CSE 6242/ CX 4242

Feb 18, 2014

Graphs / Networks

Centrality measures, algorithms, interactive applications

Duen Horng (Polo) Chau

Georgia Tech

Partly based on materials by
Professors Guy Lebanon, Jeffrey Heer, John Stasko, Christos Faloutsos, Le Song

Joint IC/CSE Seminar Announcement
Thursday (2/20), 10-11am, TSRB Banquet Hall

Design Techniques for Crowdsourcing Complex Tasks

Edith Law
Harvard University
(Graduated from CMU)

Edith is a faculty candidate; she'll talk about some of her best work
0.5% bonus point for attending.
Talk may be recorded

Recap…

- **Last time:** Basics, how to build graph, store graph, laws, etc.
- **Today:** Centrality measures, algorithms, interactive applications for visualization and recommendation

Centrality

= “Importance”

Why Node Centrality?

What can we do if we can rank all the nodes in a graph (e.g., Facebook, LinkedIn, Twitter)?

- Find **celebrities** or influential people in a social network (Twitter)
- Find "**gatekeepers**" who connect communities (headhunters love to find them on LinkedIn)
- What else?

More generally

Helps **graph analysis, visualization, understanding**, e.g.,

- let us **rank** nodes, group or study them by centrality
- only show subgraph formed by the **top 100 nodes**, out of the millions in the full graph
  - **similar to google search results** (ranked, and they only show you 10 per page)
- Most graph analysis packages already have centrality algorithms implemented. **Use them!**

Can also compute edge centrality.
Here we focus on node centrality.

Degree Centrality (easiest)

Degree = number of neighbors

For directed graphs

- in degree = # incoming edges
- out degree = # outgoing edges

Algorithms?

- Sequential scan through edge list
- What about for a graph stored in SQLite?

Computing degrees using SQL

Recall simplest way to store a graph in SQLite:

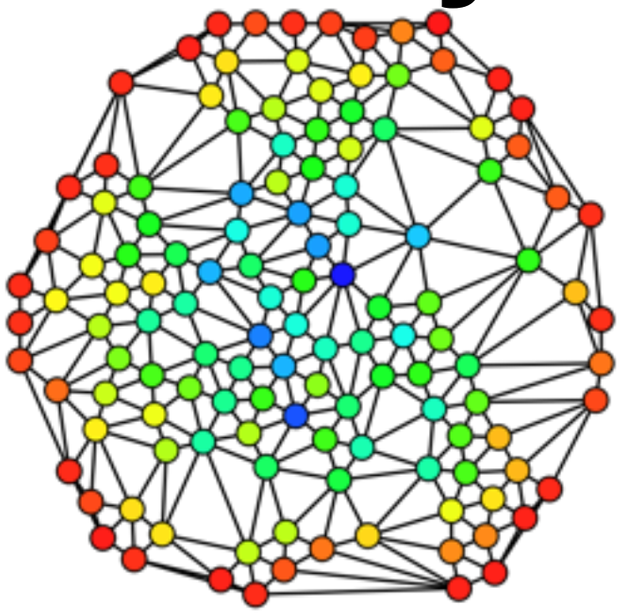
```
edges(source_id, target_id)
```

1. Create index for each column

2. Use **group by** statement to find node degrees

```
select count(*) from edges group by source_id;
```

Betweenness Centrality



High betweenness

= important “gatekeeper” or liaison

Betweenness of a node v

$$= \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Number of shortest paths between s and t that involves v

Number of shortest paths between s and t

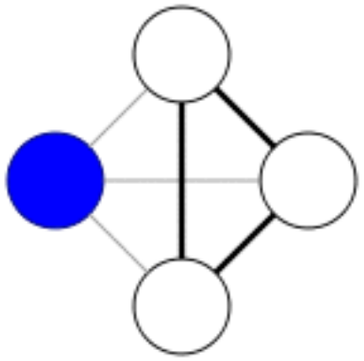
= how often a node serves as the “bridge” that connects two other nodes.

Clustering Coefficient

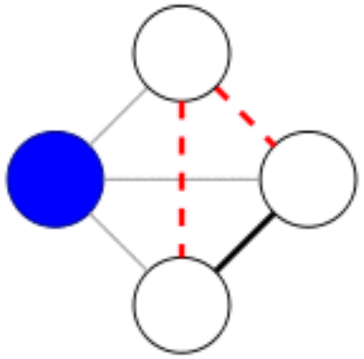
A node's clustering coefficient is a measure of how close the node's neighbors are from forming a clique.

- 1 = neighbors form a clique
- 0 = No edges among neighbors

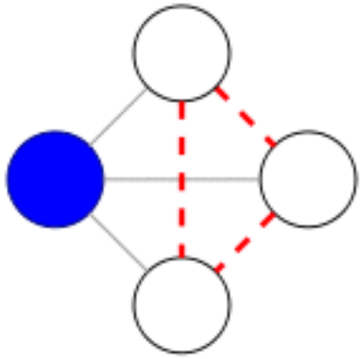
(Assuming undirected graph)



c = 1



c = 1/3



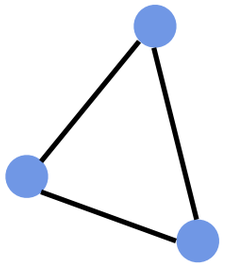
c = 0

Computing Clustering Coefficient...

Requires triangle counting

Real social networks have a lot of triangles

- Friends of friends are friends



But: triangles are expensive to compute

(3-way join; several approx. algos)

Can we do that quickly?

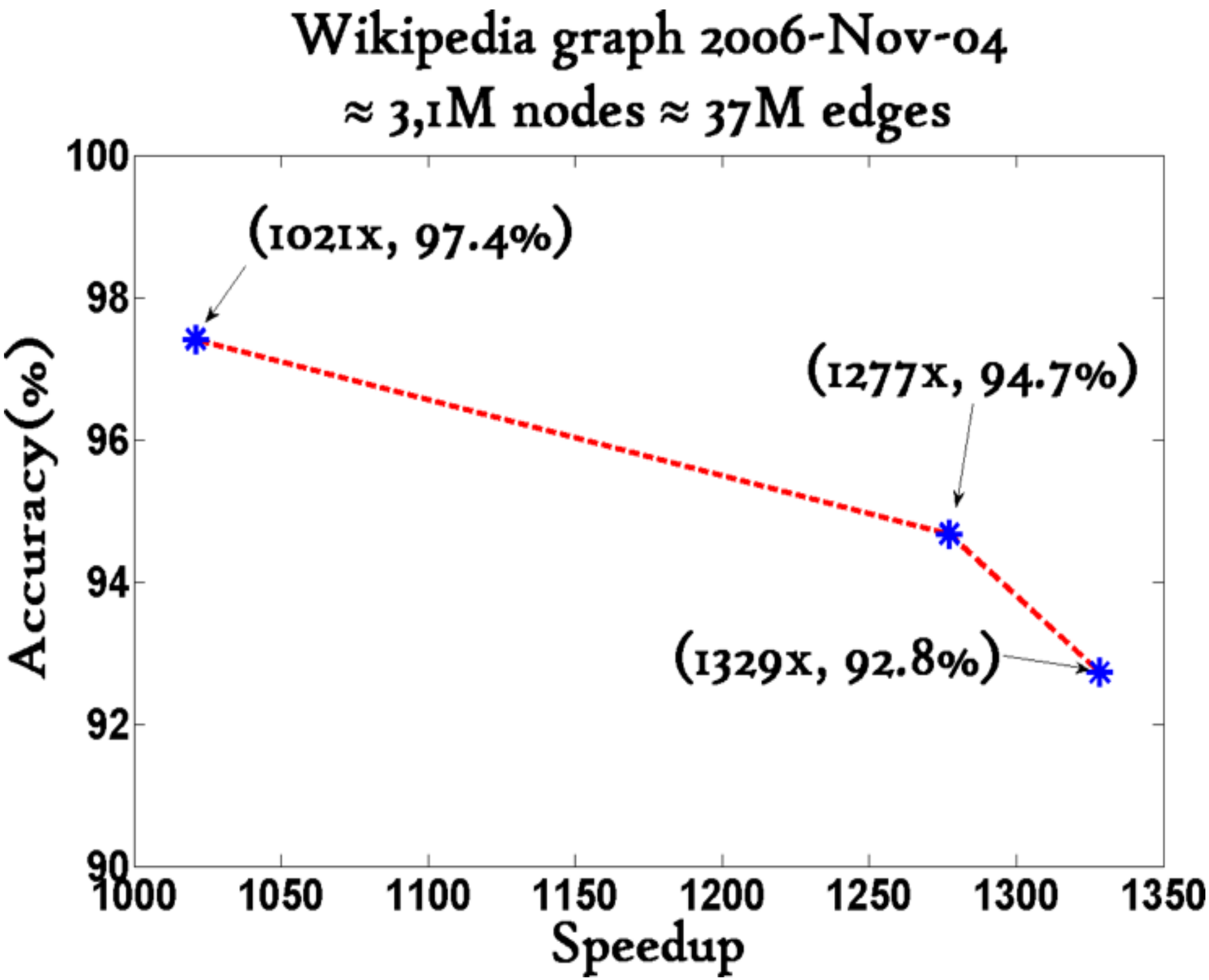
details

Super Fast Triangle Counting
[Tsourakakis ICDM 2008]

But: triangles are expensive to compute
(3-way join; several approx. algos)
Q: Can we do that quickly?
A: Yes!

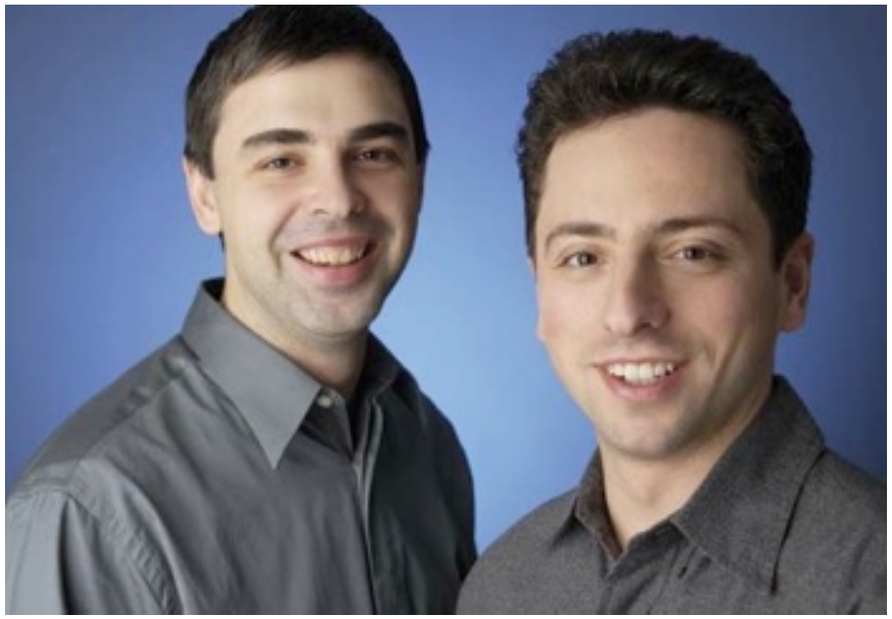
#triangles = 1/6 Sum ((λi)^3)

(and, because of skewness,
we only need the top few eigenvalues!



1000x+ speed-up, >90% accuracy

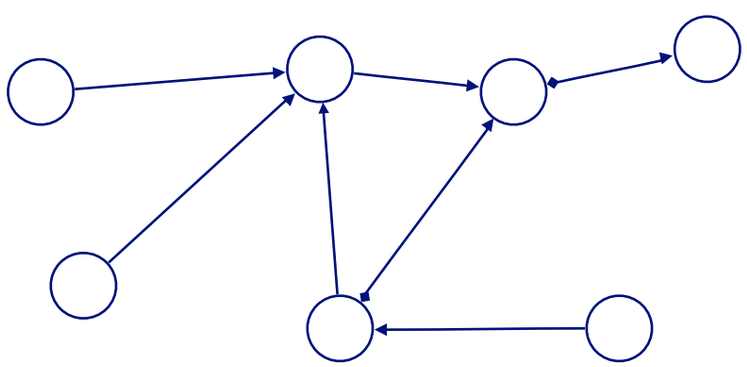
PageRank (Google)

Larry Page  Sergey Brin

Brin, Sergey and Lawrence Page (1998). *Anatomy of a Large-Scale Hypertextual Web Search Engine*. 7th Intl World Wide Web Conf.

PageRank: Problem

Given a directed graph, find its most interesting/central node

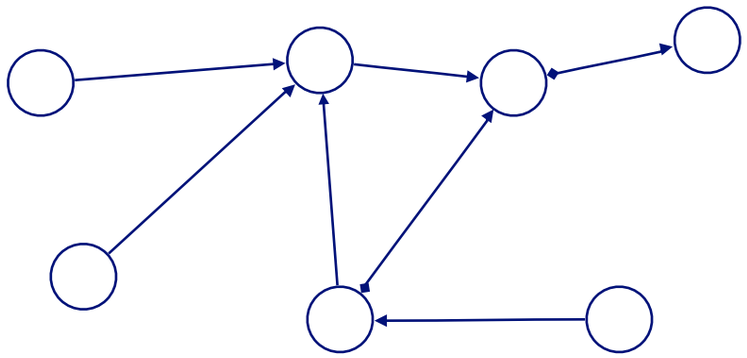


A node is important, if it is connected with important nodes (recursive, but OK!)

PageRank: Solution

Given a directed graph, find its most interesting/central node

Proposed solution: use **random walk**; spot most 'popular' node (-> steady state probability (ssp))

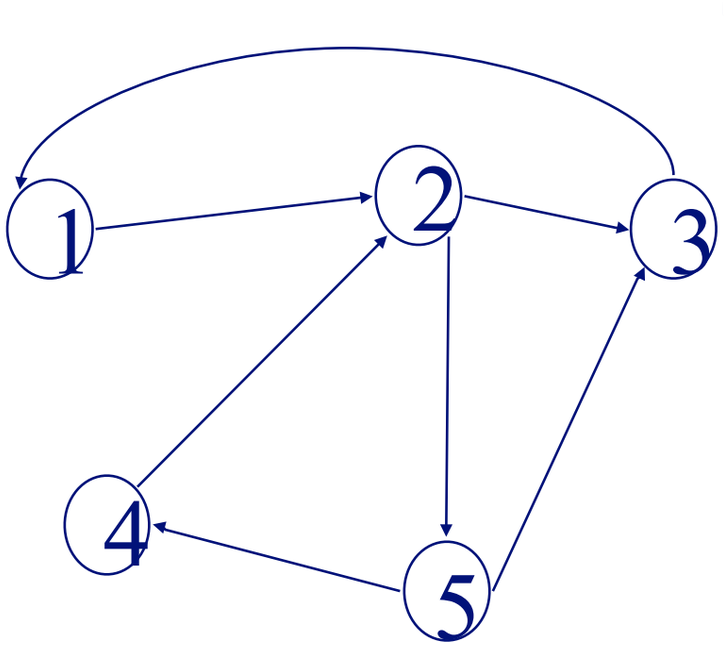


"state" = webpage

A node has high ssp, if it is connected with high ssp nodes (recursive, but OK!)

(Simplified) PageRank

Let **B** be the transition matrix:
transposed, column-normalized



From

To

B

		1		
1			1	
	1/2			1/2
				1/2
	1/2			

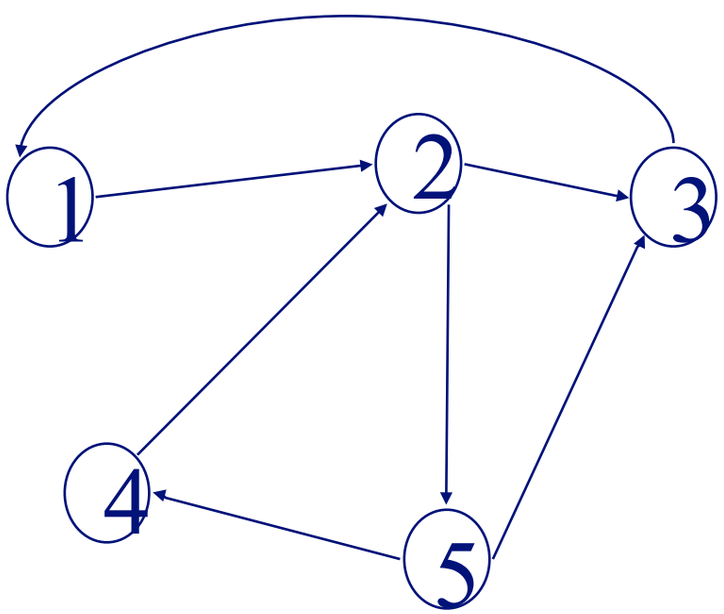
p1
p2
p3
p4
p5

=

p1
p2
p3
p4
p5

(Simplified) PageRank

B p = p



$$\mathbf{B}\ \mathbf{p} = \mathbf{p}$$

$$\begin{bmatrix} & & 1 & & \\ 1 & & & 1 & \\ & 1/2 & & & 1/2 \\ & & & & 1/2 \\ & 1/2 & & & \end{bmatrix} \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix} = \begin{bmatrix} p1 \\ p2 \\ p3 \\ p4 \\ p5 \end{bmatrix}$$

(Simplified) PageRank

- $\mathbf{B}\ \mathbf{p} = 1 * \mathbf{p}$
- thus, $\mathbf{p}$ is the **eigenvector** that corresponds to the highest eigenvalue (=1, since the matrix is column-normalized)
- Why does such a $\mathbf{p}$ exist?
  - $\mathbf{p}$ exists if $\mathbf{B}$ is nxn, nonnegative, irreducible [Perron–Frobenius theorem]

(Simplified) PageRank

- In short: imagine a particle randomly moving along the edges
- compute its steady-state probability (ssp)

Full version of algorithm:
with occasional random jumps

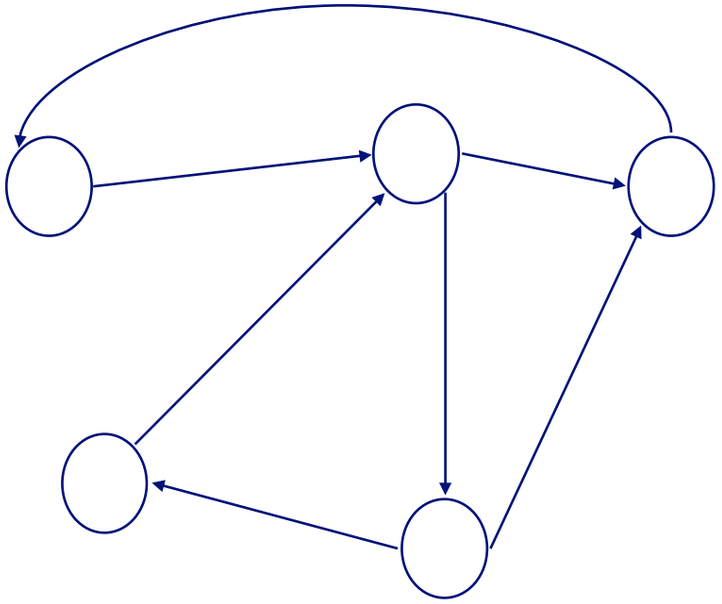
Why? To make the matrix irreducible

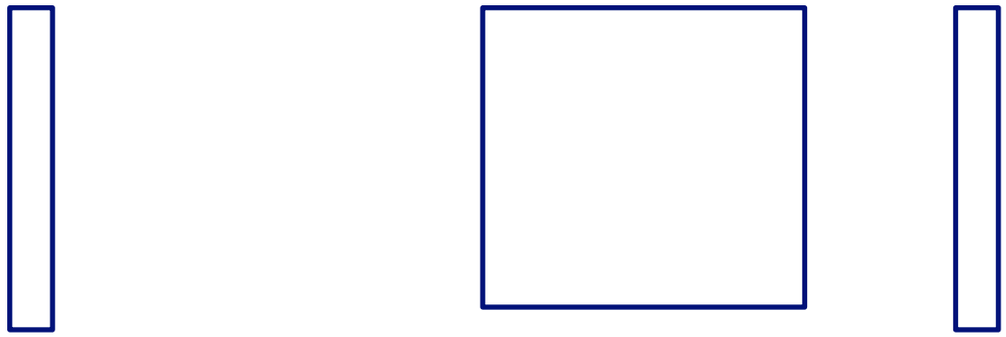
Full Algorithm

- With probability $1\text{-}c$, fly-out to a random node
- Then, we have

$$\mathbf{p} = c\,\mathbf{B}\,\mathbf{p} + (1-c)/n\ \mathbf{1} \Rightarrow$$

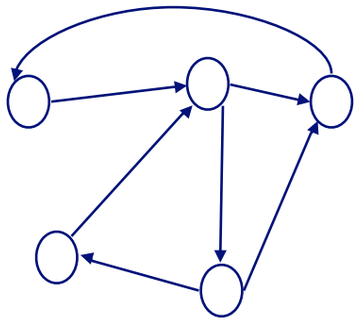
$$\mathbf{p} = (1-c)/n\ \ [\mathbf{I} - c\,\mathbf{B}]^{-1}\ \ \mathbf{1}$$





Full Algorithm

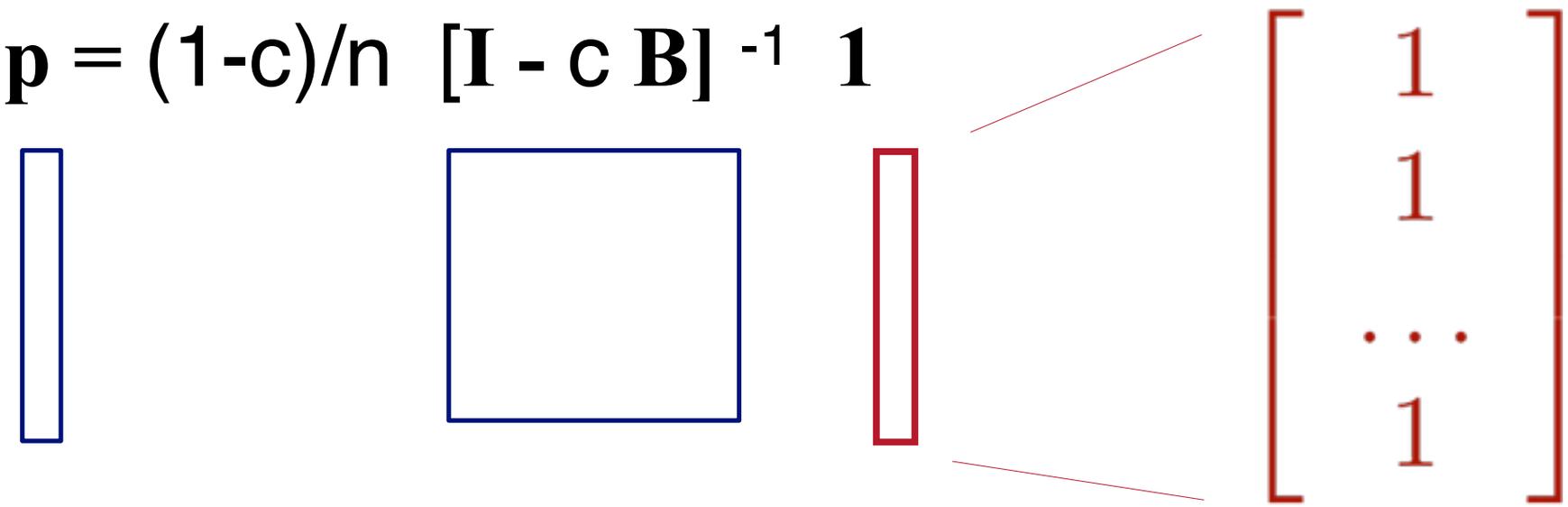
- With probability $1\text{-}c$, fly-out to a random node



- Then, we have

$$\mathbf{p} = c\,\mathbf{B}\,\mathbf{p} + (1\text{-}c)/n\ \mathbf{1} \Rightarrow$$

$$\mathbf{p} = (1\text{-}c)/n\ \ [\mathbf{I} - c\,\mathbf{B}]^{-1}\ \ \mathbf{1}$$



PageRank for graphs (generally)

You can compute PageRank for **any graphs**

Should be in your algorithm “toolbox”

- Better than simple centrality measure (e.g., degree)
- Fast to compute for large graphs (O(E))

But can be “misled” (Google Bomb)

- How?

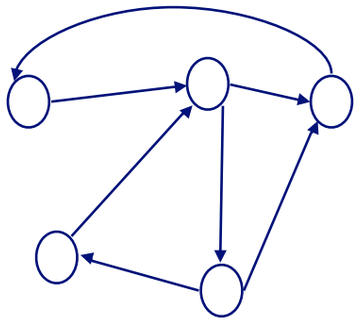
Personalized PageRank

Make one small variation of PageRank

- Intuition: not all pages are equal, some more relevant to a person's specific needs
- How?

“Personalizing” PageRank

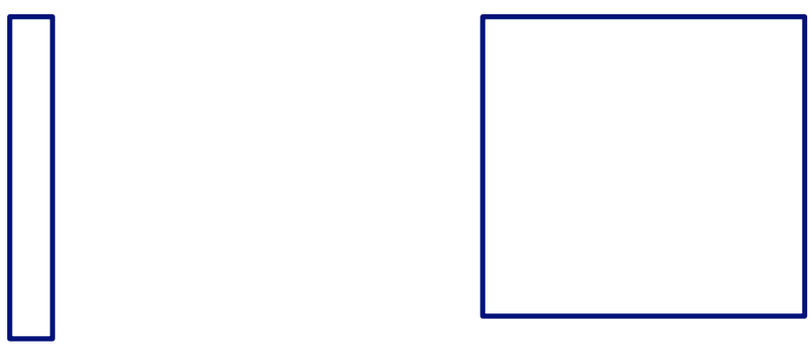
- With probability $1\text{-}c$, fly-out to ~~a random node~~ **some preferred nodes**



- Then, we have

$$\mathbf{p} = c\,\mathbf{B}\,\mathbf{p} + (1\text{-}c)/n\ \mathbf{1} \Rightarrow$$

$$\mathbf{p} = (1\text{-}c)/n\ \ [\mathbf{I} - c\,\mathbf{B}]^{-1}\ \ \mathbf{1}$$



$$\begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ 0 \\ \cdots \\ 0 \\ 1 \end{bmatrix}$$

Why learn Personalized PageRank?

Can be used for **recommendation**, e.g.,

- If I like this webpage, what would I also be interested?
- If I like this product, what other products I also like? (in a user-product bipartite graph)

- Also helps with **visualizing large graphs**
  - Instead of visualizing every single nodes, visualize the **most important ones**

Again, very flexible. Can be run on any graph.

Building an interactive application

Will show you an example application (Apolo) that uses a **"diffusion-based" algorithm** to perform recommendation on a large graph

- **Personalized PageRank**
  (= Random Walk with Restart)
- Belief Propagation
  (powerful inference algorithm, for fraud detection, image segmentation, error-correcting codes, etc.)
- "Spreading activation" or "degree of interest" in Human-Computer Interaction (HCI)
- Guilt-by-association techniques

Building an interactive application

Why diffusion-based algorithms are widely used?

- **Intuitive to interpret**
  uses “network effect”, homophily, etc.
- **Easy to implement**
  Math is relatively simple
- **Fast**
  run time linear to #edges, or better
- **Probabilistic** meaning

Human-In-The-Loop Graph Mining

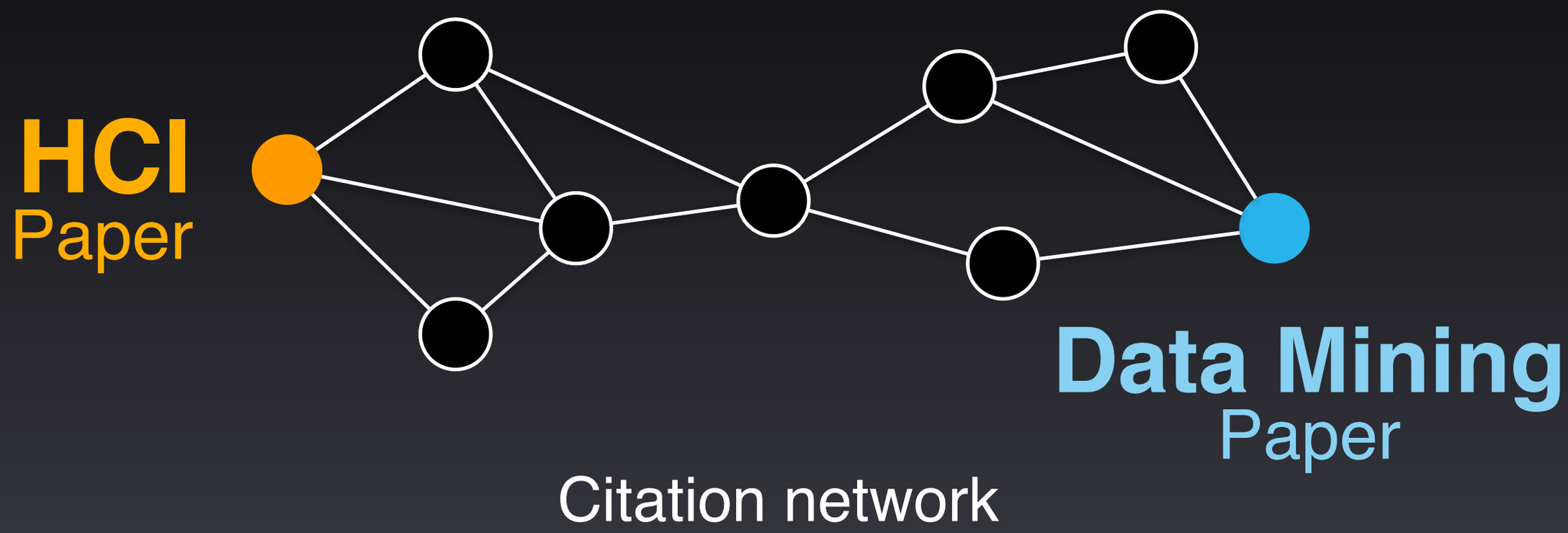
Apolo:

Machine Learning + Visualization

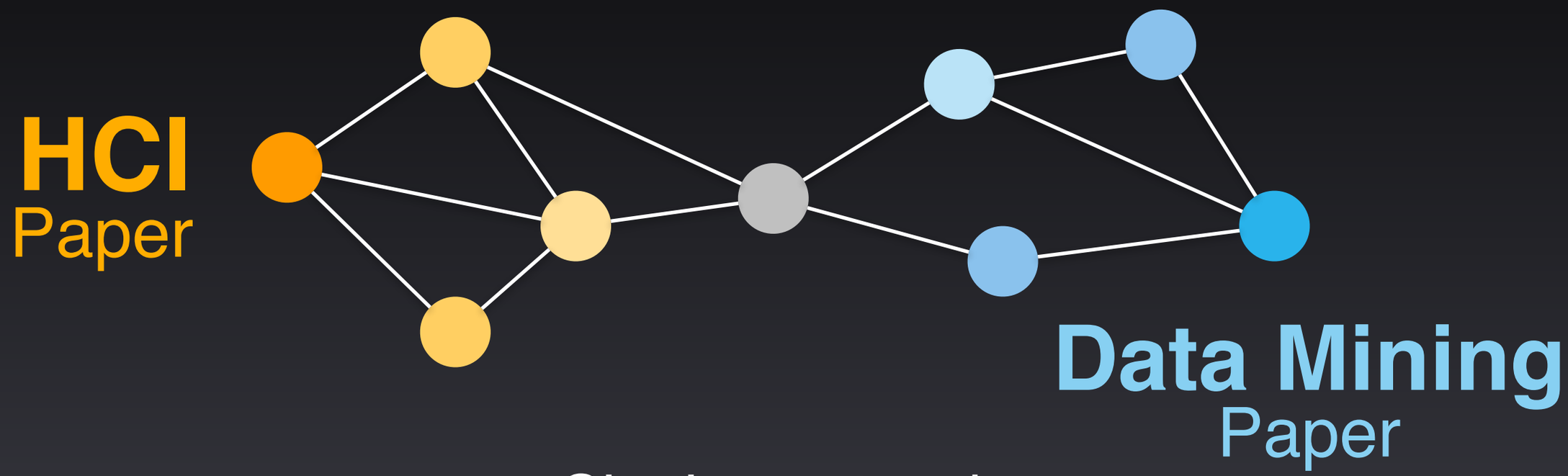
CHI 2011

Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning

Finding More Relevant Nodes

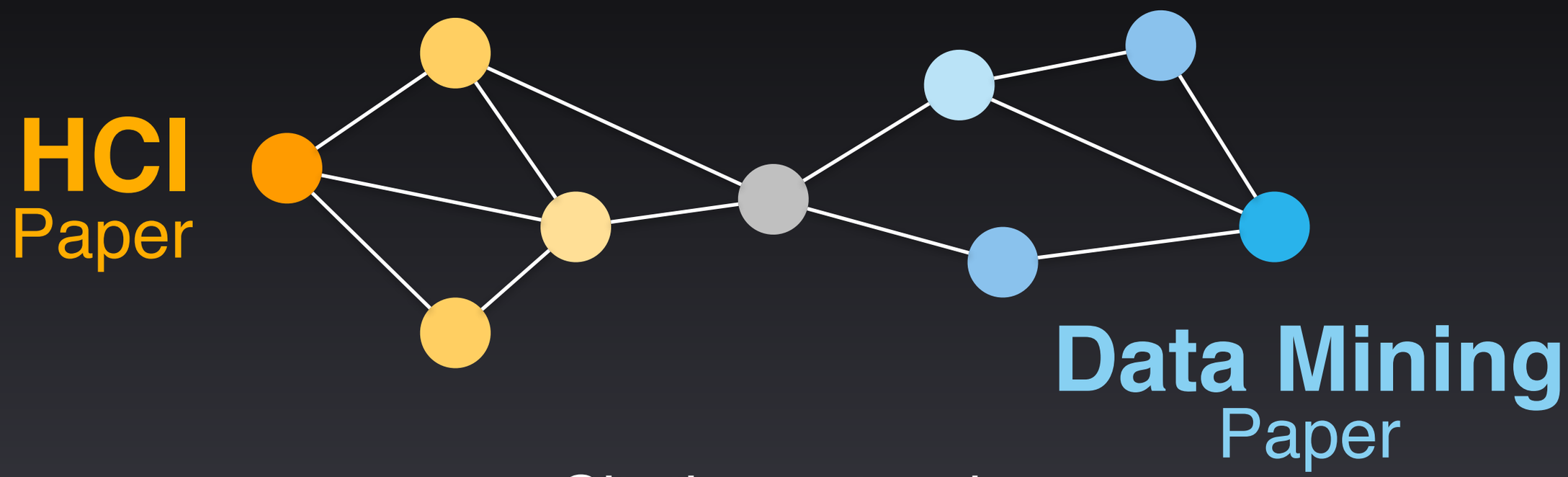


Finding More Relevant Nodes



Citation network

Finding More Relevant Nodes



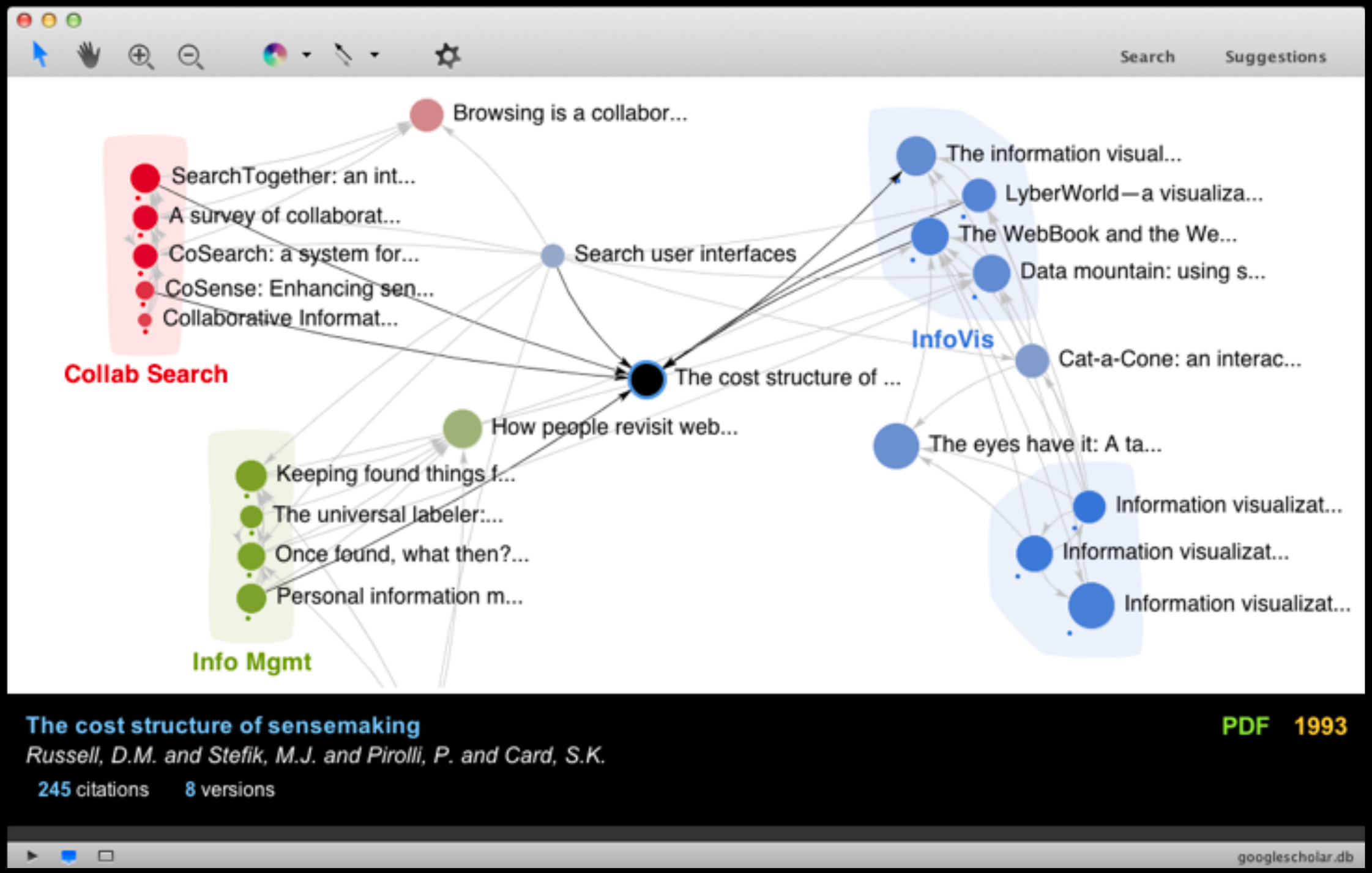
Citation network

Apolo uses **guilt-by-association**

(Belief Propagation, similar to personalized PageRank)

Demo: Mapping the Sensemaking Literature

Nodes: 80k papers from Google Scholar (node size: #citation)
Edges: 150k citations



Search | Suggestions

For **The cost structure of sensemaking**

The information visualizer, an inf...	Card, S.K. and Robertson, G.G. and Macki...	1991 532
The WebBook and the Web Forag...	Card, S.K. and Robertson, G.G. and York, W.	1996 403
LyberWorld—a visualization user...	Hemmje, M. and Kunkel, C. and Willett, A.	1994 223
The structure of the information...	Card, S.K. and Mackinlay, J.	1997 198
Information visualization	Card, S. and Mackinlay, JD and Shneiderm...	2009 180
"I"ll get that off the audio": a cas...	Moran, T.P. and Palen, L. and Harrison, S....	1997 143
An organic user interface for sear...	Mackinlay, J.D. and Rao, R. and Card, S.K.	1995 123
Using a landscape metaphor to re...	Chalmers, M.	1993 122
Personal information management	Jones, W.P. and Teevan, J.	2007 109
SearchTogether: an interface for c...	Morris, M.R. and Horvitz, E.	2007 108
Information foraging theory: Ada...	Pirolli, P.	2007 107
Investigating behavioral variabilit...	White, R.W. and Drucker, S.M.	2007 79
Jigsaw: Supporting investigative...	Stasko, J. and Görg, C. and Liu, Z.	2008 71
The cost-of-knowledge character...	Card, S.K. and Pirolli, P. and Mackinlay, J.D.	1994 54
Collaborative conceptual design:...	Potts, C. and Catledge, L.	1996 45

The cost structure of sen...

The cost structure of sensemaking PDF 1993

Russell, D.M. and Stefik, M.J. and Pirolli, P. and Card, S.K.

245 citations 8 versions

googlescholar.db

Search | Suggestions

For **The cost structure of sensemaking**

The information visualizer, an inf...	Card, S.K. and Robertson, G.G. and Macki...	1991 532
The WebBook and the Web Forag...	Card, S.K. and Robertson, G.G. and York, W.	1996 403
LyberWorld—a visualization user...	Hemmje, M. and Kunkel, C. and Willett, A.	1994 223
The structure of the information...	Card, S.K. and Mackinlay, J.	1997 198
Information visualization	Card, S. and Mackinlay, JD and Shneiderm...	2009 180
"I"ll get that off the audio": a cas...	Moran, T.P. and Palen, L. and Harrison, S....	1997 143
An organic user interface for sear...	Mackinlay, J.D. and Rao, R. and Card, S.K.	1995 123
Using a landscape metaphor to re...	Chalmers, M.	1993 122
Personal information management	Jones, W.P. and Teevan, J.	2007 109
SearchTogether: an interface for c...	Morris, M.R. and Horvitz, E.	2007 108
Information foraging theory: Ada...	Pirolli, P.	2007 107
Investigating behavioral variabilit...	White, R.W. and Drucker, S.M.	2007 79
Jigsaw: Supporting investigative...	Stasko, J. and Görg, C. and Liu, Z.	2008 71
The cost-of-knowledge character...	Card, S.K. and Pirolli, P. and Mackinlay, J.D.	1994 54
Collaborative conceptual design:...	Potts, C. and Catledge, L.	1996 45

The cost structure of sen...

The cost structure of sensemaking PDF 1993

Russell, D.M. and Stefik, M.J. and Pirolli, P. and Card, S.K.

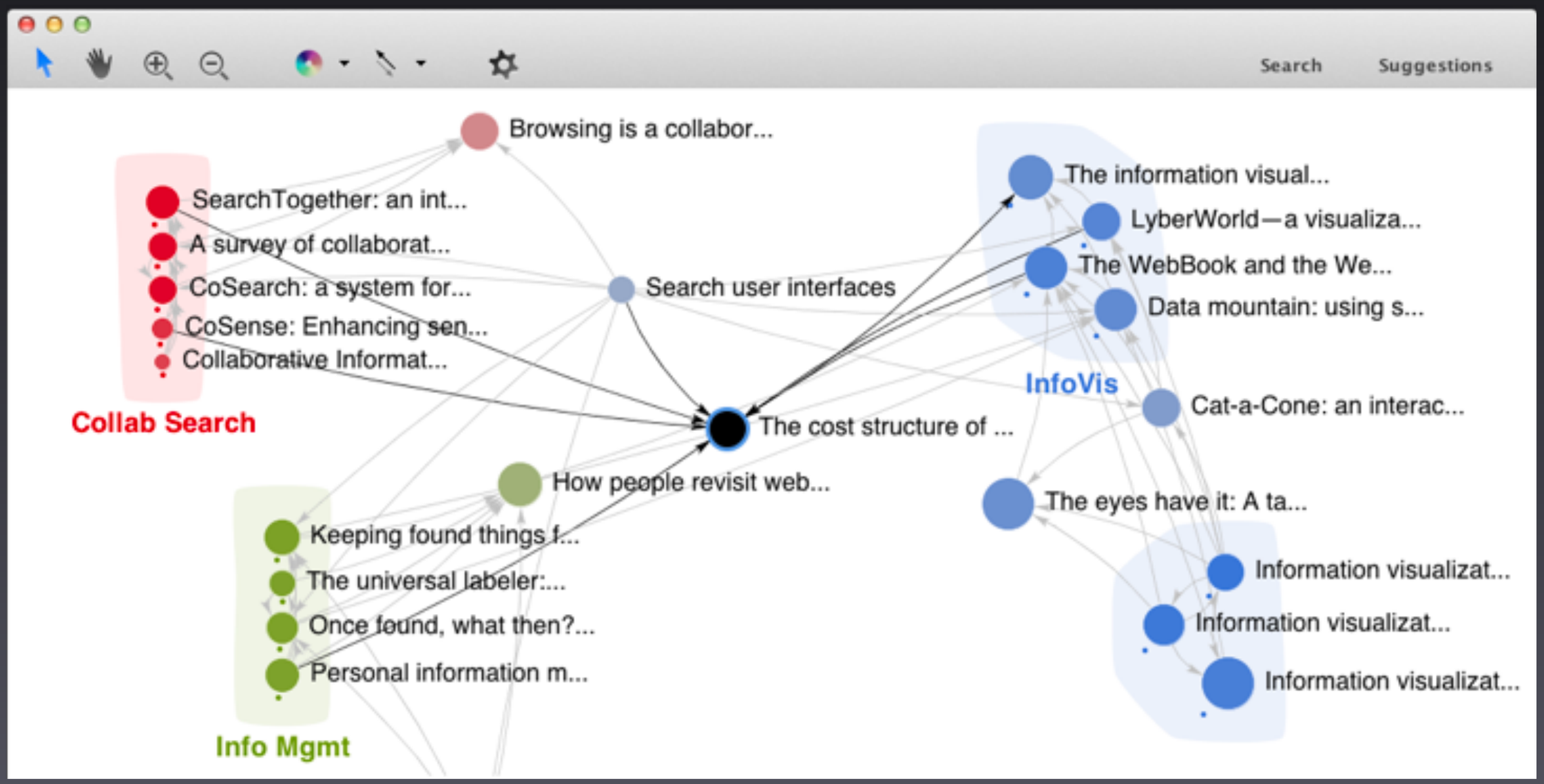
245 citations 8 versions

googlescholar.db

Key Ideas (Recap)

Specify exemplars

Find other relevant nodes (BP)



Apolo's Contributions

1 Human + Machine

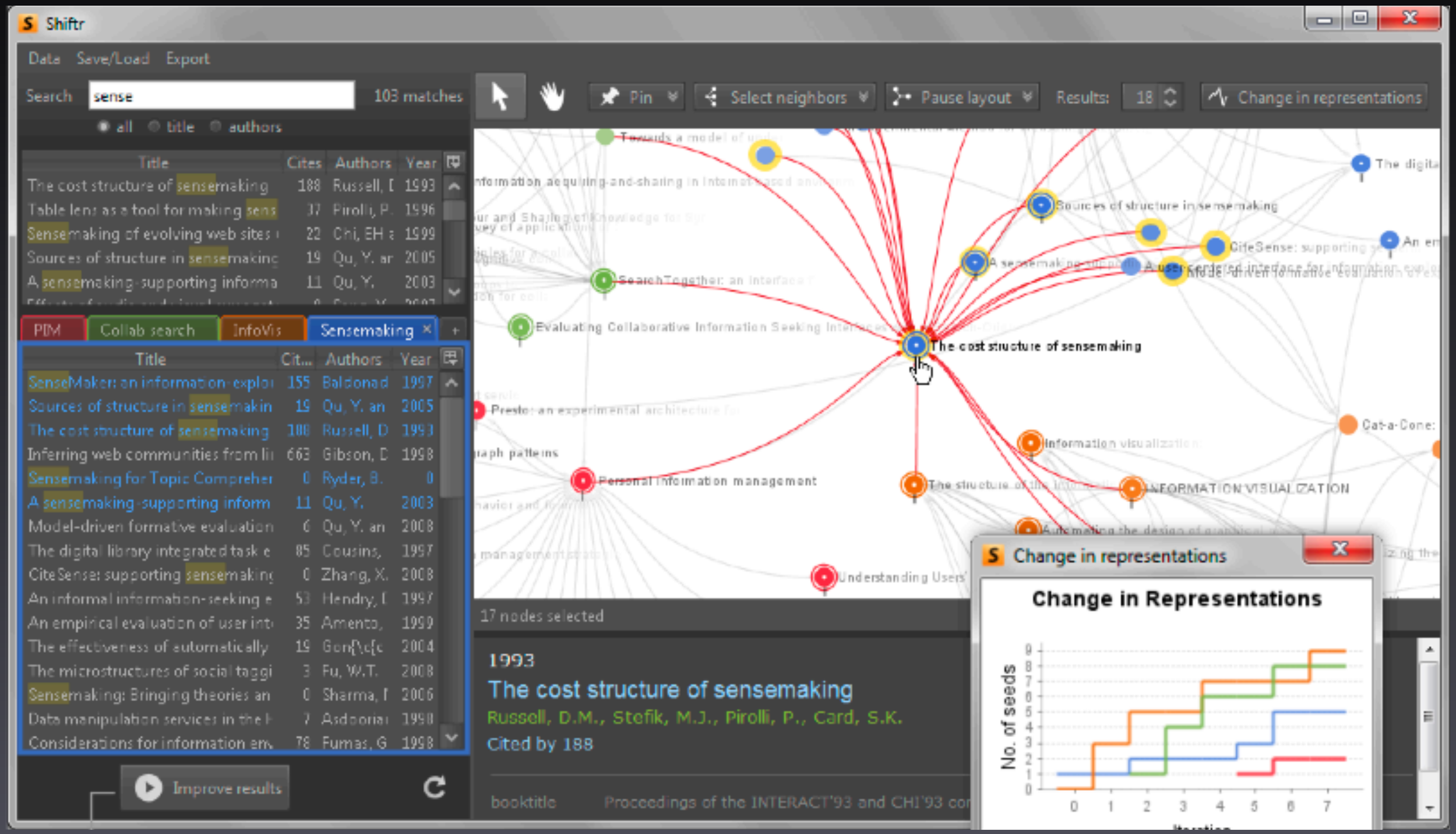
It was like having a **partnership** with the machine.

2 Personalized Landscape

Apolo 2009

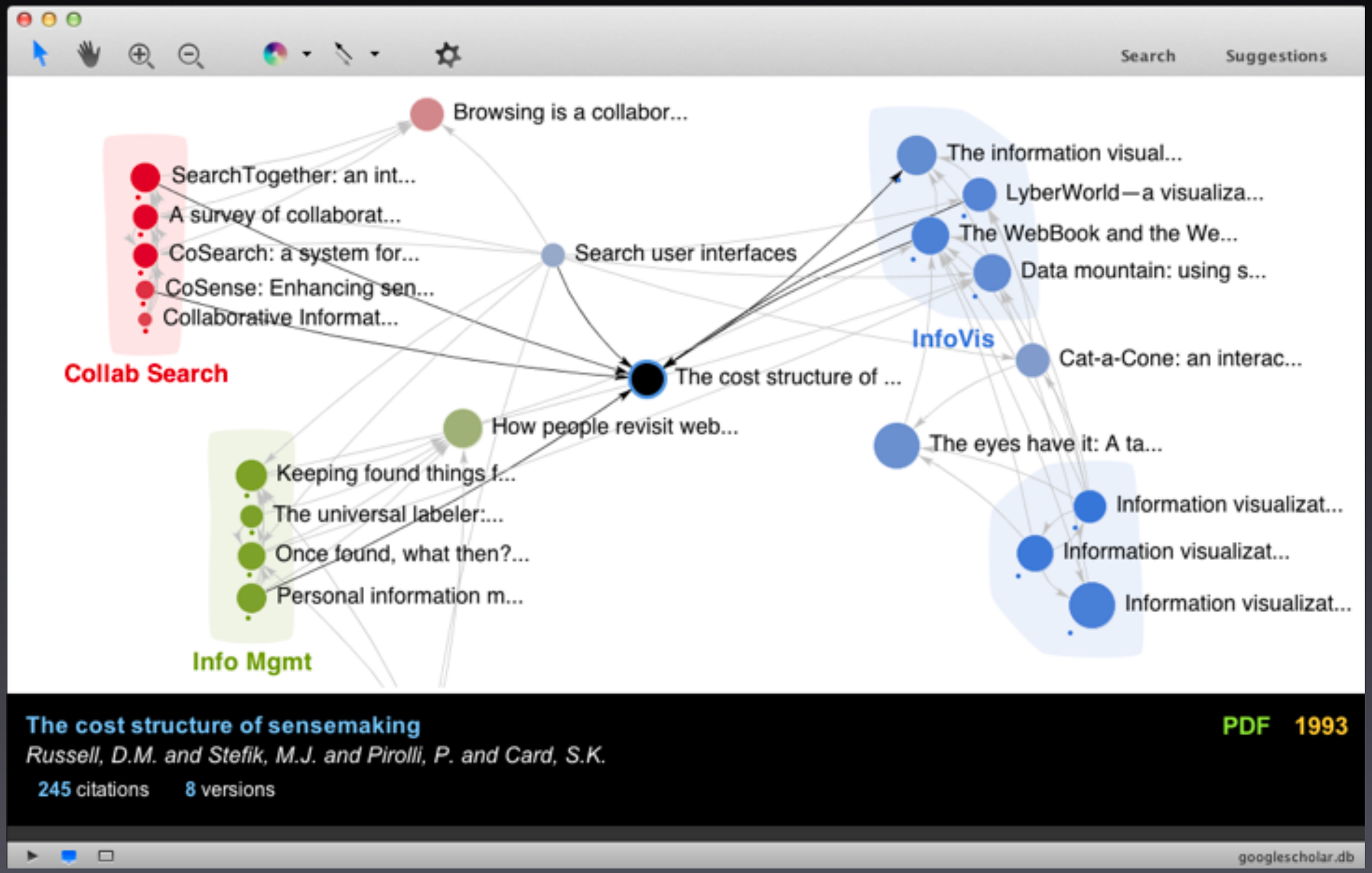

Cluster Data
Add Group
Recommendations:
Not Interested
Automatically generating user inte...
Decision-Theoretic User Interface ...
Daniel S. Weld
Krzysztof Z. Gajos
Automatically generating
Exploring the design spac
Predictability and accurac
Planning-Based Control o
End User Programming
End users creating effective softw...
End user software engineering: chi...
Invited research overview: end-us...
Brad A. Myers
Margaret M. Burnett
Mary Beth Rosson
Andrew Jensen Ko
Alan F. Blackwell
Show: All
Brad
Brad A. Myers
The garnet user interface developm...
Using HCI Techniques to Design a M...
Creating charts by demonstration.
The Amulet User Interface Developm...
Easily Adding Animations to Interfac...
Simplifying video editng using metad...
SILVER: simplifying video editing wit...
Text Entry
In-stroke word completion.
Integrating isometric joysticks into...
Eyes on the road, hands on the whe...
An alternative to push, press, and t...
Maximizing the guessability of symb...
Few-key text entry revisited: mnem...
Text entry from power wheelchairs: ...
Joystick text entry with date stamp, ...
Show: Papers
Interface Generation
Huddle: automatically generating i...
UNIFORM: automatically generatin...
Demonstrating the viability of auto...
Jeffrey Nichols
Brandon Rothrock
Duen Horng Chau

Apolo 2010


Shiftr
Data Save/Load Export
Search sense
103 matches
all title authors
Pin
Select neighbors
Pause layout
Results: 18
Change in representations
Title Cites Authors Year
The cost structure of sensemaking 188 Russell, D 1993
Table lens as a tool for making sens 37 Pirolli, P. 1996
Sensemaking of evolving web sites 22 Chi, EH 1999
Sources of structure in sensemaking 19 Qu, Y. 2005
A sensemaking-supporting informa 11 Qu, Y. 2003
PIM
Collab search
InfoVis
Sensemaking
Title Cit... Authors Year
SenseMaker: an information-explor 155 Baldonad 1997
Sources of structure in sensemakin 19 Qu, Y. 2005
The cost structure of sensemaking 188 Russell, D 1993
Inferring web communities from li 663 Gibson, D 1998
Sensemaking for Topic Compreher 0 Ryder, B. 0
A sensemaking-supporting inform 11 Qu, Y. 2003
Model-driven formative evaluation 6 Qu, Y. 2008
The digital library integrated task e 85 Cousins, 1997
CiteSense: supporting sensemaking 0 Zhang, X. 2008
An informal information-seeking e 53 Hendry, 1997
An empirical evaluation of user int 35 Amento, 1999
The effectiveness of automatically 19 2004
The microstructures of social taggi 3 Fu, W.T. 2008
Sensemaking: Bringing theories an 0 Sharma, 2006
Data manipulation services in the 7 Asdoorian 1998
Considerations for information en 78 Furnas, G 1998
Improve results
Sources of structure in sensemaking
CiteSense: supporting
SearchTogether: an interface
Evaluating Collaborative Information Seeking
The cost structure of sensemaking
Presto: an experimental architecture
Personal information management
Information visualization
INFORMATION VISUALIZATION
Understanding Users'
Cat-a-Cone:
17 nodes selected
1993
The cost structure of sensemaking
Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K.
Cited by 188
booktitle Proceedings of the INTERACT'93 and CHI'93 con
Change in representations
Change in Representations
No. of seeds
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7

Apolo 2011

22,000 lines of code. Java 1.6. Swing.
Uses SQLite3 to store graph on disk

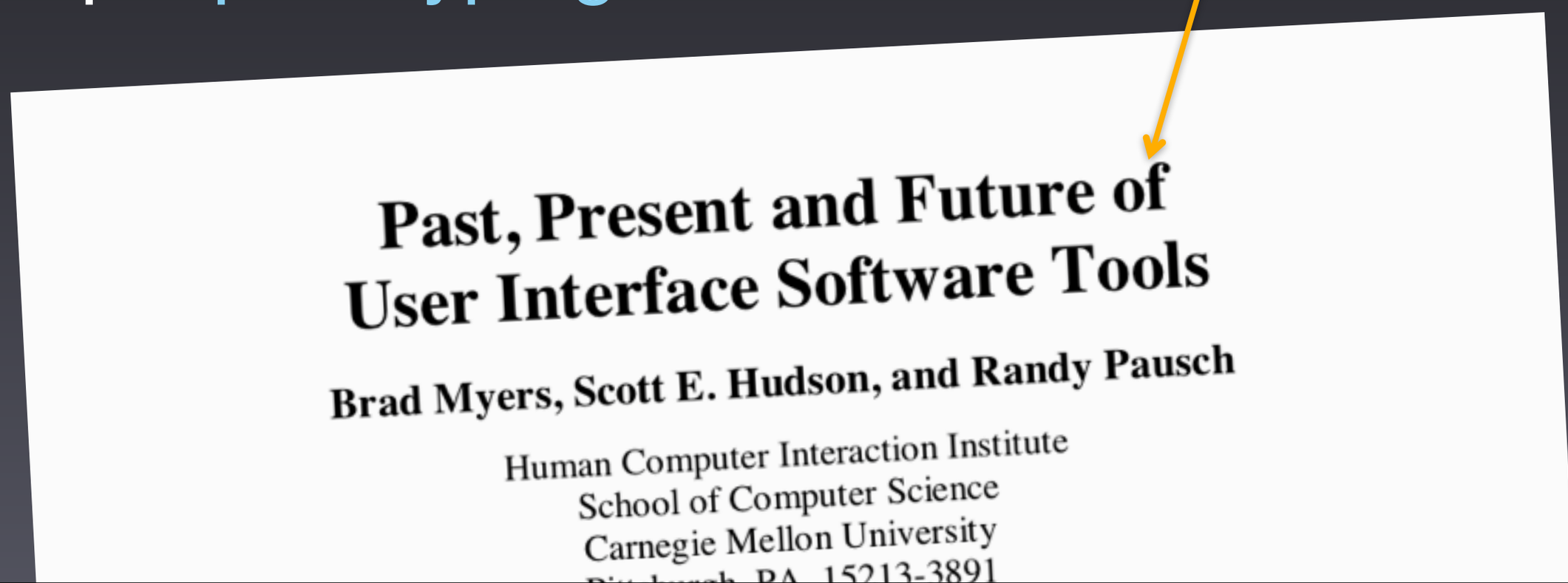


User Study

Used citation network

<u>Task</u>: Find related papers for 2 sections in a survey paper on *user interface*

- Model-based generation of UI
- Rapid prototyping tools



Past, Present and Future of User Interface Software Tools

Brad Myers, Scott E. Hudson, and Randy Pausch

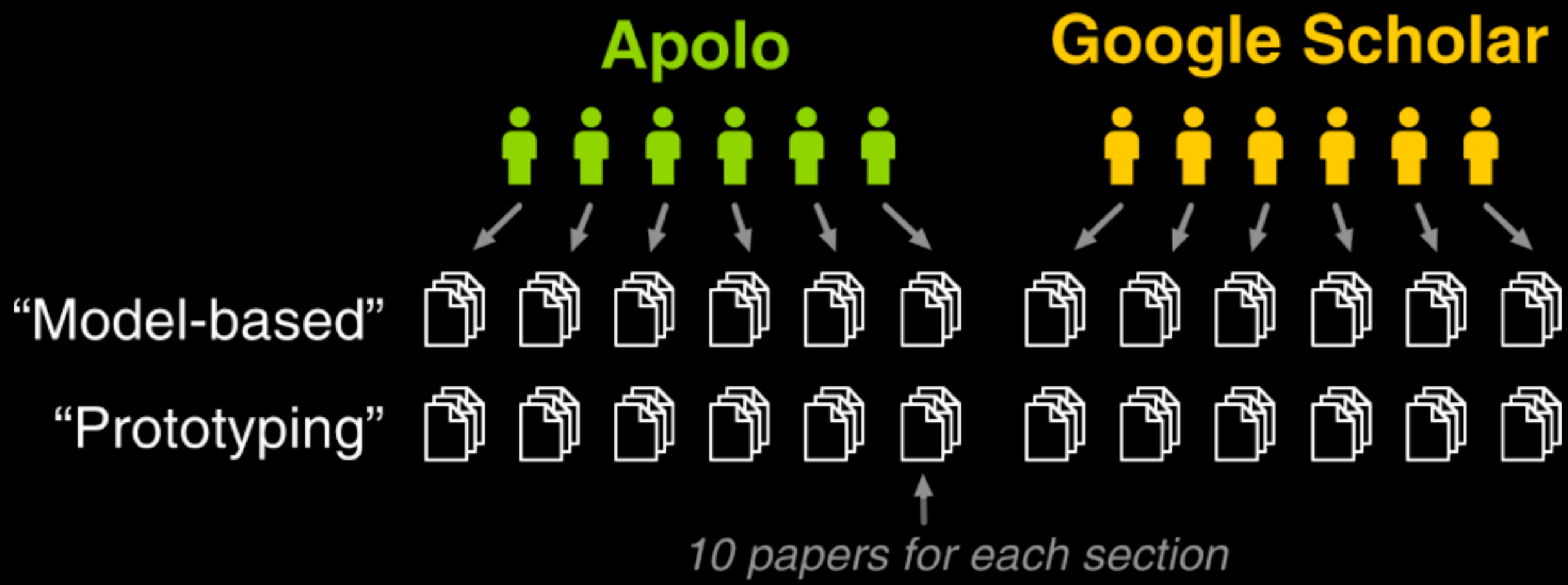
Human Computer Interaction Institute
School of Computer Science
Carnegie Mellon University

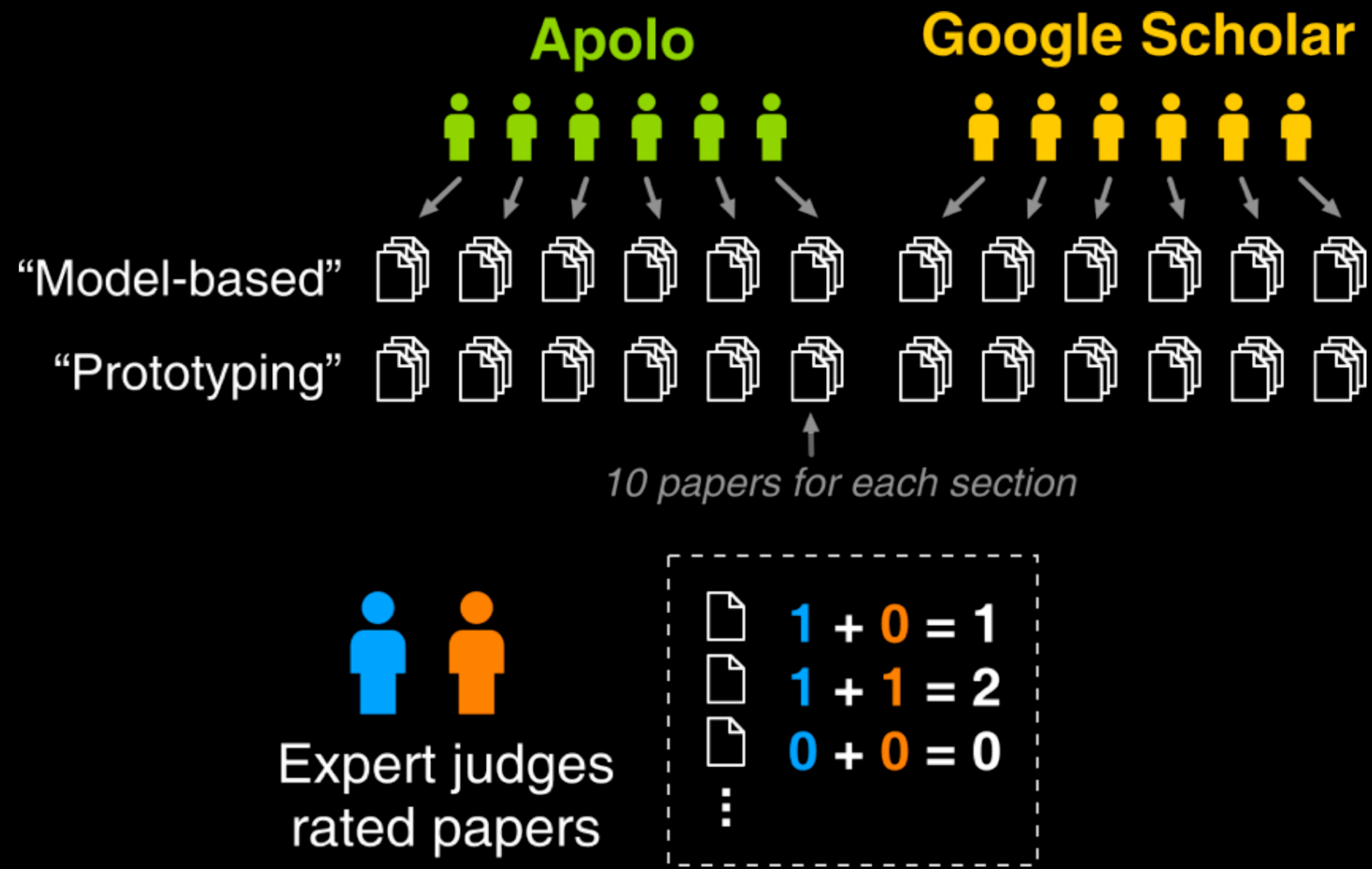
Apolo

Google Scholar

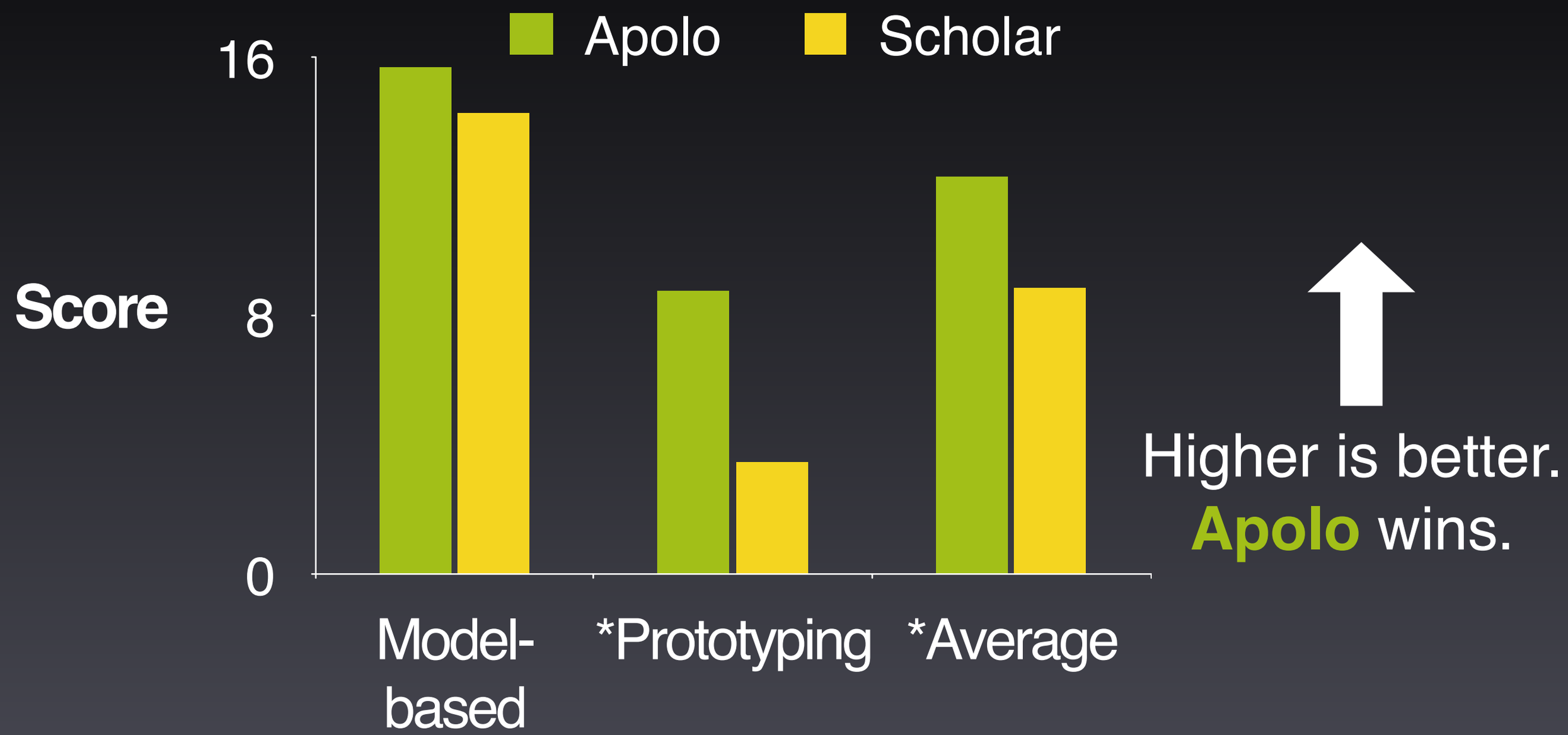
Between subjects design

Participants: **grad student** or **research staff**


Apolo
Google Scholar
"Model-based"
"Prototyping"
10 papers for each section


Apolo
Google Scholar
"Model-based"
"Prototyping"
10 papers for each section
Expert judges rated papers
1 + 0 = 1
1 + 1 = 2
0 + 0 = 0

Judges' Scores

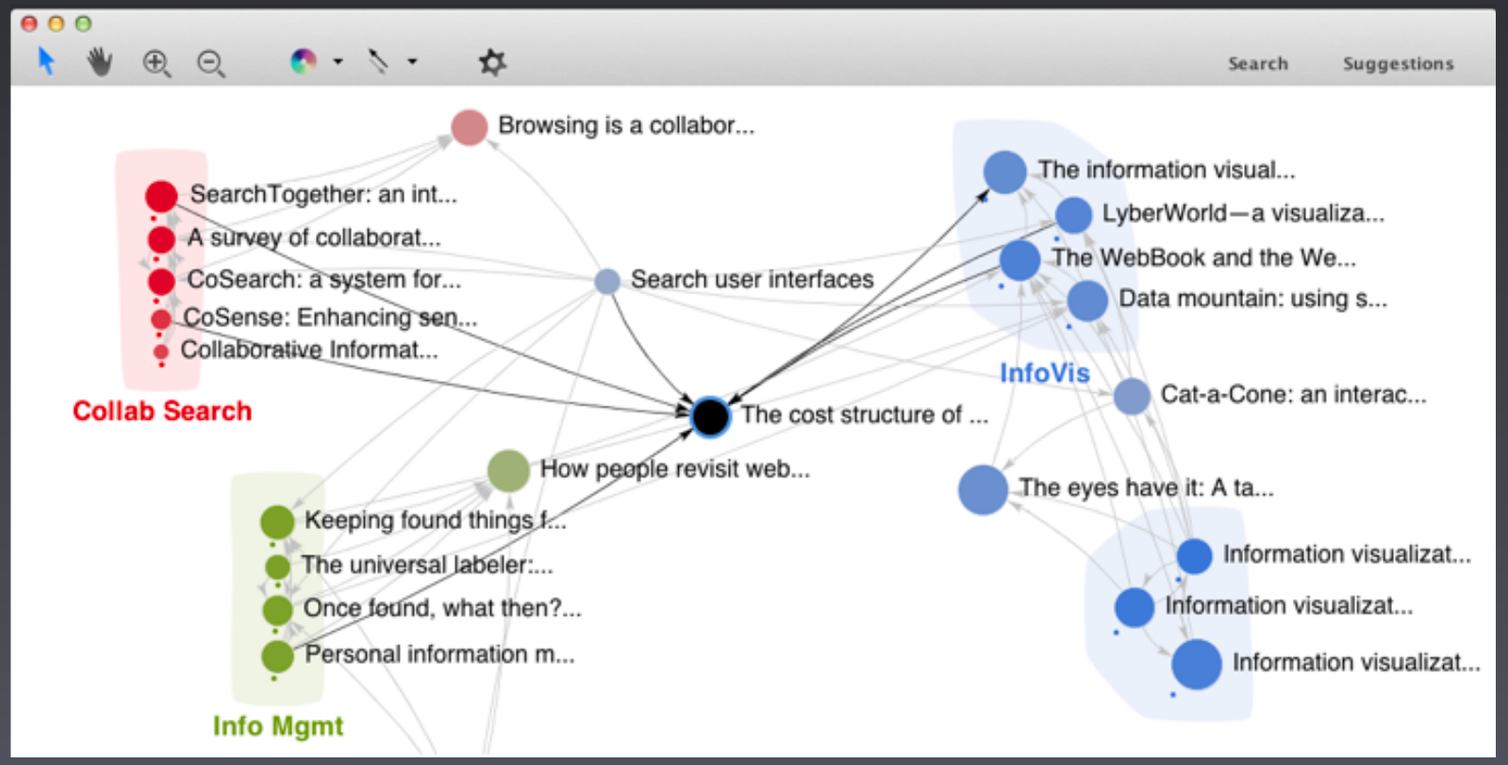


** Statistically significant, by two-tailed t test, $p < 0.05$*

Apolo: Recap

A mixed-initiative approach for exploring and creating personalized landscape for large network data

Apolo = ML + Visualization + Interaction



Practitioners' guide to building (interactive) applications?

Important that you pick a **good** problem!

- Otherwise, you solve a "non-problem", and nobody cares

Think about scalability early

- e.g., picking a scalable algorithm early on

When building interactive applications, use **iterative** design approach (as in Apolo)

- Why? It's hard to get it right the first time
- Create prototype, evaluate, modify prototype, evaluate, ...
- Quick evaluation helps you identify **important fixes early** (can **save you a lot of time**)

Practitioners' guide to building (interactive) applications?

How to do **iterative** design?

What kinds of **prototypes**?

- Paper prototype, lo-fi prototype, high-fi prototype

What kinds of **evaluation**?

- Recruit your friends to try your tools
- Lab study (controlled, as in Apolo)
- Longitudinal study (usage over months)
- Deploy it and see the world's reaction!

- To learn more:
  - CS 6750 Human-Computer Interaction
  - CS 6455 User Interface Design and Evaluation