Data Science and Analytics Curriculum development at Rensselaer (and the Tetherless World Constellation)

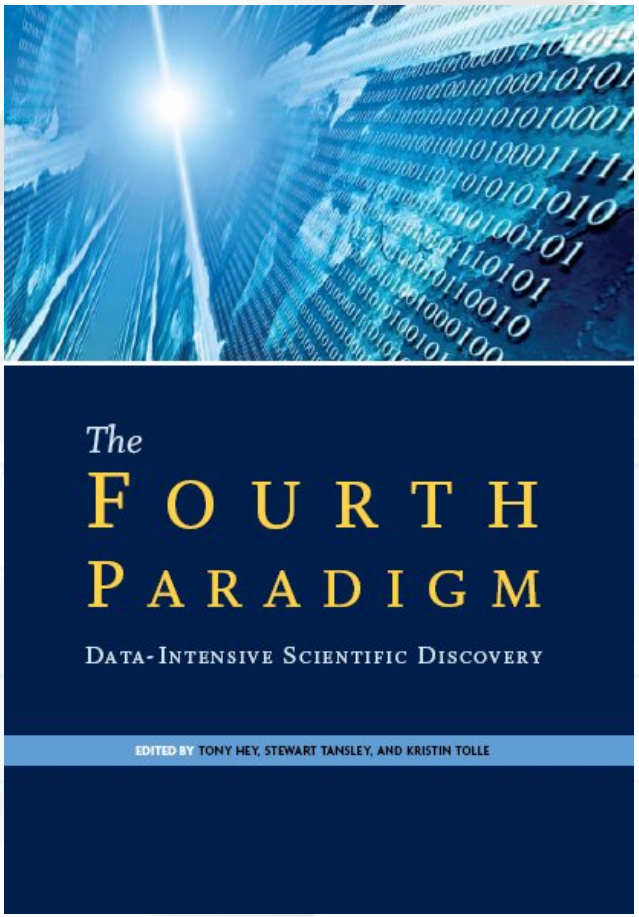
(Adapted from NRC BigData Education Was April 11-12, 2014, Washington DC)

ESIP Federation Summer Meeting,

July 10, 2014

Peter Fox (RPI and WHOI/AOP&E) pfox@cs.rpi.edu, @taswegian
Tetherless World Constellation, http://tw.rpi.edu #twcrpi
Earth and Environmental Science, Computer Science, Cognitive Science, and IT and Web Science

Data is a 1st class citizen



http://thomsonreuters.com/content/press_room/science/686112

tw.rpi.edu

Future Web
•Web Science
•Policy
•Social

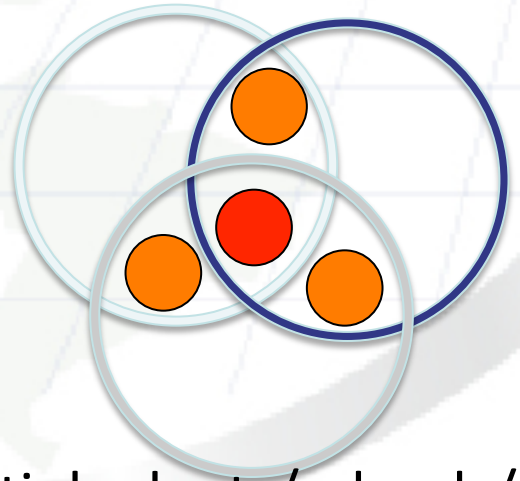
Hendler

Research Themes

Xinformatics
•Data Science
•Semantic eScience
•Data Frameworks

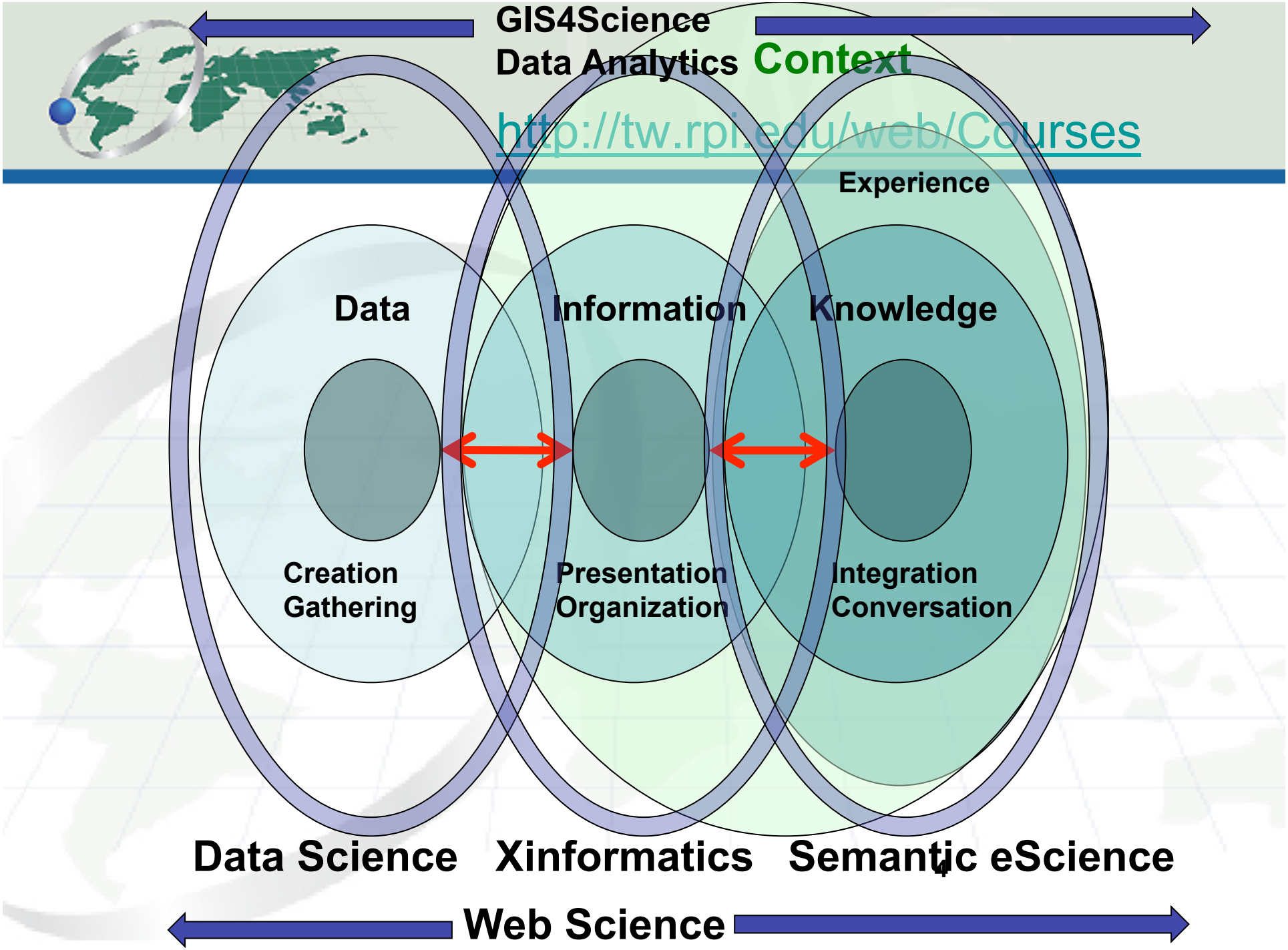
Fox

McGuinness



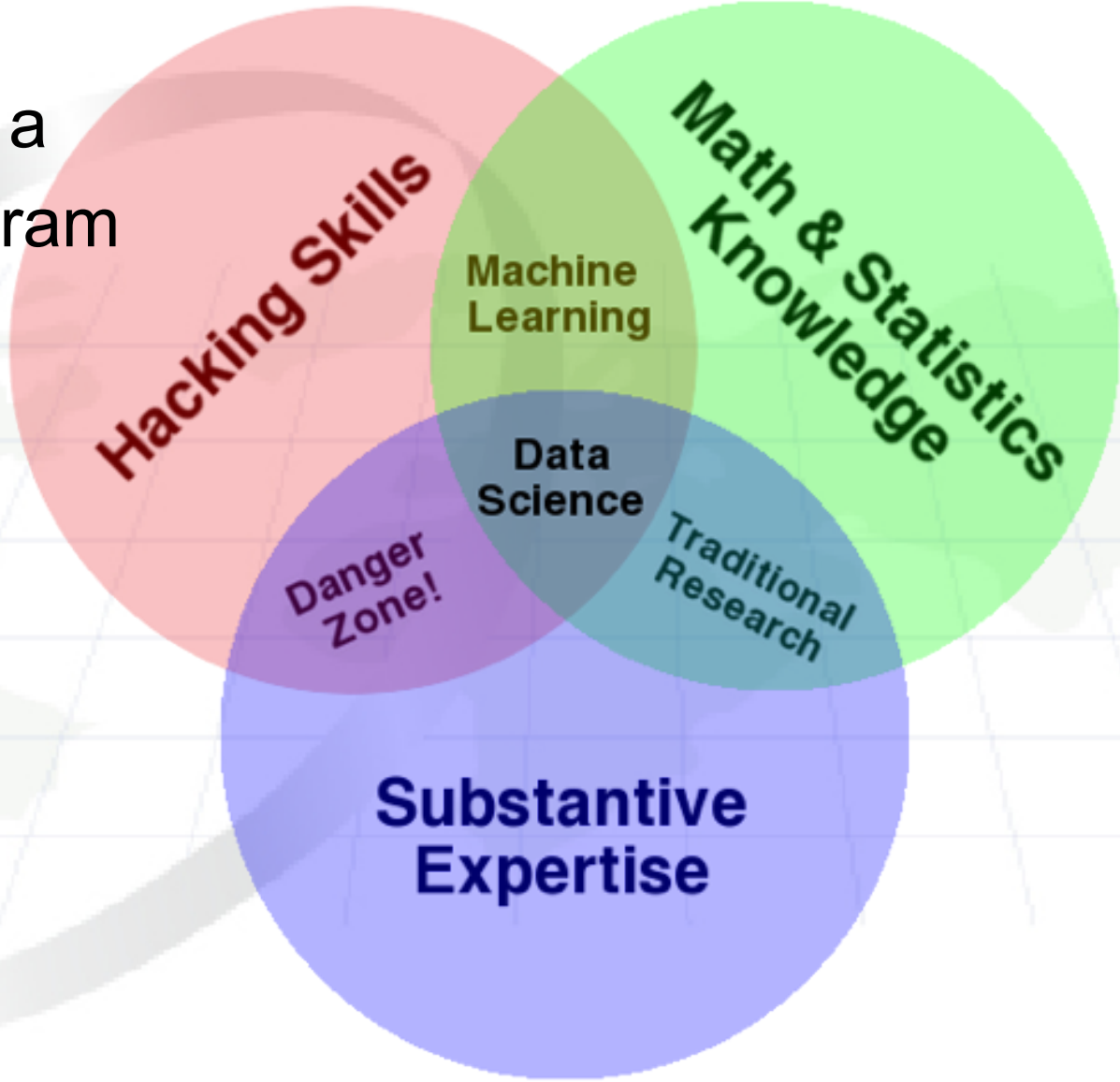
Semantic Foundations
•Knowledge Provenance
•Ontology Engineering Environments
•Inference, Trust

Multiple depts/schools/programs ~ 35 (Post-doc, Staff, Grad, Ugrad)

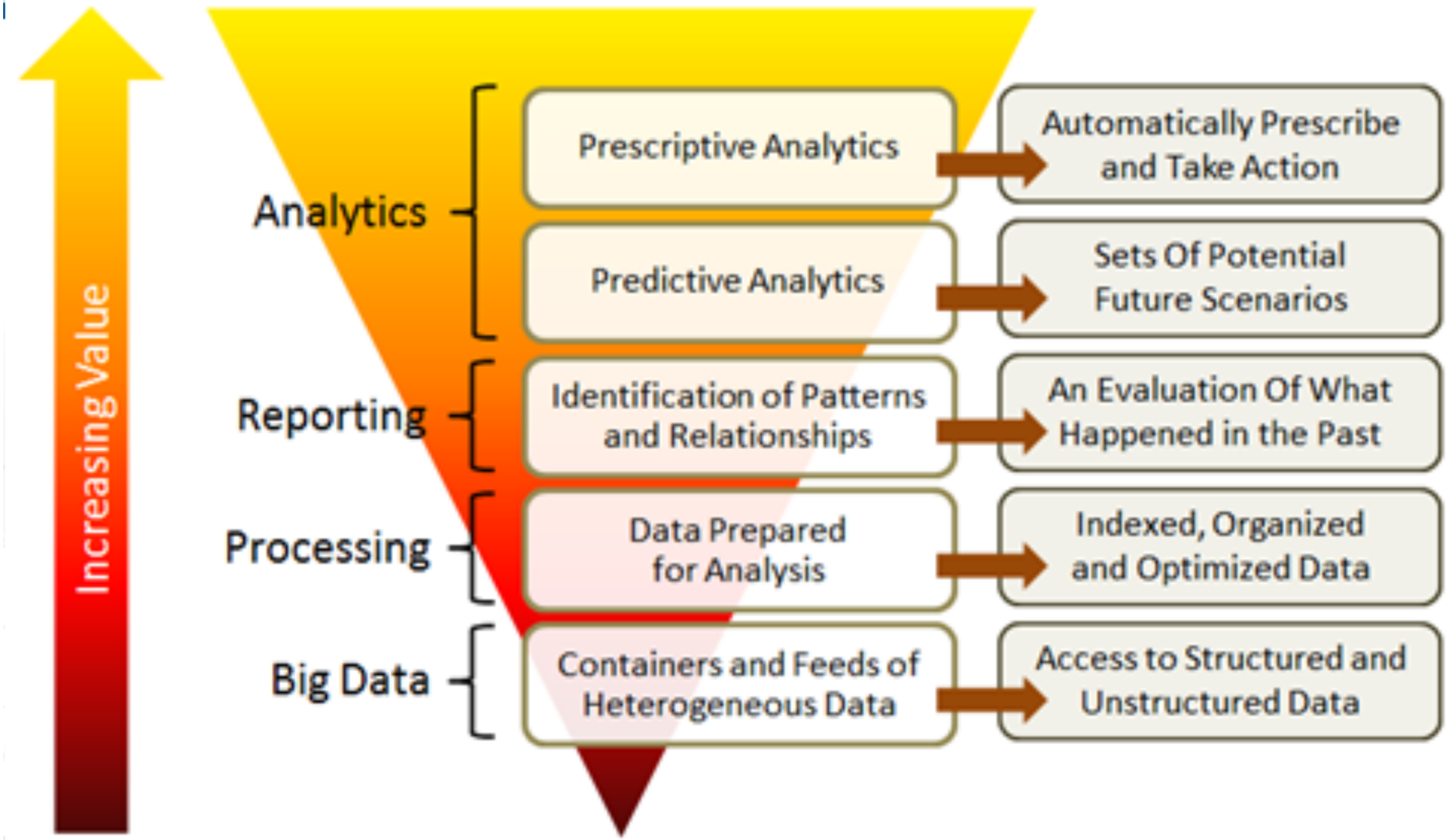

GIS4Science
Data Analytics
Context
http://tw.rpi.edu/web/Courses
Experience
Data
Information
Knowledge
Creation
Gathering
Presentation
Organization
Integration
Conversation
Data Science
Xinformatics
Semantic eScience
Web Science

Data Science

- Beyond a Venn diagram



Data Analytics Challenge



5-6 years in…

- Science and interdisciplinary from the start!
  - Not a question of: do we train scientists to be technical/data people, or do we train technical people to learn the science
  - It's a skill/ course level approach that is needed
  - We teach methodology and principles over technology *
  - Data science must be a *skill*, and natural like using instruments, writing/using codes
  - Team/ collaboration aspects are key **
  - Foundations and theory must be taught ***
- *Multi-disciplinary science program* - PhD in Data and Web Science
- DATUM: Data in Undergraduate Math! (Bennett)
- Missing – intermediate statistics

Science of Data Science

http://online.liebertpub.com/doi/pdfplus/10.1089/big.2014.0011

TABLE 1. CROSS-CUTTING DATA SCIENCE RESEARCH CHALLENGES

1. Understanding scale in systems:
   a. Inverse problems using multimodal data: scale variation, dealing with sample bias as a result
   b. State of systems complexity as scales are traversed (e.g., gene/cell/tissue/organ/organism)
   c. Big parameter spaces (~100 dimensions) that change the game in probability-based and Bayesian analytics
2. Sparse systems:
   a. Complex systems with incomplete data (sparse)
   b. Adjoint/variational data-assimilation approaches in highly nonlinear, heterogeneous, stiff systems
3. Abductive reasoning:
   a. Open-world uncertainty quantification
   b. Accommodating cognitive computing functionalities in data and information infrastructures
4. Next-generation semantic data infrastructure
   a. Tools for supporting data reuse and integration