



# SOCIAL SEARCH

# Outline

---

- Intro
- Basics of probability and information theory
- Retrieval models
- Retrieval evaluation
- Link analysis
- From queries to top-k results
- Social search
  - Overview & applications
  - Clustering & recommendation in social networks

# Social search overview (1)

---

## ➤ Background

- Rise of **Web 2.0 platforms** (e.g., Wikipedia, Facebook, Youtube, Twitter, LinkedIn, IMDB, flickr, ...) where **users are main content providers** and can give (implicitly or explicitly) recommendations, suggestions, likes, dislikes, or answers on other users' postings or questions

## ➤ Key differences to classical web search

- Users interact with the system and other users (implicitly or explicitly)
- Users provide rich information about themselves and other users (implicitly or explicitly)

## Social search overview (2)

---

- Important concepts
  - User feedback (e.g., user tags, answers, suggestions, ...)
  - Folksonomies (collaboratively derived tag taxonomies)
  - User communities (groups of users with similar properties, interests or goals)
  - “Wisdom of the Crowds” (many people can be smarter than a few)
  - Crowdsourcing (tasks difficult to solve by machines are solved by humans)
  - Information cascades (users relying on other users’ information)
  - Influential users (e.g., opinion leaders, celebrities, innovators, ...)
  - Collaborative filtering (detecting correlations between users and items)

# User feedback for social search (1)


## ➤ Search for social opinions/suggestions


WEB IMAGES VIDEOS MAPS LOCAL MORE

seattle restaurants 🔍

RECENT SEARCHES seafood · honolulu · chicago weather

176,000,000 RESULTS Price ▾ Cuisine ▾ Rating ▾

**Seattle restaurants**  Ads  
[www.DexKnows.com](http://www.DexKnows.com) · Find local restaurants in the Seattle area.

**IL Fornaio Restaurant**   
[www.ilfornaioseattle.com](http://www.ilfornaioseattle.com) · Fresh Pastas, Sauces and Breads. Private Banquet Rooms Available.  
600 Pine Street, Seattle · (206) 430-1556


**Seattle Restaurants | Urbanspoon**  
[www.urbanspoon.com/c/1/Seattle-restaurants](http://www.urbanspoon.com/c/1/Seattle-restaurants) ▾  
May 02, 2012 · Seattle restaurant reviews from critics, food blogs and fellow diners.


Ray's Boathouse ×

Recent reviews



★★★★☆ OpenTable  
Great location. Excellent food! I'll definitely choose this place again for my future guests! ... [Read more](#)



★★★★★ OpenTable  
Great place to celebrate! Melt in your mouth fish options. dreamv mashed



Bing ▾  
1648 

Ask friends... 

FRIENDS WHO MIGHT KNOW

 Jason Carlson likes Seattle, Washington 

 Lisa Miller likes Frank's Oyster House & Champagne Parlor 

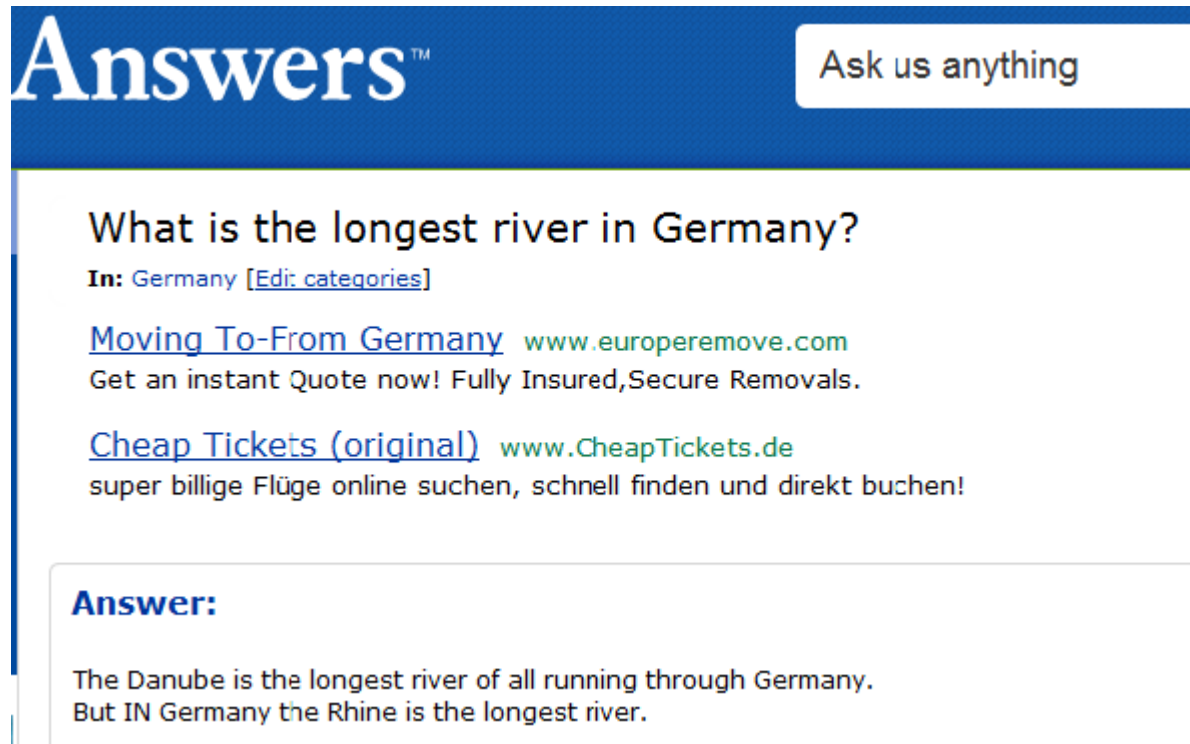
 Michele Martin likes Kirkland, Washington 

## ➤ Important questions

- How to rank user opinions (aspects to consider: user reliability, feedback quality, user history, similarity to user who issued the query...)?
- How to combine and present general search results with social results?
- How to derive explicit recommendations from implicit feedback?

## User feedback for social search (2)

### ➤ User-based question answering



The screenshot shows the 'Answers' website interface. At the top, there is a blue header with the 'Answers' logo and a search bar containing the text 'Ask us anything'. Below the header, a question is displayed: 'What is the longest river in Germany?'. Underneath the question, there is a link to 'Germany' and a link to 'Edit categories'. Two sponsored answers are shown: 'Moving To-From Germany' from www.europeremove.com and 'Cheap Tickets (original)' from www.CheapTickets.de. Below these, the 'Answer:' section is visible, containing the text: 'The Danube is the longest river of all running through Germany. But IN Germany the Rhine is the longest river.'

➤ Answers can be checked by site editors or ranked decreasingly by user ratings

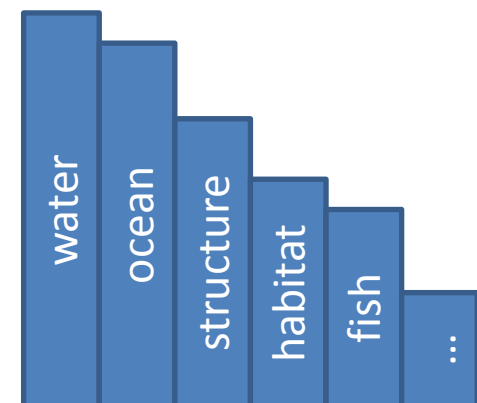
➤ Translation model for question Q and answer A: 
$$P(Q|A) = \prod_{w \in Q} \sum_{t \in \mathcal{V}} P(w|t)P(t|A)$$

# Tags and automatic tag suggestion

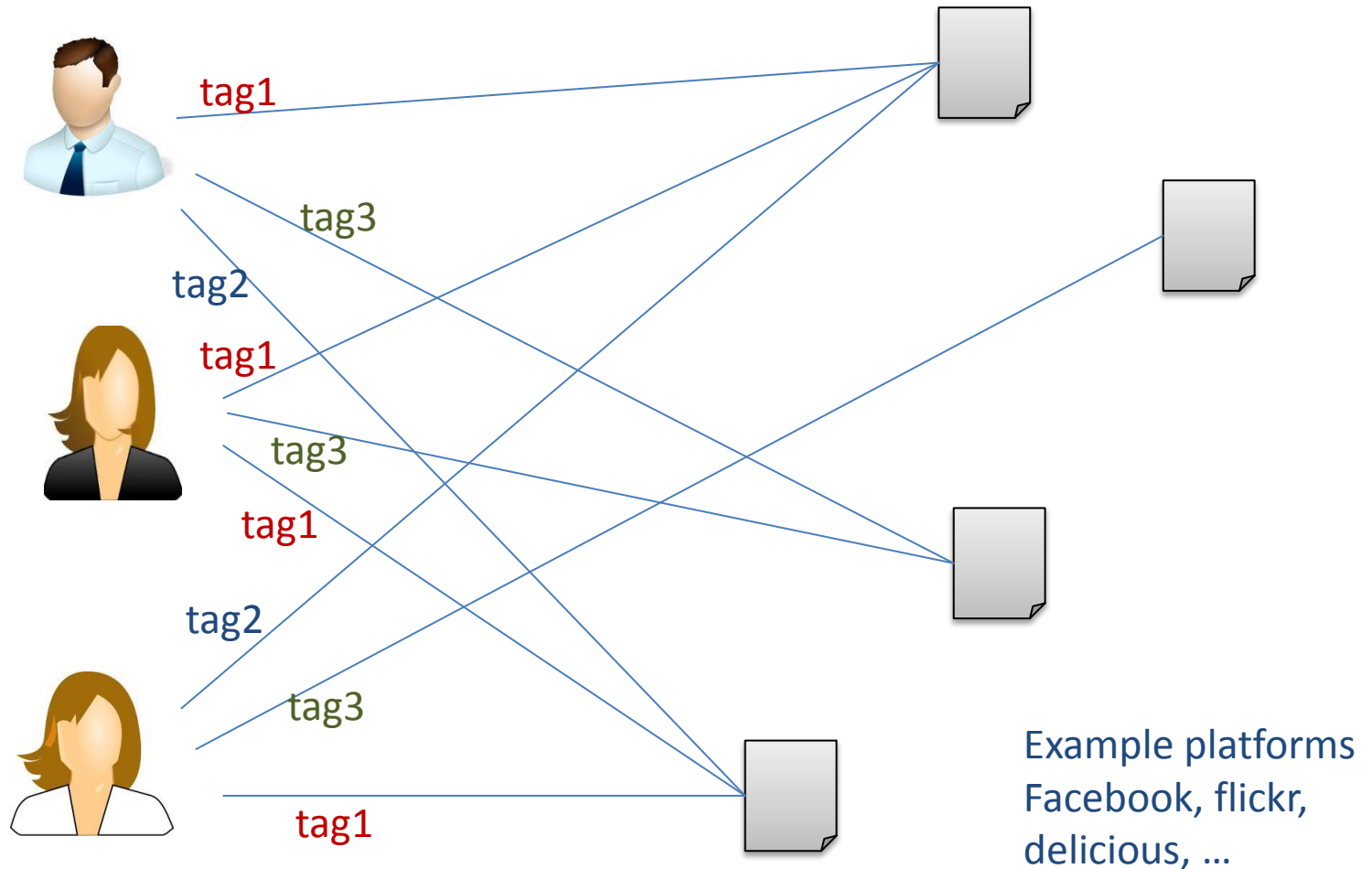
- Tags can be useful for search (as they reflect the “wisdom of the crowds”), but
  - many tags are incorrect, ambiguous, unique
  - many postings have very few or no tags at all (usually the popular ones have)
- Automatic tag suggestion (i.e., infer tags for new posting) by using
  - Tags with high tf-idf from postings of the same type (clustering or categorization needed)
  - Similarity between tag and posting and taking tag novelty, popularity, etc., into account
  - Per-tag relevance model (difficult for rare tags and difficult to scale)
  - Language models, thesauri or query logs

LM example: tag “coral reef”

$P(w | \text{tag})$

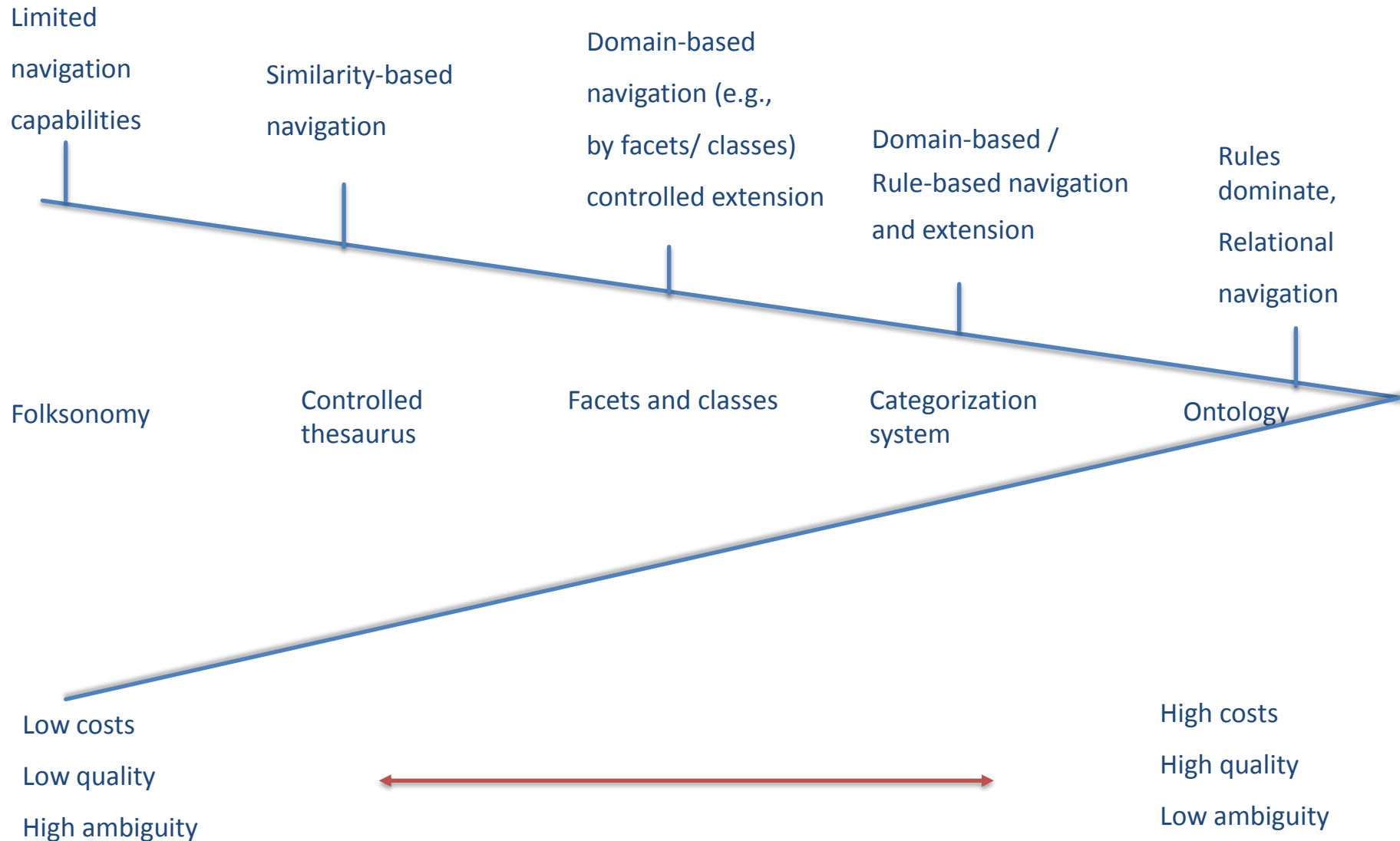


# Folksonomies: organizing items by tags





# Folksonomies



# Flickr folksonomy tag cloud

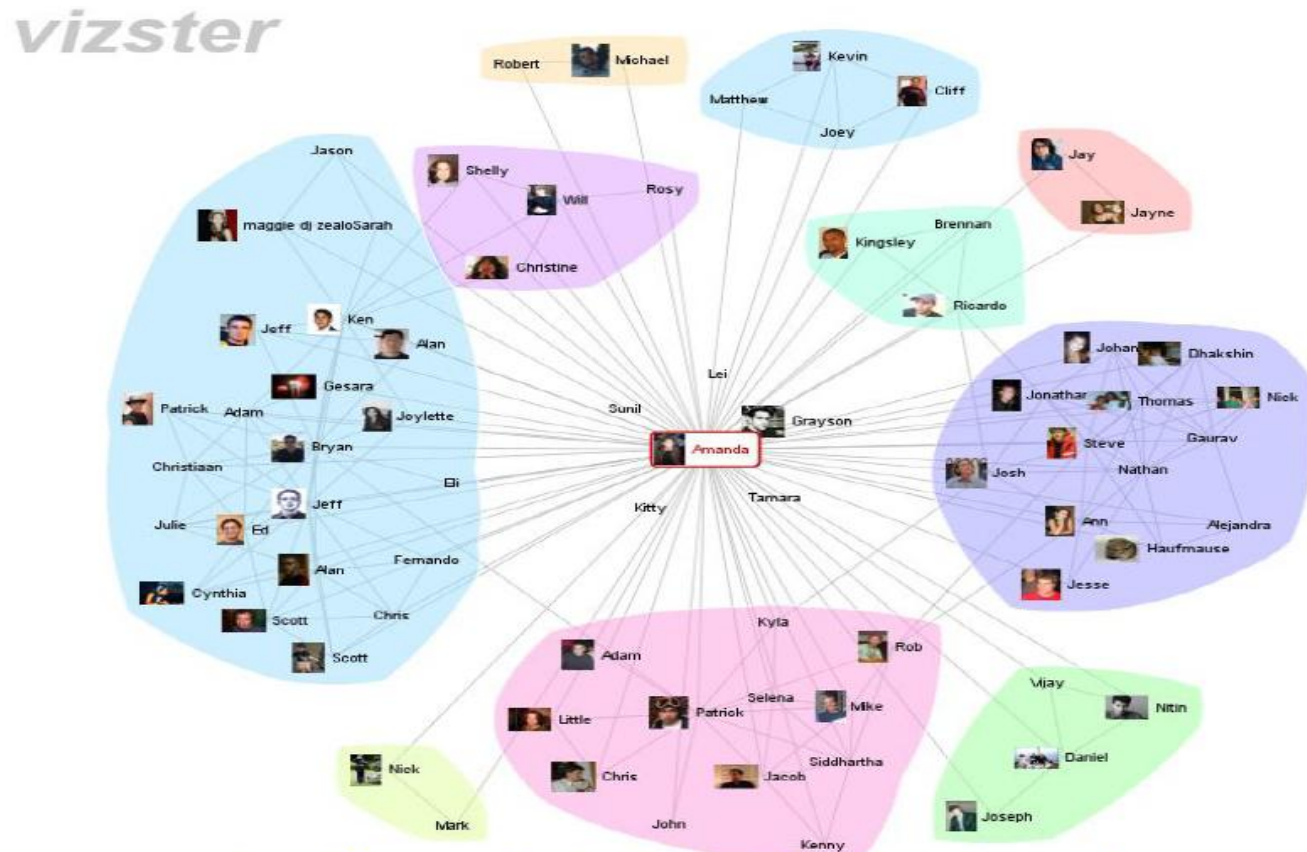
animals architecture **art** asia australia autumn baby band barcelona **beach** berlin bike bird  
birds birthday black blackandwhite blue bw **california** canada **canon** car cat  
chicago china christmas church **city** clouds color **concert** dance day de dog england  
europe fall **family** fashion **festival** film florida **flower** **flowers** food football  
**france** friends fun garden geotagged germany girl graffiti green halloween hawaii holiday  
house india **instagramapp** iphone **iphoneography** island italia italy  
**japan** kids la lake landscape light live london love macro me mexico model museum  
music nature new newyork newyorkcity night **nikon** nyc ocean old paris  
park party people photo photography photos **portrait** raw red river rock san  
sanfrancisco scotland sea seattle show sky snow spain spring **square**  
**squareformat** street **summer** sun sunset taiwan texas thailand tokyo  
travel tree trees trip uk unitedstates urban **usa** vacation vintage washington water  
**wedding** white winter woman yellow zoo

$$fontSize(t) = maxFontSize \cdot \frac{freq(t) - minFreq}{maxFreq - minFreq}$$

Source: <http://www.flickr.com/photos/tags/>

# User communities (1)

- Represented by graphs capturing interactions (through edges) between entities (i.e., nodes)
- Assumption: number of intra-community interactions is higher than that of inter-community interactions



## User communities (2)

### ➤ Detecting communities

➤ HITS algorithm assigns higher scores to nodes belonging to many v-structures and can be extended to detect communities

### ➤ Recursive graph cut

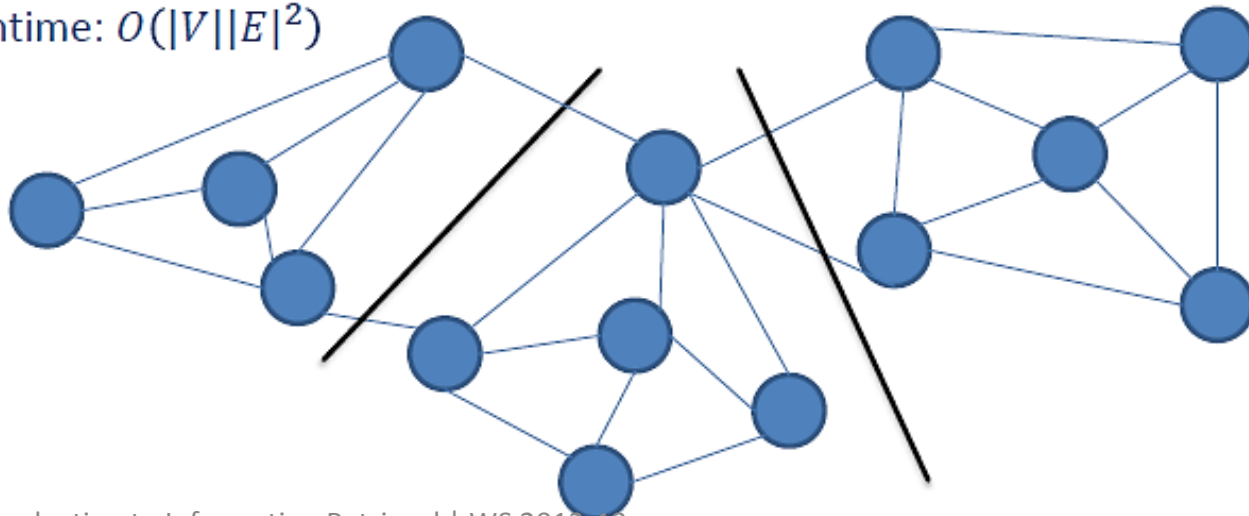
Construct similarity graph  $G(V,E)$  of  $n$  data points

Remove edges below a similarity threshold  $t$

Find set of edges  $C \subset E$  with  $w(C) = \sum_{e \in C} w(e)$  is minimum and  $G(V,E \setminus C)$  has two connected components  $G_1, G_2$  // min-cut

Run algorithm recursively on  $G_1, G_2$  until  $k$  clusters are built

Runtime:  $O(|V||E|^2)$

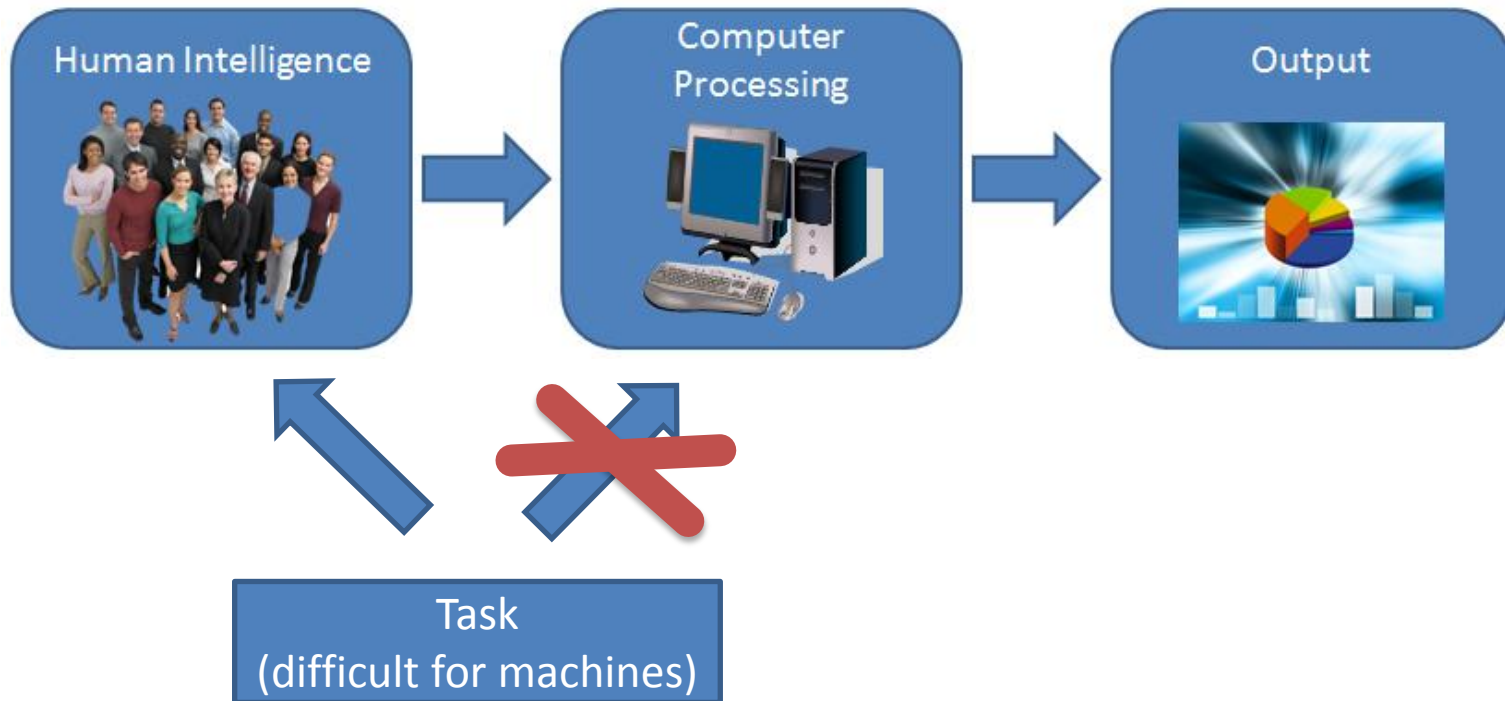


## User communities (3)

---

- More algorithms for detecting communities
  - Clustering algorithms (e.g., k-means, spectral clustering, co-clustering, latent semantic analysis, locality-sensitive hashing, and many more) to be discussed later
- Use cases
  - Detecting user interests
    - personalization
    - recommendation
  - Expert finding (combined with other authority-detection algorithms e.g., PageRank)
  - Hyperbolic browsing of “summarized interaction graph”

## Crowdsourcing: general idea

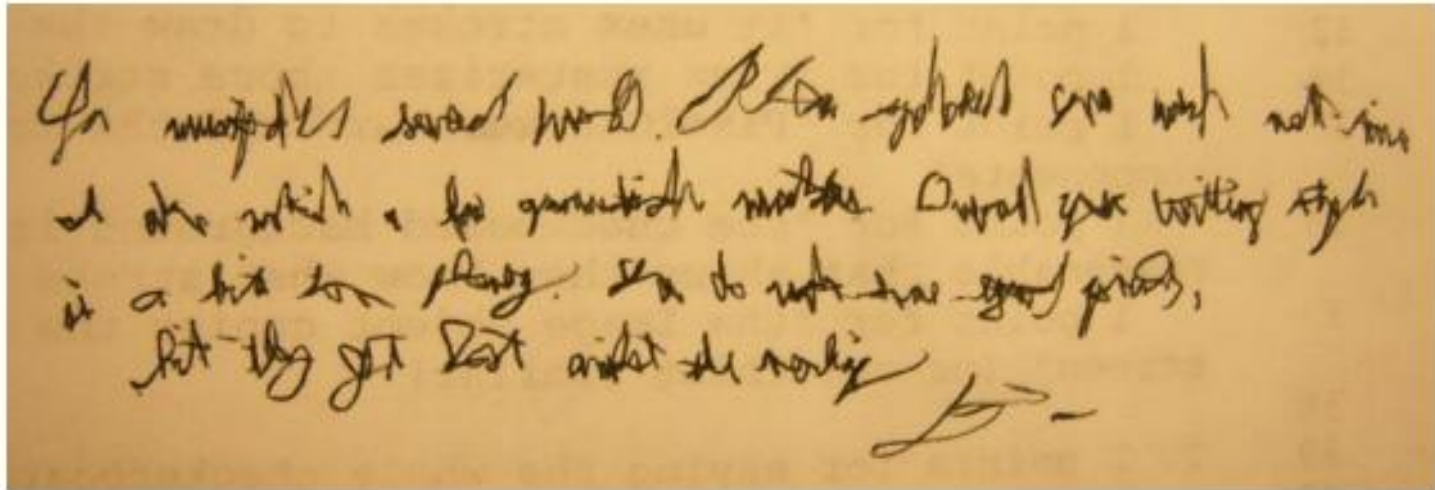


### ➤ Examples

- Relevance feedback for search engine results
- Labelling photos, annotating text, recognizing handwriting, ...
- Community-based question answering
- Proving theorems
- ...

## “Wisdom of the crowd” example in crowdsourcing

### ➤ Sample of handwritten text



### ➤ After multiple iterations on Amazon Mechanical Turk:

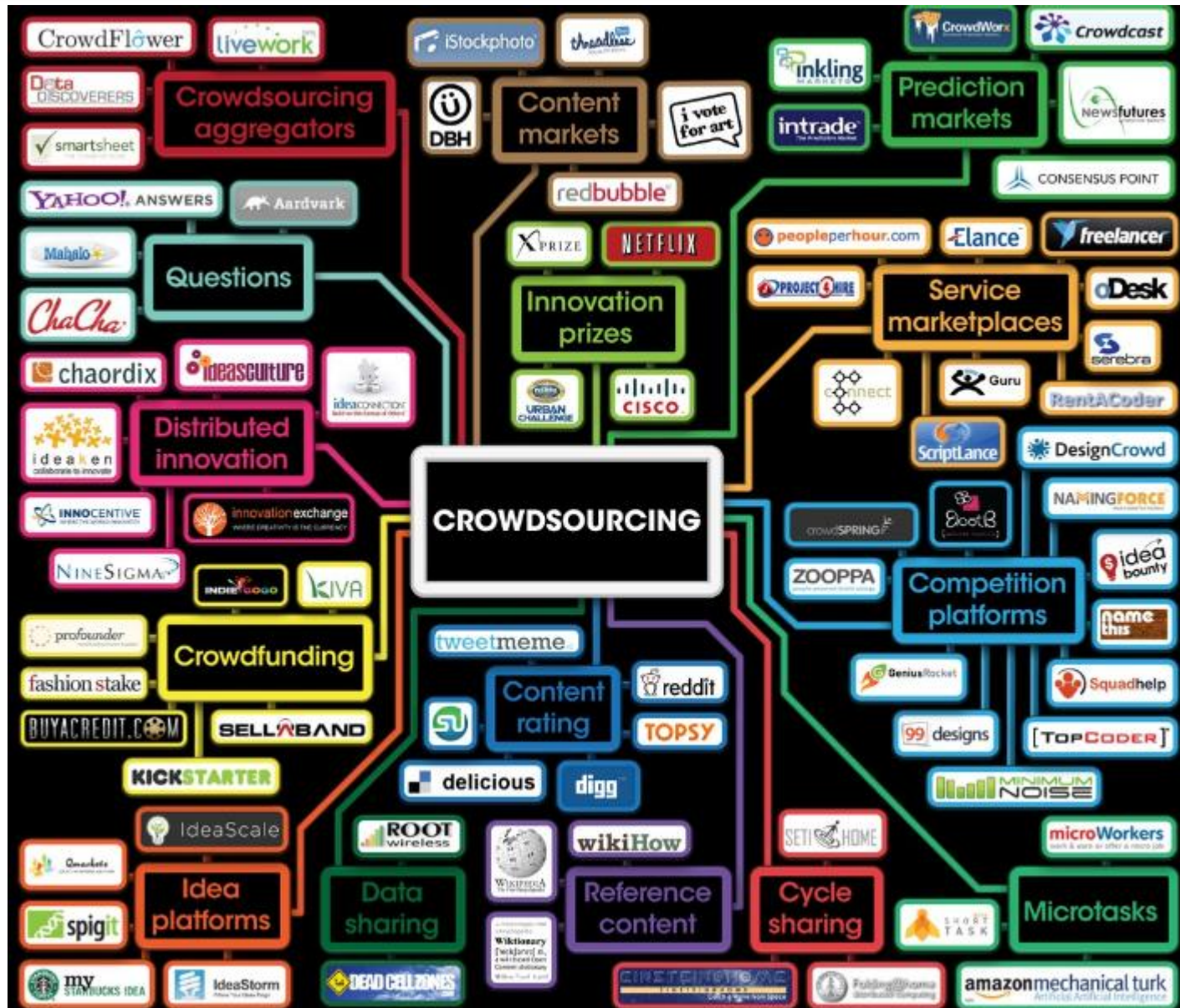
You misspelled several words. Please spellcheck your work next time. I also notice a few grammatical mistakes. Overall your writing style is a bit too phoney. You do make some good points, but they got lost amidst the writing. (signature)

Source: [G. Little et al., HCOMP 2009](#)



# Crowdsourcing landscape

Source:  
[vi.sualize.us](http://vi.sualize.us)





# Amazon Mechanical Turk (AMT)

## Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce. Workers select from thousands of tasks and work whenever it's convenient.

**248,326 HITs** available. [View them now.](#)

## 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. [Register Now](#)

As a Mechanical Turk Requester you:

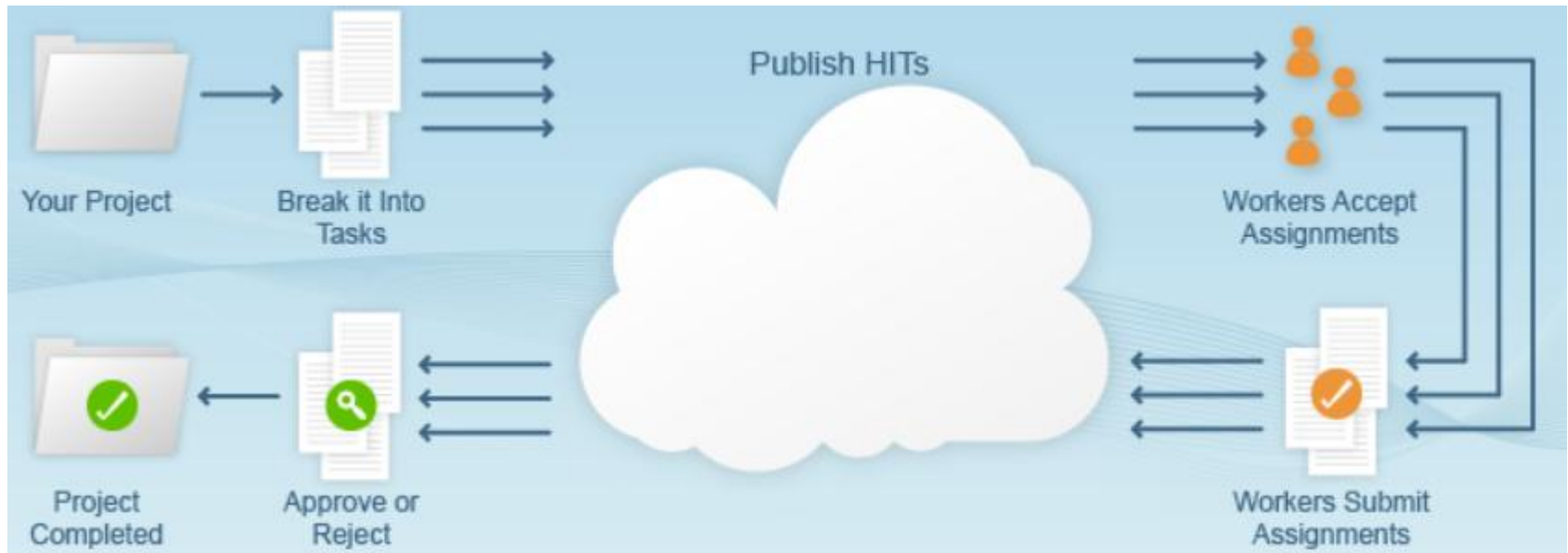
- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results



Source:

<https://www.mturk.com/mturk/welcome>

# How does AMT work?



Confidence score for worker derived from aggregation of approvals and rejections

Source: <http://www.crowdsourcingblog.de>

# HIT example: solving IQ task

Requester: Gjergji Kasneci

Reward: \$0.1 per HIT

HITs available: 0

Duration: 5 Minutes

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 98 , Number of HITs Approved greater than or equal to 100

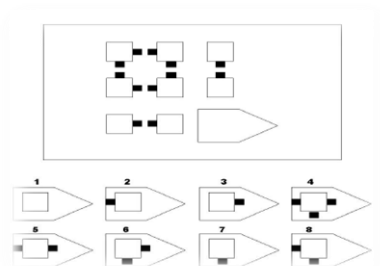
## HIT Preview

**Choose a correct answer to this computer generated reasoning problem.**

Guidelines:

- There is only one correct response
- If you don't know the answer - return the HIT.

**{Question}**



{r1}  {r2}  {r3}  {r4}  {r5}  {r6}  {r7}  {r8}

Please provide any comments you may have below, we appreciate your input!

# Measuring the crowd IQ

---

- The term “Wisdom of the Crowds” is too vague; can it be quantified?
- Can we measure the IQ of a crowd (viewed as a black-box system)?
- Approach
  - Let each crowd individual solve standardized IQ test
  - Aggregate the answers to each IQ question with the goal to boost (or maximize) the IQ of the crowd as a black-box system

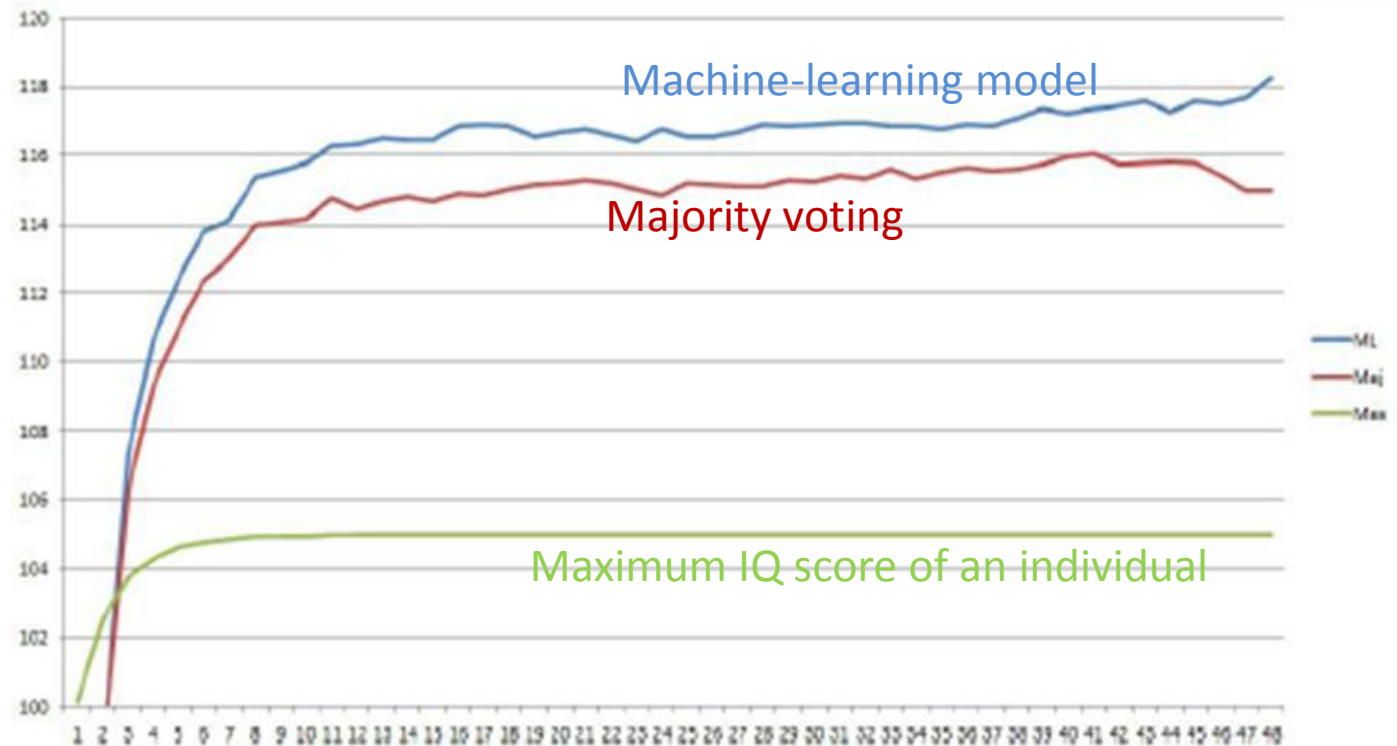
## Majority Voting on the Answer

Question \ Participant	1	2	3	4	5
Alice	A	D	B	D	C
Bob	A	D	B	B	D
Charlie	A	D	D	A	A
David	B	A	C	C	B
Eugene	A	A	C	B	D
Fiona	A	D	C	B	B
George	A	D	C	C	B
Majority	A	D	C	B	B
Correct	A	D	C	B	D

- Is it possible to outperform majority voting?
  - Learn reliability scores for each participant and compute weighted average

## Boosting the IQ in a “IQ-homogeneous” Crowd?

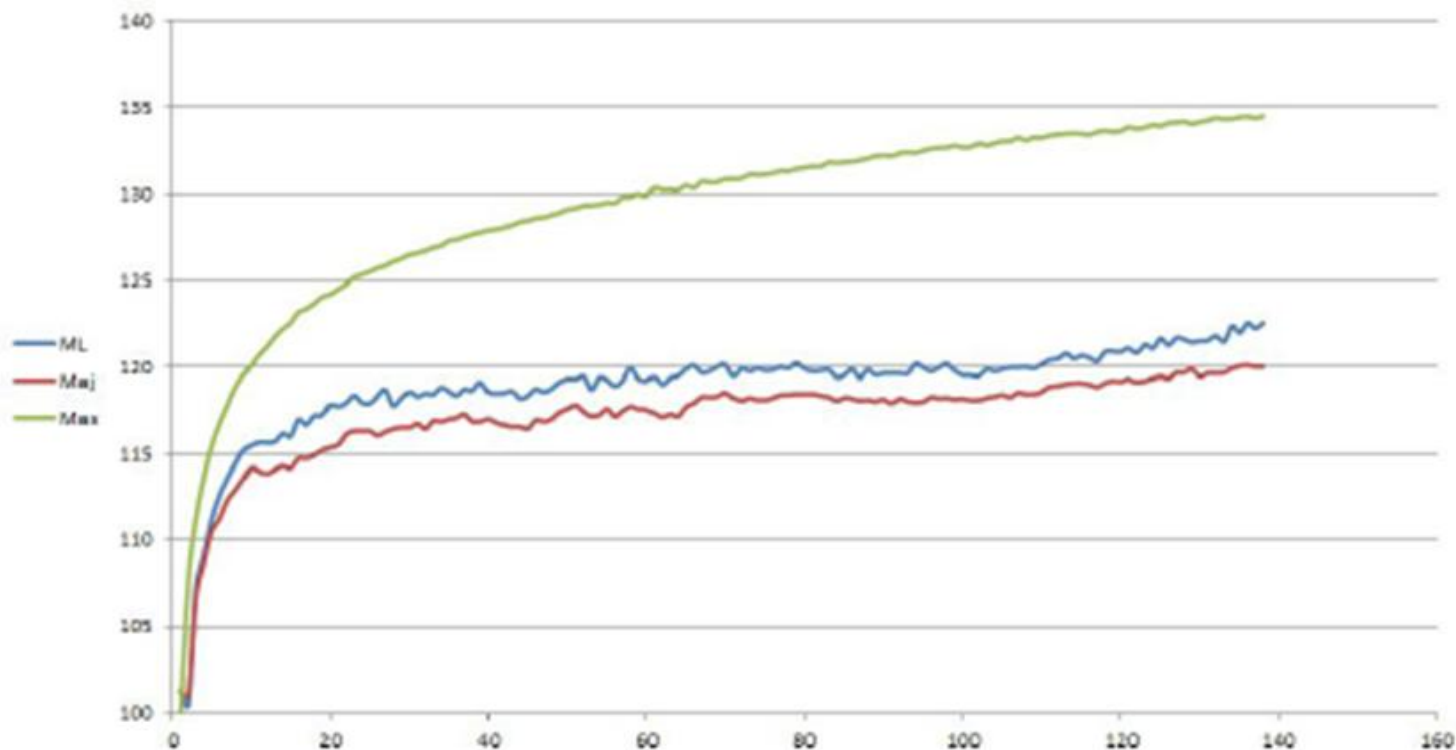
- IQ-homogeneous: small variance in individual IQ scores



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

Source: [Y. Bachrach et al., AAMAS 2012](#)

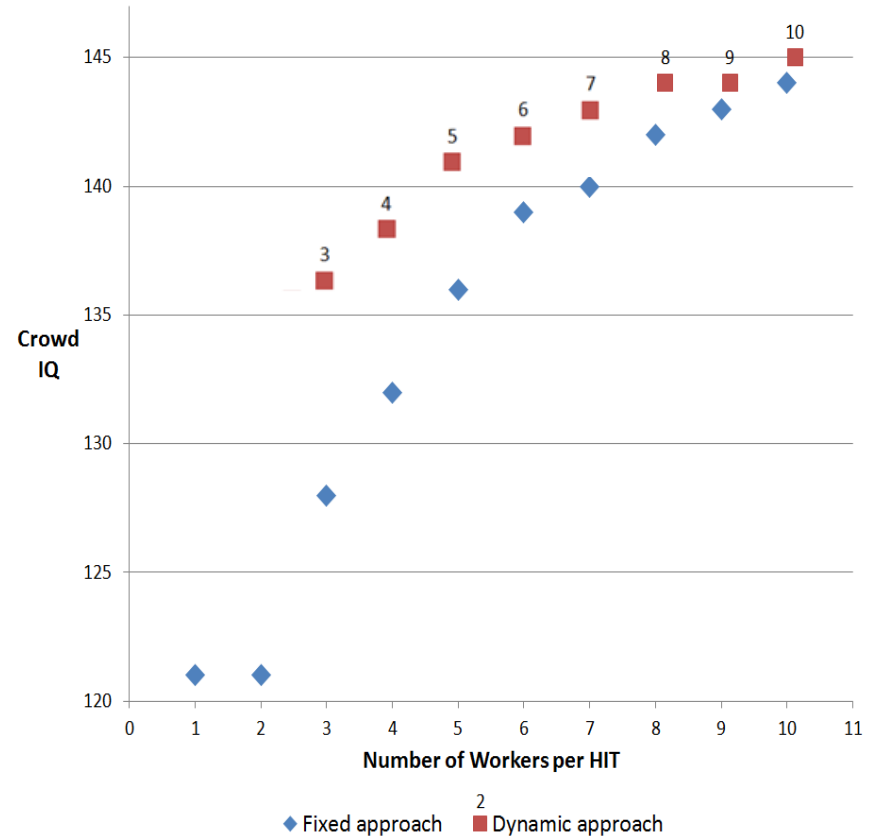
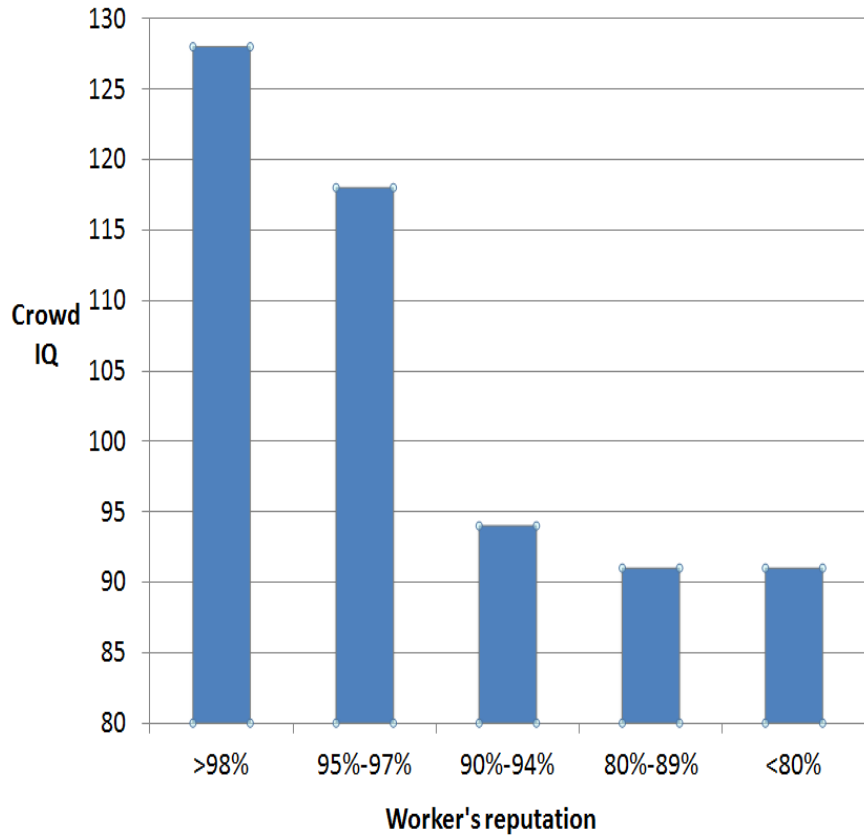
# How smart is the Crowd?



Crowd IQ and maximal IQ for the entire dataset

Source: [Y. Bachrach et al., AAMAS 2012](#)

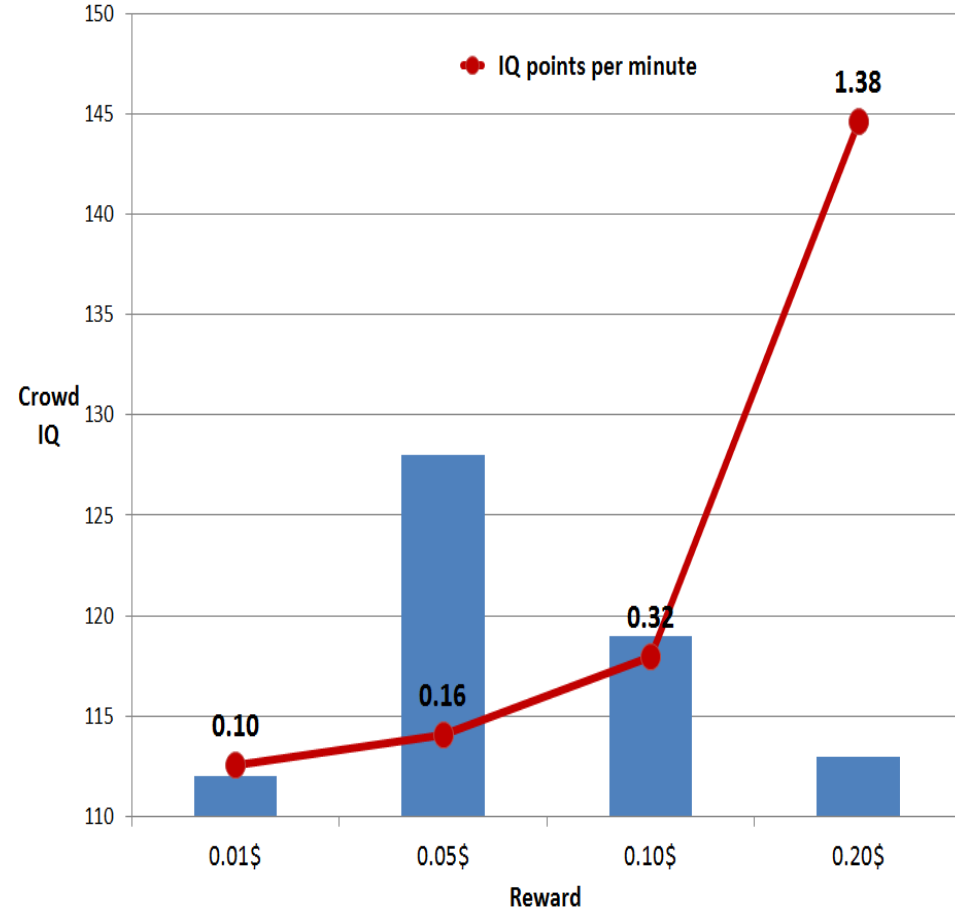
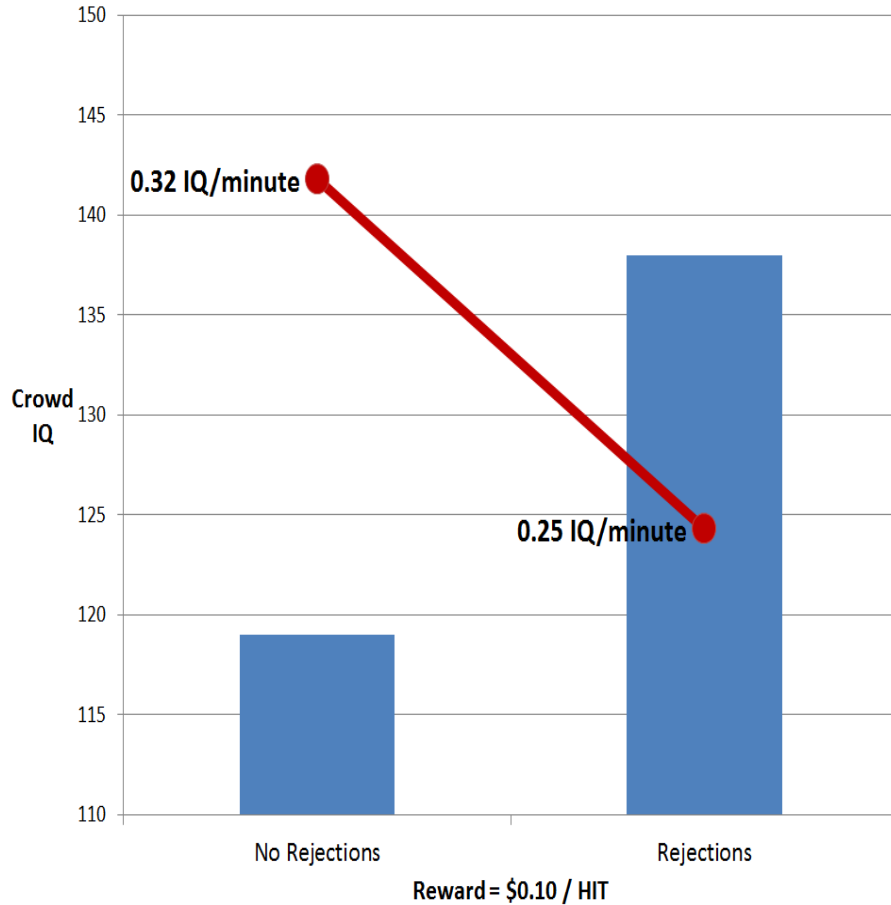
# Reputation and crowd IQ on Amazon Mechanical Turk



Source: [M. Kosinski et al., Web Science 2012](#)



# How do incentives change the crowd IQ on AMT?



Source: [M. Kosinski et al., Web Science 2012](#)

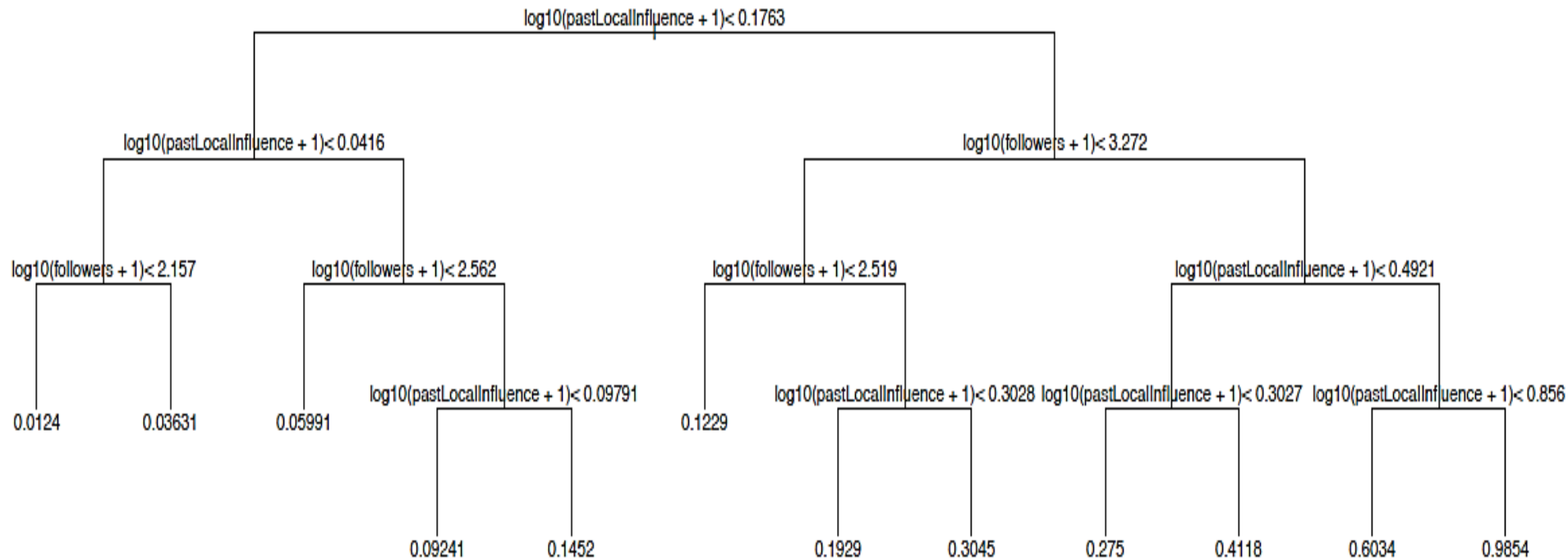
# Influential users and information cascades

---

- Influence of a user in social network
  - Number of followers in the network (can be quantified by authority-based link analysis algorithms, e.g., PageRank)
  - Influence of the user's postings (e.g., how many other users read and forward the postings?)
  - Capability of forming public opinions (difficult to measure)
  - In general: the higher the influence of a user the larger the information cascades incurred by the user

# Influential users and information cascades on Twitter

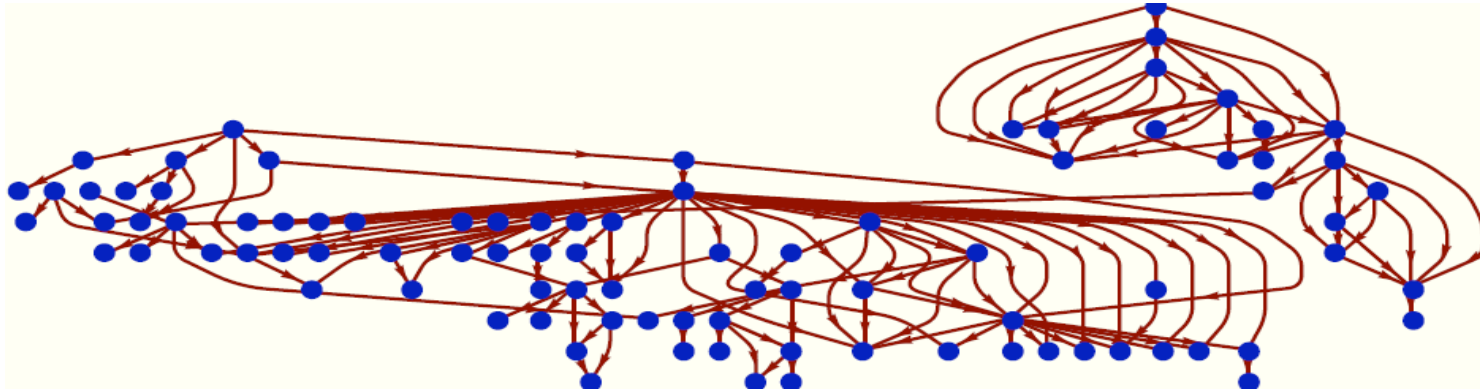
- Past local influence: average number of reposts by that user's immediate followers



Regression tree: leaf nodes give the predicted influence for the corresponding partition

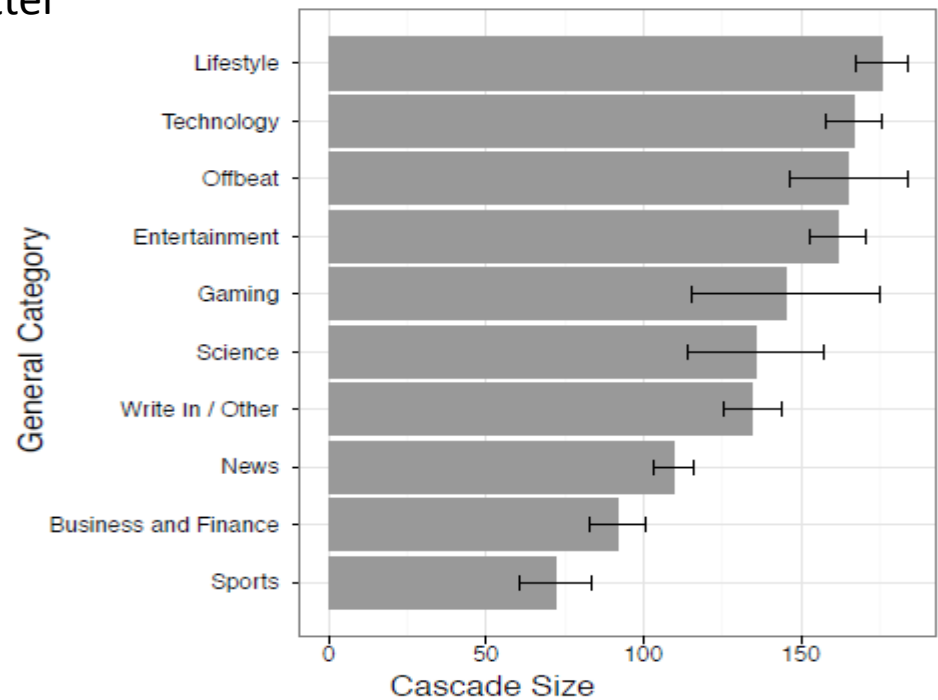
Source: [E. Bakshy et al., WSDM 2011](#)

# Information cascades on Twitter



Example of a cascade on Twitter

Source: [E. Bakshy et al., WSDM 2011](#)



# Predicting viral tweets (1)

---

- Viral tweet: tweet that spreads quickly and widely over the Twitter network
- Is it possible to predict such tweets?
  - View tweet  $\mathbf{t}$  as a vector of feature values  $\mathbf{t} = (x_1, \dots, x_n)$
  - Example
    - $x_1$ : # followers of the user who posted the tweet
    - $x_2$ : # number of URLs in tweet
    - $x_3$ : # hashtags used in tweet
    - $x_4$ : mentions in a tweet
    - $x_5$ : length of tweet
    - $x_6$ : sentiment of a tweet (i.e., positive, negative or neutral)
    - $x_7$ : tweet topic
    - ...
  - Use training set of viral and non-viral tweets to train classification model

## Predicting viral tweets (2)

- Naïve Bayes model (assumes conditional independence between features)

$$P(R(\mathbf{t}) > \tau, \mathbf{t}) = P(R(\mathbf{t}) > \tau) \cdot P(\mathbf{t} | R(\mathbf{t}) > \tau)$$

$$= P(R(\mathbf{t}) > \tau) \cdot \prod_{i=1}^n P(x_i | R(\mathbf{t}) > \tau)$$

# retweets for  $\mathbf{t}$

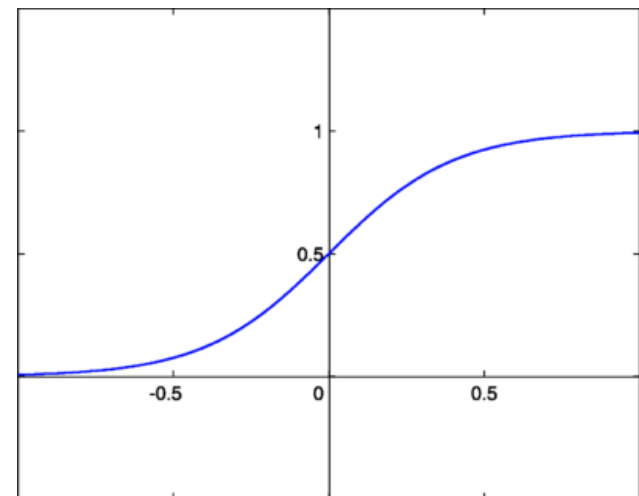
threshold

- Generalized linear model (correlations between features captured by linear combination)

$$P(R(\mathbf{t}) > \tau | \mathbf{t}) = f(w_0 + w_1 x_1 + \dots + w_n x_n)$$

unknown weights  
(can be learned from  
training corpus)

Sigmoid  
activation  
function



## Predicting viral tweets (3)

- Evaluated on dataset extracted through public Twitter API based on user IDs from the TREC 2011 microblog corpus

$T$	F-Measure NB	F-Measure GLM
50	0.916	0.936
100	0.927	0.940
500	0.947	0.963
1000	0.951	0.968

Source: [Master's Thesis](#) by M. Jenders, HPI, 2012

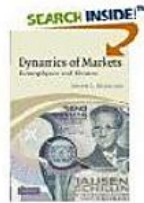
- Open questions
  - How to identify and integrate current popularity of a tweet's topic?
  - Is it possible to predict the approximate number of reposts (i.e., retweets) for a tweet?
  - ...

# Recommendation

## Amazon recommendations

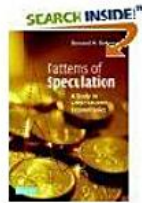
### More to Explore

You looked at

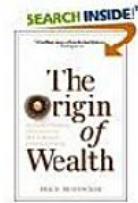


Dynamics of Markets: Econophysics and...  
Hardcover by Joseph L. McCauley  
~~\$77.92~~

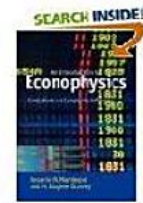
You might also consider



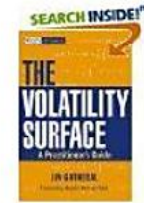
Patterns of Speculation: A Study in...  
Paperback by Bertrand M. Roehner  
~~\$39.99~~ **\$35.99**



Origin of Wealth: Evolution... Paperback  
by Eric D. Beinhocker  
~~\$46.00~~ **\$10.88**



Introduction to Econophysics...  
Paperback by Rosario N. Mantegna, H...  
**\$32.99**



The Volatility Surface: A Practitioner's...  
Hardcover by Jim Gatheral, Nassim...  
~~\$60.00~~ **\$37.80**

## Movie ratings

(rating sparsity encountered by taking feature correlations into account)

	?	👍	👍	?
	👍	?	👍	?



# Collaborative filtering

- Typically accomplished by taking ratings by other users into account

$$R_u(item) = \frac{\sum_{u' \in N(u)} sim(u, u') \cdot R_{u'}(item)}{\sum_{u' \in N(u)} sim(u, u')}$$

Users most similar to  $u$

Feature-based similarity function

- User-user similarity based on user features and user ratings
- Item-item similarity based on item features and users who rated the items
- Similarity measures
  - Cosine-similarity
  - Pearson correlation
  - Jaccard similarity
- To be continued ... (in next lecture)

- Social search overview
  - User feedback
  - Folksonomies
  - User communities
  - Crowdsourcing
  - Information cascades & influential users
  - Collaborative filtering