Big Data R&D: From Infrastructure to Applications

Dr. Chaitan Baru

Senior Advisor for Data Science,
Directorate for Computer & Information Science & Engineering

National Science Foundation

NSF's Perspective and Role

The NSF funds basic, curiosity driven research

To promote the progress of science;

to advance the national health, prosperity, and welfare;

to secure the national defense….

National Science Foundation
WHERE DISCOVERIES BEGIN

My Perspective and Role

- Recently joined NSF (September 1, 2014)
  - After 17+ years on the "other side" – as a researcher (PI) at San Diego Supercomputer Center, UC San Diego, on NSF, NIH, NASA, etc. grants
  - (a database guy at a supercomputer center)
- Senior Advisor for Data Science, CISE Directorate
  - A newly created position
  - Motivated by need to connect with applications ➔ collaborate and coordinate across boundaries
  - Link CISE with all other NSF Directorates, for Big Data
  - Work with other agencies to develop Data Science strategy
  - Involve industry
  - Communicate/interact with international activities

Big Data is well understood…

- To be a vague term
- "I wish we had a better term for it…but it's what we have"
- Evolving definition
- Is it Atad Gib ? (like the Mirror of Erised)
- InterData…? (from Wendy Wigen, NITRD)
  - If we knew what the Internet would beget, what would we have called it?
  - NSF did have a program called "DataNet"
- A tipping point

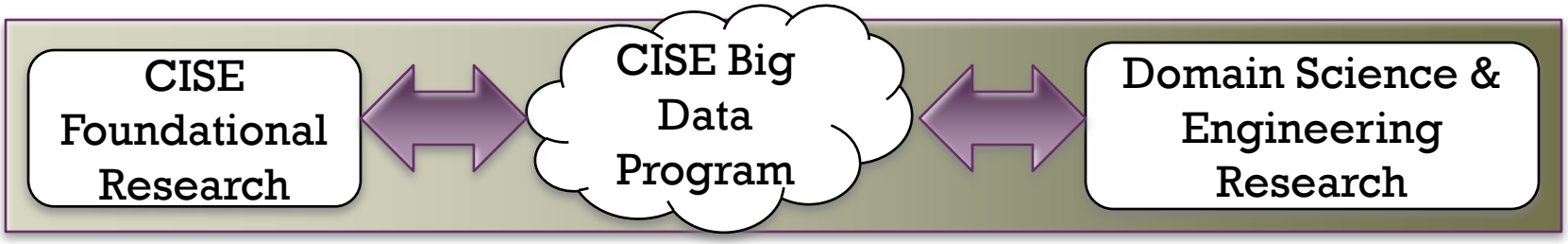
Big Data has created a greater awareness of *all* aspects of data

- It's not just about the cleaned up, structured data
- It's about
  - data through it's entire lifecycle
  - all types of data
  - as much about the metadata as the data
  - And…effective, timely use of data in end applications
- Has rejuvenated and created a new research agenda around data
  - It's great, if you are a database guy…!

NSF: Research to Infrastructure Continuum

- NSF has programs across the spectrum
  - From foundational research programs to science facilities and cyberinfrastructure
  - ACI – Advanced CyberInfrastructure is a Division in CISE
- And in education and training
- Big Data : Data Science :: Supercomputing : Computational Science

Big Data in the CISE Context



- "For us, Big Data is about 'Big'…", Dhruba Borthakur, Facebook
- Big Data is also about Data…

  → Data comes from the domains…

    → Big Data research needs to tie intimately to the domains

Examples of Big Data in NSF Domains

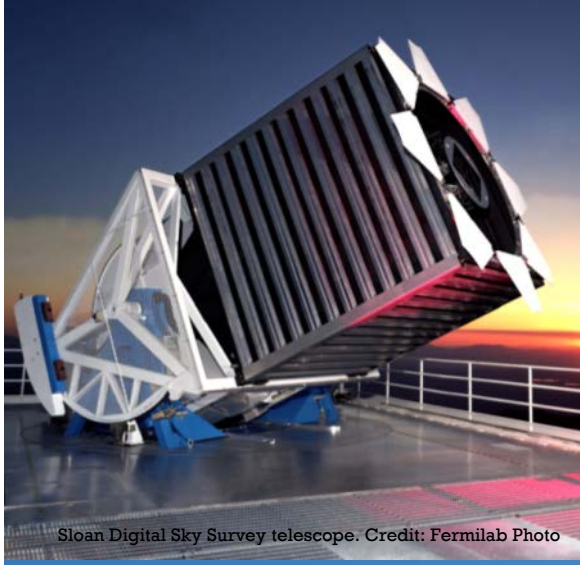
- **LIGO** – Laser Interferometer Gravitational-Wave Observatory .. MPS/Physics
  - 1PB/year
- **LHC** – Large Hadron Collider
  - 4PB/year
- **LSST** – Large Synoptic Survey Telescope … MPS/Astronomy
  - 100 PB in 10 years
  - 10+ PB catalog database
- **NCAR** – National Center for Atmospheric Research … GEO/Atmospheric Science
  - Multi-Petabytes of simulation data

Big Data in NSF Domains … Persistent data

- **EarthScope** … GEO/Earth Science
  - Seismic and Geodetic data archives at IRIS and UNAVCO
- **OOI** – Ocean Observing Initiative … GEO/Ocean Science
  - Just started. Data to be collected for 25 years
- **NEON** – National Ecological Observatory Network… BIO/Ecological Science
  - Data about to start. Collected for 30+ years.
- **MGI** – Materials Genome Initiative… MPS/Materials Science
  - Heterogeneous data collection. Novel initiative for this discipline.
- Most Are MREFC – New instrumentation projects ($100's M)
- Is Data the new instrument? – A major facility grant for Data?

LSST – Big Data in Astronomy

- The Sloan Digital Sky Survey in 2000, collected more data in its 1st few weeks than had been amassed in the entire history of astronomy
  - Within a decade, over 140 terabytes of information collected
- The Large Synoptic Survey Telescope due in Chile in 2016 will amass that quantity of data **in 10 days**

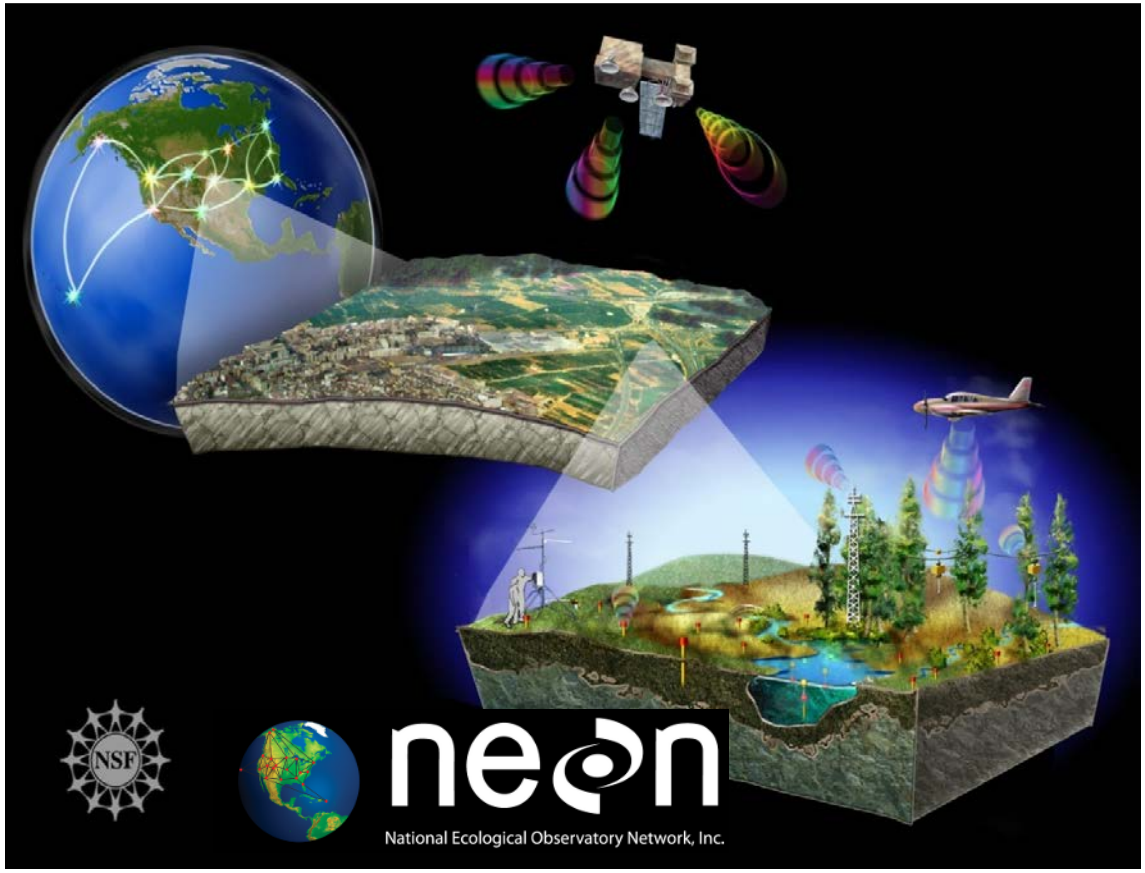

Sloan Digital Sky Survey telescope. Credit: Fermilab Photo


LSST


ALMA

National Ecological Observatory Network (NEON)

Transformational research platform and experimental facility to understand the biosphere and predict changes resulting from climate change, landuse change, and invasive species on regional to continental scales.



- Molecules to Biosphere
- Standardized multi-scale and experimental infrastructure
- Fixed, relocatable and mobile platforms
- Technology renewed throughout program lifetime
- Cutting-edge cyberinfrastructure and data products
- Open data for all
- Near real-time data access

To realize this vision, the observatory integrates fixed, flexible, and mobile sensing systems to support persistent sensing, synoptic campaigns, facilitate experiments across gradients of change, and assess episodic events.

NEON Headquarters: Control Center
CAL/VAL, Fabrication, Maintenance & Repair , and QA/QC Laboratories
Education/Outreach Portals/Tools

2 Biological Assessments: Field, Laboratory, BioArchive, Data Products

Collections of data on plants, animals, microbes (~ 3 TB data/yr)

3. Airborne Observatories (~200TB/year primary data)

1, Sensor/Instrument Packages: (~12,000 sensors generating ~30 TB data/yr)

» 60 Fundamental Instrument Units (tower, instrumentation hut, sensor nets)
- 20 Permanent – Continental Scale
- 40 Relocatable – Regional Scale

» 30 Stream Sensor Nets and 6 Lake Buoys

» 10 Experimental Stream Systems

» 10 Mobile Labs

National Ecological Observatory Network, Inc.

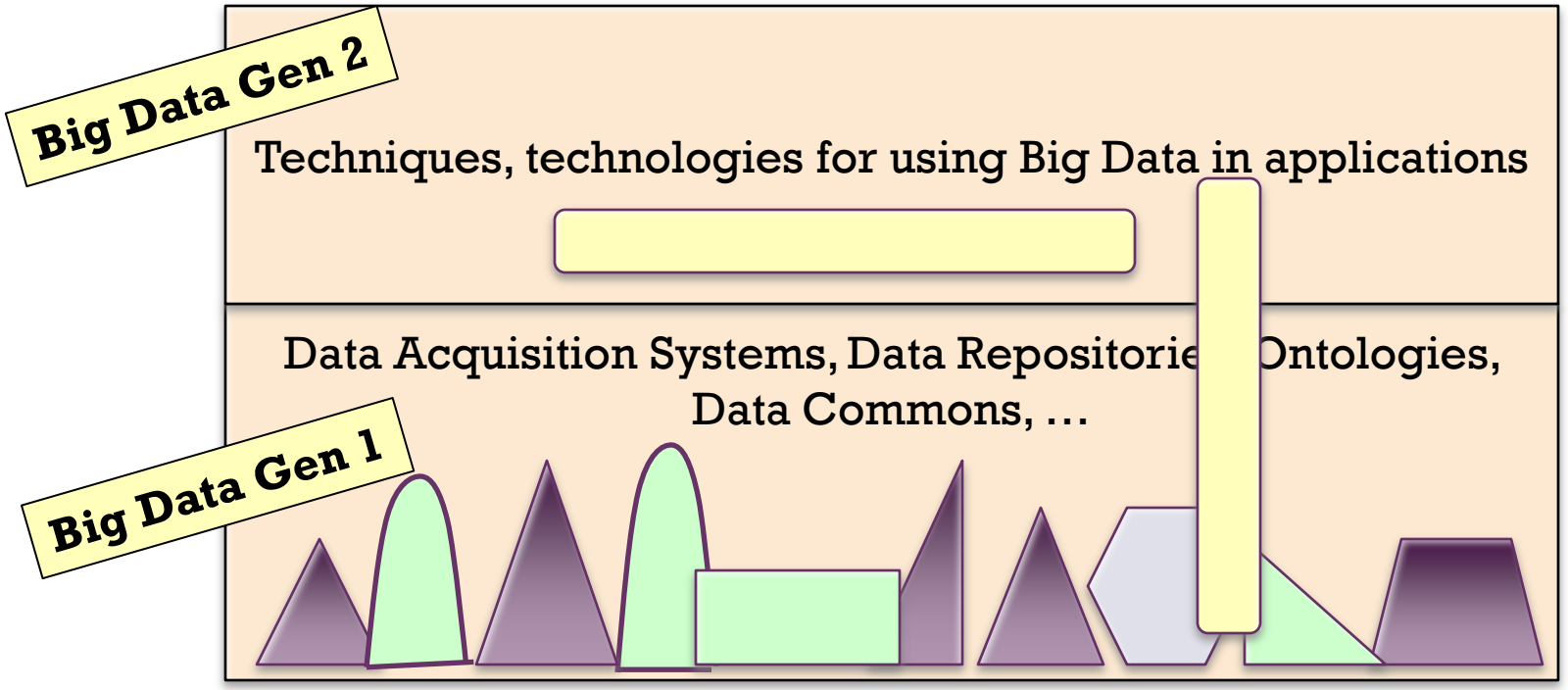
NEON Citizen Science

"Designed" vs "Found" Data

20,000 Citizen Scientists participating and growing

- **Project Budburst** is national field campaign to engage the public in the collection of important ecological data based on the timing of leafing, flowering, and fruiting of plants in relation to changes in season and climate.
- **Partnerships** - with botanic gardens, Wildlife Refuges and National Parks; National Geographic Society and others interested in contributing to a better understanding of plants and climate change.
- **Citizen Science Academy** – Professional accredited training, courses, modules, and tutorials designed to implement Citizen Science programs in educational settings
- **Provides** - key cyber technologies and applications.

From Infrastructure to Applications



Challenges

- The big payoffs are in multidisciplinary data "integration" and analysis
  - Multi-subdisciplinary;
  - Multi-disciplinary
- Requires breaking traditional silos
- Many universities are embarking down this path…
  - Joint hires
  - Incentives for multi-unit research

Challenges…Issues

- Privacy and ownership of data
  - Resolving the relationship between "data holders" and "data suppliers"
- Ethics and Data Science
  - Not just bioethics anymore
  - Big Data Ethics: Projects, articles, discussions underway
  - E.g. UK-US Data Science Summit identified ethics as the top issue
  - Opportunity for international exchange of ideas
- Value
  - Valuing Data / Big Data
  - When can you "delete" your data?

Initiatives underway …need your inputs

- RFI on Big Data Hubs
- RFI on National Big Data R&D Initiative
- Workshops focusing on new areas, e.g. Big Data and IoT (with NIST)
- Workshops with NIH (link CS researchers + Med School, Public Heath researchers)
  - Big Data Reproducibility; Data Science Training and Education
- Data Science Priority Goal for NSF
- Industry linkages are essential in all of the above

CISE RFI: Big Data Regional Innovation Hubs

Accelerating the Big Data Innovation Ecosystem

- Continue and scaleup partnerships established by the Data2Action event and stimulate, track, and help sustain new regional and grassroots partnerships around Big Data.

- Big Data Hubs could, for example:
  - Focus on solutions to specific global and societal challenges by convening stakeholders across sectors to partner in results-driven programs and projects.
  - Act as a matchmaker between the various academic, industry, and community stakeholders to help drive successful pilot programs for emerging Big Data technology.

Big Data Hubs examples…

- Coordinate across multiple regions of the country, based on shared interests and industry sector engagement to enable dialogue and share best practices.
- Aim to increase the speed and volume of technology transfer between universities, public and private research centers and laboratories, large enterprises, and SMB's.
- Facilitate engagement with opinion and thought leaders on the societal impact of Big Data technologies as to maximize positive outcomes of adoption while reducing unwanted consequences.
- Support the education and training of the entire Big Data workforce, from data scientists to managers to data end-users.

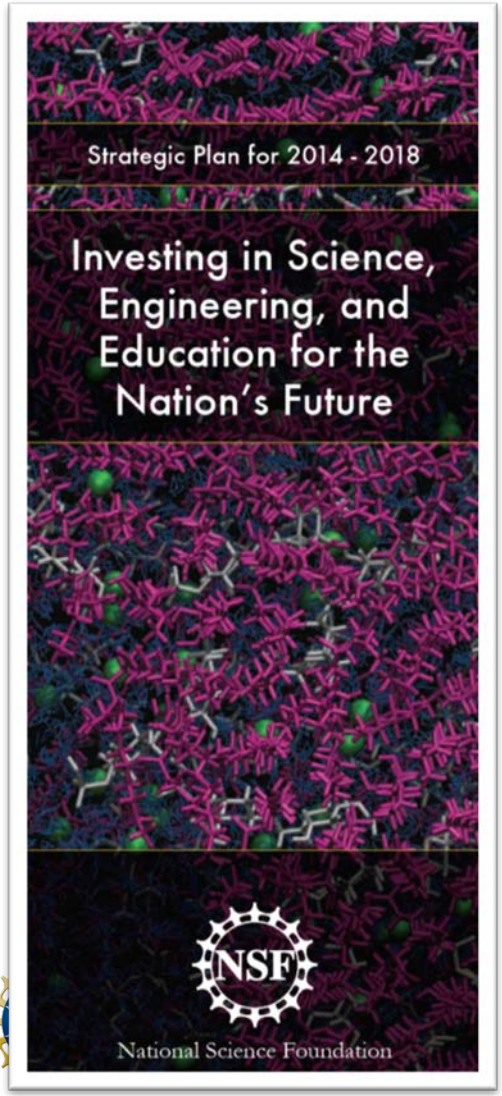
➔ Respond by Nov 1 to bigdata@nsf.gov
http://www.nsf.gov/cise/news/2014-bigdata-rfi.jsp

RFI National Big Data R&D Initiative

- Request For Input
  - …how we might best develop an overarching, comprehensive framework to support national-scale big data science and engineering research and education, discoveries and innovation."

- https://www.nitrd.gov/bigdata/rfi/02102014.aspx

Increase the Nation's Data Science Capacity – An NSF Priority Goal



- Data science education: Improve the nation's capacity in data science by investing in the development of human capital and infrastructure.
- By September 30, 2015:
  - Implement mechanisms to support the training and workforce development of future data scientists;
  - Increase the number of multi-stakeholder partnerships to address the nation's big-data challenges; and
  - Increase investments in current and future data infrastructure, extending data-intensive science into more research communities.

Thank You!

cbaru@nsf.gov