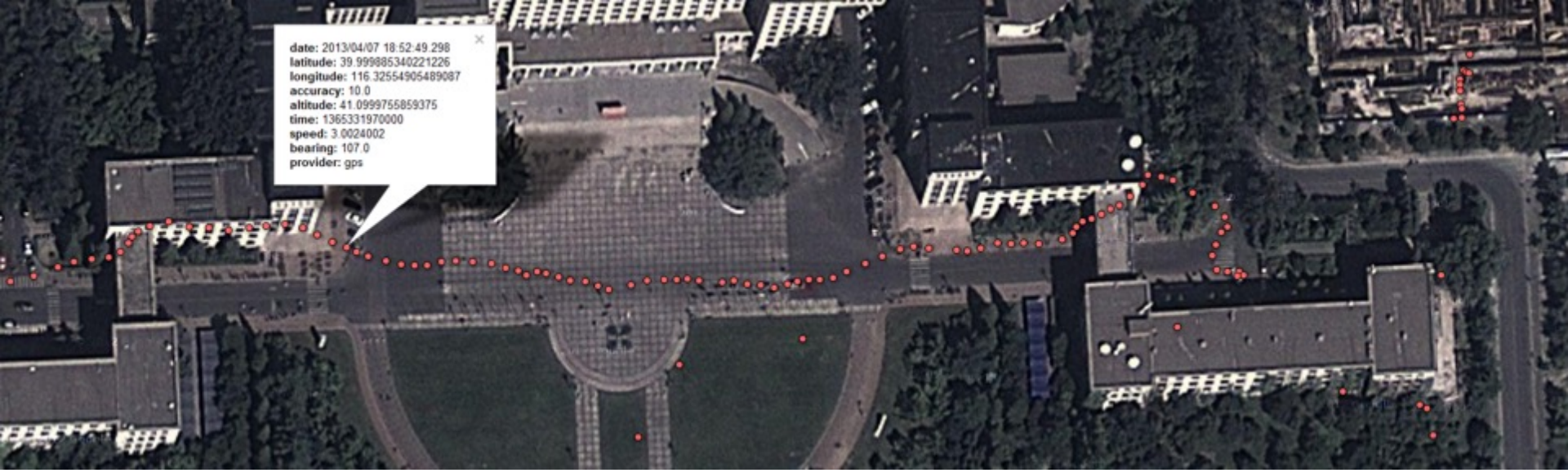


Time Series Data Cleaning

Shaoxu Song

清華大學 Tsinghua University

http://ise.thss.tsinghua.edu.cn/sxsong/

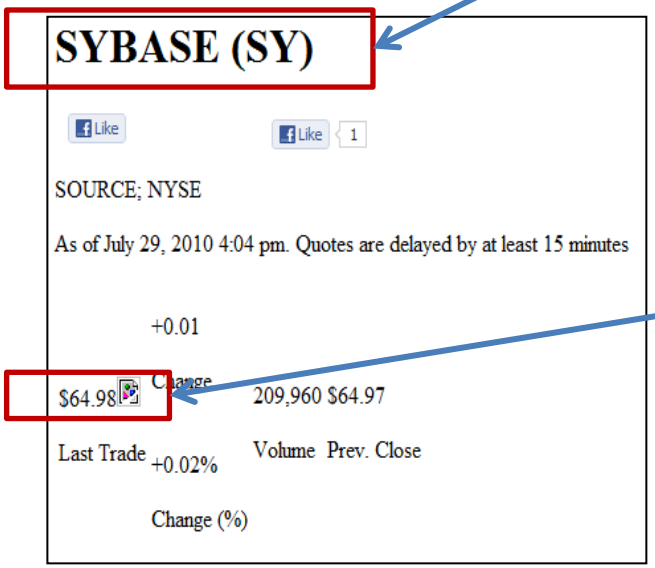
Dirty Time Series Data

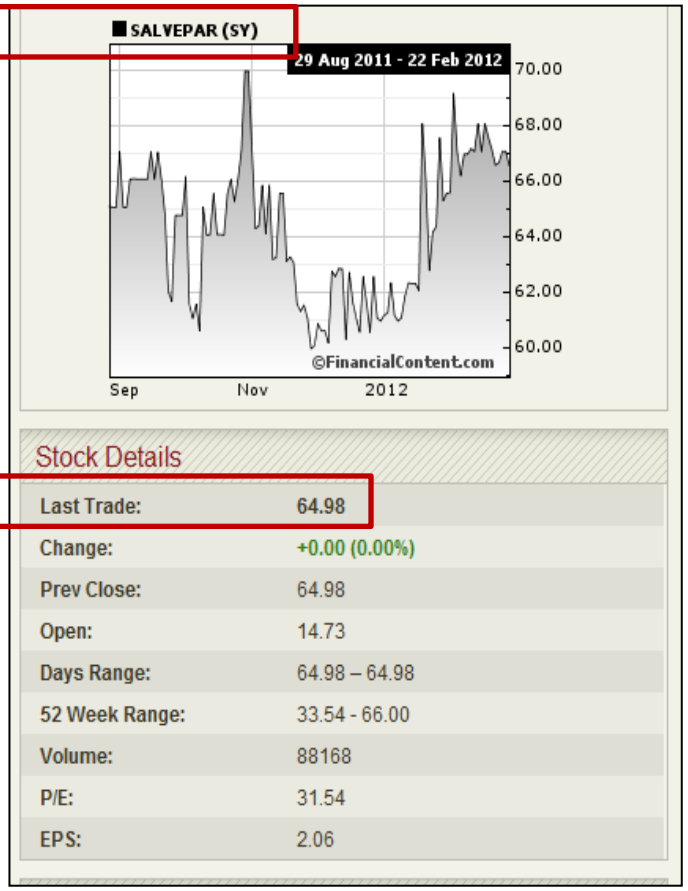
- Unreliable Readings
  - Sensor monitoring
  - GPS trajectory

J. Freire, A. Bessa, F. Chirigati, H. T. Vo, K. Zhao: Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips. IEEE Data Eng. Bull.39(2): 63-77 (2016)

Dirty Time Series Data

- Misuse



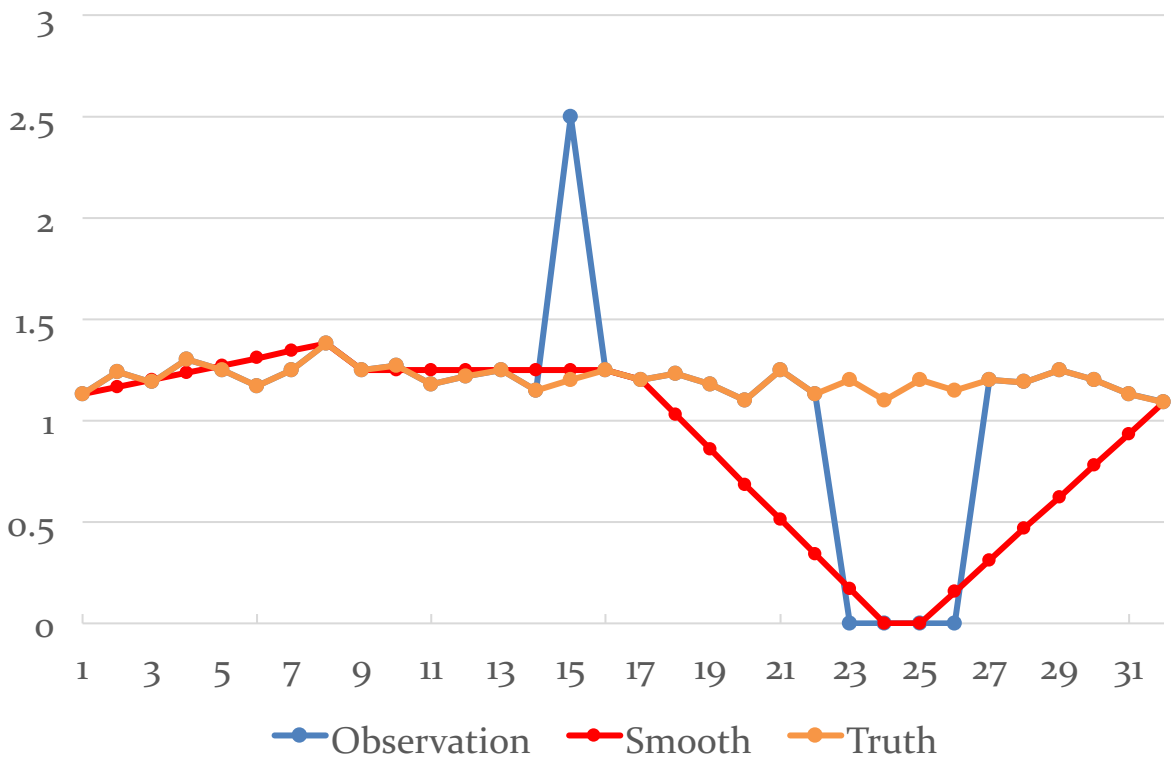


- Flight: Accuracy of Travelocity is 0.95
- Stock: Accuracy of Stock in Yahoo! Finance is 0.93

Xian Li, Xin Luna Dong, Kenneth B. Lyons, Weiyi Meng, Divesh Srivastava: Truth Finding on the deep web: Is the problem solved? PVLDB, 6(2) (2013)

Existing cleaning methods

- Smoothing Filter
  - Moving Average
  - WMA
  - EWMA



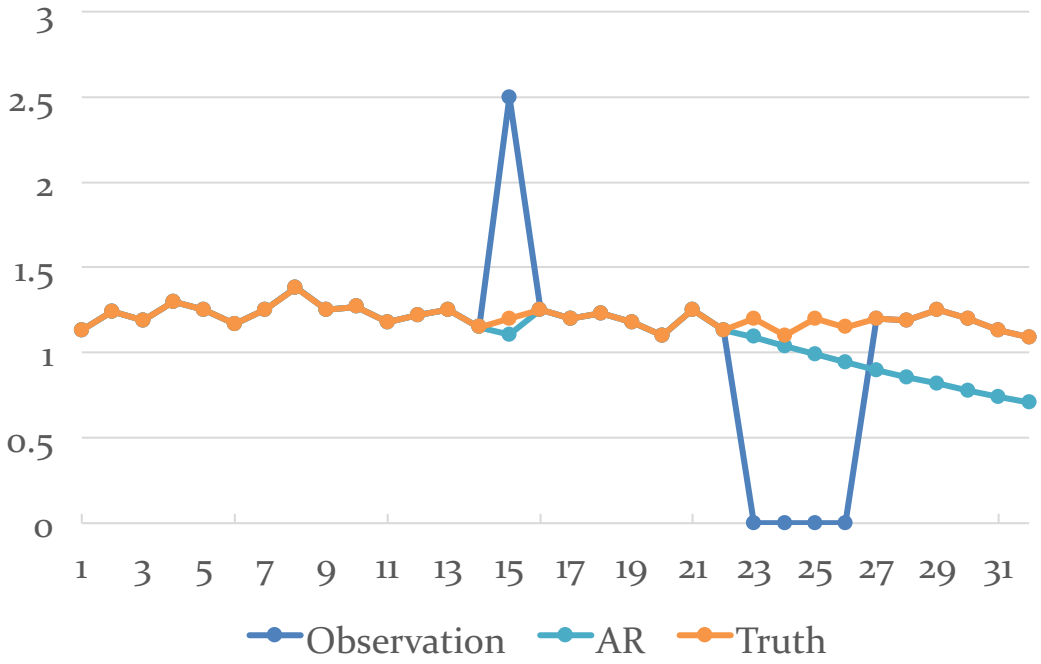
- Problem: modify almost all the data values

E. S. Gardner Jr. Exponential smoothing: The state of the art{part ii. International Journal of Forecasting, 22(4):637-666, 2006.

Existing cleaning methods

- Prediction Model
  - Modify the observation by predication
    if the predication is far distant from the observation
  - autoregressive (AR) model
  - AR(I)MA

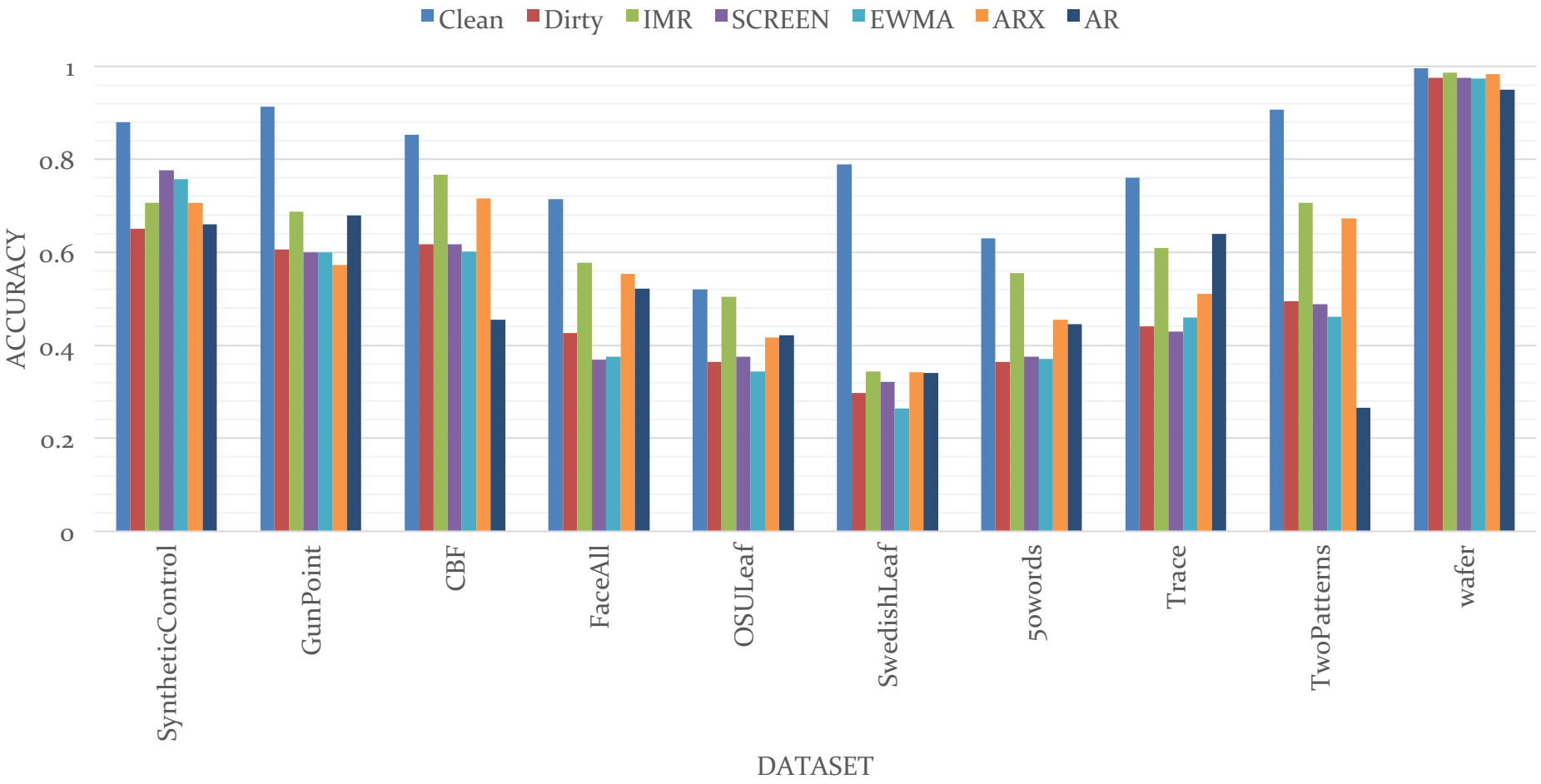
- May over change the data
  - Owing to “far distant”



Yamanishi, Kenji, and Jun-ichi Takeuchi. "A unifying framework for detecting outliers and change points from non-stationary time series data." In SIGKDD, pages 676-681, 2002

Repairing dirty data helps

- Time series classification



Constraint-based method (SIGMOD 2015)

large spike errors

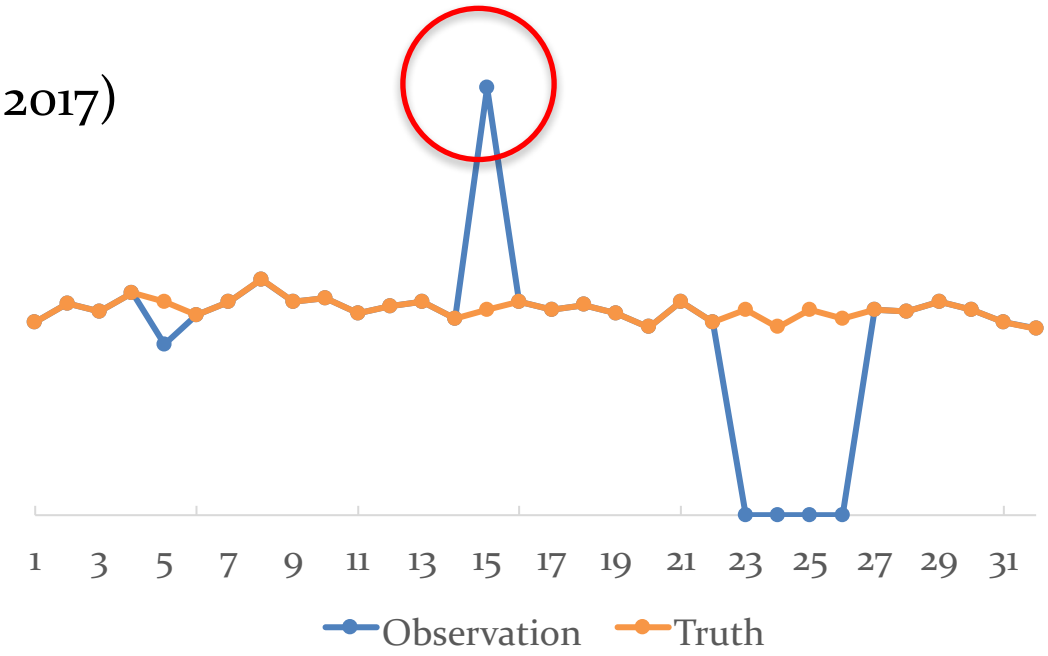
Statistical method (SIGMOD 2016)

small errors

Supervised method (VLDB 2017)

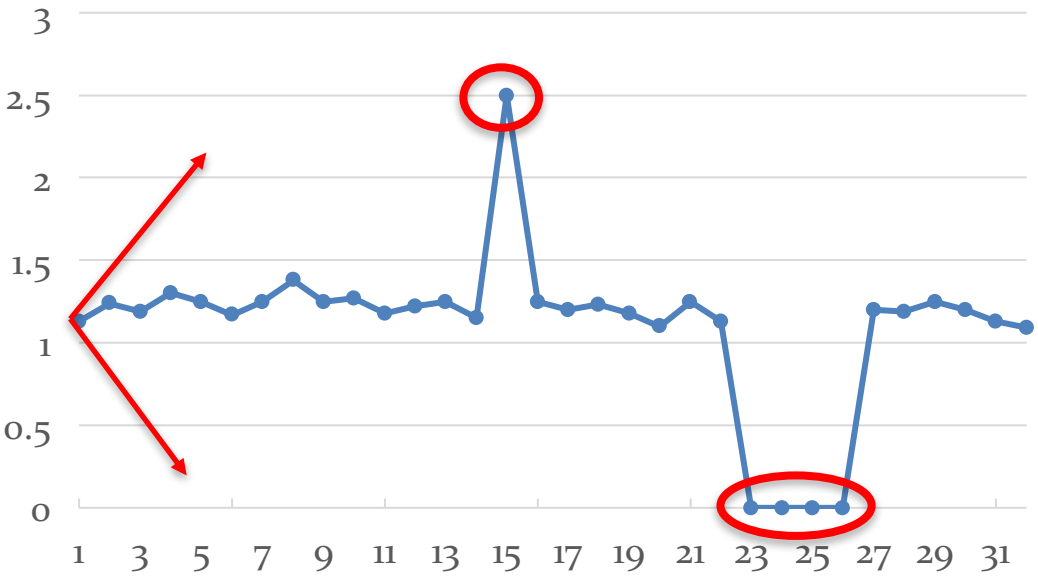
consecutive errors

Contents

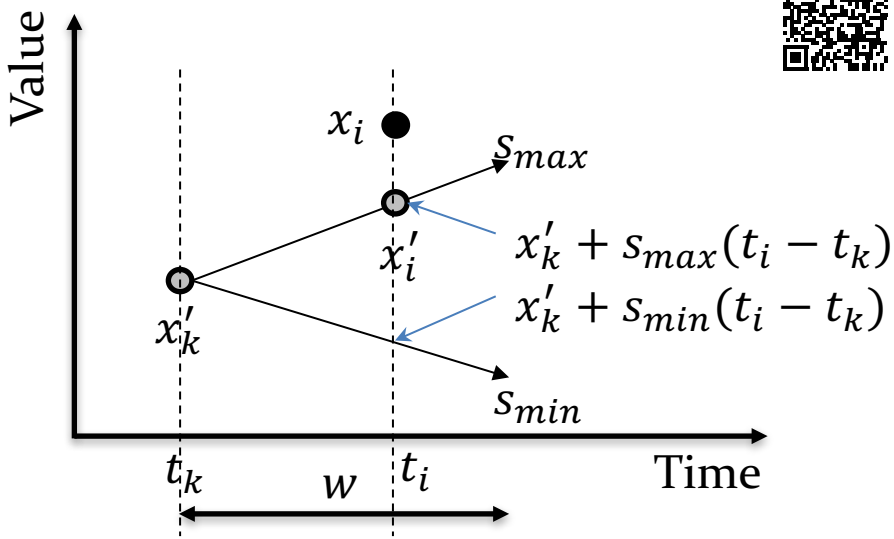


Intuition on Speed Constraints

- "Jump" of values is often constrained
  - Daily limit: in financial and commodity markets
  - Temperatures in a week
  - Fuel consumption
- Use speed constraints to identify dirty data



SCREEN Stream Data Cleaning under Speed Constraints

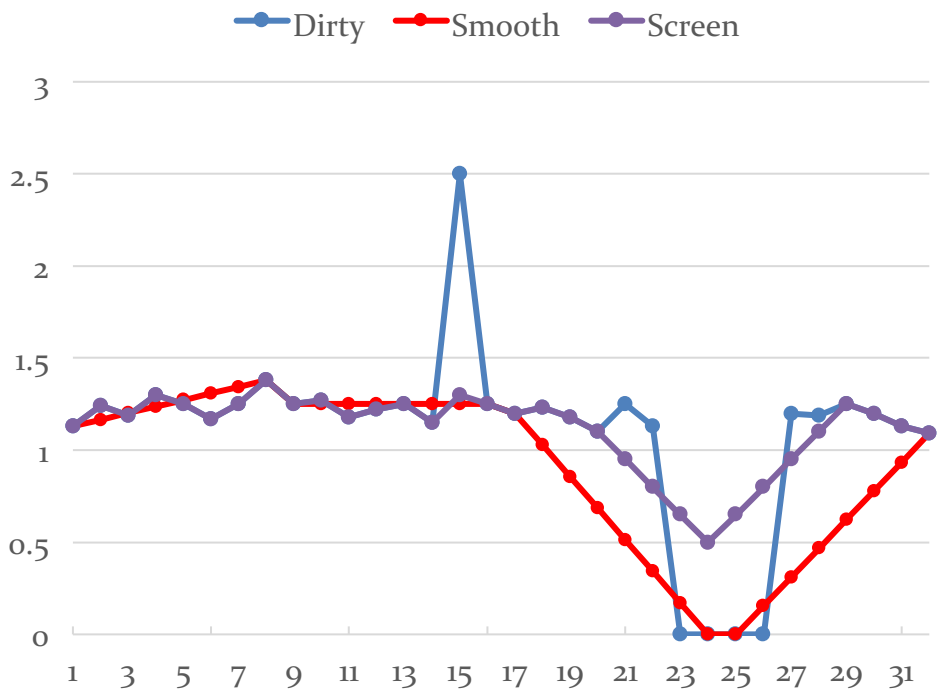


- Given
  - Time series $x = \{x[1], x[2], \dots\}$
  - Constraints $s = (s_{min}, s_{max})$ on min/max speeds
- Find repair a repair x' of x
  - Constraint satisfaction:

    $0 \le t_j - t_i \le w,$

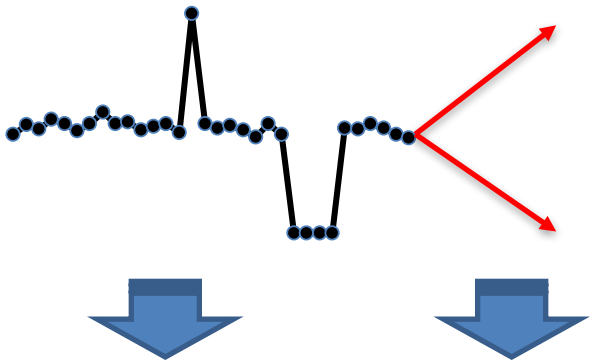
    $s_{min} \le \frac{x_j - x_i}{t_j - t_i} \le s_{max}$
  - Change minimization:

    $\sum_{x_i \in x} |x_i - x_i'|$ is minimized



Employ Existing Repairing Approach

- Holistic algorithm
  - Repairing relational data
  - Under denial constraints



ID	Timestamp	Value
1	1	1.13
2	2	1.24
3	3	1.19
4	4	1.3

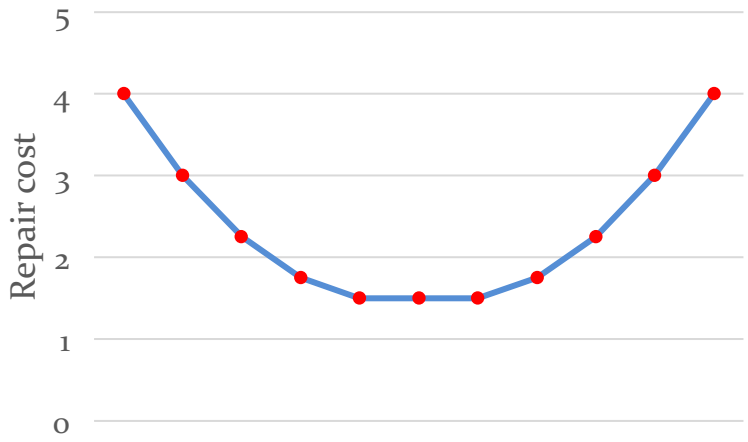
- Adaption
  - Time series as a relation
  - Express speed constraints by denial constraints roughly
- Problem
  - High computational costs
  - Not guaranteed to eliminate all violations

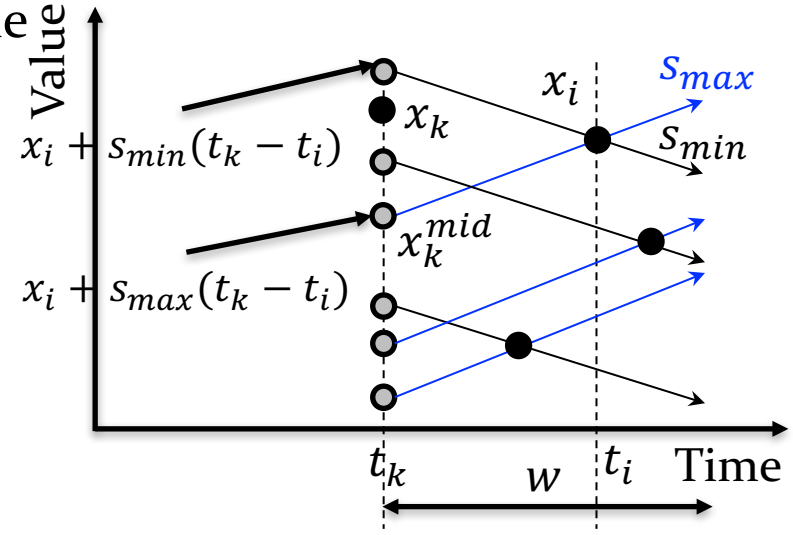
$$\neg(t_j < t_i + w \wedge x_j > x_i + (t_j - t_i) \cdot s_{max})$$
$$\neg(t_j < t_i + w \wedge x_j < x_i + (t_j - t_i) \cdot s_{\min})$$

X. Chu, I. F. Ilyas, and P. Papotti. Holistic data cleaning: Putting violations into context. In ICDE, pages 458-469, 2013.

A Lightweight Weapon

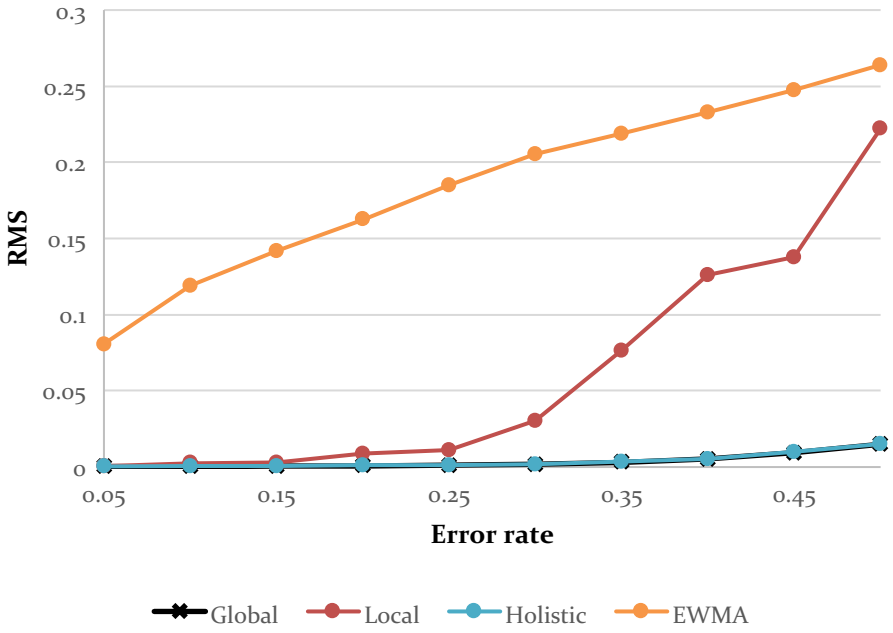
- Unlike NP-hard problems in most data repairing scenarios
- The speed constraint-based repairing can be solved
  - as a LP problem in $O(n^{3.5}L)$
  - considers the entire sequence as a whole (global optimal)
- Online computing, over streaming data
  - Consider local optimum in the current sliding window
  - Using Median Principle in $O(nw)$ time

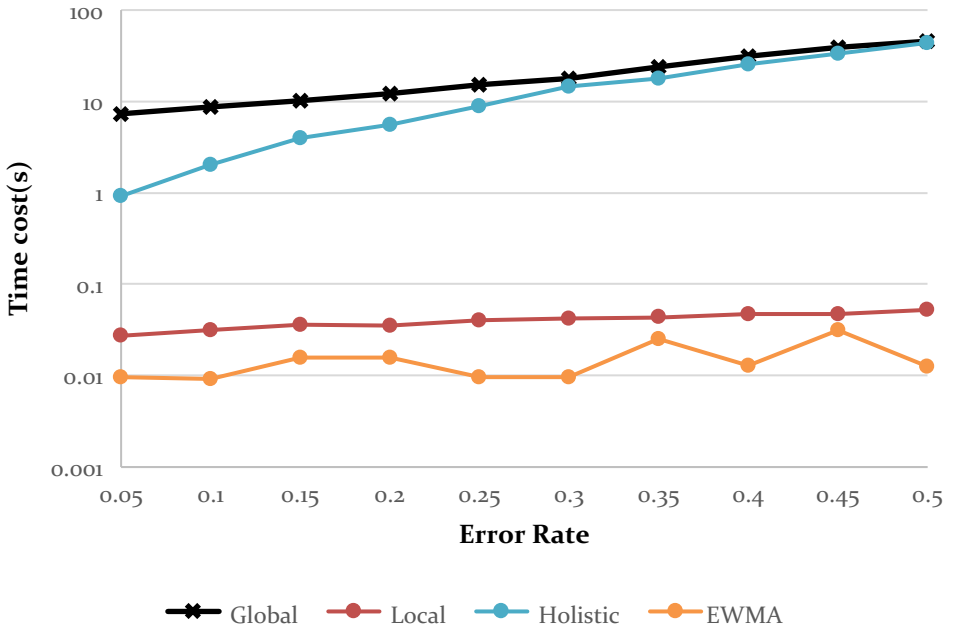




Effectiveness and Efficiency

- Global: the highest accuracy
- Local: much faster than Holistic
- Trade-off





Constraint-based method (SIGMOD 2015)

large spike errors

Statistical method (SIGMOD 2016)

small errors

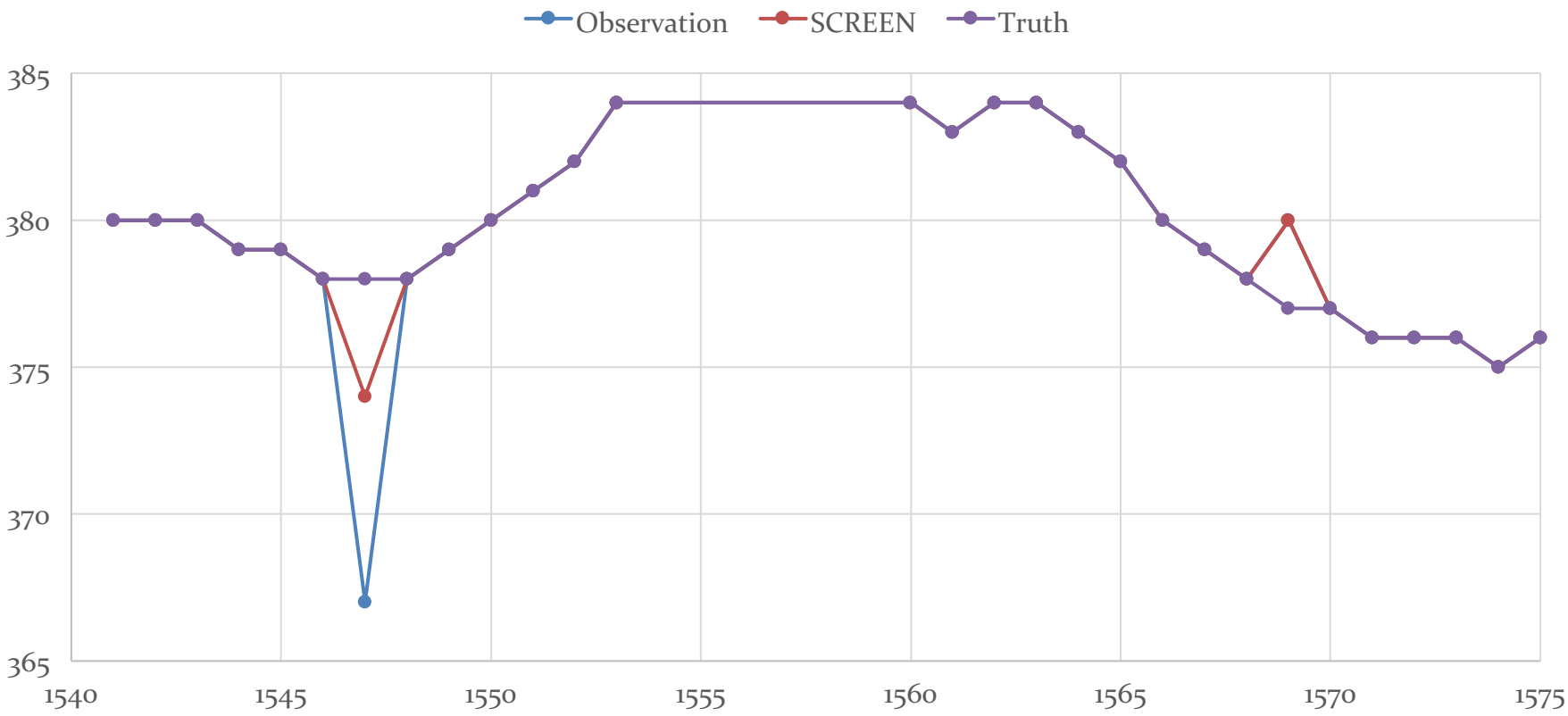
Supervised method (VLDB 2017)

consecutive errors

Contents

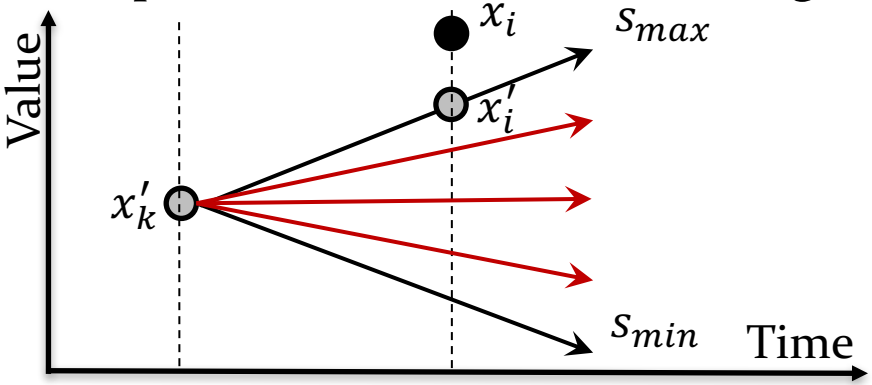
Further Issue

- Speed Constraint based method
  - Large spike error: modify to max/min values allowed
  - Small error: fail to identify

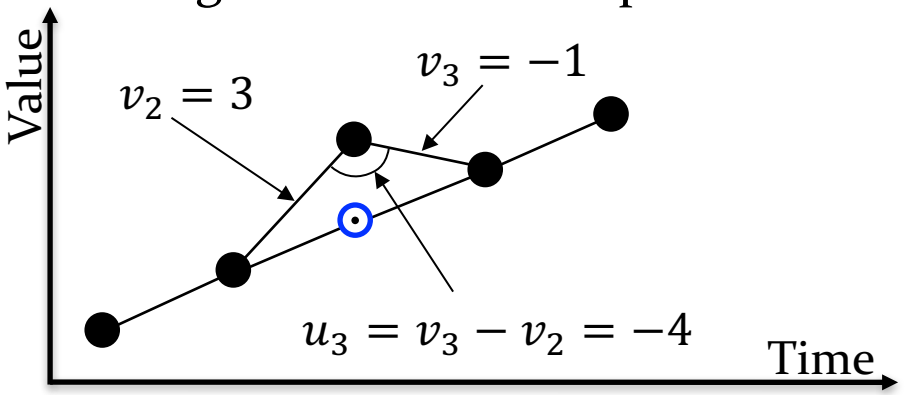


Intuition on Speed Change

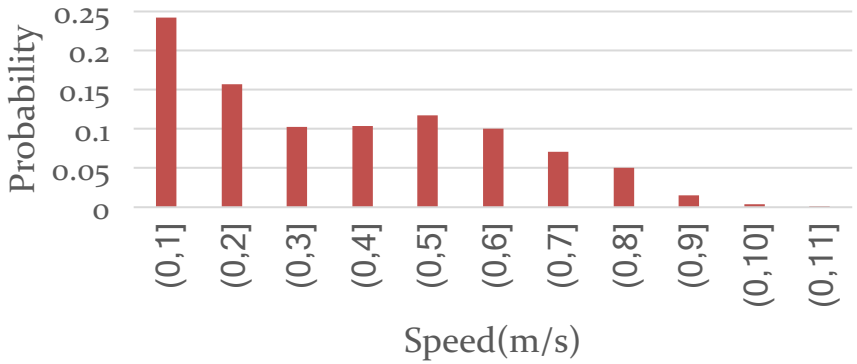
- Consider the likelihood of speeds within the allowed range

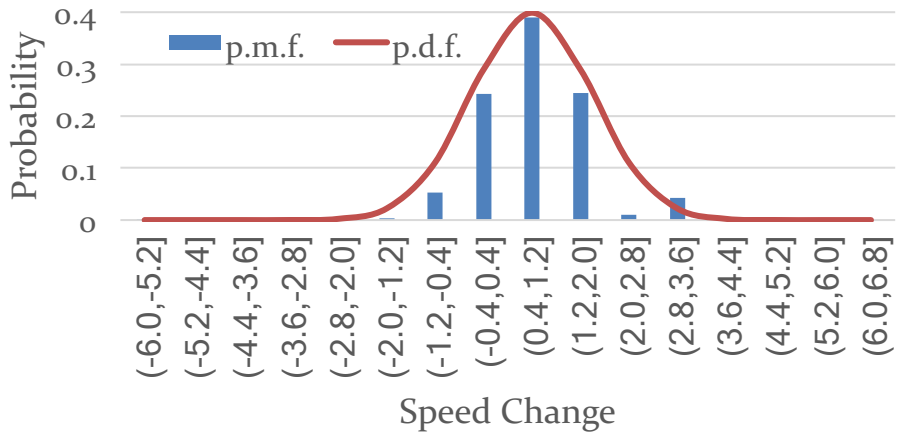


- Interesting pattern on speed changes in consecutive points



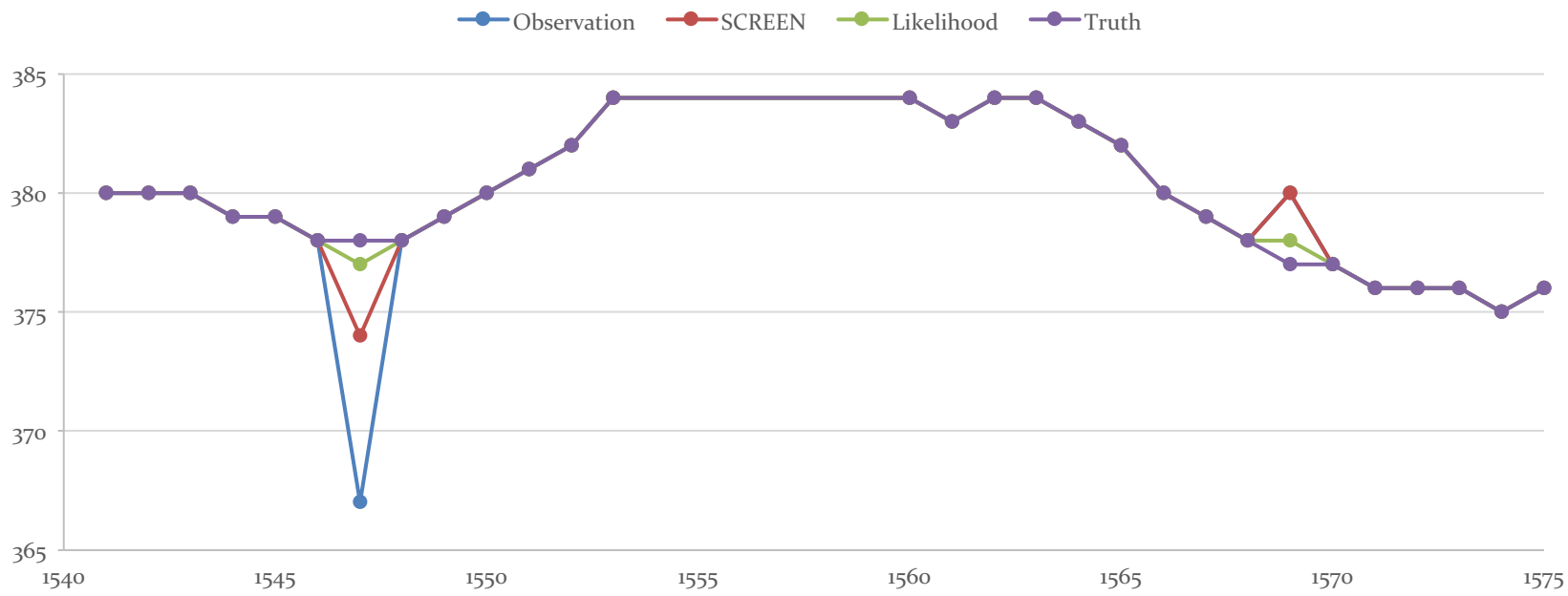
- No clear distribution pattern is observed on speeds





Statistical Approach

- Calculate the likelihood of a sequence w.r.t. the speed change
  - employ the probability distribution of speed changes
- The cleaning problem is thus to find a repaired sequence with the maximum likelihood about speed change
  - instead of minimum change towards speed constraint satisfaction



Maximum likelihood repair problem

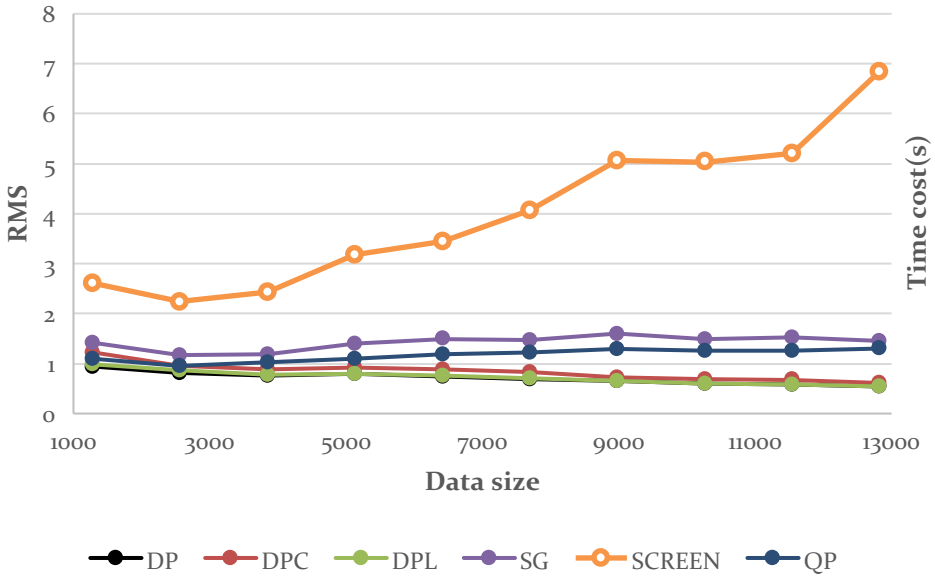
- Given
  - Time series x
  - repair cost budget δ
  - Distribution on speed changes
- Find repair a repair x' of x
  - $\Delta(x, x') \leq \delta$
  - the likelihood $L(x')$ is maximized.

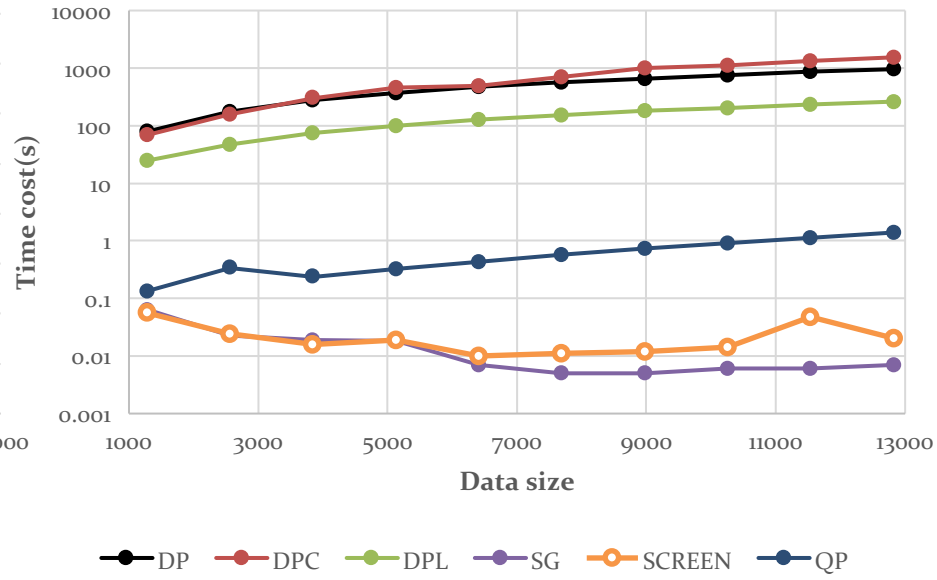
- NP-hard
- Pseudo-polynomial time solvable

DP, dynamic programming	$O(n\theta_{max}^3\delta)$	Exact
DPC, constant-factor approximation	$O(n^2\theta_{max}^3)$	Large budget
DPL, linear time heuristics	$O(nd^4)$	Fast, higher error
QP, quadratic programming		Approximate distribution
SG, simple greedy	$O(\max(n, \delta))$	Fastest

Effectiveness and Efficiency

- Significantly better accuracy than SCREEN
- SG is efficient, comparable to SCREEN, and still with better accuracy





Constraint-based method (SIGMOD 2015)

large spike errors

Statistical method (SIGMOD 2016)

small errors

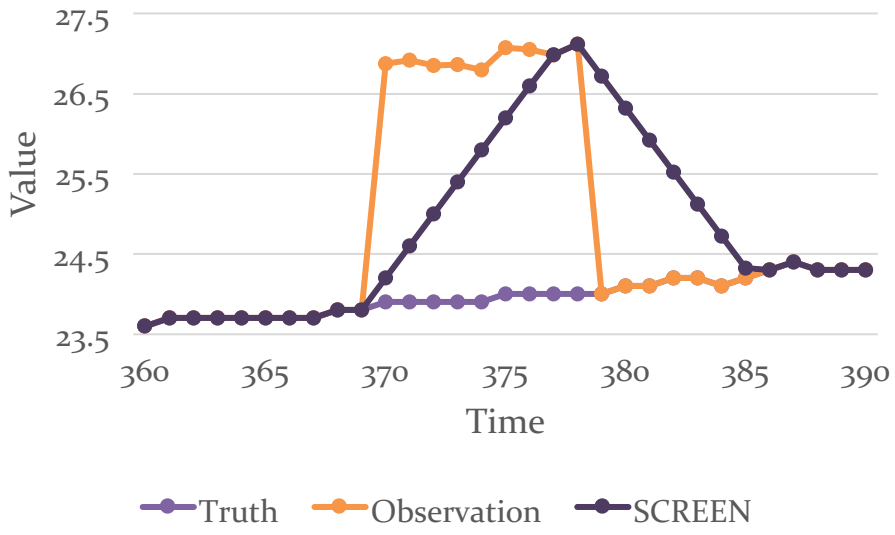
Supervised method (VLDB 2017)

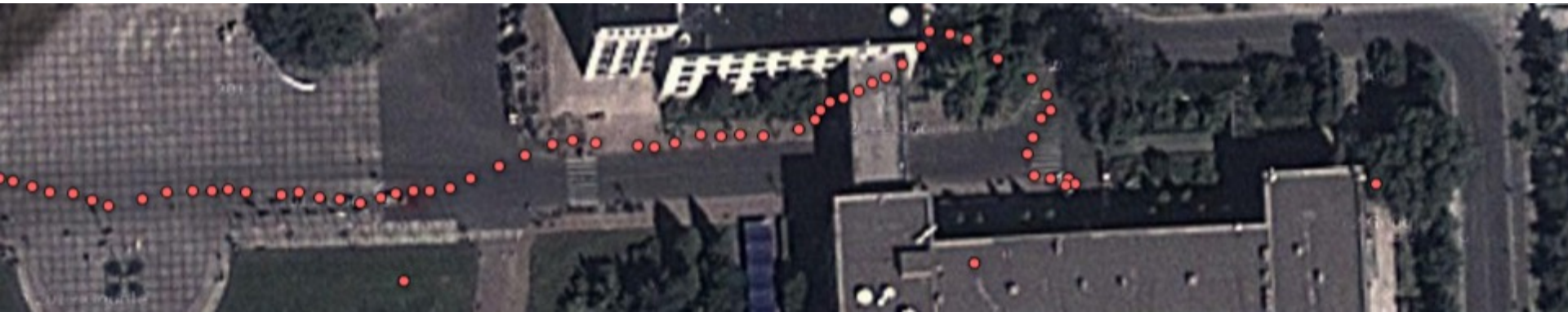
consecutive errors

Contents

Consecutive Errors

- Speed constraints handle well “Spike” errors, but not consecutive ones





Intuition

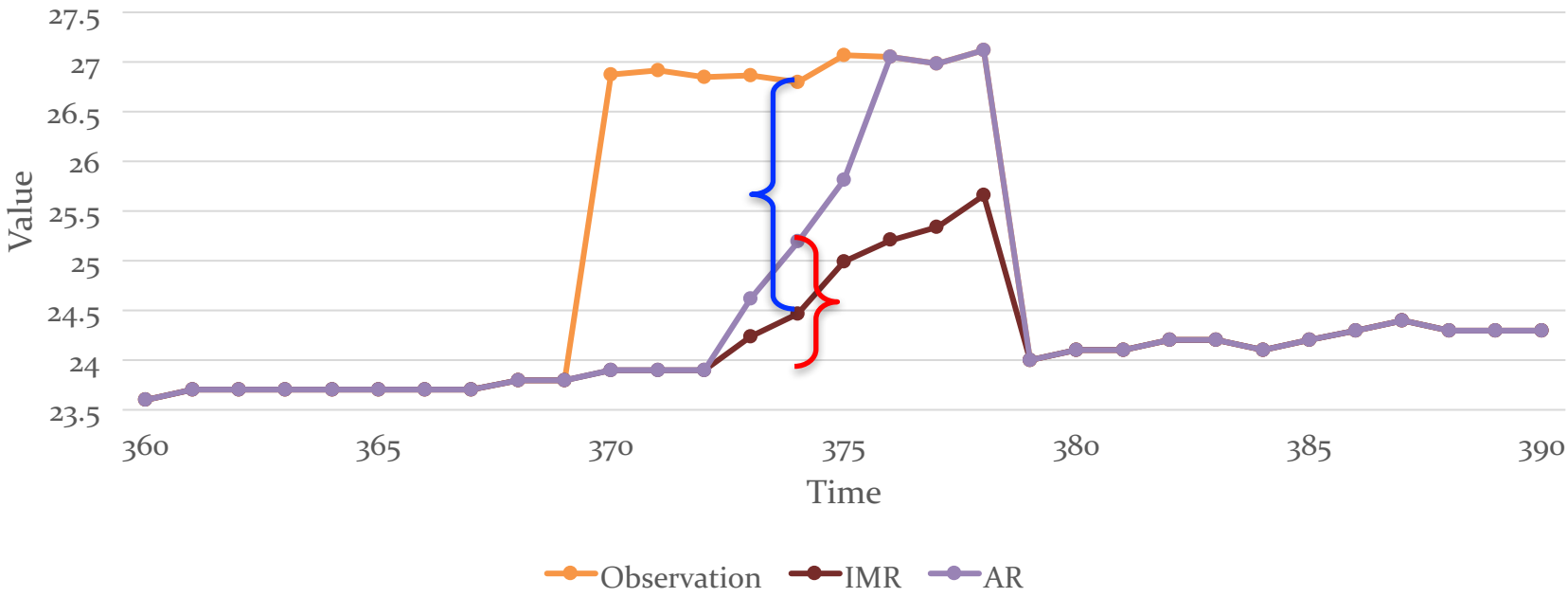
- Supervised by labeled truth of errors

- Labeling by user
  - Check-in
- Labeling by machine
  - precise equipment reports accurate air quality data in a relatively long sensing period
  - crowd and participatory sensing generates unreliable observations in a constant manner

Y. Zheng, F. Liu, and H. Hsieh. U-air: when urban air quality inference meets big data. In KDD, pages 1436–1444, 2013.

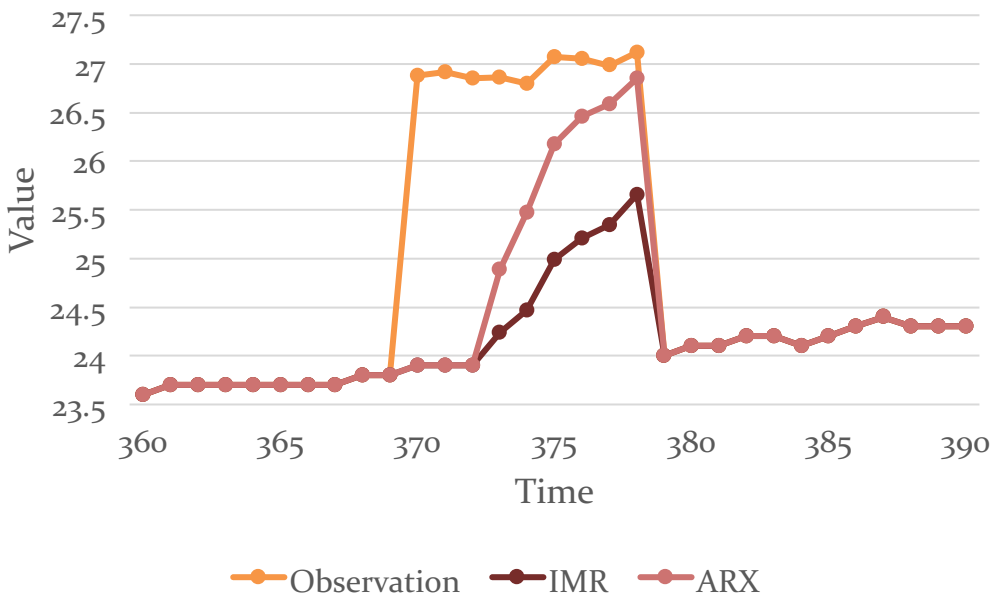
Approach

- Instead of modeling directly the values
  - by AR model (autoregression), ignoring erroneous observations
- We model and predicate the difference between errors and their corresponding labeled truths
  - by ARX model (autoregressive model with exogenous inputs)



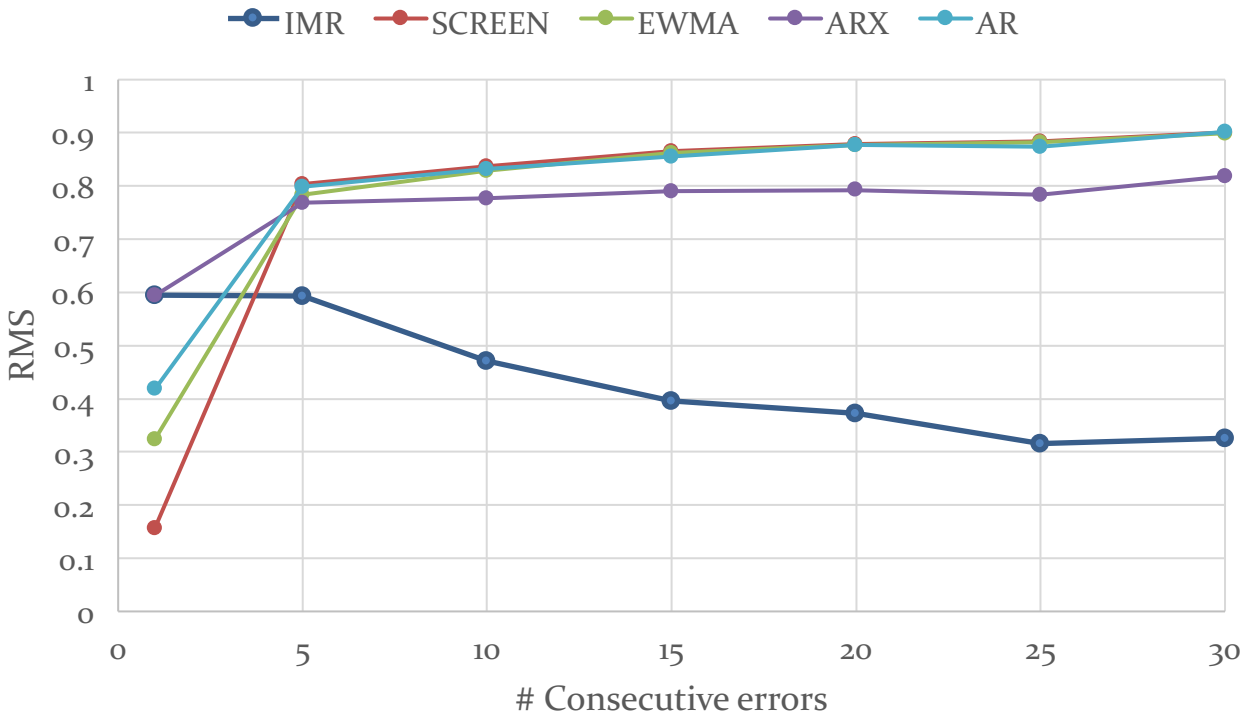
Iterative Minimum Repair (IMR)

- Rather than in chronological order
- Iterative repairing
  - minimally changes one point a time to obtain the most confident repair only
  - high confidence repairs in the former iterations could help the latter repairing
- Major concerns
  - Convergence
  - Incremental computation among iterations



Dealing with consecutive errors

- IMR shows significantly better results when there is a large number of consecutive errors



Constraint-based method (SIGMOD 2015)

large spike errors

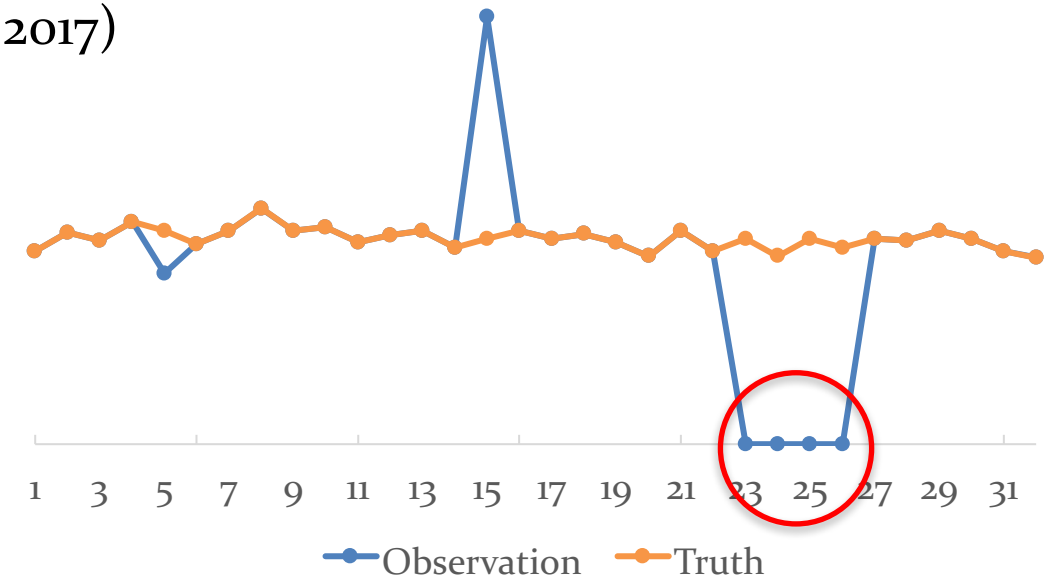
Statistical method (SIGMOD 2016)

small errors

Supervised method (VLDB 2017)

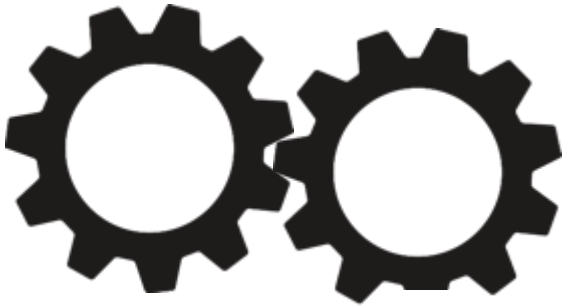
consecutive errors

Contents



Future Study

- More error types
  - Periodical



- Timestamp error
  - A single ride takes 20 years

http://gd.sina.com.cn/news/sz/2016-07-08/detail-ifxtwihp9790656.shtml

1. Shaoxu Song, Aoqian Zhang, Jianmin Wang, Philip S. Yu. SCREEN: Stream Data Cleaning under Speed Constraints. ACM SIGMOD International Conference on Management of Data, SIGMOD, 2015.
2. Aoqian Zhang, Shaoxu Song, Jianmin Wang. Sequential Data Cleaning: A Statistical Approach. ACM SIGMOD International Conference on Management of Data, SIGMOD, 2016.
3. Aoqian Zhang, Shaoxu Song, Jianmin Wang, Philip S. Yu. Time Series Data Cleaning: From Anomaly Detection to Anomaly Repairing. International Conference on Very Large Data Bases, VLDB, 2017.

Thanks

Full text available at

http://ise.thss.tsinghua.edu.cn/sxsong/