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$_X$Seer, 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.

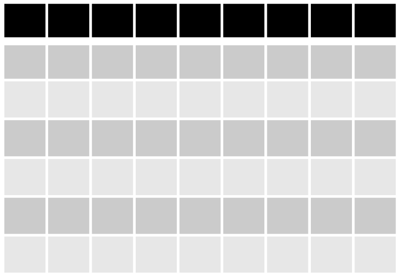
What is Scholarly Big Data

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

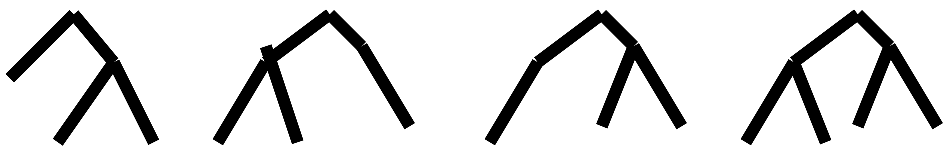
- Related data:
  - Academic/researcher/group/lab web homepages
  - Funding agency and organization grants, records, reports
  - Research laboratories reports
  - Patents
  - *Associated data*
    - presentations
    - experimental data (very large)
    - video
    - course materials
    - other
  -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

Scholarly Big Data

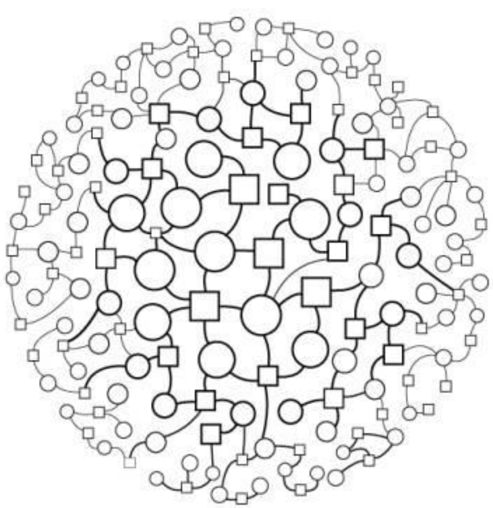
Most of the data that is available in the era of scholarly big data does not look like this



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

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

- 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

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

Lower bound estimate of size of the Web:

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

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

Freely available by scholarly field

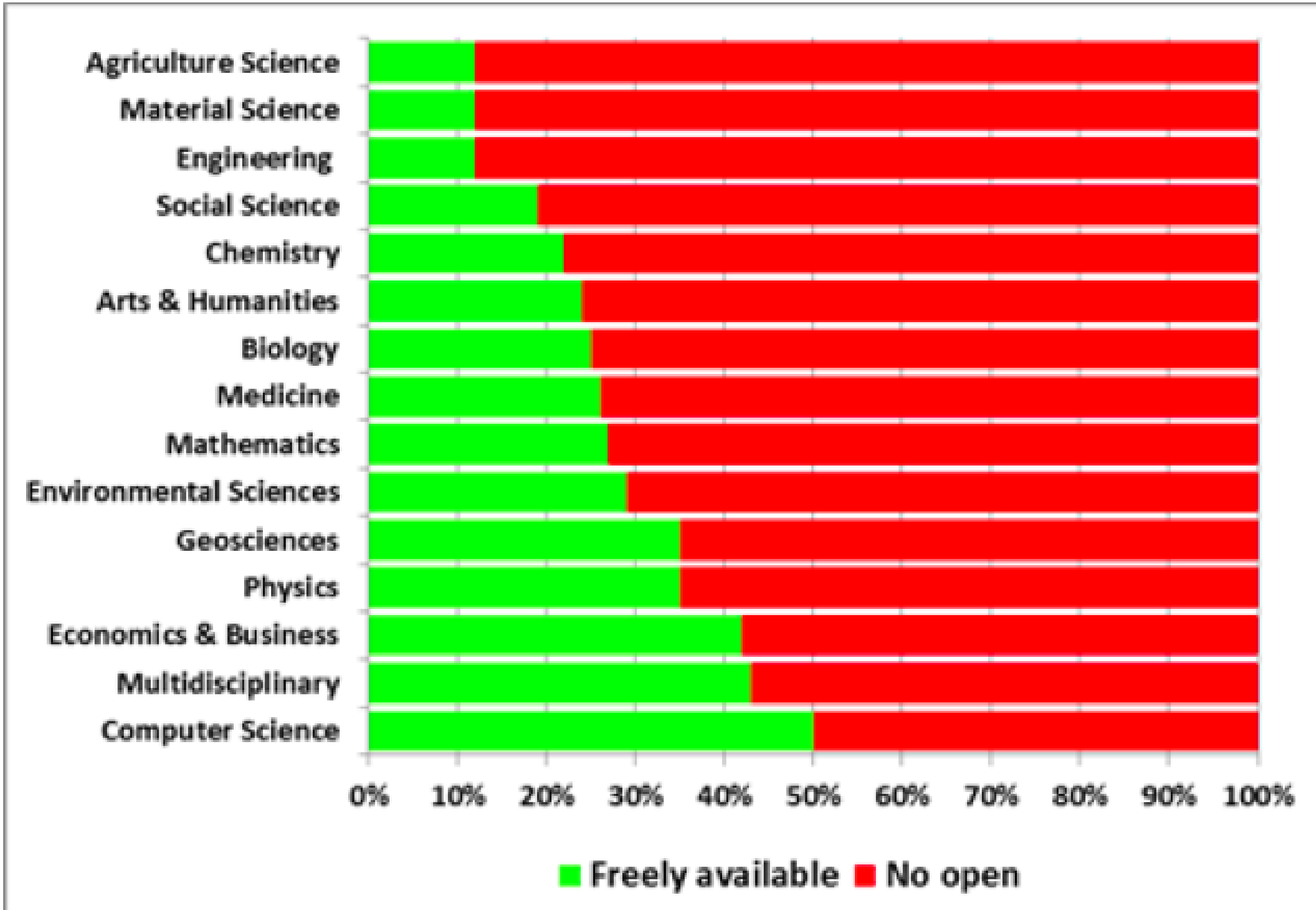


Figure 4. Percentage of publicly available documents according to scientific fields

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

Khabsa, Giles, PLoSONE, 2014

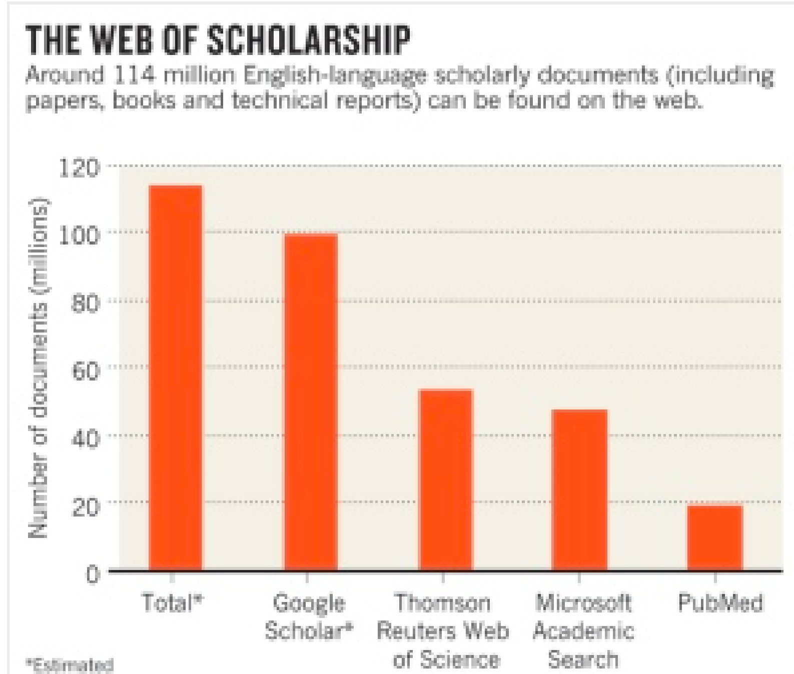
NATURE | SEVEN DAYS

عربي

Seven days: 16–22 May 2014

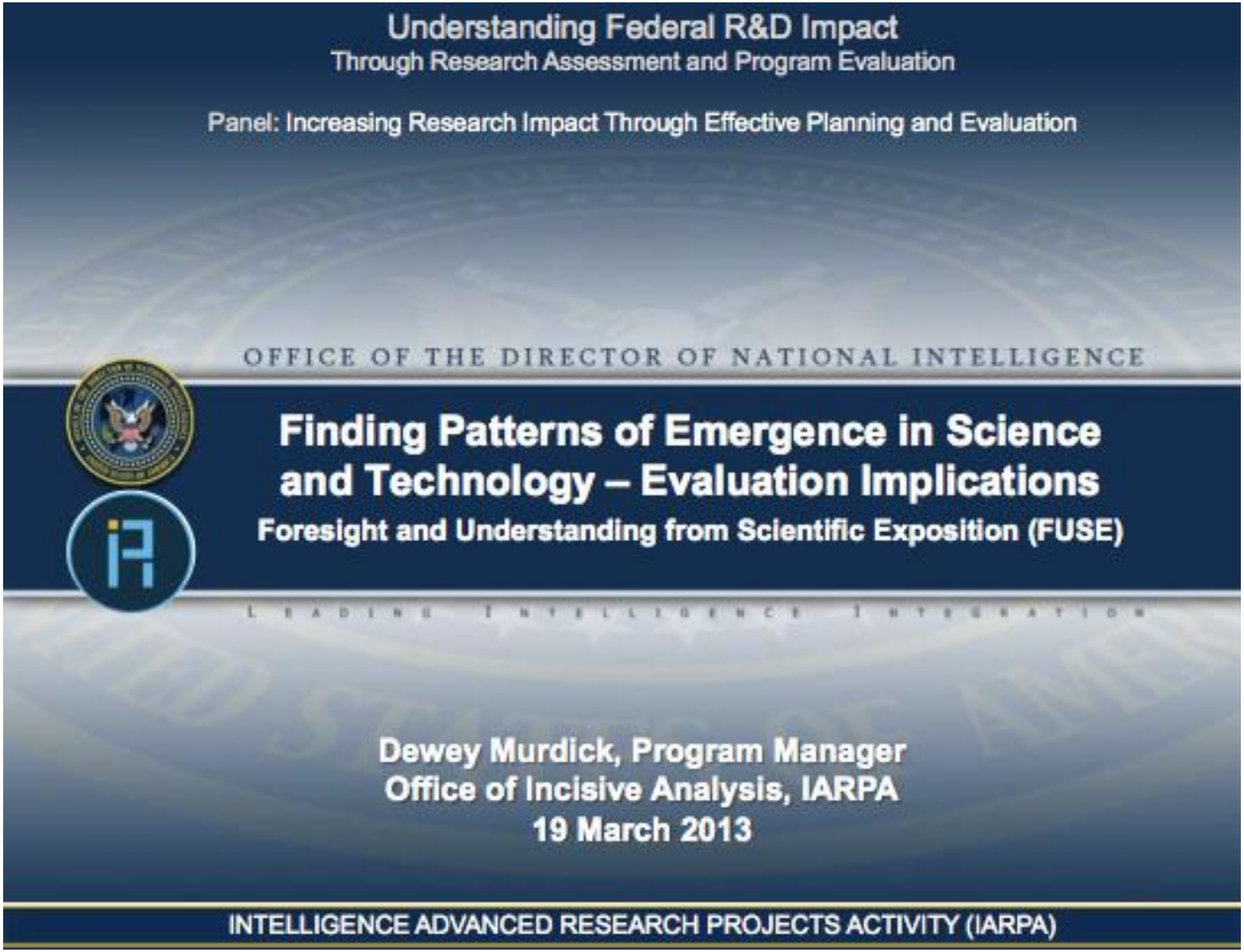
TREND WATCH

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 Khabsa at Pennsylvania State University in University Park (M. Khabsa 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.



Source: PLoS ONE/thomson reuters

IARPA FUSE Program



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

SeerSuite Toolkit & Applications

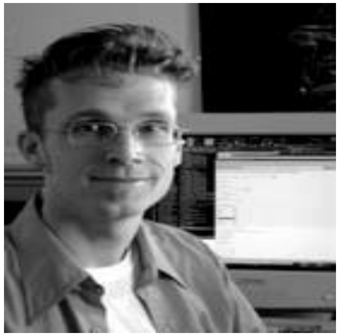
- SeerSuite: open source search engine and digital library tool kit used to build academic search engines/digital libraries
  - **CiteSeer^X , Chem$_X$Seer, AckSeer,CSSeer, CollabSeer**, **RefSeer**, etc.
  - Built on commercial grade open source tools (Solr/Lucene; mySQL)
  - Our features – automated 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

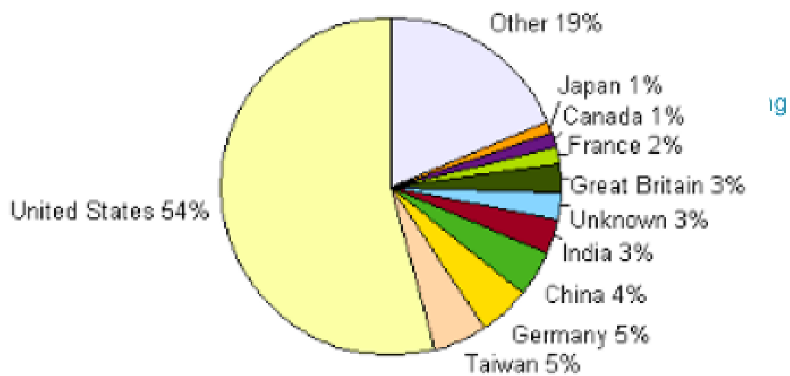
CiteSeer^X http://citeseerx.ist.psu.edu

• 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

CiteSeerX βETA



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.

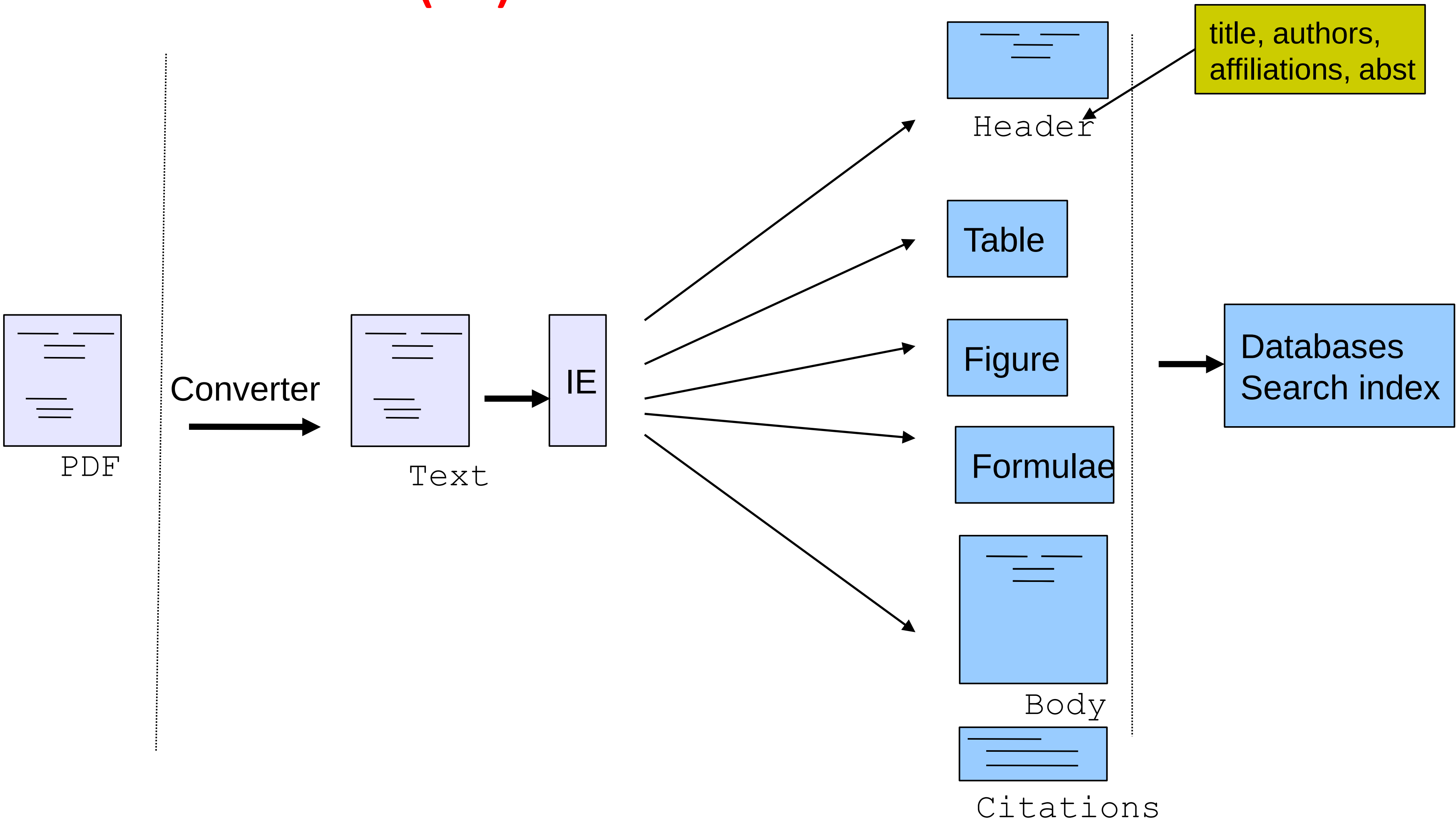
Zheng, CIKM'09; Wu Webscience'12

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

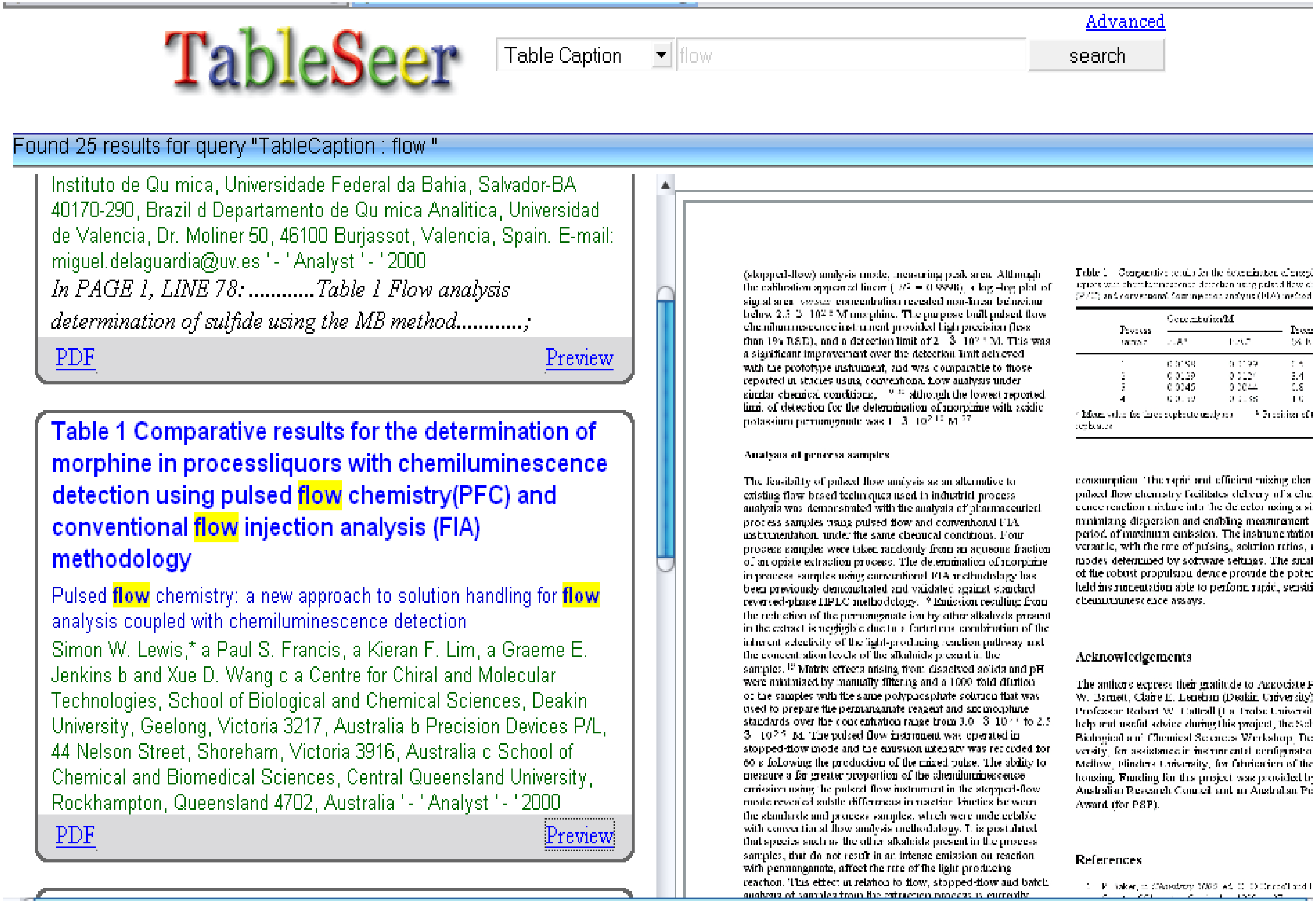


Other open source academic document metadata extractors available – workshop, metadata hackathon,

TableSeer

Liu, et al, AAAI07, JCDL06,

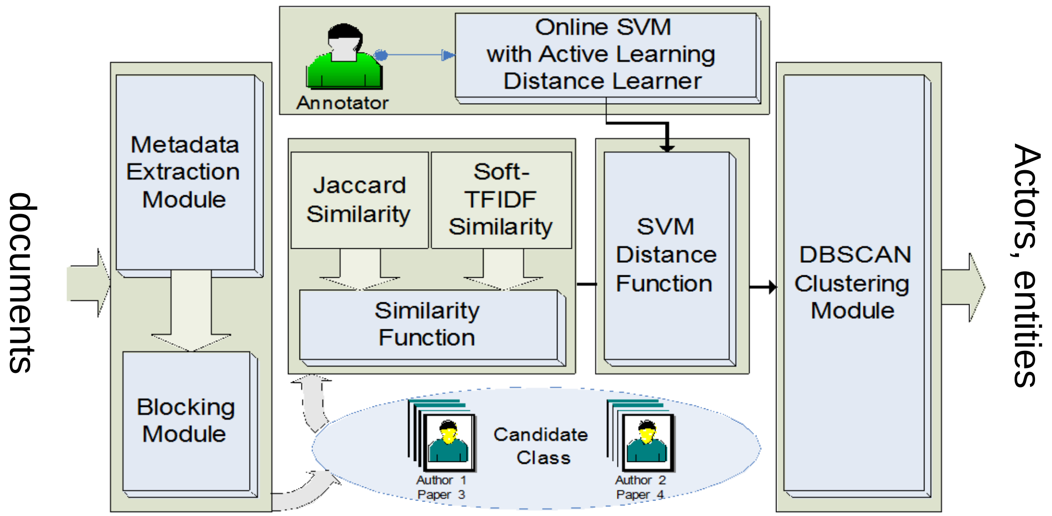
Table extraction & search engine



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

H-H Chen, JCDL 2014

- [Expert search](#) for authors

CSSeers

Search for experts in: information retrieval Submit

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

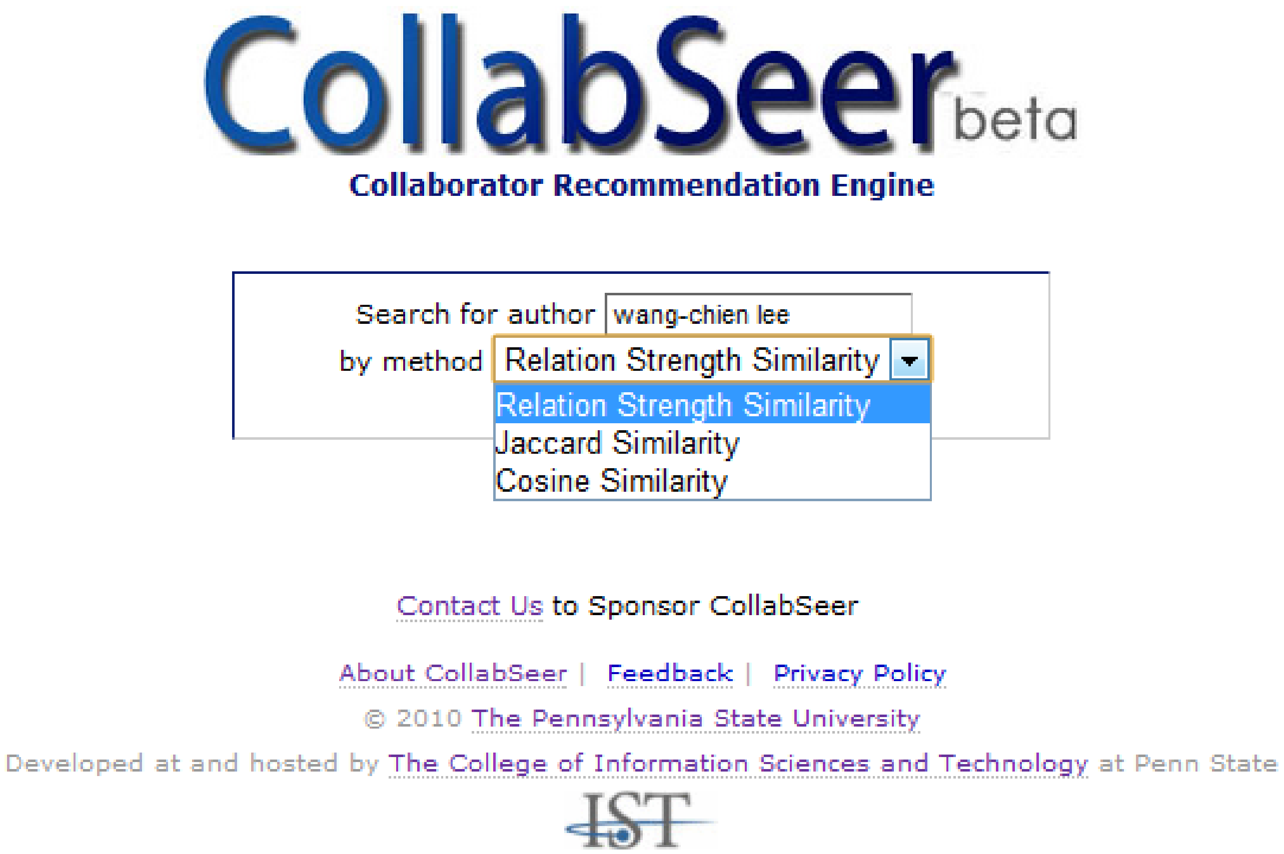
>> Related keyphrases

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

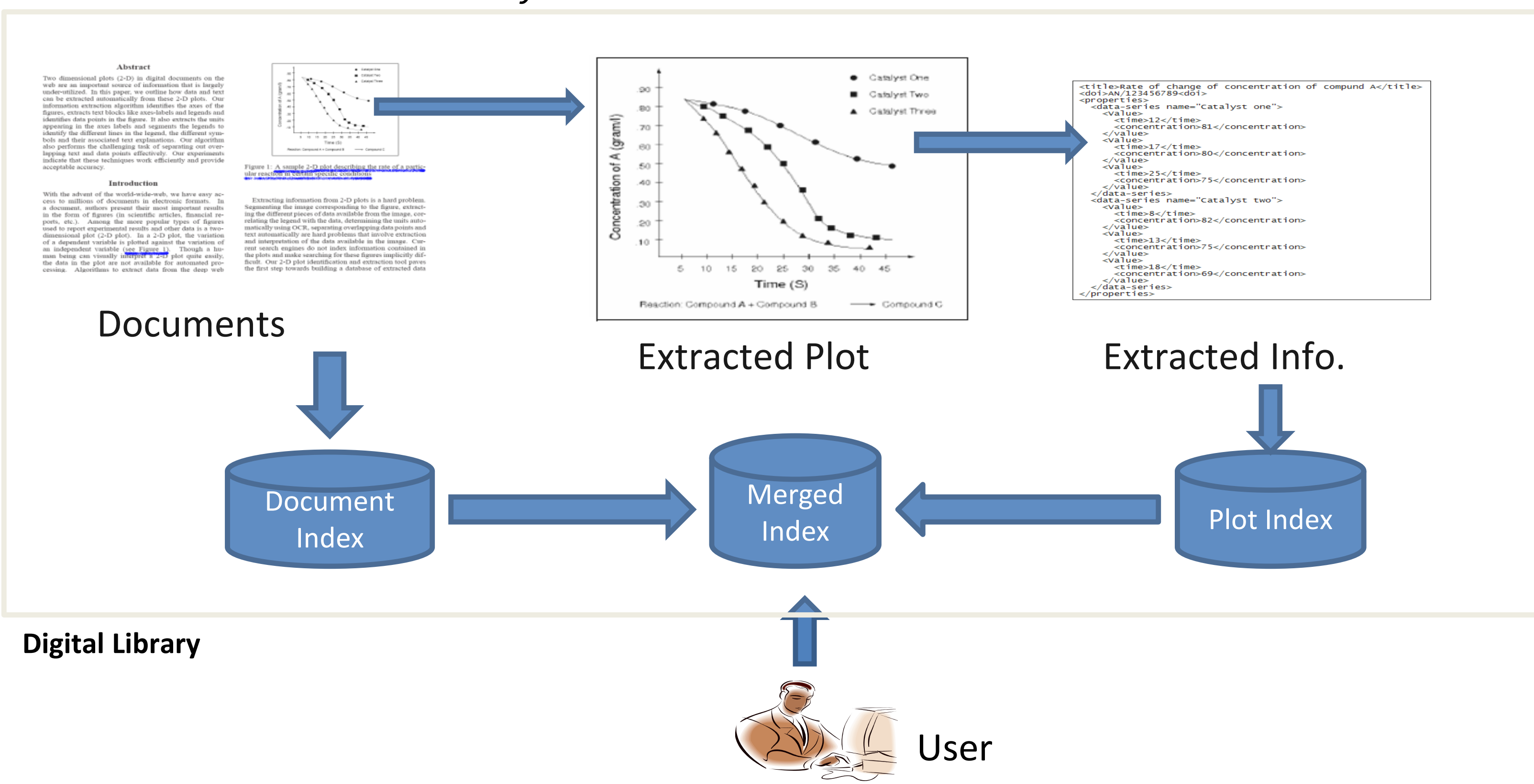
HH Chen, JCDL 2011

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)

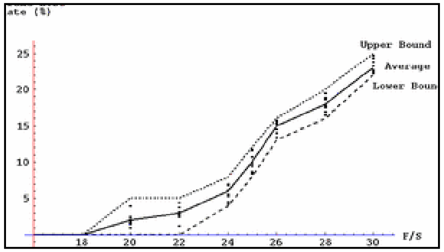
- Extraction for purposes of:
  - Further modeling using presented data
  - Indexing, meta-data creation for storage & search on figures for data reuse

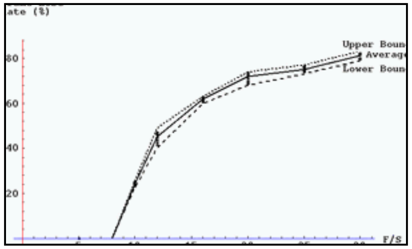
- *Current extraction done manually!*

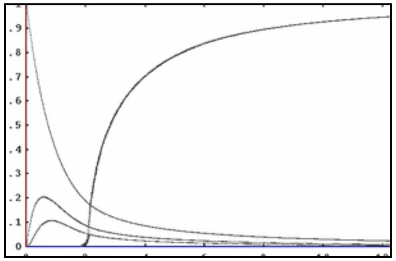


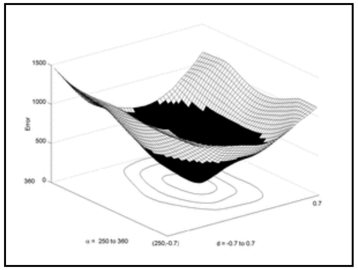
Chem$_X$Seer Figure/Plot Data Extraction and Search

Numerical data in scientific publications are often found in figures.









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

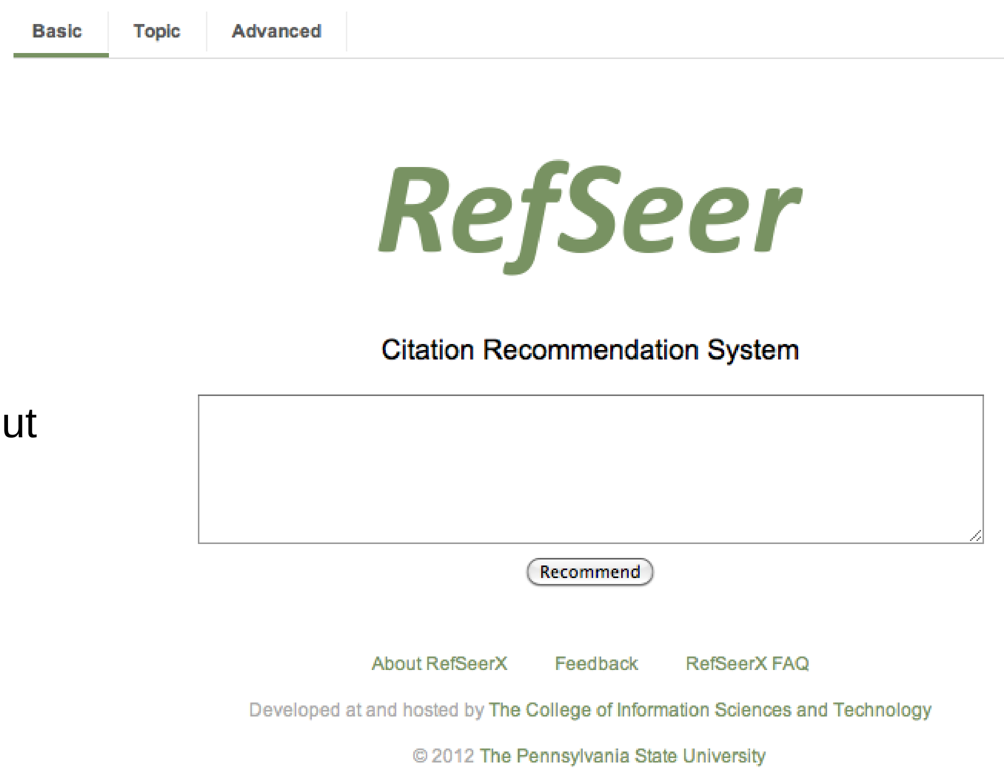
- Increases our understanding of key concepts of papers
- Provides data for automatic comparative analyses.
- Enables regeneration of figures in different contexts.
- Enables search for documents with figures containing specific experiment results.

X. Lu, et.al, JCDL 2006, Kataria, et.al, 2008
Choudhury, DocEng 2005

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



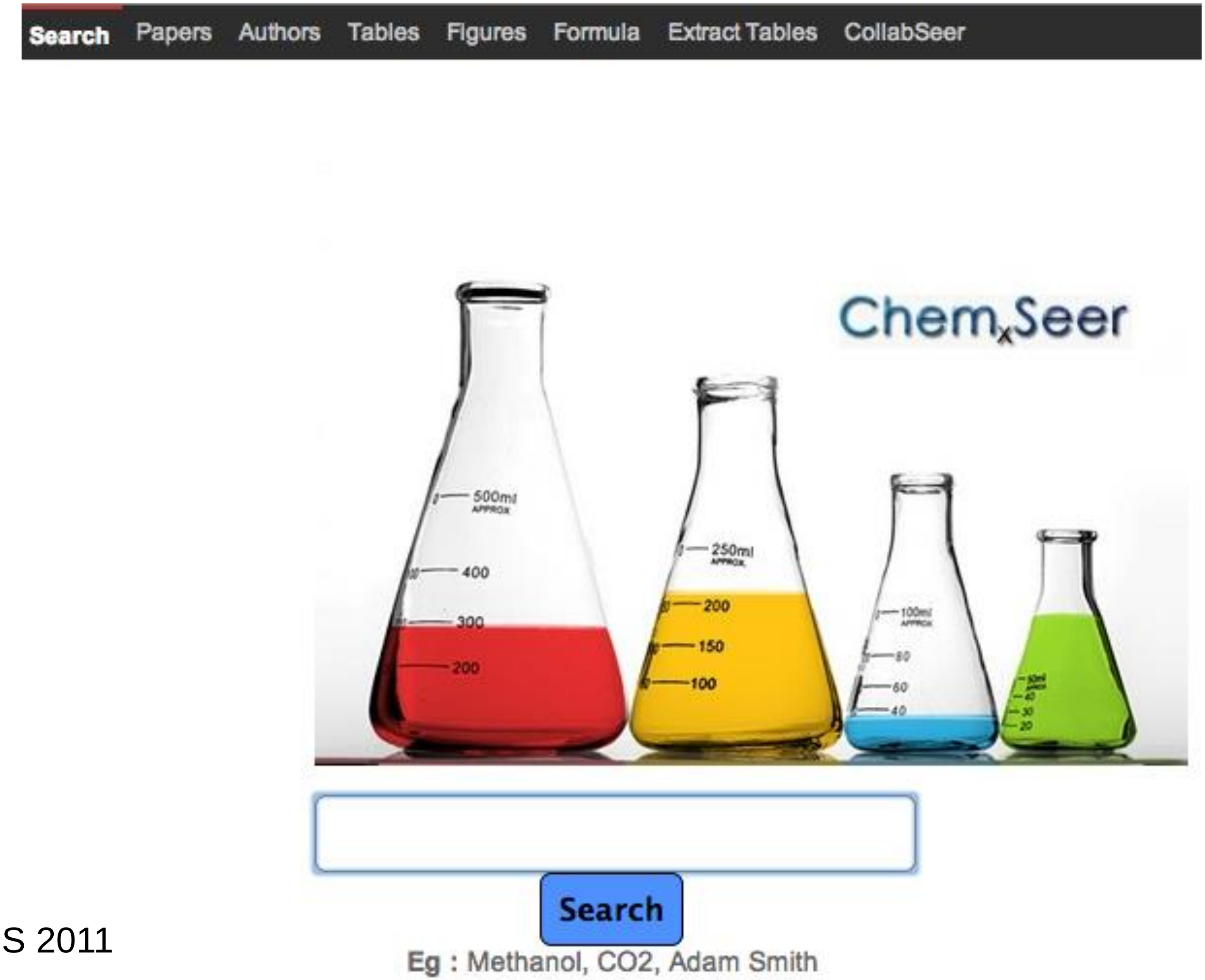
Built on top of millions of papers

Never miss a citation and know about the latest work

Several recommendations models

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

Chem$_{X}$Seer



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 <u>records</u>*** Khabsa, Giles, PLoSONE, '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

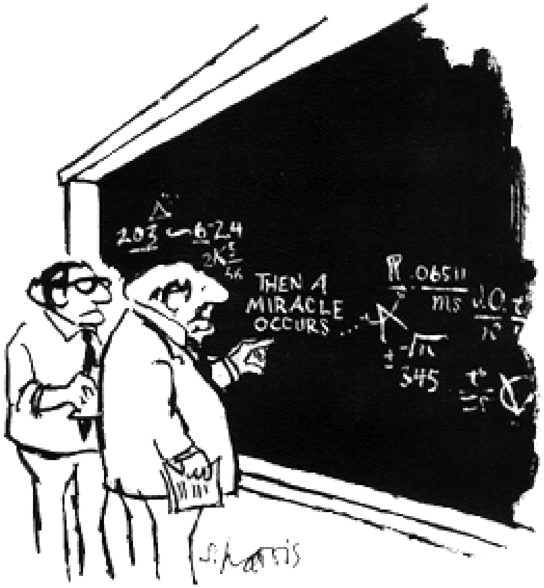
Summary/observations

- 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 – ¼ of all digital documents freely available
  - Data(s) integration issues
  - Meta analysis – "big mechanism" opportunties
- 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



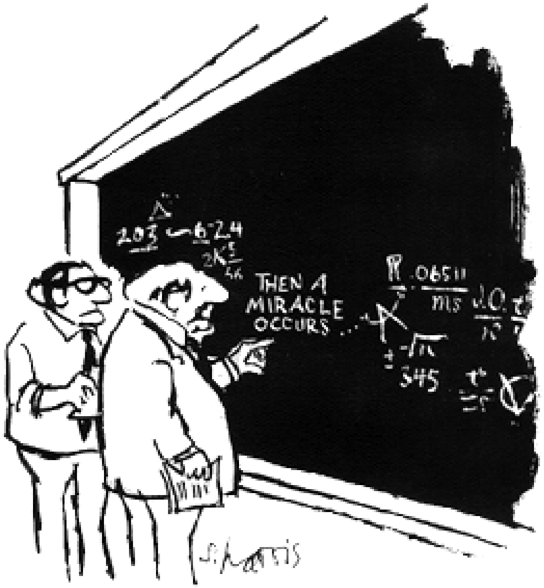
"I think you should be more explicit here in step two."

"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."