IUI 2018

# Two Tools are Better Than One: Tool Diversity as a Means of Improving Aggregate Crowd Performance

Jean Y. Song, Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, Walter S. Lasecki

# Crowdsourcing Platforms

clickworker

microtask

upwork™

spare 5

amazon mechanical turk™
Artificial Artificial Intelligence

Wikipedia
The Free Encyclopedia

Mighty Ai
Training Data as a Service

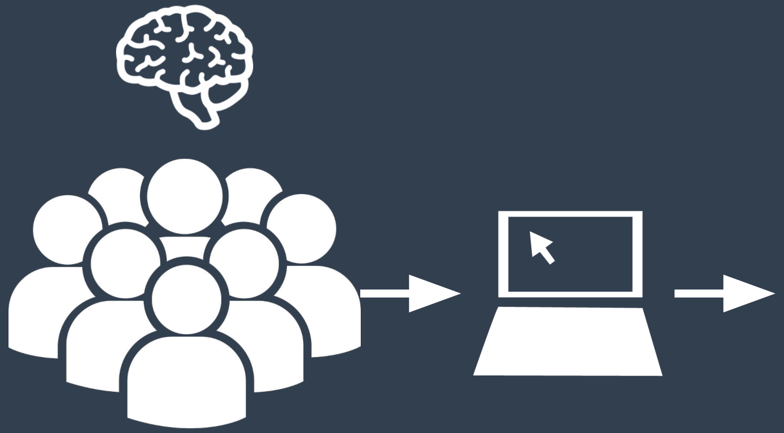
appen™

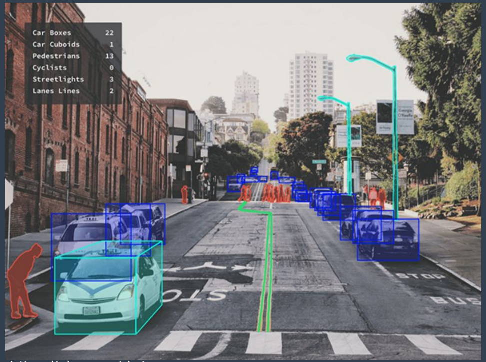
Playment

CrowdFlower

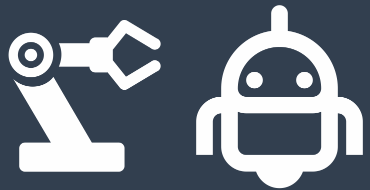
LabInTheWild

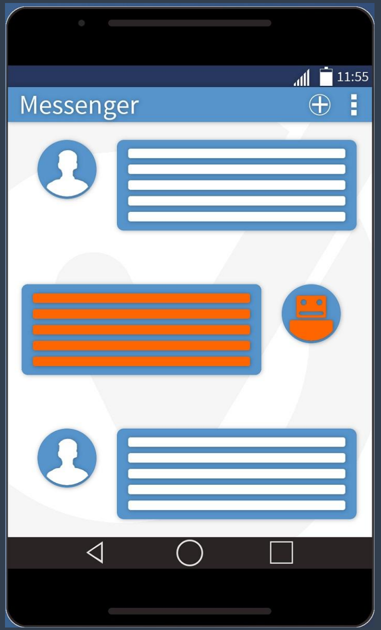
# Crowdsourcing for Human Computation





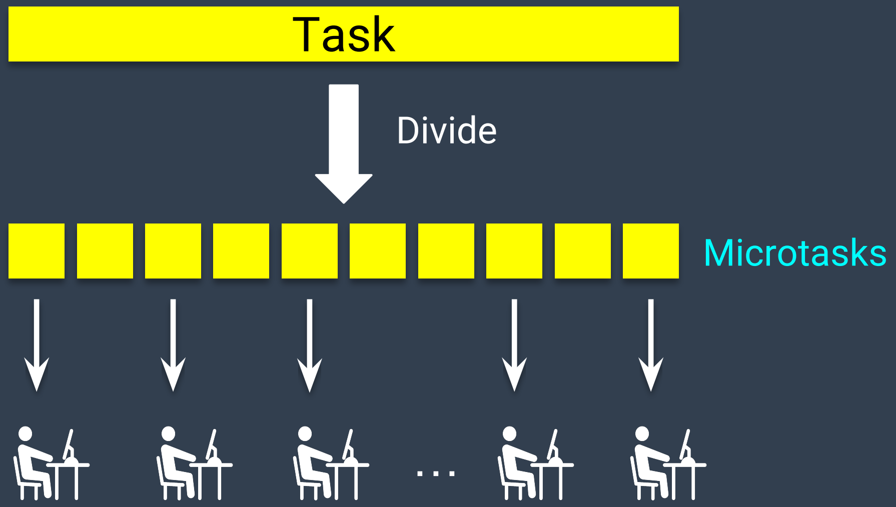
https://playment.io/



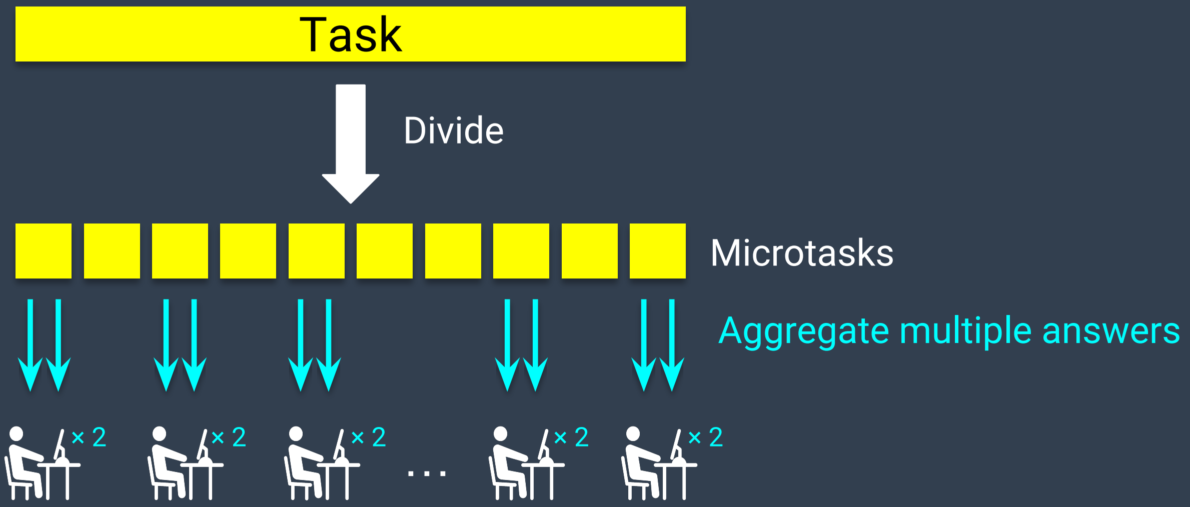


https://www.crowdguru.de/en/

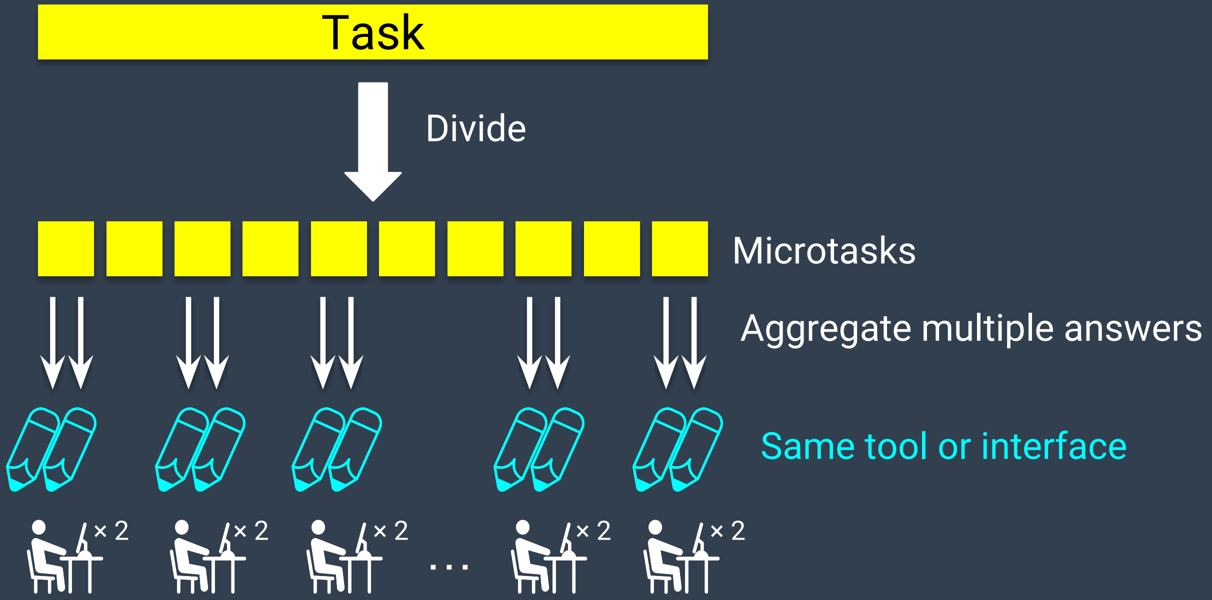
# Crowdsourcing Strategy: Microtasking



# Crowdsourcing Strategy: Aggregation



# Crowdsourcing Strategy: Using Single Tool

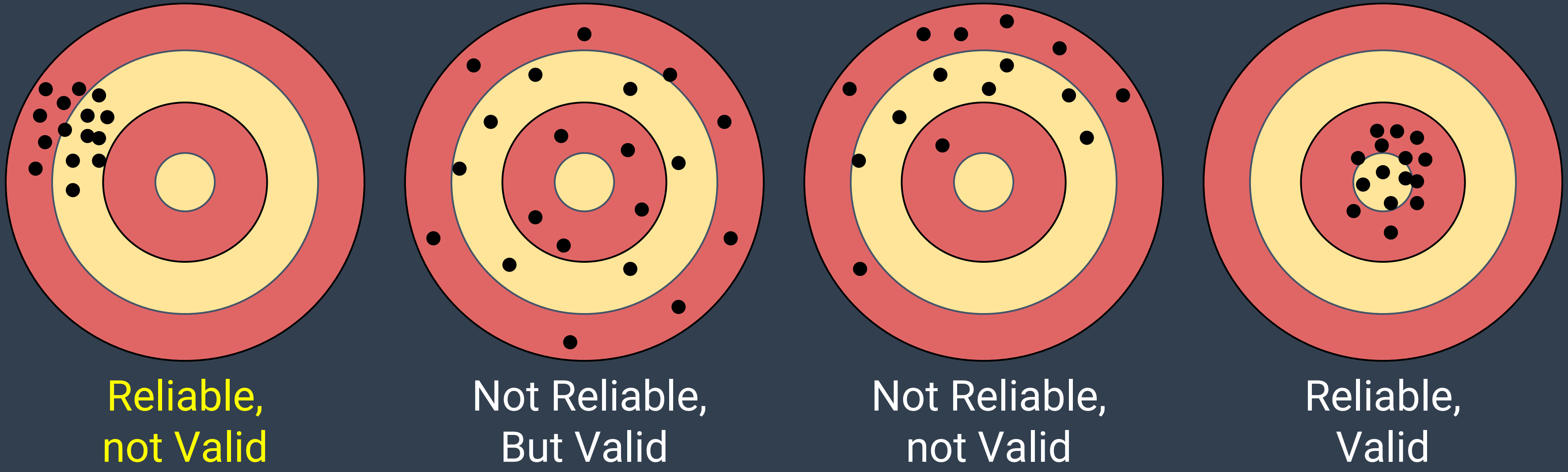


Problem with using a single tool:

Systematic bias can be accumulated, resulting in inaccurate aggregated result.

# Q. What is Systematic Bias?

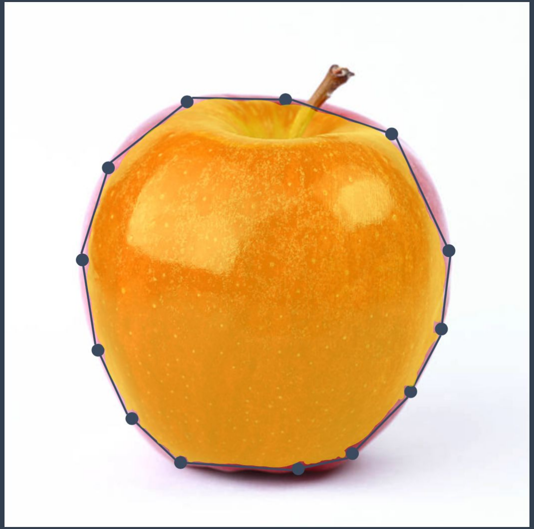
A. Reliable, but not valid performance



# Example of Systematic (Error) Bias

## Tool 1: Opensurfaces (TOG 2013)

Bell, Sean, et al. "**Opensurfaces**: A richly annotated catalog of surface appearance." *ACM Transactions on Graphics (TOG)*32.4 (2013): 111.



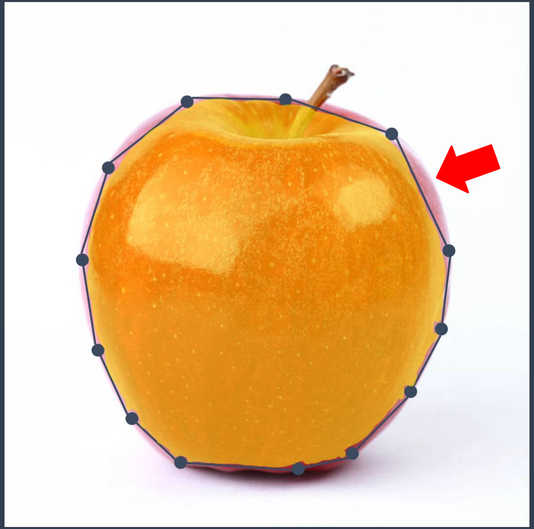
## Tool 2: Click'n'Cut (CrowdMM 2014)

Carlier, Axel, et al. "**Click'n'Cut**: Crowdsourced interactive segmentation with object candidates." *International ACM Workshop on Crowdsourcing for Multimedia*. 2014.

# Example of Systematic (Error) Bias

## Tool 1: Opensurfaces (TOG 2013)

Bell, Sean, et al. "**Opensurfaces**: A richly annotated catalog of surface appearance." *ACM Transactions on Graphics (TOG)*32.4 (2013): 111.



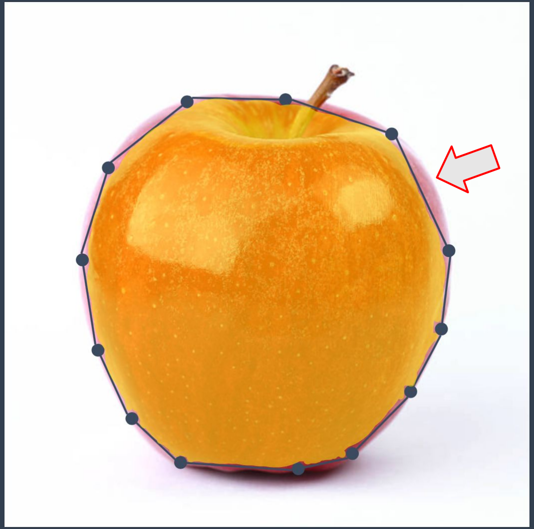
## Tool 2: Click’n’Cut (CrowdMM 2014)

Carlier, Axel, et al. "**Click'n'Cut**: Crowdsourced interactive segmentation with object candidates." *International ACM Workshop on Crowdsourcing for Multimedia*. 2014.

# Example of Systematic (Error) Bias

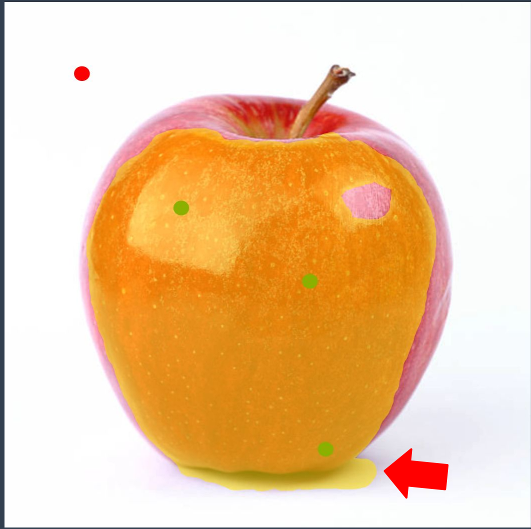
## Tool 1: Opensurfaces (TOG 2013)

Bell, Sean, et al. "**Opensurfaces**: A richly annotated catalog of surface appearance." *ACM Transactions on Graphics (TOG)*32.4 (2013): 111.



## Tool 2: Click'n'Cut (CrowdMM 2014)

Carlier, Axel, et al. "**Click'n'Cut**: Crowdsourced interactive segmentation with object candidates." *International ACM Workshop on Crowdsourcing for Multimedia*. 2014.



## Proposed Approach:

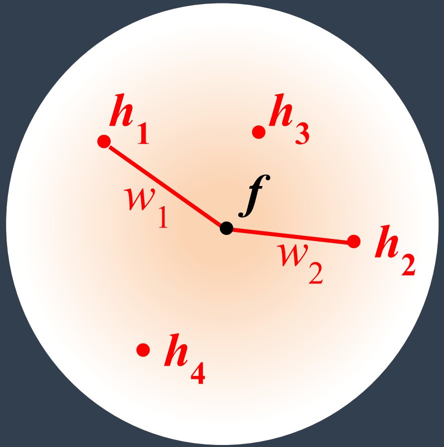
Use tool diversity as a means of improving aggregate crowd performance

## What is Tool Diversity?

A property that measures how different tools can be built in terms of their induced biases.

# Analogy to Ensemble Learning

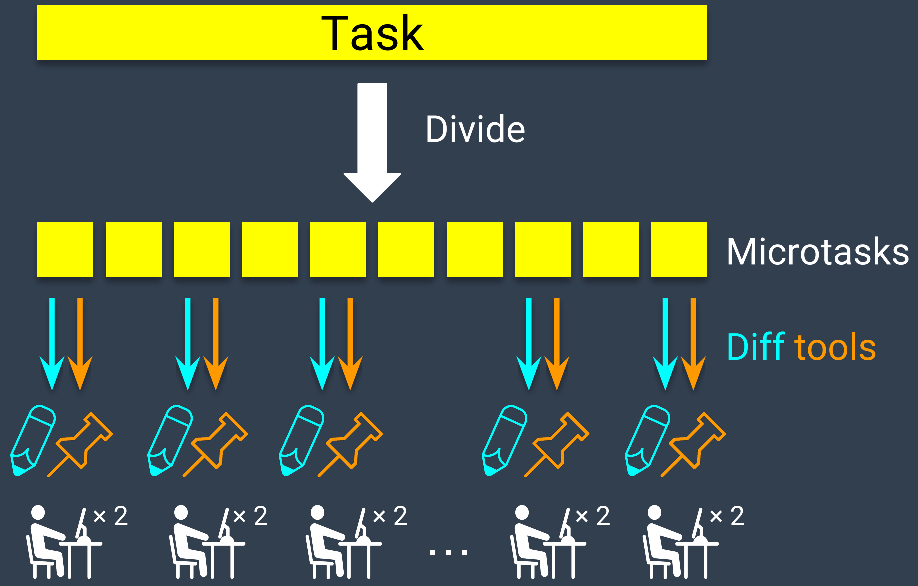
Space of hypotheses



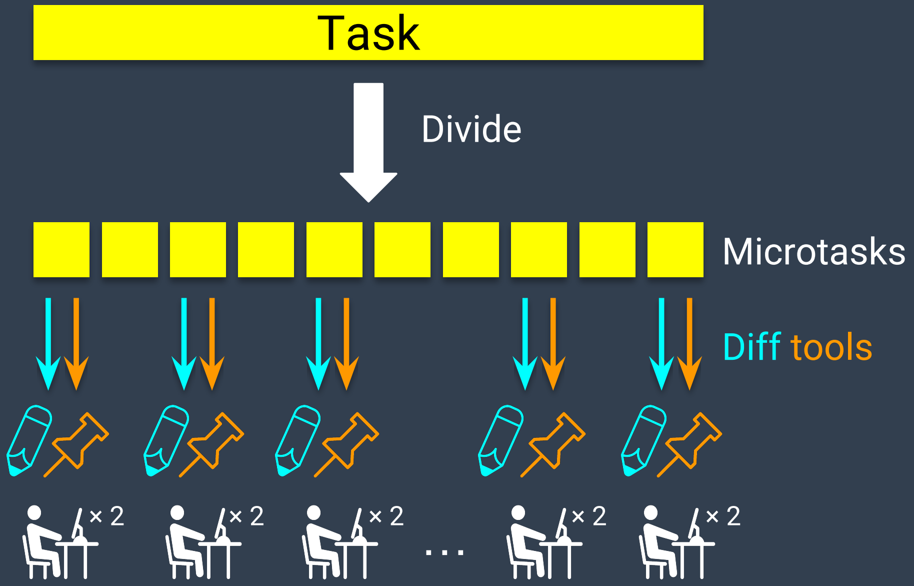

$f$ : best performing hypothesis
$h_i$ : other hypotheses
$w_i$ : weights

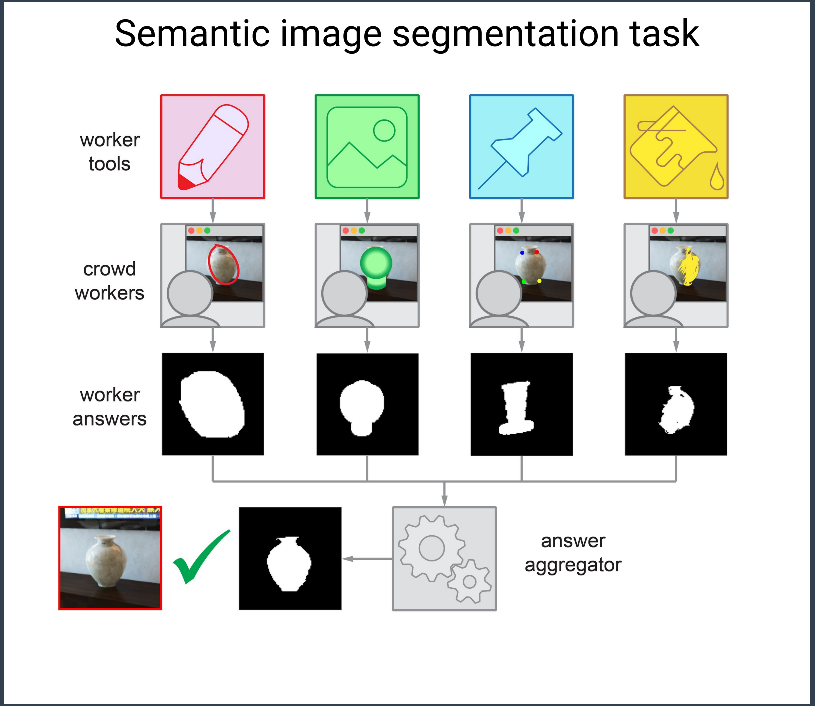
Ensemble learning constructs a combination of two alternative hypotheses $h_1$ and $h_2$ with proper weights ($w_1$ and $w_2$), and approximates the best hypothesis $f$ by averaging the two.

# Proposed Method: Leverage Tool Diversity



# Proposed Method: Leverage Tool Diversity



# Choosing the Tools

Q. How to diversify errors produced by different tool types?

# Choosing the Tools

Q. How to diversify errors produced by different tool types?

Q. What are different types of objects?

$T_1$ $T_2$

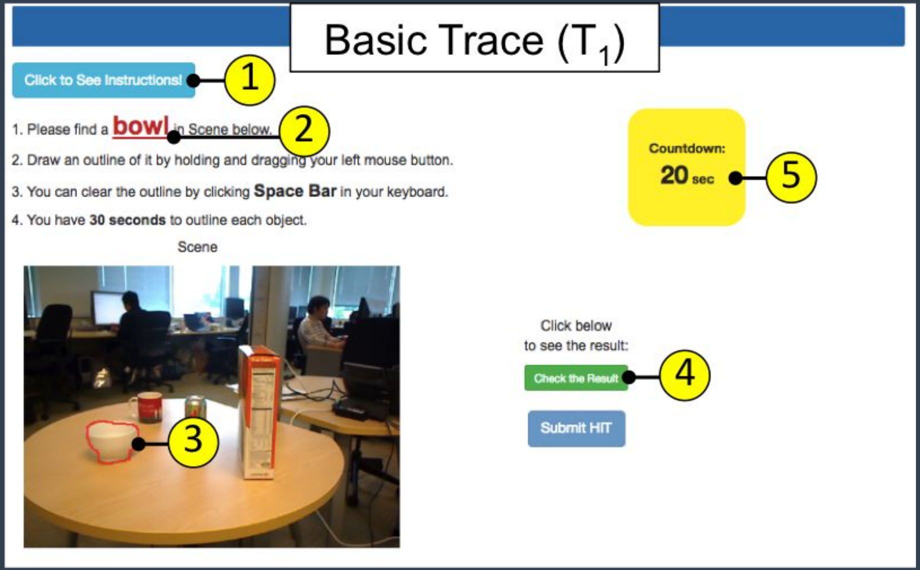
A. General objects, Fuzzy materials, plants, furry objects,

$T_3$ $T_4$

transparent objects, reflective surfaces (intuitive, deformability)

# Instructions and Worker Interface

Worker Interface :



# Instructions and Worker Interface

## Instructions :

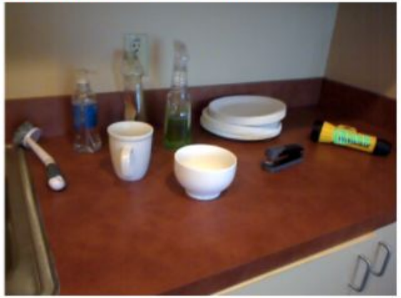
ATTENTION! Please carefully read the instructions before you ACCEPT HIT

(Please don't accept HIT if you've previously worked on this task.
You will not be paid more than onece even though you complete the HIT multiple times.)

Welcome to our object aligning system!
**The task is to align all objects in the list with icon images.** You'll have 30 seconds to align each object.


**Good example** of aligning a flashlight

**Bad examples** of aligning a flashlight



Payments and Expected time:
- Task: (< 5 minutes) task; $0.35 for successful completion.
- IMPORTANT: You must finish(Submit HIT) to get paid.

Please ACCEPT HIT to start task!

# Experiment Settings

- 12 different visual scenes
- Total 51 objects
- Six unique workers for each tool-scene pair (total 288+ workers)
- Total 1224 object segmentations
- Platform: Amazon Mechanical Turk

Each worker was paid between $0.35 and $0.60 per task, depending on the number of objects they had to segment or on the level of difficulty of given tool (a pay rate of ~$10/hr).

# Results & Discussion

# Performance of Individual Tools

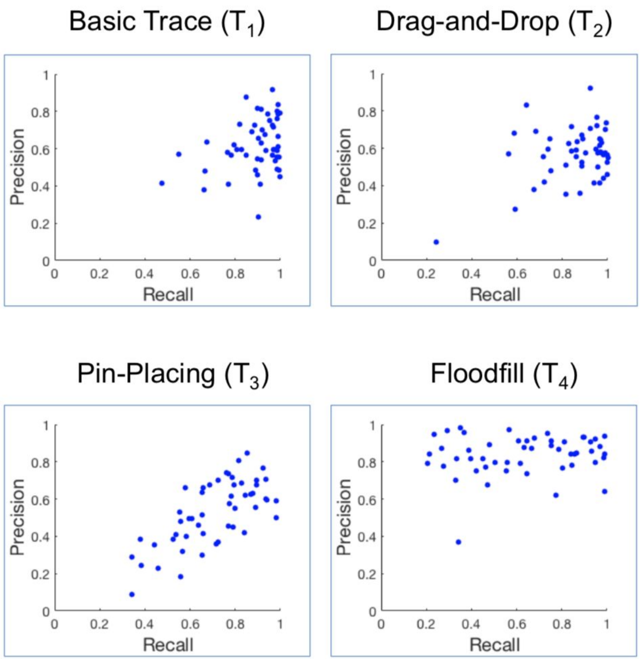
## Accuracy Metrics

$$\text{Precision (P)} = \frac{TP}{TP + FP}$$

$$\text{Recall (R)} = \frac{TP}{TP + FN}$$

$$F_1 \text{ Score} = \frac{2 \times P \times R}{(P + R)}$$

TP = True Positive
FP = False Positive
FN = False Negative




Different tools have different error patterns

# Performance of Individual Tools

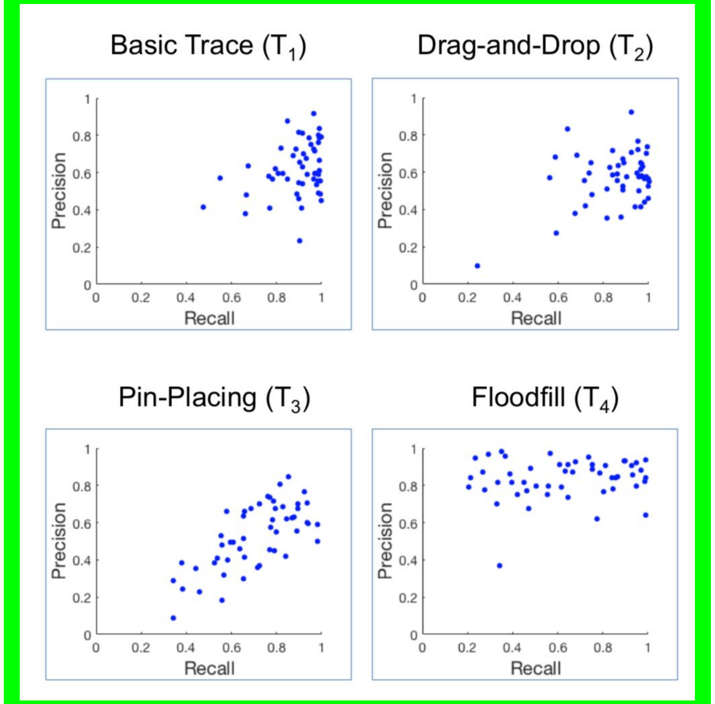
## Accuracy Metrics

$$\text{Precision (P)} = \frac{TP}{TP + FP}$$

$$\text{Recall (R)} = \frac{TP}{TP + FN}$$

$$F_1 \text{ Score} = \frac{2 \times P \times R}{(P + R)}$$

TP = True Positive
FP = False Positive
FN = False Negative




Different tools have different error patterns

# What we observed

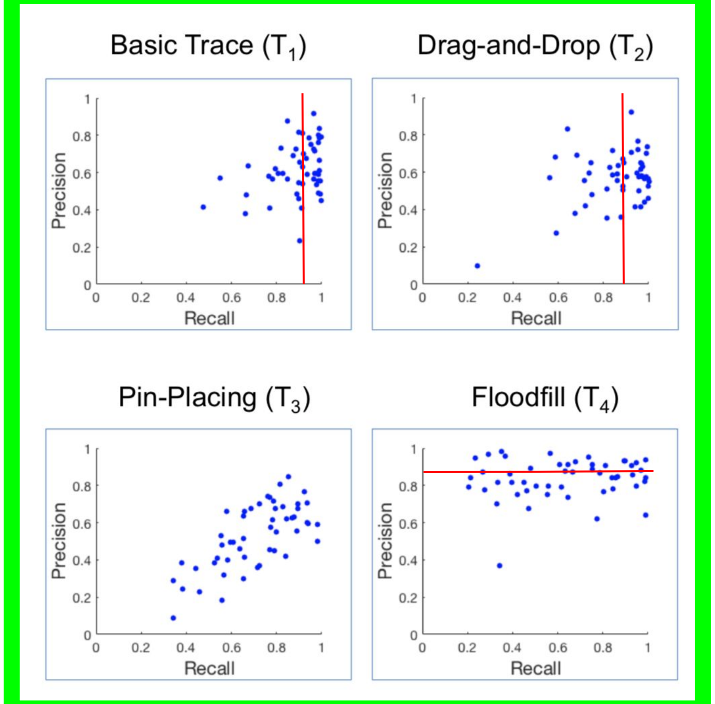
### Accuracy Metrics

$$\text{Precision (P)} = \frac{TP}{TP + FP}$$

$$\text{Recall (R)} = \frac{TP}{TP + FN}$$

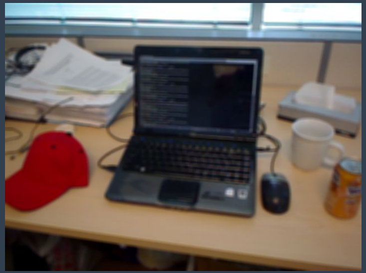
$$F_1 \text{ Score} = \frac{2 \times P \times R}{(P + R)}$$

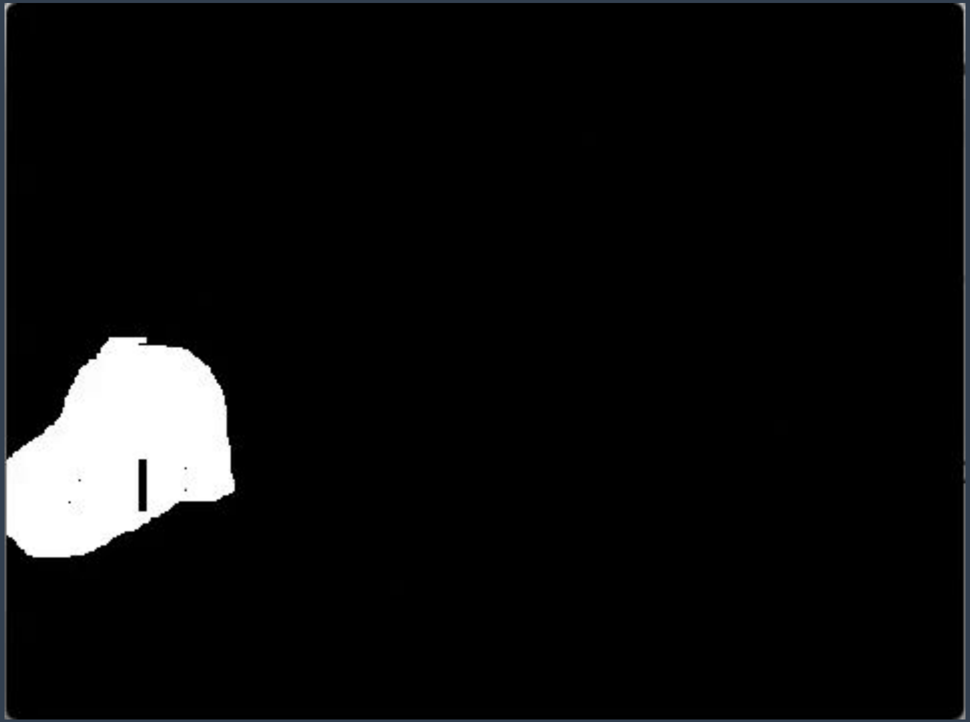
TP = True Positive
FP = False Positive
FN = False Negative




**Different tools have different error patterns**

# Some of the Answers from Workers





How can we see
the effect of leveraging tool diversity?

# Comparison of Aggregation Methods

Method 1. Single tool aggregation (Uniform majority voting): Baseline

$T_1$ → Aggregate

$T_2$ → Aggregate

# Comparison of Aggregation Methods

Method 2. Multiple tool aggregation (Uniform majority voting)

$T_1$ x $T_2$  → Aggregate

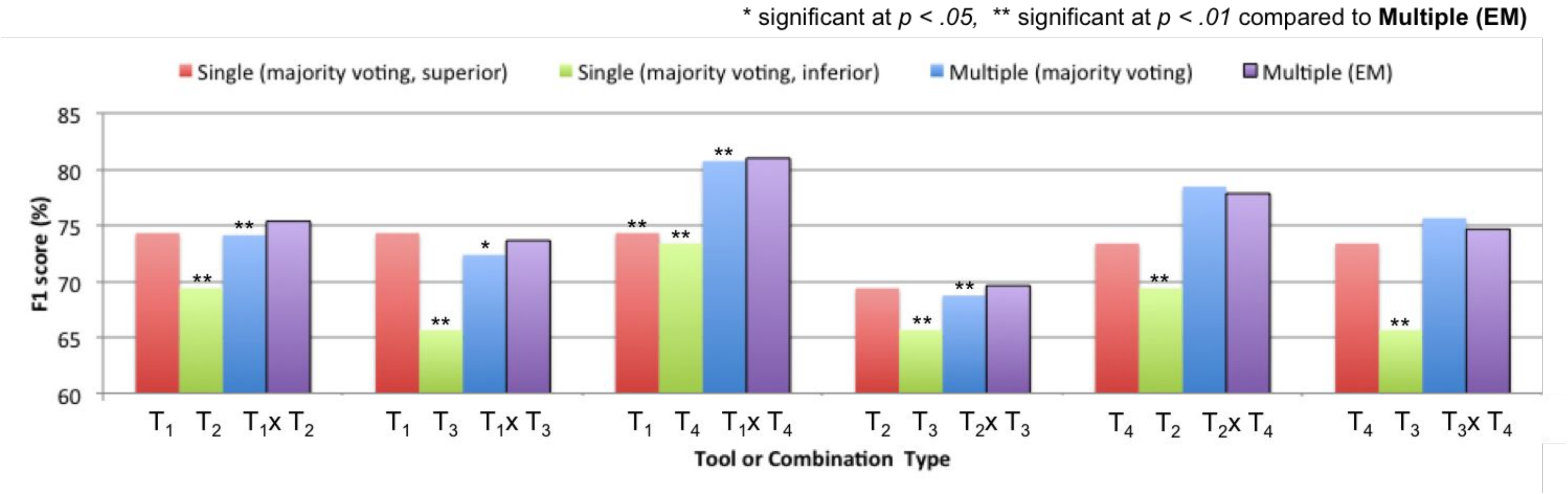
$w \quad w \quad w \quad w$

Method 3. Multiple tool aggregation (Expectation maximization)

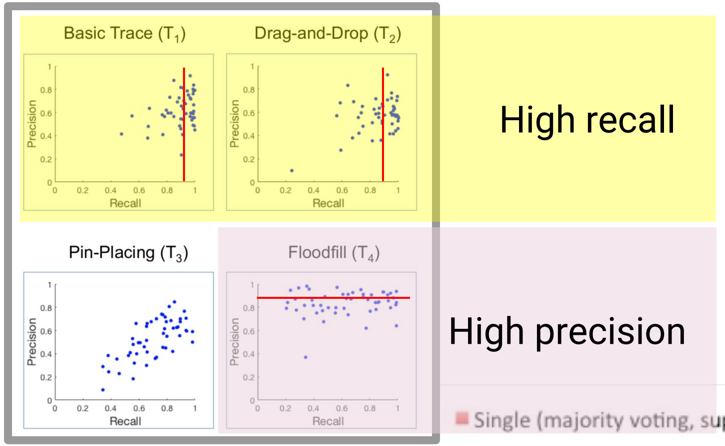
$T_1$ x $T_2$  → Aggregate

$w_1 \quad w_2 \quad w_3 \quad w_4$

# Comparison of Aggregation Methods

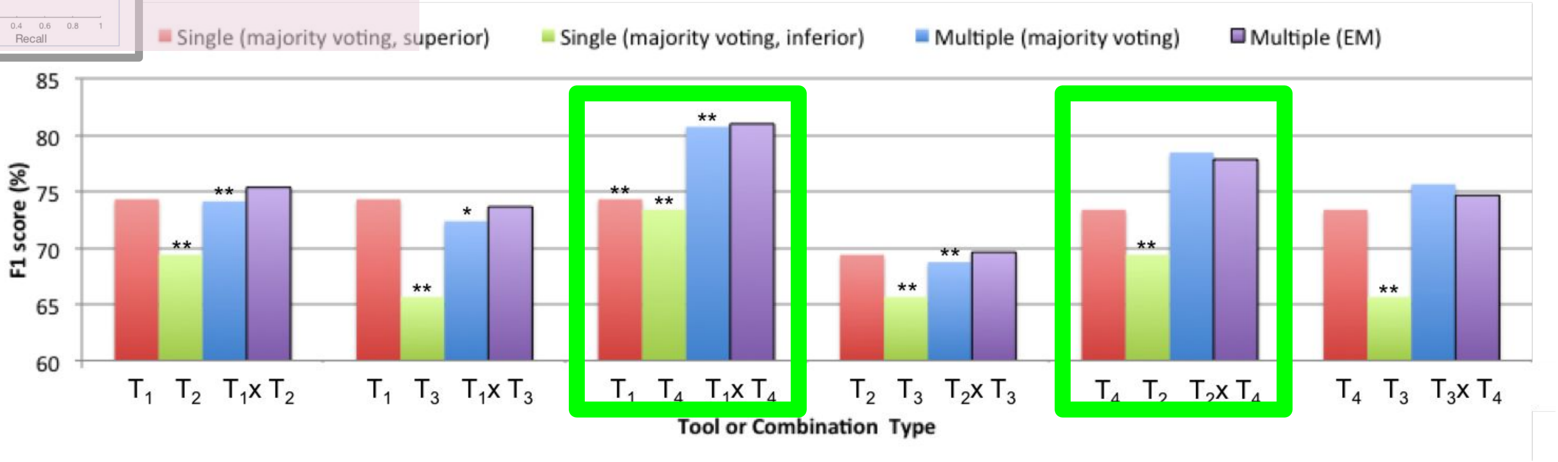


# Comparison of Aggregation Methods



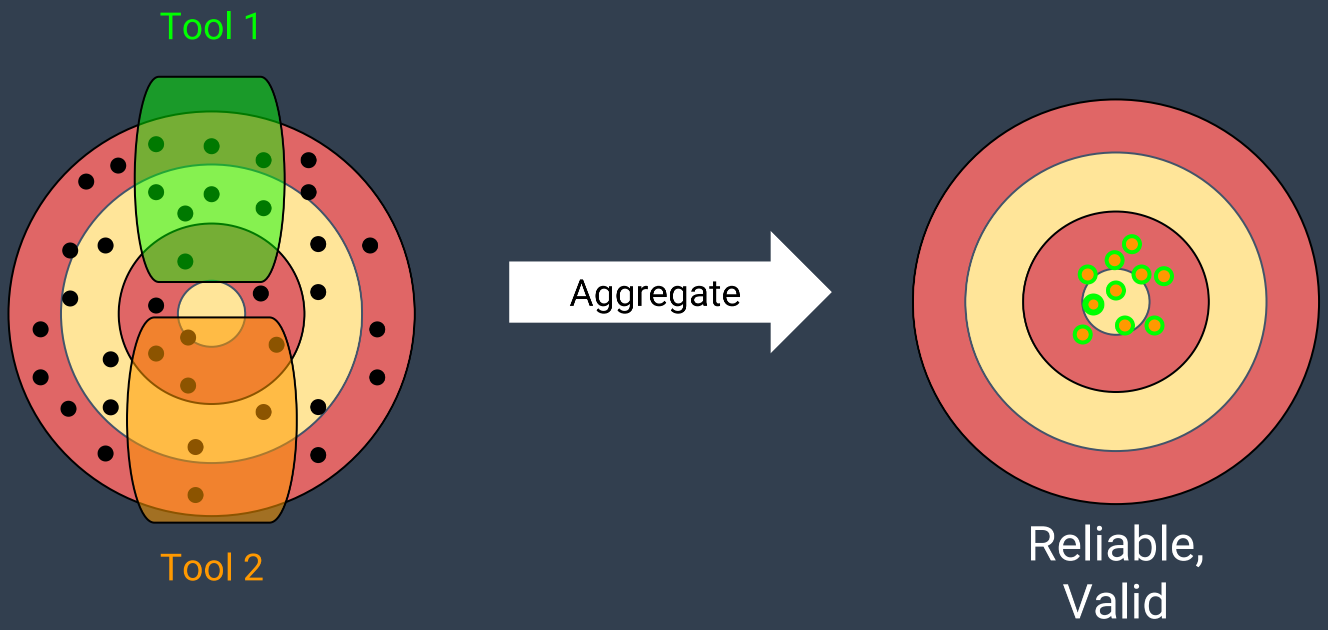

High recall + high precision pairs gave the highest performance improvement.

* significant at $p < .05$, ** significant at $p < .01$ compared to **Multiple (EM)**

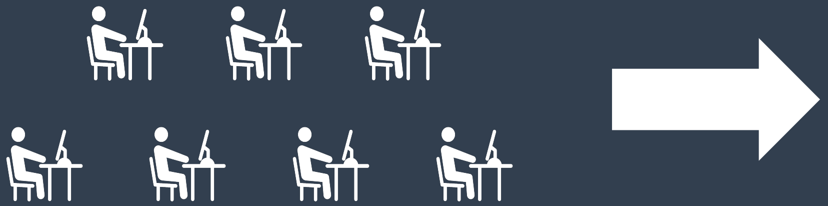


# Generalization

# Generalizability: Expected Human Error is Diverse



# Generalizability: Aggregation Improves Quality



**Quality Improves**

# Generalizability: Objective Correct Answer Exists

**Tasks with objective answers:**

Image segmentation

Live captioning

Text annotation

Handwriting recognition

Task with subjective answers:

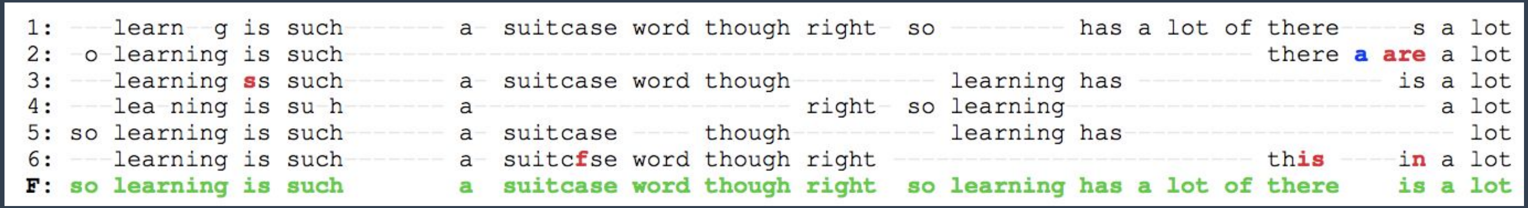
Creative writing

This paper presents Soylent, a word processing interface that uses crowd workers to help with proofreading, document shortening, editing and commenting tasks. Soylent is ~~an example of a~~ new ~~kind of~~ interactive user interface in which the end user has direct access to a crowd of workers for assistance with tasks that require human attention and common sense. Implementing these ~~kinds of~~ interfaces requires new software programming patterns ~~for interface software~~, since crowds behave differently than computer systems. We have introduced one important pattern, Find-Fix-Verify, which splits complex editing tasks into a series of identification, generation, and verification stages ~~that use independent agreement and voting~~ to produce reliable results. We evaluated Soylent with a range of editing tasks, finding and correcting 82% of grammar errors ~~when combined with automatic checking~~, shortening text to approximately 85% of original length per iteration, and executing a variety of human macros successfully.

# Generalizability: Tolerates Imperfections

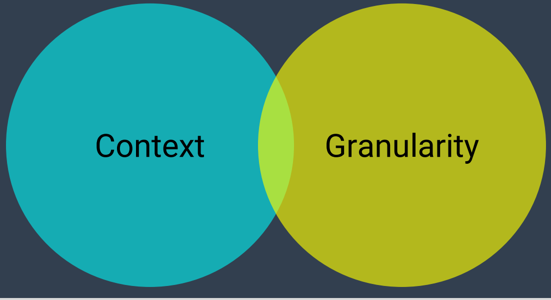
Example: Scribe (UIST 2012)

W.S. Lasecki, C.D. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, J.P. Bigham. Real-time Captioning by Groups of Non-Experts. UIST 2012.

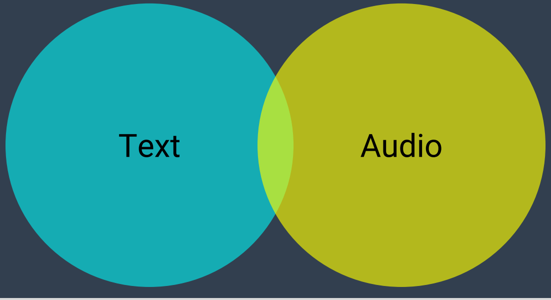


# Possible Future Applications

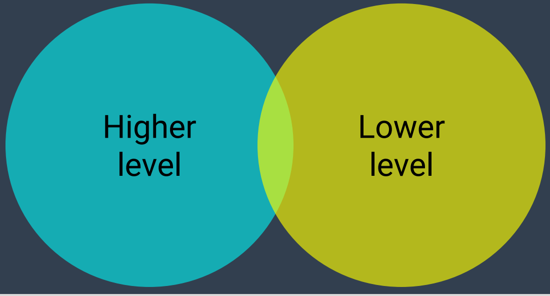
## Application1: Tagging Long Videos



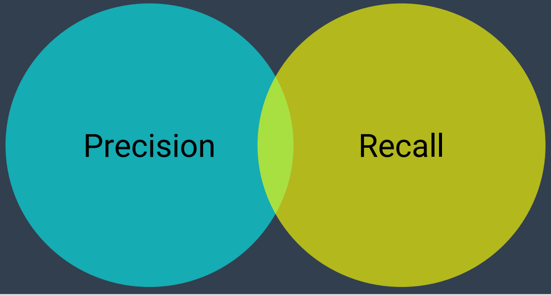

## Application2: Multichannel NLP




## Application3: Complex/Diverse Annotation




## Application4: Computer-Human Integration



# Thank you!

## Authors:


Jean Y. Song (jyskwon@umich.edu / jyskwon.github.io),
Raymond Fok, Alan Lundgard, Fan Yang, Juho Kim, Walter S. Lasecki

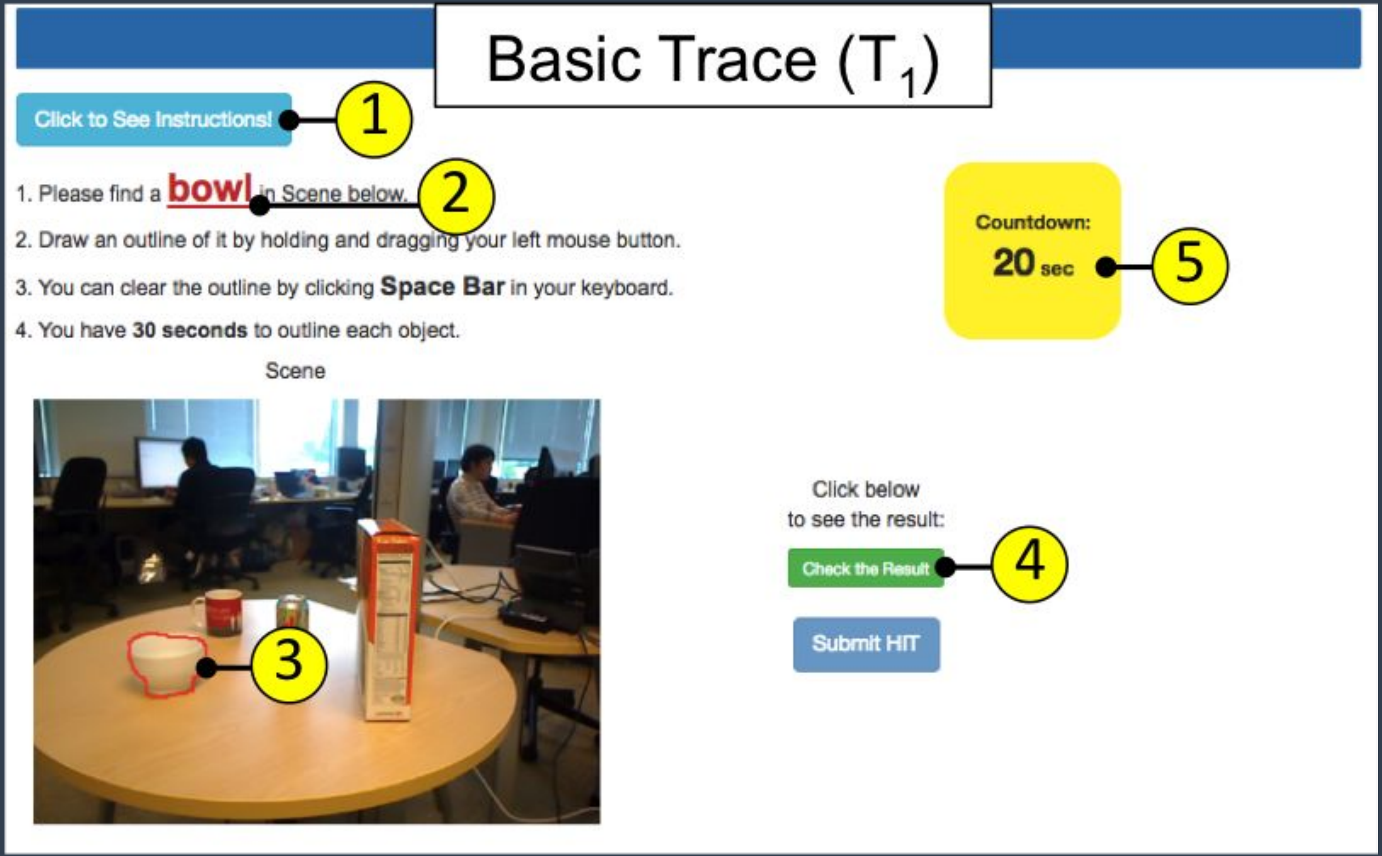

## Funding:


Denso Corporation
Toyota Research Institute
MCity at the University of Michigan
National Research Foundation of Korea


Michigan Interactive and Social Computing Group

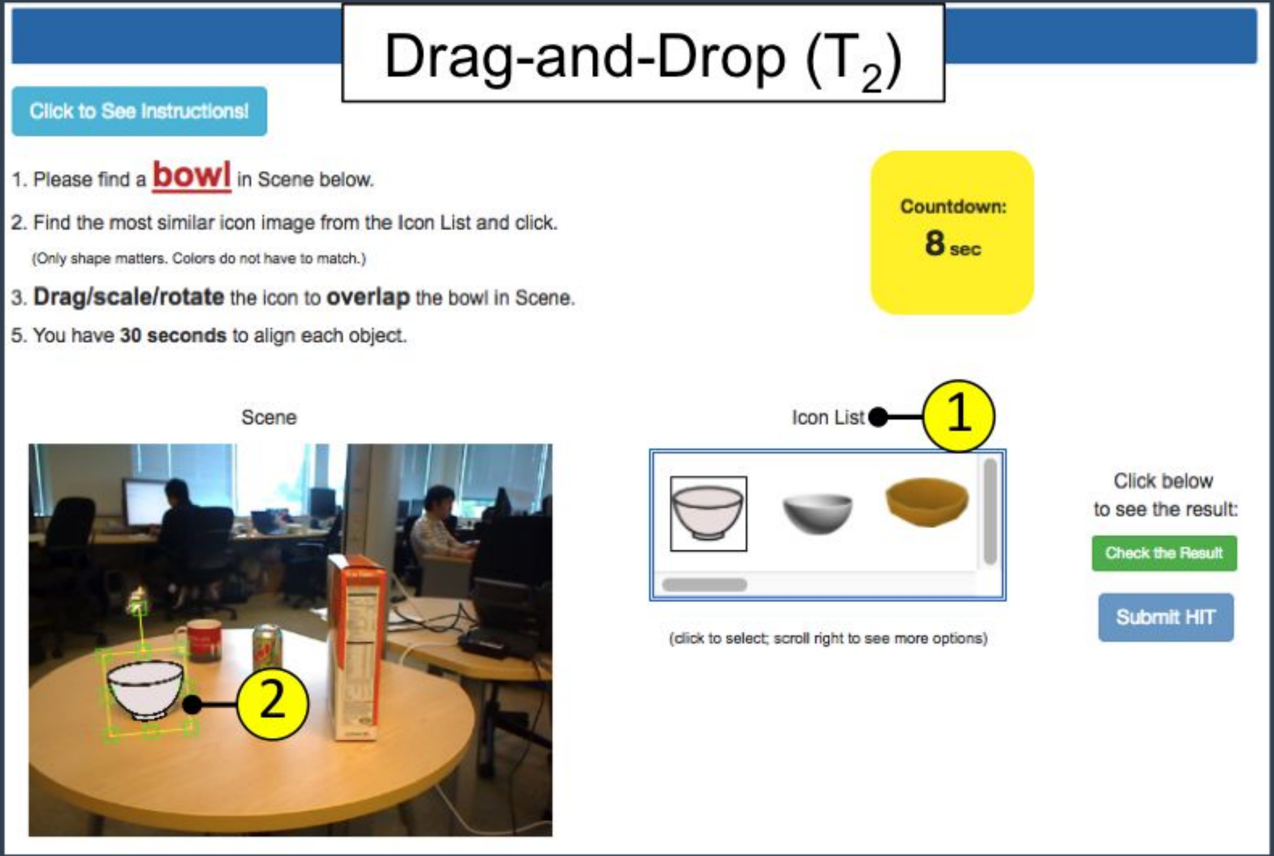
# Backup Slides

# Tool 1



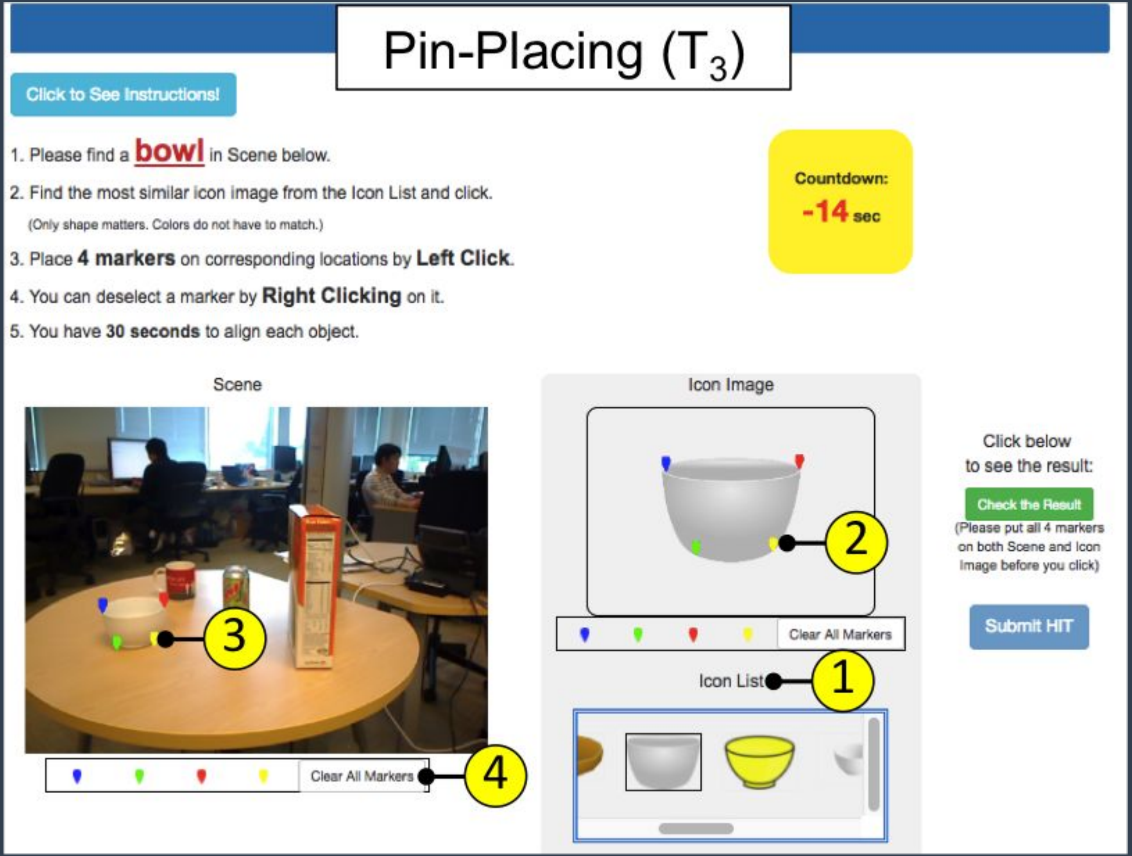
Basic Trace ($T_1$)
Click to See Instructions!
1
1. Please find a bowl in Scene below.
2
2. Draw an outline of it by holding and dragging your left mouse button.
3. You can clear the outline by clicking Space Bar in your keyboard.
4. You have 30 seconds to outline each object.
Scene
3
Countdown:
20 sec
5
Click below
to see the result:
Check the Result
4
Submit HIT

# Tool 2



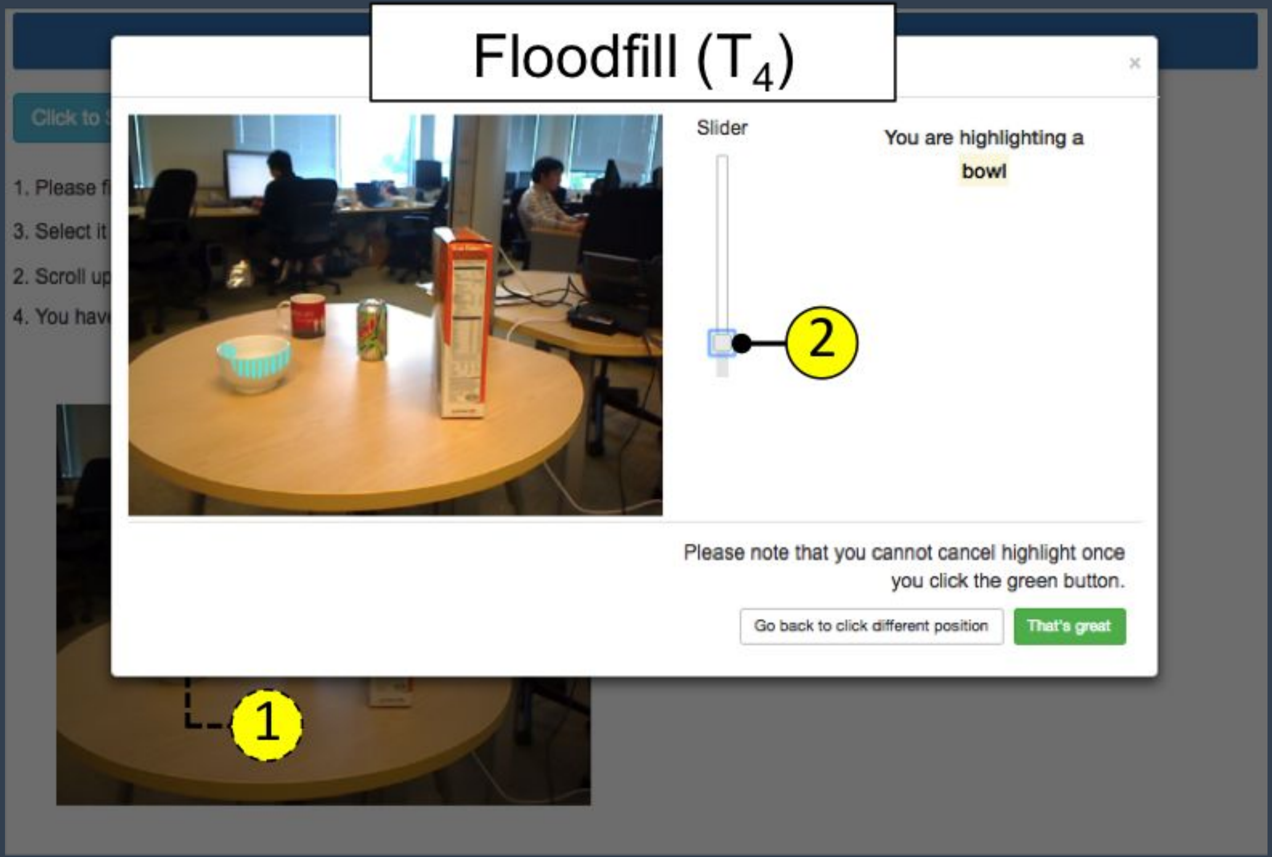
Drag-and-Drop ($T_2$)
Click to See Instructions!
1. Please find a bowl in Scene below.
2. Find the most similar icon image from the Icon List and click.
(Only shape matters. Colors do not have to match.)
3. Drag/scale/rotate the icon to overlap the bowl in Scene.
5. You have 30 seconds to align each object.
Countdown:
8 sec
Scene
Icon List
1
2
(click to select; scroll right to see more options)
Click below
to see the result:
Check the Result
Submit HIT

# Tool 3



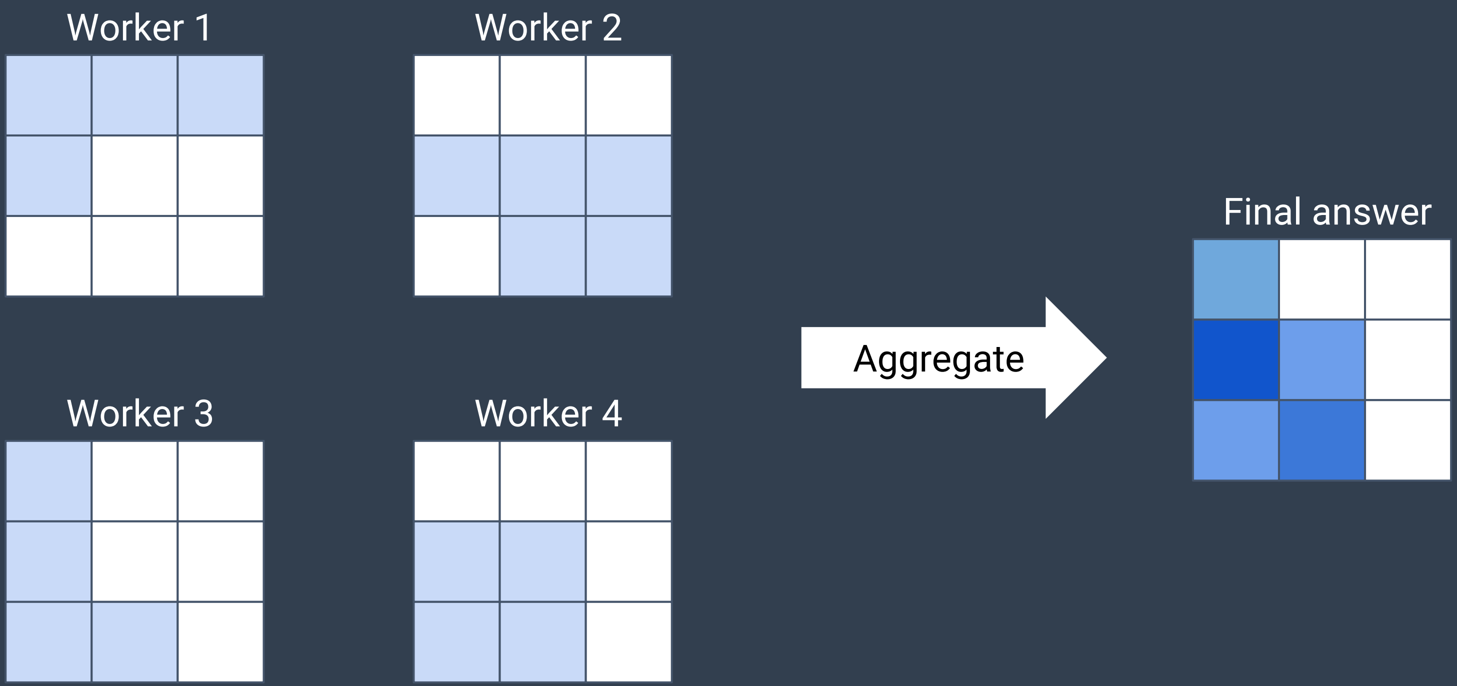
Pin-Placing ($T_3$)
Click to See Instructions!
1. Please find a bowl in Scene below.
2. Find the most similar icon image from the Icon List and click.
(Only shape matters. Colors do not have to match.)
3. Place 4 markers on corresponding locations by Left Click.
4. You can deselect a marker by Right Clicking on it.
5. You have 30 seconds to align each object.
Countdown:
-14 sec
Scene
Icon Image
Click below
to see the result:
Check the Result
(Please put all 4 markers on both Scene and Icon Image before you click)
Submit HIT
Clear All Markers
Icon List
1
2
3
4
Clear All Markers

# Tool 4



Floodfill ($T_4$)
Click to
1. Please f
3. Select it
2. Scroll up
4. You hav
Slider
You are highlighting a
bowl
2
Please note that you cannot cancel highlight once you click the green button.
Go back to click different position
That's great
1

# Pixel-Level Majority Voting (50% agreement)



# Expectation Maximization (Dawid-Skene Algorithm)

In an image, label a pixel as $1$ if it belongs to a target object, and $0$ if background.

Assume:

- image $A$ having $N$ total pixels
- $M$ crowd workers
- The label a worker $m$ assigns to each pixel is denoted as $z_{mn}$
- all labels from worker $m$ as a vector $Z_m$
- the true labels of $A$ to be estimated are denoted as a vector $Y$
- θ is the confusion matrices set to be estimated.

We can estimate the true labels $Y$ by maximizing the marginal likelihood of the observed worker labels:

$$l(\theta) := \log\Big( \sum_{Y \in \{0,1\}^n} L(\theta; Y, Z)\Big)$$

The EM algorithm works iteratively by applying the 1) expectation step and the 2) maximization step.