# The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems


## Abstract

Bringing transparency to black-box decision making systems (DMS) has been a topic of increasing research interest in recent years. Traditional active and passive approaches to make these systems transparent are often limited by scalability and/or feasibility issues. In this paper, we propose a new notion of black-box DMS transparency, named, temporal transparency, whose goal is to detect when/how the DMS policy changes over time, and is mostly invariant to the drawbacks of traditional approaches. We map our notion of temporal transparency to time-series changepoint detection methods, and develop a framework to detect policy changes in real world DMS's. Experiments on New York Stop-question-and-frisk dataset reveal a number of publicly announced and unannounced policy changes, highlighting the utility of our framework.


## 1 Introduction

In modern societies, it is widely accepted that decision making systems (DMS), particularly those whose outcomes affect people's lives, need to be *transparent*. A number of recent studies have attempted to bring transparency to black-box (opaque) decision making systems, be they driven by machines (e.g., algorithmic search and recommendation systems [7, 9]) or humans (e.g., stop and frisk decisions made by police [15, 16]). These studies attempt to reverse-engineer or infer the decision making *policy or function* ($f_{\mathrm{DMS}}$) either by (i) actively auditing the system with carefully crafted inputs and analyzing the resulting outputs [7, 9] or by (ii) passively observing the inputs and outputs of the system in operation.

The above two broad approaches to bringing transparency have their pros and cons: (i) active audits can help achieve *functional transparency*, i.e., learn the behavior of the decision function for different types of inputs, but they can be expensive and might not reveal much about the system's behavior under operational conditions (where inputs are typically drawn from specific probability distributions over the input space), (ii) passive observations of the systems' inputs and outputs, on the other hand, can help achieve *operational transparency*, but they limit studies to analyzing decision function behavior on the limited set of operational inputs.

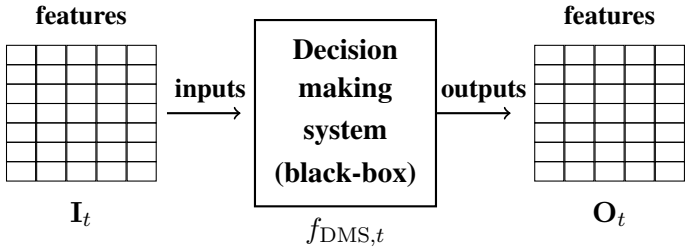


Figure 1: A traditional DMS abstraction. The algorithm is often a black-box, inputs are rarely available, while outputs may be available most of the times.

Against this background, we make the case for a new notion of transparency that we call **temporal transparency**, where the goal is to be able to detect *when and how the decision making policy or function ($f_{\mathrm{DMS}}$) changes over time*. Note that the objectives of temporal transparency are complementary but different from those of traditional functional or operational transparency. The motivating scenarios for temporal transparency are numerous.

**1. Monitoring policy change events & alerting users.** Temporal transparency enables one to track and verify when and how policies of decision making systems, such as NYPD Stop-question-and-frisk program[1] (NYPD SQF) or Facebook's newsfeed algorithm, have changed over the years [18, 17]. It would be possible to monitor whether and when an announced policy change by public or private organizations has been effected [2, 19]. Furthermore, any unannounced (or surreptitiously deployed) policy changes can be detected and used to alert civil liberties and consumer protection groups to demand greater transparency [8]. Later in this paper, we detect several instances of announced and unannounced policy changes in NYPD SQF program.

**2. Feasible when other transparency approaches aren't.** Temporal transparency can be achieved even in scenarios when traditional transparency approaches like functional or

[1] https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City

operational transparency are ineffective. For instance, consider the NYPD SQF program. It is not feasible to actively audit NYPD's decision making by generating artificial new inputs (i.e., pedestrians in NYC). One needs to rely on passively analyzing records of stops maintained by NYPD. But, as NYPD only records data for pedestrians that have been stopped and does not record data for all pedestrians that the police are observing, it is impossibly hard to infer the decision making policy (function) in its entirety. However, as we show later in the paper, these limited records are sufficient to achieve temporal transparency, i.e., robustly detect a variety of policy changes implemented by NYPD over several years.

**3. Finding targets for other transparency approaches.** By detecting the points in time when the decision making policy has changed, temporal transparency can help focus the more expensive traditional approaches to transparency (like active audits or passive input-output analysis) to the short period of time before and after the policy change events. Focusing transparency efforts on policy change events can help us better understand the magnitude and effects of the policy changes on the outcomes of the decision making system.

Intuitively, the basic idea behind detecting decision policy function changes is shown in Figure 2. We are given a time series of observed inputs and outputs to the system (where the outputs were generated from the input by the decision function) and our task is to determine if and when the decision function has changed. Any statistically significant changes in outputs over time can stem either from changes in decision function or changes in inputs. So our idea for detecting decision function changes is to look for temporal changes in outputs, where inputs remain stationary.

In this paper, we argue that the problem of detecting policy change events naturally fits existing frameworks for detecting changepoints in time-series, where the goal is to detect the optimal number of changes in a time-series of observations that best explain the observations. Time-series changepoint detection is a well-studied problem in statistics, signal processing and machine learning [3, 4, 6, 13, 20]. However, applying changepoint detection techniques to real-world datasets, subjected to noise, outliers, seasonal and weekly pattern, and different magnitudes of the detected changes, is not a straightforward task.

With the propose of tackling these challenges, we developed a framework which builds on recent advances in Bayesian changepoint detection [5, 20]. Specifically, in order to make the earlier methods robust to transitory disturbances in the observed features and aiming at detecting only significant shifts, we pose the above constraints as a *maximum a posteriori* (MAP) problem and propose a *dynamic programming* (DP) solution. Our framework also allows minimizing false positives via cross validation in the presence of ground truth (training data).

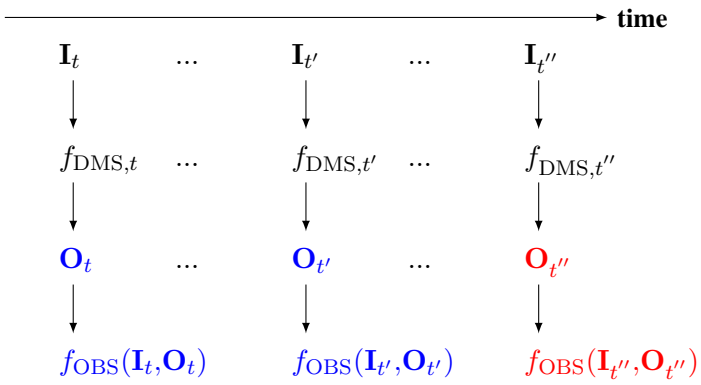


Figure 2: Temporal monitoring of a DMS. Changes in observed feature, $f_{\text{OBS}}(\mathbf{I},\mathbf{O})$, from blue to red, correspond either to changes in inputs ($\mathbf{I}$) or changes in the DMS's policy ($f_{\text{DMS}}$).

Implementation of our framework on a real world data—NYPD SQF program—provides interesting insights into the policy changes deployed by NYPD. Specifically, we show that public policy change announcements by NYPD were followed by several unannounced changes.

## 2 Detecting Policy Change Events in DMS's

Let $\mathbf{I}_t$ and $\mathbf{O}_t$ be $n\times k$ and $n'\times k'$ matrices representing the inputs and outputs of a DMS at time $t$, $n\geq n'$. Further, let $f_{\text{DMS},t}$ be the policy implemented by a DMS at time $t$, such that $\mathbf{O}_t=f_{\text{DMS},t}(\mathbf{I}_t)$. Let $x_t=f_{\text{OBS}}(\mathbf{O}_t,\mathbf{I}_t)$ be the observed feature at time $t$, where the domain of $x_t$ is dependent on $f_{\text{OBS}}$. Consider recording the observed feature $x_t$ for a period of time $[1{:}T]$. The set of observed features collected during such time period, $x_1^T$, can be considered as a time-series of data.

The problem at hand consist of finding the optimal set of changes—that is, the number of changes and their respective locations—in the time-series of the observed features, which best explain the time series $x_1^T$. This setup can leverage time-series changepoint detection frameworks. Specifically, we choose to build on recent advances in bayesian probabilistic changepoint detection setups described in [5, 20] . Adhering to the notation presented in [5], the problem above can then be formulated as:

$$({}^*m,{}^*\tau_1^{*m}) = \underset{\substack{\tau_1^m,m:1<\tau_1<\cdots<\tau_m,\\ \tau_j-\tau_{j-1}\geq d,m\in\mathcal{M}}}{\operatorname{argmax}} P(\tau_1^m,m|x_1^T), \qquad (1)$$

where ${}^*m$ and ${}^*\tau_1^{*m}$ respectively represent the optimal number of changepoints, and their location, in $x_1^T$, and $\mathcal{M}$ the set of potential number of changes.

**Detecting significant policy regimes.** Notice that, considering the characteristics of our problem—detecting *significant* policy regimes, we add an additional constraint to the traditional bayesian change detection frameworks: a minimum length $d$ of each time series segment (policy regime). The precise value of $d$ can be specified manually depending on the specific DMS under consideration.

**Solving the MAP Problem.** Let $P(x_t^s|m)$, $P(\tau_j|\tau_{j+1})$, $P(m)$ and tuning parameters, determining the sensitivity of the framework to policy changes events, be the input to the changepoint detection setting.

Right hand side of Equation 1) can be decomposed into:

$$\underset{m\in\mathcal{M}}{\operatorname{argmax}} \quad P(m)P(x_1|m) \underset{\tau_1^m:\tau_j-\tau_{j-1}\geq d}{\operatorname{argmax}} P(\tau_1^m|x_1^T,m) \quad (2)$$

The solution to the second term of 2) is yield by a dynamic program, whose recurrence relation, $j\in[1{:}m]$, is dictated by:

$$\begin{aligned} &T(j,\tau_{m-j+1}) \quad = \qquad\qquad (3)\\ &= \max_{\substack{n-(j-1)d\\ \geq\tau_{m-j+1}\geq\\ (m-j)d}} P(\tau_{m-j+1}|\tau_{m-j},x_1^T,m)T(j+1,\tau_{m-j+2}). \end{aligned}$$

The estimation of the set $\mathcal{M}$ is done through computing the cusum chart ([4, 13]) of $x_1^T$ and analyzing its first derivative. Tuning parameters influence the shape of the set $\mathcal{M}$, and therefore the sensitivity of the setting. The joint posterior probability $P(\tau_{m-j+1}|\tau_{m-j},x_1^T,m)$ is evaluated as in [5].

**Preprocessing.** In order to remove underlying noise in the input time-series $x_1^T$, and improve the reliability of the results, we apply the following preprocessing steps to $x_1^T$ before subjecting it to changepoint detection setup outlined above: identifying outliers through moving average in $x_1^T$; removing weekly patterns[2] via PCA; and smoothing $x_1^T$ using a Savitzky-Golay filter [14].

# 3 Detecting Policy Changes in NYPD SQF Program

In this section, we apply our changepoint detection framework described in Section 2 to bring temporal transparency to NYPD SQF program. The SQF program has been subject to intense public debate since its conception [1], and went through multiple publicly announced policy changes[1, 10].

[2]If needed, such as in the NYC SQF data, where the stops made by police can be affected by the time of the week

| Year | Seasonal | | Unusual inputs | Policy | |
|---|---|---|---|---|---|
| | S | W | | A | UA |
| 2006 | 1 | 1 | – | – | – |
| 2007 | 2 | 1 | 1 | – | – |
| 2008 | 1 | 1 | 1 | – | – |
| 2009 | – | 2 | – | – | 1 |
| 2010 | 1 | 1 | – | – | 2 |
| 2011 | 1 | 1 | 2 | – | 1 |
| 2012 | 2 | – | 2 | 1 | 1 |
| 2013 | – | 1 | – | – | 3 |

Table 1: List of detected changepoints from January 01 to 2006, to December 31, 2013. **S**—Summer; **W**—Winter; **A**—Announced, **UA**—Un-announced.

Our goal in this section is to track how the policy changes announced by NYPD were implemented, and explore any unannounced changes in the program. To this end, we model SQF program as a black-box DMS as follows: let $\mathbf{I}_t$ be the set of people observed at time $t$, where each individual has a set of corresponding features. Let $\mathbf{O}_t$ be a vector containing binary values indicating whether or not an observed person was stopped. Notice that we only have access to the recorded features for the set of people who were stopped. We consider the observed feature to be the number of stops per day : $x_t{=}\sum_{i\in\mathbf{I}_t}\mathbb{1}\{\mathbf{O}_t(i){=}1\}$. We focus on detecting potential policy changes by analyzing this specific observed feature $x_t$. To this end, we deployed our framework over this feature, from years 2006 to 2013. [3]

Our framework detected a total of 31 changepoints from the time period under consideration. Since the number of detected changepoints is considerably large, we systematically analyzed each of the changepoints. As a result, we were able to separate the changepoints into following categories (listed individually for each year in Table 1):

**1. Seasonal patterns.** These changepoints correspond to slight drops in number of stops made each day around mid-year (summer) and close to the end of the year (winter). This pattern persists for almost all of the years considered for the analysis. 16 out of the 31 detected changepoints fall under this category.

**2. Unusual input changes.** These changepoints *potentially* correspond to unusual changes in everyday pedestrian population of NYC. For example, we detect a changepoint on October 29, 2012, marking a consistent drop in number of stops made per day for almost one month. This drop is most

[3]The complete records of the stops made under SQF program are available at: nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_report.shtml

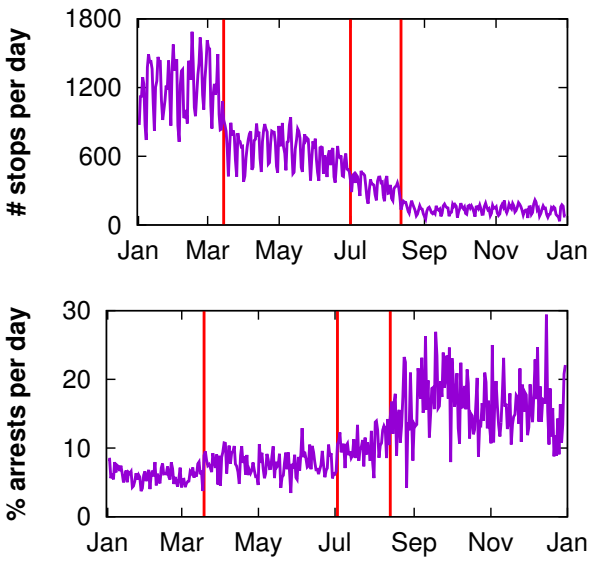


Figure 3: Changepoints detected in NYC SQF data from January 01, 2013 to December 31, 2013.

probably due to the declaration of state of emergency in NY on October 28,[4] in the face of imminent Hurricane Sandy. In fact, one day after the declaration, on October 29, 2012, the number of stops made over the city is merely 193, as compared to an average of 1147 stops per day for the previous week. Similarly, a changepoint marking a gradual increase in number of stops per day on September 22, 2011 could potentially correspond to Occupy Wall Street Movement, that started on September 17.[5] In total, 6 out of 31 changepoints map to this category.

**3. (Un)Announced Policy changes.** The changepoints that correspond to neither of the above two categories are *potentially* policy changes implemented by the NYPD. For example, we detect a gradual drop in the number of stops made per day starting March 26, 2012, which, in fact, is a consequence of a *publicly announced* policy change implemented by NYPD—where, according to NYPD, 'increased training' and staffing in 'high impact' zones results in an overall decline in number of stops [10]. Detection of this changepoint highlights the utility of our framework in verifying the policy changes announced by the governing entities.

Next, we focus on analyzing changepoints that do not map to a publicly announced policy change. In particular, we focus the year 2013. The changepoint detection framework yields 3 un-announced changes for this year. Figure 3 (top penal) shows the number of stops made per day and the detected changepoints (in the form of vertical lines). Remarkably, this series of changepoints correspond to three *abrupt* policy changes which successively brought down the number of stops per day to eventually $10\%$ of the stop rate at the beginning of the year. It is important to note that the 2013 SQF program was subject of intense debate during the 2013 Mayoral Election campaign, with a major candidate denouncing it [12] and a court stating that the SQF policy violated the constitutional rights of the citizens [11]. Consequently, these variations are likely to be associated with *unannounced* policy adjustments resulting from these events.

In addition to studying the number of stops, we also analyzed the percentage of stops leading to arrests per day in 2013. The changepoint analysis framework detects three changepoints presented in Figure 3 (bottom panel), close to the changepoints detected in the stop-rate analysis. This clear mapping between the changepoints yielded by both observed features reveals a systematic change in SQF policy by NYPD. Specifically, not only can a policy change be identified in the stop-rate, but also in the *nature* of the stops themselves.

## 4 Ongoing and Future Work

Bearing in mind the goal of exhibiting the implications of temporal transparency to real-world DMS's, the experiments carried on NYC SQF data show how to apply the framework developed to systematically detect possible policy changes events. With the hope of further generalizing the framework and catering to a wide range of real-world DMS's we plan to address the following points:

The current implementation relies on an 'offline' setting, and cannot be deployed on streaming datasets, where one might want to detect changepoints on the fly, *e.g.*, Facebook newsfeed algorithm. We are currently expanding our framework to cater to these scenarios.

The need to establish a relative degree of confidence between the changepoint detected lead us to devise a ranking mechanism. This step will allow us to direct our focus towards the change points which are most likely to correspond to policy change events.

With the objective of minimizing the occurence false positives, we explore the possibility of cross validation through the definition of an appropriate cost function.

Finally, as shown in Section 3, analyzing the structure of policy changes by jointly considering multiple observed features (number of stops, percentage of stops leading to arrests, in parallel) can provide more insights into how the DMS interplays with different features, hence revealing more information about policy changes. Hence, we plan to generalize our framework to multi-variate feature spaces.

[4] https://en.wikipedia.org/wiki/Effects_of_Hurricane_Sandy_in_New_York
[5] https://en.wikipedia.org/wiki/Occupy_Wall_Street

# References


[1] Stop-and-frisk in New York City. https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City.

[2] Al Jazeera America. Monitor: Changing NYPD stop-frisk practices a challenge. http://america.aljazeera.com/articles/2016/2/17/monitor-changing-nypd-stop-frisk-practices-a-challenge.html, February 2016.

[3] D. Barry and J. A. Hartigan. Product partition models for change point problems. *The Annals of Statistics*, pages 260–279, 1992.

[4] M. Basseville, I. V. Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

[5] P. Fearnhead. Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2):203–213, 2006.

[6] P. Fearnhead and Z. Liu. On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):589–605, 2007.

[7] A. Hannak, G. Soeller, D. Lazer, A. Mislove, and C. Wilson. Measuring price discrimination and steering on e-commerce web sites. In *Proc. IMC'14*.

[8] Huffington Post. Facebook Just Made A Pretty Awkward Change To Your Profile. http://www.huffingtonpost.com/entry/facebook-intro-work_us_57694831e4b015db1bca97c9.

[9] M. Lécuyer, G. Ducoffe, F. Lan, A. Papancea, T. Petsios, R. Spahn, A. Chaintreau, and R. Geambasu. Xray: Enhancing the web's transparency with differential correlation. In *USENIX Security'14*.

[10] New York Post. Major decline in NYPD stop-frisks. http://nypost.com/2013/02/09/major-decline-in-nypd-stop-frisks.

[11] New York Times. Judge Rejects New Yorks Stop-and-Frisk Policy. http://www.nytimes.com/2013/08/13/nyregion/stop-and-frisk-practice-violated-rights-judge-rules.html.

[12] Newsweek. Did Bill De Blasio Keep His Promise To Reform Stop-and-Frisk? http://europe.newsweek.com/did-bill-de-blasio-keep-his-promise-reform-stop-and-frisk-266310.

[13] E. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[14] A. Savitzky and M. J. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[15] R. S. Sharad Goel, Justin M. Rao. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics*, 2015.

[16] C. Simoiu, S. Corbett-Davies, and S. Goel. Testing for Racial Discrimination in Police Searches of Motor Vehicles. *SSRN abs.2811449*, 2016.

[17] The New York Times. Facebook to Change News Feed to Focus on Friends and Family. http://www.nytimes.com/2016/06/30/technology/facebook-to-change-news-feed-to-focus-on-friends-and-family.html?_r=0, June 2016.

[18] Time. Here's What Facebook's Big New Change Really Means. http://time.com/4387908/facebook-change-news-feed-update/, June 2016.

[19] C. S. Times. Chicago police and aclu agree to major changes in stop-and-frisk policy. http://chicago.suntimes.com/politics/chicago-police-and-aclu-agree-to-major-changes-in-stop-and-frisk-policy/, August 2015.

[20] X. Xiang and K. Murphy. Modeling changing dependency structure in multivariate time series. In *Proc. ICML'07*.