Community-Led Experiments (SOC 412)

Week 3 Lecture 1

Sherrerd Hall 306

J. Nathan Matias

@natematias

civilservant.io

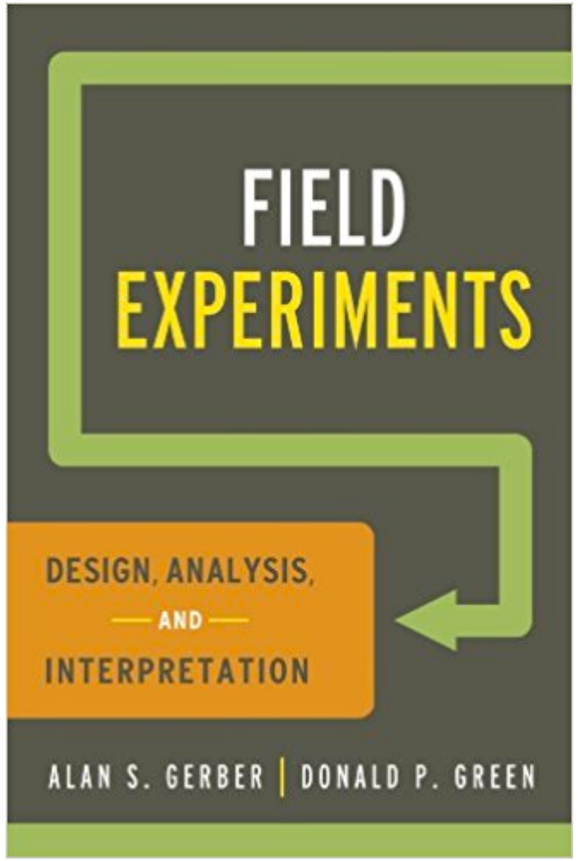
jmatias@princeton.edu

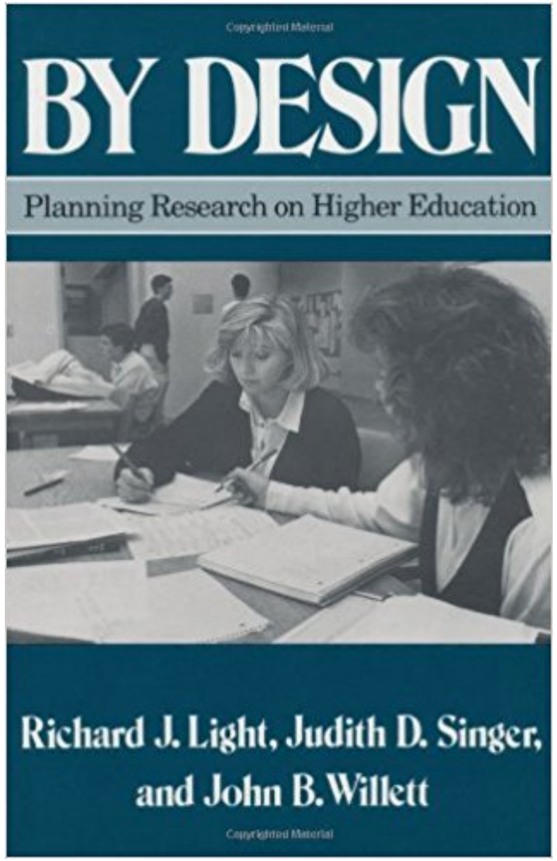
Department of Psychology
PRINCETON UNIVERSITY

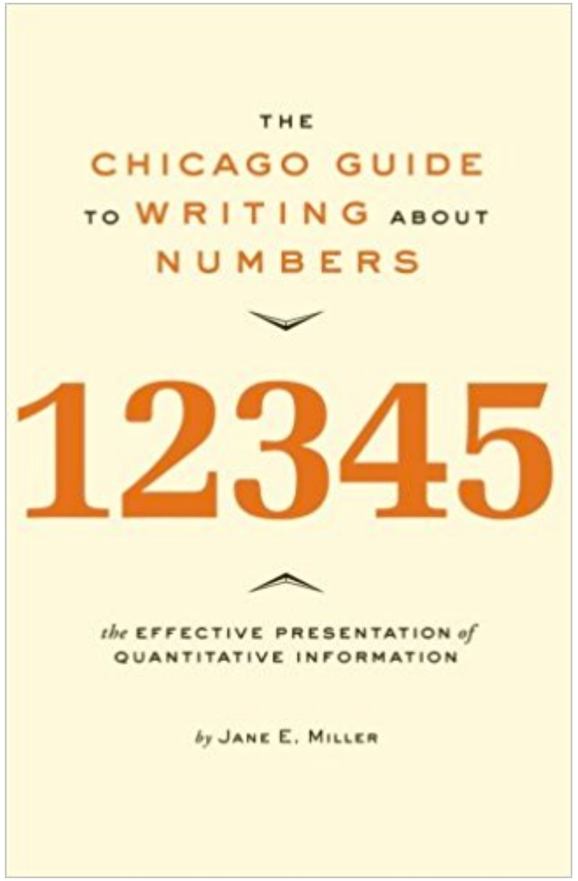
CITP

mit media lab

References to Know







Week 4: Community-Led Experiments (SOC412)

CivilServant

Citizen Behavioral Science for a fairer, safer, more understanding internet

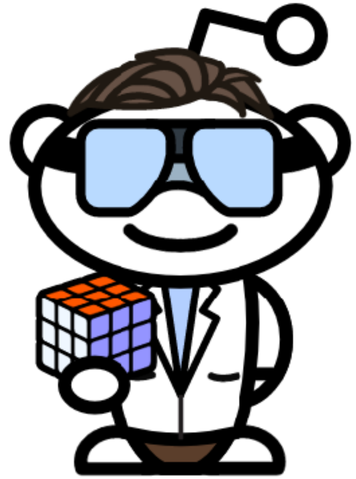
civilservant.io

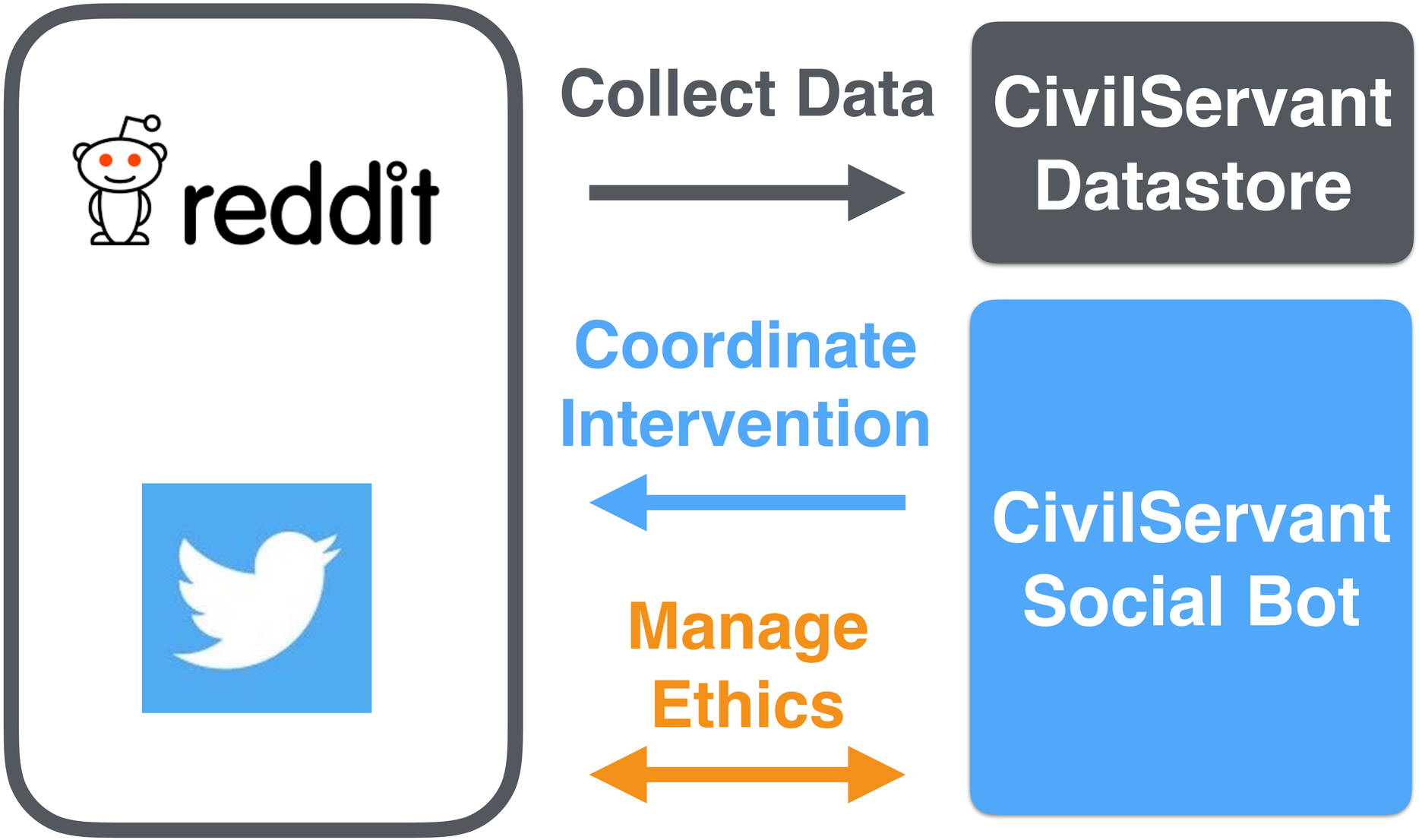
GlobalVoices

Civil Servant

Community-Led Field Experiments in
Governing Human & Machine Behavior

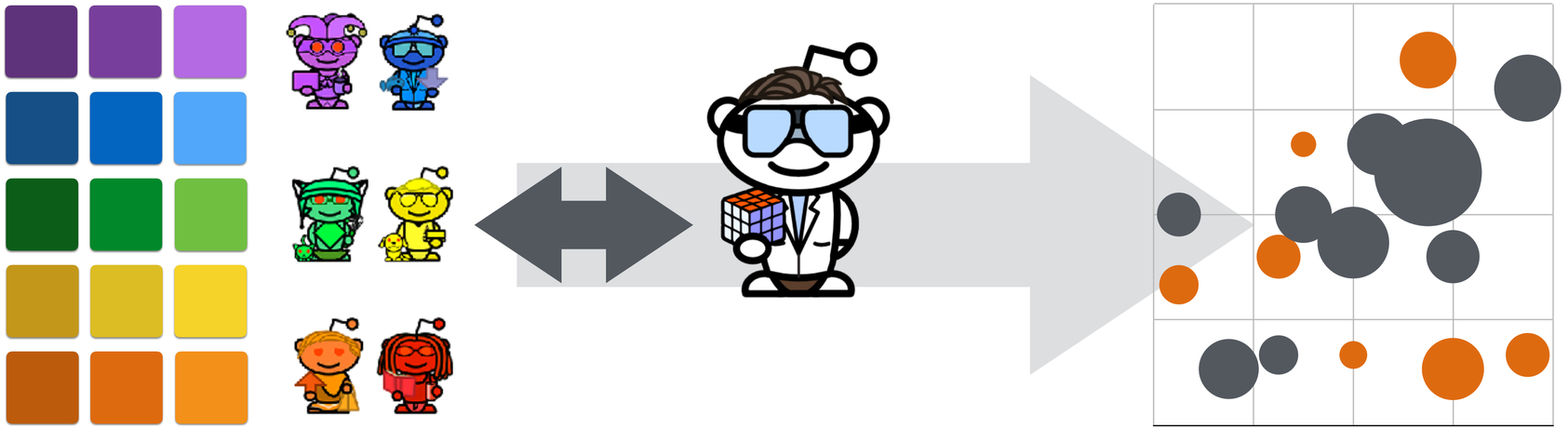
- **Co-Design Experiments**
- **Coordinate Policy Interventions**
- **Monitor Outcomes**
- **Estimate Outcomes For Discussion**
- **Support Remixes & Replications**




reddit
Collect Data
CivilServant Datastore
Coordinate Intervention
CivilServant Social Bot
Manage Ethics

Civil Servant

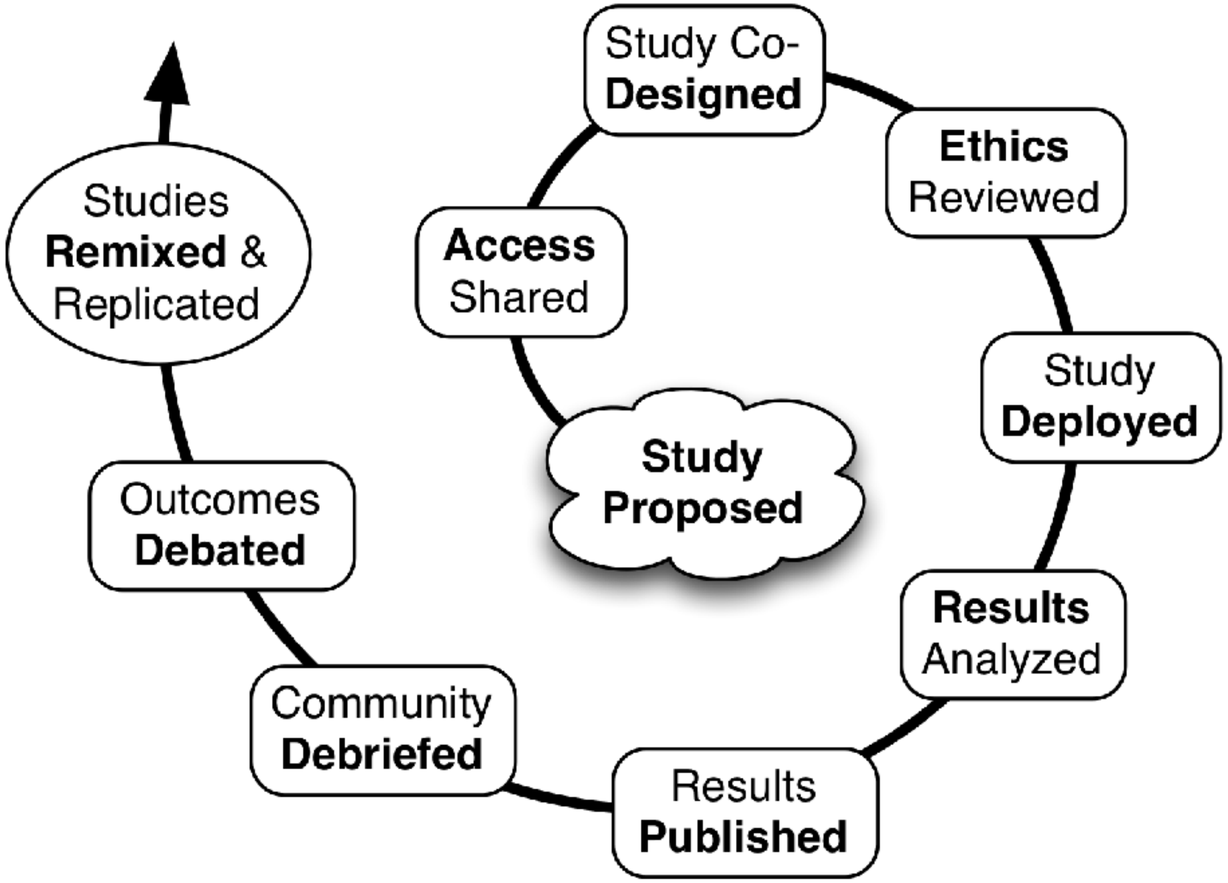
Community-Led Field Experiments in Governing Human & Machine Behavior



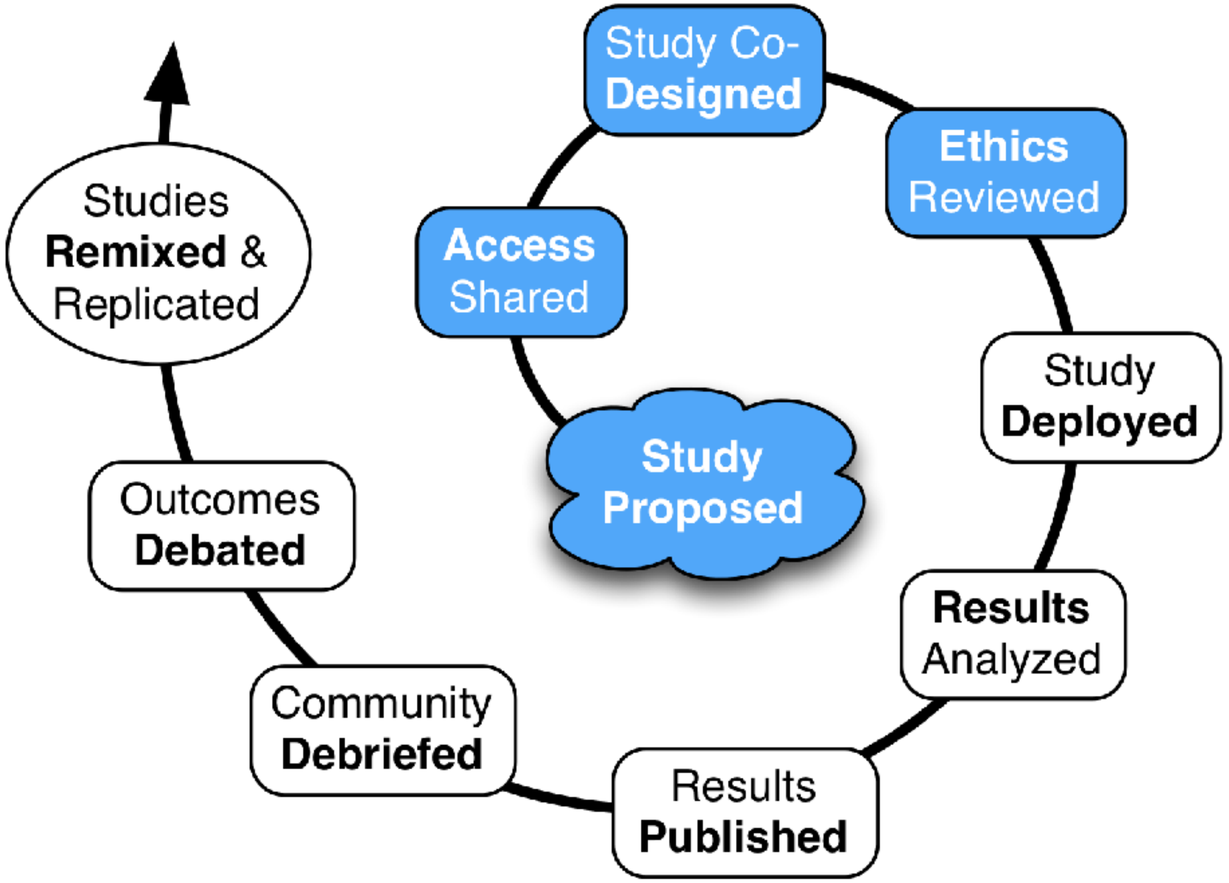
Community Experiments

Open Archive of Experiment Results

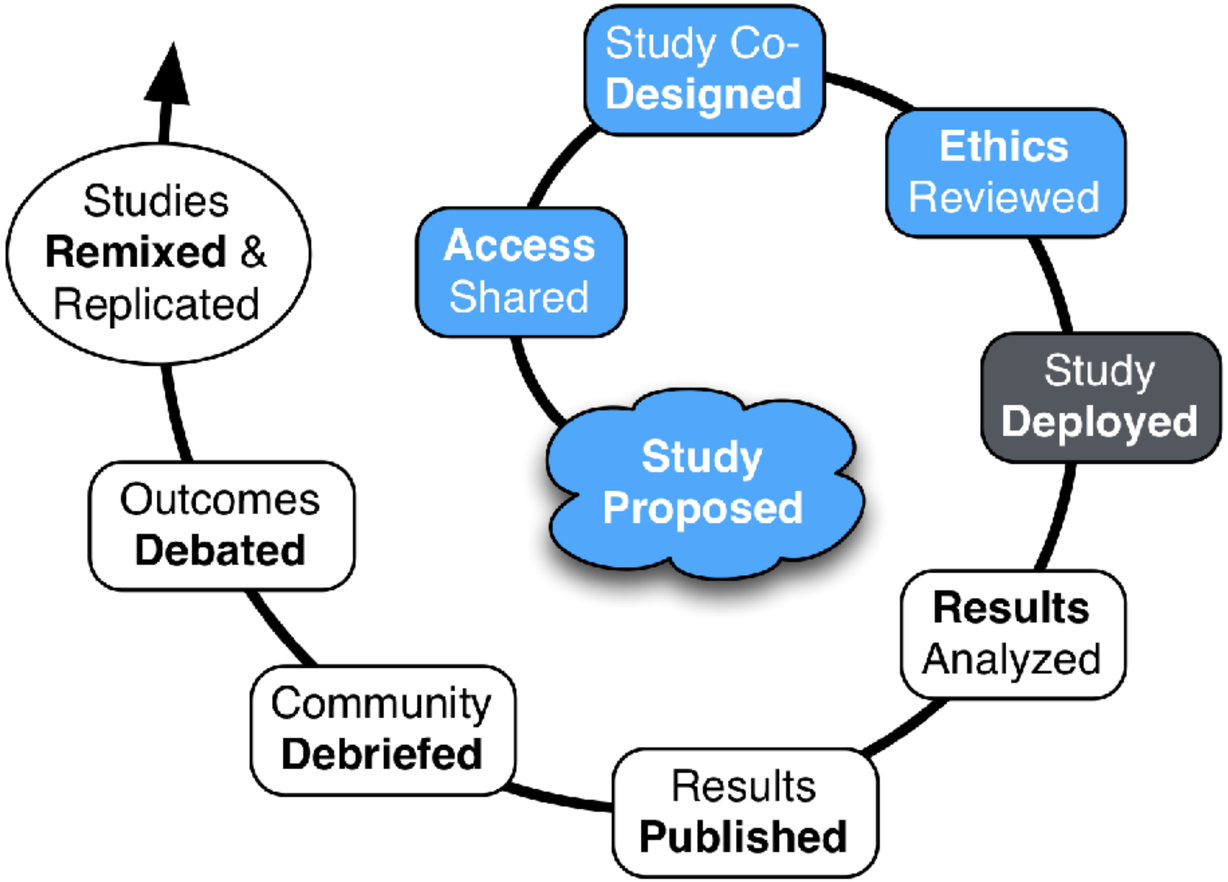
Research Process



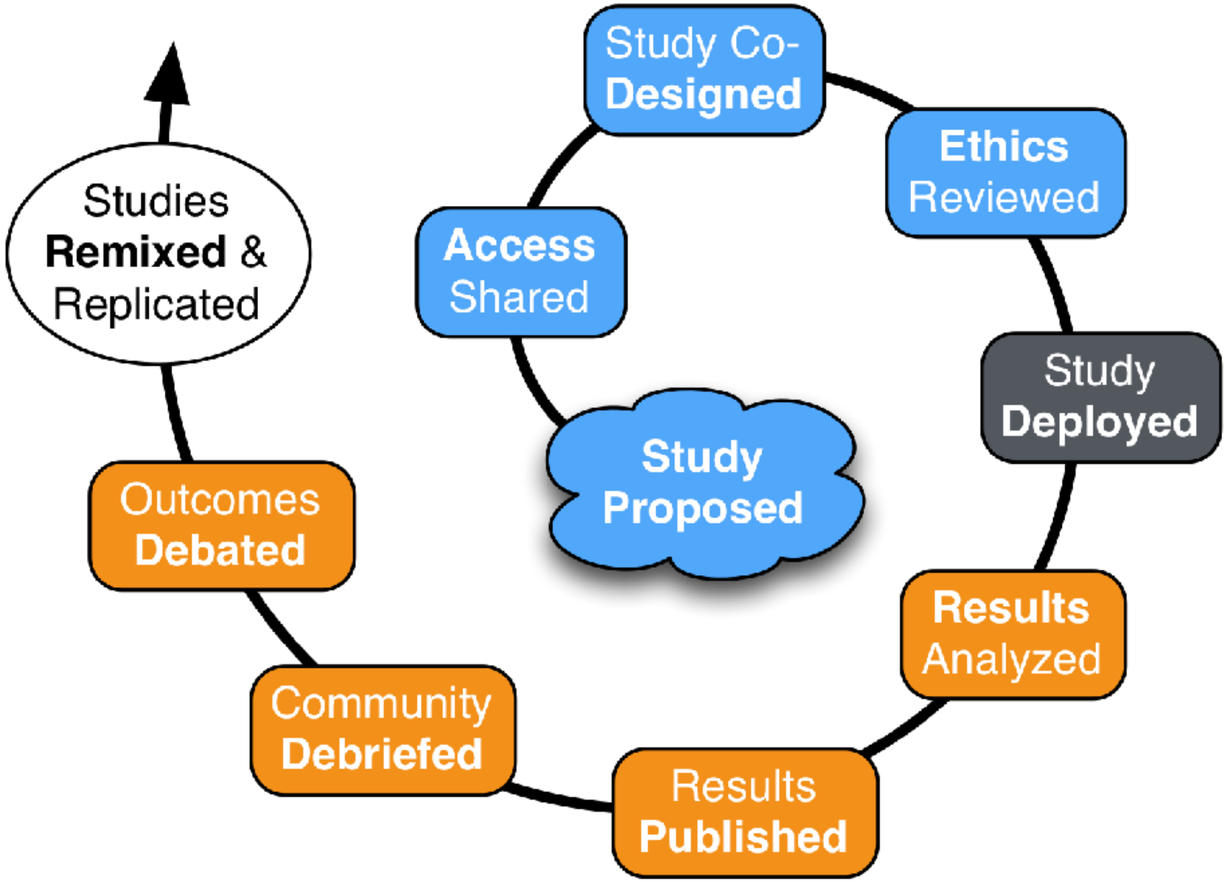
Research Process



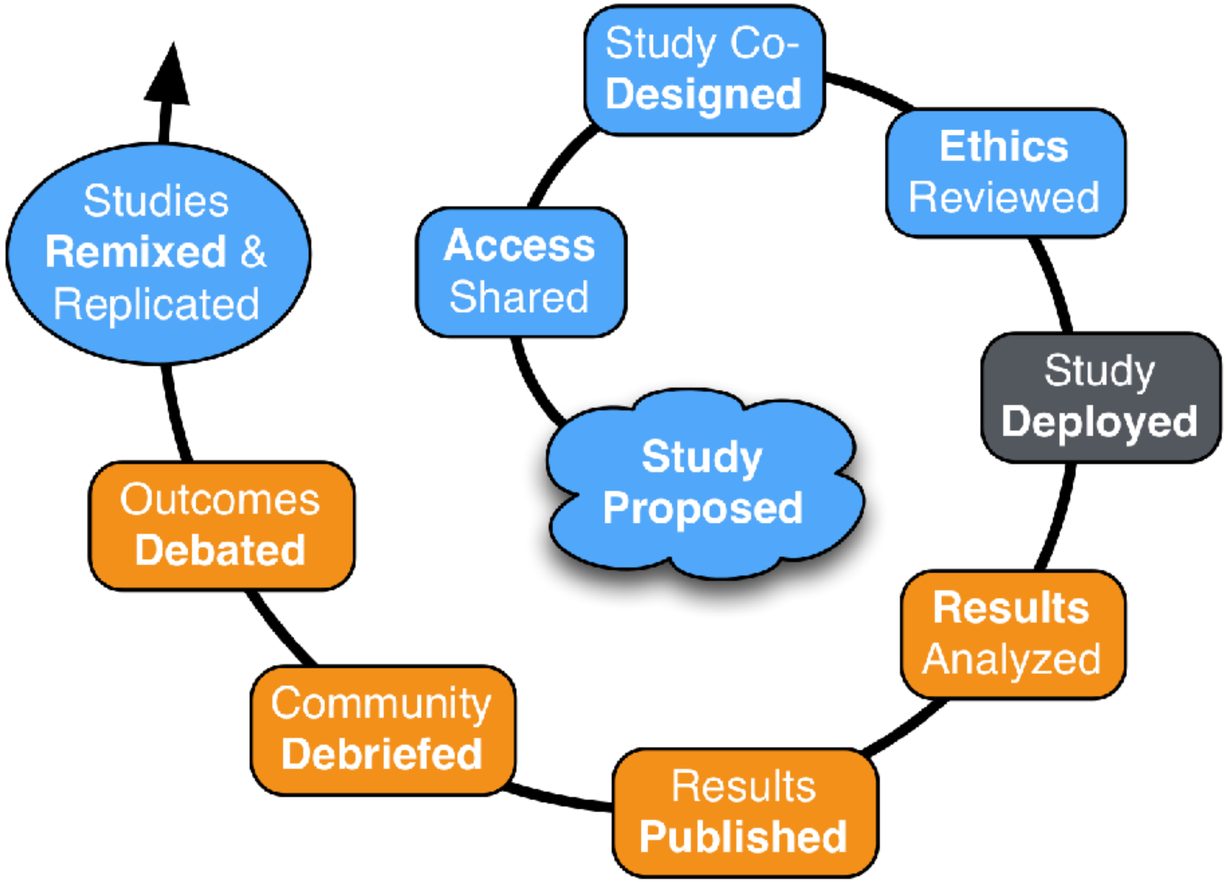
Research Process



Research Process



Research Process



mit media lab

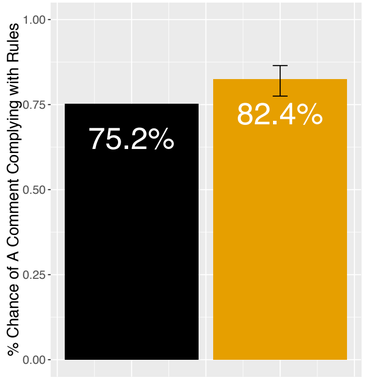
BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Ethics and Governance of Artificial Intelligence Fund

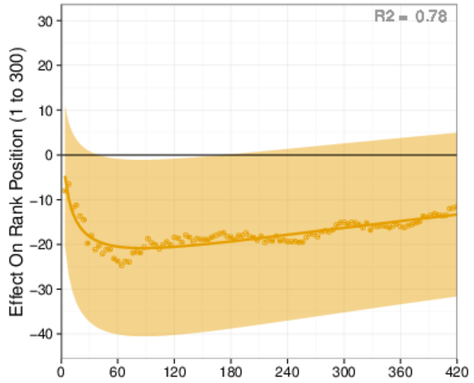
KAHNEMAN-TREISMAN
CENTER FOR BEHAVIORAL
SCIENCE & PUBLIC POLICY
PRINCETON UNIVERSITY

moz://a KF KNIGHT FOUNDATION

MacArthur Foundation



Preventing Harassment



Limiting Misinformation

Managing Conflict

Banning Accounts

Countering Sexist Attacks

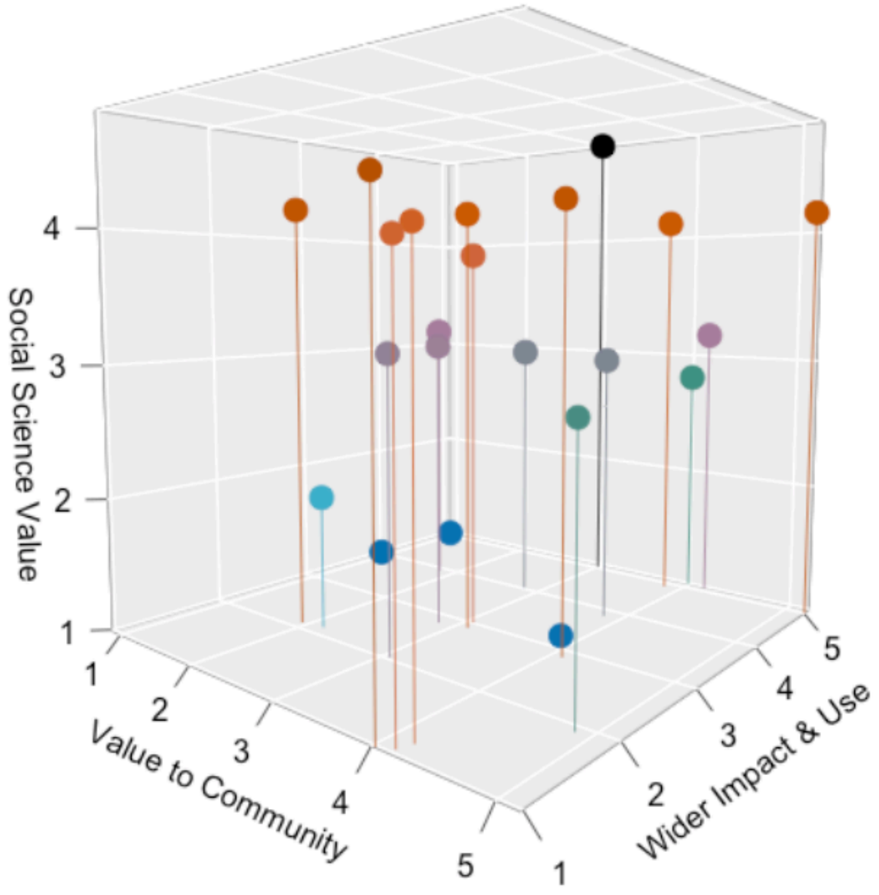
Auditing Algorithms

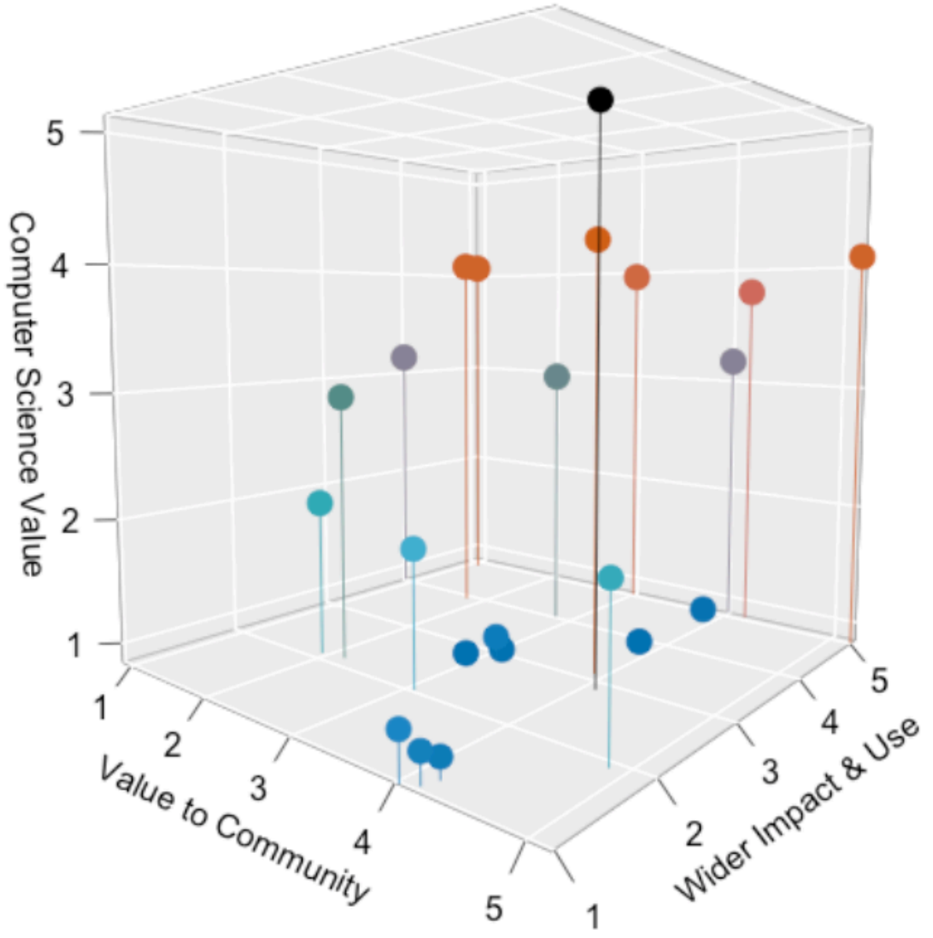
Auditing AI Law Enforcement

Accountability & Ethics

Week 4: Community-Led Experiments (SOC412)

Research proposed at the CivilServant Research Summit could **Transform Communities**, **Benefit Society**, & **Grow Knowledge** in the **Social Sciences & Computing**





Eric Pennington

Research Manager

Class Projects

- Discussing **sexual assault online**
- **Preventing harassment** in online communities
- **Improving newcomer retention** with a guestbook
- **Reduce the influence of harmful norms** on people's behavior
- **Improving moderator accuracy** with training
- The **Gray Phone Challenge**

Posting about Sexual Assault Online

"Posts about assault and rape are the ones most often downvoted, received with hostility, cross-linked, and brigaded [suppressed by opposing groups].”

Can supportive & warning messages to people who have written about rape and sexual assault improve their experience?

(community size: ~90,000 subscribers)

Preventing Unruly Behavior (replications)

What effect does posting community rules have on people's unruliness and participation?

Also, can we achieve similar effects by auto-detecting large influxes of attention and responding with more information about the rules?

(community sizes: 6M, 12M, and many more)

Guestbook & Newcomer Retention

Can a "guestbook" invitation reduce moderation actions and increase engagement among first-time participants with celebrity guests?

(community size: 18M)

Reduce Role of Norms on Behavior

When people post deeply personal stories, they are sometimes attacked by people outside of the community. Can we reduce the effect of this peer feedback on people's experience and their chance to post in the future?

(community size: 90k subscribers)

Improving Moderator Accuracy

Could basic forms of training improve the accuracy, reliability, and longevity of a moderator's content removals over time?

(community size: 18M subscribers)

The Gray Phone Challenge