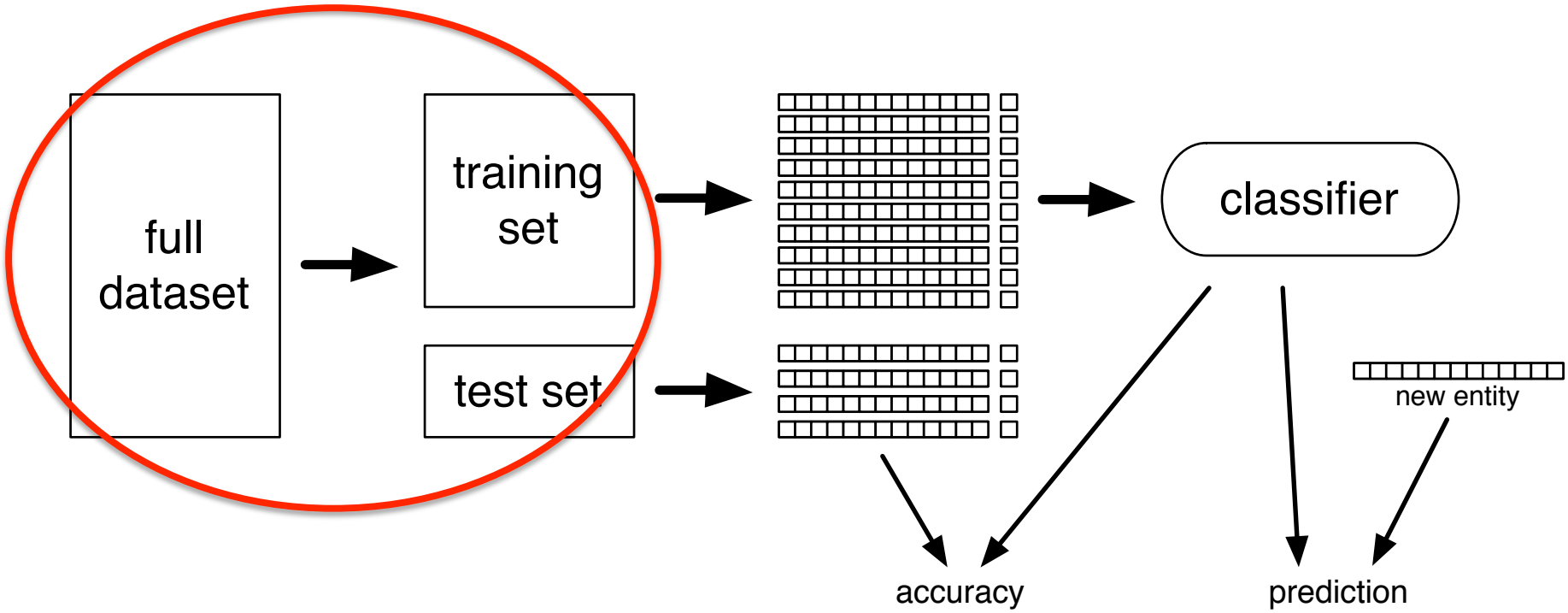
Crowdsourcing for Analytics

Tim Kraska
<kraska@cs.berkeley.edu>

Machines alone are not enough…

Classification



amplab

Machines alone are not enough…

```
SELECT Image
From Pictures
Where Image contains "Dog"
```

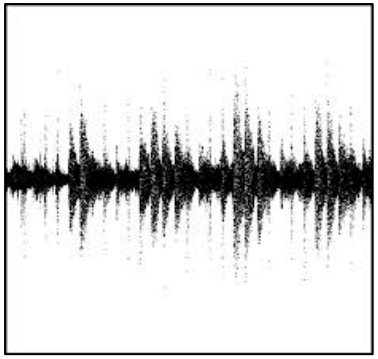
amplab

Adding People to Analytics

Data collection

America's top 10 NASDAQ companies with female CEOs

Transcription



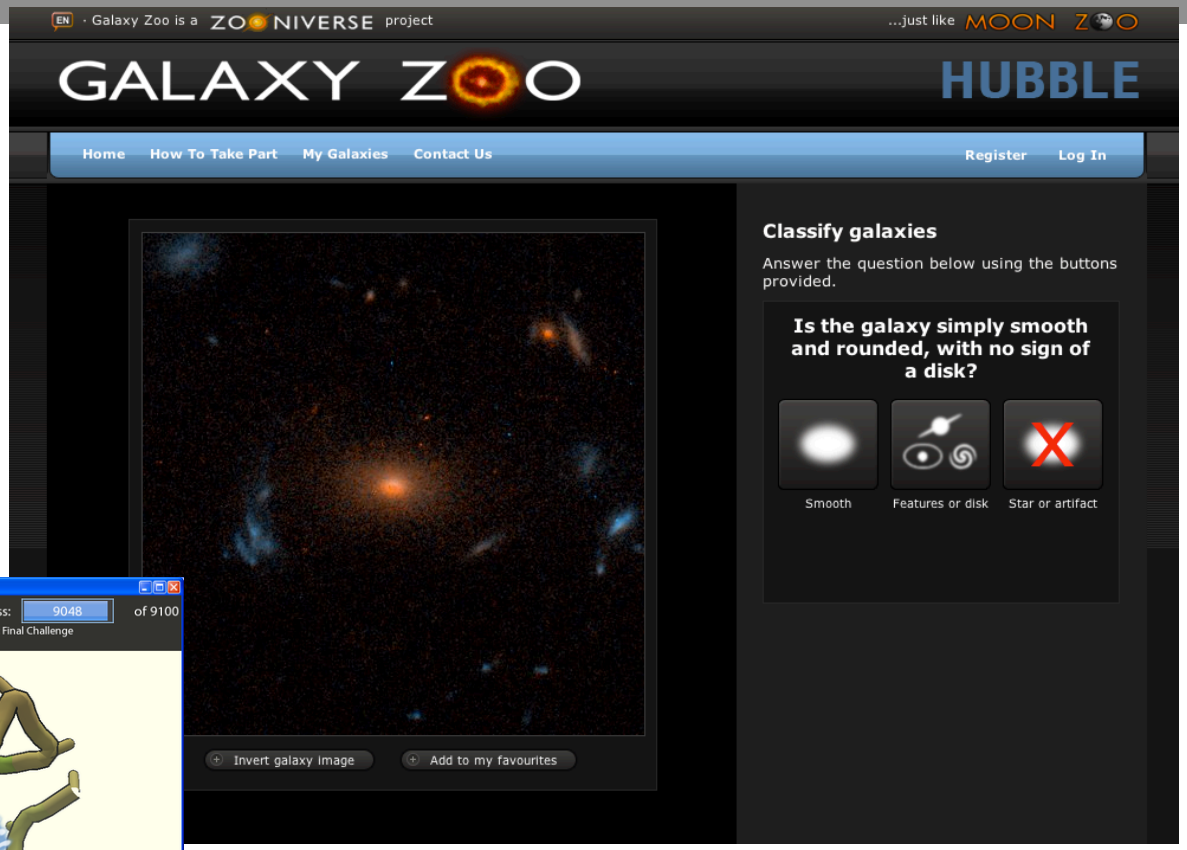
Data cleaning

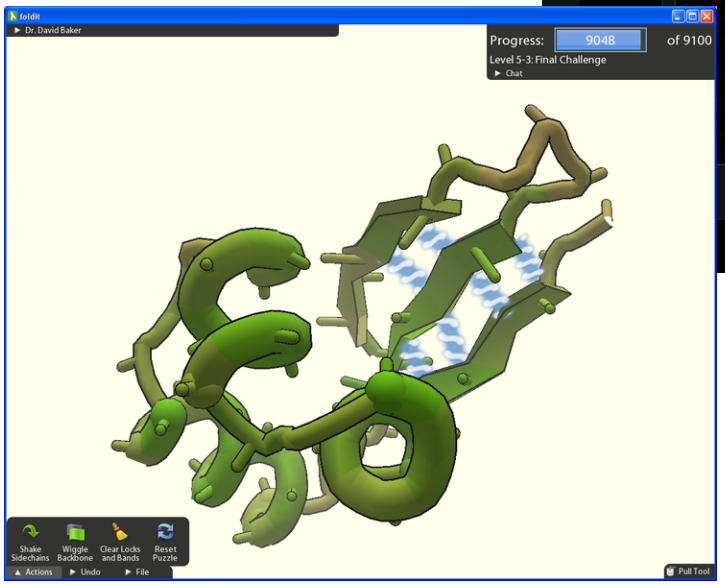
Creativity/Design/Taste

and more

amplab

Science





foldit BETA
Solve Puzzles for Science

amplab

Knowledge Basis

Welcome to Q&A for professional and enthusiast programmers — check out the FAQ!

StackExchange | log in | careers | dev days | chat | meta | about | faq | search

stackoverflow | Questions | Tags | Users | Badges | Unanswered | Ask Question

Top Questions

interesting | 237 featured | hot | week | month

votes	answers	views			
0	0	1	n Random rows for a given attribute - Postgres	sql, postgresql	44s ago Sup3rkiddo 49
1	1 answer	14	Branch descriptions in git, continued	git, branch, task-tracking	48s ago manojlds 16.2k
0	0	1	Where is hostname defined for the anchor element?	javascript	56s ago Chris Aaker 868
2	1 answer	12	User-defined Table Variables in MySQL 5.5?	mysql, stored-procedures, routines	1m ago colonel_px 11
0	2 answers	37	Closing cfpdf tag with </cfpdf> causes error	coldfusion, coldfusion-8, cfeclipse, cfpdf	1m ago Jens Wegar 81
0	0	6	cocoa memory leak by CGAffineTranform or by view	iphone, objective-c, cocoa, memory-leaks, leak	1m ago EmptyStack 9,100

Hello World!

This is a collaboratively edited question and answer site for **professional and enthusiast programmers**. It's 100% free, no registration required.

about » faq »

CAREERS 2.0 by stackoverflow

Senior PHP Engineer
Spreetales
Los Altos, CA; San Francisco, CA

Front End Software Engineer @Rdio
Rdio
San Francisco, CA

Senior Mobile Developer
American Public Media
Oakland, CA

Web Engineer
Monkey Inferno

Structured Data

Freebase | Find topics... | Data | Schema | Apps | Docs

An entity graph of people, places and things, built by a community that loves open data.

Sort by write activity

		Jul 4 – Aug 1 – last week	Facts	Topics	Top User
Featured Data	Film 80 members	43K last week	5M	641K	
Arts & Entertainment	People 87 members	11K last week	7M	2M	
Products & Services	TV 35 members	8K last week	8M	1M	
Science & Technology	Music 100+ members	771 last week	35M	10M	
Society	Business 100+ members	431 last week	2M	611K	
Special Interests	Government 47 members	240 last week	532K	135K	
Sports	Location 52 members	223 last week	10M	999K	
System	Books 47 members	108 last week	29M	6M	
Time & Space					
Transportation					
All					

Google Refine

An open source power tool to fix, discover, experiment, connect and customize your data. Learn more »

What is Freebase?

Learn what an entity graph is, what kind of information it contains, and why you should add your data!

Learn More »

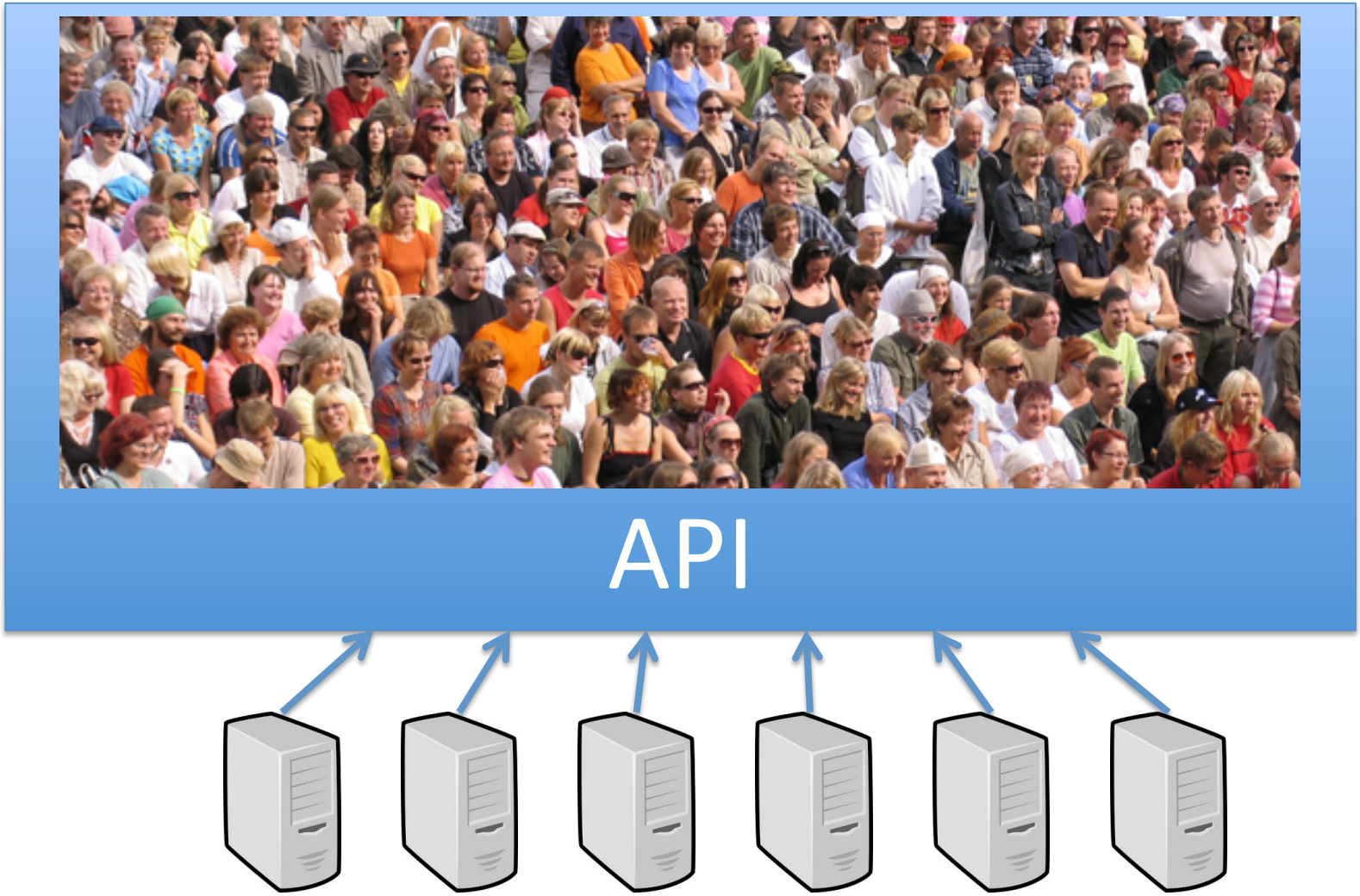
Freebase for Developers

- powerful queryable API
- JavaScript-based hosting framework
- libraries for other languages

Learn More »

amplab

Crowdsourcing for Developers 101



amplab

Micro-Task CrowdSourcing

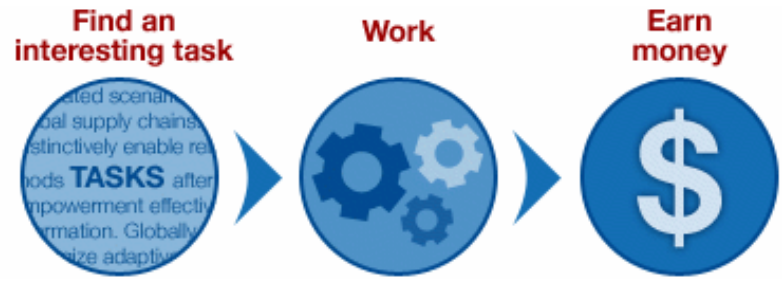
amazon mechanical turk™
Artificial Artificial Intelligence

Make Money
by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Find HITs Now

Get Results
from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

Get Started

amplab

Microtasking – Virtualized Humans

- Current leader: Amazon Mechanical Turk
- Requestors place Human Intelligence Tasks (HITs)
  - Minimum price: $0.01
  - #of replicas (assignments), expiration, **User Interface**
  - API-based: "createHit()", "getAssignments()", "approveAssignments()", "forceExpire()"
  - Requestors approve jobs and payment
- Workers (a.k.a. "turkers") choose jobs, do them, get paid

amazonmechanical turk
Artificial Artificial Intelligence

AMP Lab R | Account Settings | Sign Out | Help

Your Account | HITs | Qualifications

274,745 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Find HITs containing that pay at least $ 0.00 for which you are qualified require Master Qualification GO

All HITs

1-10 of 1157 Results

Sort by: HIT Creation Date (newest first) GO!

Show all details | Hide all details

1 2 3 4 5 Next Last

huge test — View a HIT in this group

Requester: Mr Doe	**HIT Expiration Date:**	Aug 27, 2011 (1 day 5 hours)	**Reward:** $0.01
	Time Allotted:	60 seconds	**HITs Available:** 8424

Validate Brand/Product Information from Product Picture — View a HIT in this group

Requester: Redwood Technologies	**HIT Expiration Date:**	Aug 27, 2011 (11 hours 59 minutes)	**Reward:** $0.01
	Time Allotted:	15 minutes	**HITs Available:** 6

Copy Brand/Product Information from Product Picture — View a HIT in this group

Requester: Redwood Technologies	**HIT Expiration Date:**	Aug 27, 2011 (11 hours 58 minutes)	**Reward:** $0.03
	Time Allotted:	15 minutes	**HITs Available:** 8

huge job test for refactored balancer — View a HIT in this group

Requester: Mr Doe	**HIT Expiration Date:**	Aug 27, 2011 (1 day 5 hours)	**Reward:** $0.01
	Time Allotted:	60 seconds	**HITs Available:** 12390

amplab

amazonmechanical turk
Artificial Artificial Intelligence

AMP Lab R | Account Settings | Sign Out | Help

Your Account | HITs | Qualifications

271,034 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Find HITs containing that pay at least $ 0.00 for which you are qualified require Master Qualification GO

Timer: 00:00:00 of 5 minutes

Want to work on this HIT? Accept HIT

Want to see other HITs? Skip HIT

Total Earned: $0.88
Total HITs Submitted: 80

CrowdDb Equal 7ae03d48-04f0-4259-8d81-492115e106ae

Requester: AMPLab Reward: $0.01 per HIT HITs Available: 5 Duration: 5 minutes

Qualifications Required: None

CrowdDb

'Are these two pictures of the same person?'

and

Yes ◯ No ◯

Please ACCEPT the hit before submitting.

Amount per Assignment:	$0.01	Pending Review:	1
Amount to Approve Outstanding Assignments:	$0.01 What's this?	Reviewed:	0
		Remaining:	0 Add Assignments
		Total:	1

EXPIRATION DATE

Aug 27 2011, 02:25 PM PDT Add Time

» View HIT

Pending Review (1) | Approved | Rejected

Review Submitted Assignments (showing page 1 of 1)

Select Assignments to approve or reject then click "Submit." When you approve an Assignment, the Worker is paid automatically. You will not be charged for Assignments you reject.

Approve	Reject	Worker ID	Result Less		Submission Date
All · None	All · None				
☐	☐	A5RPKDYH18EW0	crowdEqual:	no	Aug 26 2011, 02:33 PM PDT
			jsessionid:	48f4e460c2ac00c3df877f1cdbe5813d	
			crowdDbCallback:	http://128.32.45.115:8082/crowdDbEqual.do?jsessionid=48f4e460c2ac00c3df877f1cdbe5813d	
			wid:		
			id:	10202458	

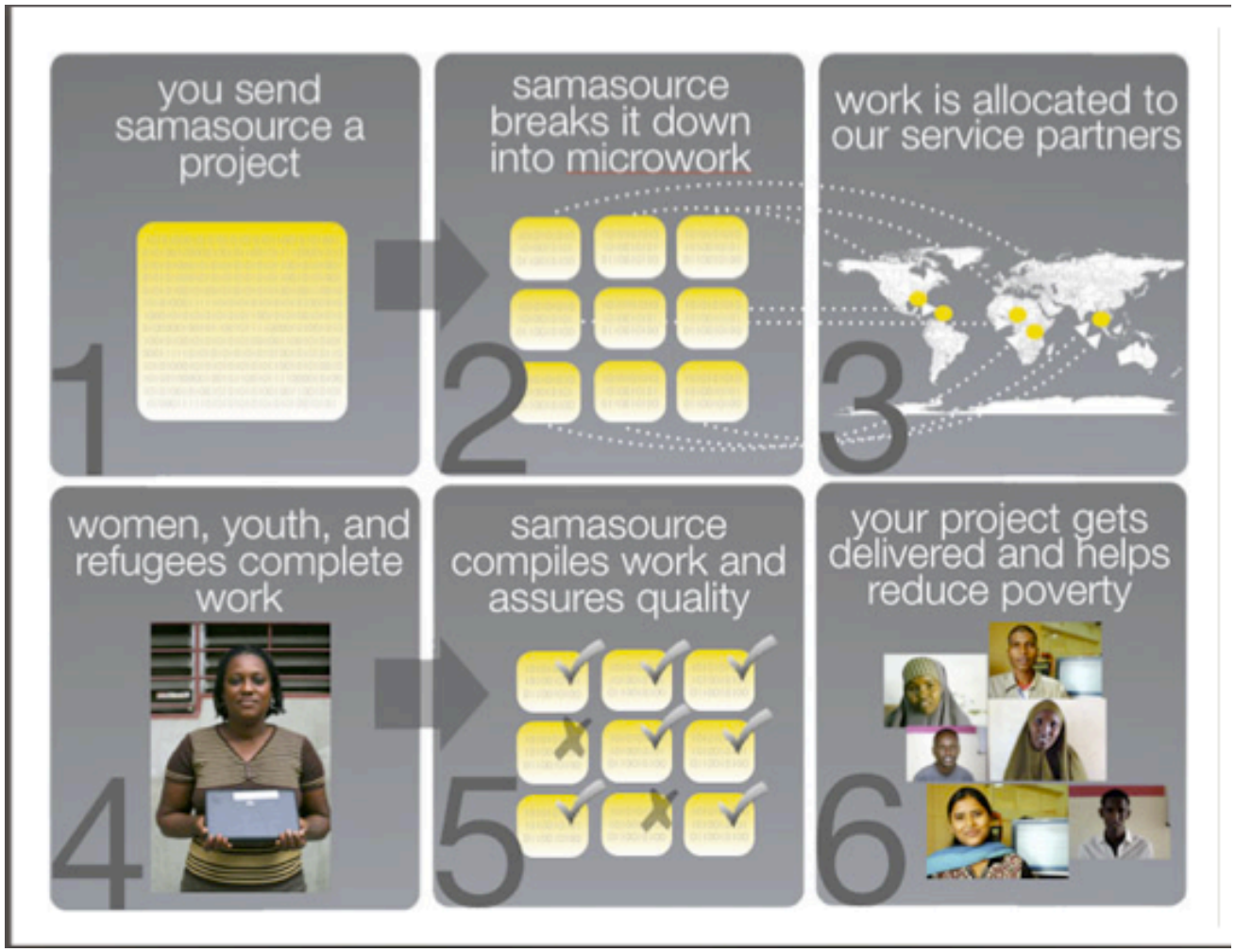
You've chosen to:

approve: none selected

reject: none selected

Submit

Samasource.org



amplab

Challenges

- Quality
- User Interface Design
- Worker motivation
- Task decomposition
- Leverage worker knowledge/capabilities
- Optimization (time/cost)
- ….

Challenges

- **Quality**
- Optimization (time/cost)
- User Interface Design
- Worker Motivation
- Task decomposition
- Leverage worker knowledge/capabilities
- ….

How Can You Trust the Crowd?

amplab

Quality Techniques

- Approval Rate / Demographic Restrictions
- Gold Sets/Honey Pots
- Redundancy
- Qualification Test
- Verification/Review
- Justification/Automatic Verification
- …

Quality Techniques

- **Approval Rate / Demographic Restrictions**
- **Gold Sets/Honey Pots**
- **Redundancy**
- Qualification Test
- Verification/Review
- Justification/Automatic Verification
- …

Approval Rate & Demographic Restrictions

Classify text about consumer electronics — View a HIT in this group

Requester: Buzz Evaluation
HIT Expiration Date: Sep 7, 2011 (1 week 6 days)
Reward: $0.02
Time Allotted: 20 minutes
HITs Available: 3966

Description: Classify text for positive, negative, mixed or neutral tone

Keywords: buzz, classify, coding, tag, sentiment, text, analysis, twitter, blog, social

Qualifications Required:
HIT approval rate (%) is not less than 95
Location is US

+ Easy to setup
+ Transparent
- Easy to defeat
- Causes a lot of trouble

amplab

Approval Rate

HIT Group » I recently did **299 HITs for this requester**.... Of the 299 HITs I completed, **11 of them were rejected** without any reason being given. **Prior to this I only had 14 rejections, a .2% rejection rate. I currently have 8522 submitted HITs, with a 0.3%** rejection rate after the rejections from this requester (25 total rejections). I have attempted to contact the requester and will update if I receive a response. Until then be very wary of doing any work for this requester, as it appears that they **are rejecting about 1 in every 27 HITs being submitted.** posted by ...

fair:2 / 5 fast:4 / 5 pay:2 / 5 comm:0 / 5

[Franklin, Kossmann, Kraska, Ramesh, Xin: CrowdDB: Answering Queries with Crowdsourcing. *SIGMOD*, 2011]

Gold Sets / Honey Pots

- Gold derived from
  - Experts
  - Crowd using high quorum
- Interject trap questions
- Block users in trap and invalidate answers

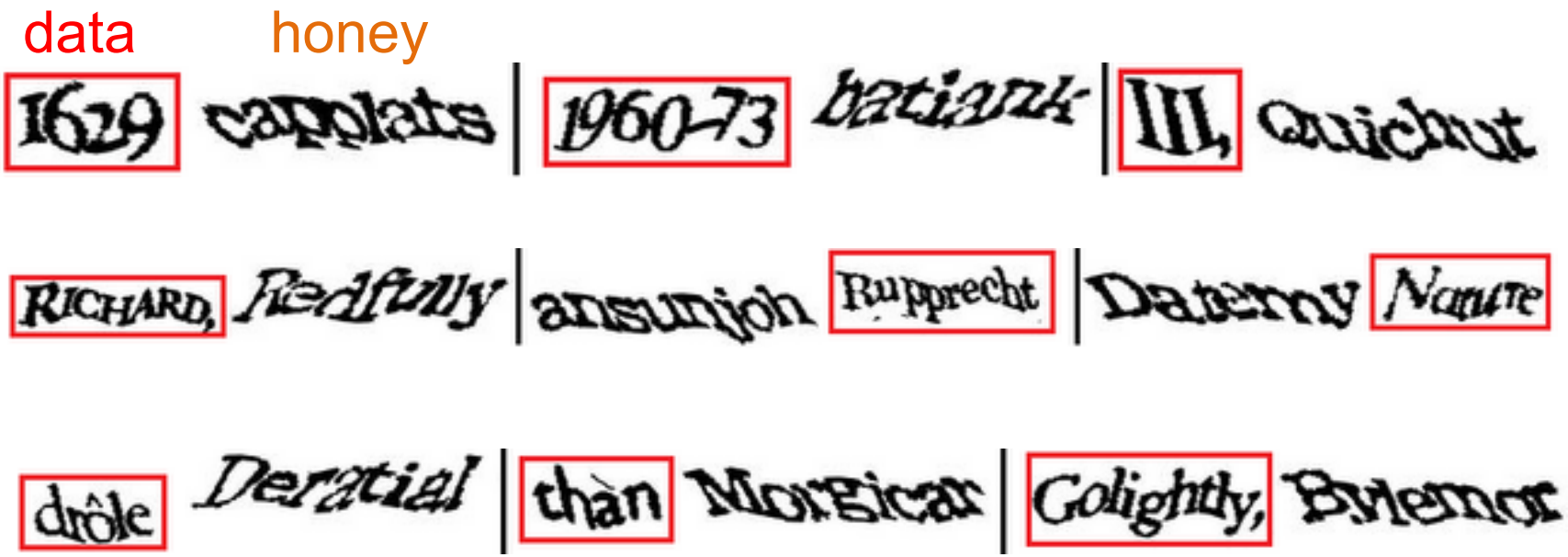
\+ **Often very effective**

\+ **Cost efficient**

\- **Not always applicable**

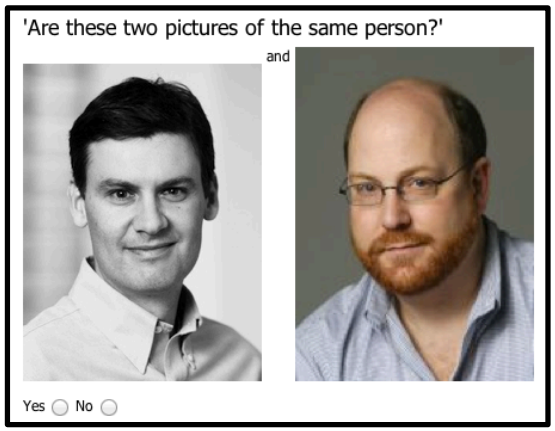
\- **Digging gold is hard**

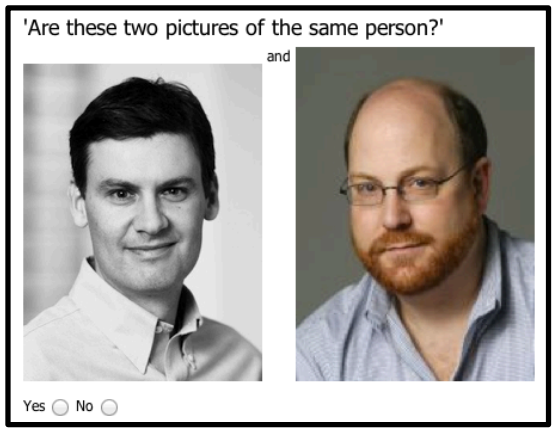
Defeating Honey Pots: reCAPTCHA

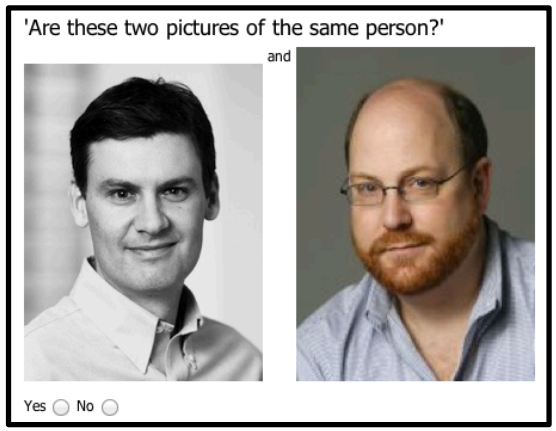


amplab

Redundancy: Quorum Votes







majority vote

result

- \+ **Easy to implement**
- \+ **Hard to defeat**
- \- **Increased cost**
- \- **Masks cases of ambiguity or diversity, "tail" behaviors**
- \- **Does not cover bias**

Challenges

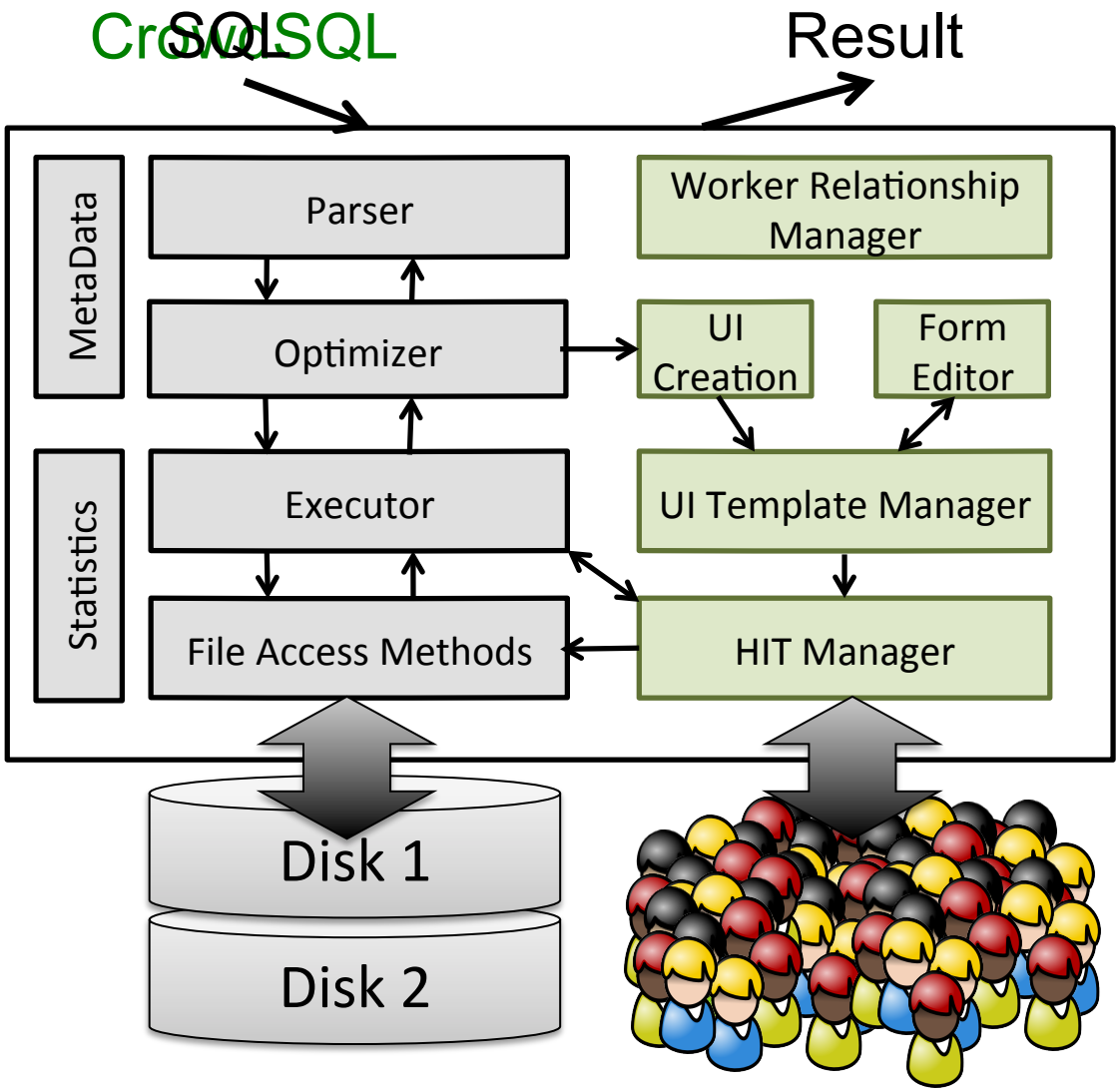
- Quality
- User Interface Design
- Worker motivation
- Task decomposition
- Leverage worker knowledge/capabilities
- Optimization (time/cost)
- ….

amplab

Use Cases

- Data collection:
  - How do my prices compare to the prices of my competitors
  - Finding job candidates (who is graduating from HPI next year)
  - Find green-tech companies in the Bay Area
  - …
- Data cleaning
  - Verifying customer addresses
  - Duplicate elimination
  - …
- Extending data
  - Labeling (spam/not_spam)
  - …

CrowdDB



amplab

CrowdSQL

DDL Extensions:

Crowdsourced columns

```
CREATE TABLE company (
  name STRING PRIMARY KEY,
  hq_address CROWD STRING);
```

Crowdsourced tables

```
CREATE CROWD TABLE department (
  university STRING,
  department STRING,
  phone_no STRING)
PRIMARY KEY (university, department);
```

DML Extensions:

CrowdEqual:

```
SELECT *
FROM companies
WHERE Name ~ "Big Blue"
```

CROWDORDER operators (currently UDFs):

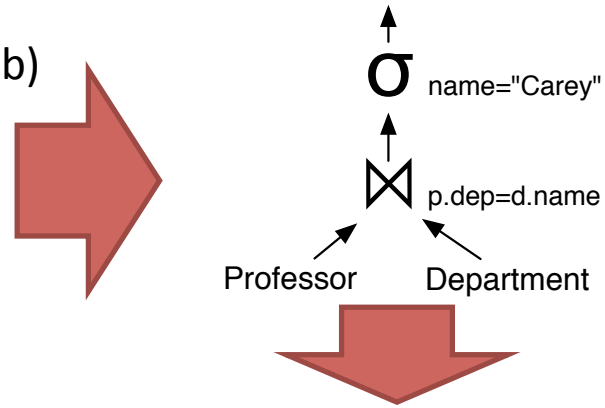
```
SELECT p FROM picture
WHERE subject =
    "Golden Gate Bridge"
ORDER BY CROWDORDER(p, "Which
pic shows better %subject");
```

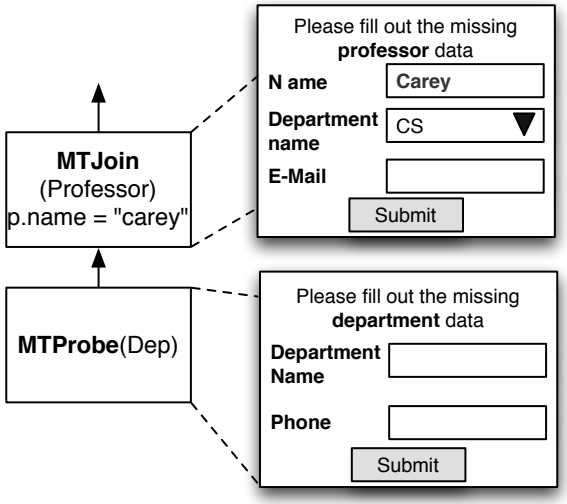

amplab

Optimization: Quality

CROWD TABLE professor(name, e-mail)
CROWD TABLE department(name, phone-nb)

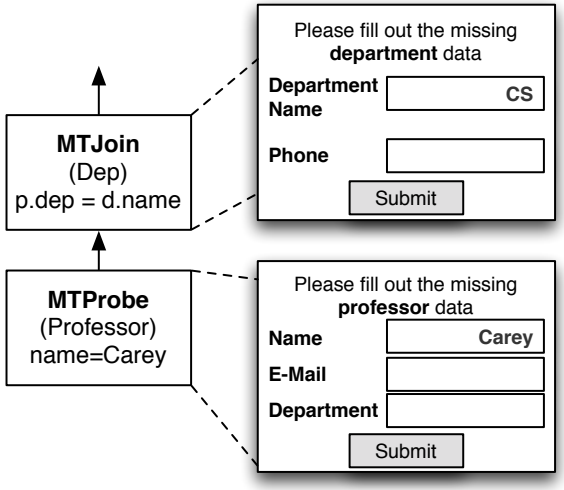
SELECT *
FROM professor p, department d
WHERE d.name = p.dep
AND p.name =“Michael J. Carey”





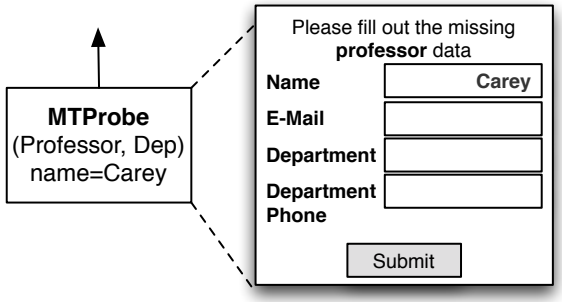
(Department first)

Inefficient



(Professor first)

≈10% Error-Rate

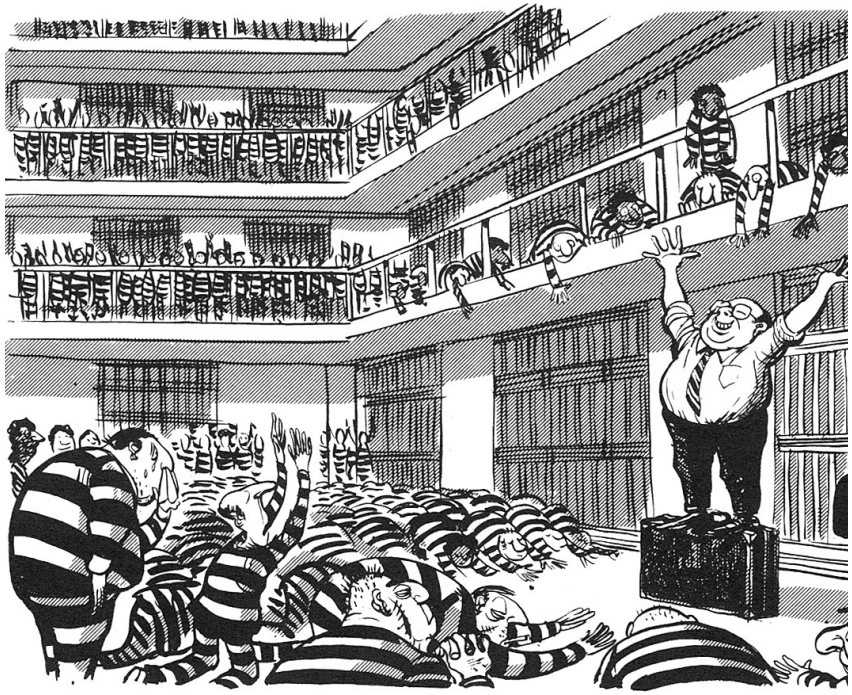


(De-normalized Probe)

≈80% Error-Rate

A Bigger(?) Underlying Issue

Closed-World

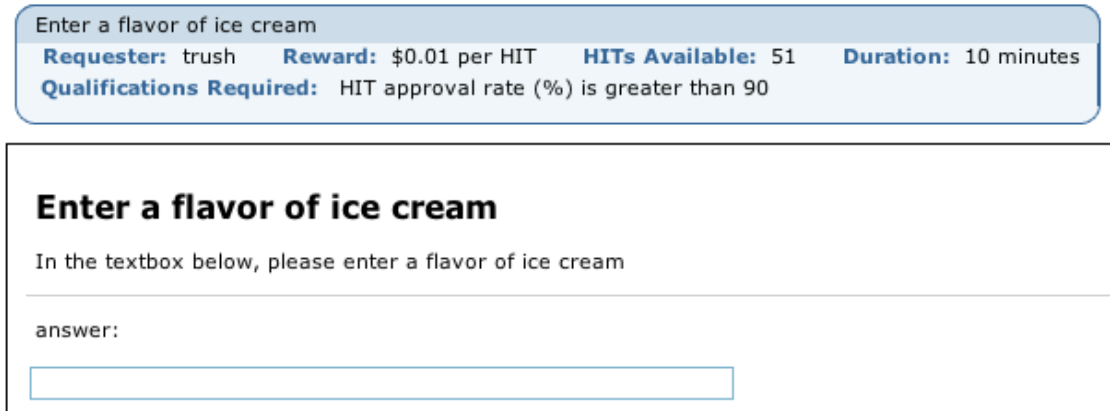


Open-World

amplab

What Does This Query Mean?

```
SELECT COUNT(*) FROM IceCreamFlavors
```



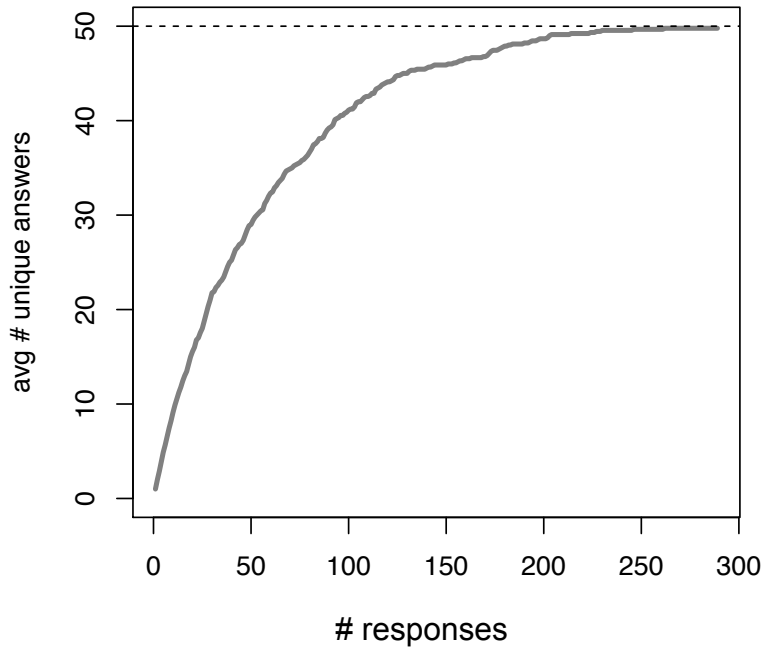
Trushkowsky *et al.* Getting it All From the Crowd, (in preparation) on arxiv

Estimating Completeness

`SELECT COUNT(*) FROM` *`US States`*

US States using Mechanical Turk

Unique items over time



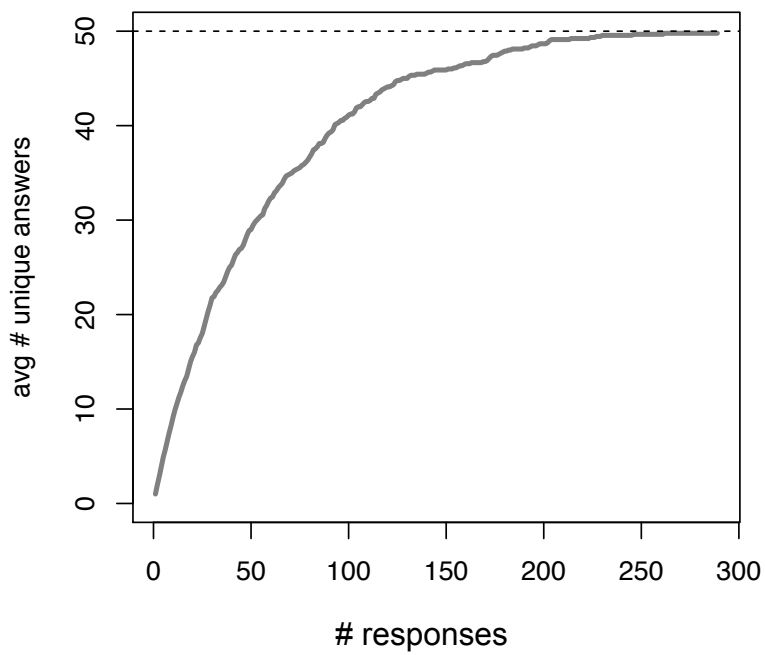
amplab

Estimating Completeness

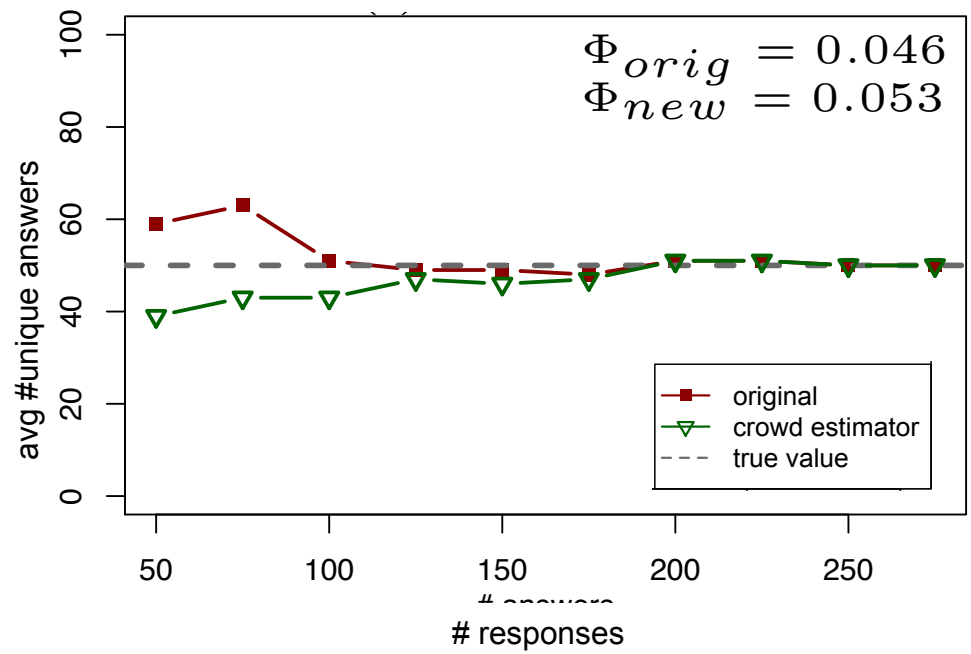
`SELECT COUNT(*) FROM` *`US States`*

US States using Mechanical Turk

Unique items over time



Crowd set-size estimation

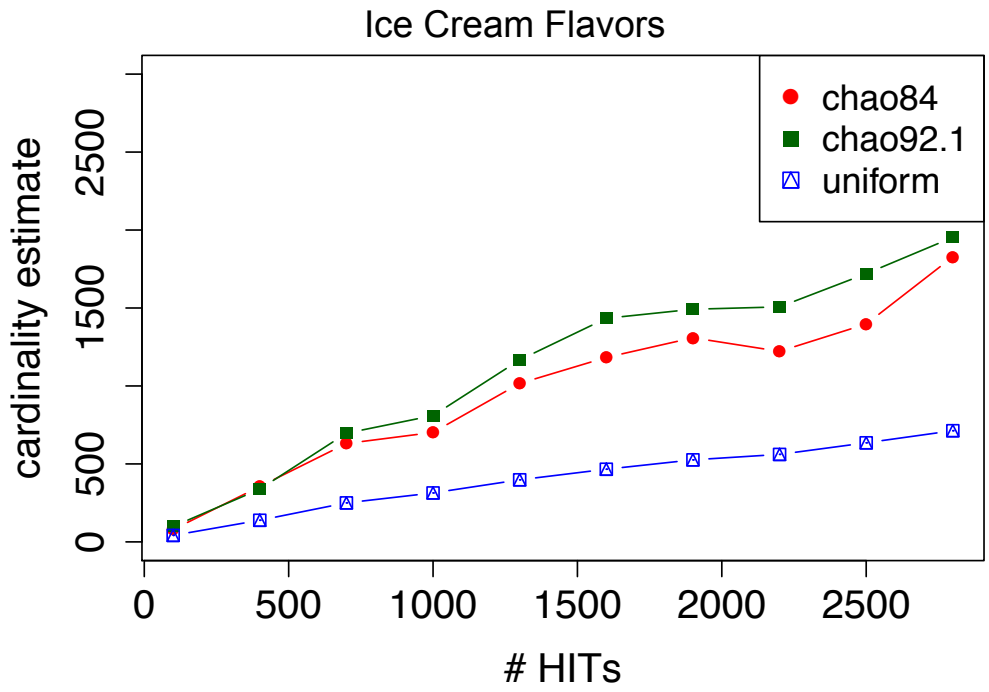


amplab

Estimating Completeness

`SELECT COUNT(*) FROM` *`IceCreamFlavors`*

- Ice Cream Flavors
  - Estimators don't converge
  - Very highly skewed (CV = 5.8)
  - Detect that # HITs insufficient (beginning of curve)



Few, short lists of ice cream flavors (e.g. "alumni swirl, apple cobbler crunch, arboretum breeze,…" from Penn State Creamery)

Pay-As-You-Go

- *"I don't believe it is usually possible to estimate the number of species... but only an appropriate lower bound for that number. This is because there is nearly always a good chance that there are a very large number of extremely rare species"*
  – Good, 1953
- So instead, can ask: "What's the benefit of *m* additional HITs?"

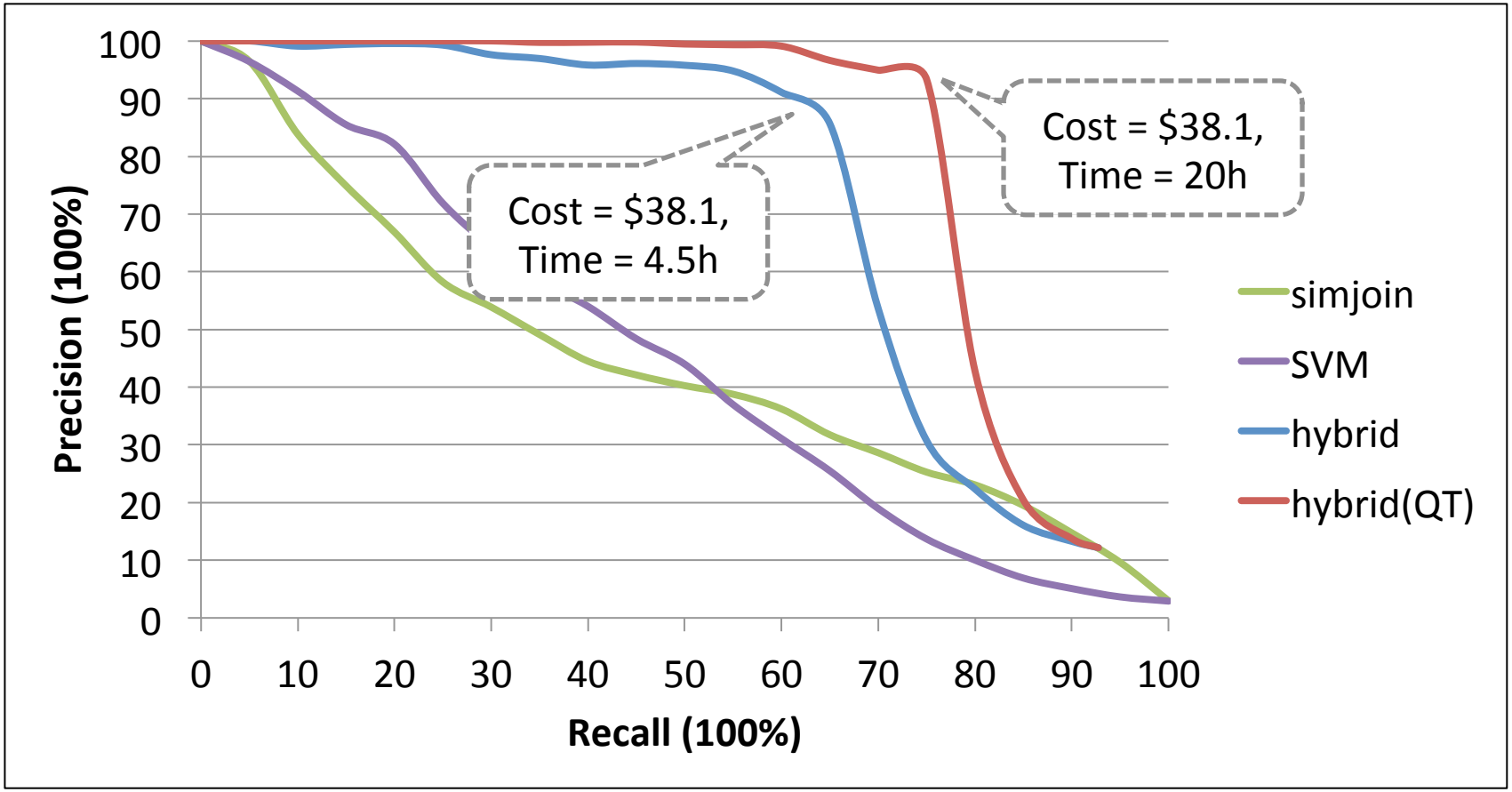
Ice Cream after 1500 HITs

m	Actual	Shen	Spline
10	1	1.79	1.62
50	7	8.91	8.22
200	39	35.4	32.9

Entity Resolution

ID	Product Name	Price
r_1	iPad Two 16GB WiFi White	$490
r_2	iPad 2nd generation 16GB WiFi White	$469
r_3	iPhone 4th generation White 16GB	$545
r_4	Apple iPhone 4 16GB White	$520
r_5	Apple iPhone 3rd generation Black 16GB	$375
r_6	iPhone 4 32GB White	$599
r_7	Apple iPad2 16GB WiFi White	$499
r_8	Apple iPod shuffle 2GB Blue	$49
r_9	Apple iPod shuffle USB Cable	$19

Hybrid Entity Resolution



J. Wang *et al.* CrowdER: Crowdsourcing Entity Resolution, P*VLDB 2012*

amplab

Human-Tolerant Computing

Adding People into the Analytics Lifecycle:

- Inconsistent answer quality
- Incentives
- Latency & Variance
- Open vs. Closed world
- Hybrid Human/Machine Design

Approaches:

- Statistical methods for error and bias
- Quality-conscious Interface design
- Cost (time, quality)-based optimization

Summary

The AMPLab looks into integrating Algorithms, Machines and People for big data analytics

- Crowdsourcing can help with Big Data analytics where machines are not enough
- CrowdDB is a first hybrid Crowd/Cloud data management system following this vision
- Full tutorial: *Crowdsourcing Applications and Platforms: A Data Management Perspective. VLDB, 2011*
- Try it out at mturk.com

Tim Kraska
kraska@cs.berkeley.edu

amplab