Web Science: Collective Intelligence & Recommender Systems

(Part 1 - Collective Intelligence)

CS 432/532

Old Dominion University

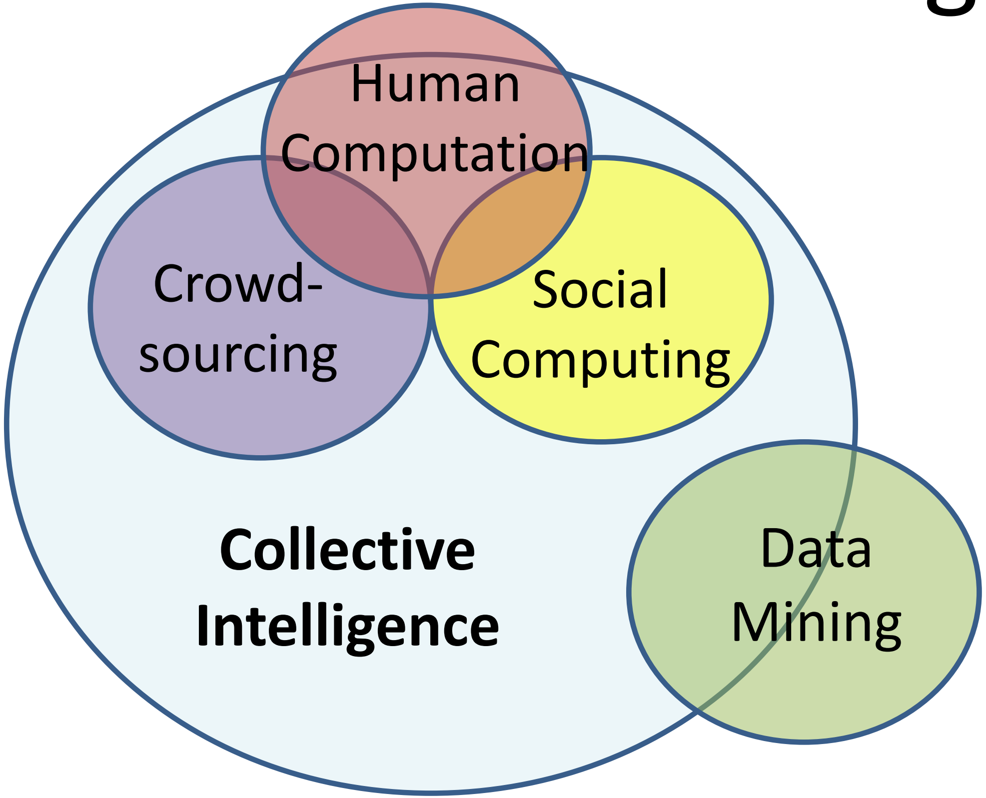
Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle

Main reference:

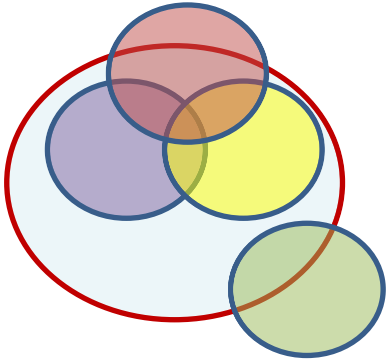
Ch 1 from *Programming Collective Intelligence* by Toby Segaran

(abbreviated as PCI)

What is Collective Intelligence?



[Quinn and Bederson](), "Human computation: A survey and taxonomy of a growing field", CHI 2011



Collective Intelligence

"groups of individuals doing things collectively that seem intelligent."

[Thomas Malone](), Director of MIT Center for Collective Intelligence, 2006

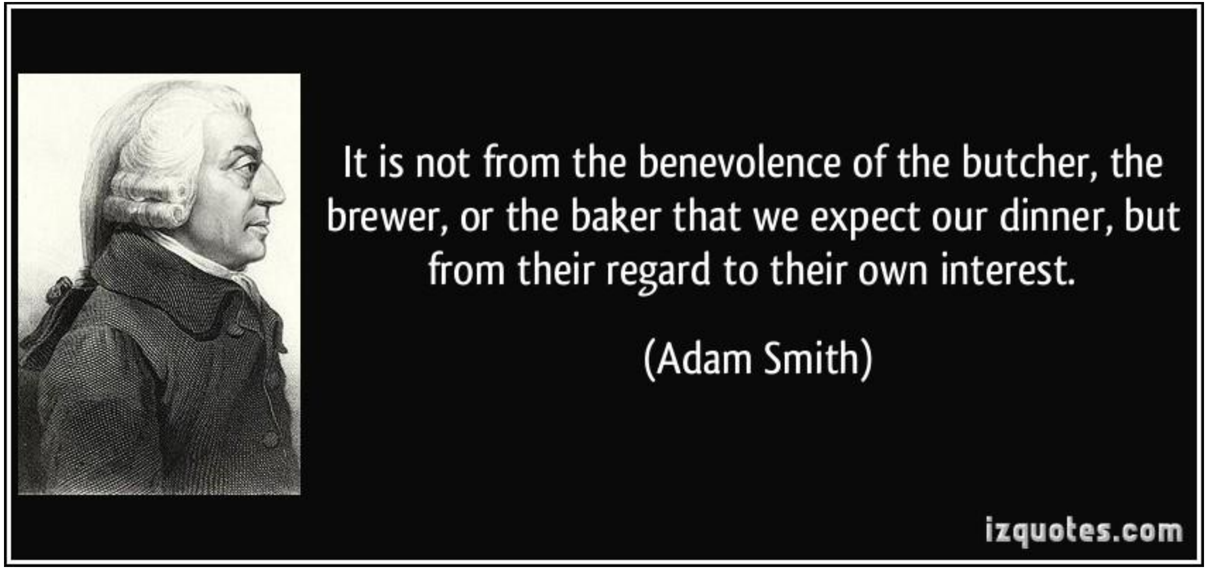
"when technologists use this phrase they usually mean the combining of behavior, preferences, or ideas of a group of people to create novel insights."

Segaran, *Programming Collective Intelligence (PCI)*, p. 2

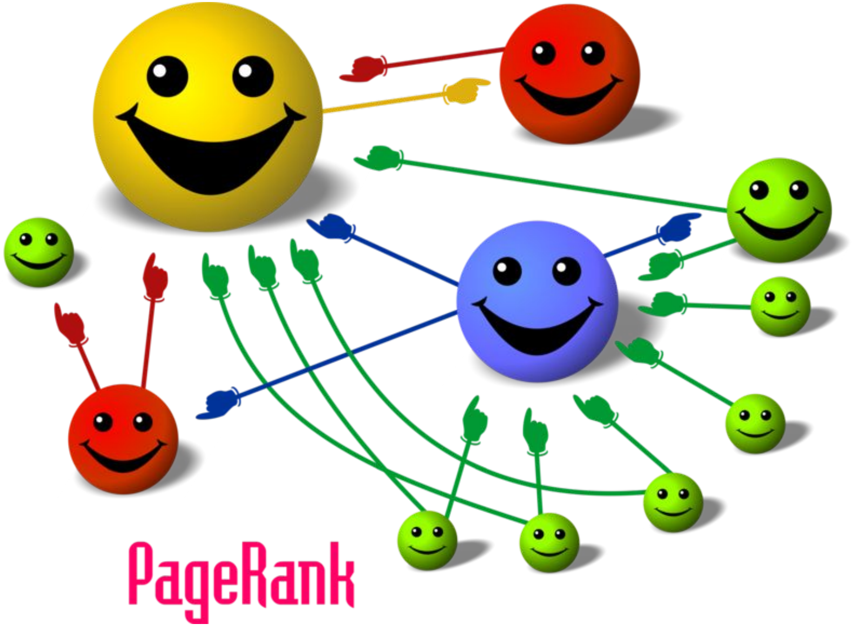
img source: https://www.shutterstock.com/image-photo/guess-how-many-jelly-beans-mason-121109632

Crowd-sourced strategies: How to Count Jelly Beans in a Jar, How to win a guess the number of jelly beans in a jar contest

Collective Intelligence in Classical Economic Theory



[Invisible hand](https://en.wikipedia.org/wiki/Invisible_hand) (Wikipedia)
The Wealth of Nations (Wikipedia)



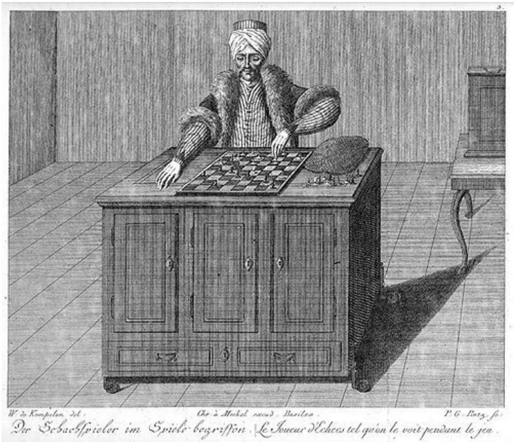
Img source: http://organicseoexpert.org/wp-content/uploads/2012/03/pagerank.png

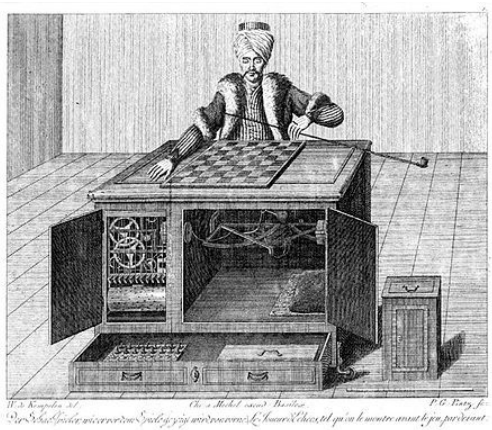
Crowdsourcing

"the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call."

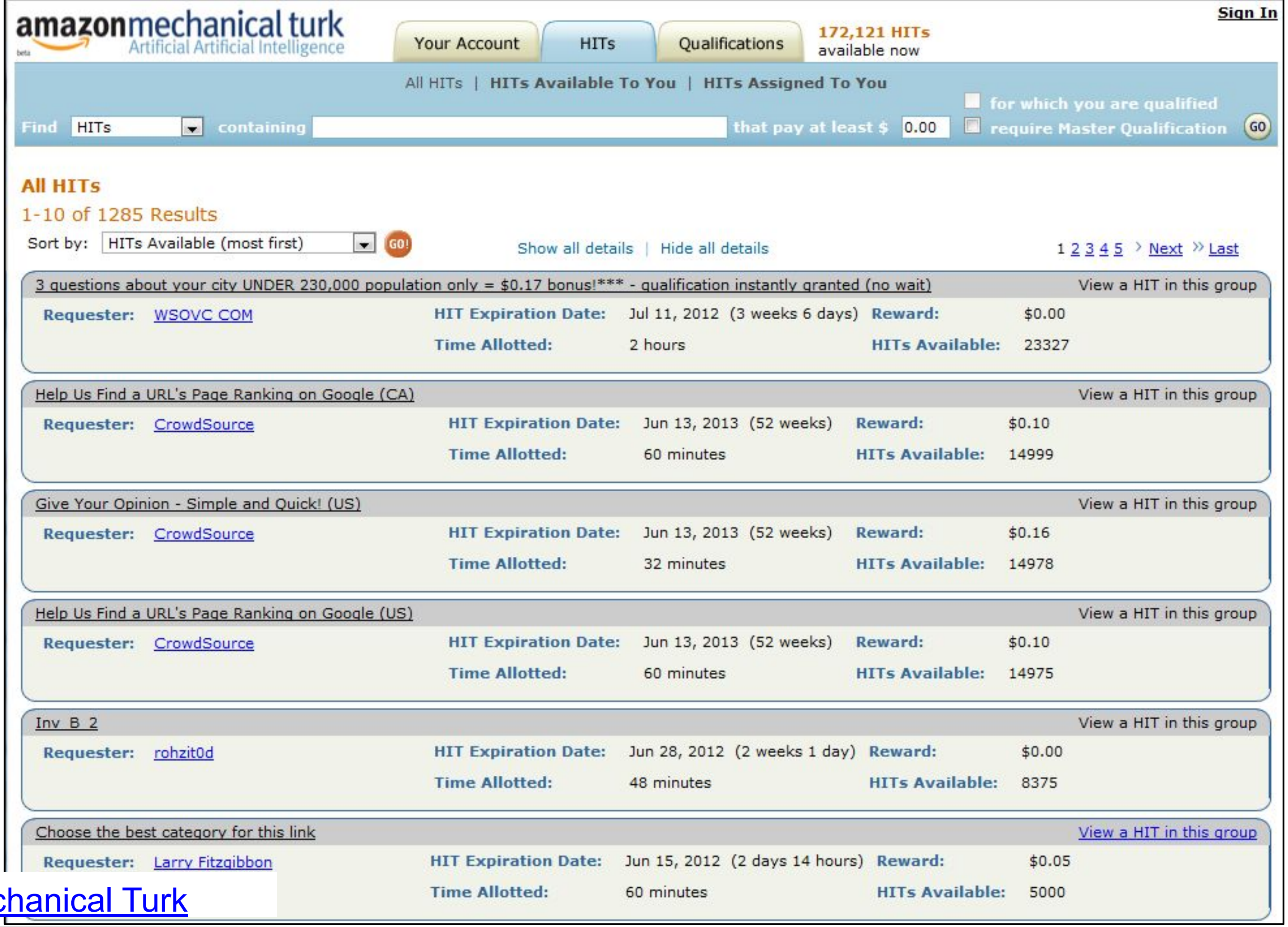
Jeff Howe, ["The Rise of Crowdsourcing"](), *Wired*, Jun 2006

Mechanical Turk



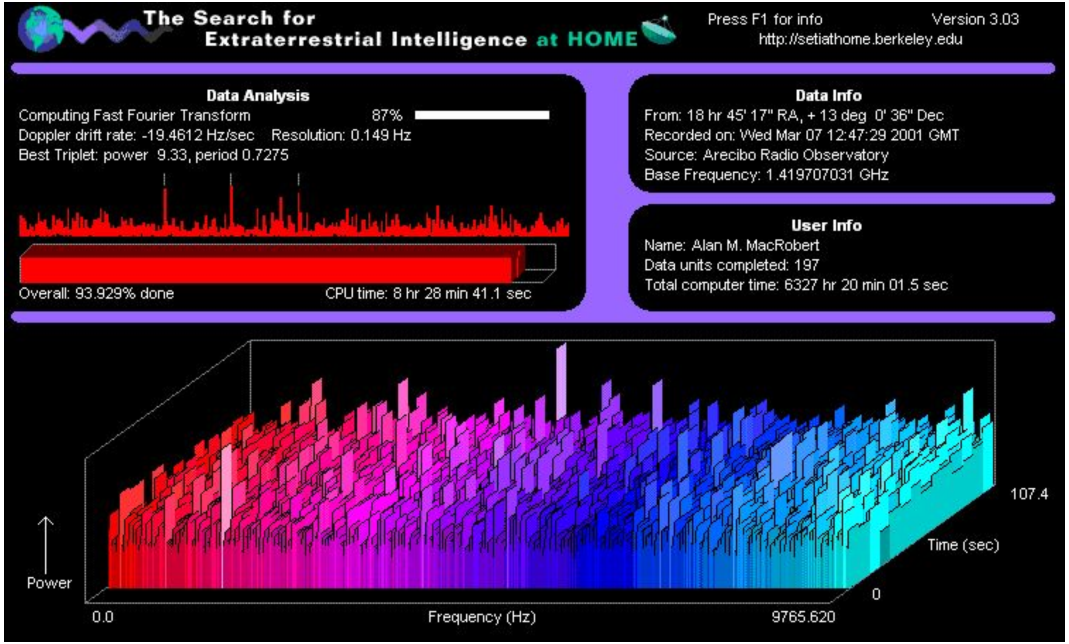


[The Turk](#) (Wikipedia)



Amazon Mechanical Turk

SETI@home Harvests Cycles From Idle Computers



SETI@home

Human Computation

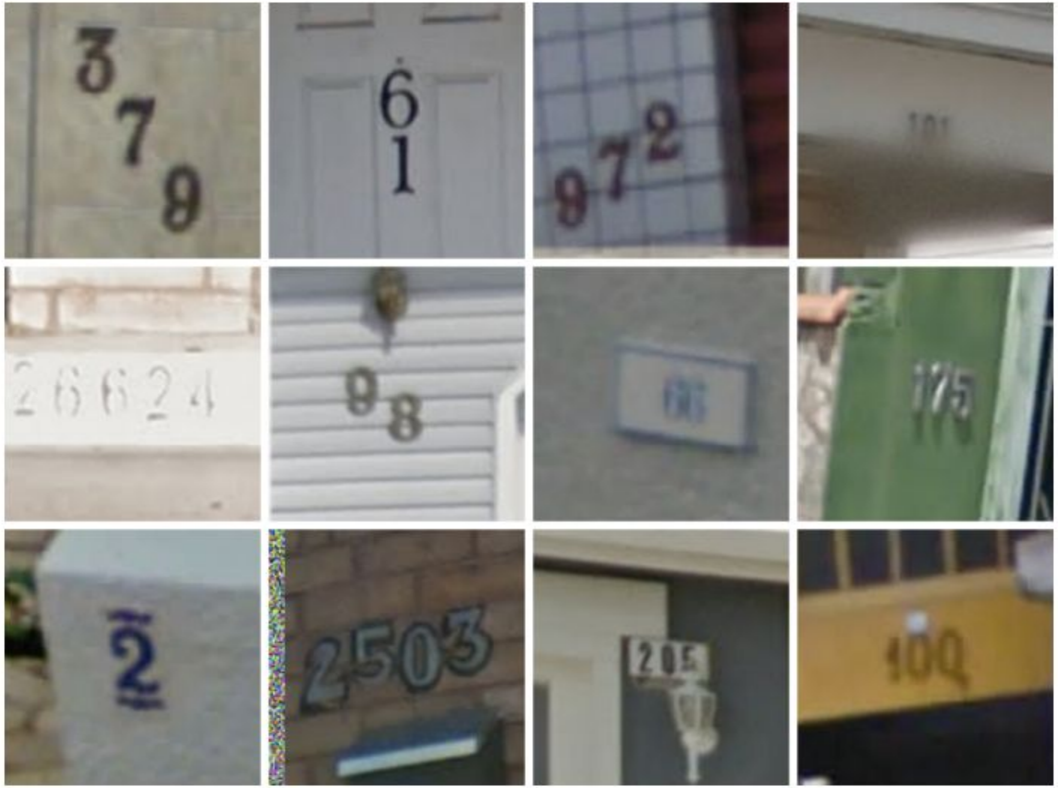
"a paradigm for utilizing human processing power to solve problems that computers cannot yet solve."

Luis van Ahn, Doctoral Dissertation at Carnegie Mellon, 2005

"Human computation:
The problems fit the general paradigm of computation, and as such might *someday* be solvable by computers.
The human participation is directed by the computational system or process."

Quinn and Bedderson, CHI 2011

[reCAPTCHA v3](), [Google's New Street View Image Recognition Algorithm Can Beat Most CAPTCHAs]()
Goodfellow et al., ["Multi-digit Number Recognition from Street View Imagery using Deep Convolutional Neural Networks"](), 2014



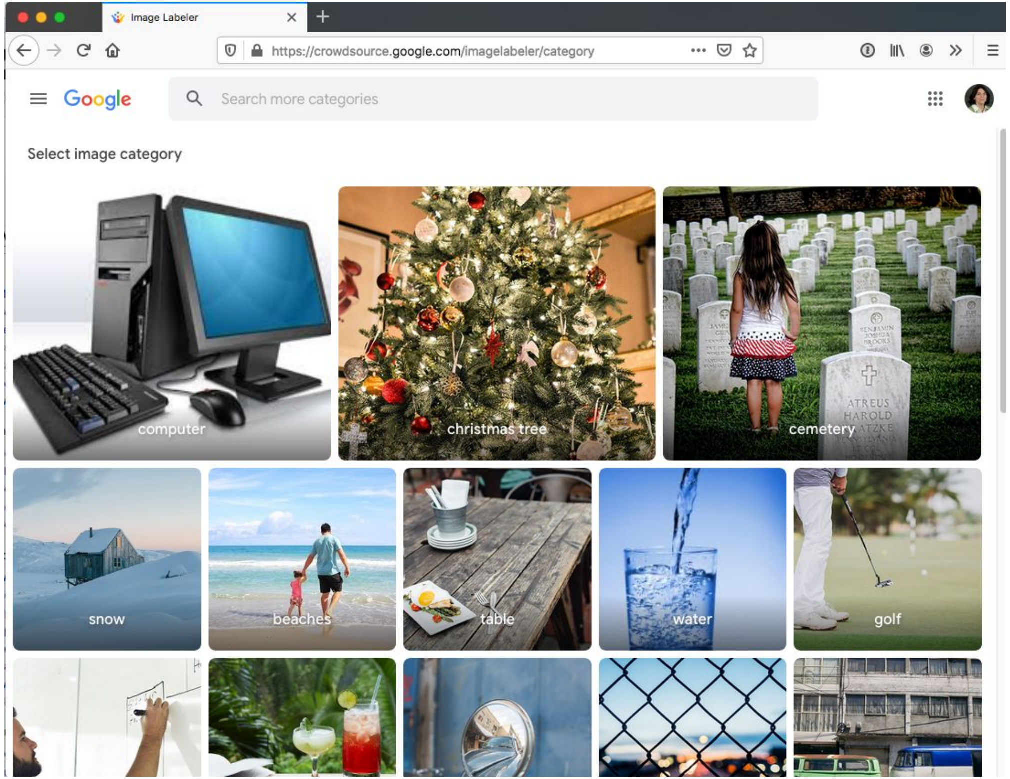
[How Google Cracked House Number Identification in Street View]()

[Tasks](#) (XKCD)

In the 60s, Marvin Minsky assigned a couple of undergrads to spend the summer programming a computer to use a camera to identify objects in a scene. He figured they'd have the problem solved by the end of the summer. Half a century later, we're still working on it.

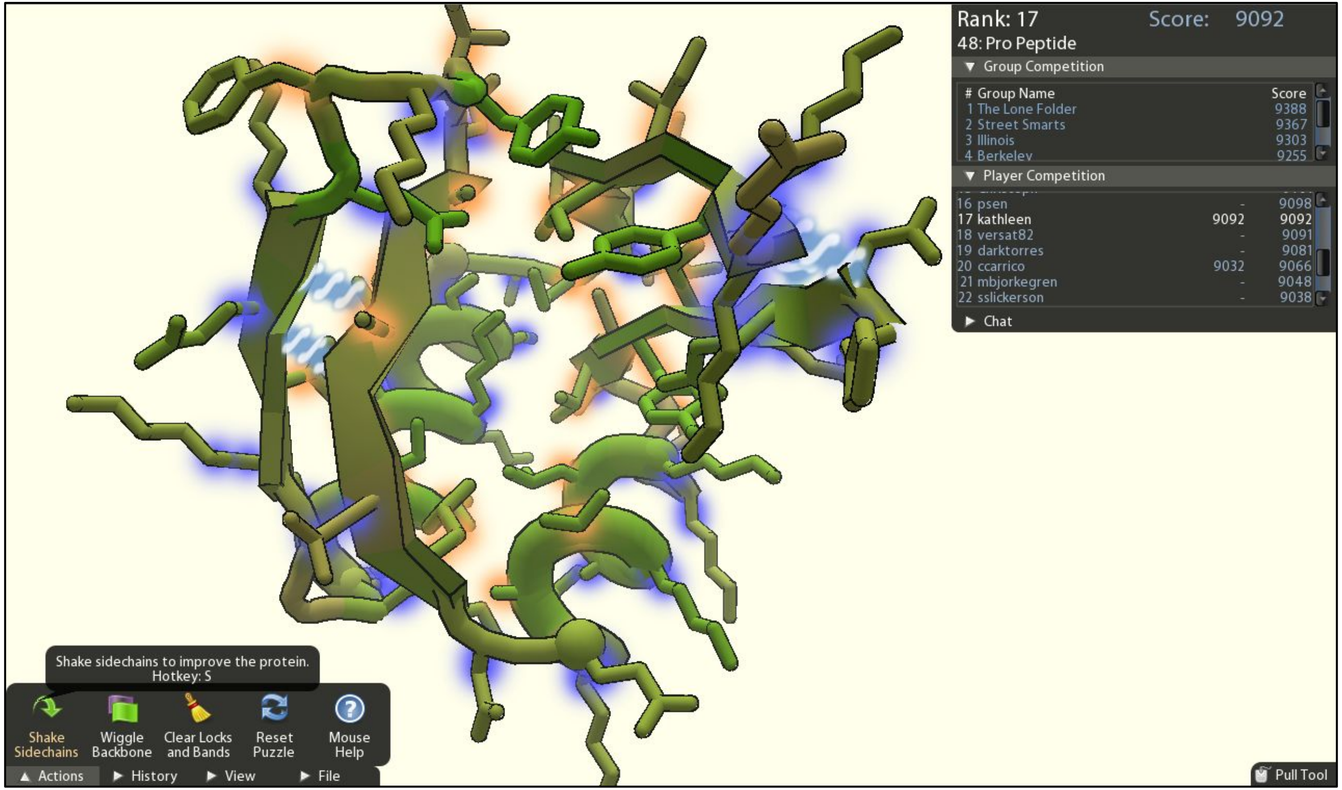
[ESP game](https://en.wikipedia.org/wiki/ESP_game) (Wikipedia)

Google licensed ESP for ImageLabeler



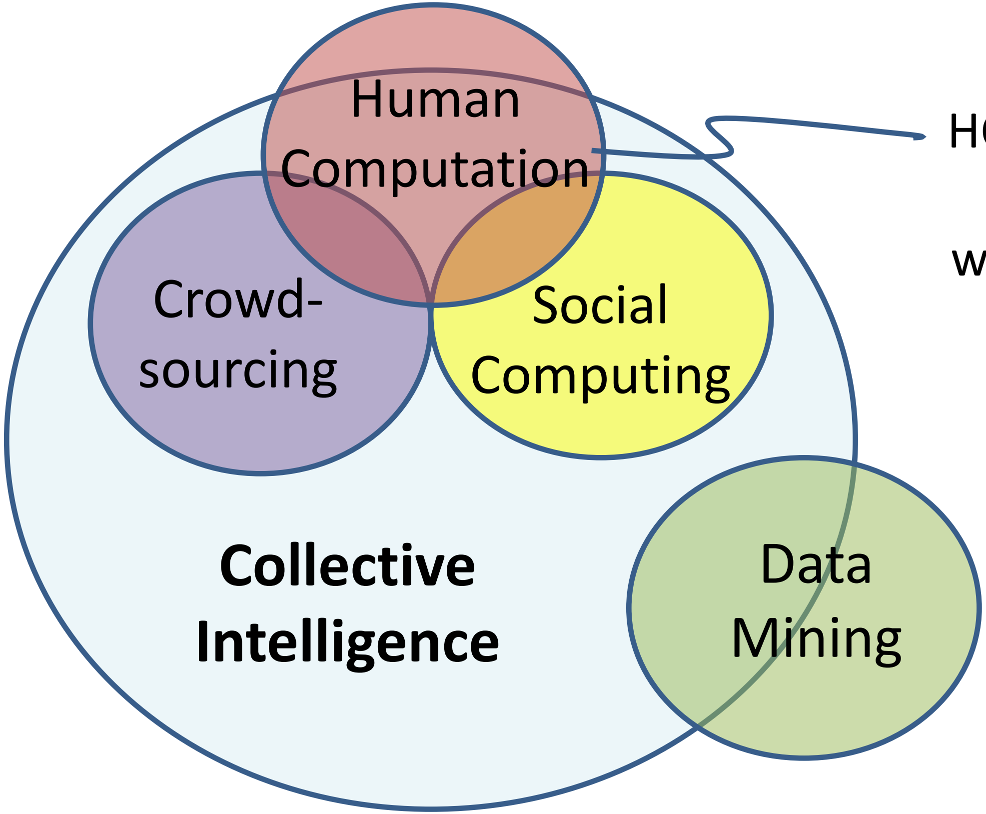
ImageLabeler (Google)

FoldIt



[Foldit: Solve Puzzles for Science]

CI not superset of HC

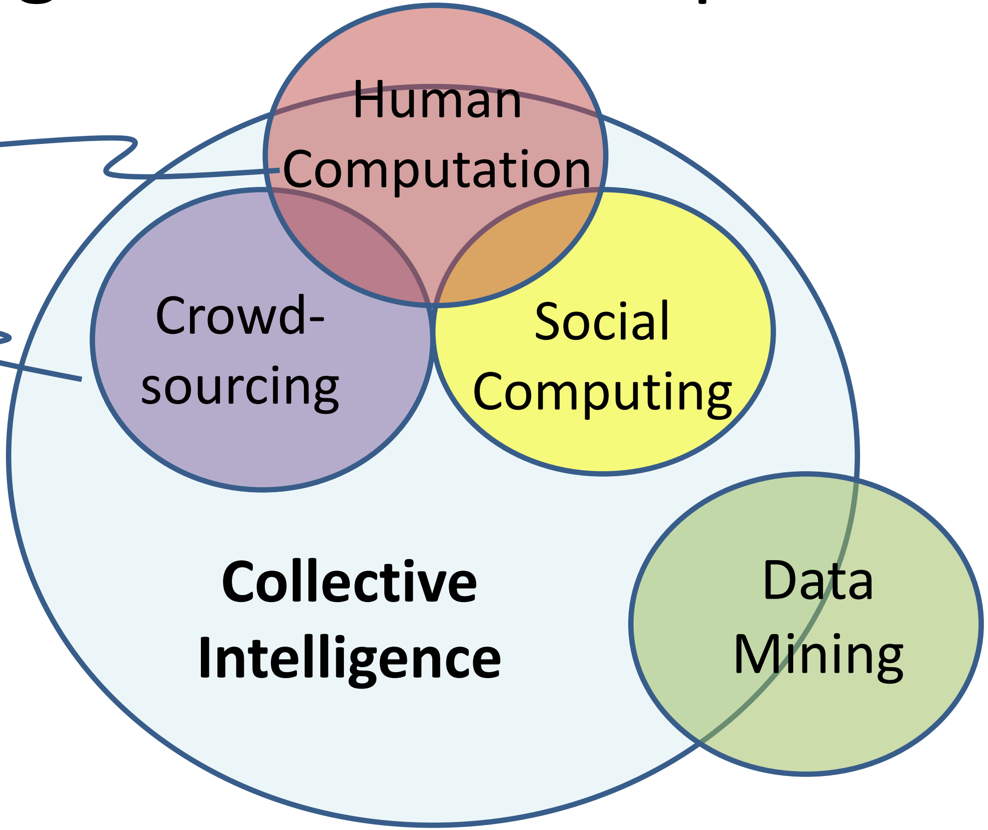


HC system could use single person working in isolation

Crowdsourcing vs. Human Computation

"Whereas *human computation* replaces computers with humans, *crowdsourcing* replaces traditional human workers with members of the public."

Quinn and Bedderson, CHI 2011



DARPA Network Challenge

DARPA Network Challenge (Wikipedia)

Challenge: Find 10 red weather balloons set up by DARPA, from 10am - 5pm EST, Dec 5, 2009. They were prepared to run the challenge for 1 week.

MIT placed first (of 10 teams). They found all 10 in less 9 hours.

MIT's Strategy: invitation-based network (with 4 initial members), finder gets $2000, with everyone in the invitation network getting 0.5 of the previous amount, with the remainder donated to charity.

MIT site: MIT Red Balloon Challenge (archived)

Social Computing

"applications and services that facilitate collective action and social interaction online with rich exchange of multimedia information and evolution of aggregate knowledge."

Parameswaran and Whinston, *Social Computing: An Overview,* CAIS 19:37, 2007

Log in Create account

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
- Help
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia

Toolbox

Print/export

Languages
- العربية
- Česky
- Deutsch
- Español
- Esperanto
- Français

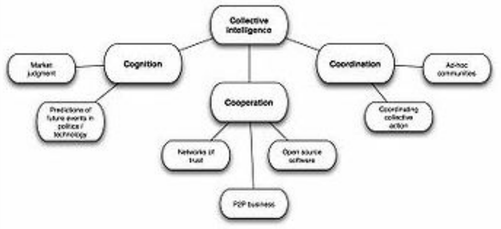
Collective intelligence

From Wikipedia, the free encyclopedia

Not to be confused with group intelligence, collaborative intelligence, or knowledge sharing.

This article **needs attention from an expert on the subject**. Please add a *reason* or a *talk* parameter to this template to explain the issue with the article. WikiProject Sociology or the Sociology Portal may be able to help recruit an expert. *(April 2010)*

Collective intelligence is a shared or group intelligence that emerges from the collaboration and competition of many individuals and appears in consensus decision making in bacteria[clarification needed], animals, humans and computer networks. The term appears in sociobiology, political science and in context of mass peer review and crowdsourcing applications. This broader definition[clarification needed] involves consensus, social capital and formalisms such as voting systems, social media and other means of quantifying mass activity. *Everything from a political party to a public wiki can reasonably be described as this loose form of collective intelligence.*



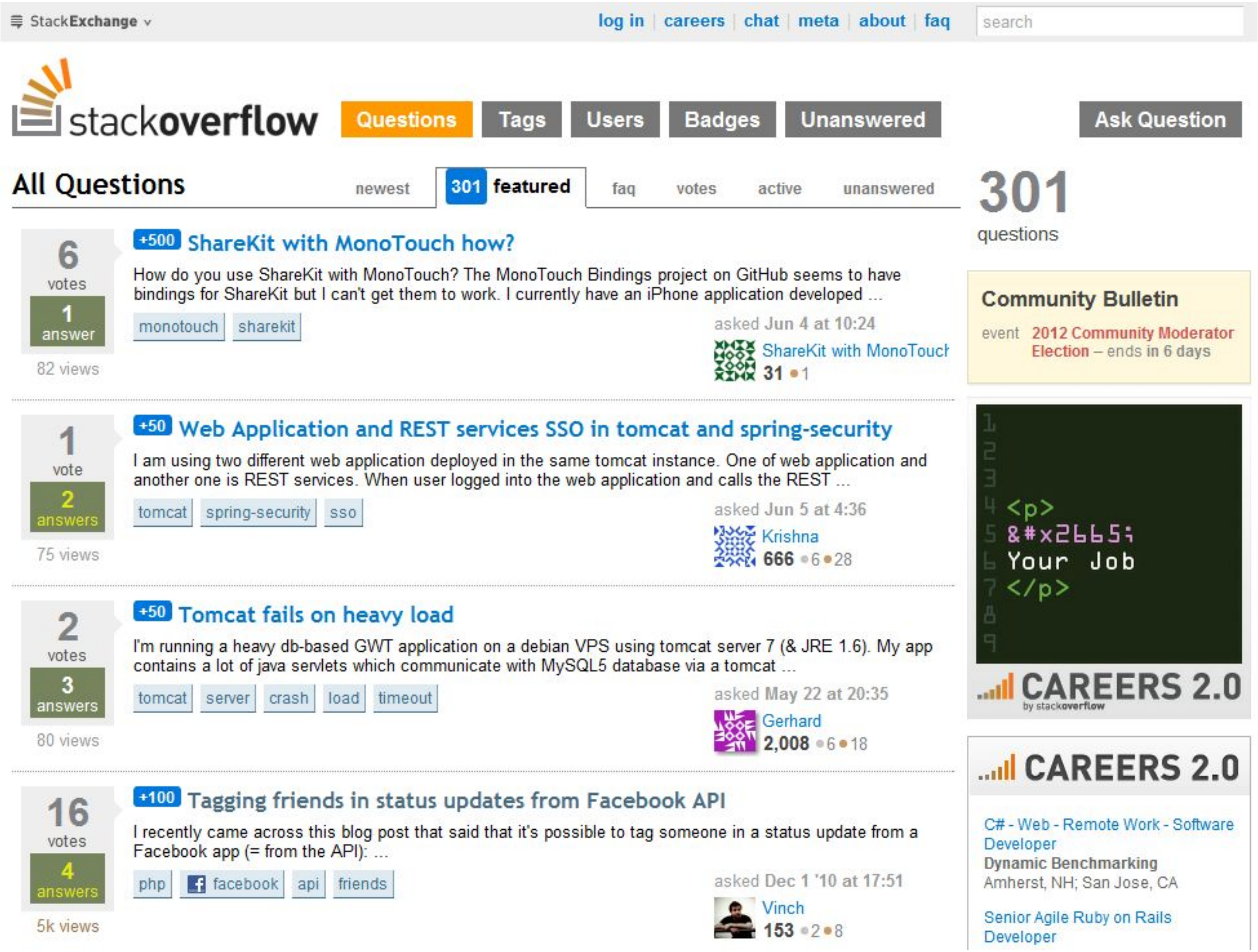
Types of collective intelligence

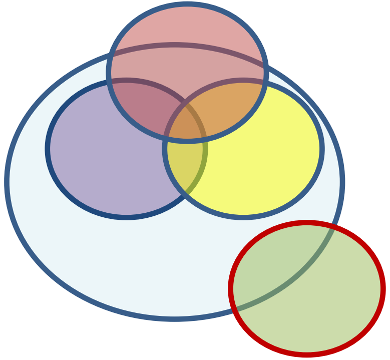
More narrowly, it can be understood as an emergent property between people and ways of processing information. [1] This notion of collective intelligence is referred to as **Symbiotic intelligence** by Norman Lee Johnson.[2] The concept is used in sociology, business, computer science and mass communications: it also appears in science fiction.

Writers who have influenced the idea of collective intelligence include Douglas Hofstadter (1979), Peter Russell (1983), Tom Atlee (1993), Pierre Lévy (1994), Howard Bloom (1995), Francis Heylighen (1995), Douglas Engelbart, Cliff Joslyn, Ron Dembo, Gottfried Mayer-Kress (2003).

Contents [hide]

Collective intelligence (Wikipedia)


StackExchange
log in | careers | chat | meta | about | faq
search
stackoverflow
Questions
Tags
Users
Badges
Unanswered
Ask Question
All Questions
newest
301 featured
faq
votes
active
unanswered
301
questions
6 votes
1 answer
82 views
+500 ShareKit with MonoTouch how?
How do you use ShareKit with MonoTouch? The MonoTouch Bindings project on GitHub seems to have bindings for ShareKit but I can't get them to work. I currently have an iPhone application developed ...
monotouch
sharekit
asked Jun 4 at 10:24
ShareKit with MonoTouch
31 •1
1 vote
2 answers
75 views
+50 Web Application and REST services SSO in tomcat and spring-security
I am using two different web application deployed in the same tomcat instance. One of web application and another one is REST services. When user logged into the web application and calls the REST ...
tomcat
spring-security
sso
asked Jun 5 at 4:36
Krishna
666 •6 •28
2 votes
3 answers
80 views
+50 Tomcat fails on heavy load
I'm running a heavy db-based GWT application on a debian VPS using tomcat server 7 (& JRE 1.6). My app contains a lot of java servlets which communicate with MySQL5 database via a tomcat ...
tomcat
server
crash
load
timeout
asked May 22 at 20:35
Gerhard
2,008 •6 •18
16 votes
4 answers
5k views
+100 Tagging friends in status updates from Facebook API
I recently came across this blog post that said that it's possible to tag someone in a status update from a Facebook app (= from the API): ...
php
facebook
api
friends
asked Dec 1 '10 at 17:51
Vinch
153 •2 •8
Community Bulletin
event 2012 Community Moderator Election – ends in 6 days
<p>
♥
Your Job
</p>
CAREERS 2.0
by stackoverflow
CAREERS 2.0
C# - Web - Remote Work - Software Developer
Dynamic Benchmarking
Amherst, NH; San Jose, CA
Senior Agile Ruby on Rails Developer

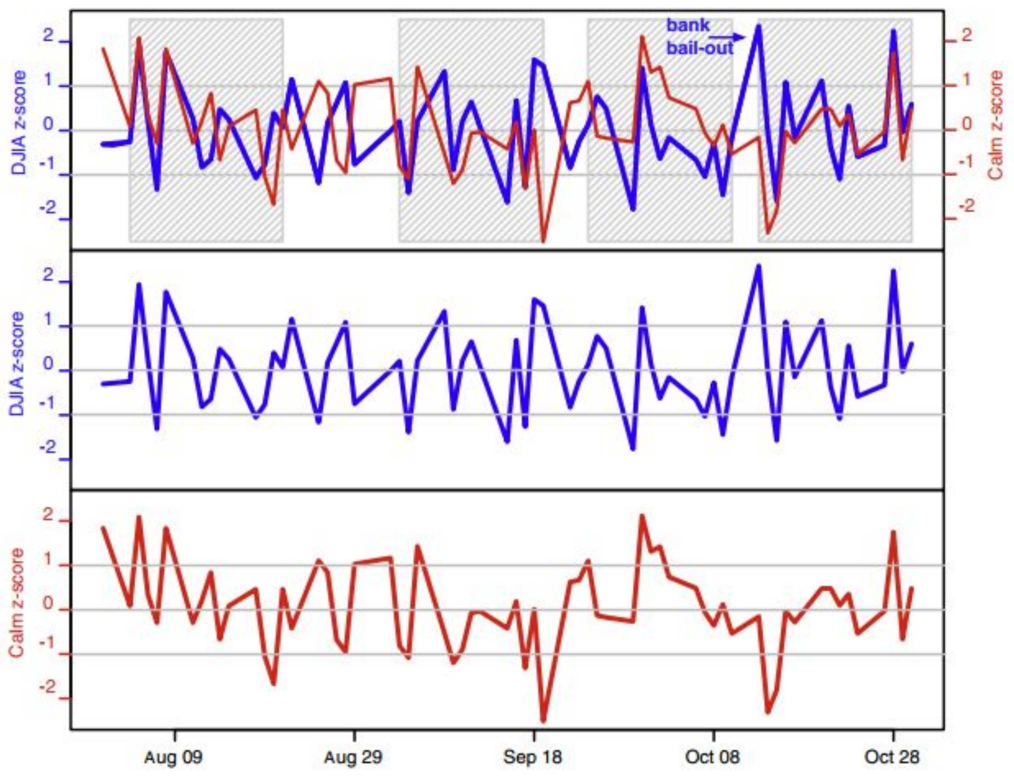


Data Mining

"the application of specific algorithms for extracting patterns from data."

Fayyad, Piatetsky-Shapiro, and Smyth,
Knowledge Discovery and Data Mining: Towards a Unifying Framework, *Proc. KDD*, 1996

J. Bollen et al. / Journal of Computational Science 2 (2011) 1–8



Bollen et al., [Twitter mood predicts the stock market](), 2011

Web Science: Collective Intelligence & Recommender Systems

(Part 2 - Intro to Recommender Systems)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle

Main reference this week:

Ch 2 from *Programming Collective Intelligence* by Toby Segaran

(abbreviated as PCI)

Recommender Systems

- Recommender (recommendation) systems recommend things to people based on their preferences or past behavior
- Two general approaches:
  - Collaborative filtering
    - You'll like *X* because other people like you also liked *X*
  - Content-based *(not covered in these slides)*
    - You'll like *X* because it is very similar to other things you like

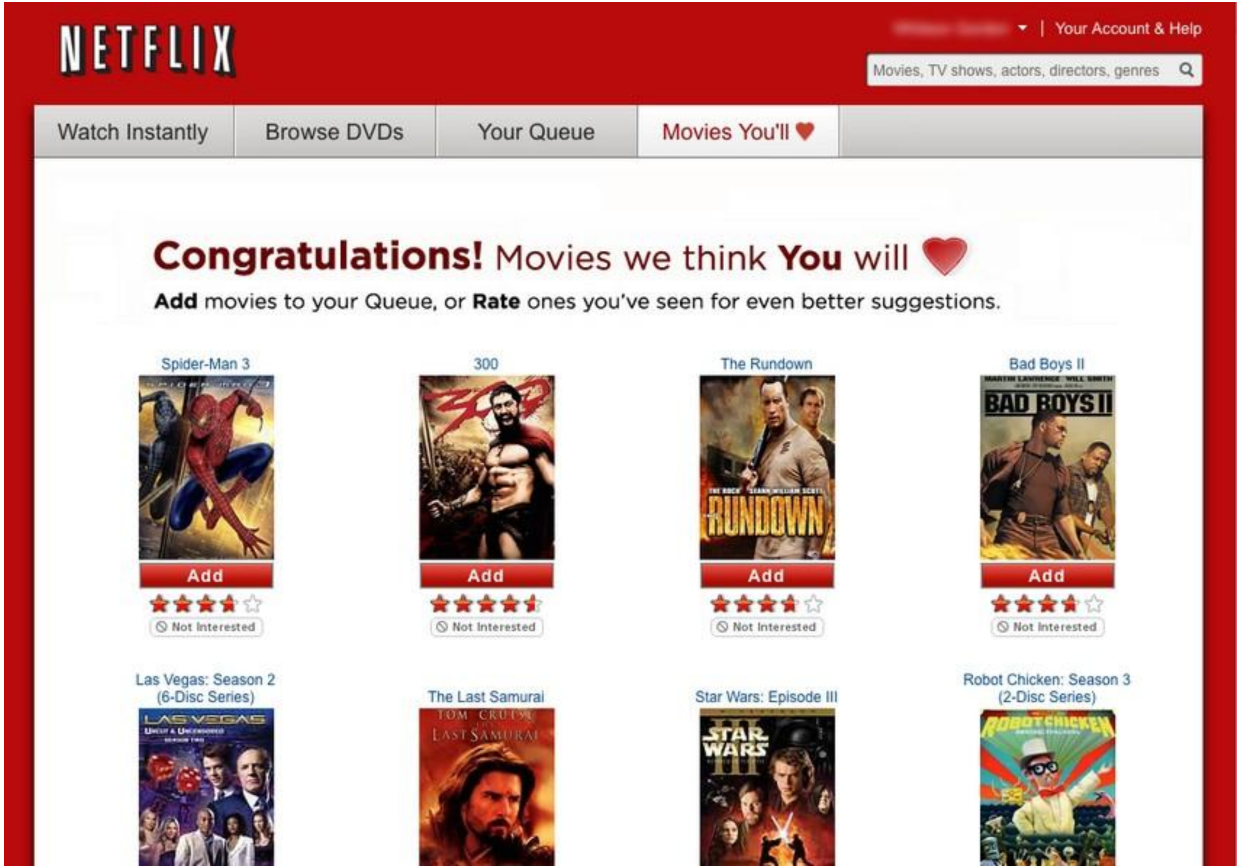


Image: http://lifehacker.com/5642050/five-best-movie-recommendation-services

Your Amazon.com › **Recommended for You**
(If you're not Frank McCown, click here.)

Just For Today

Browse Recommended

Recommendations

- Amazon Instant Video
- Appliances
- Appstore for Android
- Arts, Crafts & Sewing
- Automotive
- Baby
- Beauty
- Books
- Books on Kindle
- Camera & Photo
- Cell Phones & Accessories
- Clothing & Accessories
- Computers
- Electronics
- Grocery & Gourmet Food
- Health & Personal Care
- Home & Kitchen
- Home Improvement
- Industrial & Scientific
- Jewelry
- Kitchen & Dining
- MP3 Downloads
- Magazine Subscriptions
- Movies & TV
- Music
- Musical Instruments
- Office & School Supplies
- Patio, Lawn & Garden

These recommendations are based on items you own and more.

view: **All** | New Releases | Coming Soon

1. LOOK INSIDE!

Quiet: The Power of Introverts in a World That Can't Stop Talking
by Susan Cain (January 24, 2012)
Average Customer Review: (261)
In Stock

List Price: ~~$26.00~~
Price: $15.60
110 used & new from $14.00

Add to Cart | Add to Wish List

I own it | Not interested | Rate this item
Recommended because you added **Introverts in the Church** to your Wishlist and more (Fix this)

2. LOOK INSIDE!

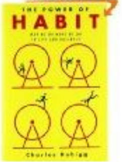
Imagine: How Creativity Works
by Jonah Lehrer (March 19, 2012)
Average Customer Review: (107)
In Stock

List Price: ~~$26.00~~
Price: $15.60
83 used & new from $13.42

Add to Cart | Add to Wish List

I own it | Not interested | Rate this item
Recommended because you added **Thinking, Fast and Slow** to your Wishlist and more (Fix this)

3. LOOK INSIDE!

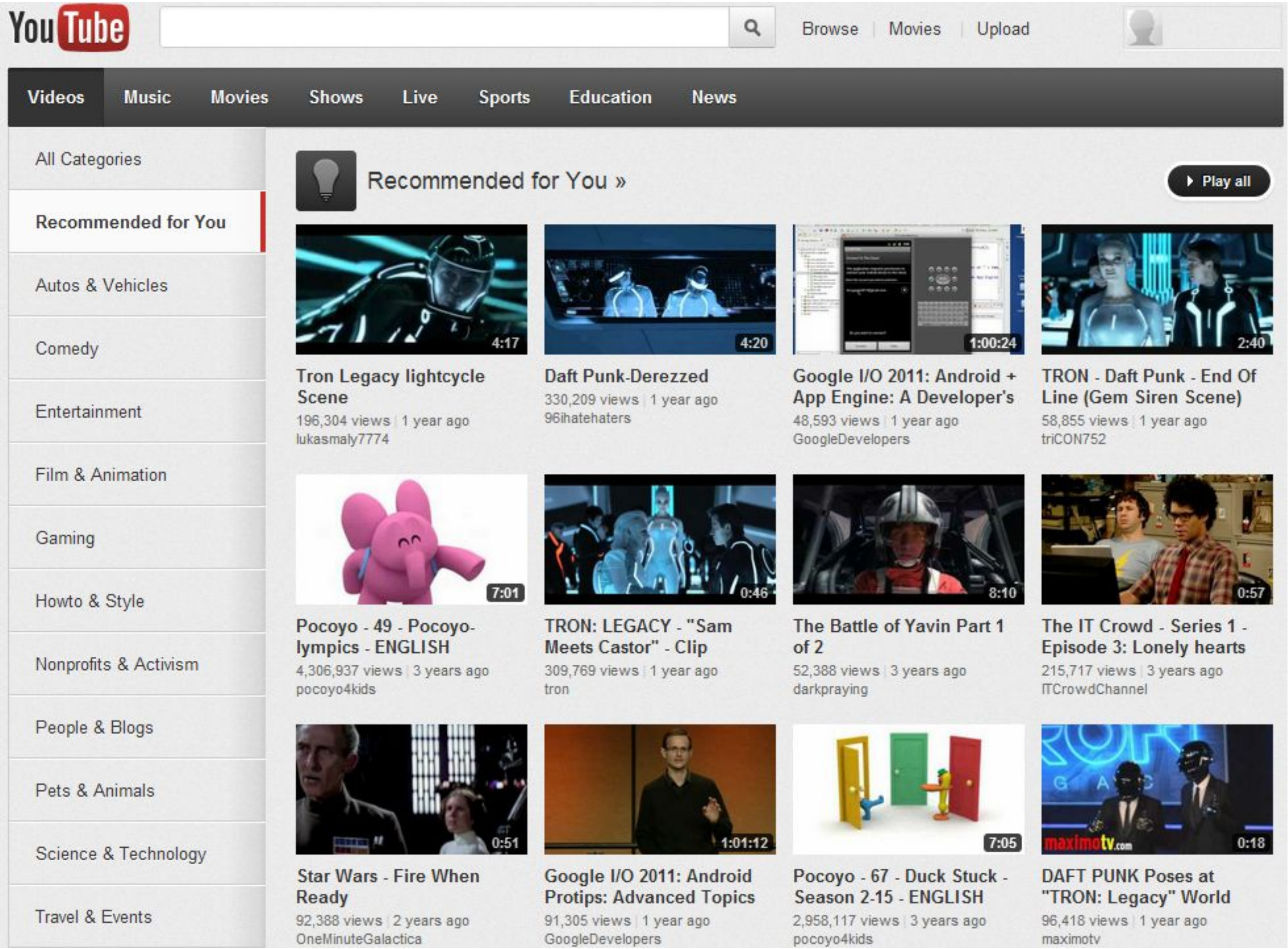


The Power of Habit: Why We Do What We Do in Life and Business
by Charles Duhigg (February 28, 2012)
Average Customer Review: (253)
In Stock

List Price: ~~$28.00~~
Price: $15.40
110 used & new from $14.91

Add to Cart | Add to Wish List

I own it | Not interested | Rate this item
Recommended because you added **Thinking, Fast and Slow** to your Wishlist and more (Fix this)



You Tube
Browse
Movies
Upload
Videos
Music
Movies
Shows
Live
Sports
Education
News
All Categories
Recommended for You
Autos & Vehicles
Comedy
Entertainment
Film & Animation
Gaming
Howto & Style
Nonprofits & Activism
People & Blogs
Pets & Animals
Science & Technology
Travel & Events
Recommended for You »
Play all
4:17
Tron Legacy lightcycle Scene
196,304 views | 1 year ago
lukasmaly7774
4:20
Daft Punk-Derezzed
330,209 views | 1 year ago
96ihatehaters
1:00:24
Google I/O 2011: Android + App Engine: A Developer's
48,593 views | 1 year ago
GoogleDevelopers
2:40
TRON - Daft Punk - End Of Line (Gem Siren Scene)
58,855 views | 1 year ago
triCON752
7:01
Pocoyo - 49 - Pocoyo-lympics - ENGLISH
4,306,937 views | 3 years ago
pocoyo4kids
0:46
TRON: LEGACY - "Sam Meets Castor" - Clip
309,769 views | 1 year ago
tron
8:10
The Battle of Yavin Part 1 of 2
52,388 views | 3 years ago
darkpraying
0:57
The IT Crowd - Series 1 - Episode 3: Lonely hearts
215,717 views | 3 years ago
ITCrowdChannel
0:51
Star Wars - Fire When Ready
92,388 views | 2 years ago
OneMinuteGalactica
1:01:12
Google I/O 2011: Android Protips: Advanced Topics
91,305 views | 1 year ago
GoogleDevelopers
7:05
Pocoyo - 67 - Duck Stuck - Season 2-15 - ENGLISH
2,958,117 views | 3 years ago
pocoyo4kids
maximotv.com
0:18
DAFT PUNK Poses at "TRON: Legacy" World
96,418 views | 1 year ago
maximotv

COMPUTING

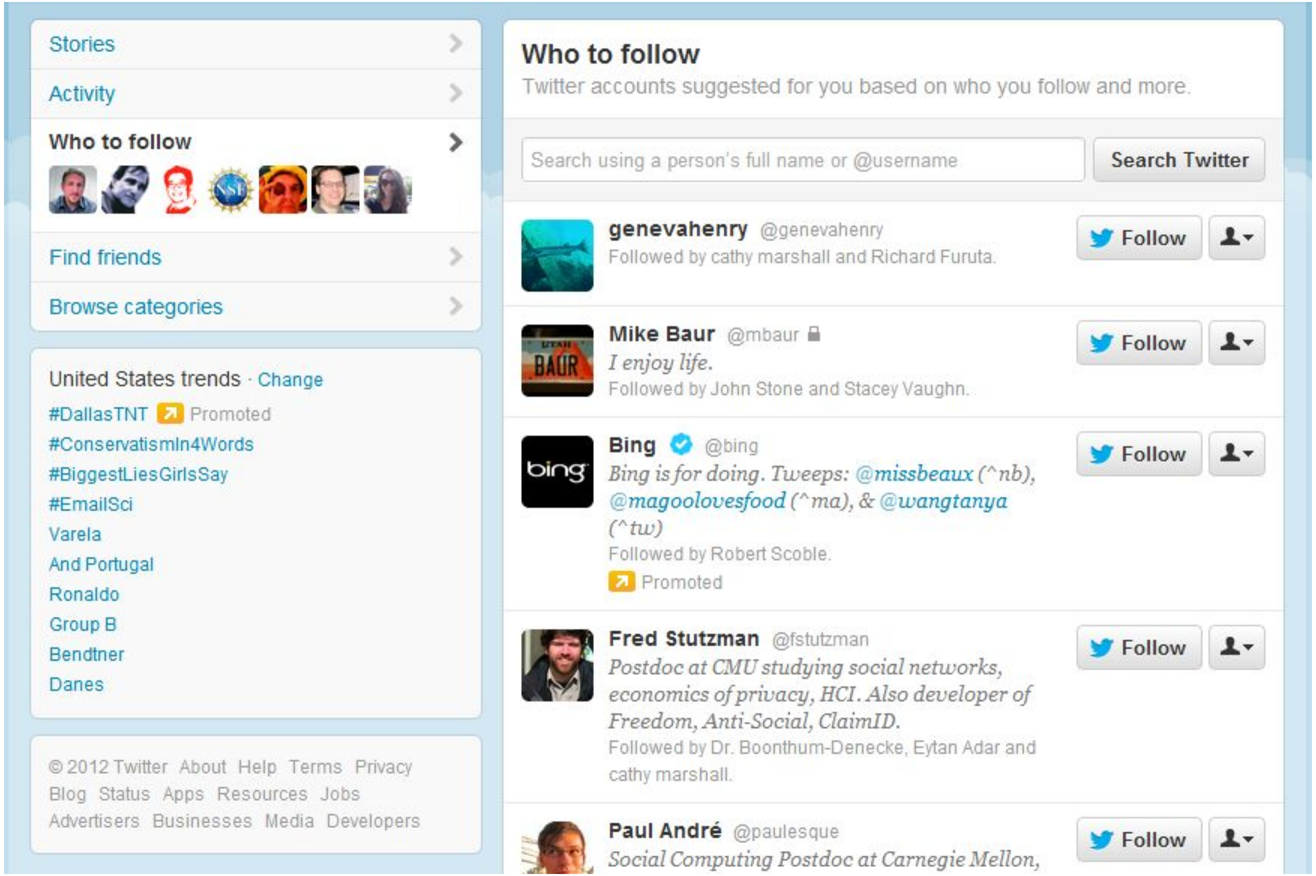
YouTube's Recommendation Algorithm Has a Dark Side

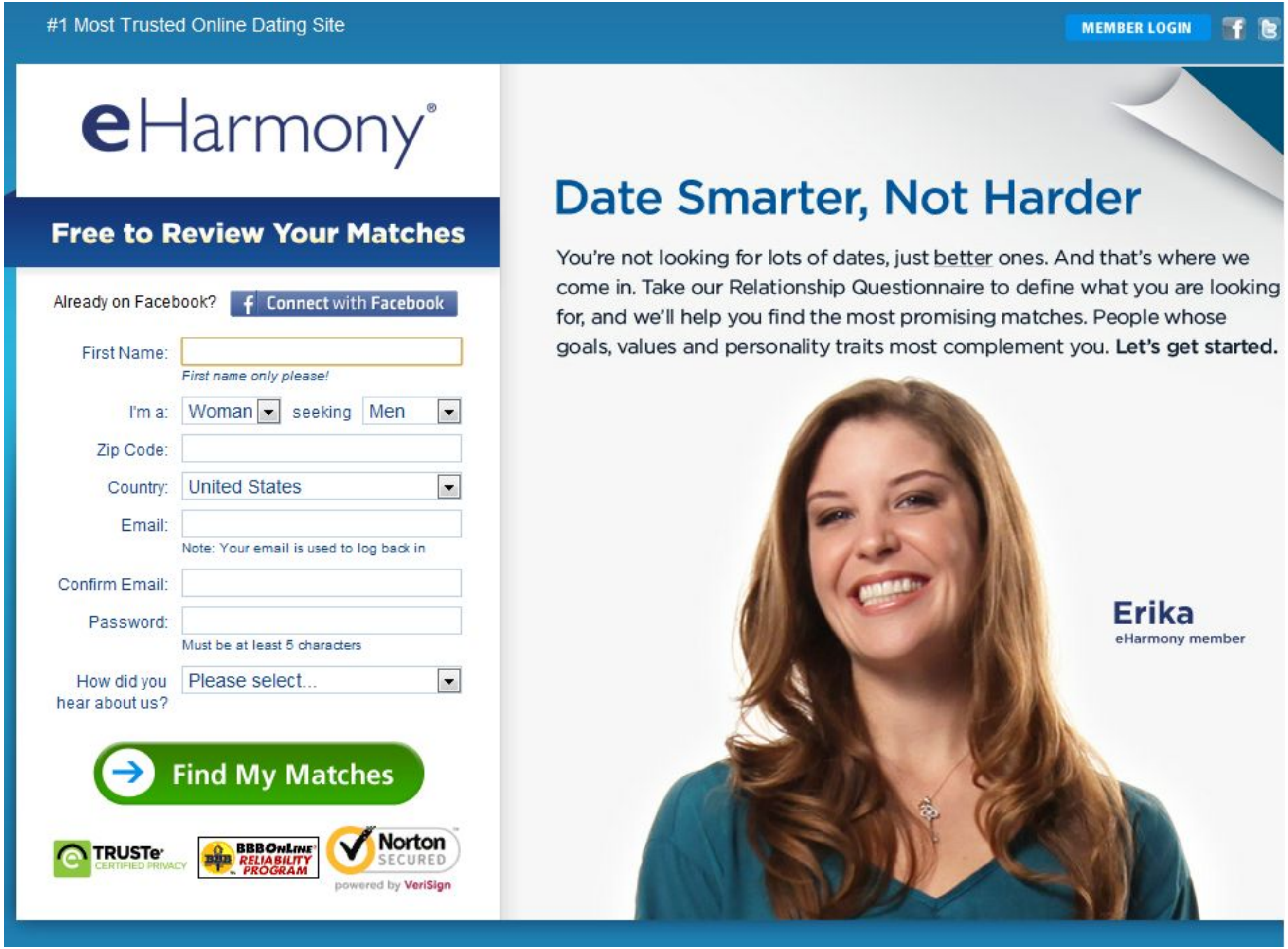
It leads users down rabbit holes

By Zeynep Tufekci on April 1, 2019 أعرض هذا باللغة العربية

> These "recommended" videos play one after the other. Maybe you finished a tutorial on how to sharpen knives, but the next one may well be about why feminists are ruining manhood, how vaccinations are poisonous or why climate change is a hoax—or a nifty explainer "proving" the *Titanic* never hit an iceberg.

Zeynep Tufekci, [YouTube's Recommendation Algorithm Has a Dark Side](), *Scientific American*, 2019


Stories
Activity
Who to follow
Find friends
Browse categories
United States trends · Change
#DallasTNT Promoted
#ConservatismIn4Words
#BiggestLiesGirlsSay
#EmailSci
Varela
And Portugal
Ronaldo
Group B
Bendtner
Danes
© 2012 Twitter About Help Terms Privacy
Blog Status Apps Resources Jobs
Advertisers Businesses Media Developers
Who to follow
Twitter accounts suggested for you based on who you follow and more.
Search using a person's full name or @username
Search Twitter
genevahenry @genevahenry
Followed by cathy marshall and Richard Furuta.
Follow
Mike Baur @mbaur
I enjoy life.
Followed by John Stone and Stacey Vaughn.
Follow
Bing @bing
Bing is for doing. Tweeps: @missbeaux (^nb), @magoolovesfood (^ma), & @wangtanya (^tw)
Followed by Robert Scoble.
Promoted
Follow
Fred Stutzman @fstutzman
Postdoc at CMU studying social networks, economics of privacy, HCI. Also developer of Freedom, Anti-Social, ClaimID.
Followed by Dr. Boonthum-Denecke, Eytan Adar and cathy marshall.
Follow
Paul André @paulesque
Social Computing Postdoc at Carnegie Mellon,
Follow


#1 Most Trusted Online Dating Site
MEMBER LOGIN
eHarmony®
Free to Review Your Matches
Already on Facebook?
Connect with Facebook
First Name:
First name only please!
I'm a:
Woman
seeking
Men
Zip Code:
Country:
United States
Email:
Note: Your email is used to log back in
Confirm Email:
Password:
Must be at least 5 characters
How did you hear about us?
Please select...
Find My Matches
TRUSTe CERTIFIED PRIVACY
BBBOnLine RELIABILITY PROGRAM
Norton SECURED
powered by VeriSign
Date Smarter, Not Harder
You're not looking for lots of dates, just better ones. And that's where we come in. Take our Relationship Questionnaire to define what you are looking for, and we'll help you find the most promising matches. People whose goals, values and personality traits most complement you. Let's get started.
Erika
eHarmony member

Collaborative Filtering

- We often seek recommendations from people we trust (family, friends, colleagues, experts)
- But who to ask for a recommendation is also based on similarity in taste
  - I trust my sister with my life, but she doesn't like the same movies I like
  - People with my tastes are likely to recommend things I will like
- CF searches a large group of people for those that like the things you like, and it combines the things they like into a ranked list of suggestions

Web Science: Collective Intelligence & Recommender Systems

(Part 3 - Recommending a Movie)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle

Main reference this week:

Ch 2 from *Programming Collective Intelligence* by Toby Segaran

(abbreviated as PCI)

GitHub repo

Example Data: Movie Ratings

	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
Rose	2.5	3.5	3.0	3.5	2.5	3.0
Seymour	3.0	3.5	1.5	5.0	3.5	3.0
Puig		3.5	3.0	4.0	2.5	4.5
LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Matthews	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	

Example from Ch 2 of *PCI* (data in code on pg. 8)

Let's visualize the data on a scatterplot…

Ratings for Snakes… and …Dupree

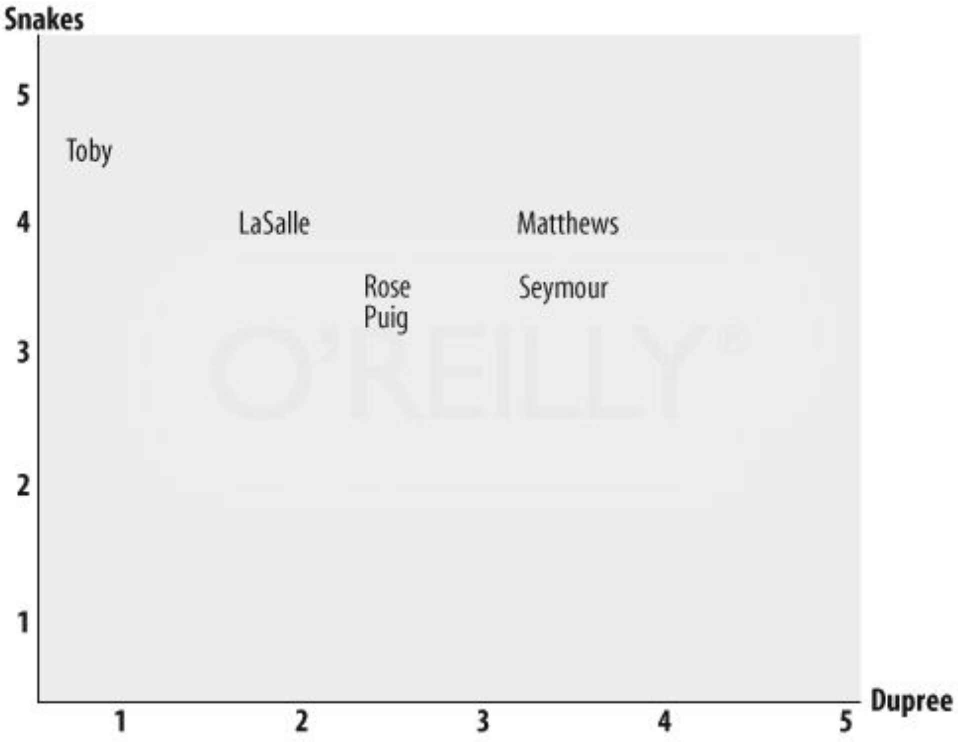
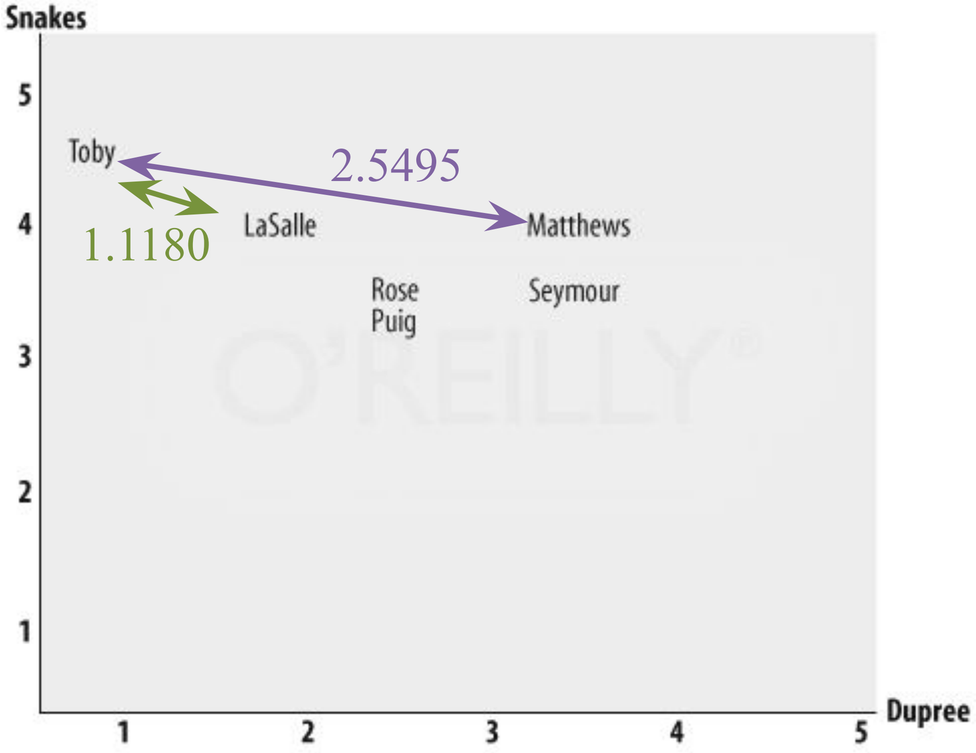


Fig 2-1, *PCI*, ["How to Find Similar Users with Python"](), archived page

Euclidean Distance Score



$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

D(Toby, Matthews)

$$= \sqrt{(1-3.5)^2 + (4.5-4)^2}$$

$= 2.5495$

D(Toby, LaSalle)

$$= \sqrt{(1-2)^2 + (4.5-4)^2}$$

$= 1.1180$

PCI, pgs. 8-15, ["How to Find Similar Users with Python"](), archived page

Modified Euclidean Distance Score

- Most similarity metrics use range [0,1]

  0 = no similarity

  1 = exactly the same

- Modify Euclidean Distance (to measure similarity)

$$D(x,y) = 1 \Big/ \left(1 + \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}\right)$$

Modified Euclidean Distance Score

	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Matthews	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	

Distance(Toby, LaSalle) - original distance value was 1.1180

$$1/(1+\sqrt{(4.5-4)^2+(4-3)^2+(1-2)^2})=0.4$$

Distance(Toby, Matthews) - original distance value was 2.5495

$$1/(1+\sqrt{(4.5-4)^2+(4-5)^2+(1-3.5)^2})=0.2675$$

Pearson Correlation Coefficient

- Problem with Euclidean distance

  Ratings for A: 5, 4, 3   Ratings for B: 3, 2, 1

  although A and B seem to share roughly same opinions, distance between ratings are significant

- Pearson's *r* corrects for "grade inflation"

- Yields values between 1 and -1

  1 = perfect correlation

  0 = no correlation

  -1 = inverse correlation

Calculating Pearson's *r*

x = User 1's ratings of all items rated by both users

y = User 2's ratings of all items rated by both users

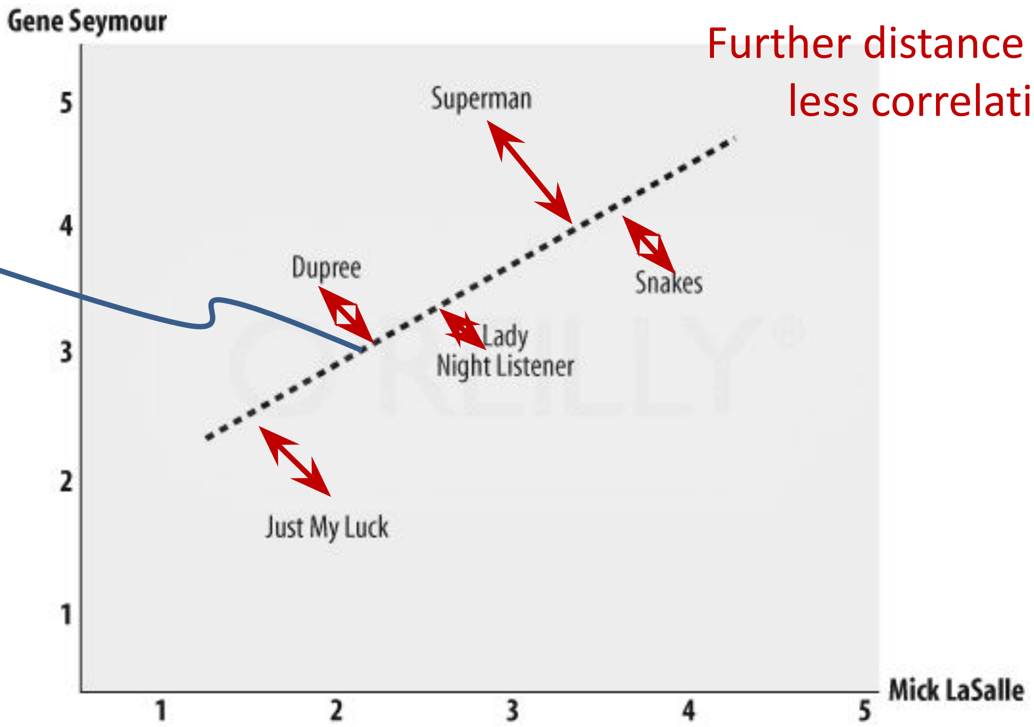
$\overline{x}$ = mean of all User 1's ratings

$\overline{y}$ = mean of all User 2's ratings

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Comparing Two Critics' Ratings

Best-fit line

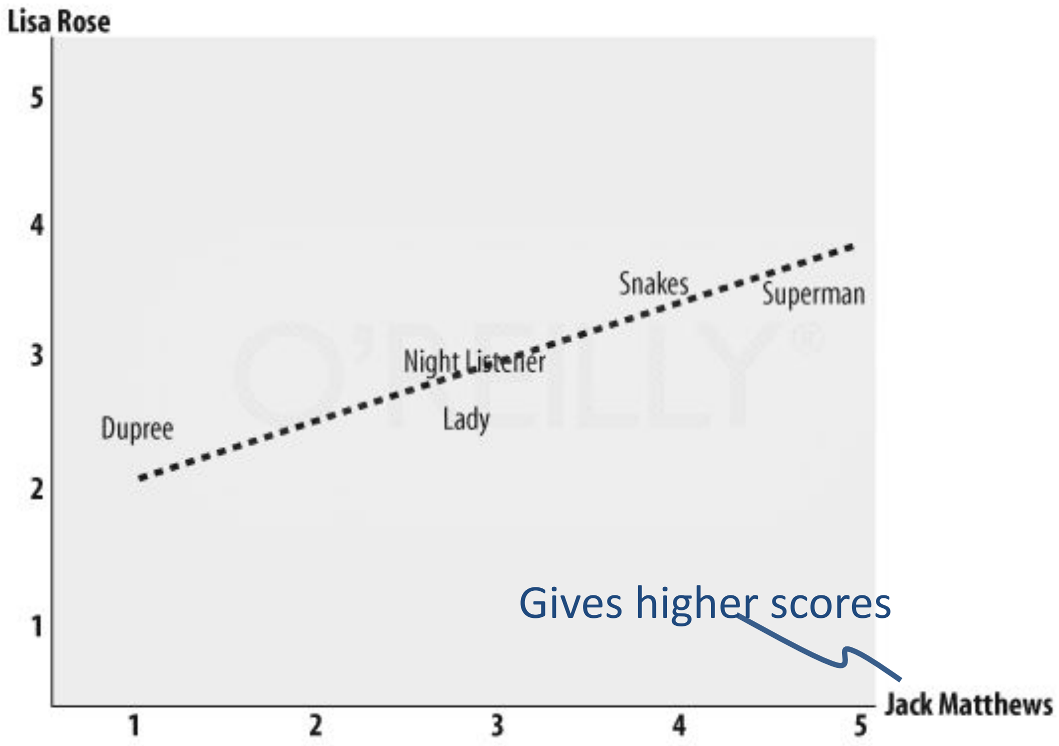


Further distance means less correlation

Pearson's $r = 0.4$

PCI, pgs. 8-15, "How to Find Similar Users with Python", archived page

Comparing Two Critics' Ratings



Pearson's

$r = 0.75$

PCI, pgs. 8-15, ["How to Find Similar Users with Python"](), archived page

Pearson's *r* in R

```
[1] cor.test(c(5, 4, 3), c(3,2,1))

        Pearson's product-moment correlation

data:  c(5, 4, 3) and c(3, 2, 1)
t = Inf, df = 1, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
sample estimates:
cor
  1
```

```
cor.test(c(5, 4, 3), c(3,1,2))

        Pearson's product-moment correlation

data:  c(5, 4, 3) and c(3, 1, 2)
t = 0.57735, df = 1, p-value = 0.6667
alternative hypothesis: true correlation is not equal to 0
sample estimates:
cor
0.5
```

[Create new R Google Colab notebook]()

Python: [scipy.stats.pearsonr — SciPy v1.5.3 Reference Guide](), and pg. 13 in *PCI*

Other Similarity Measures

- Cosine similarity (Wikipedia)
- Jaccard coefficient (Wikipedia)
- Manhattan (taxicab) distance (Wikipedia)
- Others…

Moving On…

	Lady in the Water	Snakes on a Plane	Just My Luck	Superman Returns	You, Me and Dupree	The Night Listener
Rose	2.5	3.5	3.0	3.5	2.5	3.0
Seymour	3.0	3.5	1.5	5.0	3.5	3.0
Puig		3.5	3.0	4.0	2.5	4.5
LaSalle	3.0	4.0	2.0	3.0	2.0	3.0
Matthews	3.0	4.0		5.0	3.5	3.0
Toby		4.5		4.0	1.0	

Should Toby see these movies?

PCI, pgs. 15-19

Who Should We Ask?

- To find recommendations for movies we have not seen, we could…
  1. Find a *single* critic whose taste best matches ours

     What if they haven't rated a movie we are interested in?
  2. Get *all* critics' input but give critics with *similar* tastes more impact on decision

- For option 2, use similarity metric to compare all ratings with ours, and compute average rating based on weighted similarity

Should Toby See "Lady in the Water"?

All Toby's ratings compared to other critics'

Higher values influence Weighted Avg more

	Similarity (Pearson's r)	Lady in the Water	Weighted Score
Rose	0.99	2.5	2.48
Seymour	0.38	3.0	1.14
Puig			
LaSalle	0.92	3.0	2.77
Matthews	0.66	3.0	1.99
Total	2.95		8.38

0.99×2.5

Not included in Total since no rating for Lady in the Water

Weighted Avg = Weighted Total / Similarity Total = 8.38/2.95 = **2.83**

Toby is likely to rate this movie 2.83

Product Recommendations

- Previous example used distance metric to find critics with similar taste
- What if we want to find movies that are similar to some given movie?
- Solution: Use same method but swap rows and columns

Rows and Columns Swapped

	Rose	Seymour	Puig	LaSalle	Matthews	Toby
Lady in the Water	2.5	3.0		3.0	3.0	
Snakes on a Plane	3.5	3.5	3.5	3.5	4.0	4.0
Just My Luck	3.0	1.5	3.0	2.0		
Superman Returns	3.5	5.0	4.0	3.0	5.0	4.0
You, Me and Dupree	2.5	3.5	2.5	2.0	3.5	1.0
The Night Listener	3.0	3.0	4.5	3.0	3.0	

Find movies like *Superman Returns* by comparing its row with all other rows…

Movies Similar to *Superman Returns*

	Similarity
You, Me and Dupree	0.657
Lady in the Water	0.487
Snakes on a Plane	0.111
The Night Listener	-0.179
Just My Luck	-0.422

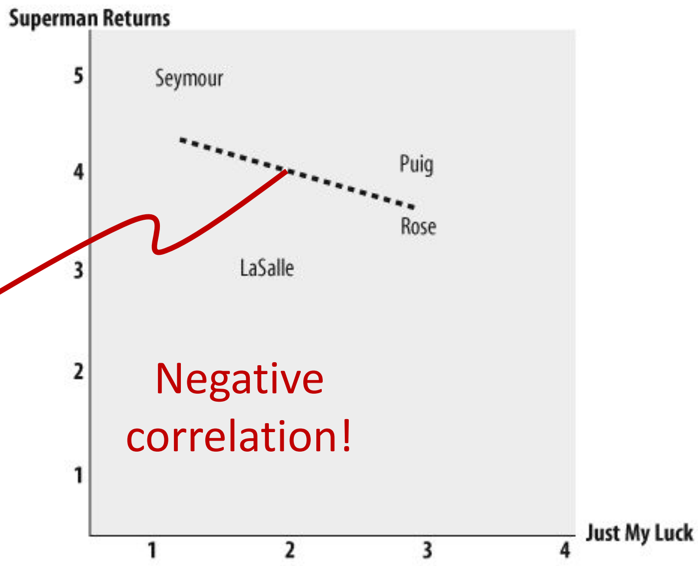


Fig 2-5 from PCI

If you like *Superman Returns,* you probably won't like *Just My Luck* (and vice versa)!

Web Science: Collective Intelligence & Recommender Systems

(Part 4 - Challenges for Collaborative Filtering)

CS 432/532

Old Dominion University

Permission has been granted to use these slides from Frank McCown, Michael L. Nelson, Alexander Nwala, Michele C. Weigle

Challenges for Collaborative Filtering

- Ratings data is often sparse
  - Large data sets required
  - Cold start problem: new users must first rate a number of items before CF can work
- Comparing all items against all users is not efficient/scalable
  - Compare against a sample
  - Use clustering algorithm to locate best neighbors

Clustering algorithms are covered in Ch 3 of *Programming Collective Intelligence*

Challenges for Collaborative Filtering

- Susceptible to cheating (Shilling attacks)
  - Rate your own products higher and your competitors lower
  - Rate like everyone else except a select few items you want recommended to others

Lam and Riedl, "Shilling Recommender Systems for Fun and Profit", WWW 2004
Img source: http://www.somethingawful.com/d/photoshop-phriday/cheating-in-sports.php

Say it isn't true!

Hidden Industry Dupes Social Media Users

Paying people to influence discussions in social media is big business in China and the U.S.

By Tom Simonite on December 12, 2011

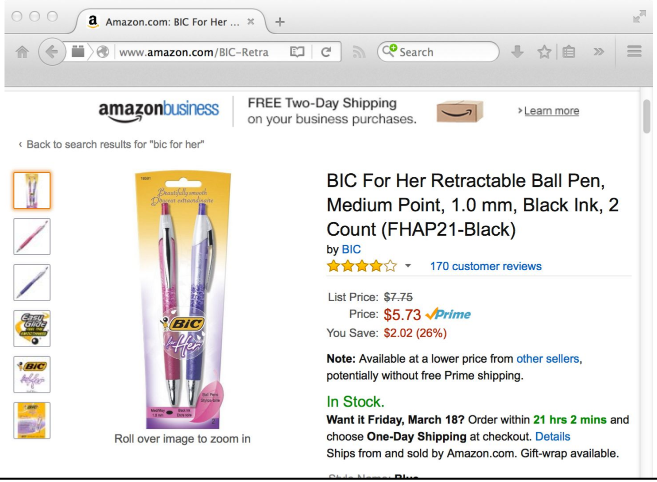
A trawl of Chinese crowdsourcing websites – where people can earn a few pennies for small jobs such as labeling images – has uncovered a multimillion-dollar industry that pays hundreds of thousands of people to distort interactions in social networks and to post spam.

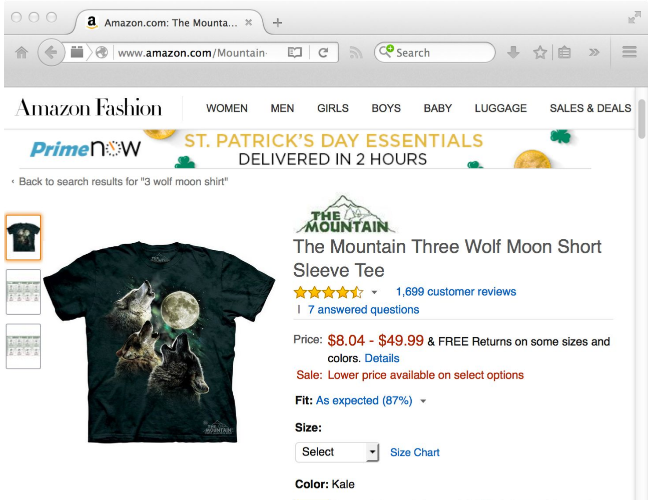
Hidden Industry Dupes Social Media Users

Astroturfing (Wikipedia)

OSoMe (Observatory on Social Media), Indiana University

Some reviews are sarcastic…





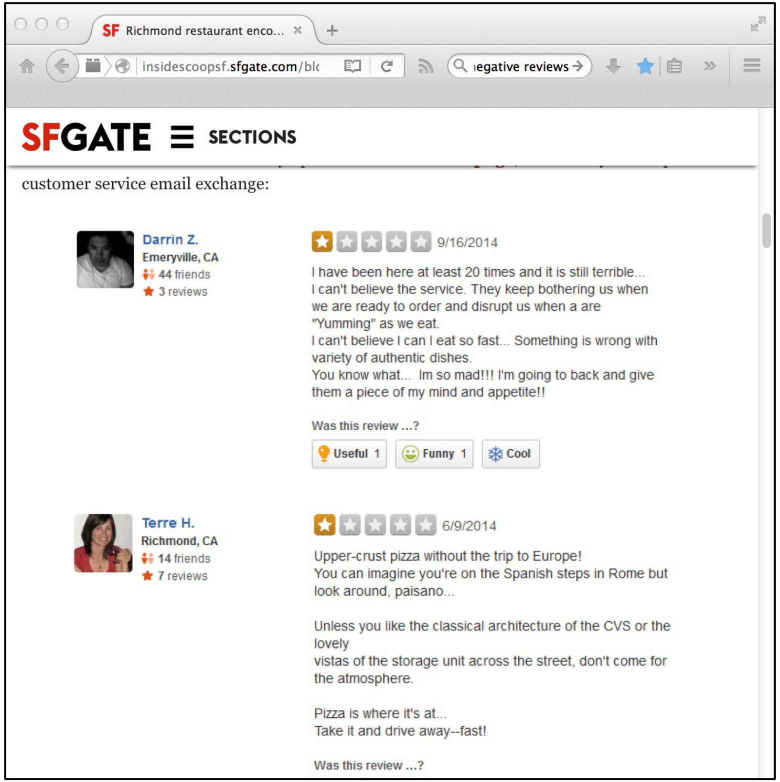
5.0 out of 5 stars Received As Wedding Gift- LOVE THEM!
By agilerton on August 28, 2012
Style Name: BlueSize: 1 - Pack
I received a set of these pens for my wedding. Unfortunately I can't write or type (my husband is actually typing for me.) However, I do so love the pretty colors. I've been learning how to write with them. …

[BIC For Her Retractable Ball Pen at Amazon.com]

5.0 out of 5 starsGreat compliment for my skin art
By overlook1977 on May 19, 2009
Size: MediumColor: Dark Green
Unfortunately I already had this exact picture tattooed on my chest, but this shirt is very useful in colder weather.

[The Mountain Men's Three Wolf Moon Short Sleeve Tee at Amazon.com]

Some are protests.



[Richmond restaurant encourages bad Yelp reviews](#)

[Botto Italian Bistro](#) (yelp)

Netflix Prize

- Netflix Prize (Oct 2006): $1M for beating Netflix's collaborative filtering algorithm by 10%
- Dataset: 100,480,507 ratings that 480,189 anonymized users gave to 17,770 movies
- Started with 50,051 contestants
- Only two teams had winning solutions by contest closing (July 2009)

Img source: http://www.wired.com/images_blogs/business/2009/09/p1010915.jpg

How Did They Do It?

- Insights that gave winning algorithms an edge:
  - People who rate a large number of movies at once are usually rating movies they saw a long time ago
  - People use different criteria when rating movies they saw a long time ago vs. recently seen
  - Some movies get better over time, others get worse
  - People rate movies differently depending on which day of the week it is
- Combined hundreds of other algorithms which were precisely weighed and tuned

How the Netflix Prize Was Won by Buskirk, Wired 2009

Netflix Prize 2

- Second prize announced in 2009
- FTC and lawsuits against Netflix about privacy
- I thought the data was **anonymized**?

Img source: http://laist.com/2008/02/10/photo_essay_ano.php

Anonymizing Data Is Hard To Do…

- In 2007 Narayanan and Shmatikov (Univ of Texas) were able to re-identify a number of the anonymous users in Netflix dataset by correlating movie ratings with IMDb ratings (full paper)
- This has happened before…
  - In 2006 users in anonymized AOL search data were re-identified
  - In 2000 Latanya Sweeney showed 87% of all Americans could be uniquely identified with only zip code, birthdate, and gender

See Why Anonymous Data Sometimes Isn't by Bruce Schneier (2007)

Netflix gave up…
no more prizes!

Objectives

- Describe and define the four components of collective intelligence.
- Explain how collaborative filtering is related to recommender systems.
- Differentiate between Euclidean distance and the Pearson correlation coefficient.
- List three challenges for collaborative filtering.