Data CrowdSourcing: Is It For Real?

Hector Garcia-Molina

(work with Steven Whang, Peter Lofgren, Aditya Parameswaran, Hyunjung Park, Vasilis Verroios, Manas Joglekar, Ming Han Teh and others)

Stanford University

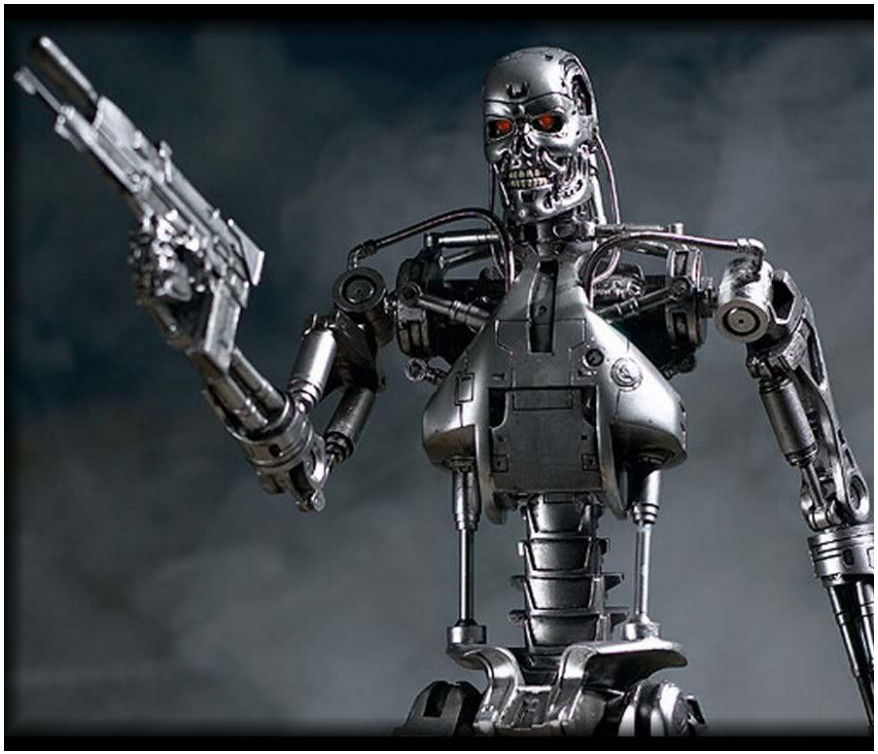
CrowdSourcing

"performing a task using
human workers that solve sub-problems"

Man/Woman vs. Machine

John Connor

>



SkyNet Terminator

CrowdSourcing

CrowdSourcing

Getting Started | Latest Headlines | Imported From Fi...

Google

john hennessy

Web | Shopping | Images | Videos | News | More | Search tools

About 3,600,000 results (0.33 seconds)

John L. Hennessy - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/John_L._Hennessy ▾ Wikipedia ▾
John LeRoy **Hennessy** (born September 22, 1952) is an American computer scientist, academician, and businessman. **Hennessy** is one of the founders of MIPS ...
Early life and career - Research - Noted publications - References

Biography: Office of the President: Stanford University
president.stanford.edu/biography/ ▾
Stanford President: **John Hennessy** John L. Hennessy joined Stanford's faculty in 1977 as an assistant professor of electrical engineering. He rose through the ...

John Hennessy - Stanford University
web.stanford.edu/~hennessy/ ▾ Stanford University ▾
May 8, 2014 - **John Hennessy** photo 001.jpg. John L. Hennessy ... Professor Hennessy initiated the MIPS project at Stanford in 1981, MIPS is a high- ...

Office of the President: Stanford University
president.stanford.edu/ ▾
John Hennessy. Stanford President. Over the years, the duties of a university president have expanded dramatically. Stanford's president remains focused on ...

Stanford's John Hennessy: MOOCs are failing students ...
www.bizjournals.com/.../stanford-head-m... ▾ South Florida Business Journal ▾
Feb 3, 2014 - Stanford president: MOOCs should not be so open, massive. Stanford President **John Hennessy** says online MOOC courses are growing too ...

John L. Hennessy: Risk Taker - IEEE Spectrum
spectrum.ieee.org/geek-life/.../john-l-hennessy-risk-taker ▾ IEEE Spectrum ▾
Apr 24, 2012 - In the 1980s, **John** L. **Hennessy**, then a professor of electrical engineering at Stanford University, shook up the computer industry by taking the ...

An Open Letter to John Hennessy, President of Stanford ...
www.mcsweeneys.net/.../an-open-letter-to-john-hen... ▾ McSweeney's Books ▾
Dear President **Hennessy**,. Forgive me! You have sent me so many thoughtful letters over the years, and I have never written back, never answered your one big ...

Amazon.com: John L. Hennessy: Books, Biography, Blog ...
www.amazon.com/John-L.-Hennessy/e/B000APA2GC ▾ Amazon.com ▾
7 Results - Computer Architecture, Fifth Edition: A Quantitative Approach (The Morgan Kaufmann Series in Computer Architecture... by **John** L. **Hennessy** and ...

John L. Hennessy

John LeRoy Hennessy is an American computer scientist, academician, and businessman. Hennessy is one of the founders of MIPS Computer Systems Inc. as well as Atheros and is the 10th President of Stanford University. Wikipedia

Born: 1953

Books: Computer Architecture: A Quantitative Approach, More

Education: State University of New York at Stony Brook, Villanova University

Awards: IEEE Medal of Honor, IEEE John von Neumann Medal, More

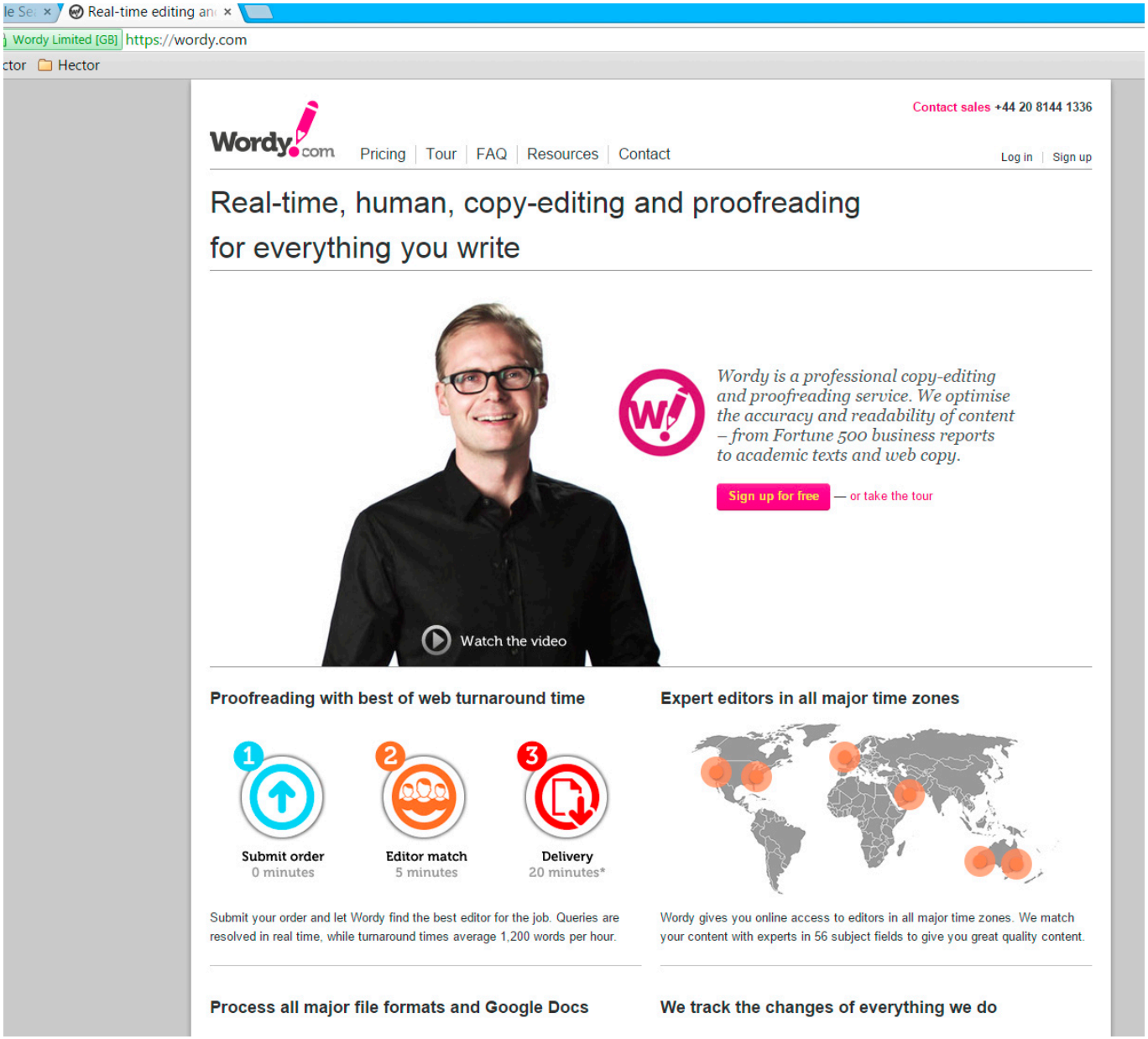
People also search for — View 15+ more

David Patterson | John Etcheme... | William Stallings | Andrew S. Tanenbaum | Jeffrey Ullman

Feedback

crowd powered

Crowd Sourcing



Crowd Funding

Crowd Funding

More CrowdSourcing Examples

Categorizing Images

Search Relevance

bing

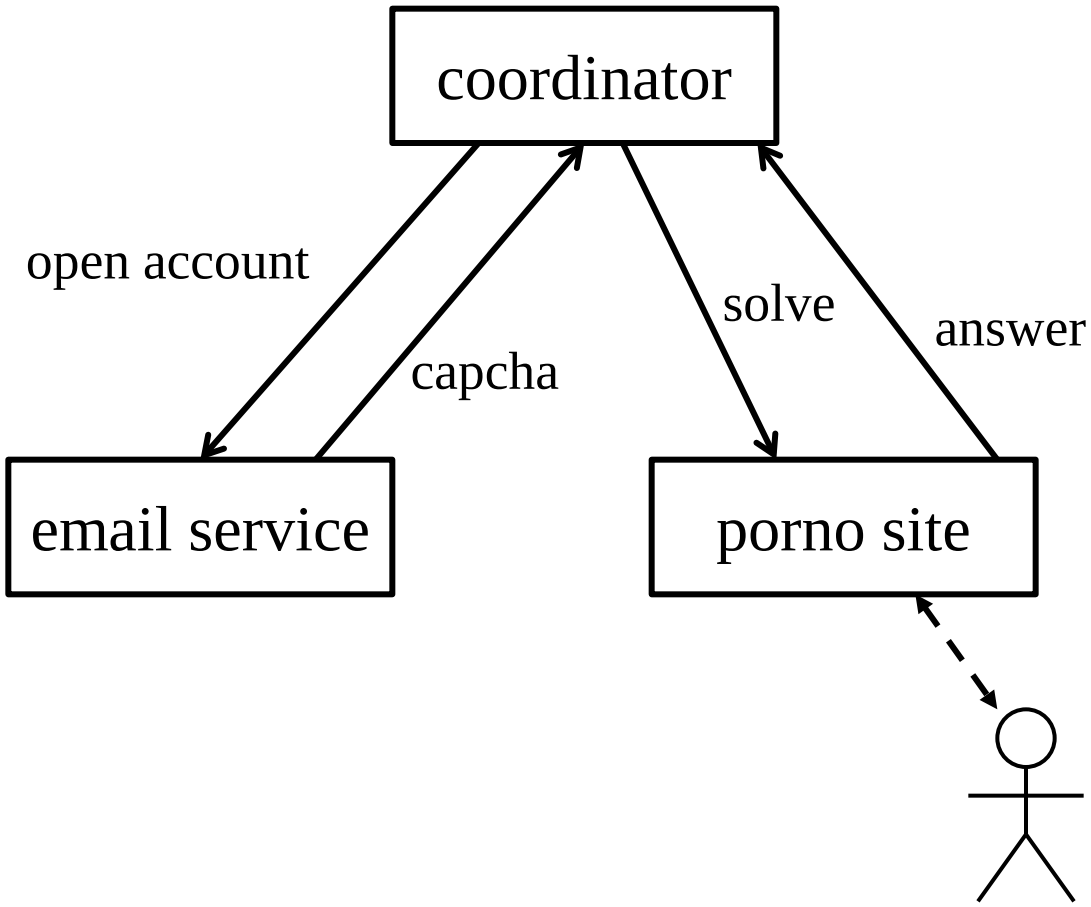
Data Gathering

CrowdFlower

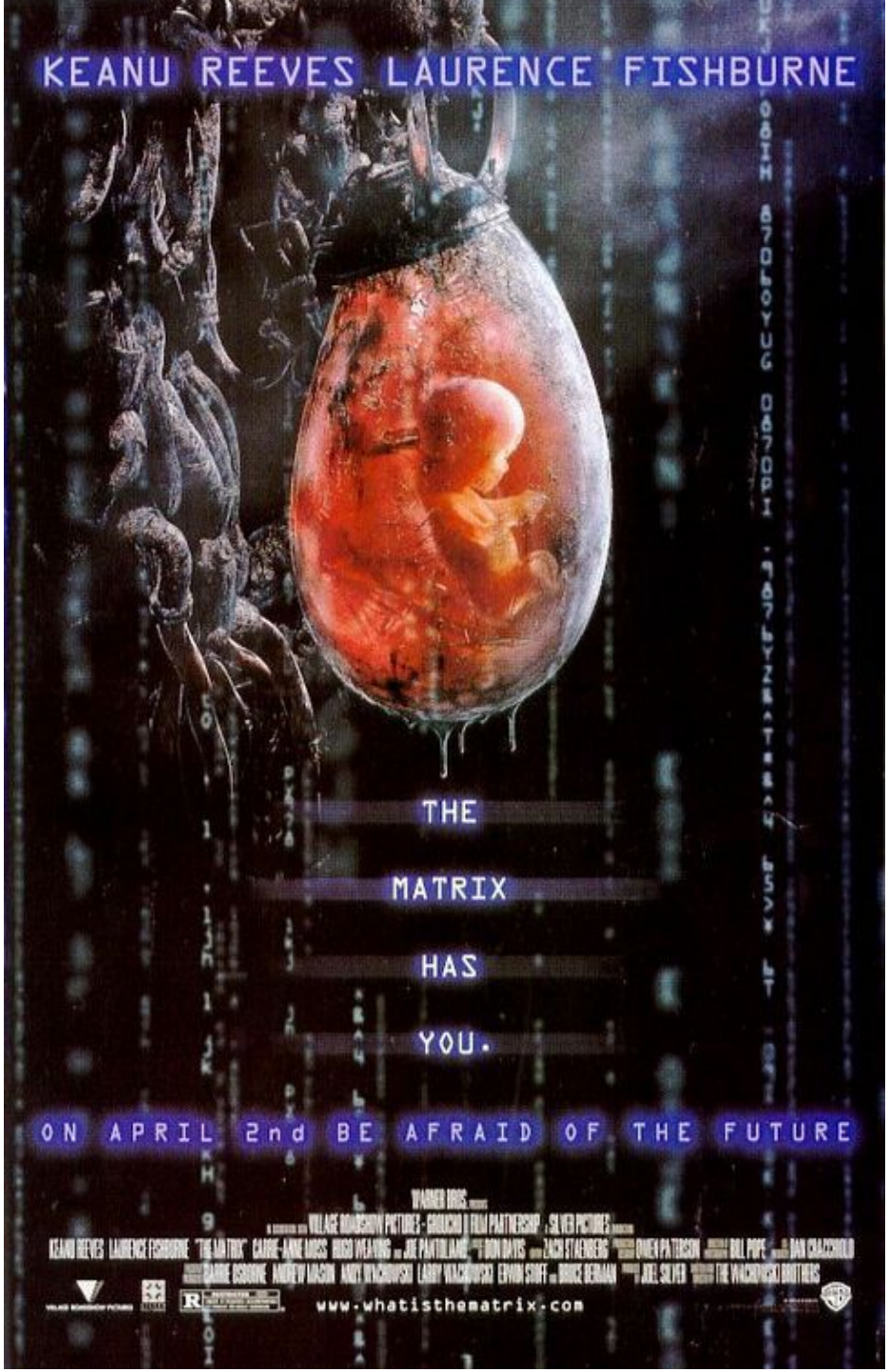
Image Matching
Translation

CrowdSourcing: Final Examples

CrowdSourcing: Spammers & Porno

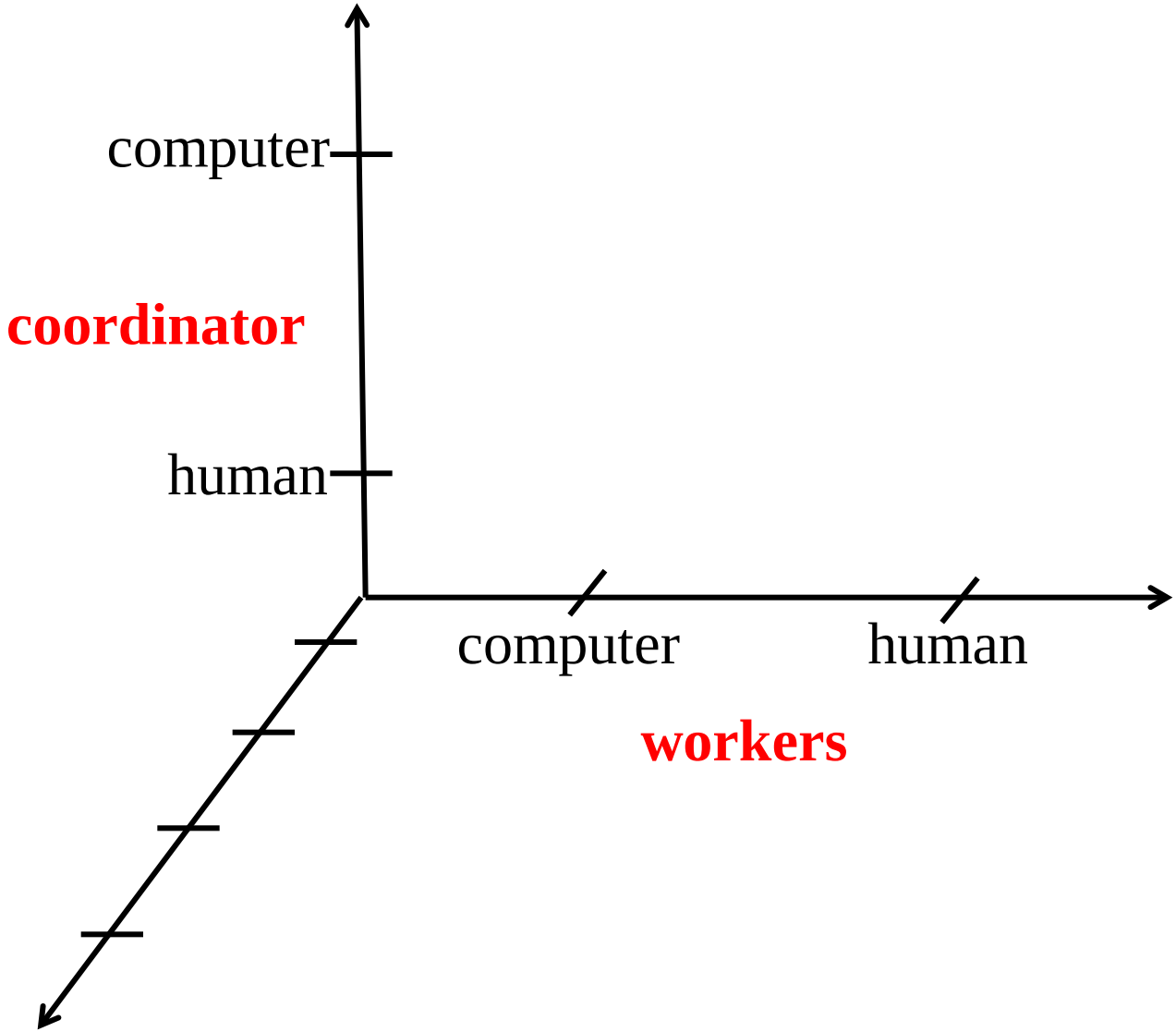


CrowdSourcing

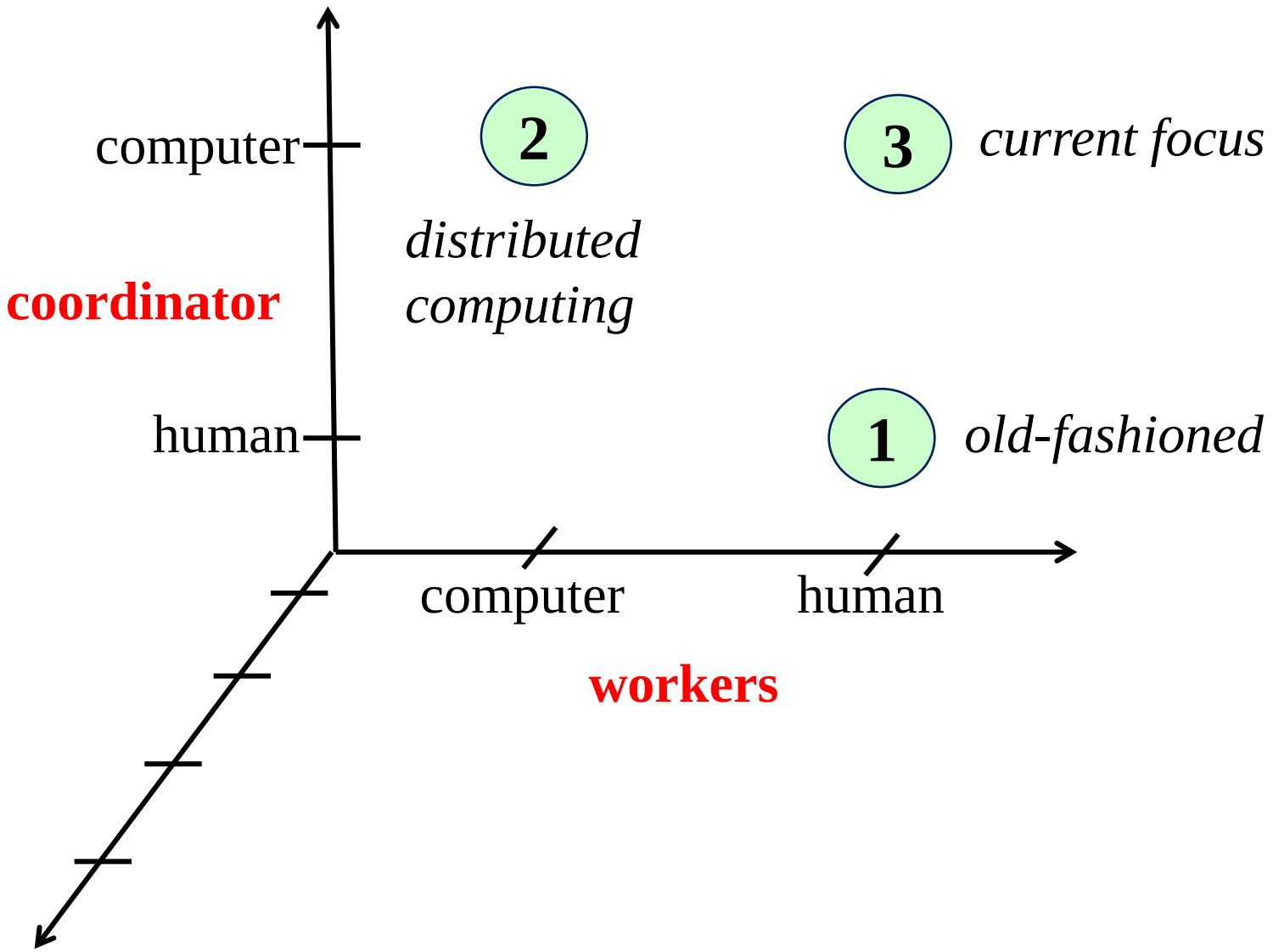


CrowdSourcing Space

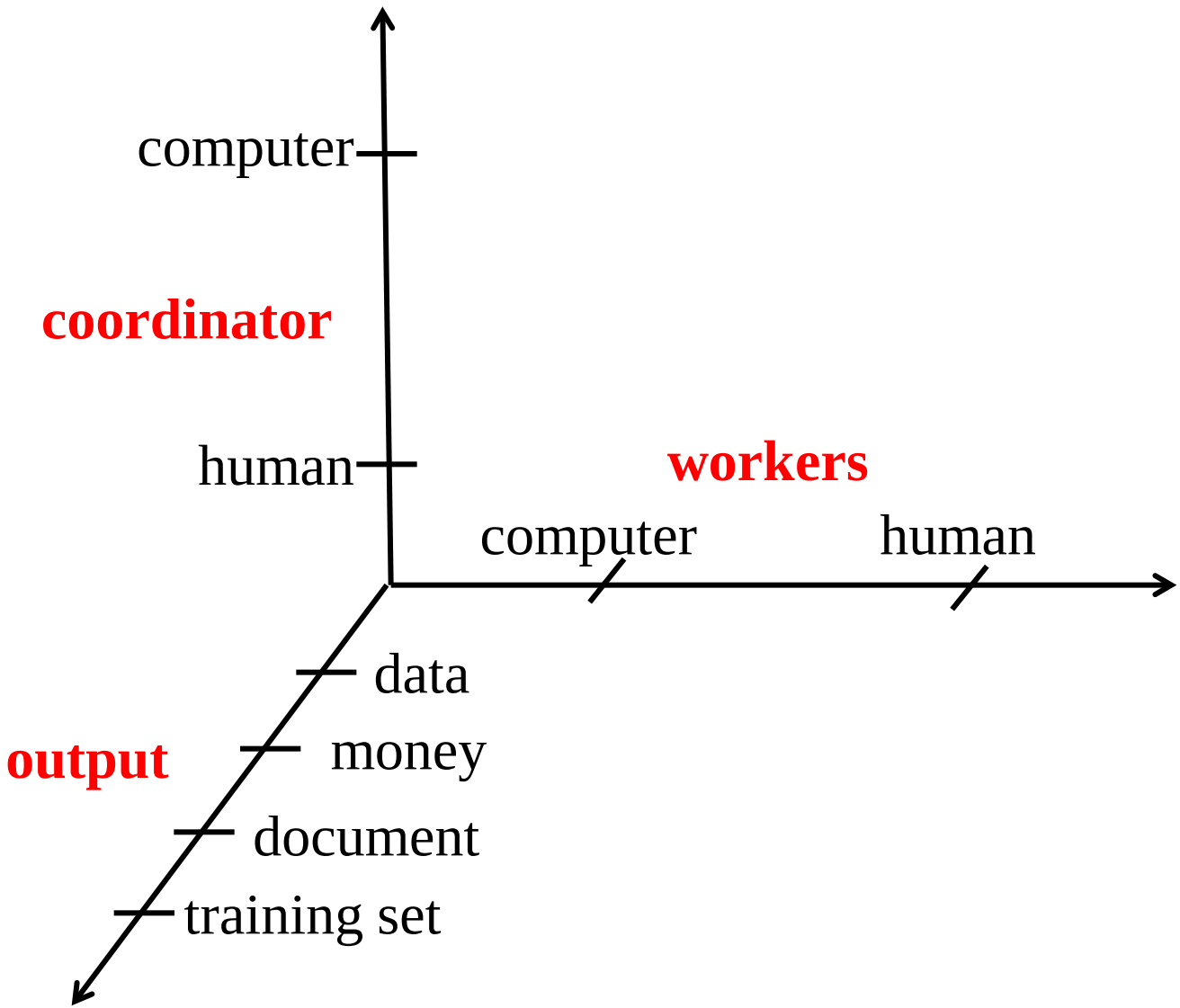
CrowdSourcing Space



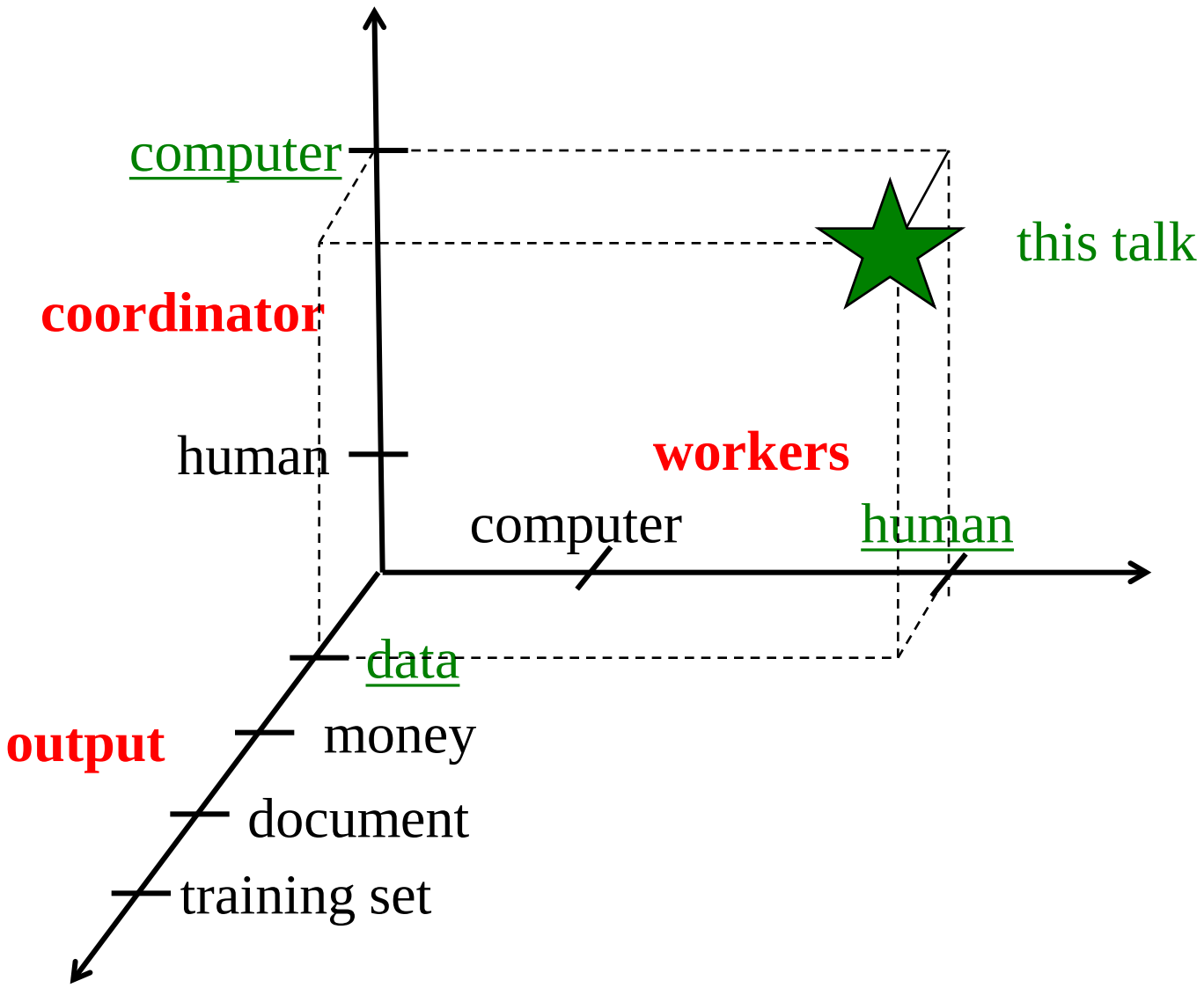
CrowdSourcing Space



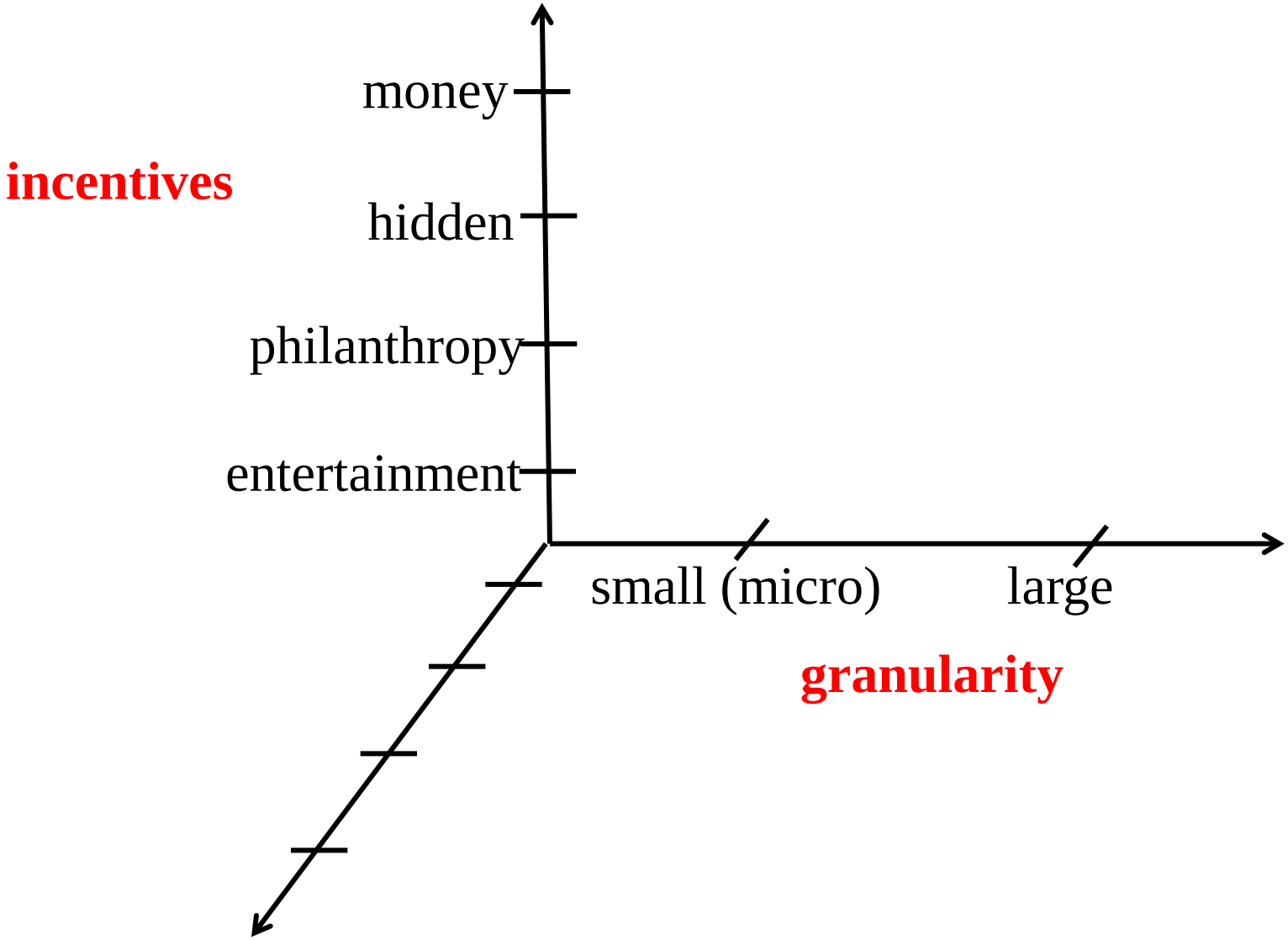
CrowdSourcing Space



CrowdSourcing Space



Two More Dimensions



So, Is CrowdSourcing for Real??

- Is it used in practice?
- Are there interesting research problems?

Many Crowdsourcing Marketplaces!

Many Research Projects!

UTest · Elance · CloudCrowd · GRABCAD · liveops · oDesk · mygengo · freelancer

amazonmechanical turk · CrowdFlower · agentanything · crowdcontent · socialvibe · DeskLancer

servio · CTR Boost · DESIGN RIDER · CAPTEO · samasource · SimpleWorkers · SpeakerText · Truveli

trada · guru · microtask · taskus · DoNanza · Lingotek · expertplanet · MOB4HIRE

peopleperhour.com · Spool Source · clickworker · PEER TO PATENT · tagasauris the smart way to tag! · mediapiston

University of Washington · Carnegie Mellon University · Cornell University · NYU · Northwestern University · Illinois · University of California Santa Cruz · Stanford · Georgia Institute of Technology · Cal

Harvard · ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich · Universität Bremen · University of California San Diego · Microsoft Research · Yahoo! Research · MIT Massachusetts Institute of Technology · Syracuse University · The University of Texas at Austin · The University of Michigan Flint · University of California Irvine

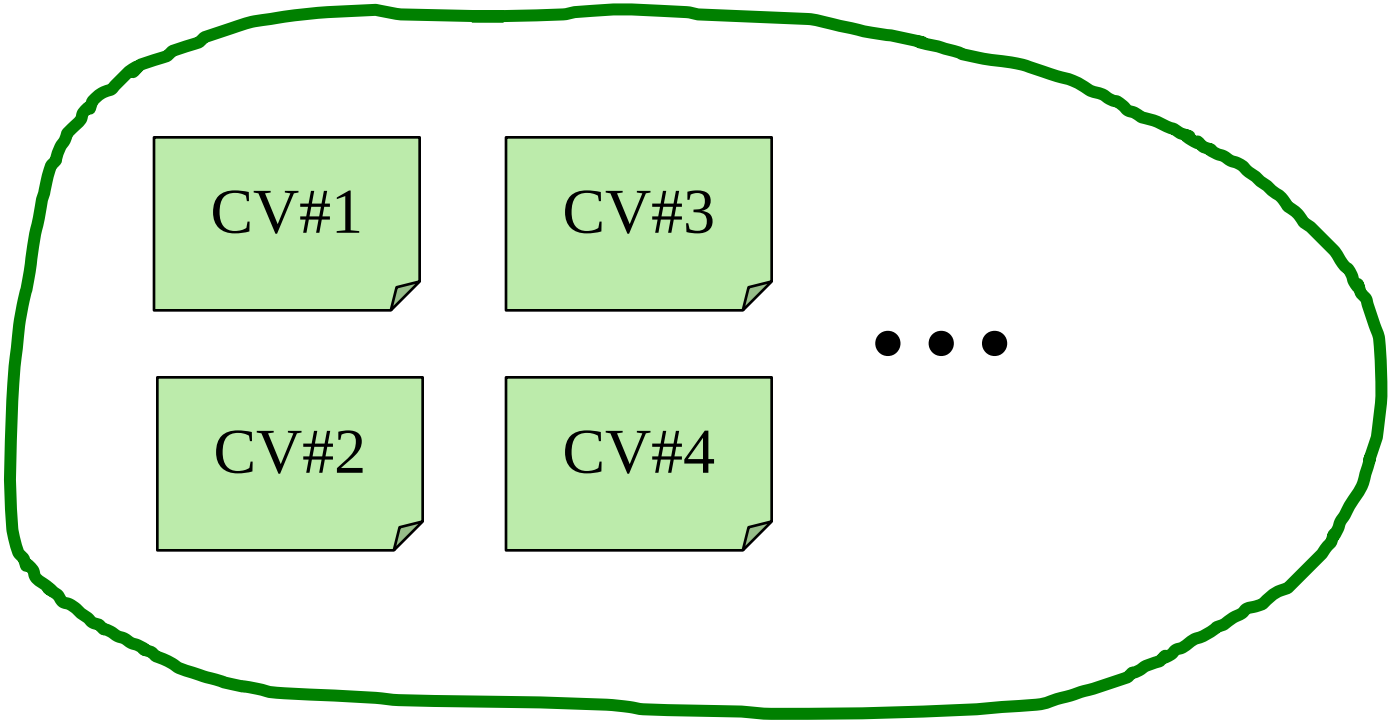
Overview: Crowd Data Management

- Data Processing
- Data Gathering
- Searching

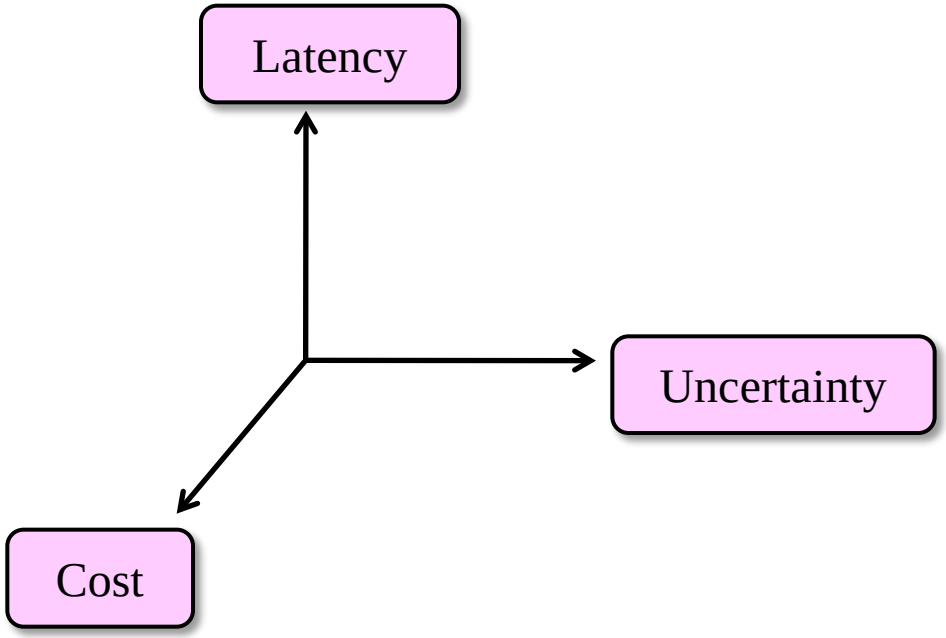
Finding the Maximum

job description

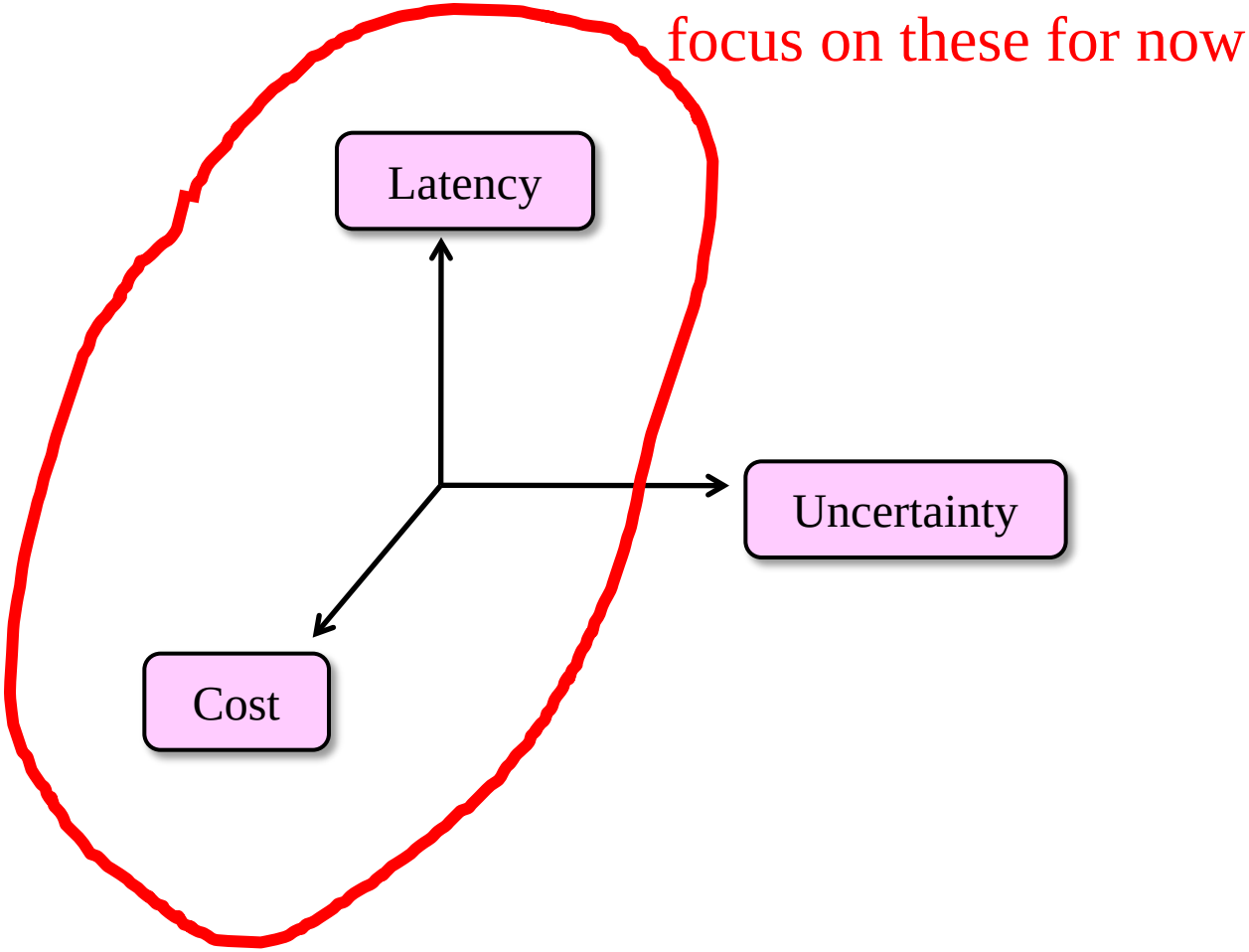
What is the best applicant for the job?



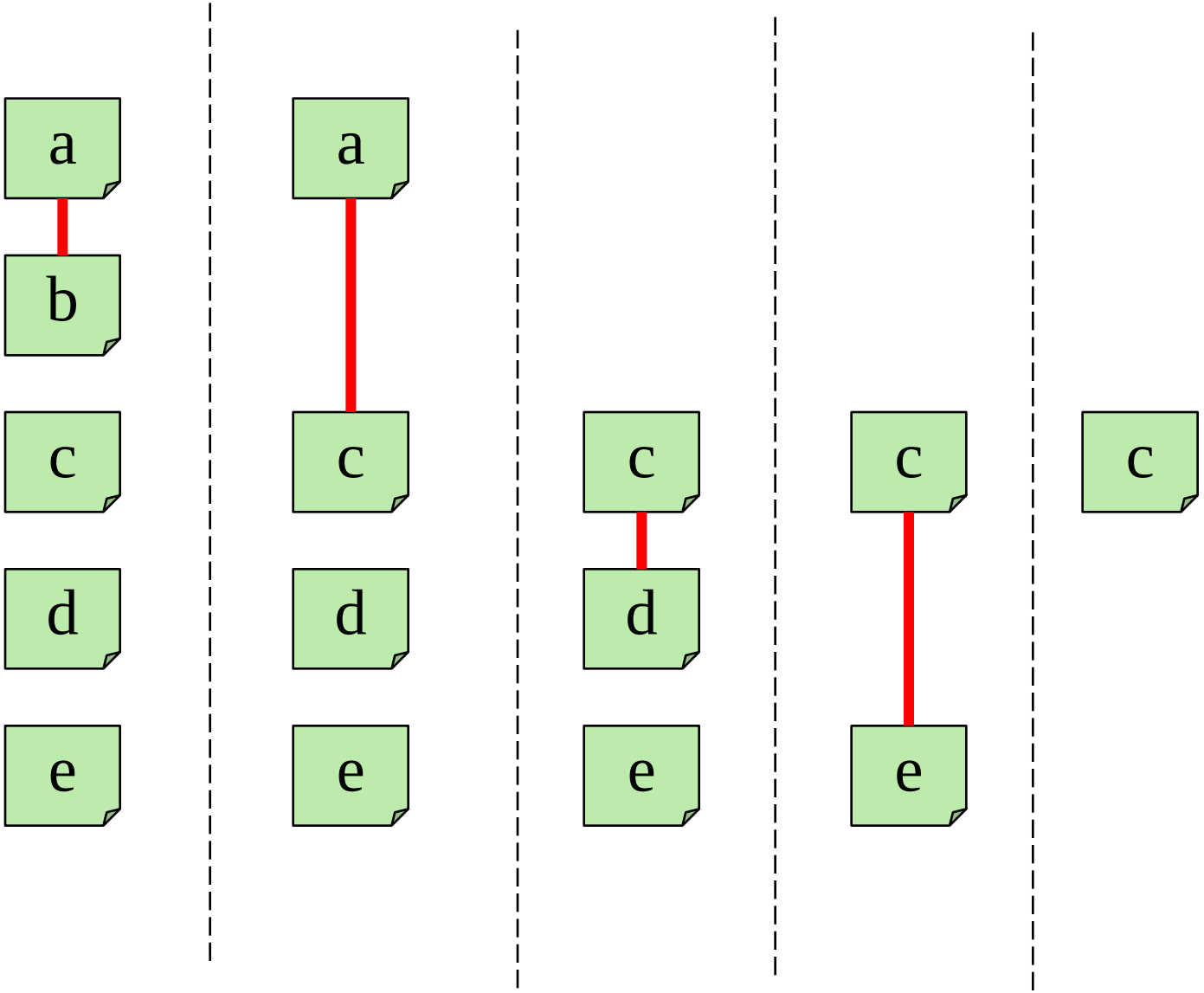
Fundamental Tradeoffs



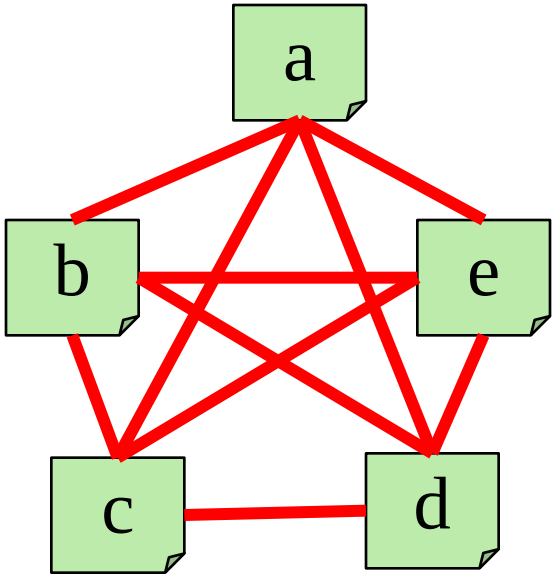
Fundamental Tradeoffs

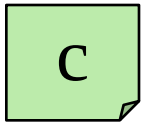


Example Max Algorithm FS

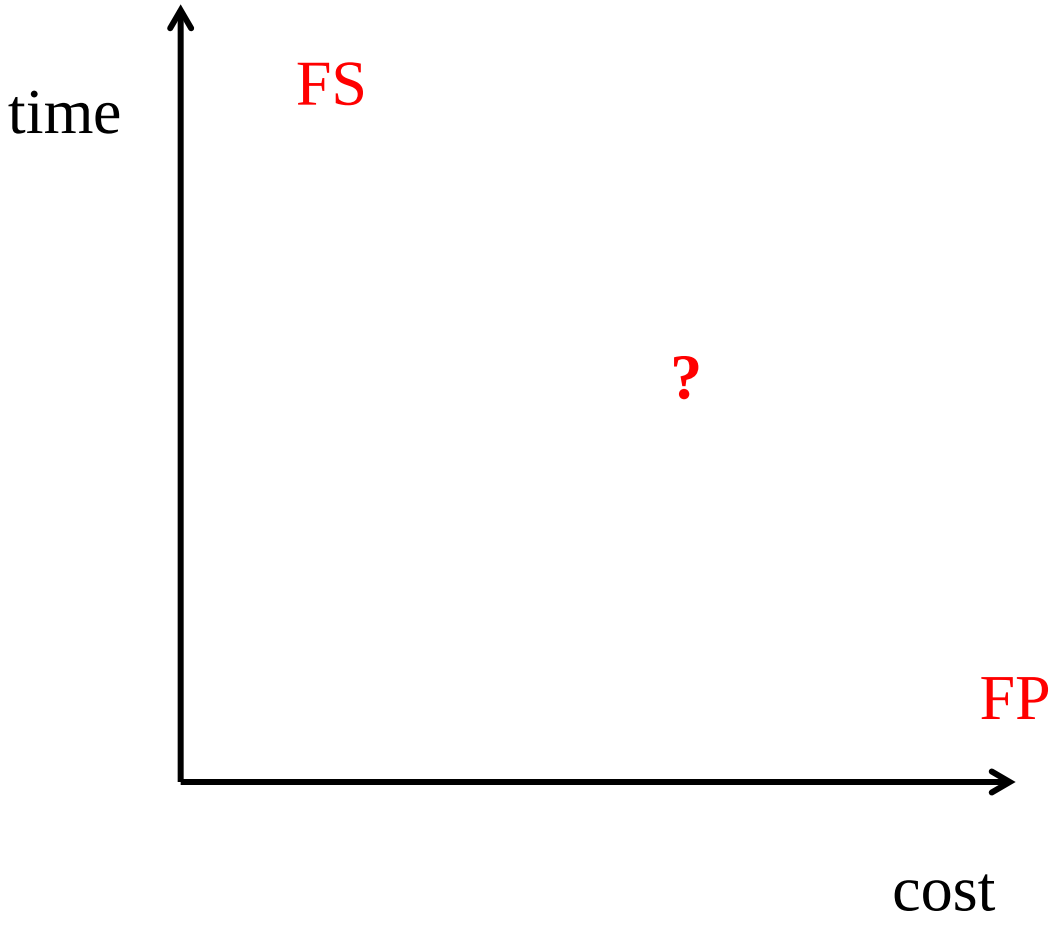


Example Max Algorithm FP

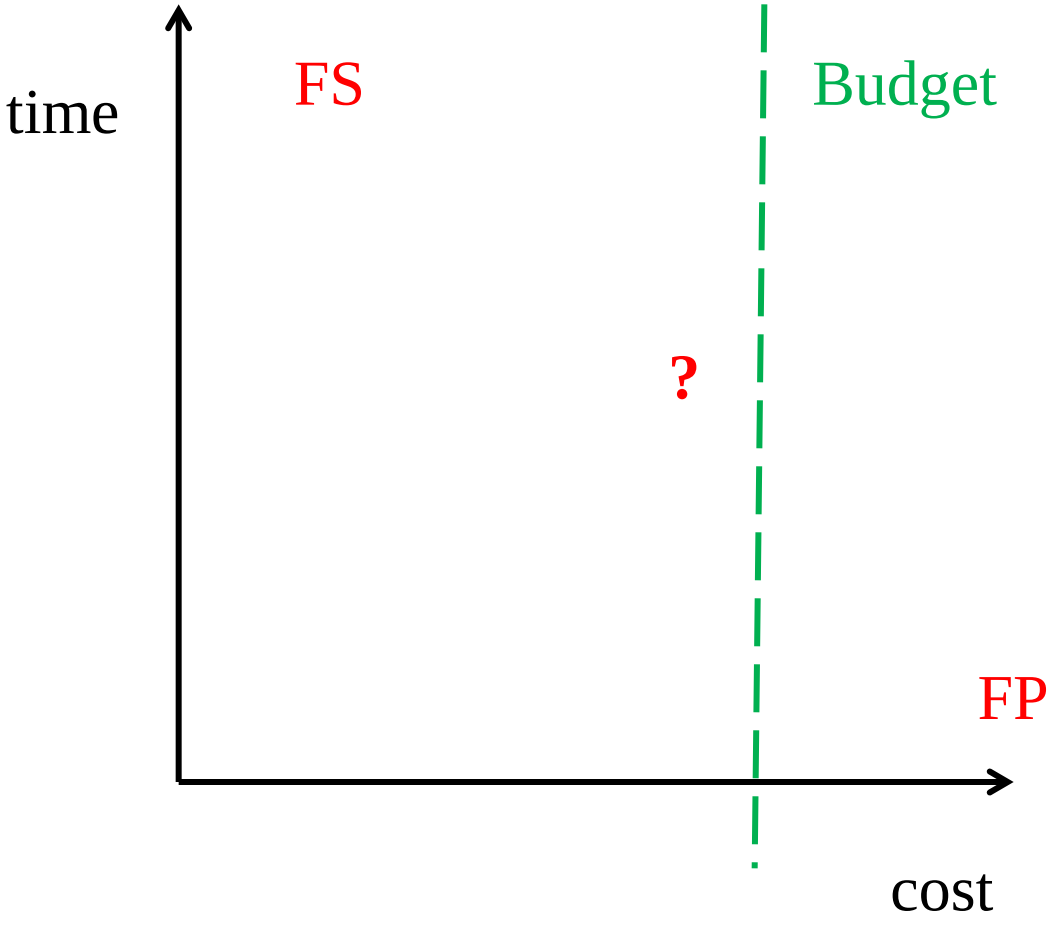




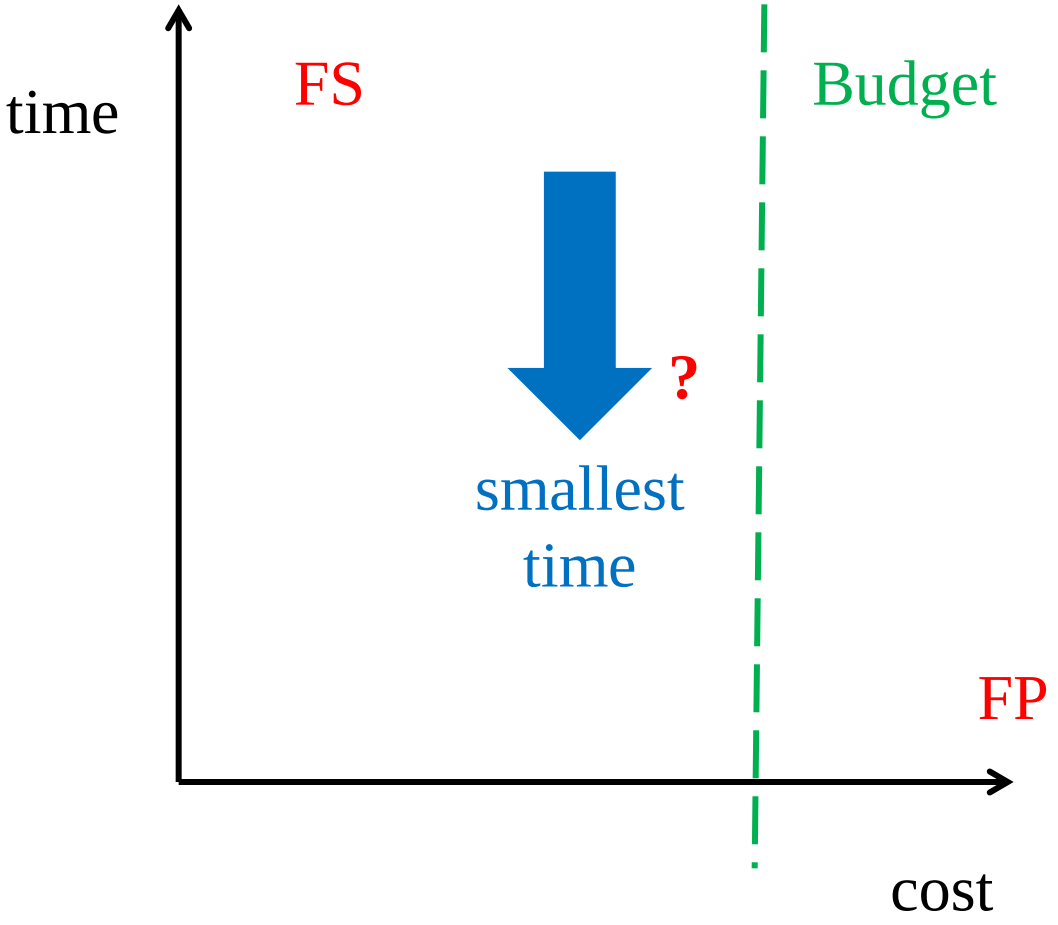
Latency-Cost Tradeoff



Latency-Cost Tradeoff



Latency-Cost Tradeoff



Min-Latency Problem

Framework

- Input:
  - Question budget *b*
  - Number of elements *c*
  - Latency function *L(q)*: time to answer *q* questions
- Reliable workers (use Reliable Worker Layer)
- Proceed in rounds
- First, select budget/round, e.g., (10, 7, 7, 5)
- Then use Question Selection Algorithm in each round

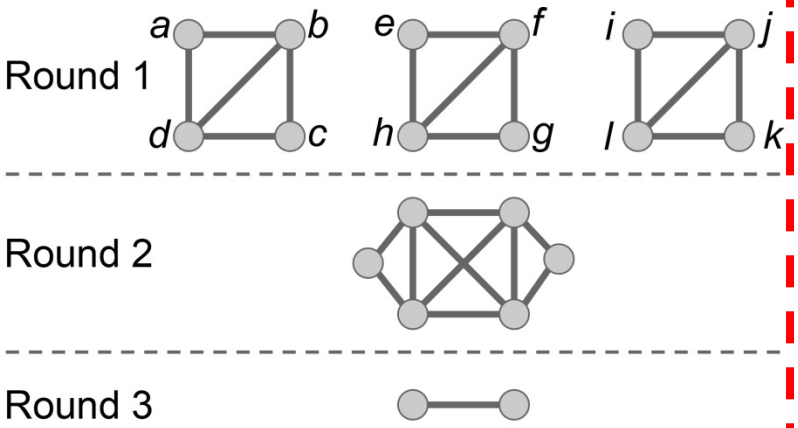
Framework

- Input:
  - Question budget *b*
  - Number of elements *c*
  - Latency function *L(q)*: time to answer *q* questions
- Reliable workers (use Reliable Worker Layer)
- Proceed in rounds
- First, select budget/round, e.g., (10, 7, 7, 5) To Do
- Then use Question Selection Algorithm in each round To Do

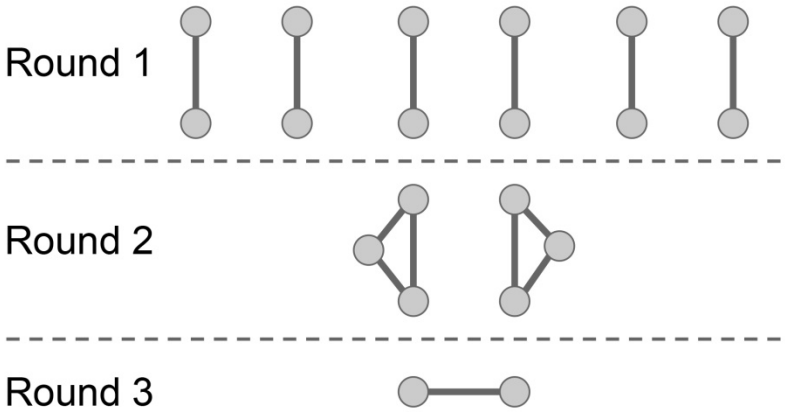
Examples

budget b= 30; elements c=12

budget vector=(15, 10, 1)



budget vector=(6,6,1)

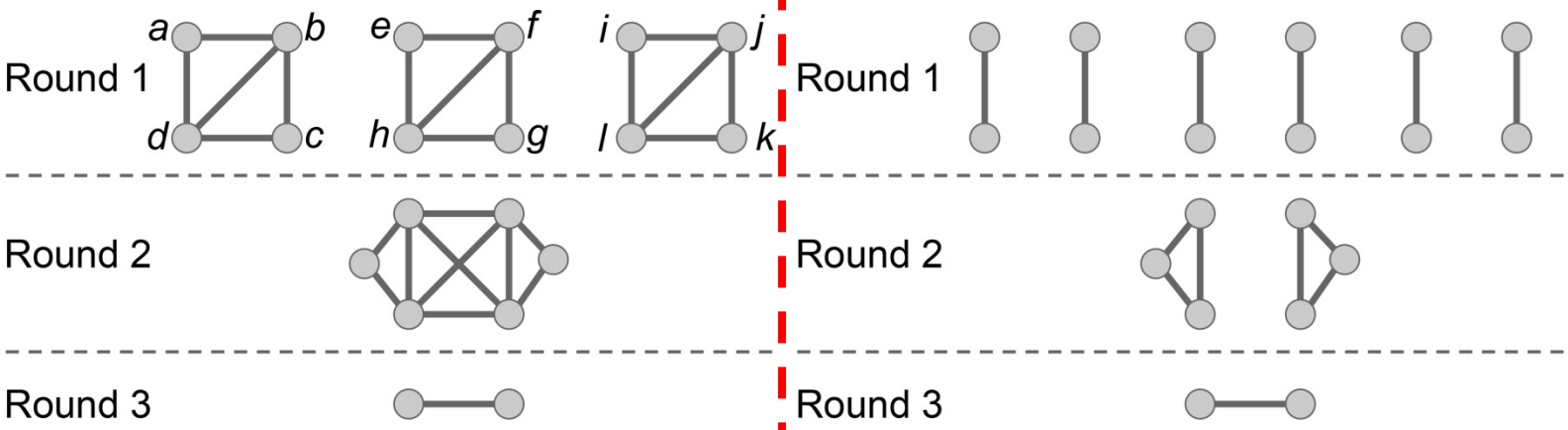


Examples

budget= 30; elements=12

budget vector=(15, 10, 1)

budget vector=(6,6,1)



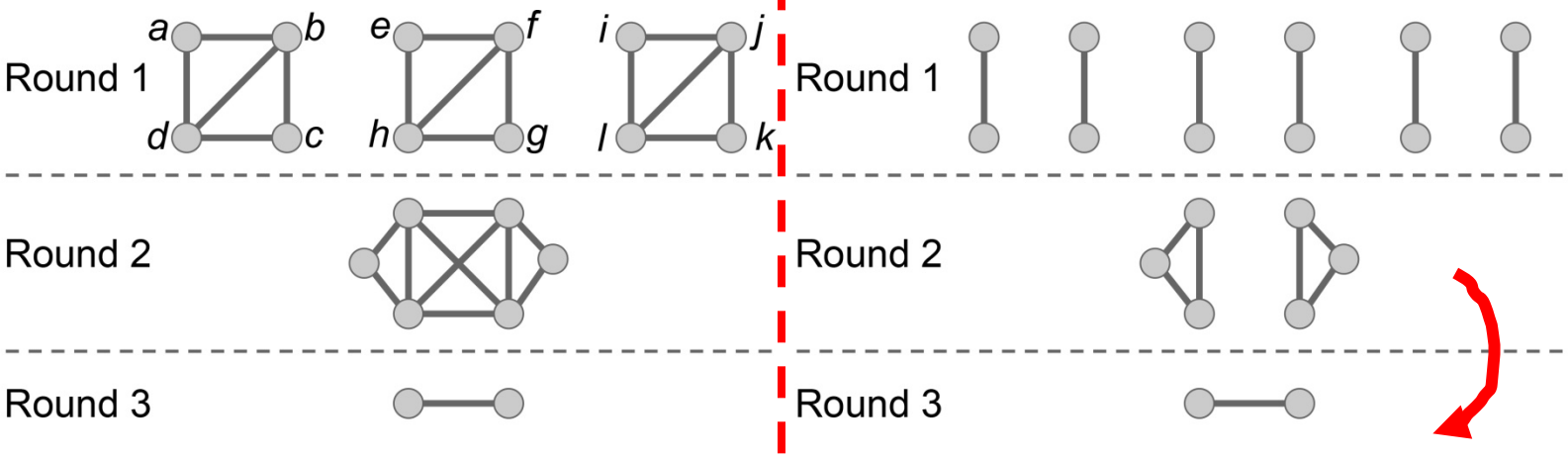
Question Selection: Tournament Graphs

Examples

budget= 30; elements=12

budget vector=(15, 10, 1)

budget vector=(6,6,1)

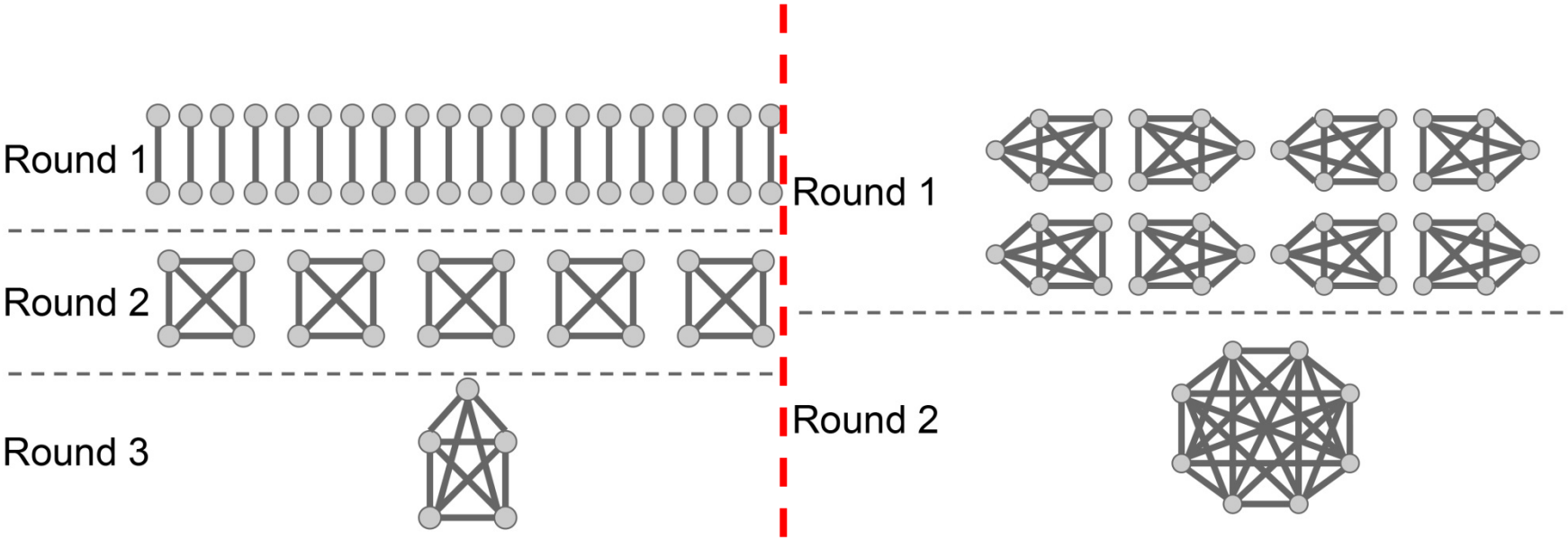


num. groups => remaining elements "memoryless"

Question Selection: Tournament Graphs

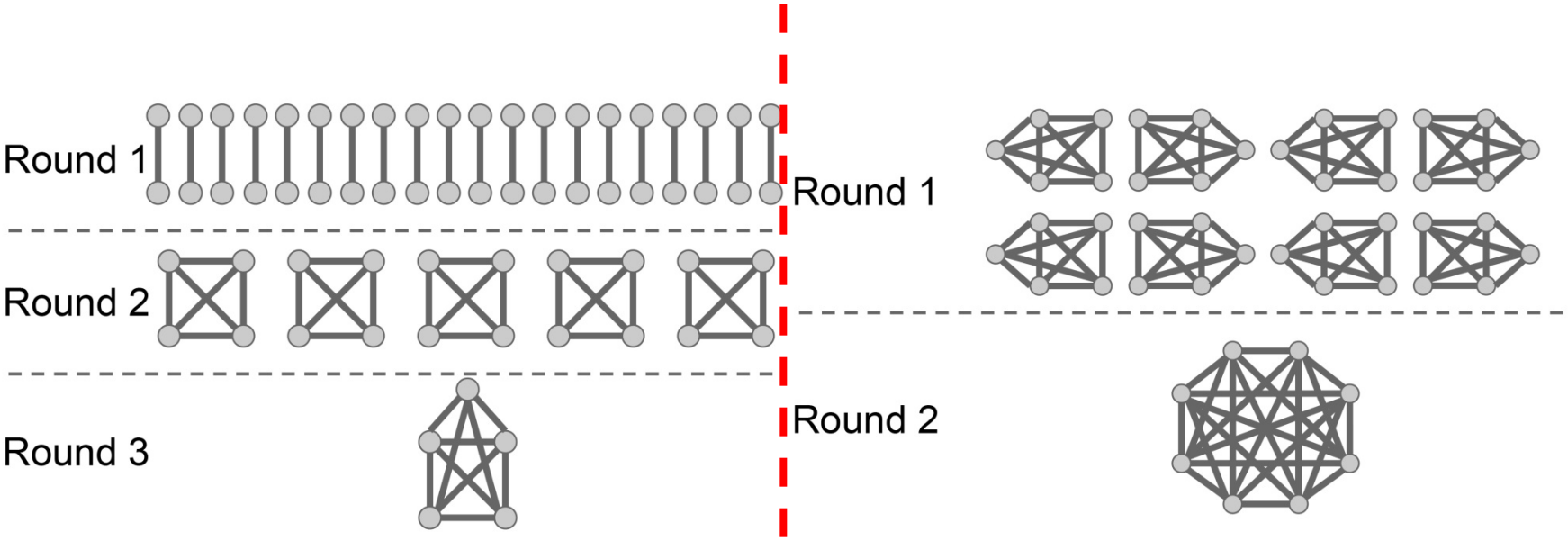
Focus on Tournament Graphs

- How to find optimal budget vector?
- Example, which is best for b=70, c=40



Focus on Tournament Graphs

- How to find optimal budget vector?
- Example, which is best for b=70, c=40



Note: Goal is not minimum questions, but minimum latency

tDP Algorithm

- Assuming tournament graph question selection, our tDP Algorithm finds optimal budget vector
- Can use dynamic programming because of nice properties of tournament graphs

How Does tDP Work (optional slide)

number of elements, c

budget, b

	1	2	3	4	5	6	7
1							
2							
3							
4							
5						y	
6							
7							
8							
9							

optimal latency

fill in table

How Does tDP Work (optional slide)

number of elements, c

budget, b

	1	2	3	4	5	6	7
1							
2			x				
3							
4							
5						y	
6							
7							
8							
9							

optimal latency

- say we start with first round that reduces elements 6->3;
- this tournament costs 3 questions, remaining 5-3=2
- y = L(3) + x

How Does tDP Work (optional slide)

number of elements, c

budget, b	1	2	3	4	5	6	7
1							
2							
3							
4							
5						y	
6							
7							
8							
9							

optimal latency

- consider all possible reductions for first round
- pick for y one that yields minimum latency

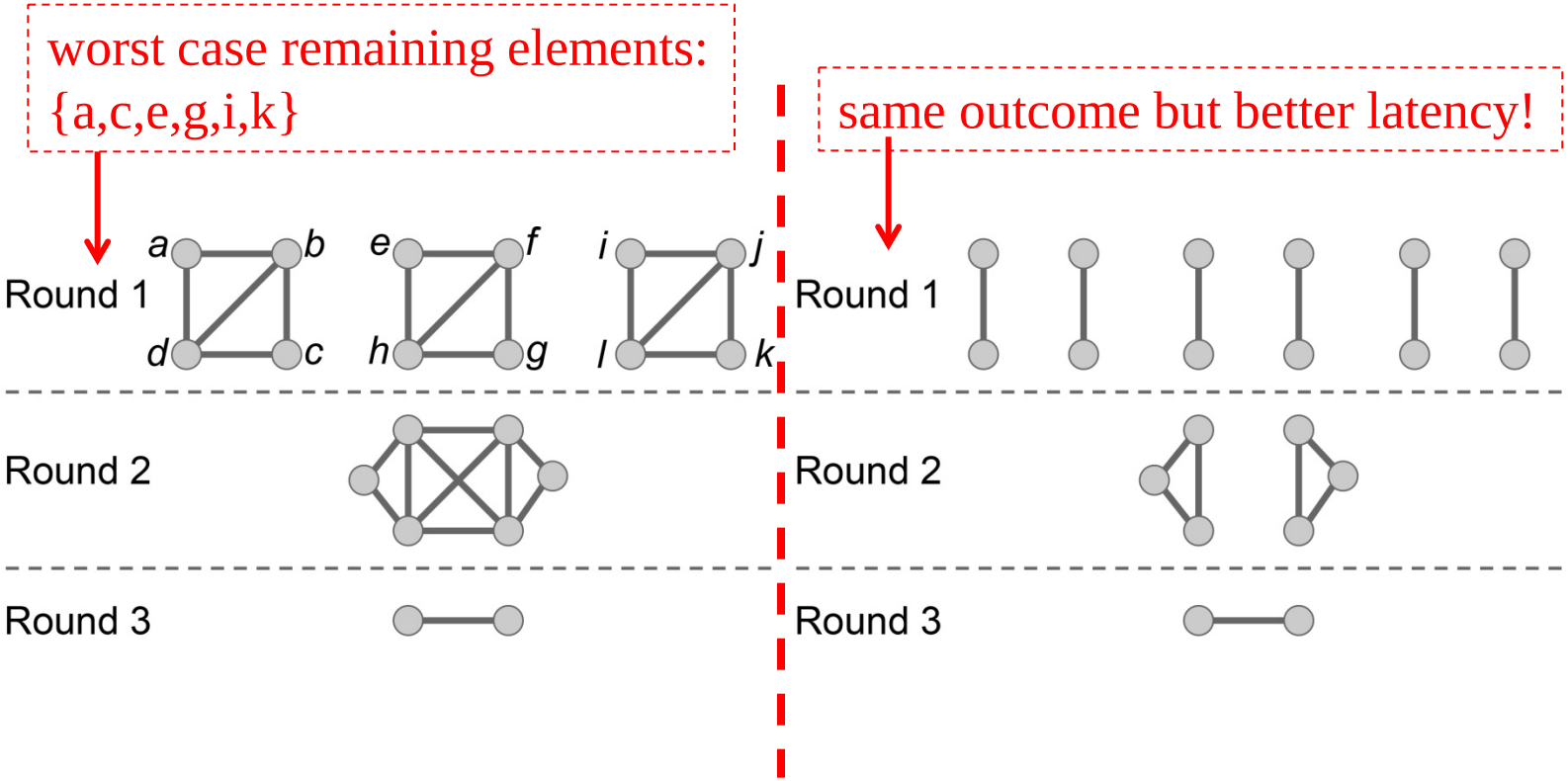
BUT wait, there is more!!

- tDP + tournament graphs has better (lower) worst case latency than any budget allocation scheme with any question selection algorithm!

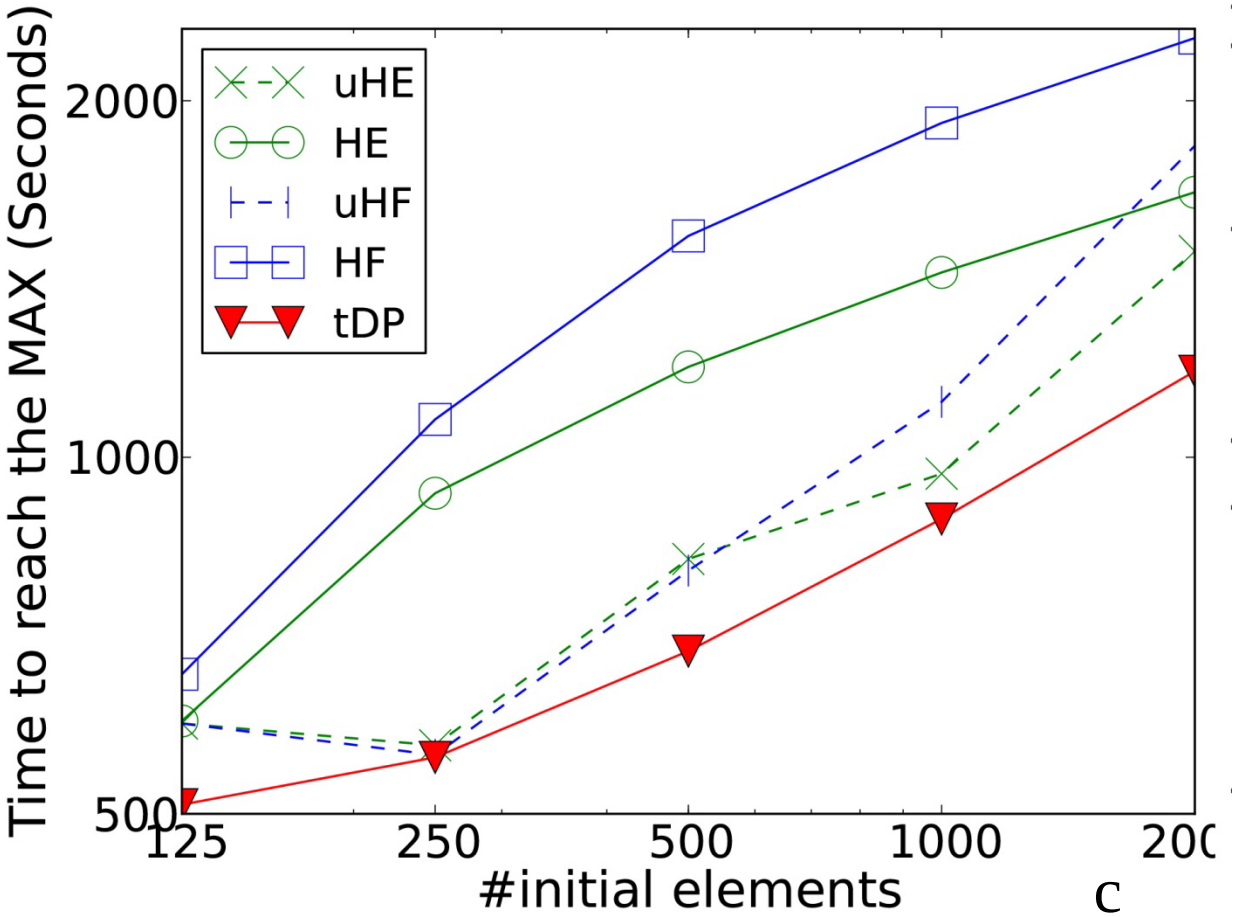
BUT wait, there is more!!

- tDP + tournament graphs has better (lower) worst case latency than any budget allocation scheme with any question selection algorithm!
- And in practice, tDP + tournament graphs is "damn good" for average case latency (see experiments)

Key Insight


worst case remaining elements:
{a,c,e,g,i,k}
same outcome but better latency!
Round 1
a
b
d
c
e
f
h
g
i
j
l
k
Round 2
Round 3
Round 1
Round 2
Round 3

Example of Experimental Results



budget b=4000, all using tournament graphs

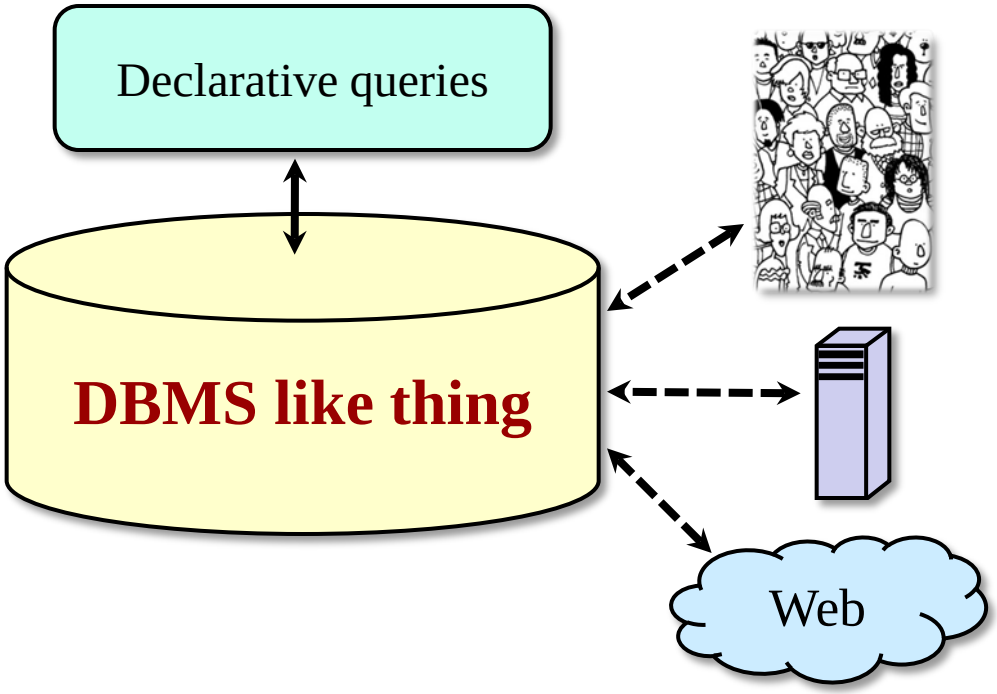
Beyond Max

- Filtering
- Sorting
- Clustering
- Entity Resolution
- Adding terms to a taxonomy
- Building a Folksonomy
- ...

Overview: Crowd Data Management

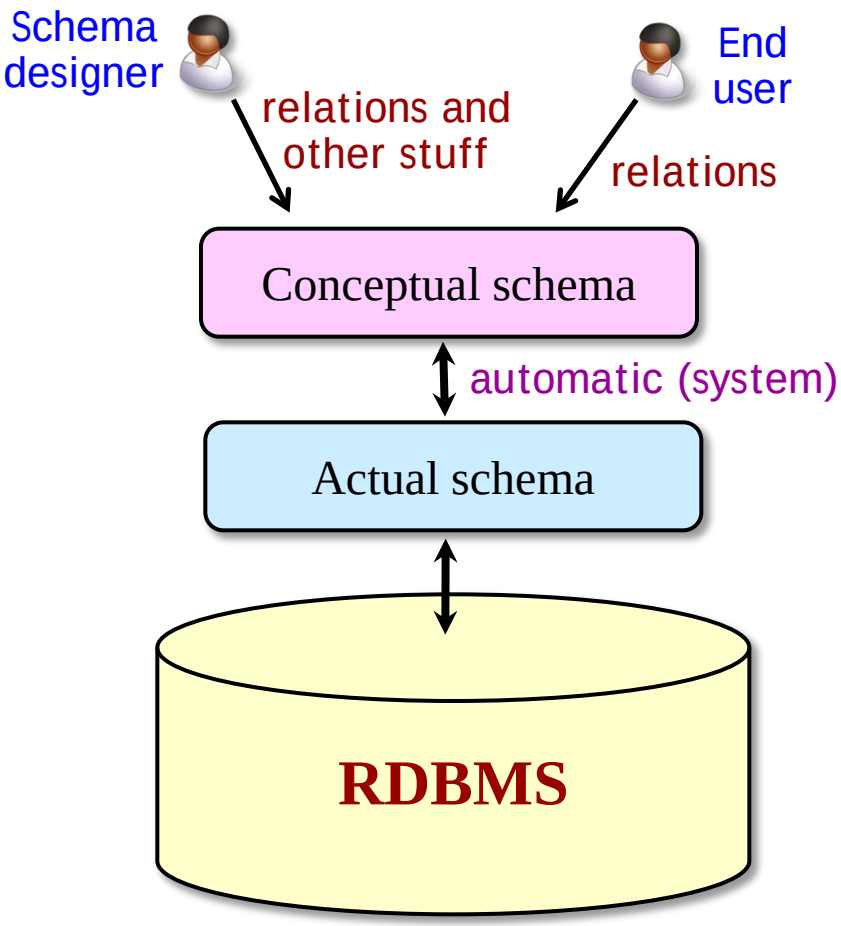
- Data Processing
- Data Gathering
- Searching

Crowd As Information Source



Example #3: DeCo*

*Work with Aditya Parameswaran, Hyunjung Park, Jennifer Widom



Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
•••	•••	•••

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
•••	•••	•••

$\bowtie_{o}$

restaurant
Chez Panisse
Bytes
•••

Anchor

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
•••	•••

Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
•••	•••
•••	•••

Dependent

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
•••	•••	•••

$\bowtie_{o}$

restaurant
Chez Panisse
Bytes
•••

Anchor

fetch rule

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
Bytes	•••

fetch rule

Dependent

restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	•••
•••	**French**

fetch rule

Dependent

fetch rule

Small Example

User view

restaurant	rating	cuisine
Chez Panisse	4.9	French
Chez Panisse	4.9	California
Bytes	3.8	California
•••	•••	•••

$\bowtie_{o}$

resolution rule

resolution rule

restaurant
Chez Panisse
Bytes
•••

Anchor

restaurant	rating
Chez Panisse	4.8
Chez Panisse	5.0
Chez Panisse	4.9
Bytes	3.6
Bytes	4.0
Bytes	•••

Dependent

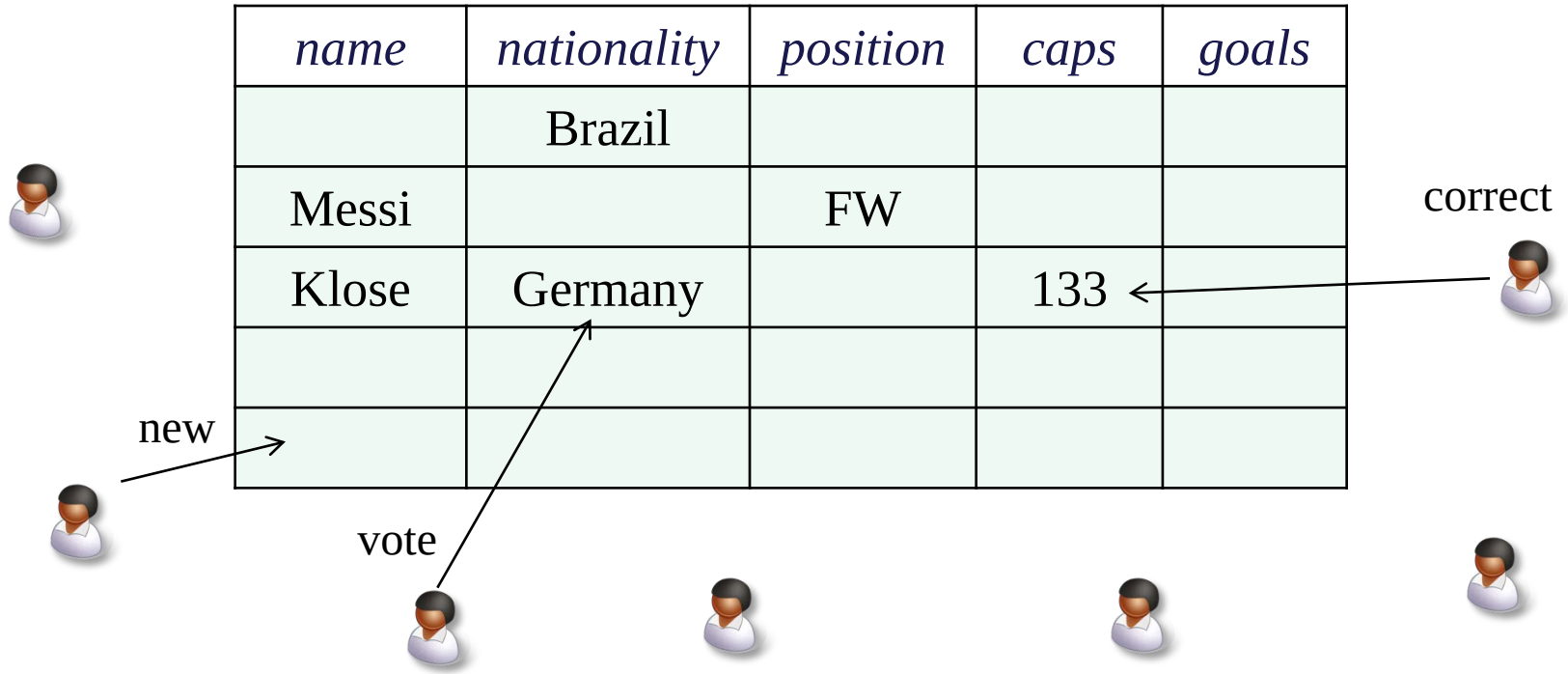
restaurant	cuisine
Chez Panisse	French
Chez Panisse	California
Bytes	California
Bytes	California
Chez Panisse	•••
•••	•••

Dependent

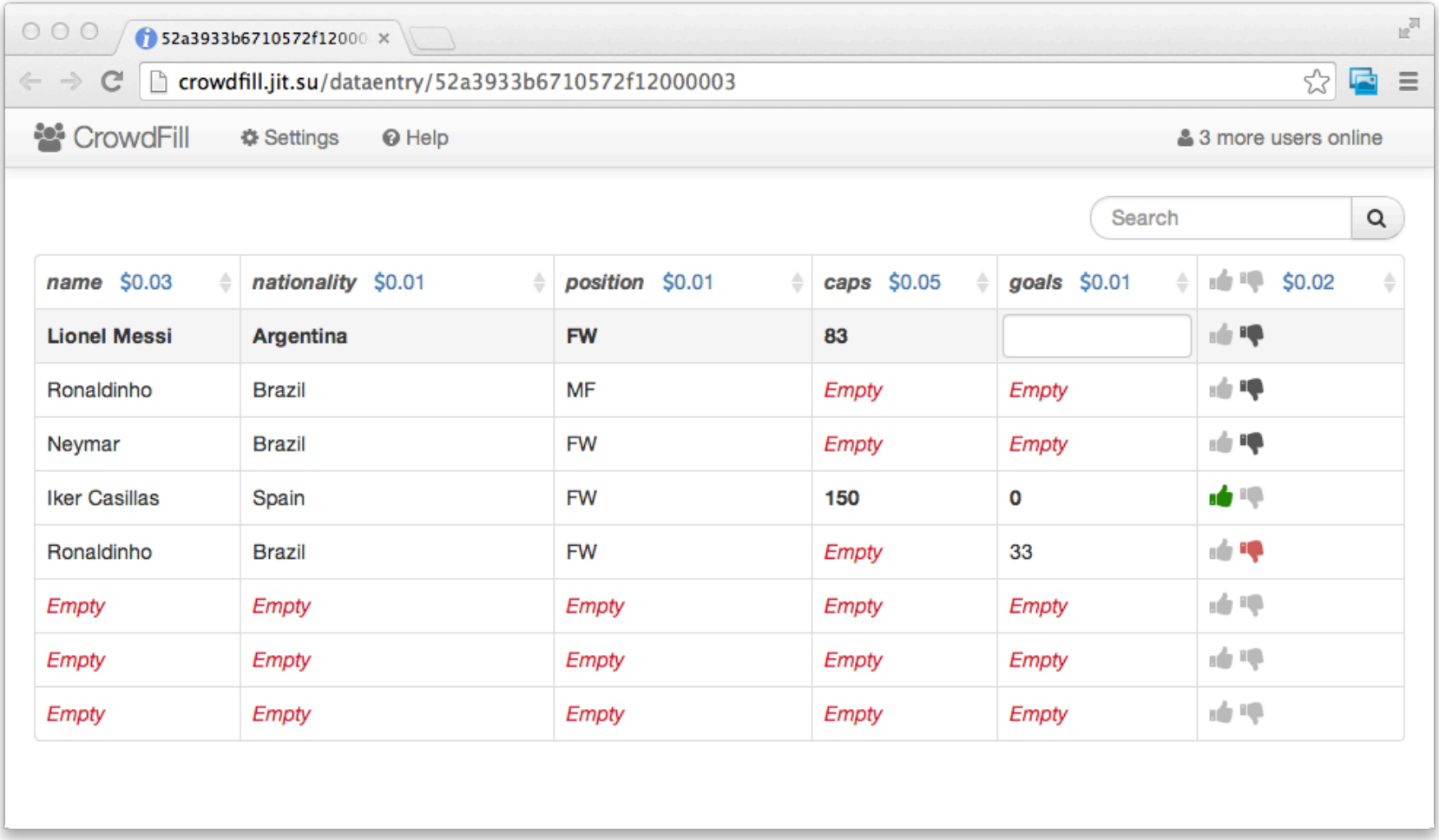
Example #4: CorwdFill*

*Hyunjung Park

- Goal: Collect high-quality structured data from the crowd, while capping total monetary cost and keeping latency low



CrowdFill Prototype



Overview: Crowd Data Management

- Data Processing
- Data Gathering
- Searching

Example #5: DataSift

- Can Your Search Engine Handle This?

buildings in the vicinity of 

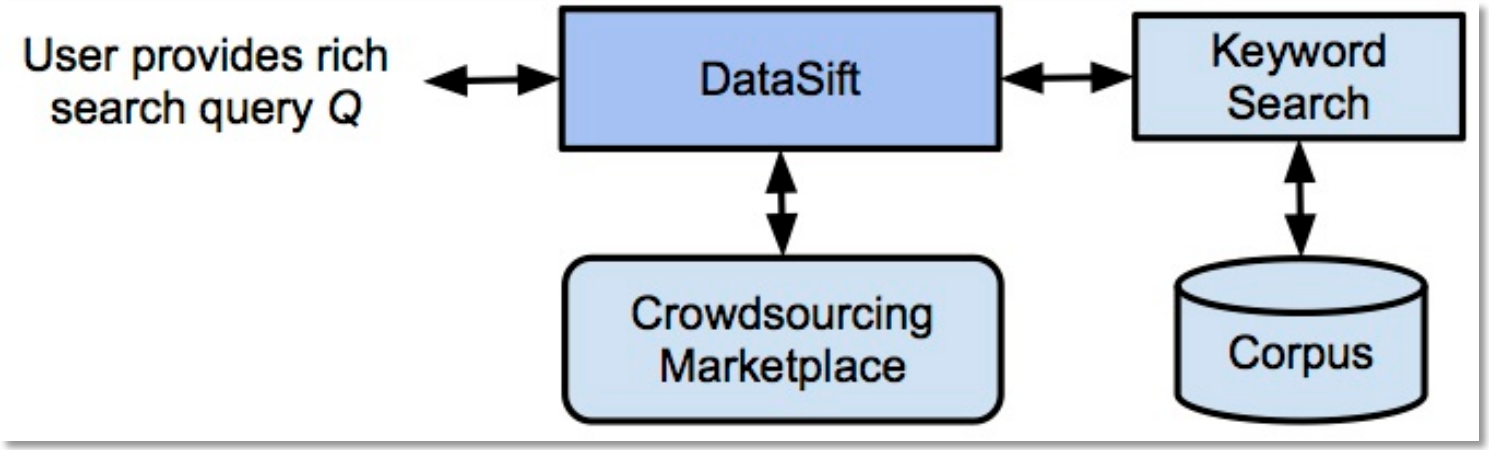
type of cable that connects to 

apartments in a good school district near Somerville, with a bus stop near by

searched for **type of cable that connects to**  using Amazon Products

DataSift Rank	Thumbnail	Product Details
1		**Mediabridge Hi-Speed USB 2.0 Cable - (6 Feet)** Product page: http://www.amazon.co/dp/B001MXLD4G Price: USD 4.99
2		**AmazonBasics USB 2.0 A-Male to B-Male Cable with Lighted Ends - Braided (6 Feet/1.8 Meters)** Product page: http://www.amazon.co/dp/B003ES5ZQE Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 6.99
3		**Epson Stylus USB Printer Cord NEW !! 2.0 A - B Cable 6'** Product page: http://www.amazon.co/dp/B0032GO0SW Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 2.88
4		**USB Printer Cable for HP DeskJet 1000 with Life Time Warranty** Product page: http://www.amazon.co/dp/B004PRXM2C Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 4.95
5		**Mediabridge Hi-Speed USB 2.0 Cable - (10 Feet)** Product page: http://www.amazon.co/dp/B001MSU1HG Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 5.49
6		**Mediabridge Hi-Speed USB 2.0 Cable - (16 Feet)** Product page: http://www.amazon.co/dp/B001MSZBNA Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 7.49

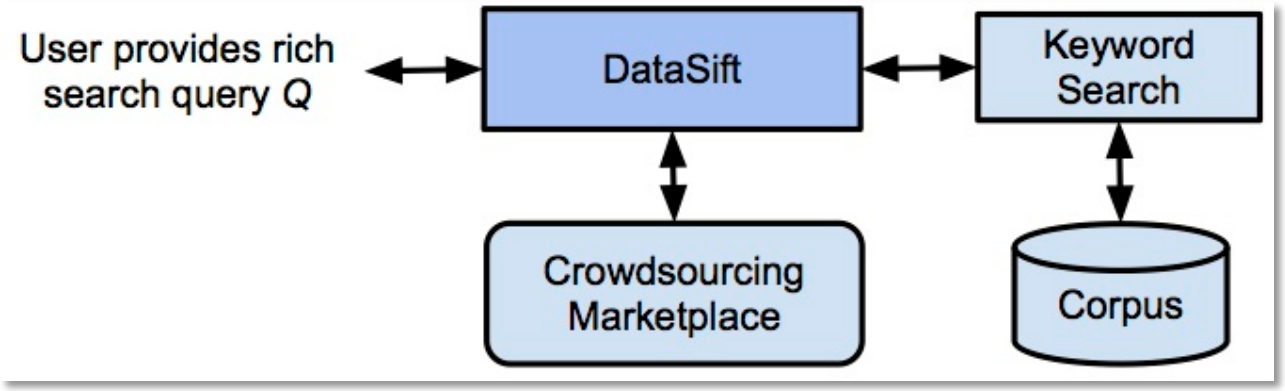
DataSift* can handle rich queries!



*work with Aditya Parameswaran and Ming-Han Teh

DataSift Steps (One Way)

- Start with rich query Q
- Ask crowd for keyword queries {K} (and target)
- Run queries {K} and get some results {D}
- Ask crowd to evaluate {D} (w.r.t. Q}
- "Calibrate" queries {K}
- Get more results



Crowd Component 1: G (Gather)

Provide 3 possible distinct queries you would issue to **Amazon Products** to find products matching the description:

---

type of cable that connects to 

You may need to use your general knowledge to summarize the requirements into a suitable search phase. (See examples)
You should click the "Try Search" button to test your query.

Query 1: Try Search

Query 2: Try Search

Query 3: Try Search

Example A: If the question asks for *"point reyes; showing lighthouse only"*, your search query might be *"point reyes lighthouse"*

Example B: If the question asks for *"SF bridge; night scene"*, your search query might be *"golden gate bridge night"*

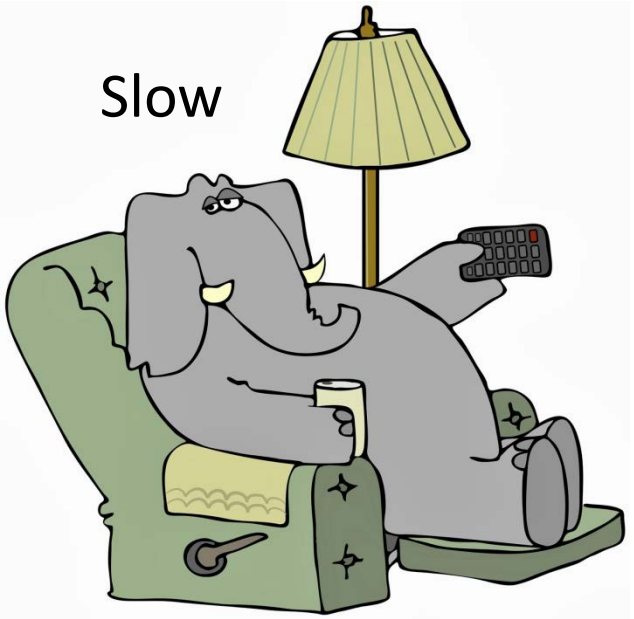
Example C: If the question asks for *"smartphone by apple"*, your search query might be *"iphone"* or *"apple smartphone"*

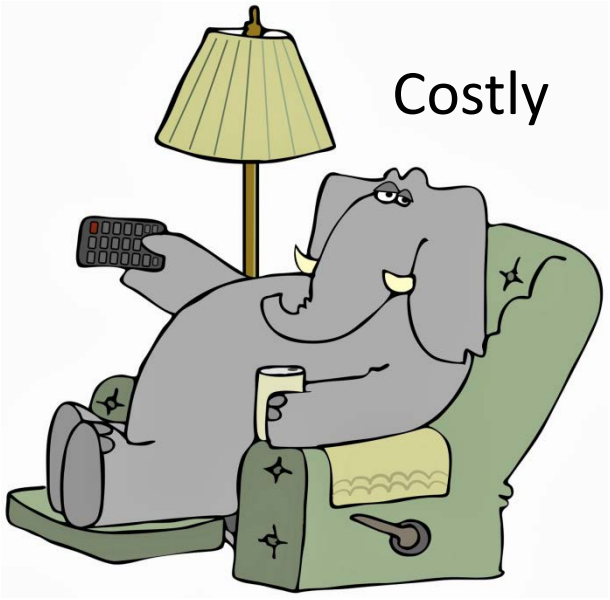
Crowd Component 2: F (Filter)

Which of the following products depicts **type of cable that connects to**  ?

Thumbnail	Product Details	Rate
	**AmazonBasics USB 2.0 A-Male to B-Male Cable with Lighted Ends - Braided (6 Feet/1.8 Meters)** Product page: http://www.amazon.co/dp/B003ES5ZQE Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 6.99 Description: 6FT Braided USB A to B Cable with Light Features: • USB A(with blue LED)--B(with blue LED) Braided Cable • Ships in Certified Frustration-Free Packaging	Yes No
	**Mediabridge Hi-Speed USB 2.0 Cable - (10 Feet)** Product page: http://www.amazon.co/dp/B001MSU1HG Reviews: http://www.amazon.com/reviews/iframe?akid=AKIAJ... Price: USD 5.49 Description: **Mediabridge High-Speed USB 2.0 A-Male to B-Male Cable** * *Full 480-Mbps Transmission Speed of The USB 2.0 Standard* * *Foil and Braid Shielding* * *Gold-Plated Copper Contacts* * *Limited 1 Year Warranty*	Yes No

The Elephant(s) in the Room…





- ***Crowds are Slow! Crowds are Costly!***
- ***Want to use DataSift selectively!***

Conclusion

- Is crowdsourcing for real??
  - YES!!
- Many interesting problems:
  - Crowd data processing
  - Crowd gathering
  - Search
  - Many others!

References

- Challenges in Data Crowdsourcing, IEEE Transactions on Knowledge and Data Engineering, 2016 (with Manas Joglekar, Adam Marcus, AdityaParameswaran, Vasilis Verroios).
- tDP: An Optimal-Latency Budget Allocation Strategy for Crowdsourced MAXIMUM Operations, 2015 ACM SIGMOD International Conference on Management of Data (SIGMOD '15), pp.1047-1062 (with Vasilis Verroios, Peter Lofgren).
- An Overview of the Deco System: Data Model and Query Language; Query Processing and Optimization, SIGMOD Record, Volume 41, Dec 2012 (with Hyunjung Park, Richard Pang, Aditya Parameswaran, Neoklis Polyzotis, and Jennifer Widom).
- Datasift: An Expressive and Accurate Crowd-Powered Search Toolkit, 1st Conf. on Human Computation and Crowdsourcing (HCOMP), Palm Springs, USA, Nov 2013 (with Aditya Parameswaran, Ming Han Teh, Jennifer Widom).