

Introduction to Crowdsourcing

Lecture 1

Gianluca Demartini

University of Sheffield

Gianluca Demartini



- B.Sc., M.Sc. at U. of Udine, Italy
 - Ph.D. 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, Aug 14
 - Tutorial on Entity Search at ECIR 2012 and RuSSIR 2015, on Crowdsourcing at ESWC 2013 and ISWC 2013
 - Research Interests
 - Information Retrieval, Semantic Web, Human Computation
- www.gianlucademartini.net

g.demartini@sheffield.ac.uk

Who are you?

Tentative Menu

- Monday
- Lecture 1 - Introduction to Crowdsourcing
 - An overview of the entire course.
 - Early examples of crowdsourcing (reCAPTCHA, ESP game).
 - Types of incentives: games with a purpose, citizen science, and community based crowdsourcing.
- Lecture 2 - Introduction to Micro-task Crowdsourcing Platforms
 - Key terminology of micro-task crowdsourcing.
 - Popular platforms such as Amazon MTurk and CrowdFlower.
 - How to use such systems as a crowd worker

Tentative Menu

- Tuesday
- Lecture 3 – How to Setup a Crowdsourcing Micro-task
 - Dimensions involved in crowdsourcing task design such as pricing, question design, and quality assurance mechanisms (e.g., honeypots).
 - Design and deploy a task during the lecture and see how to collect results back from the crowdsourcing platform.
- Lecture 4 – Micro-task Crowdsourcing Effectiveness
 - Techniques to ensure high quality in crowdsourced tasks (e.g., answer aggregation techniques, push crowdsourcing).
 - Behavior of malicious workers in crowdsourcing platforms.

Tentative Menu

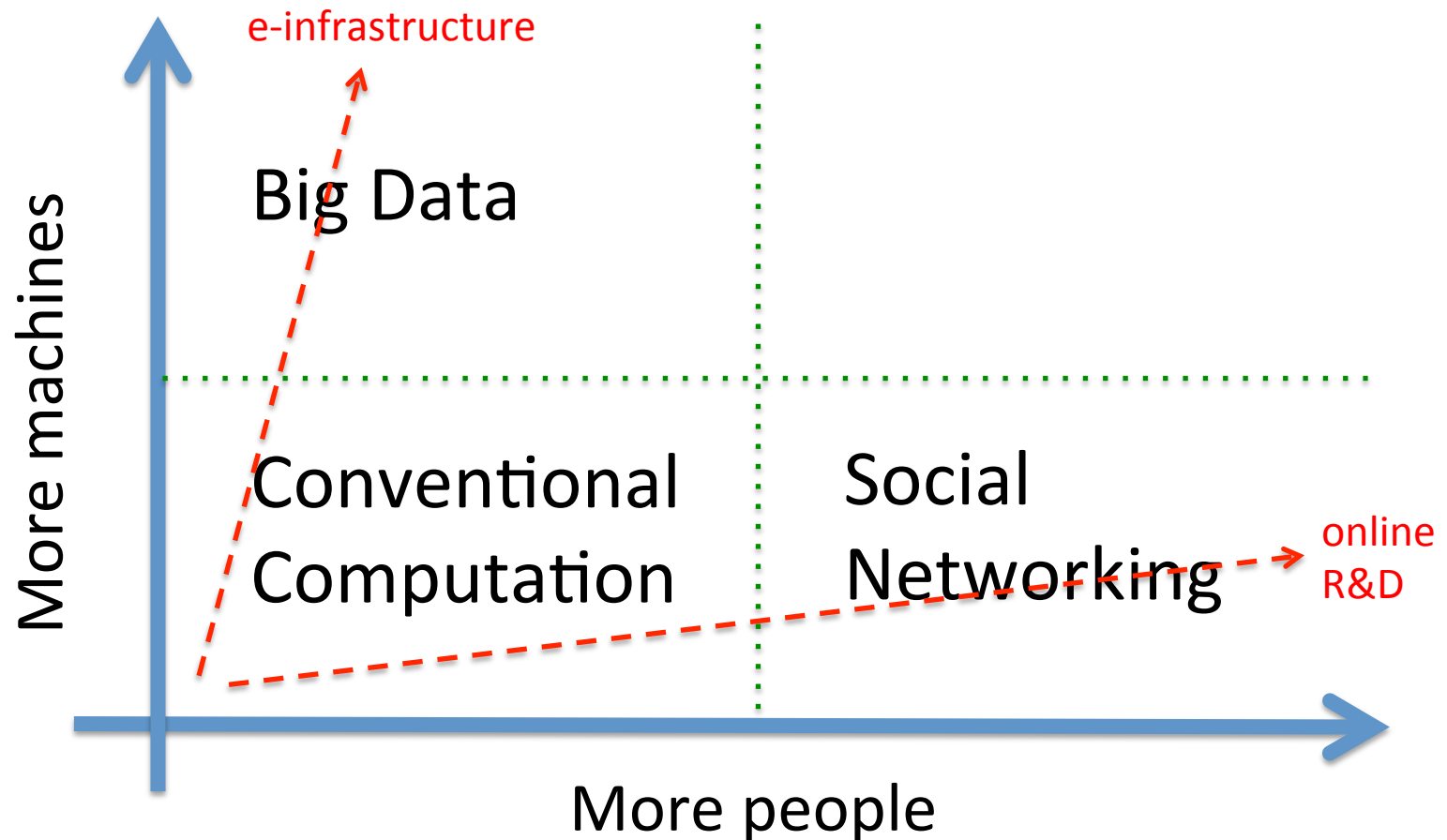
- Wednesday
- Lecture 5 - Hybrid Human-machine Systems
 - Advanced example uses of crowdsourcing.
 - Systems that combine both the scalability of machines over large amounts of data as well as the quality of human intelligence
- Lecture 6 - Micro-task Crowdsourcing Scalability
 - In hybrid human-machine systems the latency bottleneck lays on the side of the crowd.
 - Recent research results that proposed techniques to improve the latency of crowdsourcing platforms.
 - Pricing techniques, HIT scheduling

Tentative Menu

- Thursday
- Lecture 7 - Open Research Directions in Micro-task Crowdsourcing
 - In this lecture we will give an overview on open micro-task crowdsourcing research questions.
 - Summarize which communities, conferences, journal, researchers work on crowdsourcing
- Slides:
 - <http://www.gianlucademartini.net/crowdsourcing/>

Crowdsourcing

How to build social systems at scale?



Overview



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

Crowdsourcing

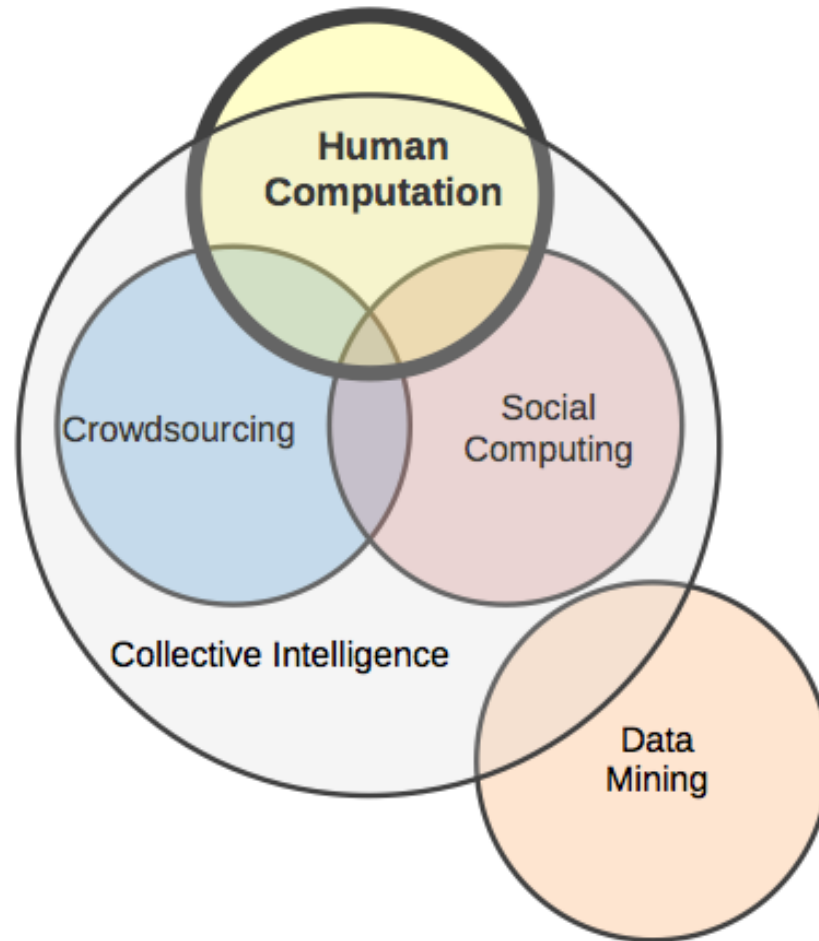
- *Portmanteau* of "crowd" and "outsourcing," first coined by Jeff Howe in a June 2006 Wired magazine article
- [Merriam-Webster] the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than 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]

One View of Crowdsourcing



From Quinn & Bederson, “Human Computation: A Survey and Taxonomy of a Growing Field”, CHI 2011.

Dimensions of human computation

See also [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)

See also [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

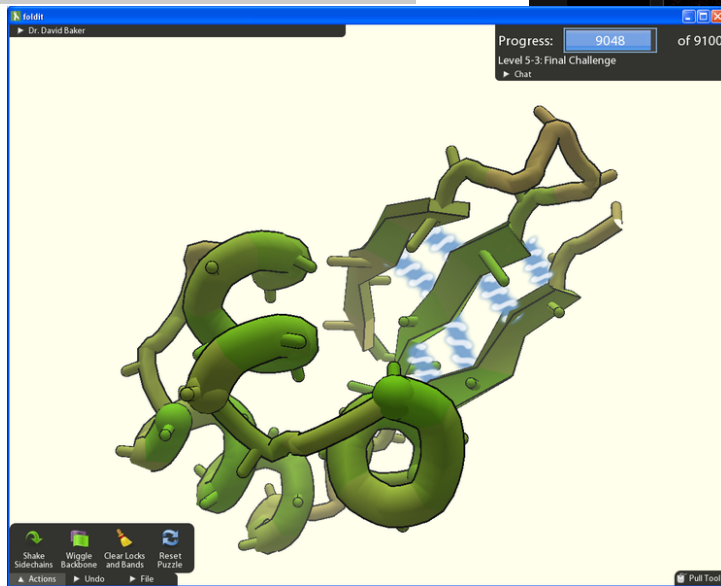
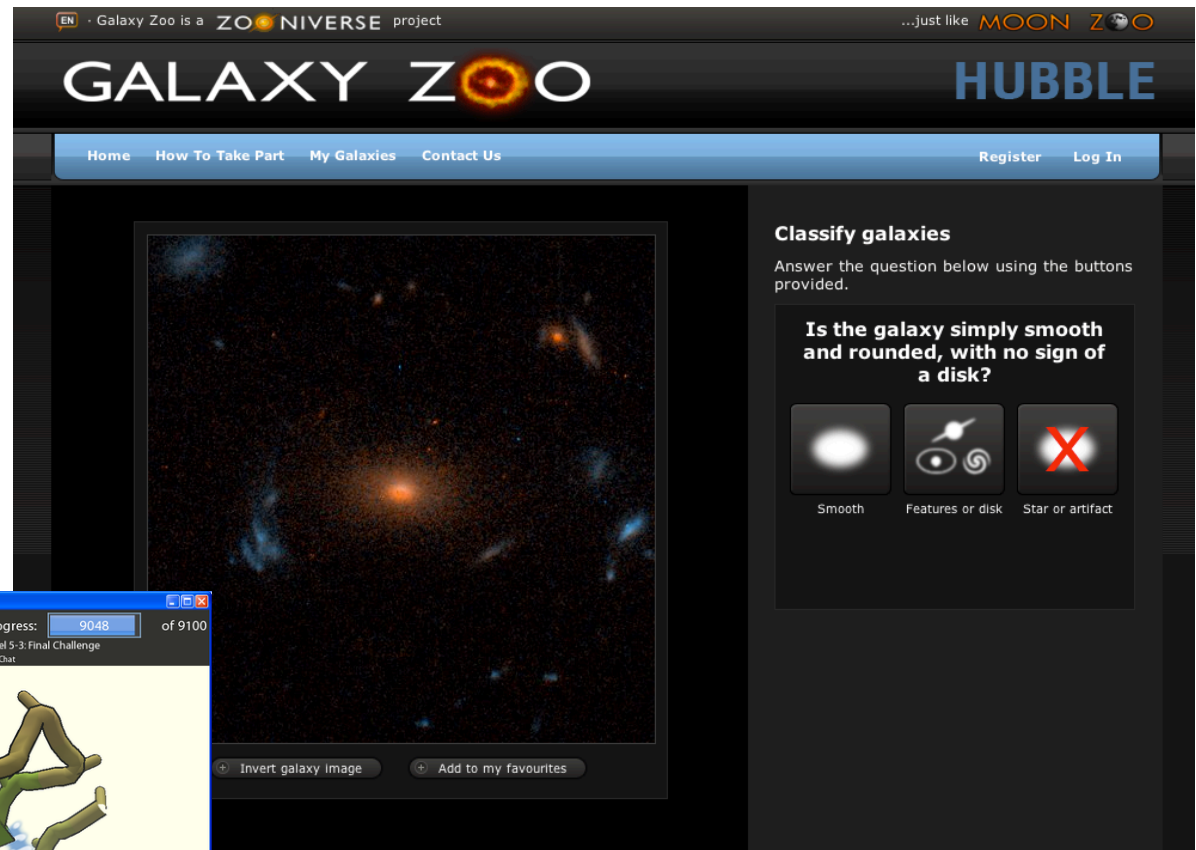
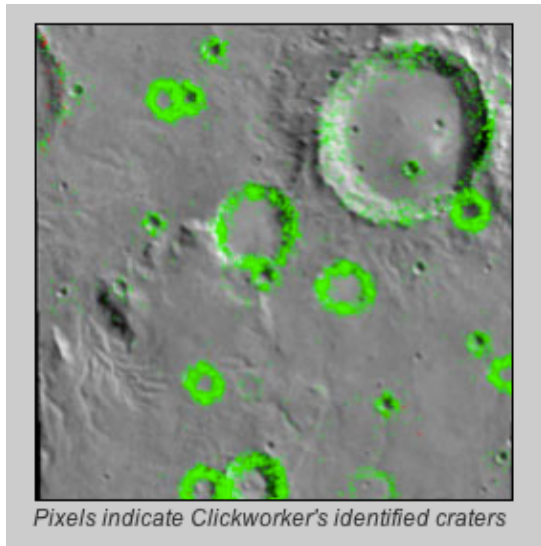
- Incentives and motivators
- Assigning tasks to people based on their skills and performance (as opposed to random assignments)
- Symbiotic combinations of human- and machine-driven computation, including combinations of different forms of crowdsourcing

Aligning incentives is essential

altruism
reputation
freedom reciprocity
self-expression
competition
community
autonomy
fun

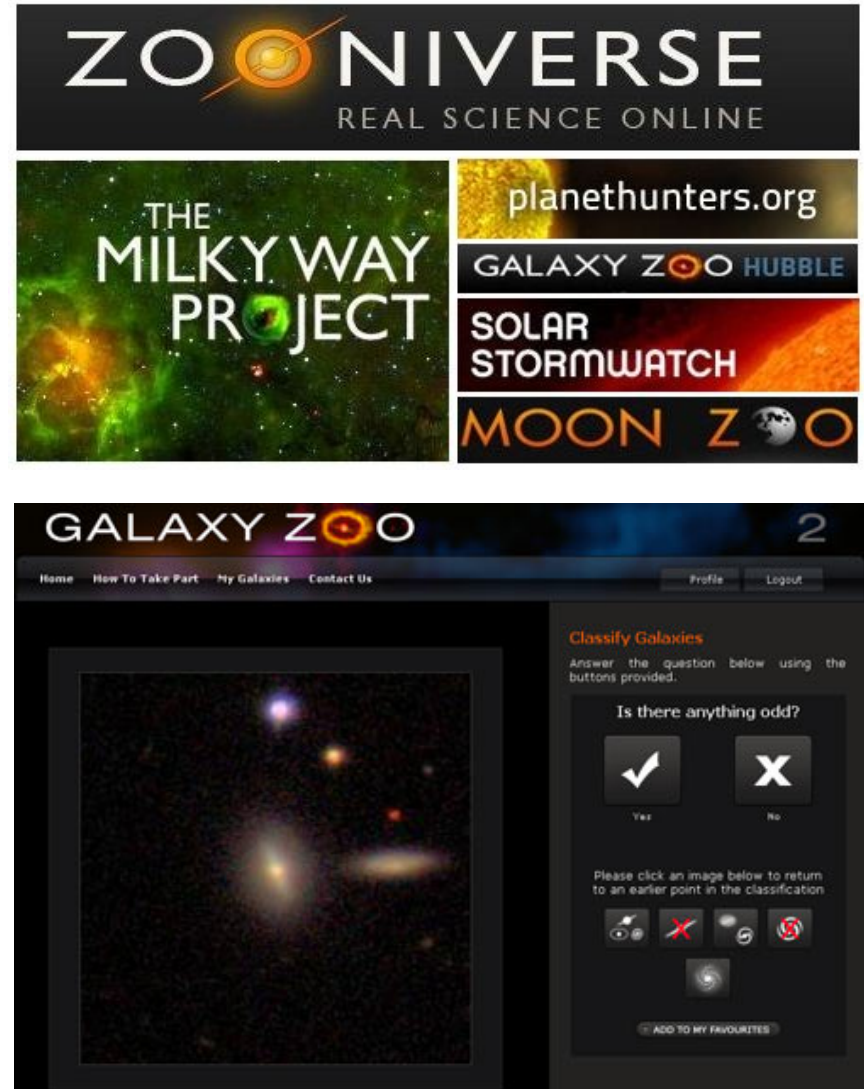
- Motivation: driving force that makes humans achieve their goals
- Incentives: ‘rewards’ assigned by an external ‘judge’ to a performer for undertaking a specific task
 - Common belief (among economists): incentives can be translated into a sum of money for all practical purposes.
- Incentives can be related to both extrinsic and intrinsic motivations.
- **Extrinsic motivation** if task is considered boring, dangerous, useless, socially undesirable, dislikable by the performer.
- **Intrinsic motivation** is driven by an interest or enjoyment in the task itself.

Citizen Science



Citizen Science

- WHAT IS OUTSOURCED
- Object recognition, labeling, categorization in media content
- WHO IS THE CROWD
- Anyone
- HOW IS THE TASK OUTSOURCED
- Highly parallelizable tasks
- Every item is handled by multiple annotators
- Every annotator provides an answer
- Consolidated answers solve scientific problems



Citizen Journalism and Participatory Sensing



innocentive.com

- Tech Innovation
- via Crowdsourcing
- Competitions
- Monetary Incentive

<https://www.innocentive.com/pavilion/NASA>




C

Challenge Title
Improved Barrier Layers ... Keeping Food Fresh in Space
Mechanism for a Compact Aerobic Resistive Exercise Device
Data-Driven Forecasting of Solar Events
Coordination of Sensor Swarms for Extraterrestrial Research
Medical Consumables Tracking
Augmenting the Exercise Experience
Simple Microgravity Laundry System

Question Answering Systems

Welcome to Q&A for professional and enthusiast programmers — check out the [FAQ!](#)

StackExchange [log in](#) [careers](#) [dev days](#) [chat](#) [meta](#) [about](#) [faq](#)

 **stackoverflow** [Questions](#) [Tags](#) [Users](#) [Badges](#) [Unanswered](#) [Ask Question](#)

Top Questions [interesting](#) [237 featured](#) [hot](#) [week](#) [month](#)

0 votes 0 answers 1 view

[n Random rows for a given attribute - Postgres](#)

[sql](#) [postgresql](#)

44s ago Sup3rkiddo 49

1 vote 1 answer 14 views

[Branch descriptions in git, continued](#)

[git](#) [branch](#) [task-tracking](#)

48s ago manojds 16.2k

0 votes 0 answers 1 view

[Where is hostname defined for the anchor element?](#)

[javascript](#)

56s ago Chris Aaker 868

2 votes 1 answer 12 views

[User-defined Table Variables in MySQL 5.5?](#)

[mysql](#) [stored-procedures](#) [routines](#)

1m ago colonel_px 11

0 votes 2 answers 37 views

[Closing cfpdf tag with </cpdf> causes error](#)

[coldfusion](#) [coldfusion-8](#) [cfeclipse](#) [cpdf](#)

1m ago Jens Wegar 81

0 votes 0 answers 6 views

[cocoa memory leak by CGAffineTranform or by view](#)


[iphone](#) [objective-c](#) [cocoa](#) [memory-leaks](#) [leak](#)

1m ago EmptyStack 9,100

Hello World!

This is a collaboratively edited question and answer site for **professional and enthusiast programmers**. It's 100% free, no registration required.

[about »](#) [faq »](#)

 **CAREERS 2.0**
by stackoverflow

[Senior PHP Engineer](#)
Spreetales
Los Altos, CA; San Francisco, CA


[Front End Software Engineer](#)
@Rdio
Rdio
San Francisco, CA

[Senior Mobile Developer](#)
American Public Media
Oakland, CA

[Web Engineer](#)
Monkey Inferno
San Francisco, CA

DB specific

- Freebase













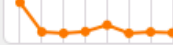





[Data](#) [Schema](#) [Apps](#) [Docs](#)


An entity graph of people, places and things, built by a community that loves open data.

Sort by write activity

Jul 4 Aug 1 last week

Facts Topics Top User

<div>Featured Data</div> <div>Arts & Entertainment</div> <div>Products & Services</div> <div>Science & Technology</div> <div>Society</div> <div>Special Interests</div> <div>Sports</div> <div>System</div> <div>Time & Space</div> <div>Transportation</div> <div>All</div>	Film 80 members		43K last week	5M	641K	
	People 87 members		11K last week	7M	2M	
	TV 35 members		8K last week	8M	1M	
	Music 100+ members		771 last week	35M	10M	
	Business 100+ members		431 last week	2M	611K	
	Government 47 members		240 last week	532K	135K	
	Location 52 members		223 last week	10M	999K	
	Books 47 members		108 last week	29M	6M	



Google Refine

An open source power tool to fix, discover, experiment, connect and customize your data. [Learn more »](#)

What is Freebase?

Learn what an entity graph is, what kind of information it contains, and why you should add your data!

[Learn More »](#)

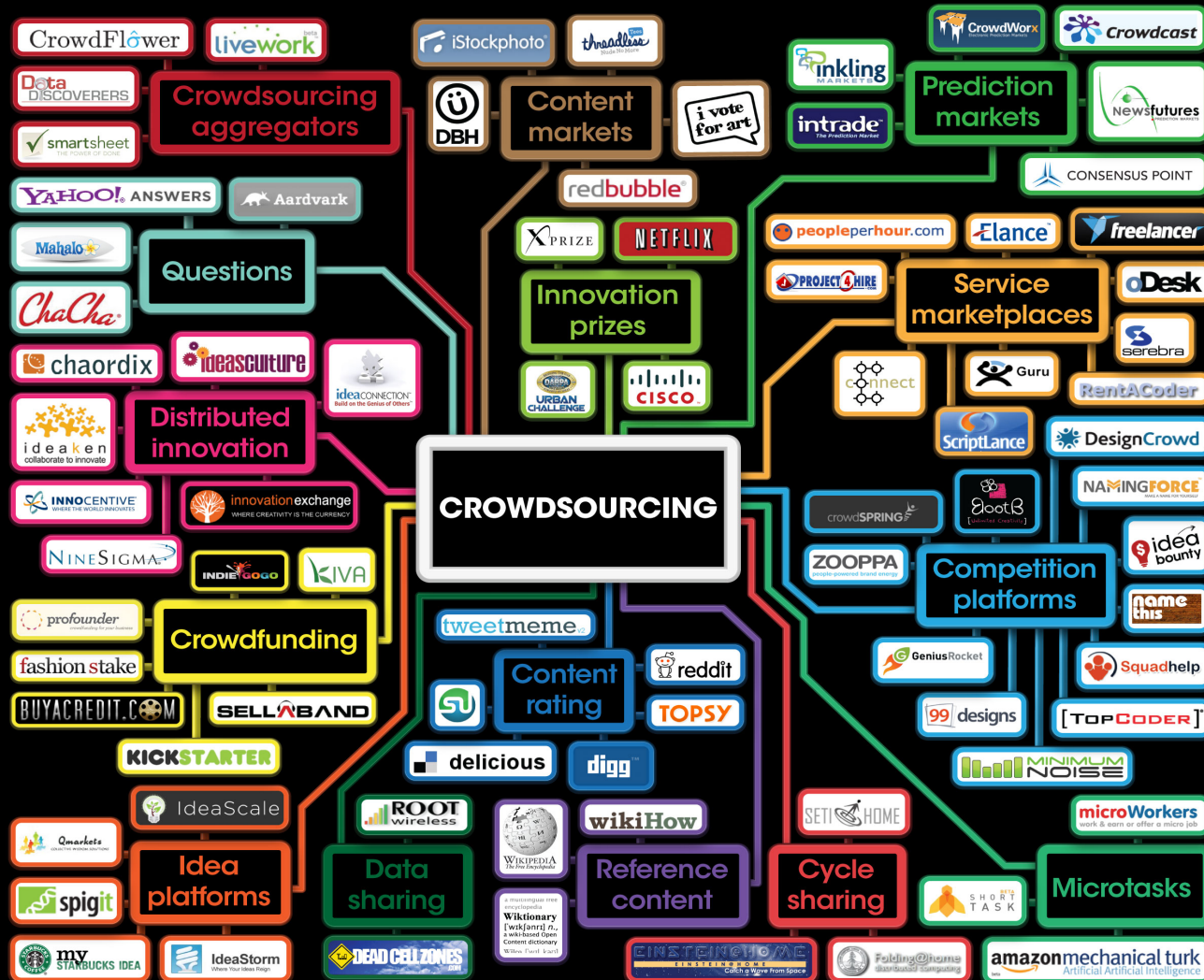
Freebase for Developers

- powerful queryable API
- JavaScript-based hosting framework
- libraries for other languages

[Learn More »](#)

The Way Industry Looks At It

CROWDSOURCING LANDSCAPE Beta v1



Common Crowdsourcing Tasks and Examples

3D object design	Thingiverse
Advertising	idea bounty
Business ideas	anadisse
Clothing	ClickAdvisor
Consumer research	ClickAdvisor
Crisis information	Ushahidi
Data analysis	SETI@HOME
Fact checking	PolitiFact.com
Graphic design	99 designs
Human reading	Capella
Investigative reporting	theguardian
Journalism	seed
Lending	zipo
Mapping	IMDb
Movie reviews	IMDb
Music	musikpitch
Observation	GALAXY ZOO
Patent research	Journalist Data Visualization
Philanthropy	philodoma
Political activism	MoveOn.org
Product design	MUJI
Proofreading	Bluewin
Scientific problems	fold
Software	RenataCoder
Software development	oTest
Software testing	oTest
Stock picking	marketocracy
Tagging	Google
Translation	facebook
Trends	TRENDHUNTER
TV programming	current
Word of mouth	BzzAgent
Writing and editing	crowdink

For details, analysis, and discussion go to:

www.crowdsourcingresults.com



Advanced
Human
Technologies

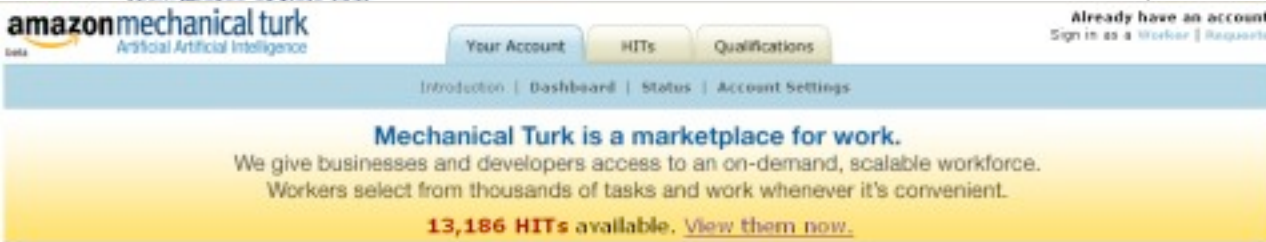
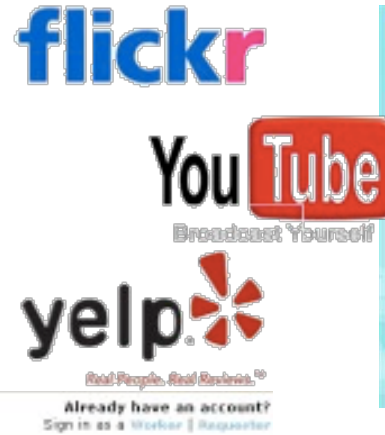
Published under a Creative Commons
Attribution-ShareAlike 2.5 License

Created by Ross Dawson
Design by Daniil Alexandrov

Taxonomies

- Doan, Halevy, Ramakrishnan; (Crowdsourcing) CACM 4/11
 - nature of collaboration (implicit vs. explicit)
 - architecture (standalone vs. piggybacked)
 - must recruit users/workers? (yes or no)
 - What do users/workers do?
- Bederson & Quinn; (Human Computation) CHI '11
 - Motivation (Pay, Altruism, Enjoyment, Reputation)
 - Quality Control (mechanisms)
 - Aggregation (how are results combined?)
 - Human Skill (Visual recognition, language, ...)
 - ...

Participatory Culture - Explicit



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



or [learn more about being a Worker](#)

Get Results from Mechanical Turk Workers

Ask workers to complete HITS - Human Intelligence Tasks - and get results using Mechanical Turk. [Get started.](#)

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



or [learn more about being a Requester](#)



Participatory Culture – Implicit

John Murrell: GM SV 9/17/09

...every time we use a Google app or service, we are working on behalf of the search sovereign, creating more content for it to index and monetize or teaching it something potentially useful about our desires, intentions and behavior.



OCR errors: reCAPTCHA



1629 capplots | 1960-73 batizak | III, Quichut

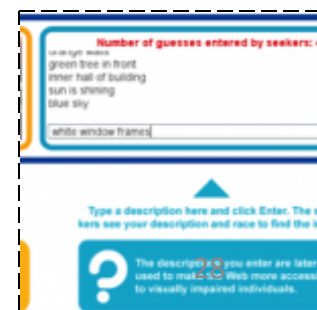
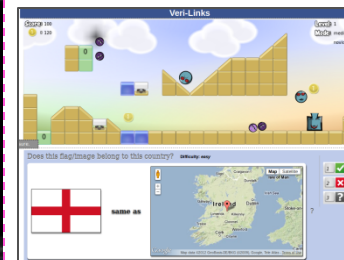
RICHARD, Redfully | ansunjon Rupprecht | Daterny Nature

drôle Deratial | than Morgicar | Golightly, Byemore

Games with a purpose (GWAP)

See also [von Ahn & Dabbish, 2008]

- Human computation disguised as casual games
- Tasks are divided into parallelizable atomic units (challenges) solved (consensually) by players
- Game models
 - Single vs. multi-player
 - Selection agreement vs. input agreement vs. inversion-problem games



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

Games with a Purpose

- Quality assurance
 - Synchronous interaction in games
 - Levels of difficulty and near-real-time feedback in games
 - Many methods applied in both cases (redundancy, votes, statistical techniques)
- Different set of incentives and motivators

Gamification

- A human-based computation technique in which a computational process performs its function by outsourcing certain steps to humans in an entertaining way

How to implement gamification

- **Cosmetic:** adding game-like visual elements or copy (usually visual design or copy-driven)
- **Accessory:** wedging in easy-to-add-on game elements, such as badges or adjacent products (usually marketing-driven)
- **Integrated:** more subtle, deeply integrated elements like % complete (usually interaction-design driven)
- **Basis:** making the entire offering a game (usually product-driven)

Transactive Search

Transactive Search

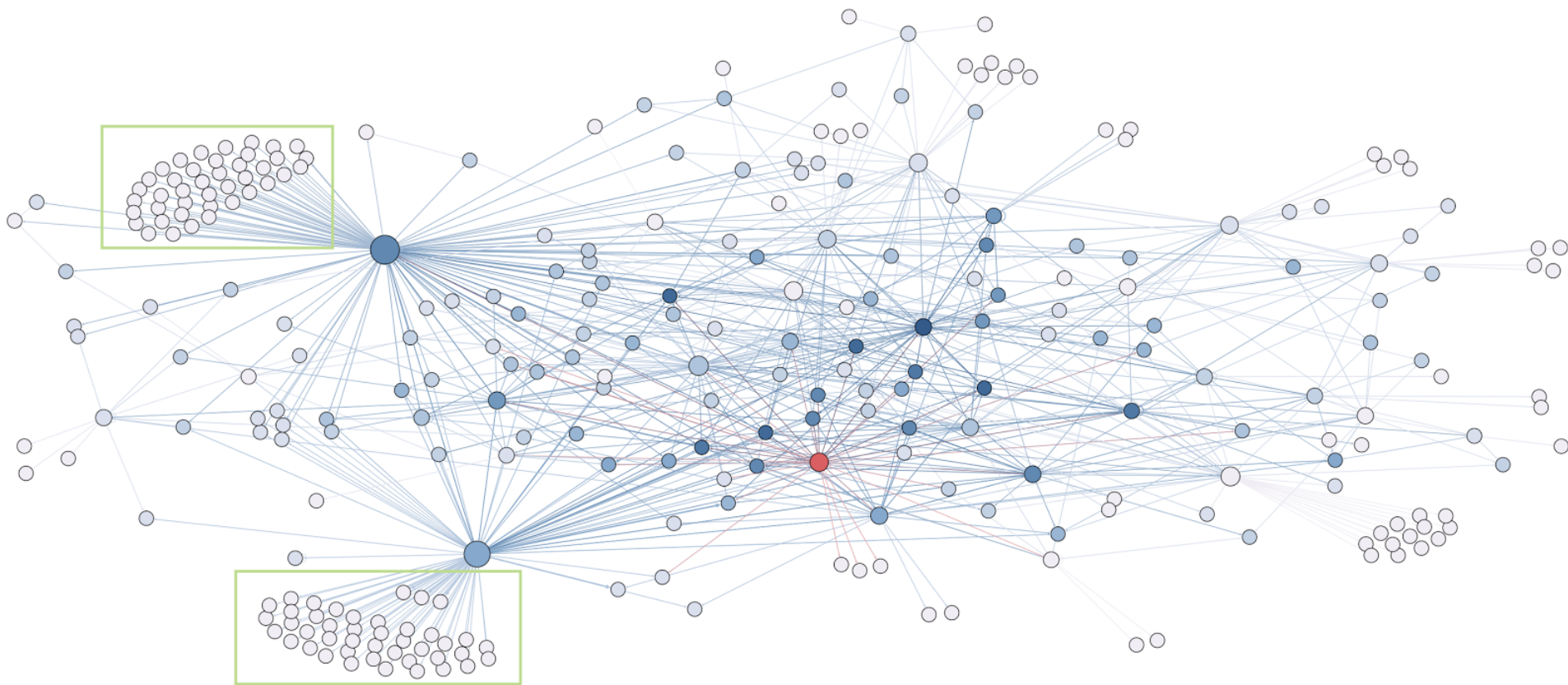
- What if the data to answer your query is not stored on any digital support?
- What if the data is just in people minds?
- Big Data ➔ No Data

Transactive Search

- Search using Transactive (group) Memories
- “Who attended the WWW 2014 conference?”
- **Machines:** Harvest the Web + Data Mining
- **Crowd:** Search twitter, look at event pictures
- **Transactive Memories:** Remember who I met

Michele Catasta, Alberto Tonon, Djellel Eddine Difallah, Gianluca Demartini, Karl Aberer, and Philippe Cudré-Mauroux. Hippocampus: Answering Memory Queries using Transactive Search. In: 23rd International Conference on World Wide Web (**WWW 2014**), Web Science Track. Seoul, South Korea, April 2014.

Transactive Search (2)



Transactive Search (3)

Approach	Precision	Recall	F-measure
Authors and Tweets	0.3048	0.6906	0.4229
SVM	0.6632	0.4532	0.5385
M5P Regression	0.6599	0.4652	0.5457
Hybrid_uncertain	0.5864	0.4964	0.5377
Hybrid_unseen	0.4884	0.6043	0.5402
Hybrid_uncertain_unseen	0.4592	0.6211	0.5280
Transactive Search	0.9006	0.7136	0.7963

Table 3: Effectiveness of machine-based, hybrid, and Transactive Search approaches using Crowdsourcing for ISWC 2013.

Discussion

- Sometime data is not on the Web
- The right group of *people* can still answer
 - Collaboratively
 - Using Transactive Search
 - Better than machines or anonymous crowds
- Open challenges
 - Incentives
 - Repeatability
 - SNA

Summary

- Crowdsourcing has very many meanings
 - Incentives
 - Explicit/Implicit participation
 - Online/Offline
- This week we will focus on (paid) micro-task crowdsourcing to improve over machine-based systems

Crowdsourcing Incentives

- Paid Crowdsourcing
 - **Competition** with others (bonus payment for best performance)
 - Surveillance (check before paying)
 - Solidarity (your **team** will receive a bonus)
 - Accuracy (**bonus** for correct answers)
 - **Agreement** with others (bonus for agreeing with the majority)
- Fun (enjoyment)
- Community (belonging, desire to help)

Paid Crowdsourcing Ethics

- People work full-time as crowd workers
- Chinese crowdsourcing platform with 5.5M workers
- Pros
 - Help developing countries
 - Provide cash fast to people == short-term satisfaction
 - Job Flexibility
- Cons
 - No job security
 - No social security
 - Long term satisfaction? Career plans?