Data- and Compute-Driven Transformation of Modern Science

Edward Seidel
Assistant Director, Mathematical and Physical Sciences, NSF
Profound Transformation of Science

**Gravitational Physics**

- Galileo, Newton usher in birth of modern science: c. 1600
- Problem: single "particle" (apple) in gravitational field (General 2 body-problem already too hard)

**Methods**
- Data: notebooks (Kbytes)
- Theory: driven by data
- Computation: calculus by hand (1 Flop/s)

**Collaboration**
- 1 brilliant scientist, 1-2 student
Part 1: Changing Cultures and Methodologies of Science…and the crises they create…
Profound Transformation of Science

Collision of Two Black Holes

1972: Hawking. 1 person, no computer 50 KB

1995: 10 people, large computer, 50MB

1998: 15 people, larger computer, 50GB
Community Einstein Toolkit

“Einstein Toolkit: open software for astrophysics to enable new science, facilitate interdisciplinary research and use emerging petascale computers and advanced CI.”

- Consortium: 67 members, 29 sites, 11 countries
- Simulation credits: Luciano Rezzolla, Max Planck Institut für Gravitationsphysik (Albert-Einstein-Institut)
Community Einstein Toolkit

“Einstein Toolkit: open software for astrophysics to enable new science, facilitate interdisciplinary research and use emerging petascale computers and advanced CI.”

- Consortium: 67 members, 29 sites, 11 countries
- Simulation credits: Luciano Rezzolla, Max Planck Institut für Gravitationsphysik (Albert-Einstein-Institut)

Many groups can do this: field explodes!
Major triumph of Computational Science---solve EEs!

Community + software + algorithms + hardware + …
Just ahead: Complexity of Universe

*LHC, Gamma-ray bursts!*

- Gamma-ray bursts!
  - Now: complex problems in relativistic astrophysics
  - Relativity, hydrodynamics, nuclear physics, radiation, neutrinos, magnetic fields: globally distributed collab!
  - Scalable algorithms, complex simulation codes, viz, PFlops*week, PB output!

- Gravity and general relativity are transformed
  - 4 centuries of small science, small data culture
  - 2-3 decades of radical change in both data (factors of 1000 per~5 years) and collaboration
New era of science after a century! Data- and compute-dominated gravitational wave astronomy!

- Scalable algorithms, complex simulation codes, viz, PFlops*week, PB output!
- Gravity and general relativity are transformed
  - 4 centuries of small science, small data culture
  - 2-3 decades of radical change in both data (factors of 1000 per~5 years) and collaboration
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics

Astronomy 1500-2010 was passive. No longer!
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes

- Simulations integrate all physics

Astronomy 1500-2010 was passive. No longer!
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics

Astronomy 1500-2010 was passive. No longer!
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics

Astronomy 1500-2010 was passive. No longer!
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics

**Astronomy 1500-2010 was passive. No longer!**
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes

- Simulations integrate all physics

Astronomy 1500-2010 was passive. No longer!
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics

*Astronomy 1500-2010 was passive. No longer!*
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
    - Indigo?
    - LSST
    - SKA = exabytes
- Simulations integrate all physics

Will require integration across disciplines, end-to-end
Transient & Data-intensive Astronomy

- New era: seeing events as they occur
  - (Almost) here now
    - ALMA, EVLA in radio
    - Ice Cube neutrinos
  - On horizon
    - 24-42m optical?
    - Indo-US transient collaboration
  - Indo-US collaboration

Simultaneous integration of all physics

Will require integration across disciplines, end-to-end

Communities need to share data, software, knowledge, in real time
Big Data vs The Long Tail of Science

- Many “Big Data” projects are “special”
  - Tend to be highly organized, have singular sources of data, professionally curated, a lot attention paid to them

- What about the “Long Tail” (the other 99%)?
  - Thousands of biologists sequencing communities of organisms
  - Thousands of chemist and materials scientists developing a “materials genome”
  - Millions of people “Tweeting”...

- Characteristics:
  - Heterogeneous, perhaps hand generated
  - Not curated, reused, served, etc...
Big Data vs The Long Tail

- Many “Big Data” projects are “special”
  - Tend to be highly organized, have singular sources of data, professionally curated, a lot of attention paid to them

- What about the “Long Tail” (the other 99%)?
  - Thousands of biologists sequencing communities of organisms
  - Thousands of chemists and materials scientists developing a “materials genome”
  - Millions of people “Tweeting”...

- Characteristics:
  - Heterogeneous, perhaps hand generated
  - Not curated, reused, served, etc...
Grand Challenge Communities Combine it All…
Where is it going to go?

Same CI useful for black holes, hurricanes
Grand Challenge Communities for

- Require many disciplines, all scales of collaborations
  - Individuals, groups, teams, communities
  - Multiscale Collaborations: Beyond teams
- Are dynamic and highly multidisciplinary
  - Time domain astronomy, emergency forecasting, metagenomics, materials genome...
- Drive sharing technologies and methodologies
- Researchers collaborate, work by sharing data. Places requirements on:
  - Software, networks, collaborative environments, data, sharing, computing, etc
  - Scientific culture, reproducibility, access, university structures
  - “Publications.” What is a modern publication?
Grand Challenge Communities for

- Require many disciplines, all scales of collaborations
  - Individuals, groups, teams, communities
  - Multiscale Collaborations: Beyond teams
- Are dynamic and highly multidisciplinary
  - Time domain astronomy, emergency forecasting, metagenomics, materials genome...
- Drive sharing technologies and methodologies
- Researchers collaborate, work by sharing data. Places requirements on:
  - Software, networks, collaborative environments, data, sharing, computing, etc
  - Scientific culture, reproducibility, access, university structures
  - "Publications." What is a modern publication?
Grand Challenge Communities for

- Require many disciplines, all scales of collaborations
  - Individuals, groups, teams, communities
  - Multiscale Collaborations: Beyond teams

- Are dynamic and highly multidisciplinary
  - Time domain astronomy, emergency forecasting, metagenomics, materials genome...

- Drive sharing technologies

- Researchers collaborate, work by sharing data. Places requirements on:
  - Software, networks, collaborative environments, data, sharing, computing, etc
  - Scientific culture, reproducibility, access, university structures
  - “Publications.” What is a modern publication?
Grand Challenge Communities for

- Require many disciplines, all scales of collaborations
  - Individuals, groups, teams, communities
  - Multiscale Collaborations: Beyond teams
- Are dynamic and highly multidisciplinary
  - Time domain astronomy, emergency forecasting, metagenomics, materials genome...
- Drive sharing technologies and methodologies
- Researchers collaborate, work by sharing data. Places requirements on:
  - Software, networks, collaborative environments, data, sharing, computing, etc
  - Scientific culture, reproducibility, access, university structures
  - “Publications.” What is a modern publication?
Grand Challenge Communities for

- Require many disciplines, all scales of collaborations
  - Individuals, groups, teams, communities
  - Multiscale Collaborations: Beyond teams

- Are dynamic and highly multidisciplinary
  - Time domain astronomy, emergency forecasting, metagenomics, materials genome...

- Drive sharing technologies and methodologies
- Researchers collaborate, work by sharing data. Places requirements on:
  - Software, networks, collaborative environments, data, sharing, computing, etc
  - Scientific culture, reproducibility, access, university structures
  - “Publications.” What is a modern publication?

Social, behavioral and economic sciences will be critical in helping us understand these issues...
Scenarios like this in all fields
Framing the Challenge: Science and Society Transformed by Data

- **Modern science**
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^{9-12}$ change!

- **Multi-disciplinary Collaborations**
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail

- **Sea of Data**
  - Age of Observation
Framing the Challenge: Science and Society Transformed by Data

- **Modern science**
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^{9-12}$ change!

- **Multi-disciplinary Collaborations**
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail
  - Sea of Data
  - Age of Observation
Framing the Challenge: Science and Society Transformed by Data

- **Modern science**
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^9-10^{12}$ change!

- **Multi-disciplinary Collaborations**
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail
  
- **Sea of Data**
  - Age of Observation
Framing the Challenge:
Science and Society Transformed by Data

- **Modern science**
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^{9-12}$ change!

- **Multi-disciplinary Collaborations**
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail
  - Sea of Data
  - Age of Observation
Framing the Challenge: Science and Society Transformed by Data

- **Modern science**
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^{9-12}$ change!

- **Multi-disciplinary Collaborations**
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail

- **Sea of Data**
  - Age of Observation
Framing the Challenge: Science and Society Transformed by Data

- Modern science
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^{9-12}$ change!

- Multi-disciplinary Collaborations
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail
  - Sea of Data
  - Age of Observation

We still think like this...
Framing the Challenge: Science and Society Transformed by Data

- Modern science
  - Data- and compute-intensive
  - Integrative, multiscale
  - 4 centuries of constancy, 4 decades $10^9$-$10^{12}$ change!

- Multi-disciplinary Collaborations
  - Individuals (Galileo!)
  - Groups, teams, Grand Challenge Communities
  - Big Data + Long Tail
  - Sea of Data
    - Age of Observation

We still think like this...

...But such radical change cannot be adequately addressed with (current) incremental approach!
We still think like this...

...But such radical change cannot be adequately addressed with (current) incremental approach!

Modern science
- Data- and compute-intensive
- Integrative, multiscale
- 4 centuries of constancy, 4 decades $10^9$-$10^{12}$ change!

Multi-disciplinary Collaborations
- Individuals (Galileo!)
- Groups, teams, Grand Challenge Communities
- Big Data + Long Tail

Sea of Data
- Age of Observation
Part 2: Crises to Deal With
Part 2: Crises to Deal With
Part 2: Crises to Deal With

Computing

Data

Software

End-to-end Networks

Instruments & Facilities

Organizational structures

Education

No, we are not...
Three Crises

- **Computing Technology**
  - Multicore: processor is new transistor
  - Programming model, fault tolerance, etc
  - New models: clouds, grids, GPUs, ... where appropriate

- **Data, provenance, and visualization**
  - Generating more data than in all of human history: preserve, mine, share?
  - How do we create “data scientists”?

- **Software**
  - Complex applications on coupled compute-data-networked environments, tools needed
  - Modern apps: $10^6+$ lines, many groups contribute, take decades
Data Crisis: Information Big Bang

![Graph showing the growth of information in billions of gigabytes from 1999 to 2009. The graph compares All Disk Storage, All Info/Yr, Unique Info/Yr, All Human Documents (40k Yrs), All Words In All Lives, and the Amount Can Store In Human Minds in 1 Yr. Sources: Lesk, Berkeley SIMS, Landauer, EMC.](image)
Data Crisis: Information Big Bang

Sources: Lesk, Berkeley SIMS, Landauer, EMC
there is a pending crisis in archiving... we have to create long-term methods for preserving information, for making it available for analysis in the future.”
80% respondents: >50 yrs; 68% > 100 yrs
Two More Crises

- **Organization for Multidisciplinary & Computational Science**
  - “Universities must significantly change organizational structures: multidisciplinary & collaborative research are needed [for US] to remain competitive in global science”
  - “Itself a discipline, computational science advances all science...inadequate/outmoded structures within Federal government and the academy do not effectively support this critical multidisciplinary field”

- **Education**
  - The CI environment is running away from us!
  - How do we develop a workforce to work effectively in this world?
  - How do we help universities transition?
Two More Crises

- Organization for Multidisciplinary & Computational Science
  - “Universities must significantly change organizational structures: multidisciplinary & collaborative research are needed [for US] to remain competitive in global science”

- Education
  - The CI environment is running away from us!
  - How do we develop a workforce to work effectively in this world?
  - How do we help universities transition?
NSF Vision and US National CI Blueprint

Software

Track 1

Campus

DataNet

Track 2

Campus

DataNet

Nets

Track 2

Campus

DataNet

Track 2

Campus

DataNet

Track 2

Campus
NSF Vision and US National CI Blueprint
NSF Vision and US National CI Blueprint
Education Crisis: I need all of this to start to solve my problem!
Education Crisis: I need all of this to start to solve my problem!

Science demands integration of all components of this ecosystem.
Education Crisis: I need all of this to start to solve my problem!
Education Crisis: I need all of this to start to solve my problem!

Science is becoming unreproducible in this environment. Validation? Provenance? Reproducibility?
The Opportunity!

- We have critical elements in place for an architecture
  - NSF XSEDE architecture can connect...
    - Campus Bridging: campus to national CI...
      - Campus Assets: MRI, Instruments, DNA sequencers...
      - Campus: InCommon, Gateways, Open Science Grid
      - Campus Applications: SI2, SAGA, etc...
      - Networks: end-to-end connectivity
      - Middleware, e.g. Globus Online
    - Facilities to campuses...
      - MREFC: telescopes, accelerators, light sources, ...“More silicon than Steel”
      - NEON, OOI, LSST, etc... how to integrate?
    - International (e.g., LHC), Commercial (e.g., Clouds) to local
  - XSEDE can enable...
Part 3: Recommendations
ACCI Task Force Reports

- Final recommendations presented to the NSF Advisory Committee on Cyberinfrastructure Dec 2010
- More than 25 workshops and Birds of a Feather sessions, 1300 people involved
- Final reports on-line

"Permanent programmatic activities in Computational and Data-Enabled Science & Engineering (CDS&E) should be established within NSF." Grand Challenges Task Force

"NSF should establish processes to collect community requirements and plan long-term software roadmaps” Software Task Force

"NSF should fund interdisciplinary research on the science of broadening participation” Cyberlearning Task Force
Part 4: Focus on Data
The Shift Towards a “Sea of Data”

Implications

- Science & society are now data-dominated
  - Experiment, computation, theory
  - Fourth paradigm
  - US mobile phone traffic exceeded 1 exabyte!

- Classes of data
  - Collections, observations, experiments, simulations
  - Software
  - Publications

- Totally new methodologies
  - Algorithms, mathematics, culture

- Data become the medium for
  - Multidisciplinarity, communication, publication...science
The Shift Towards a “Sea of Data”

Implications

- Science & society are now data-dominated
  - Experiment, computation, theory
  - Computation
  - US mobile phone traffic exceeded 1 exabyte!

- Classes of data
  - Collections, observations, experiments, simulations
  - Software
  - Publications

- Totally new methodologies
  - Algorithms, mathematics, culture

- Data become the medium for
  - Multidisciplinarity, communication, publication...science
The Shift Towards a “Sea of Data”

Implications

- Science & society are now data-dominated
  - Experiment, computation, theory
  - Fourth paradigm
  - US mobile phone traffic exceeded 1 exabyte!

- Classes of data
  - Collections, observations, experiments, simulations
  - Software
  - Publications

- Totally new methodologies
  - Algorithms, mathematics, culture

- Data become the medium for
  - Multidisciplinarity, communication, publication...science

Fundamental questions become focused around data: How to remove boundaries? How to incentivize sharing?

How do we attribute credit for this new publication form? How are data peer reviewed? What is a publication in the modern data-rich world?
Recent NSF Activities on Data Policy and Implementation
Fundamental points on data and publication policy

- Communities work together/advance through sharing of data, pubs & software (which is data)
- Publicly funded scientific data and publications should be available, and science benefits
- There has to be a place to keep data, and a way to access it
- There needs to be an affordable, sustainable cost model for this
Fundamental points on data and publication policy

- Communities work to advance through sharing of data, pubs & software (which is data)
- Publicly funded scientific data and publications should be available, and science benefits
- There has to be a place to keep data, and a way to access it
- There needs to be an affordable, sustainable cost model for this

What data must be made available? Raw data? Peer reviewed? When is it available? 6 months? 1 year? After publication?
Fundamental points on data and publication policy

- Communities work together and advance through sharing of data, publications, and software (which is data).
- Publicly funded scientific data and publications should be available, and science benefits.
- There has to be a place to keep data, and a way to access it.
- There needs to be an affordable, sustainable cost model for this.

What data must be made available? Raw data? Peer reviewed? When is it available? 6 months? 1 year? After publication?

Where is it placed? Author web site? Library? NCSA? EU repository?
Fundamental points on data and publication policy

- Communities work together/advance through sharing of data, pubs & software (which is data).
- Publicly funded scientific data and publications should be available, and science benefits.
- There has to be a place to keep data, and a way to access it.
- There needs to be an affordable, sustainable cost model for this.

What data must be made available? Raw data? Peer reviewed? When is it available? 6 months? 1 year? After publication?

Who pays? Agency? The Institution? What is the cost model? What is reasonable?

Where is it placed? Author web site? Library? NCSA? EU repository?
Fundamental points on data and publication policy

- Communities work together/advance through sharing of data, pubs & software (which is data)
- Publicly funded scientific data and publications should be available, and science benefits
- There has to be a place to keep data, and a way to access it
- There needs to be an affordable, sustainable cost model for this

What data must be made available? Raw data? Peer reviewed? When is it available? 6 months? 1 year? After publication?

Who pays? Agency? The Institution? What is the cost model? What is reasonable?

Where is it placed? Author web site? Library? NCSA? EU repository?

How long is it made available? How do we enforce/serve it post-award?
Fundamental points on data and publication policy

- Communities work together/advance through sharing of data, pubs & software (which is data)
- Publicly funded scientific data and publications should be available, and science benefits
- There has to be a place to keep data, and a way to access it
- There needs to be an affordable, sustainable cost model for this

What data must be made available? Raw data? Peer reviewed? When is it available? 6 months? 1 year? After publication?

Who pays? Agency? The Institution? What is the cost model? What is reasonable?

Where is it placed? Author web site? Library? NCSA? EU repository?

How long is it made available? How do we enforce/serve it post-award?

There is great variability in requirements across science communities: app driven concept
Changes Coming at NSF for Data!

* Data are becoming:
  - Primary means of communication through sharing
  - Major product of research (including publication)

* Long-standing NSF Policy on Data:
  - “Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data... created or gathered in the course of work under NSF grants”

* NSF now requires a Data Management Plan (DMP):
  - 2-page supplement to the proposal
    - subject to peer review; criterion for award
  - Not possible to submit proposals without DMP

* National Science Board beginning to examine policy
Changes Coming at NSF for Data!

- Data are becoming:
  - Primary means of communication through sharing
  - Major product of research (including publication)

- Long-standing NSF Policy on Data:
  - "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data... created or gathered in the course of work under NSF grants"

- NSF now requires a Data Management Plan (DMP):
  - 2-page supplement to the proposal
  - subject to peer review; criterion for award
  - Not possible to submit proposals without DMP

- National Science Board beginning to examine policy
Critical Lessons to Take Home

- Comprehensive approach needed to address complex problems of 21\textsuperscript{st} century
  - All elements must be addressed, not just a few; can’t even start to address problems without all
  - Many exponentials: data, compute, collaborate

- Data-intensive science increasingly dominant
  - Modern data-driven CI presents numerous crises, opportunities; long tail and grand challenge communities
    - Impacts CI from campus to facility
    - Policy changes needed; publication, reproducibility

- Academia and Agencies must address
  - Rethinking Academic Structures, Curriculum, P&T