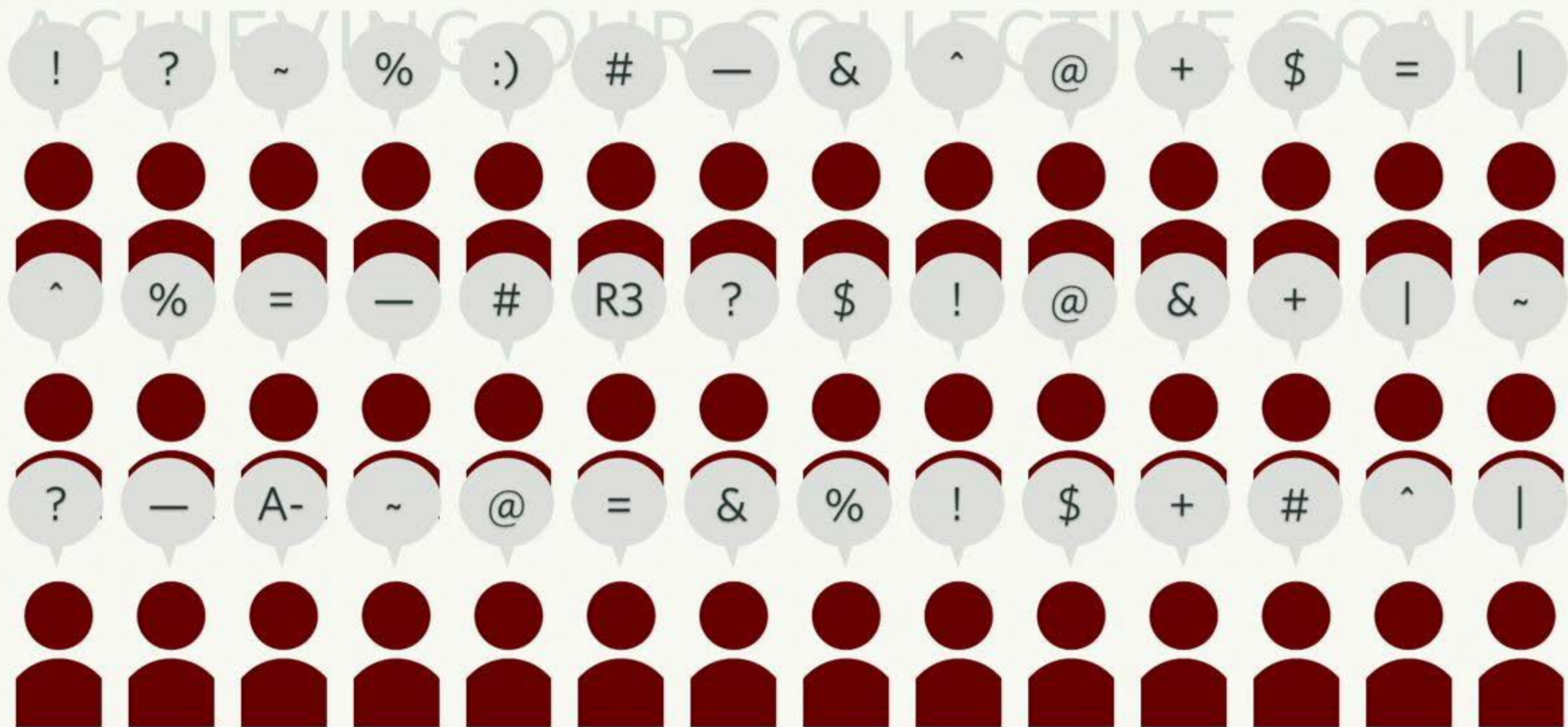


Crowds, Computation, and the Future of Work

Michael Bernstein
Stanford University

ACHIEVING OUR COLLECTIVE GOALS





COORDINATION NEGLECT: HOW LAY THEORIES OF ORGANIZING COMPLICATE COORDINATION IN ORGANIZATIONS

Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams

The Mutual Knowledge Problem and Its Consequences for Dispersed Collaboration

The team scaling fallacy: Underestimating the declining efficiency of larger teams

Who's in Charge Here? How Team Authority Structure Shapes Team Leadership

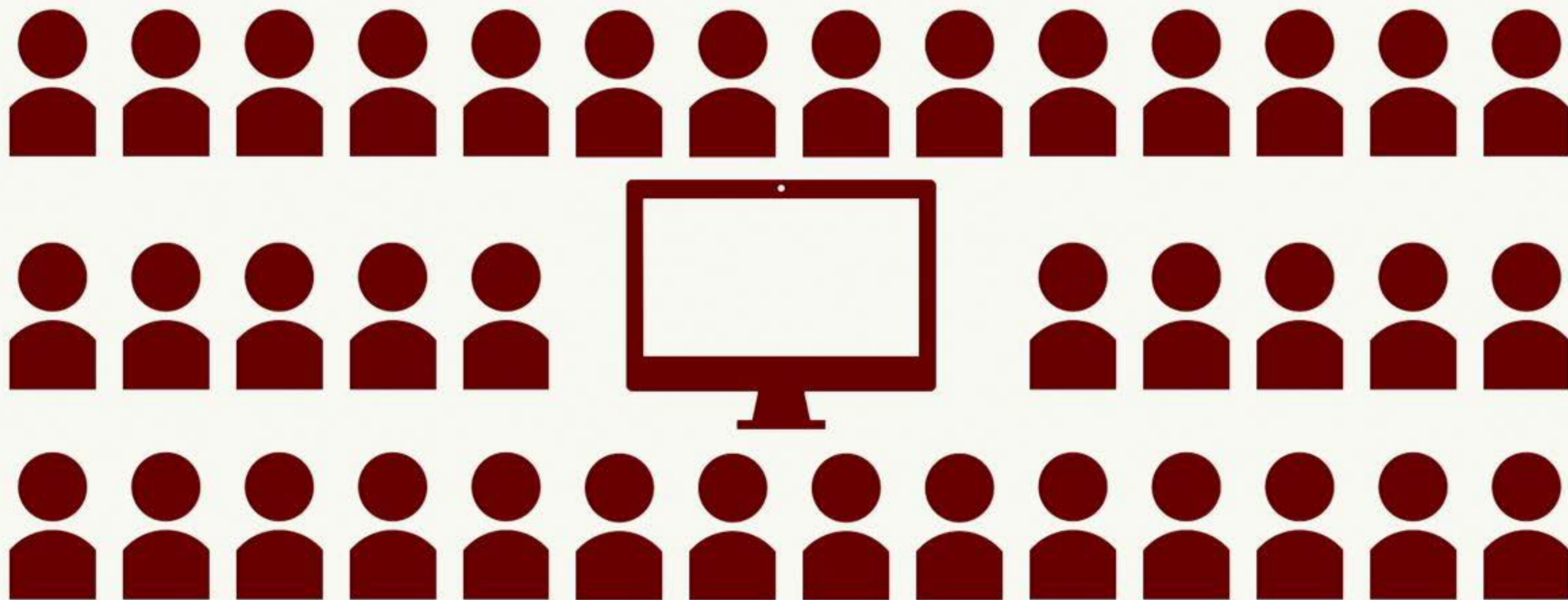
Team Familiarity, Role Experience, and Performance: Evidence from Indian Software Services

The Influence of Shared Mental Models on Team Process and Performance

Some unintended consequences of job design

Structure and Learning in Self-Managed Teams: Why "Bureaucratic" Teams Can Be Better Learners

HOW MIGHT COMPUTING AUGMENT US
IN ACHIEVING OUR COLLECTIVE GOALS?



WORKER COLLECTIVE ACTION

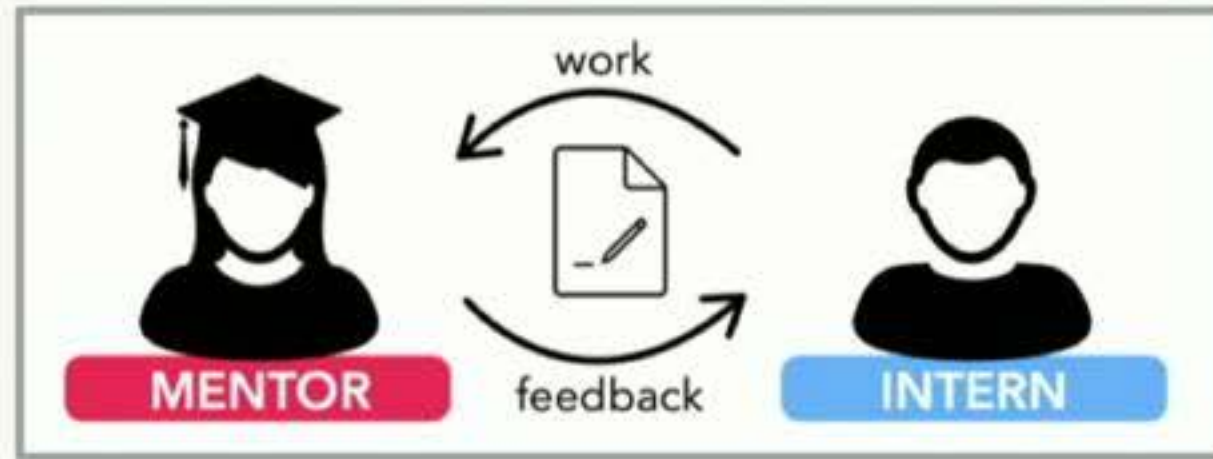
The screenshot shows the DYNAMO website interface. At the top left is the logo 'DYNAMO' with a lightning bolt icon. Navigation links include 'Home', 'Vote on new ideas!', 'How it works', and 'Forum'. A user profile 'light_dragonfly' is visible. The main heading is 'Powering change on MTurk' in a large, white, cursive font on a blue background. Below it, a sub-heading reads 'We are a community of 485 Turkers and growing...!'. The page is divided into two columns: 'Trending Campaign Ideas' on the left and 'Live Campaigns' on the right. The 'Live Campaigns' section features a campaign titled 'Dear Jeff Bezos' posted 4 months ago. The campaign text reads: 'We are writing to let you and the rest of the world know all about who we are. The intent is for you to see that Turkers are not only actual human beings, but people who deserve respect, fair treatment and open communication.' To the right of the text, it says 'This is the page that will host the letters you send to us and publicize the movement:'. The 'Trending Campaign Ideas' section shows '23 more upvotes to launch this campaign' and a blue button with an upvote arrow and the number '2', and a red button with a downvote arrow and the number '0'.

Socio-technical infrastructure for collective action amongst crowd workers

[Salehi et al. 2015]

The screenshot shows the top portion of a Guardian article. The Guardian logo is at the top, with the tagline 'Winner of the Pulitzer prize'. Below the logo is the URL 'Amazon.com'. The article title is 'Amazon's Mechanical Turk workers protest: 'I am a human being, not an algorithm''. Below the title is a sub-headline: 'A Christmas email campaign is asking Amazon's CEO Jeff Bezos to improve terms for workers providing cheap digital labour'.

FUTURE OF CROWD WORK



Micro-internships

[Suzuki et al. 2016]



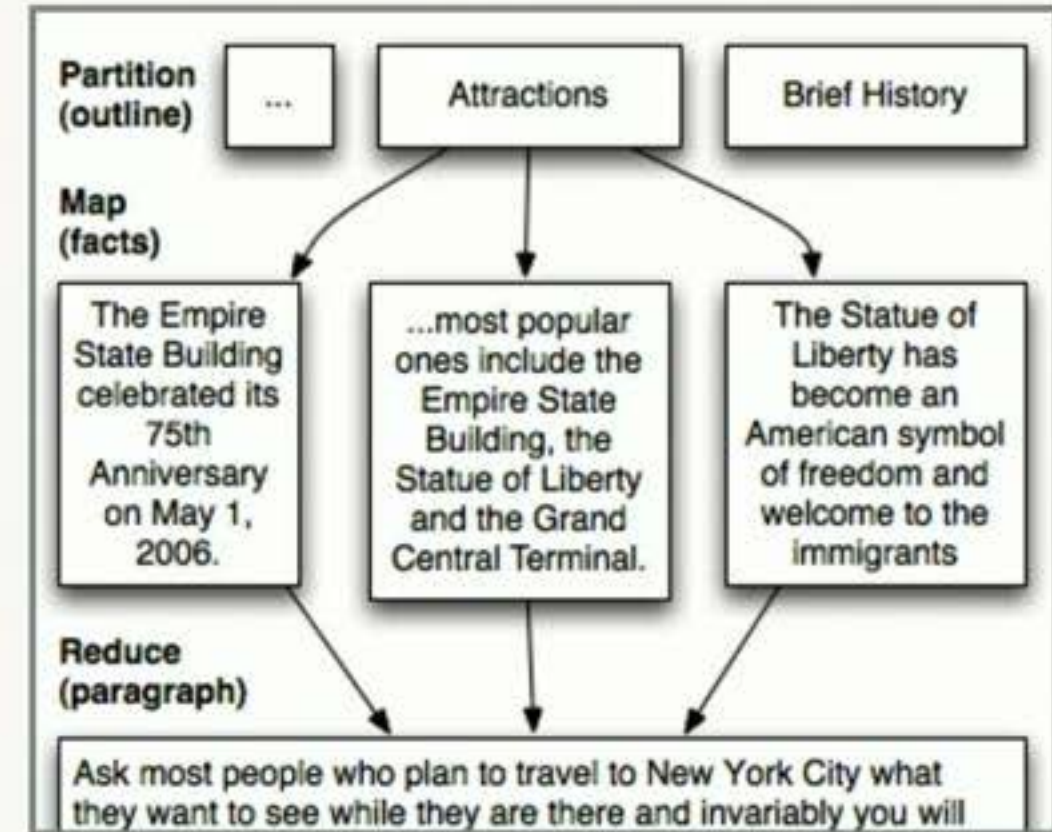
Guild-style
collective
accreditation

[Whiting et al. 2017]

DOMINANT ARCHITECTURE: ALGORITHMS

Modularize and pre-define all possible behaviors into workflows

Computation decides which behaviors are taken, when, and by whom; optimizes, error-checks, and combines submissions



[Kittur 2011]



[Little 2010; Dai, Weld 2010]

LIMITS OF ALGORITHMIC COORDINATION

So far, goals such as invention, production, and engineering have remained largely out of reach [Kittur et al. 2013]

LIMITS OF ALGORITHMIC COORDINATION

So far, goals such as invention, production, and engineering have remained largely out of reach [Kittur et al. 2013]

Reason: open-ended, complex goals are **fundamentally incompatible** with a requirement to modularize and pre-define every behavior [Van de Ven, Delbecq, and Koenig 1976; Rittel and Weber 1973; Schön 1984]

LIMITS OF OPEN SOURCE AND OPEN INNOVATION

“**Peer production is limited** not by the total cost or complexity of a project, but **by its modularity.**” [Benkler 2002]

“**With the Linux kernel [...]** we want to have a system which is as **modular as possible.** The open-source development model really requires this, because otherwise you can't easily have people working in parallel.” [Torvalds 1999]

HOW TOPCODER ATOMIZES PROJECTS INTO THEIR COMPONENTS



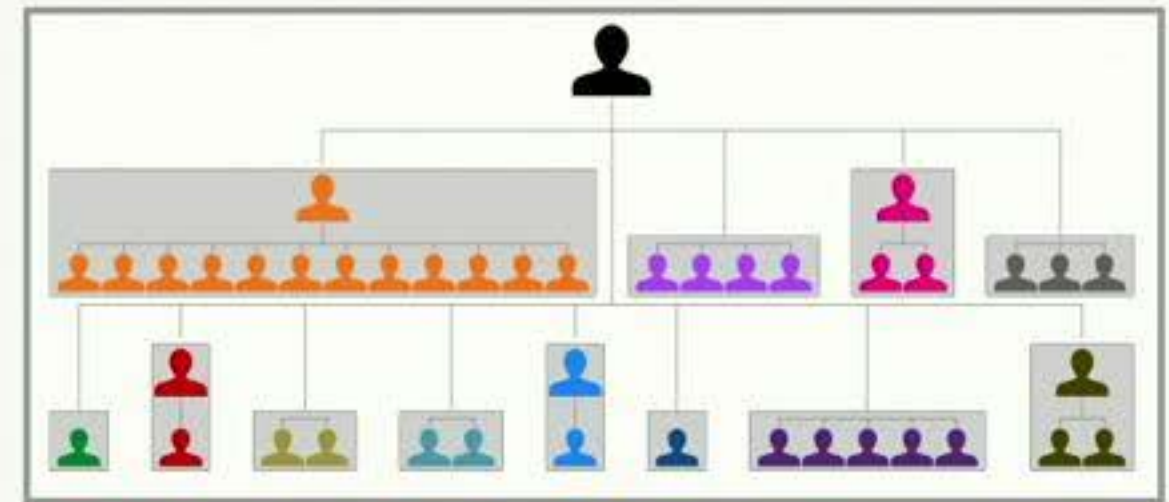
[Boudreau, Lacetera, and Lakhani 2011]

This architecture confines collaborations to goals so predictable that they can be entirely modularized and pre-defined

**An alternative architecture:
collaborations structured not
as algorithms, but as
computationally augmented
organizations**

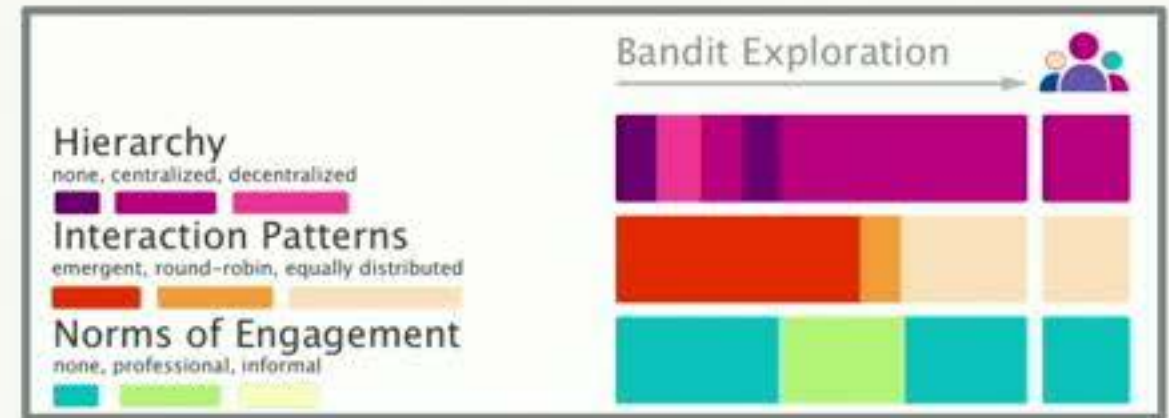
1) Flash Organizations

Create on-demand organizations capable of complex work



2) Dream Team

Find effective team structures



3) Crowd Research

Provide access to research opportunities to thousands around the world

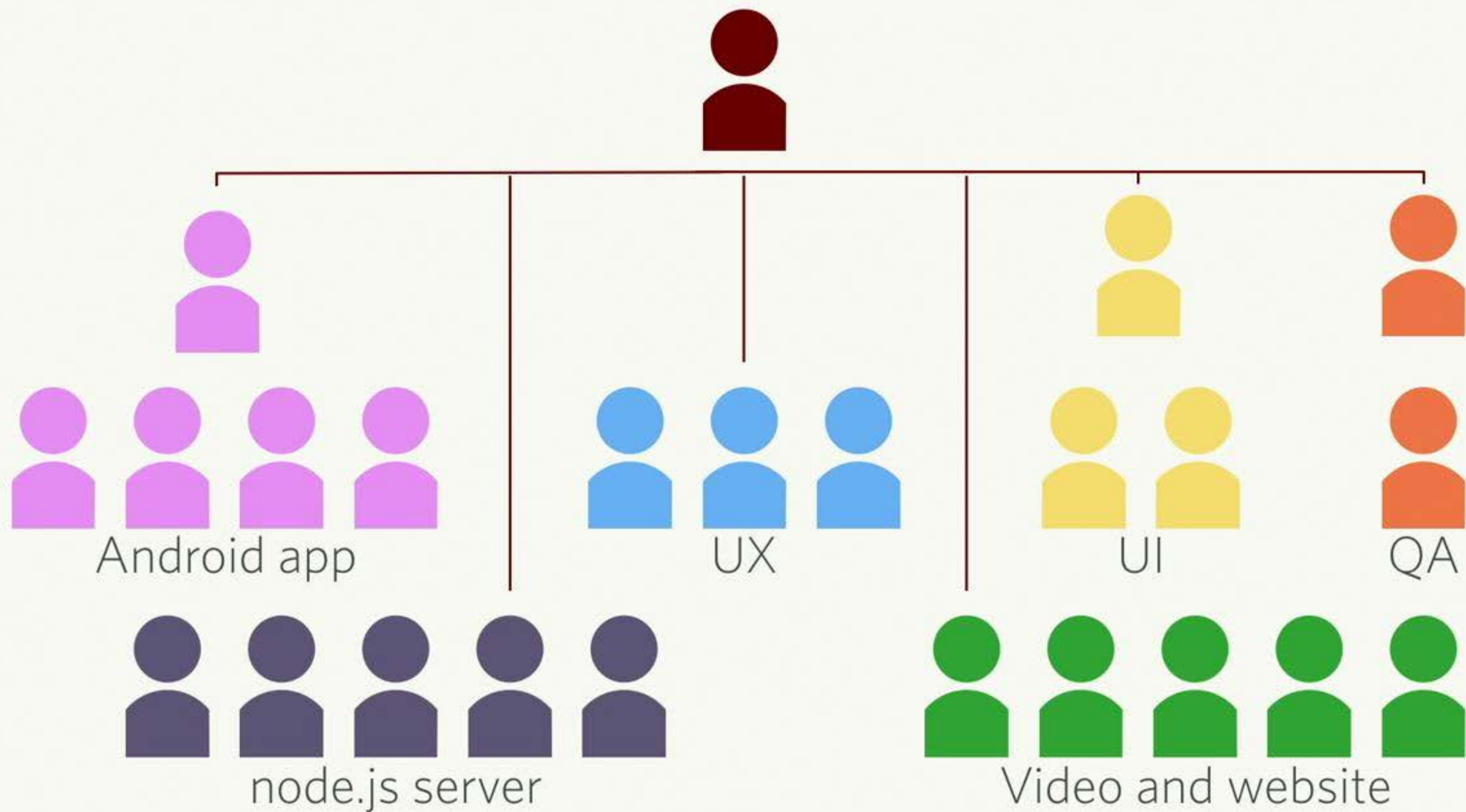


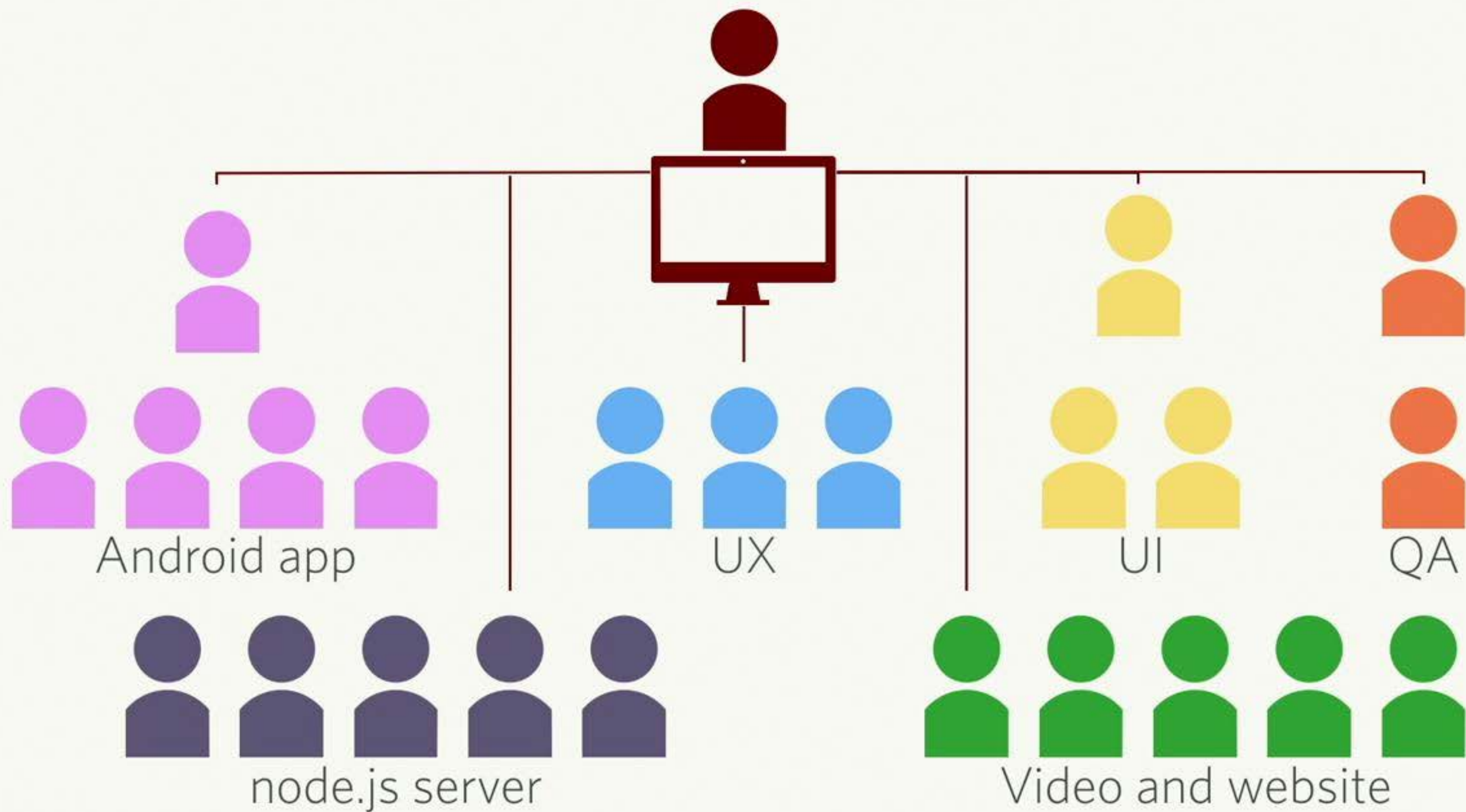
Flash organizations

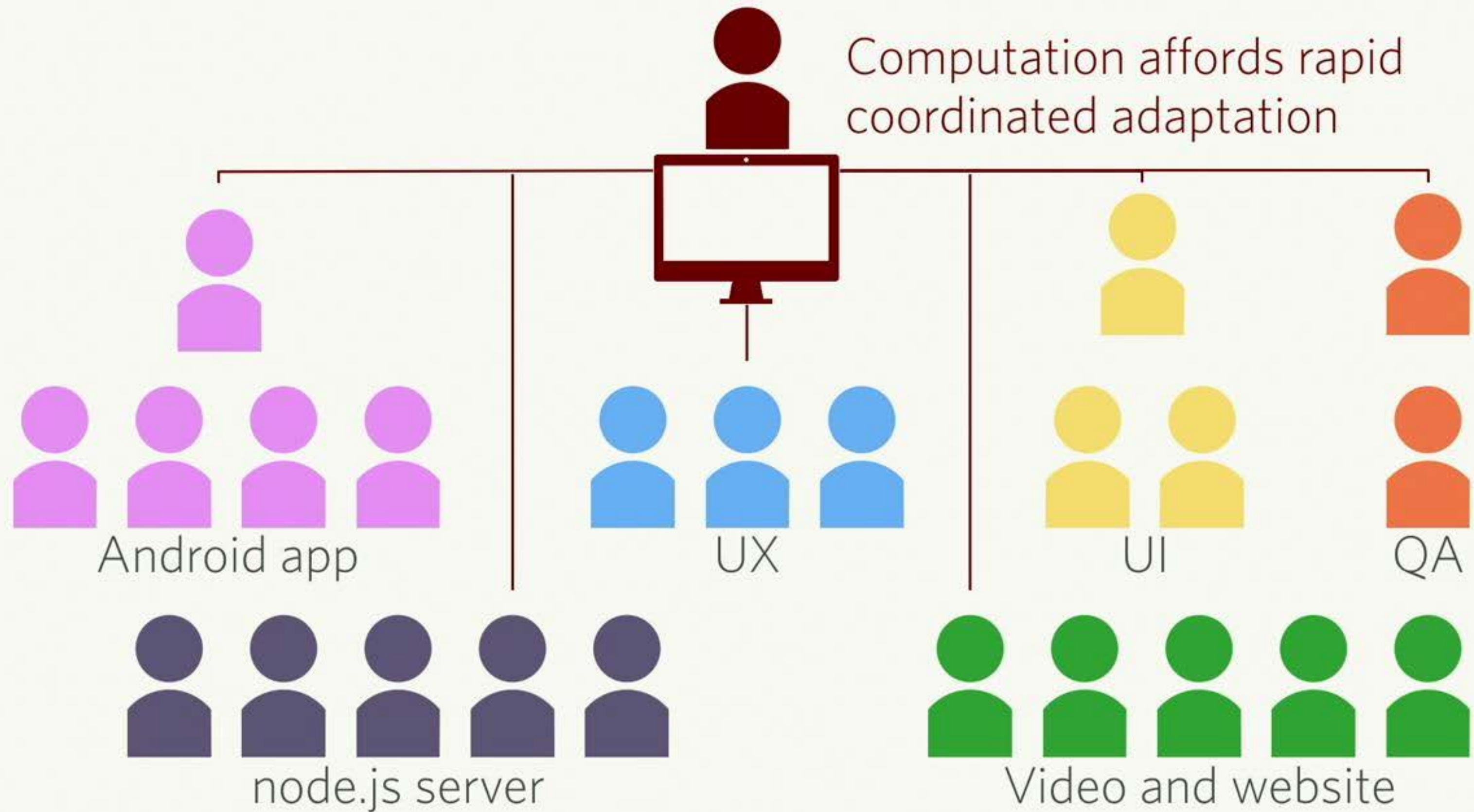
Valentine, Retelny, To, Rahmati, Doshi,
Bernstein. CHI 2017.

**Flash organizations:
rapidly assembled and
reconfigurable
organizations composed of
online collaborators**









Computation affords rapid coordinated adaptation

Android app

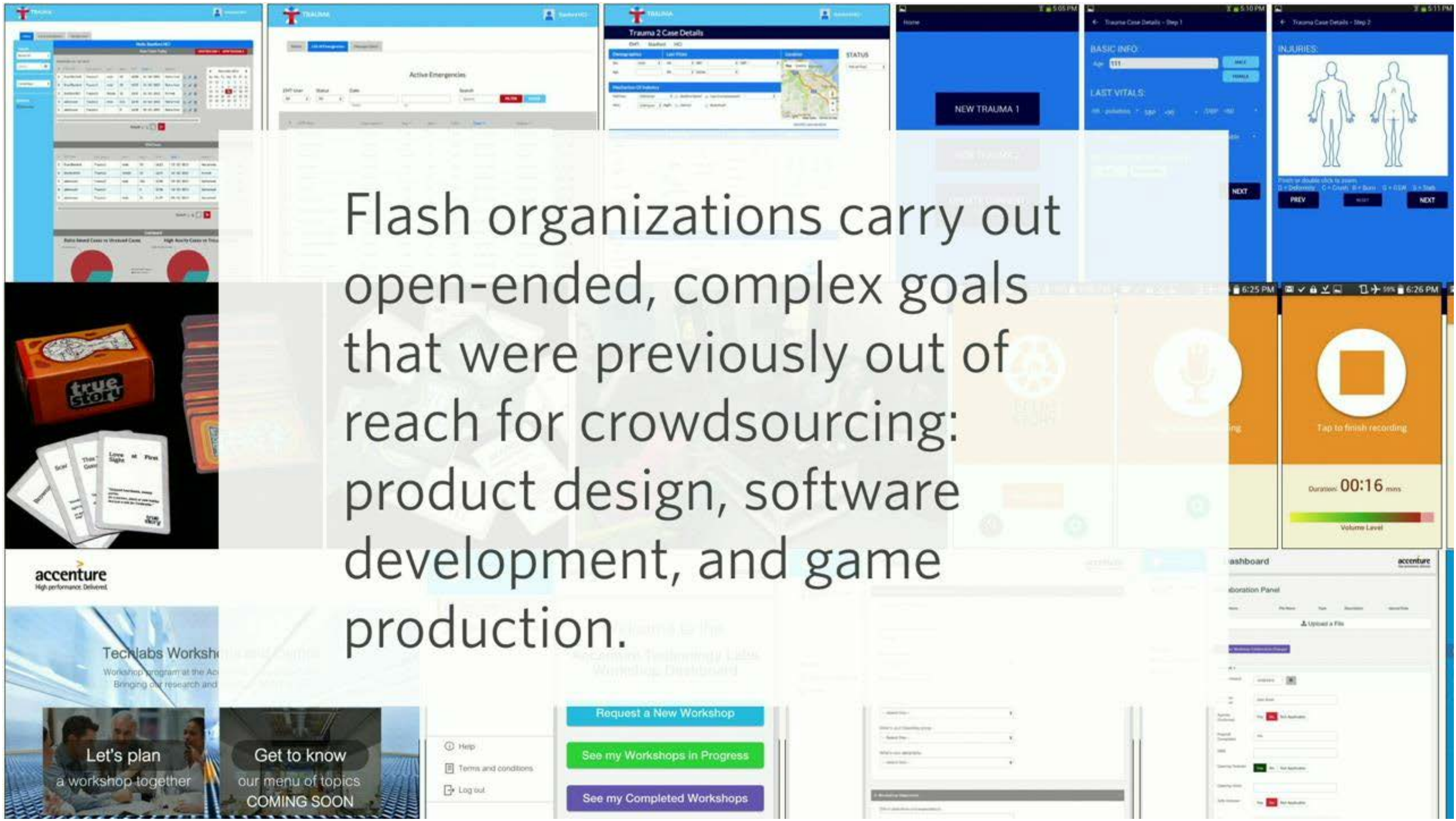
UX

UI

QA

node.js server

Video and website



Flash organizations carry out open-ended, complex goals that were previously out of reach for crowdsourcing: product design, software development, and game production.



FOUNDRY

Web platform that supports authoring, reconfiguring, and running flash organizations

The screenshot displays the Foundry web platform interface. On the left, a sidebar contains a red circular menu icon with three white lines, followed by the text "FOUNDRY". Below this, the text "QUESTION AND ANSWER WEB APPLICATION" is visible. A welcome message reads "Welcome Daniela Retelny! Your role: UI Designer - Users". A red notification states "Your task (User Profile Wireframes) is delayed." Below the notification are two buttons: a green "Complete Task" button and a blue "Take a Break" button.

The main area features a Gantt chart with a time axis from 0:00 to 9:00. A pink banner at the top of the chart area reads "Your task (User Profile Wireframes) is delayed." The chart shows several task bars:

- HOMEPAGE & LOGIN WIREFRAMES COMPLETED**: A green bar from 0:00 to 2:00.
- NEWS FEED WIREFRAMES COMPLETED**: A green bar from 2:00 to 6:00.
- QUESTION & ANSWER WIREFRAMES PAUSED**: A blue bar from 2:00 to 6:00.
- USER PROFILE WIREFRAMES -2 HRS 40 MIN**: A red bar from 2:00 to 4:40.
- INTEGRATE WIREFRAMES 1 HR 45 MIN**: A grey bar from 6:00 to 7:45.
- HEUR 2 HRS**: A grey bar from 7:00 to 9:00.

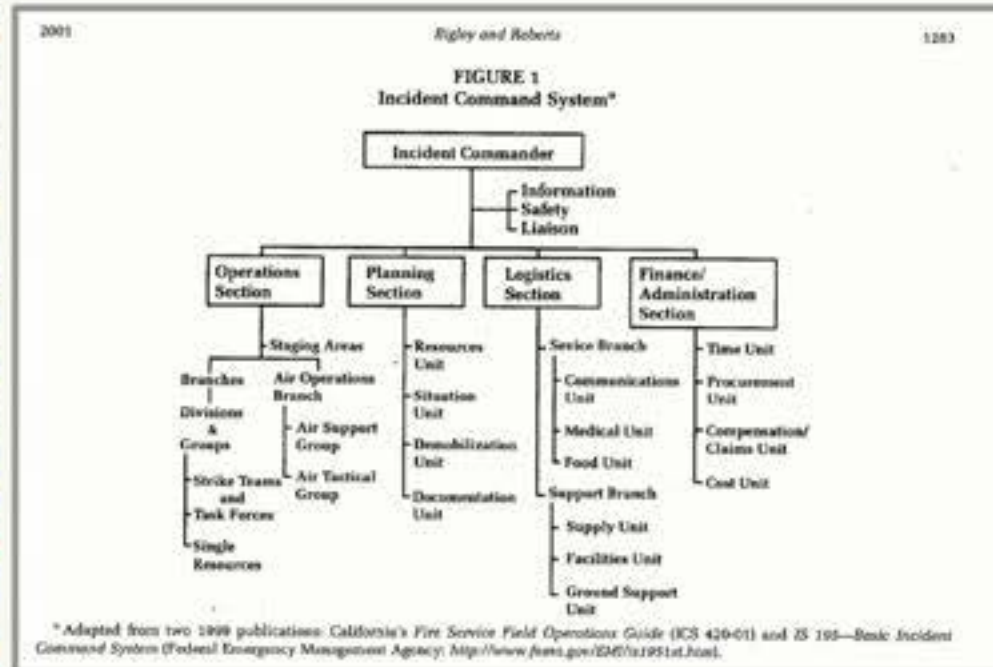
Each task bar includes a person icon, a cloud icon, and a right-pointing arrow icon.

COORDINATION SANS ALGORITHMS

Inspiration: film crews and disaster response teams

[Bigley 2001; Bechky 2006; Klein et. al 2006; Valentine & Edmondson 2015]

Role structures enable interaction based on knowledge of roles rather than asset-specific knowledge of each other

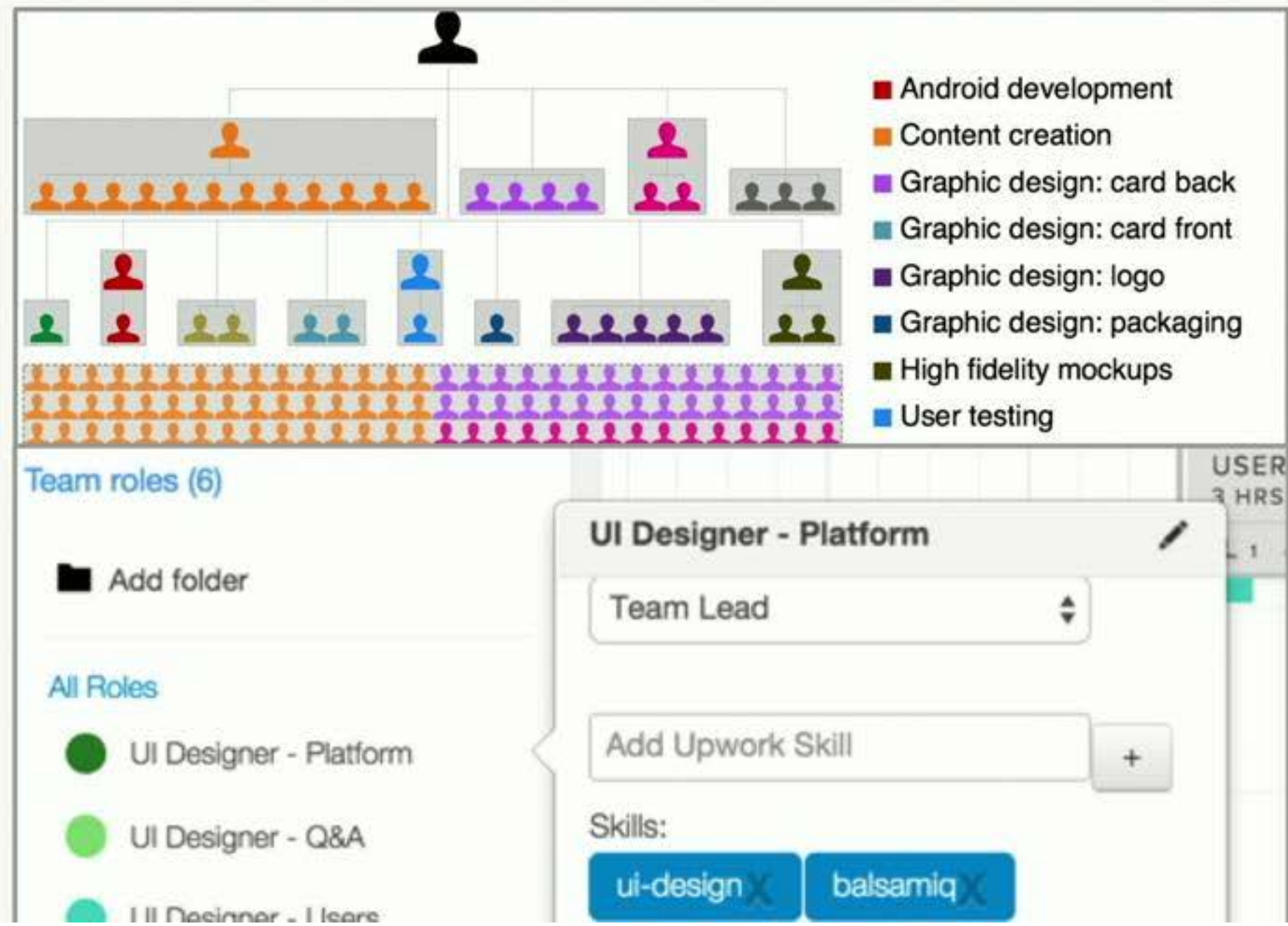


COMPUTATIONAL ORGANIZATIONAL STRUCTURES

Roles: parametrize required expertise

Teams: groups of workers with shared goal

Hierarchy: nested roles that determine decision rights



ON-DEMAND HIRING FROM UPWORK

The screenshot shows a hiring queue interface. On the left, a dark sidebar contains the following information: 'Project: Question and Answer Web Application', 'Task: Homepage & Login Wireframes', 'Position in Queue: No. 1', and 'Deadline to Accept Position: 10 minutes'. Below this are two buttons: 'Accept this position' (green) and 'Decline this position' (white). The main area is titled '> Task Available' and contains a congratulatory message: 'Congratulations! You are at No. 1 position Application project.' Below this is a warning: 'Please read the following information carefully the hiring queue. However, to reinforce again... Please do not close this page; this page will be removed from the hiring queue (only for this...'. At the bottom, there is a green square with the 'Up' logo and the text 'Project overview: answer a question'.

The screenshot shows an onboarding overlay titled 'Your Task'. The text reads: 'This is **YOUR** task. You can now end this tour, and **click on the task rectangle and click start** to read about your task, and start tracking work time. Note that time for reviewing the previous materials, etc. are accounted for as work time.' Below this, it says: 'Pay close attention to the task description, the 'inputs' (what other workers have handed off to you), and the deliverables you are expected to create.' The background shows a task card for 'USER PROFILE WIREFRAMES' with a duration of '3 HRS 45 MIN' and a '1' icon.

Automated, role-specific onboarding

Foundry hiring queue

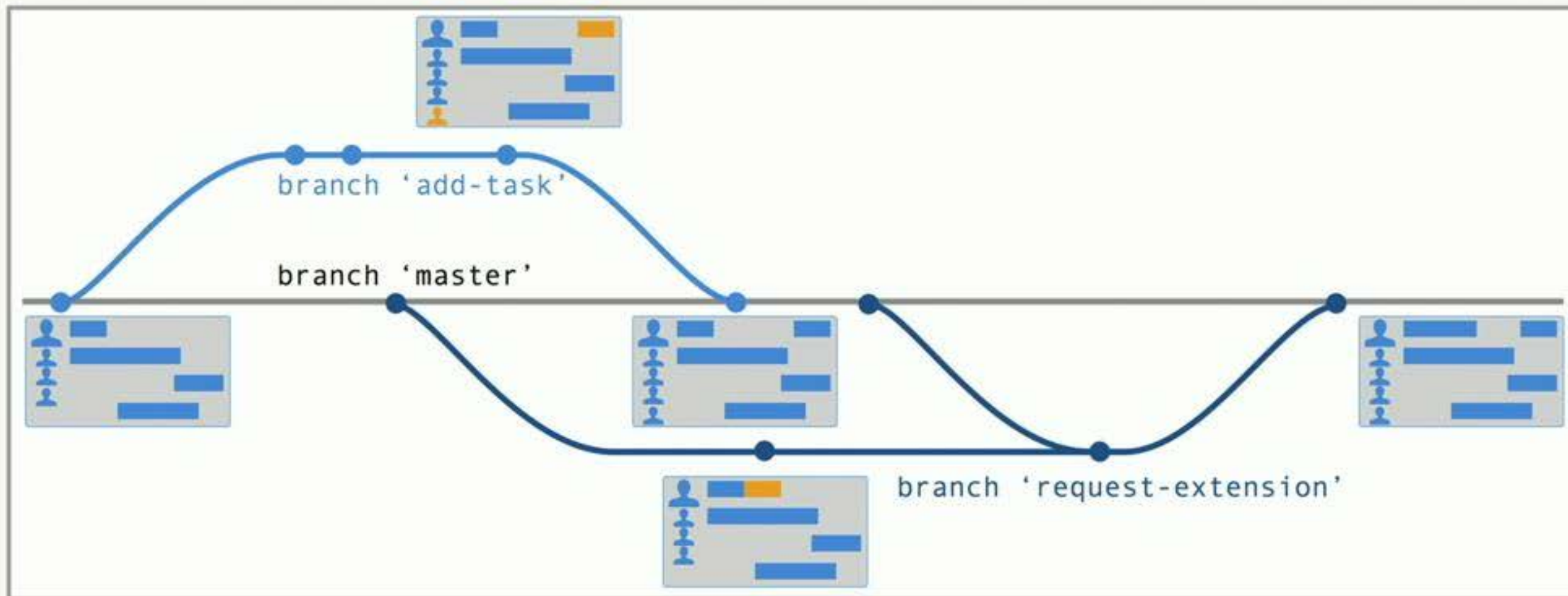
CHALLENGE: RECONFIGURATION

Organizational structures require constant reconfiguration so that the organization can adapt as it proceeds

How can a computational system keep a distributed crowd in sync as the plan evolves?

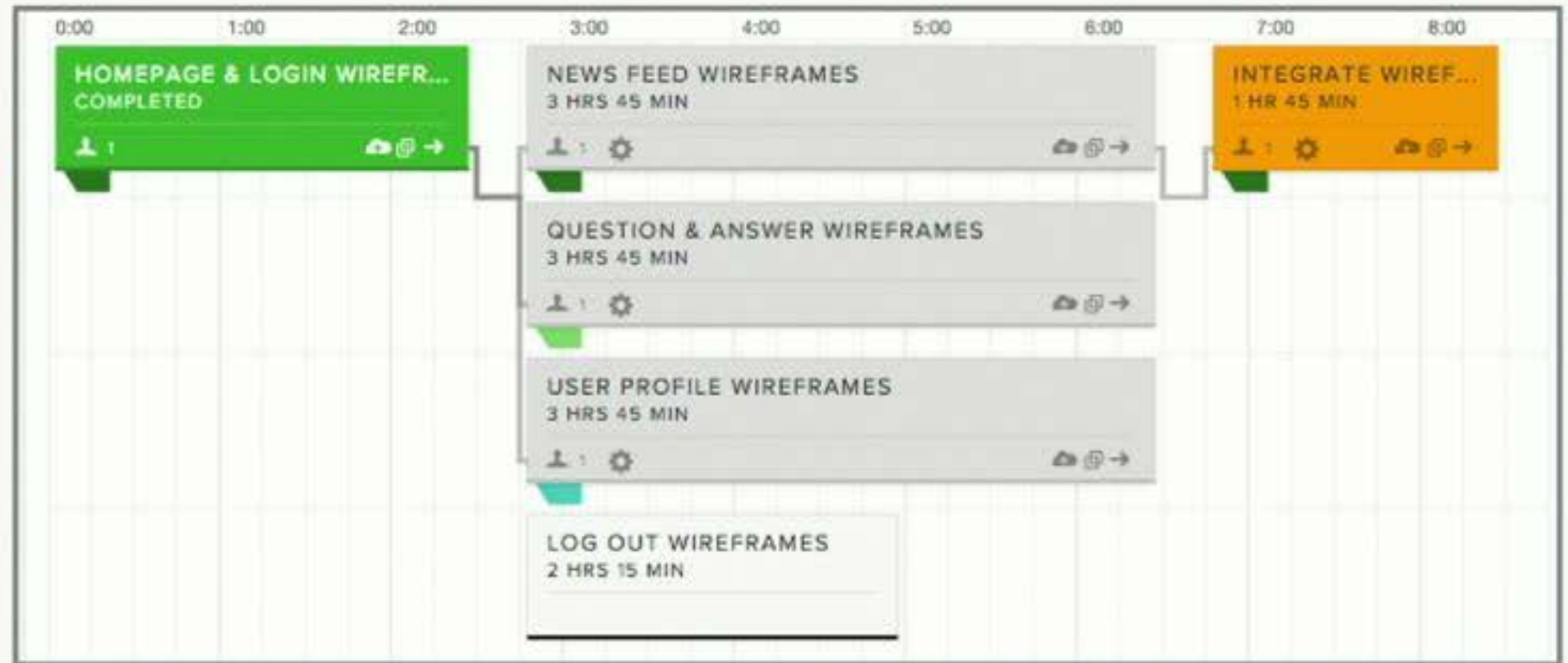
VERSION CONTROL

To enable reconfiguration of the organizational structures:
branching and merging inspired by version control



VERSION CONTROL IN FOUNDRY

Any member can **branch, edit, and make pull requests** against any organizational structure: roles, teams, hierarchy, tasks



Diff view for reviewing

Pull requests are reviewed up the hierarchy and merged through a three-way diff

EVALUATION

Field study: System deployment with outside leaders willing to crowdsource their complex open-ended goals

EMS Report

Leader

Medical resident

Open-ended goal

Develop prototype application for EMTs to transmit patient information en route to hospital

EVALUATION

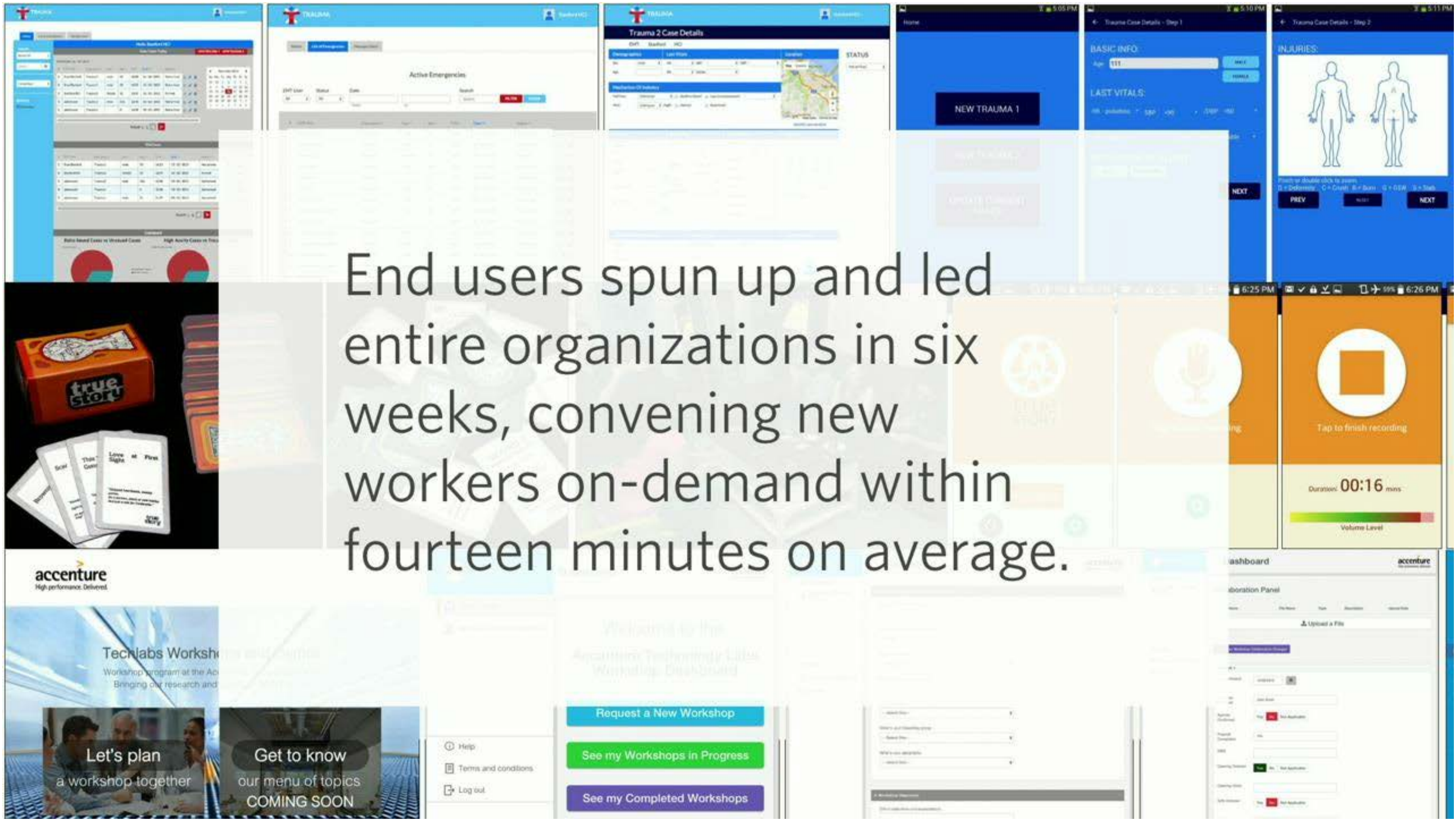
Field study: System deployment with outside leaders willing to crowdsource their complex open-ended goals

	EMS Report	True Story
Leader	Medical resident	Storytelling podcast kickstarter team
Open-ended goal	Develop prototype application for EMTs to transmit patient information en route to hospital	Design and manufacture a storytelling card game with accompanying mobile application

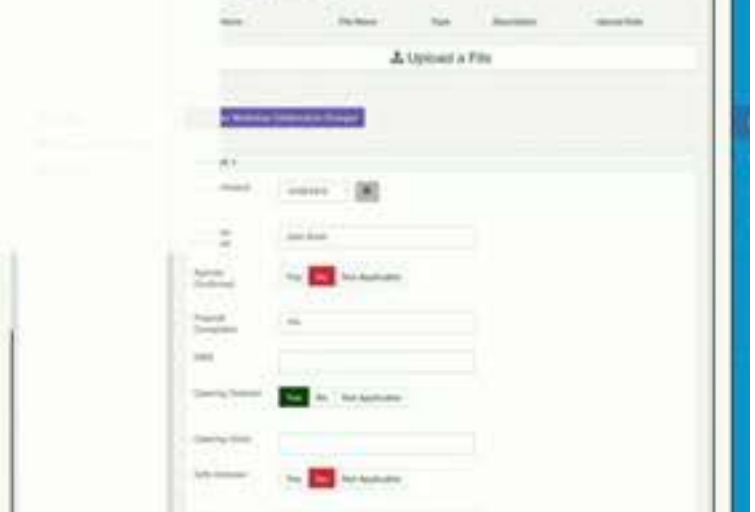
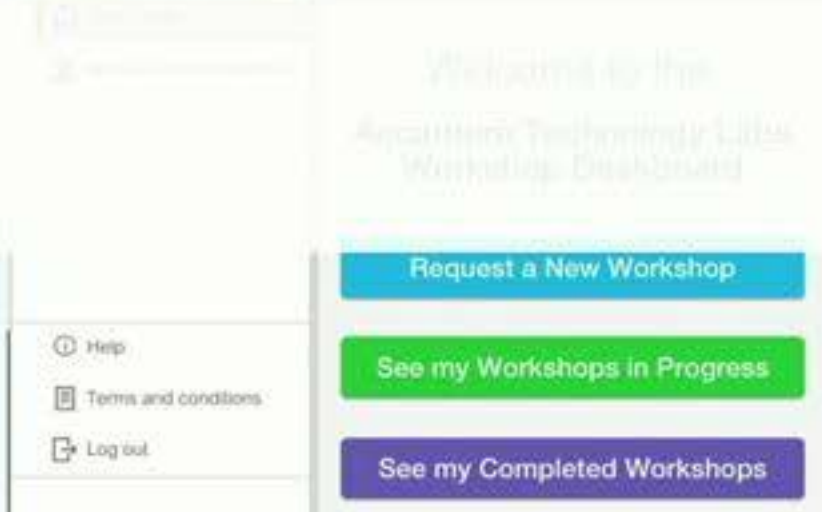
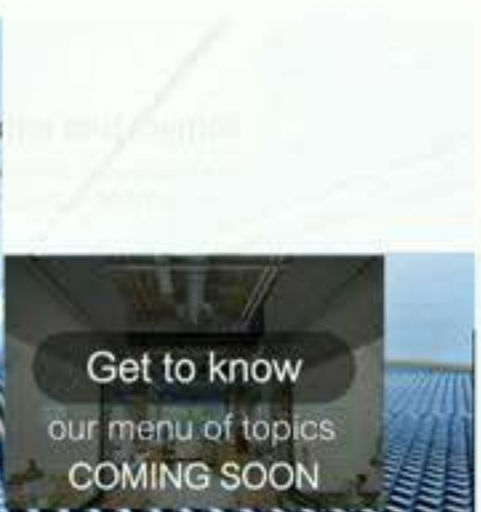
EVALUATION

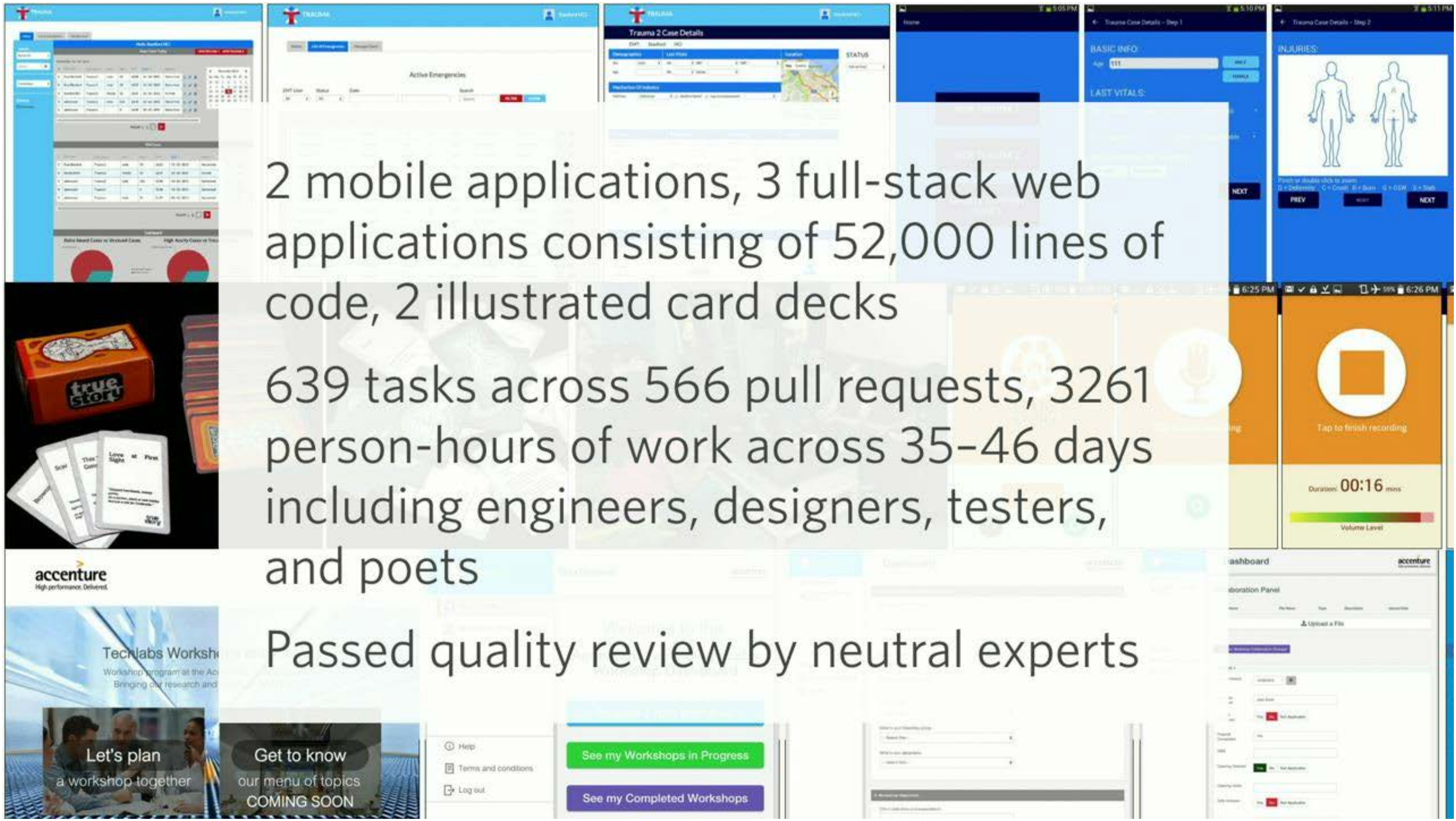
Field study: System deployment with outside leaders willing to crowdsource their complex open-ended goals

	EMS Report	True Story	Workshop Portal
Leader	Medical resident	Storytelling podcast kickstarter team	Tech lab employee of a large company
Open-ended goal	Develop prototype application for EMTs to transmit patient information en route to hospital	Design and manufacture a storytelling card game with accompanying mobile application	Develop a workshop planning portal consistent with enterprise standards and branding



End users spun up and led entire organizations in six weeks, convening new workers on-demand within fourteen minutes on average.





2 mobile applications, 3 full-stack web applications consisting of 52,000 lines of code, 2 illustrated card decks

639 tasks across 566 pull requests, 3261 person-hours of work across 35-46 days including engineers, designers, testers, and poets

Passed quality review by neutral experts

TRUE STORY GAME



TRUE STORY GAME



CRUSHING

Subtle looks, pounding pulse

However long the hover lasts

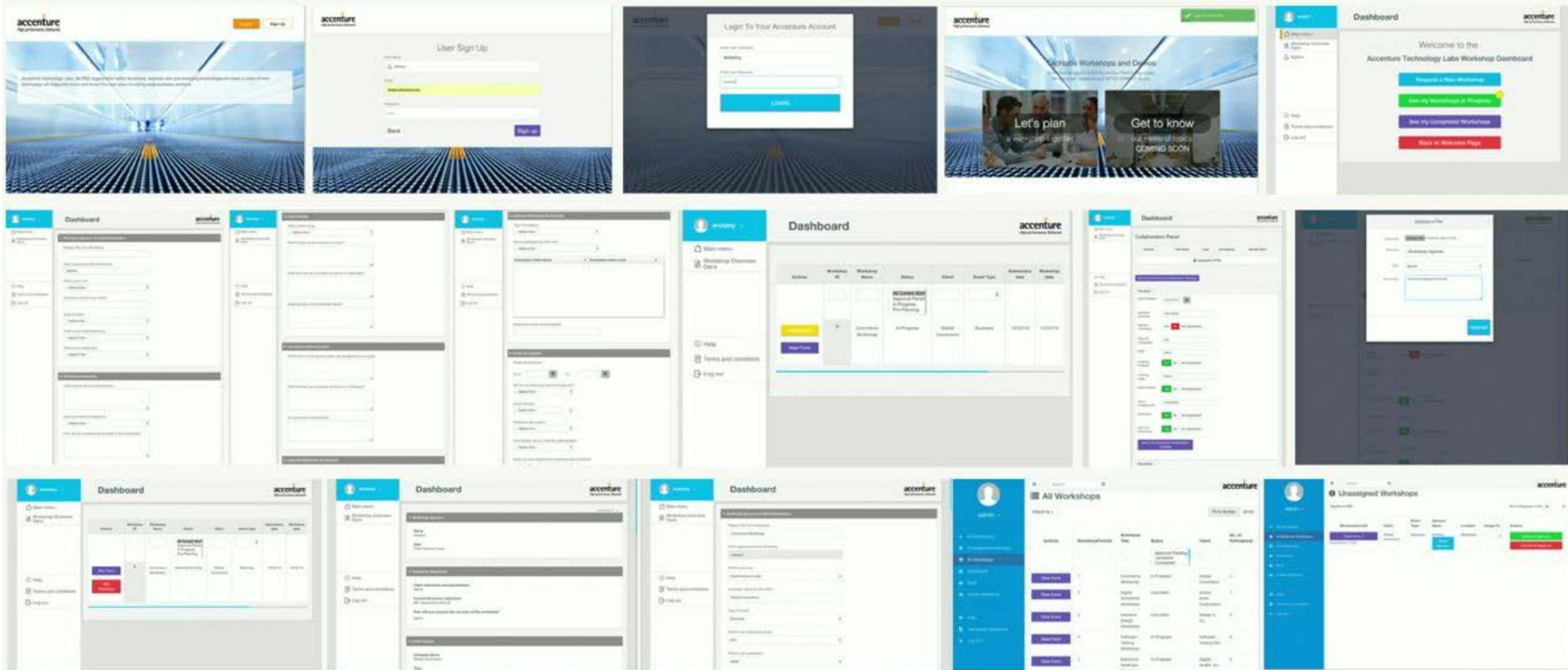
Between friend zone and fun zone

TRUE STORY GAME



Android companion app spun up in the final week

ENTERPRISE WORKSHOP PORTAL



FLASH ORGANIZATIONS: REFLECTION

Flash organizations suggest a future where organizations are **fluidly assembled and adapted** within globally networked collectives.

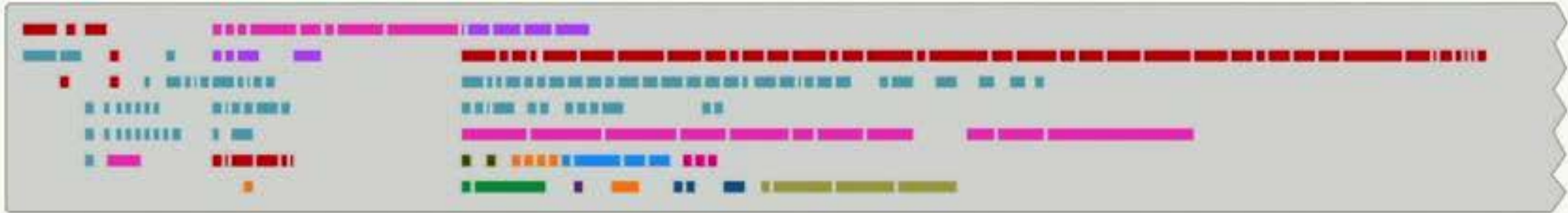
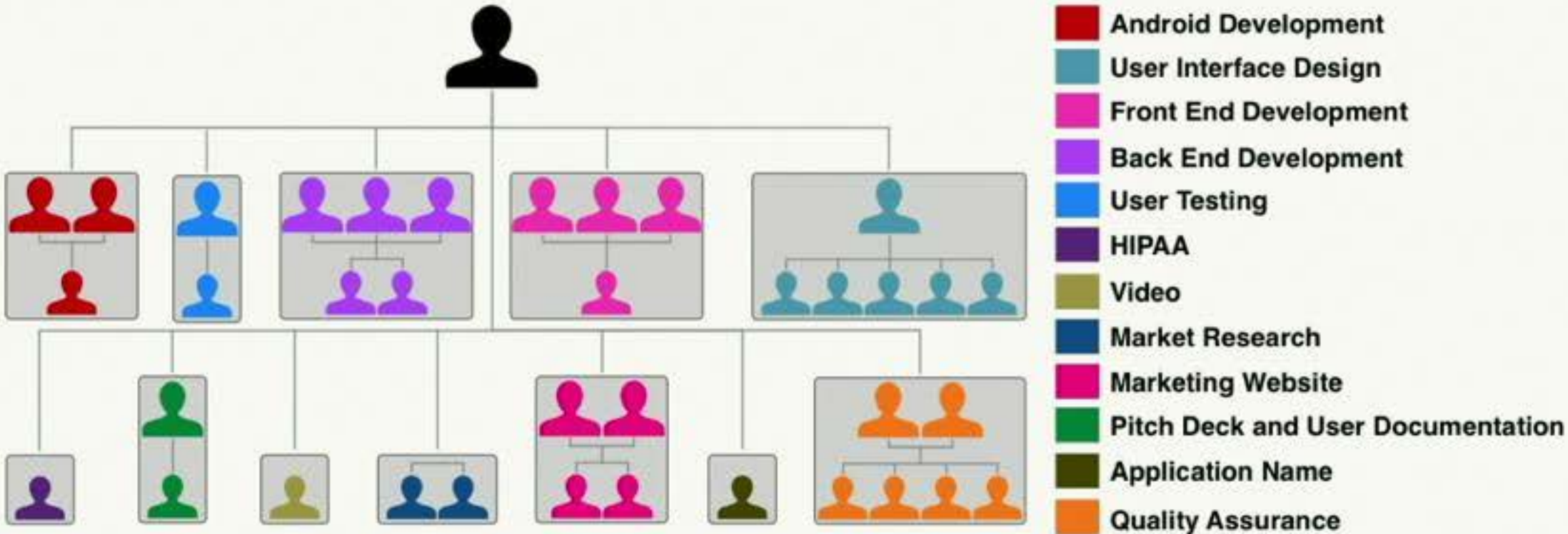
Open questions:

Might data help suggest effective organizational structures?

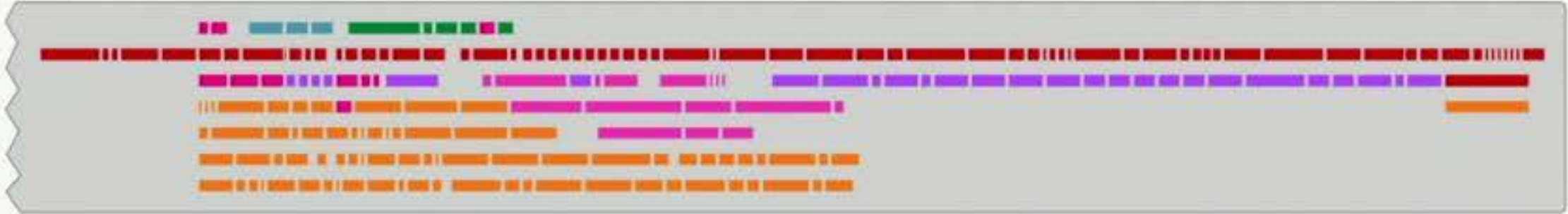
How can researchers support industry norms, labor organization, and legislation to encourage a prosocial future of work?

Do flash organizations change the transaction costs core to the Theory of the Firm?

EMS TRAUMA REPORT



Timeline



Dream Team

Zhou, Valentine, Bernstein. CHI 2018.

WHAT IS THE BEST WAY FOR TEAMS TO ORGANIZE THEMSELVES?

Organizations rely on teaming and ad-hoc collaboration
[Edmondson 2012]

WHAT IS THE BEST WAY FOR TEAMS TO ORGANIZE THEMSELVES?

Organizations rely on teaming and ad-hoc collaboration
[Edmondson 2012]

But: Should teams be flat or hierarchical? Encouraging or critical? Enforcing equal turn-taking?

WHAT IS THE BEST WAY FOR TEAMS TO ORGANIZE THEMSELVES?

Organizations rely on teaming and ad-hoc collaboration
[Edmondson 2012]

But: Should teams be flat or hierarchical? Encouraging or critical? Enforcing equal turn-taking?

These roles, norms, and interaction patterns define **team structures**
[Ilgen et al. 2005]

WHAT IS THE BEST WAY FOR TEAMS TO ORGANIZE THEMSELVES?

Organizations rely on teaming and ad-hoc collaboration
[Edmondson 2012]

But: Should teams be flat or hierarchical? Encouraging or critical? Enforcing equal turn-taking?

These roles, norms, and interaction patterns define **team structures**
[Ilgen et al. 2005]

Researchers theorize ideal structures, then build systems nudging teams toward those structures [Olsen & Olsen 2000; Ackerman 2000; Dourish & Bellotti 1992; Erickson & Kellogg 2000; Winograd 1986; Lykourantzou et al. 2017]

ORG. BEHAVIOR: THERE ARE NO UNIVERSALLY IDEAL STRUCTURES

Structural contingency theory: the best team structures depend on the task and the team members [Donaldson 1999]

ORG. BEHAVIOR: THERE ARE NO UNIVERSALLY IDEAL STRUCTURES

Structural contingency theory: the best team structures depend on the task and the team members [Donaldson 1999]

The wrong structures will doom teams to dysfunction

[Ilgen et al. 2005; Schippers, Edmondson, & West 2014; Ancona, Okhuysen & Perlow 2001]

ORG. BEHAVIOR: THERE ARE NO UNIVERSALLY IDEAL STRUCTURES

Structural contingency theory: the best team structures depend on the task and the team members [Donaldson 1999]

The wrong structures will doom teams to dysfunction
[Ilgen et al. 2005; Schippers, Edmondson, & West 2014; Ancona, Okhuysen & Perlow 2001]

Managers — who are trained and paid for choosing effective team structures — are not effective at the task
[de Brujin, Ten Heuvelhof, & In 't Veld 2002]

DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task

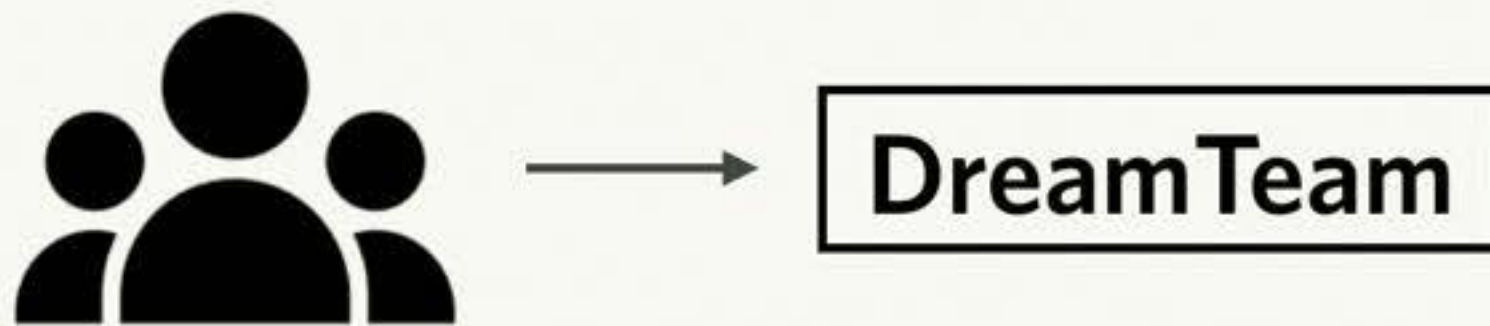
DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task



DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task



DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task

DreamTeam



DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task



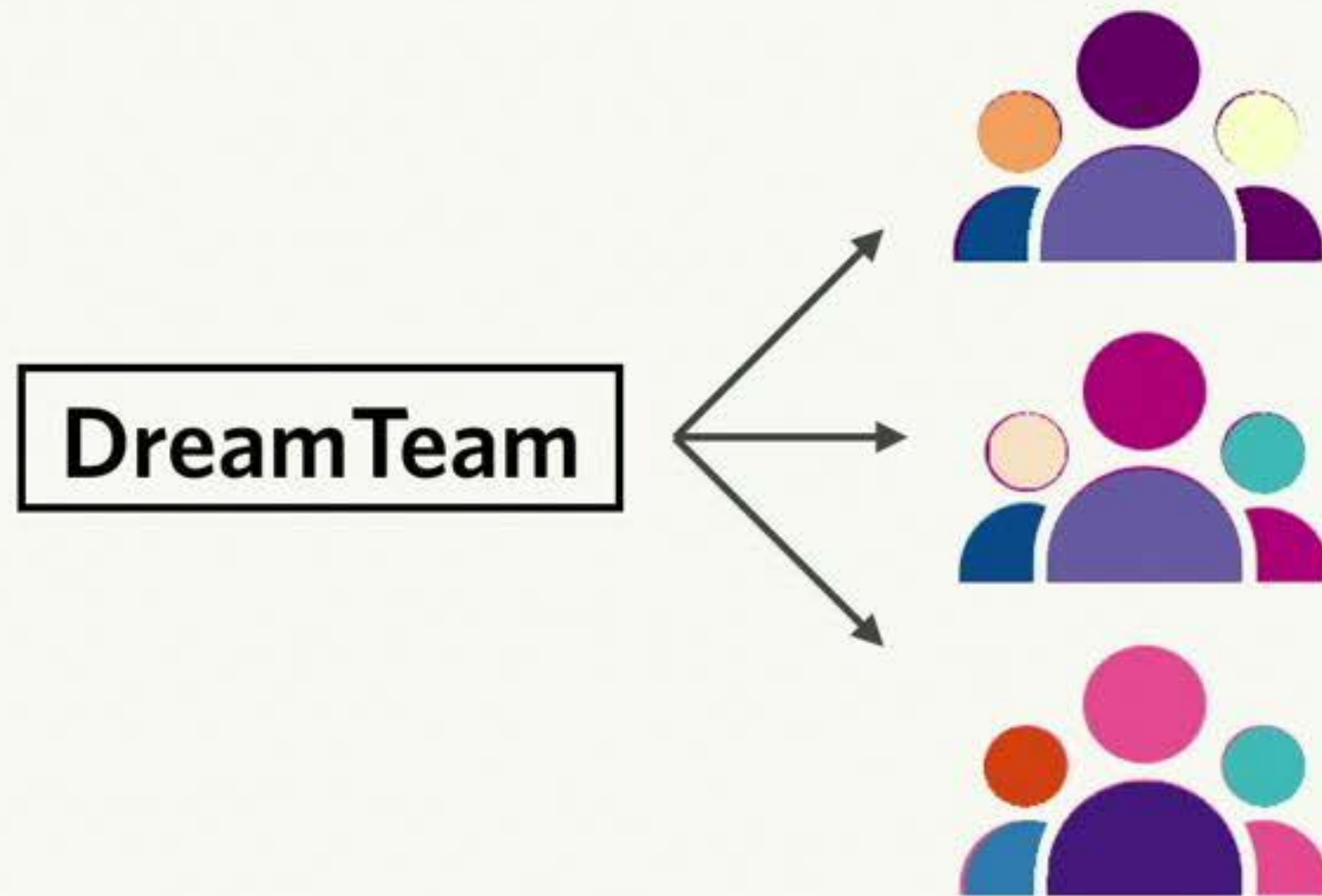
DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task



DREAMTEAM

Rapid self-experimentation with different team structures to converge on one that works well for the team and task



[Redacted]

[Redacted]

[Redacted]

[Redacted]



sdumais 9:21 AM

hello



jteevan 9:21 AM

hi



merrie 9:21 AM

Hi everyone!





sdumais 9:21 AM
hello



jteevan 9:21 AM
hi



merrie 9:21 AM
Hi everyone!



puzzle-robot APP 9:21 AM

INSTRUCTIONS

| _____
| _____
| _____

SUBMISSION

| _____
| _____

+ | _____

[Redacted]

[Redacted]

[Redacted]

[Redacted]



puzzle-robot

APP

9:21 AM

END OF ROUND

[Redacted]

[Redacted]



puzzle-robot

APP

9:21 AM

END OF ROUND



<feedback to DreamTeam system>

+



[Redacted]

[Redacted]

[Redacted]

[Redacted]



puzzle-robot

APP

9:21 AM

END OF ROUND

[Redacted]

<feedback to DreamTeam system>



dreamteam-robot

9:21 AM

This round change the following...

Be super cheery! Make sure to write encouraging comments to all your teammates, despite any losses!

[Input field with a plus sign icon on the left]

Hierarchy

None, Centralized, Decentralized



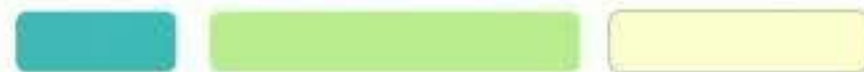
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



Time 

Hierarchy

None, Centralized, Decentralized



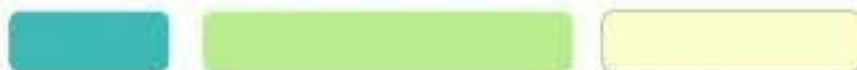
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



Time

<feedback>

Hierarchy

None, Centralized, Decentralized



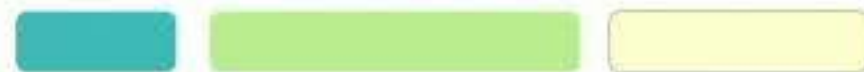
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



Time



Hierarchy

None, Centralized, Decentralized



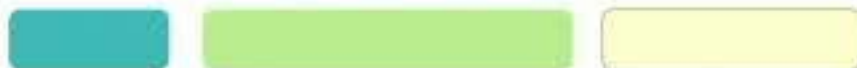
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



Time



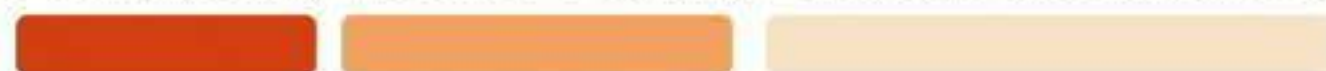
Hierarchy

None, Centralized, Decentralized



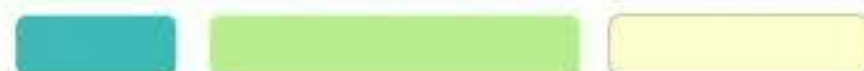
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



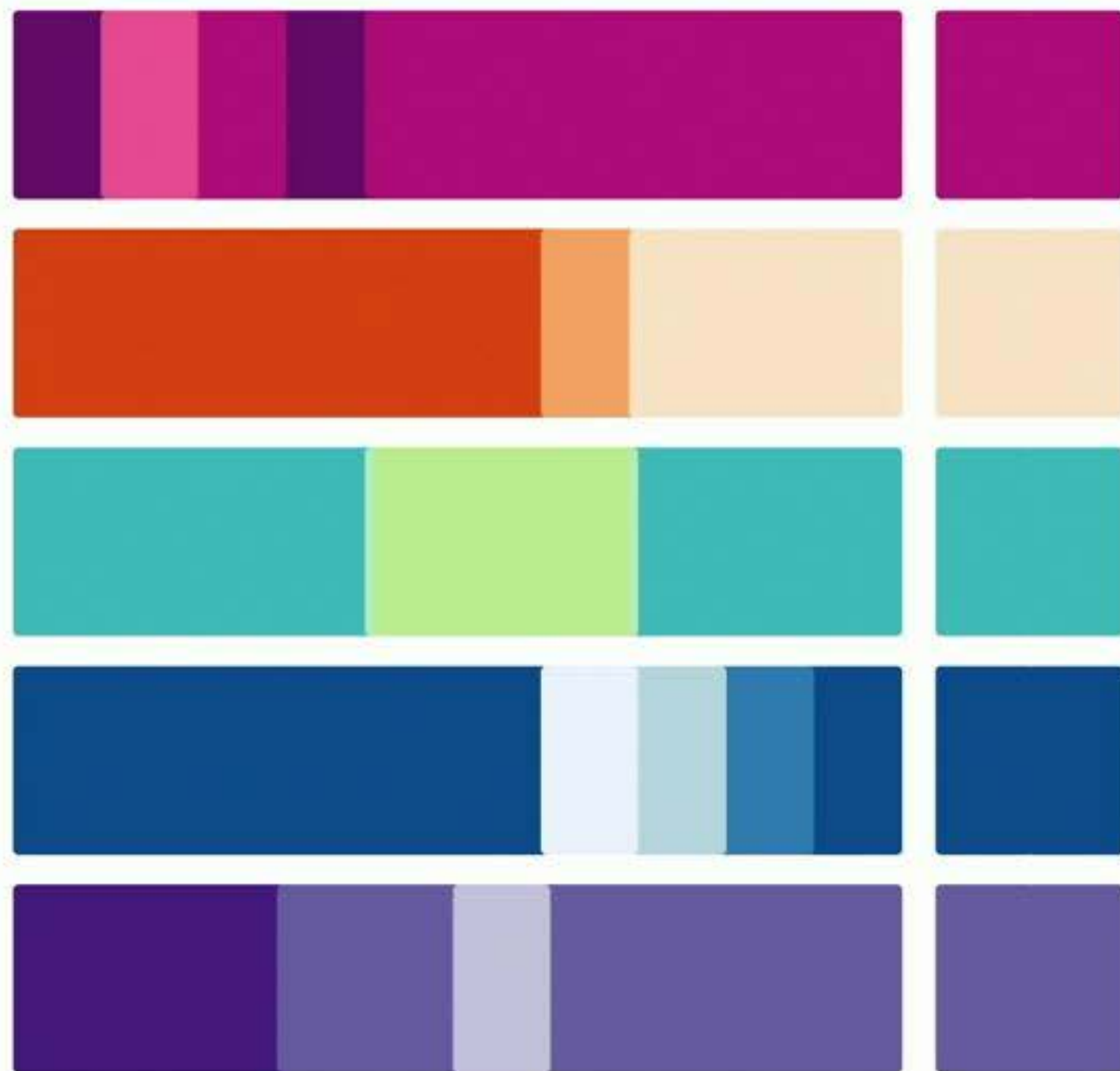
Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



Time

NETWORK OF MULTI-ARMED BANDITS

Multi-armed bandits efficiently explore multiple options over time.

However, this results in so much simultaneous change that **teams become quickly overwhelmed**



Hierarchy

None, Centralized, Decentralized



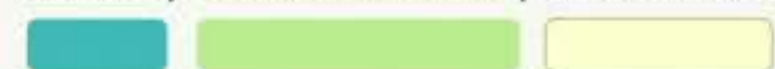
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



NETWORK OF MULTI-ARMED BANDITS

Multi-armed bandits efficiently explore multiple options over time.

However, this results in so much simultaneous change that **teams become quickly overwhelmed**



Hierarchy

None, Centralized, Decentralized



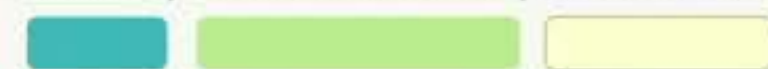
Interaction Patterns

Emergent, Round-robin, Equally distributed



Norms of Engagement

None, Professional, Informal



Decision-Making Norms

None, Divergent, Convergent, Informed, Rapid



Feedback Norms

None, Encouraging, Critical



TEMPORALLY CONSTRAINED BANDITS

Model when teams are open to change, and how much change they are open to simultaneously

e.g., teams are most open to change at the midpoint of their progress
[Okhuysen and Waller 2002]

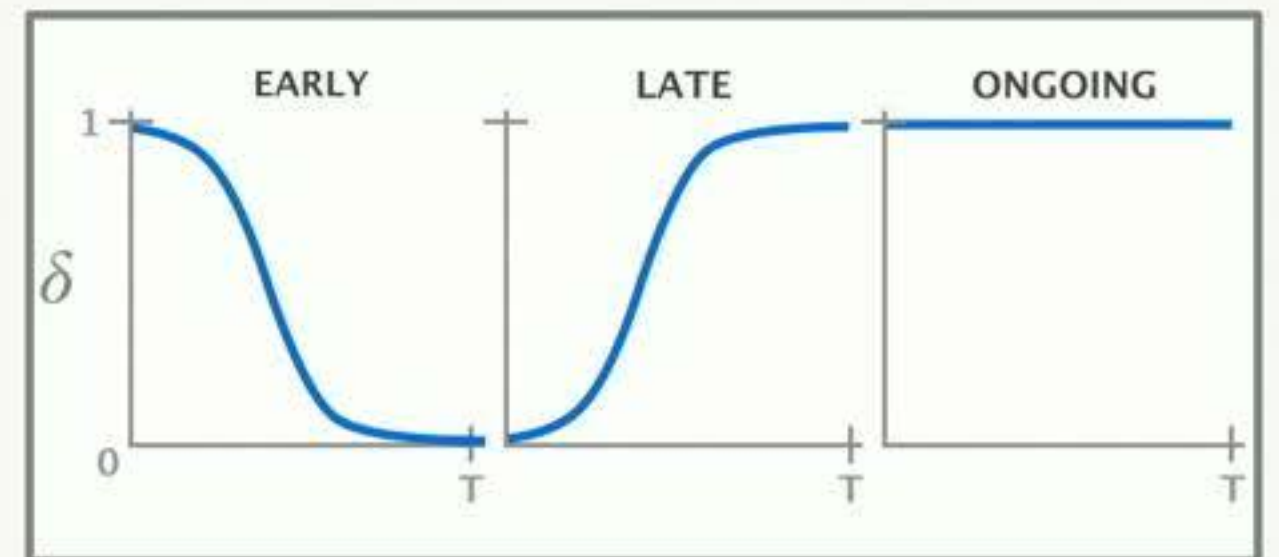
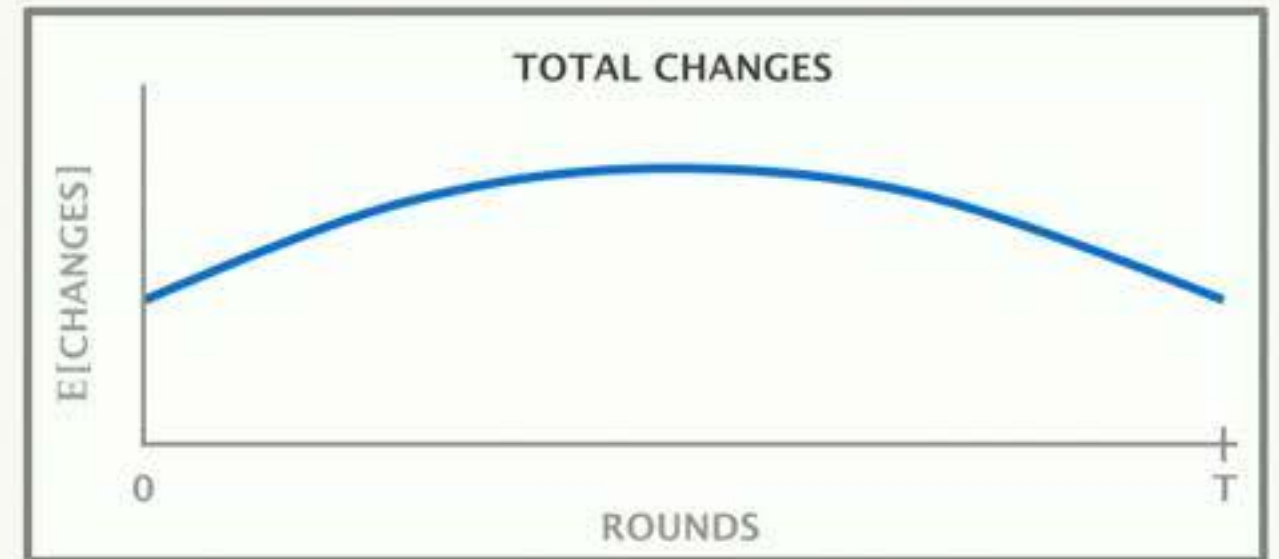
e.g., teams are resilient to exploring hierarchical structures early on, but less resilient to changing them later
[Marks, Mathieu, & Zaccaro 2001]

TEMPORALLY CONSTRAINED BANDITS

Model when teams are open to change, and how much change they are open to simultaneously

e.g., teams are most open to change at the midpoint of their progress
[Okhuysen and Waller 2002]

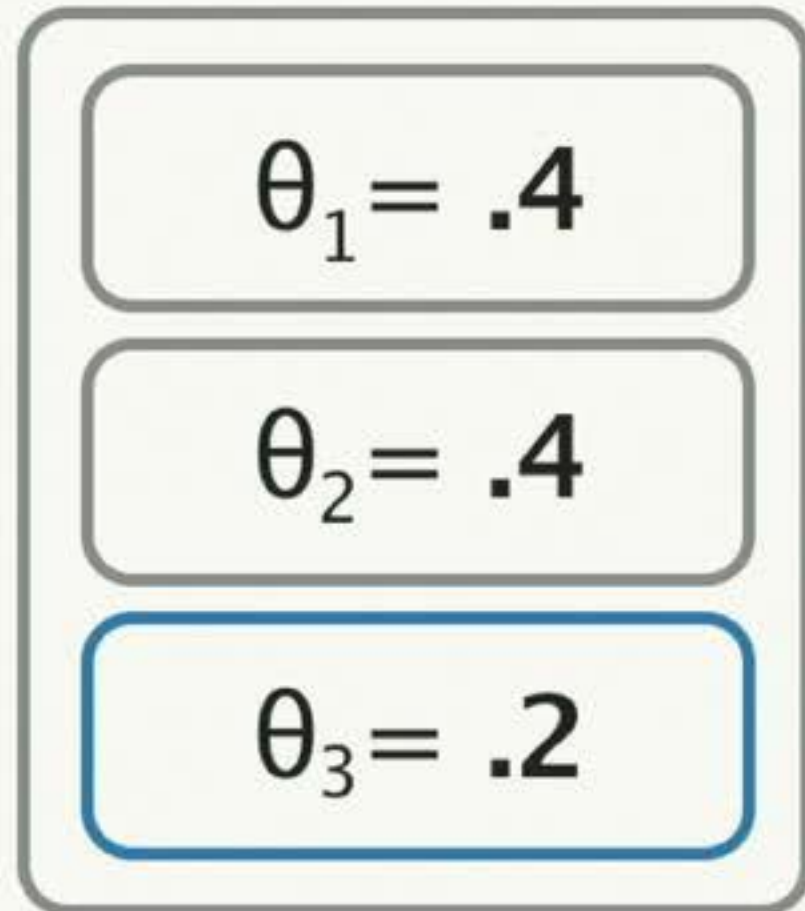
e.g., teams are resilient to exploring hierarchical structures early on, but less resilient to changing them later
[Marks, Mathieu, & Zaccaro 2001]



TEMPORALLY CONSTRAINED BANDITS

Redistribute the probability of arm selection via Thompson sampling to respect desired expected value of changes

CURRENT
ARM



TEMPORALLY CONSTRAINED BANDITS

Redistribute the probability of arm selection via Thompson sampling to respect desired expected value of changes

CURRENT
ARM

$$\theta_1 = .4$$

$$\theta_2 = .4$$

$$\theta_3 = .2$$

$$\theta_1 = .1$$

$$\theta_2 = .1$$

$$\theta_3 = .8$$

TEMPORALLY CONSTRAINED BANDITS

Redistribute the probability of arm selection via Thompson sampling to respect desired expected value of changes

CURRENT
ARM

$$\theta_1 = .4$$

$$\theta_2 = .4$$

$$\theta_3 = .2$$

$$\delta = .25$$

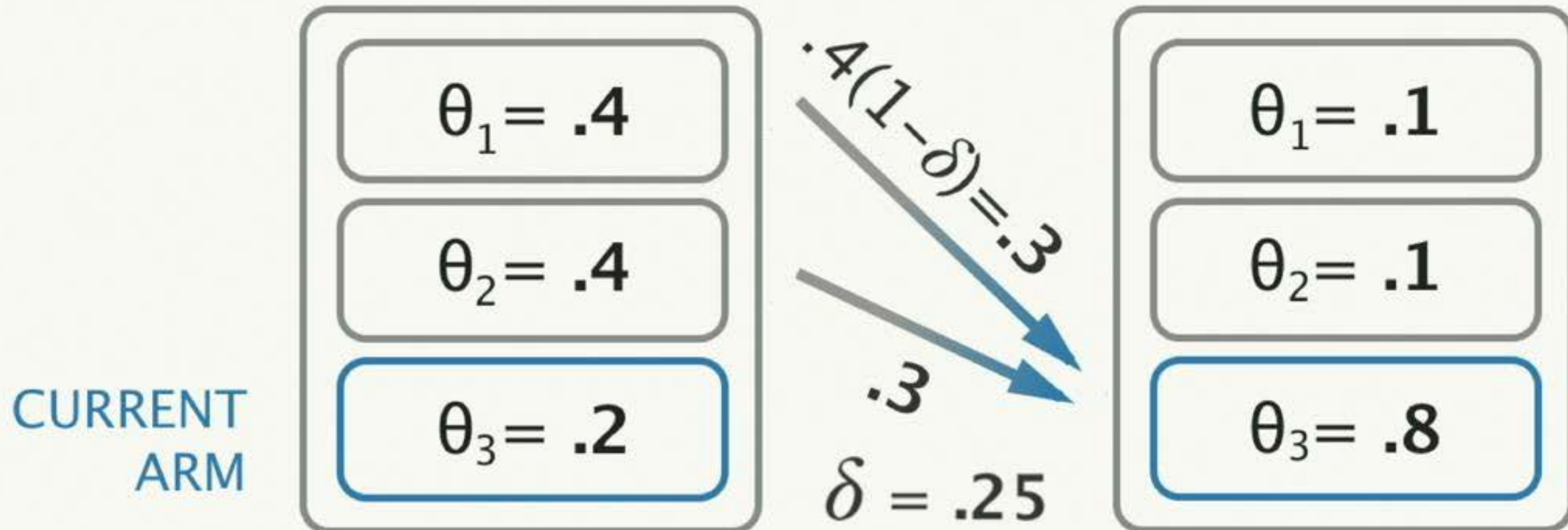
$$\theta_1 = .1$$

$$\theta_2 = .1$$

$$\theta_3 = .8$$

TEMPORALLY CONSTRAINED BANDITS

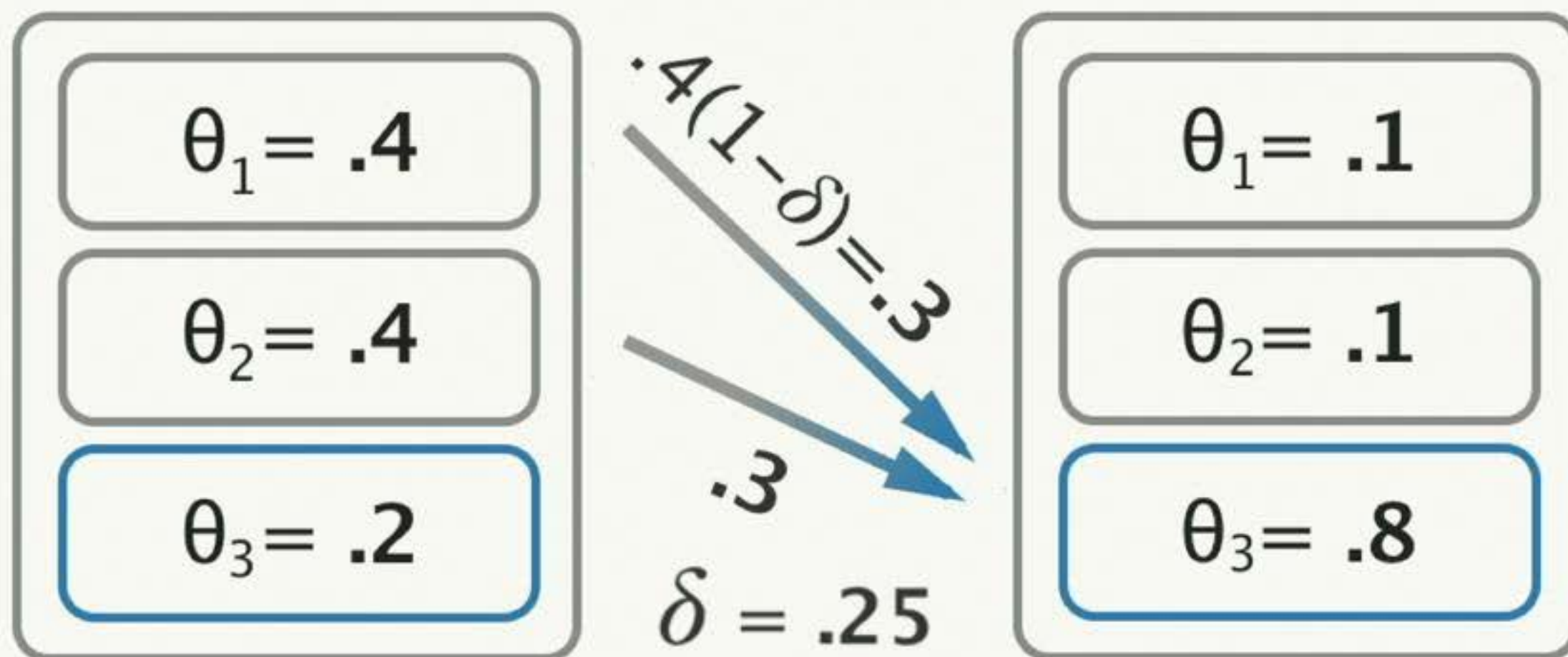
Redistribute the probability of arm selection via Thompson sampling to respect desired expected value of changes

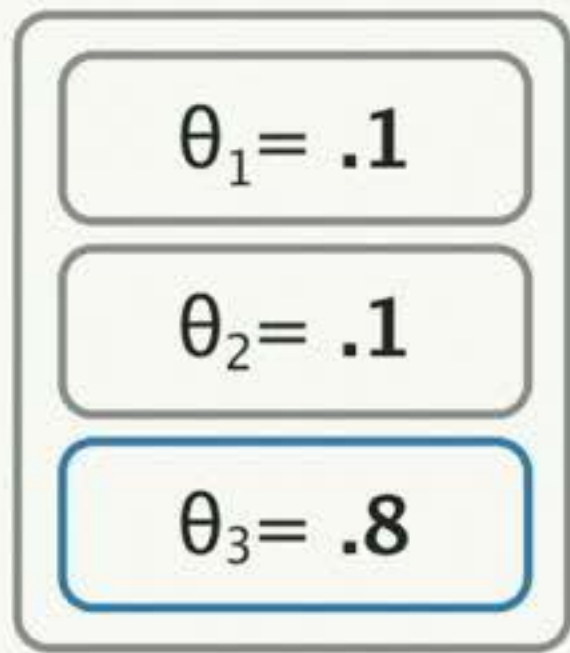


TEMPORALLY CONSTRAINED BANDITS

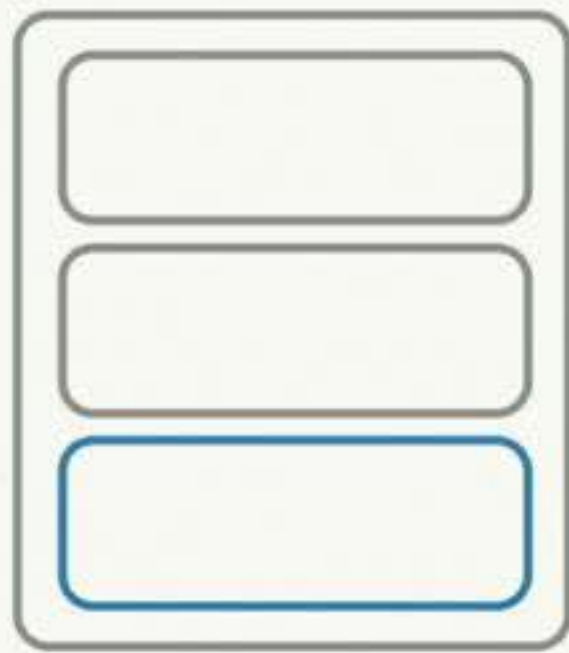
$$\theta'_i = \begin{cases} \theta_i \delta, & \text{if } i \neq c \\ \theta_i + \sum_{j \neq c} \theta_j (1 - \delta), & \text{if } i = c \end{cases}$$

CURRENT
ARM





Hierarchy



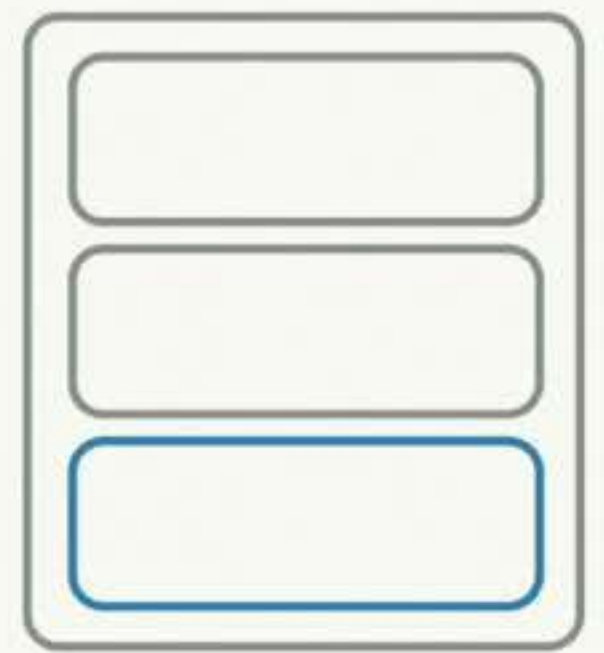
Interaction
Patterns



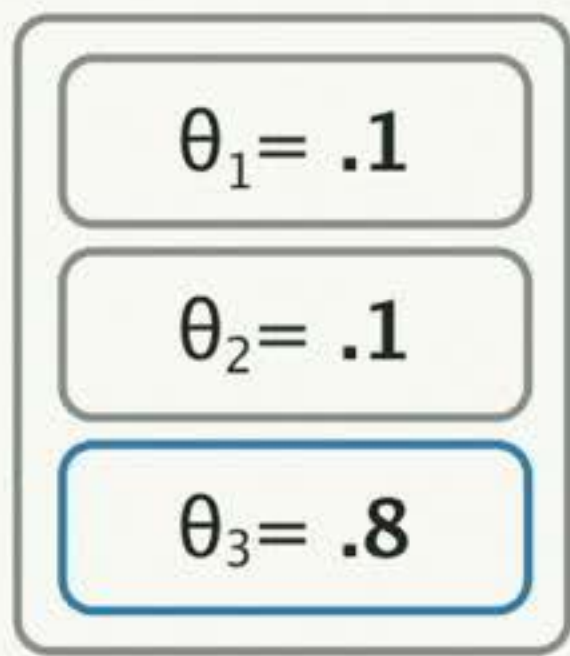
Norms of
Engagement



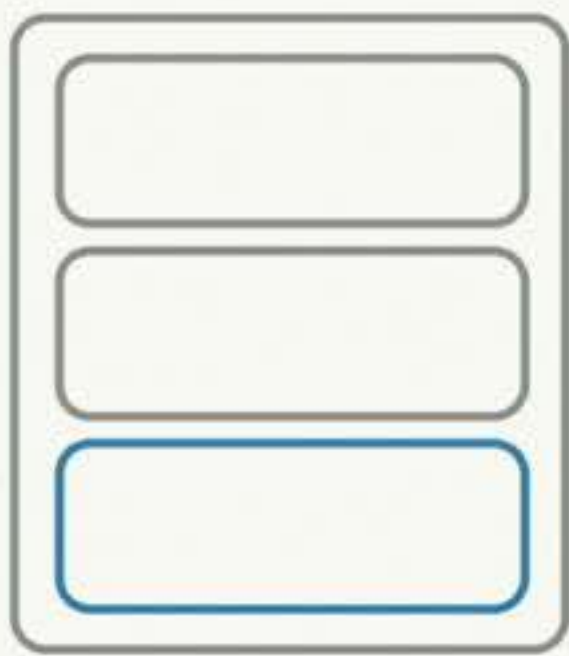
Decision
Making Norms



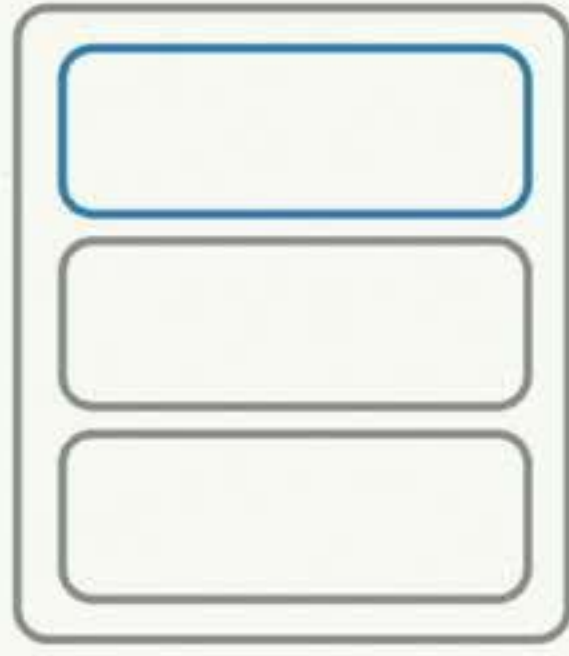
Feedback
Norms



Hierarchy



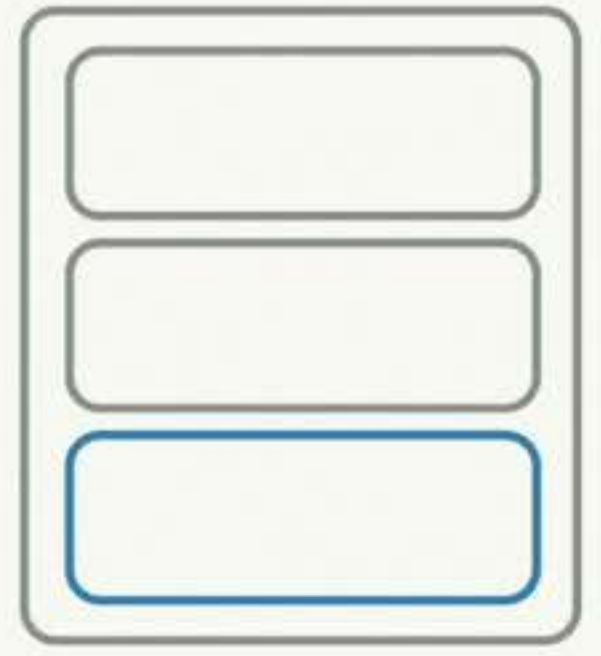
Interaction
Patterns



Norms of
Engagement



Decision
Making Norms



Feedback
Norms



Global constraint on expected number of changes

Prioritize which bandits can change and when

EVALUATION METHOD

135 workers on Mechanical Turk randomized into 45 teams

Measure & bandit reward: team performance on a collaborative intellectual task — score on Codewords puzzle

1 training round, 1 baseline round, 10 performance rounds

EVALUATION METHOD

135 workers on Mechanical Turk randomized into 45 teams

Measure & bandit reward: team performance on a collaborative intellectual task — score on Codewords puzzle

1 training round, 1 baseline round, 10 performance rounds

Conditions:

1. Manager-chosen
2. Collectively-chosen
3. Control

EVALUATION METHOD

135 workers on Mechanical Turk randomized into 45 teams

Measure & bandit reward: team performance on a collaborative intellectual task — score on Codewords puzzle

1 training round, 1 baseline round, 10 performance rounds

Conditions:

1. Manager-chosen
2. Collectively-chosen
3. Control
4. Unconstrained bandit-chosen
5. DreamTeam-chosen

DREAMTEAM TEAMS OUTPERFORM OTHER CONDITIONS BY ~40%

DreamTeam outperformed:

Manager-chosen	by 46%
Collectively-chosen	by 45%
Unconstrained bandit-chosen	by 41%
Control	by 38%

Repeated measures ANCOVA $p < 0.05$, all post-hoc Tukey pairwise comparisons to Dreamteam $p < 0.05$. $N = 45$. Non-intervention training round used as a covariate to adjust for teams' initial performance.

DREAMTEAM: REFLECTION

The heuristics we use to decide on our team structures can be risk-averse, avoiding fruitful exploration and adaptation

But, raw algorithms overcompensate and overwhelm, leading people to ignore them. Design can help.

Open questions:

How might we combine voices equitably in the reward feedback?

Can we adapt when membership changes? When tasks change? Over the long term? In traditional organizations?

DREAMTEAM TEAMS OUTPERFORM OTHER CONDITIONS BY ~40%

DreamTeam outperformed:

Manager-chosen	by 46%
Collectively-chosen	by 45%
Unconstrained bandit-chosen	by 41%
Control	by 38%

Repeated measures ANCOVA $p < 0.05$, all post-hoc Tukey pairwise comparisons to Dreamteam $p < 0.05$. $N = 45$. Non-intervention training round used as a covariate to adjust for teams' initial performance.

DREAMTEAM: REFLECTION

The heuristics we use to decide on our team structures can be risk-averse, avoiding fruitful exploration and adaptation

But, raw algorithms overcompensate and overwhelm, leading people to ignore them. Design can help.

Open questions:

How might we combine voices equitably in the reward feedback?

Can we adapt when membership changes? When tasks change? Over the long term? In traditional organizations?

Crowd research



Vaish, Gaikwad, Veit, Krishna, Ibarra, Simoiu,
Wilber, Belongie, Davis, Goel, Bernstein.
UIST 2017.

RESEARCH: THE DOMAIN OF THE PRIVILEGED FEW

Those able to attend prestigious universities can access research experiences that support open-ended inquiry and launch careers [Russell et al. 2007]

...but the vast majority of people cannot [Bowen and Bok 2016; Bianchini 2011]



Top 50 global universities, US News 2017

RESEARCH: THE DOMAIN OF THE PRIVILEGED FEW

Those able to attend prestigious universities can access research experiences that support open-ended inquiry and launch careers [Russell et al. 2007]

...but the vast majority of people cannot [Bowen and Bok 2016; Bianchini 2011]



Top 50 global universities, US News 2017

RESEARCH: THE DOMAIN OF THE PRIVILEGED FEW

Those able to attend prestigious universities can access research experiences that support open-ended inquiry and launch careers [Russell et al. 2007]

...but the vast majority of people cannot [Bowen and Bok 2016; Bianchini 2011]



A research ecosystem that **under-represents minorities and developing regions**, and a literature that overlooks their perspectives

Top 50 global universities, US News 2017

CROWD RESEARCH

A **crowdsourcing technique** enabling a global crowd to work together on an **open-ended research project**

Participants collaborate as one large team to brainstorm, execute and publish the project under the leadership of a PI

GOALS

Enable access to training and research experiences in support of upward career and educational mobility

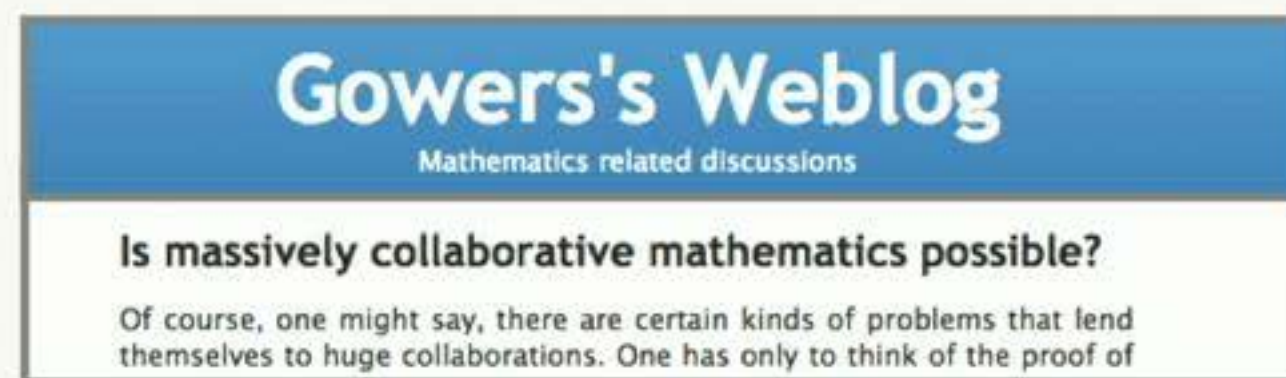
Convene hundreds or thousands of people on a single ambitious project

WE ARE NOT EQUIPPED FOR LARGE SCALE, OPEN ENDED RESEARCH

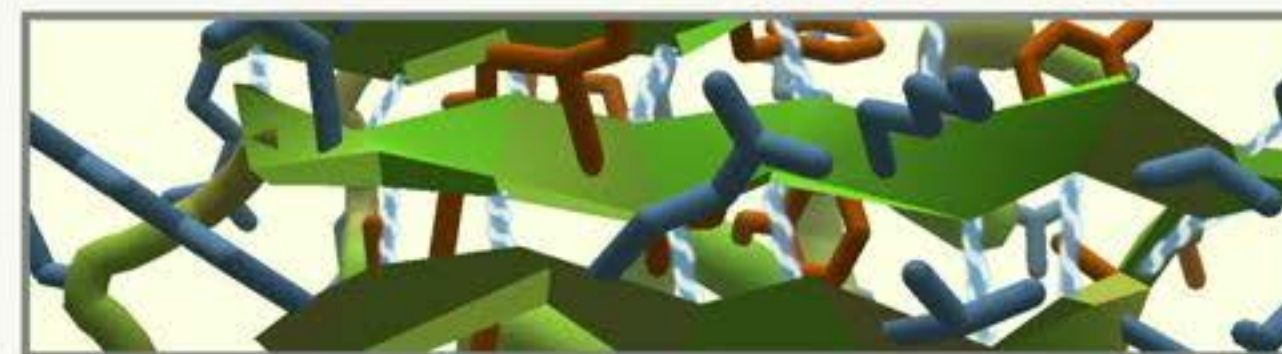
Research is not a linear path from idea to result: it is an iterative process of exploration

[Gowers 2000]

In contrast, citizen science efforts today focus on **pre-defined goals** in order to structure the crowd's contributions



[Gowers and Tao; Cranshaw and Kittur 2011]



[Cooper et al. 2010]

PROBLEMS

Coordination:

How do we prevent the project from moving in 1,000 directions at once, across easily 6,000 messages per week?

Credit:

How can we provide proof that participants made substantial contributions to the project, when no one central authority can assert this?

CROWD RESEARCH

Iterative crowdsourcing technique:

Weekly cycle of open contribution, synchronous collaboration, and peer assessment

Decentralized credit:

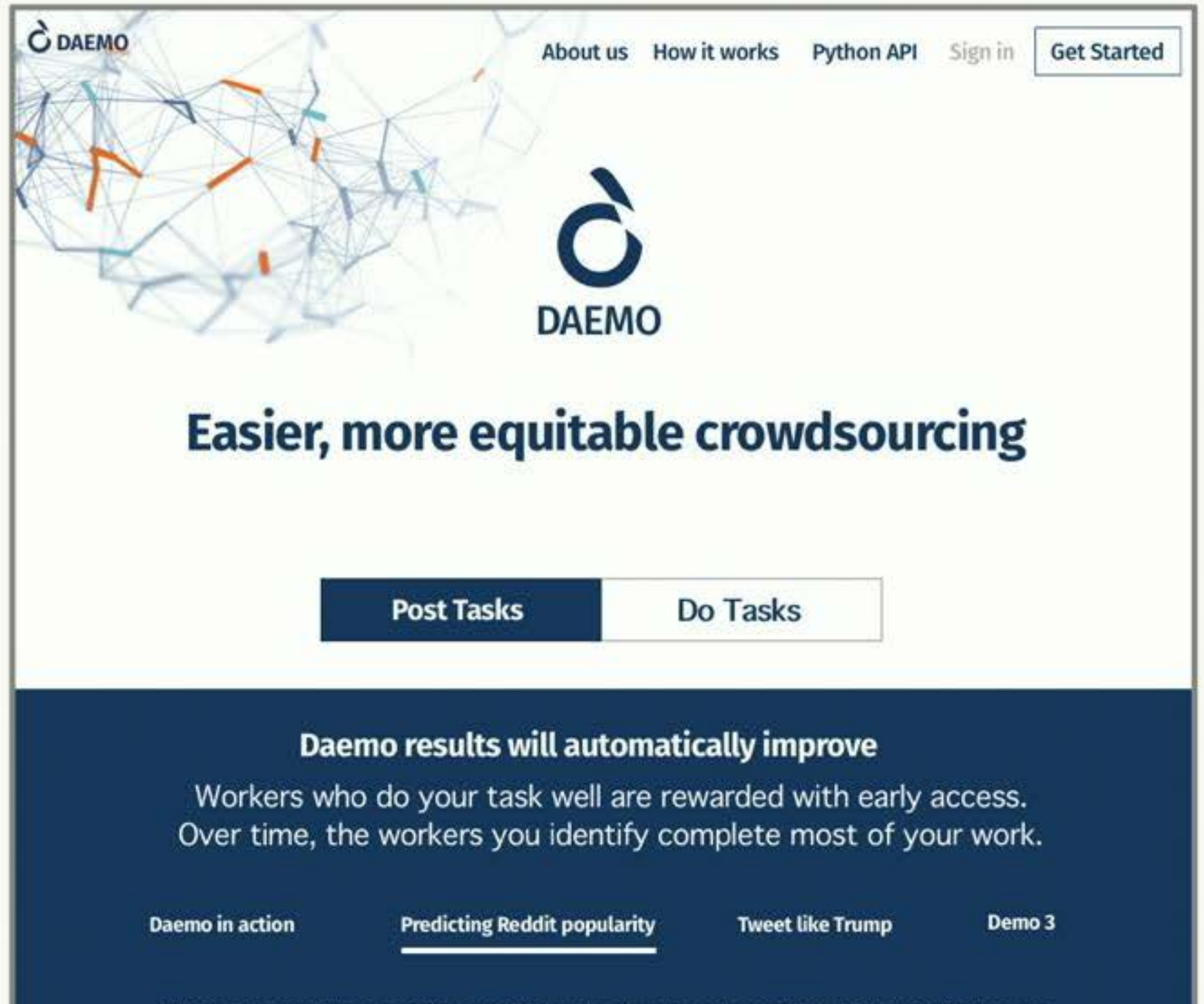
Participants allocate finite credits to each other, enabling a graph centrality algorithm to determine credit and author order

THREE PARALLEL PROJECTS

HCI

Michael Bernstein,
Stanford

Building a new crowd
marketplace



The screenshot shows the DAEMO website homepage. At the top left is the DAEMO logo. To the right are navigation links: "About us", "How it works", "Python API", "Sign in", and a "Get Started" button. The background features a network graph with blue and orange nodes and edges. Below the navigation is a large DAEMO logo and the headline "Easier, more equitable crowdsourcing". Underneath the headline are two buttons: "Post Tasks" (highlighted in dark blue) and "Do Tasks". A dark blue section below contains the text "Daemo results will automatically improve" followed by a paragraph: "Workers who do your task well are rewarded with early access. Over time, the workers you identify complete most of your work." At the bottom, there are four links: "Daemo in action", "Predicting Reddit popularity" (underlined), "Tweet like Trump", and "Demo 3".

THREE PARALLEL PROJECTS

Computer vision

James Davis, UCSC

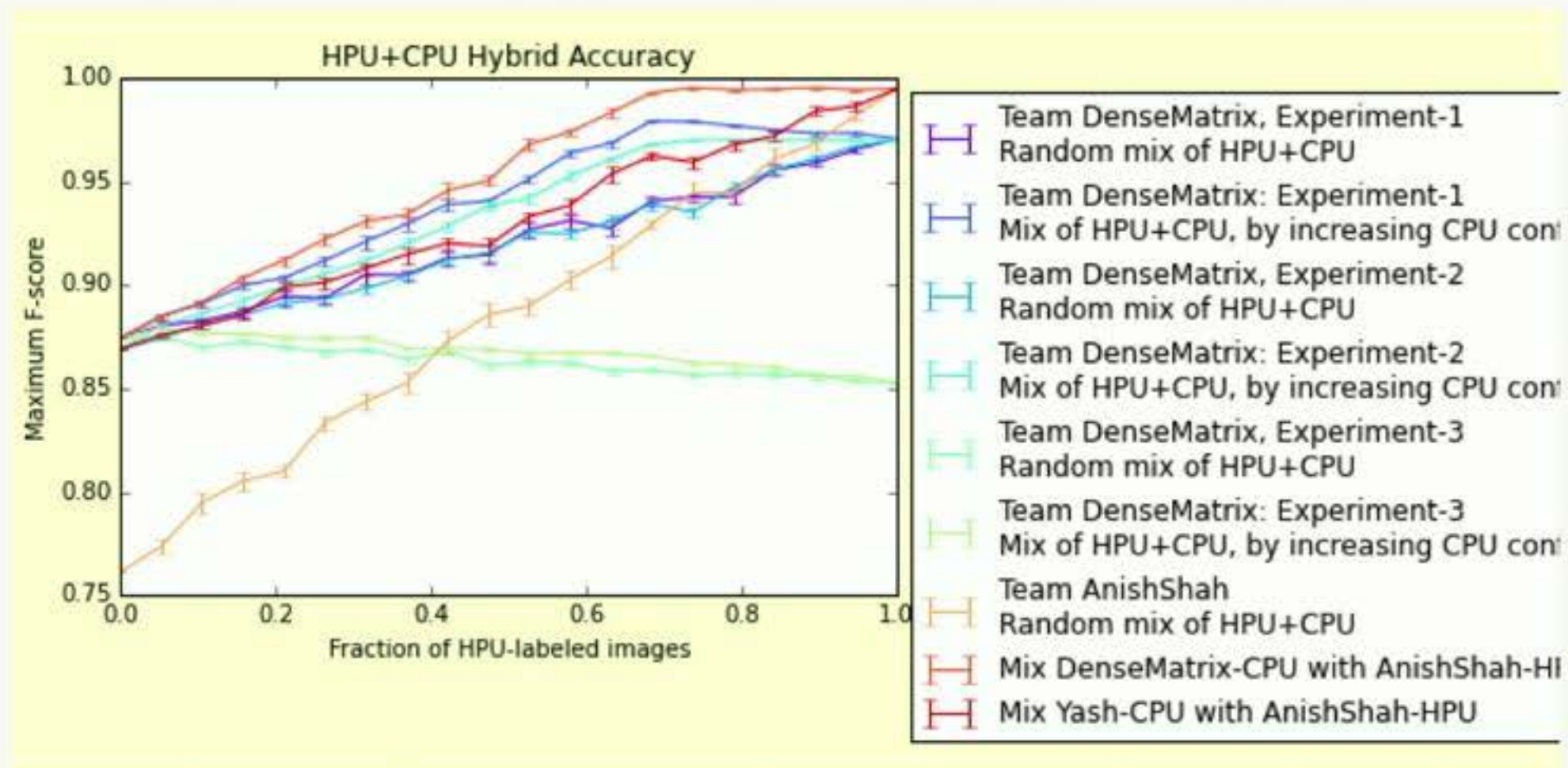
Serge Belongie,

Cornell Tech

Hybrid crowd-

computer vision

algorithms



THREE PARALLEL PROJECTS

Data science

Sharad Goel,
Stanford

Hundreds of
experiments
testing the wisdom
of the crowd

Predict the direction of penalty shot



Options:

- left
- right

Time Remaining

30

Tasks Remaining in
the domain: 15 / 20



CROWDSOURCING PROCESS

open call



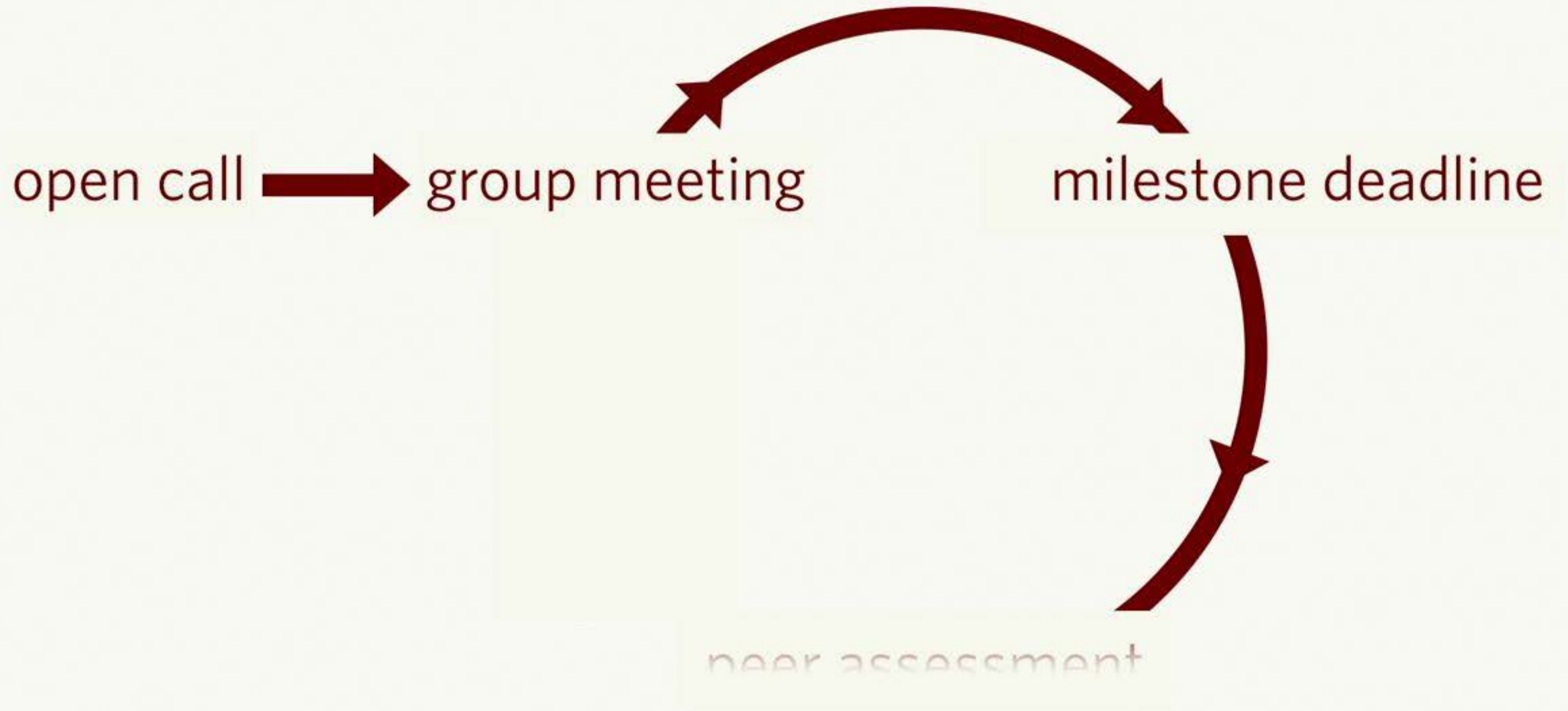
CROWDSOURCING PROCESS

open call  group meeting

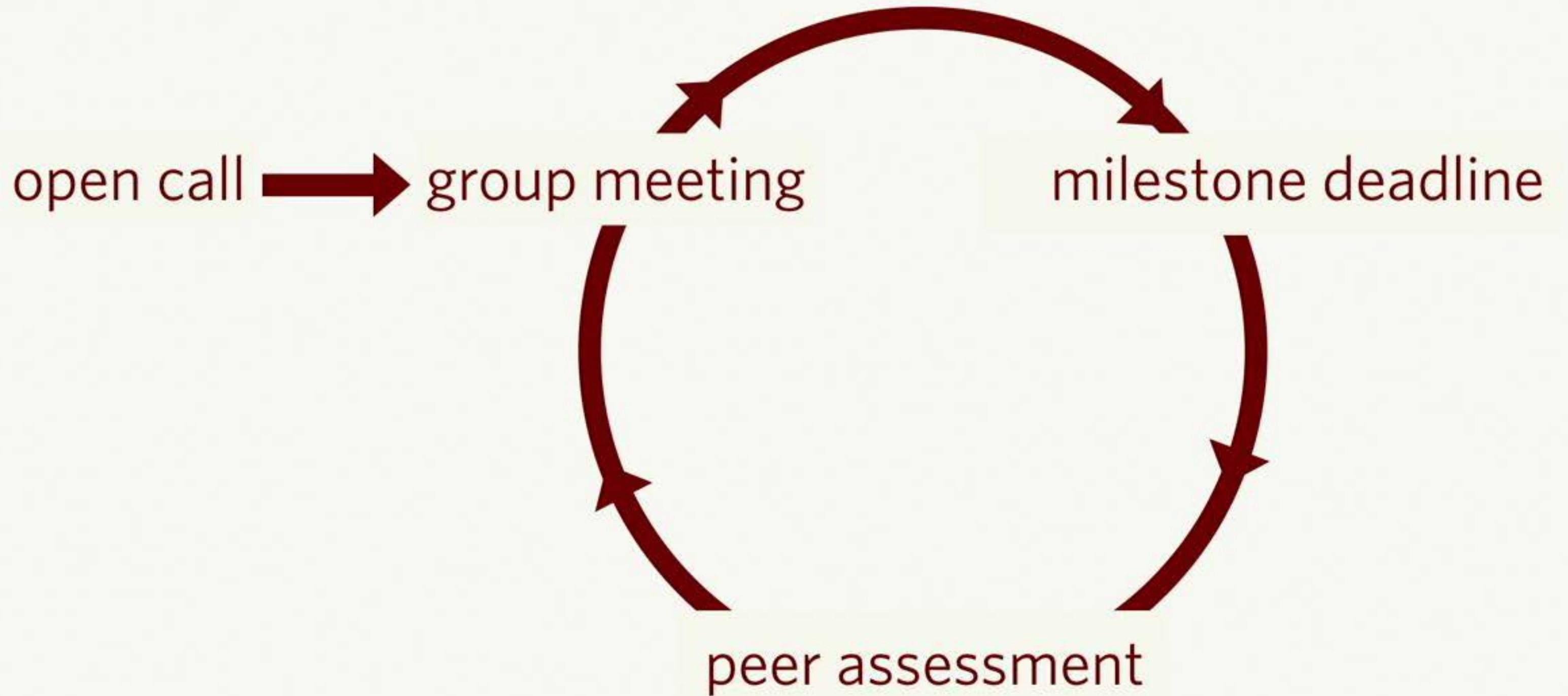
CROWDSOURCING PROCESS

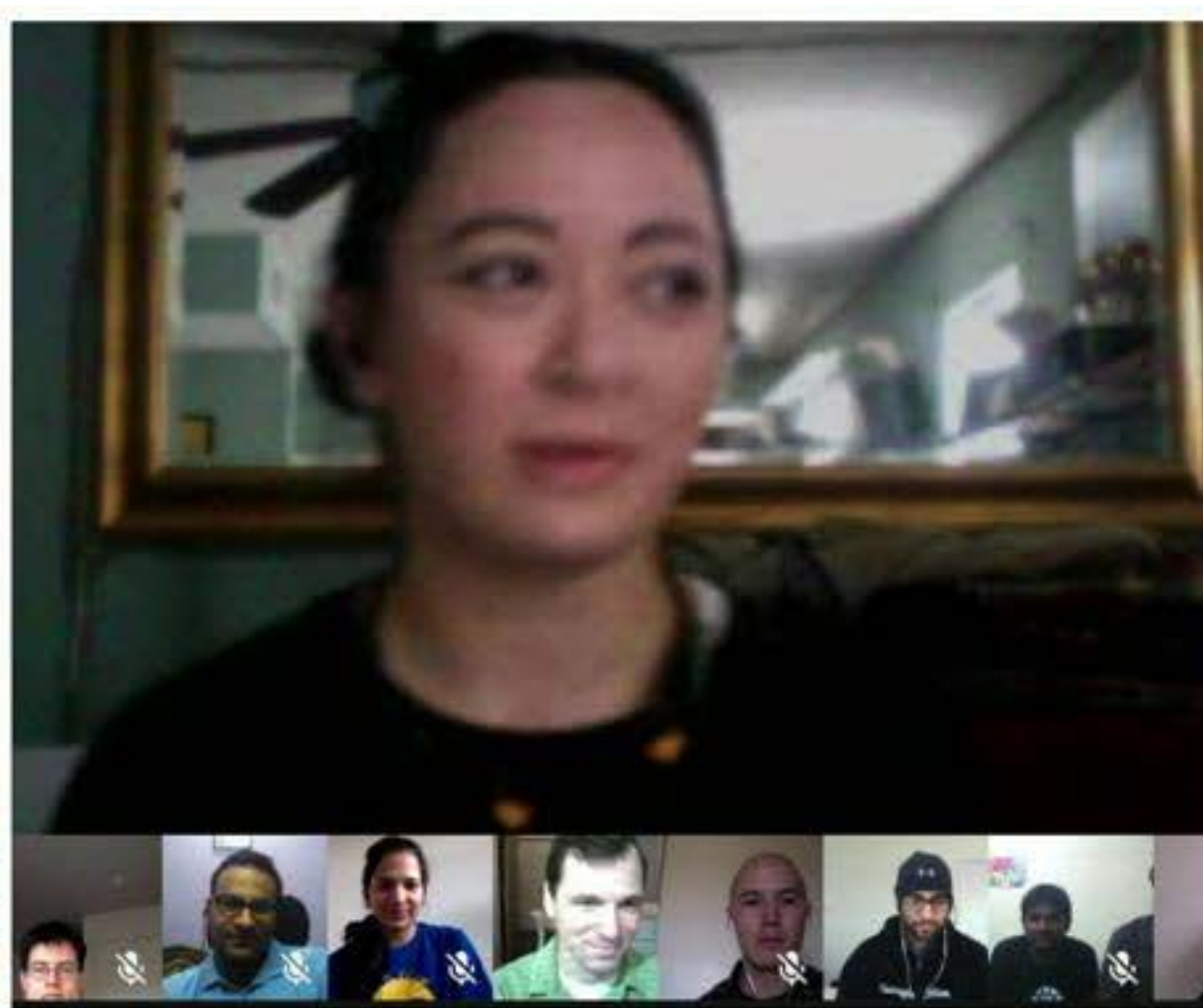







CROWDSOURCING PROCESS





CROWDSOURCING PROCESS





-  **rajanvaish** 9:01 AM
hello everyone!
-  **seasondyb** 9:02 AM
Hi!
Greetings from Seattle!
-  **csarasua** 9:02 AM
Hi!
-  **sbagroy986** 9:02 AM
hey!
-  **sujeathpareddy** 9:02 AM
Hi!



-  **meenalmandil** 9:02 AM
hi ^ ^
-  **ossolorzano** 9:02 AM
hello

12-85 weeks
500,000 Slack messages
190,000 minutes of video meetings

TASK PLANNING

MILESTONES

The screenshot shows a web application interface with a blue header and four main content panels. The header includes a 'Getting Started' button, 'Crowd Research', and 'Org Visible' options. The panels contain various tasks and milestones, each with a title, description, and a grid of user avatars. A 'May 28' date indicator is present on each task card.

Getting Started Crowd Research ☆ Org Visible

Click first: how does this work? (3 comments) May 28

Each week, you (and/or your team) sign up for at least one milestone here on Trello. See more...

Understanding lives of workers (13 comments) May 28

Try being a worker on oDesk (now Upwork.com) or other large project platforms.

Hello, world! Getting started with our code (1 comment) May 28

Work on one of our open feature requests on GitHub

Related work/papers: read and comment

Read the MobileWorks paper (17 comments) May 28

Read Flash Teams paper (5 comments) May 28

Read paper on the future of crowdwork

Category	Value	Category	Value
Approved MTs	\$1.00	Earnings Available	\$1.27
Bonus	\$0.00		
Total Earnings	\$1.00		

Task	Submitted	Approved	Rejected	Pending	Earnings
Mo. 27, 2012	0	0	0	0	\$1.00

MTs This Week Submitted	Value	Rate
MTs Submitted	0	-
Approved	0	100.0%
Rejected	0	0.0%
Pending	0	-

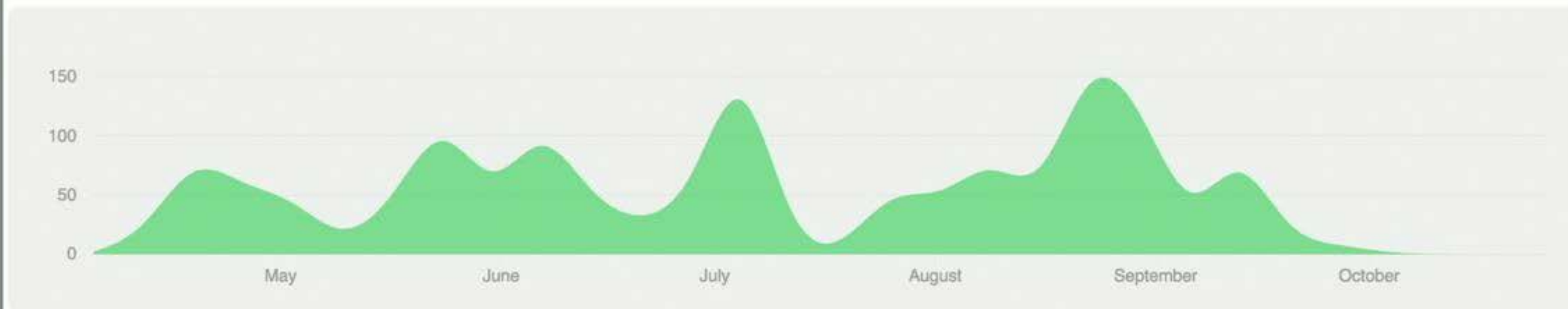
ENGINEERING

MILESTONES

Apr 5, 2015 – Oct 26, 2015

Contributions: **Commits** ▾

Contributions to develop2, excluding merge commits



dmorina

502 commits / 141,296 ++ / 365,266 --

#1



nistala

249 commits / 1,671,624 ++ / 1,442,878 --

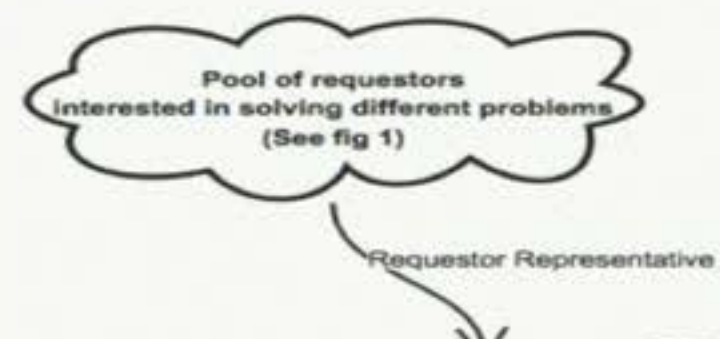
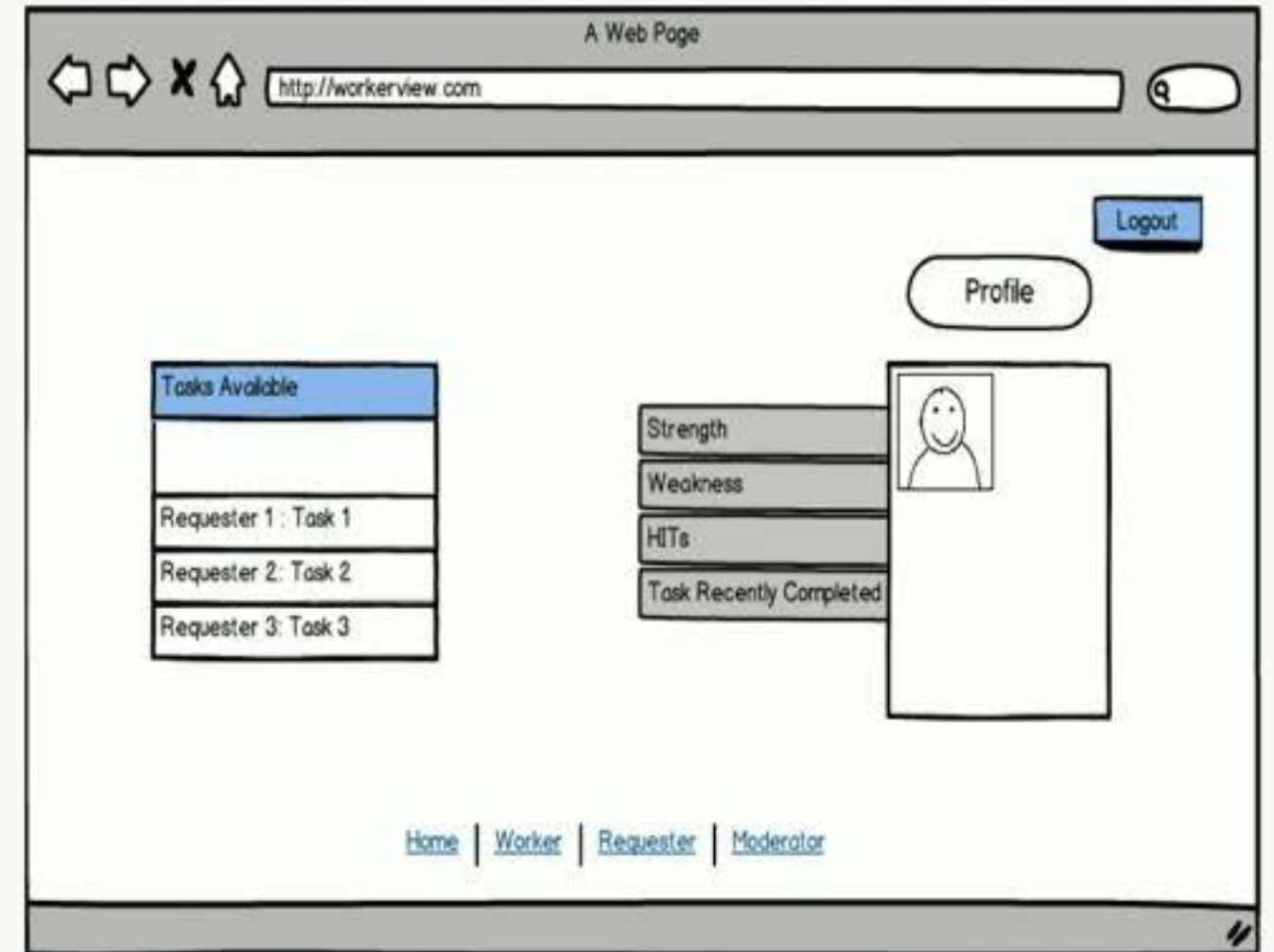
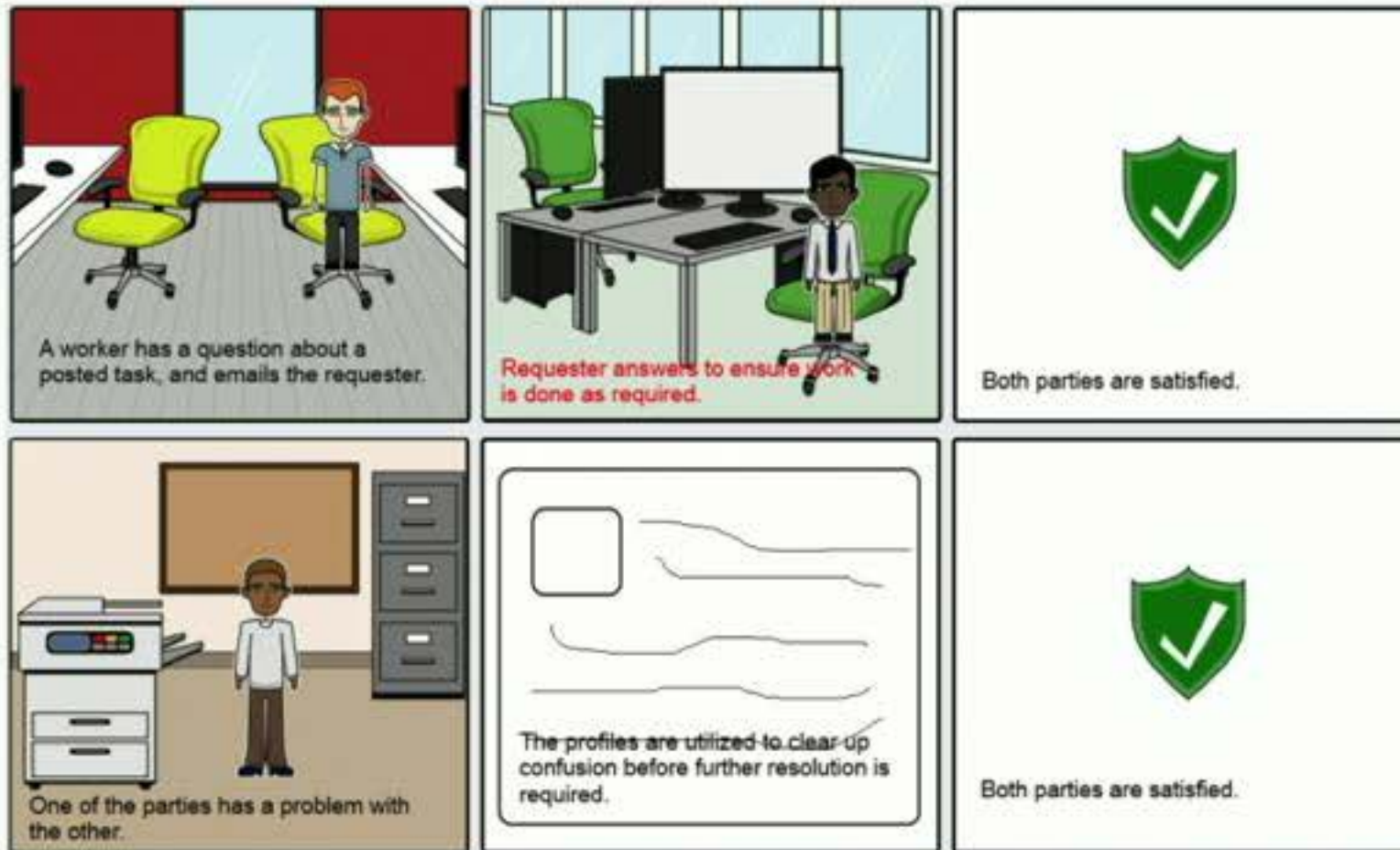
#2

50

50

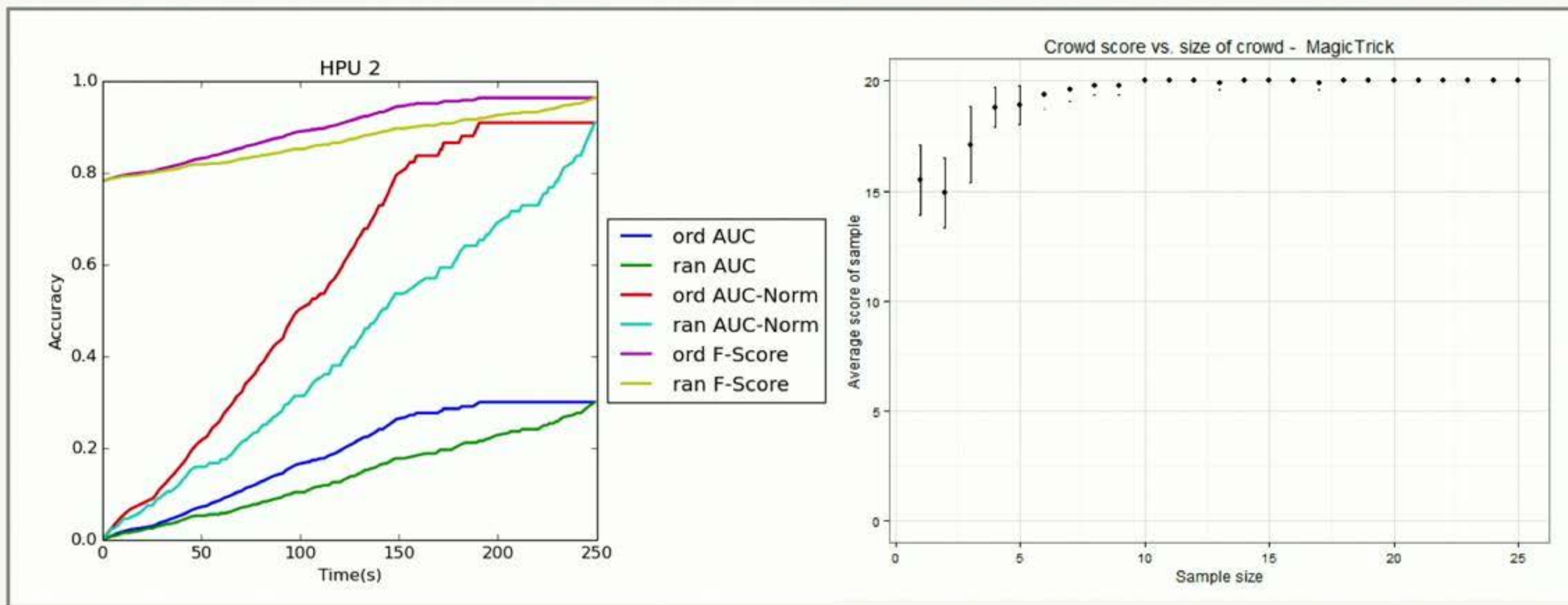
PROTOTYPING

MILESTONES



DATA ANALYSIS

MILESTONES



WRITING

MILESTONES

1. Anyone can pitch an idea. If it gets enough support, it goes to the next election and needs majority support from both workers+requesters.
 - **(original) Direct democracy:** anyone can pitch a policy idea, and once it gets past a threshold of support (e.g., 1000 votes), it goes up on a ballot. **Twice a year**, ideas go out to a direct vote for everyone on the platform. **If it gets majority support from both workers and requesters**, it passes.
2. Members get elected as worker or requester representatives (3 each) to a panel. Tiebreaking from a 7th member (jointly elected president).
 - **(original) Representative democracy:** once a year, members of the platform can be elected as **either worker or requester representatives for a small panel** (e.g, six people). Anybody can pitch a policy idea, and once it gets past a threshold of support (e.g., 1000 votes), the **elected representatives** must discuss it and vote on it.
3. Wikimocracy: the site's rules and policies are a wiki. Anyone can discuss, and if they edit, **policies change directly**.
4. Any idea that gets enough support enters a public one-month voting period. It's completely voluntary to vote. (Like a Kickstarter campaign.)
 - **Original: Fast-paced referendums:** similar concept as direct democracy, but instead of per year, you do it as vote thresholds within a month (within time of posting), and it's completely voluntary to vote. Kinda like a campaign on kickstarter. Fast pace and flexible deadlines will help the ideas continually flowing in.
5. For low-level changes, highlight the interface and suggest changes directly. Upvote/downvote directly on the interface.

majority of workers and requesters and not only one side. this could help balancing the platform.



Adam Marcus

5:05 AM May 9

Resolve

equal representation of workers and requesters? pro: seems fair, con: might run into the same sorts of paralysis issues the FEC is in now (<http://mobile.nytimes.com/2015/05/03/us/politics/fec-cant-curb-2016-election-abuse-commission-chief-says.html>)



Saloni Kogta

11:52 AM May 13
























Resolve

I am not sure how "fair" these elections would be. Money and power could play a major role here. I may be referring to a case that has extremely small possibility of occurring, but, what if the intentions of the elected members are changed or are influenced by some other party?

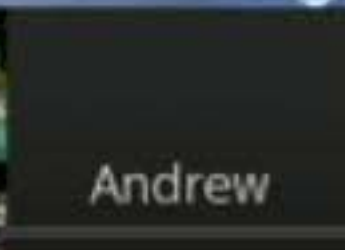
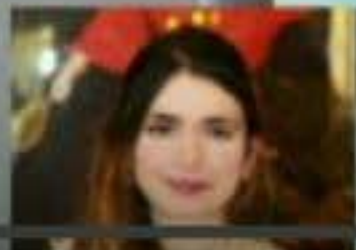
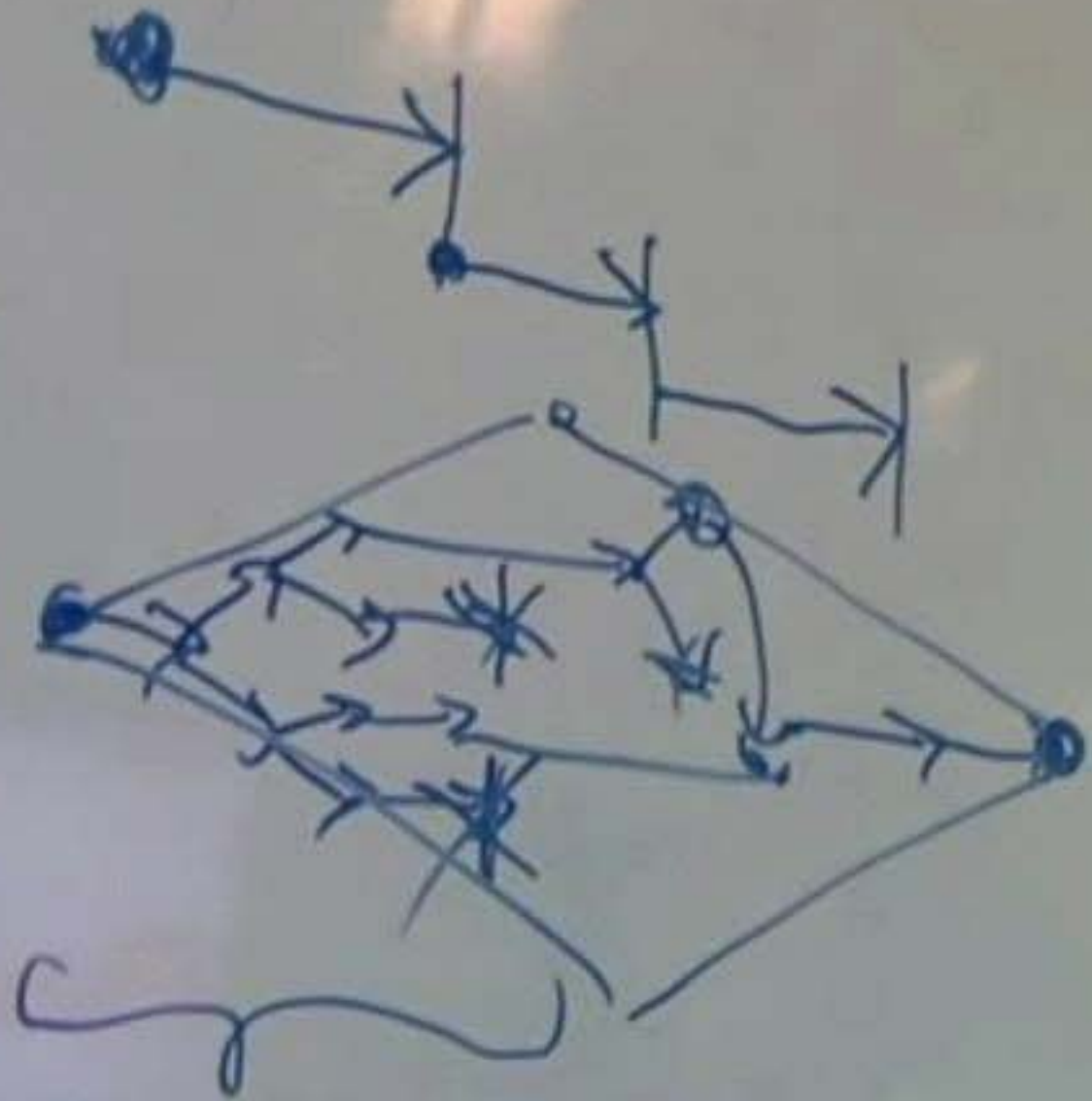
Reply...

PEER ASSESSMENT

MILESTONES

	Upvote	 Sustainable Reputation Mechanism MILESTONE-5-MOCKUP Neil 7 points 2 months ago 3 Comments		 comments
	Upvote	 Chat and Rating Prototypes for Crowdfunding Social Media Platform MILESTONE-5-STORYBOARD Kristine Hoang 7 points 2 months ago 3 Comments		 comments
	Upvote	 Leveling, Rating and Categorisation MILESTONE-5-MOCKUP Soroosh Bateni 5 points 2 months ago 2 Comments		 comments
	Upvote	 Worker Requester Mentorship MILESTONE-5-STORYBOARD Mike Young 5 points 2 months ago 4 Comments		 comments
	Upvote	 EmpathySociety: Triggering Empathy via Smart Mechanisms! MILESTONE-5-STORYBOARD Saiph Savage 4 points 2 months ago 5 Comments		 comments
	Upvote	Compensation Suggestion Tool MILESTONE-5-MOCKUP Trygve Cossette 4 points 2 months ago 2 Comments		 comments

Andrew Ng, Stanford



Peter Norvig, Google



Anant Agarwal, MIT and EdX





1500 participants from six continents
2% high school, 73% undergrad, 22% master's, 3% PhD

RECRUITMENT: PROVIDING ACCESS

Matching affiliations to Times Higher Education Global Rankings: **75% come from universities ranked below 500.**

Considering country of origin: **66% come from countries ranked over 50 in GDP per capita**



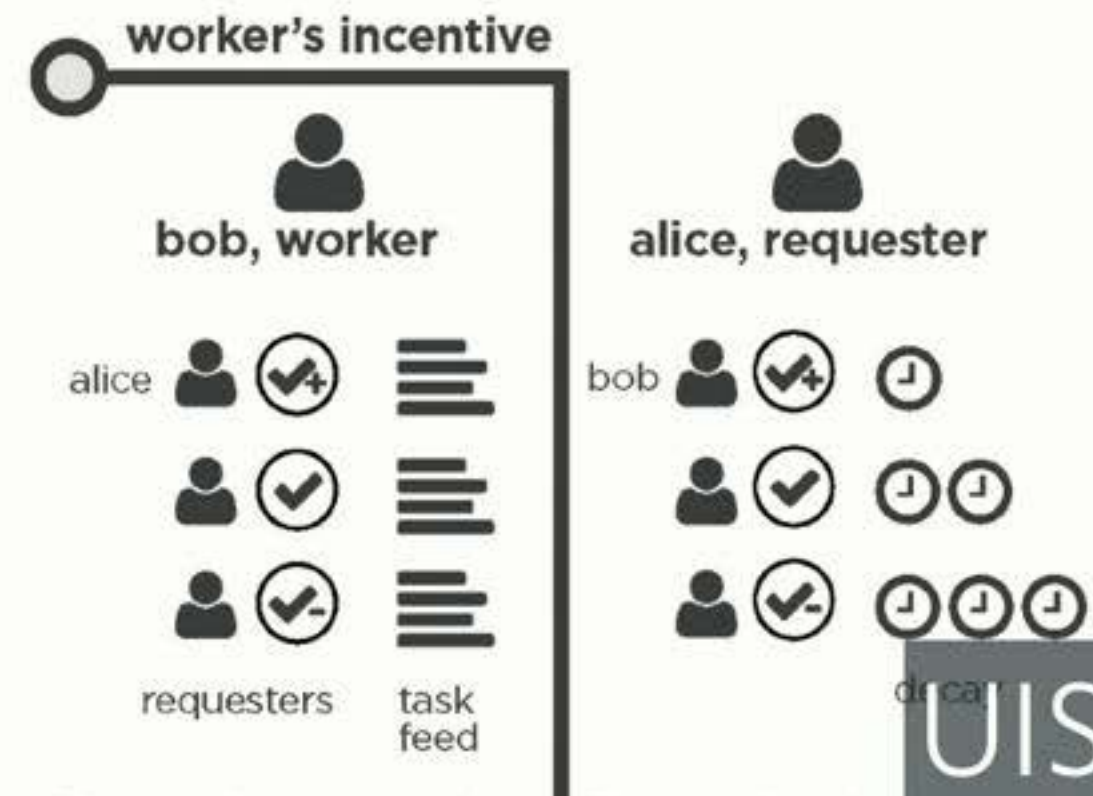
Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms

S.S. Gaikwad, D. Morina, A. Ginzberg, C. Mullings, S. Goyal, D. Gamage, C. Diemert, M. Burton, S. Zhou, M. Whiting, K. Ziulkoski, A. Ballav, A. Gilbee, S.S. Niranga, V. Sehgal, J. Lin, L. Kristianto, A. Richmond-Fuller, J. Regino, N. Chhibber, D. Majeti, S. Sharma, K. Mananova, D. Dhakal, W. Dai, V. Purynova, S. Sandeep, V. Chandrakanthan, T. Sarma, S. Matin, A. Nassar, R. Nistala, A. Stolzoff, K. Milland, V. Mathur, R. Vaish, M.S. Bernstein

Stanford Crowd Research Collective, Stanford University
daemo@cs.stanford.edu

ABSTRACT

Paid crowdsourcing platforms suffer from low-quality work and unfair rejections, but paradoxically, most workers and requesters have high reputation scores. These inflated scores, which make high-quality work and workers difficult to find, stem from social pressure to avoid giving negative feedback. We introduce Boomerang, a reputation system for crowdsourcing platforms that elicits more accurate feedback by rebounding the consequences of feedback directly back onto the person who gave it. With Boomerang, requesters find that their highly-rated workers gain earliest access to their future tasks, and workers find tasks from their highly-rated requesters at the top of their task feed. Field experiments verify that Boomerang



Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms

Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Aipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, Michael S. Bernstein
Stanford Crowd Research Collective
daemo@cs.stanford.edu

ABSTRACT

Crowd workers are distributed and decentralized. While decentralization is designed to utilize independent judgment to promote high-quality results, it paradoxically undercuts behaviors and institutions that are critical to high-quality work. Reputation is one central example: crowdsourcing systems depend on reputation scores from decentralized workers and requesters, but these scores are notoriously inflated and uninformative. In this paper, we draw inspiration from historical worker guilds (e.g., in the silk trade) to design and implement



WORKS-IN-PROGRESS

Daemo: a Self-Governed Crowdsourcing Marketplace

Stanford Crowd Research Collective^{*}
Stanford HCI Group daemo@cs.stanford.edu

ABSTRACT

Crowdsourcing marketplaces provide opportunities for autonomous and collaborative professional work as well as social engagement. However, in these marketplaces, workers feel disrespected due to unreasonable rejections and low payments, whereas requesters do not trust the results they receive. The lack of trust and uneven distribution of power among workers and requesters have raised serious concerns about sustainability of these marketplaces. To address the challenges of trust and power, this paper introduces Daemo, a self-governed crowdsourcing marketplace. We propose a prototype task to improve the work quality and open-governance model to achieve equitable representation. We envisage Daemo will enable workers to build sustainable careers and provide requesters with timely, quality labor for their businesses.

Author Keywords

crowdsourcing, crowd research, crowd work

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Computer-supported cooperative work

INTRODUCTION

Paid crowdsourcing marketplaces such as Mechanical Turk and Upwork have created opportunities for workers to supplement their income and enhance their skills, while allowing requesters to get their work completed efficiently. These marketplaces have attracted many participants globally; however, they have repeatedly failed to ensure high-quality results, fair wages, respect for workers, and convenience in authoring effective tasks [1].

^{*}This project was created via a world-wide, crowdsourced research process initiated at Stanford University: S. Gidycz, D. Morina, R. Notala, M. Agarwal, A. Cozzette, R. Bhano, S. Savage, V. Narwal, K. Rajpal, J. Regino, A. Mihal, A. Ginzberg, A. Nath, K. R. Zelenko, T. Cozzette, D. Ganage, A. Richmond-Feller, R. Soroki, J. Herrejon, K. V. Le, C. Flores-Santiago, H. Thakuram, K. Gupta, W. Dai, A. Sastry, S. Goyal, T. Rajapakse, N. Abolhasan, A. Xie, A. Reyes, S. Ingle, V. Jaramillo, M.D. Gomez, W. Angel, M. Gomez, C. Tivoli, J. Flores, A. Gupta, V. Sethia, D. Padilla, K. Midland, K. Seiyadi, N. Wejnarski, M. Batagoda, R. Cruz, J. Danson, D. Nekkan, T. Sarma, M.H. Saleh, G. Gopjes-Svartman, S. Bhatia, G. Toledo-Barra, A. Pena, R. Compton, D. Aarif, L. Palacios, M. P. Ritter, Nisha K.K., A. Kay, J. Urmesler, S. Notala, M. Fofsham, E. Bakis, C. Diemert, L. Matsunoto, M. Singh, V. Jaramillo-Lopez, K. Patel, R. Krishna, G. Kovacs, R. Vash, M. Bernstein

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author. Copyright is held by the owner/authors. 0327-7724/November 08-11, 2015, Charleston, SC, USA. ACM 978-1-4503-2817-1/15. http://dx.doi.org/10.1145/2817785.2817779



Figure 1. Task creation workflow for a requester: prototype task creation, initial submission review, and hiring high quality workers for future submissions. [http://daemo.stanford.edu]

From our interviews with requesters, it has become clear that they struggle to trust their workers. They will rerun tasks, discard gathered data, and add increasingly complex worker filters. On the other hand, workers do not trust requesters to follow through with pay and fair treatment. In response, workers often withhold their full effort unless they have an experience with the requester.

Moreover, existing marketplaces suffer from uneven distributions of power [4]. For example, requesters have the power to deny payments for finished tasks and workers have inadequate means to contest this. Operational governance and rules have been secondary considerations on markets thus far, tilted to support the focus on the commoditizing of work. This resulted in an asymmetrical relationship between workers, requesters, and the marketplace on fronts such as parity of information access, wage negotiation, and reputation. A common complaint [3] “We can be rejected yet the requesters still have our articles and sentences. Not fair”

We present Daemo, a crowd-built, self-governed crowdsourcing marketplace. To increase trust, we introduce the idea of prototype tasks, where each new task must first launch in an intermediate feedback mode where workers can comment on the task, requesters can review the submissions and qualify a subset of workers to continue. During this phase, workers and requesters work together to refine the task description and reduce errors. Daemo also adopts a representative democratic governance model to elect a leadership board. Engaging all vested parties in the governance of the marketplace gives an opportunity to create genuine worker-requester relationships and redefine the future of work.

RELATED WORK

Feedback, wages, task decomposition, and quality control are some of the fundamental elements of a successful crowdsourcing marketplace [1]. Requesters often rely on “gold standard” tasks, i.e., questions with known answers, to evaluate the performance and quality of submissions [2]. However,

On Optimizing Human-Machine Task Assignments

Organizers: Andreas Veit, Michael Wilber, Rajan Vaidh, Sergey Belongie, James Davis,
Top researchers: Vishal Anand, Anshu Arora, Prithvijit Chakrabarty, Yash Chaudak, Sudharth Chaturvedi, Chintaya Devanaj,
Ankit Dhall, Utkarsh Dwivedi, Sandip Gupta, Sharath N. Sridhar, Karthik Puga, Anuj Pabreja, Aditya Rameshbabu,
Ajaykumar Sharma, Shweta Sharma, Darpana Saha, Nisarg Thakkar, K. Balu Vignesh, Utkarsh Vinita,
Researchers: Kamniganti Abhishek, Anand Agrawal, Arya Anilwarya, Argho Bhattacharjee, Sarveshwari Dhanasekar,
Venkata Karthik Gullapalli, Shuchita Gupta, Chandana G. Kirpal Jam, Suresh Kapur, Meghana Kasarla,
Shashi Kumar, Parth Kumbharia, Utkarsh Mahur, Alankrit Mishra, Aayush Madgal, Aditya Nandipati,
Manakala Sriv Sibat, Akanksha Perival, Ayush Sagar, Ayush Shah, Vikas Sharma, Yashvardhan Sharma,
Faisal Siddiqui, Virender Singh, Abhinav S., Pradyumn Tamburkar, Rashida Tasika, Ankit Tripathi,
Anurag D Yadav

Abstract

When crowdsourcing systems are used in combination with machine inference systems in the real world, they benefit the most when the machine system is deeply integrated with the crowd workers. However, if researchers wish to integrate the crowd with “off-the-shelf” machine classifiers, this deep integration is not always possible. This work explores two strategies to increase accuracy and decrease cost under this setting. First, we show that ordering tasks presented to the human can create a significant accuracy improvement. Further, we show that greedily choosing parameters to maximize machine accuracy is sub-optimal, and joint optimization of the combined system improves performance.

Introduction

When crowdsourcing systems are deployed in the real world, the goal is often to maximize accuracy at a fixed price point or to minimize cost at a certain accuracy requirement. The best way to do this is by tightly integrating the machine and crowd worker within the overall end-to-end pipeline. For instance, the machine computation might use worker annotations as a prior to influence its results, or tasks for workers might be chosen and ordered adaptively using a Markov Decision Process (Rassakovsky, Li, and Fei-Fei 2015).

However, this tight integration is not always possible. Many real systems only provide outputs and cannot be heavily modified. In these cases, the use of crowd workers is often restricted to a post-process that attempts to correct errors in the machine computation. In this scenario, what kinds of strategies can maximize accuracy while minimizing costs?

To explore this question, we choose a representative task within the domain of computer vision: localizing objects in a large dataset. The goal is to detect all instances of certain objects of interest in the dataset. Machine systems can take images as input and automatically generate bounding boxes around objects of interest. Internal to the machine algorithm, to classify a potential detection as an object of interest or not, the algorithm employs a detection threshold such that only detections with confidence scores above the threshold

^{*}This project was created via a world-wide, crowdsourced research process initiated by UC Santa Cruz, Stanford University, and Cornell Tech. Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

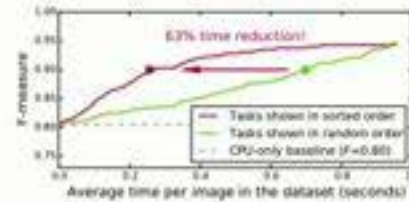


Figure 1. Consider a simple localization task where crowd workers refine the output of a machine classifier. At a threshold of 0.5, baseline accuracy starts at 0.80 (grey dotted line). If we show random tasks to human workers, accuracy improves (green), but if we order tasks by increasing machine confidence (purple), we can reduce the time requirement dramatically at a given target accuracy.

are returned. Finding many correct objects implies also detecting many false positives. Because the detection threshold determines this tradeoff, it is often treated as the primary tunable parameter of machine vision algorithms. The returned detections are then given to human workers, who we employ to remove false detections. For our experiments, we adopt the classic UIUC-Cars dataset (Agarwal, Awan, and Roth 2004). As detector, we use Support Vector Machines trained on Histograms of Ordered Gradients as a representative “out-of-the-box” machine vision system.

Our objective is to maximize the overall accuracy of the machine-crowd pipeline on the dataset given a certain time budget. We vary the time budget by presenting the humans with only a fraction of all detections. If humans look at a large fraction of detections the accuracy improvement will be large, however the average time cost per image in the dataset will also be large. If humans look at only a few images, the average accuracy of the entire dataset will show little improvement, but the time cost will be low. We plot the tradeoff between cost and accuracy as a curve.

The primary contribution of this work is a description and analysis of two strategies for improving the cost-accuracy curve. In Task Ordering we consider the impact of using the machine vision algorithm’s confidence score as a way to order human tasks. In Joint Optimization we consider how changing the machine threshold parameter impacts results.

Investigating the “Wisdom of Crowds” at Scale

Alok Shankar Mysore^{*}
PES Institute of Technology
alok.shankar@pesit.pes.edu

Vikas S Yaligar¹
National Institute of
Technology Karnataka
vikasyaligar@ieee.org

Imanol Arrieta Ibarra
Stanford University
imanol@stanford.edu

Camelia Simoiu
Stanford University
csimoiu@stanford.edu

Sharad Goel
Stanford University
sgoel@stanford.edu

Additional Authors^{*}
Various Institutions¹

ABSTRACT

In a variety of problem domains, it has been observed that the aggregate opinions of groups are often more accurate than those of the constituent individuals, a phenomenon that has been termed the “wisdom of the crowd.” Yet, perhaps surprisingly, there is still little consensus on how generally the phenomenon holds, how best to aggregate crowd judgements, and how social influence affects estimates. We investigate these questions by taking a meta wisdom of crowds approach. With a distributed team of over 100 student researchers across 17 institutions in the United States and India, we develop a large-scale online experiment to systematically study the wisdom of crowds effect for 1,000 different tasks in 50 subject

domains. These tasks involve various types of knowledge (e.g., explicit knowledge, tacit knowledge, and prediction), question formats (e.g., multiple choice and point estimation), and inputs (e.g., text, audio, and video). To examine the effect of social influence, participants are randomly assigned to one of three different experiment conditions in which they see varying degrees of information on the responses of others. In this ongoing project, we are now preparing to recruit participants via Amazon’s Mechanical Turk.

Author Keywords

Crowdsourcing; online experiment; crowd consensus

ACM Classification Keywords

H.5.m. Economics: Experimentation Design

INTRODUCTION

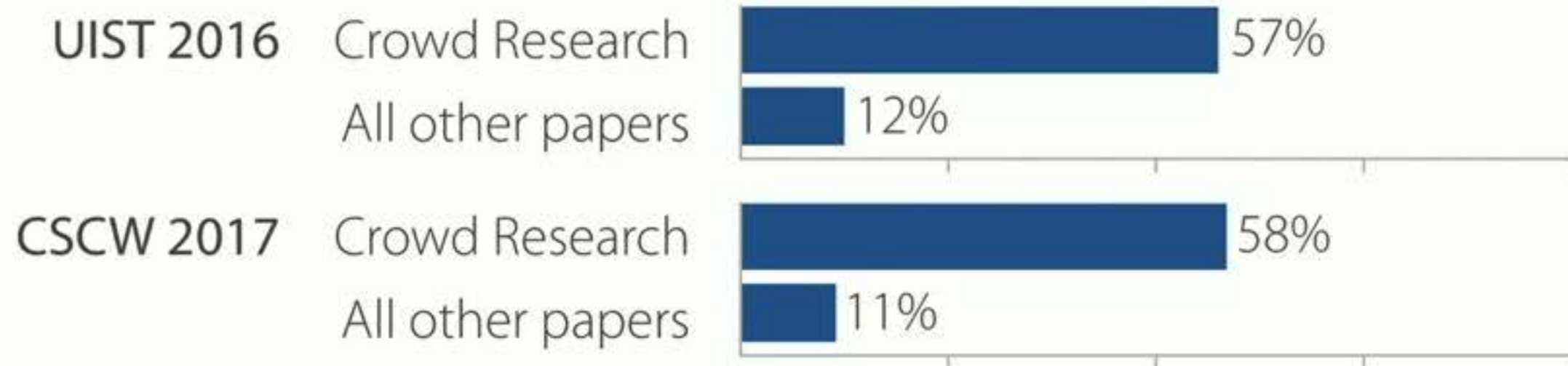
At a 1906 county fair, the statistician Francis Galton watched as eight hundred people competed to guess the weight of an ox. He famously observed that the median of the guesses, 1,207 pounds, was, remarkably, within 1% of the true weight [1].

Simple aggregation—as in the case of Galton’s ox competition, or voting in democratic elections—has been shown to be a surprisingly powerful technique for prediction, inference, and decision-making. Over the last century, there have been dozens of studies that examine this wisdom of crowds effect. For example, crowd judgements have been used to identify phishing websites [6], answer general knowledge questions [5], and forecast weather-related events [3]. In these applications, a wide variety of aggregation methods have been considered, ranging from standard measures, such as the mean and median, to more specialized, domain-specific techniques, such as those based on cognitive models of decision making [4]. However, given the diversity of experimental designs, subject pools, and analytic methods employed, it has proven difficult to compare studies and extract general principles. It is thus unclear whether these documented examples are a representative collection of a much larger space of tasks that exhibit a wisdom of crowds phenomenon, or conversely, whether they are highly specific instances of an interesting, though ultimately limited occurrence.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author. Copyright is held by the owner/authors. 0327-7724/November 08-11, 2015, Charleston, SC, USA. ACM 978-1-4503-2817-1/15. http://dx.doi.org/10.1145/2817785.2817779

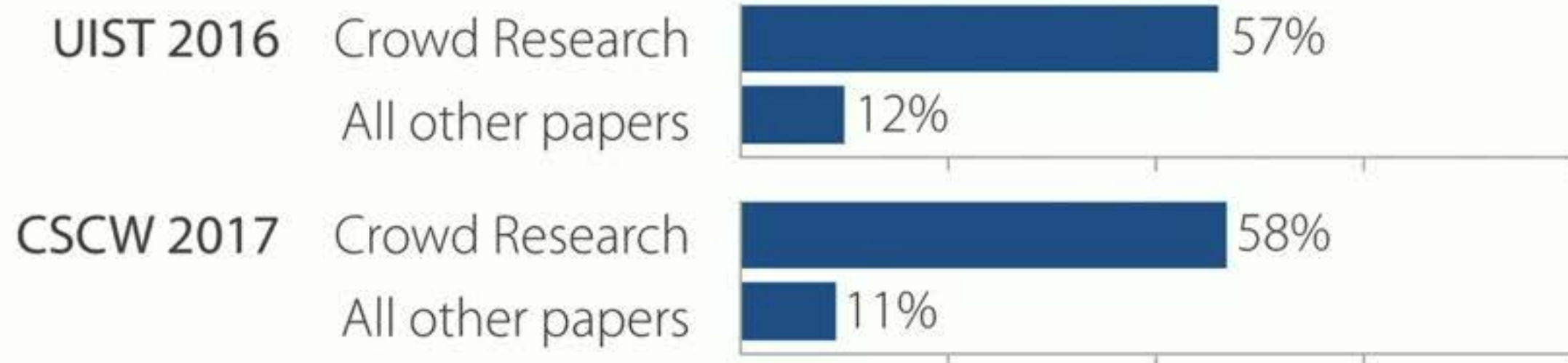
INCREASED ACCESS TO RESEARCH

Coauthors' universities that are ranked below 500 worldwide

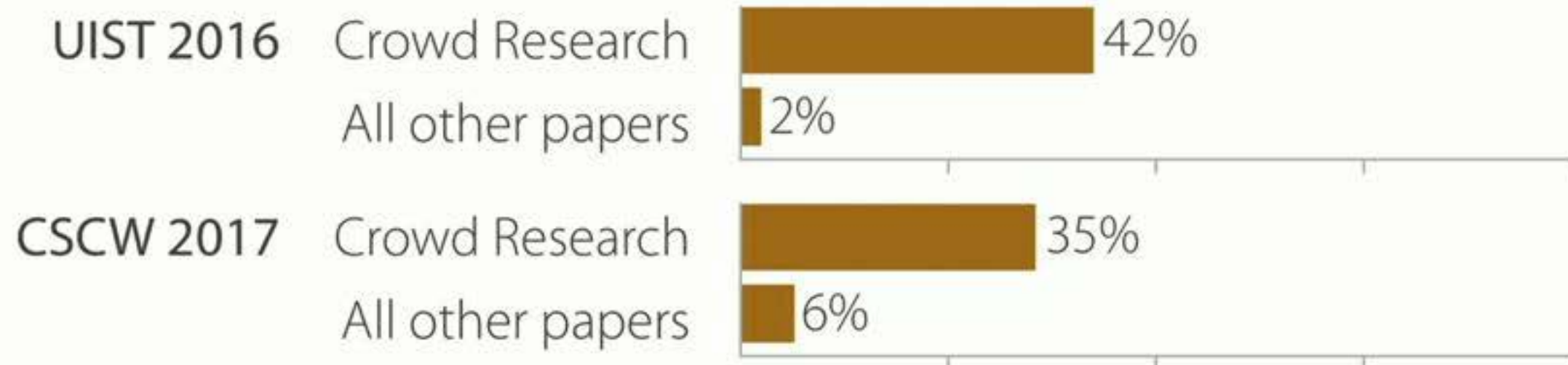


INCREASED ACCESS TO RESEARCH

Coauthors' universities that are ranked below 500 worldwide



Coauthors whose countries are ranked below 50 worldwide in GDP per capita



ATTAINMENT: PROVIDING ACCESS

Participants have gone on to programs at UC Berkeley, and Carnegie Mellon University, and MIT.

21 of 33 surveyed were admitted to at least one program, despite a **median of 0 other letter writers** from institutions ranked better than 500 worldwide.

Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms

Mark E. Whiting, Dilrukshi Gamage, Snehalkumar (Neil) S. Gaikwad, Aaron Gilbee, Shirish Goyal, Aipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, Tejas Seshadri Sarma, Varshine Chandrakanthan, Teogenes Moura, Mohamed Hashim Salih, Gabriel Bayomi Tinoco Kalejaiye, Adam Ginzberg, Catherine A. Mullings, Yoni Dayan, Kristy Milland, Henrique Orefice, Jeff Regino, Sayna Parsi, Kunz Mainali, Vibhor Sehgal, Sekandar Matin, Akshansh Sinha, Rajan Vaish, Michael S. Bernstein
Stanford Crowd Research Collective
daemo@cs.stanford.edu

ABSTRACT

Crowd workers are distributed and decentralized. While decentralization is designed to utilize independent judgment to promote high-quality results, it paradoxically undercuts behaviors and institutions that are critical to high-quality work. Reputation is one central example: crowdsourcing systems depend on reputation scores from decentralized workers and requesters, but these scores are notoriously inflated and uninformative. In this paper, we draw inspiration from historical worker guilds (e.g., in the silk trade) to design and implement



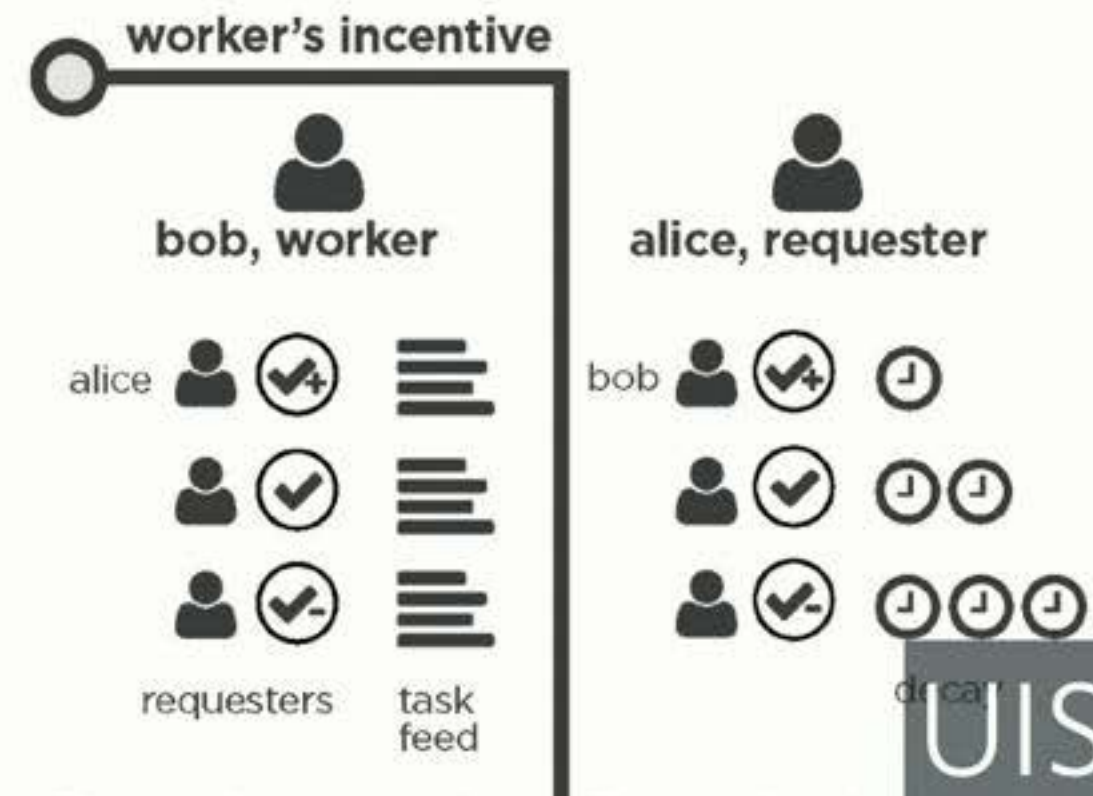
Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms

S.S. Gaikwad, D. Morina, A. Ginzberg, C. Mullings, S. Goyal, D. Gamage, C. Diemert, M. Burton, S. Zhou, M. Whiting, K. Ziulkoski, A. Ballav, A. Gilbee, S.S. Niranga, V. Sehgal, J. Lin, L. Kristianto, A. Richmond-Fuller, J. Regino, N. Chhibber, D. Majeti, S. Sharma, K. Mananova, D. Dhakal, W. Dai, V. Purynova, S. Sandeep, V. Chandrakanthan, T. Sarma, S. Martin, A. Nassar, R. Nistala, A. Stolzoff, K. Milland, V. Mathur, R. Vaish, M.S. Bernstein

Stanford Crowd Research Collective, Stanford University
daemo@cs.stanford.edu

ABSTRACT

Paid crowdsourcing platforms suffer from low-quality work and unfair rejections, but paradoxically, most workers and requesters have high reputation scores. These inflated scores, which make high-quality work and workers difficult to find, stem from social pressure to avoid giving negative feedback. We introduce Boomerang, a reputation system for crowdsourcing platforms that elicits more accurate feedback by rebounding the consequences of feedback directly back onto the person who gave it. With Boomerang, requesters find that their highly-rated workers gain earliest access to their future tasks, and workers find tasks from their highly-rated requesters at the top of their task feed. Field experiments verify that Boomerang



WORKS-IN-PROGRESS

Daemo: a Self-Governed Crowdsourcing Marketplace

Stanford Crowd Research Collective^{*}
Stanford HCI Group daemo@cs.stanford.edu

ABSTRACT

Crowdsourcing marketplaces provide opportunities for autonomous and collaborative professional work as well as social engagement. However, in these marketplaces, workers feel disrespected due to unreasonable rejections and low payments, whereas requesters do not trust the results they receive. The lack of trust and uneven distribution of power among workers and requesters have raised serious concerns about sustainability of these marketplaces. To address the challenges of trust and power, this paper introduces Daemo, a self-governed crowdsourcing marketplace. We propose a prototype task to improve the work quality and open-governance model to achieve equitable representation. We envisage Daemo will enable workers to build sustainable careers and provide requesters with timely, quality labor for their businesses.

Author Keywords

crowdsourcing, crowd research, crowd work

ACM Classification Keywords

H.5.3. Group and Organization Interfaces: Computer-supported cooperative work

INTRODUCTION

Paid crowdsourcing marketplaces such as Mechanical Turk and Upwork have created opportunities for workers to supplement their income and enhance their skills, while allowing requesters to get their work completed efficiently. These marketplaces have attracted many participants globally; however, they have repeatedly failed to ensure high-quality results, fair wages, respect for workers, and convenience in authoring effective tasks [1].

^{*} This project was created via a world-wide, crowdsourced research process initiated at Stanford University. S. Gidycz, D. Morina, R. Notala, M. Agarwal, A. Cossette, R. Bhamu, S. Savage, V. Narwal, K. Rajpal, J. Rejzino, A. Mikhal, A. Ginzberg, A. Nath, K. R. Zelenko, T. Cossette, D. Ganage, A. Richmond-Fuller, R. Soroki, J. Herrojon, K. V. Le, C. Flores-Saviaga, H. Thakurathne, K. Gupta, W. Dai, A. Sotry, S. Goyal, T. Rajapakse, N. Abolhasan, A. Xie, A. Reyes, S. Ingle, V. Jaramillo, M.D. Gomez, W. Angel, M. Gomez, C. Torib, J. Flores, A. Gupta, V. Sethia, D. Padilla, K. Midland, K. Sreyach, N. Weerasena, M. Rajagoda, R. Cruz, J. Danon, D. Nekkiam, T. Sarma, M.H. Saleh, G. Gompouz-Svartzman, S. Bhatia, G. Toledo-Barraza, A. Pena, R. Compton, D. Aarif, L. Palacios, M. P. Ritter, Nisha K.K., A. Kay, J. Uhermeister, S. Notala, M. Fofsham, E. Bakis, C. Drennet, L. Matsunoto, M. Singh, V. Jaramillo-Lopez, K. Patel, R. Krishna, G. Kovacs, R. Vash, M. Bernstein

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
Copyright held by the owner/authors.
0327-7724/15/0000-0001..2015 ACM 978-1-4503-2817-1/15
http://dx.doi.org/10.1145/2817705.2817719



Figure 1. Task creation workflow for a requester: prototype task creation, initial submissions review, and hiring high quality workers for future milestones. (<https://daemo.stanford.edu>)

From our interviews with requesters, it has become clear that they struggle to trust their workers. They will rerun tasks, discard gathered data, and add increasingly complex worker filters. On the other hand, workers do not trust requesters to follow through with pay and fair treatment. In response, workers often withhold their full effort unless they have an experience with the requester.

Moreover, existing marketplaces suffer from uneven distributions of power [4]. For example, requesters have the power to deny payments for finished tasks and workers have inadequate means to contest this. Operational governance and rules have been secondary considerations on markets thus far, fitted to support the focus on the commoditizing of work. This resulted in an asymmetrical relationship between workers, requesters, and the marketplace on fronts such as parity of information access, wage negotiation, and reputation. A common complaint [3]: “We can be rejected yet the requesters still have our articles and sentences. Not fair.”

We present Daemo, a crowd-built, self-governed crowdsourcing marketplace. To increase trust, we introduce the idea of prototype tasks, where each new task must first launch in an intermediate feedback mode where workers can comment on the task, requesters can review the submissions and qualify a subset of workers to continue. During this phase, workers and requesters work together to refine the task description and reduce errors. Daemo also adopts a representative democratic governance model to elect a leadership board. Engaging all vested parties in the governance of the marketplace gives an opportunity to create genuine worker-requester relationships and redefine the future of work.

RELATED WORK

Feedback, wages, task decomposition, and quality control are some of the fundamental elements of a successful crowdsourcing marketplace [1]. Requesters often rely on “gold standard” tasks, i.e., questions with known answers, to evaluate the performance and quality of submissions [2]. However,

On Optimizing Human-Machine Task Assignments

Organizers: Andreas Veit, Michael Wilber, Rajan Vash, Sergey Belongie, James Davis
Top researchers: Vishal Anand, Anshu Arora, Prithvijit Chakrabarty, Yash Chaudak, Sidharth Chaturvedi, Chinmaya Devanaj, Ankur Dhall, Utkarsh Dwivedi, Sandeep Gupta, Sharath N. Sridhar, Karthik Puga, Anuj Pabreja, Aditya Rameshghani, Ayush Sharma, Shweta Sharma, Darpana Saha, Nisarg Thakkar, K. Bala Vignesh, Utkarsh Vinnia, Kamniganti Abhishek, Anand Agrawal, Arya Anilwarya, Aarsh Bhattacharjee, Sarveshwaran Dhanasekar, Venkata Karthik Gullapalli, Shuchita Gupta, Chandra G. Kirpal Jam, Suresh Kapur, Meghana Kavala, Shashi Kumar, Parth Kundalya, Utkarsh Malhotra, Alankrit Mishra, Aayush Madgal, Aditya Nandipati, Manakala Srve Sihar, Akanksha Prival, Ayush Sagar, Ayush Shah, Vikas Sharma, Yashvardhan Sharma, Faizal Siddiqui, Virender Singh, Abhinav S., Pradyumn Tambwekar, Rashida Taskar, Ankit Tripathi, Anurag D Yadav

Abstract

When crowdsourcing systems are used in combination with machine inference systems in the real world, they benefit the most when the machine system is deeply integrated with the crowd workers. However, if researchers wish to integrate the crowd with “off-the-shelf” machine classifiers, this deep integration is not always possible. This work explores two strategies to increase accuracy and decrease cost under this setting. First, we show that ordering tasks presented to the human can create a significant accuracy improvement. Further, we show that greedily choosing parameters to maximize machine accuracy is sub-optimal, and joint optimization of the combined system improves performance.

Introduction

When crowdsourcing systems are deployed in the real world, the goal is often to maximize accuracy at a fixed price point or to minimize cost at a certain accuracy requirement. The best way to do this is by tightly integrating the machine and crowd worker within the overall end-to-end pipeline. For instance, the machine computation might use worker annotations as a prior to influence its results, or tasks for workers might be chosen and ordered adaptively using a Markov Decision Process (Rusakovskiy, Li, and Fei-Fei 2015).

However, this tight integration is not always possible. Many real systems only provide outputs and cannot be heavily modified. In these cases, the use of crowd workers is often restricted to a post-process that attempts to correct errors in the machine computation. In this scenario, what kinds of strategies can maximize accuracy while minimizing costs?

To explore this question, we choose a representative task within the domain of computer vision: localizing objects in a large dataset. The goal is to detect all instances of certain objects of interest in the dataset. Machine systems can take images as input and automatically generate bounding boxes around objects of interest. Internal to the machine algorithm, to classify a potential detection as an object of interest or not, the algorithm employs a detection threshold such that only detections with confidence scores above the threshold

^{*} This project was created via a world-wide, crowdsourced research process initiated by UC Santa Cruz, Stanford University, and Cornell Tech.
Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

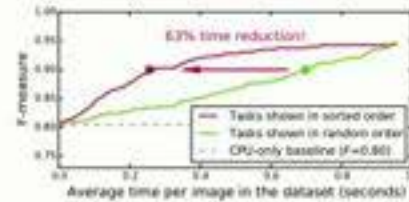


Figure 1. Consider a simple localization task where crowd workers refine the output of a machine classifier. At a threshold of 0.5, baseline accuracy starts at 0.80 (gray dotted line). If we show random tasks to human workers, accuracy improves (green), but if we order tasks by increasing machine confidence (purple), we can reduce the time requirement dramatically at a given target accuracy.

are returned. Finding many correct objects implies also detecting many false positives. Because the detection threshold determines this tradeoff, it is often treated as the primary tunable parameter of machine vision algorithms. The returned detections are then given to human workers, who we employ to remove false detections. For our experiments, we adopt the classic UIUC-Cars dataset (Agarwal, Awan, and Roth 2004). As detector, we use Support Vector Machines trained on Histograms of Ordered Gradients as a representative “out-of-the-box” machine vision system.

Our objective is to maximize the overall accuracy of the machine-crowd pipeline on the dataset given a certain time budget. We vary the time budget by presenting the humans with only a fraction of all detections. If humans look at a large fraction of detections the accuracy improvement will be large, however the average time cost per image in the dataset will also be large. If humans look at only a few images, the average accuracy of the entire dataset will show little improvement, but the time cost will be low. We plot the tradeoff between cost and accuracy as a curve.

The primary contribution of this work is a description and analysis of two strategies for improving the cost-accuracy curve. In Task Ordering we consider the impact of using the machine vision algorithm’s confidence score as a way to order human tasks. In Joint Optimization we consider how changing the machine threshold parameter impacts results.

Investigating the “Wisdom of Crowds” at Scale

Alok Shankar Mysore^{*}
PES Institute of Technology
alok.shankar@pesit.pes.edu

Vikas S Yaligar¹
National Institute of
Technology Karnataka
vikasyaligar@ieee.org

Imanol Arrieta Ibarra
Stanford University
imanol@stanford.edu

Camelia Simoiu
Stanford University
csimoiu@stanford.edu

Sharad Goel
Stanford University
sgoel@stanford.edu

Additional Authors^{*}
Various Institutions¹

ABSTRACT

In a variety of problem domains, it has been observed that the aggregate opinions of groups are often more accurate than those of the constituent individuals, a phenomenon that has been termed the “wisdom of the crowd.” Yet, perhaps surprisingly, there is still little consensus on how generally the phenomenon holds, how best to aggregate crowd judgements, and how social influence affects estimates. We investigate these questions by taking a meta wisdom of crowds approach. With a distributed team of over 100 student researchers across 17 institutions in the United States and India, we develop a large-scale online experiment to systematically study the wisdom of crowds effect for 1,000 different tasks in 50 subject

domains. These tasks involve various types of knowledge (e.g., explicit knowledge, tacit knowledge, and prediction), question formats (e.g., multiple choice and point estimation), and inputs (e.g., text, audio, and video). To examine the effect of social influence, participants are randomly assigned to one of three different experiment conditions in which they see varying degrees of information on the responses of others. In this ongoing project, we are now preparing to recruit participants via Amazon’s Mechanical Turk.

Author Keywords

Crowdsourcing; online experiment; crowd consensus

ACM Classification Keywords

H.5.m. Economics: Experimentation Design

INTRODUCTION

At a 1906 county fair, the statistician Francis Galton watched as eight hundred people competed to guess the weight of an ox. He famously observed that the median of the guesses, 1,207 pounds, was, remarkably, within 1% of the true weight [1].

Simple aggregation—as in the case of Galton’s ox competition, or voting in democratic elections—has been shown to be a surprisingly powerful technique for prediction, inference, and decision-making. Over the last century, there have been dozens of studies that examine this wisdom of crowds effect. For example, crowd judgements have been used to identify phishing websites [6], answer general knowledge questions [5], and forecast weather-related events [3]. In these applications, a wide variety of aggregation methods have been considered, ranging from standard measures, such as the mean and median, to more specialized, domain-specific techniques, such as those based on cognitive models of decision making [4]. However, given the diversity of experimental designs, subject pools, and analytic methods employed, it has proven difficult to compare studies and extract general principles. It is thus unclear whether these documented examples are a representative collection of a much larger space of tasks that exhibit a wisdom of crowds phenomenon, or conversely, whether they are highly specific instances of an interesting, though ultimately limited occurrence.

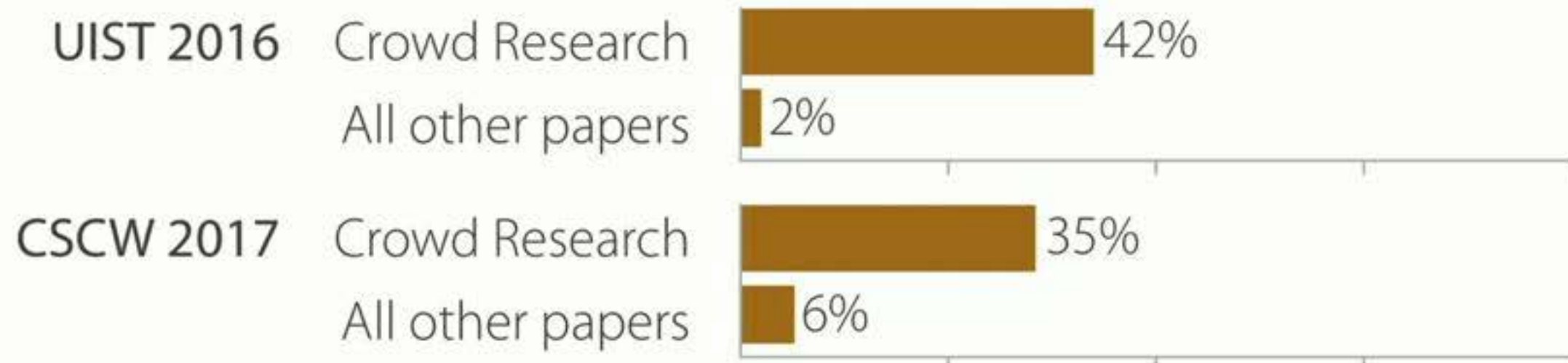
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author. Copyright held by the owner/authors.
0327-7724/15/0000-0001..2015 ACM 978-1-4503-2817-1/15
http://dx.doi.org/10.1145/2817705.2817721

INCREASED ACCESS TO RESEARCH

Coauthors' universities that are ranked below 500 worldwide



Coauthors whose countries are ranked below 50 worldwide in GDP per capita



ATTAINMENT: PROVIDING ACCESS

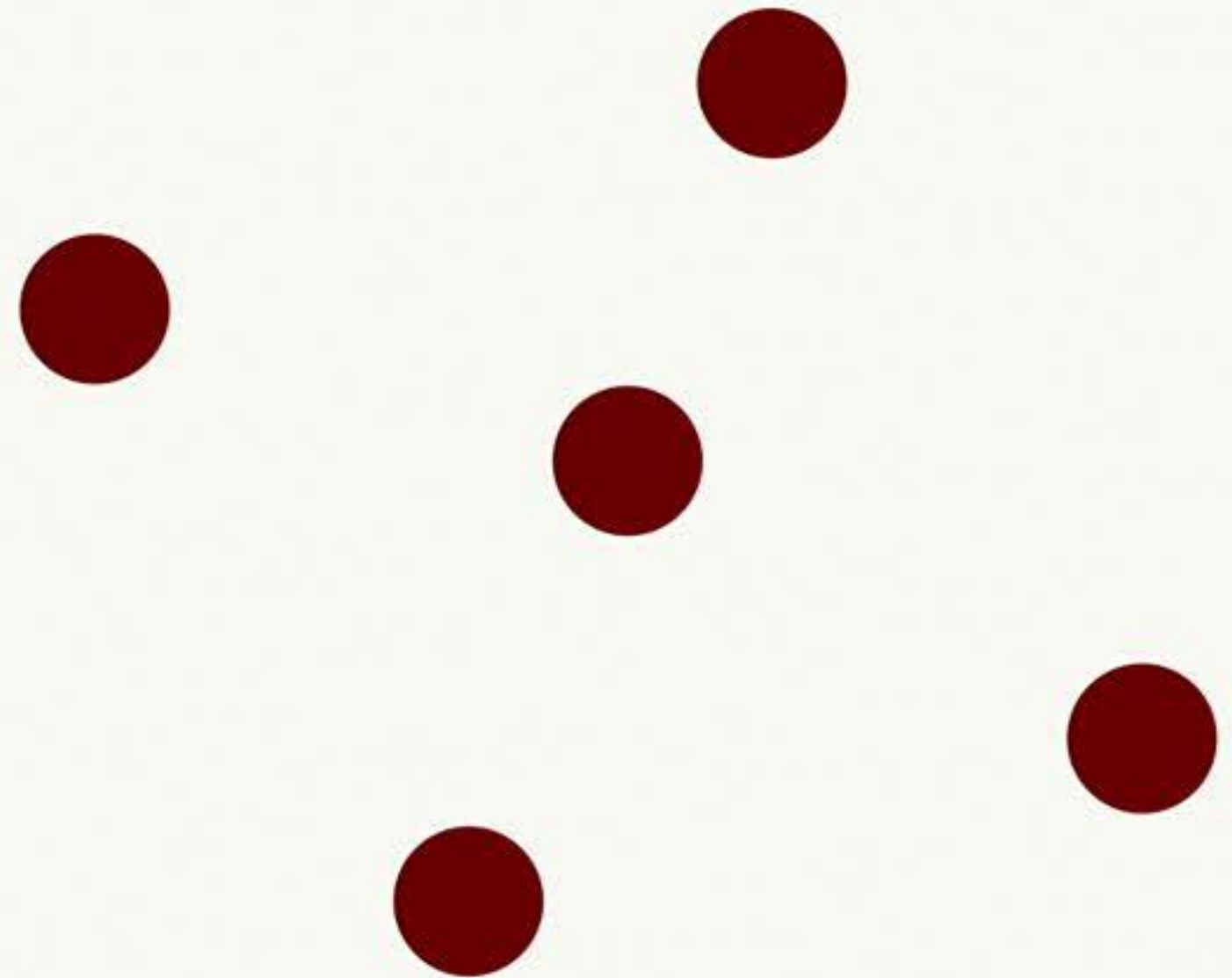
Participants have gone on to programs at UC Berkeley, and Carnegie Mellon University, and MIT.

21 of 33 surveyed were admitted to at least one program, despite a **median of 0 other letter writers** from institutions ranked better than 500 worldwide.

DECENTRALIZED CREDIT: TRANSLATE INTO GRAPH PROBLEM

Each participants allocates **credit points** to other participants based on their assessment of who impacted the project

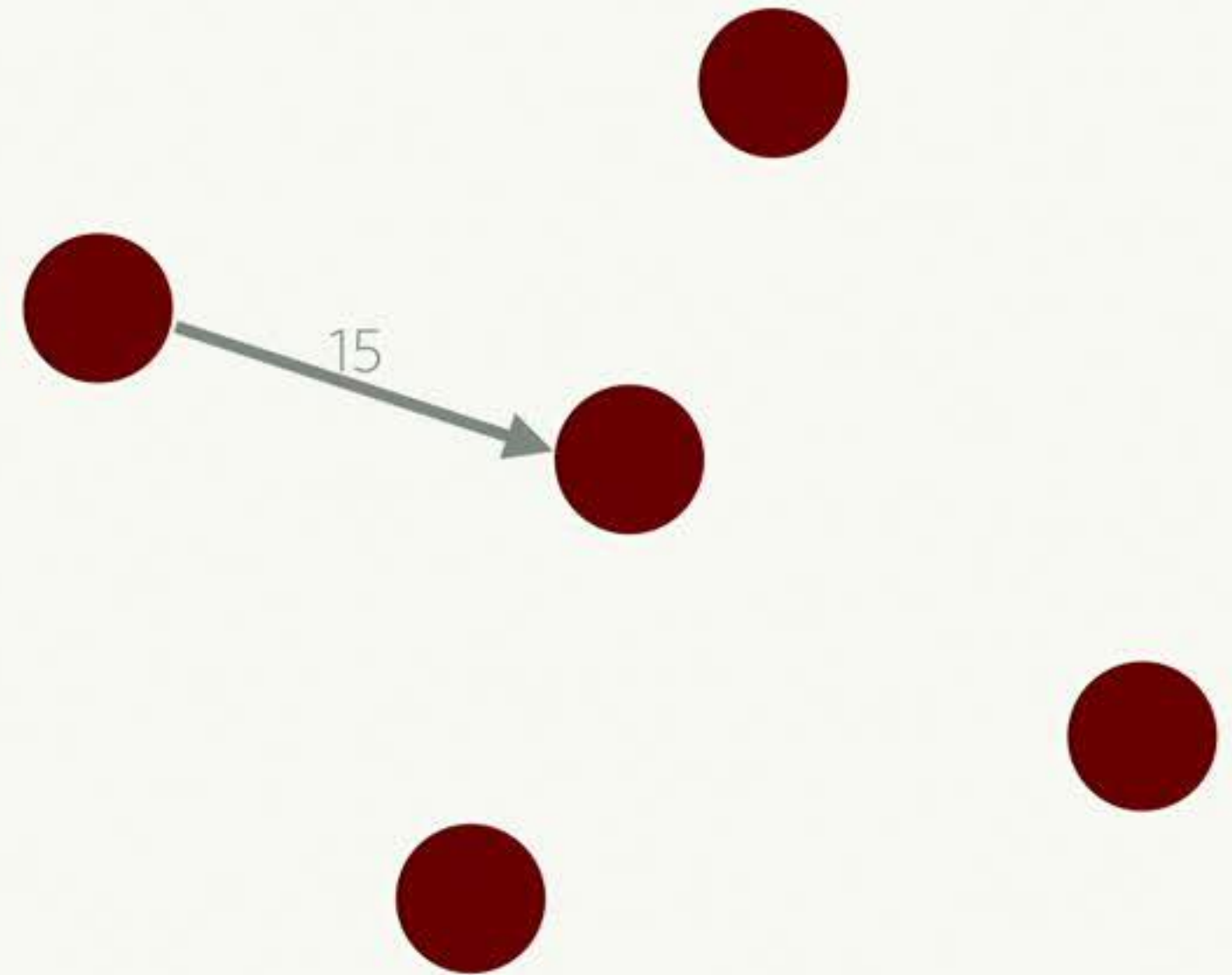
Resulted: weighted directed graph



DECENTRALIZED CREDIT: TRANSLATE INTO GRAPH PROBLEM

Each participants allocates **credit points** to other participants based on their assessment of who impacted the project

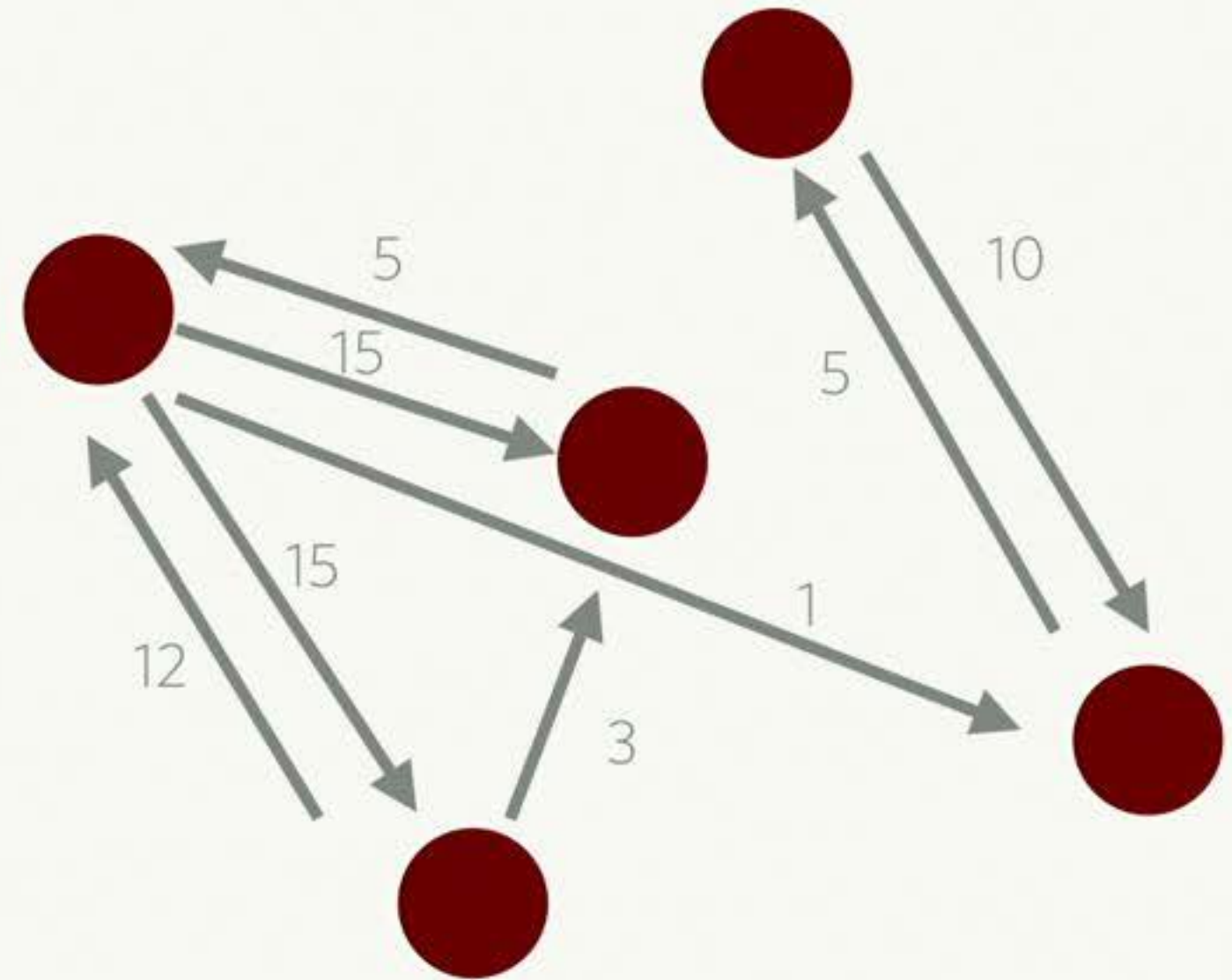
Resulted: weighted directed graph



DECENTRALIZED CREDIT: TRANSLATE INTO GRAPH PROBLEM

Each participants allocates **credit points** to other participants based on their assessment of who impacted the project

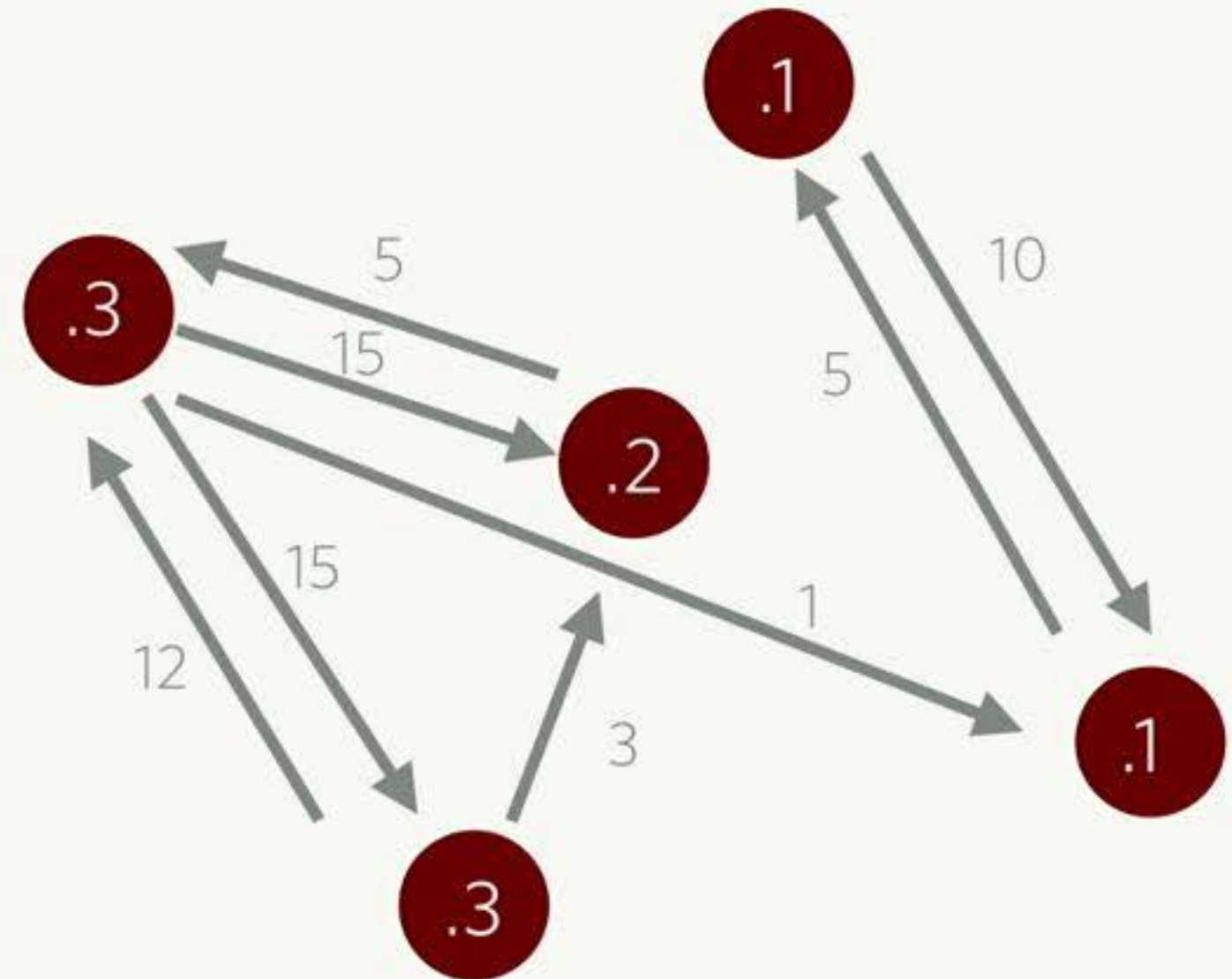
Resulted: weighted directed graph



GRAPH CENTRALITY: PAGERANK

Intuition: identify nodes that are receiving large amounts of credit, weigh those nodes' allocations heavily, and iterate until convergence

Propagate each node's score in proportion to its outgoing wedge weights



STRATEGIC BEHAVIOR

Speaking different languages or otherwise interacting with only a small part of the crowd: link ring

Strategically directing credit toward those who will return credit to you: such attacks occur in 360-degree reviews

Formulations of centrality algorithms can correct for most of these attacks

CREDIT

Give everyone a say in how much credit each person deserves. For group projects, performance reviews, paper authorship, and more.

START

HOW IT WORKS



01

Add everyone in your team



02

Team members privately score each other



03

Credit computes scores for each team member

ANALYZING CREDDIT'S EFFECT

What impact did Creddit have on credit distribution?

Method: normalize raw summed credit scores, and Creddit-adjusted scores, to sum to 1.0

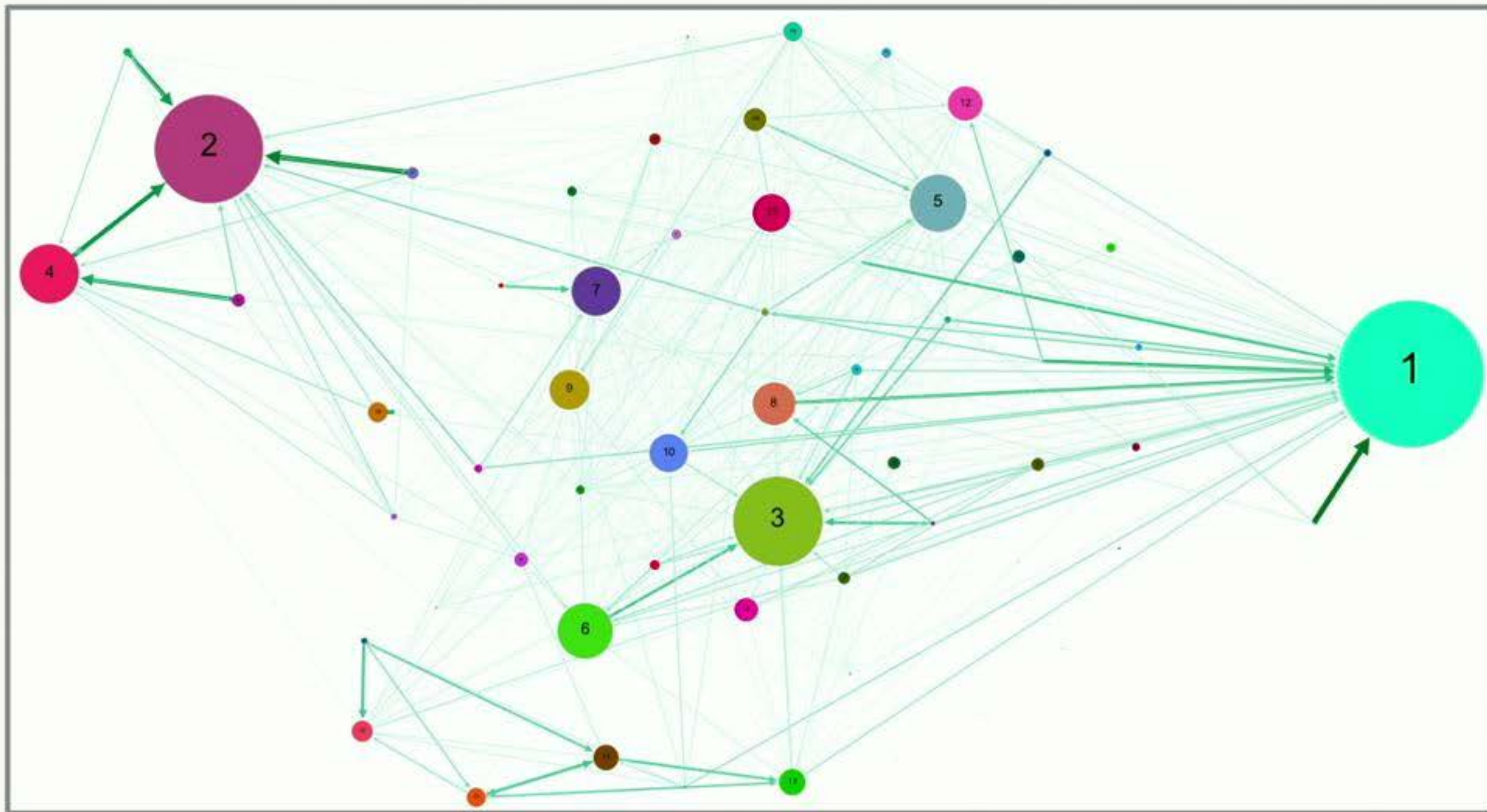
Regress both raw score and Creddit score on observable collaboration behaviors, and compare β estimates across the regressions

LESS TALKING, MORE DOING

Participation Measure	Credit: β_c	Raw Votes: β_{raw}	$\beta_c - \beta_{raw}$
# Hangouts	0.0694***	0.0438*	0.0256
# Files Uploaded	0.0352**	0.0293*	0.0059
# GitHub commits	0.0171	-0.024*	0.0411***
# Slack messages	0.0351*	0.1122***	-0.0770***
# self-organized meetings	0.0239*	0.0115	0.0123
Milestone leader (binary)	0.0360***	0.0059	0.0300**
Weeks active	0.0252*	0.0141	0.011

All variables standardized

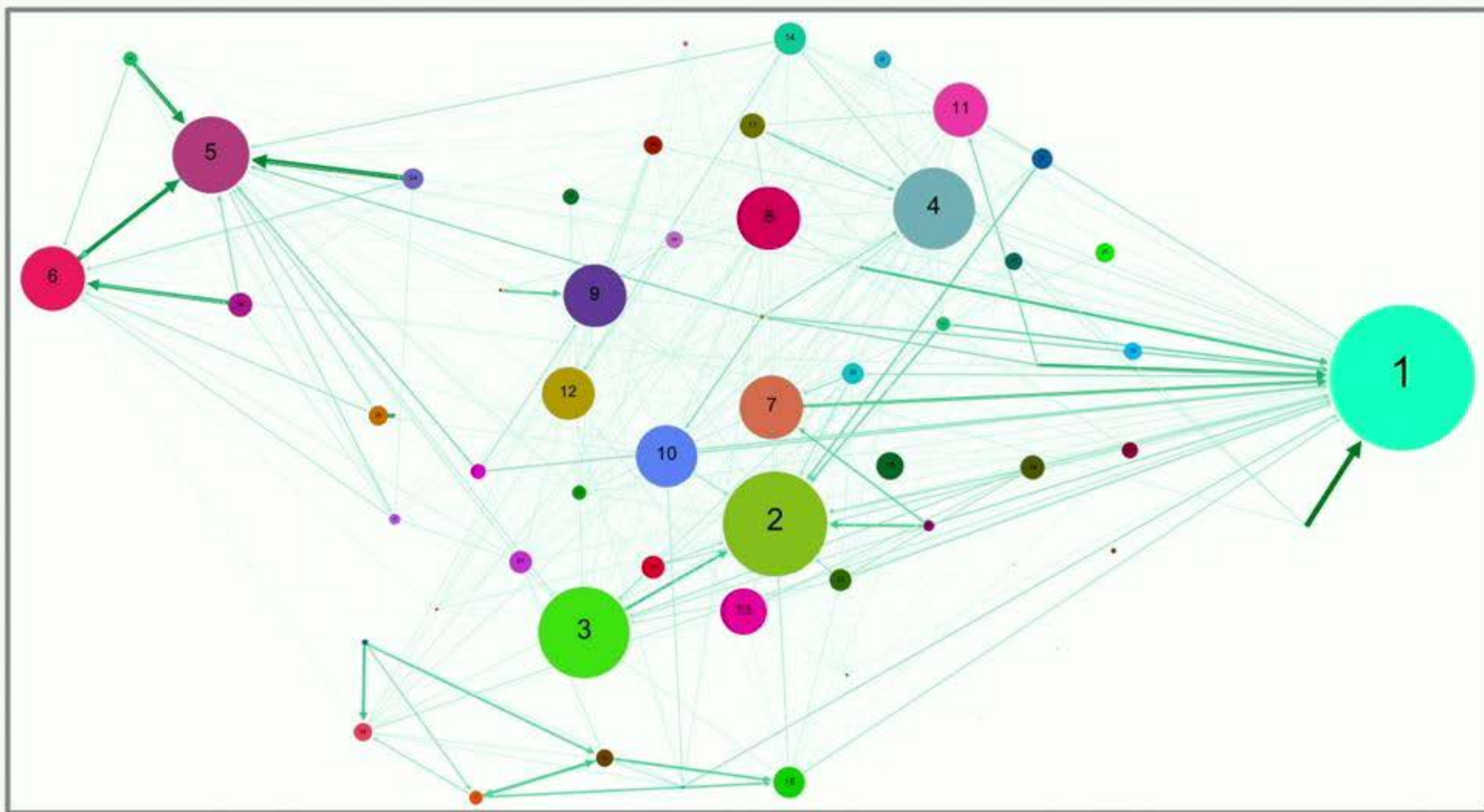
EFFECTS ON AUTHOR ORDER



Raw vote ranking

#2 and #4 have a high rank due to link ring

EFFECTS ON AUTHOR ORDER



PageRank-corrected author order

Influential coauthors reduced impact of link ring

CROWD RESEARCH: REFLECTIONS

What impact would decentralized credit have on traditional teams or organizations?

What kinds of research projects can operate at a larger technical scale than traditional CS research?

Rather than structuring the future of work as an algorithm, how might we design computationally augmented organizational structures?

Rather than structuring the future of work as an algorithm, how might we design computationally augmented organizational structures?

Organizations were originally designed with inspiration by mechanical systems. What might a computational infrastructure offer them?

Crowds, Computation, and the Future of Work

Thanks to...

Amazing students: Daniela Retelny, Niloufar Salehi, Rajan Vaish, Ranjay Krishna, Sharon Zhou

Amazing colleagues: Melissa Valentine, Fei-Fei Li, James Davis, Sharad Goel

Amazing supporters: NSF, Sloan, Accenture ATL, Microsoft, Toyota Human-Centered AI, Stanford Cyber Initiative, ONR, DARPA, Hasso Plattner Design Thinking Program

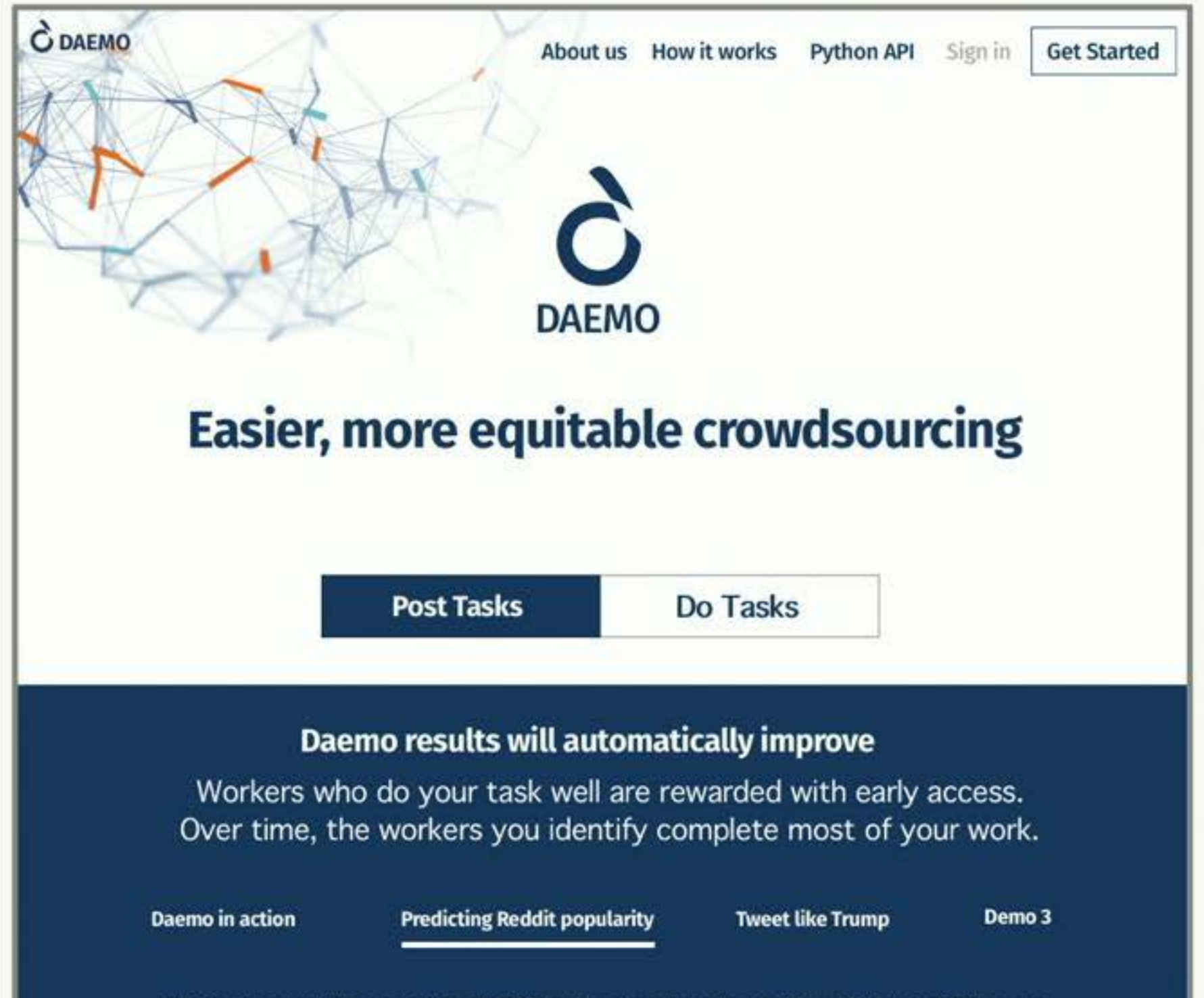
Amazing questions

THREE PARALLEL PROJECTS

HCI

Michael Bernstein,
Stanford

Building a new crowd
marketplace



The screenshot shows the DAEMO website homepage. At the top left is the DAEMO logo. To the right are navigation links: "About us", "How it works", "Python API", "Sign in", and a "Get Started" button. The background features a network graph with blue and orange nodes and edges. Below the navigation is a large DAEMO logo and the headline "Easier, more equitable crowdsourcing". Underneath the headline are two buttons: "Post Tasks" (highlighted in dark blue) and "Do Tasks". A dark blue section below contains the text "Daemo results will automatically improve" followed by a paragraph: "Workers who do your task well are rewarded with early access. Over time, the workers you identify complete most of your work." At the bottom, there are four links: "Daemo in action", "Predicting Reddit popularity" (underlined), "Tweet like Trump", and "Demo 3".

Crowds, Computation, and the Future of Work

Thanks to...

Amazing students: Daniela Retelny, Niloufar Salehi, Rajan Vaish, Ranjay Krishna, Sharon Zhou

Amazing colleagues: Melissa Valentine, Fei-Fei Li, James Davis, Sharad Goel

Amazing supporters: NSF, Sloan, Accenture ATL, Microsoft, Toyota Human-Centered AI, Stanford Cyber Initiative, ONR, DARPA, Hasso Plattner Design Thinking Program

Amazing questions