Crowd Data Sourcing: a Database Perspective

Susan Davidson

Most slides courtesy of Tova Milo and Daniel Deutch, Tel Aviv University

Data Everywhere

The amount and diversity of data being generated and collected is exploding

Web pages, Sensors data, Satellite pictures, DNA sequences, …

From Data to Knowledge

Buried in this flood of data are the keys to

- New **economic** opportunities
- Discoveries in **medicine, science and the humanities**
- Improving **productivity & efficiency**

However, raw data alone is not sufficient!!!

And sometimes we still need more input to interpret it.

People are always online

We have ubiquitous connectivity.

skype™

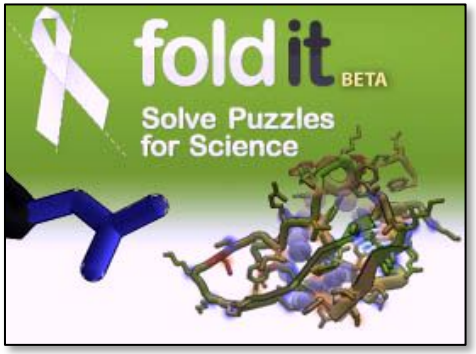
The research frontier

- Knowledge representation.
- Knowledge collection, transformation, integration, sharing.
- Knowledge discovery.

We will focus on human knowledge and what "tools" we need to mine it.

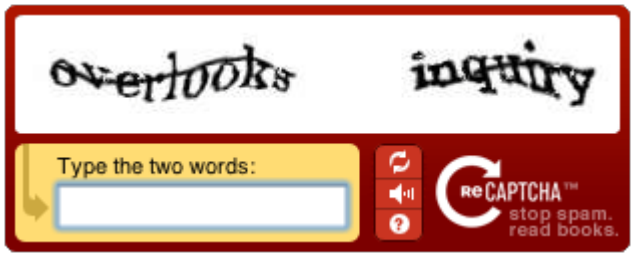
Think of humanity and its collective mind expanding…

Introducing Crowd (Data) Sourcing



The engagement of crowds of Web users for data procurement

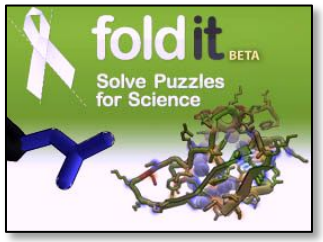
WIKIPEDIA
The Free Encyclopedia



amazon mechanical turk™
Artificial Artificial Intelligence

General goals of crowdsourcing

- Main goal
  - "Outsourcing" a task (data collection & processing) to a crowd of users



- Kinds of tasks
  - Can be performed by a computer, but inefficiently
  - Can be performed by a computer, but inaccurately
  - Tasks that can't be performed by a computer

So, are we done?

Need for principled solutions

So, are we done?

- **Analogy:** a database application can be constructed using C/[your favorite PL]
  - But you must hand code queries and implement indexing, optimizations, transactions, access control, etc
- In the same way, a crowdsourced application can always be hand-coded
  - Or simple applications can use platforms like Mechanical Turk

Challenges

- How to motivate the crowd?
  - Altruism, fun, money
- Get data, minimize errors, estimate quality?
- Direct users to contribute where it is most needed/ they are experts.
- DBMS-style generic platforms?
  - Declarative languages, optimizations

Crowd Data Sourcing

- The task is collection of data
- Two main aspects [DFKK'12]
  - Using the crowd to create better databases
  - Using database technology 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 Tutorial

Data-related Tasks (that can be) Performed by Crowds

- Data cleaning
  - E.g. repairing key violations by resolving contradictions
- Data integration
  - E.g. identify mappings
- Data mining
  - E.g. entity resolution
- Information extraction

[**Internet- Scale Collection of Human- Reviewed Data ,**
Q. Su, D. Pavlov, J. Chow, W.C. Baker, WWW '07]
[Matching Schemas in Online Communities: A Web 2.0 Approach,
R. McCann,W. Shen, A. Doan, ICDE '08]
[Amplifying Community Content Creation with Mixed Initiative Information Extraction,
R. Hoffman, S. Amershi, K. Patel, F. Wu., J. Fogarty, D. Weld, CHI '09]

Main Tasks in Crowd Data Sourcing

- What questions to ask?
- How to define the correctness of answers?
- How to clean the data?
- Who to ask? How many people?
- How to best use resources?

Declarative framework

Probabilistic data

Data cleaning

Optimizations and Incremental Computation

DB research efforts in Crowdsourcing

- CrowdDB (Berkeley and ETH Zurich)
  - Extending RDB with human-oriented query operators
- Qurk (MIT)
  - Crowd-powered workflows
- SCoOP (Stanford – Santa Cruz)
  - Optimizing crowd algorithms, declarative querying
- MoDaS (Tel Aviv University)
  - Foundations of crowdsourcing
- CrowdSearcher (Politecnico di Milano)
  - Crowdsourcing systems/platforms

Remainder of talk

- Will give a high level description of four different systems: Qurk, CrowdDB, CrowdForge, and CrowdMiner.
  - We will read detailed papers on these systems as part of the course
- Discuss course mechanics and next topics

Qurk

- Main observation: Tasks aided by MTurk can be expressed as workflows, with
  - Queries on existing data
  - "Black boxed" (User Defined Functions) that are tasks (HITs) to be performed by the Turker
- Basis: SQL3 + UDF
  - Special template language for crowd UDFs
  - Specify UI, quality control, possibly opt. hints

**[Crowdsourced Databases: Query Processing with People,
A. Marcus, E. Wu, D. R. Karger, S. Madden, R. C. Miller, CIDR 2011]**

A simple example – Crowdsourcing (Qurk)

name	Picture
Lucy	
Don	
Ken	
…	…

The goal:
Find the names of all the women
in the **people** table

```
SELECT name
FROM people p
WHERE isFemale(p)
```

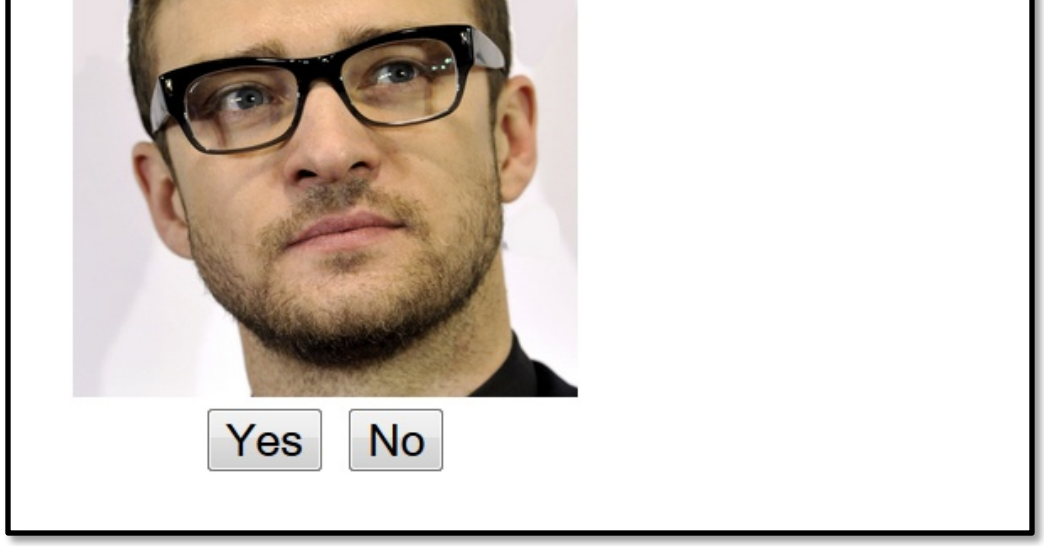
isFemale(%name, %photo)

Question: "Is %name a female?",
%photo
Answers: "Yes"/ "No"

16

A simple example – crowd data sourcing

name	Picture
Lucy	
Don	
Ken	
…	…



Issues

- Contradictions
  - The same form is presented to multiple users
  - Not everyone will have the same answer to every question
- Optimization
  - Cost of a HIT (optimized statically or at runtime)
  - Given a limited number of HITS, choose a subset
  - Evaluating predicates?

The magic is in the templates

- Templates generate UIs for different kinds of crowd-sourcing tasks
  - Filters: Yes/No questions
  - Joins: comparisons between two tuples (equality)
  - Order by: comparisons between two tuples (>=)
  - Generative: crowdsource attribute value

- Templates also specify quality control; e.g.
  - COMBINER: MajorityVote

But can you trust the crowd?

CrowdDB

- A different declarative framework for crowd datasourcing, also based on SQL
- Main difference: allows to crowdsource the generation of new tuples or attribute values
  - "closed world" assumption no longer holds
  - People are good at finding new data, and simple comparisons (e.g. entity resolution)

[CrowdDB: Answering Queries with Crowdsourcing,
M. J. Franklin, D. Kossmann ,T. Kraska, S. Ramesh, R. Xin SIGMOD '11]

CrowdDB example

- Either attributes or tables can be marked as crowdsourced

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

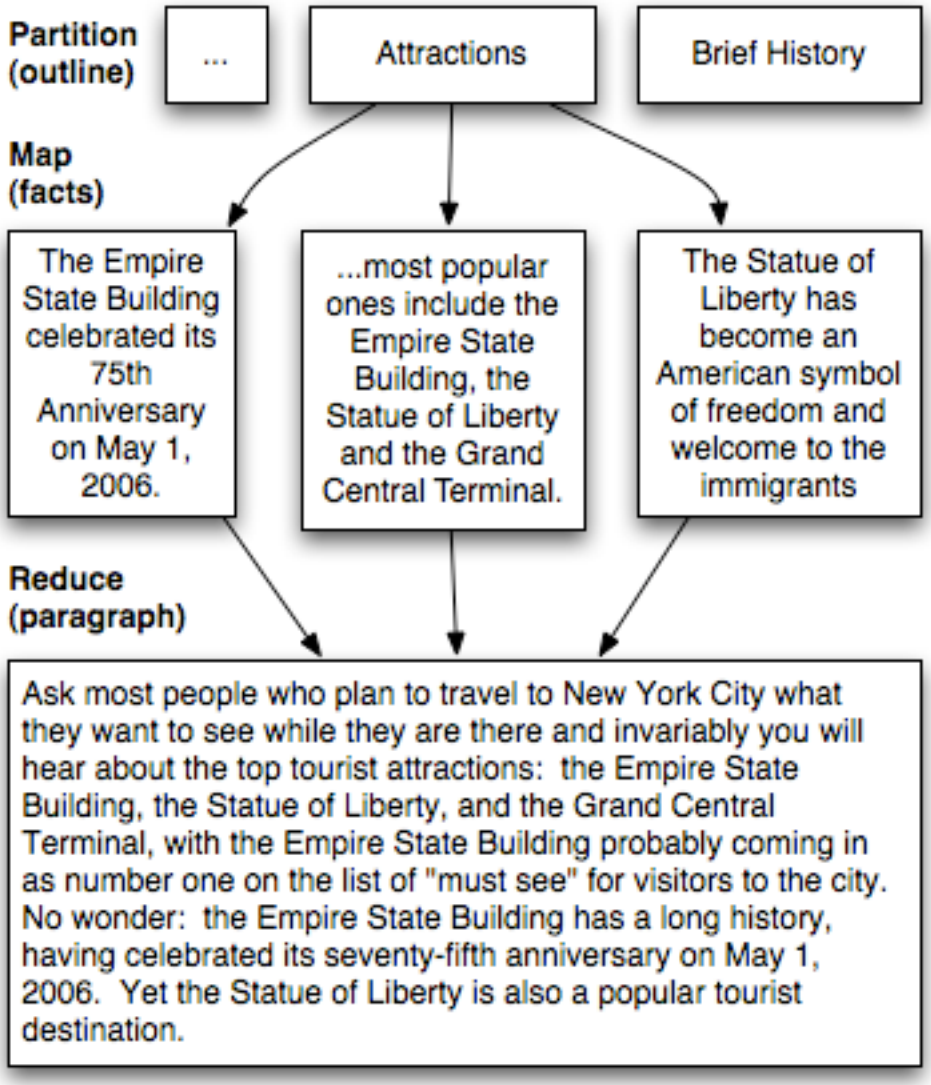
```
CREATE TABLE Department (
  university STRING,
  name STRING,
  url CROWD STRING,
  phone STRING,
  PRIMARY KEY (university, name) );
```

- Crowdsourcing is a side-effect of query processing
  - E.g. “SELECT url FROM Department WHERE name=“Math”
  - Tables are updated as a result

CrowdForge

- A declarative framework inspired by MapReduce
- Provides a small set of task primitives (partition, map, and reduce) that can be combined and nested
  - Allows to break MTurk tasks to small tasks and combine the answers
- Sub-tasks are then issued to the crowd

[CrowdForge: Crowdsourcing Complex Work,
A. Kittur, B. Smus S. Khamkar R. E. Kraut , UIST '11]


Partition
(outline)
...
Attractions
Brief History
Map
(facts)
The Empire State Building celebrated its 75th Anniversary on May 1, 2006.
...most popular ones include the Empire State Building, the Statue of Liberty and the Grand Central Terminal.
The Statue of Liberty has become an American symbol of freedom and welcome to the immigrants
Reduce
(paragraph)
Ask most people who plan to travel to New York City what they want to see while they are there and invariably you will hear about the top tourist attractions: the Empire State Building, the Statue of Liberty, and the Grand Central Terminal, with the Empire State Building probably coming in as number one on the list of "must see" for visitors to the city. No wonder: the Empire State Building has a long history, having celebrated its seventy-fifth anniversary on May 1, 2006. Yet the Statue of Liberty is also a popular tourist destination.

Crowd Mining: Crowdsourcing in an open world

- Human knowledge forms an open world
- Assume we want to find out what is *interesting* and *important* in some domain area

  Folk medicine, people's habits, …

- What questions to ask?

[Crowd Mining, Y. Amsterdamer, Y. Grossman, T. Milo, P. Senellart. SIGMOD 2013]

Back to classical databases...

- Significant data patterns are identified using data mining techniques.
- A useful type of pattern: *association rules*
  - *stomach ache → chamomile*
  - *sore throat → garlic, ginger*
- **Is it possible to mine the crowd?**

Turning to the crowd

The history of every user is modeled as a *personal database*

Treated a sore throat with garlic and oregano leaves…
Treated a sore throat and low fever with garlic and ginger …
Treated a heartburn with water, baking soda and lemon…
Treated nausea with ginger, the patient experienced sleepiness…
…

- Every case (occurrence of personal illness) = a *transaction* consisting of *items*
- Not recorded anywhere – a hidden DB
  - It is hard for people to recall many details about many transactions!
  - But … they can often provide summaries, in the form of personal rules

"To treat a sore throat I often use garlic"

Two types of questions

- Free recollection (mostly simple, prominent patterns)

  → **Open questions**

Tell me how you treat a particular illness

"I typically treat nausea with ginger infusion"

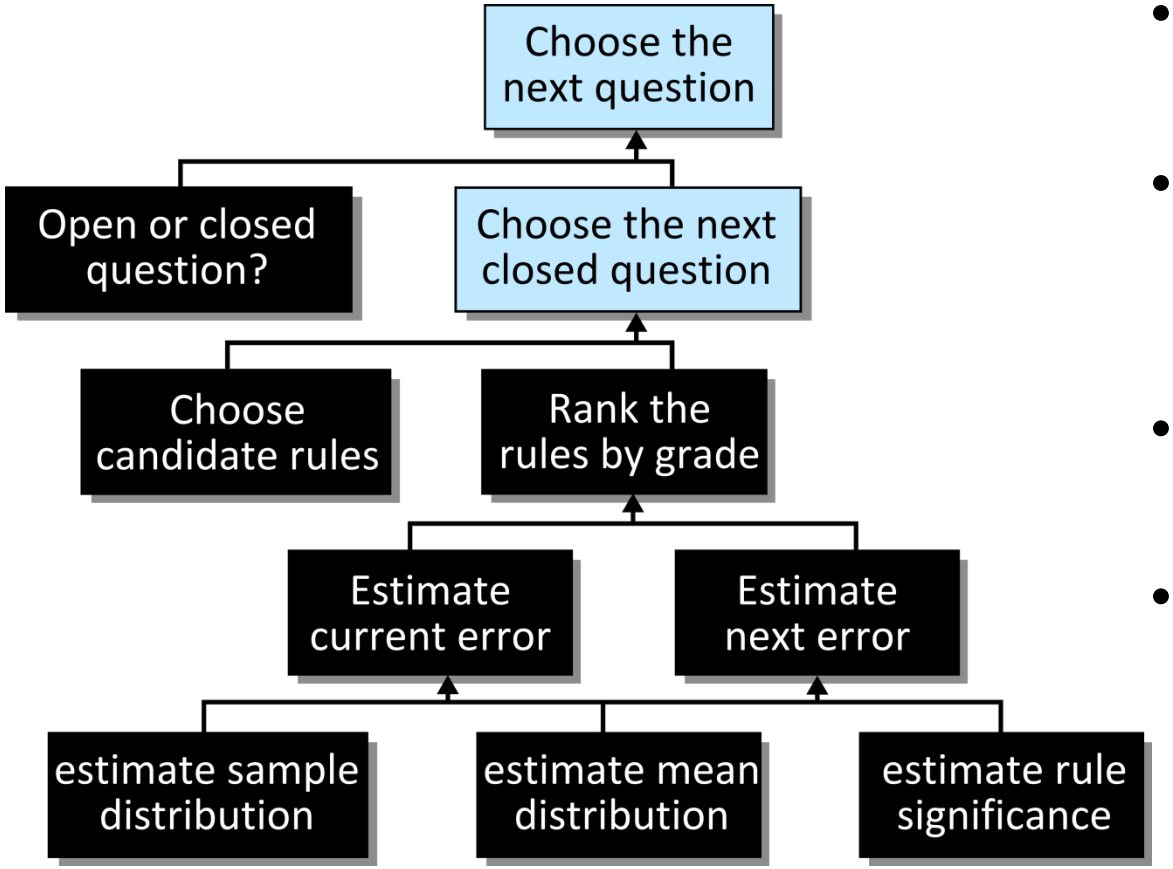
- Concrete questions (may be more complex)

  → **Closed questions**

When a patient has both headaches and fever, how often do you use a willow tree bark infusion?

CrowdMiner uses these two types interleavingly to find significant rules.

Framework components



- One generic framework for crowd-mining
- One particular choice of implementation of all black boxes
- They do not claim any optimality
- But they validate by experiments

How to measure the efficiency of crowd mining algorithms?

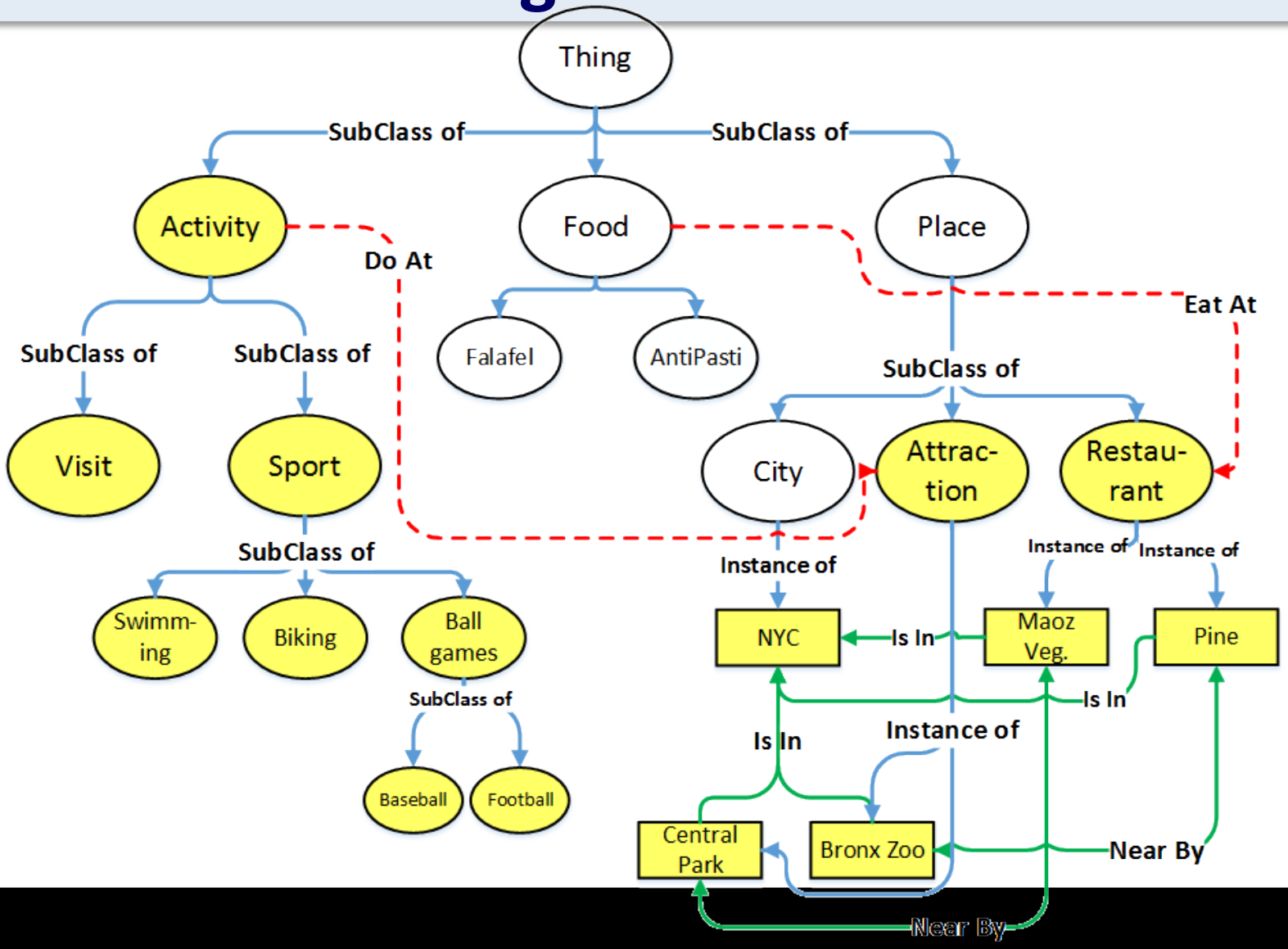
- Two distinguished cost factors:
  - **Crowd complexity:** # of crowd queries used by the algorithm
  - **Computational complexity:** the complexity of computing the crowd queries and processing the answers

[Crowd comp. lower bound is a trivial computational comp. lower bound]

- There exists a **tradeoff** between the complexity measures
  - Naïve questions selection -> more crowd questions

[On the Complexity of Mining Itemsets from the Crowd Using Taxonomies, A. Amarilli, Y. Amsterdamer, T. Milo. To appear in ICDT'14, March 2014]

Semantic knowledge can save work



Semantic knowledge can save work

Given a taxonomy of is-a relationships among items, e.g. **espresso is a coffee**

frequent({headache, espresso}) ⇒ *frequent*({headache, coffee})

Advantages

- Allows inference on itemset frequencies
- Allows avoiding semantically equivalent itemsets
  {espresso}, {espresso, coffee}, {espresso, beverage}…

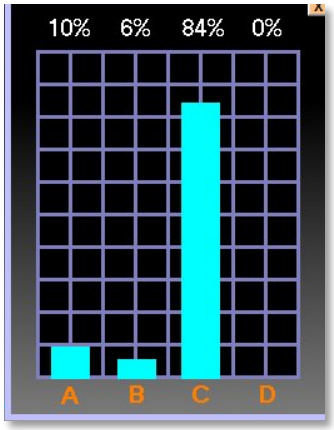
[draft publication]

MoDaS Contributions (at a very high level)

- **Formal model** for crowd mining; what questions are allowed, and how to interpret answers; personal rules and their overall significance.
- **Framework** of the generic components required for mining the crowd
- **Significance and error estimations.**
- **Crowd-mining algorithm**s

Can we trust the crowd ?

Common solution: ask multiple times



We may get different answers

- Legitimate diversity
- Wrong answers/lies

34

Things are non trivial …

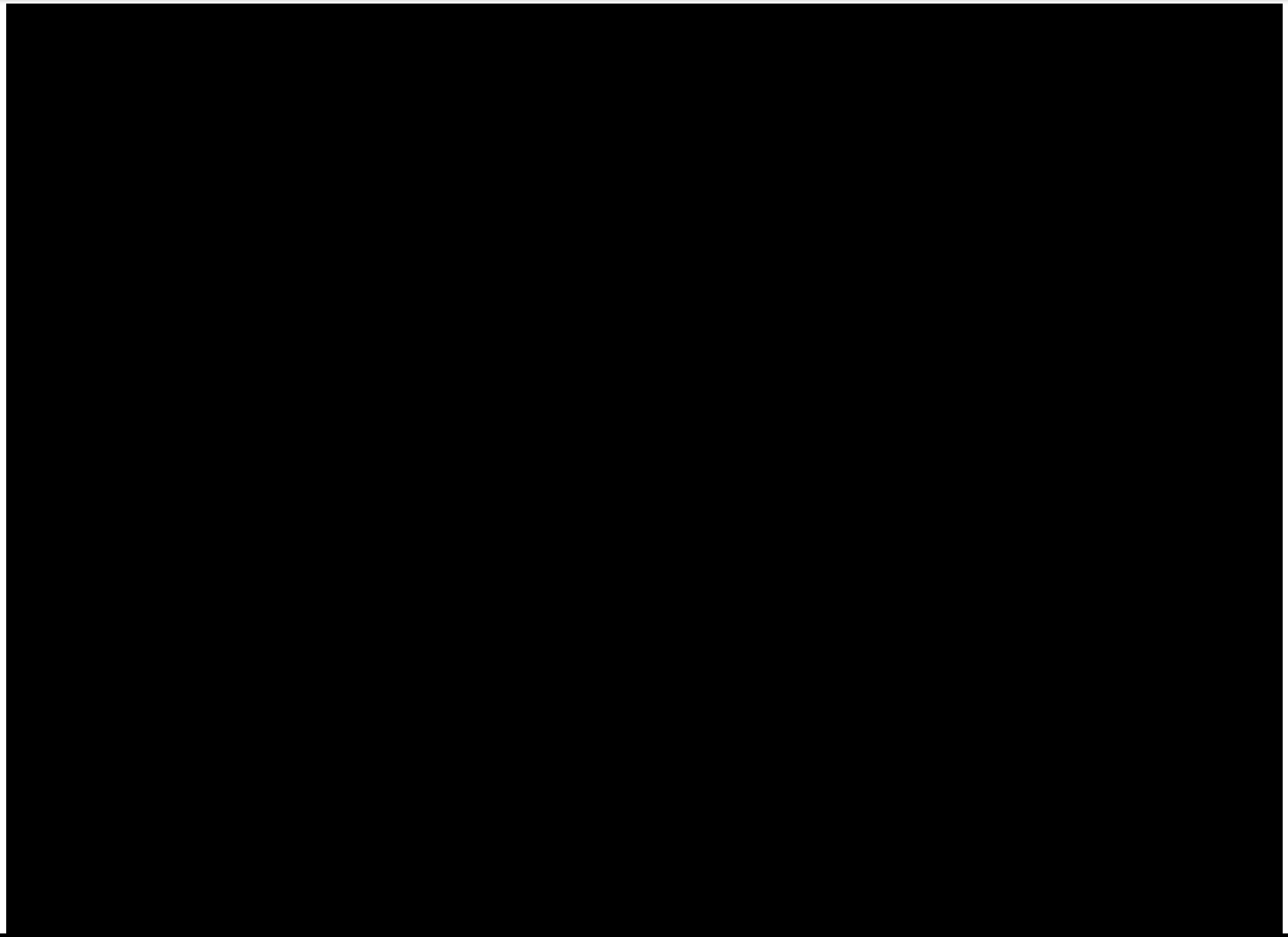
- Different experts for different areas
- "Difficult" questions vs. "simple" questions
- Data is added and updated all the time
- Optimal use of resources… (both machines and human)

Solutions based on

- Statistical mathematical models
- Declarative specifications
- Provenance

35

Can we trust the crowd ?



Summary

The crowd is an incredible resource!

"Computers are useless, they can only give you answers"

- Pablo Picasso

But, as it seems, they can also ask us questions!

Many challenges:

- (very) interactive computation
- A huge amount of data
- Varying quality and trust

This course

- Read papers associated with MoDaS, CrowdSearcher , CrowdDB, SCoOP and Qurk
  - Students pick several papers to present throughout the semester
  - Everyone must have read the paper prior to class and come prepared with questions.
- Do a project (research or implementation)
- Next time:
  - Using the crowd for top-k and group-by queries, by Davidson, Khanna, Milo and Roy. ICDT 2013.