Real-Time Aggregate Monitoring with Differential Privacy

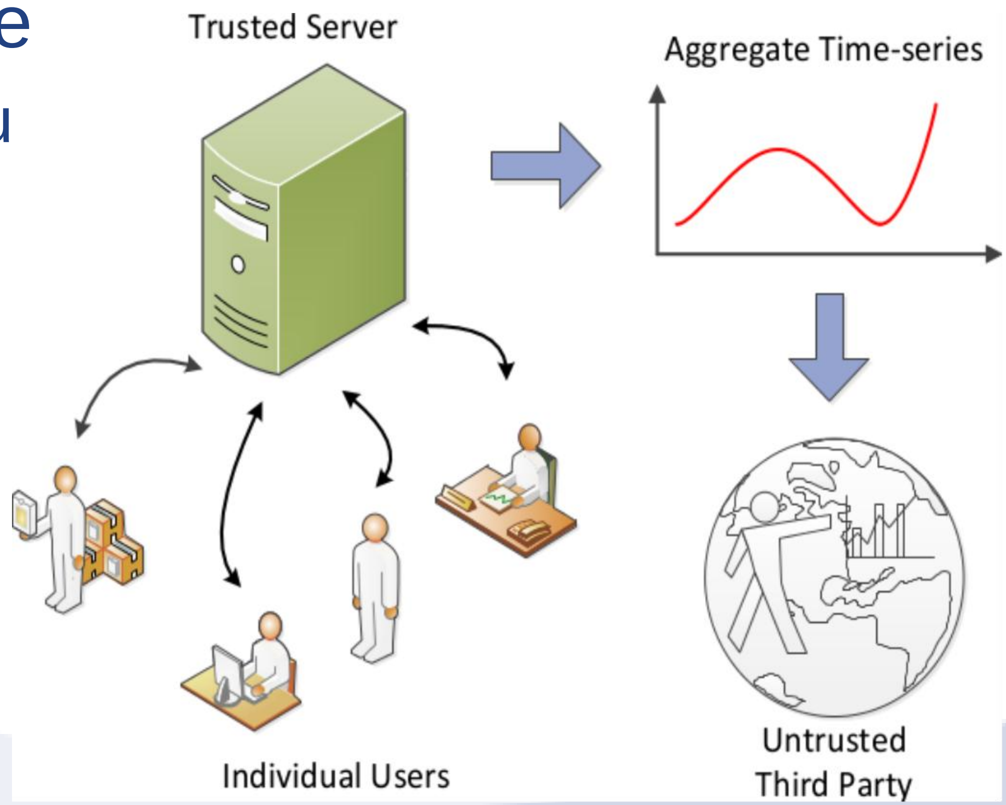
Li Xiong
Department of Mathematics and Computer Science
Department of Biomedical Informatics
Emory University

(Joint work with Liyue Fan, Vaidy Sunderam)

EMORY UNIVERSITY

Scenario

- Disease Surveillance
  - E.g. daily count of flu cases in different regions
- Traffic Monitoring
  - E.g. hourly count of vehicles at different intersections
- Single time-series
- Multi-dimensional time-series



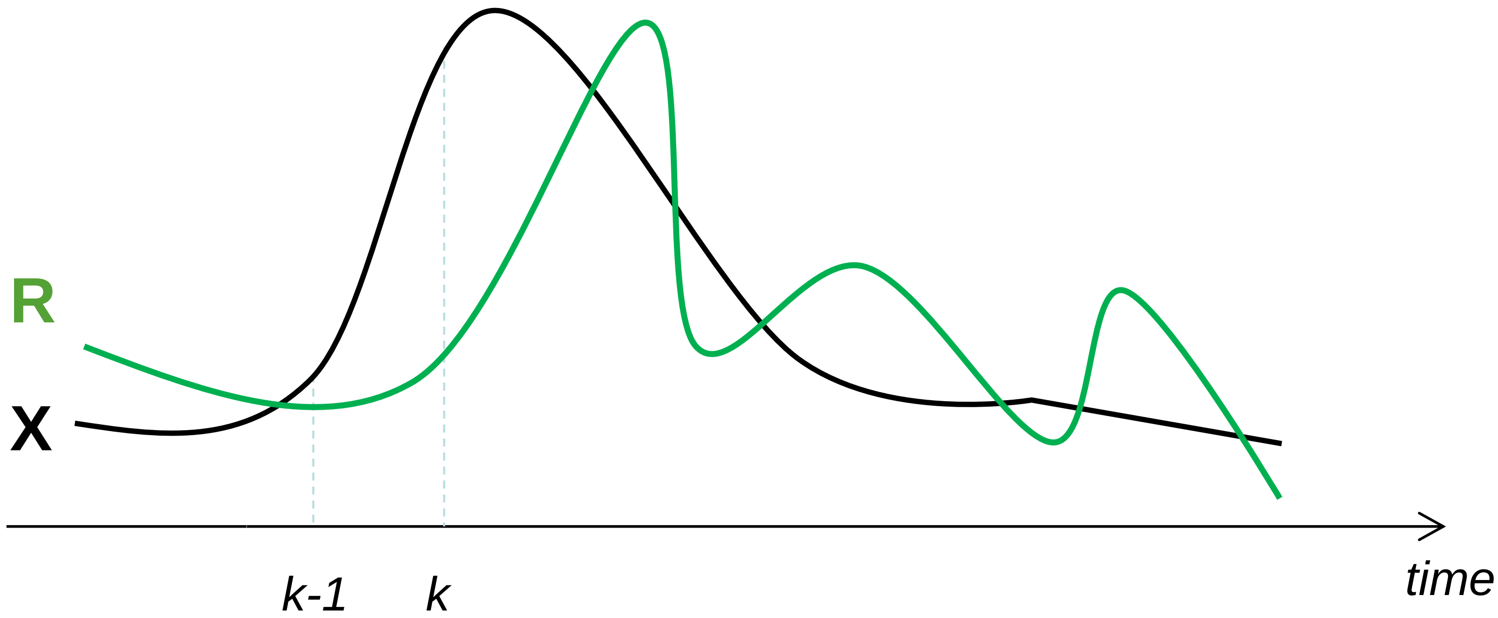
Single Time-Series: Problem Statement

- A univariate, discrete *Time-Series* $\mathbf{X} = \{x_k\}$ is a set of values of variable x observed at discrete time stamp k, where $0 \leq k < T$ and T is the lifetime of the series.

- Given time series $\mathbf{X}$ and differential privacy budget α, release α-differentially private series $\mathbf{R}$ with high utility.

EMORY
UNIVERSITY

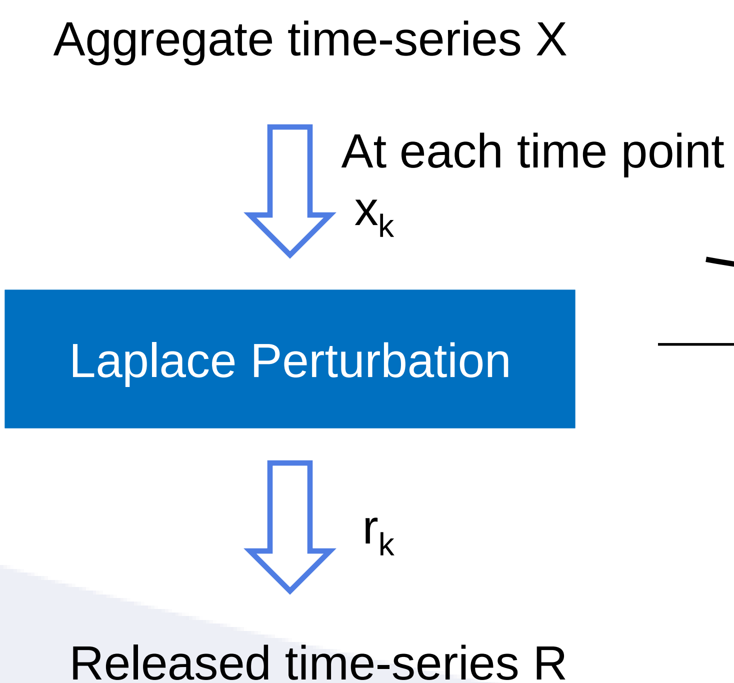
Utility

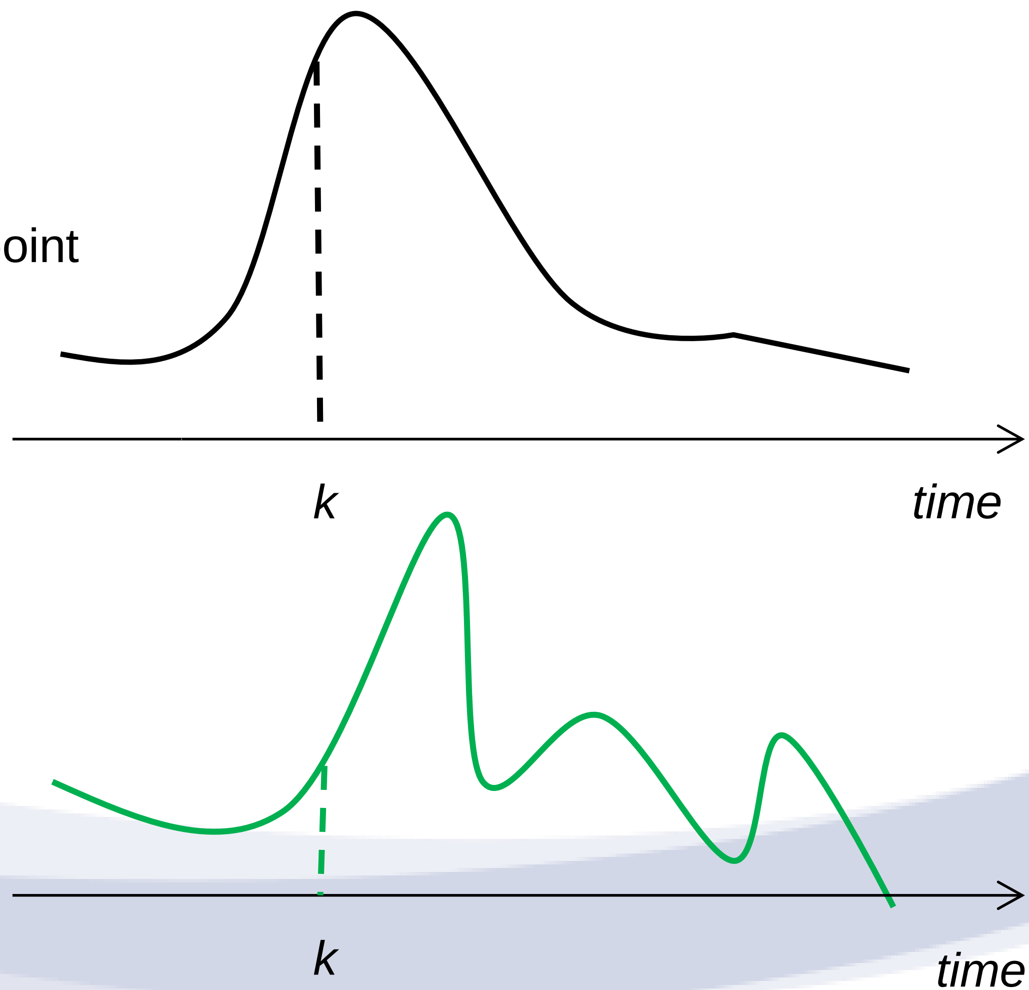


- Point-wise: average relative error
- Time-series: outbreak detection
  - Outbreak at time k: $x_k - x_{k-1} >=$ threshold
  - Specificity and sensitivity
  - Precision and recall: F1 metric

EMORY UNIVERSITY

Baseline: Laplace Perturbation Algorithm (LPA)





- High sensitivity : O(T)
- High perturbation error: O(T)

State-of-the-art: Discrete Fourier Transform [RN10]

Aggregate series **X**

Discrete Fourier Transform

Laplace Perturbation

Inverse DFT

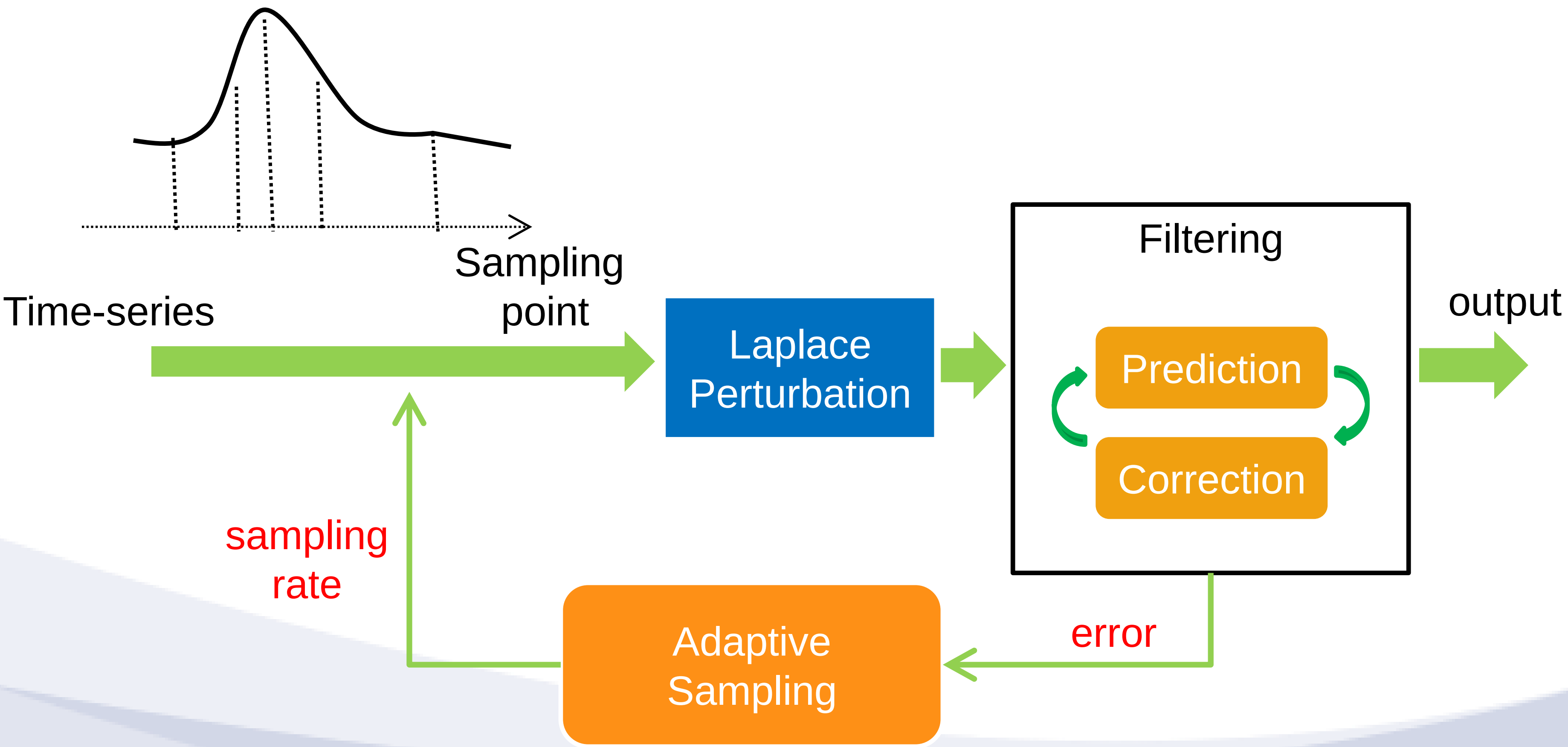
Released series **R**

Retain only the first l coefficients to reduce sensitivity

- Higher accuracy
- Offline or batch processing only

EMORY UNIVERSITY

FAST: Filtering and Adaptive Sampling for aggregate Time-series monitoring



- Filtering – posterior estimate based on prediction and perturbed values
- Adaptive sampling - reduce sensitivity

Filtering: State-Space Model

- Process Model

$$x_{k+1} = x_k + \omega$$
$$\omega \sim \mathbb{N}(0, Q)$$

Process noise

- Measurement Model

$$z_k = x_k + \nu$$
$$\nu \sim Lap(\lambda)$$

Measurement noise

- Given noisy measurement z_k, how to estimate true state x_k ?

EMORY UNIVERSITY

Filtering: Posterior Estimation

- Denote $\mathbb{Z}_k = \{z_0, \dots, z_k\}$
- Posterior estimate:

$$\hat{x}_k = E(x_k|\mathbb{Z}_k)$$

- Posterior distribution:

$$f(x_k|\mathbb{Z}_k) = \frac{f(x_k|\mathbb{Z}_{k-1})f(z_k|x_k)}{f(z_k|\mathbb{Z}_{k-1})}$$

- **Challenge**:

$f(z_k|\mathbb{Z}_{k-1})$ and $f(x_k|\mathbb{Z}_{k-1})$ are difficult to carry out when f_v is not Gaussian

EMORY UNIVERSITY

Filtering: Solutions

- Option **1**: Approximate measurement noise with Gaussian

$$\nu \sim \mathbb{N}(0, R)$$

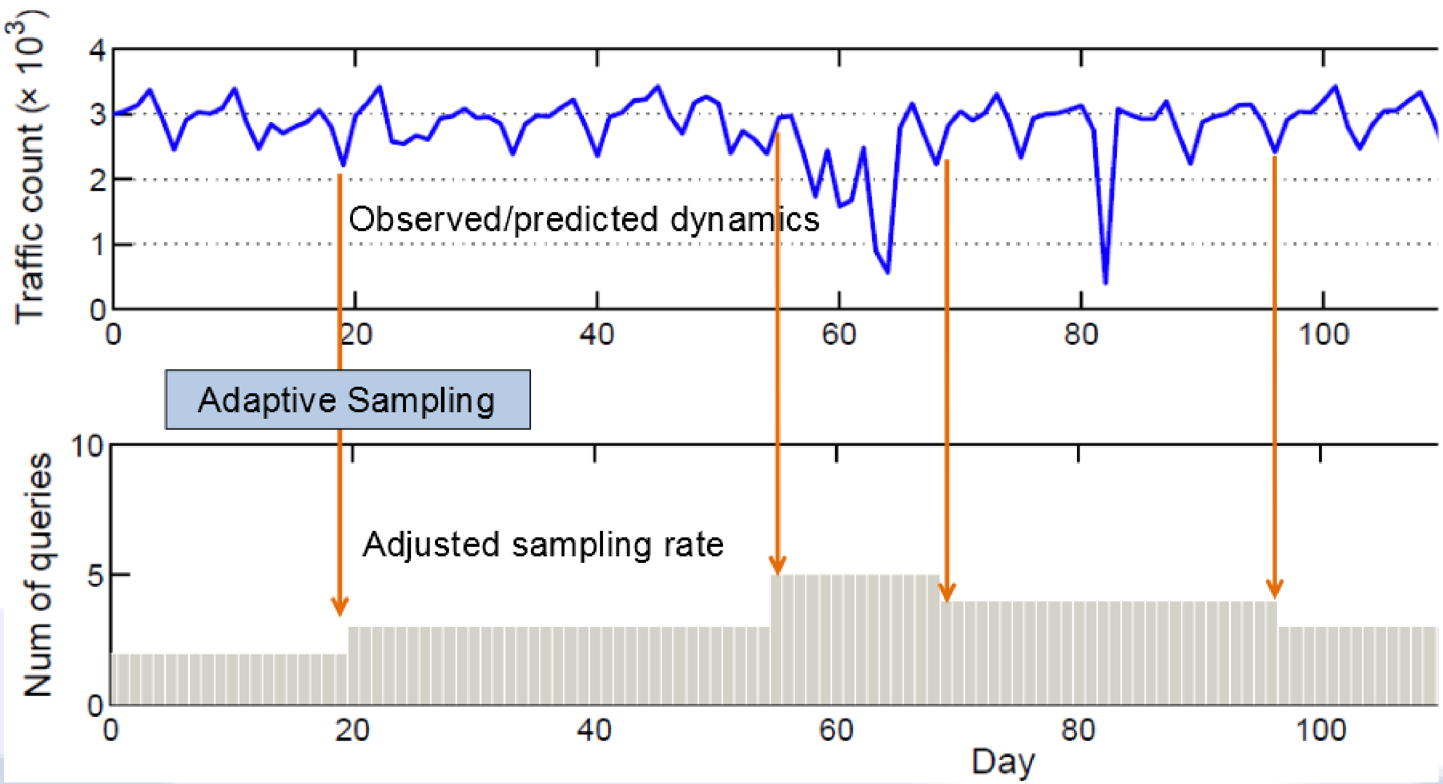
→ the Kalman filter

- Option **2**: Estimate posterior density by Monte-Carlo method

$$f(x_k|\mathbb{Z}_k) = \sum_{i=1}^{N} \pi_k^i \delta(x_k - x_k^i)$$

where $\{x_k^i, \pi_k^i\}_1^N$ is a set of weighted samples/particles.
→ particle filters

Adaptive Sampling



- Fixed sampling – difficult to select sampling rate a priori
- Adaptive sampling - adjust sampling rate based on feedback from observed data dynamics

Adaptive Sampling: PID Control

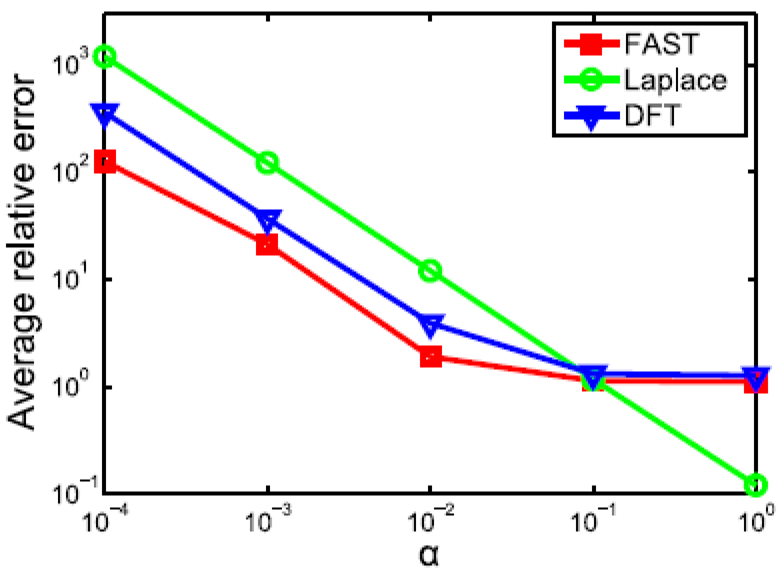
- PID error (Δ): compound of *proportional*, *integral*, and *derivative* errors
- Measures how well the constant data model describe the current trend
- Determines a new sampling interval:

$$I' = I + \theta(1 - e^{\frac{\Delta - \xi}{\xi}})$$

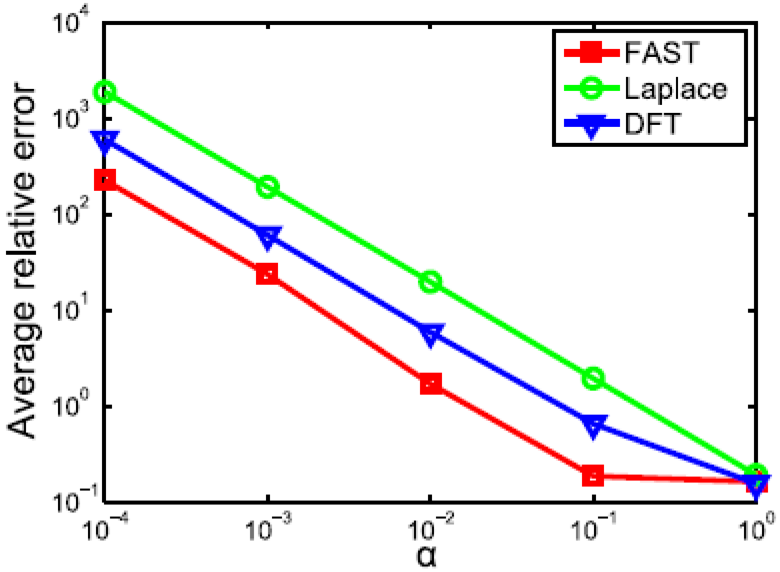
where θ represents the magnitude of change and ξ is the set point for sampling process.

EMORY UNIVERSITY

Some results: average relative error



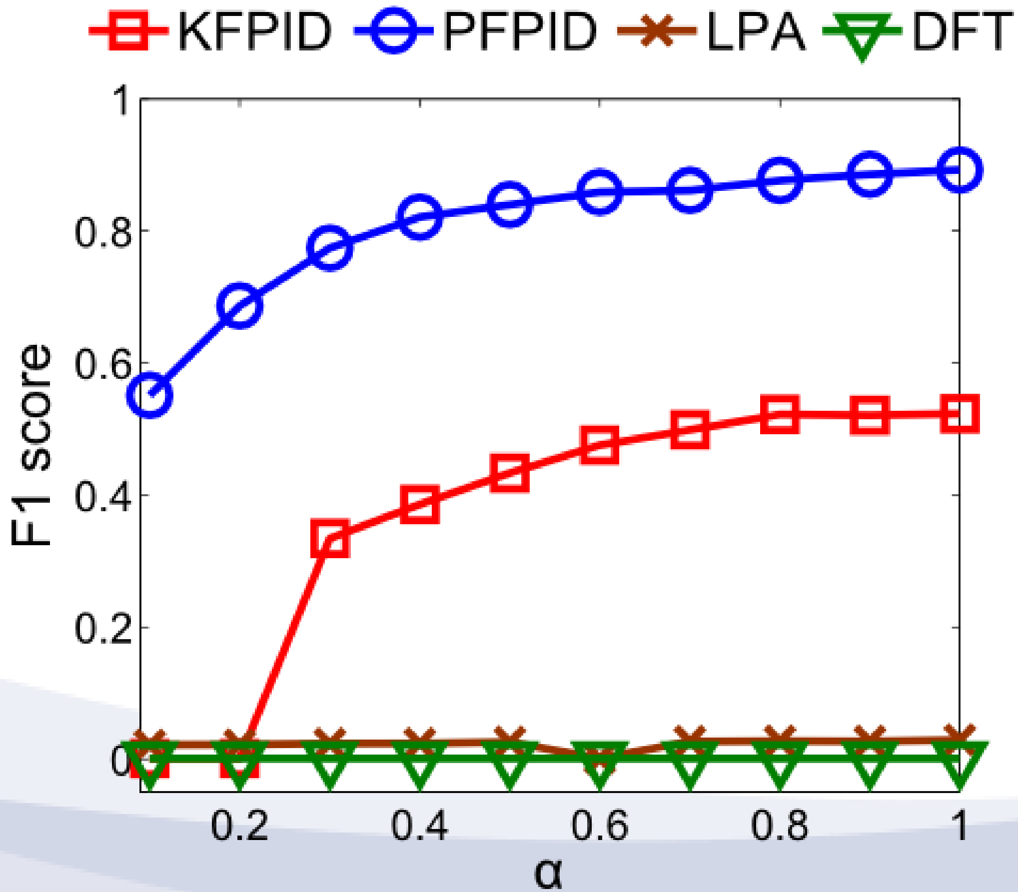
(a) Flu data set



(b) Traffic data set

- Flu dataset (CDC): weekly outpatient count of age group [5-24] from 2006 – 2010 (209 data points)
- Traffic dataset (U Washington): daily traffic count for Seattle-area highway at I-5 143.62 southbound from 2003 – 2004 (504 data points)

Some Results: F1 metric for outbreak detection



traffic data set

EMORY
UNIVERSITY

Multi-dimensional time-series: Problem statement

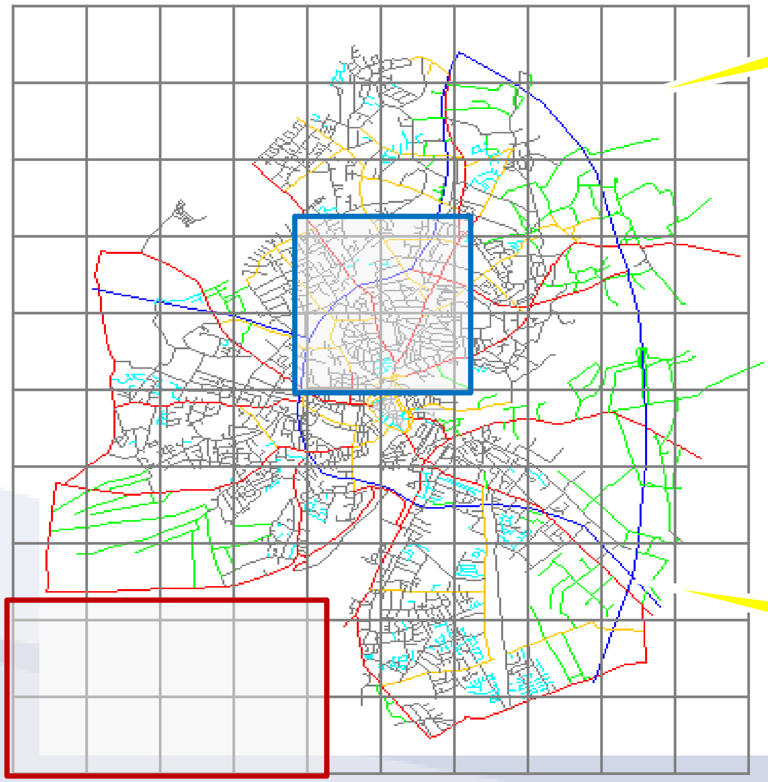
- Real-time traffic counts over a big area
- Raw aggregates obtained on a fine-grained grid
- For each cell c

$$\mathbf{X}^c = \{x_k^c\}$$

- Goal: release $\mathbf{R}^c$ for each c

EMORY UNIVERSITY

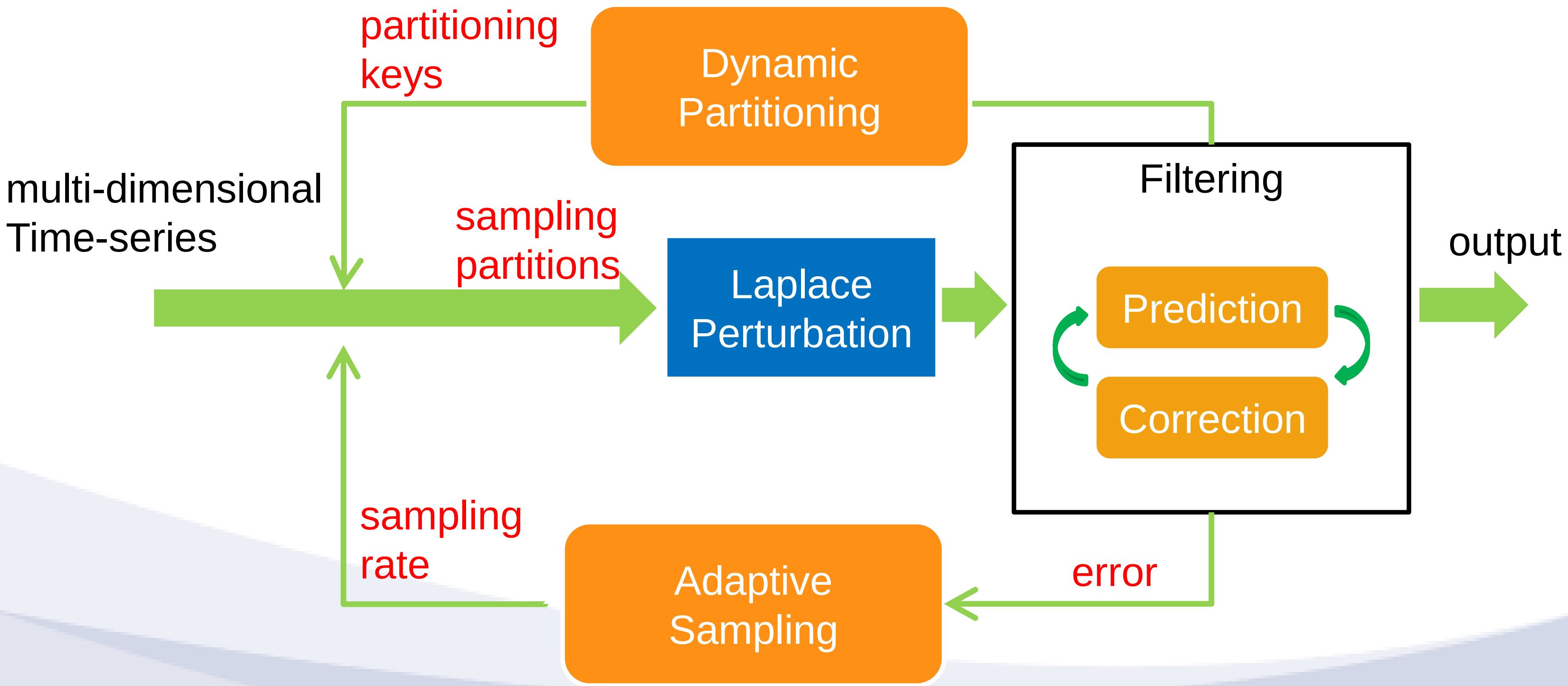
Multi-dimensional time-series: challenges and solutions



- Data has sparse and uniform regions
- Data is dynamically changing

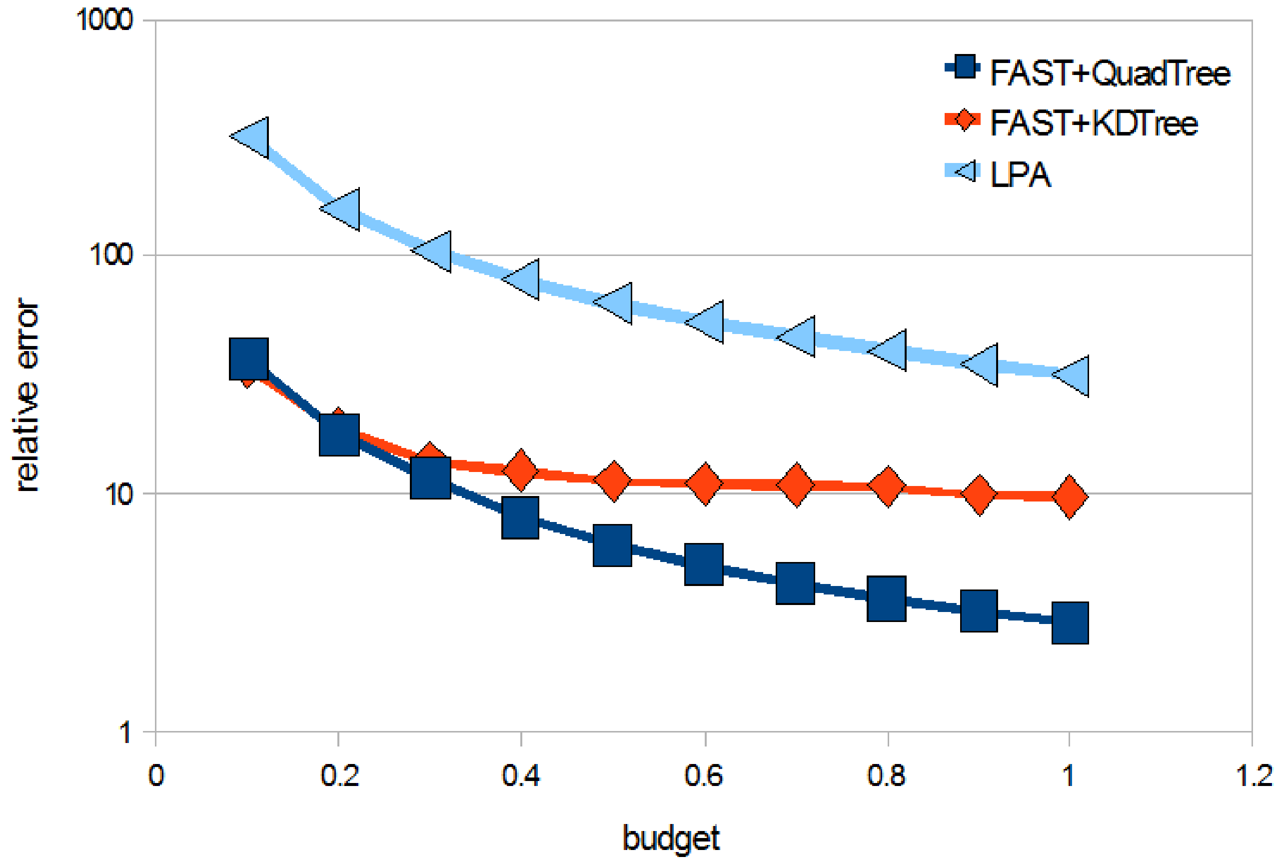
EMORY UNIVERSITY

FAST with Partitioning



- Dynamic spatial partitioning: based on KD-Tree and quad-tree

Some Results



Synthetic traffic data by Brinkhoff moving objects generator

EMORY UNIVERSITY

Closing Remarks

- Utility: time point wise relative error still high but can be useful for time-series driven applications such as outbreak detection
- Key insight: feedback loops are useful to dynamically adjust sampling, aggregation, and estimation
- Open question: how to allocate budget over time points?

EMORY UNIVERSITY

Thank you

- References
  - Liyue Fan, Li Xiong. Real-Time Aggregate Monitoring with Differential Privacy, CIKM 2012
  - Liyue Fan, Li Xiong. Adaptively sharing time-series with differential privacy, arXiv:1202.3461, 2012
- Research Support Acknowledgement
  - AFOSR: PREDICT: Privacy and Security Enhancing Dynamic Information Collection and Monitoring
  - NSF: Adaptive Differentially Private Data Release
- Contact
  - AIMS: Assured Information Management and Sharing
    http://www.mathcs.emory.edu/aims
  - lxiong@emory.edu

EMORY UNIVERSITY