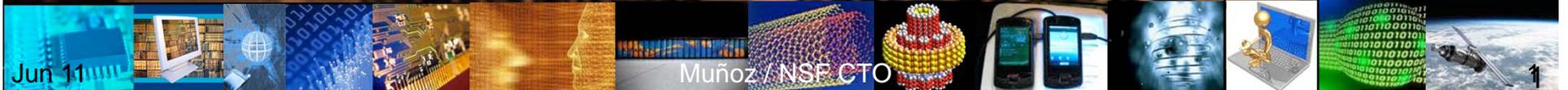


Dealing with Data in this Age

*José L. Muñoz, PhD
NSF
Chief Technology Officer*

Jun 11

Muñoz / NSF CTO





Outline

- ❖ Data Motivation
- ❖ Data: Most recent tool to the scientific method
- ❖ Data at NSF
- ❖ IT Workforce
- ❖ Opportunities at NSF
- ❖ Summary





A Data Perspective

❖ The world's "information" is doubling every two years!

➤ 1.8 zettabytes to be created and replicated in 2011



200 Billion HD movies (120 min)

1 person 47M years to view 24/7

* Zettabyte = 10^{21}

57.5B 32GB Apple iPads



DVD stack

By 2020 the amount of digital "information will have grown to 35 zettabytes!!





THE Issue

DATA \neq



INFORMATION \neq

KNOWLEDGE





Science and Society Transformed by Data

- ❖ Modern science
 - Data- and compute-intensive
 - Integrative, multi-scale
- ❖ Multi-disciplinary Collaborations for Complexity
 - Individuals, groups, teams, communities
 - ❖ Sea of Data
 - Age of Observation
 - Distributed, central repositories, sensor-driven, diverse, etc

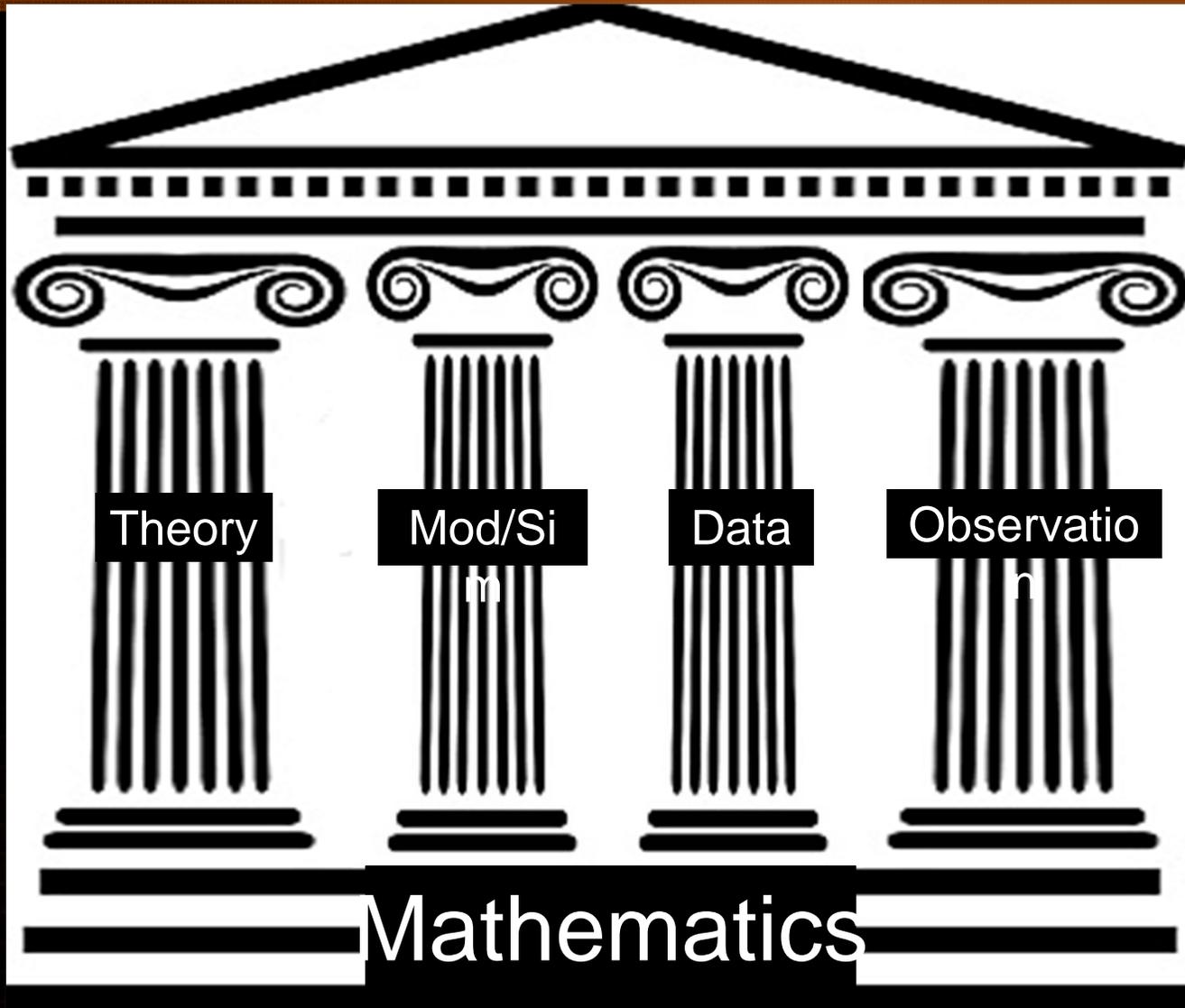


Software is also DATA





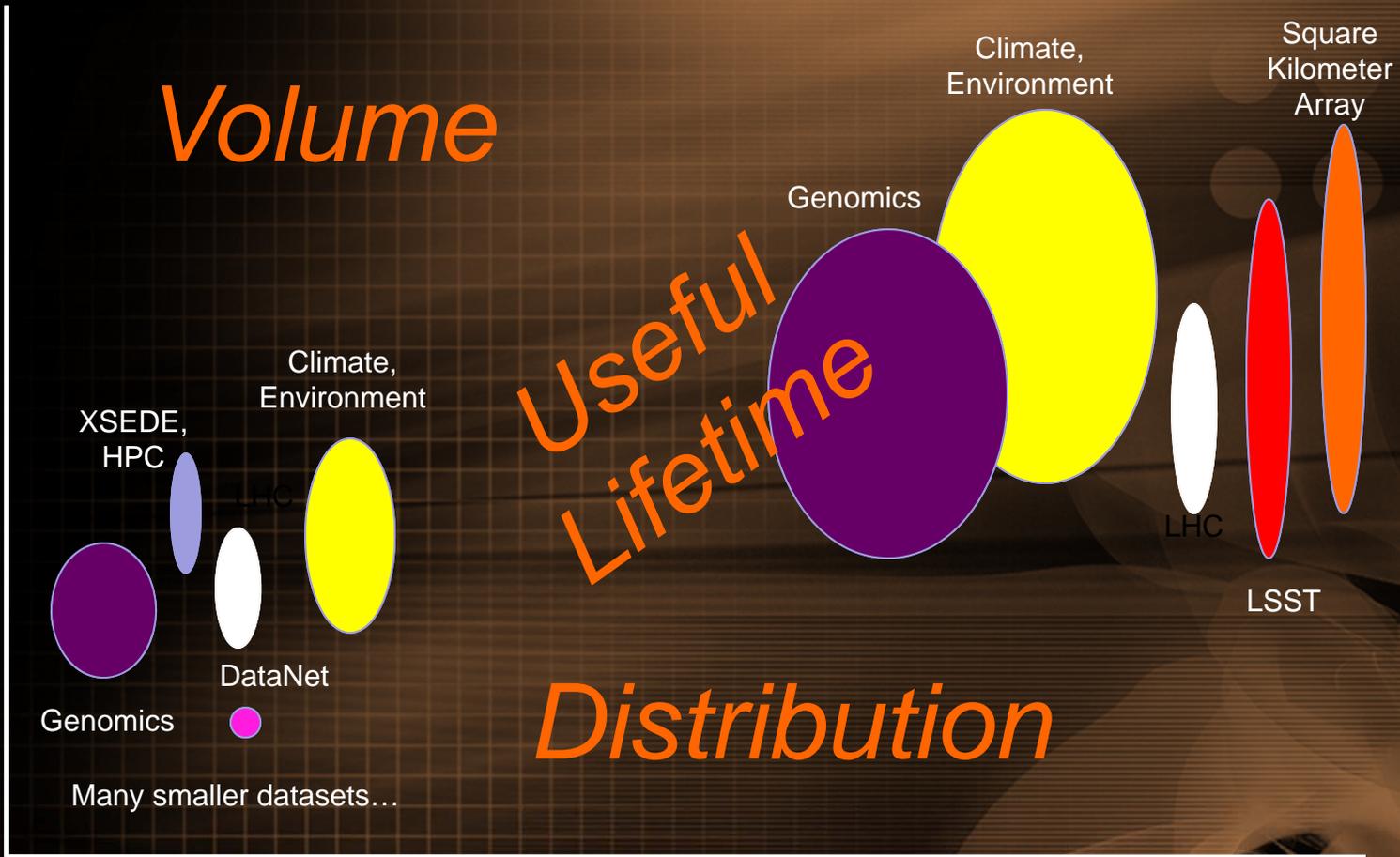
Scientific Method





Scientific Data Challenges

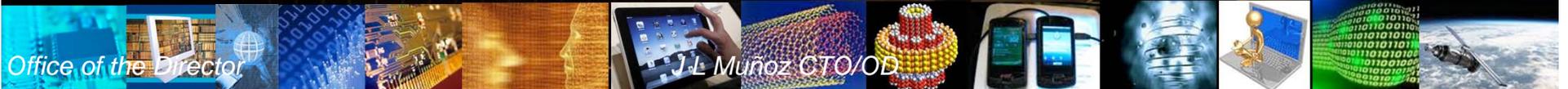
Bytes per day



2012

2020

Data Access





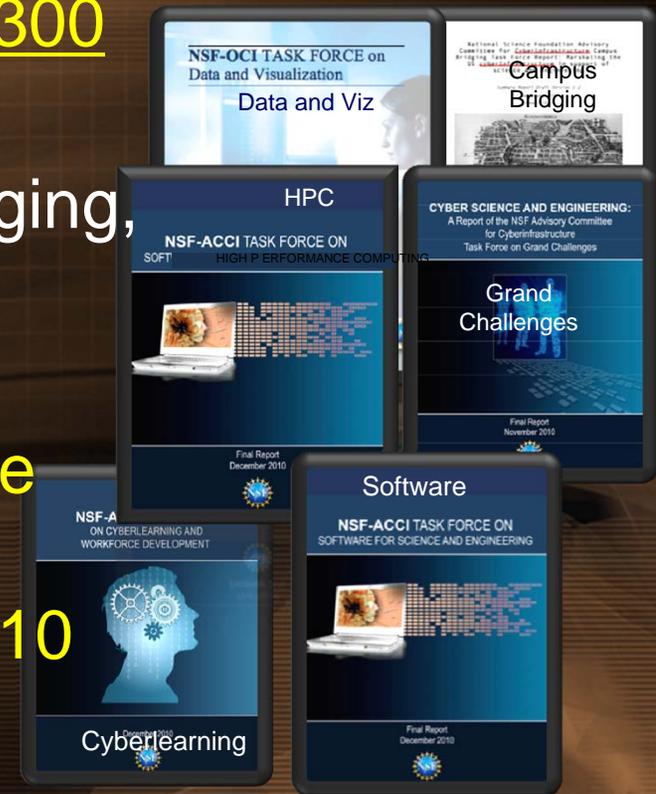
ACCI Task Force Reports: Data

- ❖ More than 25 workshops and Birds of a Feather sessions and more than 1300 people involved

- Data&Viz., HPC, Campus Bridging, Software, Grand Challenges, Cyberlearning&WFD

- ❖ Recommendations presented to the NSF Advisory Committee on Cyberinfrastructure (ACCI) Dec 2010

- ❖ Final reports on-line at:
<http://www.nsf.gov/od/oci/taskforces/>





Data Task Force Recommendations

Infrastructure:

Recognize data infrastructure and services (including visualization) as essential long term research assets fundamental to today's science

Economic sustainability:

Develop realistic cost models to underpin institutional/national business plans for research repositories/data services

Culture Change:

Emphasize expectations for data sharing; support the establishment of new citation models in which data and software tool providers and developers are recognized and credited with their contributions

Data Management Guidelines:

Identify and share best-practices for the critical areas of data management

Ethics and IP:

Train researchers in privacy-preserving data access

Data management plans: <https://dmp.cdlib.org/>





DataNet: A Multi-tiered and Multi-Disciplinary Landscape

Data-enabled Science

Genomics Communities

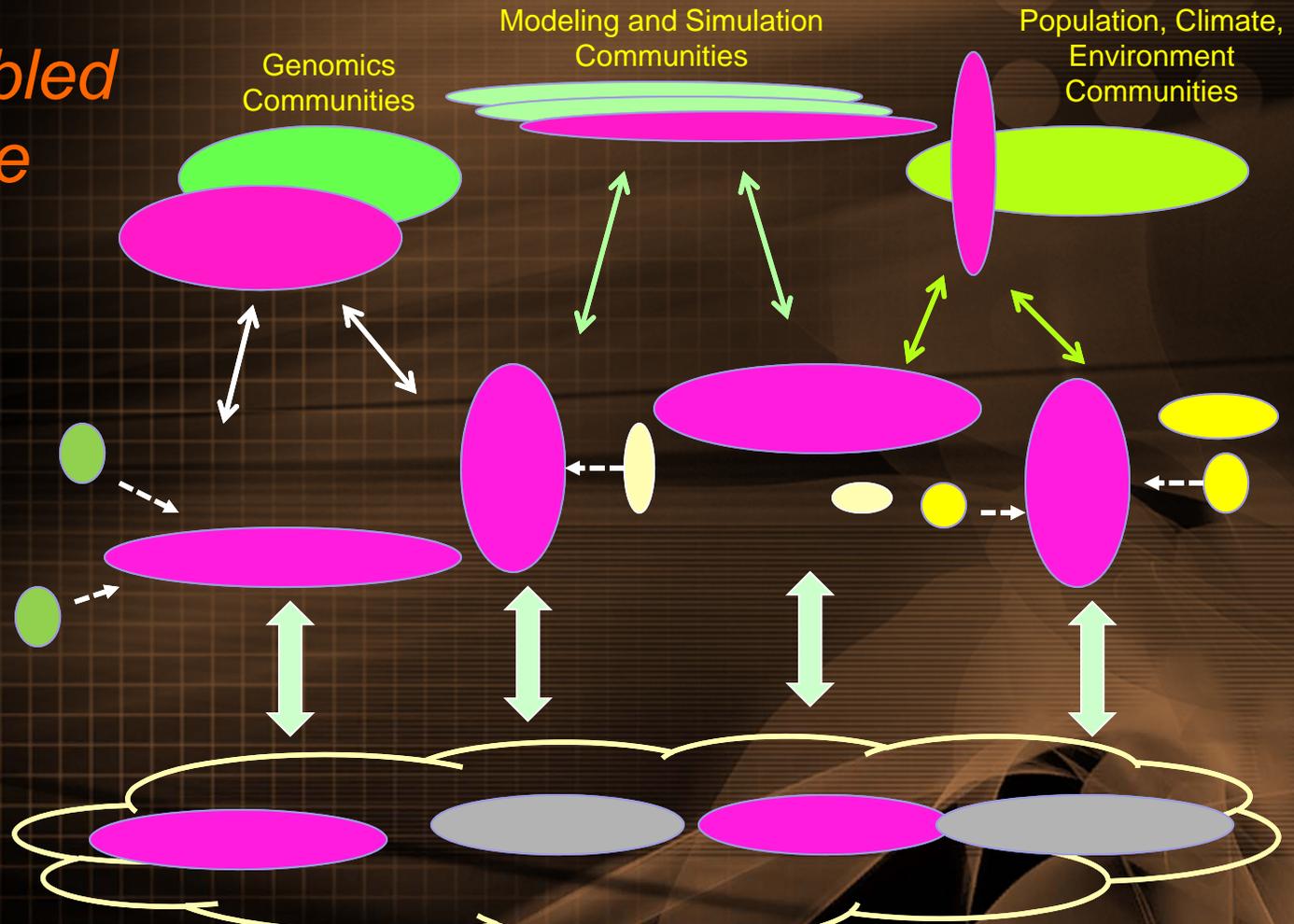
Modeling and Simulation Communities

Population, Climate, Environment Communities

Data Curation

Data Storage

 DataNet supported





Data-Enabled Science

❖ Data Services Program (*data*)

- Provide reliable digital preservation, access, integration, and analysis capabilities for science and/or engineering data over a decades-long timeline

❖ Data Analysis and Tools Program (*information*)

- Data mining, manipulation, modeling, visualization, decision-making systems

❖ Data-intensive Science Program (*knowledge*)

- Intensive disciplinary efforts
- Simulation, modeling
- Multi-disciplinary S&E





Cross Cutting Challenges

- ❖ **Balancing research into next generation's infrastructure with operation & maintenance of current capacity.**
 - Stimulate innovation and manage transitions
- ❖ **Sustainable, long term programs**
 - Technical design, development of sustainability models, and integration with the research cycle.
- ❖ **Integration**
 - Vertical – Linking low-level bit storage infrastructure to data collections, and finally to applications
 - Horizontal– Achieving connectivity and interoperability between activities that vary in scale, disciplinarity, and funding source.





NSF Opportunities

- ❖ OCI's Strategic Technologies for CyberInfrastructure would entertain data focused programs: 1 Feb 2012 (Pennington/Thompson)
- ❖ Be on the look-out for a new (re-focused) DataNet solicitation (Pennington) (1qtr FY12)
- ❖ Other Directorates and Offices are expected to have data focused programs under CIF-21





IT WORKFORCE LANDSCAPE

- ❖ 78% increase in employment in CS & math ... only 17% otherwise (GAO)
 - but STEM grads decreased from 32% to 27%
- ❖ IT careers high in demand (BLS)
- ❖ 92% of IT grads work outside the IT industry
 - health care, business, finance, manuf., ...
 - not an IT guru but rather a “versatilist”
- ❖ Imperative to stay on top of evolving IT!
- ❖ Problem solving skills (computational thinking)
- ❖ E-commerce and cyber-security!



<http://www.jsonline.com/sponsoredarticles/education/131782468.html>





Skills New IT Grads are Lacking

- ❖ An understanding of basic business functions
 - accounts receivables, marketing plans, ...
- ❖ Experience with enterprise systems integration
 - ability to integrate with other systems
- ❖ Knowledge of emerging enterprise technologies
 - technology is changing, need to keep up



Computerworld
Oct 2011



- ❖ Tech basic basics
- ❖ Familiarity of legacy systems
- ❖ Real-world perspective
 - what's best for the company?
- ❖ Ability to work as a team
 - Facebook and Twitter not the answer





Learning and Workforce Development in OCI

Catalyzing and Nurturing the Next Generation of Scientists

❖ Concentration on education, workforce development and training that complements and enhances the programs of OCI: Data, VO, Networking, Software, and HPC

- Training the next generation of CI scientists who will lead new technological developments
- Training scientists across disciplines to exploit current technologies to transform disciplinary scientific discovery and be prepared to nimbly assimilate into practice technological advancements

CE21 Focus:
Build a
computational
savvy 21st century
workforce

CI TEAM
Focus:
Undergraduate
and Graduate
Education

CI TraCS
Focus:
Postdocs and
Mid-Career





Summary

- ❖ **BIG** data is here now and only getting **BIGGER**
- ❖ Data enabled science: “data scientists” along with computational scientists will be in greater demand
- ❖ Collecting/saving data is a small part of the problem... from data to knowledge
- ❖ Positive outlook for IT in future... with caveats
 - data storage, cybersecurity, access, etc.
- ❖ NSF’s CIF-21 has participation from all NSF Directorates and Offices





THANK YOU





Zettabyte

1000 Megabytes = 1GB

1000 Gigabytes = 1TB

1000 Terabytes = 1PB

1000 Petabytes = 1EB

1000 Exabytes = 1Zettabyte



Zetta-byte (sextillion)

