The Alignment Problem: The Case of Field Experiments in the U.S. Safety Nets

Jae Yeon Kim

Research @ Johns Hopkins, Harvard
Ex-Data Science @ Code for America

June 15, 2024

Plan

Why this talk?

The scientists then in the [CMU] Engineering College neither understood engineering nor believed it could be taught. They educated engineers by giving them a lot of physics and math, hoping that their students would later be able to design safe bridges or airplanes.
— "Models of My Life," Simon (1996, 258)[1]

[1] Herbert A. Simon was an American political scientist who received the Nobel Memorial Prize in Economic Sciences in 1978 and the Turing Award in computer science in 1975.

Motivation

- In this talk, I will explain what I wish I had known about field experiments in policy contexts while I was a graduate student in political science at UC Berkeley: **balancing the rigor of research and the cost of implementation (what I called "the alignment problem").**

Motivation

- In this talk, I will explain what I wish I had known about field experiments in policy contexts while I was a graduate student in political science at UC Berkeley: **balancing the rigor of research and the cost of implementation (what I called "the alignment problem").**
- The talk is roughly based on my professional experience as a data scientist at Code for America. There, I helped design and implement field experiments with the U.S. federal, state, and local governments, especially in safety net contexts (SNAP, WIC, Medicaid, etc.).

- Why experiments? We design and run an experiment because we care about causal questions when introducing changes in practice (e.g., product release, design update, policy change).

- Why experiments? We design and run an experiment because we care about causal questions when introducing changes in practice (e.g., product release, design update, policy change).
- Experiments help answer causal questions accurately via the power of randomization (Fisher, 1935).

- Why experiments? We design and run an experiment because we care about causal questions when introducing changes in practice (e.g., product release, design update, policy change).
- Experiments help answer causal questions accurately via the power of randomization (Fisher, 1935).
  - The hidden engine of incremental innovations (i.e., product and service optimization) in Silicon Valley

"I think when people think about Silicon Valley they imagine Steve Jobs in a garage, or the invention of the iPhone or the iPad, and they think that's what Silicon Valley is. ... Most of the innovation in big tech companies is incremental. The A/B test is probably the most impactful business process innovation in a very long time."

- Susan Athey (Stanford Business School, previously Chief Economist at Microsoft)[2]

[2]Based on her 2016 interview with Russ Roberts: `https://www.econtalk.org/susan-athey-on-machine-learning-big-data-and-causation/`

- To experiment:

[3] It stands for stable unit treatment value assumption. No interference between units and no hidden variations in treatments.

- To experiment:
- You need to know the science of a randomized controlled experiment (RCT):

[3] It stands for stable unit treatment value assumption. No interference between units and no hidden variations in treatments.

- To experiment:
- You need to know the science of a randomized controlled experiment (RCT):
  - You need to know the causal inference framework and its underlying assumptions (e.g., Neyman-Rubin potential outcome model, SUTVA)[3] (Imbens and Rubin, 2015)

[3] It stands for stable unit treatment value assumption. No interference between units and no hidden variations in treatments.

- To experiment:
- You need to know the science of a randomized controlled experiment (RCT):
  - You need to know the causal inference framework and its underlying assumptions (e.g., Neyman-Rubin potential outcome model, SUTVA)[3] (Imbens and Rubin, 2015)
  - You need to know the statistical language and terms (e.g., estimands, estimators, etc.) (Lundberg, Johnson and Stewart, 2021)

[3] It stands for stable unit treatment value assumption. No interference between units and no hidden variations in treatments.

- To experiment:
- You need to know the science of a randomized controlled experiment (RCT):
  - You need to know the causal inference framework and its underlying assumptions (e.g., Neyman-Rubin potential outcome model, SUTVA)[3] (Imbens and Rubin, 2015)
  - You need to know the statistical language and terms (e.g., estimands, estimators, etc.) (Lundberg, Johnson and Stewart, 2021)
- Once implemented well, the analysis of an RCT is straightforward: e.g., creating two bar charts (treatment and control, aka "A/B" tests) and interpreting and communicating them.

[3] It stands for stable unit treatment value assumption. No interference between units and no hidden variations in treatments.

"My favorite kind of graphs are what I call big bar-little bar graphs. They're graphs that have one really little bar and one really big bar. And those are the kind of graphs that I show to C.E.O.s if I'm trying to convince them of something."

- Steve Levitt (Chicago Economics, the co-author of Freaknomics)[4]

[4] Based on his 2020 interview with Stephen Dubner:
`https://freakonomics.com/podcast/steve-levitt-im-not-as-childlike-as-id-like-to-be-bonus/`

- Often, we want to run an experiment in a field because we want both internal (causation) and external validity (real world) (Gneezy and List, 2013).

- Often, we want to run an experiment in a field because we want both internal (causation) and external validity (real world) (Gneezy and List, 2013).
- Running an experiment in a field is hard, not in the way theoretical physics is hard, but marriage or parenting is hard.

- In real estate, the key is locations, locations, and locations.

[5] The concept's name is inspired by Christian (2021), a book on aligning machine learning technologies and human values.

- In real estate, the key is locations, locations, and locations.
- In field experiments, the key is alignments, alignments, and alignments (both internal and external).

[5] The concept's name is inspired by Christian (2021), a book on aligning machine learning technologies and human values.

- In real estate, the key is locations, locations, and locations.
- In field experiments, the key is alignments, alignments, and alignments (both internal and external).
- **The alignment problem**[5]: balancing the rigor of research and the cost of implementation

[5] The concept's name is inspired by Christian (2021), a book on aligning machine learning technologies and human values.

- Data scientists (researchers) need support from other teams and individuals to conduct field experiments. The goal is to maximize rigor within these constraints.
- In the following slides, I will explain how to achieve this objective in four steps (scoping, implementation, measurement, and communication). Note that implementation starts with scoping and ends with communication.

	High rigor	Low rigor
High cost	Researcher = Happy, Implementer = Hesitant	Researcher = Frustrated, Implementer = Hesitant
Low cost	Researcher = Happy, Implementer = Happy	Researcher = Frustrated, Implementer = Happy

Table 1: Types of researcher and implementer relationships

Four Alignments: SIMC

1. **S**coping

Four Alignments: SIMC

1. **S**coping
2. **I**mplementation

Four Alignments: SIMC

1. **S**coping
2. **I**mplementation
3. **M**easurement

Four Alignments: SIMC

1. **S**coping
2. **I**mplementation
3. **M**easurement
4. **C**ommunication

Plan

Scoping

- Scoping: turning applied problems into well-defined research questions (=hypotheses)

[6] Supplemental Nutrition Assistance Program, aka "food stamp"

[7] Women, Infant, Children

[8] My wife and I received support from WIC and Medicaid (Medical) when we had our daughter as both graduate students. So, I had personal experience with these programs.

Scoping

- Scoping: turning applied problems into well-defined research questions (=hypotheses)
- Safety net programs (social insurance + anti-poverty programs): social security, SNAP[6], Medicare/Medicaid, WIC[7], etc.[8]

[6] Supplemental Nutrition Assistance Program, aka "food stamp"

[7] Women, Infant, Children

[8] My wife and I received support from WIC and Medicaid (Medical) when we had our daughter as both graduate students. So, I had personal experience with these programs.

- Many safety net programs now operate via digital services (i.e., websites and mobile applications). There are 5,300 .gov websites as of 2014. We live in a digital era, but ... :

- Many safety net programs now operate via digital services (i.e., websites and mobile applications). There are 5,300 .gov websites as of 2014. We live in a digital era, but ... :
  - The fundamental structure of these public services remains:

- Many safety net programs now operate via digital services (i.e., websites and mobile applications). There are 5,300 .gov websites as of 2014. We live in a digital era, but ... :
  - The fundamental structure of these public services remains:
    - FORM (you need to fill out a form)

- Many safety net programs now operate via digital services (i.e., websites and mobile applications). There are 5,300 .gov websites as of 2014. We live in a digital era, but ... :
  - The fundamental structure of these public services remains:
    - FORM (you need to fill out a form)
    - FLOW (you need to follow instructions)

- Therefore, the administrative burdens remain, too (cousin concepts: time tax, red tape) (Herd and Moynihan, 2019)

- Therefore, the administrative burdens remain, too (cousin concepts: time tax, red tape) (Herd and Moynihan, 2019)
  - Learning cost (e.g., navigating online forms)

- Therefore, the administrative burdens remain, too (cousin concepts: time tax, red tape) (Herd and Moynihan, 2019)
  - Learning cost (e.g., navigating online forms)
  - Compliance cost (e.g., interview requirements for SNAP)

- Therefore, the administrative burdens remain, too (cousin concepts: time tax, red tape) (Herd and Moynihan, 2019)
  - Learning cost (e.g., navigating online forms)
  - Compliance cost (e.g., interview requirements for SNAP)
  - Psychological cost (e.g., waiting anxiously for a case decision)

- The status quo of **safety net experience**: strong demand for change met by (still) mismatched supply

- The status quo of **safety net experience**: strong demand for change met by (still) mismatched supply
- **Strong demand**: Safety net applicants/participants have difficulty overcoming these challenges as poverty creates a scarcity mindset and tunnel vision (=low mental bandwidth) (Mullainathan and Shafir, 2013).

- The status quo of **safety net experience**: strong demand for change met by (still) mismatched supply
- **Strong demand**: Safety net applicants/participants have difficulty overcoming these challenges as poverty creates a scarcity mindset and tunnel vision (=low mental bandwidth) (Mullainathan and Shafir, 2013).
- **Mismatched supply**: Many safety net programs are designed in a way to help agencies comply with rules and procedures, not support their clients (contrary to the industry best practices: a.k.a. human-centered design) (Bagley, 2019; Pahlka, 2023).

- Three non-exhaustive options for burden reduction:

- Three non-exhaustive options for burden reduction:
  1. Messaging people

- Three non-exhaustive options for burden reduction:
  1. Messaging people
  2. Redesining forms and flows

- Three non-exhaustive options for burden reduction:
  1. Messaging people
  2. Redesining forms and flows
  3. Automating back-end workflows

- Three non-exhaustive options for burden reduction:
  1. Messaging people
  2. Redesining forms and flows
  3. Automating back-end workflows
- The key in scoping is balancing between implement costs and potential impacts

- 1. Messaging people (via emails/texts)

- 1. Messaging people (via emails/texts)
  - Cost: Low-cost (emails are almost free, texts are cheap, 1-5 cents per text depending on APIs)

- 1. Messaging people (via emails/texts)
  - Cost: Low-cost (emails are almost free, texts are cheap, 1-5 cents per text depending on APIs)
  - Impact: Generally not so great (avg 1.4 pp) (DellaVigna and Linos, 2022). Yet, sometimes, there are surprisingly big impacts (up to 10 pp) in unsaturated environments where information matters (De La Rosa et al., 2021; Giannella et al., 2023; Kim et al., N.d.).

- 1. Messaging people (via emails/texts)
  - Cost: Low-cost (emails are almost free, texts are cheap, 1-5 cents per text depending on APIs)
  - Impact: Generally not so great (avg 1.4 pp) (DellaVigna and Linos, 2022). Yet, sometimes, there are surprisingly big impacts (up to 10 pp) in unsaturated environments where information matters (De La Rosa et al., 2021; Giannella et al., 2023; Kim et al., N.d.).
  - Dependency: agency buy-in and support (consent, sampling frame, etc.),

- 1. Messaging people (via emails/texts)
  - Cost: Low-cost (emails are almost free, texts are cheap, 1-5 cents per text depending on APIs)
  - Impact: Generally not so great (avg 1.4 pp) (DellaVigna and Linos, 2022). Yet, sometimes, there are surprisingly big impacts (up to 10 pp) in unsaturated environments where information matters (De La Rosa et al., 2021; Giannella et al., 2023; Kim et al., N.d.).
  - Dependency: agency buy-in and support (consent, sampling frame, etc.),
  - Risk: regulations (e.g., FCC regulates texting) and inter-agency misalignment (if you want to scale)

- 2. Redesigning forms and flows (Moynihan et al., 2022): varying costs and impacts depending on scales

- 2. Redesigning forms and flows (Moynihan et al., 2022): varying costs and impacts depending on scales
  - Cost: small (modifying one section in an existing online form) to big (developing a new application website (e.g., GetCalFresh)) depending on project scopes

- 2. Redesigning forms and flows (Moynihan et al., 2022): varying costs and impacts depending on scales
  - Cost: small (modifying one section in an existing online form) to big (developing a new application website (e.g., GetCalFresh)) depending on project scopes
  - Impact: small (1-2 pp in reducing application drop-off rate) to big (reducing application completion time from 1 hour to 10 minutes)

- 2. Redesigning forms and flows (Moynihan et al., 2022): varying costs and impacts depending on scales
  - Cost: small (modifying one section in an existing online form) to big (developing a new application website (e.g., GetCalFresh)) depending on project scopes
  - Impact: small (1-2 pp in reducing application drop-off rate) to big (reducing application completion time from 1 hour to 10 minutes)
  - Dependency: agency buy-in and support (if you run experiments on your products, these problems don't exist), user research, design, and engineering supports are essential

- 2. Redesigning forms and flows (Moynihan et al., 2022): varying costs and impacts depending on scales
  - Cost: small (modifying one section in an existing online form) to big (developing a new application website (e.g., GetCalFresh)) depending on project scopes
  - Impact: small (1-2 pp in reducing application drop-off rate) to big (reducing application completion time from 1 hour to 10 minutes)
  - Dependency: agency buy-in and support (if you run experiments on your products, these problems don't exist), user research, design, and engineering supports are essential
  - Risk: gatekeepers (e.g., vendors responsible for these services)

- 3. Automating back-end workflows (e.g., Medicaid ex-parte renewals):

[9] For more information, see this blog post by Don Moynihan (2023): `https://donmoynihan.substack.com/p/using-automatic-renewals-to-reduce`

- 3. Automating back-end workflows (e.g., Medicaid ex-parte renewals):
  - Cost: It takes time to find an entry point. Other changes add another layer to the existing system. This approach brings more fundamental changes (e.g., eligibility determination), so it may also face stronger resistance (Moynihan, 2022).

[9] For more information, see this blog post by Don Moynihan (2023): `https://donmoynihan.substack.com/p/using-automatic-renewals-to-reduce`

- 3. Automating back-end workflows (e.g., Medicaid ex-parte renewals):
  - Cost: It takes time to find an entry point. Other changes add another layer to the existing system. This approach brings more fundamental changes (e.g., eligibility determination), so it may also face stronger resistance (Moynihan, 2022).
  - Impact: When implemented, it delivers HUGE impacts (e.g., million people, billion dollars)[9]

[9]For more information, see this blog post by Don Moynihan (2023): `https://donmoynihan.substack.com/p/using-automatic-renewals-to-reduce`

- 3. Automating back-end workflows (e.g., Medicaid ex-parte renewals):
  - Cost: It takes time to find an entry point. Other changes add another layer to the existing system. This approach brings more fundamental changes (e.g., eligibility determination), so it may also face stronger resistance (Moynihan, 2022).
  - Impact: When implemented, it delivers HUGE impacts (e.g., million people, billion dollars)[9]
  - Dependency: agency buy-in

[9] For more information, see this blog post by Don Moynihan (2023): `https://donmoynihan.substack.com/p/using-automatic-renewals-to-reduce`

- 3. Automating back-end workflows (e.g., Medicaid ex-parte renewals):
  - Cost: It takes time to find an entry point. Other changes add another layer to the existing system. This approach brings more fundamental changes (e.g., eligibility determination), so it may also face stronger resistance (Moynihan, 2022).
  - Impact: When implemented, it delivers HUGE impacts (e.g., million people, billion dollars)[9]
  - Dependency: agency buy-in
  - Risk: the usual gatekeepers plus changing policy and political environments on eligibility

[9] For more information, see this blog post by Don Moynihan (2023): `https://donmoynihan.substack.com/p/using-automatic-renewals-to-reduce`

- The secret power of data scientists (on insights teams) is writing.[10]

[10] Writing has helped Amazon increase the productivity of their economists supporting decision-making. See the 2022 interview of Kyle Kretschman (head of economics at Spotify, previously at Amazon): `https://causalinf.substack.com/p/s1e27-interview-with-kyle-kretschman`

- The secret power of data scientists (on insights teams) is writing.[10]
- Write a short document (one- or two-pager) to articulate the potential benefits (learning, decision, and impacts) and implementation costs (which teams and staff need to be involved, how much, and how long)

[10] Writing has helped Amazon increase the productivity of their economists supporting decision-making. See the 2022 interview of Kyle Kretschman (head of economics at Spotify, previously at Amazon): `https://causalinf.substack.com/p/s1e27-interview-with-kyle-kretschman`

- In the proposal, map problems to research designs

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)
  - Doing so takes less than an hour using a website like *Open Science Framework* (`https://osf.io/`) or *AsPredicted* (`https://aspredicted.org/`)

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)
  - Doing so takes less than an hour using a website like *Open Science Framework* (`https://osf.io/`) or *AsPredicted* (`https://aspredicted.org/`)
  - Pre-analysis benefits:

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)
  - Doing so takes less than an hour using a website like *Open Science Framework* (`https://osf.io/`) or *AsPredicted* (`https://aspredicted.org/`)
  - Pre-analysis benefits:
    - Transparency

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)
  - Doing so takes less than an hour using a website like *Open Science Framework* (`https://osf.io/`) or *AsPredicted* (`https://aspredicted.org/`)
  - Pre-analysis benefits:
    - Transparency
    - Accountability

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)
  - Doing so takes less than an hour using a website like *Open Science Framework* (`https://osf.io/`) or *AsPredicted* (`https://aspredicted.org/`)
  - Pre-analysis benefits:
    - Transparency
    - Accountability
    - Aligning learning objectives among stakeholders

- In the proposal, map problems to research designs
- You can easily expand it later and turn it into a pre-analysis plan (=analysis plan before collecting data)
  - Doing so takes less than an hour using a website like *Open Science Framework* (`https://osf.io/`) or *AsPredicted* (`https://aspredicted.org/`)
  - Pre-analysis benefits:
    - Transparency
    - Accountability
    - Aligning learning objectives among stakeholders
    - Example: the pre-analysis template of the U.S. federal Office of Evaluation Sciences (OES)

Plan

Implemetation

- Implementation: turning research objectives into research outcomes

Implemetation

- Implementation: turning research objectives into research outcomes
- We make plans, but some plans are bound to fail.

Implemetation

- Implementation: turning research objectives into research outcomes
- We make plans, but some plans are bound to fail.
- Like all surveys are imperfect (used random sampling, but non-response bias exists), almost every experiment is imperfect (not all people follow instructions or follow them as we intended).

Implemetation

- Implementation: turning research objectives into research outcomes
- We make plans, but some plans are bound to fail.
- Like all surveys are imperfect (used random sampling, but non-response bias exists), almost every experiment is imperfect (not all people follow instructions or follow them as we intended).
- Understanding the limitations of an experiment helps to scope learning (differentiating knowns from unknowns) (Wong et al., 2022).

- When best options are not available, improvise. Perfect is the enemy of good.

[11] This is also how many natural experiment designs work.

- When best options are not available, improvise. Perfect is the enemy of good.
  - **Situation**: Suppose you want to text SNAP clients randomly.

[11] This is also how many natural experiment designs work.

- When best options are not available, improvise. Perfect is the enemy of good.
  - **Situation**: Suppose you want to text SNAP clients randomly.
  - **Problem**: You need to randomize these clients in a partner agency's back-end system, but there's no randomize function (or if you try to do it, it will take substantial engineering time and effort and cause friction and misalignment).

[11] This is also how many natural experiment designs work.

- When best options are not available, improvise. Perfect is the enemy of good.
  - **Situation**: Suppose you want to text SNAP clients randomly.
  - **Problem**: You need to randomize these clients in a partner agency's back-end system, but there's no randomize function (or if you try to do it, it will take substantial engineering time and effort and cause friction and misalignment).
  - **Solution**: Consider leveraging arbitrary aspects of a program administration (e.g., odd/even application number, odd/even case number; if these identifiers are not available, you can use the last digit of timestamps)[11]

[11] This is also how many natural experiment designs work.

- Prototype, refine, and monitor implementation:

- Prototype, refine, and monitor implementation:
  - Write a user story (step-by-step hypotheses of user behaviors) and check whether users follow the scenario that you intended with engineers, designers, and user experience researchers (or caseworkers and users, depending on project scopes)

- Prototype, refine, and monitor implementation:
  - Write a user story (step-by-step hypotheses of user behaviors) and check whether users follow the scenario that you intended with engineers, designers, and user experience researchers (or caseworkers and users, depending on project scopes)
  - Do internal testing and demo(s) (=prototyping). If you redesign a form, do people navigate the form as you intended (i.e., user testing)?

- Prototype, refine, and monitor implementation:
  - Write a user story (step-by-step hypotheses of user behaviors) and check whether users follow the scenario that you intended with engineers, designers, and user experience researchers (or caseworkers and users, depending on project scopes)
  - Do internal testing and demo(s) (=prototyping). If you redesign a form, do people navigate the form as you intended (i.e., user testing)?
  - Expect failures and make the implementation plan robust. e.g., were these messages delivered successfully if you message people? Are there any security issues?

- Don't peak data

- Don't peak data
  - Monitoring implementation is not the same as stopping an experiment when you find a statistically significant result (e.g., based on a p-value).

- Don't peak data
  - Monitoring implementation is not the same as stopping an experiment when you find a statistically significant result (e.g., based on a p-value).
  - Peeking data causes false positive findings (even if you run an A/A test, if you run it for a while, you will find a significant outcome) (Johari, Pekelis and Walsh, 2015).

- Don't peak data
  - Monitoring implementation is not the same as stopping an experiment when you find a statistically significant result (e.g., based on a p-value).
  - Peeking data causes false positive findings (even if you run an A/A test, if you run it for a while, you will find a significant outcome) (Johari, Pekelis and Walsh, 2015).
- Resist the temptation! Reliable null findings are more valuable than noisy false positives. These findings help prioritize decisions by not doing low to zero-impact projects.

Plan

Measurement

- Measure what matters (Doerr, 2018)

Measurement

- Measure what matters (Doerr, 2018)
  - In safety net contexts: people, burden, and benefits

Measurement

- Measure what matters (Doerr, 2018)
  - In safety net contexts: people, burden, and benefits
- Set primary (e.g., direct impacts) and secondary learning objectives (e.g., indirect impacts)

Measurement

- Measure what matters (Doerr, 2018)
  - In safety net contexts: people, burden, and benefits
- Set primary (e.g., direct impacts) and secondary learning objectives (e.g., indirect impacts)
- Set stopping conditions (e.g., **no harm!**)

Measurement

- Measure what matters (Doerr, 2018)
  - In safety net contexts: people, burden, and benefits
- Set primary (e.g., direct impacts) and secondary learning objectives (e.g., indirect impacts)
- Set stopping conditions (e.g., **no harm!**)
  - Some experiments are better to be stopped if they (unintentionally) harm rather than help people.

Measurement

- Measure what matters (Doerr, 2018)
  - In safety net contexts: people, burden, and benefits
- Set primary (e.g., direct impacts) and secondary learning objectives (e.g., indirect impacts)
- Set stopping conditions (e.g., **no harm!**)
  - Some experiments are better to be stopped if they (unintentionally) harm rather than help people.
  - See Code for America GetCalFesh team's test and learn principles

Measurement

- Measure what matters (Doerr, 2018)
  - In safety net contexts: people, burden, and benefits
- Set primary (e.g., direct impacts) and secondary learning objectives (e.g., indirect impacts)
- Set stopping conditions (e.g., **no harm!**)
  - Some experiments are better to be stopped if they (unintentionally) harm rather than help people.
  - See Code for America GetCalFesh team's test and learn principles
  - See The Urban Institute's The Do No Harm Project

- Details matter: Always pay attention to the units when measuring impacts from safety net field experiments. Benefit applications (cases) != applicants. One case may have several dependents (so you must observe or infer household sizes). These are also related to calculating benefit amounts.

- Easier metrics do not imply better (or more important) metrics (Muller, 2018).

- Easier metrics do not imply better (or more important) metrics (Muller, 2018).
- Some impacts have direct monetary values (e.g., dollar amounts), others not (e.g., time saved).

- Administrative data are essential in impact tracking but far from perfect.

- Administrative data are essential in impact tracking but far from perfect.
- Everyone knows pieces but not the whole picture. You need to put the puzzles together.

- Administrative data are essential in impact tracking but far from perfect.
- Everyone knows pieces but not the whole picture. You need to put the puzzles together.
  - Often, there is no dictionary.

- Administrative data are essential in impact tracking but far from perfect.
- Everyone knows pieces but not the whole picture. You need to put the puzzles together.
  - Often, there is no dictionary.
  - Even if a data dictionary exists, it could be outdated or incomplete.

- Administrative data are essential in impact tracking but far from perfect.
- Everyone knows pieces but not the whole picture. You need to put the puzzles together.
  - Often, there is no dictionary.
  - Even if a data dictionary exists, it could be outdated or incomplete.
  - Check the quality of administrative data by talking to technical (e.g., IT/business intelligent people) and domain experts (e.g., case workers).

- Document and share knowledge.

- Document and share knowledge.
- Build your knowledge base of the data systems (for yourself, your team, and partners):

- Document and share knowledge.
- Build your knowledge base of the data systems (for yourself, your team, and partners):
  - Document the data access process (this helps your transition, too)

- Document and share knowledge.
- Build your knowledge base of the data systems (for yourself, your team, and partners):
  - Document the data access process (this helps your transition, too)
  - Document the SQL queries (if you directly query the agency's database) so that someone else can replicate and build upon your workflows.

- Document and share knowledge.
- Build your knowledge base of the data systems (for yourself, your team, and partners):
  - Document the data access process (this helps your transition, too)
  - Document the SQL queries (if you directly query the agency's database) so that someone else can replicate and build upon your workflows.
  - Document other key insights you've gained from your work. If you want your project to be sustained (policy adoption) and expanded (policy diffusion), your implicit knowledge must become explicit.

- Integrate insights to improve measurements:

- Integrate insights to improve measurements:
  - **Scaling evidence**: Experiments (evaluation) often come after qualitative / user experience research (discovery). Early-stage discoveries help understand contexts and accurately measure and interpret quantitative outcomes.

- Integrate insights to improve measurements:
  - **Scaling evidence**: Experiments (evaluation) often come after qualitative / user experience research (discovery). Early-stage discoveries help understand contexts and accurately measure and interpret quantitative outcomes.
  - **Bridging evidence**: Experiments are often insufficient to understand the underlying mechanism (if texting helped people show up for SNAP interview, why?). A follow-up survey (and in-depth interviews) could be useful in answering such questions.

Plan

Communication

- Communication creates values. Data don't speak for themselves.

Communication

- Communication creates values. Data don't speak for themselves.
- Basic structure: stories + evidence

Communication

- Communication creates values. Data don't speak for themselves.
- Basic structure: stories + evidence
  - It is important to demonstrate evidence and tell stories of the people behind numbers (their proportions vary by context).

Communication

- Communication creates values. Data don't speak for themselves.
- Basic structure: stories + evidence
  - It is important to demonstrate evidence and tell stories of the people behind numbers (their proportions vary by context).
  - Example: Why Californians need food assistance: The stories behind the numbers (Code for America)

- There are different types of deliverables suitable for different audiences.
  - Fellow data scientists: R/Python notebooks (code + analysis)
  - Stakeholders and partners: **slide deck (!!!)** + briefs (recommendations + key findings)
  - Academic audience (if you end up writing papers): journal articles/conference proceedings (more theories or methods or more rigor to evidence)
  - General audience (data scientists usually don't handle this): blog posts, newspaper articles, interviews

Plan

Takeaways

- Data science helps improve the U.S. safety net experience by sizing opportunities, designing rigorous research, and measuring impacts for continuous product and service improvement.[12]

[12] Since I worked at Code for America, the examples I provided in these slides focus on Code for America, but there are many other organizations active in this policy space, such as United States Digital Service, GSA's Office of Evaluation Sciences, Nava (PBC), Georgetown's Better Government Lab, Urban Institute, Mathematica and others.

- Running a field experiment is challenging because it requires internal and external alignments on scoping, implementation, measurement, and communication.

- Running a field experiment is challenging because it requires internal and external alignments on scoping, implementation, measurement, and communication.
- Caveat: We didn't discuss responsible data access and sharing in these slides, but they are critical for legal and ethical reasons (see The Urban Institute's Safe Data Technologies project and Bowen (2021)).

Thank you

Comments or questions?
E-mail: jkim638@jhu.edu

Plan

Bagley, Nicholas. 2019. "The procedure fetish." *Michigan Law Review* pp. 345–401.

Bowen, Claire McKay. 2021. *Protecting your privacy in a data-driven world*. Chapman and Hall/CRC.

Christian, Brian. 2021. *The alignment problem: How can machines learn human values?* Atlantic Books.

De La Rosa, Wendy, Eesha Sharma, Stephanie M Tully, Eric Giannella and Gwen Rino. 2021. "Psychological ownership interventions increase interest in claiming government benefits." *Proceedings of the National Academy of Sciences* 118(35):e2106357118.

DellaVigna, Stefano and Elizabeth Linos. 2022. "RCTs to scale: Comprehensive evidence from two nudge units." *Econometrica* 90(1):81–116.

Doerr, John. 2018. *Measure what matters: How Google, Bono, and the Gates Foundation rock the world with OKRs*. Penguin.

Fisher, Ronald. 1935. *The Design of Experiments*. Oliver and Boyd.

Giannella, Eric, Tatiana Homonoff, Gwen Rino and Jason Somerville. 2023. Administrative burden and procedural denials: Experimental evidence from SNAP. Technical report National Bureau of Economic Research Cambridge, MA.

Gneezy, Uri and John A List. 2013. *The why axis: Hidden motives and the undiscovered economics of everyday life*. Random House.

Herd, Pamela and Donald P Moynihan. 2019.

Administrative burden: Policymaking by other means. Russell Sage Foundation.

Imbens, Guido W and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Johari, Ramesh, Leo Pekelis and David J Walsh. 2015. "Always valid inference: Bringing sequential analysis to A/B testing." *arXiv preprint arXiv:1512.04922* .

Kim, Jae Yeon, Pamela Herd, Sebastian Jilke, Donald Moynihan and Kerry Rodden. N.d. "Administrative Checkpoints, Burdens, and Human-centered Design: Increasing Interview Access to Raise SNAP Participation." *Working paper*. Forthcoming.

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is your estimand? Defining the

target quantity connects statistical evidence to theory." *American Sociological Review* 86(3):532–565.

Moynihan, Donald, Eric Giannella, Pamela Herd and Julie Sutherland. 2022. "Matching to categories: Learning and compliance costs in administrative processes." *Journal of Public Administration Research and Theory* 32(4):750–764.

Moynihan, Donald P. 2022. "How Can Scholars Help to Embed Institutions of Public-Sector Change?(Or Things I Wish I'd Known When I Was a Grad Student)." *Perspectives on Public Management and Governance* 5(4):276–287.

Mullainathan, Sendhil and Eldar Shafir. 2013. *Scarcity: Why having too little means so much*. Macmillan.

Muller, Jerry. 2018. *The tyranny of metrics*. Princeton University Press.

Pahlka, Jennifer. 2023. *Recoding America: why government is failing in the digital age and how we can do better*. Metropolitan Books.

Simon, Herbert A. 1996. *Models of my life*. MIT press.

Wong, Jeffrey, Jasmine Nettiksimmons, Jiannan Lu and Katherine Livins. 2022. "Addressing Hidden Imperfections in Online Experimentation." *arXiv preprint arXiv:2209.00649* .