

# Collective Intelligence: Crowdsourcing groundtruth data for large scale evaluation in Information Retrieval

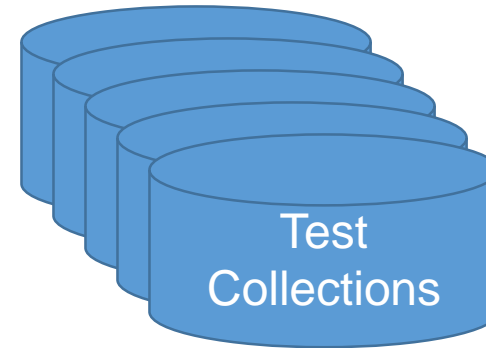
Sergej Zerr

S.Zerr@soton.ac.uk

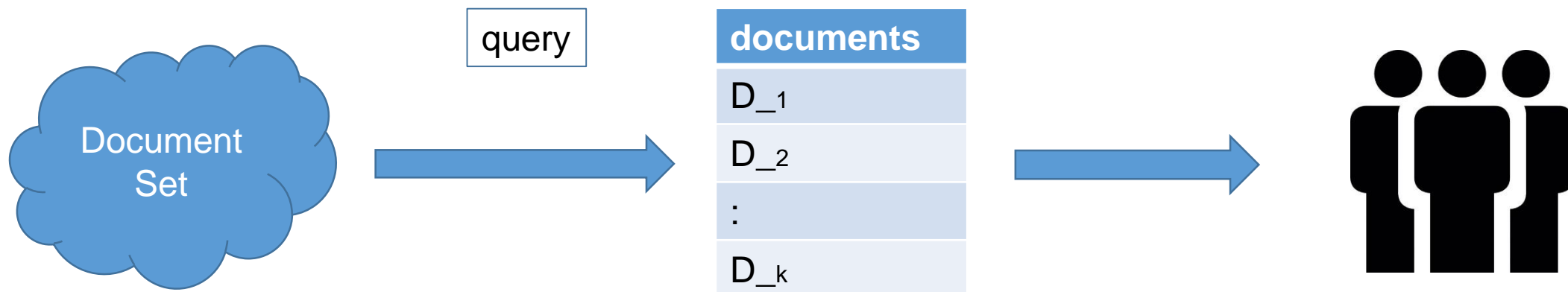
# Challenges in IR Evaluation

## ➤ BigData

- Heterogeneity (larger annotation demand)
- Dynamicity (updates required)
- Novel tasks (no test collections)

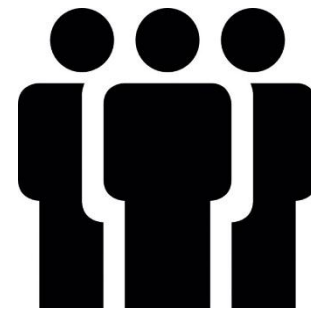


- Relevance ranking
- Search result diversification
- Temporal retrieval
- etc.



# Challenges in IR Evaluation

- Better human accessibility
  - WiFi, Mobile Networks, Portable gadgets (larger crowd)
- Challenges:
  - How to motivate the crowd to work?
  - How to obtain meaningful results from the individuals?
  - How to aggregate the crowdsourced results?
  - How to evaluate the output?

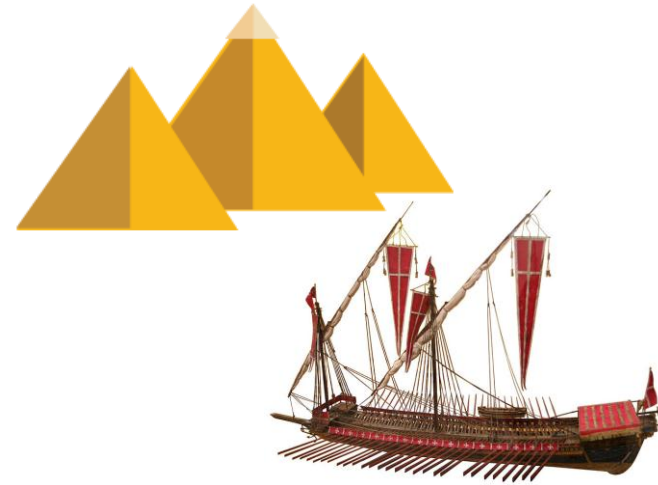


# Outline

- **Collaborative Advantages**
  - The wisdom of crowds
  - Conditions for a successful collaboration
- Obtaining collaborative knowledge
  - Crowd motivation
  - Scalability/Efficiency
  - Own work
- Input/Output Evaluation
  - Users and Data
  - Quality assurance
- Discussion

# Collaboration

Often we need more than one hand



Also more than one brain

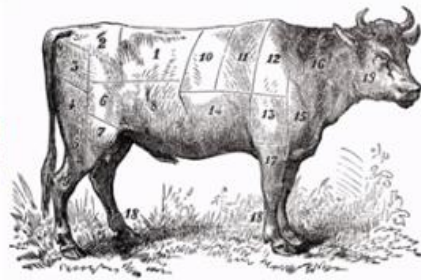
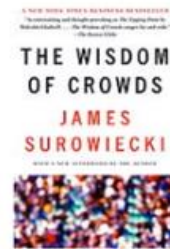
*“Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations”*

James Surowiecki



.. The Wisdom of Crowds ..

## The Wisdom of Crowds



*average of 800 guesses = 1,197*  
*actual weight of the ox = 1,198*

93

In 1906, the statistician Francis Galton observed a competition at a country fair. The crowd accurately guessed the weight of an ox when their individual guesses were averaged (the average was closer to the ox's true butchered weight than the estimates of most experienced crowd members)

# Crowd IQ: aggregating opinions to boost performance

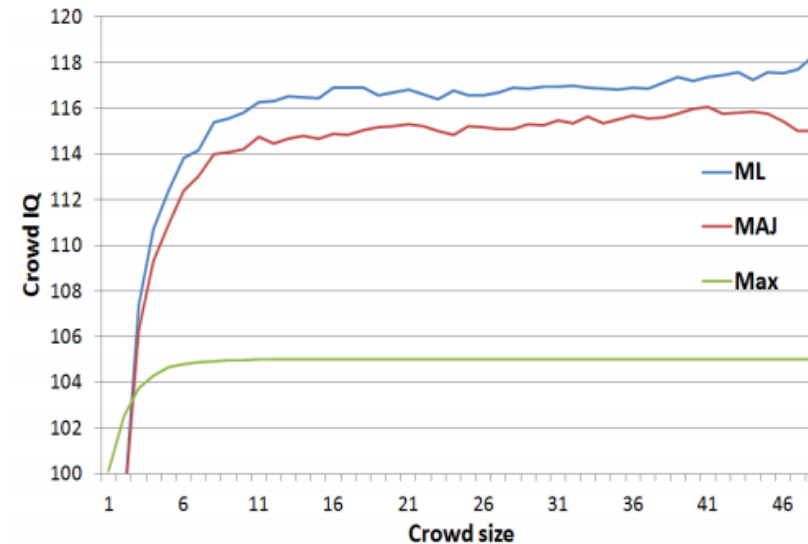
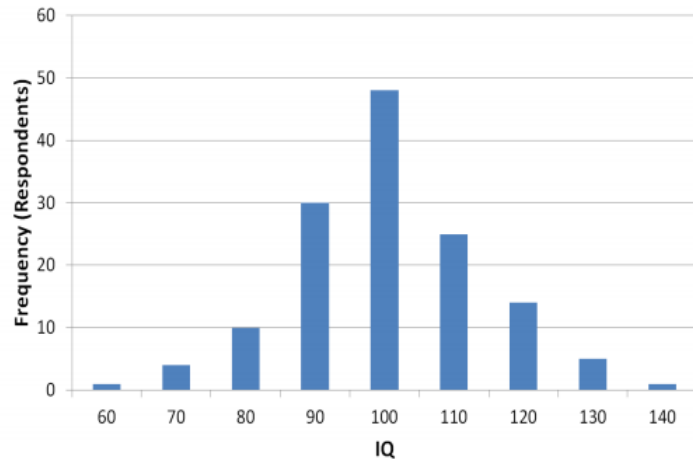
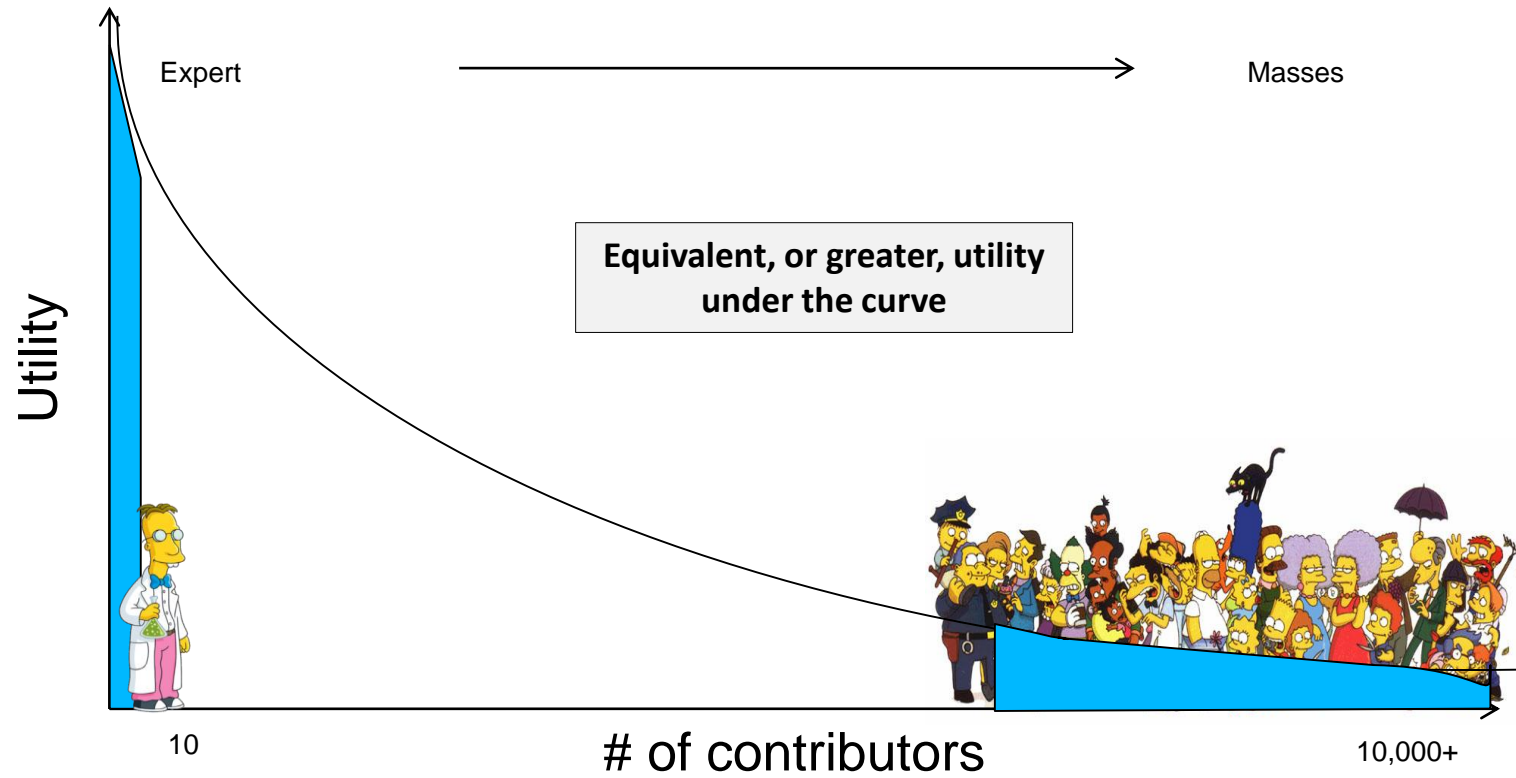


Figure 7: Crowd IQ and maximal IQ for  $P_{[95,105]}$

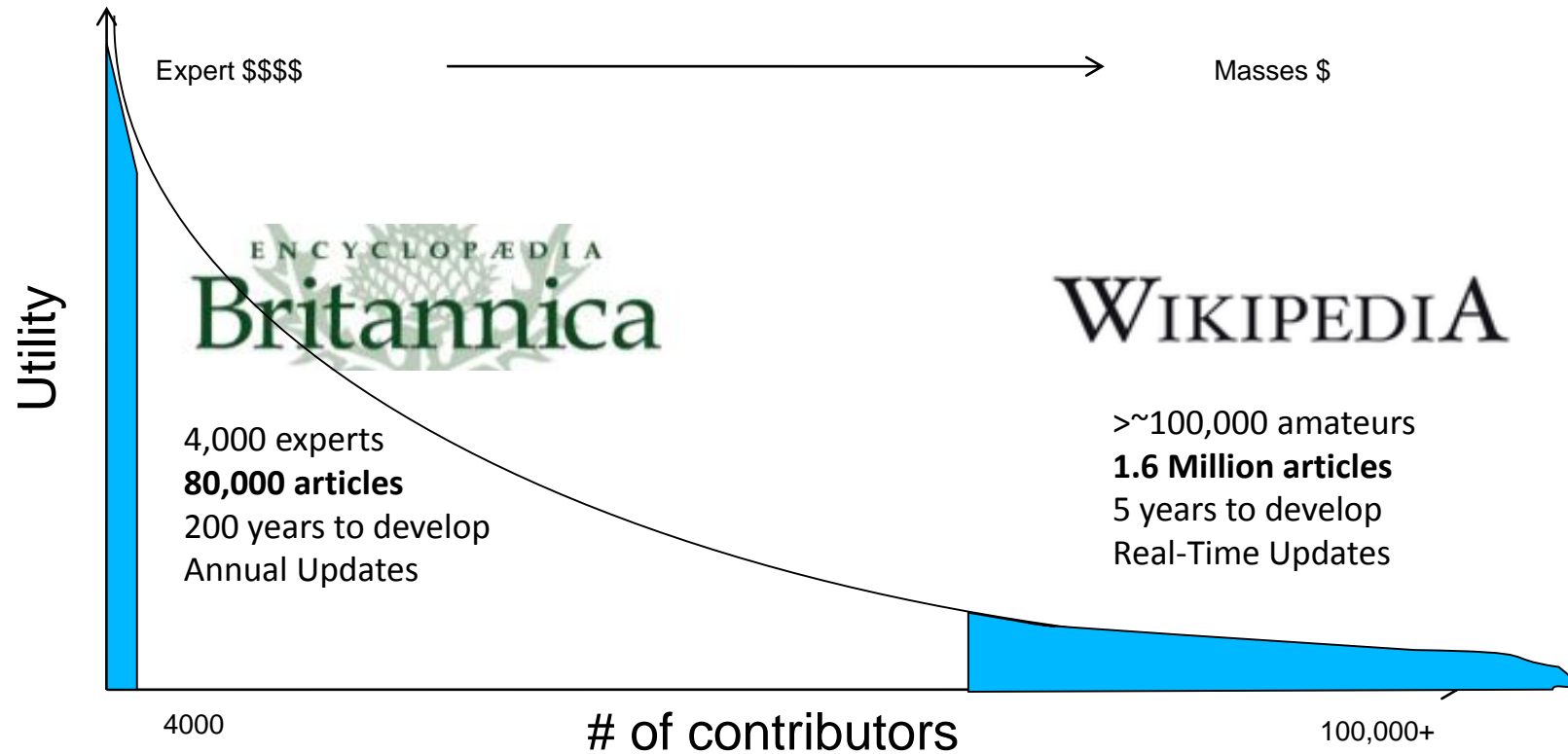
Yoram Bachrach, Thore Graepel, Gjergji Kasneci, Michal Kosinski, Jurgen Van Gael: Crowd IQ: aggregating opinions to boost performance. AAMAS 2012

# United Brains





# United Brains



# (In,-) Direct Collaboration in IR can be used:

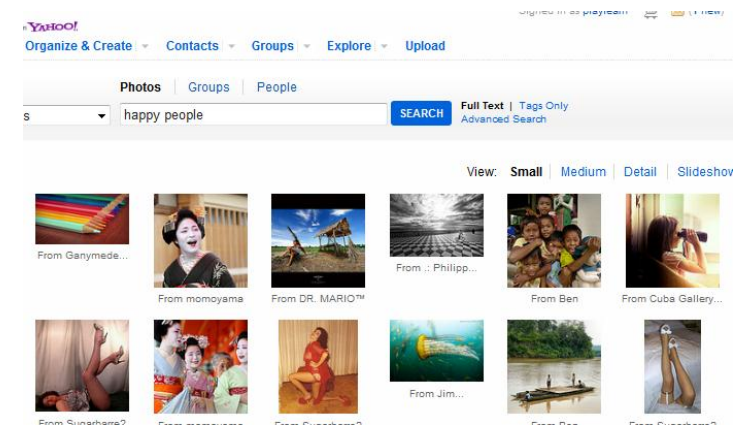
- Collaborative tagging,
- Favorite assignments,
- Click logs,
- Data partitioning,
- Recommendations,
- ect., ect.,ect....



1. Rating: 3.4/5 (14 votes cast)

Tags: Rainbow, Sea, Island, Green, Palm tree, Maui

# Google

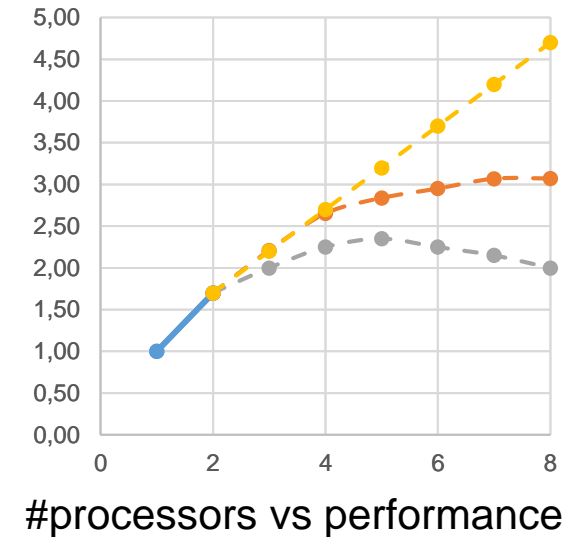


# Collaboration: Paradox

- Using „Wisdom of Crowds” is not always straight-forward to achieve.
- Collaborative work needs to be managed efficiently
- Kasparov won against the world in 1999

[http://en.wikipedia.org/wiki/Kasparov\\_vsus\\_the\\_World](http://en.wikipedia.org/wiki/Kasparov_vsus_the_World)

ORACLE speedups. <https://docs.oracle.com/cd/E19205-01/819-5265/bjael/index.html>



# Collaboration: Success Criteria

Criteria	Description
<b>Diversity of opinion</b>	Each person should have “private” information.
<b>Independence</b>	People's opinions aren't determined by the opinions of those around them.
<b>Decentralization</b>	People are able to specialize and draw on local knowledge.
<b>Aggregation</b>	Effective mechanism exists for turning private judgments into a collective



# Groupthink Symptoms:

Irving Lester Janis (26 May 1918 - 15 November 1990)

- Collective rationalization
- Self-censorship
- Direct pressure on dissenters
- Self-appointed 'mindguards'



<https://www.youtube.com/watch?v=fulXiXqv978>

# Collaboration

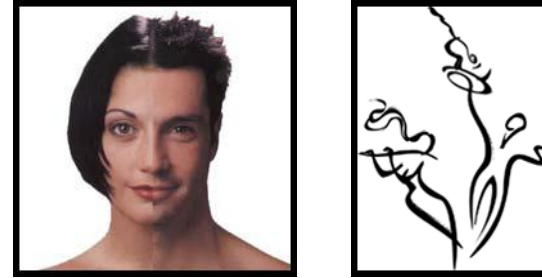
“The best collective decisions are the product of disagreement and contest, not consensus or compromise.”

“The best way for a group to be smart is for each person in it to think and act as independently as possible.”



# Outline

- Collaborative Advantages
  - The Wisdom of Crowds
  - Conditions for a successful collaboration
- **Obtaining collaborative knowledge**
  - Crowd motivation
  - Scalability/Efficiency
  - Own work
- Input/Output Evaluation
  - Users and Data
  - Quality assurance
- Discussion



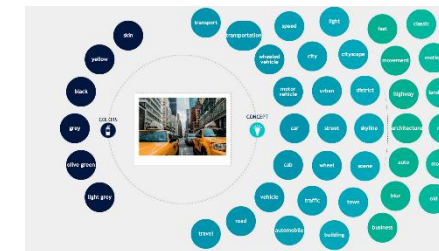
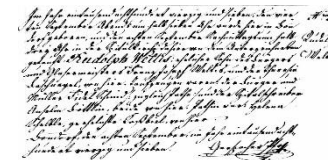
Two different images that share the same labels:  
**man and woman**



# Machine Vs. Human

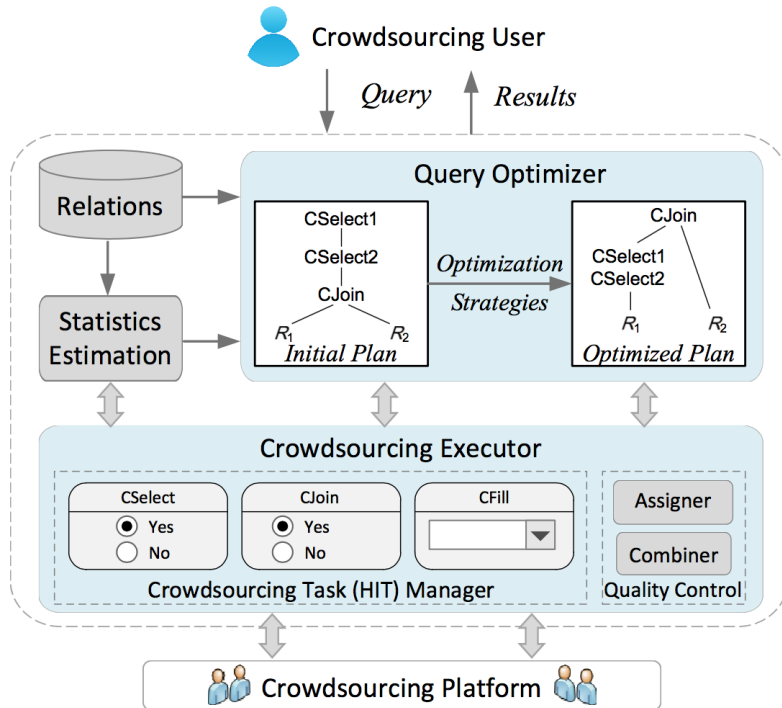
Humans can (yet) solve some tasks more efficient and/or accurate as a machine would do.

- Captcha (OCR)
- Classification
- Image tagging
- Speech recognition
- Face/emotion recognition



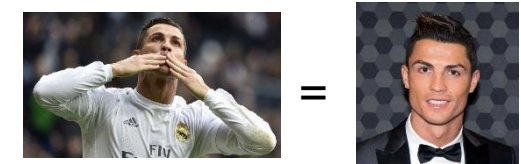


# Declarative Crowdsourcing Systems



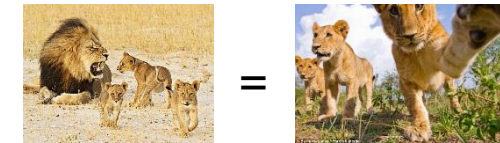
```

SELECT c.name
FROM celeb c JOIN photos p
ON samePerson(c.img,p.img)
TASK samePerson(f1, f2) TYPE EquiJoin:
SingularName: "celebrity"
LeftPreview: "<img src='%s'>",tuple1[f1]
RightPreview: "<img src='%s' >",tuple2[f2]
Combiner: MajorityVote
    
```



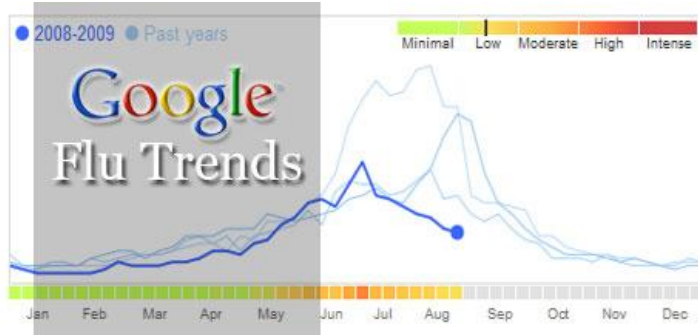
```

SELECT image i FROM serengety
ORDER BY CROWDORDER (i, "Which image
contains more baby animals");
    
```



- J. Fan et al: CrowdOp: Query Optimization for Declarative Crowdsourcing Systems, TKDE, 2015.
- Michael J. Franklin et al: CrowdDB: answering queries with crowdsourcing, SIGMOD 2011
- A. Marcus et al: Human-powered Sorts and Joins, VLDB 2011

# Gathering Input, Reusing “Natural” Human Power



# Human Computation Platforms and Motivation for Participation

## Citizen Science



- Helping/Contribute to something important
- Social pressure
- Virtual goods
- Competitions
- Gaming
- Money



## Paid Crowdsourcing

# Monetary based Motivation



amazonmechanical turk Artificial Intelligence [Your Account](#) [HITS](#) [Qualifications](#) 110,288 HITS available now [Sign In](#)

All HITS | [HITS Available To You](#) | [HITS Assigned To You](#)

Search for  containing  that pay at least \$  for which you are qualified

**All HITS**  
1-10 of 1895 Results

Sort by:   [Show all details](#) | [Hide all details](#) 1 2 3 4 5 > [Next](#) >> [Last](#)

<b>Image Tagging - Answer questions about ONE image. Great images!</b> <a href="#">View a HIT in this group</a>			
Requester: <a href="#">TagCow</a>	HIT Expiration Date: Oct 24, 2010 (2 weeks 5 days)	Reward: \$0.02	
	Time Allotted: 20 minutes	HITS Available: 14019	
<b>Find Restaurant Web Addresses</b> <a href="#">View a HIT in this group</a>			
Requester: <a href="#">Dolores Labs</a>	HIT Expiration Date: Oct 12, 2010 (6 days 23 hours)	Reward: \$0.07	
	Time Allotted: 60 minutes	HITS Available: 8773	
<b>Product Search Relevance</b> <a href="#">View a HIT in this group</a>			
Requester: <a href="#">Amazon Requester Inc.</a>	HIT Expiration Date: Oct 6, 2010 (1 day 21 hours)	Reward: \$0.01	
	Time Allotted: 10 minutes	HITS Available: 7867	
<b>Verify Restaurant Websites</b> <a href="#">View a HIT in this group</a>			
Requester: <a href="#">Dolores Labs</a>	HIT Expiration Date: Oct 11, 2010 (6 days 23 hours)	Reward: \$0.05	
	Time Allotted: 60 minutes	HITS Available: 6760	
<b>Find Business Web Addresses</b> <a href="#">View a HIT in this group</a>			
Requester: <a href="#">Dolores Labs</a>	HIT Expiration Date: Oct 11, 2010 (6 days 22 hours)	Reward: \$0.07	
	Time Allotted: 60 minutes	HITS Available: 5290	

*Human Intelligence Task (HIT)*

# Mturk: IR Example – Snippet Evaluation

- Study on summary lengths
- Determine preferred result length
- Asked workers to evaluate snippet quality
  
- Payment between \$0.01 and \$0.05 per HIT
- 12,790 queries - 40K judgments 400\$-2000\$ (300h of work)

M. Kaiser, M. Hearst, and L. Lowe. "Improving Search Results Quality by Customizing Summary Lengths", ACL/HLT, 2008. July 24, 2011 Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 50

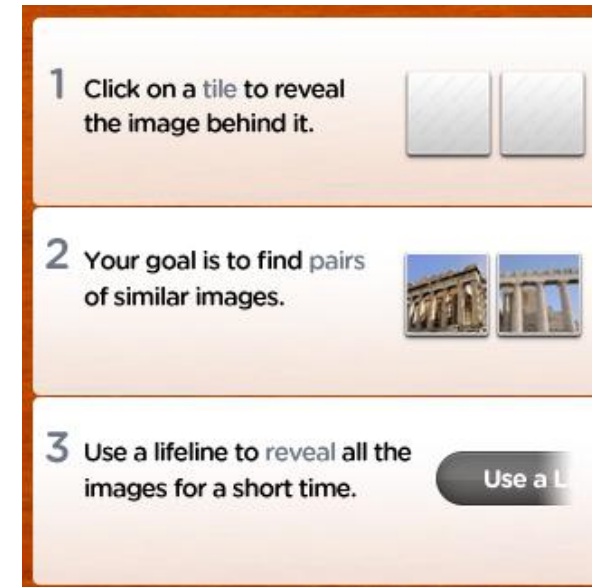
# IR Example – Relevance Assessment

- Replace TREC-like relevance assessors with MTurk?
- Selected topic “space program” (011)
- Modified original 4-page instructions from TREC
- Workers more accurate than original assessors!
- 40% provided justification for each answer
  
- Payment between \$0.02 per HIT
- 1 topic, 29 documents - 290 judgments (6\$)

O. Alonso and S. Mizzaro. “Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment”, SIGIR Workshop on the Future of IR Evaluation, 2009. July 24, 2011 Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 51

# Games

- ESP Game: label images
  - Image retrieval by text
- Squigl: match the labels to areas
  - Object recognition
- Matchin: find the better image
  - Image ranking
- Fliplt: memory with similar images
  - Near duplicate detection

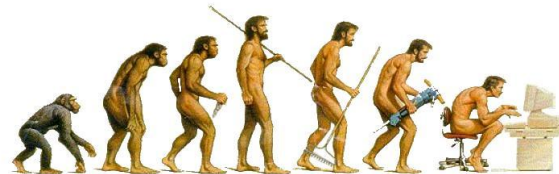


- Other areas covered as well: label songs, find synonyms, describe videos
- See: [www.gwap.com](http://www.gwap.com) by Luis von Ahn

## Useful human power for annotating the Web

- 5000 people playing simultaneously could label all images on Google in 30 days!
- Individual games in Yahoo! and MSN average over 5,000 players at a time

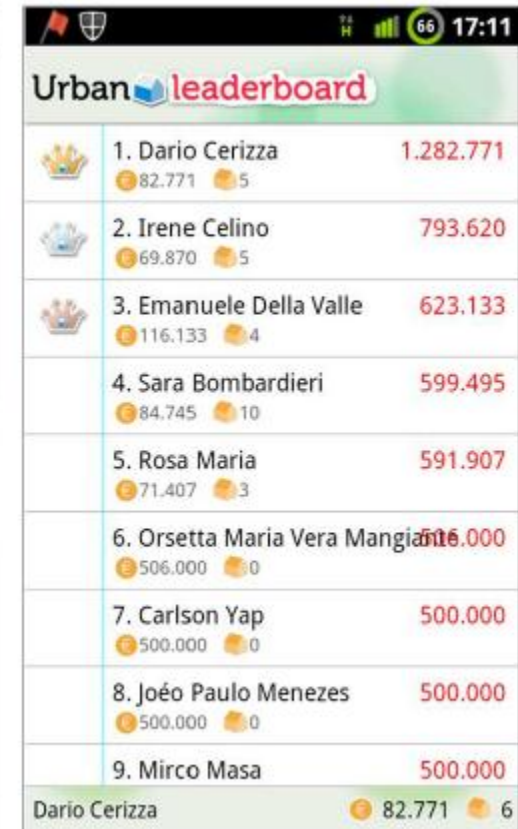
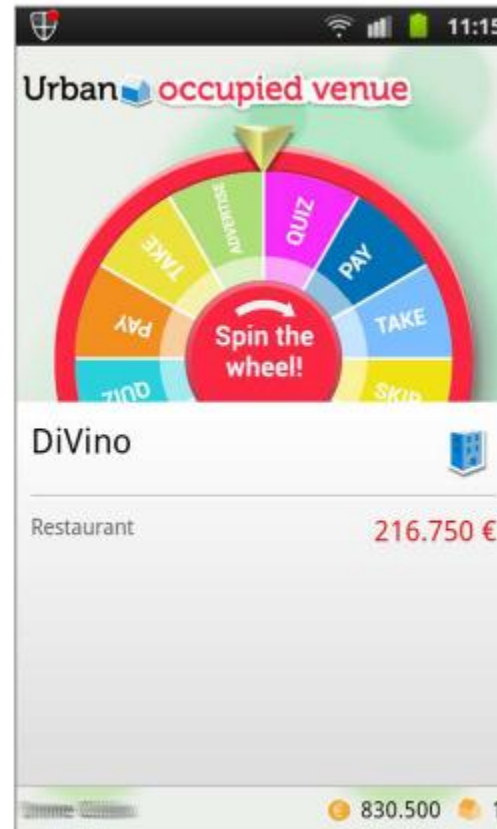
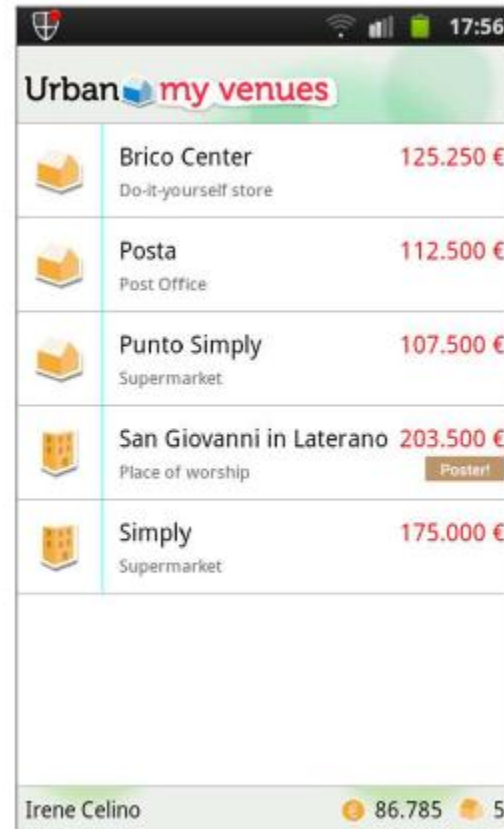
Evolution



(OR is it?)



# Urbanopoly



I. Celino et al., "Urbanopoly -- A Social and Location-Based Game with a Purpose to Crowdfund Your Urban Data," Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)

# Competition based Motivation (Image Privacy)

**Private**



**Public**



**Work**

**Sea**


**Winter**

**Water**

# Gathering average community notion of privacy


- We crawled “most recently uploaded” Flickr photos (2 Months)
- Started a social annotation game (over the course of 2 weeks)
- 81 users (colleagues, social networks , forum users) , 6 teams
- Collected around 30K annotated photos

Points: **223**



private undecidable public

„Private are photos which have to do with the private sphere (like self portraits, family, friends, your home) or contain objects that you would not share with the entire world (like a private email). The rest is public. In case no decision can be made, the picture should be marked as undecidable.“



private undecidable public

**The Best**

Rank	User	Score
1	D.A.	2984
2	Omer	2169
3	garret	1592
-----		
9	hendrick	223

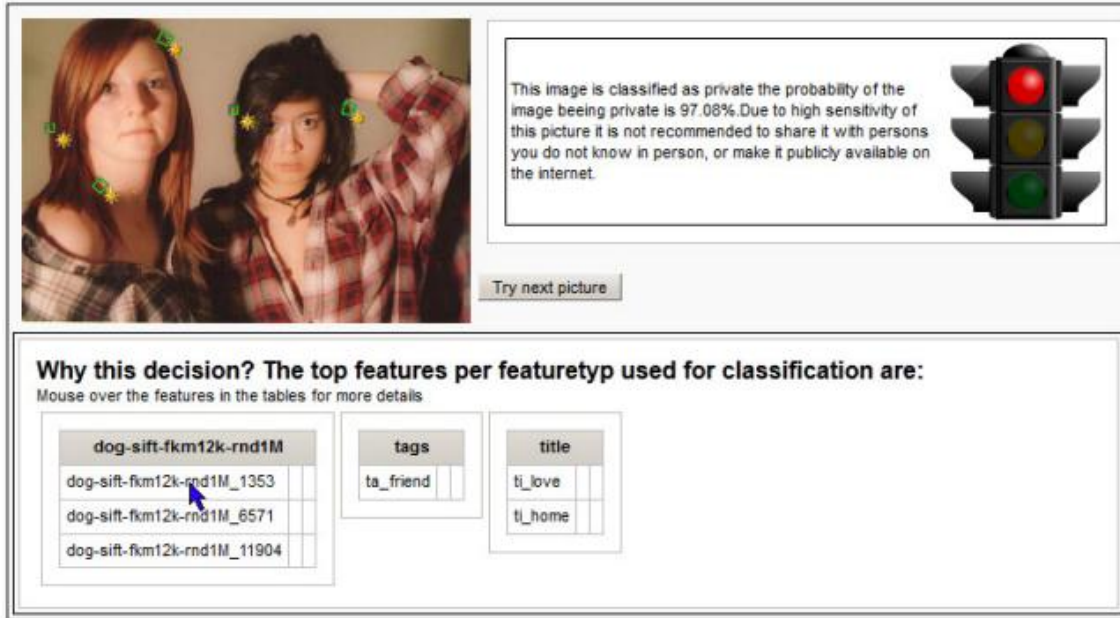
**Best teams**

Rank	Team	Score
1	vingrad	177599
2	world	101732
3	l3s	56735
-----		
4	sbnt	11221

Sergej Zerr , Stefan Siersdorfer , Jonathon Hare , Elena Demidova Privacy-Aware Image Classification and Search , SIGIR'12



# The GUI for Privacy Aware Image IR



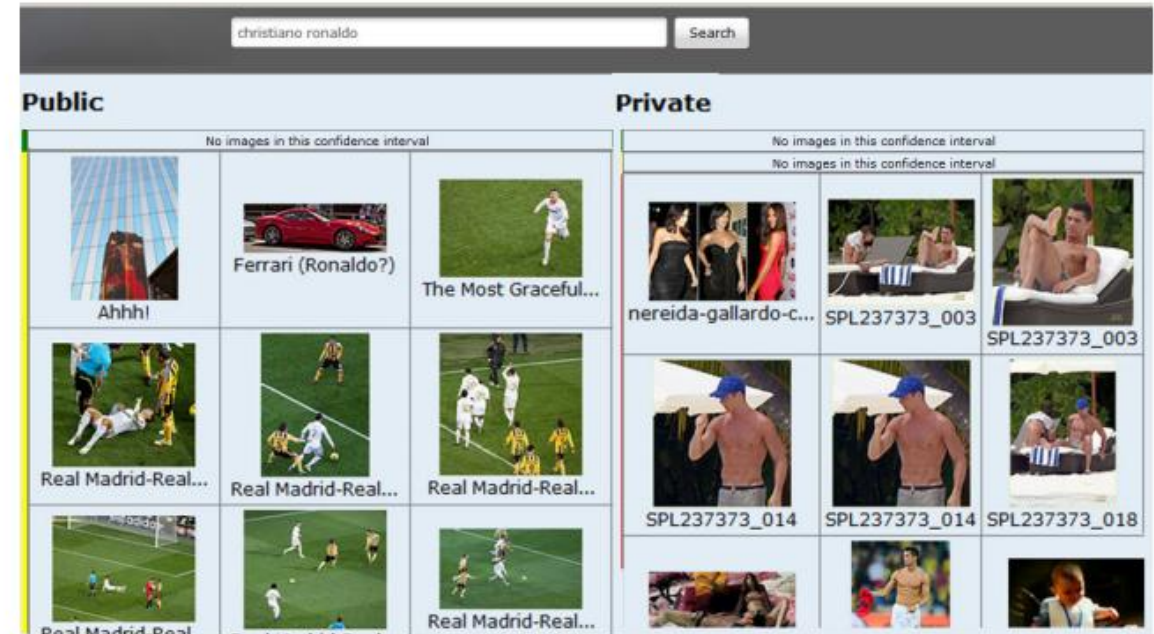
This image is classified as private the probability of the image being private is 97.08%. Due to high sensitivity of this picture it is not recommended to share it with persons you do not know in person, or make it publicly available on the internet.

Try next picture

**Why this decision? The top features per featurtyp used for classification are:**  
Mouse over the features in the tables for more details

dog-sift-fkm12k-rnd1M	tags	title
dog-sift-fkm12k-rnd1M_1353	ta_friend	ti_love
dog-sift-fkm12k-rnd1M_6571		ti_home
dog-sift-fkm12k-rnd1M_11904		

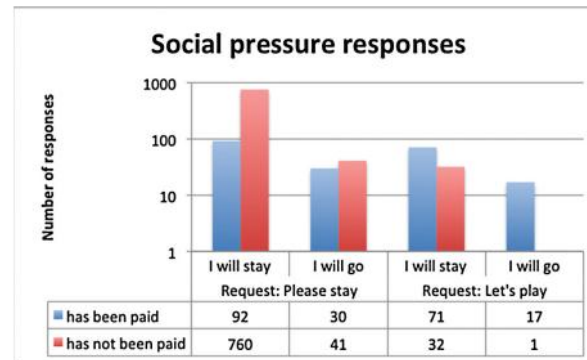
(a) Web service GUI for privacy-oriented image classification.



Search results for the query "cristiano ronaldo" (06/06/12).

(b) Search results for the query "cristiano ronaldo" (06/06/12).

# Motivation: Add Social Pressure



Oluwaseyi Feyisetan, Elena Simperl: Please Stay vs Let's Play: Social Pressure Incentives in Paid Collaborative Crowdsourcing. ICWE 2016

# Combine Gamification, Competition and Money

- **Problem:** improve time aware cost effectiveness of crowdsourcing

## Individual reward mechanisms



Competitive game designs for improving the cost effectiveness of crowdsourcing  
**CIKM'14**



## Team-based reward mechanisms



Groupsourcing: Team competition designs for crowdsourcing  
**WWW'15**



## Temporal-based crowdsourcing performance



Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions  
**WWW'16**



# Reward Distribution 1: “Pay-per-Task” (Baseline)

## Reward Distribution 1: “Pay-per-Task” (Baseline)

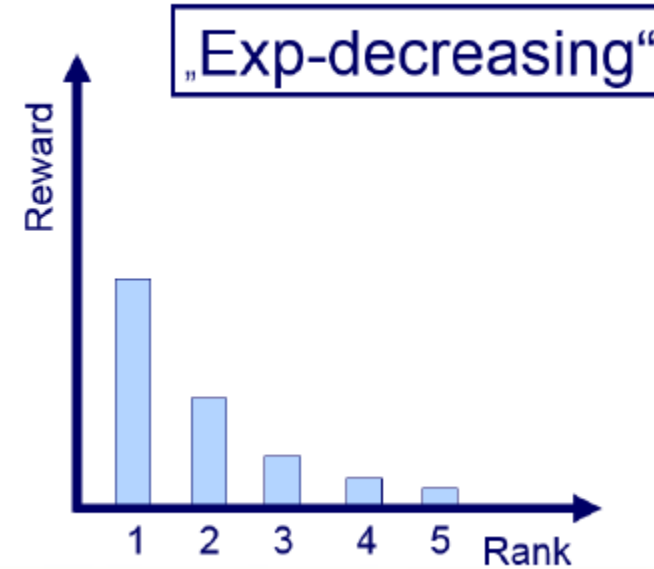
- Fixed reward rate  $c$  (\$ per task) for each worker
- Reward of workers proportional to value produced by worker (e.g. no. of annotations, ratings, etc.)



M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. CIKM'14

# Reward Distribution 2: Competitions

- Workers compete during limited time period
- Workers obtain scores based on their performance (e.g. no. of tasks fulfilled)
  - Ranking of the workers based on their performance
  - Distributing of the rewards according to the rank





# Workers' View: Tasks

**Captcha Competition** your user code: 18C77EFF17FC4C30BB0685A456F2D7E4 markus  
Logout

Solve Captchas in order to gain points

qaragixev

rimikecera

Yexifamas

ofiyegimek

kedejavape

Submit and make more points


This game will end in 1 day, 19 hours and 11 minutes.

[leave a comment](#)






Highscore		
Rank	User	Points
1	Tiger	14082
2	Duck	12912
3	Ferret	8614
4	markus	8256
5	Kraken	5251
6	Gnu	3982
7	Llama	3272






**Face Identification Competition** markus  
Llama: Ok, I'll do my best

Select the image that most clearly shows the same person or one of the persons as the following reference image:



(If there is more than one image showing the person, pick the best one.)

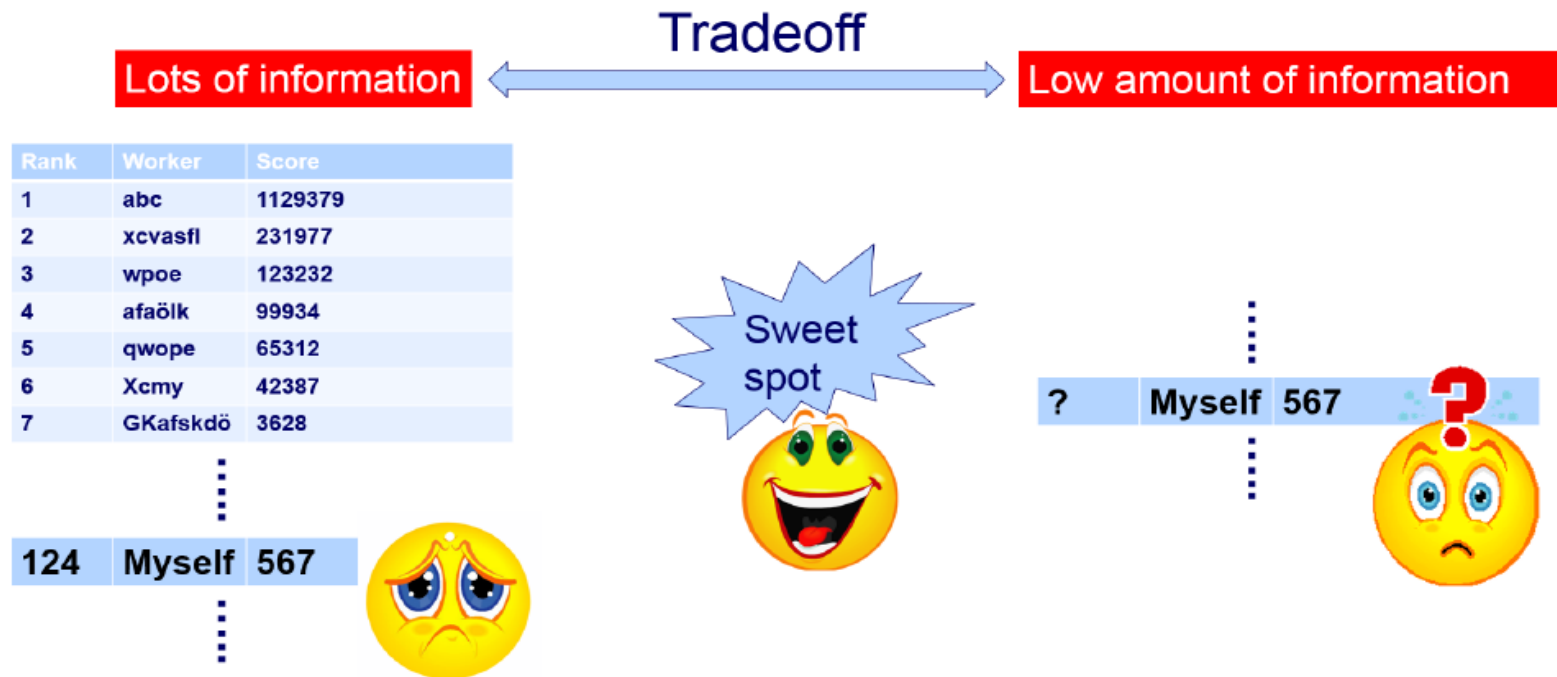
show team details and messages new team message(s)

Progress in your current batch: (1/100)

Highscore		
Rank	User	Points
1	Tiger	14080
2	Duck	12920
3	Ferret	8600
4	markus	8460
5	Kraken	5220
6	Gnu	3980
7	Llama	3280

# Information Policies

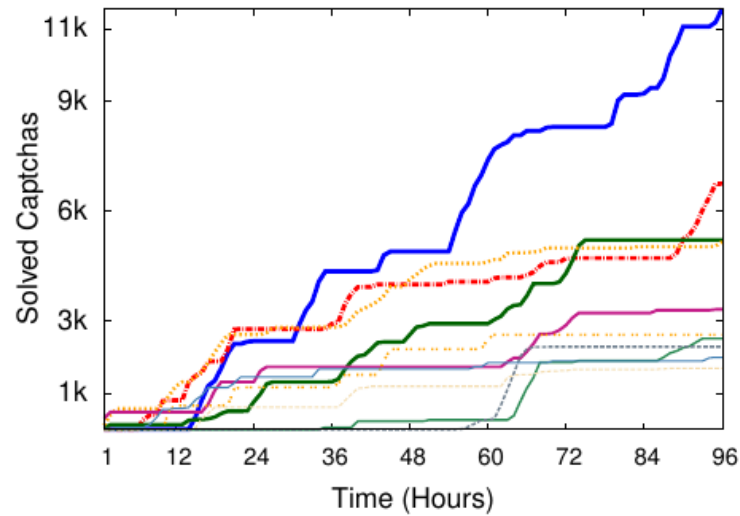
- How much information about fellow workers to provide during competition?
- Information: scores, rank



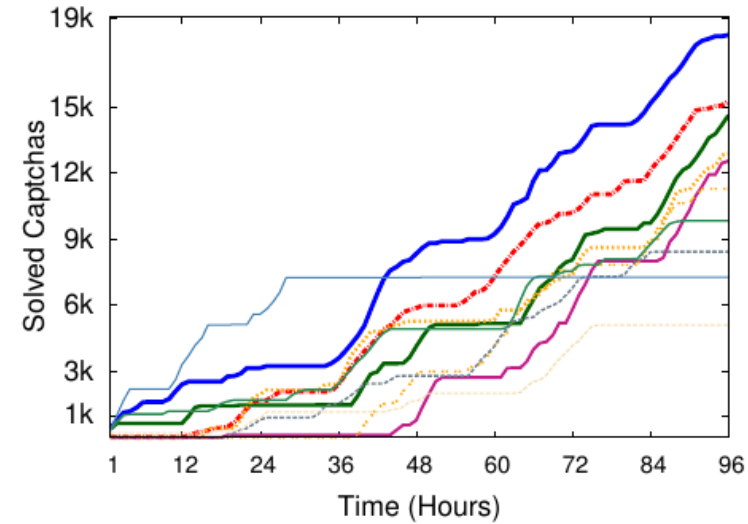
# Performance of Strategies (Captcha Task)

Experiment	No. Captchas	USD/Hour	Cent/Captcha
<b>Performance-based payment</b>			
exp-open	67,951	0.300	0.074
<b>exp-med</b>	154,188	0.138	0.032
exp-res	25,853	0.605	0.193
<b>pay-per-task</b>	58,635	0.352	0.077

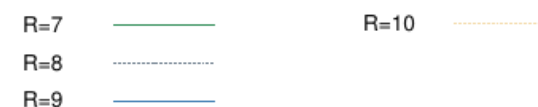
# Competition among Top-10 Workers



(a) Baseline

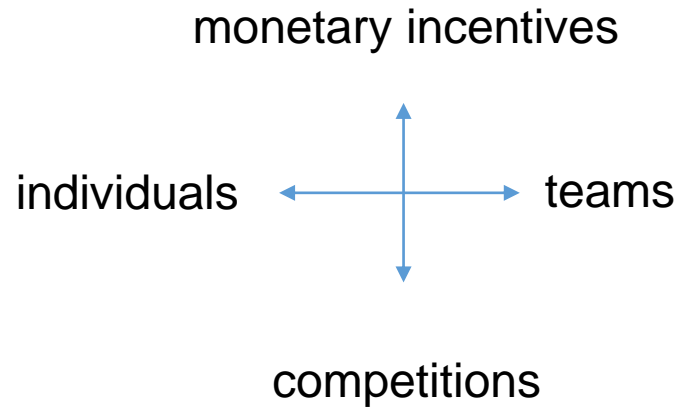


(b) Exp-med



# Team-based reward mechanisms

Can we use work groups to further improve the performance?



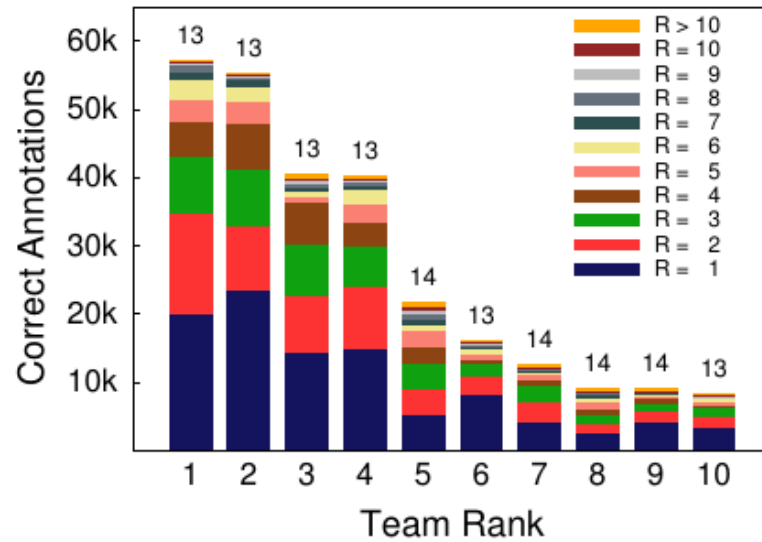
M. Rokicki, S. Zerr, and S. Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. WWW'15

- Rewards:
  - Non-linear distribution among teams
  - Individual share proportional to contribution
- Communication:
  - Team chats with notifications
- Combinations with individual reward
  - **balanceTS**
  - **ind-balanceTS**

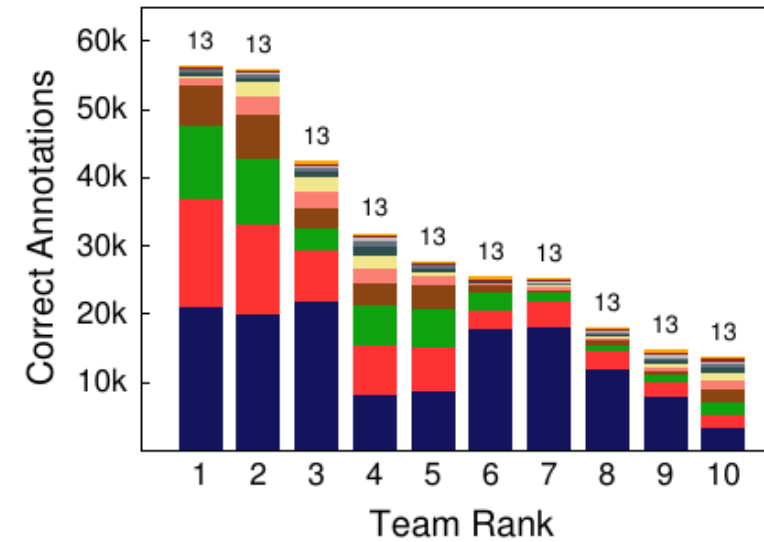
# Performance of the Strategies

Experiment	No. Images	Cent / Hour	Cent / 100 Images
<b>Baseline</b>			
ind	298,332	9.895	3.352
<b>Balanced Teams</b>			
balanceTS	327,073	8.967	3.057
<b>ind-balanceTS</b>	391,620	8.059	2.553

# Results: Team Contributions



(a) balanceTS



(b) ind-balanceTS



# Workers Interaction

- Communication in team chats
  - 2,500 messages by over 200 participants
- Encouragement
- Help and clarification of rules
- Discussing strategy
- Democratic team administration
- Discussing our strategies

Lets go team !!! we are 5, team A are 3.  
We can reach them !!!

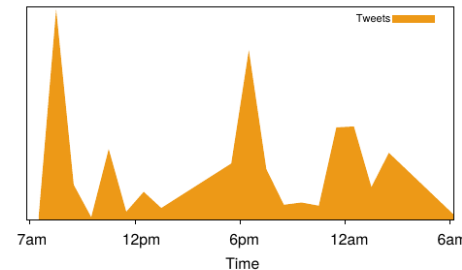
**user 1**  
What if I answer wrong?  
**user 2**  
we will lose 20 points :)

**user 1**  
Im trying to get to number 5 spot because he/she stopped clicking.  
**user 2**  
Yeah but u need 2000 thousand more buddy,  
and you know that he/she will be careful now ./  
she will check again to see if you will attack and  
then he/she will start doing more [...]  
**user 1**  
good point

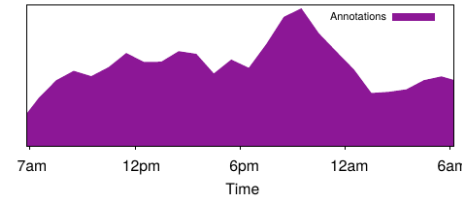
[...] this system.. its stable and  
perfect.. all in our hands(public) but  
not of system automatically selecting  
arranging them in teams..

# Temporal-based crowdsourcing performance

Can we control the crowd to annotate at right times?



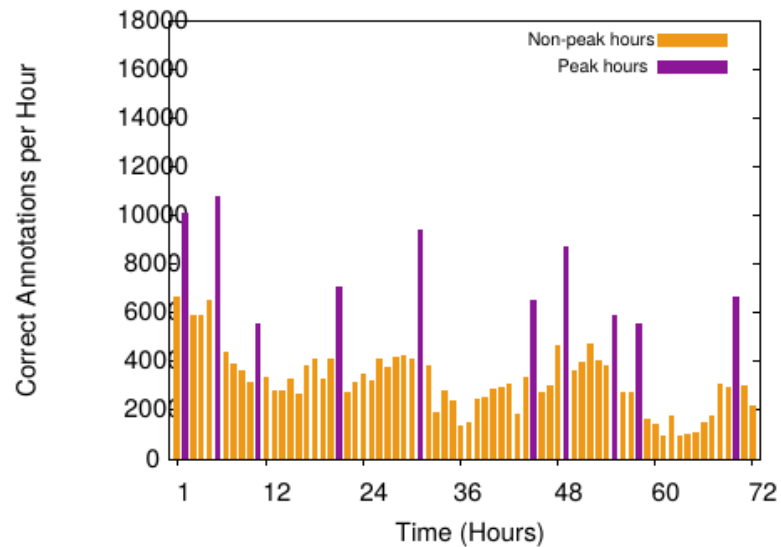
(a) Tweet volumes



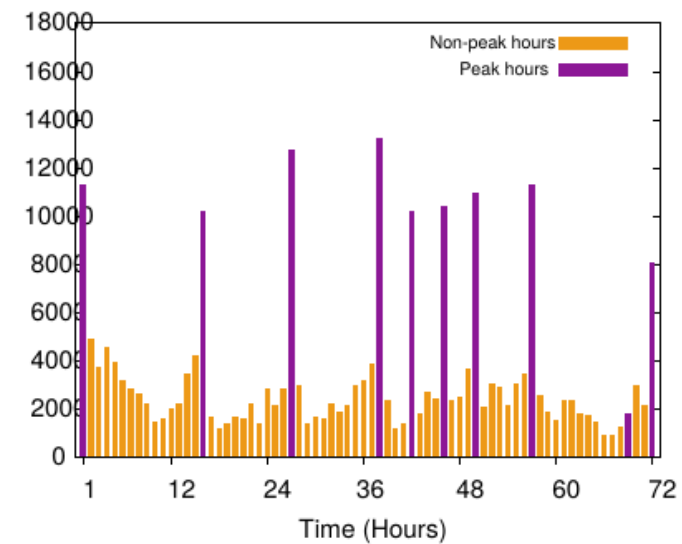
(b) Annotation output

M. Rokicki, S. Zerr, and S. Siersdorfer. Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions. WWW'15

# Peak vs NonPeak

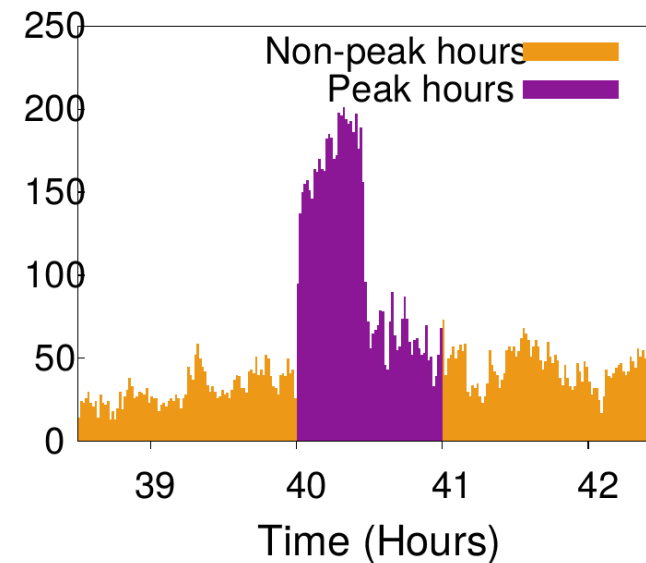
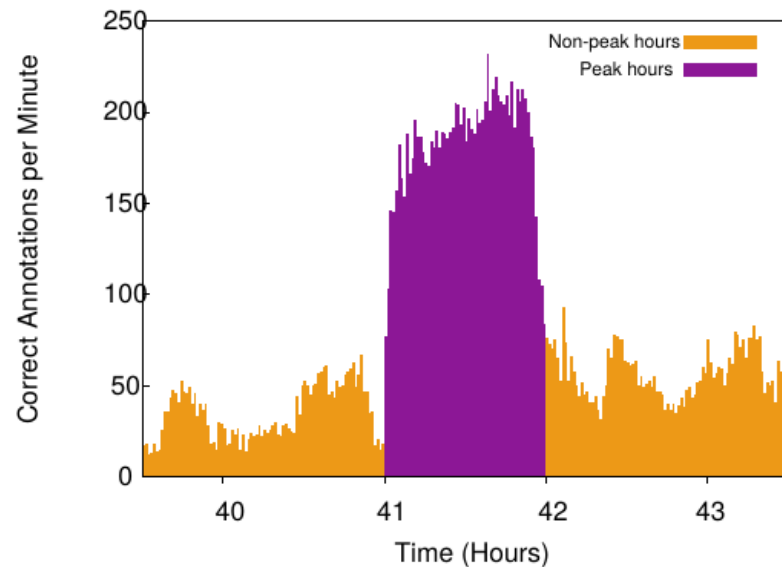


(a) short notice, low bonus

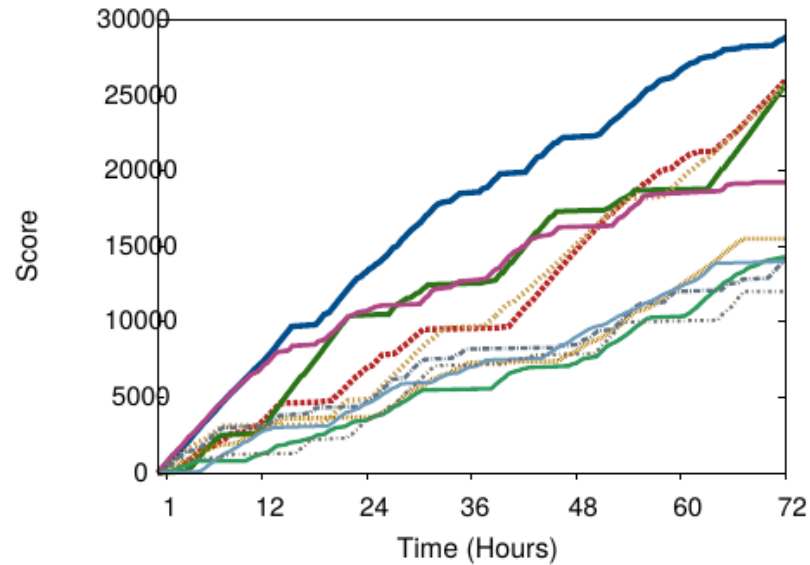


(f) long notice, high bonus

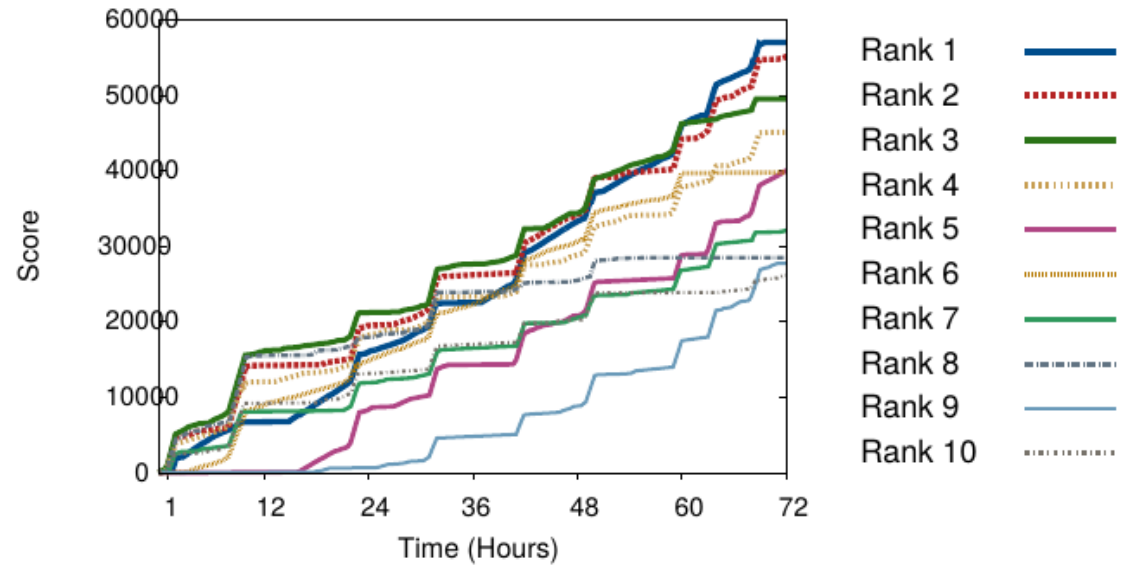
# Number of correct annotations per minute around a typical bonus hour.



# Annotation dynamics



(a) Baseline individual scores



(b) Long notice, medium bonus individual scores

# Workers Interaction

- Communication in team chats
  - 3,400 messages by over 200 participants

Bonus coming 14 minutes from now.  
prepare everyone, we must try that first  
place :D

our rank fall from rank 4-5 to 12 bcoz  
the other team work on time of bonus

as soon as they announce the time  
of the next bonus I will email you

# Contributions

## Individual reward mechanisms



Competitive game designs for improving the cost effectiveness of crowdsourcing  
**CIKM'14**



Medium Information policy and exponential rewards significantly increase crowd performance by 300%

## Team-based reward mechanisms



Groupsourcing: Team competition designs for crowdsourcing  
**WWW'15**



Balanced teams + individual rewards further increase performance and make the work more attractive (+30%)

## Temporal-based crowdsourcing performance



Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions  
**WWW'16**



Framework using our strategies additionally increased output in peak times by more than 300%

# Motivation: Contribute to Science (“Zooniverse”)

The screenshot shows the Zooniverse website interface. At the top, there is a navigation bar with icons for different disciplines: All Disciplines (selected), Arts, Biology, Climate, History, Language, Literature, Medicine, and Nature. Below the navigation bar, there is a search bar with the text "Most Recently Launched" and a dropdown menu. To the right of the search bar, it says "Showing 1-20 of 42 projects found." and there is a "Name:" search field. Below the search bar, there are three numbered tabs (1, 2, 3). The main content area displays a grid of project cards. Each card features a representative image and a title. The projects shown are:

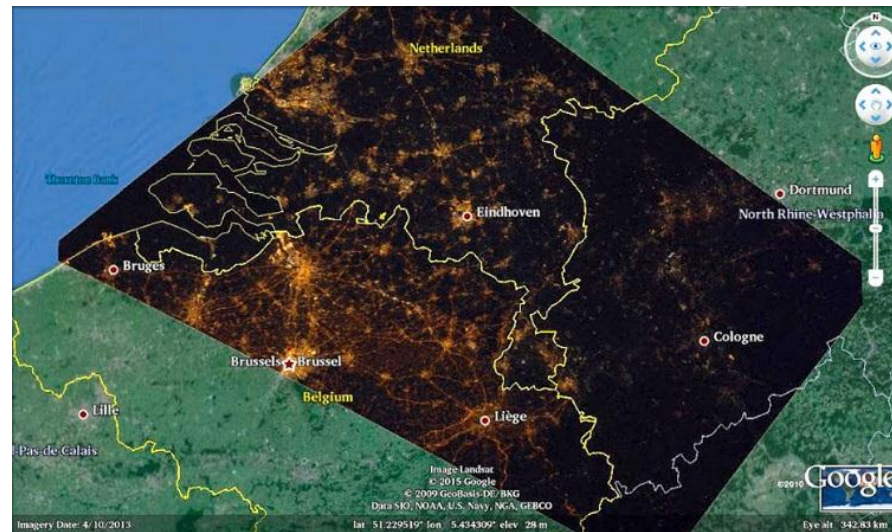
- DECODING THE CIVIL WAR**: Image of a document with the text "U. S. Military Telegraph. Strictly Private".
- NOTES FROM NATURE**: Image of a forest with a grid overlay.
- PATTERN PERCEPTION**: Image of a colorful, abstract pattern resembling an eye.
- SNAPSHOT WISCONSIN**: Image of three small icons (a tree, a sun, and a cloud). Text: "Welcome to Snapshot Wisconsin. Help us identify animals in trail camera images."
- COMPUTER VISION: SERENGETI**: Image of a gazelle in a field.
- WESTERN SHIELD — CAMERA WATCH**: Image of a brown squirrel.
- MICROSCOPY MASTERS**: Image of a colorful, abstract, textured object.
- POPPIN' GALAXY**: Image of a galaxy in space.
- SNAPSHOTS AT SEA**: Image of a whale breaching the ocean surface.
- COMET HUNTERS**: Image of a bright comet in a dark sky.

<https://www.zooniverse.org/>



# Motivation: Contribute to Science (“Cities at Night”)

- Classification of night photos from ISS to estimate artificial light pollution in cities
- Observe temporal development, measure impact on citizens and biosphere

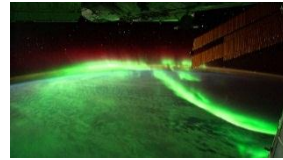


<http://stars4all.eu>

<http://www.citiesatnight.org/>

# Motivation: Contribute to Science (“Cities at Night”)

- Task 1: “Dark skies” – Find night cities in a photo stream (over 100K annotated)

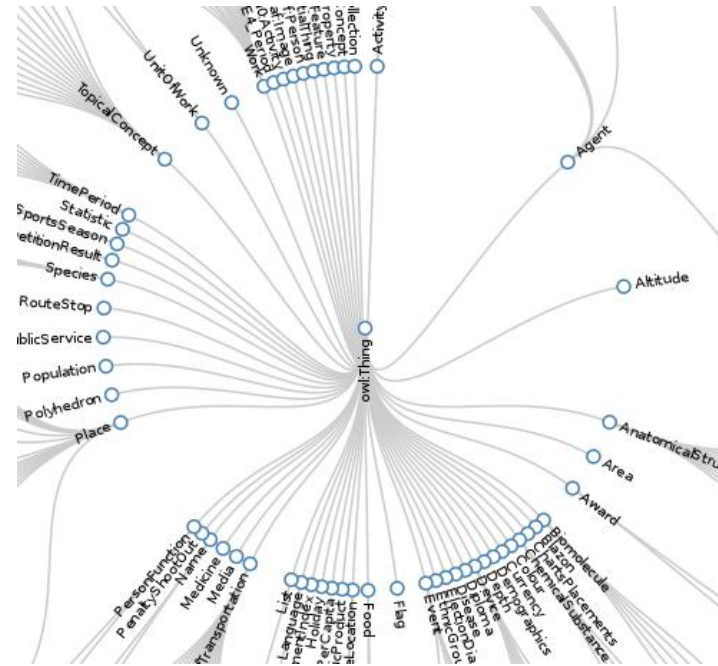
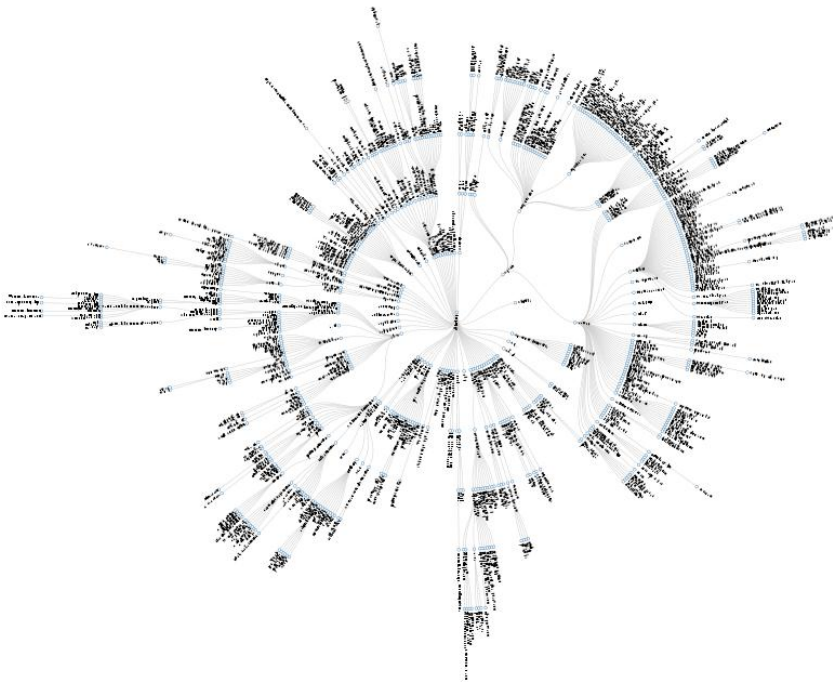


- Task 2: “Lost at night” – identify the city on the photograph (around 500 identified)



- Task 3: “Night cities” – position, rotate and scale the image to the map.

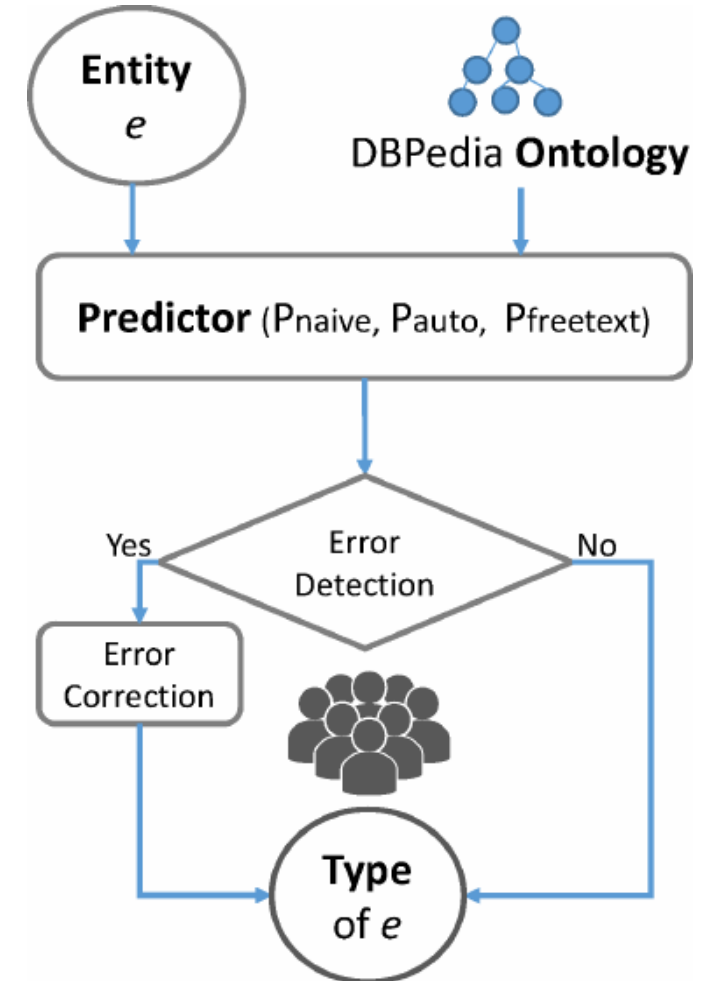
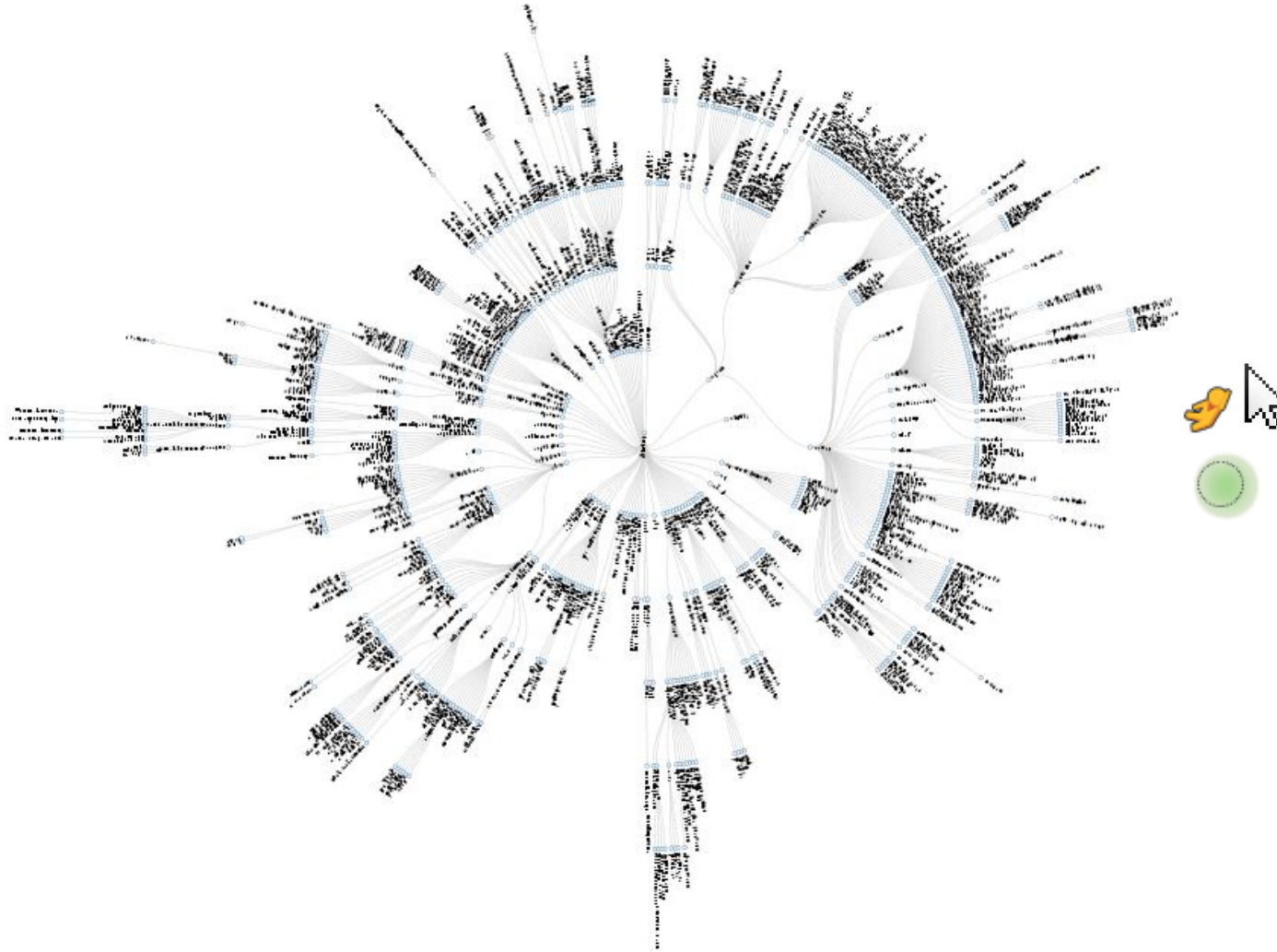
# Combine Human and Machine Input



- Assign DBpedia class to entities
- Baby food, Petroleum industry in Nigeria, Light infantry
- "Region", "Locality", "Settlement"

Using microtasks to crowdsource Dbpedia entity classification: A study in workflow Design. Qiong Bu, Elena Simperl, Sergej Zerr and Yunjia Li

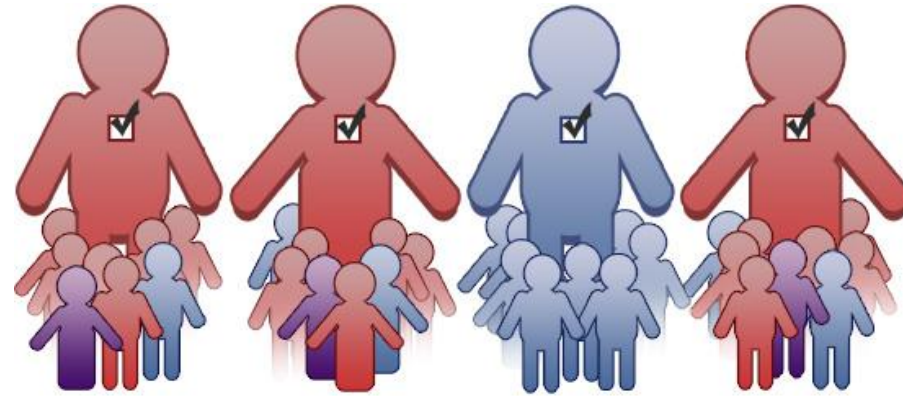
# Combine Human and Machine Input





# Output Aggregation

- Statistical models
  - Majority voting
- Graphical models
- Optimization models



# Outline

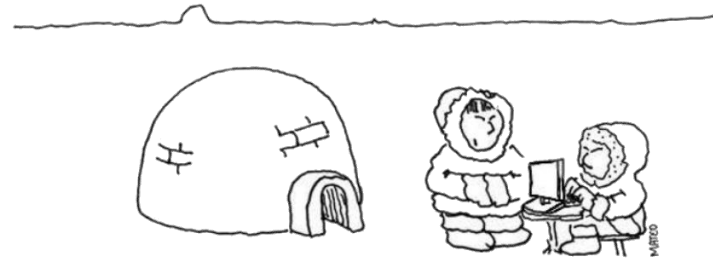
- Collaborative Advantages
  - The Wisdom of Crowds
  - Conditions for a successful collaboration
- Obtaining collaborative knowledge
  - Gathering Data from Social Web / Mechanical Turk
  - From Data to knowledge (Applications)
  - Own work
- **Input/Output Evaluation**
  - Users and Data
  - Quality assurance
- Discussion

# Asking questions

- Ask / formulate the right questions
- Part art, part science
- Instructions are key
- Workers may not be IR experts (don't assume the same understanding in terms of terminology)
- Show examples

N. Bradburn, S. Sudman, and B. Wansink. Asking Questions: The Definitive Guide to Questionnaire Design, Jossey-Bass, 2004

# Quality: Ambiguity and Subjectivity



What is relevant?

*"Snow. Snow is relevant."*



Nederland, netherlands, holland, dutch  
Rotterdam, wielrennen, cycling, duck  
le grand depart, tour de france,  
Reklame, caravan, Funny Fotos

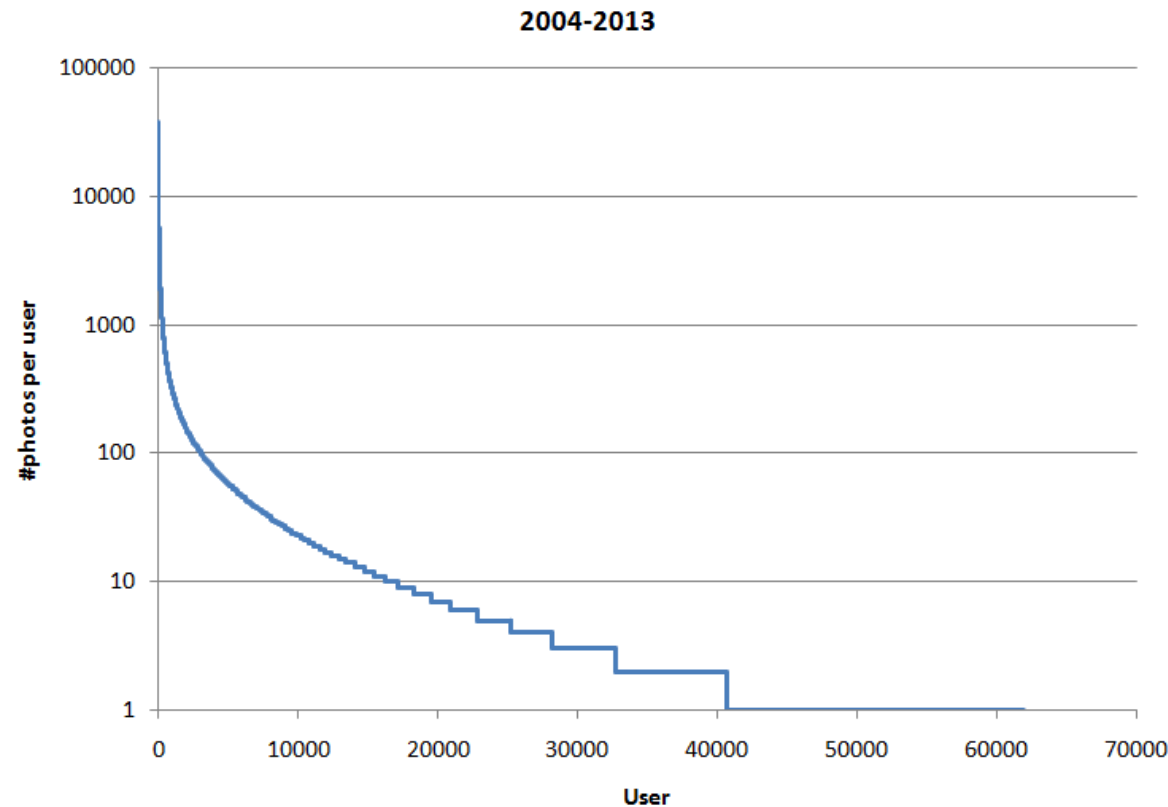
„Alice saw Bob  
with the binoculars“





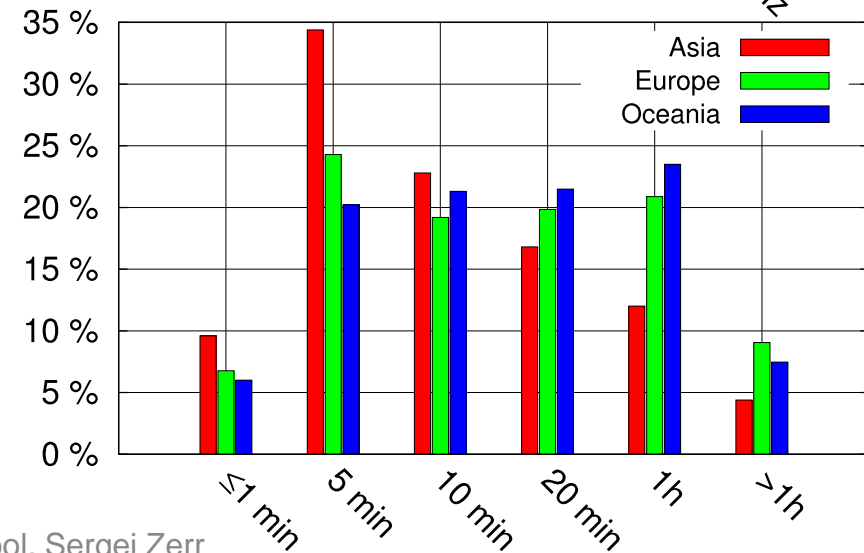
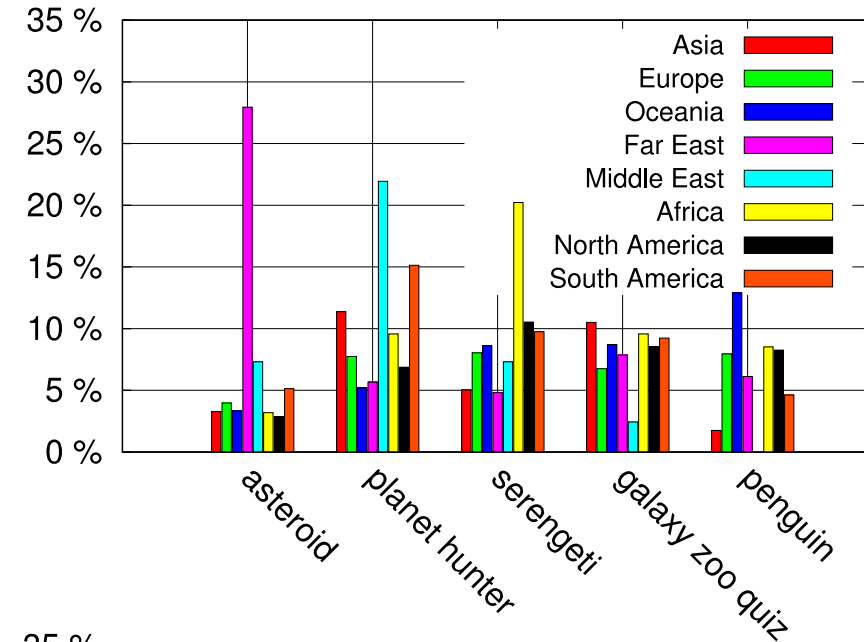
# Quality: Data from Social Web

- Simple random sample can result in a set dominated by few power user












# Demographic Bias (Zooniverse)

Region	Classifications
Europe (UK, Germany, France)	3688453 (48.2%)
North America (USA, Canada, Mexico)	3071134 (40.2%)
Oceania (Australia, New Zealand, Tanzania)	347818 (4.6%)
Asia (Singapore, India, Japan)	277536 (3.6%)
Far East	37278 (0.5%)
Middle East	15318 (0.2%)
South America (Brazil, Argentina, Chile)	154807 (2.0%)
Africa, Egypt, Kenya)	50045 (0.7%)



S. Zerr, R. Tinati, M. Luczak-Roesch, and E. Simperl: Investigating the Global Adoption of Citizen Science. Collective Intelligence 2016

# Rater Reliability “Where is the cat?”

						
V	V	X	V	V	X	
X	X	X	X	X	X	
V	X	X	V	X	V	
V	X	X	V	X	X	<b>Results</b>

# Quality Assurance

- Qualification Tests
- Test questions
- „Static“ Honeypots
- „Dynamic“ honeypots
- Workers' reputation mechanisms
- Inter-rater agreement

# Test Questions

Throw the coin and tell us the result

- Head ●
- Tail ●

Results

- **Head 61**
- **Tail 39**

People often tend  
just to select the  
first option 😞



**Better:** Some preliminary textual answer

- **Coin type?**
- **Head or tail.**

Matthew Lease and Omar Alonso: <http://de.slideshare.net/mattlease/crowdsourcing-for-search-evaluation-and-socialalgorithmic-search>

# Honeypots

- Static honeypots
  - Let the workers perform the task. Reject the results with honeypot errors
- Dynamic batches with injected honeypots
  - Only reject the low quality batches



# Measure the Inter-Rater Reliability

		A		total
		Yes	No	
B	Yes	2		
	No		1	
total				

Naive approach: 3 cases out of 6 = 0.5 agreement

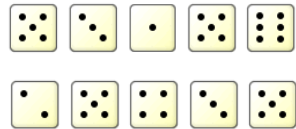
Kilem L. Gwet, Handbook of inter-rater reliability 2010

	MR. GOOD A	MR. GOOD B	<b>Attractive?</b>
	1	1	
	0	1	
	0	0	
	1	0	
	1	0	
	1	1	

# Statistic Significance

## First experiment:

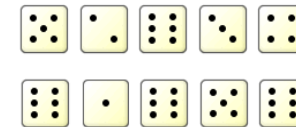
Throw the dices with the right hand  
10 times.



Compute the average  
3.7

## Second experiment:

Throw the dices with the left hand  
10 times.



Compute the average  
4.4

Claim that the left hand is better.....



# Inter-Rater Reliability: Cohen's Kappa(1960)

➤ Idea: We need to remove agreement achieved just by chance

$$\hat{\gamma}_{\kappa} = \frac{p_a - p_e}{1 - p_e}$$

		A		total
		Yes	No	
B	Yes	35	20	55
	No	5	40	45
total		40	60	100

		A		total
		Yes	No	
B	Yes	$n_{11}$	$n_{12}$	
	No	$n_{21}$	$n_{22}$	
total				

$$p_a = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} = \frac{35 + 40}{100} = 0.75$$

$$p_e = \frac{55}{100} * \frac{40}{100} + \frac{45}{100} * \frac{60}{100} = 0.49$$

$$\hat{\gamma}_{\kappa} = \frac{0.75 - .49}{1 - .49} = .51$$

# Inter-Rater Reliability: Missing Values

- Idea: Use partial ratings to estimate the marginal probability only

		A			total
		Yes	No	X	
B	Yes	30	18	2	50
	No	5	34	3	42
	X	5	3	0	8
total		40	55	5	100

$$p_a = \frac{n_{11} + n_{22}}{n - (n_{x1} + n_{x2} + n_{1x} + n_{2x})} = \frac{30 + 34}{100 - (5 + 8)} = .74$$

$$p_e = \frac{50}{100} * \frac{40}{100} + \frac{42}{100} * \frac{55}{100} = 0.431$$

		A			total
		Yes	No	X	
B	Yes	$n_{11}$	$n_{12}$	$n_{x1}$	
	No	$n_{21}$	$n_{22}$	$n_{x2}$	
	X	$n_{1x}$	$n_{2x}$	0	
total					

$$\hat{\gamma}_k = \frac{0.74 - .431}{1 - .431} = .54$$

# Inter-Rater Reliability: Extensions

- **Multiple Raters/Categories:**
  - Fleiss 1971 – Average over random pairs of raters for random objects
  
- **Adjustment for Ordinal and Interval Data, Weighting:**
  - weight judgments using distances between categories.
  - Measures:  $AC_1$ ,  $AC_2$  (ordinal and interval data)
  
- **Check for statistical significance:**
  - The number of categories and/or raters matters.

Kilem L Gwet: andbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters, 2014

# Inter-Rater Reliability: Kappa Interpretations

Koch	
Kappa	Strenght of Agreement
<0.0	Poor
0.0 – 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 100	Almost Perfect

Fleiss	
Kappa	Strenght of Agreement
0.0-0.40	Poor
0.41 – 0.75	Intermediate to Good
>0.75	Excellent

Altman	
Kappa	Strenght of Agreement
<0.20	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good
0.81 - 100	Very Good

**Please note:** These interpretations were proven to be usefull mostly in medical domain (diagnosis)

# Summary

- Wisdom of the Crowd: Collective Intelligence and Groupthinking
- Obtaining Collaborative Knowledge: Motivation in Paid Crowdsourcing and Citizen Science
- Result Aggregation and Quality Assurance

## Discussions / Questions / Remarks



# Outline

- Collaborative Advantages
  - The Wisdom of Crowds
  - Conditions for a successful collaboration
- **Small experiment**
  - Can we collaborate?
- Obtaining collaborative knowledge
  - Crowd motivation
  - Scalability/Efficiency
  - Own work
- Input/Output Evaluation
  - Users and Data
  - Quality assurance
- Discussion

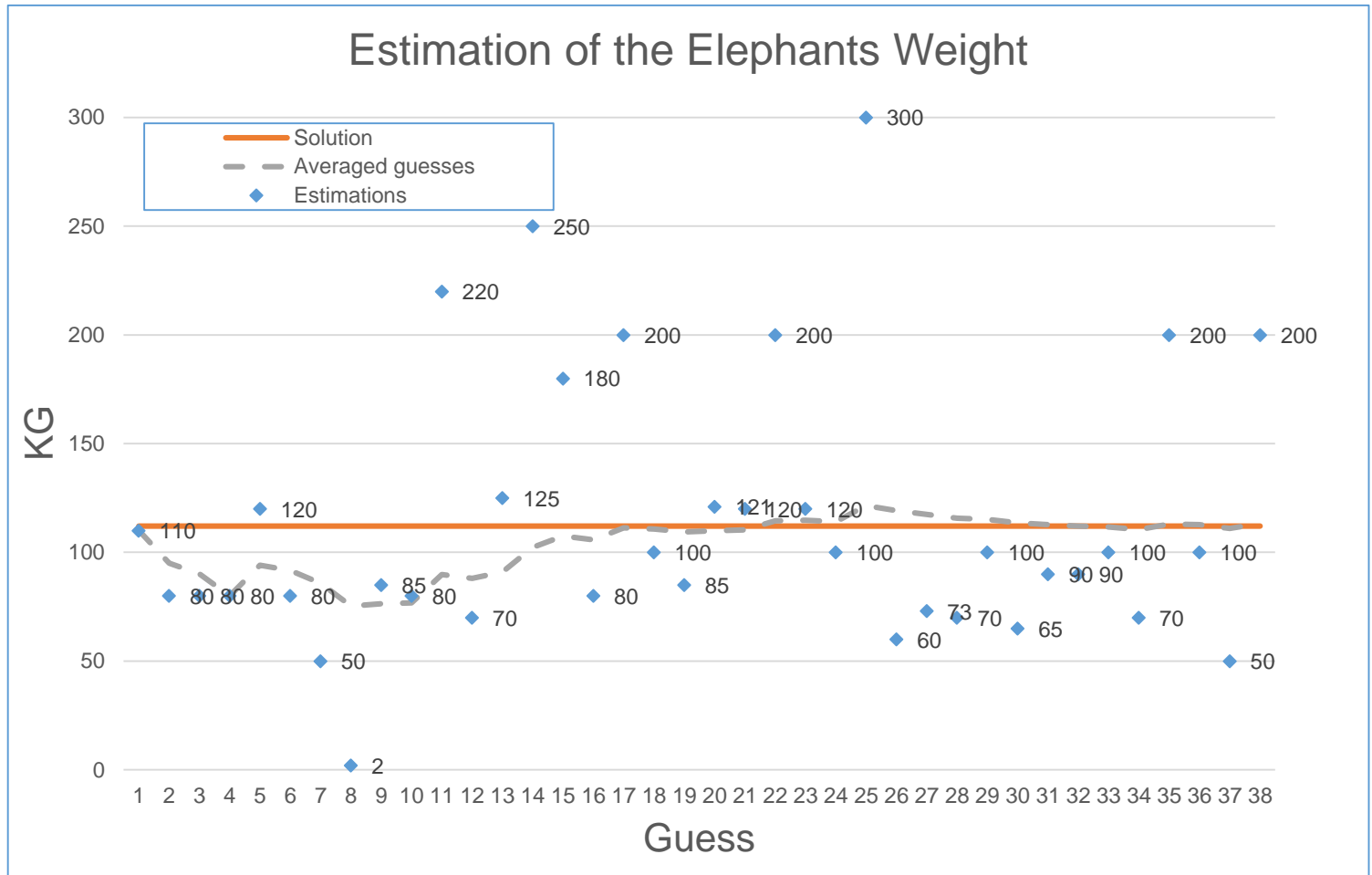


# Results of the Experiment:

Baby Elephant: [http://www.zimbio.com/pictures/zrf\\_WCjHyqn/Baby+Elephant+Born+Munich+Zoo/H\\_xQAzpvSP9](http://www.zimbio.com/pictures/zrf_WCjHyqn/Baby+Elephant+Born+Munich+Zoo/H_xQAzpvSP9)

- The real weight of the Babyphant: 112 KG
- Average of the 38 estimates: **113,32 KG**
- Max/Min guesses: 300/2

The graph shows the single estimations as blue points, the average after each estimate as the grey dotted line and the real value as the orange line.



# References

- Oluwaseyi Feyisetan, Elena Simperl: Please Stay vs Let's Play: Social Pressure Incentives in Paid Collaborative Crowdsourcing. ICWE 2016
- J. Fan et al: CrowdOp: Query Optimization for Declarative Crowdsourcing Systems, TKDE, 2015.
- Kilem L Gwet: Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters, 2014
- Lei Chen (HKUST), Dongwon, and MeihuiZhang: Crowdsourcing in Information and Knowledge Management, 2014 CIKM Tutorial
- Yoram Bachrach, Thore Graepel, Gjergji Kasneci, Michal Kosinski, Jurgen Van Gael: Crowd IQ: aggregating opinions to boost performance. AAMAS 2012
- I. Celino et al., "Urbanopoly -- A Social and Location-Based Game with a Purpose to Crowdsourc Your Urban Data," Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)
- Matthew Lease, Omar Alfonso: Crowdsourcing for Search Evaluation and Social-Algorithmic Search, 2012 SIGIR Tutorial
- Omar Alfonso, Matthew Lease : Crowdsourcing for Information Retrieval: Principles, Methods, and Applications, 2011
- Michael J. Franklin et al: CrowdDB: answering queries with crowdsourcing, SIGMOD 2011
- A. Marcus et al: Human-powered Sorts and Joins, VLDB 2011
- M. Kaisser, M. Hearst, and L. Lowe. "Improving Search Results Quality by Customizing Summary Lengths", ACL/HLT, 2008.July 24, 2011
- Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 50
- O. Alonso and S. Mizzaro. "Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment", SIGIR Workshop on the Future of IR Evaluation, 2009. July 24, 2011 Crowdsourcing for Information Retrieval: Principles, Methods, and Applications 51
- N. Bradburn, S. Sudman, and B. Wansink. Asking Questions: The Definitive Guide to Questionnaire Design, Jossey-Bass, 2004
- James Surowiecki : The Wisdom of Crowds, 2004
- Sergej Zerr , Stefan Siersdorfer , Jonathon Hare , Elena Demidova Privacy-Aware Image Classification and Search , SIGIR'12
- M. Rokicki, S. Chelaru, S. Zerr, and S. Siersdorfer. Competitive game designs for improving the cost effectiveness of crowdsourcing. CIKM'14
- M. Rokicki, S. Zerr, and S. Siersdorfer. Groupsourcing: Team competition designs for crowdsourcing. WWW'15
- M. Rokicki, S. Zerr, and S. Siersdorfer. Just in Time: Controlling Temporal Performance in Crowdsourcing Competitions. WWW'15