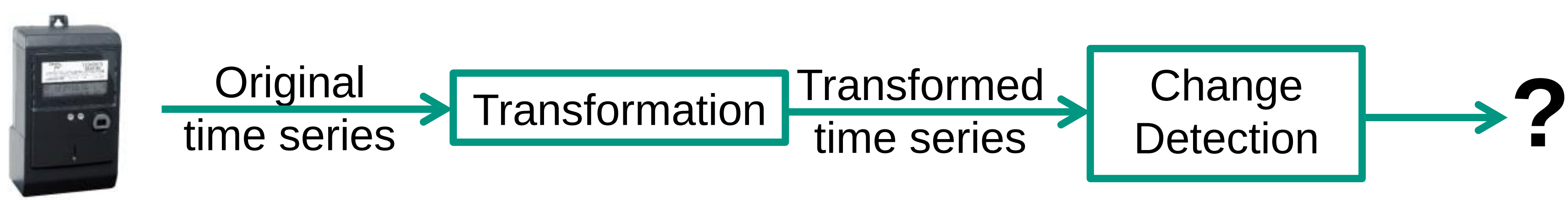
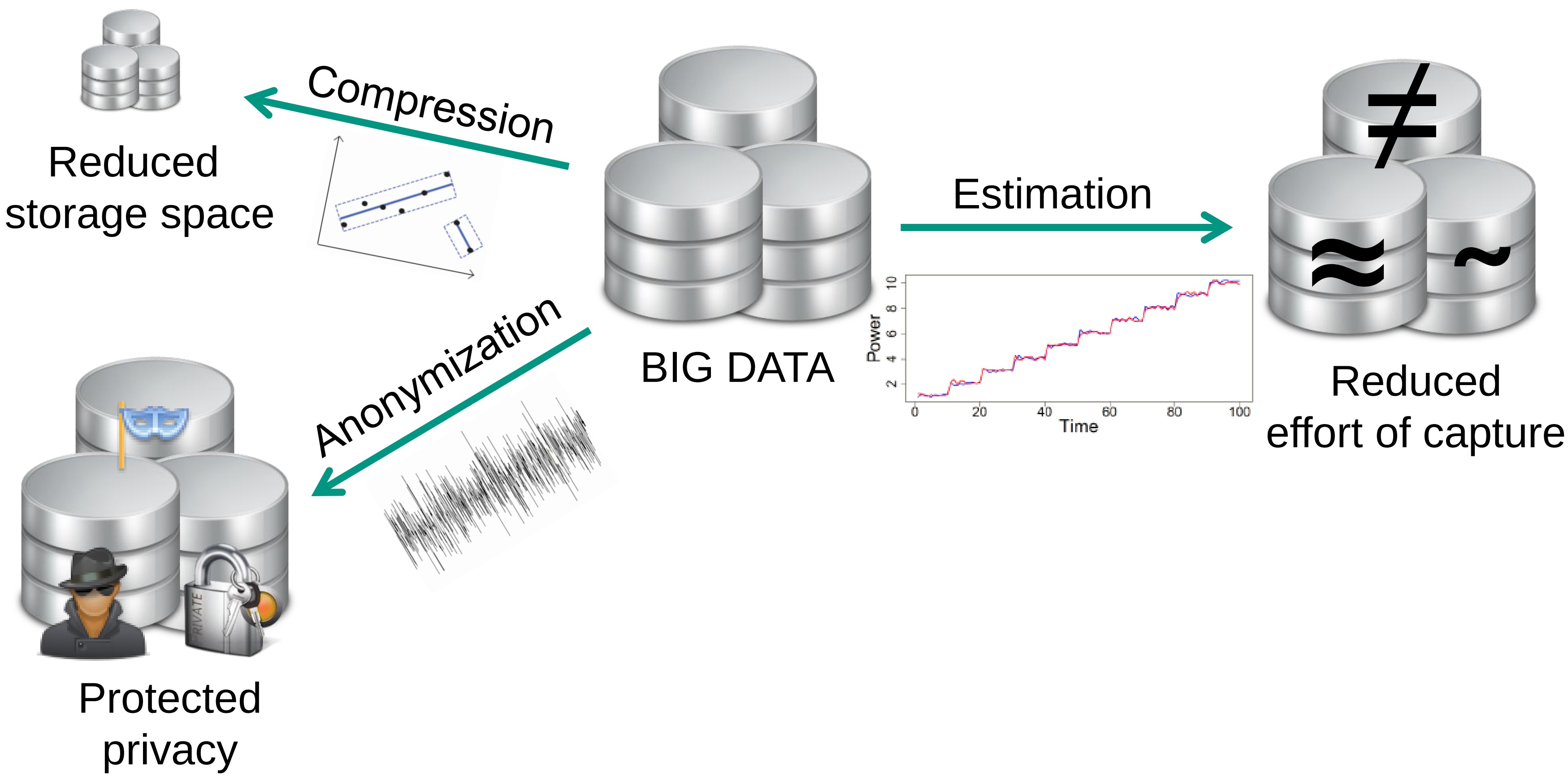
How to Quantify the Impact of Lossy Transformations on Change Detection

Pavel Efros, Erik Buchmann, Adrian Englhardt, Klemens Böhm | June 30

27th International Conference on Scientific and Statistical Database Management (SSDBM) 2015



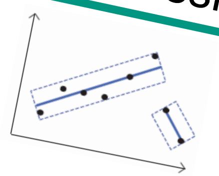
Lossy Transformation



Lossy Transformation

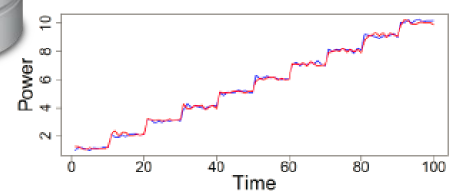
Reduced storage space

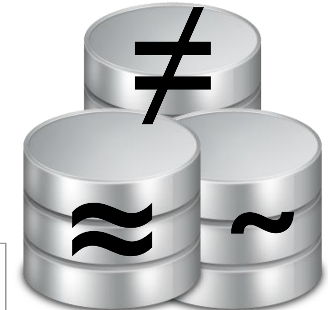
Compression



BIG DATA

Estimation





Reduced effort of capture

Anonymization

Protected privacy

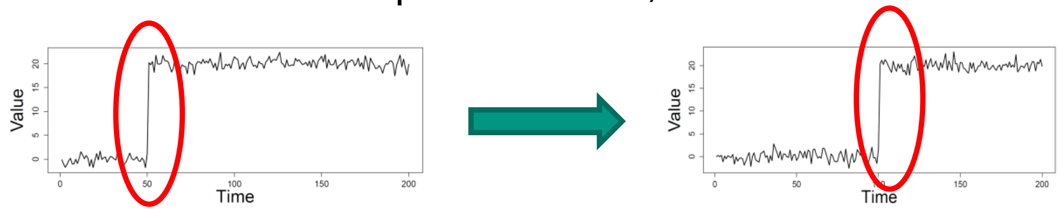
How do lossy transformations affect subsequent data analysis?

- Example: Change Detection

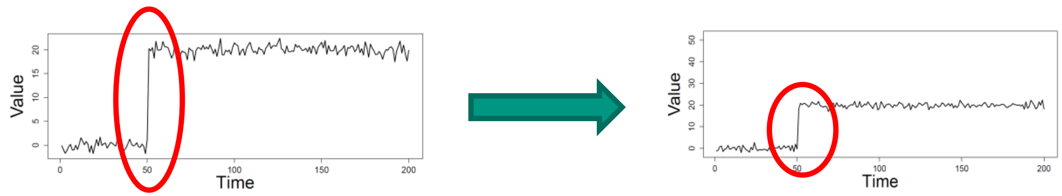
Original time series → Transformation → Transformed time series → Change Detection → ?

Challenges

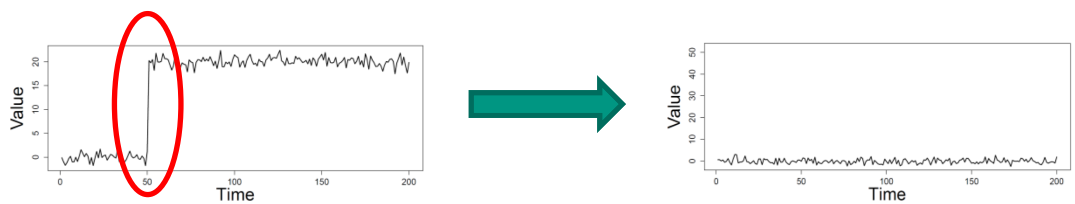
- Impact of transformations on changes is manifold
  - Existing changes may
    - be detected at different points in time,



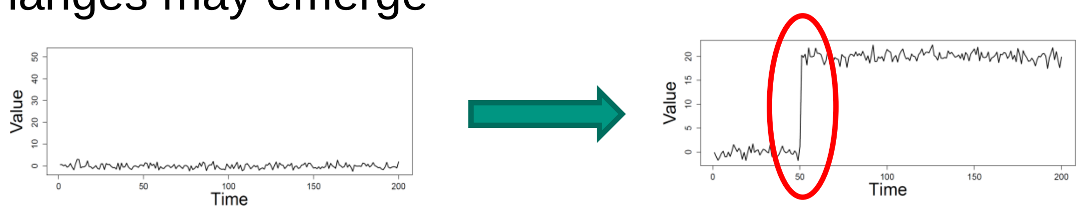
    - have their significance altered,



    - go undetected



  - New changes may emerge



Requirements

- Customization to subsequent application is necessary:

Producers

Consumers

Requirements

- Customization to subsequent application is necessary:

Producers

Consumers

- Disappearance of changes is crucial
  - => should have higher weight than a newly emerged one

Agenda

- MILTON: A **M**easure for quantifying the **I**mpact of **L**ossy **T**ransformations on subsequent change detecti**ON**
  - Sets
  - Definition
- Evaluation
  - Compression Scenario
  - Anonymization Scenario
- Conclusions & Outlook

MILTON sets of changes

- PC: pairs of corresponding changes
- MISS: changes which disappear due to Transformation
- FP: changes which appear due to Transformation

Original

Time	Score
15	0.7
25	0.5
30	0.4
40	0.9
50	0.8

$\Delta_t = 2, \Delta_s = -0.1$

$s = 0.4$

$s = 0.7$

Transformed

Time	Score
17	0.6
26	0.6
38	0.8
45	0.7
52	0.7

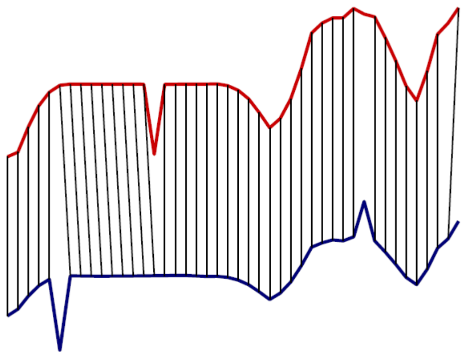
Calculating PC

- Goal
  - Find best-matching subsequences $O' \in$ *Original* and $T' \in$ *Transformed*
- Requirements
  - consider all distances between matched changes
  - allow leaving unmatched changes (MISS and FP)

- Define distance between changes x and y

$$dist(x, y) = g(f_{TIME}(\Delta_t(x, y)), f_{SCORE}(\Delta_s(x, y)))$$

- Solution: Optimal Subsequence Bijection
  - may skip changes (adds penalty instead)
  - minimizes the sum of distances



Calculating MISS & FP

- MISS
  - Loop over changes in $Original$
    - If change not in PC → add to MISS
- FP
  - Loop over changes in $Transformed$
    - If change not in PC → add to FP

$Original$

Time	Score
15	0.7
25	0.5
30	0.4
40	0.9
50	0.8

$\Delta_t = 2, \Delta_s = -0.1$

$s = 0.4$

$s = 0.7$

$Transformed$

Time	Score
17	0.6
26	0.6
38	0.8
45	0.7
52	0.7

MILTON Definition

Karlsruhe Institute of Technology

PC

$$errTIME = \sum_{(x,y)\in PC} f_{TIME}(\Delta_t(x,y)) \qquad errSCORE = \sum_{(x,y)\in PC} f_{SCORE}(\Delta_s(x,y))$$

$$errPC = errTIME + errSCORE$$

MISS

$$errMISS = \sum_{(t,s)\in MISS} f_{MISS}(s)$$

FP

$$errFP = \sum_{(t,s)\in FP} f_{FP}(s)$$

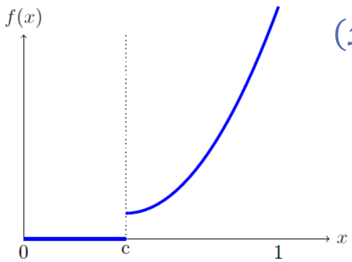
$$MILTON = \frac{errMISS + errPC + errFP}{|PC| + |MISS| + 1}$$

MILTON Definition

PC

$$errTIME = \sum_{(x,y) \in PC} f_{TIME}(\Delta_t(x,y))$$

$$errSCORE = \sum_{(x,y) \in PC} f_{SCORE}(\Delta_t(x,y))$$



$$errPC = errTIME + errSCORE$$

MISS

$$errMISS = \sum_{(t,s) \in MISS} f_{MISS}(s)$$

FP

$$errFP = \sum_{(t,s) \in FP} f_{FP}(s)$$

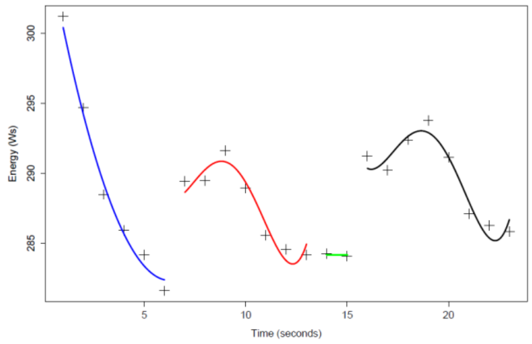
$$MILTON = \frac{errMISS + errPC + errFP}{|PC| + |MISS| + 1}$$

Evaluation: Compression Scenario

- Setup
  - Datasets
    - Reference Energy Disaggregation Dataset (REDD)
    - Smart Home Dataset (Smart)
  - Lossy Compression Methods
    - Adaptive Piecewise Constant Approximation (APCA)
    - Piecewise Linear Histogram (PWLH)
    - Adaptive Polynomial Piecewise Compression (APP)
  - Change Detection: CUSUM

Common parameter ε

Norm : $L_\infty = ||x - x'||_\infty = \max_{i=1,\dots,n} |x_i - x_i'|$

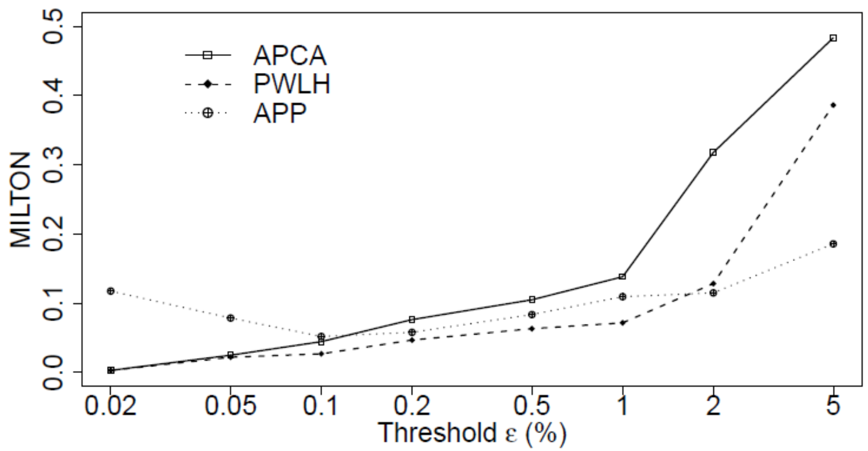


- MILTON Parameters

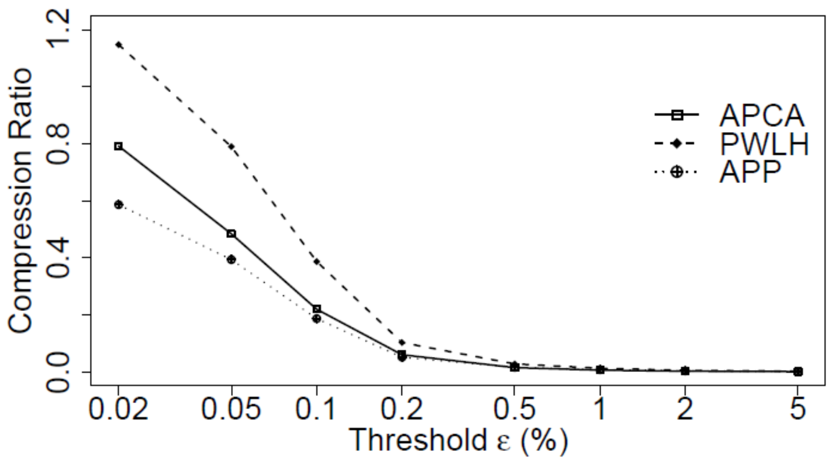
$$f_{TIME}(\Delta_t(x,y)) = |\Delta_t|$$
$$f_{SCORE}(\Delta_s(x,y)) = 0$$
$$f_{MISS}(s) = s^2 + 1$$
$$f_{FP}(s) = s$$

Results: Compression Scenario

Smart dataset



MILTON vs. Threshold ε



Compression Ratio vs. Threshold ε

Evaluation: Anonymization Scenario

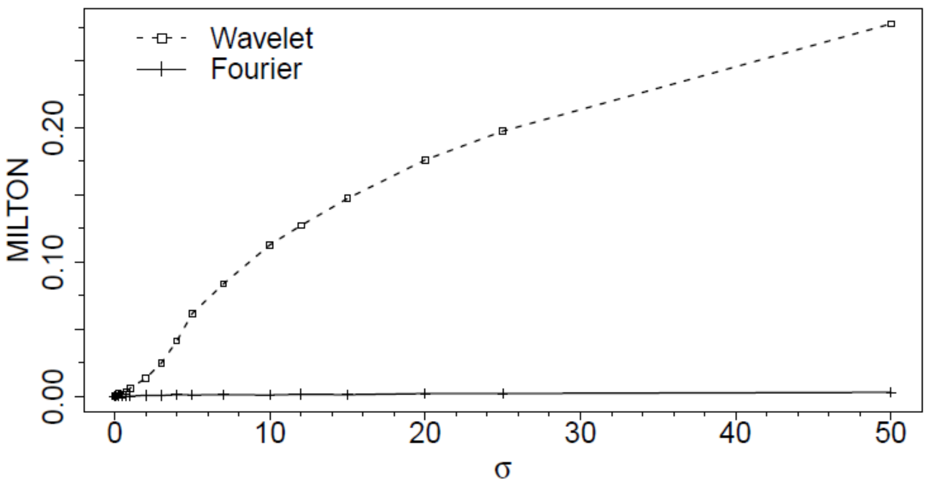
- Setup
  - Datasets
    - Reference Energy Disaggregation Dataset (REDD)
    - Smart Home Dataset (Smart)
  - Perturbation Methods
    - Fourier-based
    - Wavelet-based
  - Change Detection: CUSUM
- MILTON Parameters

Common parameter σ
=
amount of perturbation added

$$f_{TIME}(\Delta_t(x,y)) = \frac{1}{2} \cdot |\Delta_t|$$
$$f_{SCORE}(\Delta_s(x,y)) = \frac{1}{2} \cdot |\Delta_s|$$
$$f_{MISS}(s) = s$$
$$f_{FP}(s) = s$$

Results: Anonymization Scenario

REDD dataset



MILTON vs. Perturbation σ

	σ	PC	$MISS$	FP	$errPC$	$errMISS$	$errFP$
Fourier	1	160	0	0	0.001	0.000	0.000
	2	160	0	0	0.002	0.000	0.000
	5	156	4	4	0.004	0.394	0.394
	10	159	1	1	0.011	0.209	0.209
	25	158	2	3	0.033	0.148	0.168
	50	157	3	5	0.071	0.213	0.257
Wavelet	1	153	7	9	0.025	1.109	1.164
	2	157	3	6	0.065	0.370	0.434
	5	149	11	120	0.044	1.965	4.931
	10	152	8	336	0.048	1.171	11.606
	25	148	12	469	0.038	2.158	26.093
	50	152	8	498	0.047	0.984	38.317

MILTON Components

Conclusions & Outlook

- MILTON
  - is general and flexible measure,
  - identifies an appropriate lossy transformation method.

- Future Work
  - evaluation with broad range of CD methods,
  - investigation of parameter initialization for MILTON.