

How, When and Why to Conduct Behavioral Experiments

Andrew Mao

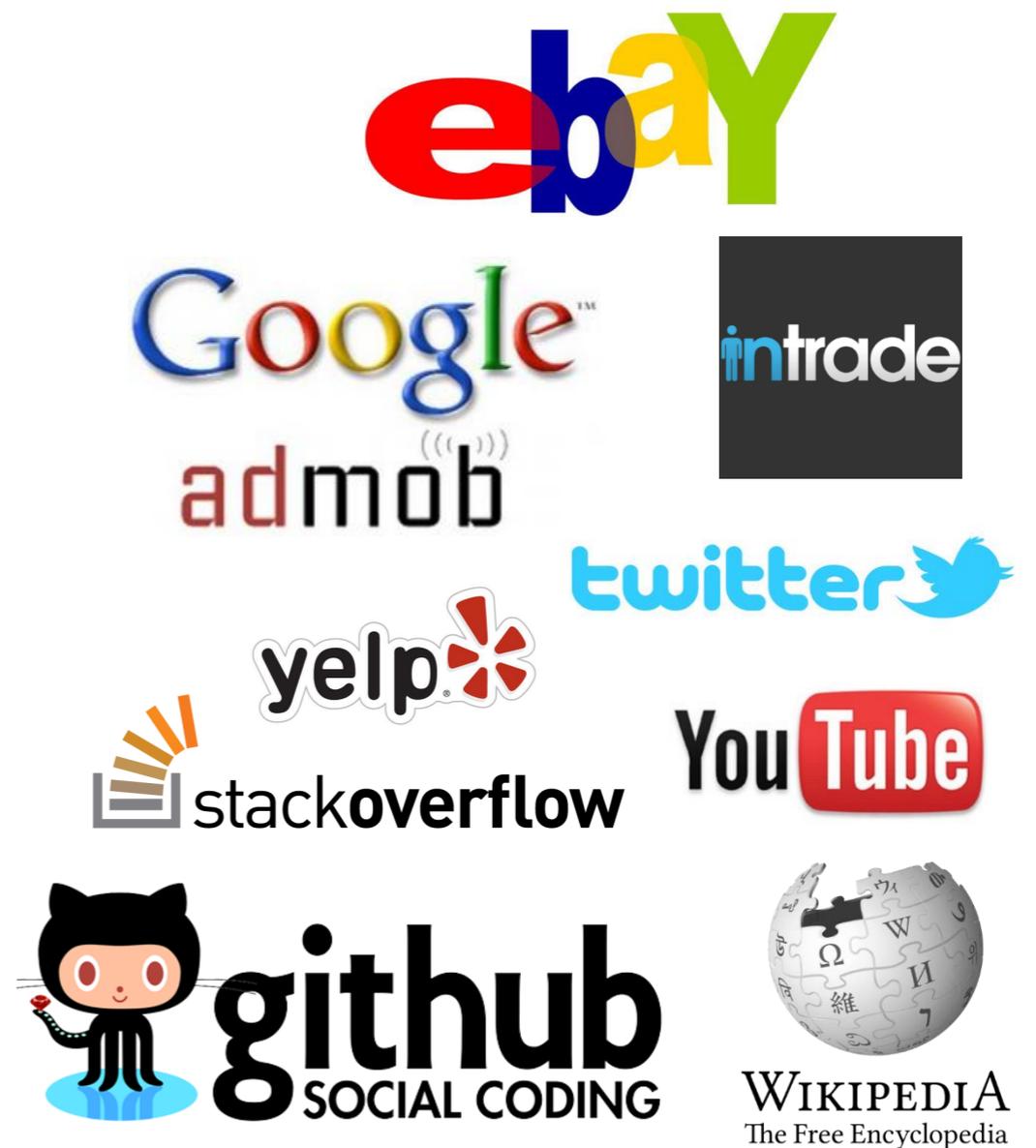
▶ Harvard University

Siddharth Suri

▶ Microsoft Research, New York City

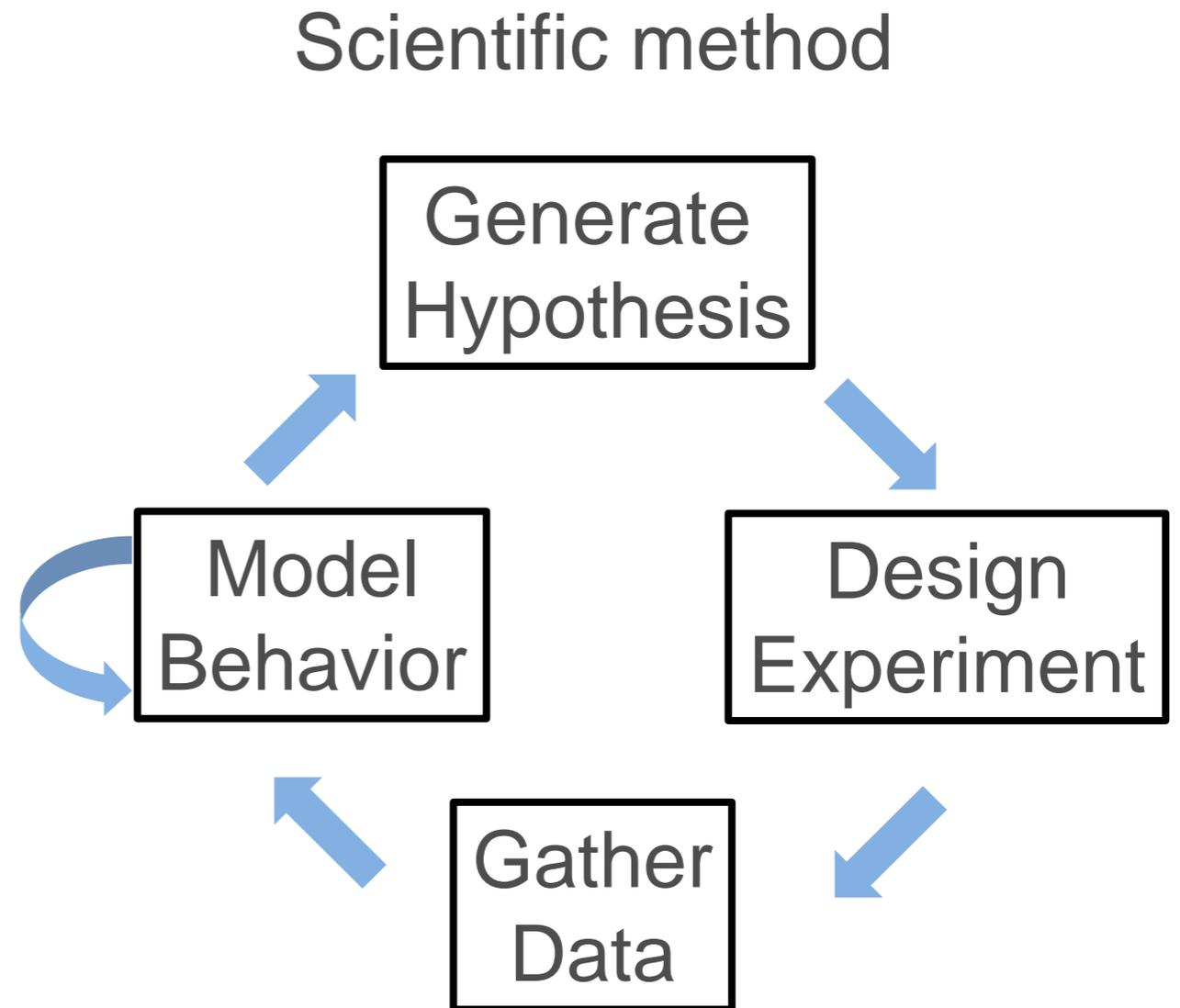
Why Study Human Behavior?

- Computer science enabled the construction of all of these technological systems
 - Humans use these systems
- Computer scientists need to understand how people use these systems to improve them, construct new systems
- Also gives insight into human behavior



How Can We Study Behavior?

- Various methods used in computer science:
 - Theory
 - Simulations
 - Data Mining
- Fields such as psychology, economics, but not CS, use behavioral experiments



Why Conduct a Behavioral Experiment? Causality

- Controlled, randomized experiments are the gold standard method for establishing causality about human behavior
 - Since only a single variable changes between treatment and control, and participants were randomly assigned, any effect must be due to treatment
- Need causality to design interventions
 - If I eat more vegetables will I live longer?
 - Which advertising campaign is more effective?
 - Are smaller classes better for students?

Why Use Experiments? Online Behavior is Far From Understood

- A new area of research: online, interconnected behavior, with anonymity and electronically mediated communication
 - In-person, physical world behavior may not apply
 - Greater numbers of people and more data, yet less is understood
- Electronic setting makes the online environment less accessible to fields that have traditionally studied behavior
- Social scientists realize this too!

Why Use Experiments? Fewer Barriers to Doing Them

Barriers to doing behavioral experiments in computer science that are coming down:

- Computer science not interested in human behavior
 - *Algorithmic game theory, mechanism design, participation in large online communities & software systems*
- Experiment design is not in the typical CS curriculum
 - *this tutorial, summer school, forthcoming papers*
- Doing experiments is time-consuming, expensive, and not well established as a method
 - *more trailblazers, better tools to execute experiments*

Why Use Experiments? Test Models and Design Systems

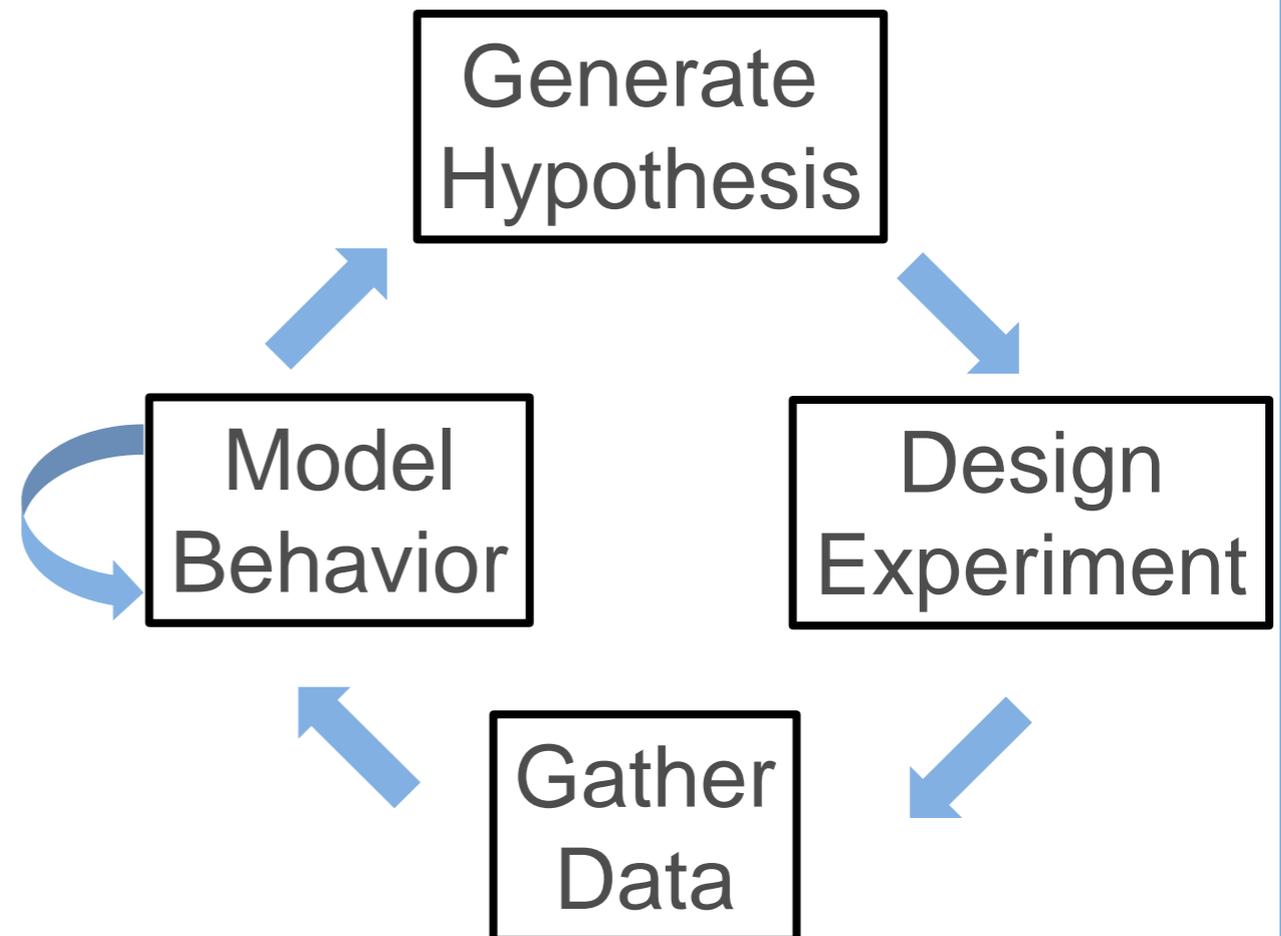
- Economists design models and study them empirically
 - *Computer scientists also design models, but often don't study empirically*
- Computer scientists build large systems with many human participants (peer-to-peer networks, display advertising, online user-generated content)
 - *but often in an ad hoc way or without understanding behavior*
- Understanding human behavior is important for
 - *Testing models and theory*
 - *Designing mechanisms that incorporate human input*
 - *Improving systems for online collaboration, communication, human computation, etc.*

Why Use Experiments? We Have Some of the Best Tools

- Studying online behavior requires a synthesis of skills:
 - Asking interesting questions and proper experiment design
 - Building and deploying an appropriate experiment interface
 - Producing useful insights from lots of data
- As a whole, **computer scientists are uniquely equipped to study online behavior using tools that we already have:**
 - Programming and system-building experience
 - Computational and statistical tools to analyze empirical data
 - Techniques to develop new models and theory from the results

Why Use Experiments? Because we can!

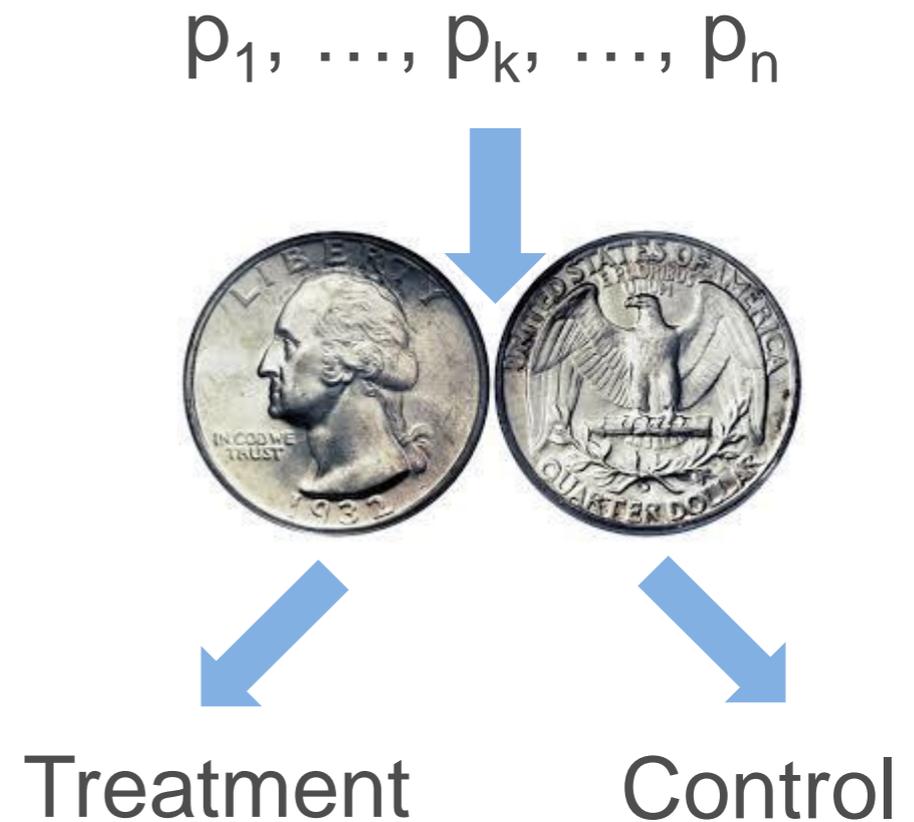
- CS already has data mining, modeling, and theory expertise
- Add in experiments and we can complete the scientific method.
- If we can go all the way around this cycle to understand behavior, then we should!





What is a Behavioral Experiment?

- Behavioral experiment means randomized, controlled experiment
- Each participant is randomly assigned uniformly and independently to treatment or control
- Manipulate a single variable while controlling the rest
- For example: clinical trial



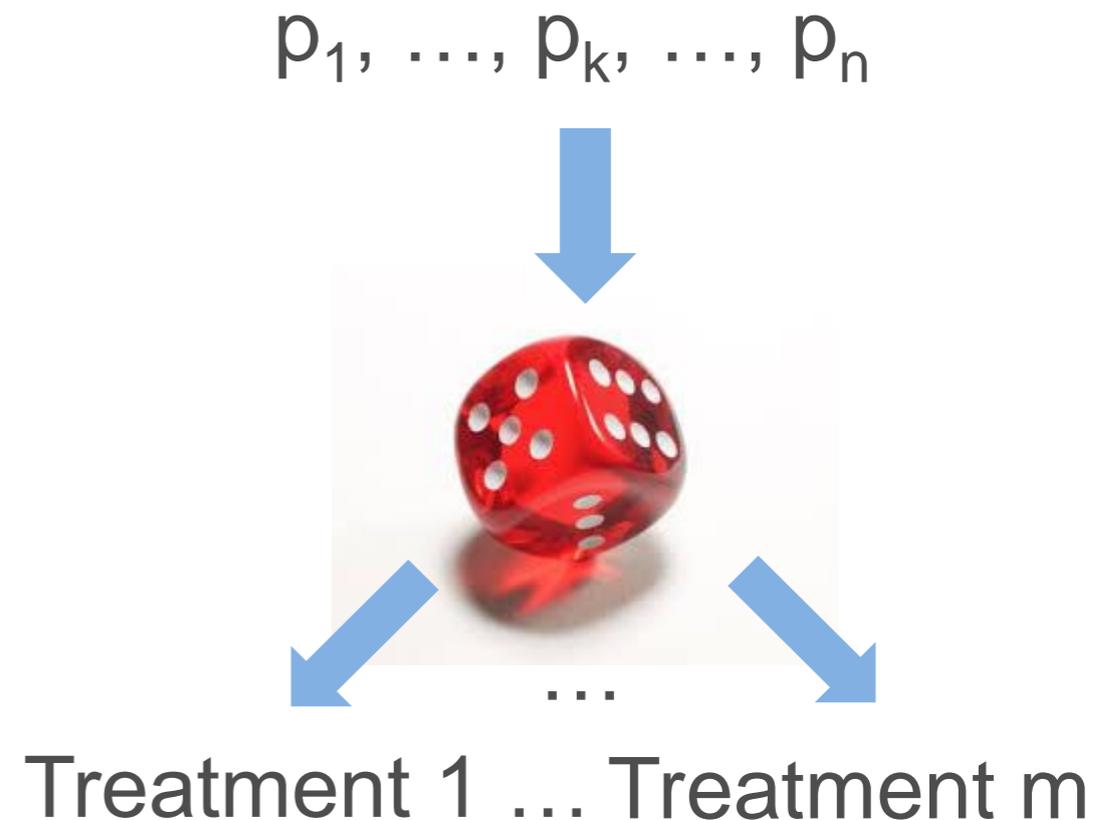
What is a Behavioral Experiment?

- Behavioral experiment means randomized, controlled experiment
 - Each participant is randomly assigned uniformly and independently to a treatment
 - Manipulate a single variable while controlling the rest
 - For example: honesty study [Suri, Goldstein, Mason 2011]



What is a Behavioral Experiment?

- Behavioral experiment means randomized, controlled experiment
 - Each participant is randomly assigned uniformly and independently to a treatment
 - Manipulate a single variable while controlling the rest
 - For example: behavioral graph coloring [Kearns, Suri, Montfort '2006]



Pros and Cons of Behavioral Experimentation

- + Establish causality
- + Can put people in a situation that might not exist
- Demographic limitations: Can be hard to generalize from subject pool to general population
- External validity: does the experimental setting map to a real world setting
- Can be hard to collect lots of data (mitigated by online experiments)
- Only changing 1 thing between treatments is slow, tedious

Studying Behavior Using Data Mining

- All of these systems generate huge log files of human behavior.
- Data mining finds interesting correlations in data sets
- [Burt '92, '04] found that people who bridge different communities are more successful
 - Are they successful because they are bridges?
 - Do they bridge because they are successful?

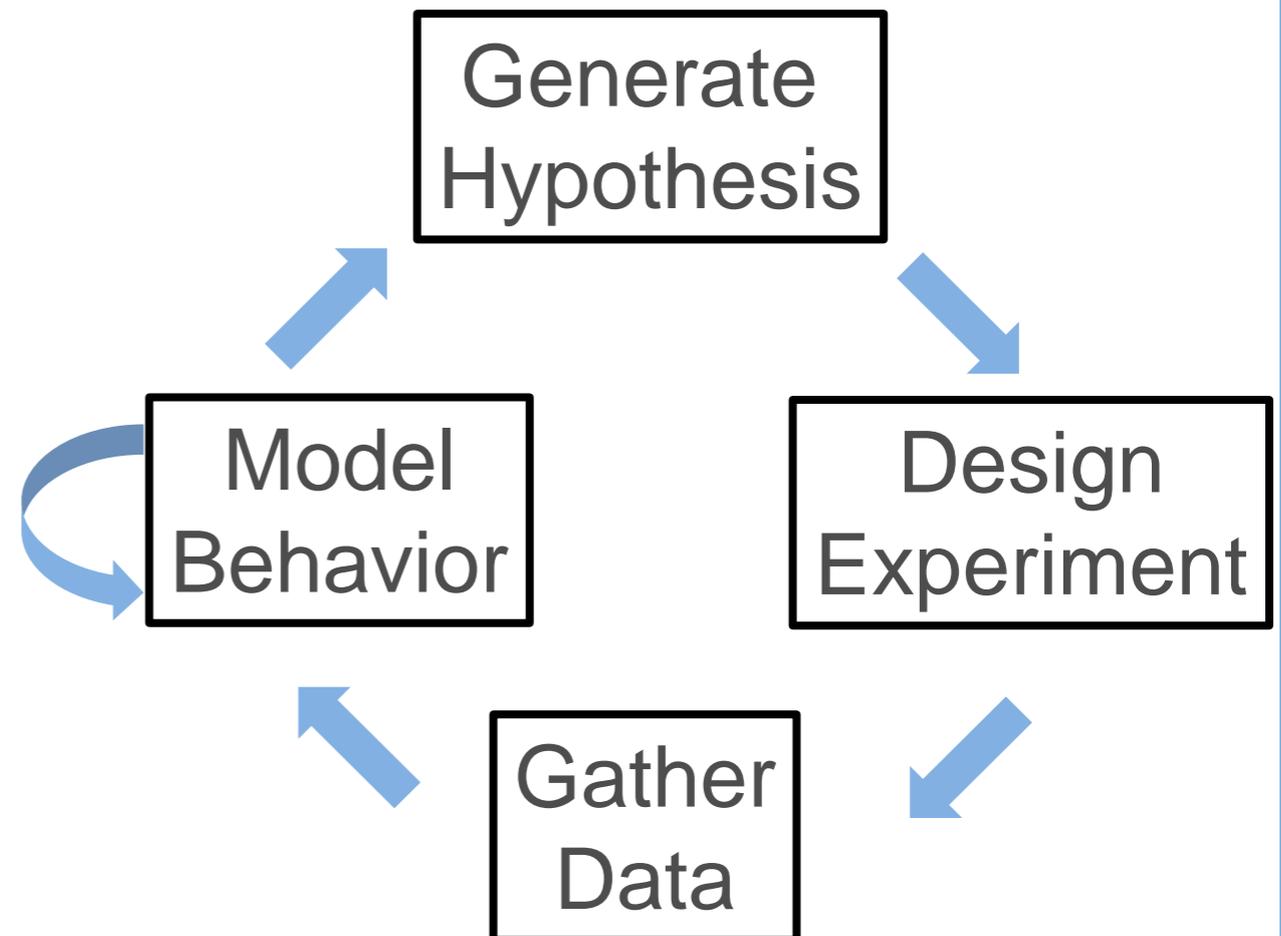


Studying Behavior Using Data Mining

- + Logs of human behavior on the internet means lots of data available
 - + Human behavior recorded at an unprecedented scale
- + Computational power (CPUs, storage) is cheap
- + Real data to test and build models from
 - + Starting to see agent-based models based on real behavioral data (empirical agent-based models)
- Can't always get the data you want (e.g. Facebook graph, Twitter graph)
- Must take the data as you get it: a log of some type of human behavior
 - Probably won't match exactly with the research question you originally had
- Can have many phenomenon occurring at the same time
- Can be hard to observe behavior in certain situations
- Only correlations! No causality

Incorporating Behavioral Experiments with Data Mining

- Data mining provides correlations
 - Can test for causality using experiments
- Data mining provides external validity for experiments
- Data mining casts a wide net, experiments are a microscope
- Massive noisy data vs. small, less noisy data
 - **Big data → big experiments:** design your data instead of passively collecting it

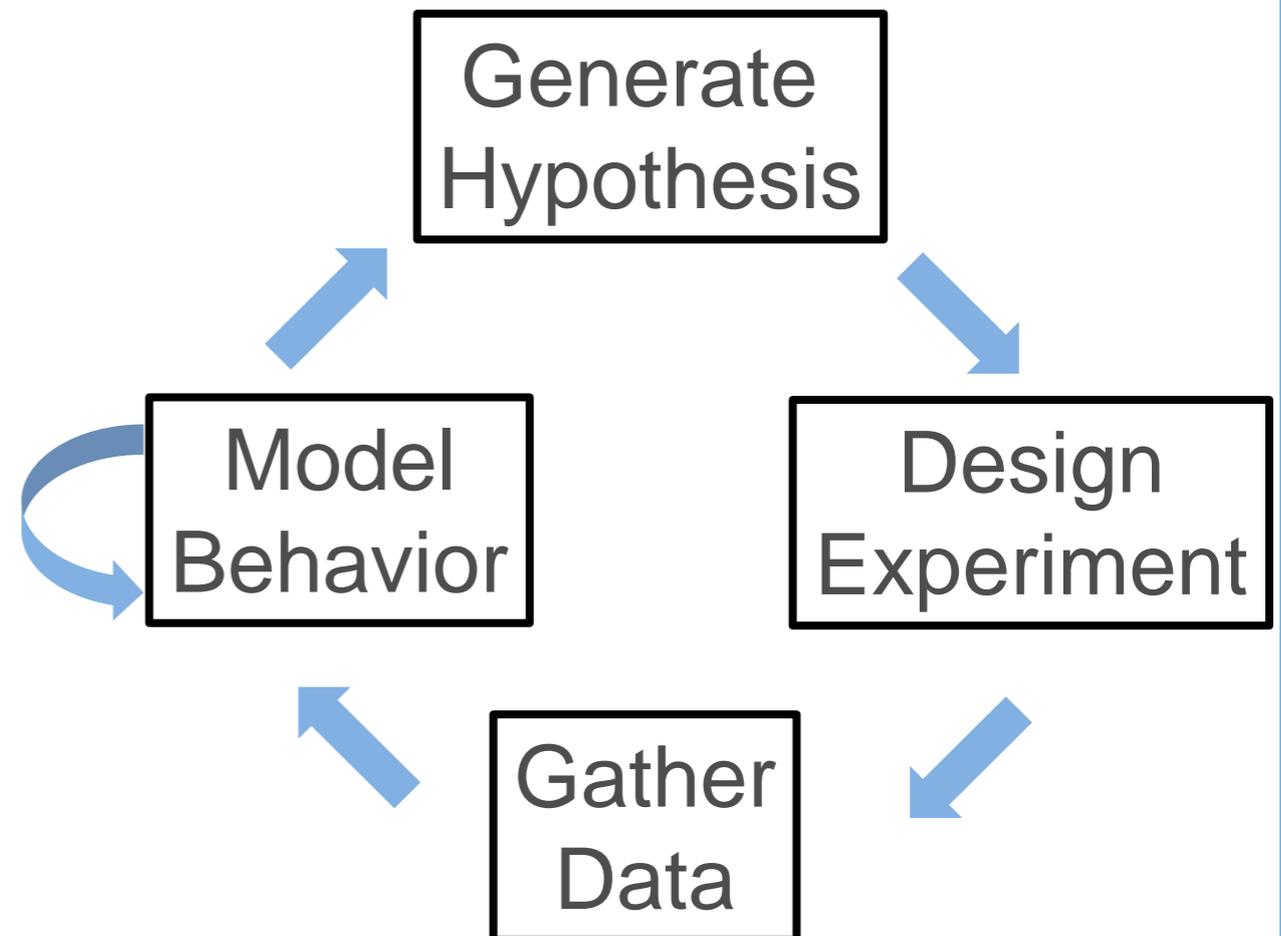


Studying Behavior with Theory & Simulations

- + Knowing the worst case (running time, Price of Anarchy, approximation ratio, etc.) is important
- + Covers situations that may not be observable empirically
- + Makes predictions
- + Computer scientists are good at it (competitive advantage)
- Often requires strong assumptions for mathematical tractability (e.g. rationality, quasi-linear utility etc.)
- Models need to strip away “messy” details (e.g. reputation, learning, etc.)
- Since models may not be realistic, predictions may be incorrect

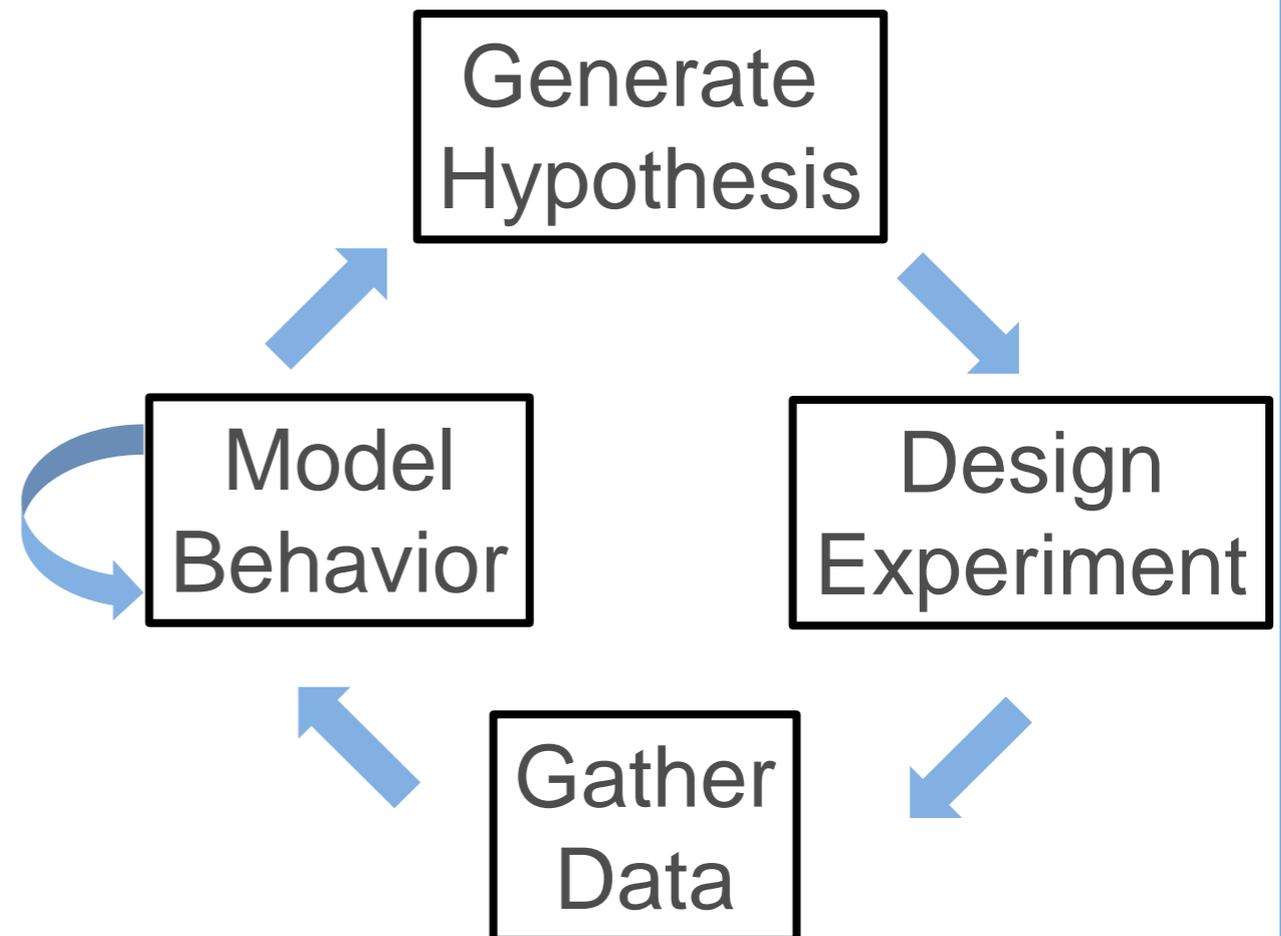
Incorporating Behavioral Experiments with Theory

- Theoretical C.S. derives many models of human behavior
- Can test predictions using experiments, iterate
- E.g. If a mechanism predicts a behavior, try it!
- Where to start? Theory or experiment



Incorporating Behavioral Experiments with Simulations

- Simulations and agent based models can check many parameter settings
 - Often assume some type of behavior
- Can test predictions
- Can generate data to base models on
 - Empirical Agent Based Models



When to Conduct a Behavioral Experiment

- Conduct a behavioral experiment when you want to make a causal claim about human behavior
 - Perhaps you want to design an intervention
- Want to test a prediction from a theory or simulation
- Want to test a correlation found in data
- What to test how people behave in a situation that does not yet exist
- **As a means for studying how people behave on the Internet, in a natural setting**
- **A promising complement to existing established methods**

The Camerer* Test

- Look for experiments where the result is interesting no matter which direction it comes out.
 - Perhaps one theory predicts one direction and another theory predicts the opposite direction
 - Or, both a positive and negative result would be interesting observations
- Behavioral experiments are a lot of work
 - This ensures that only risk is not enough data to detect effect
 - Can minimize this risk if you do the experiment online

* Told to me by Colin Camerer

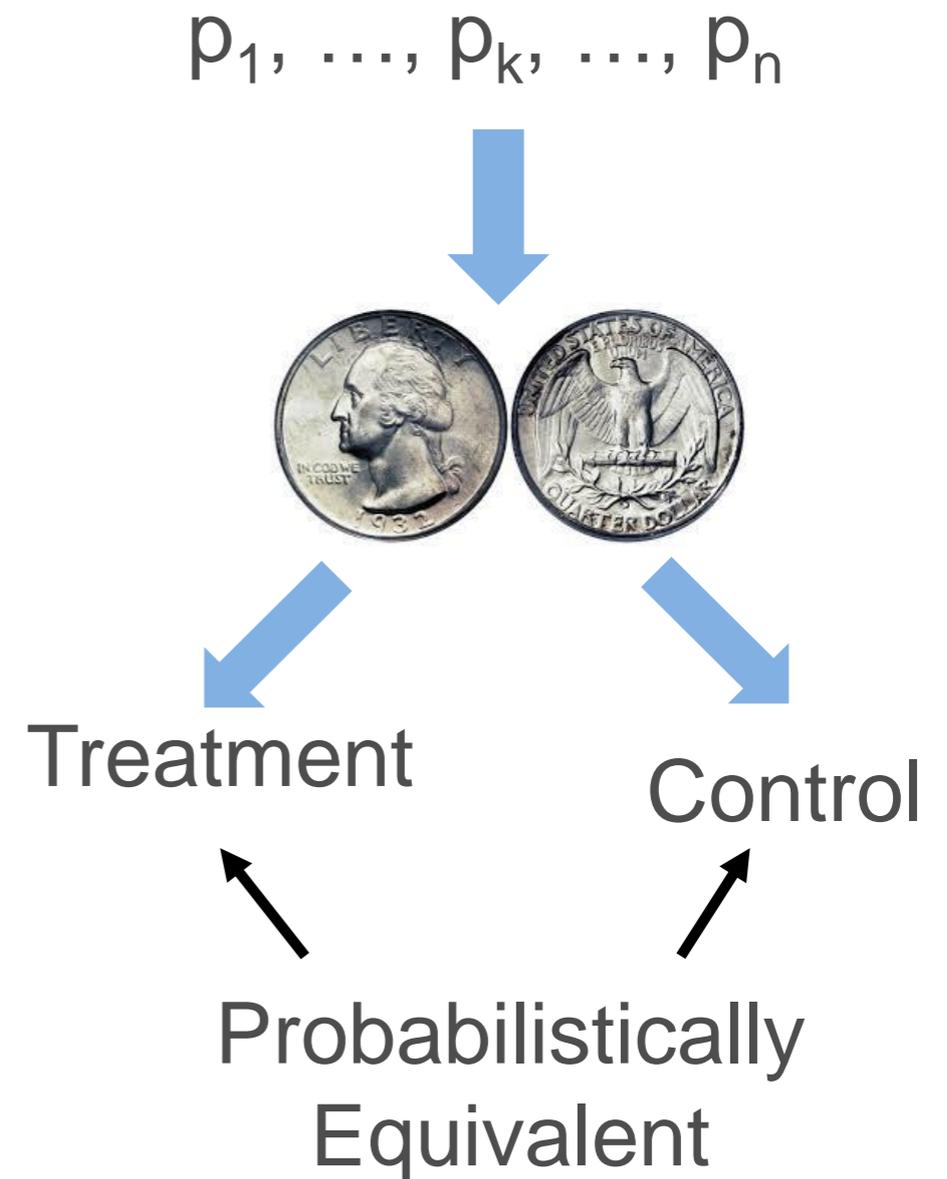
Designing Behavioral Experiments

Terminology

- Independent variable: what the experimenter manipulates
- Dependent variable: what the experimenter presumes to be affected by the independent variable
- Hypothesis: statement of prediction
- Participants or subject: human beings who do the experiment
- Treatment: Specific condition applied to a group of participants

The Importance of Randomization

- Assignment to treatment or control is independent of the subject
 - Results in two groups of subjects that are equal in all aspects
 - Comparing treatment group and control group ensures that factors other than treatment operate equally on both groups
 - Difference in average effect must be due to treatment or bad luck
 - Statistics (e.g. t-test) bounds this



Correlation \neq Causality

- Observation: Students who take SAT prep classes do better on the SAT
- Could match students on observable attributes (gender, age, race, GPA, etc.) and compare
- Can't observe everything!! E.g. motivation
- Experiment: Randomly assign students to take SAT class or gym class, compare test scores
 - Random assignment ensures all attributes (observable or not) will be the same in either class

SAT class



better score

High
motivation



SAT
class

Better
score

Correlation \neq Causality

- If X and Y are correlated, they might not have any causal relation between them!
- If X , Y and Z are variables and $X \sim Y$ then either:
 - $X \rightarrow Y$
 - $Y \rightarrow X$
 - $Z \rightarrow X, Y$

Confound: The Enemy of Causality

- Confound: when more than one thing is different on average between the treatment and control
 - The goal of experimental design is to conduct experiments that avoid confounds
- Random assignment to treatment is the best way to ensure that all confounding variables are equal on average between groups
 - As we will see, randomization is not enough by itself

Example Confound: Selection Effects

- Research question: Do people cooperate more with their friends than strangers?
- Design: Facebook game which allows users to choose to play prisoner's dilemma with a friend or a random player
- Say people cooperate more with their friends
 - Is it because they played a friend?
 - Or that those who choose to play a friend are more cooperative

Validity

- Internal validity is the ability to make causal conclusions
 - Avoid confounds
- External validity: can I generalize the effect I found to other populations, places, times, settings
 - What does this experiment say about the world?
- Construct validity: does what I am manipulating and measuring map to a real world phenomenon?
 - Did I accurately operationalize what I am trying to study?

Analyzing Data

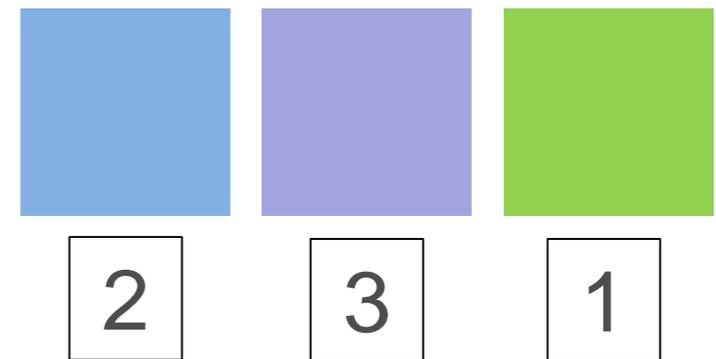
- Know what analyses you want to do with data before you gather data: how will you make a causal claim?
 - What type of regression or statistical test?
 - What model to use for learning from or fitting to the data?
- Are you making the best choice for a dependent variable?
- Will give example later

User Interface Design

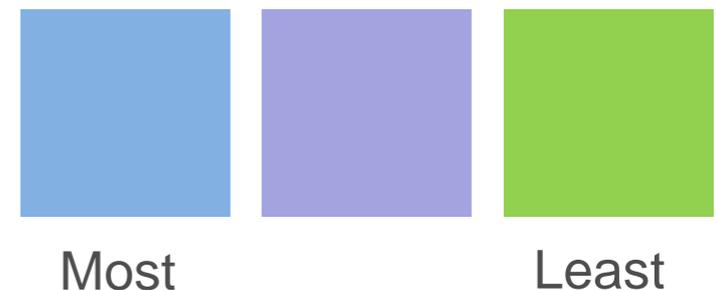
Reducing unnecessary variance in data by...

- Designing good tutorials
 - 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 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!

Additional Factors in Web-based Experiments

- Latency from remote connections
- Language or cultural barriers and diverse demographics
- Exogenous communication/chatter between participants
- Can't ensure that you will have undivided attention:
 - Bathroom breaks
 - Picking up kids
 - Walking the dog
- (but can check if attention was lost)

Institutional Review Boards (IRBs) and Research Ethics

- Researchers' self-policing was not enough to prevent serious harm:
 - Tuskegee syphilis experiment, Milgram obedience experiment, Stanford prison experiment, etc. *
- Institutional Review Boards
 - Review whether the potential benefits of research are worth the risks
 - Determines whether deception or other harm is justified, especially for vulnerable subjects
 - Reviews the informed consent process for an experiment
- Many archival venues require IRB approval for experiments
 - Most online studies can be reviewed under an expedited process
 - Online research ethics standards are not yet standardized and vary widely across institutions; expect convergence as online experiments become more prominent

* http://en.wikipedia.org/wiki/Unethical_human_experimentation_in_the_United_States

Exit Surveys

- Ask participants
 - If they understood the instructions
 - If they understood the task
 - How they approached the task: strategies, beliefs, etc.
 - **If they observed bugs or unexpected results**
- Debrief participants
 - If deception was involved in the experiment
 - To explain the purpose of the research, if not part of the informed consent process

Pilot Experiments

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

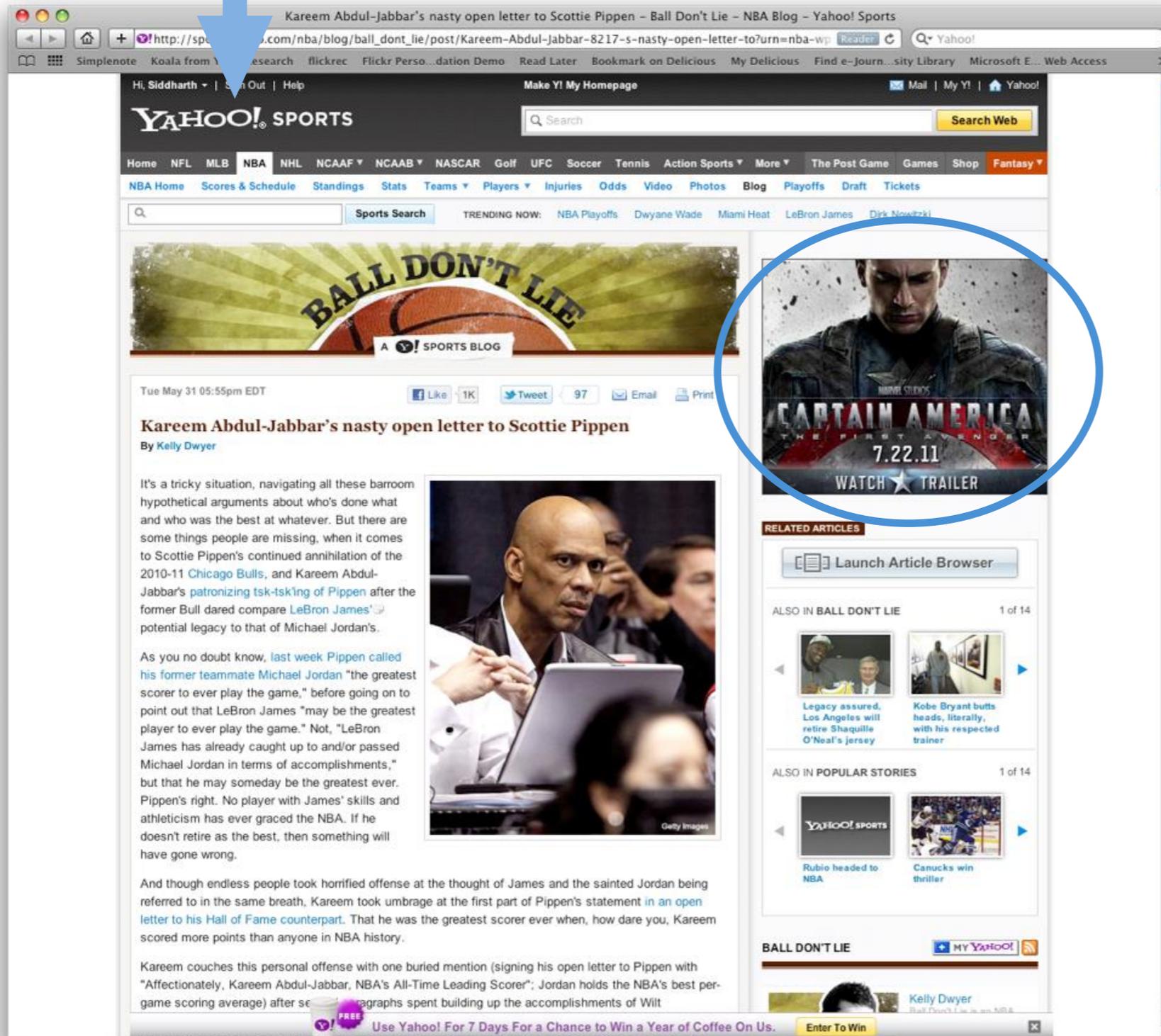
- Run pilot studies with collaborators, co-workers, and small samples from the intended subject pool
- Make sure to check
 - Are the instructions easily understood or confusing?
 - Is the interface intuitive to use or not?
 - Is all the necessary data being collected?

Experiment Design: Examples

Publisher: Yahoo!

One download of ad is one impression

Display Ad Advertiser: Marvel Studios



Research Question

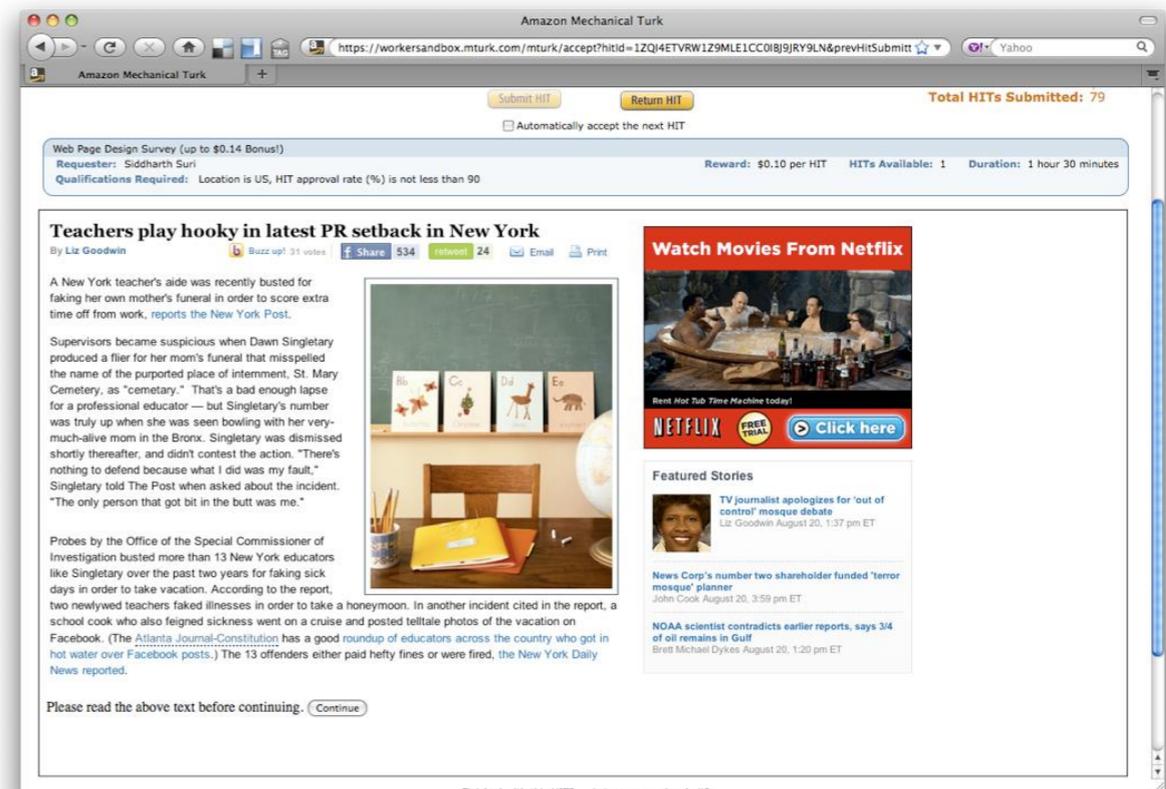
- ▶ Should display ads be sold by impression or by exposure time?
- ▶ Want to charge based on advertiser value.
 - ▶ Advertisers value brand recall and brand recognition
 - ▶ Recall and recognition affect consumers consideration set
- ▶ Will time of exposure causally influence recall/recognition for display ads? Not clear for two reasons
 - ▶ Maybe users initially scan the page and then focus on content
 - ▶ Banner blindness suggests time does not matter [Benway, '98]

Time Based Display Advertising Design 1

(spot the confounds)

Experimental Design

- ▶ Mechanical Turk experiment
 - ▶ [Mason & Suri 2011]
- ▶ Instructions: read web page and answer questions about it
- ▶ Payment: \$0.50 flat rate plus \$0.10 per question answered
- ▶ Ad would disappear after a few seconds
- ▶ [Goldstein, McAfee, Suri '11]



Experimental Manipulation

- ▶ Long and short versions of article manipulate time ad is in view
 - ▶ Long version is 2 screens of text
 - ▶ Short version is $1+\epsilon$ screens of text
 - ▶ Avoid scrolling, one screen full is about 600 pixels
 - ▶ Keep the number of impressions constant
 - ▶ 2 x 2 x 2 design, randomized assignment

article	length	ad
{NY teachers, oil spill}	{short, long}	{netflix, jeep}
- ▶ Factorial design: 3 factors, each with 2 levels

NY Teachers article, Jeep Ad

Teachers play hooky in latest PR setback in New York
 By Liz Goodwin | Buzz up! 31 votes | Share 534 | retweet 24 | Email | Print

A New York teacher's aide was recently busted for faking her own mother's funeral in order to score extra time off from work, [reports the New York Post](#).

Supervisors became suspicious when Dawn Singletary produced a flier for her mom's funeral that misspelled the name of the purported place of interment, St. Mary Cemetery, as "cemetary." That's a bad enough lapse for a professional educator — but Singletary's number was truly up when she was seen bowling with her very-much-alive mom in the Bronx. Singletary was dismissed shortly thereafter, and didn't contest the action. "There's nothing to defend because what I did was my fault," Singletary told The Post when asked about the incident. "The only person that got bit in the butt was me."

Probes by the Office of the Special Commissioner of Investigation busted more than 13 New York educators like Singletary over the past two years for faking sick days in order to take vacation. According to the report, two newlywed teachers faked illnesses in order to take a honeymoon. In another incident cited in the report, a school cook who also feigned sickness went on a cruise and posted telltale photos of the vacation on Facebook. (The *Atlanta Journal-Constitution* has a good roundup of educators across the country who got in hot water over Facebook posts.) The 13 offenders either paid hefty fines or were fired, the *New York Daily News* reported.




Featured Stories

 **TV journalist apologizes for 'out of control' mosque debate**
Liz Goodwin August 20, 1:37 pm ET

News Corp's number two shareholder funded 'terror mosque' planner
John Cook August 20, 3:59 pm ET

NOAA scientist contradicts earlier reports, says 3/4 of oil remains in Gulf
Brett Michael Dykes August 20, 1:20 pm ET

Page 1

This latest report just compounds a recent string of bad PR for New York teachers in local and national media. Earlier this month, [the Post reported that a first-year Brooklyn teacher threw herself down the stairs in order to avoid a performance review](#).



Featured Stories



TV journalist apologizes for 'out of control' mosque debate
Liz Goodwin August 20, 1:37 pm ET

News Corp's number two shareholder funded 'terror mosque' planner
John Cook August 20, 3:59 pm ET

NOAA scientist contradicts earlier reports, says 3/4 of oil remains in Gulf
Brett Michael Dykes August 20, 1:20 pm ET

Page 2 (short)

This latest report just compounds a recent string of bad PR for New York teachers in local and national media. Earlier this month, [the Post reported that a first-year Brooklyn teacher threw herself down the stairs in order to avoid a performance review](#). Ilene Feldman denied that she faked the fall, but quickly resigned after confronted with the report.

Feldman had received a bad [performance review](#) and was set to be observed by a supervisor for a second review. In New York, teachers can be fired for a bad performance review in their first three years on the job before they gain tenure.

And [critical coverage](#) of the city's "rubber room" — where teachers accused of wrongdoing were sent, doing no work but drawing full salaries while awaiting a lengthy disciplinary process — led [City Hall](#) and the city teachers union to agree to shut the facility down in April. Now the teachers will be assigned administrative work while they await hearings.

[Video: [Should parents have the right to know which teachers are bad?](#)]

When announcing the "rubber room" closure, Mayor [Mike Bloomberg](#) acknowledged that the image of city teachers had suffered.

"Given the amount of press that this subject has gotten, to say that this is a big deal is probably an understatement," [Bloomberg said at a news conference, according to the New York Times](#). "This was an absurd and expensive abuse of tenure. We've been able to solve what was one of the most divisive issues in our school system."

But Diane Ravitch, an education historian who strongly opposes the Obama administration's test-score-based approach to education reform, says that these reports are unfair and part of a larger "attack" on teachers and teachers unions in New York and across the nation.

"New York City has 1.1 million children in its public schools, and about 80,000 or so teachers. The article identifies 13 out of 80,000 or so teachers who used sick days for vacation time. Why does this get turned into an indictment of all teachers?" she wrote in an e-mail.



Featured Stories



TV journalist apologizes for 'out of control' mosque debate
Liz Goodwin August 20, 1:37 pm ET

News Corp's number two shareholder funded 'terror mosque' planner
John Cook August 20, 3:59 pm ET

NOAA scientist contradicts earlier reports, says 3/4 of oil remains in Gulf
Brett Michael Dykes August 20, 1:20 pm ET

Page 2 (long)

Oil spill article, Netflix Ad

White House edits stain its reliance on science

By DINA CAPIELLO, Associated Press – 21 mins ago

WASHINGTON – The oil spill that damaged the Gulf of Mexico's reefs and wetlands is also threatening to stain the Obama administration's reputation for relying on science to guide policy.

Academics, environmentalists and federal investigators have accused the administration since the April spill of downplaying scientific findings, misrepresenting data and most recently misconstruing the opinions of experts it solicited.

Meanwhile, the owner of the rig that exploded in the Gulf of Mexico, Transocean Ltd., is renewing its argument that federal investigators are in danger of allowing the blowout preventer, a key piece of evidence, to corrode as it awaits forensic analysis. Testing had not begun as of last week, the company says, some two months after it was raised from the seafloor.

The blowout preventer could be a key piece of evidence in lawsuits filed by victims, survivors and others. Transocean was responsible for maintaining it while it was being used on BP's well. Investigators agreed to flush the control pods with fluid on Sept. 27 to prevent corrosion. But a Transocean lawyer wrote in his Nov. 3 letter that there have been no further preservation steps on the blowout preventer since then.

RELATED QUOTES

RIG	69.86	+2.04
*GSPC	1,218.71	+5.31
*IXIC	2,578.78	+15.80

Watch Movies From Netflix

Rent Hot Tub Time Machine today!

NETFLIX FREE TRIAL Click here

Featured Stories

TV journalist apologizes for 'out of control' mosque debate
Liz Goodwin August 20, 1:37 pm ET

News Corp's number two shareholder funded 'terror mosque' planner
John Cook August 20, 3:59 pm ET

NOAA scientist contradicts earlier reports, says 3/4 of oil remains in Gulf
Brett Michael Dykes August 20, 1:20 pm ET

Page 1

The latest complaint from scientists comes in a report by the Interior Department's inspector general, which concluded that the White House edited a drilling safety report in a way that made it falsely appear that scientists and experts supported the administration's six-month ban on new deep-water drilling.

AP – Oil spill workers continue the process of cleaning tar balls and oil from the beaches of Orange Beach, ...

Watch Movies From Netflix

Rent Hot Tub Time Machine today!

NETFLIX FREE TRIAL Click here

Featured Stories

TV journalist apologizes for 'out of control' mosque debate
Liz Goodwin August 20, 1:37 pm ET

News Corp's number two shareholder funded 'terror mosque' planner
John Cook August 20, 3:59 pm ET

NOAA scientist contradicts earlier reports, says 3/4 of oil remains in Gulf
Brett Michael Dykes August 20, 1:20 pm ET

Page 2 (short)

The latest complaint from scientists comes in a report by the Interior Department's inspector general, which concluded that the White House edited a drilling safety report in a way that made it falsely appear that scientists and experts supported the administration's six-month ban on new deep-water drilling. The AP obtained the report early Wednesday.

AP – Oil spill workers continue the process of cleaning tar balls and oil from the beaches of Orange Beach, ...

The inspector general said the editing changes by the White House resulted "in the implication that the moratorium recommendation had been peer reviewed." But it hadn't been. Outside scientists were asked only to review new safety measures for offshore drilling.

"There are really only a few people that know what they are talking about" on offshore drilling," said Ford Brett, managing director of Petroskills, a Tulsa, Okla.-based petroleum training organization. "The people who make this policy do not ... so don't misrepresent me and use me for cover," said Brett, one of seven experts who reviewed the report.

Last month, staff for the presidential oil spill commission said that the White House's budget office delayed publication of a scientific report that forecast how much oil could reach the Gulf's shores. Federal scientists initially used a volume of oil that did not account for the administration's various cleanup efforts, but the government ultimately cited smaller amounts of oil.

The same report said that President Barack Obama's energy adviser, Carol Browner, mischaracterized on national TV a government analysis about where the oil went, saying it showed most of the oil was "gone." The report said it could still be there. It also said that Browner and the head of the National Oceanic and Atmospheric Administration, Jane Lubchenco, contributed to the public's perception the report was more exact than it was by emphasizing peer review.

Watch Movies From Netflix

Rent Hot Tub Time Machine today!

NETFLIX FREE TRIAL Click here

Featured Stories

TV journalist apologizes for 'out of control' mosque debate
Liz Goodwin August 20, 1:37 pm ET

News Corp's number two shareholder funded 'terror mosque' planner
John Cook August 20, 3:59 pm ET

NOAA scientist contradicts earlier reports, says 3/4 of oil remains in Gulf
Brett Michael Dykes August 20, 1:20 pm ET

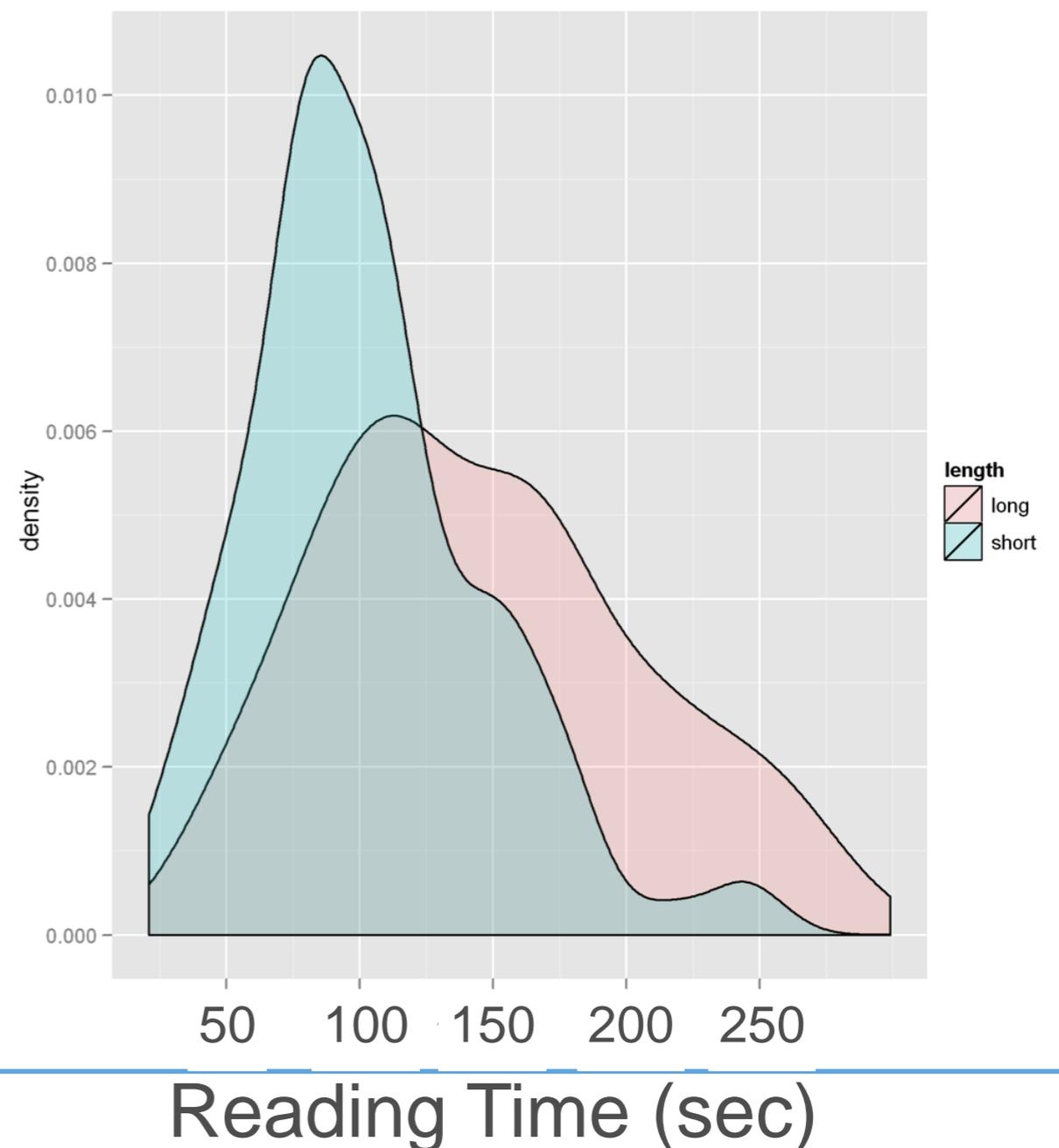
Page 2 (long)

Confounds

- Maybe increased whitespace makes ad more salient
- Manipulated ad exposure time through text length
 - Manipulation is endogenous, depends on participants reading speed
 - Say fast readers are sloppy so they miss ad
 - Say slow readers are careful so they see ad
 - Would look like manipulation had an effect, but it is spurious

Manipulation Check

- ▶ Average reading times:
 - ▶ short: 103 ± 6 sec
 - ▶ long: 146 ± 9 sec
- ▶ 30% someone in the short condition spent longer reading than someone in the long condition
- ▶ Manipulation is not strong enough

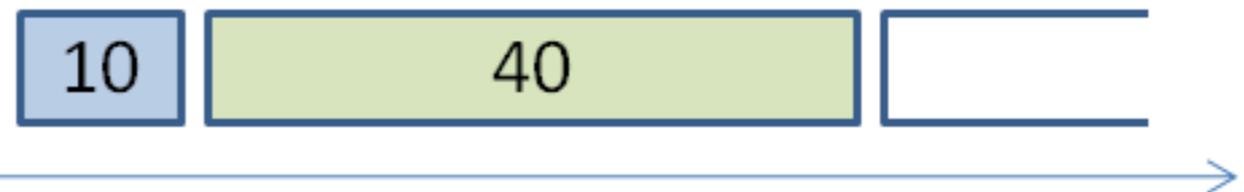


Time Based Display Advertising Design 2

(check your analysis)

Four Time Treatments

- ▶ Exogenously varied how long the ad was in view



- ▶ Users randomly placed in one of 4 treatments



- ▶ [Goldstein, McAfee, Suri '11]



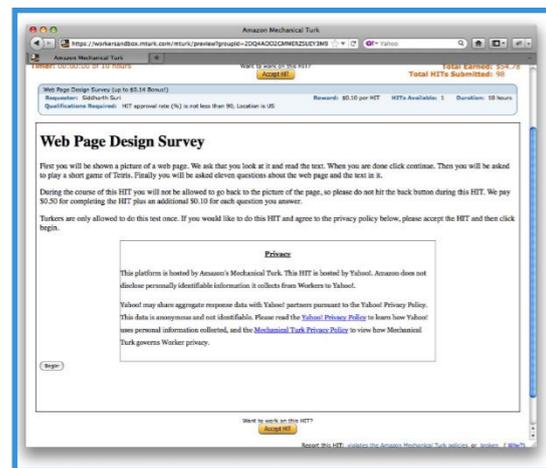
Time (seconds)

Two Ad Treatments

- ▶ Randomly varied which ad was shown first
- ▶ 4 x 2 design



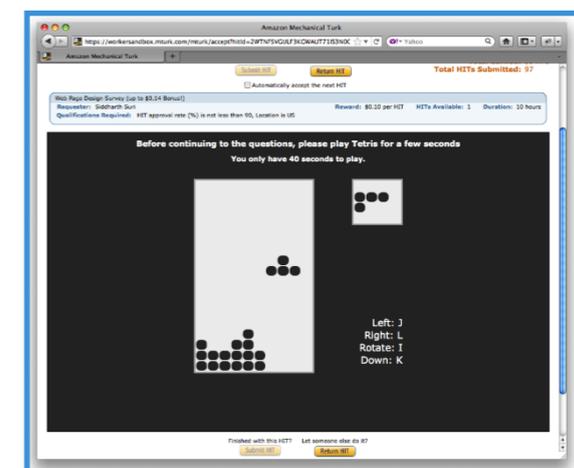
Participants' Perspective



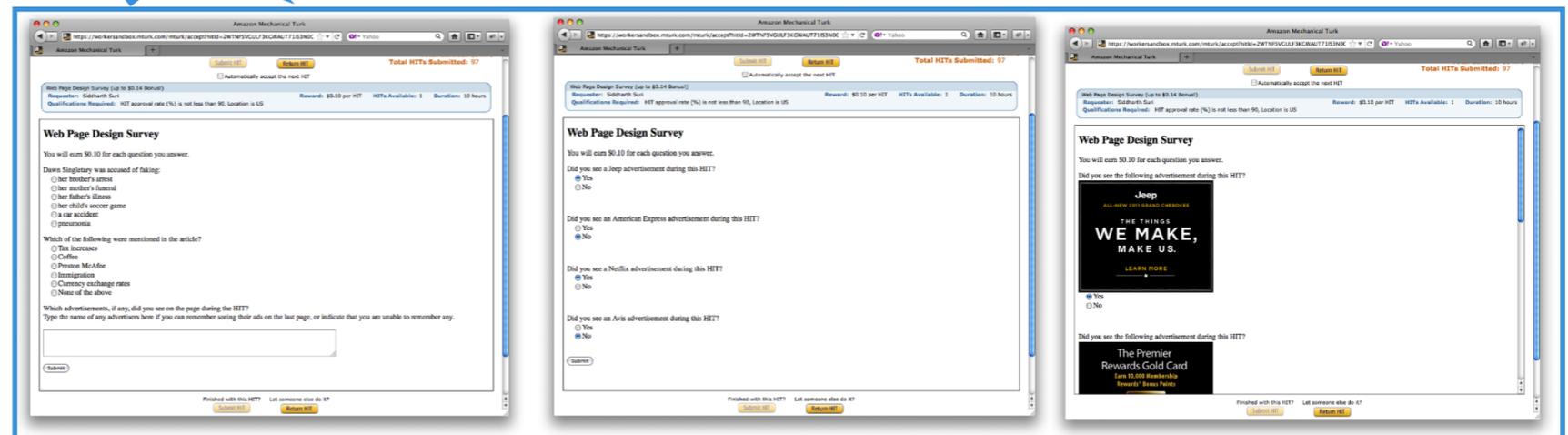
Instructions



Reading



Tetris



Quiz

9 Question Quiz

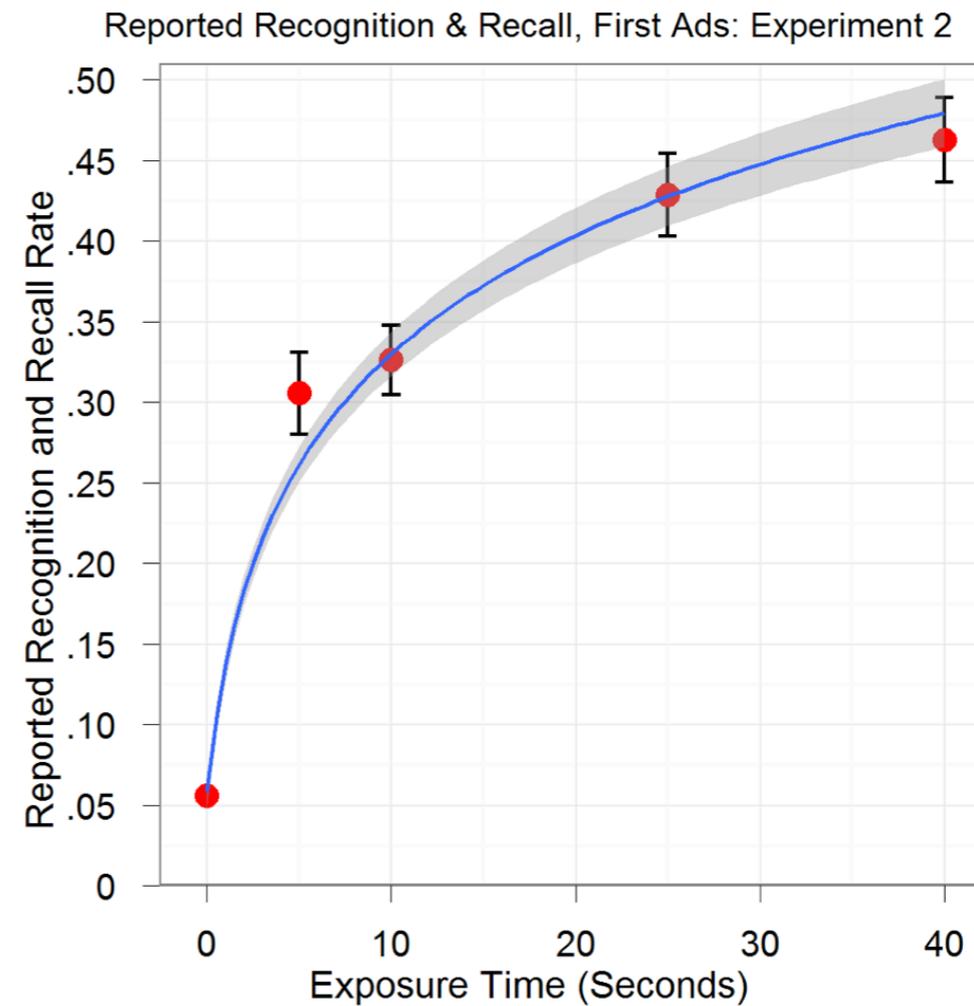
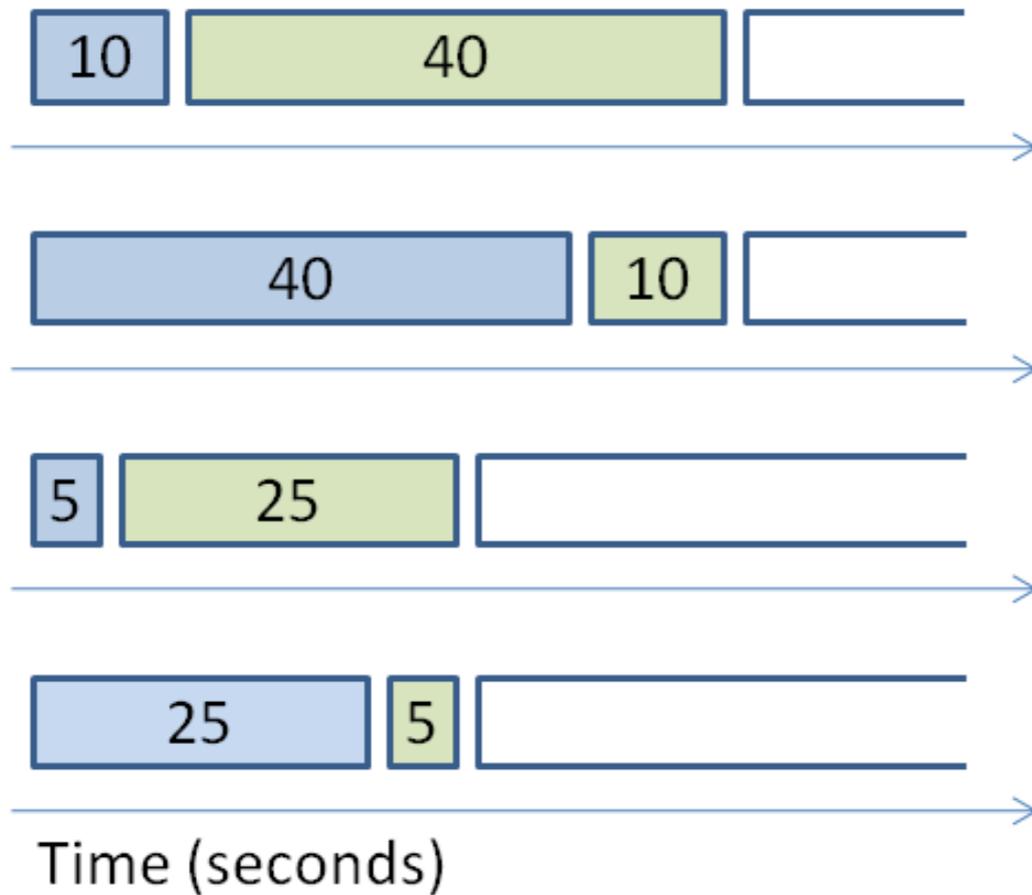
- ▶ Unaided recall: “Which advertisements, if any, did you see on the page during this experiment?”
- ▶ Text Recognition: “Did you see a Netflix/Jeep/Avis/Amex ad during this experiment?”
- ▶ Image Recognition: “Did you see the following ad during this experiment?”
- ▶ followed by image of a Netflix/Jeep/Avis/Amex ad
- ▶ Repeated measures can boost confidence



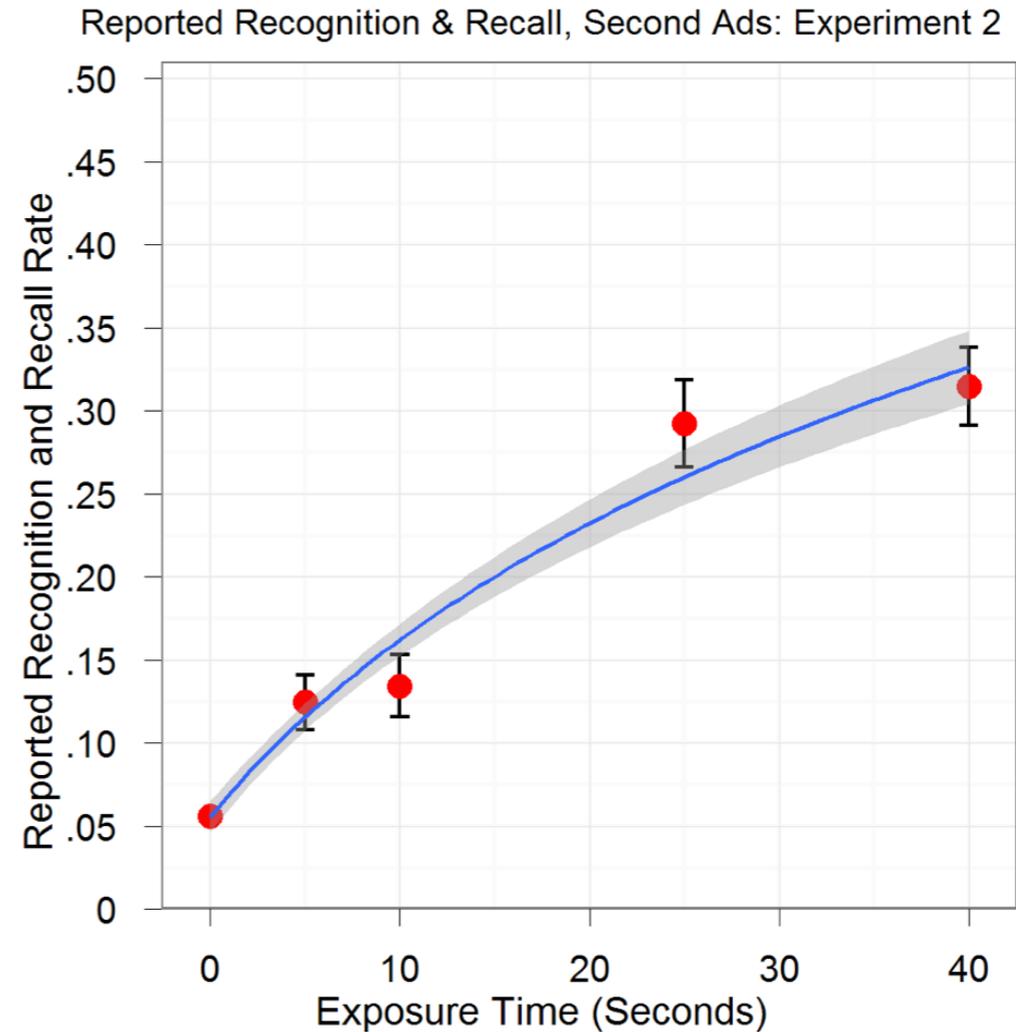
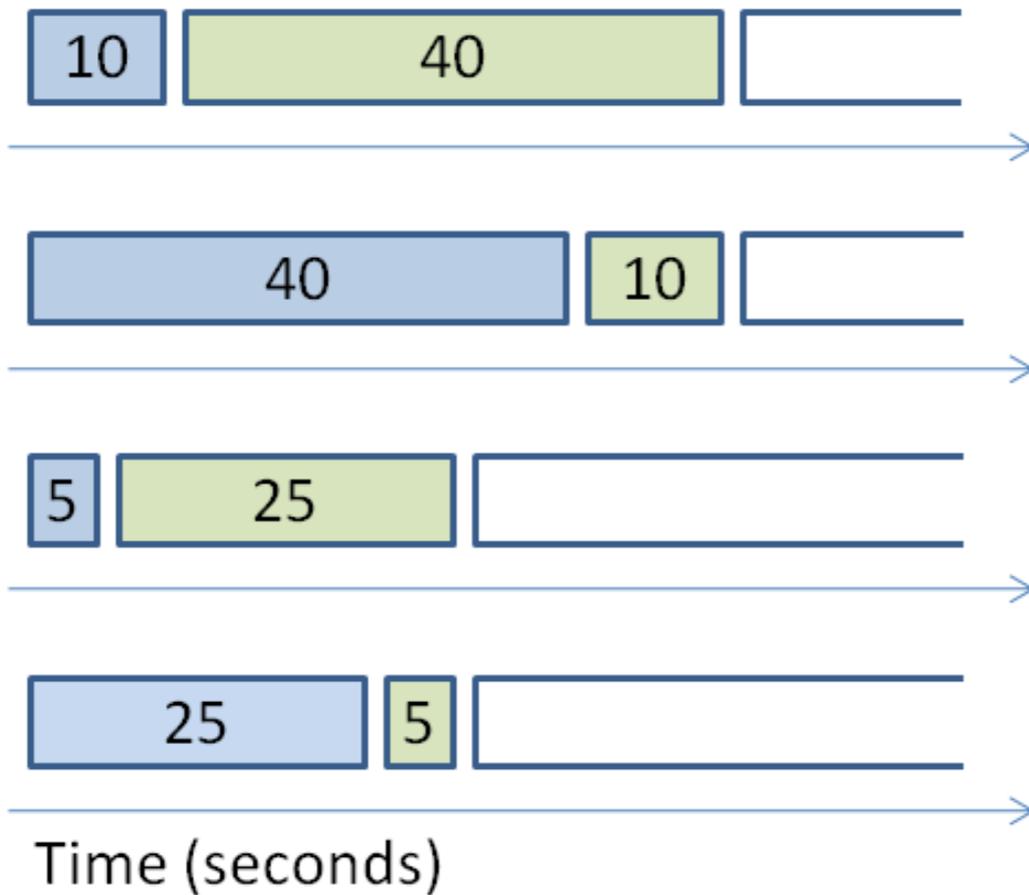
Ads



Lures



- ▶ 550 responses
- ▶ Curve for memory of first ad (all metrics combined)
- ▶ Clear impact of exposure time



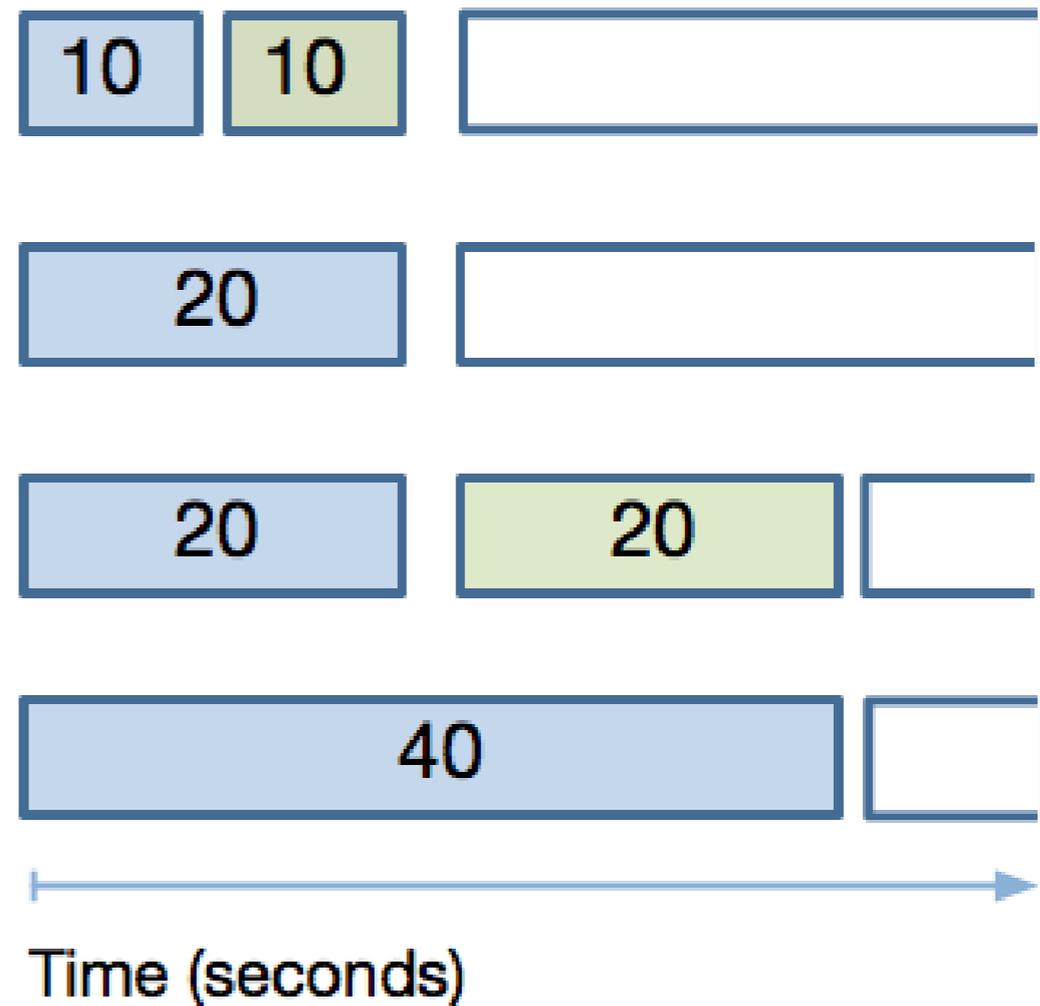
- ▶ 550 responses
- ▶ Curve for memory of second ad (all metrics combined)
- ▶ Confounded with onset time
- ▶ Know the analyses you want to run at design time.

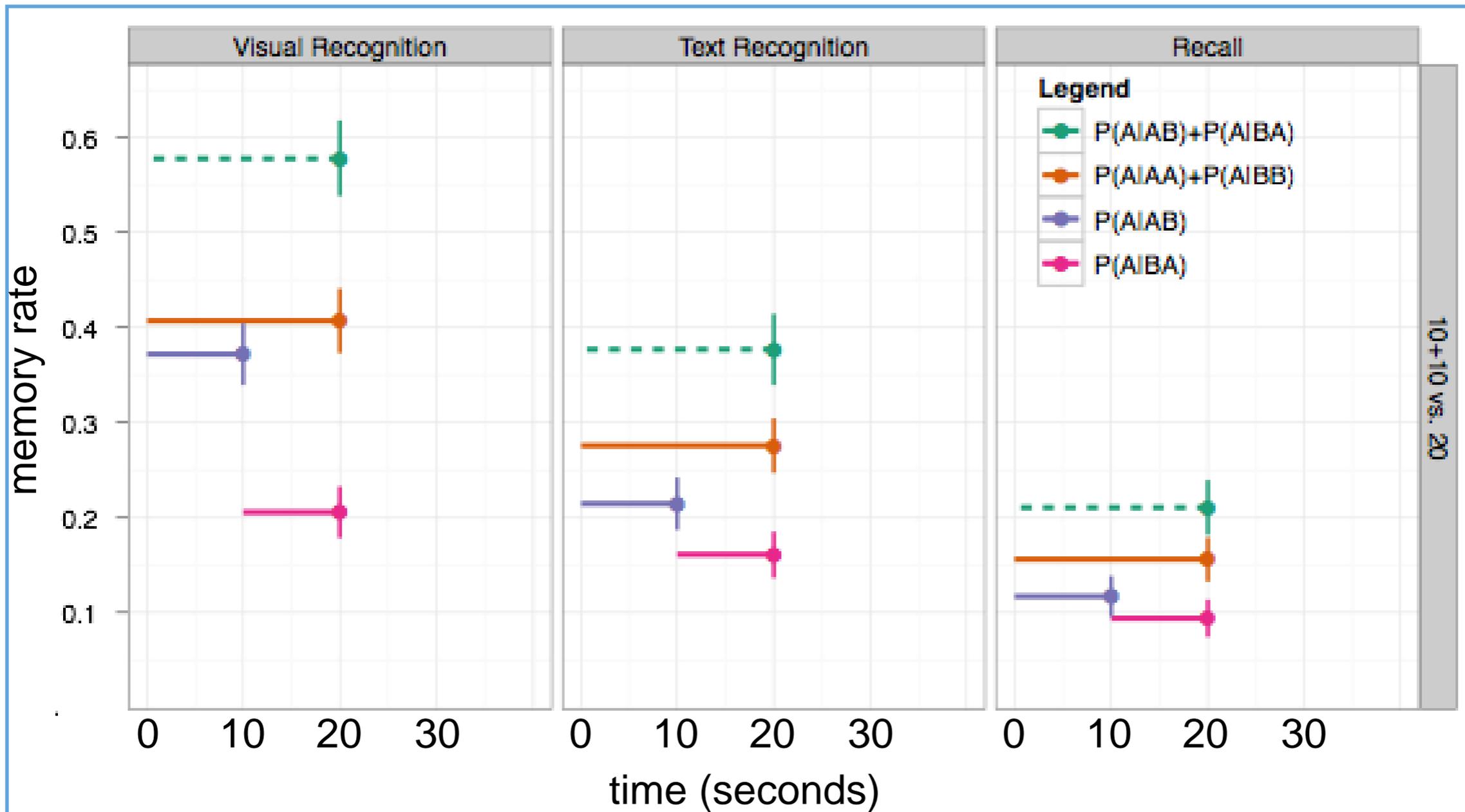
Time Based Display Advertising Design 3

(deceptively simple design)

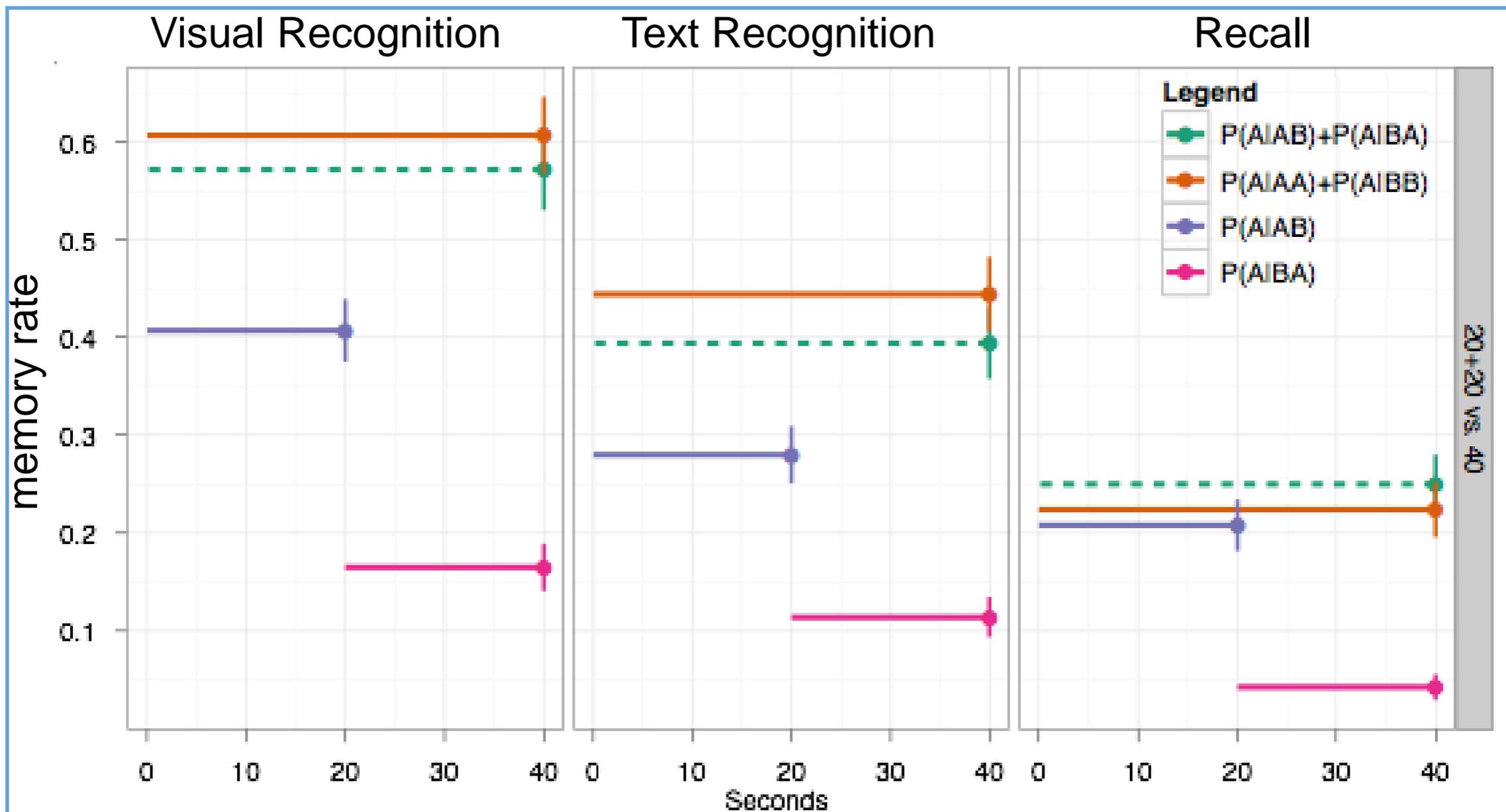
Four Time Treatments

- ▶ Exogenously varied ad duration
 - ▶ $t = 10, 20$
 - ▶ Showed one ad for $2t$ seconds
 - ▶ Showed two ads for t seconds
- ▶ Users randomly placed into one of four time treatments
- ▶ [Goldstein, McAfee, Suri '12]





- ▶ Left endpoint: when ad appeared, right endpoint: when ad disappeared
- ▶ Height: memory rate
- ▶ Onset time impacts memory



- ▶ Left endpoint: when ad appeared, right endpoint: when ad disappeared
- ▶ Height: memory rate
- ▶ Big effect of onset time

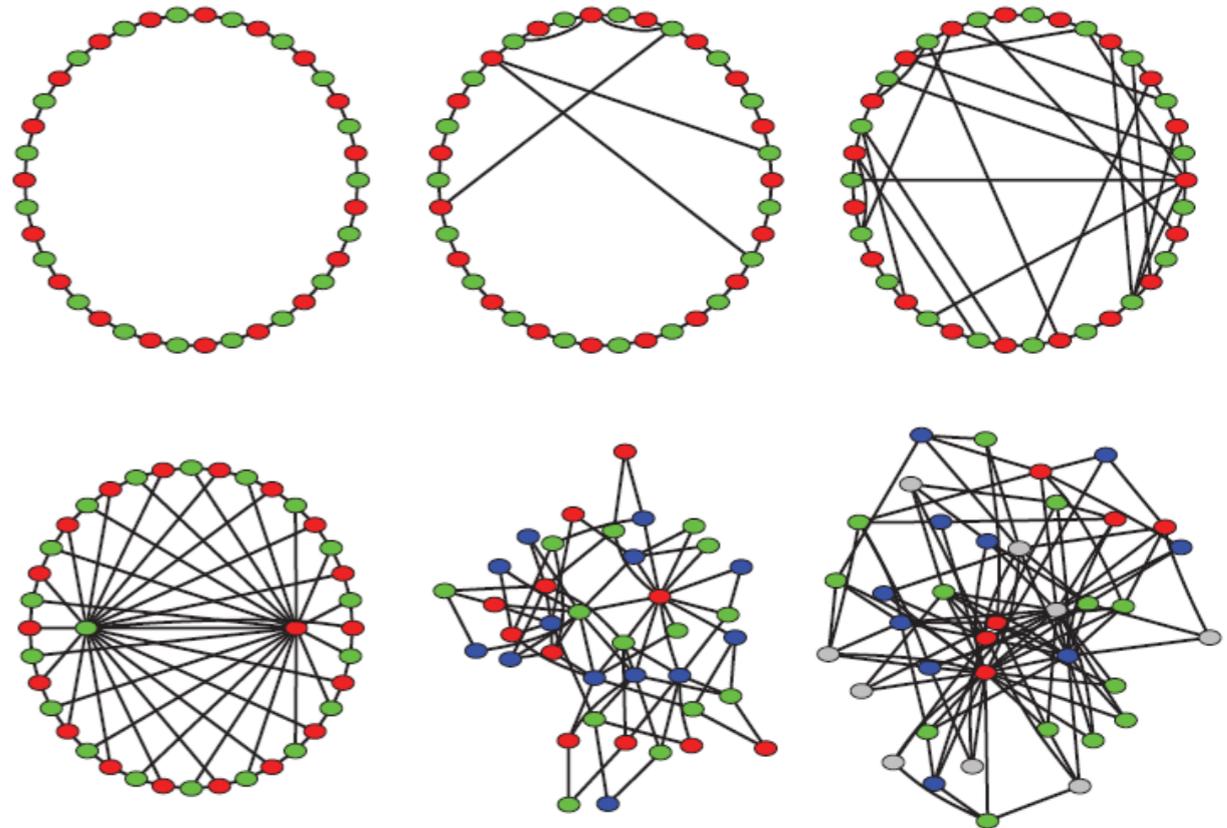
Algorithmic Graph Coloring

(be careful with methodology)

Graph Coloring

- Kearns, Suri, Montfort '06: graph coloring with humans
 - Each user controls one node
 - Different topologies/incentives affect time to solve the graph
- Graph coloring is a distributed constraint satisfaction problem, solvable with algorithms from AI

Research Question: Can we get people to color a graph faster by using algorithmic “hints”?



Adapting Algorithms for People

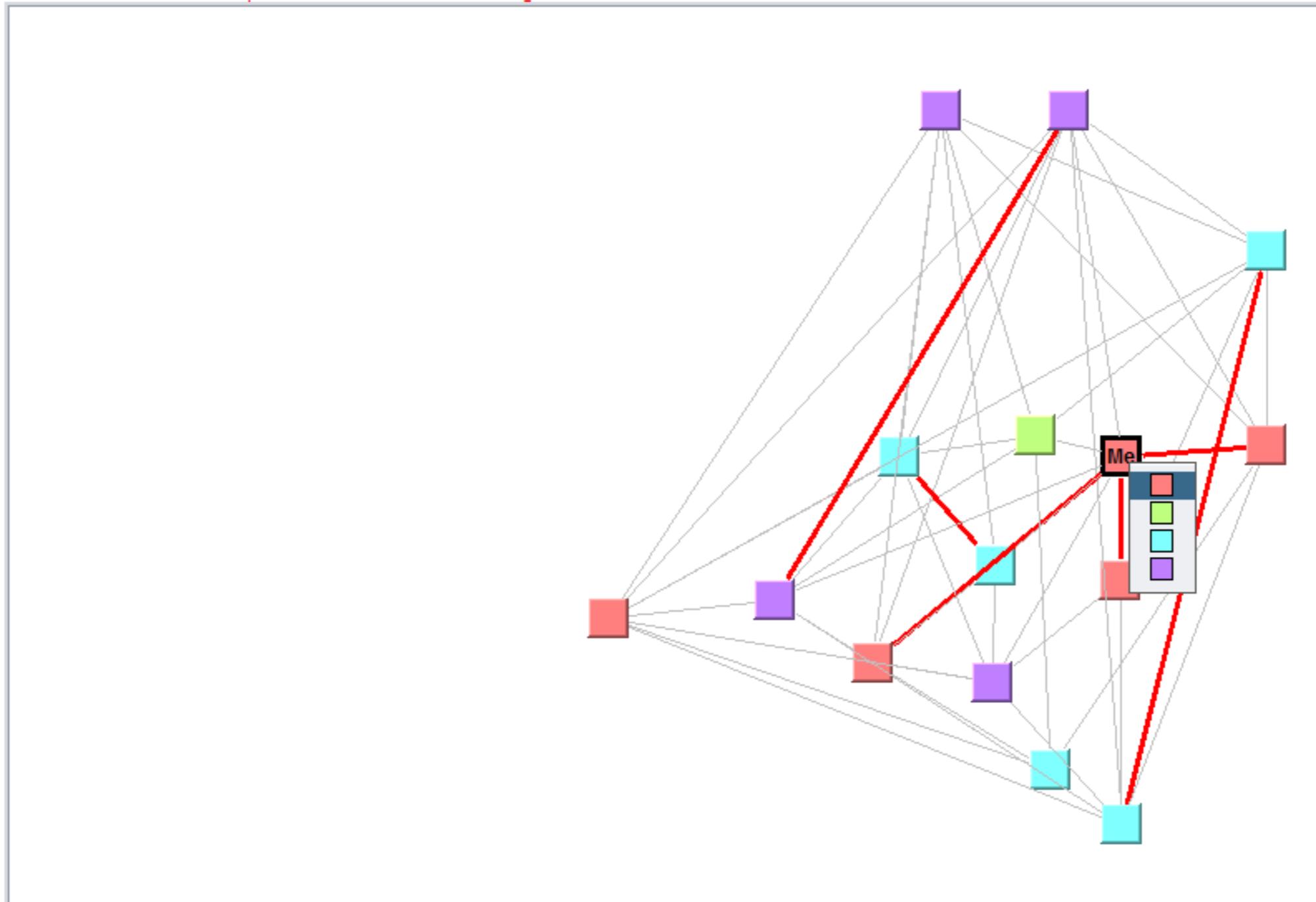
Humans:

- Don't always follow instructions
- Can communicate or signal to one other, causing side effects
- May become frustrated or lose focus if confused

Thus, the algorithmic “hints” should

- be robust to deviations or independent actions – or even take advantage of them
- show information that is easy to understand and act upon

User Interface



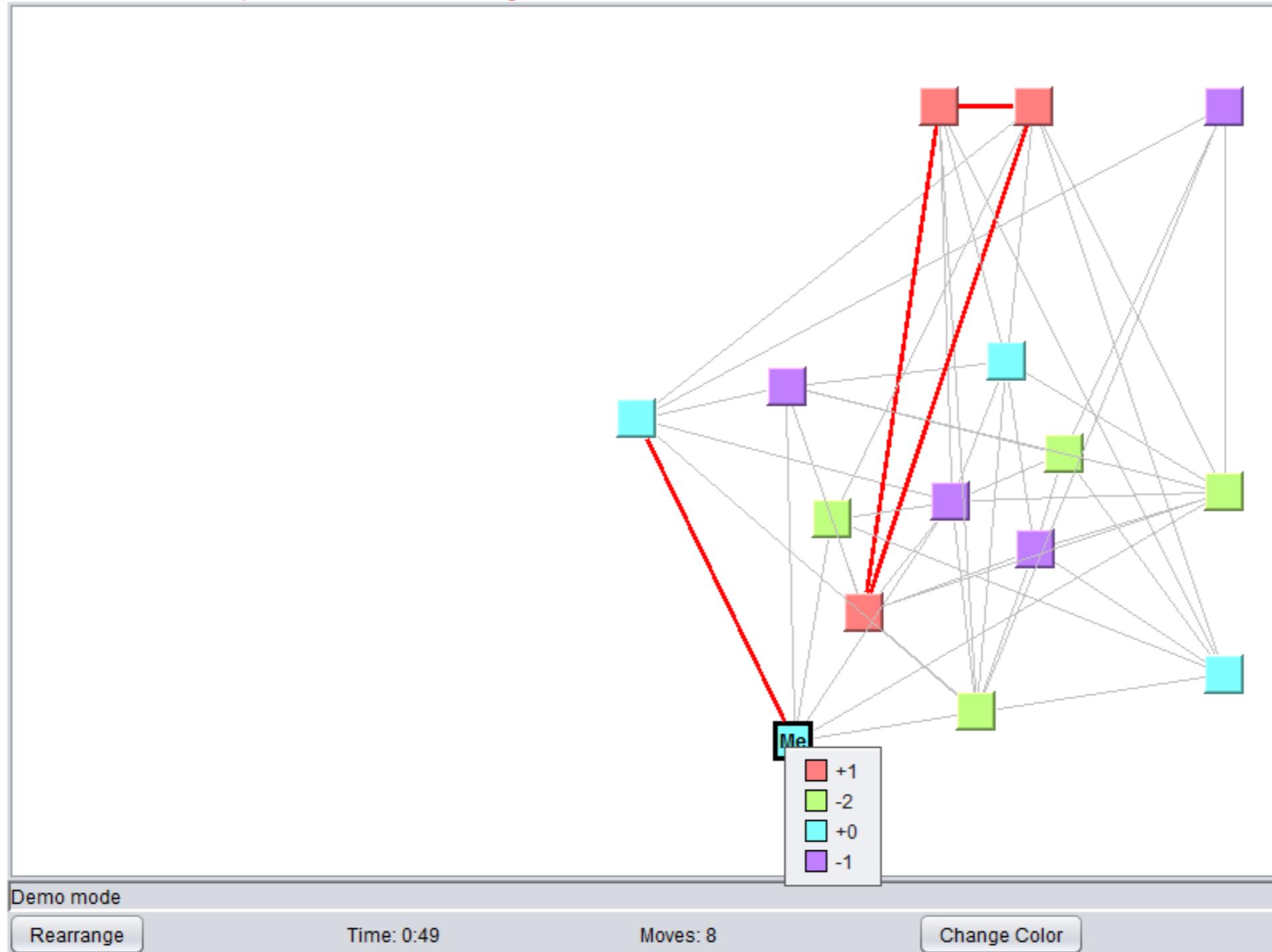
Rearrange Time: 1:01 Moves: 10 Change Color

Demo mode

Algorithmic Idea: Distributed Breakout

- Each constraint is assigned a weight
 - “Objective function”: Sum of weights of violated constraints
- Score for an action:
 - weights of violated constraints it would resolve, minus the weights of new violations it creates
 - improvement in objective
- If no action has a positive score, increase the weight of all violated constraints

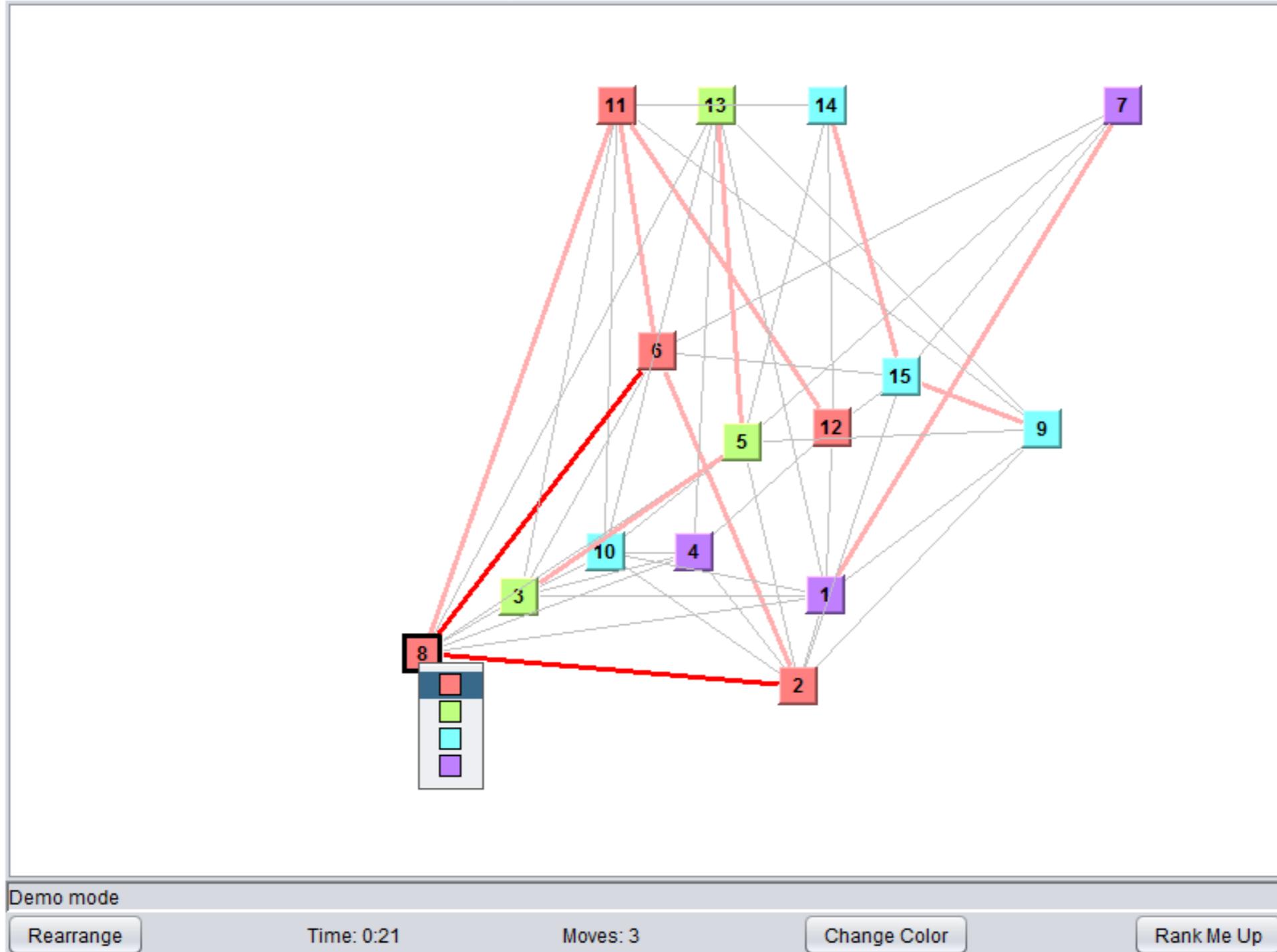
DB Interface



Algorithmic Idea: Asynchronous Weak Commitment

- Each agent has a rank from 1..n, where a lower number is higher rank
- Choose an action that doesn't conflict with higher-ranked neighbors
 - If possible, minimize conflicts with lower-ranked neighbors
- If no choices work, 'grab' the highest rank among all neighbors

AWC Interface



Recruiting Participants

- Experiment requires 15-20 people to be present at the same time, waiting in a lobby
- If waiting around for too long, they get bored and leave

Solution:

- Run smaller games that are easy to fill, and give an “opt-in” for future experiments
- Send an e-mail out to notify workers of an experiment “session”
- Can also reiterate specific instructions for the session in the e-mail

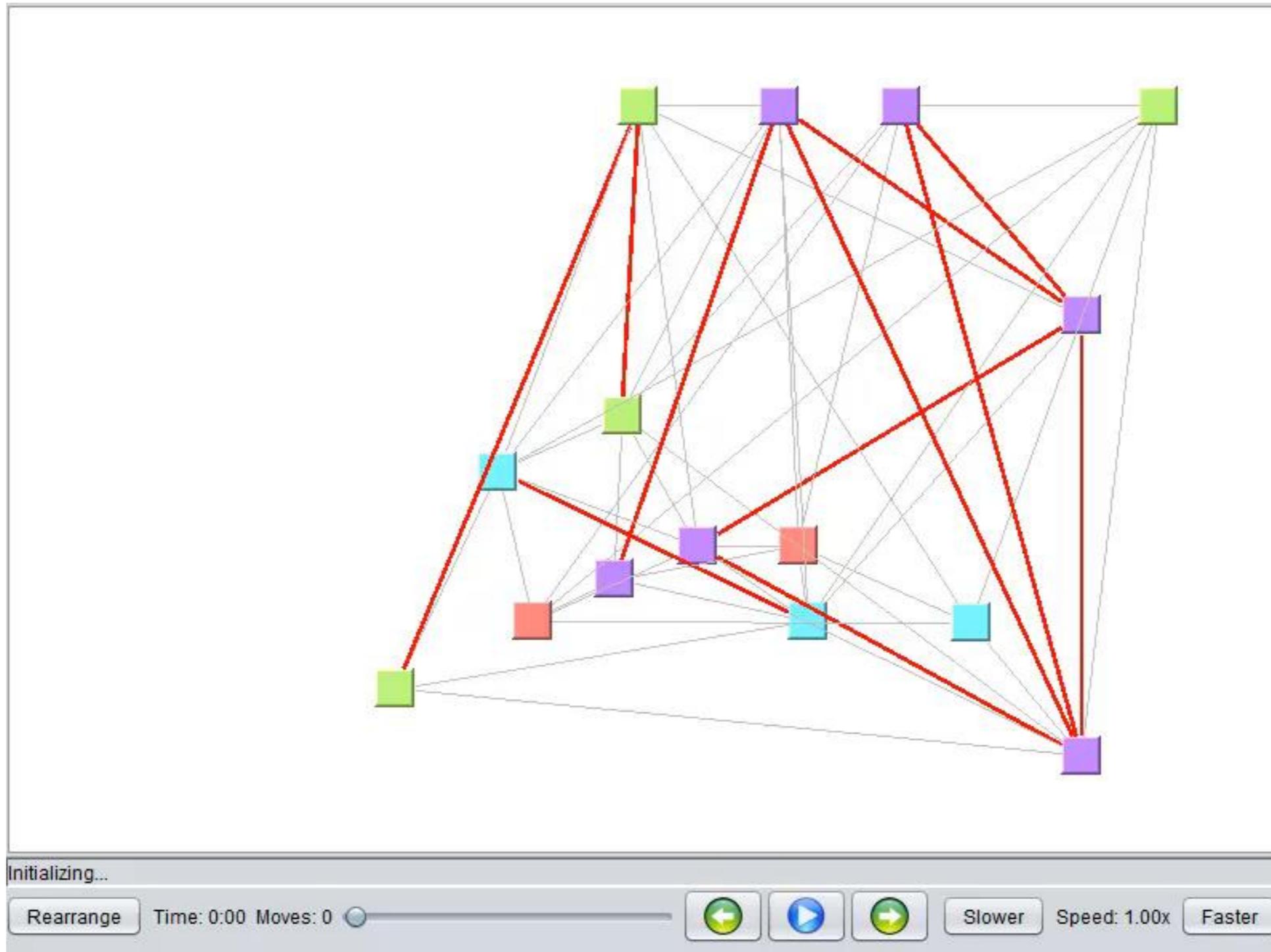
A Suitable Experiment on MTurk

- Fast and high throughput
 - After recruitment, large experiment sessions easy to schedule
 - 30 experiments of 15 people each: ~20 minutes
- Cheap
 - At \$0.20 per HIT, games cost \$3.00 each
 - Compare to lab experiments: \$10 or more per participant, per hour
- Controlled participants
 - Users are trained on the game before controlled experiments

Experiment Design

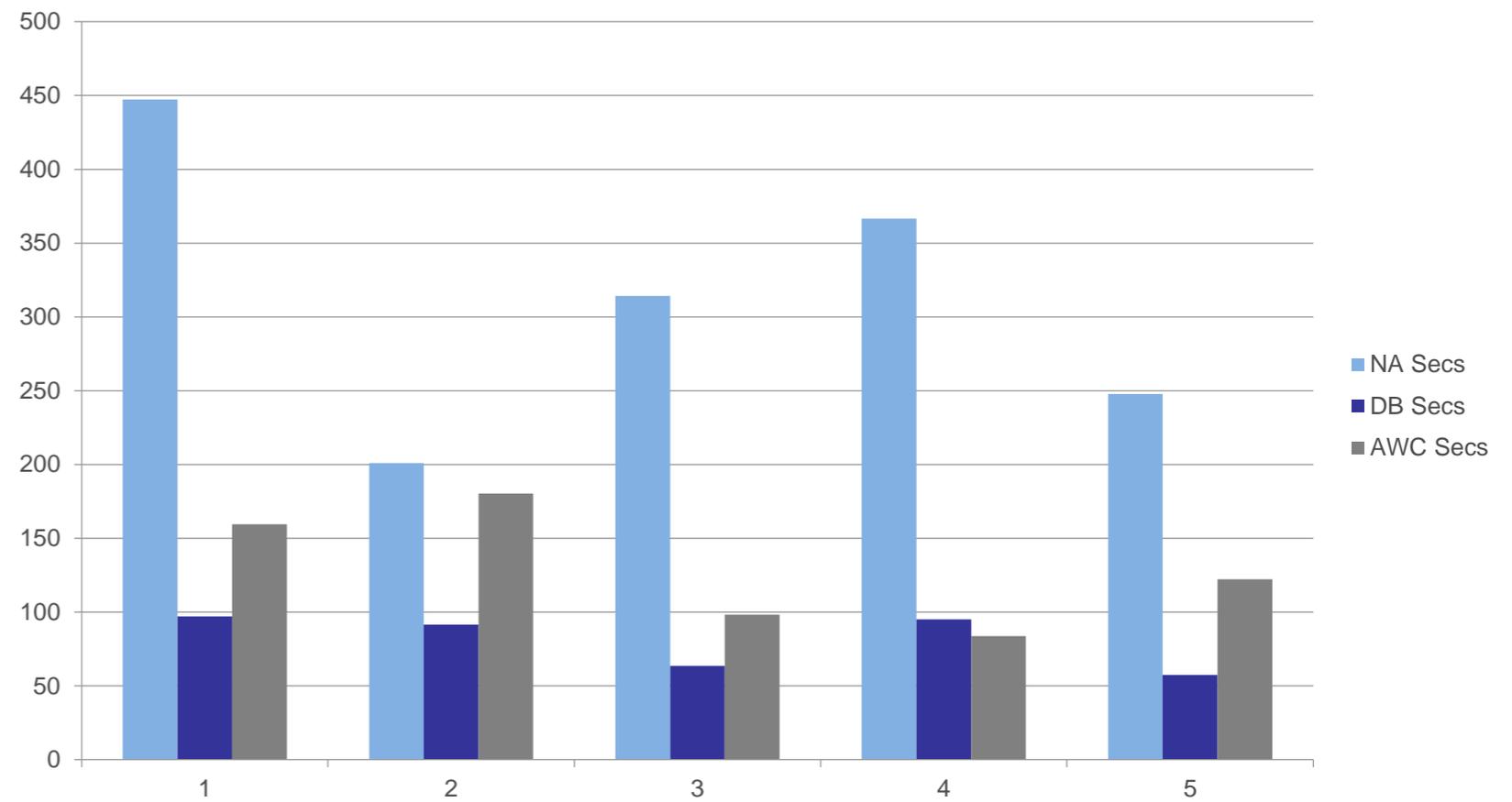
- Five pre-generated graphs:
 - 15 vertices, 50 edges, 4-colorable
 - fixed initial assignment of colors
- For each experiment:
 - graph gets random permutation of original coloring
 - players randomly assigned to vertices
- For each player:
 - List of color choices is randomly shuffled
 - Random layout of the graph on screen

Example

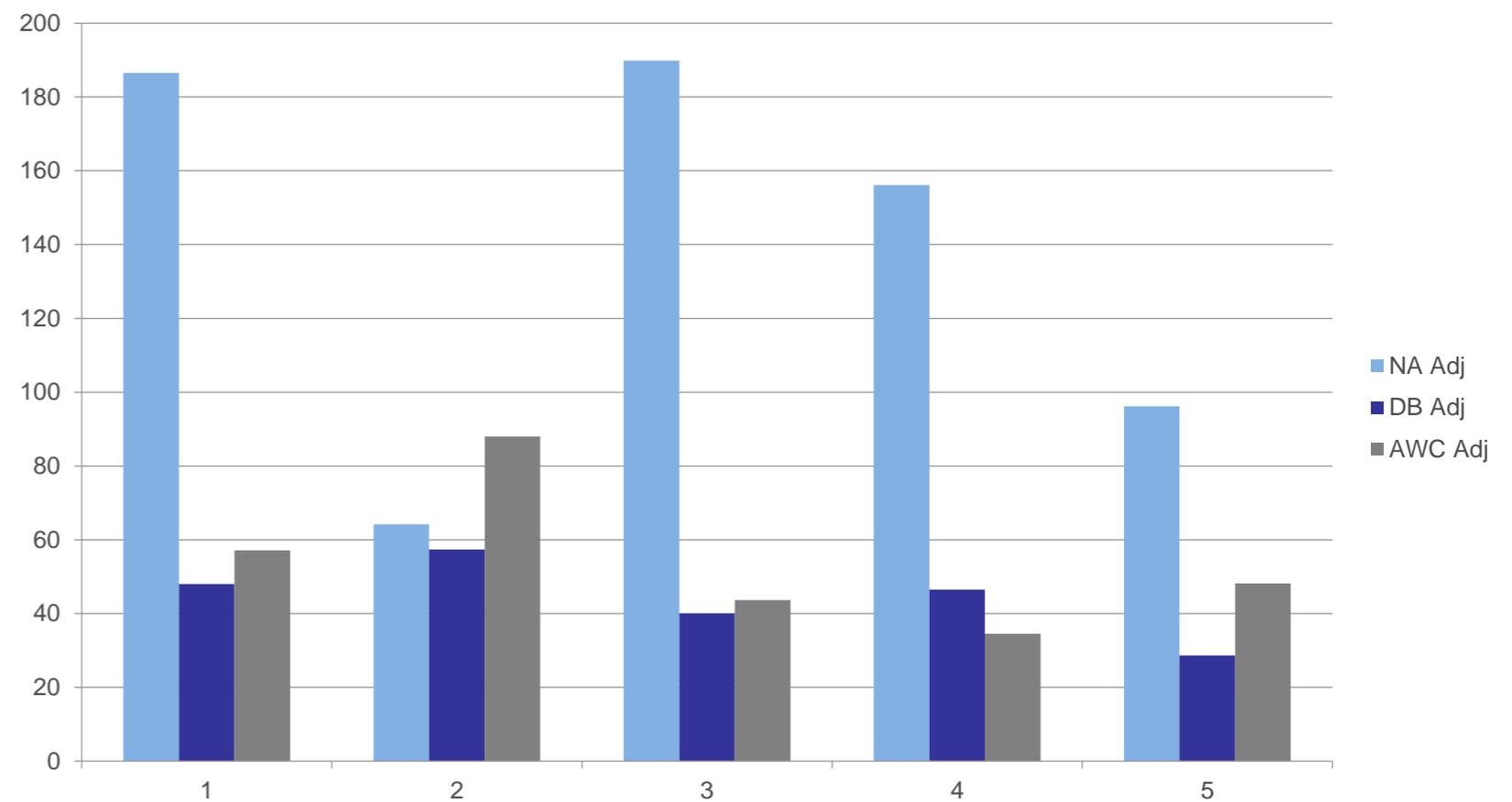


Initial Results

Average Solve Time



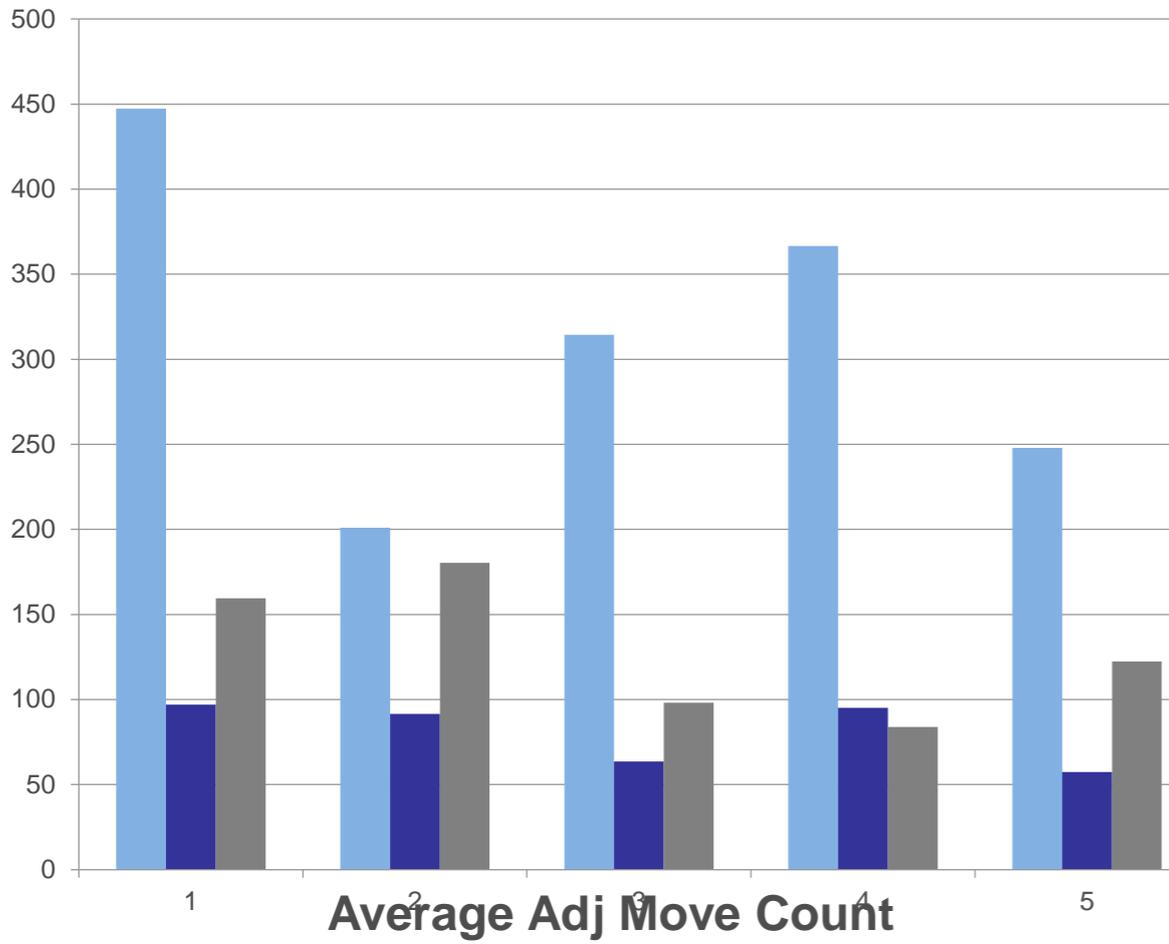
Average Adjusted Move Count



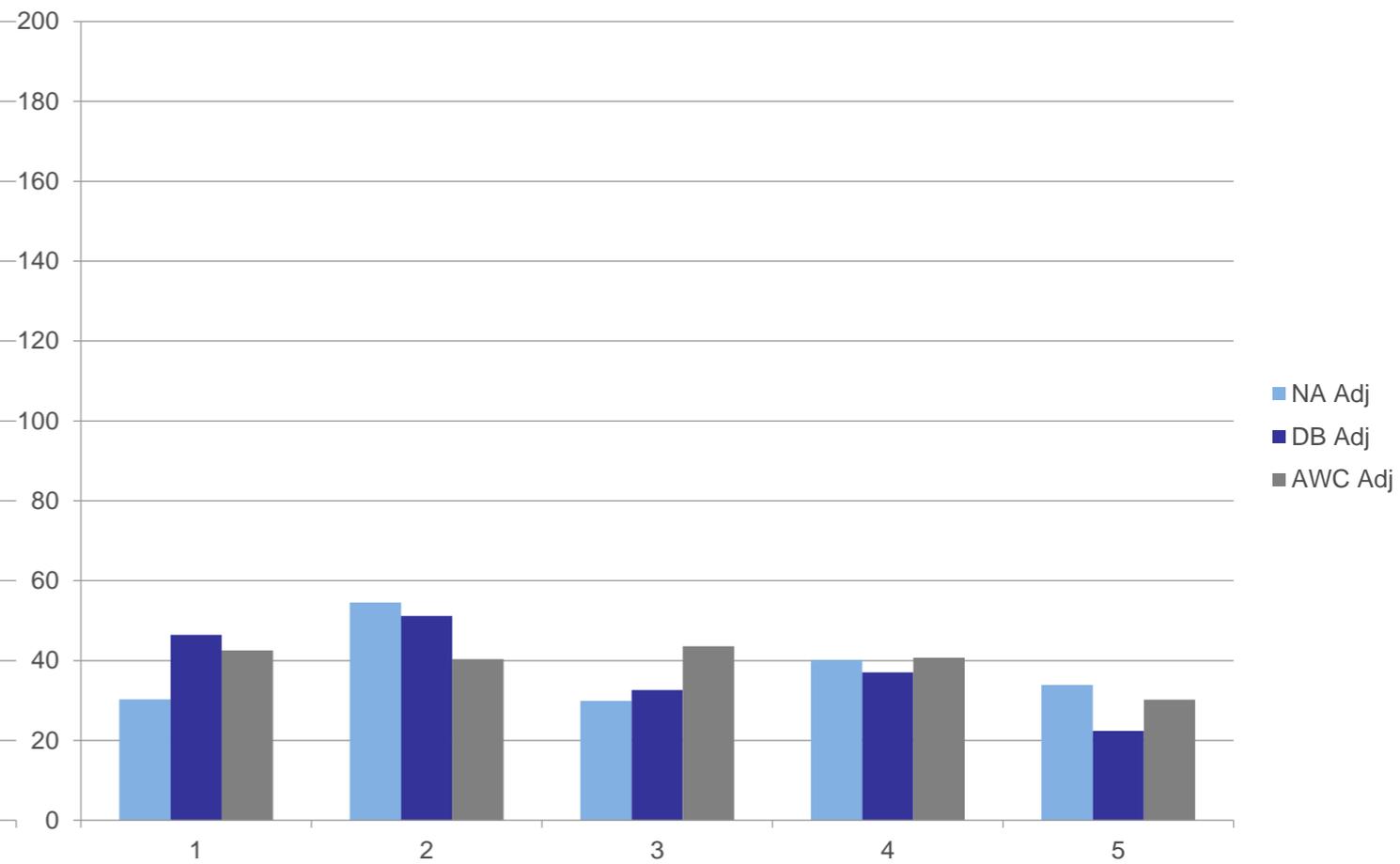
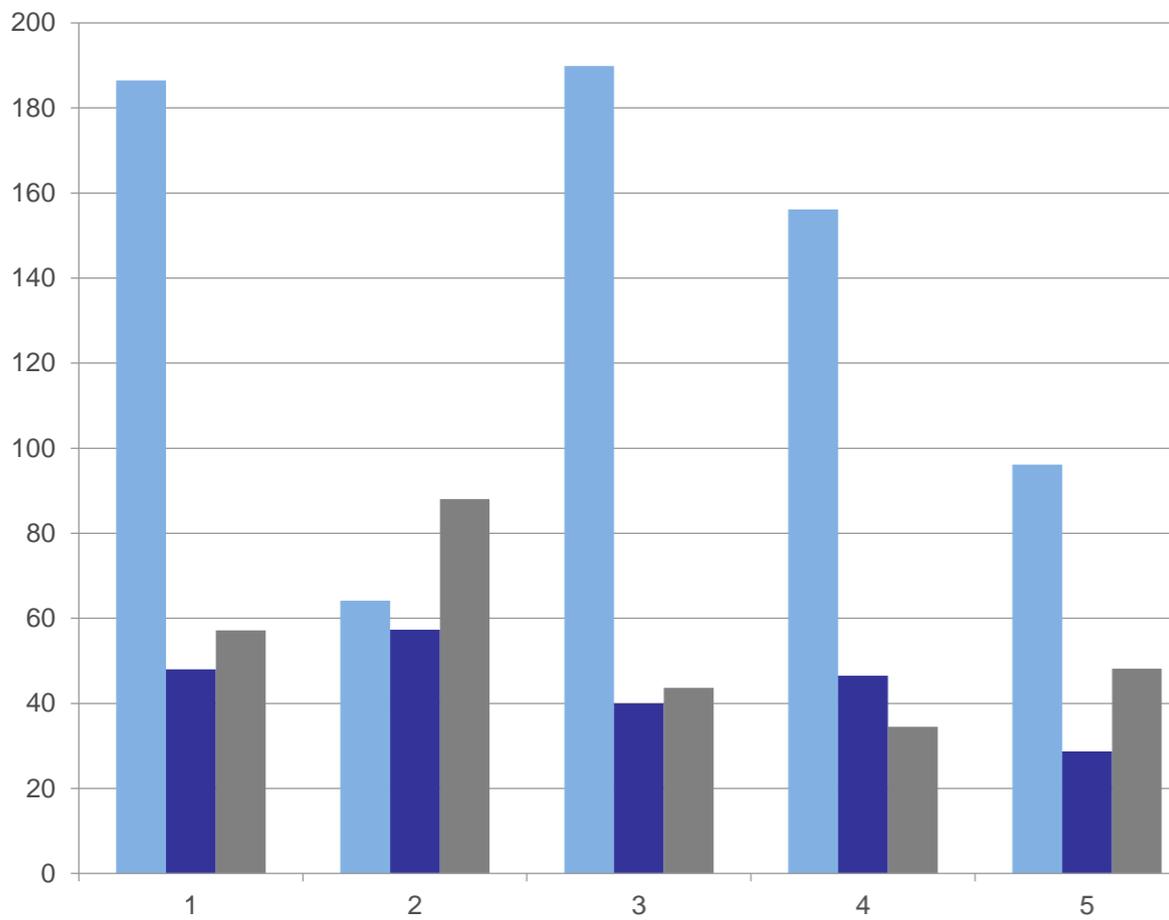
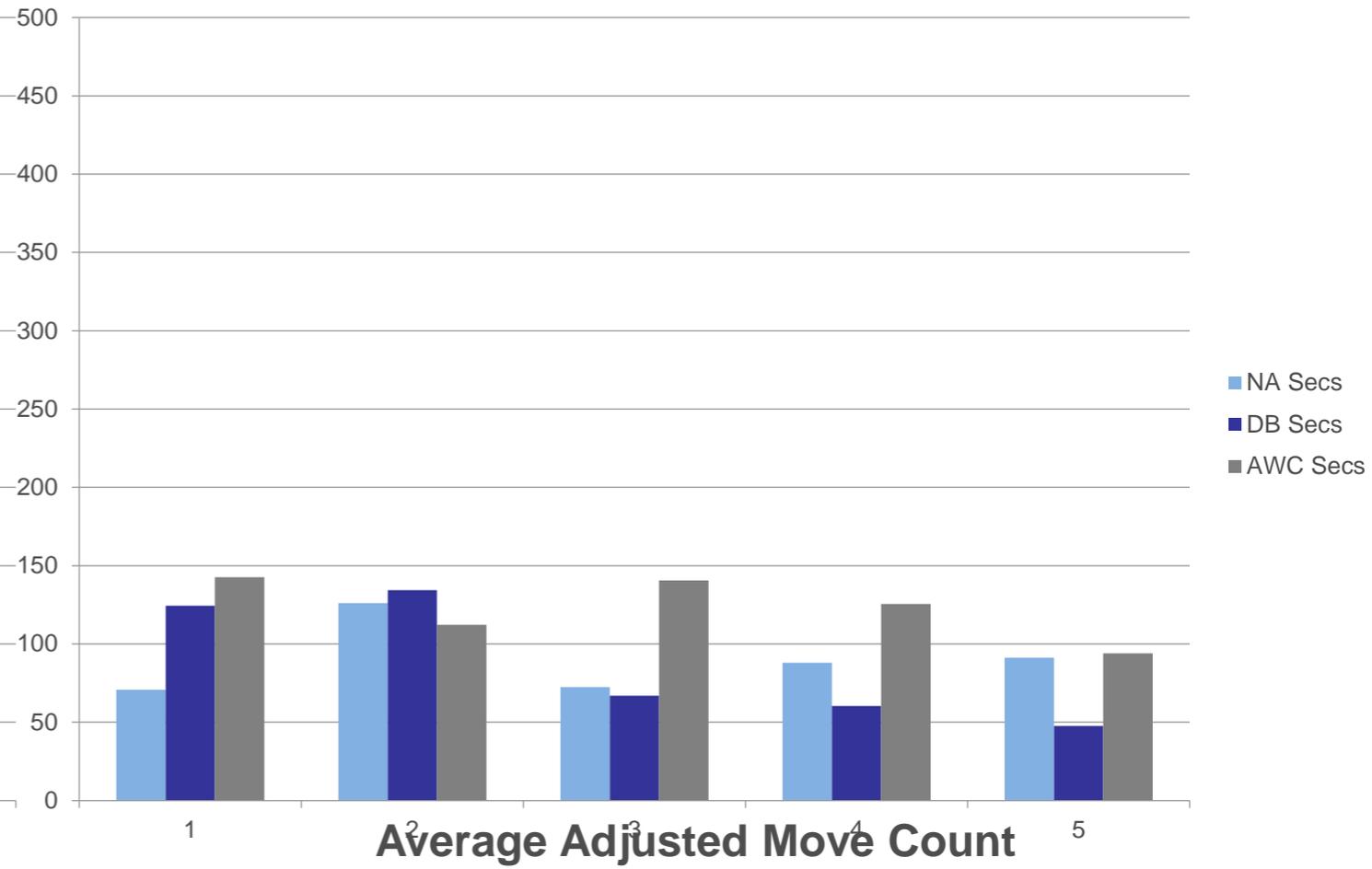
Too good to be true?

- For normal Turk tasks, the morning and afternoon (EST) are the busiest
 - Mix of Indian and US workers in the morning
 - Primarily US workers in the afternoon
 - We didn't use region controls and ran at different times of day. Does this affect us?
- New controls:
 - US-registered workers only
 - Sessions at same time every day
 - Force people to pass a quiz about the instructions

Average Solve Time



Average Solve Time



What did we learn?

- Treatment order was not randomized and run with different populations
 - Region and time-of-day controls are very important
 - When networked, latency and bandwidth affect performance: higher latency and lower bandwidth to India
 - Indian workers may react differently to English instructions
- Poor choice of dependent variable: time and num. moves are very noisy
- Learning effect: workers start to use their own heuristic after repeated participation
- Using quiz increases adherence to instructions

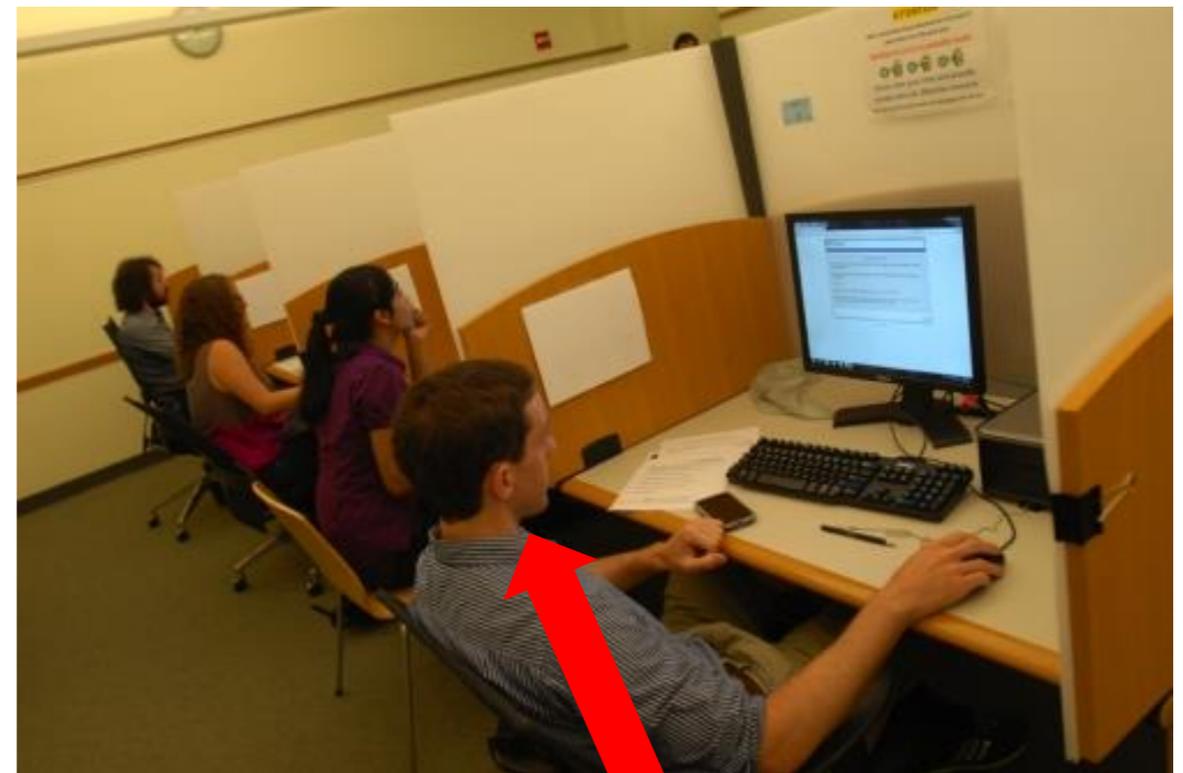
Executing Online Experiments

Where Can We Conduct (Causal) Experiments?

- Behavioral experiment labs in universities
- Field experiments (in real settings)
 - Health care, education, fundraising, blood/organ donation
 - Experiments on websites: e.g. eBay
- **Recruiting subjects via the Internet**
 - Crowdsourcing & online labor markets
 - Custom web sites and web applications

The Behavioral Experiment Lab

- Great control and monitoring
- Subjects are WEIRD:
 - Western
 - Educated
 - Industrialized
 - Rich
 - Democratic
- Not necessarily a representative sample of populations we may want to study
 - Can result in external validity issues



undergrad

Field Experiments

- Field experiment: randomized study conducted in a real-world setting.
 - Subjects often don't even know they are in an experiment
 - Can be done online, for example:
 - A/B testing employed in many web sites (Google, Bing, Facebook)
- + Studies behavior in a more natural environment
- + Allows potentially larger scale than a lab
- Can be hard to find the right setting and data
- Possible sacrifice of control and enforcement

Online Experiments: Recruiting from the Internet

Potentially combines benefits of lab and field experiments:

- Access to a larger or more diverse population
- Connect many more people together than can fit in a lab
- Allows for good control, yet flexibility in experiment design
- For online systems, studies the natural environment



Amazon Mechanical Turk

- Also known as MTurk; widely gaining acceptance for
 - Surveys: reaching a large number of people
 - Social science (esp. psychology studies *)
 - HCI experiments: used to test different interfaces
- Demographics and meta-studies are well-known
- Established techniques for conducting experiments (Ipeirotis '10, Mason & Suri '12)

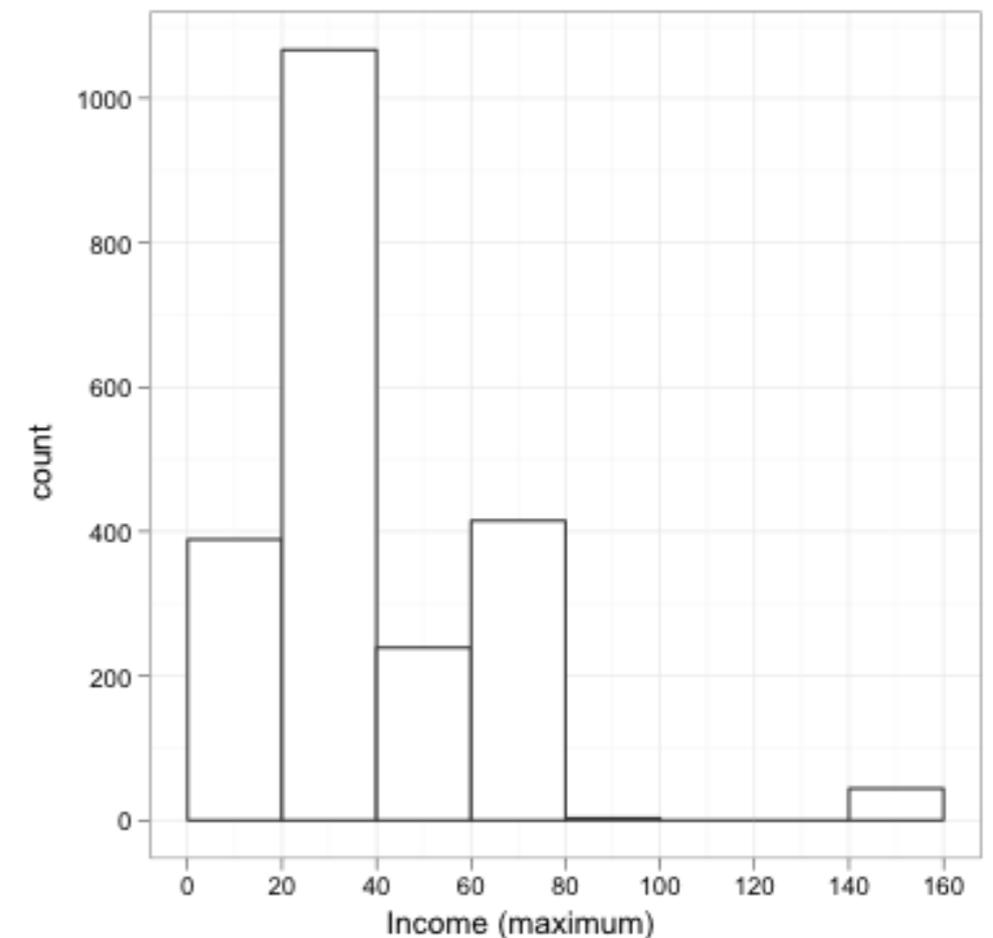
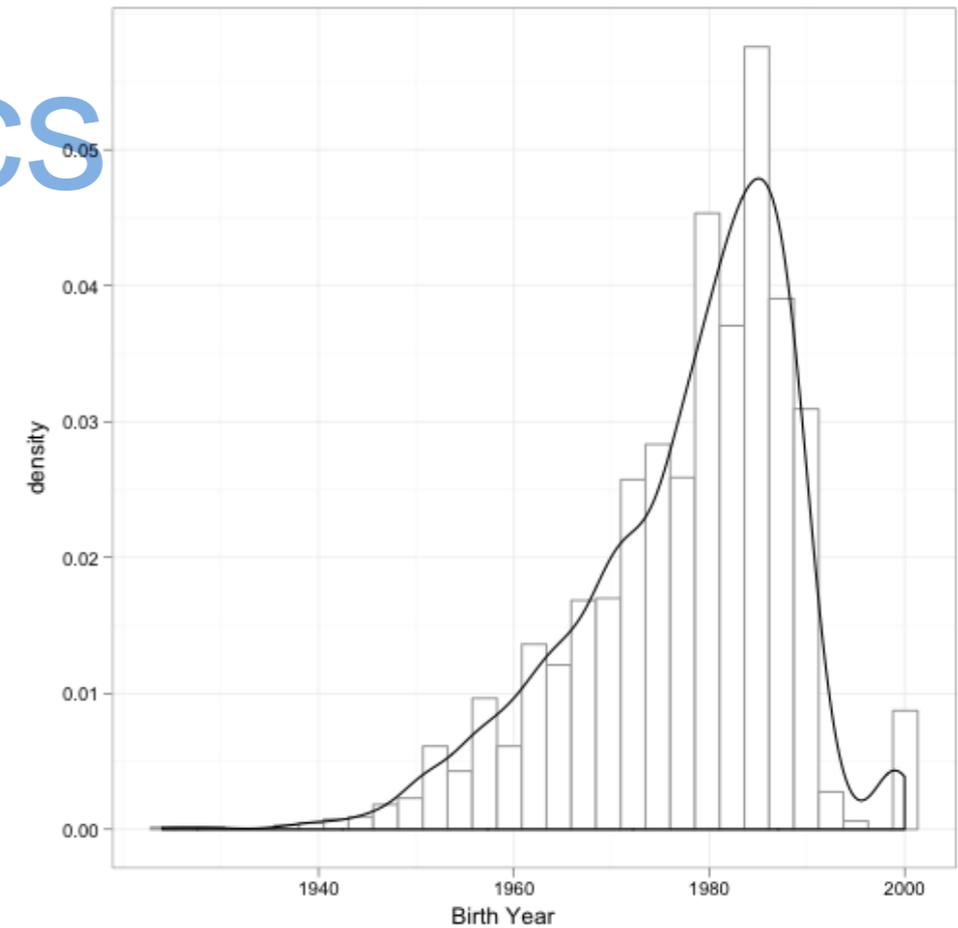
* Buhrmester et al. "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6.1 (2011): 3-5.

Why Mechanical Turk?

- Subject pool size
 - Central place for > 100,000 workers (Pontin '07)
- Subject pool diversity
 - Open to anyone *globally* with a computer, internet connection
- Low cost
 - Reservation Wage: \$1.38/hour (Chilton et al '10)
 - Effective Wage: \$4.80/hour (Ipeirotis, '10)
- Faster theory/experiment cycle
 - Less coordination than getting subjects in a lab
 - Not on semester cycle

Worker Demographics

- Self reported demographic information from 2,896 workers over 3 years (MW '09, MW '11, SW '10)
- 55% Female, 45% Male
 - Similar to other internet panels (e.g. Goldstein)
- Age:
 - Mean: 30 yrs,
 - Median: 32 yrs
- Mean Income: \$30,000 / yr
- Similar to Ipeirotis '10, Ross et al '10



Internal Consistency of Demographics

- 207 out of 2,896 workers did 2 of our studies
 - Only 1 inconsistency on gender, age, income (0.4%)
- 31 workers did ≥ 3 of our studies
 - 3 changed gender
 - 1 changed age (by 6 years)
 - 7 changed income bracket
- Strong internal consistency

Turker Community

Asymmetry in reputation mechanism:

- Reputation of Workers given by approval rating, # tasks done
 - Requesters can reject work
 - Requesters can refuse workers with low approval rates
- Requester reputation of is not built in to MTurk
 - **TurkOpticon** (browser plugin): Workers rate requesters
 - TurkerNation, mTurkForum, [reddit.com/r/mturk](https://www.reddit.com/r/mturk/): HIT notifications and discussion for workers

Anatomy of a HIT

- **HITs** (human intelligence tasks) are the basic unit of work on MTurk
- HITs are broken up into **Assignments**
- A worker cannot do more than 1 assignment of a HIT

Amazon Mechanical Turk - All HITs

Search for containing that pay at least \$ for which you are qualified

All HITs
31-40 of 1765 Results

Sort by: [Show all details](#) | [Hide all details](#) [First](#) << [Previous](#) < [2](#) [3](#) [4](#) [5](#) [6](#) > [Next](#) >> [Last](#)

Telugu to English Translation Requester: Chris Callison-Burch	HIT Expiration Date: Jan 2, 2011 (12 weeks 5 days) Time Allotted: 24 hours	Reward: \$0.70 HITs Available: 663	Request Qualification Take Qualification test (Why?) View a HIT in this group
Given URL, find the company name Requester: TurkUser12345	HIT Expiration Date: Oct 11, 2010 (6 days) Time Allotted: 10 minutes	Reward: \$0.01 HITs Available: 594	View a HIT in this group
Acquire Information From Tax Assessment Database Requester: ProPublica	HIT Expiration Date: Oct 7, 2010 (2 days 3 hours) Time Allotted: 30 minutes	Reward: \$0.02 HITs Available: 542	Take Qualification test (Why?) View a HIT in this group
Evaluate title-to-function mappings Requester: Crowd Task	HIT Expiration Date: Oct 19, 2010 (2 weeks) Time Allotted: 20 hours	Reward: \$0.50 HITs Available: 534	Request Qualification (Why?) View a HIT in this group
Finden Sie den Firmennamen und die Kontaktinformationen für die genannte Website heraus. Requester: Amazon Requester Inc. (Product Ads)	HIT Expiration Date: Oct 18, 2010 (1 week 5 days) Time Allotted: 60 minutes	Reward: \$0.05 HITs Available: 503	Not Qualified to work on this HIT (Why?) View a HIT in this group
Rank Machine Translation Output (for speakers of English) Requester: WMT Admin	HIT Expiration Date: Nov 1, 2010 (3 weeks 6 days) Time Allotted: 60 minutes	Reward: \$0.05 HITs Available: 500	Not Qualified to work on this HIT (Why?) View a HIT in this group

Anatomy of a HIT

- **HITs** (human intelligence tasks) are the basic unit of work on MTurk
- HITs are broken up into **Assignments**
- A worker cannot do more than 1 assignment of a HIT

The screenshot displays the Amazon Mechanical Turk interface. At the top, the browser address bar shows the URL: <https://www.mturk.com/mturk/preview?groupId=15M5U73SZPDX5H3KVYF2R92>. The page header includes the Amazon Mechanical Turk logo, navigation tabs for 'Your Account', 'HITs', and 'Qualifications', and a notification for '106,201 HITs available now'. The user's name 'Sid Suri' and links for 'Account Settings', 'Sign Out', and 'Help' are visible in the top right.

The main content area features a search bar with 'HITs' selected, a search filter 'containing' followed by an empty input field, and a filter for 'that pay at least \$ 0.00 for which you are qualified'. Below the search bar, there is a timer 'Timer: 00:00:00 of 2 hours' and two buttons: 'Accept HIT' and 'Skip HIT'. On the right side, the 'Total Earned: \$0.27' and 'Total HITs Submitted: 4' are displayed.

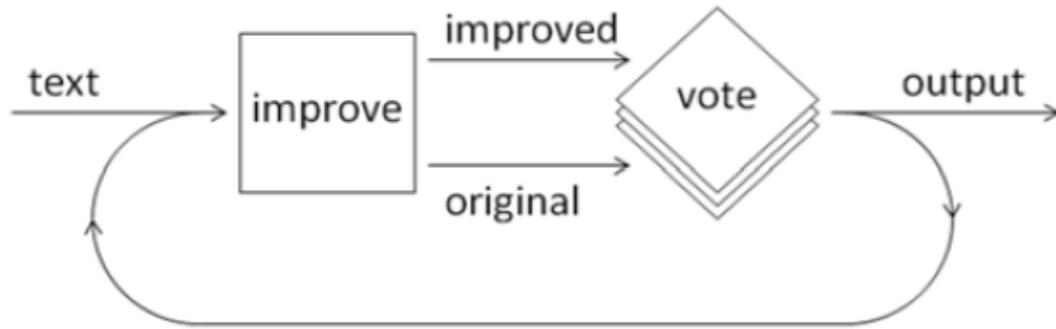
The task description is highlighted in yellow and reads: **Identify street:** You are given a set of panoramic pictures showing 360-degree field of view for some places. Your task is to identify which place is a **street**. Please indicate your choice by clicking the check symbol (or the image) to highlight the green check. If no place is a **street** in this set, check the checkbox at the bottom of the page to clearly indicate this.

The task includes a grid of panoramic images. Each image has a green checkmark to its left, indicating that the worker has selected it as the correct answer. The images show various outdoor scenes, including a grassy field, a paved area with a building, a dirt road, a large crowd of people, and a brick building.

Lifecycle of a HIT

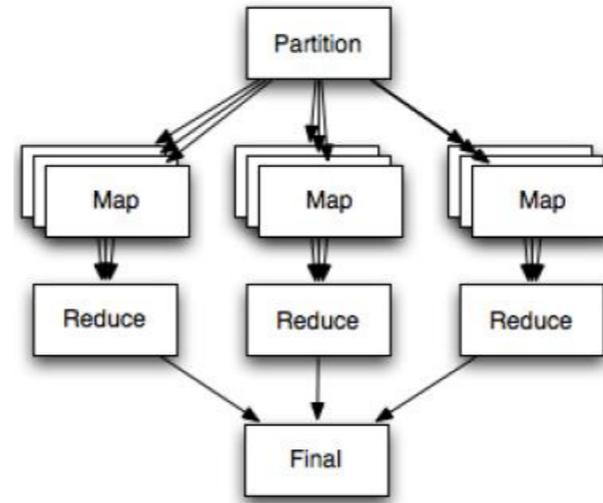
- Requester builds a HIT
 - Internal HITs are hosted by Amazon
 - External HITs are hosted by the requester
 - HITs can be tested on {requester, worker}sandbox.mturk.com
- Requester posts HIT on mturk.com
 - Can post as many HITs as account can cover
- Workers do HIT and submit work
- Requester approves/rejects work
 - Payment is rendered
 - Amazon charges requesters 10%
- HIT completes when it expires or all assignments are completed

To summarize: MTurk is great!

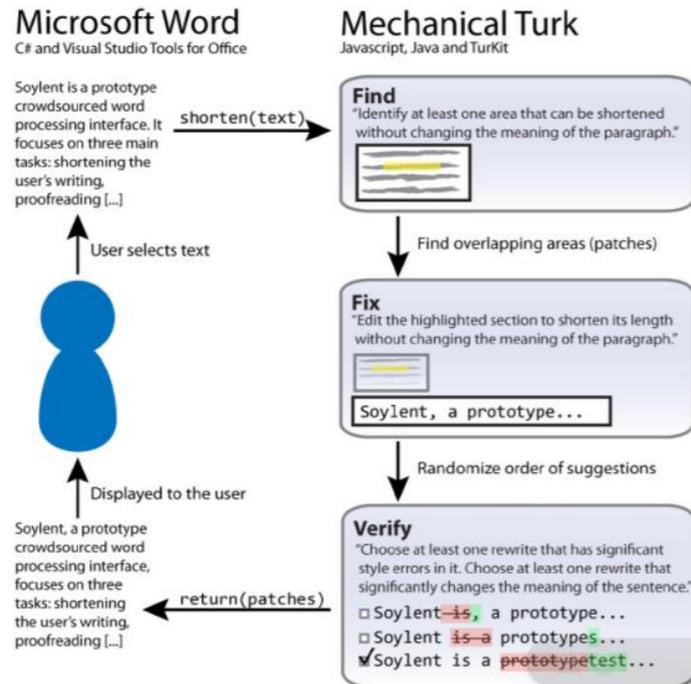


TurkKit

Little et. al (HCOMP 2009)

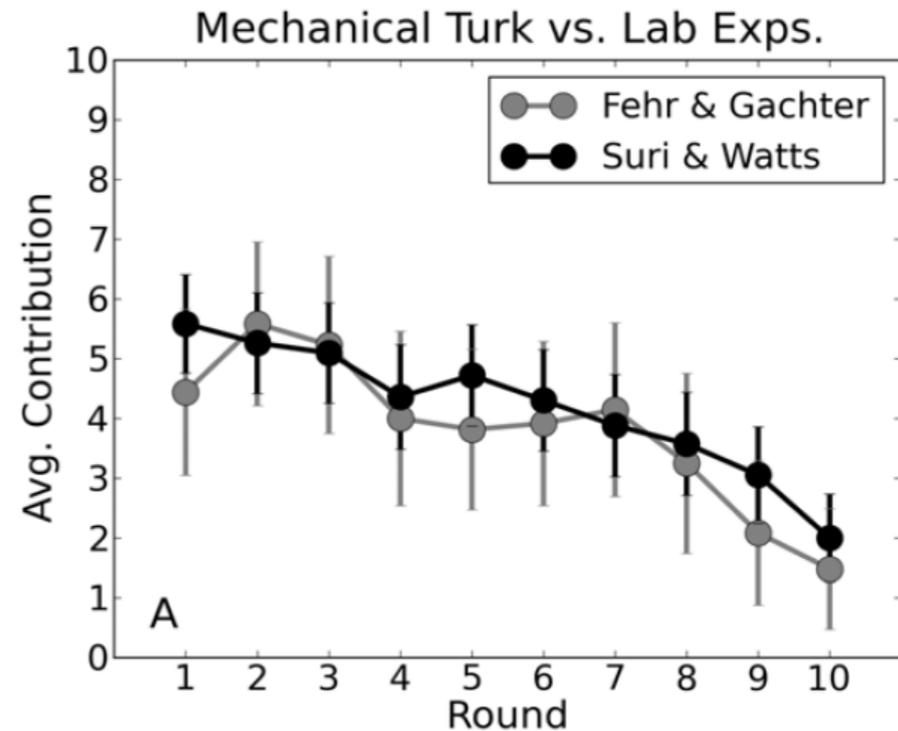


CrowdForge
Kittur et. al (UIST 2011)



Soylent

Bernstein et. al (UIST 2010)

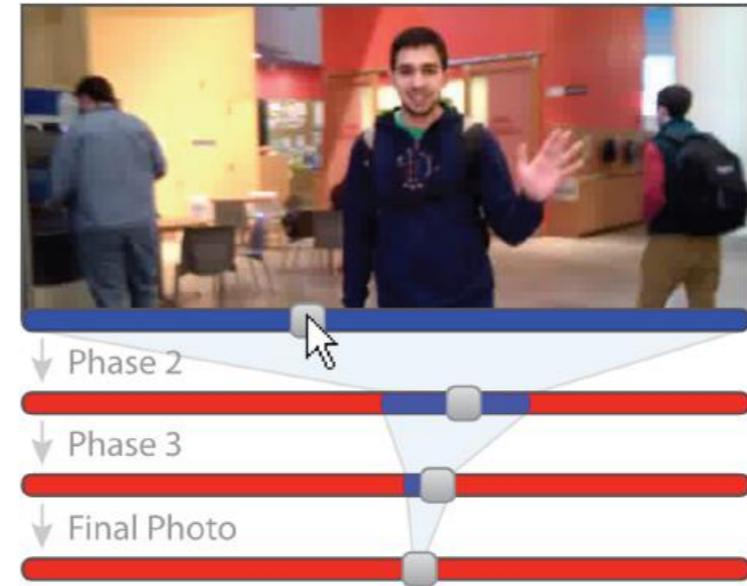


Public Goods Games
Suri & Watts (PloS ONE 6(3),
2011)

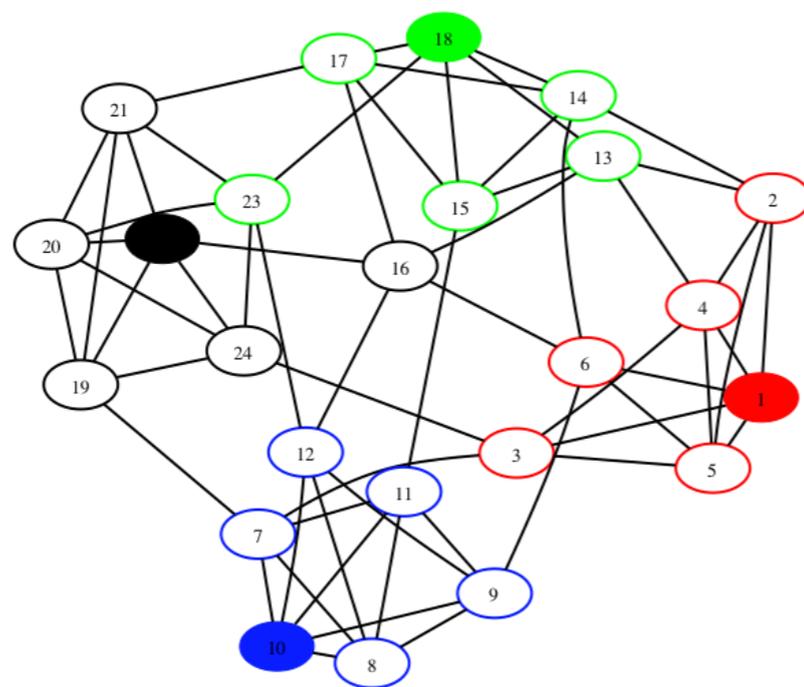
But what if you wanted to do these?



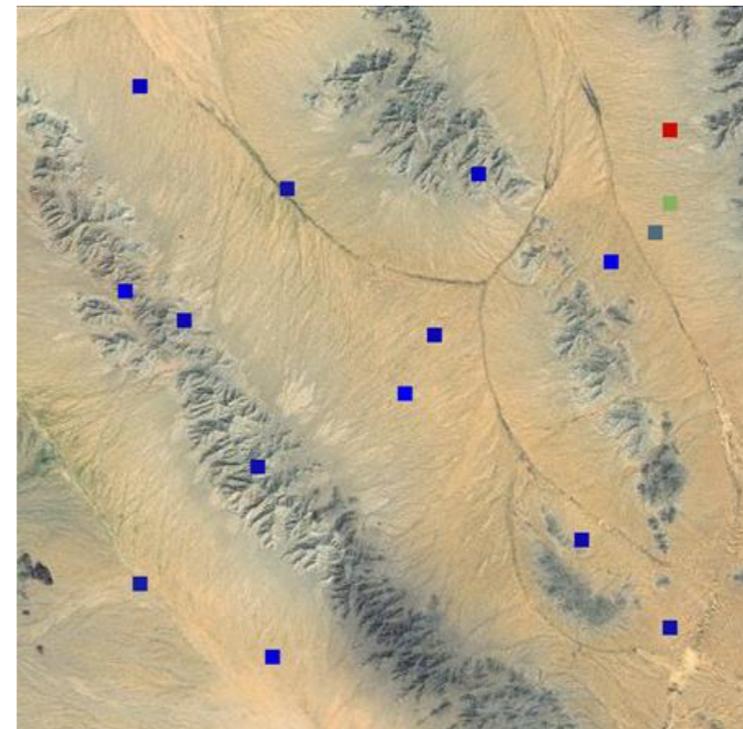
Zhang et. al (CHI 2012)



Bernstein et. al (UIST 2011)



Suri & Watts (PLoS ONE 2011)



Mason & Watts (PNAS 2011)

TurkServer – A framework for deploying MTurk experiments

- Deploying a real-time or synchronous experiment to MTurk or online has many moving parts:
 - Enforce participation limits and proper randomization
 - Tracking user actions and inattention
 - Monitoring progress and logging data
 - Building a complex software system
- TurkServer* handles many of these issues and is the result of several years of wrestling with online (especially **real-time and interactive**) experiments

* currently developing third version: <https://github.com/HarvardEconCS/turkserver-meteor>

TurkServer Features

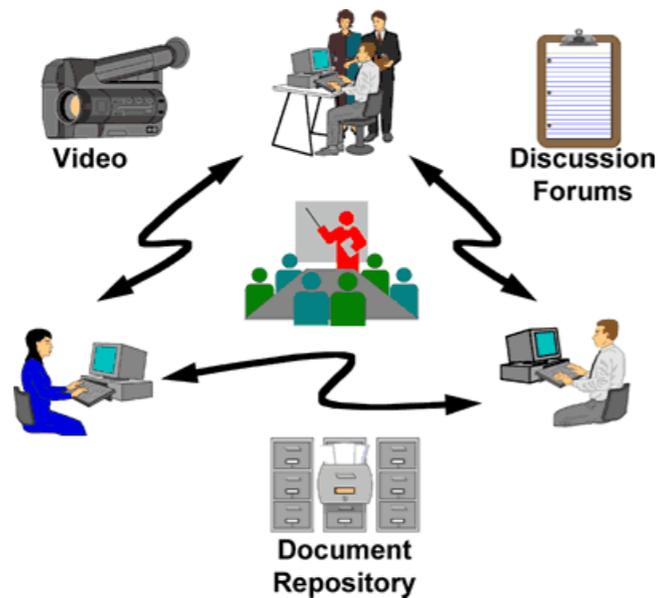
- **Real-time** server-client communication, allowing for continuous interaction between multiple workers
 - Leverages the revolutionary Meteor (<http://www.meteor.com/>) web framework
- Handles calls to the MTurk API, smart HIT management
- Worker tracking and enforcement of participation limits
- Integrated user instrumentation / data robustness tracking
- Deploy and monitor experiments via a web interface

Why Use TurkServer?

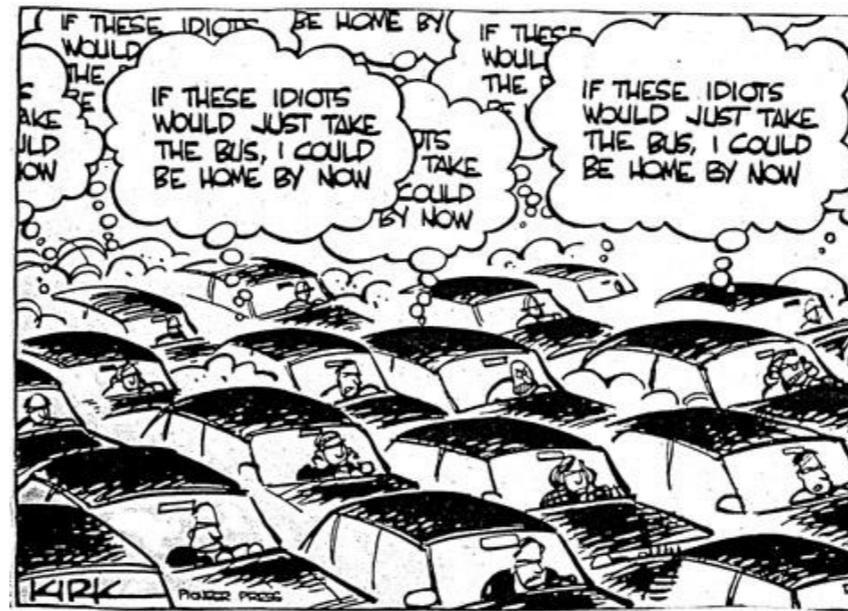
- Quickly prototype real-time or interactive experiments and deploy them to MTurk
 - Easy to test different treatments, conditions and settings
- High degree of automation and other conveniences
 - **No need to re-invent the wheel** for every online experiment
 - Current implementation is the product of many existing experiments

Skip the part where you shoot yourself in the foot and take advantage of the fact that we have done it already!

Human Computation Systems

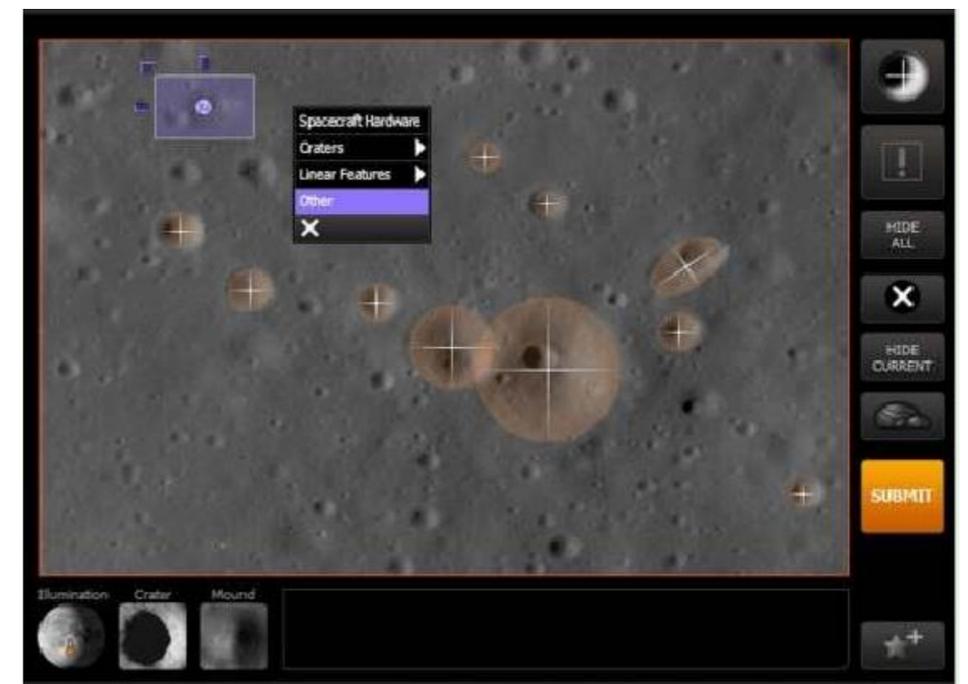
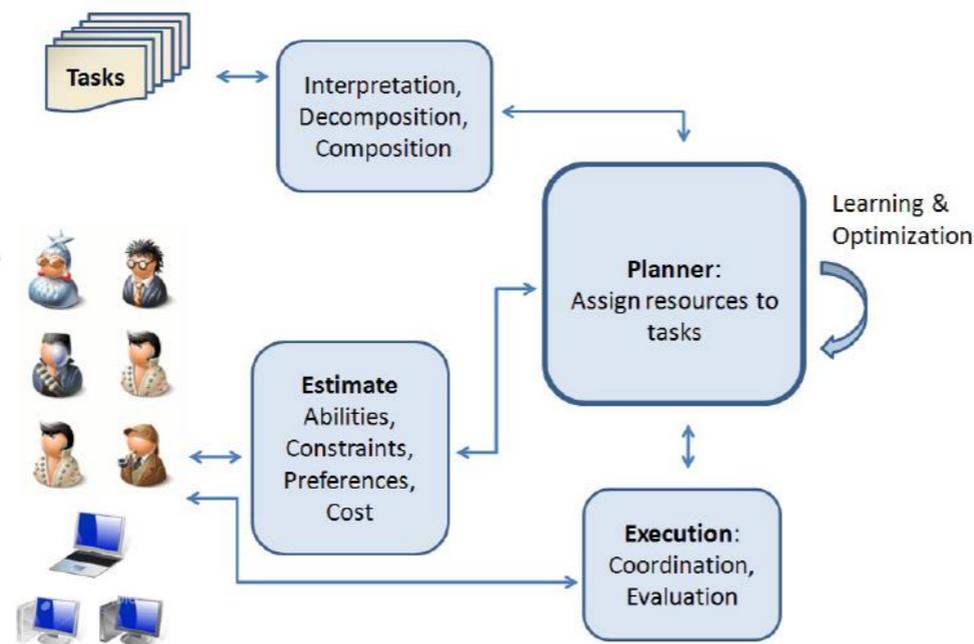


Design and testing of collaborative software



Coordination and interaction in human computation

Social incentives (“fun actor”) and adaptive systems for human computation



Synchronous mechanisms for crowdsourcing

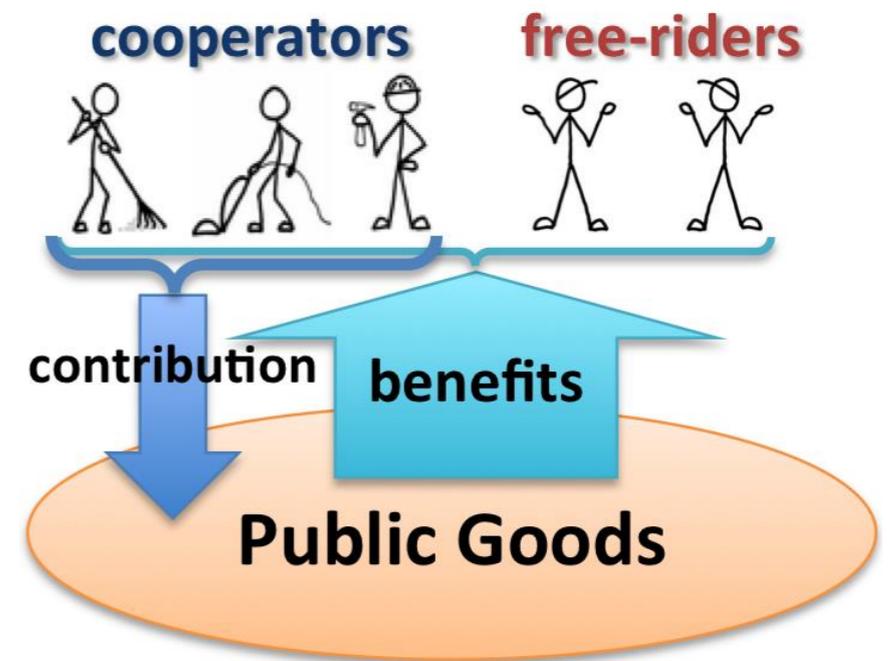
Online, Interactive Social Science

Intrade elections futures as of July 12, 2008 source: Intrade.com

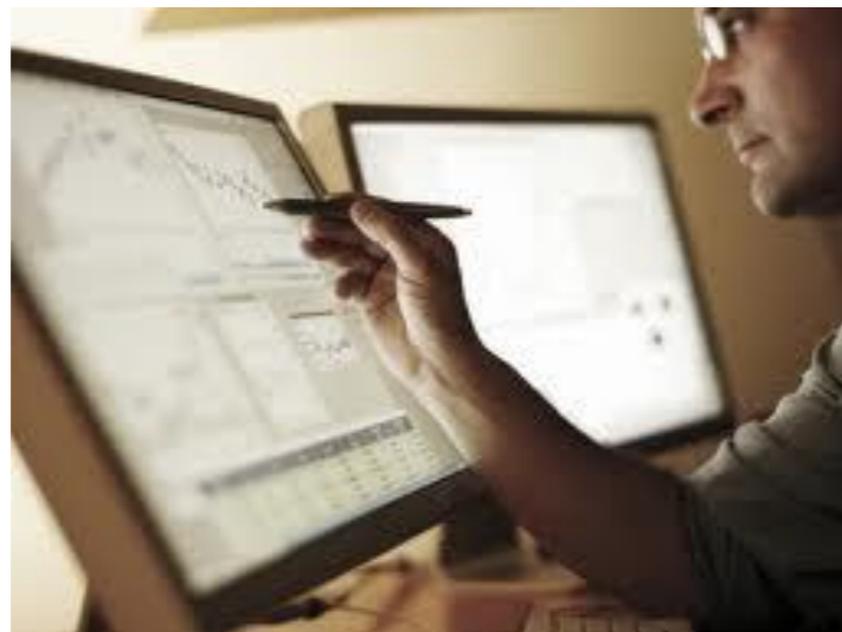


Empirical testing of mechanisms for financial and prediction markets

The Public Goods Game



Information aggregation; trading patterns and behavior in markets

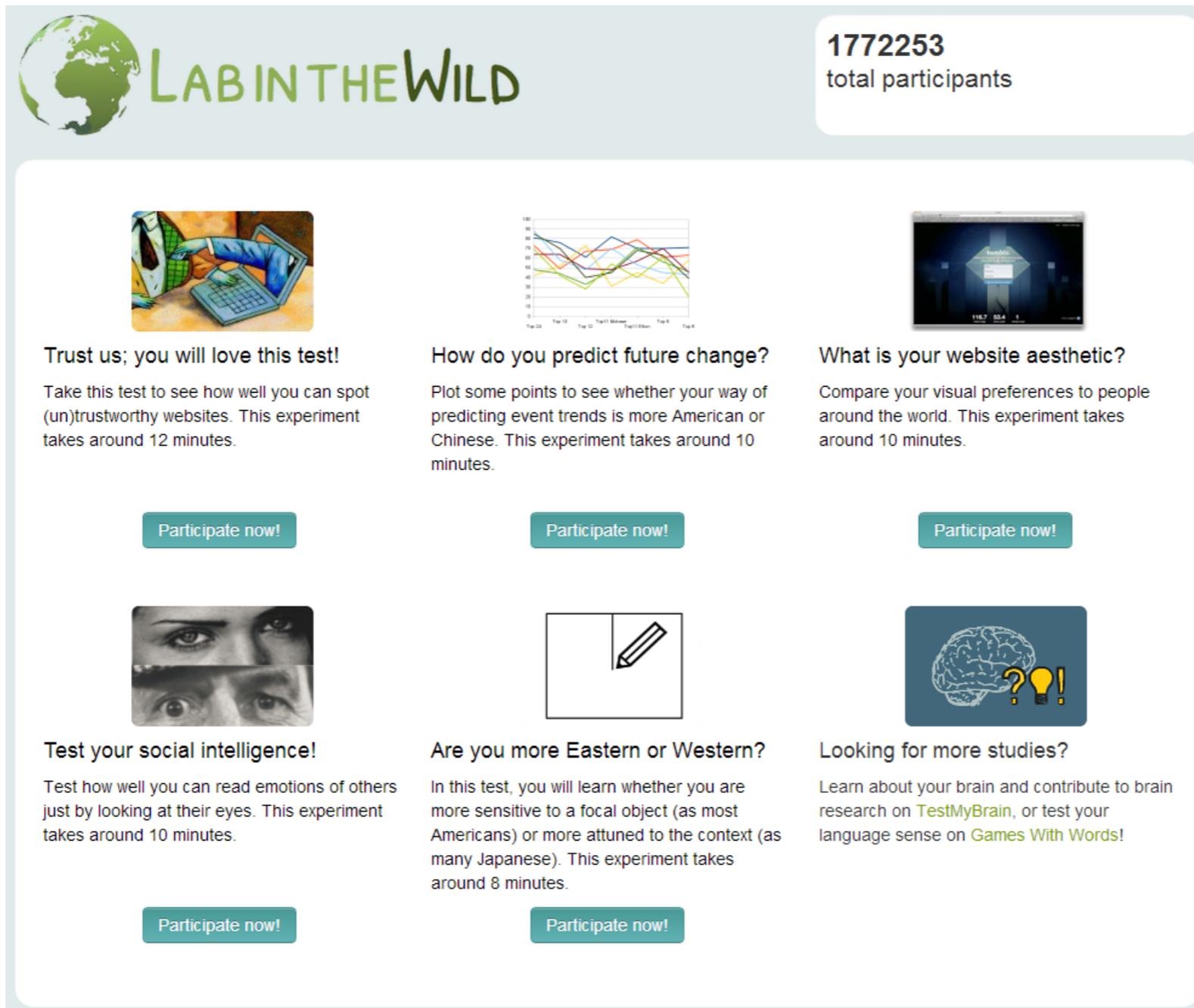


Large-scale behavioral experiments

Running Your Own Lab

- Anyone can use a website can start recruiting human subjects online for experiments
- Starting from scratch is a lot of work
 - Must actually get people to come to your website
 - Building hardware and software infrastructure
 - Keeping track of your subject pool
- Potential for future “unified online lab” that centralizes subject recruitment and tracks existing studies to reduce conflicts

Large-Scale Online Labs: LabInTheWild



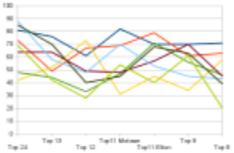
 **LABINTHEWILD**

1772253
total participants



Trust us; you will love this test!
Take this test to see how well you can spot (un)trustworthy websites. This experiment takes around 12 minutes.

Participate now!



How do you predict future change?
Plot some points to see whether your way of predicting event trends is more American or Chinese. This experiment takes around 10 minutes.

Participate now!



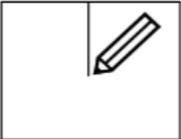
What is your website aesthetic?
Compare your visual preferences to people around the world. This experiment takes around 10 minutes.

Participate now!



Test your social intelligence!
Test how well you can read emotions of others just by looking at their eyes. This experiment takes around 10 minutes.

Participate now!



Are you more Eastern or Western?
In this test, you will learn whether you are more sensitive to a focal object (as most Americans) or more attuned to the context (as many Japanese). This experiment takes around 8 minutes.

Participate now!

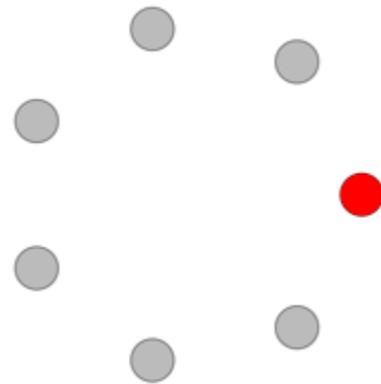


Looking for more studies?
Learn about your brain and contribute to brain research on [TestMyBrain](#), or test your language sense on [Games With Words!](#)

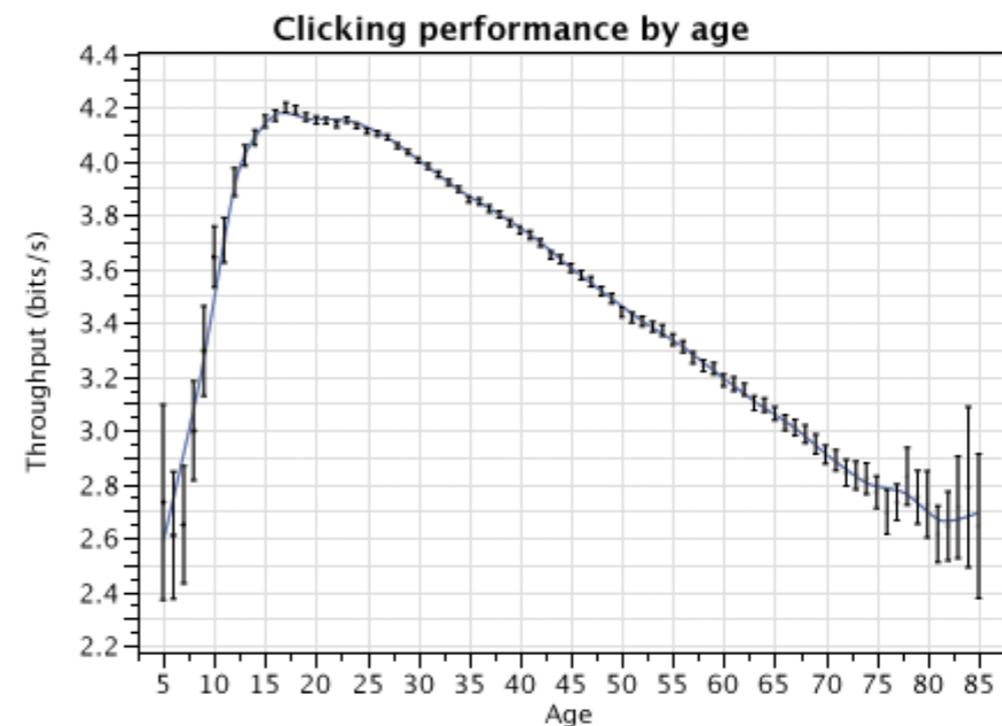
LabInTheWild: Age Guessing Game

Click on a few dots and our program will guess your age!

Human motor system changes as we age. Our program will guess your age by analyzing how you click.



- > 1,000,000 visitors
- > 500,000 participants from 218 countries
- ~ 350,000 participants who provided full demographic info



Large-Scale Online Labs: TestMyBrain

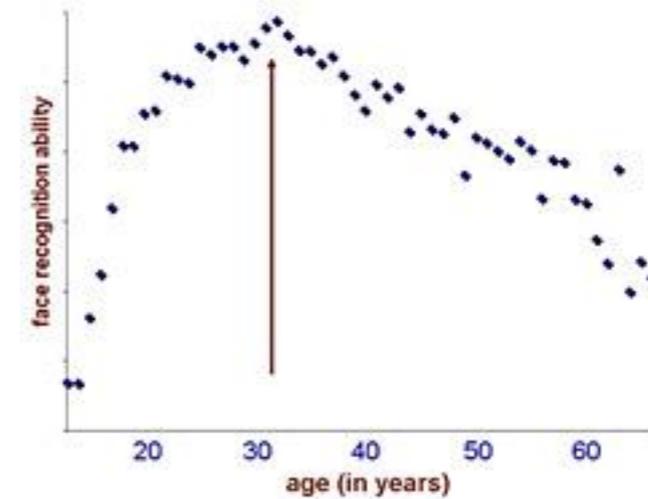
TESTMYBRAIN

Why test? TestMyBrain aims to engage and collaborate with citizen scientists like you, by providing tools to help you learn about yourself. When you test yourself and build your brain profile, you contribute to brain research.

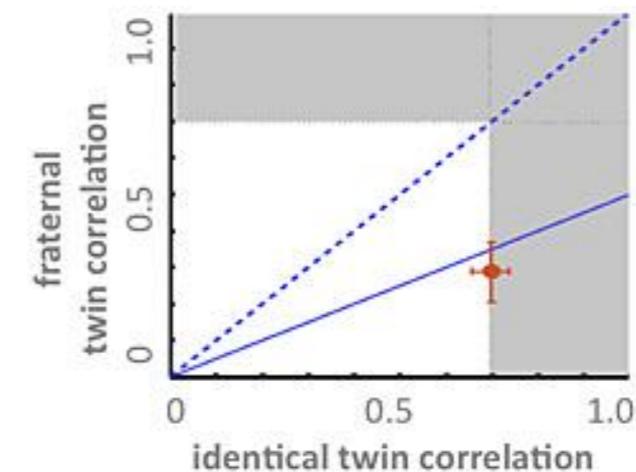
Brain tests

Cognitive Style
 In this test, we look at two measures of your cognitive style and personality.
 Estimated time to complete: 15 minutes
 3577 brains

Matching Faces In Photographs
 These tests look at how good you are at matching two photographs of the same person.
 Estimated time to complete: 20 minutes
 3904 brains



Face recognition peaks at age 30



Face recognition ability is genetic

Comparison of web-based and lab-based perception studies:

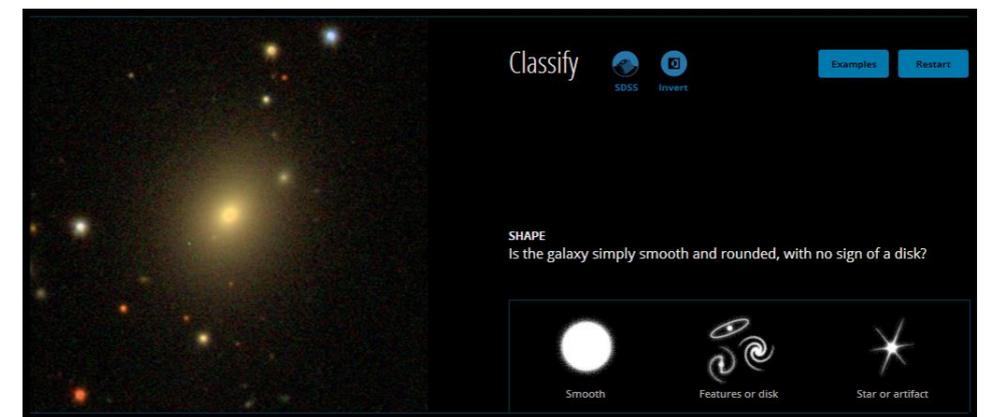
L. Germine et al. *Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments.* **Psychonomic Bulletin & Review** 19.5 (2012): 847-857

Citizen Science: the Zooniverse

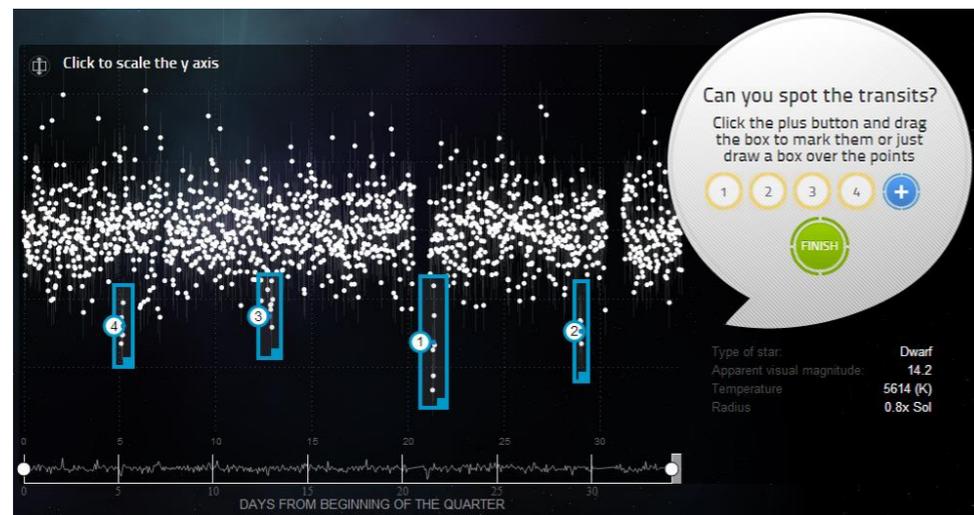
LOG of the UNITED STATES *Steamer Bear*
Left Cape York & Making Passage to Cape Sabine.

Hour.	Knots.	Fathoms.	Course steered.	WINDS.			BAROMETER.			TEMPERATURE.			State of the Weather, by symbols.	Form of Clouds, by symbols.
				Direction.	Force.	Velocity.	Height in inches.	Ther. at't'd.	Air Dry Bulb.	Air Wet Bulb.	Water at surface.			

Old Weather



Galaxy Zoo



Planet Hunters



Snapshot Serengeti

Exciting news: field experiments coming to the Zooniverse over next year

Examples and Discussion

Let's Design an Experiment...

Think of a project of interest to you, with...

- ...theory that makes a prediction about human behavior?
- ...data mining results showing some correlation?
- ...agent based modeling or simulation predictions?

Generating Content on StackOverflow

The screenshot shows a Stack Overflow page for the question "How to start learning Ajax?". The question is at the top, followed by six answers. The top answer is the most voted, with 12 votes. The answers are sorted by the number of votes. The page includes navigation links for Questions, Tags, Users, Badges, and Unanswered. The question text asks for the best way to start learning Ajax, and the answers provide various resources and explanations.

Question

Answers, sorted by # votes

- Someone asks a question
- Anyone can post an answer
- Answers can be voted up, which generates +5 “reputation” (points) for the answerer
- Answers can be voted down, which costs -2 rep to the answerer and -1 to the downvoter
- How did we arrive at this system?

A Concise History of StackOverflow Design

- FGITW (Fastest Gun in the West)
 - First design: answers ordered by upvotes, then by oldest first
 - Result: Users post answers **as fast as possible** to get the oldest spot, then leave as-is or edit for more detail
 - Fix: don't order answers by age
- SCITE (Slowest Cheater in the East)
 - Second design: answers ordered by upvotes, ties randomly broken
 - Result: Users **downvote** others' answers so they appear first, garnering more upvotes, then remove downvotes later
 - Fix: lock in downvotes after 5 minutes
- Should we design a system in this ad hoc fashion? How would you design a Q&A system? Experimental ideas?

Collecting Ratings for User Generated Content

Suppose we are designing a system for gathering ratings for artifacts to provide recommendations and feedback, a la



NETFLIX



- How should we collect the ratings?
- What induces people to provide the most information?
- What question should we use for elicitation?
- How can we find out what users enjoy using the most?
- What is the best way to provide feedback to users?

Thanks!

Questions?

- Feedback, comments, complaints? Send any quick thoughts to
 - mao@seas.harvard.edu
 - suri@microsoft.com
- Would you be interested in a one week course?