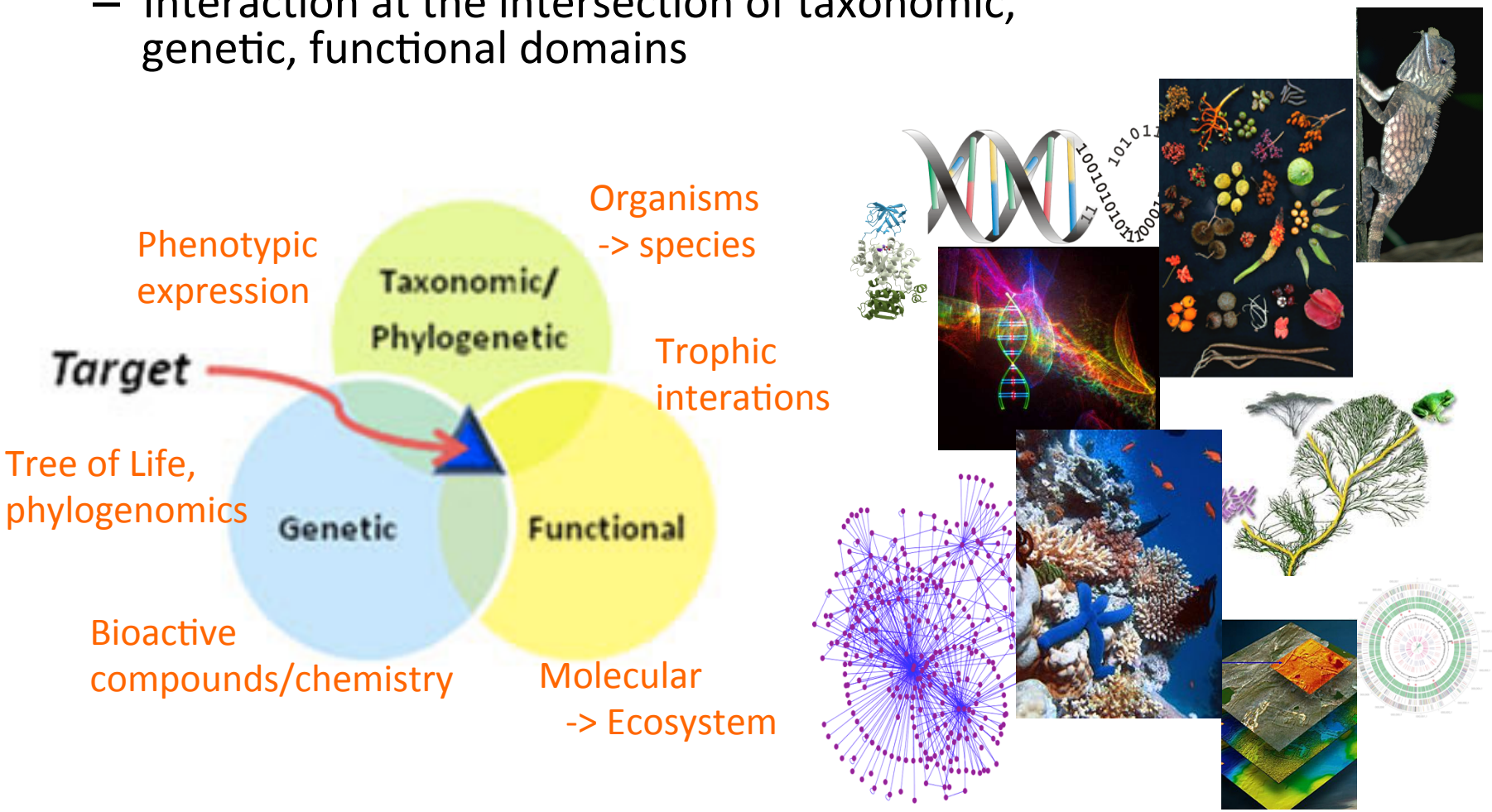
Linking collections to related resources: Multi-scale, multi-dimensional, multi-disciplinary collaborative research in biodiversity. Is this a "Big Data" paradigm?

Reed Beaman,
Florida Museum of Natural History,
University of Florida, Gainesville, FL, USA

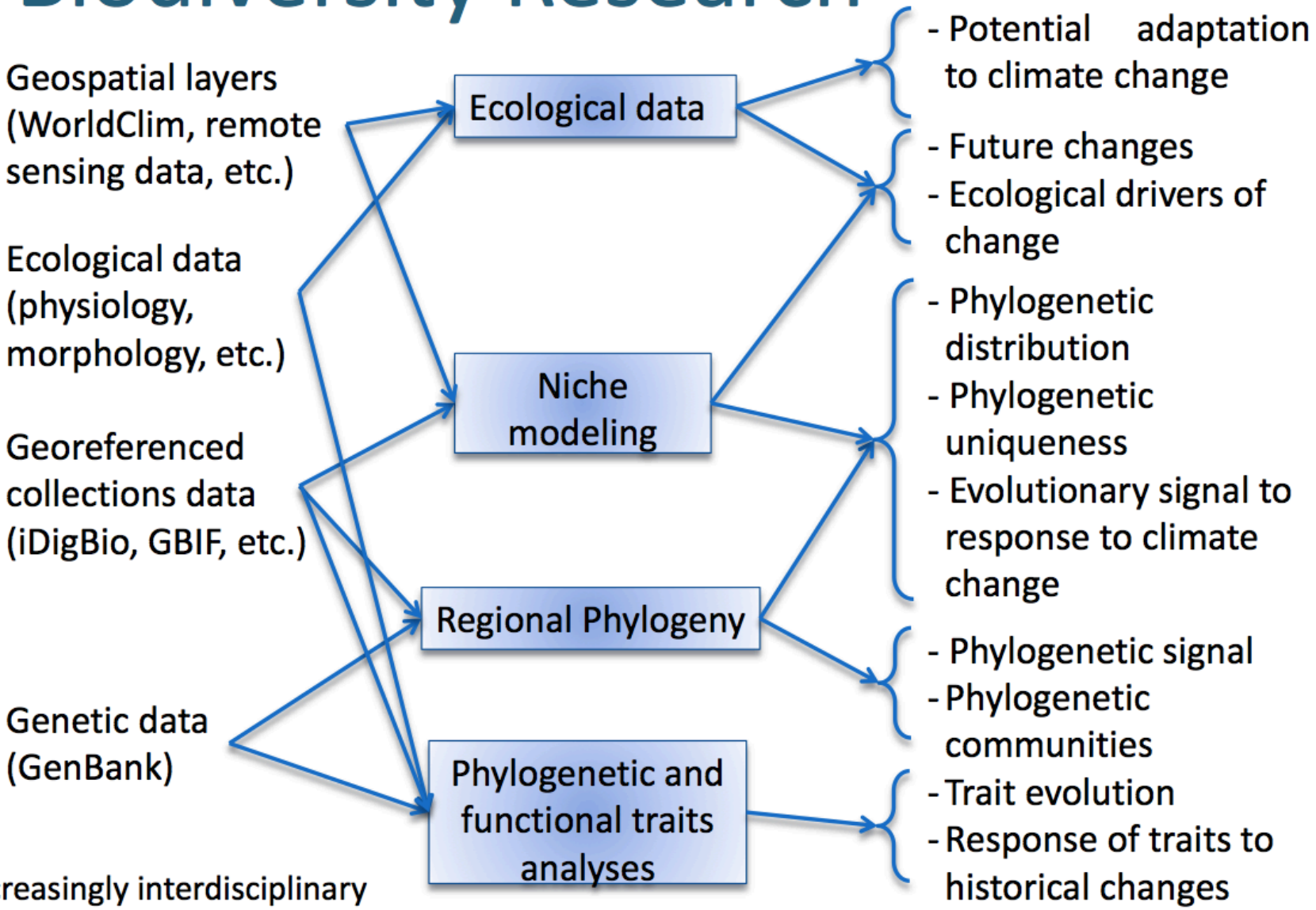
27 March 2014

Integrative Biodiversity: It's about the science questions

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



Biodiversity Research

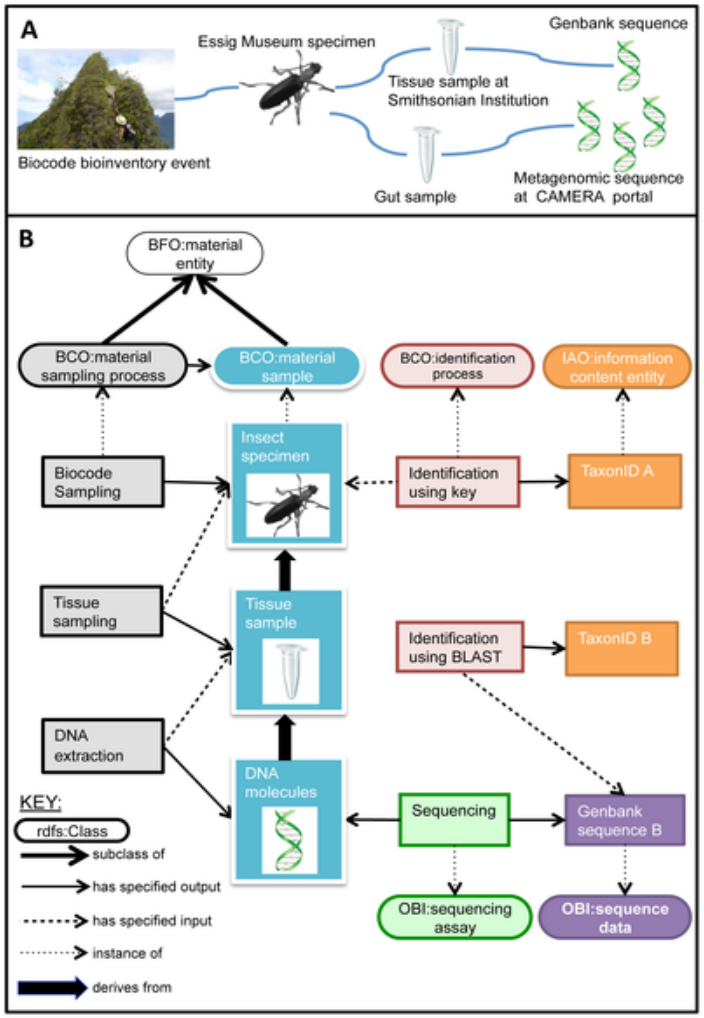


Increasingly interdisciplinary

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

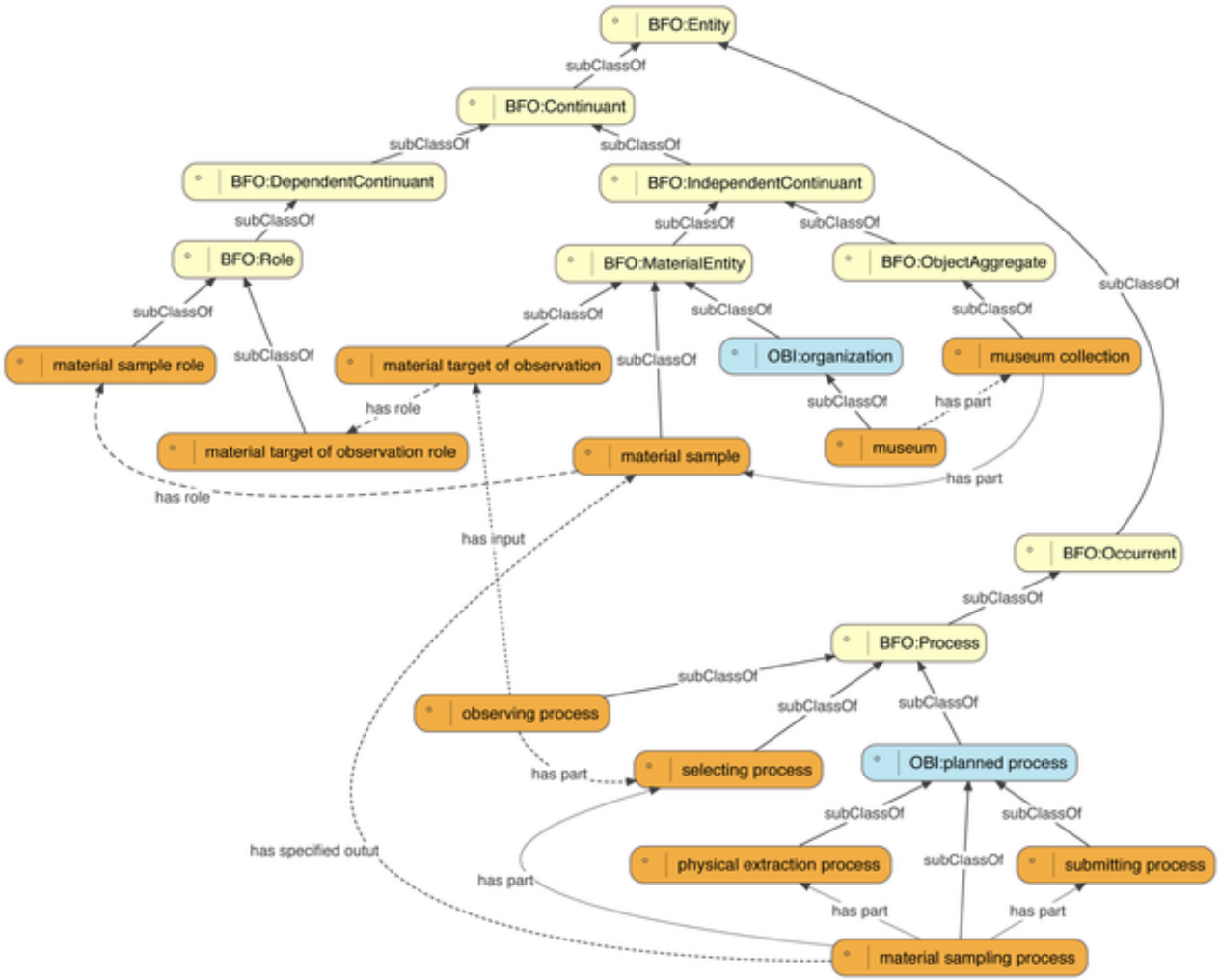
BiSciCol (Biological Science Collections Tracker) use case:

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

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](http://www.plosone.org/article/info:doi/10.1371/journal.pone.0089606)

GUID Land

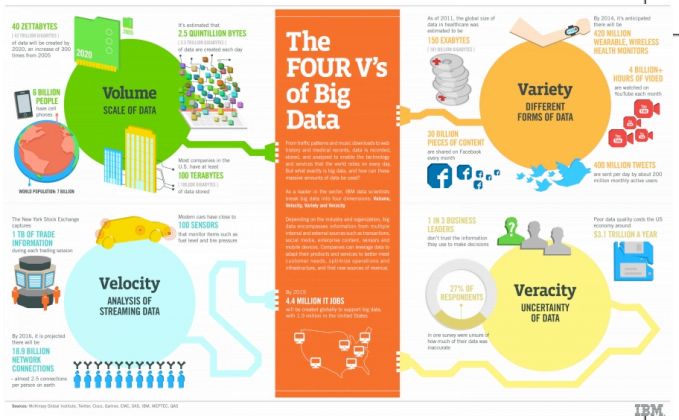
From the iDigBio MISC-Authority-File-Working-Group

AlternativePersonGUID
AuthorityListGUID
AuthorOfTaxonGUID
CentroidGeoreferencedByGUID
CollectingEventAttributeID
CollectingEventGUID
CollectingProtocolReferenceGUID
CollectingTripID
CollectionAlternativeGUID
CollectionCuratorGUID
CollectionDirectorGUID
CollectionGUID
CollectionManagerGUID
CollectionObjectCollectorGUID
CollectionObjectDeterminerGUID
CollectionObjectGUID
CollectionTechnitianGUID
CollectorGUID
CurrentGeographyGUID
CurrentNameGUID
DisciplineID
EventGeographyGUID
ExpertInTaxonGUID
GeographyGUID
GeologyGUID
GeoreferenceProtocolGUID
GeoreferencesGeoreferencedByGUID
HigherCollectionGUID
HigherGeographyGUID
HigherGeologyGUID
HigherTaxonGUID
HostForColletionObjectGUID
HostForTaxonGUID
HostsOnCollectionObjectGUID
HostsOnTaxonGUID
HybridParent1GUID
HybridParent2GUID
iDigBioProviderGUID
iDigBioPublisherGUID
IGSN
InstitutionGUID
MediaGUID
NamePublishedInGUID
PersonGUID
ProjectGUID
ProviderCreatedByGUID
ProviderCreatedByGUID
ProviderModifiedByGUID
ProviderModifiedByGUID
SpecimensGUID
SynonymTaxonGUID
TaxonNameGUID

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?

The 3 or 4 Vs of Big Data



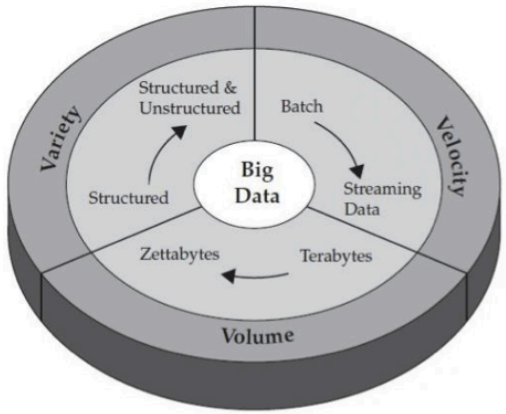
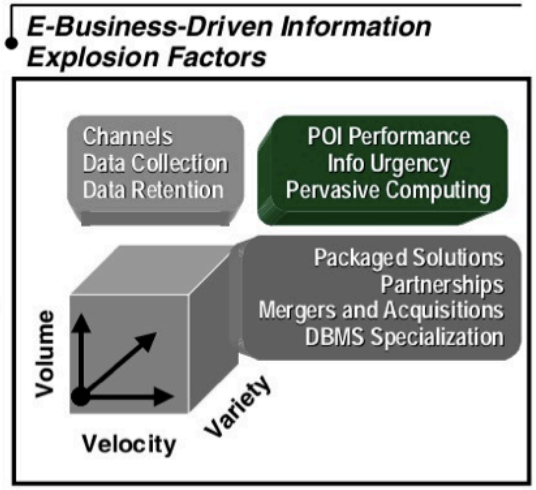


Figure 1-1 *IBM characterizes Big Data by its volume, velocity, and variety—or simply, V³.*

Figure 1 — Data Management Solutions

- **Volume**
  - Tiered storage/hub and spoke
  - Selective data retention
  - Statistical sampling
  - Redundancy elimination
  - Offload "cold" data
  - Outsourcing
- **Velocity**
  - Operational data stores
  - Data caches
  - Point-to-point data routing
  - Balance data latency with decision cycles
- **Variety**
  - Inconsistency resolution
  - XML-based "universal" translation
  - Application-aware EAI adapters
  - Data access middleware and ETLM
  - Distributed query management
  - Metadata management



Extending data management options enables greater returns on information assets

Source: META Group

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

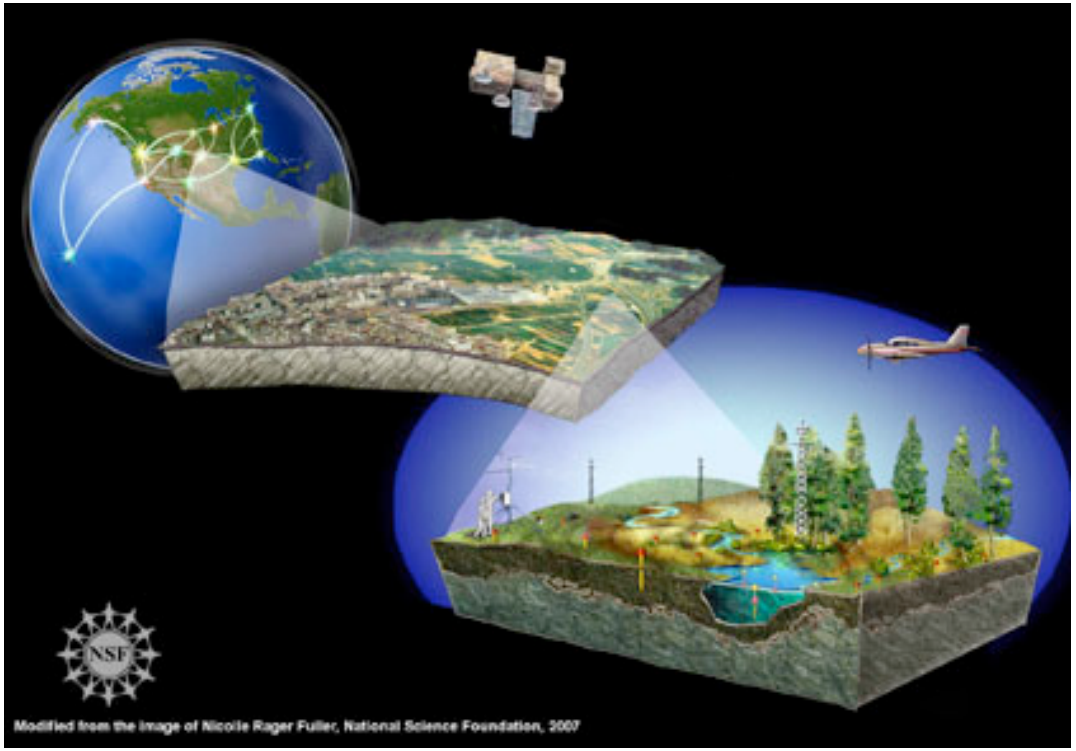
4Vs for Biodiversity Big data

- 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?

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



- The expeditions of Cook, Darwin, Wallace, Beccari, and others were the moonshots of the 18th and 19th centuries.



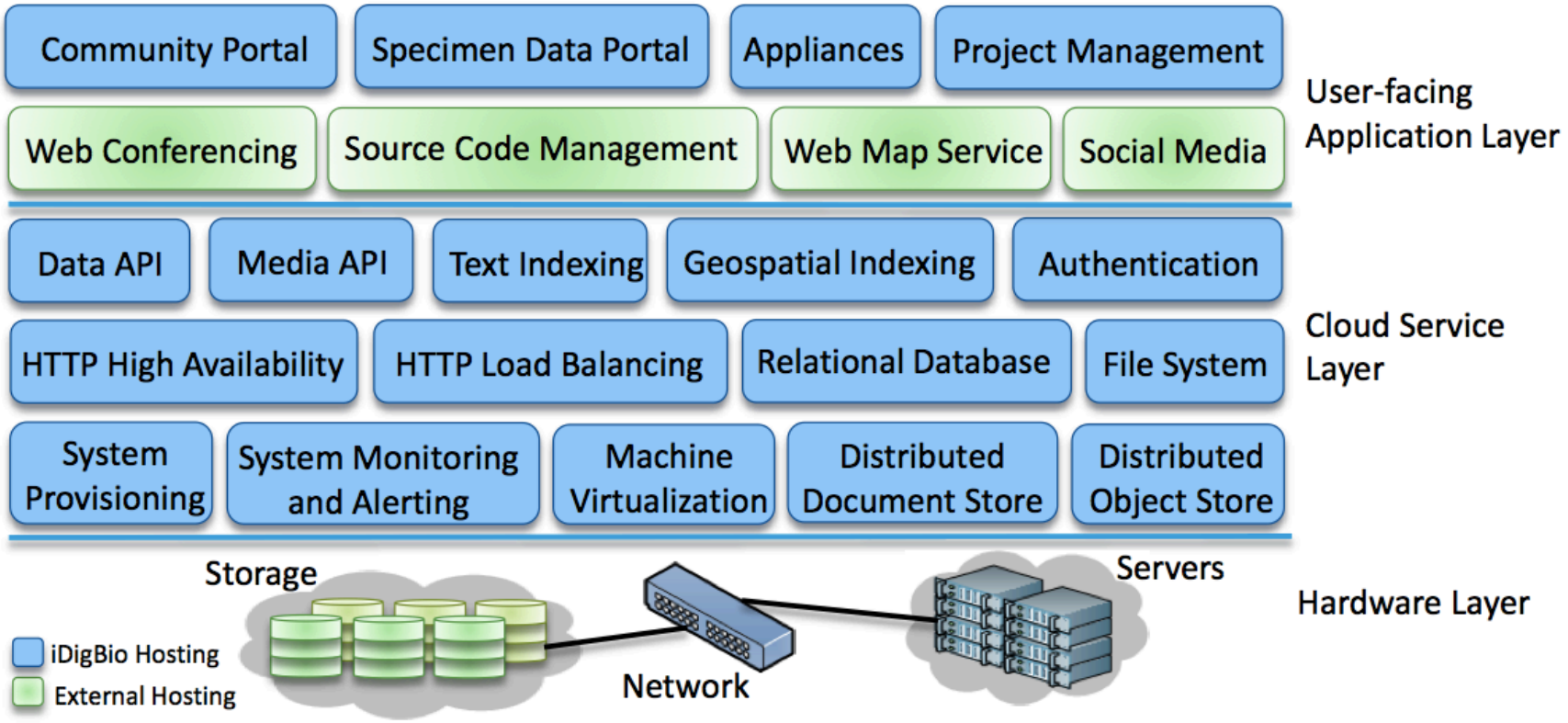

- Taxonomic classification dating to Linnaeus, and more recently phylogenetic systematics are a treasure trove of data, which we can now leverage in a big data paradigm.
- As is local, indigenous knowledge of natural systems going even further back in time.

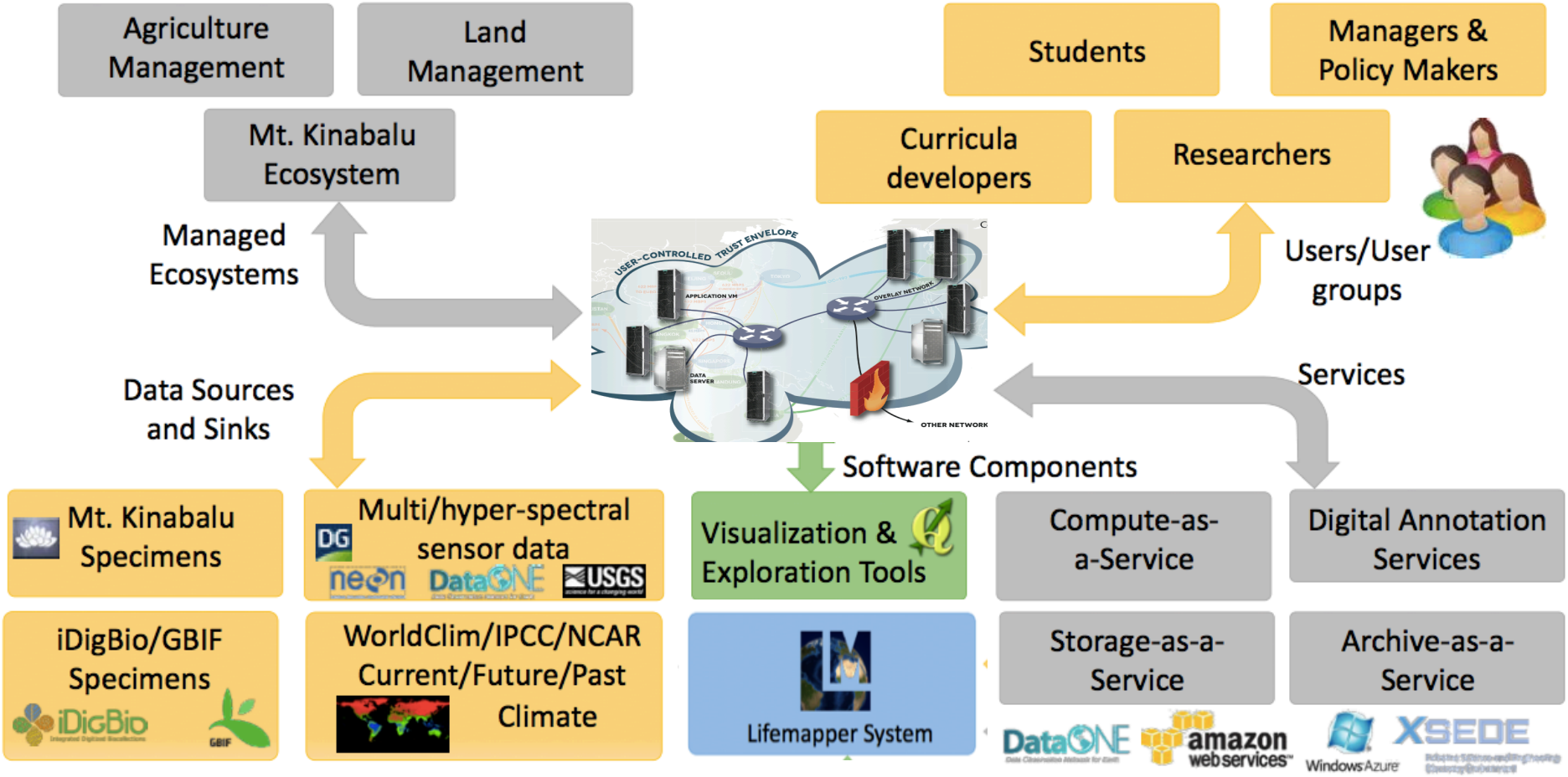
4Cs of Biodiversity Big Data

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

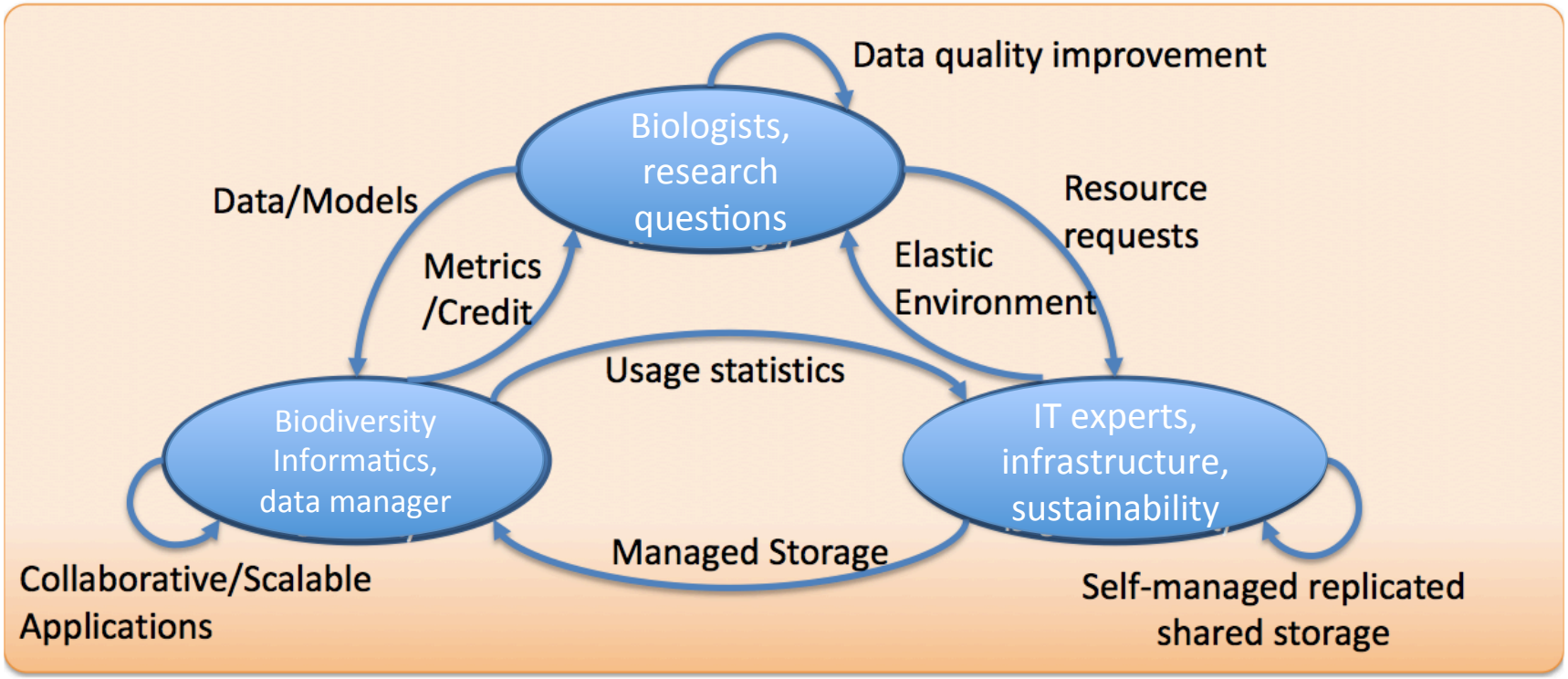
iDigBio Cyberinfrastructure

- Developed in consultation with stakeholders
- Implementation determined internally
- Feedback on prototypes solicited from community




Agriculture Management
Land Management
Students
Managers & Policy Makers
Mt. Kinabalu Ecosystem
Curricula developers
Researchers
Managed Ecosystems
Users/User groups
USER-CONTROLLED TRUST ENVELOPE
APPLICATION VM
OVERLAY NETWORK
DATA SERVER
OTHER NETWORK
Data Sources and Sinks
Services
Software Components
Mt. Kinabalu Specimens
Multi/hyper-spectral sensor data
Visualization & Exploration Tools
Compute-as-a-Service
Digital Annotation Services
iDigBio/GBIF Specimens
WorldClim/IPCC/NCAR Current/Future/Past Climate
Lifemapper System
Storage-as-a-Service
Archive-as-a-Service
DataONE
amazon web services
Windows Azure
XSEDE

Socio-technical challenges and feedback



Modified from Fortes et al.