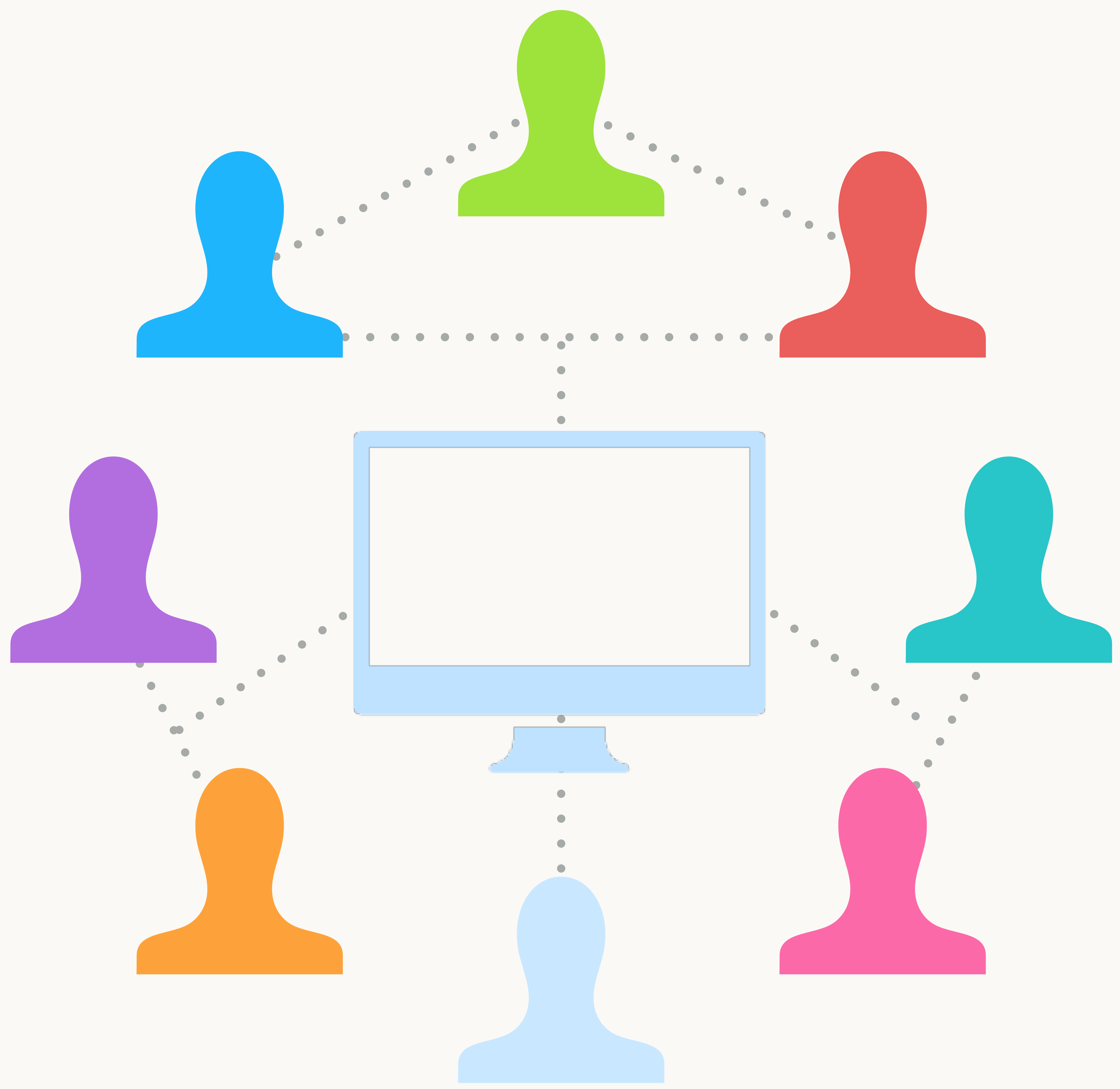
In A Flash: Crowdsourcing Organizations, Collaboration, and Research

Michael Bernstein
Stanford University

A personal story

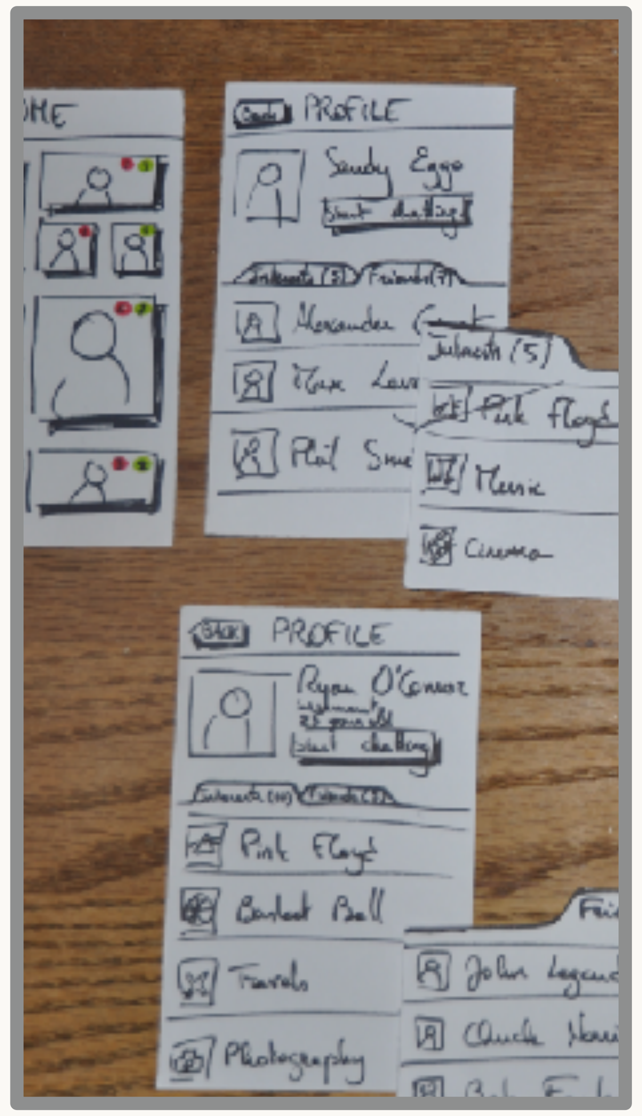
ACHIEVING MORE TOGETHER

Crowdsourcing is a technology for amplifying human effort



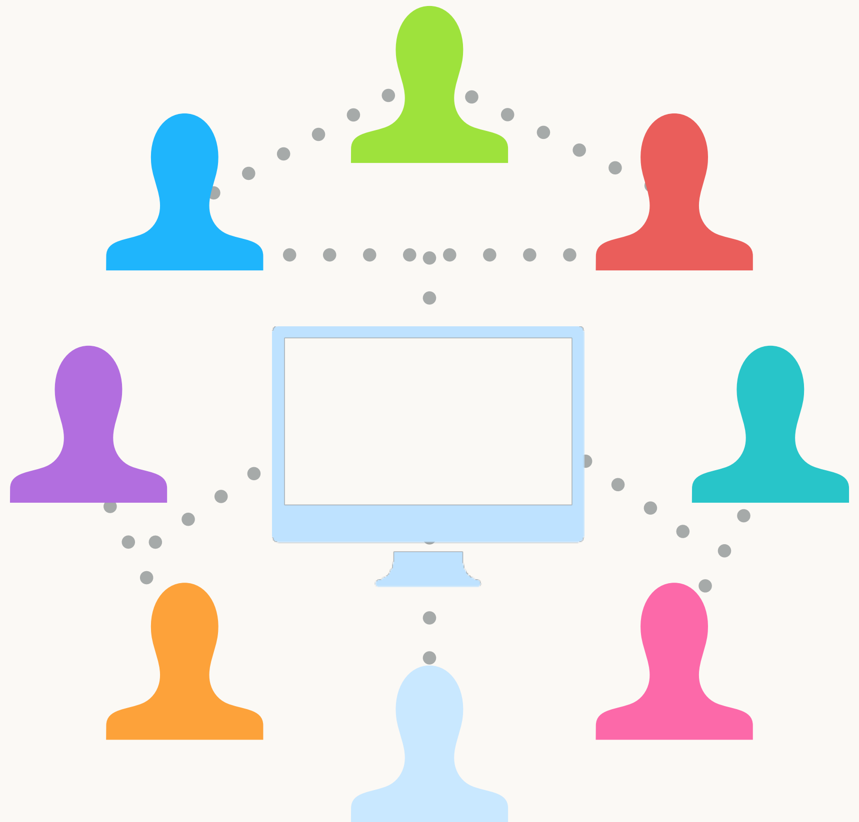
CREATING INFRASTRUCTURES FOR COMPLEX GOALS

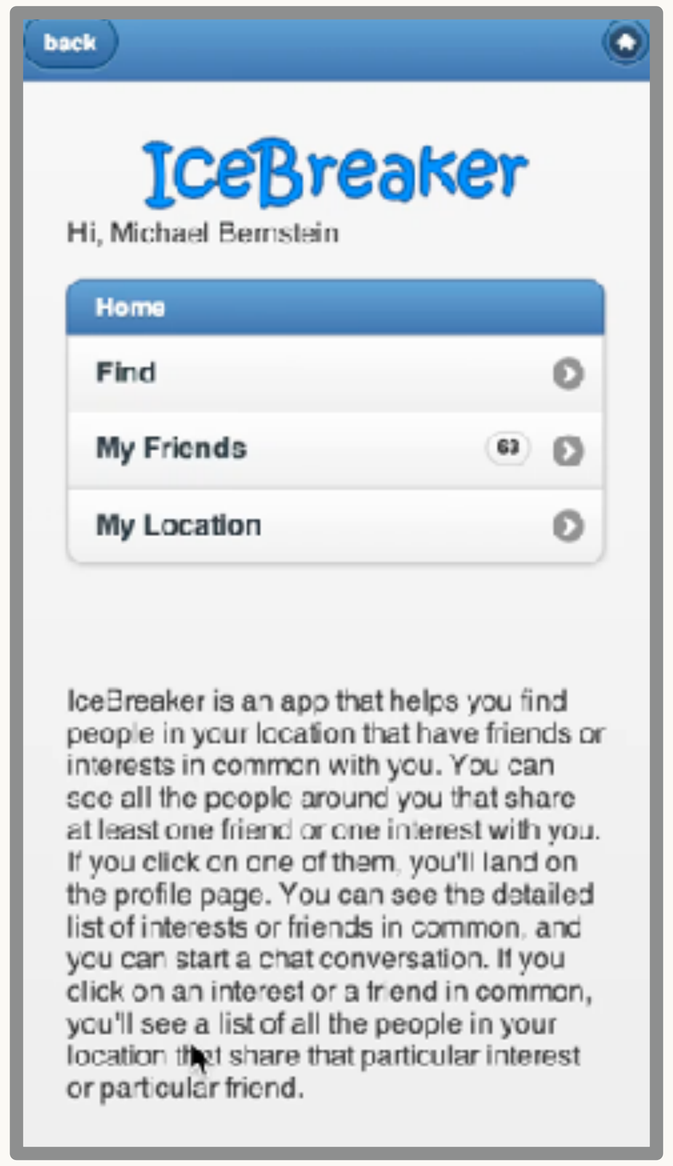
Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]



NAPKIN SKETCH

24 HOURS



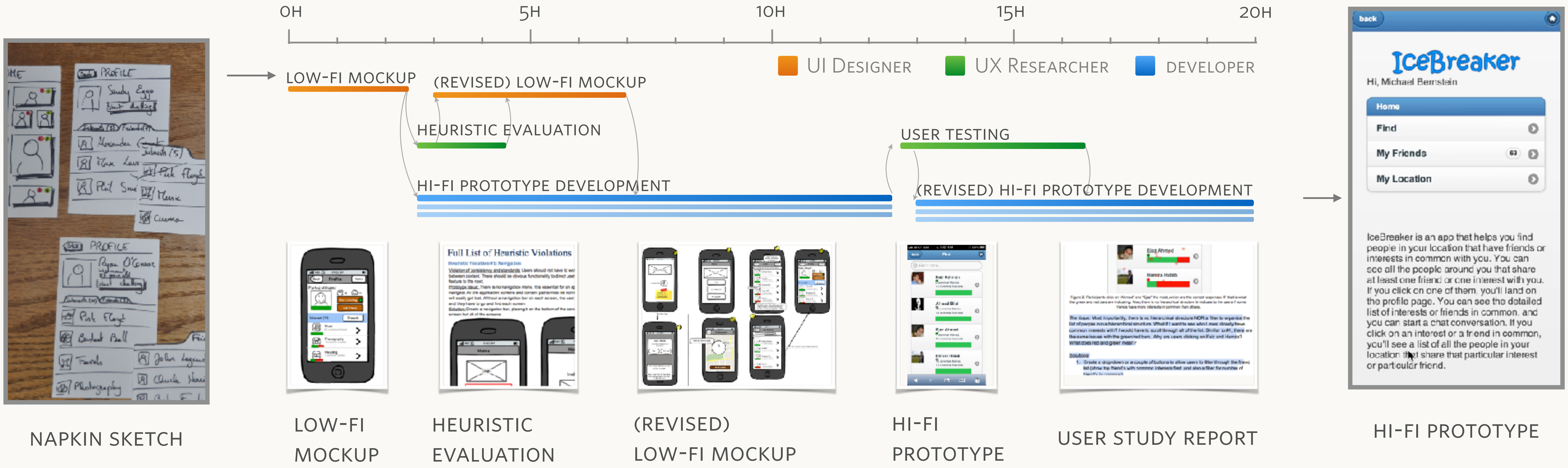


HI-FI PROTOTYPE

CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork

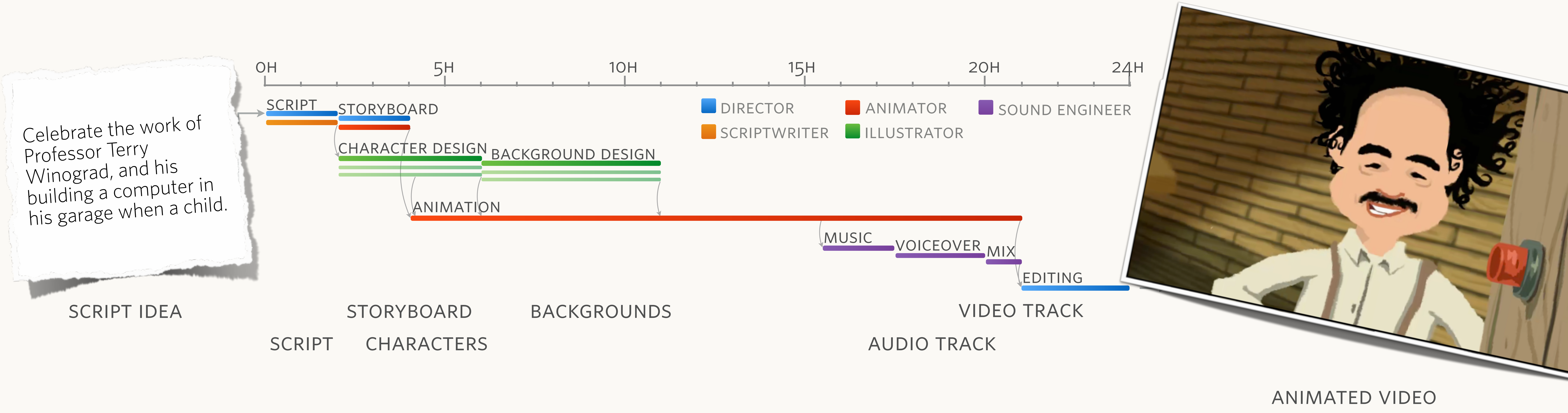
[Retelny et al. 2014]



CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork

[Retelny et al. 2014]



CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]

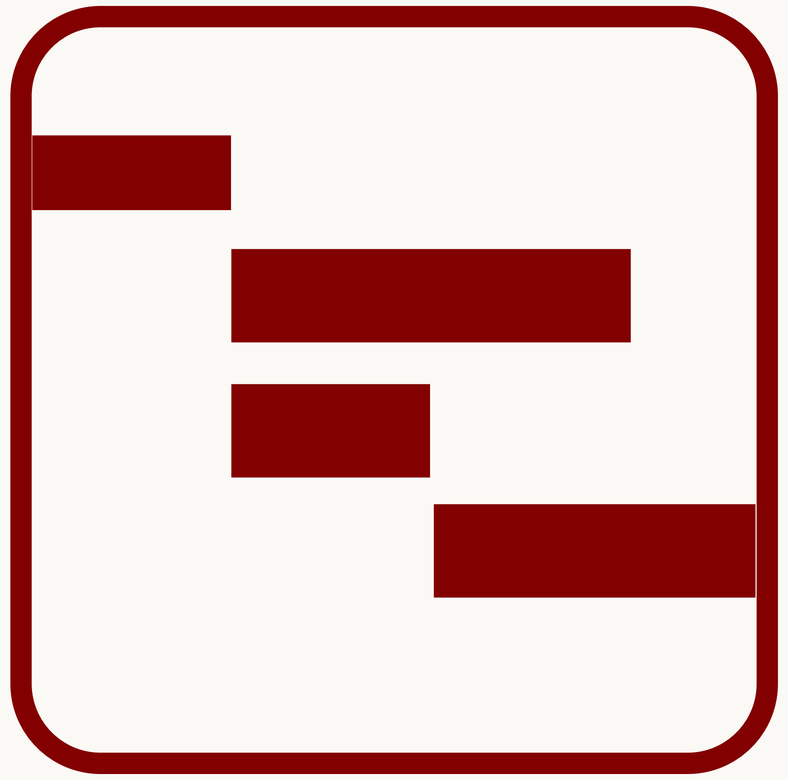
Computational crowdsourcing techniques enable...

Modularity

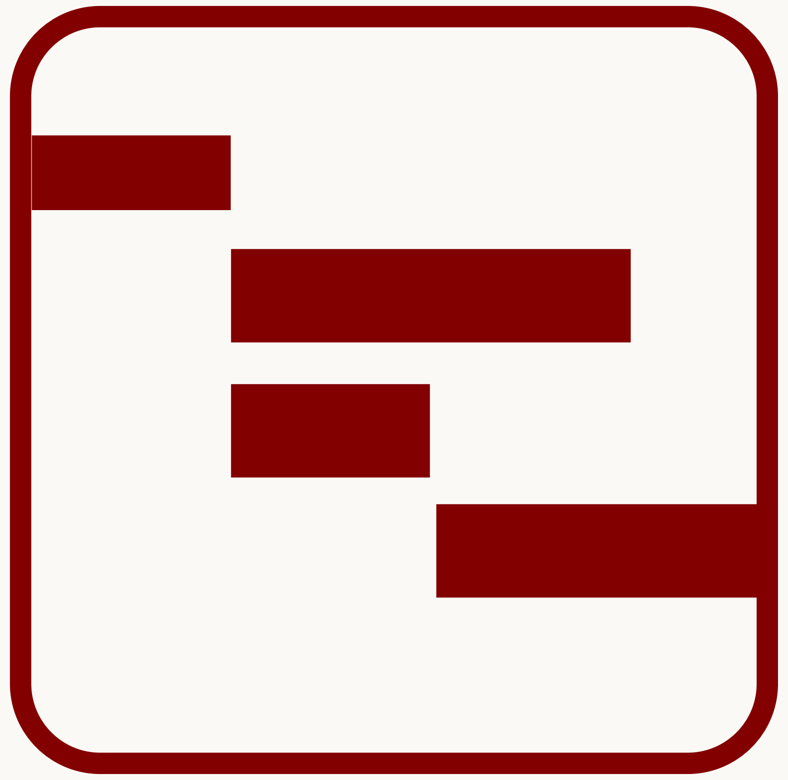
Elasticity

Pipelining

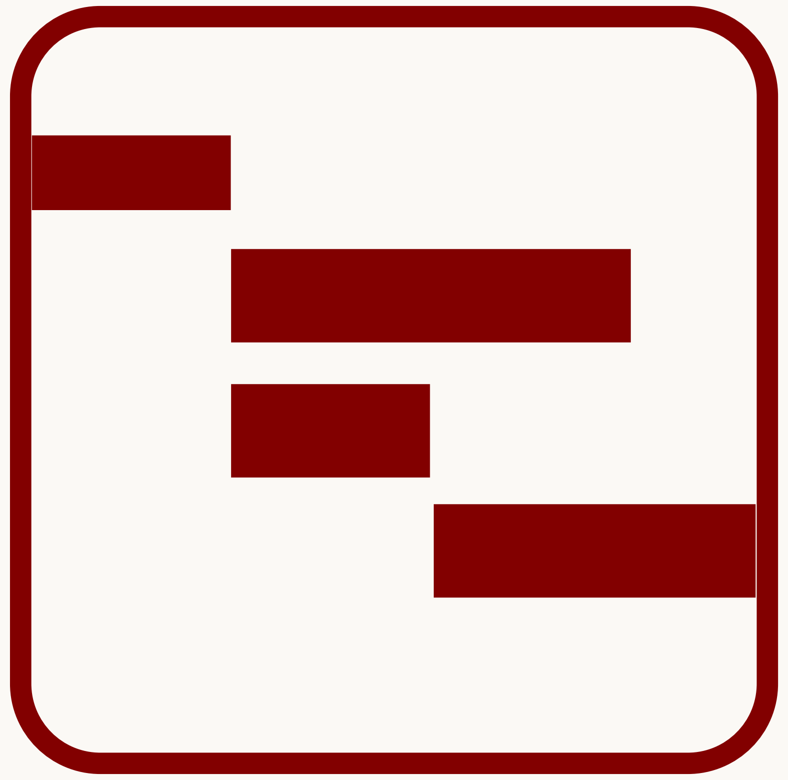
Automatic creation



DESIGN



DESIGN



DESIGN

CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]

Computational crowdsourcing techniques enable...

- Modularity
- **Elasticity**
- Pipelining
- Automatic creation

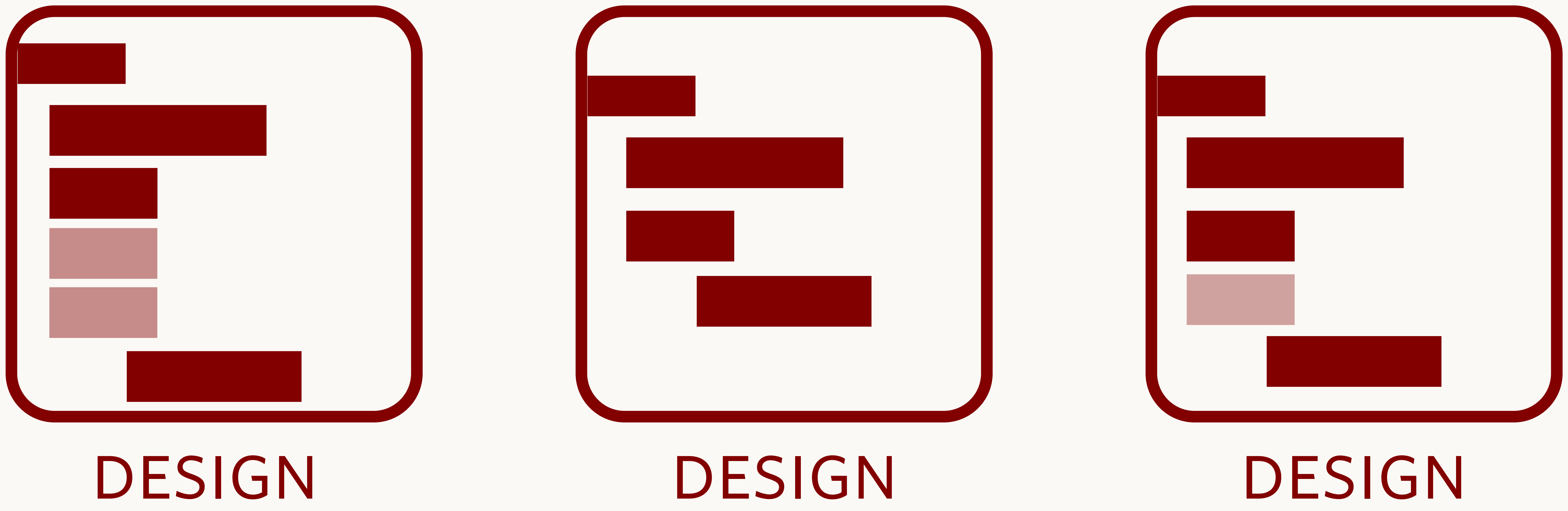
DESIGN DESIGN DESIGN

CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]

Computational crowdsourcing techniques enable...

- Modularity
- Elasticity
- **Pipelining**
- Automatic creation



CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]

Computational crowdsourcing techniques enable...

- Modularity
- Elasticity
- Pipelining
- **Automatic creation**

CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]

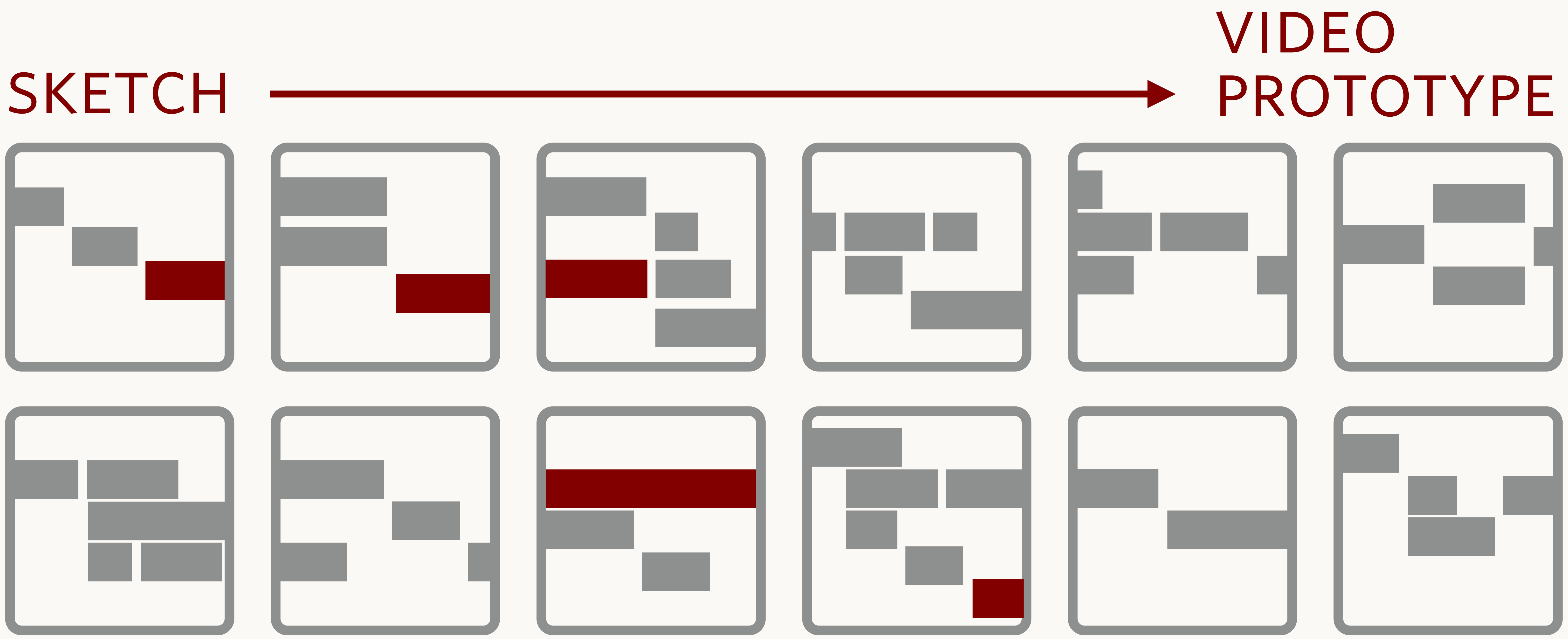
Computational crowdsourcing techniques enable...

Modularity

Elasticity

Pipelining

Automatic creation

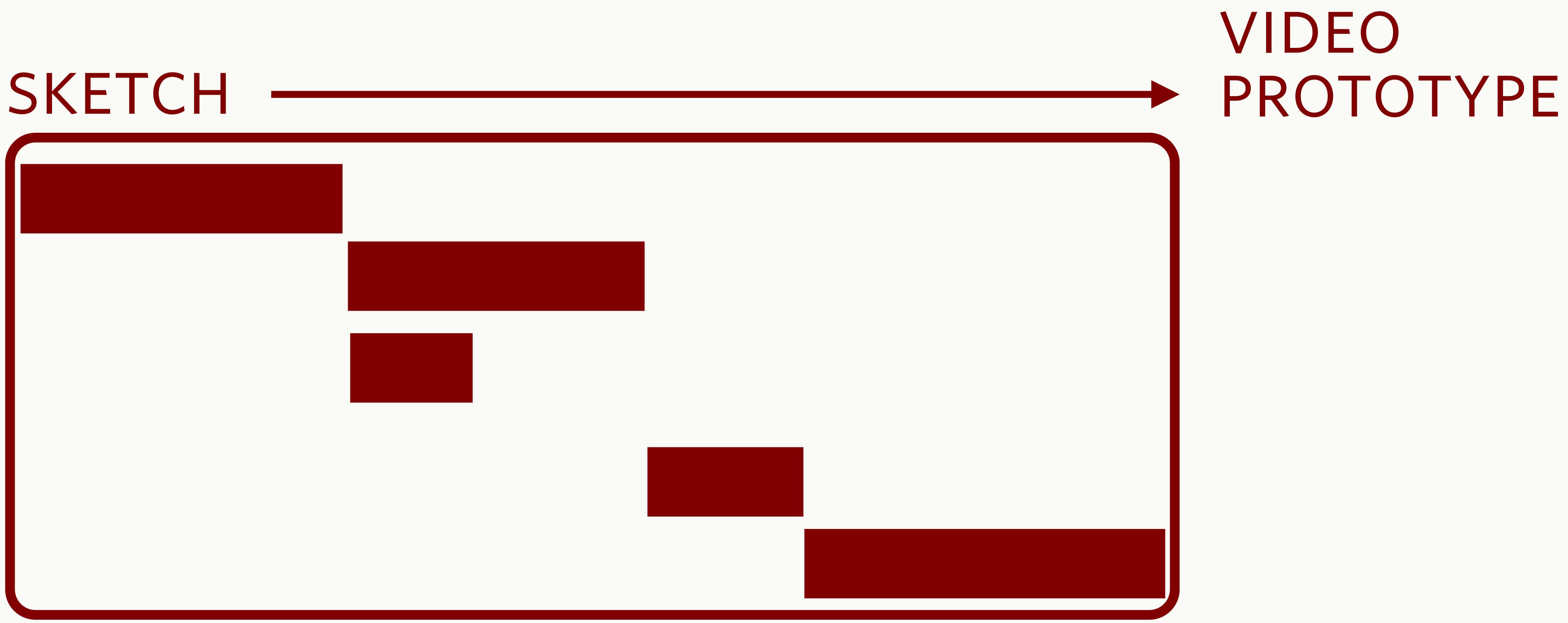


CREATING INFRASTRUCTURES FOR COMPLEX GOALS

Crowdsourcing on-demand groups of experts from Upwork
[Retelny et al. 2014]

Computational crowdsourcing techniques enable...

- Modularity
- Elasticity
- Pipelining
- **Automatic creation**



NARROW VS. OPEN-ENDED GOALS

These teams were restricted in what they can achieve

- Interface iteration, not product design
- Rapid prototypes, not software engineering
- Animating a prompt, not film or game production

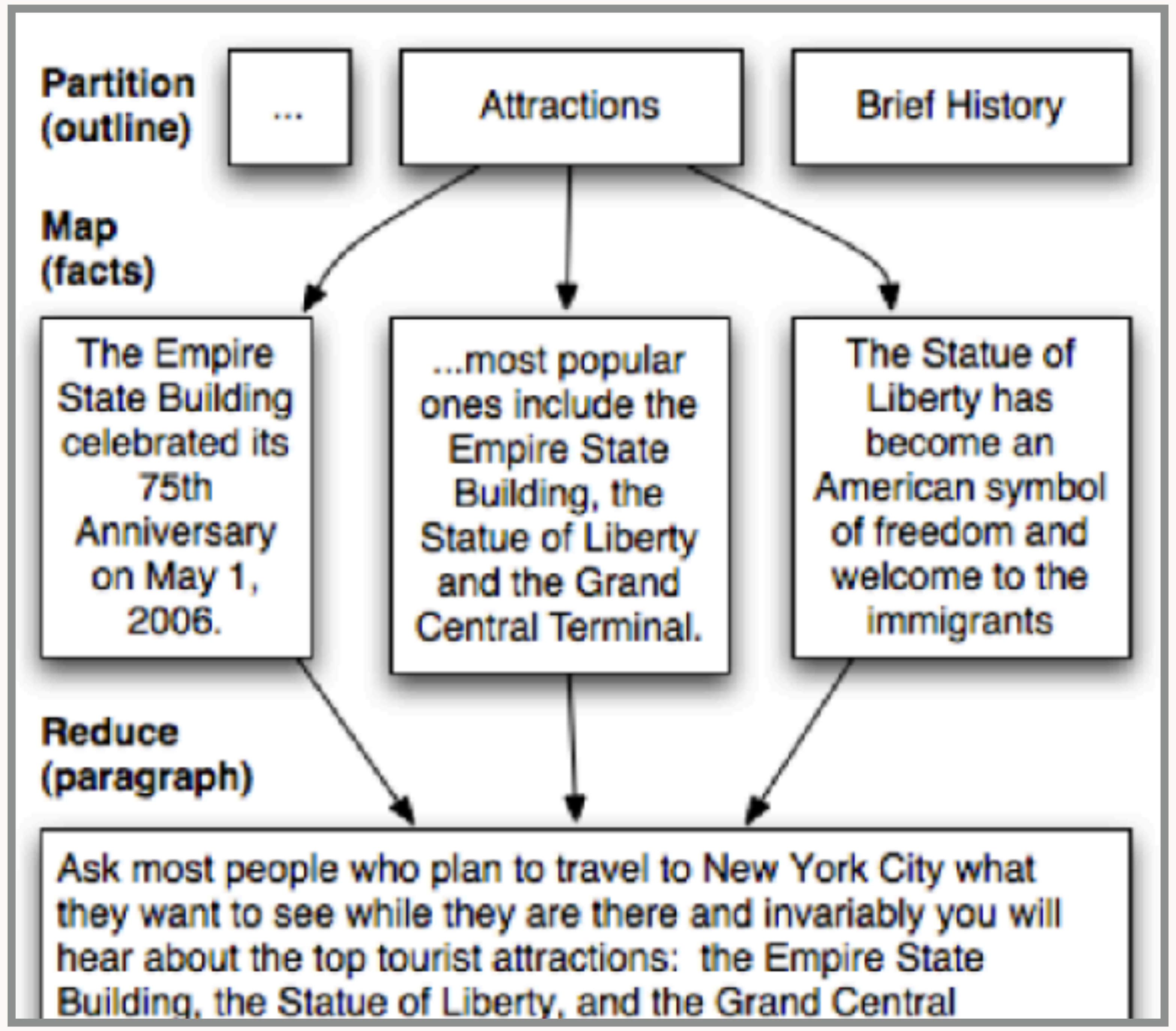
Could we achieve open-ended, complex goals such as product design, software development, and game production?

No. We couldn't create a process structured enough for computation to help, without over-constraining it.

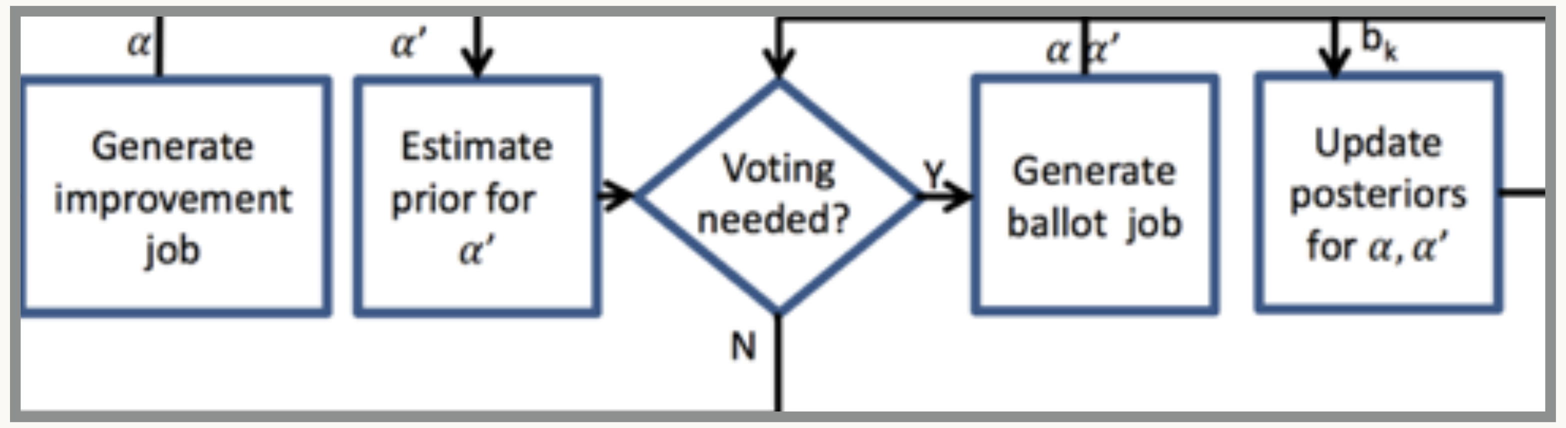
INFRASTRUCTURE: CROWD ALGORITHMS

Crowdsourcing's infrastructure is based on algorithmic primitives

- Modularize and pre-define all actions
- This abstraction allows computation to decide which behaviors are taken, when, and by whom; optimize, error-check, combine submissions, and more



[Kittur et al. 2011]



[Dai et al. 2010]

[Little et al. 2010]

THE LIMITS OF ALGORITHMS

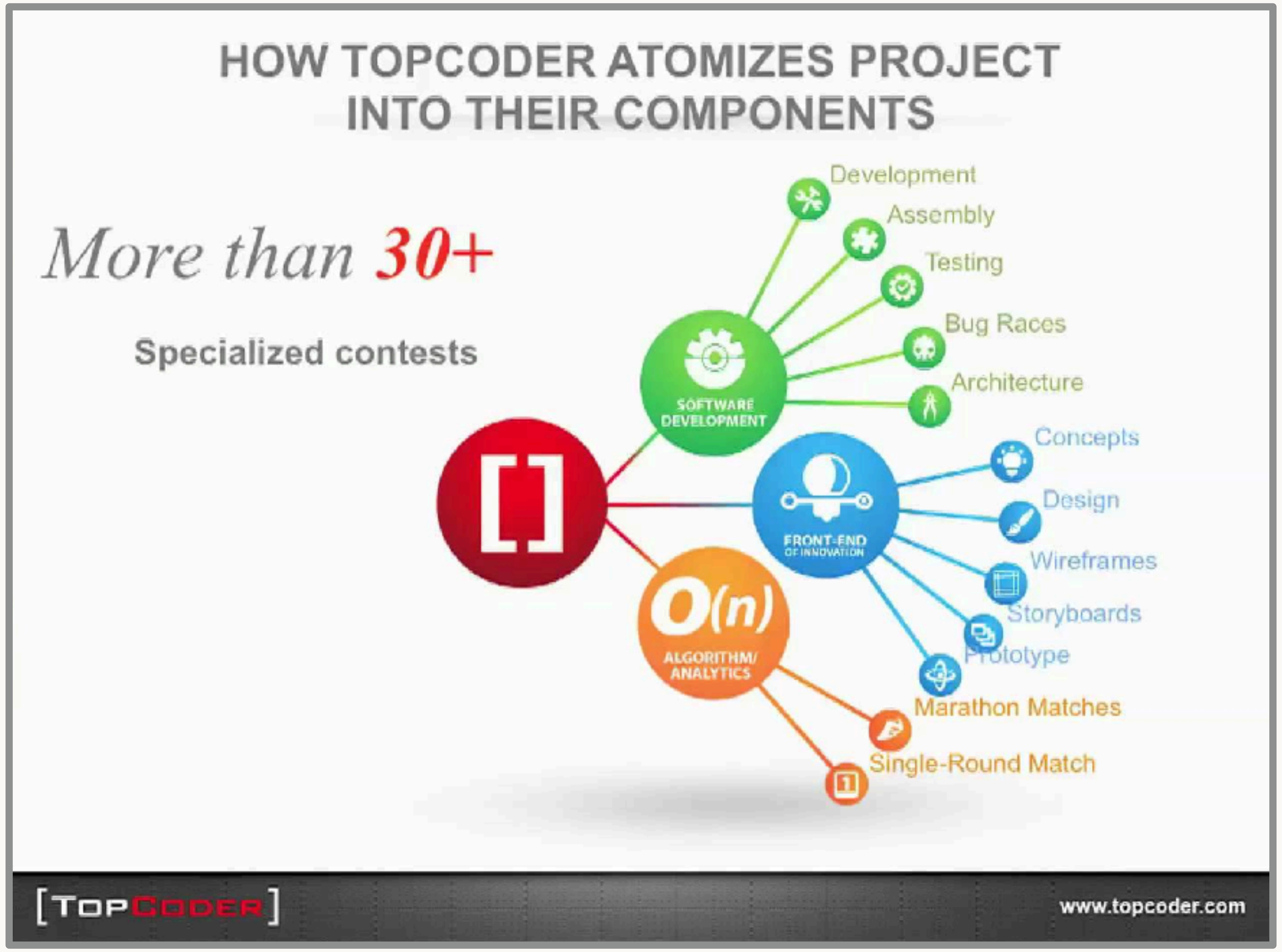
Open-ended and complex goals are fundamentally incompatible with a requirement to pre-define all behaviors
[Van de Ven, Delbecq, and Koenig 1976; Rittel and Weber 1973; Schön 1984]

This infrastructure confines crowdsourcing to goals so predictable that they can be entirely pre-defined

OPEN SOURCE AND OPEN INNOVATION SUFFER TOO

"Peer production is limited not by the total cost or complexity of a project, but by its modularity, the granularity of its components, and the cost of integration."

[Benkler 2002]



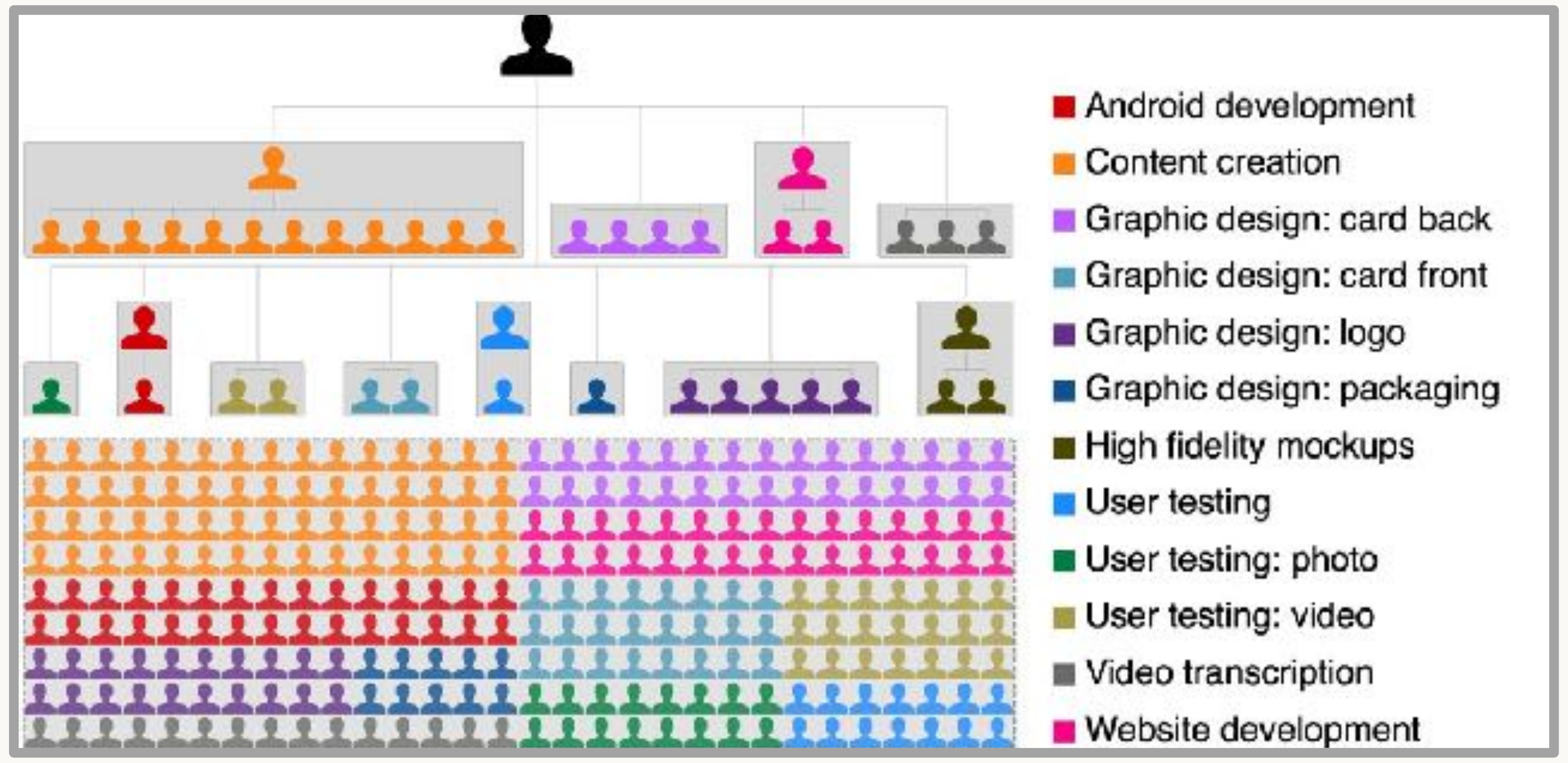
[Boudreau, Lacetera, and Lakhani 2011]

THE CHALLENGE

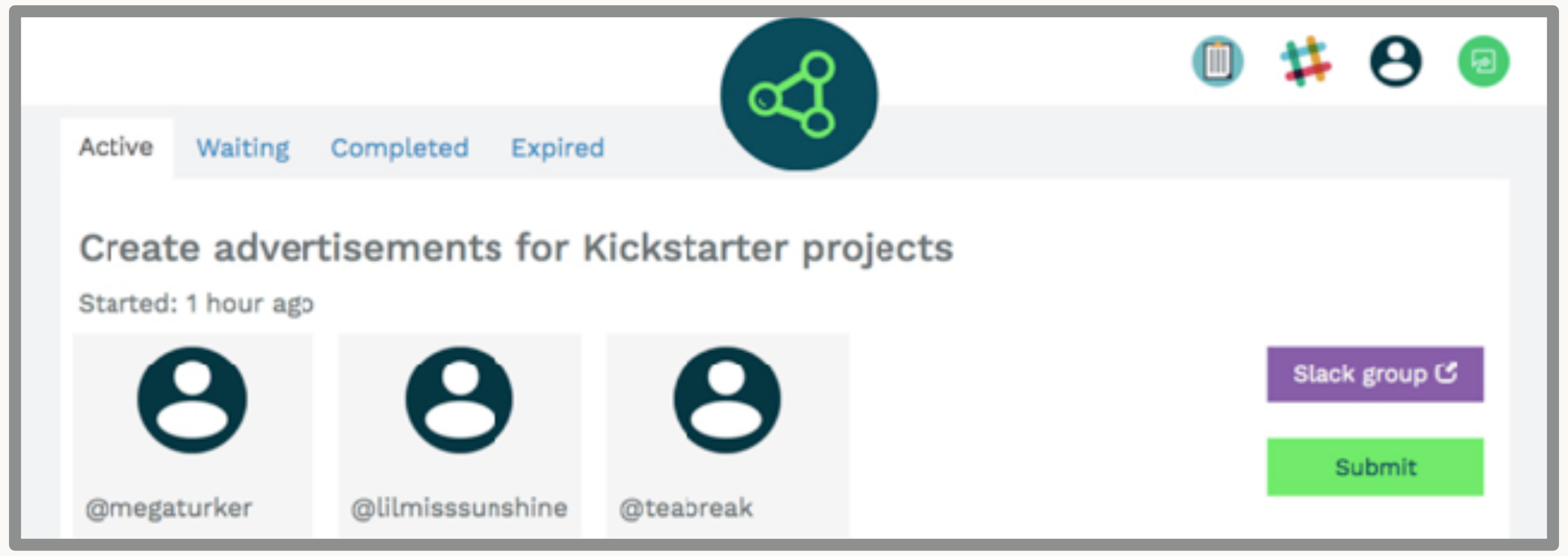
The very thing that gives crowdsourcing systems their leverage is also preventing them from achieving complex and open-ended outcomes

An alternative approach: crowds structured not like algorithms, but like organizations

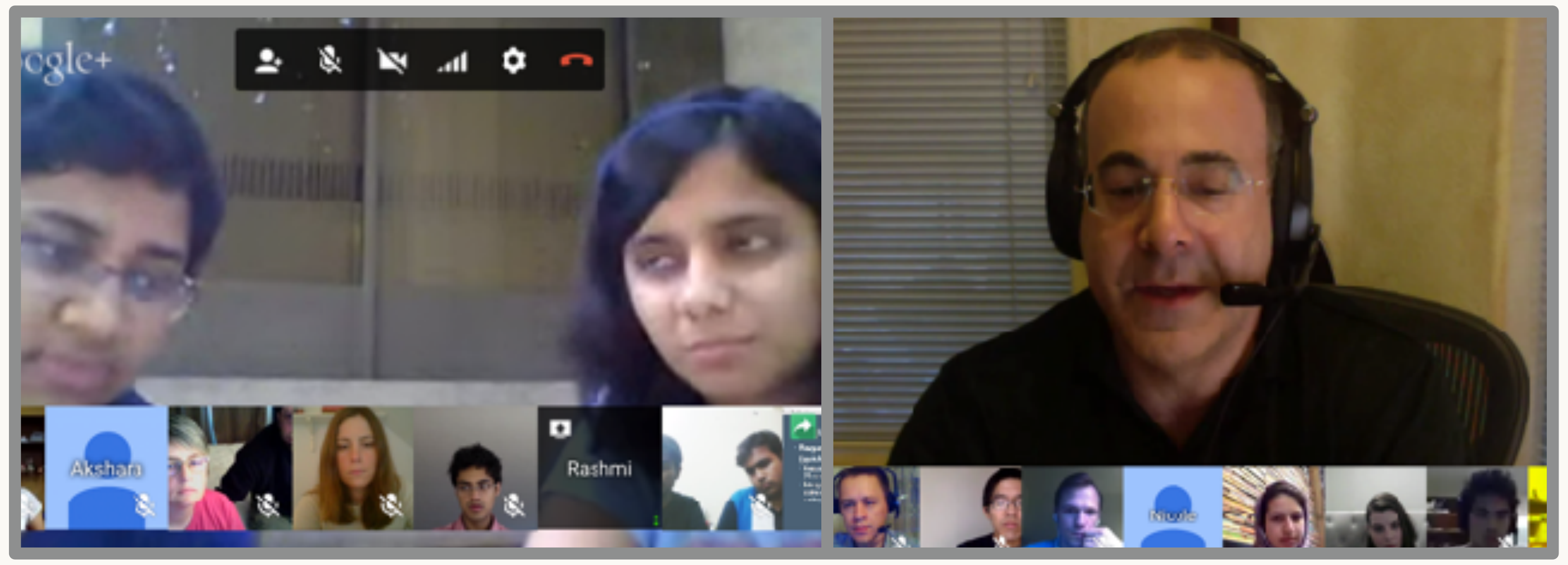
1) Enable crowd collectives to achieve complex and open-ended goals



2) Recruit effective collaborators despite unpredictable availability



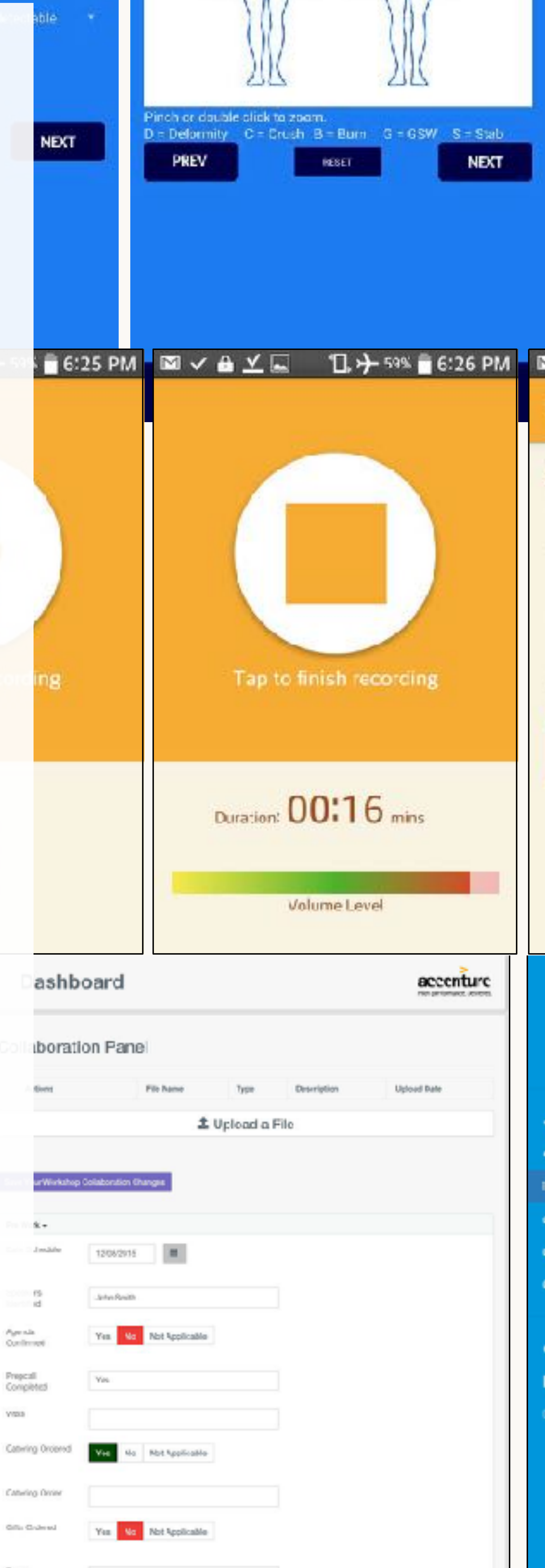
3) Crowdsource research itself, providing global access to upward mobility



Flash organizations

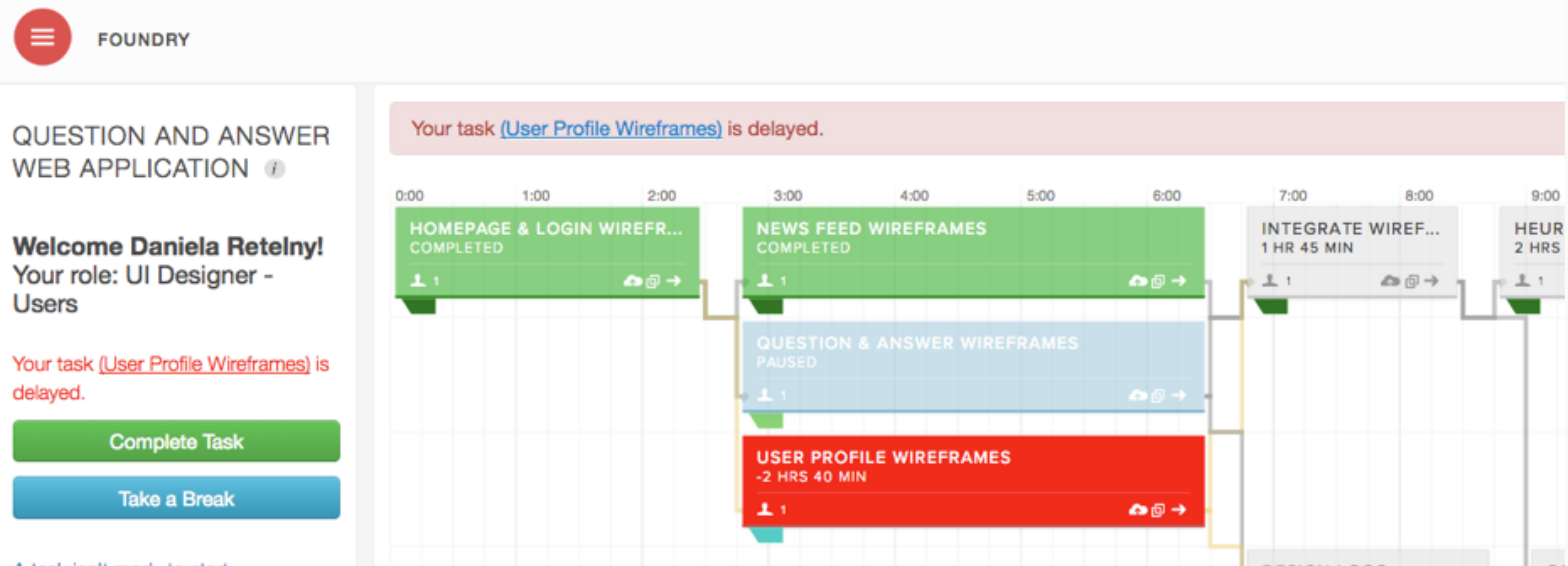
Valentine, Retelny, To, Rahmati, Doshi, Bernstein. CHI 2017.

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



CHALLENGES

1) Organizations assume **asset specificity**: people developing effective collaboration patterns over time
[Williamson 1976]

...but on-demand crowds do not offer asset specificity

CHALLENGES

1) Organizations assume **asset specificity**: people developing effective collaboration patterns over time
[Williamson 1976]

...but on-demand crowds do not offer asset specificity

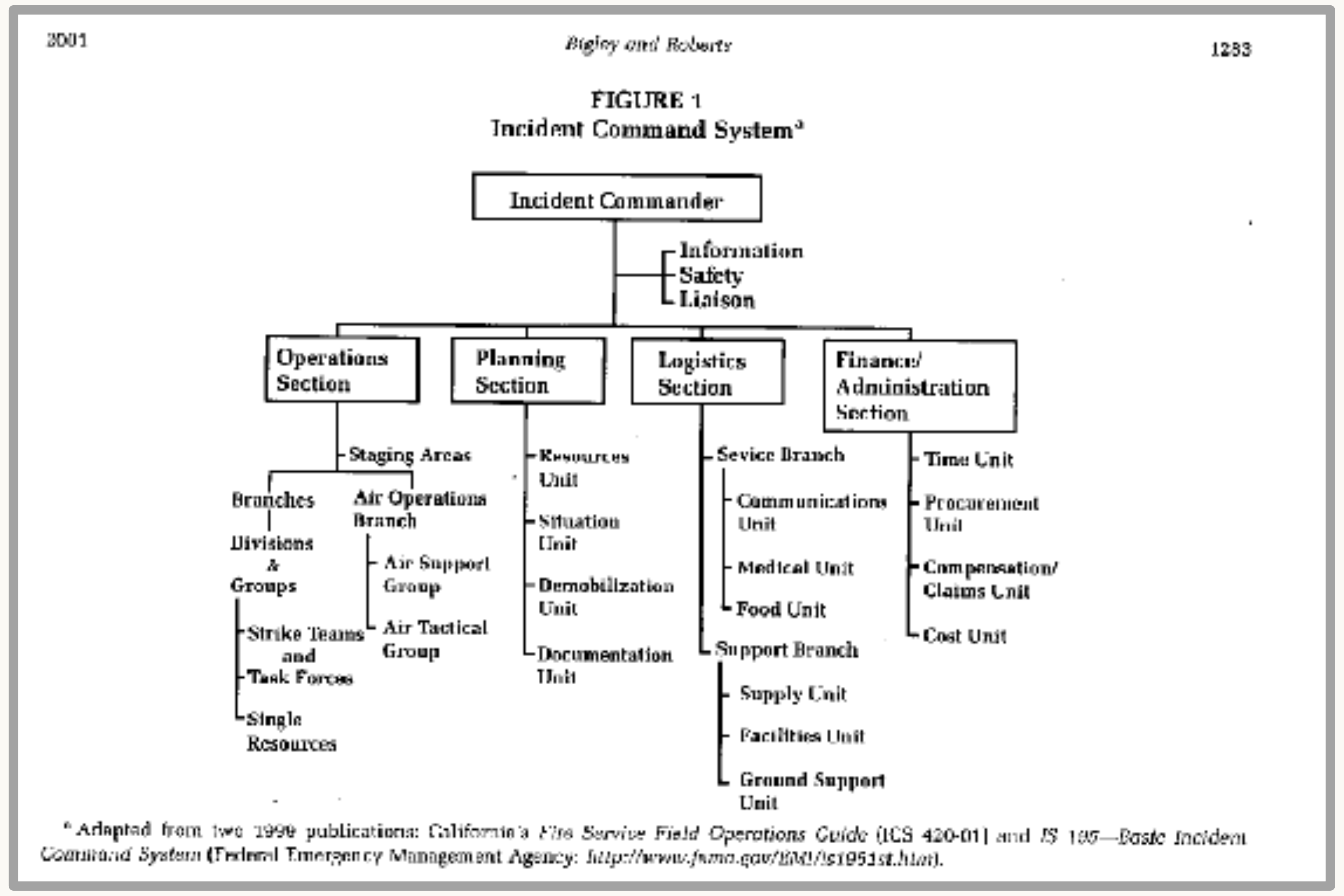
2) Organizational structures require **constant reconfiguration** so that the organization can adapt as it proceeds

...but algorithmic models have not allowed for open-ended reconfiguration

APPROACH: ROLE STRUCTURES

Inspired by 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

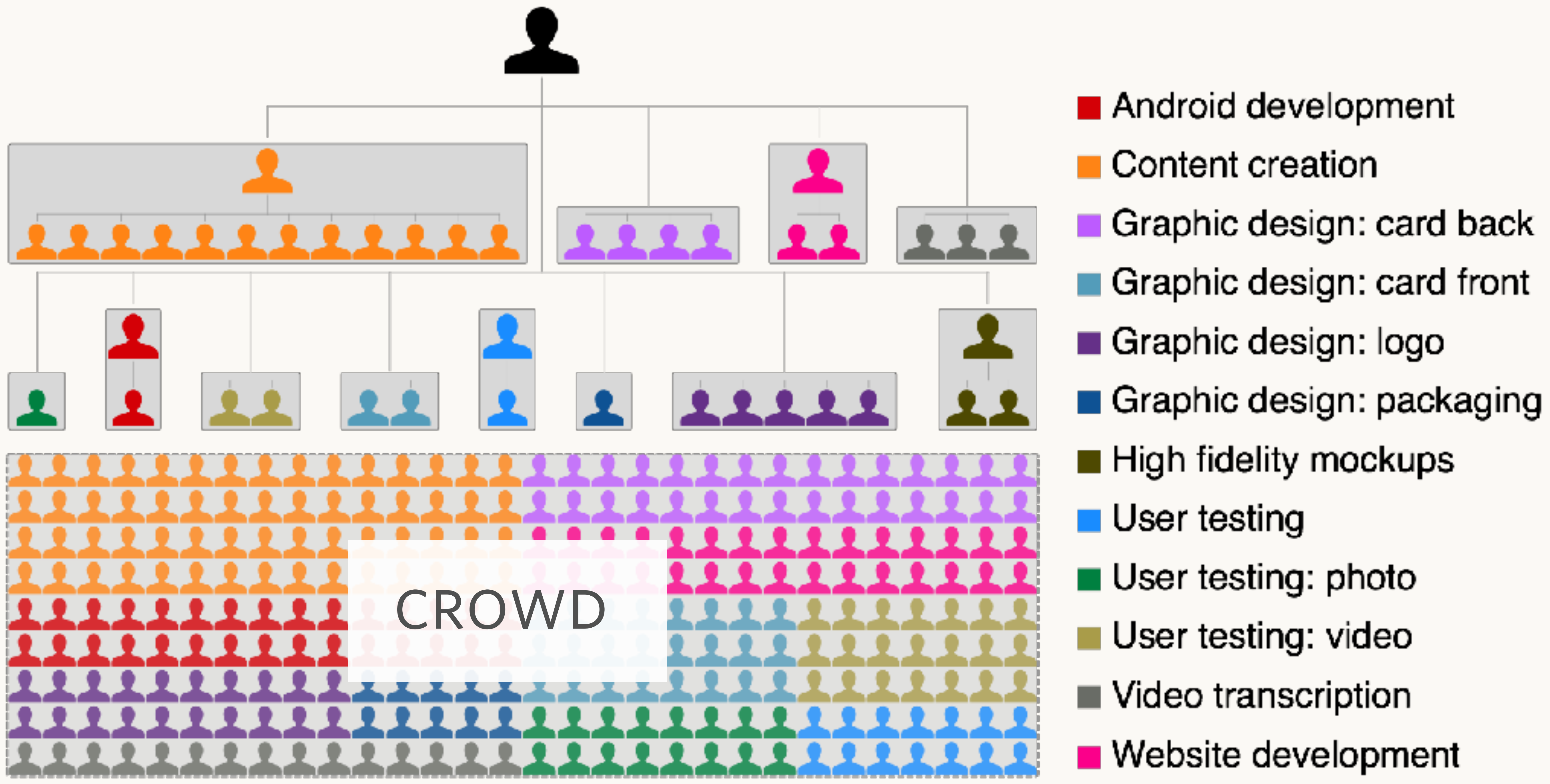


COMPUTATIONAL ORGANIZATIONAL STRUCTURES

Roles: parametrize required expertise

Teams: groups of workers with shared goal

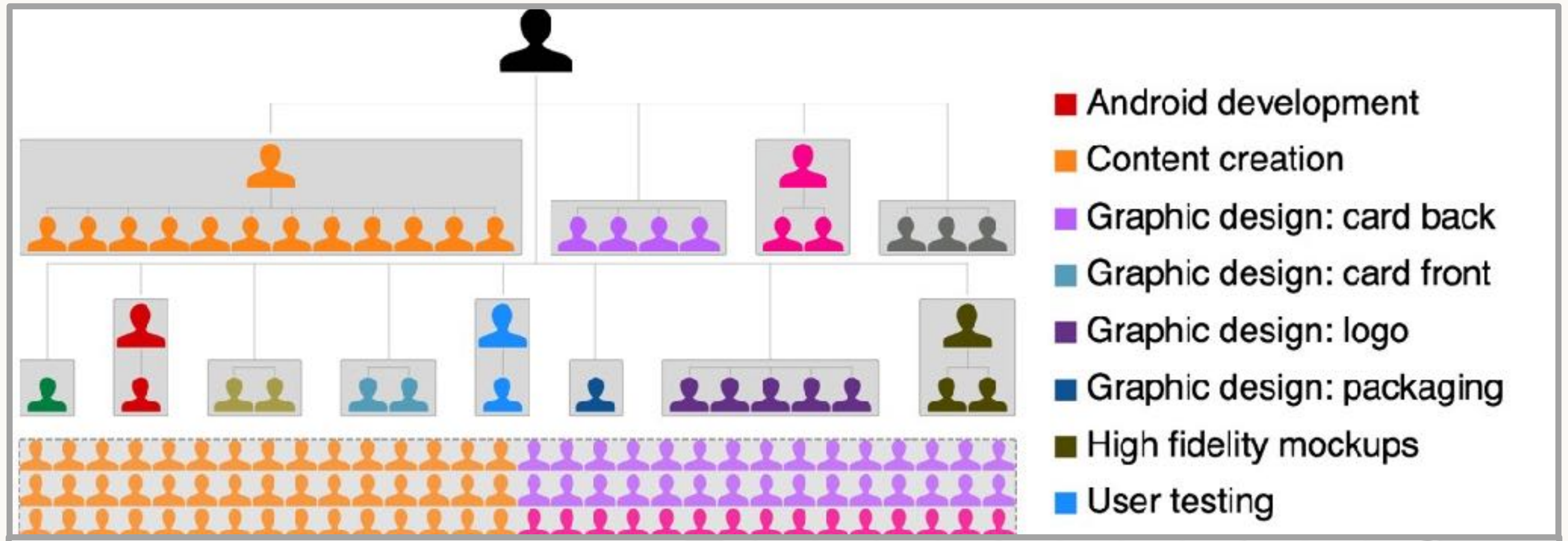
Hierarchy: nested roles that determine decision rights

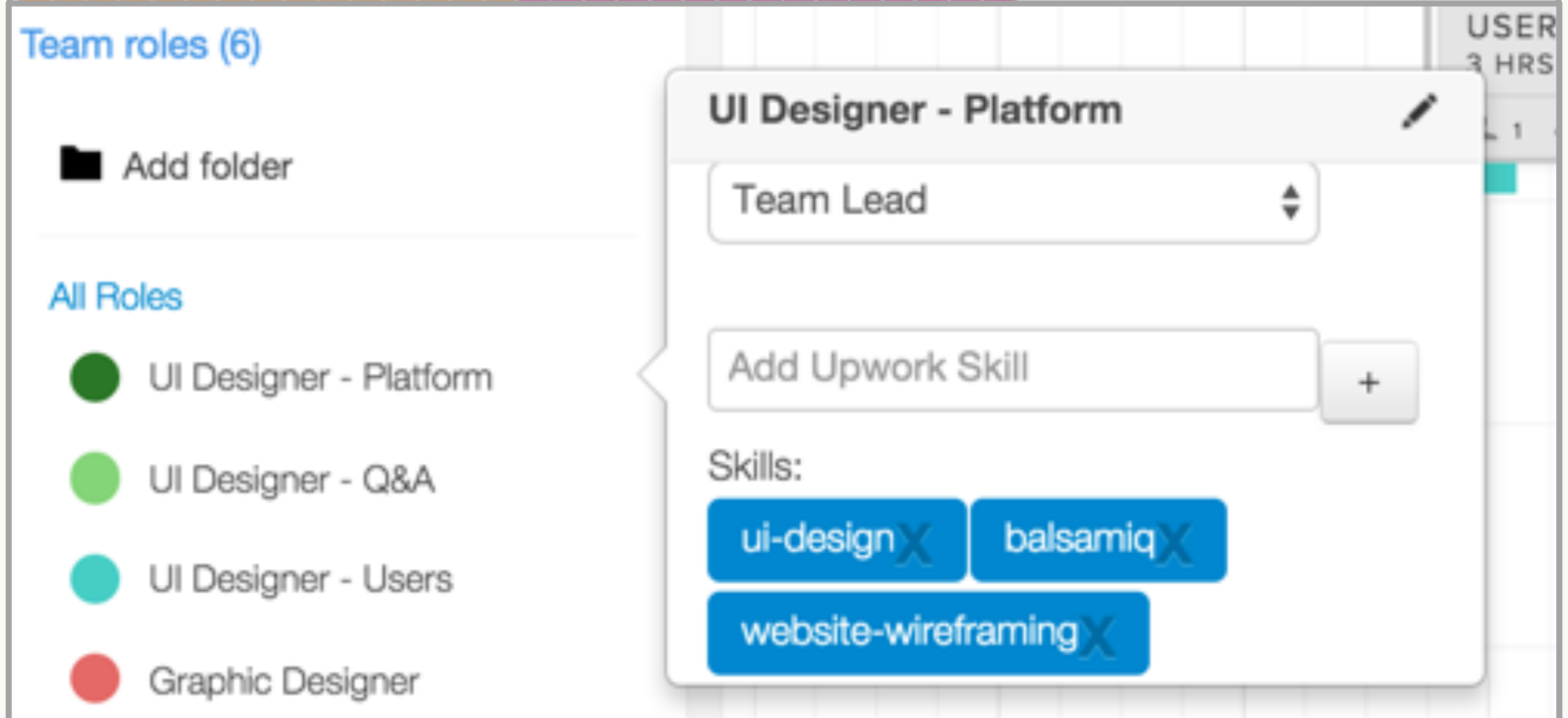


FOUNDRY ROLE STRUCTURES

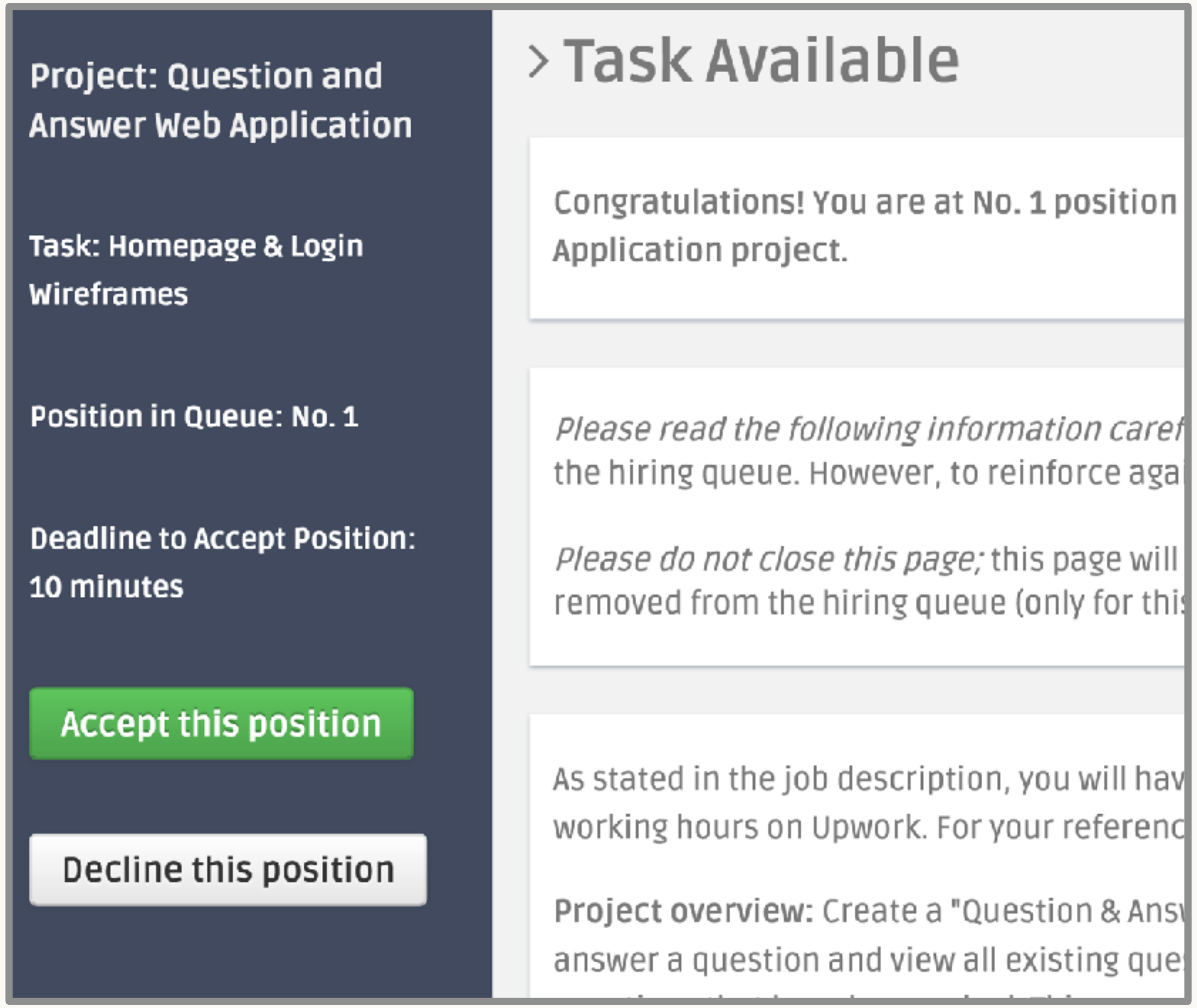
Map each role onto a skill in the Upwork labor market

Nest roles into teams to indicate hierarchy

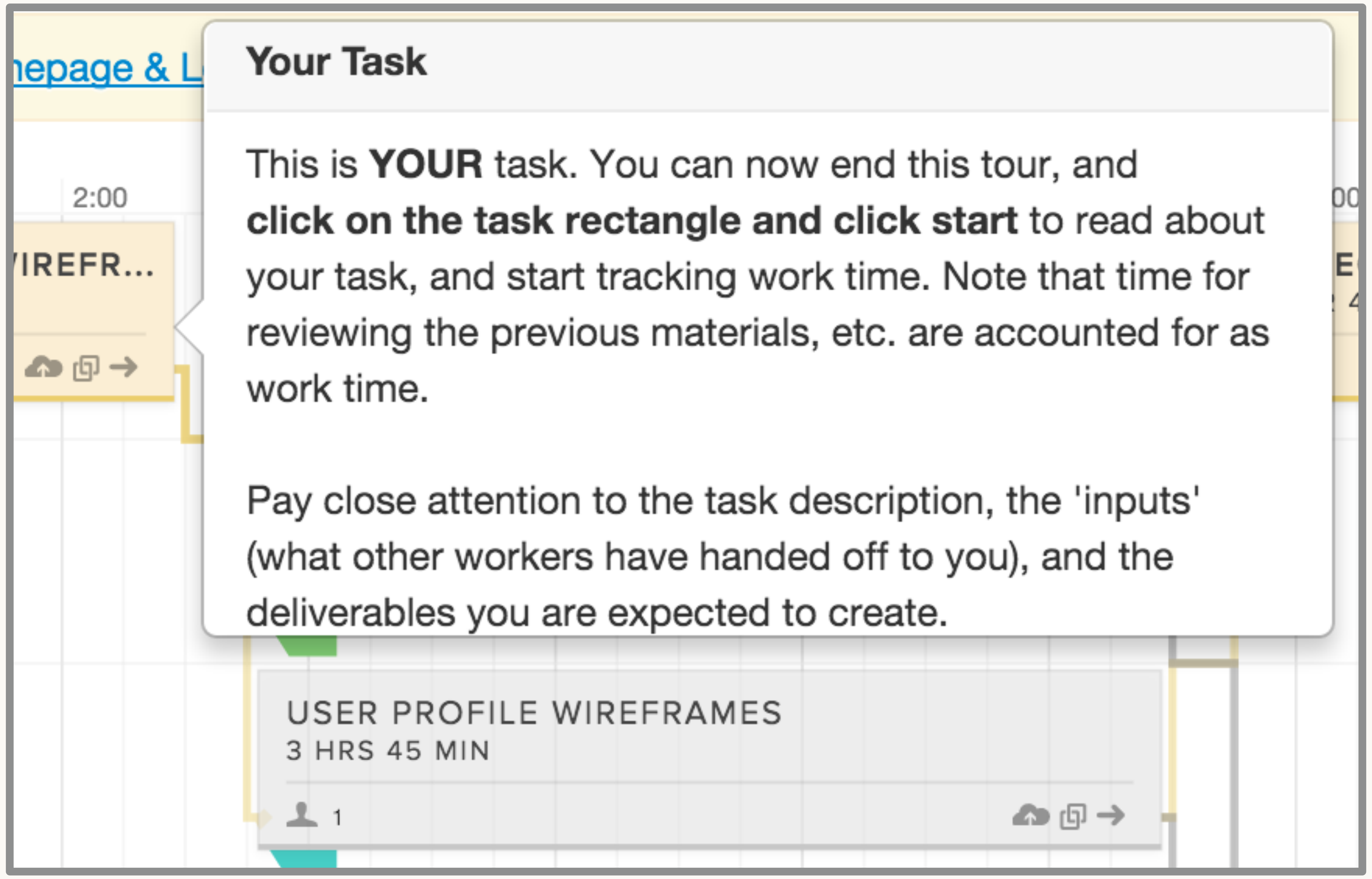




ROLE-BASED HIRING+ONBOARDING



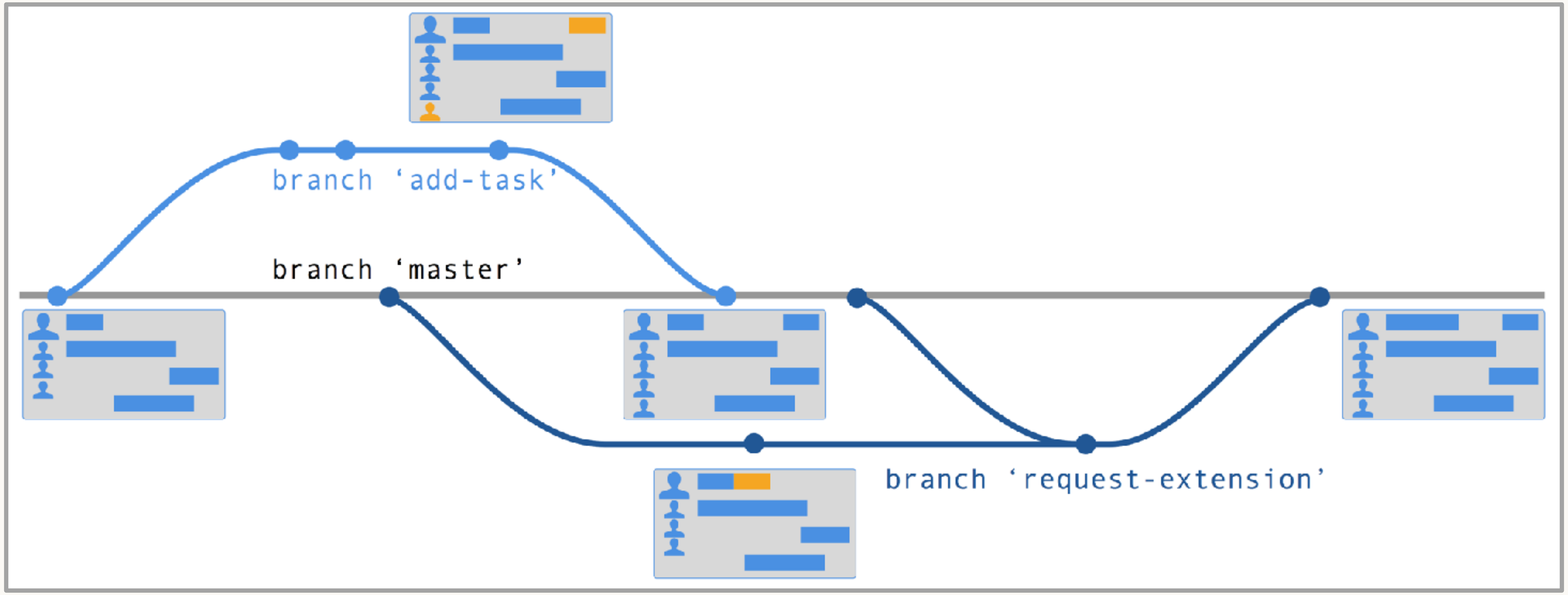
On-demand hiring from the labor market



Automated, role-specific onboarding

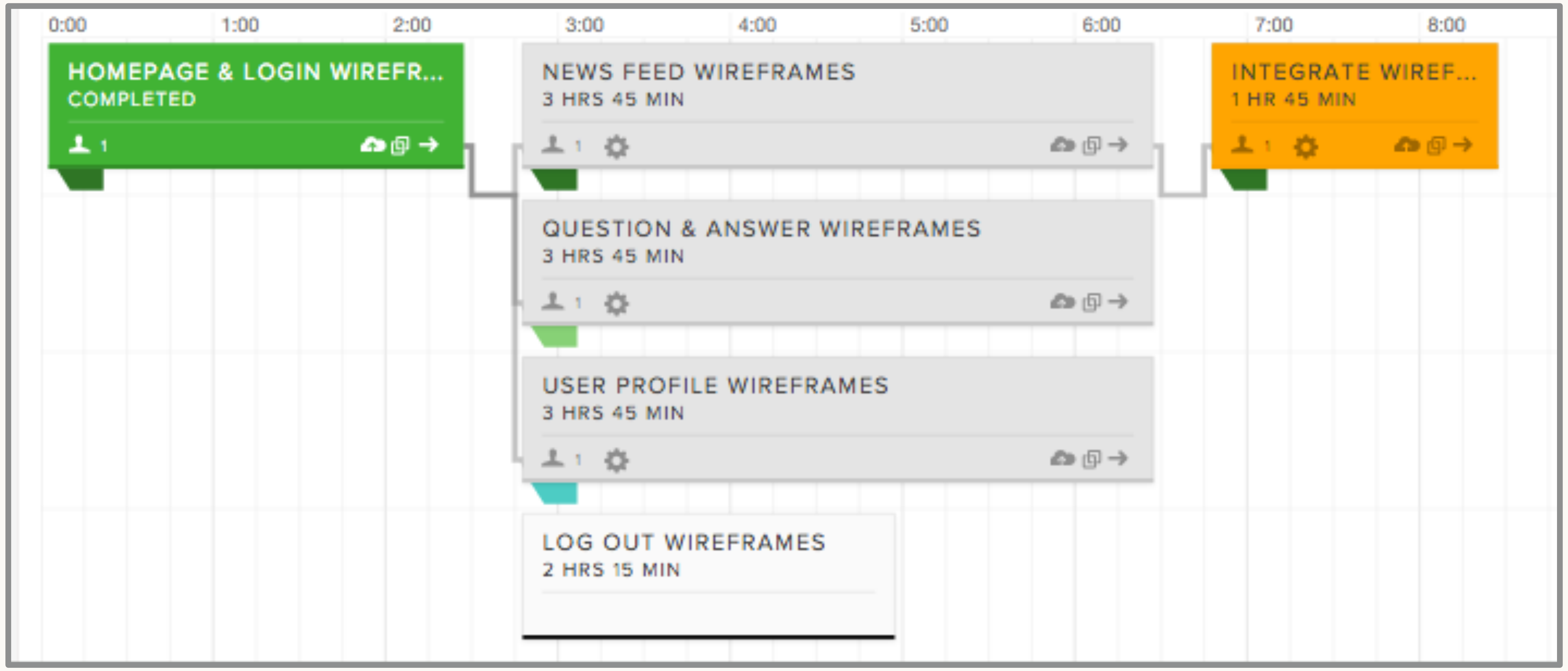
APPROACH: RECONFIGURATION

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



VERSION CONTROL IN FOUNDRY

Any member can branch, edit, and issue 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

COMPUTATIONAL AFFORDANCES

1) Asset specificity → **Hierarchical role structures**

Rapid hiring and onboarding

2) Adaptation → **Reconfigurable tasks, teams, and hierarchy: top-down and bottom-up**

Branch+merge version control-style reconfiguration

EVALUATION

Field study: recruit outside leaders to pursue open-ended goals that have remained out of reach for crowdsourcing

	EMS Report		
Leader	Medical resident		
Open-ended goal	Develop prototype application for EMTs to transmit patient information en route to hospital		

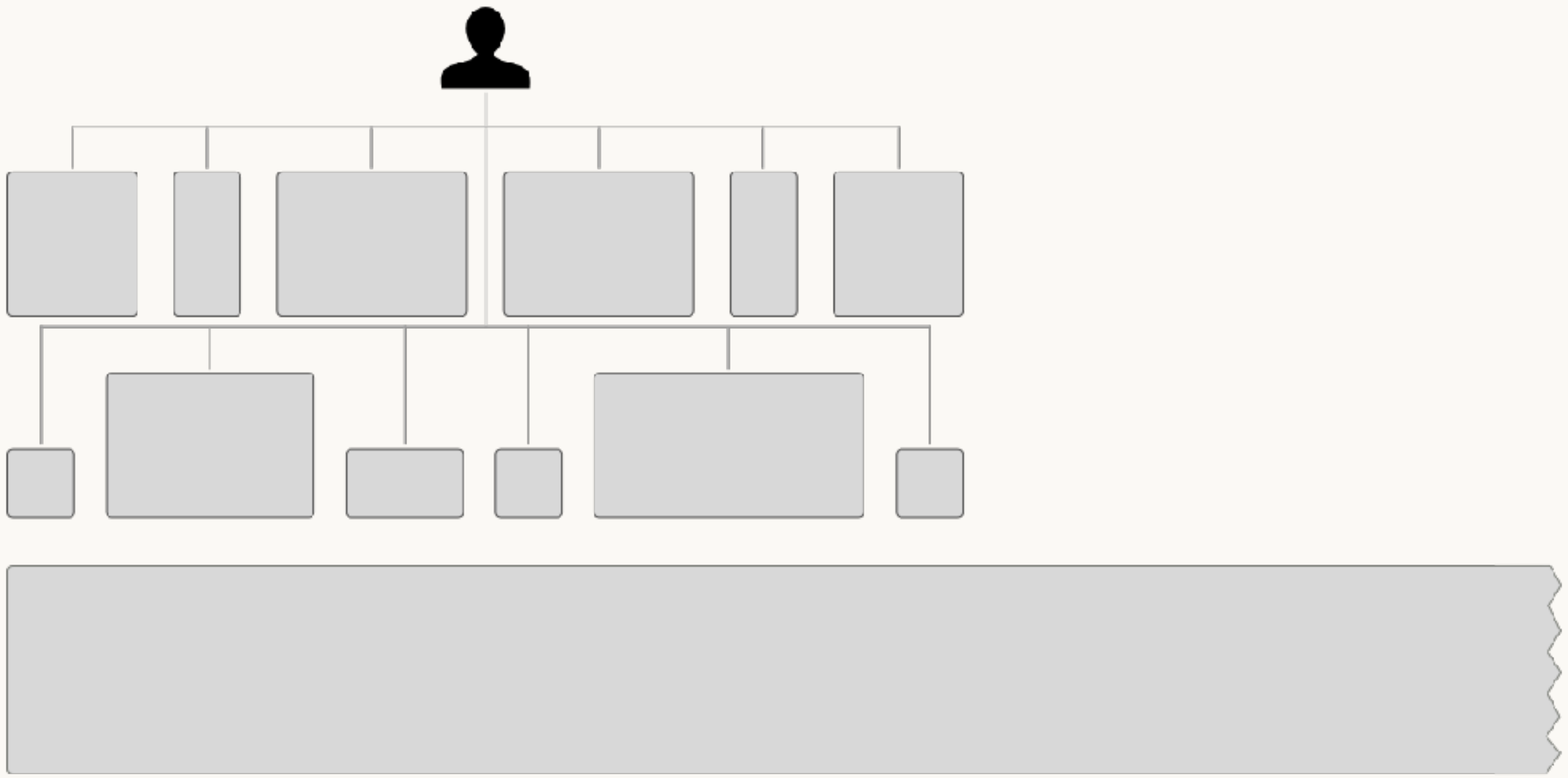
End users spun up and led entire organizations in six weeks, convening new workers on-demand within fourteen minutes on average.

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

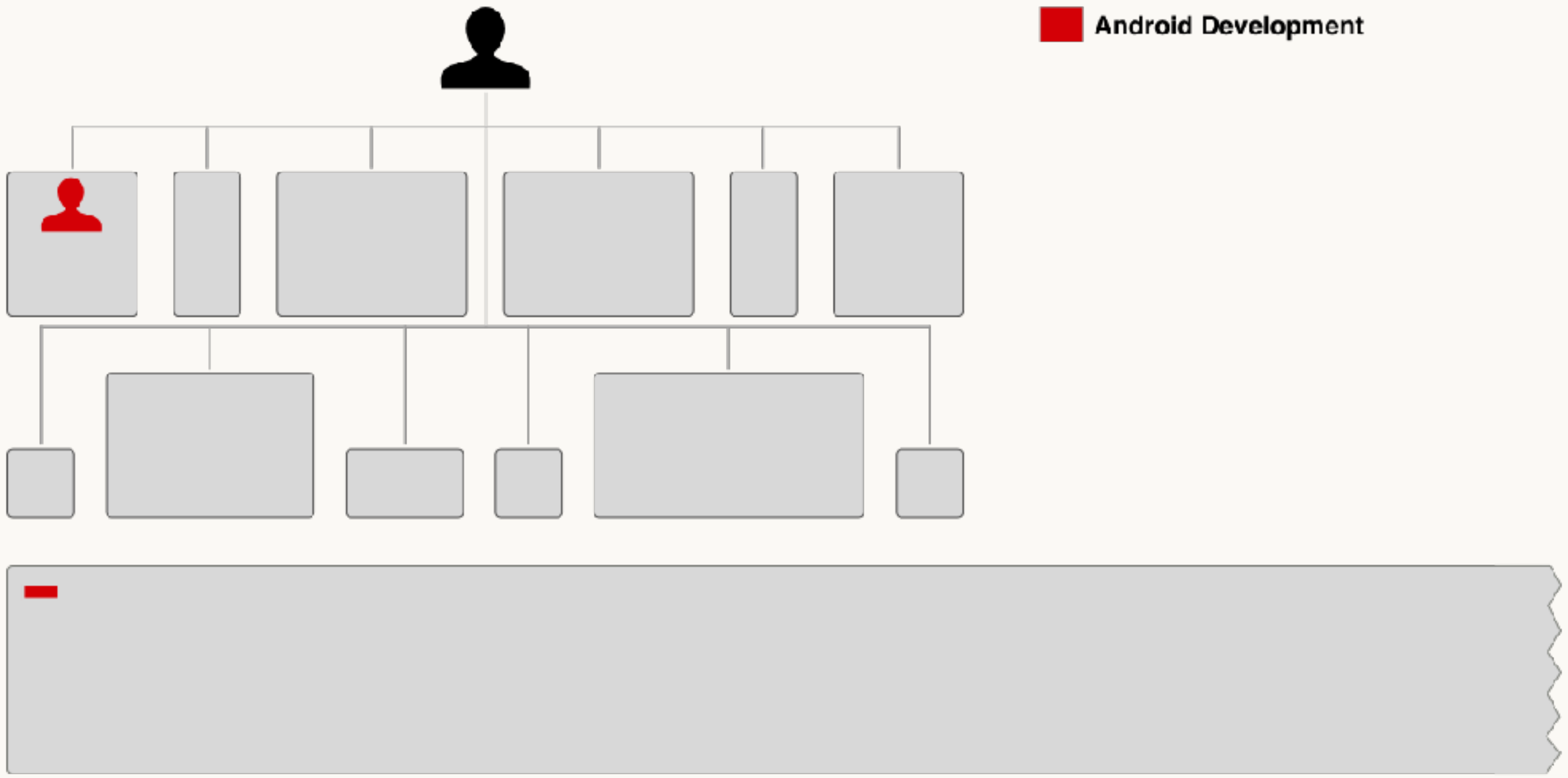
639 tasks, 3261 person-hours of work across 35–46 days from engineers, designers, testers, poets, and others

Passed quality review by neutral experts

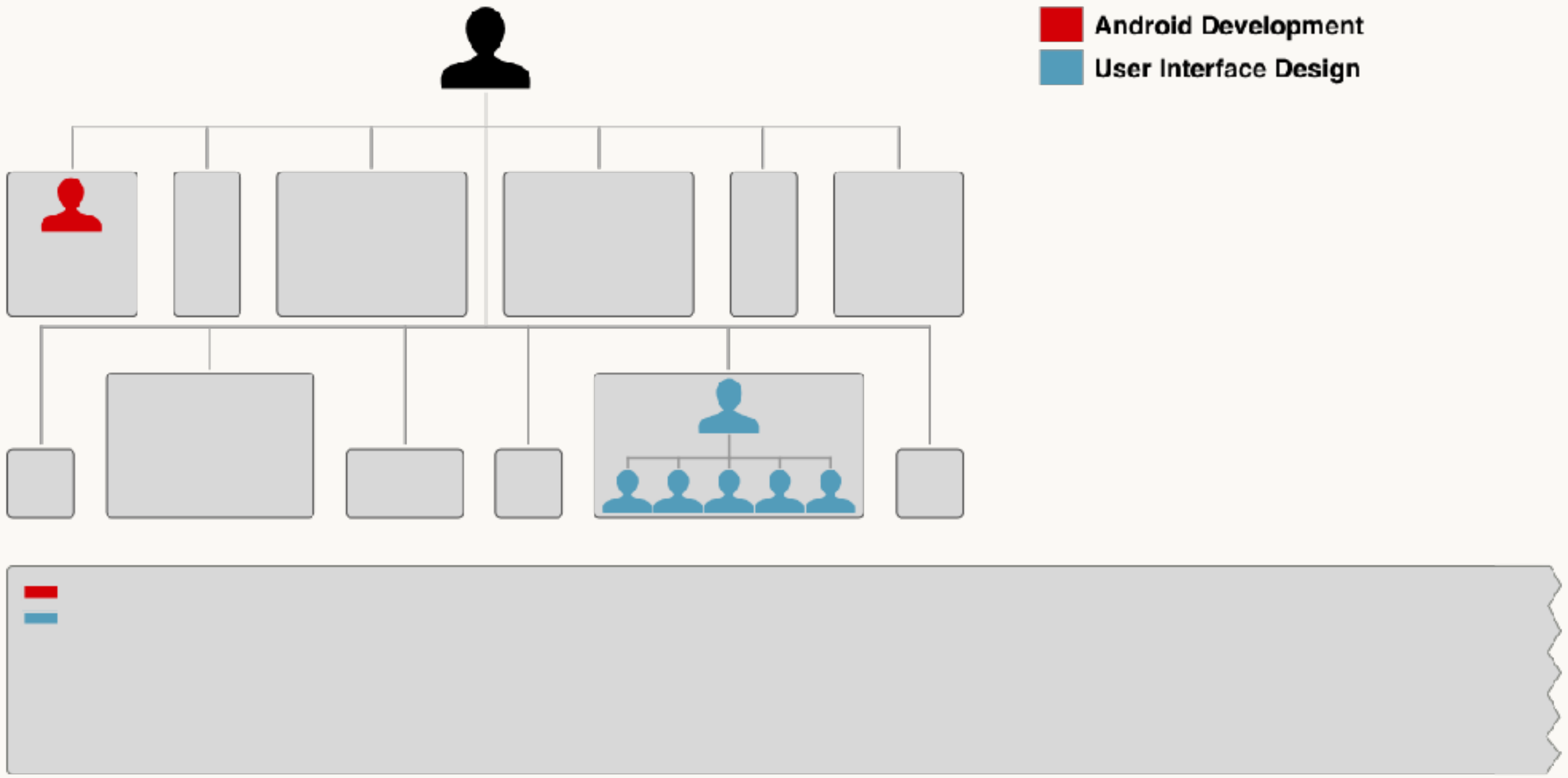
EMS TRAUMA REPORT



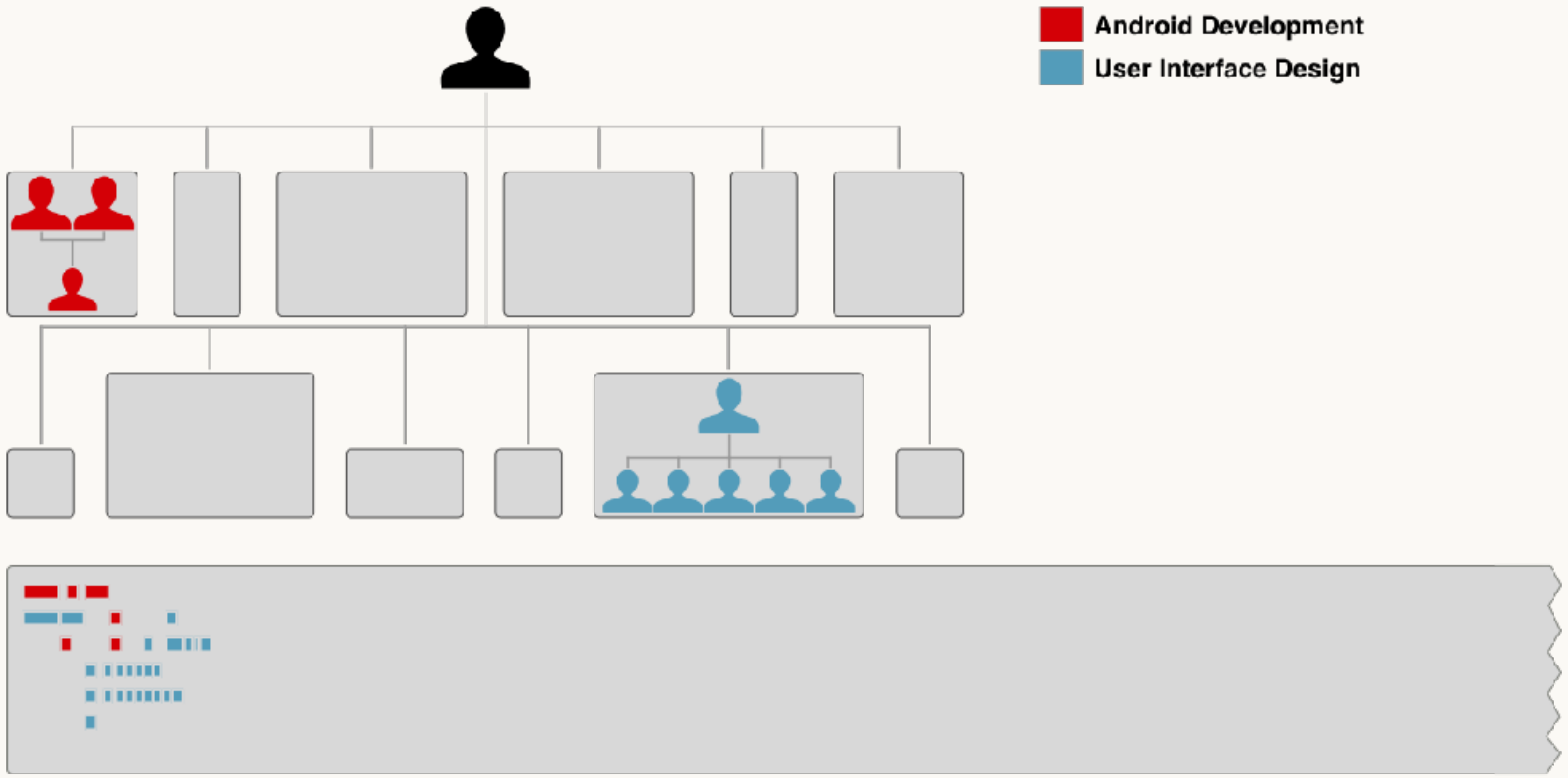
EMS TRAUMA REPORT



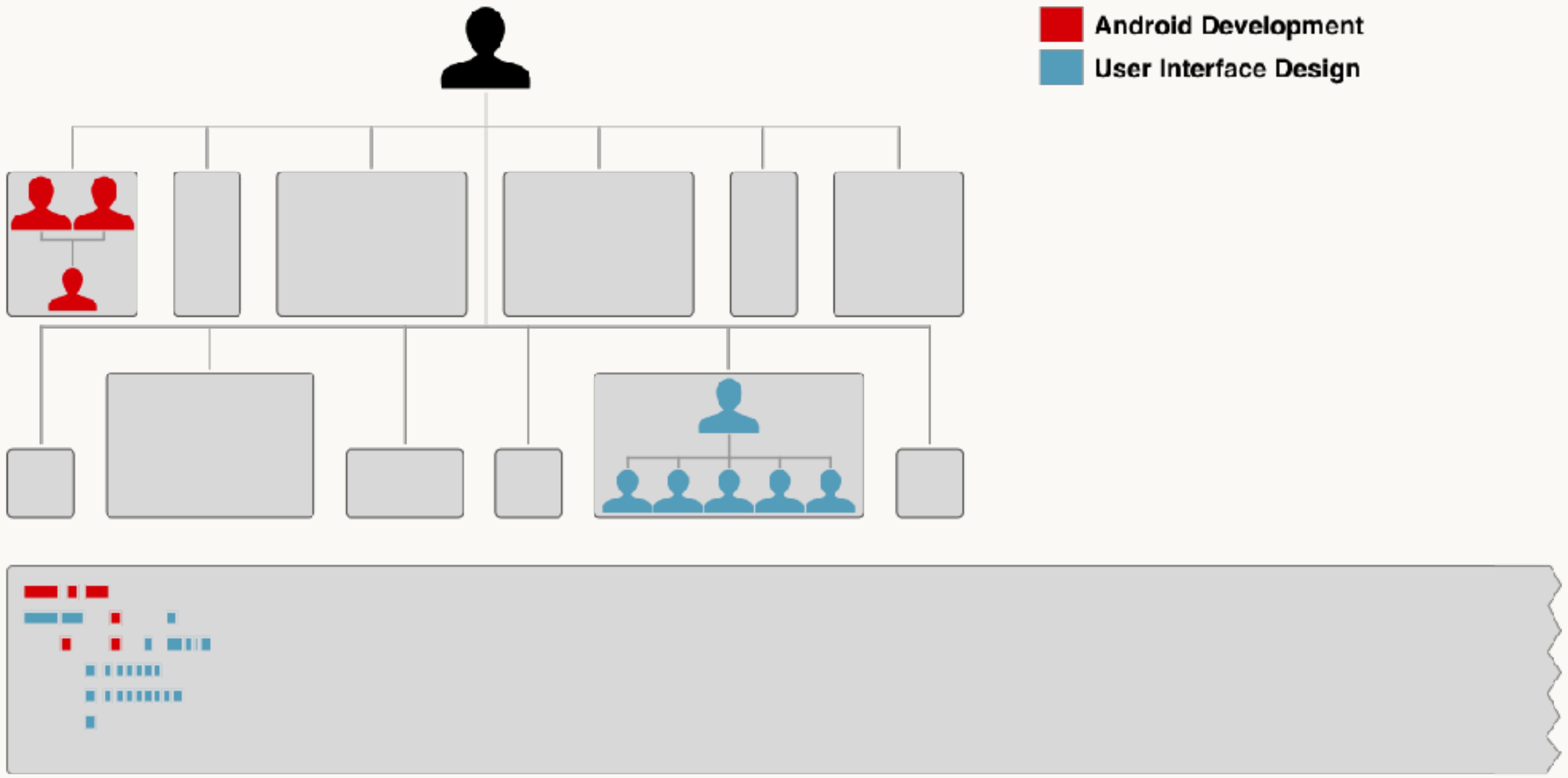
EMS TRAUMA REPORT

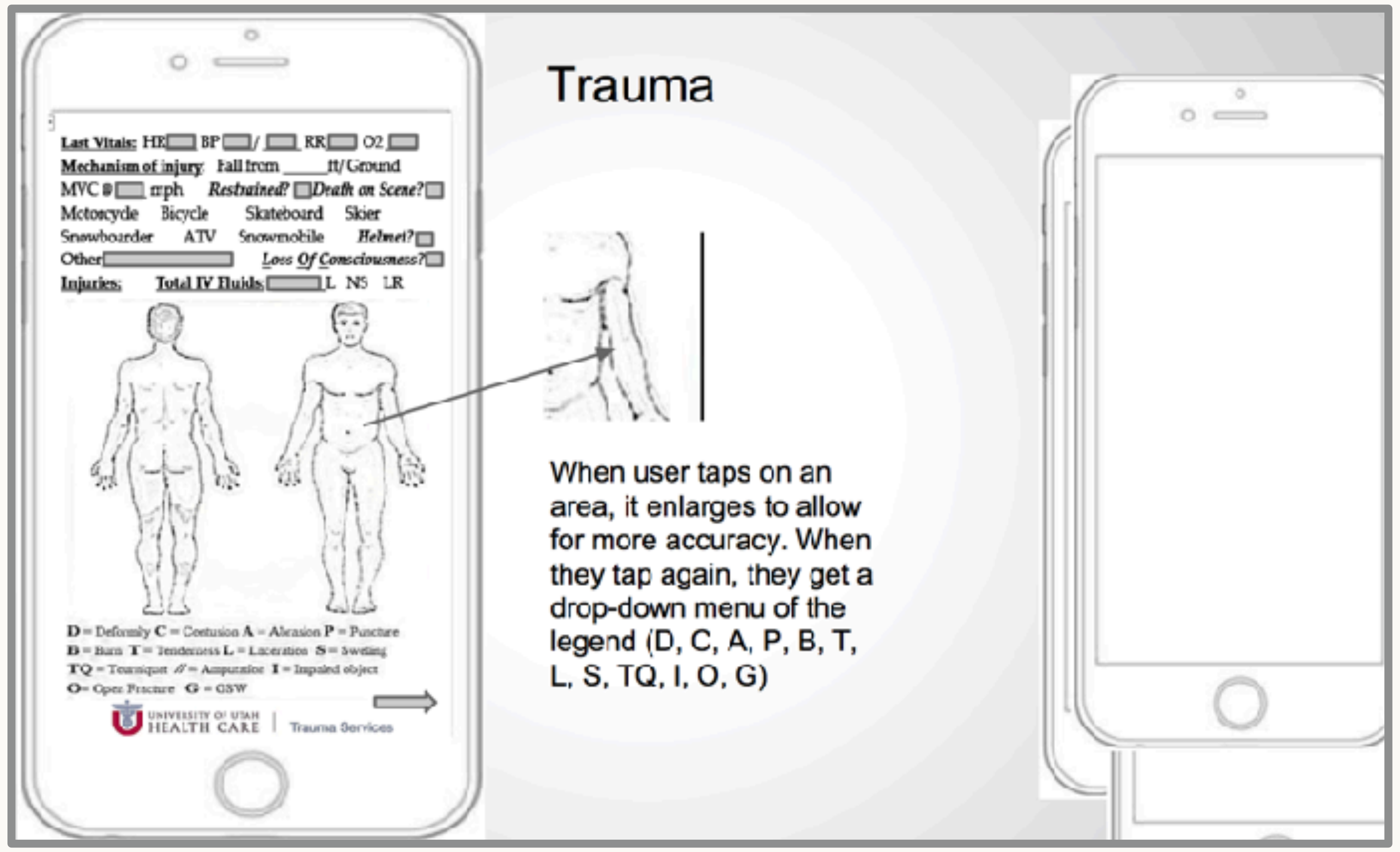


EMS TRAUMA REPORT

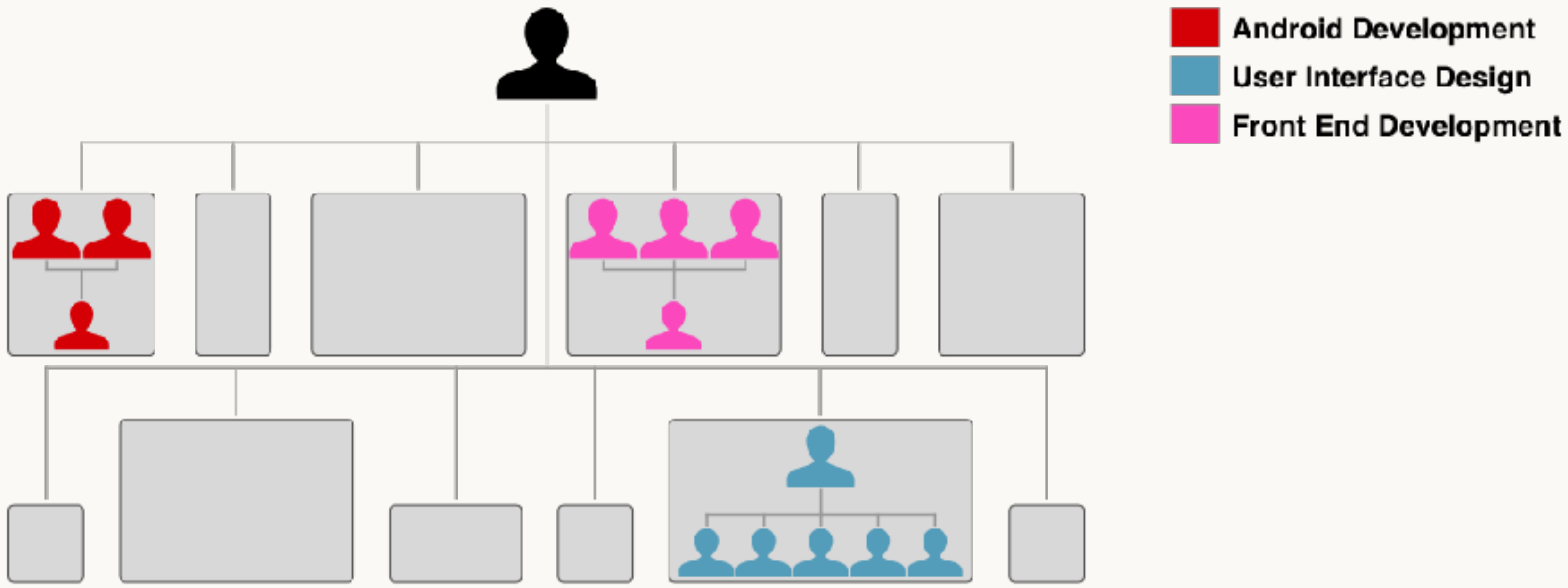


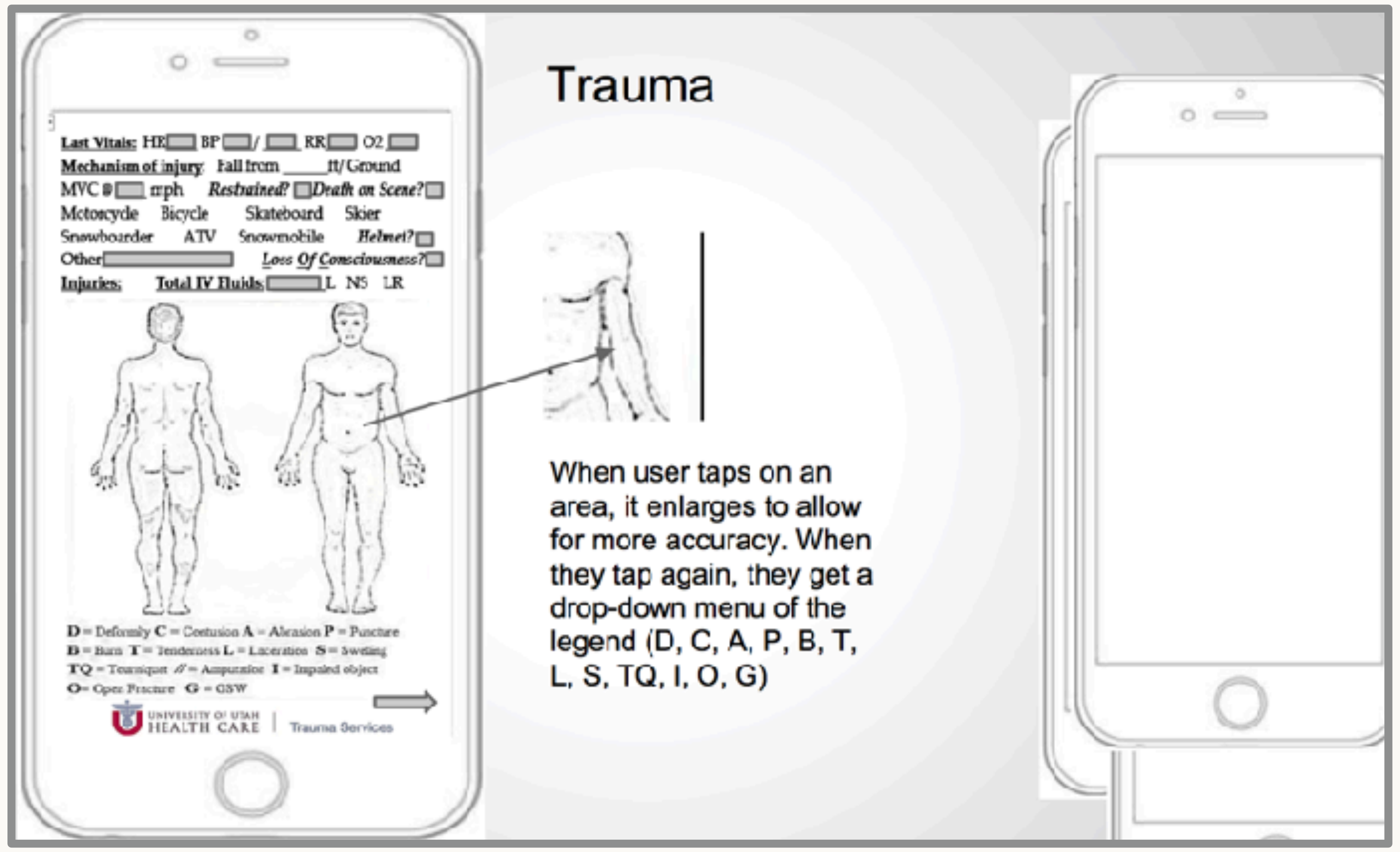
EMS TRAUMA REPORT



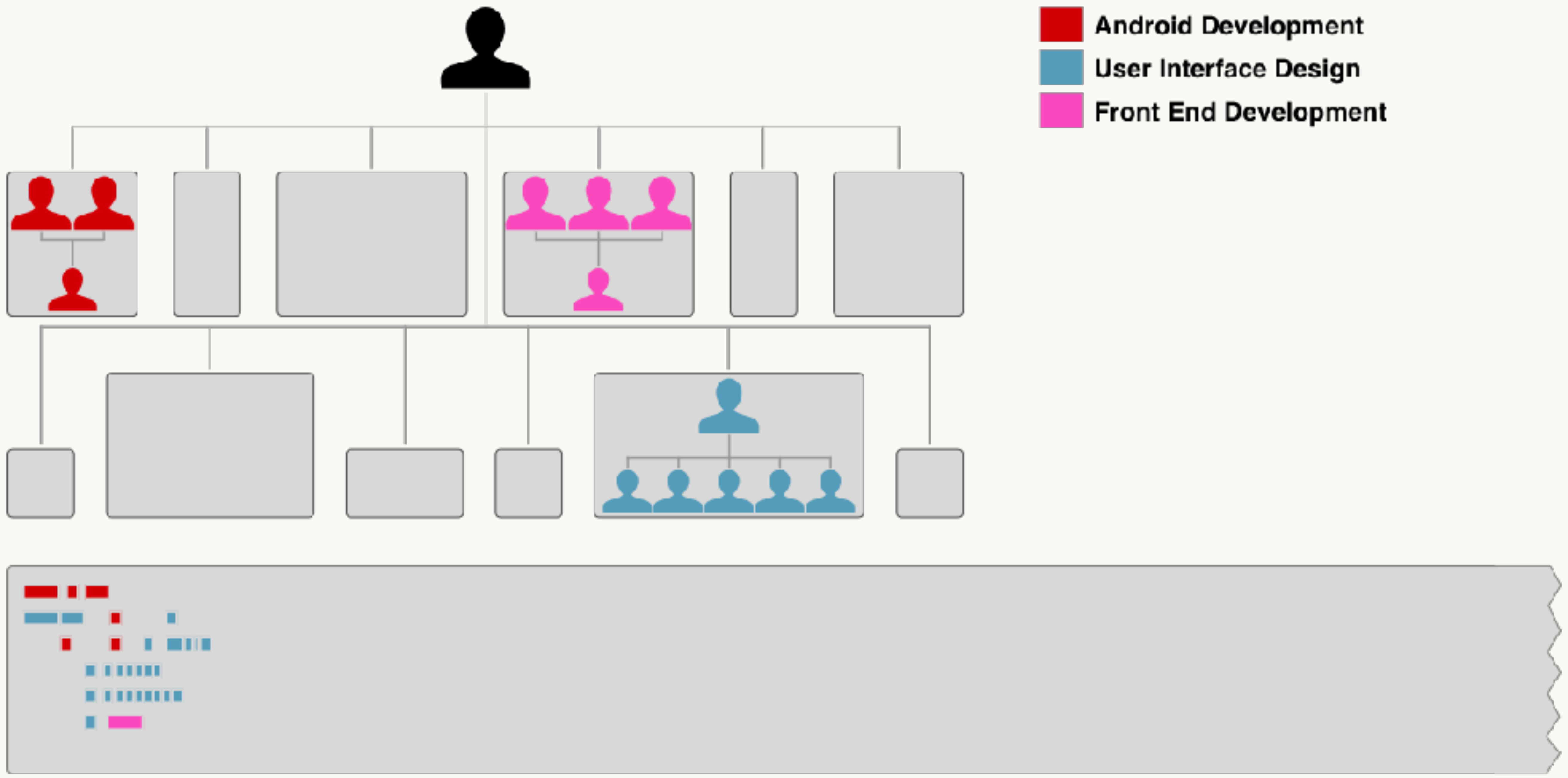


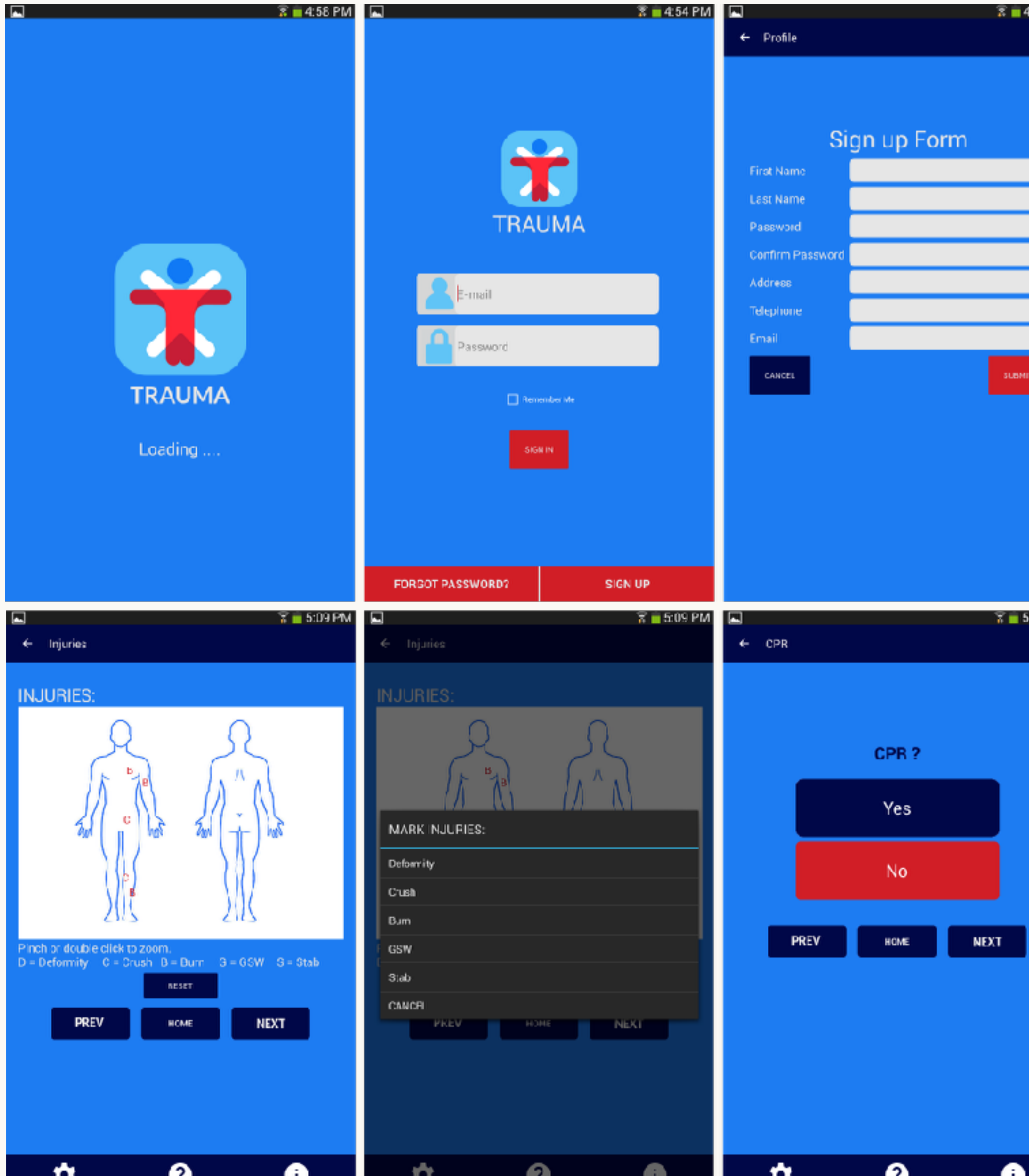
EMS TRAUMA REPORT



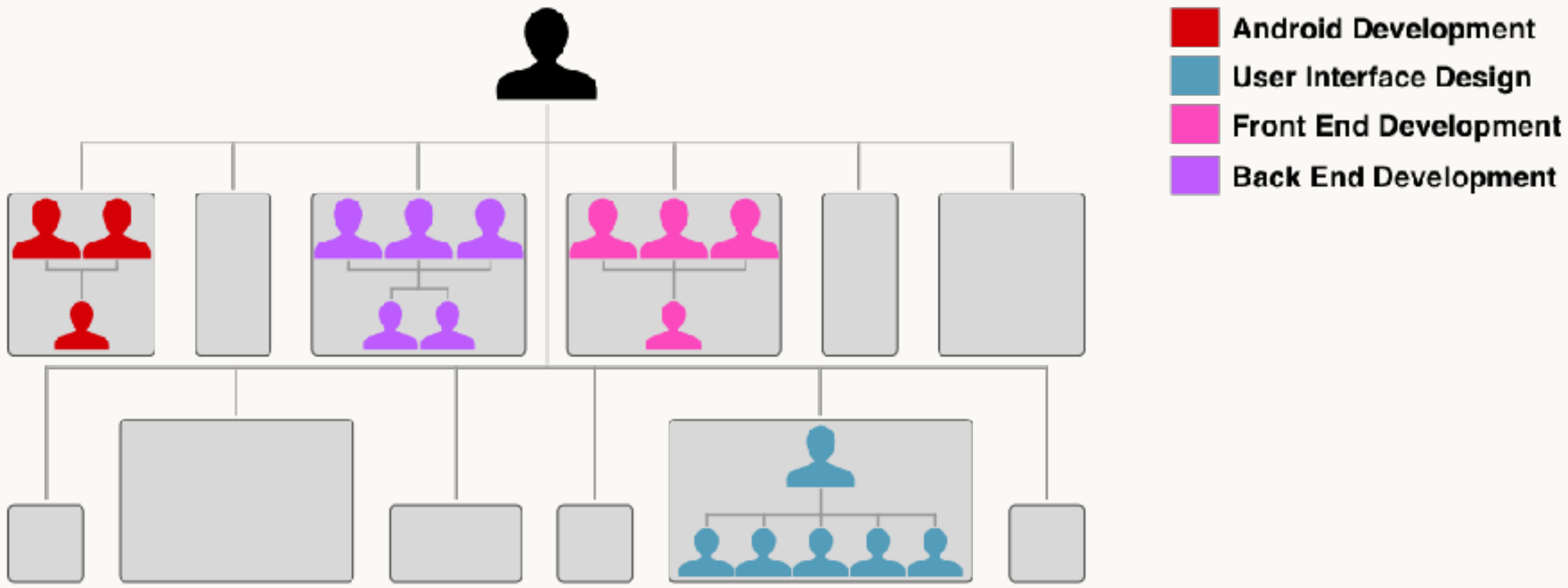


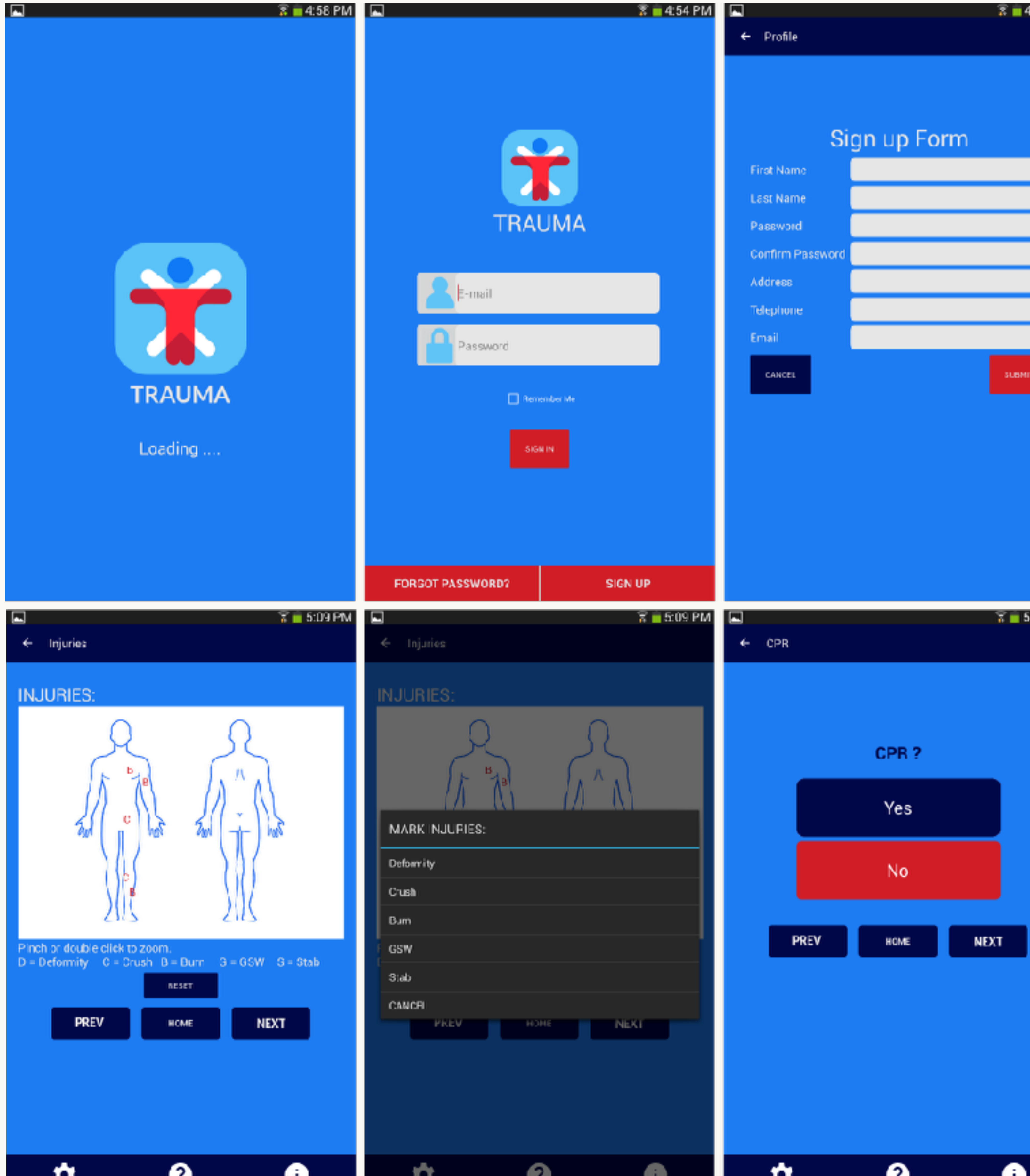
EMS TRAUMA REPORT



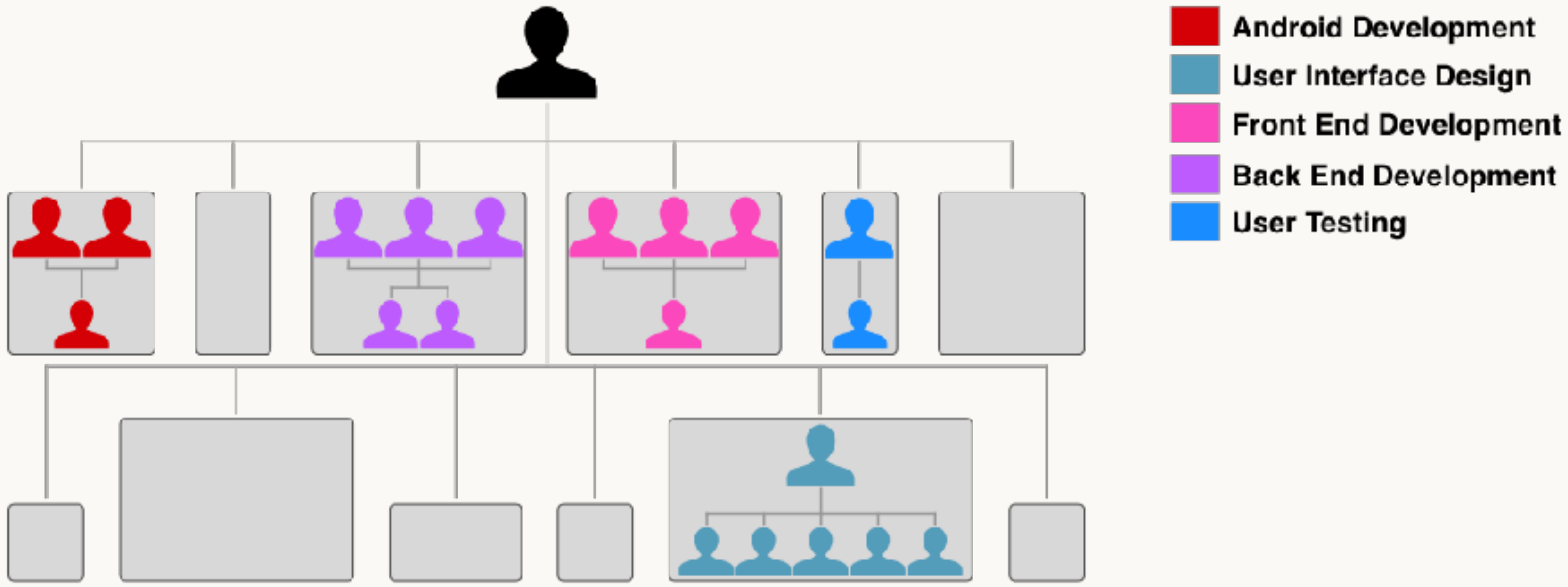


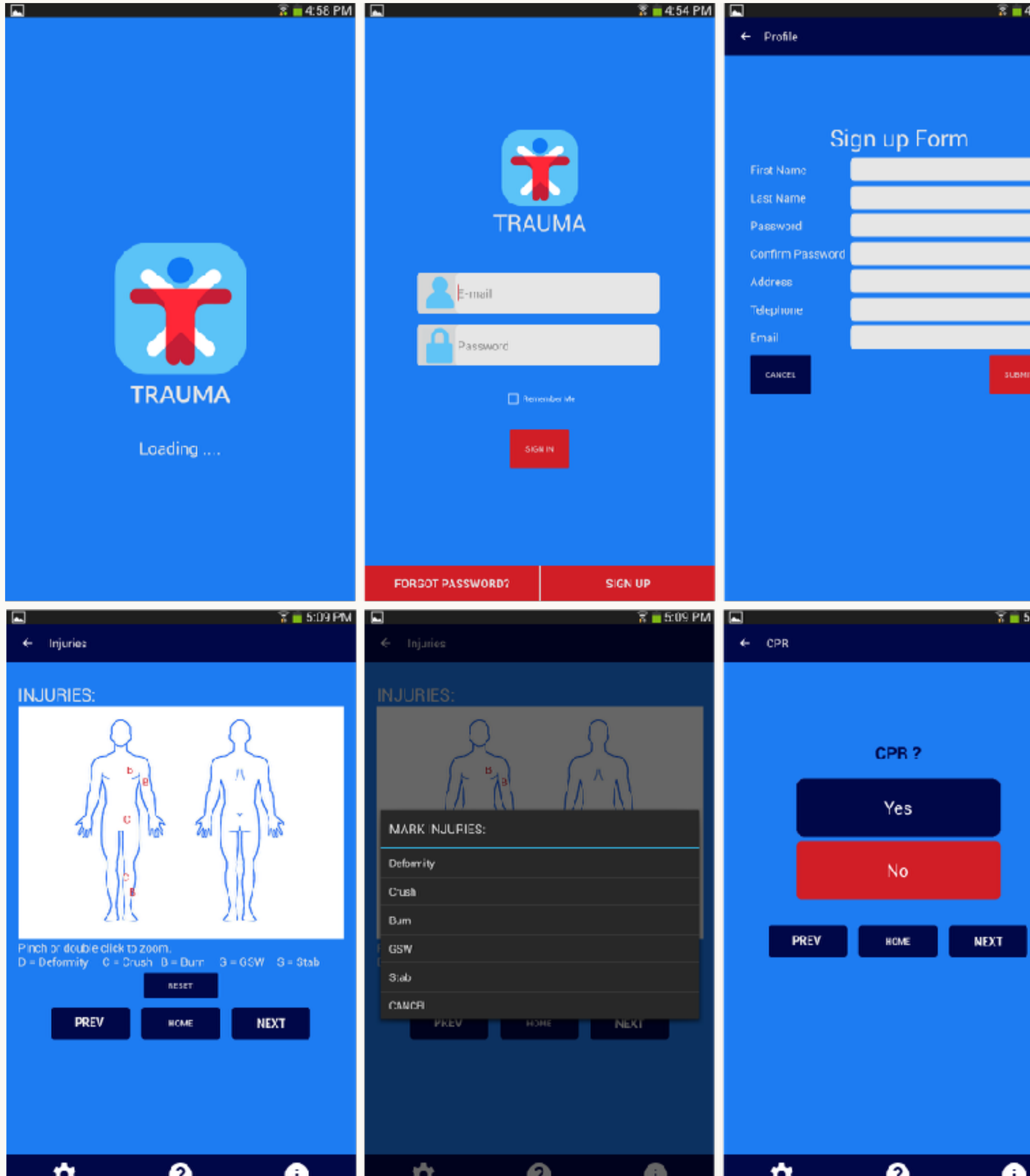
EMS TRAUMA REPORT



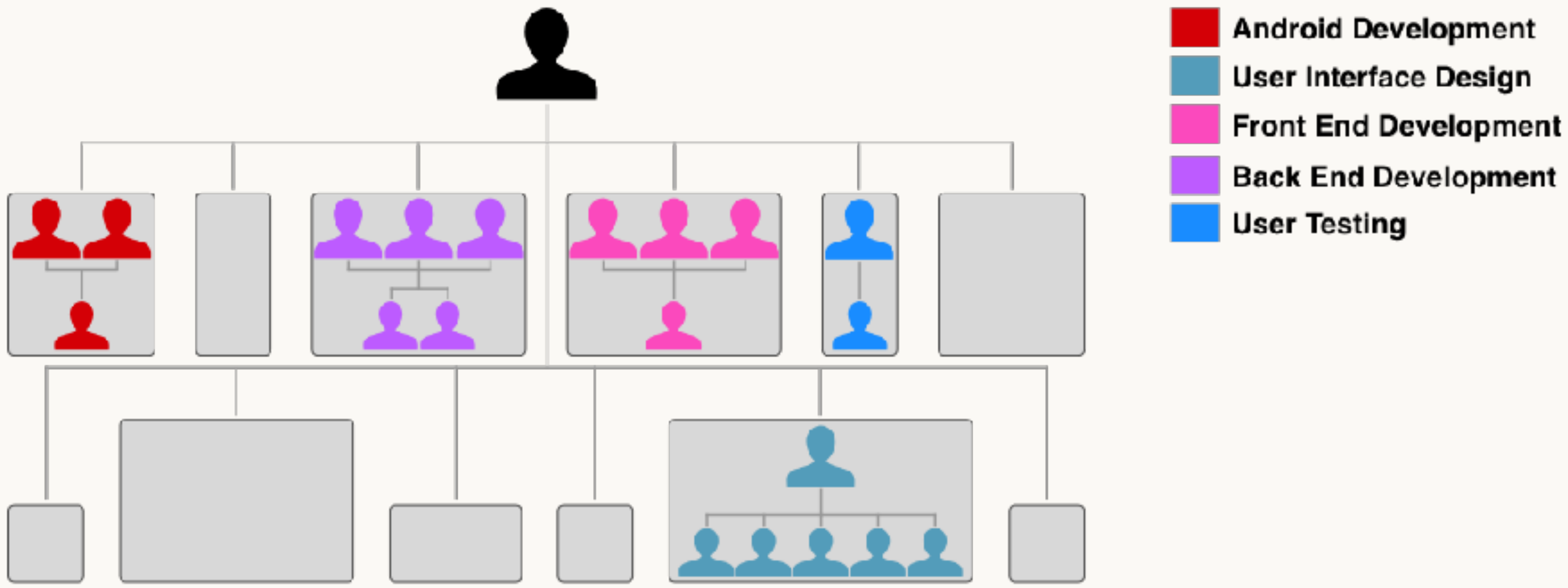


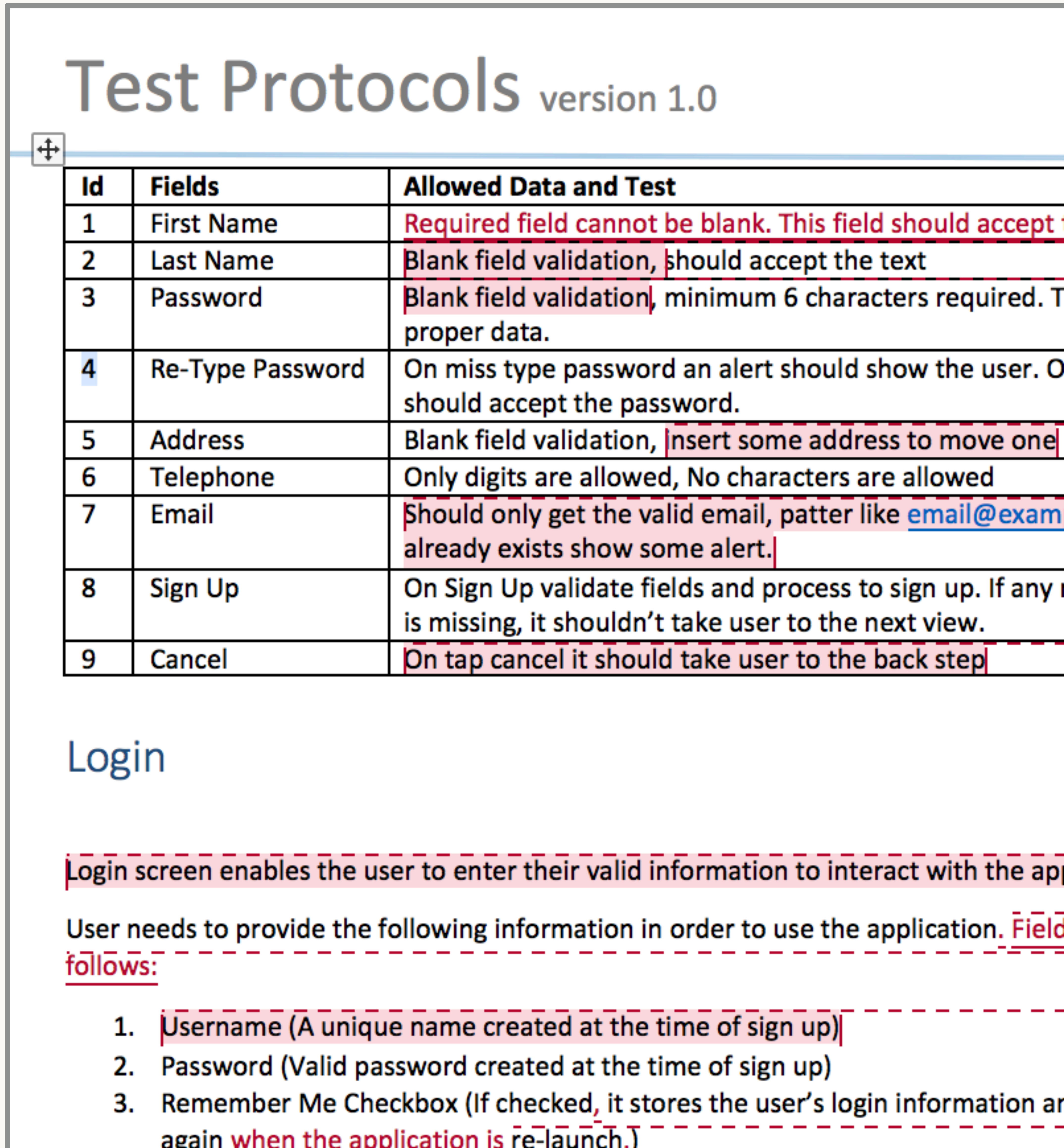
EMS TRAUMA REPORT





EMS TRAUMA REPORT





Test Protocols version 1.0

Id	Fields	Allowed Data and Test
1	First Name	Required field cannot be blank. This field should accept
2	Last Name	Blank field validation, should accept the text
3	Password	Blank field validation, minimum 6 characters required. T proper data.
4	Re-Type Password	On miss type password an alert should show the user. O should accept the password.
5	Address	Blank field validation, Insert some address to move one
6	Telephone	Only digits are allowed, No characters are allowed
7	Email	Should only get the valid email, patter like email@exam already exists show some alert.
8	Sign Up	On Sign Up validate fields and process to sign up. If any is missing, it shouldn't take user to the next view.
9	Cancel	On tap cancel it should take user to the back step

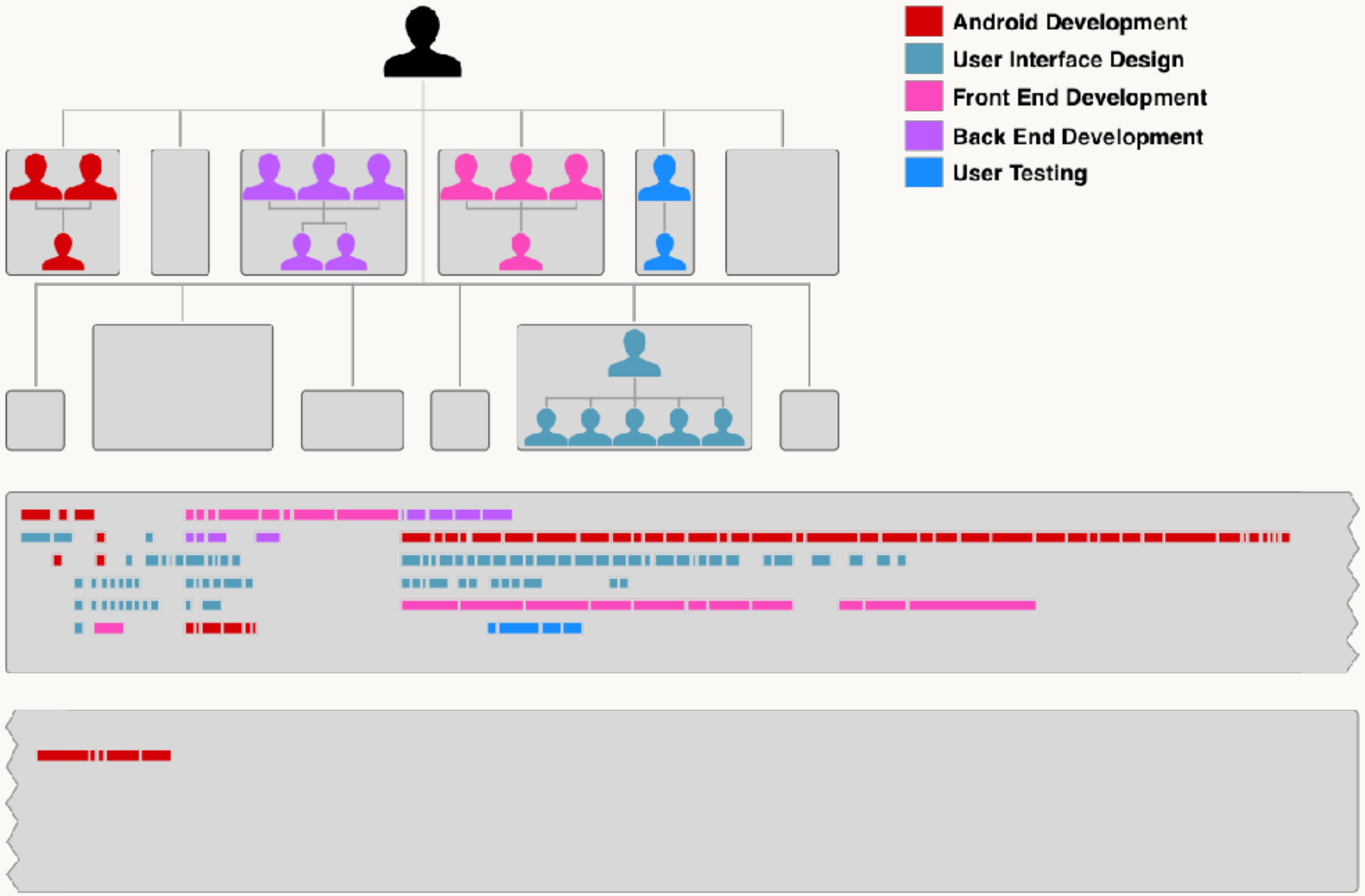
Login

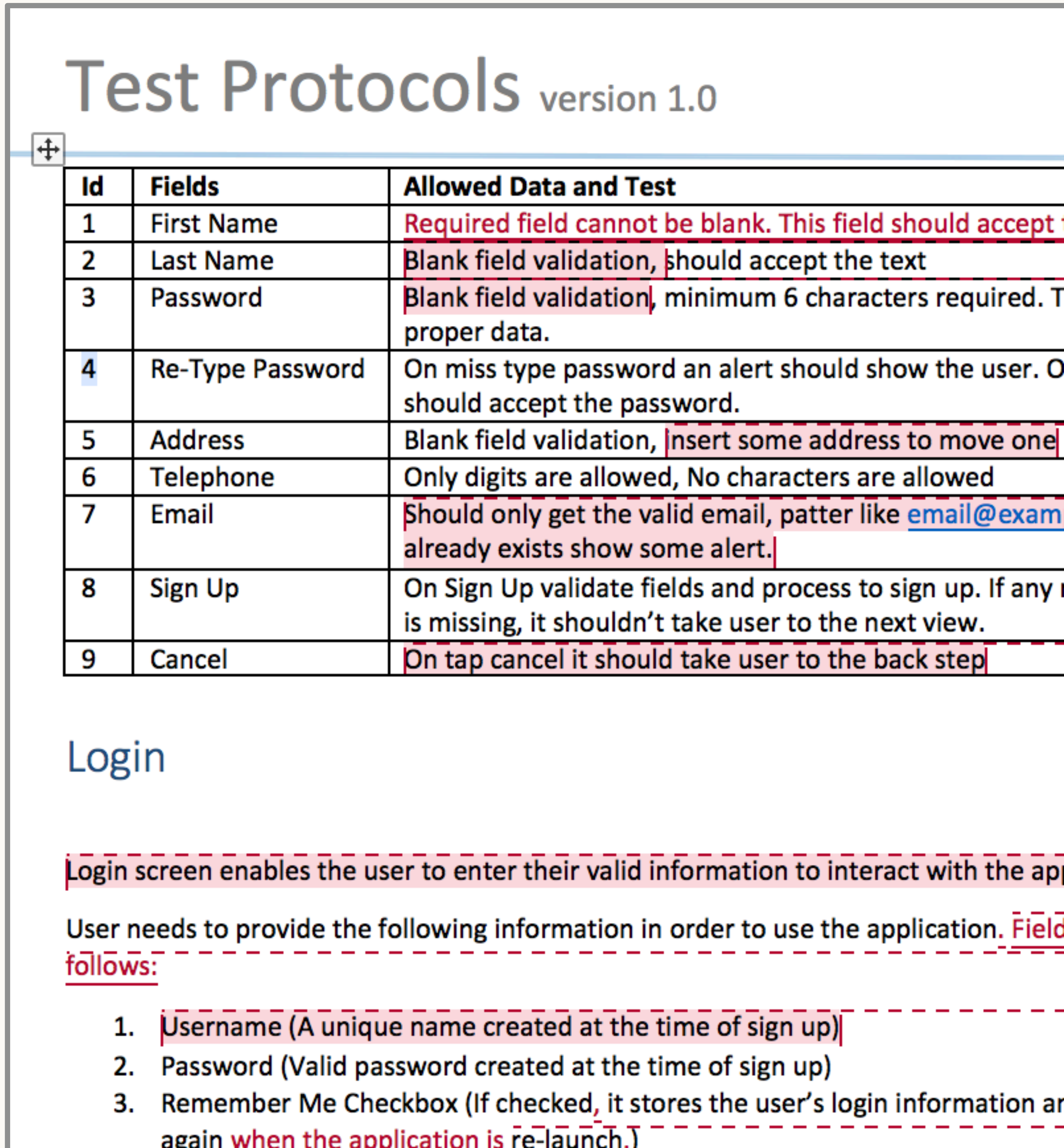
Login screen enables the user to enter their valid information to interact with the ap

User needs to provide the following information in order to use the application. Field follows:

1. Username (A unique name created at the time of sign up)
2. Password (Valid password created at the time of sign up)
3. Remember Me Checkbox (If checked, it stores the user's login information ar again when the application is re-launch.)

EMS TRAUMA REPORT





Test Protocols version 1.0

Id	Fields	Allowed Data and Test
1	First Name	Required field cannot be blank. This field should accept t
2	Last Name	Blank field validation, should accept the text
3	Password	Blank field validation, minimum 6 characters required. T proper data.
4	Re-Type Password	On miss type password an alert should show the user. O should accept the password.
5	Address	Blank field validation, Insert some address to move one
6	Telephone	Only digits are allowed, No characters are allowed
7	Email	Should only get the valid email, patter like email@exam already exists show some alert.
8	Sign Up	On Sign Up validate fields and process to sign up. If any is missing, it shouldn't take user to the next view.
9	Cancel	On tap cancel it should take user to the back step

Login

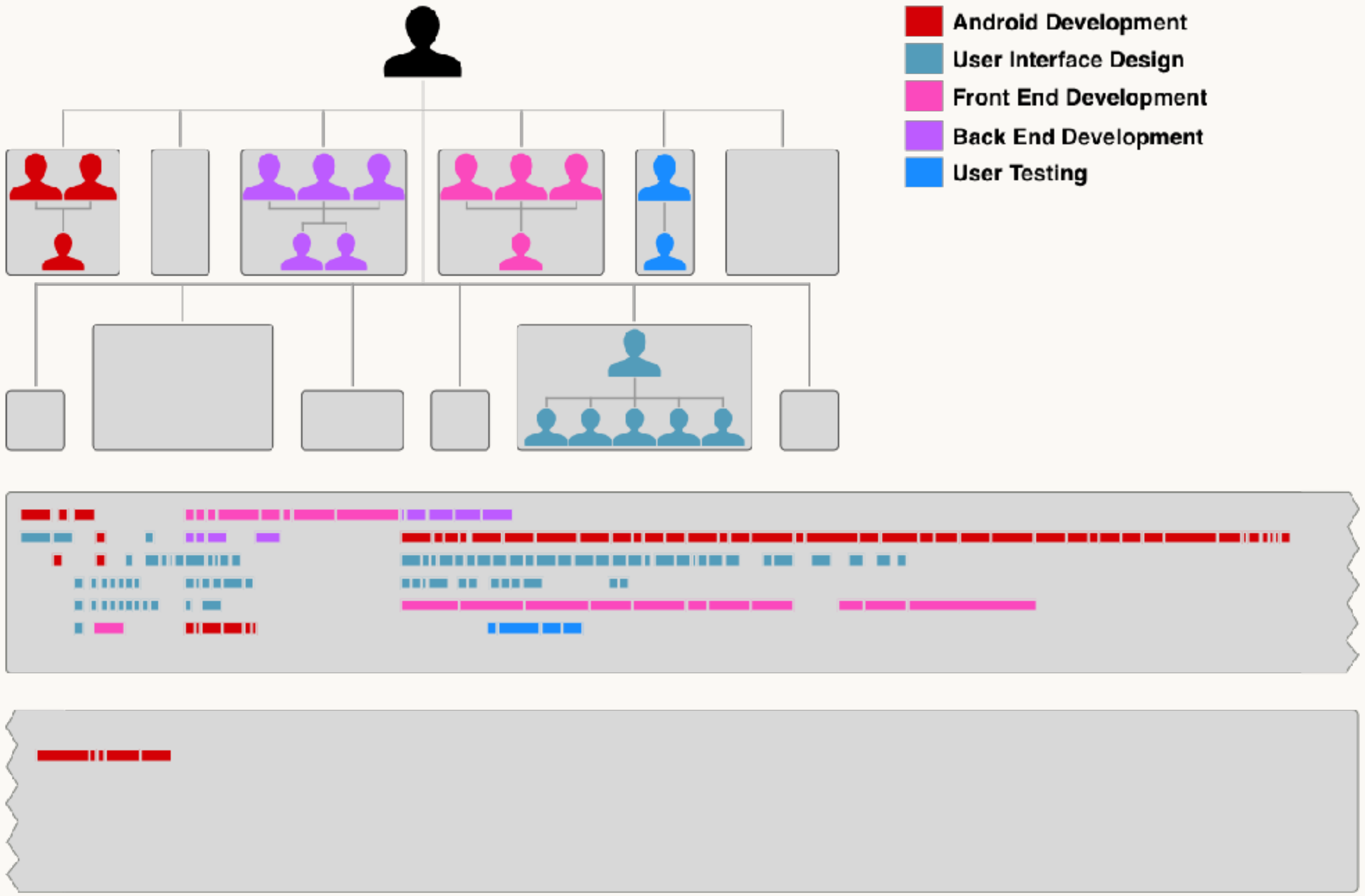
Login screen enables the user to enter their valid information to interact with the ap

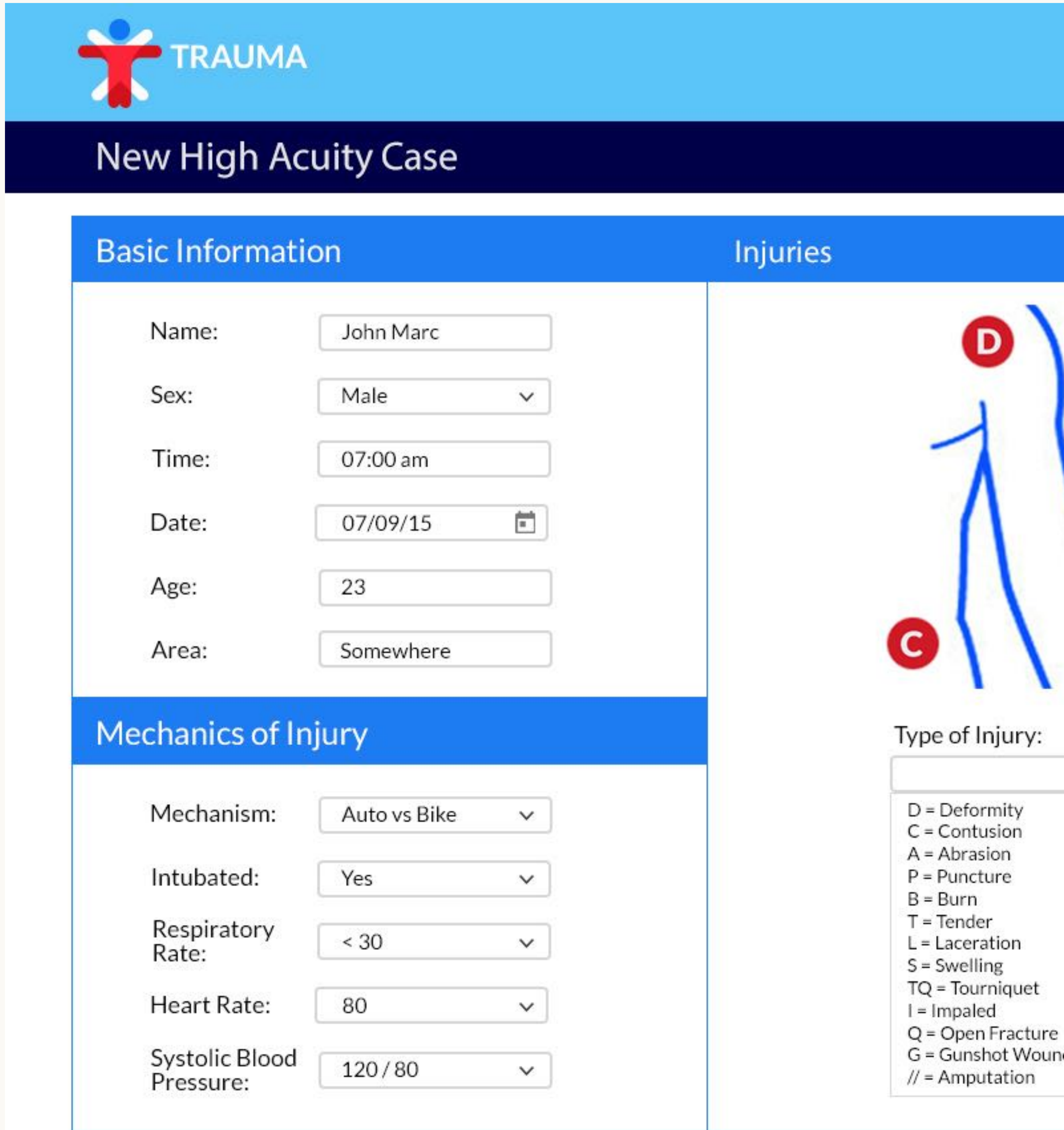
User needs to provide the following information in order to use the application. Field follows:

1. Username (A unique name created at the time of sign up)
2. Password (Valid password created at the time of sign up)
3. Remember Me Checkbox (If checked, it stores the user's login information ar again when the application is re-launch.)

TOP-DOWN RECONFIGURATION

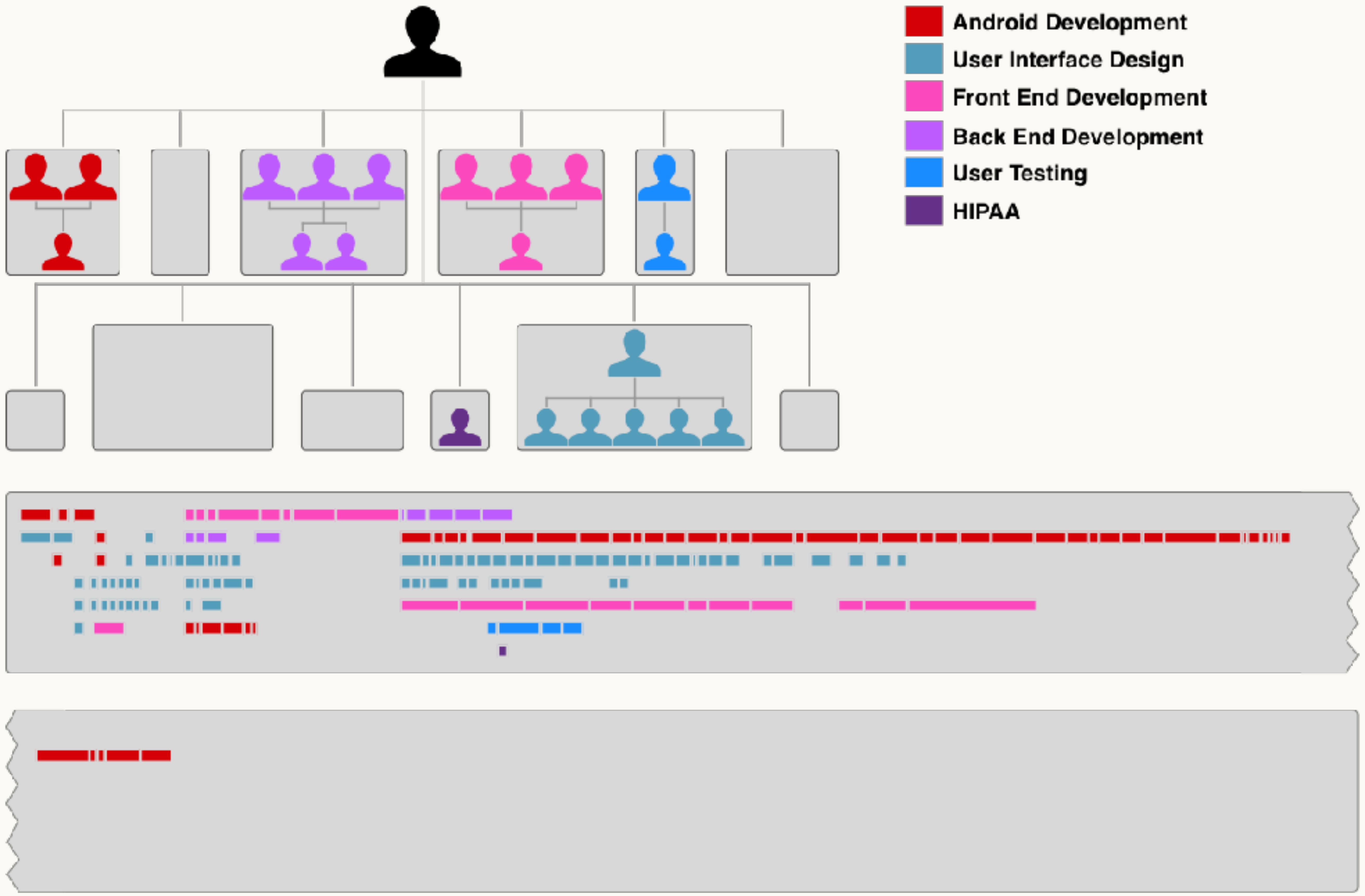
EMS TRAUMA REPORT

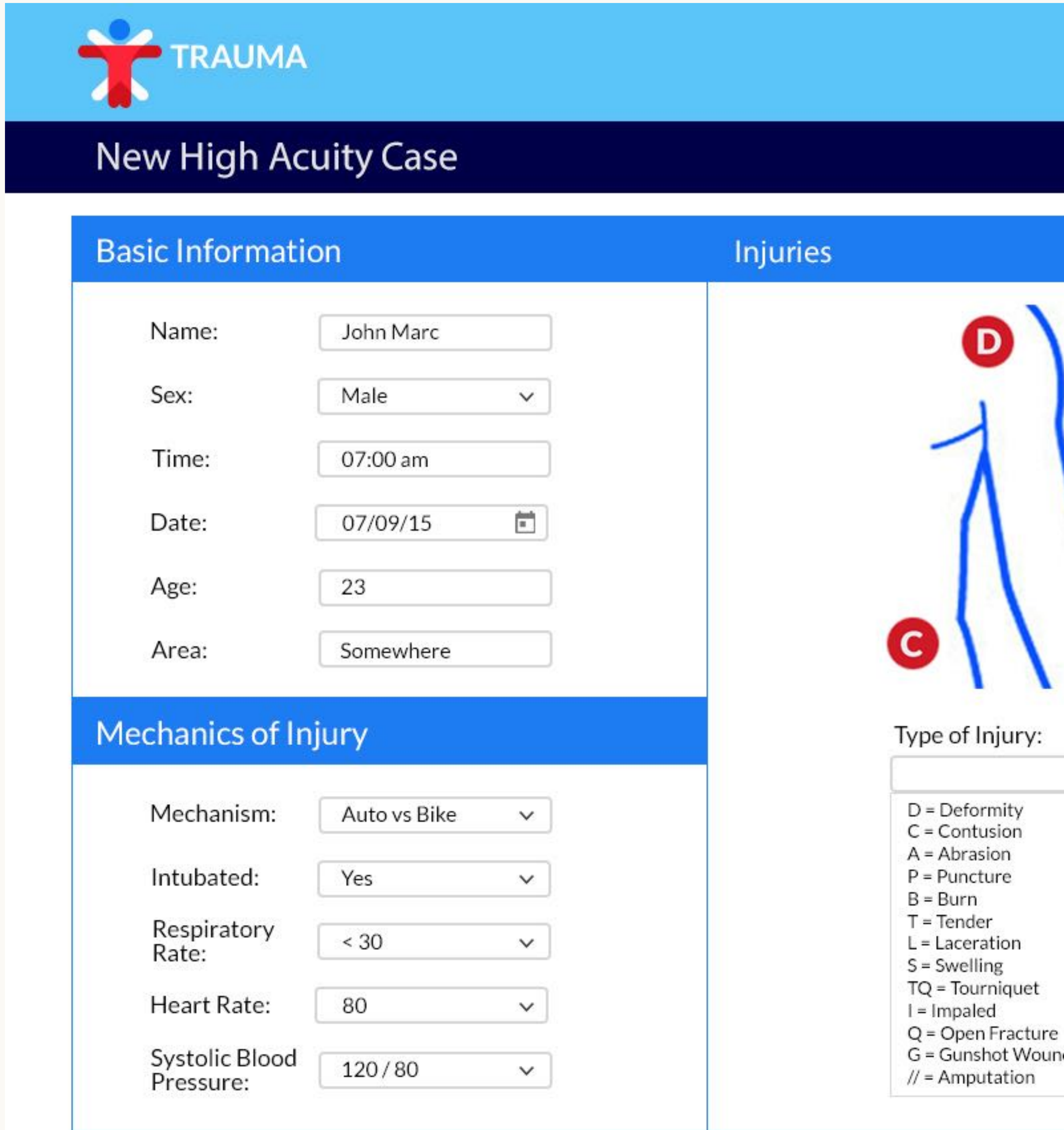




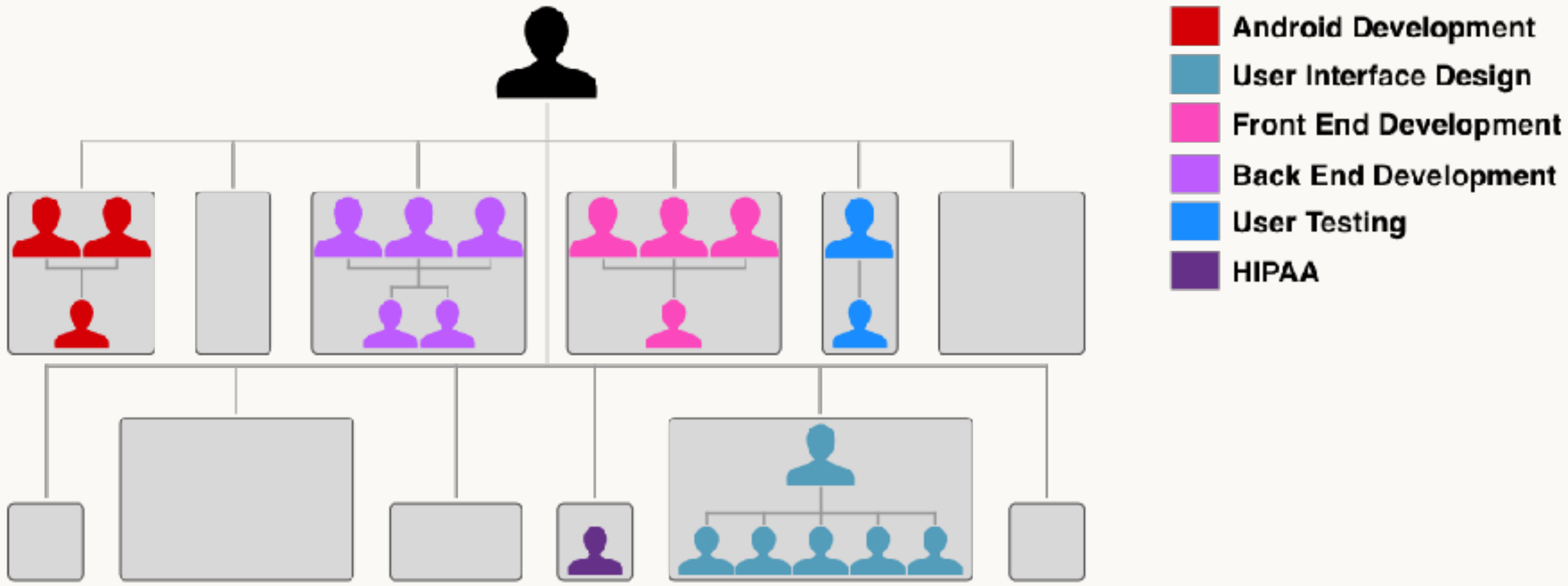
TOP-DOWN RECONFIGURATION

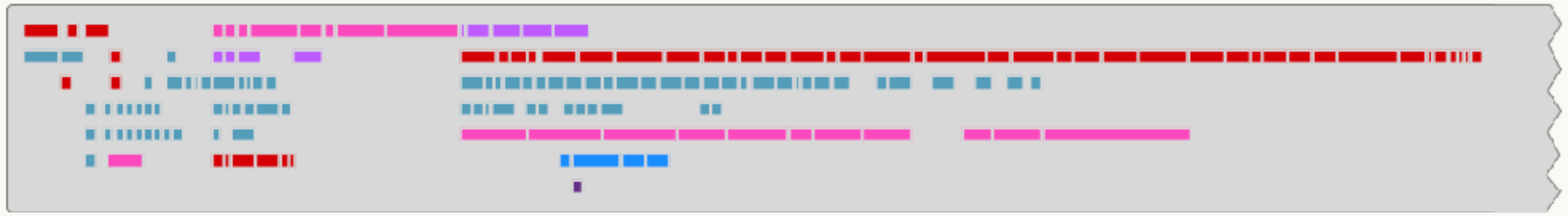
EMS TRAUMA REPORT





EMS TRAUMA REPORT



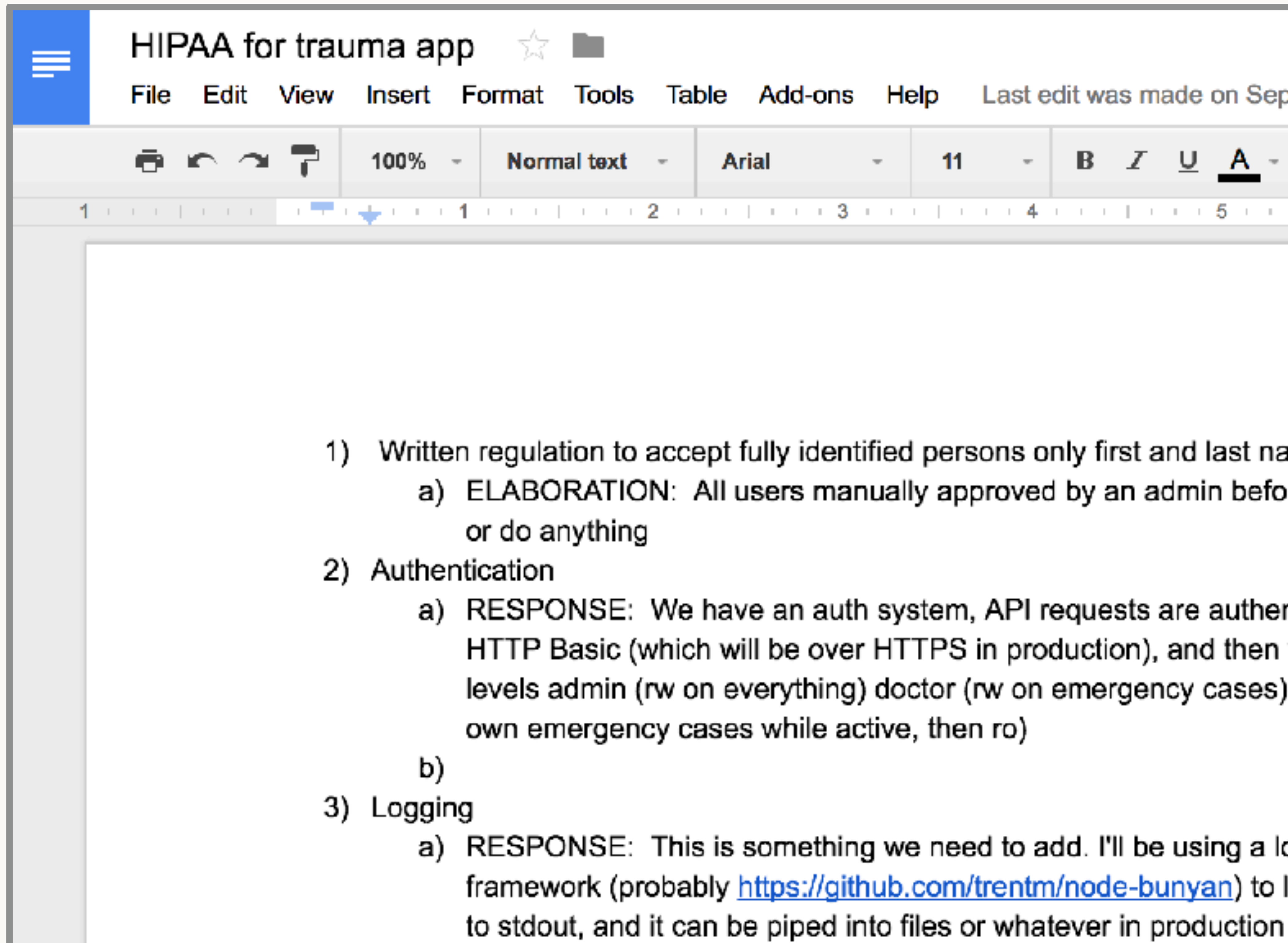


How do I become HIPAA compliant? (a checl

By Jason Wang / Published on October 30, 2013

A little housekeeping before we answer the question. This article is not a definitive list of what is req should assign a Privacy Officer to review each rule in its entirety. This article is intended to point yo

So you have determined that you are handling protected health information (PHI) and that you need next? What steps need to be taken in order to become HIPAA compliant?



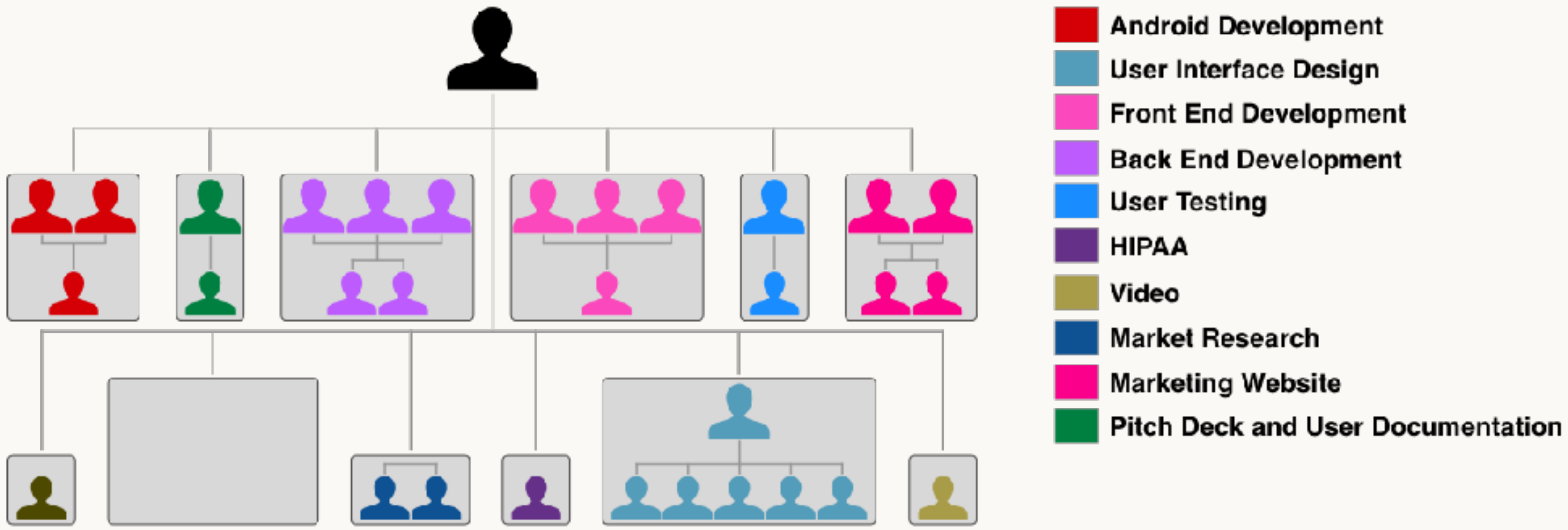
HIPAA for trauma app

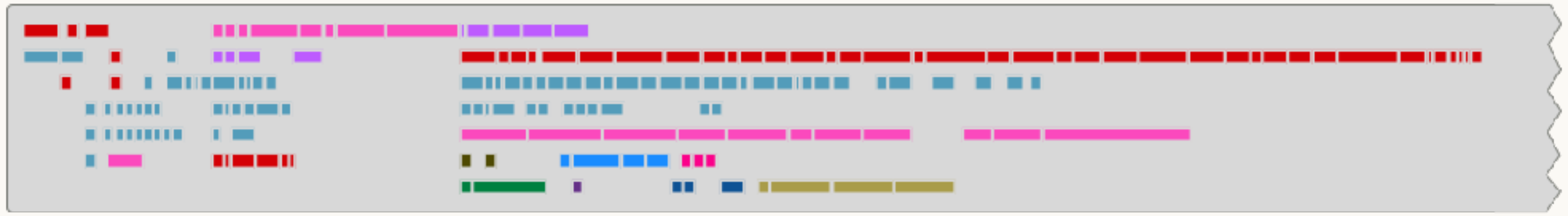
File Edit View Insert Format Tools Table Add-ons Help Last edit was made on Sep

100% Normal text Arial 11 B I U A

1) Written regulation to accept fully identified persons only first and last na
   a) ELABORATION: All users manually approved by an admin befo or do anything
2) Authentication
   a) RESPONSE: We have an auth system, API requests are auther HTTP Basic (which will be over HTTPS in production), and then levels admin (rw on everything) doctor (rw on emergency cases) own emergency cases while active, then ro)
   b)
3) Logging
   a) RESPONSE: This is something we need to add. I'll be using a lo framework (probably https://github.com/trentm/node-bunyan) to l to stdout, and it can be piped into files or whatever in production

EMS TRAUMA REPORT



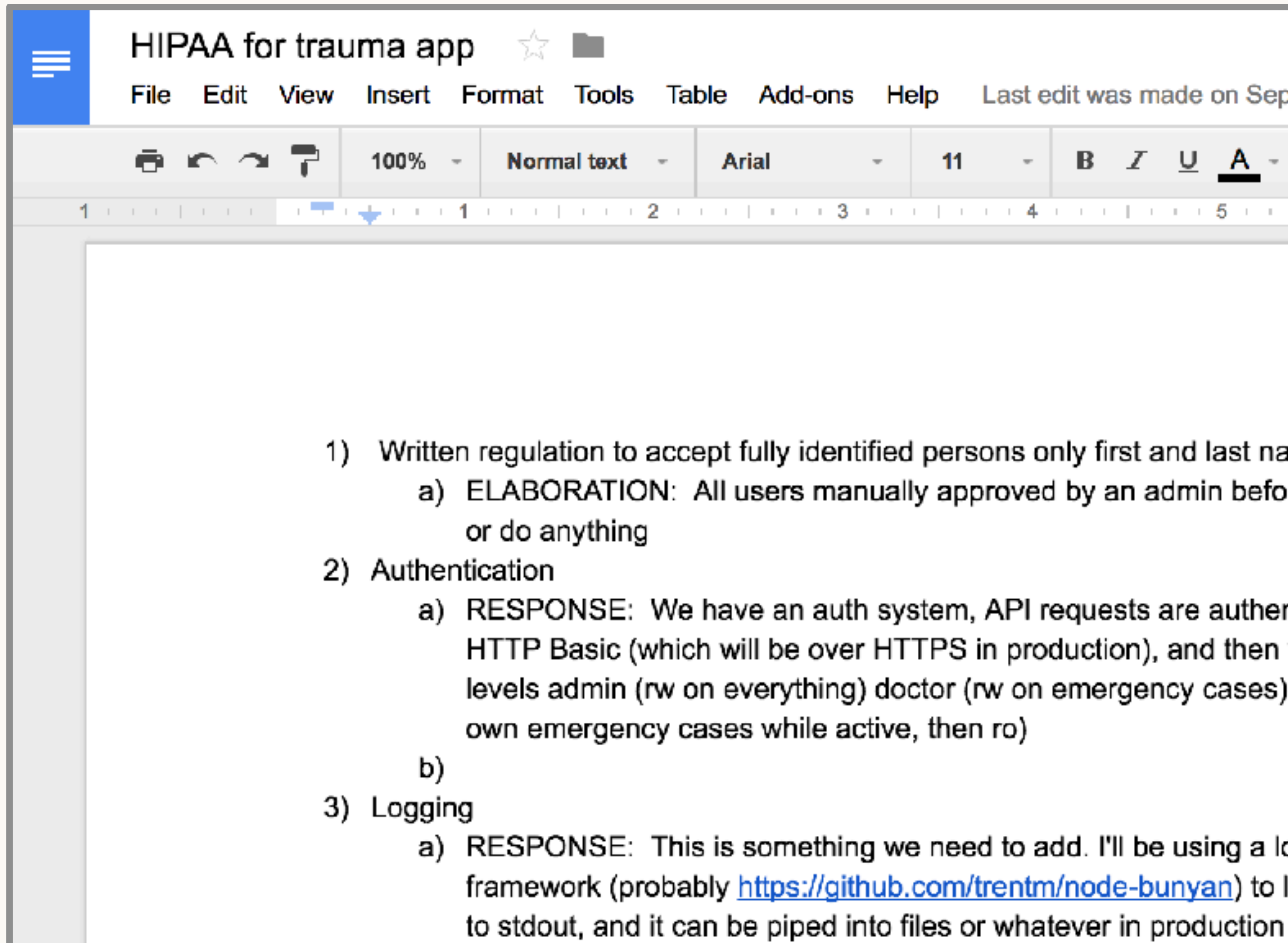


How do I become HIPAA compliant? (a chec

By Jason Wang / Published on October 30, 2013

A little housekeeping before we answer the question. This article is not a definitive list of what is re
should assign a Privacy Officer to review each rule in its entirety. This article is intended to point yo

So you have determined that you are handling protected health information (PHI) and that you nee
next? What steps need to be taken in order to become HIPAA compliant?



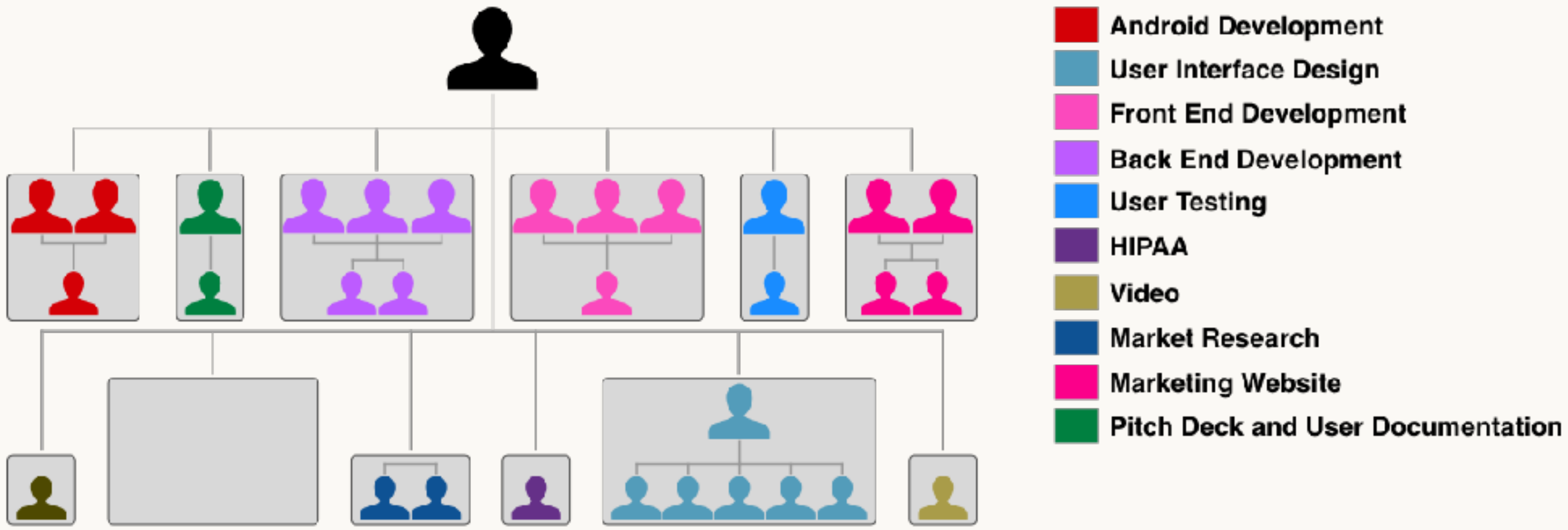
HIPAA for trauma app

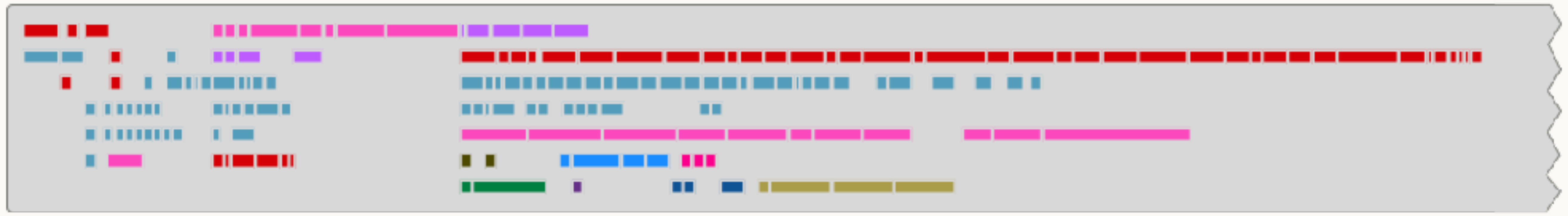
File Edit View Insert Format Tools Table Add-ons Help Last edit was made on Sep

100% Normal text Arial 11 B I U A

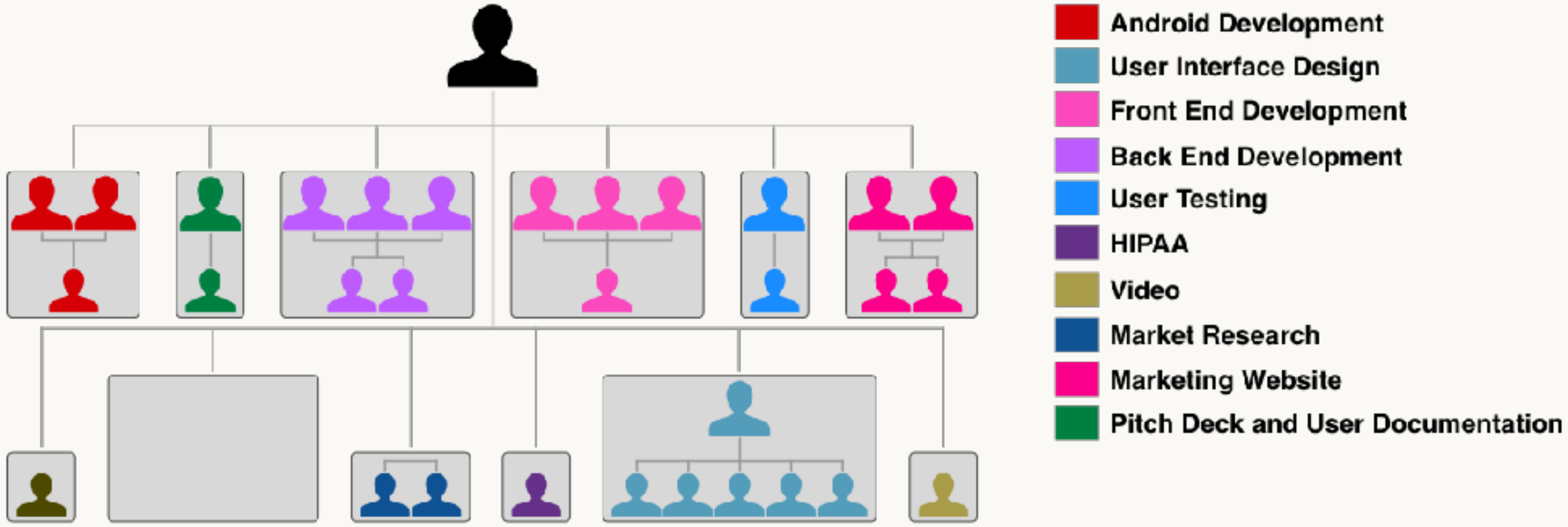
1) Written regulation to accept fully identified persons only first and last na
   a) ELABORATION: All users manually approved by an admin befo
      or do anything
2) Authentication
   a) RESPONSE: We have an auth system, API requests are auther
      HTTP Basic (which will be over HTTPS in production), and then
      levels admin (rw on everything) doctor (rw on emergency cases)
      own emergency cases while active, then ro)
   b)
3) Logging
   a) RESPONSE: This is something we need to add. I'll be using a lo
      framework (probably https://github.com/trentm/node-bunyan) to l
      to stdout, and it can be piped into files or whatever in production

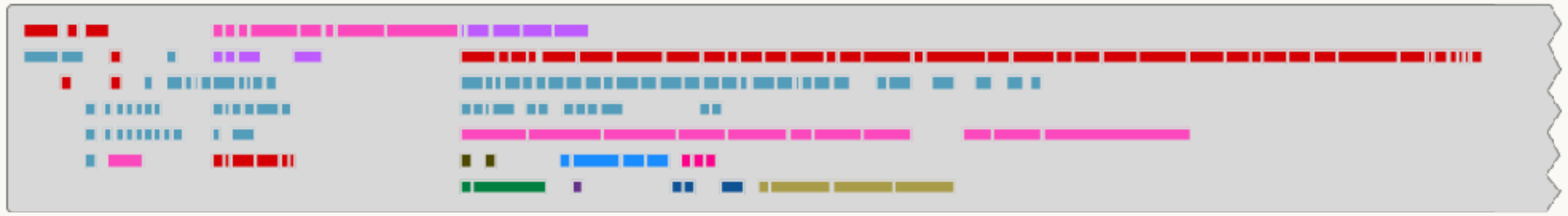
EMS TRAUMA REPORT

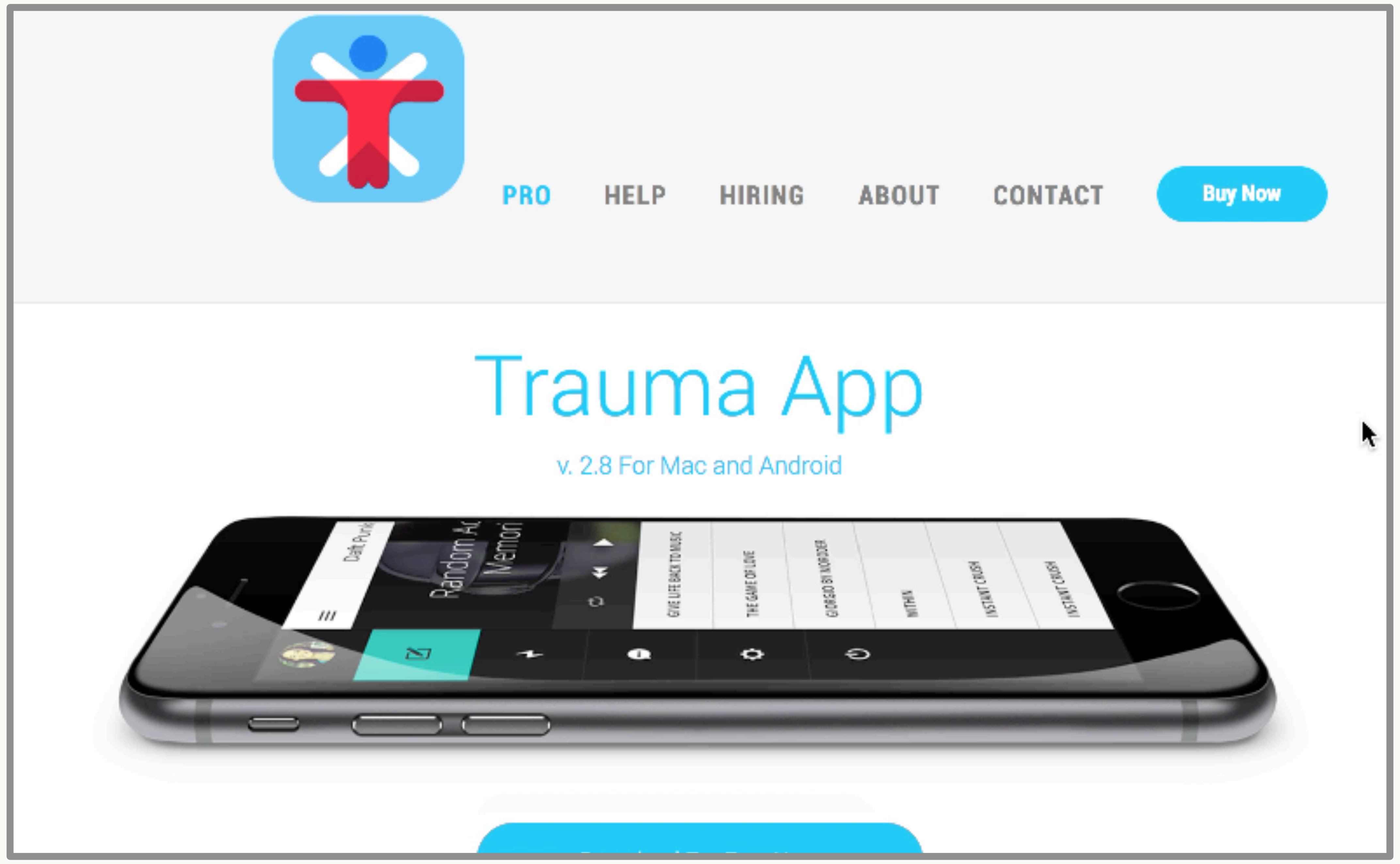




EMS TRAUMA REPORT

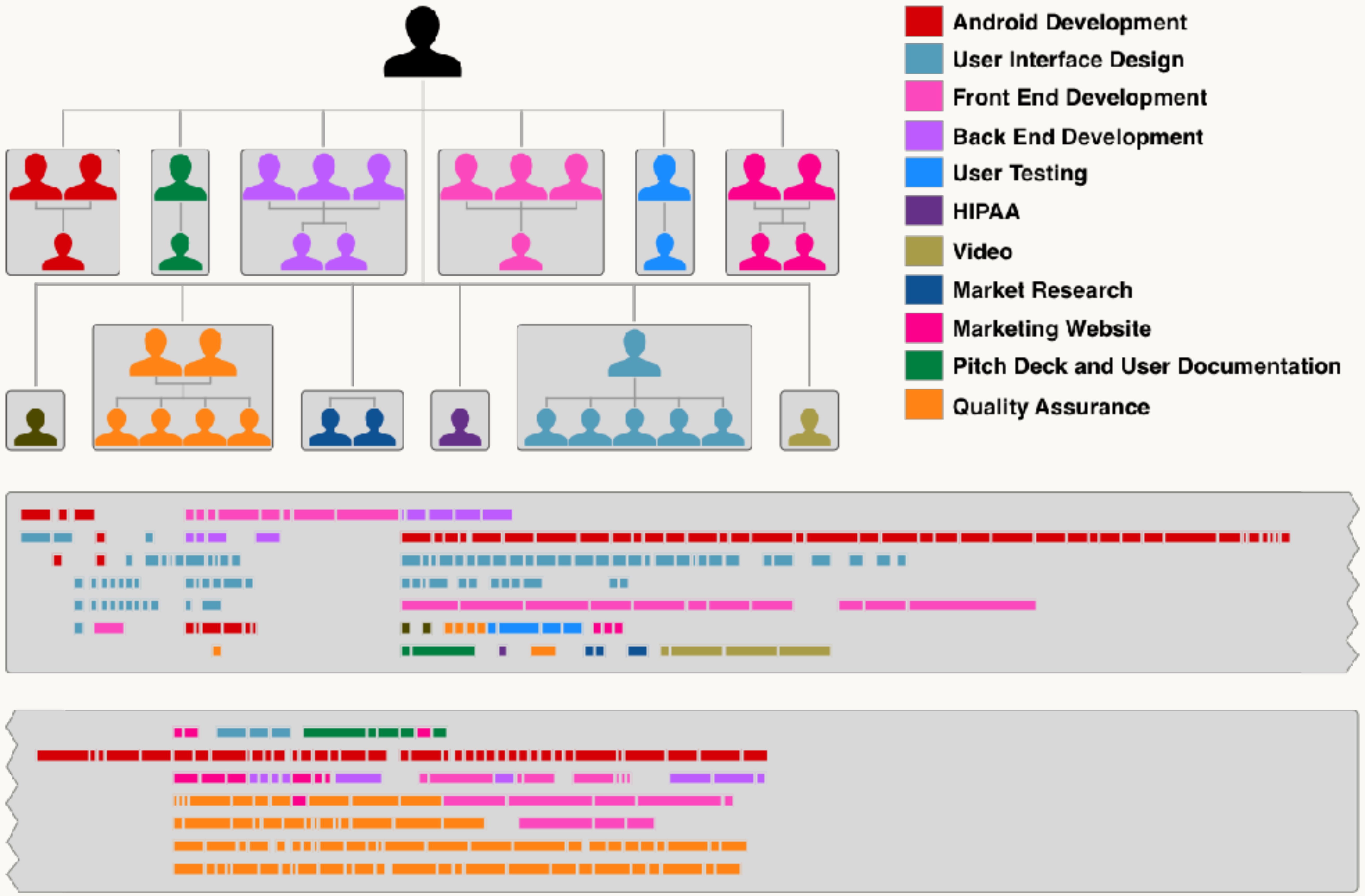




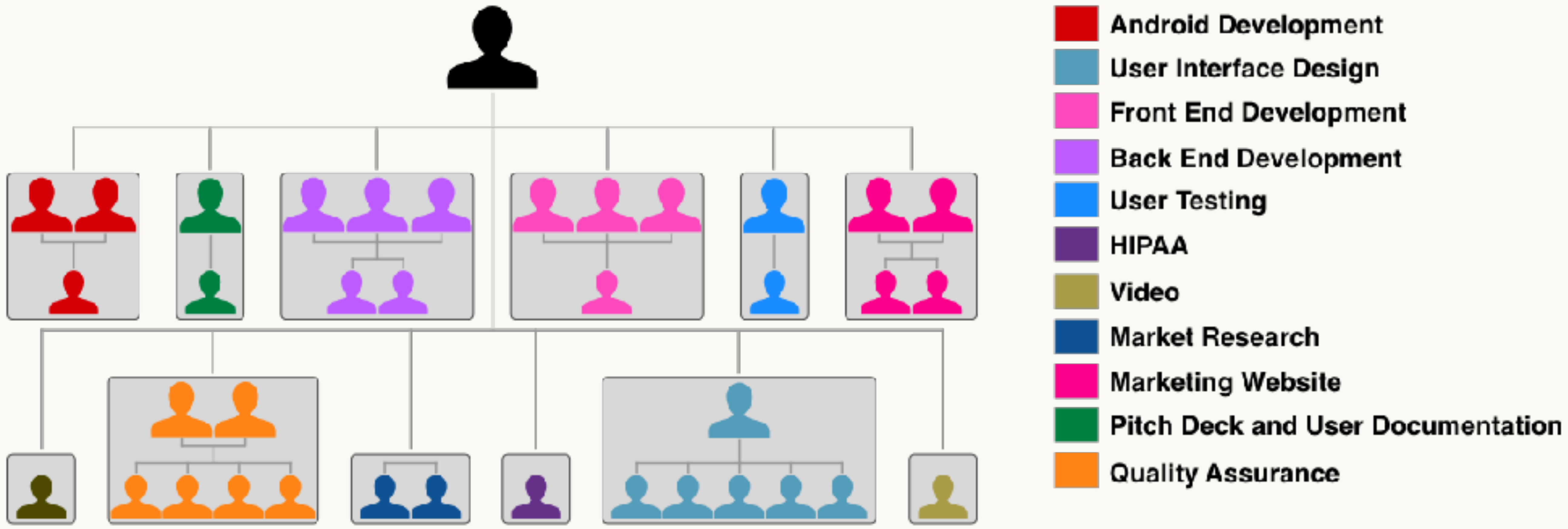


BOTTOM-UP RECONFIGURATION

EMS TRAUMA REPORT



EMS TRAUMA REPORT



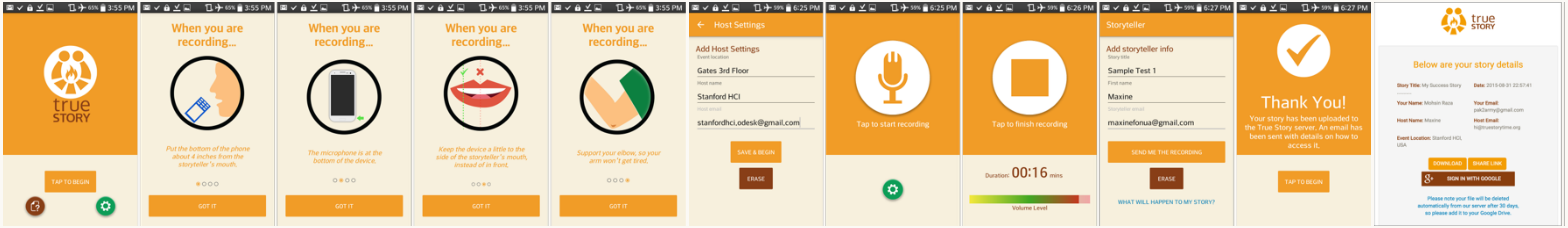
EMS TRAUMA REPORT

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

ROLE HIRES IN <14 MINUTES

	EMS Report	True Story	Enterprise Portal	All Projects
Median time (mm:ss)	13:40	12:40	15:13	13:40

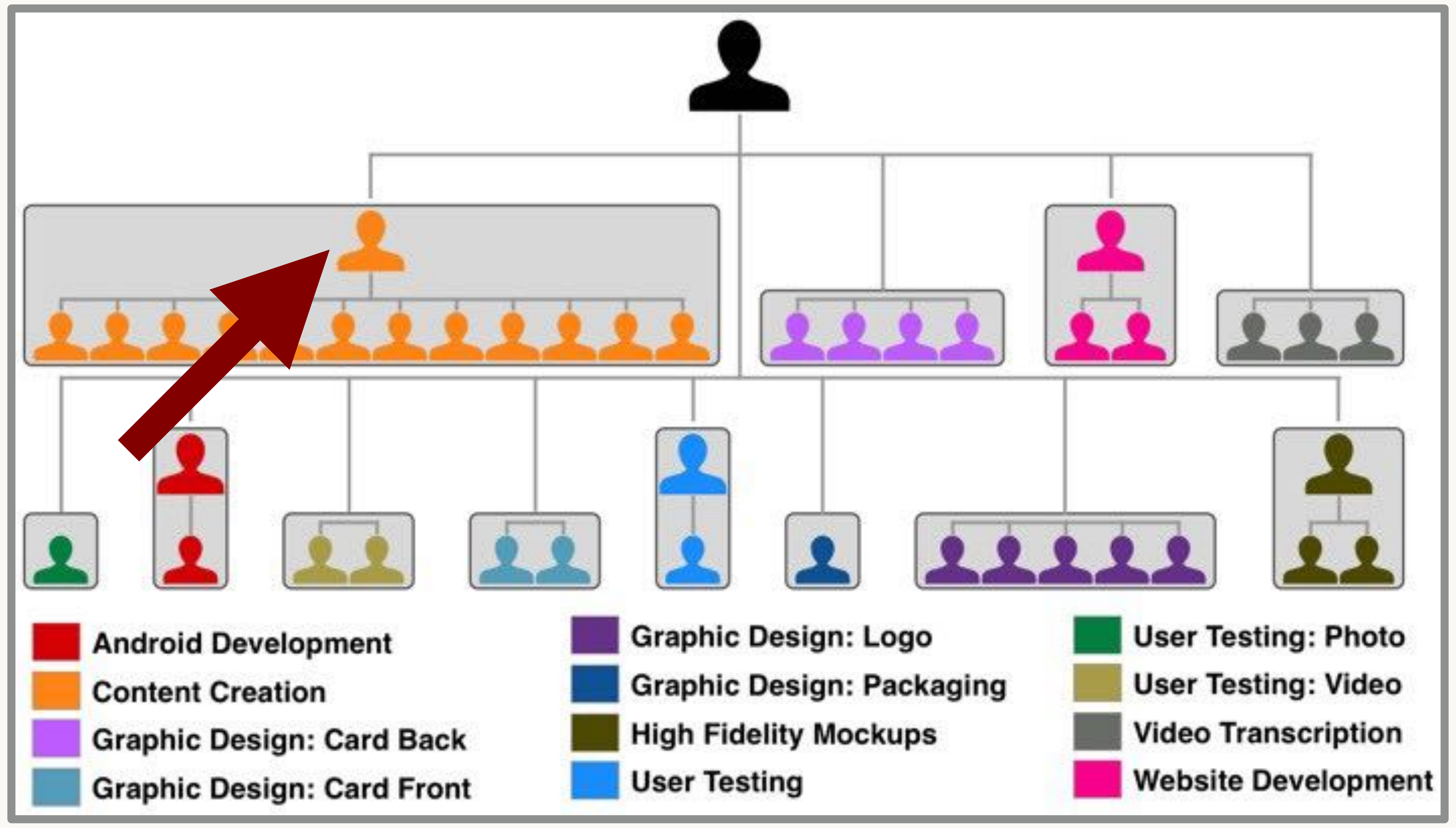
20 manual hires in a median 889 minutes (~15 hours)

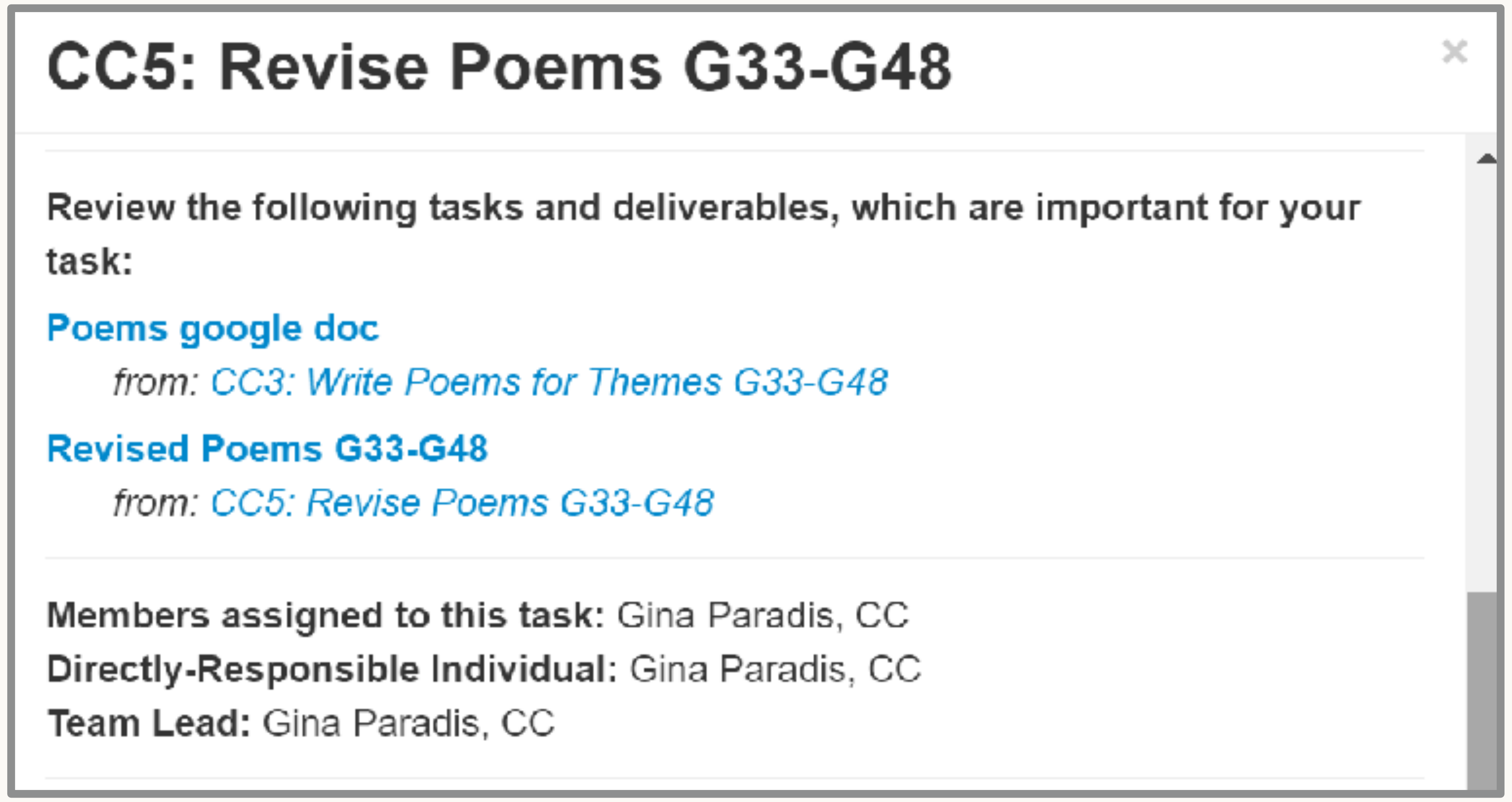
RECONFIGURED ORG STRUCTURES

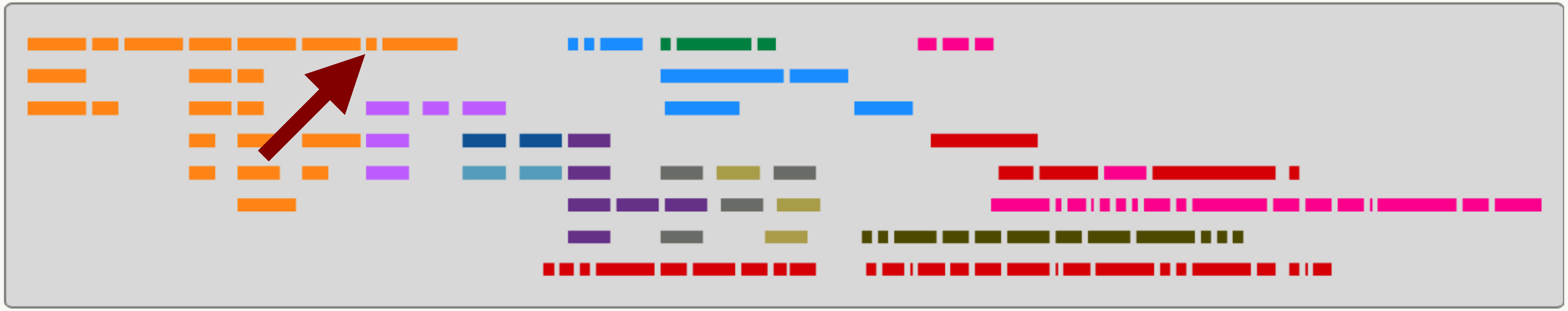
	EMS Report	True Story	Enterprise Portal	All Projects
# of pull requests	335	113	118	566
Mean pull requests per day	7.3	2.8	3.4	4.5

CHIEF POETRY OFFICER

RECONFIGURED ROLES

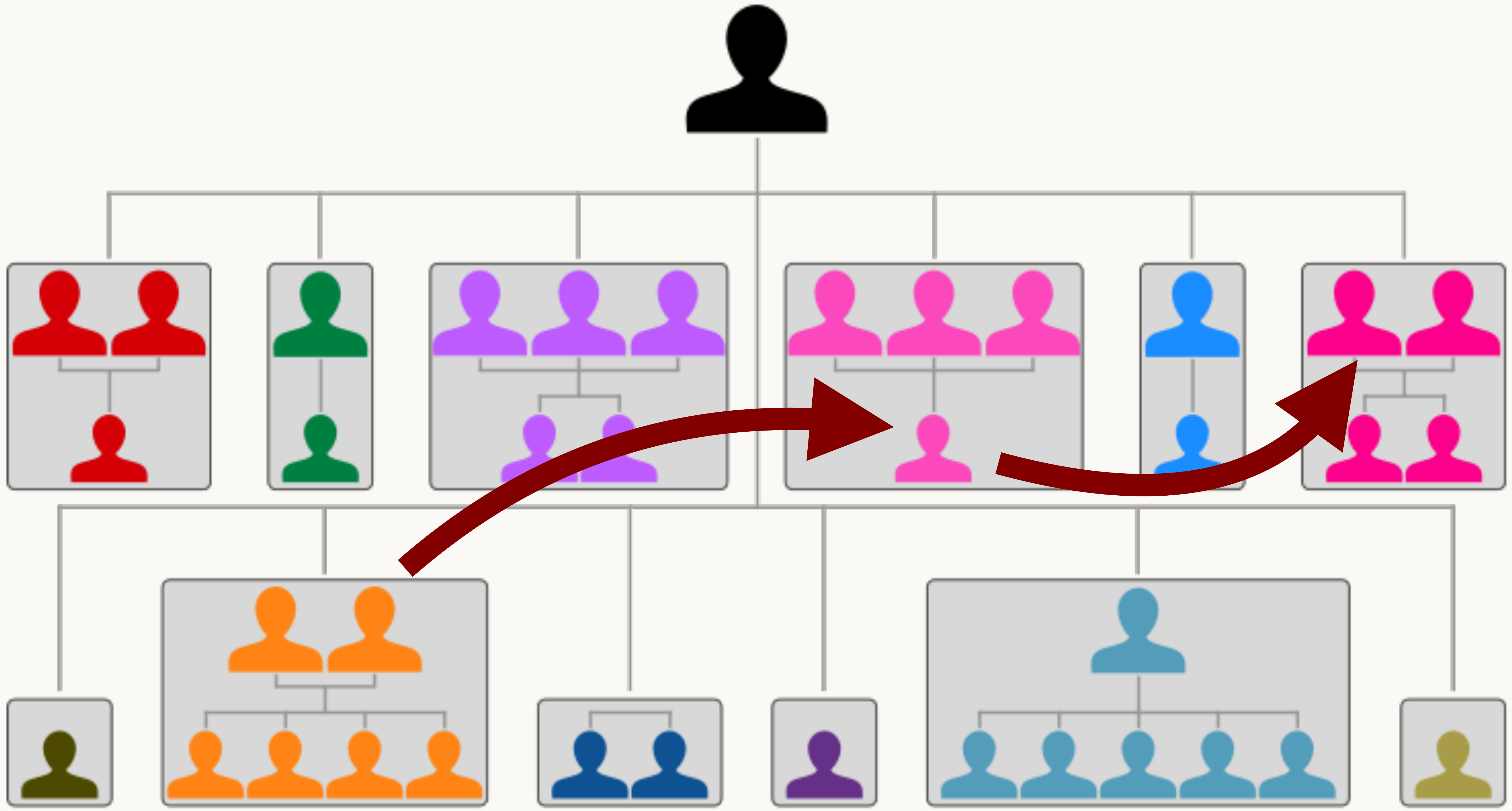






ACCRETION AND ROTATION

RECONFIGURED ROLES



While organizations could conceivably hire anew for each role, in practice they also accreted members and rotated those members into new roles

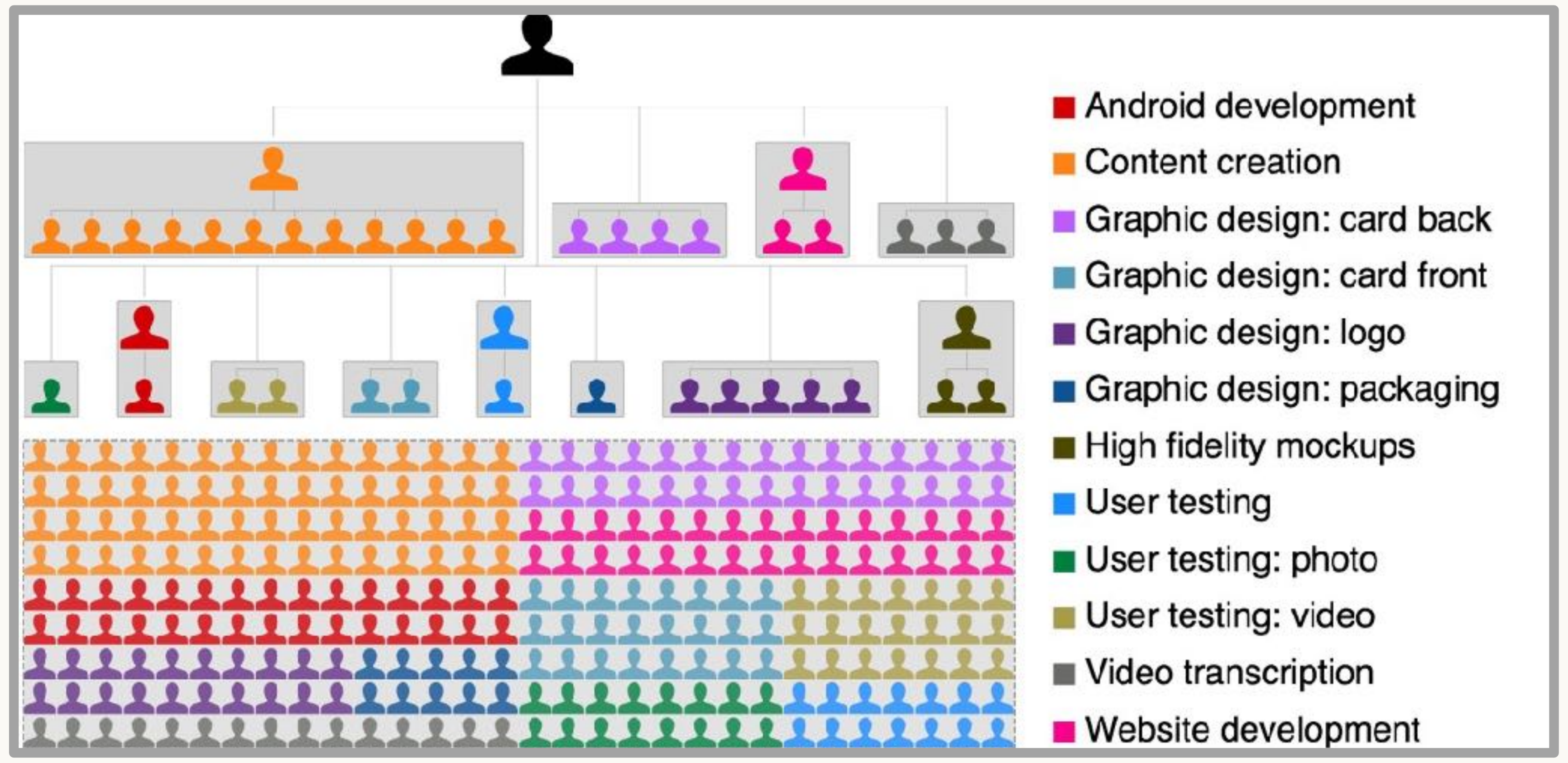
REFLECTIONS

When computation is a mediating layer for work, we can design it to help guide and support peoples' goals

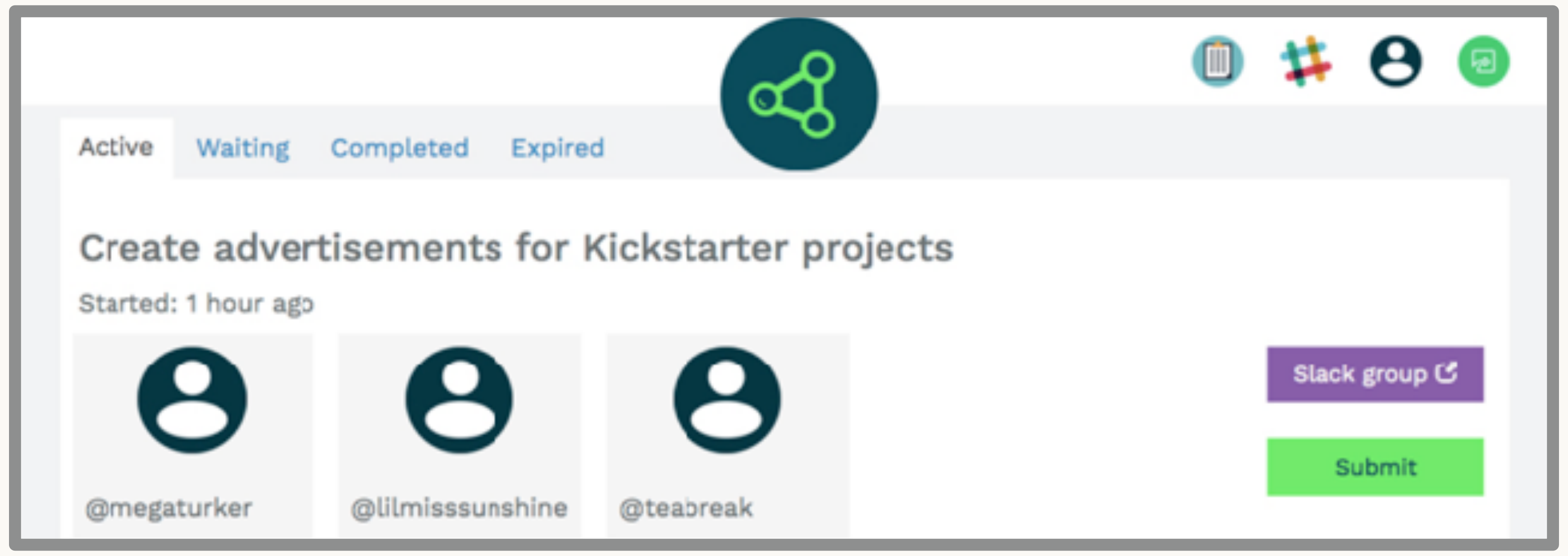
Current and future contributions:

- Use data, theory, and experimentation to encourage more effective organizational practices
- Change the transaction costs core to the Theory of the Firm
- Extract crowdsourcing from the microtasking swamp

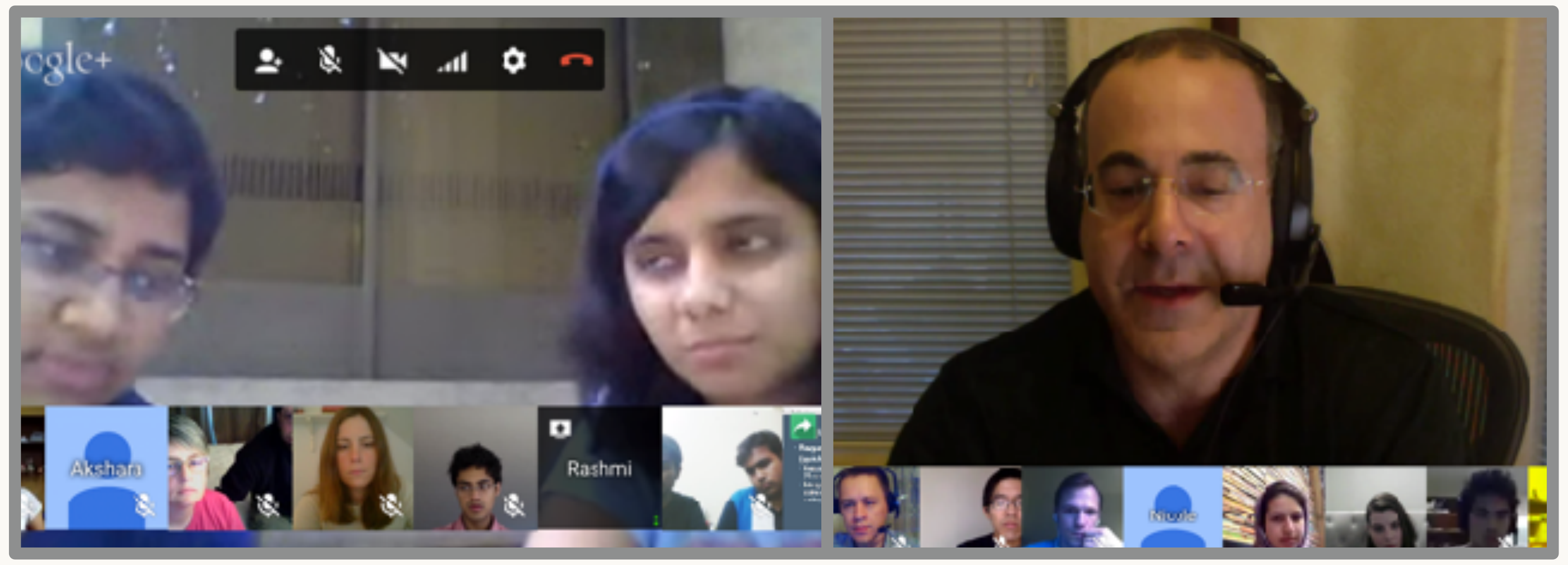
1) Enable crowd collectives to achieve complex and open-ended goals



2) Recruit effective collaborators despite unpredictable availability



3) Crowdsource research itself, providing global access to upward mobility



Huddler

Salehi, Valentine, Bernstein. CSCW 2017.

FLASH ORGANIZATIONS SACRIFICE FAMILIARITY TO GAIN SPEED

Revisiting asset specificity…

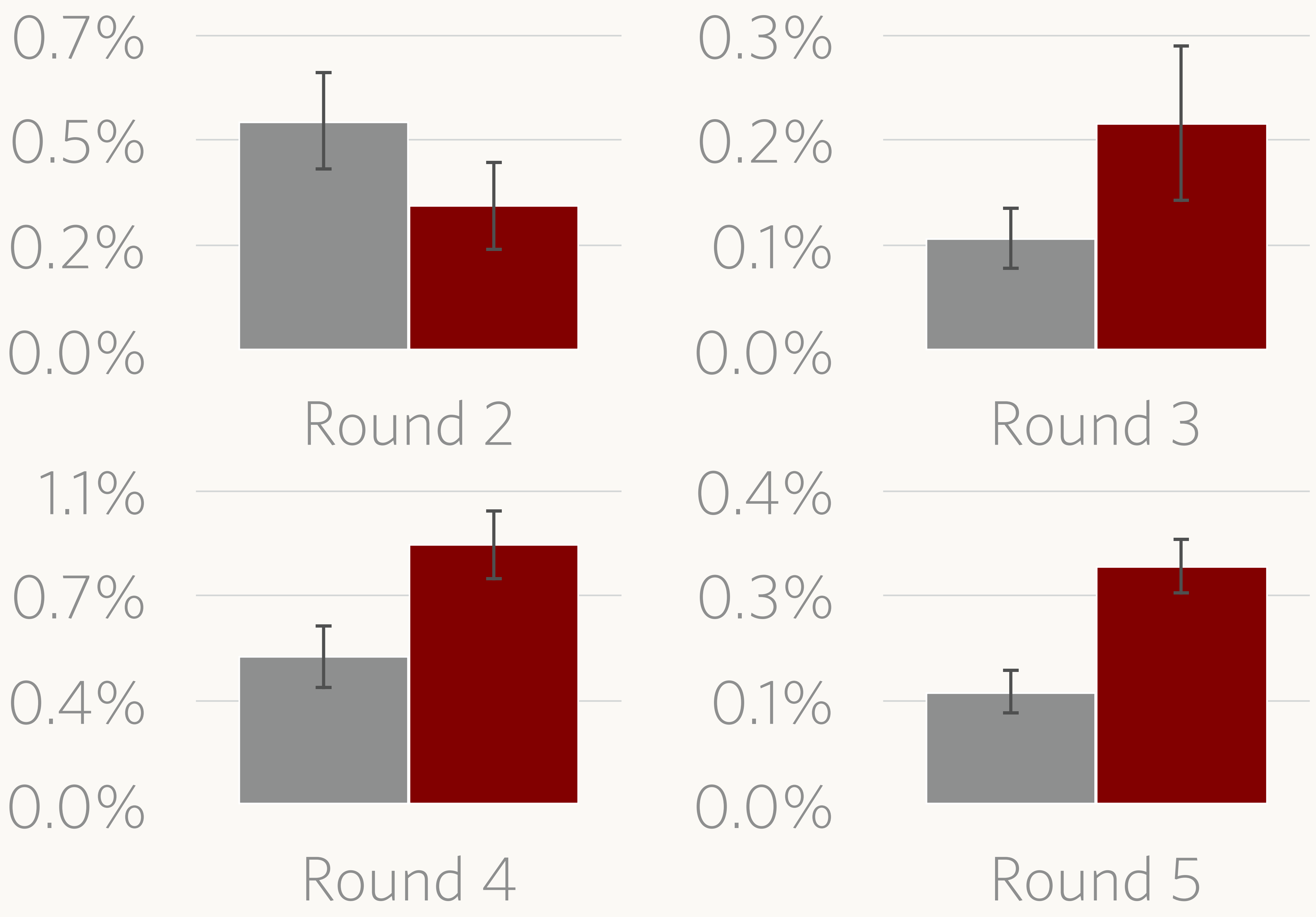
Team-based coordination requires that team members become familiar with each other by working together over time [Huckman, Staats, and Upton 2009; Reagans, Argote, and Brooks 2005]

FAMILIAR TEAMS PERFORM BETTER

Teams from AMT authored creative ads for Kickstarter projects

Manipulation: Team membership was random in each round, or kept familiar by maintaining the team across rounds

Measure: AdWords CTR



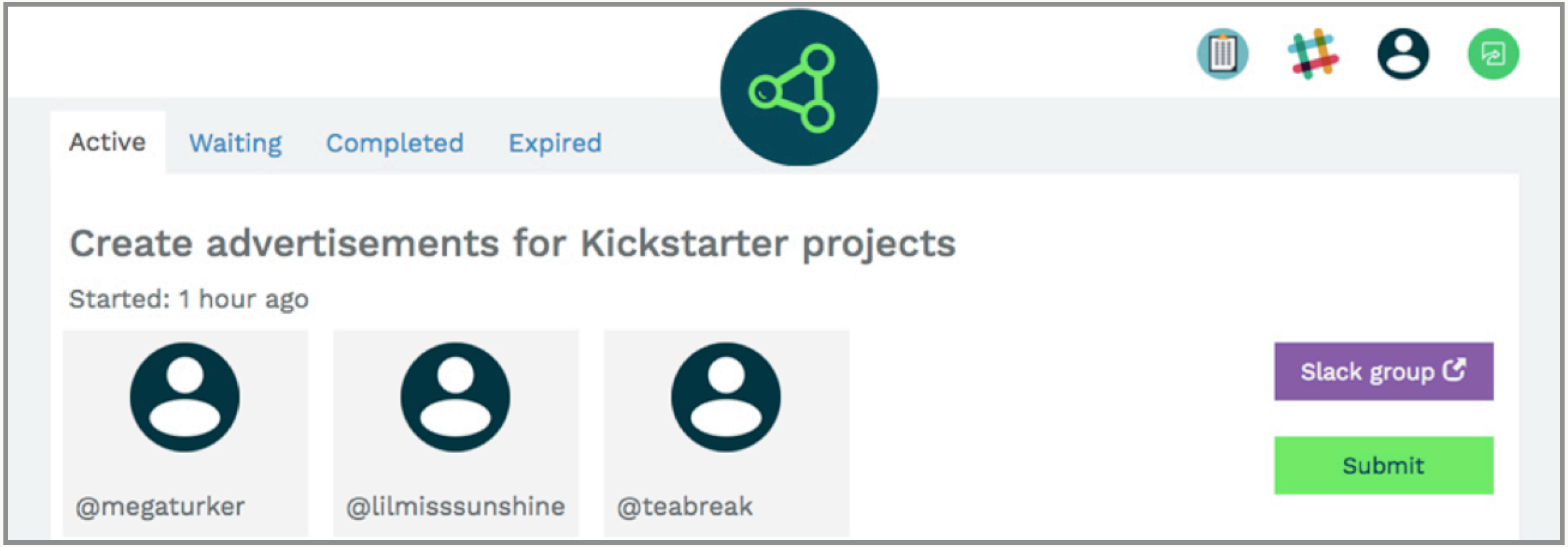
By Task 5, familiar teams had twice the CTR of random teams: $t(31)=3.37$, $p<.01$, $d=1.2$

CROWDSOURCING IS AT ODDS WITH FAMILIARITY

On-demand crowdsourcing would seem to make building familiarity infeasible

Goal: a system that enables assembly of familiar crowd teams, even under unpredictable availability and strict time constraints

HUDDLER


Active
Waiting
Completed
Expired
Create advertisements for Kickstarter projects
Started: 1 hour ago
@megaturker
@lilmisssunshine
@teabreak
Slack group
Submit

Given a time constraint t and a current set of team members, find a schedule of people $p_{1 \dots n}$ to invite and wait times $t_{1 \dots n}$ to wait for each person to respond

Maximize the expected familiarity of the resulting team, given invitees' probability of being available

$$\underset{p_i, t_i,\ i=1,\dots,n}{\text{maximize}} \quad \sum_{i=1}^{n} \left[\text{availability}(p_i, t_i) \cdot \sum_{p_m \in \text{team}} \text{familiarity}(p_i, p_m) \right]$$

$$\text{subject to} \quad \sum_{i=1}^{n} t_i \leq t$$

Planning who to ask and how long to wait before moving on is a combinatorial problem with an exponential number of alternatives

Dynamic program: recursively compute sub-solutions

$$\sum_{i=1}^{n}\left[\text{availability}(p_i, t_i) \cdot \sum_{p_m \in \text{team}} \text{familiarity}(p_i, p_m)\right]$$

$$= E([p_1, \ldots, p_n], t)$$

$$= \max_{0 \leq t_1 \leq t} E([p_2, \ldots, p_n], t - t_1) + E(p_1, t_1)$$

EVALUATION

N=280 teams assembled from Amazon Mechanical Turk over two weeks to perform Kickstarter advertisement design tasks, randomized across condition

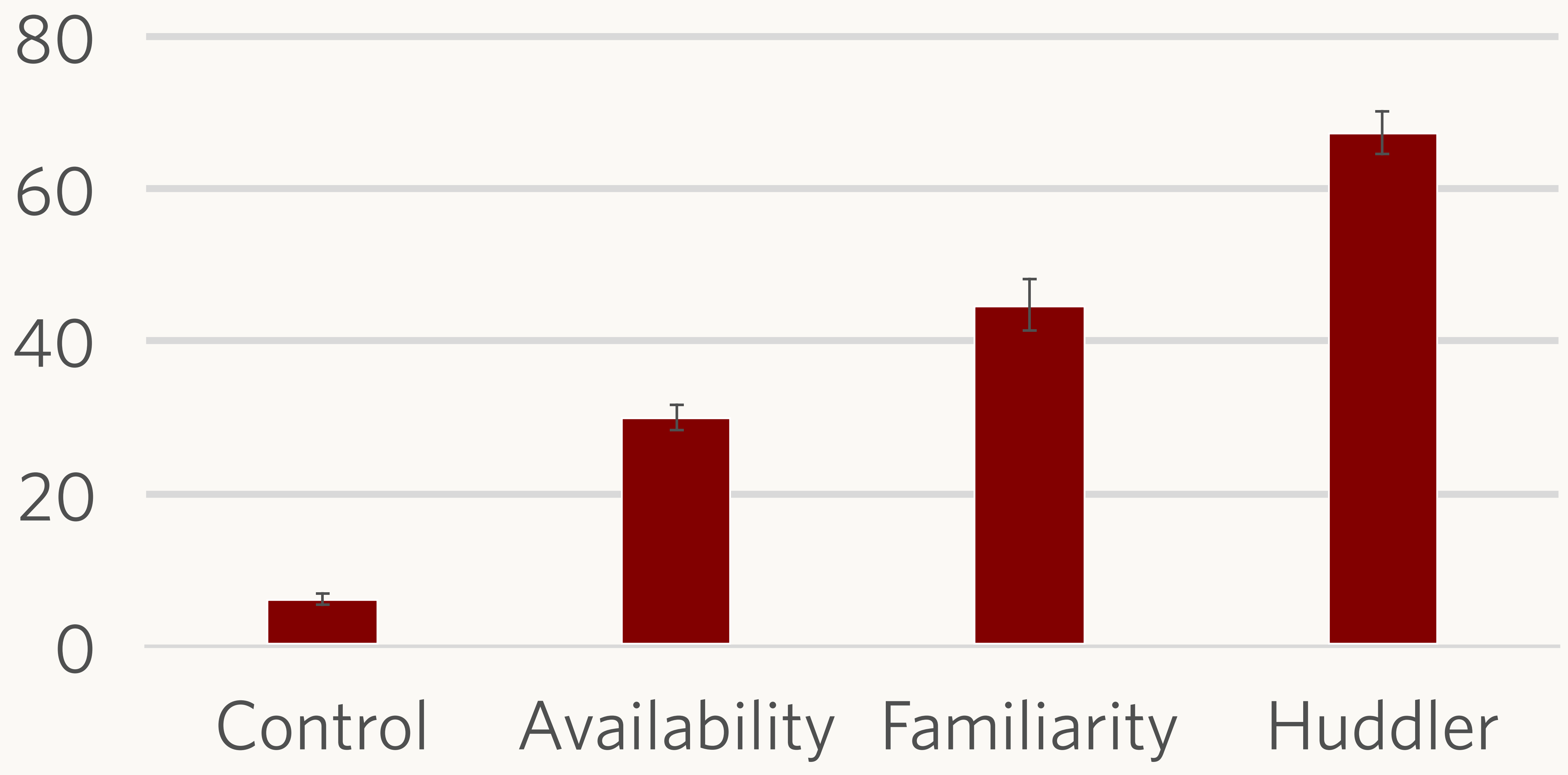
Measured: pairwise tasks between team members, minutes to form team

Control	Availability only
Familiarity only	Huddler

RESULTS

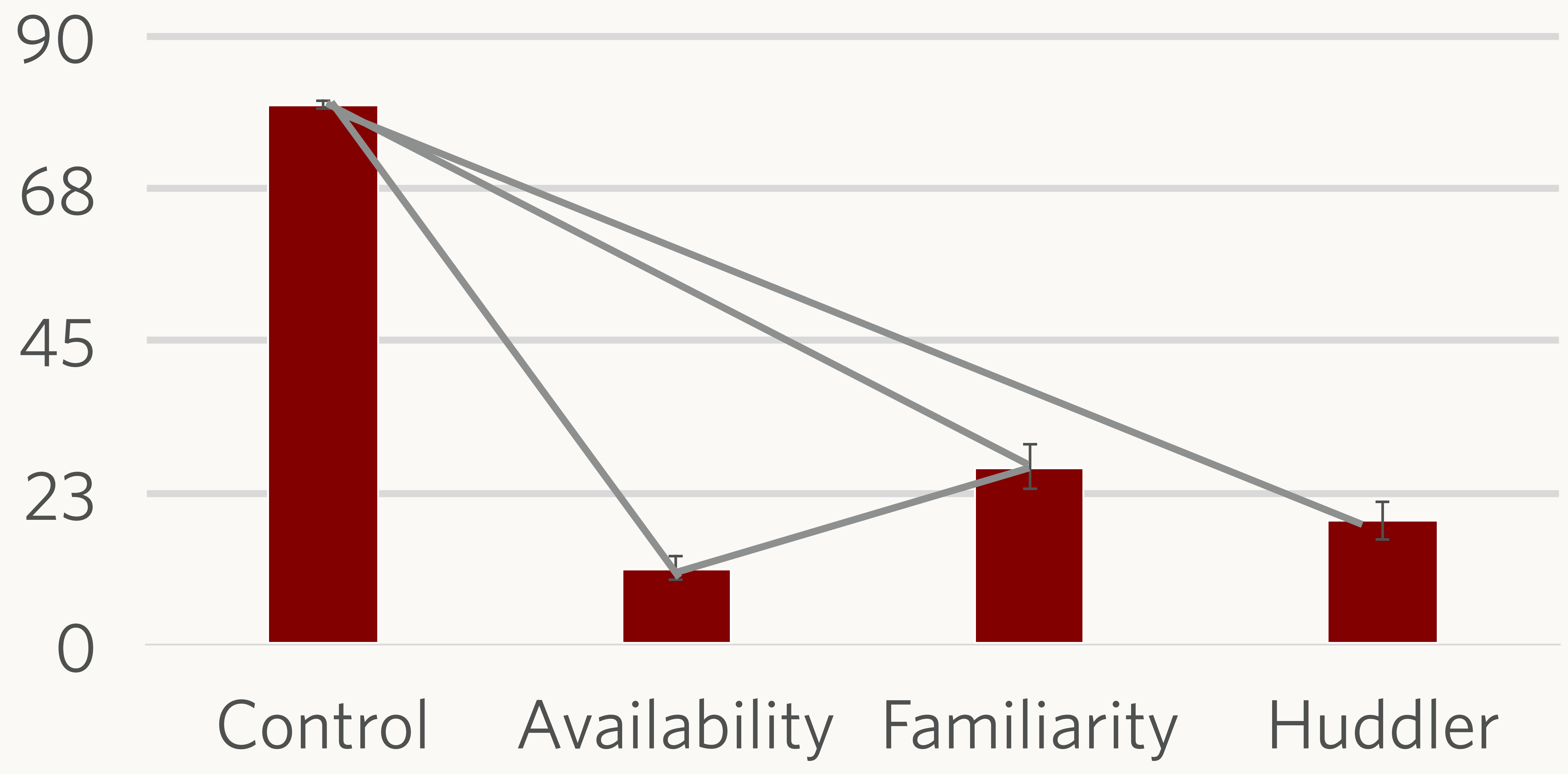
Huddler convened highly familiar teams nearly as quickly as when only trying to optimize for speed

Pairwise tasks between members at end of study



Two-way ANOVA: significant main effects (p<.01), no significant interaction

Minutes to form teams



Two-way ANOVA: significant main effects (p<.01), significant interaction (p<.01)

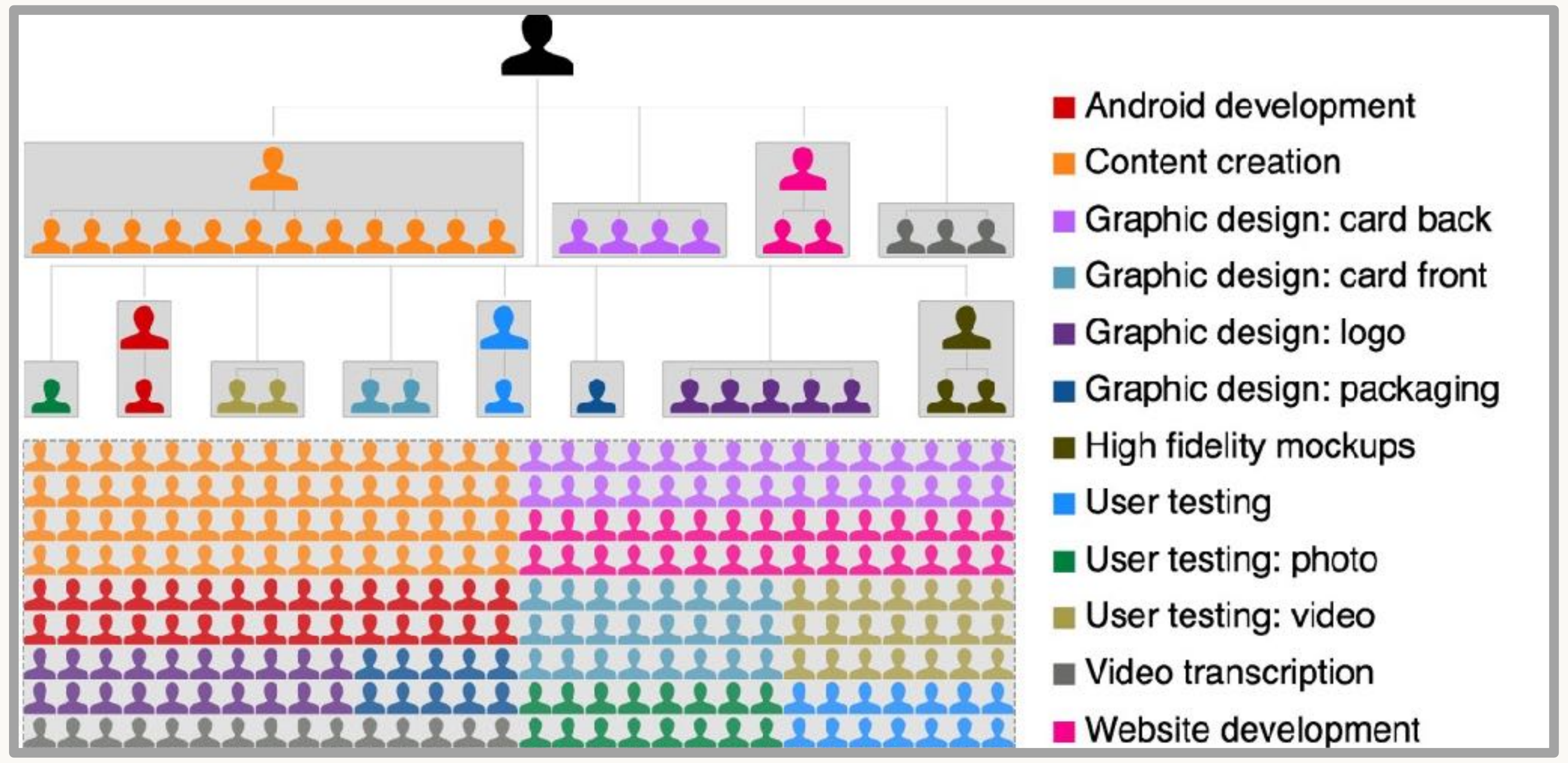
REFLECTIONS

Crowdsourcing does not need to give up the social fabric of teamwork in order to achieve rapid, responsive efforts

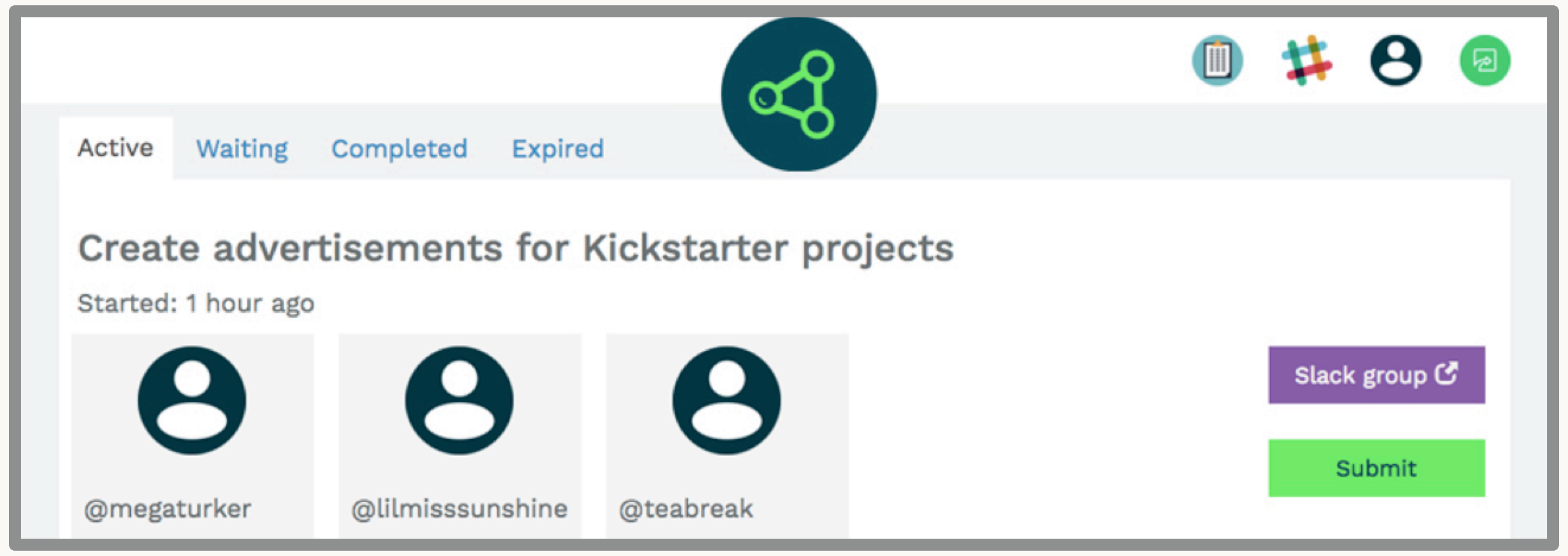
Current and future contributions:

- Adapting Huddler's value function to span other goals: personality balancing, diverse expertise, predicted performance, satisfaction…
- Countering Huddler's risk aversion from availability — the system exploits early teams rather than build a deeper network

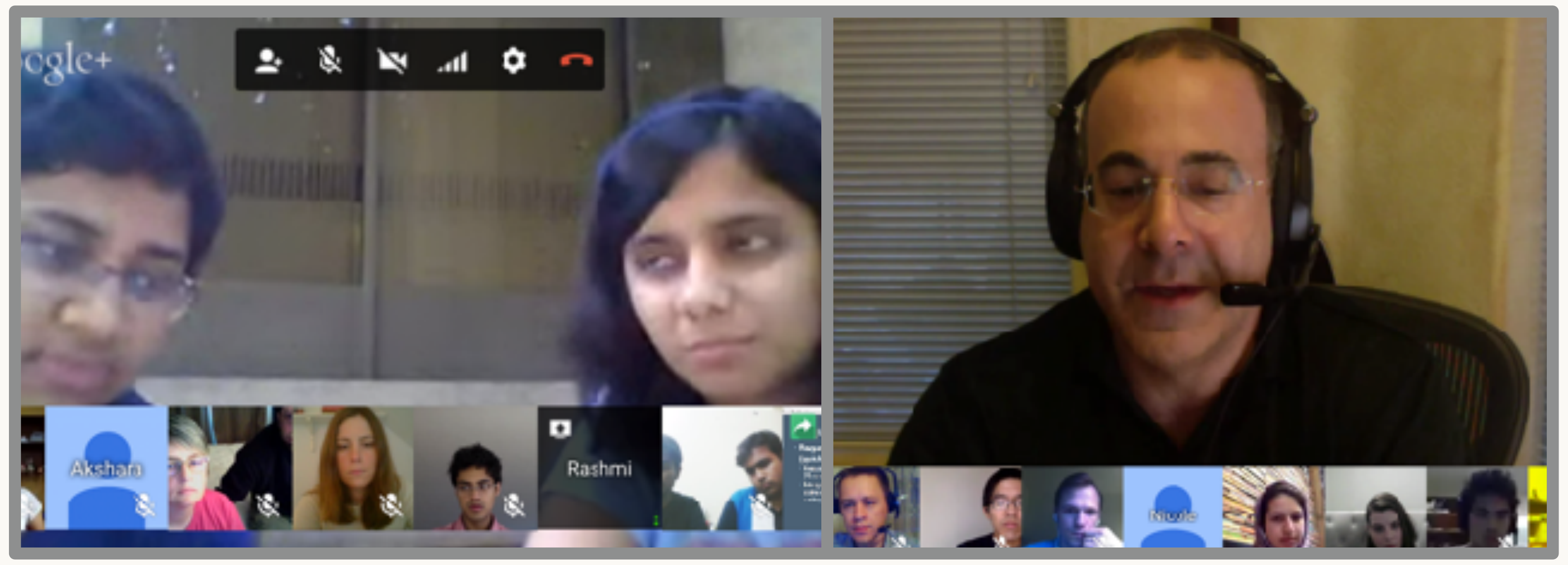
1) Enable crowd collectives to achieve complex and open-ended goals



2) Recruit effective collaborators despite unpredictable availability



3) Crowdsource research itself, providing global access to upward mobility



Crowd research

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

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

…but the vast majority of people cannot

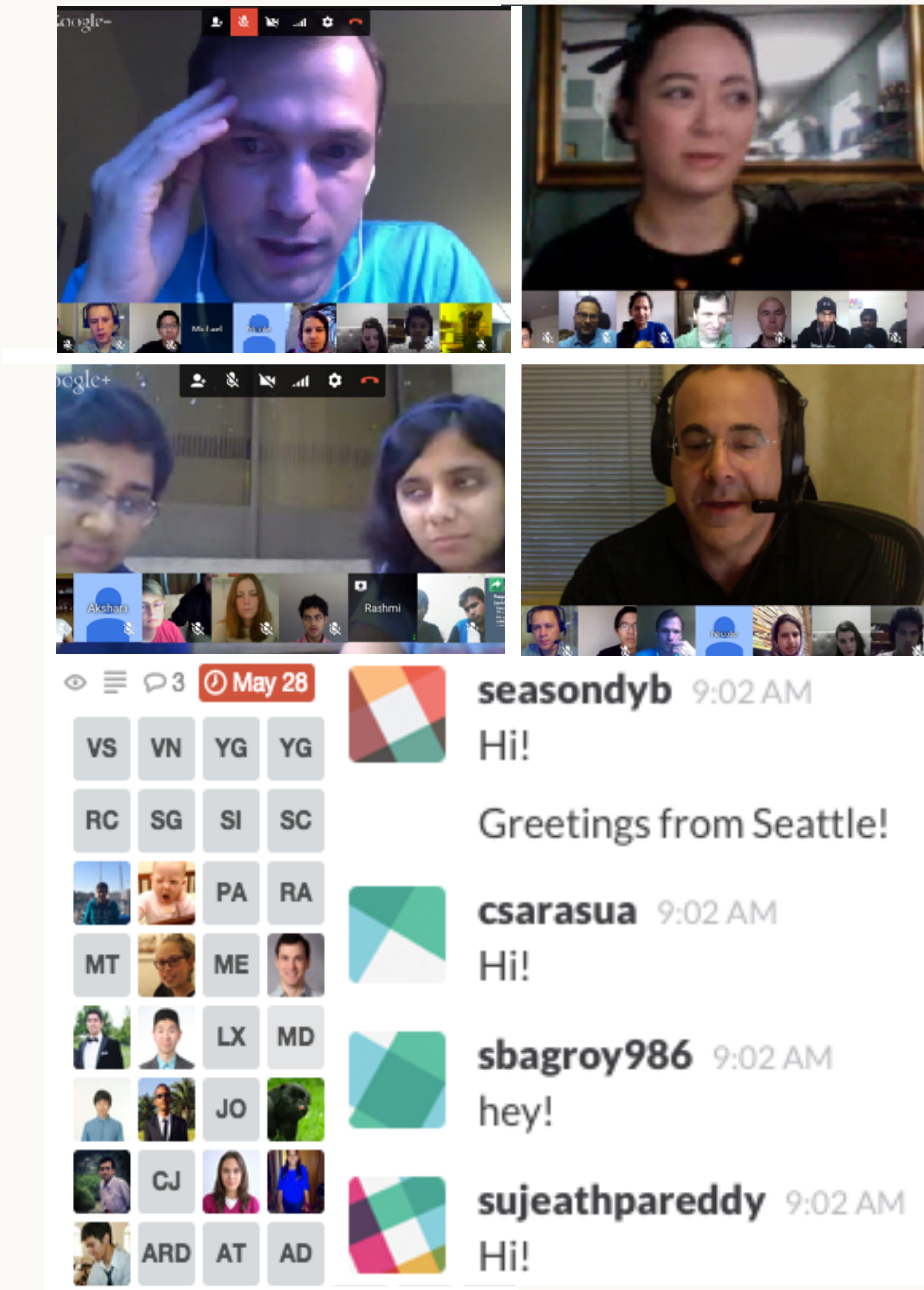
A research ecosystem that under-represents minorities and developing regions, and a literature that overlooks their perspectives

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

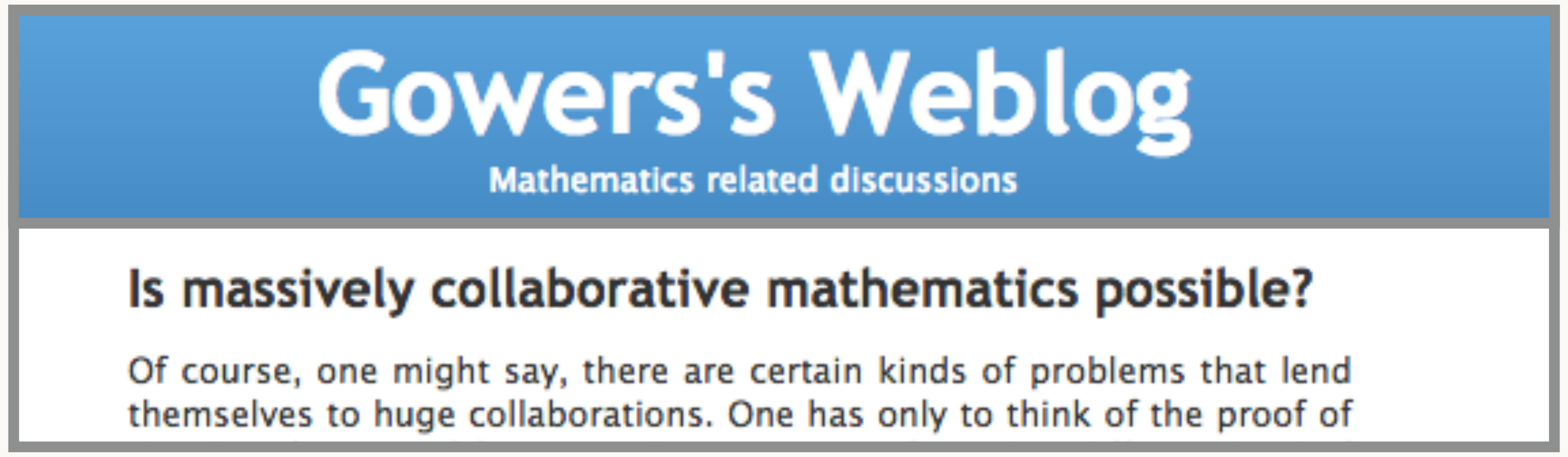
Give access to training and research experiences that can enable upward career and educational mobility

Convene hundreds or thousands of people on a single ambitious project

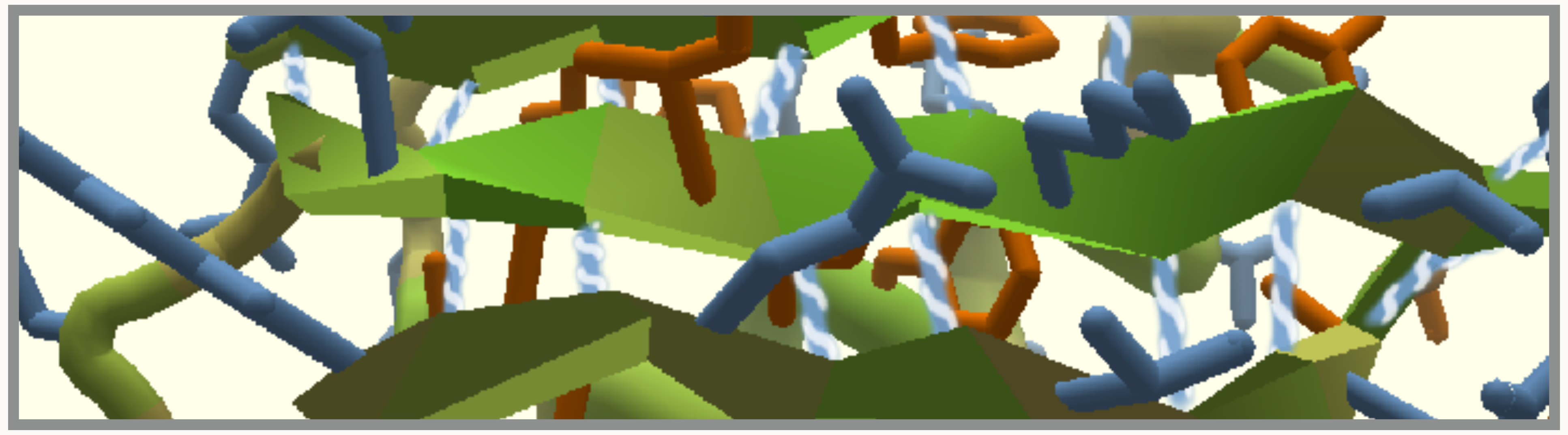
WE ARE NOT EQUIPPED FOR LARGE-SCALE OPEN-ENDED RESEARCH

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

CROWDSOURCING PROCESS

open call → group meeting

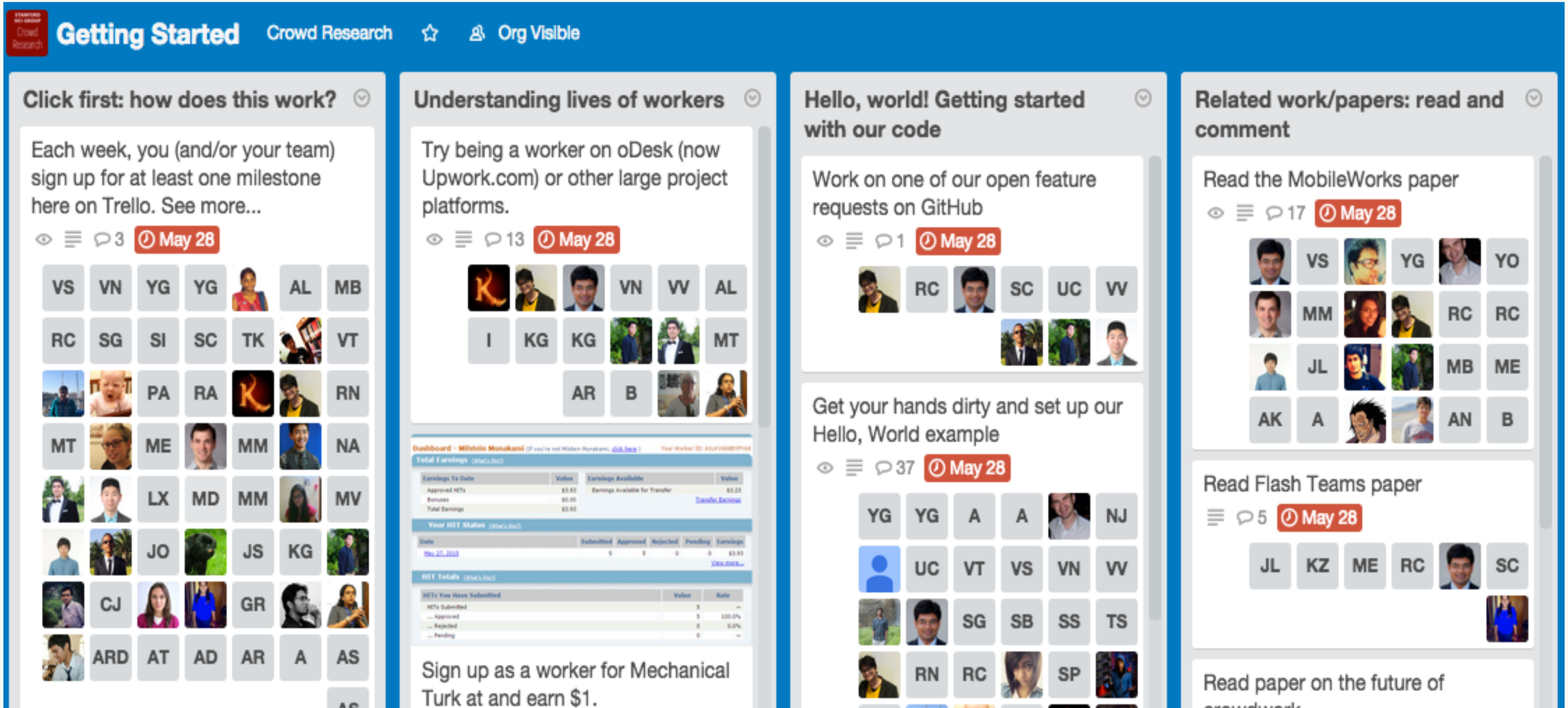
milestone deadline

peer assessment


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!
hi ^_^
meenalmandil 9:02 AM
ossolorzano 9:02 AM
hello
Michael
Nicole
Akshara
Rashmi
Nicole
12-85 weeks
500,000 Slack messages
190,000 minutes of video meetings

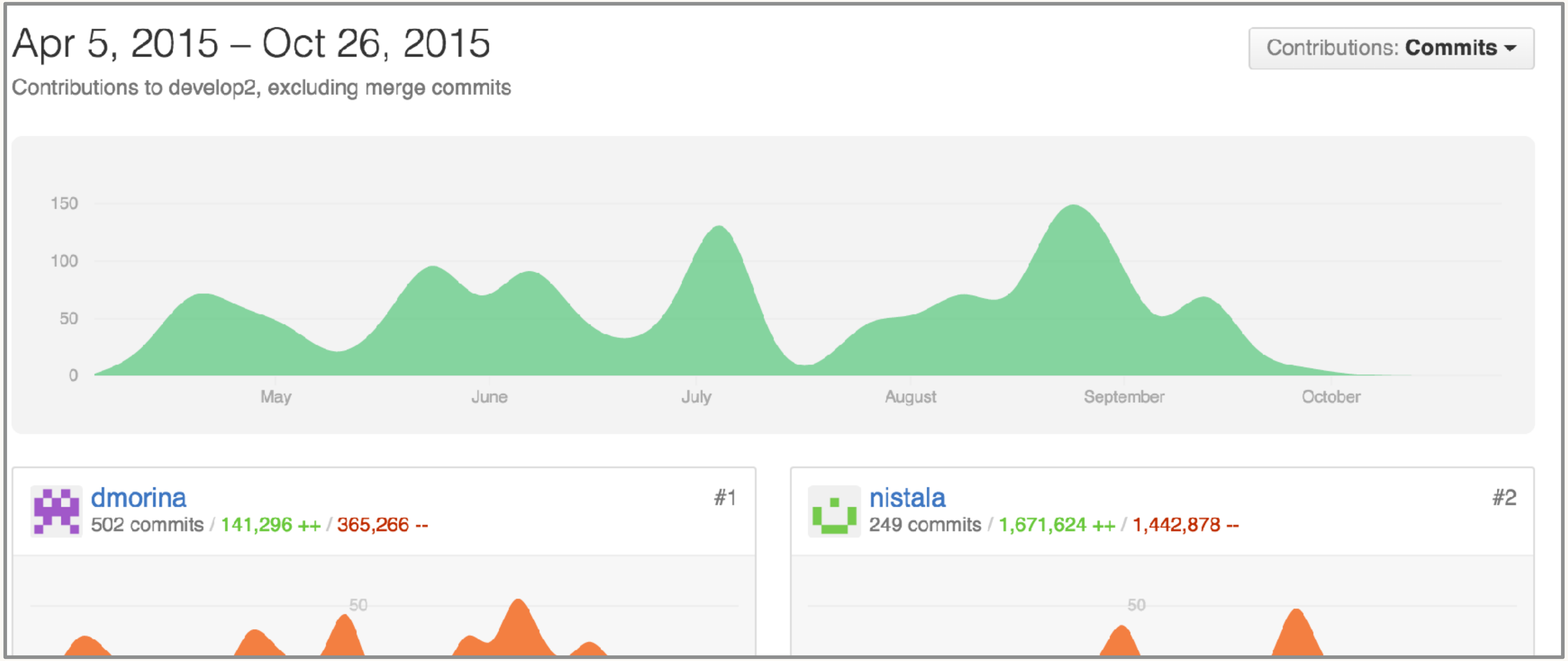
TASK PLANNING

MILESTONES



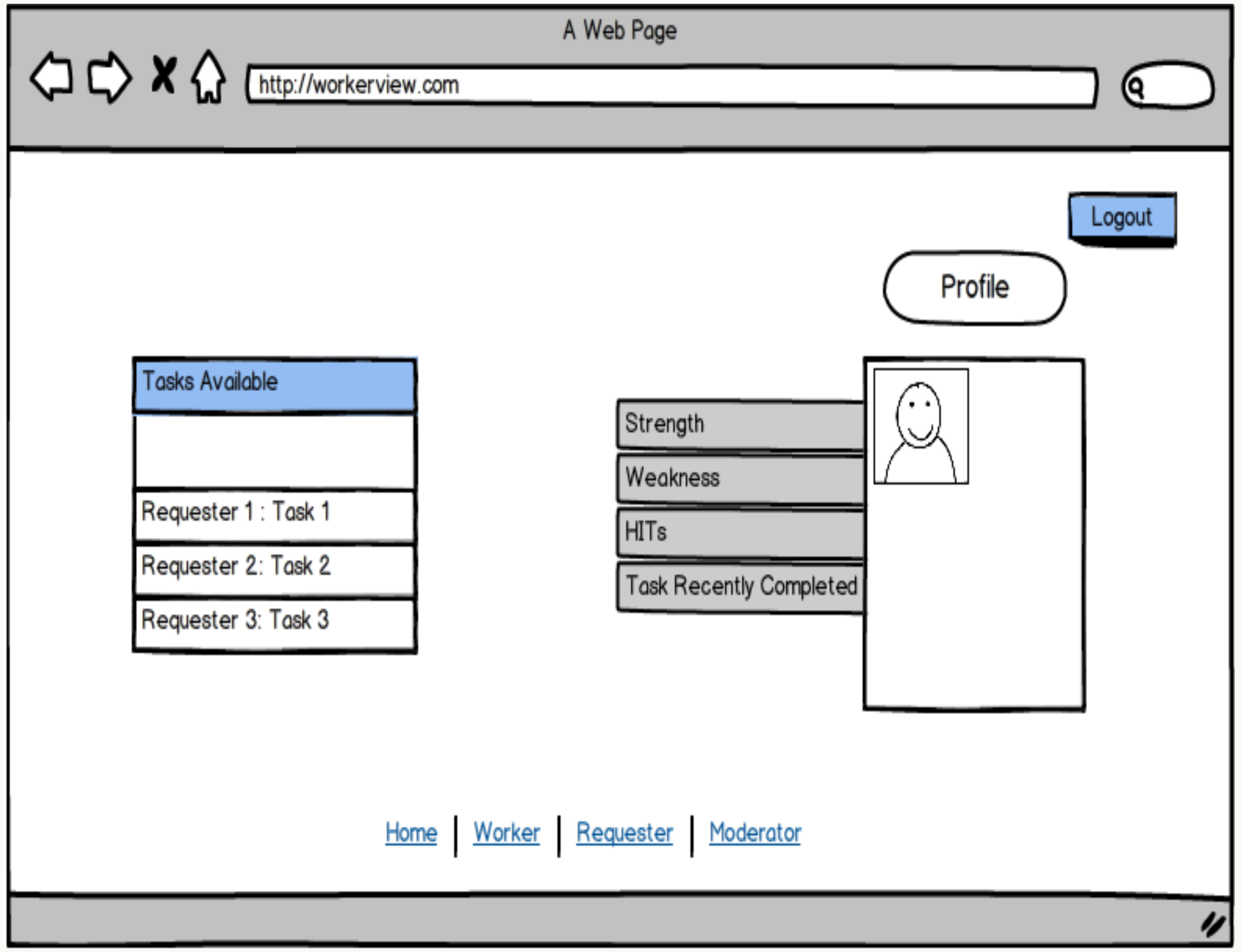
ENGINEERING

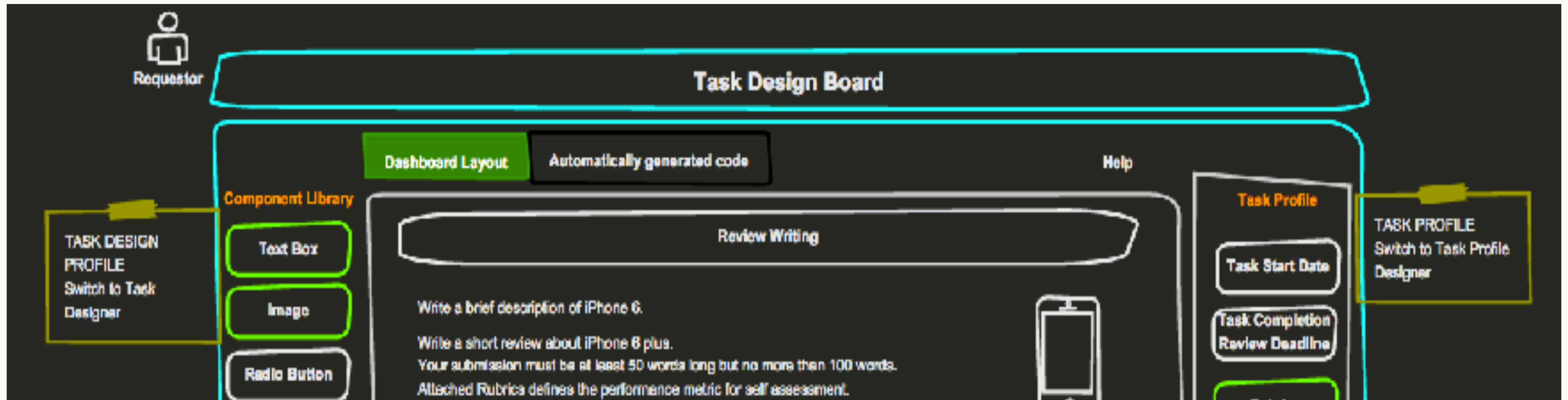
MILESTONES



PROTOTYPING

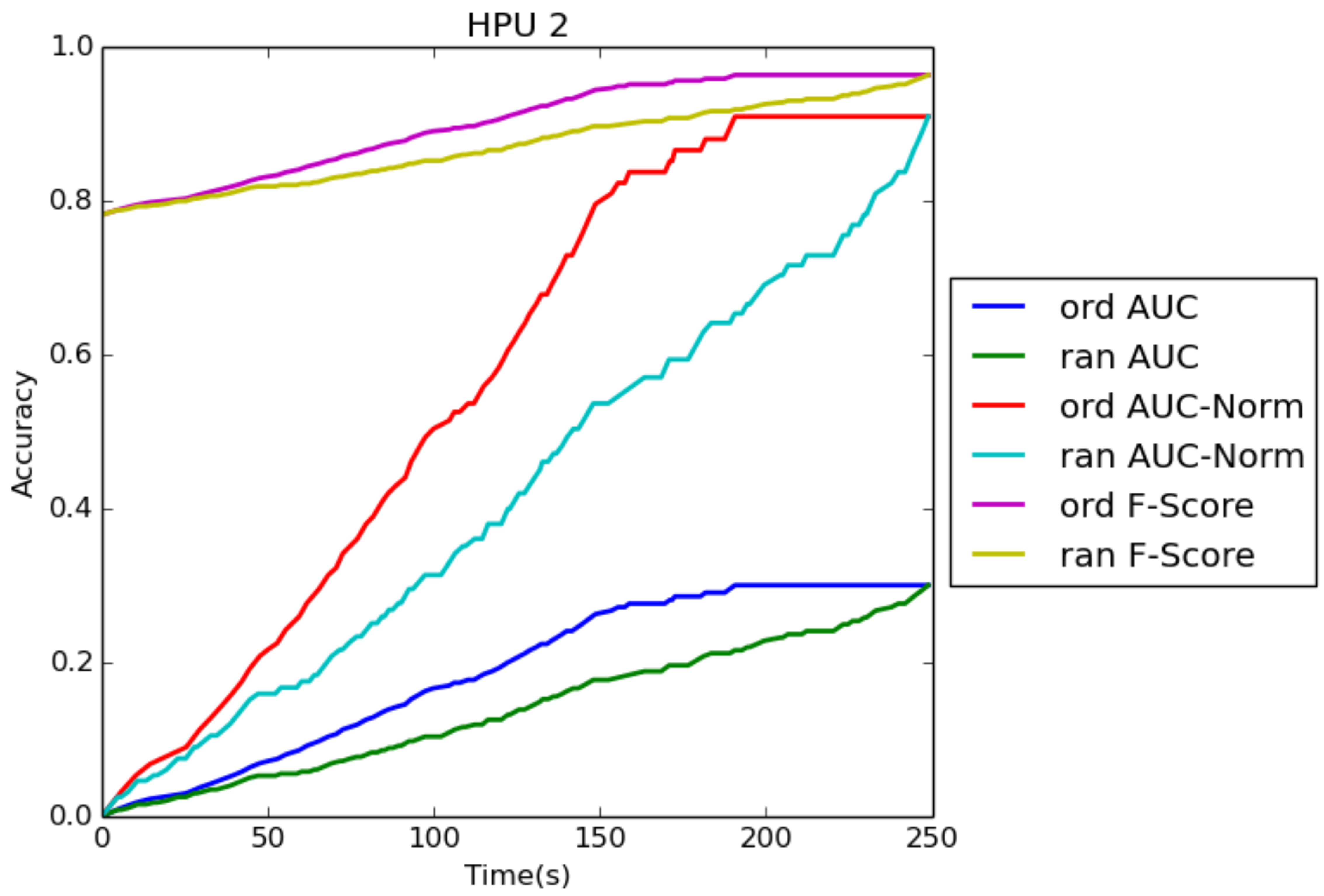
MILESTONES

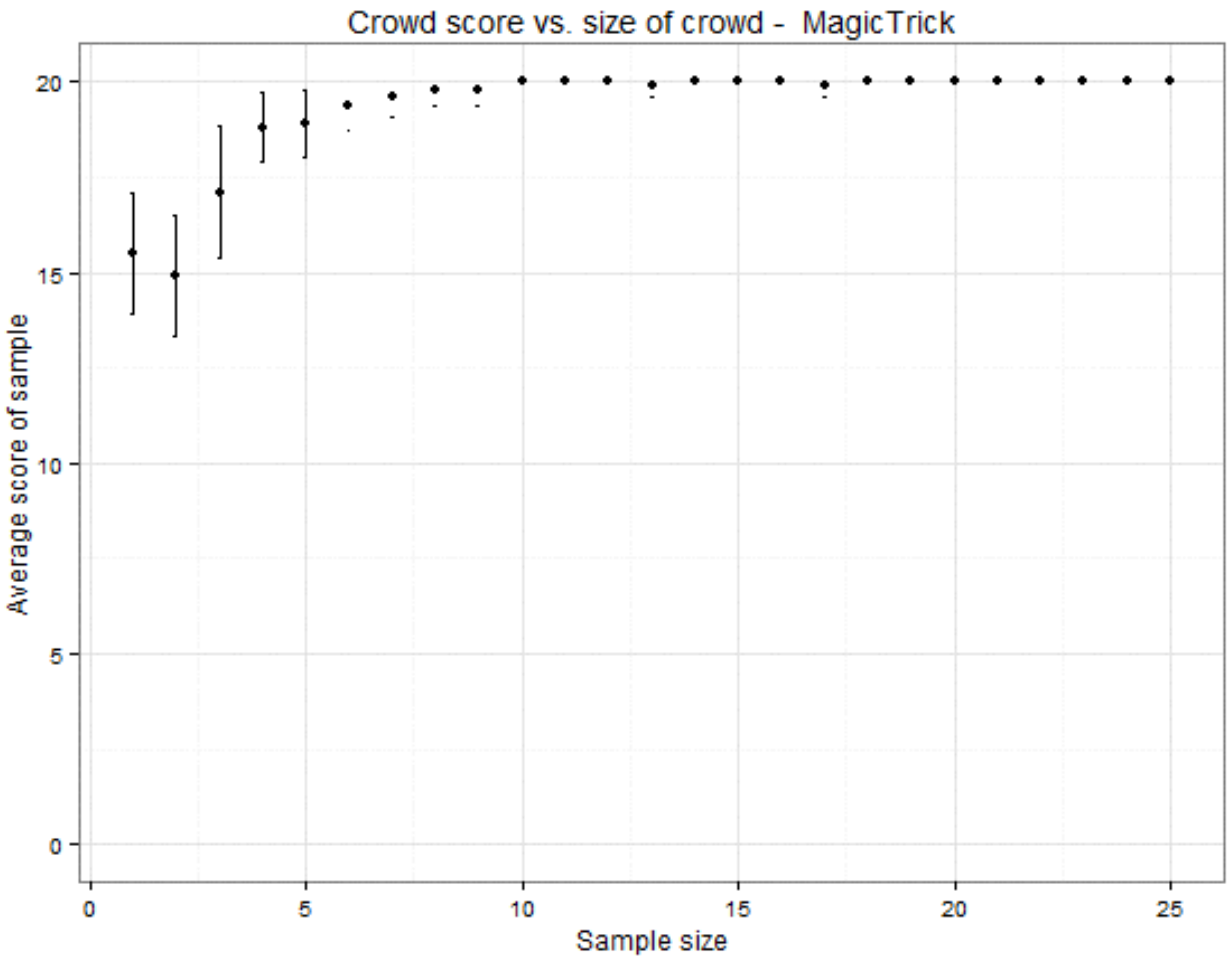




DATA ANALYSIS

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 requestors 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

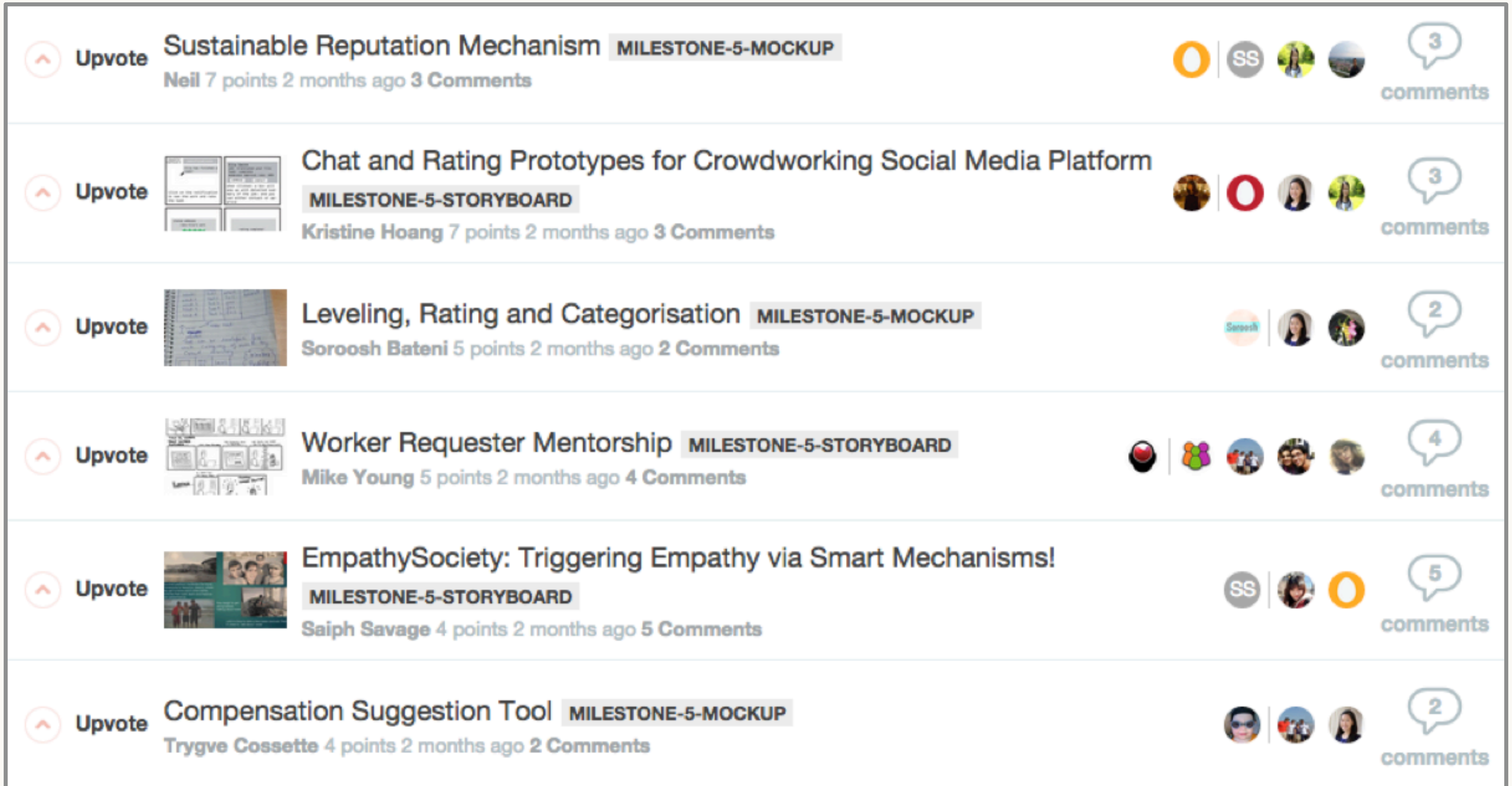
I am not sure how "fair" these elections would be. Money and power could play a major role here. I may be referring to a case that has extremely small possibility of occurring, but, what if the intentions of the elected members are changed or are influenced by some other party?

Reply...

Anonymous

PEER ASSESSMENT

MILESTONES




Andrew Ng, Stanford and Baidu Research
Andrew

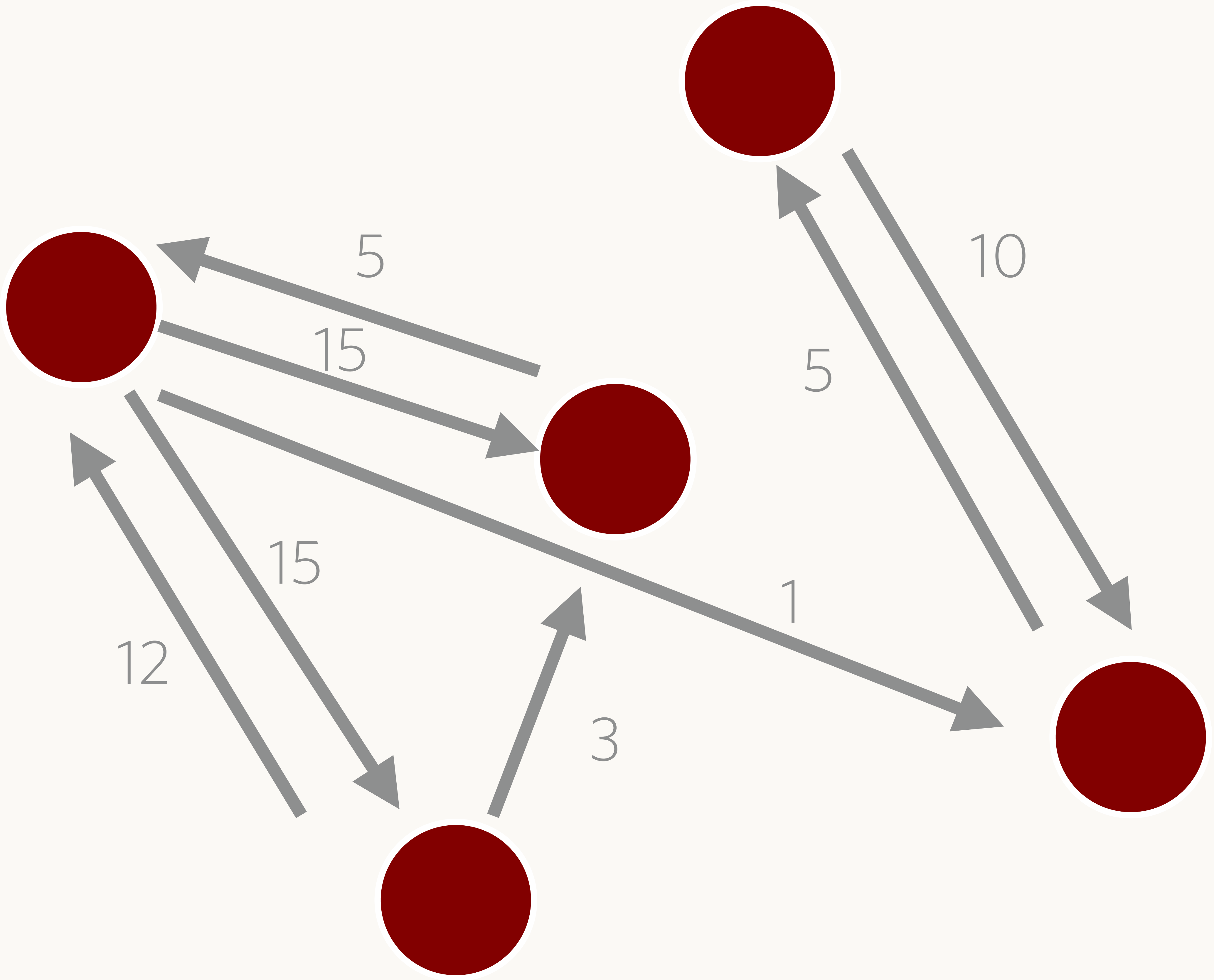

Peter Norvig, Google
Peter

Anant Agarwal, MIT and EdX

DECENTRALIZED CREDIT: TURN IT INTO A GRAPH PROBLEM

Each participants allocates 100 credit points to other participants based on their assessment of who impacted the project

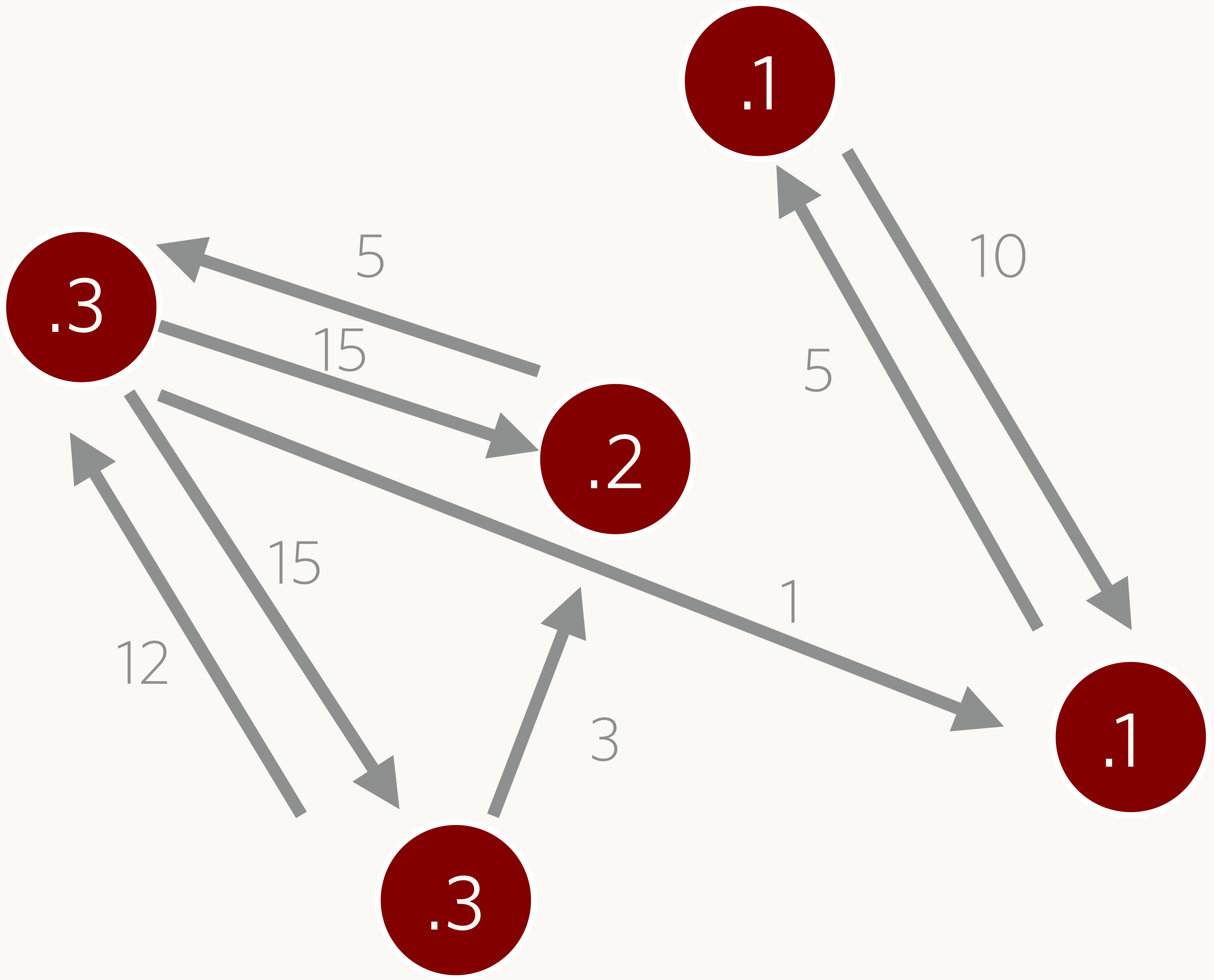
Resulted: weighted directed graph



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 such as PageRank can correct for most of these attacks

1500 participants from six continents

2% high school, 73% undergrad, 22% master's, 3% PhD

RECRUITMENT: PROVIDING ACCESS

Matching names to DBLP: **90% with no prior research experience**

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

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

LARGE-SCALE PROJECTS

Design and develop a new paid crowdsourcing platform

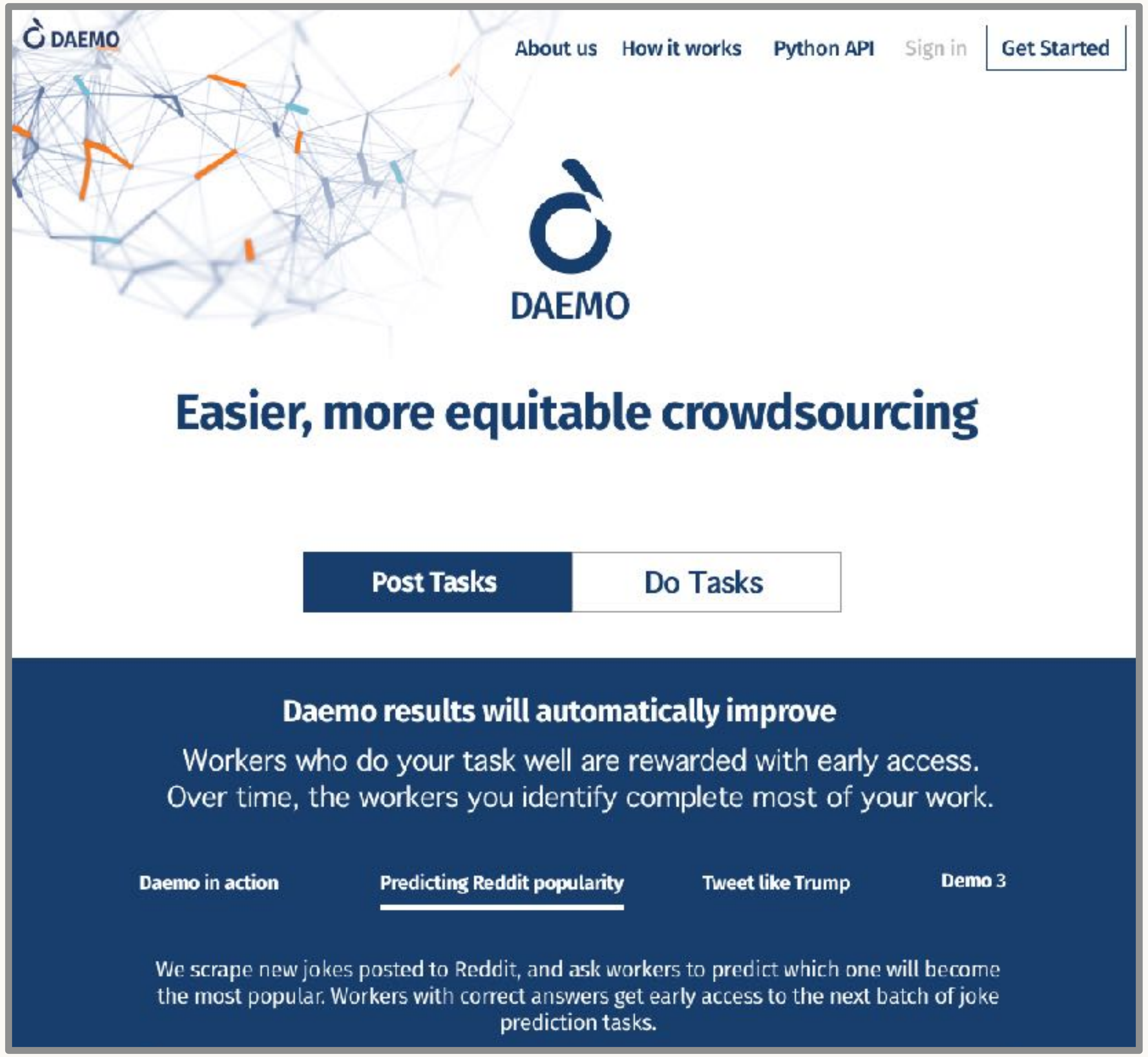
- Michael Bernstein, Stanford, HCI

Run hundreds of parallel experiments

- Sharad Goel, Stanford, Data Science

Create new hybrid human-computer vision algorithms

- James Davis, UCSC, and Serge Belongie, Cornell Tech, Computer Vision



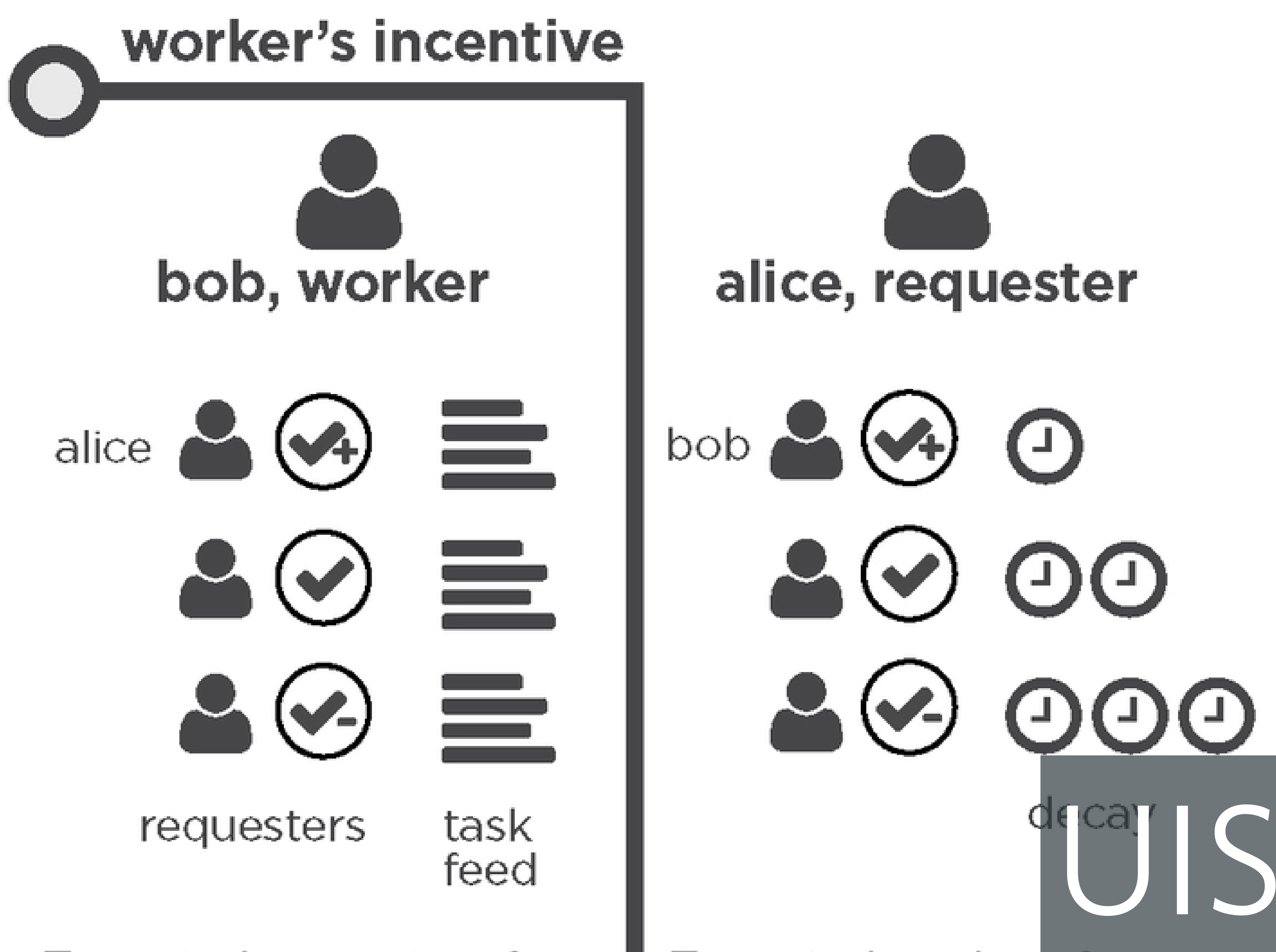
Boomerang: Rebounding the Consequences of Reputation Feedback on Crowdsourcing Platforms

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

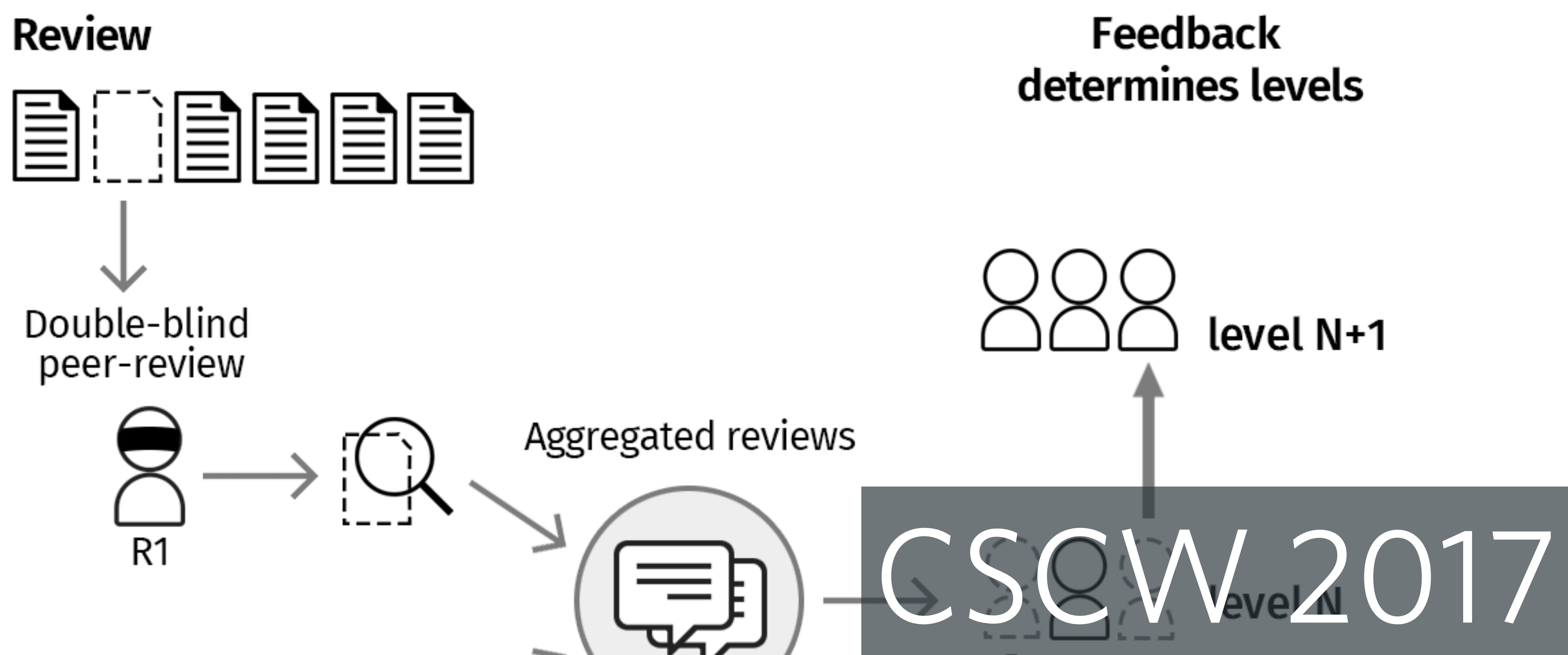
Paid crowdsourcing platforms suffer from low-quality work and unfair rejections, but paradoxically, most workers and requesters have high reputation scores. These inflated scores, which make high-quality work and workers difficult to find, stem from social pressure to avoid giving negative feedback. 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, Alipta 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. Gaikwad, D. Morina, R. Nistala, M. Agarwal, A. Cossette, R. Bhanu, S. Savage, V. Narwal, K. Rajpal, J. Regino, A. Mithal, A. Ginzberg, A. Nath, K. R. Ziulkoski, T. Cossette, D. Gamage, A. Richmond-Fuller, R. Suzuki, J. Herrejon, K. V. Le, C. Flores-Saviaga, H. Thilakarathne, K. Gupta, W. Dai, A. Sastry, S. Goyal, T. Rajapakshe, N. Abolhassani, A. Xie, A. Reyes, S. Ingle, V. Jaramillo, M.D. Godinez, W. Angel, M. Godinez, C. Toxtli, J. Flores, A. Gupta, V. Sethia, D. Padilla, K. Milland, K. Setyadi, N. Wajirasena, M. Batagoda, R. Cruz, J. Damon, D. Nekkanti, T. Sarma, M.H. Saleh, G. Gongora-Svartzman, S. Bateni, G. Toledo-Barrera, A. Pena, R. Compton, D. Aariff, L. Palacios, M. P. Ritter, Nisha K.K., A. Kay, J. Uhrmeister, S. Nistala, M. Esfahani, E. Bakiu, C. Diemert, L. Matsumoto, M. Singh, V. Jaramillo-Lopez, K. Patel, R. Krishna, G. Kovacs, R. Vaish, M. Bernstein

UIST '15 Adjunct, November 08-11, 2015, Charlotte, NC, USA.
ACM 978-1-4503-3780-9/15/11
http://dx.doi.org/10.1145/2815585.2815739

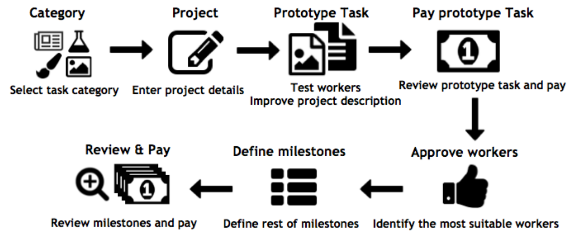


Figure 1. Task creation workflow for a requester: prototype task creation, initial submissions review, and hiring high quality workers for future milestones. [`https://daemo.stanford.edu`]. Icon courtesy Font Awesome by Dave Gandy - http://fontawesome.io

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 Vaish, Serge Belongie, James Davis.
Top researchers: Vishal Anand, Anshu Aviral, Prithvijit Chakrabarty, Yash Chandak, Sidharth Chaturvedi, Chinmaya Devaraj, Ankit Dhall, Utkarsh Dwivedi, Sanket Gupte, Sharath N. Sridhar, Karthik Paga, Anuj Pahuja, Aditya Raisinghani, Ayush Sharma, Shweta Sharma, Darpana Sinha, Nisarg Thakkar, K. Bala Vignesh, Utkarsh Verma,
Researchers: Kanniganti Abhishek, Amod Agrawal, Arya Aishwarya, Aurgho Bhattacharjee, Sarveshwaran Dhanasekar, Venkata Karthik Gullapalli, Shuchita Gupta, Chandana G, Kinjal Jain, Simran Kapur, Meghana Kasula, Shashi Kumar, Parth Kundaliya, Utkarsh Mathur, Alankrit Mishra, Aayush Mudgal, Aditya Nadimpalli, Munakala Sree Nihit, Akanksha Periwal, Ayush Sagar, Ayush Shah, Vikas Sharma, Yashovardhan Sharma, Faizal Siddiqui, Virender Singh, Abhinav S., Pradyumna Tambwekar, Rashida Taskin, 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 reordering 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.

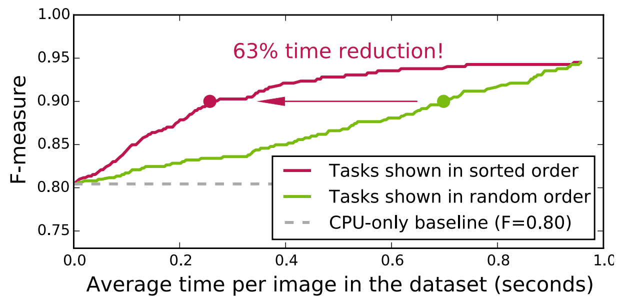


Figure 1: Consider a simple localization task where crowd workers refine the output of a machine classifier. At a threshold of 0.5, baseline accuracy starts at 0.80 (gray dotted line). If we show random tasks to human workers, accuracy improves (green), but if we order tasks by increasing machine confidence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

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 (Russakovsky, 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 scenerio, 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 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.

*This project was created via a world-wide, crowdsourced research process initiated by UC Santa Cruz, Stanford University, and Cornell Tech.

Investigating the "Wisdom of Crowds" at Scale

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

Vikas S Yaligar‡
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 Institutions †

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.

Ramesh Arvind, Chiraag Sumanth *, Arvind Srikantan*, Bhargav HS*, Mayank Pahadia‡, Tushar Dobhal‡, Atif Ahmed‡, Mani Shankar‡, Himani Agarwal, Rajat Agarwal, Sai Anirudh-Kondaveeti, Shashank Arun-Gokhale, Aayush Attri, Arpita Chandra, Yogitha Chilukuri, Sharath Dharmaji, Deepak Garg, Naman Gupta, Paras Gupta, Glincy Mary Jacob, Siddharth Jain, Shashank Joshi, Tarun Khajuria, Sameeksha Khillan, Sandeep Konam, Praveen Kumar-Kolla, Sahil Loomba, Rachit Madan, Akshansh Maharaja, Vidit Mathur, Bharat Munshi, Mohammed Nawazish, Venkata Neehar-Kurukunda, Venkat Nirmal-Gavarraju, Sonali Parashar, Harsh Parikh, Avinash Paritala, Amit Patil, Rahul Phatak, Mandar Pradhan, Abhilasha Ravichander, Krishna Sangeeth, Sreecharan Sankaranarayanan, Vibhor Sehgal, Ashrith Sheshan, Suprajha Shibiraj, Aditya Singh, Anjali Singh, Prashant Sinha, Pushkin Soni, Bipin Thomas, Lokesh Tuteja, Kasyap Varma-Dattada, Sukanya Venkataraman, Pulkit Verma, Ishan Yelurwar
†Jaypee Institute Of Information Technology; BITS; National Institute of Technology Karnataka, Indian Institute of Technology Delhi; PES Institute of Technology; International Institute of Information Technology; BMS College of Engineering; Bhagwan Parshuram Institute of Technology; Indian Institute of Technology Guwahati; College of Engineering; College of Engineering Chengannur; Bharati Vidyapeeth's College Of Engineering; Maharaja Agrasen Institute of Technology; Cluster Innovation Centre; Rajiv Gandhi University of Knowledge Technologies; Indraprastha Institute of Information Technology

UIST'15 Adjunct, November 08-11, 2015, Charlotte, NC, USA ACM 978-1-4503-3780-9/15/11. http://dx.doi.org/10.1145/2815585.2815725

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].

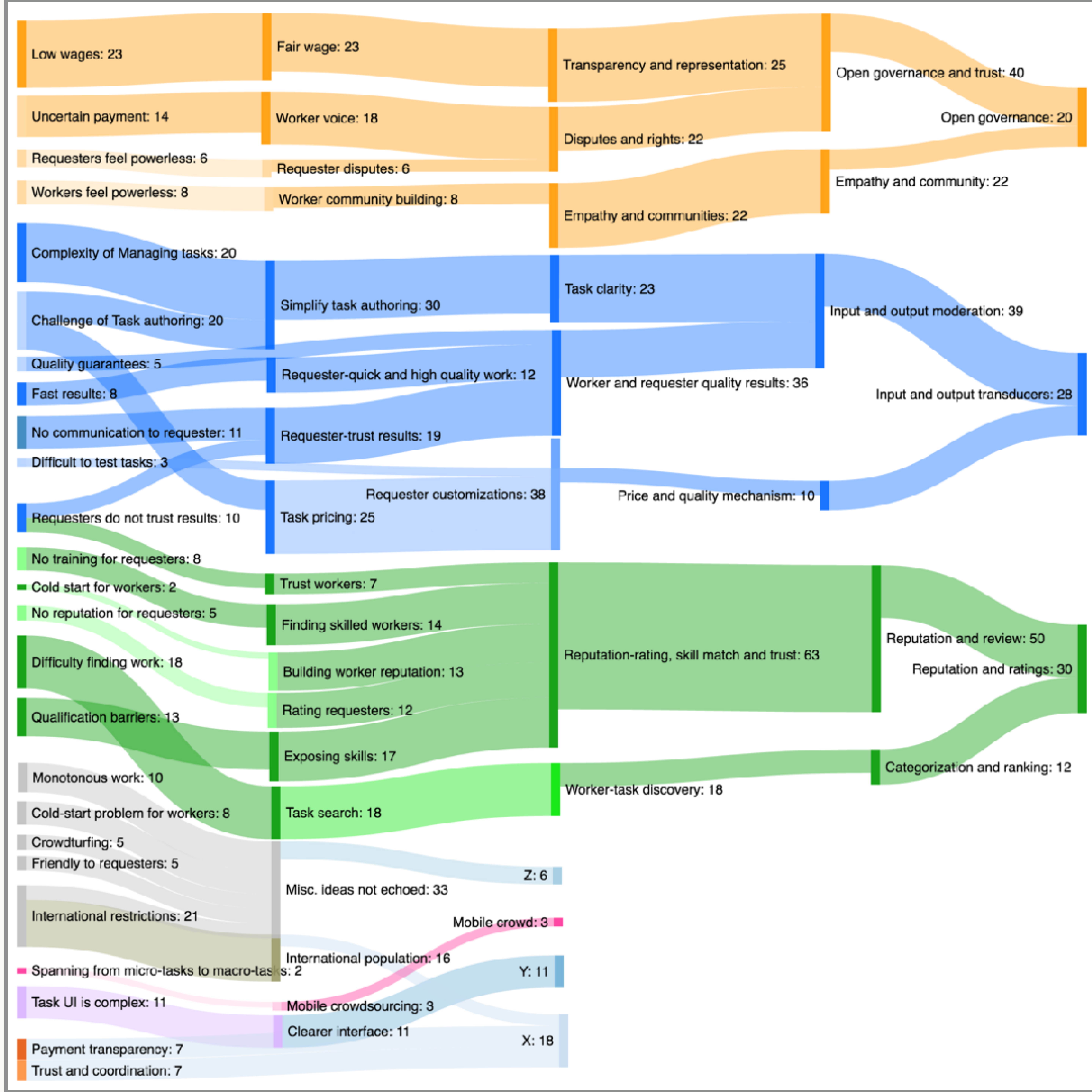
Simple aggregation—as in the case of Galton's ox competition, or voting in democratic elections—has been shown to be a surprisingly powerful technique for prediction, inference, and decision-making. Over the last century, there have been dozens of studies that examine this wisdom of crowds effect. For 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.

HCI Computer Vision Data Science

THE CROWD LED IDEATION

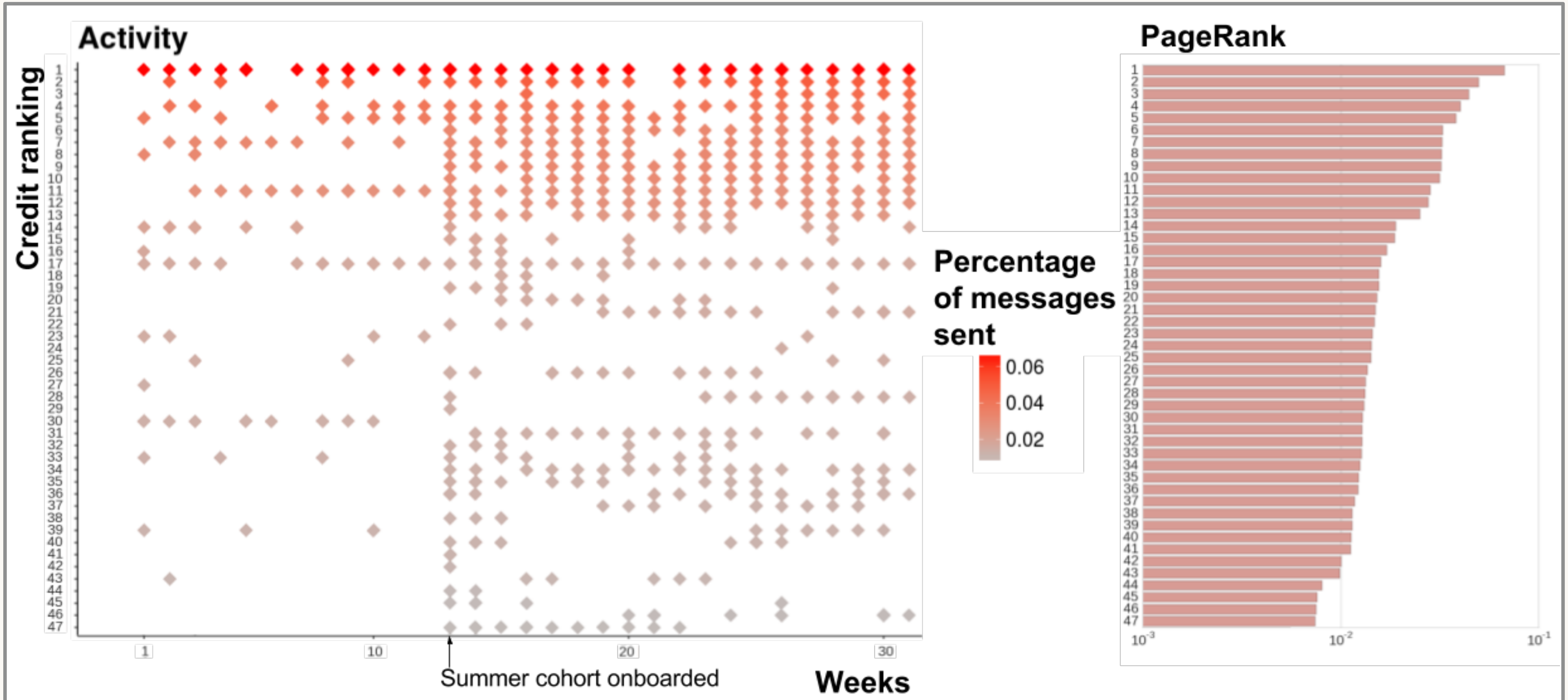
RESULTS

Thematic coding of milestone submissions across weeks



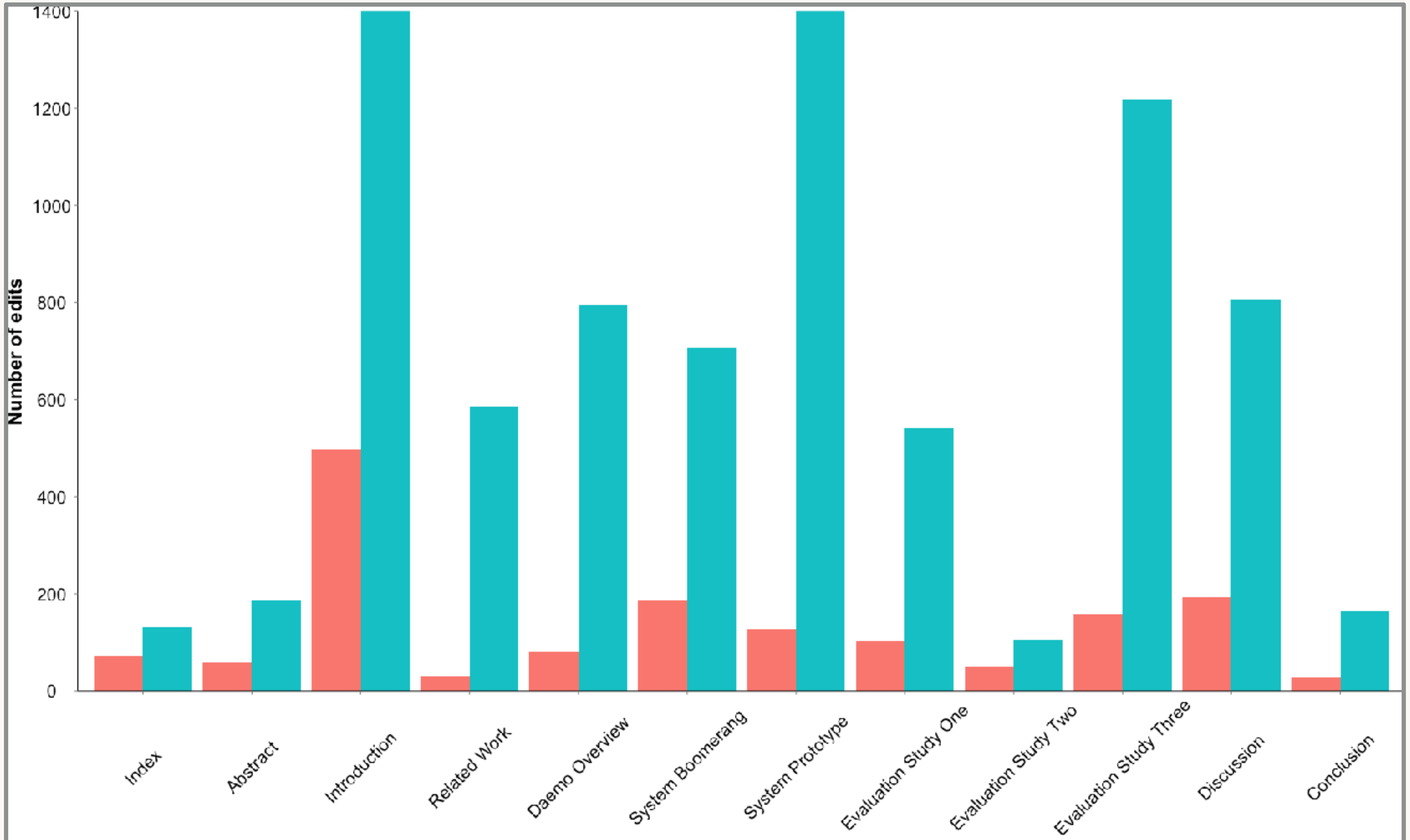
THE CROWD INVESTED TIME

RESULTS



THE CROWD LED WRITING

RESULTS



Count of number of edits to shared document

Crowd: 80%

Principal investigator: 20%

ANALYZING PAGERANK'S EFFECT

What impact did PageRank have on credit distribution?

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

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

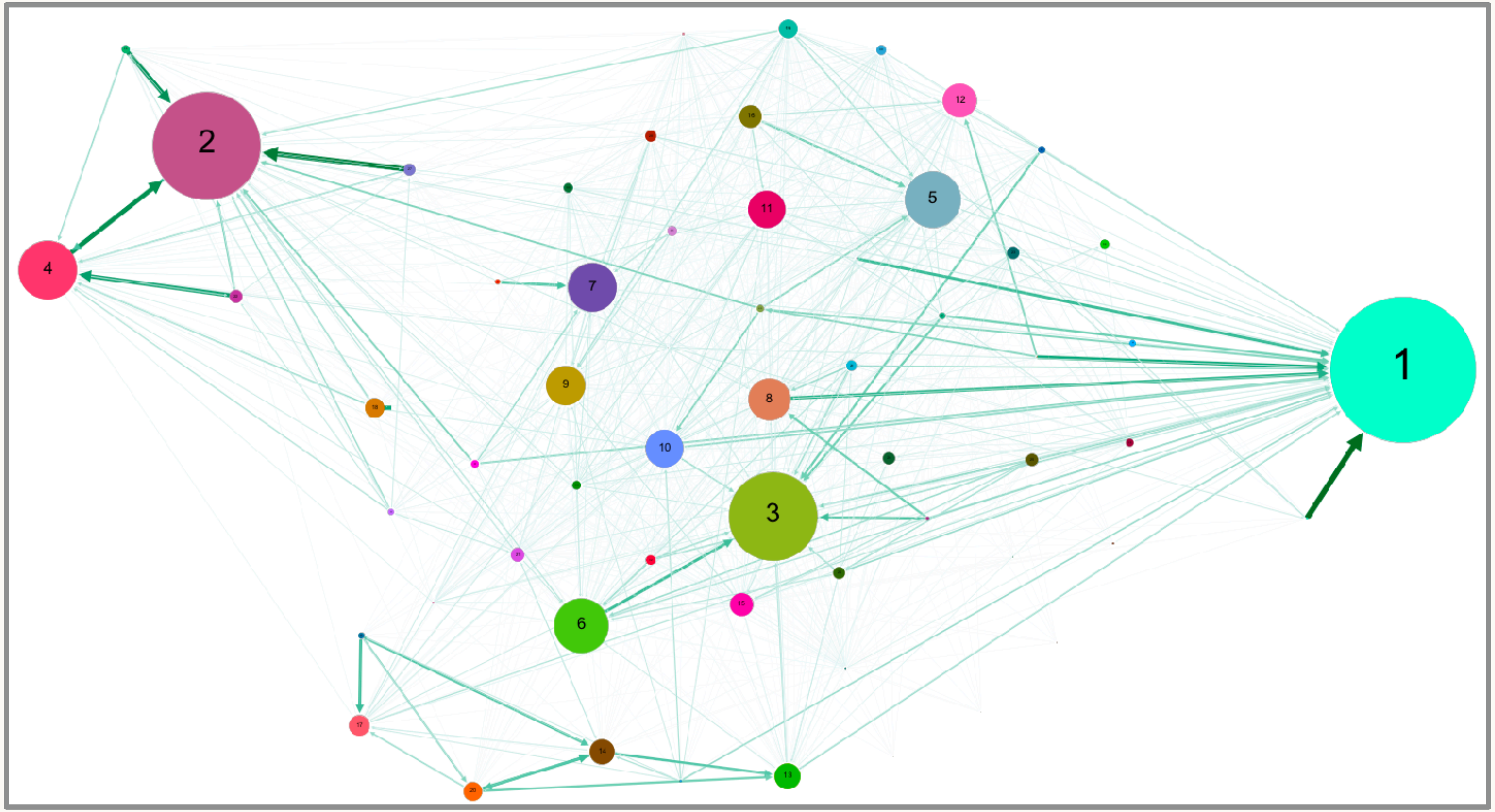
LESS TALKING, MORE DOING

Participation Measure	PageRank: β_{PR}	Raw Votes: β_{raw}	$\beta_{PR}-\beta_{raw}$
# 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

PAGERANK

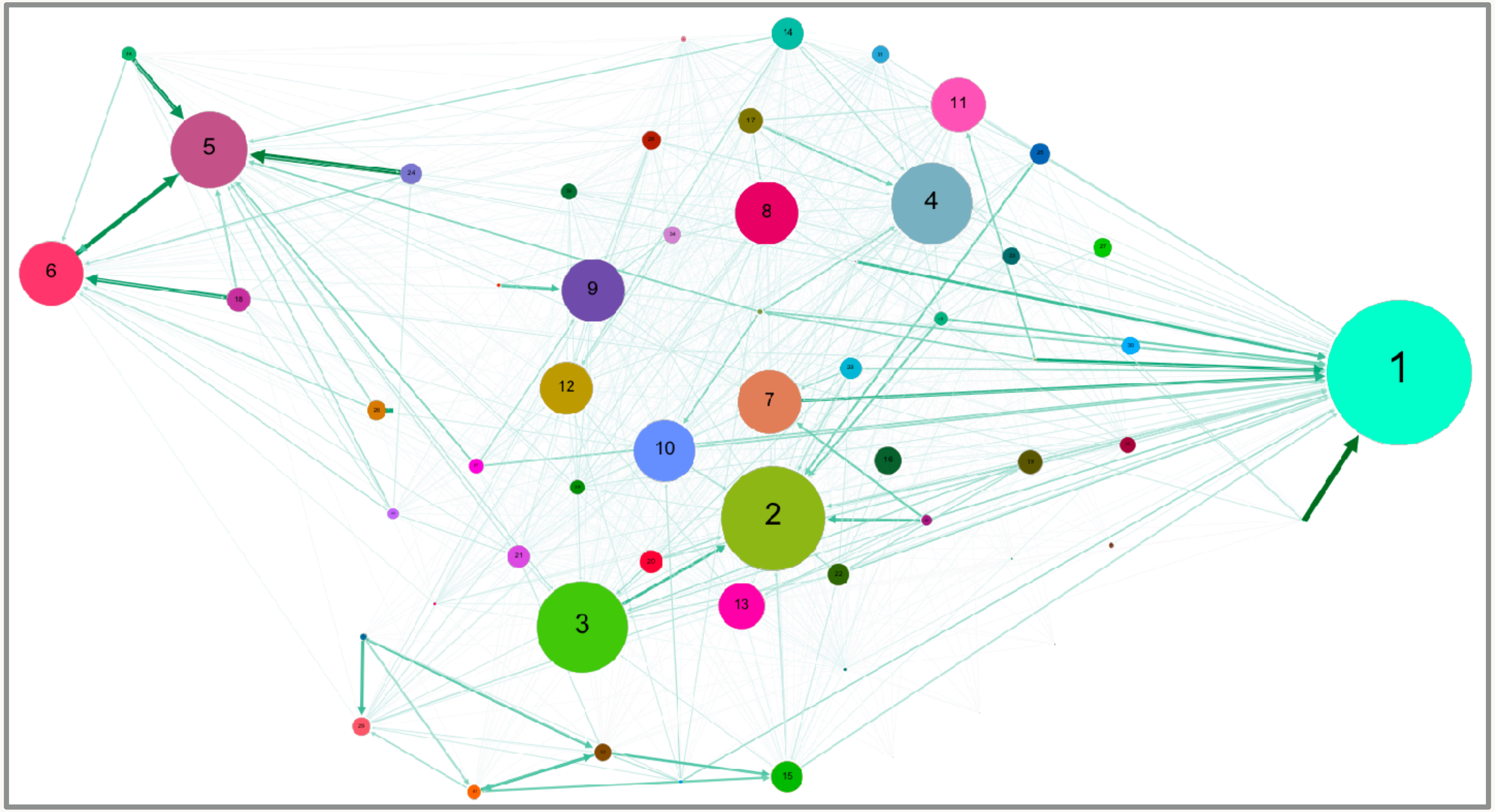


Raw vote ranking

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

EFFECTS ON AUTHOR ORDER

PAGERANK



PageRank-corrected author order

Influential coauthors reduced impact of link ring

REFLECTIONS

Computation and crowdsourcing can scale not just the teaching of new skills and the execution of research, but the experience of research and upward career mobility as well

Current and future contributions:

- Decentralized evaluation could help even traditional groups escape the tyranny of top-down review
- Projects that not only reach more people, but operate at a larger technical scale than traditional CS research

Rather than structuring crowds like algorithms, let's structure them like organizations.

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

In A Flash: Crowdsourcing Organizations, Collaboration, and Research

Thanks to…

Amazing students: Daniela Retelny, Niloufar Salehi, Rajan Vaish

Amazing colleagues: Melissa Valentine, Sharad Goel, James Davis

Amazing supporters: NSF, Sloan, Accenture, Microsoft, Toyota, Stanford Cyber Initiative, ONR, HPDTRP

Amazing questions