SHONAN MEETING

Intelligent Information Processing – Chances of Crowdsourcing

Wolf-Tilo Balke, Takahiro Hara,
Seung-won Hwang, Christoph Lofi

November 18-21, 2013

Goals of this talk

- Provide an overview of **Social Information Processing** and **Crowdsourcing**
  - Challenges
  - Classification of scenarios
  - Applications

Social Information Processing

- Currently **social information processing** is a hot and emerging paradigm
  - Vaguely defined concept:
    "an activity through which collective human actions organize knowledge"
  - Obviously more complex information processing needs intelligence!
  - But…human intelligence?!

Social Information Processing

- Examples:
  - Building complex artefacts
    - Knowledge: Wikipedia.org
    - Software: Linux, Apache
  - Content Creation
    - YouTube, Flickr
  - User opinions
    - IMDb, Netflix, Amazon
  - Networking
    - Facebook. LinkedIn
  - etc.

Anhai Doan, R. Ramakrishnan, A.Y. Halevy: ***Crowdsourcing Systems on the World-Wide Web***. Communications of the ACM (CACM), No. 54, 2011.

Crowdsourcing

- Crowdsourcing has the power to **flexibly** add a certain degree of human intelligence to digital tasks
- Four challenges need to be overcome
  - How to **recruit** and retain users?
  - What **contributions** can users make?
  - How to **combine** the contributions to solve the target problem?
  - How to **evaluate** users and their contributions?

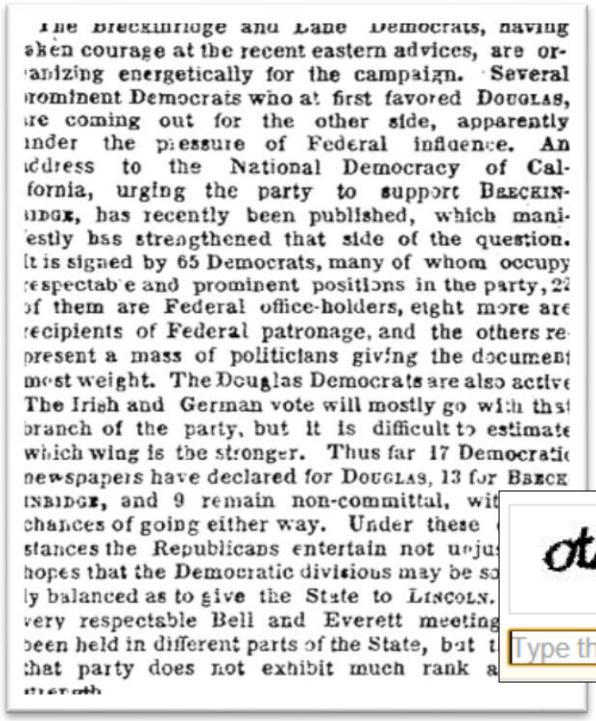
Crowdsourcing

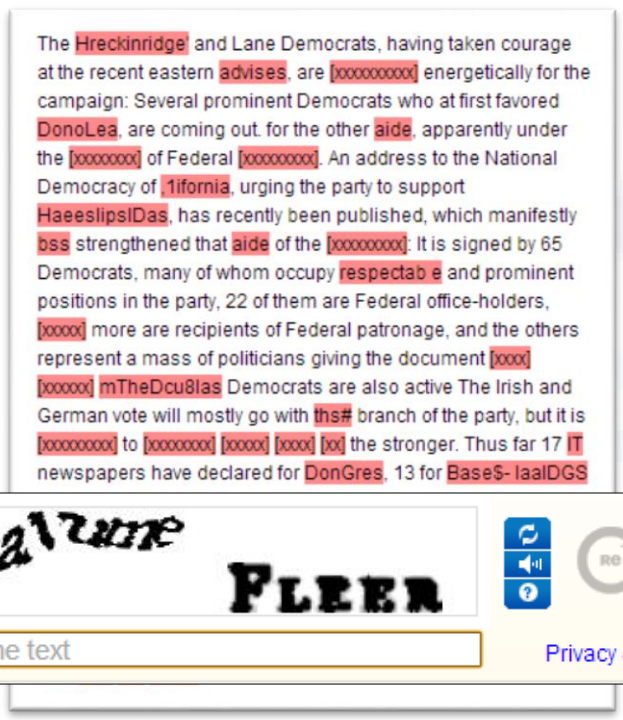
- Community platforms rely on **volunteers**
  - Intrinsically motivated
    - Users believe in the mission of the platform
    - Users somehow profit from the platform

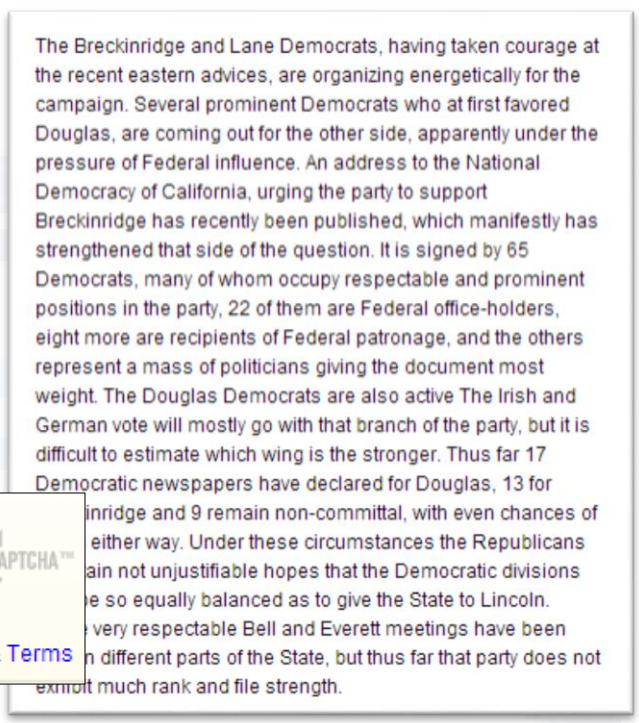
- **Problem:**
  - Mission cannot easily be changed, only specialized tasks solvable on each platform
  - Communities have to be carefully fostered and are hard to control

Crowdsourcing

- Possible solution: **Piggybacking**
  - **reCAPTCHA**
    correct faulty OCR with human help while providing anti-spam functionality to websites







Professional Crowdsourcing

- **Generic Task-Based Crowdsourcing**
  - General purpose platforms can facilitate virtually any task for anybody
    - Workers are attracted and retained by **paying money**

CrowdFlower

samasource
give work

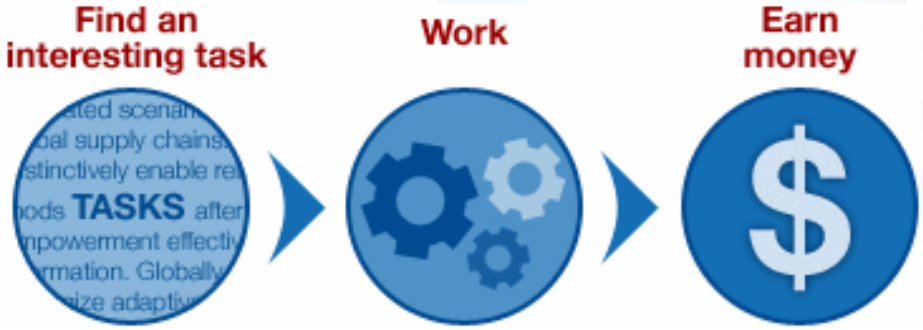
amazon mechanical turk™
Artificial Artificial Intelligence

Professional Crowdsourcing

– **Clients** can initiate a large crowd-sourcing task
  - Define the user interface
  - Define how the task is broken down to individual work packages: **HITs** (Human Intelligence Tasks)
  - Define the overall workload
  - Define how individual results are aggregated
  - Define payment per HIT

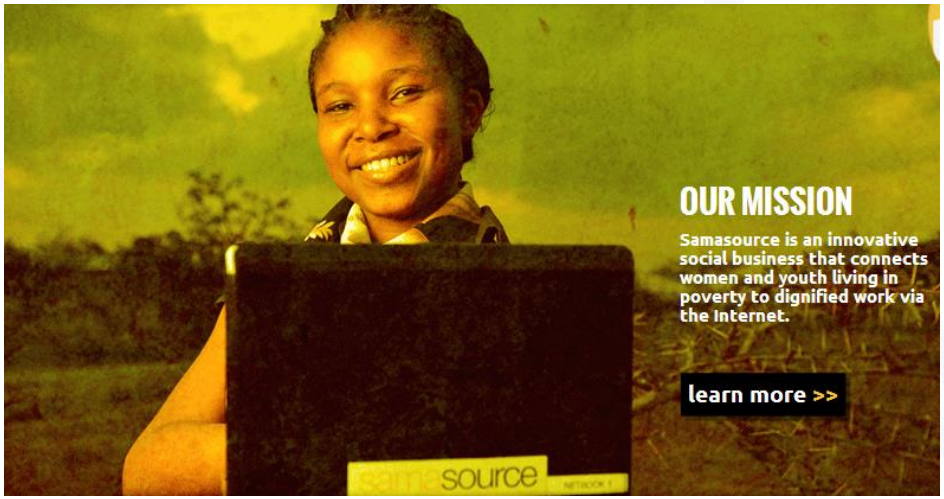
Professional Crowdsourcing

– **Workers** solve task
  - Short description of task
  - Transparent **payment** per HIT
  - Solves task using user interface provided by client
  - Can provide feedback with respect to task and its initiator



Professional Crowdsourcing

- Can crowdsourcing have a social impact?
  - Can crowdsourcing break traditional work patterns in a positive way?
    - Potential to change the way we work?
  - What kind of moral obligations do we have when issuing crowd-sourcing tasks?

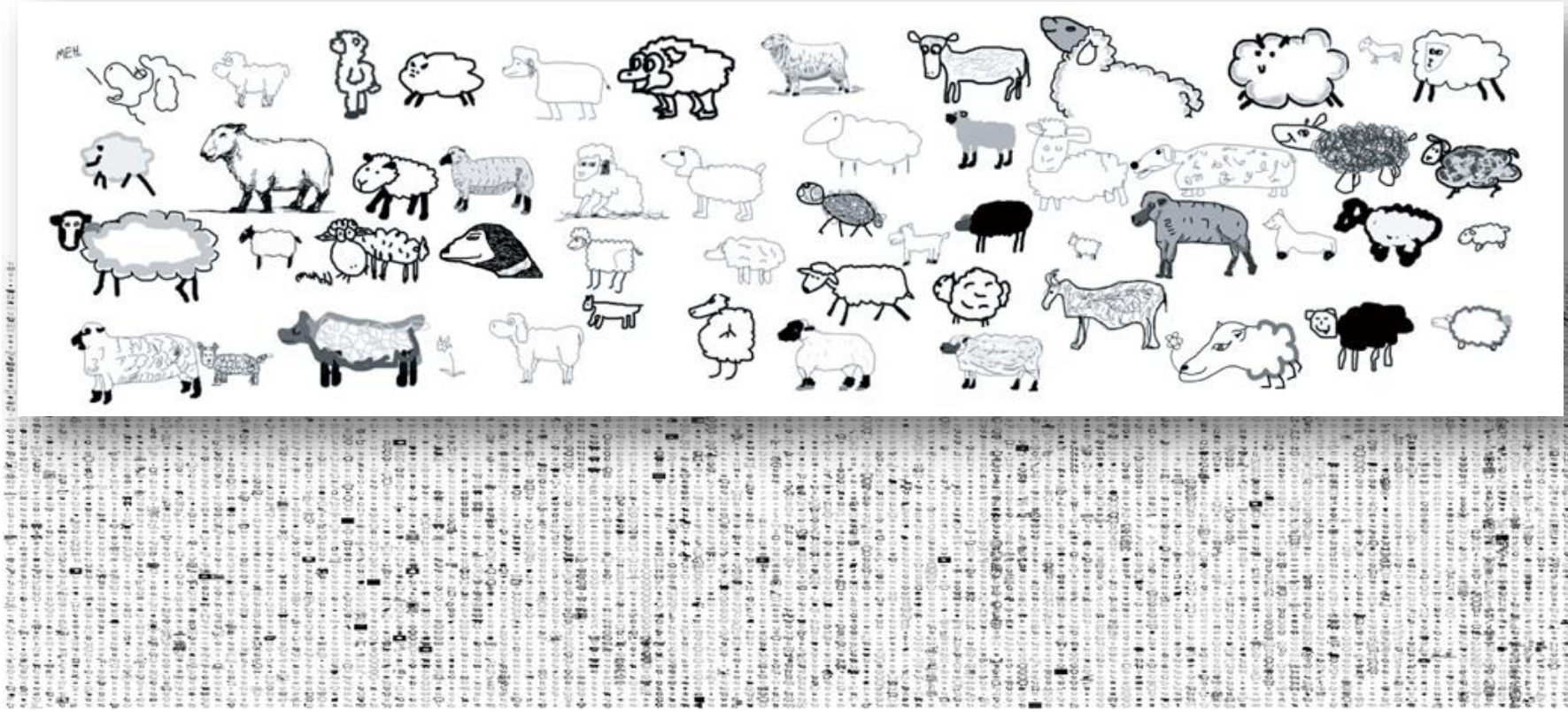


Success Story

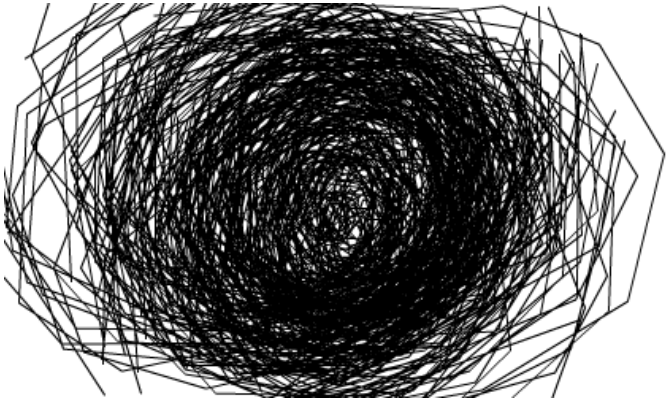
- Popular example from art: Aaron Koblin
  - http://www.thesheepmarket.com/
  - Laboral Centro de Arte, Gijon, Spain
    Japan Media Arts Festival, Tokyo, Japan
    Apex Gallery, New York, USA
    ElectroFringe, New Castle, Australia
    Media Art Friesland, The Netherlands

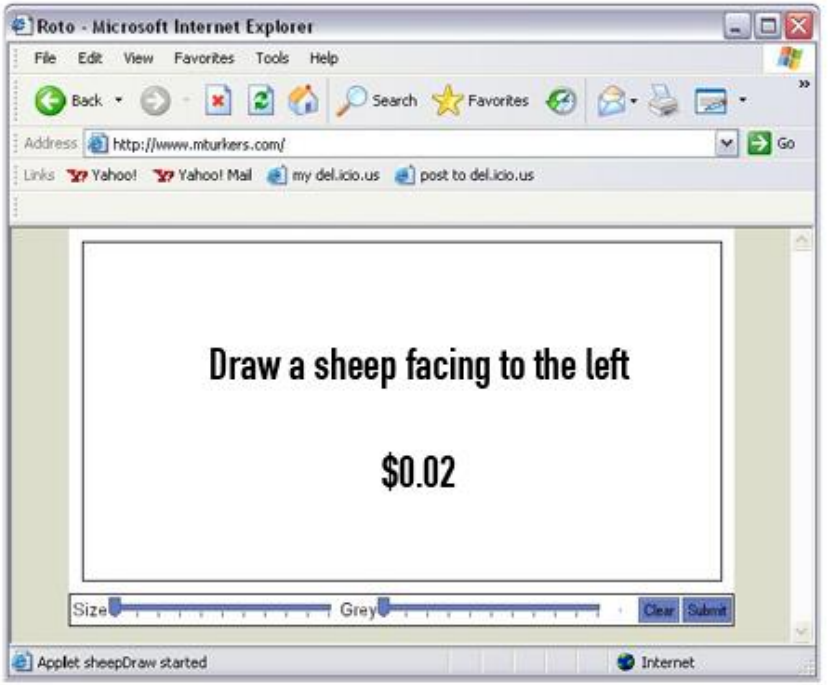
Success Story

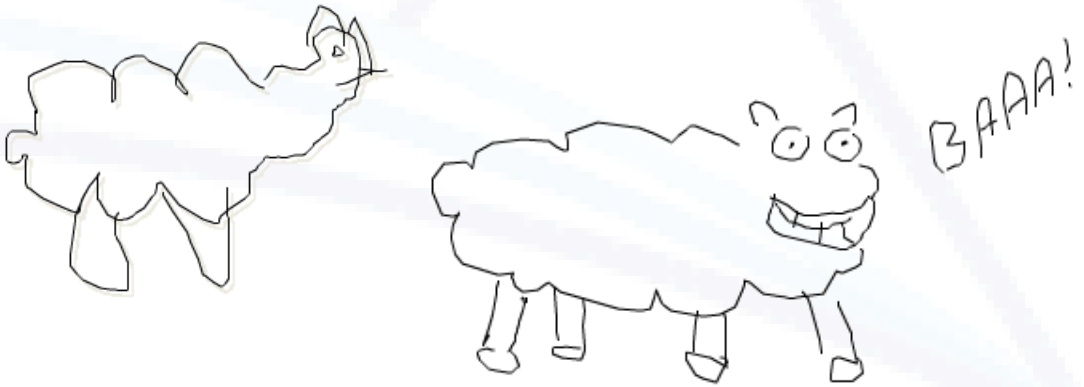
- You get what you pay for…
  - 10 000 sheep = 200 USD



Success Story



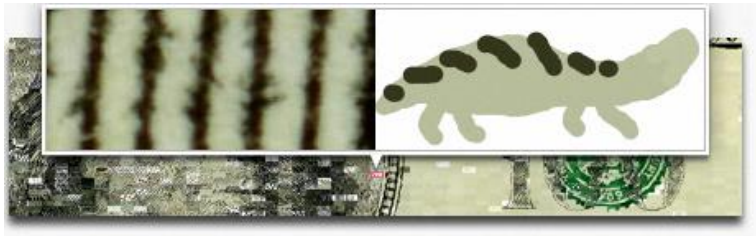




Success Story

- Popular examples from art reloaded
  - How about more detailed instructions?
  - www.tenthousandcents.com/

Success Story

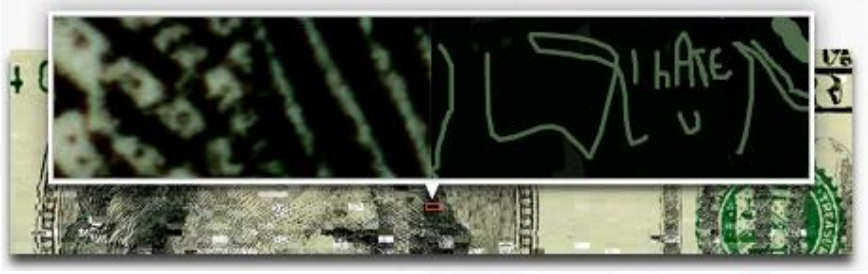


Collection Period: 2007/11 - 2008/03

Total paid labor: 10000 cents

Number of countries involved: 51

	Country	Avg. Time Spent/User	Percent Unique Visitors
1	United States	00:02:48	83.35%
2	India	00:11:32	75.31%
3	China	00:23:52	10.61%
4	Canada	00:01:57	93.88%
5	Philippines	00:10:05	60.00%
6	Egypt	00:31:54	3.12%
7	United Kingdom	00:01:24	93.75%
8	Germany	00:01:51	76.92%
9	Netherlands	00:01:11	100.00%
10	Poland	00:02:29	75.00%



Towards General Frameworks

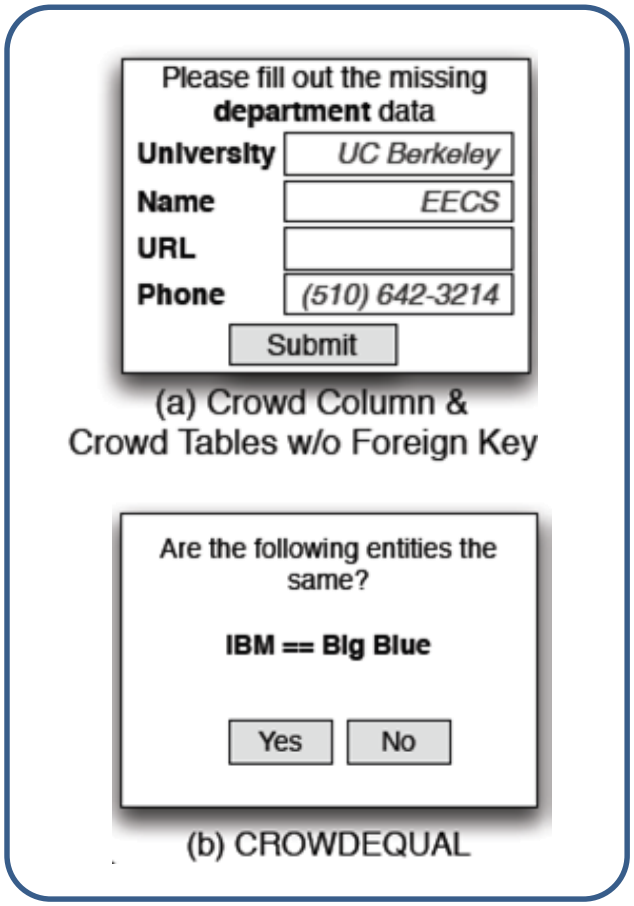
- **Crowd-Enabled Databases**
  - Core idea: Build a **database engine** which can dynamically crowdsource certain operations
    - **Complete missing** data during query time
      - Incomplete tuples (CNULL values)
      - Elicit completely new tuples

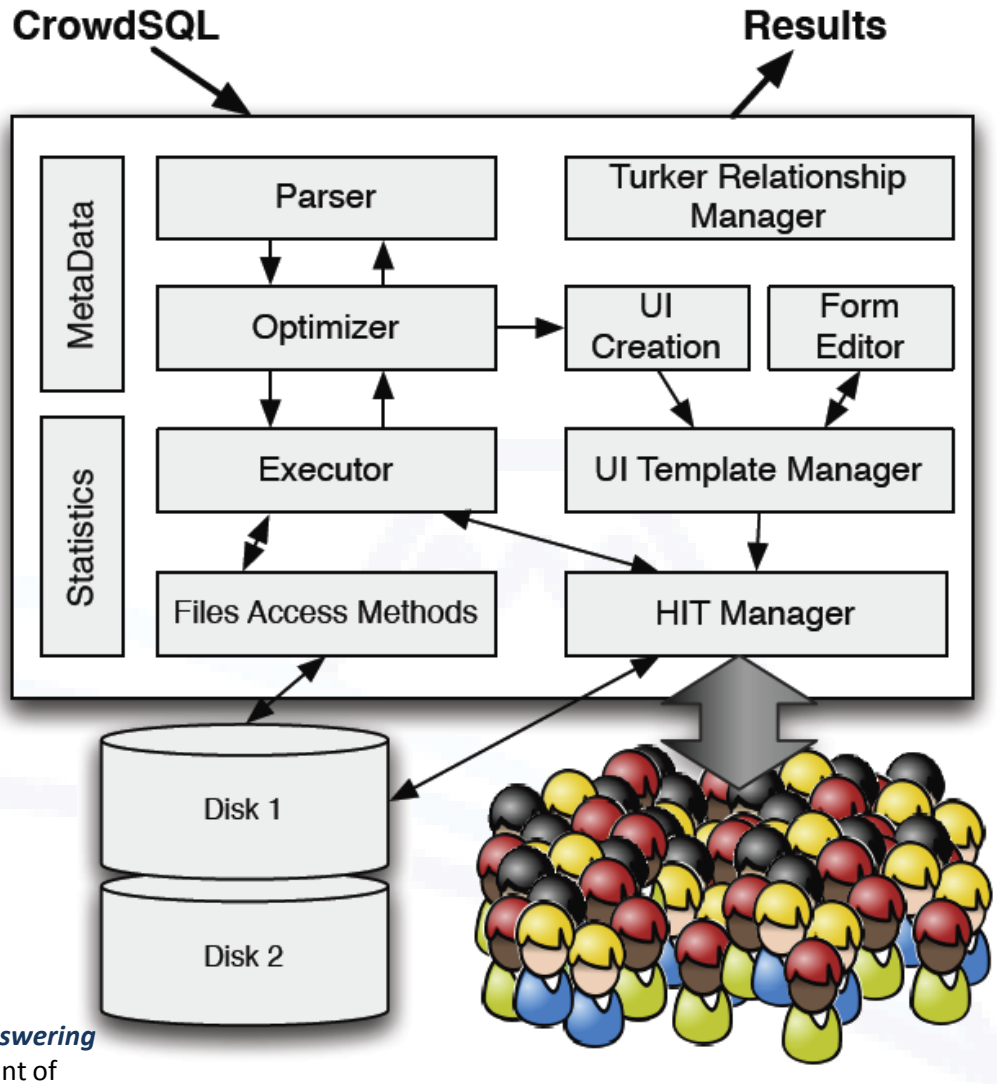
```
CREATE TABLE Department (
  university STRING,
  name STRING,
  url  CROWD STRING,
```

    - Use human intelligence **operators**
      - Entity resolution
      - Similarity rankings
      - etc.

```
SELECT market_capitalization FROM company
WHERE name = "I.B.M.";
```

Crowd-Enabled DB





M. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin: ***CrowdDB: Answering Queries with Crowdsourcing***. ACM SIGMOD Int. Conf. on Management of Data, Athens, Greece, 2011.

Classification of CS Tasks

- The ease-of-use and reliability of crowdsourcing tasks **varies** with the respective use case
- In general, three variables have to be controlled
  - **Answer/Solution Quality**, impacted by…
    - Worker diligence
    - Worker maliciousness
    - Worker quality and skills
  - **Execution Time**
    - Job attractiveness (payment vs. time)
    - Worker pool size
  - **Costs**
    - Number of HITs
    - costs per HIT (affected by time and skill needed)
    - Quality control overhead

Classification of CS Tasks

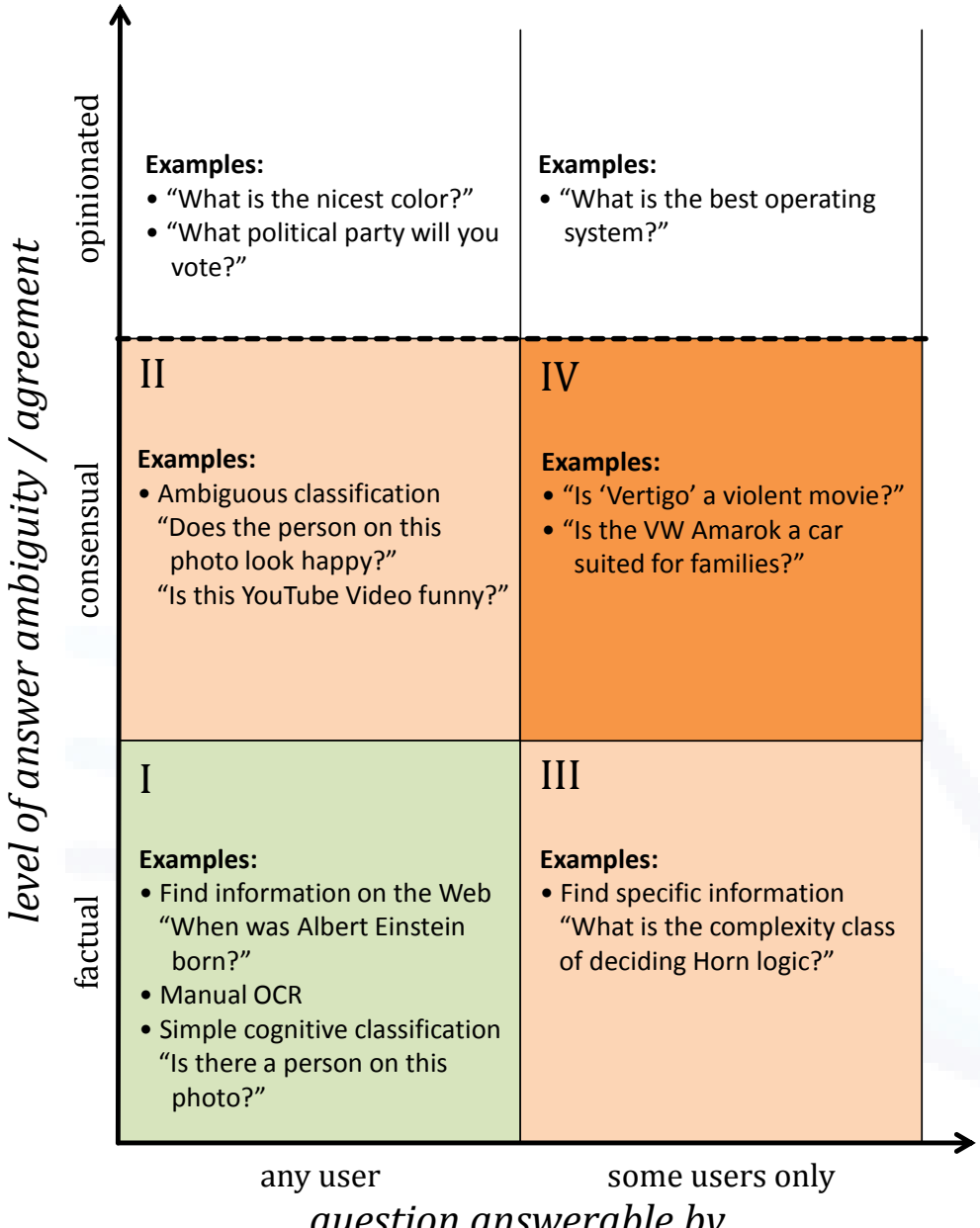
- Two general **discriminating properties** impacting these variables can be identified
  - **Ambiguity of the tasks solutions**
    - For a given solution, can we indisputably decide if it is correct or wrong?
      - **Factual tasks** (best case)

    - Can we at least reach a community consensus?
      - i.e. answer is considered correct by most people
      - **Consensual tasks** (not-so-good case)
    - Is there no correct answer? Answers completely subjective?
      - **Opinionated tasks** (luckily, uninteresting case for most computer science tasks)

Classification of CS Tasks

- **Required level of worker expertise / skill**
  - **Can anybody solve the tasks?**
    - General worker pool can be used
  - **Are special skills / background knowledge required?**
    - Worker pool must be filtered
    - Expert users must be found

Classification of CS Tasks



Lofi, C., J. Selke, and W. - T. Balke, ***"Information Extraction Meets Crowdsourcing: A Promising Couple***", Datenbank-Spektrum, vol. 12, no. 2: Springer, 05/2012

1: Factual Tasks / No Skill

1: Factual tasks not requiring any special skills

- Finding something in the Web, manual OCR, etc.
- **Each HIT is very simple**
  - No special skills required
  - No background knowledge required
- **Answers are not ambiguous**
  - i.e.: two non-malicious workers will give the same answer
- **Quality Control is easy to perform**
  - Need to catch and remove malicious users
  - Need to catch and correct oversights and mistakes

I: Factual Tasks / No Skill

- Quality Control with **majority votes**
  - Suitable for fixing minor oversights
  - Can be adjusted dynamically
  - Increases costs
  - In case of **malicious users**, dramatically increases costs!

1: Factual Tasks / No Skill

- Quality Control using **Gold Questions**
  - Tasks where correct answer is known upfront
  - Mix Gold questions into regular tasks
    - Workers **cannot distinguish** Gold Questions
    - Best practice: 10% Gold Questions
  - Mark users as being **malicious** if they fail Gold Questions
    - Malicious users are excluded from the tasks
      - Their previous results are discarded
    - **Malicious users will not receive payment**
      - Payment is also retrospectively renounced

1: Factual Tasks / No Skill

- Experiment 1: **Classify Movies by Genre**
  - Task: **Is a given movie a comedy movie or not?**
    - Special constraint: **Look-up movie in IMDb**

1: Factual Tasks / No Skill

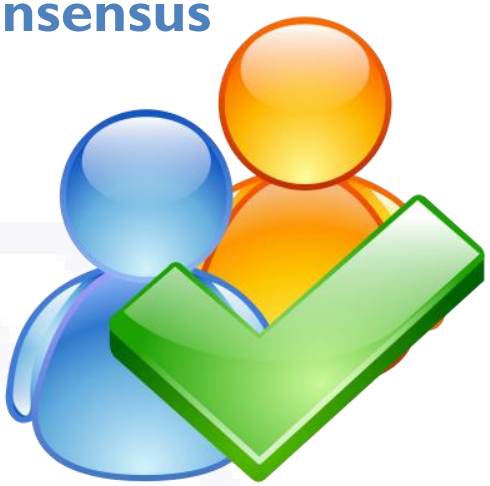
- **Settings**:
  - Amazon Mechanical Turk
  - Look-up 1,000 movies in IMDb
  - Majority vote of 10 workers each
  - 10% Gold questions
  - $0.03 per HIT with 10 movies?
    - Higher than later experiments, look-ups are time-consuming

- **Result** (stop after $30; 10,000 look-ups incl. Gold)
  - 562 minutes (9:22 hours)
  - 96% classified
    - 93.5% of those movies are classified correctly
    - Result quality / costs acceptable under certain constraints

2: Consensual Tasks / No Skill

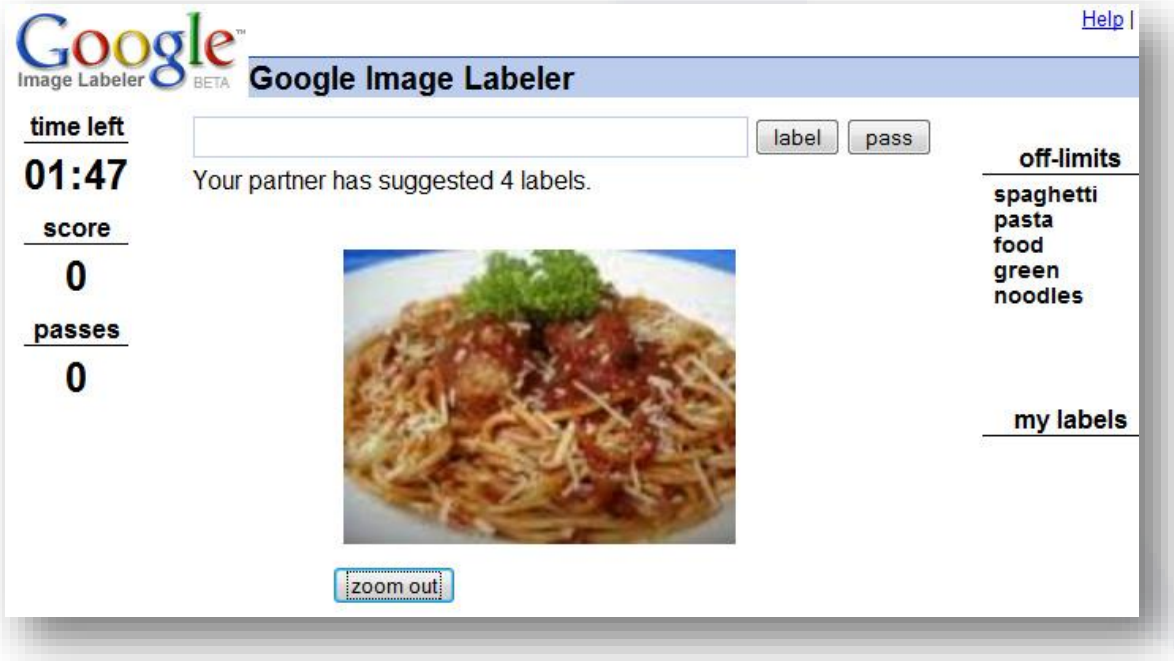
2: Consensual tasks not requiring special skills

- **There are no clearly "correct" results**
  - **"Correctness" is given by community consensus**
  - e.g.: "Is this YouTube video funny?"
- Quality control more challenging
  - More difficult to reach clear majority votes
    - Increased costs
  - Gold questions are difficult to use
    - How to obtain "correct" Gold values?
    - What is a good threshold for failing Gold questions?
    - Can users be punished for not sharing an opinion?
    - Result: Gold questions either not possible or very ineffective
  - Workers know this → **Higher incentive for cheating!**



2: Consensual Tasks / No Skill

- Example: ESP Game & Google Image Labeler
  - Idea: "**Games with a purpose**"
    - Image Labeling: Guess your partner's tags, and both score.
    - No payment necessary
    - Lower incentive for cheating? (Still happened a lot)



ESP Game and
Google Image Labeler
are now offline.

3: Factual Tasks / Special Skill

3: Factual tasks requiring special skills

- Answers are factual, i.e. clearly right or wrong
  - Quality control with majority votes, Gold questions possible
- But: Some **background knowledge** or **special skills** are required to solve task
- **Challenge**:
  - Find and retain workers which possess the required skills

3: Factual Tasks / Special Skill

- **Filter workers before task execution**
  - **Worker Self-Assessment**
    - Prone to abuse
    - only suitable for honest workers
  - **Reputation systems**
    - Workers gain reputation for successfully solving complex tasks
    - Not offered by most CS platforms
  - **Expert Communities**
    - There are expert communities for nearly any topic in the social web
      - But their expertise cannot be tapped easily !

3: Factual Tasks / Special Skill

- **Filter workers during task execution**
  - **"I don't know option"**
    - Should be offered when not all workers can solve all tasks
      - If not, users will guess or provide wrong answers
    - Should still be paid
      - If not, users will protest against task and initiator
    - Can be easily abused!

4: Consensual / Special Skill

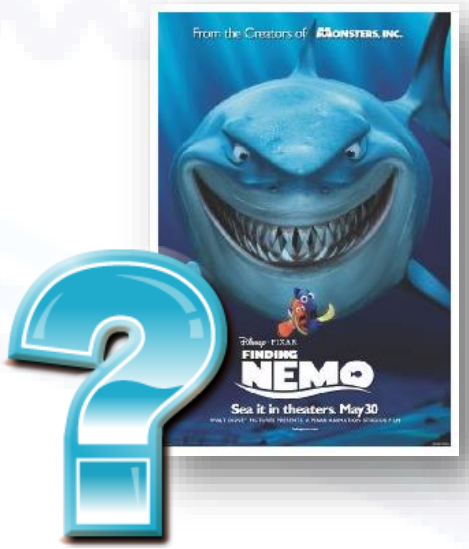
4: Consensual tasks requiring special skills

- Combines all challenges
  - Difficult to find suitable workers
  - quality hard to control
  - high rate of abuse

- Experiment 2: **Classify Movies by Genre**
  - Task: **Is a given movie a comedy movie or not?**
    - No internet look-up!
    - If the movie is known, subjective judgement should be provided
      - Background knowledge required
      - Otherwise: "I don't know this movie"

4: Consensual / Special Skill

- **Settings**:
  - Amazon Mechanical Turk
  - Judge 1000 random movies
    - Consider only movies which have consensual genre classifications in IMDb, Rotten Tomatoes, and Netflix
      - Only 10,562 movies overall
      - Use these movies as "truth"
  - Majority vote of 10 workers each
  - No Gold questions
  - $0.02 per HIT with 10 movies



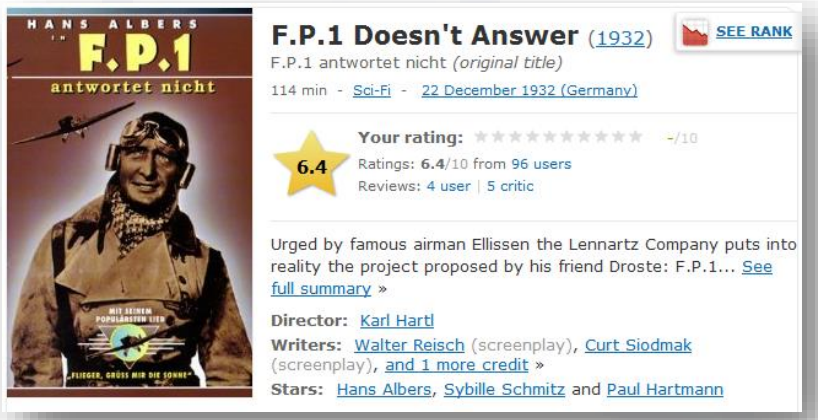
4: Consensual / Special Skill

- **Result** (stop after $20; 10,000 answers)
  - 105 minutes (1:45 hours)
  - **89% reached a consensus**
    - 59% of these movies are classified correctly

- **What went wrong?**
  - Malicious workers!
    - 62% selected "comedy" (first choice in form)
      - 30% of all movies in test set are indeed comedies
    - 24% selected "no comedy"
      - 70% of all movies in test set are no comedies
    - 14% selected "I don't know this movie"

4: Consensual / Special Skill

- Observation: the test set contains some **very obscure movies**
  - Quick survey among students: knew only 10%-20%
  - But: **Many workers claimed to know all movies**
    - Judged 56% of all movies as comedies, 44% as no comedy
    - Originate just from two distinct countries
  - All others workers:
    - Knew only 26% of all movies
    - 32% comedy
    - 68% no comedy
    - Realistic values!



4: Consensual / Special Skill

- Experiment 3:
  - Similar two experiment 2, but exclude all workers from the two offending countries
    - Hopefully, only trustworthy workers remain
- **Result** (stop after $20; 10,000 look-ups)
  - 116 minutes (1:56 hours)
  - **63% of all movies reached consensus**
    - Of those, **79% are classified correctly**
- **Result still disappointing**
  - Obscure movies do not reach consensus
  - Consensus still not reliable

Hybrid Approaches

- **How to perform better?**
  - Employ hybrid techniques combining machine-based heuristics with occasional help of humans
- Tackle the following challenges
  - **Performance**
    - Drastically speed up crowdsourcing times (not everything needs to be crowd-sourced)
  - **Costs**
    - Require just few crowdsourcing HITs for obtaining a large number of judgements
  - **Data Quality**
    - Circumvent the impact of malicious workers
    - Reliably obtain judgements for even obscure and rare items

Challenges and Future Visions

- How can crowd-sourcing help to solve **current problems** encountered in data processing?
- How can this be achieved **efficiently** and **reliably**?
  - Hybrid approaches? How can hybridization be designed in a structured fashion (patterns)?
  - How can result quality be measured and increased?
  - How can workers be recruited and retained?
  - How can workers be involved in the tasks?
    - e.g., for providing training during system setup? As on-demand workers during system execution?

Meeting Overview

Monday

- Today's Schedule

07:30-09:00	Breakfast
09:00-09:10	Shonan Introduction by Staff
09:10-12:00	Seminar Session with Coffee Break - Opening briefing from organizers - Position talks from participants
12:00-14:00	Lunch with Photo Shooting
14:00-18:00	Seminar Session with Coffee Break - Position talks from participants (continued) - Discussion to categorize the issues addressed by the participants
18:30-19:30	Dinner
19:30-	Free Time

Tuesday

- Tomorrow:
  - Break-out Sessions
    - Topic-based, application-based,…
  - Discussion of challenges and relevant issues in smaller groups
    - Result: group presentation

Wednesday

- Wednesday
  - Morning: break-out session result presentation
  - Afternoon: excursion to Kamakura
  - Banquet Dinner

Thursday

- Thursday
  - Idea marketplace and incubator for sparking collaborations
  - Final organizer wrap up

- **Let's have a good and inspiring meeting!**