How Big is “Big Data” Across Disciplines: Workshop 1 Analysis and lessons learned

Gul Kremer, Program Officer
Division of Undergraduate Education

Elizabeth Burrows, AAAS Big Data Fellow
Division of Mathematical Sciences

June 1, 2015

Observations

- Efficiency in scientific discovery through curation, analyses and interpretation of massive datasets
- Uptake level and concentration on “Big Data” opportunities are varied across disciplines
  - The nature of the data needed within disciplinary communities
  - Characterization using Velocity, Variety, Veracity, Volume typology

Definition

“Big Data consists of extensive datasets primarily in the characteristics of volume, variety, velocity, and/or variability that require a scalable architecture for efficient storage, manipulation, and analysis.”

NIST

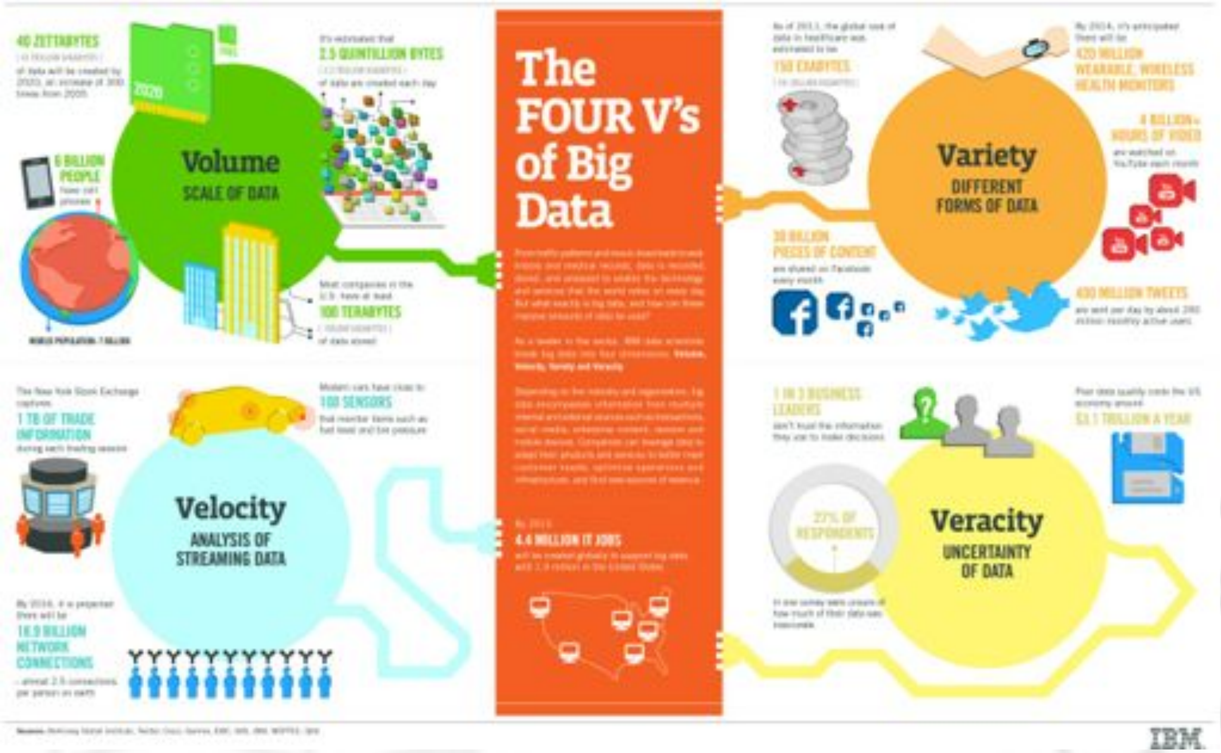
Workshop 1 Overview

- **5 case studies** of effective partnerships **outside of education** btw producers & consumers
  - Earth Science
  - Biology
  - Astronomy
  - Health
  - Computer Science
- Breakouts
- Methods & Analytics

Focus of Case Studies

For each case study we identified:

- What makes this "big data"?



From: http://www-1.ibm.com/software/data/bigdata/

Projects that exemplify the V's

Volume - LSST - Climate	Variety - S & C Health - Plant Genomics
Velocity - Reality Deck - SARS Outbreak - LSST	Veracity - Climate

How do these projects relate to yours?

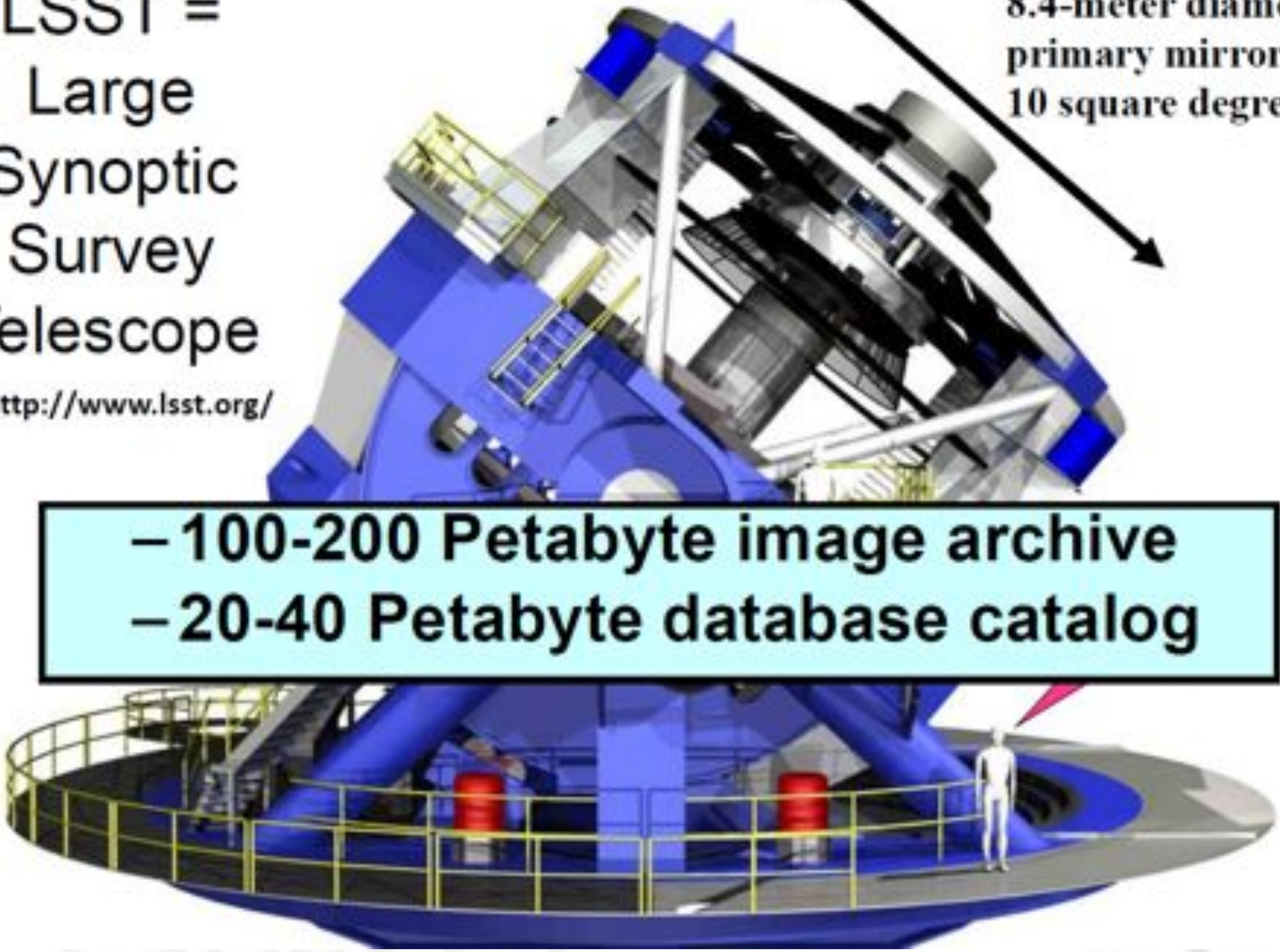
Volume (from Borne):

LSST =
Large
Synoptic
Survey
Telescope

http://www.lsst.org/

(mirror funded by private donors)

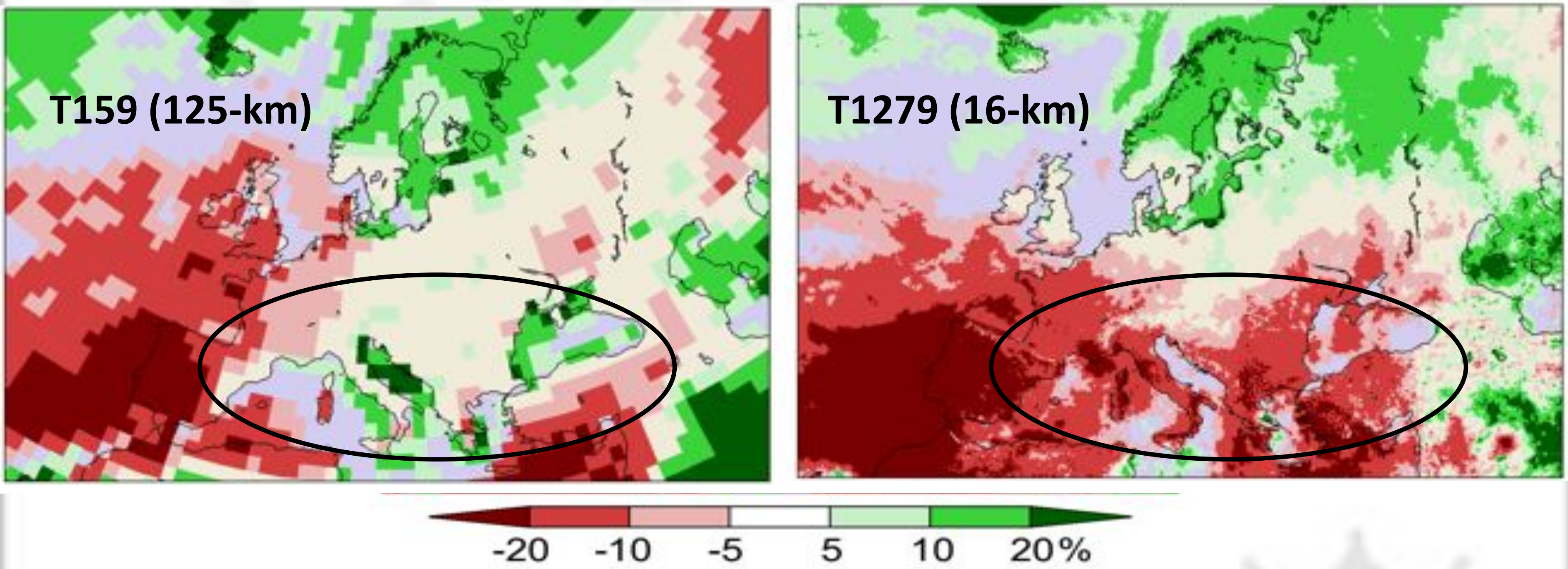
8.4-meter diameter
primary mirror =
10 square degrees!



– 100-200 Petabyte image archive
– 20-40 Petabyte database catalog

Volume (from Cash, Project Minerva):

Growing Season Precip. Change: 20th C - 21st C



- IBM iDataplex, 72,280 cores, **1.5 petaFLOPS peak** performance
- #17 on June 2013 Top500 list of supercomputing sites
- **10.7 PB disk capability**
- **10x increase in FLOPS, 100x increase in storage over Athena**

This much data breaks everything: H/W, systems management policies, networks, apps S/W, tools, and shared archive space

Velocity (from Borne):

LSST Key Science Drivers: Mapping the Dynamic Universe

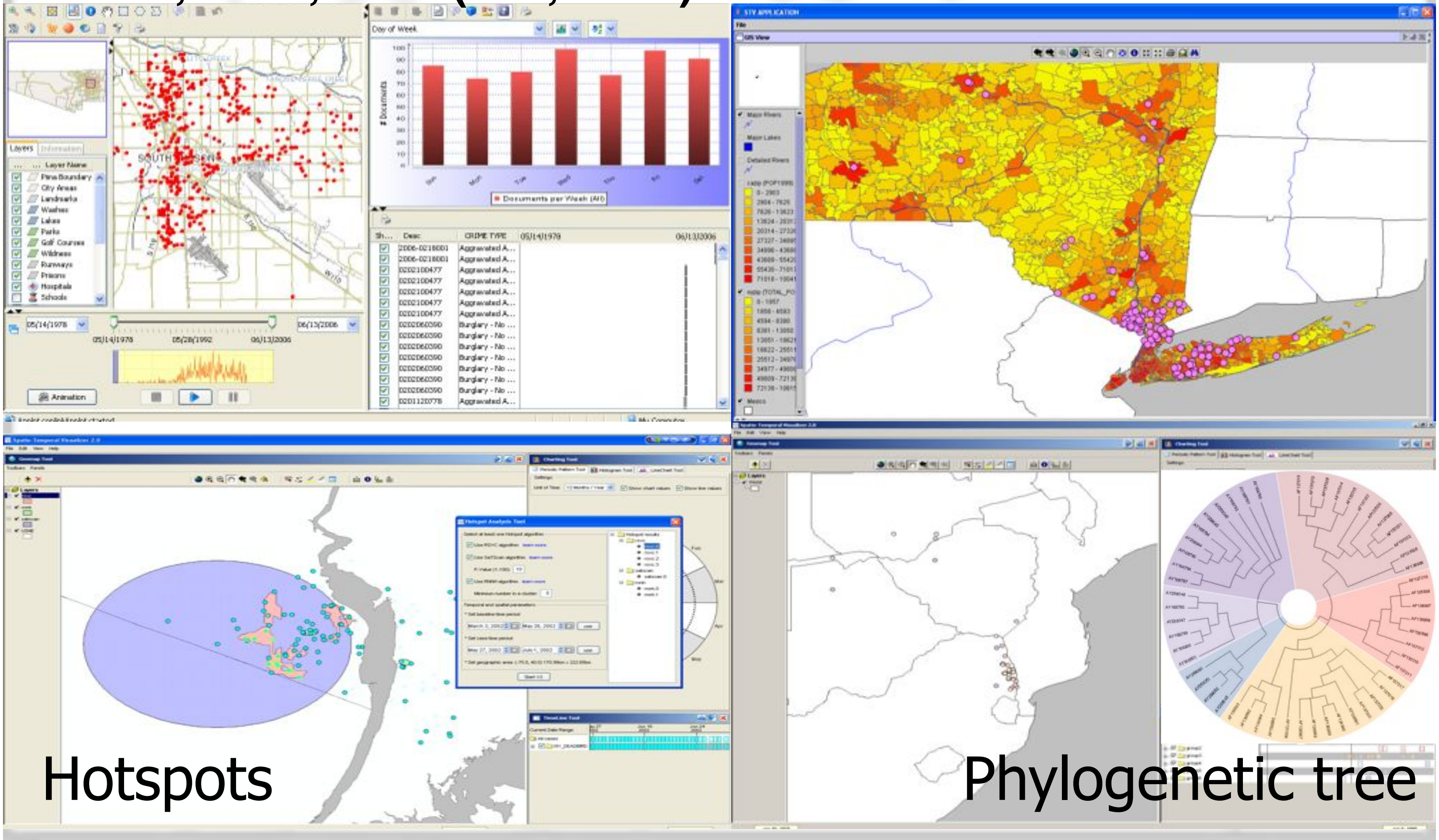
- Complete inventory of the Solar System in **REAL-TIME**
  - **EVENT MINING:** ~10 million events per night, every night, for 10 years!
    - – Follow-up observations required to classify these
    - – Which ones should we follow up? …

    … Decisions! Decisions! Data-to-Decisions!
  - Repeat images of the entire night sky every 3 nights
  - One 6-GB image every 20 seconds
    - Near-Earth Objects; killer asteroids???
    - Exploding supernovae

Antofagasta
Copiapo
La Serena
Argen
Valparaiso
Santiago
Rancagua
South America
Chile
Region de Coquimbo
Summit of Cerro Pachon - The Gemini South Observatory

Velocity (from Chen):

BioPortal: Infectious Disease Tracking and Visualization, SARS, WNV, FMD (ISR, 2009)



Sh...	Desc	CRIME TYPE	05/14/1978	06/13/2006
☑	2006-0218001	Aggravated A...		
☑	2006-0218001	Aggravated A...		
☑	0202100477	Aggravated A...		
☑	0202100477	Aggravated A...		
☑	0202100477	Aggravated A...		
☑	0202100477	Aggravated A...		
☑	0202100477	Aggravated A...		
☑	0202060390	Burglary - No ...		
☑	0202060390	Burglary - No ...		
☑	0202060390	Burglary - No ...		
☑	0202060390	Burglary - No ...		
☑	0202060390	Burglary - No ...		
☑	0202060390	Burglary - No ...		
☑	0202060390	Burglary - No ...		
☑	0201120078	Aggravated A...		

Hotspots

Phylogenetic tree

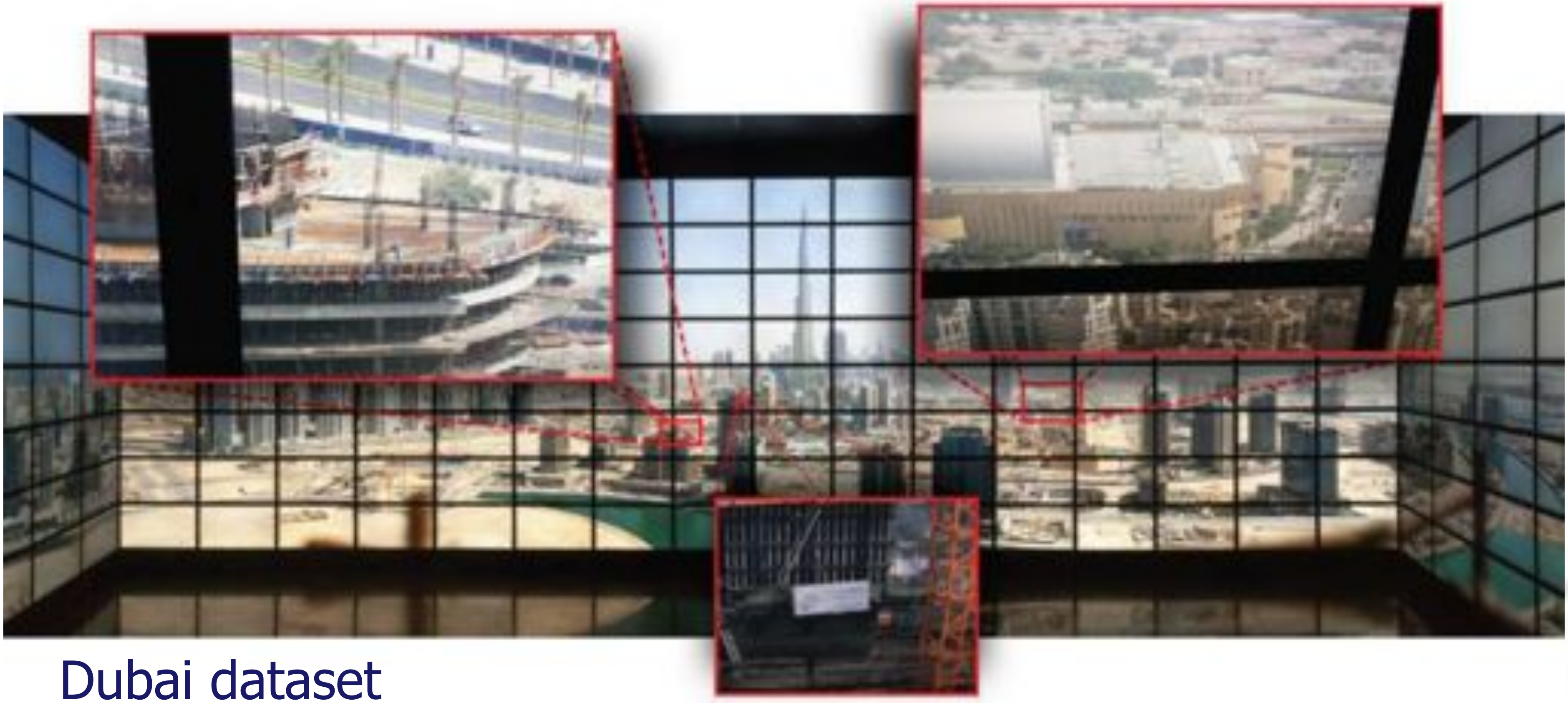
Velocity (from Kaufman):

The Reality Deck at Stony Brook University

Velocity (from Kaufman):

Reality Deck numbers:

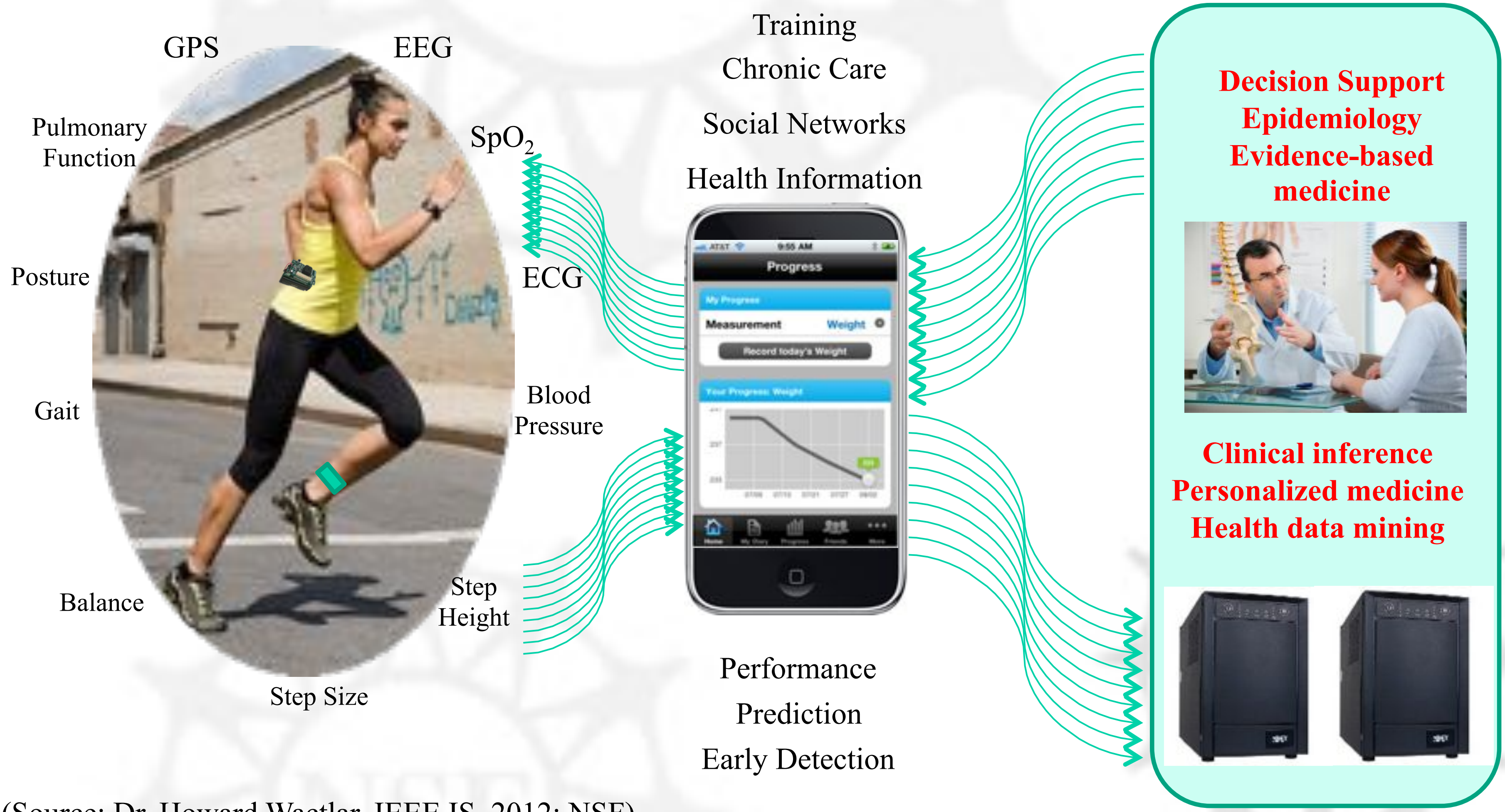
- 1.5 Gigapixels
- 240 CPU cores: 2.3 TFLOPS, 1.2 TB distributed memory



Dubai dataset

Variety (from Chen):

Smart & Connected Health: From Medicine to Health



GPS

EEG

Pulmonary Function

SpO_2

Posture

ECG

Gait

Blood Pressure

Balance

Step Height

Step Size

Training

Chronic Care

Social Networks

Health Information

Performance

Prediction

Early Detection

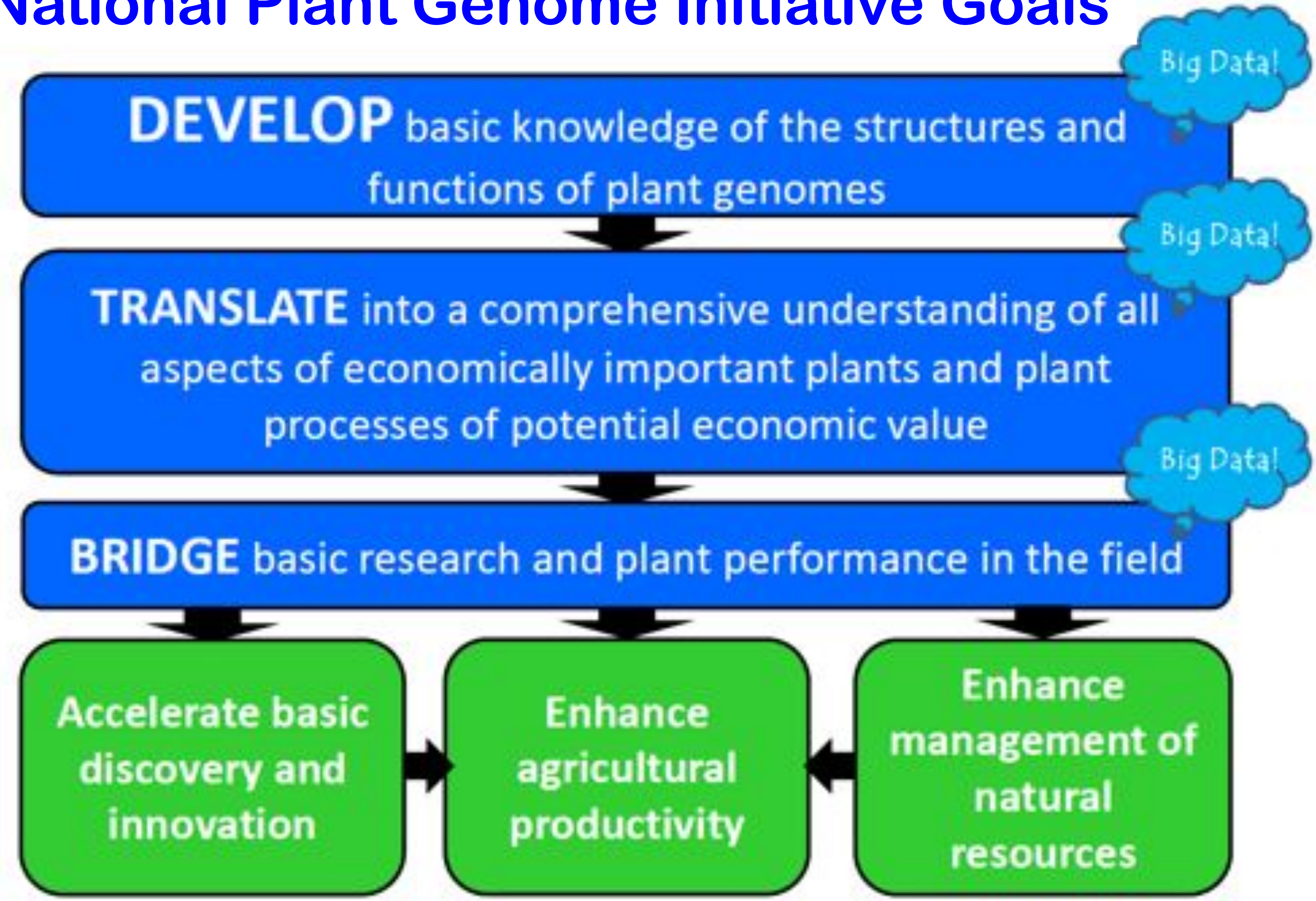
Decision Support
Epidemiology
Evidence-based medicine

Clinical inference
Personalized medicine
Health data mining

(Source: Dr. Howard Wactlar, IEEE IS, 2012; NSF)

Variety (from Okamuro):

National Plant Genome Initiative Goals



DEVELOP basic knowledge of the structures and functions of plant genomes

TRANSLATE into a comprehensive understanding of all aspects of economically important plants and plant processes of potential economic value

BRIDGE basic research and plant performance in the field

Accelerate basic discovery and innovation

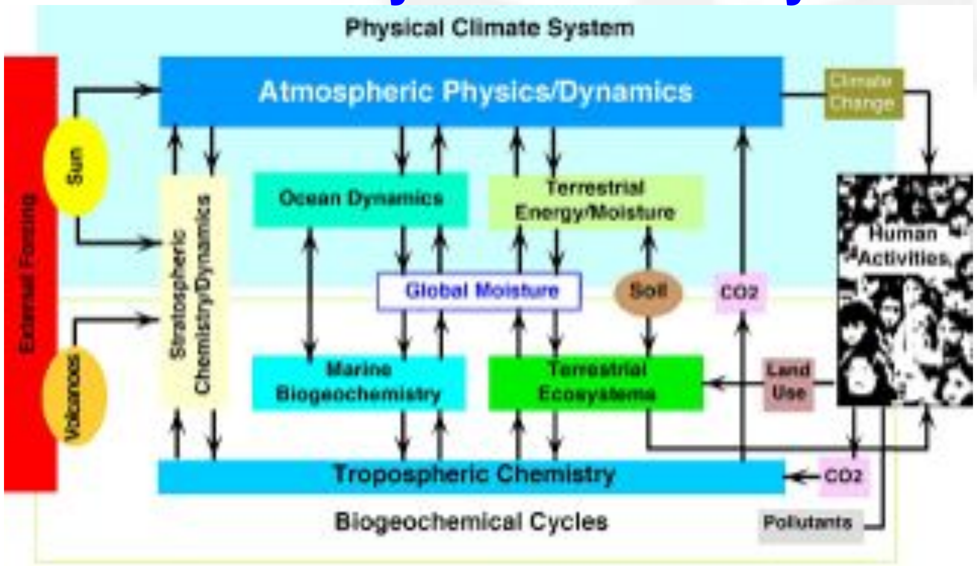
Enhance agricultural productivity

Enhance management of natural resources

Veracity (from Cash):

NMME The U.S. National Multi-Model Ensemble

Total Climate System – Earth System



(from Earth System Science: An Overview, NASA, 1988)

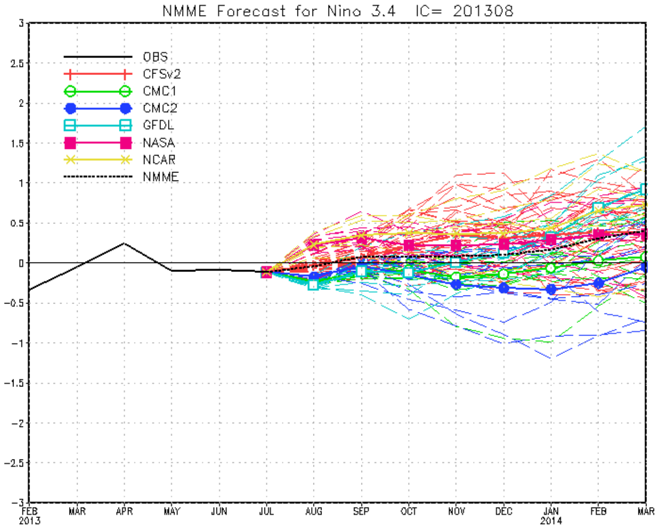


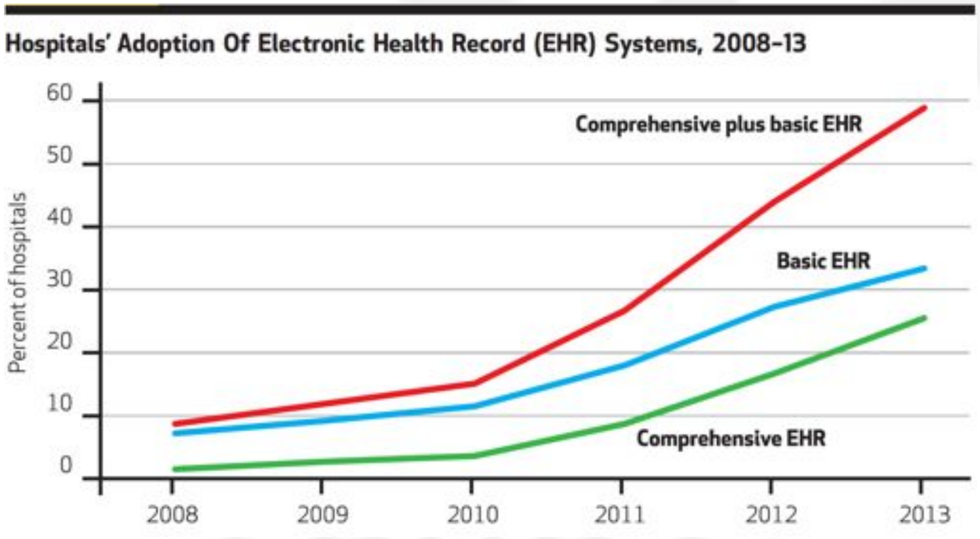
Fig. 12. Real-time Niño-3.4 predictions

Kirtman et al., 2014

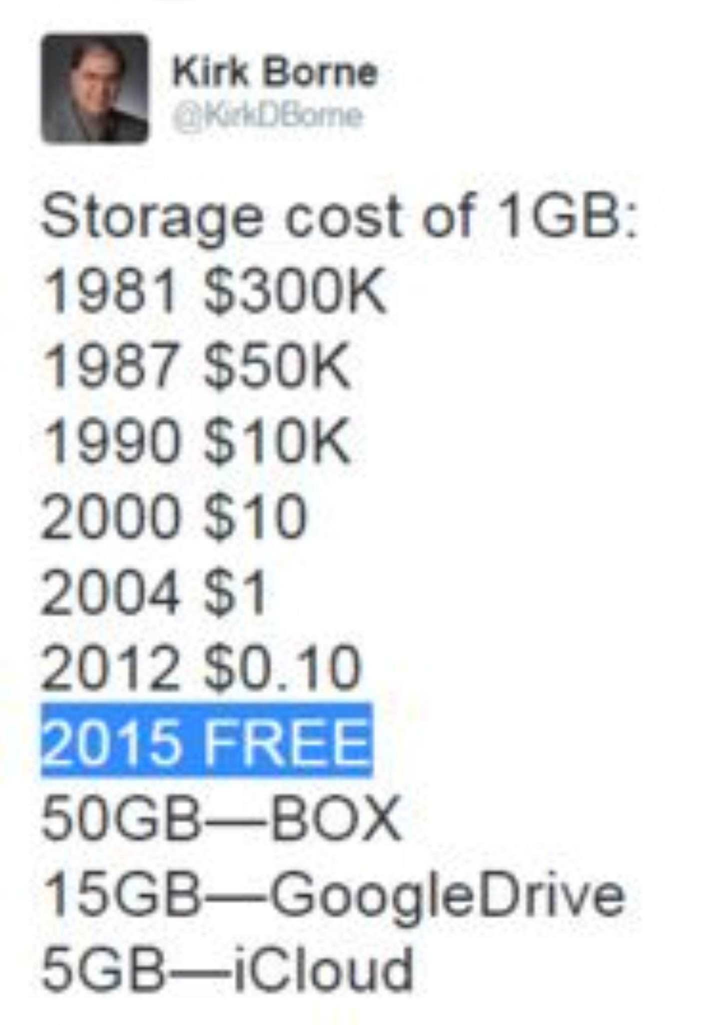
Funded by NOAA, NSF, NASA, & DOE

Some Lessons Learned

- Collaborate!
- Landscape rapidly Δing
- Get exaflood insurance ☺



Source: Adler-Milstein, J., DesRoches, C. M., Furukawa, M. F., Worzala, C., Charles, D., Kralovec, P., Stalley, S., and Jha, A. K. 2014. "More Than Half of US Hospitals Have At Least A Basic EHR, But Stage 2 Criteria Remain Challenging For Most," Health Affairs (33:9), pp. 1664–1671.



Kirk Borne
@KirkDBorne

Storage cost of 1GB:
1981 $300K
1987 $50K
1990 $10K
2000 $10
2004 $1
2012 $0.10
2015 FREE
50GB—BOX
15GB—GoogleDrive
5GB—iCloud

Thank you!

Questions?

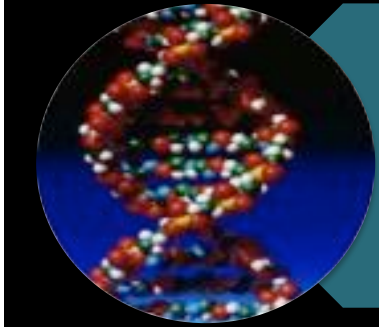
Discussion may continue at the Producer/ Consumer Relationships breakout tomorrow

Earth Sciences Lessons Learned

- Spatial resolution alone is not a panacea
- Validating high-resolution, high-complexity data pushes and in some cases exceeds observational capabilities
- Data from simulations can inform the veracity of the observational data.

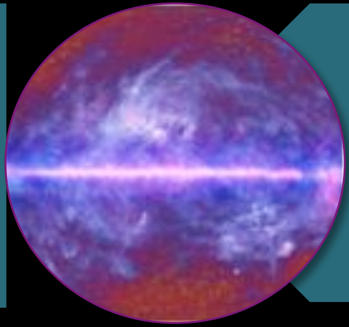
“Big Data” Challenges in Science

Volume, velocity, variety, and veracity



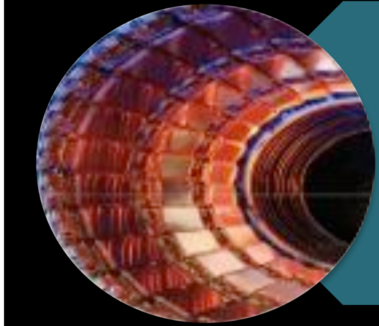
Biology

- *Volume:* Petabytes now; computation-limited
- *Variety*: multi-modal analysis on bioimages



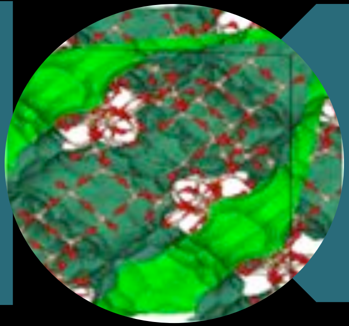
Cosmology & Astronomy:

- *Volume:* 1000x increase every 15 years
- *Variety:* combine data sources for accuracy



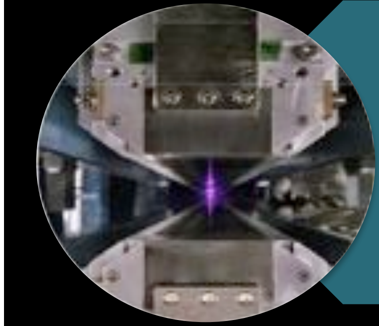
High Energy Physics

- *Volume*: 3-5x in 5 years
- *Velocity*: real-time filtering adapts to intended observation



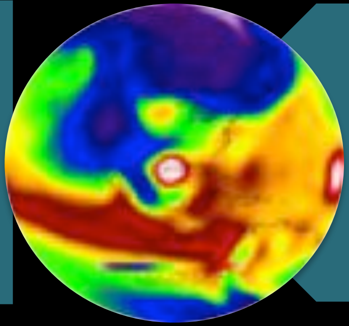
Materials:

- *Variety:* multiple models and experimental data
- *Veracity:* quality and resolution of simulations



Light Sources

- *Velocity:* CCDs outpacing Moore’s Law
- *Veracity:* noisy data for 3D reconstruction

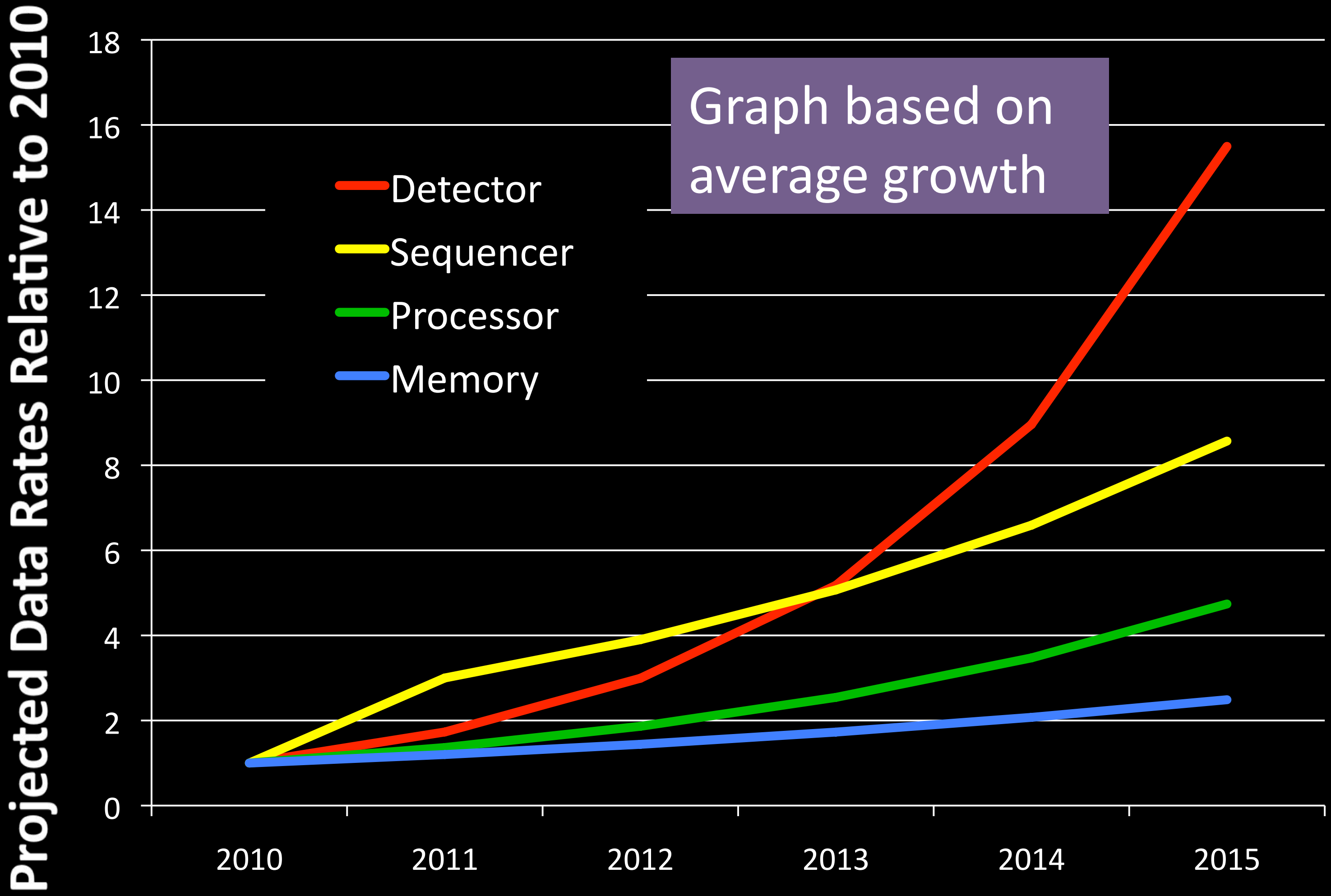


Climate

- *Volume:* Hundreds of exabytes by 2020
- *Veracity:* Reanalysis of 100-year-old sparse data

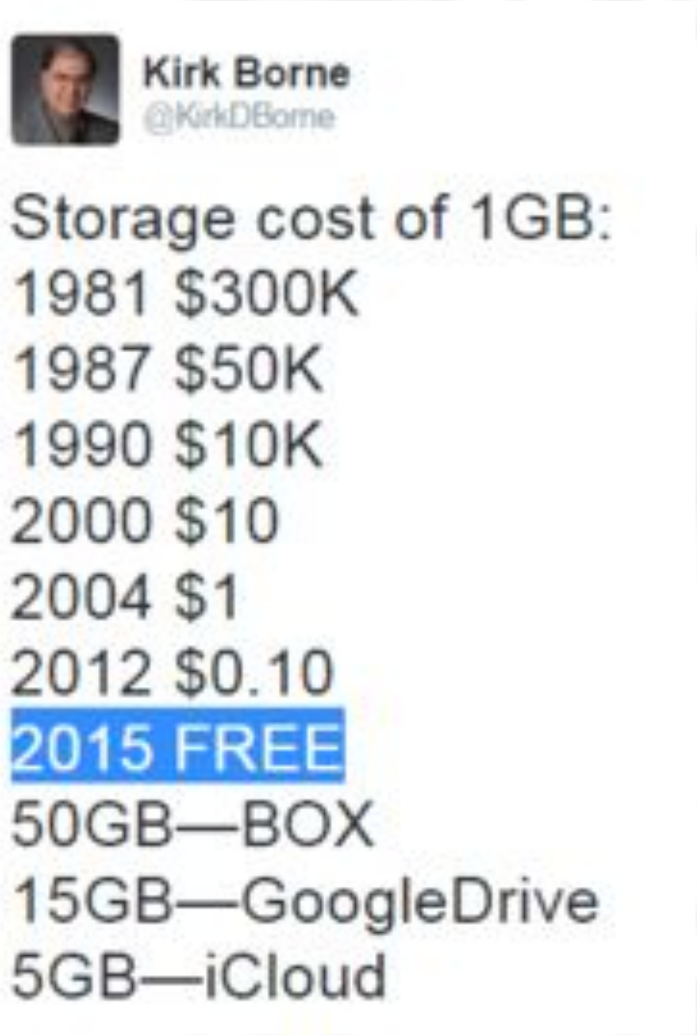
Source: Kathy Yelick

Data Growth is Outpacing Computing Growth



Graph based on average growth

Source: Kathy Yelick



Kirk Borne
@KirkDBorne

Storage cost of 1GB:
1981 $300K
1987 $50K
1990 $10K
2000 $10
2004 $1
2012 $0.10
2015 FREE
50GB—BOX
15GB—GoogleDrive
5GB—iCloud

2006	*CMIP3 (IPCC AR4)*	36 TB
2010	*Project Athena*	1.2 **PB**
2011	*CMIP5 (IPCC AR5)*	3 **PB**
2014	*Project Minerva*	3+ **PB**
2011-	*NMME*	1 **PB**
2015-	COLA storage	1 **PB**

Participants: ~50, 13 non-NSF

Workshop Goals

- Build capacity at NSF for using big data in education
- Articulate the conditions for success for effective usage of big data
- Study models of effective partnerships between sources of big data and its consumers
- Publish a volume that describes insights from the workshops

Focus of Case Studies

- Types of data wanted
- What makes this “big data”?
- Infrastructure, funding, policies needed
- Issues of data standards & interoperability
- Issues of privacy, security, & ethics
- Methods used by producers & consumers
- Issues of limited capacity
- Partnerships involved
- How has big data changed your field?
- Advice for others

Biology

- Doreen Ware, USDA & Cold Spring Harbor
  - Biology has become an information science
  - Genome sequencing – all big data issues apply
  - MAKER-P and Gramene project

- Diane Okamuro, Plant Genome Research Program
  - National Plant Genome Initiative – 17 yrs
    - 2014-2018 goals focused on open access and enhancing usability, big data
    - Companies are now ready to collaborate

Astronomy

- Lucy Fortson, University of Minnesota
  - Zooniverse
    - Citizen science – galaxy classification
    - Not producer/consumer – seamless knowledge discovery system

- Kirk Borne, George Mason University
  - Data Literacy For All: Astrophysics and Beyond
  - Undergrad data science program since 2007
  - Creating and storing data as fast as capabilites
  - LSST: Large Synoptic Survey Telescope

Methods & Analytics

- Barry Sloane, EHR/DRL
  - Analysis of numerical data
- Piotr Mitros, edX
  - Machine learning, analysis of non-numerical data

Breakout sessions

- Infrastructure, sharing, & standards
  - Notetakers: Al, Jay, Brandi
- Privacy, security, & ethics
  - Notetaker: Renata
- Capacity & producer/ consumer relationships
  - Notetakers: Quincy & Lida

Follow-up Activities

- Slides will be posted
    - Document produced
- Fellows will be invited to discussions
- Larger, education-based workshop May 13-15
- Volume published