PURCHASE SIGNATURES OF RETAIL CUSTOMERS

DECADE 11/10/2017

Clément Gautrais, P. Cellier, R. Quiniou, T. Guyet, A. Termier

Lacodam/SemLIS team

Motivations

- Retailers have a lot of data on customers purchases
- Detecting individual customer habits is crucial
  - Personalized marketing
  - Attrition detection/characterization
- Challenges
  - Customers are not perfectly regular
  - Dataset size (~300 GB)

Motivations

- How often does a customer replenish his/her products?
  - Give coupon on the right product at the right time
  - Strong attrition signal on favourite products
- Find the favourite products of a customer
- Find the replenishment period

Existing methods

- Pattern mining methods
  - Top-k [6]
  - Periodic pattern [5]
  - Frequent itemsets [1]
  - Episode mining [9]

- Item recommendation methods

- Drawbacks
  - Many results
  - Regularity definition too strict or too loose
  - Products have to be bought in the same transaction
  - Non interpretable models

Proposed model: signatures

- Find favourite products of a customer
  - Bought several times
  - Not necessarily in the same transaction

- Find recurrent symbols and their occurrences in a symbolic sequence, with no predefined period
  - A set of product and its occurrences as results
  - Period adapts to the sequence rhythm

Signature model – Sequence segmentation

- k-segmentation [8]: split a sequence of n transactions into k segments

Timestamp	Receipts
1	Bread, Milk, Orange Juice, Soup
2	Butter, Apple, Soup, Orange Juice
3	Bread, Sponge
4	Bread, Butter, Soup
5	Orange Juice, Eggs
6	Bread, Milk, Eggs

Timestamp	Receipts
1	Bread, Milk, Orange Juice, Soup
2	Butter, Apple, Soup, Orange Juice
3	Bread, Sponge
4	Bread, Butter, Soup
5	Orange Juice, Eggs
6	Bread, Milk, Eggs

A 3-segmentation of a customer purchase sequence

Signature model – Sequence segmentation

- Segment representative: $\mu(S_i) = \bigvee_{t \in S_i} t$

	Timestamp	Receipts
S1	1	Bread, Milk, Orange Juice, Soup
	2	Butter, Apple, Soup, Orange Juice
S2	3	Bread, Sponge
	4	Bread, Butter, Soup
S3	5	Orange Juice, Eggs
	6	Bread, Milk, Eggs

Segment index	Segment representatives $\mu(S_i)$
1	Bread, Milk, Orange Juice, Soup, Butter, Apple
2	Bread, Butter, Soup, Sponge
3	Bread, Orange Juice, Eggs, Milk

Signature model – Sequence segmentation

- Adequation: $A(\alpha, S) = \left|\bigwedge_{S_i \in S} \mu(S_i)\right|$

Segment index	Segment representatives $\mu(S_i)$
1	Bread, Milk, Orange Juice, Soup, Butter, Apple
2	Bread, Butter, Soup, Sponge
3	Bread, Orange Juice, Eggs, Milk

- $A(\alpha, S) = \left|\bigwedge_{S_i \in S} \mu(S_i)\right| = |\{Bread, Milk, Orange\ Juice, Soup, Butter, Apple\} \cap \{Bread, Butter, Soup, Sponge\} \cap \{Bread, Orange\ Juice, Egges, Milk\} = |\{Bread\}| = 1$

Segment index	Segment representatives $\mu(S_i)$
1	**Bread**, Milk, Orange Juice, Soup, Butter, Apple
2	**Bread**, Butter, Soup, Sponge
3	**Bread**, Orange Juice, Eggs, Milk

Signature model – Sequence segmentation

- $S_{opt}(\alpha, k) = arg \max_{S \in \mathcal{S}_{n,k}} A(\alpha, S)$

Timestamp	Receipts
1	Bread, Milk, Orange Juice, Soup
2	Butter, Apple, Soup, Orange Juice
3	Bread, Sponge
4	Bread, Butter, Soup
5	Orange Juice, Eggs
6	Bread, Milk, Eggs

$+ \qquad k = 3$

- Solve $S_{opt}(\alpha, k)$

Signature model – Sequence segmentation

	Timestamp	Receipts
S1	1	Bread, Milk, Orange Juice, Soup
	2	Butter, Apple, Soup, Orange Juice
S2	3	Bread, Sponge
	4	Bread, Butter, Soup
S3	5	Orange Juice, Eggs
	6	Bread, Milk, Eggs

Segment index	Segment representatives $\mu(S_i)$
1	**Bread**, Milk, Orange Juice, Soup, Butter, Apple
2	**Bread**, Butter, Soup, Sponge
3	**Bread**, Orange Juice, Eggs, Milk

$$A(\alpha, S) = |\{Bread\}| = 1$$

Signature model – Sequence segmentation

	Timestamp	Receipts
S1	1	Bread, Milk, Orange Juice, Soup
S2	2	Butter, Apple, Soup, Orange Juice
	3	Bread, Sponge
	4	Bread, Butter, Soup
S3	5	Orange Juice, Eggs
	6	Bread, Milk, Eggs

Segment index	Segment representatives $\mu(S_i)$
1	**Bread**, Milk, **Orange Juice**, Soup
2	**Bread**, Apple, Sponge, **Orange Juice**, Butter, Soup
3	**Bread**, **Orange Juice**, Eggs, Milk

$$A(\alpha, S) = |\{Bread, Orange\ Juice\}| = 2$$

Signature model – Sequence segmentation

	Timestamp	Receipts
S1	1	Bread, Milk, Orange Juice, Soup
S2	2	Butter, Apple, Soup, Orange Juice
	3	Bread, Sponge
S3	4	Bread, Butter, Soup
	5	Orange Juice, Eggs
	6	Bread, Milk, Eggs

Segment index	Segment representatives $\mu(S_i)$
1	**Bread**, Milk, **Orange Juice**, **Soup**
2	**Bread**, Apple, Sponge, **Orange Juice**, Butter, **Soup**
3	**Bread**, **Orange Juice**, Eggs, Milk, **Soup**

$$A(\alpha,S) = |\{Bread, Orange\ Juice, Soup\}| = 3 = arg \max_{S\in\mathcal{S}_{6,3}} A(\alpha,S)$$

Signature model – Sequence segmentation

- Mining algorithms: exact approaches
  - Dynamic programming $O(n^2k)$
  - Pattern growth $O(2^{|I|})$

- Mining algorithms: other approaches
  - Greedy algorithms $O(n * log(n))$
  - Non exact algorithms with bounded error $O(n^{\frac{4}{3}}k^{\frac{5}{3}})$

Signature model – Sequence segmentation

- T_i is a boolean vector
  - (p_1, p_2)=(1,1,0,0) with 4 products
- $\mu(S_i) = \bigvee_{t \in S_i} t$
- $A(\alpha, S) = \left| \bigwedge_{S_i \in S} \mu(S_i) \right|$
- $S_{opt}(\alpha, k) = arg \max\limits_{S \in \mathcal{S}_{n,k}} A(\alpha, S)$ →optimized with dynamic programming [8]

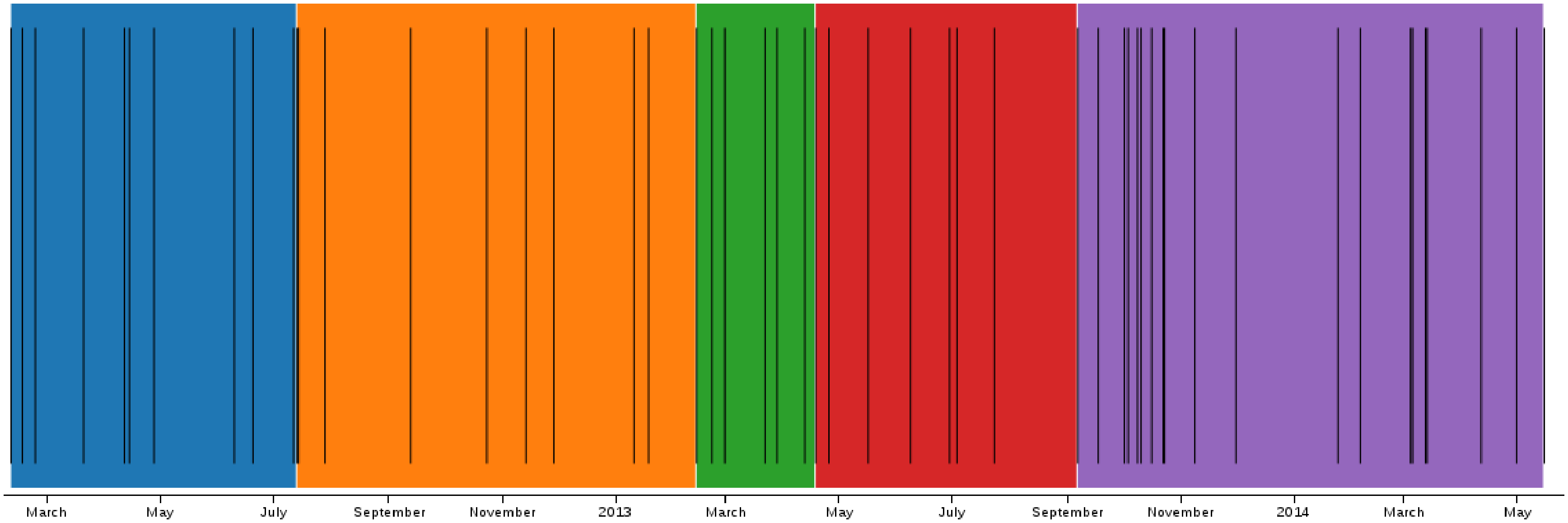
Timestamp	Receipts
1	Bread, Milk, Orange Juice, Soup
2	Butter, Apple, Soup, Orange Juice
3	Bread, Sponge
4	Bread, Butter, Soup
5	Orange Juice, Eggs
6	Bread, Milk, Eggs

$+\ \{Bread, Orange\ Juice, Soup\}$

Signature - example

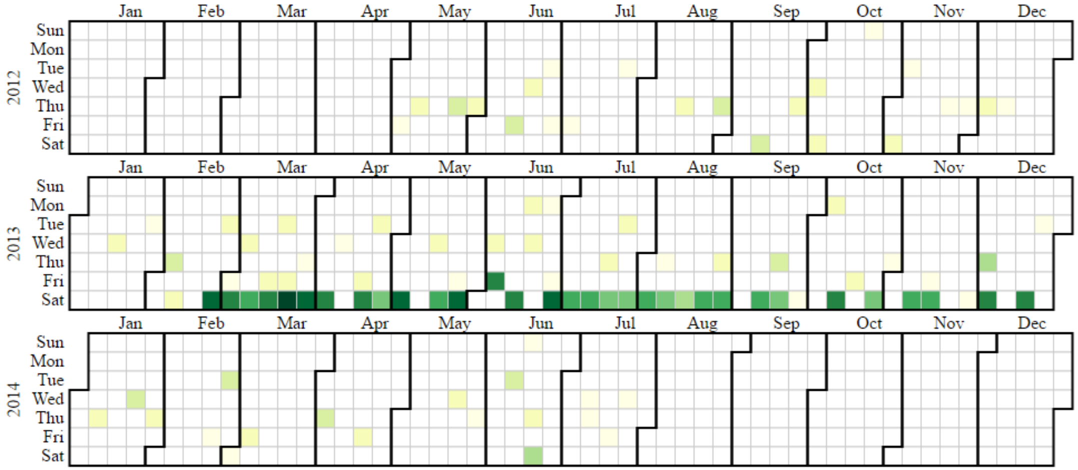
- JOKER MULTIFRUIT BRK OVALINE1L
- SIROP SPORT CITROR BTL 1L
- BRETS CHIPS POULET BRAISE 6X25
- RANOU ROTI PORC 6TR 240G
- MINI BABYBEL X12 264G
- IDS CREME CASSIS 20D 70CL
- MT BLANC VANILLE MINI 6X125G
- J.ROZE S.HACHE LETENDR X10 1K
- 1ER PRIX BEURRE 1/2S PQ 500G
- ECR/AD COLOSSE CHOC.BLC4X120
- RANOU ROTI DE PORC 4TR 160G
- PASQUIER BISCOTTE MINC.36T 300
- RANOU JBON MON PARIS DD6T270G
- KINDER PINGUI CHOCOLAT 8X30G
- PASQUIER 12 CROISSANTS 480G

Customer from a dataset of 149 942 customers of a French retailer

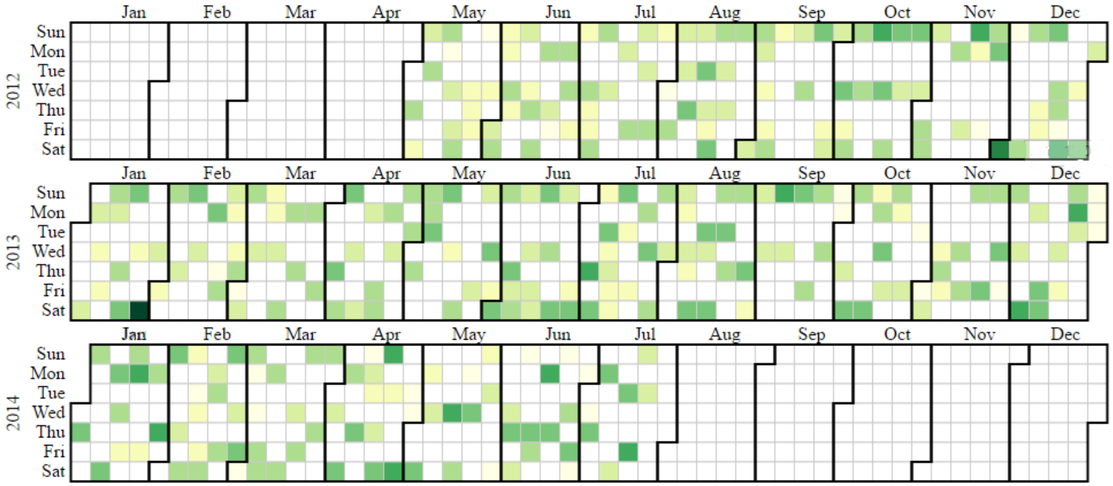


Signature advantages

- Find regularities in seemingly no regular data
- No window size
- Simple output



Periodic works, signature works



Periodic does not work, signature works

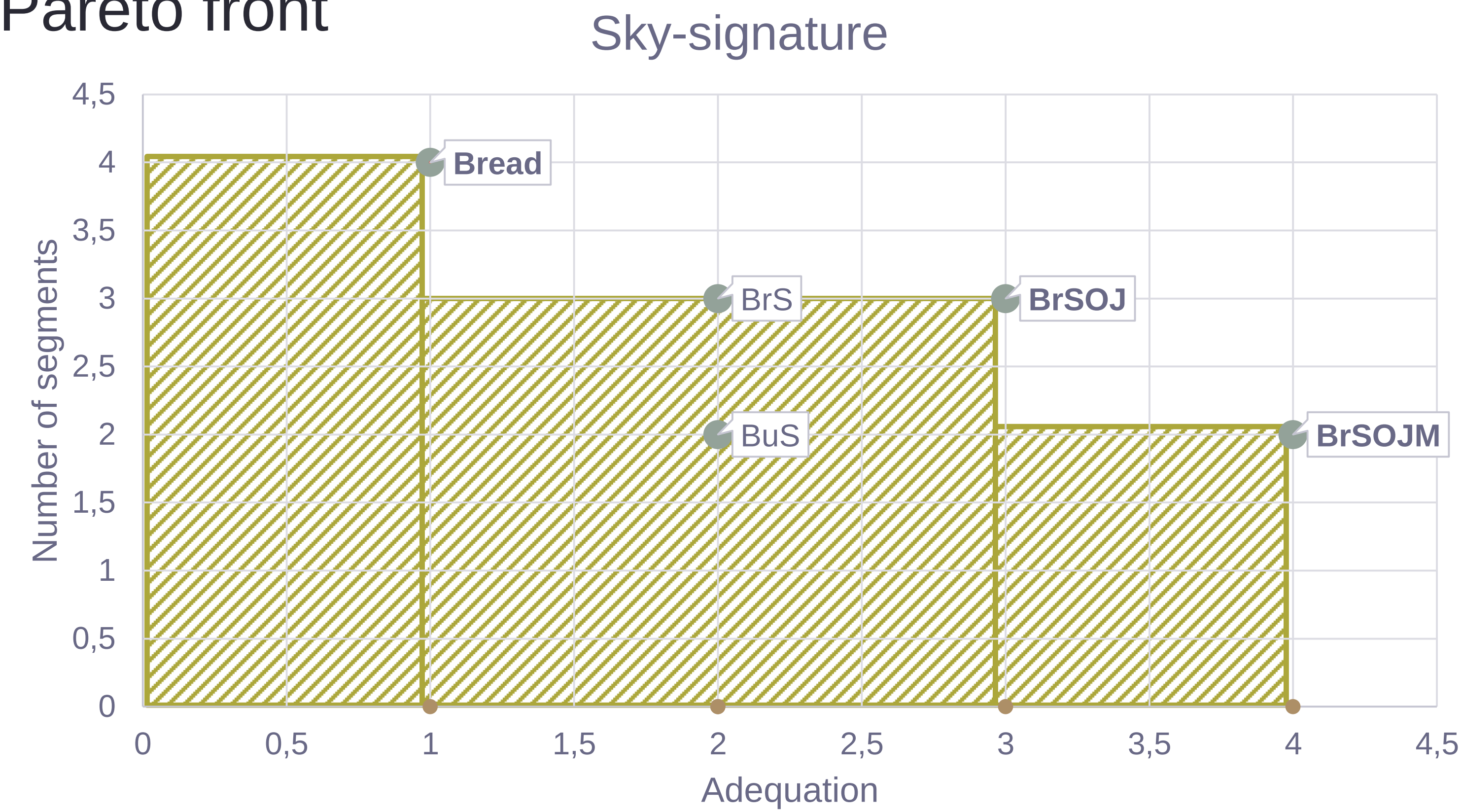
Sky-signature

- Extension of the signature model
  - How to choose the right number of repetitions?
  - Don’t choose, try them all
    - Too many results
    - Pattern selection with a skyline [7]

Timestamp	Receipts
1	Bread, Milk, Orange Juice, Soup
2	Butter, Apple, Soup, Orange Juice
3	Bread, Sponge
4	Bread, Butter, Soup
5	Orange Juice, Eggs
6	Bread, Milk, Eggs

Sky-signature

- Sky-signature
  - Compromise between adequation and number of segments
- = Pareto front



Sky-signature

- Algorithm based on dynamic programming
  - Similar to the sequence segmentation
  - Same complexities as classic signature with $k = \max\_freq(I)$

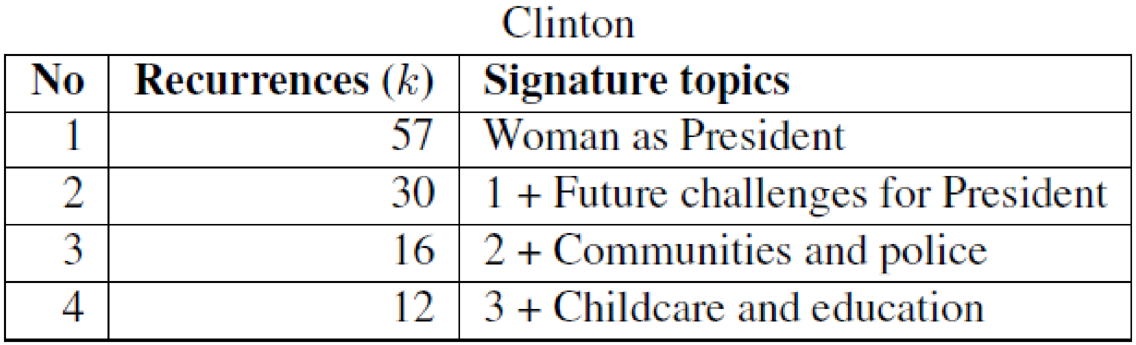
- Algorithm based on pattern mining approach
  - Pattern-growth approach $O(2^{|I|})$

Sky-signature use case

- Dataset
  - Speeches of D.Trump and H.Clinton in the 2016 presidential campaign
- Objective
  - Find the recurrent topics of each candidate
- Analysis pipeline
  - Apply topic modeling methods on the dataset to get a more abstract overview of each speech main topics
  - Compute the sky-signature on each candidate series of speeches
  - Analyze!

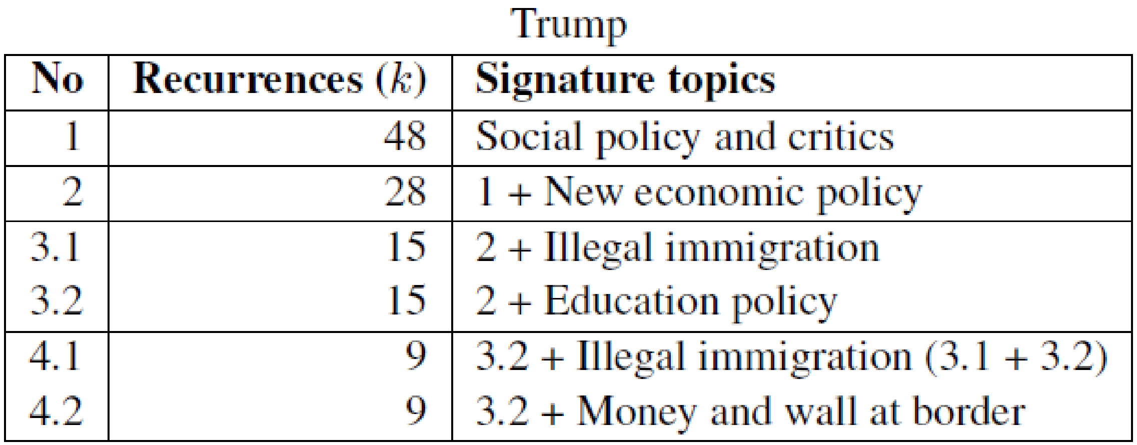
Sky-signature use case

- Politician signatures
- “Hierarchy” of main topics



Clinton

No	Recurrences (k)	Signature topics
1	57	Woman as President
2	30	1 + Future challenges for President
3	16	2 + Communities and police
4	12	3 + Childcare and education

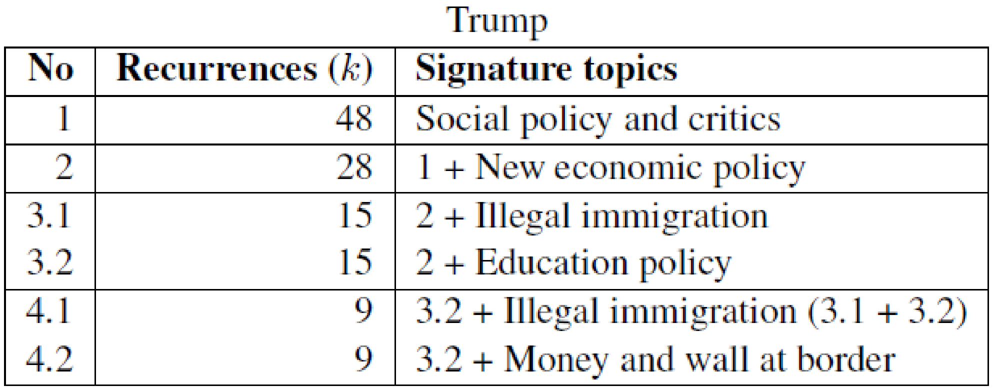


Trump

No	Recurrences (k)	Signature topics
1	48	Social policy and critics
2	28	1 + New economic policy
3.1	15	2 + Illegal immigration
3.2	15	2 + Education policy
4.1	9	3.2 + Illegal immigration (3.1 + 3.2)
4.2	9	3.2 + Money and wall at border

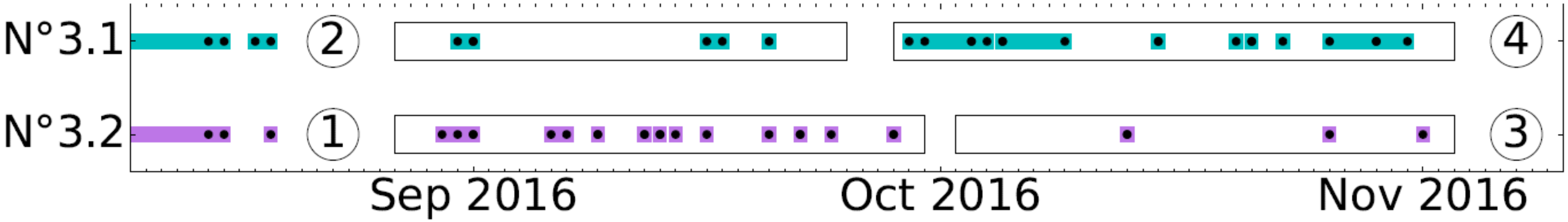
Sky-signature use case

- Information from segments



Trump

No	Recurrences (k)	Signature topics
1	48	Social policy and critics
2	28	1 + New economic policy
3.1	15	2 + Illegal immigration
3.2	15	2 + Education policy
4.1	9	3.2 + Illegal immigration (3.1 + 3.2)
4.2	9	3.2 + Money and wall at border



- Segment size and frequency provides information

Conclusion

- Signatures
  - Find regularities in data, with no constraint on the periodicity
  - No window size
- Sky-signatures
  - Removes the frequency parameter
  - More complex model
- Applied signatures on real use cases
  - Retail use case
  - Natural language processing

Perspectives

- Add quantities in the model

- Get rid of the number of segments parameter
  - First steps with MDL encoding

Thank you for your attention

Questions?