CROWD DATA SOURCING & CROWDSOURCING

DATABASE QUESTIONS
HUMAN FACTORS
OPEN QUESTIONS

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes

Roma Tre course
6 May 2019

(2 HOURS)
- INTRODUCTION TO CROWDSOURCING
- DATABASE QUERY PROCESSING WITH THE CROWD
- TASK DEPLOYMENT STRATEGIES

(1 HOUR)
YOUR ASSIGNMENT

(2 HOURS)
- CROWDSOURCING AND HUMAN FACTORS
- OPEN QUESTIONS

Crowdsourcing: a definition

- The act of posting an open call to hire *cheap, immediate, skilled, and easily accessible* labor online
- A place where one finds work, possibly with *remuneration*
- Micro-tasks often easier to complete by humans than by machines

1. Disaster Management in CrowdMap Ushahidi

ALL CATEGORIES

DAMAGE IMAGES

AID IMAGES (REPORTED DURING DAMAGE SEARCH)

TRUSTED REPORTS

OTHER LAYERS [HIDE]

TYPHOON TRAJECTORY

TOMNOD CROWDRANK

HOW TO REPORT

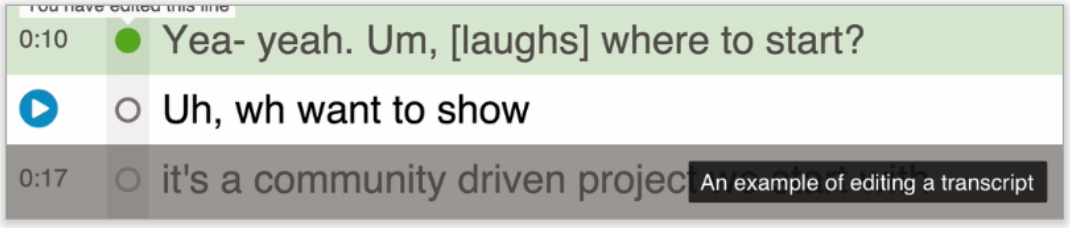
By using an app:
iPhone
Android

By sending an email:

2. Audio Transcription
NYPL Lab

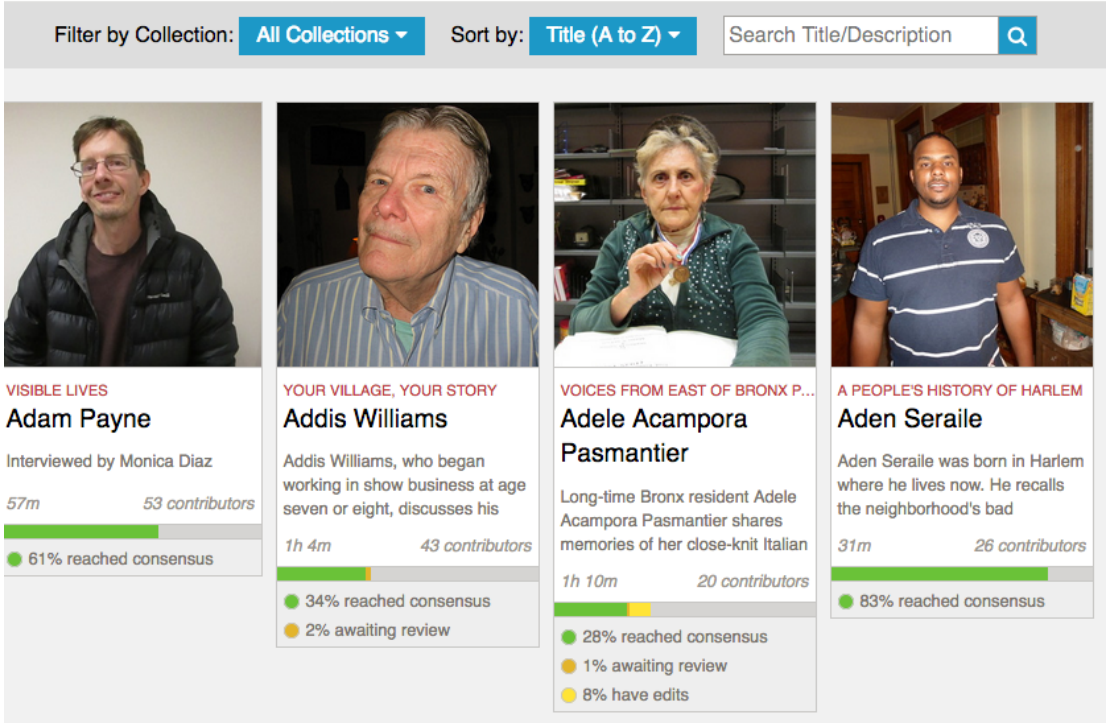
Together We Listen

Help The New York Public Library fix computer-generated transcripts from hundreds of stories from the library's Community Oral History Project.

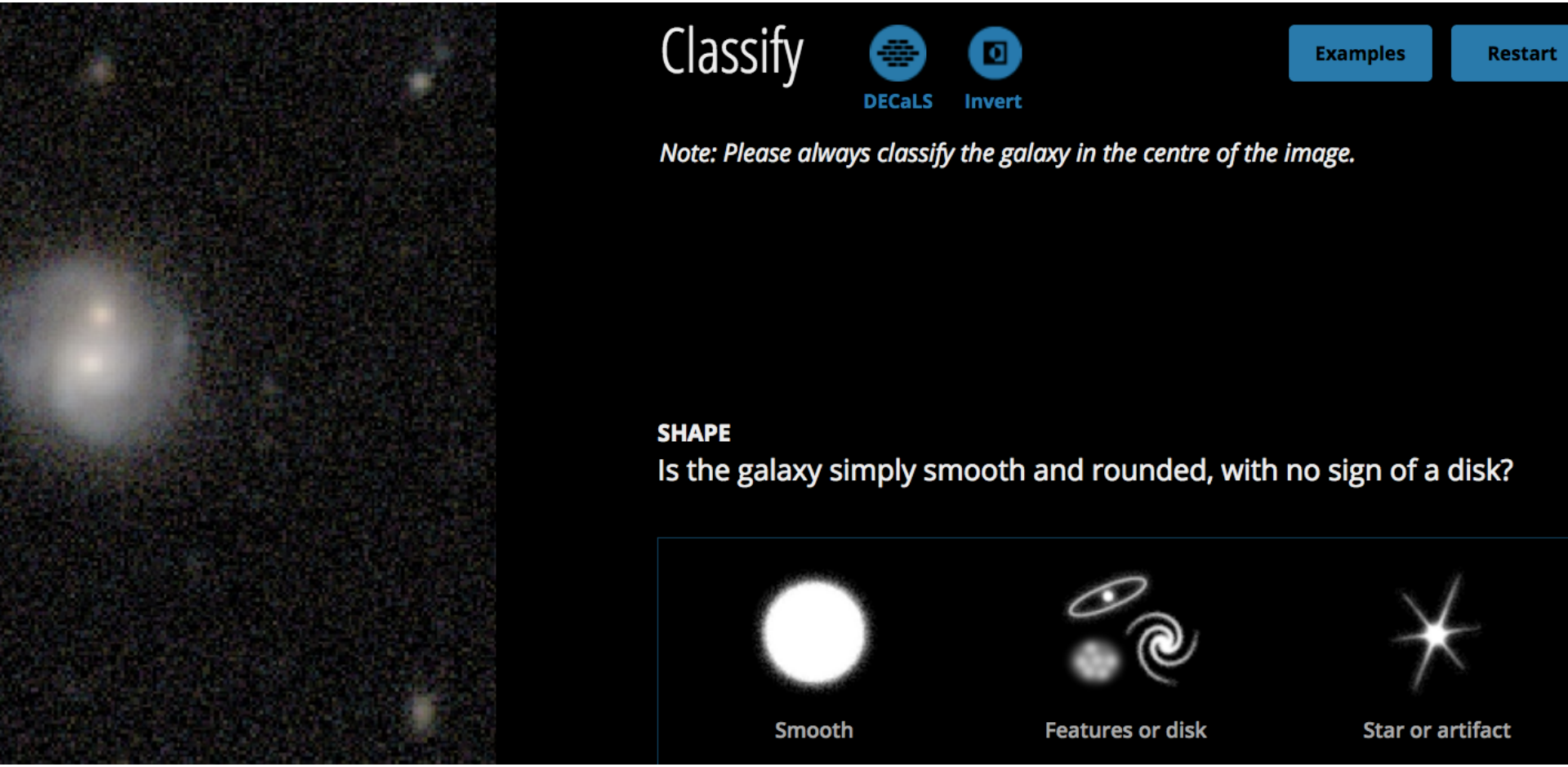


An example of how the transcript editor works (click for sound)

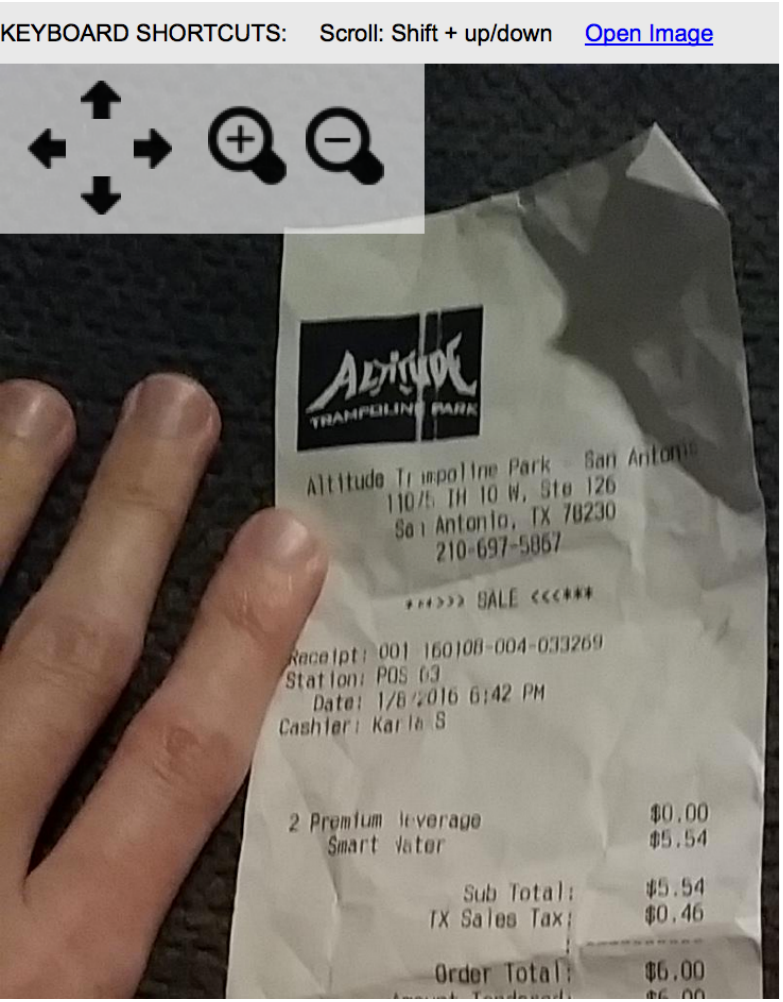
Select an interview to get started.



3. Galaxy classification
Galaxy Zoo



4. Receipt Transcription on AMT



Classify Receipt

Hit Reward: $0.02

Real readable original receipt | Not a receipt or not readable

The following details can often be found at the top or bottom of the receipt. Enter as much information as you can find.

Find and enter the business phone number:

Phone

Example: (888) 555-1234 or 8885551234

Find and enter the business address:

Address

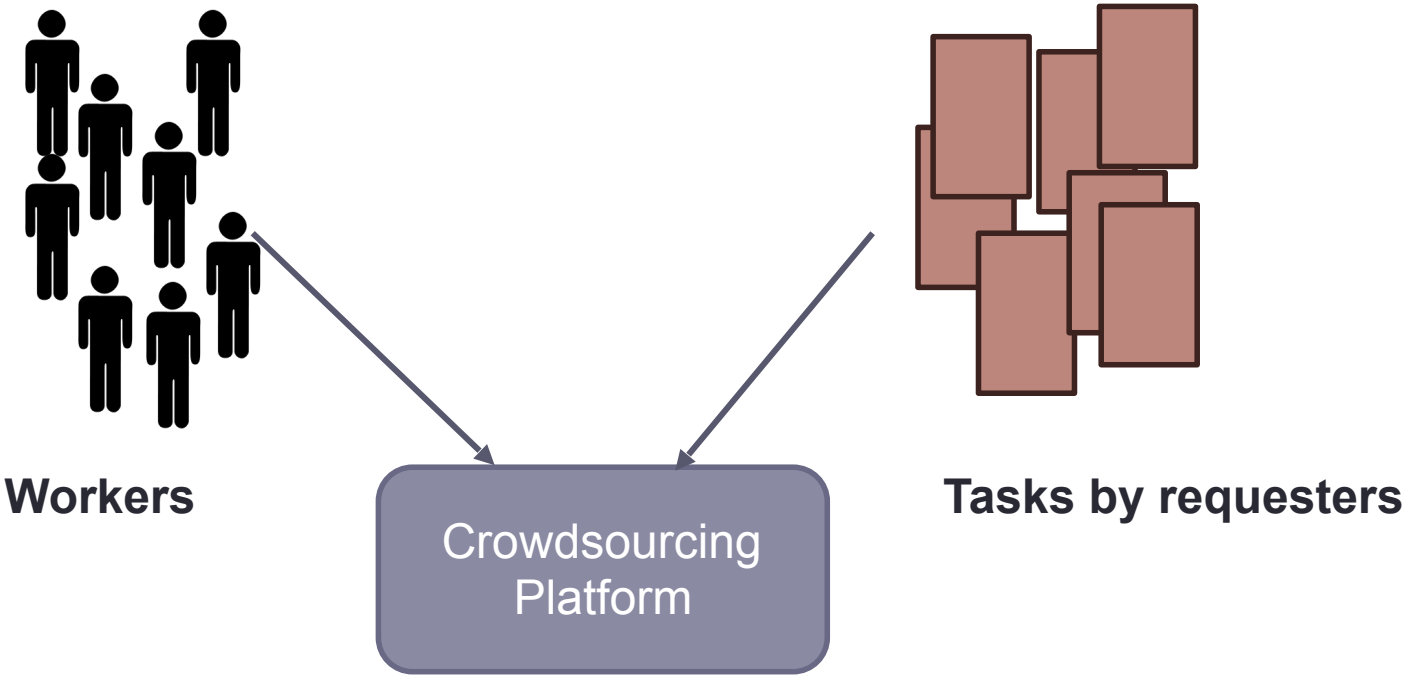
City

State | Postal code

Example: 321 Fake Street, Los Angeles, CA, 90210

Next

5. Generic/Horizontal Crowdsourcing



pybossa

foulefactory

Prolific

amazon mechanical turk

Crowd4U

CrowdFlower

figure eight

THE UPCOMING MATERIAL WAS PREPARED BY **TOVA MILO (PROF. TEL AVIV U.)** WITH WHOM I GAVE A VARIANT OF THIS COURSE AT THE EDBT SUMMER SCHOOL LAST SEPTEMBER

CROWD **DATA** SOURCING

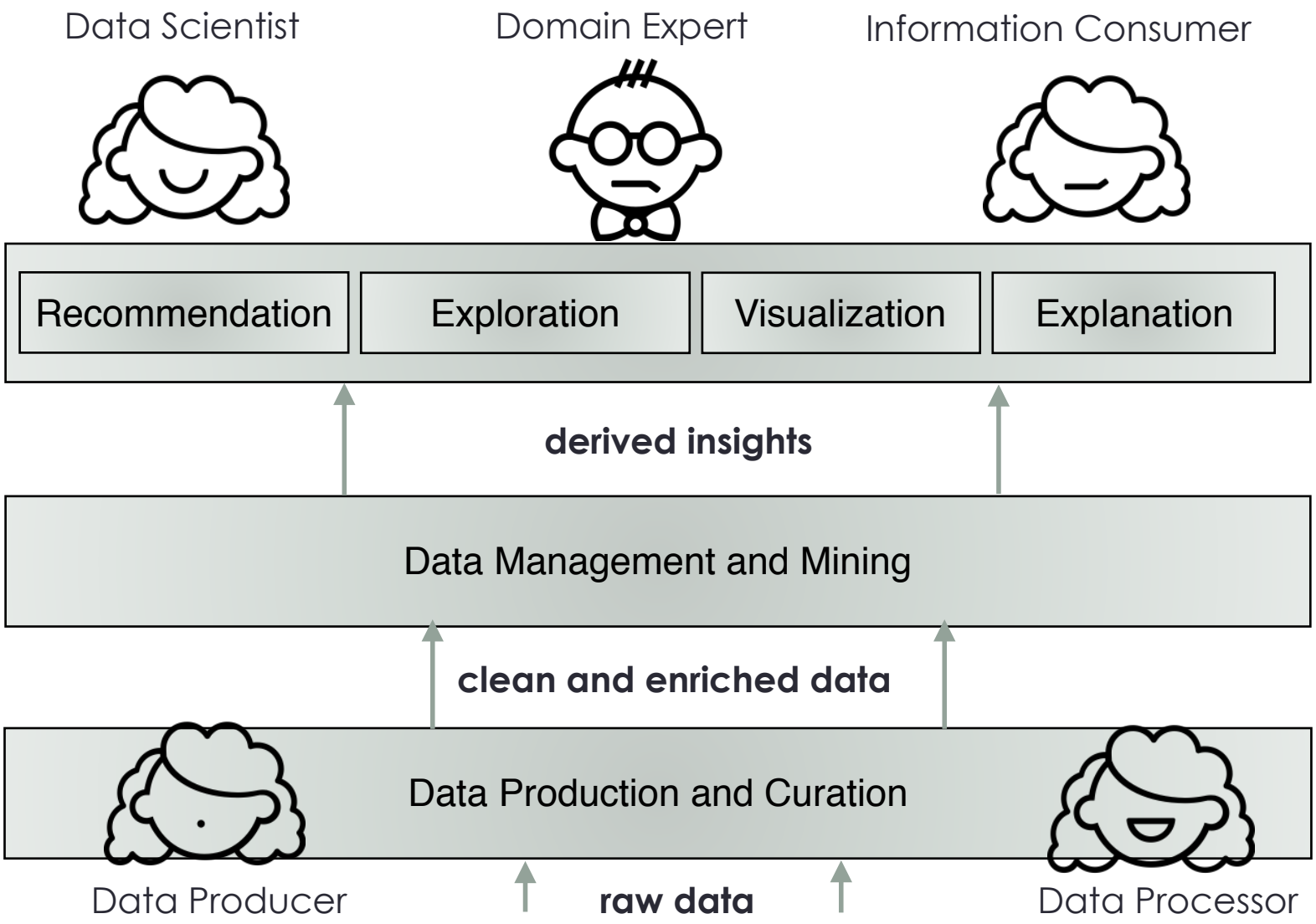
Outsourcing data collection to the crowd

- When people can provide the data
- When people are the only source of data
- When people can efficiently clean and/or organize the data

- Two main aspects [DFKK'12]:
  - Using the crowd to create better databases
  - Using database technologies to create better crowd data sourcing applications

[DFKK'12]: Crowdsourcing Applications and Platforms: A Data Management Perspective, A.Doan, M. J. Franklin, D. Kossmann, T. Kraska, VLDB 2011

Crowd Data Sourcing&Crowdsourcing



Crowdsourcing research groups

An incomplete list of groups working on crowdsourcing:

- Qurk (MIT)
- CrowdDB (Berkeley and ETH Zurich)
- Deco (Stanford and UCSC)
- CrowdForge (CMU)
- HKUST DB Group
- WalmartLabs
- HoloClean (Waterloo)
- MoDaS (Tel Aviv University)
- GACS (U. of Grenoble Alps)

…

Crowdsourcing applications

- Data cleaning/integration (ProPublica)
- Finding missing people (Haiti, Fossett, Gray)
- Translation/Transcription (SpeakerText)
- Word Processing (Soylent)
- Outsourced insurance claim processing
- Data journalism (The Guardian)

Some relevant research

- **Crowdsourced Databases, Query evaluation, Sorts/ joins, Top-K**
  - CrowdDB, Qurk, Deco
- **Crowdsourced Data Collection/Cleaning**
  - AskIt, QOCO,….
- **Crowdsourced Data Mining**
  - CrowdMining, OASSIS, …
- **Cleaning with the crowd**
  - HoloClean (Tamr)
- **Image tagging, media meta-data collection**
- **Crowdsourced recommendations and planning**

(2 HOURS)
- INTRODUCTION TO CROWDSOURCING
- DATABASE QUERY PROCESSING WITH THE CROWD
- TASK DEPLOYMENT STRATEGIES

(1 HOUR)
YOUR ASSIGNMENT

(2 HOURS)
- CROWDSOURCING AND HUMAN FACTORS
- OPEN QUESTIONS

CROWD SOURCED DATABASES

- **Motivation**: Why we need crowdsourced databases?

- There are many things (queries) that cannot be done (answered) with a classical DB approach
- They are called **DB-Hard queries**

- Examples…

DB-Hard Queries (1)

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View CA	$210Bn
Intl. Business Machines	Armonk, NY	$200Bn
Microsoft	Redmond, WA	$250Bn

SELECT Market_Cap **FROM** Companies **WHERE** Company_Name = 'I.B.M'

- Result: 0 rows
- Problem: Entity Resolution

DB-Hard Queries (2)

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View CA	$210Bn
Intl. Business Machines	Armonk, NY	$200Bn
Microsoft	Redmond, WA	$250Bn

```
SELECT Market_Cap FROM Companies WHERE Company_Name =
  'Apple'
```

- Result: 0 rows
- Problem: Closed World Assumption

DB-Hard Queries (3)

SELECT Image **FROM** Images
WHERE Theme = 'Business Success' **ORDER BY** relevance

- Result: 0 rows
- Problem: Missing Intelligence

CrowdDB (Berkeley)

- Use the crowd to answer DB-Hard queries
  - Use the crowd when:
    - Looking for new data (Open World Assumption)
    - Doing a fuzzy comparison
    - Looking to recognize patterns
  - **Don't** use the crowd when:
    - Doing anything the computer already does well

CrowdSQL – Crowd column

- DDL Extension:

```
CREATE TABLE Department (
                    university STRING ,
                    name STRING ,
                    url CROWD STRING ,
                    phone STRING ,
                    PRIMARY KEY ( university , name )
);
```

CrowdSQL – Example #1

INSERT INTO Department (university, name) **VALUES** (“TAU”, “CS”);

- Result:

University	Name	Url	Phone
TAU	CS	**CNULL**	NULL

CrowdSQL – Example #2

```
SELECT url FROM Department WHERE name = "Math";
```

- Side effect of this query:
  - Crowdsourcing of CNULL values of Math departments

CrowdSQL – Crowd table

- DDL Extension:

```
CREATE CROWD TABLE Professor(
                    name STRING PRIMARY KEY ,
                    email STRING UNIQUE ,
                    university STRING ,
                    department STRING ,
                    FOREIGN KEY ( university , department )
                              REF Department ( university , name )
);
```

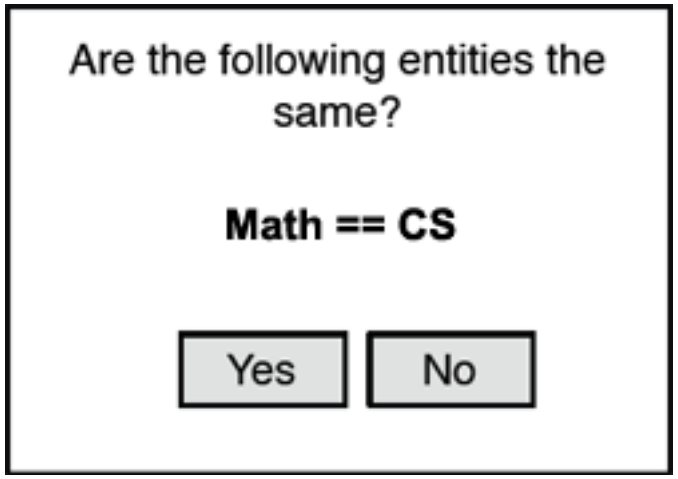
CrowdSQL – Subjective comparisons

- Two functions
  - CROWDEQUAL
    - Takes 2 parameters and asks the crowd to decide if they are equal
    - ~= is a syntactic sugar
  - CROWDORDER
    - Used when we need the help of the crowd to rank or order results

CROWDEQUAL Example

```
SELECT profile FROM department WHERE name ~= "CS";
```

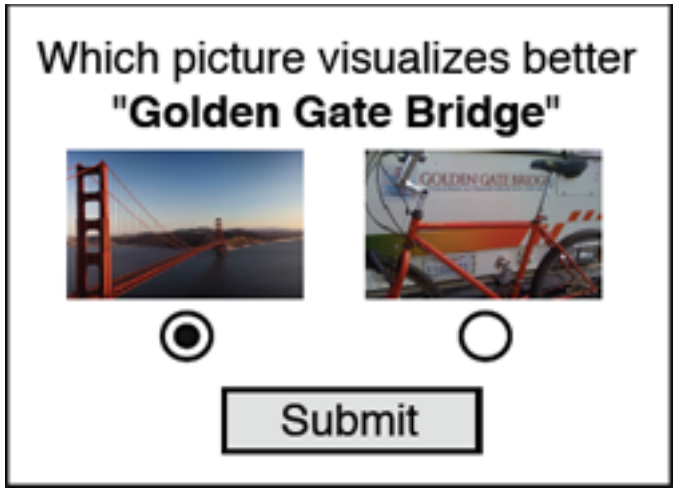
To ask for all "CS" departments, the query asks the crowd to do entity resolution with possibly different names of Computer Science in the database.



CROWDORDER Example

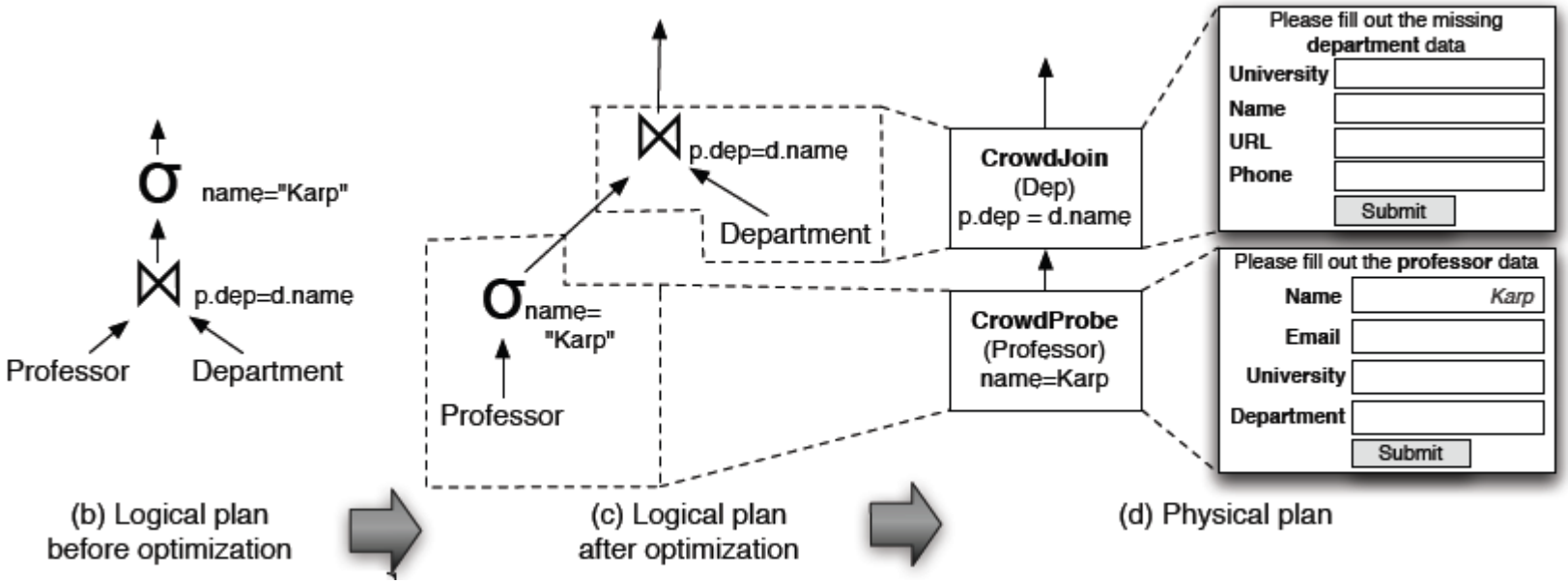
SELECT p **FROM** Picture **WHERE** subject = “Golden Gate Bridge”
ORDER BY CROWDORDER (p, “Which picture visualizes better %subject”);

The CrowdSQL query asks for ranking of pictures with regard to how well they depict the Golden Gate Bridge.



Query plan generation

Query: SELECT * FROM d Professor p, Department d
WHERE d.name = p.dep AND p.name = “Karp”



Note:
Three levels of optimization!

- Global query optimizations
  - Operator reordering
- Operator optimization (we will see this next)
  - What questions should be asked?
- Question optimization
  - How many times each question should be asked?
  - To whom, and in which order?

- **Qurk** (MIT): Declarative workflow management system that allows human computation over data

(human is part of query execution)

QURK

- Schema

  celeb(name **text**, img **url**)

- Query

  SELECT c.name **FROM** celeb **AS** c **WHERE** isFemale(c)

- UDF(User Defined Function) - isFemale:

```
TASK isFemale(field) TYPE Filter:
    Prompt: "<table><tr> \
            <td><img src='%s'></td> \
            <td>Is the person in the image a woman?</td> \
             </tr></table>", tuple[field]
    YesText: "Yes"
    NoText: "No"
    Combiner: MajorityVote
```

isFemale function (UI)

JOIN

- Schema

  celeb(name **text**, img **url**)

  photos(img url)

- Query

  SELECT c.name **FROM** celeb c **JOIN** photos p

  ON samePerson(c.img, p.img)

- samePerson:

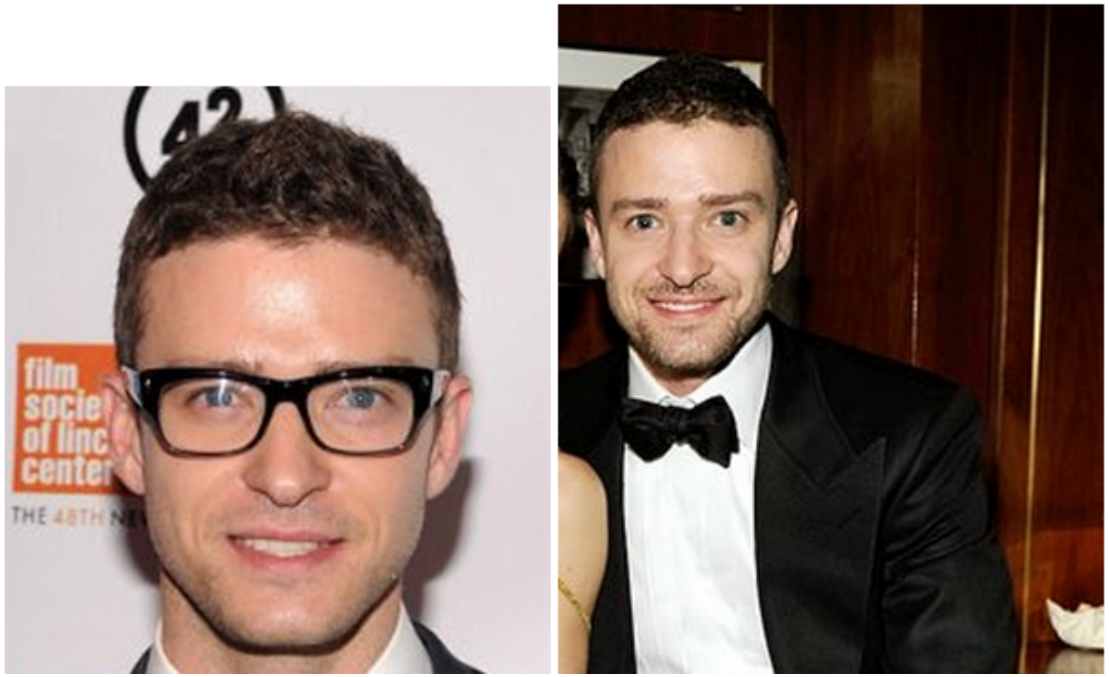
```
TASK samePerson(f1, f2) TYPE EquiJoin:
 SingluarName: "celebrity"
 PluralName: "celebrities"
 LeftPreview: "<img src='%s' class=smImg>",tuple1[f1]
 LeftNormal: "<img src='%s' class=lgImg>",tuple1[f1]
 RightPreview: "<img src='%s' class=smImg>",tuple2[f2]
 RightNormal: "<img src='%s' class=lgImg>",tuple2[f2]
 Combiner: MajorityVote
```

Join – UI example

- # of HITs = |R| * |S|

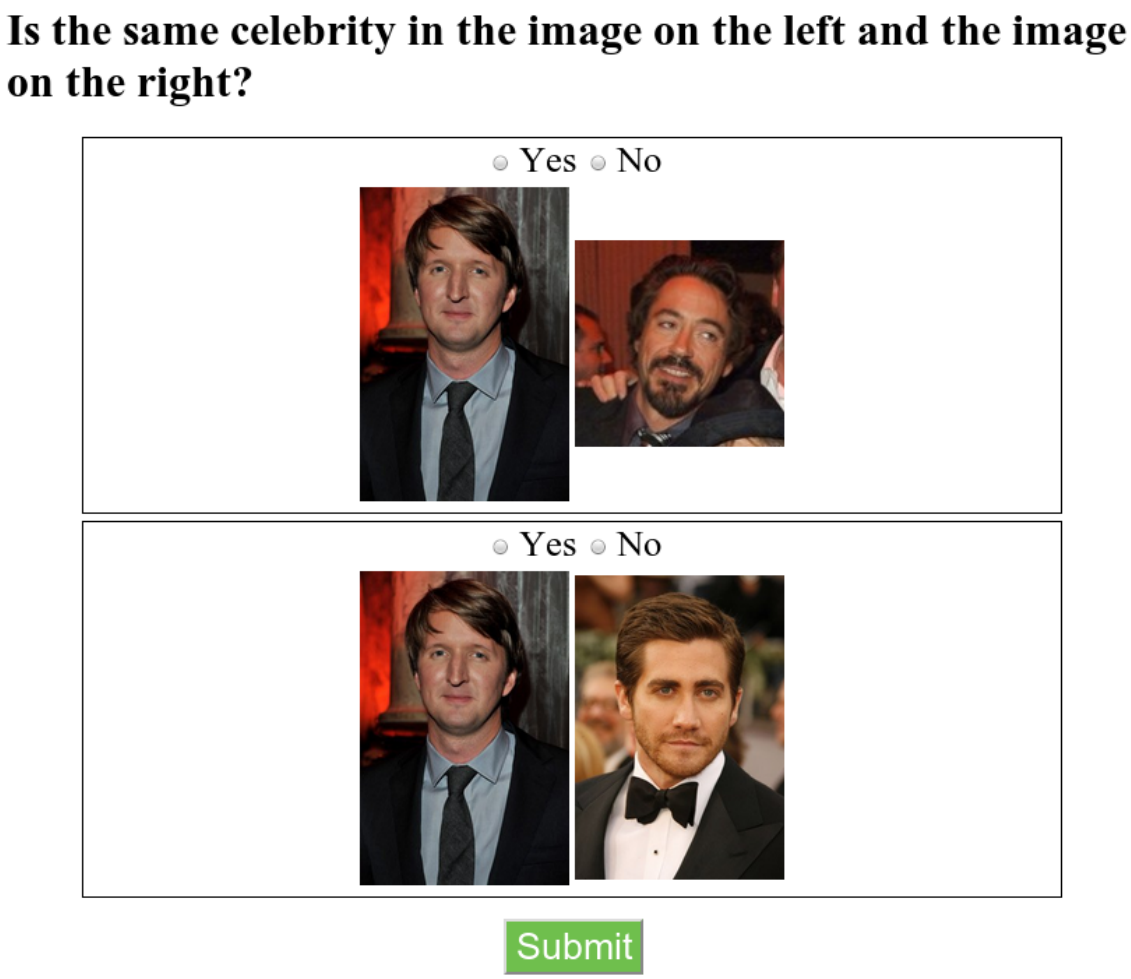
Is the same celebrity in the image on the left and the image on the right?

Yes No



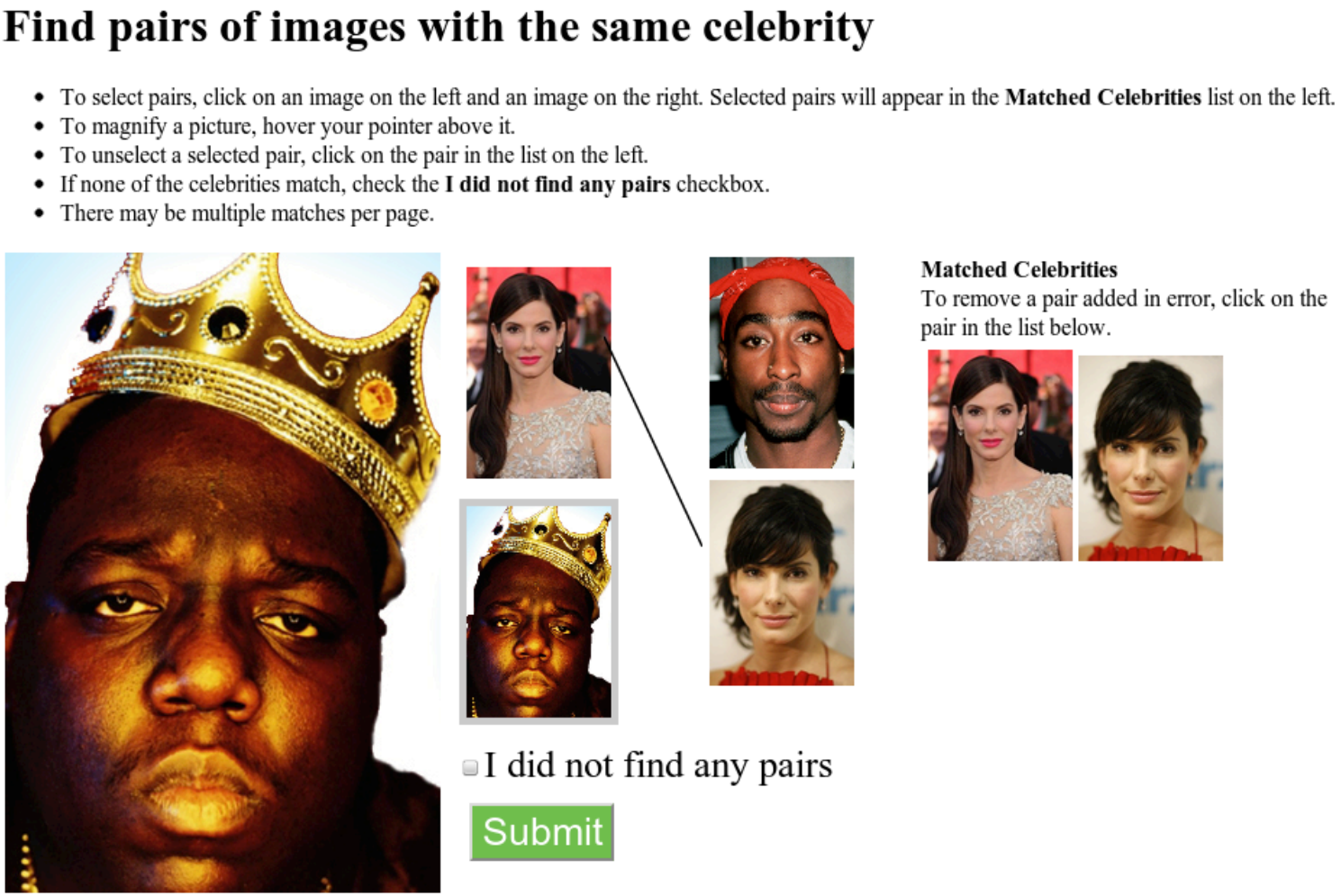
Join – Naïve Batching

- # of HITs = (|R| * |S|) / b



Join – Smart Batching

- # of HITs = (|R| * |S|) / (r * s)



Feature extraction

SELECT c.name **FROM** celeb c **JOIN** photos p
ON samePerson(c.img,p.img)
AND POSSIBLY gender(c.img) = gender(p.img)
AND POSSIBLY hairColor(c.img) = hairColor(p.img)
AND POSSIBLY skinColor(c.img) = skinColor(p.img)

```
TASK gender(field) TYPE Generative:
   Prompt: "<table><tr> \
                <td><img src='%s'> \
                <td>What this person's gender? \
           </table>", tuple[field]
   Response: Radio("Gender",
                ["Male","Female",UNKNOWN])
   Combiner: MajorityVote
```

Economics of feature extraction

- Dataset: Table1 [20 rows] x Table2 [20 rows]
- Join with no filtering (Cross Product): 400 comparisons
- Filtering on 1 parameter (say gender):
  - +40 extra HITS
  - For example: 11 females, 9 males in Table1
  - 10 females, 10 males in Table2
- Join after filtering: ~200 comparisons
- No-Filter/Filter HITs ratio: 400/140
- Decrease the number of HITs ~ **3x**

POSSIBLY filters selection

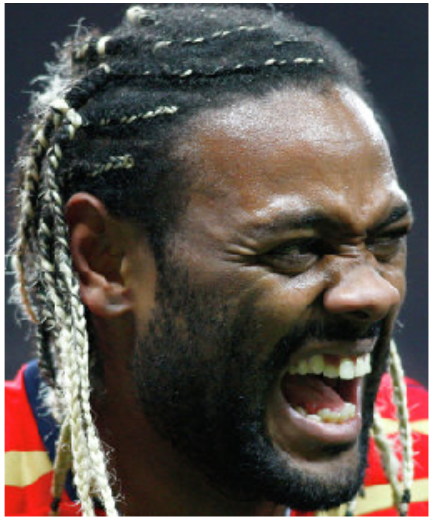
- Gender?

POSSIBLY filters selection

- Skin color?

POSSIBLY filters selection

- Hair color?



QURK – more features: counting with crowd

Given a dataset of images, run queries on it (filtering, aggregation).

Images are unlabeled
No prior knowledge on distribution.

crowd-powered selectivity estimation

50%

1%

Motivating example #1

- Schema

  people (name varchar2(32), photo img)

- Query

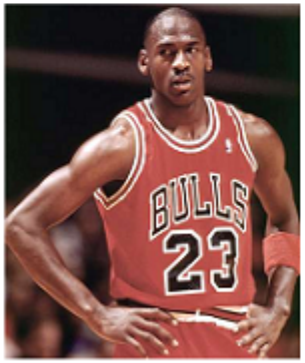
  SELECT * FROM people

  WHERE gender="Male" AND hairColor="red"

Motivating example #1





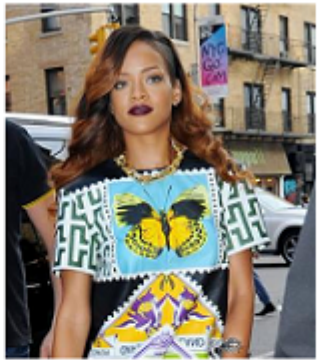






Motivating example #1

- Filter by gender(photo) = ‘male’



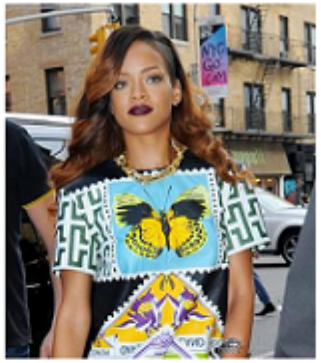



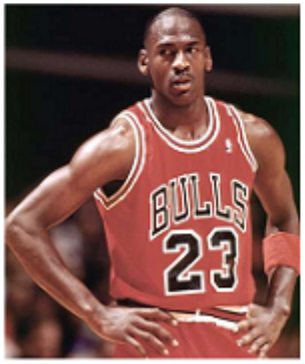



Motivating example #1

- Filter by hairColor(photo) = ‘red’








Motivating example #1

- Filter by gender(photo) = ‘male’ , then by hairColor(photo) = ‘red’
  - First pass: 10 HITs (result – 5 photos)
  - Second pass: 5 HIT
  - Total: **15 HITs**
- Filter by hairColor(photo) = ‘red’, then by gender(photo) = ‘male’
  - First pass: 10 HITs (result – 1 photo)
  - Second pass: 1 HIT
  - Total: **11 HITs**

how many males/females?

interface: labeling

There are 2 people below. Please identify the gender of each.

What is the gender of this person?

○ male ◉ female

What is the gender of this person?

○ male ◉ female

Submit

interface: counting

There are 10 people below. Please provide rough estimates for how many of the people have various properties.

About how many of the 10 people are male? 4

About how many of the 10 people are female?

Submit

Estimating counts

- Can’t show all the images to every user
- Show random sample
  - Sampling error
  - Worker error
  - Dependent samples

Counting vs labeling

- Dataset: 500 images
- Labeling
  - 10 images per HIT (can be 5 – 20)
  - 5 workers per HIT (majority) (can be 3 – 7)
  - Total HITs = 500/10 * 5 = **250**
- Counting
  - 75 images per HIT (can be 50 – 150)
  - 1 worker per HIT (spammer detection algo – later)
  - Total HITs = 500/75 = **7**

- x37.5 times cheaper!

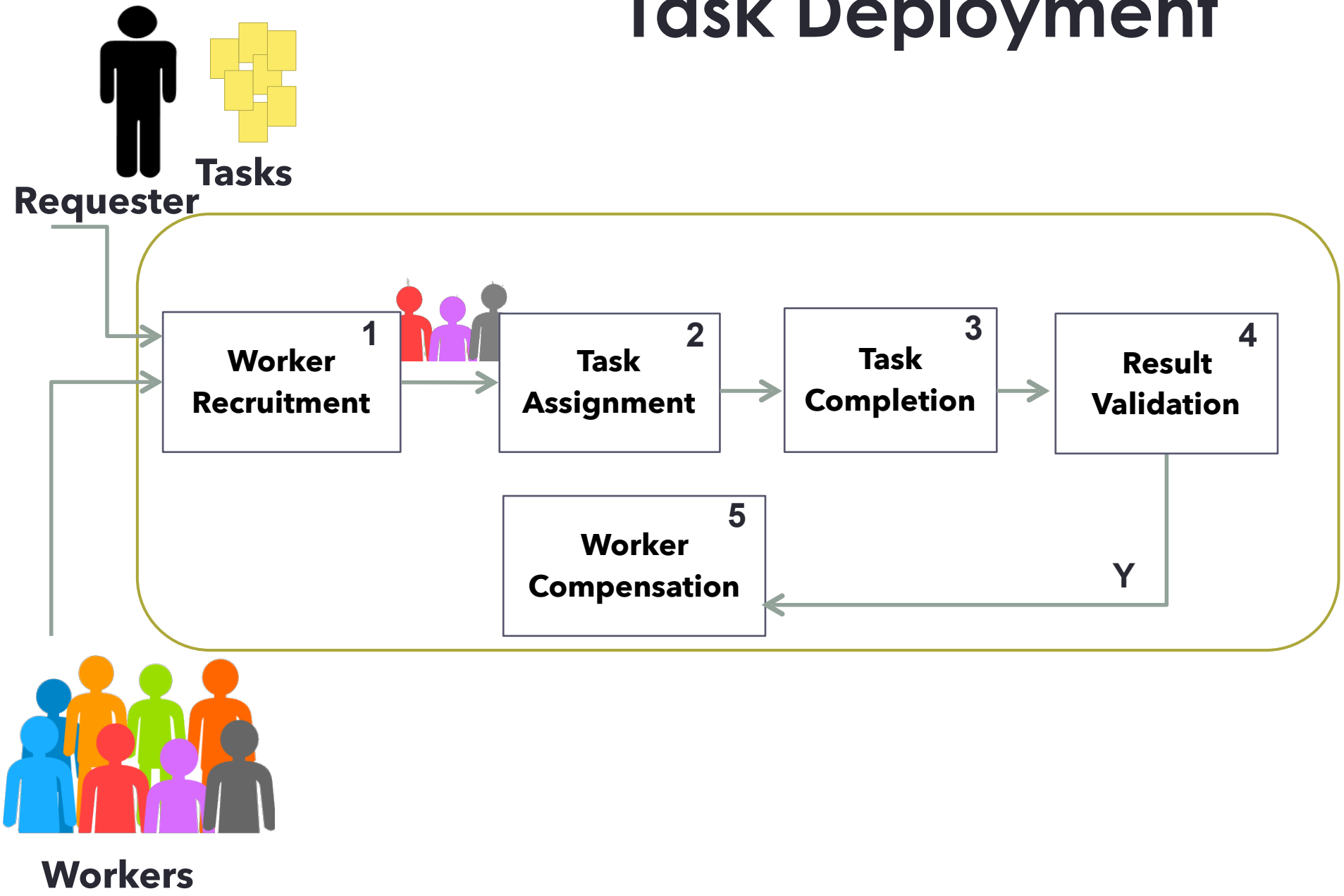
END OF BORROWED MATERIAL
THANK YOU TOVA

(2 HOURS)
- INTRODUCTION TO CROWDSOURCING
- DATABASE QUERY PROCESSING WITH THE CROWD
- TASK DEPLOYMENT STRATEGIES

(1 HOUR)
YOUR ASSIGNMENT

(2 HOURS)
- CROWDSOURCING AND HUMAN FACTORS
- OPEN QUESTIONS

Task Deployment



Task deployment strategies

Translation (Eng > Fr)

"Philosophy is not a bauble of the intellect, but a power from which no man can abstain. Anyone can say that he dispenses with a view of reality, knowledge, the good, but no one can implement this credo….."
- 500 Words

How Many Worker Assignments ?

Reward per Assignment ?

Time Alloted per Assignment ?

What Quality to Expect ?

Should Workers be Organised ?

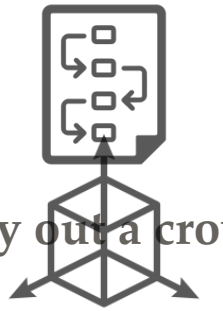
How to Deal With Worker Outputs ?

Expectations with respect to Worker Organisation ?

Ria Mae Borromeo, Thomas Laurent, Motomichi Toyama, Maha Alsayasneh, Sihem Amer-Yahia, Vincent Leroy: Deployment strategies for crowdsourcing text creation. Inf. Syst. 71: 103-110 (2017)

Crowdsourcing Deployment Strategies

Definition



A plan on how to carry out a crowdsourcing task

Work Structure

Sequential vs Simultaneous

Work Style

Hybrid vs Crowd-Only

Work Organisation

Independent vs Collaborative

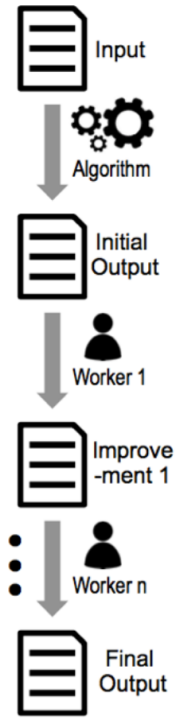
Worker Affinity

Aware vs Unaware

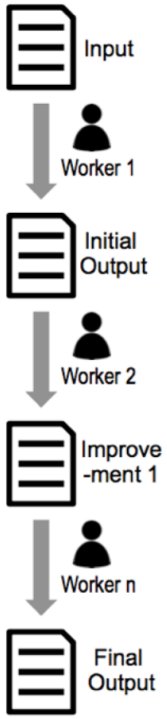
Crowdsourcing Deployment Strategies

Strategies

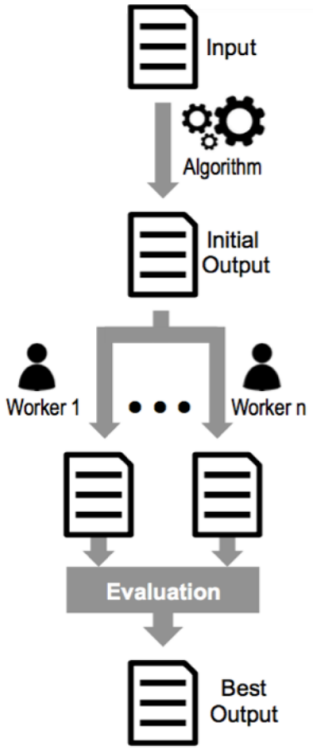
SEQ-IND-HYB



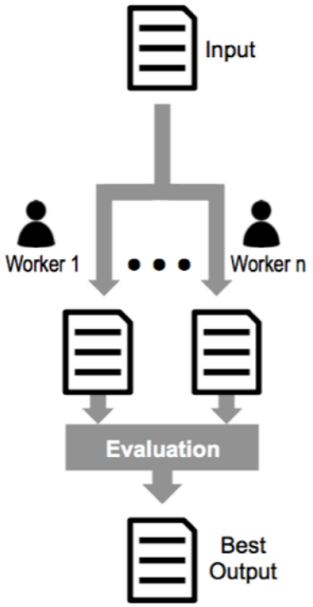
SEQ-IND-CRO



SIM-IND-HYB



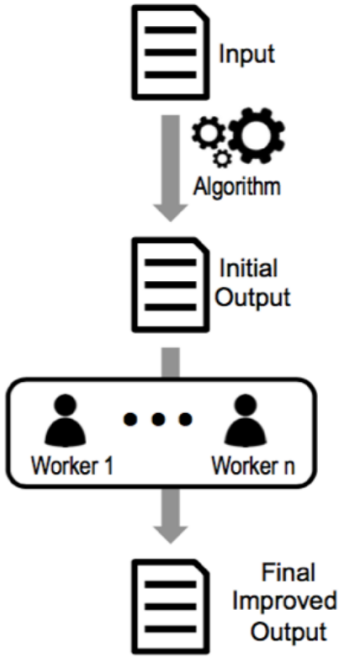
SIM-IND-CRO



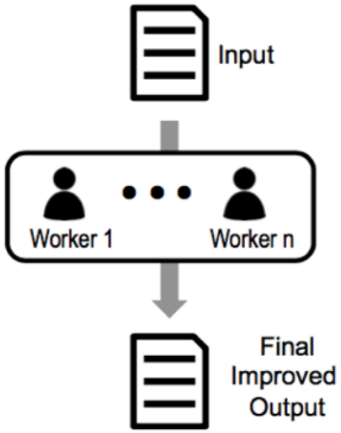
Crowdsourcing Deployment Strategies

Strategies

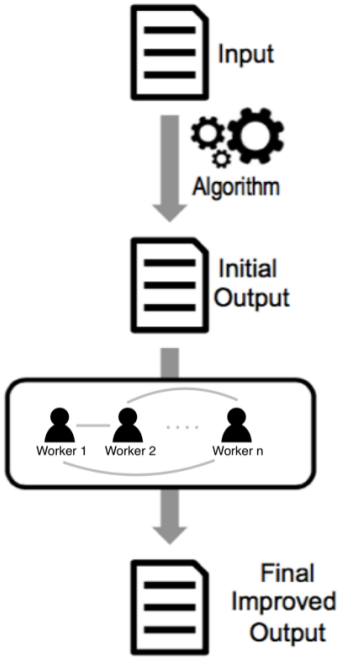
SIM-COL-HYB



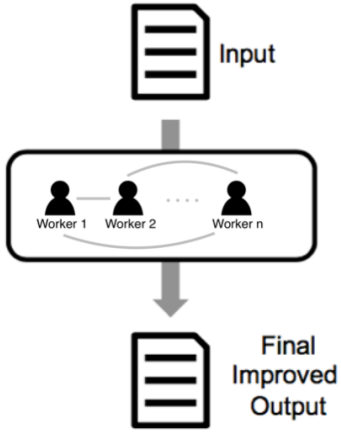
SIM-COL-CRO



SIM-COLAFF-HYB

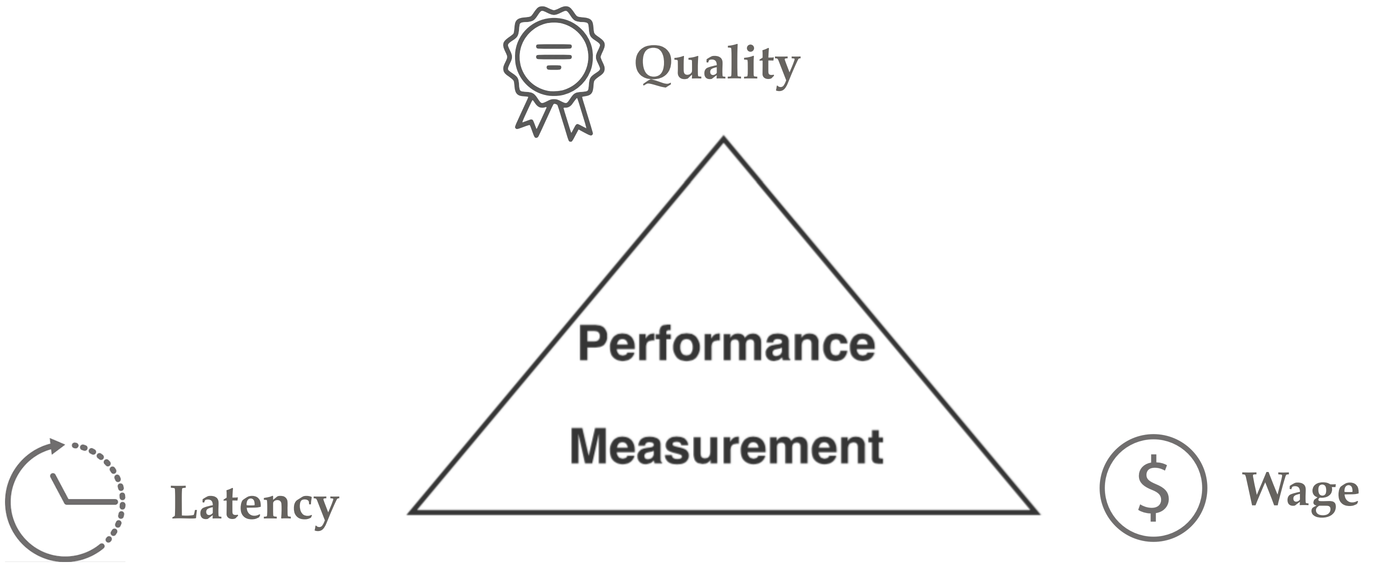


SIM-COLAFF-CRO



Crowdsourcing Deployment Strategies

Performance Dimensions



[illegible]

Possible outcome of optimization

Example

Translation (Eng > Fr)

"Philosophy is not a bauble of the intellect, but a power from which no man can abstain. Anyone can say that he dispenses with a view of reality, knowledge, the good, but no one can implement this credo….."
- 500 Words

For 3 workers with skill values 5, 7 and 8 :

Quality Range : [5, 9]

Latency Range : [15, 30]

Wage Range : [8, 12]

Requester Constraint: Minimal Quality 8

Quality Range : [8, 9]

Latency Range : [20, 30]

Wage Range : [10, 12]

Deployment strategies Desiderata

- Help requesters choose:
  1. **Workforce structure**
     - Sequential vs simultaneous
  2. **Workforce organization**
     - Independent vs collaborative
  3. **Work style**
     - Hybrid vs crowd-only

- **Challenges**
  - **need to estimate performance dimensions for one workers**
  - **for a set of workers**
  - **need to optimize**

Estimating performance dimensions

Strategy				Task	Estimated Quality	Estimated Latency	Estimated Wage
Str	Org	Style	Aff				
SEQ	IND	CRO	-	Creation	$eQuality(w_1,T,t) + \sum_{w\in W\setminus w_1} e^{-\lambda\times eQuality(w,T,t)}$	$eLatency(w_1,T,t) + \sum_{w\in W\setminus w_1} e^{-\lambda\times eLatency(w,T,t)}$	$\sum_{w\in W} eWage(w,T,t)$
				Translation			
				Editing	$max_{w\in W} eQuality(w,T,t)$	$\sum_{w\in W} eLatency(w,T,t)$	
				Labeling			
		HYB		Creation	$machineQuality + \sum_{w\in W} e^{-\lambda\times eQuality(w,T,t)}$	$machineLatency + \sum_{w\in W} e^{-\lambda\times eLatency(w,T,t)}$	$machinePrice + \sum_{w\in W} eWage(w,T,t)$
				Translation			
				Editing	$machineQuality + max_{w\in W} eQuality(w,T,t)$	$machineLatency + \sum_{w\in W} eLatency(w,T,t)$	
				Labeling			
SIM	IND	CRO	-	Creation	$max_{w\in W} eQuality(w,T,t)$	$max_{w\in W} eQuality(w,T,t)$	$\sum_{w\in W} eWage(w,T,t)$
				Translation			
				Editing			
				Labeling			
		HYB		Creation	$machineQuality + max_{w\in W} eQuality(w,T,t)$	$machineLatency + max_{w\in W} eQuality(w,T,t)$	$machinePrice + \sum_{w\in W} eWage(w,T,t)$
				Translation			
				Editing			
				Labeling			
	COL	CRO	Unaware	Creation	$\prod_{w\in W} eQuality(w,T,t)$	$e^{\lambda\times\sum_{w\in W} eLatency(w,T,t)}$	$\sum_{w\in W} eWage(w,T,t)$
				Translation			
				Editing			
				Labeling	10		
			Aware	Creation	$\prod_{w_i,wj\in W} Aff(w_i,w_j)\times mean(eQuality(w_i,T,t), eQuality(w_j,T,t)))$		
				Translation			
				Editing			
				Labeling	10		
		HYB	Unaware	Creation	$machineQuality + \prod_{w\in W} eQuality(w,T,t)$		
				Translation			
SIM		HYB				$machineLatency + \sum_{w\in W} e^{\lambda * eLatency(w,T,t)}$	$machinePrice + \sum_{w\in W} eWage(w,T,t)$

Deployment strategies Research opportunity for DB

- Help requesters choose:
  1. **Workforce structure**
     - Sequential vs simultaneous
  2. **Workforce organization**
     - Independent vs collaborative
  3. **Work style**
     - Hybrid vs crowd-only

- **(Multi-criteria?) optimization formulation** to account for *wage, quality, latency, worker retention…*

- For different task types: e.g., idea generation, algorithm design, puzzle solving

(2 HOURS)
- INTRODUCTION TO CROWDSOURCING
- DATABASE QUERY PROCESSING WITH THE CROWD
- TASK DEPLOYMENT STRATEGIES

(1 HOUR)
YOUR ASSIGNMENT

(2 HOURS)
- CROWDSOURCING AND HUMAN FACTORS
- OPEN QUESTIONS

ASSIGNMENT

1. form groups

2. each group chooses a paper to summarize in a Google doc (paper list to be provided)