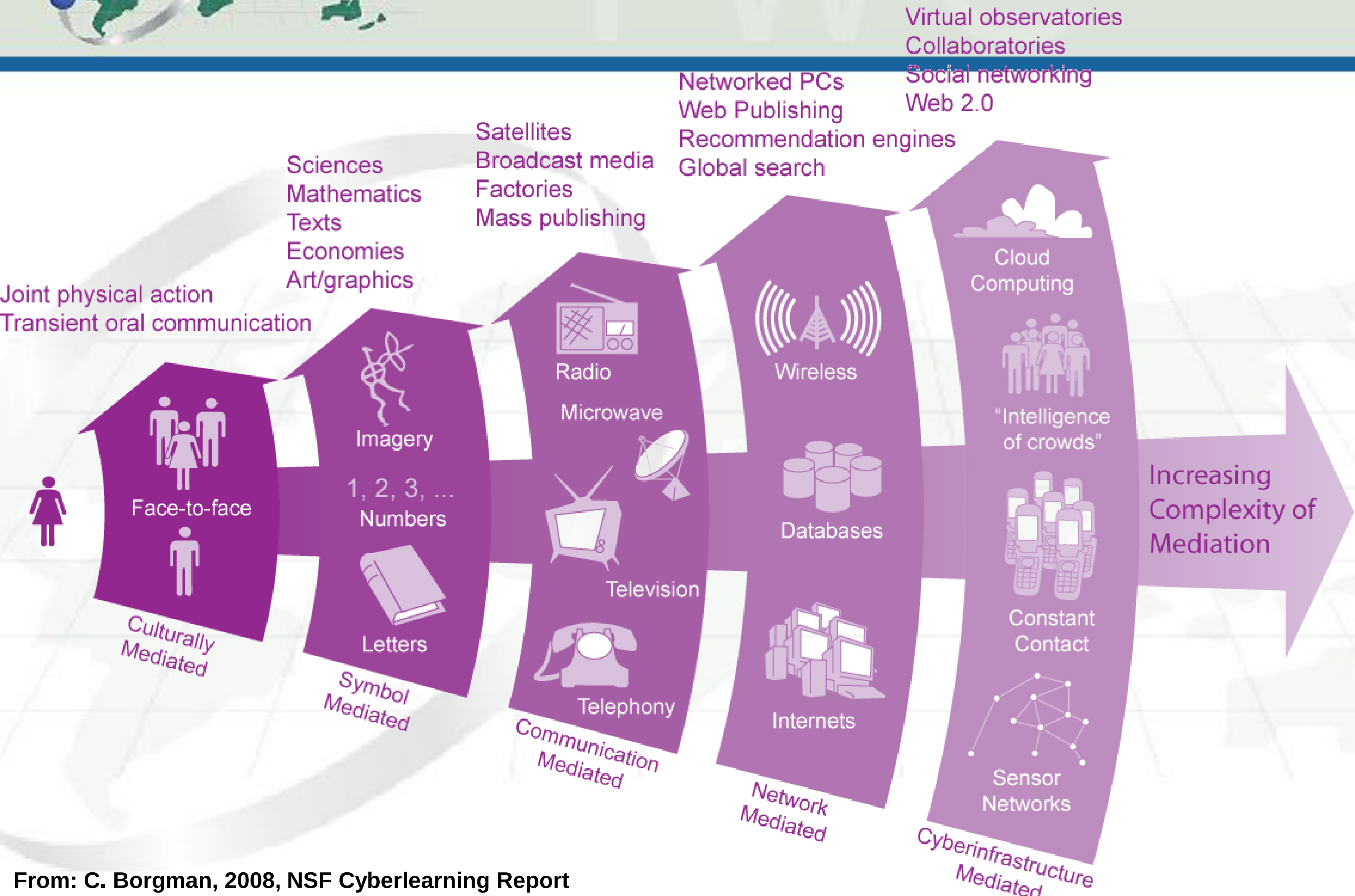
Virtual Observatories Working premise

Scientists – actually ANYONE - should be able to access a global, distributed knowledge base of scientific data that:

- appears to be integrated
- appears to be locally available

Data intensive – volume, complexity, mode, discipline, scale, heterogeneity

Technical advances



Joint physical action
Transient oral communication

Face-to-face

Culturally Mediated

Sciences
Mathematics
Texts
Economies
Art/graphics

Imagery

1, 2, 3, ...
Numbers

Letters

Symbol Mediated

Satellites
Broadcast media
Factories
Mass publishing

Radio

Microwave

Television

Telephony

Communication Mediated

Networked PCs
Web Publishing
Recommendation engines
Global search

Wireless

Databases

Internets

Network Mediated

Virtual observatories
Collaboratories
Social networking
Web 2.0

Cloud Computing

"Intelligence of crowds"

Constant Contact

Sensor Networks

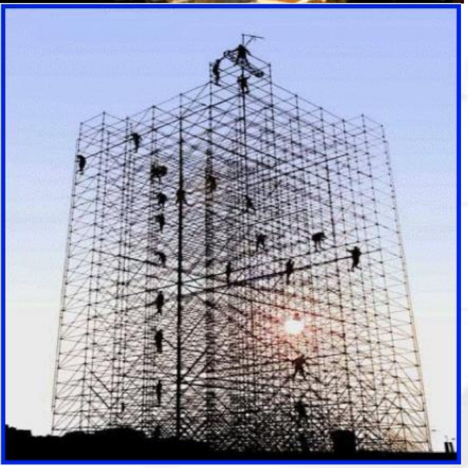
Cyberinfrastructure Mediated

Increasing Complexity of Mediation

From: C. Borgman, 2008, NSF Cyberlearning Report

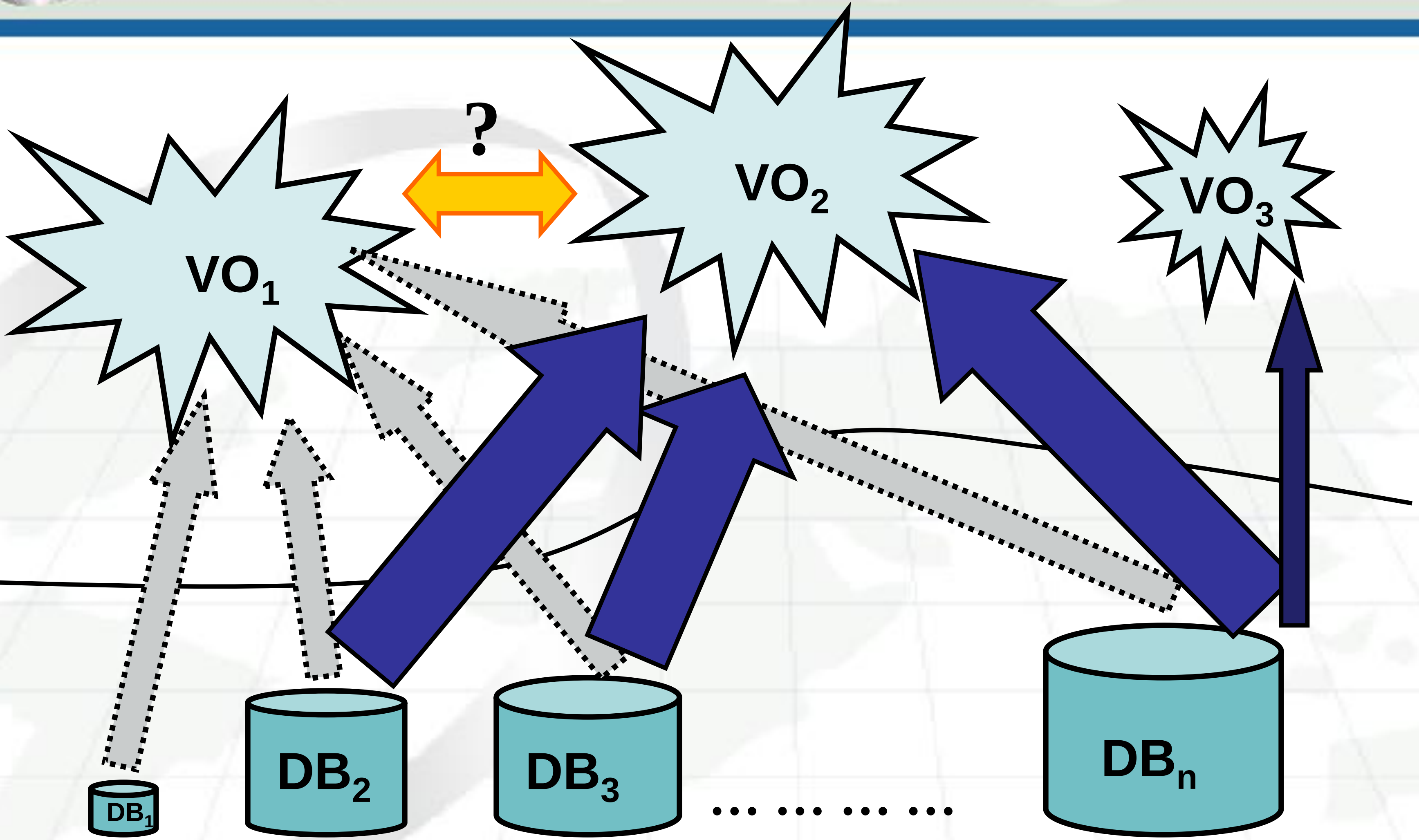
Prior to 2005, we built *systems*

- Rough definitions

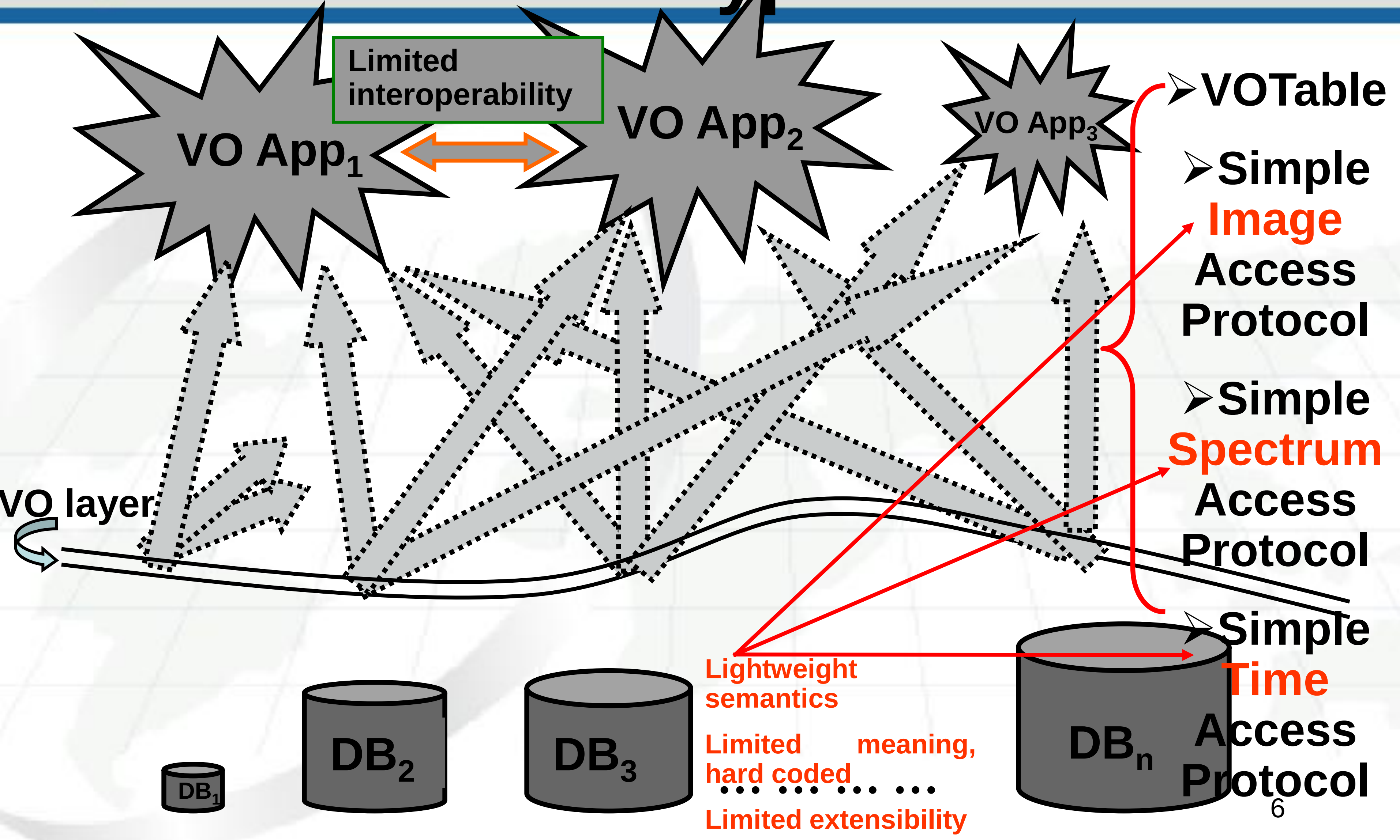


  - Systems have very well-define entry and exit points. A user tends to know when they are using one. Options for extensions are limited and usually require engineering
  - Frameworks have many entry and use points. A user often does not know when they are using one. Extension points are part of the design
  - Modern platforms are built on frameworks

Early days of VxOs



The Astronomy approach; **data-types** as a service



- VOTable
- Simple **Image** Access Protocol
- Simple **Spectrum** Access Protocol
- Simple **Time** Access Protocol

Lightweight semantics

Limited meaning, hard coded

… … … …

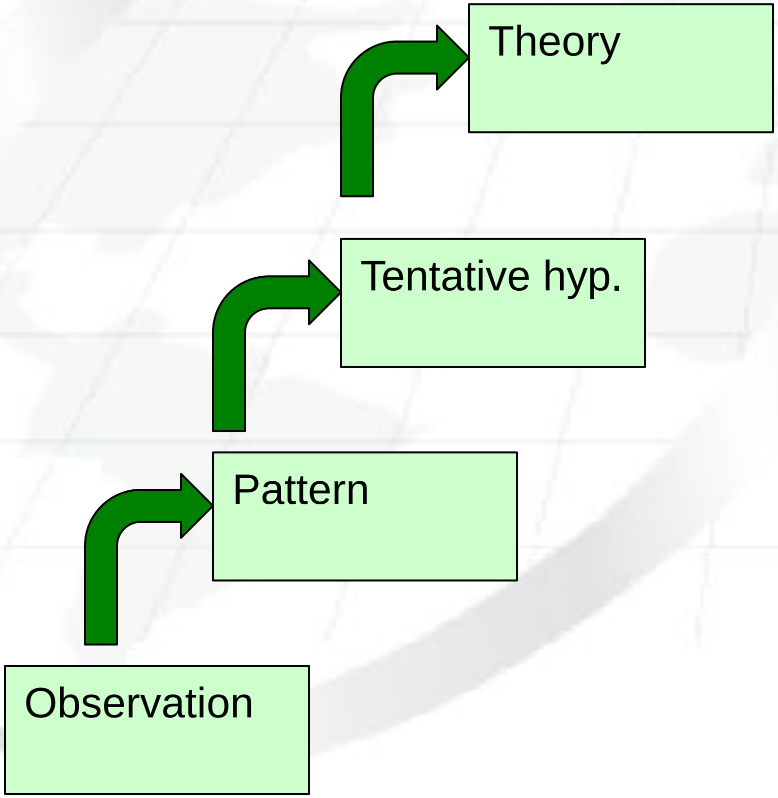
Limited extensibility

VO Standards

- Creation – largely technical activity
- Adoption – largely cultural activity

Means of conduct of research*

- Induction



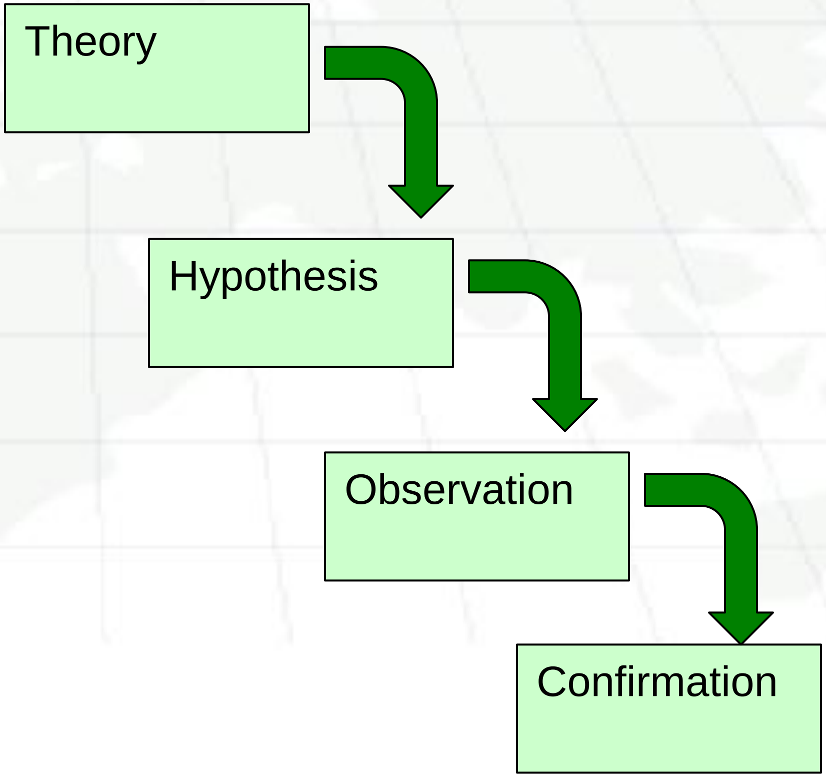
Observation

Pattern

Tentative hyp.

Theory

- Deduction



Theory

Hypothesis

Observation

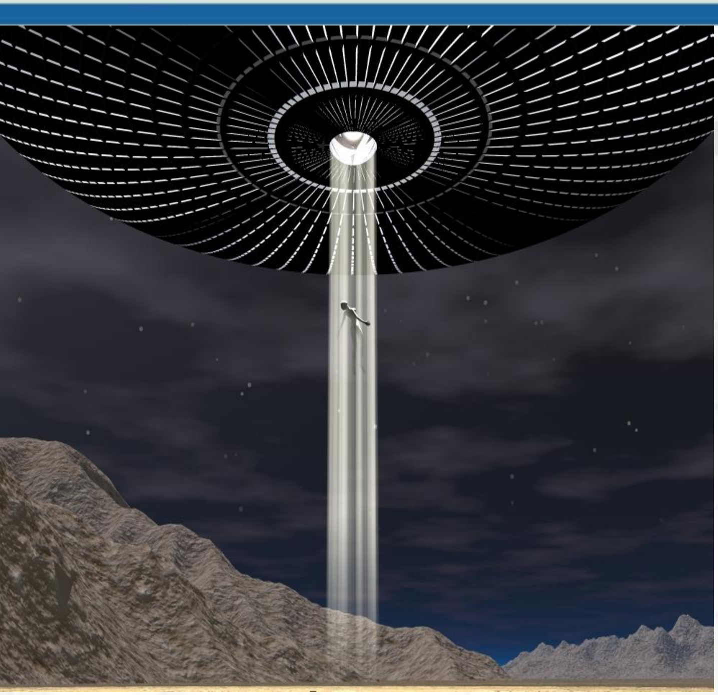
Confirmation

Fundamentally though

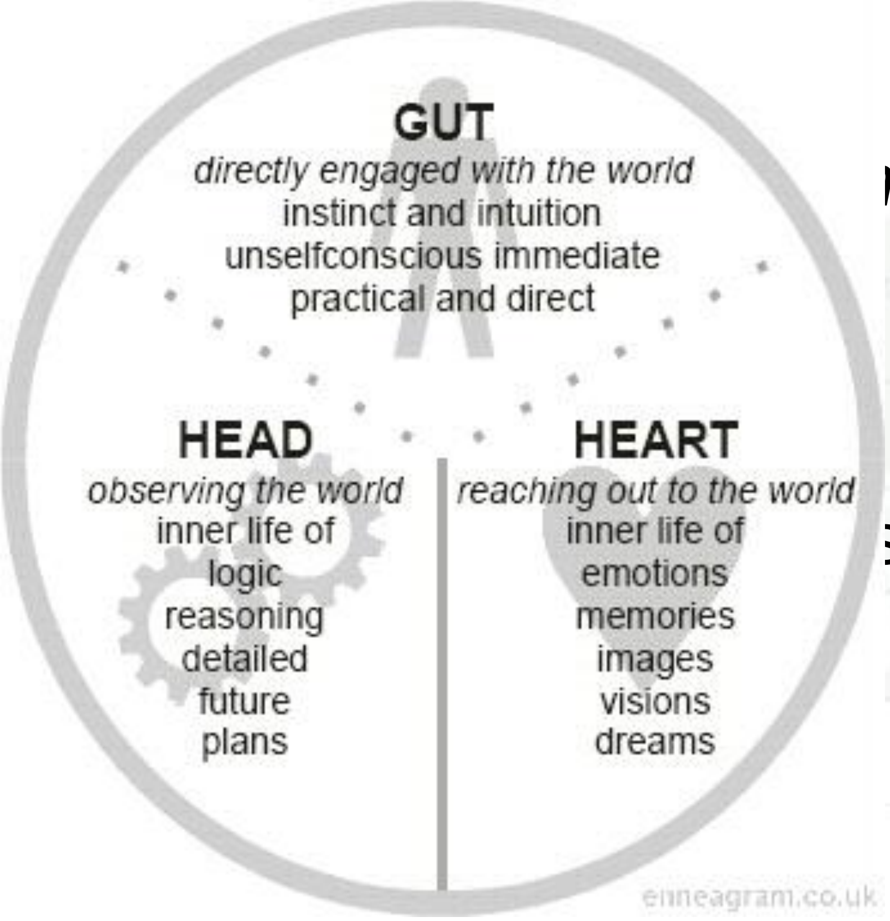
We’ve built capabilities in VOs to support induction or deduction and sometimes both, but does this really enable the breadth of science discoveries we seek in the ERA of bIG dATA?

Edges? In-betweens? Discipline mashups? Accidental? …

For real discovery – we need abduction!



- *ace*
i
v
i
v s
t



Importantly - human intuition is needed in interacting with large-scale data

What should a VO do? Circa 2006

- Make “standard” scientific research much more efficient.
  - Even the principal investigator (PI) teams should want to use them.
  - Must improve on existing services (mission and PI sites, etc.). VOs will not replace these, but will use them in new ways.
  - Access for young researchers, non-experts, other disciplines,
- Enable new, global problems to be solved.
  - Rapidly gain integrated views, e.g. from the solar origin to the terrestrial effects of an event.
  - Find meaningful data related to any particular observation or model.
  - (Ultimately) answer “higher-order” queries such as “Show me the data from cases where a large coronal mass ejection observed by the Solar-Orbiting Heliospheric Observatory was also observed *in situ*.” (science-speak) or “What happens when the Sun disrupts the Earth’s environment” (general public)”

What should a VO do? Circa 2011

- Data science research is routine.
  - Anyone can and does use them.
  - Resource requisition is transparent, anything on demand.
  - Are part of the research infrastructure and data providers do not have to work hard to have many uses of their data.
  - There is a new breed of data publishers.
  - Support full life cycle of data.
- Enable new, global problems to be solved.
  - Services are easily assembled to analyze multiple data streams.
  - Pose questions without an initial detailed knowledge of data sources or origin. Verification and validation is built in.
  - Society relevant use of science data and applications such as planning and decision support are easily supported.

Issues for Next Generation (2-3 yrs) Virtual Observatories

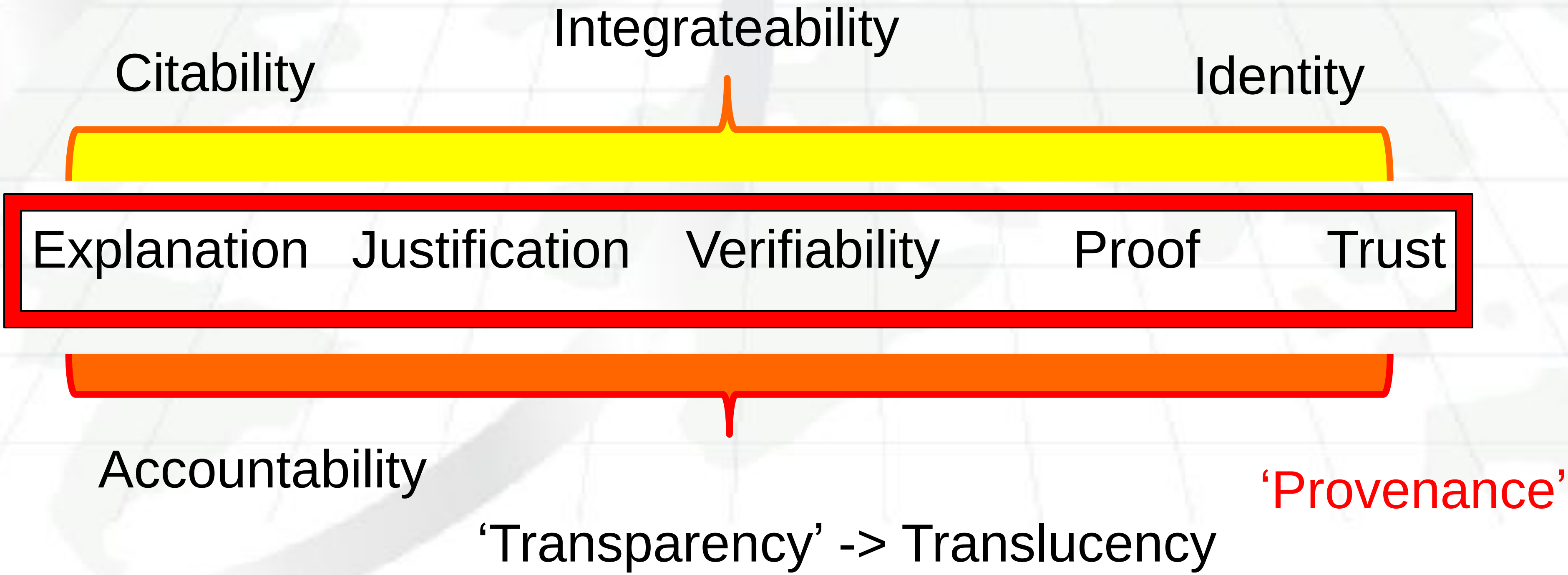
- Unintended and non-specialist use (appropriate and inappropriate)
- Scaling to large numbers of data providers and redefining the role(s) and relations
- Data publication
- Crossing discipline boundaries
- Security, access to resources, policy awareness and enforcement
- Need to survive ‘user testing’, evaluation

Issues for Next Generation Virtual Observatories

- Branding and attribution (where did this data come from and who gets the credit, is it the correct version, is this an authoritative source?)
- Provenance/derivation (propagating key information as it passes through a variety of services, copies of processing algorithms, …)
- Role in data quality, acquisition, curation and preservation

Science ecosystems

- Elements are what enable scientists to explore/ confirm/ deny their research
- Abduction versus induction and deduction



Integrateability

Citability

Identity

Explanation Justification Verifiability Proof Trust

Accountability

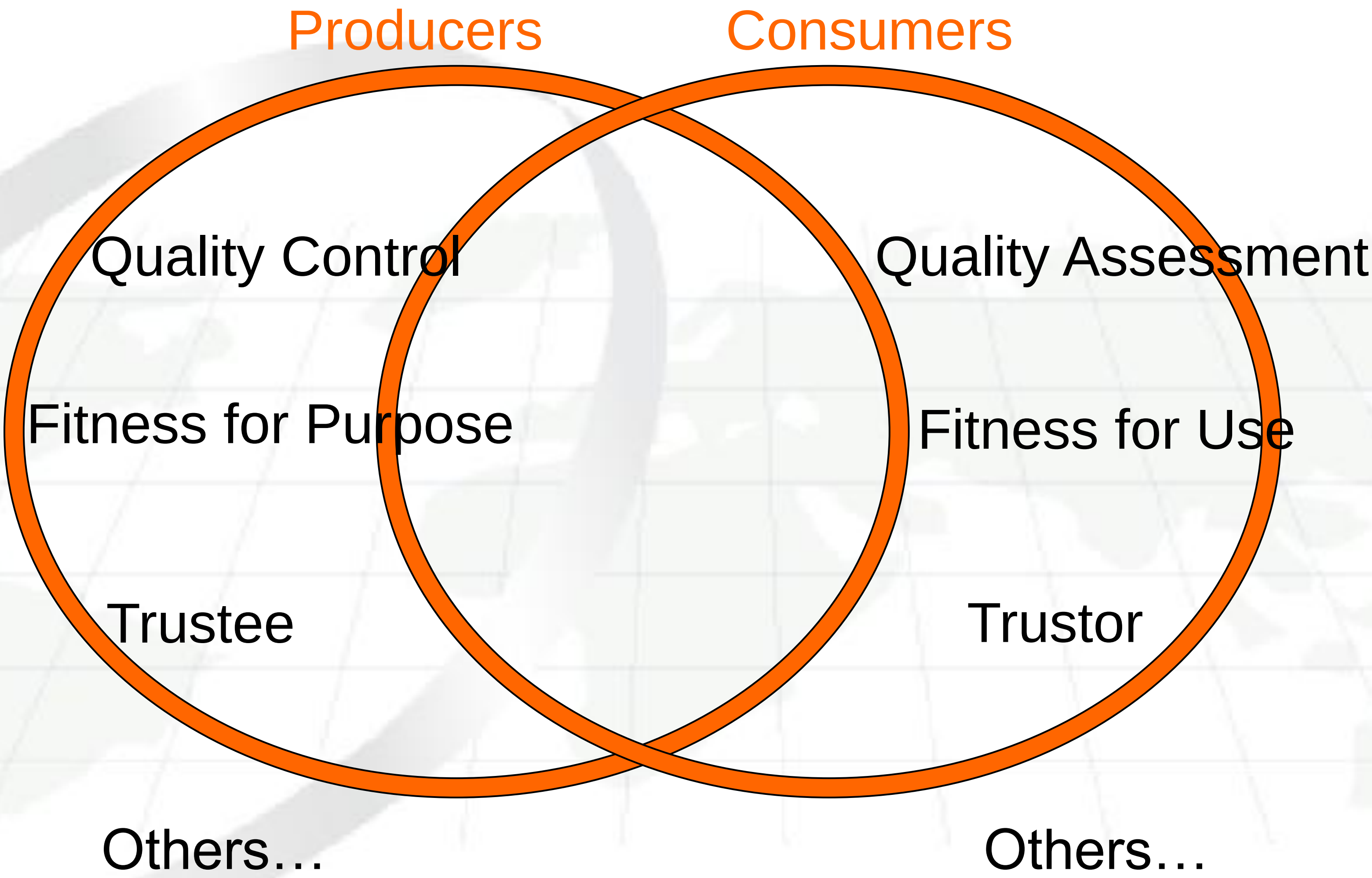
‘Provenance’

‘Transparency’ -> Translucency

Provenance

- Origin or source from which something comes, intention for use, who/what generated for, manner of manufacture, history of subsequent owners, sense of place and time of manufacture, production or discovery, documented in detail sufficient to allow reproducibility or who, what, where, why, when…
- Knowledge provenance; enrich with ontologies and ontology-aware tools
- Provenance presentation is a challenge

VOs and modern curation



Producers

Consumers

Quality Control

Quality Assessment

Fitness for Purpose

Fitness for Use

Trustee

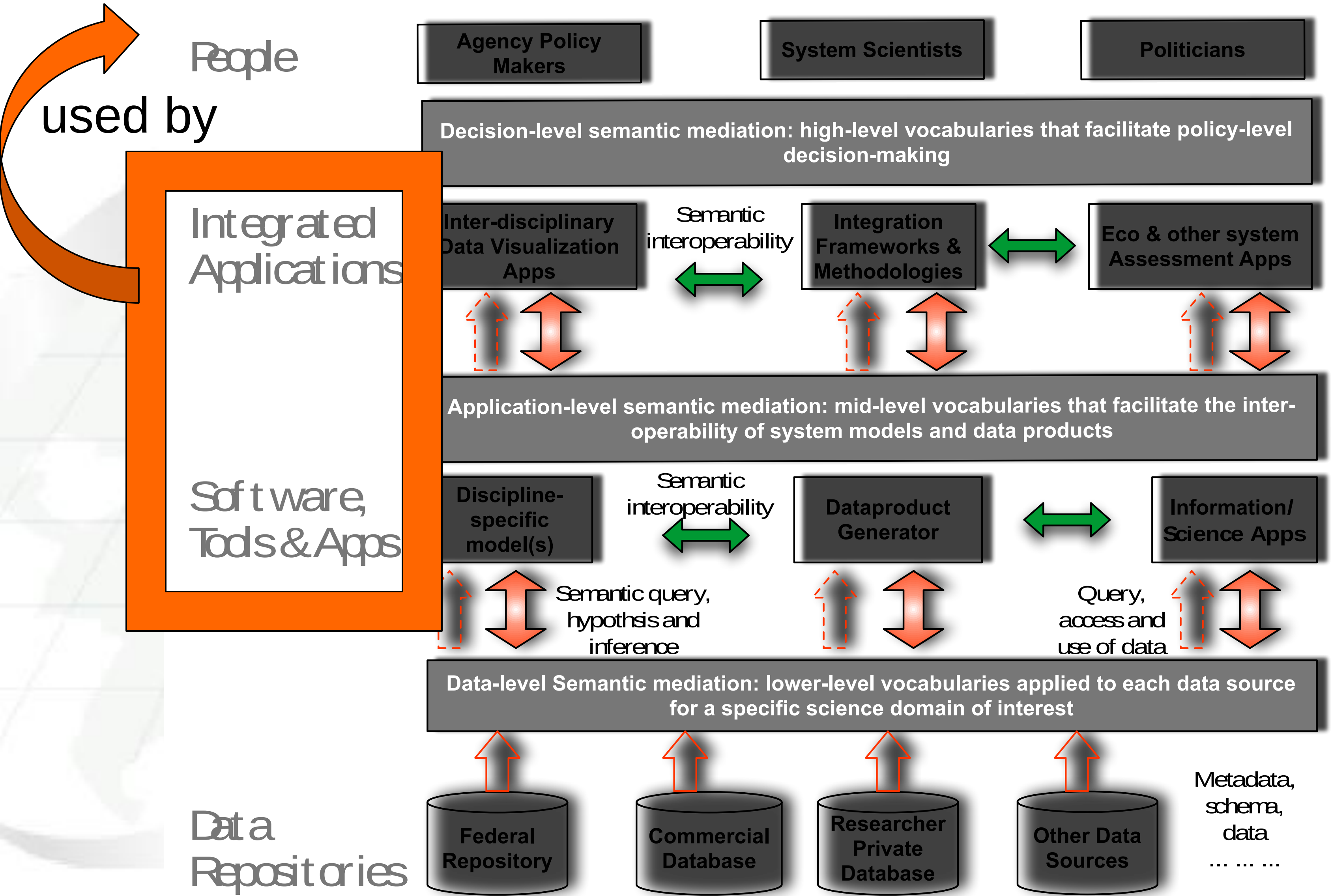
Trustor

Others…

Others…

Skill/ tools?

VOs Architectures – Multi-tiered interoperability



People

used by

Integrated Applications

Software, Tools & Apps

Data Repositories

Agency Policy Makers

System Scientists

Politicians

Decision-level semantic mediation: high-level vocabularies that facilitate policy-level decision-making

Inter-disciplinary Data Visualization Apps

Semantic interoperability

Integration Frameworks & Methodologies

Eco & other system Assessment Apps

Application-level semantic mediation: mid-level vocabularies that facilitate the inter-operability of system models and data products

Discipline-specific model(s)

Semantic interoperability

Dataproduct Generator

Information/ Science Apps

Semantic query, hypothsis and inference

Query, access and use of data

Data-level Semantic mediation: lower-level vocabularies applied to each data source for a specific science domain of interest

Federal Repository

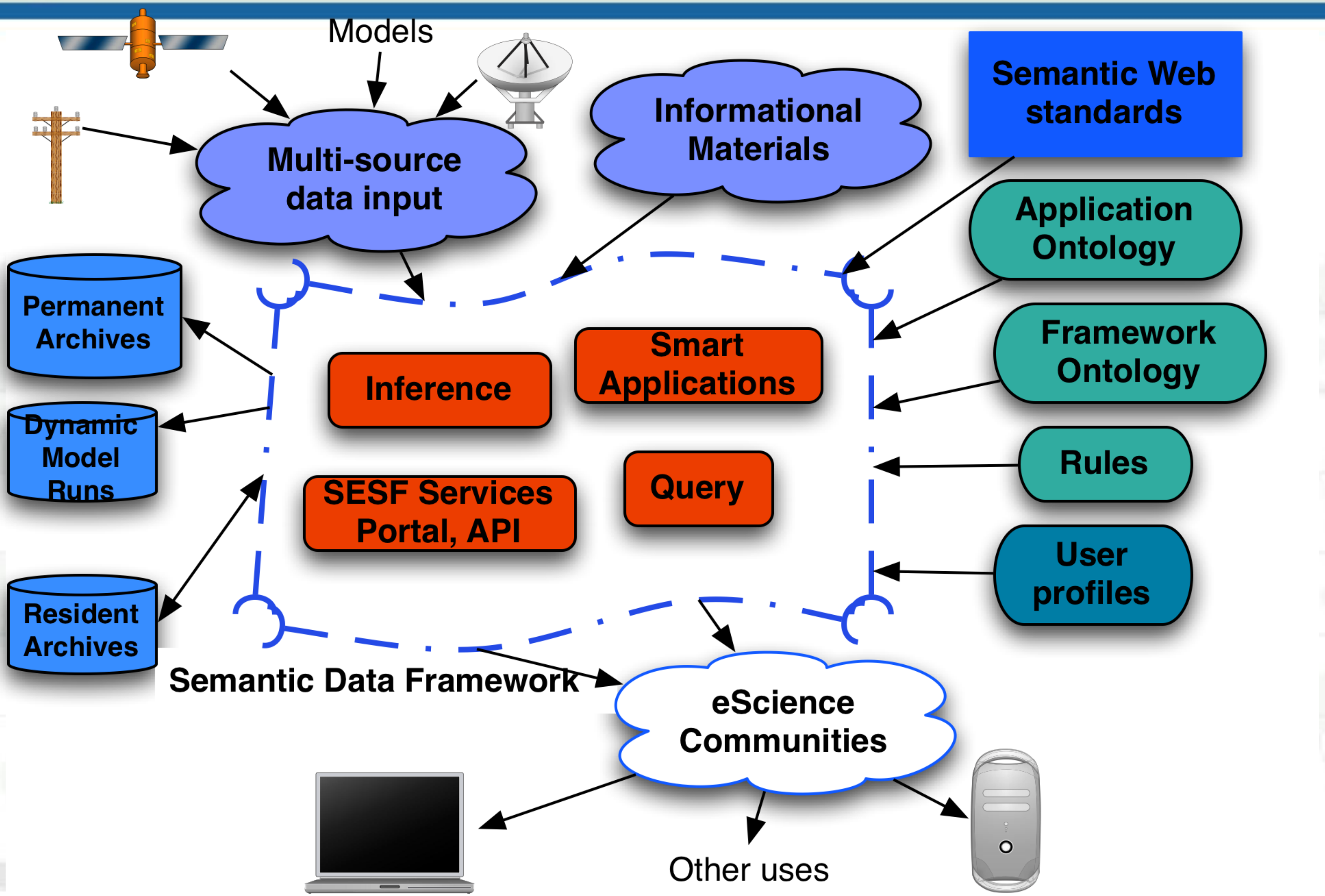
Commercial Database

Researcher Private Database

Other Data Sources

Metadata, schema, data … … …

VO framework…



Discussion

- Significant opportunities for VOs and data as service approaches to ‘scale’ for big data
- Focus on delivering ‘products’ allows analytics on the back end, but tools to plug into a framework are lacking
- Encapsulation is good: hides a lot of inner workings, and bad: opaque, hampers transparency
- Next generation VOs must accommodate: abduction, transparency, interactivity and retain what they do well!
- Thanks. @taswegian, pfox@cs.rpi.edu