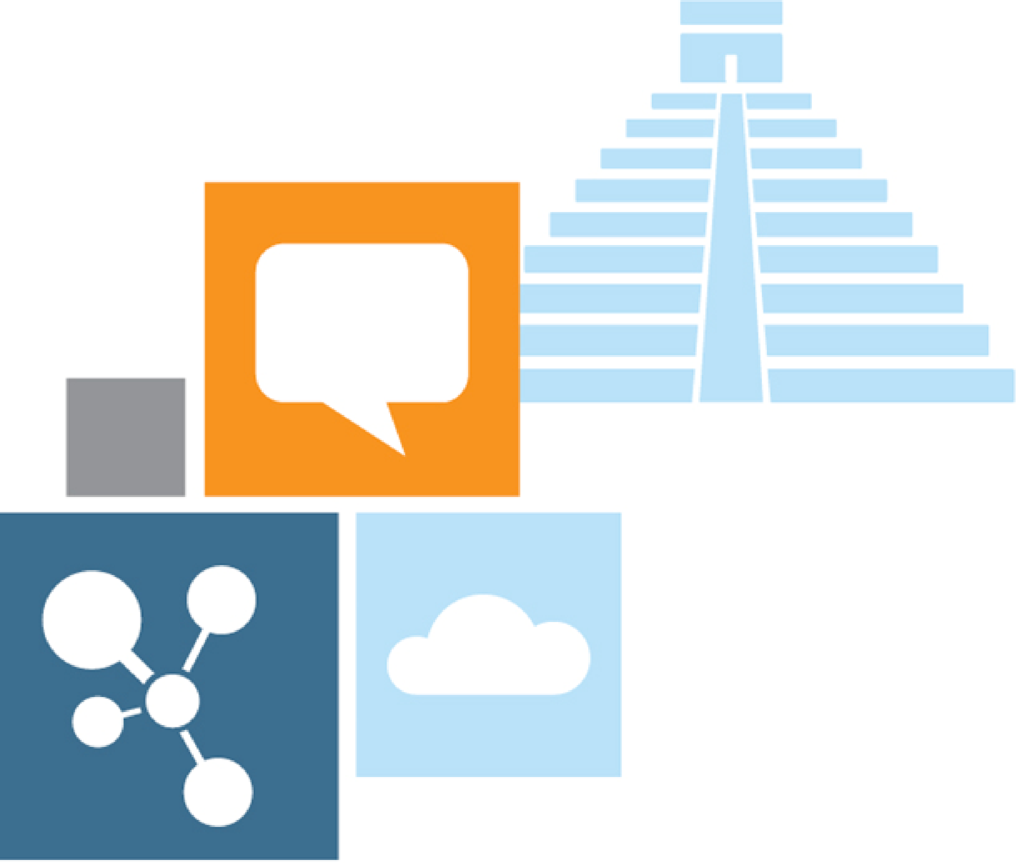
Microsoft



Microsoft® Research

Faculty Summit 2012

Riviera Maya, Mexico | May 23-25 | In partnership with CONACYT

Advancing Environmental Understanding: the Role of eScience

Dan Fay
Director – Earth, Energy and Environment
dan.fay@microsoft.com

MSR eScience Workshop 2011

Looking Back 8 yrs to the Beginning

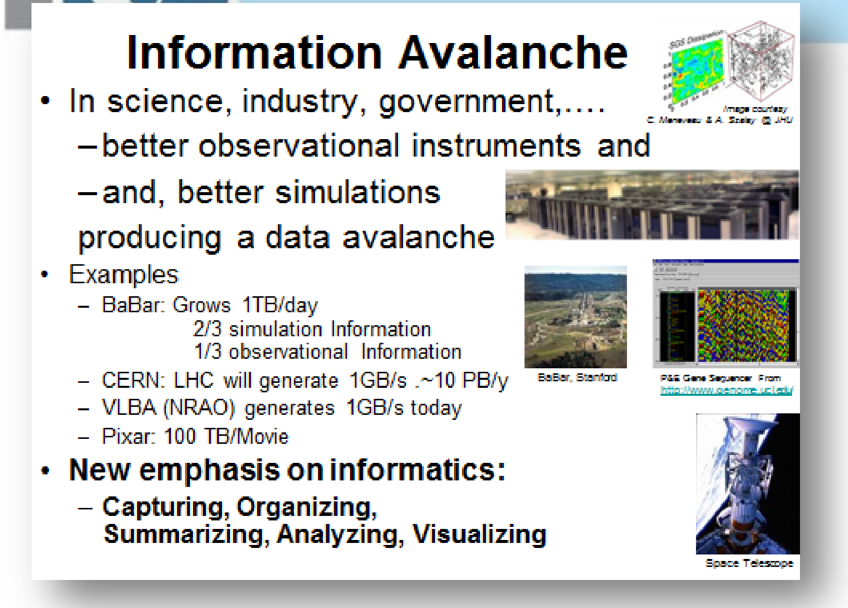
Scientific Data Intensive Computing Workshop 2004

- 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
  - Pixar: 100 TB/Movie
- **New emphasis on informatics:**
  - **Capturing, Organizing, Summarizing, Analyzing, Visualizing**

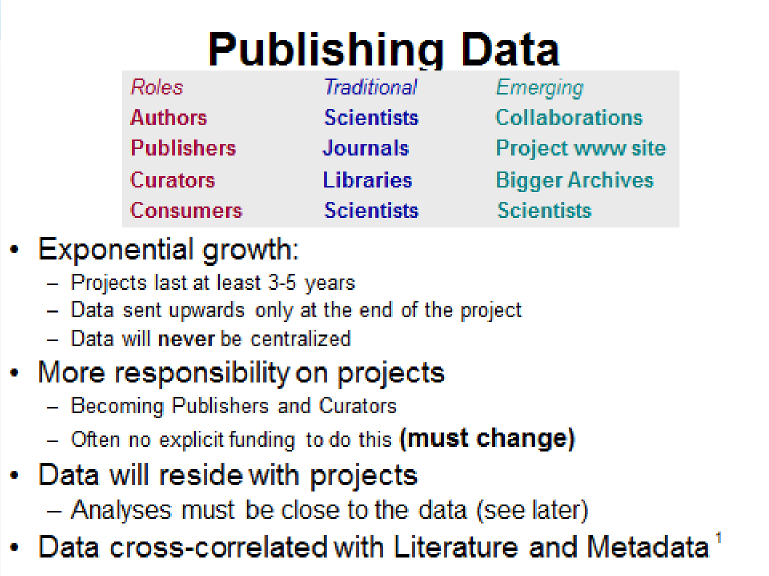
Image courtesy C. Meneveau & A. Szalay @ JHU

BaBar, Stanford

P&G Gene Sequencer From http://www.genome.uci.edu/

Space Telescope

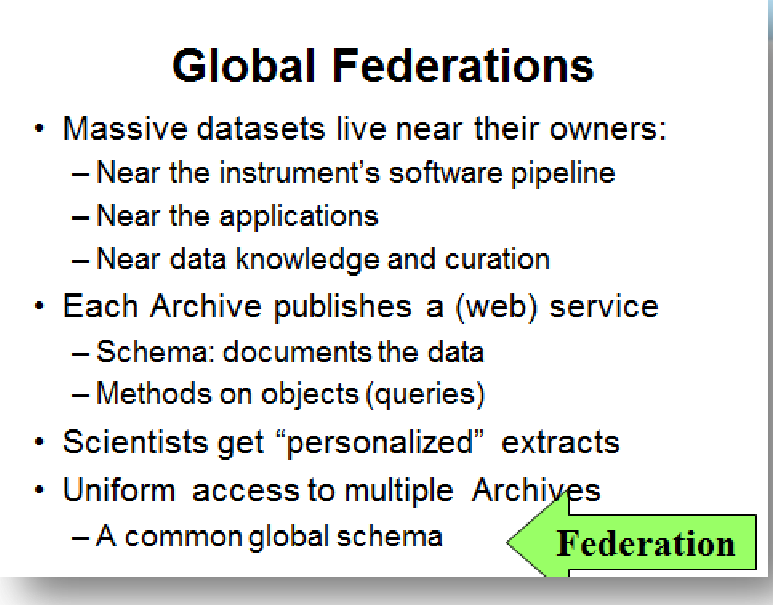
SGS Dissipation



Publishing Data

Roles	*Traditional*	*Emerging*
Authors	Scientists	Collaborations
Publishers	Journals	Project www site
Curators	Libraries	Bigger Archives
Consumers	Scientists	Scientists

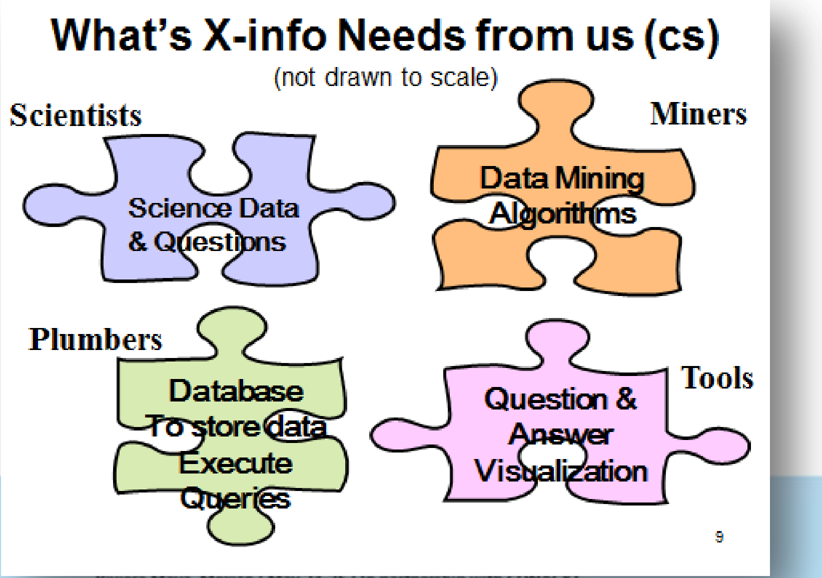
- Exponential growth:
  - Projects last at least 3-5 years
  - Data sent upwards only at the end of the project
  - Data will **never** be centralized
- More responsibility on projects
  - Becoming Publishers and Curators
  - Often no explicit funding to do this **(must change)**
- Data will reside with projects
  - Analyses must be close to the data (see later)
- Data cross-correlated with Literature and Metadata [1]



Global Federations

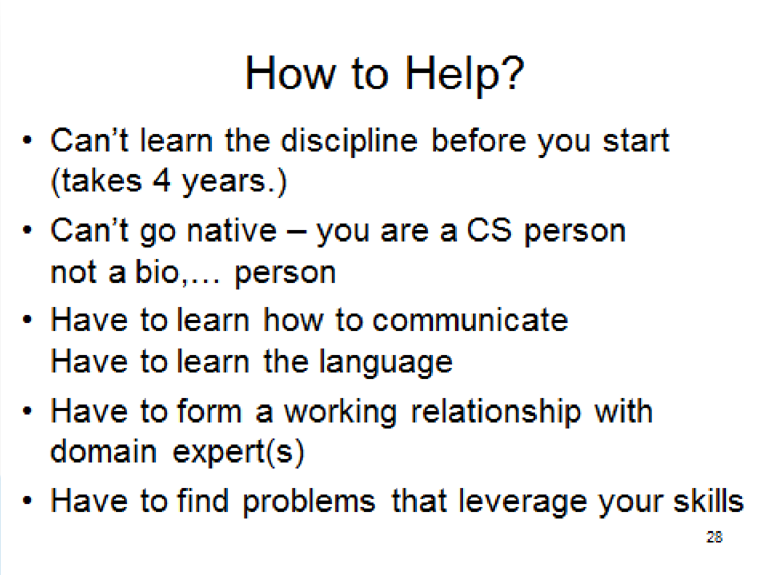
- Massive datasets live near their owners:
  - Near the instrument's software pipeline
  - Near the applications
  - Near data knowledge and curation
- Each Archive publishes a (web) service
  - Schema: documents the data
  - Methods on objects (queries)
- Scientists get "personalized" extracts
- Uniform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

Federation



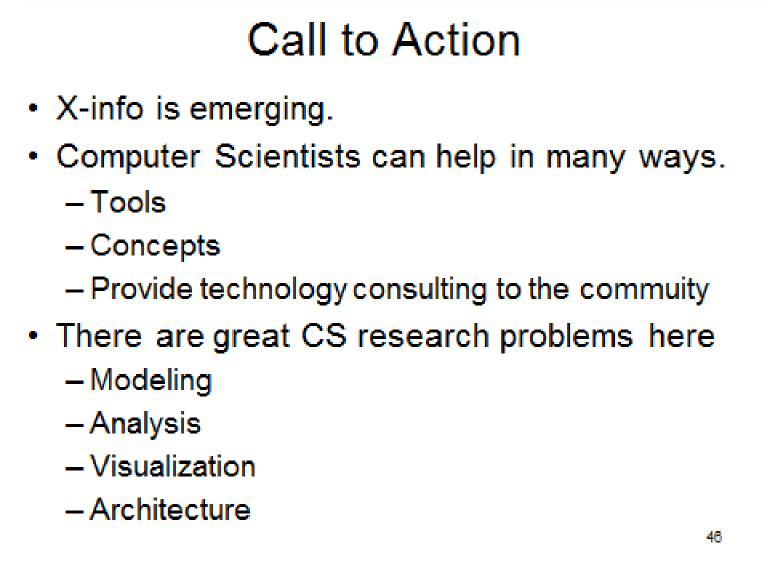
What's X-info Needs from us (cs)

(not drawn to scale)



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 commuity
- There are great CS research problems here
  - Modeling
  - Analysis
  - Visualization
  - Architecture

A Tidal Wave of Scientific Data

COMMUNICATIONS OF THE ACM

CACM.ACM.ORG 12/08 VOL.51 NO.12

Surviving the Data Deluge

The Economist

Obama the warrior

Misgoverning Argentina

The economic shift from West to Eas

Genetically modified crops blossom

The right to eat cats and dogs

FEBRUARY 27TH–MARCH 5TH 2010 Economist.com

The data deluge

TO HANDLE IT: A 14-PAGE SPECIAL REPORT

ieee spectrum INSIDE TECHNOLOGY

MAGAZINE MULTIM

AEROSPACE BIOMEDICAL COMPUTING CONSUMER ELECTRONICS ENERGY

AT WORK / INNOVATION

OPINION

The Coming Data Deluge

As science becomes more data intensive, so does our language

By PAUL MCFEDRIES / FEBRUARY 2011

✉ Email Print Share

Illustration: gluekit

WIRED SUBSCRIBE » SECTIONS » BLOGS » REVIEWS » VIDEO »

Sign In | RSS

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson ✉ 06.23.08

Illustration: Marian Bantjes

nature

THE BITER BIT
Viral infections for viruses

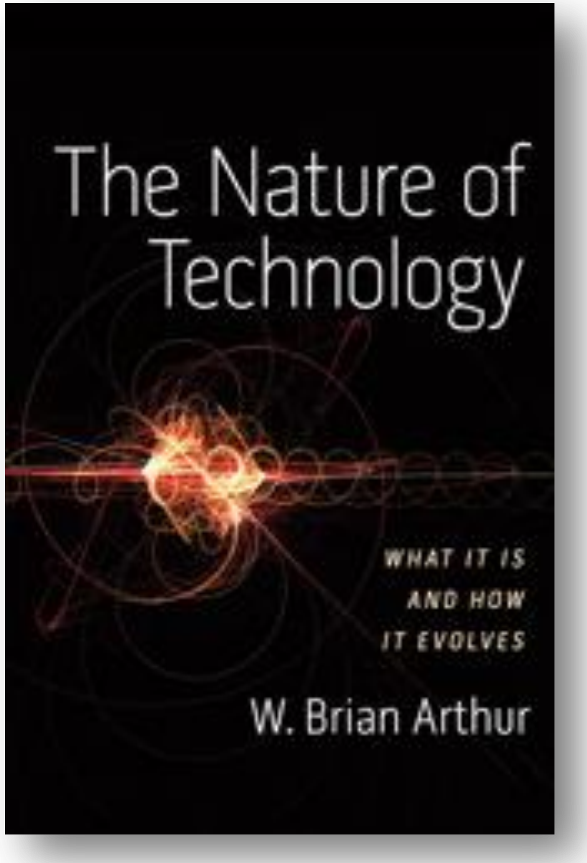
TROPICAL CYCLONES
The strong get stronger

BLACK HOLE PHYSICS
A new window on the Galactic Centre

BIG DATA

SCIENCE IN THE PETABYTE ERA

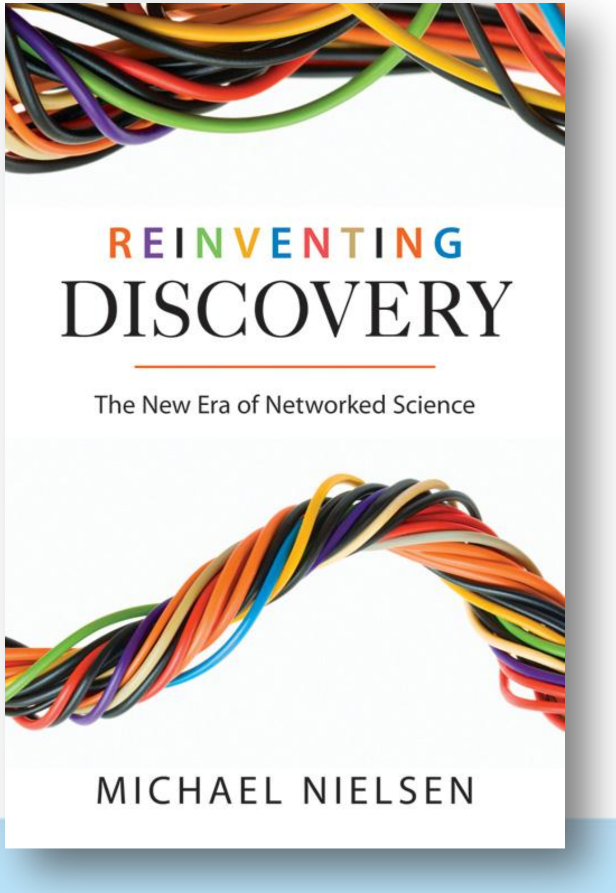
Interesting Thinking



The Nature of Technology

WHAT IT IS AND HOW IT EVOLVES

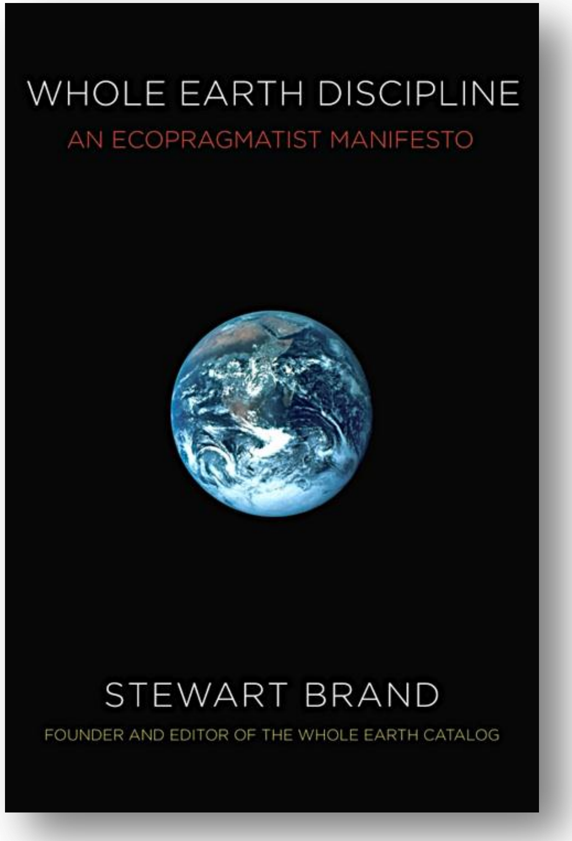
W. Brian Arthur



REINVENTING DISCOVERY

The New Era of Networked Science

MICHAEL NIELSEN



WHOLE EARTH DISCIPLINE

AN ECOPRAGMATIST MANIFESTO

STEWART BRAND

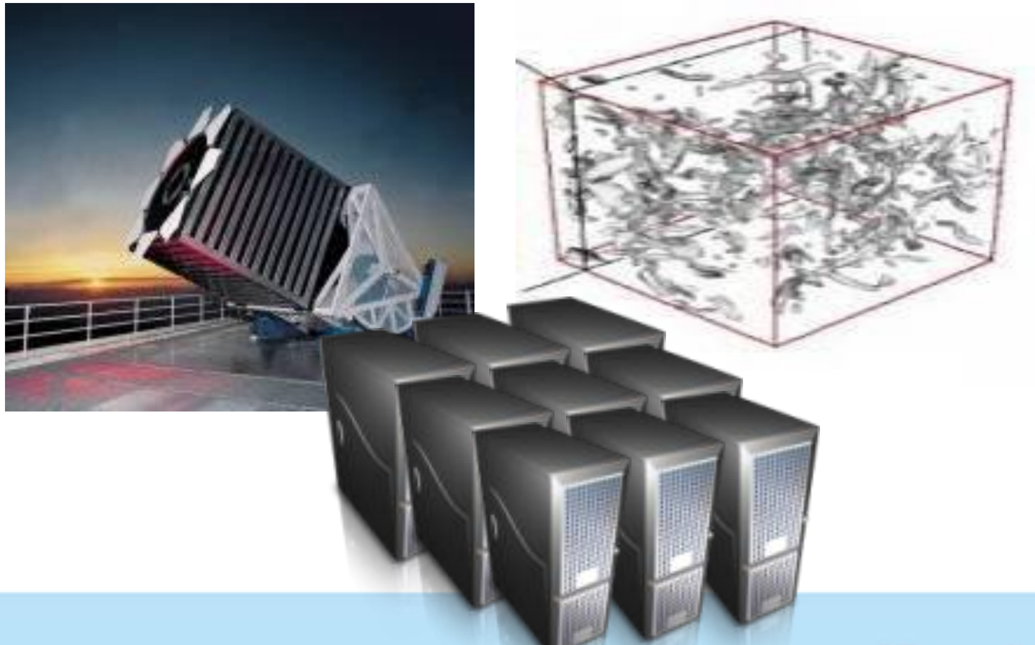
FOUNDER AND EDITOR OF THE WHOLE EARTH CATALOG

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

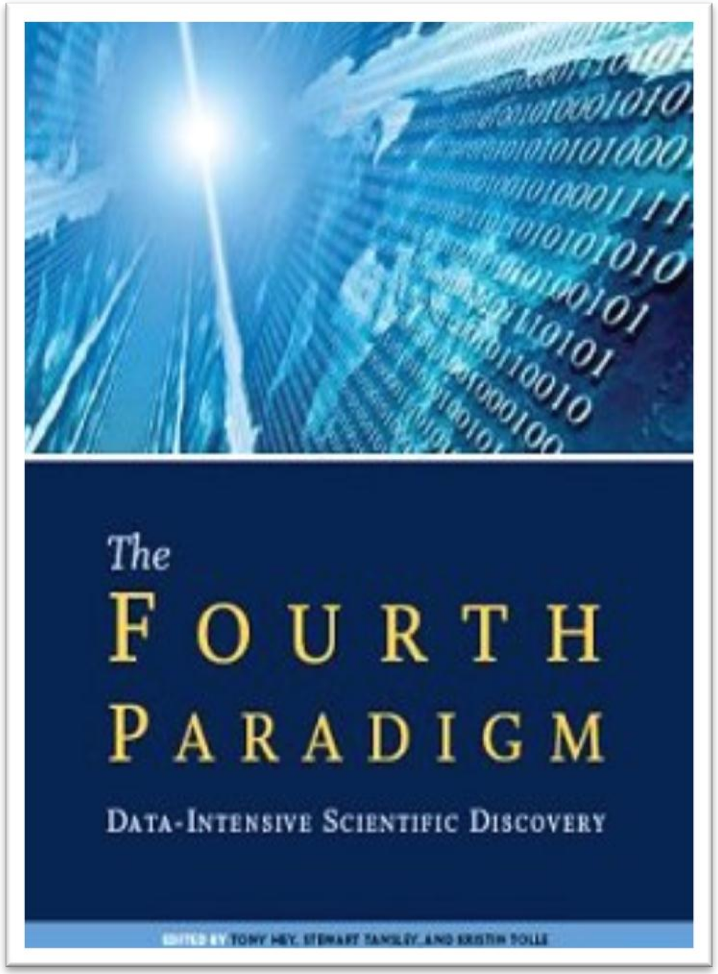
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\,\frac{c^2}{a^2}$$



Jim Gray 2007

Changing Nature of Discovery

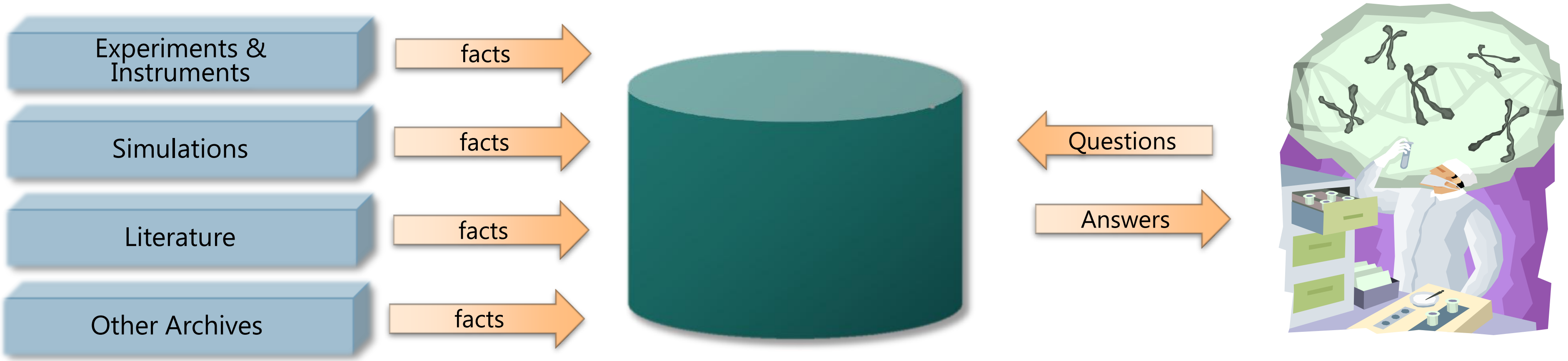
- Complex models
  - Multidisciplinary interactions
  - Wide temporal and spatial scales
- Large multidisciplinary data
  - Real-time steams
  - Structured and unstructured
- Distributed communities
  - Virtual organizations
  - Socialization and management
- Diverse expectations
  - Client-centric and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


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

http://fourthparadigm.org

The Problem for the e-Scientist

How to codify and represent our knowledge



Experiments & Instruments

Simulations

Literature

Other Archives

facts

facts

facts

facts

Questions

Answers

The Generic Problems

- Data ingest
- Managing a petabyte
- Common schema
- How to organize it
- How to *re*organize 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)

What is eScience?

Definition of eScience (Wikipedia)

E-Science (or **eScience**) is computationally intensive science that is carried out in highly distributed network environments, or science that uses immense data sets that require grid computing; the term sometimes includes technologies that enable distributed collaboration

Definition from Microsoft Research

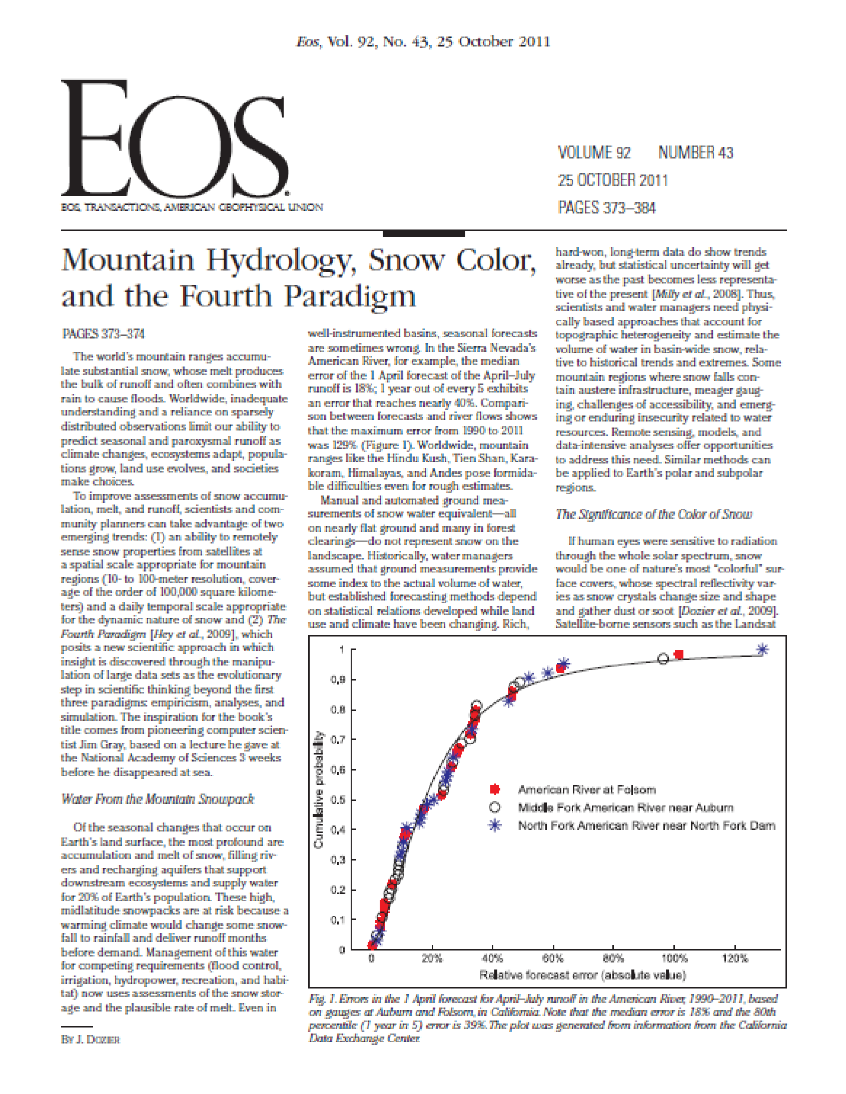
How computing technologies can help address scientific challenges
eScience efforts at Microsoft Research seek to further the understanding of these challenges, support the developing community, develop computational tools that will enable the advancement of scientific research, and catalyze discovery through funded collaborative research.

What it really means

How can current and future products and technologies can be applied to scientific challenges to help with scientific insight in a easy to use system

Technology in support of Science

EOS Article: *Mountain Hydrology, Snow Color, and the Fourth Paradigm* by Jeff Dozier



Eos, Vol. 92, No. 43, 25 October 2011

EOS
EOS, TRANSACTIONS, AMERICAN GEOPHYSICAL UNION

VOLUME 92 NUMBER 43
25 OCTOBER 2011
PAGES 373–384

Mountain Hydrology, Snow Color, and the Fourth Paradigm

PAGES 373–374

The world's mountain ranges accumulate substantial snow, whose melt produces the bulk of runoff and often combines with rain to cause floods. Worldwide, inadequate understanding and a reliance on sparsely distributed observations limit our ability to predict seasonal and paroxysmal runoff as climate changes, ecosystems adapt, populations grow, land use evolves, and societies make choices.

To improve assessments of snow accumulation, melt, and runoff, scientists and community planners can take advantage of two emerging trends: (1) an ability to remotely sense snow properties from satellites at a spatial scale appropriate for mountain regions (10- to 100-meter resolution, coverage of the order of 100,000 square kilometers) and a daily temporal scale appropriate for the dynamic nature of snow and (2) *The Fourth Paradigm* [*Hey et al.*, 2009], which posits a new scientific approach in which insight is discovered through the manipulation of large data sets as the evolutionary step in scientific thinking beyond the first three paradigms: empiricism, analyses, and simulation. The inspiration for the book's title comes from pioneering computer scientist Jim Gray, based on a lecture he gave at the National Academy of Sciences 3 weeks before he disappeared at sea.

Water From the Mountain Snowpack

Of the seasonal changes that occur on Earth's land surface, the most profound are accumulation and melt of snow, filling rivers and recharging aquifers that support downstream ecosystems and supply water for 20% of Earth's population. These high, midlatitude snowpacks are at risk because a warming climate would change some snowfall to rainfall and deliver runoff months before demand. Management of this water for competing requirements (flood control, irrigation, hydropower, recreation, and habitat) now uses assessments of the snow storage and the plausible rate of melt. Even in well-instrumented basins, seasonal forecasts are sometimes wrong. In the Sierra Nevada's American River, for example, the median error of the 1 April forecast of the April–July runoff is 18%; 1 year out of every 5 exhibits an error that reaches nearly 40%. Comparison between forecasts and river flows shows that the maximum error from 1990 to 2011 was 129% (Figure 1). Worldwide, mountain ranges like the Hindu Kush, Tien Shan, Karakoram, Himalayas, and Andes pose formidable difficulties even for rough estimates.

Manual and automated ground measurements of snow water equivalent—all on nearly flat ground and many in forest clearings—do not represent snow on the landscape. Historically, water managers assumed that ground measurements provide some index to the actual volume of water, but established forecasting methods depend on statistical relations developed while land use and climate have been changing. Rich, hard-won, long-term data do show trends already, but statistical uncertainty will get worse as the past becomes less representative of the present [*Milly et al.*, 2008]. Thus, scientists and water managers need physically based approaches that account for topographic heterogeneity and estimate the volume of water in basin-wide snow, relative to historical trends and extremes. Some mountain regions where snow falls contain austere infrastructure, meager gauging, challenges of accessibility, and emerging or enduring insecurity related to water resources. Remote sensing, models, and data-intensive analyses offer opportunities to address this need. Similar methods can be applied to Earth's polar and subpolar regions.

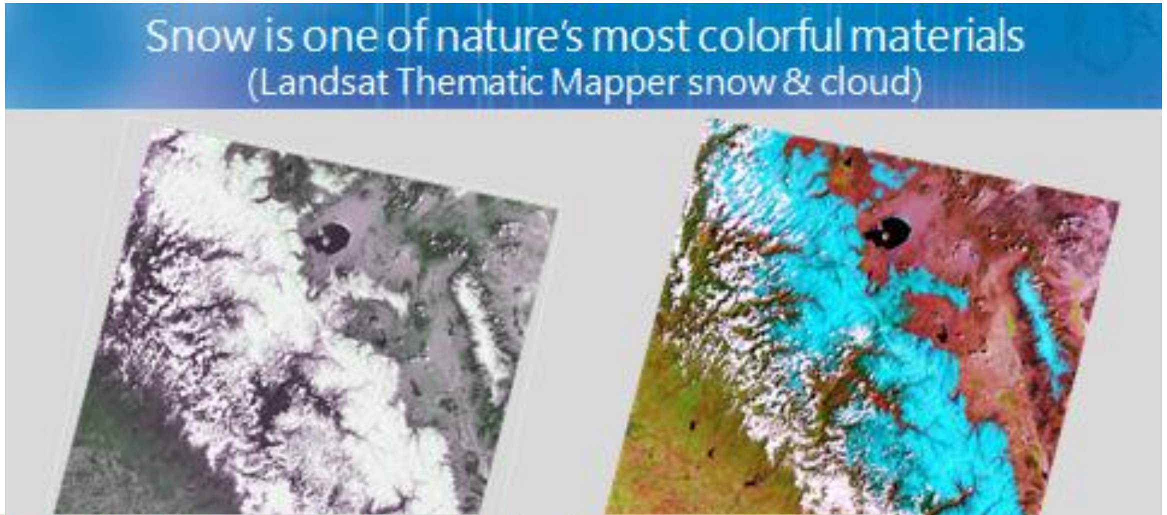
The Significance of the Color of Snow

If human eyes were sensitive to radiation through the whole solar spectrum, snow would be one of nature's most "colorful" surface covers, whose spectral reflectivity varies as snow crystals change size and shape and gather dust or soot [*Dozier et al.*, 2009]. Satellite-borne sensors such as the Landsat

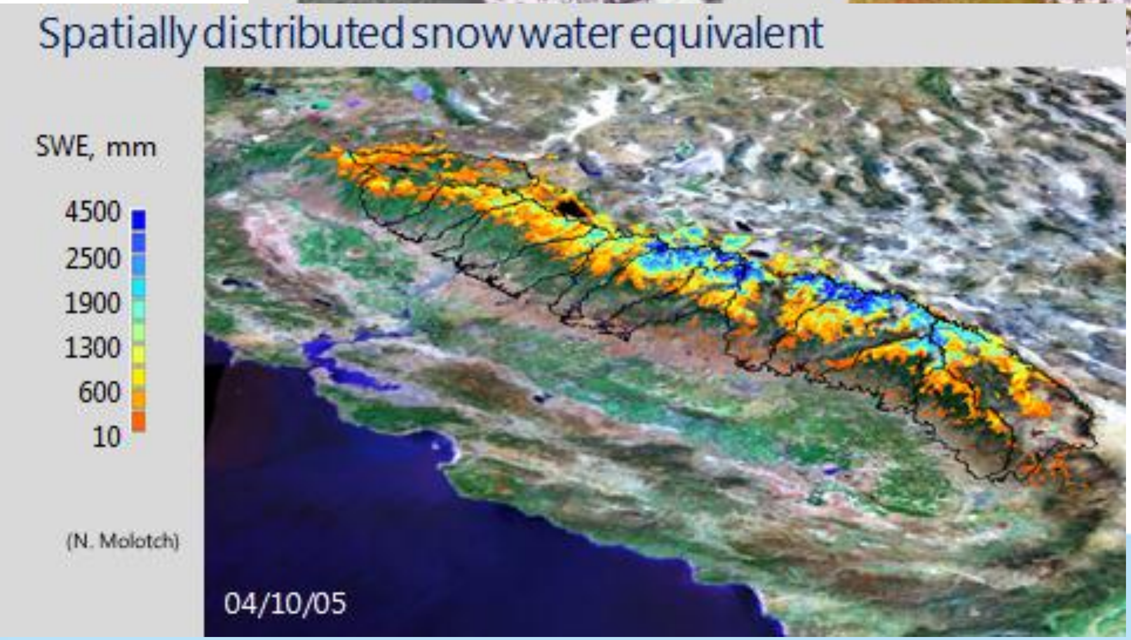
1
0,9
0.8
0,7
0,6
0.5
0,4
0,3
0.2
0,1
0
Cumulative probability
0
20%
40%
60%
80%
100%
120%
Relative forecast error (absolute value)
● American River at Folsom
○ Middle Fork American River near Auburn
✳ North Fork American River near North Fork Dam

Fig. 1. Errors in the 1 April forecast for April–July runoff in the American River, 1990–2011, based on gauges at Auburn and Folsom, in California. Note that the median error is 18% and the 80th percentile (1 year in 5) error is 39%. The plot was generated from information from the California Data Exchange Center.

By J. Dozier



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



Spatially distributed snow water equivalent

Information about water is more useful as we climb the value ladder

Data >>> Information >>> Insight

>>> Increasing value >>>

Forecasting

Reporting

Analysis

Integration

Done poorly,
but a few notable
counter-examples

Distribution

Aggregation

Done poorly to moderately,
not easy to find

Quality assurance

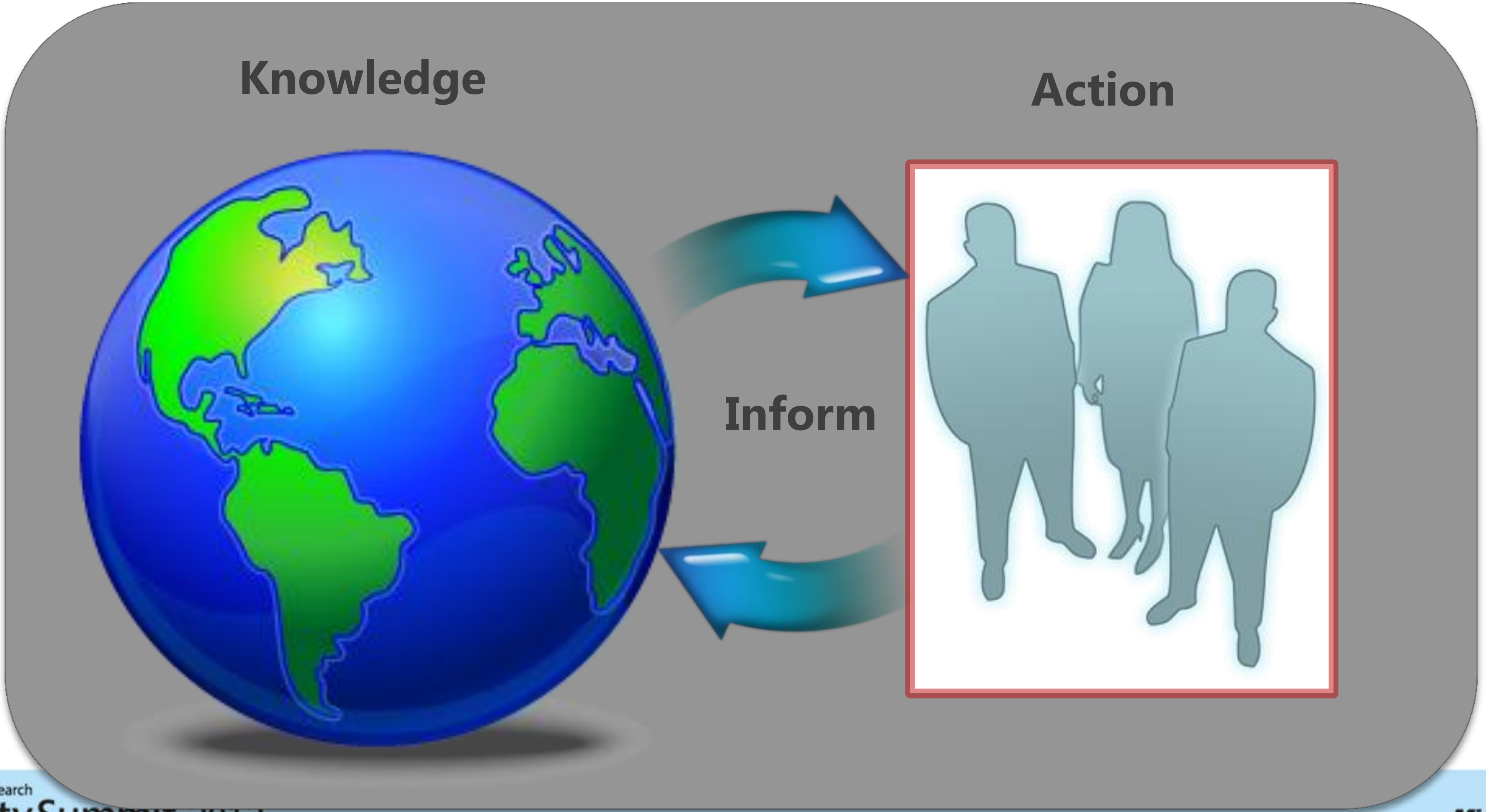
Collation

Monitoring

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

(I. Zaslavsky & CSIRO, BOM, WMO)

Environmental Ecosystem

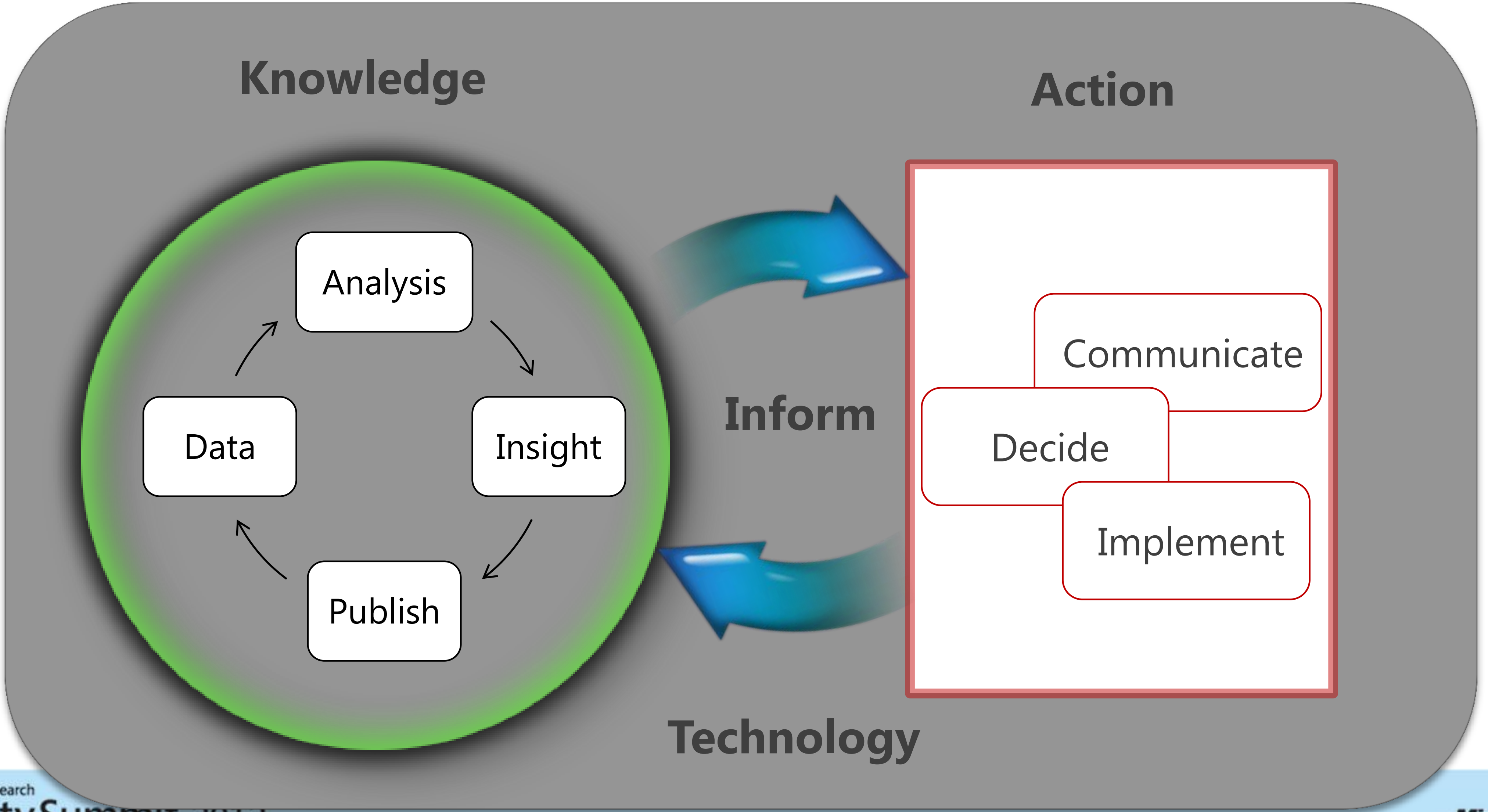


Knowledge

Action

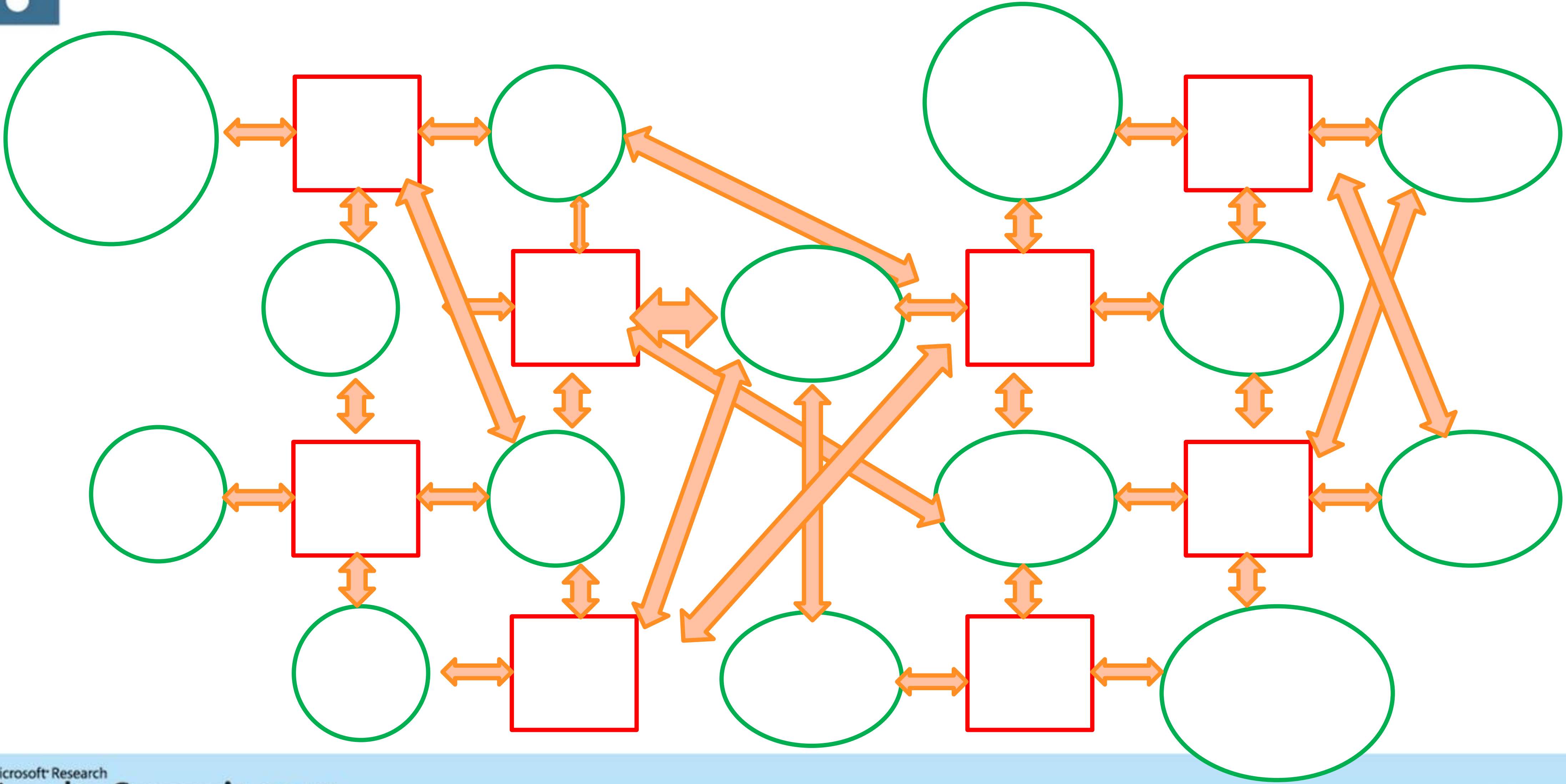
Inform

Environmental Ecosystem

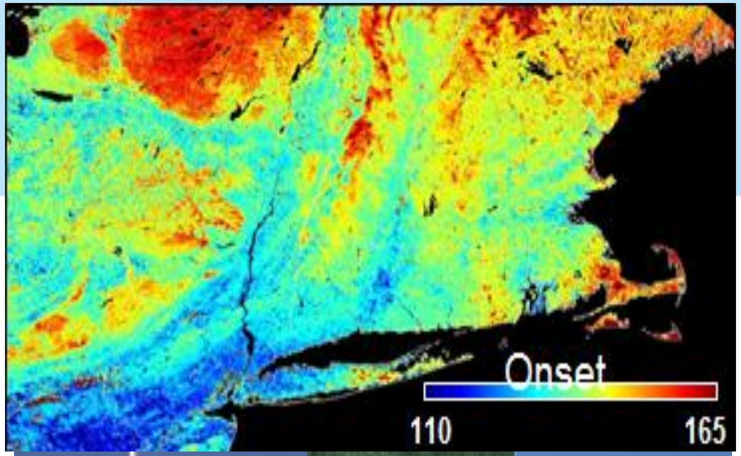


Information ecosystem:

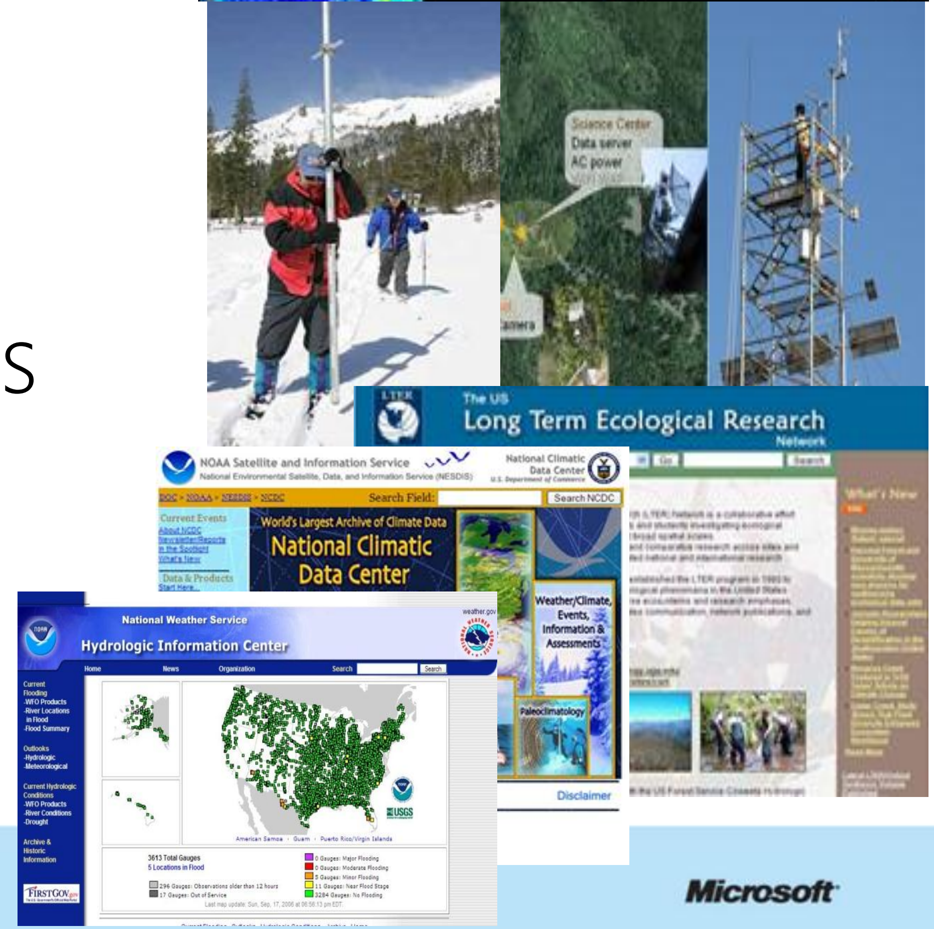
It is chaotic, unstructured and ad hoc



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

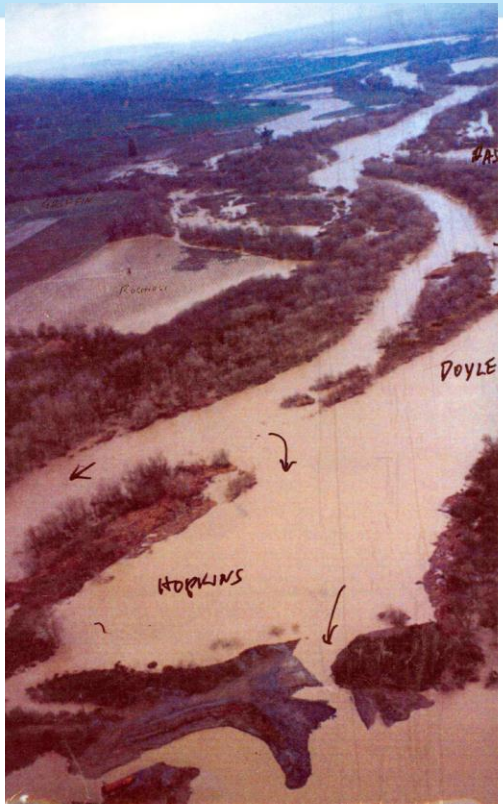

Science Center
Data server
AC power
The US
Long Term Ecological Research
Network
NOAA Satellite and Information Service
National Environmental Satellite, Data, and Information Service (NESDIS)
National Climatic Data Center
U.S. Department of Commerce
Search Field:
Search NCDC
World's Largest Archive of Climate Data
National Climatic Data Center
Weather/Climate Events, Information & Assessments
Paleoclimatology
Disclaimer
National Weather Service
Hydrologic Information Center
Home
News
Organization
Search
3613 Total Gauges
5 Locations in Flood
FIRSTGOV

Data Variety – The Spice of Life

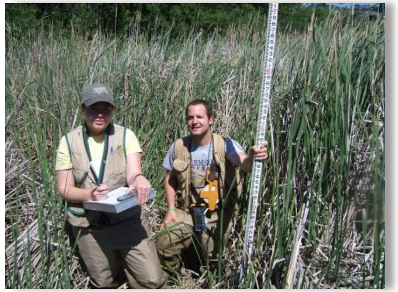
Manual Measurement

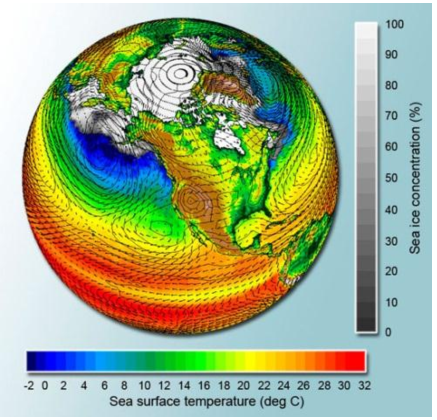
Automated Measurement

Sample Collection

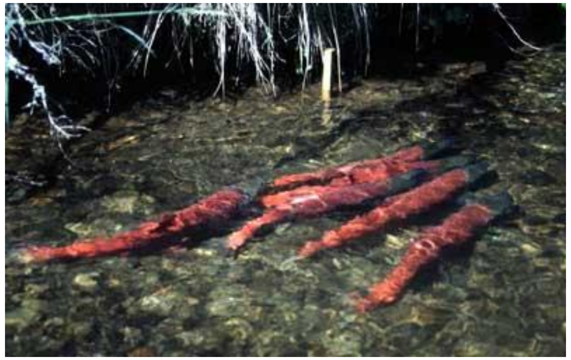


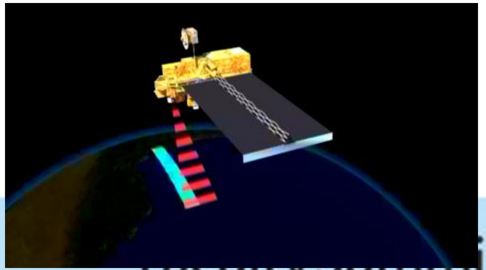
Historical Photographs


Typing



Model Output


Counting

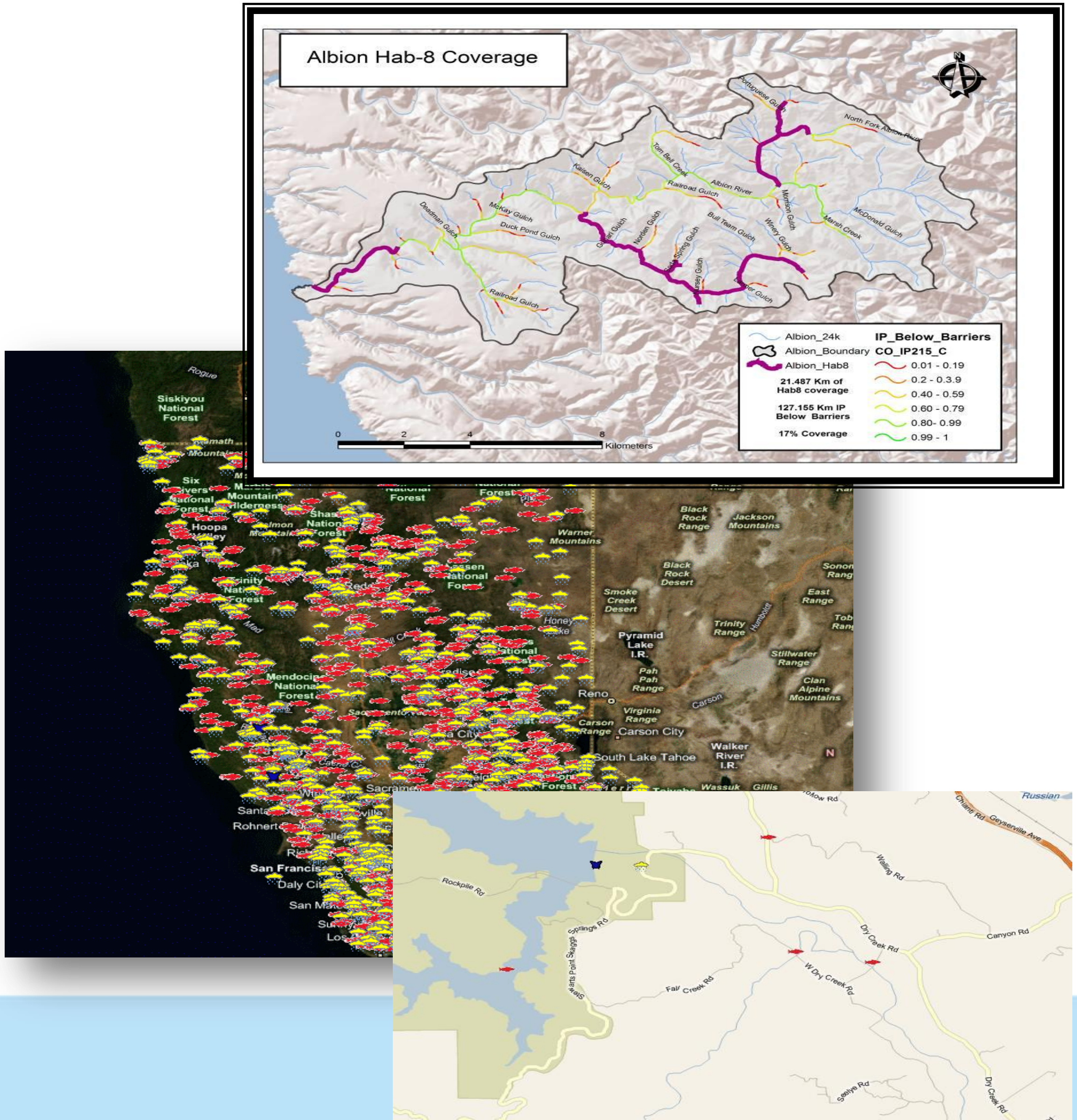

Satellite

Aircraft Surveys

Relatively Ubiquitous Motes

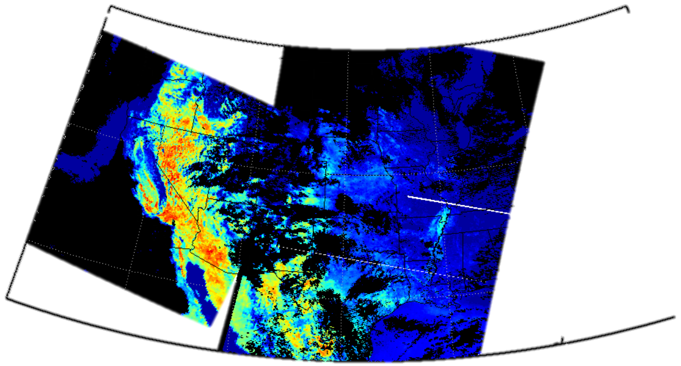
Data Integration Challenges

- 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

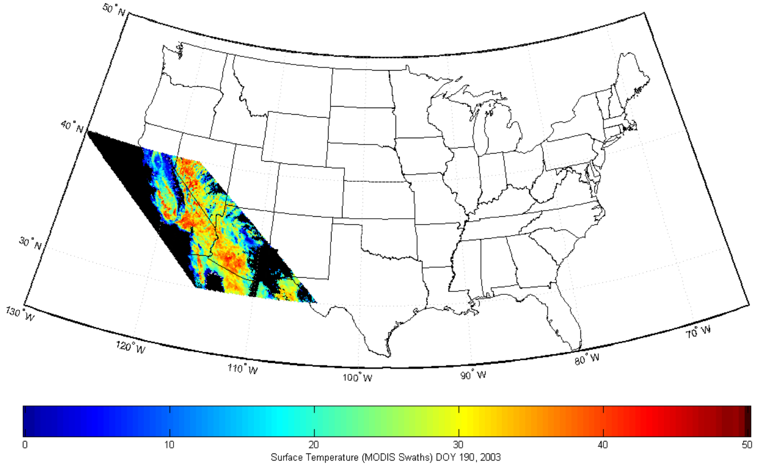


Tiling: Do Scientists Have to be Computer Scientists?

- Reprojection
  - Converts one geo-spatial representation to another.
  - Example is converting from latitude-longitude swaths to sinusoidal cells.
- Spatial resampling
  - Converts one spatial resolution to another.
  - Example is converting from 1 KM to 5 KB pixels.
- Temporal resampling
  - Converts one temporal resolution to another.
  - Example is converting from daily observation to 8 day averages.
- Gap filling
  - Assigns values to pixels without data either due to inherent data issues such as clouds or missing pixels introduced by one of the above.
- Masking
  - Eliminates uninteresting or unneeded pixels.
  - Examples are eliminating pixels over the ocean when computing a land product or eliminating pixels outside a spatial feature such as a watershed.

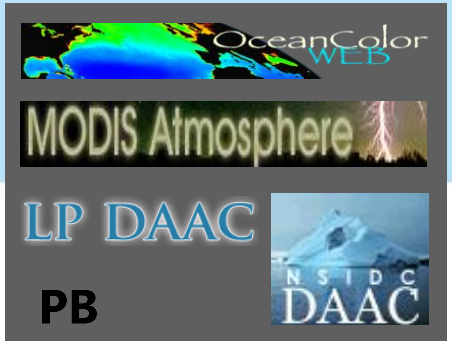


Source Data (Swath format)

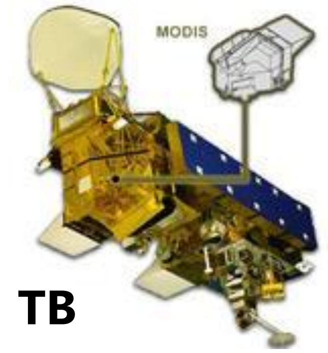


Reprojected Data (Sinusoidal format)

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)

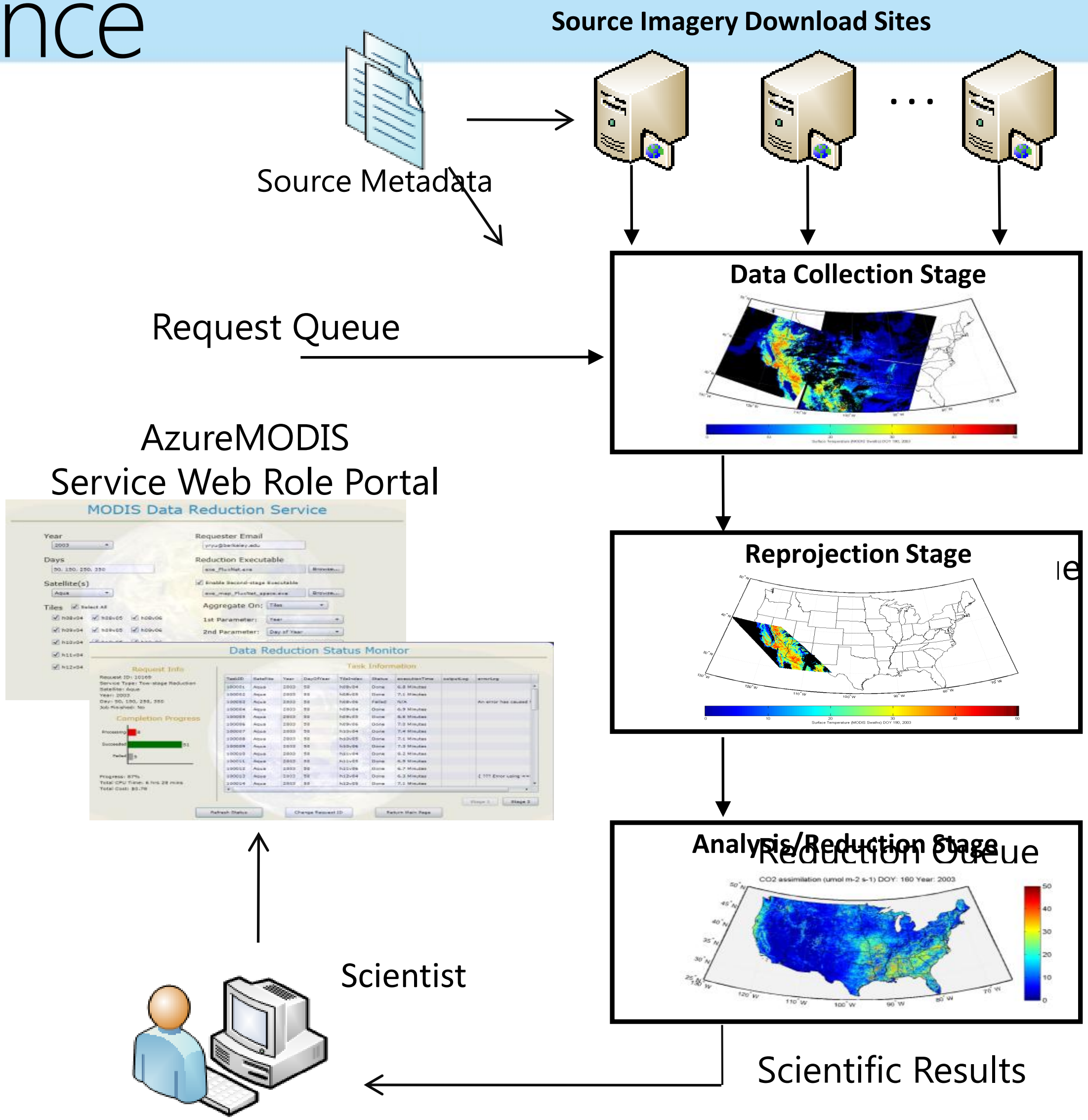
Science happens when PBs, TBs, GBs, and KBs can be mashed up simply

Complex and Heavy process by experts

Ad hoc observations and models

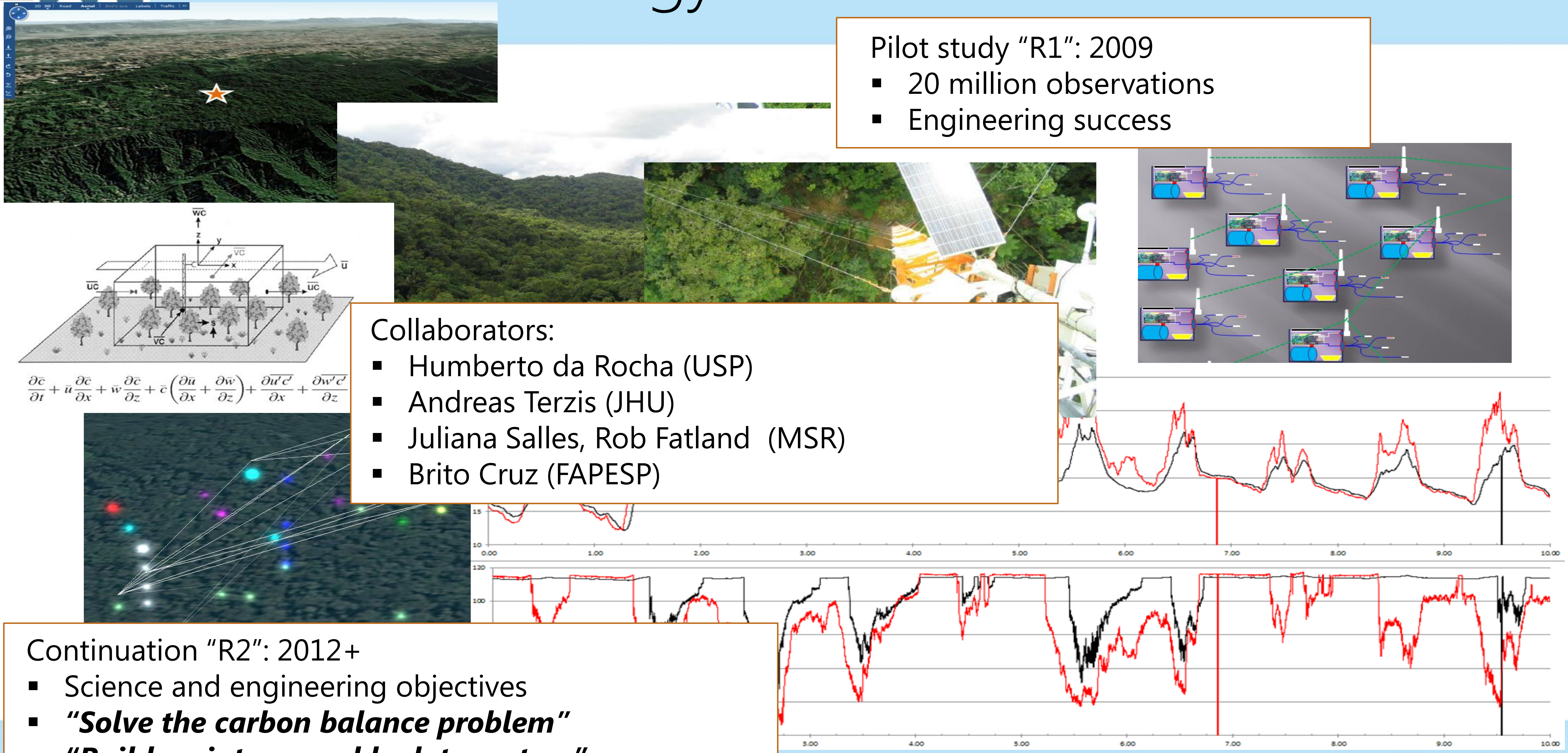
AzureMODIS – Azure Service for Remote Sensing Geoscience

- 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)

Micrometeorology



$$\frac{\partial \bar{c}}{\partial t} + \bar{u}\frac{\partial \bar{c}}{\partial x} + \bar{w}\frac{\partial \bar{c}}{\partial z} + \bar{c}\left(\frac{\partial \bar{u}}{\partial x} + \frac{\partial \bar{w}}{\partial z}\right) + \frac{\partial \overline{u'c'}}{\partial x} + \frac{\partial \overline{w'c'}}{\partial z}$$

Pilot study “R1”: 2009

- 20 million observations
- Engineering success

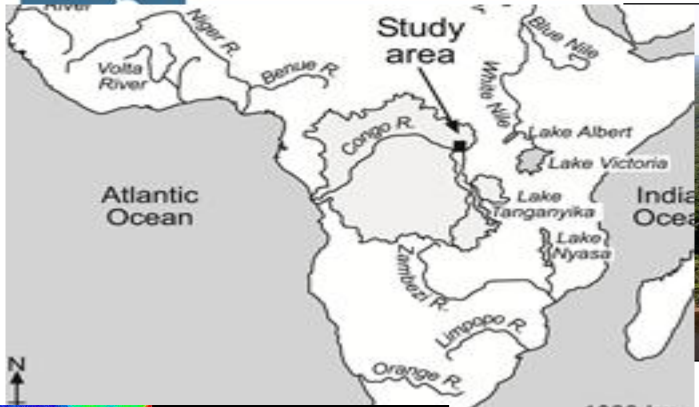
Collaborators:

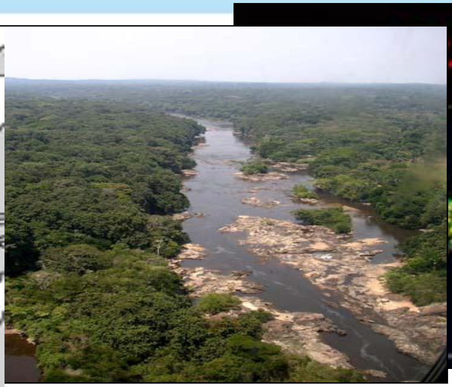
- Humberto da Rocha (USP)
- Andreas Terzis (JHU)
- Juliana Salles, Rob Fatland (MSR)
- Brito Cruz (FAPESP)

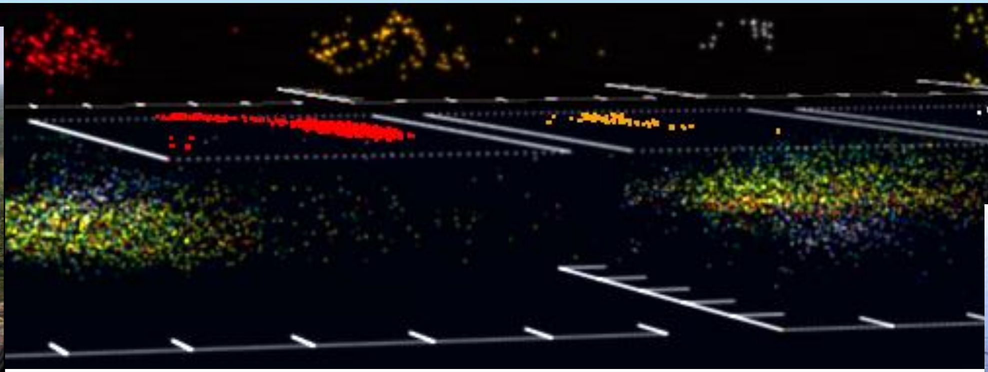
Continuation “R2”: 2012+

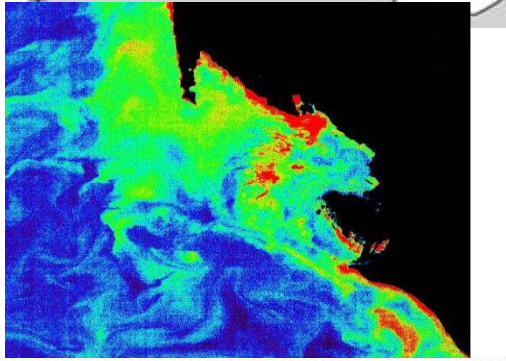
- Science and engineering objectives
- ***“Solve the carbon balance problem”***
- ***“Build an interoperable data system”***

Biogeochemistry: Carbon+





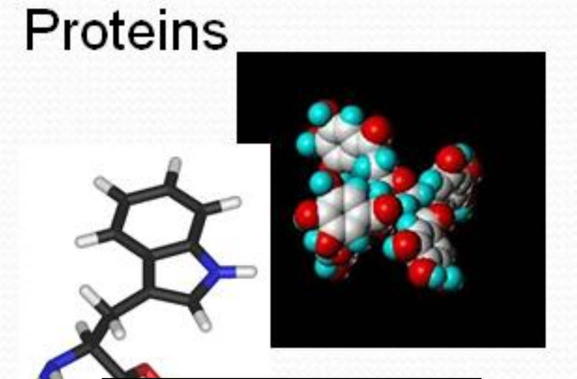


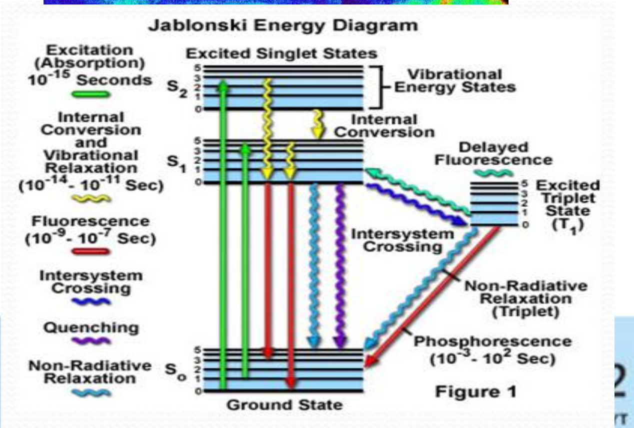


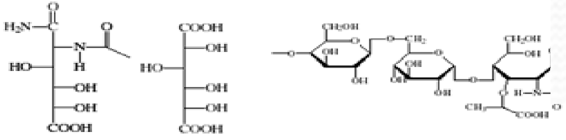
First Objective: Characterize fate of terrigenous carbon

- Multiple spectral analysis methods
- Data reduction: From correlation to machine learning

- Second objective: Library
  - Follow Environmental Information Framework
  - Contribute merit to Data Publishers
  - Address { curation, versioning, provenance }




Jablonski Energy Diagram
Excitation (Absorption) 10^-15 Seconds
Excited Singlet States
Vibrational Energy States
Internal Conversion and Vibrational Relaxation (10^-14 - 10^-11 Sec)
Internal Conversion
Delayed Fluorescence
Excited Triplet State (T1)
Fluorescence (10^-9 - 10^-7 Sec)
Intersystem Crossing
Non-Radiative Relaxation (Triplet)
Quenching
Phosphorescence (10^-3 - 10^2 Sec)
Non-Radiative Relaxation
S2
S1
S0
Ground State
Figure 1



Discoverability

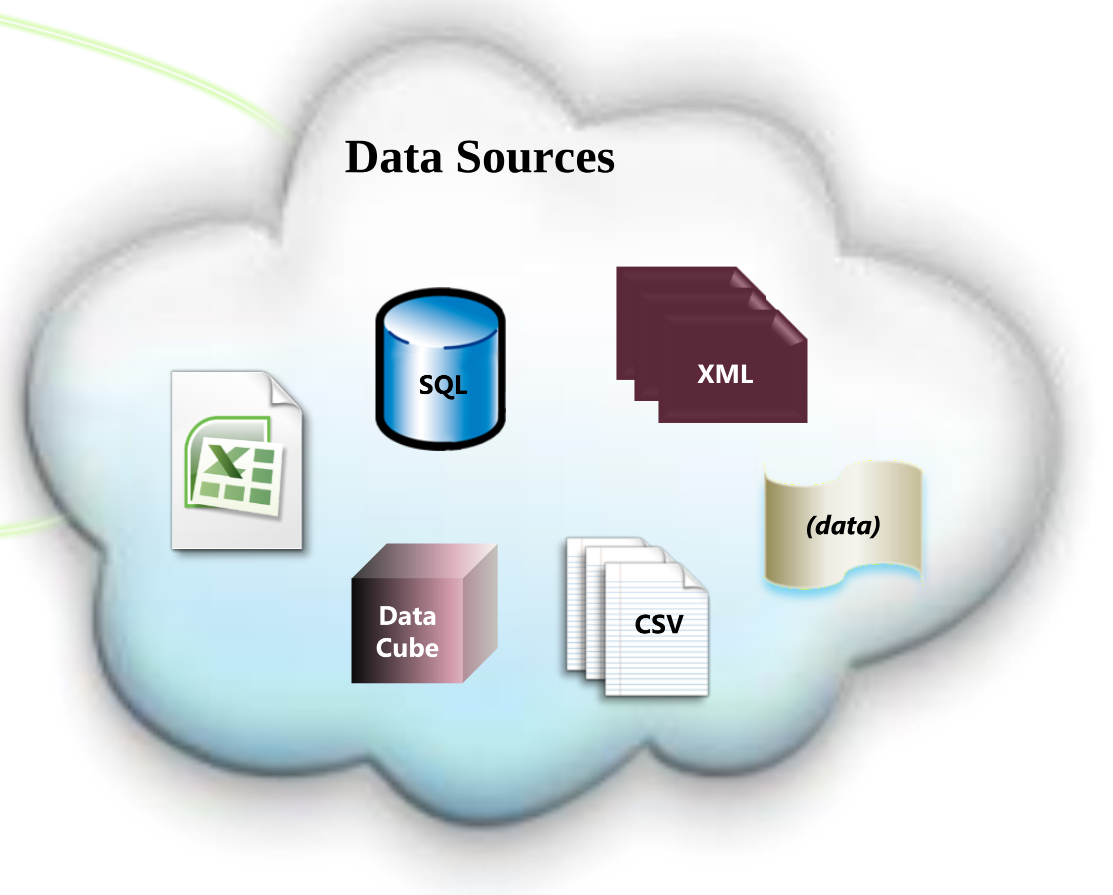
Accessibility

Consumeability

Environmental Informatics Framework (EIF)

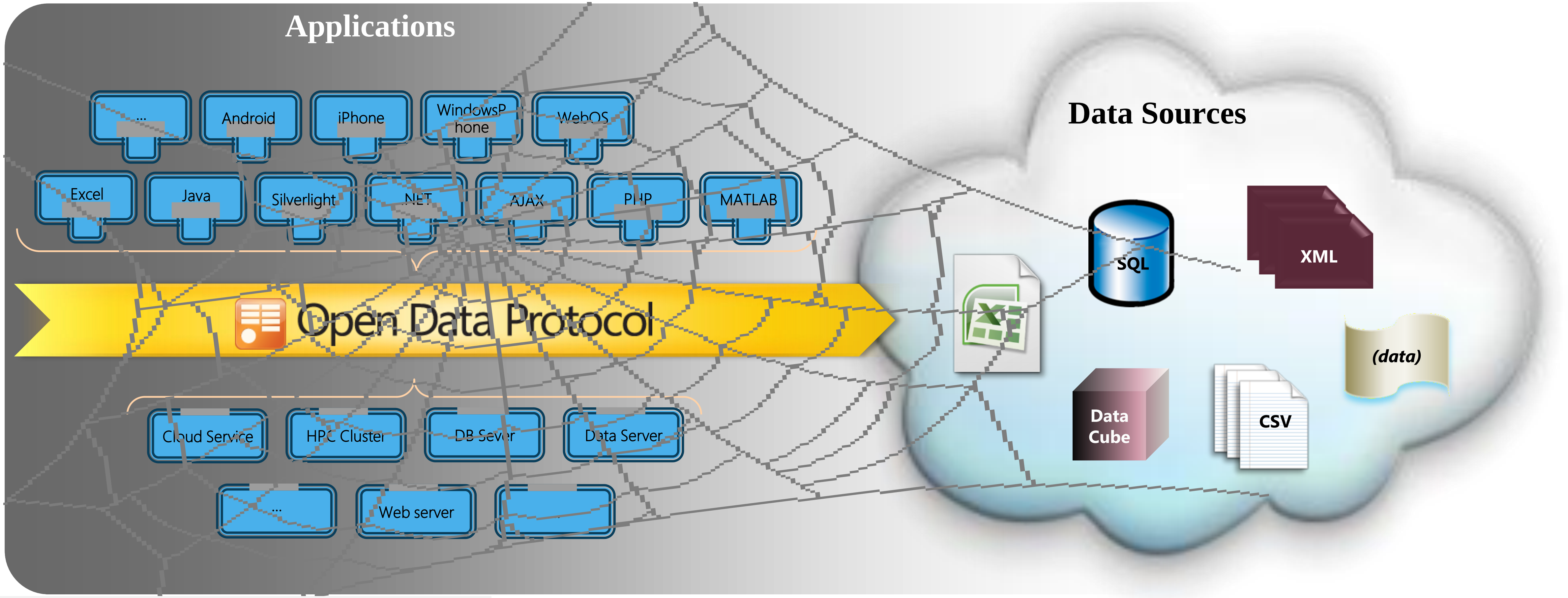
Common Problems with Data

- To use data from different sources
  - Non-standard formats, scales, and units
  - Lack of data quality control
  - Lack of metadata
  - Difficult to repurpose data for different (my) tools
- To share data
  - Lack of incentive (no credit)
  - Need extra resources and tools
- Hidden problems, seldom addressed
  - Versioning
  - Provenance
  - Curation



Environmental Informatics Framework (EIF)

Current State of Data Ecosystem



Environmental Informatics Framework (EIF)

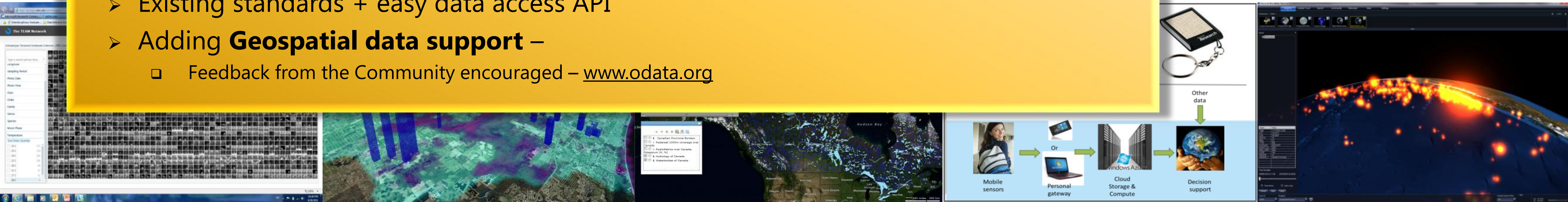
Advance data discoverability, accessibility, and consumability

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
- An application of a set of internet standards:
  - HTTP,
  - Atom (RFC 4287),
  - AtomPub (RFC 5023),
  - REST semantics
- Existing standards + easy data access API
- Adding **Geospatial data support** –
  - Feedback from the Community encouraged – www.odata.org



“Data Explorer”

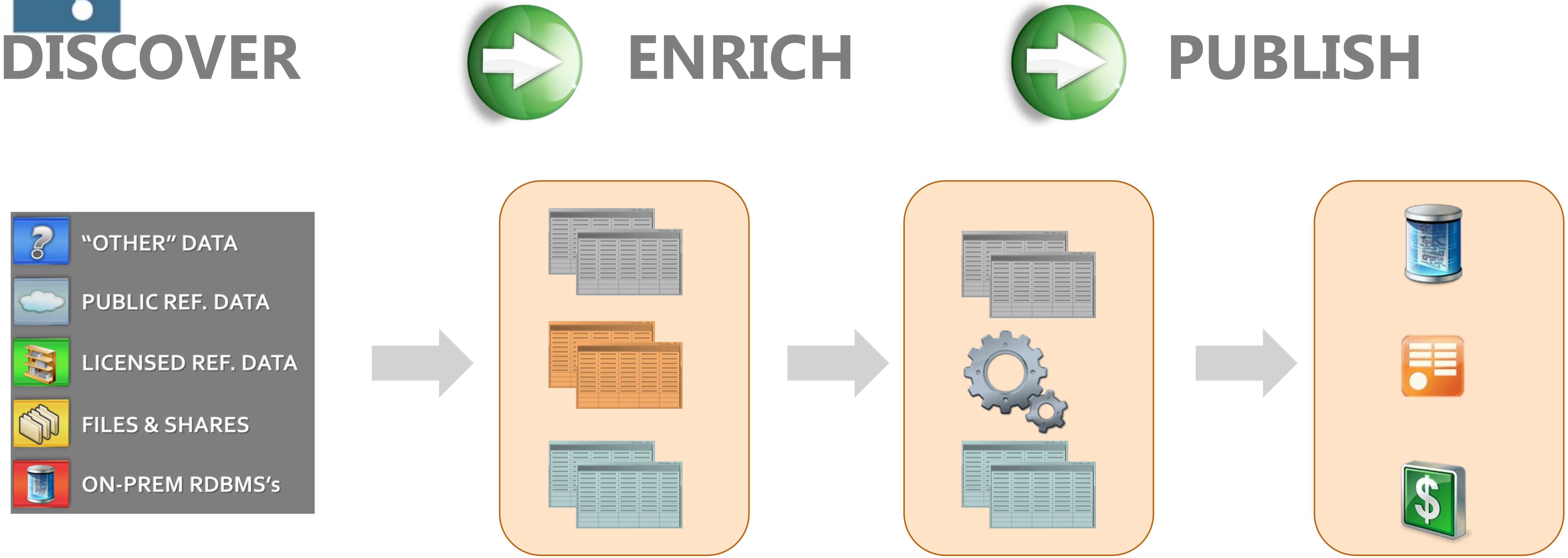
- A new SQL Azure Lab that provides a new way to organize, manage, mashup and gain new insights from your data.
- Data Explorer provides capabilities for data curation, collaboration, classification and mashup, opening new capabilities and opportunities around the data that you own or want to work with.

DISCOVER

ENRICH

PUBLISH

Microsoft Codename “Data Explorer” capabilities



DISCOVER

ENRICH

PUBLISH

“OTHER” DATA

PUBLIC REF. DATA

LICENSED REF. DATA

FILES & SHARES

ON-PREM RDBMS’s

Add & Manage
Data Sources

Classify
Understand
Recommend

Transform
Mash up
Cleanse

Snapshot
Publish
Sell

Data Tsunami or Scientific Data Hording

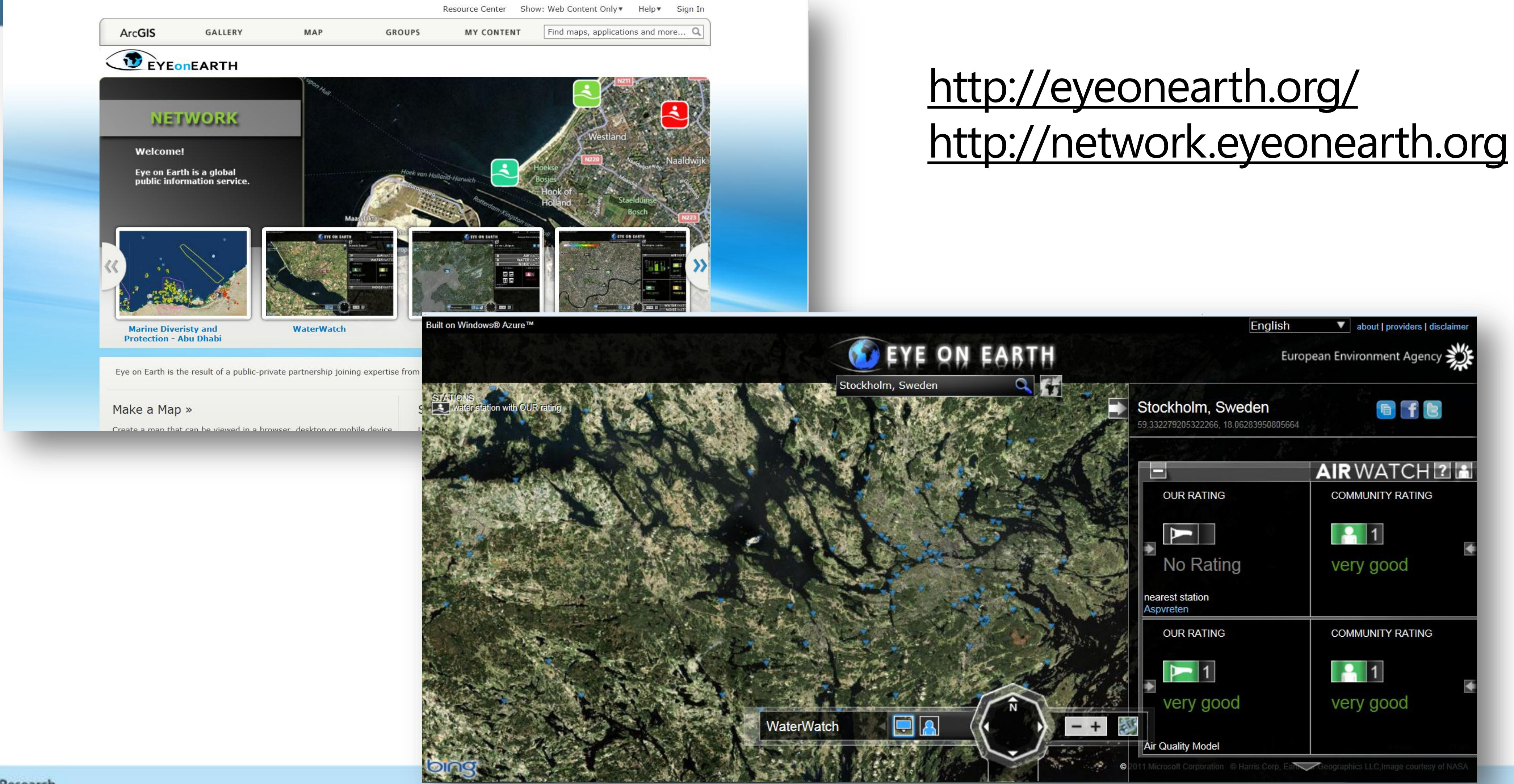
- Technology Trends
  - Compute and Storage make it easy to Run and Keep
  - Does it get used? Could it be mined?

- Data Collaboration reuse needs
- Discoverability
  - Catalogs, etc
- Accessibility
  - Protocols
- Consumability
  - Tooling support

New ways to analyze and communicate data

EYEonEARTH Network

http://eyeonearth.org/
http://network.eyeonearth.org



Resource Center Show: Web Content Only▾ Help▾ Sign In

ArcGIS GALLERY MAP GROUPS MY CONTENT Find maps, applications and more...

EYEonEARTH

NETWORK

Welcome!

Eye on Earth is a global public information service.

Marine Diveristy and Protection - Abu Dhabi

WaterWatch

Eye on Earth is the result of a public-private partnership joining expertise from

Make a Map »

Create a map that can be viewed in a browser, desktop or mobile device

Built on Windows® Azure™

English ▼ about | providers | disclaimer

EYE ON EARTH

European Environment Agency

Stockholm, Sweden

STATIONS
water station with OUR rating

Stockholm, Sweden
59.332279205322266, 18.06283950805664

AIR WATCH

OUR RATING	COMMUNITY RATING
No Rating	1 very good
nearest station Aspvreten	
1 very good	1 very good
Air Quality Model	

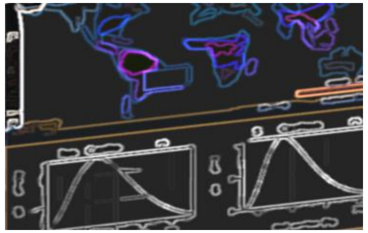
WaterWatch

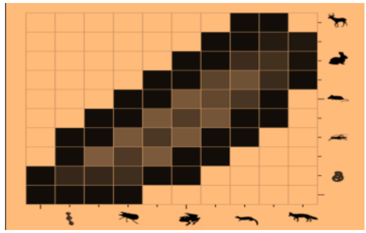
© 2011 Microsoft Corporation © Harris Corp, Earthstar Geographics LLC,Image courtesy of NASA

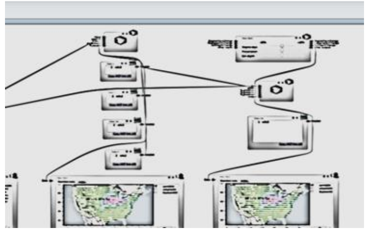
Computational Ecology and Environmental Science Group

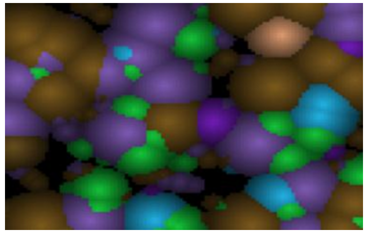
- Developing new concepts, methods and tools to enable better information and predictions about our planet
- Joining up theory, with data, via statistics, to produce useful predictive models
- Targeting predictive models at key policy requirements to help develop better informed solutions

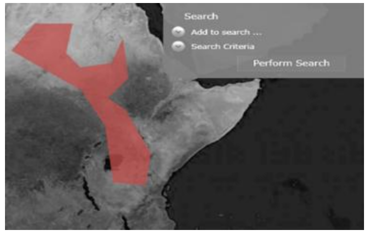
http://research.microsoft.com/ecology/


Global Carbon Model

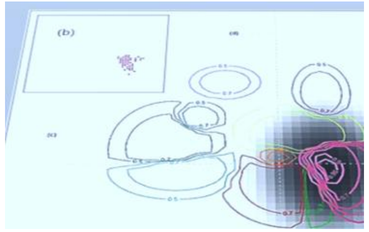

Ecosystem function


Food security

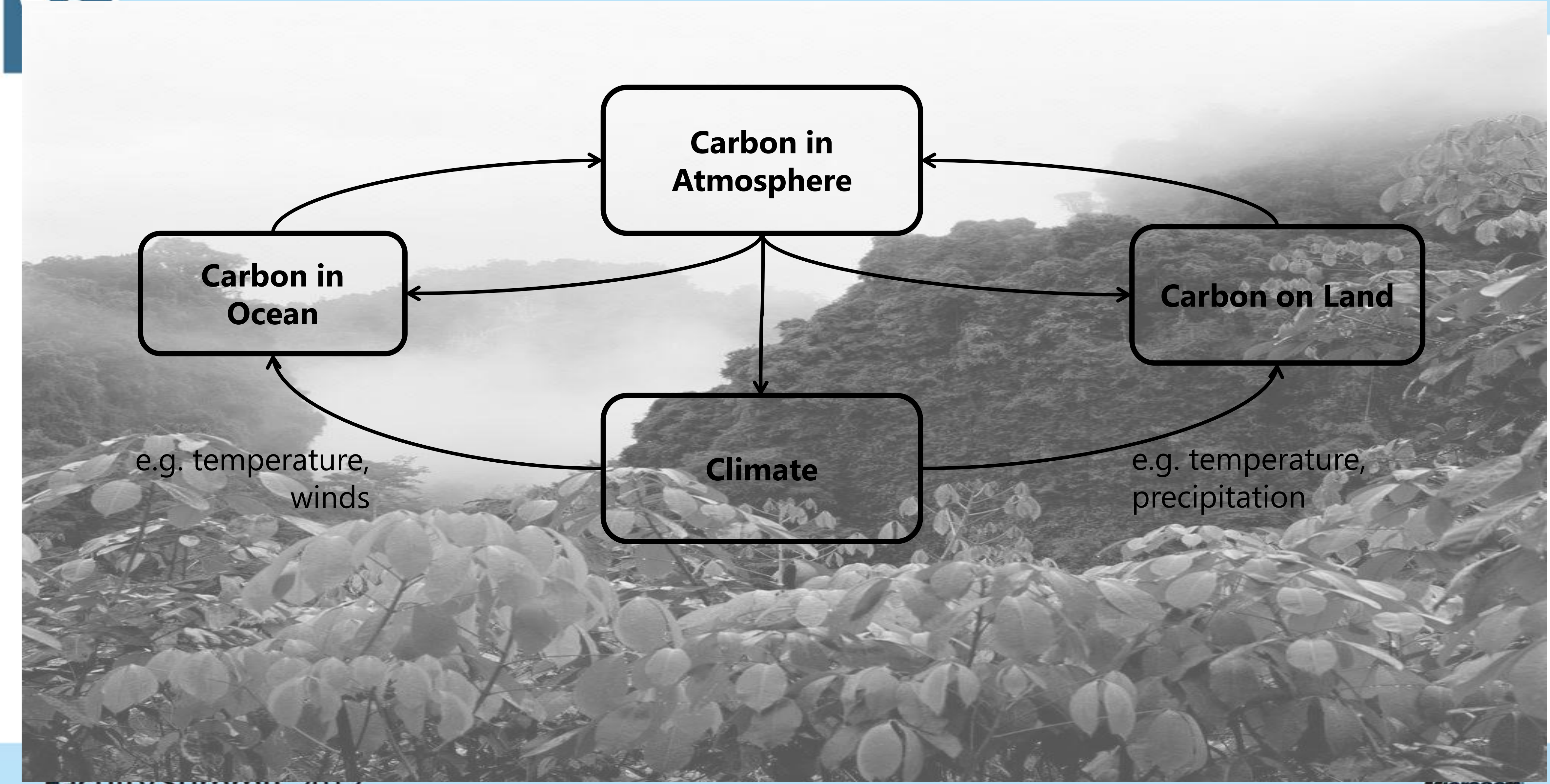

Forest dynamics



Threats to biodiversity


Species distributions

Global Carbon-Climate Feedback Model



Carbon in Atmosphere

Carbon in Ocean

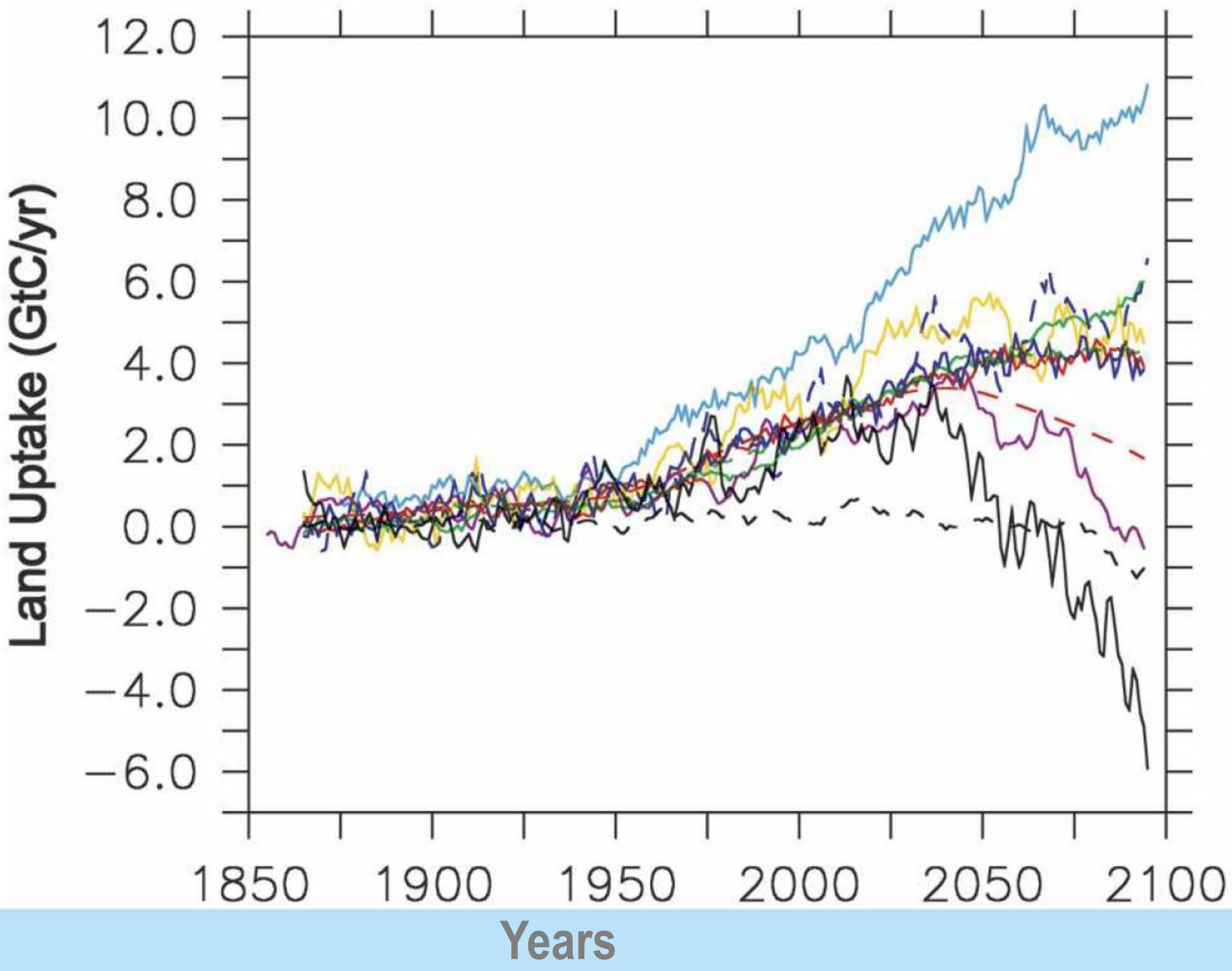
Carbon on Land

Climate

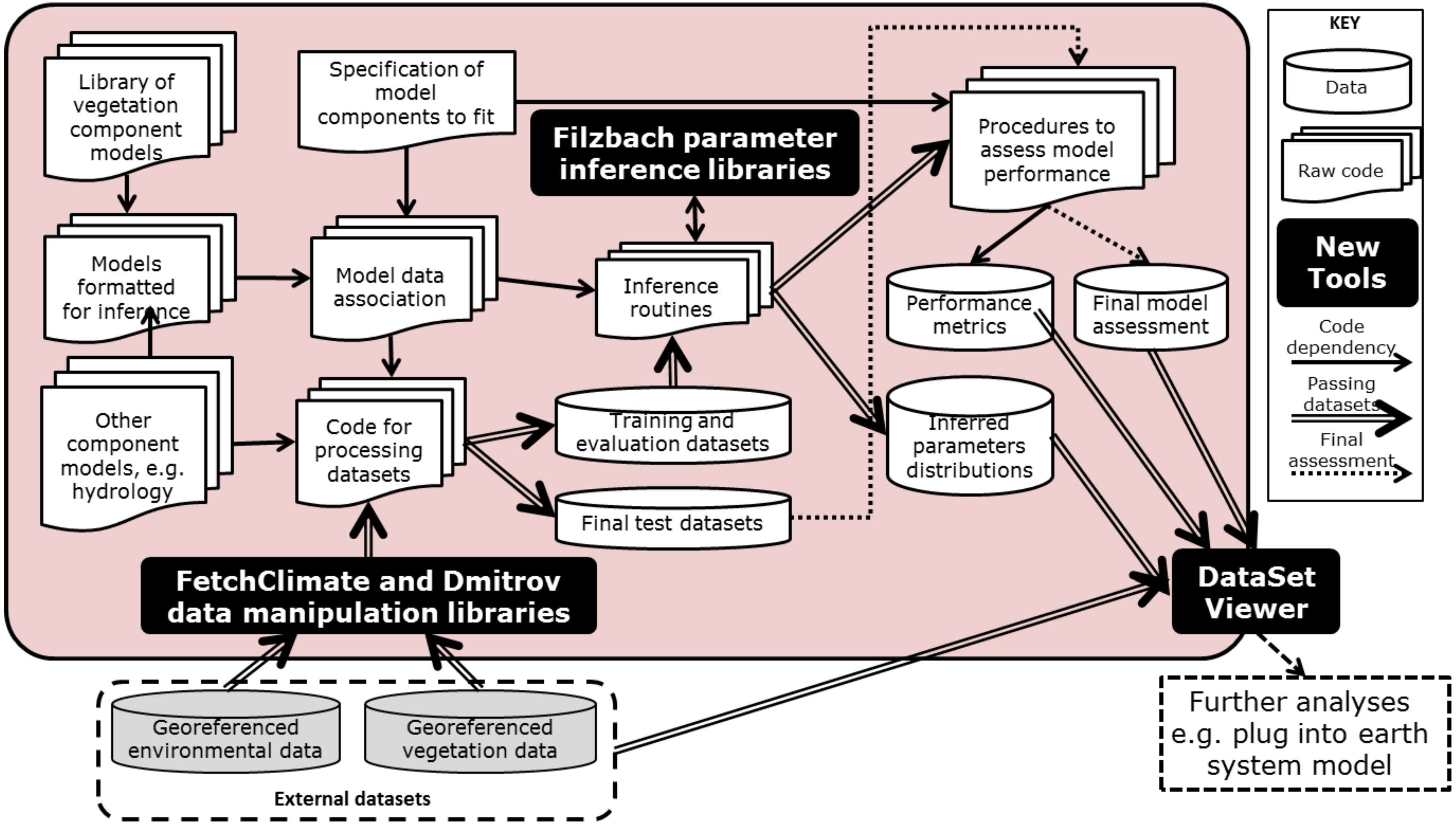
e.g. temperature, winds

e.g. temperature, precipitation

Current models are limited

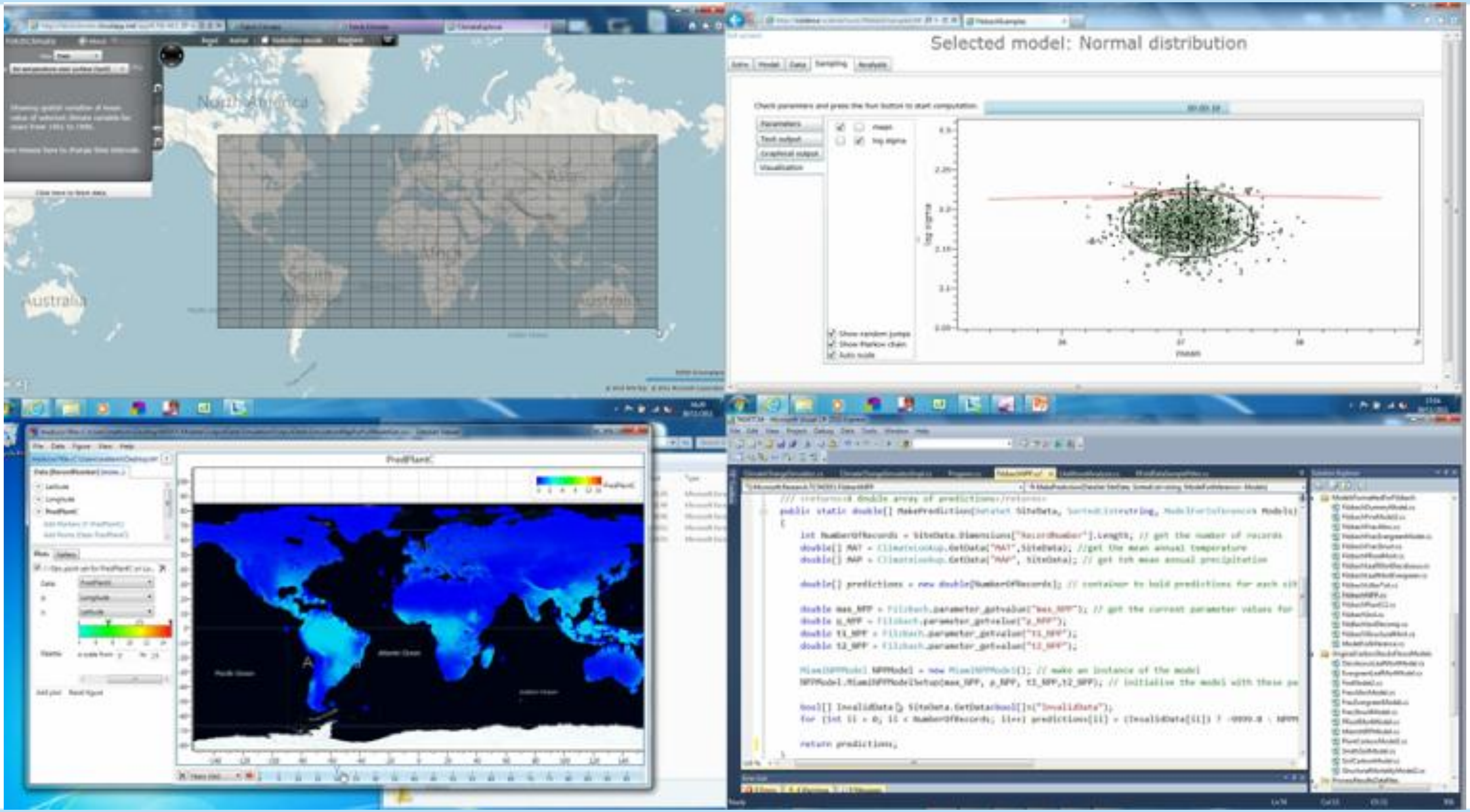


Multi-Component Model Framework



Library of vegetation component models
Specification of model components to fit
Filzbach parameter inference libraries
Procedures to assess model performance
Models formatted for inference
Model data association
Inference routines
Performance metrics
Final model assessment
Other component models, e.g. hydrology
Code for processing datasets
Training and evaluation datasets
Inferred parameters distributions
Final test datasets
FetchClimate and Dmitrov data manipulation libraries
DataSet Viewer
Georeferenced environmental data
Georeferenced vegetation data
External datasets
Further analyses e.g. plug into earth system model
KEY
Data
Raw code
New Tools
Code dependency
Passing datasets
Final assessment

Bringing Together Data for Models



```
/// <returns>A double array of predictions</returns>
public static double[] MakePrediction(DataSet SiteData, SortedList<string, ModelForInference> Models)
{
    int NumberOfRecords = SiteData.Dimensions["RecordNumber"].Length; // get the number of records
    double[] MAT = ClimateLookup.GetData("MAT",SiteData); //get the mean annual temperature
    double[] MAP = ClimateLookup.GetData("MAP", SiteData); // get teh mean annual precipitation

    double[] predictions = new double[NumberOfRecords]; // container to hold predictions for each sit

    double max_NPP = Filzbach.parameter_getvalue("max_NPP"); // get the current parameter values for
    double s_NPP = Filzbach.parameter_getvalue("s_NPP");
    double t1_NPP = Filzbach.parameter_getvalue("t1_NPP");
    double t2_NPP = Filzbach.parameter_getvalue("t2_NPP");

    MiamiNPPModel NPPModel = new MiamiNPPModel(); // make an instance of the model
    NPPModel.MiamiNPPModelSetup(max_NPP, s_NPP, t1_NPP,t2_NPP); // initialise the model with these pa

    bool[] InvalidData = SiteData.GetData<bool[]>("InvalidData");
    for (int ii = 0; ii < NumberOfRecords; ii++) predictions[ii] = (InvalidData[ii]) ? -9999.0 : NPPM

    return predictions;
}
```

http://research.microsoft.com/science/tools

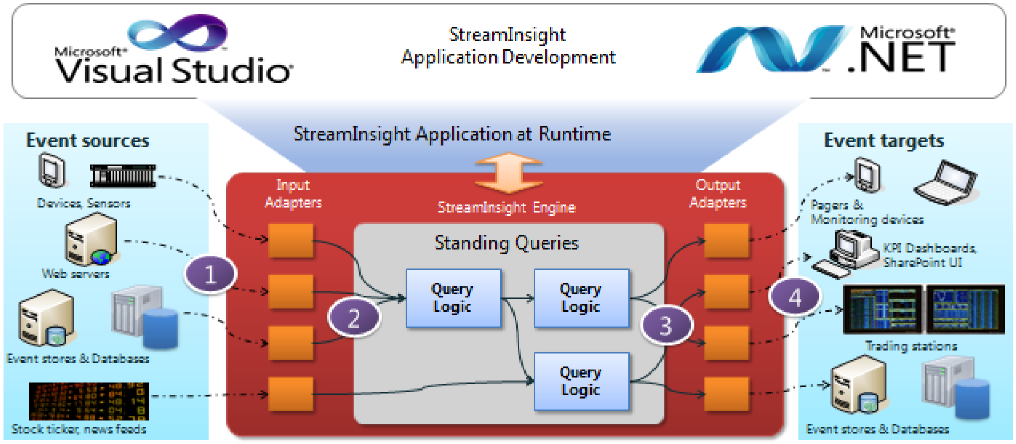
Complex event processing

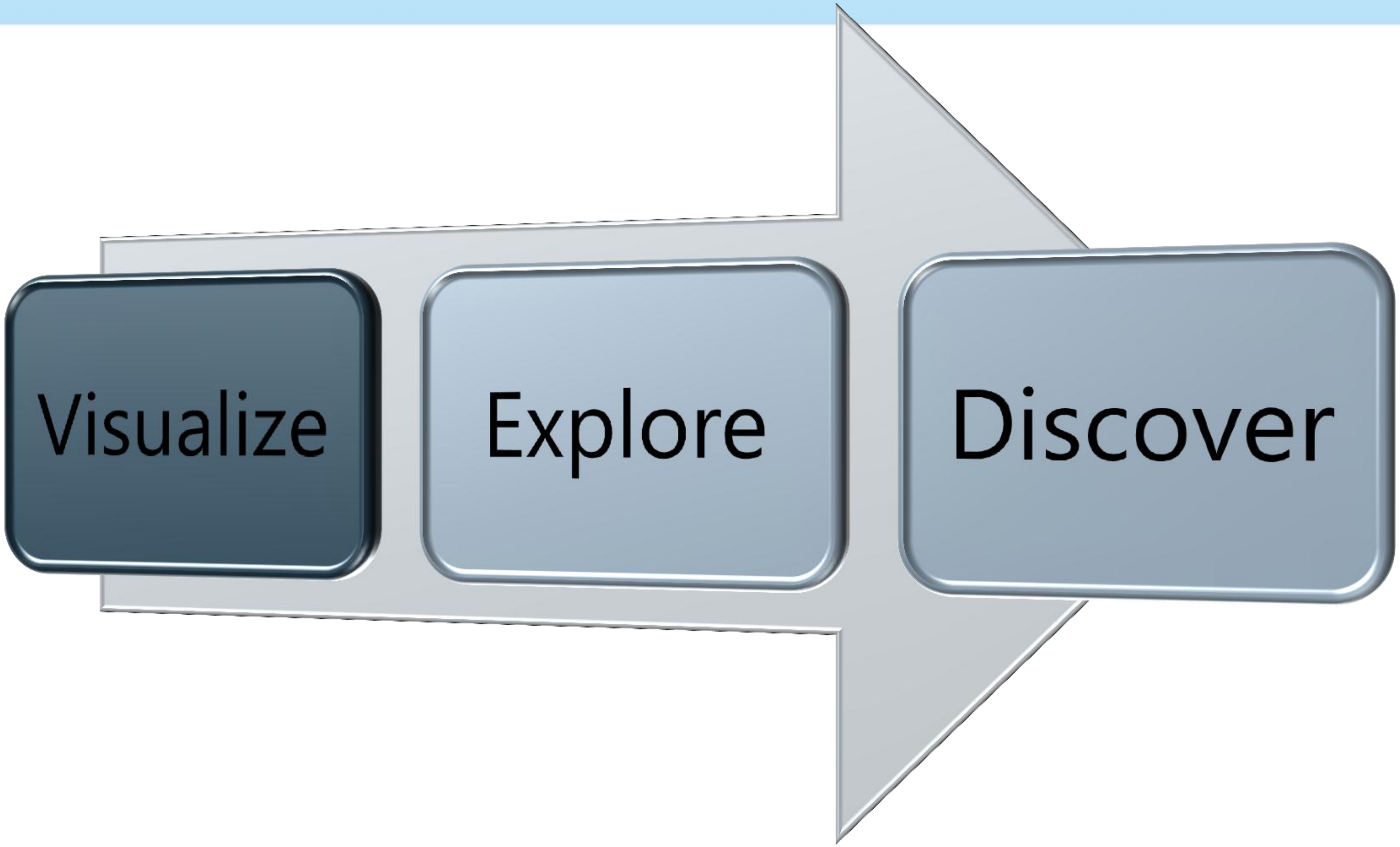
- Event-driven computations
- Reason about time
- Detect interesting patterns
- Connect to and correlate heterogeneous sources
- Process late data
- Process lots of data
- Re-use existing functions and algorithms

Microsoft StreamInsight

- API to build CEP applications
- Development in .NET & LINQ
- Maturing tool support
  - Event Flow Debugger
  - LINQPad
- Flexible integration of data sources / sinks
- Extensibility model
- Needs a SQL Server 2008 R2 License

StreamInsight Platform





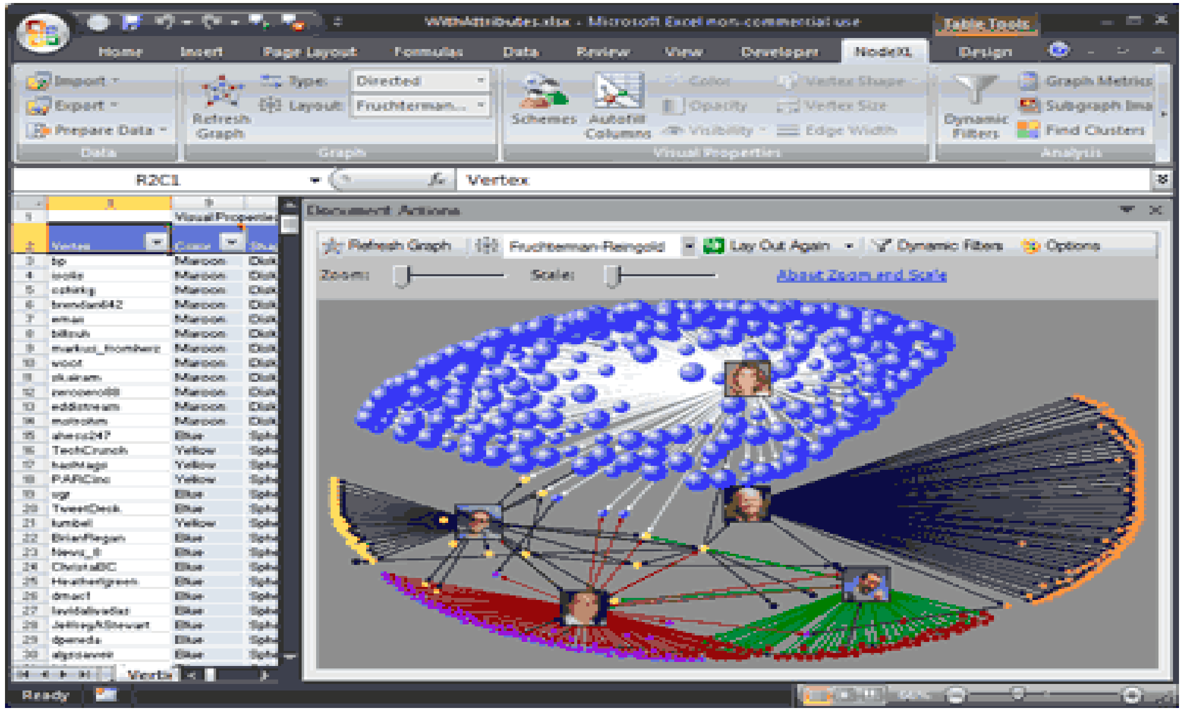
Visualize

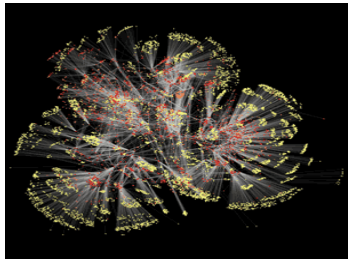
Explore

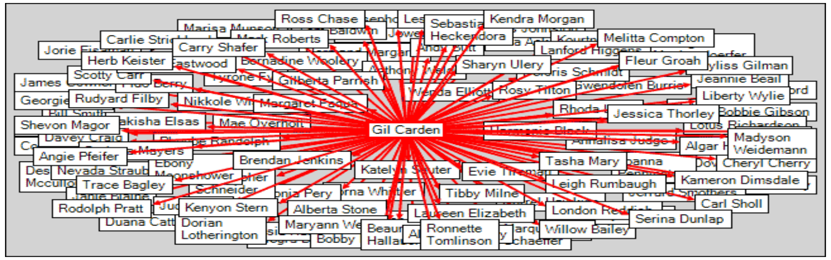
Discover

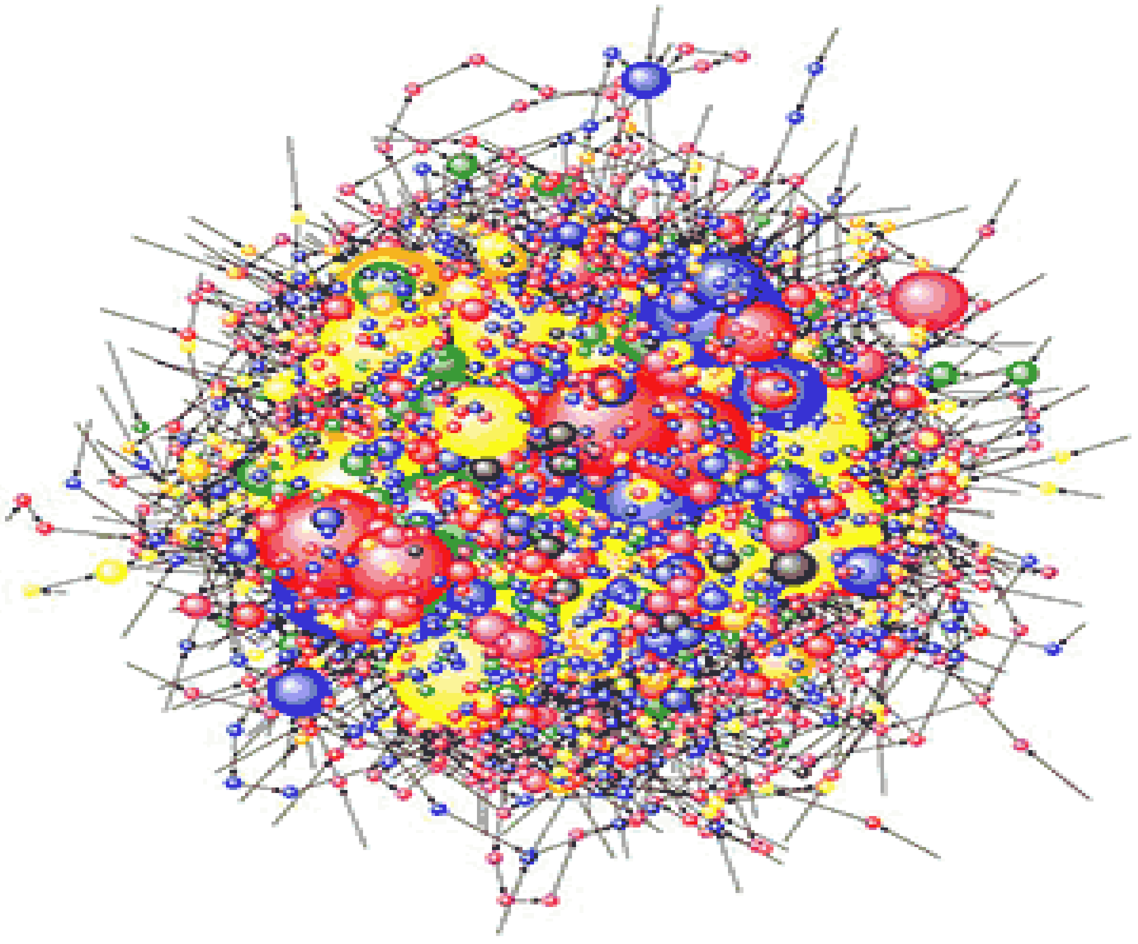
NodeXL

Network graph visualization



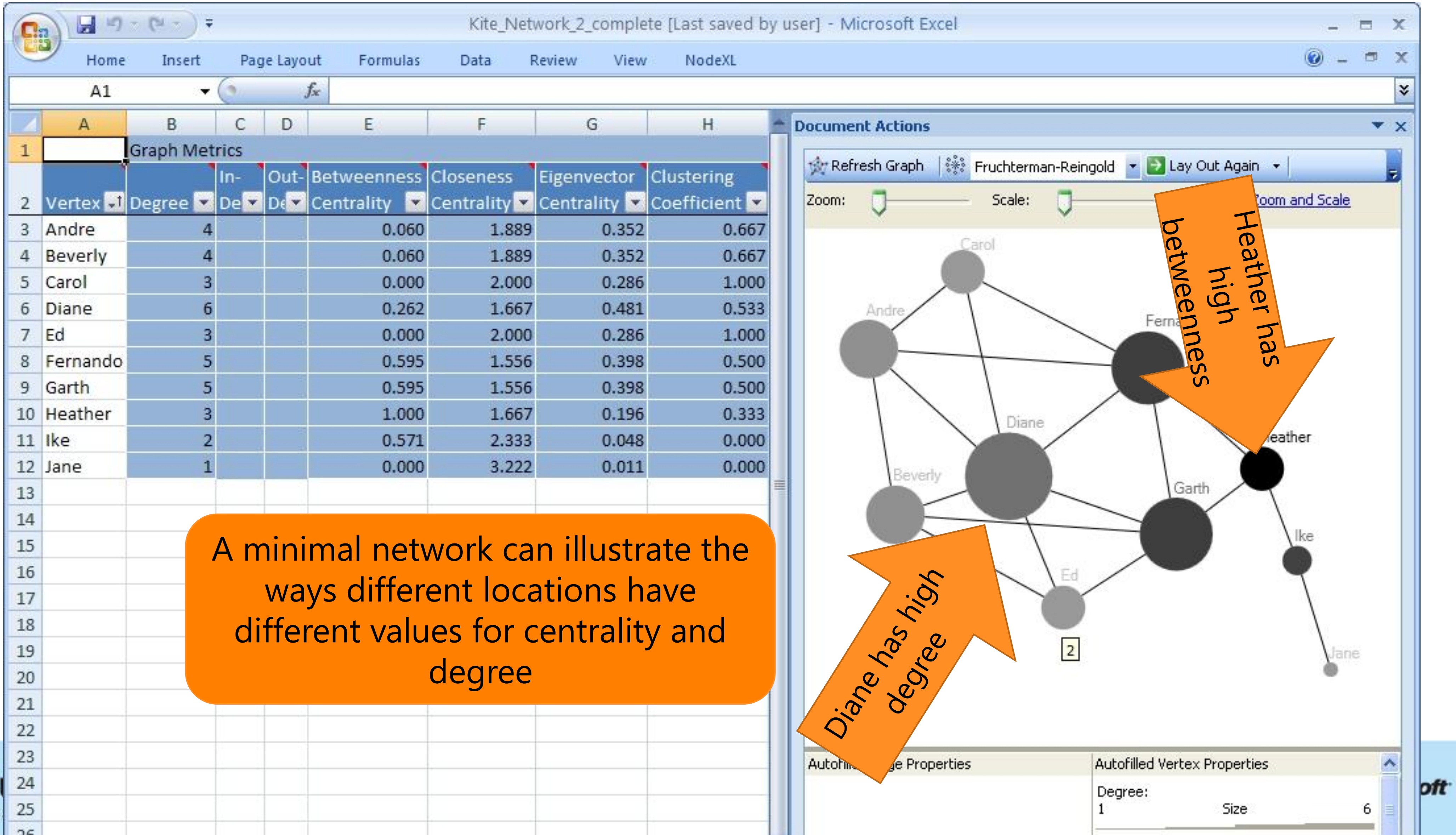






Binary and source code:
http://nodexl.codeplex.com

NodeXL - Network Overview Discovery and Exploration add-in for Excel 2007/2010

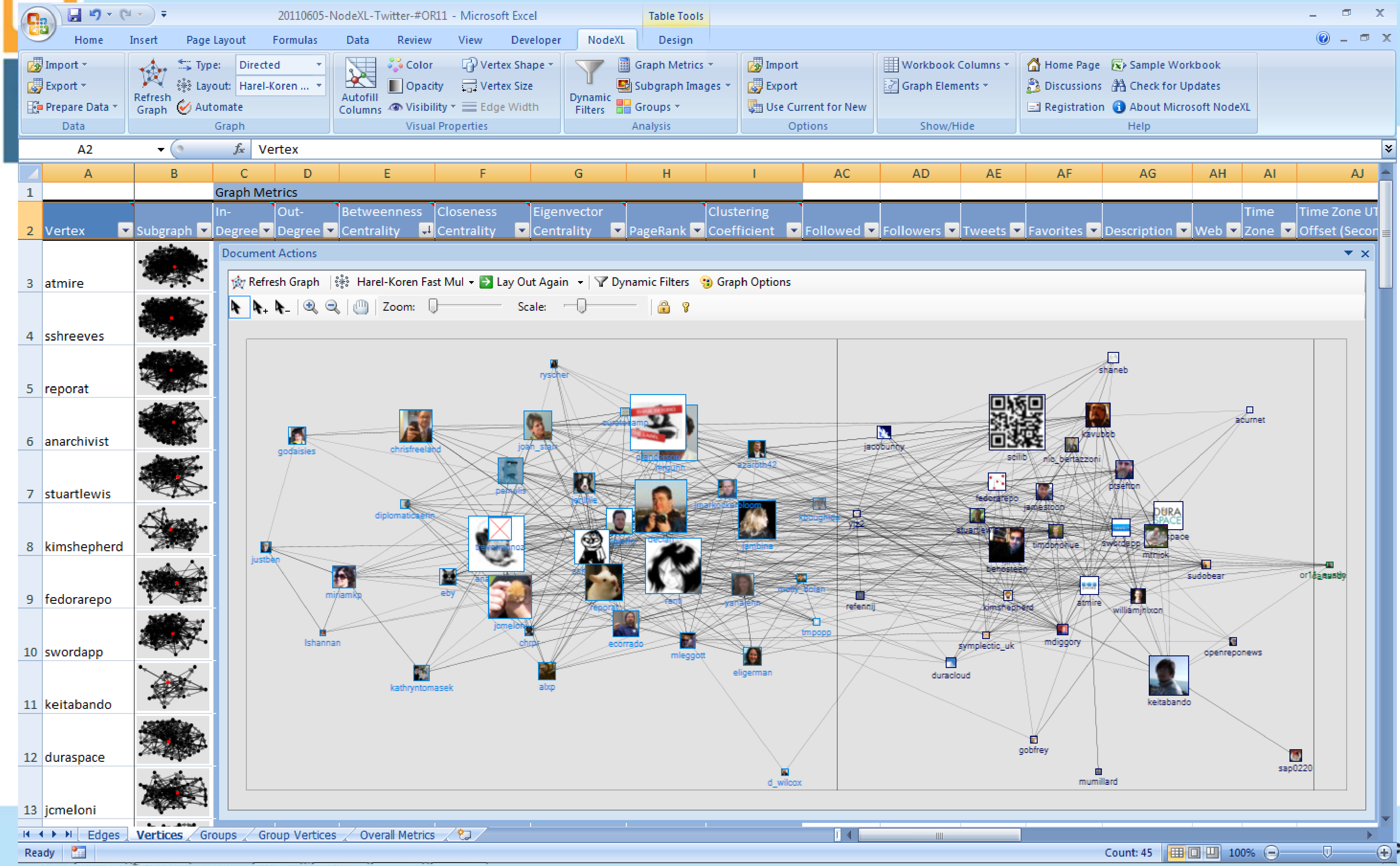


	Graph Metrics						
Vertex	Degree	In- De	Out- De	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	Clustering Coefficient
Andre	4			0.060	1.889	0.352	0.667
Beverly	4			0.060	1.889	0.352	0.667
Carol	3			0.000	2.000	0.286	1.000
Diane	6			0.262	1.667	0.481	0.533
Ed	3			0.000	2.000	0.286	1.000
Fernando	5			0.595	1.556	0.398	0.500
Garth	5			0.595	1.556	0.398	0.500
Heather	3			1.000	1.667	0.196	0.333
Ike	2			0.571	2.333	0.048	0.000
Jane	1			0.000	3.222	0.011	0.000

A minimal network can illustrate the ways different locations have different values for centrality and degree

Heather has high betweenness

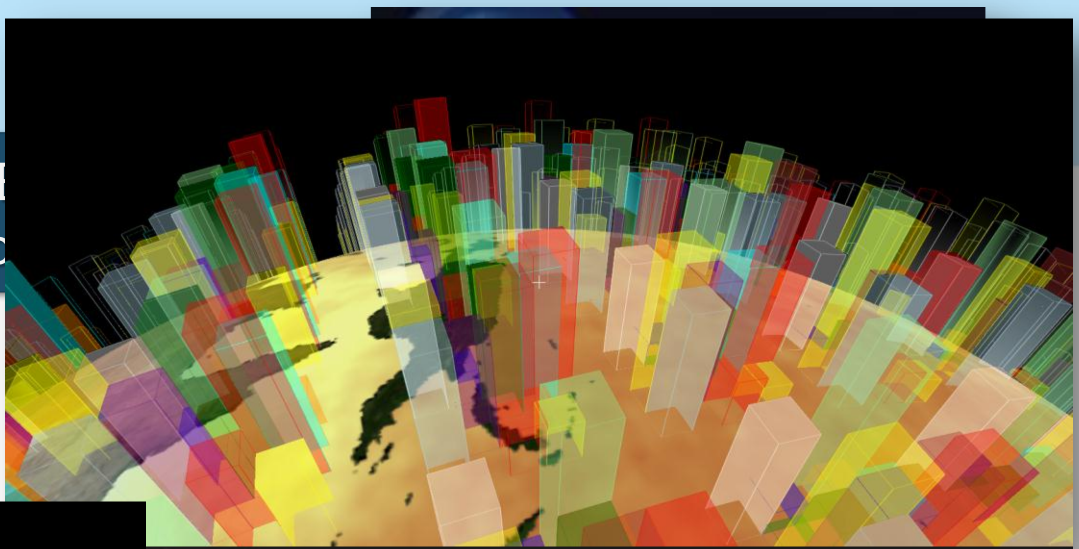
Diane has high degree



	A	B	C	D	E	F	G	H	I	AC	AD	AE	AF	AG	AH	AI	AJ
1			Graph Metrics														
2	Vertex	Subgraph	In-Degree	Out-Degree	Betweenness Centrality	Closeness Centrality	Eigenvector Centrality	PageRank	Clustering Coefficient	Followed	Followers	Tweets	Favorites	Description	Web	Time Zone	Time Zone UT Offset (Secon
3	atmire																
4	sshreeves																
5	reporat																
6	anarchivist																
7	stuartlewis																
8	kimshepherd																
9	fedorarepo																
10	swordapp																
11	keitabando																
12	duraspace																
13	jcmeloni																

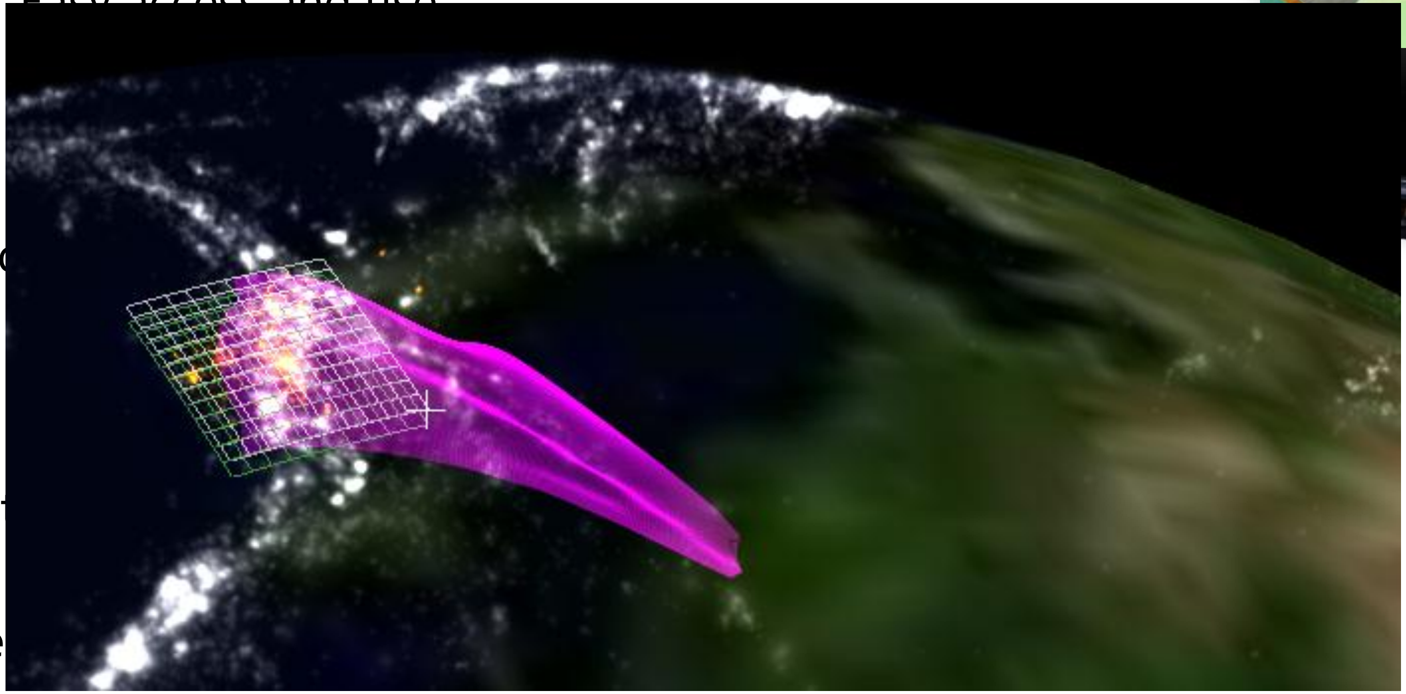
World Wide Telescope

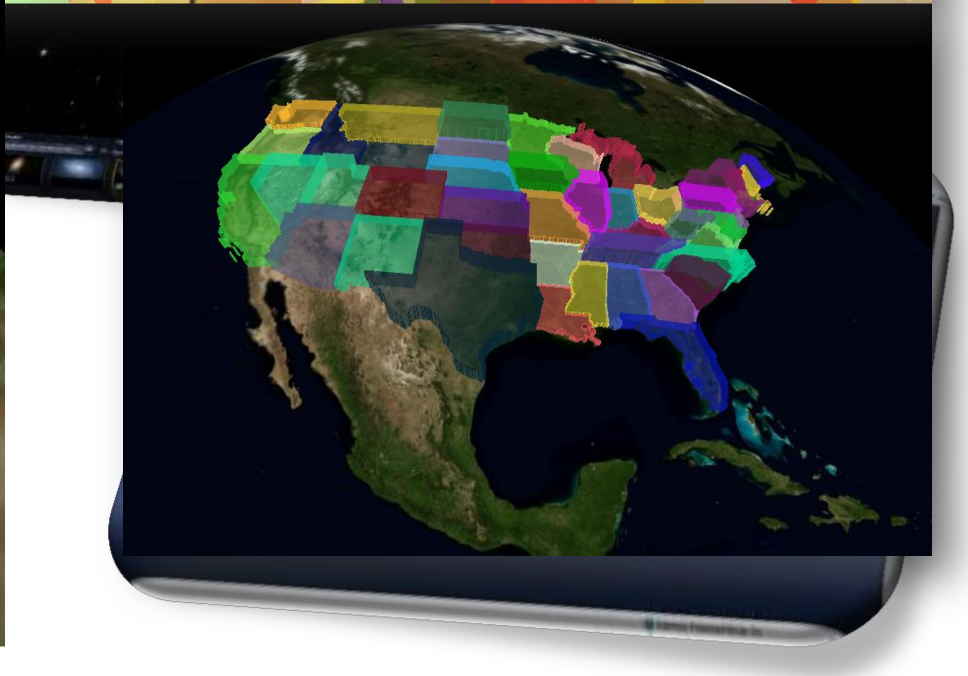
Seamless Rich Social Media Virtual Sky and E
Web application for science and educatic



Goals

- Integration of data sets and one-click contextual access
- Easy access and use
-
-





Up

-
-

No

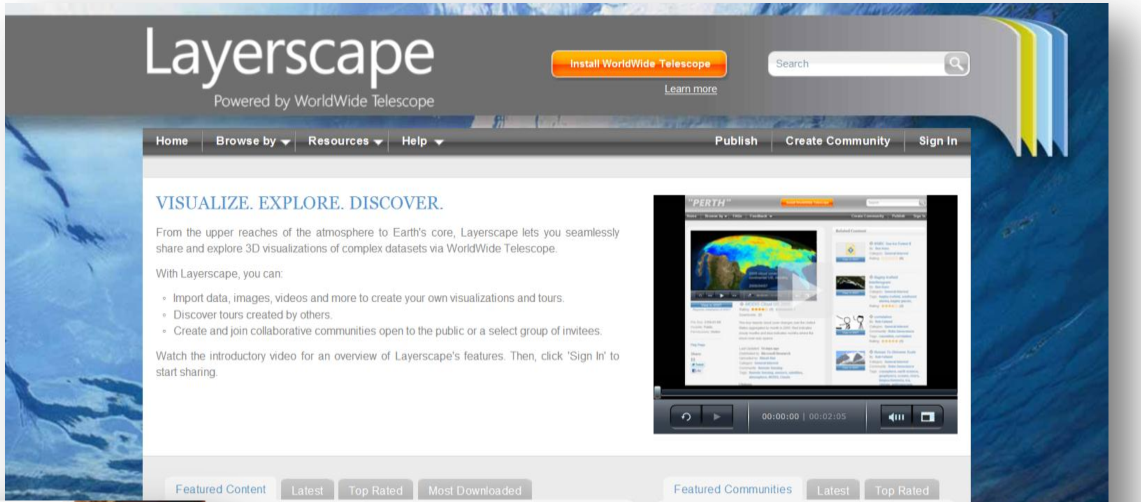
We

Layerscape

Powered by WorldWide Telescope

www.layerscape.org

- Community Site for WWT Tours and Layers (Data)
- Sharing by groups/individuals



VISUALIZE. EXPLORE. DISCOVER.

From the upper reaches of the atmosphere to Earth's core, Layerscape lets you seamlessly share and explore 3D visualizations of complex datasets via WorldWide Telescope.

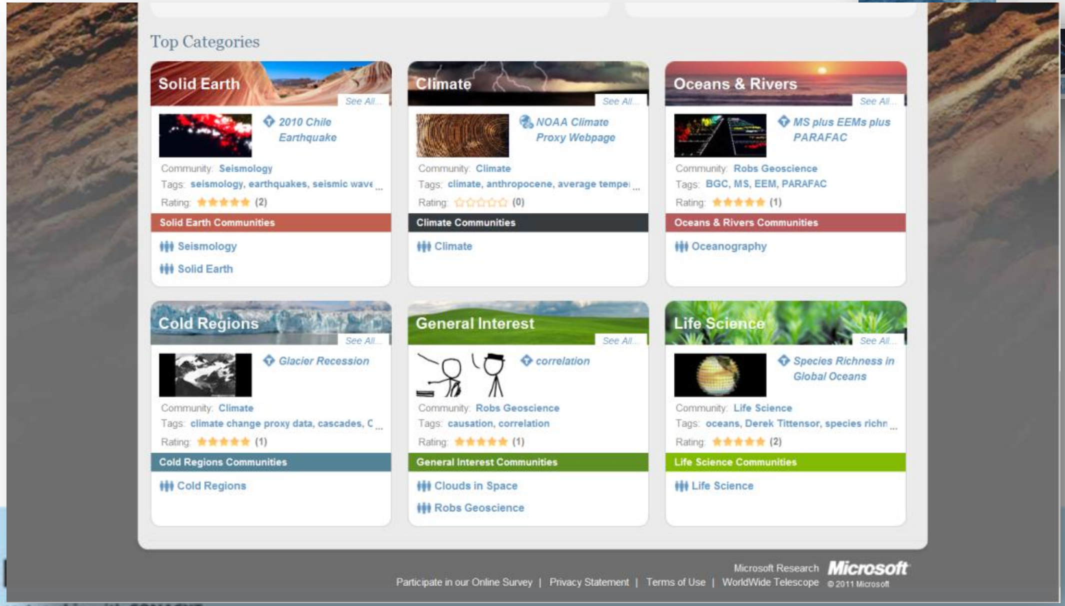
With Layerscape, you can:

- Import data, images, videos and more to create your own visualizations and tours.
- Discover tours created by others.
- Create and join collaborative communities open to the public or a select group of invitees.

Watch the introductory video for an overview of Layerscape's features. Then, click 'Sign In' to start sharing.

MBARI Dorado Drifter Tour 1 CTD Salinity
By: Rob Fatland
Category: Oceans & Rivers
Community: Oceanography

Robs Geoscience
By: Rob Fatland
Rating: ★★★★☆ (2)



Top Categories

Solid Earth See All...

2010 Chile Earthquake
Community: Seismology
Tags: seismology, earthquakes, seismic wave ...
Rating: ★★★★★ (2)

Solid Earth Communities
- Seismology
- Solid Earth

Climate See All...

NOAA Climate Proxy Webpage
Community: Climate
Tags: climate, anthropocene, average tempe ...
Rating: ☆☆☆☆☆ (0)

Climate Communities
- Climate

Oceans & Rivers See All...

MS plus EEMs plus PARAFAC
Community: Robs Geoscience
Tags: BGC, MS, EEM, PARAFAC
Rating: ★★★★★ (1)

Oceans & Rivers Communities
- Oceanography

Cold Regions See All...

Glacier Recession
Community: Climate
Tags: climate change proxy data, cascades, C ...
Rating: ★★★★★ (1)

Cold Regions Communities
- Cold Regions

General Interest See All...

correlation
Community: Robs Geoscience
Tags: causation, correlation
Rating: ★★★★★ (1)

General Interest Communities
- Clouds in Space
- Robs Geoscience

Life Science See All...

Species Richness in Global Oceans
Community: Life Science
Tags: oceans, Derek Tittensor, species richn ...
Rating: ★★★★★ (2)

Life Science Communities
- Life Science

Participate in our Online Survey | Privacy Statement | Terms of Use | WorldWide Telescope
Microsoft Research Microsoft
© 2011 Microsoft

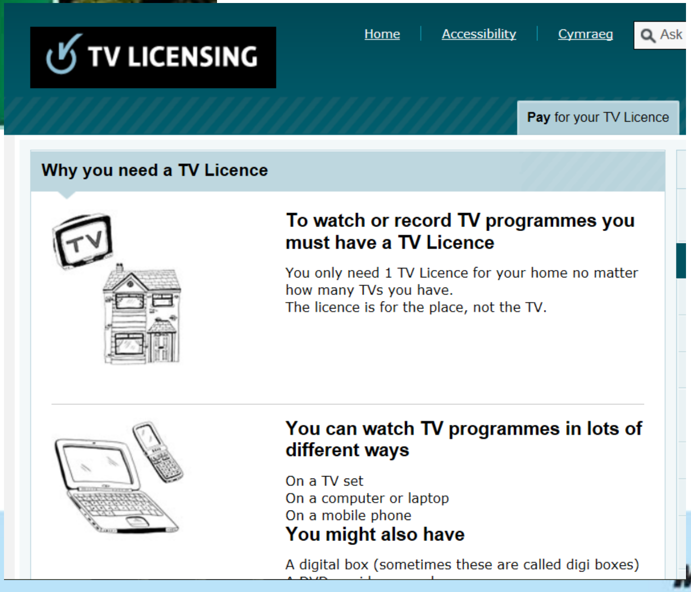
Natural User Interfaces (NUI) Kinect SDK and WWT

KINECT™
for Windows®

- 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

Data Storage Sustainability?

- Digital Data can be open – who should pay the cost?
- Spinning Disks, Bandwidth, Cooling, etc



Why you need a TV Licence

To watch or record TV programmes you must have a TV Licence

You only need 1 TV Licence for your home no matter how many TVs you have.
The licence is for the place, not the TV.

You can watch TV programmes in lots of different ways

On a TV set
On a computer or laptop
On a mobile phone

You might also have

A digital box (sometimes these are called digi boxes)

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

eScience in Action

Microsoft eScience Workshop 2012

October 8–9, 2012 | Chicago, Illinois, United States

http://research.microsoft.com/events/escience2012

Microsoft®

© 2012 Microsoft Corporation. All rights reserved. Microsoft, Windows, and other product names are or may be registered trademarks and/or trademarks in the U.S. and/or other countries.
The information herein is for informational purposes only and represents the current view of Microsoft Corporation as of the date of this presentation. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information provided after the date of this presentation. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS PRESENTATION.

Microsoft®

Research Connections

Presentation Fonts/Typography

What are the font choices and sizes?

Any font size 28pt or larger should use Segoe UI Light
Any type that is less than 28pt should use Segoe UI
On one slide, try to use a maximum of 3 font sizes

Where to start?

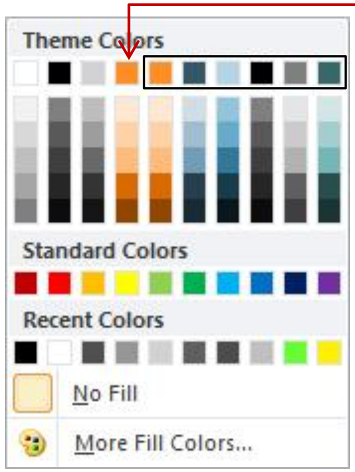
Control focus by using scale versus bullets
Use color to draw focus when necessary
Start with main topic at 40pt and subtopics at 20pt

Title/sentence casing and periods

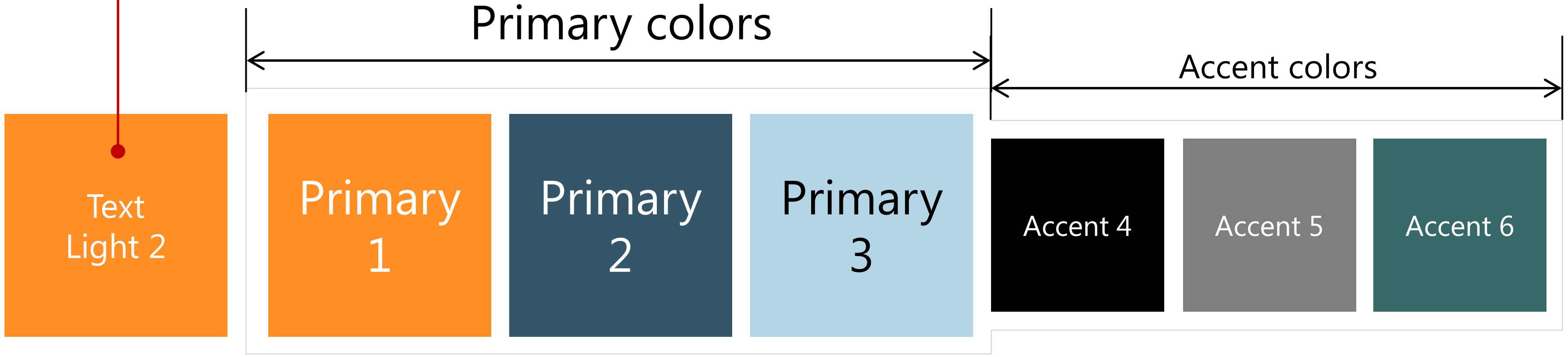
Title text should be "title caps" including a, is, of, & and
All supporting slide text should be sentence case
Periods should only be used on the title & main topics for complete sentences

Slide Palette Info

The PowerPoint palette for this template has been built for you and is shown below. Avoid using too many colors in your presentation.



Color text. Select the 4th color from the left for subheads and 1st level non-bulleted text color, or wherever “color” text is preferred over the default black/white text



Primary colors

Accent colors

Text Light 2

Primary 1

Primary 2

Primary 3

Accent 4

Accent 5

Accent 6

Use **Primary 1** as the main color. Use **Primary 2** and **Primary 3** when additional colors are needed.

Use **Accents 4-6** sparingly – only when more colors are necessary.

Preferred Text Layout (No Bullets)

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Same Color Text Layout (No Bullets)

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

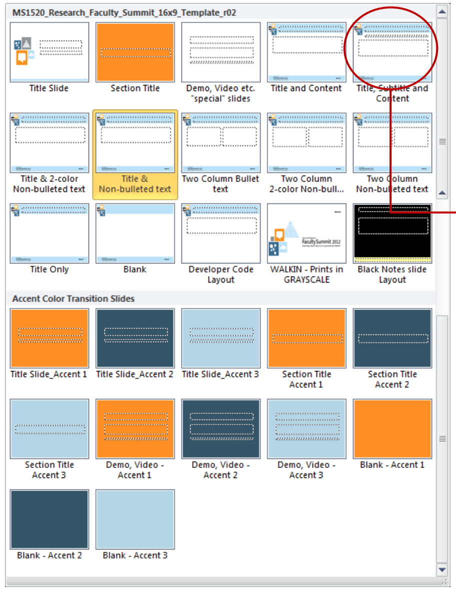
Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics



This is an almost identical layout to the previous slide, except that it has **all white text**. Sometimes you may prefer not to use colored text – for example if your list is only top level points, all white might look better. You can choose the layout you prefer.

Here's how to select different layouts:

1. Click on the Home tab at the top (if not already selected)
2. Click on Layout. A drop down list similar to the one shown on the left will appear. Notice that the layout for the slide you are on is **highlighted**. This slide uses a layout called "Title & Non-bulleted text"
3. Try clicking on the Layout to the left of it, called "Title & 2-color Non-bulleted text". Notice how the 1st level subheads change to a color.
4. Next try clicking on the layout called "Title and Content" layout. This is a the bulleted layout used on the next slide.
5. Use Layouts to set up new slides or to change existing slide layouts.

Adjusting List Levels

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

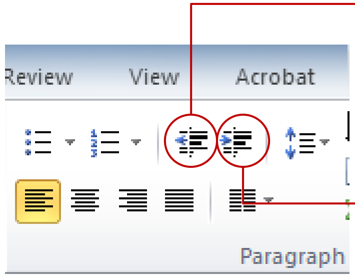
Size 20pt for the subtopics

Size 20pt for the subtopics

Use the “Increase List Level” and “Decrease List Level” tools on the Home Menu to change text levels.

Try this:

1. Place your cursor in any row of text to the left that says “Size 20pt for subtopics”
2. Next click the Home tab, and then on the “Decrease List level” tool. Notice how the line jumps up a level in size.
3. Now try placing your cursor in one of the “Main topic…” lines of text. Click the “Increase List Level” tool and see how the text is pushed down one level



Use these 2 tools to adjust your text levels as you work

Preferred Two Column Layout

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Same Color Two Column Layout

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 1: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 2: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Main topic 3: size 40pt

Size 20pt for the subtopics

Size 20pt for the subtopics

Demo Title

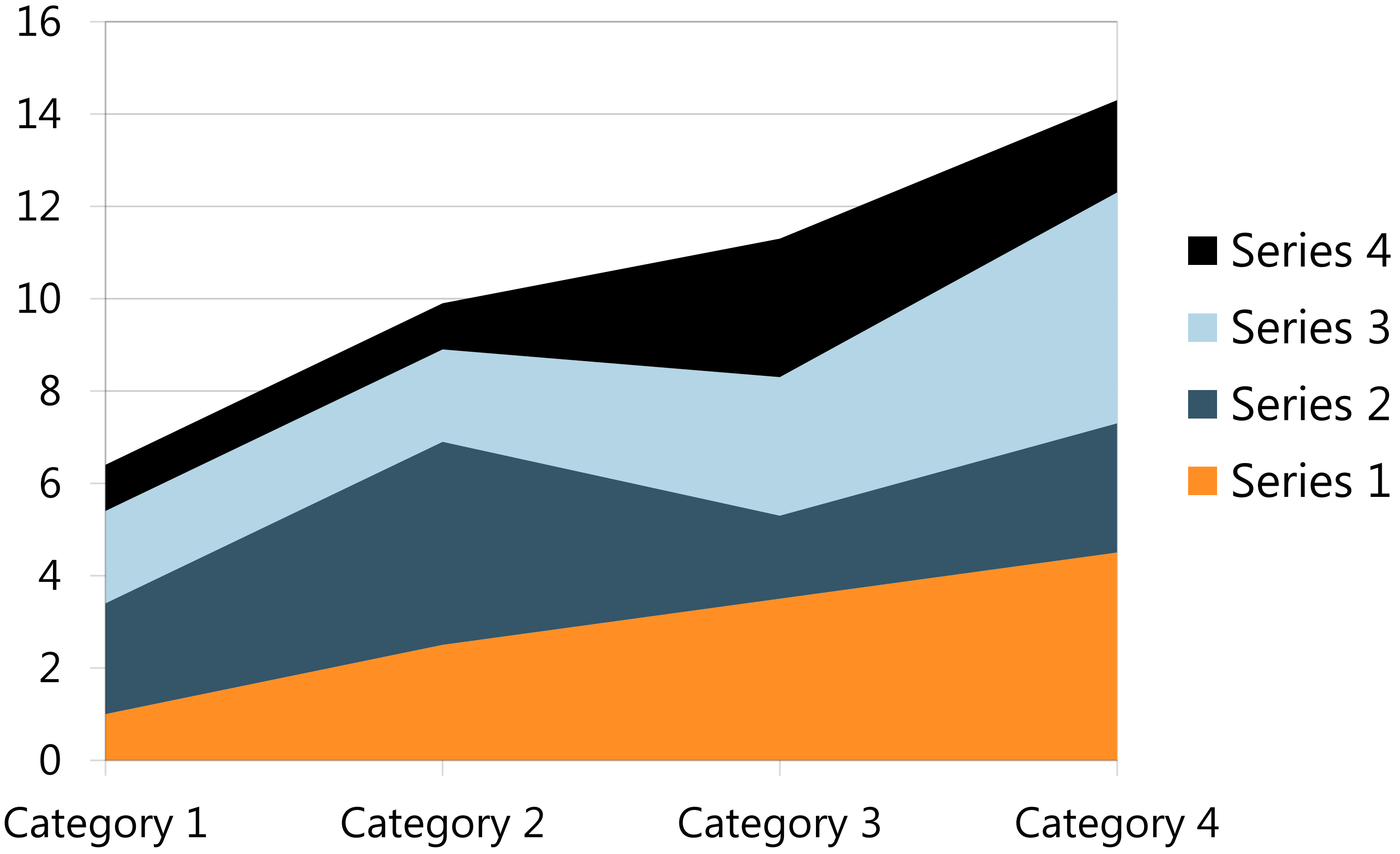
demo

Name
Title

Video Title

video

Chart Example



Slide for Showing Software Code

```
Add code here
```