DATA-CENTERED CROWDSOURCING WORKSHOP

PROF. TOVA MILO

SLAVA NOVGORODOV

TEL AVIV UNIVERSITY 2014/2015

ADMINISTRATIVE NOTES

Course goal:

Learn about crowd-data sourcing and prepare a final project (improvement of existing problem's solution / solving new problem)

Group size: ~4 students

Requirements: DataBases (SQL) is recommended, Web programming (we will do a short overview), (optionally) Mobile development (we will not teach it)

ADMINISTRATIVE NOTES (2)

Schedule:

- **3 intro meetings**
  - 1st meeting – overview of crowdsourcing
  - 2nd meeting – open problems, possible projects
  - 3rd meeting – Web programming overview, SDK
- **Mid-term meeting**
- **Final meeting and projects presentation**
- **Dates:** http://slavanov.com/teaching/crowd1415b/

WHAT IS CROWDSOURCING?

Crowdsourcing = Crowd + Outsourcing

Crowdsourcing is the act of sourcing tasks traditionally performed by specific individuals to a group of people or community (crowd) through an open call.

CROWDSOURCING

Main idea: Harness the crowd to a "task"

- Task: solve bugs
- Task: find an appropriate treatment to an illness
- Task: construct a database of facts

…

Why now?

- Internet and smart phones …

  We are all connected, all of the time!!!

THE CLASSICAL EXAMPLE

WIKIPEDIA

English
The Free Encyclopedia
3 907 000+ articles

日本語
フリー百科事典
799 000+ 記事

Español
La enciclopedia libre
879 000+ artículos

Deutsch
Die freie Enzyklopädie
1 383 000+ Artikel

Русский
Свободная энциклопедия
838 000+ статей

Français
L'encyclopédie libre
1 230 000+ articles

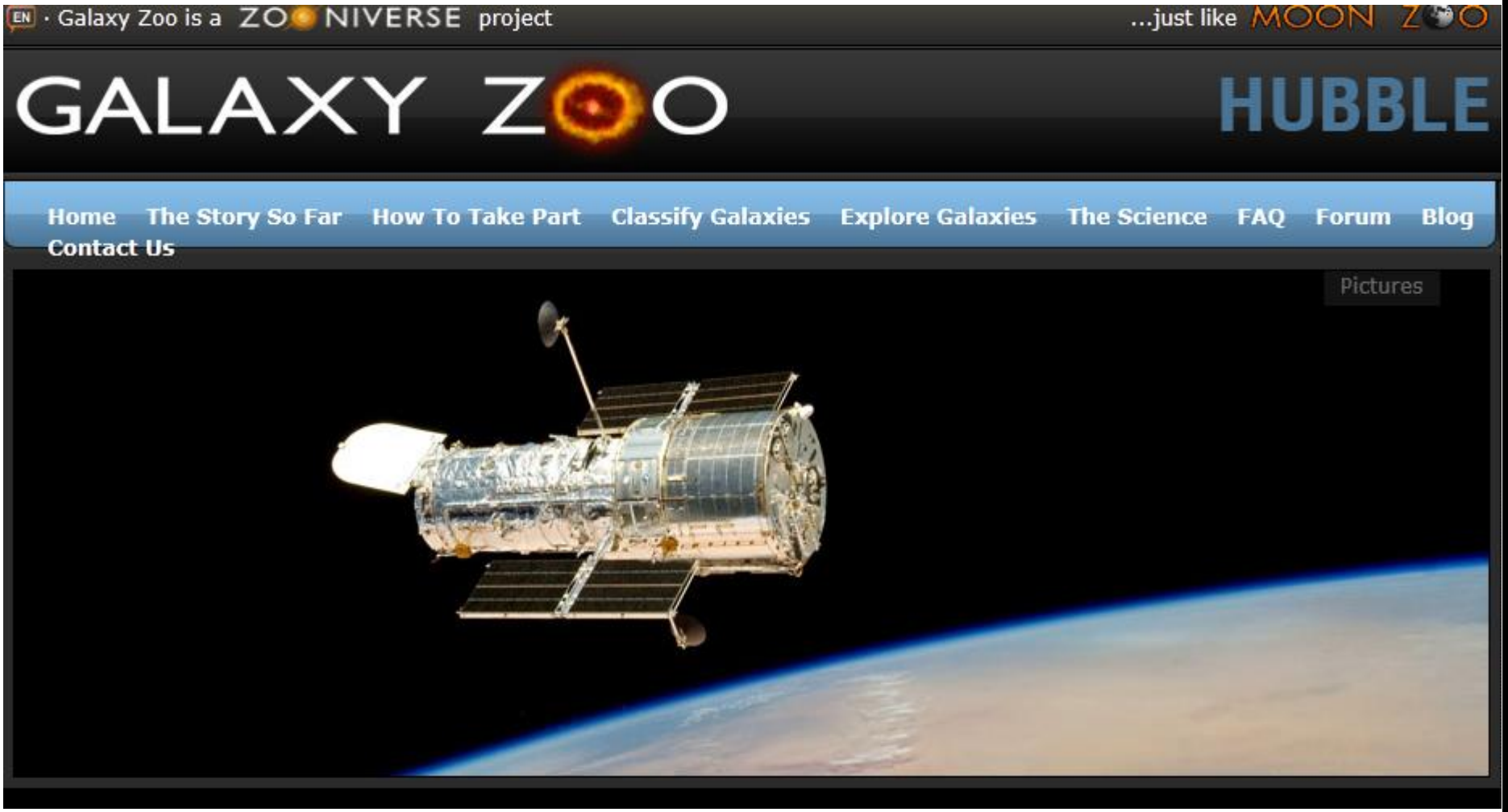
Italiano
L'enciclopedia libera
905 000+ voci

Polski
Wolna encyklopedia
887 000+ haseł

Português
A enciclopédia livre
718 000+ artigos

中文
自由的百科全書
429 000+ 條目

GALAXY ZOO


· Galaxy Zoo is a ZOONIVERSE project
...just like MOON ZOO
GALAXY ZOO
HUBBLE
Home
The Story So Far
How To Take Part
Classify Galaxies
Explore Galaxies
The Science
FAQ
Forum
Blog
Contact Us
Pictures

MORE


fold it
Solve Puzzles for Science


Dr. David Baker
Several sheets are lining up in this protein, but one is out of place. Pull the misaligned sheet back in to form hydrogen bonds! Don't forget you can control-click to lock, and use Shake and Wiggle.
Repeat Introduction
Clear Labels
Progress: 8899 of 10000
Level 4-2: A Sheet Out of Place
Chat
Shake Sidechains
Wiggle Backbone
Clear Locks
Reset Puzzle
Actions
History
File
Pull Tool

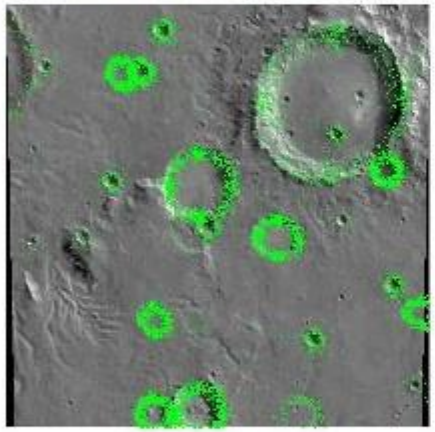
AND EVEN MORE

Mars Clickworkers

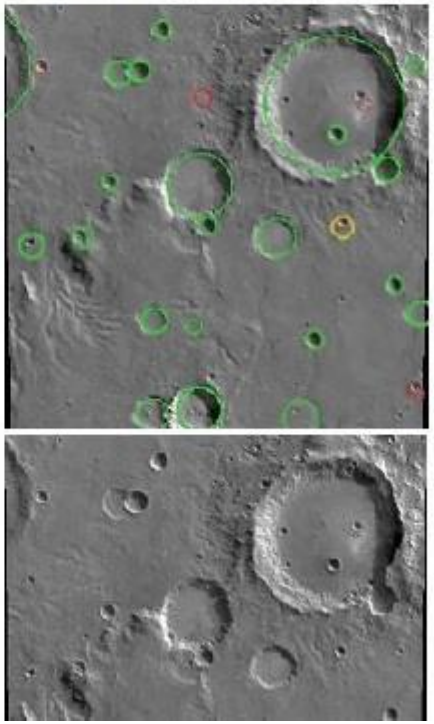
Sample result

We try to have several people cover each region on Mars so that we can compute a consensus, throwing out any mistaken or frivilous entries and averaging out the inaccuracies. Here's an example of one region selected as a test case for intensive coverage:

Here are all the clicks we received for this region. (The red ones were rejected for non-circularity or size.)



Here is the consensus. Green circles have a high degree of agreement. Red circles have little agreement, and yellow ones are marginal.



It may be hard to see some of the fainter craters in the marked-up images. Here's the image with nothing in the way:

(http://clickworkers.arc.nasa.gov/sample-results)

CROWDSOURCING: UNIFYING PRINCIPLES

Main goal

- “Outsourcing” a task to a crowd of users

Kinds of tasks

- Tasks that can be performed by a computer, but inefficiently
- Tasks that can’t be performed by a computer

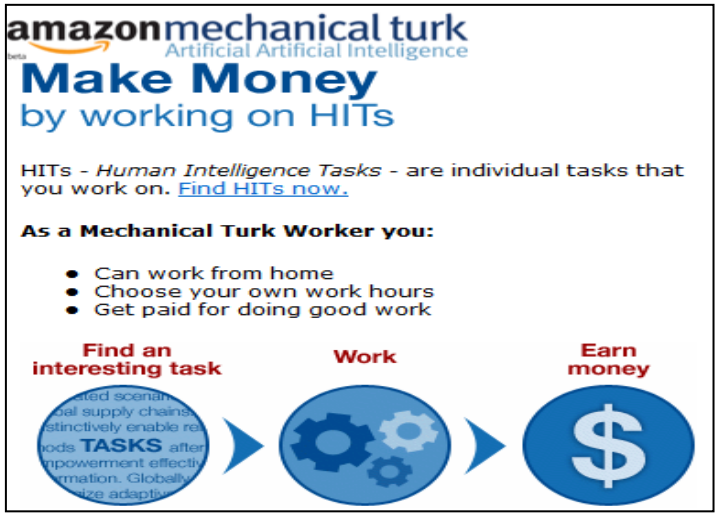
Challenges

- How to motivate the crowd?
- Get data, minimize errors, estimate quality
- Direct users to contribute where is most needed \ they are experts

MOTIVATING THE CROWD

Altruism

Fun



Money

CROWD DATA SOURCING

Outsourcing data collection to the crowd of Web users

- 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 datasourcing applications

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

MY FAVORITE EXAMPLE

ReCaptha

- 100,000 web sites
- 40 million words/day

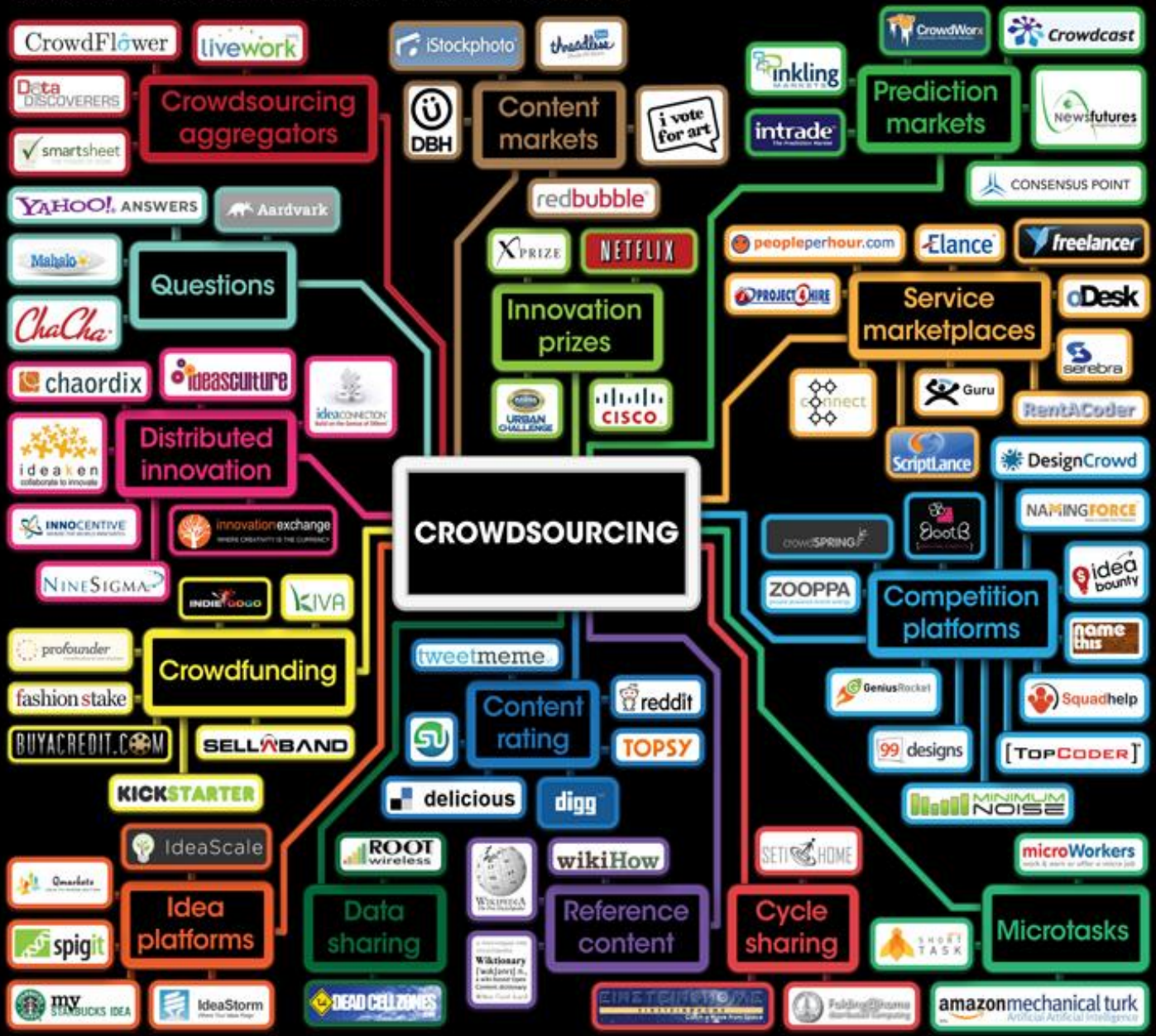
CROWDSOURCING RESEARCH GROUPS

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**
- **MoDaS (Tel Aviv University)**

…

CROWDSOURCING LANDSCAPE Beta v1



Common Crowdsourcing Tasks and Examples

- 3D object design — Thingiverse
- Advertising — idea bounty
- Business ideas
- Clothing — threadless
- Consumer research
- Crisis information — Ushahidi
- Data analysis — SETI@HOME
- Fact checking — PolitiFact.com
- Graphic design — 99designs
- Human reading — reCAPTCHA
- Investigative reporting — theguardian
- Journalism — seed
- Lending
- Mapping
- Movie reviews — IMDb
- Music — muzicpitch
- Observation — GALAXY ZOO
- Patent research
- Philanthropy
- Political activism — MoveOn.org
- Product design — MUJI
- Proofreading
- Scientific problems — fold.it
- Software
- Software development — RentACoder
- Software testing — uTest
- Stock picking
- Tagging — Google
- Translation — facebook
- Trends
- TV programming — current
- Word of mouth — BzzAgent
- Writing and editing

For details, analysis, and discussion go to:
www.crowdsourcingresults.com

Advanced Human Technologies

Created by Ross Dawson
Design by Daniil Alexandrov

crowdsourcing.org
The Industry Website

May 2011 - www.crowdsourcing.org

Crowdsourcing Industry Landscape

Crowdfunding

Financial contributions from online investors, sponsors or donors to fund for-profit or non-profit initiatives or enterprises.

Collective Knowledge

Development of knowledge assets or information resources from a distributed pool of contributors.

Tools

Applications, platforms and tools that support collaboration, communication and sharing among distributed groups of people.

Collective Creativity

Tapping of creative talent pools to design and develop original art, media or content.

Cloud Labor

Leveraging of a distributed virtual labor pool, available on-demand to fulfill a range of tasks from simple to complex.

Community Building

Development of communities through active engagement of individuals who share common passions, beliefs or interests.

Civic Engagement

Collective actions that address issues of public concern.

Open Innovation

Use of sources outside of the entity or group to generate, develop and implement ideas.

CROWDSOURCING MARKETPLACES

CrowdFlower

clickworker

amazon mechanical turk
beta
Artificial Artificial Intelligence

oDesk
Love the way you work.

samasource

DesignCrowd

mobileworks

microWorkers
work & earn or offer a micro job

MECHANICAL TURK

Please Copy Text from Business Card:

Please **select/crop** company logo or image from the business card above.
Click + Drag to select the company logo.

Your Current Quality Score is: --

If you have a high enough score, you will be considered for promotion to a Trusted Worker. ?

Name ?

Title | Company

Email | Website

Address: ?

Address Line 1

add line

City | State | Zip Code

Phone: click here if not a U.S. phone number ?

Work | Ext.

Mobile

Fax

add phone

MECHANICAL TURK

Requestor places Human Intelligence Tasks (HIT)

- Min price: $0,01
- Provide expiration date and UI
- # of assignments

Requestor approve jobs and payments

- Special API

Workers choose jobs, do them and getting money

USES OF HUMAN COMPUTATION

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

TYPES OF TASKS

Task Type	*Characteristics*	*Examples*
Micro Tasks	Generally very high volume, extremely low pay rate per task, and heavily automated throughout the process.	- find email addresses or company websites - translate a product description to another language - find prices for competitive products
Macro Tasks	Generally high volume, low rate of pay, mostly automated throughout the process.	- choose a category from a new catalog structure - write a product review - test this website and provide feedback - fill in the missing research citations in this report - build a list of universities conducting energy research
Simple Projects	Low volume or single tasks with moderate rate of pay, often requiring some direct contact with the worker.	- design a branded website - prepare an outline for a conference presentation - contact all confirmed attendees for an event
Complex Projects	Single project with high rate of pay, typically requiring a substantial amount of time and direct interaction with the worker.	- program a software module - design a new edible adhesive - develop a new security algorithm - develop an eCommerce website - inbound/outbound calls (sales, mrkt research, support)

Source: “Paid Crowdsourcing”, SmartSheet.com

OVERVIEW OF RECENT RESEARCH

- **Crowdsourced Databases, Query evaluation, Sorts/joins, Top-K**
  - CrowdDB, Qurk, Deco,
- **Crowdsourced Data Collection/Cleaning**
  - AskIt, QOCO,….
- **Crowd sourced Data Mining**
  - CrowdMining, OASSIS, …
- **Image tagging, media meta-data collection**
- **Crowdsourced recommendations and planning**

CROWD SOURCED DATABASES

- **Motivation:**

  Why we need crowdsourced databases?

- **There are many things (queries) that cannot be done (answered) in classical DB approach**
- **We call them: 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

Use the crowd to answer DB-Hard queries

- Use the crowd when:
  - Looking for new data (Open World Assumption)
  - Doing a fuzzy comparison
  - Recognize patterns

- **Don't** use the crowd when:
  - Doing anything the computer already does well

CLOSED WORLD VS OPEN WORLD

OWA

- Used in Knowledge representation

CWA

- Used in classical DBMS

Example:

- Statement: Marry is citizen of France
- Question: Is Paul citizen of France?
  - CWA: No
  - OWA: **Unknown**

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

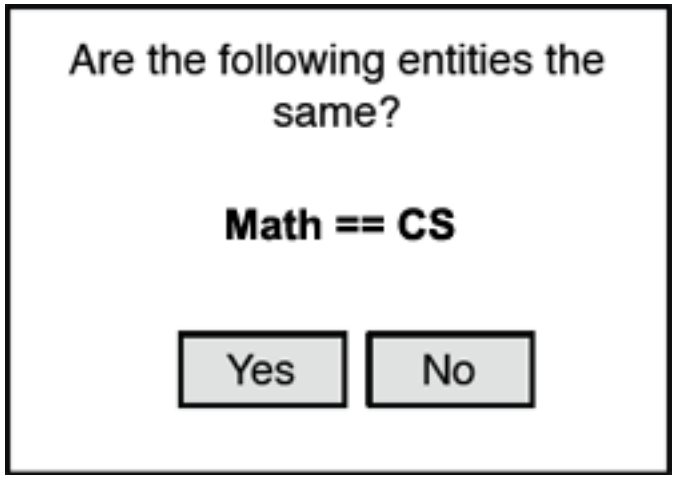
Two functions

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

CROWDEQUAL EXAMPLE

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

To ask for all "CS" departments, the following query could be posed. Here, the query writer asks the crowd to do entity resolution with the possibly different names given for Computer Science in the database.

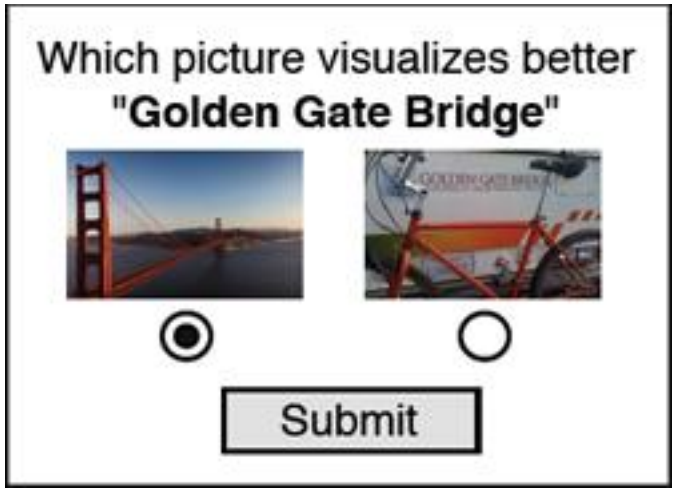


CROWDORDER EXAMPLE

SELECT p FROM Picture WHERE subject = “Golden Gate Bridge”

ORDER BY CROWDORDER (p, “Which picture visualizes better %subject”);

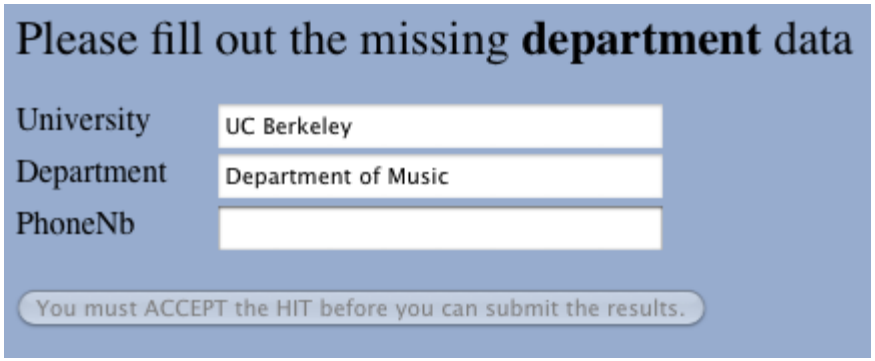
The following CrowdSQL query asks for a ranking of pictures with regard to how well these pictures depict the Golden Gate Bridge.



UI GENERATION

Clear UI is key to quality of answers and response time

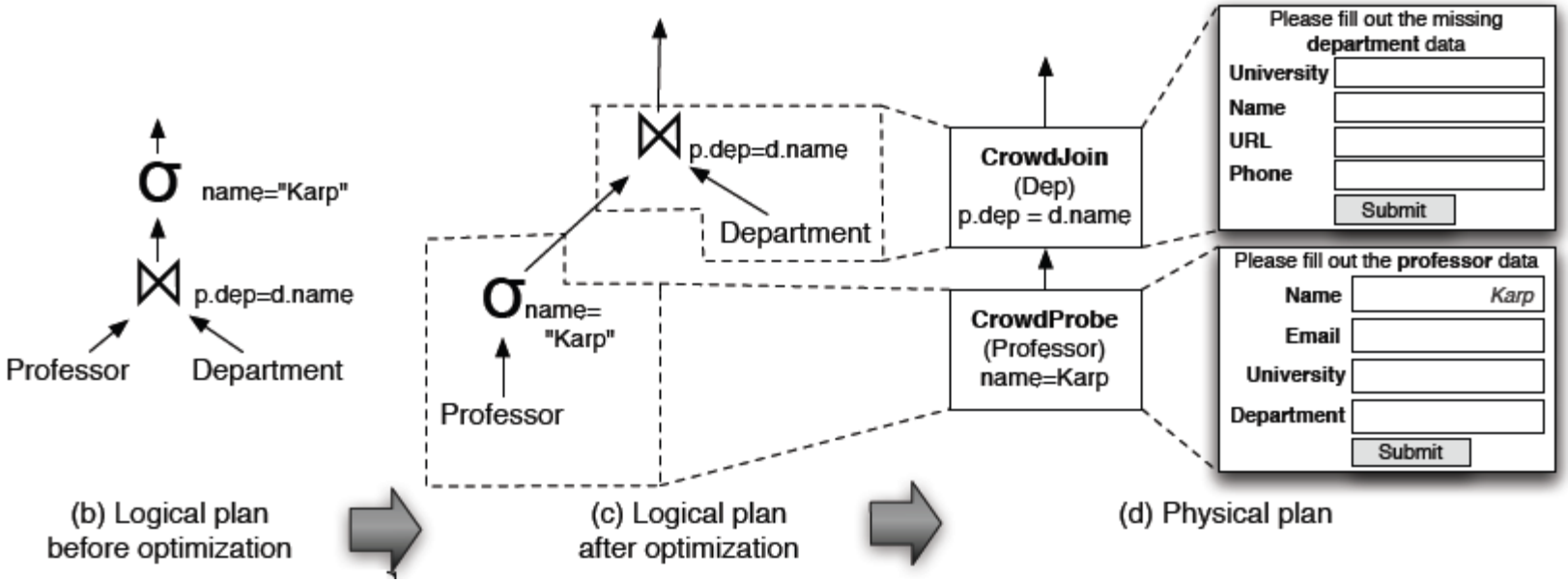
SQL Schema to auto-generated UI



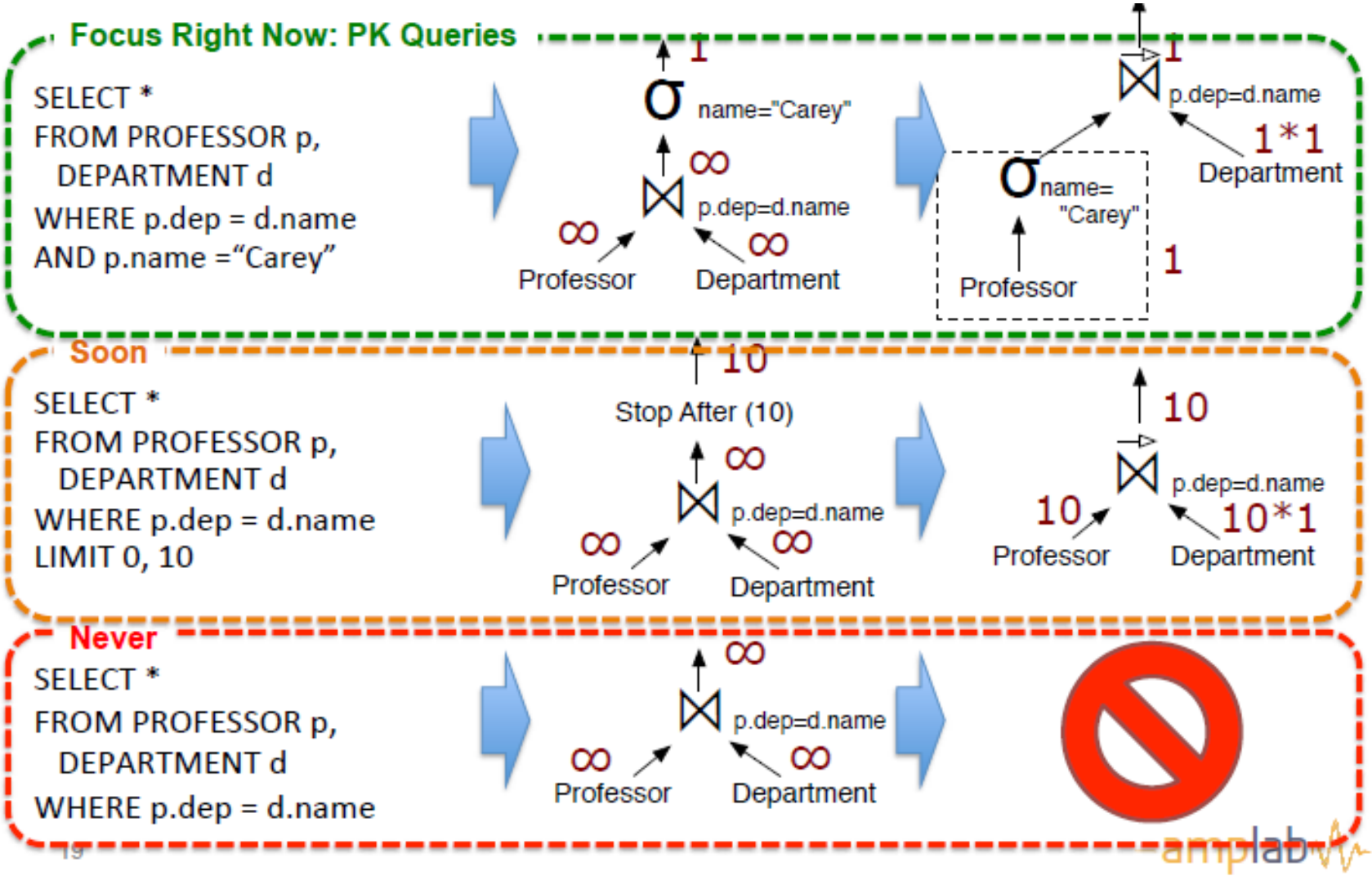
QUERY PLAN GENERATION

Query: SELECT * FROM d Professor p, Department d

WHERE d.name = p.dep AND p.name = “Karp”



DEALING WITH OPEN-WORLD



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

(human is a part of query execution)

QURK: THE BEGINNING

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

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

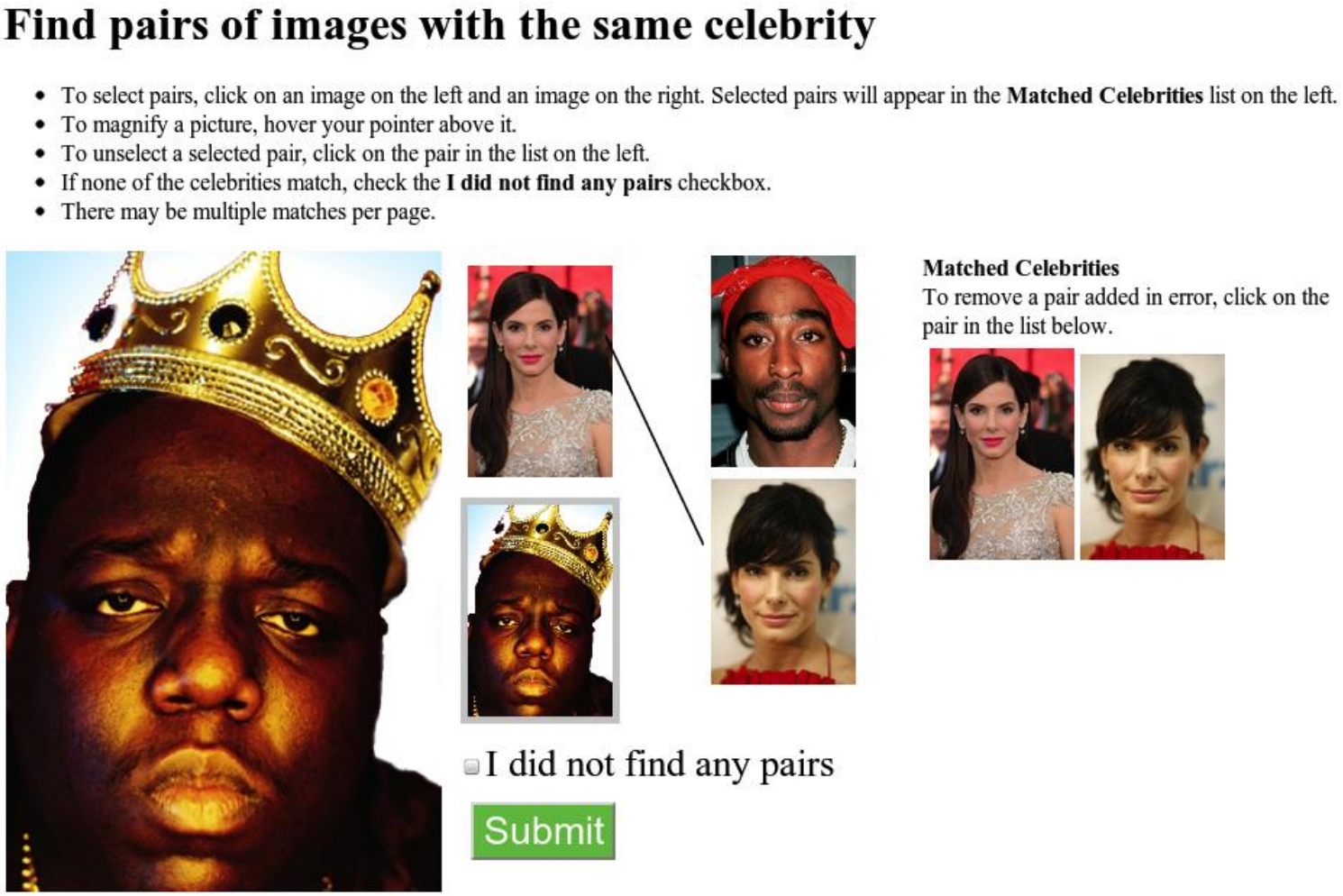
JOIN – NAÏVE BATCHING

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

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

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

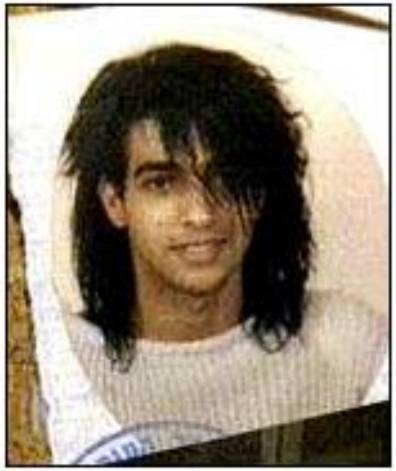
Join after filtering: ~100 comparisons

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)

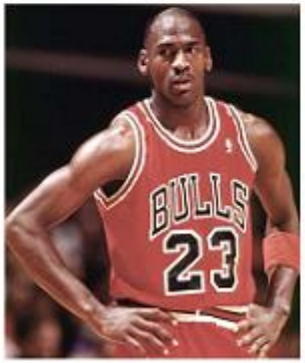
Query

```
SELECT * FROM people
WHERE gender="Male" AND hairColor="red"
```

MOTIVATING EXAMPLE #1





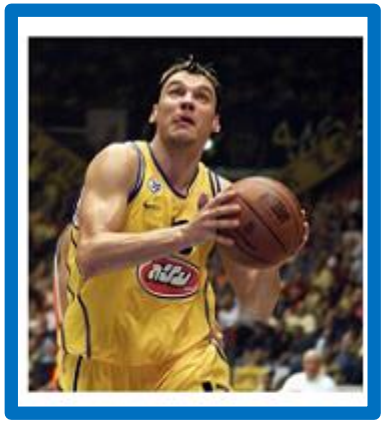
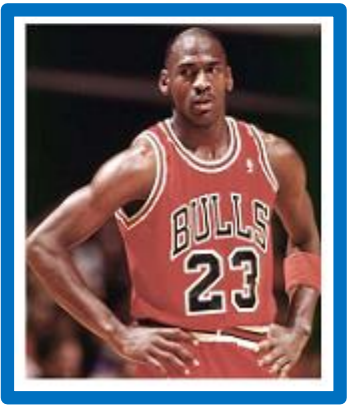



MOTIVATING EXAMPLE #1

Filter by gender(photo) = ‘male’





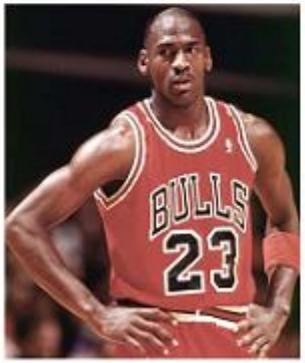



MOTIVATING EXAMPLE #1

Filter by hairColor(photo) = ‘red’








MOTIVATING EXAMPLE #1

- **Filter by gender(photo) = ‘male’ , the 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?

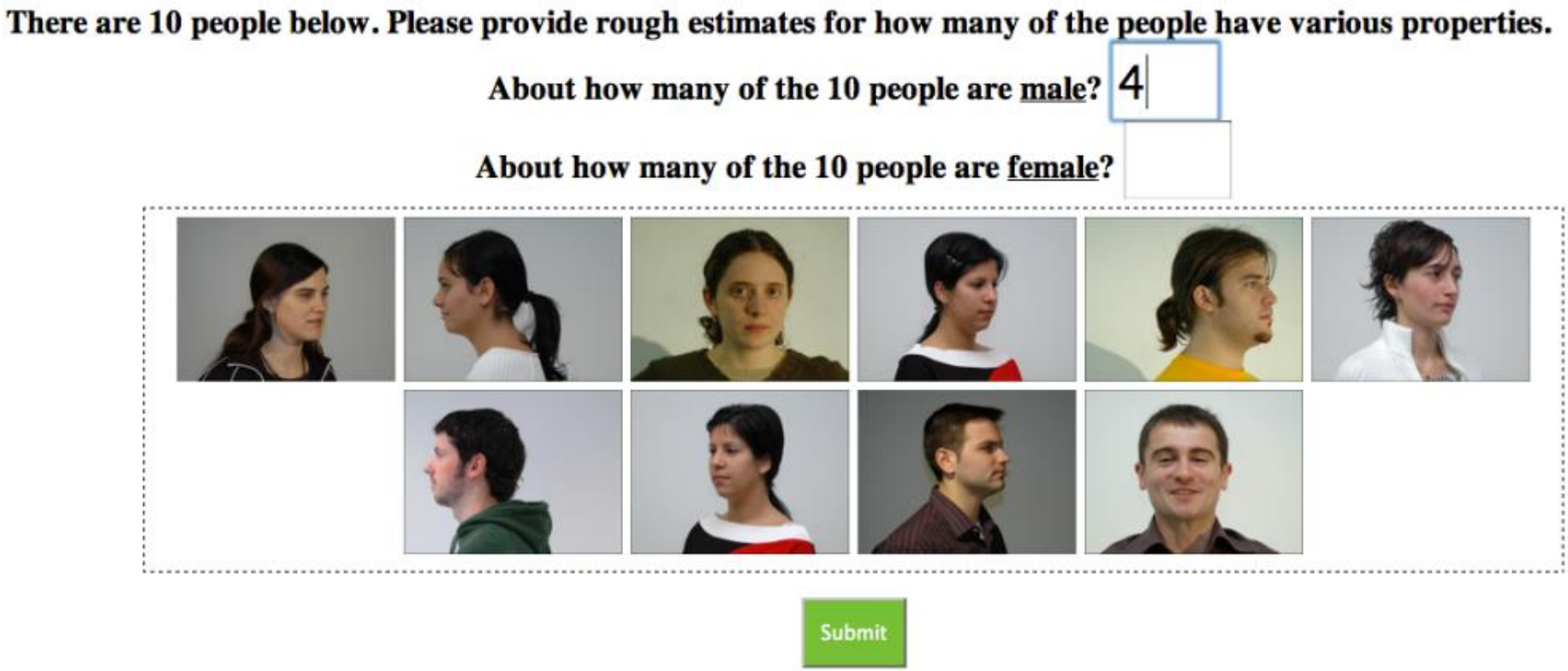
○ male ◉ female

What is the gender of this person?

○ male ◉ female

Submit

INTERFACE: COUNTING



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

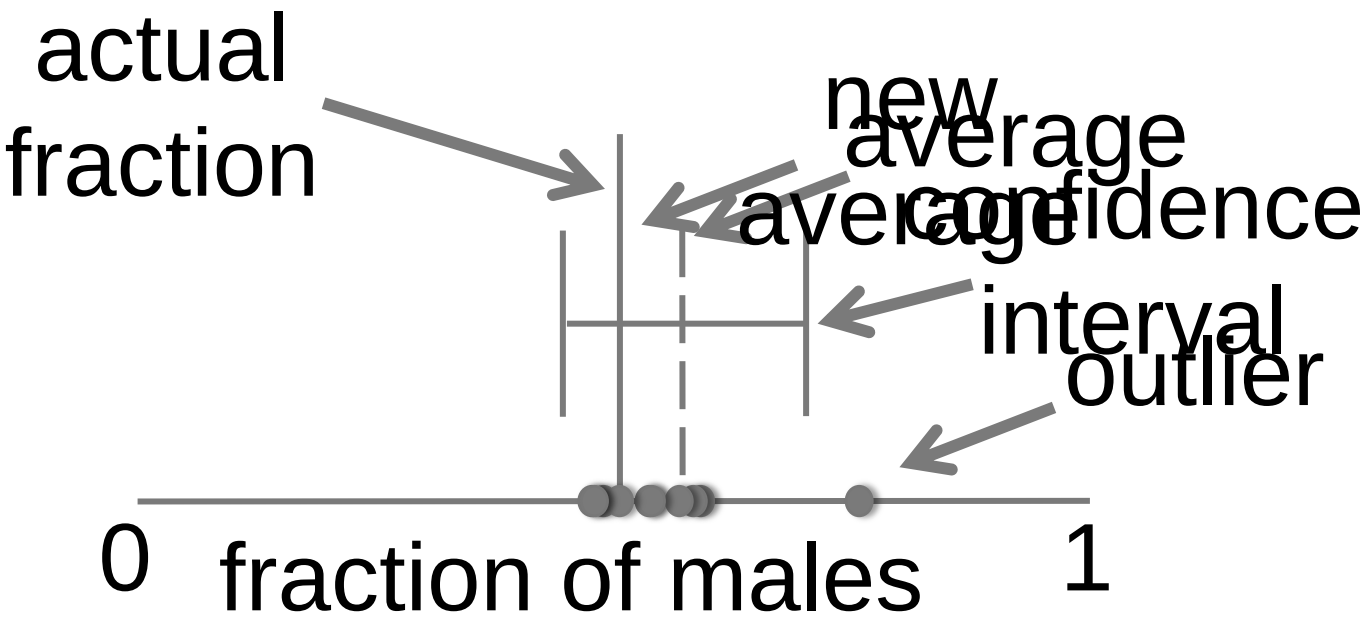
- **Labeling**
  - 10 images per HIT (can be 5 – 20)
  - 5 workers per HIT (majority) (can be 3 – 7)
  - Total HITs = 500/10 * 5 = **250**
- **Counting**
  - 75 images per HIT (can be 50 – 150)
  - 1 worker per HIT (spammer detection algo – later)
  - Total HITs = 500/75 = **7**

x37.5 times cheaper!

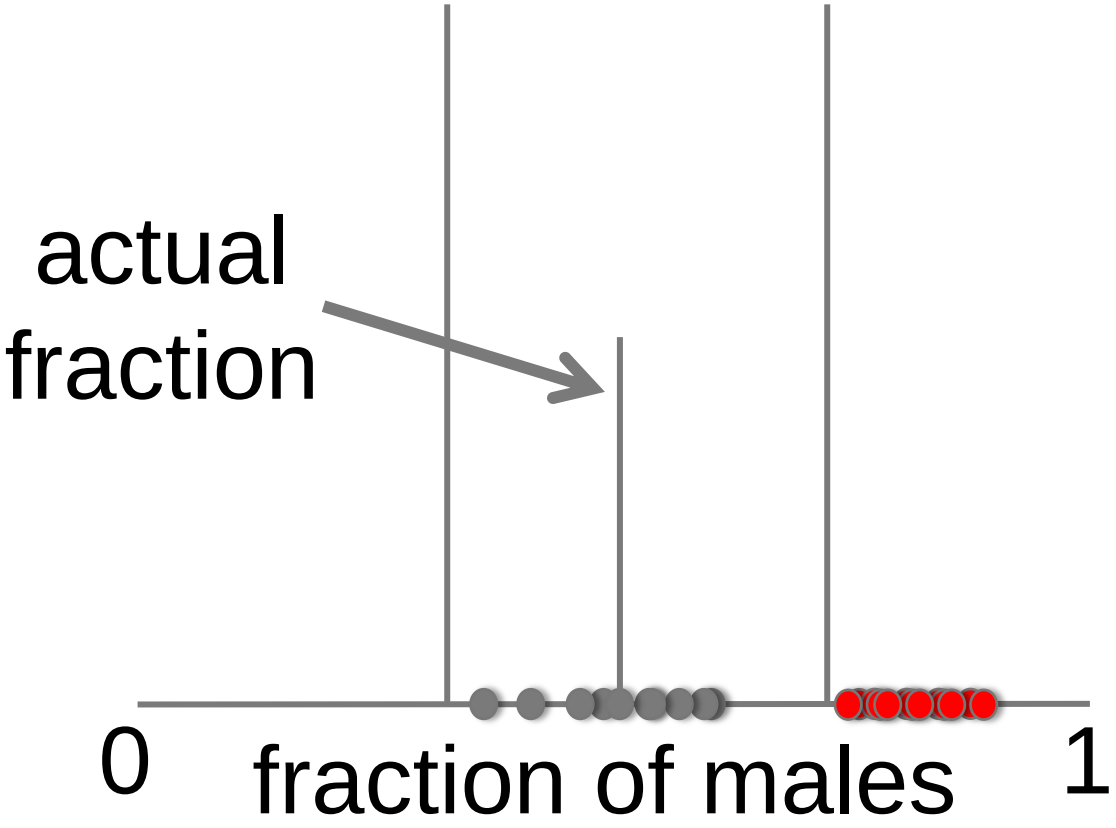
AVOIDING SPAMMERS: FORMAL

- If no spammers, just average all the results
- Average the contribution of each user = Fi (really helps!)
- Initialize: $\hat{F} = \frac{\sum_i F_i}{N}$
- Define: $\theta_i = \begin{cases} 1 - |F_i - \hat{F}|, & \text{if } |F_i - \hat{F}| < \lambda \\ 0, & \text{otherwise} \end{cases}.$
- Iterate: $\hat{F} = \frac{\sum_i \theta_i \hat{F}_i}{\sum_i \theta_i}.$
- Finally: $\hat{F}_{\text{final}} = \frac{\sum_{i,j} \theta_i F_{ij}}{\sum_{i,j} \theta_i}.$

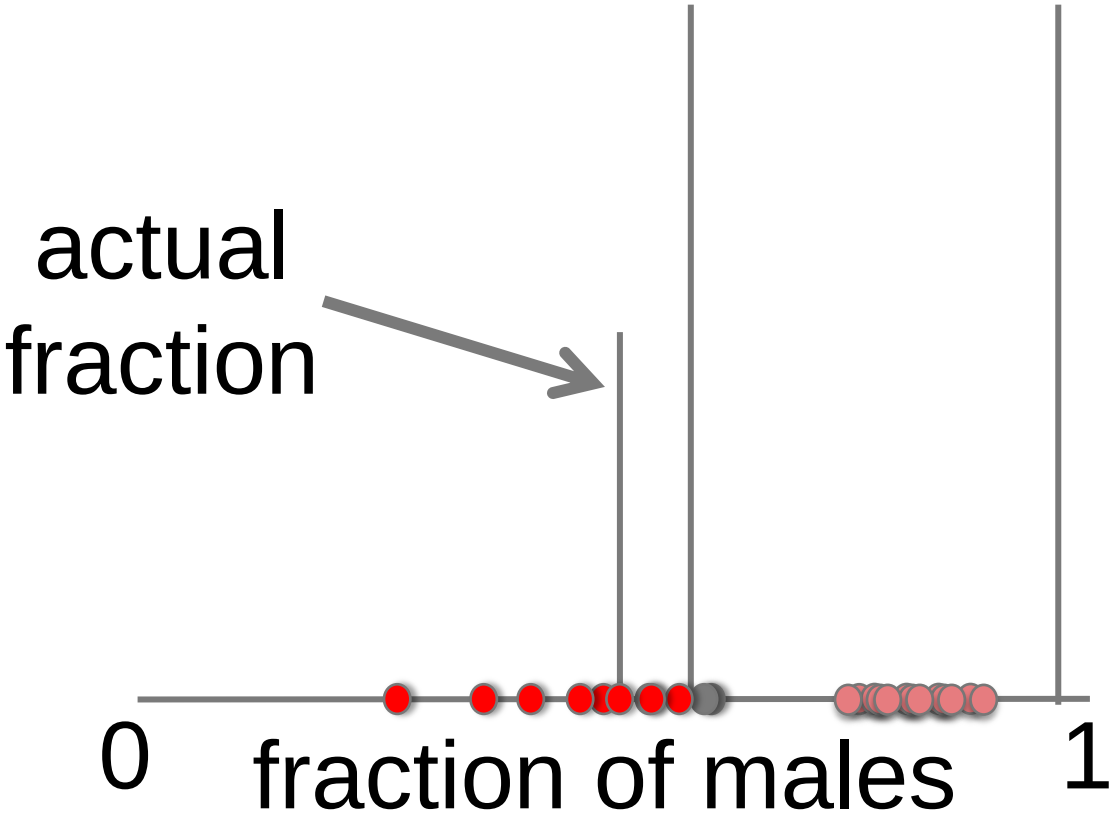
AVOIDING SPAMMERS: DEMONSTRATION



COORDINATED ATTACKS



COORDINATED ATTACKS



SOLUTION:
ADD RANDOM KNOWN RESULTS

GOLD STANDARD IMAGES

Not only 1 “golden truth” task

- Each worker complete only 1-2 tasks
- Spammers can identify those tasks

But distribute “golden truth” over all tasks

Old approx.: F = C/R

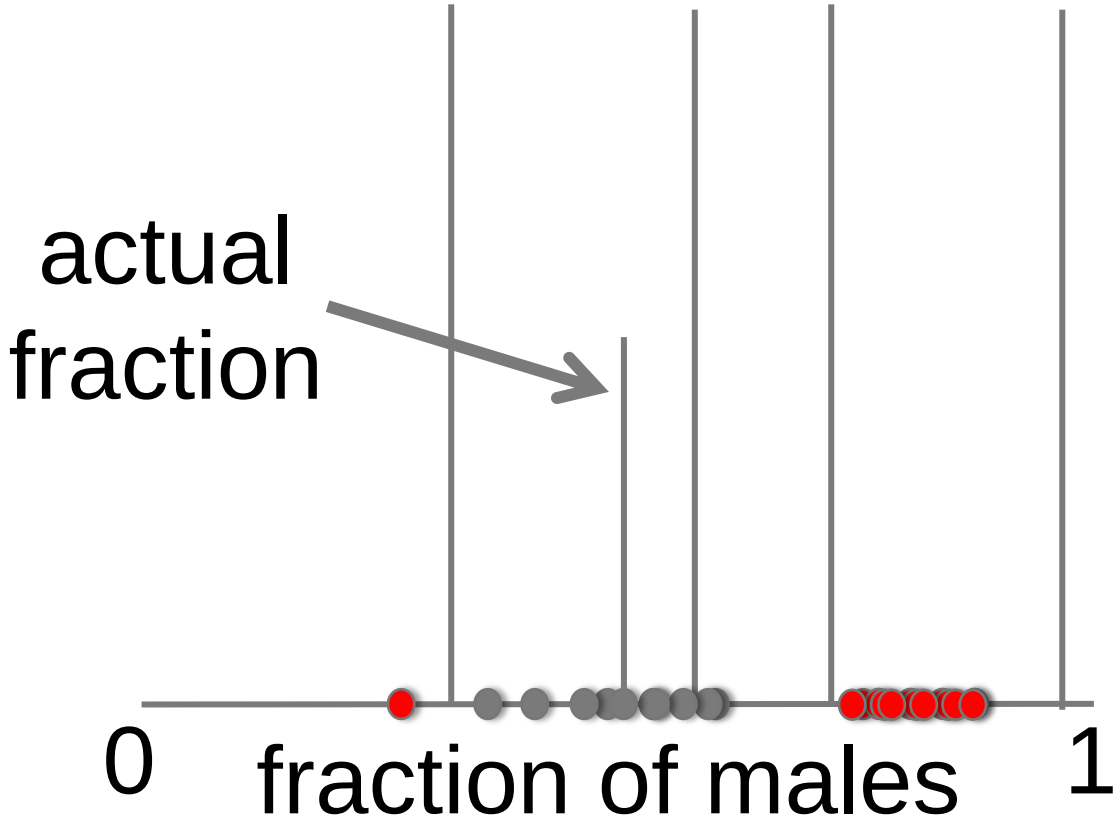
New approx.: F = (C-G)/(R-G)

C – count provided by worker

R – number of items

G – number of golden truth images

RANDOM RESULTS WEAKEN COORDINATED ATTACKERS



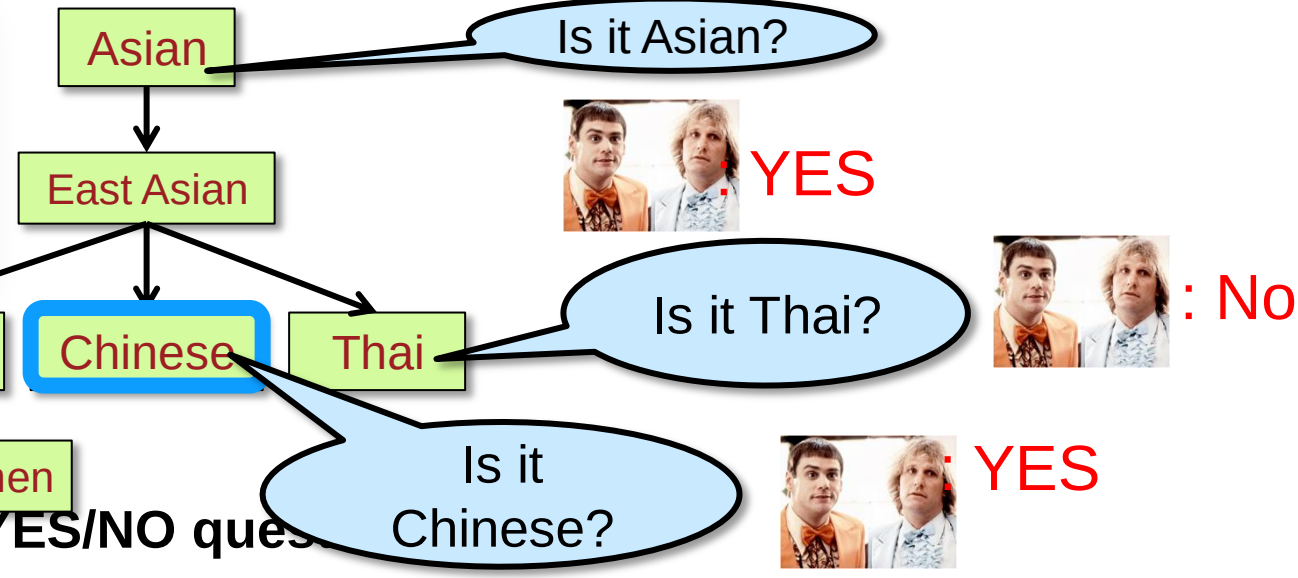
TRADEOFF

can withstand extreme coordinated attacks (>70% attackers)

in exchange for commensurate number of known labels

BEST USE OF RESOURCES: HUMAN ASSISTED GRAPH SEARCH

Given a DAG and some unknown target(s)



We can ask YES/NO ques

E.g. reachability •

HumanAssisted Graph Search: It's Okay to Ask Questions,
A. Parameswaran, A. D. Sarma, H. G. Molina, N. Polyzotis, j. Widom, VLDB '11

THE OBJECTIVE

Find an optimal set of questions to find the target nodes

- **Optimize cost:** Minimal # of questions
- **Optimize accuracy:** Minimal # of possible targets

Challenges

- Answer correlations (Falafel → Middle Eastern)
- Location in the graph affects information gain (leaves are likely to get a NO)
- Asking several questions in parallel to reduce latency

PROBLEM DIMENSIONS

Single target/Multiple targets

Online/Offline

- Online: one question at a time
- Offline: pre-compute all questions
- Hybrid approach

Graph structure

IMPORTANCE OF (GOOD) UI

- Good UI – better results
- Good UI – faster results

- Bad UI – inaccurate results
- Bad UI – workers leave without completing the task

CHALLENGES

Open vs. closed world assumption

Asking the right questions

Estimating the quality of answers

Incremental processing of updates

MORE CHALLENGES

Distributed management of huge data

Processing of textual answers

Semantics

More ideas?

RESEARCH AGENDA

- Data Model/Query Language
- Query Execution/Query Optimization
- Quality Control
- Storage/Caching
- User Interfaces
- Worker Behavior/Worker Relationship Management
- Interactivity
- Platform design
- Hybrid Human/Machine algorithms

(from VLDB'11 Tutorial by Doan, Franklin, Kossman, Kraska)

TEASERS

(1) - **User panel**. Used for logout and sh
A bit more about rating:
= 5 answers
=
=

(2) - **Never go/eat there**. Used to specify
not relevant for you and the system should
question about this place/restaurant

(3) **Submit**. First, you should select the f
current pair and you can click on this butto

OR

(4) **Specify**. Used if you want to give a sp
example. The frequency of more general q
submitted.

Got the rules, let me play!

TEASERS


score
6100
ESP Game
Concentrate...
time
0:14
What do you see?
taboo words
forest
meadow
guesses
tree
grass
Matched on: tree
submit
pass

TEASERS


AskIt!
Asking the Right Quetions
AskIt!
Admin View
Statistics
Logout
(5/10)
Hosagrahar V Jagadish
University of Michigan
DBLP
Who does Hosagrahar V Jagadish resemble most
Kevin Spacey
David Beckham
Barack Obama
Is Hosagrahar V Jagadish theoretician or practitioner?
Theoretician
Practitioner
Both
Is Hosagrahar V Jagadish senior or junior researcher?
Senior
Junior
In between

TEASERS


Score
50
Tag a Tune
Hear Here
Timer
1:48
Bonus
Describe the tune ...
0:11
Listening to the same tune?
same
different
1 in a row!
your descriptions
alarm
your partner's descriptions
wierd
wierd wind instrument
alien
submit
pass

TEASERS


01 home
02 learned rules
03 user answers
04 answer questions
05 about
macademo
crowdsourcing well-being
search
search
rules of the day
morning -> jogging
the flu -> chicken soup
academia
Full Paper
TAU DB Group
The MoDaS Project
the flu -> chicken soup
Correlation
Items correlated with the flu :
garlic
Items correlated with chicken soup :
a cold
the flu
You might find these interesting...
Key Facts About Flu and Flu Vaccine
Scientific American – Influenza
25 Comforting Chicken Soup Recipes
Chicken Soup Scientifically Proven to Help Cure Colds

CROWDPLANR

PLANNING MADE EASY WITH CROWD

BUILD ME A TRIP | **CONTRIBUTE** | STATISTICS | SEARCH TRIPS | HELP

Welcome, **Dennis**
(2 points)

Region: **Central Europe**
Period: **September**
Duration: **1 month**
Traveller: **2 friends**
Level: **cities**
Request: **We hired a car and want to do a series of short (one day) hikes in the Alpine scenery. We land and fly out from Milan, Italy.**

Current path: **Canazei - Meyerhofen - Konigsee - ...**

Please advise the next step:

I

Interlaken, Switzerland
[Stop at Interlaken]

Innsbruck, Austria
[Stop at Innsbruck]

I - Add new!

More – next week…

Questions?

REFERENCES

This presentation partially based on:

- “Counting with Crowd” slides by Adam Marcus
- Crowdsourcing tutorial from VLDB’11 by Doan, Franklin, Kossman and Kraska
- “Introduction to Crowdsourcing” slides by Tova Milo

References:

https://users.soe.ucsc.edu/~alkis/papers/ivd.pdf

https://amplab.cs.berkeley.edu/wp-content/uploads/2011/06/CrowdDB-Answering-Queries-with-Crowdsourcing.pdf

http://db.csail.mit.edu/pubs/mturk-cameraready.pdf