Experiments of collective social behavior in the “virtual lab”

Andrew Mao

Microsoft Research NYC

IC2S2 Tutorial – June 23, 2016

Outline

Part 1: **the bigger picture**

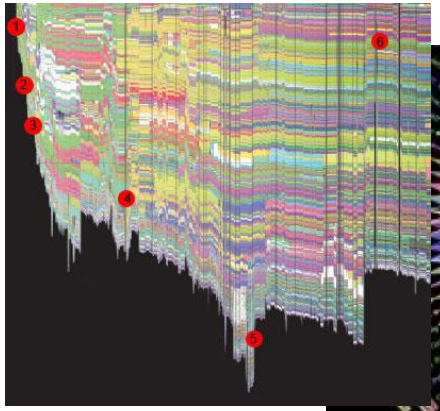
- Experiments as a part of computational social science
- Some interesting examples of online social experiments
- Live demo of experiment on TurkServer, our open-source platform

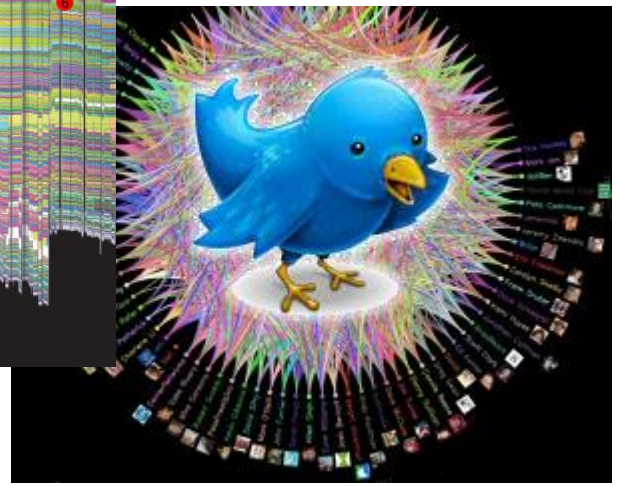
Part 2: **the nitty-gritty**

- Modern web programming and architecture of TurkServer
- Design and logistics for social experiments with crowdsourcing participants
- Questions, discussion, and brainstorming

Computational Social Science

Observational studies, ethnographic work



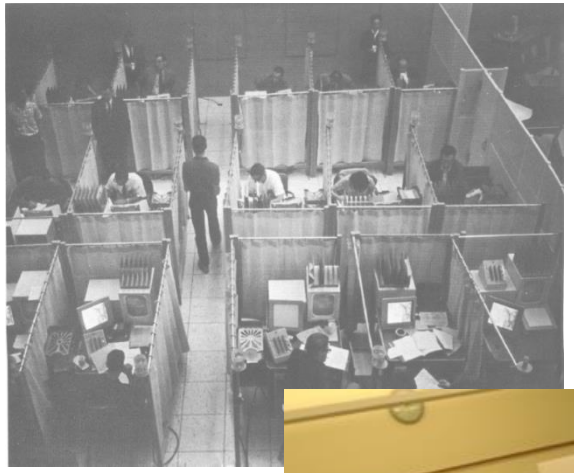


Data mining
Network analysis
Trend detection
…

Lab and field experiments

??

A brief history of the behavioral lab



ca. 1960s

ca. 2000s

- High degree of procedural control
- **Optimized for causal inference**

But, **many limitations**:

- Artificial environment
- Simple tasks, demand effects
- Homogeneous (WEIRD)* subject pools
- Time/scale limitations
- Expensive, difficult to set up

Poor generalization, expensive, slow

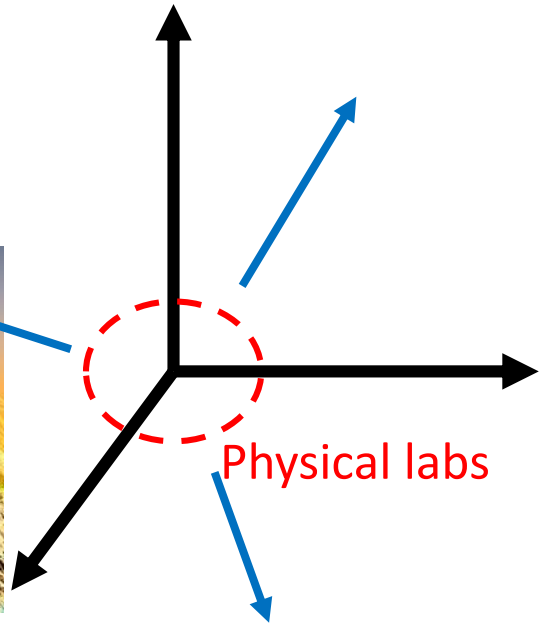
* [Henrich et al. 2010]

Bringing the lab closer to the real world

Complexity, Realism

- Realistic *vs.* abstract, simple tasks
- More precise instrumentation

Using the Internet as a behavioral lab



Duration, Participation

- Longer periods of time
- Fewer constraints on location

Size, Scale

- More samples of data
- Large-scale social interaction

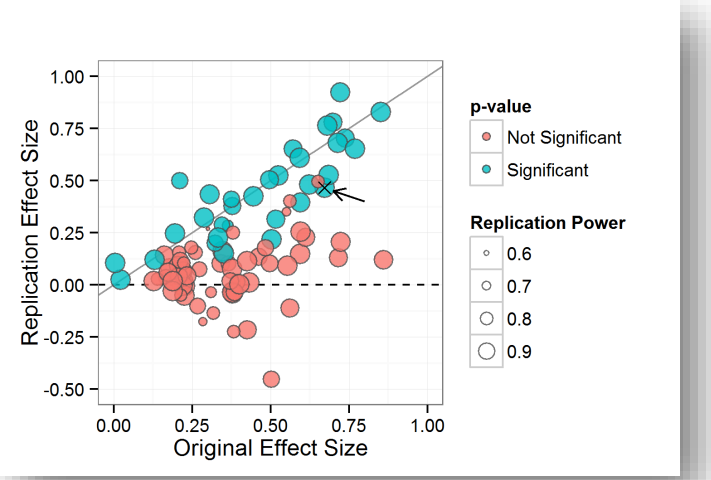
Benefits of the online lab

Larger, more diverse participant pool

Lower barriers to designing and conducting experiments

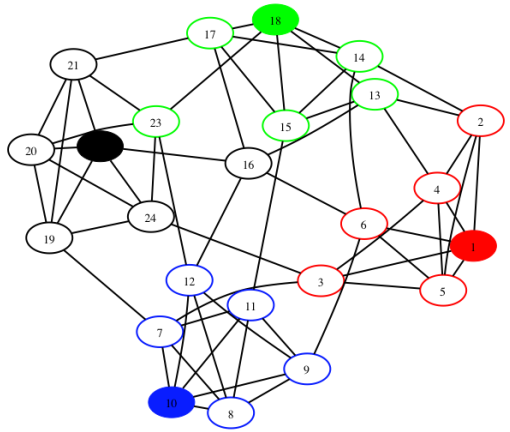
Data instrumentation for complex group interaction

Participation over longer time, broader space

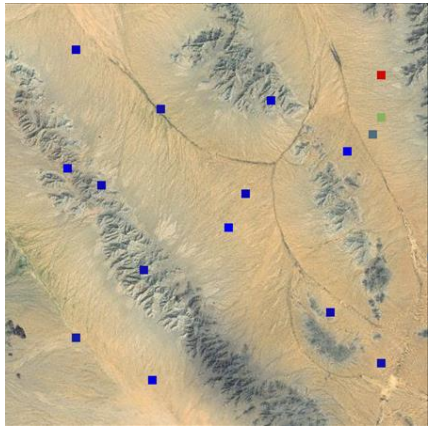


Easier replication, variation of existing work

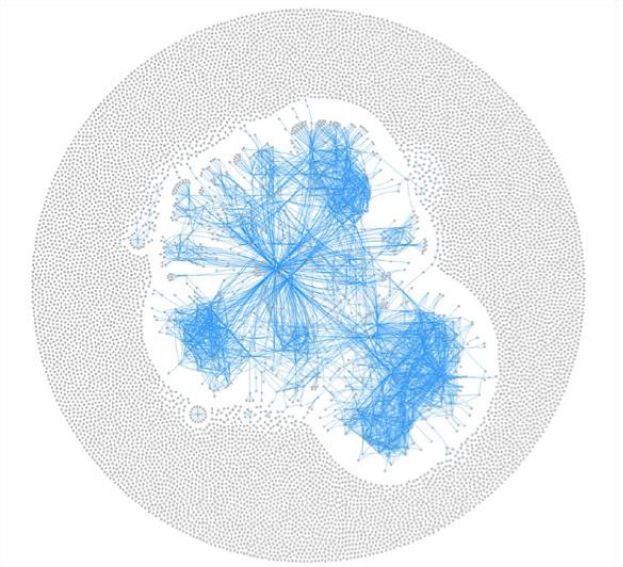
Today’s focus: online social experiments



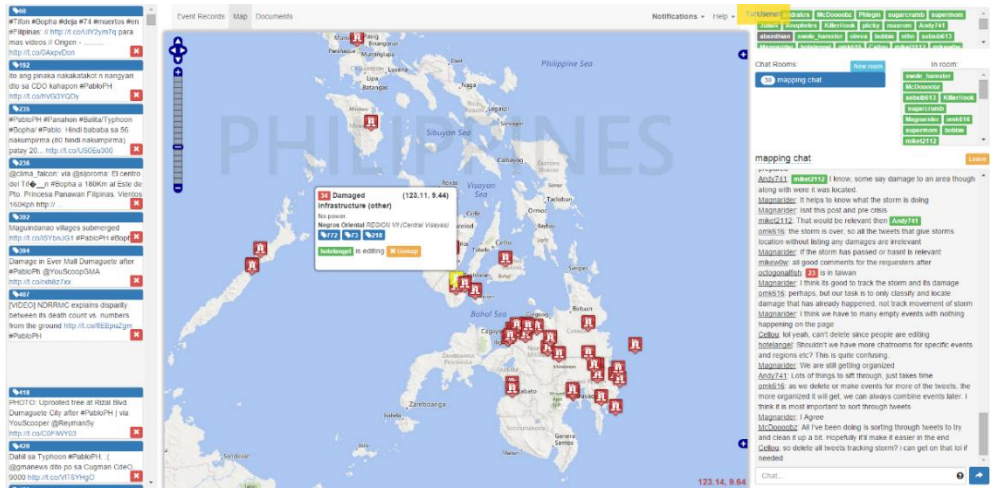
Suri and Watts (2011)



Mason and Watts (2011)

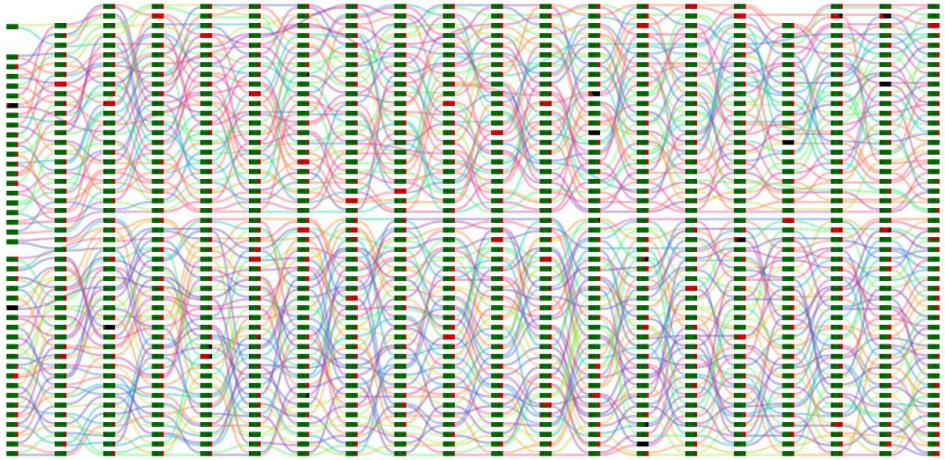


Yin, Gray, Suri, and Vaughan (2016)



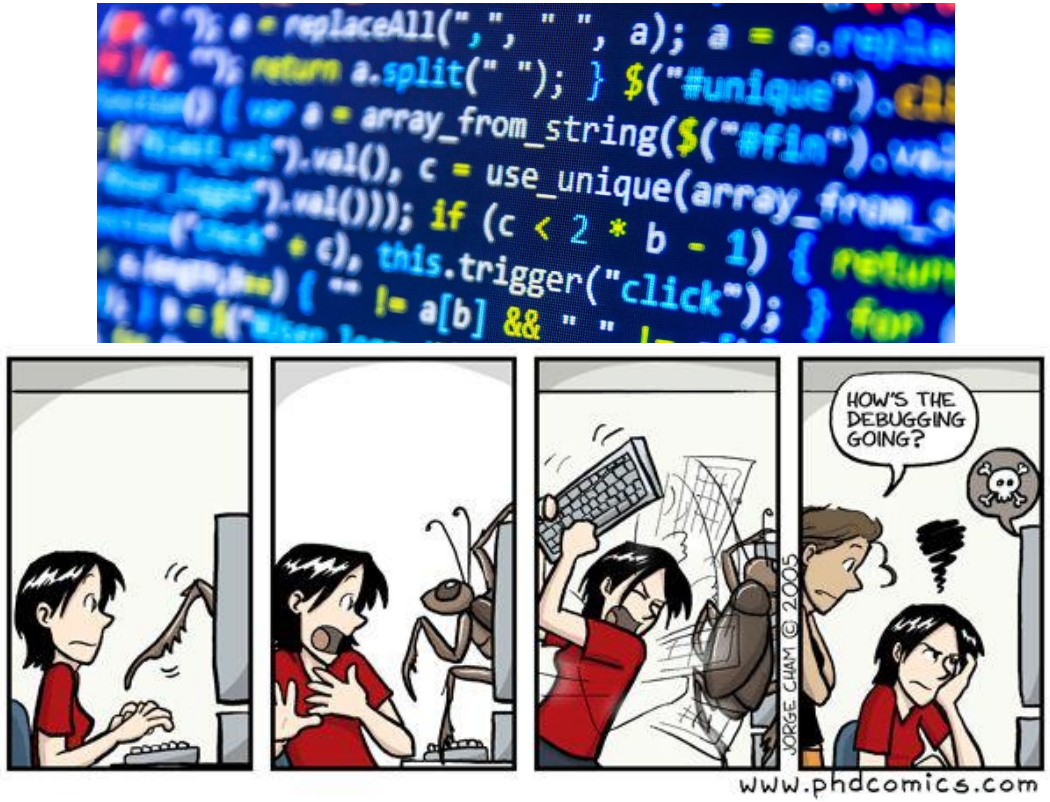
M., Mason, Suri, and Watts (2016)

M., Dworkin, Suri, and Watts (2016)

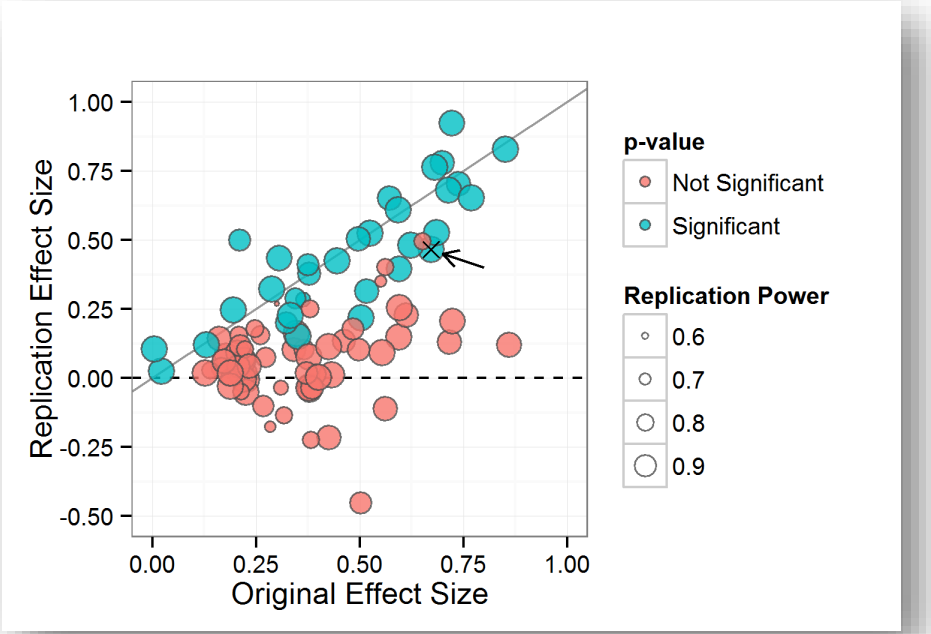


But, experiments are still pretty hard…

They're a lot of work, especially for studying social interaction.

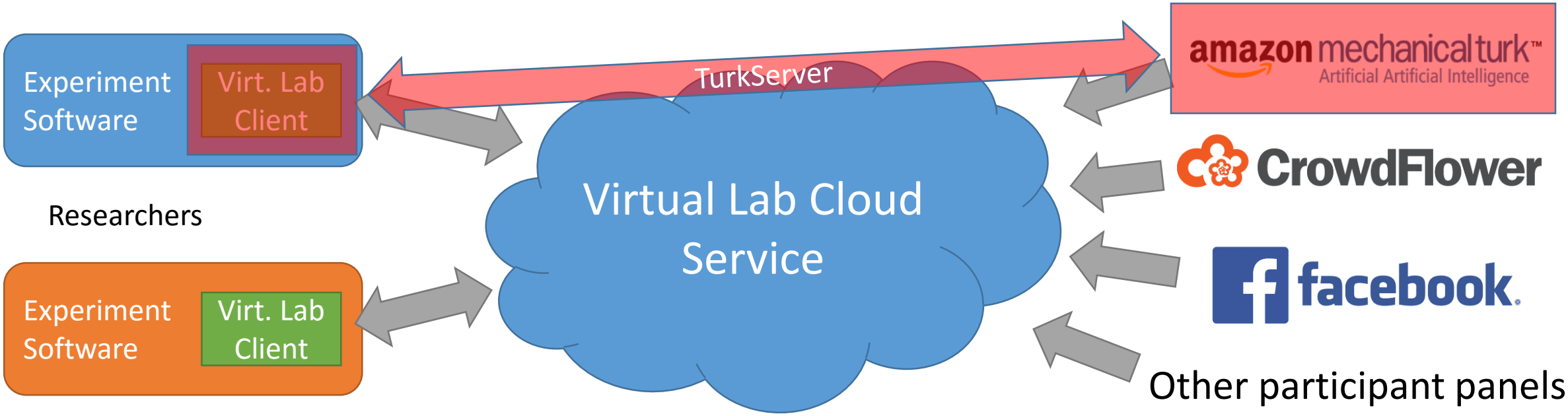


This hard work is discarded, or difficult to share and build upon.



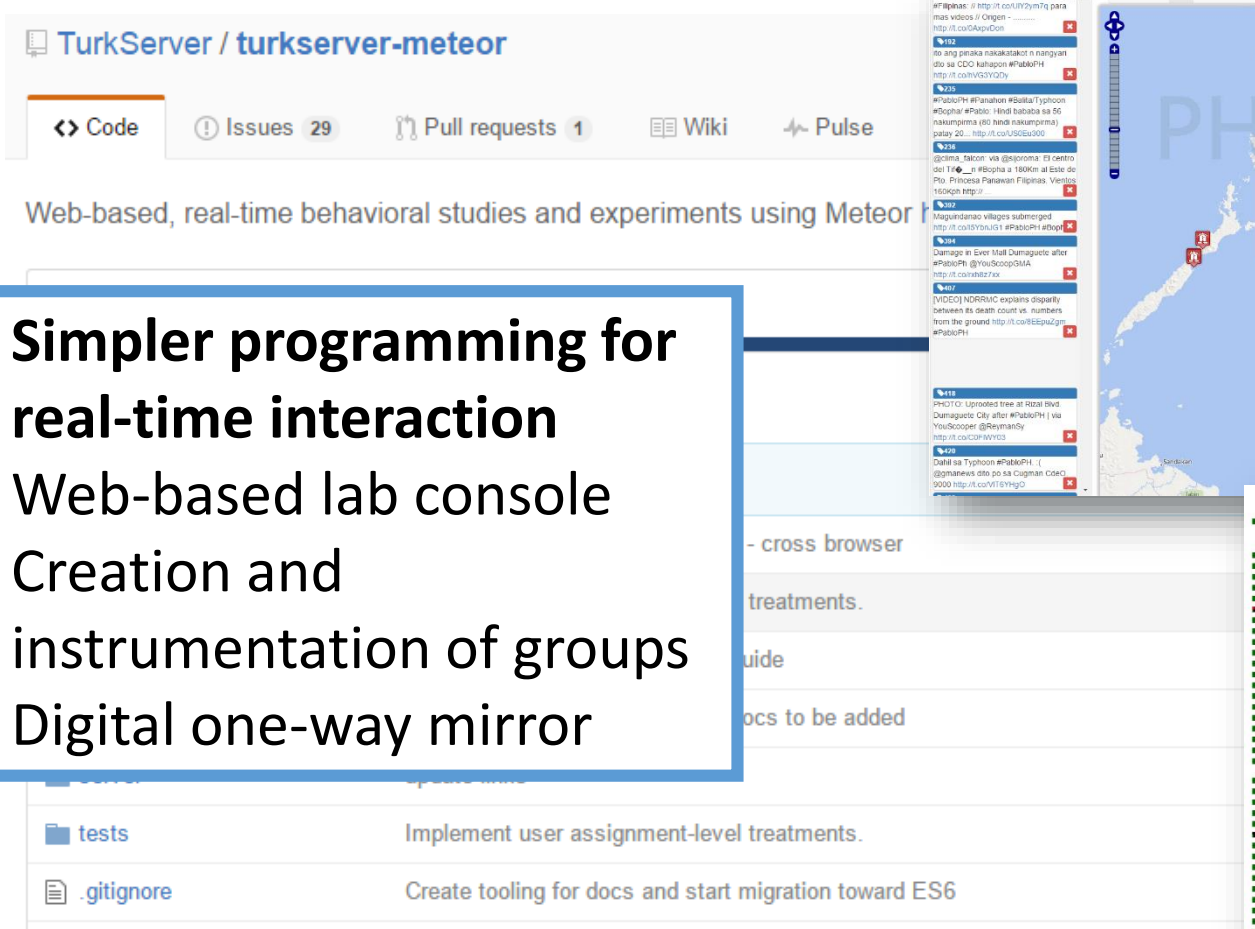
Estimating the Reproducibility of Psychological Science (2015)

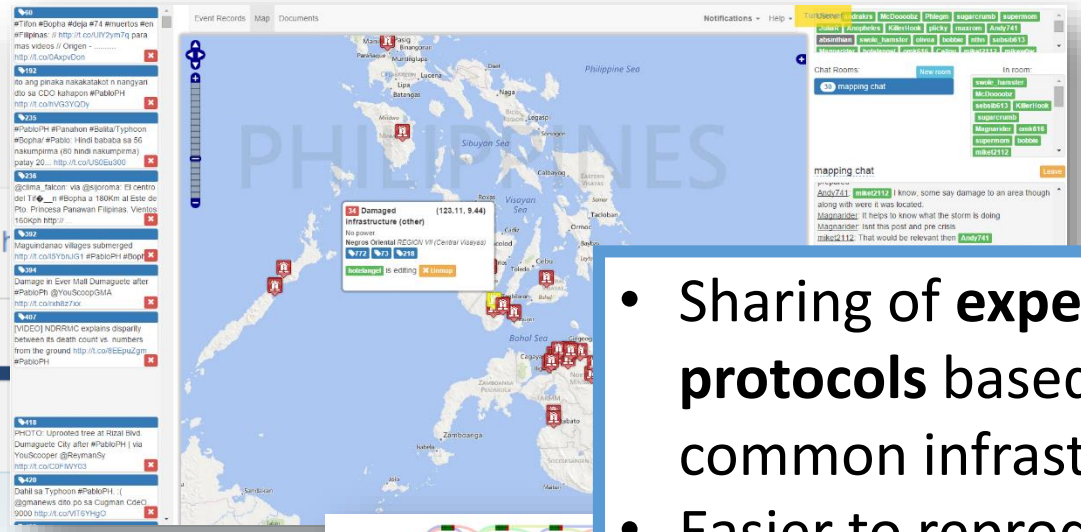
What would we like to have? One idea:

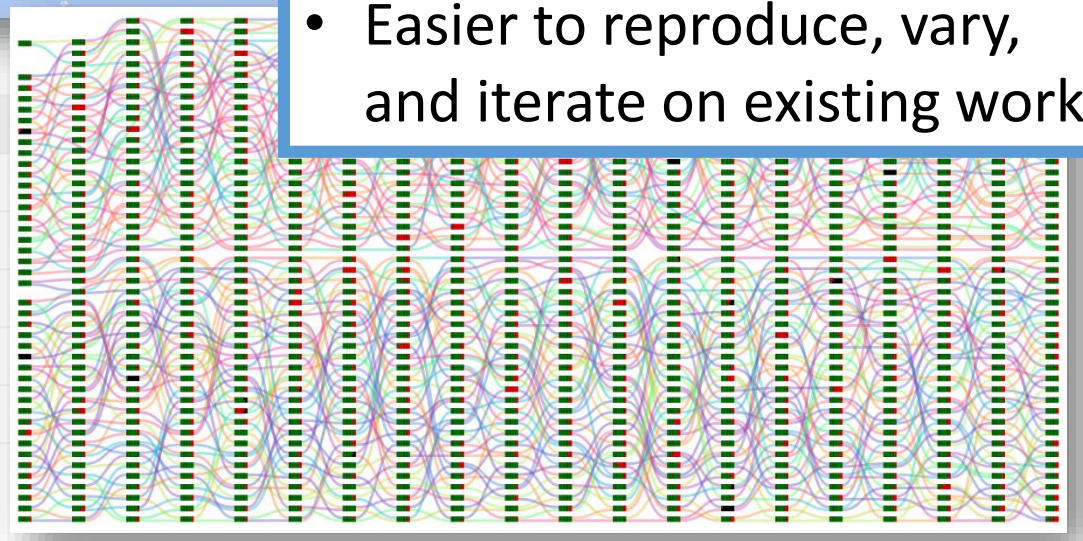


- Standardized OSS virtual lab interface
- Faster iteration, sharing, variation
- Consistent, scalable participant pool
- Demographic/experience tracking

TurkServer: OSS platform + experiments







- **Simpler programming for real-time interaction**
- Web-based lab console
- Creation and instrumentation of groups
- Digital one-way mirror

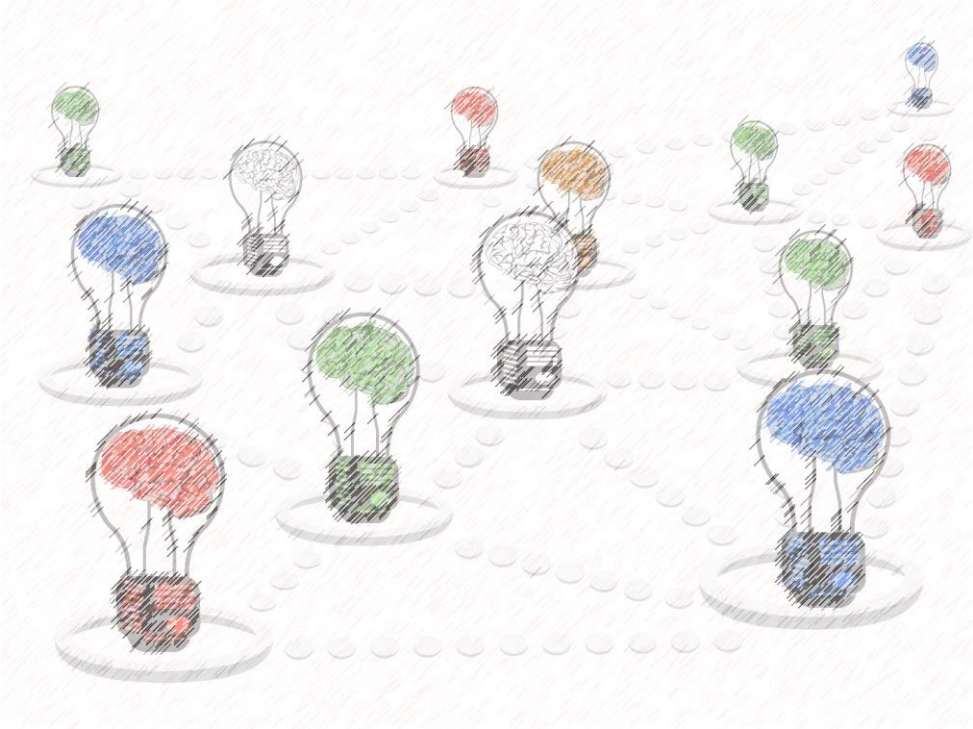
- Sharing of **experiment protocols** based on a common infrastructure
- Easier to reproduce, vary, and iterate on existing work

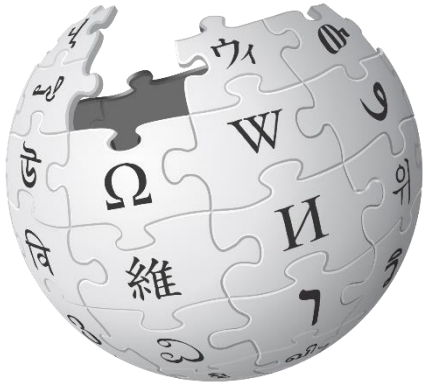
https://www.github.com/TurkServer/turkserver-meteor

Two examples of interesting social experiments

- Controlled, instrumented study of teamwork and collective intelligence
- A hundred people playing prisoner’s dilemma for one month of time

Teamwork and collective intelligence





Wikipedia

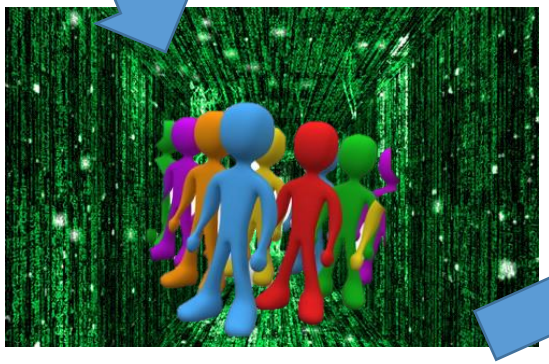
Open-source software

Libya crisis map, 2011

- Decentralized (or even distributed)
- Self-organized
- Complex problems

Crisis mapping

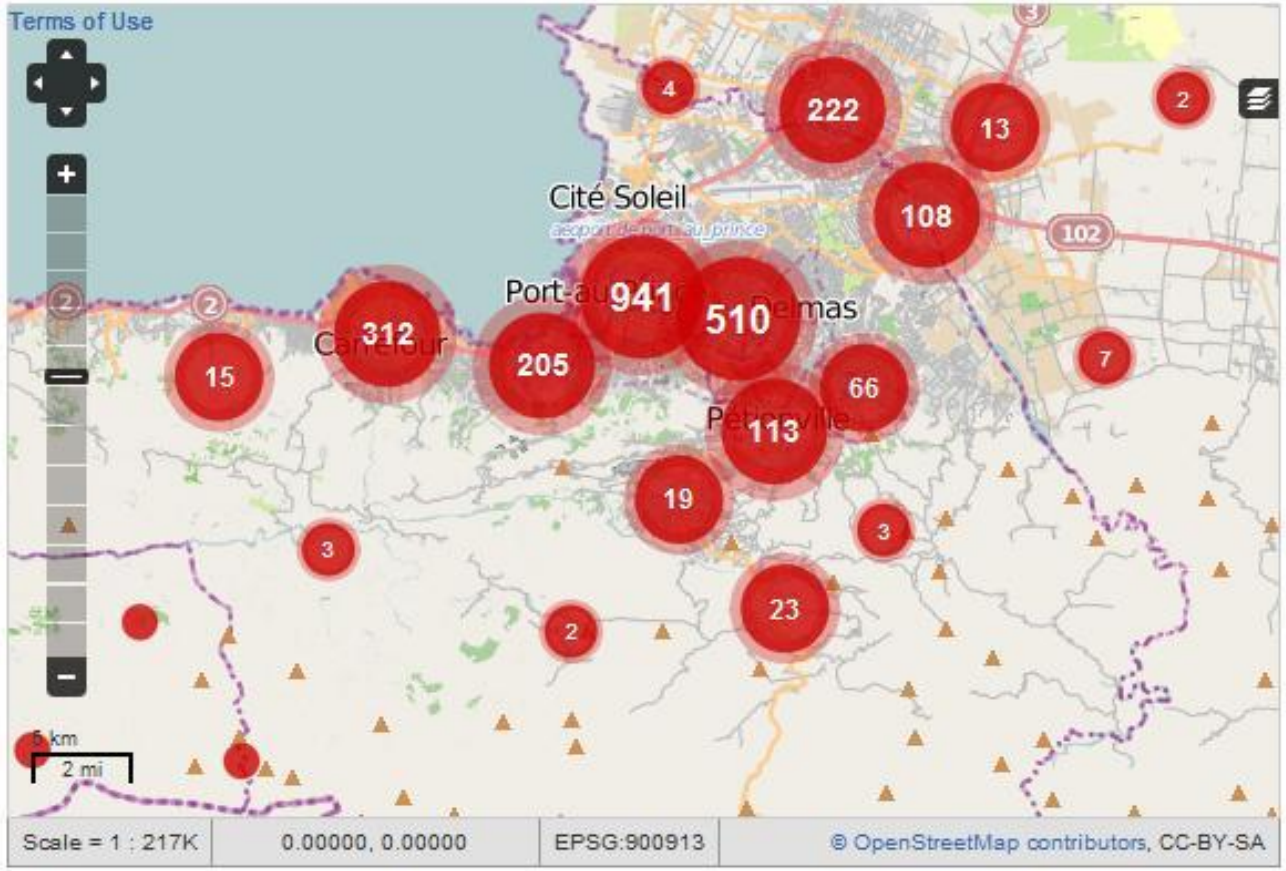
Haiti Earthquake, 2010



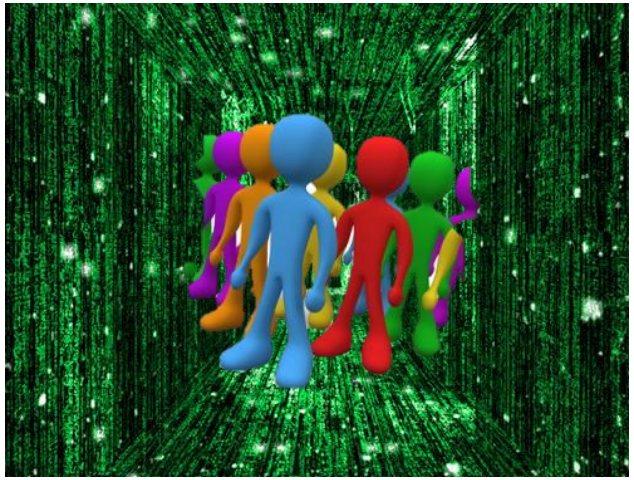
The Standby Task Force

We believe that digital volunteers are the future of humanitarian response

Haiti crisis map

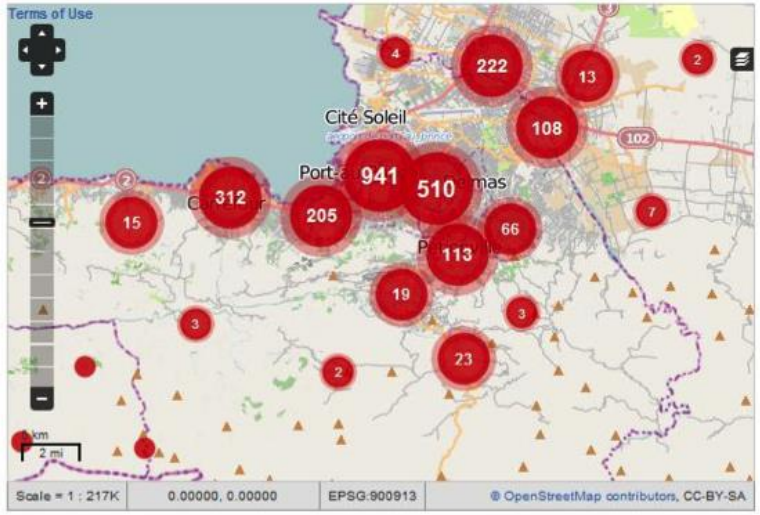


Crisis mapping:
A “model problem” for studying teamwork?



Online, distributed members

Crisis mapping tools

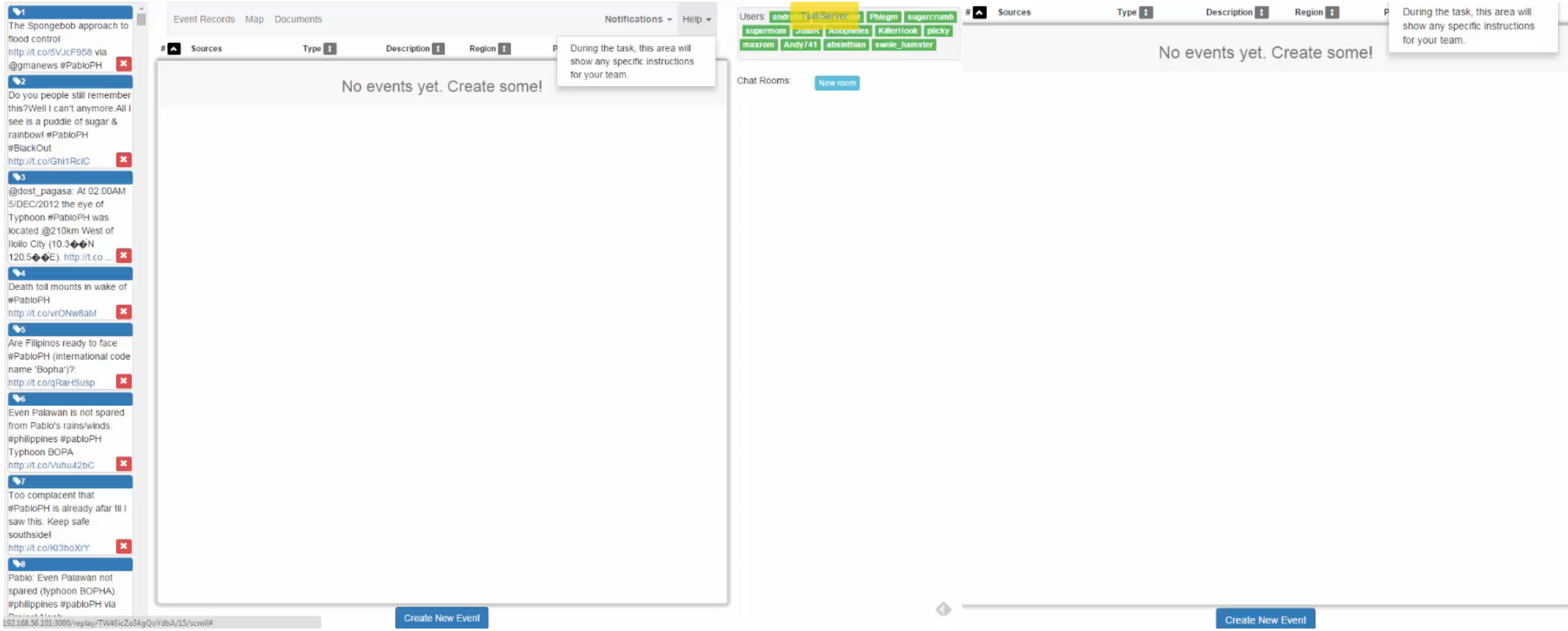


(Haiti 2010)

Complex output

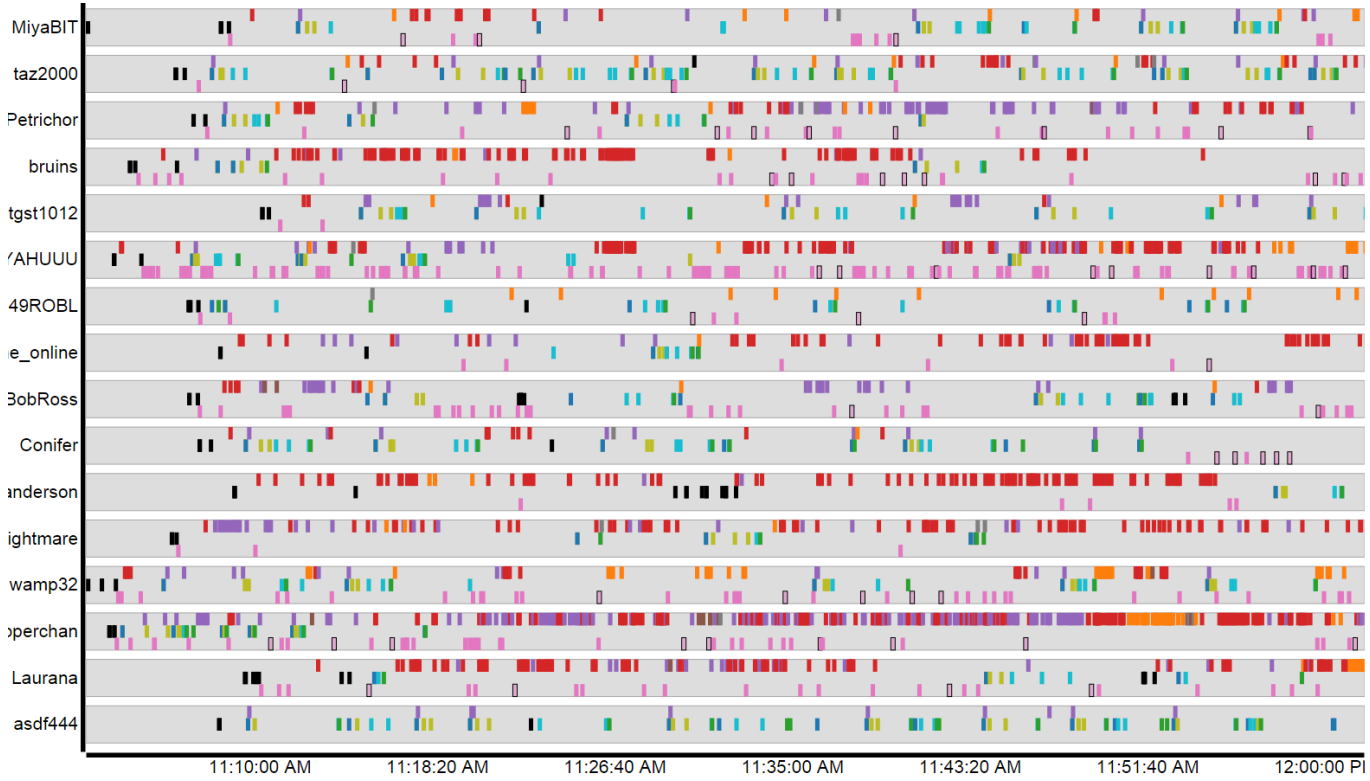
M., Mason, Suri, Watts (2016). PLoS ONE 11(4): e0153048

Studying teamwork and collective intelligence

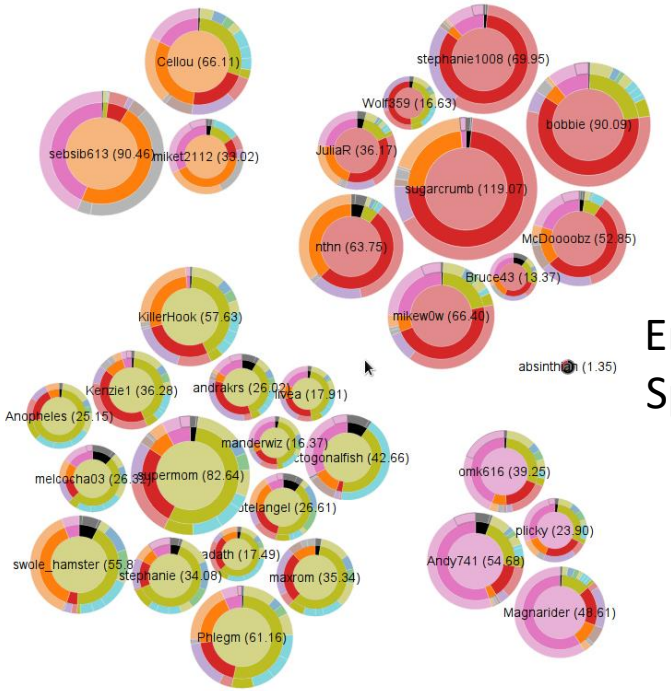


M., Mason, Suri, Watts (2016). PLoS ONE 11(4): e0153048

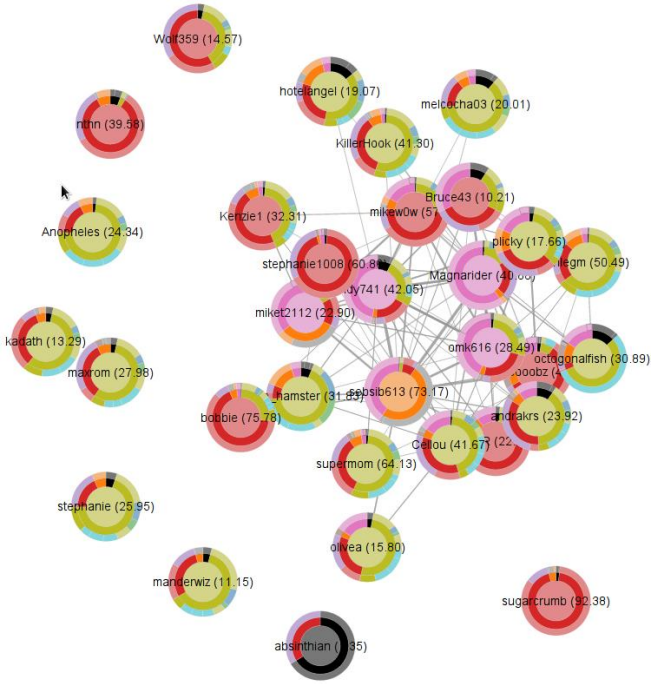
Fine-grained data instrumentation



Timeline of users and actions

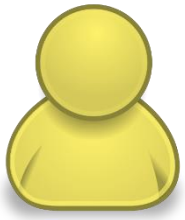


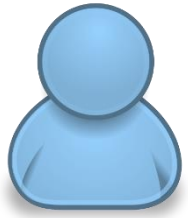
Emergence of Specialization



Communication Network

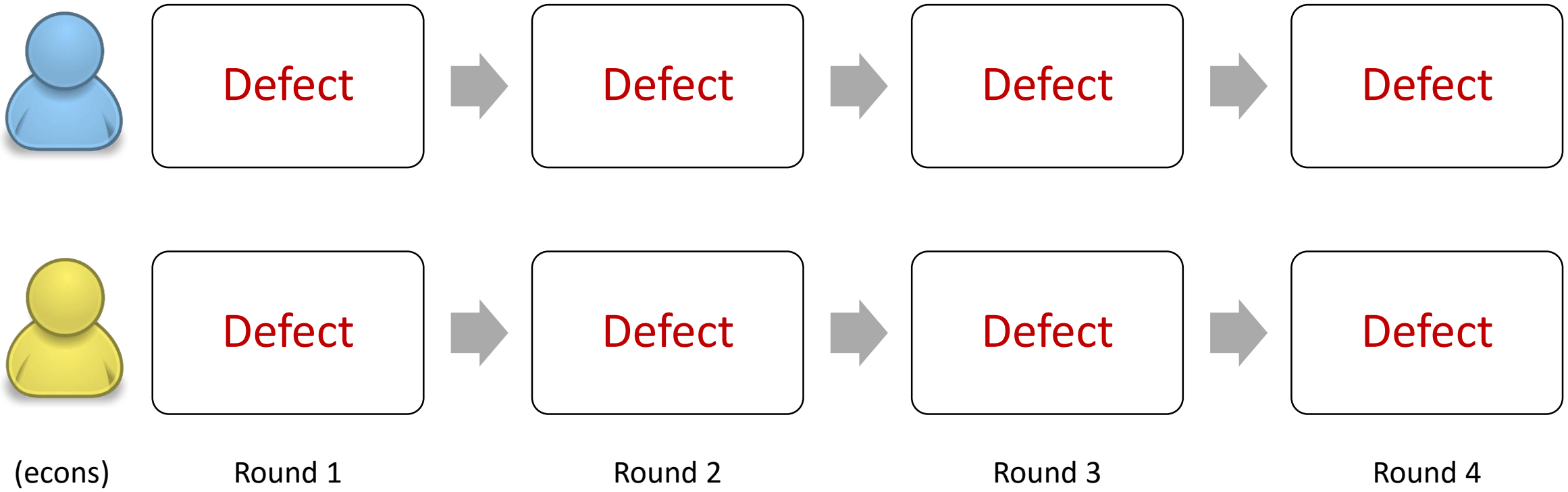
Prisoner’s Dilemma



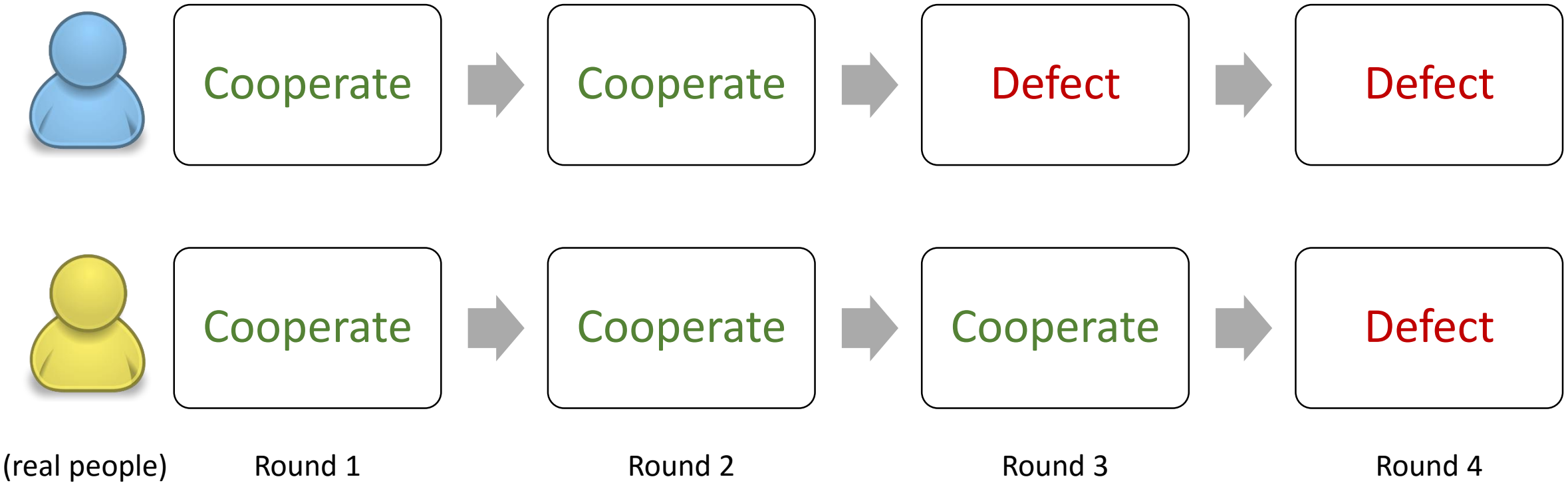


	Defect	Cooperate
Defect	**3**, *3*	**7**, *1*
Cooperate	**1**, *7*	**5**, *5*

Repeated Prisoner's Dilemma (*in theory*)



Repeated Prisoner's Dilemma (*in practice*)



see: **Selten and Stoecker** [1986]; **Andreoni and Miller** [1993]; **Dal Bo** [2005]; **Bereby-Meyer and Roth** [2006]; **Friedman and Oprea** [2012], **Embrey, Fréchette, and Yuksel** [2015]

Would cooperation unravel with experience?

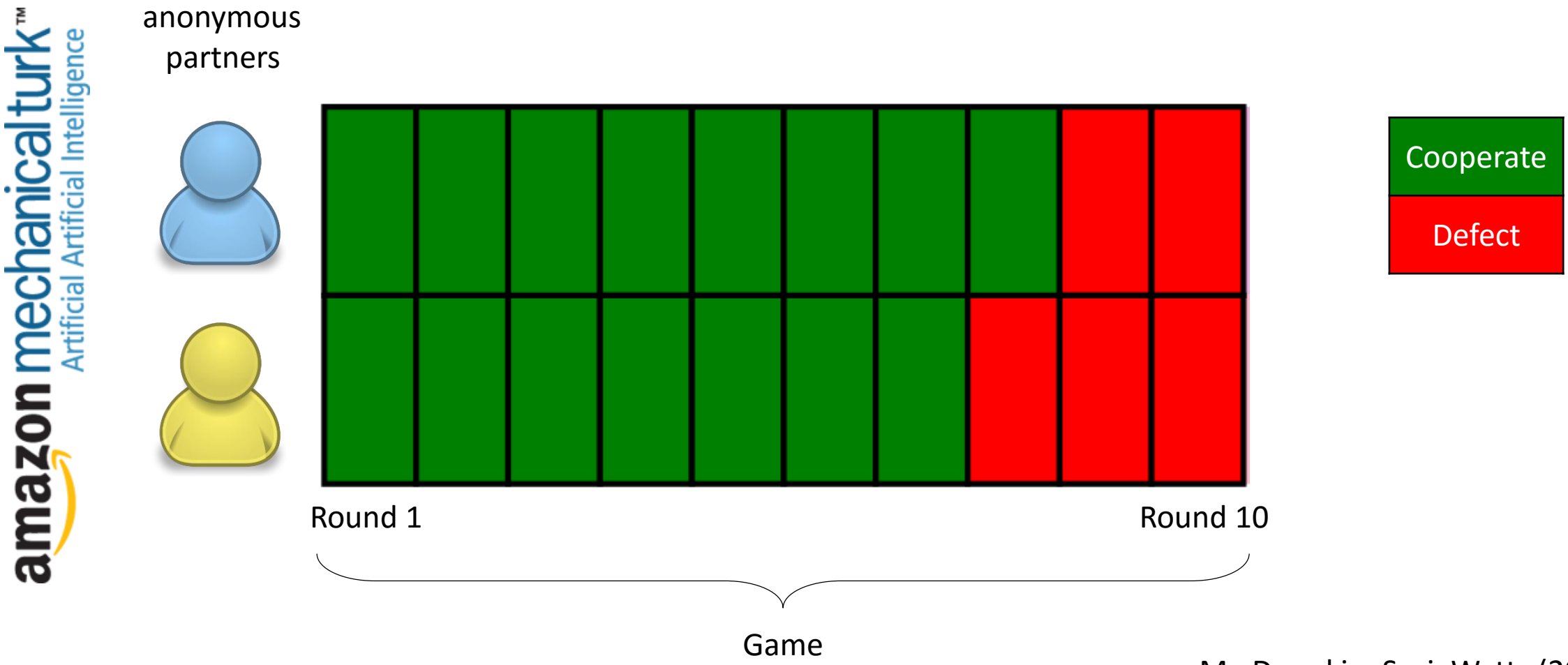
- *"… we conjecture that convergence to Nash would require in excess of 200 games of 10 rounds each." [Mason et al. 2014]*

- *"Although … unravelling is at work in all treatments, the process is slow enough that … it is not plausible to observe cooperation rates to decline to negligible levels in an amount of time that is reasonable to spend* ***in a laboratory****." [Embrey et al. 2015]*

A experimental study of cooperation <u>over time</u> **would:**

- (maybe) resolve conflict between theory and empirical data
- be closer to the real world

A very long prisoner’s dilemma experiment

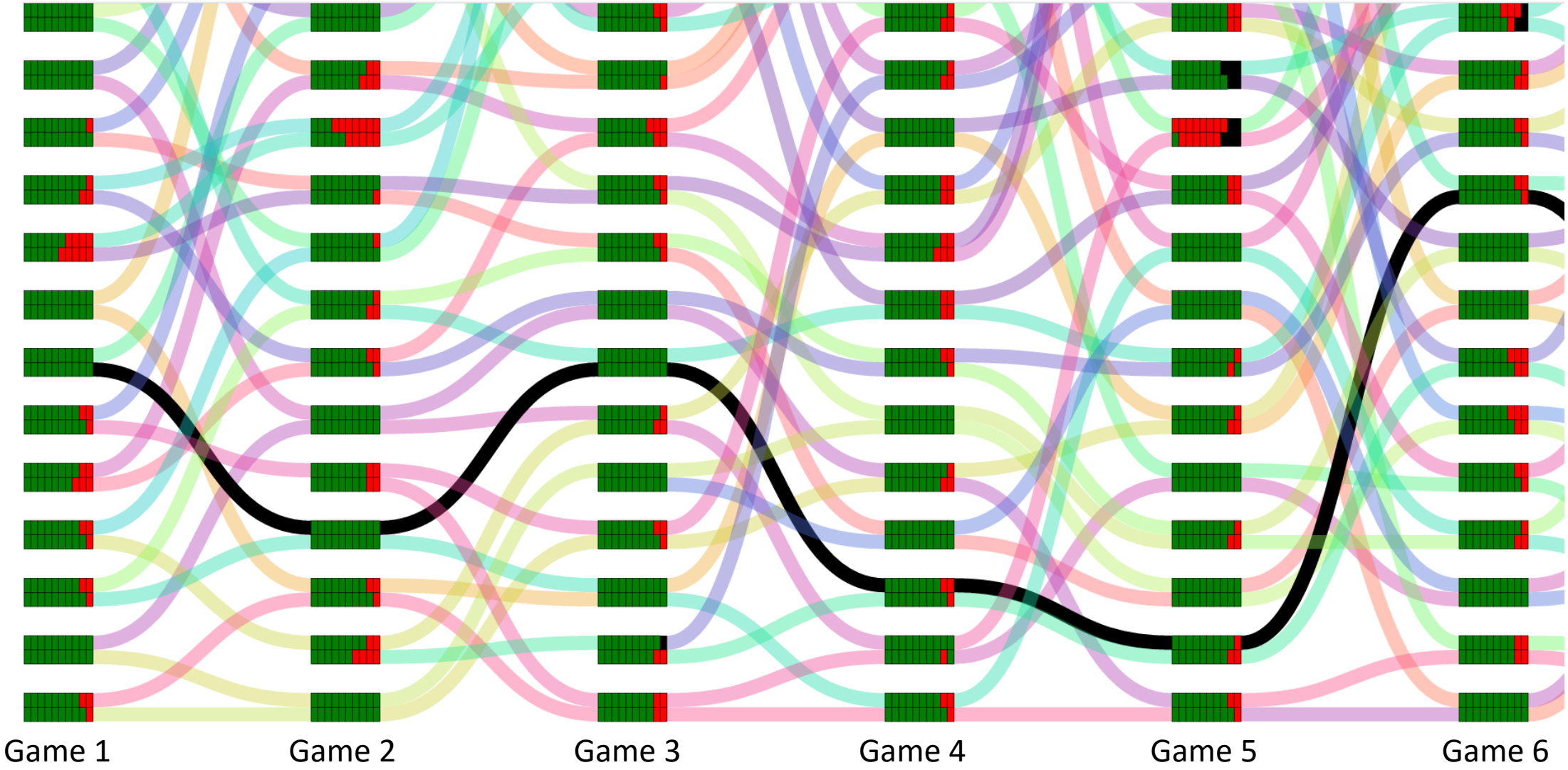


M., Dworkin, Suri, Watts (2016).

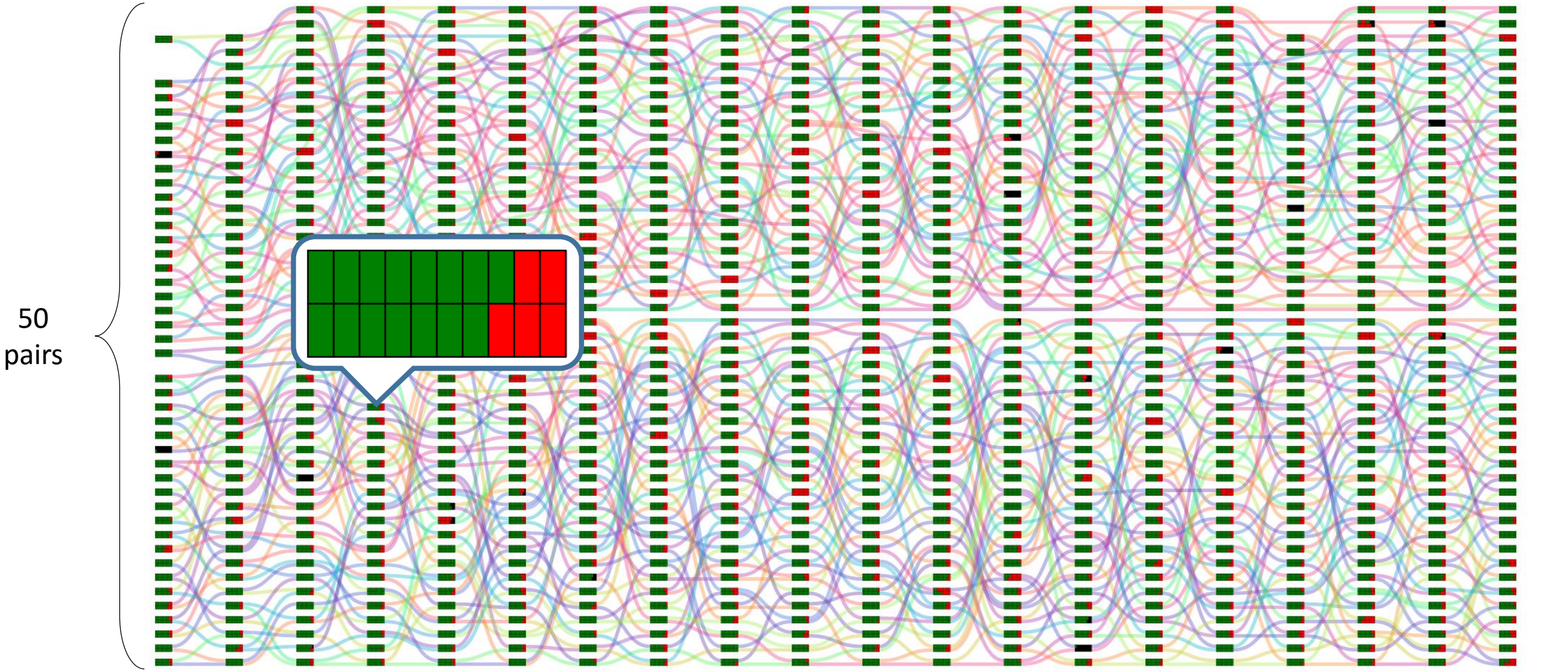
Random rematching across games

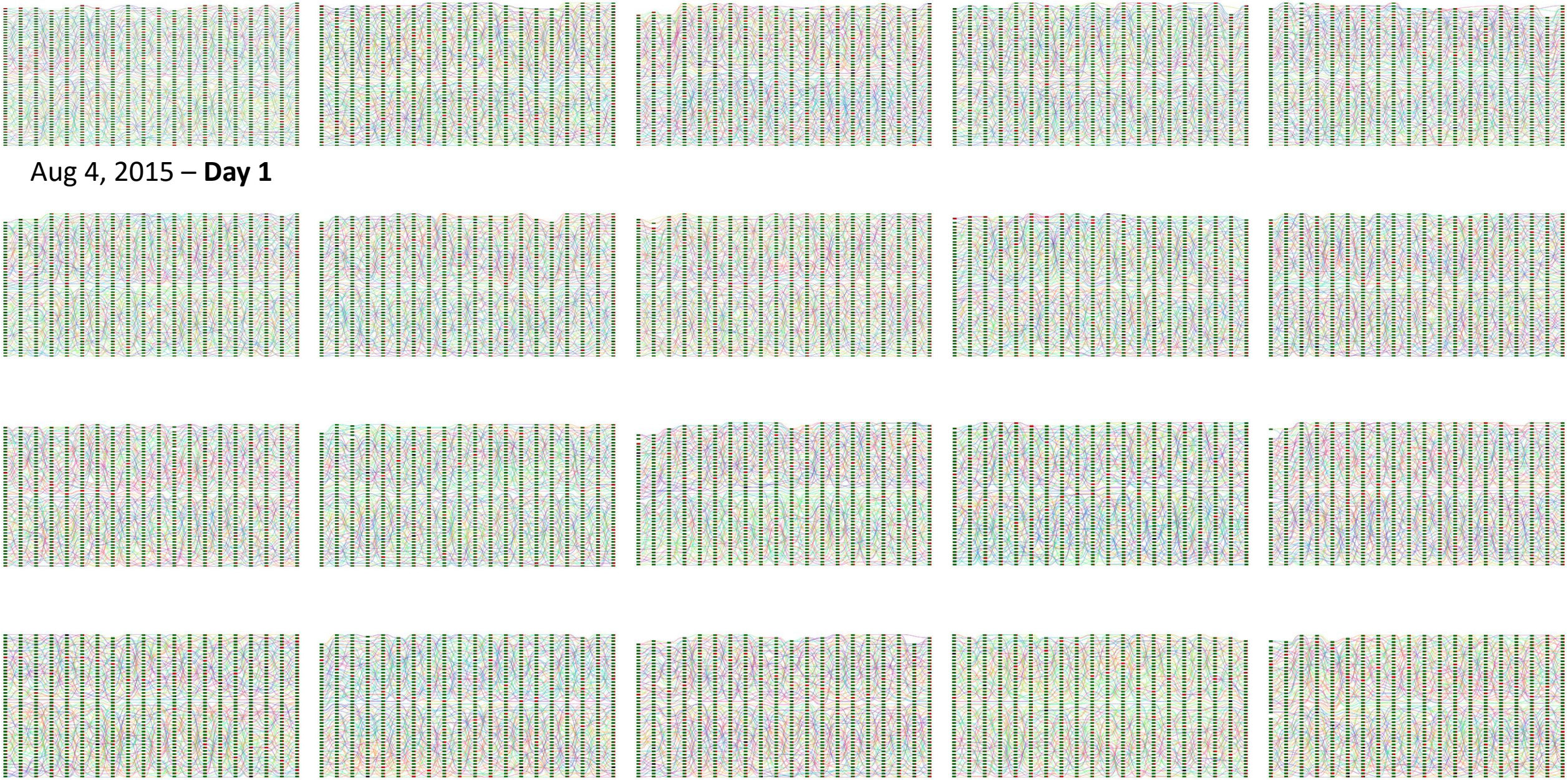
Game 1

Random rematching across games



20 games per day





Aug 4, 2015 – **Day 1**

Aug 31, 2015 – **Day 20**

Demo time!

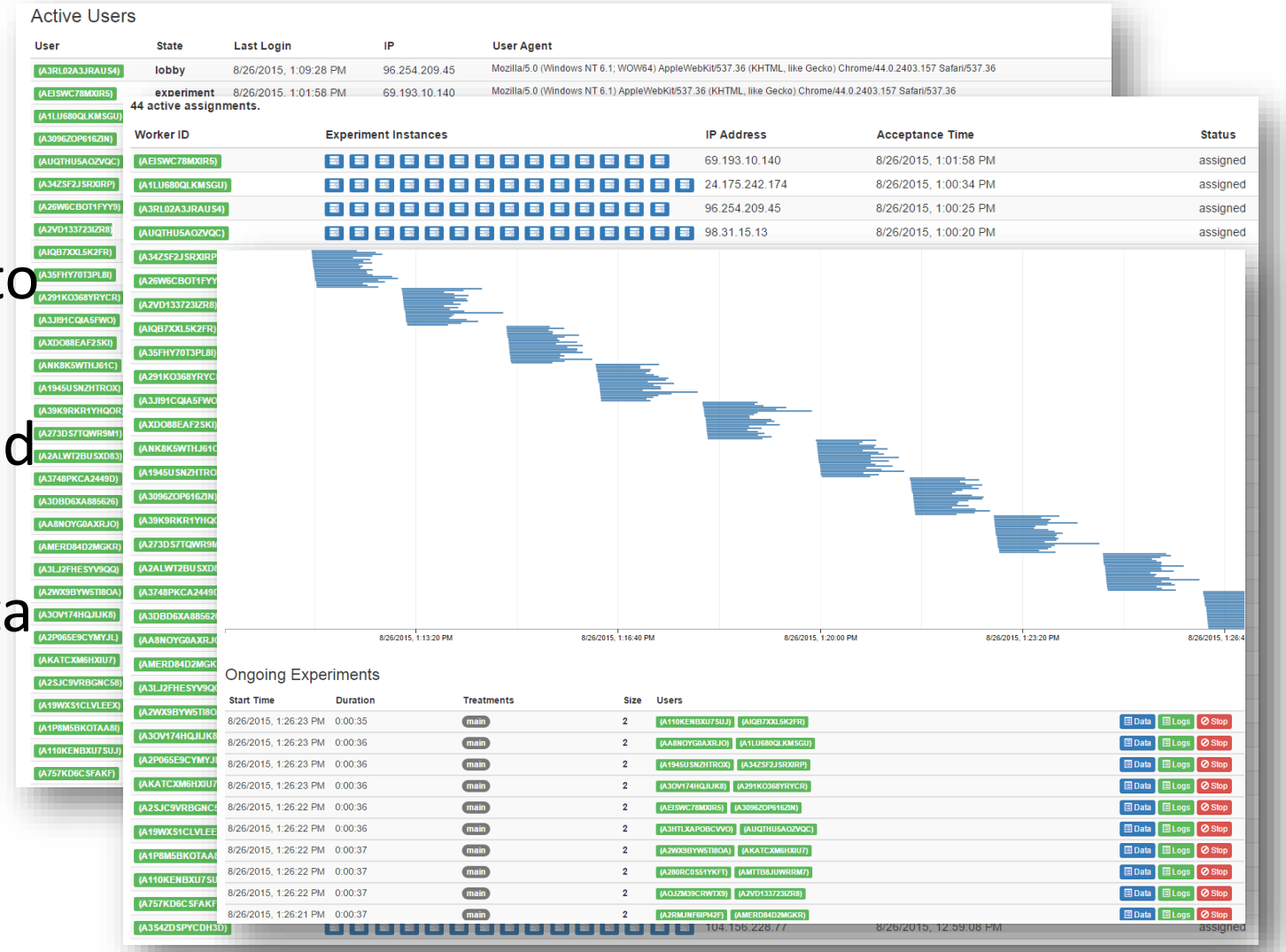
You, the esteemed audience, will play prisoner's dilemma with each other!

Navigate your browsers to:

http://turkserver.github.io

Web-based virtual lab console

- All connected users, their metadata and their state
- Participation history, ability to contact users in real time
- Live view of active worlds and progress
- Real-time view of logged data from any world



Random rematching, experimenter view

Current Lobby

Viewing lobby users in batch **pilot**

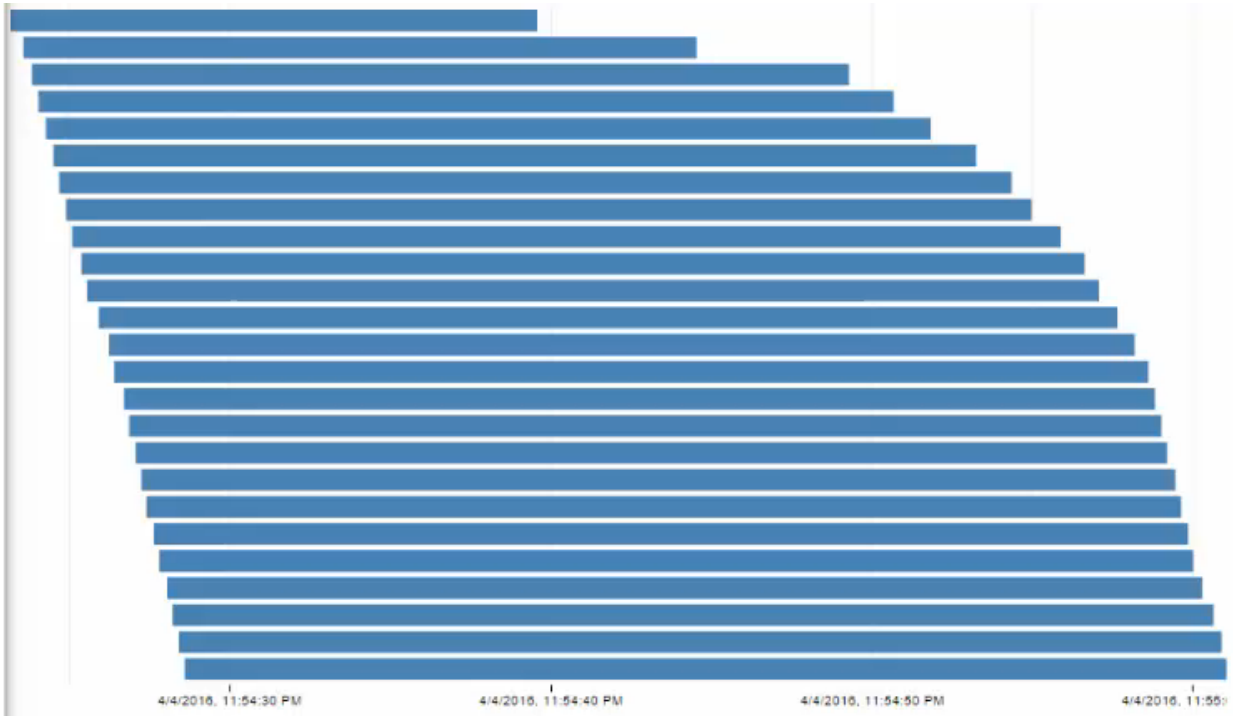
50 users currently in lobby.

50 ready users currently in lobby.

next-game | Trigger Lobby Event

User	Status
(SawQ:JXW9hGb5864a_Worker)	READY
(o66S2JWDwZ5LjTXhe_Worker)	READY
(3wq6JaQDtR2WzLayJ_Worker)	READY
(TjyhgiBeXsJJPDKoZ_Worker)	READY
(6jcFaFY3z85qRAJor_Worker)	READY
(BRK45vxFtqDhTEWHM_Worker)	READY
(KsHop3629CnuJ4YRN_Worker)	READY
(cwcRn2aMv56rDbkBK_Worker)	READY
(uv9k8LSYwR4oGpWvb_Worker)	READY
(kyPeaKE7jNzyRcbwJ_Worker)	READY
(NDHYtFokAztYJpWBo_Worker)	READY
(rqH6aRJtLS6Au4BNe_Worker)	READY
(ZkYvPjh3a3aJfcK3p_Worker)	READY
(gHh5jrvtTEBRaxGiW_Worker)	READY
(qYqqBmYNZ8eZXwmaK_Worker)	READY
(4ZKYy2MechbyJ9N4W_Worker)	READY
(NNxPzaiMqZqwJj7r_Worker)	READY
(iE5gCdW6G4FumemYL_Worker)	READY
(6d7bjq8AZdfM8WL4t_Worker)	READY
(tGM9deyD2ekJSwbD_Worker)	READY
(Kt6aREBHfoh5hD7s_Worker)	READY
(HjGLWrMQBvyJY47rZ_Worker)	READY
(wKnNtZ7GfyHfWAnJb_Worker)	READY
(nfQDdC62dpa9pC6At_Worker)	READY
(Twsfu8ZPy8l4hWcP9_Worker)	READY
(XJEbJ5HbGmQRRZLoH_Worker)	READY
(v8ufMPJF3uNW5tdtG_Worker)	READY

FvdLgx3xB2LKSLRat
Username:
(none)
WorkerId:
TjyhgiBeXsJJPDKoZ_Worker
Last login:
4/4/2016, 11:53:05 PM *from* 127.0.0.1;
State:
lobby
Contact:
Connected



Ongoing Experiments

Start Time	Duration	Treatments	Size	Users

Completed Experiments

Start Time	Duration	Treatments	Size	Users	
4/4/2016, 11:54:28 PM	0:00:32	main	2	(p2Ify7TD9yXWRERMko_Worker) (6FP4BeJJHoumHfJTt_Worker)	Logs
4/4/2016, 11:54:28 PM	0:00:32	main	2	(R3EwQJY55GXjPevvy_Worker) (EPKJwE3pzaveQ4ZPx_Worker)	Logs
4/4/2016, 11:54:28 PM	0:00:32	main	2	(AzP3pLAt3gaKACqXq_Worker) (56c7Clb4kwFrEQur6_Worker)	Logs
4/4/2016, 11:54:28 PM	0:00:32	main	2	(MGMQ25Fz9ZrCwPMCq_Worker) (M3fgJ8hA6r9XvgapB_Worker)	Logs
4/4/2016, 11:54:27 PM	0:00:32	main	2	(f4JJHyF48Q4XfMonq_Worker) (ixaNr9t5mZrFNpuxy_Worker)	Logs
4/4/2016, 11:54:27 PM	0:00:32	main	2	(LuuHGaNqp4ZTJbYr_Worker) (kbHG82M7MxFz49TfW_Worker)	Logs
4/4/2016, 11:54:27 PM	0:00:32	main	2	(KoaRQkWxnqR5fZ2qi_Worker) (xL7kKvq6Y4LfMAskt_Worker)	Logs

Real-time interaction among 100 people

Lobby

100 people

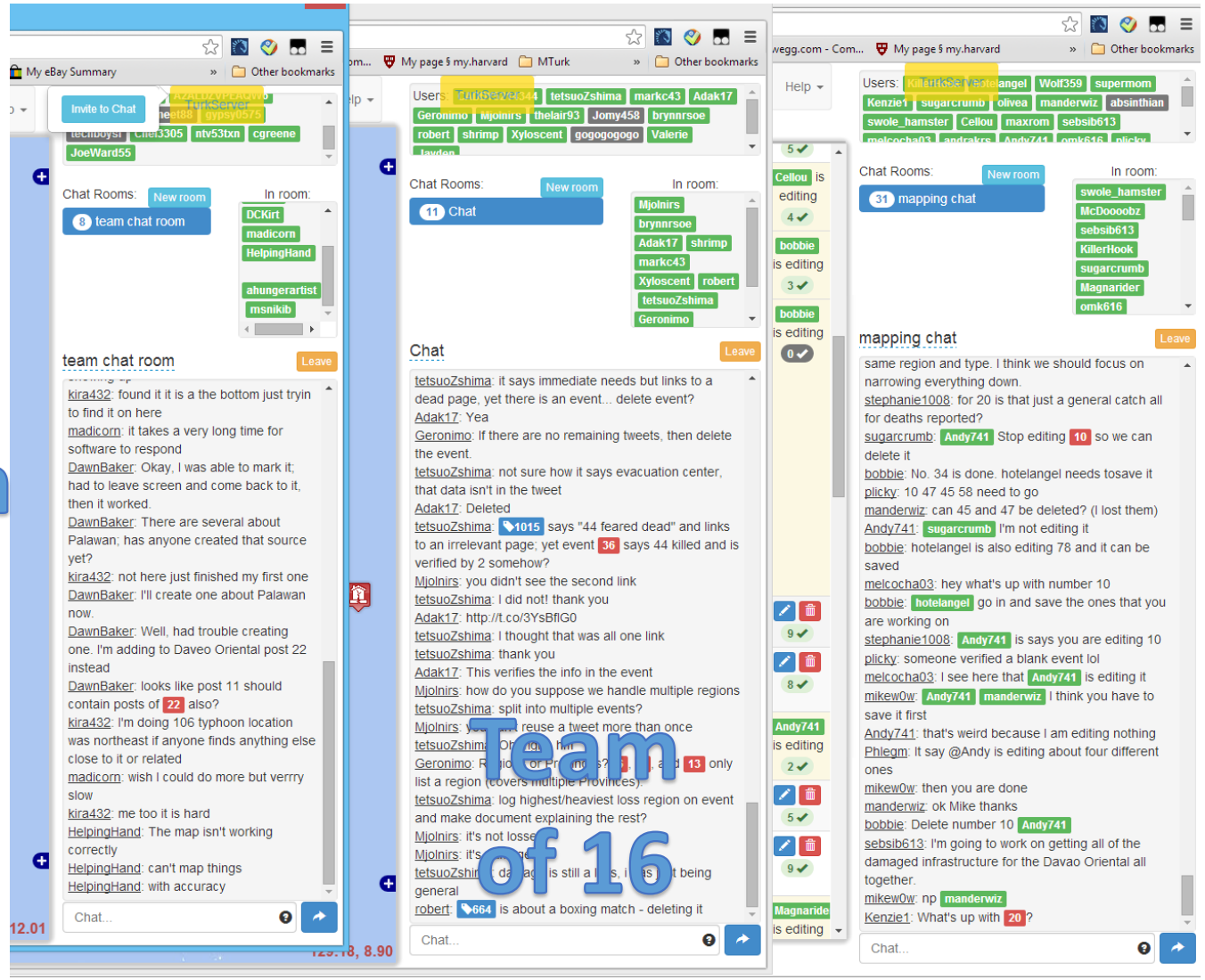
16

32

Start Time	Duration	Treatments	Size	Users	
8/13/2014 2:21:50 PM	0:05:50	parallel_worlds	8	ntv53txn, DawnBaker, langford, kira432, Tenkei, madicorn, DCKirt, techboysf	Watch, Logs, Stop
8/13/2014 2:20:33 PM	0:07:07	group_1, parallel_worlds	1	eolive	Watch, Logs, Stop
8/13/2014 2:17:35 PM	0:10:05	group_1, parallel_worlds	1	gms5002	Watch, Logs, Stop
8/13/2014 2:16:58 PM	0:10:41	group_2, parallel_worlds	2	Jennifer, Spyle07	Watch, Logs, Stop
8/13/2014 2:12:05 PM	0:15:35	group_4, parallel_worlds	4	b0nk444, Nicks7, mrwilliams, mikejamo	Watch, Logs, Stop
8/13/2014 2:11:35 PM	0:16:04	group_2, parallel_worlds	2	plrs199, bjones76nc	Watch, Logs, Stop
8/13/2014 2:11:35 PM	0:16:05	group_1, parallel_worlds	1	stsers	Watch, Logs, Stop
8/13/2014 2:11:35 PM	0:16:05	group_1, parallel_worlds	1	CatsMeow	Watch, Logs, Stop
8/13/2014 2:11:32 PM	0:16:07	group_16, parallel_worlds	15	shrimp, Mjolnirs, Valerie, brynnrsoe, Adak17, markc43, CADDY5214344, tetsuoZshima, Jayden, gogogogogo, Xyloscent, robert, thelair93, Geronimo, Jomy458	Watch, Logs, Stop
8/13/2014 2:11:32 PM	0:16:07	group_8, parallel_worlds	8	Andrewmatt, EricD35, JohnRocker, Presto, Nathan, Kelby19, keezay, ryanawall	Watch, Logs, Stop
8/13/2014 2:11:32 PM	0:16:08	group_1, parallel_worlds	1	sdfnioagji45	Watch, Logs, Stop
8/13/2014 2:11:32 PM	0:16:08	group_1, parallel_worlds	1	Klasens	Watch, Logs, Stop
8/13/2014 2:11:32 PM	0:16:08	group_4, parallel_worlds	4	Bandista, Rhelton117, ccarman, code3banker	Watch, Logs, Stop
8/13/2014 2:11:32 PM	0:16:08	group_2, parallel_worlds	2	mrb123, gaviidae	Watch, Logs, Stop
8/13/2014 2:11:31 PM	0:16:08	group_2, parallel_worlds	2	arsi741, limcid	Watch, Logs, Stop
8/13/2014 2:11:31 PM	0:16:08	group_32, parallel_worlds	32	andrakrs, McDoooobz, sugarcrumb, Phlegm, supermom, JuliaR, Anopheles, KillerHook, plicky, maxrom, Andy741, absinthian, swole_hamster, olivea, bobbie, nthn, sebsib613, Magnarider, hotelangel, omk616, Cellou, miket2112, mikew0w, octogonalfish, Bruce43, kadath, stephanie1008, Wolf359, Kenzie1, manderwiz, melcocha03, stephanie	Watch, Logs, Stop
8/13/2014 2:11:31 PM	0:16:09	group_1, parallel_worlds	1	mrmiyagisr	Watch, Logs, Stop

Simultaneous one-way mirror on multiple worlds

Team of 8

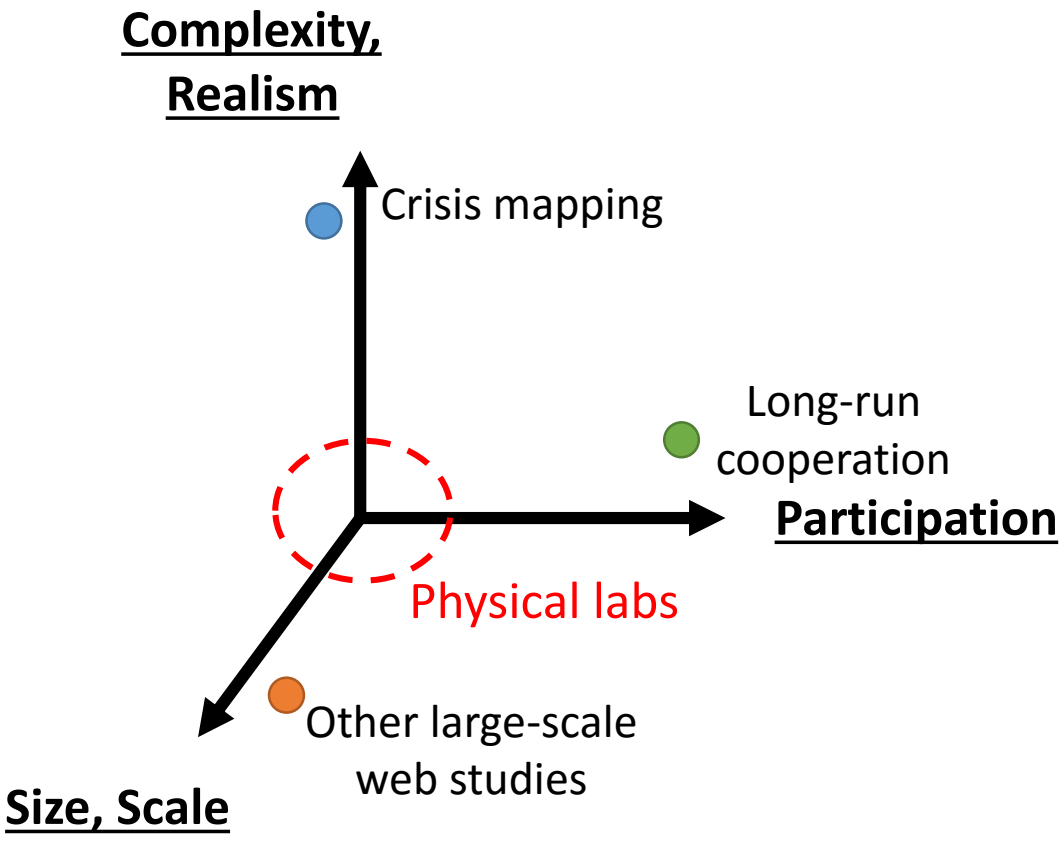


Team of 16

Team of 32

Actual Screenshot, Aug. 2014

Opportunities in the online lab



Pushing these boundaries can answer novel & otherwise inaccessible research questions.

- Highly instrumented group interaction
- "Longitudinal" studies of social interaction over time

- Mobile devices and sensors
- Algorithmic and computational interventions

First Half: Takeaways

TurkServer aims for two main goals for the future online or “virtual” lab:

- It allows us to answer **novel scientific questions** by making experiments more powerful
- It makes experiments easier to **build, share and iterate** upon

https://github.com/TurkServer/turkserver-meteor

Part 2:
The nitty-gritty of doing online social experiments

- Web programming and architecture of TurkServer
- Designing experiments and logistics of using crowd workers
- Additional information: http://turkserver.readthedocs.io

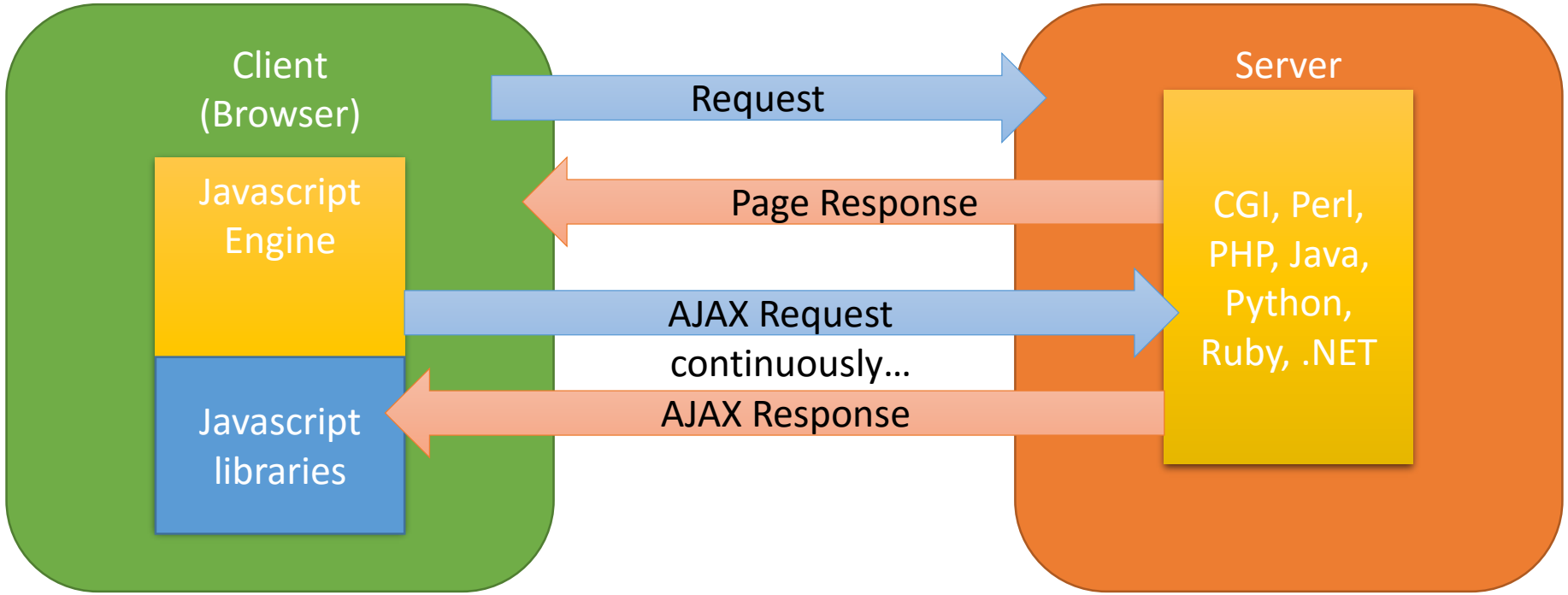
Prerequisite: The Experimental Method

- Why are experiments good for establishing causality?
  - Why is randomization important?
- When to use experiments vs. other methods of causal inference?
- How to operationalize a research question as an experiment?
- How will I analyze my data to establish causality?

(*very important, but won't be covered in this tutorial*)

A simplified history of web programming

The web is now the ultimate application platform…

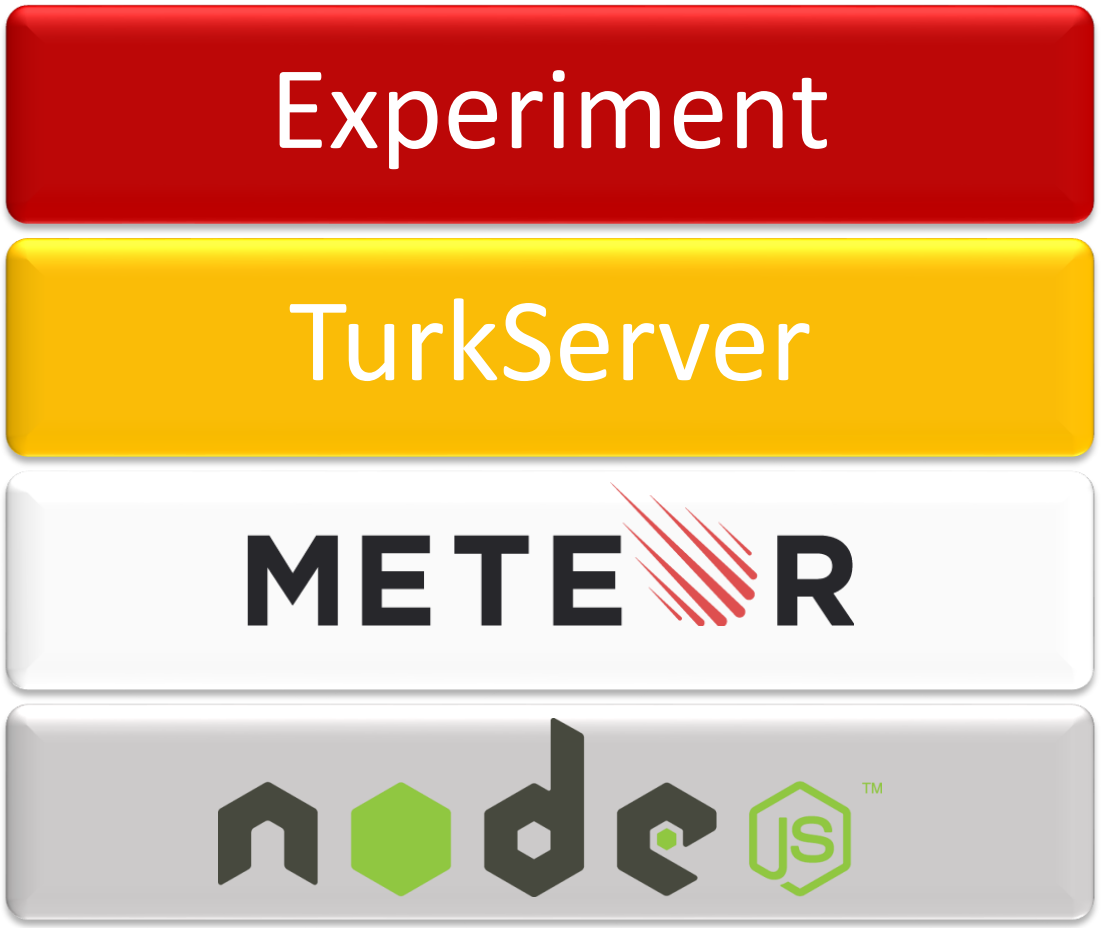


… and it’s quite a mess.

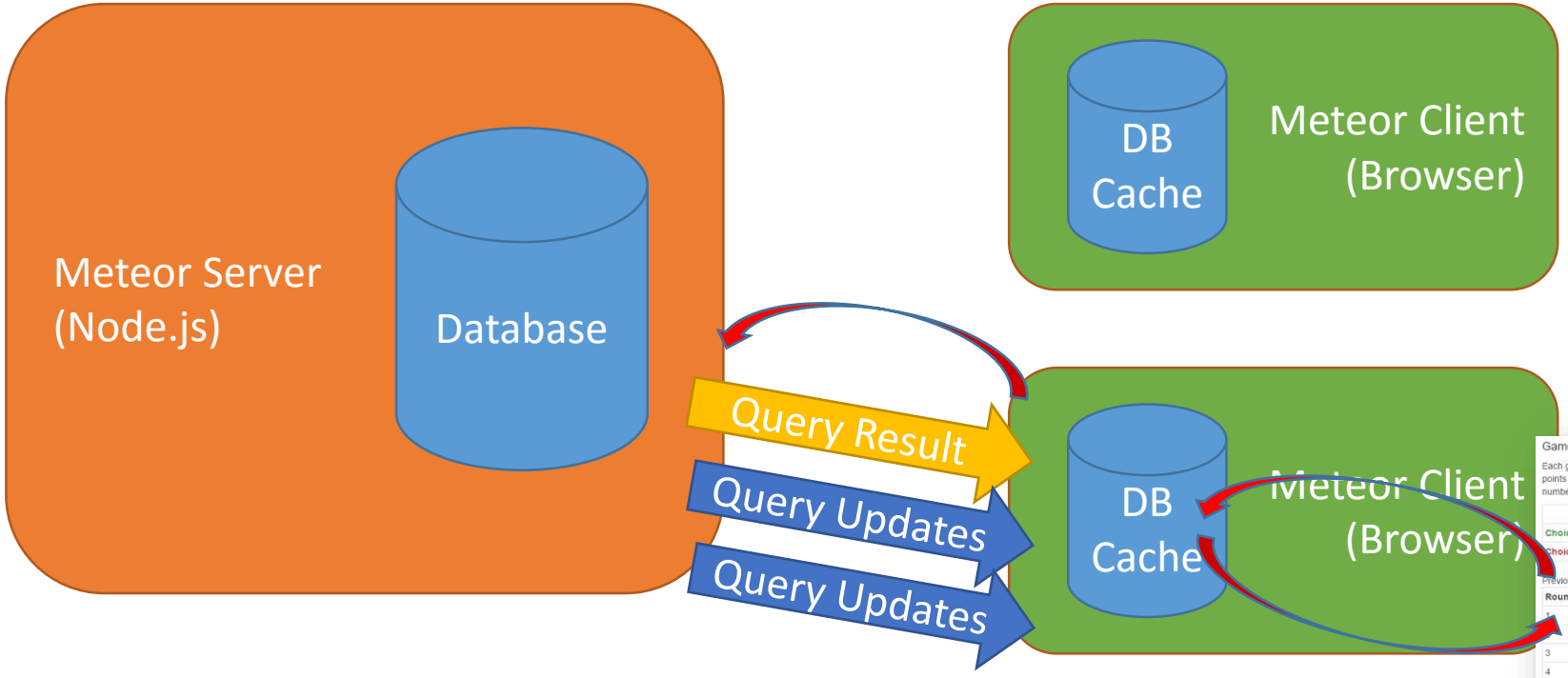
TurkServer is built on METEOR (www.meteor.com)

Why Meteor?

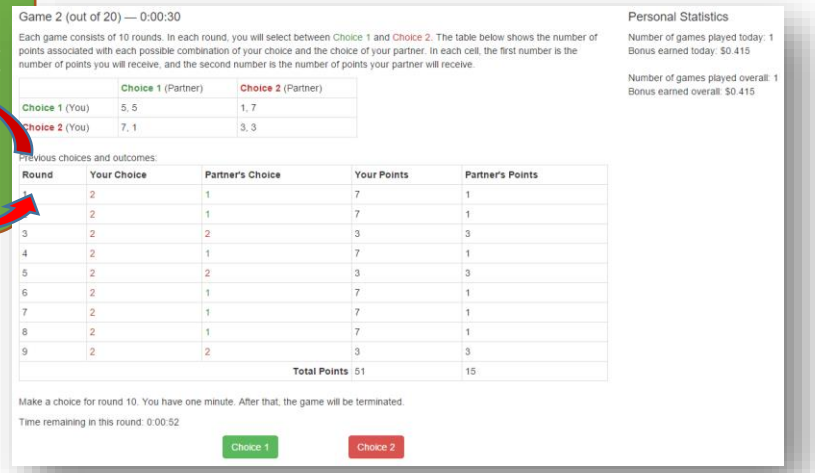
- One language (Javascript)
- Simpler abstractions for real-time interaction with the server or among multiple clients
- Easy hosting and deployment
- Open-source, well-documented, with an active community



Simplified architecture of METEOR (www.meteor.com)



Reactive UI: DOM updates automatically as data changes



Game 2 (out of 20) — 0:00:30

Each game consists of 10 rounds. In each round, you will select between Choice 1 and Choice 2. The table below shows the number of points associated with each possible combination of your choice and the choice of your partner. In each cell, the first number is the number of points you will receive, and the second number is the number of points your partner will receive.

	Choice 1 (Partner)	Choice 2 (Partner)
Choice 1 (You)	5, 5	1, 7
Choice 2 (You)	7, 1	3, 3

Previous choices and outcomes:

Round	Your Choice	Partner's Choice	Your Points	Partner's Points
1	2	1	7	1
2	2	1	7	1
3	2	2	3	3
4	2	1	7	1
5	2	2	3	3
6	2	1	7	1
7	2	1	7	1
8	2	1	7	1
9	2	2	3	3
		Total Points	51	15

Make a choice for round 10. You have one minute. After that, the game will be terminated.

Time remaining in this round: 0:00:52

Choice 1 Choice 2

Personal Statistics

Number of games played today: 1
Bonus earned today: $0.415

Number of games played overall: 1
Bonus earned overall: $0.415

Live queries: most recent set of data is always on the client

Fast prototyping with Meteor: Chat example

```
Messages = new Mongo.Collection("chat");

Meteor.publish("chatData", function() {
    return Messages.find();
});
```

Server code

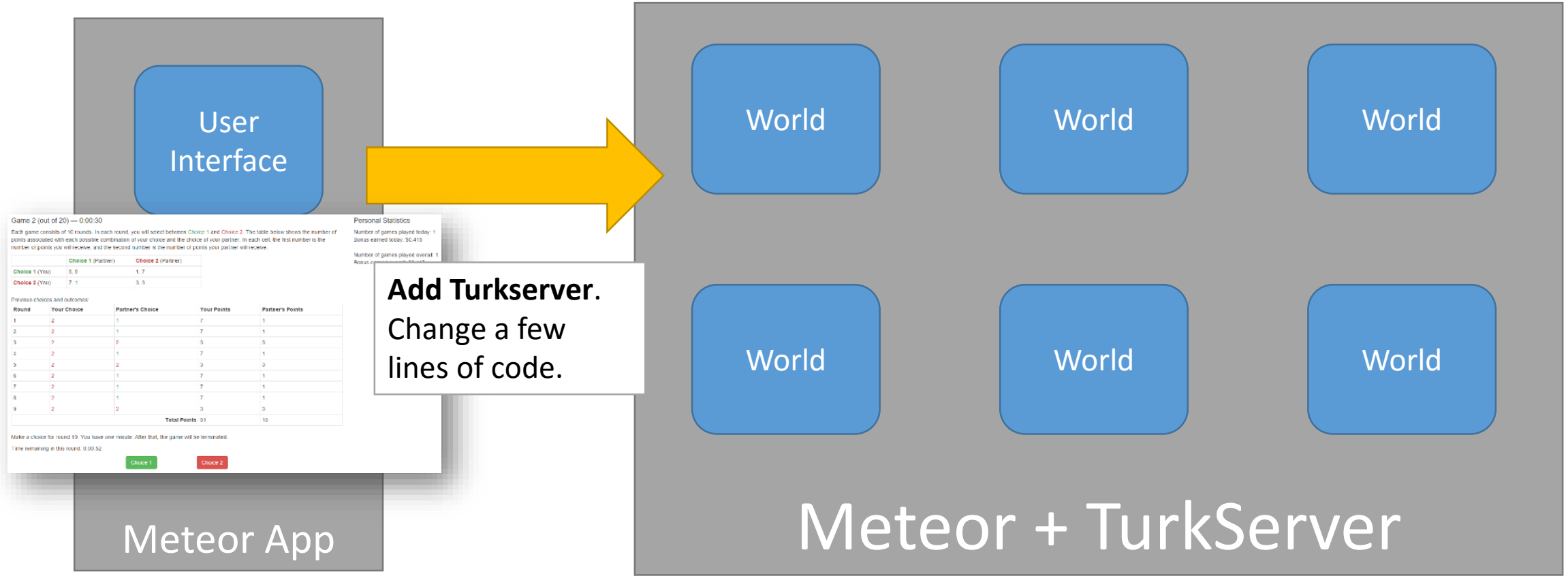
```
Meteor.subscribe("chatData");

Template.chat.helpers({
    messages: function() {
        return Messages.find({},
        sort: {timestamp: -1}});
```

EricD35: "Tree Dents Pool".....NOOOOOOOOO! The desctruction!
Presto: lol EricD35 priorities yu know
Andrewmatt: haha
EricD35: The summer BBQ is ruined!
Andrewmatt: At least the worlds largest croc survived the storm

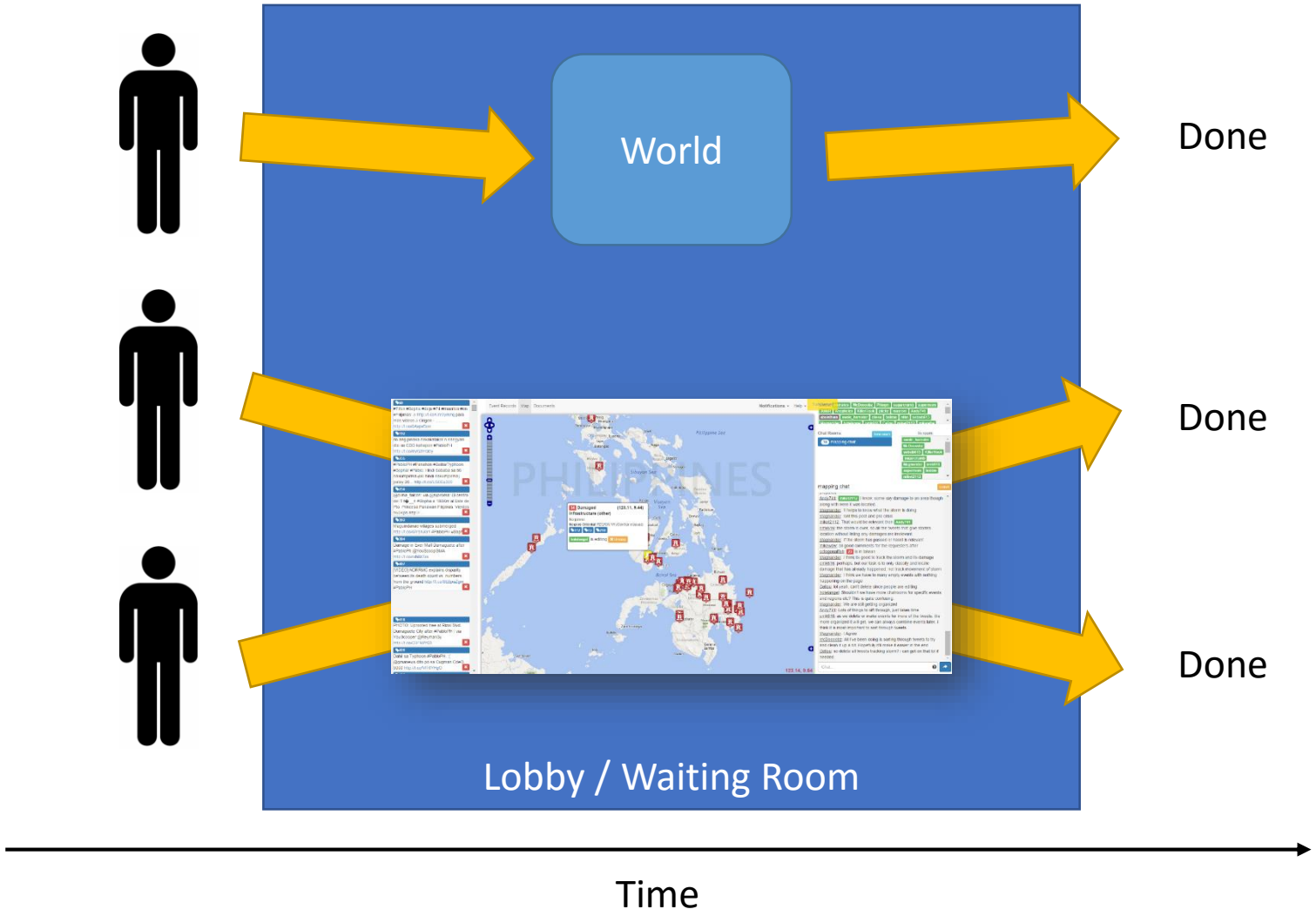
```
e="chat">

essages}}
    <li>{{username}}: {{text}}</li>
    {{/each}}
    </ul>
</template>
```

Client code

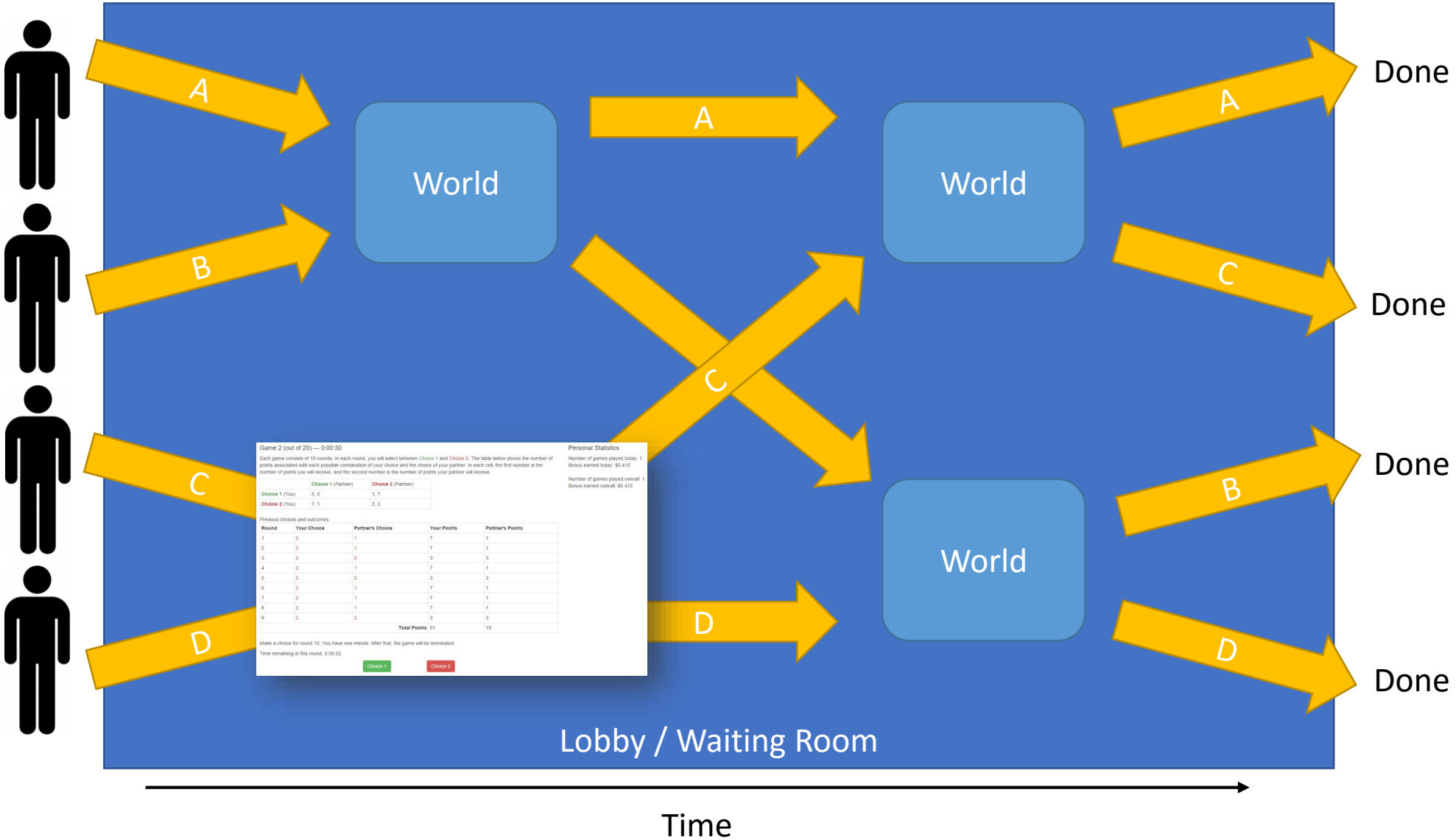
From a prototype to multiple worlds



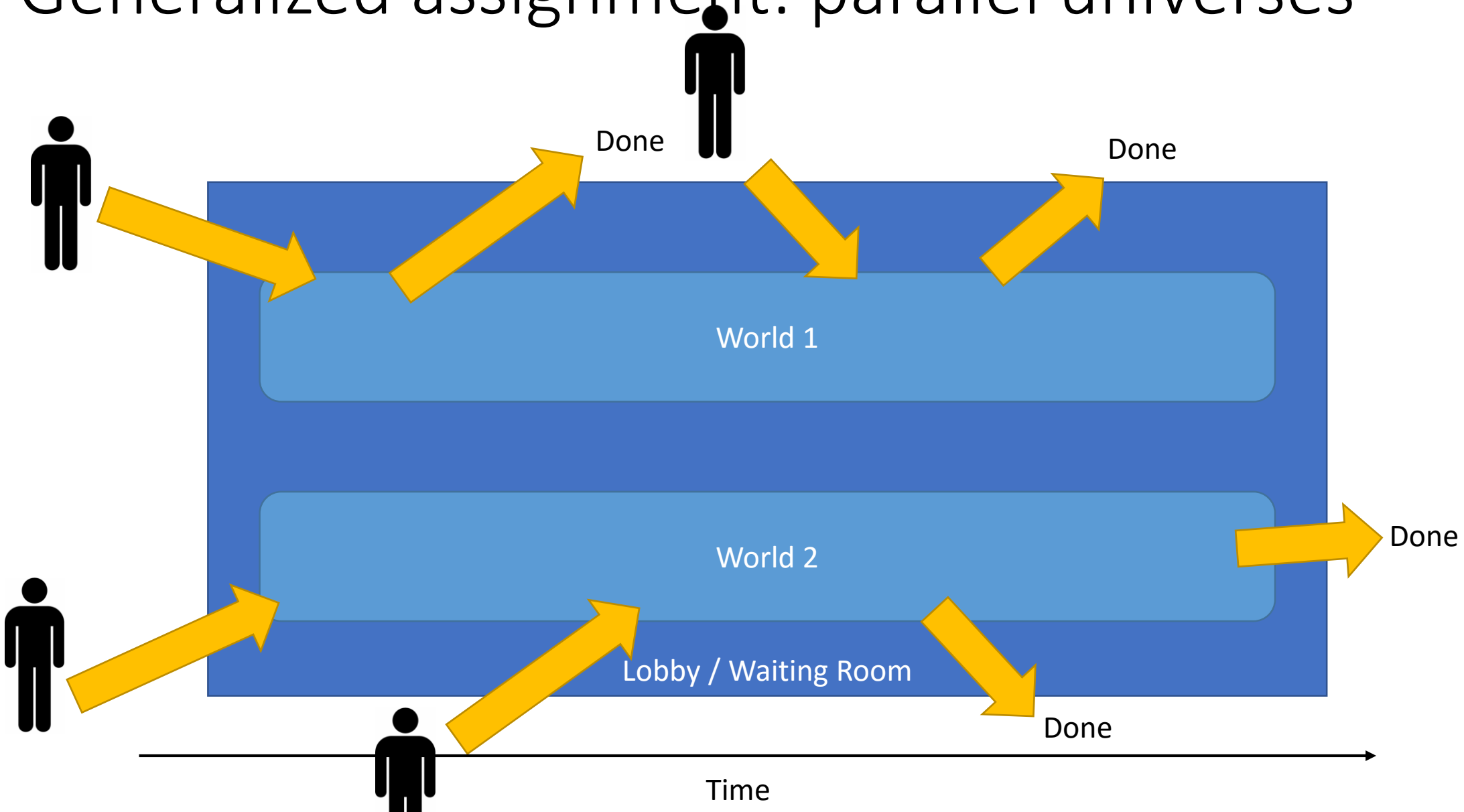
Generalized assignment mechanism



Generalized assignment: rematching



Generalized assignment: parallel universes



Typical Workflow for using TurkServer

1. Find a **good research question**, suitable for an experiment
2. **Prototype** your experiment design in a **standalone Meteor app**, for a single unit of interaction (e.g. one team or one pair).
   - *Use Meteor's fast development capabilities to quickly iterate on feasibility*
3. **Add TurkServer** to your project; set up assignment of users to worlds; think through logistics of running the experiment
4. **Test, debug, and pilot;** then test some more
5. **Run the experiment**, analyze the data, write the paper
6. **Share** your experiment protocol via open-source software

Designing experiments with crowdsourced participants

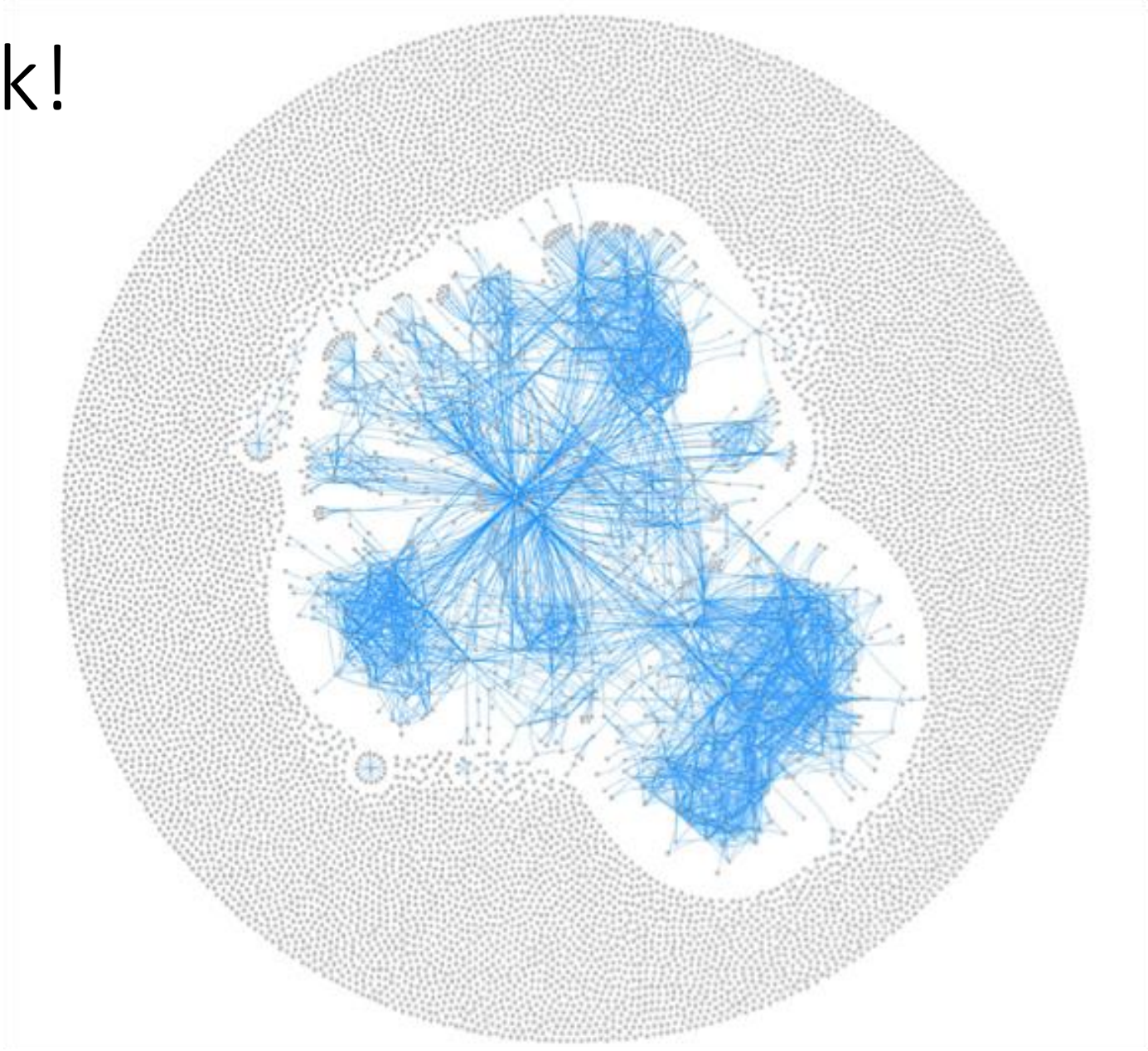
Best practices, things to consider

The crowd is a network!

- 2-week study of MTurk workers and their connections to each other

- 10,354 respondents
- 5,268 edges added

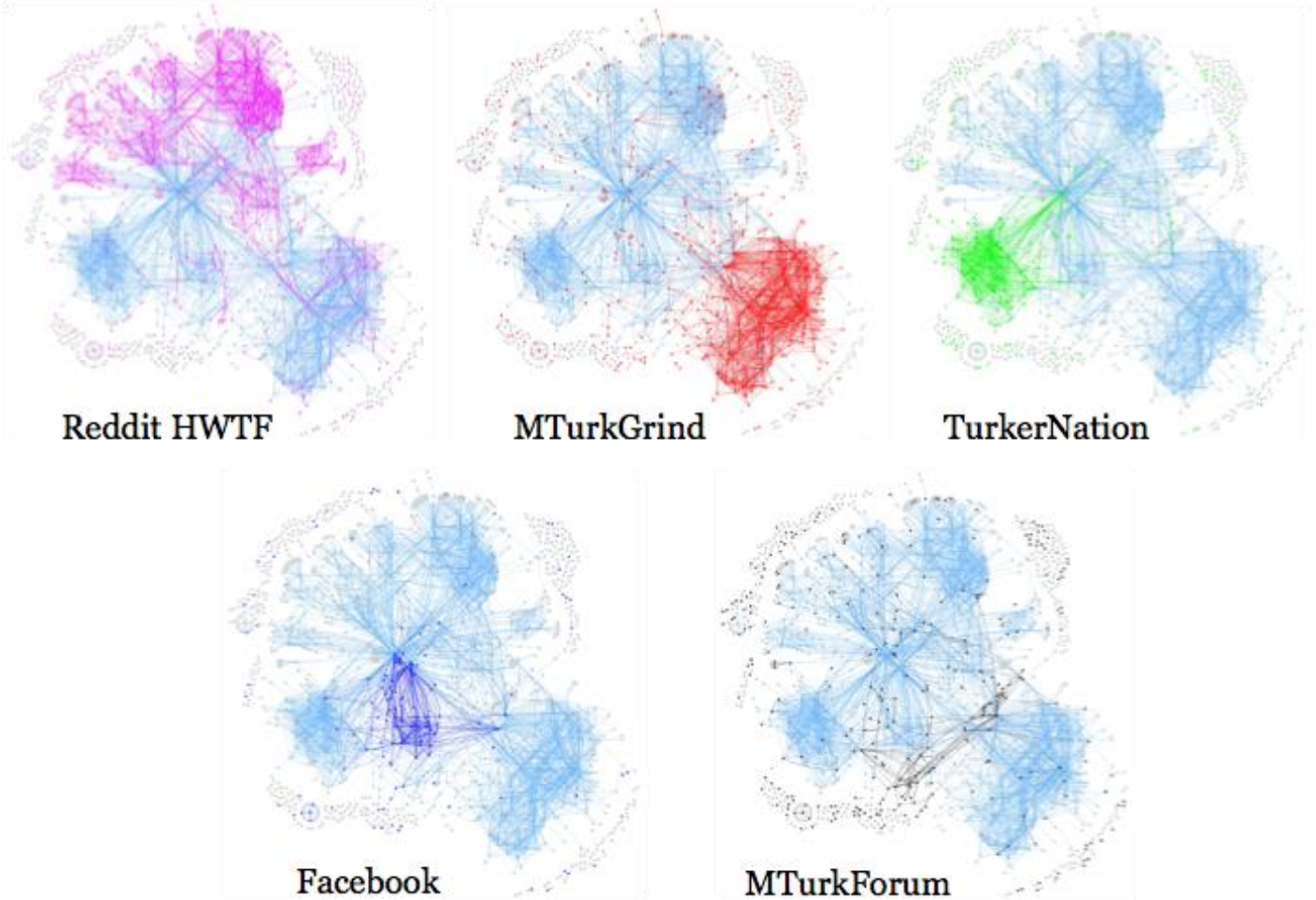
- 1,389 workers (13.4%) added at least 1 edge (called **connected**)



[Yin, Gray, Suri, and Vaughan, 2016]

- 59% of all workers and 83% of **connected** workers reported using at least one forum.

- 90% of all edges are between pairs of workers who communicate via forums
  - 86% are between pairs communicate exclusively through forums.



Worker forums

- The “water cooler” of online crowdsourcing
- Sharing of good and bad HITs, requesters
- Many forums have moderators, rules to protect integrity of research studies
- Engaging with workers on forums: find bugs, get feedback, manage relationships

08-11-2014, 08:10 PM #694

mizzao
Newbie
member

Join Date:	Jun 2014
Posts:	23
Points:	29,477
Thanks:	13

Thanked 201 Times in 22 Posts

Originally Posted by karimi53

Originally Posted by ChicagoK

Finally got to do crisis mapping today and it was awesome!

If memory serves me, that's the one the requester asks that no one discusses anywhere.

The first rule of Crisis Mapping is: You do not talk about Crisis Mapping.

The second rule of Crisis Mapping is: You do not talk about Crisis Mapping.

Jokes aside, we just don't want people discussing the HIT extensively in forums because it can mess up our research. The ultimate goal of this HIT is to be able get good workers to effectively respond to real crises, and for that we need to be able to do simulated mapping in a controlled way and improve the design of the system. Once we figure out how everything works, we'll hopefully be able to respond to real disasters in the future. I can already say I've been honestly impressed with some of the teams so far.

Feel free to say that crisis mapping was awesome. Just don't say too much more ☺

08-11-2014, 08:13 PM #700

loki3404
Master
member

Join Date:	Apr 2014
Location:	NYC
Posts:	603
Points:	488,290
Thanks:	2,464

Thanked 2,282 Times in 570 Posts

Originally Posted by mizzao

The first rule of Crisis Mapping is: You do not talk about Crisis Mapping.

The second rule of Crisis Mapping is: You do not talk about Crisis Mapping.

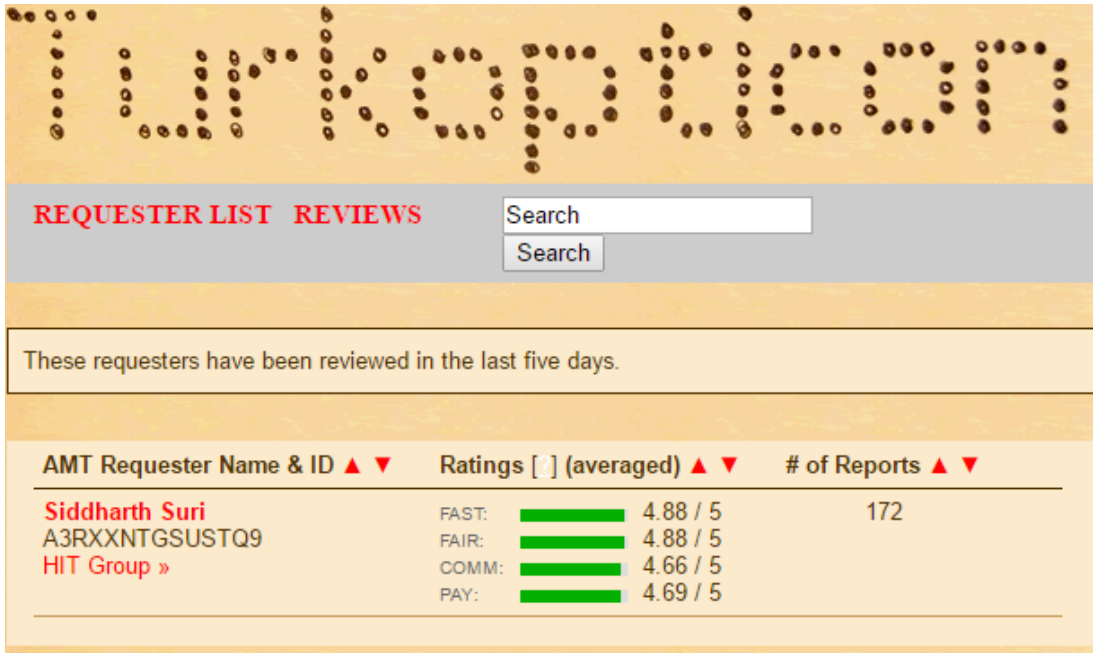
Jokes aside, we just don't want people discussing the HIT extensively in forums because it can mess up our research. The ultimate goal of this HIT is to be able get good workers to effectively respond to real crises, and for that we need to be able to do simulated mapping in a controlled way and improve the design of the system. Once we figure out how everything works, we'll hopefully be able to respond to real disasters in the future. I can already say I've been honestly impressed with some of the teams so far.

Feel free to say that crisis mapping was awesome. Just don't say too much more ☺

You are quickly turning into my favorite requester.

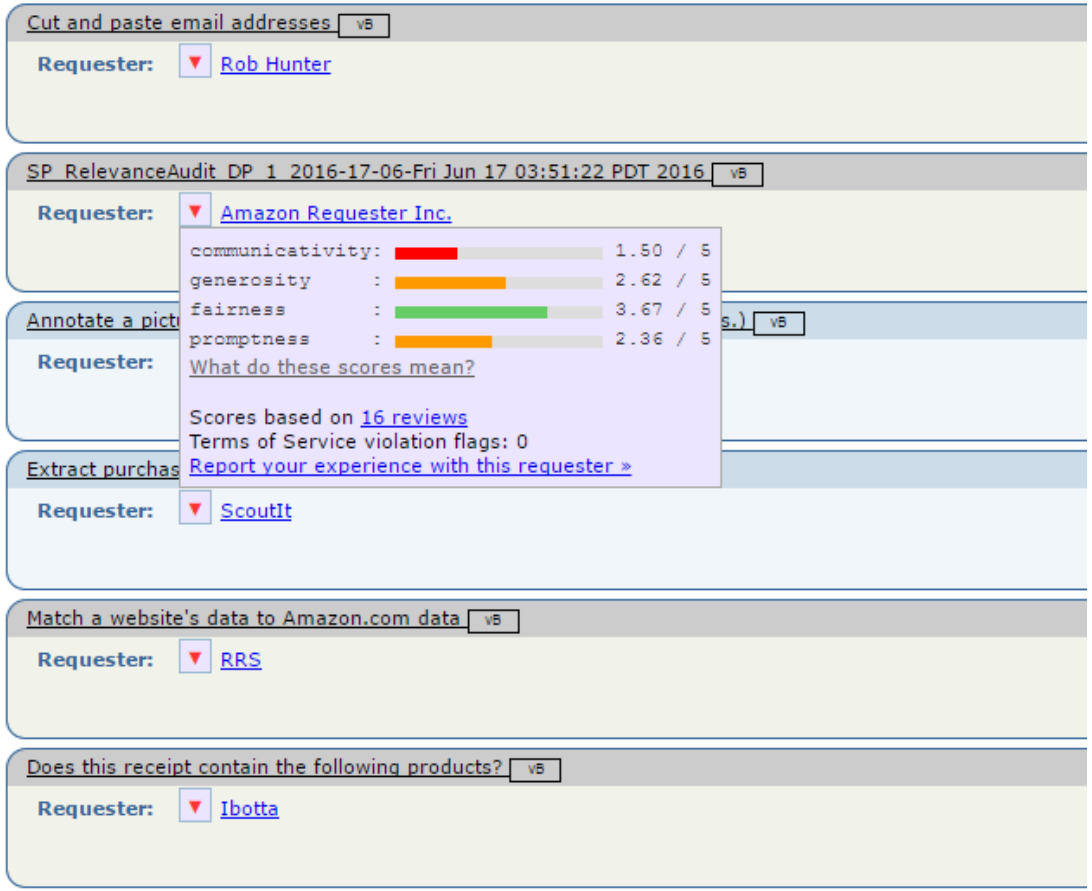
www.reddit.com/r/HITsWorthTurkingFor, www.mturkgrind.com, turkernation.com, etc.

TurkOpticon – 3rd-party requester reviews



With better reputation comes more diligent and helpful workers, faster recruitment, etc.

Check your reputation: https://turkopticon.ucsd.edu/



Most workers use a browser extension showing reviews inline

Attention, disconnection, and attrition

When designing experiments, consider that:

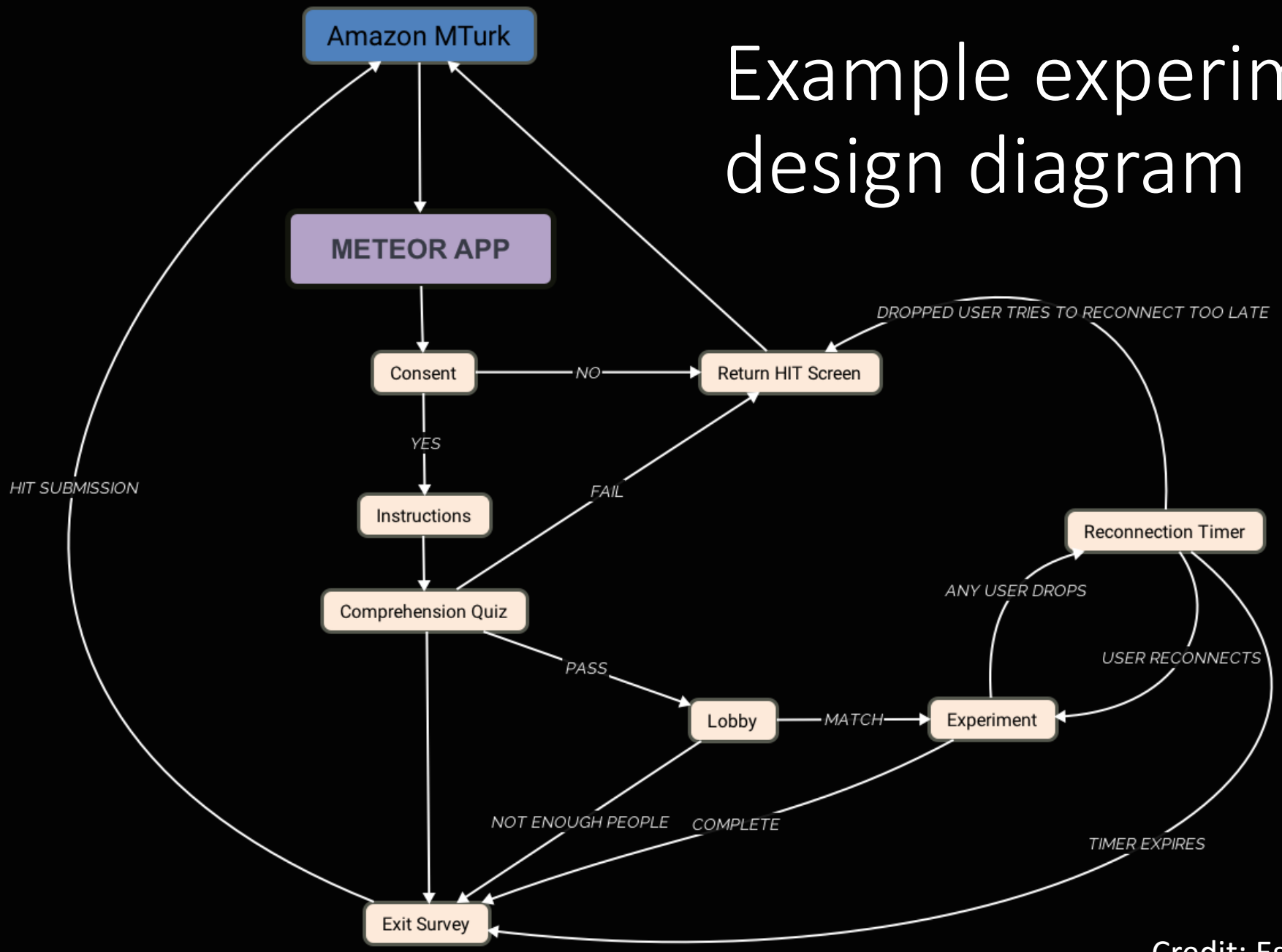
- Participants may not give their full attention
- They may lose connection briefly or go idle for some time
- They can leave the experiment altogether

This affects:

1. The quality of your data,
2. The experience of other participants

TurkServer handles reconnections, and can record inattention. The rest is up to your experiment design

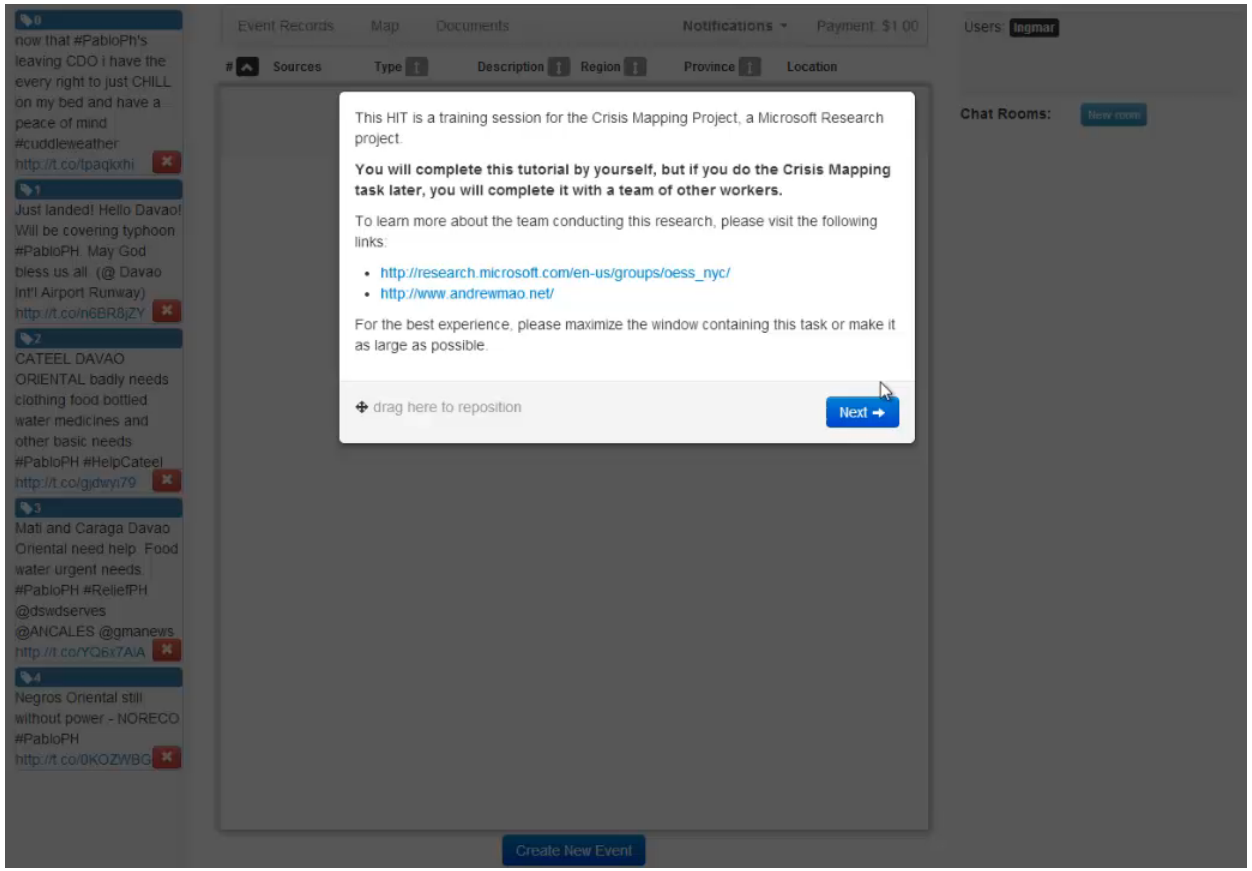
Example experiment design diagram



Credit: Eshin Jolly

Giving clear instructions

- Explain clearly and concisely: you always know your experiment better than the participants!
- Make sure people can't just "click through" to finish and get paid
- Make any unusual requirements of the experiment **known upfront**



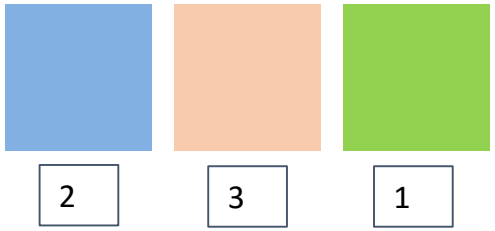
Interactive training for Meteor apps: https://github.com/mizzao/meteor-tutorials

Designing user interfaces

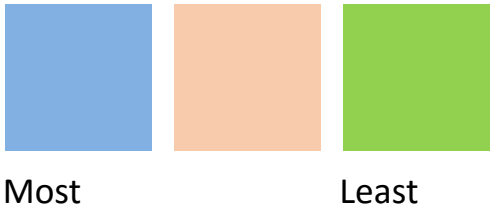
Reduce unnecessary variance in your data:

- Check for comprehension of instructions
  - Check for understanding with a quiz
  - Check if workers are using all the features of the interface
- Making information easy to process
- Making interactions easy to perform
  - Drag and drop
  - Reduce excessive buttons/text entry where possible

Enter numbers for your preference:



Drag and drop to indicate your preference:

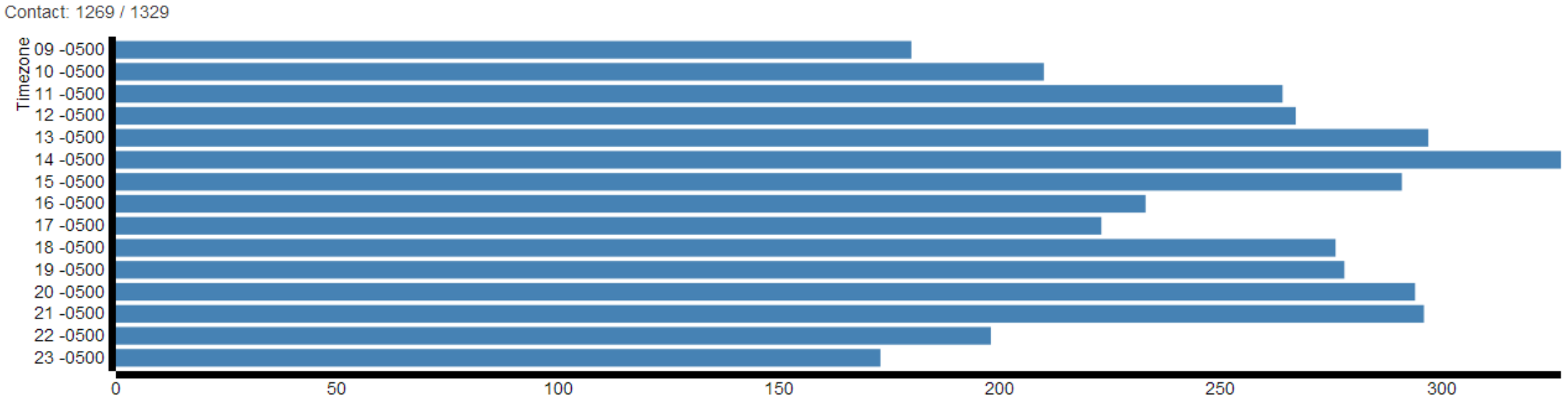


A small amount of laziness results in a huge change in data quality!

Panel recruiting of large groups

- For large simultaneous groups, schedule sessions in advance [Mason and Suri 2012]
- For unusual requirements on participation, set guidelines upfront and allow people to opt-in until there are enough users

Collecting panel time availability using TurkServer:



2pm ET/ 11am PT

9pm ET/ 6pm PT

Testing your app, pilot experiments

It's rare to get experiments completely right the first time!

- Run pilot studies:
  - Project collaborators
  - Co-workers who didn't design the experiment
  - Small samples from the intended subject pool
- Make sure to check:
  - Are the instructions clear?
  - Is the user interface effective?
  - What happens (*to other users; data*) if users reconnect or drop out?
  - Is all of the relevant data being collected and stored properly?

Exit Surveys

- Ask participants
  - If they understood the instructions
  - If they understood the task
  - How they approached the task: strategies, beliefs, etc.
    - Qualitative observations can contribute significantly to quantitative analysis
  - **If they observed bugs or unexpected events**
- Debrief participants
  - To explain the purpose of the research, if not part of the informed consent process
  - *If any deception was involved in the experiment*

Managing a live experiment

- Supporting dozens/hundreds of active users can be frenetic: Plan your logistics beforehand
  - What times will you run the experiment?
  - Prepare a checklist (**like launching a spaceship**)
  - Have a backup plan
  - Divide up responsibility among team members
- Allow time for communication with participants (workers), including responding in forums and answering e-mails
- Take notes of bugs or issues to fix later
- Pay workers promptly

Acknowledgments

Sid Suri

Winter Mason

Duncan Watts

Lili Dworkin

Eshin Jolly

Kevin Gao

Thank You!

Additional resources (contributions welcome!)

- TurkServer: github.com/TurkServer/turkserver-meteor
- Guide: turkserver.readthedocs.io
- Simple example: github.com/TurkServer/tutorial

Contact: mao@microsoft.com; mizzao@gmail.com

Twitter: @mizzao

Microsoft Research

Questions, discussion, and brainstorming

- Any missing details that you are particularly interested in?
- Discussion and comparison of crowdsourced, social experiments to other approaches?
- Feasibility of potential experiment designs?