Design & Governance of Data-Driven Algorithms

Sarah H. Cen
MIT EECS
Advised by Aleksander Mądry and Devavrat Shah

October 2, 2023
Young Researchers Workshop, Cornell

The Role of Data-Driven Algorithms

Should data-driven algorithms intervene on high-stakes decisions?

Should social media platforms choose how election-related content is disseminated?

When should the government regulate the AI industry, and when should it abstain?

Normative Factors

Practical Considerations

Data-Driven Algorithms

Design

What is feasible from
a design perspective?
(e.g., when do trade-offs exist)

[**C**S'22, **C**IM'23, Z**C**S'23,
A**C**SY'23, **C**S'21, **C**MS'23]

Governance

What is feasible from a
governance perspective?
(e.g., what can be regulated)

[**C**S'21, **C**IM'22, **C**R'23,
CSFMM'23, **C**HIMSV'23, CMS'23]

Today: Auditing

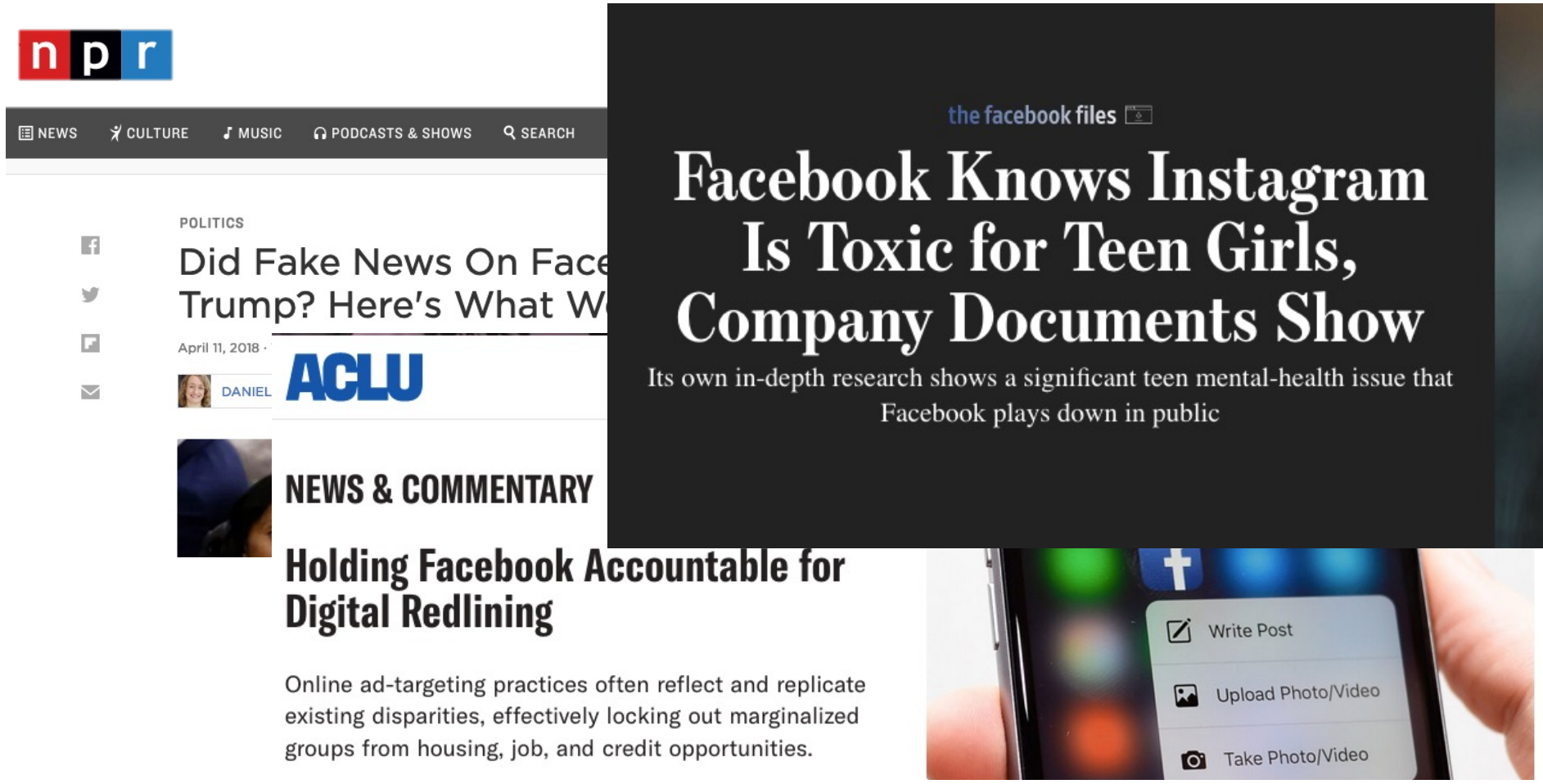
Goal: Verify that a data-driven algorithm satisfies some criteria.

Input: Criteria (e.g., given by law)
Constraint: Black-box access

Questions:
1. To what extent does the audit test for the given criteria?
2. How much data does the audit need?
3. Are there side effects?

Case Study: Auditing Social Media

Cen and Shah, 2021



There are rising calls for social media regulations.

Calls to Regulate

Ex 1: Ads not be based on user's sexual orientation.

Ex 2: Info on public health (e.g., COVID-19) not reflect political affiliation.

Ex 3: Not sway voting preferences beyond serving as a social network.

Translating desiderata → audit is difficult

- Performance cost
- Censorship
- User privacy
- Trade secrets

Main contribution: data-driven auditing procedure

- Strong statistical guarantees
- Not necessarily a performance-audit trade-off
- Incentivizes platform to inject content diversity
- Requires only black-box access
- Does not remove content or require user data

Problem setup

Inputs $\boldsymbol{x}$

(user attributes, advertisers, etc.)

Filtering algorithm $\mathcal{F}$

Feed $Z = \mathcal{F}(\boldsymbol{x})$

$$Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_m\}$$

$$\boldsymbol{z}_i \sim p_z(\,\cdot\,; \theta(\boldsymbol{x}))\ ,\ \theta(\boldsymbol{x}) \in \Theta$$

Black-box access: Run $\mathcal{F}$ on $\{\boldsymbol{x}_j\}$ and observe $\{Z_j\}$.

Auditor's task

Given a CF desideratum & black-box access to $\mathcal{F}$, check if platform complies.

Counterfactual desideratum

hypothetical!

"Algorithm $\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."

Articles with medical advice on COVID-19 must be robust to user's political affiliation.

Articles shown by $\mathcal{F}$ that have medical advice on COVID-19 should be **similar** whether a user is left- or right-leaning.

Counterfactual desideratum

hypothetical!

"Algorithm $\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."

The platform should not sway voting beyond serving as a social network.

Election-related posts that $\mathcal{F}$ injects should be **similar** to those a user would see from its social network (without filtering).

What is an appropriate notion of "similarity" ?

$$\|Z_i - Z_i'\|_q < \delta \ ?$$

Observation: Algorithmic filtering is powerful (sometimes harmful) because information influences decisions.

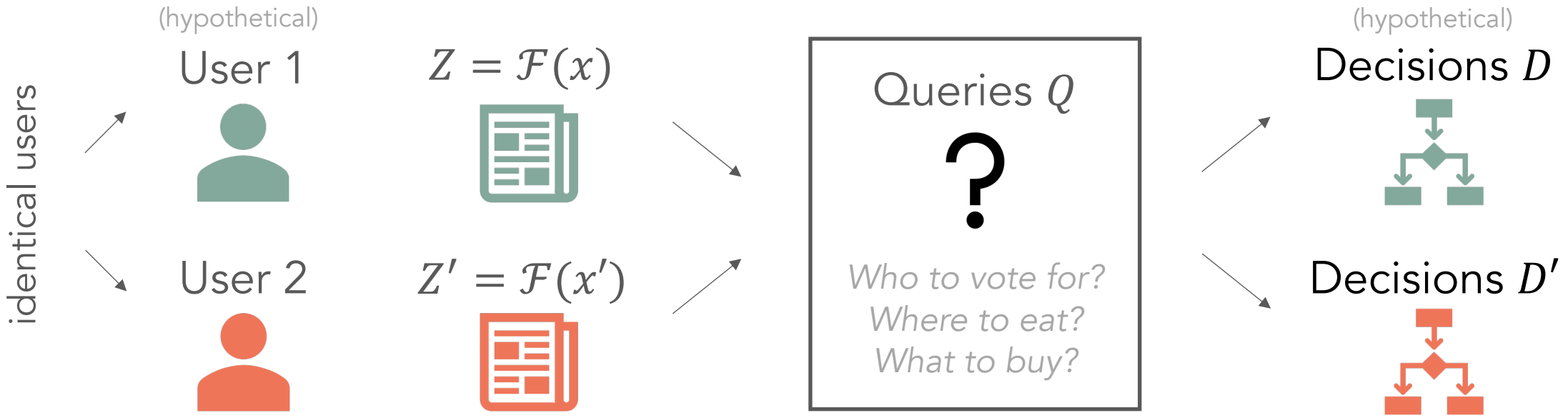
Examples. The content that $\mathcal{F}$ filters affects ...

- What a user eats
- What they buy
- How the user votes

Implication. Should enforce **similarity** between $\mathcal{F}(\boldsymbol{x})$ and $\mathcal{F}(\boldsymbol{x}')$ w.r.t. the outcome of interest: **the users' decisions**.

Decision robustness

CF Reg: "$\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."



$\mathcal{F}$ is *decision-robust* to $(\boldsymbol{x}, \boldsymbol{x}')$ if and only if, for any Q, one cannot confidently determine that $\boldsymbol{x} \neq \boldsymbol{x}'$ from D and D'.

↳ can formalize as hypothesis test

Auditing procedure

"Algorithm $\mathcal{F}$ must behave similarly under $\boldsymbol{x}$ and $\boldsymbol{x}'$ for all $(\boldsymbol{x}, \boldsymbol{x}') \in S$."

Inputs: $\mathcal{F}$ $\boldsymbol{x}$ $\boldsymbol{x}'$ Θ ϵ

$\tilde{\boldsymbol{\theta}} \leftarrow \mathcal{L}^{+}(\mathcal{F}(\mathbf{x}))$;

$\tilde{\boldsymbol{\theta}}' \leftarrow \mathcal{L}^{+}(\mathcal{F}(\mathbf{x}'))$;

if $(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')^{\top} I(\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}') \geq \frac{2}{m}\chi_r^2(1-\epsilon)$ **then**

  Does not pass the test for $(\mathbf{x}, \mathbf{x}')$;

end

Passes the test for $(\mathbf{x}, \mathbf{x}')$;

Minimum-variance unbiased estimator (MVUE)

Advantages

- Only needs black-box access to $\mathcal{F}$.
- Does not require access to users or their personal data.
- Modular. Can scale up for any (x, x') pairs.
- Intuitive tunable parameter. ϵ is false positive rate.
- No content removal.

Guarantee on how well the audit enforces the regulation.

Theorem (informal). If the filtering algorithm $\mathcal{F}$ passes the audit, then $\mathcal{F}$ is guaranteed to be approximately asymptotically decision-robust.

Alternative statement

If $\mathcal{F}$ does not pass the audit, then the auditor is $(1 - \epsilon)$-confident that $\mathcal{F}$ is not decision-robust as $m \to \infty$.

Takeaways

- The audit enforces strong similarity between $\mathcal{F}(\boldsymbol{x})$ and $\mathcal{F}(\boldsymbol{x}')$.
- ϵ is the allowable false positive rate: increasing ϵ increases strictness.

Why the MVUE?

Proposition (informal). Faced with a decision between a finite number of options, the decision of the hypothetical user whose belief after viewing content Z is given by the MVUE is more sensitive to Z than any other user.

Takeaway

MVUE = user whose decisions are **most sensitive** to the content they see.

The MVUE allows us to reason about how content affects users *without access to users' decisions* → expensive or unethical to obtain.

There isn't a regulation-performance trade-off.

Theorem (informal). Consider a finite feed. If performance is independent of elements in $\boldsymbol{\theta}$ that can increase the Fisher information and the available content is diverse, then there is no regulation-performance trade-off.

Takeaway

Platform can pass audit without sacrificing performance.

Content diversity can reduce the cost of regulation

The lower the content diversity of Z and Z', the more easily an auditor can distinguish between how $\mathcal{F}$ behaves under $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Design & Governance of Human-Facing Algorithms

Case Study: Auditing Social Media

- Black-box auditing procedure
- Audit is consistent with existing laws

Extensions

- Instantiated a viable regulation [**C**MS ’23]
- Running a live audit this month
- Auditing from dataset [A**C**SY ’23]

Here, we asked: What is feasible from an auditing perspective?

Interplay between design & governance is going to be important

Thank you!

shcen@mit.edu