Lecture 2

The humans-in-the-loop

Ranjay Krishna | ranjay@cs.washington.edu

Course logistics

Assignment 1 due in 2 days.

- It should be easy and not take much time.
- I am looking for you to be insightful. It’s quite open ended.
- 3-4 minute presentation for class.
- 1 min for QA.

Assignment 1: Reflections on personal AI use

Your goal is to reflect on your personal usage of AI applications. You can approach this assignment a number of ways. Feel free to be creative! Here are some example ways of completing the assignment:

- you could take a data-driven approach to track or **measure** some aspect of your reliance on an AI application for a week.
- You could do a retrospective analysis of your **own interactions** with AI systems or that of a **community** that you are active in.
- You could **spend time attempting to interact with an AI model** in some way, such as switching to a new technology and reporting back on the experience.
- You could **interview or survey an AI engineer or AI product designer**.
- You could talk about the possible **societal or behavioral implications** of a new emerging technology.

Slack and canvas - our two main forms of communication

We have a slack channel for discussions.

- If you are not part of it, email Jiafei (duanj1@cs)
- We will redundantly make announcements on both slack and canvas

A space where you can organize yourselves for discussions and projects

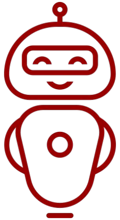
Course project

Project teams:

- If you are looking for a team or want a team member, please post on #project-team-search

- start thinking about course project ideas. Feel free to message us with questions

Recap: looking at how the fields evolved together



Artificial Intelligence

Goal: to create an artificial rival to human intelligence

Artifact: models of human intelligence

Long time horizon

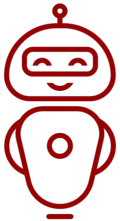
Human-Computer Interaction

Goal: To improve applications as they approach widespread use

Artifact: designs for mass market products

Short time horizon

AI is now finally in mass market use



Artificial Intelligence

Goal: to align AI with human intelligence

Artifact: models for mass market use

No longer for long time horizon

The three AI winters and how HCI thrived. Perhaps this time, both will.

Happening last night

[link]

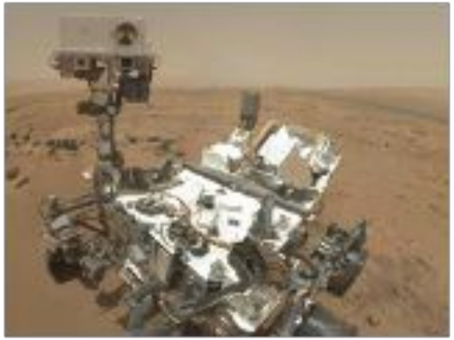
Lecture 2

The humans strike back,
The humans-in-the-loop

Ranjay Krishna | ranjay@cs.washington.edu

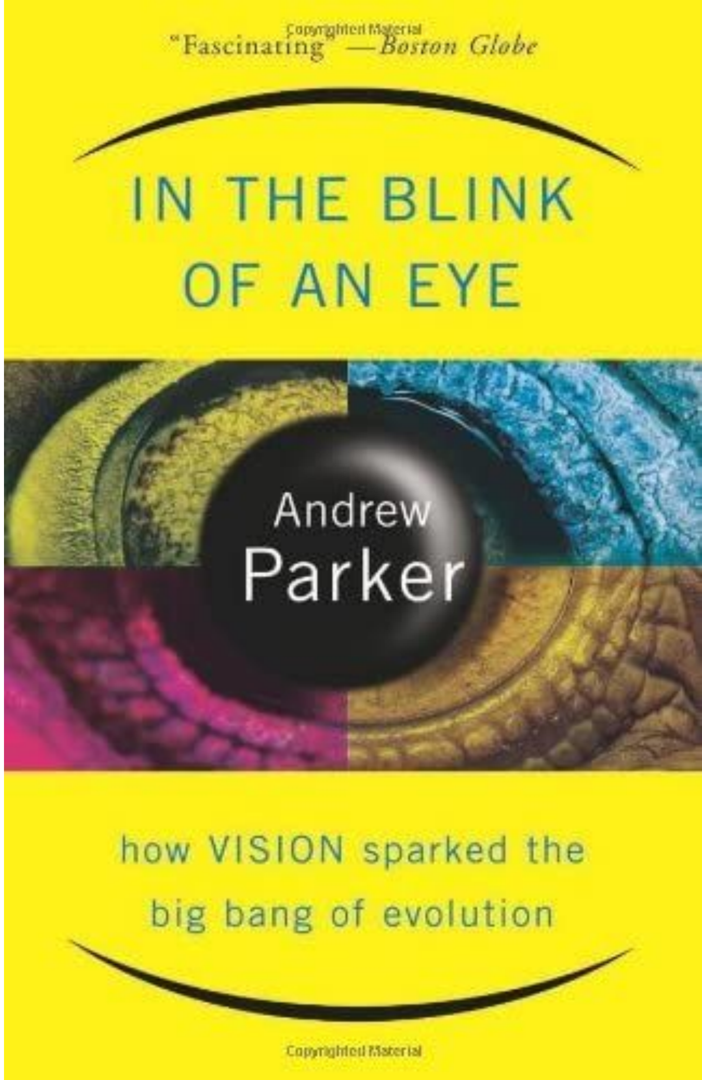
Humans in the loop?


The European Camera of the Year 1983.
Fortunately, fine photography is a universal language.
PENTAX
PENTAX



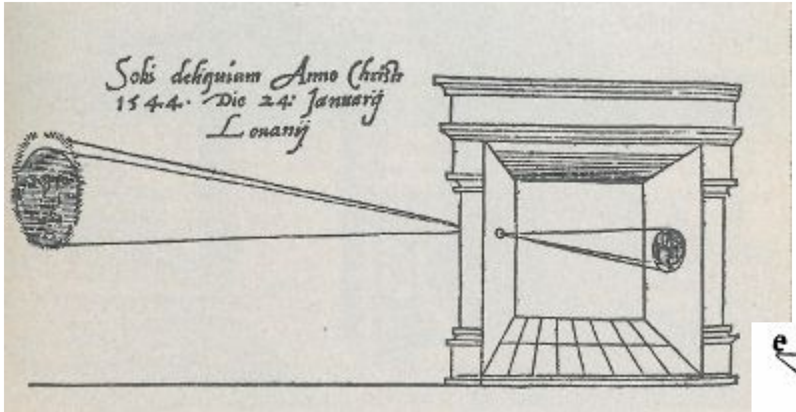
Vision is core to the evolution of intelligence

543 million years ago.



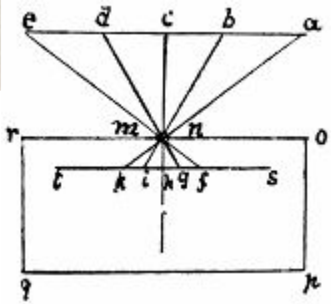
Ranjay Krishna | ranjay@cs.washington.edu

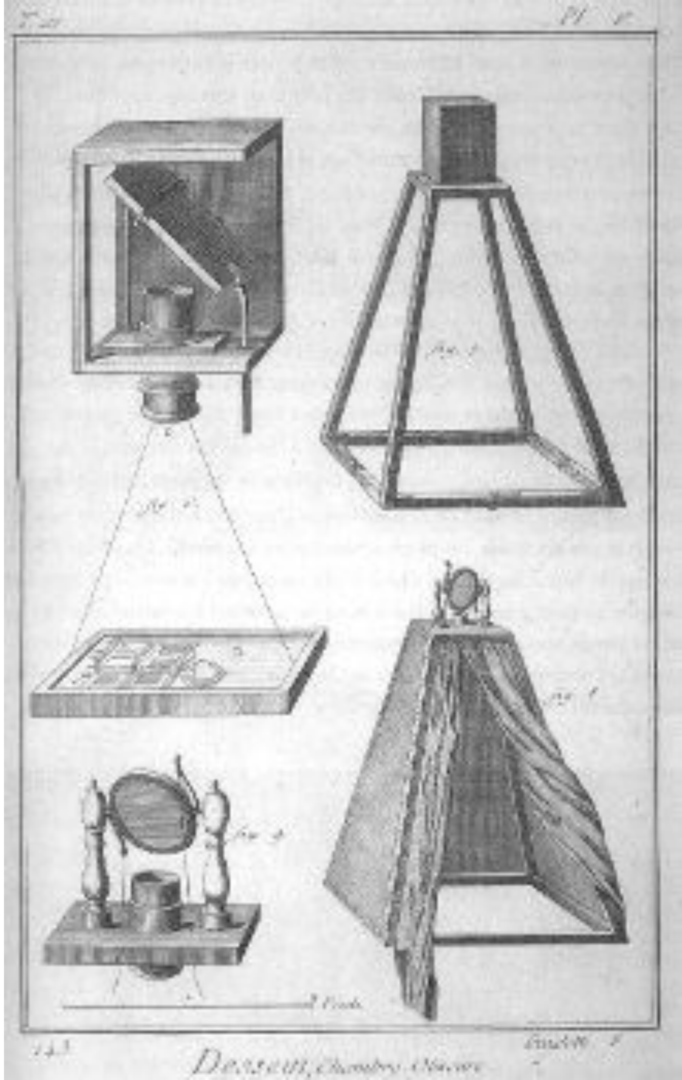
The first attempts at capturing the visual world



Camera obscura by Gemma Frisius, 1545

Inspired Leonardo da Vinci,
16th Century AD





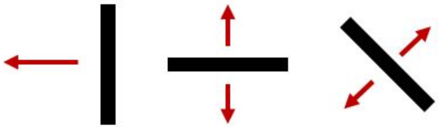
Examples from 18th century Encyclopedia

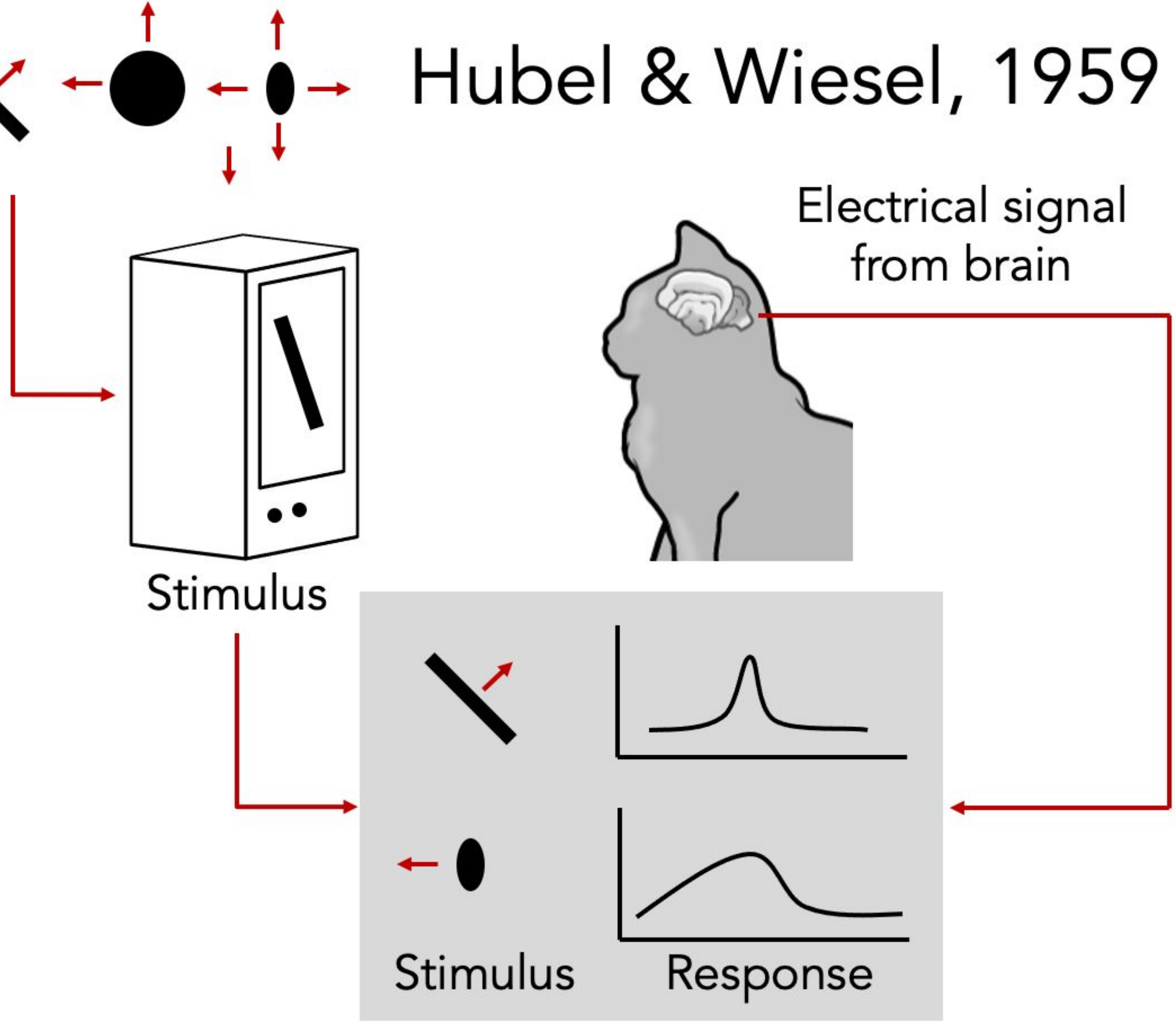
Ranjay Krishna | ranjay@cs.washington.edu

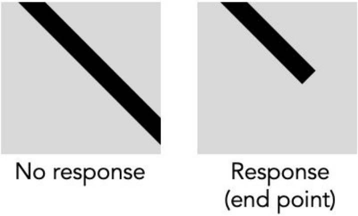
Hubel & Wiesel, 1959

How does animal vision work?

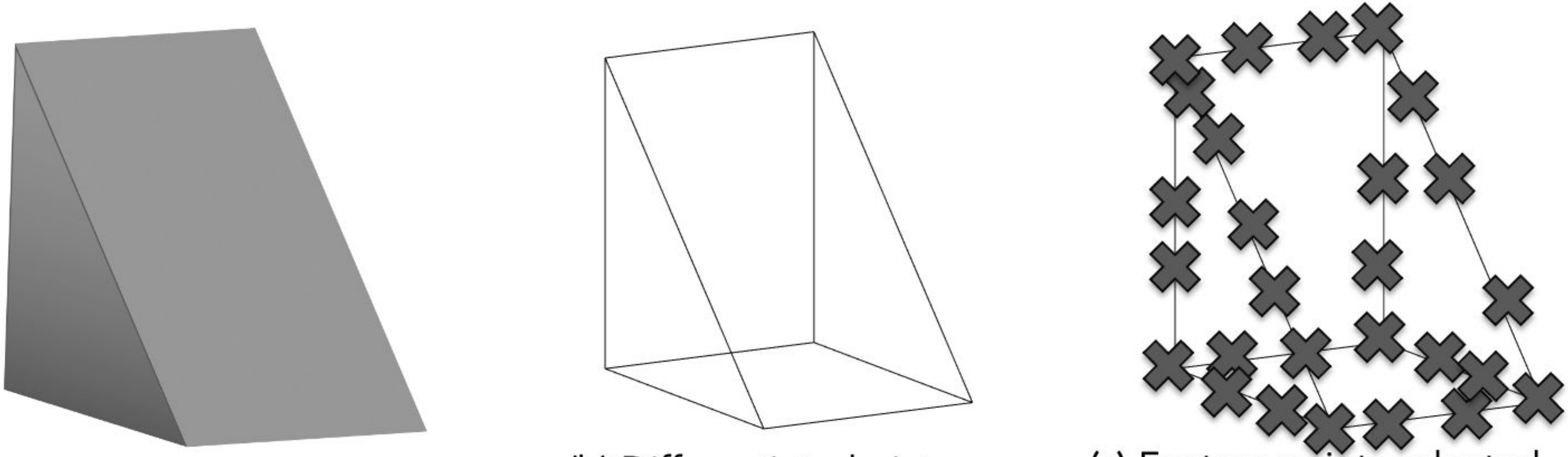
Won Nobel Prize in 1981
Visual processing is hierarchical, involving recognizing simpler structures, edges, etc.







Larry Roberts - Father of computer vision



(a) Original picture (b) Differentiated picture (c) Feature points selected

Synthetic images, building up the visual world from simpler structures

The summer vision project

Organized by Seymour Papert

Computer vision was meant to be just a simple summer intern project

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966

Vision Memo. No. 100.

THE SUMMER VISION PROJECT

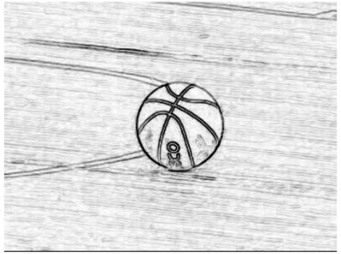
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

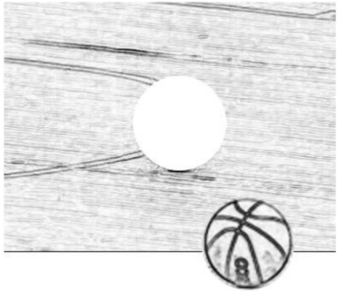
Input image

This image is CC0 1.0 public domain

Edge image



2 ½-D sketch



3-D model

This image is CC0 1.0 public domain

Input Image	Primal Sketch	2 ½-D Sketch	3-D Model Representation
Perceived intensities	Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries	Local surface orientation and discontinuities in depth and in surface orientation	3-D models hierarchically organized in terms of surface and volumetric primitives

David Marr, Stages of Visual Representation, 1970

Until the 90s,
computer vision was not broadly applied to real world images

The focus was on algorithms!

Shi & Malik, *Normalized Cut*, 1997

First commercial success of computer vision

It came from embracing machine learning in 2001.

Does anyone know what it was?

First commercial success of computer vision

Real time face detection using using an algorithm by Viola and Jones, 2001

- Fujifilm face detection in cameras
- [HP patent]() immediately

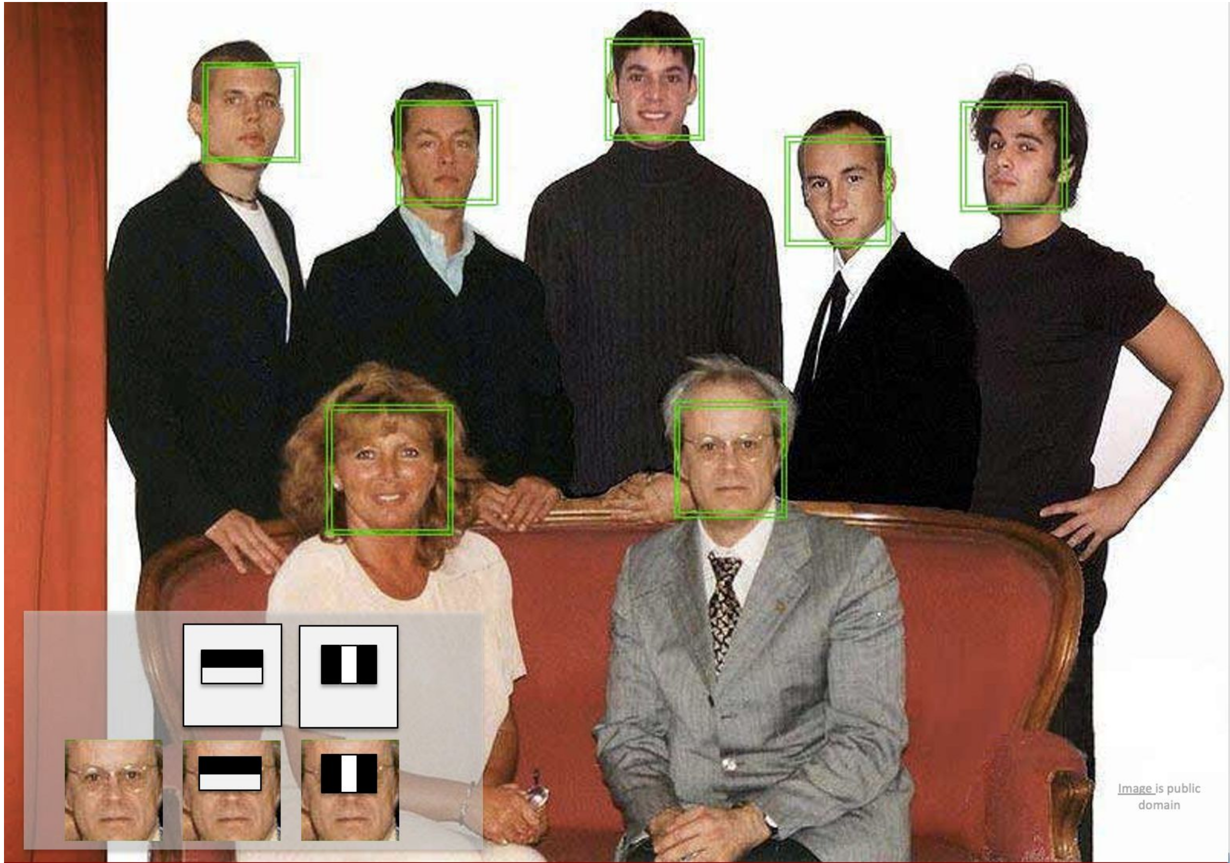


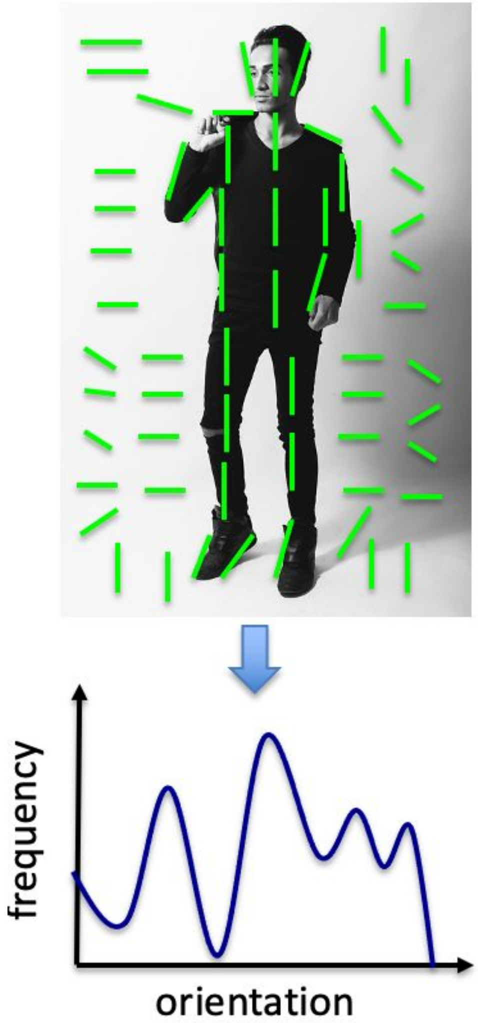
Image is public domain

Designing better feature extraction became the focus

HoG features

- Histogram of oriented gradients
- Handcrafted

[Dalal & Triggs, HoG. 2005]



IMAGENET

www.image-net.org

22K categories and **14M** images

- Animals
  - Bird
  - Fish
  - Mammal
  - Invertebrate
- Plants
  - Tree
  - Flower
- Food
- Materials
- Structures
- Artifact
  - Tools
  - Appliances
  - Structures
- Person
- Scenes
  - Indoor
  - Geological Formations
- Sport Activities

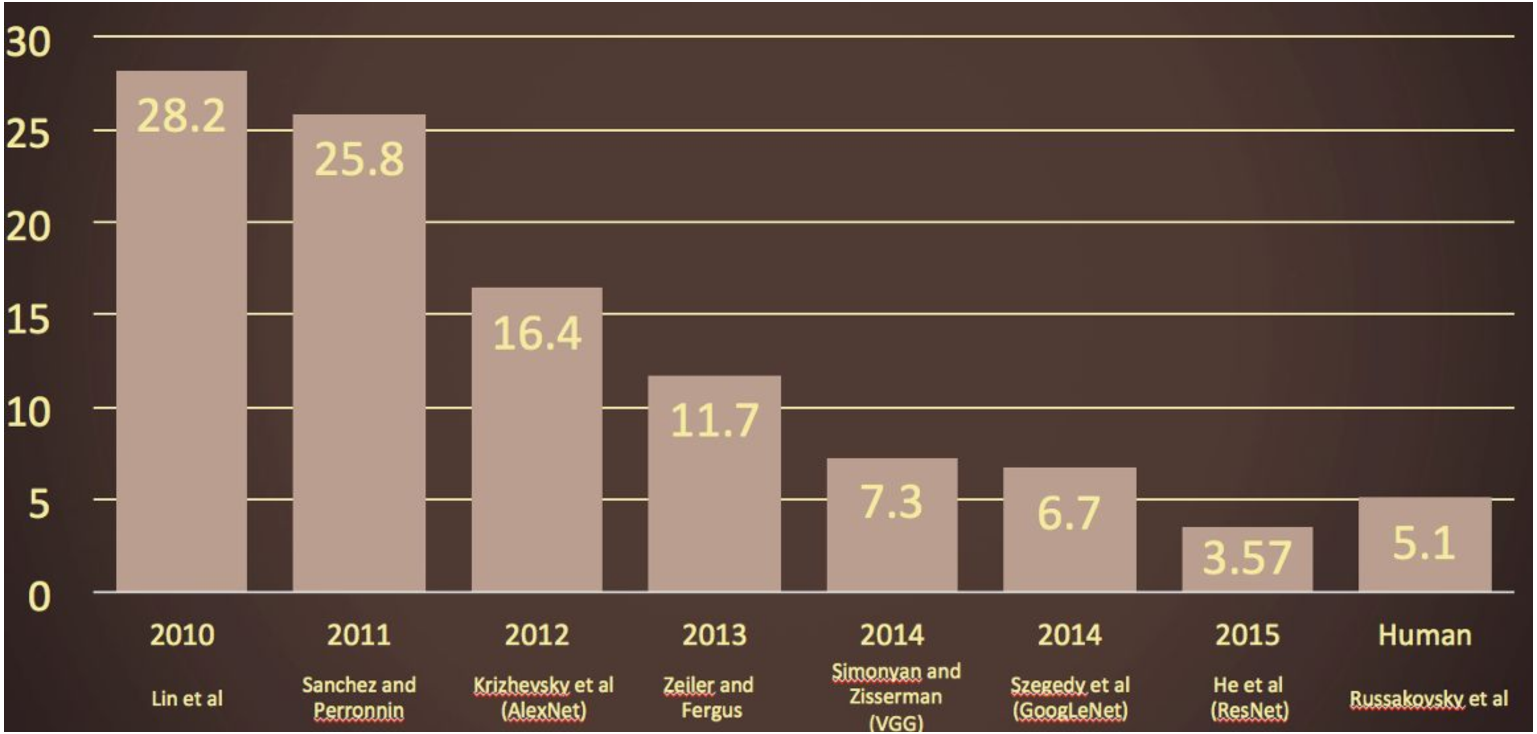
Deng, Dong, Socher, Li, Li, & Fei-Fei, 2009

Hypothesis behind ImageNet

- A child sees nearly 3K unique objects by the age of 6
- Calculated by Irving Biederman
    - [Biederman. Recognition-by-components: a theory of human image understanding. 1983]

- But computer vision algorithms are trained on a handful of objects.

Object recognition accuracy drops year after year



Year 2010

NEC-UIUC

Dense descriptor grid: HOG, LBP

↓

Coding: local coordinate, super-vector

↓

Pooling, SPM

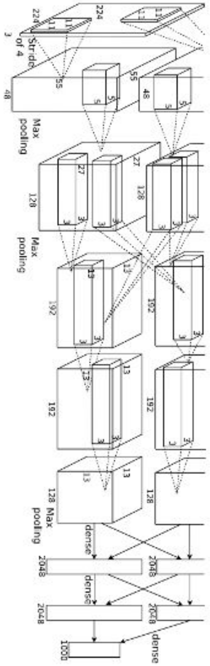
↓

Linear SVM

[Lin CVPR 2011]

Year 2012

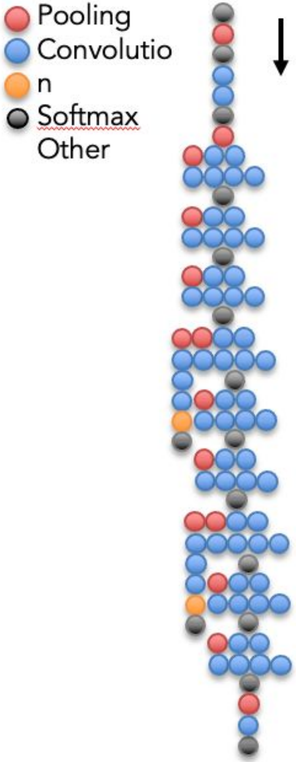
SuperVision



[Krizhevsky NIPS 2012]

Year 2014

GoogLeNet



[Szegedy arxiv 2014]

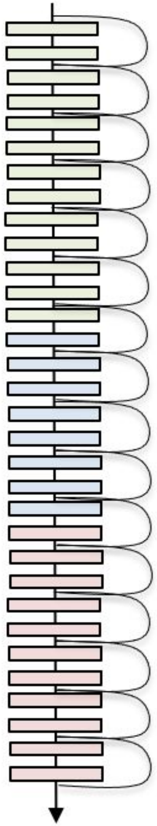
VGG

- Image
- conv-64
- conv-64
- maxpool
- conv-128
- conv-128
- maxpool
- conv-256
- conv-256
- maxpool
- conv-512
- conv-512
- maxpool
- conv-512
- conv-512
- maxpool
- fc-4096
- fc-4096
- fc-1000
- softmax

[Simonyan arxiv 2014]

Year 2015

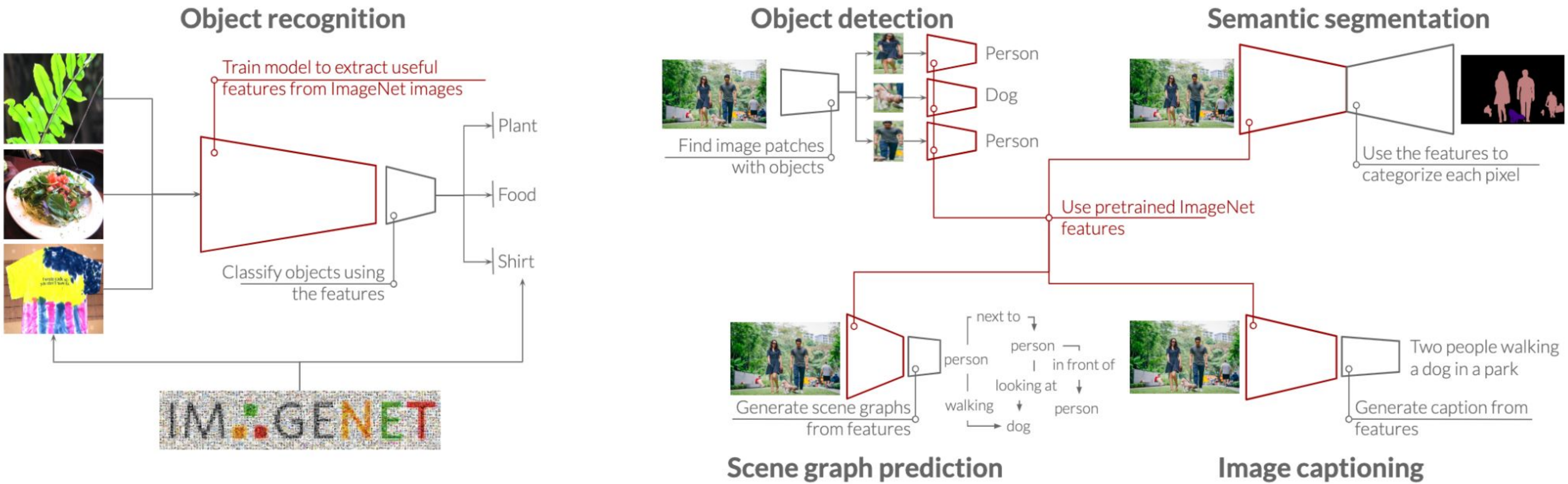
MSRA



[He ICCV 2015]

Data hungry machine learning models are now everywhere

Pretraining on ImageNet for object classification → Transfer ImageNet features for many other tasks:



What we don’t often talk about

How was ImageNet created?

50K human workers!!

1. Create set of search terms

cat : cat feline, cat mammal, cat carnivore, 猫 (chinese), kat (Dutch), gatto/gatta (Italian), gato/gata (Spanish), …

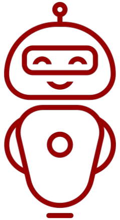
2. Search for images on Google, MSN, Yahoo, Flickr

3. Hire 50K annotators to verify each image

Final dataset with 500-1000 images per category

The humans-in-the-loop

The humans-in-the-loop: two perspectives



Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Human-Computer Interaction

Goal: To support a labor force achieve their financial and career goals

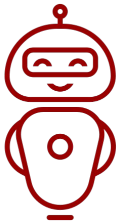
Artifact: automations that structure work

Impacts across long time horizon

Ranjay Krishna | ranjay@cs.washington.edu

The humans-in-the-loop
from an AI perspective

The humans-in-the-loop: two perspectives



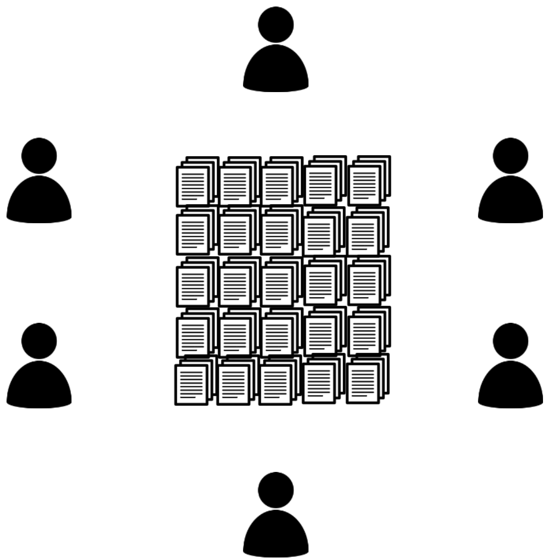
Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Hundreds of thousands of data labeling tasks are completed everyday.



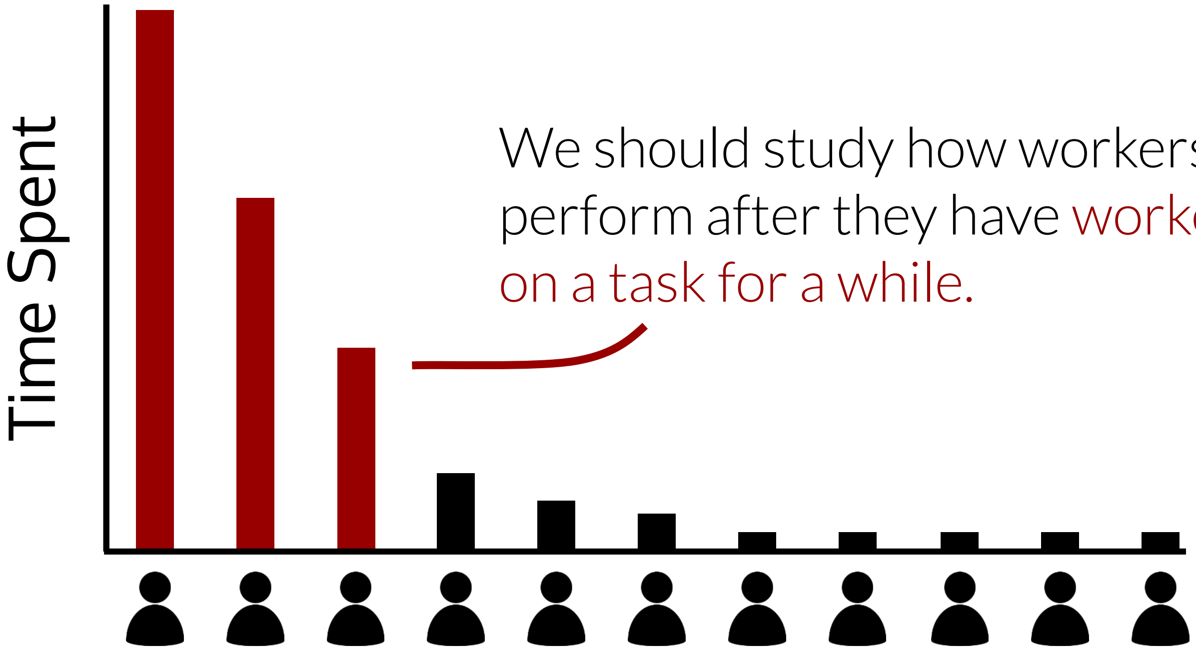
[Little. 2009]

A few workers do most of the work.

These workers could spend
hours, days, or weeks.

[Little. 2009]

Most crowd work is collected by workers who have already completed many of the same task.



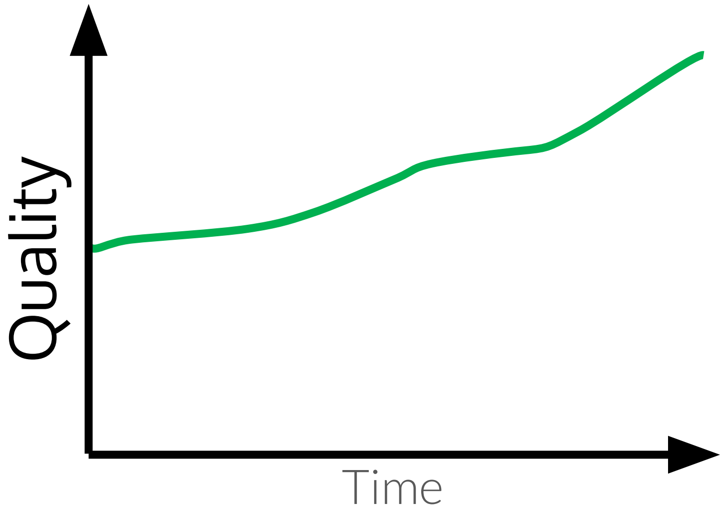
We should study how workers perform after they have worked on a task for a while.

Humans-in-the-loop from an AI perspective:

How does a worker's quality on a certain task change over long periods of time?

[Hata et al. A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality. CSCW 2017]

Conflicting hypotheses from previous work

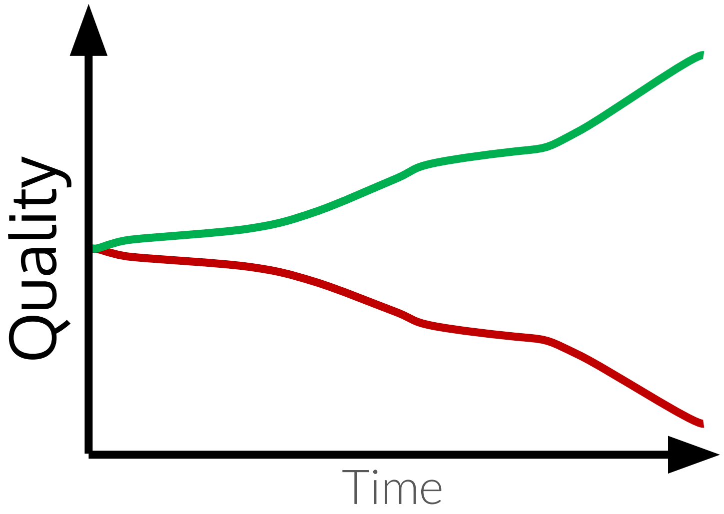


Quality increases over time:

Familiarity with a task builds expertise.
Retaining good workers improves quality.

[Ho et al. 2015] [Dai et al. 2013]

Conflicting hypotheses from previous work



Quality increases over time:

Familiarity with a task builds expertise.
Retaining good workers improves quality.

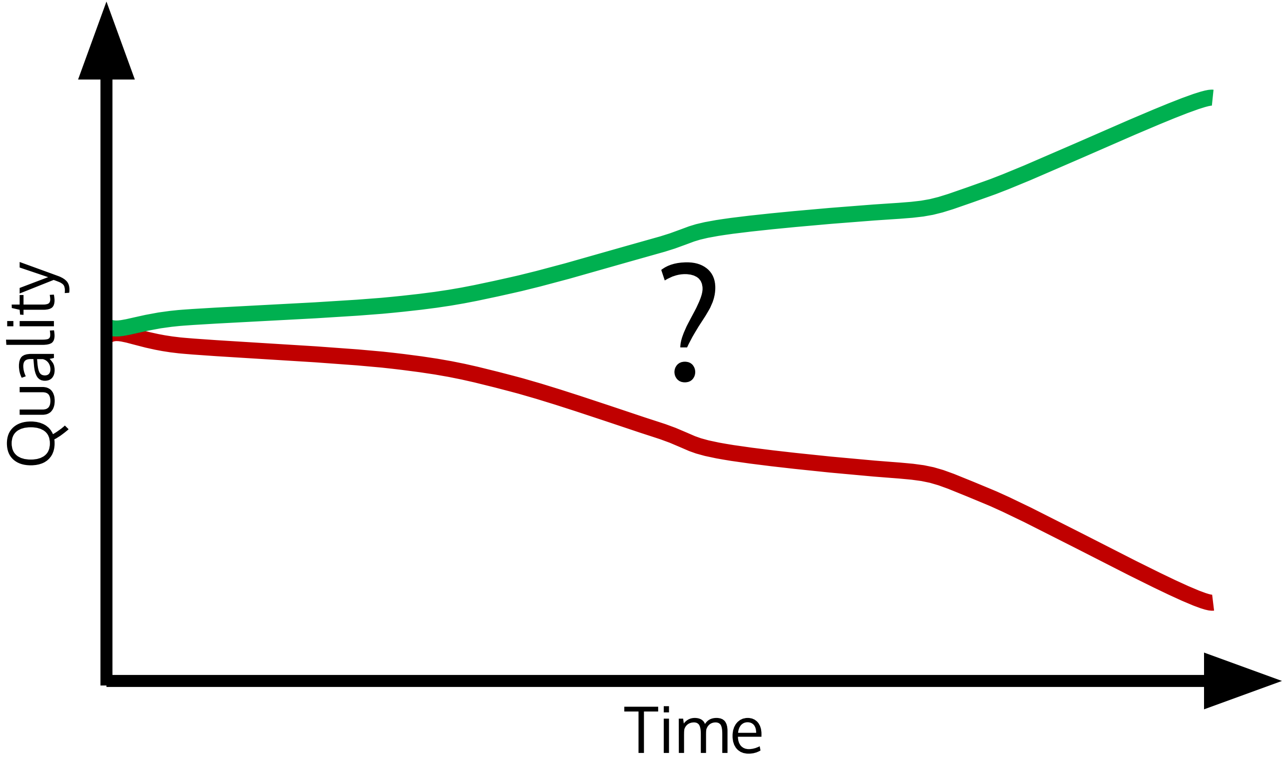
[Ho et al. 2015] [Dai et al. 2013]

Quality decreases over time:

Fatigue reduces productivity and performance.
Workers cannot identify fatigue easily.

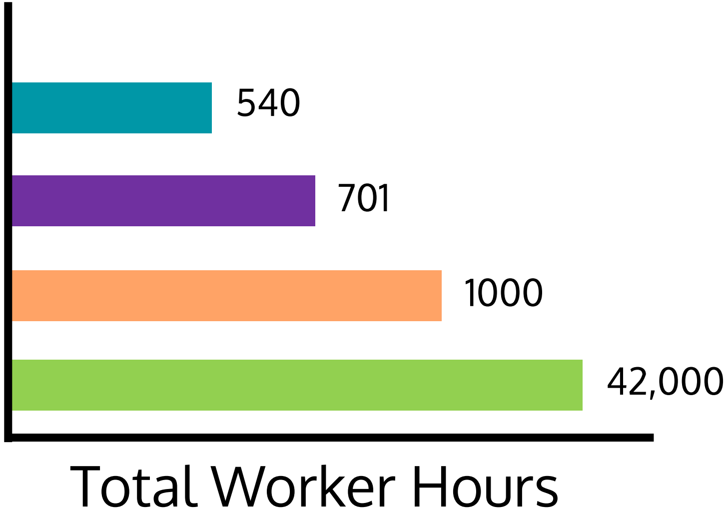
[Perelli. 1980] [Boksem et al. 2008] [Henning et al. 1989]

What does every think? Which theory is correct?



We collected 42K hours of work over several months

Previous Work	Workers	Time Per Worker
Dai et al.	270	1 – 2 hours
Chandler et al.	2471	20 minutes
Law et al.	496	1 – 2 hours
Our study	**815**	**5 – 350+ hours**



[Dai et al.. 2013] [Chandler et al. 2013] [Law et al. 2016]

[Hata et al. A Glimpse Far into the Future: Understanding Long-term Crowd Worker Quality. CSCW 2017]

We analyzed three types of tasks:

Image Descriptions

A dog wearing a hat.

Question-Answer Pairs

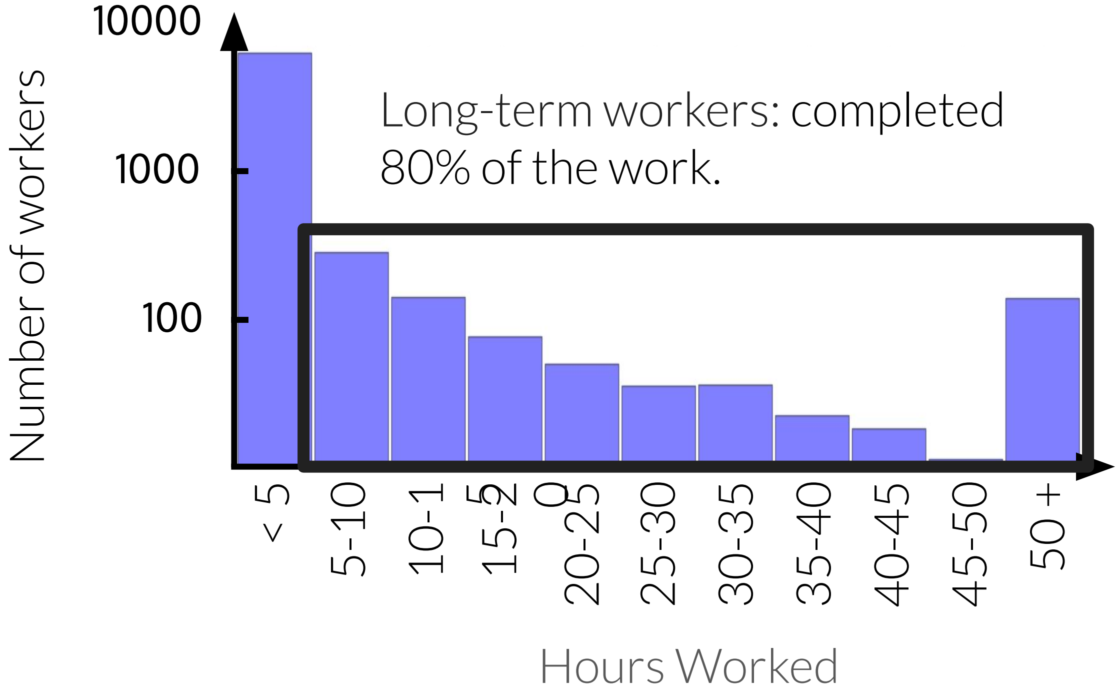
Q: What is that hat made of?

A: Corduroy.

Verification

Voted true to above question-answer pair.

Long-term worker statistics

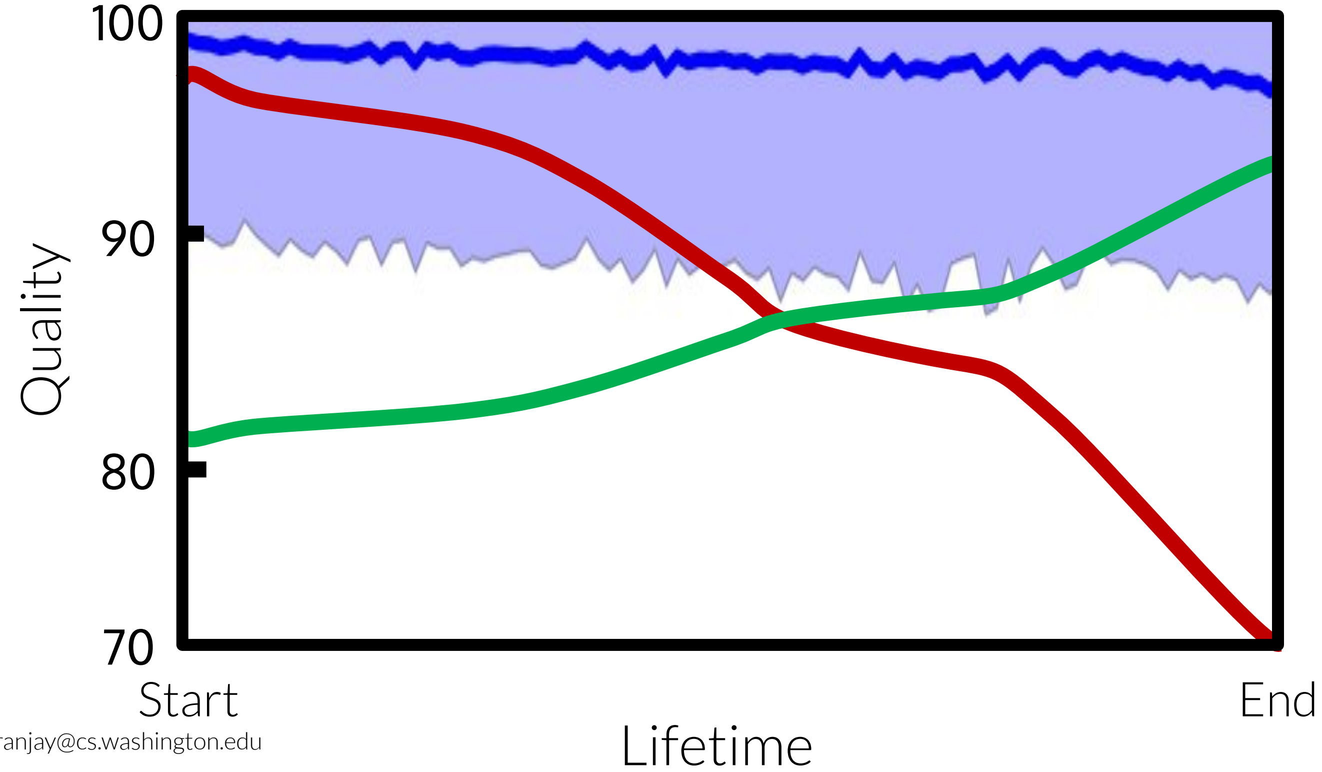


815 long-term workers

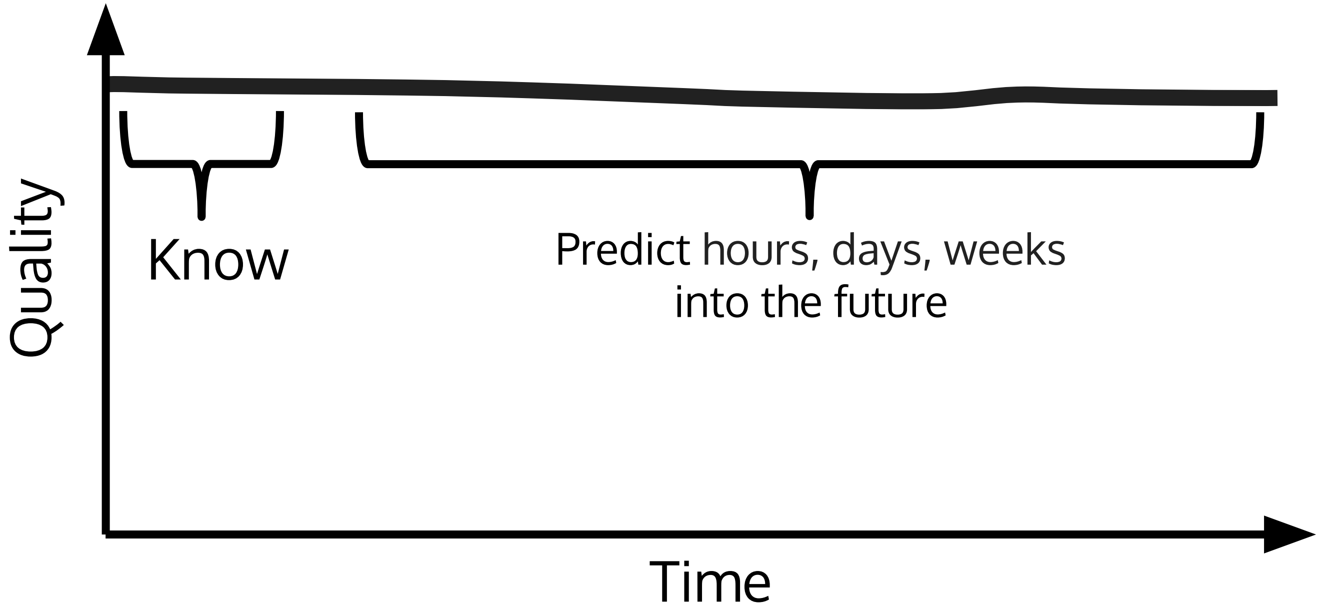
Each worked 5 – 350+ hours

Median of 20 hours

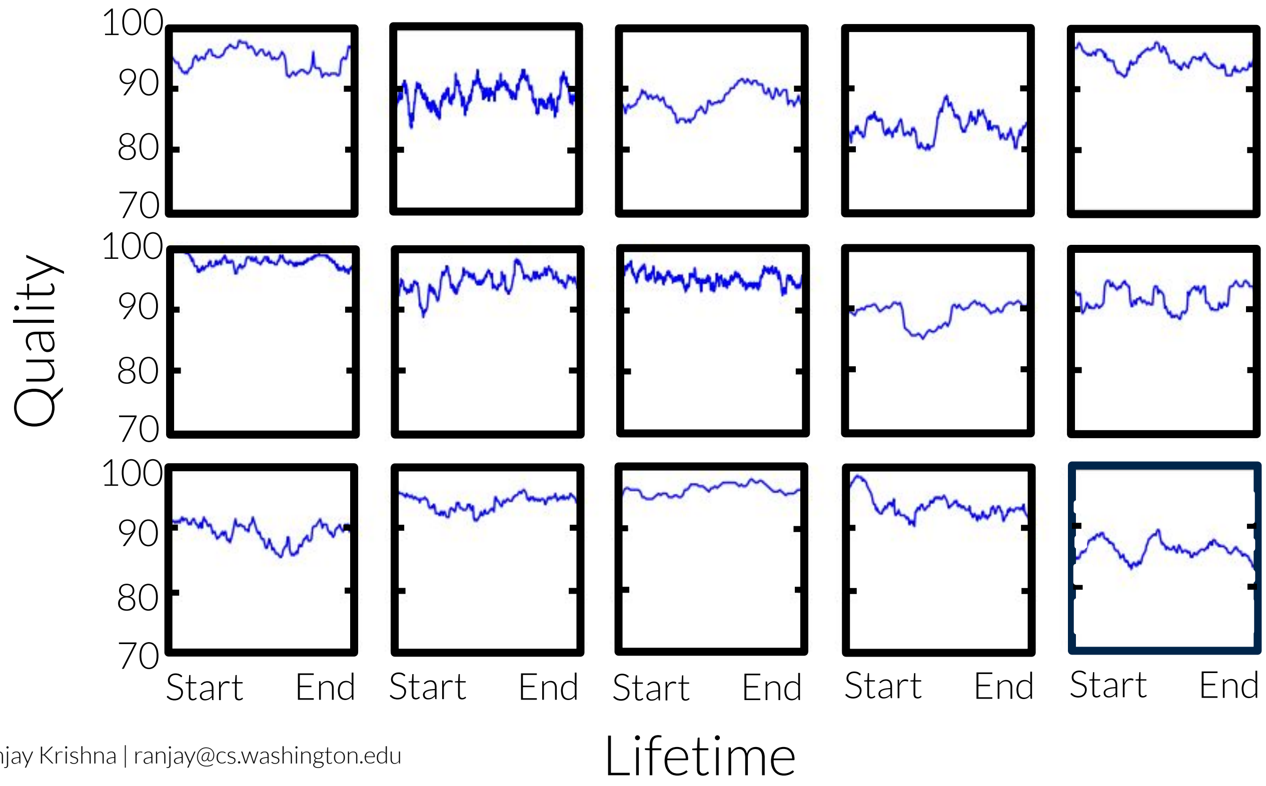
Prior work gives mixed hypotheses



Surprise: crowd workers are surprisingly consistent, allowing us to make accurate quality predictions

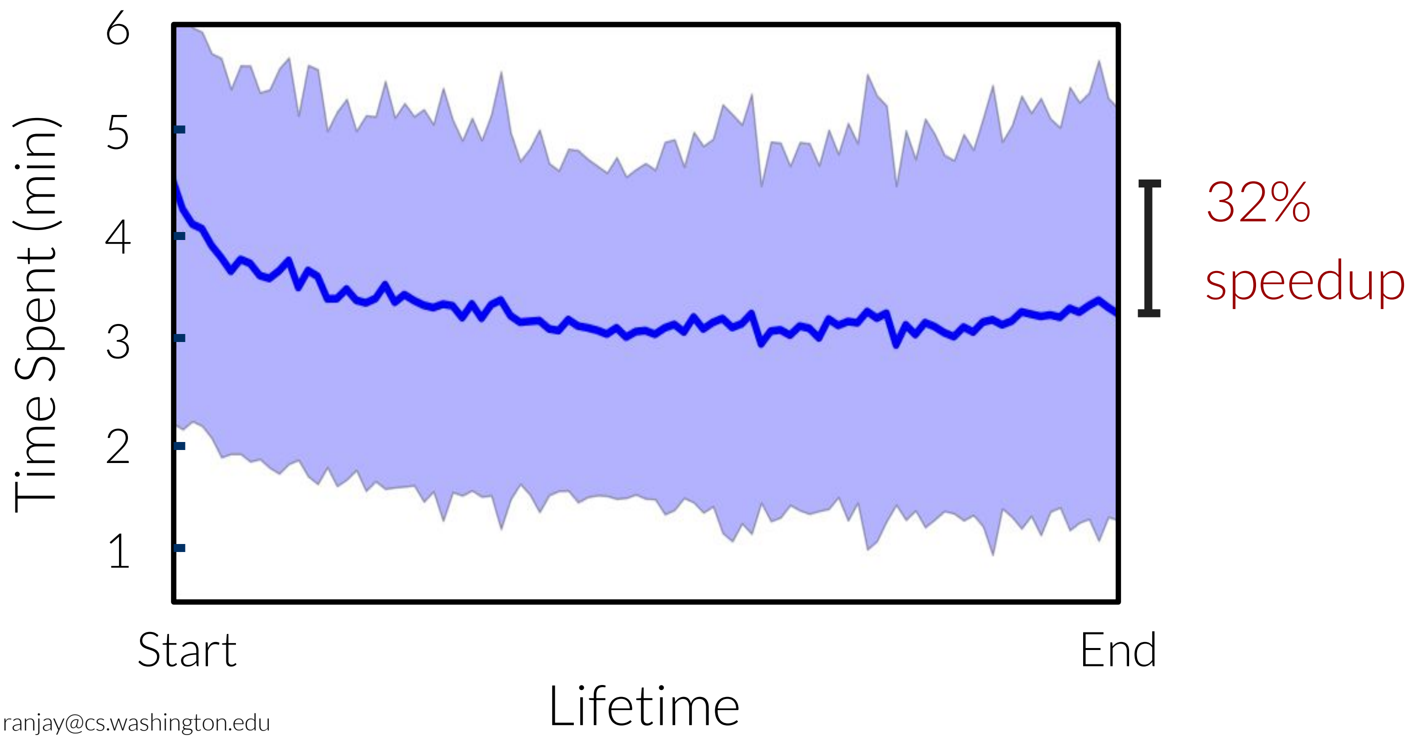


Individual workers are consistent.



Each worker, on average, deviated 3% from their mean quality.

Time spent per task decreases.



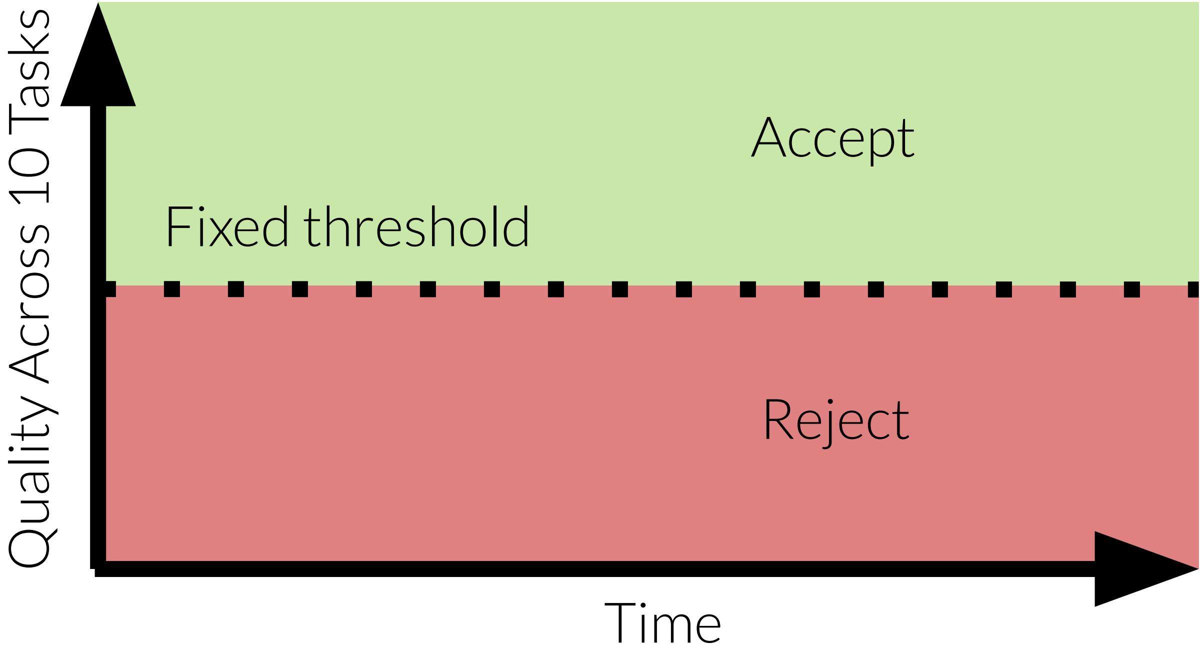
Was the consistency due to the task design?

Crowd workers often do the minimal amount of work required for acceptance.

Was the observed consistency due to strict acceptance criteria?

[Mason et al. 2009] [Chandler et al. 2013]

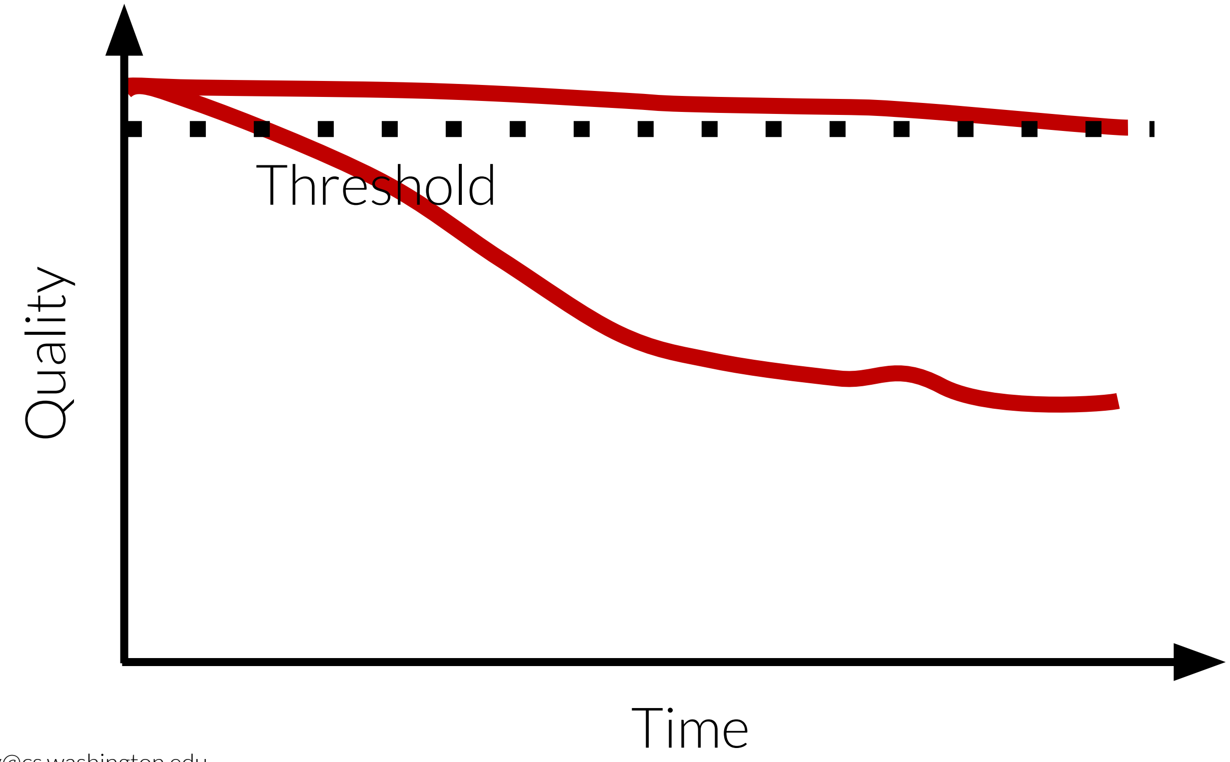
Controlled experiment - work accepted if average of past 10 tasks is above threshold



Collected data from 1134 workers.

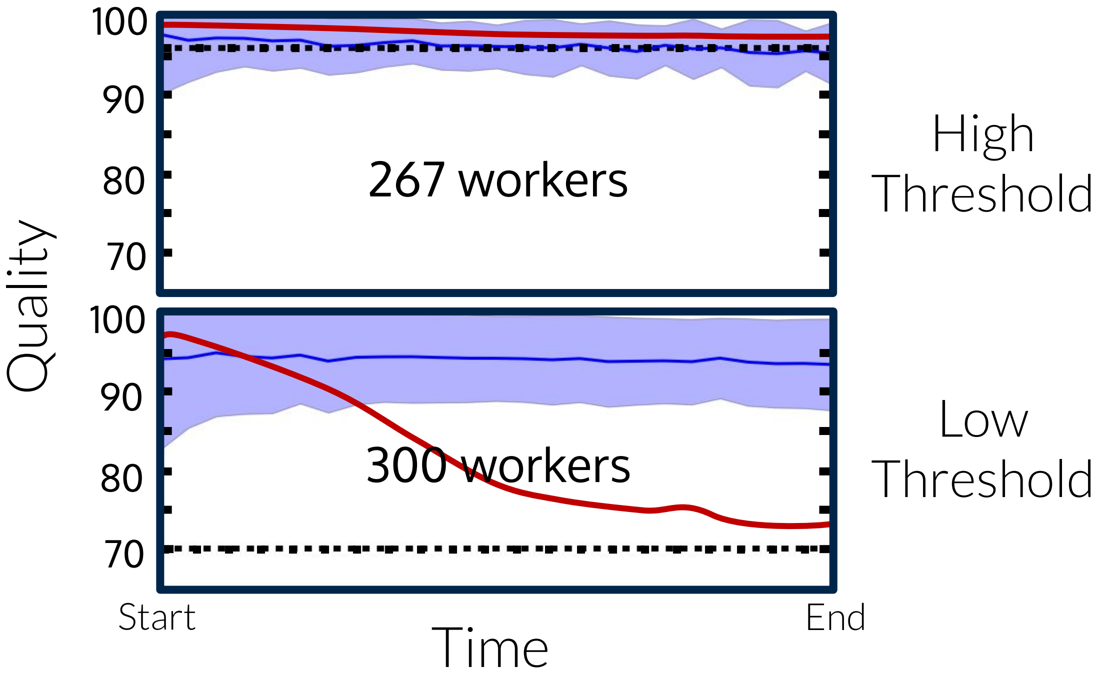
Each worked from 1 – 12 hours.

How responsive are workers to the threshold?

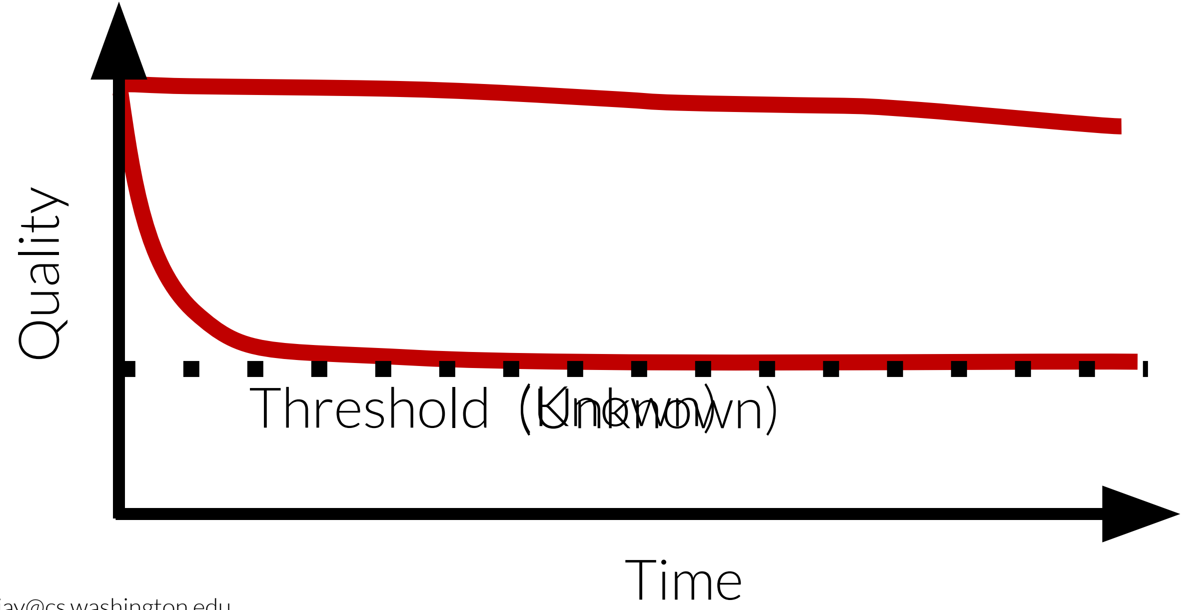


Do workers drop to the threshold?

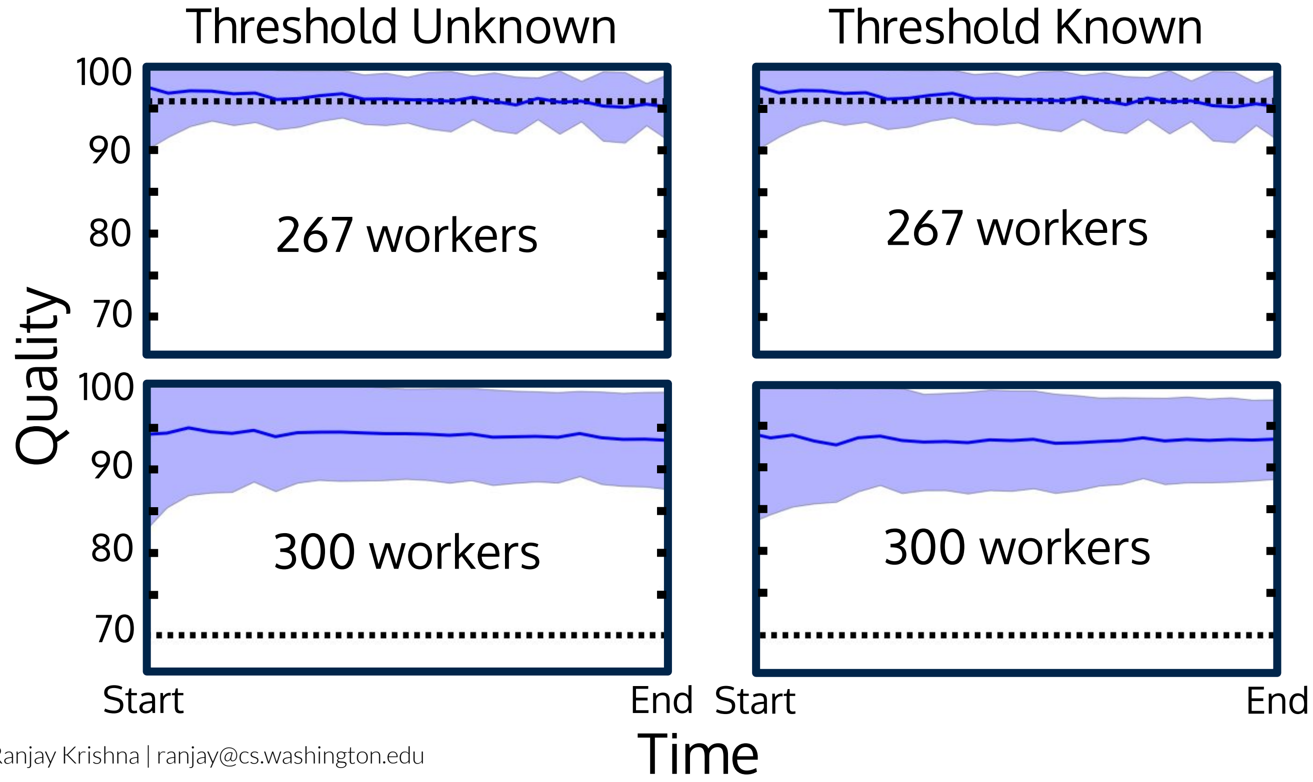
Quality remains consistent!



Does knowing their performance relative to the threshold matter?



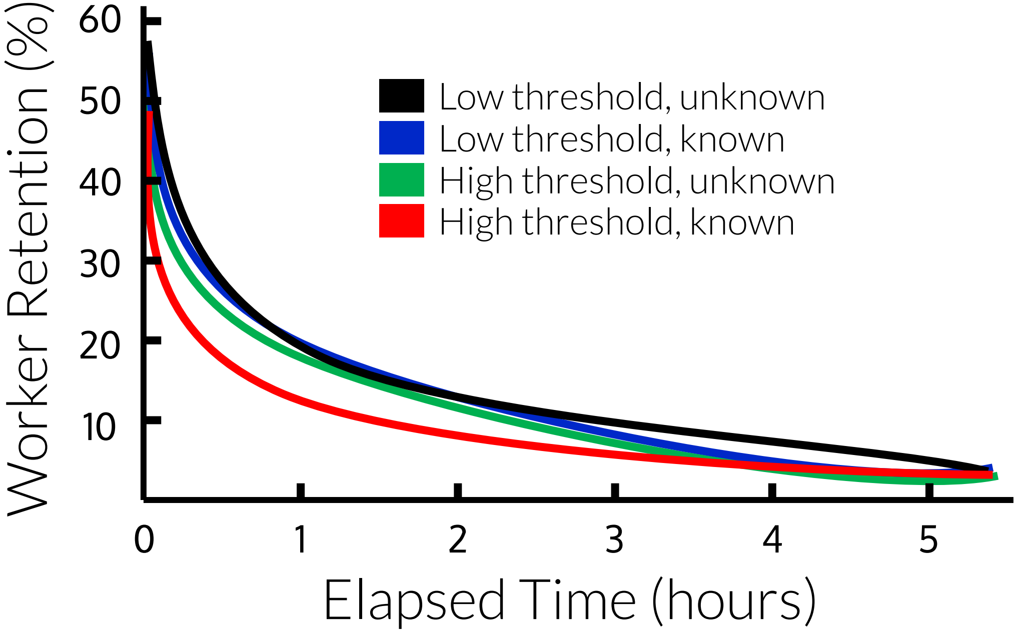
Quality remains consistent even if workers know the threshold



ANOVA

Threshold (p = 0.45)
Visibility (p = 0.13)
Interaction (p = 0.62)

Workers drop out at a higher rate when they know they are assigned to difficult tasks.



ANOVA

Threshold ($p < 0.001$)

Visibility ($p < 0.001$)

Interaction ($p < 0.001$)

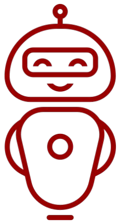
Implications and Future Work

- Retaining good workers will maintain a consistently high quality.
- Person-centric strategies may be more effective.

Limitations

- Does consistency hold in complex tasks? For non vision tasks? For effortful tasks? For tasks that involve more learning?
- What about observing workers across multiple requesters?

The humans-in-the-loop: two perspectives



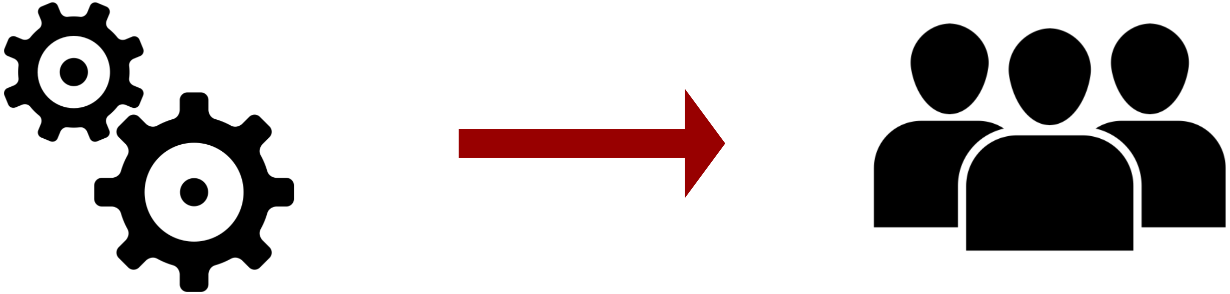
Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Workers were consistent because they were slow?



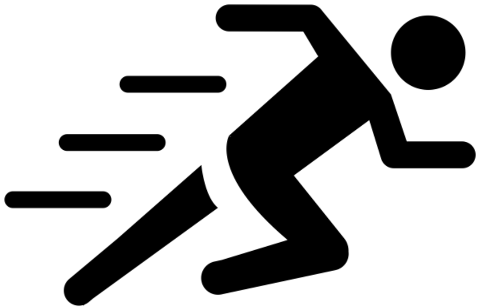
Crowdsourcing platforms
punish errors

Crowdworkers do
slow, deliberate work

Irani et al. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. CHI 2013
Martin et al. Being a Turker. CSCW 2014
Sheng et al. Get another label? improving data quality and data mining using multiple, noisy labelers. KDD 2008

Can you guess how long it takes a crowd worker to answer?

Does this contain a dog?



We want to allows workers to go faster and make errors, and even encourage it

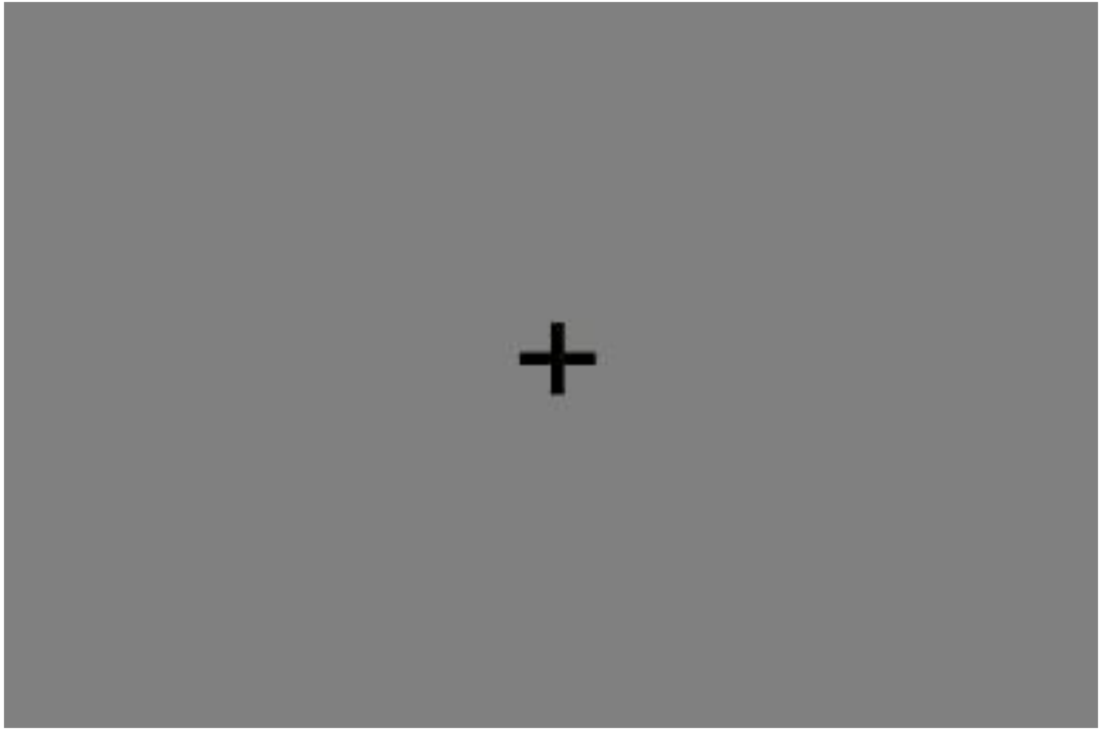
We want design a technique that is tolerant to the errors

Humans-in-the-loop from an AI perspective:

Can we speed up the annotation of vision data?

[Krishna et al. Embracing Error to Enable Rapid Crowdsourcing. CHI 2016]

Human visual processing is extremely rapid



Fei-Fei, Iyer, Koch, Perona, *J. Vision*, 2007

RSVP: Rapid Serial Visual Presentation

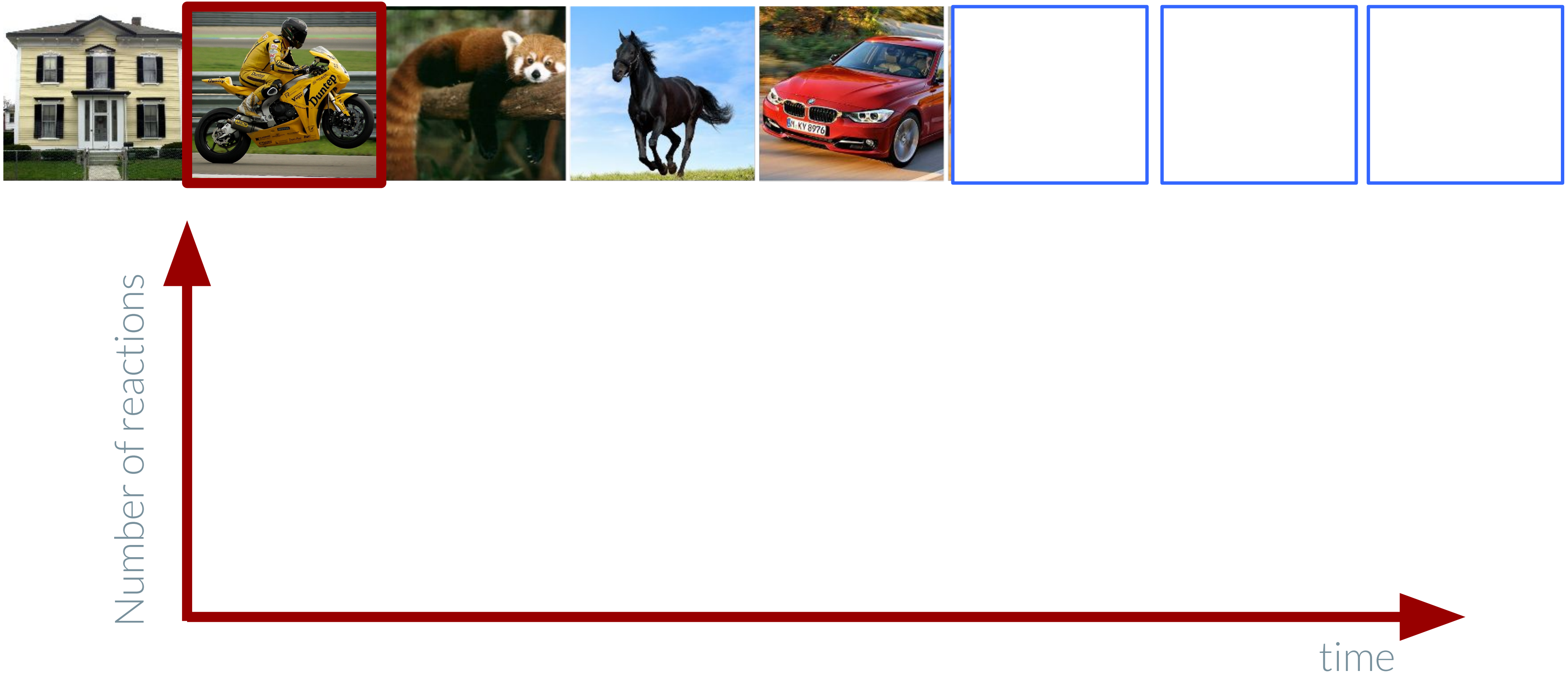
- Potter et al. 1976. Short-term conceptual memory for pictures

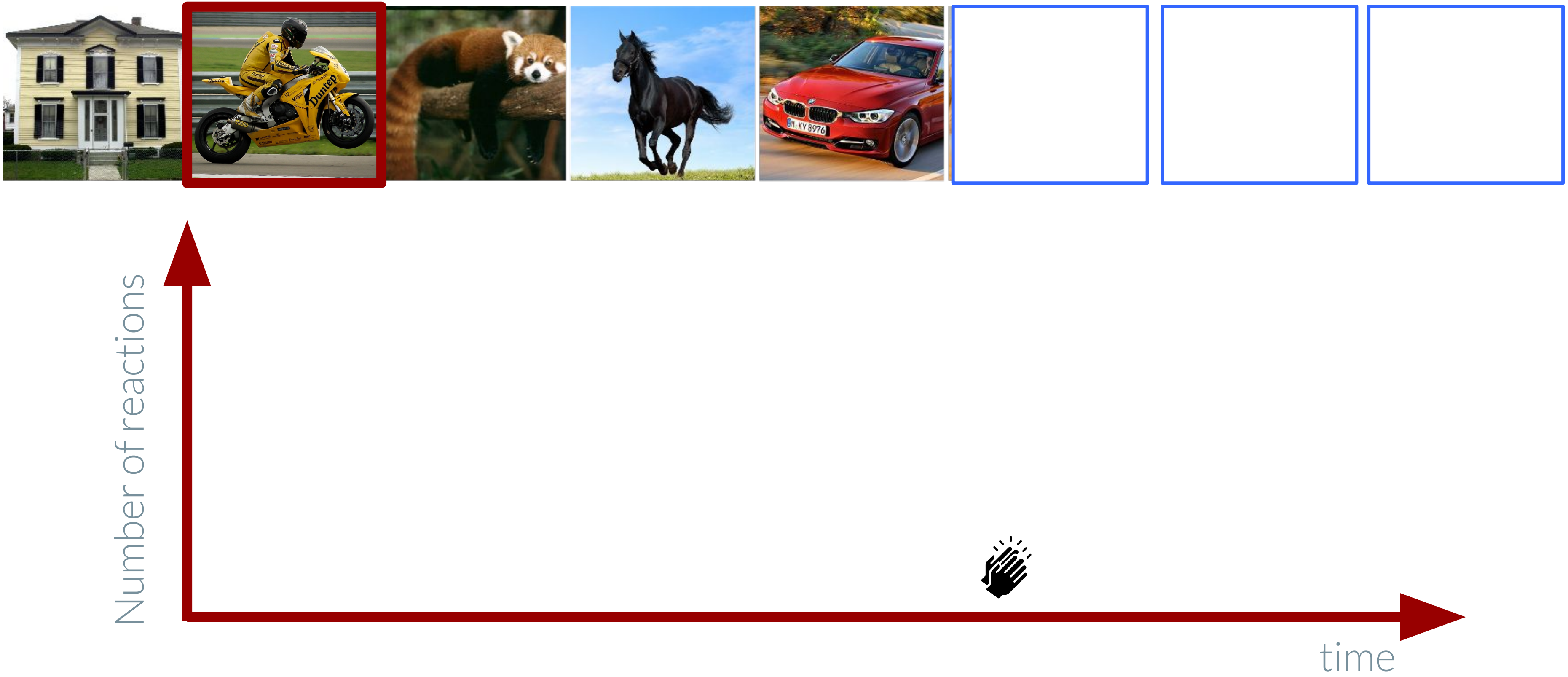
- Fei-Fei et al. What do we perceive in a glance of a real-world scene?


Duntep
Ten Kate
Veljer
DUNLOP
MOTUL
ARROW
KLASMANN
Trans-Peat
BWC
VEMAR


Damen - und
Herren - Salon

👏👏 are delayed and noisy…

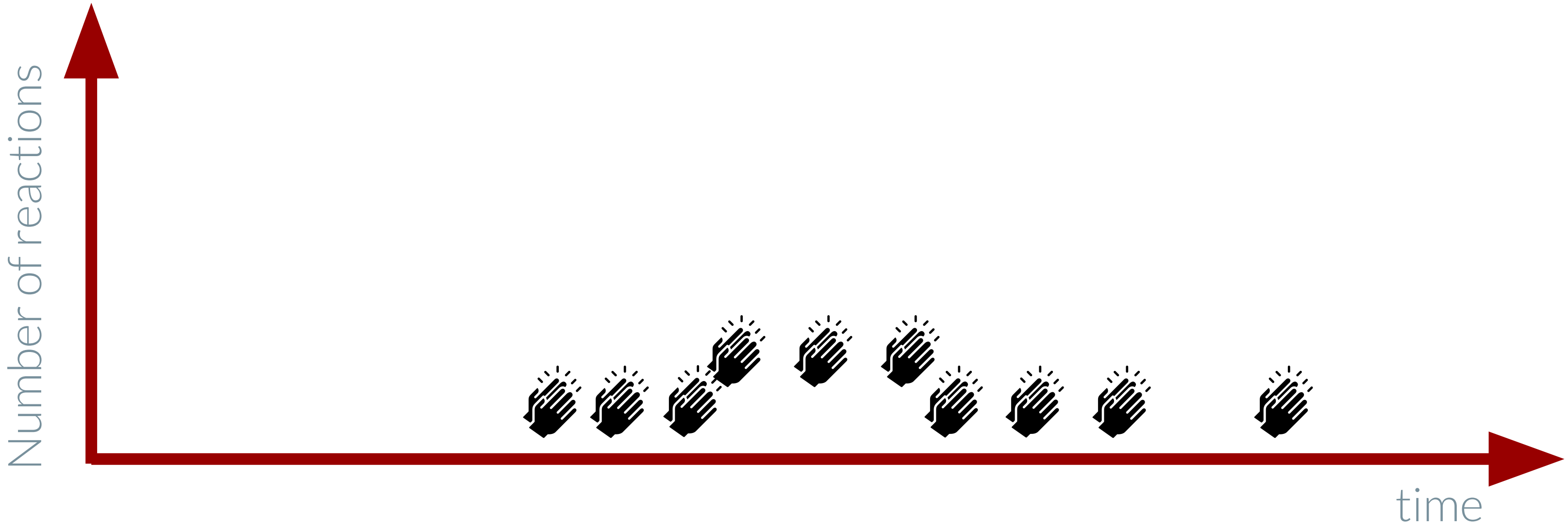

Number of reactions
time

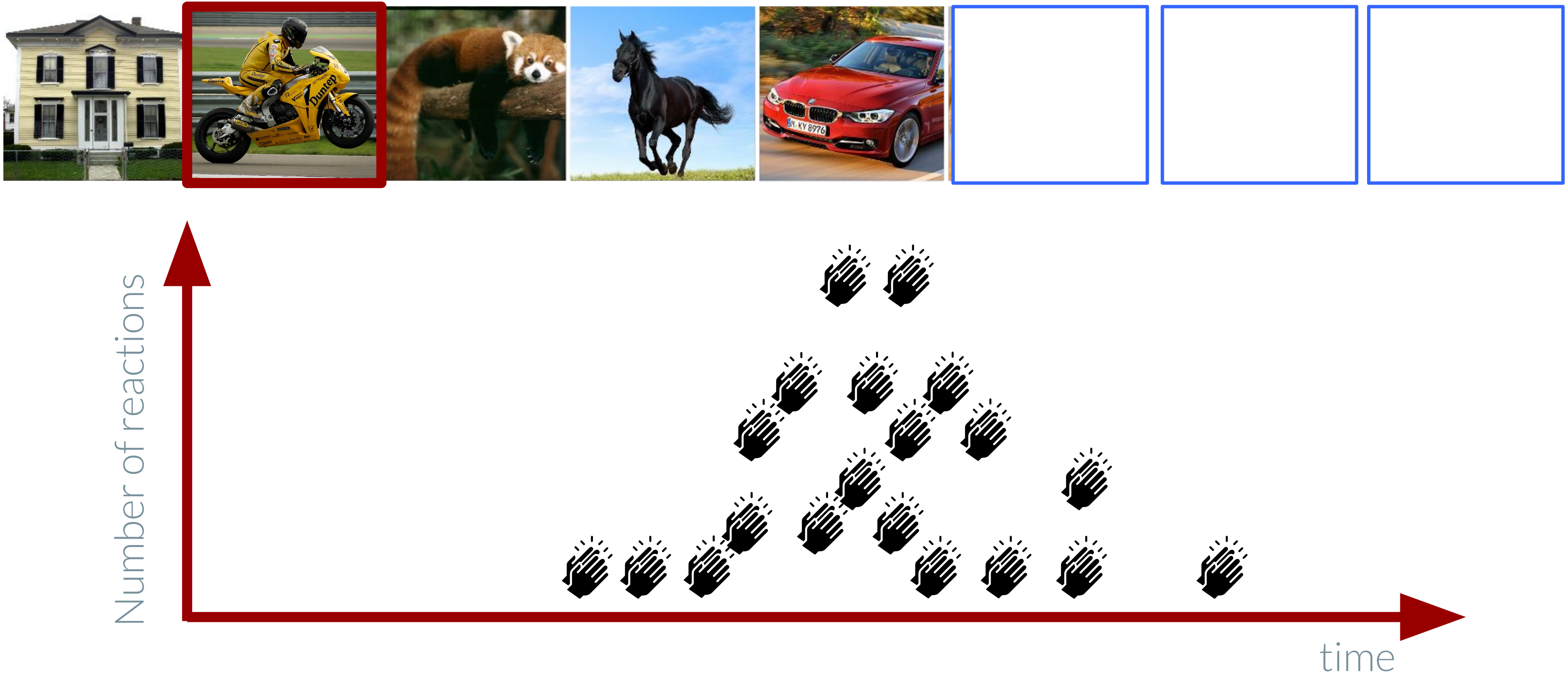


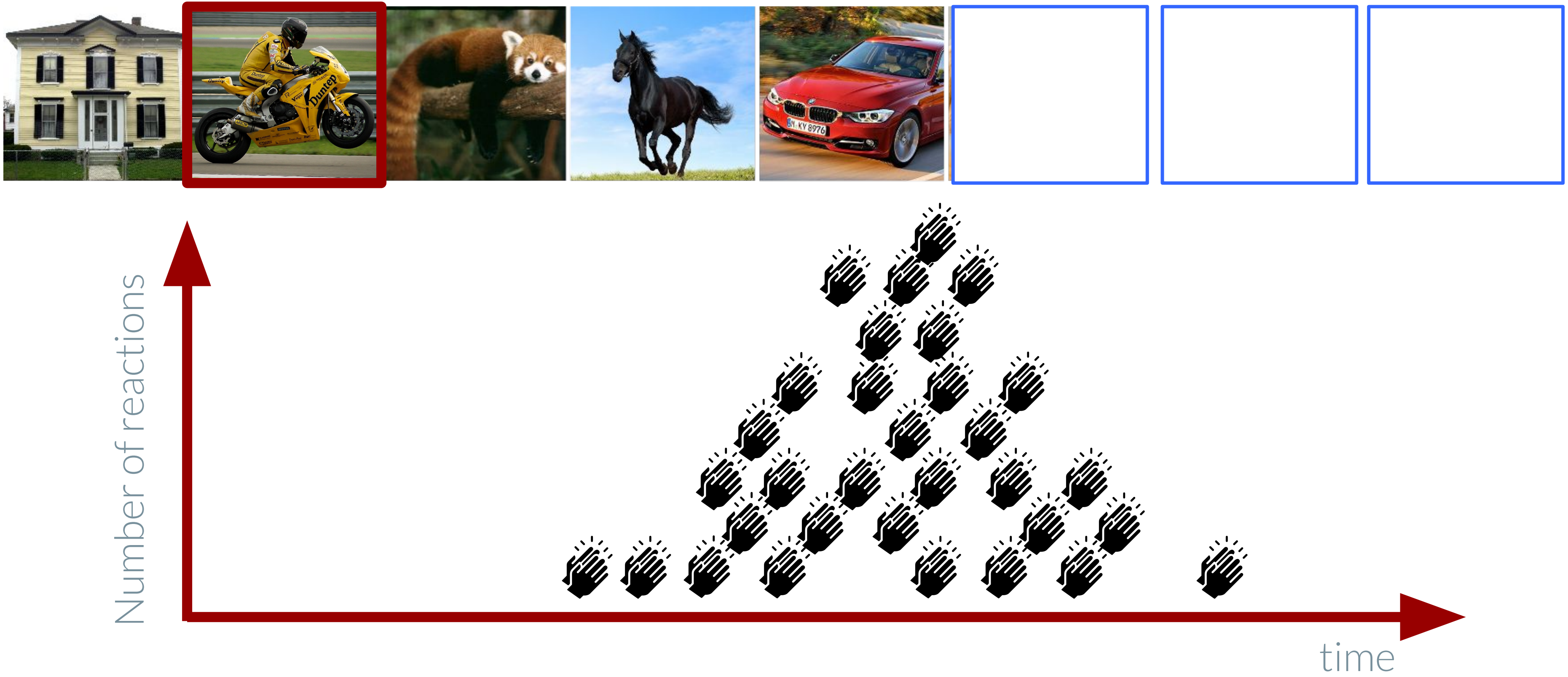
Number of reactions
time

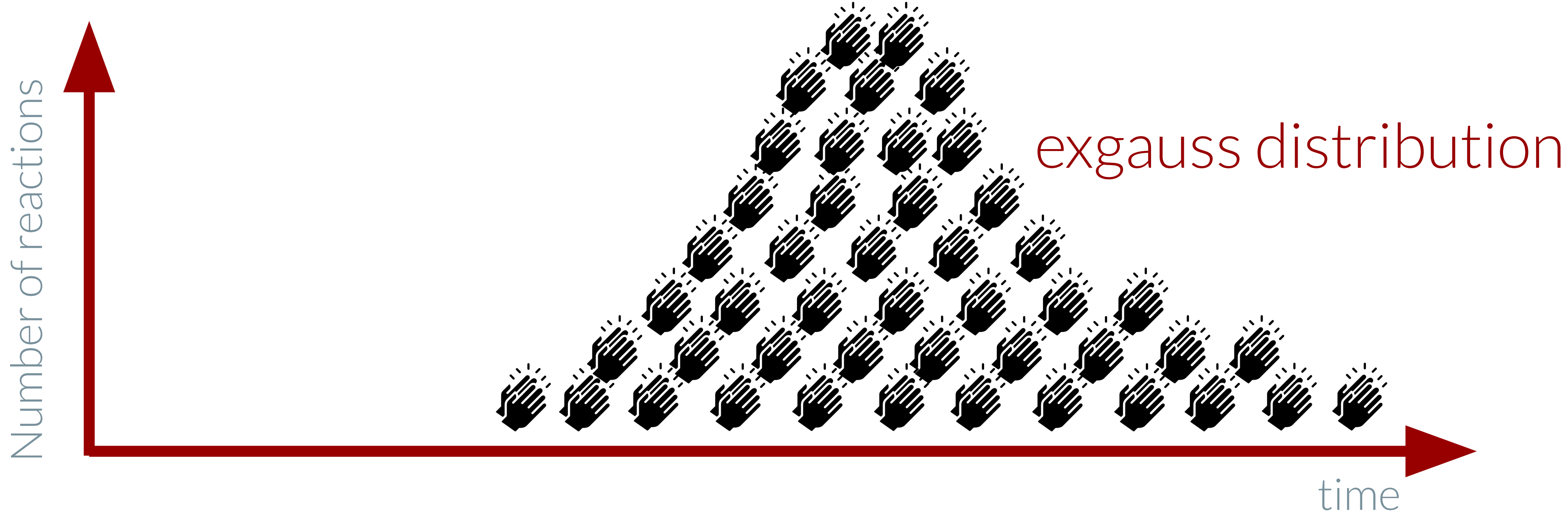

Number of reactions
time

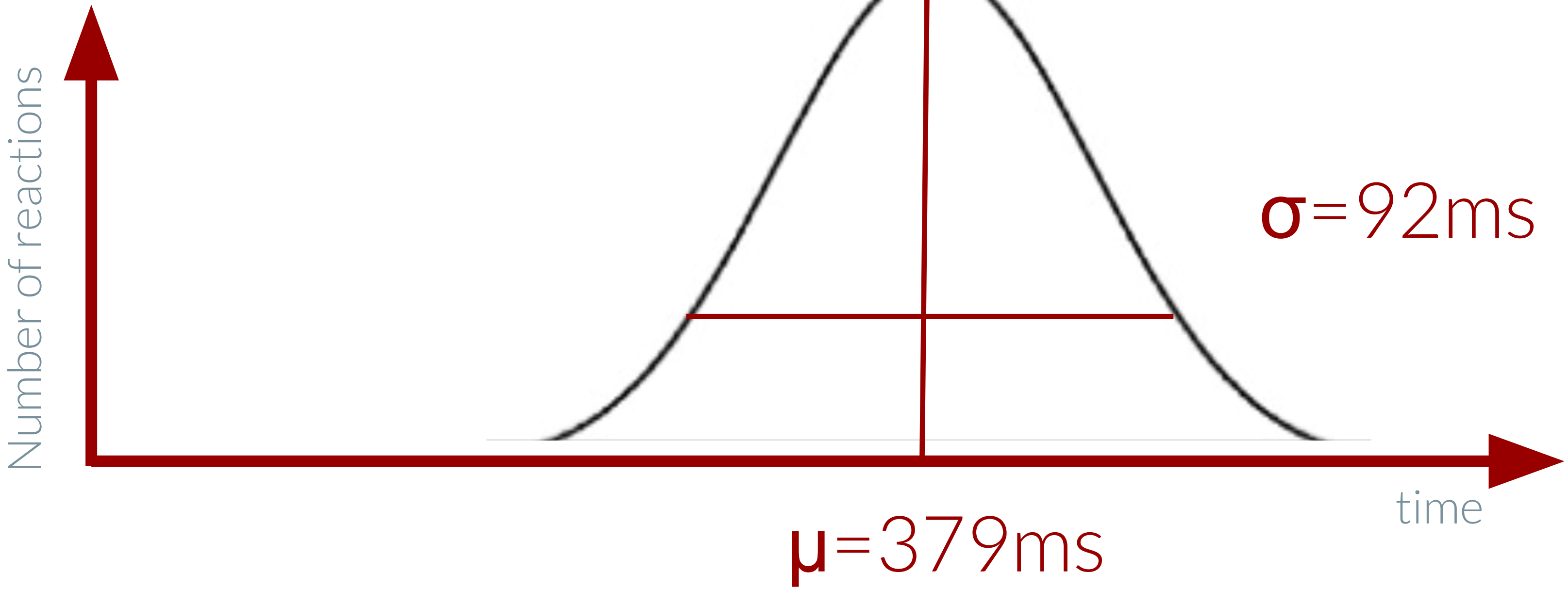

Number of reactions
time


Number of reactions
time


Number of reactions
time


Number of reactions
time


exgauss distribution
Number of reactions
time



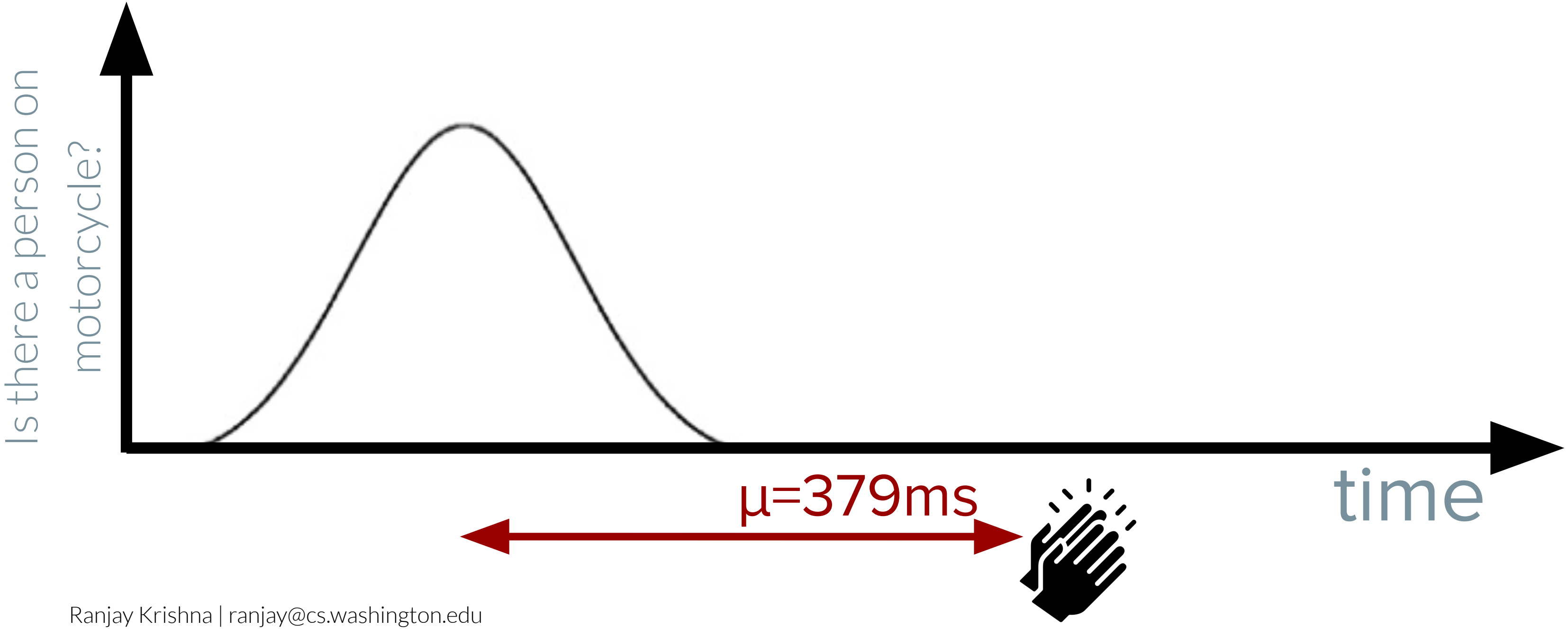
Number of reactions
σ=92ms
time
μ=379ms


Is there a person on motorcycle?
time

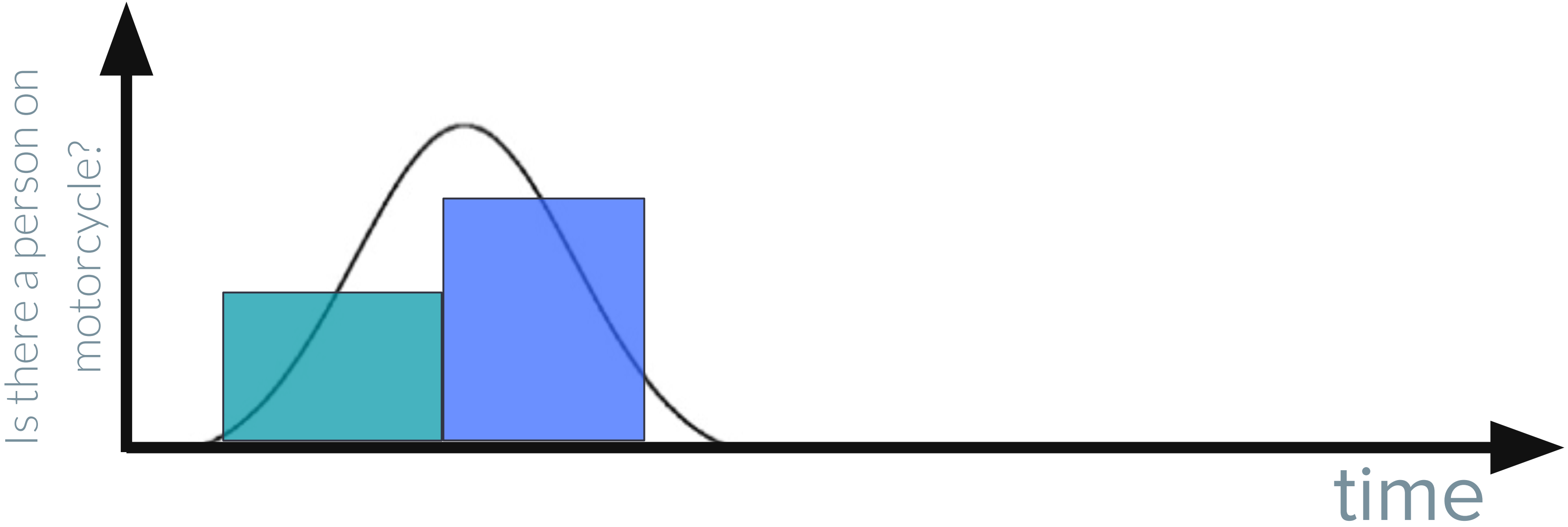
    

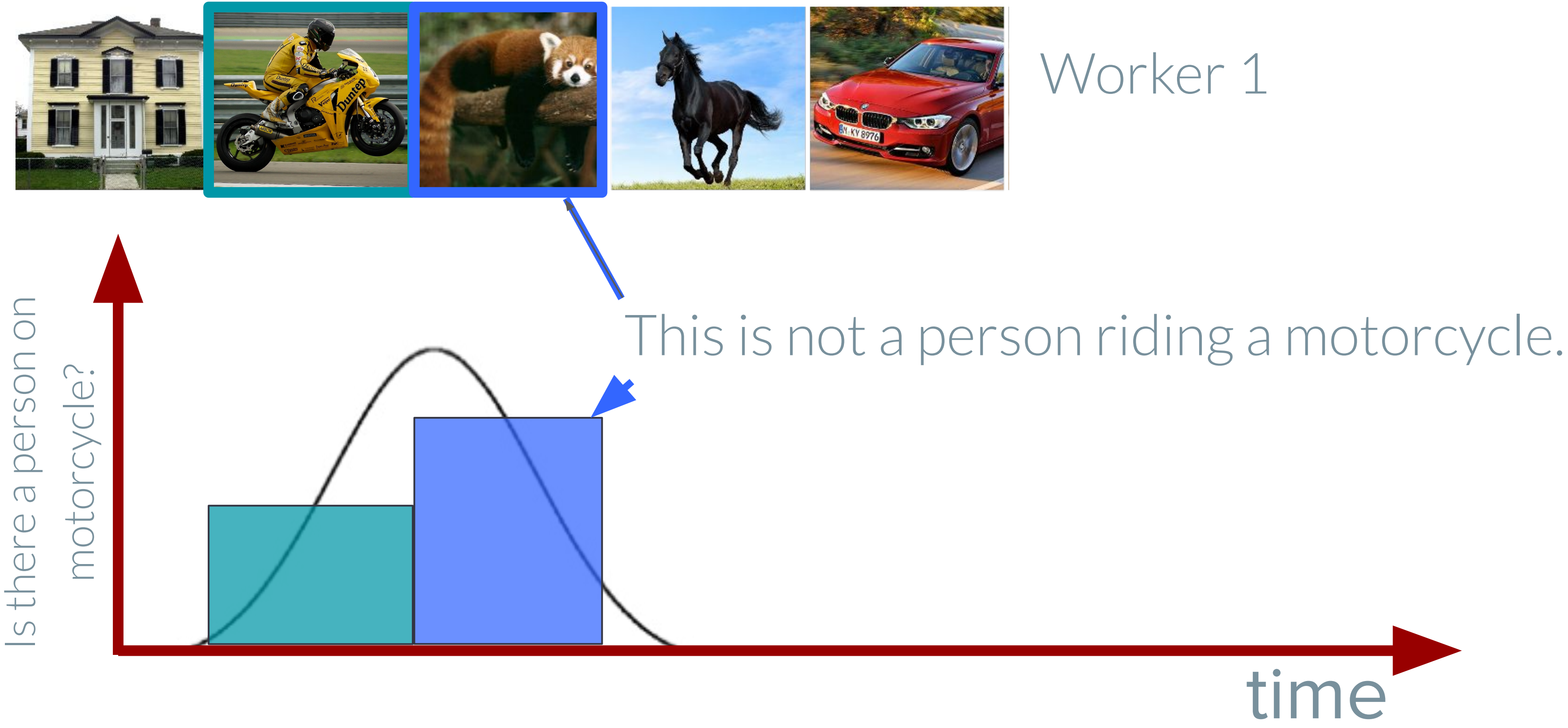
Worker 1

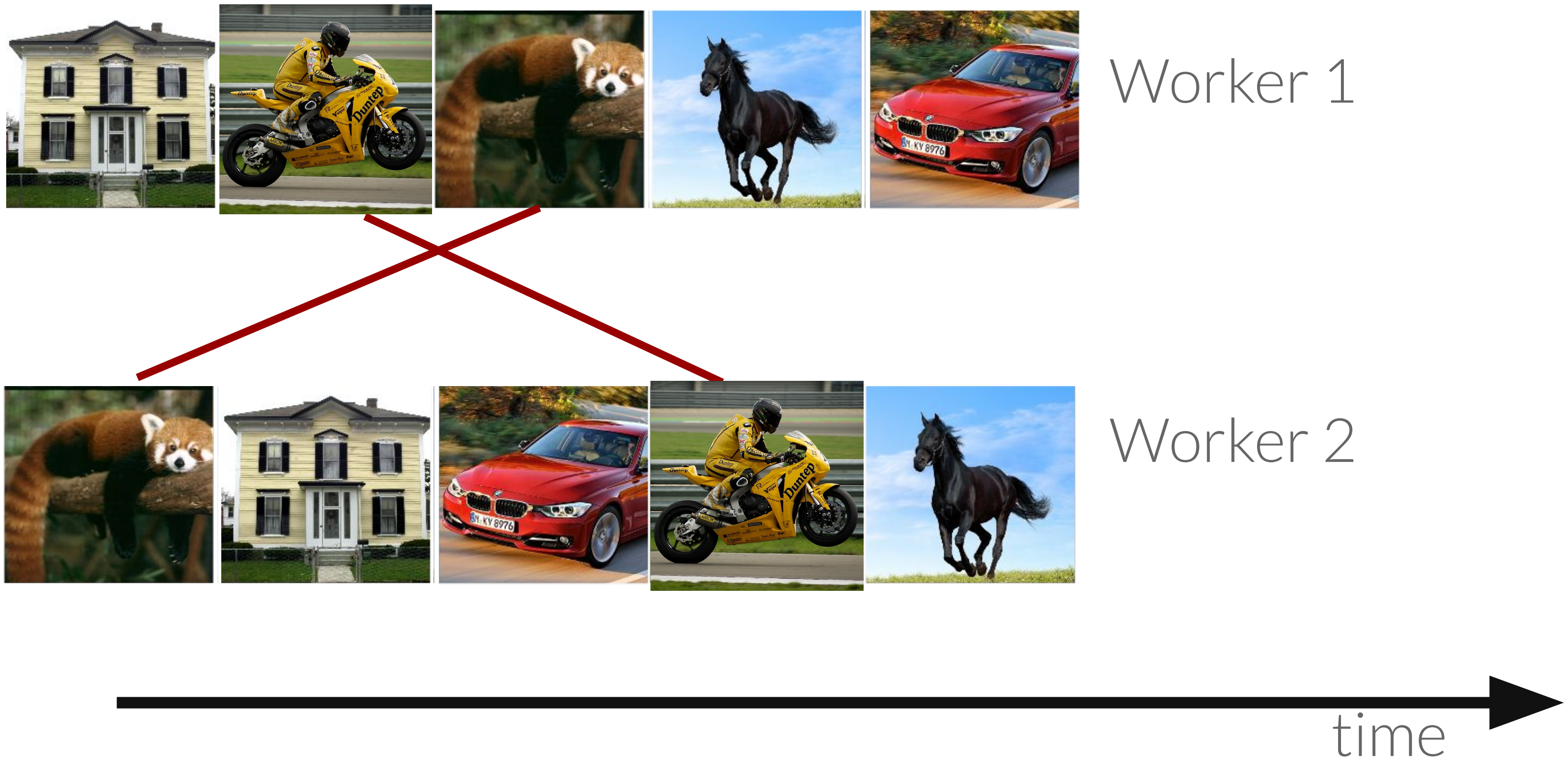
Worker 1



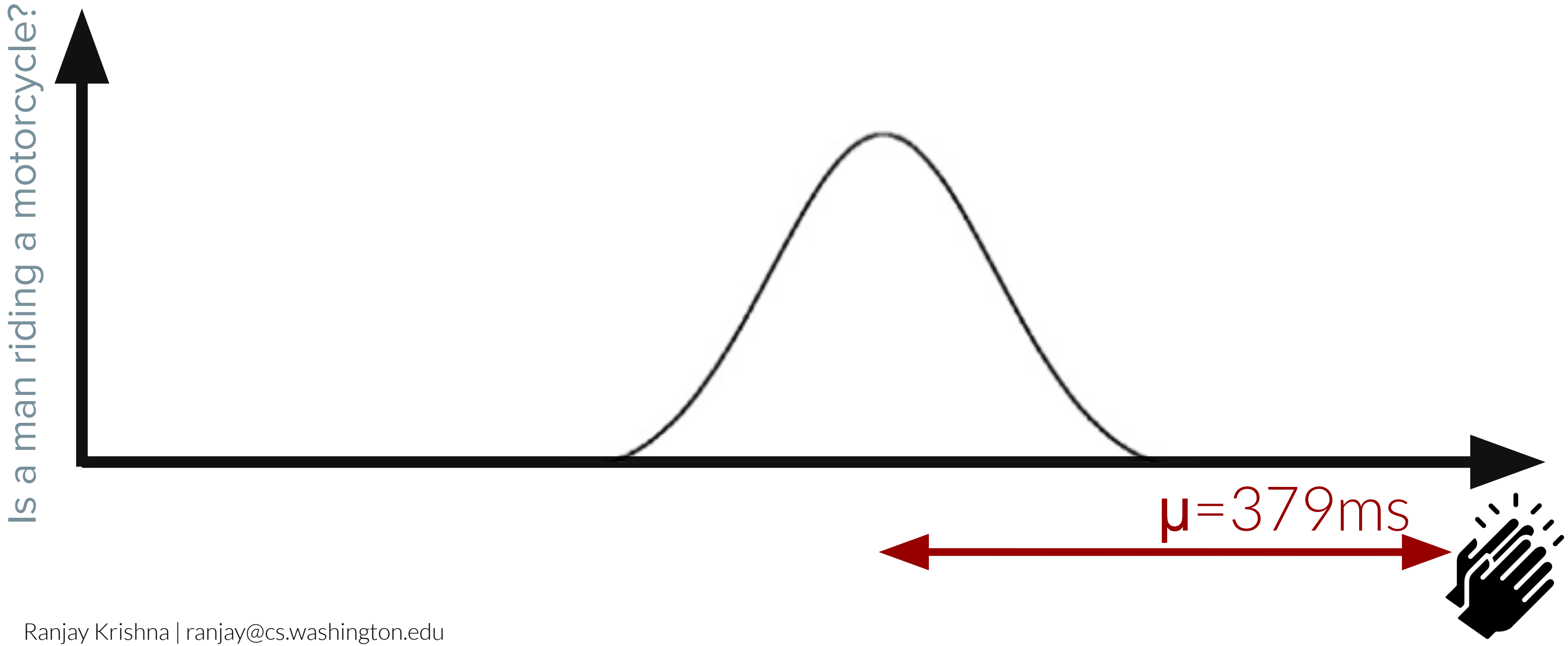
Worker 1



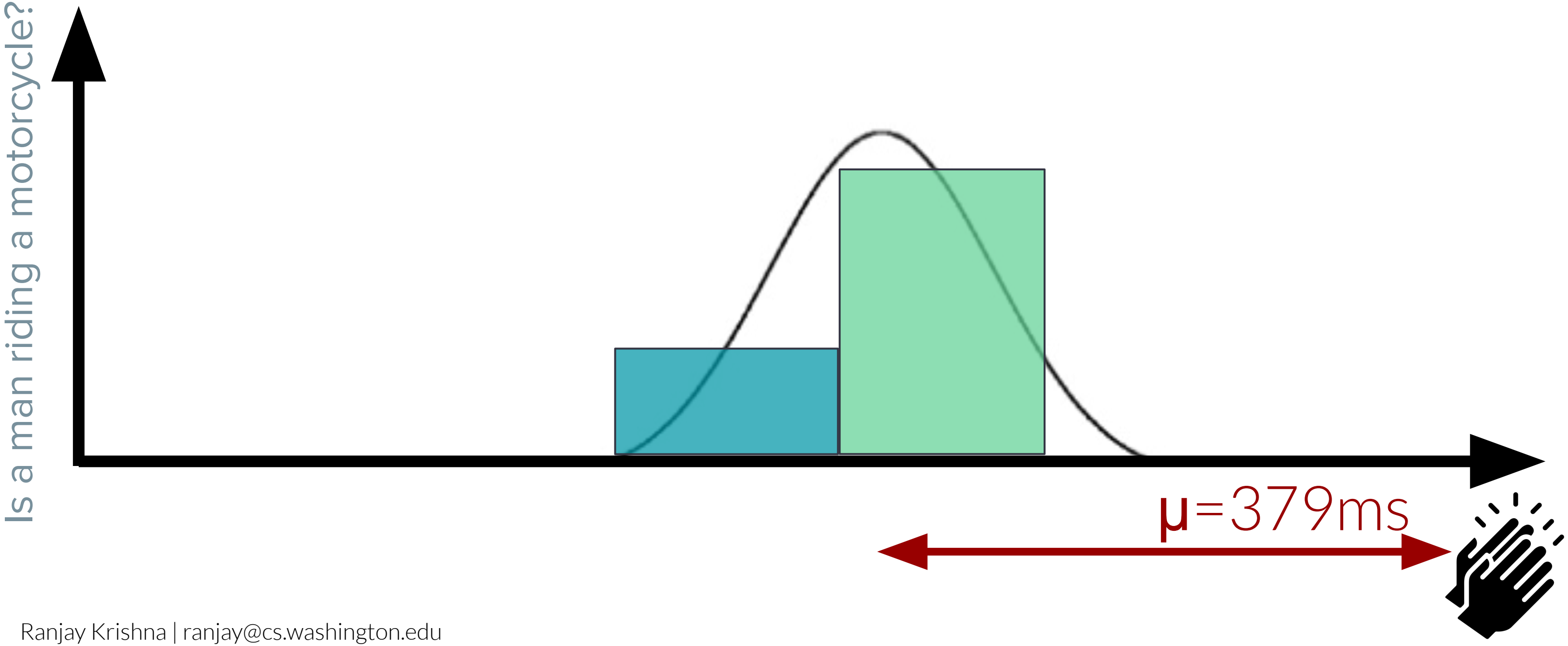

Worker 1
This is not a person riding a motorcycle.
Is there a person on motorcycle?
time


Worker 1
Worker 2
time

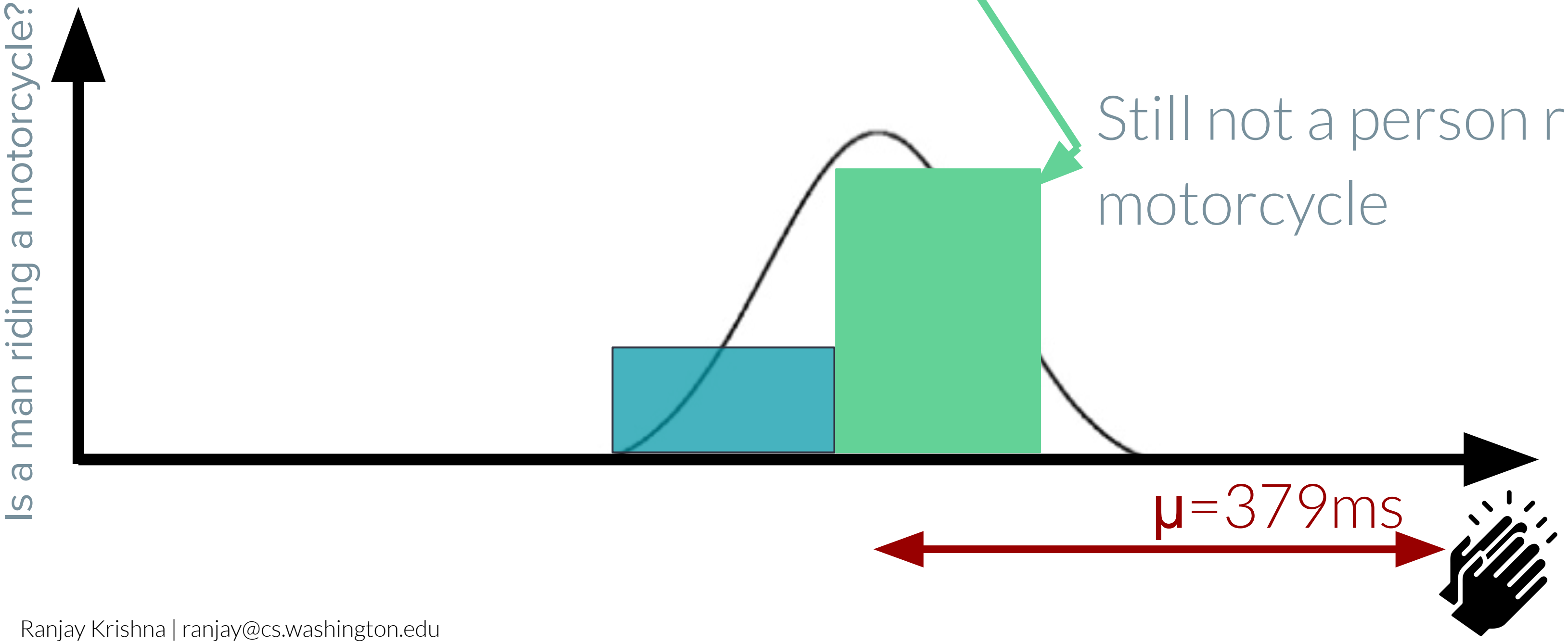
Worker 2



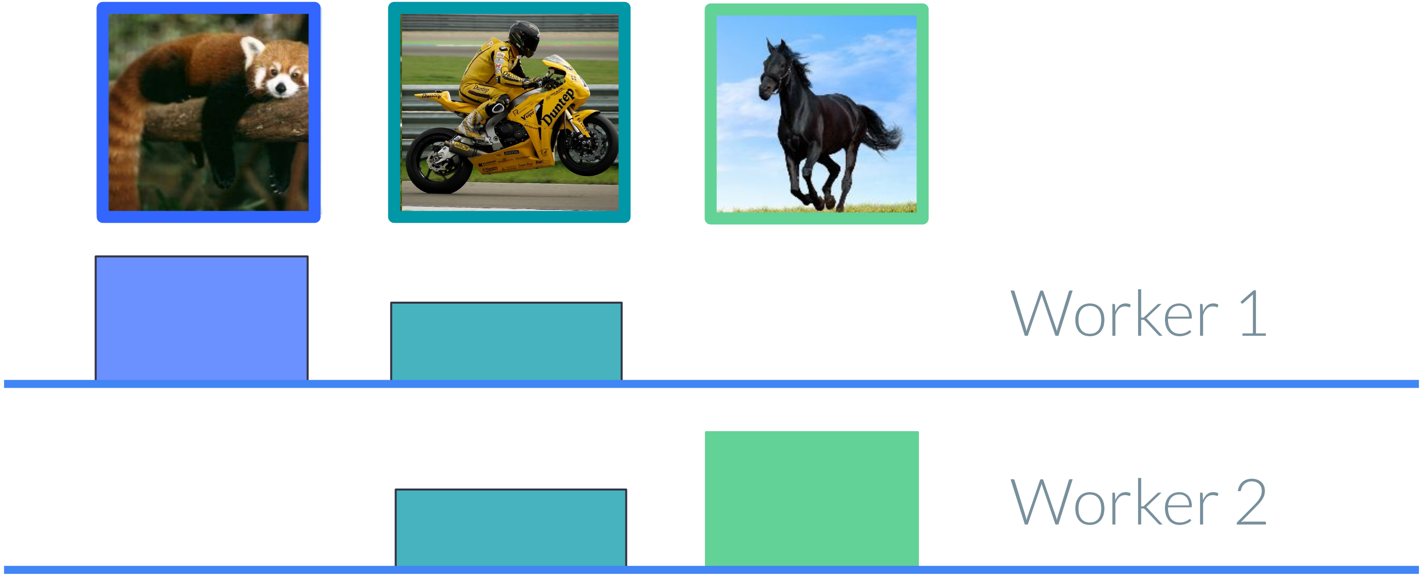
Worker 2

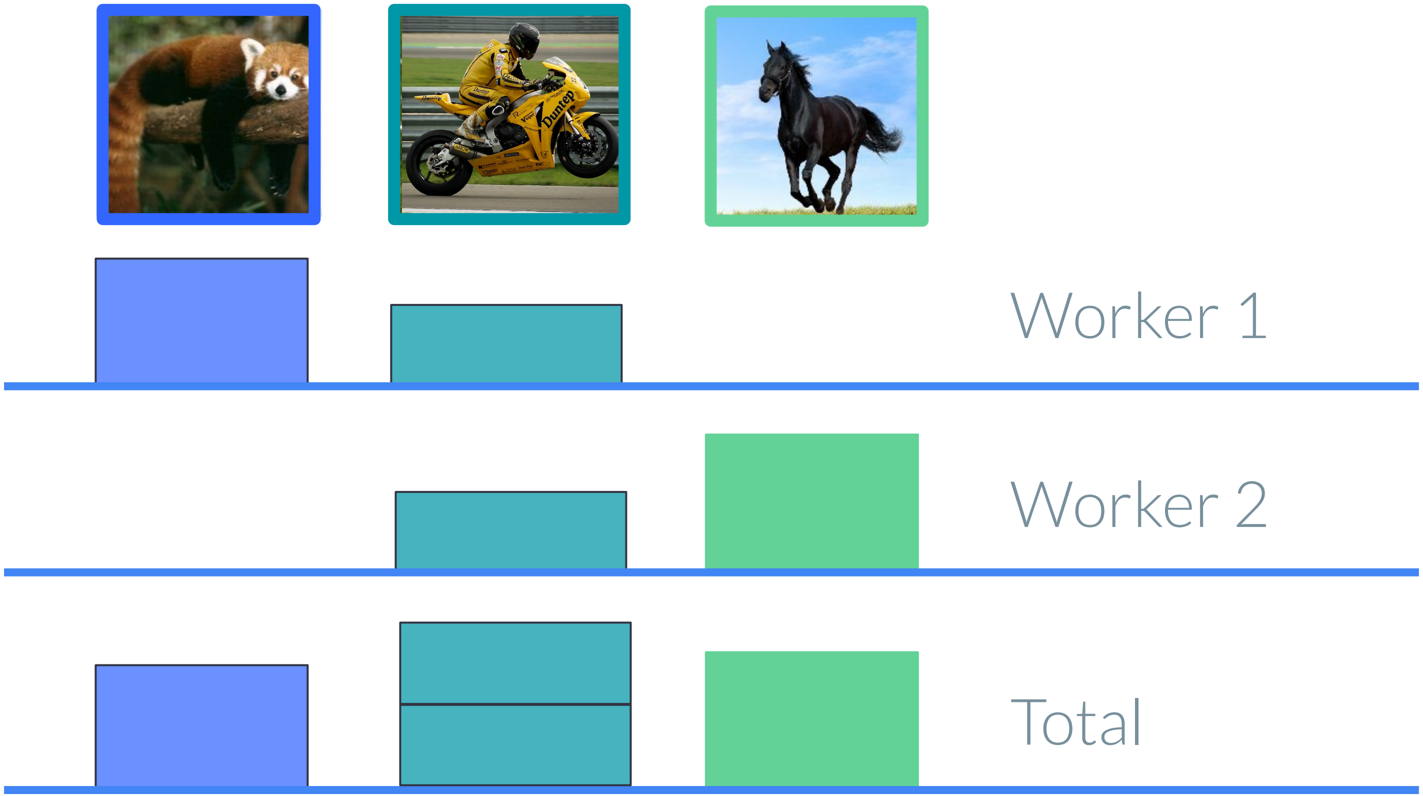


Worker 2



Still not a person riding a motorcycle


Worker 1
Worker 2


Worker 1
Worker 2
Total

By randomizing task ordering and asking multiple workers, our model is able to perform binary classification

For a set of images: $\mathcal{I} = \{I_1, \ldots, I_n\}$

Each worker gives us a set of reactions: $C^w = \{c_1^w, \ldots, c_k^w\}$

Our goal is to measure the probability of an image being positive:

$$P(I_i|C^w) = \frac{P(C^w|I_i)P(I_i)}{P(C^w)}$$

We assume that each worker reaction is independent:

$$P(C^w|I_i) = P(c_1^w, \ldots, c_k^w|I_i) = \prod_k P(C_k^w|I_i)$$

By asking multiple workers, we calculate which images are positive:

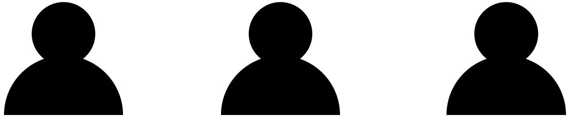
$$P(I_i) = \sum_w P(I_i|C^w)P(C^w)$$

Evaluation criteria: speedup

Control approach:
majority voting with 3 workers

Evaluation criteria: speedup

Control approach:
majority voting with 3 workers



1.7s 1.7s 1.7s

Total time per image: 5.1s

Evaluation criteria: speedup

Control approach:
majority voting with 3 workers

1.7s 1.7s 1.7s

Total time per image: 5.1s

RSVP:
at the same precision

0.1s 0.1s 0.1s 0.1s 0.1s

Total time per image: 0.5s

Evaluation criteria: speedup

Control approach:
majority voting with 3 workers

1.7s 1.7s 1.7s

Total time per image: 5.1s

RSVP:
at the same precision

0.1s 0.1s 0.1s 0.1s 0.1s

Total time per image: 0.5s

That’s a order of magnitude
speed up of > 10X

Recall suffered for long streams

RSVP worked for NLP tasks: sentiment analysis

4.25 → 0.25 seconds per tweet

Play

Natsume, you dont get it, do you? I dont want a story in Harvest Moon, I wanna farm, not spend my time looking for Sunstones and things.

RSVP worked for NLP tasks: word similarity

6.23 → 0.60 seconds per word

Find synonyms for wide

broad	hushing
	crunch
	short

Play

2

RSVP worked for NLP tasks: topic detection

14.33 → 2.00 seconds per article

Find articles related to “housing”

Sales of previously owned homes dropped 14.5% in January to a seasonally adjusted annual rate of 3.47 min units, the national association of realtors ….

Limitations: fine grained detection

Sayornis

Gray Kingbird

Limitations: Influence of typicality

Typicality score: 0.9

Typicality score: 0.1

Iordan et al. Basic level category structure emerges gradually across human ventral visual cortex. 2011

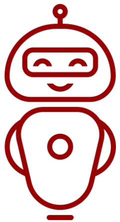
Implications and Future Work

- Allowing Embrace errors can speed them up if algorithms can recover the errors
- RSVP can speed up vision and NLP tasks.

Limitations

- There is a tradeoff between recall and speed
- It doesn’t work for fine grained differences

The humans-in-the-loop: two perspectives



Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Human-Computer Interaction

Goal: To support a labor force achieve their financial and career goals

Artifact: automations that structure work

Impacts across long time horizon

Ranjay Krishna | ranjay@cs.washington.edu

The humans-in-the-loop
from an HCI perspective

Ranjay Krishna | ranjay@cs.washington.edu

The humans-in-the-loop: two perspectives

Human-Computer Interaction

Goal: To support a labor force achieve their financial and career goals

Artifact: automations that structure work

Impacts across long time horizon

Ranjay Krishna | ranjay@cs.washington.edu

A new online economy of labelers to support machine learning

Paradox of automation’s last mile

“As ML techniques automate some work, they create new types of work that depend on human expertise.”

- Mary Gray. Ghost Work, 2019

Gig work necessary to support AI infrastructures

It leads to Ghost Work conditions that devalue the humans-in-the-loop

It's not going away

Dismantling of full-time employment for on-demand work

amazon mechanical turk beta

figure eight

V7

aws

V7 Darwin

Automated Image Annotation

shaip

Better AI Data | Better Results

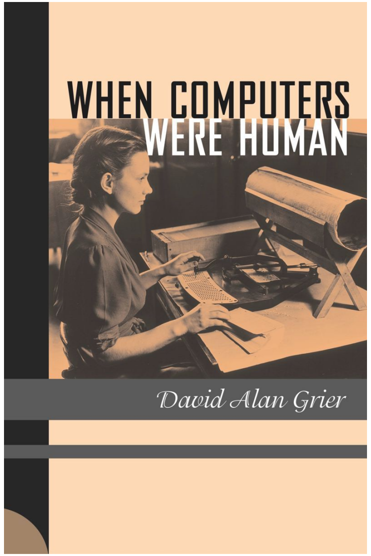
HIVE

Looking back at ghost work through the lens of piece work

The idea that complex tasks can be broken down into simpler tasks for individuals

Roots in intellectual work in the 18th century

- Astronomers hired teenage men to calculate equations



Alkhatib et al. Examining Crowd Work and Gig Work Through The Historical Lens of Piecework. CHI 2017

Industrial revolution adopted piecework- Cars in 93 mins

Job Characteristic Model

Hackman & Oldham, 1980

Core Job Characteristics → Critical Psychological States → Outcomes

Core Job Characteristics	Critical Psychological States	Outcomes
Skill variety Skill identity Skill significance	Experience meaningfulness of the work	High internal work motivation High "growth" satisfaction High general job satisfaction High work effectiveness
Autonomy	Experience responsibility of the outcomes of the work	
Feedback	Knowledge of the actual results of the work	

Existing platforms do not support these job characteristics

Requester	Title	HITs	Reward	Created	Actions	
James Billings	Market Research Survey	25,571	$0.05	9m ago	Preview	Accept & Work
Research Rewards	Quick Market Research Survey	22,826	$0.02	6m ago	Preview	Accept & Work
Mayanksoniphd	Generate praise, given a persona.	6,655	$0.03	15d ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	1,150	$0.01	11s ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	1,121	$0.02	4h ago	Preview	Qualify
minsVA	Draw a polygon around the tailgate of the requested cars	915	$0.10	4h ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	811	$0.03	3h ago	Preview	Qualify
VacationRentalAPI CA	Address Identification - 10207 - Kelowna, BC	676	$7.50	5h ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	628	$0.05	16h ago	Preview	Qualify
minsVA	Draw a polygon around the front hood of the requested cars	616	$0.10	4h ago	Preview	Qualify
Shopping Receipts	Extract General Data & Items From Shopping Receipt	554	$0.04	12h ago	Preview	Qualify
VacationRentalAPI	Address Identification - 10227 - Minneapolis, MN	405	$2.50	5h ago	Preview	Qualify
VacationRentalAPI	Address Identification - 10243 - New Listing Mix	371	$2.00	3h ago	Preview	Qualify
str11223344	Tell us what this item is - General Contents - Batch ID #44814	353	$0.08	6d ago	Preview	Qualify
VacationRentalAPI	Address Identification - 10242 - New Listing Mix	353	$2.00	4h ago	Preview	Qualify
Alexander Gutin	Run a query in ChatGPT	326	$0.02	11d ago	Preview	Qualify
VacationRentalAPI CA	Address Identification - 10200 - Brampton, ON	321	$7.50	5h ago	Preview	Qualify
Company	Company Logos	297	$0.01	17s ago	Preview	Accept & Work
Shopping Receipts	Extract Data From Shopping Receipt	294	$0.01	1m ago	Preview	Qualify
VacationRentalAPI CA	Address Identification - 10201 - Burnaby, BC	258	$7.50	5h ago	Preview	Qualify

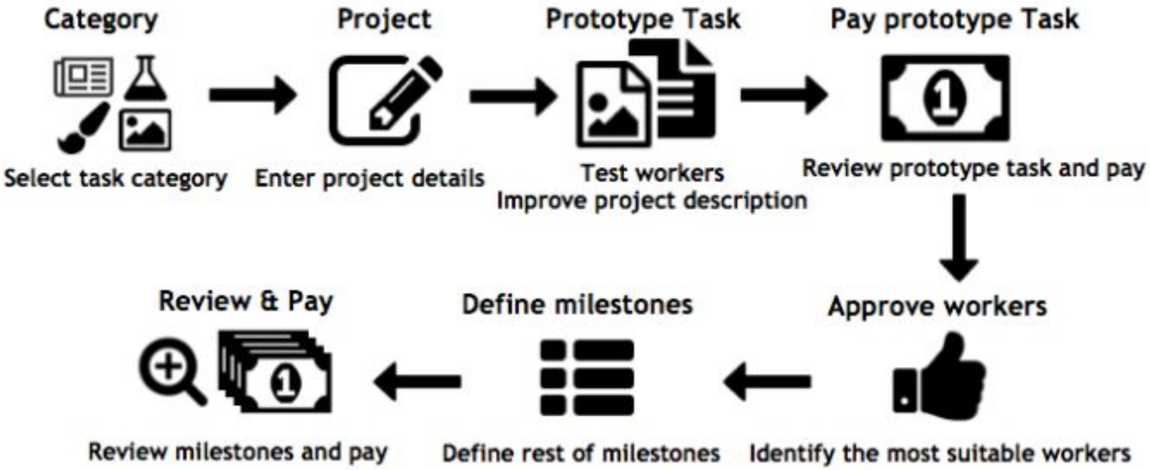
Humans-in-the-loop from an HCI perspective:
Can we develop a platform that supports worker needs?

Daemo: a Self-Governed Crowdsourcing Marketplace

V1:

Launched with prototype tasks

-



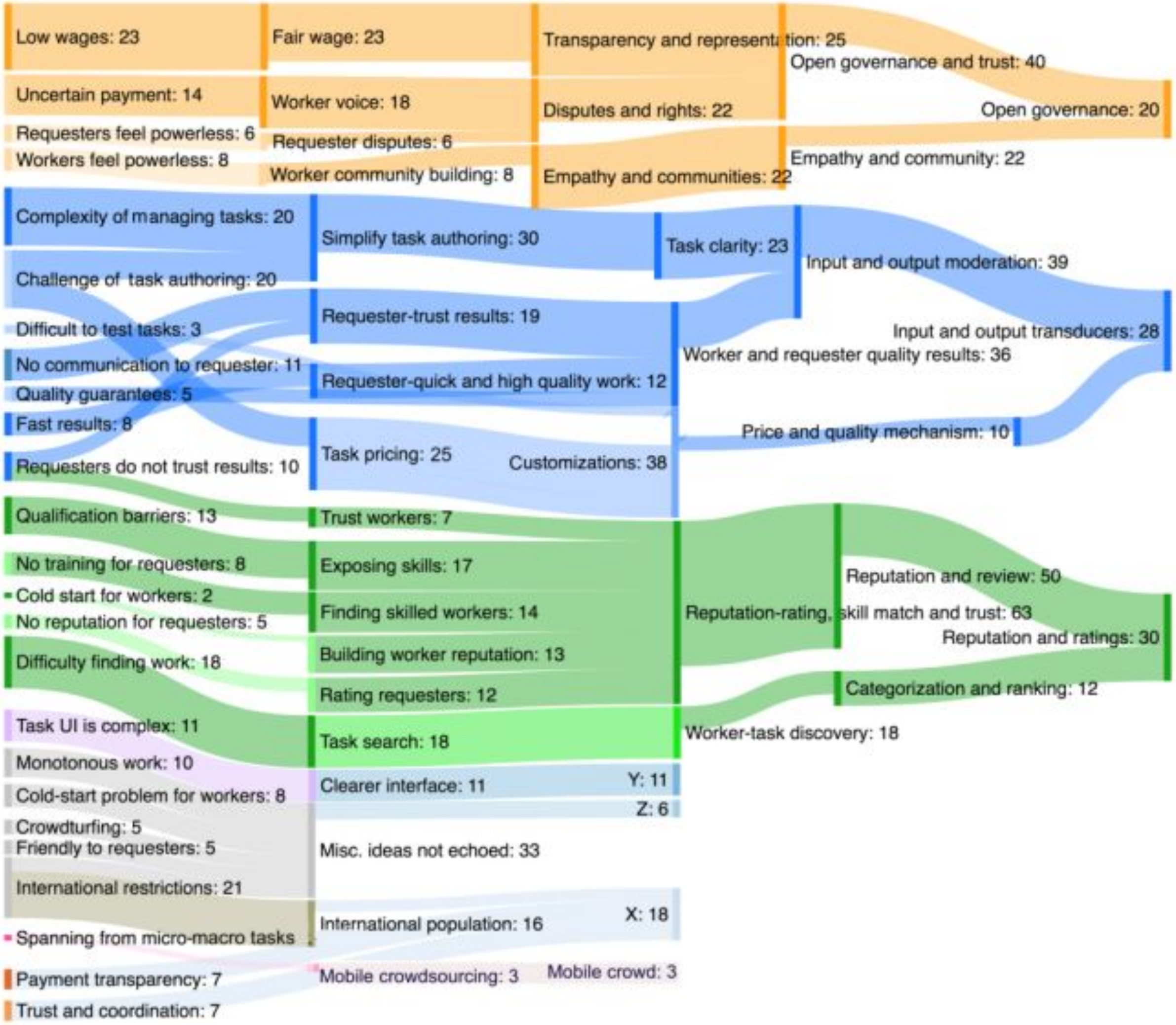
Open governance

- 3 workers
- 3 requesters
- 1 researcher

Gaikwad et al. Daemo: a Self-Governed Crowdsourcing Marketplace. UIST 2017

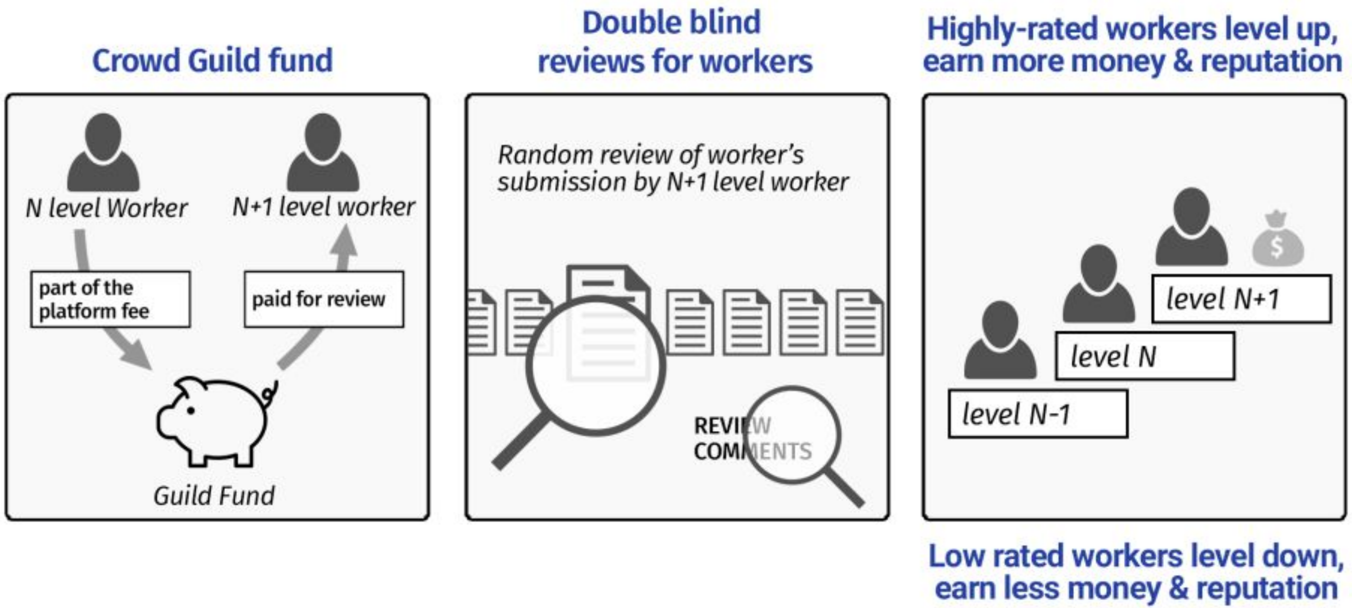
Ideas

Changes to the platform were ideated on transparently and collectively prioritized



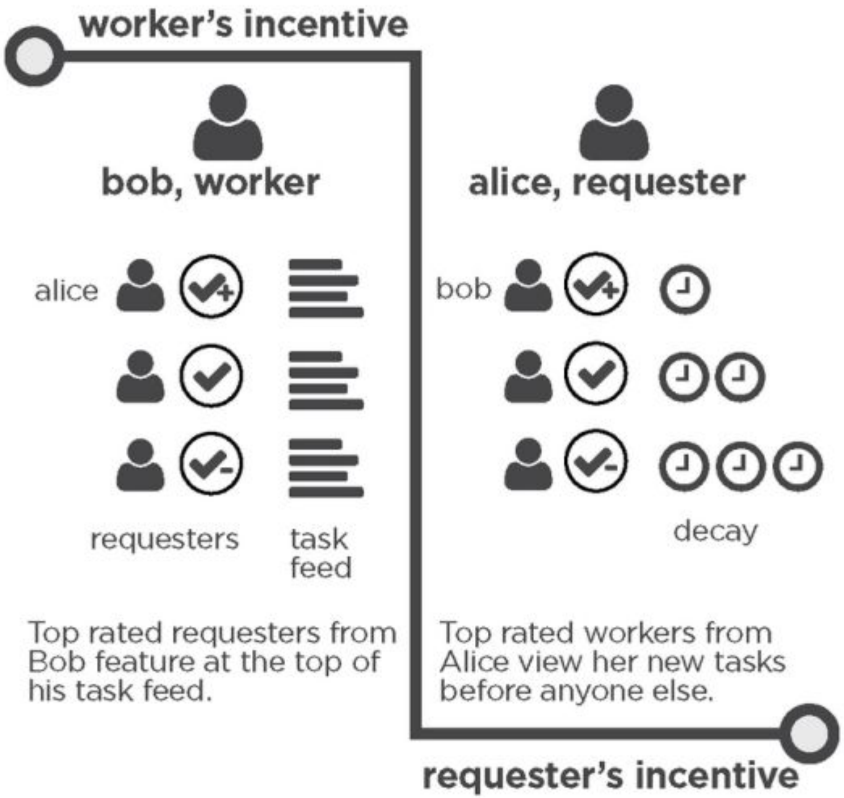
Ranjay Krishna | ranjay@cs.washington.edu

A reputation protocol: workers received feedback



Whiting et al. Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms. CSCW 2017

A rating system: To trade off skill variety of identity



Gaikwad et al. Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms. UIST 2016

Building a new decentralized crowdsourcing system with a crowd of researchers

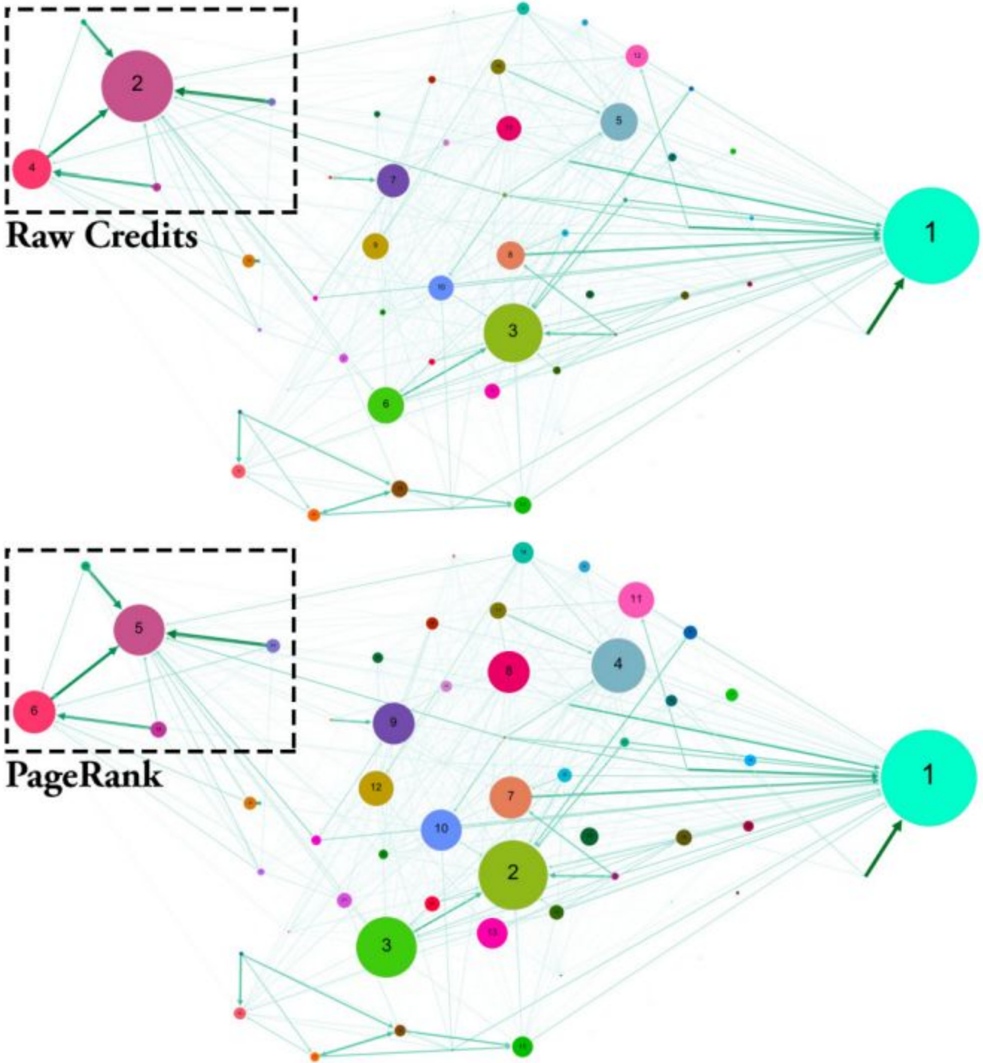
Achieve upward educational mobility while creating research systems and co-authoring papers

Vaish et al. Crowd Research: Open and Scalable University Laboratories. UIST 2017

Author order determined using crowdsourced points and page rank

Potential challenges:

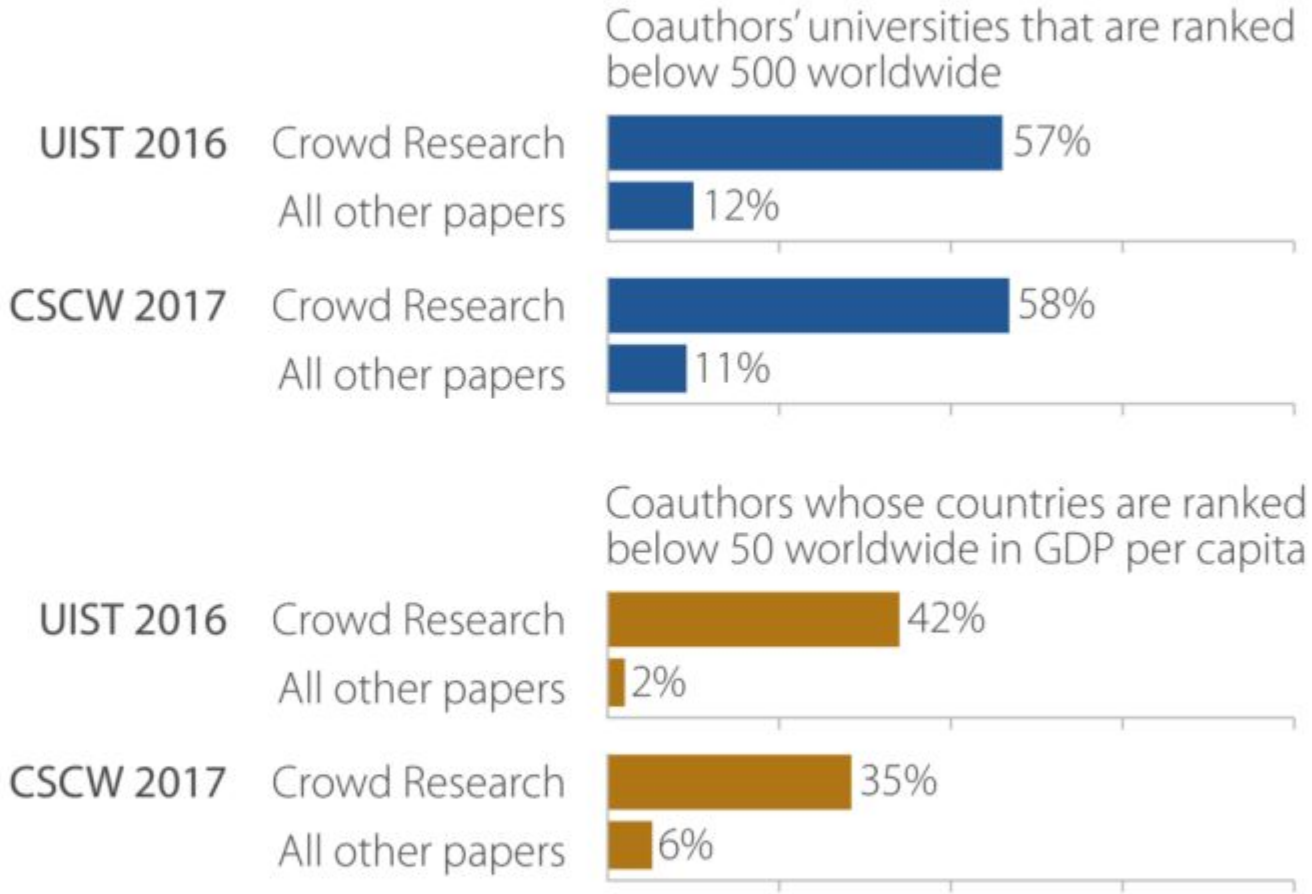
- Link ring
- Quid-proquo strategy



Ranjay Krishna | ranjay@cs.washington.edu

Supporting upward mobility

Our authors were more diverse than those from other papers at the same venue



Job Characteristic Model

Hackman & Oldham, 1980

Core Job Characteristics → Critical Psychological States → Outcomes

Skill variety
Skill identity
Skill significance

Experience meaningfulness of the work

Autonomy

Experience responsibility of the outcomes of the work

Feedback

Knowledge of the actual results of the work

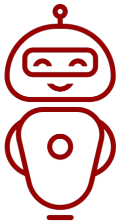
High internal work motivation

High "growth" satisfaction

High general job satisfaction

High work effectiveness

The humans-in-the-loop: two perspectives



Artificial Intelligence

Goal: To produce high quality labels as efficiently as possible

Artifact: training data for models

Impacts across short time horizon

Human-Computer Interaction

Goal: To support a labor force achieve their financial and career goals

Artifact: automations that structure work

Impacts across long time horizon

Future lectures will look at other humans-in-the-loop: the users