Crowdsourcing for Data Management

Gianluca Demartini
University of Queensland

http://gianlucademartini.net
@eglu81
Gianluca Demartini

• B.Sc., M.Sc. at U. of Udine, Italy
• Ph.D. at U. of Hannover, Germany
  • Entity Retrieval
• Worked at the University of Sheffield (UK), eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
• Senior Lecturer in Data Science at the University of Queensland, since 2017.
• Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013, ISWC 2013, ICWSM 2016, WebSci 2016, Facebook

www.gianluca demartini.net
Research Interests

• **Entity-centric Information Access** (2005-now)
  - Structured/Unstructured data (SIGIR 12), TRank (ISWC 13, WSemJ 16)
  - NER in Scientific Docs (WWW 14), Prepositions (CIKM 14)
  - IR Evaluation (CIKM 2017, ECIR 16 Best Paper Award, IRJ 2015)

• **Hybrid Human-Machine Systems** (2012-now)
  - ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  - Human Memory based Systems (WWW 14, PVLDB)
  - Hybrid systems overview (COMNET, 2015)

• **Better Crowdsourcing Platforms** (2013-now)
  - Platform Dynamics (WWW 15)
  - Pick-a-Crowd (WWW 13), **Malicious Workers** (CHI 15)
  - Scale-up Crowdsourcing (HCOMP 14), Scheduling (WWW 16)
  - **Timeout** (HCOMP 16), **Complexity** (HCOMP 16)
Course Outline

• Micro-task Crowdsourcing
  • Examples
  • Dimensions
  • Platforms (Amazon MTurk)

• Hybrid human-machine systems
  • Examples
  • Challenges: Efficiency / Effectiveness

• Effectiveness
  • Quality control
  • Task assignment

• Efficiency
  • Scheduling
  • Pricing / Timeouts
Crowdsourcing

from http://www.bbc.co.uk/news/magazine-32993891
Crowdsourcing

• *Portmanteau* of "crowd" and "outsourcing," first coined by Jeff Howe in a June 2006 Wired magazine article

• [Merriam-Webster] the practice of obtaining needed services, ideas, or content by soliciting *contributions from a large group of people* and especially from the online community rather than from traditional employees or suppliers
Crowdsourcing

• "Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential laborers."

[Howe, 2006]
Incentives in Crowdsourcing

- **Extrinsic motivation** if task is considered boring, dangerous, useless, socially undesirable, dislikable by the performer.
  - Paid Crowdsourcing

- **Intrinsic motivation** is driven by an interest or enjoyment in the task itself.
  - Fun (enjoyment) / Games with a purpose
  - Community (belonging, desire to help)
  - Citizen Science
Dimensions of Human Computation

[Quinn & Bederson, 2012]

What is outsourced

• Tasks based on human skills not easily replicable by machines (visual recognition, language understanding, knowledge acquisition, basic human communication etc)

Who is the crowd

• Open call
• Call may target specific skills and expertise
• Requester typically knows less about the workers than in other work environments

How is the task outsourced

• Explicit vs. implicit participation
• Tasks broken down into smaller units undertaken in parallel by different people
• Coordination required to handle cases with more complex workflows
• Partial or independent answers consolidated and aggregated into complete solution
Dimensions of Human Computation (2)

[Quinn & Bederson, 2012]

**How are the results validated**
- Solutions space closed vs. open
- Performance measurements/ground truth
- Statistical techniques employed to predict accurate solutions
- May take into account confidence values of algorithmically generated solutions

**How can the process be optimized**
- Incentives and motivators
- Assigning tasks to people based on their skills and performance (as opposed to random assignments)
- Symbiotic combinations of human- and machine-driven computation, including combinations of different forms of crowdsourcing
Games with a Purpose

• Tasks leveraging common human skills, appealing to large audiences
  • Selection of domain and task more constrained in games to create typical UX
• Tasks decomposed into smaller units of work to be solved independently
• Complex workflows
  • Creating a casual game experience vs. patterns in microtasks
  • Single vs. multi-player
Paid Micro-Task Crowdsourcing

A Crowdsourcing Platform allows requesters to publish a crowdsourcing request (batch) composed of multiple tasks (HITs).

Programmatically Invoke the crowd with APIs or using a website.

Workers in the crowd complete tasks and obtain a monetary reward.

The platform takes a fee (30% of the reward).
Example use of micro-task crowdsourcing

• Relevance judgments
• Ontologies
• Sentiment Analysis in Social Media
• http://www.thesheepmarket.com/

"draw a sheep facing to the left."
Case-Study: Amazon MTurk

• Micro-task crowdsourcing marketplace
• On-demand, scalable, real-time workforce
• Online since 2005 (still in “beta”)
• Currently the most popular platform
• Developer’s API as well as GUI
Amazon MTurk

Make Money by working on HITs

HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:
- Can work from home
- Choose your own work hours
- Get paid for doing good work

Get Results from Mechanical Turk Workers

Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now

As a Mechanical Turk Requester you:
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results
MTurk is a Marketplace for HITs
You must **accept** this HIT before working on it.

**Data Collection Instructions!**

Find the postal address for this Australian company.

- Search on Google, the company’s website, YellowPages or Facebook to find the correct postal address for the company below.
- Enter the *full Australian postal address* for the business.
- You may use the research links provided to help.
- **Do not enter incomplete or incorrect details!**

<table>
<thead>
<tr>
<th>Company name:</th>
<th>Stellar Electrical And Solar Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location:</td>
<td>Australia</td>
</tr>
<tr>
<td>Company website:</td>
<td></td>
</tr>
<tr>
<td>Company YellowPages:</td>
<td></td>
</tr>
<tr>
<td>Company Facebook:</td>
<td></td>
</tr>
</tbody>
</table>

**Australian Street Address (ONLY this field is required if complete):**

Start typing Australian Street Address...
You must accept this HIT before working on it.

Receipt Transcription Instructions (Click to expand)

<table>
<thead>
<tr>
<th>Item Description</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subscription to Quip Business - Monthly</td>
<td>$108.00</td>
</tr>
<tr>
<td>Subtotal</td>
<td>$108.00</td>
</tr>
<tr>
<td>Coupons</td>
<td>-$30.00</td>
</tr>
<tr>
<td>Credit for first five users - QUIP3120 ($30.00 off)</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$78.00</td>
</tr>
</tbody>
</table>

Is the receipt legible?
- Legible
- NOT legible

Issuer name: Company Inc.

Invoice number: IV2348977374

Invoice Date: 2017-05-13

Currency (3 digits): USD / EUR / ...

Content and Cost:
- Content 1: 0.00
Crowdsourcing Ontology Mapping

• Find a set of mappings between two ontologies

• Micro-tasks:
  • Verify/identify a mapping relationships:
    • Is concept A the same as concept B
    • A is a kind of B
    • B is a kind of A
    • No relation

Crowdsourcing Ontology Mapping

• Crowd-based outperforms purely automatic approaches
Crowdsourcing Ontology Engineering

• Ask the crowd to create/verify subClassOf relations
  • “Car” is a “vehicle”

• Does it work for domain specific ontologies?
  • A “protandrous hermaphroditic organism” is a “sequential hermaphroditic organism”

• Workers perform worse than experts

• Workers presented with concept definitions perform as good as experts

Jonathan Mortensen, Mark A. Musen, Natasha F. Noy: Crowdsourcing the Verification of Relationships in Biomedical Ontologies. AMIA 2013
MTurk is a Marketplace for HITs

<table>
<thead>
<tr>
<th>Requester name</th>
<th>hits</th>
<th>reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speechpad</td>
<td>23857</td>
<td>$172,994.63</td>
</tr>
<tr>
<td>Percy Liang</td>
<td>883</td>
<td>$7,320.48</td>
</tr>
<tr>
<td>Princeton Vision</td>
<td>51187</td>
<td>$5,762.44</td>
</tr>
<tr>
<td>Stanford GSB Behavioral Lab</td>
<td>3749</td>
<td>$2,110.70</td>
</tr>
<tr>
<td>Chris Callison-Burch</td>
<td>8157</td>
<td>$2,064.29</td>
</tr>
<tr>
<td>RC.org Mechanical Turk</td>
<td>6591</td>
<td>$2,011.33</td>
</tr>
<tr>
<td>VacationrentalAPI</td>
<td>399</td>
<td>$1,373.50</td>
</tr>
<tr>
<td>Med Expertise</td>
<td>869</td>
<td>$1,303.50</td>
</tr>
<tr>
<td>Bluejay Labs</td>
<td>13613</td>
<td>$1,288.59</td>
</tr>
<tr>
<td>YL Testing</td>
<td>1051</td>
<td>$1,236.83</td>
</tr>
</tbody>
</table>
Demographics of MTurk workers in 2009

Country of residence
- United States: 46.80%
- India: 34.00%
- Miscellaneous: 19.20%

2013 Statistics:
1M workers
10% active
Distribution of *Batch Size*

"Power-law"
Reward Distribution
Distribution of HIT Types

Less Content Access batches

Content Creation being the most popular
Some findings of our longitudinal study

• HIT reward has increased over time
• Audio transcription is the most popular task
• Demand for Indian workers has decreased
• Surveys are most popular for US workers
• 1000 new requesters per month join
• 10K new HITs arrive and 7.5K HITs get completed every hour

• Check #mturkdynamics for the main findings
Hybrid Human-Machine Systems

• Use Machines to scale over large amounts of data
• Keep humans in the loop
  • By means of Crowdsourcing
  • To make sure the quality of the data processing is good
• Crowd for Pre-processing vs Post-processing

Hybrid Image Search

Yan, Kumar, Ganesan, CrowdSearch: Exploiting Crowds for Accurate Real-time Image Search on Mobile Phones, Mobisys 2010.
Example: Hybrid Data Integration

- **Generate plausible matches**
  - paper = title, paper = author, paper = email, paper = venue
  - conf = title, conf = author, conf = email, conf = venue

- **Ask users to verify**

<table>
<thead>
<tr>
<th>paper</th>
<th>conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration</td>
<td>VLDB-01</td>
</tr>
<tr>
<td>Data mining</td>
<td>SIGMOD-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>email</th>
<th>venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLAP</td>
<td>Mike</td>
<td>mike@a</td>
<td>ICDE-02</td>
</tr>
<tr>
<td>Social media</td>
<td>Jane</td>
<td>jane@b</td>
<td>PODS-05</td>
</tr>
</tbody>
</table>

Does attribute **paper** match attribute **author**?

<table>
<thead>
<tr>
<th>paper</th>
<th>conf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data integration</td>
<td>VLDB-01</td>
</tr>
<tr>
<td>Data mining</td>
<td>SIGMOD-02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>title</th>
<th>author</th>
<th>email</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLAP</td>
<td>Mike</td>
<td>mike@a</td>
</tr>
<tr>
<td>Social media</td>
<td>Jane</td>
<td>jane@b</td>
</tr>
</tbody>
</table>

Yes  No  Not sure
CrowdDB

Use the crowd to answer DB-hard queries

Where to use the crowd:
- Find missing data
- Make subjective comparisons
- Recognize patterns

But not:
- Anything the computer already does well

M. Franklin, D. Kossmann, T. Kraska, S. Ramesh and R. Xin.
CrowdDB: Answering Queries with Crowdsourcing, SIGMOD 2011
Crowd DB query

The Volvo S80 is the flagship model of this brand...

Is the review positive?

SELECT review
FROM car_review
WHERE sentiment ~= "pos";

Which one is better?

SELECT image i
FROM car_image
WHERE subject = "Volvo S60"
ORDER BY CROWDORDER("clarity");
CrowdDB – Missing Data

**Missing Columns**

<table>
<thead>
<tr>
<th>review</th>
<th>make</th>
<th>model</th>
<th>sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>Volvo</td>
<td>S80</td>
<td>?</td>
</tr>
</tbody>
</table>

**Missing Tuples**

<table>
<thead>
<tr>
<th>make</th>
<th>model</th>
<th>style</th>
<th>color</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

CREATE TABLE car_review
(
  review STRING,
  make STRING,
  model STRING,
  sentiment STRING
);

CREATE CROWD TABLE car
(
  make STRING,
  model STRING,
  color STRING,
  style STRING,
  PRIMARY KEY (make, model)
);
CrowdDB - Joins and Sorts

Are the following entities the same?

IBM == Big Blue

Yes  No

Which picture visualizes better "Golden Gate Bridge"

Submit
Crowdsourcing for Entity Linking
Facebook is not waiting for its initial public offering to make its first big purchase.

In its largest acquisition to date, the social network has purchased Instagram, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday.

HTML:
<p>Facebook is not waiting for its initial public offering to make its first big purchase.</p>

RDFa enrichment

```html
<p><span about="http://dbpedia.org/resource/Facebook" cite property="rdfs:label">Facebook</span> is not waiting for its initial public offering to make its first big purchase. <span about="http://dbpedia.org/resource/Instagram" cite property="rdfs:label">Instagram</span>, the popular photo-sharing application, for about $1 billion in cash and stock, the company said Monday. </p>
```
ZenCrowd

• Combine both algorithmic and manual linking
• Automate manual linking via crowdsourcing
• Dynamically assess human workers with a probabilistic reasoning framework
ZenCrowd Architecture

Entity Factor Graphs

- **Graph components**
  - Workers, links, clicks
  - Prior probabilities
  - Link Factors
  - Constraints

- **Probabilistic Inference**
  - Select all links with posterior prob $\geq \tau$

2 workers, 6 clicks, 3 candidate links
Experimental Evaluation

• Worker Selection

![Graph showing worker precision vs. number of tasks for US and IN workers.]
ZenCrowd Summary

• ZenCrowd: Probabilistic reasoning over automatic and crowdsourcing methods for entity linking
• Standard crowdsourcing improves 6% over automatic
• 4% - 35% improvement over standard crowdsourcing
• 14% average improvement over automatic approaches

• Follow up-work (VLDBJ, 2013):
  • Also used for instance matching across datasets
  • 3-way blocking with the crowd
Blocking for Instance Matching

- Find the instances about the same real-world entity within two datasets
- Avoid Comparison of all possible pairs
  - Step 1: cluster similar items using a cheap similarity measure
  - Step 2: n*n comparison within the clusters with an expensive measure
3-steps Blocking with the Crowd

• Crowdsourcing as the most expensive similarity measure
Lessons Learnt

• Crowdsourcing + Prob reasoning works!
• But
  • Different worker communities perform differently
  • Many low quality workers
  • Completion time may vary (based on reward)
• Need to **find the right workers** for your task (see WWW2013 and CHI2015 papers)
• Need to make sure **high priority tasks** are completed fast (see WWW2016 paper)
Behavioral Patterns of Malicious Workers

**Ineligible Workers (IW)**

*Instruction:* Please attempt this microtask ONLY IF you have successfully completed 5 microtasks previously.
*Response:* ‘this is my first task’

eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.

*Response:* ‘What’s your task?’ , ‘adasd’, ‘fgfgf gsd lijlkj’

**Fast Deceivers (FD)**

*Instruction:* Identify 5 keywords that represent this task (separated by commas).
*Response:* ‘survey, tasks, history’, ‘previous task yellow’

**Rule Breakers (RB)**

*Instruction:* Identify 5 keywords that represent this task (separated by commas).
*Response:* ‘one, two, three, four, five’

**Smart Deceivers (SD)**

These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions!

Turker Contribution and Errors

[Franklin, Kossmann, Kraska, Ramesh, Xin: CrowdDB: Answering Queries with Crowdsourcing. SIGMOD, 2011]
Scheduling HITs

Platform throughput is dominated by large HIT batches.
**HIT-Bundle**

**Definition**

- Scheduling requires control over the serving process of tasks
- A **HIT-Bundle** is a batch that contains heterogeneous tasks
- All tasks that are generated by the system are published through the HIT-Bundle
Scheduling HITs

• Fair Scheduling
  • Priority of HITs but avoid starvation
  • Assign HITs of the same type (no context switch)

## Overview of hybrid systems

<table>
<thead>
<tr>
<th>Year</th>
<th>Cit.</th>
<th>Domain</th>
<th>Data Type</th>
<th>Human role</th>
<th>Incentive</th>
<th>Time constrains</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>62</td>
<td>Web</td>
<td>Images</td>
<td>Pre-p.</td>
<td>Fun</td>
<td>Batch</td>
</tr>
<tr>
<td>2007</td>
<td>35</td>
<td>Science</td>
<td>Images</td>
<td>Pre-p.</td>
<td>Community</td>
<td>Batch</td>
</tr>
<tr>
<td>2008</td>
<td>64</td>
<td>Web</td>
<td>Images</td>
<td>Post-p.</td>
<td>Access</td>
<td>Batch</td>
</tr>
<tr>
<td>2011</td>
<td>52</td>
<td>Database</td>
<td>Graph</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2011</td>
<td>30</td>
<td>Database</td>
<td>Struct. data</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2011</td>
<td>5</td>
<td>Filtering</td>
<td>Video</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2012</td>
<td>54</td>
<td>Database</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2012</td>
<td>19</td>
<td>Web</td>
<td>Unstruct. text</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2012</td>
<td>56</td>
<td>Data Integration</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2012</td>
<td>66</td>
<td>Entity Resolution</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2012</td>
<td>68</td>
<td>Entity Resolution</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2012</td>
<td>8</td>
<td>Search</td>
<td>Unstruct. text</td>
<td>Post-p.</td>
<td>Community</td>
<td>Real-time</td>
</tr>
<tr>
<td>2012</td>
<td>42</td>
<td>Captioning</td>
<td>Video</td>
<td>Pre-p.</td>
<td>Community</td>
<td>Real-time</td>
</tr>
<tr>
<td>2013</td>
<td>34</td>
<td>Info Extraction</td>
<td>Unstruct. text</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>20</td>
<td>Entity Resolution</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>67</td>
<td>Entity Resolution</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>21</td>
<td>Database</td>
<td>Struct. data</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>44</td>
<td>Database</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2013</td>
<td>48</td>
<td>Biomedical</td>
<td>Ontology</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
<tr>
<td>2013</td>
<td>43</td>
<td>Personal assistance</td>
<td>Unstruct. text</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2013</td>
<td>27</td>
<td>Biomedical</td>
<td>Unstruct. text</td>
<td>Post-p.</td>
<td>Fun</td>
<td>Batch</td>
</tr>
<tr>
<td>2014</td>
<td>53</td>
<td>Search</td>
<td>Image</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2014</td>
<td>49</td>
<td>Database</td>
<td>Struct. data</td>
<td>Post-p.</td>
<td>Monetary</td>
<td>Real-time</td>
</tr>
<tr>
<td>2014</td>
<td>51</td>
<td>Cult. Heritage</td>
<td>Image</td>
<td>Pre-p.</td>
<td>Monetary</td>
<td>Batch</td>
</tr>
</tbody>
</table>
Overview of hybrid systems

• Balance between systems that use the human component as pre-processing or post-processing of data (11 vs 13)
• Mostly monetary reward
• Majority of systems perform batch data processing rather than real-time jobs
• In 2014 we can observe a decreased number of hybrid human-machine systems being propose: focus on solving core problems rather than building new systems
Want to know more?

• SIGMOD 2017 Tutorial (3 hours):  

• Adam Marcus and Aditya Parameswaran. Crowdsourced data management
  industry and academic perspectives. Foundations and Trends in Databases, 
  2015.

• Gianluca Demartini. Hybrid Human-Machine Information Systems: 
  Challenges and Opportunities. In: Computer Networks, Volume 90, page 
  5-13 (2015), Elsevier.

• Edith Law and Luis von Ahn. Human Computation. Synthesis Lectures on 
  Artificial Intelligence and Machine Learning. June 2011

• “An introduction to Hybrid Human-Machine Information Systems” in 
  Foundations and Trends® in Web Science (coming soon)

• Open Research topics:  
  http://www.gianlucademartini.net/research/openquestions.html
Summary

- **Hybrid human-machine systems** can
  - Scale over large amounts of data
  - Reach high accuracy by keeping humans in the loop

- Entities are the new entry point to Web content
  - “Things not string”
  - Google Knowledge Vault (but also Bing, Yahoo!, Yandex)

- Users can benefit from **entity-centric search**, browsing, and exploration of the Web

Gianluca Demartini
http://gianlucademartini.net
@eglu81