

Scholarly Big Data: CiteSeerX Insights

C. Lee Giles

The Pennsylvania State University

University Park, PA, USA

giles@ist.psu.edu

<http://clgiles.ist.psu.edu>

Funded in part by NSF & Qatar.

Contributors/Collaborators: recent past and present (incomplete list)

Projects: CiteSeer, CiteSeer^x, Chem_xSeer, ArchSeer, CollabSeer, GrantSeer, SeerSeer, RefSeer, CSSeer, AlgoSeer, AckSeer, BotSeer, YouSeer, ...

- P. Mitra, V. Bhatnagar, L. Bolelli, J. Carroll, I. Councill, F. Fonseca, J. Jansen, D. Lee, W-C. Lee, H. Li, J. Li, E. Manavoglu, A. Sivasubramaniam, P. Teregowda, J. Yen, H. Zha, S. Zheng, D. Zhou, Z. Zhuang, J. Stribling, D. Karger, S. Lawrence, K. Bollacker, D. Pennock, J. Gray, G. Flake, S. Debnath, H. Han, D. Pavlov, E. Fox, M. Gori, E. Blanzieri, M. Marchese, N. Shadbolt, I. Cox, S. Gauch, A. Bernstein, L. Cassel, M-Y. Kan, X. Lu, Y. Liu, A. Jaiswal, K. Bai, B. Sun, Y. Sung, Y. Song, J. Z. Wang, K. Mueller, J. Kubicki, B. Garrison, J. Bandstra, Q. Tan, J. Fernandez, P. Treeratpituk, W. Brouwer, U. Farooq, J. Huang, M. Khabsa, M. Halm, B. Urgaonkar, Q. He, D. Kifer, J. Pei, S. Das, S. Kataria, D. Yuan, S. Choudhury, H-H. Chen, N. Li, D. Miller, A. Kirk, W. Huang, S. Carman, J. Wu, L. Rokach, C. Caragea, K. Williams. Z. Wu, S. Das, A. Ororbia, others.

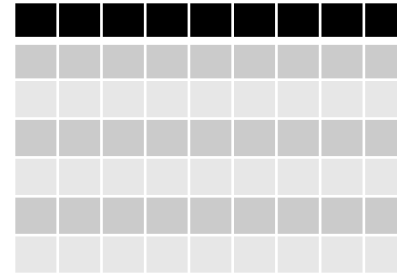
What is Scholarly Big Data

All academic/research documents (journal & conference papers, books, theses, TRs)

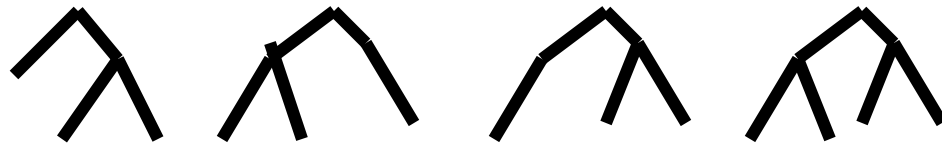
- Related data:
 - Academic/researcher/group/lab web homepages
 - Funding agency and organization grants, records, reports
 - Research laboratories reports
 - Patents
 - Associated data
 - presentations
 - experimental data (very large)
 - video
 - course materials
 - other
 - Social networks
- Examples: **Google Scholar**, *Microsoft Academic Search*, Publishers/repositories, CiteSeer, ArnetMiner, Funding agencies, Universities, Mendeley, others

Scholarly Big Data

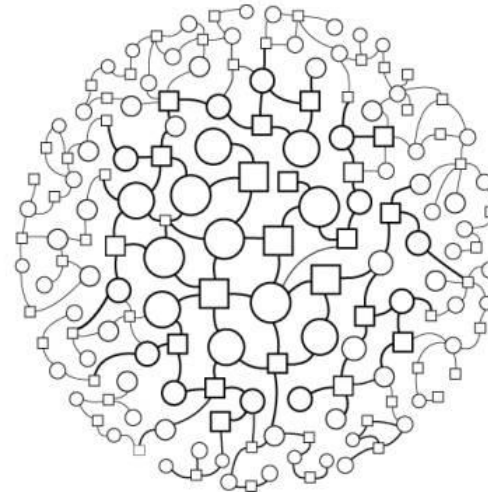
Most of the data that is available in the era of scholarly big data does not look like this



Or even like this



It looks more like this



Courtesy Lise Getoor NIPS'12

Where do you get this data?

- Web (Wayback machine, Heritrix)
- Repositories (arXiv, CiteSeer)
- Bibliographic resources (PubMed, DBLP)
- Funding sources/laboratories
- Publishers
- Data aggregators (Web of science)
- Patents
- API's (Microsoft Academic)

How much is there & how much freely available?

Estimate of Scholarly Big Data

- Use two public academic search engines
 - estimate the size of scholarly articles on the web using capture/recapture (Lincoln Petersen) methods
 - Google Scholar
 - Microsoft Academic Search
- Find a paper that both search engines have
- Extract the list of citations for that paper in both search engines and compare overlap
- The list of citations for a paper is representative of the coverage of a search engine
- Using the size of one of the search engines, estimate the total size on the web
- Limit to English articles only

Consider the web page coverage of search engines a and b

- p_a probability that engine a has indexed a page, p_b for engine b , $p_{a,b}$ joint probability

$$p_{a,b} = p_{a|b} p_b \geq p_a p_b$$

- s_a number of unique pages indexed by engine a ; N number of web pages

$$p_a = \frac{s_a}{N} \quad \frac{s_{a,b}}{N} \geq \frac{s_a}{N} \frac{s_b}{N} \quad N \geq s_a \frac{s_b}{s_{a,b}}$$

- n_b number of documents returned by b for a query, $n_{a,b}$ number of documents returned by both engines a & b for a query

$$\left\langle \frac{s_b}{s_{a,b}} \right\rangle \cong \left\langle \frac{n_b}{n_{a,b}} \right\rangle_{\text{queries}}$$

Lower bound estimate of size of the Web:

$$\hat{N} \geq s_{a_o} \left\langle \frac{n_b}{n_{a,b}} \right\rangle_{\text{queries}} ; s_{a_o} \text{ known}$$

- random sampling assumption
- extensions - bayesian estimate, more engines (Bharat, Broder, WWW7 '98), etc.

Freely available by scholarly field

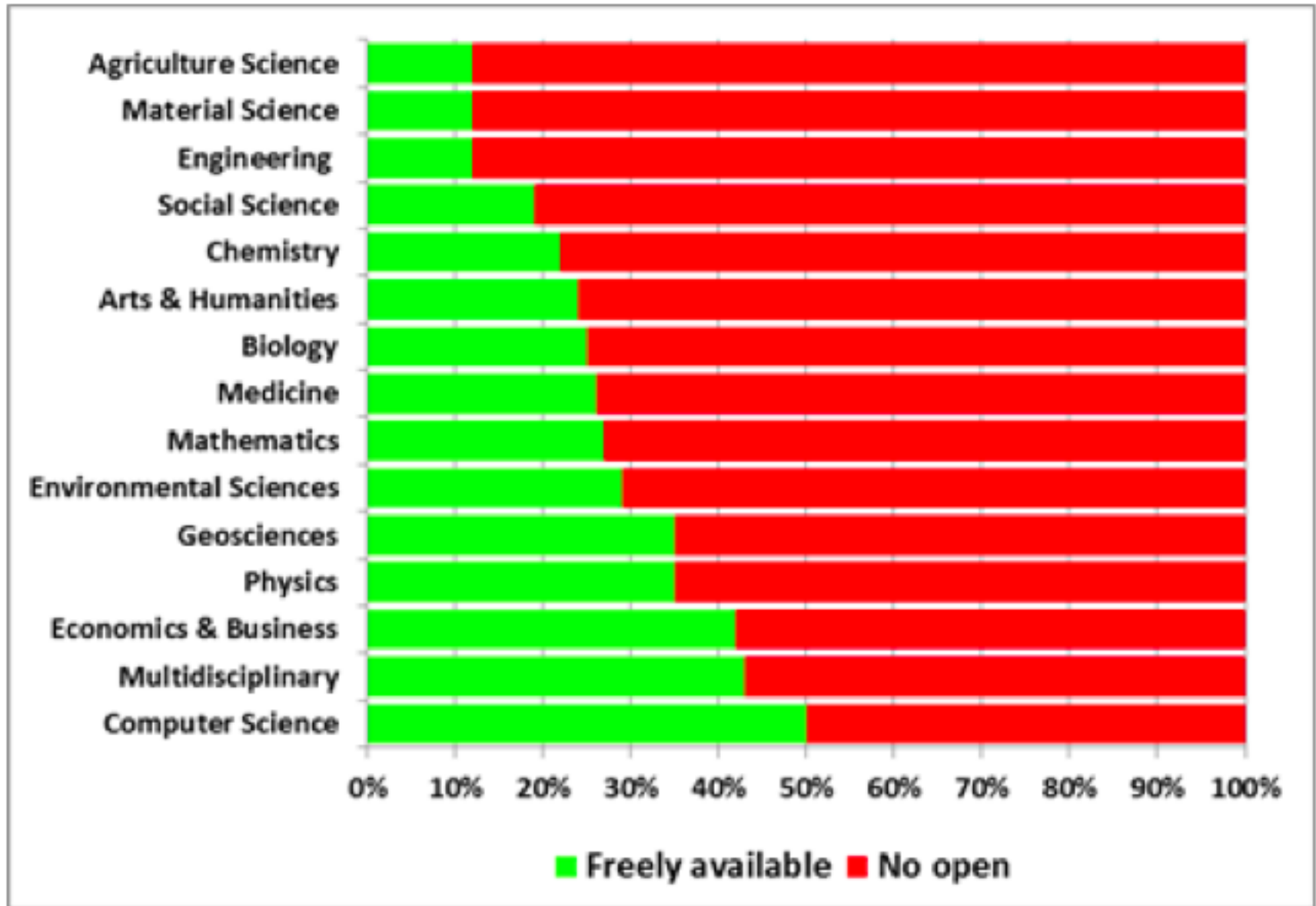


Figure 4. Percentage of publicly available documents according to scientific fields

Data source: re-elaborated from Khabsa & Giles (2014)

Lower bound on potential sources of scholarly data

- At least 114 million scholarly articles available on the web
- At least 24% of them are publicly available
 - 27 million
 - Varies significantly based on academic field
 - Computer science!
- Google Scholar has nearly 100 million articles
- Other things to do:
 - Distinguish between publication types: paper, thesis, tech report, etc
 - More estimates
 - Longitudinal and geographical study
 - Duplicates
 - Languages besides english

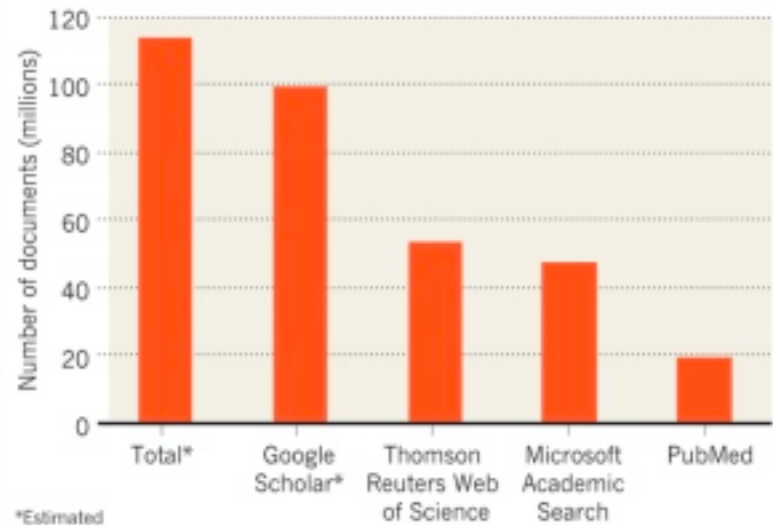
Seven days: 16–22 May 2014

TREND WATCH

The academic search engine Google Scholar can find about 88% of all English-language scholarly documents on the World Wide Web, according to an estimate by computer scientists Lee Giles and Madian Khabza at Pennsylvania State University in University Park (M. Khabza and C. L. Giles *PLoS ONE* 9, e93949; 2014). The duo studied the coverage of Google Scholar and a competitor, Microsoft Academic Search. At least 24% of documents are freely available, they add. See go.nature.com/matsio for more.

THE WEB OF SCHOLARSHIP

Around 114 million English-language scholarly documents (including papers, books and technical reports) can be found on the web.



Source: *PLoS ONE*/thomson reuters



IARPA FUSE Program

Understanding Federal R&D Impact
Through Research Assessment and Program Evaluation

Panel: Increasing Research Impact Through Effective Planning and Evaluation

OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE



Finding Patterns of Emergence in Science and Technology – Evaluation Implications

Foresight and Understanding from Scientific Exposition (FUSE)

LEADING INTELLIGENCE INTEGRATION

Dewey Murdick, Program Manager
Office of Incisive Analysis, IARPA
19 March 2013

INTELLIGENCE ADVANCED RESEARCH PROJECTS ACTIVITY (IARPA)

IARPA FUSE Program



OFFICE OF THE DIRECTOR OF NATIONAL INTELLIGENCE
LEADING INTELLIGENCE INTEGRATION



Goal: Validated, early detection of technical emergence

Enable reliable, early detection of emerging scientific and technical capabilities across disciplines and languages found within the full-text content of scientific, technical, and patent literature

Special focus from the outset on multiple languages

- | | |
|----------------|--|
| Novelty | → Discover <u>patterns</u> of emergence and <u>connections</u> between technical concepts at a speed, scale, and comprehensiveness that exceeds human capacity |
| Usage | → <u>Alert analyst</u> of emerging technical areas with sufficient explanatory evidence to support further exploration |

SeerSuite Toolkit & Applications

- SeerSuite: open source search engine and digital library tool kit used to build academic search engines/digital libraries
 - **CiteSeer^x** , **Chem_xSeer**, **AckSeer**, **CSSeer**, **CollabSeer**, **RefSeer**, etc.
 - Built on commercial grade open source tools (**Solr/Lucene**; **mySQL**)
 - Our features – automated specialized metadata extraction
 - Information extraction tools for PDF documents
 - Authors, titles, affiliations, citations, acknowledgements, etc
 - Entity disambiguation
 - Tabular data
 - Figures & graphs
 - Chemical formulae
 - Equations
 - Author ethnicity detection
 - Data and search built on top of these
 - CiteSeerX – open source Google Scholar (ScholarSeer)
 - ChemXSeer – chemical search engine
 - RefSeer – citation recommendation system
 - CSSeer – expert recommendations
 - CollabSeer- collaboration recommendation
 - AckSeer – acknowledgement search
 - ArchSeer – archaeology map search
- Wu, IAAI 2014
Teregowda, IC2E 2013
Teregowda, USENIX 2010

CiteSeer (aka ResearchIndex)

- Project of NEC Research Institute
- Hosted at Princeton, from 1997 – 2004
- Moved to Penn State after collaborators left NEC
- Provided a broad range of unique services including
 - Automatic metadata extraction
 - Autonomous citation indexing
 - Reference linking
 - Full text indexing
 - Similar documents listing
 - Several other pioneering features
- Impact
 - First scholarly search engine?
 - Changed access to scientific research
 - Shares code and data



C. Lee Giles



Kurt Bollacker



Steve Lawrence

CiteSeer^X

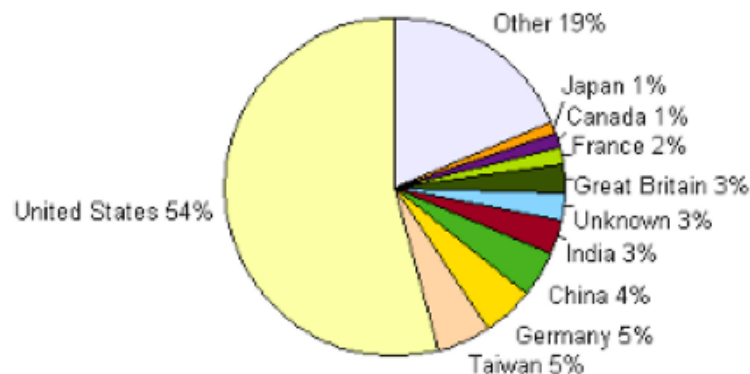
<http://citeseerx.ist.psu.edu>

- CiteSeer^X **actively** crawls researcher homepages & archives on the web for scholarly papers, formerly in computer science
 - Converts PDF to text
 - Automatically extracts and tags OAI metadata and **other data**
 - Automatic citation indexing, links to cited documents, creation of document page, author disambiguation
 - Software open source – can be used to build other such tools
 - All data shared

- 5+ M documents
- Ms of files
- 87 M citations
- 12 M authors
 - 1.3 M disambig
- 2 to 4 M hits day
- 100K documents added monthly
- 300K document downloaded monthly
- 800K individual users
- ~40 Tbytes

Search

Include Citations [Advanced Search](#)



Cite
Seer
X BETA

Focused Crawling – getting the documents

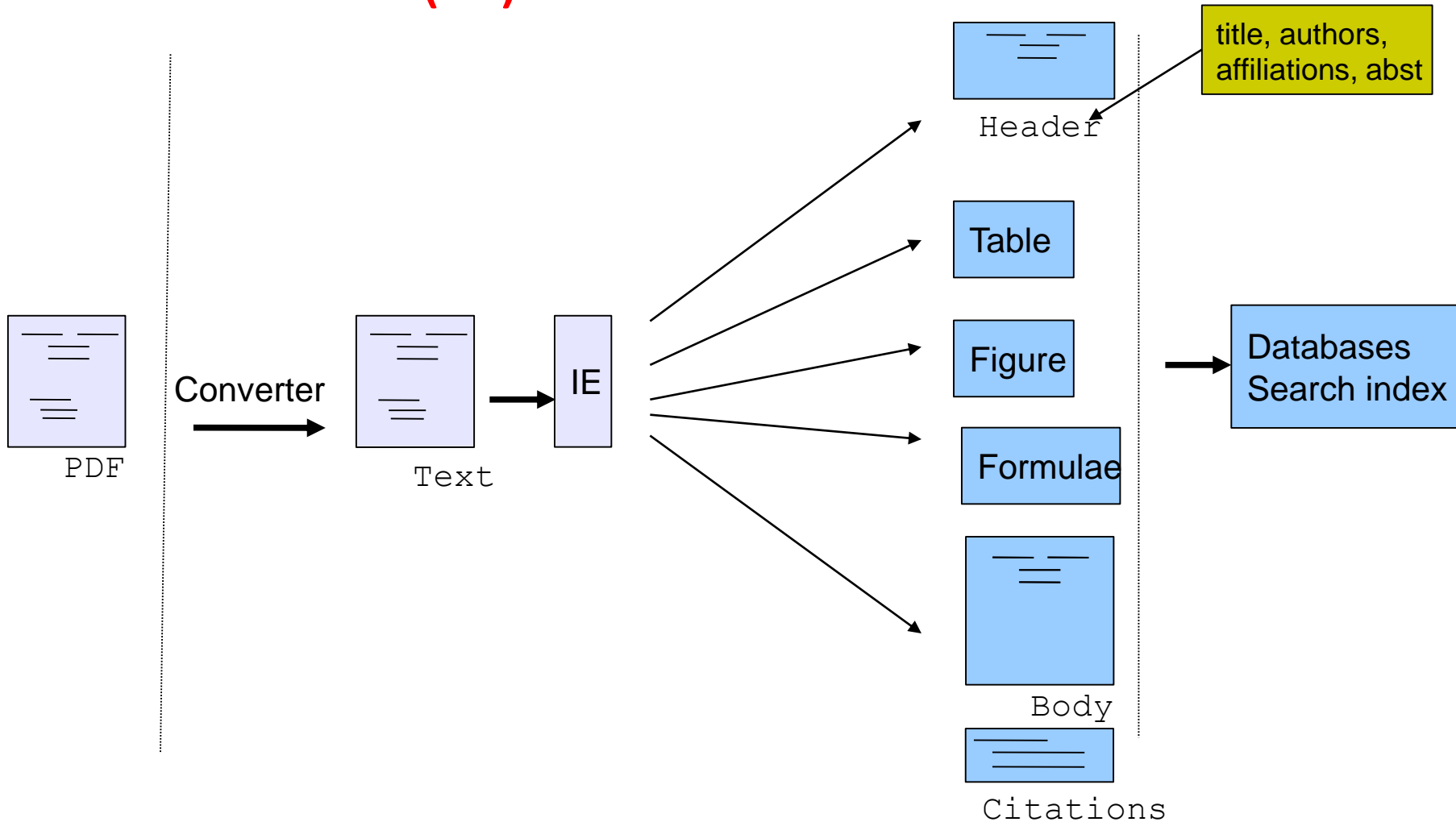
- Maintain a list of parent URLs where documents were previously found
 - Parent URLs are usually academic homepages.
 - >1,000,000 unique parent URLs, as of summer 2013
 - Parent URLs are stored in a database
 - Crawled weekly.
- Crawling process starts with the scheduler selecting all parent URLs
- Crawling batch with Heritrix
 - Most discovered documents have been crawled before.
 - Use hash table comparison for detection of new documents
 - Normally retrieve a 10K NEW documents per day, sometimes less than 1K.
- Very ethical crawler
 - Use whitelist and blacklist policy.

Highlights of AI/ML Technologies in CiteSeerX

- Document Classification
- Document Deduplication and Citation Graph
- Metadata Extraction
 - Header Extraction
 - Citation Extraction
 - Table Extraction
 - Figure Extraction
- Author Disambiguation

Wu, et.al IAAI 2014

Automatic Metadata Information Extraction (IE) - CiteSeerX



Other open source academic document metadata extractors available –
workshop, metadata hackathon,

TableSeer

Liu, et al, AAAI07, JCSDL06,

Table extraction & search engine

TableSeer

Table Caption ▼ flow

[Advanced](#)

search

Found 25 results for query "TableCaption : flow "

Instituto de Química, Universidade Federal da Bahia, Salvador-BA 40170-290, Brazil d Departamento de Química Analítica, Universidad de Valencia, Dr. Moliner 50, 46100 Burjassot, Valencia, Spain. E-mail: miguel.delaguardia@uv.es '-' Analyst '-' 2000

In PAGE 1, LINE 78:.....Table 1 Flow analysis determination of sulfide using the MB method.....;

[PDF](#)

[Preview](#)

Table 1 Comparative results for the determination of morphine in process liquors with chemiluminescence detection using pulsed flow chemistry (PFC) and conventional flow injection analysis (FIA) methodology

Pulsed flow chemistry: a new approach to solution handling for flow analysis coupled with chemiluminescence detection

Simon W. Lewis,* a Paul S. Francis, a Kieran F. Lim, a Graeme E. Jenkins b and Xue D. Wang c a Centre for Chiral and Molecular Technologies, School of Biological and Chemical Sciences, Deakin University, Geelong, Victoria 3217, Australia b Precision Devices P/L, 44 Nelson Street, Shoreham, Victoria 3916, Australia c School of Chemical and Biomedical Sciences, Central Queensland University, Rockhampton, Queensland 4702, Australia '-' Analyst '-' 2000

[PDF](#)

[Preview](#)

(stopped-flow) analysis mode, measuring peak area. Although the calibration appeared linear ($r^2 = 0.9996$), a log-log plot of signal area versus concentration revealed non-linear behaviour below 2.5×10^{-7} M and above. The purpose built pulsed flow chemiluminescence instrument provided high precision (less than 1% RSD), and a detection limit of 2.3×10^{-7} M. This was a significant improvement over the detection limit achieved with the prototype instrument, and was comparable to those reported in studies using conventional flow analysis under similar chemical conditions,^{9,10} although the lowest reported limit of detection for the determination of morphine with acidic potassium permanganate was 1.3×10^{-7} M.¹⁷

Analysis of process samples

The feasibility of pulsed flow analysis as an alternative to existing flow based techniques used in industrial process analysis was demonstrated with the analysis of pharmaceutical process samples using pulsed flow and conventional FIA instrumentation under the same chemical conditions. Four process samples were taken randomly from an aqueous fraction of an opiate extraction process. The determination of morphine in process samples using conventional FIA methodology has been previously demonstrated and validated against standard reversed-phase HPLC methodology.⁹ Interference resulting from the reduction of the permanganate ion by other alkaloids present in the extract is negligible due to a further rate contribution of the inherent selectivity of the light-producing reaction pathway and the concentration levels of the alkaloids present in the samples.¹⁸ Matrix effects arising from dissolved acids and pH were minimized by manually filtering and a 1000 fold dilution of the samples with the same polyphosphate substrate that was used to prepare the permanganate reagent and six morphine standards over the concentration range from 3.0×10^{-11} to 2.5×10^{-7} M. The pulsed flow instrument was operated in stopped-flow mode and the emission intensity was recorded for 60 s following the production of the mixed pulse. The ability to measure a far greater proportion of the chemiluminescence emission using the pulsed flow instrument in the stopped-flow mode revealed subtle differences in reaction kinetics between the standards and process samples, which were made reliable with conventional flow analysis methodology. It is postulated that species such as the other alkaloids present in the process samples, that do not result in an intense emission on reaction with permanganate, affect the rate of the light producing reaction. This effect, in relation to flow, stopped-flow and batch analysis of samples from the extraction process is currently

Table 1 Comparative results for the determination of morphine using chemiluminescence detection using pulsed flow (PFC) and conventional flow injection analysis (FIA) method

Process sample	Concentration/M		Flow rate/ml
	FIA	PFC	
1	0.0000	0.0000	1.0
2	0.0020	0.0020	1.4
3	0.0040	0.0040	0.8
4	0.0010	0.0010	1.0

* Morphine for chemiluminescence analysis ^b Detection of substrate

consumption. This rapid and efficient mixing and pulsed flow chemistry facilitates delivery of a chemical reaction mixture into the detector using a minimizing dispersion and enabling measurement period of minimum emission. The instrument is versatile, with the rate of pulsing, injection ratios, modes determined by software settings. The small of the robust propulsor device provide the potential for instrumentation able to perform rapid, versatile chemiluminescence assays.

Acknowledgements

The authors express their gratitude to Associate E. W. Jenett, Claire L. Leubsdorf (Deakin University) Professor Robert W. Carroll (Deakin University) help and useful advice during this project, the 6th Biomedical and Chemical Sciences Workshop. The society for assistance in instrumental configuration at Deakin University, for fabrication of the housing. Funding for this project was provided by Australian Research Council and an Australian Post Award (for PRP).

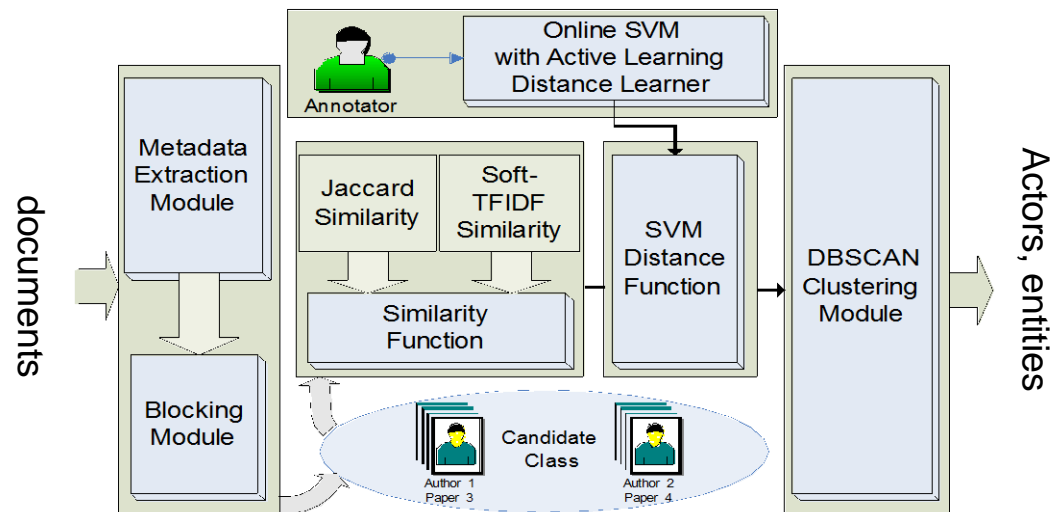
References

1. P. K. Liu, J. Chromatogr. 2000, 883, 113-117

Efficient Large Scale Author Disambiguation

CiteSeer^x & PubMed

- Must scale!!
- Motivation
 - **Correct attribution**
- Manually curated databases still have errors – DBLP, medical records
- Entity disambiguation problem
 - Determine the real identity of the authors using metadata of the research papers, including co-authors, affiliation, physical address, email address, information from crawling such as host server, etc.
 - Entity normalization



- Challenges
 - Accuracy
 - Scalability
 - Expandability

Han, et.al JCDL 2004
Huang, et.al PKDD 2006
Treeratpituk, et.al JCDL 2009
Khabsa, et.al JCDL 2015

- Key features
 - Learn distance function
 - Random Forest
 - others
 - DBSCAN clustering
 - Ameliorate labeling inconsistency (transitivity problem)
 - Efficient solution to find name clusters
 - N logN scaling

Recently all of PubMed authors, 80M mentions

csseer.ist.psu.edu

CSSeer^{beta}

Expert Recommender for Computer Science

Search for

Examples: [information retrieval](#), [data structure](#), [database](#)

[Contact Us to Sponsor CSSeer](#)

[About CSSeer](#) | [Privacy Policy](#)

© 2013 [The Pennsylvania State University](#)

Developed at and hosted by [The College of Information Sciences and Technology](#) at Penn State



- [Expert search](#) for authors

H-H Chen, JCDL 2014

Search for experts in:

Examples: *operating system* , *database* , *nonparametric statistics*

>> Related keyphrases

[dirichlet-multinomial distribution](#) [vector space](#) [natural language processing](#) [information seeking](#)

[search engine](#) [chinese restaurant process](#) [document classification](#)

[natural language](#) [digital library](#) [user profile](#) [decryption oracle](#)

[information overload](#) [audio mining](#) [information science](#) [world wide web](#)

>> List of experts

1. **W. Bruce Croft**
Dept. of Computer Science, University of Massachusetts
2. **Jamie Callan**
Language Technologies Institute, School of Computer Science, Carnegie Mellon University
3. **Alan F. Smeaton**
Centre for Digital Video Processing
4. **Eyal Kushilevitz**
Computer Science Dept., Technion
5. **Yuval Ishai**
Computer Science Dept., Technion

Experimental Collaborator recommendation system

CollabSeer^{beta} Collaborator Recommendation Engine

Search for author
by method

[Contact Us](#) to Sponsor CollabSeer

[About CollabSeer](#) | [Feedback](#) | [Privacy Policy](#)

© 2010 [The Pennsylvania State University](#)

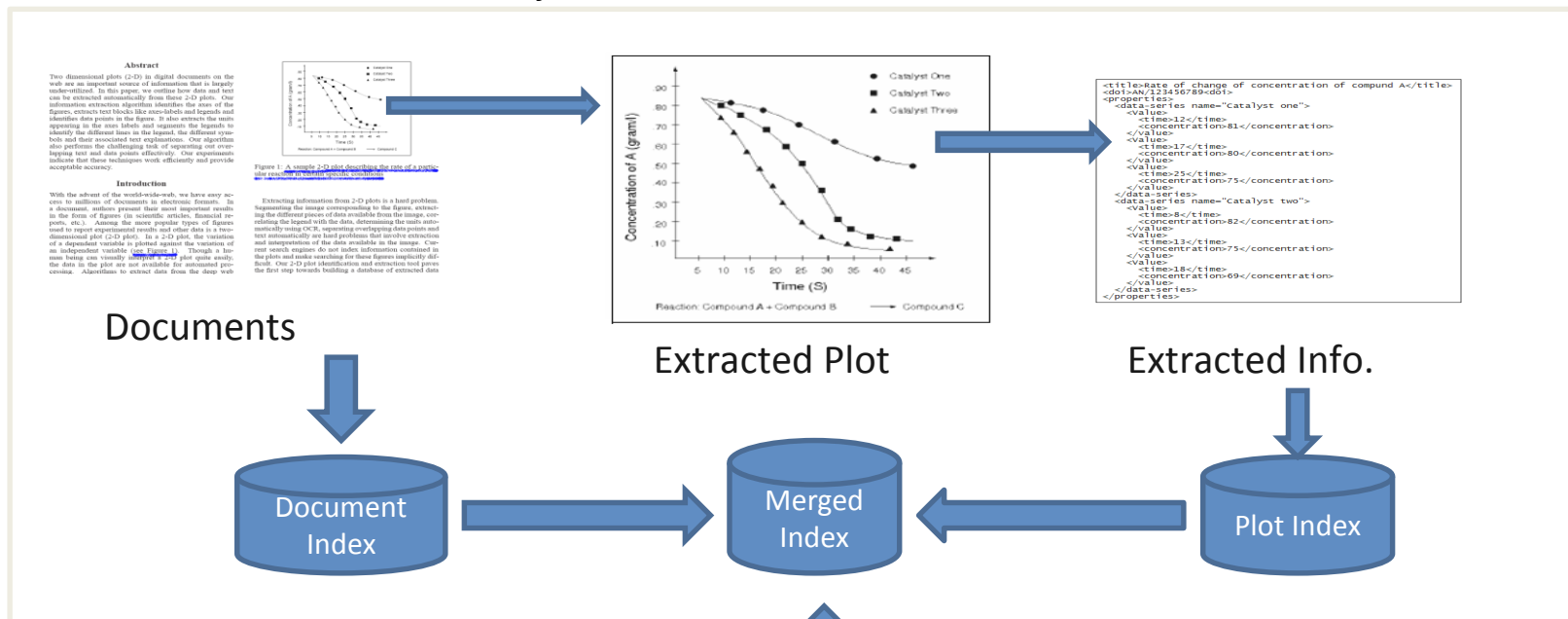
Developed at and hosted by [The College of Information Sciences and Technology](#) at Penn State



- CollabSeer currently supports 400k authors
- <http://collabseer.ist.psu.edu>

Automated Figure Data Extraction and Search

- Large amount of results in digital documents are recorded in figures, time series, experimental results (eg., NMR spectra, income growth)
- Extraction for purposes of:
 - Further modeling using presented data
 - Indexing, meta-data creation for storage & search on figures for data reuse
- *Current extraction done manually!*



Abstract
Two dimensional plots (2-D) in digital documents on the web are an important source of information that is largely underutilized. In this paper, we outline how data and text can be extracted automatically from these 2-D plots. Our information extraction algorithm identifies the axes of the figures, extracts text blocks like axis-labels and legends and identifies data points in the figure. It also extracts the units appearing in the axis labels and segments the legends to identify the different lines in the legend, the different symbols and their associated line equations. Our algorithm also performs the challenging task of separating out overlapping text and data points effectively. Our experiments indicate that these techniques work efficiently and provide acceptable accuracy.

Introduction
With the advent of the world-wide-web, we have easy access to millions of documents in electronic formats. In a document, authors present their most important results in the form of figures (in scientific articles, financial reports, etc.). Among the more popular types of figures used to report experimental results and other data is a two-dimensional plot (2-D plot). In a 2-D plot, the variation of a dependent variable is plotted against the variation of an independent variable (see Figure 1). Through a technique being currently developed, plot data can be used. The data in the plot are not available for automated processing. Algorithms to extract data from the drawn web

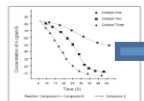
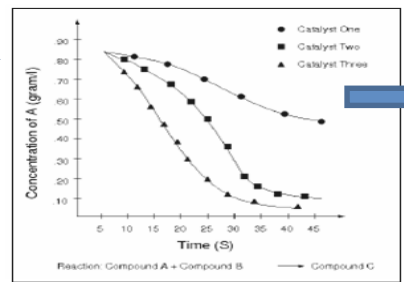


Figure 1: A sample 2-D plot describing the rate of a specific reaction in a catalytic reaction.

Extracting information from 2-D plots is a hard problem. Separating the image corresponding to the figure, extracting the different pieces of data available from the image, correlating the legend with the data, determining the units automatically using OCR, separating overlapping data points and text automatically are hard problems that involve extraction and interpretation of the data available in the image. Current search engines do not index information contained in the plots and make searching for these figures practically difficult. Our 2-D plot identification and extraction tool paves the first step towards building a database of extracted data



```
<title>Rate of change of concentration of compound A</title>
<id>AN_123456789</id>
<properties>
<data-series name="Catalyst one">
<value>
<time>12/<time>
<concentration>81</concentration>
</value>
<value>
<time>17/<time>
<concentration>80</concentration>
</value>
<value>
<time>25/<time>
<concentration>75</concentration>
</value>
</data-series>
<data-series name="Catalyst two">
<value>
<time>8/<time>
<concentration>82</concentration>
</value>
<value>
<time>13/<time>
<concentration>75</concentration>
</value>
<value>
<time>18/<time>
<concentration>69</concentration>
</value>
</data-series>
</properties>
```

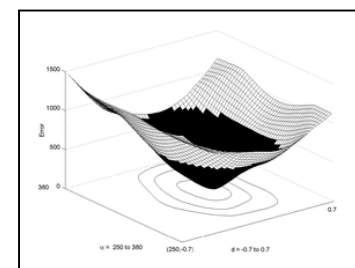
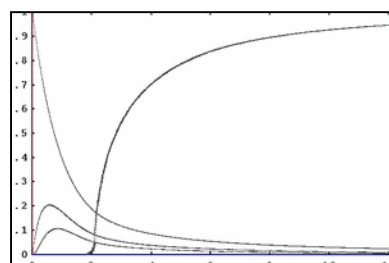
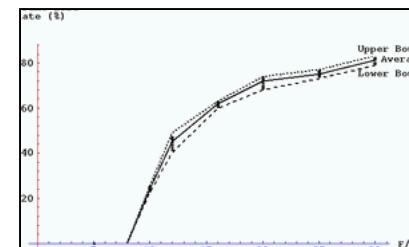
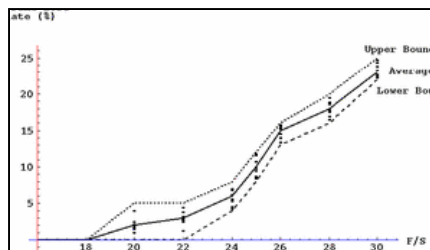
Digital Library



User

Chem_xSeer Figure/Plot Data Extraction and Search

Numerical data in scientific publications are often found in figures.



Tools that automate the data extraction from figures provide the following:

- Increases our understanding of key concepts of papers
- Provides data for automatic comparative analyses.
- Enables regeneration of figures in different contexts.
- Enables search for documents with figures containing specific experiment results.

An Approach to Plot Data Extraction

- Identify and extract figures from digital documents
 - Ascii and image extraction (xpdf)
 - OCR - bit map, raster pdfs
- Identify figures as images of 2D plots using SVM (Only for Bit map images)
 - Hough transform
 - Wavelets coefficients of image
 - Surrounding text features
- Binarization of the 2D plots identified for preprocessing (No need for Vectorized Images)
 - Adaptive Thresholding
- Image segmentation to identify regions
 - Profiling or Image Signature
- Text block detection
 - Nearest Neighbor
- Data point detection
 - K-means Filtering
- Data point disambiguation for overlapping points
 - Simulated Annealing

Automatic Citation (or paper) Recommendation

Basic

Topic

Advanced

RefSeer

Citation Recommendation System

Built on top of millions of papers

Never miss a citation and know about
the latest work

Several recommendations models

Recommend

[About RefSeerX](#)

[Feedback](#)

[RefSeerX FAQ](#)

Developed at and hosted by The College of Information Sciences and Technology

© 2012 The Pennsylvania State University

Huang, AAI 2015
Huang, CIKM 2013
He, WWW 2010

Chem_xSeer

Search Papers Authors Tables Figures Formula Extract Tables CollabSeer



Search

Eg : Methanol, CO₂, Adam Smith

Sun TOIS 2011

Scholarly Document Size & Numbers

- Large # of academic/research documents, all containing lots of data
 - Many millions of documents
 - 50M records – Microsoft Academic (2013)
 - 25M records, 10 million authors, 3 times mentions – PubMed
 - Google scholar (english) estimated to be **~100M records**
 - ***Total online estimate ~120M records*** [Khabsa, Giles, PLoS ONE, '14](#)
 - ***~25 million full documents freely available***
 - 100s of millions of authors, affiliations, locations, dates
 - Billions of citation mentions
 - 100s millions of tables, figures, math, formulae, etc.
 - Related & linked data
 - Raw data > petabytes

Challenges

- Tables, figures, formula, equations, methodologies, etc.
 - How do we effectively integrate and utilize this data for search?
 - Natural language generation?
- Ontologies for scholarly data
 - Scholarly “knowledge vault”
- “Big Mechanism” approaches

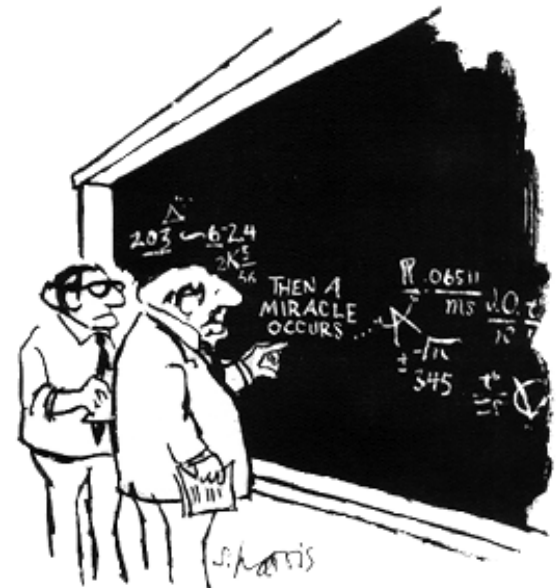
Summary/observations

- Scholarly big data – petabytes, billions of objects
 - Rich content – large related data
- Applications more common than most realize
 - Science, research, education, patents, policy, sociology, economics, business, MOOCs, etc
 - Growth of associated data: tables, figures, chemical & drug entities, equations, methodologies, slides, video, etc
- Many issues – AI and ML very useful:
 - Focused NLP
 - Information extraction still an art; domain dependent
 - Data is not always easy to move around or share
 - Some data still not readily available but is changing – $\frac{1}{4}$ of all digital documents freely available
 - Data(s) integration issues
 - Meta analysis – “big mechanism” opportunities
- Observations
 - Large amount and growing scholarly related data
 - Big scholarly data creates new research opportunities
 - *Big scholarly data creates other big data*

**“The future ain’t what it used to be.”
Yogi Berra, catcher, NY Yankees.**

For more information

- clgiles.ist.psu.edu
- giles@ist.psu.edu
- SourceForge.com



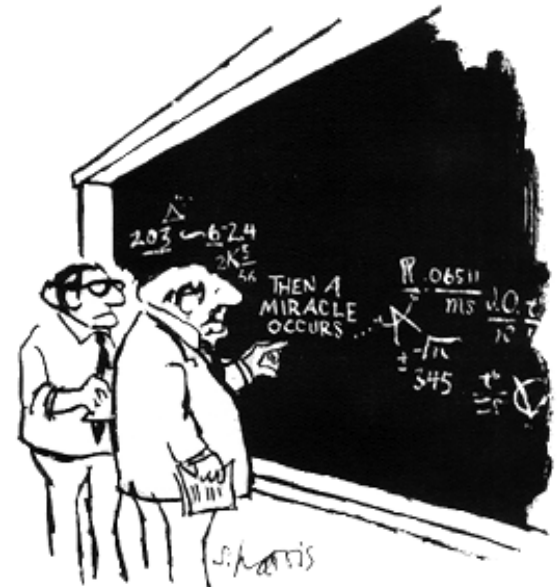
"I think you should be more explicit here in step two."

“Online or invisible,” *Nature*, ‘01, Steve Lawrence, Google Desktop creator

“5 times more likely to be cited if your paper is freely available online”

For more information

- <http://clgiles.ist.psu.edu>
 - Most of our papers
- giles@ist.psu.edu
- SourceForge.com (github)



"I think you should be more explicit here in step two."