Sciences Economiques et Sociales de la Santé
& Traitement de l'Information Médicale

sesstim.univ-amu.fr

Bastien RANCE

Université Paris Descartes - Faculté de médecine - AP-HP ; Hôpital Européen Georges Pompidou
INSERM ; Centre de Recherche des Cordeliers - Equipe Burgun

Temporality in health data warehouses: the issue of data quality

février 2019

Cliquez ici pour voir l'intégralité des ressources associées à ce document

Qualité temporelle dans les entrepôts de données cliniques

Bastien Rance[1,2,3]

Vincent Looten[1,2,3]

1- Université Paris Descartes - Faculté de médecine
2- AP-HP ; Hôpital Européen Georges Pompidou
3- INSERM ; Centre de Recherche des Cordeliers ; Equipe Burgun

HEGP

Ouvert en 2000

HIMMS niveau 6

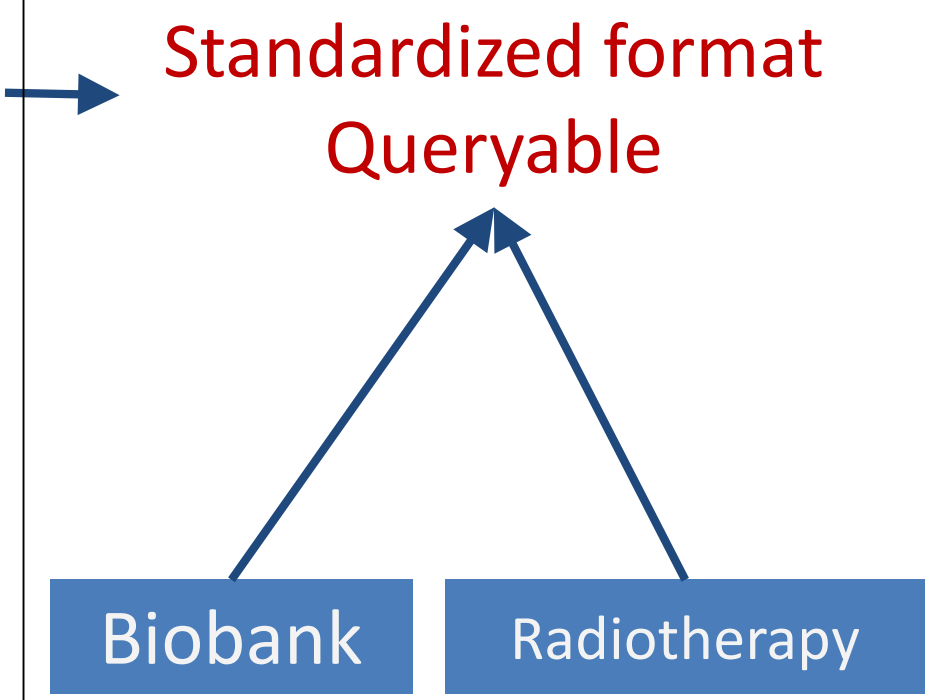
(http://www.himss.eu/node/1116)

Clinical Data Warehouse

Electronic Health Record (EHR)

Diagnosis
Clinical items
Billing codes
Biology (lab)
Nurse transmission
Imaging reports
Pathology reports
Drug prescription
Chemotherapy

Clinical Data Warehouse (CDW)

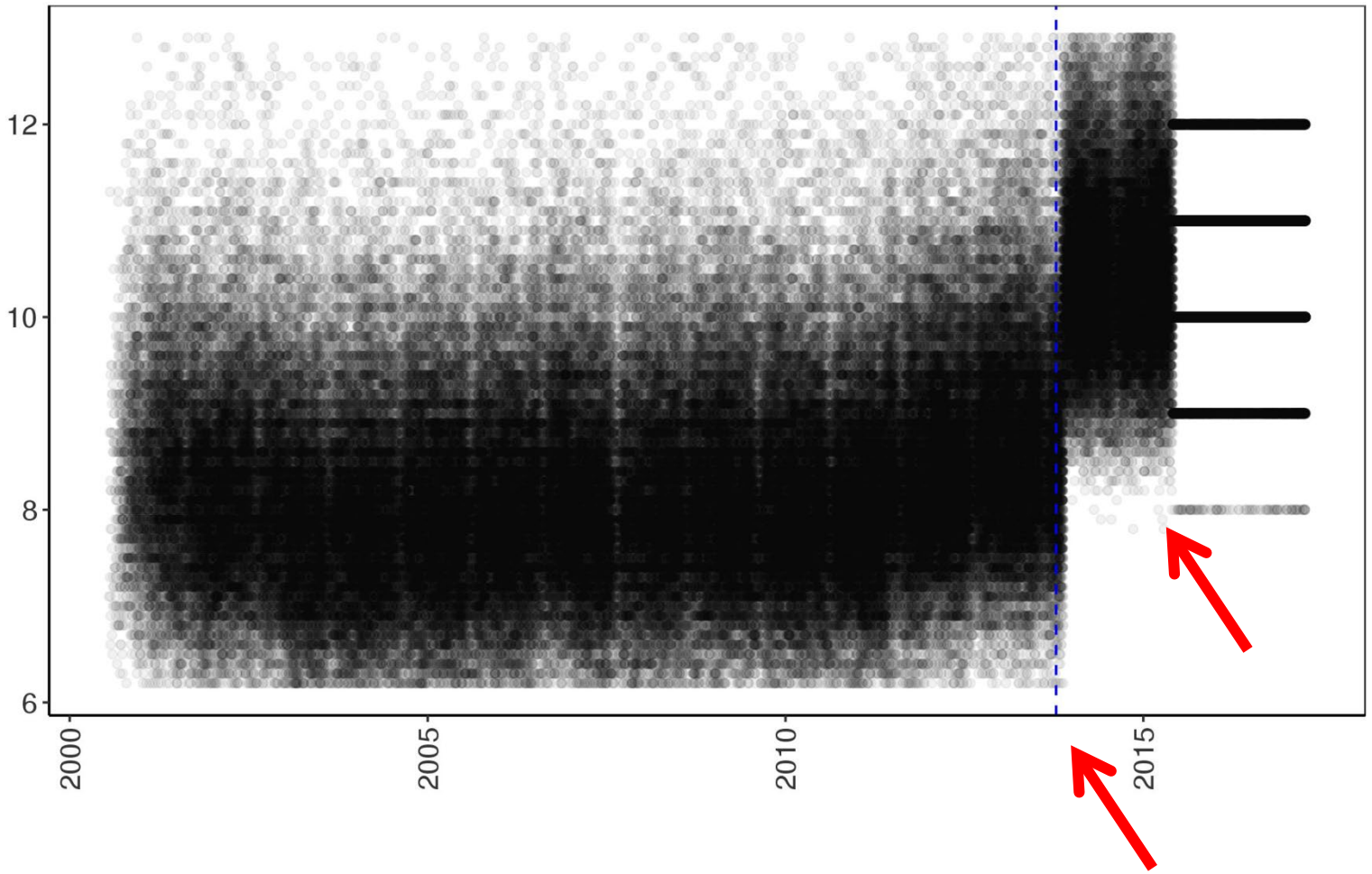


Clinical Data Warehouse at HEGP

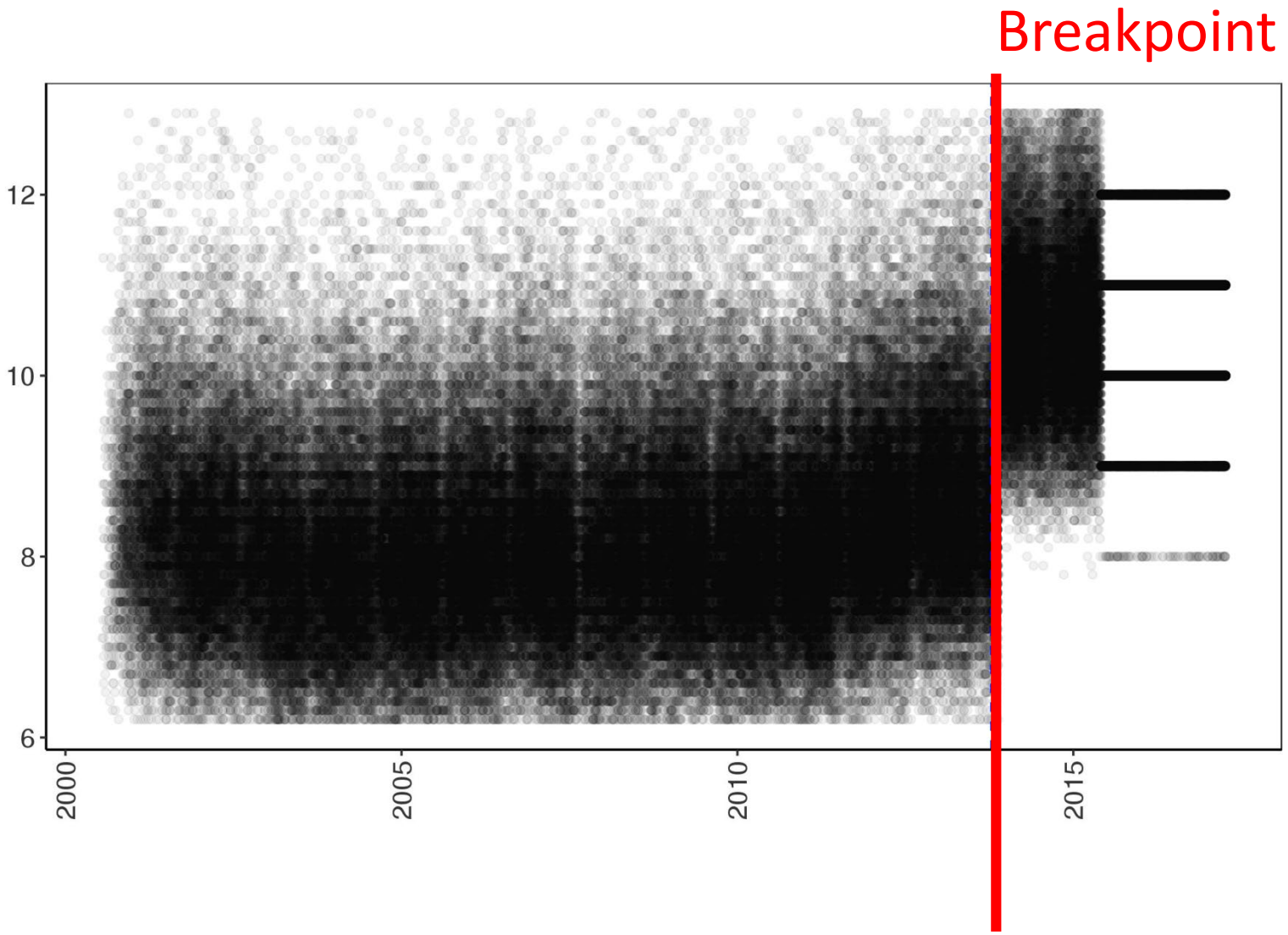
Concept	# patients	# observations
EHR concepts	602,198	124,852,989
Biology (Laboratory)	452,006	**132,525,661**
Nursing transmission	309,322	18,495,958
Billing (disease) codes	396,285	8,183,118
Rx prescription	191,531	7,243,484
Text reports	546,725	4,039,333
Imaging reports	351,702	1,325,270
Pathology codes	98,401	1,496,635

Unstructured data: transformation is needed before reuse

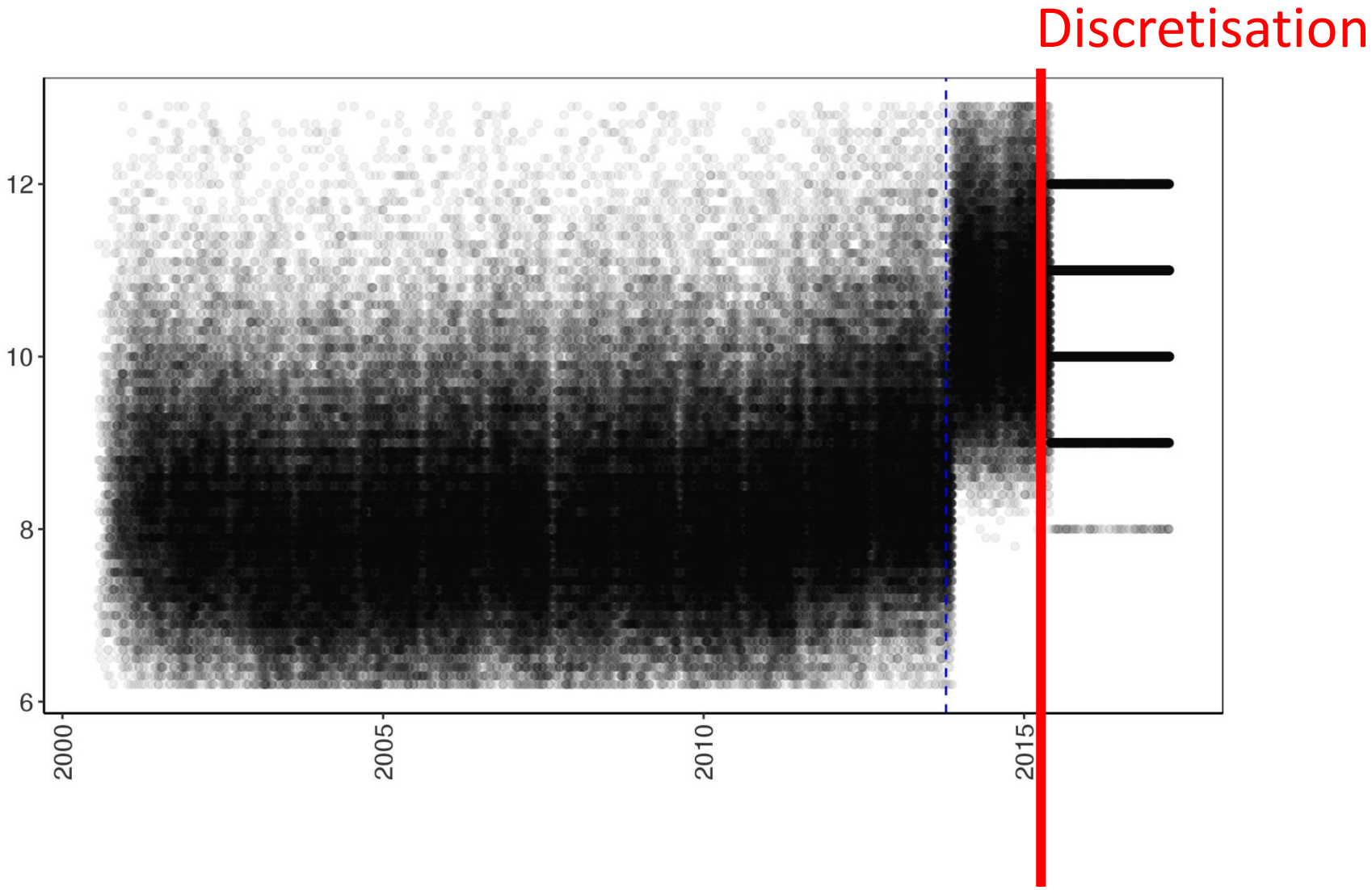
Pour commencer, visualisation des données



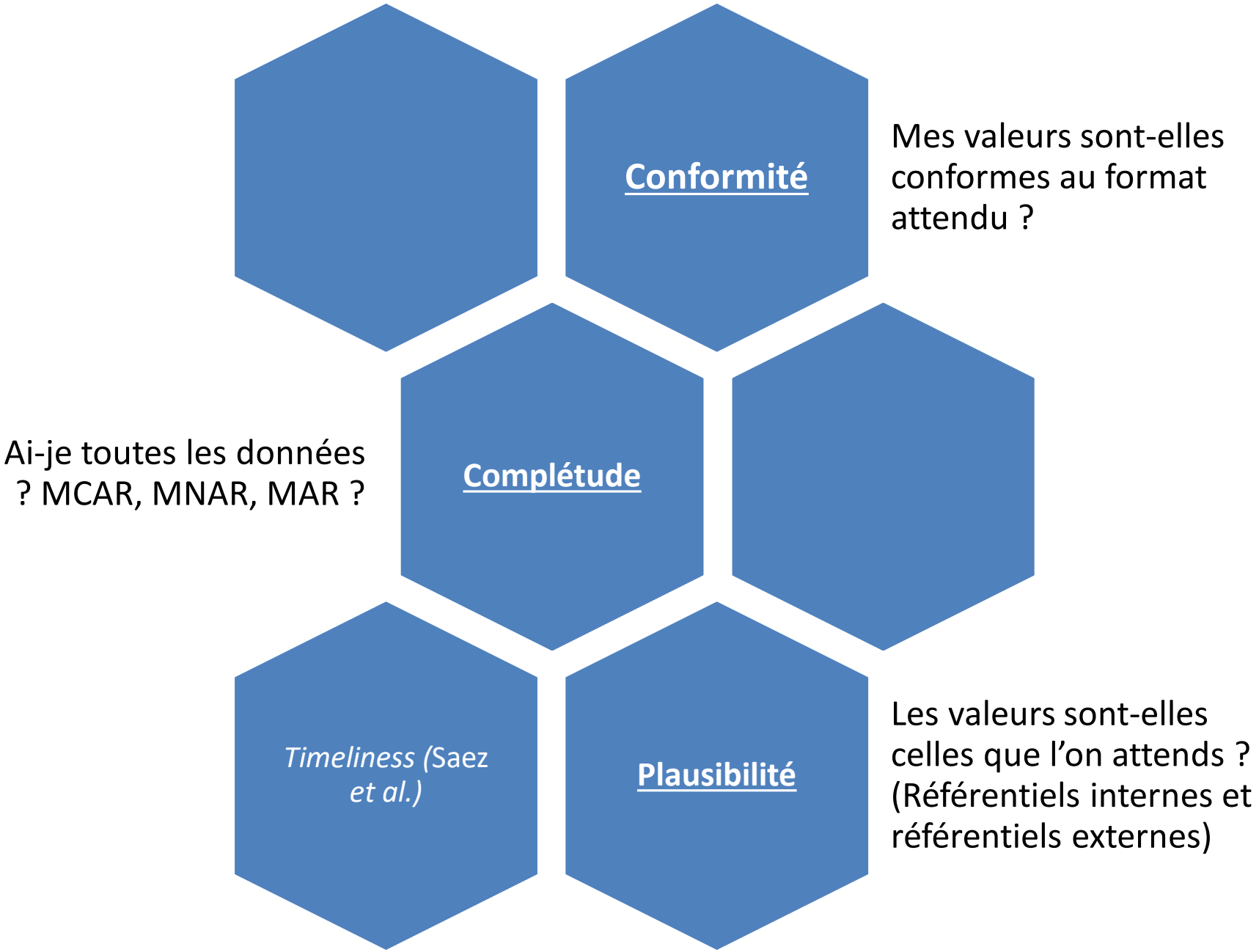
Des événements remarquables



Des événements remarquables



Kahn *et al.*A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data, EGEMs. doi:10.13063/2327-9214.1244.



Etude monosite – données numériques

PARTIE 1 – DONNÉES BIOLOGIQUES

What can Millions of Laboratory Test Results Tell Us about the Temporal Aspect of Data Quality? Study of Data Spanning 17 Years in a Clinical Data Warehouse

V. Looten, L. Kong Win Chang, A. Neuraz, MA. Landau-Loriot, B. Vedie, JL. Paul, L. Mauge, N. Rivet, A. Bonifati, G. Chatellier, A. Burgun, B. Rance

Computer Methods and Programs in Biomedicine, 2018
ISSN 0169-2607, doi:10.1016/j.cmpb.2018.12.030

- **Quelles données ?**

  Données de **biologie**

  Issues de l'entrepôt de données de l'HEGP

  Entre 2000 et 2017

- **Qu'est ce que l'on propose d'explorer ?**

  Décrire leurs évolutions

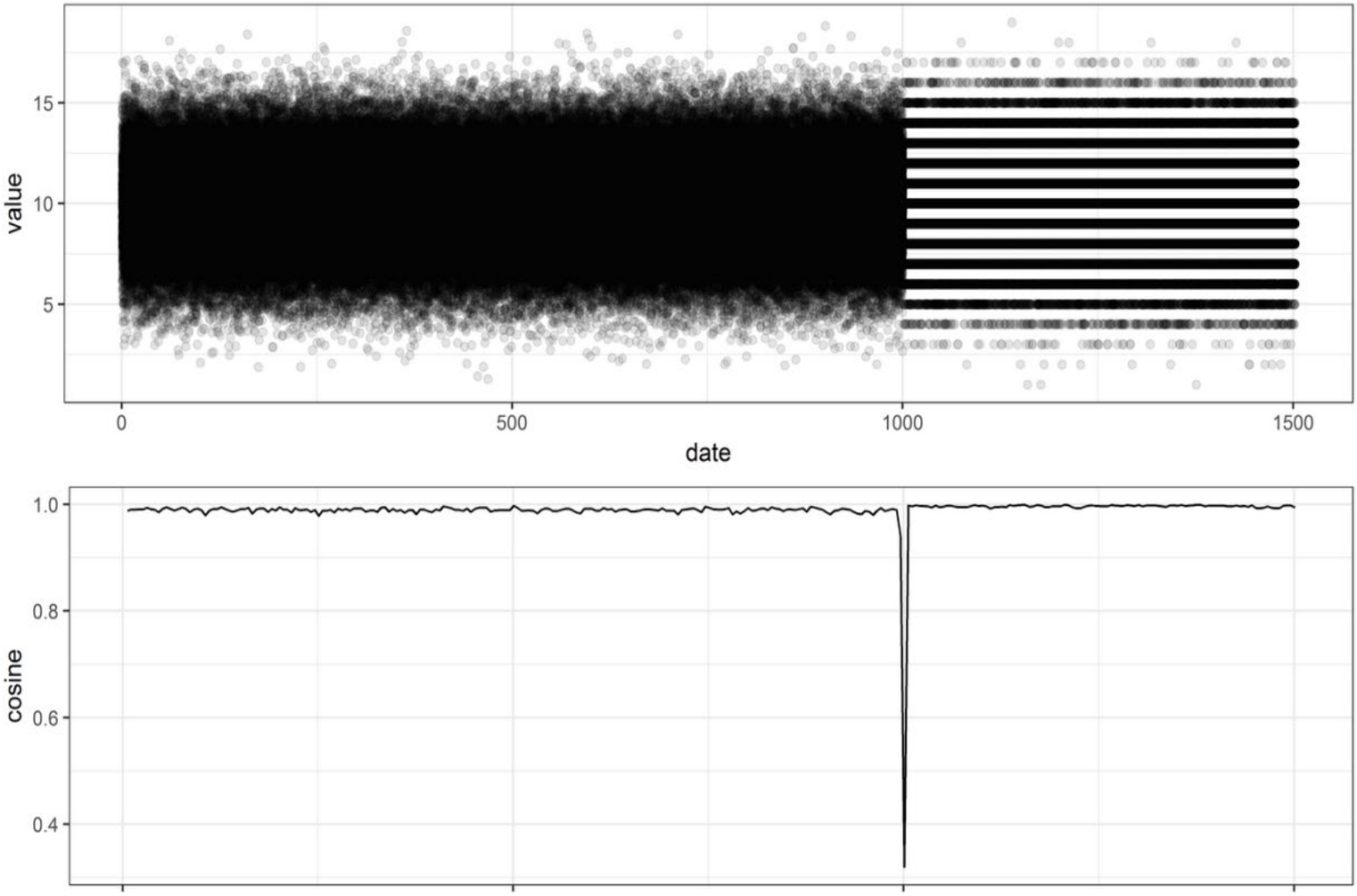
  Chercher des patterns d'évolution

Pattern temporelle et données de biologie : méthodes

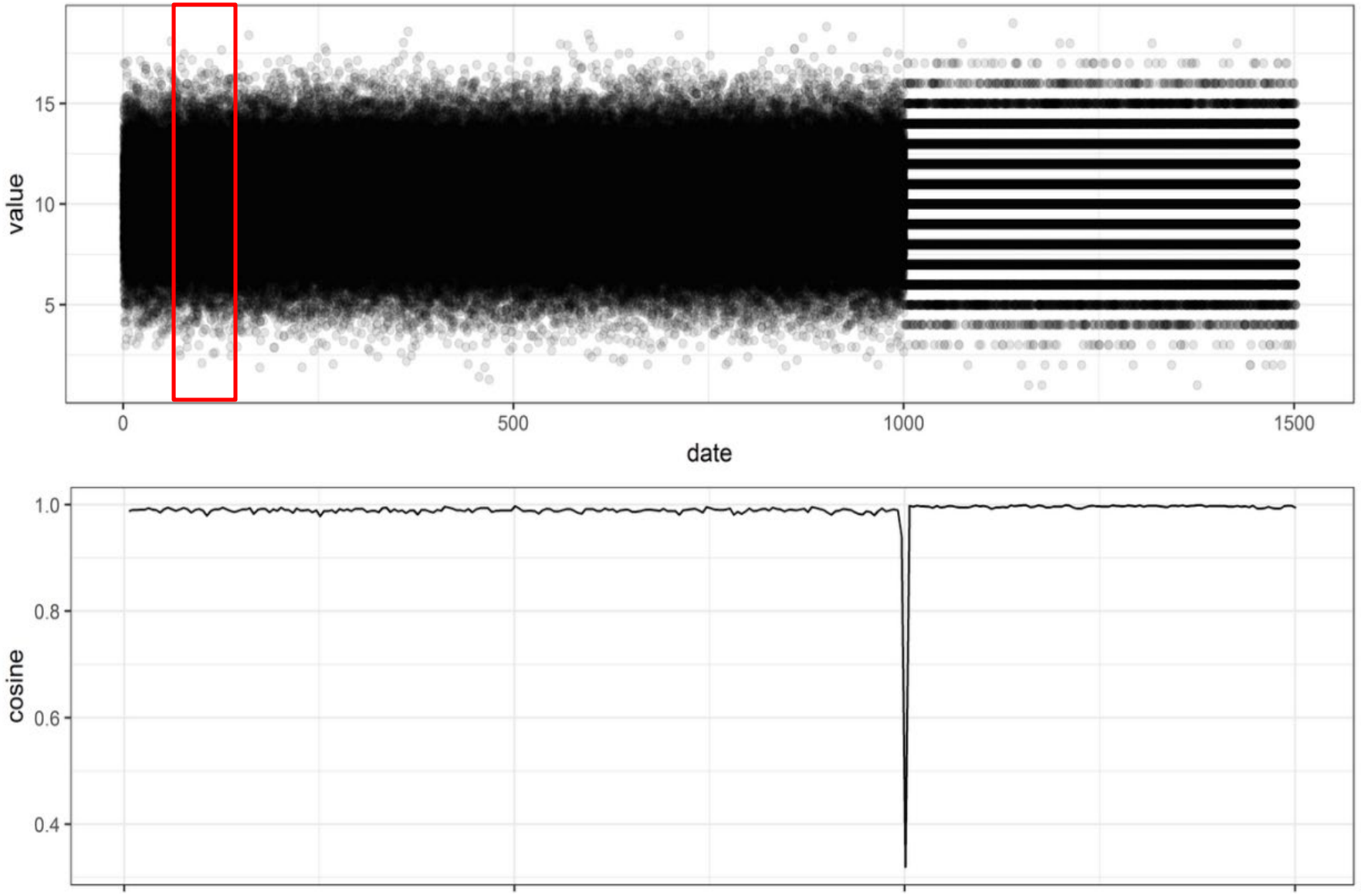
- Détection des phénomènes de discrétisation
  - Calcul des fréquences de chiffres selon leur position
  - Détection via une mesure cosine s'appuyant sur la loi de Benford
- Détection des breakpoints :
  - Pruned Exact Linear Time (PELT) de Killick *et al.*
  - Un algorithme récursif s'appuyant sur une fonction de coût
  - On peut détecter les changements de moyennes ou de variances.
- Détection des tendances
  - Régressions part partie sur la médiane

Killick R, Fearnhead P, Eckley IA. Optimal Detection of Changepoints With a Linear Computational Cost. J
DOI: 10.1080/01621459.2012.737745

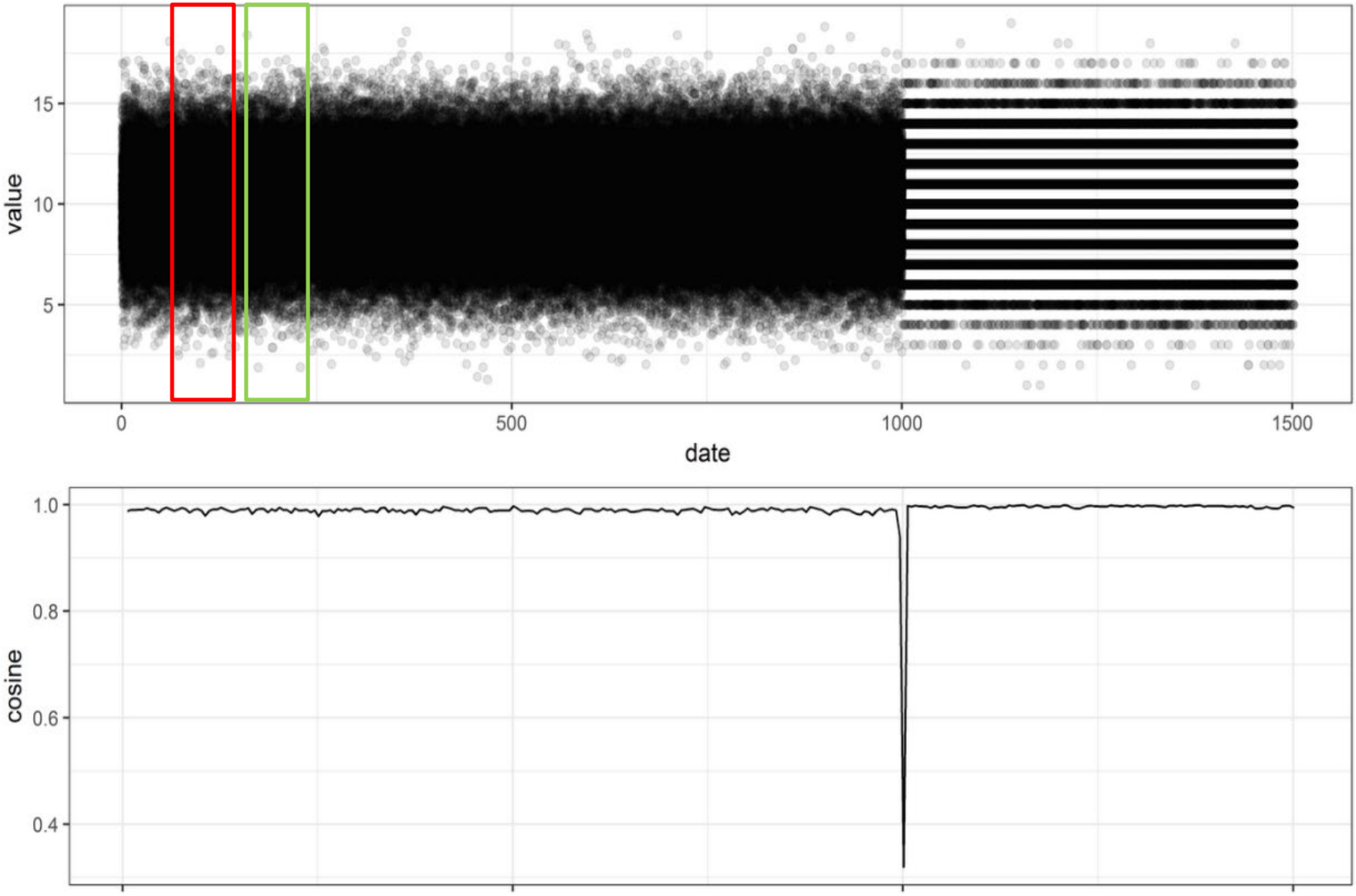
Détection des phénomènes de discrétisation



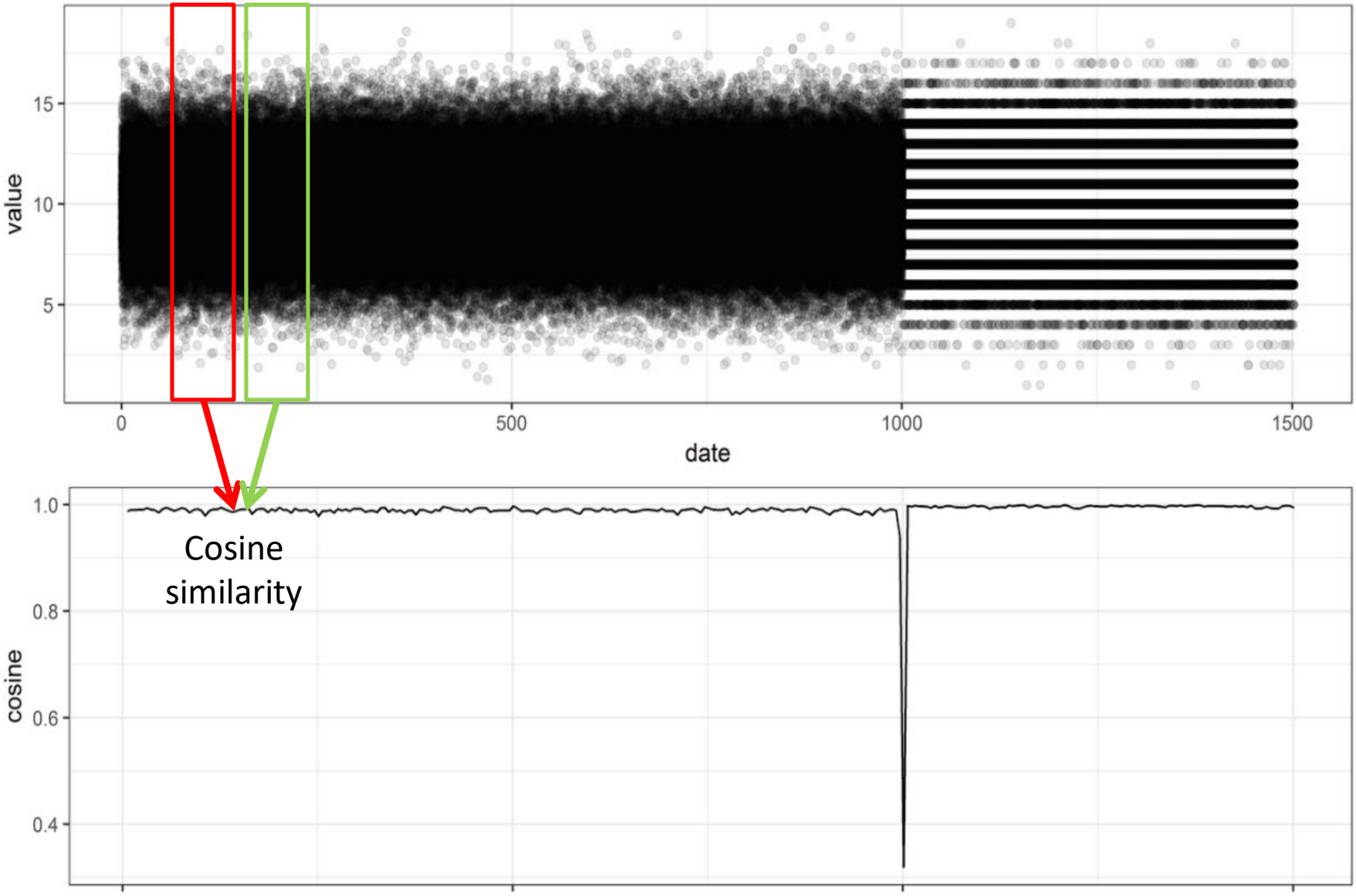
Détection des phénomènes de discrétisation



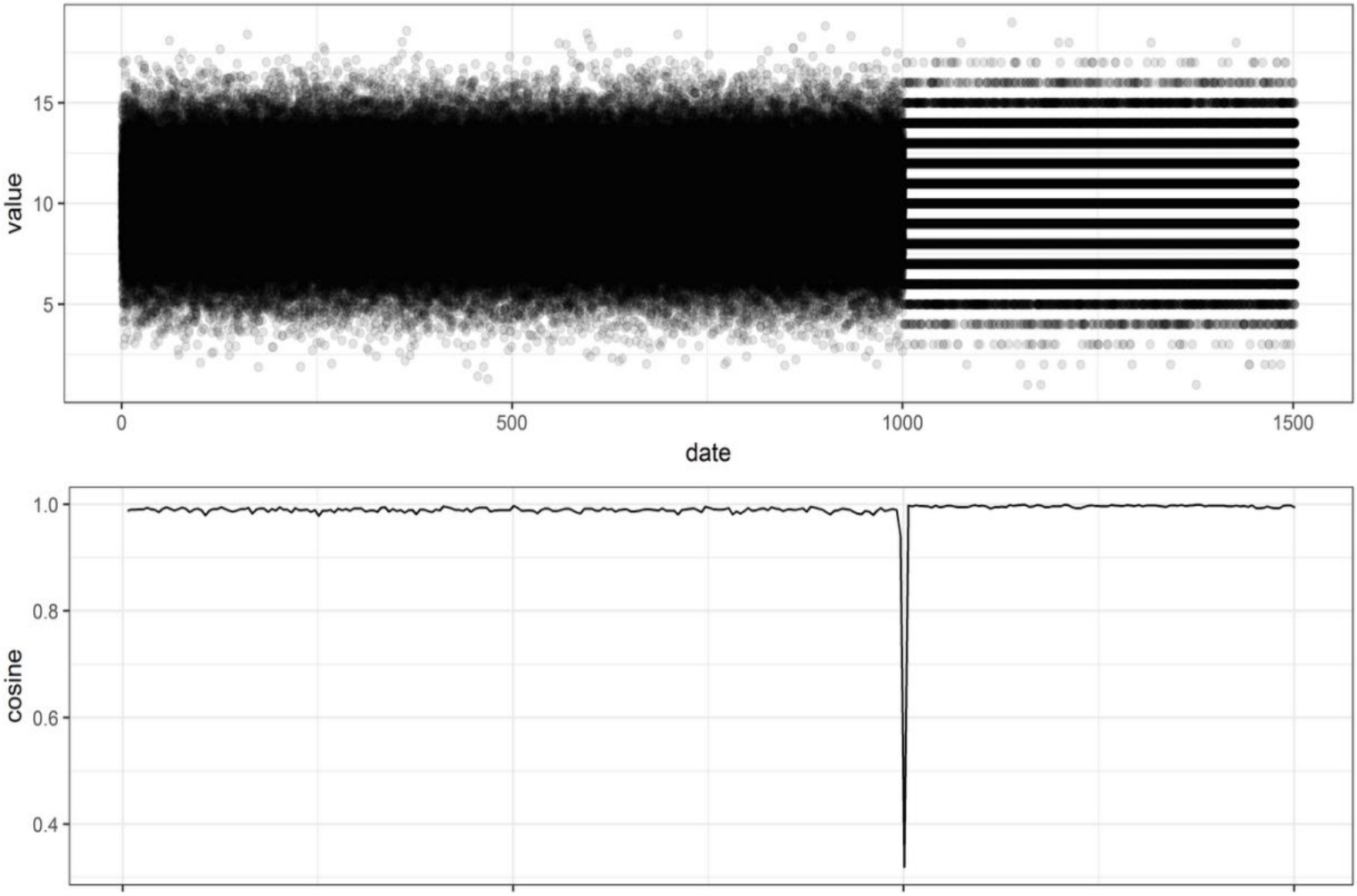
Détection des phénomènes de discrétisation



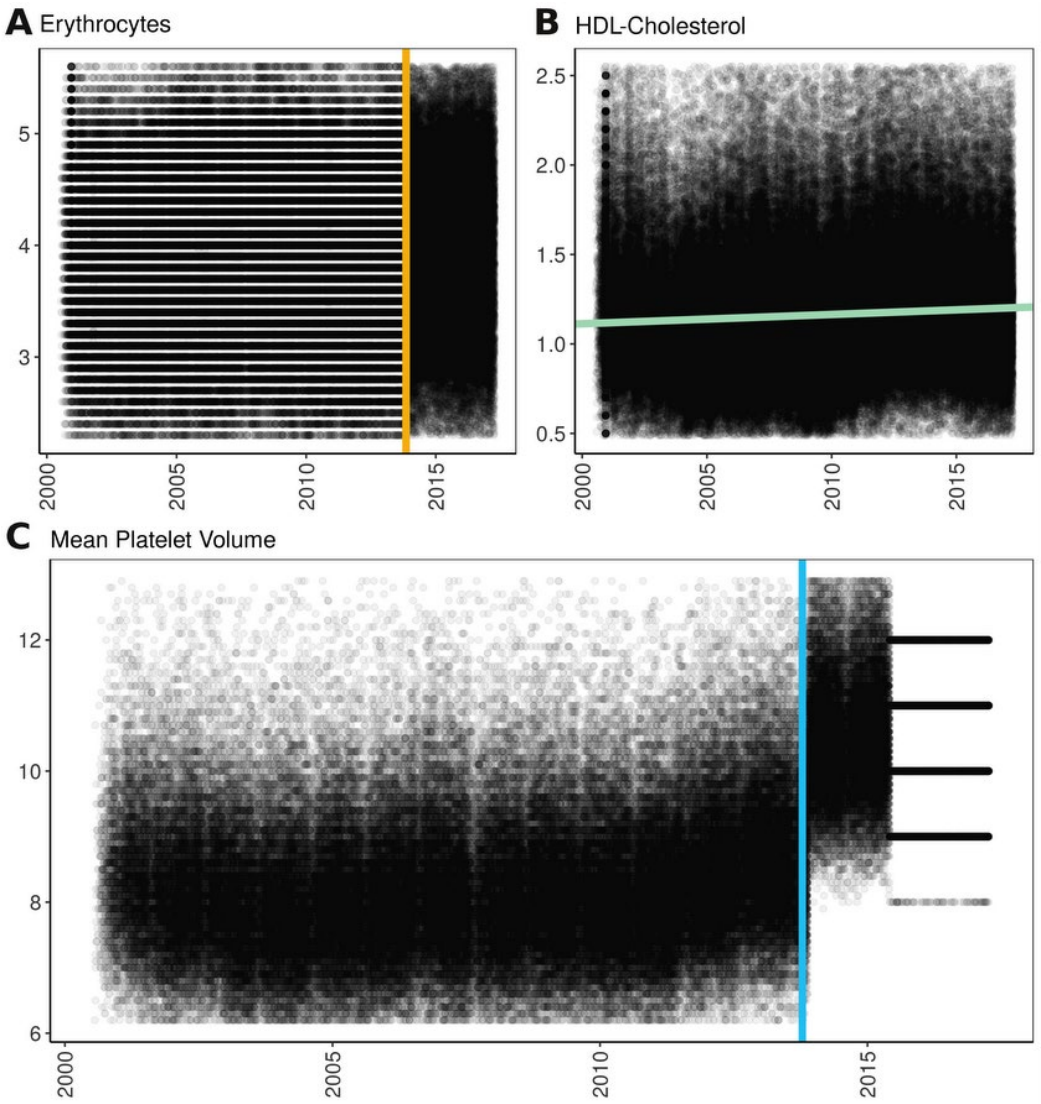
Détection des phénomènes de discrétisation



Détection des phénomènes de discrétisation

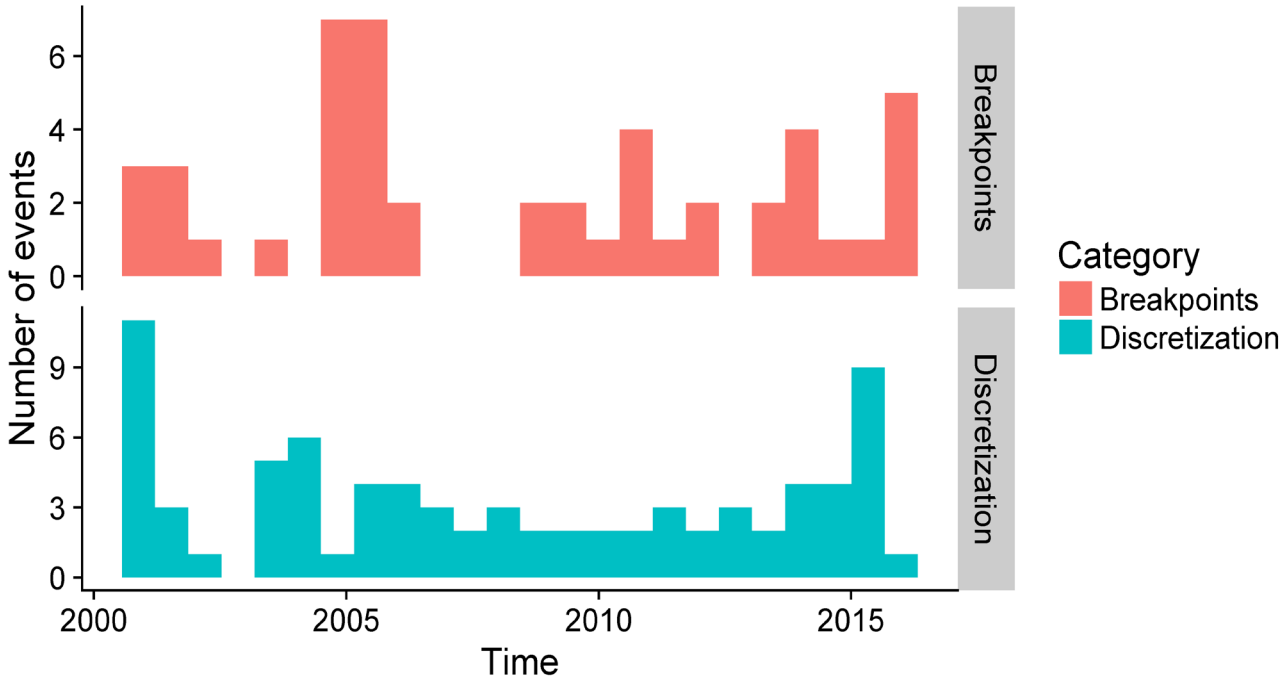


Résultats


A Erythrocytes
5
4
3
2000
2005
2010
2015
B HDL-Cholesterol
2.5
2.0
1.5
1.0
0.5
2000
2005
2010
2015
C Mean Platelet Volume
12
10
8
6
2000
2005
2010
2015

Résultats

Category*	# of biological parameters	Example of laboratory parameter impacted
Discretization	32 (16.7%)	Erythrocytes
Breakpoints	30 (15.6%)	Mean platelet volume
Trends	79 (41.1%)	HDL Cholesterol



Identification des causes ?

- Causes extrinsèques
  - Changement d'automate de mesure
  - Changement d'algorithme de calcul
  - Changement de SGL
  - …
- Causes intrinsèques
  - Evolution de la pratique médicale
  - Evolution de la population
  - …

Conclusion Partie 1

- Approche pragmatique pour détecter des patterns d'évolution
  - Approche visuelle
  - Algorithmes de détection dédiés
- Détectable sur un site unique
  - Importance de la maitrise de l'historique
- La qualité longitudinale concerne-t-elle d'autres types de donnée ?

Etude nationale – données catégorielles

PARTIE 2 - DONNÉES ADMINISTRATIVES

Exploration of the Timeliness of ICD codes in Administrative Databases

Marie SIMON, Bastien RANCE, Sandrine KATSAHIAN, Karim BOUNEBACHE, Grégoire REY, Gilles CHATELLIER, Antoine NEURAZ, Anita BURGUN, Vincent LOOTEN

Quelles données ?

Codes CIM tronqués à 3 caractères

Issues des RUM du PMSI national

Entre 2008 et 2017

Qu'est ce que l'on propose d'explorer ?

La **timeliness** (stabilité temporelle)

Etudier les variations et essayer de les classer

Exploration of the Timeliness of ICD codes in Administrative Databases

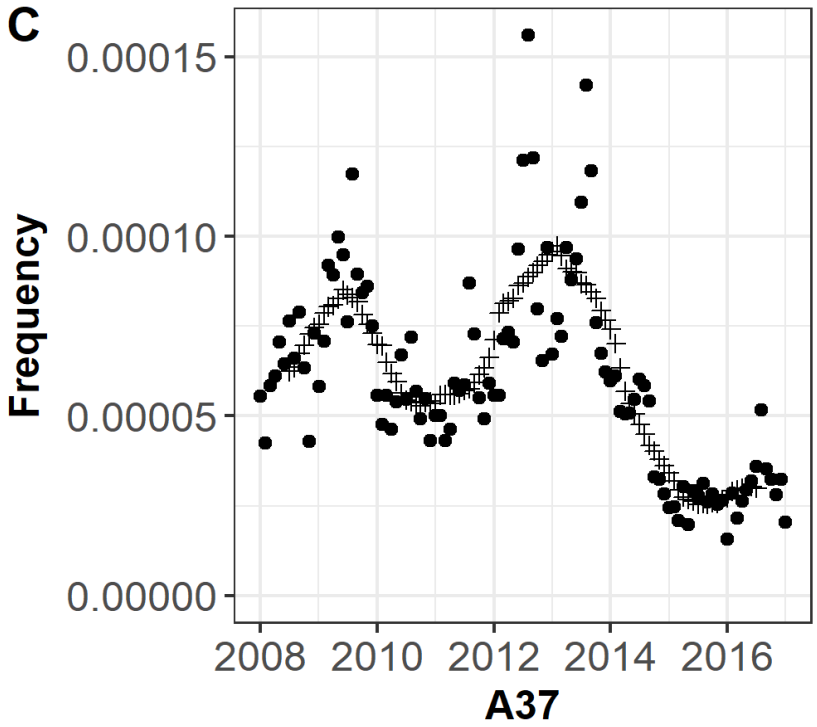
Comment ?

- Fréquences mensuelles des codes rapportées au nombre de patients distincts
- On propose également un lissage

$$Freq_{ICD\ code}(t) = \frac{\text{distinct number of patients with the code } at\ the\ time\ t}{\text{distinct number of patients with any code } at\ the\ time\ t}$$

Est-ce que ça marche ?

- Exemple de la coqueluche



Exploration of the Timeliness of ICD codes in Administrative Databases

$$\text{Relative Amplitude}_{\text{ICD code}} = \frac{|\text{min relative frequency code} - \text{max relative frequency code}|}{\text{Mean of min and max relative frequencycode}}$$

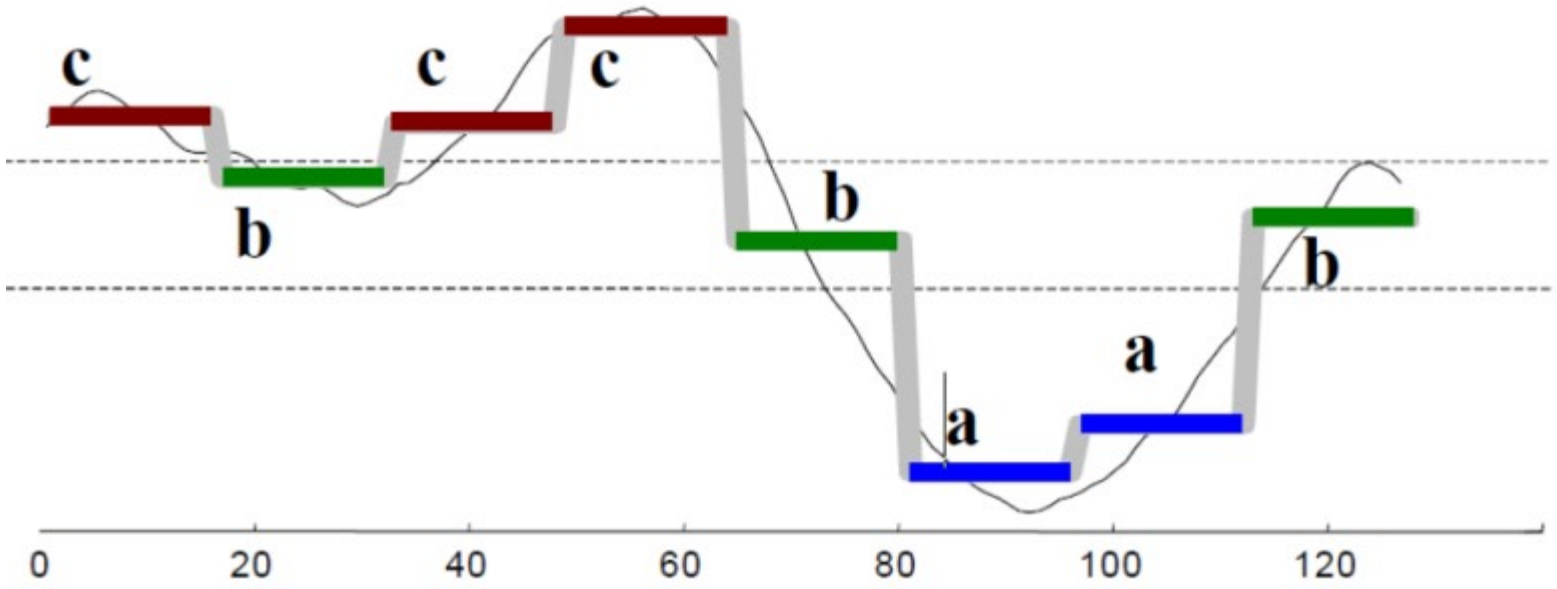
- Variation d'amplitude concernant les codes CIM
  - Inférieure à 50% pour 1 006 codes
  - Entre 50 et 100% pour 510 codes
  - Supérieure à 100% pour 521
- Parmi ces 2037 codes, 1758 sont régulièrement utilisés sur toute la période d'étude

Clustering temporelle

- Enjeu : explorer les « formes » des principales variations
- On choisit de réaliser le clustering sur les données lissées ce qui exclut les atypies.
- Une approche *model-free*:
  - Symbolic Aggregate approXimation (SAX) (Lin *et al.*)
  - On normalise les courbes
  - On découpe les sequences en fenêtres de tailles fixe
  - On attribute à chaque “cadran” une etiquette constituent ainsi un alphabet
  - Ce qui permet d’attribuer à chaque courbe un vecteur

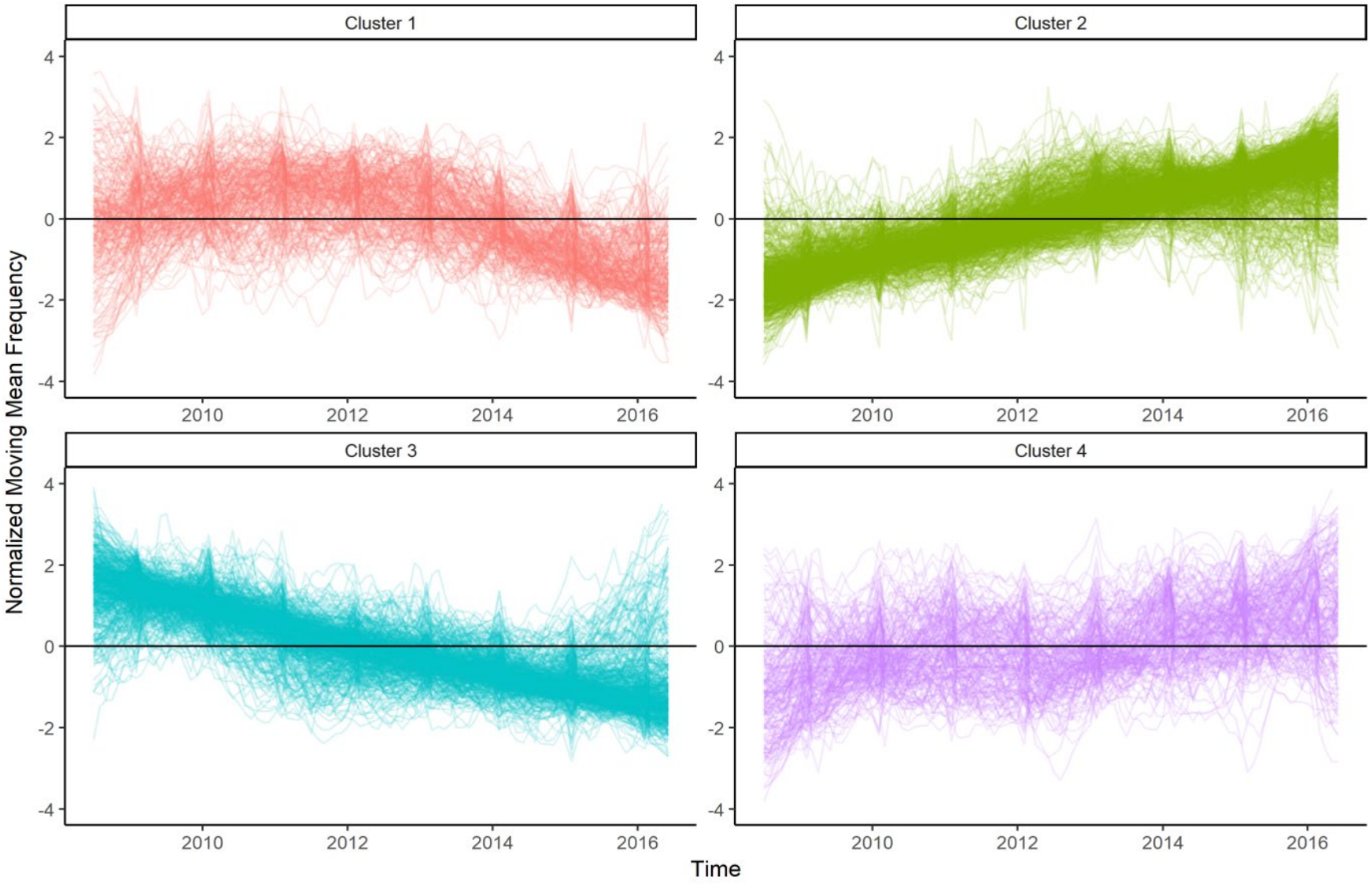
Lin *et al.*, A symbolic representation of time series, with implications for streaming algorithms. doi:10.1145/882082.882086

Symbolic Aggregate approXimation (SAX)

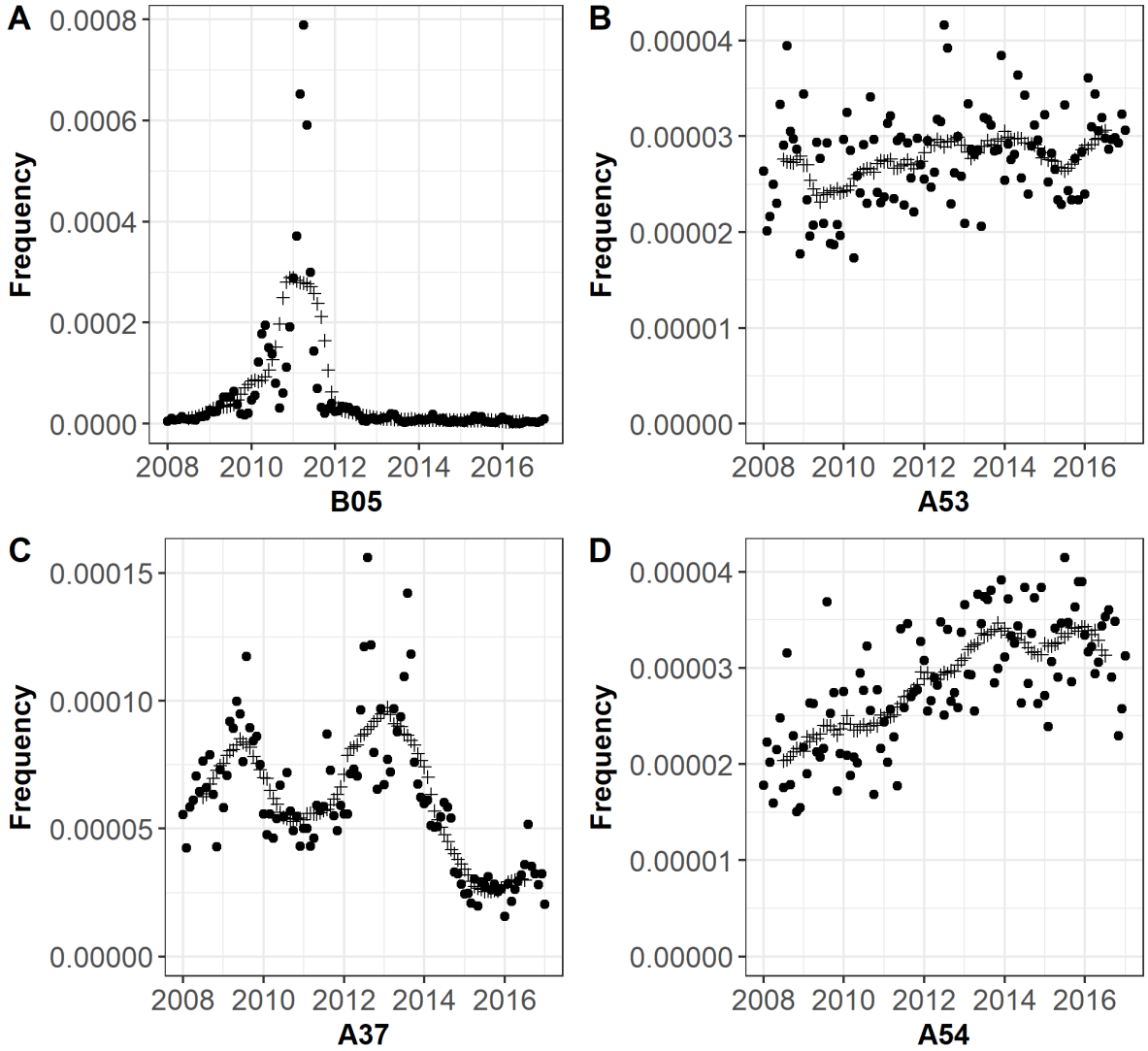


From: The use of Bioinformatics Techniques for Time-Series Motif-Matching: A Case Study. ADVCOMP 2012

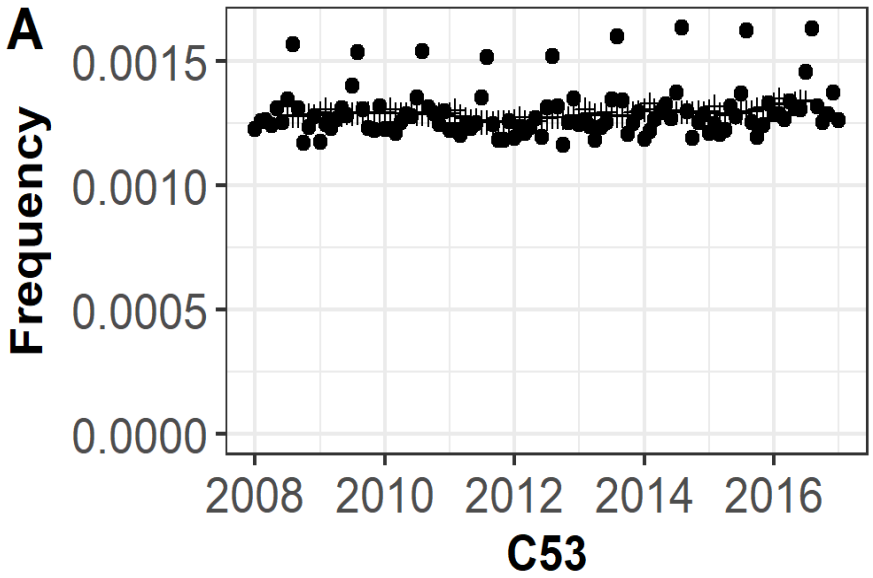
Clustering temporelle : résultats

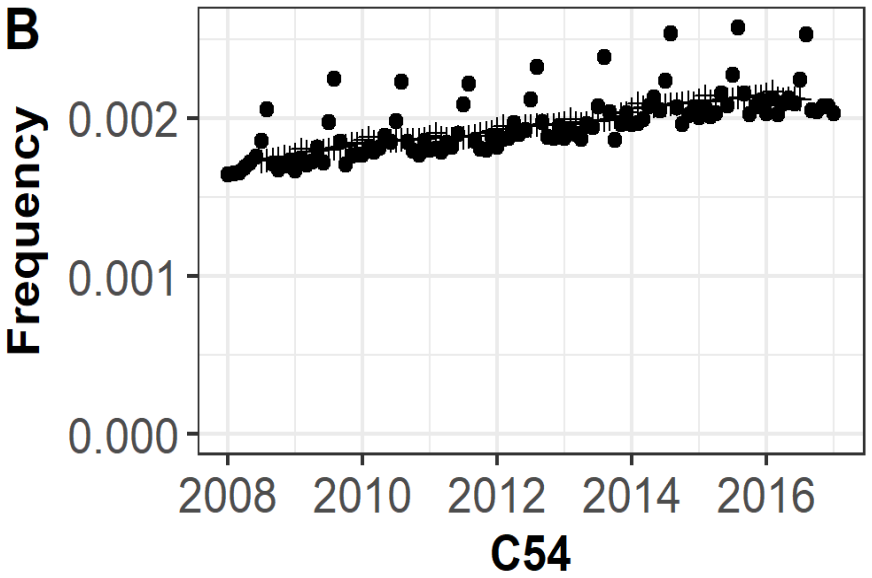


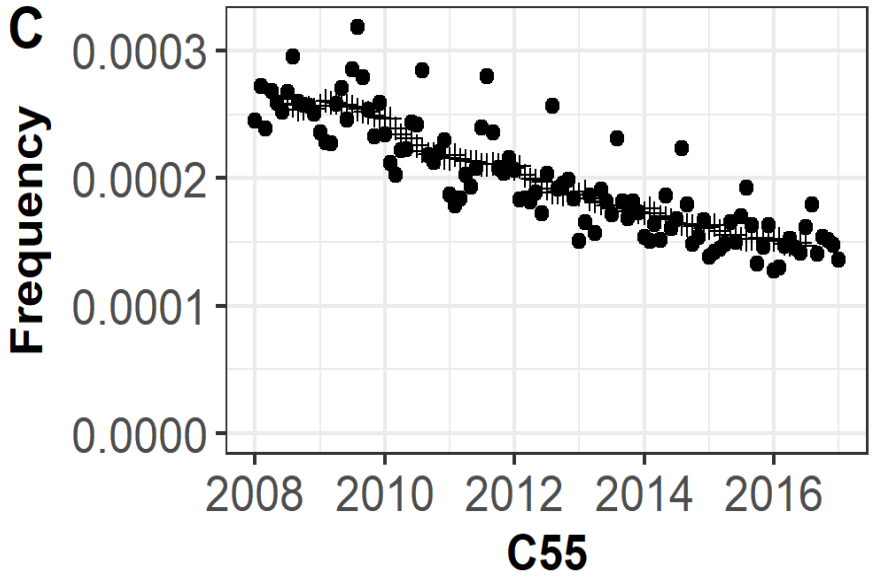
Des facteurs intrinsèques …



Des facteurs extrinsèques …







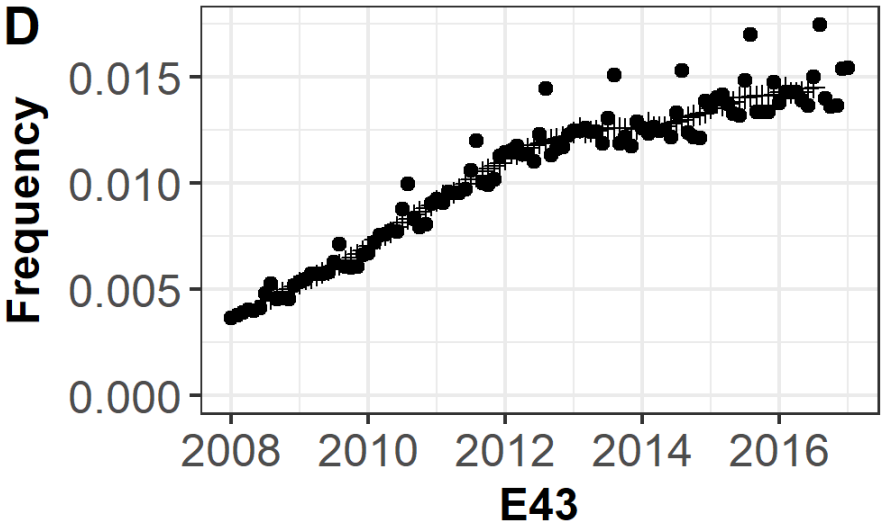
C53 : Tumeur maligne du col de l'utérus
C54 : Tumeur maligne du corps de l'utérus
C55 : Tumeur maligne de l'utérus, partie non précisée

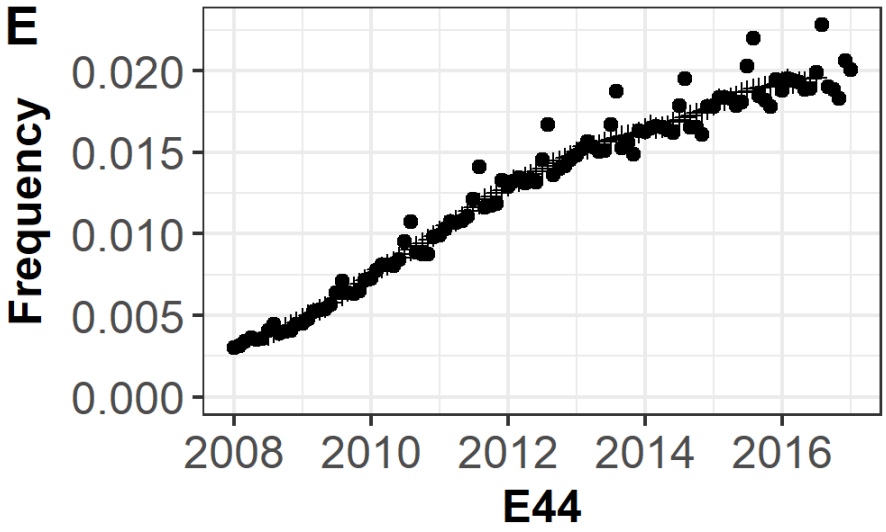
E43: Malnutrition protéino-énergétique grave, sans précision (Sev. 3)

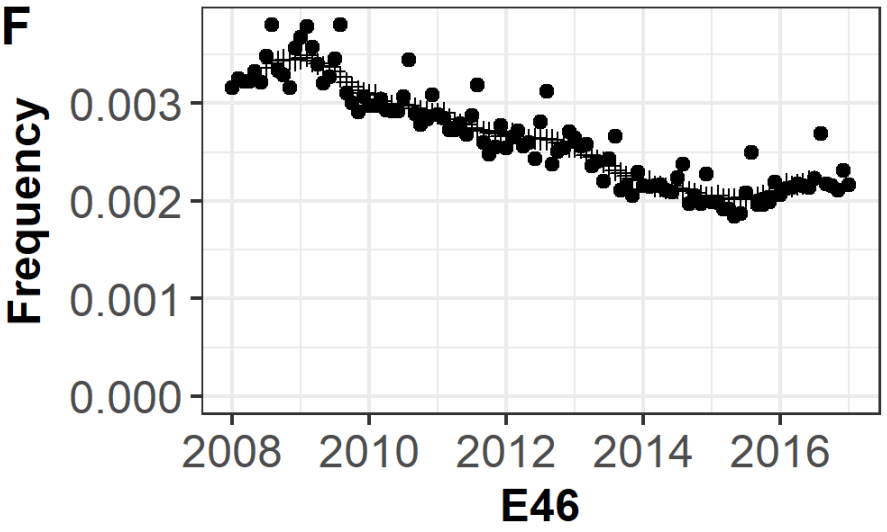
E44 : Malnutrition protéino-énergétique légère ou modérée

(dont E440 (Sev. 3) et E442 (Sev.2))

E46 : Malnutrition protéino-énergétique, sans précision (Sev. 2)





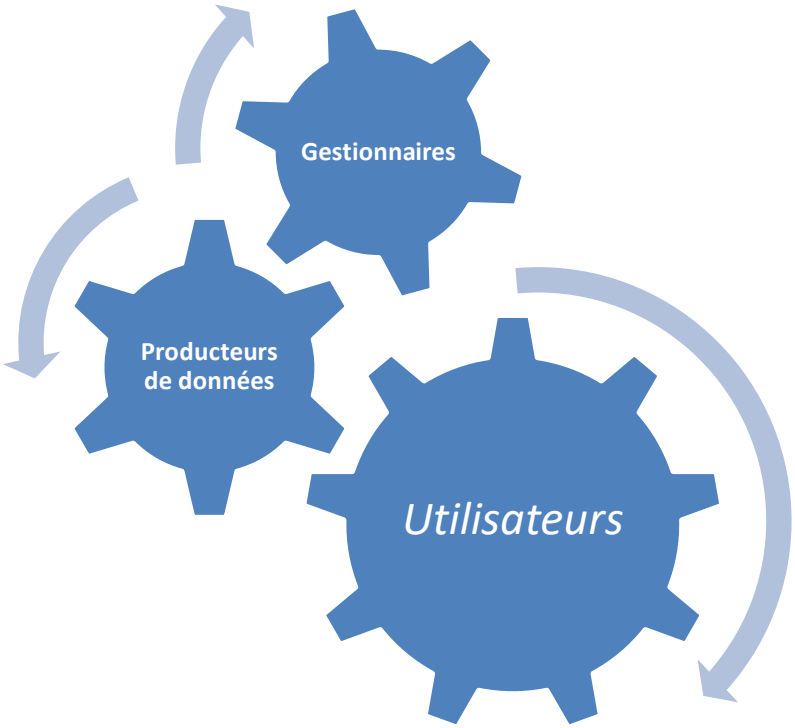


Enjeux sur la réutilisation des données : besoin d'annotations

Facteurs externes	Facteurs internes (épidémiologie)
• Pratiques de codage et facturation • Organisation des soins : nationale, régionale, locale	• Epidémie • Evolution des traitements • Phénomènes épidémiologiques globaux vs. Phénomènes locaux

Quelles conclusions sur l'exploration de la qualité des données dans les entrepôts

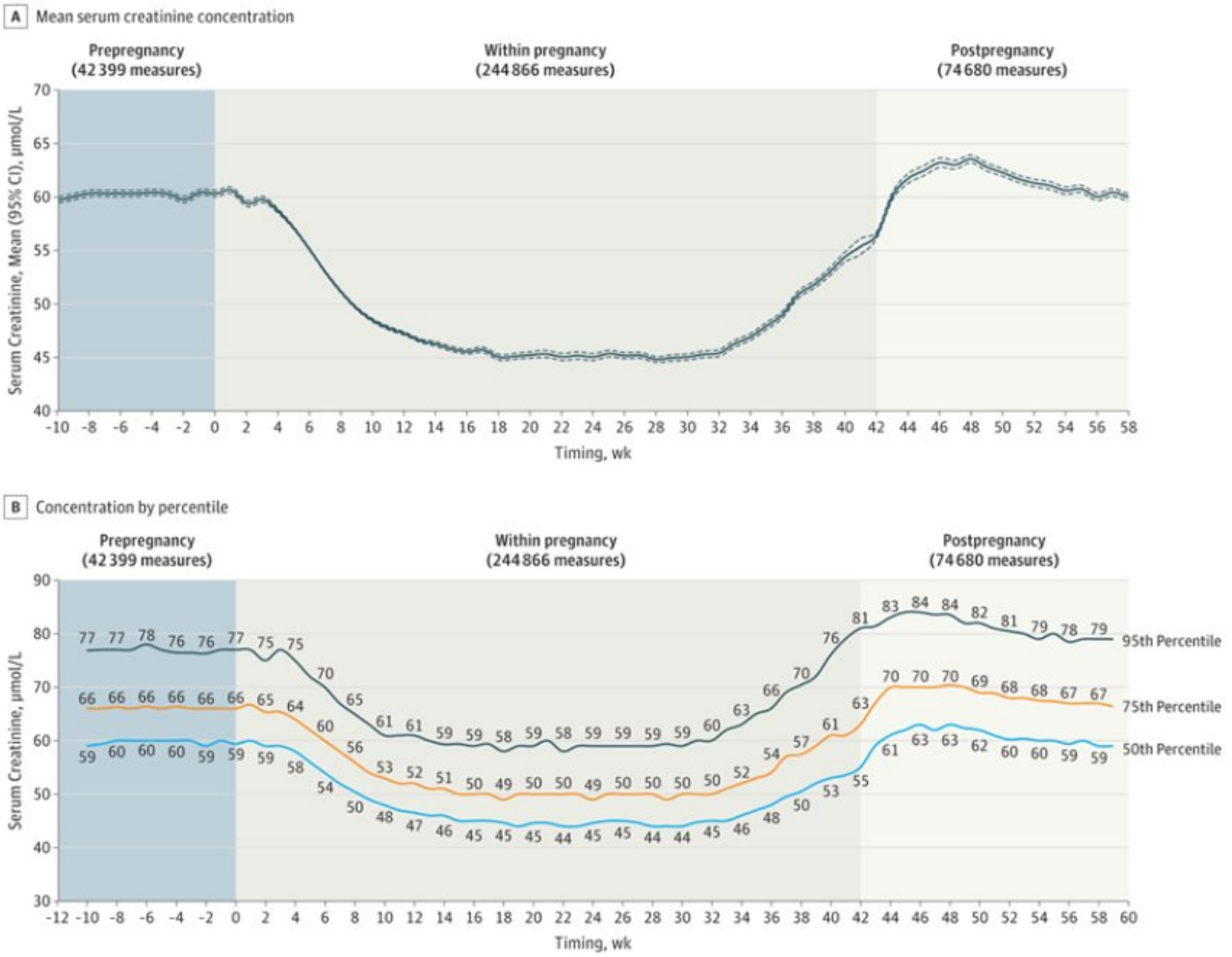
- Proximité entre les données, les producteurs de données et les utilisateurs



- Une quantité de données importante et des méthodes qui se cherchent

Qu’est ce qu’une valeur normale ?

Serum Creatinine Levels Before, During, and After Pregnancy.
Harel Z, McArthur E, Hladunewich M, et al. *JAMA.* 2019;321(2):205–207. doi:10.1001/jama.2018.17948



Remerciements

HEGP - Necker
Informatique médicale
- **V. Looten**
- A. Neuraz
- M. Simon
- S. Katsahian
- G. Chatellier
- A. Burgun

Université Lyon 1
- **L. Kong Win Chang**
- A. Bonifati

HEGP
Département de Biologie
- MA. Landau-Loriot
- B. Vedie
- JL. Paul
- L. Mauge
- N. Rivet

CépiDC
- K. Bounebache
- G. Rey