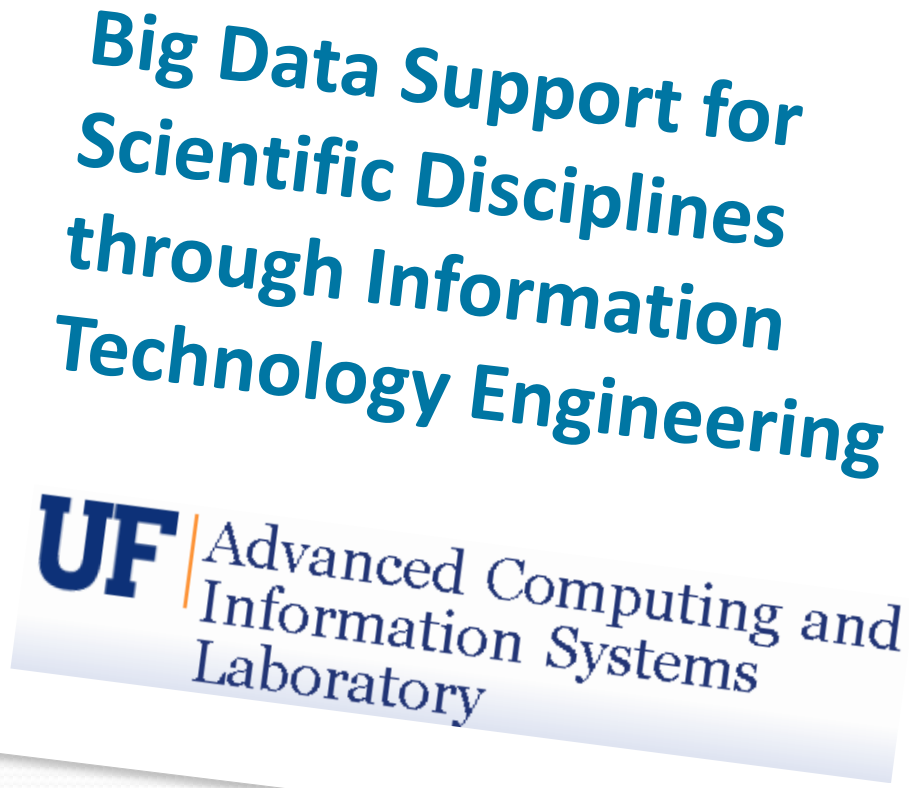
Big Data Support for Scientific Disciplines through Information Technology Engineering

Andréa Matsunaga
Maurício Tsugawa

ACIS Laboratory
Department of Electrical
and Computer Engineering
University of Florida



UF Workshop on Dense, Intense, and Complex Data

209 Emerson Alumni Hall
June 19th, 2013

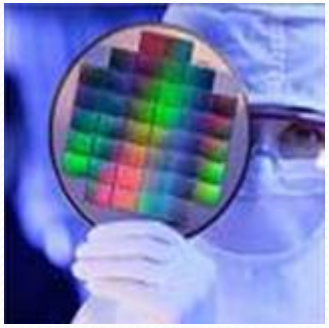
Supporting Scientific Disciplines

Biodiversity Research

iDigBio

PRAGMA

Semiconductor Manufacturing



Samsung

Computer Science/Engineering

FutureGrid

Disaster Mitigation and Recovery

RAPID-Japan

RAPID-Thai

Bioinformatics Research

Health Care Research

REPAIR

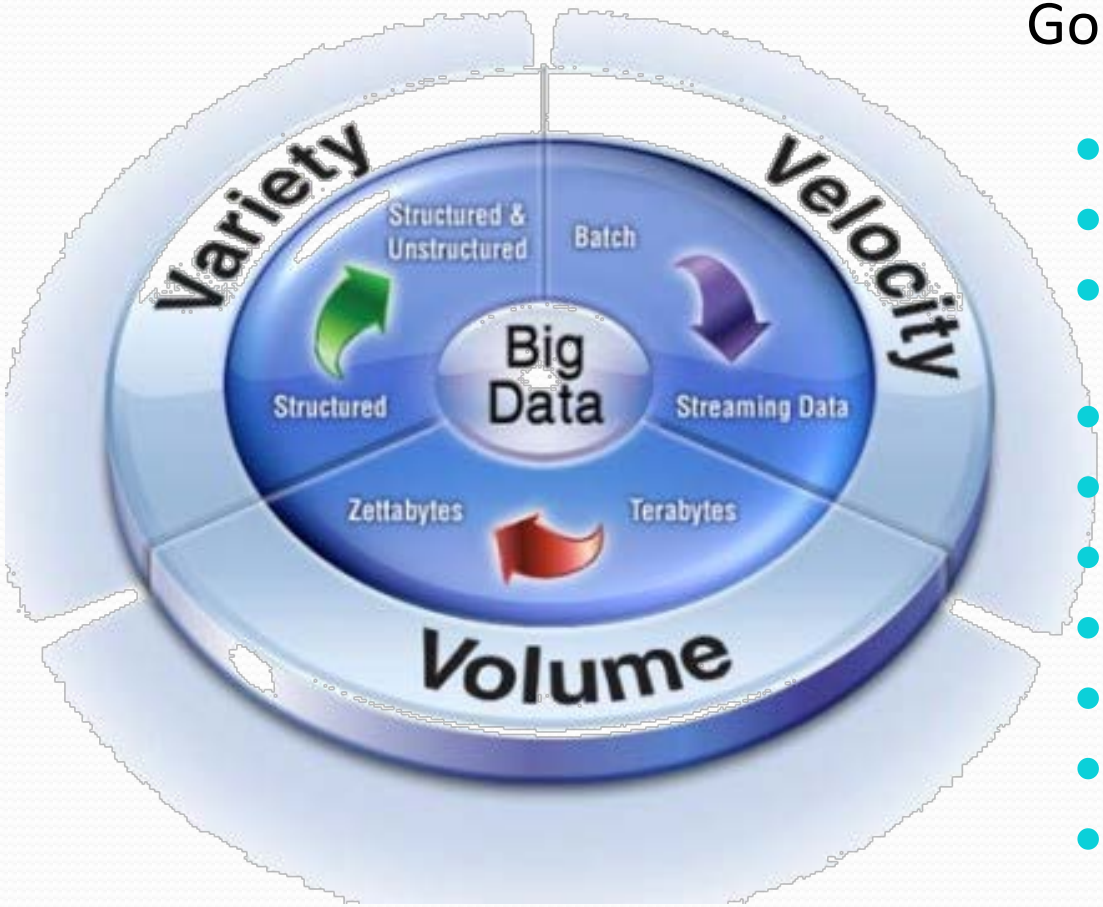
AIR

Area of Research:

Computer Science/Engineering

Conduct fundamental and applied research on systems that integrate computing and information processing

What is Big Data?



Goal: Extract Value

- Scalable
- Resilient to failures
- Single general purpose big data infrastructure
- Autonomous
- Standardized interfaces
- Secure
- Ad-hoc analytics/visualization
- Traceability
- Long-term
- Low cost

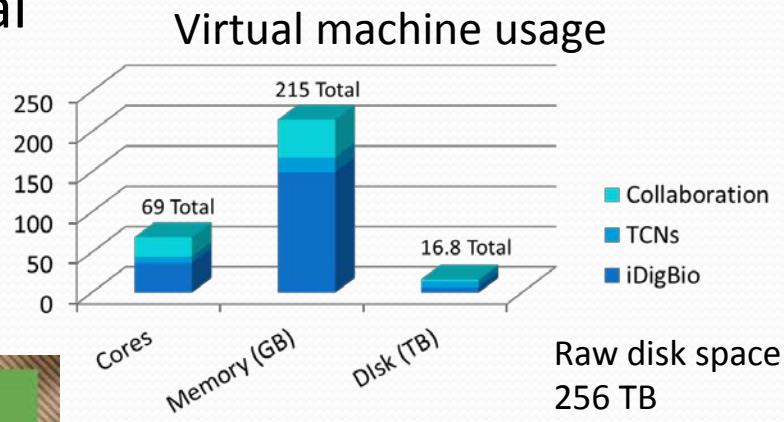
Qualifier: Traditional approaches cannot handle

3-D Data Management: Controlling Data Volume, Velocity and Variety. (Doug Laney, Gartner, 2001)

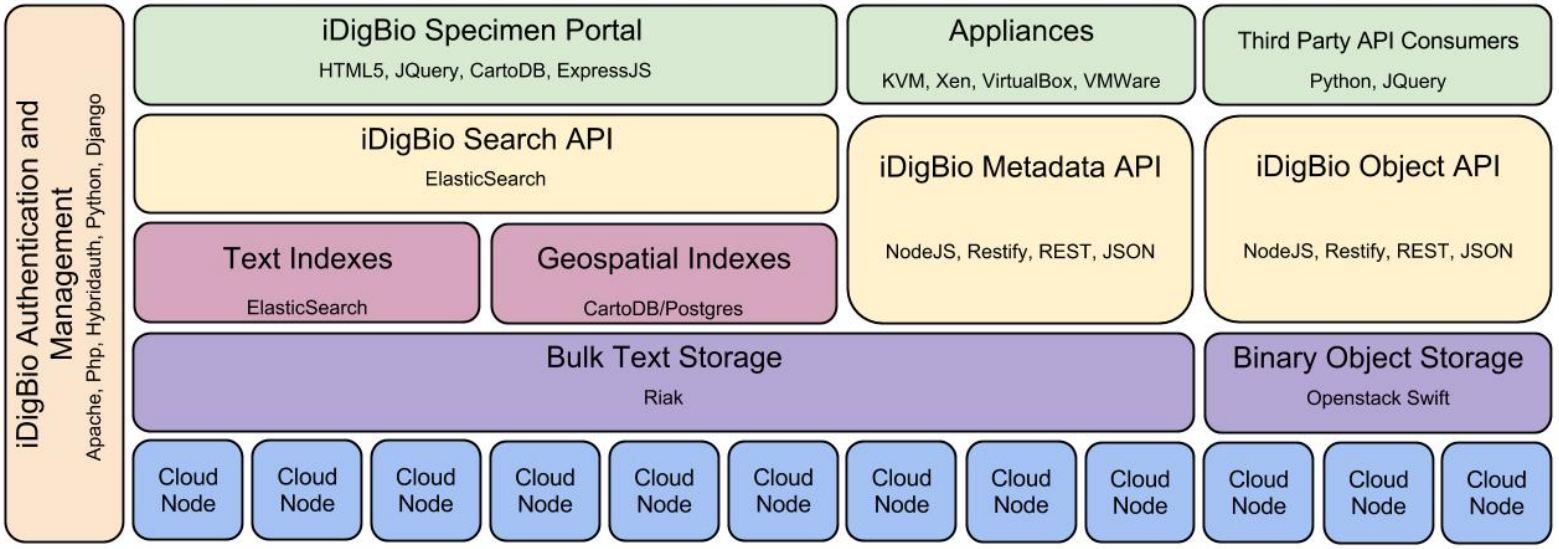
iDigBio

iDigBio
Integrated Digitized Biocollections
http://idigbio.org

- Make available millions of biological specimens in digital format: text and media (images, vocalizations, videos, 3D-models, etc.)



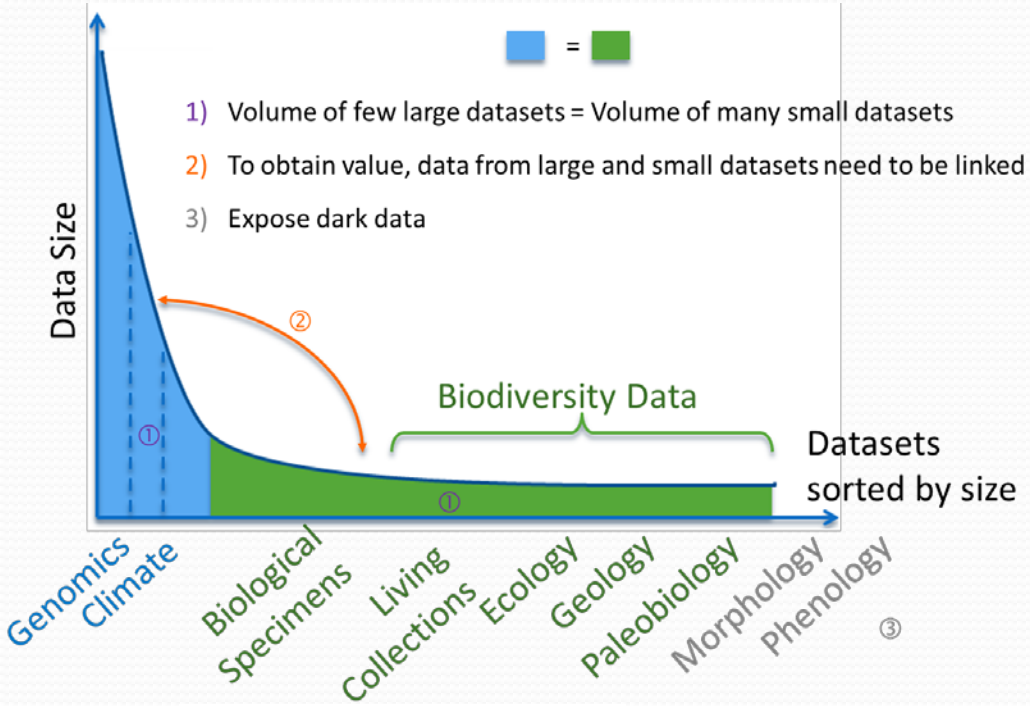
Raw disk space: 256 TB

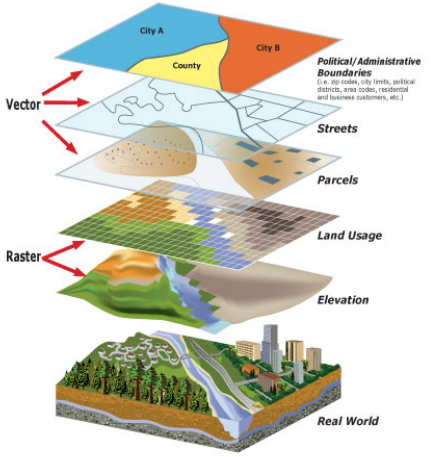


iDigBio

- Need to deal with a variety of sources:
  - Structured, semi-structured, unstructured, sensors, objects
- Linking data (cross-domain research)
  - Within iDigBio data concepts
  - Across iDigBio and other biodiversity data
    - e.g., genetic material, scientific publications, mapping information and ecological information

Potential for long-tail big data science

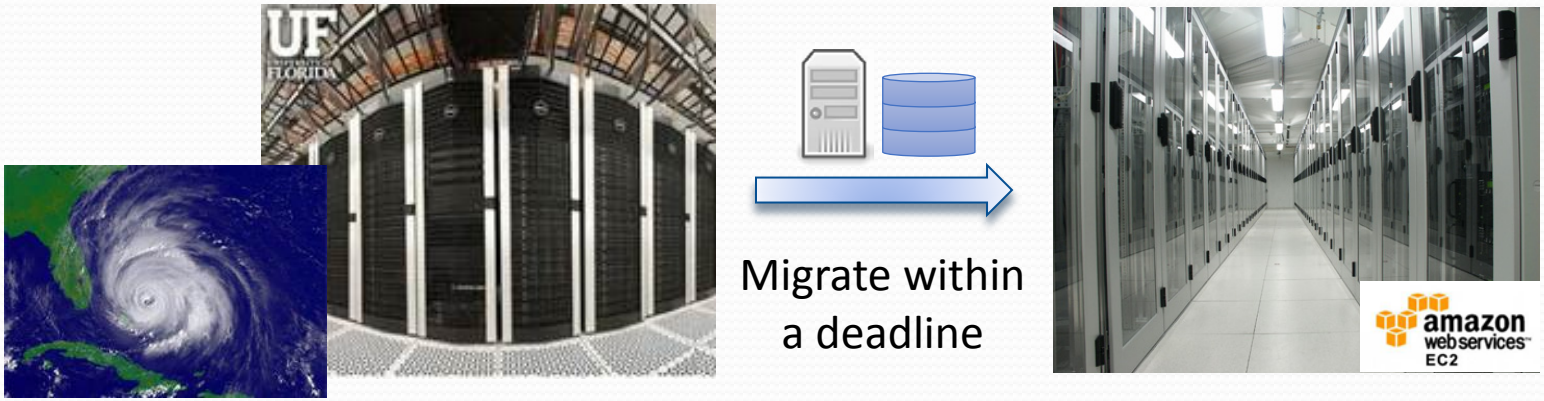




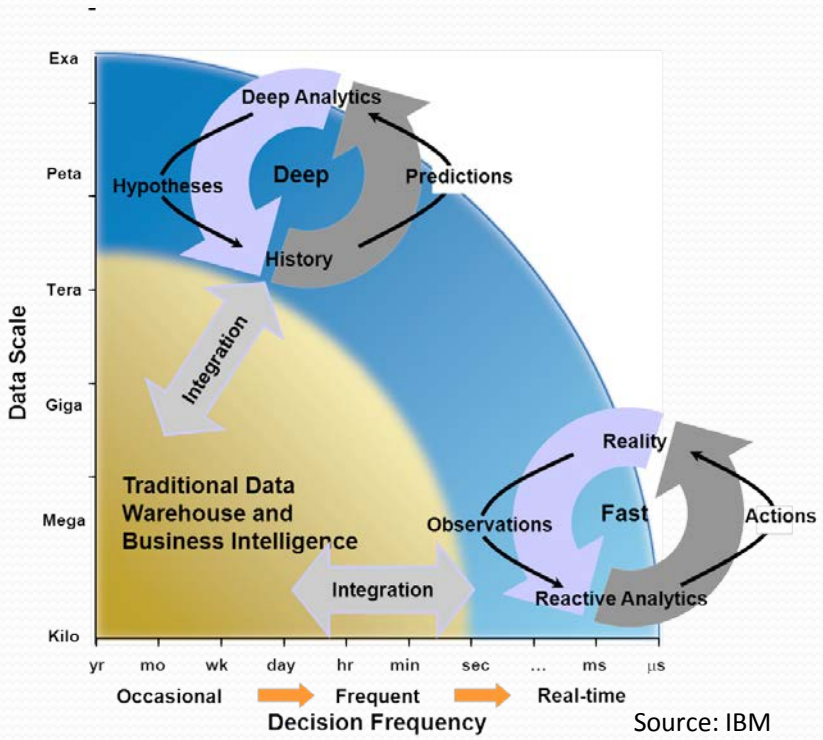
Big data science questions:

How did species evolve?
How should land be managed to avoid loss of biodiversity?
How do invasive species spread?

Disaster Mitigation and Recovery



- Feedback control mechanism to maximize system evacuation throughput while minimizing per-VM migration time and adapting to changing network conditions and VM activity



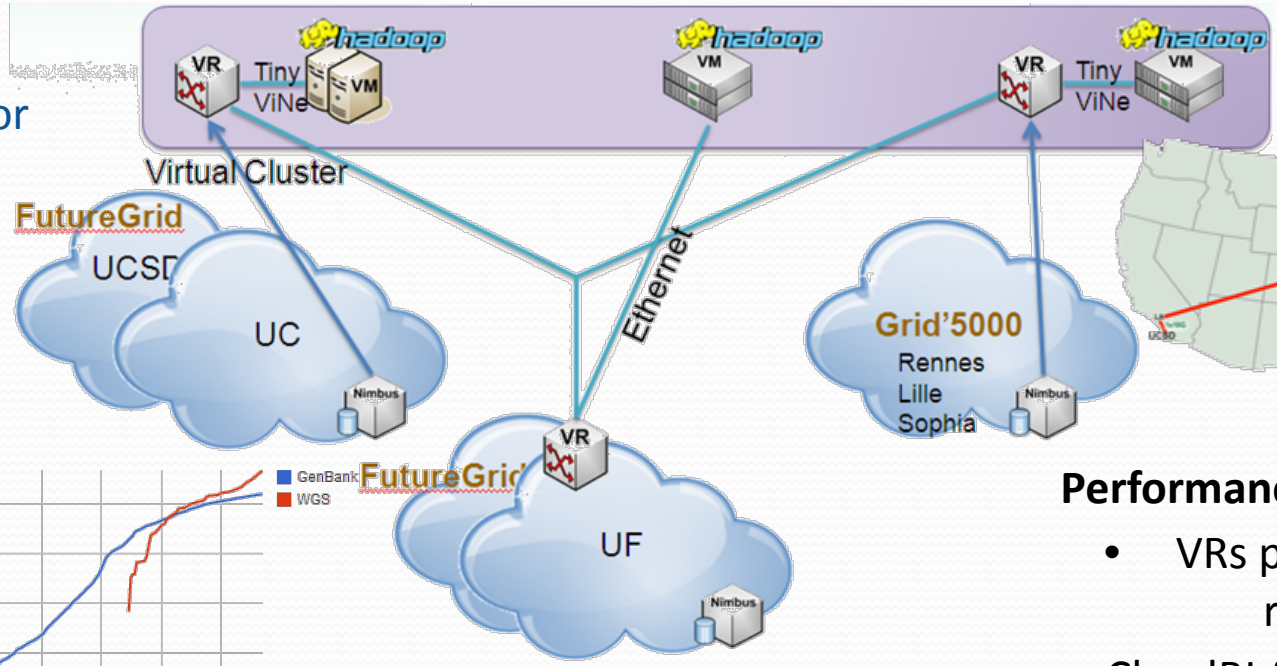
Source: IBM

Potential for real-time big data

FutureGrid – Intercloud Research

- **Offers**: Geographically distributed testbed for computer scientists to conduct research and development of cloud technologies
- **Case Study**: Execution of CloudBLAST on 750 VMs (across 3 FG sites and 3 Grid'5000 sites), and 1500 cores
- **Objective**: Efficiently combine cloud technologies (ViNe, Nimbus, Hadoop, etc) to form an intercloud virtual cluster

Potential testbed for big data analysis



Performance and Connectivity

- VRs process IP packets at rates over 850 Mbps
- CloudBLAST speedup: ~870X
- Firewall traversal

Opportunities in Big Data

- Big data address challenges that are naturally inter-disciplinary
  - UF has expertise in many domains
  - Computer engineers need to understand interests, differences and commonalities of big data challenges in different domains
  - Pilot close collaborations within UF
  - Minimize barriers for data sharing (technical and non-technical)

Thank you!