Crowdsourcing for Data Processing and Search

Dr Gianluca Demartini
University of Sheffield
http://gianlucademartini.net
Gianluca Demartini

• BSc, MSc at U. of Udine, Italy
• PhD at U. of Hannover, Germany
  – Entity Retrieval
• Worked at the eXascale Infolab U. Fribourg (Switzerland), UC Berkeley (on Crowdsourcing), Yahoo! (Spain), L3S Research Center (Germany)
• Lecturer in Data Science at the iSchool, U. of Sheffield
• Tutorials on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 13, ISWC 13, SearchSolutions 2015

www.gianlucaademartini.net
Research Interests

• **Entity-centric Information Access** (2005-now)
  – Structured/Unstructured data (SIGIR 12), TRank (ISWC 13)
  – NER in Scientific Literature (WWW 14) Prepositions (CIKM 14)

• **Hybrid Human-Machine Systems** (2012-now)
  – ZenCrowd (WWW 12, VLDBJ), CrowdQ (CIDR 13)
  – Memory-based Information Systems (WWW 14, PVLDB)

• **Better Crowdsourcing Platforms** (2013-now)
  – Pick-a-Crowd (WWW 13), Malicious Workers (CHI 15)
  – Scale-up Crowdsourcing (HCOMP 14), Dynamics (WWW 15)
Learning Objectives

• Demonstrate an understanding of crowdsourcing applications to search problems with its opportunities as well as its limitations;

• Demonstrate knowledge of the common techniques to be used in crowdsourced task design to improve the quality of the collected data;

• Discuss how crowdsourcing can be leveraged in combination with machine-based algorithms for data processing problems and to answer complex search queries;

• Discuss the benefits and challenges of applying crowdsourcing solutions for search within the enterprise.

Slides Available here:
www.gianlucademartini.net/crowdsourcing/searchsolutions
Introductions

• Name, role

• Interest / experience in Crowdsourcing / Data Processing / Search
Tutorial Outline

• Part 1
  – Introduction to Crowdsourcing (30min)
  – Ensuring Quality in Paid Crowdsourcing (60min)

• Part 2
  – Hybrid Human-Machine Data Integration (30min)
  – Crowd-Powered Search (30min)
  – Enterprise Crowdsourcing for Search (30min)
Introduction to Crowdsourcing
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

- Leverage human intelligence at scale to solve
  - Tasks simple for humans, complex for machines
  - With a large number of humans (the Crowd)
  - Small problems: micro-tasks (Amazon MTurk)

- Examples
  - Wikipedia, Image tagging

- Incentives
  - Financial, fun, visibility

- See also longer tutorial at ISWC 2013
Crowdsourcing Incentives

• Paid Crowdsourcing
• Fun (enjoyment)
  – Gamification
• Community (belonging, desire to help)
  – For example, Wikipedia
The Way Industry Looks At It
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

Find an interesting task Work Earn money

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

Fund your account Load your tasks Get results

Find HITs Now

Get Started
Amazon MTurk

• Requesters create tasks (HITs)
• The platform takes a fee (30% of the reward)
• Workers preview, accept, submit HITs
• Requesters approve, download results

• If the results are approved, workers are paid
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
Demographics of MTurk workers in 2009

Household Income for US workers

Household Income for Indian workers

Year of Birth for US workers

Year of Birth for Indian workers
Demographics of MTurk workers in 2009

http://www.mturk-tracker.com/
For bugs reports or feature requests, please contact Panos Ipeirotis
If you want to cite this website, please cite the paper Analyzing the Amazon Mechanical Turk Marketplace, P. Ipeirotis, ACM XRDS, Vol 17, Issue 2, Winter 2010, pp 16-21.
5-year Analysis of MTurk workload

• Mturk-tracker.com
  – Collects metadata about each visible batch (Title, description, rewards, required qualifications, HITs available, etc), that is, set of similar tasks or HITs
  – Records batch progress (every ~20 minutes)
  – Covers 130M tasks

Country-Specific HITs

Workers from US, India and Canada are the most sought after.
Distribution of *Batch Size*

“Power-law”
Batch Size over time

Very large batches start to appear
How much are HITs paid?

5-cents is the new 1-cent
Requesters and Reward over time

Increasing number of New and Distinct Requesters
# One month of MTurk Requesters

## Top-1000 Requesters, report for October 25, 2015 to November 24, 2015

<table>
<thead>
<tr>
<th>Requester name</th>
<th>hits</th>
<th>reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speechpad</td>
<td>32114</td>
<td>$288,834.35</td>
</tr>
<tr>
<td>CastingWords</td>
<td>11727</td>
<td>$6,817.26</td>
</tr>
<tr>
<td>Chris Callison-Burch</td>
<td>18812</td>
<td>$5,597.21</td>
</tr>
<tr>
<td>p9r</td>
<td>76873</td>
<td>$4,239.22</td>
</tr>
<tr>
<td>Stanford GSB Behavioral Lab</td>
<td>3262</td>
<td>$2,579.85</td>
</tr>
<tr>
<td>Jon Brelig</td>
<td>46459</td>
<td>$2,483.66</td>
</tr>
<tr>
<td>Farhan Memon</td>
<td>9177</td>
<td>$1,835.40</td>
</tr>
<tr>
<td>OCMP5</td>
<td>33243</td>
<td>$1,651.25</td>
</tr>
<tr>
<td>nada hashmi</td>
<td>457</td>
<td>$1,623.00</td>
</tr>
<tr>
<td>VidAngel</td>
<td>126</td>
<td>$1,583.80</td>
</tr>
</tbody>
</table>
Top requesters

SIMPLY THE BEST HUMAN-GENERATED TRANSCRIPTIONS. DELIVERED ON TIME, EVERY TIME. GUARANTEED!

$1.00 PER MINUTE OF AUDIO OR VIDEO DELIVERED IN ONE WEEK GUARANTEED

$1.50 PER MINUTE OF AUDIO OR VIDEO DELIVERED IN 48 HOURS GUARANTEED

$3.00 PER MINUTE OF AUDIO OR VIDEO DELIVERED IN 24 HOURS GUARANTEED
Transcription made fast & easy

UPFRONT PRICING · QUALITY RESULTS

3 EASY STEPS
CHOOSE A PRODUCT TO BEGIN

BUDGET
$1.00/minute

1 WEEK
$1.50/minute

1 DAY
$2.50/minute

INTERNATIONAL
$1.75/minute
Watch Movies Your Way - However the **BLEEP** you want!

NEW RELEASES

$1

PER NIGHT WITH SELLBACK
Distribution of HIT Types

Less Content Access batches

Content Creation: the most popular

Classify HITs into types (Gadiraju et al. 2014)
- Information Finding (IF)
- Verification and Validation (VV)
- Interpretation and Analysis (IA)
- Content Creation (CC)
- Surveys (SU)
- Content Access (CA)
Is the Market Elastic?

Intercept = 2.5
Slope = 0.5%

20% of new work gets completed within an hour
Summary

• HIT reward has increased over time
• **Audio transcription**: 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 #mturk dynamics for more findings
Why Crowdsourcing for IR Evaluation?

• Easy, cheap and fast labeling
• Ready-to use infrastructure – MTurk payments, workforce, interface widgets – CrowdFlower quality control mechanisms, etc.
• Allows early, iterative, frequent experiments – Iteratively prototype and test new ideas – Try new tasks, test when you want & as you go
• Proven in major IR shared task evaluations – CLEF image, TREC, INEX, WWW/Yahoo SemSearch
Gamification of IR Evaluation

• GeAnn: http://www.geann.org/

• Relevance judgments with Gamification:
  – Text relevance
  – Image relevance

Tutorial Outline

• Part 1
  – Introduction to Crowdsourcing (30min)
  – Ensuring Quality in Paid Crowdsourcing (60min)

• Part 2
  – Hybrid Human-Machine Data Integration (30min)
  – Crowd-Powered Search (30min)
  – Enterprise Crowdsourcing for Search (30min)
Ensuring Quality in Paid Crowdsourcing
A Crowdsourcing Task

Choose the best category for this image

View Instructions

Select the room location in home for this picture. Seating areas outside are outside not living. Offices or dens are living not bedrooms. Bedrooms should contain a bed in the picture.

- kitchen
- living
- bath
- bed
- outside
High-level Issues in Crowdsourcing

• Process
  – Experimental design, annotation guidelines, iteration

• Choose crowdsourcing platform (or roll your own!)

• Human factors
  – Payment / incentives, interface and interaction design, communication, reputation, recruitment, retention

• Quality Control / Data Quality
  – Trust, reliability, spam detection, consensus labeling
Task Design

• Ask the right questions
• Workers may not be experts so don’t assume the same understanding in terms of terminology
• Instructions matter!
• Show examples
• Hire a technical writer
  – Engineer writes the specification
  – Writer communicates
Task Design - UI

• Generic tips
  – Experiment should be self-contained.
  – Keep it short and simple. Brief and concise.
  – Be very clear with the task.
  – Engage with the worker. Avoid boring stuff.
  – Always ask for feedback (open-ended question) in an input box.
Bad Example

- Asking too much, task not clear, “do NOT/reject”
- Worker has to do a lot of stuff

Help us describe How-To Videos! Earn $2.50 bonus for every 25 videos entered!

Watch a how-to video, and write a keyword-friendly synopsis describing the video.

1. Click on the link to watch the Film & Theater how-to video —> 332492 Get a 35mm film look with a depth of field adapter
2. Write a description of the video linked in 4 or more sentences.
3. Be detailed in your description. Describe how the procedure is done.
4. Description should be at least 100 words.
5. Description should be fewer than 2000 characters.
6. Use the character and word counters below to help you stay within the limits.
7. You must complete 25 video descriptions in order to earn the $2.50 bonus. Bonuses are distributed after HITs have been completed. The more HITs completed and approved, the more you will earn.
8. It is not necessary to repeat the headline in your entry. It will NOT count toward your word count.
9. Do NOT describe the following: the format, where the video comes from, or how long the video is. This information is IRRELEVANT.
10. Do NOT describe the video in the following manner. “She turns around to face the camera. Then she faces left.” Follow the examples below.

Current Word Count: 0  Current Character Count: 0 / 2000

Criteria for REJECTION:

1. Entries with obvious and multiple spelling or grammatical errors will be rejected.
2. Entries with fewer than 100 words will be automatically rejected.
3. Text copied from the web or other places will be rejected. Multiple plagiarized answers will lead to being BLOCKED. You may use a quotation, but the majority of your content must be ORIGINAL.
4. Incomplete and blank answers will be rejected. Multiple blank answers will result in being blocked.
5. Tasks submitted without descriptions will be rejected.
6. Tasks submitted with inaccurate descriptions will be rejected as well.
7. Do NOT add any personal opinions. Entries with personal opinions or reviews will be automatically REJECTED.
8. If you notify us that a link is broken, we appreciate it but will not be able to accept the submission. The notification will result in rejection.
9. Entries that transcribe the video will be REJECTED.
Good Example

- All information is available
  - What to do
  - Search result
  - Question to answer
Form and Metadata

• Form with a close question (binary relevance) and open-ended question (user feedback)
• Clear title, useful keywords
• Workers need to find your task
How Much to Pay?

• Price commensurate with task effort
  – Ex: $0.02 for yes/no answer + $0.02 bonus for optional feedback

• Ethics & market-factors
  – e.g. non-profit SamaSource contracts workers refugee camps

• Uptake & time-to-completion vs. Cost & Quality
  – Too little $$, no interest or slow
  – too much $$, attract spammers

• Accuracy & quantity
  – More pay = more work, not better (W. Mason and D. Watts, 2009)
Quality Control

• Extremely important part of the experiment
• Approach as “overall” quality; not just for workers
• Bi-directional channel
  – You may think the worker is doing a bad job.
  – The same worker may think you are a lousy requester.
Quality Control

• Approval rate: easy to use, & just as easily defeated
• Mechanical Turk Masters
  – Recent addition, only for specific tasks
• Qualification test
  – Pre-screen workers’ ability to do the task (accurately)
• Assess worker quality as you go
  – Trap questions with known answers (“honey pots”)
  – Measure inner-annotator agreement between workers
Qualification tests: pros and cons

• Advantages
  – Great tool for controlling quality
  – Adjust passing grade

• Disadvantages
  – Extra cost to design and implement the test
  – May turn off workers, hurt completion time
  – Refresh the test on a regular basis
  – Hard to verify subjective tasks like judging relevance

• Try creating task-related questions to get worker familiar with task before starting task in earnest
Other quality heuristics

• Justification/feedback as quasi-captcha
  – Should be optional
  – Automatically verifying feedback was written by a person may be difficult (classic spam detection task)

• Broken URL/incorrect object
  – Leave an outlier in the data set
  – Workers will tell you
  – If somebody answers “excellent” for a broken URL => probably spammer
Dealing with bad workers

• Pay for “bad” work instead of rejecting it?
  – Pro: preserve reputation, admit if poor design at fault
  – Con: promote fraud, undermine approval rating system

• Use bonus as incentive
  – Pay the minimum $0.01 and $0.01 for bonus
  – Better than rejecting a $0.02 task

• If spammer “caught”, block from future tasks
  – May be easier to always pay, then block as needed
Build Your Reputation as a Requestor

• Word of mouth effect
  – Workers trust the requester (pay on time, clear explanation if there is a rejection)
  – Experiments tend to go faster
  – Announce forthcoming tasks (e.g. tweet)
Crowd Worker Communities

Turkopticon.com
Mturkforum.com
Turkernation.com
## 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*’ |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Fast Deceivers (FD)     | eg: Copy-pasting same text in response to multiple questions, entering gibberish, etc.  
Response: ‘*What’s your task?*, ‘*adasd*, ‘*fgfgf gsd ljlkj*’ |
| Rule Breakers (RB)      | **Instruction:** Identify 5 keywords that represent this task (separated by commas).  
Response: ‘*survey, tasks, history*, ‘*previous task yellow*’ |
| Smart Deceivers (SD)    | **Instruction:** Identify 5 keywords that represent this task (separated by commas).  
Response: ‘*one, two, three, four, five*’ |
| Gold Standard Preys (GSP)| These workers abide by the instructions and provide valid responses, but stumble at the gold-standard questions! |

OpenTurk.com

• Yet another a platform? Build on top of Mturk!
• Chrome Extension for push / notification
• 400+ users
• Open source: https://github.com/openturk/extension
Majority Vote

- Ask N workers and pick the most popular answer
- Works for multiple-choice questions
  - Relevance judgments
  - Sentiment analysis / supervised machine learning
- For other task use *iterations*
  - Audio transcription
  - Ask one worker to transcribe, the next to correct, etc.
- Learning weights for workers
Entity Factor Graphs

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

- **Probabilistic Inference**
  - Select all links with posterior prob > \( \tau \)

\[
\begin{align*}
\text{Graph components:} & \\
\text{Workers, links, clicks} & \\
\text{Prior probabilities} & \\
\text{Link Factors} & \\
\text{Constraints} & \\
\text{Probabilistic Inference:} & \\
\text{Select all links with posterior prob > } \tau
\end{align*}
\]
Aggregation based on worker similarity

- “Community-Based Bayesian Aggregation Models for Crowdsourcing”, Venanzi et al., WWW2014.
- Community-based Bayesian aggregation model
- Group workers by the type of errors they do
SQUARE

• A benchmark for crowd answer aggregation
  – Binary choices (e.g., sentiment)
  – Multiple-choices (e.g., relevance, word-sense disambiguation)

• Compares a number of aggregation techniques over a number of tasks

http://ir.ischool.utexas.edu/square/
Other benchmarks

- Simulations
  - BATC - A Benchmark for Aggregation Techniques in Crowdsourcing
  - Understand effect on efficiency and effectiveness
  - Set aggregation parameters
Tutorial Outline

• Part 1
  – Introduction to Crowdsourcing (30min)
  – Ensuring Quality in Paid Crowdsourcing (60min)

• Part 2
  – Hybrid Human-Machine Data Integration (30min)
  – Crowd-Powered Search (30min)
  – Enterprise Crowdsourcing for Search (30min)
Hybrid Human-Machine Data Integration
Example: Hybrid Data Integration

<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>

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

- **Ask users to verify**

  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
Example: Hybrid Query Processing

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

Facebook Buys Instagram for $1 Billion

by Evelyn Rusli

2:02 p.m. | Updated

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>

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.

RDFa enrichment

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.
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 $\gt \tau$

2 workers, 6 clicks, 3 candidate links
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 WWW13 paper)
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

http://exascale.info/zencrowd/

• Follow up-work (VLDBJ):
  – Also used for instance matching across datasets
  – 3-way blocking with the crowd
ZenCrowd Architecture

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
tamr.com

- Data Integration solutions: algorithms+experts

Tutorial Outline

• Part 1
  – Introduction to Crowdsourcing (30min)
  – Ensuring Quality in Paid Crowdsourcing (60min)

• Part 2
  – Hybrid Human-Machine Data Integration (30min)
  – Crowd-Powered Search (30min)
  – Enterprise Crowdsourcing for Search (30min)
Crowd-Powered Search
Slow Search

• “Not All Searches Need to Be Fast”
  – Planning a vacation
  – Medical diagnosis

• Use additional time for human computation


Crowd-powered Search

• Search process
  – Understand query
  – Retrieve
  – Understand results

• Machines are good at operating at scale

• People are good at understanding
Extract Direct Answers w/ Crowdsourcing

Bernstein et al., Direct Answers for Search Queries in the Long Tail, CHI 2012.
birthdate of the mayor of the capital city of Italy

Asmara - Wikipedia, the free encyclopedia
Jump to Italian Eritrea - ... and when it was occupied by Italy in 1889 and was made the capital city of Eritrea in preference to Massawa by Governor Martini ...

Turin - Wikipedia, the free encyclopedia
Jump to City centre - Via Roma crosses one of the main squares of the city: the pedestrianised ... senate and, for few years, the Italian senate after the Italian unification), the ... to Saint John the Baptist, which is the major church of the city.

Milan - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Milan ▸ Wikipedia ▸
Its business district hosts the Borsa Italiana (Italy's main stock exchange) and the headquarters of the largest national banks and companies. The city is a major ...

Rome - Wikipedia, the free encyclopedia
capital city of Italy

Rome
Italy, Capital

Rome - Wikipedia, the free encyclopedia
Ignazio Marino

The outgoing Mayor of Rome, Gianni Alemanno (PdL), stood for election for a second term as mayor. The center-left candidate, heart surgeon Ignazio Marino was be chosen by a multi-party primary election on 7 April 2013. Control of the 15 municipi of the Italian capital was decided in the election.

March 10, 1955 (age 60 years)

Ignazio Marino, Date of birth

Gianni Alemanno
March 3, 1958

Nicola Zingaretti
October 11, 1965

Salvatore Buzzi
November 15, 1955

Ignazio Roberto Maria Marini is a transplant surgeon and the mayor of Rome. He is a member of the Democratic Party and has been a member of the Italian Senate since 2006. He was born on March 10, 1955. He attended the Catholic University of Sacred Heart (1979) and is a member of the Democratic Party.
Motivation

• Web Search Engines can answer simple factual queries directly on the result page

• Users with complex information needs are often unsatisfied

• Purely automatic techniques are not enough
• We want to solve it with Crowdsourcing!
CrowdQ

- CrowdQ is the first system that uses crowdsourcing to
  - *Understand* the intended meaning
  - *Build* a structured query template
  - *Answer* the query over Linked Open Data
Cities in Italy | Italy Travel Guide
www.italylogue.com/italian-cities
Learn about the best cities in Italy to visit, and some Italian cities you might never have heard of before. These cities in Italy are all great for visitors.

Top Ten Cities for Visitors to Italy - Top Italian Cities to See
goitaly.about.com/od/planningandinformation/tp/topcities.htm
Italy has many beautiful and historic cities that are well worth a visit. Here are our picks for the ten best cities for visitors to Italy.

Italian Cities and Towns - Italy
en.comuni-italiani.it/
Information and statistics on Italian Regions, Provinces, and Municipalities, All Cities...
CrowdQ Architecture

**Off-line**: query template generation with the help of the crowd

**On-line**: query template matching using NLP and search over open data
Hybrid Human-Machine Pipeline

Q= birthdate of actors of forrest gump

Query annotation

Noun

Noun

Named entity

Verification

Is forrest gump this entity in the query?

Entity Relations

Which is the relation between: actors and forrest gump

Starring

<dbpedia-owl:starring>

Schema element

Verification

Is the relation between:
Indiana Jones – Harrison Ford
Back to the Future – Michael J. Fox
of the same type as
Forrest Gump - actors
Structured query generation

Q = birthdate of actors of Forrest Gump

SELECT ?y ?x
WHERE {
  ?z <rdfs:label> ‘Forrest Gump’
}

Results from BTC09:

<http://dbpedia.org/resource/Robin_Wright_Penn> 1966-04-08
<http://dbpedia.org/resource/Tom_Hanks> 1956-07-09
<http://dbpedia.org/resource/Sally_Field> 1946-11-06
<http://dbpedia.org/resource/Gary_Sinise> 1955-03-17
<http://dbpedia.org/resource/Mykelti_Williamson> 1960-03-04
## 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>2011</td>
<td>[52]</td>
<td>Database</td>
<td>Graph</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>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>[48]</td>
<td>Biomedical</td>
<td>Ontology</td>
<td>Post-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>
</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
Summary

• Crowdsourcing big data can make you go bankrupt! -> hybrid systems
• When to ask a human, when to trust the machine
• Hybrid systems (human in the loop)
  – Pre-processing: training data for ML
  – Post-processing: based on confidence scores
  – Mix: active learning

Tutorial Outline

• Part 1
  – *Introduction to Crowdsourcing* (30min)
  – *Ensuring Quality in Paid Crowdsourcing* (60min)

• Part 2
  – *Hybrid Human-Machine Data Integration* (30min)
  – *Crowd-Powered Search* (30min)
  – *Enterprise Crowdsourcing for Search* (30min)
Enterprise Crowdsourcing for Search
Enterprise Crowdsourcing

• Internal crowd
  – Employees of the company
  – Full-time annotators
  – Casual crowd workers

• Pro: Trust, Domain Knowledge

• Contra: Incentives, Load-balancing
Crowds for Enterprise Crowdsourcing

• Internal Crowd
  – IBM
  – Microsoft
  – Google

• External Crowd
  – Amazon MTurk
  – Yandex Toloka toloka.yandex.com
Crowds for Enterprise Crowdsourcing

• Mixed
  – NDA Crowds by Crowdflower
  – Top Tolokers become Yandex employees

• Tamr.com
  – Internal Expert-sourcing for data integration
Use of Crowdsourcing for data cleaning / extraction

• Locu / GoDaddy
  – http://www.oreilly.com/pub/e/3298
  – “learnings from 17 conversations with companies that make heavy use of crowd work”
Conclusions

• Crowdsourcing: a way to get manual data annotation / cleaning / processing at scale
• Applications to search
  – Evaluation / relevance judgments
  – Complex query understanding
  – Information Finding (e.g., customer care phone no)
  – Result extraction and aggregation in tabular format
Conclusions

• Challenges
  – **Quality** if public crowds are used
  – Many techniques can be used to guarantee high quality, commercial services are coming up
  – **Deadlines**: it is difficult to predict crowd execution time
  – Task reward can be used as a means to speed-up execution
  – **Cost**: can be reduced thanks to hybrid human-machine systems