Sprout: Crowd-Powered Task Design for Crowdsourcing

Jonathan Bragg
University of Washington
(Stanford University)

Mausam
Indian Institute of Technology Delhi

Daniel S. Weld
University of Washington

UW CrowdLab

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Crowdsourcing Task Platforms

Volunteer

Paid

amazon mechanical turk

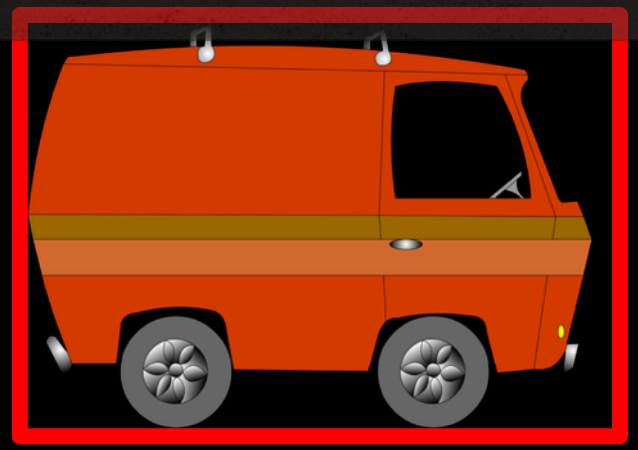
upwork™
formerly oDesk

Instructions:

Is this a car?

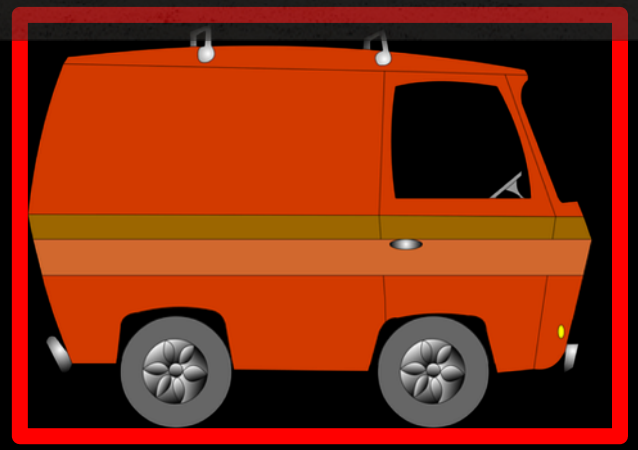
Instructions:

Is this a car?



Instructions:

Is this a real (not cartoon) car?



Instructions:

Is this a real (not cartoon) car that a person can drive on a street?

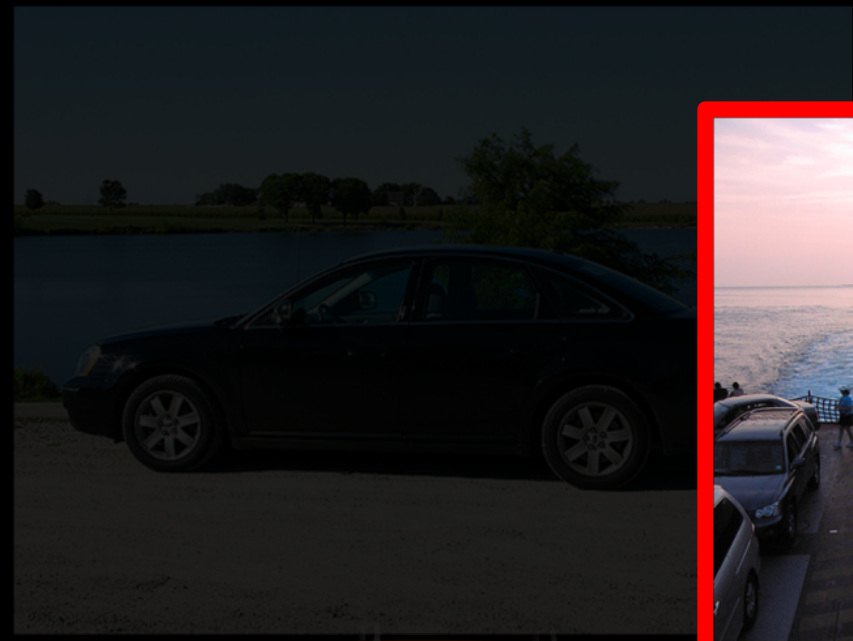
Low Inter-annotator Agreement

Yes

No

Instructions:

Is this a single real (not cartoon) car that a person can drive on a street?



Instructions:

Is this a single real (not cartoon) car that a person can drive on a street?
The whole car must be visible for it to count.

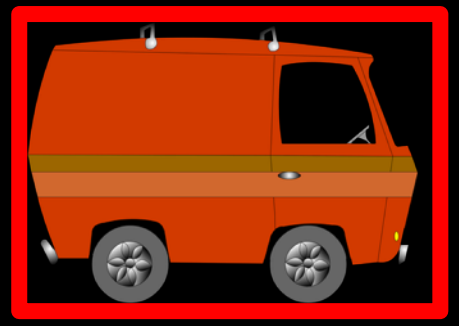
Iterative Task Design Loop

Revise Instructions

Revise Training & Testing

<u>Perform (explore) the task</u>

- Task definition changes
- Concept evolution [Kulesza et al., CHI ‘14]



<u>Run (debug) the task</u>

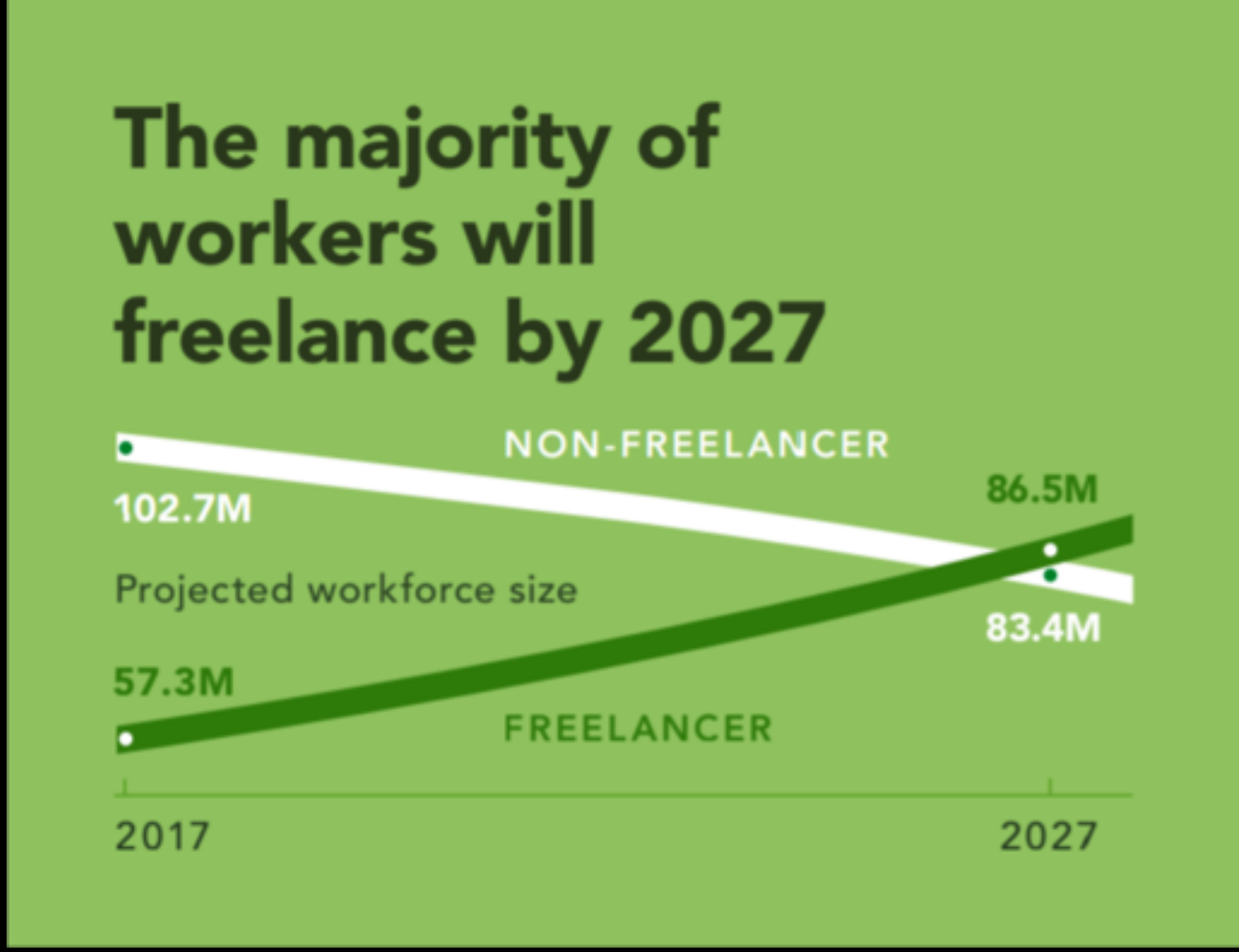
- Worker mistakes initially unknown
- Model mismatch / debugging

Low Inter-Annotator Agreement

Task Design is a Major Problem

- Difficulty for Requesters; Importance for Work Quality
  - [Alonso & Mizzaro, Information Processing and Management ‘12; Papoutsaki et al. HCOMP ‘15; Liu et al., NAACL ‘16]
- Risk to Workers: Task Design Flaws & Unfair Rejections
  - [McInnis et al., CHI ‘16; Gadiraju et al., HT ’17; Wu & Quinn, HCOMP ‘17]

…and it matters



[Freelancing in America: 2017. Upwork and Freelancers Union]

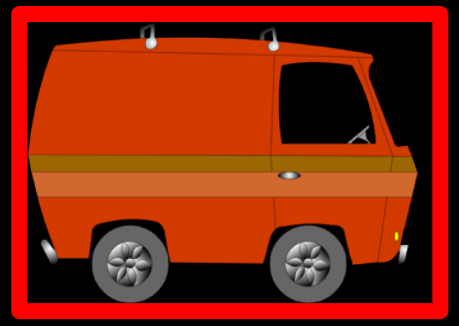
How can we design tools for task design?

Revise Instructions

Revise Training & Testing

<u>Perform (explore) the task</u>

- Task definition changes
- Concept evolution [Kulesza et al., CHI ‘14]

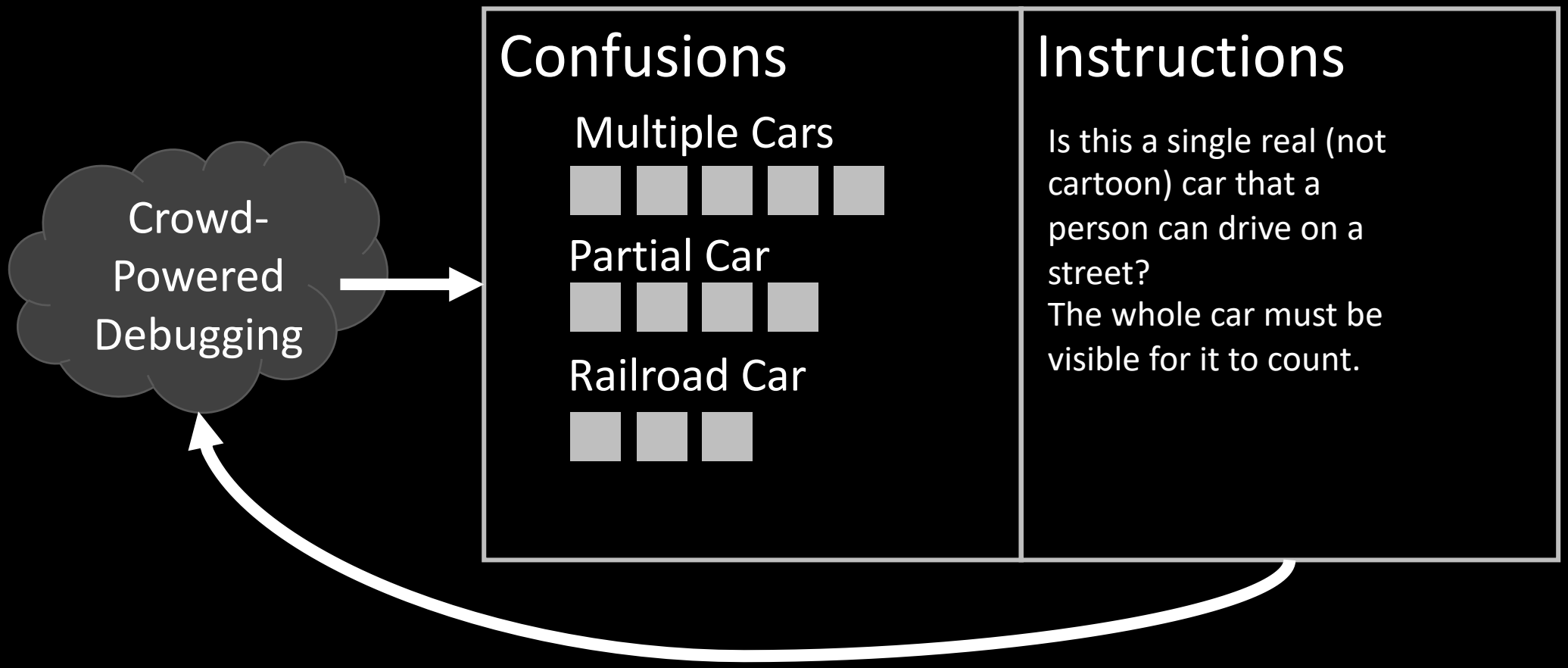


<u>Run (debug) the task</u>

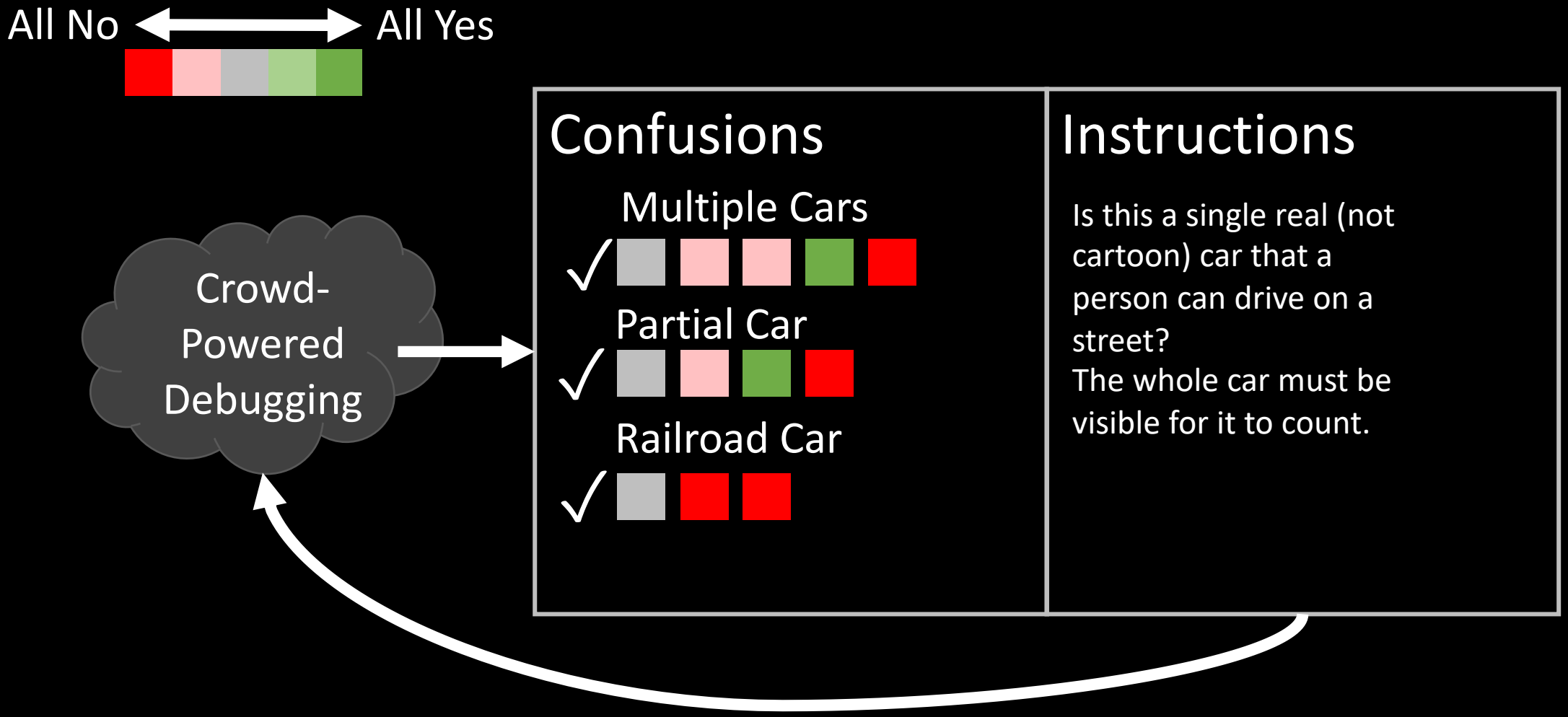
- Worker mistakes initially unknown
- Model mismatch / debugging

Low Inter-Annotator Agreement

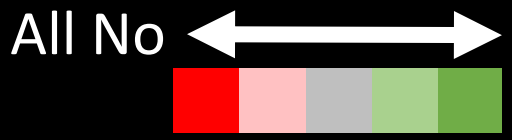
Sprout: Debugging-Prioritized Exploration



Sprout: Debugging-Prioritized Exploration



Sprout: Debugging-Prioritized Exploration



Test Q's

Crowd-Powered Debugging

Confusions

Multiple Cars ✓

Partial Car ✓

Railroad Car ✓

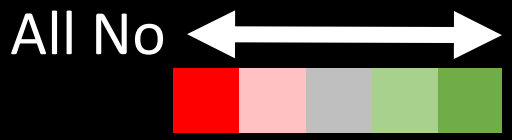
Instructions

Is this a single real (not cartoon) car that a person can drive on a street?
The whole car must be visible for it to count.

Yes

No

Sprout: Debugging-Prioriti



Suggested Test Question:

Q's

Crowd-Powered Debugging

Confusions

Multiple Cars ✓

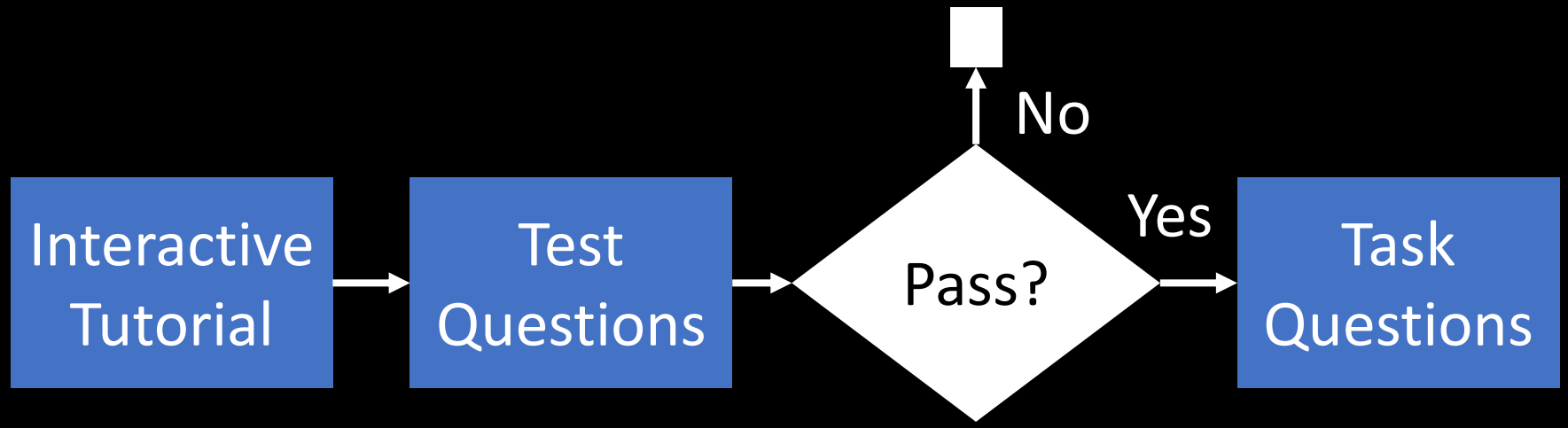
Partial Car ✓

Railroad Car ✓

Is this a single real (not cartoon) car that a person can drive on a street?
The whole car must be visible for it to count.

No

Sprout Compiles Test Q's into Training & Testing



Gated Instruction [Liu et al., NAACL '16]

Confusions viewed

14

Confusions

a

No ? Yes

- ✔is a train car
- ✔is multiple cars
- ✔is a truck
- ✔is a train
- ✔is a bumper car
- ✔has a car as the main subject
- ✔is a military vehicle
- ✔is the interior of a car
- is an automobile
- ✔is a car on a ferry boat

349 444

Related

- is car on a boat (1 in common)
- ✔includes more than one (1 in common)

444

- is a street car
- is a trolley car
- is a subway train car
- is a toy car

Preview

b

444

Answer	Reason
?	is a car on a ferry boat
?	includes more than one
yes	There are several clearly visible vehicles in the photograph.

Similar items

Instructions

Your instructions for workers go here. Use twitter mention notation to reference items, for example, `@10` refers to item 10 and will preview as . You may also use other types of Markdown to format your instructions, like

- `- This` will be a list item
- This will be `**bold**`

v0 v1

Write Preview

c

Is this an image of a car?

Select "yes" even if there are multiple cars (e.g., @444)

Recommended test questions

because you mentioned 444 in the instructions

d

Test questions

yes

no

Submit (87:21 remaining)

Crowd-Powered Debugging

Write a change to our instructions (or choose one written by another worker) that:

- will make the question have a single correct answer
- will make this group of HITs more clear to workers

Better instructions

Is this an image of a car?

- If it is a railroad car you should answer no

Meta W

Work

Run the main task

Diagnose

Diagnose uncertain questions

Clarify

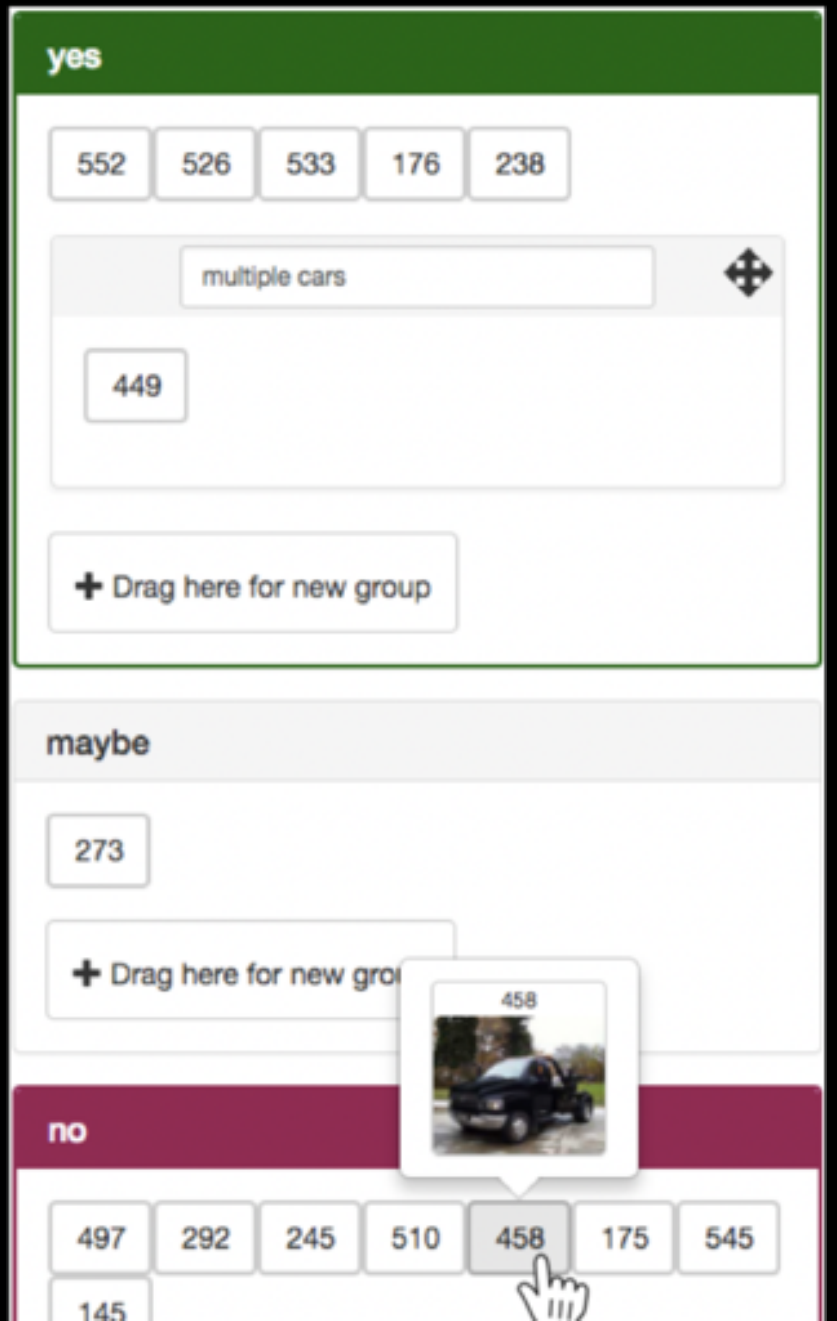
Suggest task changes

GenTest

Create test questions

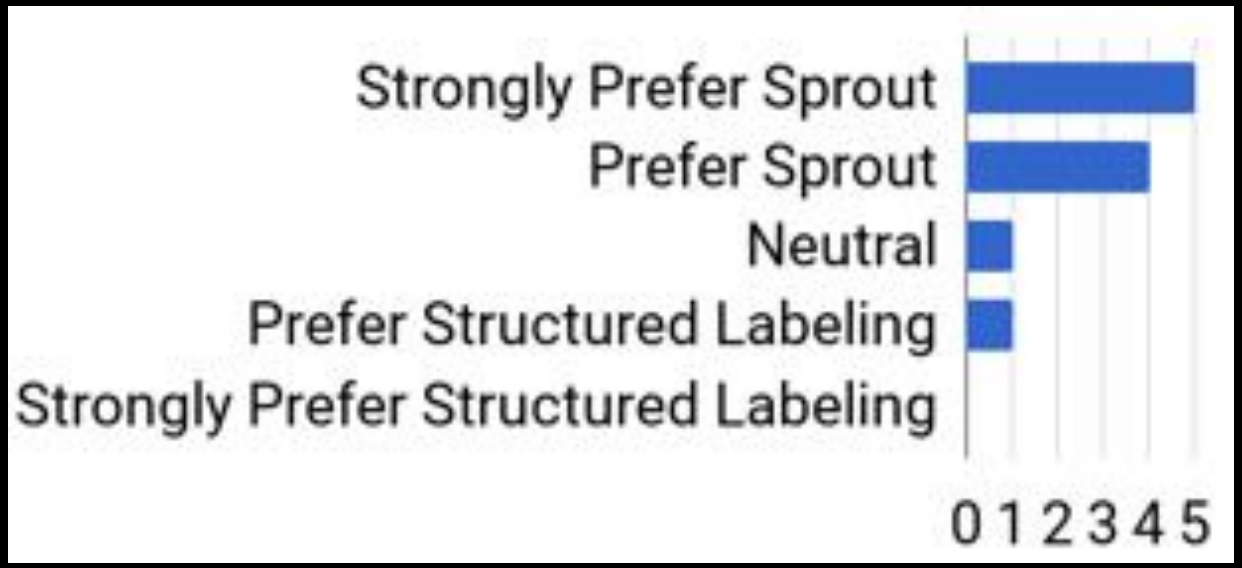
Requester User Study

- Goal: improve underspecified instructions prompts
- 2 domains from prior work (car images and travel websites)
- Baseline no-crowd interface: Structured Labeling [Kulesza et al. CHI 2014]
- Within-subjects (counterbalanced interface)
- 11 participants
  - varying crowdsourcing experience
  - graduate and undergraduate students
  - 5 male, 5 female, 1 other



Sprout preferred by requesters

"Which interface did you prefer for creating instructions?"



"[c]ategorizing the inputs, showing me the cases where there was confusion, etc., made it SUPER easy to identify cases that needed clarification." (P1)

"The similarity metrics seem to be working great and the suggested items are great for testing the points I emphasized in the instructions." (P2)

20-29% of test questions from suggestions

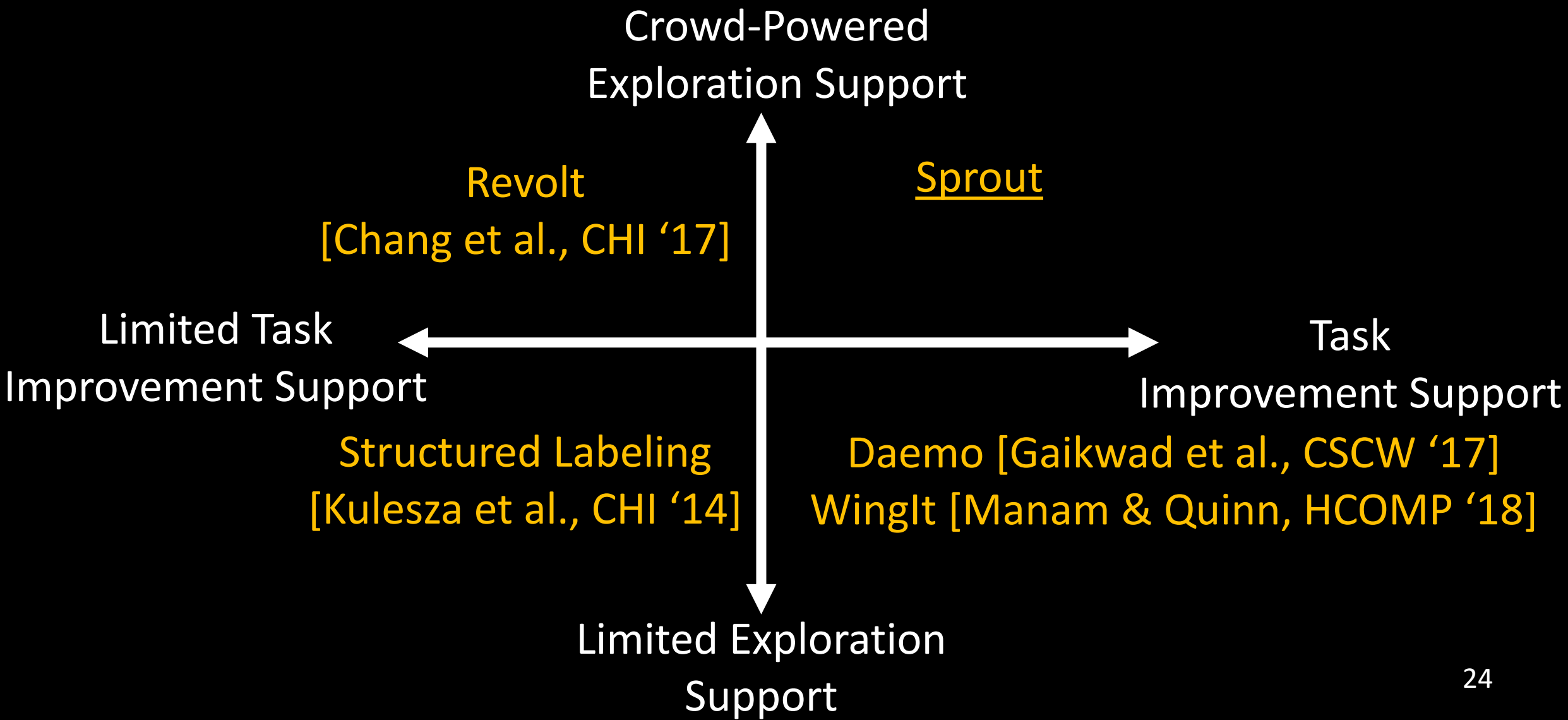
Sprout aids instruction comprehensiveness

- Longer instructions and more examples (mean=4.6 vs 2.6 on travel task)
- More resolved ambiguous categories (mean=4.0 vs. 2.8 on cars task)
  - 2 experts coded number of distinct categories

Some requesters wanted to use both tools

- Structured labeling benefits
  - No bias (P2)
  - “Let me try doing the task myself” (P3)
- Tradeoff: “It sucks that you have to start from a completely blank slate. [SPROUT] gave you some more support.” (P4)

Task Design Tool Landscape



Future Work

- Expanded roles for workers & algorithms
  - More automated instructions: Infer requester decisions (“legal precedent”)
  - Engaging, progressive tutorial
  - Interface design
  - Workflow design (task decomposition)
- More studies
  - Field deployments
- Beyond labeling tasks
  - Open-ended & creative tasks

Thanks!

https://jonathanbragg.com

Sprout: Crowd-Powered Task Design for Crowdsourcing

Jonathan Bragg
UW
(Stanford)

Mausam
IIT Delhi

Dan Weld
UW

NSF
ONR
WRF
Google
IBM
Bloomberg
Government of India

UW CrowdLab

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING