The Changing Scientific Landscape

*Star Scientist -> Research Teams*: In former times, science was driven by key scientists. Today, science is driven by effectively collaborating co-author teams often comprising expertise from multiple disciplines and several geospatial locations (Börner, Dall'Asta, Ke, & Vespignani, 2005; Shneiderman, 2008).

*Users -> Contributors*: Web 2.0 technologies empower anybody to contribute to Wikipedia and to exchange images and videos via Fickr and YouTube. WikiSpecies, WikiProfessionals, or WikiProteins combine wiki and semantic technology in support of real time community annotation of scientific datasets (Mons et al., 2008).

*Cross-disciplinary*: The best tools frequently borrow and synergistically combine methods and techniques from different disciplines of science and empower interdisciplinary and/or international teams of researchers, practitioners, or educators to fine-tune and interpret results collectively.

*One Specimen -> Data Streams*: Microscopes and telescopes were originally used to study one specimen at a time. Today, many researchers must make sense of massive streams of multiple types of data with different formats, dynamics, and origin.

*Static Instrument -> Evolving Cyberinfrastructure (CI)*: The importance of hardware instruments that are rather static and expensive decreases relative to software infrastructures that are highly flexible and continuously evolving according to the needs of different sciences. Some of the most successful services and tools are decentralized increasing scalability and fault tolerance.

*Modularity*: The design of software modules with well defined functionality that can be flexibly combined helps reduce costs, makes it possible to have many contribute, and increases flexibility in tool development, augmentation, and customization.

*Standardization*: Adoption of standards speeds up development as existing code can be leveraged. It helps pool resources, supports interoperability, but also eases the migration from research code to production code and hence the transfer of research results into industry applications and products.

*Open data and open code*: Lets anybody check, improve, or repurpose code and eases the replication of scientific studies.
Just as the **microscope** empowered our naked eyes to see cells, microbes, and viruses thereby advancing the progress of biology and medicine or the **telescope** opened our minds to the immensity of the cosmos and has prepared mankind for the conquest of space, **macroscopes** promise to help us cope with another infinite: the infinitely complex. Macroscopes give us a ‘vision of the whole’ and help us ‘synthesize’. They let us detect patterns, trends, outliers, and access details in the landscape of science. Instead of making things larger or smaller, macroscopes let us observe what is at once too great, too slow, or too complex for our eyes.

---

**Desirable Features of Plug-and-Play Macroscopes**

**Division of Labor:** Ideally, labor is divided in a way that the expertise and skills of computer scientists are utilized for the design of standardized, modular, easy to maintain and extend “core architecture”. Dataset and algorithm plugins, i.e., the “filling”, are initially provided by those that care and know most about the data and developed the algorithms: the domain experts.

**Ease of Use:** As most plugin contributions and usage will come from non-computer scientists it must be possible to contribute, share, and use new plugins without writing one line of code. Wizard-driven integration of new algorithms and data sets by domain experts, sharing via email or online sites, deploying plugins by adding them to the ‘plugin’ directory, and running them via a Menu driven user interfaces (as used in Word processing systems or Web browsers) seems to work well.

**Plugin Content and Interfaces:** Should a plugin represent one algorithm or an entire tool? What about data converters needed to make the output of one algorithm compatible with the input of the next? Should those be part of the algorithm plugin or should they be packaged separately?

**Supported (Central) Data Models:** Some tools use a central data model to which all algorithms conform, e.g., Cytoscape, see Related Work section. Other tools support many internal data models and provide an extensive set of data converters, e.g., Network Workbench, see below. The former often speeds up execution and visual rendering while the latter eases the integration of new algorithms. In addition, most tools support an extensive set of input and output formats.

**Core vs. Plugins:** As will be shown, the “core architecture” and the “plugin filling” can be implemented as sets of plugin bundles. Answers to questions such as: “Should the graphical user interface (GUI), interface menu, scheduler, or data manager be part of the core or its filling?” will depend on the type of tools and services to be delivered.

**Supported Platforms:** If the software is to be used via Web interfaces then Web services need to be implemented. If a majority of domain experts prefers a stand-alone tool running on a specific operating system then a different deployment is necessary.
The Network Workbench (NWB) tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

In Aug. 2009, the tool provides more 160 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 40 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 30,000 times since Dec. 2006.


Project Details

Investigators: Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert

Software Team: Lead: Micah Linnemeier

Members: Patrick Phillips, Russell Duhon, Tim Kelley & Ann McCranie

Previous Developers: Weixia (Bonnie) Huang, Bruce Herr, Heng Zhang, Duygu Balcan, Mark Price, Ben Markines, Santo Fortunato, Felix Terkhorn, Ramya Sabbineni, Vivek S. Thakre & Cesar Hidalgo

Goal: Develop a large-scale network analysis, modeling and visualization toolkit for physics, biomedical, and social science research.

Amount: $1,120,926, NSF IIS-0513650 award


Website: http://nwb.slis.indiana.edu
**NWB Tool: Supported Data Formats**

**Personal Bibliographies**
- Bibtex (.bib)
- Endnote Export Format (.cnw)

**Data Providers**
- Web of Science by Thomson Scientific/Reuters (.isi)
- Scopus by Elsevier (.scopus)
- Google Scholar (access via Publish or Perish save as CSV, Bibtex, EndNote)
- Awards Search by National Science Foundation (.nsf)

**Scholarly Database** (all text files are saved as .csv)
- Medline publications by National Library of Medicine
- NIH funding awards by the National Institutes of Health (NIH)
- NSF funding awards by the National Science Foundation (NSF)
- U.S. patents by the United States Patent and Trademark Office (USPTO)
- Medline papers – NIH Funding

**Network Formats**
- NWB (.nwb)
- Pajek (.mat)
- GraphML (.xml or .graphml)
- XGMML (.xml)

**Burst Analysis Format**
- Burst (.burst)

**Other Formats**
- CSV (.csv)
- Edgelist (.edge)
- Pajek (.mat)
- TreeML (.xml)
NWB Tool: Algorithms (July 1st, 2008)
See https://nwb.slis.indiana.edu/community and handout for details.

---

NWB Tool: Output Formats

- NWB tool can be used for data conversion. Supported output formats comprise:
  - CSV (.csv)
  - NWB (.nwb)
  - Pajek (.net)
  - Pajek (.mat)
  - GraphML (.xml or .graphml)
  - XGMML (.xml)

- GUESS
  Supports export of images into common image file formats.
  - Horizontal Bar Graphs
  - saves out raster and ps files.
Exemplary Analyses and Visualizations

Individual Level
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

Scientific Field Level
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.
Data Acquisition from Web of Science

Download all papers by
- Eugene Garfield
- Stanley Wasserman
- Alessandro Vespignani
- Albert-László Barabási

from
- Science Citation Index Expanded (SCI-EXPANDED) --1955-present
- Social Sciences Citation Index (SSCI) --1956-present
- Arts & Humanities Citation Index (A&HCI) --1975-present

Comparison of Counts
No books and other non-WoS publications are covered.

<table>
<thead>
<tr>
<th>Age</th>
<th>Total # Cites</th>
<th>Total # Papers</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eugene Garfield</td>
<td>82</td>
<td>1,525</td>
<td>672</td>
</tr>
<tr>
<td>Stanley Wasserman</td>
<td></td>
<td>122</td>
<td>35</td>
</tr>
<tr>
<td>Alessandro Vespignani</td>
<td>42</td>
<td>451</td>
<td>101</td>
</tr>
<tr>
<td>Albert-László Barabási</td>
<td>40</td>
<td>2,218</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>16,920</td>
<td>159</td>
</tr>
</tbody>
</table>
Extract Co-Author Network

Load *yourwbdirectory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi’ using 'File > Load and Clean ISI File'.
To extract the co-author network, select the ‘361 Unique ISI Records’ table and run 'Scientometrics > Extract Co-Author Network’ using isi file format:

The result is an undirected network of co-authors in the Data Manager. It has 247 nodes and 891 edges.
To view the complete network, select the network and run 'Visualization > GUESS > GEM'. Run Script > Run Script… And select Script folder > GUESS > co-author-nw.py.

Comparison of Co-Author Networks

Eugene Garfield

Stanley Wasserman

Alessandro Vespignani

Albert-László Barabási
Joint Co-Author Network of all Four NetsSci Researchers

Paper-Citation Network Layout

Load `yournwbdirectory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi` using 'File > Load and Clean ISI File'.

To extract the paper-citation network, select the '361 Unique ISI Records' table and run 'Scientometrics > Extract Directed Network' using the parameters:

The result is a directed network of paper citations in the Data Manager. It has 5,335 nodes and 9,595 edges.

To view the complete network, select the network and run 'Visualization > GUESS'. Run 'Script > Run Script …' and select 'yournwbdirectory*/script/GUESS/paper-citation-nw.py'.
Exemplary Analyses and Visualizations

**Individual Level**

A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.

B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

**Institution Level**


**Scientific Field Level**

D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.
### NSF Awards Search Results

<table>
<thead>
<tr>
<th>Name</th>
<th># Awards</th>
<th>First A. Starts</th>
<th>Total Amount to Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geoffrey Fox</td>
<td>27</td>
<td>Aug 1978</td>
<td>12,196,260</td>
</tr>
<tr>
<td>Michael McRobbie</td>
<td>8</td>
<td>July 1997</td>
<td>19,611,178</td>
</tr>
<tr>
<td>Beth Plale</td>
<td>10</td>
<td>Aug 2005</td>
<td>7,224,522</td>
</tr>
</tbody>
</table>

**Disclaimer:**

Only NSF funding, no funding in which they were senior personnel, only as good as NSF’s internal record keeping and unique person ID. If there are ‘collaborative’ awards then only their portion of the project (award) will be included.
Using NWB to Extract Co-PI Networks

- Load into NWB, open file to count records, compute total award amount.
- Run ‘Scientometrics > Extract Co-Occurrence Network’ using parameters:
  - Sl "Ednk"
d  "Alin Alik Tlki"
- Select “Extracted Network ..” and run ‘Analysis > Network Analysis Toolkit (NAT)’
- Remove unconnected nodes via ‘Preprocessing > Delete Isolates’.
- ‘Visualization > GUESS’, layout with GEM
- Run ‘co-PI-nw.py’ GUESS script to color/size code.
Exemplary Analyses and Visualizations

Individual Level
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

Scientific Field Level
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

Active NSF Awards on 11/07/2008:

- Indiana University 257
  (there is also Indiana University at South Bend Indiana University Foundation, Indiana University Northwest, Indiana University-Purdue University at Fort Wayne, Indiana University-Purdue University at Indianapolis, Indiana University-Purdue University School of Medicine)
- Cornell University 501
  (there is also Cornell University – State, Joan and Sanford I. Weill Medical College of Cornell University)
- University of Michigan Ann Arbor 619
  (there is also University of Michigan Central Office, University of Michigan Dearborn, University of Michigan Flint, University of Michigan Medical School)

Active NSF Awards on 09/10/2009:

- Stanford University 429

Save files as csv but rename into .nsf.
Or simply use the files saved in ‘*yournwbdirectory*/sampledata/scientometrics/ nsf/’.

Extracting Co-PI Networks

Load NSF data, selecting the loaded dataset in the Data Manager window, run ‘Scientometrics > Extract Co-Occurrence Network’ using parameters:

Two derived files will appear in the Data Manager window: the co-PI network and a merge table. In the network, nodes represent investigators and edges denote their co-PI relationships. The merge table can be used to further clean PI names.

Running the ‘Analysis > Network Analysis Toolkit (NAT)’ reveals that the number of nodes and edges but also of isolate nodes that can be removed running ‘Preprocessing > Delete Isolates’.

Select ‘Visualization > GUESS’ to visualize. Run ‘co-PI-nw.py’ script.
Extract Giant Component

Select network after removing isolates and run ‘Analysis > Unweighted and Undirected > Weak Component Clustering’ with parameter

Indiana’s largest component has 19 nodes, Cornell’s has 67 nodes, Michigan’s has 55 nodes.

Visualize Cornell network in GUESS using same .py script and save via ‘File > Export Image’ as jpg.
Top-10 Investigators by Total Award Money

for i in range(0, 10):
    print str(nodesbytotalawardmoney[i].label) + "": " +
    str(nodesbytotalawardmoney[i].totalawardmoney)
Stanford University
429 active NSF awards on 09/10/2009

Largest component
39 nodes

Stanford U:
218 nodes, 285 edges, 49 components
157 isolate nodes were deleted
Top-10 Investigators by Total Award Money

for i in range(0, 10):
    print str(nodesbytotalawardmoney[i].label) + str(nodesbytotalawardmoney[i].totalawardmoney)

Stanford University
Dan Boneh: 11,837,800
Rajeev Motwani: 11,232,154
Hector Garcia-Molina: 10,577,906
David Goldhaber-Gordon: 9,792,029
Kathryn Moler: 7,870,029
John C. Mitchell: 7,290,668
Alfred Spormann: 6,803,000
Gordon Brown: 6,158,000
Jennifer Widom: 5,661,311
3. Exemplary Analyses and Visualizations

**Individual Level**
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

**Institution Level**

**Scientific Field Level**
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.
### Top-10 burst terms from abstracts of the AI search results

<table>
<thead>
<tr>
<th>Medline</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>medical</td>
<td>17</td>
<td>299.7924</td>
<td>1983</td>
<td>1999</td>
</tr>
<tr>
<td>knowledge</td>
<td>5</td>
<td>293.3575</td>
<td>1991</td>
<td>1995</td>
</tr>
<tr>
<td>knowledge</td>
<td>6</td>
<td>215.2407</td>
<td>1997</td>
<td>2002</td>
</tr>
<tr>
<td>expert</td>
<td>15</td>
<td>171.0442</td>
<td>1985</td>
<td>1994</td>
</tr>
<tr>
<td>systems</td>
<td>15</td>
<td>170.3326</td>
<td>1985</td>
<td>1999</td>
</tr>
<tr>
<td>intelligence</td>
<td>21</td>
<td>123.9784</td>
<td>1983</td>
<td>2001</td>
</tr>
<tr>
<td>patient</td>
<td>21</td>
<td>123.9297</td>
<td>1962</td>
<td>2002</td>
</tr>
<tr>
<td>sage</td>
<td>12</td>
<td>106.5522</td>
<td>1999</td>
<td>2001</td>
</tr>
<tr>
<td>registration</td>
<td>5</td>
<td>104.8189</td>
<td>2005</td>
<td></td>
</tr>
<tr>
<td>knowledge-based</td>
<td>16</td>
<td>98.83778</td>
<td>1987</td>
<td>2002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NIH</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>Place</td>
<td>8</td>
<td>117.2205</td>
<td>1993</td>
<td>2000</td>
</tr>
<tr>
<td>commercial</td>
<td>9</td>
<td>87.5718</td>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>proposed</td>
<td>9</td>
<td>87.5718</td>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>case</td>
<td>5</td>
<td>42.25052</td>
<td>1978</td>
<td>1990</td>
</tr>
<tr>
<td>protein</td>
<td>1</td>
<td>72.15968</td>
<td>1988</td>
<td>1988</td>
</tr>
<tr>
<td>networks</td>
<td>4</td>
<td>71.252</td>
<td>1993</td>
<td>1996</td>
</tr>
<tr>
<td>patterns</td>
<td>3</td>
<td>66.44626</td>
<td>1977</td>
<td>1979</td>
</tr>
<tr>
<td>being</td>
<td>8</td>
<td>66.2954</td>
<td>1971</td>
<td>1978</td>
</tr>
<tr>
<td>reasoning</td>
<td>2</td>
<td>65.68178</td>
<td>1984</td>
<td>1985</td>
</tr>
<tr>
<td>expert</td>
<td>4</td>
<td>60.49955</td>
<td>1987</td>
<td>1990</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NSF</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>theory</td>
<td>6</td>
<td>47.05097</td>
<td>1999</td>
<td></td>
</tr>
<tr>
<td>gray</td>
<td>2</td>
<td>28.39208</td>
<td>2009</td>
<td>2001</td>
</tr>
<tr>
<td>learning</td>
<td>2</td>
<td>27.40728</td>
<td>1997</td>
<td>1998</td>
</tr>
<tr>
<td>human</td>
<td>5</td>
<td>25.4525</td>
<td>2009</td>
<td></td>
</tr>
<tr>
<td>control</td>
<td>2</td>
<td>24.07277</td>
<td>1992</td>
<td>1993</td>
</tr>
<tr>
<td>knowledge</td>
<td>1</td>
<td>21.48756</td>
<td>1992</td>
<td>1996</td>
</tr>
<tr>
<td>students</td>
<td>1</td>
<td>21.07674</td>
<td>1997</td>
<td>1997</td>
</tr>
<tr>
<td>problem</td>
<td>2</td>
<td>20.71133</td>
<td>1998</td>
<td>1999</td>
</tr>
<tr>
<td>more</td>
<td>2</td>
<td>19.98819</td>
<td>2000</td>
<td>2001</td>
</tr>
<tr>
<td>we</td>
<td>1</td>
<td>19.38503</td>
<td>2001</td>
<td>2001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>USPTO</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>3</td>
<td>19.03537321</td>
<td>2004</td>
<td>2006</td>
</tr>
<tr>
<td>video</td>
<td>3</td>
<td>15.32735425</td>
<td>1998</td>
<td>2000</td>
</tr>
<tr>
<td>disclosed</td>
<td>2</td>
<td>14.06616781</td>
<td>1999</td>
<td>2000</td>
</tr>
<tr>
<td>neural</td>
<td>3</td>
<td>13.3019596</td>
<td>2004</td>
<td>2006</td>
</tr>
<tr>
<td>current</td>
<td>2</td>
<td>12.4336047</td>
<td>1999</td>
<td>2000</td>
</tr>
<tr>
<td>unit</td>
<td>2</td>
<td>12.33743838</td>
<td>2002</td>
<td>2003</td>
</tr>
<tr>
<td>material</td>
<td>1</td>
<td>12.08420935</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>feedback</td>
<td>1</td>
<td>12.0730195</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>role</td>
<td>1</td>
<td>12.0730195</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>elevator</td>
<td>4</td>
<td>11.8351827</td>
<td>1991</td>
<td>1994</td>
</tr>
</tbody>
</table>

---

**Science of Science Cyberinfrastructure (PORTAL)**

*Provided by the Cyberinfrastructure for Network Science Center at Indiana University.*

**Introduction**

E. O. Wilson writes in *Consilience: The Unity of Knowledge* (1978) that the integration of science and the humanities is one of the great challenges of the future. The aim of the Science of Science Cyberinfrastructure (PORTAL) project is to provide the tools and resources necessary to support this integration.

**Needs Analysis**

All ASI sites will use a combination of social and natural resources to support the integration of science and the humanities. A special journal issue of the *Science of Science: Conceptualizations and Models of Science* edited by E. O. Wilson will be published in the *Journal of Cyberinfrastructure* in January 2009.

**Cyberinfrastructures**

The Cyberinfrastructures of the Network Working Group (NWGB) provides a unique distributed, shared environment for large-scale network analysis, modeling, and visualization. Thomas Stocker, R. S. Smolin, and Google Scholar data, Enrich and Bibo data, or NSF awards can be read and diverse networks can be visualized in a single environment.

[http://sci.slis.indiana.edu](http://sci.slis.indiana.edu)
Macrosope Outlook

CIShell/OSGi is at the core of different CIs and a total of 169 unique plugins are used in the
Information Visualization (http://iv.slis.indiana.edu),
Network Science (http://nwb.slis.indiana.edu),
Science Policy (http://sci.slis.indiana.edu), and
Epidemics (http://epic.slis.indiana.edu) research communities.

Most interestingly, a number of other projects recently adopted OSGi and one adopted CIShell:
Cytoscape (http://www.cytoscape.org) lead by Trey Ideker, UCSD is an open source bioinformatics
software platform for visualizing molecular interaction networks and integrating these interactions
with gene expression profiles and other state data (Shannon et al., 2002).
Taverna Workbench (http://taverna.sourceforge.net) lead by Carol Goble, University of Manchester,
UK is a free software tool for designing and executing workflows (Hull et al., 2006). Taverna allows
users to integrate many different software tools, including over 30,000 web services.
MAEviz (https://wiki.ncsa.uiuc.edu/display/MAE/Home) managed by Shawn Hampton, NCSA is an
open-source, extensible software platform which supports seismic risk assessment based on the Mid-
America Earthquake (MAE) Center research.
TEXTrend (http://www.textrend.org) lead by George Kampis, Eötvös University, Hungary develops a
framework for the easy and flexible integration, configuration, and extension of plugin-based
components in support of natural language processing (NLP), classification/mining, and graph
algorithms for the analysis of business and governmental text corpuses with an inherently temporal
component.

As the functionality of OSGi-based software frameworks improves and the number and diversity of
dataset and algorithm plugins increases, the capabilities of custom tools or macrosopes will expand.

All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.slis.indiana.edu