Data Quality on Biological Data Series: an Empirical Study †

Angela Bonifati and Bastien Rance

Université Lyon 1 - CNRS LIRIS and HEGP Paris

December 11, 2017







†. Supported by CNRS Mastodons MedClean (2016-2017)

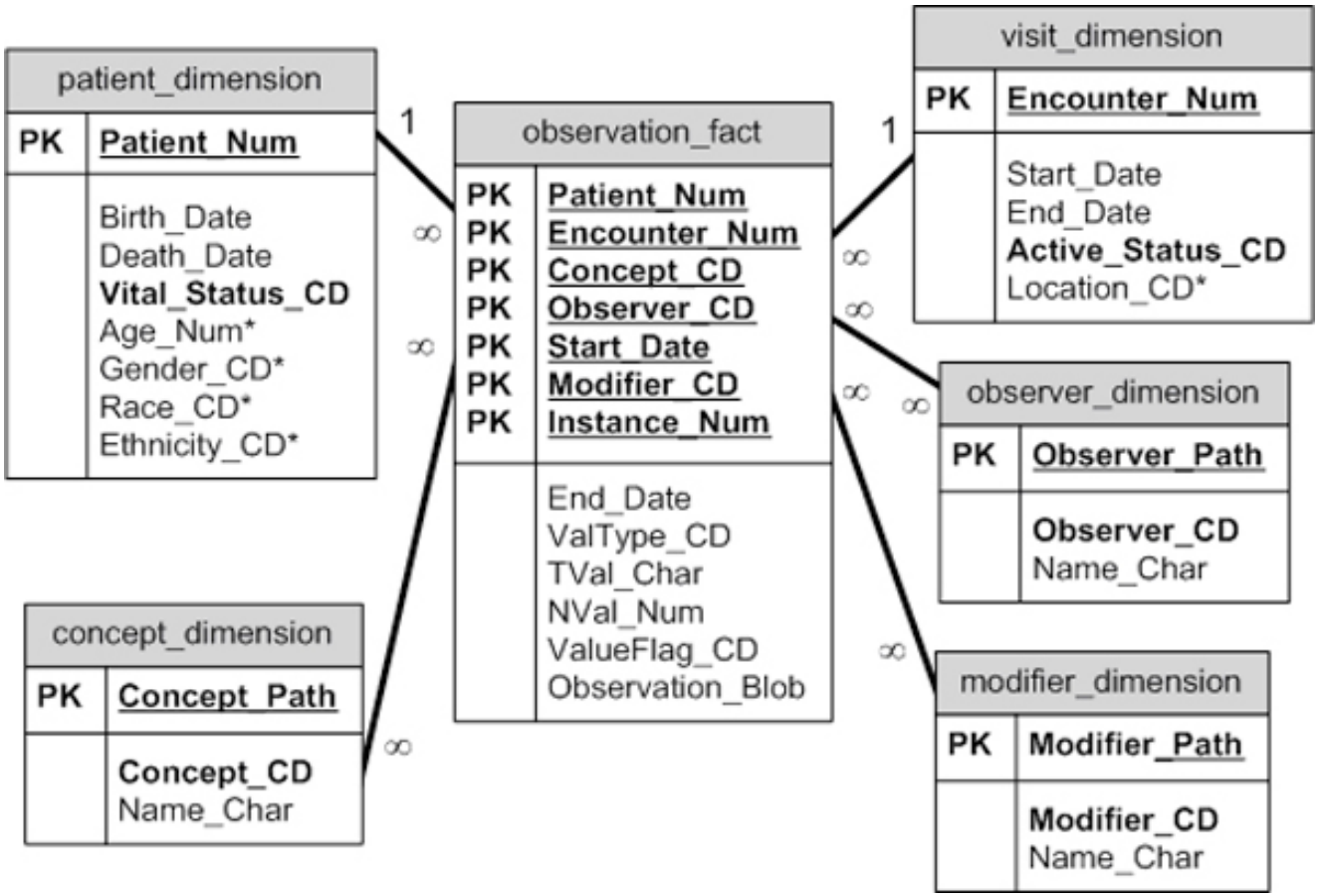
Data Quality in Healthcare

Underlying Setting : Clinical Data warehouse

- The data warehouse i2b2 (at HEGP) contains massive heterogeneous healthcare data (concerning roughly 700.000 patients and amounting to $7.7M$ diagnosis codes, $3.7M$ clinical reports, $1.5M$ clinical image reports and $125M$ of laboratory results).
- The data need to be annotated and sanitized in order to be used for clinical studies.

We focus in the remainder on a subset of the i2b2 data : temporal data (over 17 years) representing 202 different laboratory measurements (for a total of $90M$ laboratory results).

Schema of the data warehouse



Glitches in biological temporal data

i2b2 and biological data

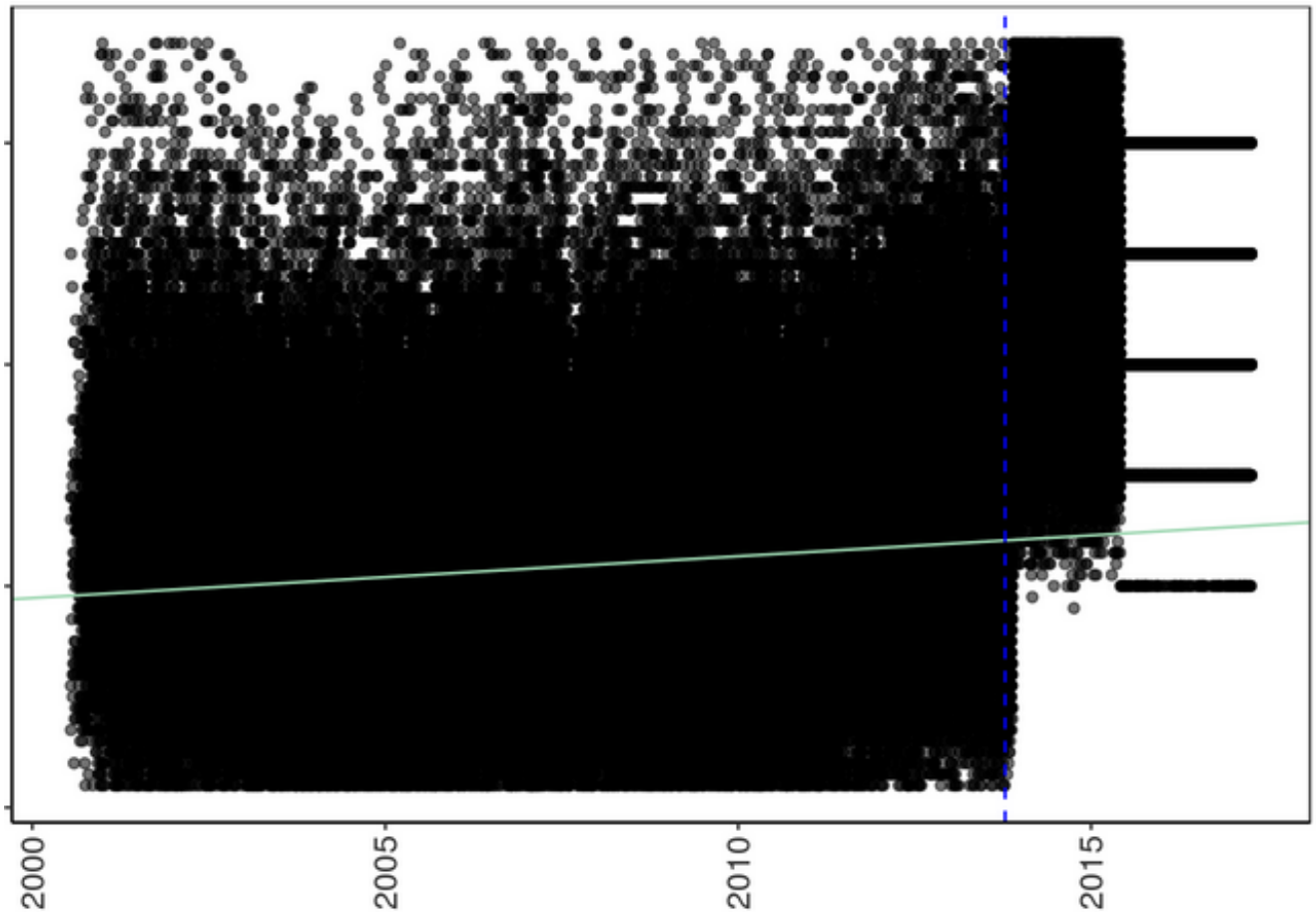
- Biological temporal data series contain millions of numerical values grouped by concept (laboratory exam).
- They contain several distribution glitches that have to be taken into account in secondary use of clinical data [a] for translational medicine [b].

---

a. e.g. using data collected during patient care for clinical research

b. Translational Research focuses on original investigations in the fields of laboratory, clinical and health care research by conducting preclinical research with potential for application to human disease and/or leveraging preliminary human experimentation with potential to refine the understanding of biological principles underpinning human disease.

Motivation for this study



Plan

1. Introduction
2. Glitches under the lenses
3. Detection of glitches
4. Conclusion

Plan

1 Introduction

2 Glitches under the lenses

3 Detection of glitches

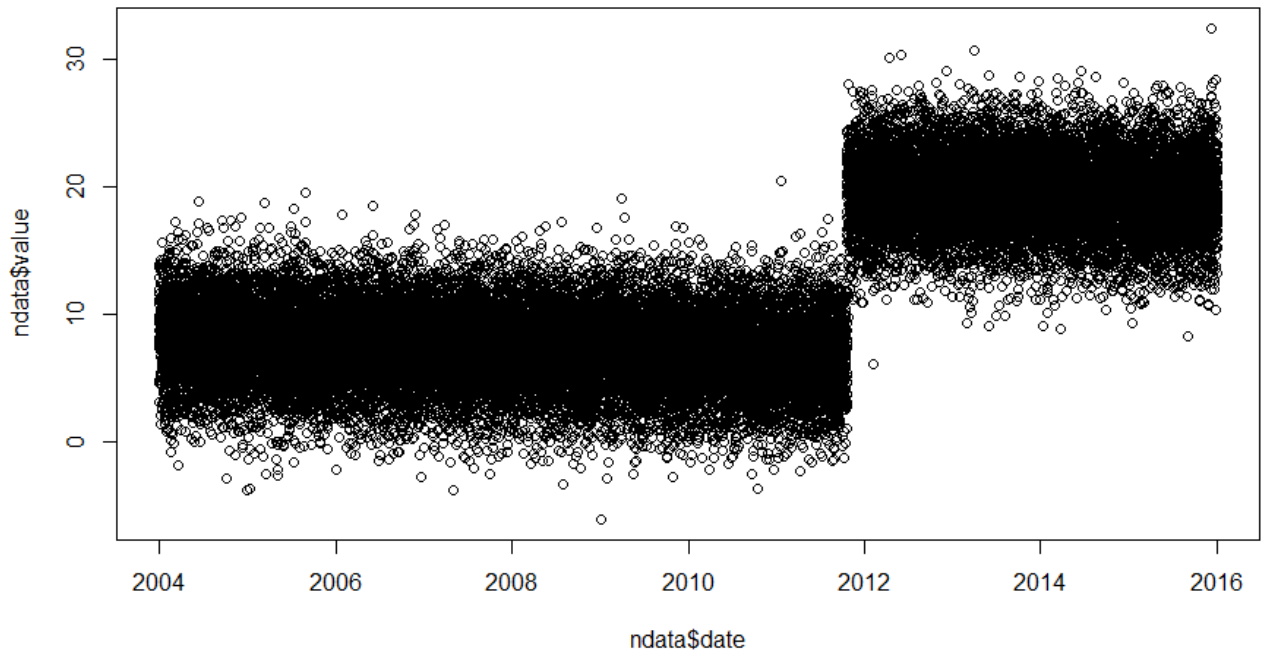
4 Conclusion

Glitches considered in the analysis

Definition of a Glitch

- Abrupt discontinuity in the value distributions, temporary at times.
- There exist several classes, each of which has its peculiarities in the detection process.
- Among them, in our study we mainly focused on (1) breakpoints ; (2) discretizations ; (3) missing data ; (4) trends.

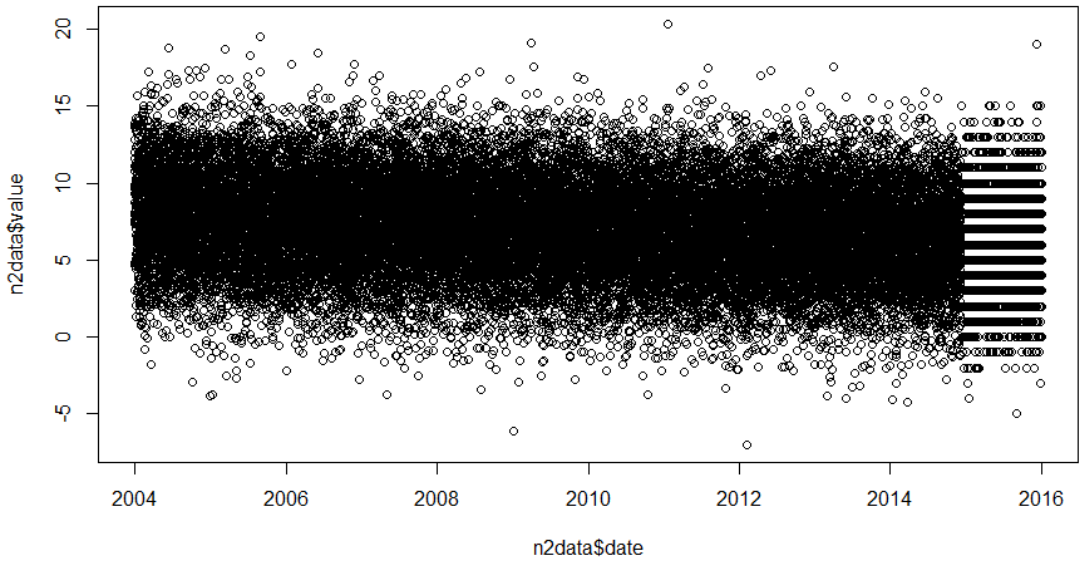
Breakpoints



Informal definition

- A breakpoint is an abrupt drop or increase of the mean of the distribution. It can also be called a distribution shift.

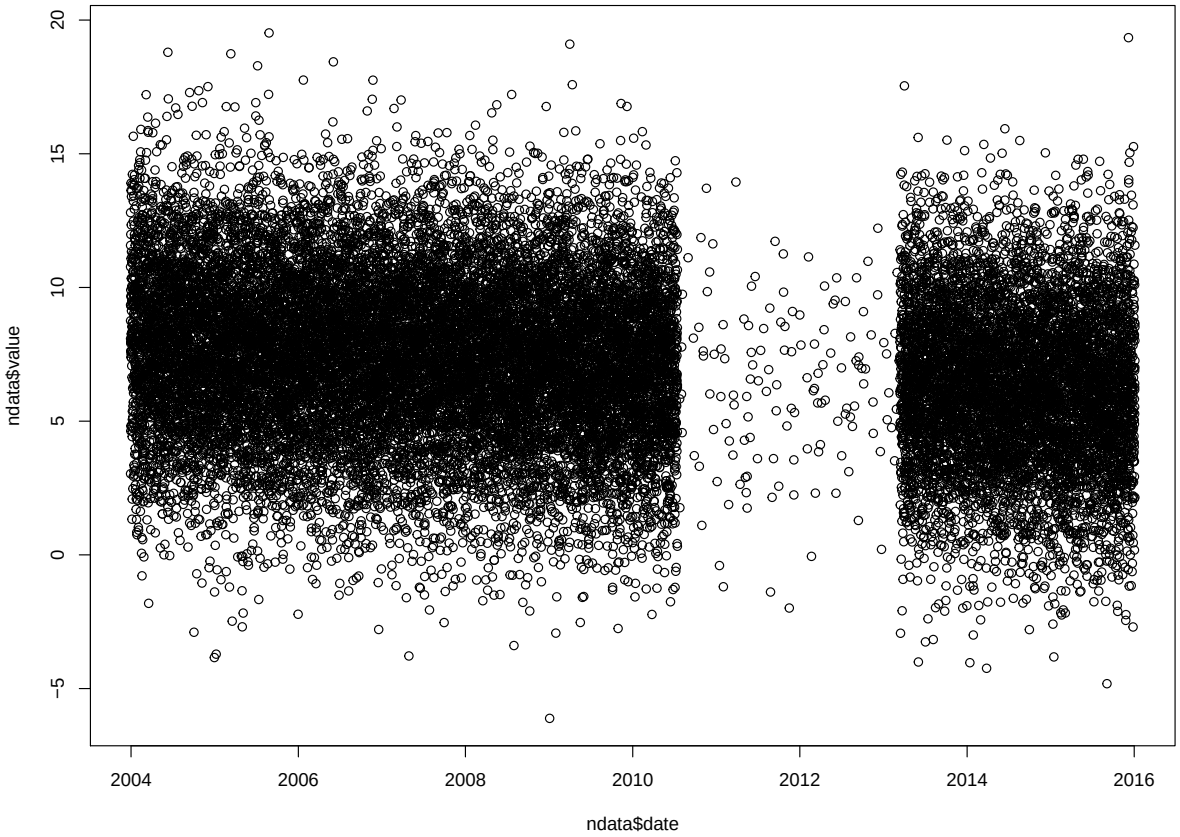
Discretization



Informal definition

- Discretization corresponds to a change in the distribution of decimals (from 0.001 to 0.01,...) of data values.
- The inverse is equally interesting : the distribution goes from discrete to continuous data.

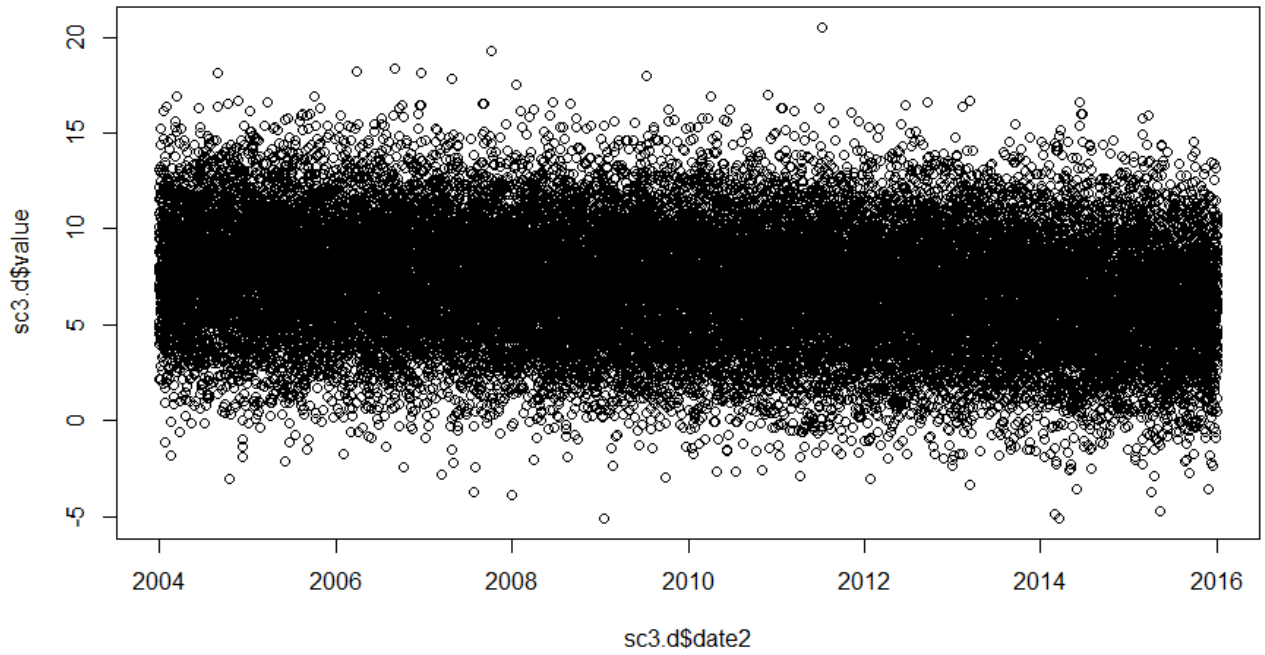
Missing data



Informal definition

- Partial or complete disappearance of pieces of the distributions.

Trends



Informal definition

- Progressive evolution of the data distribution over time : identical standard deviation but evolving average.

Plan

1 Introduction

2 Glitches under the lenses

3 Detection of glitches

4 Conclusion

Breakpoint

Adopted Algorithm

- Data standardization by computing a z-score : $z = \frac{x-\mu}{\sigma}$, where x is the raw value, μ is the mean of the population and σ is the standard deviation of the population.
- To avoid outliers and random noise, a median is computed over a 30 days window : for a day d, a median is computed by considering measurements on $d - 15$ days and $d + 15$ days.
- Detection of *distribution shifts* : adoption of the PELT (Pruned Exact Linear Time) method [a] on the standardized time series.
- The output of the algorithm are the dates for which breakpoints are detected.
- Usage of a sliding window of convenient size for the detection

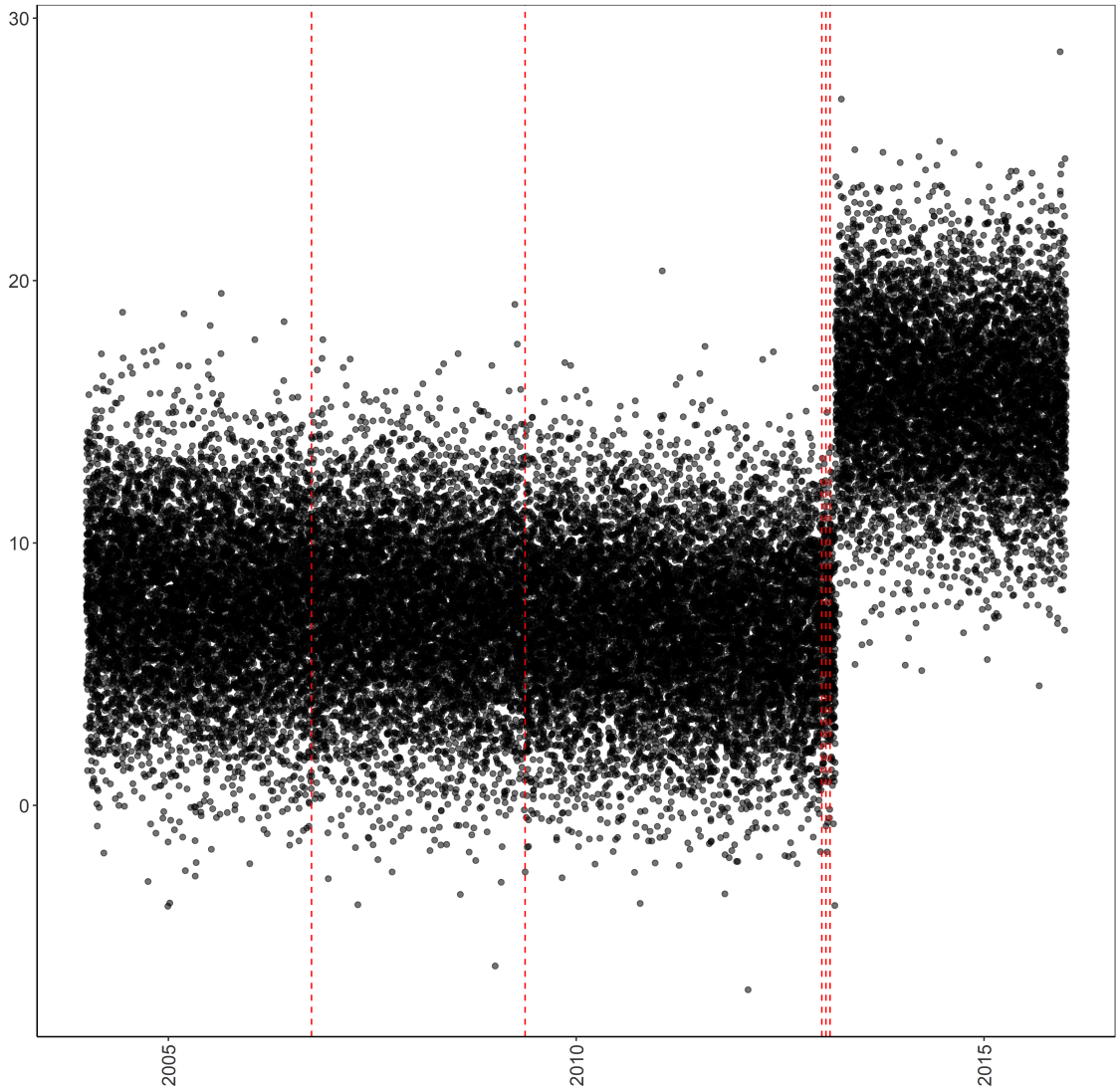
a. Killick R, Fearnhead P, Eckley IA (2012) Optimal detection of changepoints with a linear computational cost, JASA107(500), 1590 ?1598

Breakpoint : cons of the applied algorithm

Limitations

- False positives : a breakpoint is interpreted as such whereas it should not be.
- A slight breakpoint is not captured.
- The size of the sliding window is arbitrarily (and empirically) chosen.

Breakpoint



Discretization : A New Algorithm

Adopted approach

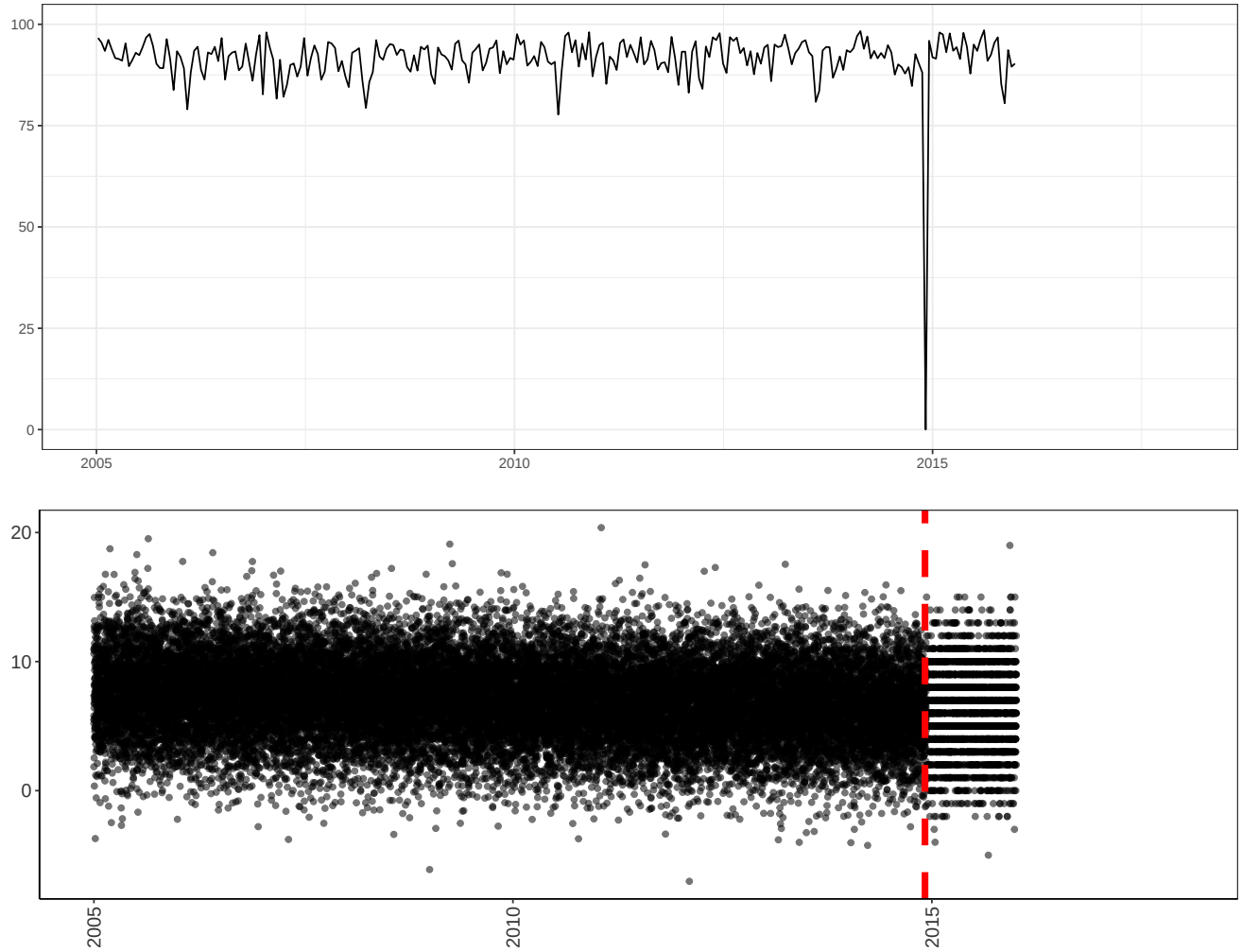
- Detection of the changes of the distribution of values from continue value to discrete values : number of digits and last digit are used to detect the changes.
- Computation of the differences via the cosine between the values of two adjacents sliding windows.

Discretization : an example

- Consider the following distribution : (1, 5, 15, 3, 25, 21.1). We define three classes : one digit numbers, two digits numbers and one decimal numbers.

Class	Last dec./digit	# occurrences in the last position	# occurrences in the class	Ratio #last_digit/#class
Int_1	1	1	3	33.33%
Int_1	3	1	2	33.33%
Int_1	5	1	2	33.33%
Int_2	5	2	2	100%
Dec_1	1	1	1	100%

Detection of evolution over time : adoption of the cosine similarity of the vectors of ratios



Missing data

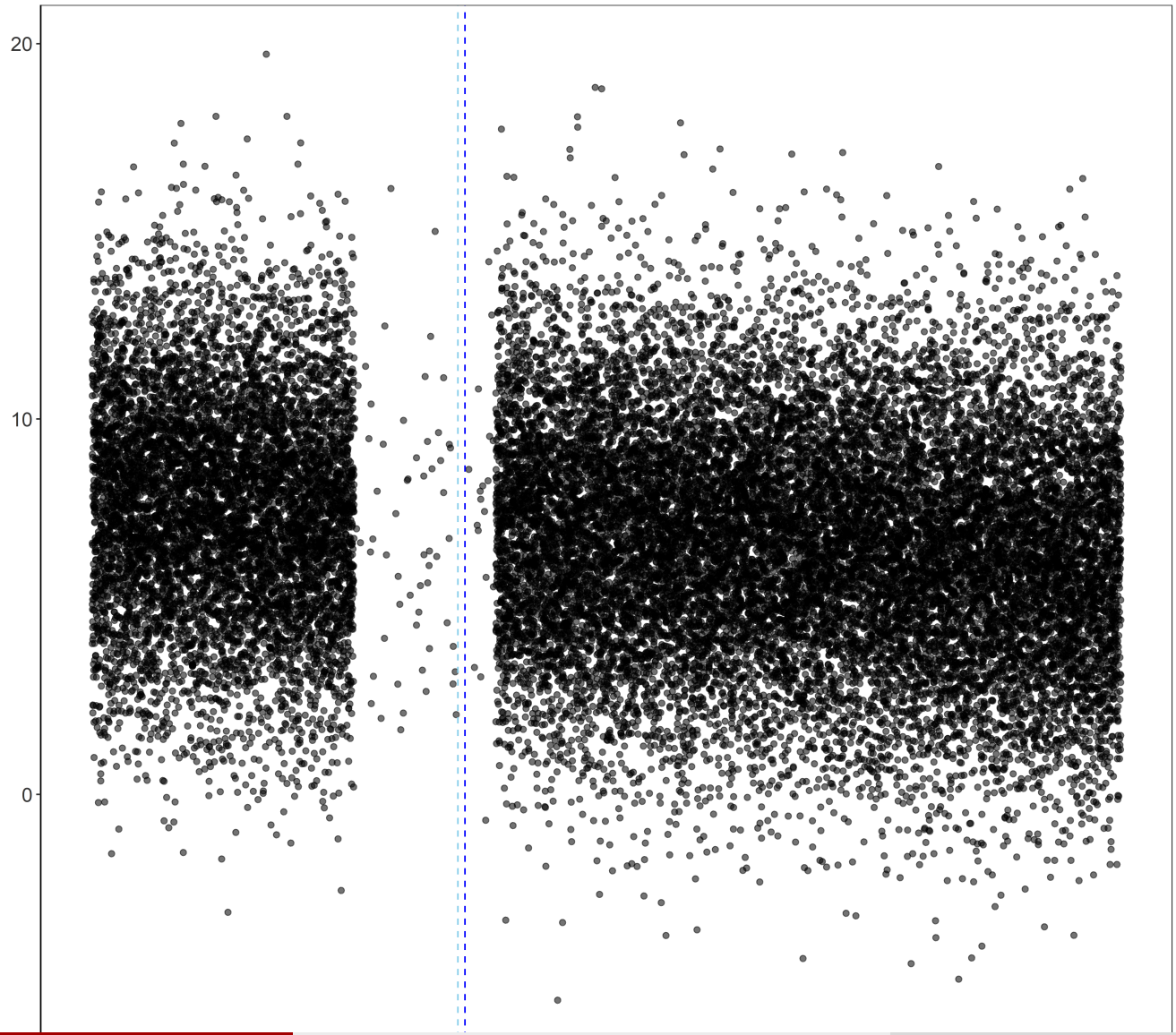
Adopted approach

- At present, naïve implementation based on counts of the number of data values.

Limitations

- Threshold defined arbitrarily for the detection of changes
- Does not detect small variations

Missing data



Trends

Adopted approach

- Computation of the trend of the normal probability plot and of the inferior and superior quartiles (Koenker and d'Orey method, Bonferroni correction)

Limitations

- Arbitrarily chosen threshold.
- Uncertain/imprecise detection in some cases.

Future Perspectives : Other glitches/combinations of glitches

- In the above empirical study, we considered one glitch at a time. What about considering other glitches (e.g. outliers) ?
- Study the masking effect of missing data.

Possible directions

- Focus on masking indexes in the spirit of related work [a] on the effect of missing data on outliers.

---

a. Laure Berti-Equille, Ji Meng Loh, Tamraparni Dasu : A masking index for quantifying hidden glitches. Knowl. Inf. Syst. 44(2) : 253-277 (2015)

Future Perspectives : Constraint Discovery

Preliminary investigation

- We ran Metanome [a] on healthcare data to discover functional dependencies. We could not find meaningful results on our local setting.

---

a. Thorsten Papenbrock, Tanja Bergmann, Moritz Finke, Jakob Zwiener, Felix Naumann : Data Profiling with Metanome. PVLDB 8(12) : 1860-1863 (2015)

Possible directions

- Restrict to conditional functional dependencies or restricted forms of denial constraints.
- Is there a connection between statistical glitches and logical constraints ?

Future Perspectives : Provenance and context

- Quality metrics are difficult to establish in a context-independent fashion.

Possible directions

- Addition of metadata [a] in order to characterize and leverage context information on top of data distributions.
- Usage of such metadata to better detect and explain the glitches.

---

a. Melanie Herschel, Ralf Diestelkamper, Houssem Ben Lahmar : A survey on provenance : What for ? What form ? What from ? VLDB J. 26(6) : 881-906 (2017)

Plan

1 Introduction

2 Glitches under the lenses

3 Detection of glitches

4 Conclusion

Conclusion

Summary

- Preliminary study on detection of glitches in healthcare.
- Plenty of open problems in both CS and medicine.

Other results and Improvements

- Findings around the adoption of non arbitrary thresholds (and meaningful ones !).
- Generator of distribution glitches : as a byproduct of this work, we have implemented a generator that can be applied to clean data and reproduce the same glitches observed in healthcare data.

MedClean : other lines of research

- Data Quality : User-guided repairing of Knowledge Bases (will be presented tomorrow).
- Log Analysis : An Analitycal Study of Large Graph Query Logs (including BioMed and BioPortal queries) [a]

a. A.Bonifati, W.Martens, T.Timm : An Analytical Study of Large SPARQL Query Logs. PVLDB 11(2) : 149-161 (2017)

Thanks for your attention.

Thanks to our co-authors [a] : V. Looten, L. Kong-Win-Chang, A. Neuraz, A. Burgun and the HEGP Biochemistry and Hematology teams.

a. V. Looten, L. Kong-Win-Chang, A. Neuraz, N. Rivet, L. Mauge, B. Vedie, J.L. Paul, M.A. Loriot, A. Burgun, A. Bonifati, B. Rance. Tracking changes in 202 laboratory measurements over 17 years : the challenge of temporal data quality in a clinical data warehouse (under preparation).