

and human computation!

Crowdsourcing^y and

Its Application in Cultural Heritage

Andrew Huynh

August 20, 2013

the
BIG
picture

Who cares?

Why Crowdsourcing?



\$30k in rewards per day
150k tasks completed per day

<http://mturk.com>



Over 240,000 players
Millions of folded proteins

<http://fold.it>



Over 10,000 players
2.4+ million annotations
3.4 man years of exploration.

<http://exploration.nationalgeographic.com>

Why Crowdsourcing?

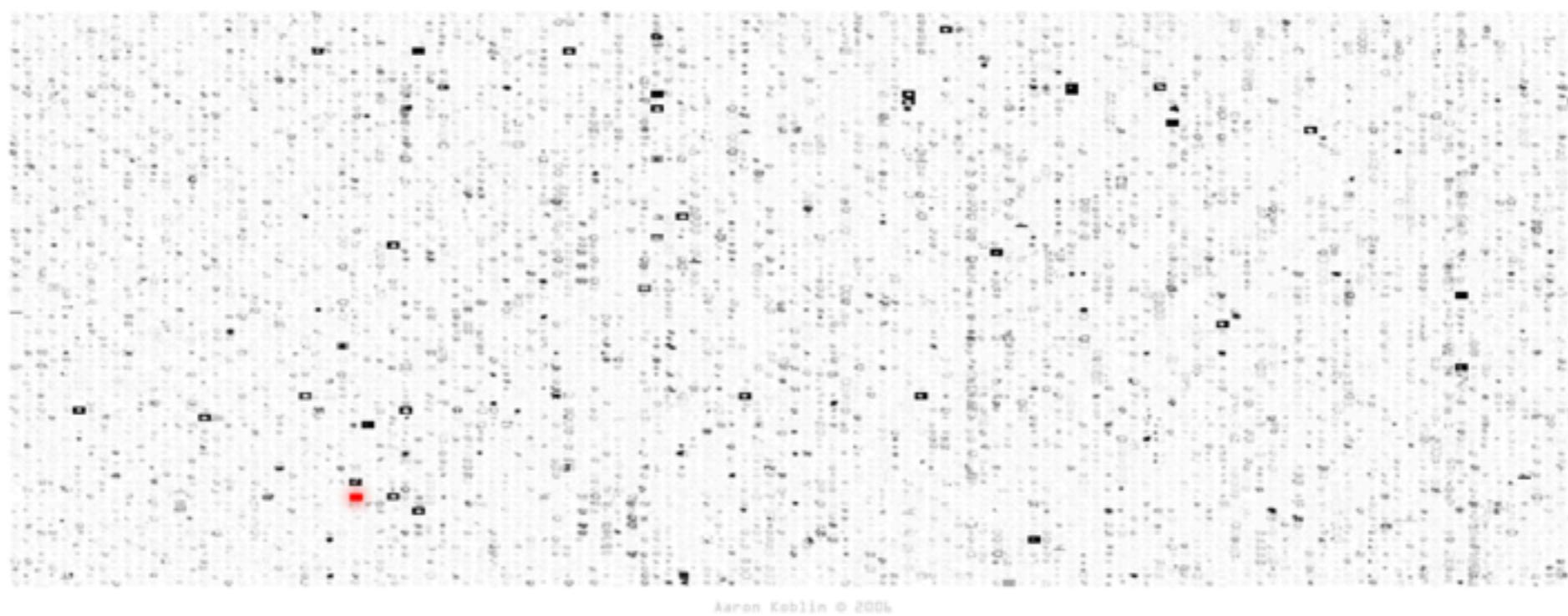
My personal favorite

THE SHEEP MARKET

10,000 sheep created
by online workers.
[More...](#)



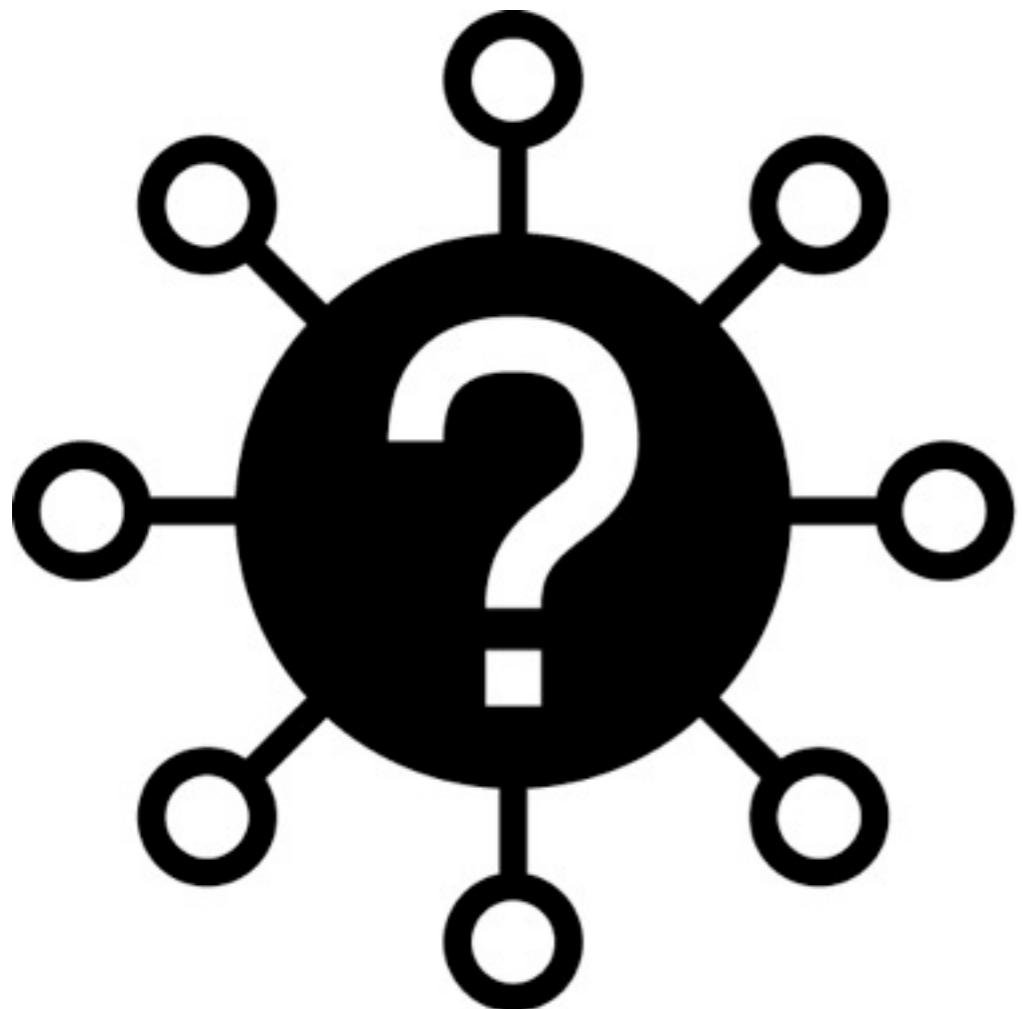
2228 / 10000



40 days. 10,000 sheep. 11 sheep/hour.

Why Crowdsourcing?

What is crowdsourcing?



Outsourcing + crowd

Distribution of tasks

Harness human computation

Why Crowdsourcing?

What is human computation?



Intuition

Creativity

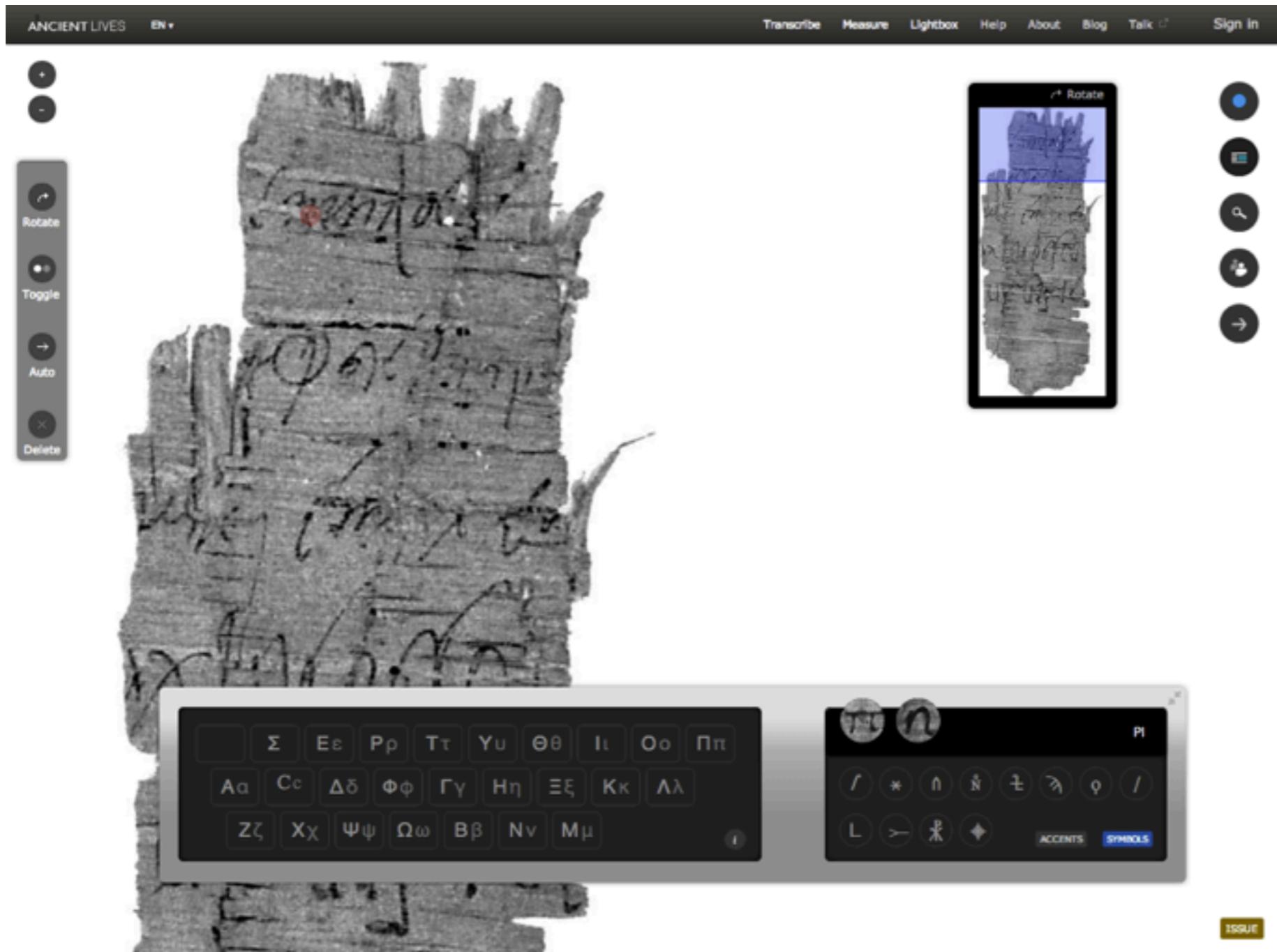
Symbolic Reasoning

Why Cultural Heritage?



Why Cultural Heritage?

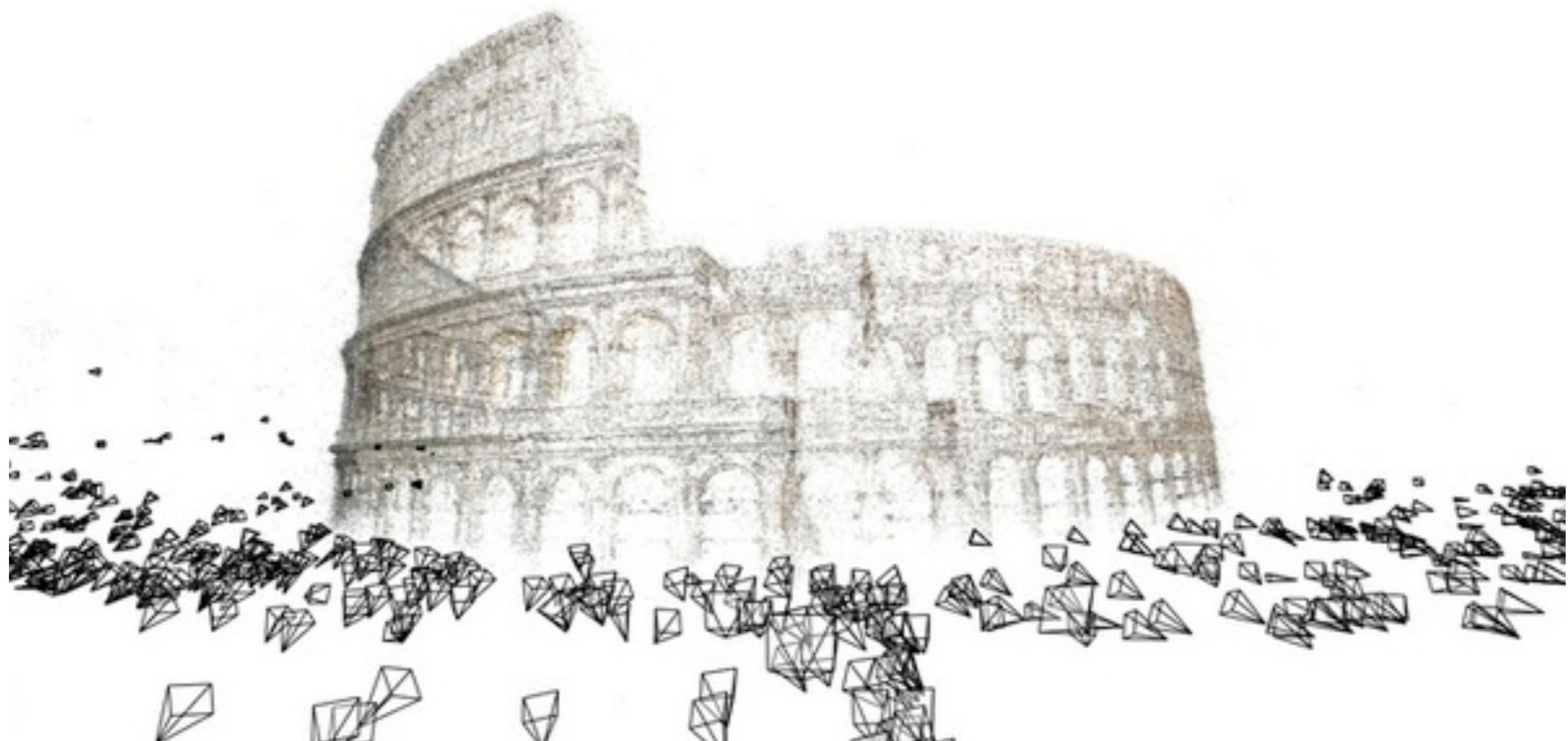
Digitization & Preservation



<http://ancientlives.org>

Why Cultural Heritage?

Digitization & Preservation



Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., & Szeliski, R.
Building Rome in a day. ICCV (2009)

Why Cultural Heritage?

Exploration & Discovery

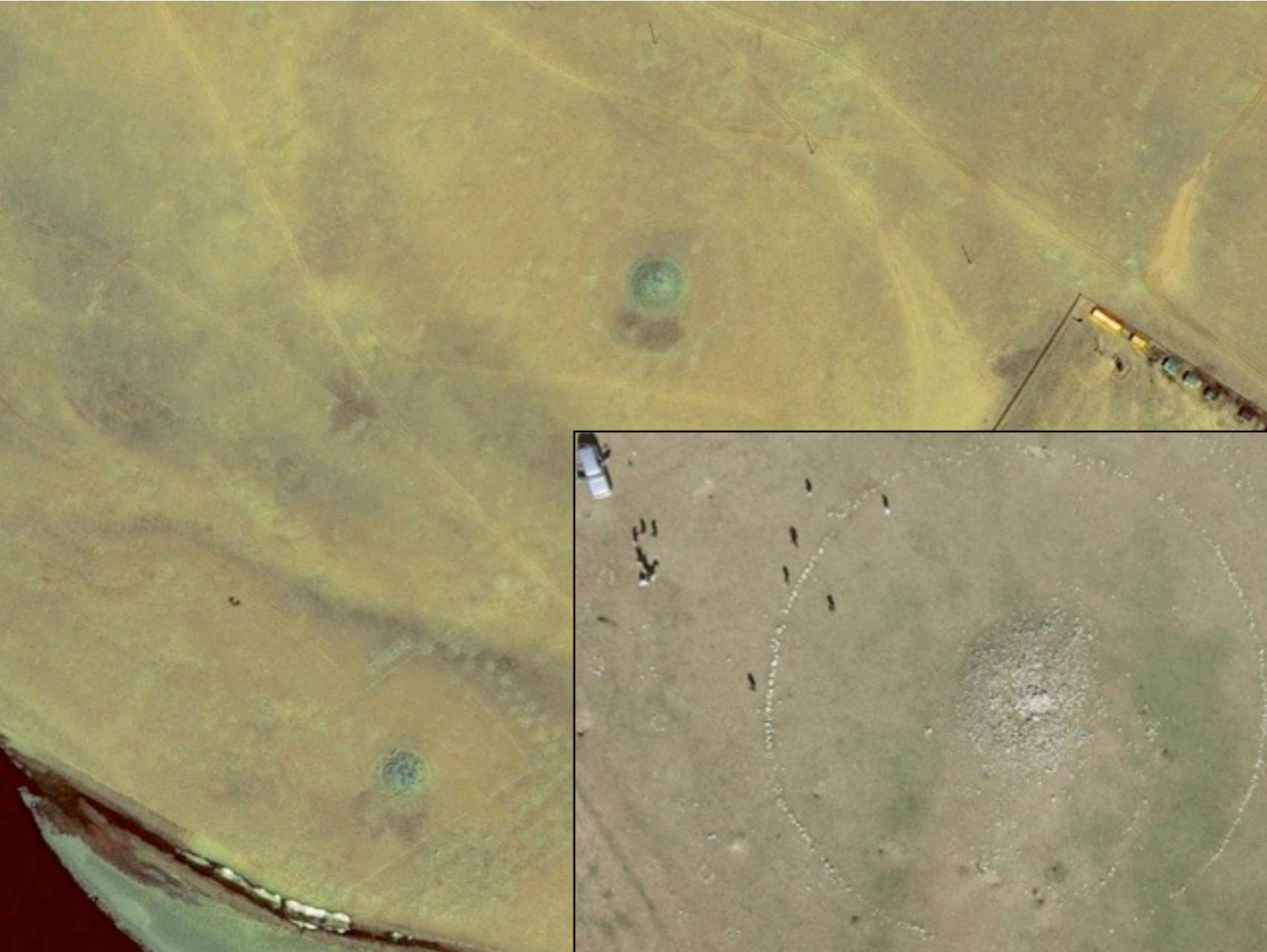


Corrie, Robert K.

Detection of ancient Egyptian archaeological sites using satellite remote sensing and digital image processing. Proc. of SPIE Vol. Vol. 8181. (2011)

Why Cultural Heritage?

Exploration & Discovery



Lin, A., Huynh, A., Lanckriet, G., Barrington, L.

Crowdsourcing the Unknown: The Search for Genghis Khan. 2013 (In review).

Active Crowdsourcing

Crowd Engagement

Data Quality Control

Passive Crowdsourcing

Parasitic Computing

Crowdsensing

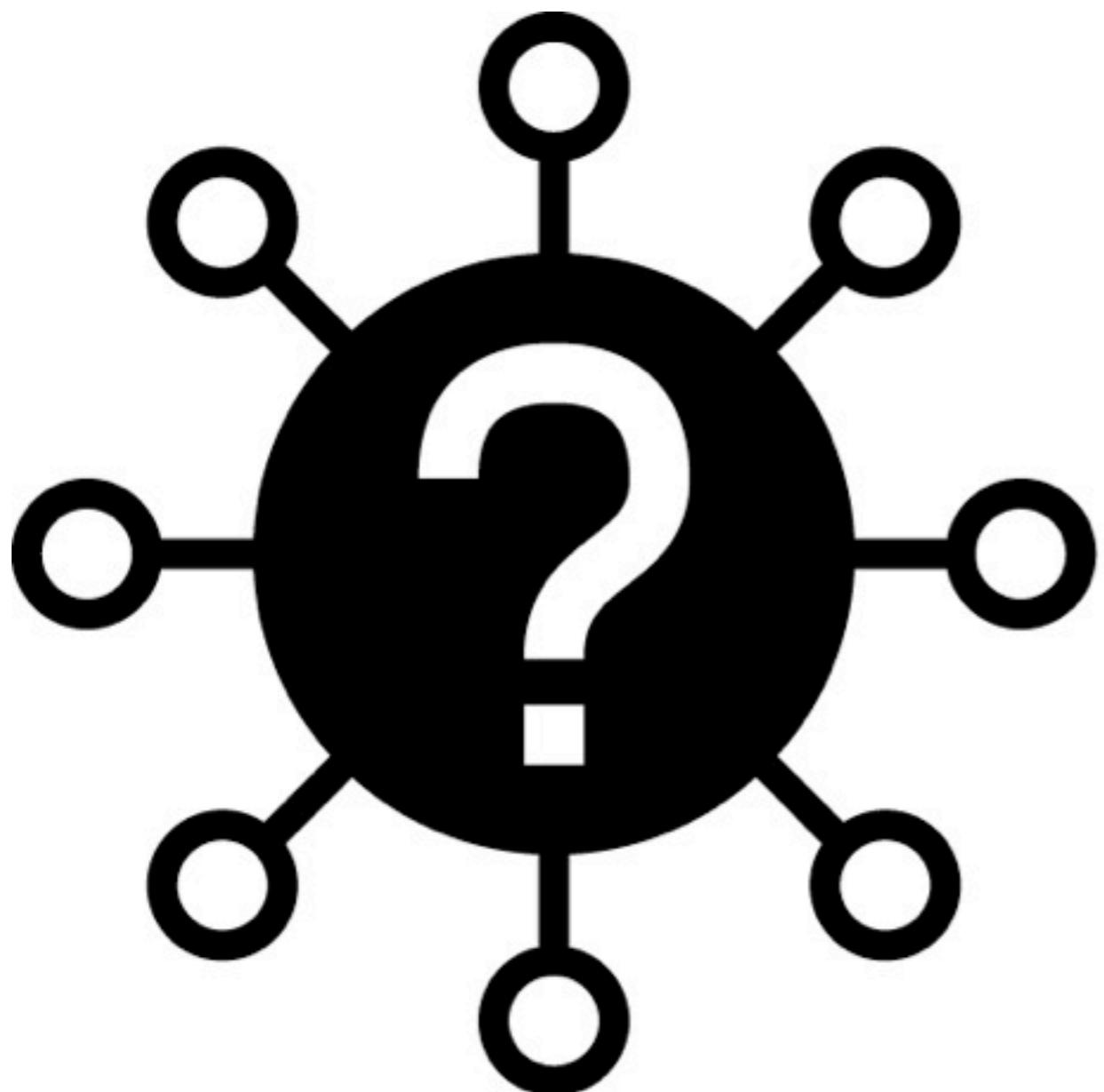
Learning from the Crowd

Active Learning

Closing the Loop

Applications in Cultural Heritage

Active Crowdsourcing

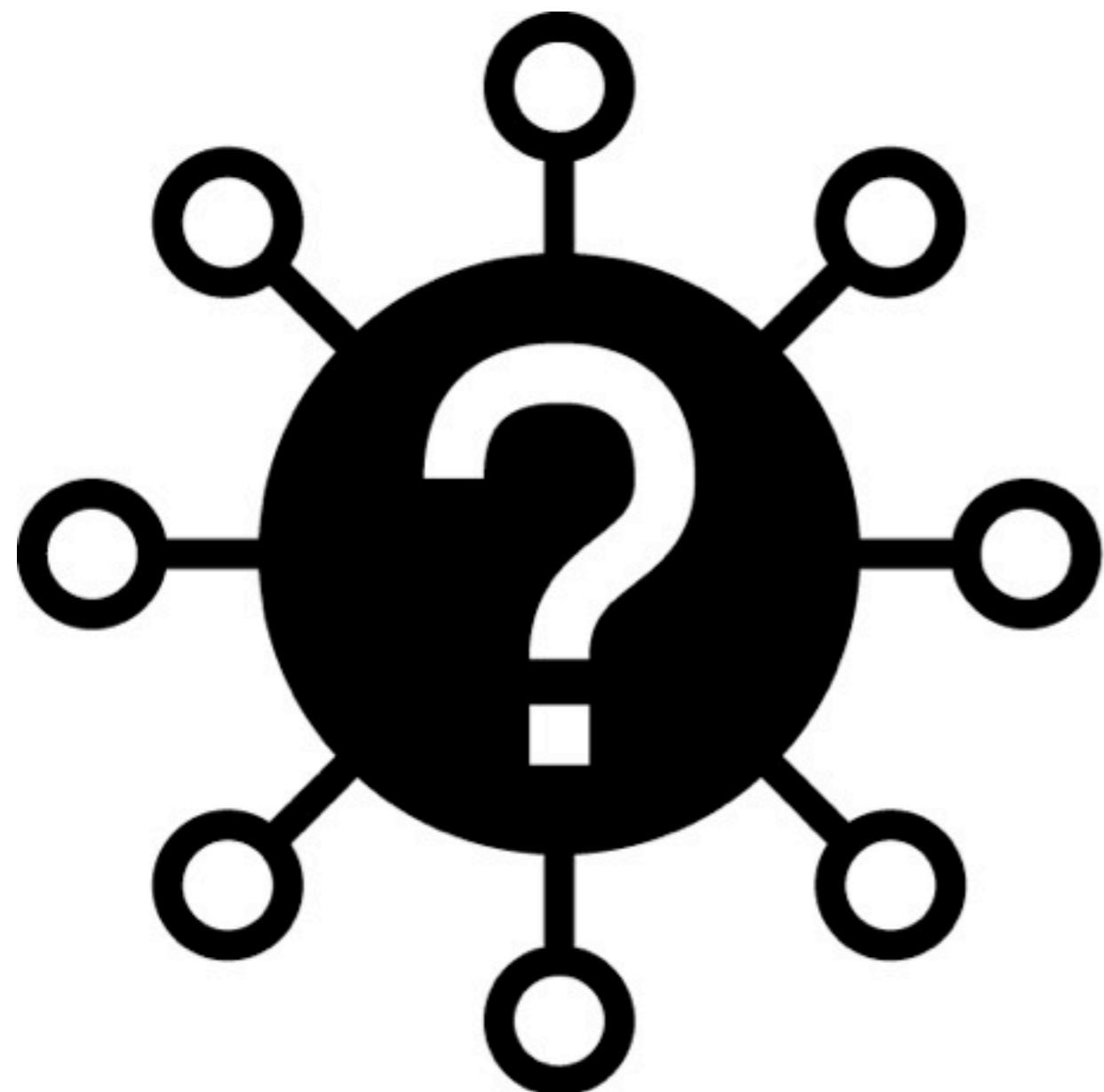


Distributed tasks

Crowd = Pool of labor

Aggregate solutions

Active Crowdsourcing



Distributed tasks

Creating engaging tasks

Crowd = Pool of labor

Keeping the crowd motivated

Aggregate solutions

Data quality control

Active Crowdsourcing

Crowd Engagement & Motivation



Value



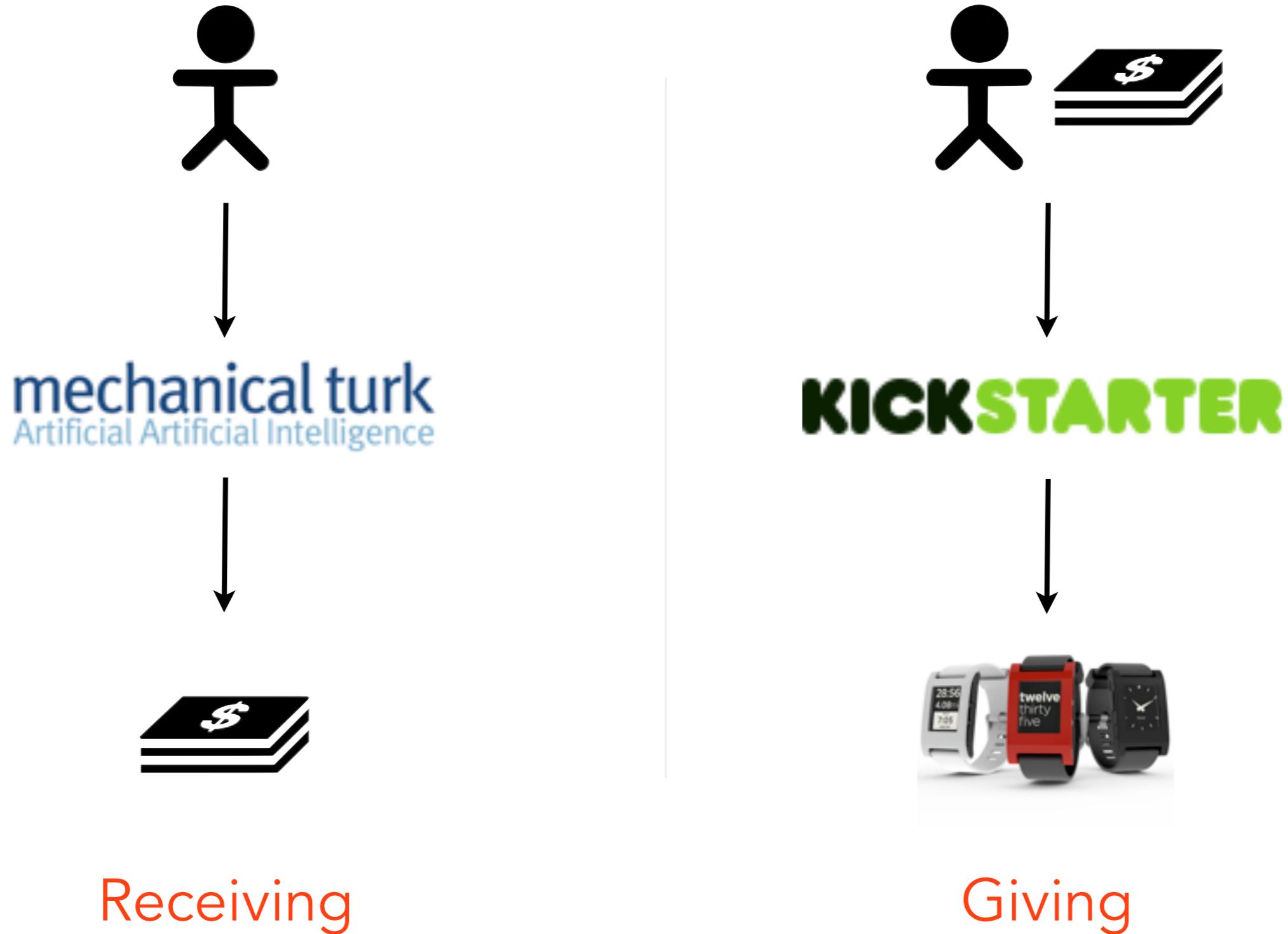
Greater Good



Games

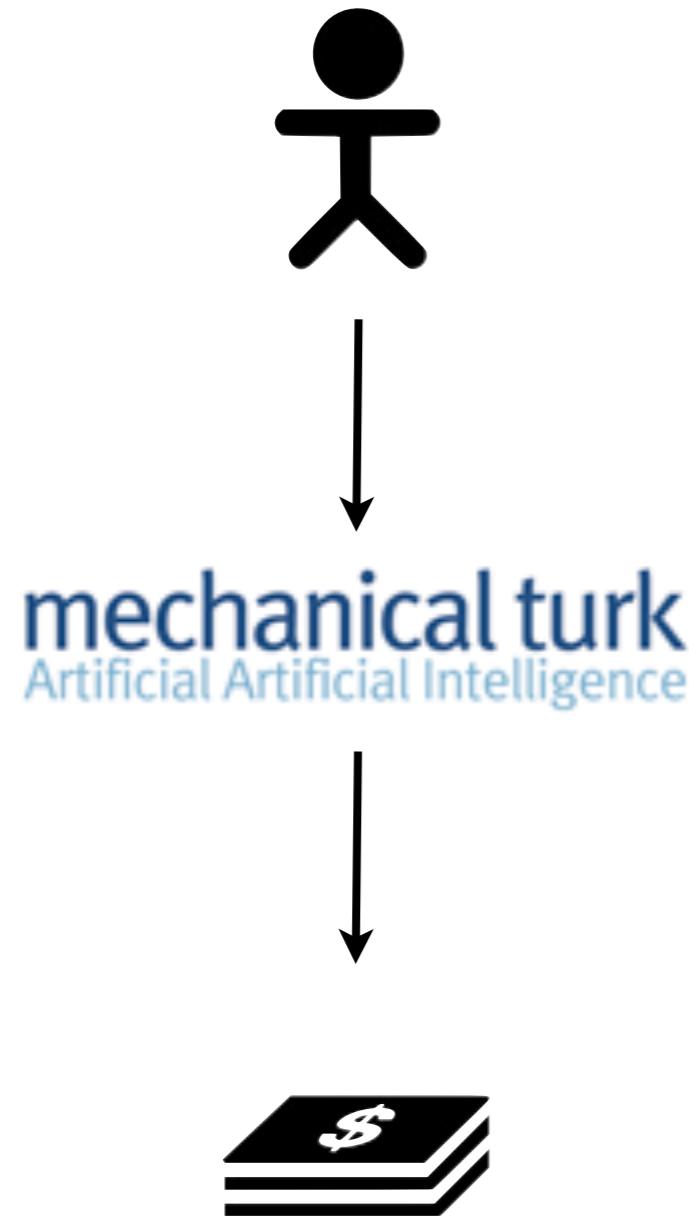


Value Incentives





Value Incentives



Can be a great motivator [1]



Quality **does not** increase
with additional money [2]

[1] Tang, J. C., Cebrian, M., Giacobe, N. a., Kim, H.-W., Kim, T., & Wickert, D. "Beaker." Reflecting on the DARPA Red Balloon Challenge. Communications of the ACM. 2011

[2] Mason, W., & Watts, D. J. Financial incentives and the performance of crowds. ACM SIGKDD. 2010

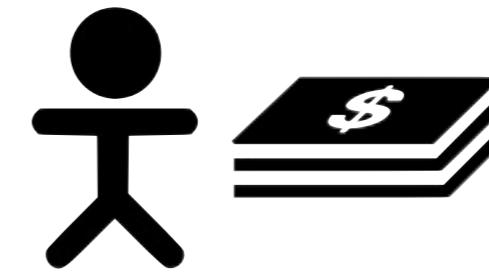


Value Incentives

Fund products, projects, services, anything!

Collective pocketbook [1]

Enhanced customer experience [2]



KICKSTARTER

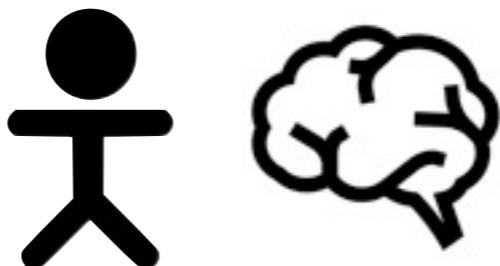


[1] Howe, J. Crowdsourcing: How the power of the crowd is driving the future of business. 2008

[2] Belleflamme, P. Crowdfunding: Tapping the right crowd. AFFI. 2011

Crowd Engagement & Motivation

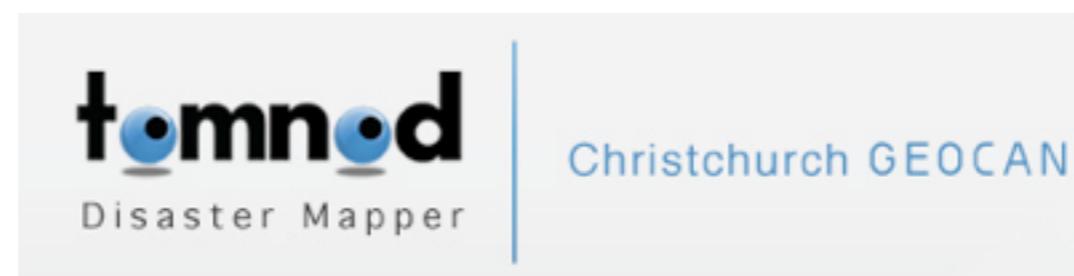
The Greater Good



“Citizen Scientists”

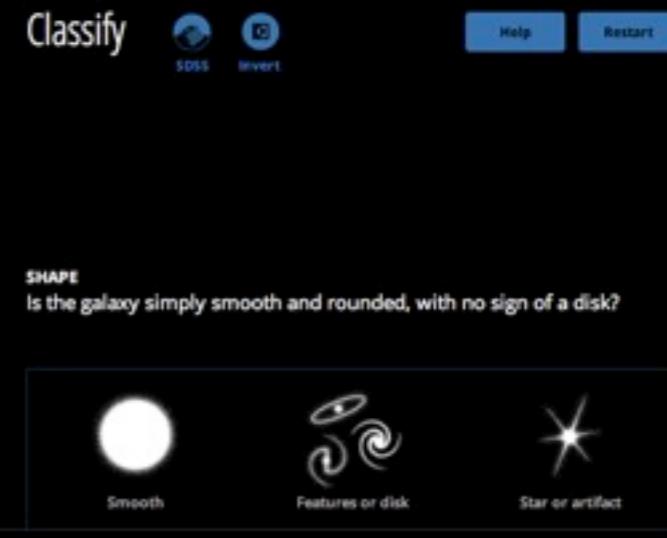
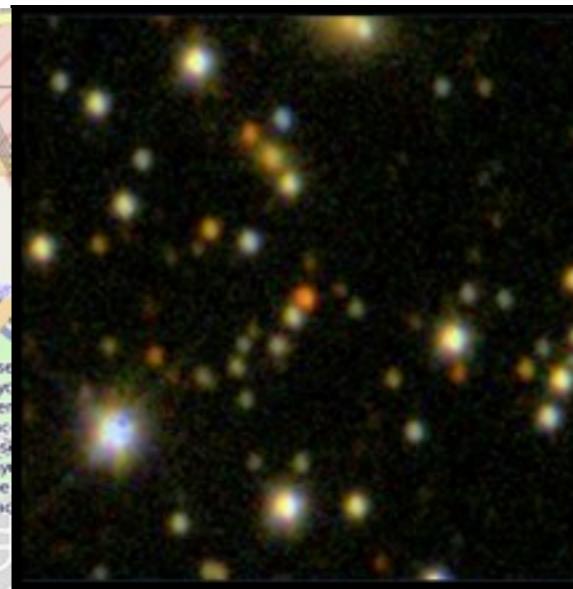


OpenStreetMap



Crowd Engagement & Motivation

The Greater Good



Improved data collection [1]

Analyzing and interpreting data [1]

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012).

- [1] The future of citizen science: emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*.

Crowd Engagement & Motivation

Games With A Purpose [1]



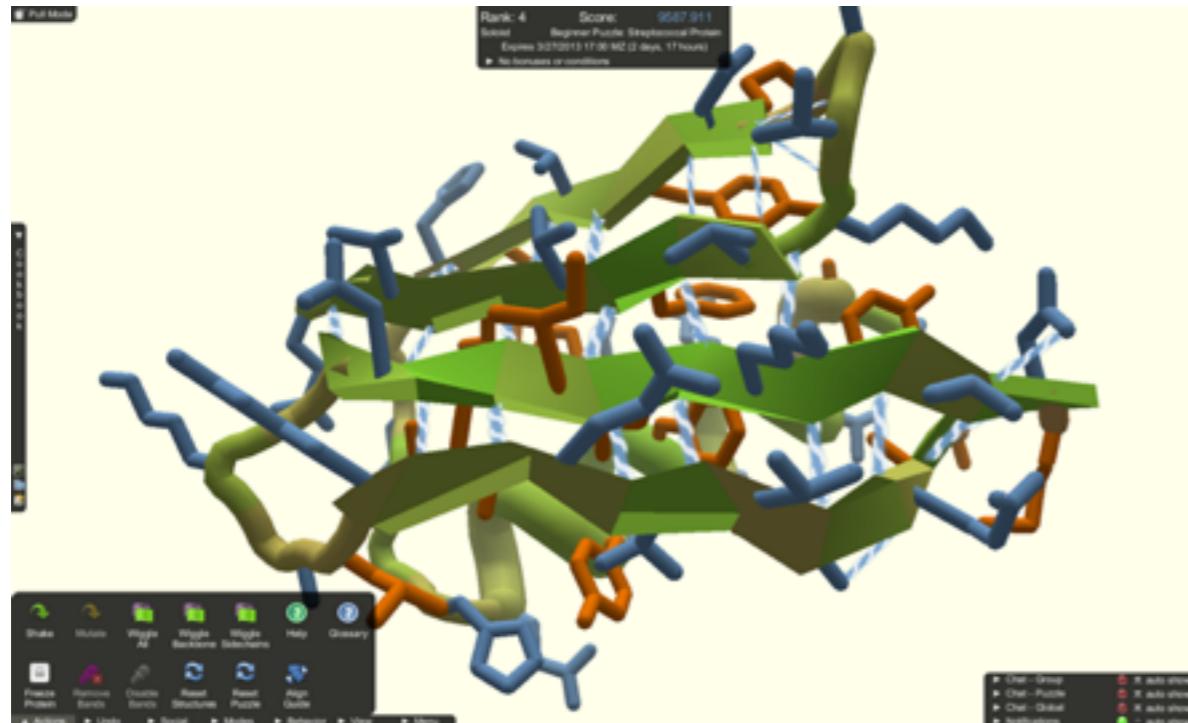
Engagement using a game

- [1] Ahn, L. von, & Dabbish, L. (2008).
Designing games with a purpose. Communications of the ACM.

Active Crowdsourcing

Discussion: Engagement & Motivation

Increase quantity and quality of participants

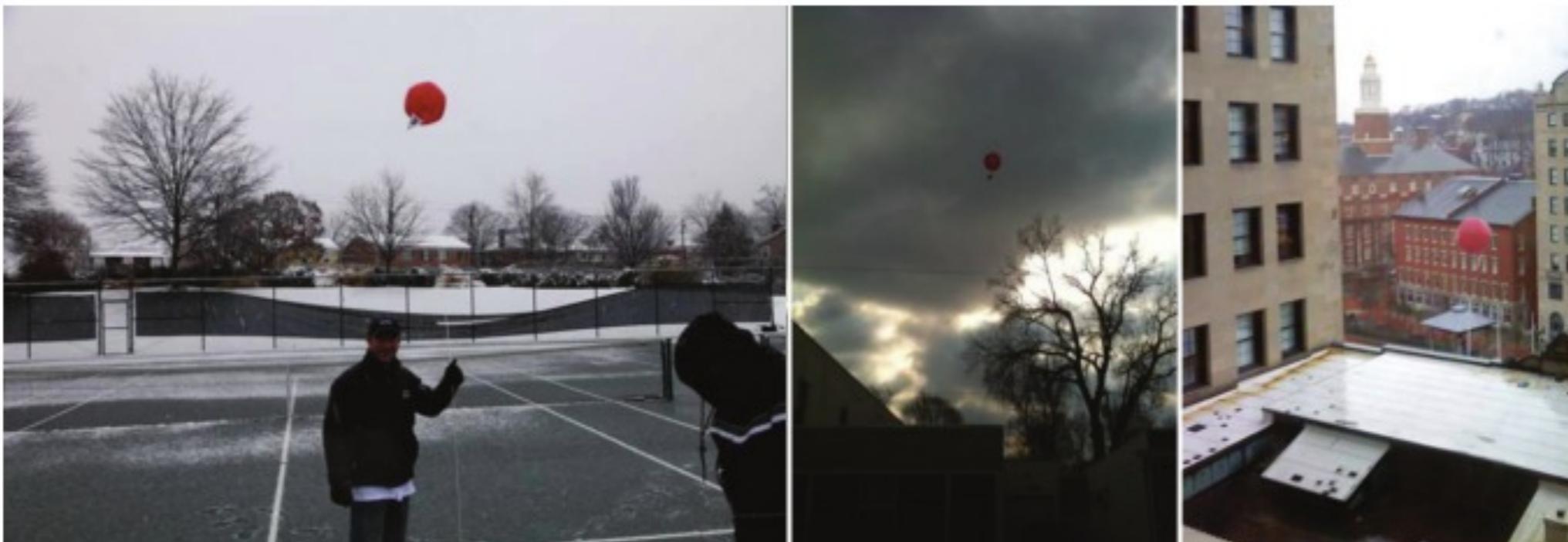


GWAP +
Greater Good

Combine different engagement methods

Discussion: Engagement & Motivation

Increase quantity and quality of participants



Real

Fake

Fake

Financial incentives increase **quantity** not **quality** [1]

[1] Mason, W., & Watts, D. J.
Financial incentives and the performance of crowds. 2010

Discussion: Engagement & Motivation



Value



Greater Good



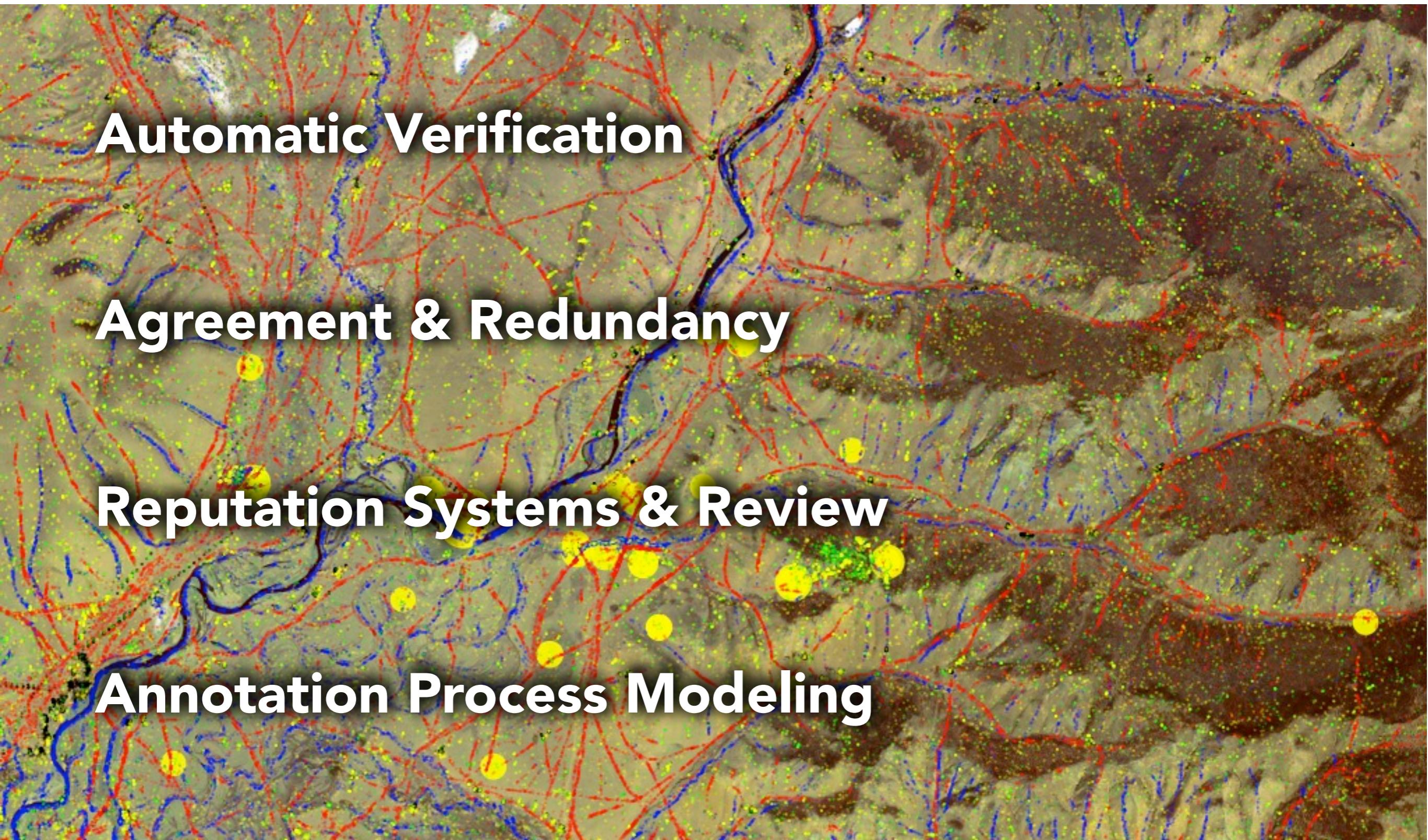
Games

Crowd is engaged and motivated

How do we make sure that the data coming in is good?

Active Crowdsourcing

Quality Control



Automatic Verification

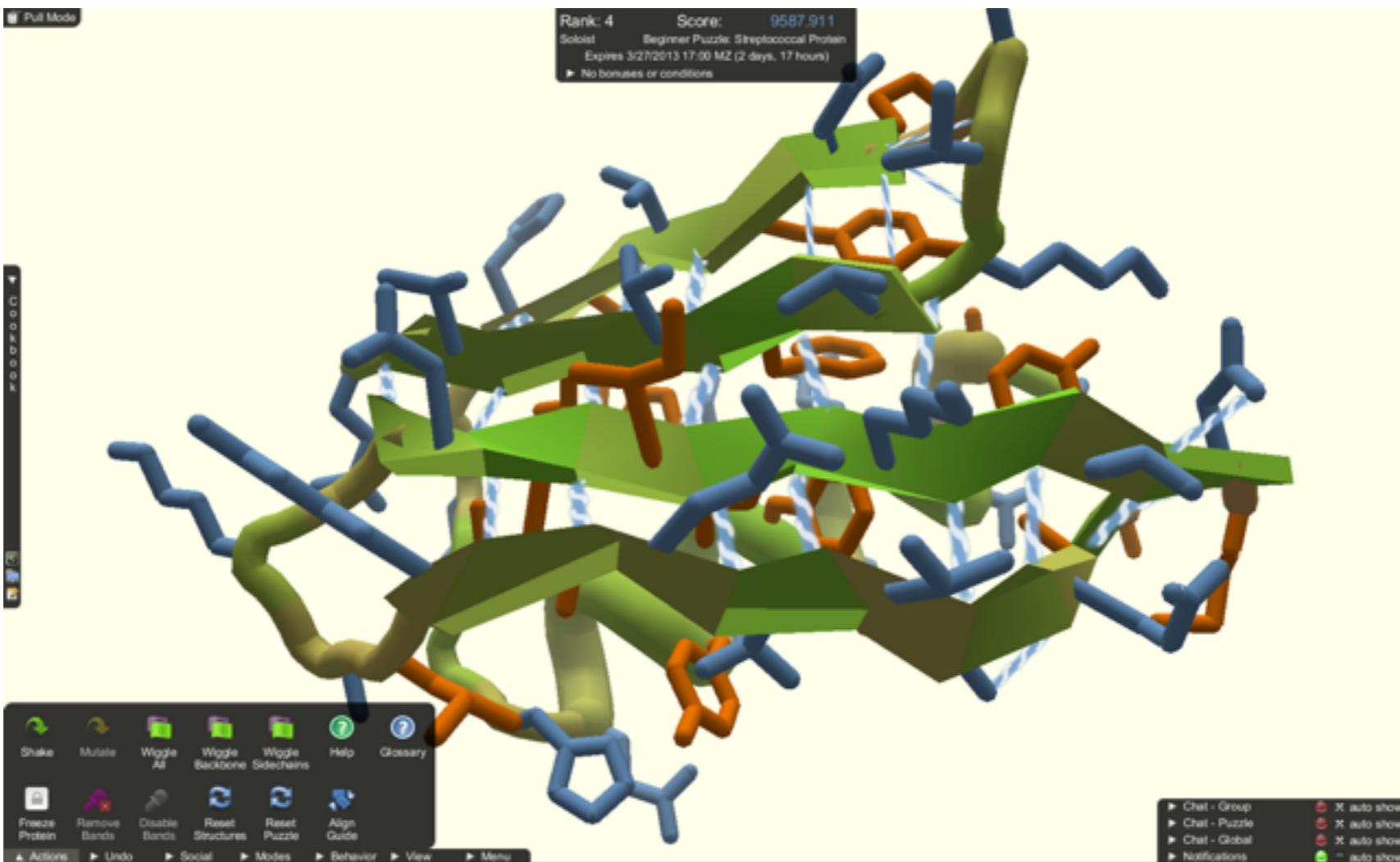
Agreement & Redundancy

Reputation Systems & Review

Annotation Process Modeling

Quality Control

Automatic Verification

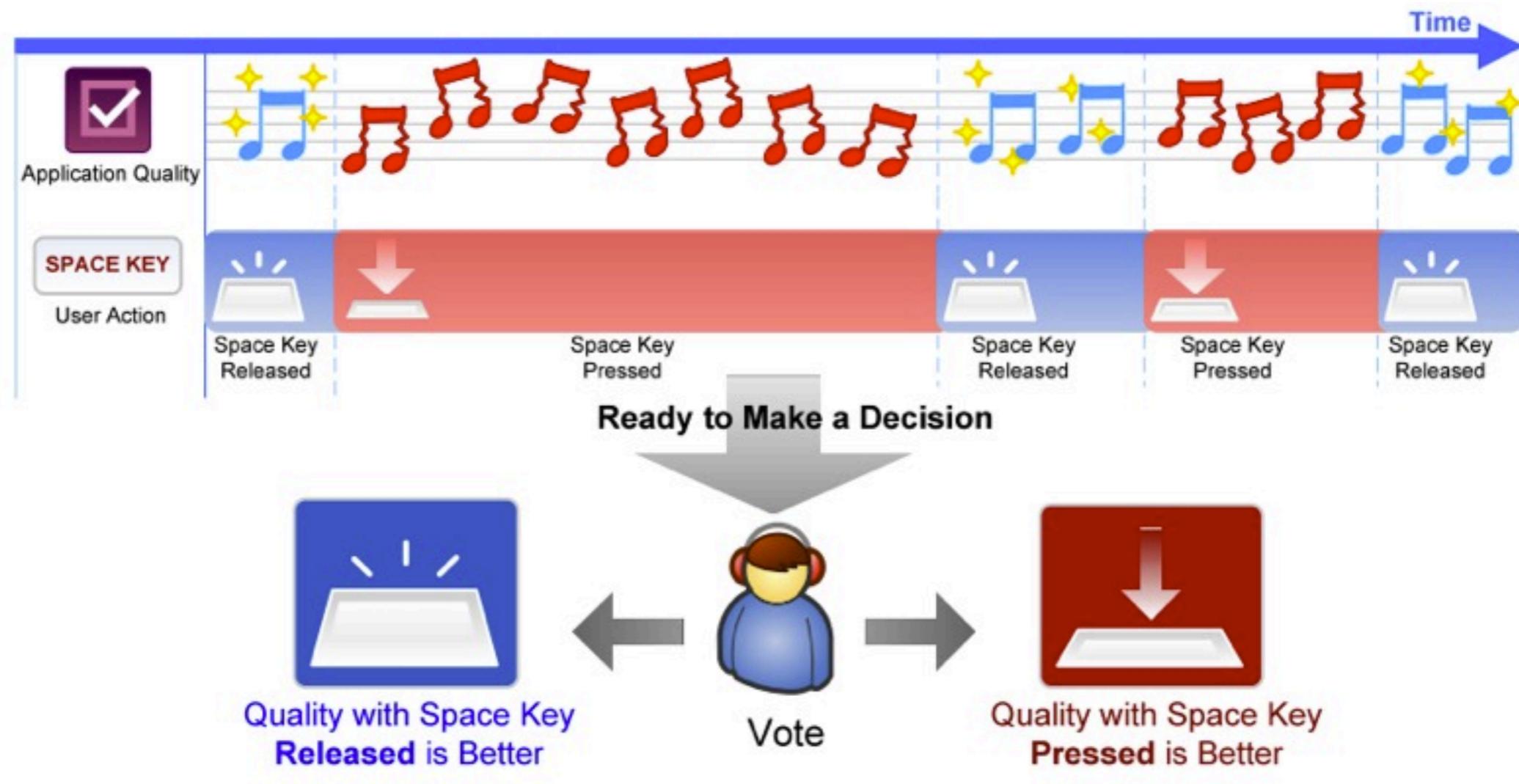


Automatic

<http://fold.it>

Quality Control

Statistical Filtering



QoE can be filtered using the known distribution [1]

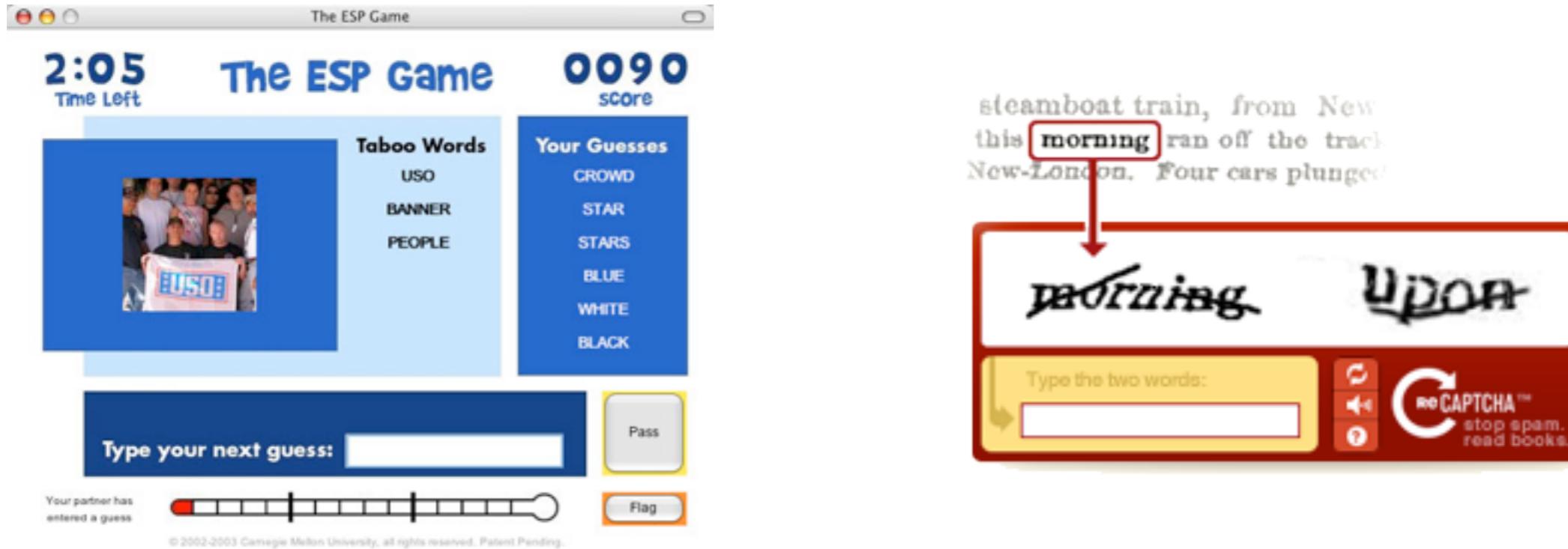
[1]

Chen, K.-T., Wu, C.-C., Chang, Y.-C., & Lei, C.-L.

A crowdsourceable QoE evaluation framework for multimedia content. 2009

Quality Control

Agreement & Redundancy



Agreement [1]

All participants agree

Redundancy [2]

Majority voting or averaging

[1] Ahn, L. von, & Dabbish, L. Labeling images with a computer game. 2004

[2] Ahn, L. von, Maurer, B., McMillen, C., Abraham, D., & Blum, M. reCAPTCHA: human-based character recognition via Web security measures. 2008

Quality Control

Reputation Systems

Assign Qualification Type

Select which of your Qualification Types to assign to this Worker. You will be prompted to provide a score between 0 and 100. (Assign up to 5 Qualification Types at a time.)

Spanish speaker
 Legal
 Good tagger
Score:
 Good editor

Assign or [Cancel](#)

Based on **previous** input

Set a minimum qualification limit

Filter participants by “qualifications”

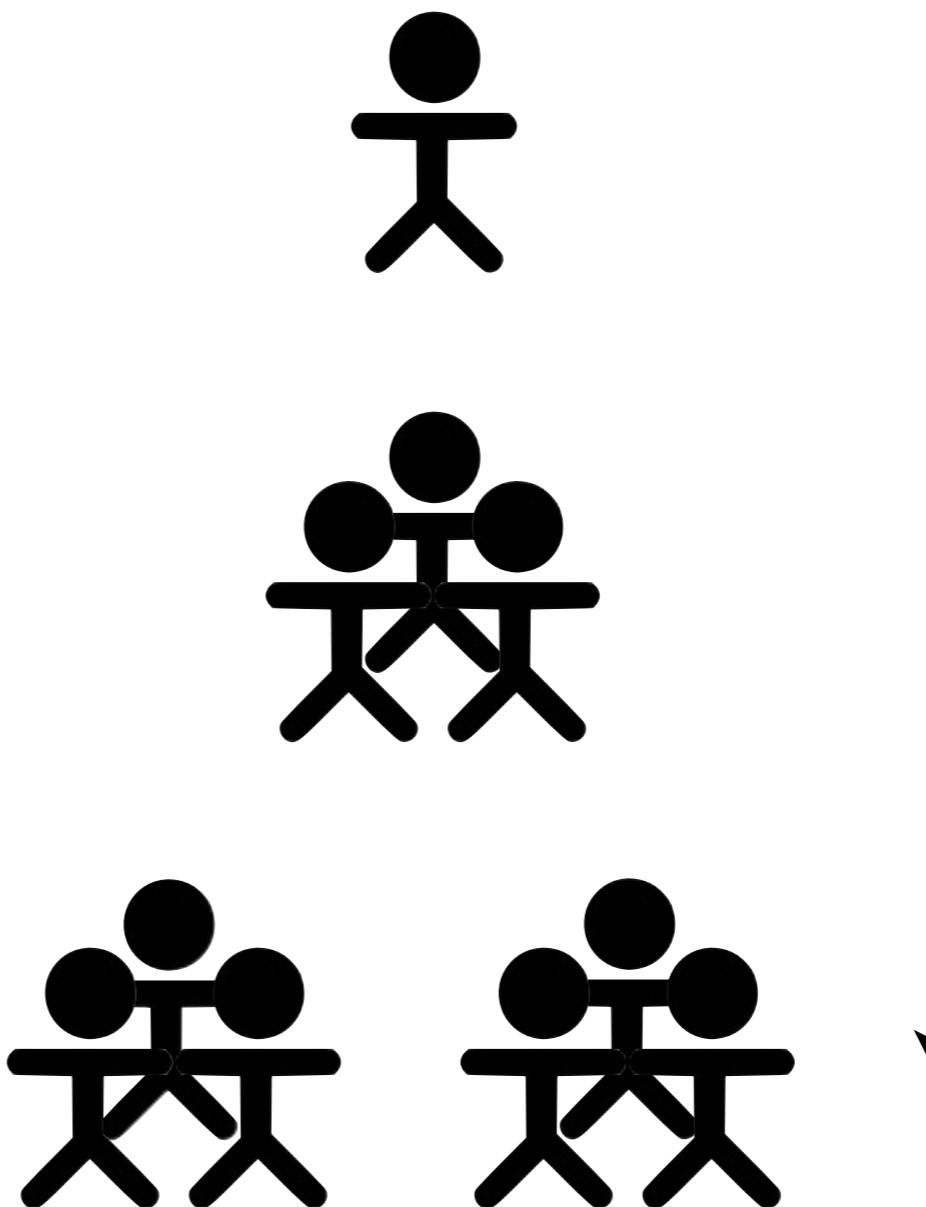
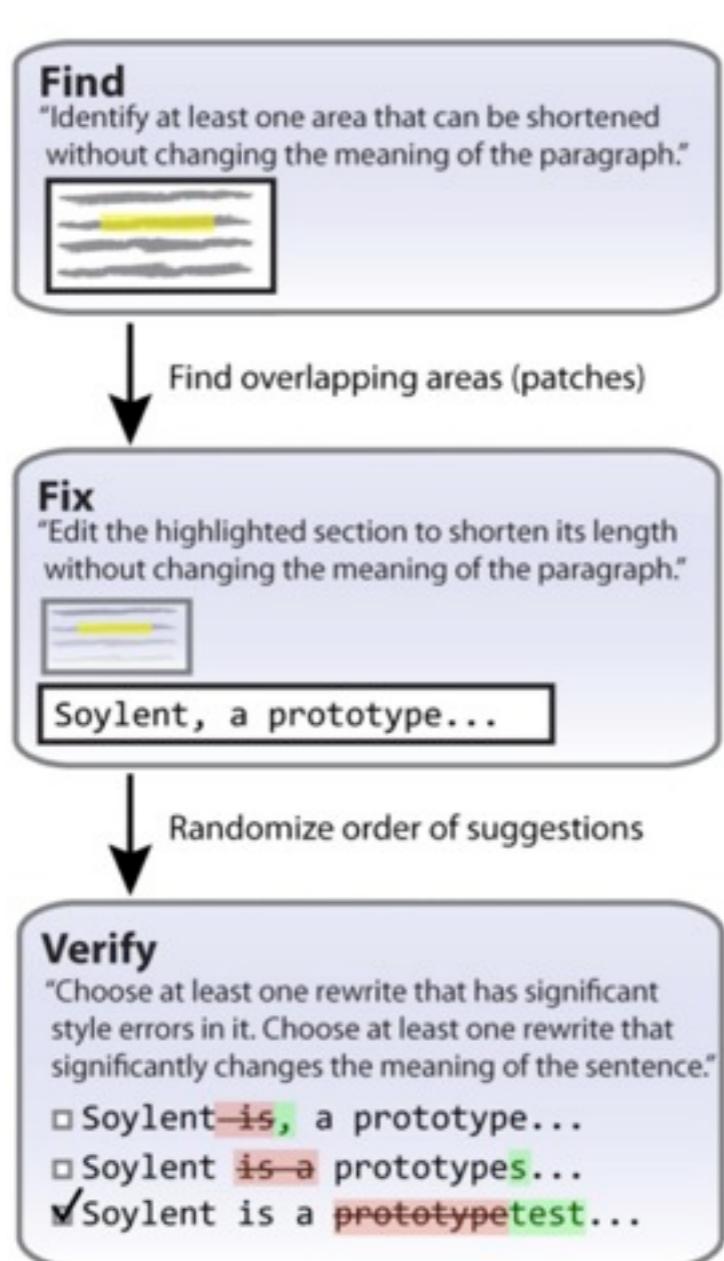
Worker's Qualifications For Your Work

[Assign Qualification Type](#)

Qualifications			
	Name	Description	Score
<input checked="" type="checkbox"/>	A plus work	This Worker does excellent work.	98 edit

Quality Control

Expert/Crowdsourced Review



Bernstein, M., Little, G., & Miller, R.

[1] Soylent: a word processor with a crowd inside. Proceedings of the 23rd annual ACM symposium on User interface software and technology. 2010

Quality Control

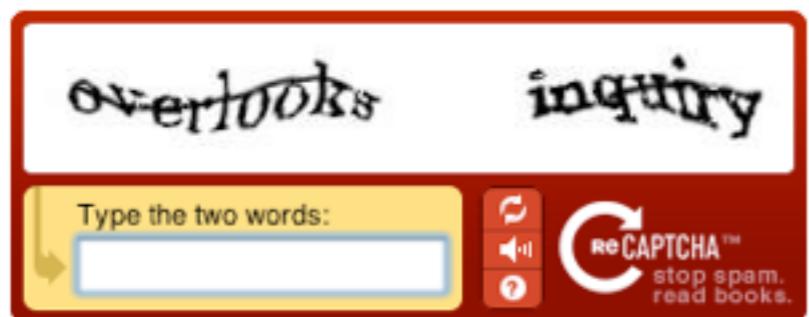
Annotation Modeling



VS



Varying degrees of competence

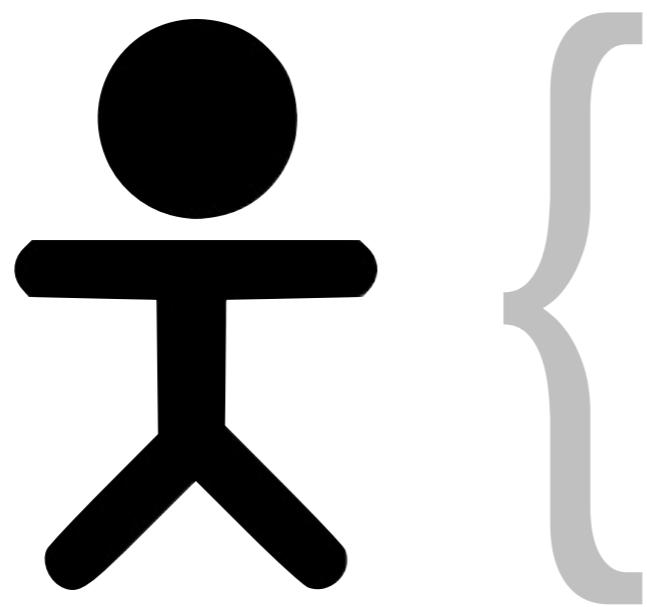


VS



Some tasks are easier than others

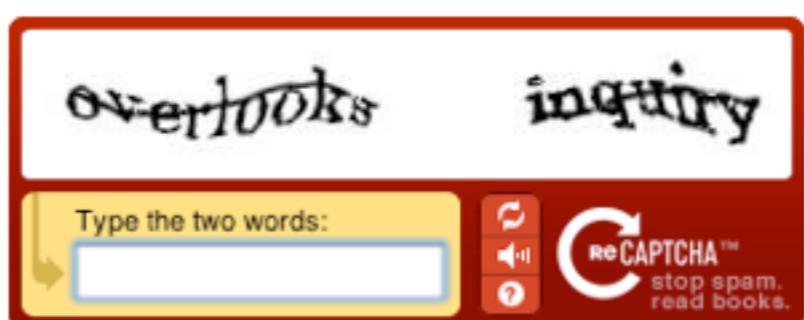
Representing the annotation process [1]



Bias

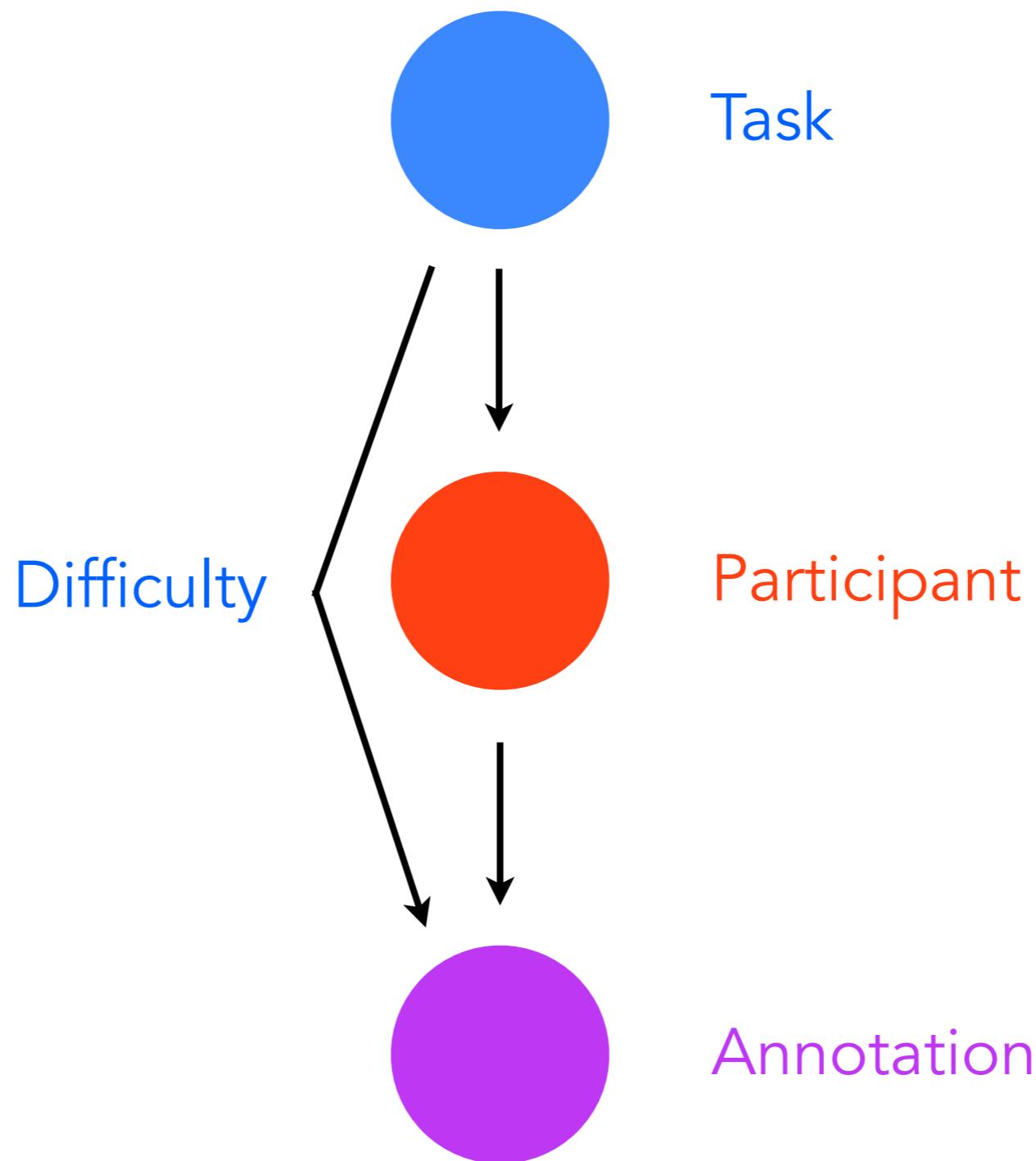
Expertise

Competency



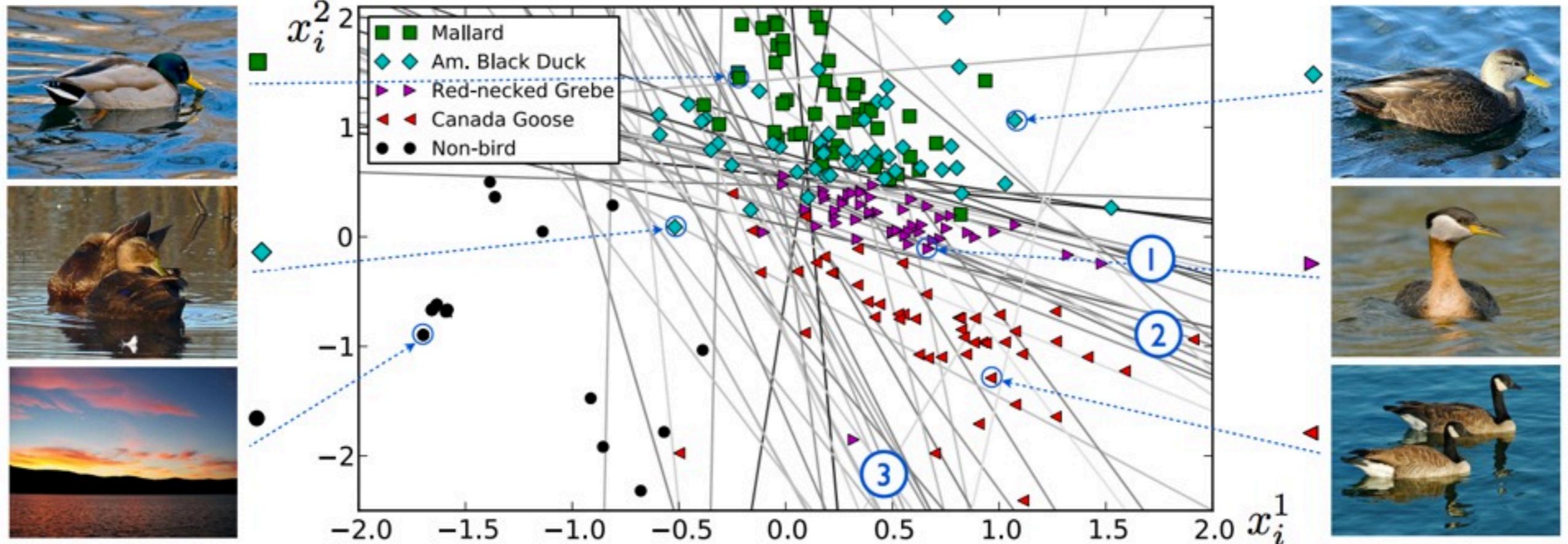
Task Difficulty

Representing the annotation process



Annotation Modeling

Representing the annotation process

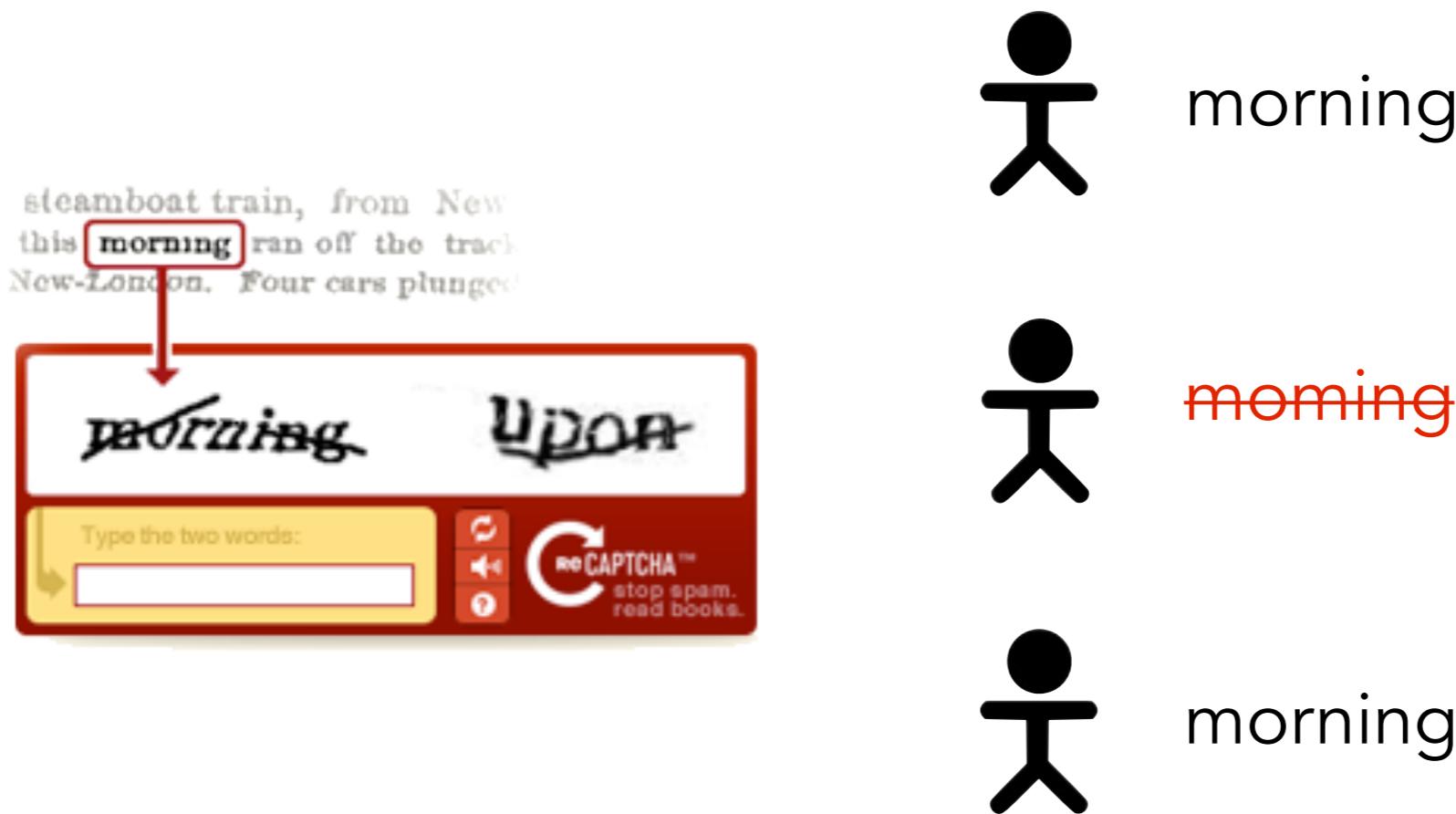


Users (and the data they contribute) can be understood.

These different “schools of thoughts” can be used to weight the data to retrieve a more accurate classification.

Discussion: Quality Control

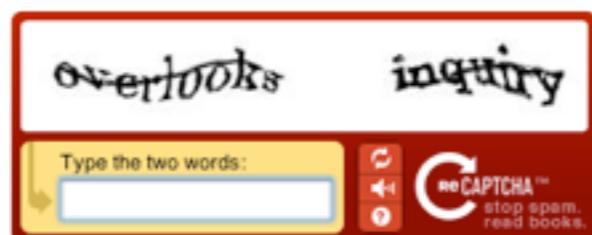
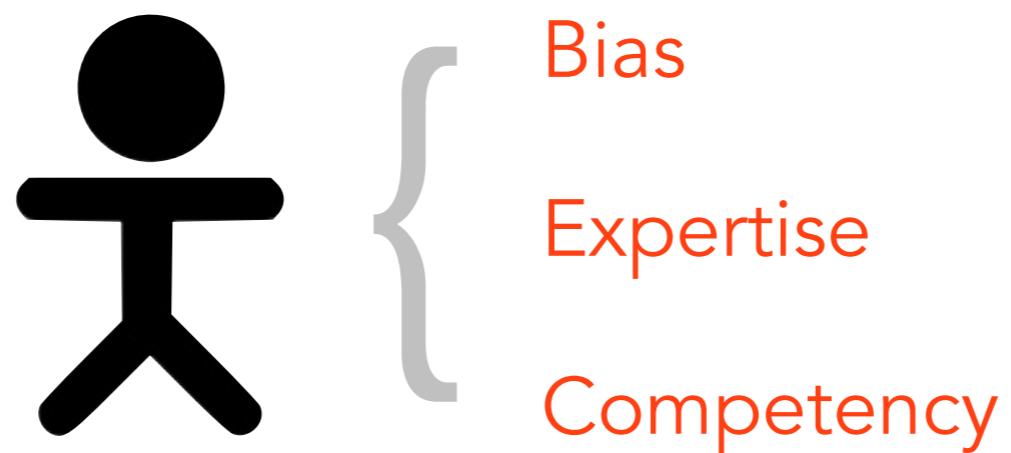
Increase the quality of data from participants



Depending on task, simple voting may suffice

Discussion: Quality Control

Increase the quality of data from participants



Task Difficulty

Annotation process modeling improves accuracy
with additional computational cost

Passive Crowdsourcing

Active

Distributed tasks

Crowd = Pool of labor

Aggregate solutions

Passive

Distributed sensors
Idle computers, mobile devices.

Crowd = Source of data
Harness existing behavior.

“Work for nothing” [1]
No need for motivation/
engagement

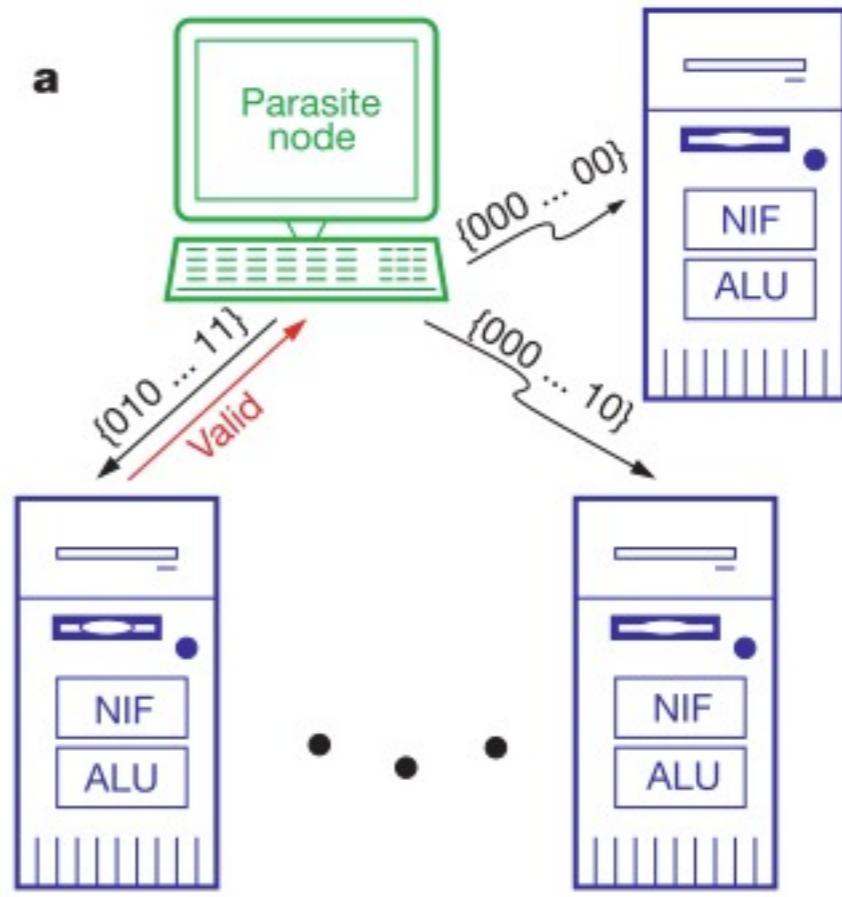
[1] Adar, E.
Why I hate Mechanical Turk research (and workshops). 2011

Parasitic Computing [1]

Computations using TCP checksums.

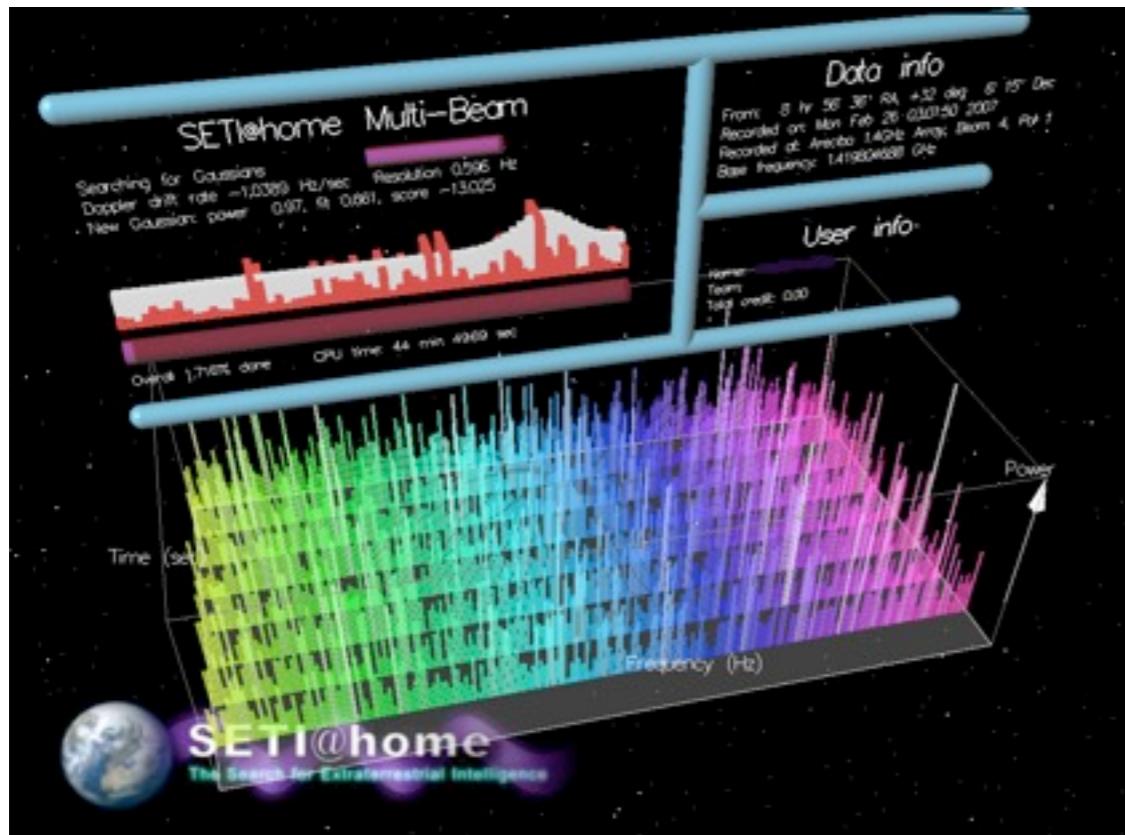
Failures are dropped.
Success have responses.

Distributed the computation
of a satisfiability problem.



Passive Crowdsourcing

SETI@Home



Problems with high computing to data ratio.

Computers often idle.

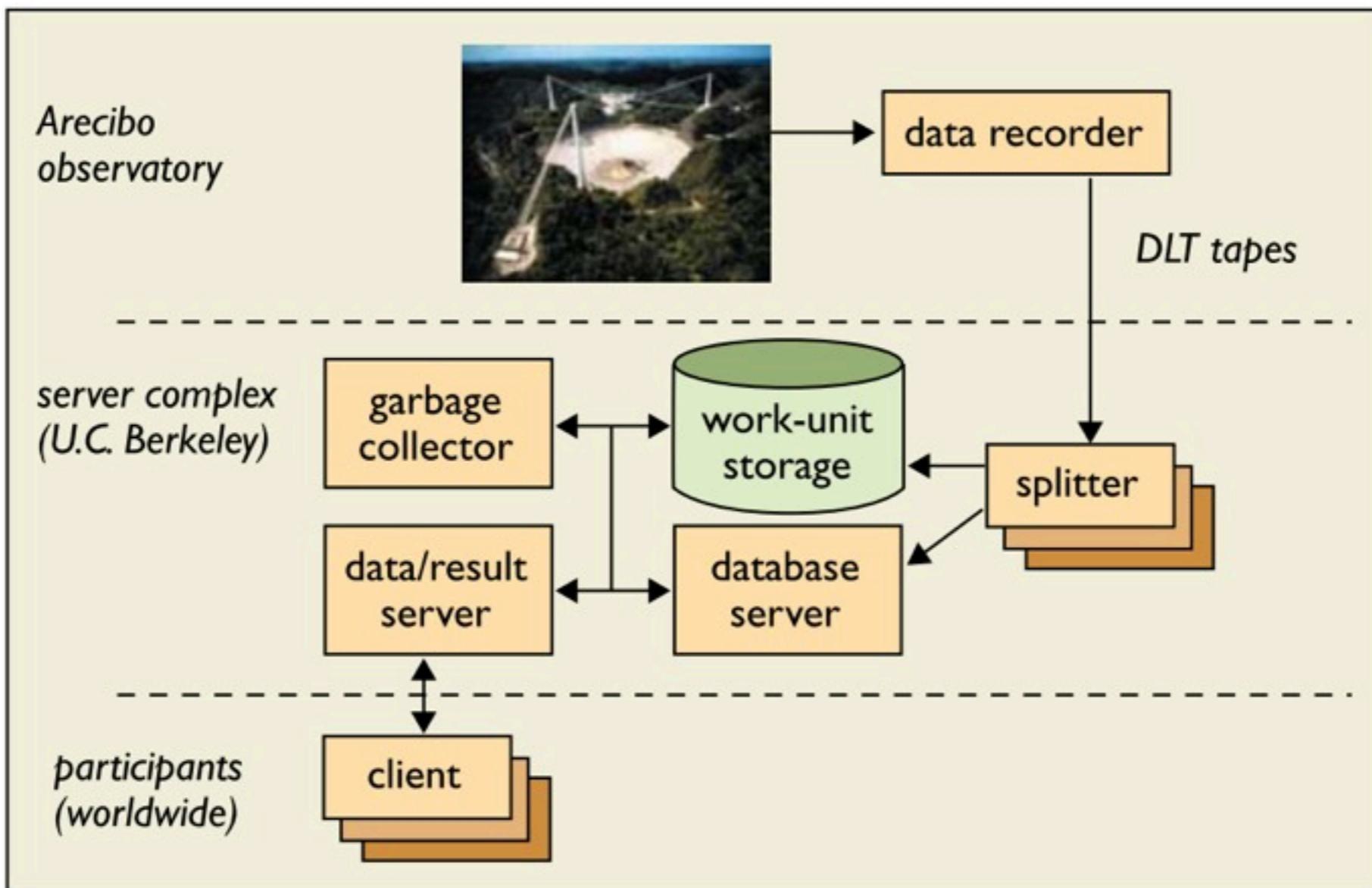
Idle computing time harnessed [1].

[1]

Anderson, D., Cobb, J., & Korpela, E.
SETI@ home: an experiment in public-resource computing. 2002

Passive Crowdsourcing

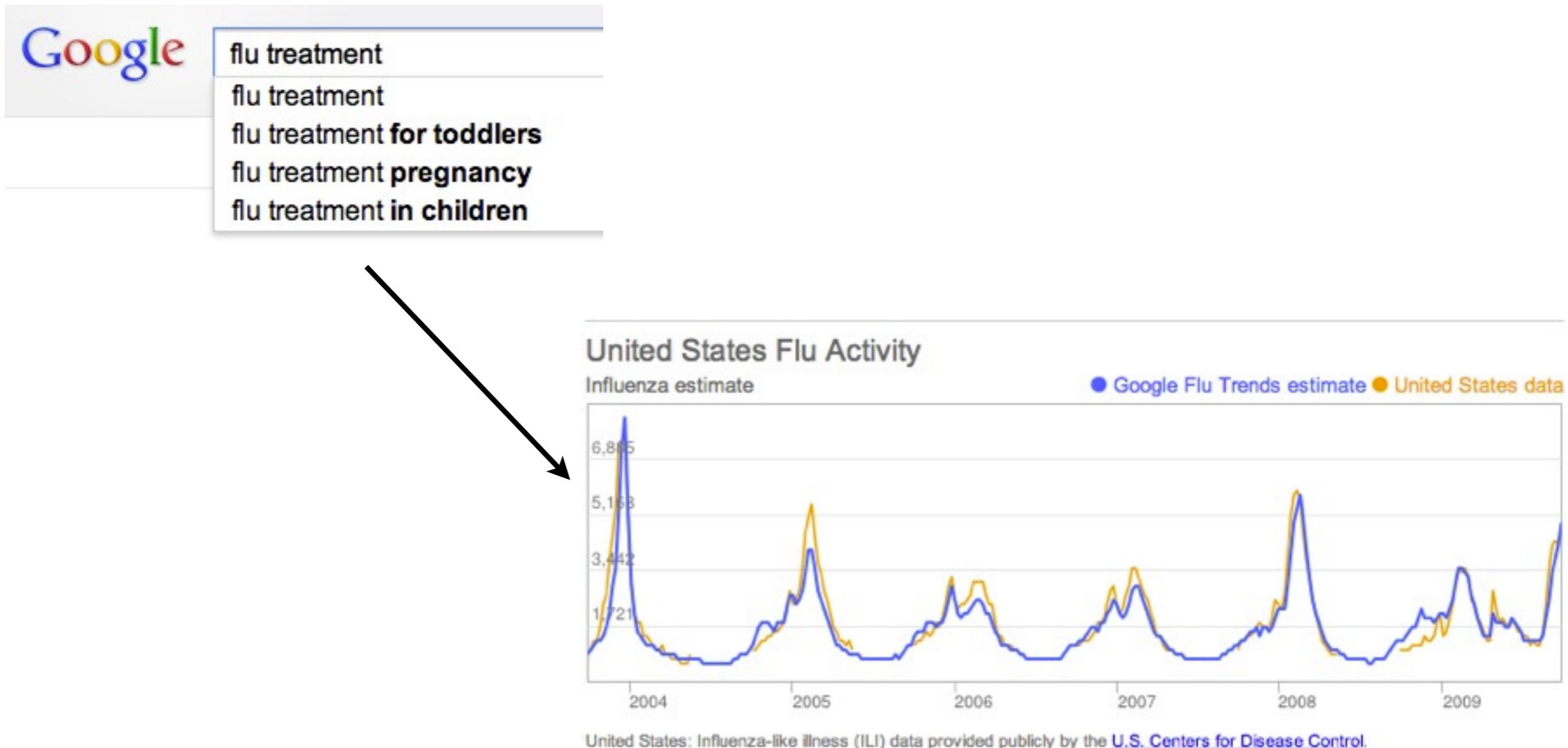
SETI@Home



145,000 active computers. 668 TeraFLOPS
Around 40th on TOP500 supercomputers

Passive Crowdsourcing

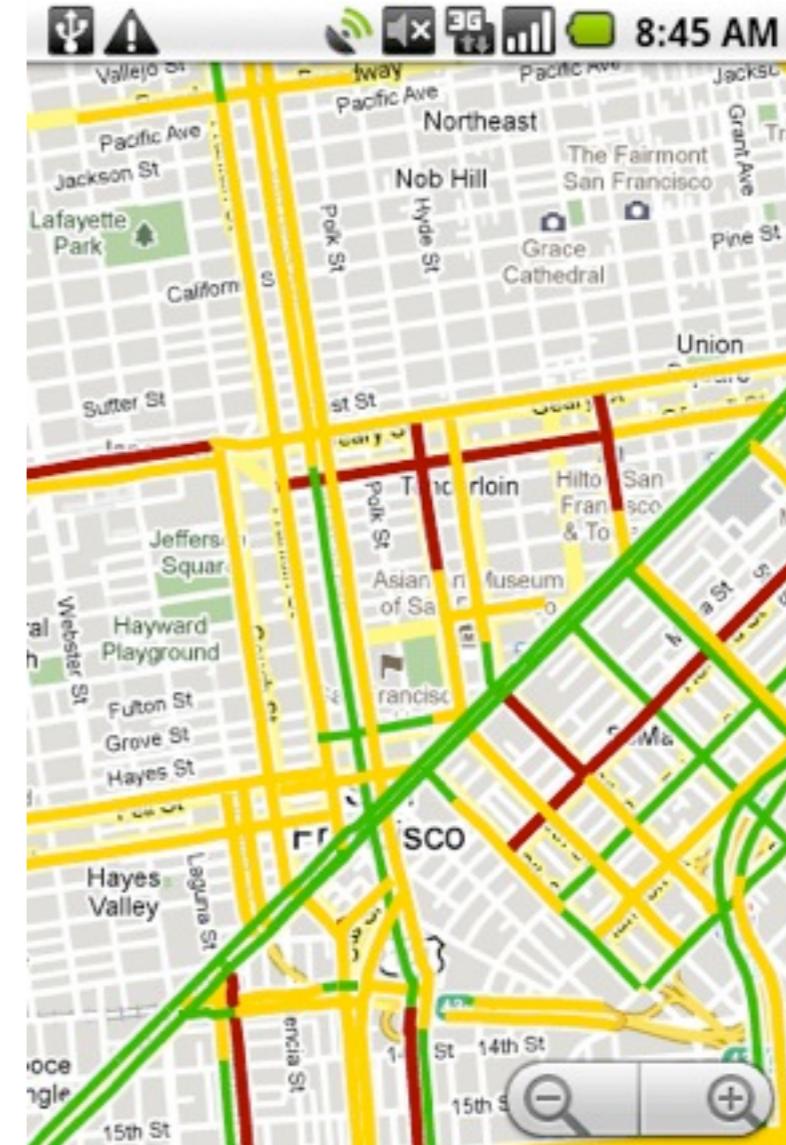
Crowdsensing



Based on everyday search queries, clear correlations with influenza activity.

Passive Crowdsourcing

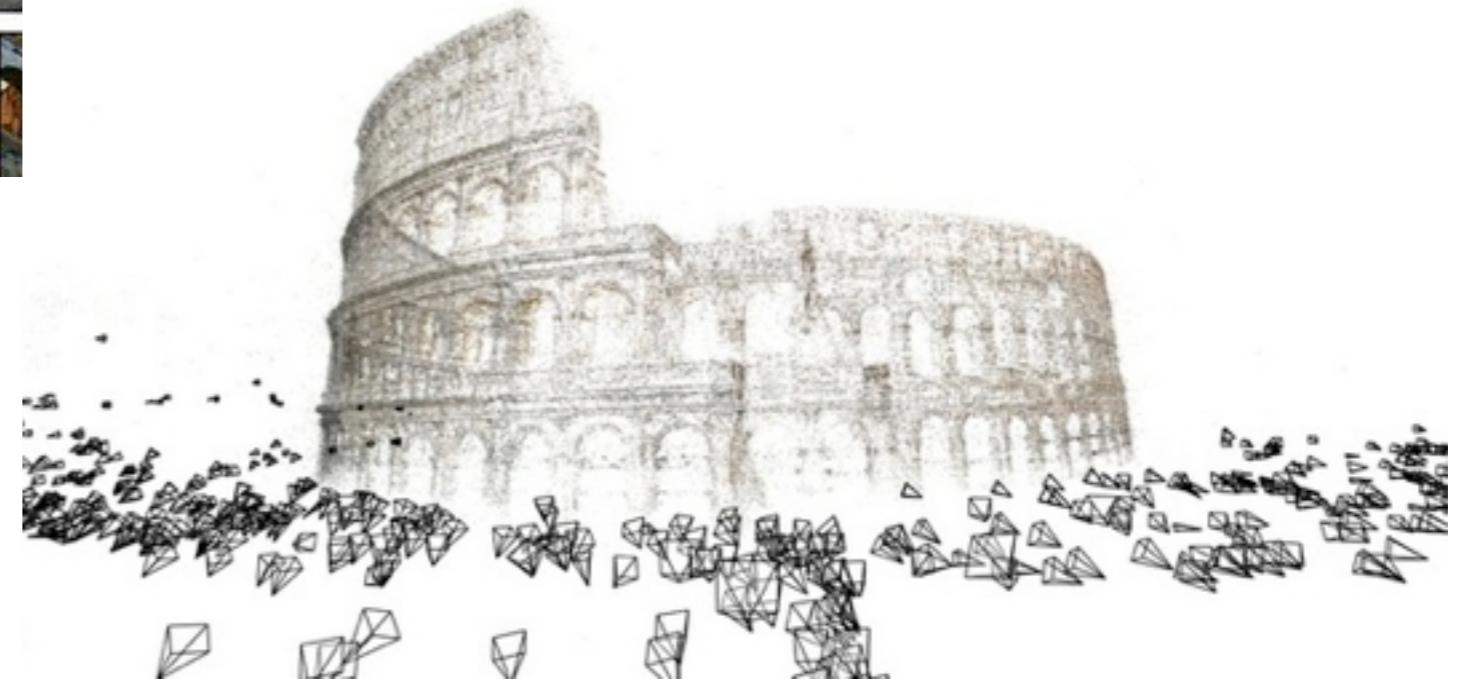
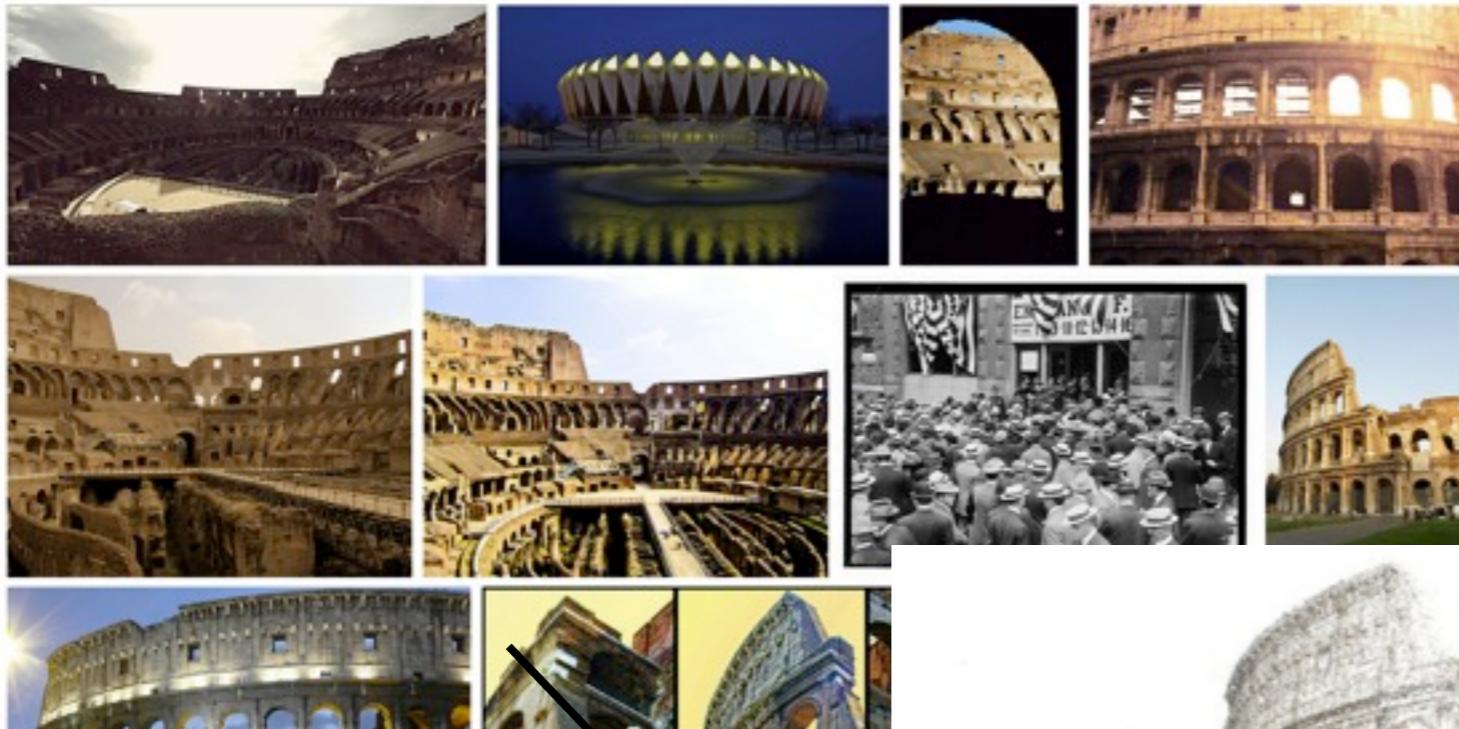
Crowdsensing



Gather data from Google Maps app as people sit in traffic

Passive Crowdsourcing

Crowdsensing



Public tourist pictures translated into a 3D model [1]

[1] Agarwal, S., Snavely, N., Simon, I., Seitz, S. M., & Szeliski, R.
Building Rome in a day. ICCV, 72–79. 2009

Passive Crowdsourcing

Discussion: Passive Crowdsourcing

Tap into existing behavior to collect data/run computations

Motivation/engagement is no longer a major problem

Untapped source of behavioral data.

Medical diagnostics, detecting activities such as running, walking, etc.

Security?

Privacy?

the
BIG
data

Learning from the Crowd



Training Set → Machine Learning → Classification

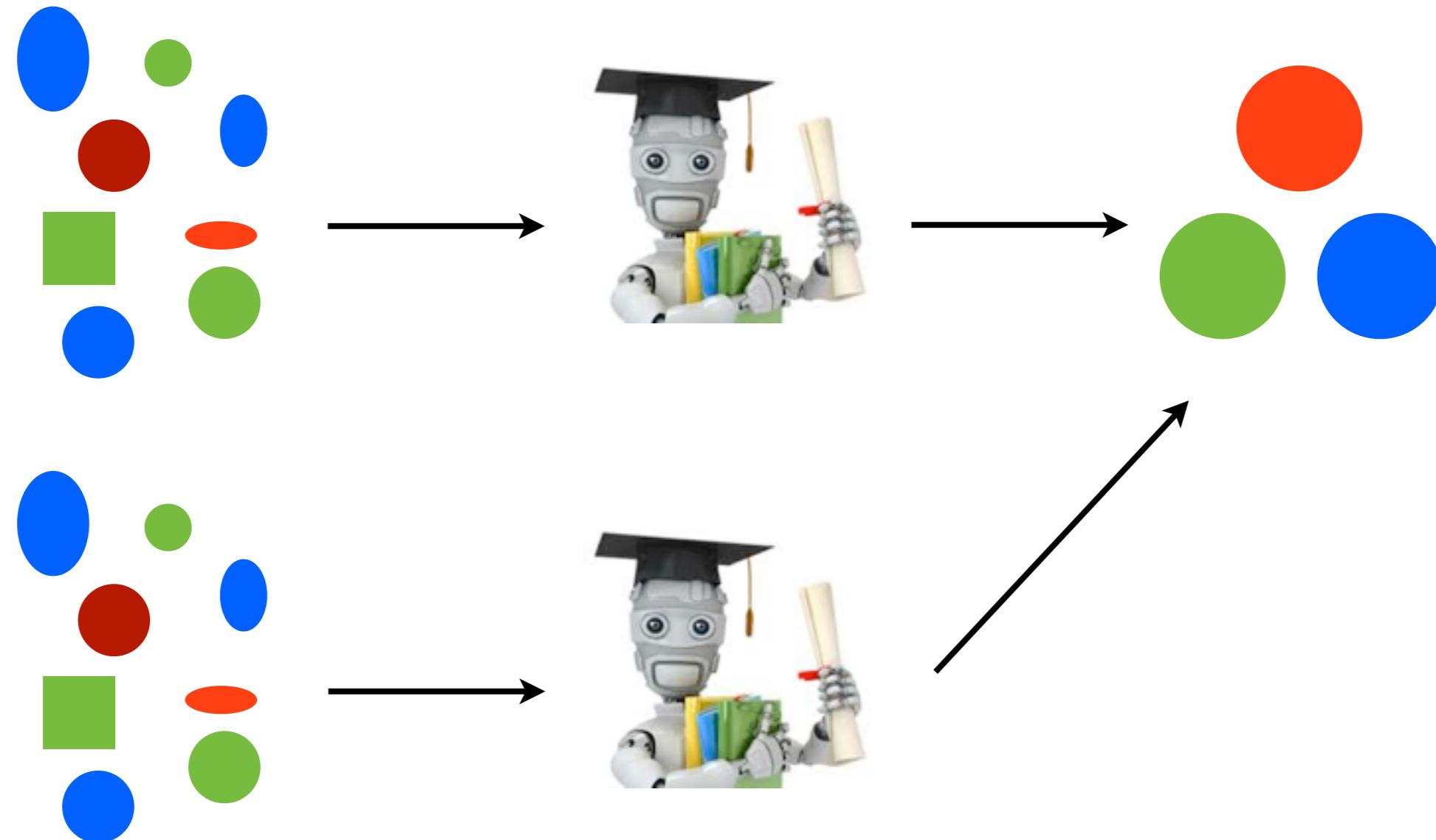
Learning from the Crowd



Crowd Data → Machine Learning → Classification

Learning from the Crowd

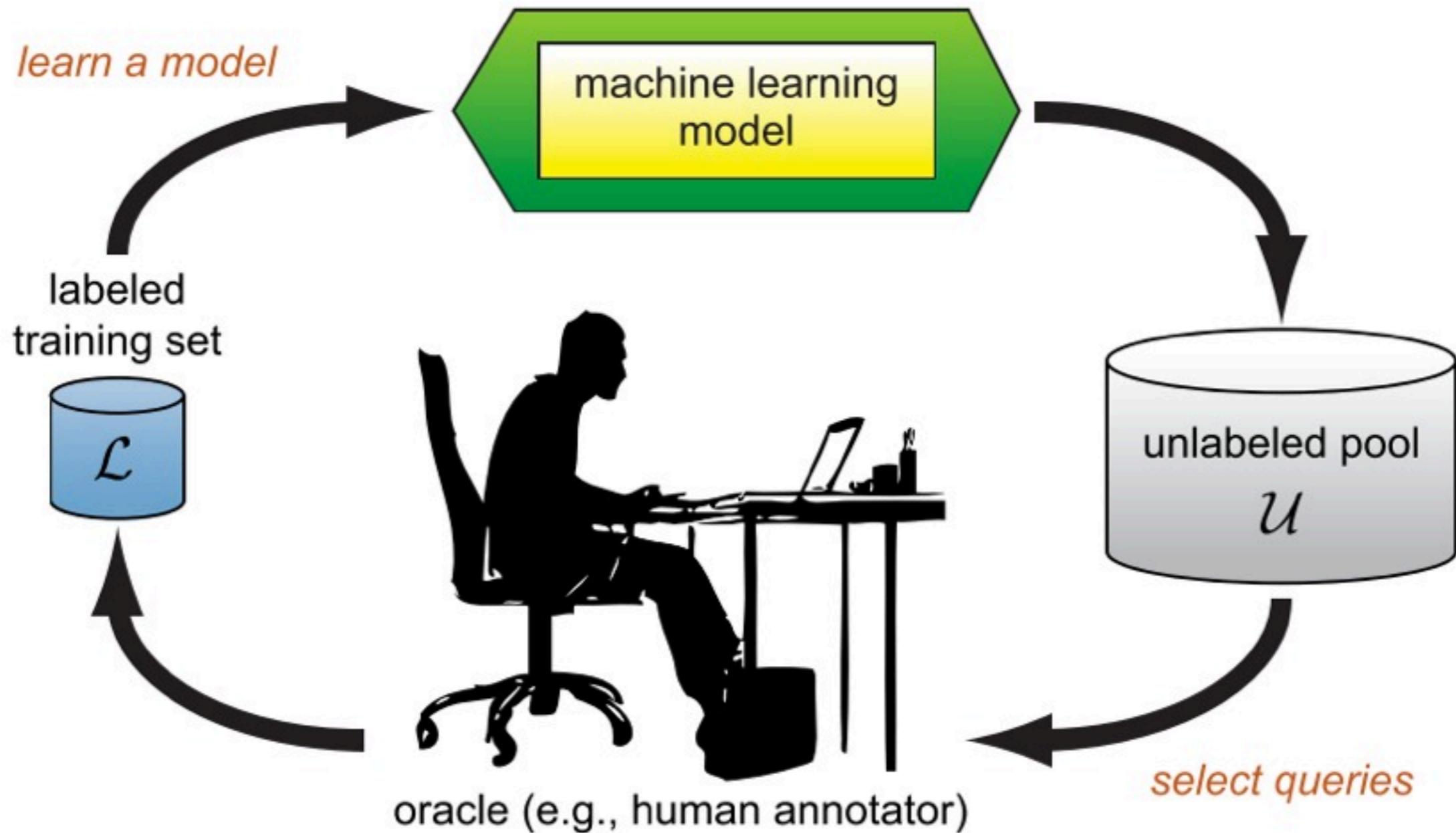
Active Learning



Continually improve model with additional training data

Learning from the Crowd

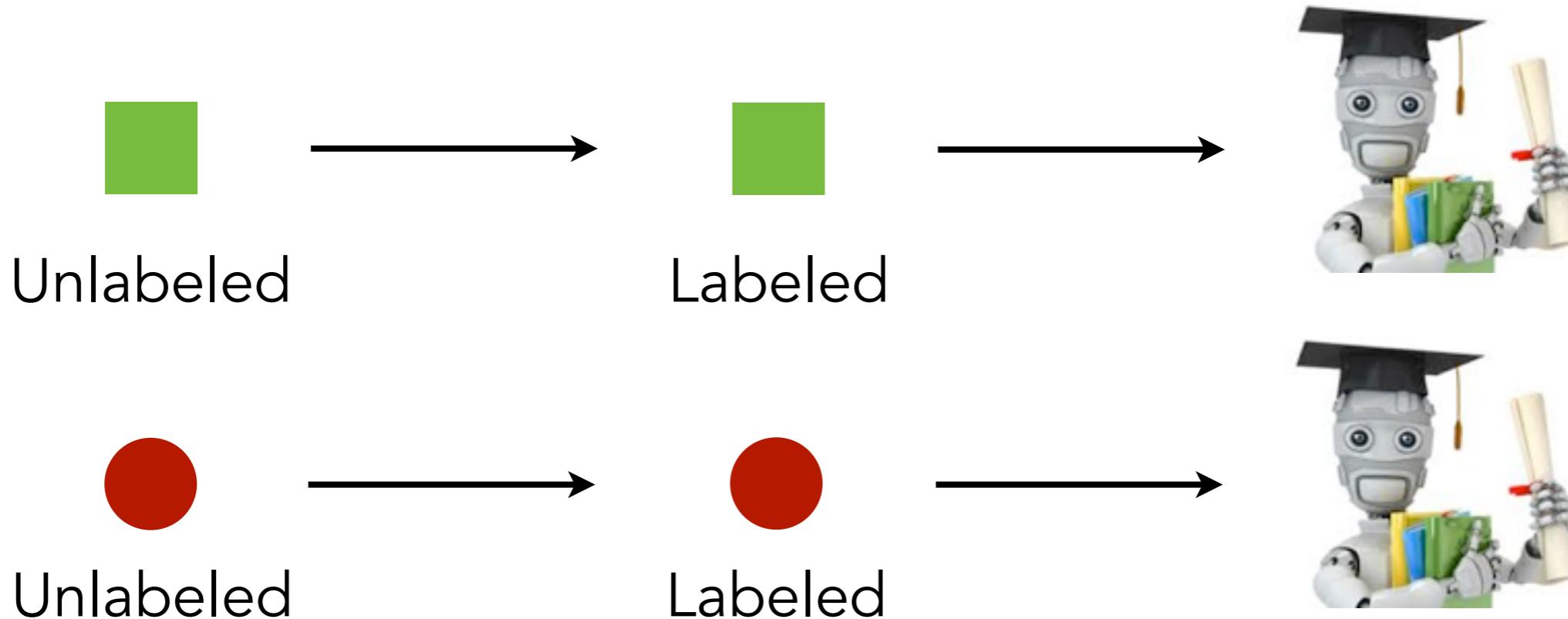
Active Learning



Active Learning

Sequential/Online approaches

Ask for a single unlabeled data at each iteration



Instances where we have a stream of unlabeled data.

Each instance is drawn one at a time, query for it's label or discard.

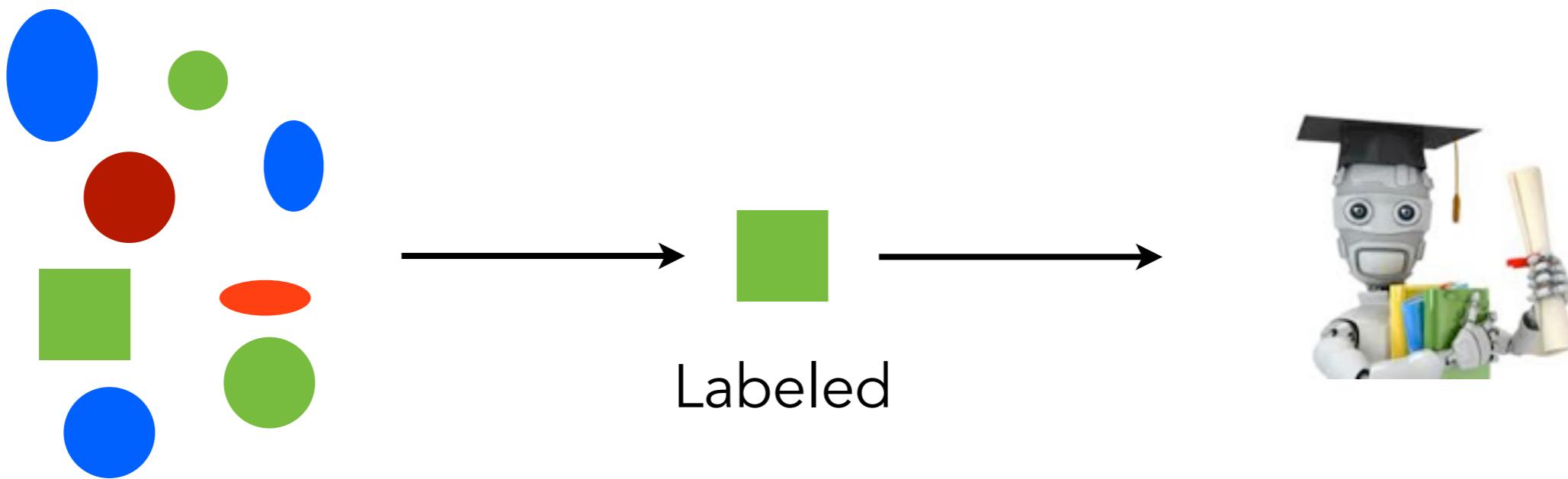
Select instances that will be more informative to the model.

Active Learning

Pool-based approaches

Large collections of unlabeled data

Small collection of labeled data



Queries from the pool using some informativeness measure

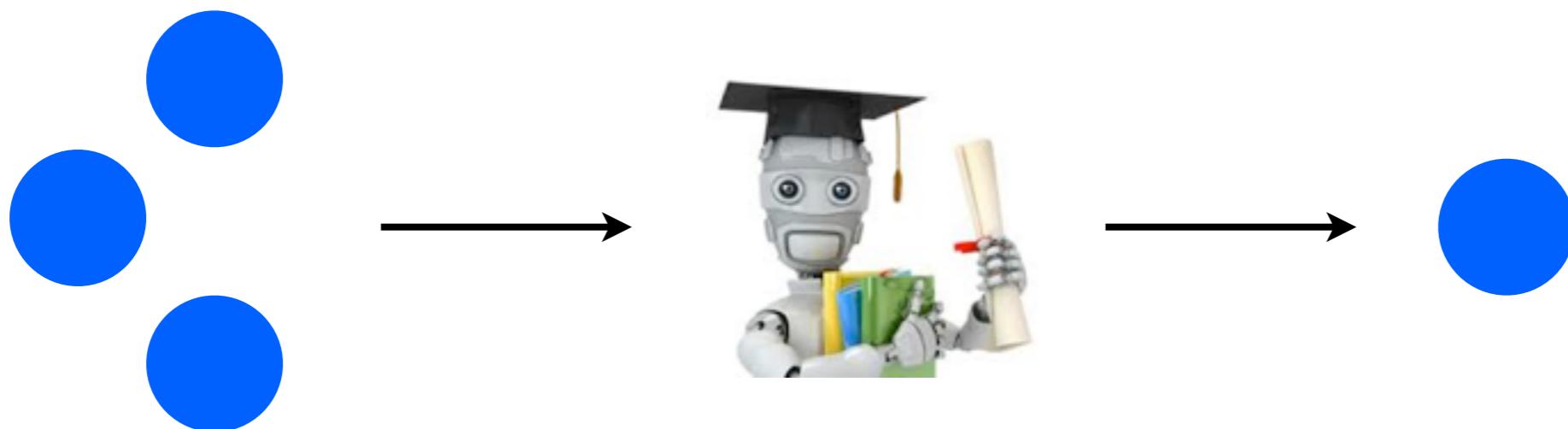
More computational expensive than online approach. Ranks the entire pool before making decisions.

Active Learning

Discriminative vs. Generative



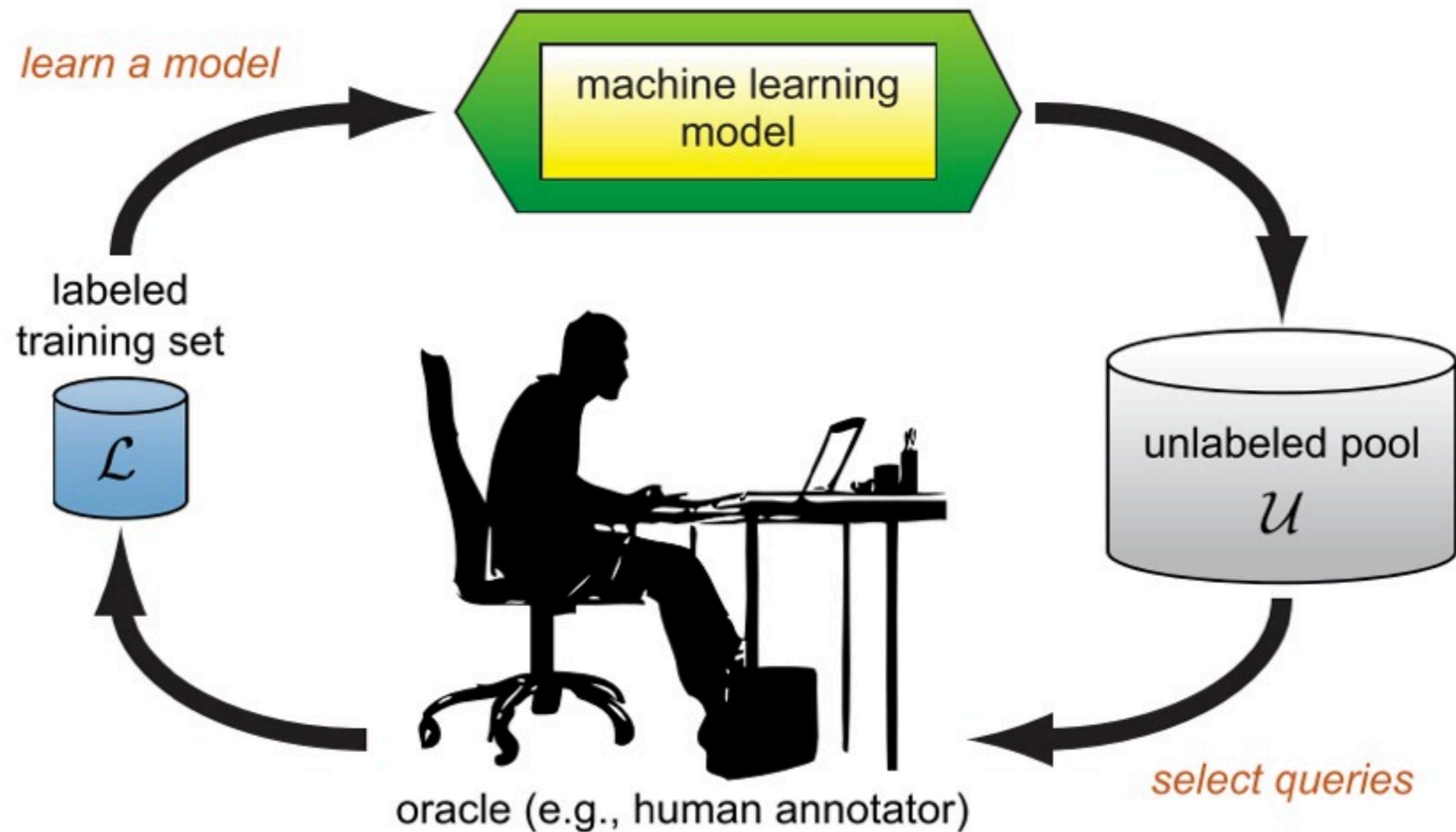
Discriminative uses positive and negative examples for training



Generative require only positively labeled examples

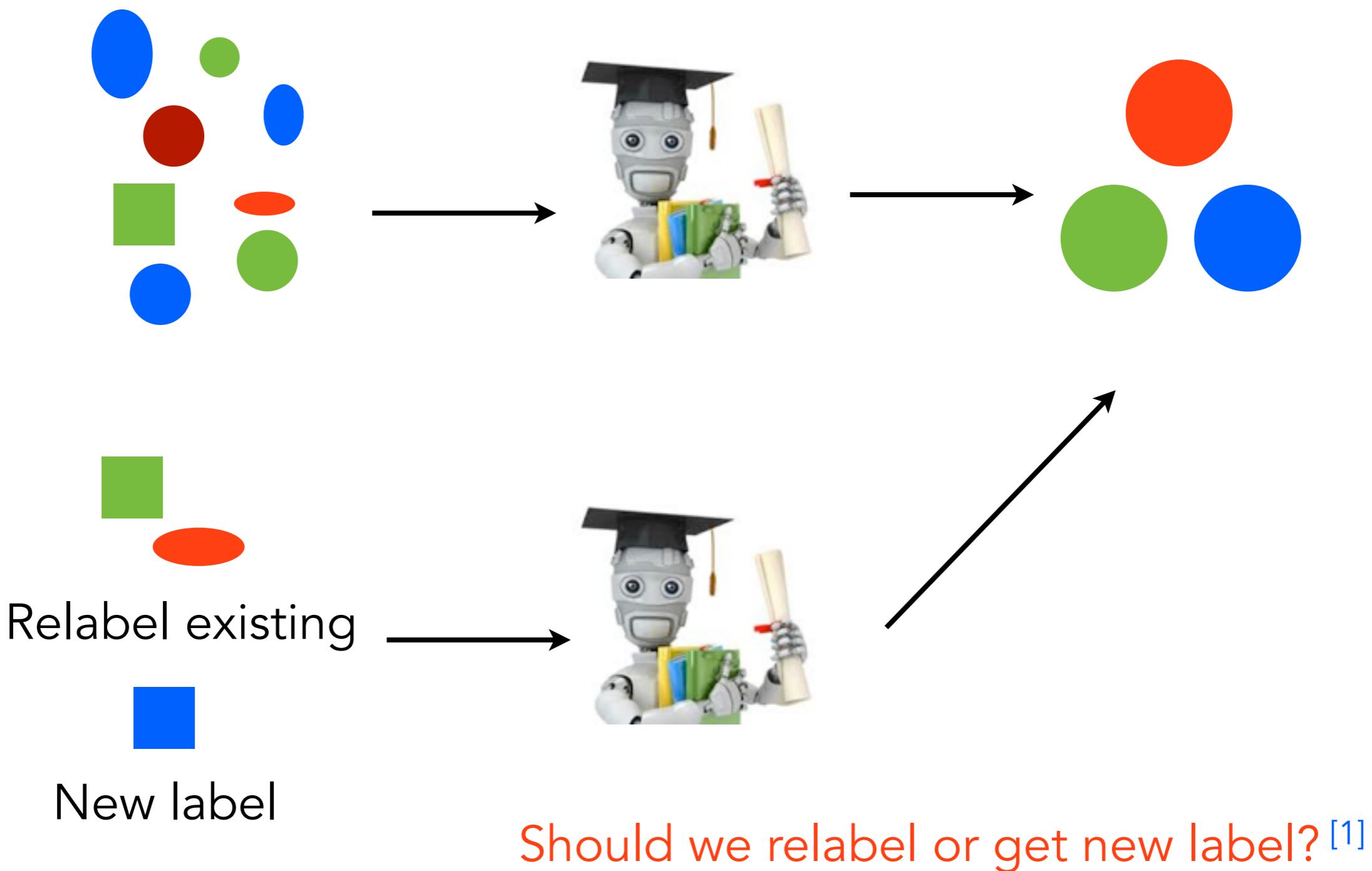
Learning from the Crowd

Closing the Loop



Learning from the Crowd

Closing the Loop

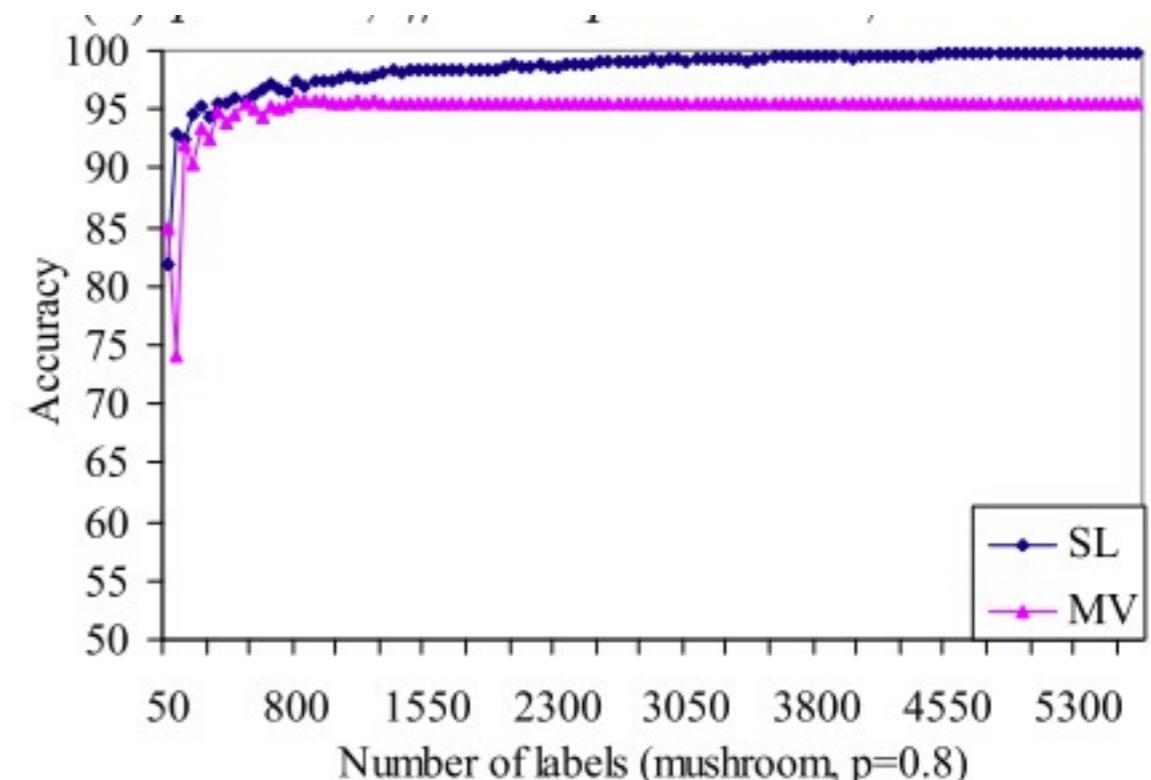
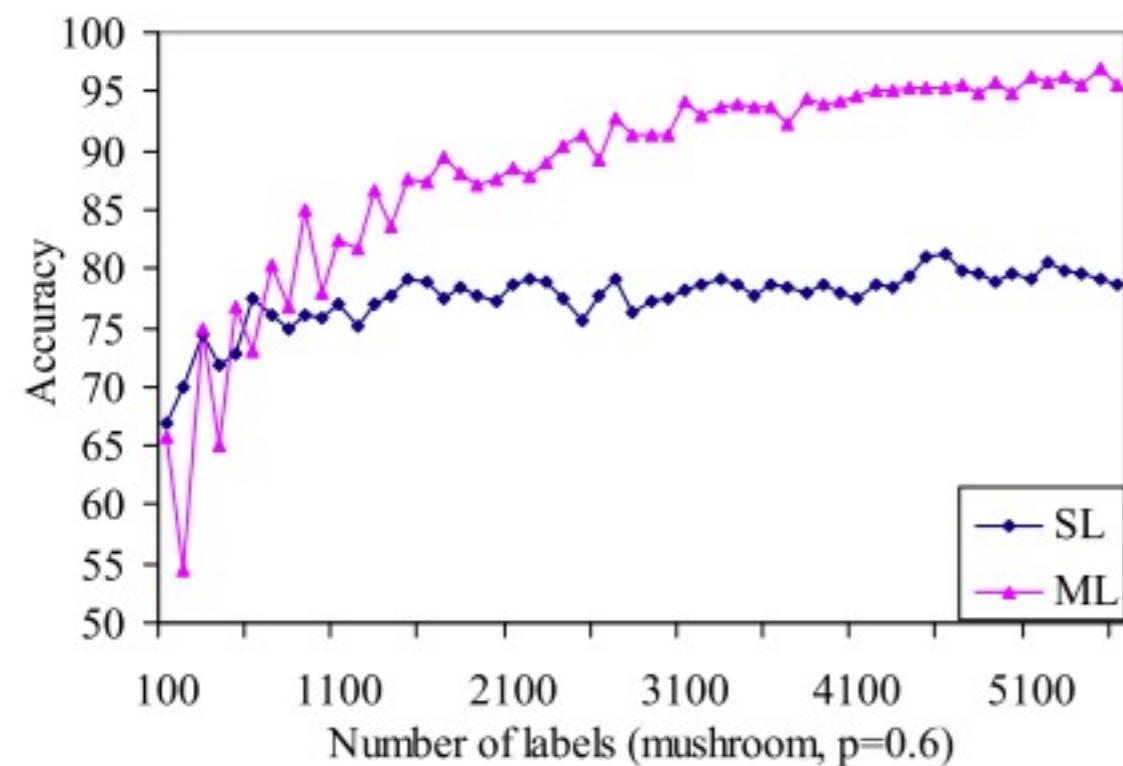


[1] Settles, B. Active learning literature survey. 2010

Learning from the Crowd

Closing the Loop

Repeated labeling improves the overall data collected. [1]



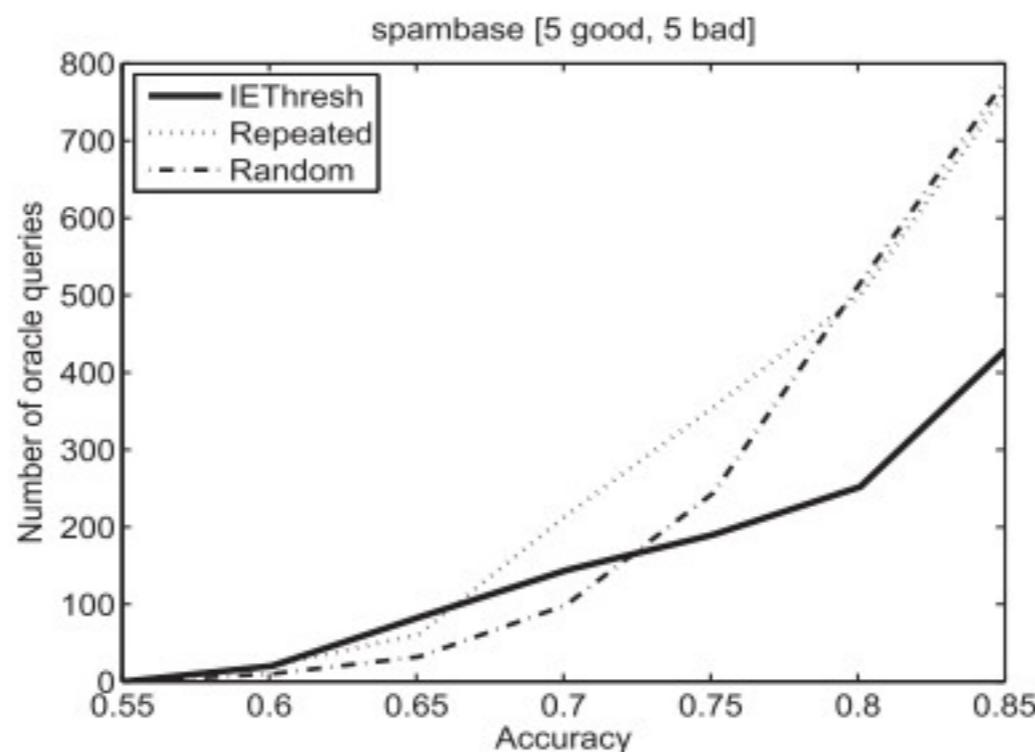
Repeated labeling when labeling is **noisy**

Took into account model and label uncertainty

[1] Sheng, V., Provost, F., & Ipeirotis, P.
Get another label? improving data quality and data mining using multiple, noisy labelers. 2008

Closing the Loop

Can further improve accuracy by choosing expert labelers [1]



Compute a confidence level for each participant

On next iteration, pick a better participant for a unlabeled instance

[1] Donmez, P., Carbonell, J. G., & Schneider, J.
Efficiently learning the accuracy of labeling sources for selective sampling. 2009

Discussion

Active learning gives a methodology to iteratively improve models with humans in the loop.

Inexpensive

Crowdsourcing gives us access to inexpensive pools of unlabeled data and noisy labelers.

Improving Uncertainty

We can begin to answer questions that normally had no clear model through uncertainty sampling and repeated labeling.

Scalable

As soon as we have a model we can apply that model to find solutions where before we had to use human labelers

Applications in Cultural Heritage



Applications in Cultural Heritage

Increased digitization

Increasing digitization,
preservation, exploration efforts



Engage participants with the
reason why the digital collections
exists

Applications in Cultural Heritage

Interpretation

No Knowledge ?



Some Knowledge ?



More Than a Little Knowledge ?



Brooklyn Museum: Click! Crowd curation of photography ^[1]

[1] Proctor, N. (2010).

Digital: Museum as Platform, Curator as Champion, in the Age of Social Media.

Preservation - Transcription



Using symbolic reasoning
to transcribe ancient greek
texts



CAPTCHAs uses to digitize
old texts

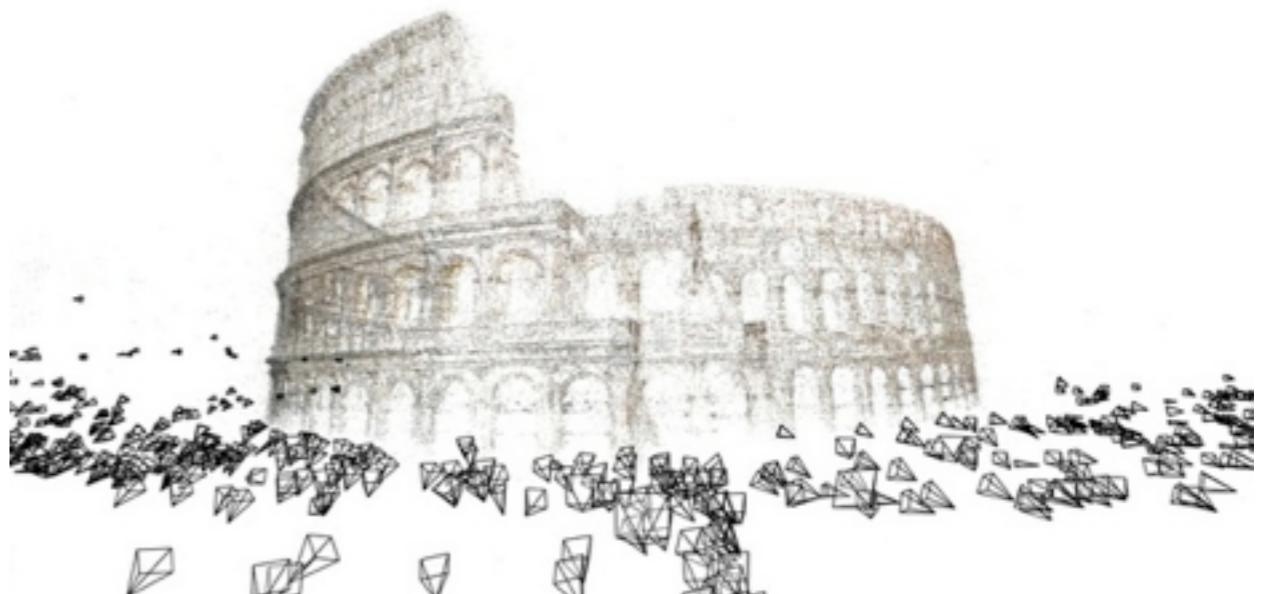
Applications in Cultural Heritage

Preservation - Modeling



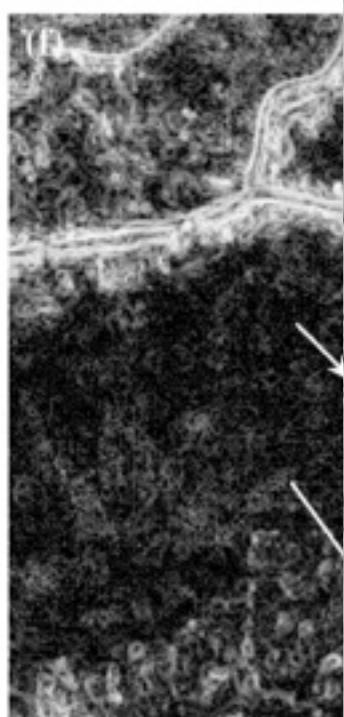
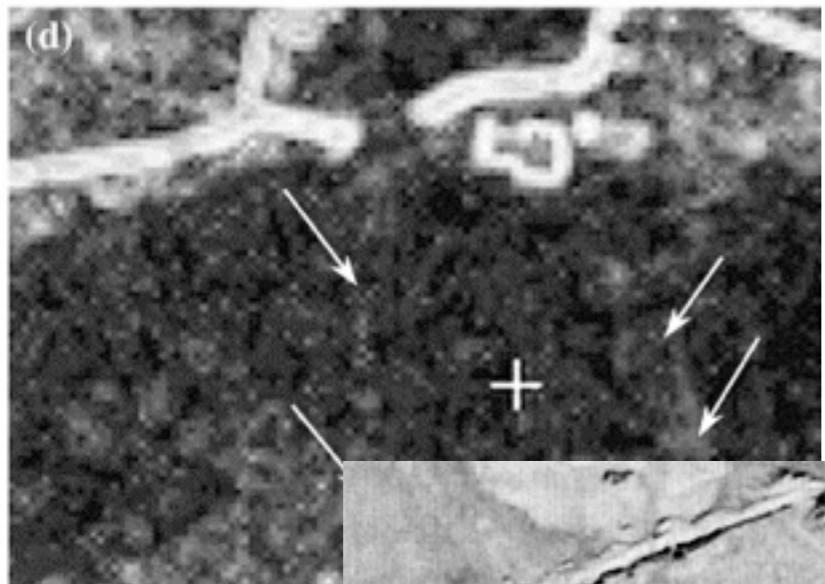
Traditional methods expensive
and time consuming.

Structure from motion utilizing
crowdsourced images has
great potential.



Applications in Cultural Heritage

Exploration & Discovery

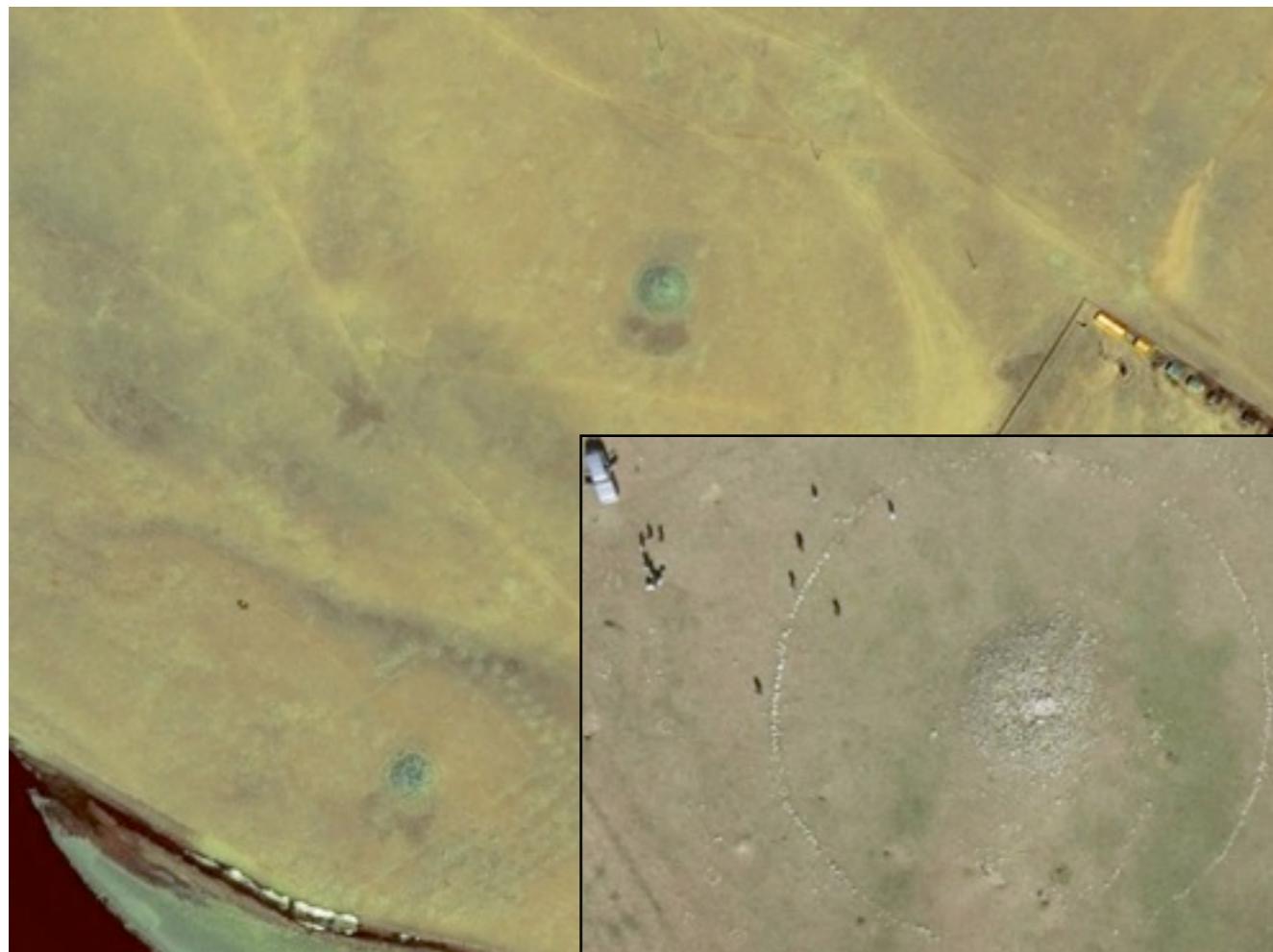


Large amounts of satellite data available for analysis



Exploration & Discovery

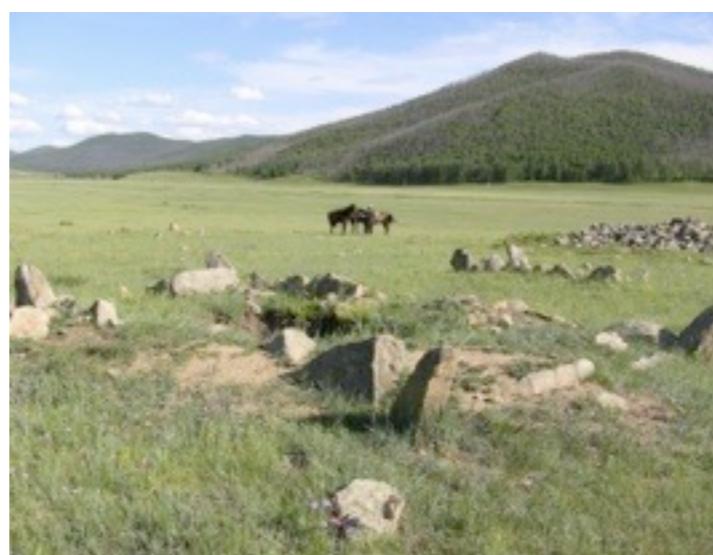
Valley of the Khans



100 sites visited.

55 archaeological sites.

3.4 man years of exploration.



Active Crowdsourcing

Crowd Engagement

Data Quality Control

Passive Crowdsourcing

Parasitic Computing

Crowdsensing

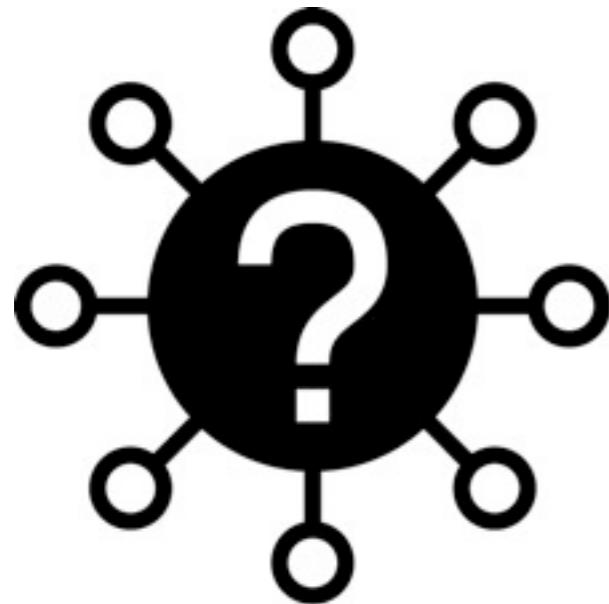
Learning from the Crowd

Active Learning

Closing the Loop

Applications in Cultural Heritage

Towards the Future



Numerous success over the years

Crowdsourcing gives us a platform to collect data

POTENTIAL FOR

Increasing use of passive crowdsourcing

More integration between humans and machine learning

Increased use of human cognition to solve problems



Towards the Future



Increased digitization of cultural heritage

More involvement from beneficiaries

POTENTIAL FOR

Crowd-curation of digital collections

Passive digitization of cultural heritage

Discovering new sites of interest



Thank you for listening!

Questions, comments, concerns.



All References

Corrie, Robert K.

Detection of ancient Egyptian archaeological sites using satellite remote sensing and digital image processing. Proc. of SPIE Vol. Vol. 8181. 2011

Lin, A., Huynh, A., Lanckriet, G., Barrington, L.

Crowdsourcing the Unknown: The Search for Genghis Khan. 2013 (In review).

Tang, J. C., Cebrian, M., Giacobe, N. a., Kim, H.-W., Kim, T., & Wickert, D. "Beaker."

Reflecting on the DARPA Red Balloon Challenge. Communications of the ACM. 2011

Mason, W., & Watts, D. J.

Financial incentives and the performance of crowds. ACM SIGKDD. 2010

Howe, J.

Crowdsourcing: How the power of the crowd is driving the future of business. 2008

Belleflamme, P.

Crowdfunding: Tapping the right crowd. AFFI. 2011

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K.

The future of citizen science: emerging technologies and shifting paradigms. Frontiers in Ecology and the Environment. 2012

Ahn, L. von, & Dabbish, L.

Labeling images with a computer game. 2004

All References

Ahn, L. von, Maurer, B., McMillen, C., Abraham, D., & Blum, M.

reCAPTCHA: human-based character recognition via Web security measures. 2008

Bernstein, M., Little, G., & Miller, R.

Soylent: a word processor with a crowd inside. Proceedings of the 23nd annual ACM symposium on User interface software and technology. 2010

Chen, K.-T., Wu, C.-C., Chang, Y.-C., & Lei, C.-L.

A crowdsourceable QoE evaluation framework for multimedia content. 2009

Welinder, P., Branson, S., Belongie, S., & Perona, P.

The Multidimensional Wisdom of Crowds. 2010

Adar, E.

Why I hate Mechanical Turk research (and workshops). 2011

Barabási, a L., Freeh, V. W., Jeong, H., & Brockman, J. B.

Parasitic computing. 2001

Anderson, D., Cobb, J., & Korpela, E.

SETI@ home: an experiment in public-resource computing. 2002

Settles, B.

Active learning literature survey. 2010

Sheng, V., Provost, F., & Ipeirotis, P.

Get another label? improving data quality and data mining using multiple, noisy labelers. 2008

All References

Donmez, P., Carbonell, J. G., & Schneider, J.

Efficiently learning the accuracy of labeling sources for selective sampling. 2009

Proctor, N.

Digital: Museum as Platform, Curator as Champion, in the Age of Social Media. 2010

Image Credits

Pac Man designed by Paulo Volkova from The Noun Project

Brain designed by Arjun Adamson from The Noun Project

Passive Crowdsourcing

Parasitic Computing

a

$$P = (x_1 \oplus x_2) \wedge (x_3 \oplus x_4) \wedge (x_5 \oplus x_6) \wedge (x_7 \oplus x_8) \wedge (x_9 \wedge x_{10}) \wedge (x_{11} \oplus x_{12}) \wedge (x_{13} \wedge x_{14}) \wedge (x_{15} \oplus x_{16})$$

b

X	Y	$X \oplus Y$	$X \wedge Y$	$X + Y$
0	0	0	0	00
0	1	1	0	01
1	0	1	0	01
1	1	0	1	10

2SAT variables

c $M = \boxed{0x_1 \ 0x_3 \ 0x_5 \ 0x_7 \ 0x_9 \ 0x_{11} \ 0x_{13} \ 0x_{15}}$ $\boxed{0x_2 \ 0x_4 \ 0x_6 \ 0x_8 \ 0x_{10} \ 0x_{12} \ 0x_{14} \ 0x_{16}}$
 $E = \boxed{01 \ 00 \ 01 \ 01 \ 00 \ 01 \ 01 \ 01}$ $\boxed{00 \ 00 \ 01 \ 00 \ 01 \ 01 \ 01 \ 00}$

S_1

S_2

d

$$\begin{array}{ccccccccc} 0x_1 & 0x_3 & 0x_5 & 0x_7 & 0x_9 & 0x_{11} & 0x_{13} & 0x_{15} \\ 0x_2 & 0x_4 & 0x_6 & 0x_8 & 0x_{10} & 0x_{12} & 0x_{14} & 0x_{16} \end{array}$$

$$\begin{array}{cccccccc} \oplus & \wedge & \oplus & \wedge & \oplus & \wedge & \oplus & \wedge \\ \hline 01 & 10 & 01 & 01 & 10 & 01 & 10 & 01 \end{array}$$

$$\begin{array}{ccccccccc} S_1 & 01 & 00 & 01 & 01 & 00 & 01 & 01 & 01 \\ S_2 &) & 00 & 00 & 01 & 00 & 01 & 01 & 00 \end{array}$$

$$\begin{array}{ccccccccc} SUM & 01 & 00 & 10 & 01 & 01 & 10 & 10 & 01 \\ SUM & 10 & 11 & 01 & 10 & 10 & 01 & 01 & 10 \end{array}$$

(Real checksum)

10 01 10 10 01 10 01 10

T_c

Construct message

f

Transmitted message

1001101001100110|0100010100010101|0000010001010100

T_c

S_1

S_2