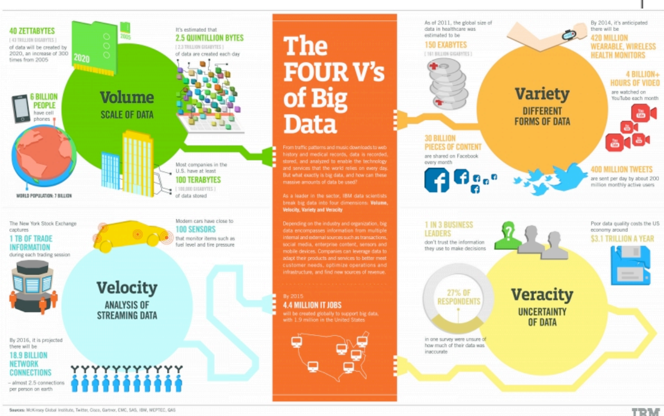
Linking collections to related resources: Multi-scale, multi-dimensional, multi-disciplinary collaborative research in biodiversity. Is this a "Big Data" paradigm?

Reed Beaman,
, University of Florida, Gainesville, FL, USA

PRAGMA 26: 10 April 2014

The 3 or 4 Vs of Big Data



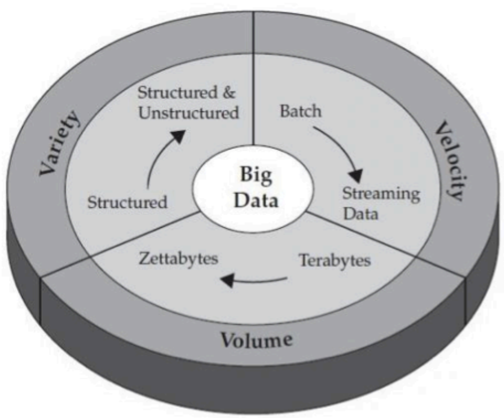
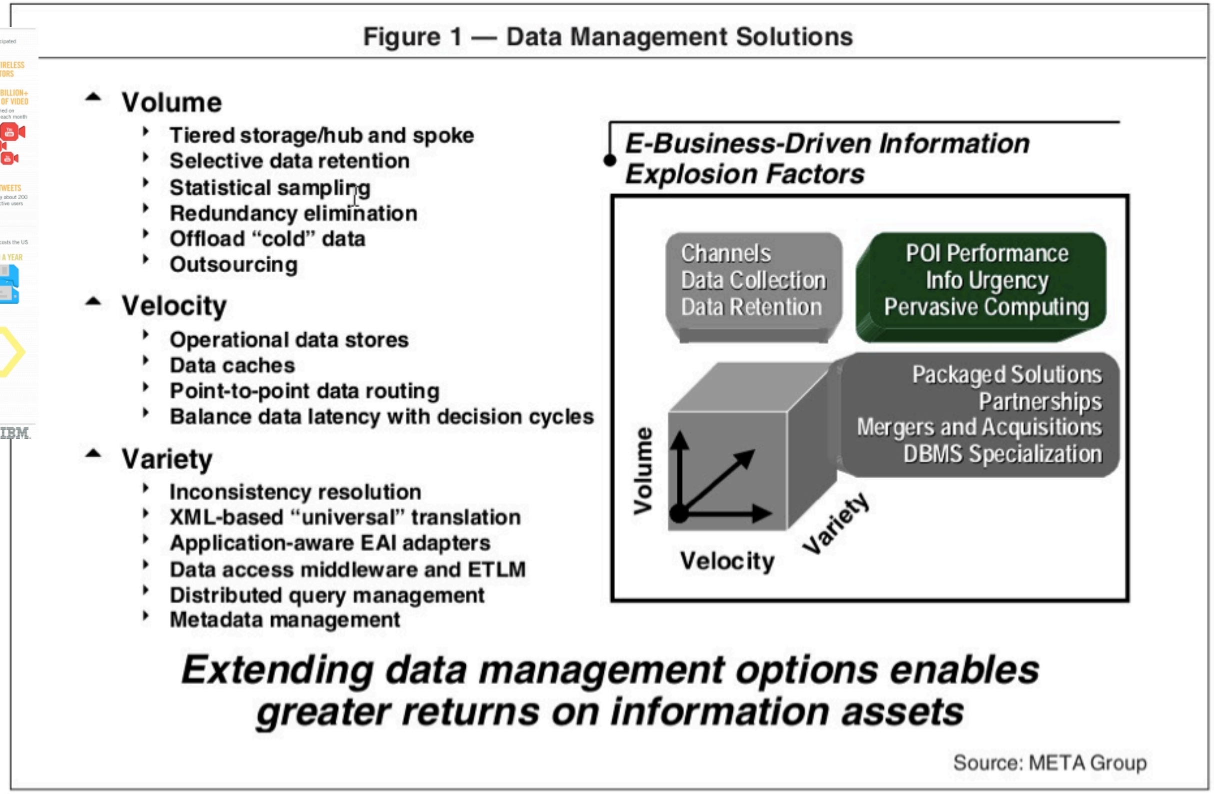


Figure 1-1 *IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.*



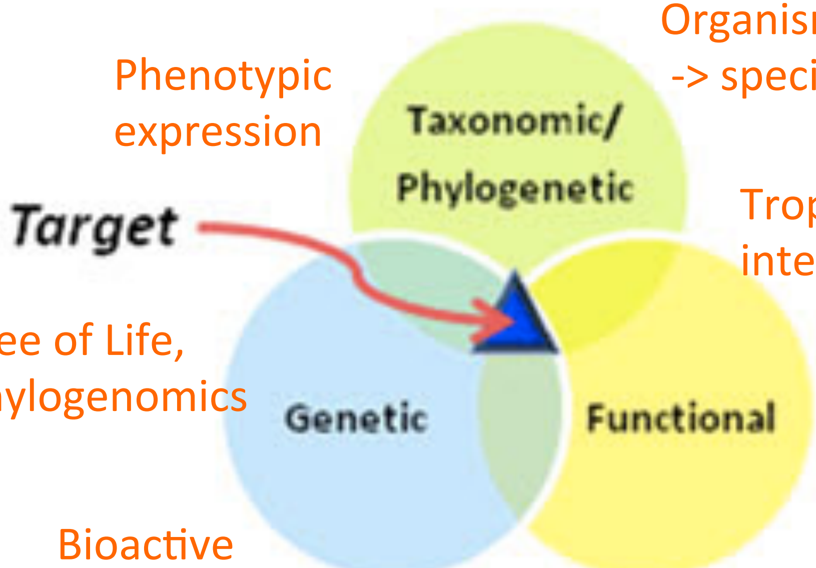
"Big data is data that's an order of magnitude bigger than you're accustomed to, Grasshopper." Doug Laney, Gartner

Integrative Biodiversity: Multiscale, Multi-disciplinary

- US NSF Dimensions of Biodiversity program)
  - Interaction at the intersection of taxonomic, genetic, functional domains

Phenotypic expression

Organisms -> species

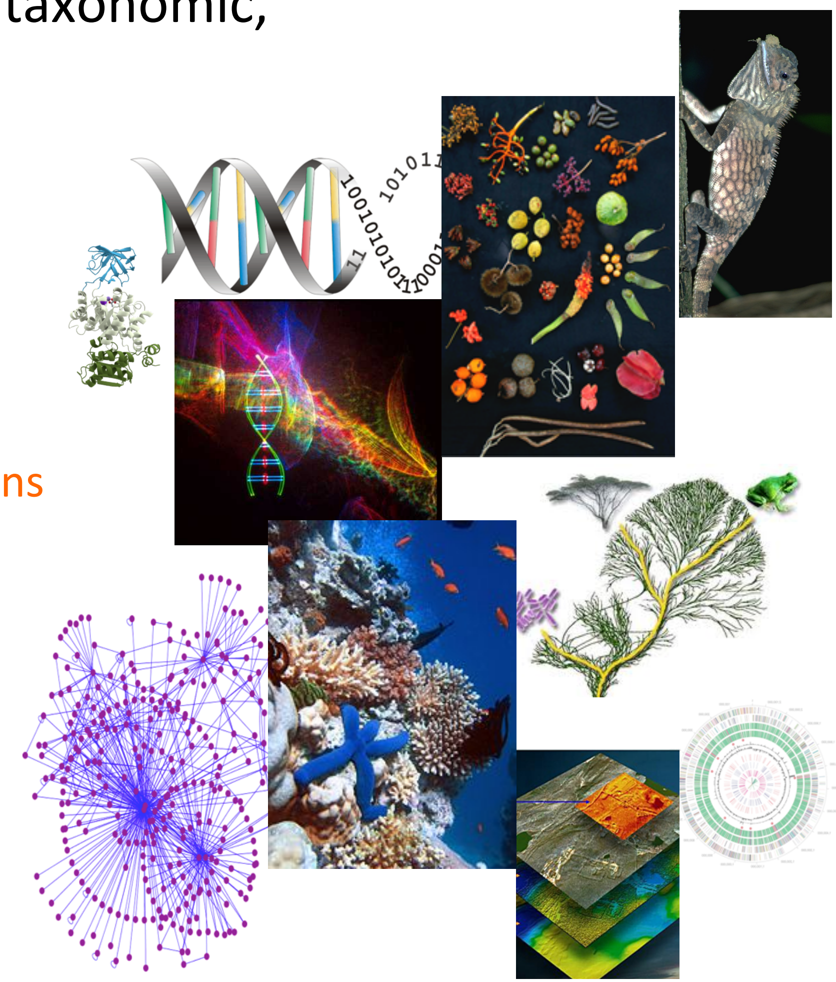


Trophic interations

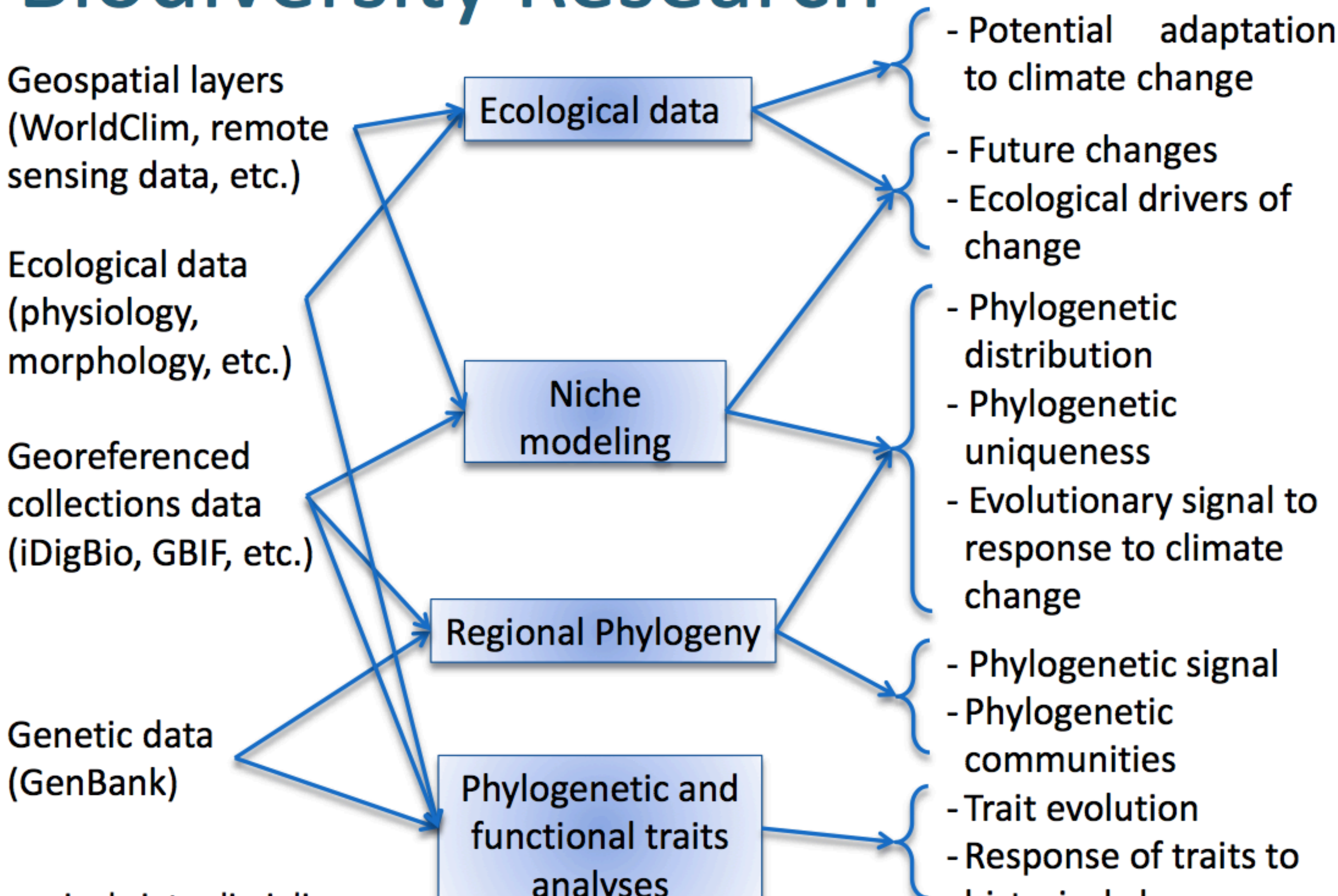
Tree of Life, phylogenomics

Bioactive compounds/chemistry

Molecular -> Ecosystem



Biodiversity Research



Increasingly interdisciplinary

Source: Andrea Matsunaga

Big data is a given for genomics, high throughput sequencing, analysis, and visualization

What about all the other data that ***relates*** to genetic and genomic data?

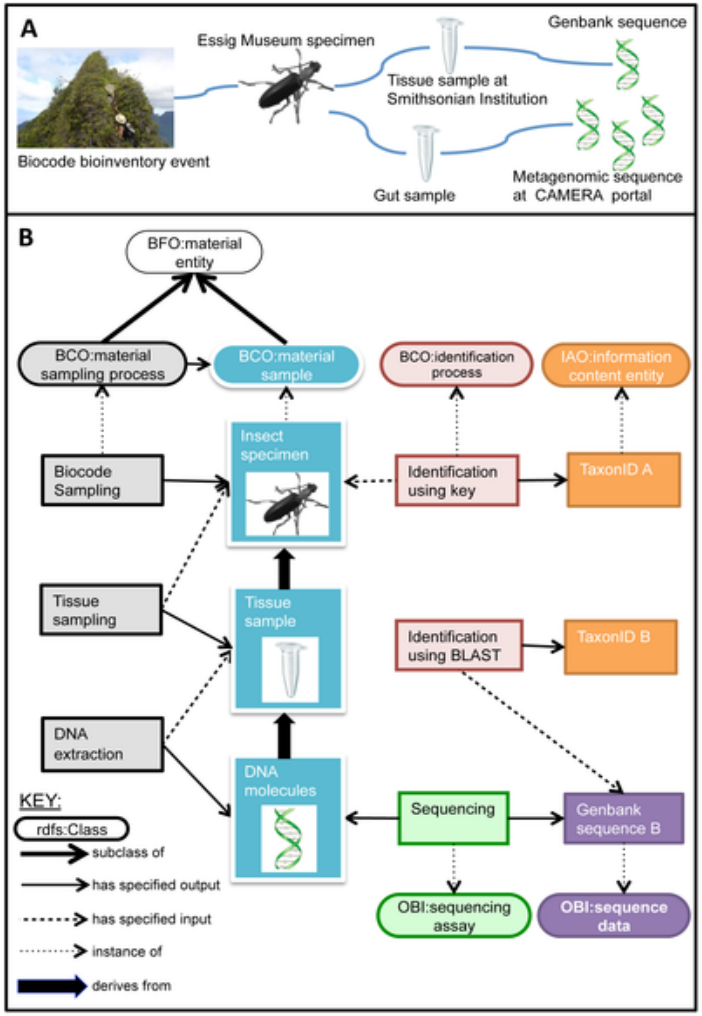
4Vs for Biodiversity Big data

- Volume: billion or more specimens, 2-10 million species (excluding microbial), 10 billion plus related edges
- Velocity: Snail's pace? 250 year long-tail legacy of taxonomic data -> rapid digitization <-> large scale genomic sequencing
- Variety: Occurrences, sequences, morphological, geospatial; structured and unstructured
- Veracity: Very challenging to validate?

Figure 3. Linking samples and derivatives from the Moorea Biocode project.

BiSciCol (Biological Science Collections Tracker) use case:

Every specimen links to a multitude of parent and derivative data. Users of biodiversity data need to be able to ***easily and quickly*** see these relationships

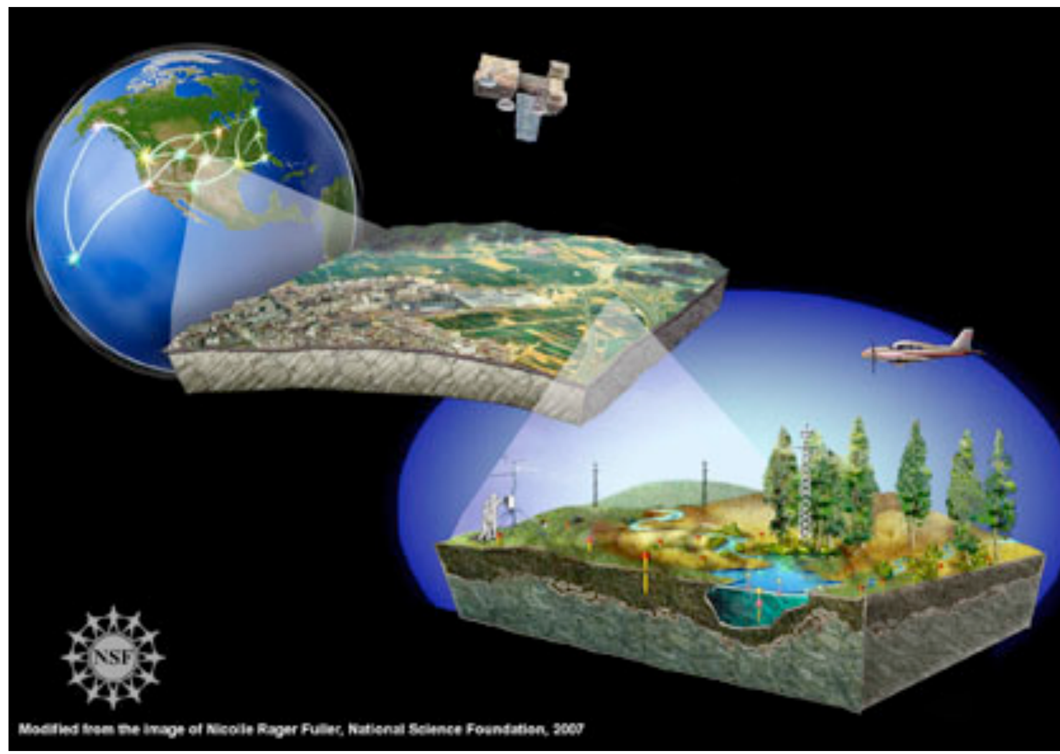


Citation: Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, et al. (2014) Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PLoS ONE 9(3): e89606. doi:10.1371/journal.pone.0089606 http://www.plosone.org/article/info:doi/10.1371/journal.pone.0089606

The "Big" in Ecological Big Data

The defining aspect of ecological Big Data is not raw size but another dimension: complexity.

Dave Schimel, (former) NEON Chief Scientist

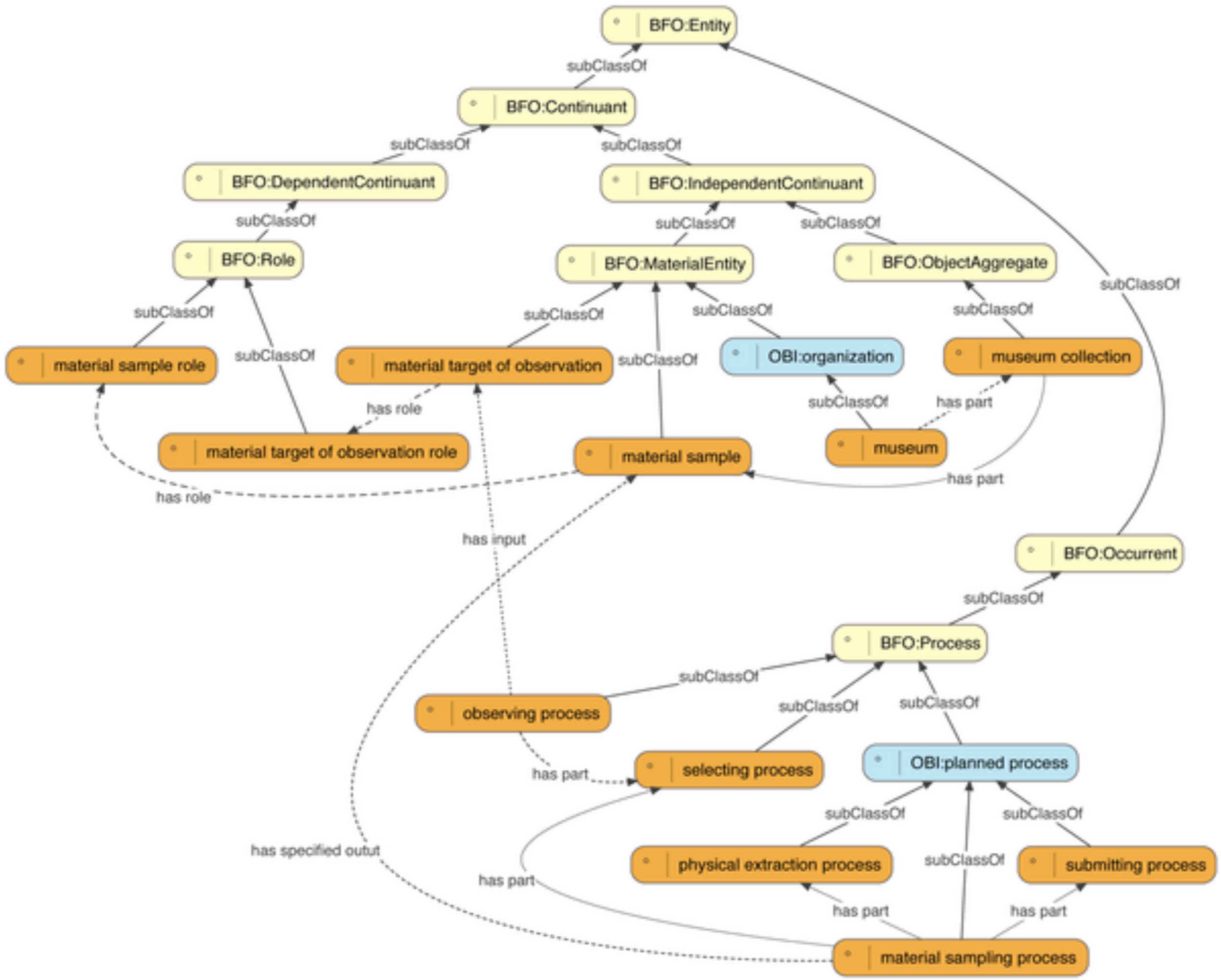


Modified from the image of Nicolle Rager Fuller, National Science Foundation, 2007

4Cs of Biodiversity Big Data

- Complexity: scale, interactions (e.g., food webs) between individuals, populations, species, environments (cf. story lines)
- Collaboration: International and multidisciplinary
- Citizen Science: Increasing as a solution to digitization
- Completeness: Will we always be 10% complete, and can we validate and create the linkages?

Figure 2. Core terms of the Biological Collections Ontology (BCO) and their relations to upper ontologies.

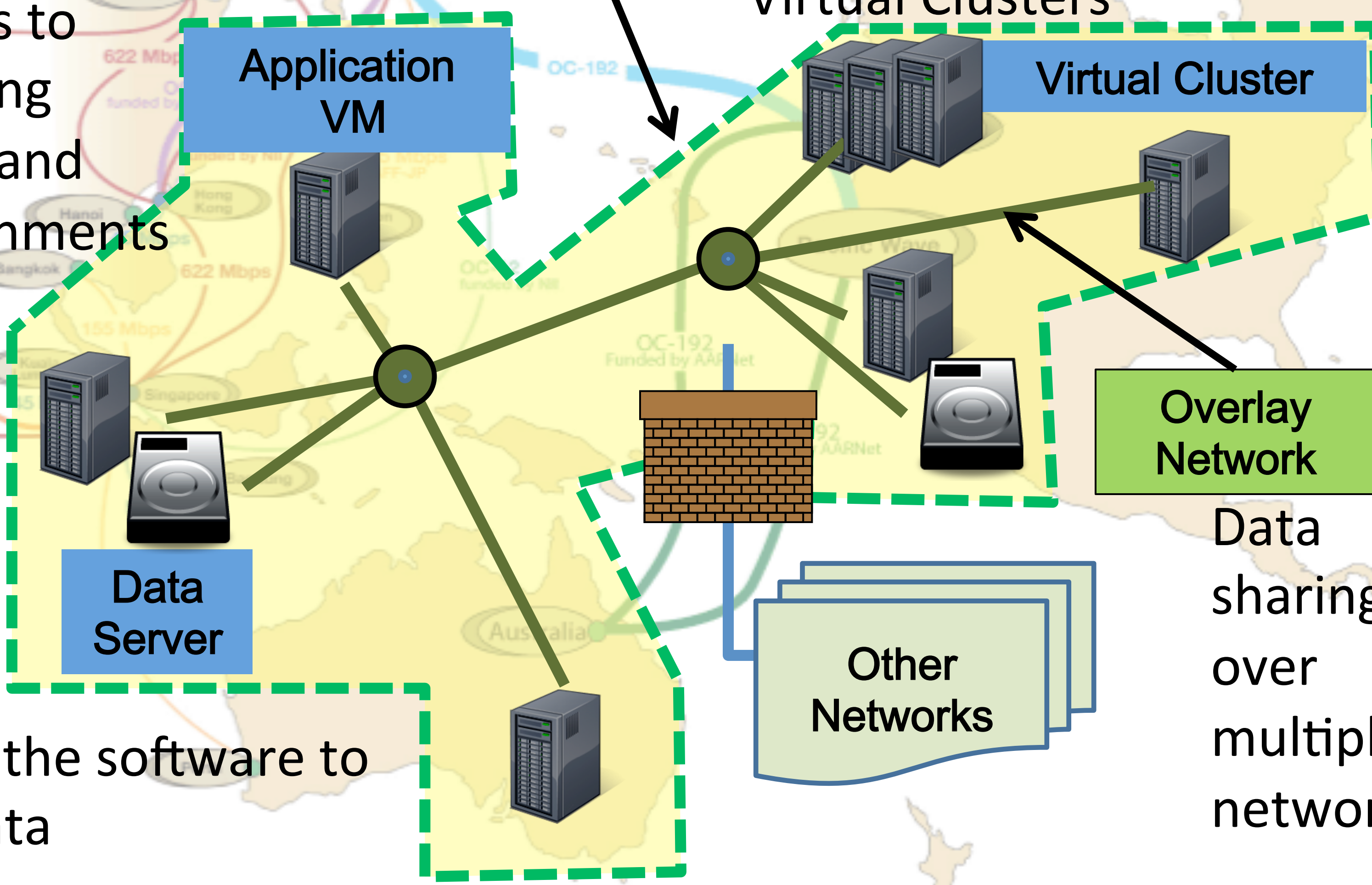


Citation: Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, et al. (2014) Semantics in Support of Biodiversity Knowledge Discovery: An Introduction to the Biological Collections Ontology and Related Ontologies. PLoS ONE 9(3): e89606. doi:10.1371/journal.pone.0089606
httphttp://www.plosone.org/article/info:doi/10.1371/journal.pone.0089606

Software Defined System

From Virtual Machines to Virtual Clusters

Adjusts to changing needs and environments

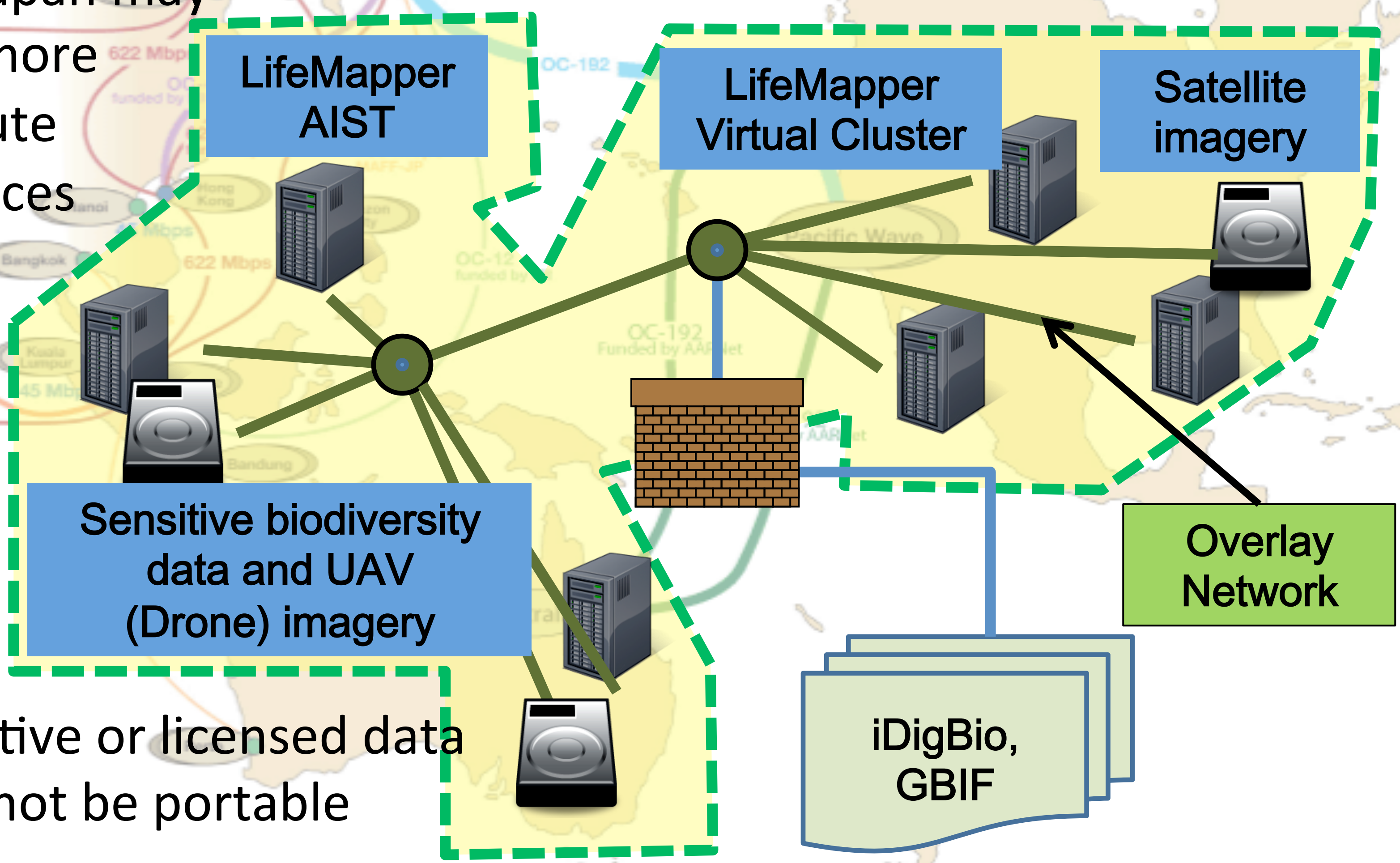


Data sharing over multiple networks

Move the software to the data

Trust Envelope

AIST Japan may have more compute resources



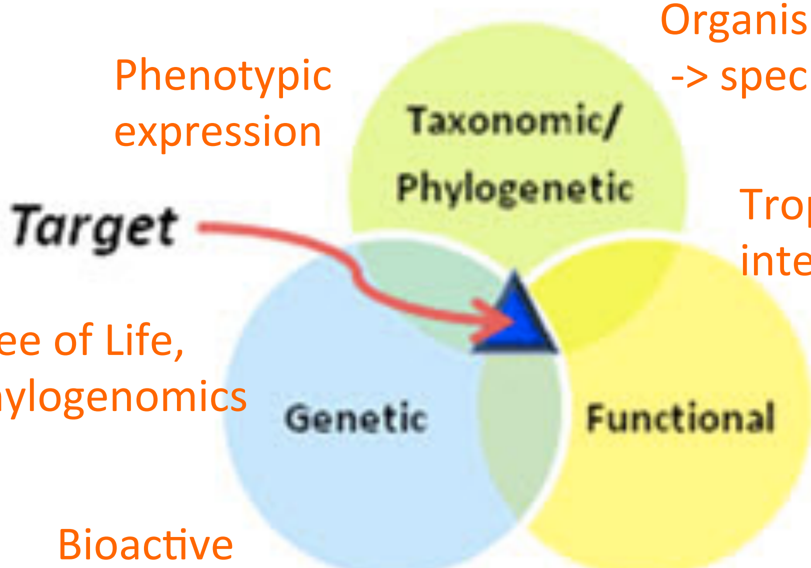
Sensitive or licensed data may not be portable

Integrative Biodiversity

- Grand challenge science: Big data is about asking big questions

Phenotypic expression

Organisms -> species



Trophic interations

Tree of Life, phylogenomics

Bioactive compounds/chemistry

Molecular -> Ecosystem