



EPL646 – Advanced Topics in Databases

Lecture 17

Introduction to Crowdsourcing

Demetris Zeinalipour

<http://www.cs.ucy.ac.cy/~dzeina/courses/epl646>

Lecture Outline



- ***Introduction to Crowdsourcing***
 - *Definitions, Stakeholders, Incentives, Landscape, Challenges, Smartphone Era, Previous Tutorials*
- ***Web & DB Crowdsourcing***
 - *Implicit Crowdsourcing: reCAPTCHAs and ESP Game*
 - *Explicit Crowdsourcing: Contests, Microwork, Declarative CS, Wisdom-of-the-Crowd, Crowdfunding, Crowdvoting, Q/A*

Crowdsourcing (CS) Definitions



- **Crowdsourcing = Crowd + Outsourcing**
 - Jeff Howe (2006). "The Rise of Crowdsourcing". Wired
- **Merriam-Webster Definition:**
 - "the practice of obtaining needed **services, ideas, or content** by **soliciting contributions** from a **large group** of people, and **especially** from an **online community**, rather than from **traditional employees or suppliers**."
 - URL: <http://www.merriam-webster.com/dictionary/crowdsourcing>
- **From our recent work:**
 - Crowdsourcing refers to a **distributed problem-solving model** in which a **crowd** of **undefined size** is engaged in the task of **solving a complex problem** through an **open call ... for monetary or ethical benefit**.

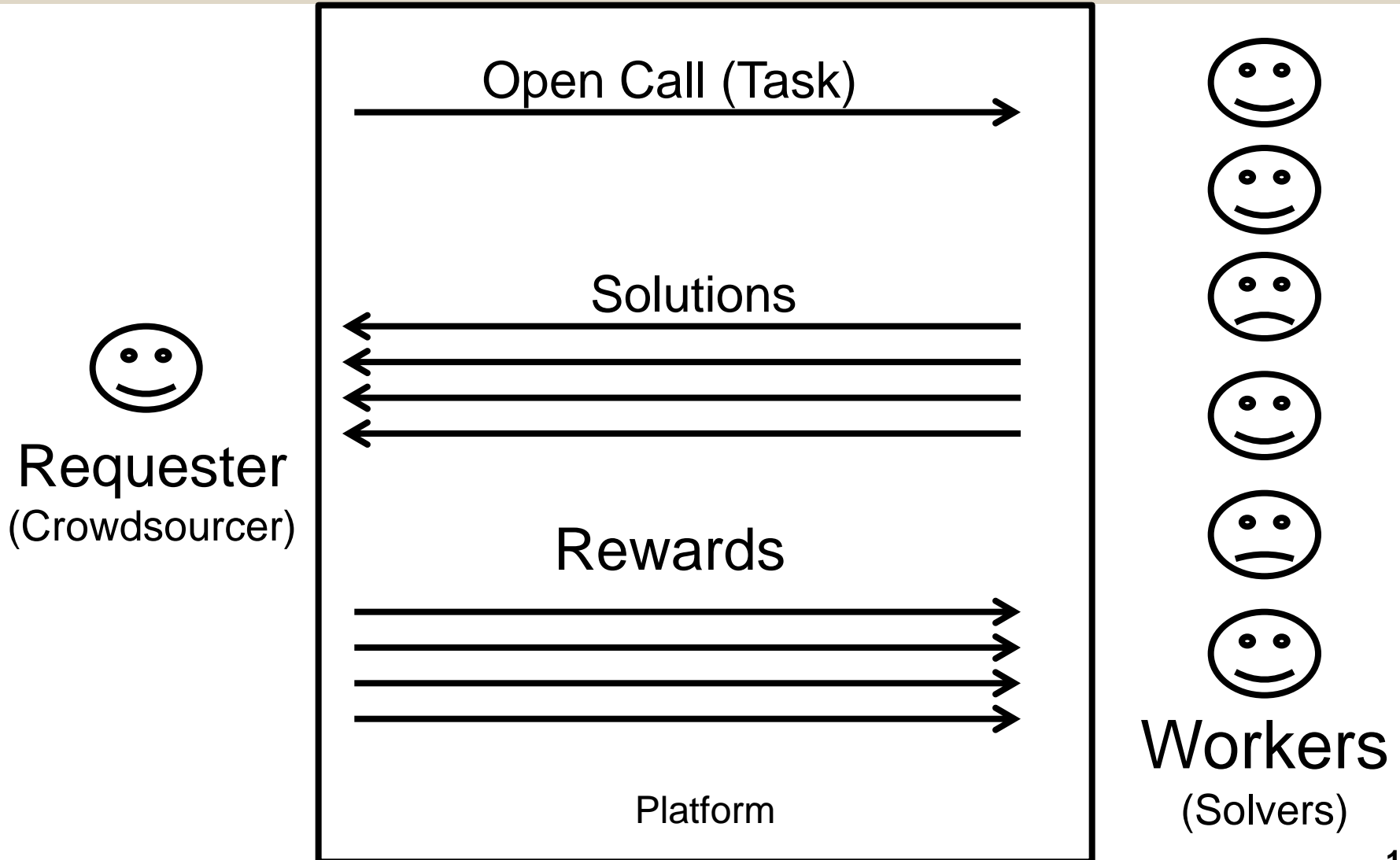
"Crowdsourcing with Smartphones", Georgios Chatzimiloudis, Andreas Konstantinidis, Christos Laoudias, Demetrios Zeinalipour-Yazti, IEEE Internet Computing, Special Issue: Sep/Oct 2012 - Crowdsourcing, May 2012. IEEE Press, Volume 16, Pages: 36-44, 2012.

Other Faces of CS



- **Crowdsourcing:** New buzzword with old meaning:
 - *"peer production, user-powered systems, user-generated content, collaborative systems, community systems, social systems, social search, social media, collective intelligence, wikinomics, crowd-wisdom, smart-mobs, mass collaboration, and human computation. "*
- **Many consider the following to be part of the greater Crowdsourcing picture:**
 - Wikipedia, Linux, Yahoo Answers!, etc.
- Crowdsourcing involves **real users, connected through the Internet** that **collaborate to solve problems** that computers can't (see next flow) for some incentives (not only free of charge).

Crowdsourcing StakeHolders



Crowdsourcing Incentives



- Tangible (Monetary) Incentives
 - Cash, Credit or Gifts (MTurk, Kickstarter)
 - Unintended or as-a-by-product (reCaptchas)
- Ethical Incentives
 - Socialize & Fun
 - Earn Prestige
 - Altruism
 - Learn something New
- Usually a combination of several incentives

Crowdsourcing Challenges



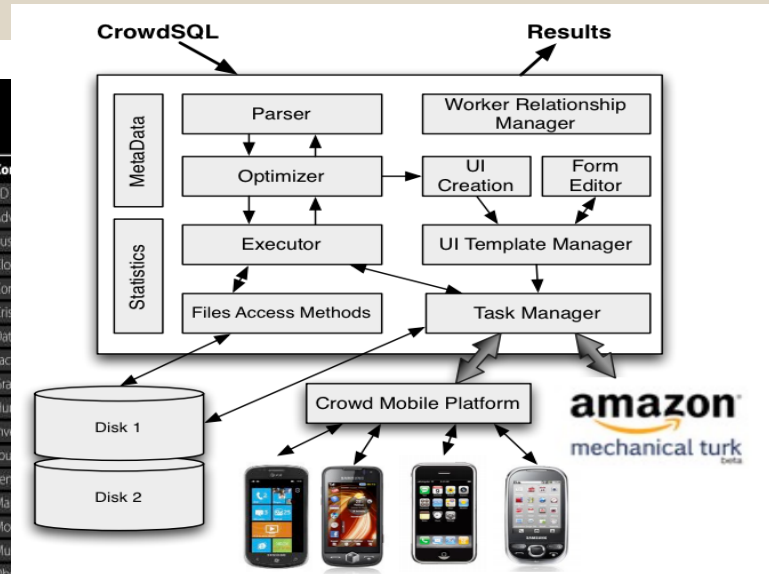
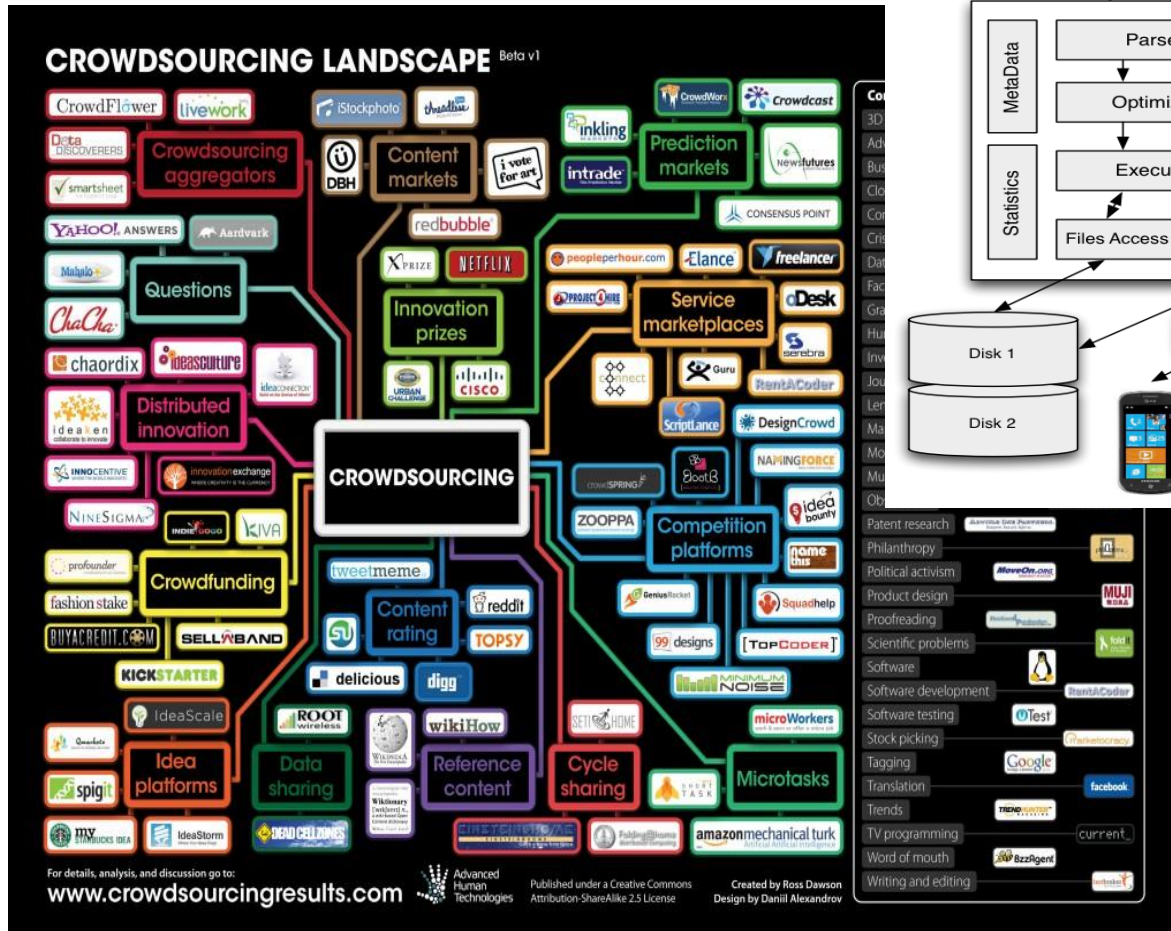
- Web Crowdsourcing Challenges
 - How to Recruit Contributors (randomly, marketplaces?) / What the Contributors Can Do (qualifications, tests)?
 - How to **Combine** their Contributions?
 - How to Manage **Abuse**?
 - How To **Scale/Manage** Complex/Larger Tasks?
 - **Openness** / Quality?
 - **Disclosure Issues** (Privacy related to Tasks, NDAs?)
 - **Minimum Wages** & Social Contributions?
- Many **open questions** that **can not** be answered through this **overview seminar** to the field.
- We will attempt to **provide an intuition** to these by using an **example-driven** approach.

Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing systems on the World-Wide Web. *Commun. ACM* 54, 4 (April 2011), 86-96.

Web Crowdsourcing Landscape



Industrial:



Academic:
Databases
WWW
IR

Courtesy of: <http://www.qualitativemind.com/trend-tamer/crowdsourcing/>

Previous CS Tutorials



Below tackle Information Management perspectives

- **VLDB'11:** *"Crowdsourcing Applications and Platforms: A Data Management Perspective"*, AnHai Doan, Michael J. Franklin, Donald Kossmann, Tim Kraska.
 - Anhai Doan, Raghu Ramakrishnan, and Alon Y. Halevy. 2011. Crowdsourcing systems on the World-Wide Web. Commun. ACM 54, 4 (April 2011), 86-96.
- **WWW'11:** *"Managing crowdsourced human computation: a tutorial"*, Panagiotis G. Ipeirotis, Praveen K. Paritosh
- **SIGIR'11:** *"Crowdsourcing for information retrieval: principles, methods, and applications"*, Omar Alonso and Matthew Lease.
- **SIGMOD'12:** *"Mob Data Sourcing"* D. Deutch, T. Milo.

Types of Crowdsourcing



- **Implicit crowdsourcing:** users solve a problem as a side effect (passively) of something else they are doing.
 - **Standalone** (solve CAPTCHAs) | **Piggyback** (spell correction improvement **from search traces**)
- **Explicit Crowdsourcing:** users work together (actively) to evaluate, share, and build specific tasks.
 - Crowdvoting, Crowdsourcing creative work
 - Crowdsourcing language-related data collection
 - Crowdfunding, "Wisdom of the crowd", Microwork, Declarative Crowdsourcing, Contests, etc.

Implicit Crowdsourcing

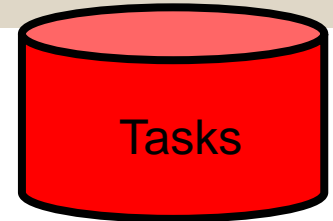
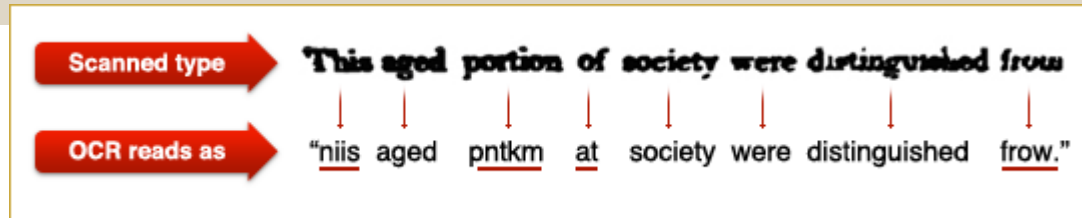
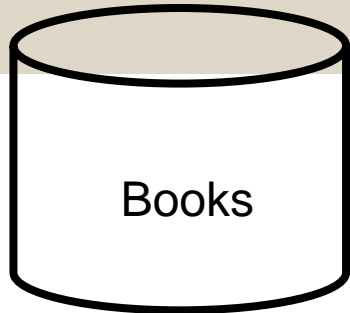


- **reCAPTCHA: Stop Spam, Read Books.**
 - Used by websites to prevent abuse.
 - **Implicit (background) task:** digitize books, newspapers and old time radio
 - Started as a CMU Project
 - Over 200M CAPTCHAs solved per day (i.e., 150,000 hours / day)

known



Implicit Crowdsourcing



- **reCAPTCHA how?**

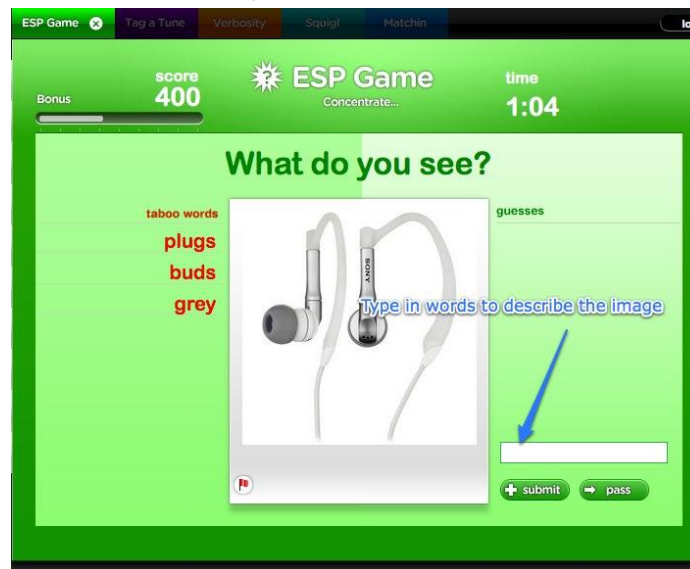


- **Service:** Provides 2 words to each user
 - 1 word is known to the service and 1 unknown
- **User:** Types in both words
- **Service:**
 - Answer to known word is taken as an indication of a "correct" answer.
 - "Aggregating" several "correct" answers yields the final correct answer.

Implicit Crowdsourcing



- **Gwap ESP Game: Play while Labeling Images (GWAP: Game With A Purpose)**
 - **Implicit (background) task:** Image Recognition / Create useful metadata for images.
 - **Application:** Better image search
 - Started as a CMU Project but licensed to Google



Implicit Crowdsourcing



- **ESP Game how?**

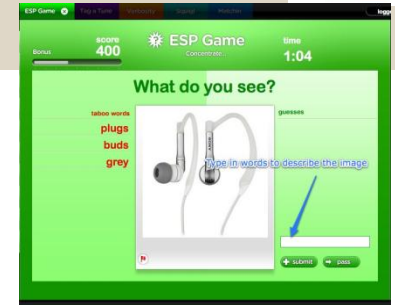
- **Service:** Provides image to 2 users

- **User:**

- Enter possible word (label) for image seen
 - Once a word is entered by both partners (not necessarily at the same time) that word becomes a label for the image
 - Once they agree on a word, they are shown another image.
 - They have two and a half minutes to label 15 images.

- **Service:**

- **Quality:** Might provide “taboo” words, labels from previous rounds that cannot be entered as possible labels.
 - **Spamming:** Provide test images to thwart spamming. Only store answers if tests were answered and if N have agreed upon.



Explicit Crowdsourcing / Contests

[TOPCODER][®]

- TopCoder is a company which **administers contests in computer programming founded in 2001** (well before the crowdsourcing term was coined)
- The work in **design and development produces useful software** which is **licensed for profit** by TopCoder.
- **Competitors** involved in the creation of these components are **paid royalties** based on these **sales**.

Explicit Crowdsourcing / Contests

- **How do I decide how much money to offer “the crowd”?**
 - Topcoder suggests to use Historic data (completed competitions) *Ebook: 10 Burning Questions on Crowdsourcing, Topcoder.com*
- **Price Examples:**
 - **Netflix Contest** – Improve Recommendation Algorithm – Price \$1M.
 - **Mastercard – Data Cleansing Competition @ Kaggle** – Price: \$100K (private competition closed to participants high in leaderboard).
 - **KDD'13 Cup Competition** – Identify which authors correspond to the same author – MS Academic Search Data – Price: \$7.5K
 - **Yelp Recommendation Service Recruiting Competition @ Kaggle** – Find useful votes for a review! – Price: Fast track Job Interview!



Explicit Crowdsourcing / Contests

[TOPCODER]

- How long should my contests run for?
 - **Contest that are either too short or too long in duration hinder participation!**
 - Algorithms: 2 hours - Code, Submit, Challenge Others Code! (+50pts for successful challenge, -25pts for unsuccessful challenge)
 - Software Design Specification from User Requirements: 1 week.
 - Development of Software Specs: 1 week.
 - Marathon Matches - Complex algorithmic problems: 1-2 weeks
- How do I pick up the winners?
 - Use **quantitative metrics** (e.g., time, throughput, etc.)
 - **Fairness** and **Consistency** in judging is paramount.
(subjective metrics are difficult to adhere to fairness standards).

Explicit Crowdsourcing / Microwork



- **Microwork** is a small task users do for low amounts of money, for which computers lack aptitude. Managed through some platform.

- **Amazon Mechanical Turk (mTurk)**



- "The Turk," a chess-playing "automaton" (hoax) of the 18th century chess master hidden in a special compartment)
 - >1M workers, 250K hits available, still in beta, less than 10% of market (the lion share in virtual currencies of games)!
 - Mostly for Language and Linguistic tasks (translate, transcribe, annotate, experiments...) http://www.crowdscientist.com/wp-content/uploads/2011/08/start_of_the_art.pdf
 - **Jan. 2007:** users searched satellite images for images of a boat in order to find Jim Gray.
- **Other Platforms: Crowdfunder, CloudCrowd, Livework, Clickworker, SmartSheet, uTest, ...**

Explicit Crowdsourcing / Microwork



Select Human Intelligence Task: Little Test

amazonmechanical turk
Artificial Artificial Intelligence

Your Account **HITS** Qualifications 253,412 HITS available now

All HITS | HITS Available To You | HITS Assigned To You

Find containing that pay at least \$ ☐ for which you ☐ require Master

All HITS

1-10 of 32004 Results

Sort by:

[Show all details](#) | [Hide all details](#)

Each worker solves a Hit once (3-5 assignment per hit) to enable majority voting (groups also available)

Classify Arabic Tweets Dialects SEE REVISED HIT

Requester: Chris Callison-Burch	HIT Expiration Date: Aug 4, 2013 (9 weeks 6 days)	Reward: \$0.00
	Time Allotted: 60 minutes	HITS Available: 23530

Identify whether the phrase/keywords belongs to the category provided?

Requester: InterestProfiler	HIT Expiration Date: Jun 21, 2013 (3 weeks 4 days)	Reward: \$0.01
	Time Allotted: 10 minutes	HITS Available: 17969

Search: Keywords on Google.com (US)

Requester: CrowdSource	HIT Expiration Date: May 27, 2014 (52 weeks)	Reward:
	Time Allotted: 16 minutes	HITS Available: 14969

Price

Explicit Crowdsourcing / Microwork



Microworks with not-so-micro prices!

Timer: 00:00:00 of 3 days

Want to work on this HIT?

Accept HIT

Transcribe a 40 minute mp3 audio file of Italian conversation

Requester: Transcribing Larcobaleno

Qualifications Required: Masters has been granted

Reward: \$40.00 per HIT

Listen to a 40 minute Italian conversation and transcribe what is said.

- Listen to a 40 minute mp3 audio file and transcribe what is said in Italian.
- Do not include "hmm" and "errs" in the transcription.
- Do not correct for grammar mistakes but transcribe as spoken.
- Use punctuation where appropriate please.
- Indicate different speakers with "speaker 1" or "speaker 2".
- Audio file is hosted at: http://annabanana.toppingdesign.com/AmbraMay25_2013.mp3
- Must be fluent in Italian to accept job.

Qualifications

Explicit Crowdsourcing / Microwork



- **Requester Best Practices**

- **Divide Project into Steps** that can be parallelized by workers (e.g., collect address, phone, owner for one provided company)
- **Keep HITs focused** (e.g., categorize, verify, etc.) to match the user skill (qualification)
- Be specific and **concise with instructions** (e.g., "is photo offensive?" Vs. "Does the photo contain nudity?")
- **Pay fairly** and the same amount for all HITs in a project (otherwise certain HITs may remain unworked)

Requester Best Practices Guide: mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf

CS Types: Explicit / Declarative



- **Declarative Crowdsourcing:** express the logic of the expected crowdsourcing task without describing its control flow (e.g., SQL or Map-Reduce like languages)
 - e.g., the following DDL + DML created at VLDB'11 a HIT requesting scenario through which Mturk workers to fill in the abstract for \$0.03 in 60 minutes:

```
CREATE TABLE Talk (  
  title STRING PRIMARY KEY,  
  abstract CROWD STRING, #default: CNULL  
  nb_attendees CROWD INTEGER  
); # Crowd Tables: capture complete record
```

```
SELECT abstract  
FROM talk  
WHERE title = "CrowdDB";
```

fuzzy for a DB query

amazonmechanicalturk
Your Account | HITs | Qualifications | 95,437 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
Timer: 00:00:00 of 60 minutes Want to work on this HIT? Total Earned: \$0.14 Total HITs Submitted: 16
CrowdBDB
Requester: AMPLab
Qualifications Required: HIT approval rate (%) is not less than 95 Reward: \$0.03 per HIT HITs Available: 1 Duration: 60 minutes
Please fill out the missing VLDB Talk data
Title: CrowdDB
Abstract: []
You must ACCEPT the HIT before you can submit the results.
Want to work on this HIT? Report this HIT: violates the Amazon Mechanical Turk policies or broken (Why?)
FAQ | Contact Us | Careers at Amazon | Developers | Press | Policies | Blog
©2005-2011 Amazon.com, Inc. or its Affiliates An amazon.com company

Explicit Crowdsourcing / Declarative



- **Qurk (MIT) – [SQL]**

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

- **CrowdDB (Berkeley and ETH Zurich) [SQL]**

- CrowdDB: Answering Queries with Crowdsourcing, M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin, SIGMOD'11 & VLDB'11 Demo
- *"Crowdsourcing Applications and Platforms: A Data Management Perspective"*, AnHai Doan, Michael J. Franklin, Donald Kossmann, Tim Kraska., VLDB'11 Tutorial

Overviewed next

- **Deco (Stanford and UCSC)[SQL]**

- Deco: Declarative Crowdsourcing, A. Parameswaran, H. Park, H.G. Molina, N. Polyzotis, J. Widom, Stanford Infolab Technical Report, 2011

- **MoDaS (Tel Aviv University)**

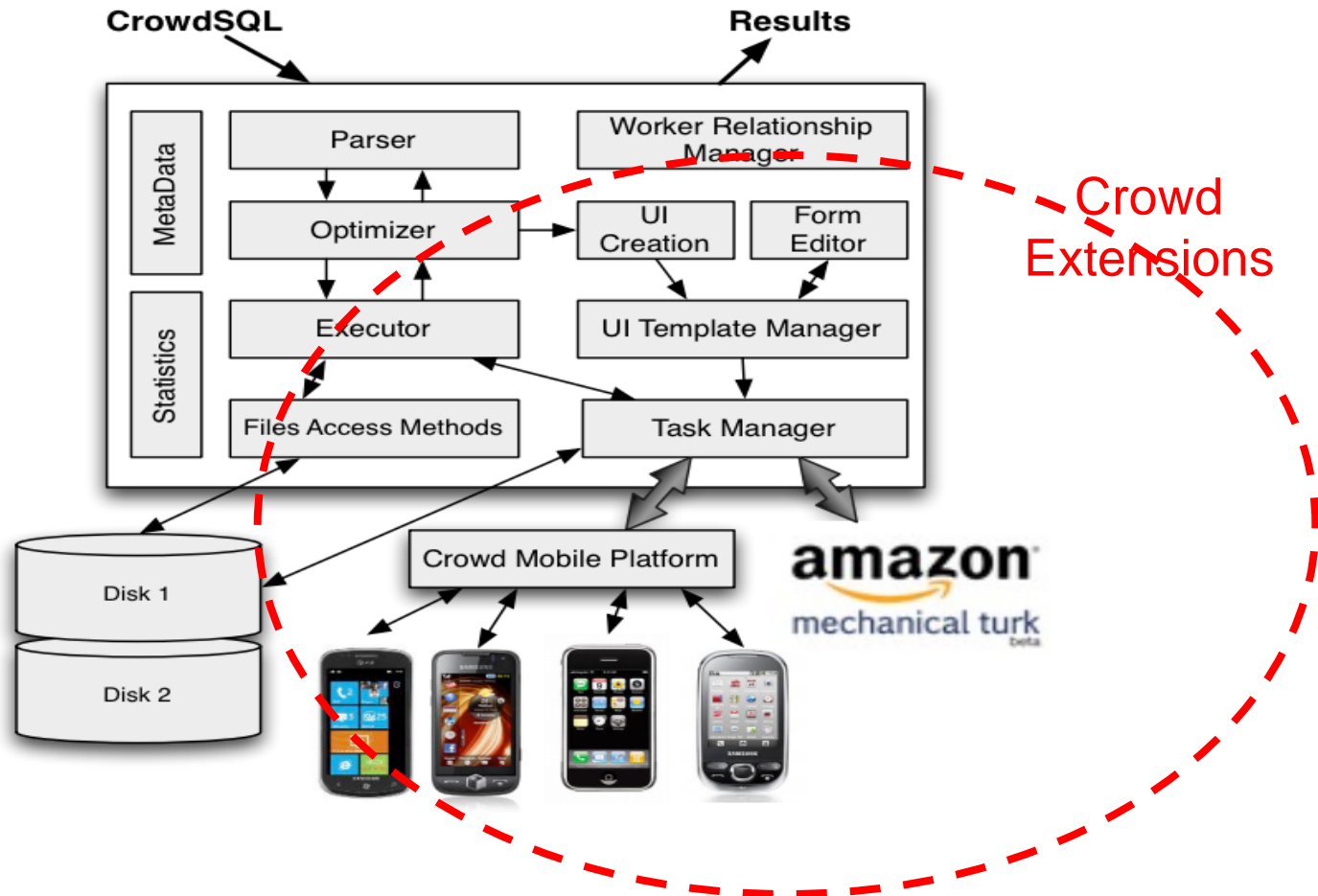
- SIGMOD'12: *"Mob Data Sourcing" (tutorial)* D. Deutch, T. Milo.

- **CrowdForge (CMU) – [MapReduce]**

Explicit Crowdsourcing / Declarative



- CrowdDB Architecture



Explicit Crowdsourcing / Declarative



- **CrowdSQL in Action:**

- Finding Missing Data / Generate GUIs on-the-fly

- `SELECT * FROM companies WHERE name = "IBM";`

- Fuzzy Matching (subjective comparison)

- `SELECT * FROM companies WHERE`
 - `name ~="IBM";` (nor regex-oriented LIKE)

Are the following entities the same?

IBM == Big Blue

(b) CROWDEQUAL

- Fuzzy Ranking (subjective):

- `SELECT image FROM pictures`
 - `ORDER BY novel_idea LIMIT 10`

Which picture visualizes better
"Golden Gate Bridge"

☒ ☐

- Fuzzy Aggregation (subjective), Join and π
typical DBMS operators (group-by, index scans, etc)

- CrowdDB: Answering Queries with Crowdsourcing, M. J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin, SIGMOD'11 & VLDB'11 Demo

Explicit Crowdsourcing / Declarative



- **Challenges for Declarative CS :**
 - Mobile Issues (VLDB'11 Demo)
 - Quality Assessment & Improvem.
 - Latency & Scheduling
 - Cost Optimization
- **Other Applications:** Structuring Data, Linking Data, Schema Matching, Graph Search, ...

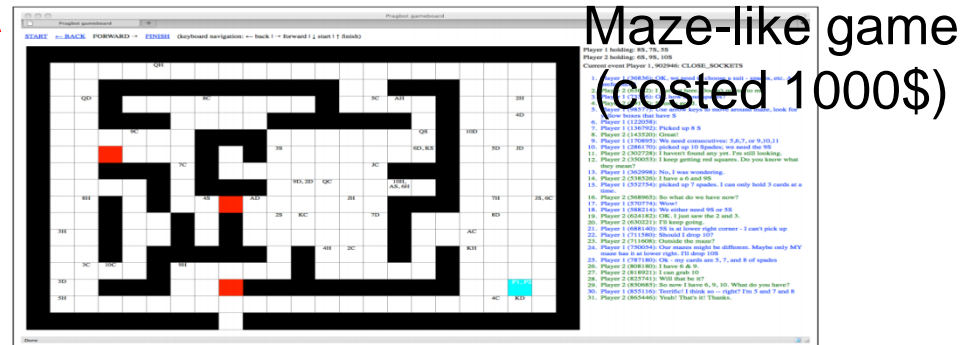


- *Crowdsourcing Applications and Platforms: A Data Management Perspective*, AnHai Doan, Michael J. Franklin, Donald Kossmann, Tim Kraska., VLDB'11 Tutorial

Explicit Crowdsourcing / Wisdom-of-the-Crowd



- **Wisdom-of-the-Crowd** is the process of taking into account the **collective opinion** of a **group of individuals** rather than a **single expert** to answer a question.
 - Collective better than Individual Intelligence!



David Clausen and Christopher Potts (Stanford), Collecting task-oriented dialogues, Workshop on Crowdsourcing Technologies for Language and Cognition Studies, Boulder, July 27, 2011 (aim: to study how people coordinate with chat to solve a problem: 1\$ payment + \$0.5 bonus)



Explicit Crowdsourcing / Wisdom-of-the-Crowd

- Dec. 2009, 10 red **weather balloons** were deployed from locations throughout the USA by DARPA
- **Task:** Find balloon coordinates the quickest.
- **Price:** 40,000 \$!
- **Solution:** Referral marketing (like magazine subscription) solution by MIT!
 - Solution: 2000\$ finder, 1000\$ inviter, 500\$ inviter of inviter, etc., i.e., $n + n/2 + \dots + 2 + 1 = 2^n - 1 = 2^8 - 1 = 255$ / balloon = 39,990 for all ten balloons
 - Invitee not necessary to see the balloon to get the price!
 - After 8 hours and 52 minutes all balloons were found
 - Runner-up: Georgia Tech, Used Twitter with Altruism incentive (i.e., donate prize money to the American Red Cross) but that incentive did not work the same well.



Source: <http://web.mit.edu/newsoffice/2011/red-balloons-study-102811.html>

Explicit Crowdsourcing / Crowdfunding



- Crowdfunding is the process of funding your projects by a **multitude of people** contributing a **small amount** in order to attain a certain **monetary goal**.

KICKSTARTER

BUYACREDIT.COM

KIVA



Explicit Crowdsourcing / Crowdfunding

Kickstarter's Microfinance Workflow: **KICKSTARTER**

- **Requester** choose a deadline (up-to-60-days) and a minimum funding goal.
- **Workers:** "pledge" projects (min:25\$, avg:70\$) and select some **tangible reward** (e.g., limited edition CD, custom T-shirt, initial run pledged at retail price, etc.)
- US account necessary to acquire funding
- Kickstarter charges 5% + 3-5% by Amazon Payments (only if project reaches funding goal)
- No guarantee that product will realize
 - Oct 2012: 73,620 projects (mostly technology and gaming) with 43% succeeding their funding goal, \$380M raised
 - 11% never receive a single pledge.

Explicit Crowdsourcing / Crowdfunding




- First project raising over \$1M

KICKSTARTER

Elevation Dock: The Best Dock For iPhone
by Casey Hopkins + ElevationLab

Home Updates **23** Backers **12,521** Comments **2,445** Portland, OR Product Design

Funded! This project successfully raised its funding goal on Feb 11, 2012.



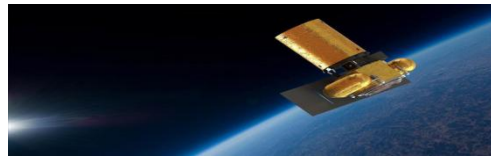
12,521
backers
\$1,464,706
pledged of \$75,000 goal
0
seconds to go

Funding period
Dec 13, 2011 - Feb 11, 2012 (60 days)

Project by
Casey Hopkins + ElevationLab
Portland, OR
[Contact me](#)

CS Types: Explicit / Crowdfunding

- **Bogota Colombia:** BD Bacatá project raised over \$200 M to build largest skyscraper in Colombia's history
- **Crowd:** More than 3,500 investors, allowing them to take part in "claimed" **profitable project** previously accessible to superrich people.
- **Other Projects:**
 - Crowdfund a whole city!
 - Crowdfund a space telescope (currently 96K pledged, needs 1B!, incentives: shot video into space)



<http://www.bdbacata.com/newsite/>

Explicit Crowdsourcing / Q&A



- Q&A sites are **interest-based** social portals that allow users **question, answer, edit and organize** the constructed information

- **Task:** Question
- **Price:** Ethical Benefit
 - Socialize & Fun
 - Earn Prestige



Mobile extend



Explicit Crowdsourcing / Buy-Sell Skills



- People **buy** or **sell** (outsource) **skills** through an **open-call**, which is propagated to an undefined crowd (e.g., mobile users on the go).
- 250,000 active users
 - 180,000 freelancers, 70,000 clients.
 - Wired UK: "Europe's 100 Hottest Startups of 2012"
- Similar Marketplaces:
 - Elance (since '99: \$772M with 2M freelancers),
 - ODesk and Freelancer.com

