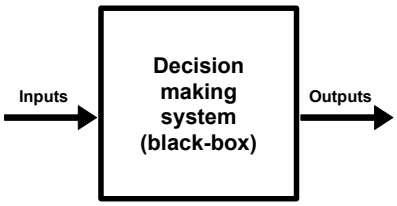
The Case for Temporal Transparency: Detecting Policy Change Events in Black-Box Decision Making Systems

Miguel Alves Ferreira, Muhammad Bilal Zafar and Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS)

1. DMS design and implementation



Examples: Police stops, load approval, rec. systems

Need to Bring transparency to DMSs

2. Opaqueness of DMSs

- DMS algorithm proprietary.
- Idea! Reverse-engineer
  - Monitor different input/output pairs
- Problems!
  - Providing new inputs not possible (police stops)
  - Existing inputs not available (privacy)
  - Not easy to scale

Hard to understand blackbox DMSs

This work: Achieve transparency without having to reverse-engineer the DMS

3. New notion: Temporal transparency

Key insight: We might not know the current policy (or DMS algorithm), but we can detect changes in it!

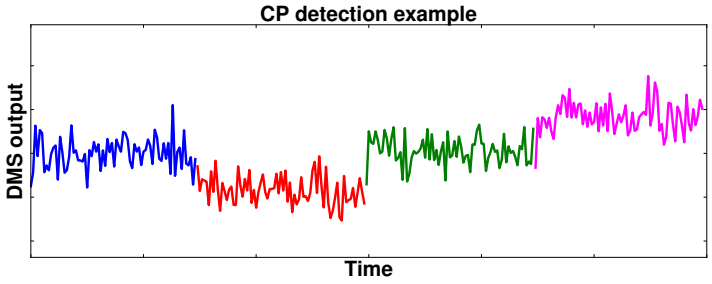
- Change in DMS output over time ⇔ change in inputs *OR* change in decision making policy
- Temporal transparency: Focus on detecting changes in DMS output over time
- Can be leveraged for targeted implementation of reverse-engineering techniques

4. *Tetra:* Methodology for detecting policy change events

- Model DMS output as time-series data consisting of different probability distributions (policy regimes)
- Changepoints (CPs) in time-series data correspond to boundaries of these policy regimes
- Leveraged Bayesian CP detection frameworks

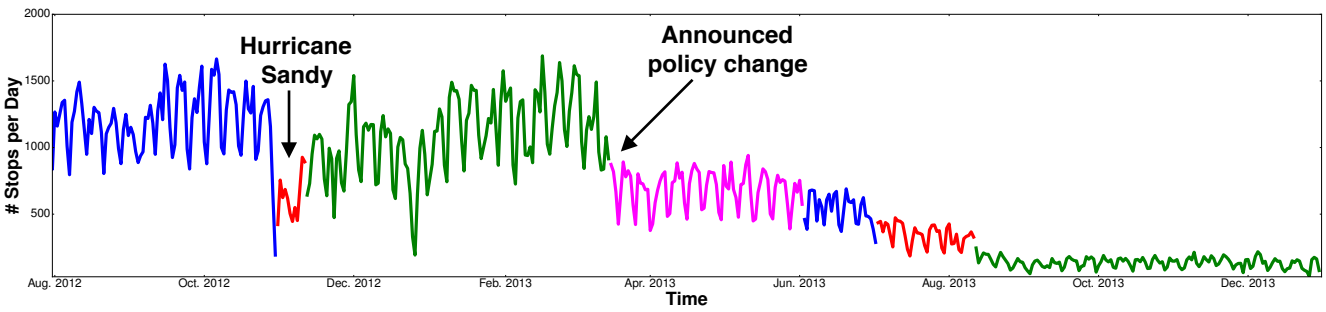
***Tetra* framework:**

- Solve MAP optimization problem to recover underlying distributions
- Optimal solution yields the location of CPs
- Framework can be fine-tuned to adjust the *significance* of the detected CPs



Different colors represent different distributions.
Find the (optimal) parameters of these distributions.

5. Application of *Tetra* on NYC SQF data



Three kind of changes:

1. Seasonal patterns (not shown)
2. Unusual input changes (Hurricane Sandy)
3. (Un)Announced policy changes

Going forward:

- Multivariate feature analysis
- Online CP detection
- Other application domains (ad recommendation)