# End-to-end Auditing of Decision Pipelines

Benjamin Laufer, Emma Pierson and Nikhil Garg

Cornell Tech

## Background

### Many consequential policy decisions are multi-faceted and distributed over time.

- Policies often have to deal with both **allocation** and **scheduling**.
- Policies often happen along a **pipeline** consisting of a sequence of decisions [1, 2].
- Examples of such policies exist in many critical domains with equity concerns, including **education** and **criminal justice**.

### Pipelines are difficult to study empirically.

- Pipelines can be **long and complex**, often spanning many years and multiple decision-makers
- There is **substantial unobserved confounding** between stages.

## Domain: NYC Dept. of Parks and Recreation (DPR)

### Cities use *resident crowdsourcing.*

- The public reports problems such as downed trees or power lines to the government.
- NYC's 311 system received over 2.6 million requests in 2021.

### The Parks Department fields requests on *street trees.*

- **Street Trees are important**: NYC's 700,000 street trees provide life-saving temperature reductions, and when they fall they can cause significant damage, disruption and death.
- **Requests trigger a pipeline of decisions**: From 100,000 annual requests, DPR makes a sequence of bureaucratic decisions: an *inspection* involving an agency member visiting the incident location, and then a *work order* to fix the issue if necessary [3].
- **Domain Advantages**: Decision pipelines are centralized and short (weeks/months). There is arguably little unobserved confounding. Regular conversations with DPR officials provide us with vital context and an avenue to *change* operations.

## References

[1] Eshwar Ram Arunachaleswaran, Sampath Kannan, Aaron Roth, and Juba Ziani. Pipeline interventions. *arXiv preprint arXiv:2002.06592*, 2020.

[2] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[3] Zhi Liu and Nikhil Garg. Equity in resident crowdsourcing: Measuring under-reporting without ground truth data. *arXiv preprint arXiv:2204.08620*, 2022.

[4] Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*, 2018.

## Main Contributions

- We develop **a method for auditing decision pipelines end-to-end** with conditional parity tests at each stage.
- Using data on NYC DPR's decision pipeline, preliminary evidence suggests there are **socio-economic disparities in allocation and scheduling decisions**.

## Parity Definitions

All definitions are provided for group attribute $g$ and pipeline events `report`, `insp`, `work order`, and `work completed`.

### Equity in Allocation Decisions

- **Inspection Parity**: Compare $\mathbb{P}(\texttt{insp}|g, \texttt{report})$
- **Work Order Parity**: Compare $\mathbb{P}(\texttt{work order}|g, \texttt{insp})$
- **Work Completion Parity**: Compare $\mathbb{P}(\texttt{work completed}|g, \texttt{work order})$

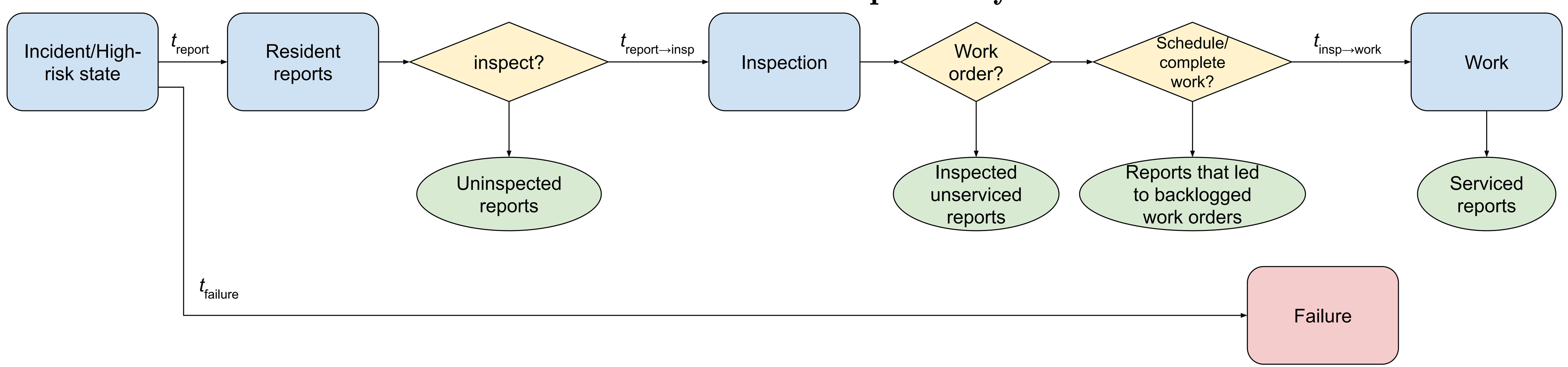
### Equity in Scheduling Decisions

- **Inspection Time Parity**: Compare $\mathbb{E}[t_{\texttt{report}\rightarrow\texttt{insp}}|g]$
- **Work Time Parity**: Compare $\mathbb{E}[t_{\texttt{insp}\rightarrow\texttt{work}}|g]$

### Risk-adjusted Regression Tests

In addition to demographic parity, we use risk-adjusted regression tests [4] in which we include the (predicted or observed) risk as a regressor in order to directly compare parity among reports that are of the same risk level

**Figure 1 (below): Pipeline of NYC DPR Decisions**




## Preliminary Results

### DPR inspection allocation benefits low-income neighborhoods.

- In regression tests with and without adjusting for risk, DPR's first decision of whether or not to inspect directs extra attention to low-income census tracts.
- This observation may be explained by existing audits being directed at the inspection stage.

### However, each subsequent pipeline decision disadvantages low-income neighborhoods.

- In ordering work and completing work, the DPR is significantly less likely to allocate to low-income census tracts.
- In scheduling decisions, similar biases are observed. Reports from lower-income census tracts wait longer before being inspected and worked on, on average, compared to reports of the same risk from higher-income census tracts.

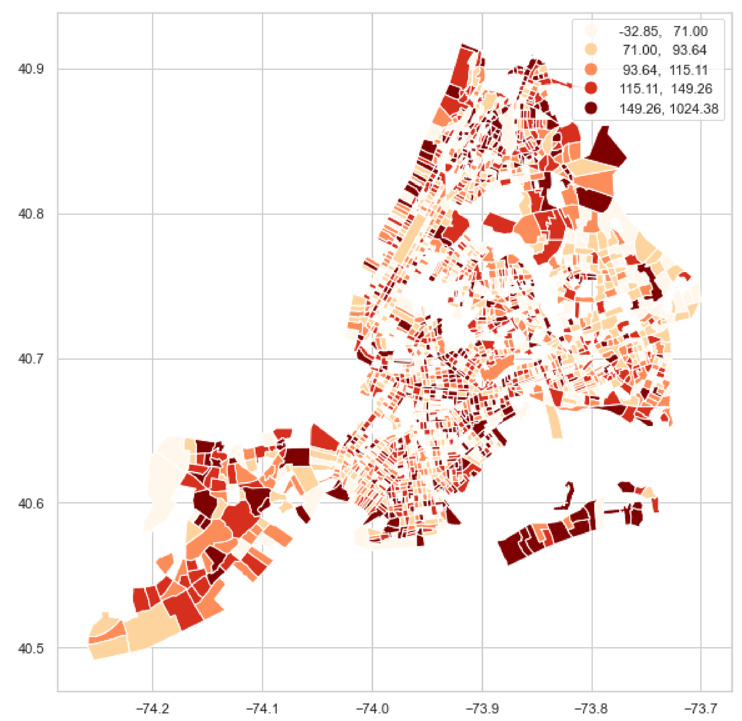



**Figure 2 (above): Delay from Report to Work Completion by Census Tract**