Recommender Systems & Collective Intelligence COMP47580

Barry Smyth
http://barrysmyth.me, @barrysmyth

Part 0 - Introduction

Overview

Housekeeping

Course structure, assessment, timetabling, expectations etc.

Framing Collective Intelligence

What is collective intelligence? How does it work? Why does it work? When does it work?

Games with a Purpose

Incentives for participation. Games & fun + collective intelligence => Games-with-a-purpose (GWAPs). Lots of examples. Building your own GWAP.

Housekeeping

Assessment

Assessment

100% continuous assessment throughout the semester.

Game-with-a-purpose (GWAP) Project, 50%

Group-based (x3), prototyping and presentation (report & seminar)

Recommender Systems Project

Coding (Java), evaluation, report.

Lectures & Labs

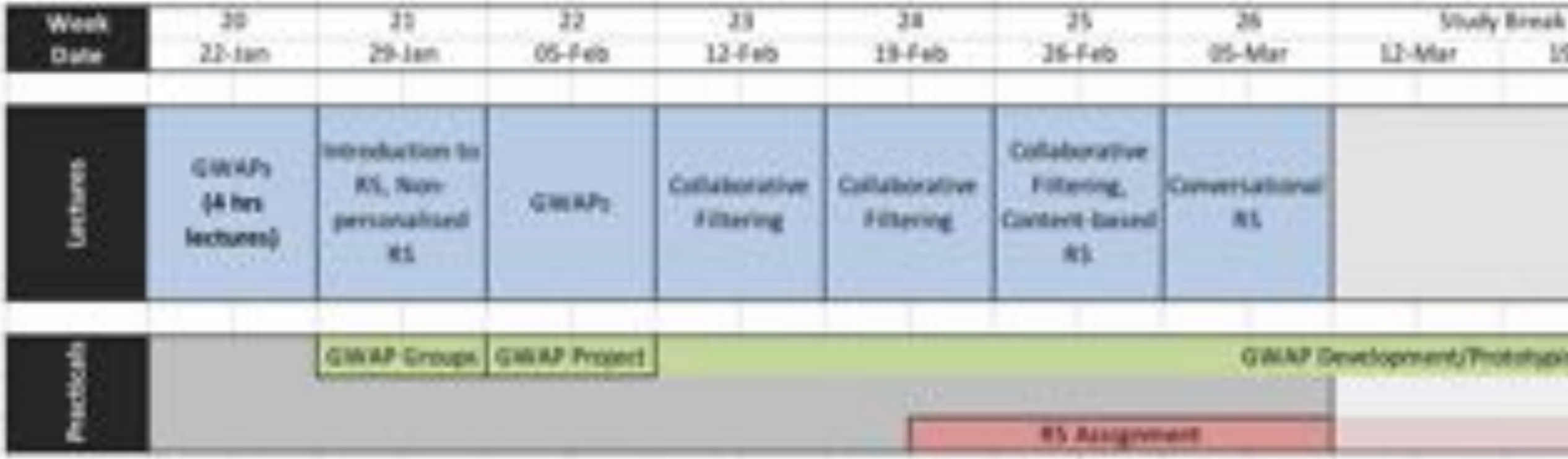
Lectures

11am-1pm Thursdays in B.002.
Slides will on Moodle each week (***Enrollment Key = COMP47580-17-18-S2***).

Laboratories

3-5pm Wednesdays in B.002.
Project discussion, demonstrator support etc.

There are some important exceptions (see next slides)



Lectures

Week 20: 4 hrs lectures (GWAPs)
Week 21: Lab (wed): GWAP groups.
Week 22: 2-4hrs (GWAPs) Wed/Thurs
Week 23: no lecture
All other weeks: 2 hrs lectures

GWAP Assignment

Week 21: GWAP Group Formation
Week 22: GWAP Detailed Assignment
Week 31: GWAP Final Submission
Week 32: GWAP Presentations #1
Week 33: GWAP Presentations #2

RS Assignment

Week 24: handed out
Week 25: 1st deliverable (UBCF)
Week 26: 2nd deliverable (IBCF)
Week 29: 3rd deliverable (CB)
Week 30: 4th deliverable (MF)
Week 31: 5th deliverable (MF, report)

	25 16-Feb	26 05-Mar	Study Break 12-Mar – 19-Mar	29 26-Mar	30 02-Apr	31 09-Apr	32 16-Apr	33 23-Apr
ive g	Collaborative Filtering, Content-based RS	Conversational RS		Robustness & Trust	Opinionated RS, Explanations	Opinionated RS, Explanations	Summary & Conclusions	No Lecture
	GWAP Development/Prototyping						GWAP Pres.	GWAP Pres.
	RS Assignment			RS Assignment contd				

RS Assignment

Week 24: handed out

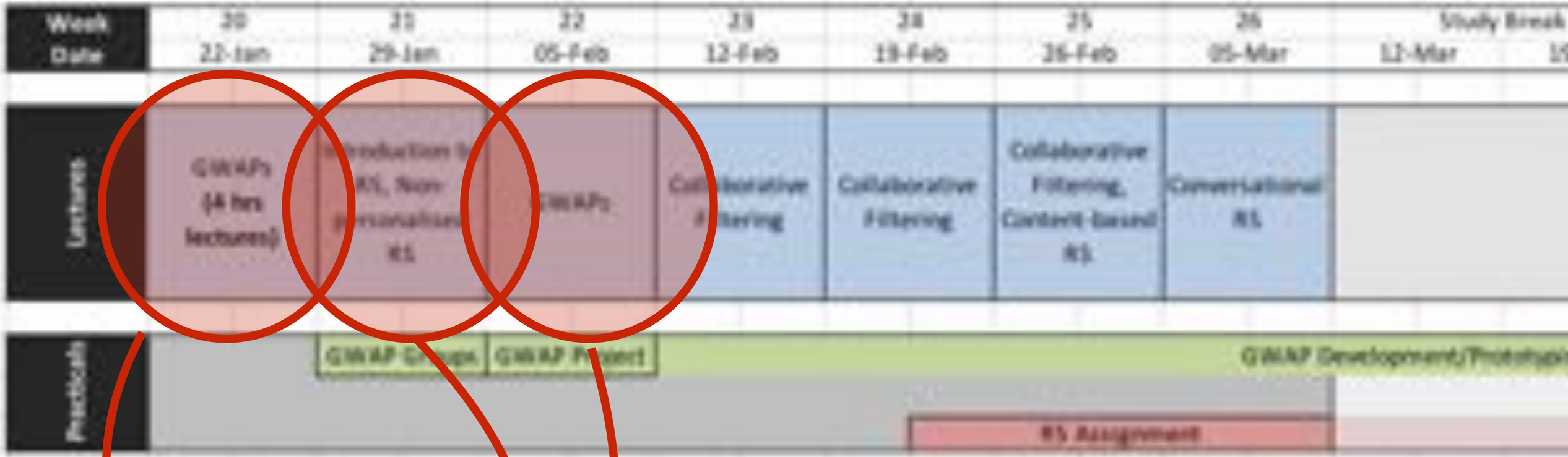
nent Week 25: 1st deliverable (UBCF)

Week 26: 2nd deliverable (IBCF)

Week 29: 3rd deliverable (CB)

Week 30: 4th deliverable (MF)

Week 31: 5th deliverable (MF, report)



Lectures

Week 20: 4 hrs lectures (GWAPs)
Week 21: Lab (wed): GWAP groups.
Week 22: 2-4hrs (GWAPs) Wed/Thurs
Week 33: no lecture
All other weeks: 2 hrs lectures

GWAP Assignment

Week 21: GWAP Group Formation
Week 22: GWAP Detailed Assignment
Week 31: GWAP Final Submission
Week 32: GWAP Presentations #1
Week 33: GWAP Presentations #2

RS Assignment

Week 24: handed out
Week 25: 1st deliverable (UBCF)
Week 26: 2nd deliverable (IBCF)
Week 29: 3rd deliverable (CB)
Week 30: 4th deliverable (MF)
Week 31: 5th deliverable (MF, report)

Summary GWAP Timetable

Week 20: *Lectures @ 3pm on Wednesday & 11am on Thursday*

Week 21: *Labs @ 3pm on Wednesday (Finalise GWAP Groups)*

Week 22: *Lectures @ 3pm on Wednesday / Labs 11am Thursday*

Weeks 23/24: *Labs @ 11am Thursday (GWAP Project)*

Week 31: *GWAP Submission (Details to be provided)*

Weeks 32/33: *GWAP Group Project Presentations*

Expectation Setting & Grading

With 100% continuous assessment this module requires a significant level of continuous effort throughout the semester.

Each week involves lectures, labs, and project work.

5 Credits ~ 70 hours of project effort => 35 hours per student per group for the GWAP assignment (~ 4 hours per week).

Computer Science Marking/Grading Scheme available at:
https://www.cs.ucd.ie/Grading/

Grade	Min	Max	Average
A+	95	100	97.5
A	90	95	92.5
A-	85	90	87.5
B+	80	85	82.5
B	75	80	77.5
B-	70	75	72.5
C+	65	70	67.5
C	60	65	62.5
C-	55	60	57.5
D+	50	55	52.5
D	45	50	47.5
D-	40	45	42.5

D+	50	55	52.5
D	45	50	47.5
D-	40	45	42.5
E+	35	40	37.5
E	30	35	32.5
E-	25	30	27.5
F+	20	25	22.5
F	15	20	17.5
F-	10	15	12.5
G+	8	10	9
G	5	8	6.5
G-	2	5	3.5
NG	0	0	0

Grade	Criteria more relevant to levels* 0, 1 and 2 Knowledge, understanding, application	Additional criteria more relevant to levels** 3, 4, 6 and 7 Analysis, synthesis, evaluation
A	**Excellent** A comprehensive, highly-structured, focused and concise response to the assessment task, consistently demonstrating • an extensive and detailed knowledge of the subject matter • a highly-developed ability to apply this knowledge to the task set • evidence of extensive background reading • clear, fluent, stimulating and original expression • excellent presentation (spelling, grammar, graphical) with minimal or no presentation errors	A deep and systematic engagement with the assessment task, with consistently impressive demonstration of a comprehensive mastery of the subject matter, reflecting; • a deep and broad knowledge and critical insight as well as extensive reading; • a critical and comprehensive appreciation of the relevant literature or theoretical, technical or professional framework • an exceptional ability to organise, analyse and present arguments fluently and lucidly with a high level of critical analysis, amply supported by evidence, citation or quotation; • a highly-developed capacity for original, creative and logical thinking

		creative and logical thinking
B	**Very Good** A thorough and well-organised response to the assessment task, demonstrating • a broad knowledge of the subject matter • considerable strength in applying that knowledge to the task set • evidence of substantial background reading • clear and fluent expression • quality presentation with few presentation errors	A substantial engagement with the assessment task, demonstrating • a thorough familiarity with the relevant literature or theoretical, technical or professional framework • well-developed capacity to analyse issues, organise material, present arguments clearly and cogently well supported by evidence, citation or quotation; • some original insights and capacity for creative and logical thinking
C	**Good** An adequate and competent response to the assessment task, demonstrating • adequate but not complete knowledge of the subject matter • omission of some important subject matter or the appearance of several minor errors • capacity to apply knowledge appropriately to the task albeit with some errors	An intellectually competent and factually sound answer with, marked by, • evidence of a reasonable familiarity with the relevant literature or theoretical, technical or professional framework • good developed arguments, but more statements of ideas • arguments or statements adequately but not well supported by evidence, citation or quotation • some critical awareness and analytical

	presentation errors	
C	**Good** An adequate and competent response to the assessment task, demonstrating • adequate but not complete knowledge of the subject matter • omission of some important subject matter or the appearance of several minor errors • capacity to apply knowledge appropriately to the task albeit with some errors • evidence of some background reading • clear expression with few areas of	An intellectually competent and factually sound answer with, marked by, • evidence of a reasonable familiarity with the relevant literature or theoretical, technical or professional framework • good developed arguments, but more statements of ideas • arguments or statements adequately but not well supported by evidence, citation or quotation • some critical awareness and analytical qualities • some evidence of capacity for original and logical thinking

GWAP Project Outline

Design and prototype a plausible GWAP.

Details to be provided - major design/protyping project involving the creation of a plausible game-with-a-purpose to address a realistic collective intelligence task.

Project Outputs

‘Working’ prototype (eg MarvelApp, Web, etc.), screencast, group collaboration log, project report, project presentation.

GWAP Project Groups

The project is designed ideally for groups of 3 students, ***working together*** on the design, prototyping, reporting.

Ideally students should self organise into groups but for those who cannot groups will be assigned (***Week 21***).

Groups will be expected to maintain a ***Group Log*** to document meetings, decisions, contributions etc.

Any ***problems*** that occur should be raised as early as possible so that appropriate action can be taken. Highlighting a problem at the end of the project is unlikely to produce a satisfactory result for all concerned.

Demonstrators will be on-hand during laboratory hours to help with any questions or issue related to the project

A Note on Plagiarism

Plagiarism is a serious academic offence

See Section 6.2 of Student Code or UCD Registry Plagiarism Policy or the School's Plagiarism Policy & Procedures document.

A proactive approach will be taken to detect incidents

Suspected incidents will be referred directly to the school's Plagiarism Sub-Committee who will interview and investigate those involved.

Penalties

Typically 0% or NG for a first offence. 2nd offence may be referred to UCD's Disciplinary Committee.

UCD Assessment Submission Form

All students will be required to provide a signed Assessment Submission Form with their GWAP submissions.

Student’s who enable plagiarism are normally viewed as equally responsible…

… please, just don't do it!

Questions?

Part 1 - Collective Intelligence

Vox Populi & the Wisdom of the Crowd

The West of England Fat Stock and Poultry Exhibition, Plymouth 1906

Guess the weight of the ox …

6d per entry. Approx 800 entrants, including butchers, farmers, but also the general public, etc.

787 legitimate guesses (13 eliminated due to legibility problems)

How well do you think the crowd did?

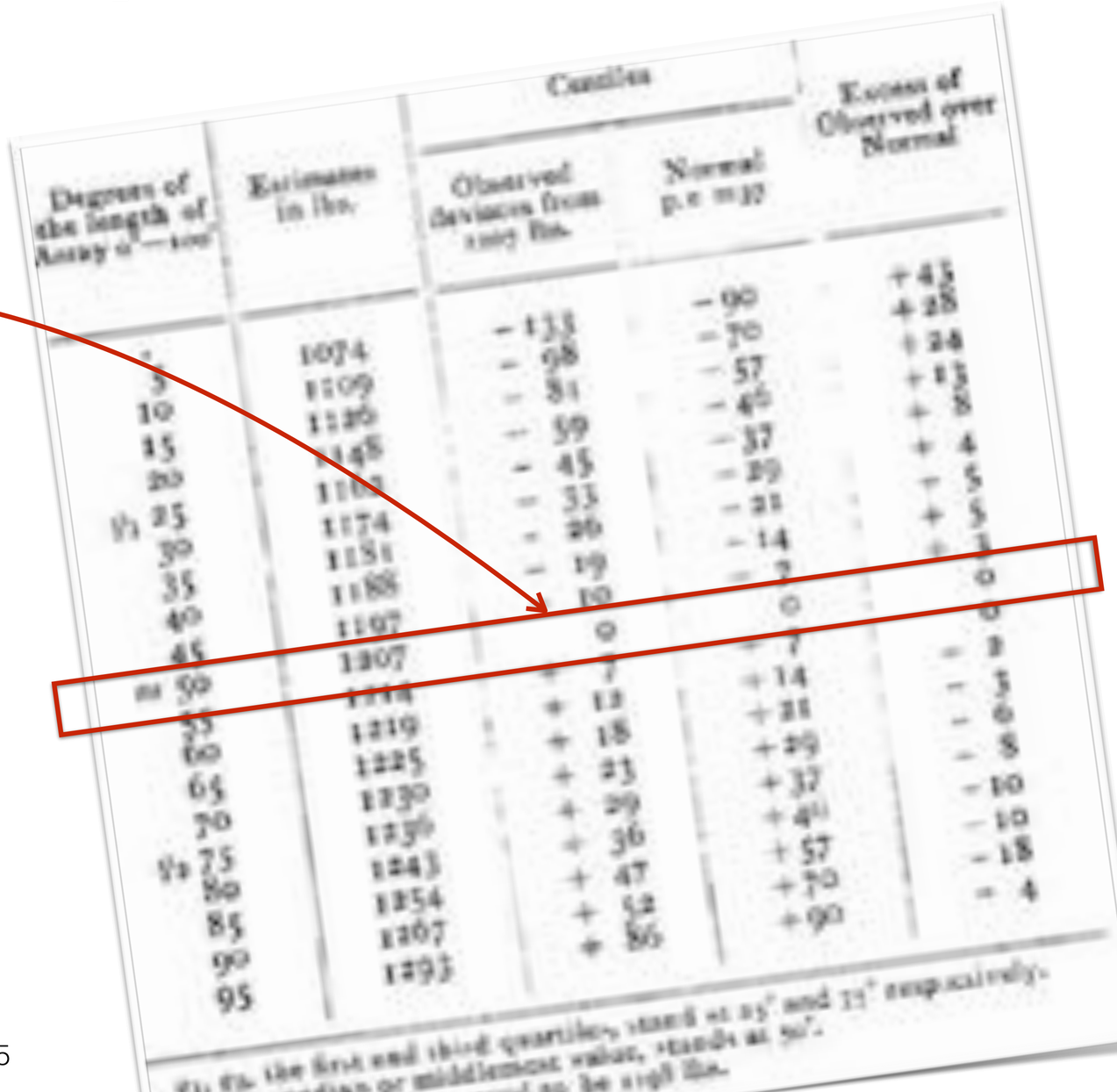
Vox Populi, Nature (1907), No. 1949, Vol. 75,

Distribution of estimates after conversions to *lbs.*

Median guess: *1,207 lbs*

Correct weight: *1,198 lbs*

In other words the crowd's guess fell within 1% of the true weight of the ox!



Degrees of the length of Assay = 100	Estimates in lbs.	Centiles		Excess of Observed over Normal
		Observed deviation from 1207 lbs.	Normal p.e. 21.37	
1	1074	−133	−90	+43
5	1109	− 98	− 70	+28
10	1126	− 81	− 57	+24
15	1145	− 59	− 46	+13
20	1162	− 45	− 37	+ 8
½ 25	1174	− 33	− 29	+ 4
30	1181	− 26	− 21	+ 5
35	1188	− 19	− 14	+ 5
40	1197	− 10	− 7	+ 3
45	1207	0	0	0
m 50	1214	+ 7	+ 7	0
55	1219	+ 12	+ 14	− 2
60	1225	+ 18	+ 21	− 3
65	1230	+ 23	+ 29	− 6
70	1236	+ 29	+ 37	− 8
½ 75	1243	+ 36	+ 47	− 10
80	1254	+ 47	+ 57	− 10
85	1267	+ 52	+ 70	− 18
90	1293	+ 86	+ 90	− 4
95				

the first and third quartiles, stand at 25° and 75° respectively.
middlemost value, stands at 50°.
the right line.

Crowd wisdom or a lucky guess?

Was this just a lucky guess or is this type of accuracy genuine example of crowd wisdom?

What factors tend to influence crowd accuracy?

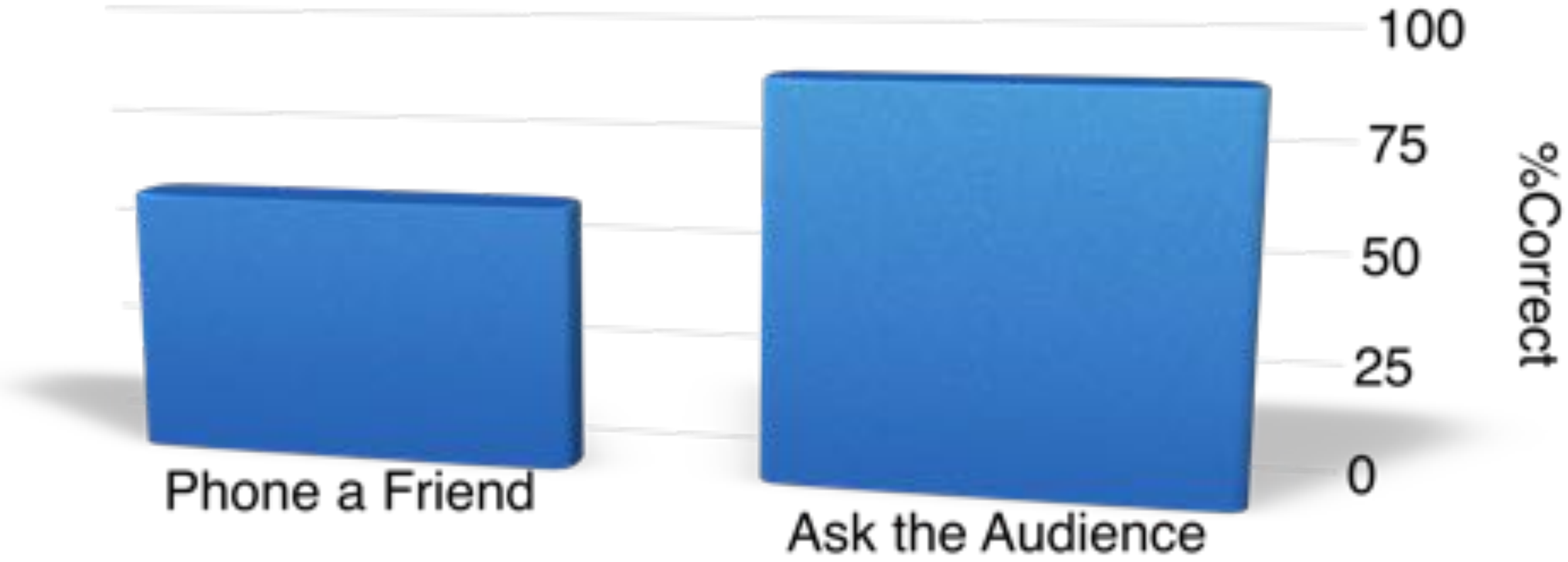
What is more important: a crowd of experts vs a crowd of diverse non-experts? *Expertise vs Diversity?*

Can we rely on crowd wisdom in other situations?


WHO WANTS TO BE A
MILLIONAIRE
WHO WANTS TO BE A

Phone a Friend *vs.* Ask the Audience

Official Game Stats



Why does this work?

Mistakes cancel & correct answers rise to the top…

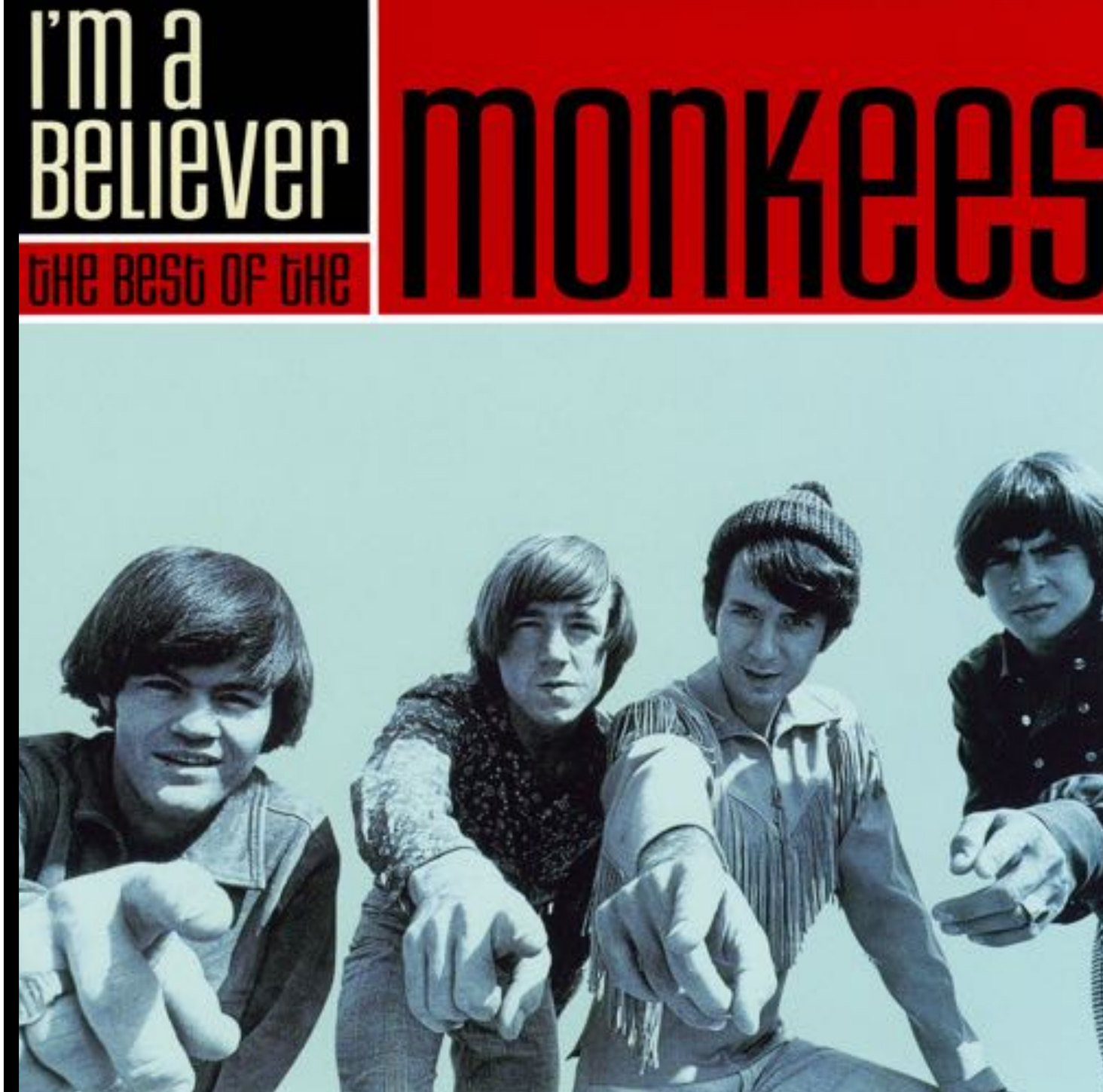
Identify the non-monkee

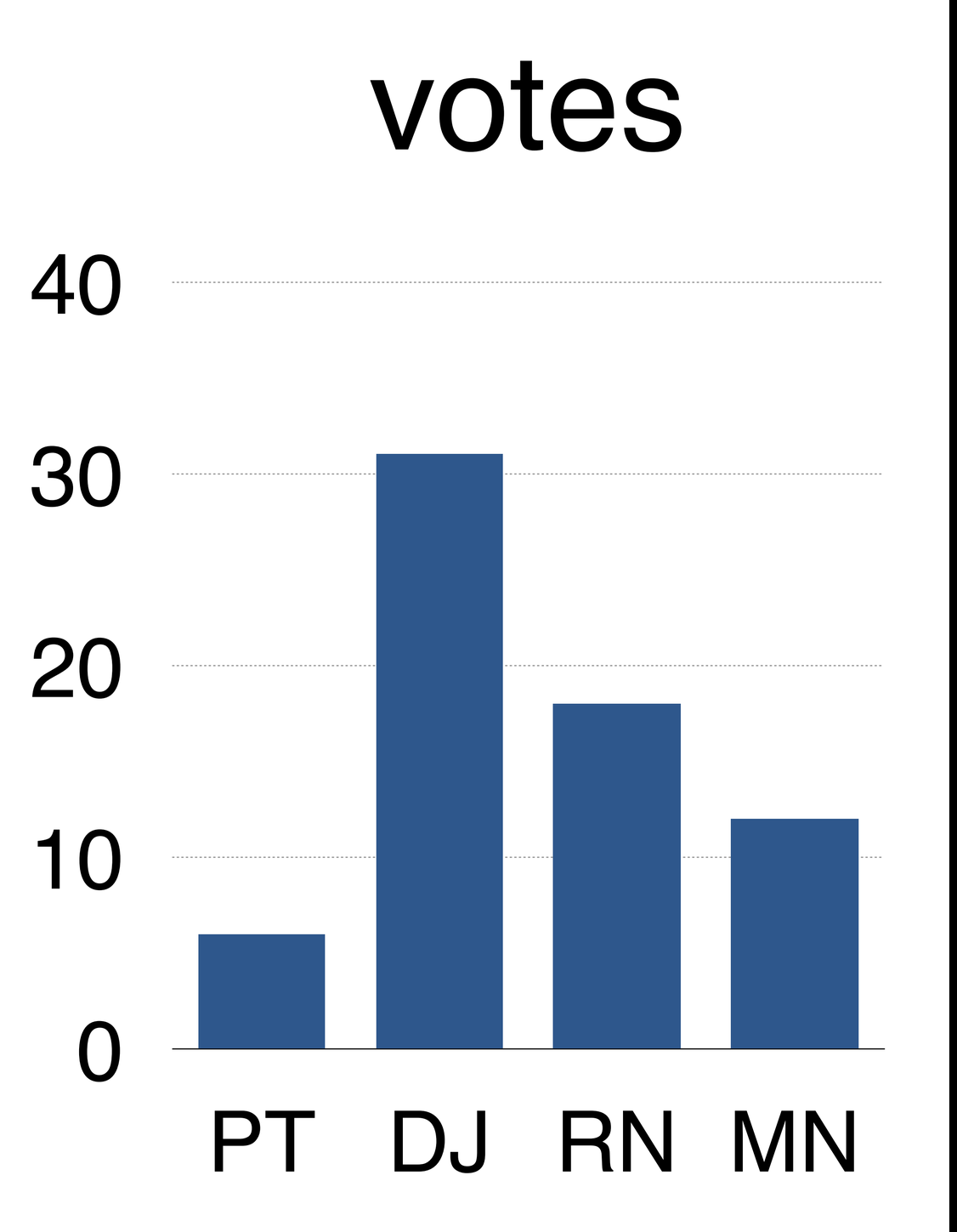
Peter Tork

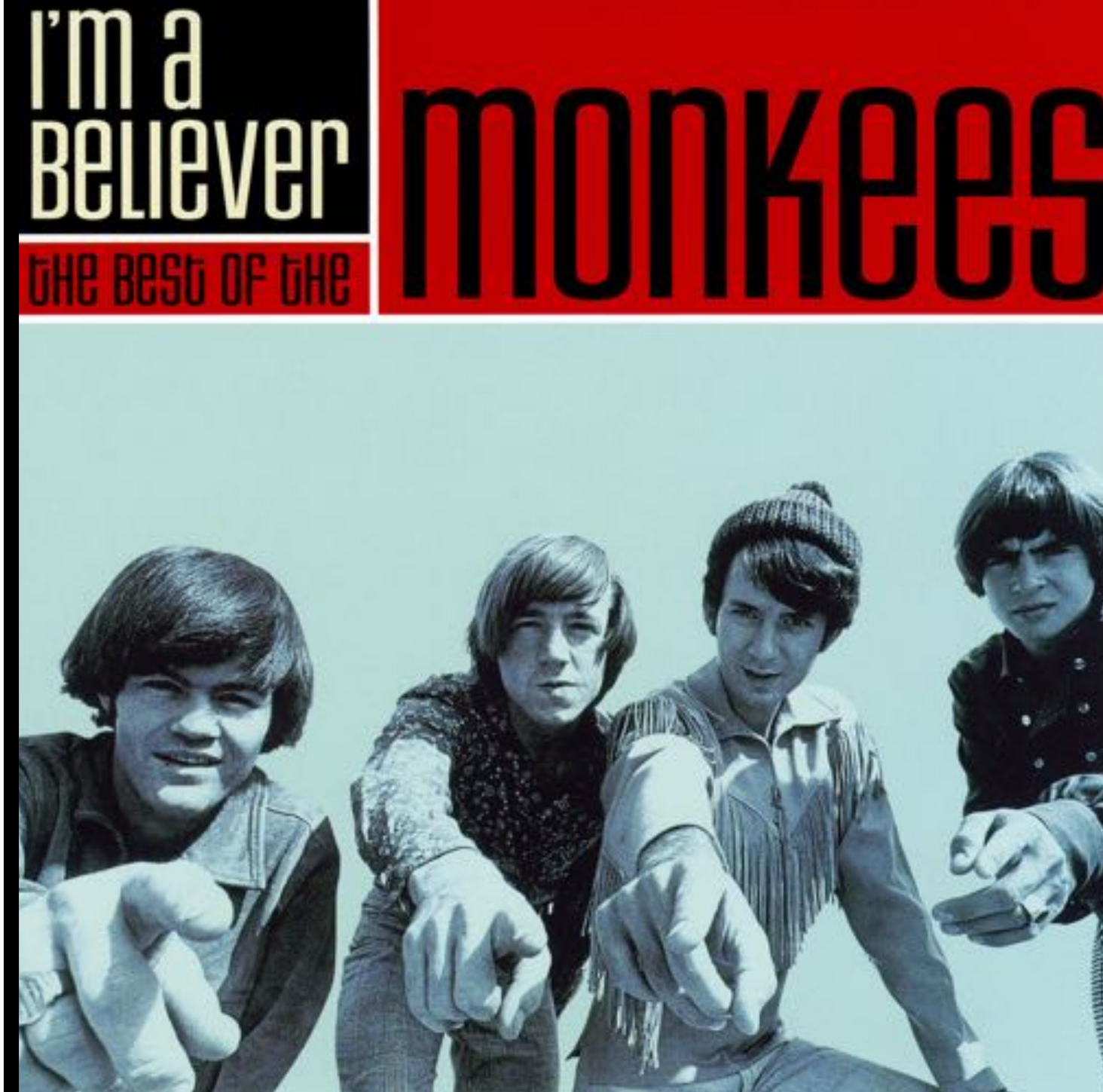
Davy Jones

Roger Noll

Michael Nesmith




votes
40
30
20
10
0
PT
DJ
RN
MN


I'm a Believer
the best of the
monkees

Assume 100 people

Assume 20 know none of the Monkees …

… they select at random so we can expect about 5 votes for each option.

Assume 10 know the Monkees & therefore the non-Monkee…

… therefore RN receives these 10 correct votes.

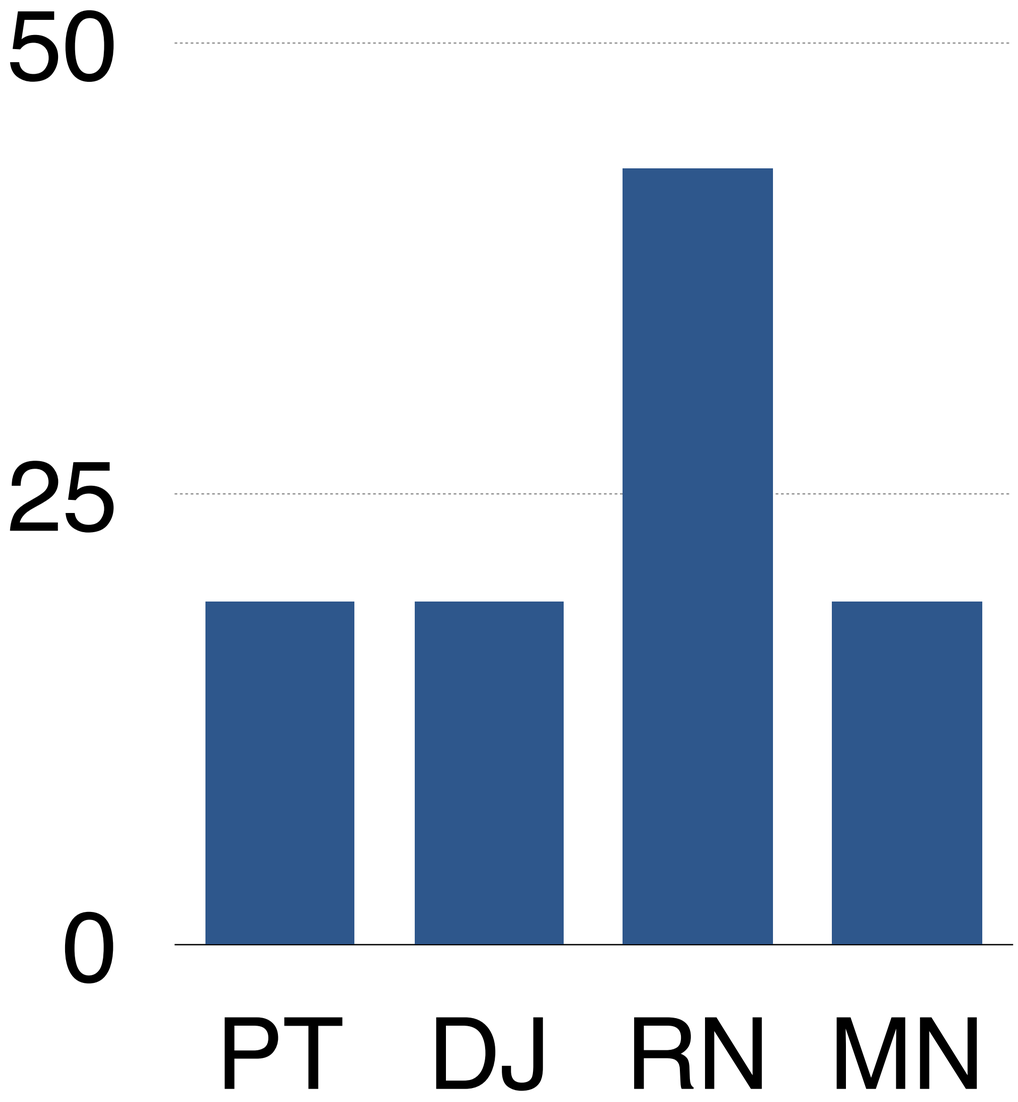
Let’s say 30 know 2 of the Monkees …

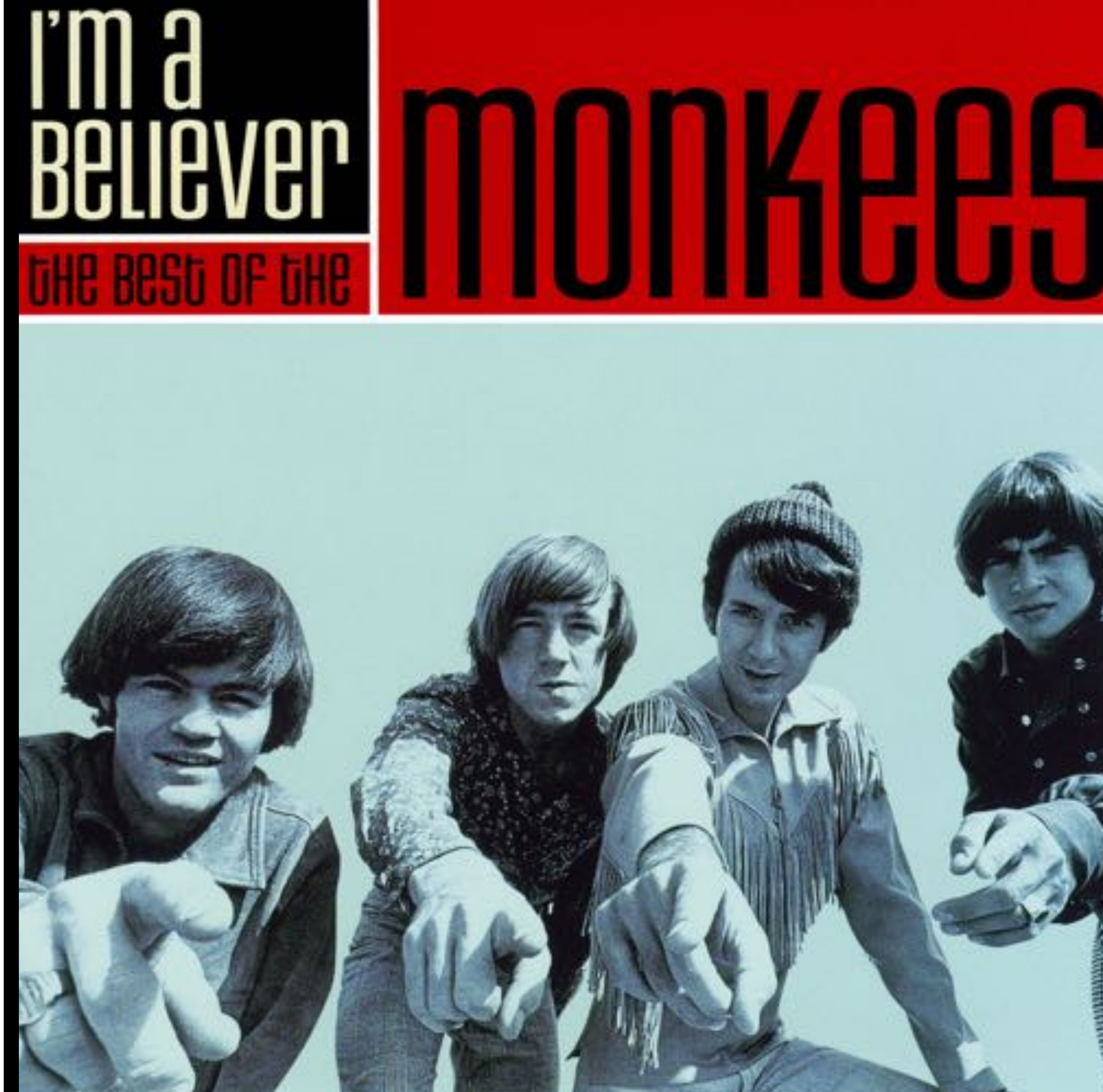
… so their votes are shared between the 2 they don’t know; RN gets 15 of these votes and, all other things being equal, the others get 5 each.

Finally there are 40 who know just one of the Monkees …

… RN gets about 13 (40/3) votes & the others get 9 each (27/3).

votes





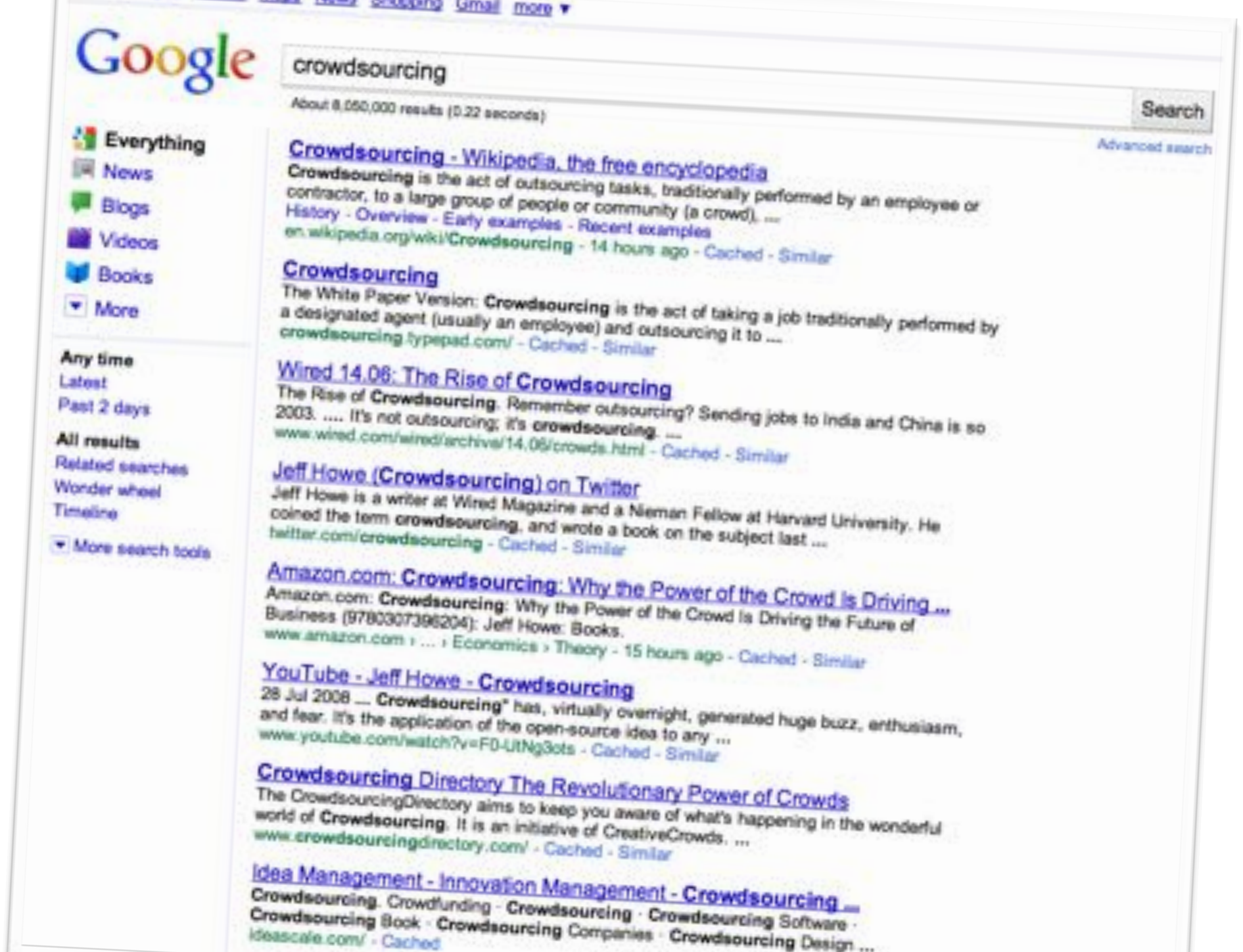
Even if no-one knows the correct answer…

Even if no-one knows the correct answer the crowd prediction may still be correct?

To see this imagine that no-one knows the correct answer but 40 people suspect that it is either RN or one of the others. The other 60 guess at random.

In this case RN will attract 35 of the votes (40/20 + 60/4) whereas the others will only attract about 22 votes (20/3 + 60/4).

Some Scene Setting Examples


Shopping
Gmail
more
Google
crowdsourcing
Search
About 8,060,000 results (0.22 seconds)
Advanced search
Everything
News
Blogs
Videos
Books
More
Any time
Latest
Past 2 days
All results
Related searches
Wonder wheel
Timeline
More search tools
Crowdsourcing - Wikipedia, the free encyclopedia
Crowdsourcing is the act of outsourcing tasks, traditionally performed by an employee or contractor, to a large group of people or community (a crowd). ...
History - Overview - Early examples - Recent examples
en.wikipedia.org/wiki/Crowdsourcing - 14 hours ago - Cached - Similar
Crowdsourcing
The White Paper Version: Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to ...
crowdsourcing.typepad.com/ - Cached - Similar
Wired 14.06: The Rise of Crowdsourcing
The Rise of Crowdsourcing. Remember outsourcing? Sending jobs to India and China is so 2003. .... It's not outsourcing; it's crowdsourcing. ...
www.wired.com/wired/archive/14.06/crowds.html - Cached - Similar
Jeff Howe (Crowdsourcing) on Twitter
Jeff Howe is a writer at Wired Magazine and a Nieman Fellow at Harvard University. He coined the term crowdsourcing, and wrote a book on the subject last ...
twitter.com/crowdsourcing - Cached - Similar
Amazon.com: Crowdsourcing: Why the Power of the Crowd Is Driving ...
Amazon.com: Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business (9780307396204): Jeff Howe: Books.
www.amazon.com › ... › Economics › Theory - 15 hours ago - Cached - Similar
YouTube - Jeff Howe - Crowdsourcing
28 Jul 2008 ... Crowdsourcing" has, virtually overnight, generated huge buzz, enthusiasm, and fear. It's the application of the open-source ideas to any ...
www.youtube.com/watch?v=F0-UtNg3ots - Cached - Similar
Crowdsourcing Directory The Revolutionary Power of Crowds
The CrowdsourcingDirectory aims to keep you aware of what's happening in the wonderful world of Crowdsourcing. It is an initiative of CreativeCrowds. ...
www.crowdsourcingdirectory.com/ - Cached - Similar
Idea Management - Innovation Management - Crowdsourcing ...
Crowdsourcing. Crowdfunding · Crowdsourcing · Crowdsourcing Software · Crowdsourcing Book · Crowdsourcing Companies · Crowdsourcing Design ...
ideascale.com/ - Cached


amazon.co.uk
Hello Barry Smyth. We have recommendations for you. (Not Barry?)
Barry's Amazon.co.uk | Deals of the Week | Gift Cards | Gifts & Wish Lists
Your Account | Help
Basket
Wish List
Shop All Departments
Search All Departments
Your Amazon.co.uk
Page You Made
Recommended For You
Rate These Items
Improve Your Recommendations
Your Profile
Learn More
Barry, Welcome to Your Amazon.co.uk (If you're not Barry Smyth, click here.)
Page 1 of 30
Today's Recommendations For You
Here's a daily sample of items recommended for you. Click here to see all recommendations
What the Dog Saw: and other... (Paperback) by Malcolm Gladwell £5.00
Fix this recommendation
Animal Spirits: How Human P... (Hardcover) by George A. Akerlof £9.51
Fix this recommendation
Free: The Future of a Radical... (Hardcover) by Chris Anderson £10.79
Fix this recommendation
High Performance Web Sites: Ess... (Paperback) by Steve Souders £11.95
Fix this recommendation
The Men Who Stare At Goats [D... DVD ~ Kevin Spacey £7.99
Fix this recommendation
Burn Notice - Season 1 [DV... DVD ~ Jeffrey Donovan £14.95
Fix this recommendation
From Your Wish List
My Tiny Life: Crime and Passion in the Virtual World (Paperback) by Julian Dibbell
House - Season 3 [Hugh Laurie] [DVD] [2006] (DVD) ~ Hugh Laurie
Prison Break - Season 3 - Complete [DVD] [2006] (DVD) ~ Wentworth Miller
Improve Your Recommendations
The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies (New Edition)
Rate this item
Don't use for recommendations
Items you own
Items you've rated
New for You
Sherlock Holmes [DVD] [2009]
The Men Who Stare At Goats [D... DVD ~ Kevin Spacey
The Book Of Eli [DVD] [2009] DVD ~ Ray Stevenson £5.99
Alice in Wonderland [DVD] [2010] DVD ~ Johnny Depp £9.91
Fix this recommendation

News, Images, Videos

Most Recent Top in 24 Hr 7 Days 30 Days 365 Days

161 diggs — digg


Aww Nice Wedding ...Oh Wait, Married Among Sharks?!?
youtube.com — ***** (Submitted by gorgeousasand)
9 Comments Share Bury Made popular 8 min ago

127 diggs — digg
U.K. Porn Star Ranks Irish Men Number One After Sex Tour
thehoorse.com — Irish men come out on top it seems according to Liverpool porn star Tanya Tate. (SFW)
(Submitted by TimBuck1)
13 Comments Share Bury Made popular 22 min ago


490 diggs — digg
TOYOTA
What is Toyota doing to ensure your safety?
toyota.com/safety - Find out how Toyota is enhancing the safety and technology of its vehicles.
Share Bury

173 diggs — digg
The New Segregation Debate: Single-Sex Classrooms
newsweek.com — Single-sex classes have increased by 4,000 percent in less than a decade. Can educating girls and boys separately fix our public schools, or does it reinforce outmoded gender stereotypes? (Submitted by LBUCHA)
12 Comments Share Bury Made popular 22 min ago

82 diggs — digg
WSJ.com - Americans Lose Confidence in U.S., Obama
online.wsj.com — Americans are more pessimistic about the state of the country and less confident in Obama's leadership than at any point since he entered the White House, according to a new WSJ/NBC News poll.
(Submitted by rw2y)
62 Comments Share Bury Made popular 51 min ago


219 diggs — digg
Internet Explorer 9 destroys Chrome 6 in speed test (Vid)
downloadsquad.com — IE9 is some orders of magnitude faster than H7ML5 when it comes to hardware-accelerated canvas rendering. In some other initial benchmarks, see video IE9 is about 30% slower than Chrome 6 in the SunSpider JavaScript benchmark -- and about 10% faster than Firefox 3.7.
54 Comments Share Bury Made popular 1 hr ago


Best of Digg in your Email! Daily Weekly
Your email... Subscribe

Top in All Topics all news videos images

1586 YouTube wins case against Viacom!

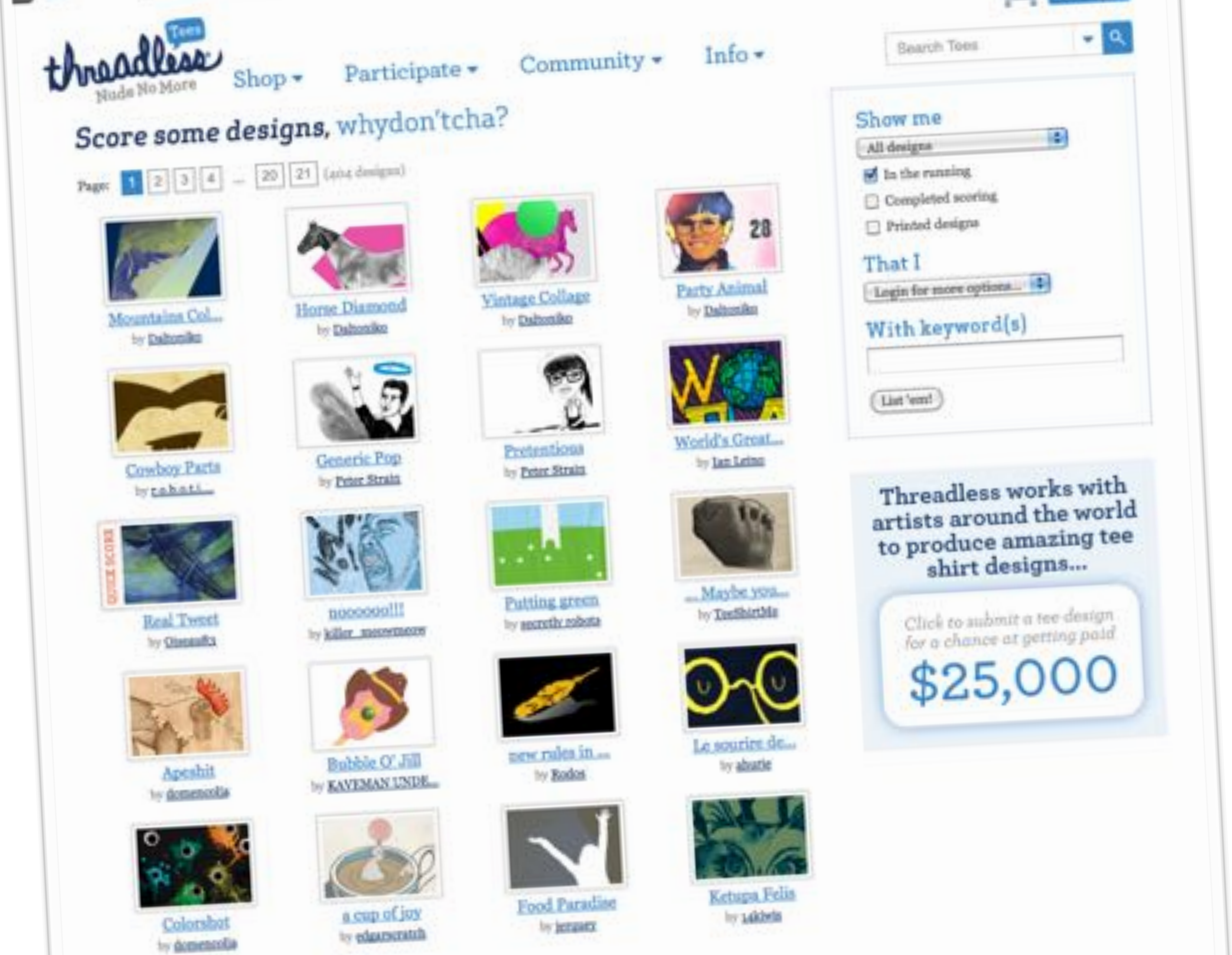
1359  Goddamn Vuvuzela [GIF]

908  The 20 Worst Names Real People Have (With Proof)

880 BREAKING NEWS: Toronto, Ottawa, Montreal Report Earthquake

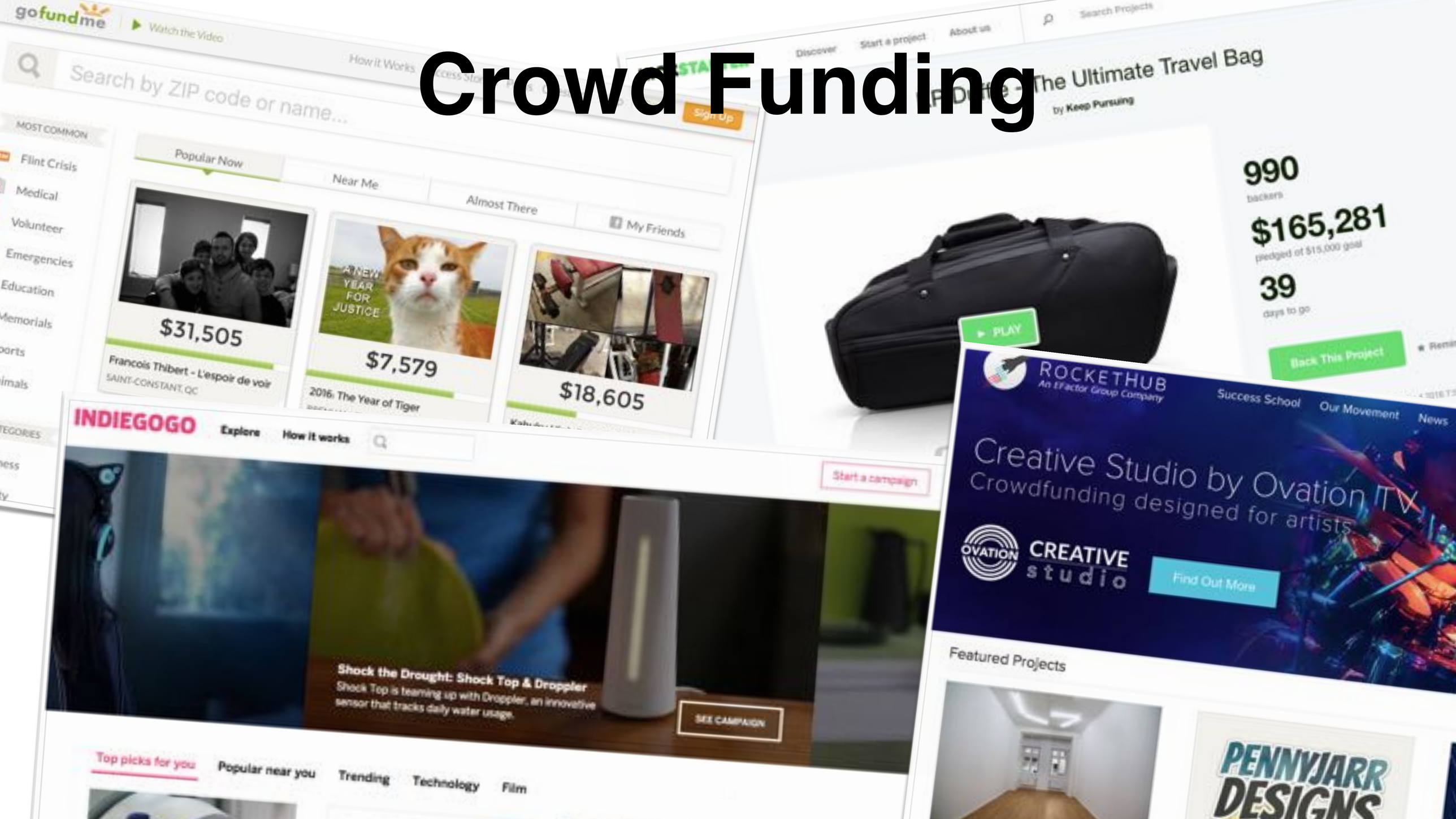
864 War on drugs worse than drugs

790  Monster Crab Exits Shell (Time-Lapse Video)


Tees
threadless
Nude No More
Shop
Participate
Community
Info
Search Tees
Score some designs, whydon'tcha?
Page: 1 2 3 4 ... 20 21
Mountains Col...
by Daltoniko
Horse Diamond
by Daltoniko
Vintage Collage
by Daltoniko
Party Animal
by Daltoniko
Cowboy Parts
Generic Pop
by Peter Strain
Pretentious
by Peter Strain
World's Great...
Real Tweet
noooooo!!!
Putting green
...Maybe you...
Apeshit
Bubble O' Jill
new rules in...
Le sourire de...
Colorshot
a cup of joy
Food Paradise
Ketupa Felis
Show me
All designs
In the running
Completed scoring
Printed designs
That I
Login for more options...
With keyword(s)
List 'em!
Threadless works with artists around the world to produce amazing tee shirt designs...
Click to submit a tee design for a chance at getting paid
$25,000


INNOCENTIVE
WHERE THE WORLD INNOVATES
Home
Products
Seekers
Solvers
Challenge Center
My InnoCentive
About Us
News & Events
Blog
Help
Contact Us
Register
Login
WHAT IS INNOCENTIVE?
InnoCentive harnesses collective brainpower around the world to solve problems that really matter.
Learn more
OPEN CHALLENGES
Show Newest Challenges
View all Challenges
$40,000 USD
34 active solvers
Deadline - Aug 22, 2010
Advanced Material Coatings for High Strength, High Conductivity Substrates
$10,000 USD
18 active solvers
Deadline - Aug 21, 2010
Seeking N-substituted 2-imino-1,3-dithietanes
Varies
17 active solvers
Deadline - Aug 18, 2010
Sustainable Packaging Materials for the Developing World
WINNING SOLVERS
View all Winning Solvers
View Challenge
Trevor Rose
Read Bio
View Challenge
John Davis
Read Bio
NETWORK FACTS
CA US State with most wins
81% Solvers w/advanced degrees
1044 Total Challenges posted
SOLVER NETWORK
View Map in detail
FEATURED CHALLENGE
CHALLENGE CENTER
Emergency Response 2.0 : Solutions to Respond to Oil Spill in the Gulf of Mexico
Recently, an explosion on an offshore oil platform in the Gulf of Mexico caused both loss of life and a sizable and ongoing oil spill. We are asking Solvers worldwide to respond quickly with ideas and approaches to react to this very serious environmental threat.
Can you make a difference? Yes. InnoCentive's work with the Oil Spi...
Learn More
Reward: See details
Type: Ideation
Deadline - Jun 30, 2010
INNOVATION PARTNERS
The Economist
nature.com
FOLLOW INNOCENTIVE
twitter
Join the conversation! Follow InnoCentive.
Visit Twitter
Seeking Solutions?
Find solutions for your
Solvers Wanted!

Crowd Funding



Predicting/Tracking Influenza

Google Flu Trends

Geo-coded search terms as indicators of human activity ...

- flu remedy
- cure flu
- flu shot, etc.

Strong correlation between flu-related searches and illness.

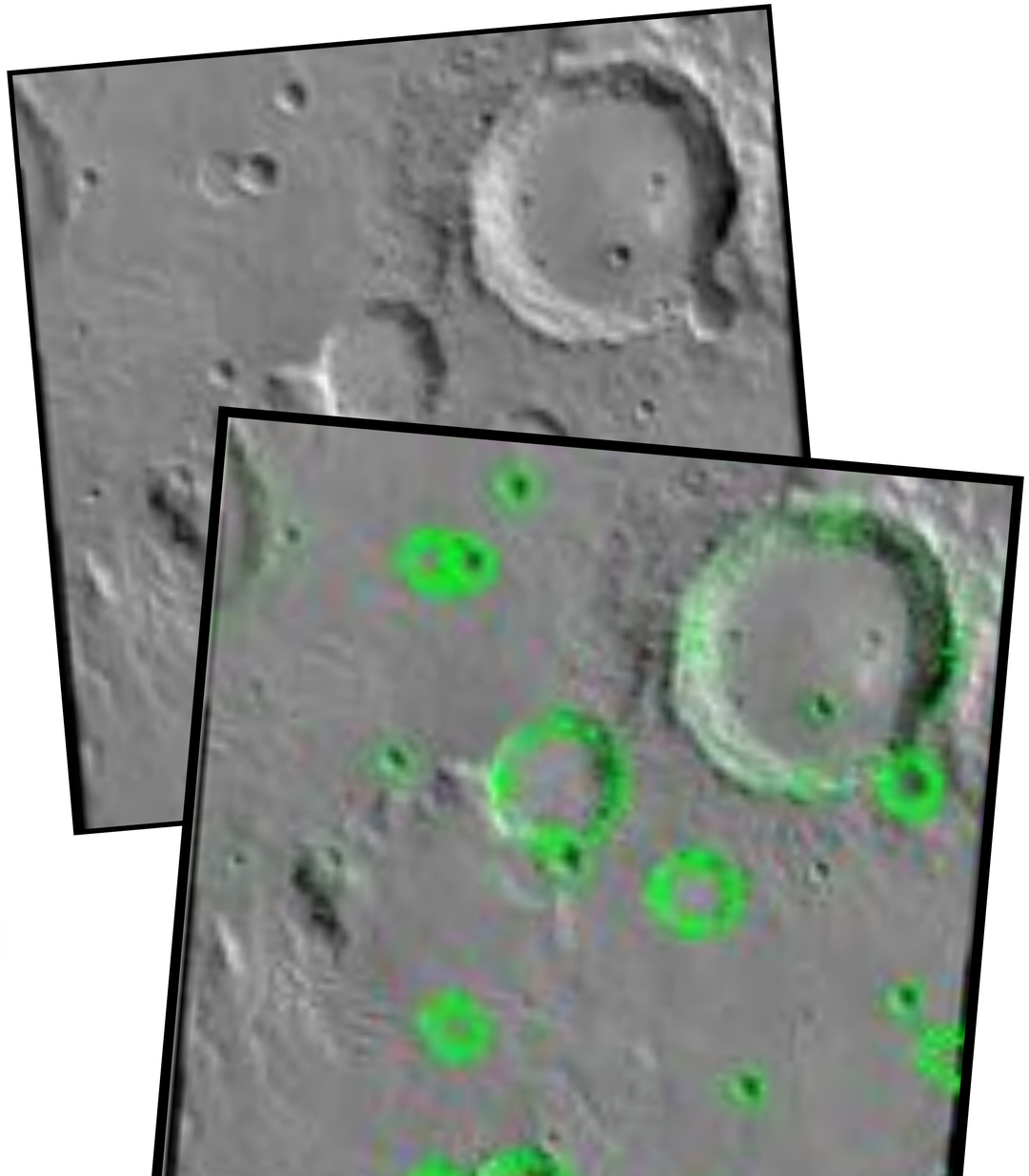
NASA’s Clickworkers

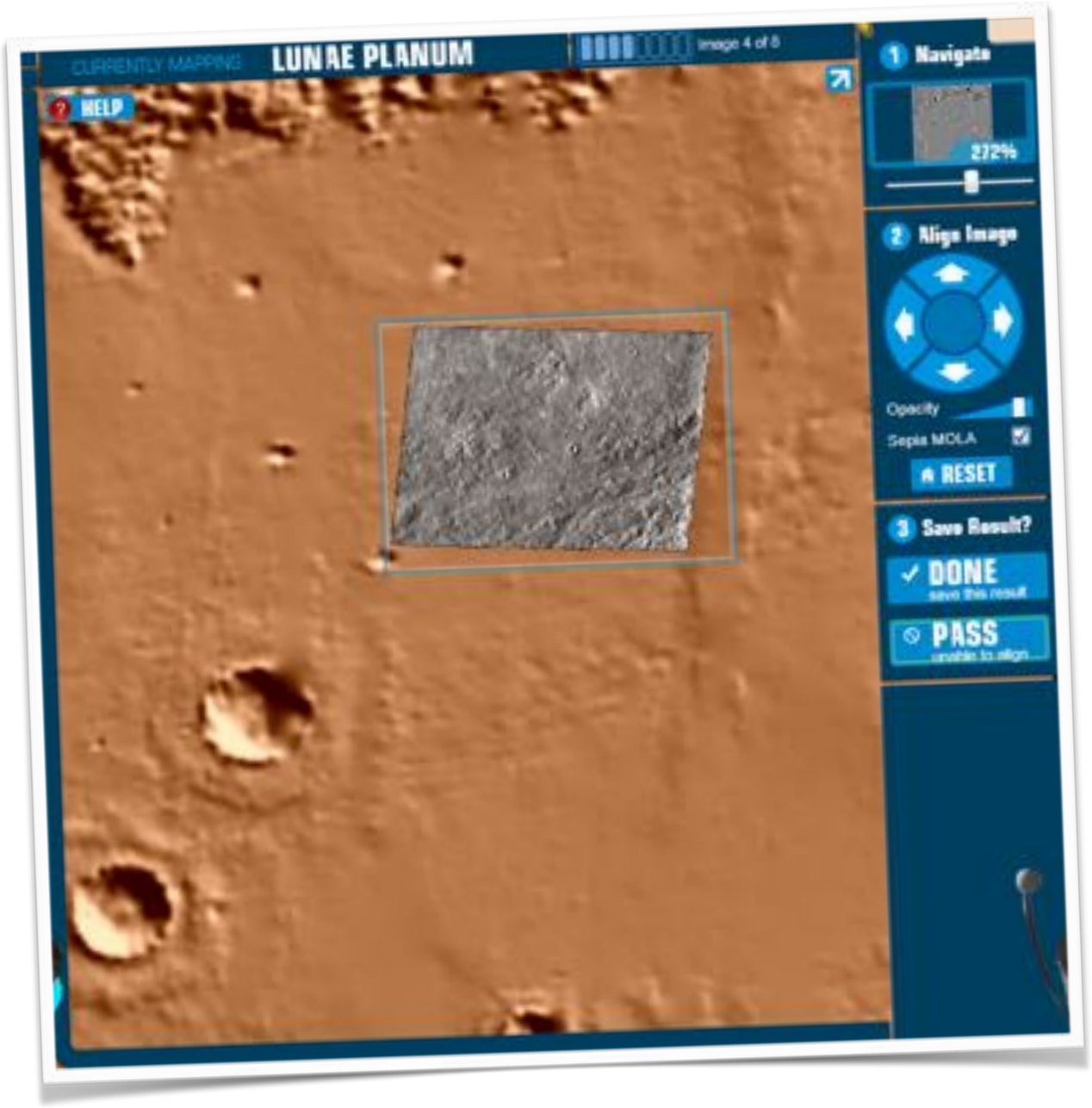
The life of a planetary geologist at NASA

identifying and measuring geological landforms (craters, ridges, valleys) from satellite imagery.
Tedious, error-prone, labour-intensive (80k landforms ≈ 2 person-years)

The Clickworkers Experiment

NASA put the entire Viking-Mars image database online and invited amateur astronomers to perform the same analysis task online.
Individual contributions are aggregated.
Within a month the entire DB was completed to a comparable degree of accuracy by a few thousand contributors ... 37% were one-time contributors!




CURRENTLY MAPPING
LUNAE PLANUM
image 4 of 5
HELP
1 Navigate
272%
2 Align Image
Opacity
Sepia MOLA
RESET
3 Save Result?
DONE
save this result
PASS
unable to align

PolyMath

Mathematician Tim Bower's Blog Post…

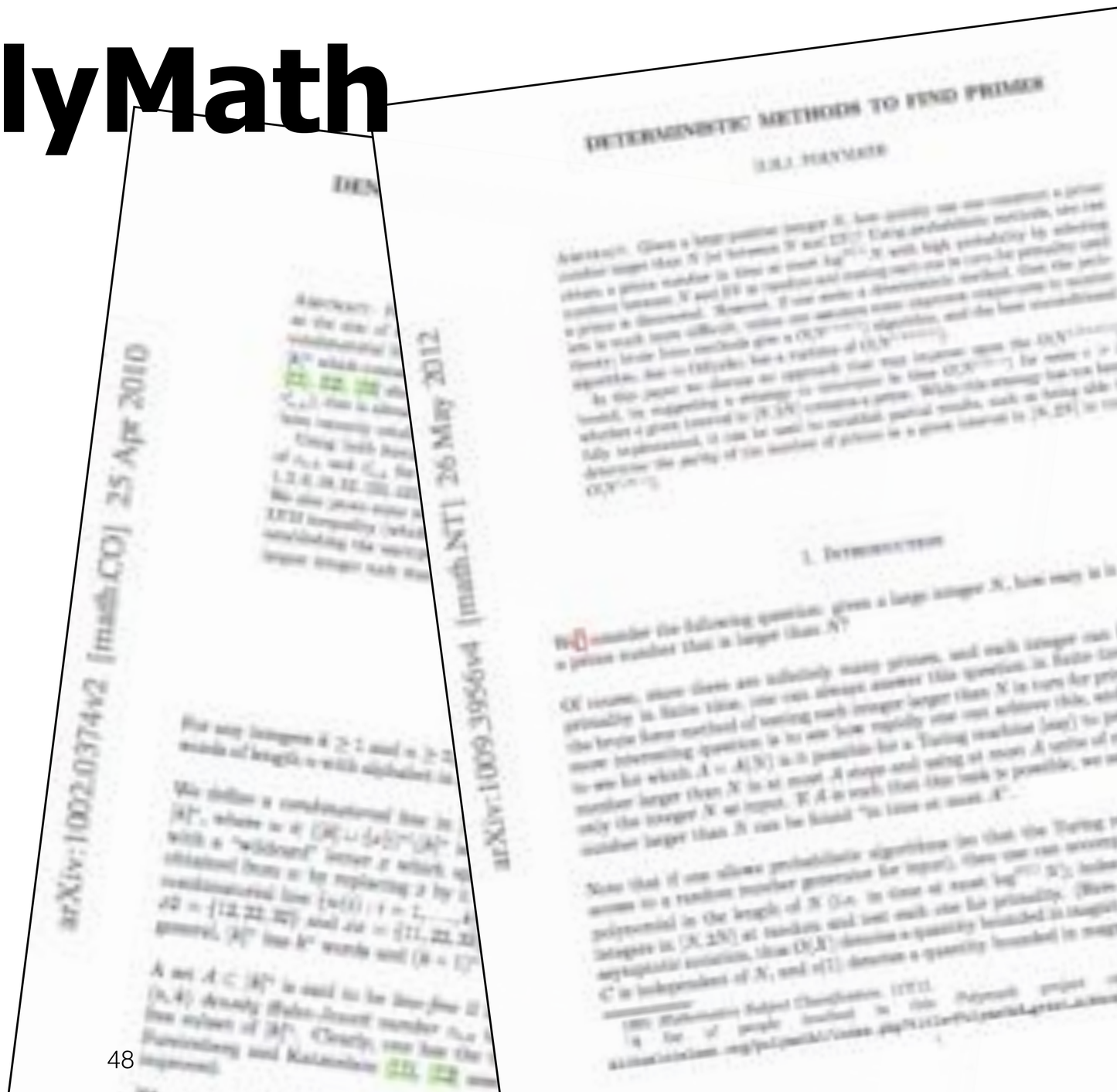
Find a new combinatorial proof to the density version of the Hales–Jewett theorem

A Social Maths Experiment

Using blogs and wikis to coordinate and amplify

Solution Success

After 7 weeks: problem was solved and involved the contributions of >40 people. Polymath 2 -8 spawned.



Kasparov vs the World

Gary Kasparov
World #1 since 1985.

The World
Players from around the world voted on moves; some strong players but far below GK. 50k people voted during the game.

The Game
Complex 4-month, 62-move game. Kasparov eventually won, with supreme effort. The World played a game a t a level far greater than any of its individual players.

Amplifying Micro-Expertise
The move of Irina Krush…


Framing
Collective
Intelligence

Why/When are the Many Smarter than the Few?

Private Information

People need to be acting on their own private information.

Diversity of Opinion

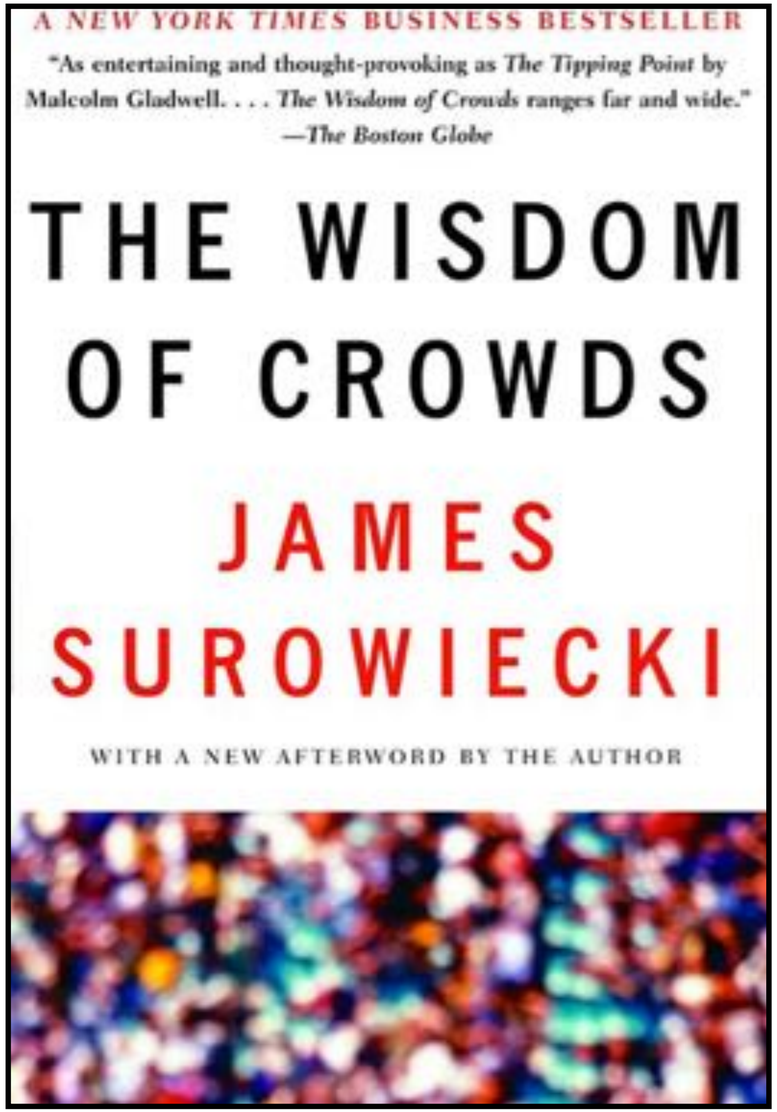
Differing opinions based on private/local information matter.

Independence of Opinion

Opinions forms independently of others matter.

Aggregation Mechanism

There needs to be a mechanism for turning private judgements into a collective decision.



What is Collective Intelligence?

What is Collective Intelligence?

"… machines mimicking tasks that humans are good at"

Artificial Intelligence

"… machines solving tasks that humans are <u>not</u> good at"

Data-Driven Intelligence

"amplying human intelligence by … combining machine & human intelligence"

Collective Intelligence

What is Collective Intelligence?

"… machines mimicking tasks that humans are good at"

Artificial Intelligence

machine learning
planning
expert systems
perception & robotics

"… machines solving tasks that humans are not good at"

Data-Driven Intelligence

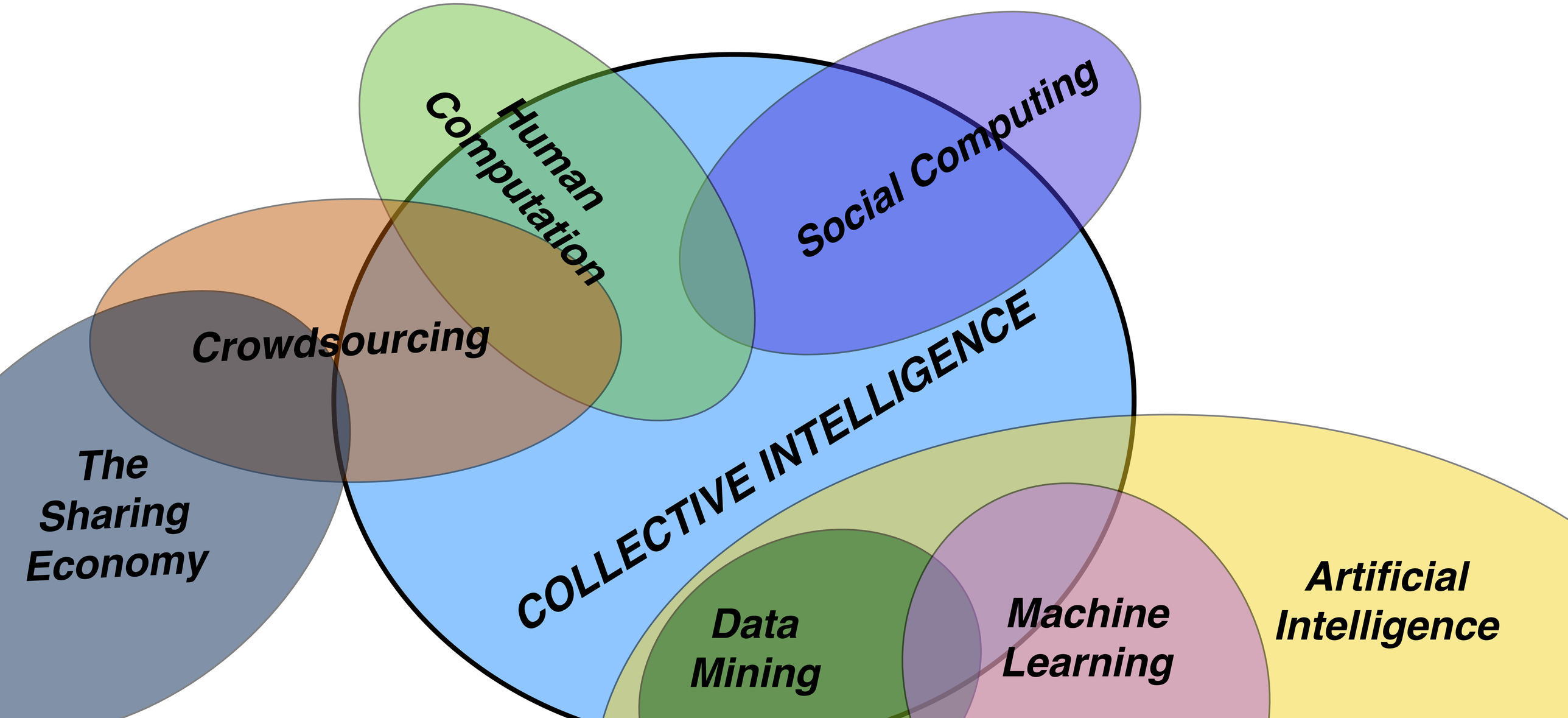
data mining
data analytics
big data
the semantic web
linked data

"amplying human intelligence by … combining machine & human intelligence"

Collective Intelligence

crowdsourcing
human computation
social computing
collaboration

The Collective Intelligence Landscape



How is this changing the way that we think about computation?

Machine vs Human?

Towards a New Computational Paradigm

We are living in an increasingly connected world.

Where ubiquitous computation is available at near-zero cost.

People are inherently social and collaborative.
Collectively our fragmented contributions add to a lot.

Some problems are better suited to machines …

… while others require human intelligence.

A New Computational Paradigm

Picking the right problem.

What type of problems are suited to a collective intelligence approach?

Motivating and incentivising the crowd.

What makes for a suitable crowd and how do we attract/motivate them?

Amplifying the wisdom of the crowd.

How can we guide and amplify the wisdom of the crowd.

Ensuring correctness.

Can correctness be guaranteed?

Part 2 - CaPTCHAs

Are you Human?

Alan Turing (1912 - 1954)

Mathematician, Cryptanalyst, Computer Scientist, Bio-Informatician,...

Bletchley Park, Breaking the Enigma Machine (the Turing-Welcheman Bombe)

The Turing Machine

Abstract, tape-based computing formalism that laid key elements of the groundwork for the theoretical underpinnings of modern computer science.

A Founding Father of AI

Turing's Chess playing 'programme'

The Turing Test (see also Searle's Chinese Room)

... Morphogenesis and other oddities ...

Turing, A.M. (1950). Computing machinery and intelligence. Mind, 59, 433-460.

The Turing Test

Ok ... but isn’t the Turing Test just an interesting philosophical problem?

What’s the practical application?

When might it be useful to know whether you are dealing with a human or a machine?

Turing Test Applications?

Authenticating Human Voters in Online Polls

Preventing Comment Spam in Blogs and other Social Media

Protecting Website/Email Registrations

Protecting Plaintext Email Addresses from Scrapers

Blocking Search Engine Spiders/Bots from Restricted Site Regions

Preventing Dictionary Attacks at Login Time

Authenticating members of my Social Graph (e.g. eliminating Spam followers)

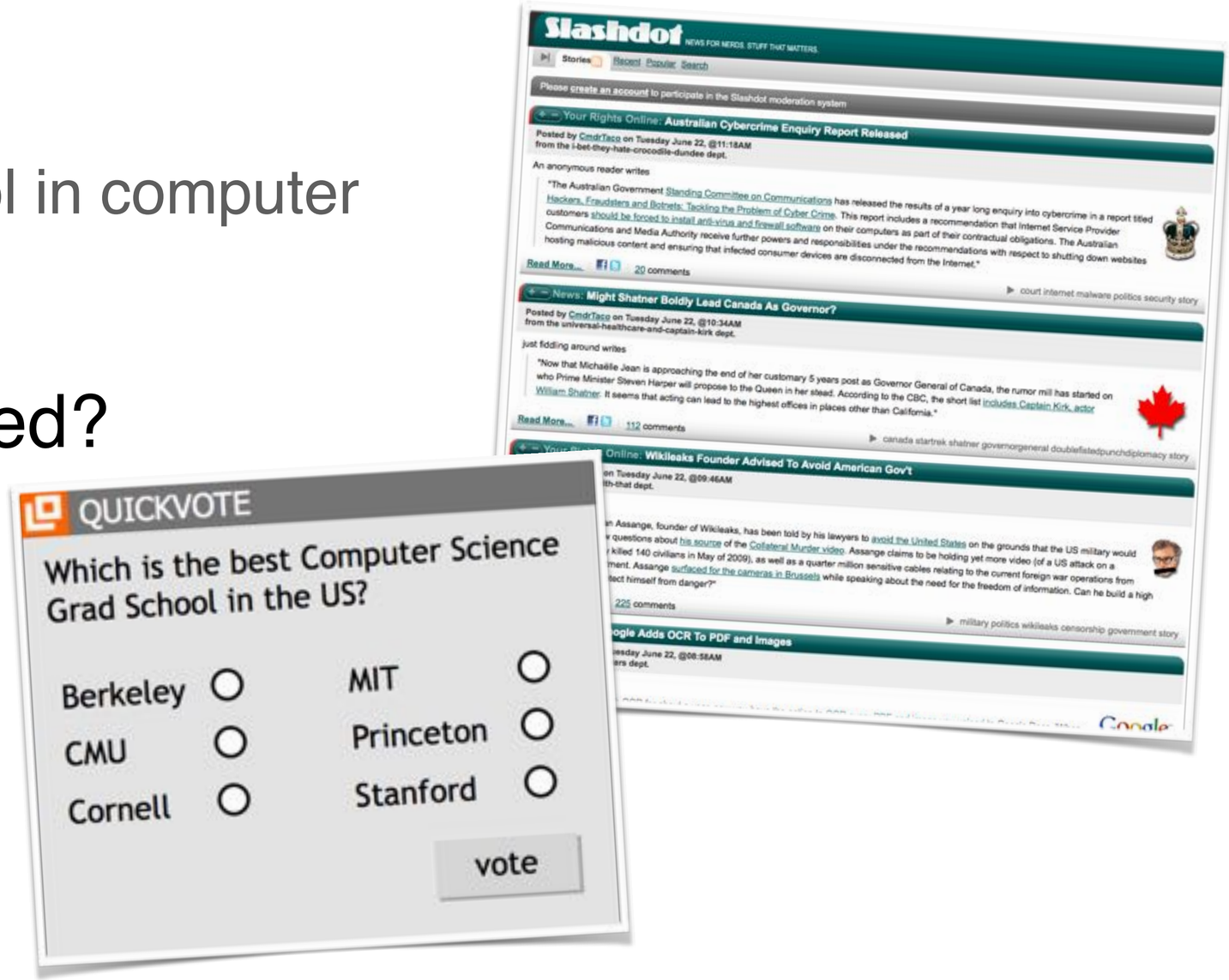
Identifying Internet ChatBots (see also http://www.loebner.net/Prizef/loebner-prize.html)

Authenticating Votes in Online Polls

1999, Slashdot Poll

What's the best graduate school in computer science?

What do you think happened?

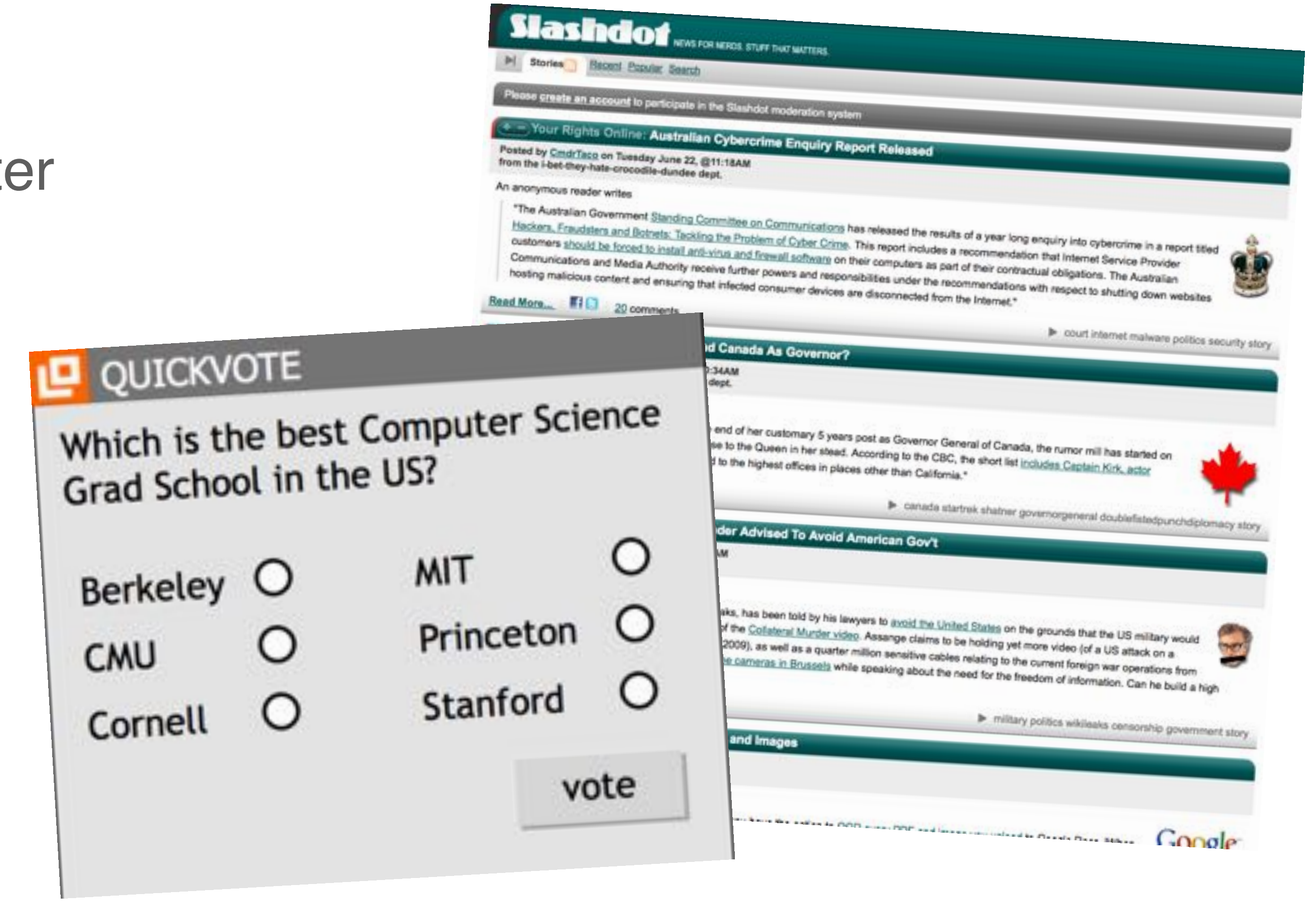


Authenticating Votes in Online Polls

1999, Slashdot Poll

What's the best graduate school in computer science?

What do you think happened?



CMU & MIT Poll Bots

Distributed programmes across multiple machines to stuff the ballots

Standard IP verification insufficient.

End Result ...

MIT @ 21,156 votes vs CMU @ 21,032 votes ... everyone else with <1000 votes.

Which is the best graduate school?

You've Got (Junk) Mail...

In 2000 Yahoo! had a BIG problem .... Email Spam!

Spammers love free email accounts, especially ones that come from trusted brands like Yahoo!, Google, Microsoft etc.

Around 2000 it was easy and free to write simple programmes that would automatically register millions of new email accounts, which could then be used to send spam around the world.

Yahoo, as one of the leading web-based email providers, needed a solution ...

... in short they needed some way to determine when an account was being created by a human versus a machine.

How Lazy Cryptographers do AI

Need an Automated Turing Test

A test to distinguish humans from computers ...

... one that is easy for humans to pass but difficult for machines ...

... but which can be graded by a computer.

A seeming paradox?

See: Luis von Ahn, Manuel Blum and John Langford, *How Lazy Cryptographers do AI*. Communications of the ACM, Feb. 2004

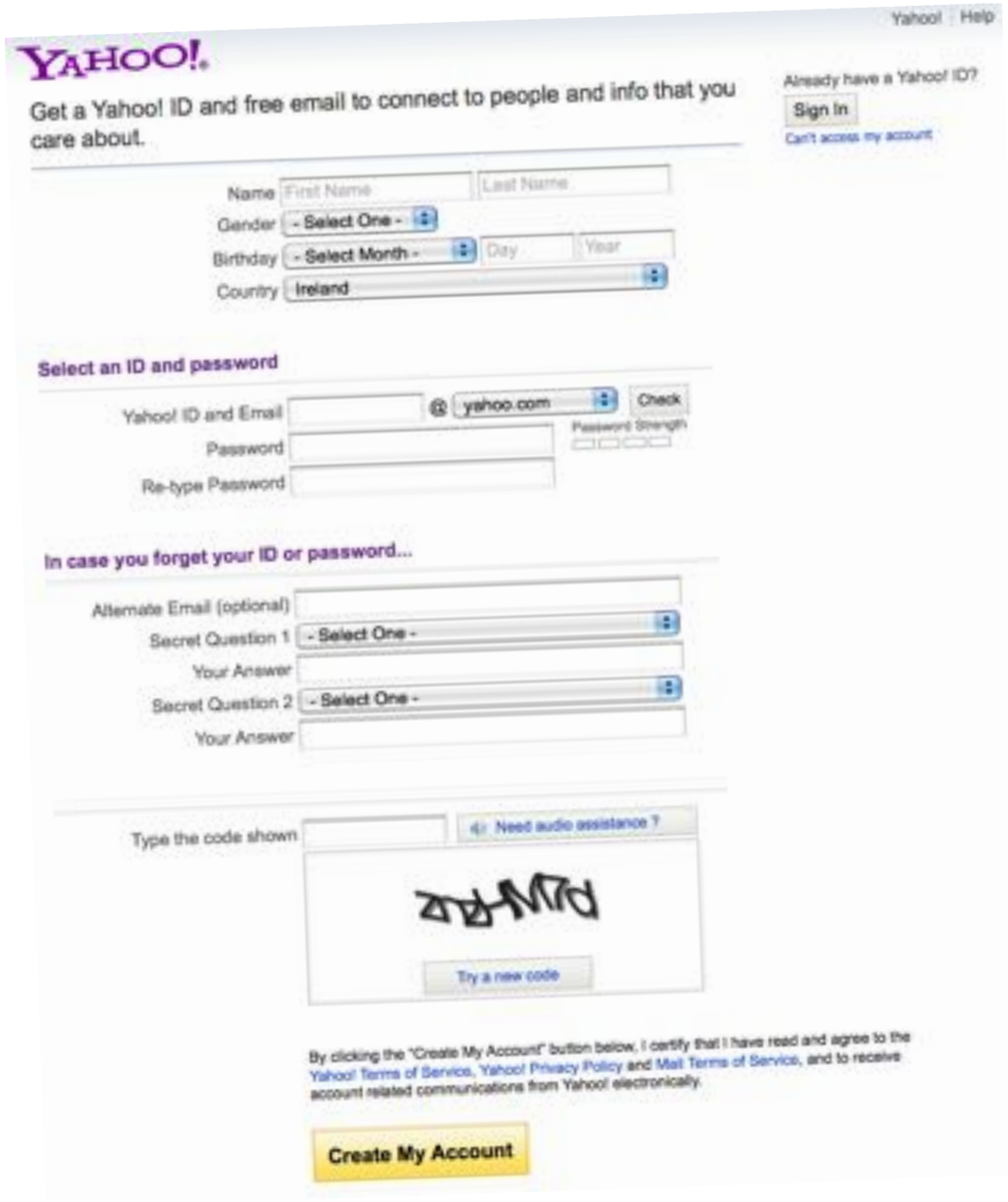
Luis Von Ahn

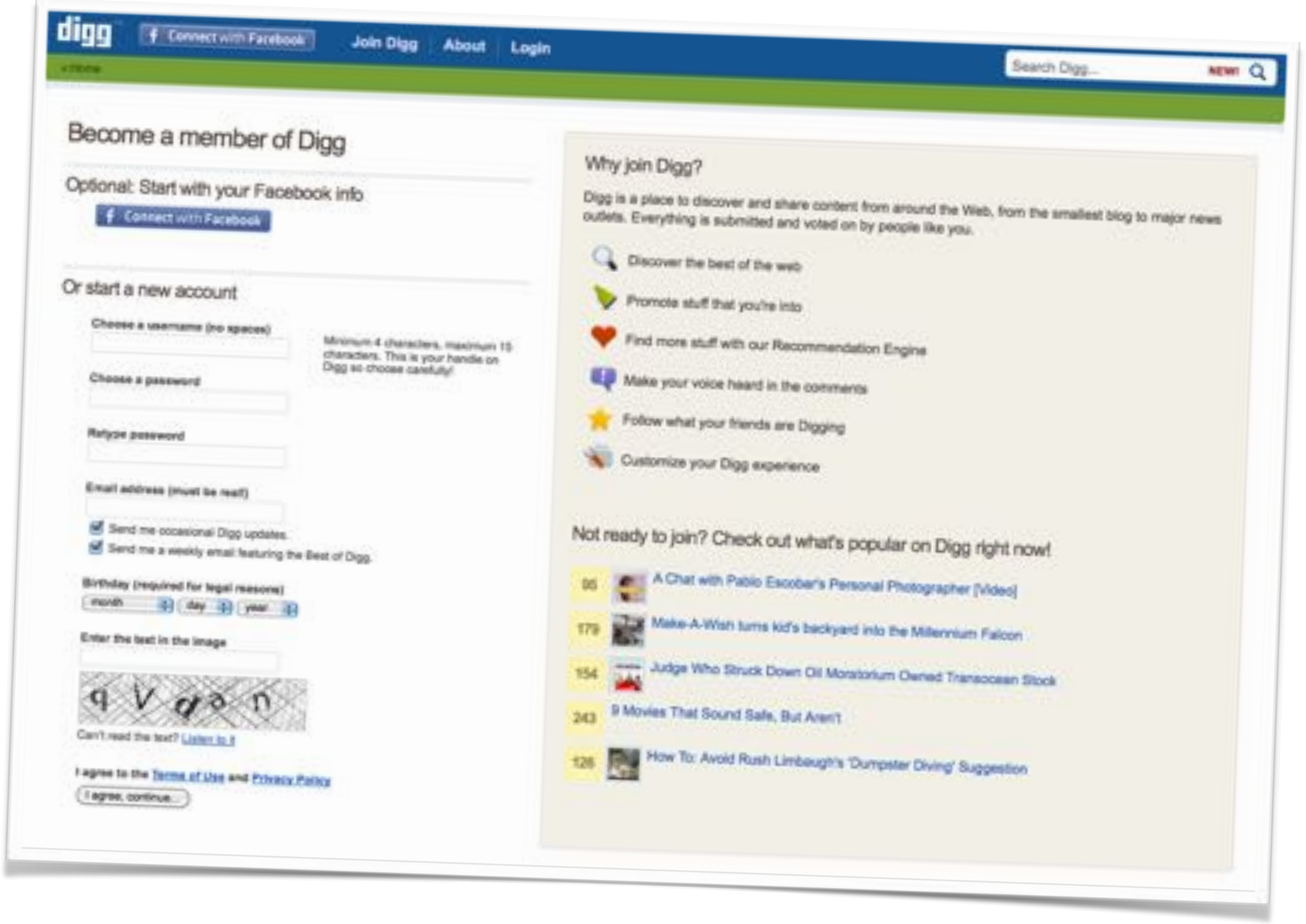
Manuel Blum

What types of things/problems/tests are humans good at but machines are poor at?

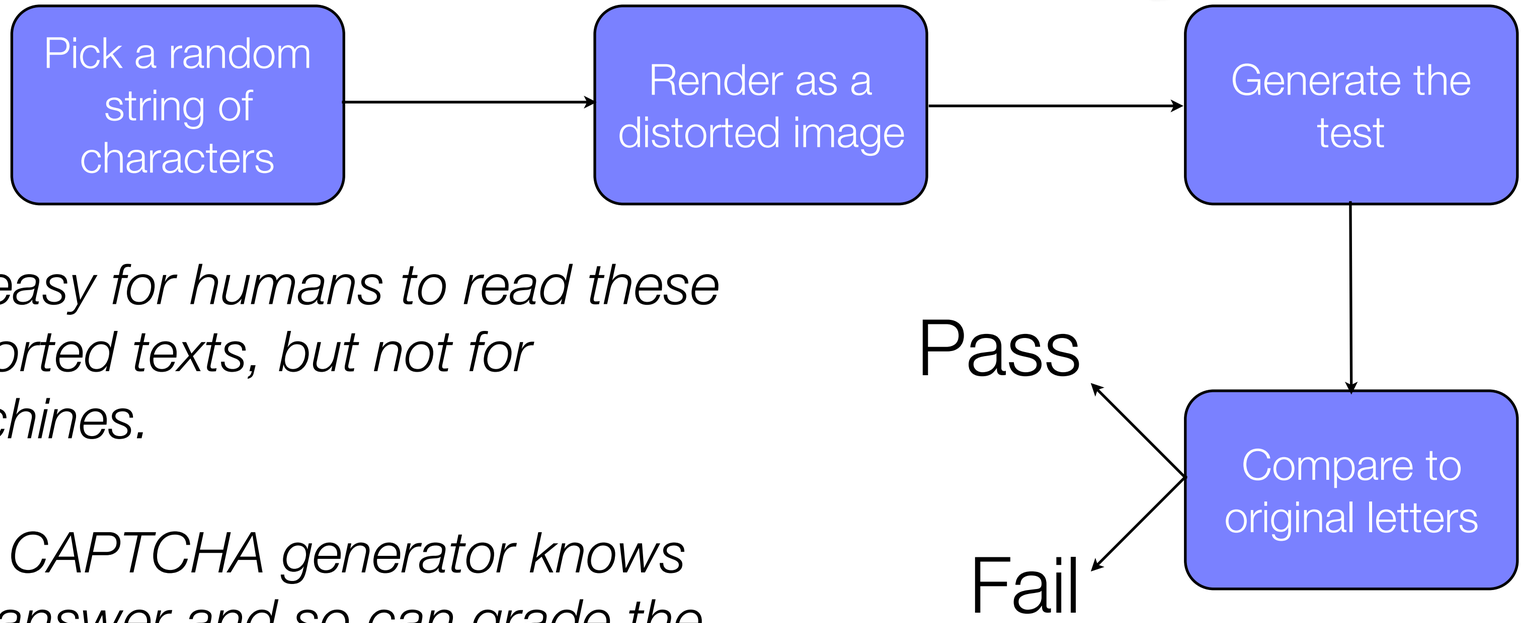
(Completely Automated Public Turing Test to tell Computers and Humans Apart)

Luis von Ahn, Manuel Blum, John Langford: Telling humans and computers apart automatically. Commun. ACM 47(2): 56-60 (2004)


Yahoo! Help
YAHOO!
Get a Yahoo! ID and free email to connect to people and info that you care about.
Already have a Yahoo! ID?
Sign In
Can't access my account
Name First Name Last Name
Gender - Select One -
Birthday - Select Month - Day Year
Country Ireland
Select an ID and password
Yahoo! ID and Email @ yahoo.com Check
Password Strength
Password
Re-type Password
In case you forget your ID or password...
Alternate Email (optional)
Secret Question 1 - Select One -
Your Answer
Secret Question 2 - Select One -
Your Answer
Type the code shown
Need audio assistance ?
Try a new code
By clicking the "Create My Account" button below, I certify that I have read and agree to the Yahoo! Terms of Service, Yahoo! Privacy Policy and Mail Terms of Service, and to receive account related communications from Yahoo! electronically.
Create My Account



CAPTCHAs on Digg

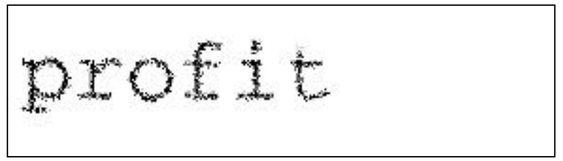


It's easy for humans to read these distorted texts, but not for machines.

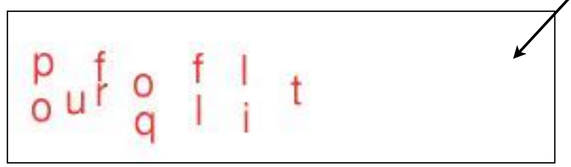
The CAPTCHA generator knows the answer and so can grade the solution – even though it couldn't pass the test itself!

Creating a CAPTCHA

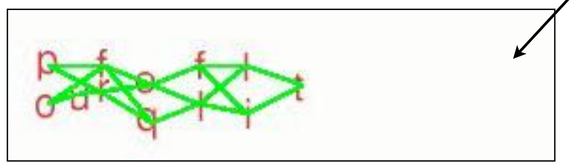
Breaking a CAPTCHA



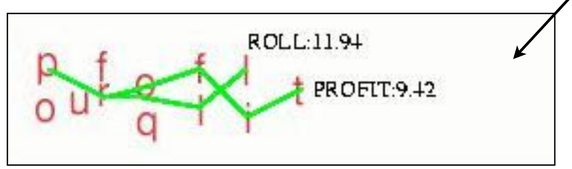
Hypothesize candidate letters. Use shape matching techniques to compare randomly selected points against known letter templates.



Identify consistent letters. Analyze pairs of letters to see whether they can be used consecutively to form a word. Green lines imply consistent letter sequences.



Rank plausible words. Score and rank consistent letter sequences according to their word-plausibility (and matching strength). Select top-ranking word as solution.



80%+ success rate at breaking (EZ-GIMP) CAPTCHAs.

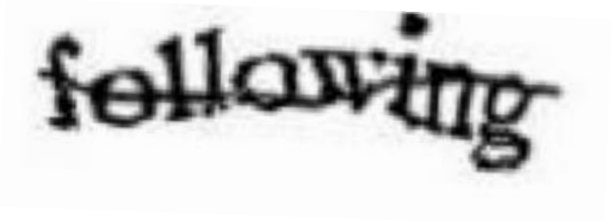
Greg Mori, Jitendra Malik: Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA. CVPR (1) 2003: 134-144

Improved OCR-based CAPTCHAs

Improved contrast for human readability.

Limited per-character distortions.

Character-level font randomization.



Limited background noise.

Focus on frustrating image segmentation techniques

E.g. Adding an angled cross-cutting line frustrates character segmentation.

Squashing individual symbols together maximizes connected components to frustrate character segmentation.

CAPTCHA Variations

Types of CAPTCHA

OCR-based CAPTCHAs (based on the difficulty of reading distorted text)

Random-letter strings, simple distortions.
Single-word & multi-word strings. E.g. GIMPY picks 7 words from a dictionary and asks the user to type any 3 of their distorted images; FYI, EZ-GIMPY relies on a single word.
Animated GIFs, ASCII art, 3D-reliefs etc.
Non-word-based CAPTCHAs. E.g. solving arithmetic problems.

Pattern-Recognition CAPTCHAs (based visual pattern recognition problems)

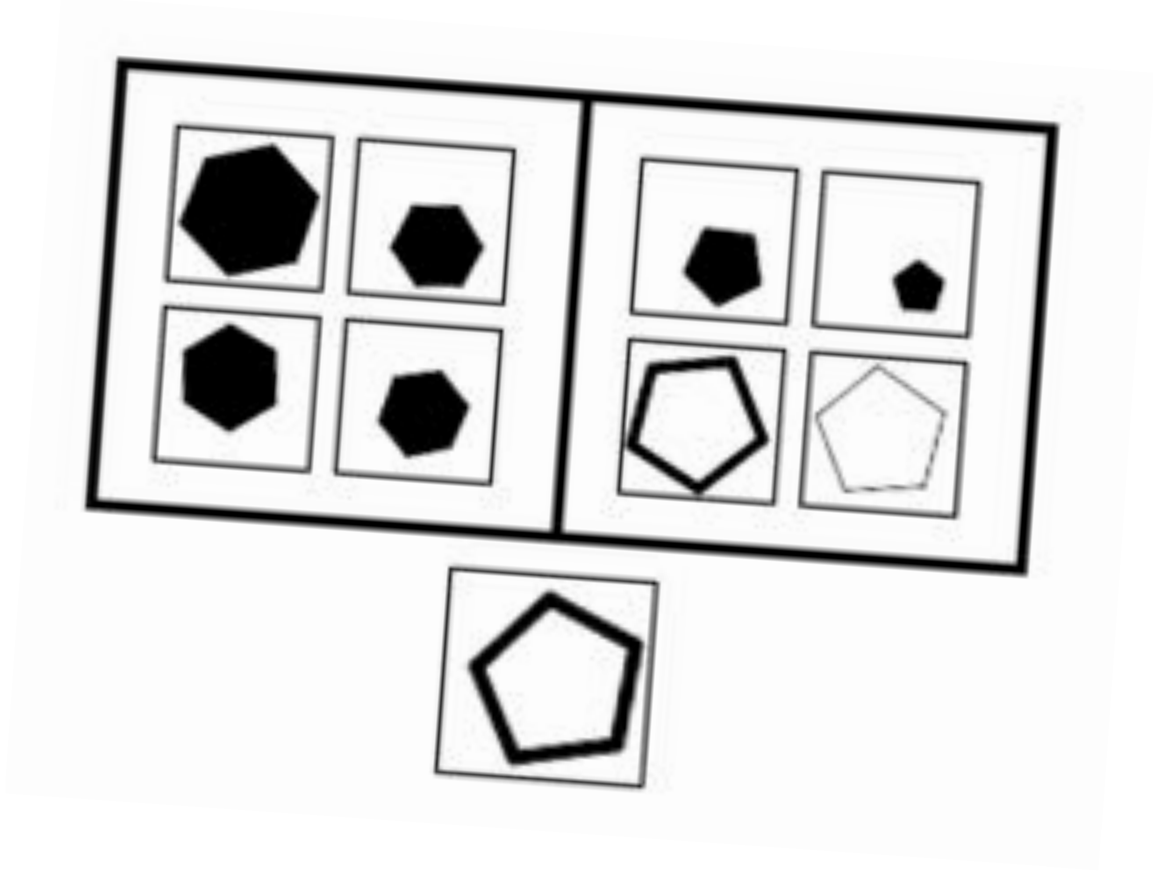
Image Understanding/Recognition CAPTCHAs (based on the ability of humans to recognise images.

PIX (CMU) - Identifying commonalities between images.
Asirra (Microsoft) - Given a set of images, identify all those containing a specific object.

Non-Visual CAPTCHAs

Audio / Sound-based CAPTCHAs render words/digits as distorted sound clips motivated by speech recognition problems analogous to the OCR problems that facilitate visual CAPTCHAs.

Which side does the isolated block belong to?



PIX - Identify commonalities between images.

Image Recognition CAPTCHAs

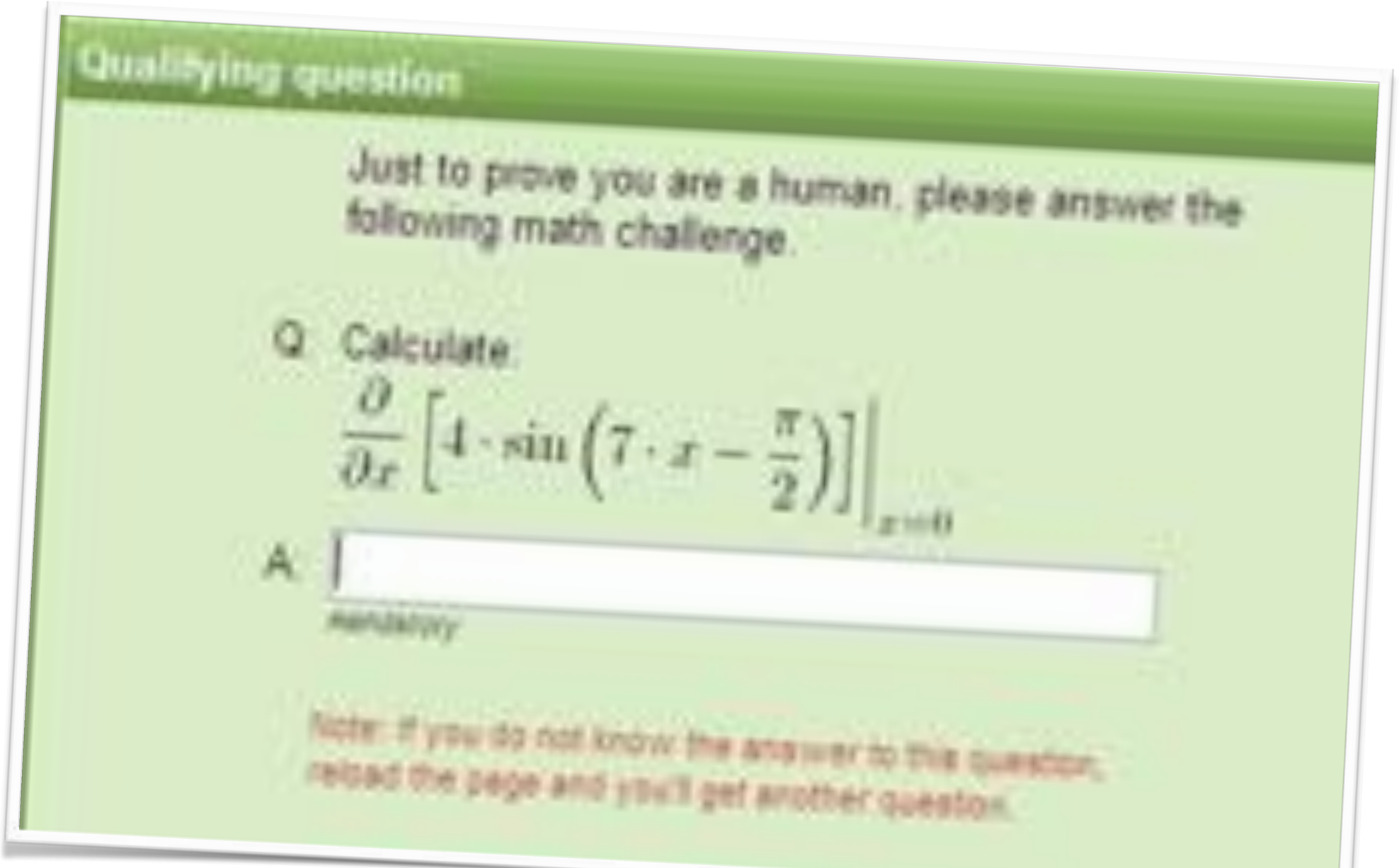
A 3D Object CAPTCHA

Enter letters corresponding to:

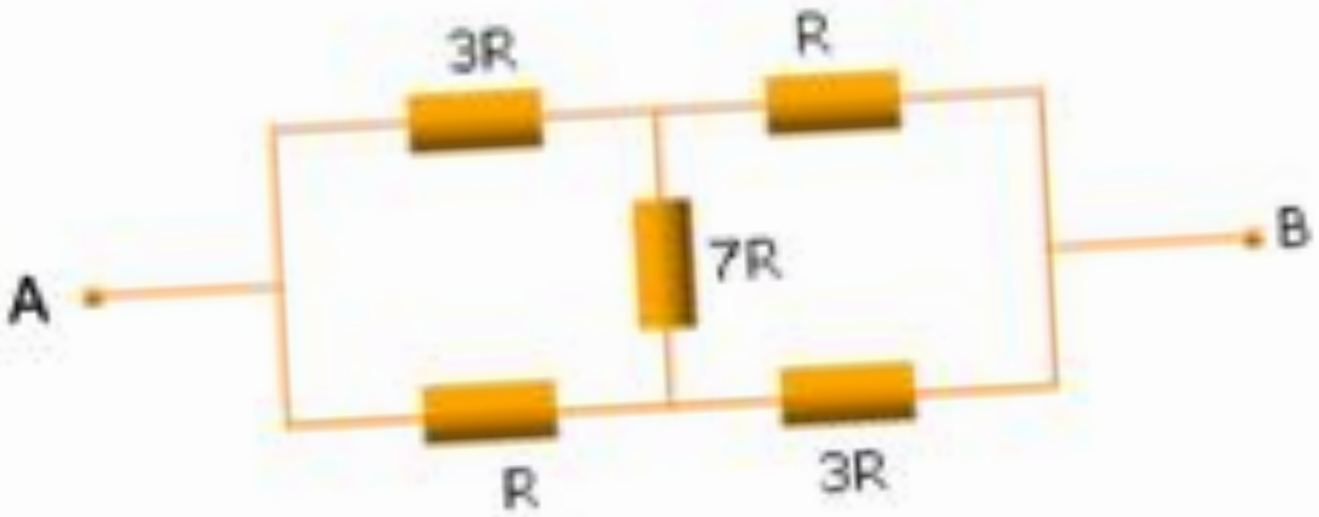
1. The head of the walking man.
2. The vase.
3. The back of the chair.

C K T

Mathematical CAPTCHAs (ahem!)



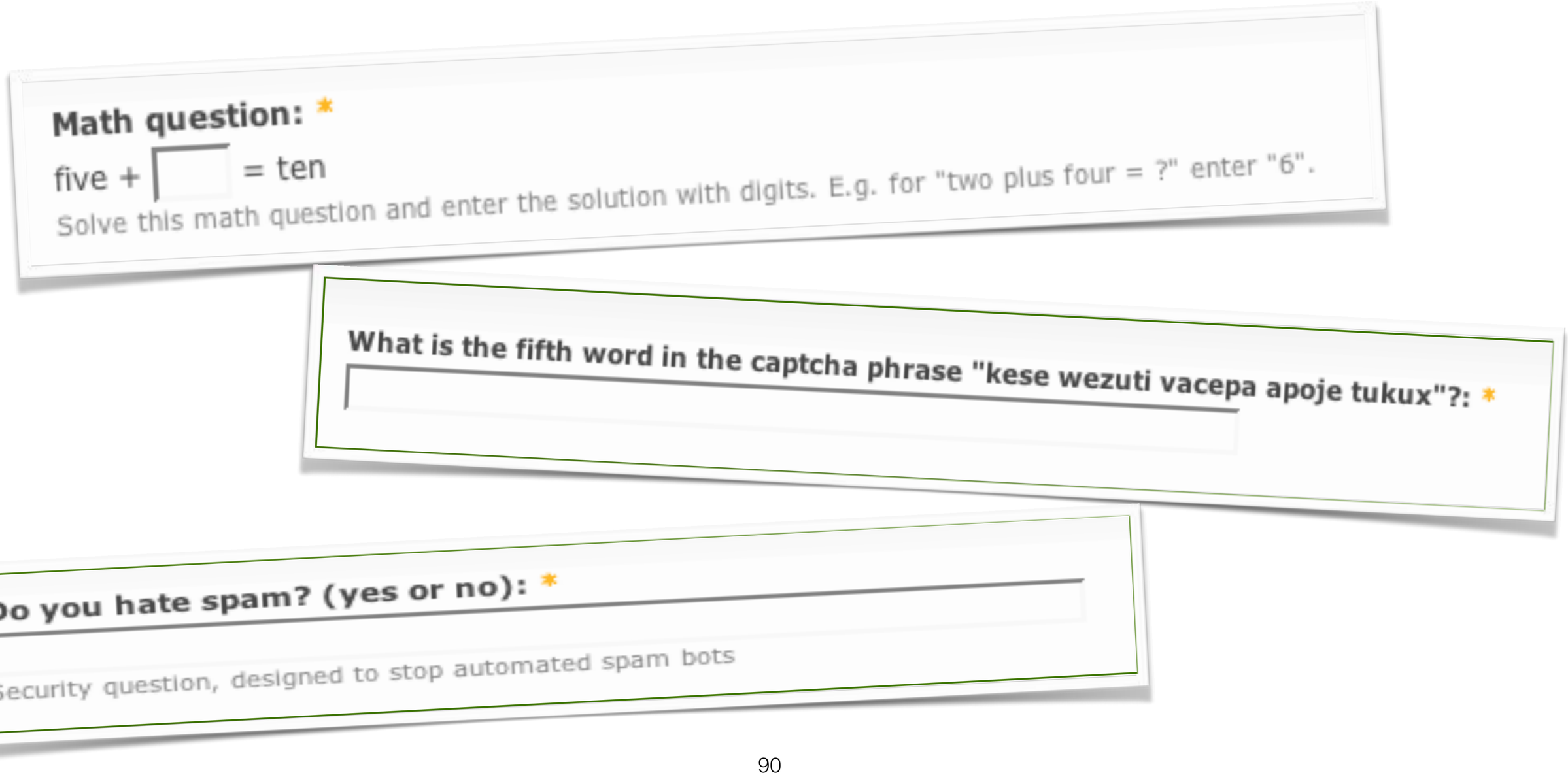
Eeek !



$R_{AB} = \frac{\square}{\square} R$

जमा करें

Some DRUPAL CAPTCHAs



When CAPTCHAs Go Wrong.

"Now what? I've been waiting for the last 20 minutes!"

CAPTCHA Guidelines

Accessibility

Visual CAPTCHA's are not accessible to the visually impaired, which contravenes national site accessibility regulations ⇒ Need to provide alternatives (e.g. audio) for such user groups.

Robustness

Obviously care must be taken to generate (random) distortions that are genuinely challenging. Many CAPTCHAs produce only minor modifications to the underlying image/sound which leaves them vulnerable to attack by automated methods.

Security

Even the toughest CAPTCHAs are useless unless they are integrated into a site/service in a secure way. E.g. (1) Encoding the CAPTCHA answer within the web page script in plaintext is a common mistake; (2) Repeating CAPTCHAs are vulnerable to "replay attacks" once their solution is known. Also the widespread adoption of certain classes of CAPTCHAs may motivate others to break them by automated methods – e.g. Fixed sets of text-based questions.

More on Breaking CAPTCHAs ...

Improving OCR/Speech Recognition Software

Already seen examples of how CAPTCHAs can be broken using AI techniques - improved image analysis/OCR, improved speech recognition. In other words, CAPTCHAs are helping to push the bounds of AI ...

What About Harnessing Human Labour?

The CAPTCHA Sweatshop with dedicated CAPTCHA solvers

CAPTCHA Backscratching Solution with an unholy alliance between spammers and porn sites.

CAPTCHA Sweatshops

SPAM companies hire cheap-labour to solve CAPTCHAs ALL DAY LONG!

The Economics if a CAPTCHA Sweatshop

Assume $2.50 per human solver per hour ...
... and a solving rate of 750 CAPTCHAs per solver per hour ...
... to generate new email accounts at a cost of 0.33 cent per account!

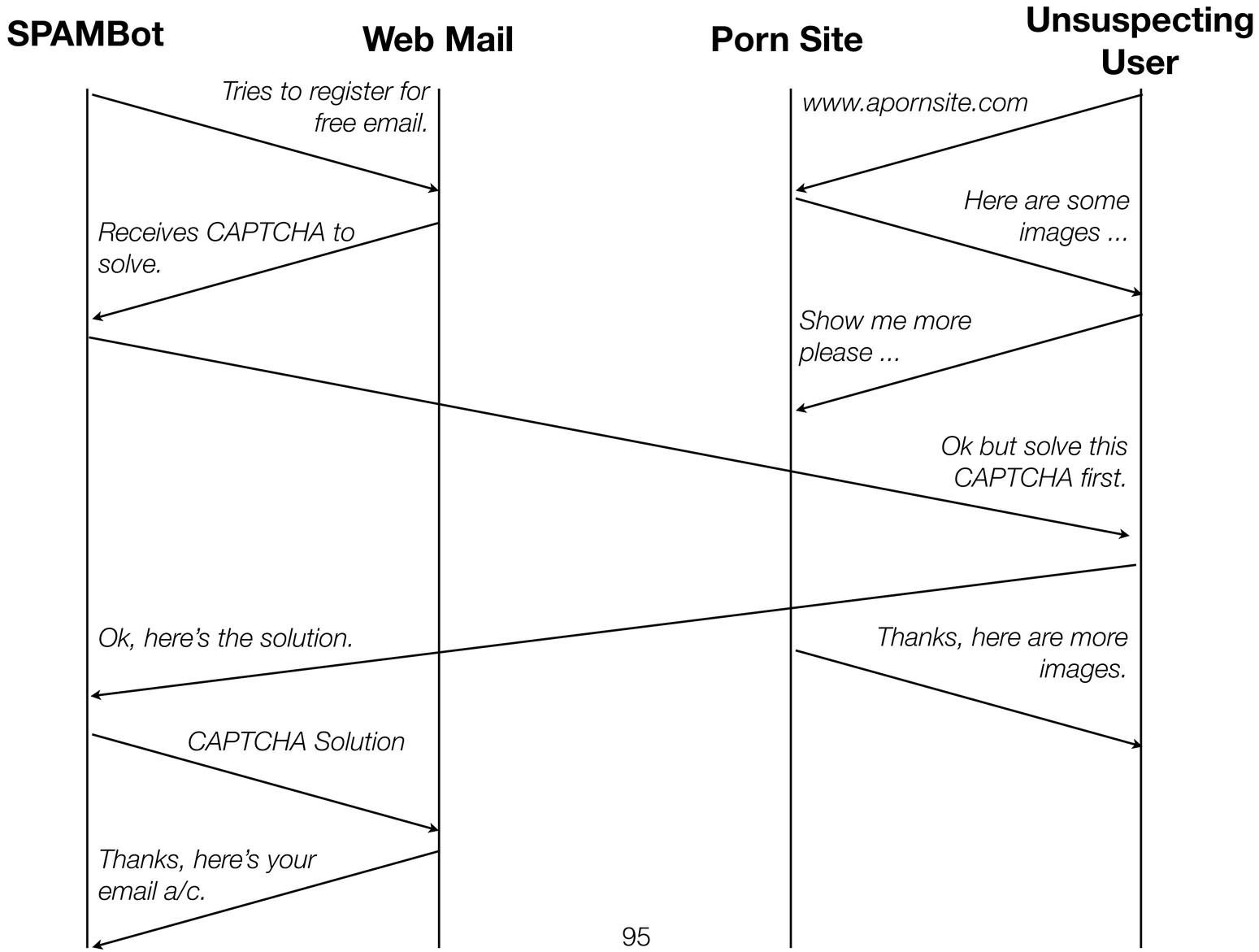
But creating say 1m fresh email accounts per day will require over 100 people working 12 hour shifts.

Given the returns on SPAM email the CAPTCHA sweatshop offers some questionable economics ...

... but at least it is now costing the spammers something to produce new email accounts...

... and jobs are being created in underdeveloped countries.

CAPTCHA Backscratching



CAPTCHA Scale

About 200 million CAPTCHAs are solved around the world every day (source: www.captcha.net).

Each CAPTCHA takes about 10 seconds of uniquely human effort ... not a lot individually ...

... but, collectively this means that CAPTCHAs are consuming about 150,000 hours of (wasted?) work every single day!

To put this in context the Empire State Building took 7m human-hours of effort to build, that's less that 50-days of global CAPTCHAs...

Hmmm ... might it be possible to harness all of this human computation to do something more useful?

Reading Text

Recognising Text ... the AI Way

Type vs Handwritten

Fonts, styles, size, noise, penmanship, ...

Bitmap ⇒ ASCII

Document Analysis

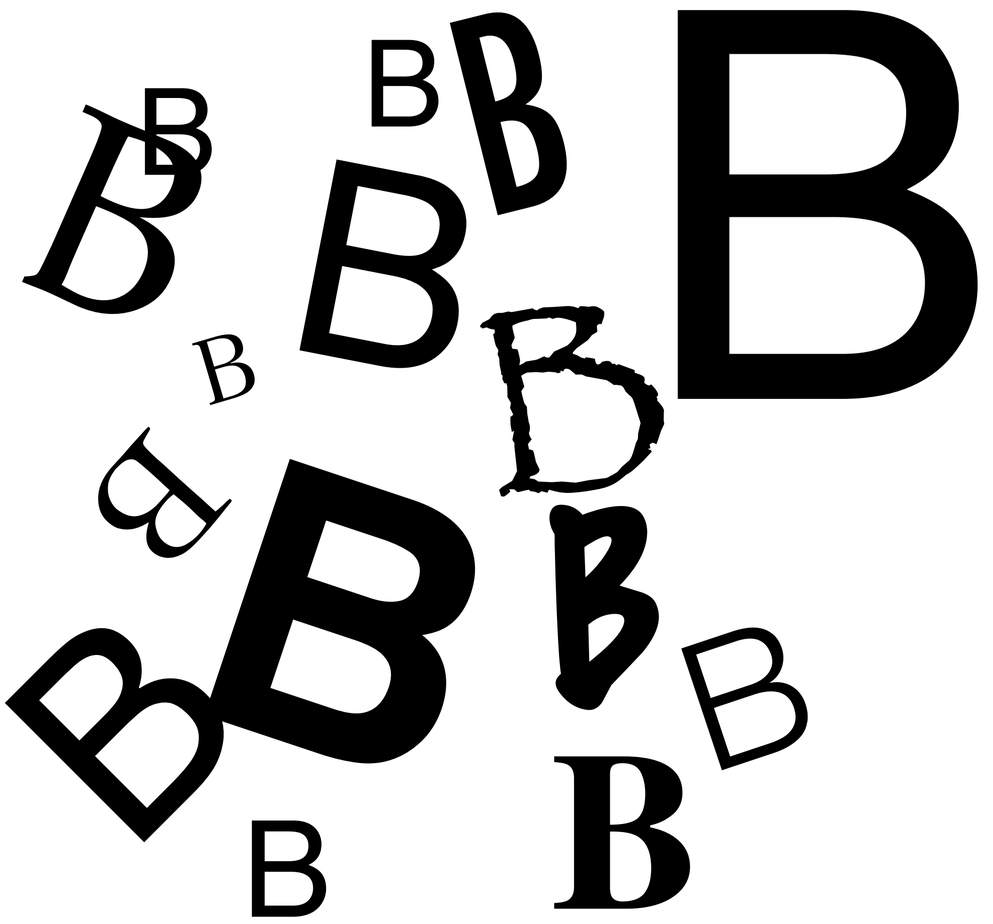
Alignment, Block Classification etc.

Segmentation

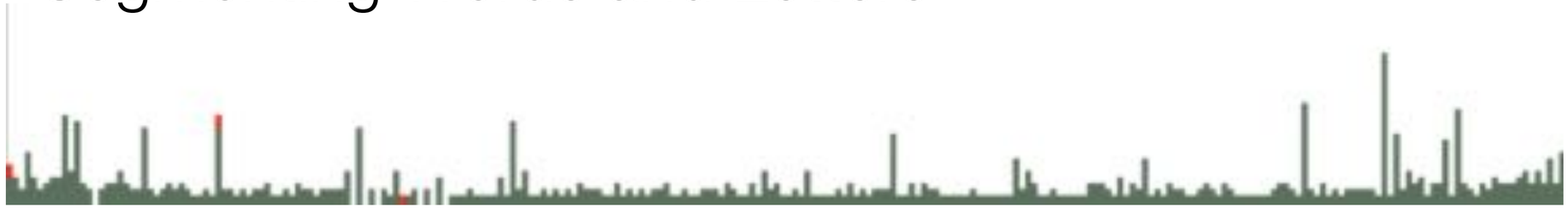
Line ⇒ Word ⇒ Letter

Letter Recognition

Feature extraction ⇒ letter classification

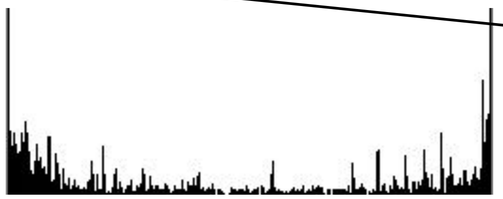


Segmenting Words and Letters

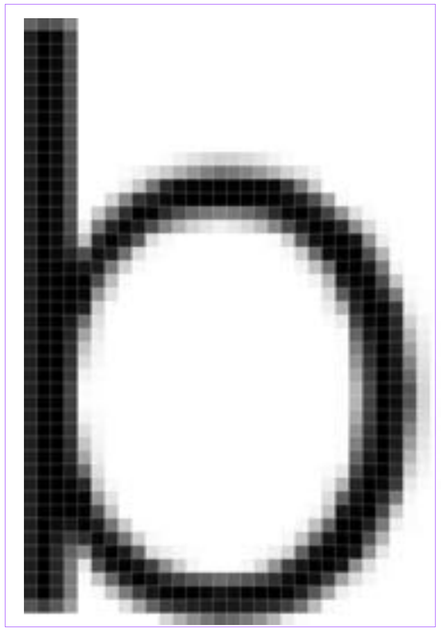


The quick brown fox jumped over the lazy dog.

brown



Horizontal histogram
(counting on-pixels per horizontal position)



Feature Extraction

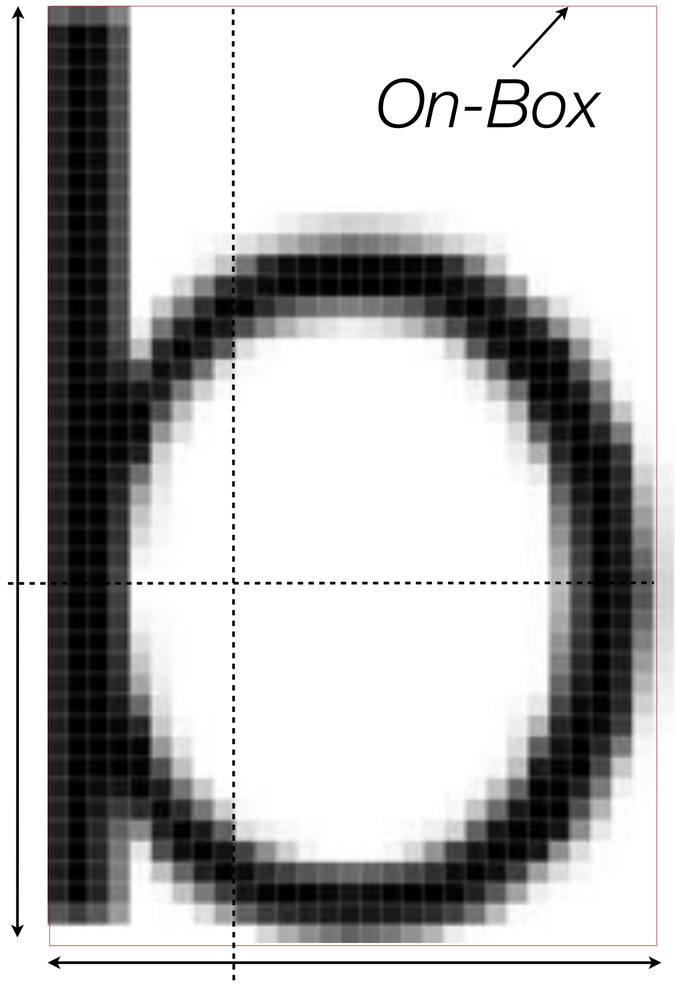
P. W. Frey, D. J. Slate, (1991) Letter recognition using Holland-style classifiers, Machine Learning

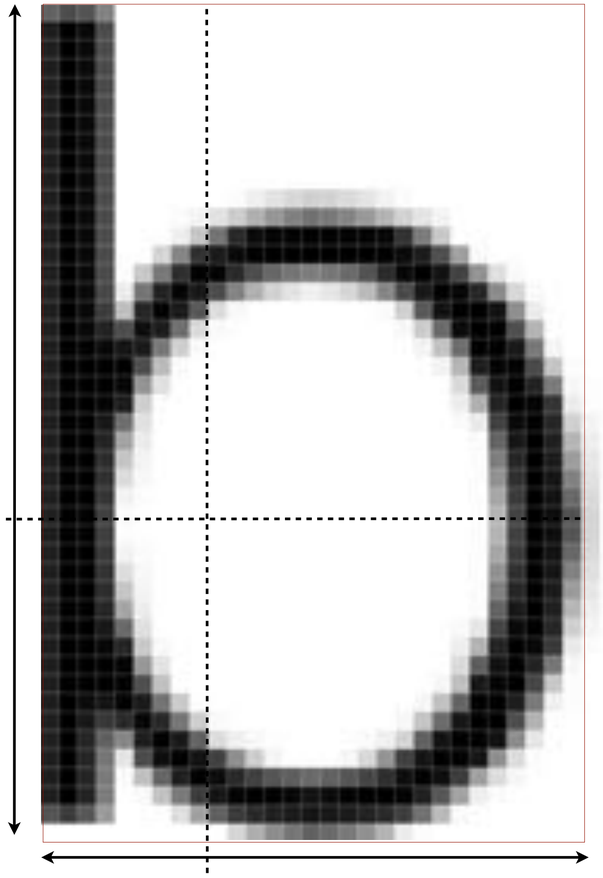
Key concept: On-Box - smallest rectangular box such that all on-pixels are contained.

Identify primitive features of pixel distributions from the on-box...

- Horizontal position
- Vertical position
- Width, height
- # on-pixels
- Mean horizontal/vertical position of on-pixels

Eg: B,4,2,5,4,4,8,7,6,6,7,6,6,2,8,7,10






Horizontal Position
T,2,8,3,5,1,8,13,0,6,6,10,8,0,8,0,8
I,5,12,3,7,2,10,5,5,4,13,3,9,2,8,4,10
D,4,11,6,8,6,10,6,2,6,10,3,7,3,7,3,9
N,7,11,6,6,3,5,9,4,6,4,4,10,6,10,2,8
G,2,1,3,1,1,8,6,6,6,6,5,9,1,7,5,10
S,4,11,5,8,3,8,8,6,9,5,6,6,0,8,9,7
B,4,2,5,4,4,8,7,6,6,7,6,6,2,8,7,10
A,1,1,3,2,1,8,2,2,2,8,2,8,1,6,2,7
J,2,2,4,4,2,10,6,2,6,12,4,8,1,6,1,7
M,11,15,13,9,7,13,2,6,2,12,1,9,8,1,1,8
X,3,9,5,7,4,8,7,3,8,5,6,8,2,8,6,7
O,6,13,4,7,4,6,7,6,3,10,7,9,5,9,5,8
C,4,9,6,7,6,7,8,6,2,6,5,11,4,8,7,8
M,6,9,8,6,9,7,8,6,5,7,5,8,8,9,8,6
R,5,9,5,7,6,6,11,7,3,7,3,9,2,7,5,11
F,6,9,5,4,3,10,6,3,5,10,5,7,3,9,6,9
O,3,4,4,3,2,8,7,7,5,7,6,8,2,8,3,8
C,7,10,5,5,2,6,8,6,8,11,7,11,2,8,5,9
T,6,11,6,8,5,6,11,5,6,11,9,4,3,12,2,4
J,2,2,3,3,1,10,6,3,6,12,4,9,0,7,1,7
Target Class
Letter Features

Learning to Classify

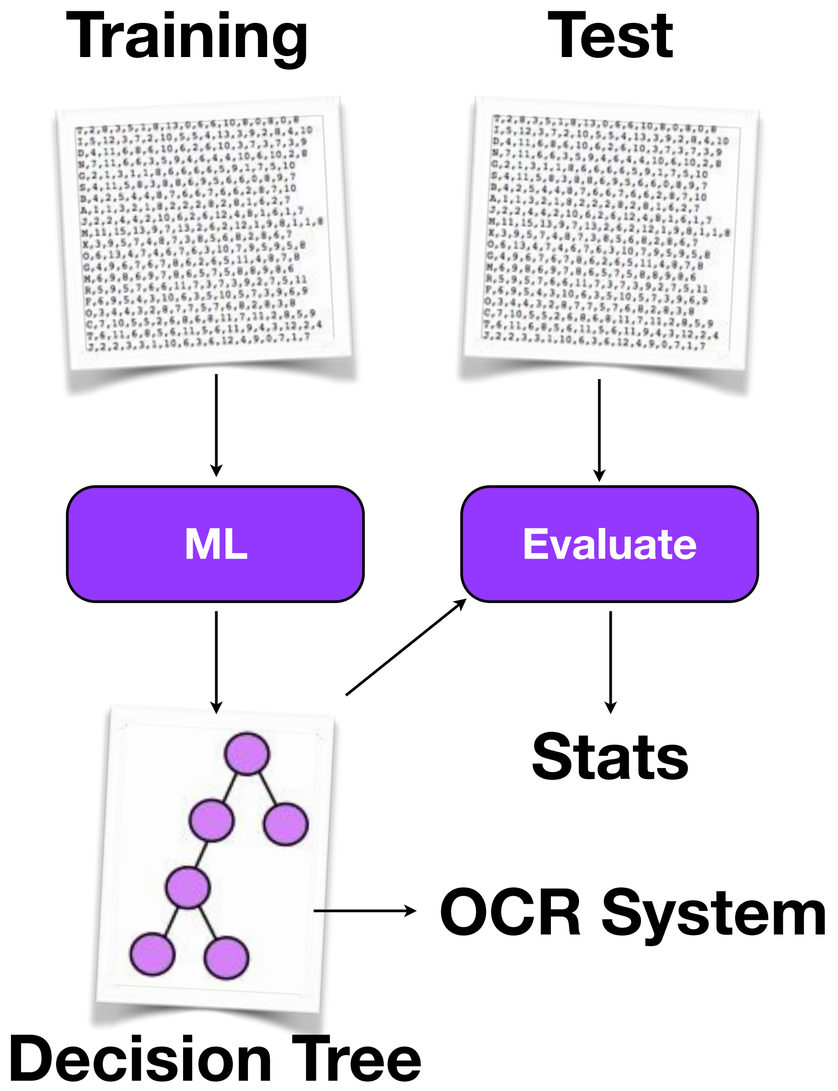
Generate labelled training data from pre-classified letters ⇒ ‘000’s of instances.

Wide range of ML techniques to convert training data into a classification model.

Neural Networks
Decision Trees
Naive Bayes
...

Test resulting model on unseen test data or use cross-fold validation techniques.

Resulting classifier can be used to classify unknown letters.



Issues

Conventional AI Solutions are Brittle

Highly dependent on representation (features), learning algorithm, and quality of the available training data.

Feature Extraction is Hard

There is no universal set of “good features” that are always guaranteed to lead to accurate letter recognition.

Choosing the Right Learning Algorithm is non-Trivial

Wide variety of machine learning techniques to cope with a diverse space of classification/prediction problems.

The State-of-the-Art

99% accuracy on clean, printed text (≈ 25-50 errors per page of text).
80%-90% accuracy on clean, hand-printed text; <80% accuracy for cursive script (dozens of errors per page)

So OCR/handwriting recognition is hard!

Is there a better way?

ReCAPTCHA

Based on CAPTCHA tests for validating human-users.

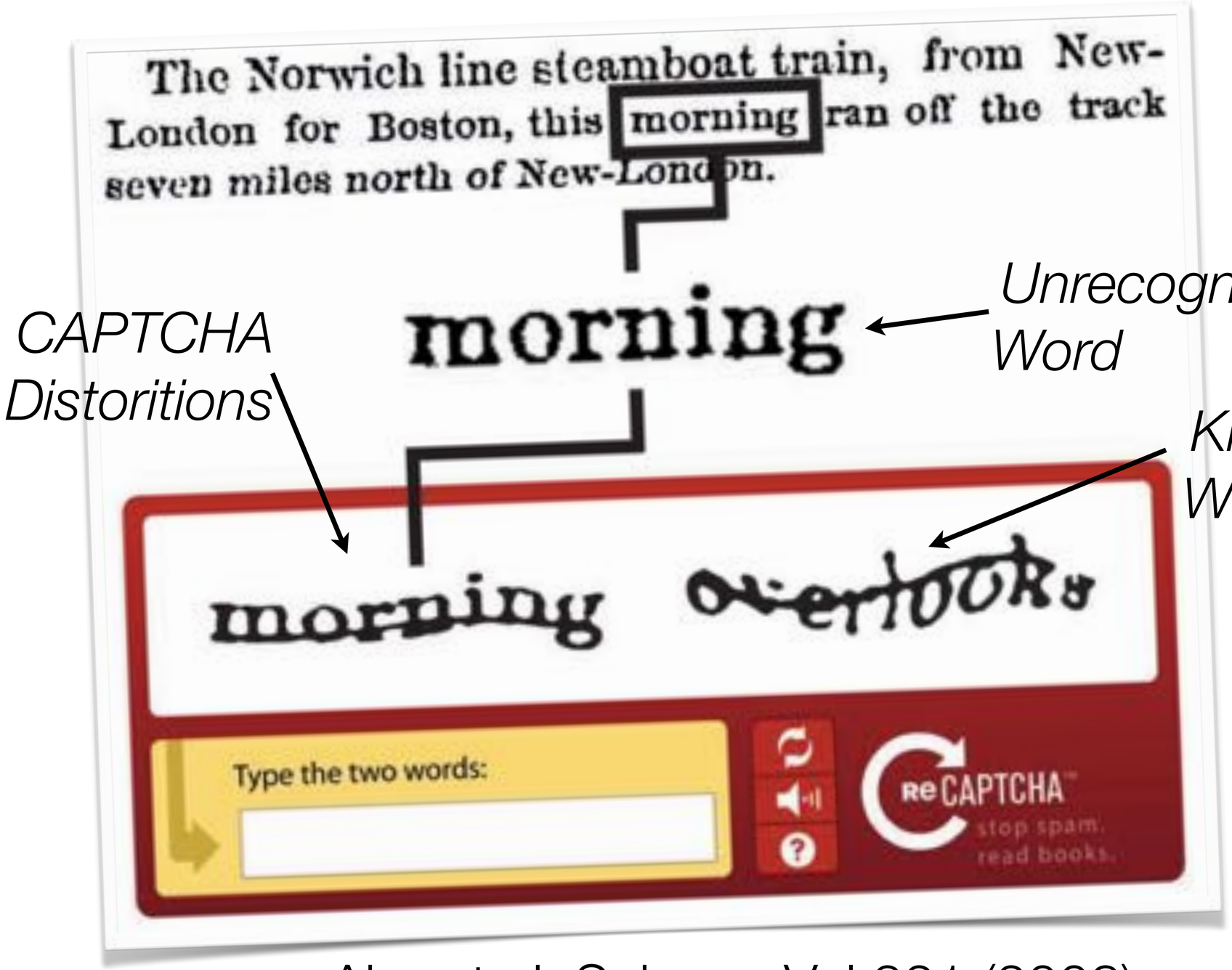
Instead of presenting a single word to decipher, reCAPTCHA's present two words. One words is known to the CAPTCHA but one is not.

The unknown word is typically an OCR failure.

The user does not know which word is known and which is unknown and so must attempt to decipher both words correctly.

The user passes the CAPTCHA if she correctly deciphers the known word.

Her attempt at the unknown word is stored and if enough people decipher the unknown word in the same way then a confident prediction can be made for it.

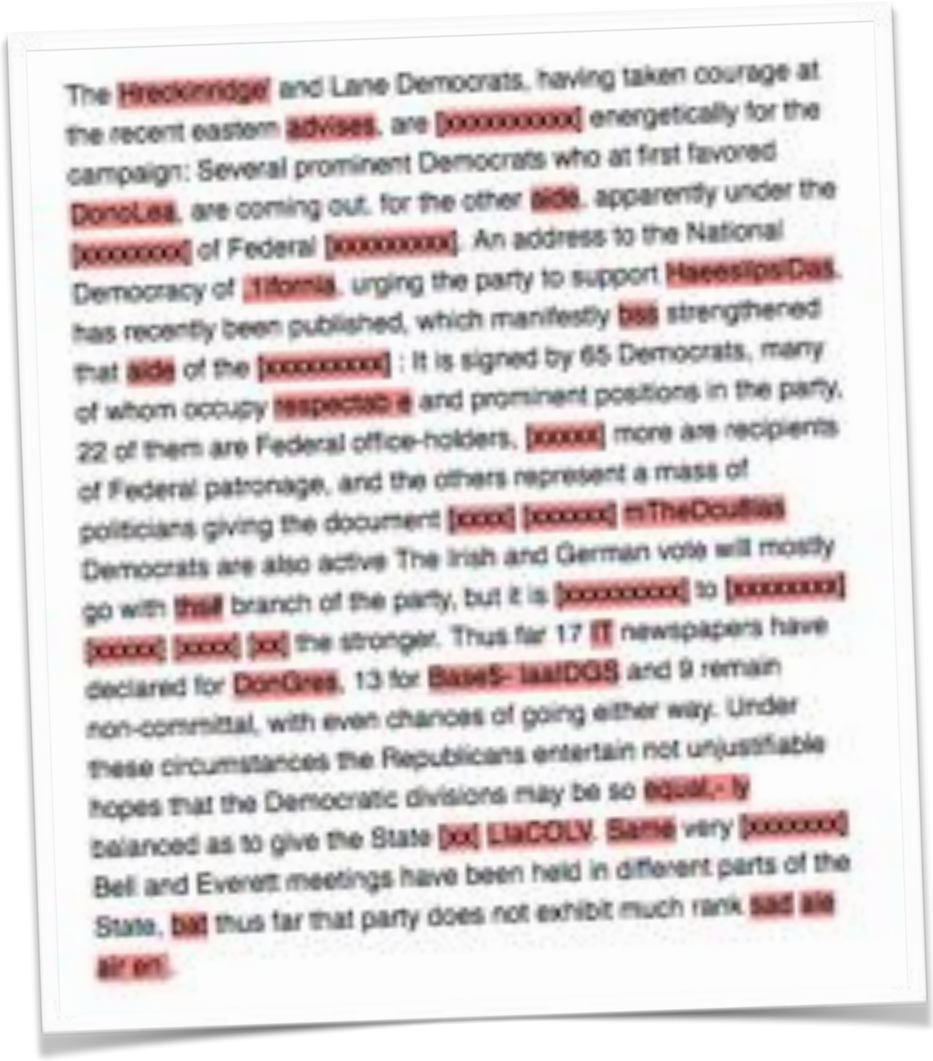


von Ahn et al. Science Vol 321 (2008)

Original

The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge, has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge, and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

OCR



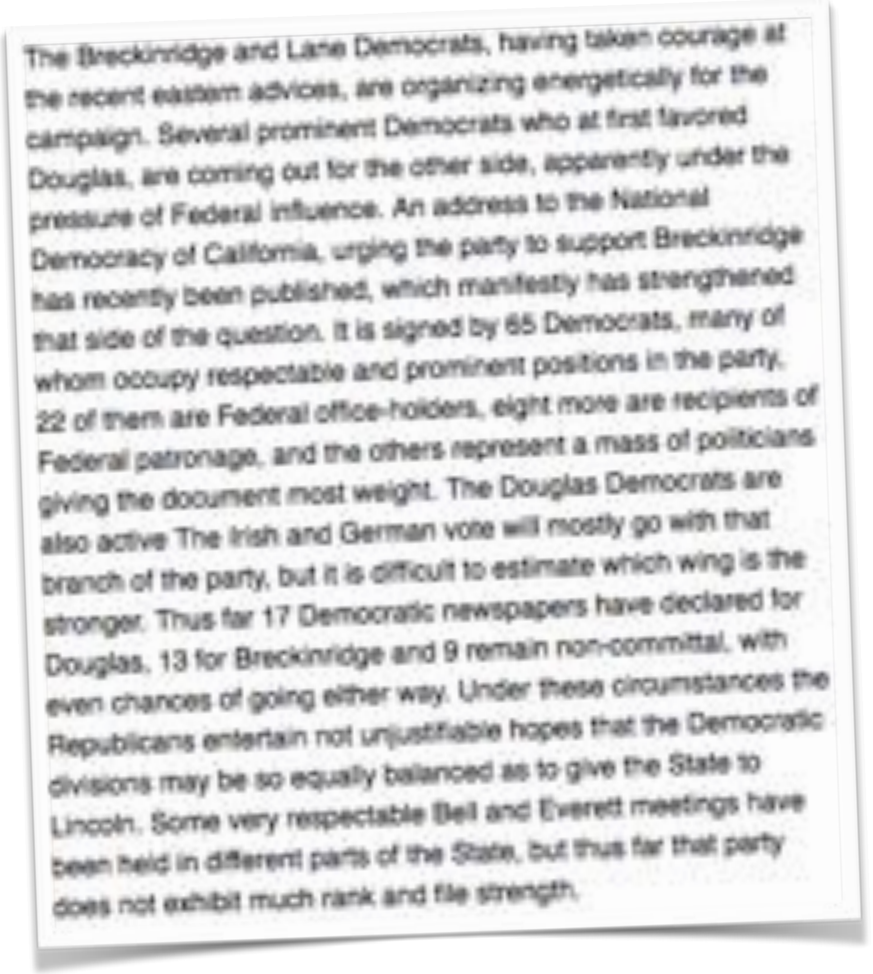
The Hreckinridge and Lane Democrats, having taken courage at the recent eastern advises, are [xxxxxxxxx] energetically for the campaign: Several prominent Democrats who at first favored DonoLea, are coming out, for the other aide, apparently under the [xxxxxxx] of Federal [xxxxxxxx]. An address to the National Democracy of 1ifornia, urging the party to support HaeesiipsiDas, has recently been published, which manifestly bss strengthened that aide of the [xxxxxxxx] : It is signed by 65 Democrats, many of whom occupy respectab e and prominent positions in the party, 22 of them are Federal office-holders, [xxxxx] more are recipients of Federal patronage, and the others represent a mass of politicians giving the document [xxxx] [xxxxxx] mTheDouttas Democrats are also active The Irish and German vote will mostly go with thsil branch of the party, but it is [xxxxxxxxx] to [xxxxxxxx] [xxxxx] [xxxx] [xx] the stronger. Thus far 17 IT newspapers have declared for DonGres, 13 for BaseS- laaIDGS and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equal.- ly balanced as to give the State [xx] LtaCOLV. Same very [xxxxxxx] Bell and Everett meetings have been held in different parts of the State, bat thus far that party does not exhibit much rank sad ale air en.

From www.recaptcha.net

Original

The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge, has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party, 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge, and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

reCAPTCHA



The Breckinridge and Lane Democrats, having taken courage at the recent eastern advices, are organizing energetically for the campaign. Several prominent Democrats who at first favored Douglas, are coming out for the other side, apparently under the pressure of Federal influence. An address to the National Democracy of California, urging the party to support Breckinridge has recently been published, which manifestly has strengthened that side of the question. It is signed by 65 Democrats, many of whom occupy respectable and prominent positions in the party. 22 of them are Federal office-holders, eight more are recipients of Federal patronage, and the others represent a mass of politicians giving the document most weight. The Douglas Democrats are also active The Irish and German vote will mostly go with that branch of the party, but it is difficult to estimate which wing is the stronger. Thus far 17 Democratic newspapers have declared for Douglas, 13 for Breckinridge and 9 remain non-committal, with even chances of going either way. Under these circumstances the Republicans entertain not unjustifiable hopes that the Democratic divisions may be so equally balanced as to give the State to Lincoln. Some very respectable Bell and Everett meetings have been held in different parts of the State, but thus far that party does not exhibit much rank and file strength.

From www.recaptcha.net

Key Findings

Large-scale deployments ...

ReCAPTCHA generated accuracy levels of 99.1% at the word level, compared to 83.5% for state-of-the-art OCR.

Moreover, professional manual deciphering of the test set of documents only delivered accuracy rates of 99.2%, barely more than ReCAPTCHA.

Within a year of launching ReCAPTCHA, humans 1.2 billion CAPTCHAs, correctly deciphering over 440 million words which is the equivalent to transcribing more than 17,000 books!

By 2008 ReCAPTCHA was deciphering the equivalent of 160 books/day

Further Information

Find out more about CAPTCHAs at http:///www.captcha.net

Recommended Reading List

Turing, A.M. (1950). Computing machinery and intelligence. Mind, 59, 433-460.
Greg Mori, Jitendra Malik: Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA. CVPR (1) 2003: 134-144
Luis von Ahn, Manuel Blum, John Langford: Telling humans and computers apart automatically. Commun. ACM 47(2): 56-60 (2004)
Greg Mori, Jitendra Malik: Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA. CVPR (1) 2003: 134-144
Kobus Barnard, David A. Forsyth: Learning the Semantics of Words and Pictures. ICCV 2001: 408-415
Luis von Ahn, Manuel Blum, Nicholas J. Hopper, John Langford: CAPTCHA: Using Hard AI Problems for Security. EUROCRYPT 2003: 294-311

Part 3 - Games with a Purpose

There are many problems that computers cannot currently solve…

Lets look at some examples ...

Example 1 - Image Labeling

Labeling Images with Words

Labeling images with a correct set of words continues to be an open problem in AI....

Labeling Images with Words

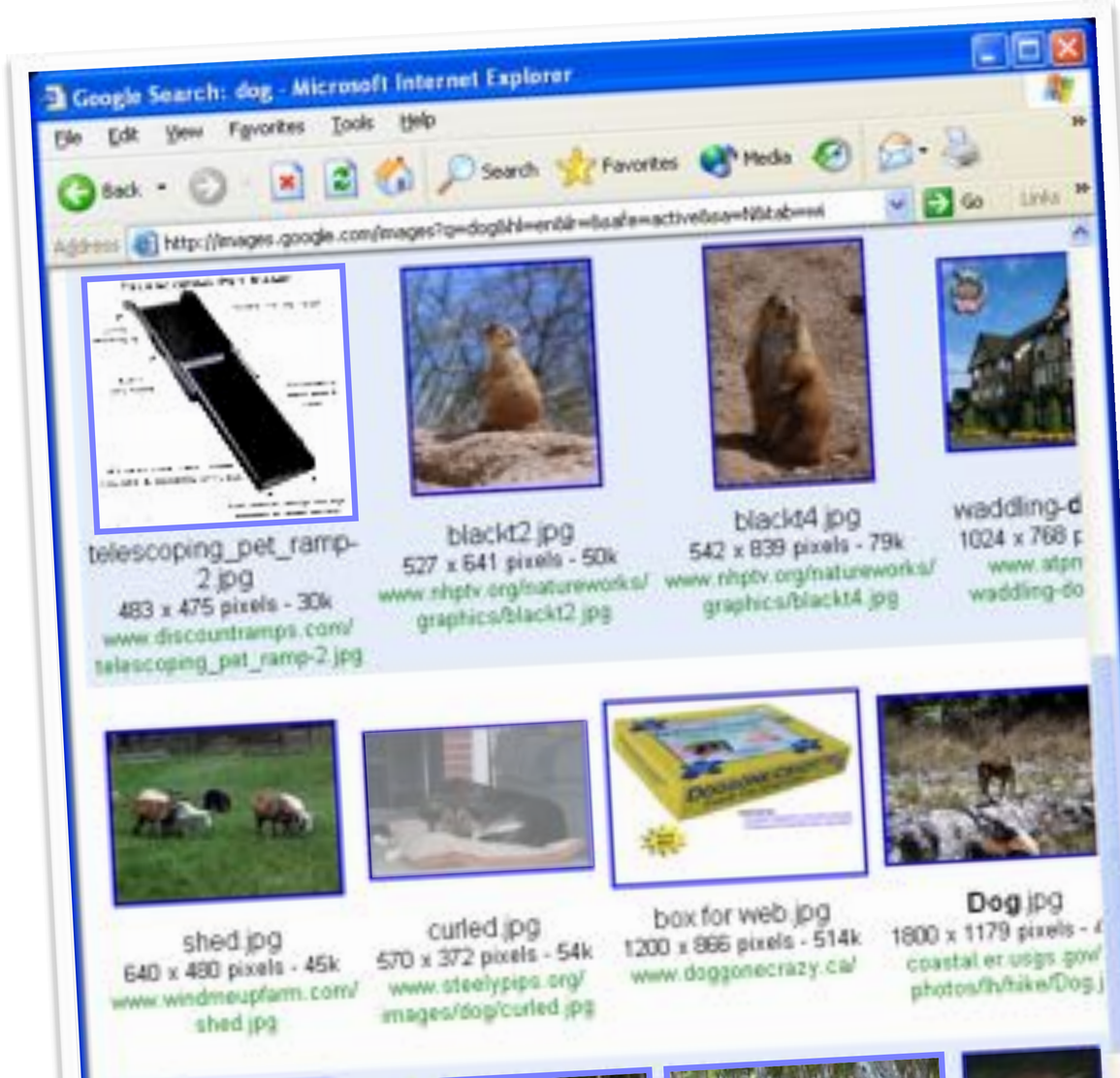
Bertie

Politician

Tea-Cup

…

Image Search on the Web



Conventional Image Search

There is no algorithm capable of taking an arbitrary image and labeling it with a correct set of terms/words that can be used for image search. So how does/did image search on the web work?

Typically image search engines operate by harnessing the text that is associated with images on the web.

Image filenames, hyperlink text, the works in the web page surrounding the image etc.

These terms can be used to index images and so act as a source of query matching during image retrieval; but of course this does not always work very well ...

How can we build a better image search engine?

... by creating a better image labeler!

In other words, we want a method that can label all images on the web ...

…but it has to be fast, cheap, and correct.

Okay, but how do we create this better image labeler?

Use humans ...

We could pay people (cheap?)

How do we find willing labelers? (fast?)

What if we could attract people to label images for free...

The secret?

Make it fun!

Games-with-a-Purpose Overview

What are Games-with-a-Purpose?

Lots of examples of different types of GWAPs.

A framework for understanding GWAPs.

Some GWAP Gotchas.

Designing and building your own GWAP.

What are GWAPs?

Simply put …

…a GWAP is a computer game (in the sense that it's primary objective is *entertaining play*) like any other computer game.

But, as a by-product of play …

…some useful information is created that can contribute to the solution of some secondary (and presumably challenging) problem/task.

The hope is that a fun game will attract many players to produce problem solving data at scale…

…but designing a GWAP is not as easy as it might first seem.

Pay Attention! Ask Questions!

You will receive your GWAP project specification in the coming weeks.

You will be designing (5-Credit) and building (10-Credit) your own GWAP.

The next few lectures will cover a lot of GWAP ground and the more inquisitive you are now the more likely you are to develop a strong GWAP idea of your own…

The ESP Game

The ESP Game

An very enjoyable image labeling game. As a side-effect of game-play, people are labeling images... and lots of them.

The labels that they are generating are likely to be relevant and useful for tagging images.

Image labels are generated extremely quickly, so much so that is the game was placed on a high-traffic gaming site then all of the images on the web could potentially be labeled in a matter of weeks.

How does it work?

Symmetric ,***two-player*** online game. Players are randomly paired at the start of the game. Players do not know each other and cannot communicate with each other during game-play.

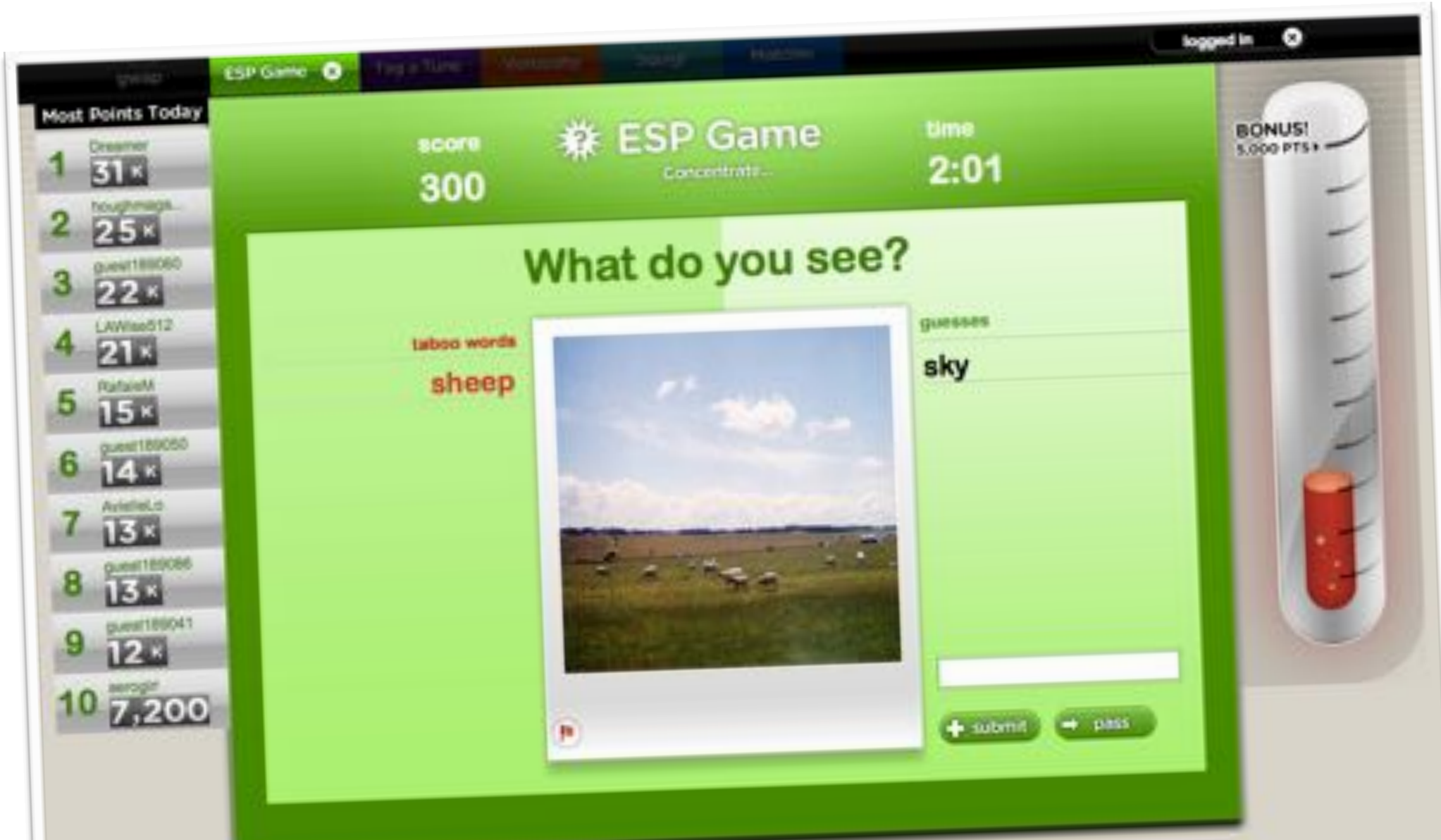
Each player is shown the same single image and the goal of the game is for both partners to type the exact same word.

Both players (intuitively) type words related to the image and the players ***win*** the round when they have both typed the same word for the image.

And because the words come from two independent sources they are usually very good labels for the image in question.


ESP Game
Top Scores
Today
This Month
All Time
1 lubeckm 51
2 greggiq 5 209
3 Catwoman 4 477
4 pook 3 283
5 knowitall 2 540
Got it, Let's Play!

The Anatomy of the ESP Game



What are Taboo Words?

Taboo words are a key element of the game. These are words that the players are not allowed to enter, or that at least do not count in a particular round.

Usually these are selected from words that are commonly used by past players. A simple strategy might be to use previously agreed labels as Taboo words, for example.

So taboo words are generated from the game itself and they accumulate as images are labelled.

Taboo words encourage players to seek out less obvious labels for images and improve the labeling coverage

When is an image done?

As images pass through the ESP game they accumulate several/lots of labels.

It is meaningful to ask the question:

when has an image has been fully labeled.

When is an image done?

As images pass through the ESP game they accumulate several/lots of labels.

It is meaningful to ask the question: when has an image has been *fully labeled*.

In the ESP game the principle at work is to consider an image as fully labeled when it is no longer enjoyable for the players when it is included in the game.

A signal that this is the case is when players repeatedly pass on an image.

This will typically occur when an image has accumulated lots of labels and therefore lots of taboo words so that future agreement is just unlikely. Players will spend a long time in a given round and ultimately pass in frustration.

Label Thresholds

What words are attached as labels to the image?

The basic idea is that agreement by two independent labelers probably suggests that the agreed word is a good image label.

To improve the quality of labels it is straightforward to mediate labeling by the addition of a good label threshold.

In this case a agreed word, *w*, is only used as a label for the target image, if it has been agreed on, for that image, by at least *T* player-pairs.

So if *T = 1* then a single instance of agreement is enough at the image is labeled with w. Obviously higher values of *T* lead to higher quality labels.

What if it is not possible to pair two players?

Player Selection & Pre-Recorded Game Play

What happens if it is not possible to pair two players?

In practice the ESP game can be played by a single human player paired with a computer player (bot).

But where does the bot come from, since we known that it is impossible to algorithmically label images?

Player Selection & Pre-Recorded Game Play

The ESP game records player actions (guesses and timings) during game play …

… and this information can be replayed as a *bot*.

In this way a new player can effectively play against a previous player over the same set of images that the earlier player labeled.

Does this mean that labeling effectively stops? ...

Cheating

Players cannot communicate by design. If they could then they could easily cheat to build high scores and damage the labeling.

In theory however, cheating could be possible, if for example, a large group of players adopted some null labeling strategy such as always labeling an image with some agreed word or letter.

This is unlikely because of the random selection of players from a large pool of players.

Other controls include monitoring of IP addresses and timing information. If players seem to find rapid agreement then bot-players are often used to prevent cheating.

Global taboo words, which persist across a game session, can also be used.

Image Selection

Where do images come from?

In the early days the ESP game selected images randomly from across the web with controls on image size , aspect ration etc so that they would fit the game window.

Laterly images were selected from the Google image repository and currently image are selected from Flickr.

There are many strategies that could be used to deliver a more interesting gameplay: selecting interesting and/or top-rated images from Flickr; pairing players from a similar location and selecting images from that location.

Significant attention needs to be given to the selection of appropriate image content. E.g. avoiding images from pages containing inappropriate terms.

Context Specific Labels

Presenting images that have been selected at random to a wide-ranging audience is likely to result in labels that are very generic (e.g. house, dog, cat, baby...)

More specialised labels can be more useful, especially if labeled images are to be used in image search.

Ask players to declare interests as part of their gaming profile, pairing players with overlapping interests, and then selecting images based on these interests.

Original ESP creators considered the idea of Theme Rooms for this reason.

Evaluation

Does it work? What does that mean?

Do people enjoy playing the game? How often and for how long?

How many images are labeled and how many labels to the acquire?

Are the labels good ones?

Key Results (circa 2004)

During the first 4 months of gameplay (August - December, 2003)...

13,630 people played the game during this time, generating 1,271,451 labels for 293,760 different images.

Over 80% of the people played on more than one occasion (i.e., more than 80% of the people played on multiple dates). Furthermore, 33 people played more than 1,000 games (this is over 50 hours of playing!).

The average number of labels collected per minute by a pair of individuals is 3.89 (std. dev. = 0.69).

At this rate, 5,000 people playing the ESP game 24 hours a day would label all images on Google (425,000,000 images) in 31 days.

Search Precision

Examine the results of searching for all images associated to particular labels.

Chose 10 labels at random from the set of all labels collected using the game. We chose from labels that occurred in more than 8 images.

All (100%) of the images retrieved made sense with respect to the test labels. In more technical terms, the precision of searching for images using our labels is extremely high.

This should be surprising, given that the labels were collected not by asking players to enter search terms, but by recording their answers as they tried to maximize their score in the ESP game.

E.g. Opposite, all images labeled with car.

Manual Label Assessment

15 unique participants rated the quality of the labels generated using the game.

Twenty images were chosen at random out of the first 1,023 images that had more than 5 labels associated to them by the game.

All 15 participants were presented with each of the 20 images in randomized order and shown the first six words that were agreed on for that image during the game. For each of the 20 image-word sets they were asked to answer the following questions:

Manual Label Assessment (2)

How many of the words above would you use in describing this image to someone who couldn't see it.

How many of the words have nothing to do with the image (i.e., you don't understand why they are listed with this image)?

For question 1, the mean was 5.105 words (std. dev. 1.0387), indicating that a majority (or 85%) of the words for each image would be useful in describing it.

The mean for question 2 was 0.105 words (std. dev. 0.2529), indicating that for the most part subjects felt there were few (1.7%) if any labels that did not belong with each image.

Google Image Search and Image Labeler

Let’s look at another challenging image-related problem...

Example 2 - Locating Objects in Images

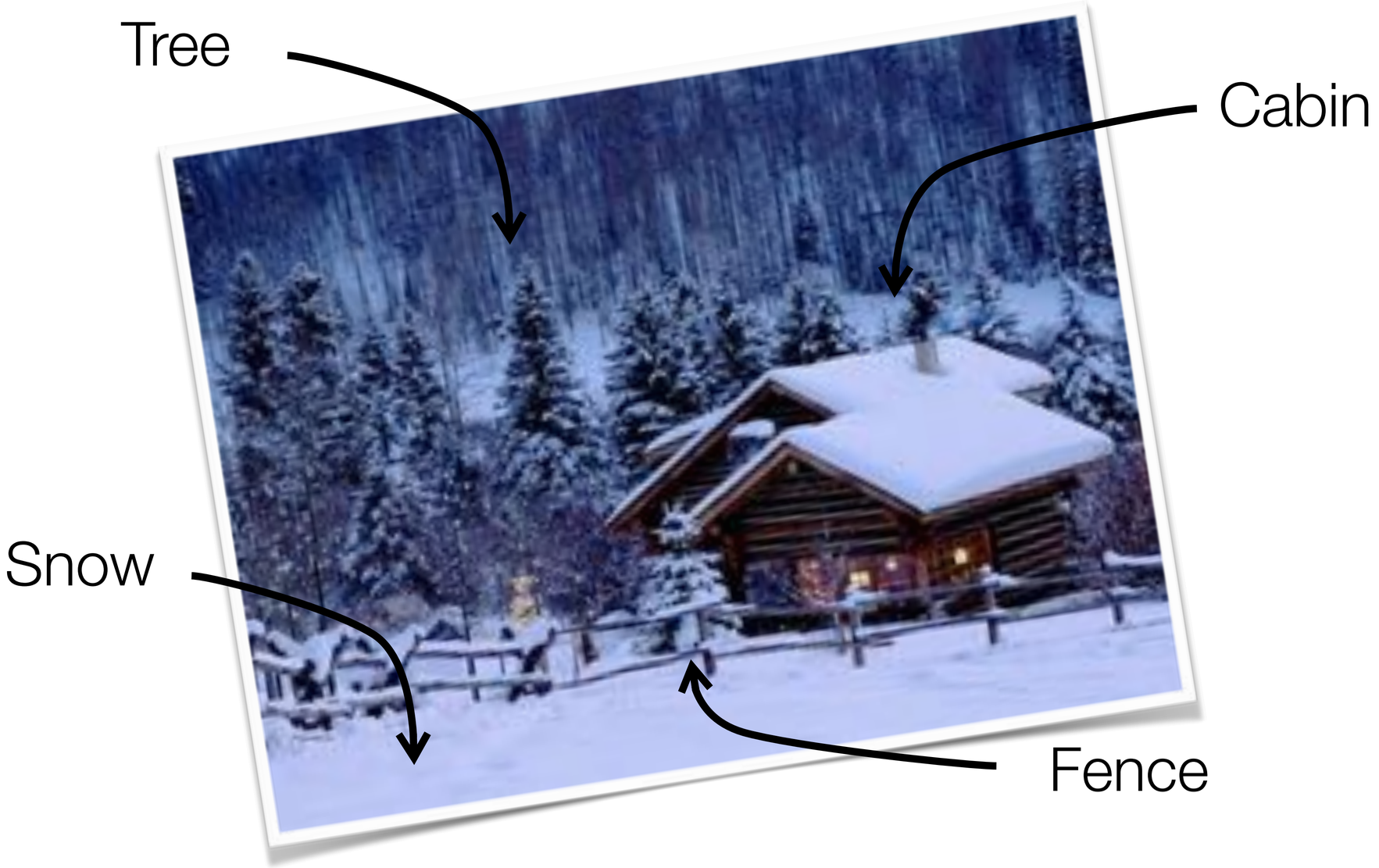
Object Labeling

The ESP game focused on generating labels for ***images***. Generating labels for ***objects in*** **images** is different …

The task is to generate labels to identify all of the objects in an image as well as associated those labels with the appropriate object in the image.

Significantly more challenging that simply generating image labels?

The ESP game can, in principle, be used to identify what objects are in an image but it does not help us locate the objects within the image.


Tree
Cabin
Snow
Fence

Peek-a-Boom

How does it work?

Asymmetric two-player image annotation game. Each player adopts one of two roles: ***peek*** or ***boom***. The goal of the game is for boom to gradually reveal parts of the image to peek, and for peek to guess the correct label for this part.

Boom reveals parts of the image by clicking on the image; each click reveals a 20-pixel radius area of the image.

Meanwhile peek types in guesses and boom responds by indicating whether those guesses are hot or cold.

When peek correctly guesses the object, both players receive points & switch roles.

Boom is naturally motivated to only reveal those parts of the image that relate to a specific object because this facilitates Peek's guesses.

The Anatomy of Peek-a-Boom

Hints & Pings

Sometimes it can be hard for boom to localise peek's guesses.

By right-clicking on a part of the object, boom can create a ripple at this location for peek.

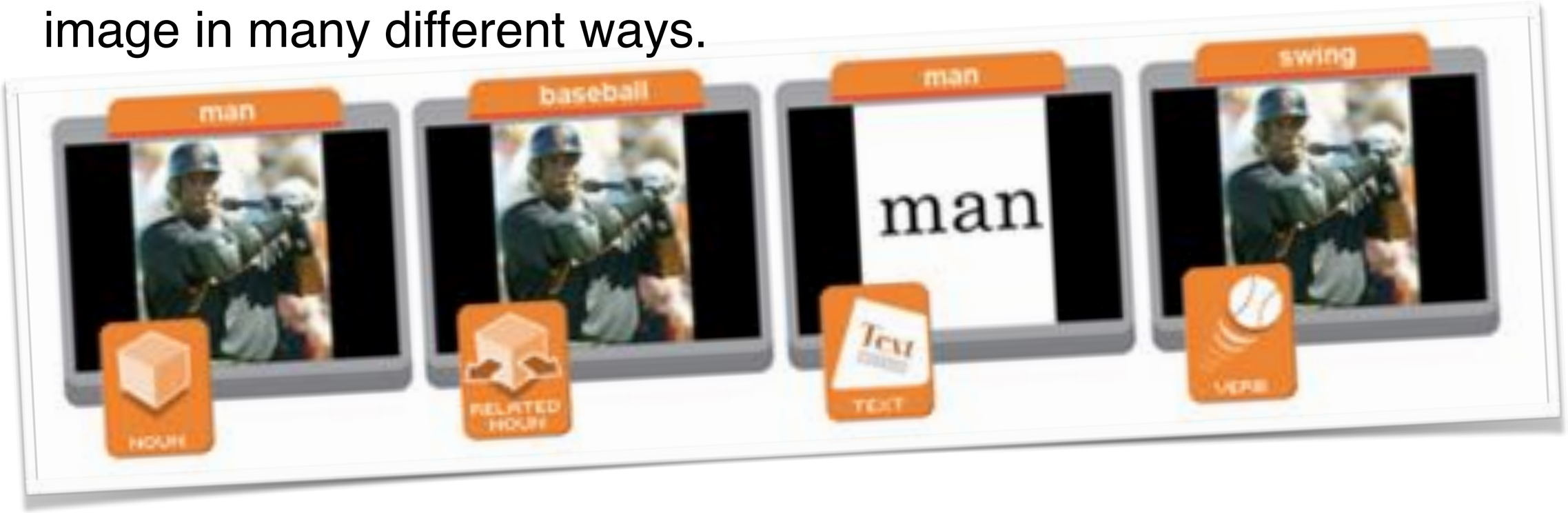
In this way, boom can guide peek's guesses towards a particular part of the object.

E.g. in this game the task is to identify the trunk of the elephant and so boom pings the trunk section to help peek.

More Hints

Peek-a-boom provides boom with access to a set of pre-defined hint buttons to give peek clues about the type of label that is being sought.

This helps peek to disambiguate between words that relate to an image in many different ways.



Where do the images and labels come from?

Each peek-a-boom round consists of an image and a label.

Where do these data come from?

The simple answer is that data is drawn from the output of the ESP game. That is, the most common labels collected for images of the ESP game are used as the basis for peek-a-boom games.

Scoring & Bonuses

Both boom and peek score equal points whenever peek correctly identifies the target object.

Additional points are given for this use of the hint buttons. This is counter-intuitive. Why was this adopted?

Every time a player completes 4 images they are sent to a timed bonus round. In this round each player is simply asked to click on the image containing a given object and the players score points according to how close their clicks are.

Why use a bonus round? Reinforces gameplay fun by adjusting the style of play. It also has a *leveling-up* effect creating the sense of game progress. And it collects more object data.

Uses for Peek-A-Boom Data

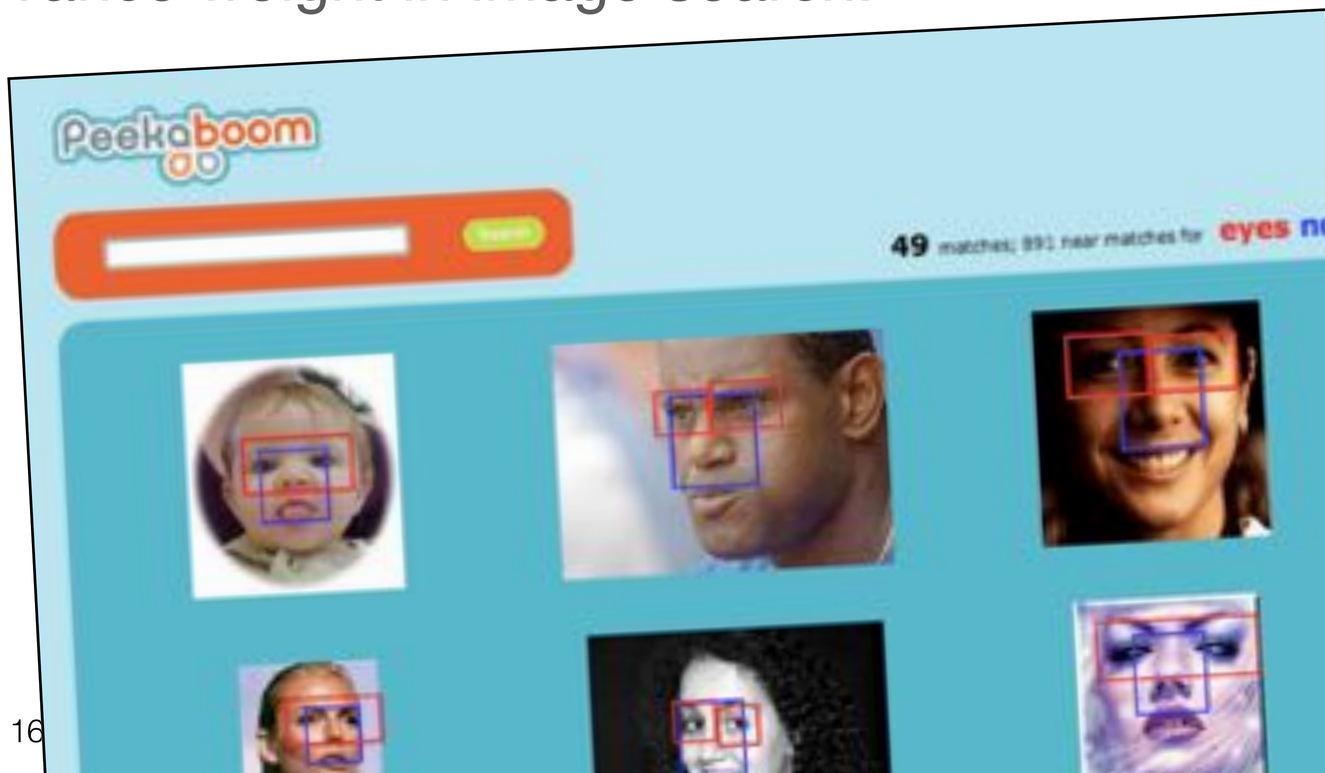
Machine learning training data for object recognition tasks.

Improving image search results.

Peek-a-boom gives an accurate assessment of how much of an image is related to a given term which can be used as a relevance weight in image search.

Object bounding Boxes

Peek-a-boom can be used to automatically generate labeled object-bounding boxes similar to those used in Flickr by aggregating different plays of the same image-word pairs.



Does it work?

Usage Statistics (circa 2006)

August 1, 2005 to September 1, 2005: 14,153 different people played the game generating 1,122,998 pieces of labeled object data.

On average each person played on 158.68 images over an average of 72.96 minutes.

Over 90% of the people played on more than one occasion (that is, more than 90% of the people played on different dates).

Furthermore, every player in the top scores list played over 800 games (that’s over 53 hours without including the time they spent waiting for a partner!). This undoubtedly attests to how enjoyable the game is.

The Accuracy of Bounding-Box Data

Are peek-a-boom bounding boxes as good as bounding boxes people would make around an object in a non-game setting?

Selected at random 50 image-word pairs from the data pool that had been successfully played on by at least two independent pairs of people.

For each image, peek-a-boom data was used to calculate object bounding boxes using the method explained in previous sections.

4 volunteers made bounding boxes around the objects for each image, providing 200 bounding boxes drawn by volunteers.

The average degree of pixel overlap between the peek-a-boom bounding boxes and those created by the volunteers was 0.754 (std dev 0.109)

Squigl - Another Object Segmentation Game

Symmetric two-player game.

Each player sees the same image-label pair and must trace the outline of the labeled object in the image.

When both players have completed their trace, the receive a score based on the overlap between their traces.

Let’s see an example ...

gwap | ESP Game | Tag a Tune | Verbosity | Squigl | Matchin | Flipit | PopVideo | Refer for 15,000 pts! | guest189085

Squigl

Ready, set... trace!

How to Play

1 You and a partner see the same image and word.

2 Hold down the mouse and trace the object described by the word.

3 Click submit.
You get points for matching your partner's trace.

+ submit

Got it, Let's Play!

View Top Scores

Example 3 – Labeling Music

Tag-a-Tune

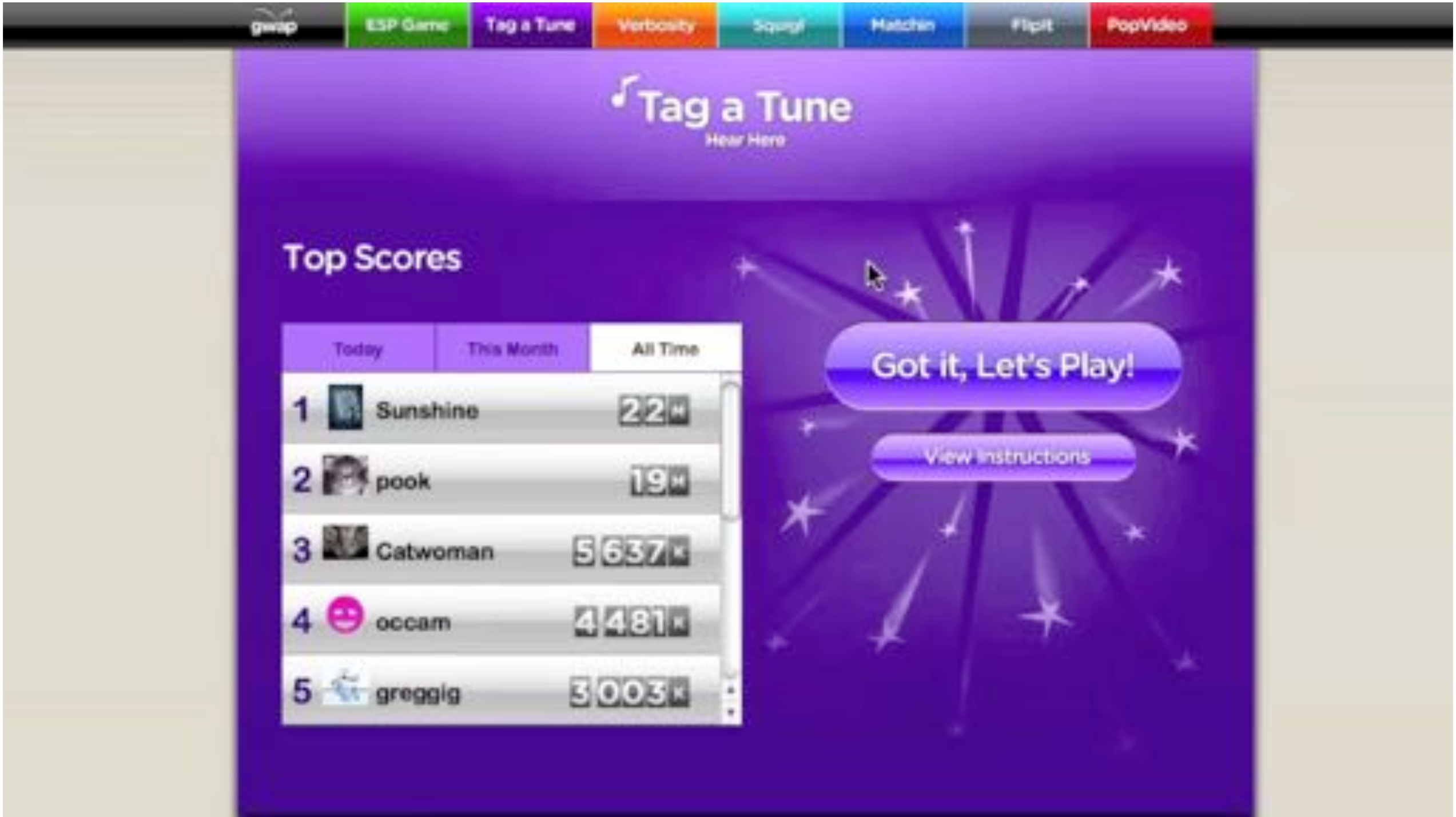
How does it work?

Firstly, an early version of producing an audio version of the ESP game did not work well; players did not enjoy the experience.

Instead, tag-a-tune was developed, based on the ESP game, but with two important variations: (1) ***Both players can see each other's guesses***; (2) ***Players may or may not be listening to the same song***.

The objective of the game is to ***decide whether both players are listening to the same tune***; note that unlike the ESP game, in this case the system knows the ***correct answer***, that is, whether both players are listening to the same tune.

If both players guess correctly, they score points and move on to a new round.


gwap
ESP Game
Tag a Tune
Verbosity
Squigl
Matchin
Flipit
PopVideo
Tag a Tune
Hear Here
Top Scores
Today
This Month
All Time
1 Sunshine 22M
2 pook 19M
3 Catwoman 5 637K
4 occam 4 481K
5 greggig 3 003K
Got it, Let's Play!
View Instructions

Scoring Mechanism

Tag-a-tune is a cooperative game and players only score points if they both guess correctly; neither gains any points if only one guesses correctly.

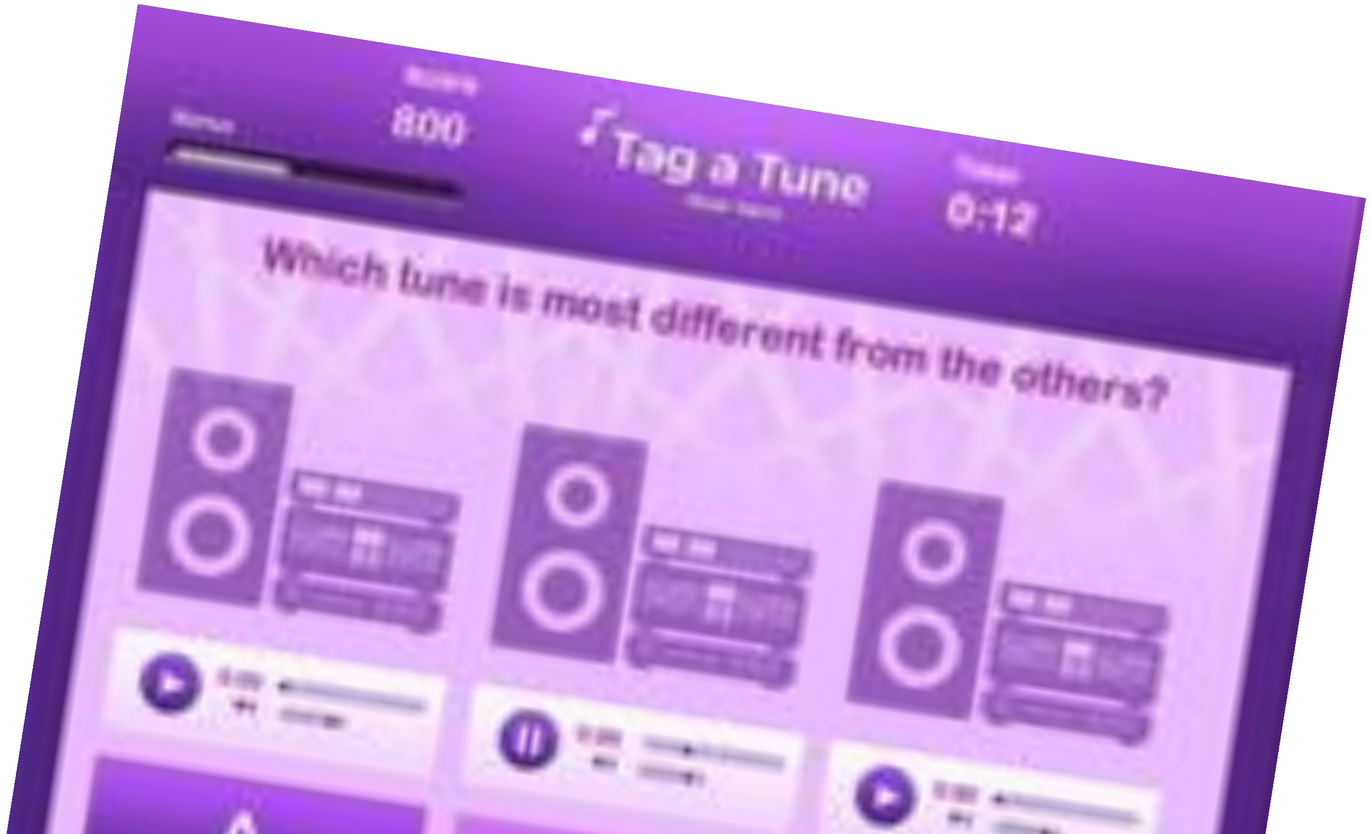
This provides a natural incentive for players to behave honestly and act truthfully which in turn leads to labeled data that is more likely to be accurate.

The Bonus Round at 1000 points

Players are asked to listen to three pieces of music and must decide which one of the three clips is most different from the other two. If they agree, they both obtain points.

Bonus rounds produce 2 types of new data

Similarity data for music is useful for powering and improving music recommendation systems. The similarity between songs is potentially a good indication of the level of difficulty that a particular pair of songs would present during a normal round of tag-a-tune.



Evaluation

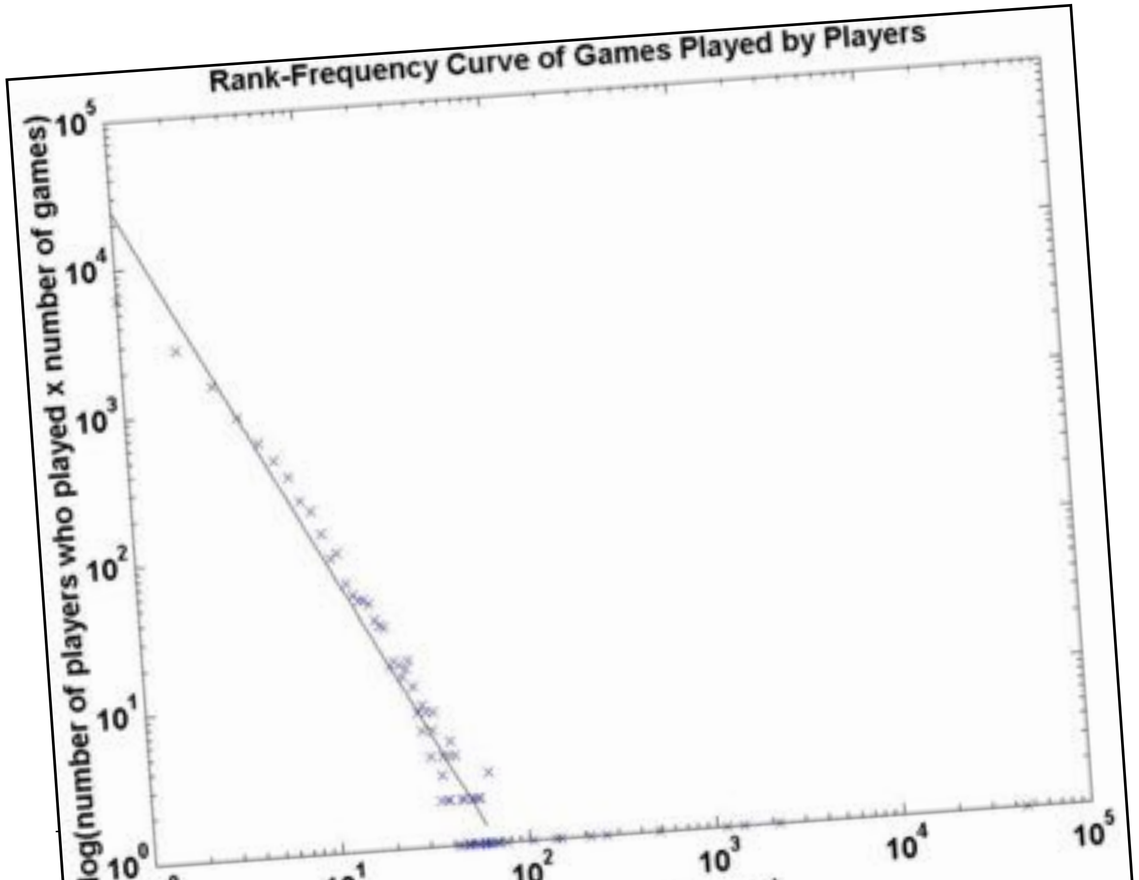
Based on seven months worth of data from May 2008.

A total of 49,088 unique games played by 14,224 unique players, corresponding to 439,760 rounds.

The number of games each person played ranged from 1 to 6,286, and the total time each person spent in game play ranged from three minutes to 420 hours.

The average number of games played was four.

Players only passed on 0.5% of rounds.



Successful vs Failed Rounds

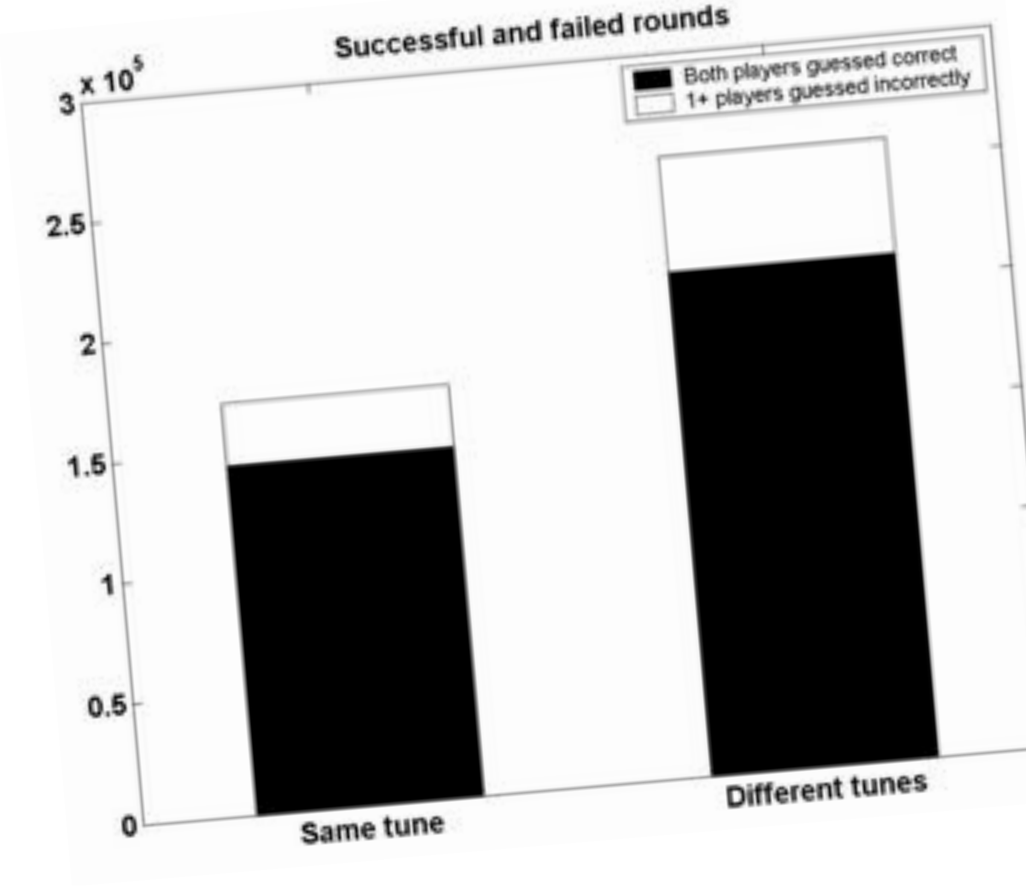
Completed vs Missed Rounds

97.36% of the rounds, both players voted same or different before the end of the round. The remaining 2.64% are called missed rounds.

80% of completed rounds were successful meaning that both players guessed correctly.

The success rate for rounds in which the tunes were the same was 85%.

The success rate for rounds in which the tunes were different was 81% ⇒ harder to distinguish between tunes that are different.



Tag Statistics

512,770 tags collected, of which 108,558 were verified by at least two players and 70,908 were unique.

Based on this, the average number of tags generated per minute of play is approximately four.

Most common tags are for genre, instrumentation, or aspects of the music itself.

Some non-relevant ***communication*** tags (no, same, diff).

Some ***negation*** tags describing what is not present in the music.

Tag	Count	Tag	Count
classical	37,781	no vocals	6,126
guitar	30,093	soft	5,642
piano	27,718	sitar	5,413
violin	19,525	no vocal	5,285
slow	18,485	classic	5,228
strings	17,484	male	5,216
rock	17,413	singing	5,059
techno	15,627	solo	5,047
opera	14,512	vocals	5,014
drums	13,667	cello	4,966
same	12,610	loud	4,957
flute	12,149	woman	4,321
fast	11,435	pop	4,213
diff	11,046	male vocal	3,951
electronic	10,333	choir	3,576
ambient	8,733	violins	3,454
beat	7,683	new age	3,390
yes	7,352	beats	3,387
harpsichord	7,261	no voice	3,252
indian	7,255	harp	3,172
female	7,071	voice	3,080
vocal	6,964	weird	3,056
no	6,659	instrumental	2,946
synth	6,530	dance	2,896
quiet	6,167	female vocal	2,873

Table 1. Head List: top 50 most frequently used tags

Tune-Based Statistics

30,237 audio clips annotated and 108,558 verified (confirmed by at least two players) tags collected.
Note: verified is used to refer to tags that have high confidence (because they have been independently generated by multiple players) and unverified refers to tags that have low confidence.

92% of the audio clips have been annotated by two or more players, 61% have been annotated by ten or more players, and 26% have been annotated by 20 or more players.

Even using a simple random selection strategy for picking songs to present to players, most songs are evaluated by multiple players.

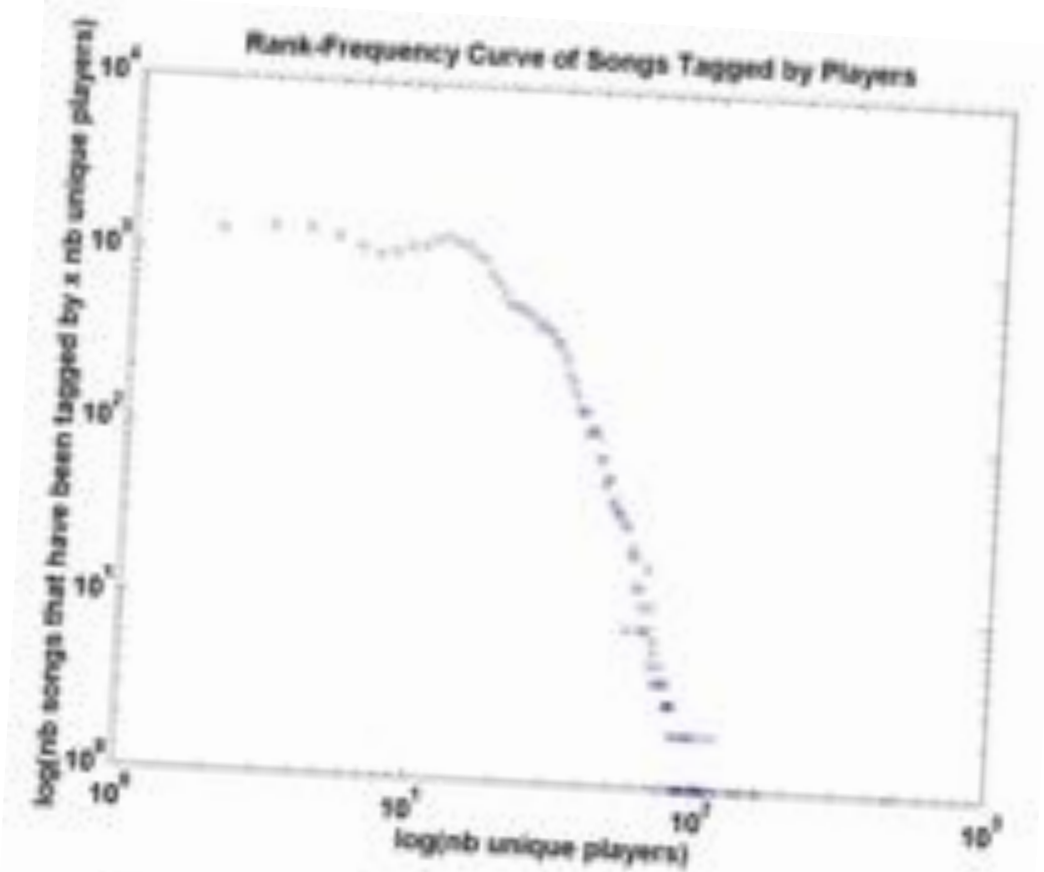


Figure 11. Number of songs that are tagged by x number of unique players

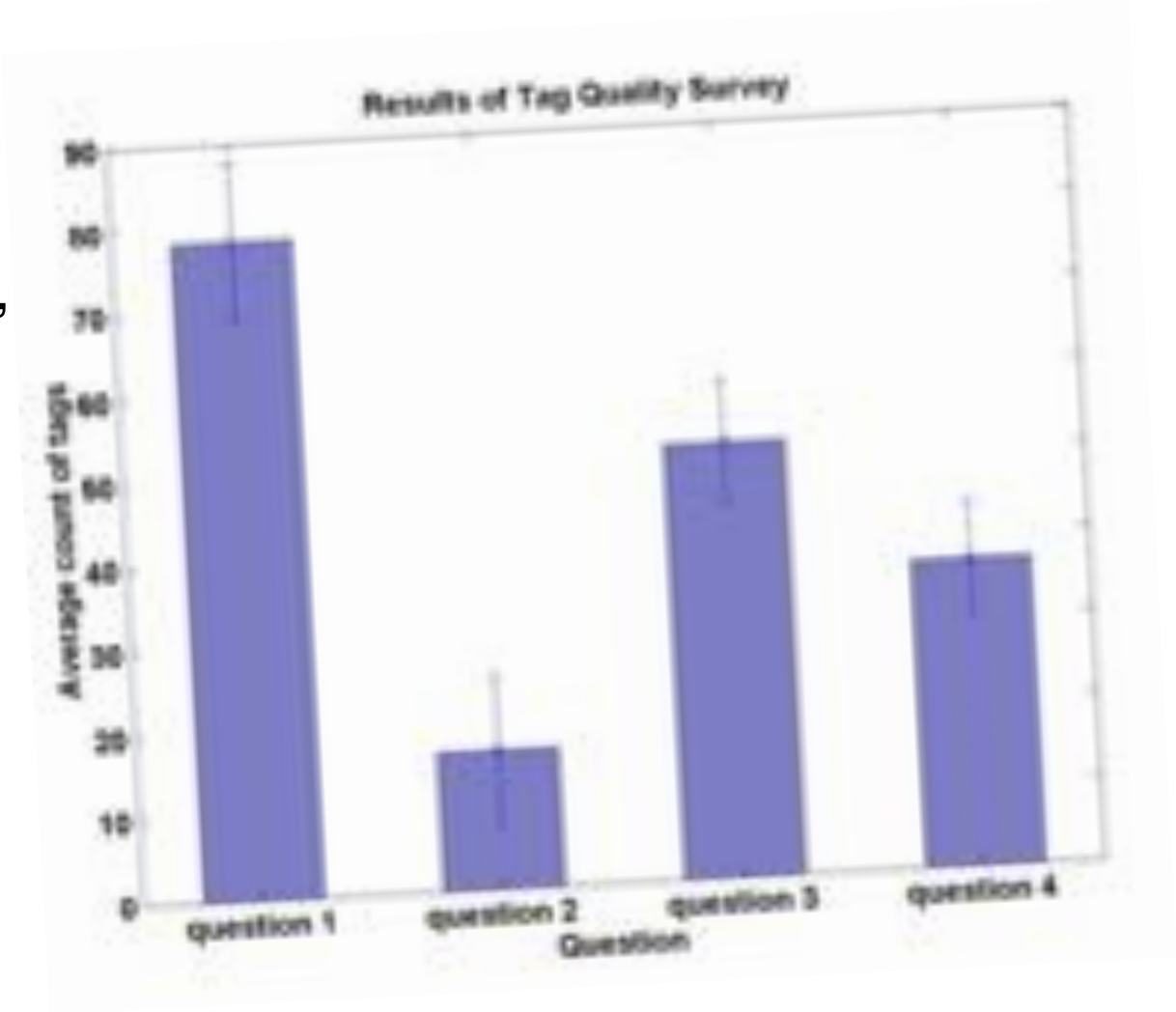
Tag Quality

20 clips with at least five verified tags chosen at random, each with an average of 7 verified tags and 17 unverified tags.

100 participants asked to answer 4 questions, 2 for verified and 2 for unverified tags:

1. Which of the following tags would you use to describe the piece of music to someone who could not hear it?
2. Which of the following tags have *nothing* to do with the piece of music (i.e., you don't understand why they are listed with this piece of music)?

Q1 > Q3 means a larger proportion of verified tags than unverified tags are useful to describe the clips.



Example 4 – Bookmarking & Tagging

The Dogear Game

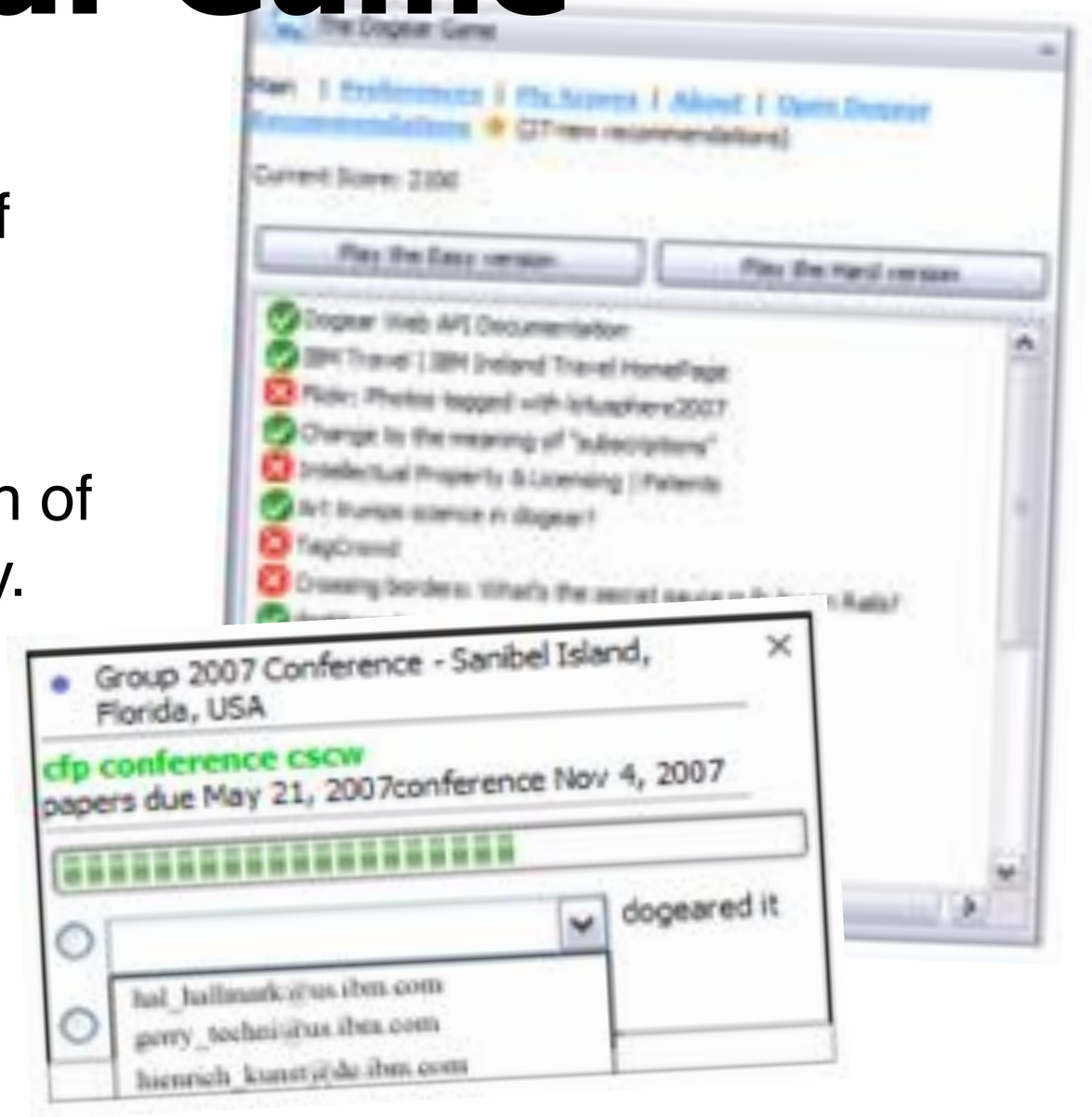
Aim is to generate human-sourced bookmarked recommendations to users of the Dogear social bookmarking system.

The player is presented with a bookmark (URL and tags) and asked to predict which of their social connections submitted it, if any.

Bookmarks sourced from Dogear.

Players win points by guessing the correct source.

What do incorrect guesses provide?



Collabio - Annotating/Tagging People

Tagging game embedded in Facebook.

The basic idea of Callabio is to encourage people to annotate/tag their friends.

Players see a tag-cloud for their friend but with the tags obscured. The object of the game is for the player to guess the tags that relate to their friend.

If they guess correctly, their tag is revealed and they score points. If they guess incorrectly then this provides a new tag for a future game by another player.

Scores are used to encode a “People you know best” type feature which motivated others to play.

Example


Greg Smith
Stanford Alumnus/Alumna
Microsoft
Tag Greg to reveal each hidden item. One point for each tag, another point for each other friend who used the same tag to describe Greg!
Tag!
Greg's friends have tagged him with:
band
cruise
dogs
hacker
microsoft
msrs
msr
poker
smoky
stanford
vibe
Choose someone else:
Start typing a friend's name
Go
People who know Greg best:
Amy Karlson
96 points
You
85 points
Raman Sarin
83 points
My Score: 85 points
microsoft
12 points

Example 5 – Avoiding Web Spam

The Web Spam Game

The objective is to try to identify web pages / search results that are poor quality and ultimately to collect voting information to decide whether or not to move the page up or down the search rankings.

The game is based around a series of independent questions consisting of a query, Q, and a piece of snippet text from a random result retrieved for Q.

Players do not see the page in question or know of its rank in the result-list for Q.

The players must decide whether the page appears to be relevant or not relevant. if both players agree they receive points. If they disagree they loose points. if they pass there is no point change.

The votes of players are aggregated to determine ultimate page relevance.

For example...

Query: ice age 2
Snippet:
Ice Age 2
Official site. Help the Scrat find enough acorns to survive the Ice Age. Meet the Sub Zero Heroes, view a trailer, download desktop wallpaper, ...

Highly relevant | Not Relevant | Pass

Query: ice age 2
Snippet:
Strange Horizons Articles: Interview: Glen Cook, by Donald Mead
DM: Why do you think the Black Company series is so popular among soldiers? ... There's also an ice age encroaching, which is making world sea levels drop ...

Highly relevant | Not Relevant | Pass

Example 6 – Acquiring Common-Sense Knowledge

Common-Sense Reasoning and AI

Many major efforts to assemble a knowledge-base of common sense facts to support automated common-sense reasoning.

E.g. Cyc was a major decade-long effort to represent and encode common-sense facts from paid experts ⇒ 1M+ facts but still does not provide sufficient coverage

except in niche domains.

Collective Intelligence Approaches ⇒ OpenMind and MindPixels projects provide

a collaborative platform for ordinary web users to participate in common-sense knowledge capture. And of course Wikipedia can be viewed as a related initiative.

What about building a GWAP to capture common-sense knowledge?

Verbosity

Online, two-player game.

One player is the *Describer* or *Narrator* while the other is the *Guesser*; these roles switch after each round.

The narrator must try to describe a target word using a set of predefined descriptive templates. And the guesser tries to guess the word. The templates help when it comes to disambiguation, categorization, and parsing and also add to the gameplay.

A cooperative scoring system is used: both players score points when the guessers enters the correct word.

A player-bot can be used to facilitate a single player game.


Verbosity
It's common sense.
Top Scores
Today
This Month
All Time
1 PlasticBiddy 181
2 Catwoman 153
3 Jeff 137
4 webstercats 133
5 misterf 72
Got it, Let's Play!
View Instructions

Evaluation

Enjoyability & Playability

A total of 267 people played the game in a period of 1 week, generating 7,871 facts.
On average, each player contributed 29.47 facts and each person played for an average of 23.58 minutes in one sitting, and some played for over 3 hours.

Quality of Data

Selected at random 200 facts collected using Verbosity and asked the following question to six different raters: Is this sentence true?
Overall, 85% of the sentences collected were rated as true by all six raters.
Many of the sentences not rated as true by all were debatable; for example, Buddha is a kind of god.
Strong evidence to support the notion that Verbosity if capable of collecting high-quality common-sense facts.

Example 7 – Protein Folding

FoldIt - Solving Puzzles for Science

An online game about protein folding.

The process by which protein molecules “fold” into a functional three dimensional structure is complex and only partially understood.
The general process and basic constraint framework is known but the protein structure prediction is much more demanding.

The Game Structure

FoldIt players manipulate 3D molecular models in an effort to product structures that conform to the know rules and constraints of protein folding.
Players are led through a series of tutorials to familiarise themselves with these rules on simple structures.
Ultimately the game relies on the natural ability of humans to manipulate 3D structures under a set of complex structural constraints.
Players are scored based on their designs and compete for high-scores etc.


Full Mode
Score:
5571
of 8110
Hide the Hydrophobics
Orange sidechains are HYDROPHOBIC.
They want to be buried inside the protein.
Tell me more...
Reset Puzzle

Does it work?

Nature Structural and Molecular Biology (2011)

NSMB published an article (Khatib et al, Nature Structural & Molecular Biology 18, 1175-1177 (2011)) that revealed the structure of an enzyme used by retroviruses similar to HIV. The achievement was widely viewed as a breakthrough. Who solved the riddle? A bunch of FoldIt gamers!

Scientific American (2012)

Reported that the Foldit gamers achieved the first crowdsourced redesign of a protein. The protein is an enzyme which catalyses the Diels-Alder reactions widely used in synthetic chemistry. A team including David Baker in the Center for Game Science at University of Washington in Seattle computationally designed this enzyme from scratch but found the potency needing improvement. The Foldit players reengineered the enzyme by adding 13 amino acids and increased its activity by more than 18 fold.

Example 8 – Eliciting User Preferences

Learning Recommendation Knowledge

Collecting recommendation knowledge is costly and difficult (e.g. user preferences, item descriptions, user similarities …)

Can we use ideas from GWAPs to capture recommendation knowledge as a side-effect of gameplay?

What type of recommendation knowledge could we learn?

If it’s new knowledge how will we validate it?

Basic Gameplay


Time:
20s
player's overall score for this game
Score:
40
movie posters float across screen
player's score when they match a movie and a friend correctly (with exploding popcorn)
+10
player's friends

Friends(u)
Likes(u)
Users
Likes(u)
Likes(f)
Recommendation Engine
Game Engine
match(p,m,f)
knows(p,f)
interest(m,f)
Game
match(p,m,u)
interest(m,u)
Poster(m)
TopMovies
Movies
SimMovies(m)
Poster(m)
SimMovies(m)
match(p,m,f)
score(p,m,f)
Likes(f), PastMatches(m,f)

Basic Gameplay

In each game a player, ***p***, is presented with a set of friends' avatars, f_1, …, f_n (currently $n = 5$) Moreover, a set of movies, $m_1,\ldots,m_k$, are selected based on the likes of friends and including additional popular movies not among the friends' likes.

To make the game more challenging, the movie posters follow different trajectories across the screen, becoming more erratic as the game progresses.

The player has a limited time to make as many matches as they can. Each match is rewarded with a graph- ical and audible flourish (the friend's avatar explodes in a fountain of popcorn) and the player receives a variable score.

Matches

Obviously the objective of the game is for ***p*** to ***match*** a movie ***m*** with some friend ***f***, ***match(p,m,f)***, and to generate as many of these matches as possible.

The set of matches generated during a game is given by ***Matches(p)***, as below.

$$Matches(p) = \bigcup_{\forall f \in Friends(p)} \{(p, m, f) : match(p, m, f)\}$$

From Matches to Recommendation Data

Each match, ***match(p,m,f)***, is either a ***known*** match or an ***unknown*** match.

match(p,m,f) is a ***known*** match if ***m*** $\in$ ***Likes(f)*** *and* ***match(p,m,f)*** is an ***unknown*** match otherwise.

What do we learn from known and unknown matches from a recommendation viewpoint?

Known Matches

Known matches tell us about ***p's*** understanding of ***f's*** movie interests.

$$knows(p,f) = \frac{|KnownMatches(p,f)|}{|FriendMatches(p,f)|}$$

If ***p*** generates a lot of known matches for f then she must have a good understanding of ***f's*** movie preferences.

$$KnownMatches(p,f) = \{(p,m,f) \in FriendMatches(p,f) : m \in Likes(f)\}$$

We can use the proportion of known matches to estimate ***knows(p,f).***

$$FriendMatches(p,f) = \{(p,m,f^*) \in Matches(p) : f = f^*\}$$

Unknown Matches

Unknown matches are not failed matches; ***Likes(f)*** is not exhaustive.

If ***match(p,m.f)*** then we can assume that ***p*** believes ***f*** will like ***m***.

m *is a potential (and novel)* ***recommendation candidate*** for ***f.***

If other players also match ***m*** with ***f*** then this increases the likelihood that ***m*** is relevant to ***f,*** especially if these players know ***f*** well.

$$interest(m,f) = \sum_{\forall p : match(p,m,f) \wedge m \notin Likes(f)} knows(p,f)$$

Scoring

Scoring is key in GWAPs. Scoring is a signalling system.

How a player's score changes during the game provides important feedback to the player and incentivises the right type of play.

Scoring must be aligned with the type of play that is desired.

We need to encourage the player to make lots of matches and generate new recommendation data.

Need to credit both known and unknown matches.

Known matches are easy to score (known to be correct) but arguably less valuable that unknown matches, which are by definition more difficult to score.

Scoring

Balancing known and unknown match scores with α (= 0.5).

Known matches score 5 points; unknown matches score up to 5 points (for the first player that creates the match) but fewer points as more players generate the same unknown match.

$$score(p,m,f) = 10 \bullet \left(\alpha \bullet 1[m \in Likes(f)] + \frac{1-\alpha}{1+PastMatches(m,f)} \right)$$

Recommendation Strategies

So far focused on using gameplay to generate (match) data.

Next look at how this data can be used to estimate user relationship strengths and recommendation candidates as part of a crowdsourced recommendation strategy.

Compare to benchmark collaborative filtering and content-based approaches.

Crowdsourced Recommendations

Using crowdsourced data to generate and rank candidate recommendations for some target user u_t.

These candidates are novel movies that were matched with u_t during gameplay by some player-friend, p.

$$CS\_Candidates(u_t) = \bigcup_{\forall p \in Friends(u_t)} \{m : match(p, m, u_t) \wedge m \notin Likes(u_t)\}$$

Finally, we rank these candidates by decreasing value of $interest(m,u_t)$.

Collaborative Filtering Recommendations

Implement variation on ACF by choosing movies from the ***Likes*** of u_t's friends, but with are not yet liked by u_t.

$$CF\_Candidates(u_t) = \bigcup_{\forall f \in Friends(u_t)} Likes(f) - Likes(u_t)$$

Normally ACF ranks based on ***sim(f,u_t)*** but here we use ***knows(f,u_t)*** as a proxy for similarity so that once again sort by decreasing ***interest(m,u_t)***.

Content Based Recommendations

Content-based recommendation implemented using Rotten Tomatoes API and ***movie_similar(m)*** function; returns 5 movies similar to ***m***.

Using u_t's own ***Likes*** as seeds to generate a set of similar movies from RT.

$$CB\_Candidates(u_t) = \bigoplus_{m \epsilon Likes(u_t)} movie\_similar(m)$$

Then select movies at random from this set for recommendation; note duplicates from RT more likely to be selected.

Evaluation

27 participants (18-30 year-old males and females, undergrads and post-grads) based on 3 groups.

Group 1 = 15 friends; Group 2 = 6 friends; Group 3 = 6 peers (that is, not close friends).

Each group acted as players and friends to each other.

Phase 1 - gameplay & data collection.

Each player was asked to play as many games as they liked with a different selection of 5 friends where possible.

Phase 2 - recommendation testing based on 3x strategies.

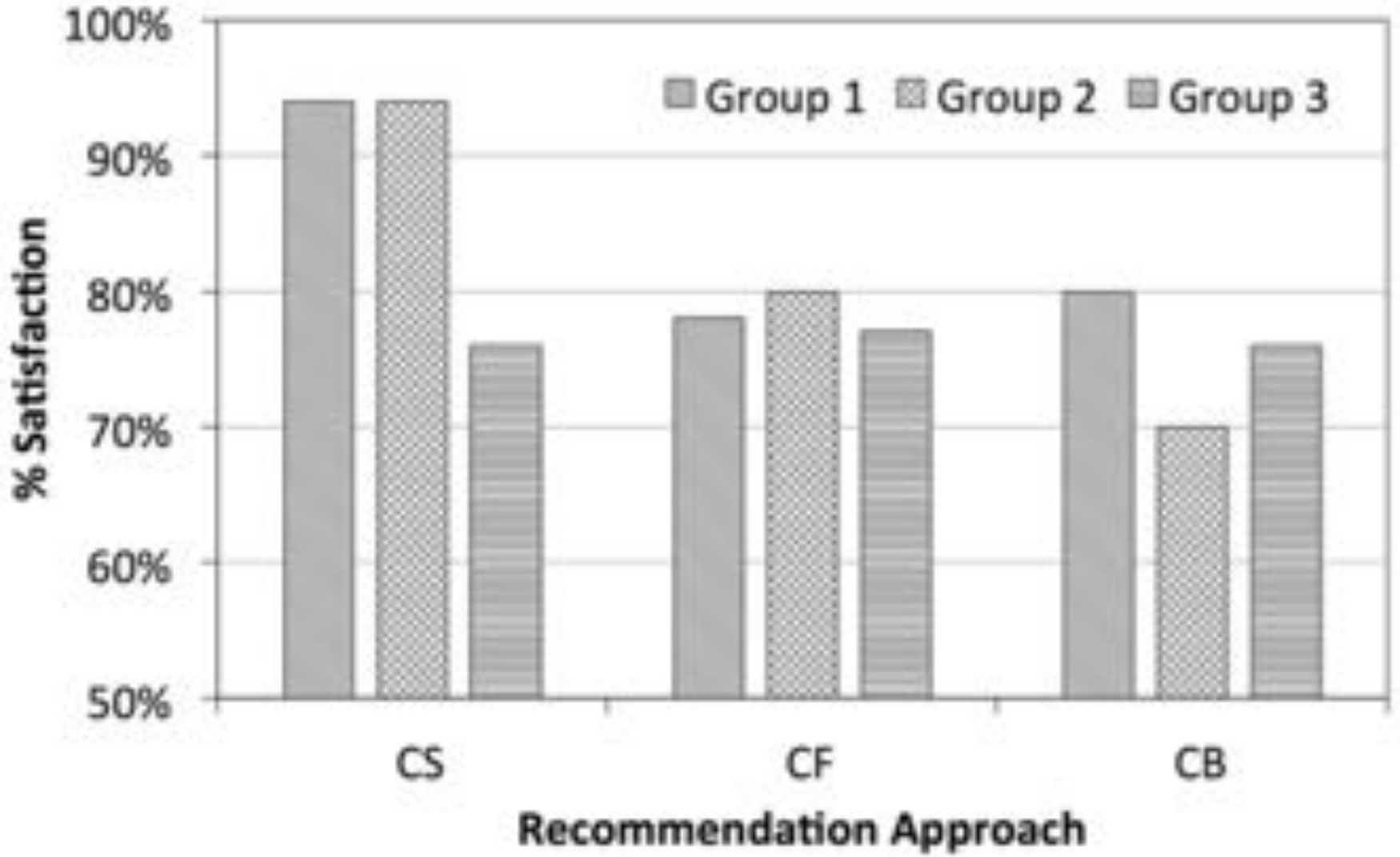
Each user received balanced mix of 18 recommendations and were asked to rate their satisfaction with each recommendation.

Phase 1 - Gameplay Stats

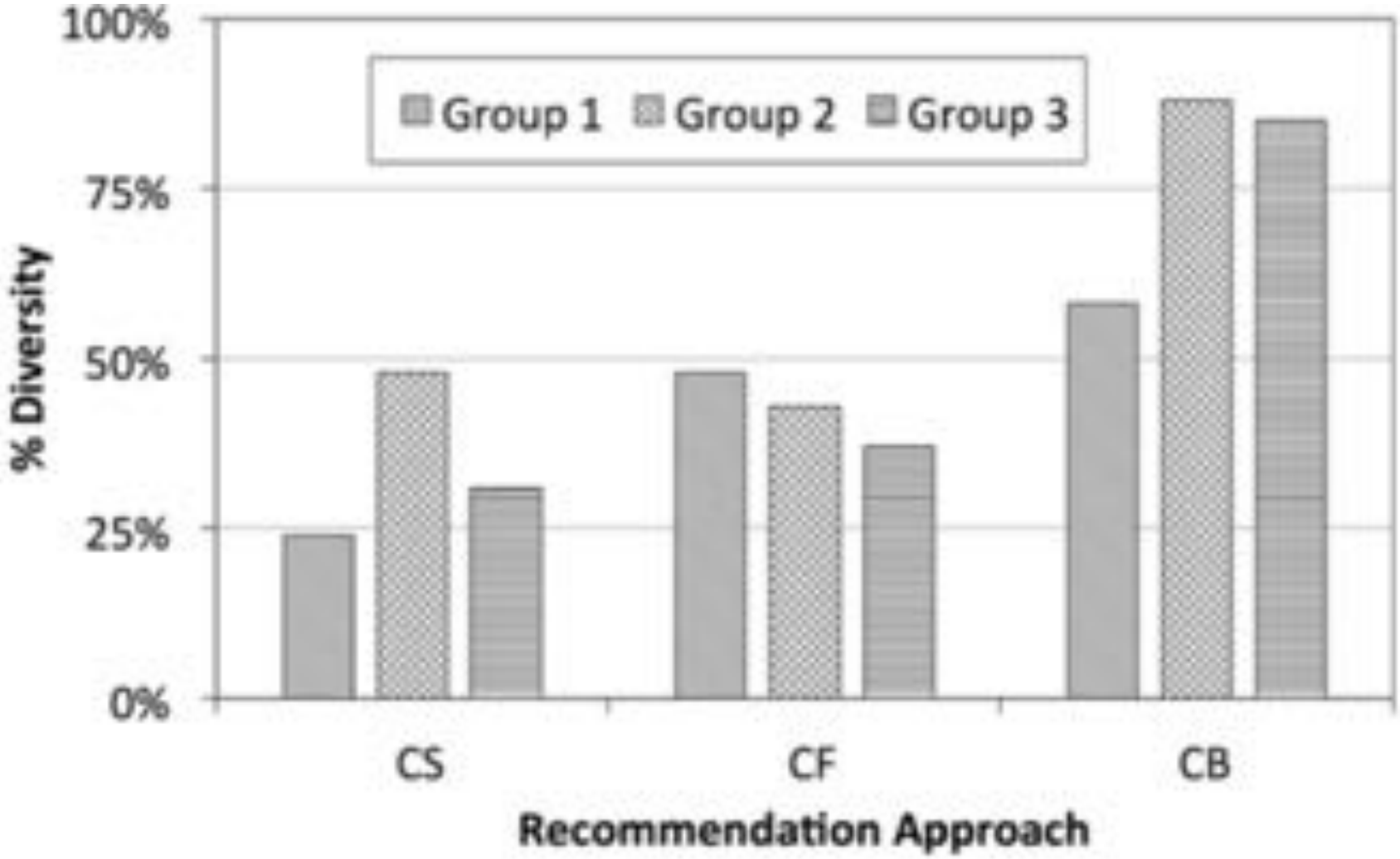
Gameplay Stats	G1	G2	G3	Mean
Members/Group	15	6	6	-
Movies/Profile	20	11	13	16.4
Games/Player	15	5	7	11
Matches/Game/Player	10.85	11.03	13.2	11.4
Known/Game/Player	2.76	3.93	3.07	3

Phase 2 - Evaluating Recommendations

Satisfaction Scores



Recommendation Diversity



Conclusions

Explore the use of a GWAP to collect recommendation data and user preferences as a side effect of casual gameplay

Implemented/tested a simple movie matching game & described the recommendation data that we can collect from its gameplay.

Data can be used in a recommendation context and we have demonstrated its potential as part of a live-user trial.

Lots of limitations …

Incomplete prototype, single domain, small-scale evaluation, etc.

… but promising first step and well received by recommender systems research community.

Parting Thoughts

Creating a compelling, enjoyable, and useful GWAP is not as easy as it looks!

Matching a challenging problem with compelling gameplay is far from easy.

And motivating users with the right mixture of gameplay and incentives is more art than science.

But, … getting it right has a significant upside in terms of the potential scale of popular games.

Thinking about Scale

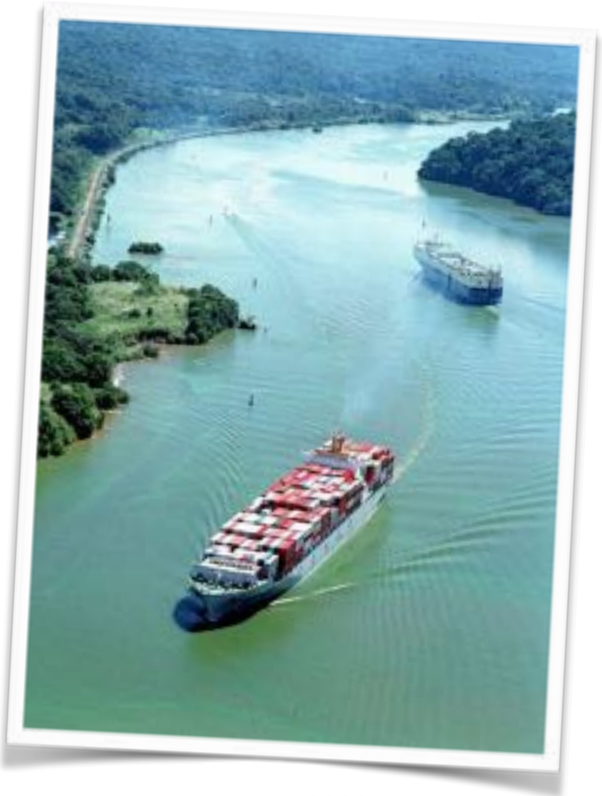
People love games!

We spend 3Bn+ hours/week playing online games…

… that's a significant amount of intellectual effort!

Imagine if some of this game-play could be harnessed to solve really hard problems!

7 million person hours



20 million person hours

Watch and Learn!

Jane McGonigal: Gaming can make a better world

http://www.ted.com/talks/jane_mcgonigal_gaming_can_make_a_better_world.html

Gabe Zichermann: How games make kids smarter

http://www.ted.com/talks/lang/en/gabe_zichermann_how_games_make_kids_smarter.html

Jesse Schell: When games invade real life

http://www.ted.com/talks/jesse_schell_when_games_invade_real_life.html

Seth Priebatsch: The game layer on top of the world

http://www.ted.com/talks/seth_priebatsch_the_game_layer_on_top_of_the_world.html

Reading List

Edith Law, Luis von Ahn: Input-agreement: a new mechanism for collecting data using human computation games. CHI 2009: 1197-1206

Severin Hacker, Luis von Ahn: Matchin: eliciting user preferences with an online game. CHI 2009: 1207-1216

Luis von Ahn, Laura Dabbish: Designing games with a purpose. Commun. ACM 51(8): 58-67 (2008)

Luis von Ahn, Ruoran Liu, Manuel Blum: Peekaboom: a game for locating objects in images. CHI 2006: 55-64

Luis von Ahn, Shiry Ginosar, Mihir Kedia, Ruoran Liu, Manuel Blum: Improving accessibility of the web with a computer game. CHI 2006: 79-82

Luis von Ahn: Games with a Purpose. IEEE Computer 39(6): 92-94 (2006)