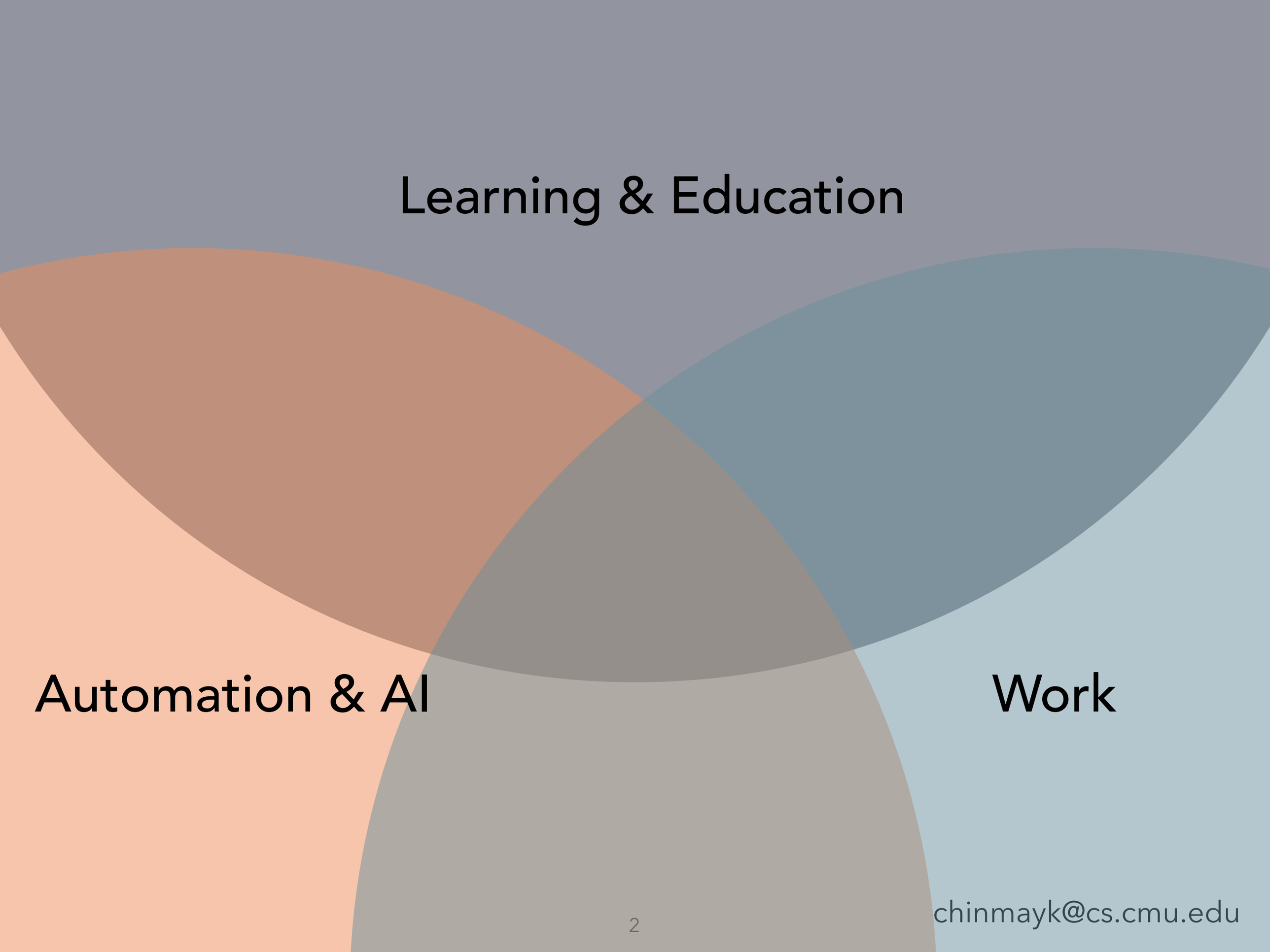
Human-machine partnerships for accelerating learning in the wild

Carnegie Mellon University

chinmayk@cs.cmu.edu



chinmayk@cs.cmu.edu

People

Learning & Education

My research

Automation & AI

Work

chinmayk@cs.cmu.edu

People

MOOCs
Peer assessment
Global discussions

My research

Impartial algorithms
Algorithmic framing
eloquent.ai

Labor markets
Crowdsourcing

chinmayk@cs.cmu.edu

Human-machine partnerships for accelerating learning in the wild

chinmayk@cs.cmu.edu

Online freelancer markets offer opportunities to hire experts for short tasks

software, writing, design, voiceover,
translations.... more than
$360 million of work every year

. A. Agrawal, J. Horton, N. Lacetera, and E. Lyons. Digitization and the contract labor market: A research agenda. Technical report, National Bureau of Economic Research, 2013.

Employer challenges in hiring in online labor markets

- Hiring in a domain needs domain expertise

  ! How does a hairdresser hire a web-designer?

  ! How do you hire a translator in a language you don't speak?

- Hiring manually is laborious and time-consuming

  ! Hiring on Upwork takes 3 days (i.e. longer than many tasks)

Horton & Golden. Reputation inflation: Evidence from an online labor market. 2015.

Worker challenges in hiring in online labor markets

- Hiring in a domain needs reputation in domain

  ! But you have no reputation when starting out

  ! Reputations are inflated (91% of contracts on Upwork had *perfect* feedback scores in 2014)

- No visibility into what skills are needed to succeed

  ! Freelancers not hired get little feedback

. A. Pallais. Inefficient hiring in entry-level labor markets. The American Economic Review, 104(11): 3565–3599, 2014.

Hiring uncertainties lead to a downward spiral

If employers can't find good workers quickly, they either hire less, or offer lower wages, which further discourages qualified workers

M. Silberman, J. Ross, L. Irani, and B. Tomlinson. Sellers' problems in human computation markets. In Proceedings of the acm SigKDD workshop on human computation, pages 18–21. ACM, 2010.

Photo by Ludde Lorentz on Unsplash

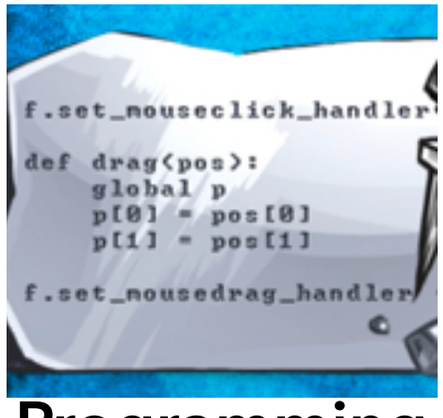
What if we could...

- Develop a method for hiring experts (workers) in online outsourcing markets that requires:
  - no involvement from employers
  - does not rely on reputations

Past work: Large scale peer assessment for MOOCs

Human-computer Interaction

Design



Programming in Python

Code

Introduction to Philosophy

Essays

Teaching character

Management

Child Nutrition

Recipes

Social Psychology

Essays

Constitutional law

Arguments

World Music

Music

Approach: peer assessed hiring of experts

Requester submits crowd expert request

This should be impartial
(workers shouldn’t be able to manipulate their rank)

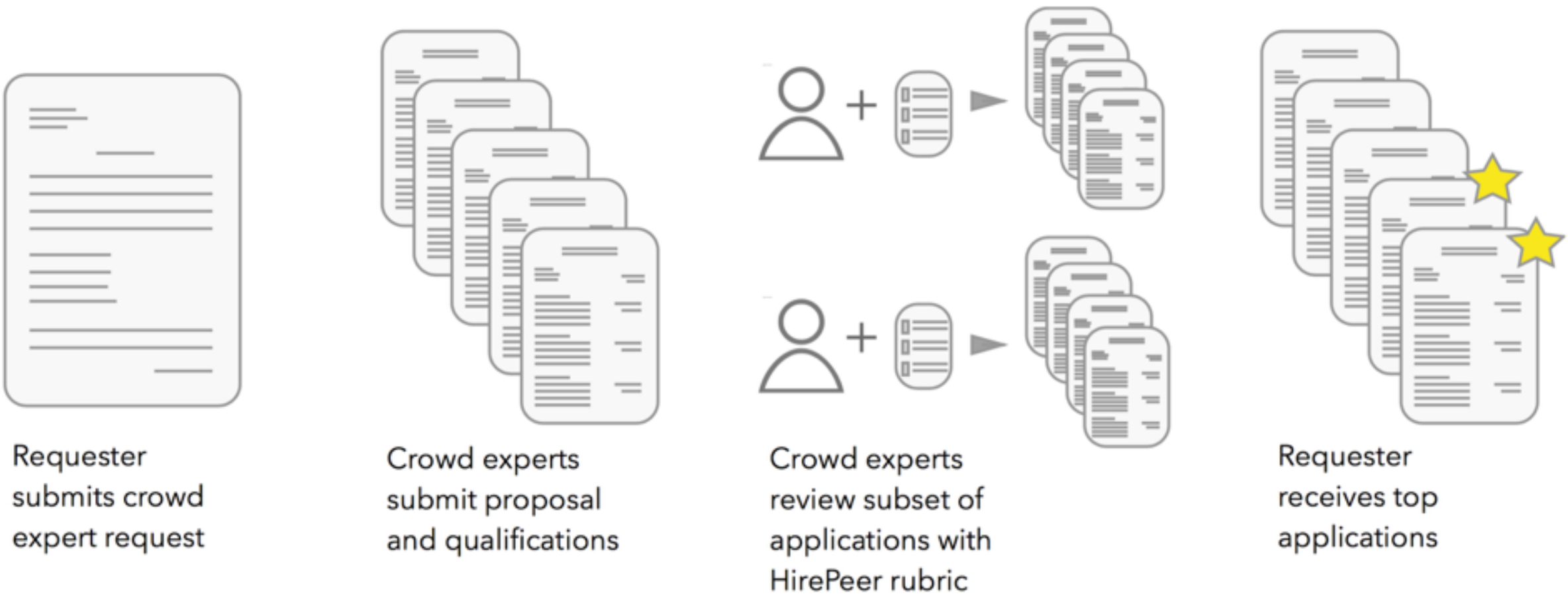
Past work: key results

- Peers in classroom can assess similarly accurately as course staff (Kulkarni et al., 2013)

- Peers can provide *rapid* feedback and assessment (median < 7 minutes for N > 200) (Kulkarni et al., 2015)

- Open question: can peers assess each other in cases of direct conflict of interest; i.e. in hiring?

Approach: peer assessed hiring of experts



Requester submits crowd expert request

Crowd experts submit proposal and qualifications

Crowd experts review subset of applications with HirePeer rubric

Requester receives top applications

This should be impartial (workers shouldn't be able to manipulate their rank)

Impartiality (our defn.)

- No player *i* can affect his probability of being ranked in position j, for all *i, j* ∈ [n].

- Equivalently, we may assume that each player *i* has a value v_{ij} for being ranked in position *j;* then impartiality would mean no player can affect his expected value for the outcome

- This is not the same as *societal* definition

A naive impartial algorithm

1. Choose a candidate C at random, and assign them the last position.

2. Use C's ranking as the final ranking for everyone else

… Fair, but bad if you're C.

We can do better!

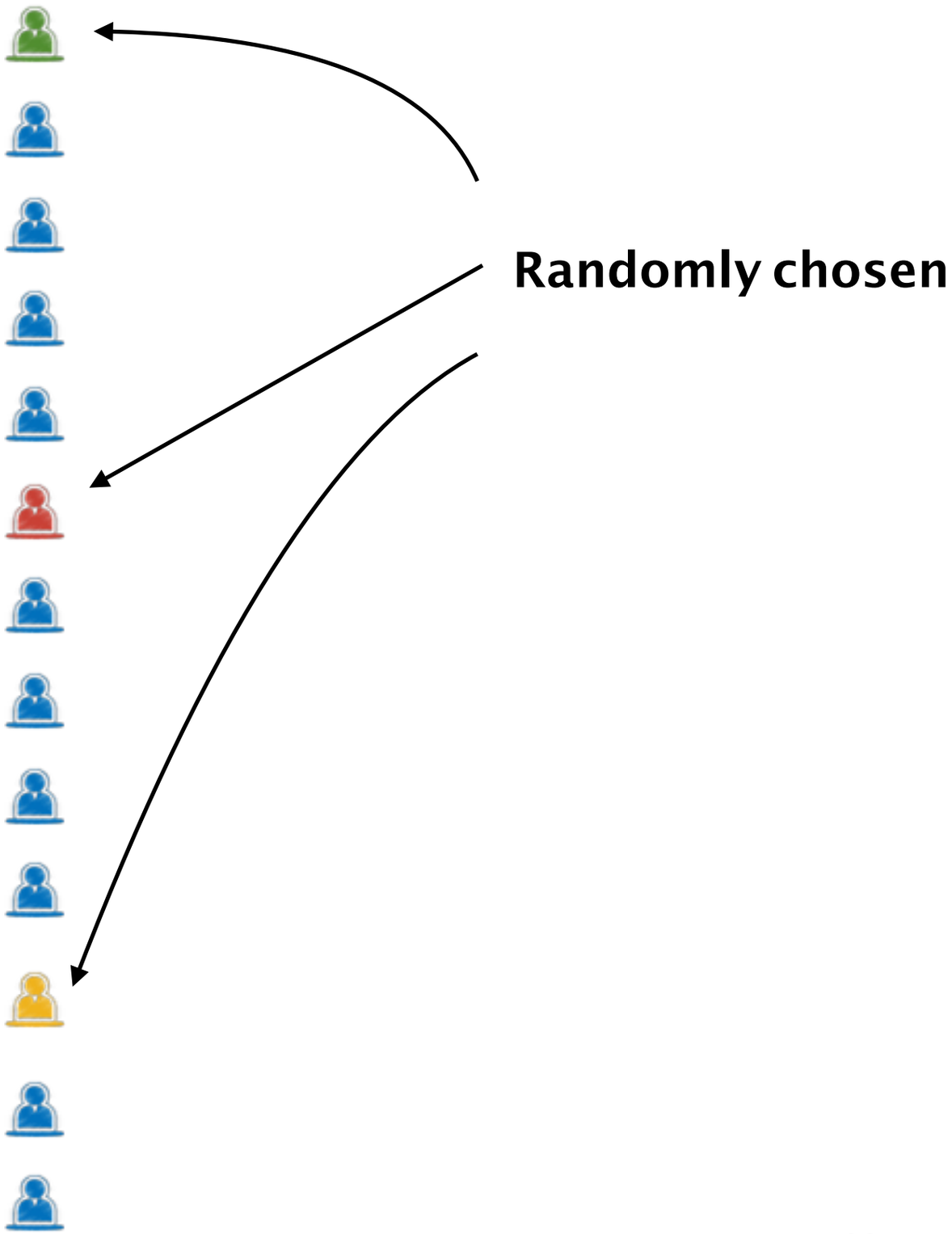
Our contribution: Two impartial algorithms with bounded error guarantees

- COMMITTEE
- K-PARTITE

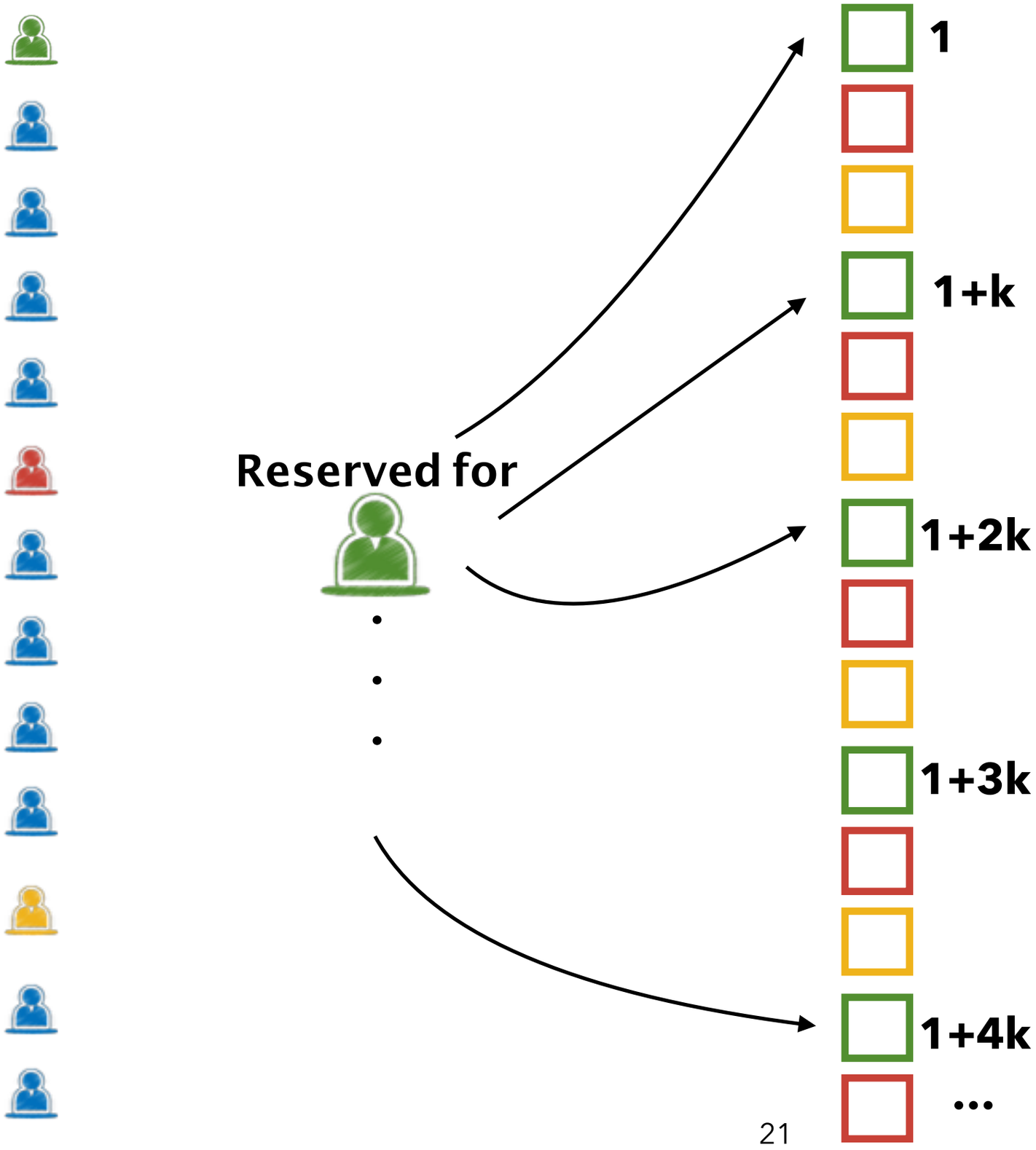
See Kahng, A., Kotturi, Y., Kulkarni, C., Kurokawa, D., & Procaccia, A. D. *(2017). Ranking wily people who rank each other.* AAAI 2017.

COMMITTEE algorithm

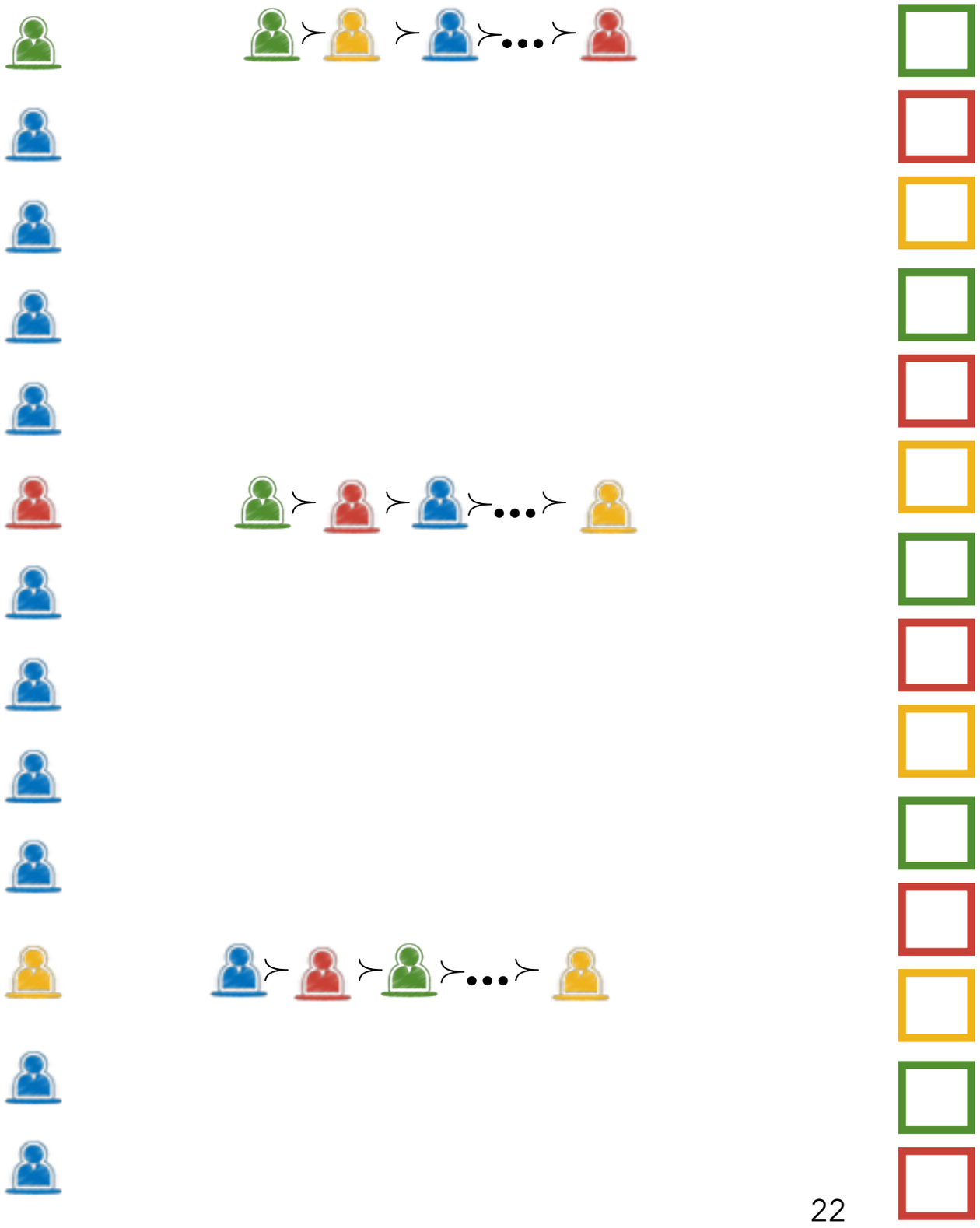
COMMITTEE algorithm



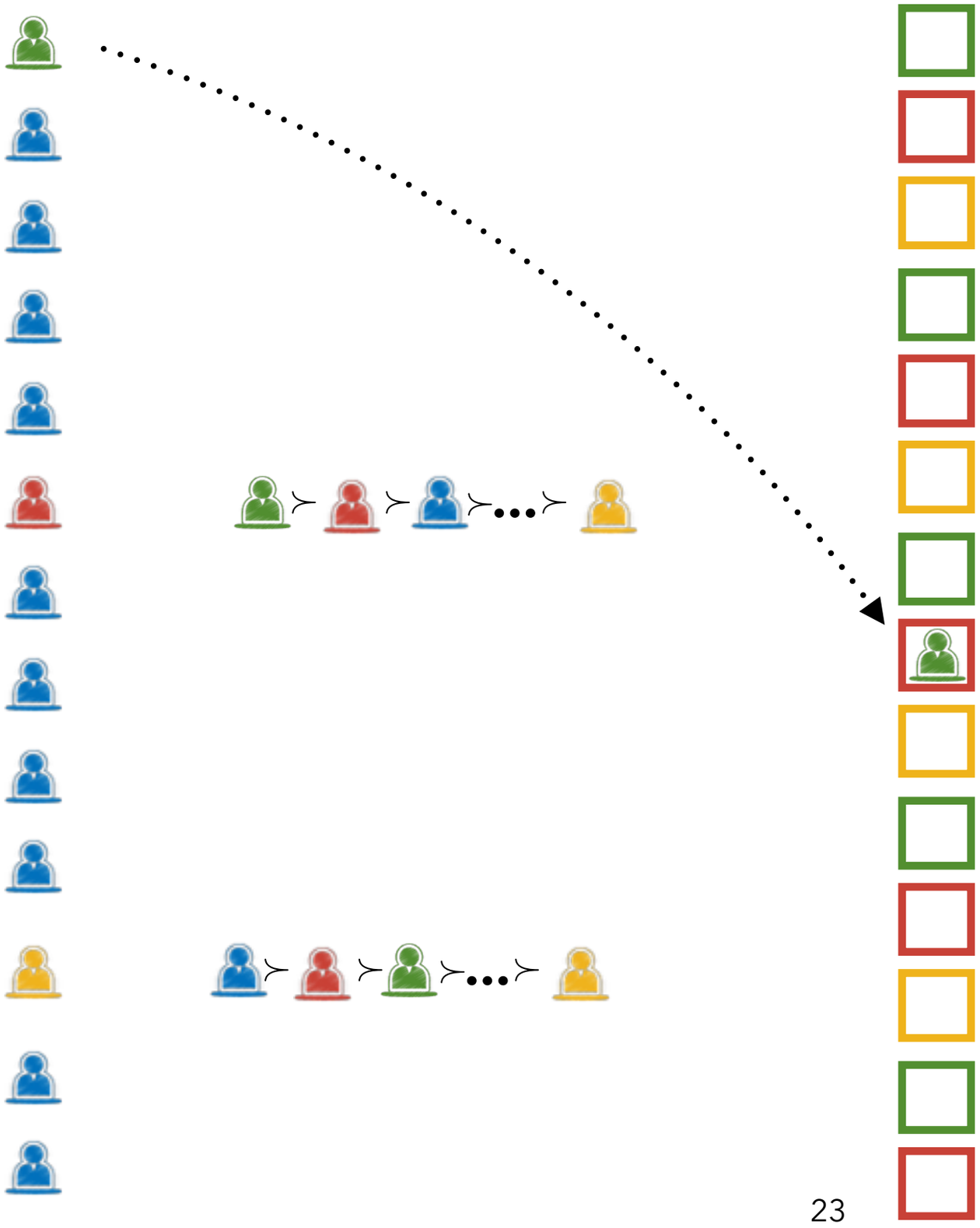
COMMITTEE algorithm



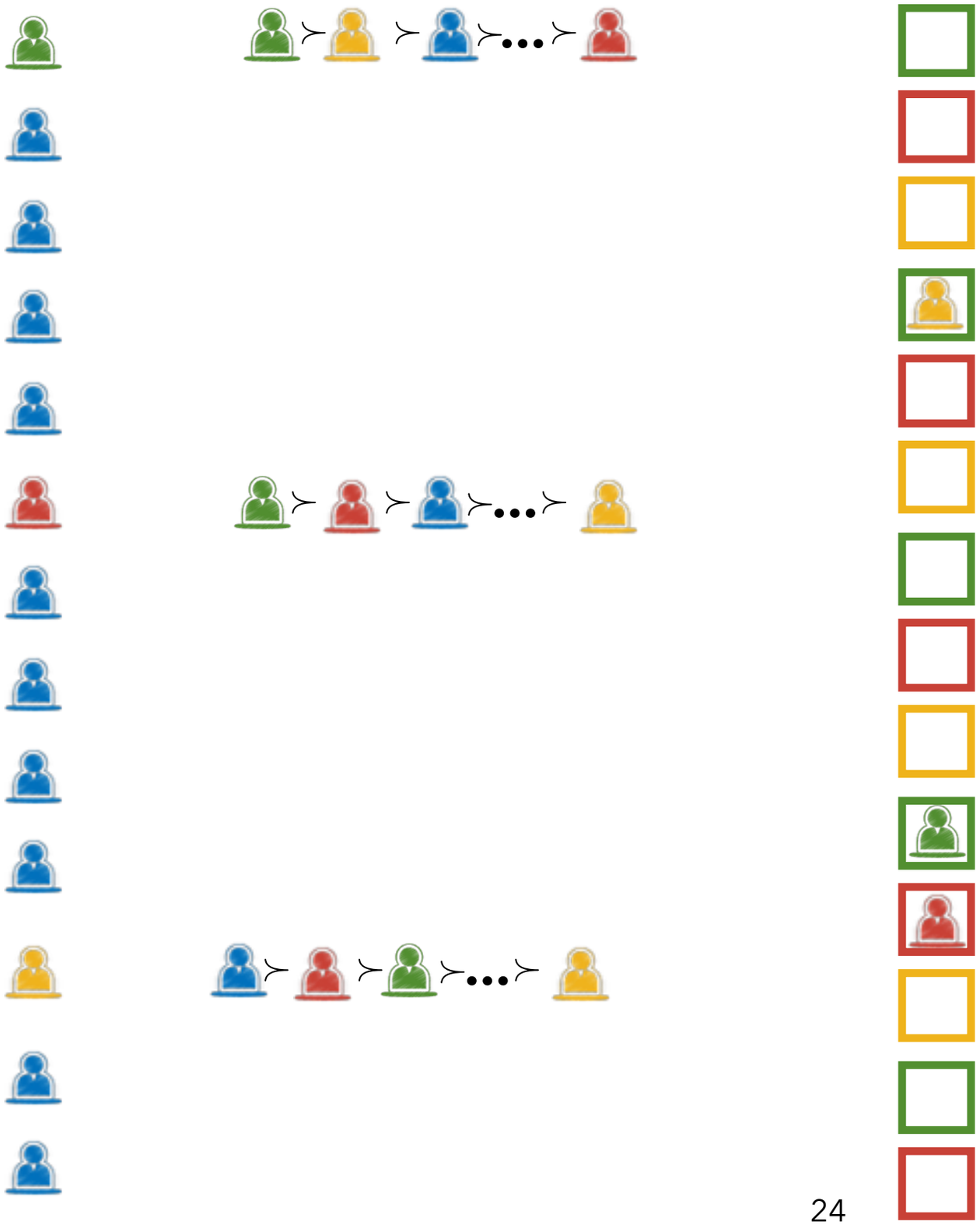
Committee algorithm



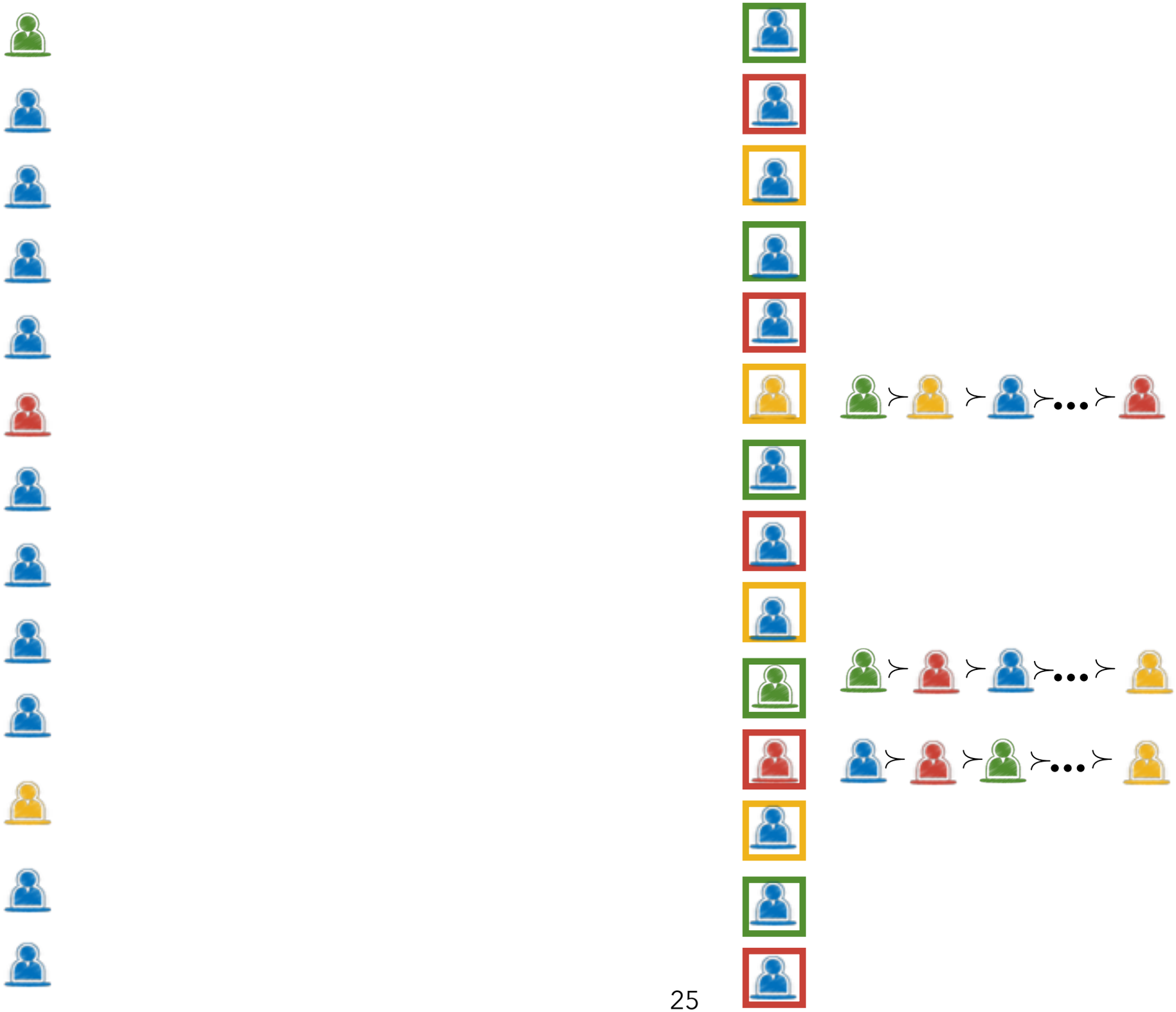
Committee algorithm



COMMITTEE algorithm



COMMITTEE algorithm



Committee algorithm

- Obviously impartial
- Has boundedly low *backward error* and *forward error*
  - i.e. every player *i* is placed in a rank that had the players altered their opinions slightly, *i* would be in the correct rank; with high probability (low backward error)
  - but *i* may not *exactly* maintain same position given a small perturbation in input rankings (non-zero forward error)

k-Partite Algorithm

- Split applicants into partitions, rank using input from other partitions only
- No forward error, but larger backward error
- Details in paper

Theorem 4.1. *k*-PARTITE *is impartial, and, for every* $f \in \mathcal{P}$ *and* $\vec{\sigma} \in \Pi^n$, *if* $k = \lfloor (n/\ln n)^{1/3} \rfloor$, *it gives at most*

$$(4/k, 4/k) \in \left(O\left(\left(\frac{\ln n}{n} \right)^{1/3} \right), O\left(\left(\frac{\ln n}{n} \right)^{1/3} \right) \right)$$

backward error with respect to f.

Theorem 5.1. COMMITTEE *is impartial, and, for every* $f \in \mathcal{P}$, $\vec{\sigma} \in \Pi^n$, *and* $\varepsilon > 0$, *if*

$$k = 1 + \frac{2}{\varepsilon^2} \ln \left(\frac{n^3}{\varepsilon} \right),$$

it gives at most $(\varepsilon, \varepsilon, (k+1)/n)$ *mixed error with respect to* f.

How do you communicate an impartial algorithm?

- Complex algorithms are increasingly ubiquitous, but remain hard to explain

- “Interpretable algorithms” struggle to articulate consistent metrics of success [Lipton 2016]

  - Interpretability doesn’t automatically lead to trust or acceptance

- In our experience, it is more useful instead to use framing effects to communicate algorithms

chinmayk@cs.cmu.edu

Which instructions lead to least exaggeration?

1. **Control**: “Be sure to read the instructions carefully.”

2. **Self-concept maintenance**: “For your protection, we prevent others from cheating using an impartial algorithm.” (Greenwald 1980; Griffin and Ross 1991; Sanitioso, Kunda, and Fong 1990)

3. **Policing**: “To prevent you from cheating, we’ve implemented an impartial algorithm” (Allingham and Sandmo,1972)

4. **Behavior**: “Your own ranking will not affect the final ranking of your item as we use an impartial algorithm.” (Mazar, Amir and Ariely, 2008)

chinmayk@cs.cmu.edu

Experiment: How to communicate impartial algorithms?

N=170, Amazon Mechanical Turk (worldwide)

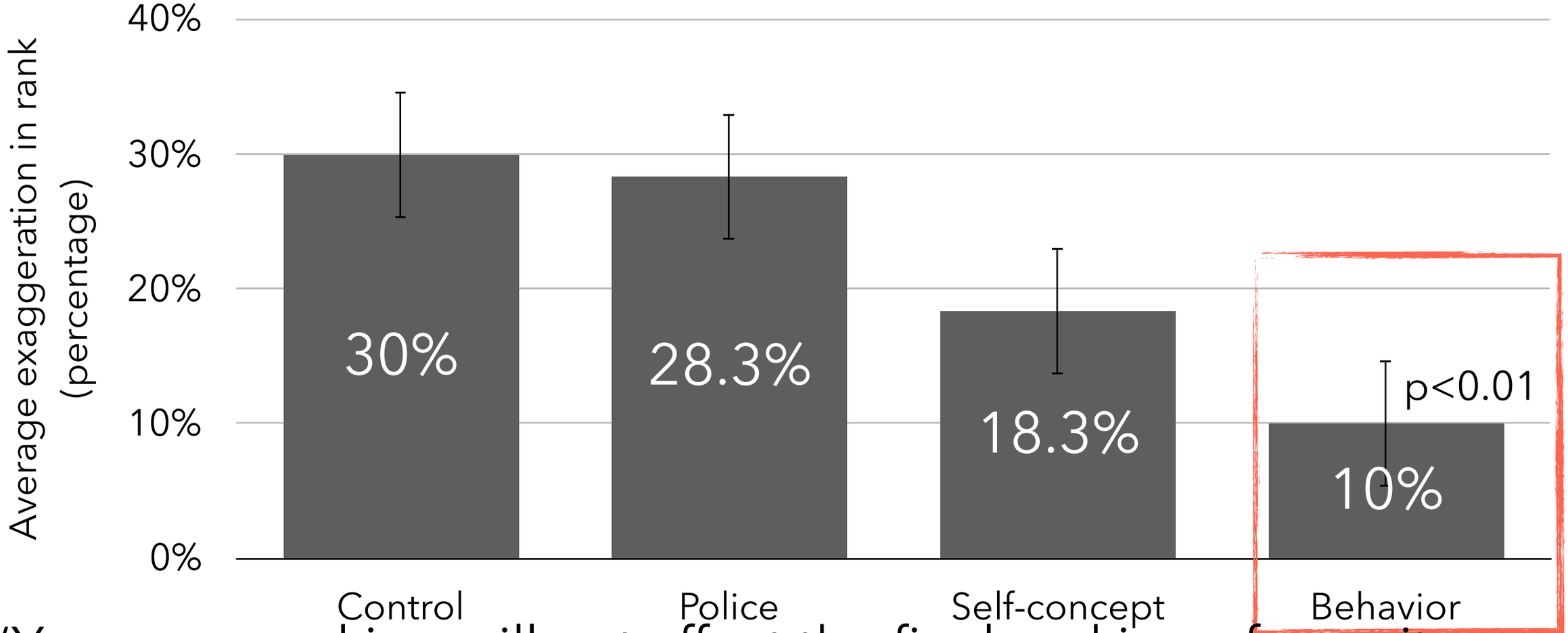
Task: fix typos in a product review; rank 12 related reviews. Bonus if your review is in top k=5.

Everyone gets same review. Independent judges rank = #6.

Results

N=170, Amazon Mechanical Turk (worldwide)

Exaggeration = (Actual Rank - Reported Rank)/#items



"Your own ranking will not affect the final ranking of your item as we use an impartial algorithm."

chinmayk@cs.cmu.edu

Story 2: effects of impartiality on behavior

Requester submits crowd expert request

Crowd experts submit proposal and qualifications

Crowd experts review subset of applications with HirePeer rubric

Requester receives top applications

This should be impartial
(workers shouldn't be able to manipulate their rank)

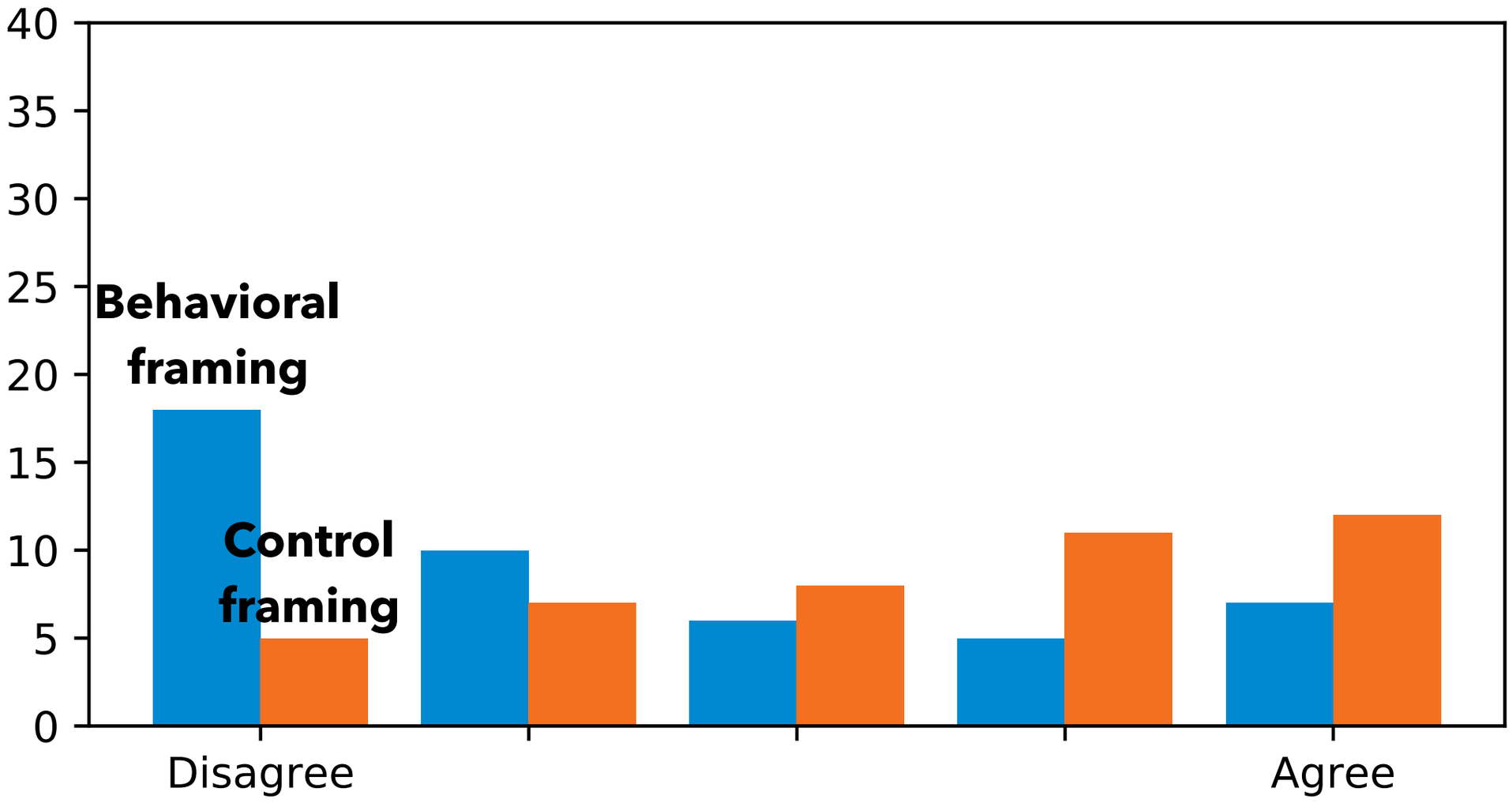
All results nearly same, but impartial versions do slightly worse

Aggregation \ Framing	Behavioral	Control
Partial	0.9566	0.9665
Impartial (NAIVE-BIPARTITE)	0.8884	0.9259
Impartial (COMMITTEE)	0.8044	0.8375
Impartial (k-PARTITE)	0.7831	0.8092

Table 2: Average accuracy for each condition and aggregation scheme. In the Study 2 setup, impartial aggregation reduced accuracy by 7% (NAIVE-BIPARTITE). An impartial framing also likely reduced quality of rating data slightly (Control better than Behavioral framing by 1%.

Impartial aggregation might reduce effort

"My effort in rating others affects my final ranking"



Despite these comments… some workers saw benefits to comparison

> “It’s great to see 2 submissions side by side, because you have a **point of reference** to grade. Also, it’s a learning stimulation, as you tend to jump quickly from submission A to B and vice versa to **identify the good/best answer…**”

Others were not so sure…

> Comparative assessment **cannot ever lead to fair and accurate assessment** of another's work. It's very much like asking someone if they prefer fried or boiled potatoes. Very much a matter of **personal taste**.

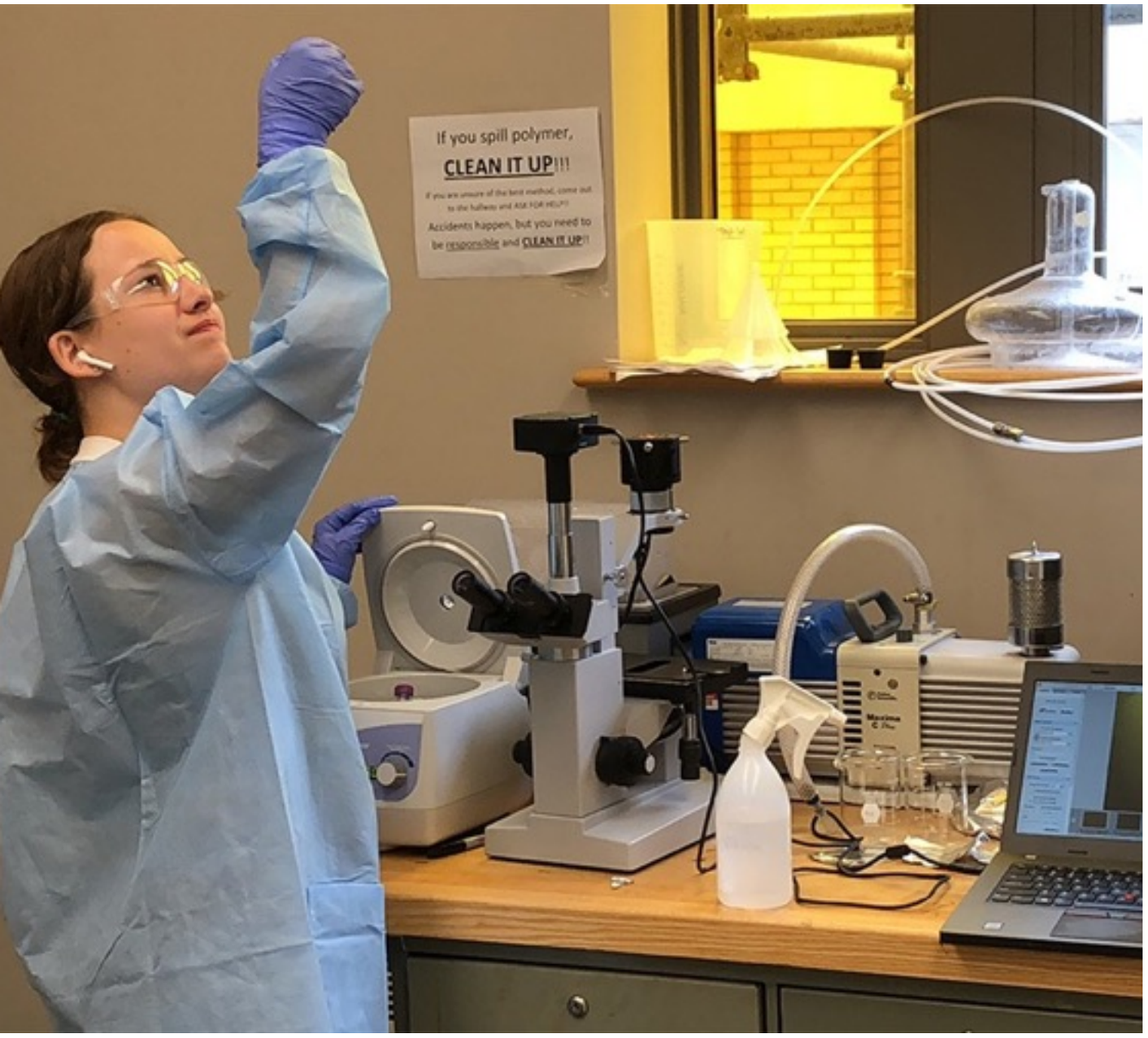
> I didn't really like the fact that we need to compare works by two people; it's **not always possible to choose**. Sometimes, the two works are **equally good, or bad**.

Story 3: Building trust in automation

Could you trust a voice assistant to help you with longer and more complex tasks? Like…

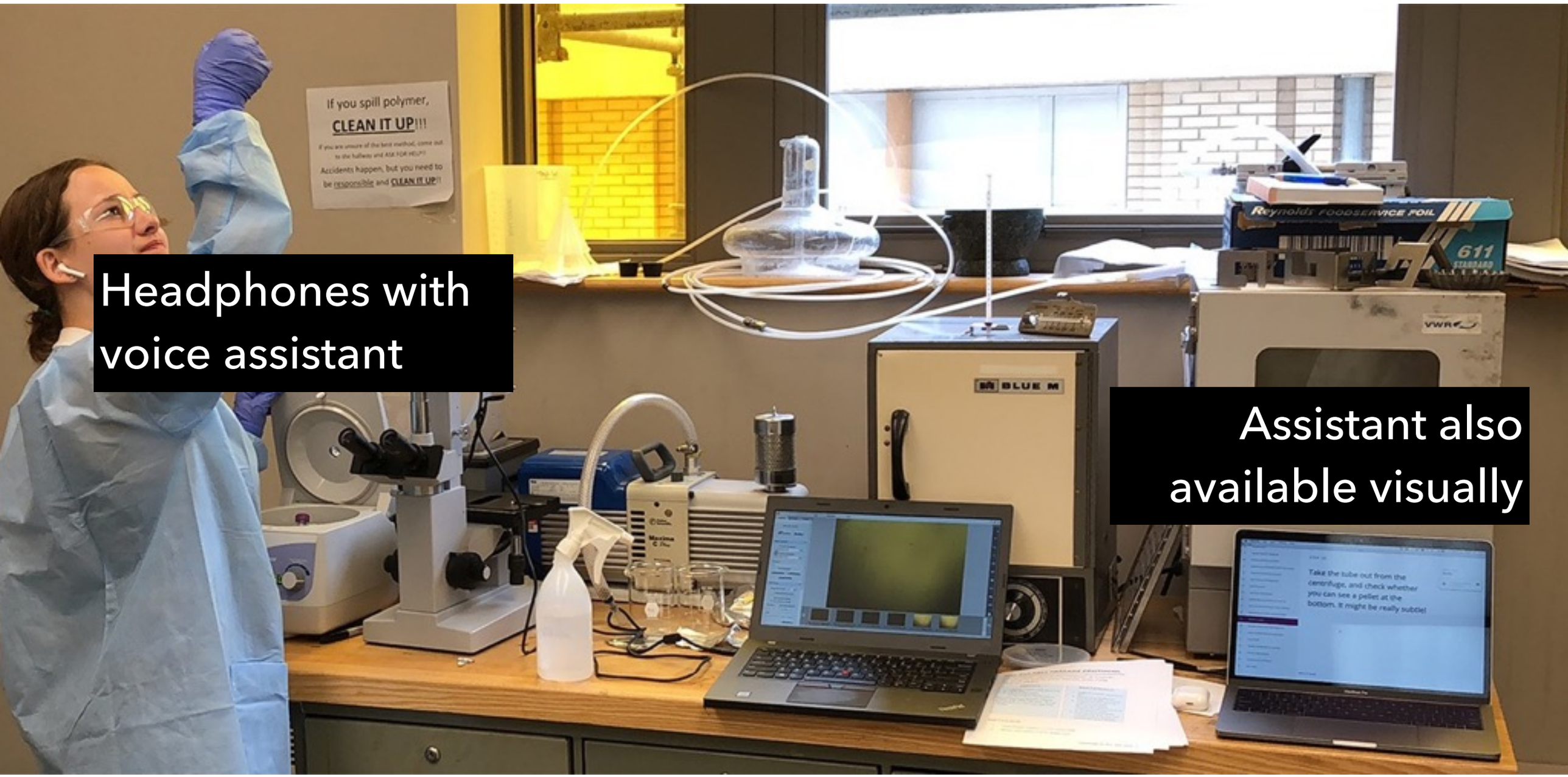
- Structuring an argument/essay?
- Asking your manager for a raise?
- Helping biologists culture stem-cells more effectively? (And pave the way for personalized medicine)

chinmayk@cs.cmu.edu


If you spill polymer,
CLEAN IT UP!!!
Accidents happen, but you need to be responsible and CLEAN IT UP!!


"Vitro, help me culture my fibroblast cells"
amazon
all-new
echo
$99.99

Using voice to guide cell culture in BSL-1 lab

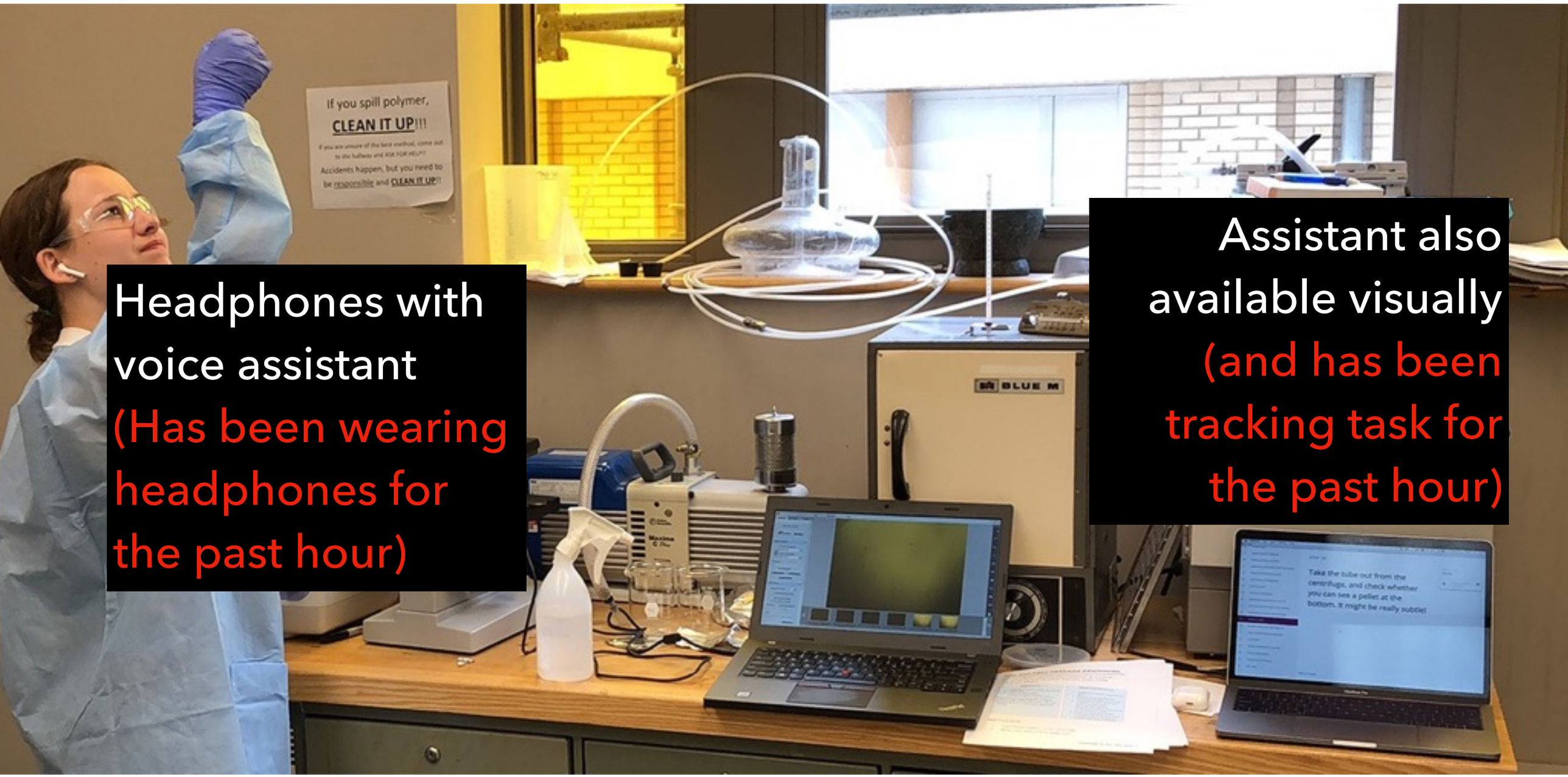


Headphones with voice assistant

Assistant also available visually

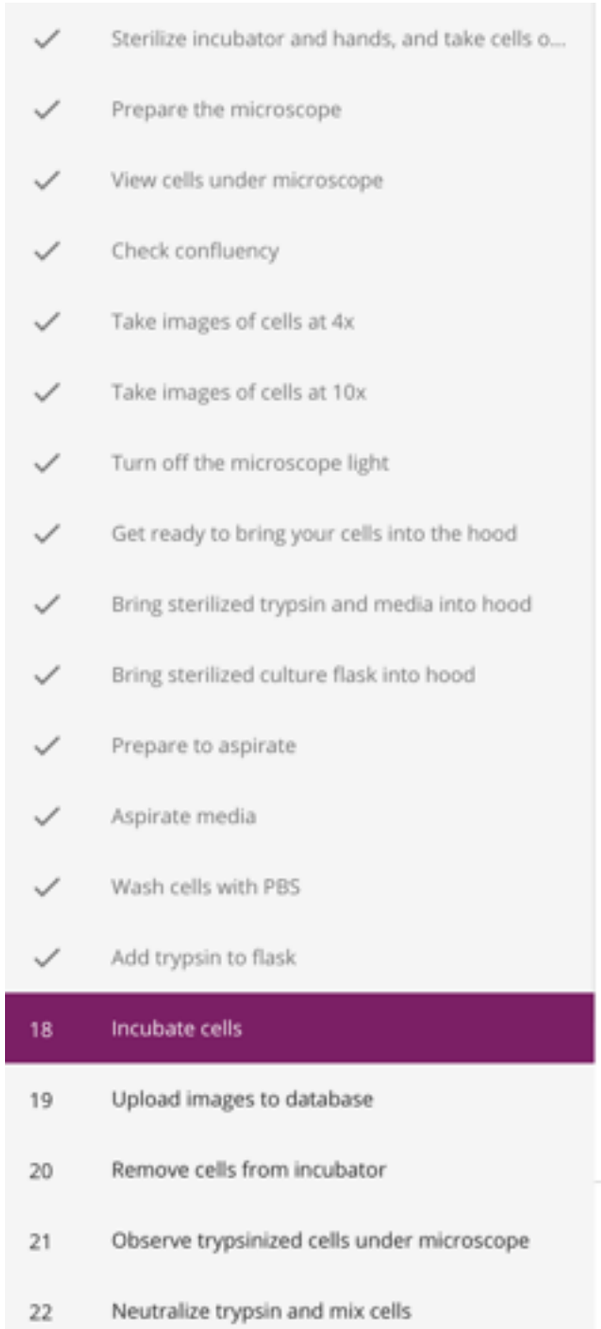
chinmayk@cs.cmu.edu

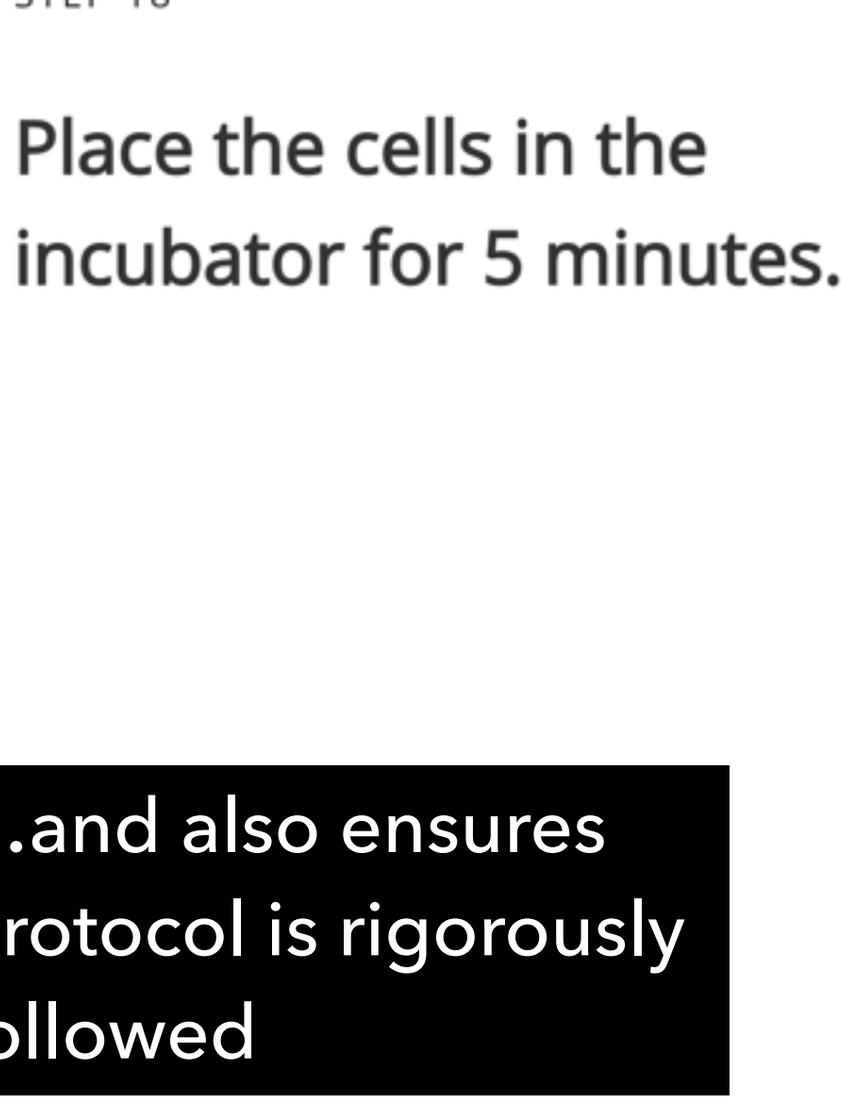
Continuous and long term task support



chinmayk@cs.cmu.edu

Task assistance in a 30+ step protocol





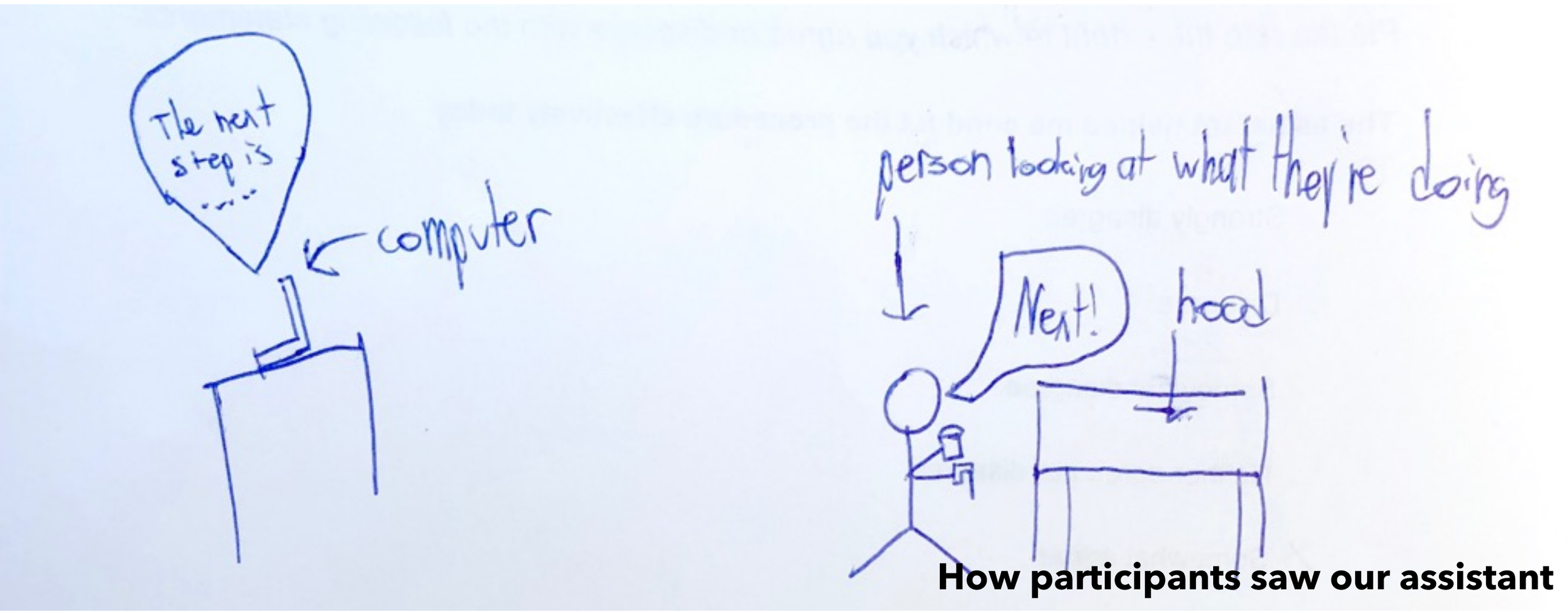
chinmayk@cs.cmu.edu

Design implications from assisting long, complex tasks

- Tasks are *long!* (30 min+), and “Hey Siri!” becomes tiring
  - So our assistant is always listening
- Participants had eyes and hands busy, only looked at the screen when something went wrong
  - So we use the screen for non-focal tasks
- Participants saw themselves as experts
  - Assistant was designed as a coach, not a supervisor

chinmayk@cs.cmu.edu

Our solution: Make assistant seem limited and admit mistakes



How participants saw our assistant

- Participants didn't think our assistant was very smart.
- But they uniformly believed they were in control, not the assistant

chinmayk@cs.cmu.edu

A simple model builds trust

- “I feel like in terms of Alexa and Siri, they just have a lot stored in them, whether through the cloud or through their hardware. Whereas I wasn't exactly sure how much was stored in here, and **it's probably just a very small percentage of what something like [Alexa and Siri] would have the capacity for**.” (P1)

chinmayk@cs.cmu.edu

...But limited functionality still needs to be reliable

“ Like when I first walked in I was just like, ``well, I don't know if I can trust it.'' ... Only after a couple of steps I realized that [the spoken protocol] was fairly concrete, it was fairly descriptive that then I could let my guard down.” (P5)

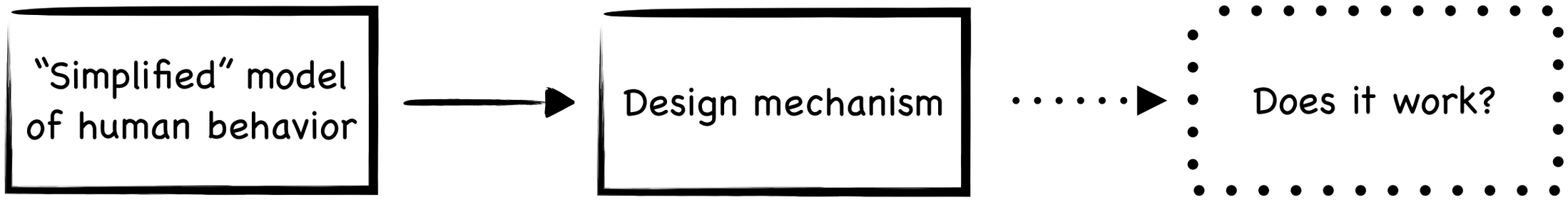
chinmayk@cs.cmu.edu

Some takeaways

- Automation will increasingly mediate work and coordination
    - The *same* algorithm has different effects depending on how it's communicated
    - Automation has emergent effects that are non-obvious, but might be more important than the designed effects
- Could automation as a coach be a metaphor in other human-coordination tasks?

...and one more thing!

Traditional mechanism design



A more HCI-inspired mechanism design?