Science Informatics

Dan Fay
Director – Earth, Energy and Environment
dan.fay@microsoft.com
MSR eScience Workshop
Looking Back 8 yrs to the Beginning

Scientific Data Intensive Computing Workshop 2004

- Keynote: 20 Questions to a Better Application – Jim Gray
  Online Science the New Computational Science
- Talk: Data Explosion: Astrophysics with Terabytes of Data
  - Alex Szalay
**Online Science the New Computational Science**

**Information Avalanche**
- In science, industry, government,...
  - better observational instruments and
  - and, better simulations
  - producing a data avalanche
- **Examples**
  - BaBar: Grows 1TB/day
  - 2/3 simulation Information
  - 1/3 observational Information
  - CERN: LHC will generate 1GB/s, ~10 PB/y
  - VLBA (NRAO) generates 1GB/s today
  - Pixel: 100 TB/Move
- **New emphasis on informatics:**
  - Capturing, Organizing, Summarizing, Analyzing, Visualizing

**Publishing Data**
- **Roles**
  - Traditional: Scientists
  - Emerging: Collaborations, Project, www site, Bigger, Archives, Scientists
- **Exponential growth:**
  - Projects last at least 3-5 years
  - Data sent upwards only at the end of the project
  - Data will never be centralized
- **More responsibility on projects**
  - Becoming Publishers and Curators
  - Often no explicit funding to do this (must change)
- **Data will reside with projects**
  - Analyses must be close to the data (see later)
- **Data cross-correlated with Literature and Metadata**

**Global Federations**
- **Massive datasets live near their owners:**
  - Near the instrument's software pipeline
  - Near the applications
  - Near data knowledge and curation
- **Each Archive publishes a (web) service**
  - Schema: documents the data
  - Methods on objects (queries)
- **Scientists get "personalized" extracts**
- **Uniform access to multiple Archives**
  - A common global schema

**What's X-info Needs from us (cs)**
- **Not drawn to scale**
- **Scientists**
  - Science Data & Questions
  - Data Mining Algorithms
- **Miners**
  - Database
  - To store data
  - Execute Queries
- **Plumbers**
  - Question & Answer Visualization
- **How to Help?**
  - Can't learn the discipline before you start (takes 4 years.)
  - Can't go native — you are a CS person not a bio,... person
  - Have to learn how to communicate
  - Have to learn the language
  - Have to form a working relationship with domain expert(s)
  - Have to find problems that leverage your skills

**Call to Action**
- X-info is emerging.
- Computer Scientists can help in many ways.
  - Tools
  - Concepts
  - Provide technology consulting to the community
- There are great CS research problems here
  - Modeling
  - Analysis
  - Visualization
  - Architecture
Microsoft Research

Not all moods are created equal

Learn how researchers are mining social media to infer people's moods >>

Highlights

- Microsoft Research Opens a Lab in New York City
- Video: Opportunities for the Future of Computer Research

Spotlight


Feature Stories (1/3)

Kinect Launches a Surgical Revolution

Microsoft Research Cambridge researcher Helena Mentis is part of a U.K. research team exploring a way for surgeons to manipulate medical images with a wave of a hand. Read more...

Energy-Capping System Holds Promise of Greener Data Centers

Worldwide Locations

- Europe
- Asia
- Americas

More feature stories »
Earth, Energy & Environment

- **Visualizing and Experiencing E³ Data + Information**: Provide a unique experience to reduce time to insight and knowledge through visualizing data and information.

- **Accessible Data**: Ensure $E^3$ data (remote and local sensing) is easily discoverable, accessible and consumable in the scientists domain.

- **Enabling Scientific Collaboration**: Look at new ways to enable collaboration in scientific virtual organizations.
Emergence of a Fourth Paradigm

- Thousand years ago – **Experimental Science**
  - Description of natural phenomena
- Last few hundred years – **Theoretical Science**
  - Newton’s Laws, Maxwell’s Equations...
- Last few decades – **Computational Science**
  - Simulation of complex phenomena
- Today – **Data-Intensive Science**
  Scientists overwhelmed with data sets from many different sources
  - Data captured by instruments
  - Data generated by simulations
  - Data generated by sensor networks
- eScience is the set of tools and technologies to support data federation and collaboration
  - For analysis and data mining
  - For data visualization and exploration
  - For scholarly communication and dissemination

\[
\left( \frac{a}{a} \right)^2 = \frac{4\pi G \rho}{3} - K \frac{c^2}{a^2}
\]

Jim Gray 2007
Changing Nature of Discovery

• Complex models
  • Multidisciplinary interactions
  • Wide temporal and spatial scales
• Large multidisciplinary data
  • Real-time steams
  • Structured and unstructured
• Distributed communities
  • Virtual organizations
  • Socialization and management
• Diverse expectations
  • Client-centric and infrastructure-centric

http://fourthparadigm.org
The Problem for the e-Scientist

How to codify and represent our knowledge

The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to reorganize it
- How to share with others

- Query and Vis tools
- Building and executing models
- Integrating data and Literature
- Documenting experiments
- Curation and long-term preservation

(With thanks to Jim Gray)
Mountain Hydrology, Snow Color, and the Fourth Paradigm

Snow is one of nature's most colorful materials (Landsat Thematic Mapper snow & cloud)

Spatially distributed snow water equivalent

SWE, mm

4500
4000
3500
3000
2500
2000
1500
1000
500
100

(N. Molotch)

04/10/05
Information about water is more useful as we climb the value ladder

- **Monitoring**
- **Collation**
- **Quality assurance**
- **Aggregation**
- **Analysis**
- **Distribution**
- **Integration**
- **Forecasting**
- **Reporting**

---

**Done poorly, but a few notable counter-examples**

**Done poorly to moderately, not easy to find**

**Sometimes done well, generally discoverable and available, but could be improved**

(I. Zaslavsky & CSIRO, BOM, WMO)
Environmental Ecosystem

Knowledge

Inform

Action
Environmental Ecosystem

Knowledge
- Analysis
- Data
- Insight
- Publish

Technology

Inform

Action
- Communicate
- Decide
- Implement
Information ecosystem:
It is chaotic, unstructured and ad hoc
The Ecological Data Flood

• We’re living in a perfect storm of remote sensing, cheap ground-based sensors, internet data access, and commodity computing
• Yet deriving and extracting the variables needed for science remains problematic
  • Specialized knowledge for algorithms, internal file formats, data cleaning, etc, etc
  • Finding the right needle across the distributed heterogeneous and very rapidly growing haystacks
Data Variety – The Spice of Life

- Manual Measurement
- Automated Measurement
- Sample Collection
- Typing
- Aircraft Surveys
- Model Output
- Counting
- Historical Photographs
- Relatively Ubiquitous Motes
- Satellite
Data Integration Challenges

- Regular rasters, points, and spatial features
- Time series and intermittent
- Vocabulary meanings (ontology)
- Sparse in time, duration, or location
- Science variable derivation
- Gaps
- Spatial/temporal harmonization
Why Make this Distinction?

Provenance and trust widely varies
Data acquisition, early processing, and reporting ranges from a large government agency to individual scientists.
Smaller data often passed around in email; big data downloads can take days (if at all)

Data sharing concerns and patterns vary
Open access followed by (non-repeatable and tedious) pre-processing
True science ready data set but concerns about misuse, misunderstanding particularly for hard won data.

Computational tools differ.
Not everyone can get an account at a supercomputer center
Very large computations require engineering (error handling)
Space and time aren’t always simple dimensions

Complex shared detector
Simple instrument (if any)

Science happens when PBs, TBs, GBs, and KBs can be mashed up simply
Complex and Heavy process by experts
Ad hoc observations and models
AzureMODIS – Azure Service for Remote Sensing Geoscience

- Science pipeline for download, initial processing, and reduction of satellite imagery. Developed by MSR, UVA, UCB.
- Dramatically lowers resource and complexity barriers to use satellite imagery for terrestrial hydrology and geoscience.
  - Common imagery location determination and upload from diverse sources
  - Optional scientist-provided reduction algorithm (.NET, Java, or MatLab)
  - On-demand scalability beyond local desktop or cluster
- In use now to compute 10 year continental scale water balance for North America. Per year:
  - 500 GB (~60K files) upload of 9 different source imagery products from 15 different locations
  - 400 GB reprojected harmonized imagery consuming ~3500 cpu hours
  - 5 GB reduced science result leveraging reported field data aggregates consuming ~60 cpu hour

Catharine van Ingen (Microsoft Research), Jie Li, Marty Humphreys (UVA), Youngryel Ryu (UCB), Deb Agarwal (BWC/LBL)
Microsoft Codename "Cloud Numerics"
numerical and data analytics library for data scientists, quantitative analysts, and others

Microsoft Codename "Data Explorer"
organize, manage, mash up and gain new insights from your data.

Microsoft Codename "Data Hub"
Organizations to curate and publish its data on a private data marketplace

Microsoft Codename "Trust Services"
data encryption services for cloud applications so that they can roam encryption keys in a secure way.
Fetch Climate

Tools BETA new tools for new science

FetchClimate

Inspiration | Features | Try online | Download | Case studies | People | Acknowledgements

Retrieve climatic and environmental information with the click of a button or a few lines of code

FetchClimate is a fast, free, intelligent climate information service that operates over the cloud to return exactly the information you need. FetchClimate can be accessed either through a simple web interface, or via a few lines of code inside any .NET program. FetchClimate is intended to make it easy for you to retrieve information for any geographical region, at any grid resolution: from global, through continental, to a few kilometres, and for any range of years (1900 – 2010), days within the year, and/or hours within the day. FetchClimate can also report the uncertainty associated with the values it returns and list data sources used to fulfil the request. When multiple sources could potentially provide information on the same environmental variable, FetchClimate automatically selects the most appropriate data source. Finally, the entire query you ran can be shared as a single URL, enabling others to retrieve the identical information.

Inspiration

Many environmental science research and applications require information about climate and

http://fetchclimate.cloudapp.net/
Common Problems with Data

➢ To use data from different sources
  o Non-standard formats, scales, and units
  o Lack of data quality control
  o Lack of metadata
  o Difficult to repurpose data for different (my) tools

➢ To share data
  o Lack of incentive (no credit)
  o Need extra resources and tools

➢ Hidden problems, seldom addressed
  o Versioning
  o Provenance
  o Curation
Environmental Informatics Framework (EIF)

Current State of Data Ecosystem
Advance data discoverability, accessibility, and consumability

Open Data Protocol (OData)

http://www.odata.org

It allows you to form URLs based on what you know about the underlying data

- A Web protocol for querying and updating data
  - provides a way to unlock your data and free it from data silos
  - does this by building upon Web technologies such as HTTP, Atom Publishing Protocol (AtomPub) and JSON to provide access to information from a variety of applications, services, and stores.

- In Open Source/Specifications Promise – being submitted to OASIS

- An application of a set of internet standards:
  - HTTP,
  - Atom (RFC 4287),
  - AtomPub (RFC 5023),
  - REST semantics

- Existing standards + easy data access API

- Added Geospatial data support –
  - Feedback from the Community encouraged – www.odata.org
New ways to analyze and communicate data
The ‘Cosmic Genome’ Project

- The Sloan Digital Sky Survey is the first major astronomical survey project:
  - 5 color images of ¼ of the sky
  - Pictures of 300 million celestial objects
  - Distances to the closest 1 million galaxies
- Jim Gray from Microsoft Research worked with astronomer Alex Szalay to build the public ‘SkyServer’ archive for the survey
- New model of scientific publishing
  - Have to publish the data before astronomers publish their analysis
Public Use of the SkyServer

- **Posterchild in 21st century data publishing**
  - 380 million web hits in 6 years
  - 930,000 distinct users vs 10,000 astronomers
  - 1600 scientific papers
  - Delivered 50,000 hours of lectures to high schools
  - Delivered 100B rows of data

- **Citizen Science: GalaxyZoo**
  - Goal of 1 million visual galaxy classifications by the public
  - Allows general public to search for photographs and classify different types of galaxies
World Wide Telescope

Seamless Rich Social Media Virtual Sky and Earth
Web application for science and education

Goals
- Integration of data sets and one-click contextual access
- Easy access and use
- Tours for sharing information/insights
- Spatial/temporal visualization of 3D datasets

Updates
- API for extensibility
- Excel Add-in for easy data integration

Not just for Astronomy – Earth Focused
Being used in Planetariums – Cal Academy

We invite you to experience it!
www.worldwidetelescope.org
• Community Site for WWT Tours and Layers (Data)
• Sharing by groups/individuals
Excel to Visualize in seconds

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Index</td>
<td>lat</td>
<td>lon</td>
<td>alt</td>
<td>color</td>
<td>magnitude</td>
<td>startdate</td>
<td>enddate</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>72.5</td>
<td>-70</td>
<td>100000</td>
<td>white</td>
<td>0.25</td>
<td>1/1/2000</td>
<td>1/4/2000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>72.43387</td>
<td>-70.0101</td>
<td>99500</td>
<td>white</td>
<td>0.25</td>
<td>1/2/2000</td>
<td>1/5/2000</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>72.40775</td>
<td>-70.0202</td>
<td>99900</td>
<td>white</td>
<td>0.25</td>
<td>1/3/2000</td>
<td>1/6/2000</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>72.36162</td>
<td>-70.0303</td>
<td>99900</td>
<td>white</td>
<td>0.25</td>
<td>1/4/2000</td>
<td>1/7/2000</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>72.33549</td>
<td>-70.0404</td>
<td>99500</td>
<td>white</td>
<td>0.25</td>
<td>1/5/2000</td>
<td>1/8/2000</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>72.30847</td>
<td>-70.0505</td>
<td>99000</td>
<td>white</td>
<td>0.25</td>
<td>1/6/2000</td>
<td>1/9/2000</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>72.28137</td>
<td>-70.0606</td>
<td>99000</td>
<td>white</td>
<td>0.25</td>
<td>1/7/2000</td>
<td>1/10/2000</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
<td>72.25427</td>
<td>-70.0707</td>
<td>99300</td>
<td>white</td>
<td>0.25</td>
<td>1/8/2000</td>
<td>1/11/2000</td>
</tr>
<tr>
<td>10</td>
<td>8</td>
<td>72.21218</td>
<td>-70.0808</td>
<td>99200</td>
<td>white</td>
<td>0.25</td>
<td>1/9/2000</td>
<td>1/12/2000</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
<td>72.17318</td>
<td>-70.1001</td>
<td>98900</td>
<td>white</td>
<td>0.25</td>
<td>1/10/2000</td>
<td>1/13/2000</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>72.12418</td>
<td>-70.1101</td>
<td>99100</td>
<td>white</td>
<td>0.25</td>
<td>1/11/2000</td>
<td>1/14/2000</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>72.06518</td>
<td>-70.1202</td>
<td>98900</td>
<td>white</td>
<td>0.25</td>
<td>1/12/2000</td>
<td>1/15/2000</td>
</tr>
<tr>
<td>14</td>
<td>12</td>
<td>72.00618</td>
<td>-70.1303</td>
<td>98700</td>
<td>white</td>
<td>0.25</td>
<td>1/13/2000</td>
<td>1/16/2000</td>
</tr>
<tr>
<td>15</td>
<td>13</td>
<td>6.94718</td>
<td>-70.1404</td>
<td>98600</td>
<td>white</td>
<td>0.25</td>
<td>1/14/2000</td>
<td>1/17/2000</td>
</tr>
<tr>
<td>16</td>
<td>14</td>
<td>7.80422</td>
<td>-70.1415</td>
<td>98600</td>
<td>white</td>
<td>0.25</td>
<td>1/15/2000</td>
<td>1/18/2000</td>
</tr>
<tr>
<td>17</td>
<td>15</td>
<td>7.8081</td>
<td>-70.1516</td>
<td>98500</td>
<td>white</td>
<td>0.25</td>
<td>1/16/2000</td>
<td>1/19/2000</td>
</tr>
<tr>
<td>18</td>
<td>16</td>
<td>7.76197</td>
<td>-70.1616</td>
<td>98400</td>
<td>white</td>
<td>0.25</td>
<td>1/17/2000</td>
<td>1/20/2000</td>
</tr>
<tr>
<td>19</td>
<td>17</td>
<td>7.71584</td>
<td>-70.1717</td>
<td>98300</td>
<td>white</td>
<td>0.25</td>
<td>1/18/2000</td>
<td>1/21/2000</td>
</tr>
<tr>
<td>20</td>
<td>18</td>
<td>7.66571</td>
<td>-70.1818</td>
<td>98200</td>
<td>white</td>
<td>0.25</td>
<td>1/19/2000</td>
<td>1/22/2000</td>
</tr>
<tr>
<td>21</td>
<td>19</td>
<td>7.62539</td>
<td>-70.1919</td>
<td>98100</td>
<td>white</td>
<td>0.25</td>
<td>1/20/2000</td>
<td>1/23/2000</td>
</tr>
<tr>
<td>22</td>
<td>20</td>
<td>7.57746</td>
<td>-70.2020</td>
<td>98000</td>
<td>white</td>
<td>0.25</td>
<td>1/21/2000</td>
<td>1/24/2000</td>
</tr>
</tbody>
</table>

To map additional columns, include the column in your selection range and click the "Update Range" button available above.
Natural User Interfaces (NUI)  
Kinect SDK and WWT

- Rethinking ways in which people will interact with computers/technologies of the future
- Re-evaluating everything from their (non-) physical design to the human needs and interaction models
- Revolutionize the way we think about technology and what it can do on our behalf
Our Next Steps

• Continue to work with Scientists
  • Implement features from feedback
    • I.e. Netcdf support
  • New clients and User Interface
  • Make it easier to use/create tours

• Cloud: Windows Azure integration - Write, Run, or Use Software
  • Write: Platform as a Service
    • Web Sites
    • Storage
    • Languages - .Net, Java, Python, PHP, Ruby, NodeJS, C++
  • Run: Infrastructure as a Service
    • Virtual Machines – Windows Servers & Linux
    • Connect to on-premises
  • Use: Software as a Service
    • Azure Media Services
    • HPC and Big Data – Hadoop, etc
    • Windows Azure Marketplace
eScience in Action
Microsoft eScience Workshop 2012
October 8–9, 2012 | Chicago, Illinois
http://research.microsoft.com/events/escience2012
Data Storage Sustainability?

• Digital Data can be open – who should pay the cost?
• Spinning Disks, Bandwidth, Cooling, etc
No Silver Bullet - What is needed?

- Algorithms that scale
- Data Management from the Start
- Automatic Ancillary Data capture
- Thinking about the Data, and retention
- Data sharing is natural from the start
- Visualization for everyone
- Best practices – insights and challenges shared amongst domains
  - Ie. eScience Workshop, etc
Challenges

• Balancing
  Data Acquisition | Bandwidth | Storage/Processing

• Cross Discipline Collaboration – Knowledge sharing
• The data deluge - How to manage and analyze information?
• New types of Scientists:
  • Data Collectors & Data Analysis
• Riding the commodity curve
• Technology/Computing in support of Science