



# **Traversing the Academic Graph**

**Matt Luongo**



**@mhlungo**

**[github.com/mhlungo](https://github.com/mhlungo)**

**Matt Luongo**



**SCHOLR.LY**

# Search for research.

SCHOLR.LY Beta



SCHOLR.LY

"graph database"



Search



CURRENTLY SERVING COMPUTER SCIENCE

## Featured Publications



**Open information extraction from the web**

**Michael J. Cafarella, Stephen Soderl, Oren Etzioni**

In Proceedings of the International Joint Conferences on Artificial Intelligence 2007



**Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories**

## Featured Authors



**Alex Pentland**

**MIT Media Lab  
Massachusetts Institute of  
Technology**

**225** TOTAL NUMBER OF  
PAPERS PUBLISHED



**Charles Isbell**

**College of Computing**

# Find relevant papers and authors.

SCHOLR.LY Beta



"graph database"



Search

## Publications

### Survey of graph database models

Renzo Angles, Renzo Angles, Claudio Gutierrez, Claudio Gutierrez  
2001

Graph database models can be characterized as those where data structures for the schema and instances are modeled as graphs or generalizations of them, and data manipulation is expressed by graph-oriented operations and type constructors. These models flourished in the eighties and early nineties in parallel to object oriented models and their influence gradually faded with ...

### GRACE: A Graph Database System

GRACE is a database management system for managing data that is in the form of labeled undirected graphs. A large amount of data generated today, like road maps, protein-structures, citation-data, etc. are in the form of undirected labeled graphs. Also, problems involving such data require new forms of queries that are run on the structural properties of the graph. Finding st...

### A novel spectral coding in a large graph database

Lei Chen, Jeffrey Xu Yu, Yansheng Lu  
2008

Retrieving related graphs containing a query graph from a large graph database is a key issue in many graph-based applications, such as drug discovery and structural pattern recognition. Because sub-graph isomorphism is a NP-complete problem [4], we have to employ a filter-and-verification framework to speed up the search efficiency, that is, using an effective and

## Authors

### Wei Wang

Simon Fraser University, Canada

RELATED PUBLICATION:

Computerized Biological Brain Phantom for Evaluation of PET and SPECT Reconstruction



### Jun Huan

1; Computer Science Department; 2; The Laboratory for Molecular Modeling, School of Pharmacy; University of North Carolina at Chapel Hill

RELATED PUBLICATION:

Special Section on Automated Function Prediction



### Lei Chen

2; Department of Computer Science, Hong Kong University of Science and Technology

RELATED PUBLICATION:

OPHMR: An Optimized Polymorphic Hybrid Multicast Routing Protocol for MANET



### Jeffrey Xu Yu

# Explore an author's work.

SCHOLR.LY Beta



Search

20

PAPERS

168

CITATIONS



Website

## Jun Huan

1; Computer Science Department; 2;  
The Laboratory for Molecular Modeling,  
School of Pharmacy; University of North  
Carolina at Chapel Hill

### TOP KEYWORDS

Structure Motif  
Graph Mining  
Clique  
Protein Structure Comparison

## Dashboard

### Publications

#### 2008 GRAPH WAVELET ALIGNMENT KERNELS FOR DRUG VIRTUAL SCREENING

Gerald Lushington, Jun Huan, Aaron Smalter, Gerald Lushington, Jun Huan,  
Aaron Smalter, Aaron Smalter, Gerald Lushington, Jun Huan

2008

In this paper we introduce a novel graph classification algorithm and demonstrate its efficacy in drug design. In our method, we use graphs to model chemical structures and apply a wavelet analysis of graphs to create features capturing graph local topology. We design a novel graph kernel function to utilize the created feature to build predictive models for chemicals. We call the new graph kernel

### Co-authors



#### Wei Wang

Simon Fraser University, Canada



#### Jan Prins

Department of Computer Science,  
Department of Computer Science,; Duke  
University, University of North Carolina,

# Motivation

# Motivation





# **Search & Graphs**



Octopart is a search engine for electronic parts

---

[Browse Categories](#)

[Semiconductors and Actives](#)  
[Test Equipment](#)  
[Enclosures](#)  
[Industrial Control](#)  
[Sound Input/Output](#)

[Cables and Wire](#)  
[Tools and Supplies](#)  
[Optoelectronics](#)  
[Connectors and Adapters](#)

[Current Filtering](#)  
[Passive Components](#)  
[Indicators and Displays](#)  
[Power Products](#)

Filter by: ☐ In stock ☐ Lead free ☐ RoHS compliant

[Share results »](#)  
[Start search over »](#)

## Categories

- Cables and Wire
- Connectors and Adapters
- Current Filtering
- Enclosures
- Indicators and Displays
- Industrial Control
- Optoelectronics
- Passive Components
- Power Products
- Semiconductors and Actives (3)
- Sound Input/Output
- Test Equipment
- Tools and Supplies

## Manufacturer

- ☐ Atmel (6)

## Supplier

- ☐ Avnet Express (3)
- ☐ Arrow (2)
- ☐ Avnet Express Asia (2)
- ☐ Avnet Express Europe (2)
- ☐ Digi-Key (2)
- [more...](#)

## FLASH Memory Size

### Results 1-6 of 6 for ATMEGA168-20AU:

sort by: **relevance** | [price](#)

#### [Atmel - ATMEGA168-20AU](#)

datasheet: [pdf](#)

ic,atmega168-20au,avr 8 bit r i sc, 16kb flash...

*Newark	<a href="#">68T2692</a>	\$4.33	avail:1872	<a href="#">Buy Now</a>
*Verical	<a href="#">ATMEGA168-20AU</a>	\$2.7543	avail:713	<a href="#">Buy Now</a>
*Arrow	<a href="#">ATMEGA168-20AU</a>	\$3.03	avail:713	<a href="#">Buy Now</a>
*Avnet Express	<a href="#">ATMEGA168-20AU</a>	\$3.09	avail:449	<a href="#">Buy Now</a>
*Mouser	<a href="#">556-ATMEGA168-20AU</a>	\$3.03	avail:2621	<a href="#">Buy Now</a>
<a href="#">more...</a>				
*authorized				



jameco.com

#### [Atmel - ATMEGA168-20AUR](#)

datasheet: [pdf](#)

Mcu Avr 16kb Flash 20mhz 32tqfp

*Arrow	<a href="#">ATMEGA168-20AUR</a>	\$3.12	avail:0	
*Avnet Express	<a href="#">ATMEGA168-20AUR</a>	\$2.51	avail:2000	<a href="#">Buy Now</a>
*Digi-Key	<a href="#">ATMEGA168-20AUR</a>	\$2.472	avail:3636	<a href="#">Buy Now</a>
*Nu Horizons	<a href="#">ATMEGA168-20AUR</a>		avail:0	
*Avnet Express Europe	<a href="#">ATMEGA168-20AUR</a>	\$3.176	avail:0	
<a href="#">more...</a>				
*authorized				

#### [Atmel - ATMEGA168-20AU-SL383](#)

datasheet: [pdf](#)

Semiconductors and Actives, Microprocessors, Microcontrollers

*Avnet Express	<a href="#">ATMEGA168-20AU SL383</a>	\$2.45	avail:0	
*authorized				

#### [Atmel - ATMEGA168-20AUPBF](#)

[Upload datasheet »](#)

XS Micro	<a href="#">ATMEGA168-20AUPBF</a>		avail:contact	
*none authorized				

#### [Atmel - ATMEGA168-20AUTQFP32XPB](#)

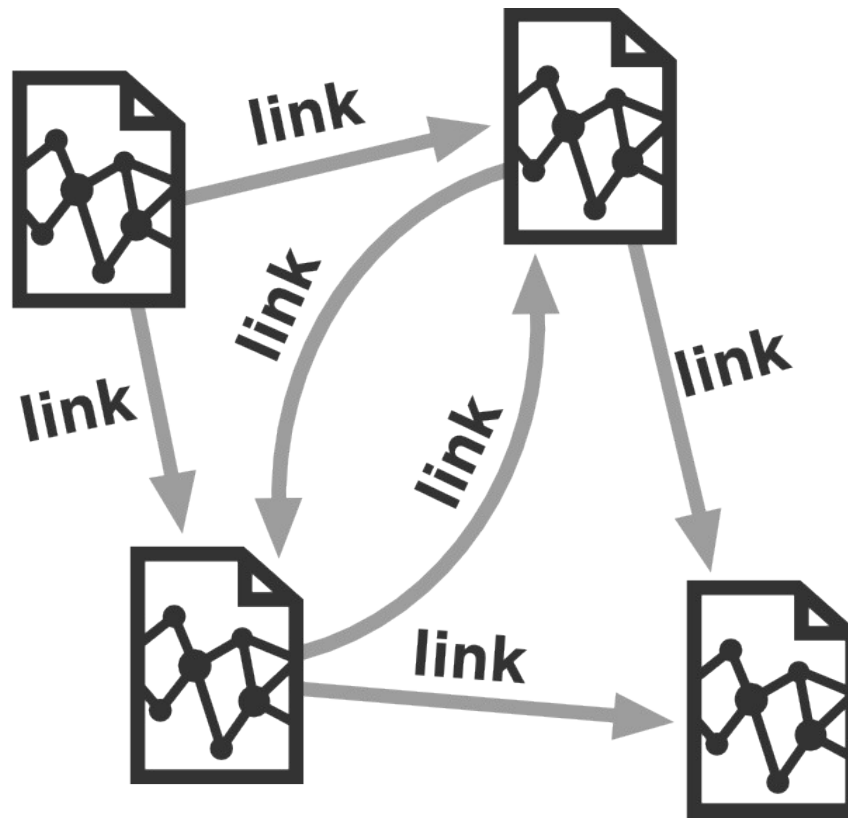
[Upload datasheet »](#)

XS Micro	<a href="#">ATMEGA168-20AUTQFP32XPB</a>		avail:contact	
----------	---	--	---------------	--

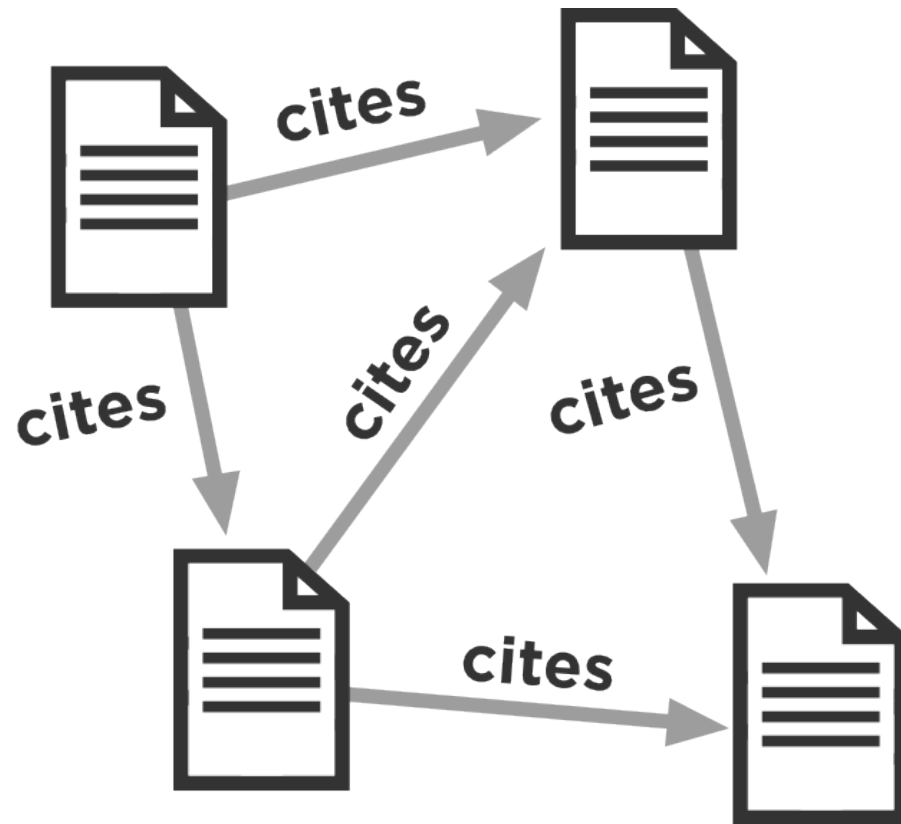
**On the other hand...**



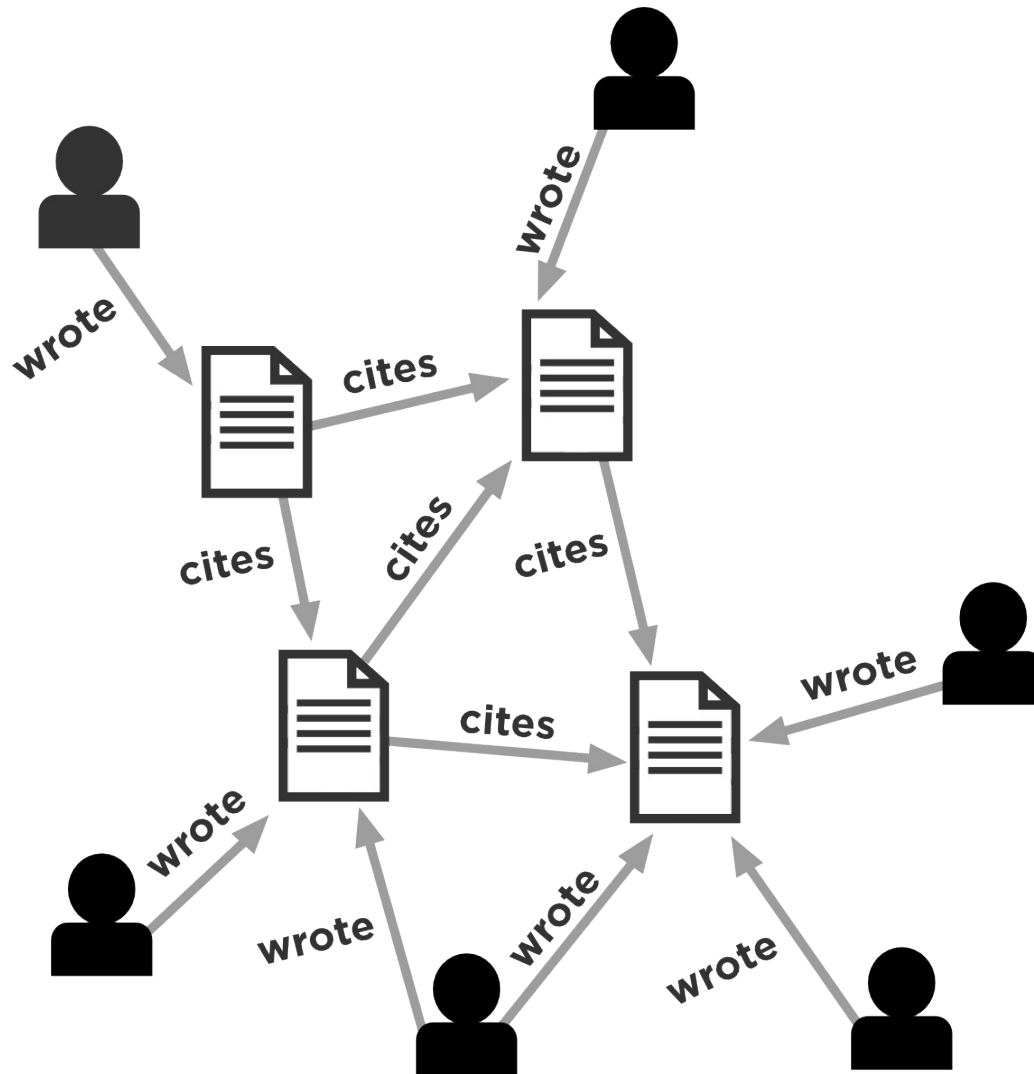
# The Web



# Academic Papers



# Academic Papers





# “Deep belief networks”

SCHOLR.LY Beta



deep belief networks



Search

## Publications

### 1 Deep Belief Networks are compact universal approximators

Nicolas Le Roux, Yoshua Bengio

Deep Belief Networks (DBN) are generative models with many layers of hidden causal variables, recently introduced by Hinton et al. (2006), along with a greedy layer-wise unsupervised learning algorithm. Building on Le Roux and Bengio (2008) and Sutskever and Hinton (2008), we show that deep but narrow generative networks do not require more parameters than shallow ones to ach...

### Investigation of full-sequence training of deep belief networks for speech recognition

Abdel-rahman Mohamed, Dong Yu, Li Deng  
2010

Recently, Deep Belief Networks (DBNs) have been proposed for phone recognition and were found to achieve highly competitive performance. In the original DBNs, only framelevel information was used for training DBN weights while it has been known for long that sequential or full-sequence information can be helpful in improving speech recognition accuracy. In this paper we inves...

### Deep Belief Networks

Ruslan Salakhutdinov, Geoffrey Hinton  
1985

Recently, Hinton and Salakhutdinov (2006) introduced a new method for training deep generative stochastic artificial neural networks that can be used for a variety of tasks. The model consists of two fully connected layers of hidden units, each of which is connected to a layer of visible units. The hidden units are trained using a greedy layer-wise unsupervised learning algorithm. The model is then trained using a greedy layer-wise unsupervised learning algorithm. The model is then trained using a greedy layer-wise unsupervised learning algorithm.

## Authors

### Geoffrey Hinton

University of Toronto, 6 King's College Rd.

RELATED PUBLICATION:

Restricted Boltzmann machines for collaborative filtering



### Ruslan Salakhutdinov

Brain and Cognitive Sciences and CSAIL; Massachusetts Institute of Technology

RELATED PUBLICATION:

Restricted Boltzmann machines for collaborative filtering



### Abdel-rahman Mohamed

Department of Computer Science, University of Toronto

RELATED PUBLICATION:

Deep Belief Networks for phone recognition

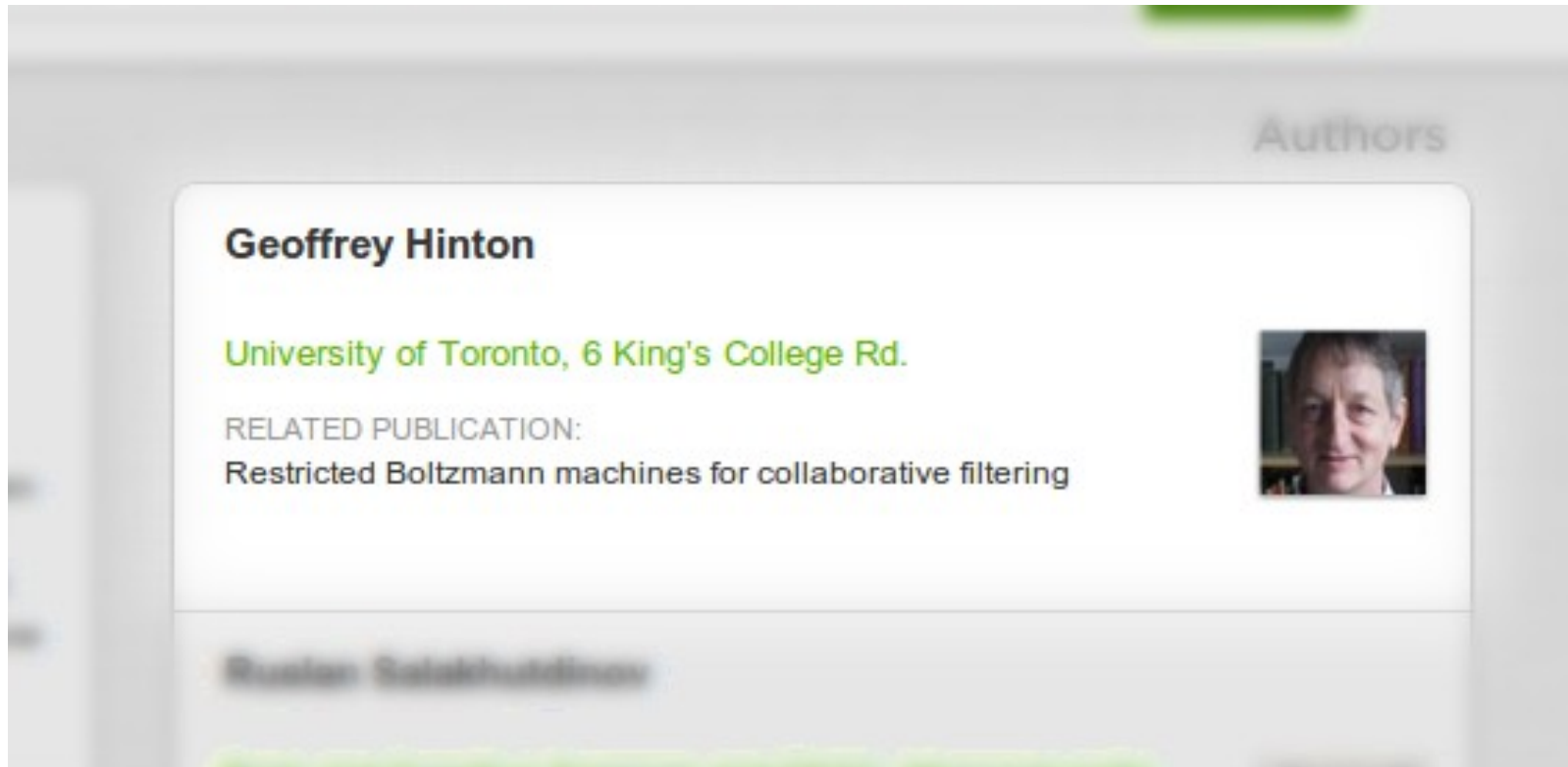


### Iain Murray

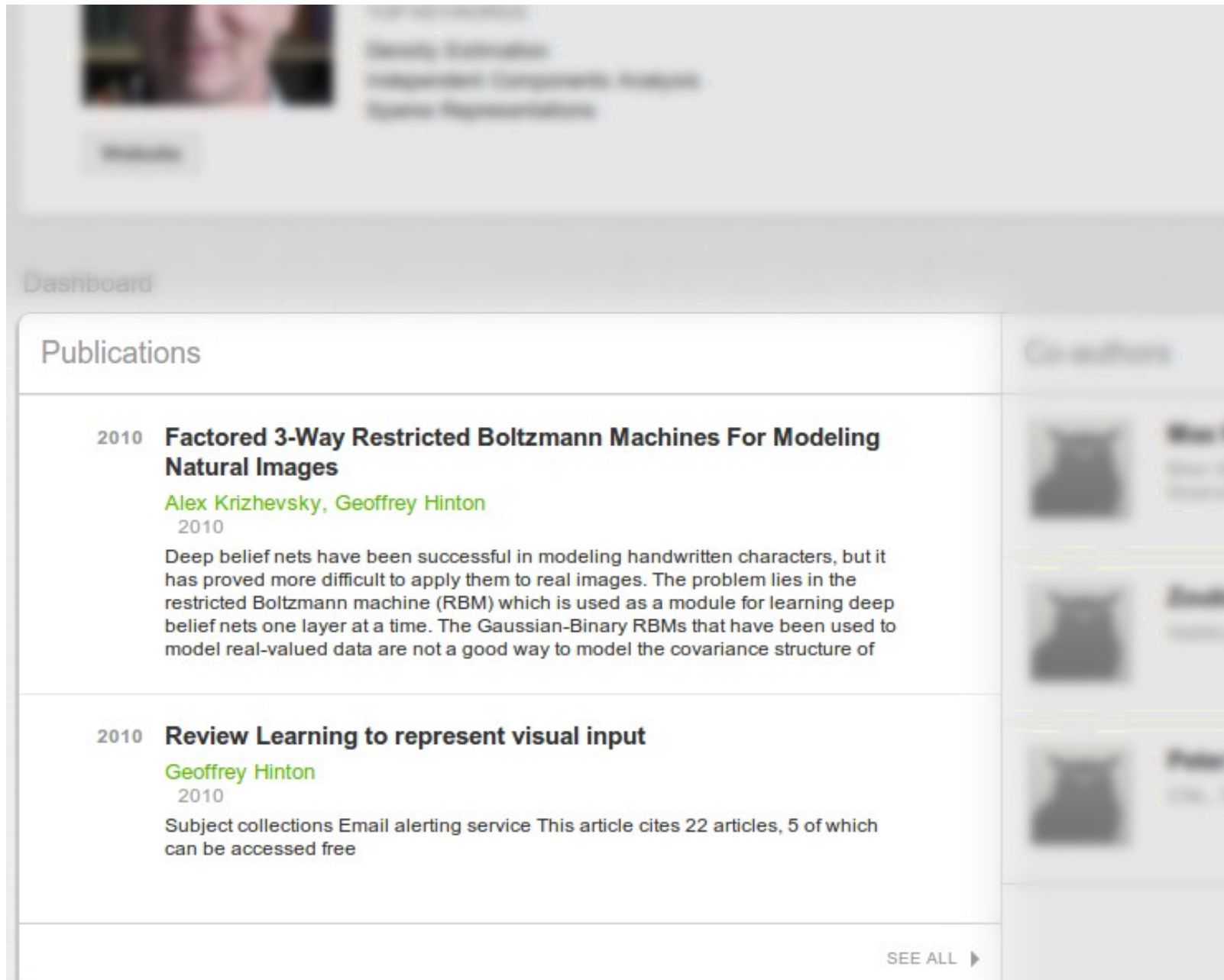
Department of Computer Science, University of Toronto



# Geoffrey Hinton, deep learning pioneer



# Geoffrey Hinton, deep learning pioneer



The image is a screenshot of a research profile page, likely from a university or research institution. At the top, there is a header section with a profile picture and some text, which is partially blurred. Below this is a "Dashboard" section. The main content area is titled "Publications" and lists two entries:

**2010 Factored 3-Way Restricted Boltzmann Machines For Modeling Natural Images**  
[Alex Krizhevsky](#), [Geoffrey Hinton](#)  
2010  
Deep belief nets have been successful in modeling handwritten characters, but it has proved more difficult to apply them to real images. The problem lies in the restricted Boltzmann machine (RBM) which is used as a module for learning deep belief nets one layer at a time. The Gaussian-Binary RBMs that have been used to model real-valued data are not a good way to model the covariance structure of

**2010 Review Learning to represent visual input**  
[Geoffrey Hinton](#)  
2010  
Subject collections Email alerting service This article cites 22 articles, 5 of which can be accessed free


At the bottom right of the publications list, there is a link that says "SEE ALL ►". To the right of the publications list, there is a sidebar with a heading "Co-authors" and a list of small profile pictures.

# Professor Hinton seems like a busy guy

**Ruslan Salakhutdinov**

Brain and Cognitive Sciences and CSAIL; Massachusetts Institute of Technology


RELATED PUBLICATION:  
Restricted Boltzmann machines for collaborative filtering



**Abdel-rahman Mohamed**


Department of Computer Science, University of Toronto

RELATED PUBLICATION:  
Deep Belief Networks for phone recognition



**Iain Murray**

Department of Computer Science, University of Toronto





**Let's Get Technical**

**Why  
Neo4j?**

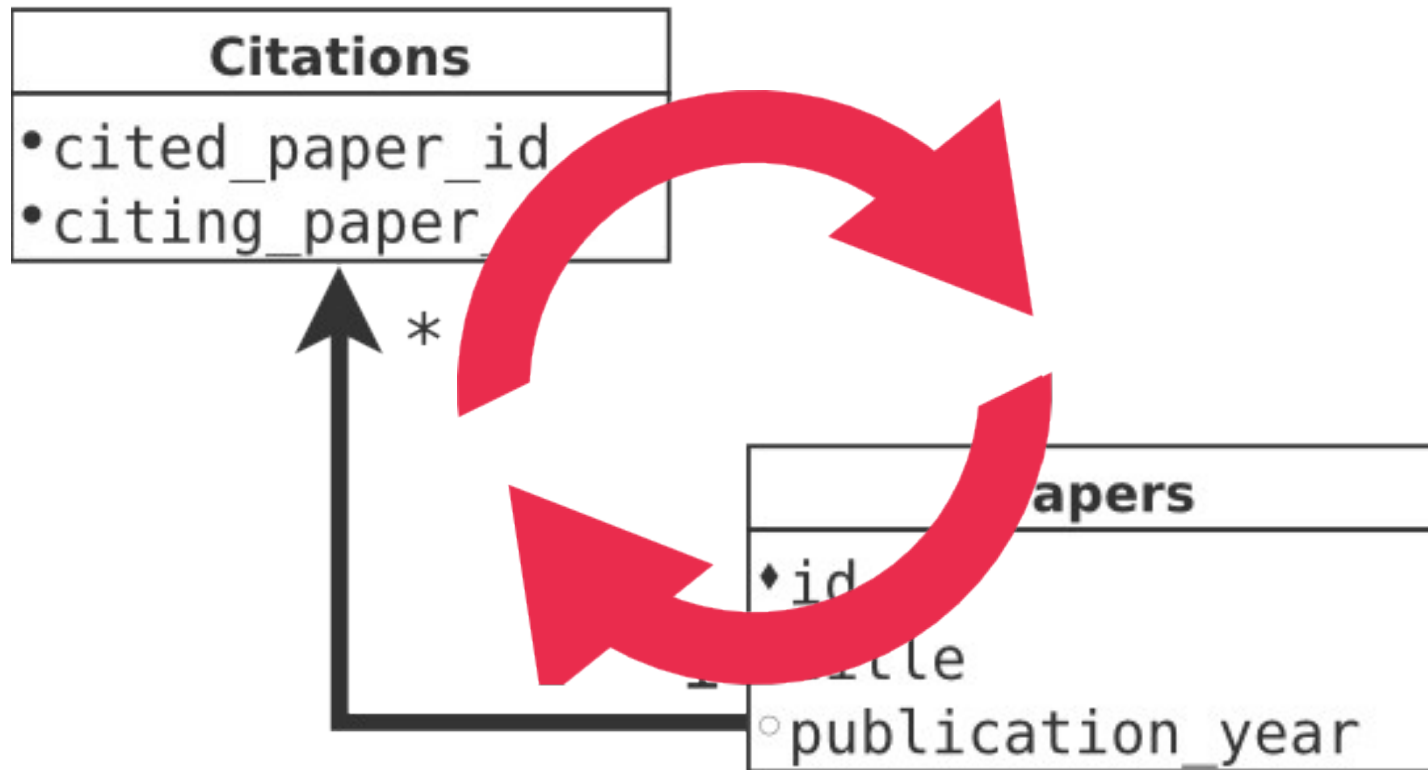
Citations
<ul style="list-style-type: none"> <li>•cited_paper_id</li> <li>•citing_paper_id</li> </ul>



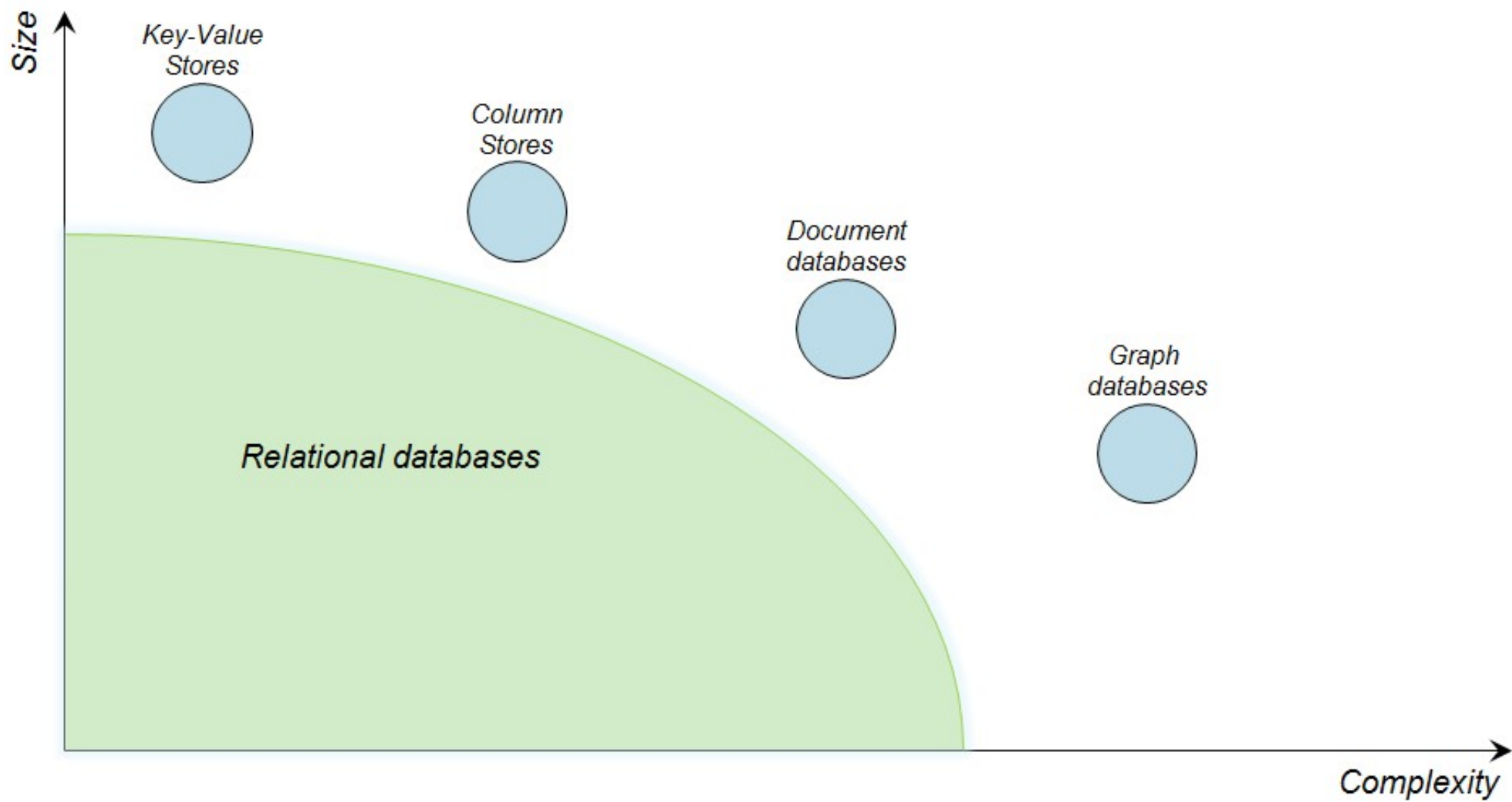
\*

1

Papers
<ul style="list-style-type: none"> <li>♦id</li> <li>○title</li> <li>○publication_year</li> </ul>



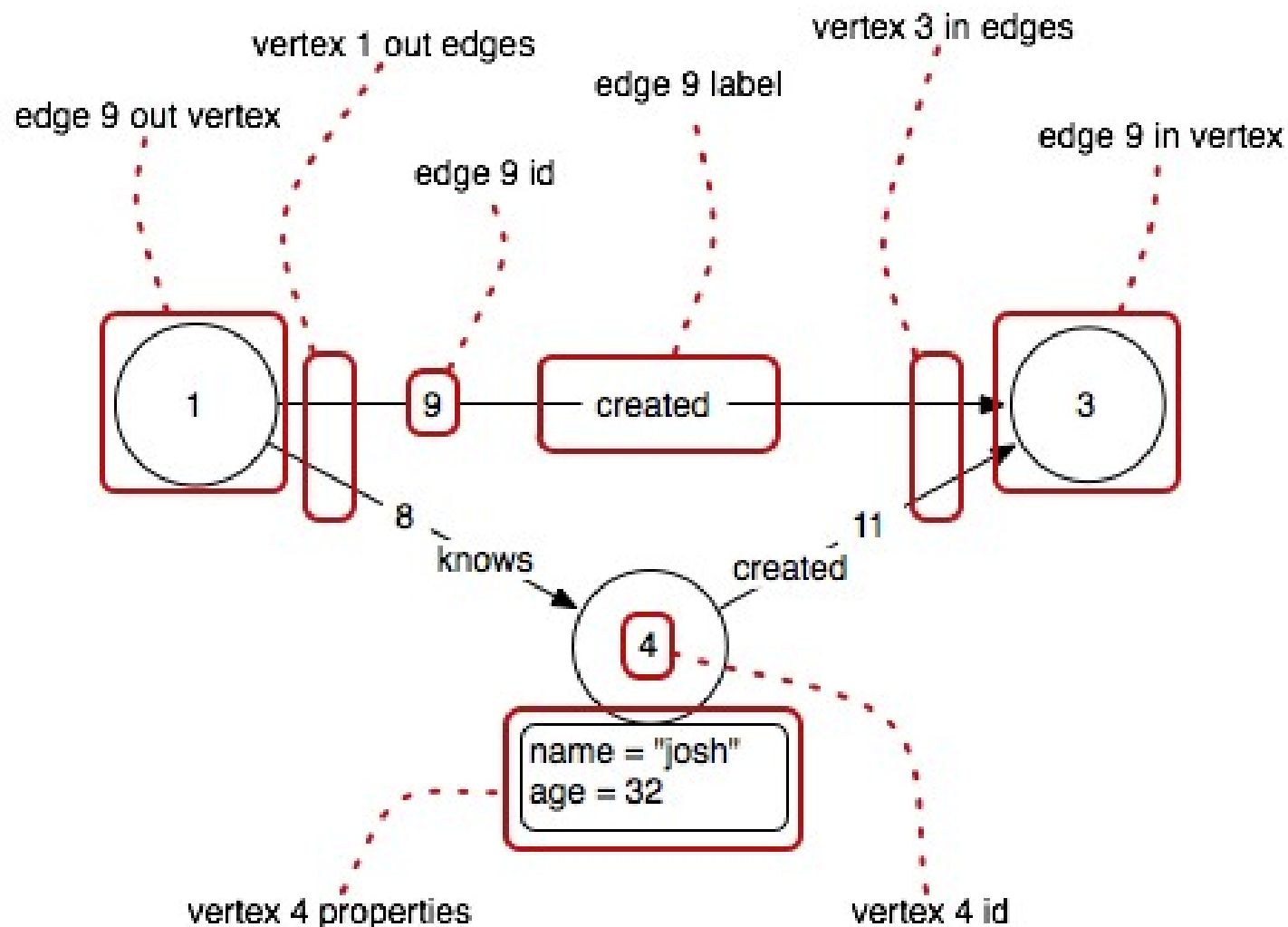






# Use Cases

# A comment on property graphs



# A comment on query examples

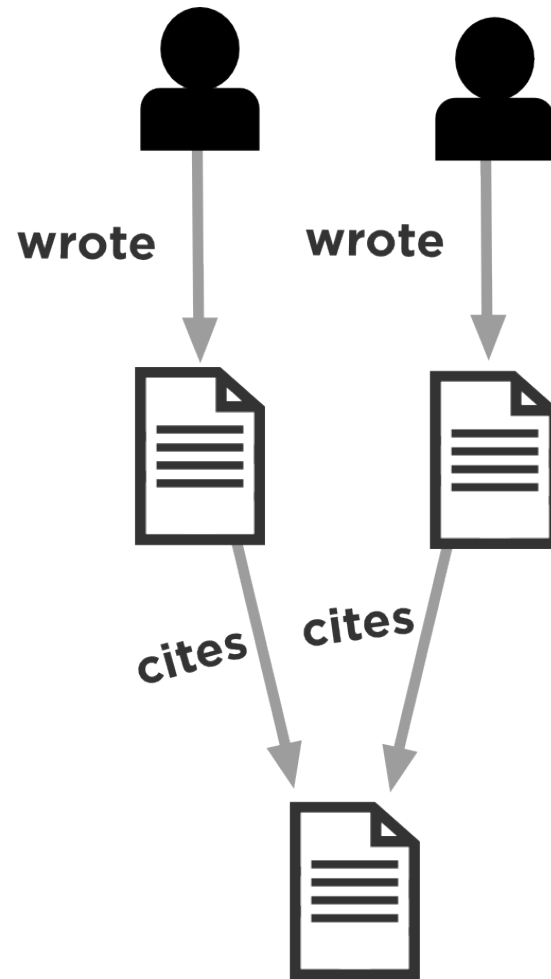
- **Cypher**
  - **Declarative**
  - **SQL-like**
  - **Easy, smooth pattern matching**
  - **Neo4j only**
- **Gremlin**
  - **DSL atop JVM languages like Groovy**
  - **Lower-level, but more powerful**
  - **Cross-database**

**A comment on query examples**

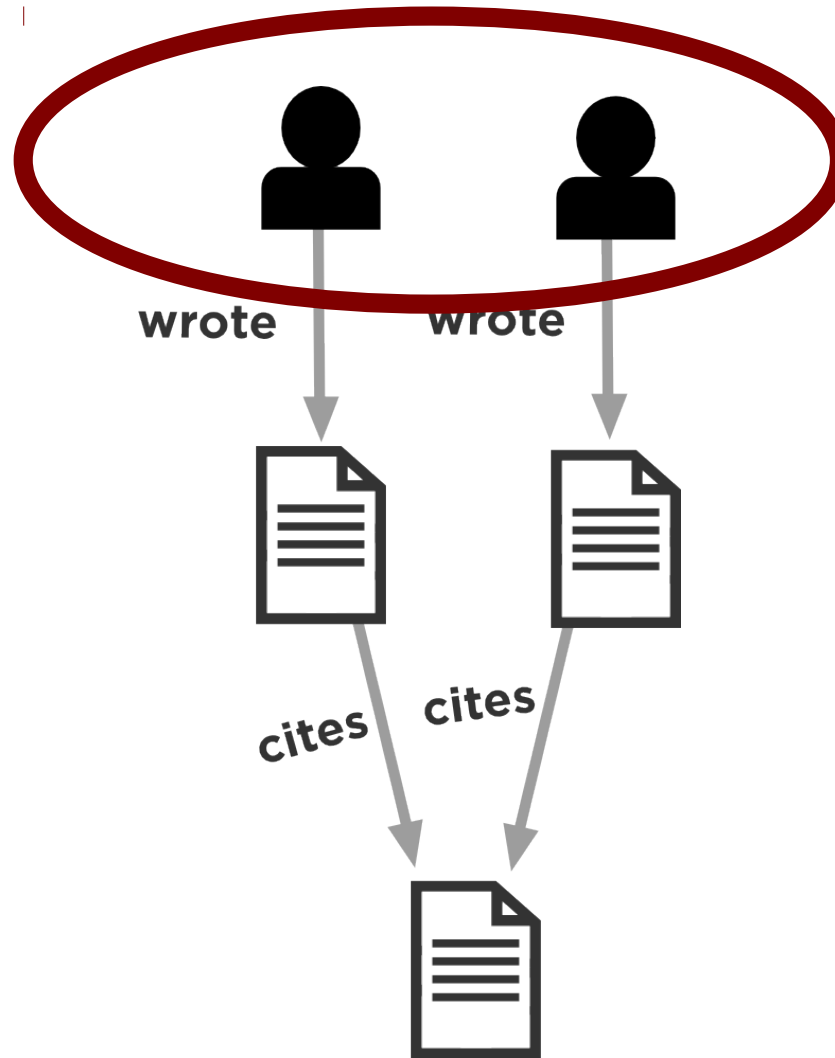
**These snippets are untested.**

# Similar Profiles

# Similar Profiles



# Similar Profiles

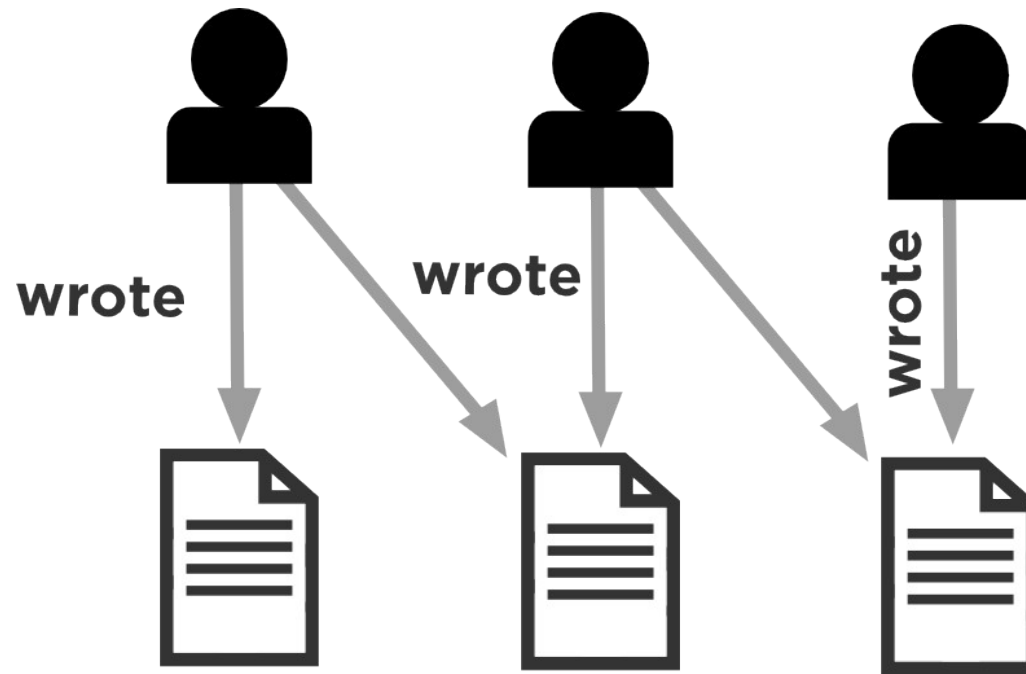




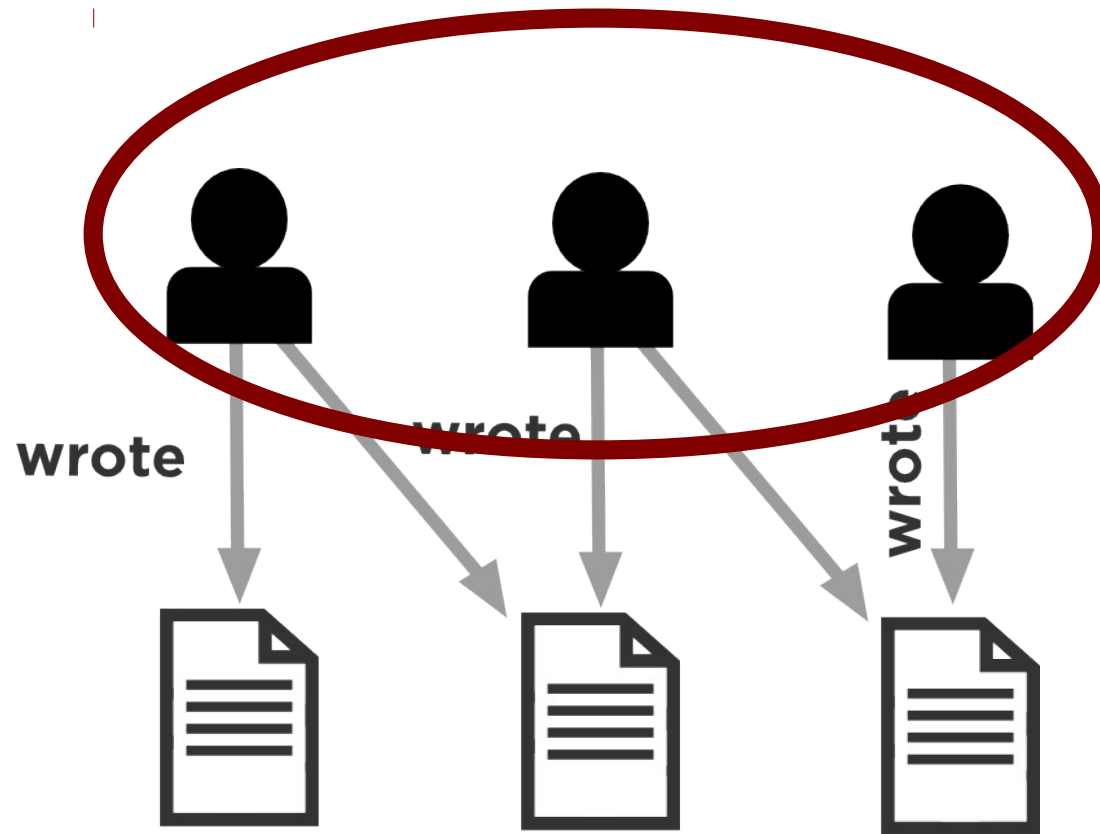
# Similar Profiles

```
START author=node(123)
MATCH author-[:wrote]->(work)-[:cites]->(cited_work) \
        <-[:cites]-(other_works)<-[:wrote]-(other_author)
WITH author, other_author,
        COUNT(cited_work) AS work_in_common
ORDER BY work_in_common
RETURN other_author
```

# Similar Profiles



# Similar Profiles



# Similar Profiles

```
START author=node(123)
MATCH author-[:wrote]->(work)<-[:wrote]-(coauthor) \
        -[:wrote]->(other_work) \
        <-[:wrote]-(second_coauthor)
WITH author, second_coauthor,
        COUNT(coauthor) AS shared_coauthors
ORDER BY shared_coauthors
RETURN second_coauthor
```

# Entity Resolution

# Entity Resolution

**How do we reconcile different data sources?**

**What happens when people share names?**

**How do we know who's who?**

# Entity resolution is an active area of research.

SCHOLR.LY Beta



entity resolution



Search

## Publications

### Abstract A Two-Step Classification Approach to Unsupervised Record Linkage

Peter Christen

Linking or matching databases is becoming increasingly important in many data mining projects, as linked data can contain information that is not available otherwise, or that would be too expensive to collect manually. A main challenge when linking large databases is the classification of the compared record pairs into matches and non-matches. In traditional record linkage, c...

### The Role of Asserted Resolution in Entity Identity Information Management

Yinle Zhou, John R. Talburt

Abstract – This paper introduces the concept of asserted resolution as a technique for entity resolution. In asserted resolution trusted information sources are used to force the equivalence (or non-equivalence) of entity references and identity structures regardless of matching conditions. The paper proposes five specific forms of assertion to support entity identity informa...

### Identifying Graphs From Noisy and Incomplete Data

Galileo Mark, S. Namata, Lise Getoor

There is a growing wealth of data describing networks of various types, including social networks, physical networks such as transportation or communication networks, and biological networks. At the same time, there is a growing interest in analyzing these networks, in order to uncover general laws

## Authors

### Lise Getoor

Department of Computer Science; University of Maryland, College Park



RELATED PUBLICATION:

Learning Probabilistic Models of Relational Structure

### George Varghese

RELATED PUBLICATION:

Multiway Range Trees: Scalable IP Lookup with Fast Updates



### John N Zigman

Real-Time Entity Resolution. August 2009.

RELATED PUBLICATION:

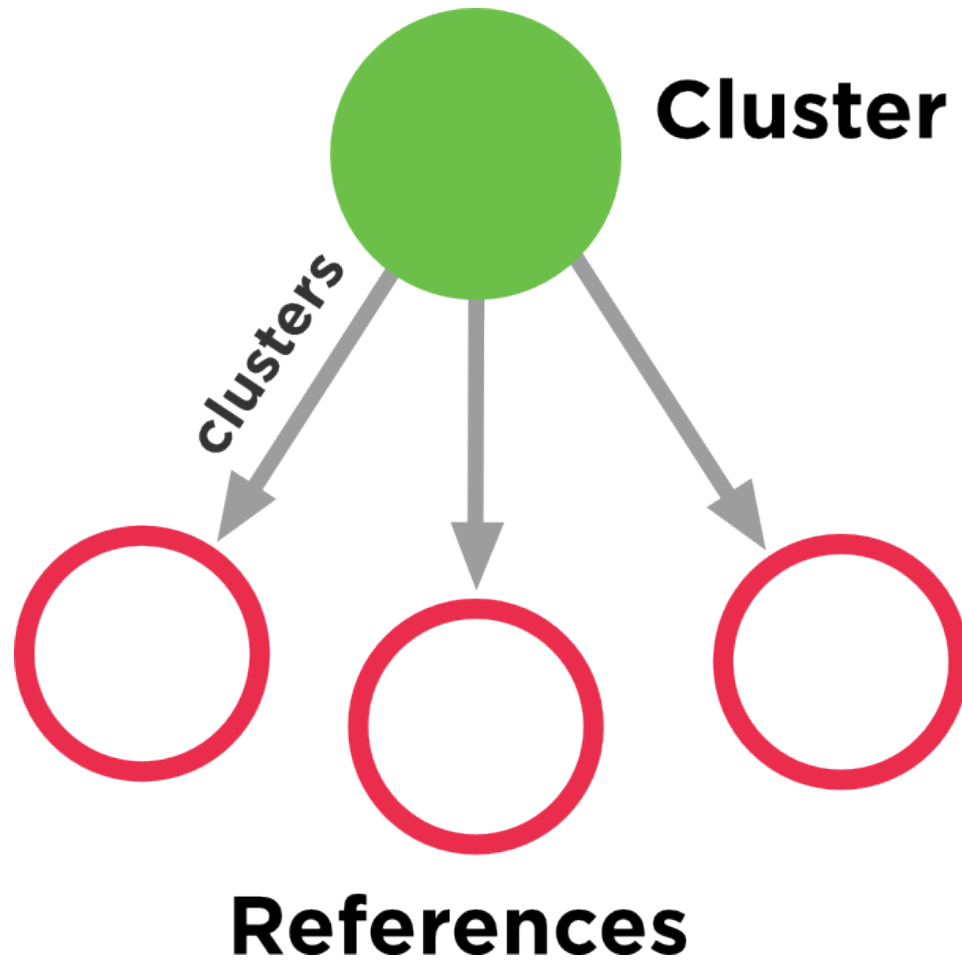
Countries – A Pilot Study



### Peter Christen

Department of Computer Science, The Australian National University,



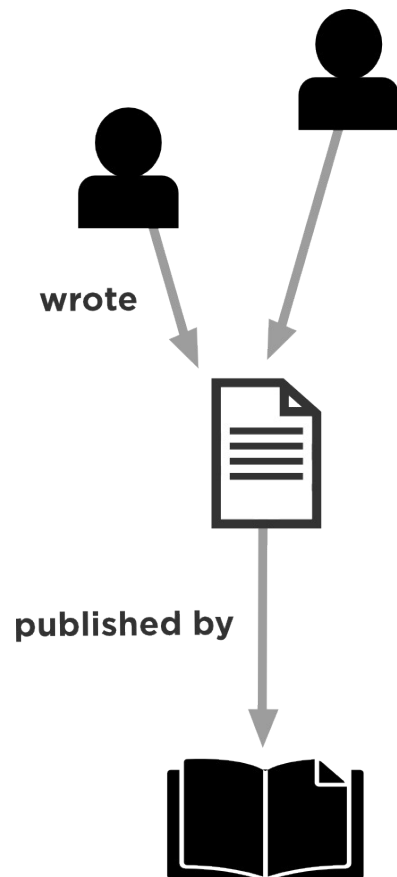




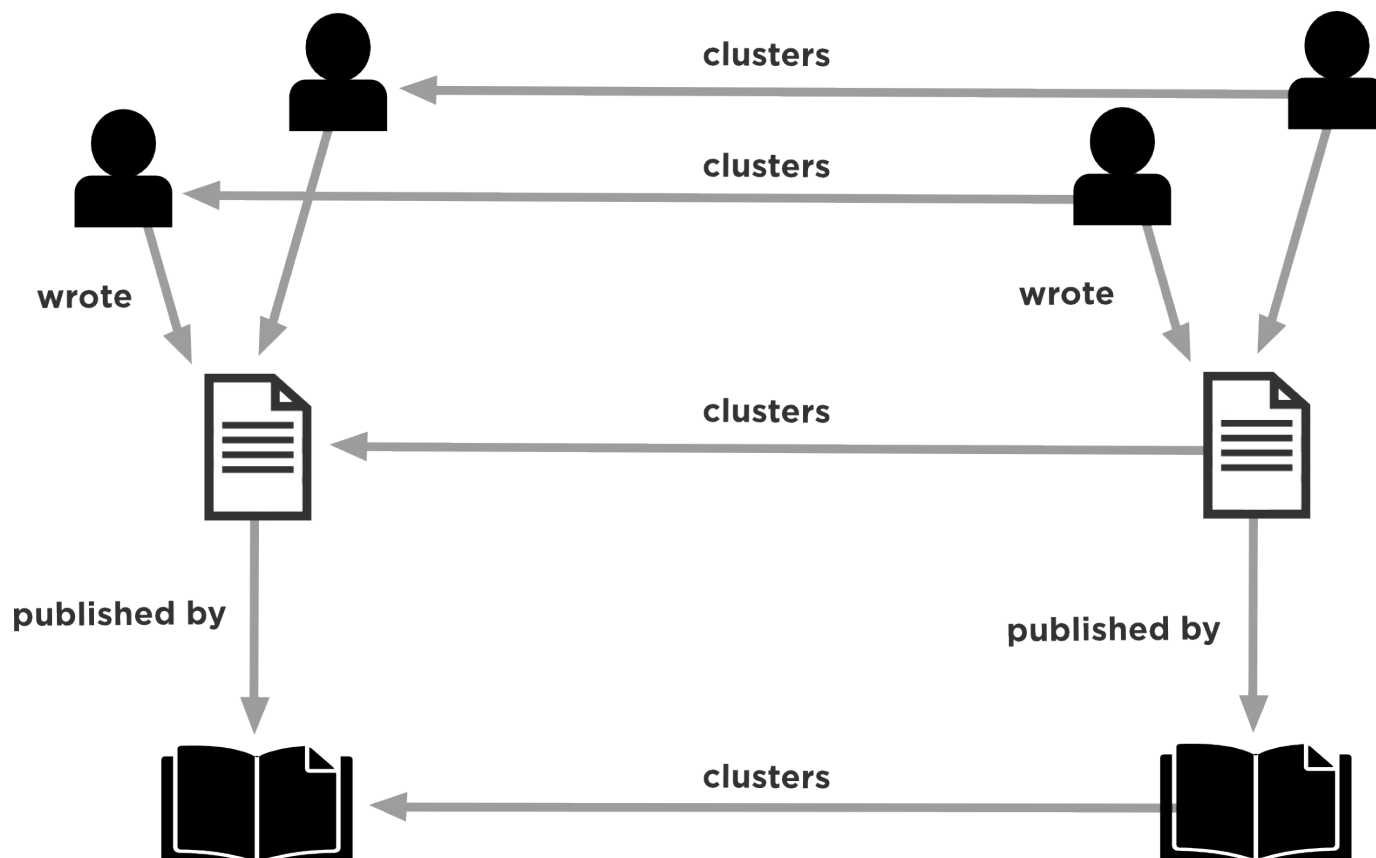
**E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.**

**E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.**

**E. Agichtein** and **L. Gravano**. **Snowball: Extracting relations from large plain-text collections**. In **Proceedings of the Fifth ACM International Conference on Digital Libraries**, 2000.



**E. Agichtein and L. Gravano. Snowball: Extracting relations from large plain-text collections. In Proceedings of the Fifth ACM International Conference on Digital Libraries, 2000.**



# Entity Resolution

```
votes = [:]

g.v(clusterIds).out('clusters').map.each{ properties ->
    properties.each{
        votes[it.key] = (votes[it.key] ? : [:])
        votes[it.key][it.value] = \
            (votes[it.key][it.value] ? : 0) + 1
    }
}

newClusterProperties = votes.collectEntries{prop,
valueVotes ->
    [prop, valueVotes.sort{-it.value}[0].key]
}
```

# Search

# Search

**We'd like to show  
the expected  
publication results  
on the left.**

# Search

We'd like to show  
the expected  
publication results  
on the left.

**On the right, we  
want to show  
influencers based  
on the publication  
results.**



# Users can search for a name + a topic.

SCHOLR.LY Beta



c lee giles digital libraries



Search

## Publications

### Libraries

Alan F. Smeaton

Widespread use of the internet has resulted in digital libraries that are increasingly used by diverse communities of users for diverse purposes and in which sharing and collaboration have become important social elements. As such libraries become commonplace, as their contents and services become more varied and as their patrons become more experienced with computer technolog...

### e-Science and its implications for the library community Tony

Purpose: To explain the nature of the 'e-Science' revolution in 21 st century scientific research and its consequences for the library community.

Design/methodology/approach: The concepts of e-Science are illustrated by a discussion of the CombeChem, eBank and SmartTea projects. The issue of open access is then discussed with reference to arXiv, PubMed Central and EPrints. T...

### Image Libraries and the Internet

Harold S. Stone

1999

IEEE Communications Magazine . January 1999 100 Internet also supports access to audio and video clips, but these are infrequently used at present. By another measure, bits of information, graphics well exceeds the volume of

## Authors

### C. Lee Giles

The Pennsylvania State University

RELATED PUBLICATION:

Winners don't take all: Characterizing the competition for links on the web



### Alan F. Smeaton

2; CLARITY: Centre for Sensor Web Technologies

RELATED PUBLICATION:

Challenges for Content-Based Navigation of Digital Video in the Fischlár Digital Library



### Prasenjit Mitra

College of Information Sciences and Technology and  
Department of Computer Science and Engineering, The  
Pennsylvania State University

RELATED PUBLICATION:

Tableseer: Automatic table metadata extraction and searching in digital libraries



### Seyda Ertekin

1: Department of Computer Science and Engineering; 2:



# Search

```
authorCounts = [:]
```

```
g.v(publicationIds).in('WROTE').\  
    groupCount(authorCounts).iterate()
```

```
return authorCounts.collect{author, count →  
    [author, count * authorBoost]}.sort{-it[1]}  
}
```

# Search

```
authorCounts = [:]  
coauthorCounts = [:]  
  
g.v(publicationIds).in('WROTE') \  
    .groupCount(authorCounts).out('WROTE') \  
    .in('WROTE').groupCount(coauthorCounts) \  
    .iterate()  
  
// IMAGINE - poor man's "histogram"  
  
totalAuthorCounts = [:]  
  
return totalAuthorCounts.sort(-it.value}
```

# Search

```
citedAuthorCounts = [:]
```

```
g.v(publicationIds).out('cite').in('wrote') \  
    .groupCount(citedAuthorCounts)
```

**Still with me?**

**<http://scholr.ly>**



**Get Involved**

**neo4j.org**

**[meetup.com/graph-database-austin](https://meetup.com/graph-database-austin)**

# Upcoming Events

**April 15<sup>th</sup> – Austin Graph DB Meetup**

**April 16<sup>th</sup> – Austin Neo4j Tutorial**

**More details at  
[meetup.com/graph-database-austin](https://meetup.com/graph-database-austin)**



**Questions?**





**Thanks!**

# **Bibliography**

**Nicholas Menghini and Alex Fuller from the Noun Project – thanks for the icons!**

**TinkerPop – thanks for the graphic!**