Open Data and Open Code for BIG Science of Science Studies

Robert P. Light[+], David E. Polley[+], and Katy Börner[+*]

[+] CNS & ILS, SOIC, Indiana University, Bloomington, Indiana, USA
[*] Royal Netherlands Academy of Arts and Sciences, Amsterdam, The Netherlands
http://cns.iu.edu

14th ISSI Conference
Vienna, Austria

Thursday July 18, 2013

Goals of the Paper/Structure of this Talk

Inspire the development of "Open Data and Open Code for BIG Science of Science Studies" **see ISSI 2013 Workshop on "Standards for Science Mapping and Classifications"**

Introduce a database-tool infrastructure designed to support big SoS studies:
- The open access Scholarly Database (SDB) (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records.
- The open source Science of Science (Sci2) tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies, **see ISSI 2013 Tutorial on Workshop on "Sci2: A Tool of Science of Science Research and Practice"**

Showcase scalability of the infrastructure:
- temporal analyses scale linearly with the number of records and file size.
- geospatial algorithm show quadratic growth.
- network science algorithms scale with the number of edges rather than nodes.

Motivation

Historically,

- science of science studies were/are performed
- by single investigators or small teams using
- proprietary data and
- proprietary software tools.

Few results can be replicated.

Big science of science studies

- utilize "big data", i.e., large, complex, diverse, longitudinal, and/or distributed datasets that might be owned by different stakeholders
- apply a systems science approach to uncover hidden patterns, bursts of activity, correlations, laws, etc.
- make available open data and open code in support of
- replication of results, iterative refinement of approaches and tools, and education.

Motivation

nature Vol 464|25 March 2010

OPINION

Let's make science metrics more scientific

To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity, says **Julia Lane**.

SUMMARY

- Existing metrics have known flaws
- A reliable, open, joined-up data infrastructure is needed
- Data should be collected on the full range of scientists' work
- Social scientists and economists should be involved

Scientometricians, Webometricians, Infometricians should also be involved.

Goals of the Paper/Structure of this Talk

Inspire the development of "Open Data and Open Code for BIG Science of Science Studies" **see ISSI 2013 Workshop on "Standards for Science Mapping and Classifications"**

Introduce a database-tool infrastructure designed to support big SoS studies:

- The open access Scholarly Database (SDB) (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records.
- The open source Science of Science (Sci2) tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies, **see ISSI 2013 Tutorial on Workshop on "Sci2: A Tool of Science of Science Research and Practice"**

Showcase scalability of the infrastructure:

- temporal analyses scale linearly with the number of records and file size.
- geospatial algorithm show quadratic growth.
- network science algorithms scale with the number of edges rather than nodes.

5

Goals of the Paper/Structure of this Talk

Inspire the development of "Open Data and Open Code for BIG Science of Science Studies" **see ISSI 2013 Workshop on "Standards for Science Mapping and Classifications"**

Introduce a database-tool infrastructure designed to support big SoS studies:

- The open access Scholarly Database (SDB) (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records.
- The open source Science of Science (Sci2) tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies, **see ISSI 2013 Tutorial on Workshop on "Sci2: A Tool of Science of Science Research and Practice"**

Showcase scalability of the infrastructure:

- temporal analyses scale linearly with the number of records and file size.
- geospatial algorithm show quadratic growth.
- network science algorithms scale with the number of edges rather than nodes.

6

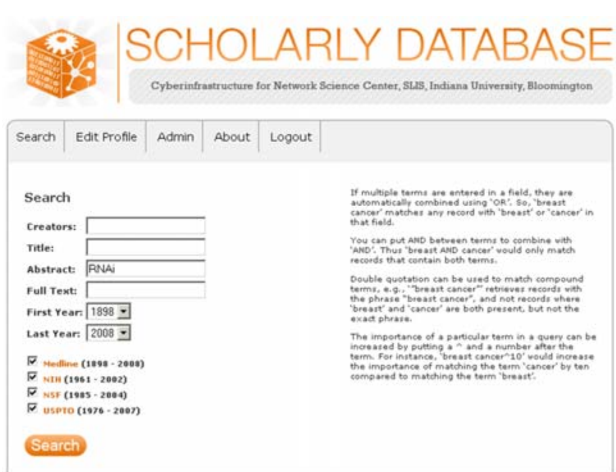
Data Access & Preprocessing Challenges

- Different datasets by diverse providers: need to align formats and their changes over time.
- MS Excel can load a maximum of 1,048,576 rows of data by 16,384 columns per sheet or a max of 2 gigabytes. Larger datasets need to be stored in a DB.
- Preprocessing comprises identification of uniqueX, geocoding, science coding, extraction of networks, among others.
- Data cleaning & preprocessing easily consumes 80 percent of project effort.

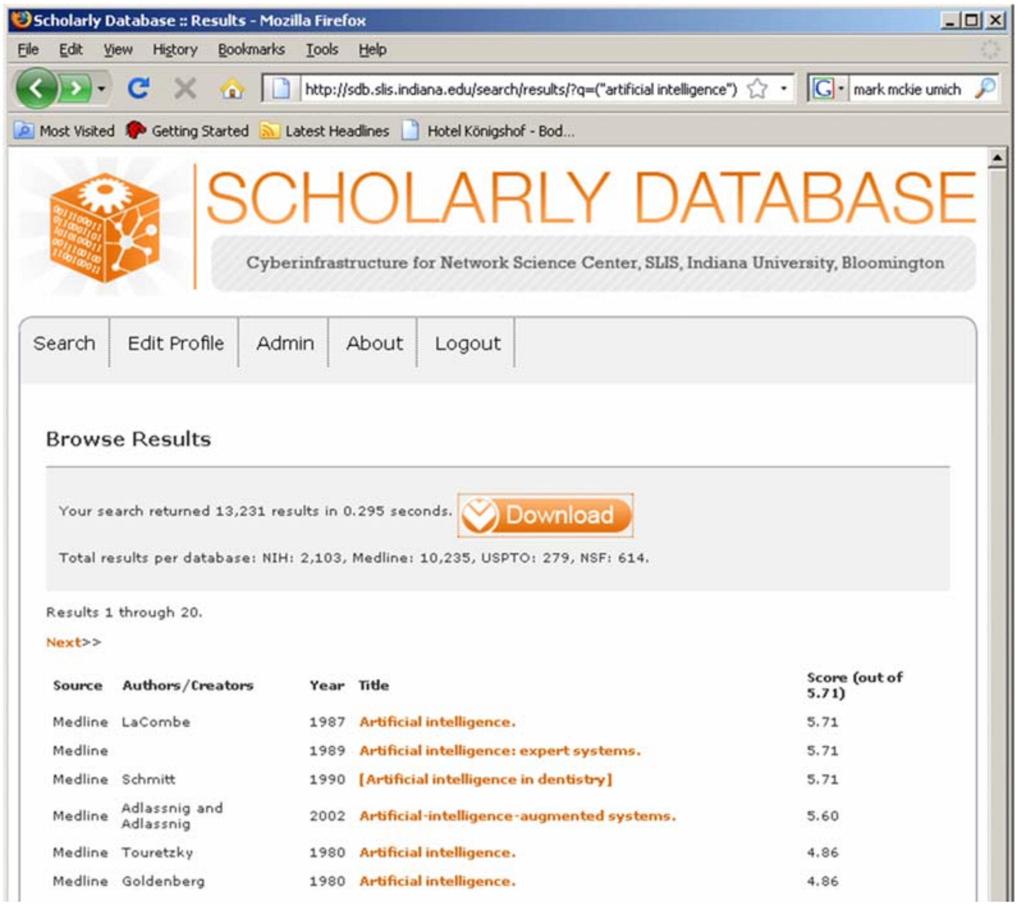
For many researchers, the effort to compile ready-to-analyze-and-visualize data is extremely time consuming and challenging and sometimes simply insurmountable.

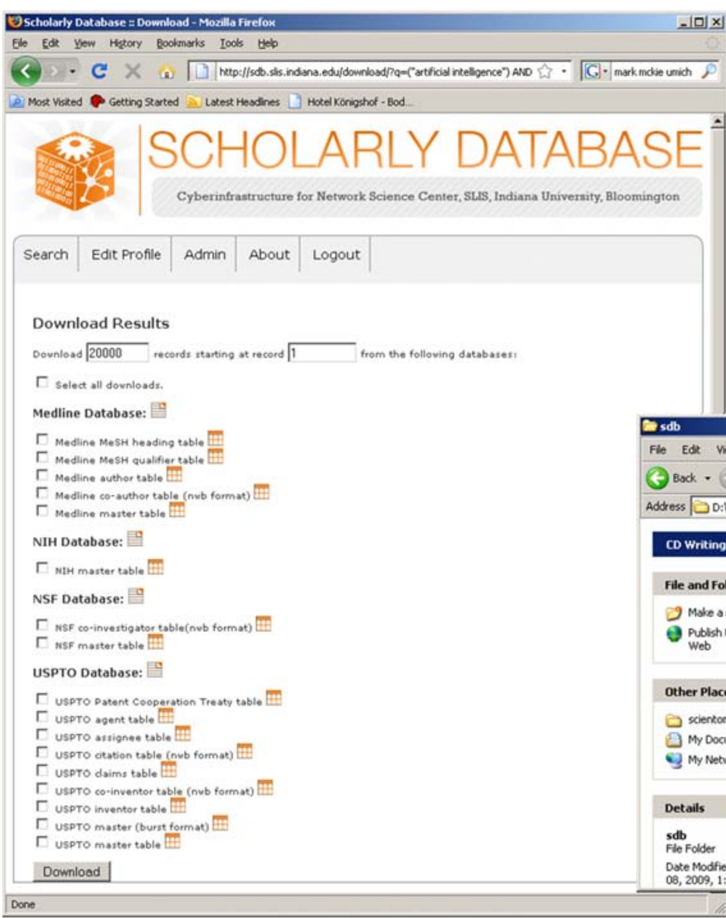
Scholarly Database at Indiana University
http://sdb.wiki.cns.iu.edu

Supports federated search of 26 million publication, patent, clinical trials, and grant records. Results can be downloaded as data dump and (evolving) co-author, paper-citation networks.



Register for free access at http://sdb.cns.iu.edu





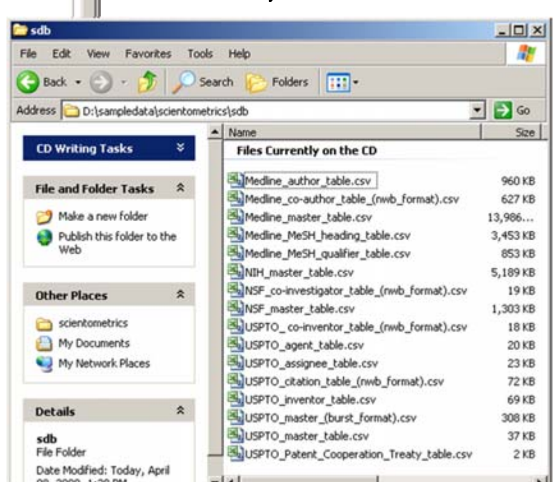
Since March 2009:

Users can download networks:

- Co-author
- Co-investigator
- Co-inventor
- Patent citation

and tables for

burst analysis in NWB.



SDB: Unique Features

- *Open Access:* SDB is free to researchers. No copyright or proprietary issues.
- *Ease of Use:* One-stop data access experience reducing the time spent on parsing, searching, and formatting data=more time for research!
- *Federated Search* across datasets powered by a Solr index.
- *Bulk Download* of data records; data linkages—co-author, patent citations, grant-paper, grant-patent; burst analysis files.
- *Unified File Formats:* SDB source data comes in different file formats but can be downloaded in easy-to-use file formats, e.g., comma-delimited tables for use in spreadsheet programs and common graph formats for network analysis and visualization.
- *Well-Documented:* SDB publishes data dictionaries, sample files, baseline stats, see SDB Wiki at http://sdb.wiki.cns.iu.edu.

Goals of the Paper/Structure of this Talk

Inspire the development of "Open Data and Open Code for BIG Science of Science Studies" **see ISSI 2013 Workshop on "Standards for Science Mapping and Classifications"**

Introduce a database-tool infrastructure designed to support big SoS studies:

- The open access Scholarly Database (SDB) (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records.
- The open source Science of Science (Sci2) tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies, **see ISSI 2013 Tutorial on Workshop on "Sci2: A Tool of Science of Science Research and Practice"**

Showcase scalability of the infrastructure:

- temporal analyses scale linearly with the number of records and file size.
- geospatial algorithm show quadratic growth.
- network science algorithms scale with the number of edges rather than nodes.

Sci² Tool: Download, Install, and Run

Sci2 Tool v1.0 Alpha (June 13, 2012)

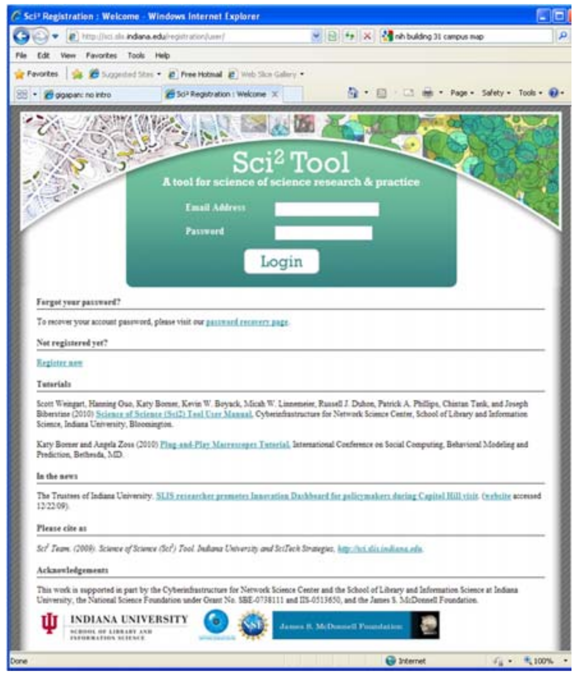
Can be freely downloaded for all major operating systems from http://sci2.cns.iu.edu

Select your operating system from the pull down menu and download.
Unpack into a /sci2 directory.
Run /sci2/sci2.exe

Sci2 Manual is at http://sci2.wiki.cns.iu.edu

Cite as

Sci² Team. (2009). Science of Science (Sci²) Tool. Indiana University and SciTech Strategies, http://sci2.cns.iu.edu

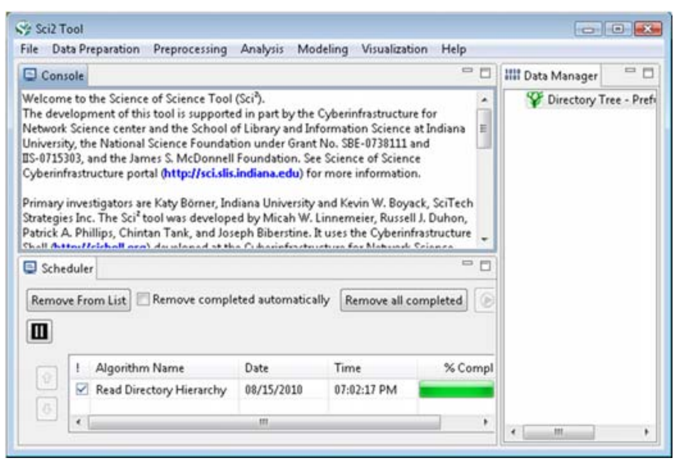


Sci2 Tool Interface Components

See also http://sci2.wiki.cns.iu.edu/2.2+User+Interface

Use

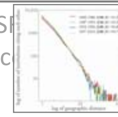
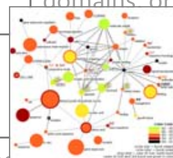
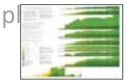
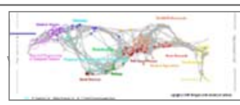
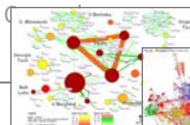
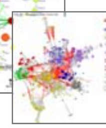
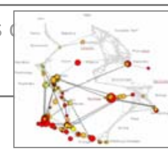
- **Menu** to read data, run algorithms.
- **Console** to see work log, references to seminal works.
- **Data Manager** to select, view, save loaded, simulated, or derived datasets.
- **Scheduler** to see status of algorithm execution.



All workflows are recorded into a log file (see /sci2/logs/…), and soon can be re-run for easy replication. If errors occur, they are saved in a error log to ease bug reporting.

All algorithms are documented online; workflows are given in tutorials, see Sci2 Manual at http://sci2.wiki.cns.iu.edu

Type of Analysis vs. Level of Analysis

	Micro/Individual (1-100 records)	***Meso/Local (101–10,000 records)***	***Macro/Global (10,000 < records)***
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NS ... A, all of scienc ... 
Temporal Analysis (When)	Funding portfolio of one individual	 ...ic bursts ...f *PNAS*	113 years of p... research 
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a s... intellectual la... 	*PNAS* p...  
Topical Analysis (What)	 ...s.	Knowledge f... chemistry research	
Network Analysis (With Whom?)	NSF ... ork of one 	C... k  	NIH's ... cy 

Sci² Tool – Supported Data Formats

Input:

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

Scientometric Formats
- ISI (*.isi)
- Bibtex (*.bib)
- Endnote Export Format (*.enw)
- Scopus csv (*.scopus)
- NSF csv (*.nsf)

Other Formats
- Pajek Matrix (*.mat)
- TreeML (*.xml)
- Edgelist (*.edge)
- CSV (*.csv)

Output:

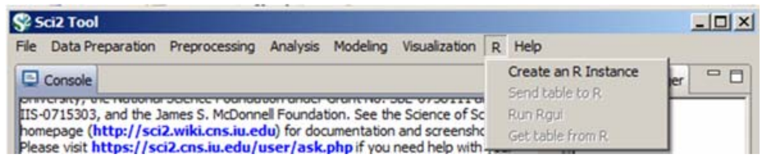
Network File Formats
- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)

Image Formats
- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps)

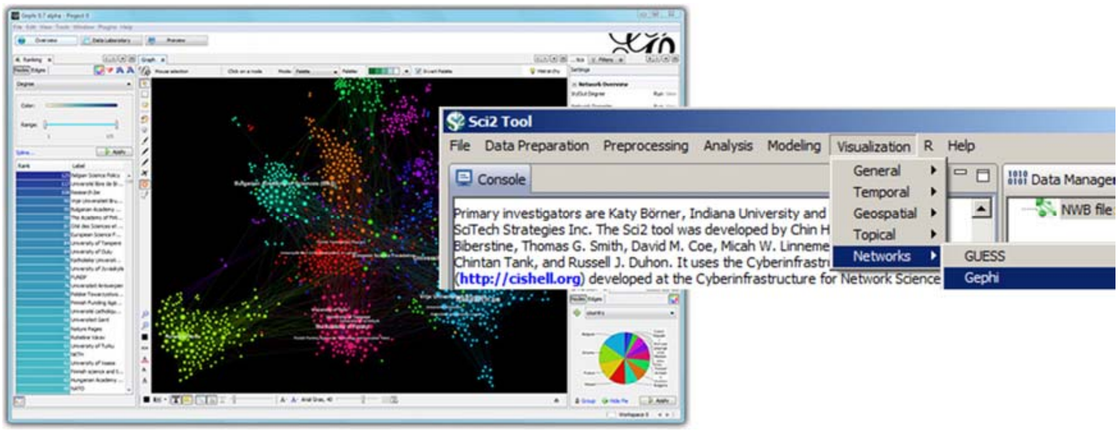
Formats are documented at http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.3+Data+Formats.

Sci² Tool – Bridged Tools

R statistical tool bridging



Gephi visualization tool bridging



Sci2 Tool v1.0 Alpha (June 13, 2012)

Major Release
featuring a Web services compatible CIShell v2.0 (http://cishell.org)

New Features

- Google Scholar citation reader
- New visualizations such as
  - geospatial maps
  - science maps
  - bi-modal network layout
- R statistical tool bridging
- Gephi visualization tool bridging
- Comprehensive online documentation

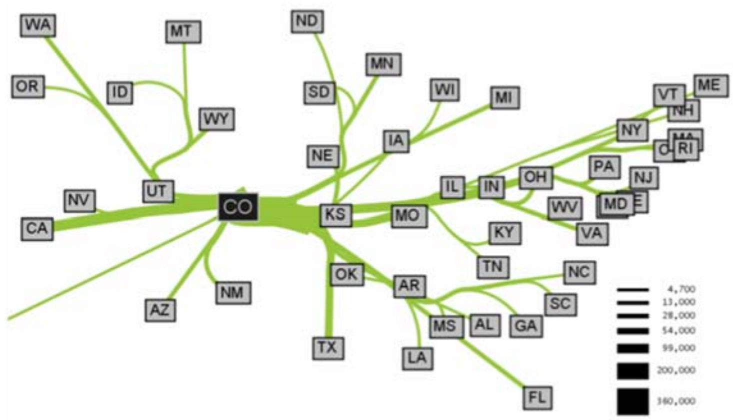
Release Note Details
http://wiki.cns.iu.edu/display/SCI2TUTORIAL/4.4+Sci2+Release+Notes+v1.0+alpha

Sci2 Tool v1.1 Alpha (planned for August 2013)

New Features

- Twitter, Facebook, and Flickr readers
- Bing Geocoder
- Flow map visualization, see below
- Comprehensive online documentation

Bug fixes



OSGi/CIShell Adoption

A number of other projects recently adopted OSGi and/or CIShell:

USA

- *Cytoscape (http://cytoscape.org)* Led by Trey Ideker at the University of California, San Diego 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).
- *MAEviz (https://wiki.ncsa.uiuc.edu/display/MAE/Home)* Managed by Jong Lee at NCSA is an open-source, extensible software platform which supports seismic risk assessment based on the Mid-America Earthquake (MAE) Center research.

Europe

- *Taverna Workbench (http://taverna.org.uk)* Developed by the myGrid team (http://mygrid.org.uk) led by Carol Goble at the University of Manchester, U.K. 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.
- *TEXTrend (http://textrend.org)* Led by George Kampis at Eötvös Loránd University, Budapest, Hungary supports natural language processing (NLP), classification/mining, and graph algorithms for the analysis of business and governmental text corpuses with an inherently temporal component.
- *DynaNets (http://www.dynanets.org)* Coordinated by Peter M.A. Sloot at the University of Amsterdam, The Netherlands develops algorithms to study evolving networks.
- *SISOB (http://sisob.lcc.uma.es)* An Observatory for Science in Society Based in Social Models.

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 will expand.

CIShell – Integrate New Algorithms

About the Cyberinfrastructure Shell

The Cyberinfrastructure Shell (CIShell) is an open source, community-driven platform for the integration and utilization of datasets, algorithms, tools, and computing resources. Algorithm integration support is built in for Java and most other programming languages. Being Java based, it will run on almost all platforms. The software and specification is released under an Apache 2.0 License.

CIShell is the basis of Network Workbench, TexTrend, Sci² and the upcoming EpiC tool.

CIShell supports remote execution of algorithms. A standard web service definition is in development that will allow pools of algorithms to transparently be used in a peer-to-peer, client-server, or web front-end fashion.

CIShell Features

A framework for easy integration of new and existing algorithms written in any programming language

Using CIShell, an algorithm writer can fully concentrate on creating their own algorithm in whatever language they are comfortable with. Simple tools are provided to then take their algorithm and

Learn More...

- CIShell Papers
- CIShell Powered Tools
- Algorithms
- Plugins (coming soon)
- Misc. Tool Documentation
- CIShell Web Services (coming soon)
- Screenshots

Getting Started...

- Documentation & Developer Resources
- Download

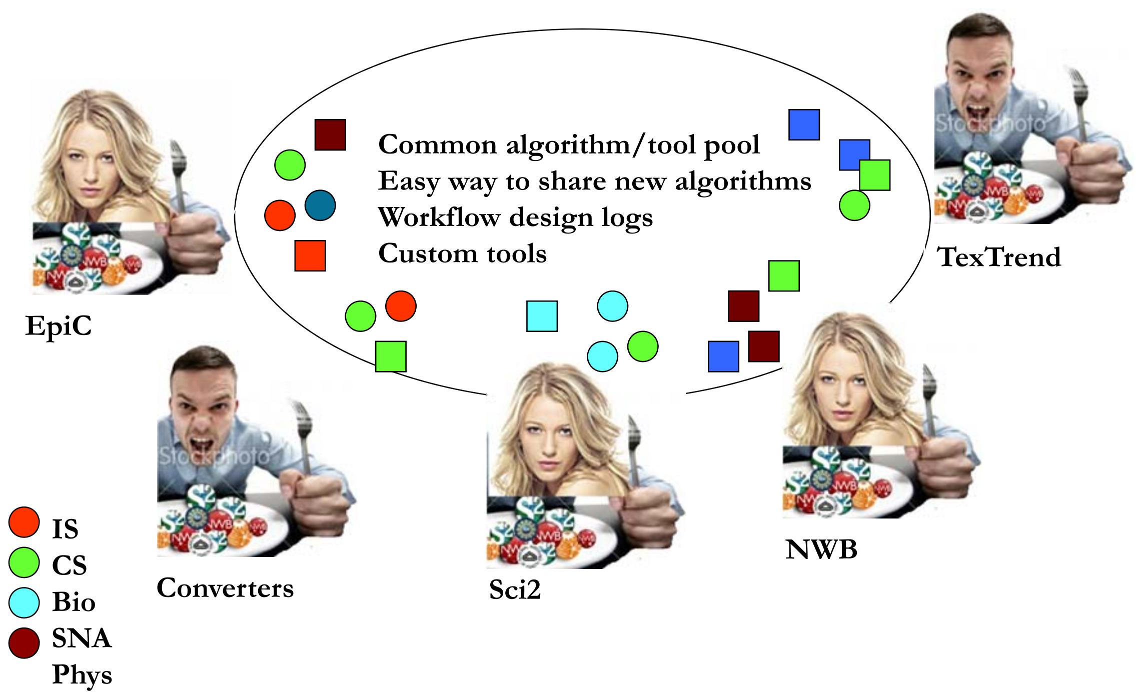
Getting Involved...

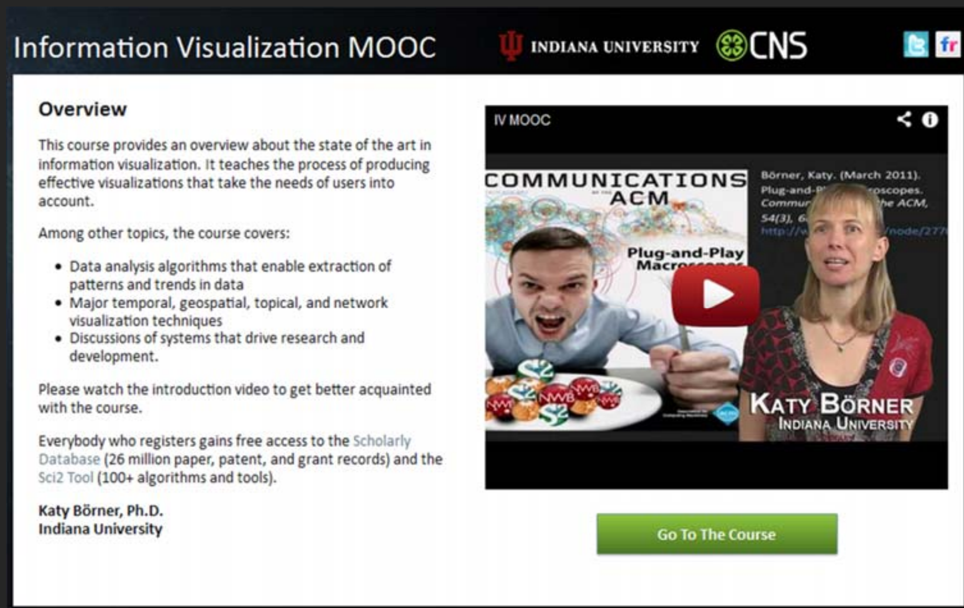
- Contact Us

CIShell Developer Guide is at http://cishell.wiki.cns.iu.edu

Additional Sci2 Plugins are at http://sci2.wiki.cns.iu.edu/3.2+Additional+Plugins

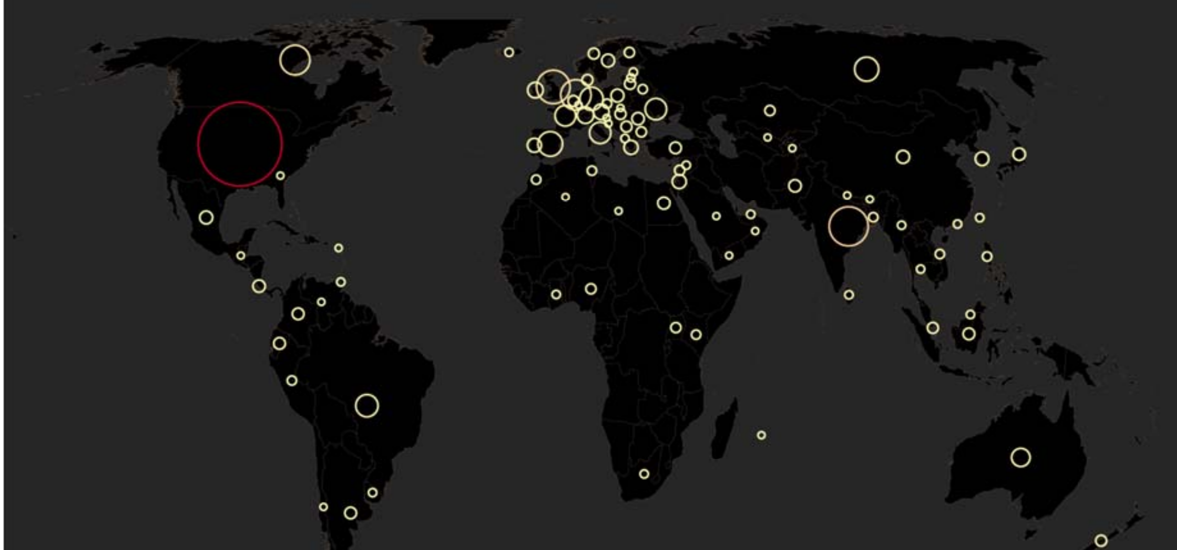
OSGi/CIShell-Powered Tools Support Algorithm Sharing





ivmooc.cns.iu.edu

The Information Visualization MOOC
ivmooc.cns.iu.edu



Students come from 93 countries
300+ faculty members
#ivmooc

Course Schedule

Course started on January 22, 2013

- **Session 1** – Workflow design and visualization framework
- **Session 2** – “When:” Temporal Data
- **Session 3** – “Where:” Geospatial Data
- **Session 4** – “What:” Topical Data

Mid-Term

Students work in teams with clients.

- **Session 5** – “With Whom:” Trees
- **Session 6** – “With Whom:” Networks
- **Session 7** – Dynamic Visualizations and Deployment

Final Exam

Grading

All students are asked to create a personal profile to support working in teams.

Final grade is based on Midterm (**30%**), Final (**40%**), Client Project (**30%**).

- Weekly self-assessments are not graded.
- Homework is graded automatically.
- Midterm and Final test materials from theory and hands-on sessions are graded automatically.
- Client work is peer-reviewed via online forum.

All students that receive more than **80%** of all available points get an official certificate/badge.

Goals of the Paper/Structure of this Talk

Inspire the development of "Open Data and Open Code for BIG Science of Science Studies" **see ISSI 2013 Workshop on "Standards for Science Mapping and Classifications"**

Introduce a database-tool infrastructure designed to support big SoS studies:

- The open access Scholarly Database (SDB) (http://sdb.cns.iu.edu) provides easy access to 26 million paper, patent, grant, and clinical trial records.
- The open source Science of Science (Sci2) tool (http://sci2.cns.iu.edu) supports temporal, geospatial, topical, and network studies, **see ISSI 2013 Tutorial on Workshop on "Sci2: A Tool of Science of Science Research and Practice"**

Showcase scalability of the infrastructure:

- temporal analyses scale linearly with the number of records and file size.
- geospatial algorithm show quadratic growth.
- network science algorithms scale with the number of edges rather than nodes.

Scalability Types

Data Scalability

- Most tools work well for **micro** and **meso** level studies (up to 100,000 records).
- Few scale to **macro** level big-data studies with millions or even billions.

Analysis Scalability

Many data mining algorithms have a high complexity, e.g., betweenness centrality is $O(n^3)$, pathfinder network scaling $O(n^2)$- $O(n^4)$, Fruchterman-Reingold layout $O(n^2)$ per iteration. Do you know the complexity of the algorithms you use? **How many of you use parallel computing?**

Visual Scalability (ease of use and ease of interpretation)

- How to communicate temporal trends/activity burst over a 100 year time span?
- How to depict the geospatial or topical locations of millions of records?
- Most visualizations of million node networks resemble illegible spaghetti balls—do advanced network analysis algorithms scale and help to derive insights?

Scalability: Four Exemplary Workflows

Each consists of several steps:

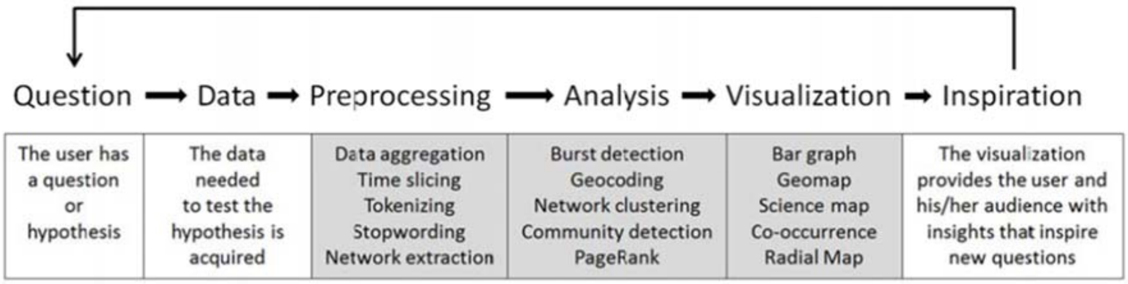


Figure 2: General Sci2-based visualization creation workflow (tool-specific tasks in gray).

Overall run time is strongly impacted by the slowest algorithm!

Scalability: Data Used & Process

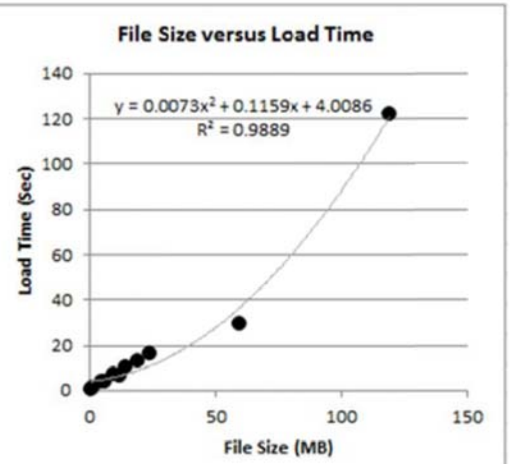
- Synthetic datasets with pre-defined properties generated in Python.
- Datasets retrieved from the Scholarly Database:
  - NIH data at 3.4GB, NSF data at 489MB, NIH data at 139MB, and NEH data at 12.1MB data prepared for temporal analysis.
  - Data from NIH, NSF, MEDLINE, UPSTO, and Clinical Trials at 11.5 MB and MEDLINE data at 1GB to be used in geospatial analysis.
  - MEDLINE data at 514KB for topical analysis.
  - NSF data at 11.9MB and UPSTO data at 1.04GB network analysis.

- For each test, we calculated the average for 10 trials.
- Tests were performed on a common system: an Intel(R) Core(TM) Duo CPU E8400 3.00GHz processor and 4.0GB of memory running a 64bit version of Windows 7 and a 32bit version of Java 7. Memory allotted to Sci2 was extended to 1500 MB.

Scalability: Data Load Times
Synthetic Datasets

Obviously data load time depends on the number of records and file size.

Records	Columns	Size (MB)	Load Time (sec)	SD (sec)
50,000	2	0.48	0.72	0.06
100,000	2	0.95	1.08	0.04
500,000	2	4.77	3.75	0.05
1,000,000	2	9.54	7.14	0.14
1,500,000	2	14.31	10.26	0.08
2,000,000	2	19.07	13.26	0.17
2,500,000	2	23.84	16.47	0.13
50,000	25	5.96	3.56	0.07
100,000	25	11.92	6.44	0.05
500,000	25	59.61	29.62	0.91
1,000,000	25	119.21	122.36	0.64
1,500,000	25	178.81	-TF*	
2,000,000	25	238.42	-TF*	
2,500,000	25	298.02	-TF*	



Java heap space error (-TF*)

Figure 5: Comparison of load times, measured in seconds, across standardized datasets, tabulated (left) and plotted with quadratic regression line (right).

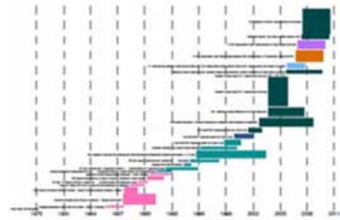
Scalability: Data Load Times
SDB Datasets

Table 1: Comparison of load times, measured in seconds, across nine different datasets.

Dataset	Size	Number of Records	Mean	Standard Deviation	Minimum	Maximum
NIH (year, title, abstract)	3.4GB	2,490,837	-TF*			
USPTO (patent, citations)	1.04GB	57,902,504	-TF*			
MEDLINE (geospatial)	1.0GB	9,646,117	-TF*			
NSF (year, title, abstract)	489MB	453,740	64.54	0.991	63.2	65.9
NIH (title, year)	139MB	2,490,837	83.86	1.32	82.3	85.6
NEH (year, title, abstract)	12.1MB	47,197	2.05	0.070	1.9	2.1
NSF (co-author network)	11.9MB	341,110	4.52	0.063	4.4	4.6
Combined geo-spatial	11.5MB	11,549	1.91	0.056	1.8	2.0
MEDLINE journals	0.5MB	20,775	0.44	0.096	0.3	0.6

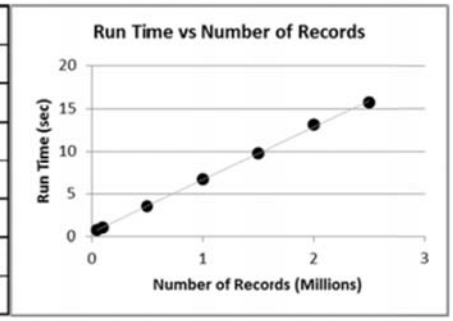
Scalability: Burst Analysis

Synthetic & SDB Datasets



Highly scalable:

Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	0.48	0.75	0.07
100,000	0.95	1.03	0.05
500,000	4.77	3.55	0.07
1,000,000	9.54	6.67	0.07
1,500,000	14.31	9.76	0.18
2,000,000	19.07	13.15	0.17
2,500,000	23.84	15.73	0.22



NIH: Lowercase, Tokenize, Stem, and Stopword Text algorithm failed to terminate .

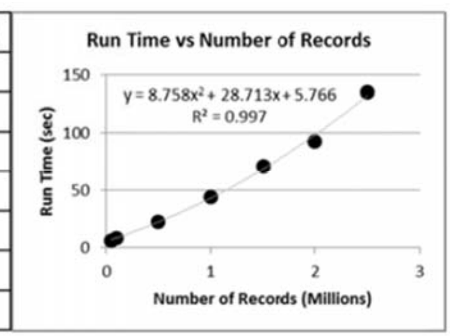
Burst Detection						
Dataset	**Size**	**Rows**	**Mean**	**SD**	**Min**	**Max**
NSF	489 MB	453,740	13.64	0.648	12.9	14.8
NIH	139 MB	2,490,837	-NT*			
NEH	12.1 MB	47,197	1.57	0.094	1.4	1.7

Scalability: Geospatial Map

Synthetic & SDB Datasets

Highly scalable (but about 10x slower than burst).

Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	1.82	6.26	0.25
100,000	3.66	8.86	0.45
500,000	18.71	22.71	2.00
1,000,000	37.52	44.37	5.21
1,500,000	56.81	70.73	2.15
2,000,000	76.09	92.93	5.63
2,500,000	95.38	134.69	2.78



11,848 SDB records related to gene therapy funding (NIH, NSF), publications (MEDLINE), patents (USPTO), and clinical trials were geolocated. 299 records had no geolocation data and were removed resulting in 11,549 rows at 11.5MB.

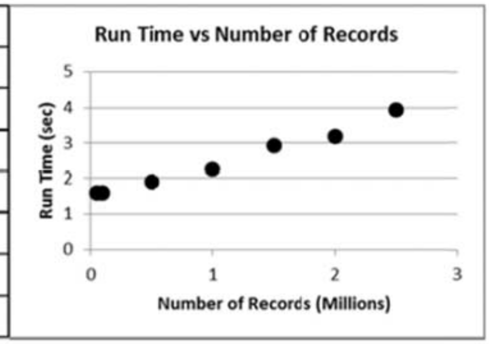
Dataset	Size	Rows	Mean	SD	Min	Max
Pre-located	11.5 MB	11,549	4.37	0.125	4.2	4.6

Scalability: UCSD Science Map

Synthetic & SDB Datasets

Highly scalable (and about 5x FASTER than burst).

Records	Size (MB)	Run Time (sec)	SD (sec)
50,000	1.33	1.59	0.13
100,000	2.67	1.58	0.09
500,000	13.79	1.89	0.09
1,000,000	27.66	2.25	0.07
1,500,000	42.02	2.92	0.16
2,000,000	56.40	3.19	0.03
2,500,000	70.77	3.93	0.26

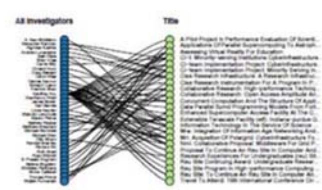


MEDLINE data was obtained from SDB comprising all 20,773 unique journals indexed in MEDLINE and the number of articles published in those journals.

Dataset	Size	Rows	Mean	SD	Min	Max
MEDLINE journals	514 KB	20,773	7.84	0.096	7.7	8.0

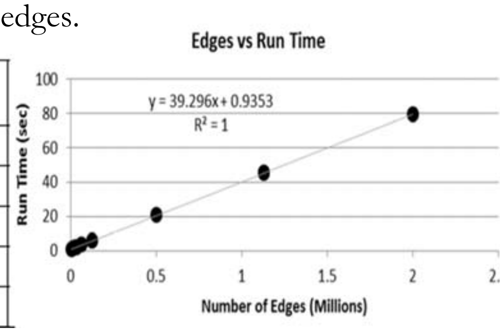
Scalability: Network

Synthetic Dataset



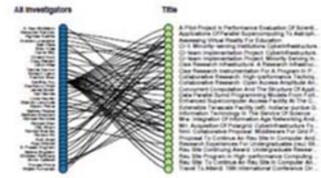
Complexity depends on number of nodes and edges.

Records	% Conn	Edges	Size (MB)	Run (sec)	SD (sec)
500	2	5,000	0.017	1.13	0.05
500	5	12,500	0.045	1.44	0.07
500	10	25,000	0.093	1.92	0.04
500	25	62,500	0.247	3.46	0.08
500	50	125,000	0.546	5.89	0.1



Records	% Conn	Edges	Size (MB)	Run (sec)	SD (sec)
250	50	31,250	0.124	1.86	0.05
500	50	125,000	0.546	5.89	0.1
1,000	50	500,000	2.28	20.74	0.12
1,500	50	1,125,000	5.21	45.28	0.44
2,000	50	2,000,000	9.33	79.41	0.62

Scalability: Network
SDB Dataset



All 6,206 USPTO patents that cite patents with numbers 591 and 592 in the patent number field were retrieved.

Extract Network:

Extract Directed Network algorithm was run, creating a network pointing from the patent numbers to the numbers those patents reference in the dataset.

Dataset	Size in MB	Nodes	Edges	Mean	SD	Min	Max
U.S. Patent References	0.147	12,672	7,940	7.88	0.103	7.7	8.1

Layout:

Neither Cytoscape nor GUESS could render the network in a Fruchterman-Reingold layout.

Gephi loaded the network in 2.1 seconds and rendered it in about 40 seconds—due to its ability to leverage GPUs in computing intensive tasks.

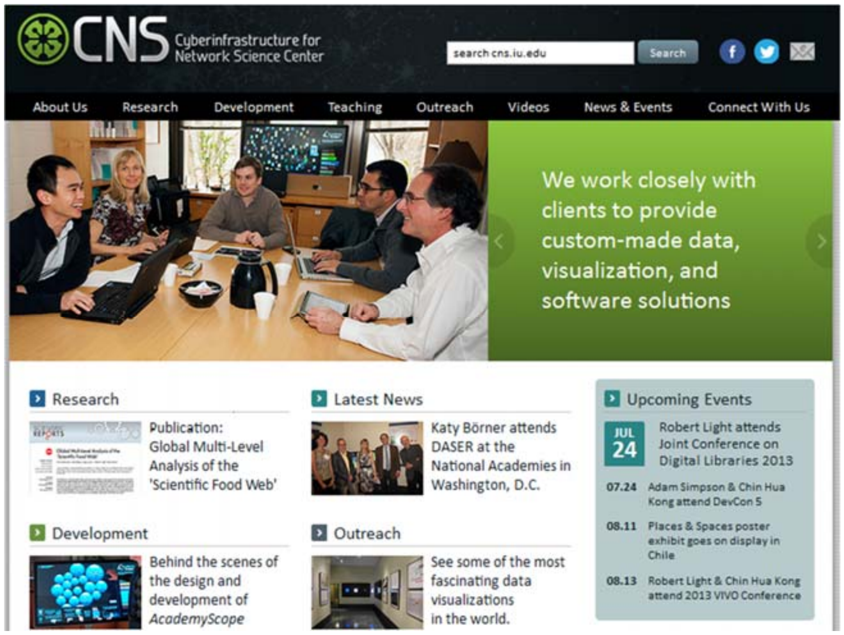
Scalability: Discussion & Outlook

- Most run-times scale linearly or exponentially with file size.
- The number of records impacts run-time more than file size.
- Files larger than 1.5 million records (synthetic data) and 500MB (SDB) cannot be loaded and hence not be analyzed or visualized.
- Run times for rather large datasets are commonly less than 10 seconds.
- Only large datasets combined with complex analysis require more than one minute to execute.

Scalability tests are time consuming, this paper took more than 1000 workflow runs.

They are important to understand, optimize, improve time complexity.

The Sci2 Tool and selected workflows can now be run as Web services and a similar study is desirable for those.



All papers, maps, tools, talks, press are linked from http://cns.iu.edu

CNS Facebook: http://www.facebook.com/cnscenter
Mapping Science Exhibit Facebook: http://www.facebook.com/mappingscience