Your Customers are not a Static Picture

A Dynamic Understanding of Customer Behavior Processes Based on Self-Organizing Maps and Sequence Mining

Dr. Seppe vanden Broucke, KU Leuven

Data Mining for Business Intelligence – 2016
Ben-Gurion University of the Negev
May 19, 2016

KU LEUVEN

About the presenter

Seppe vanden Broucke

- Studied at KU Leuven (University of Leuven)
- PhD in Applied Economic Sciences
- Postdoctoral researcher at department of Management Informatics

Research

- Process and Data Mining
- Process Conformance Analysis
- Sequence Analysis
- Artificial Negative Events
- Process Discovery Algorithms
- Evolutionary Computing

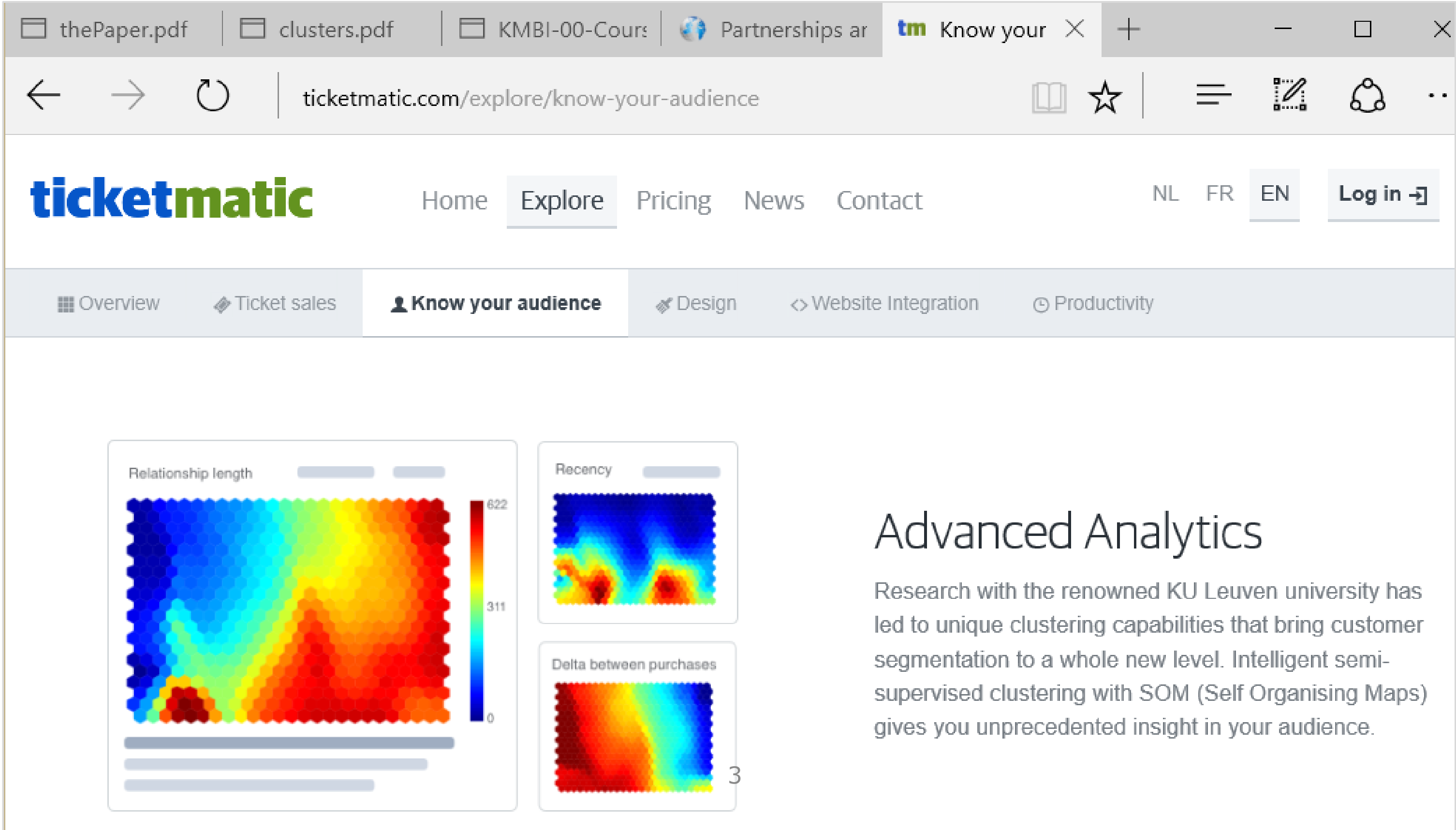
Contact

- www.seppe.net (mail, LinkedIn, …)
- seppe.vandenbroucke@kuleuven.be
- www.dataminingapps.com

datamining
apps

Introduction

- Work together with Alex Seret, Bart Baesens, Jan Vanthienen
- Ticketmatic: Netherlands and Belgium based vendor of ticketing and marketing software. Sales, CRM, marketing, analytics, after-sale support



Advanced Analytics

Research with the renowned KU Leuven university has led to unique clustering capabilities that bring customer segmentation to a whole new level. Intelligent semi-supervised clustering with SOM (Self Organising Maps) gives you unprecedented insight in your audience.

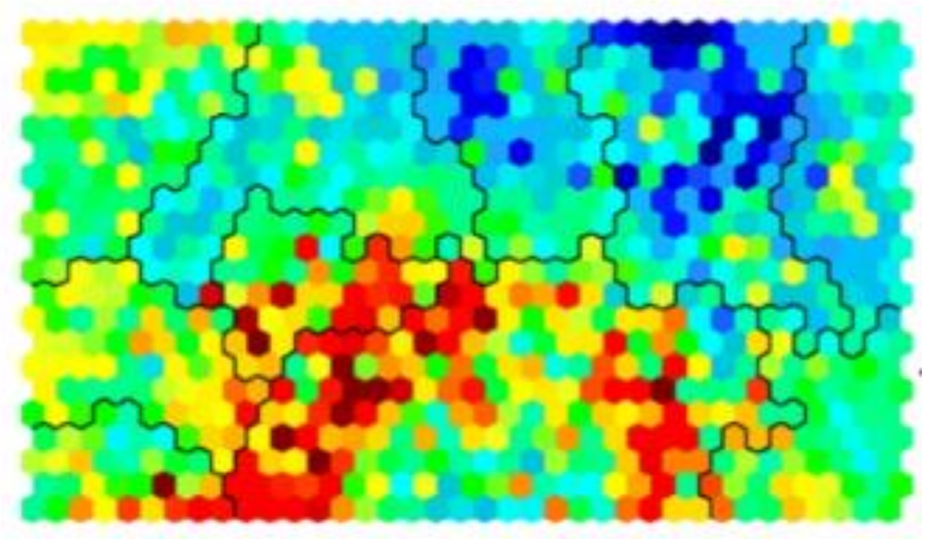
Introduction

ticketmatic

- Venue and event organisers are interested in customer insights
- “Traditional” BI: drill-up, selecting, slicing and dicing
- Also more advanced techniques such as customer segmentation
- How can we improve unsupervised data exploration?

Self-organizing maps

- Also called a Kohonen map. Introduced in 1981 by prof. Teuvo Kohonen
- Can be formalized as a special type of artificial neural networks
- Produces a low-dimensional representation of the input space
- Useful for visualizing low-dimensional views of high-dimensional data
- Topologic properties of the input space are maintained in map



Self-organizing maps

- The two main objectives of the SOM algorithm are vector quantization and vector projection

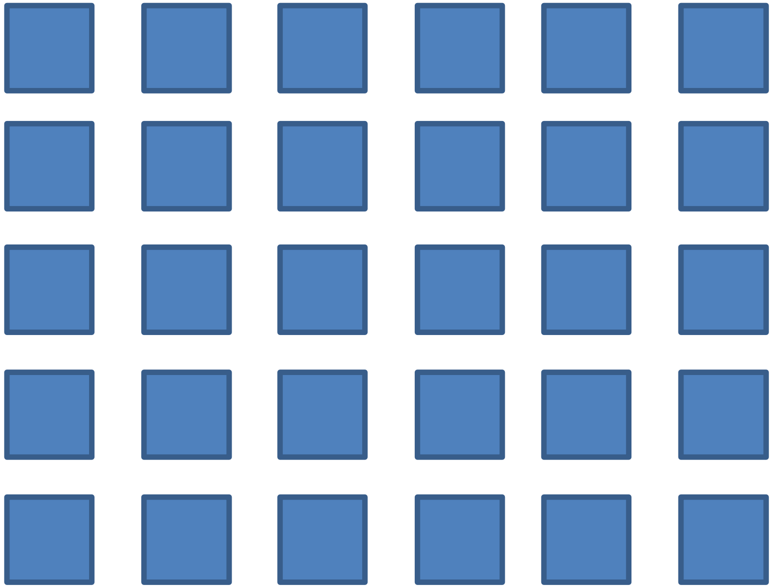
- Vector quantization aims at summarizing the data by dividing a large set of data points into groups having approximately the same number of points closest to them. The groups are then represented by their centroid points

- Vector projection aims to reduce the dimensionality of the data points by projection onto lower dimensional maps. Typically, a projection to two-dimensional maps is performed

Self-organizing maps

Start by laying out a group of output nodes

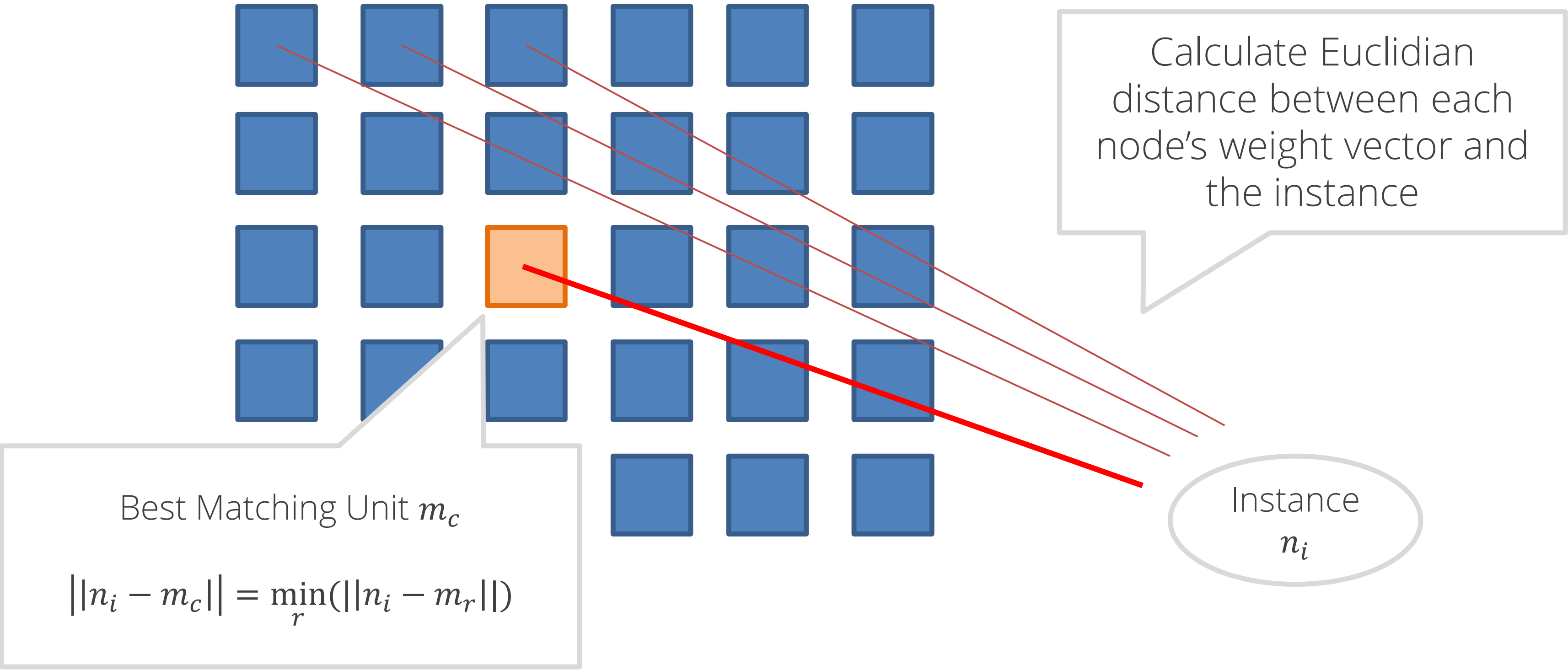
(In most cases 2d, rectangular or hexagonal grid)



Each node gets initialized with a weight vector (e.g. random) with same length as instances

Self-organizing maps

Next, we iterate over all instances, and for every instance, we check each node



Calculate Euclidian distance between each node's weight vector and the instance

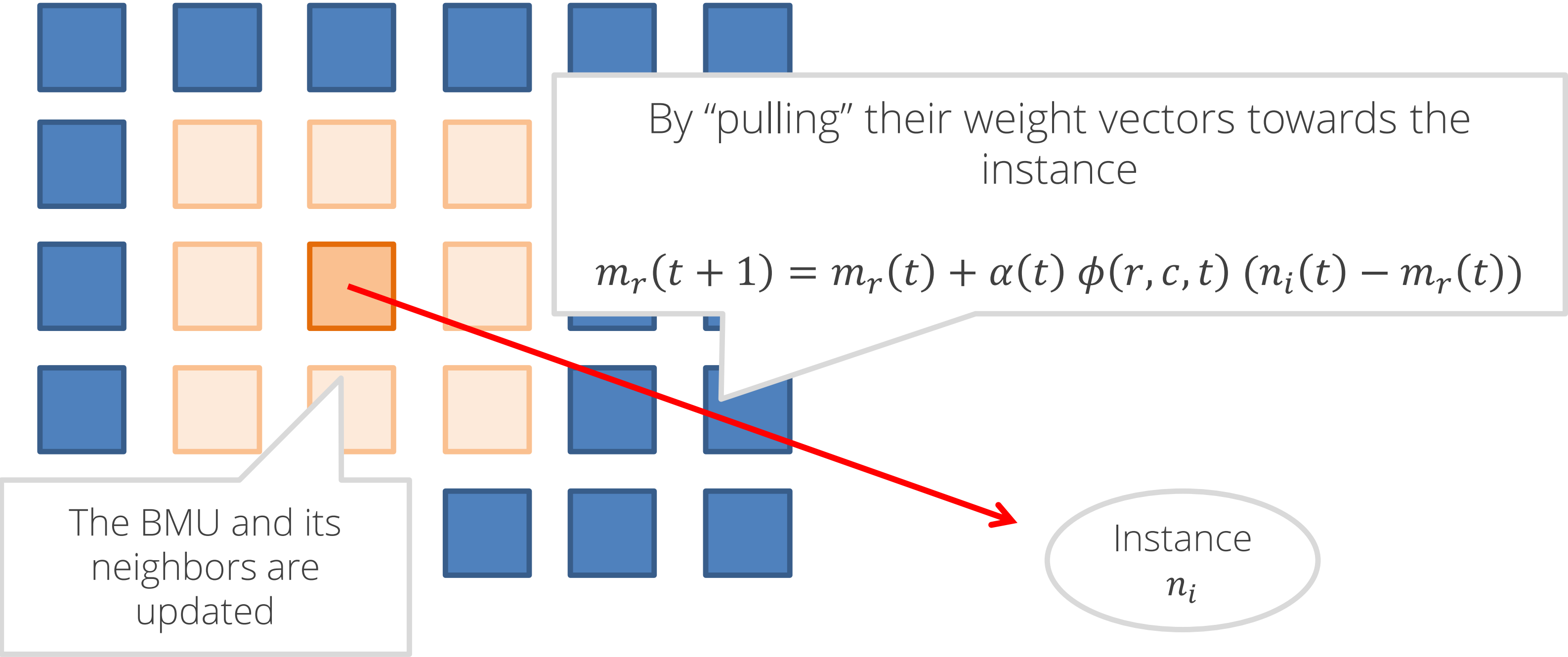
Best Matching Unit m_c

$$||n_i - m_c|| = \min_r(||n_i - m_r||)$$

Instance n_i

Self-organizing maps

Next, the weights of the BMU and neighbors are adjusted



By “pulling” their weight vectors towards the instance

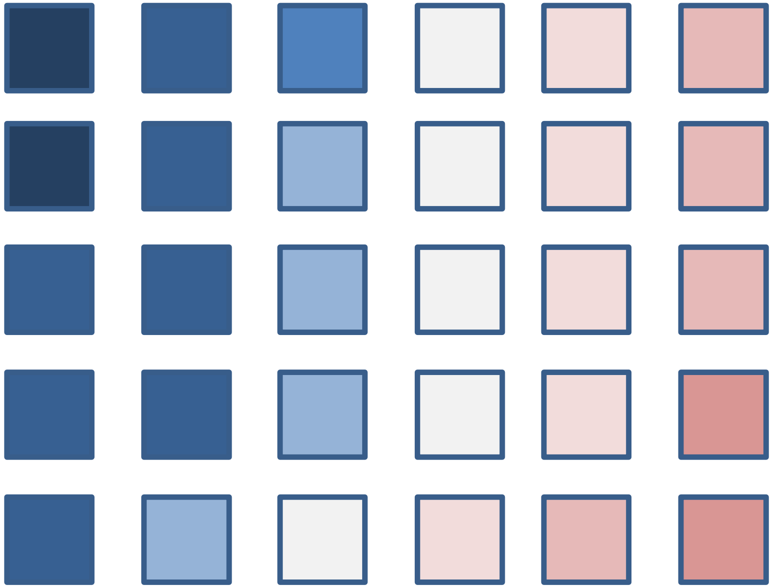
$$m_r(t+1) = m_r(t) + \alpha(t)\ \phi(r,c,t)\ (n_i(t) - m_r(t))$$

The BMU and its neighbors are updated

Instance
n_i

Self-organizing maps

At the end, every node has a resulting weight vector



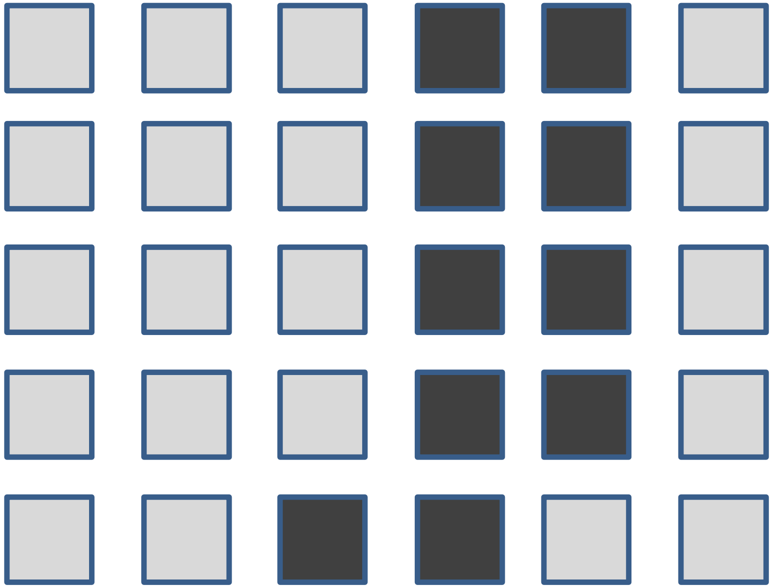
Visualize the values for the *n*'th element in the weight vectors

Groups and regions appear

We know the position, weigh vector and associated input instances for each node

Self-organizing maps

At the end, every node has a resulting weight vector

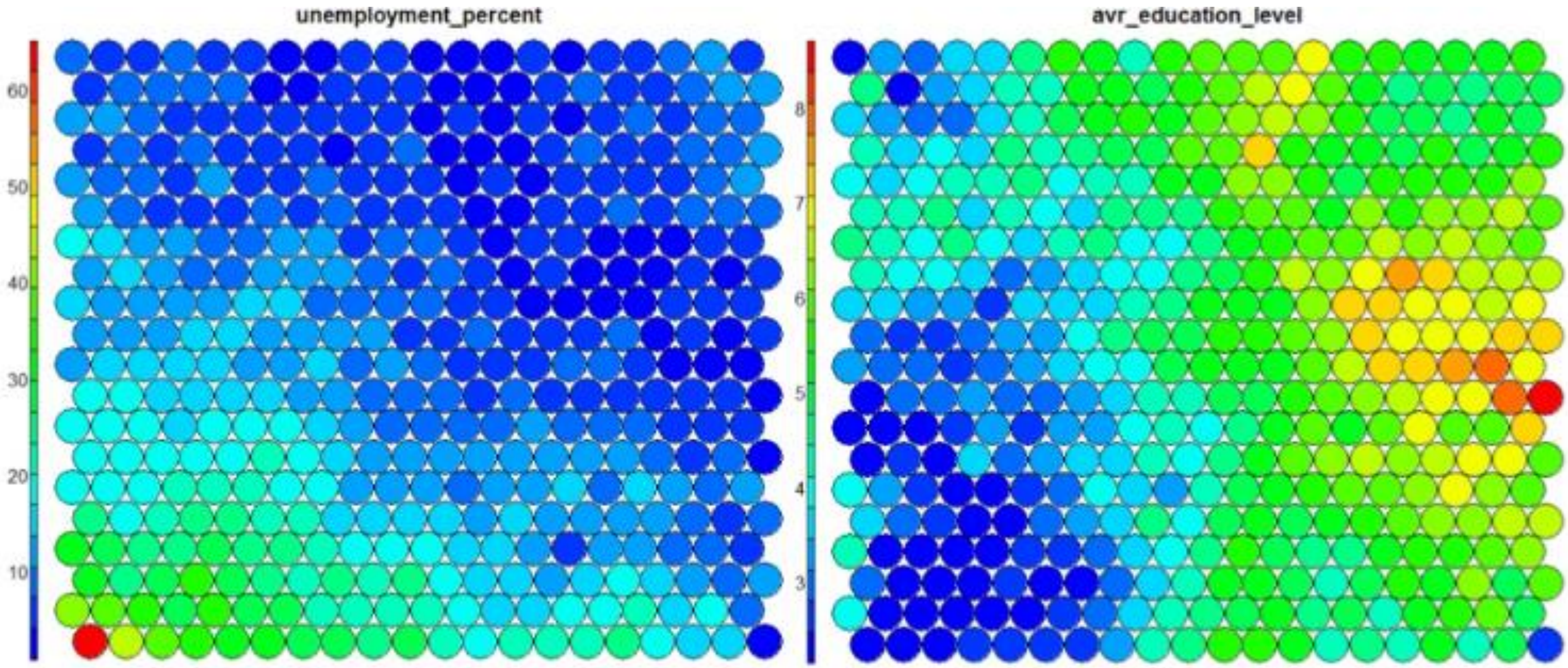


Visualize mean distance between node and its neighbors

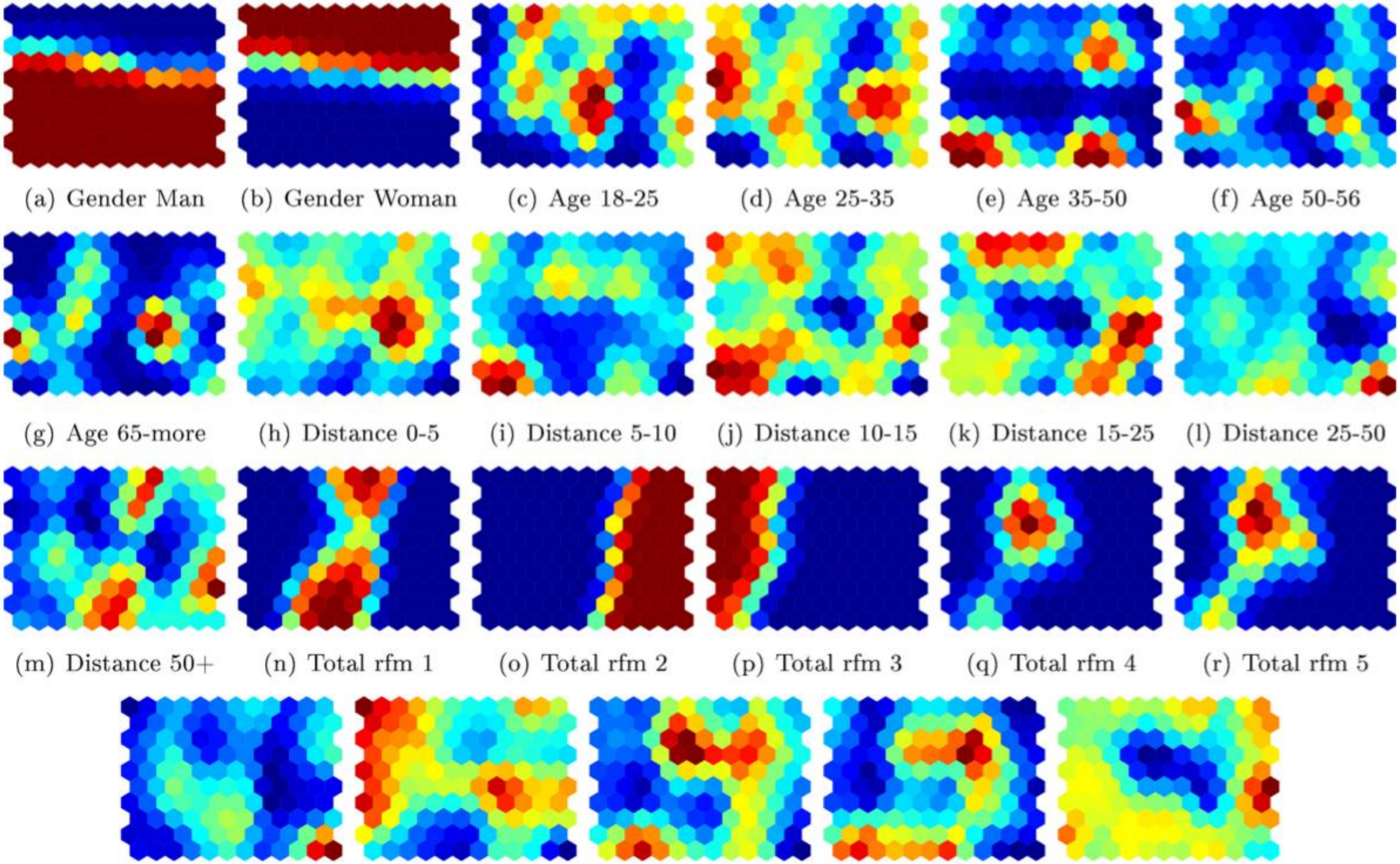
Boundaries and edges appear

U-matrix

Self-organizing maps



Self-organizing maps



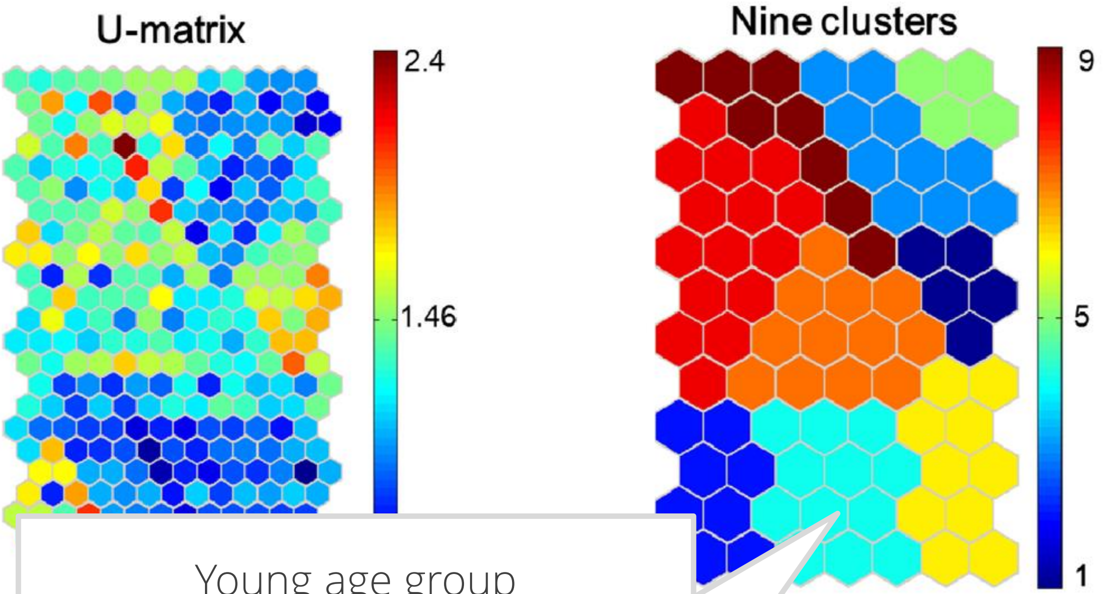
(a) Gender Man (b) Gender Woman (c) Age 18-25 (d) Age 25-35 (e) Age 35-50 (f) Age 50-56

(g) Age 65-more (h) Distance 0-5 (i) Distance 5-10 (j) Distance 10-15 (k) Distance 15-25 (l) Distance 25-50

(m) Distance 50+ (n) Total rfm 1 (o) Total rfm 2 (p) Total rfm 3 (q) Total rfm 4 (r) Total rfm 5

(s) The Concert 1 (t) The Concert 2 (u) The Concert 3 (v) The Concert 4 (w) The Concert 5

Self-organizing maps



Young age group

Mostly students

Higher spending on concerts than average

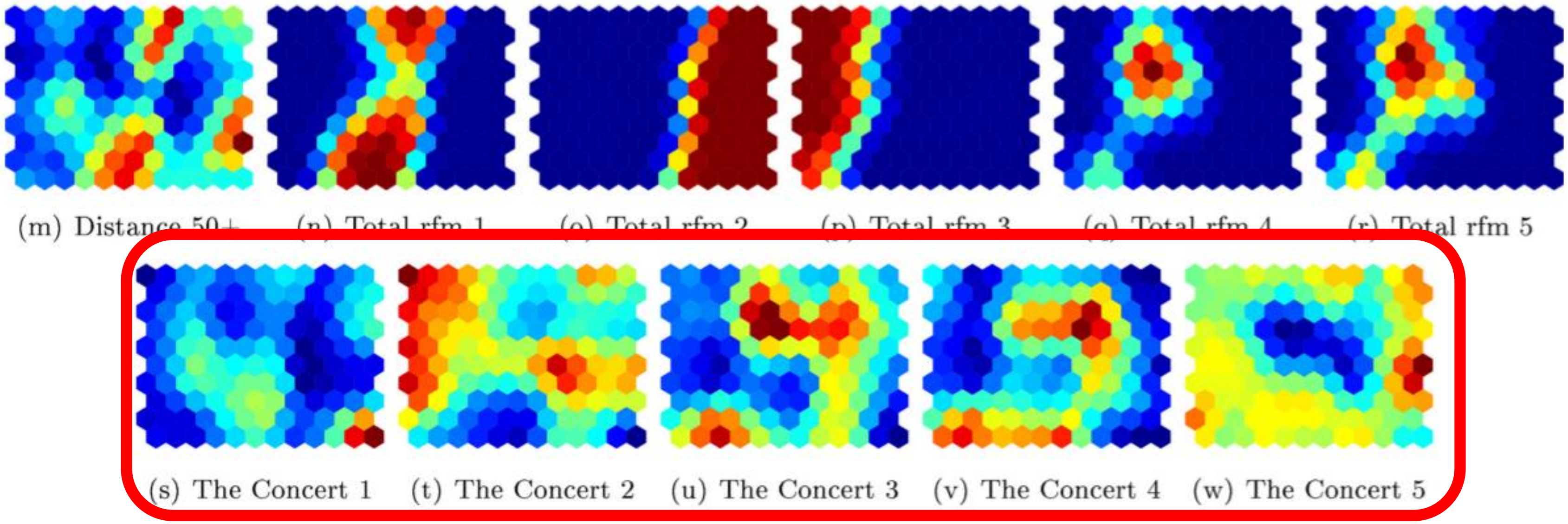
“Young culture fanatics”

Automatic labeling possible (e.g. using salient dimension)

Problem one: prioritizing variables

- We wish to incorporate business knowledge in the segmentation exercise

- “I’m interested to expect the groups based on to which concert people went, but they’re not topologically close”



(m) Distance 50+ (n) Total rfm 1 (o) Total rfm 2 (p) Total rfm 3 (q) Total rfm 4 (r) Total rfm 5

(s) The Concert 1 (t) The Concert 2 (u) The Concert 3 (v) The Concert 4 (w) The Concert 5

Problem one: prioritizing variables

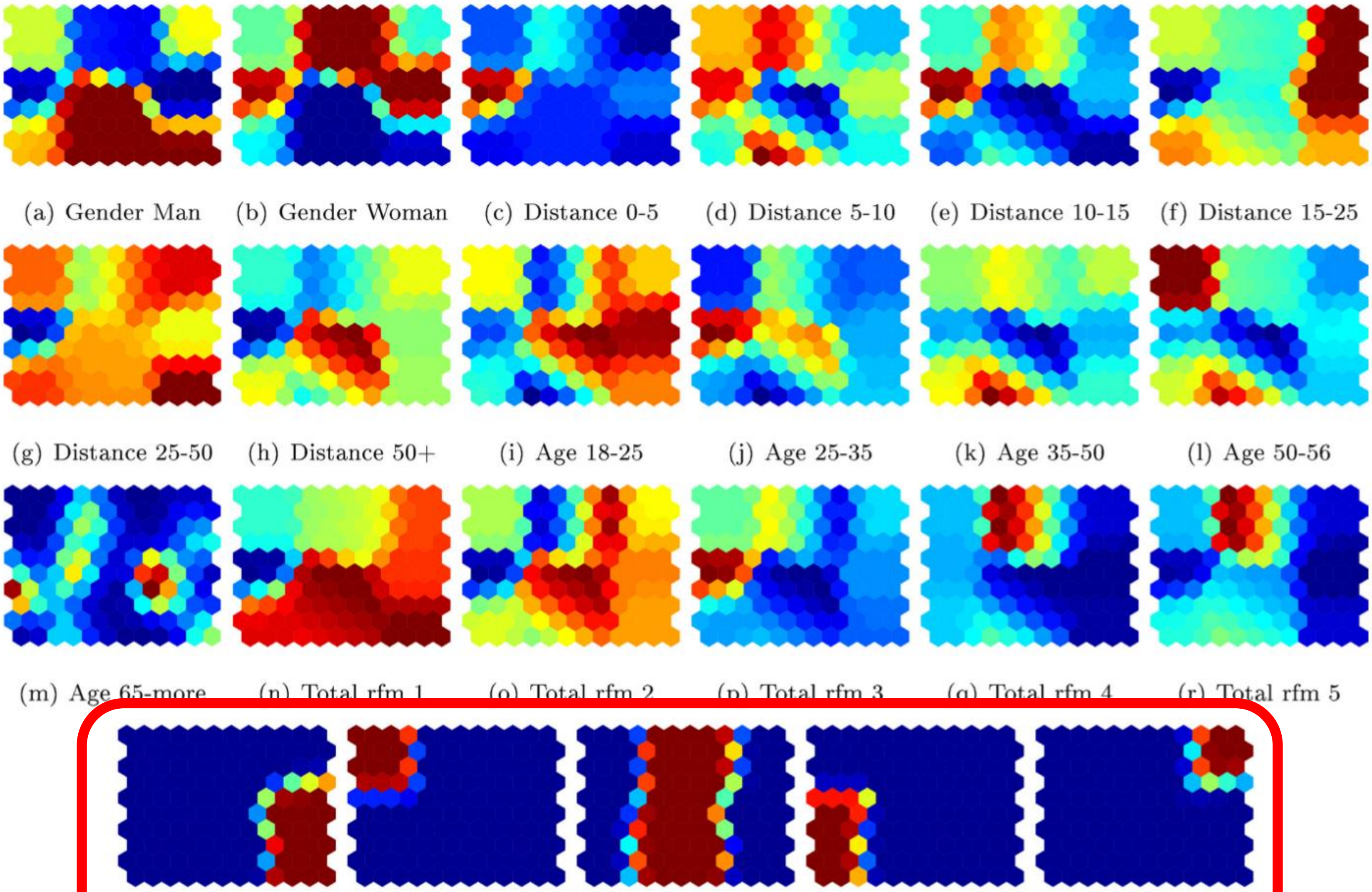
- BMU identification step is modified

$$m_c \; with \; c = argmax_r(\sqrt{\sum_{j=1}^{d} w_{d_j}\left(n_{id_j} - m_{rd_j}\right)^2})$$

Weight assigned to variable d_j

The higher the weight, the higher the impact in the resulting clustering

Problem one: prioritizing variables



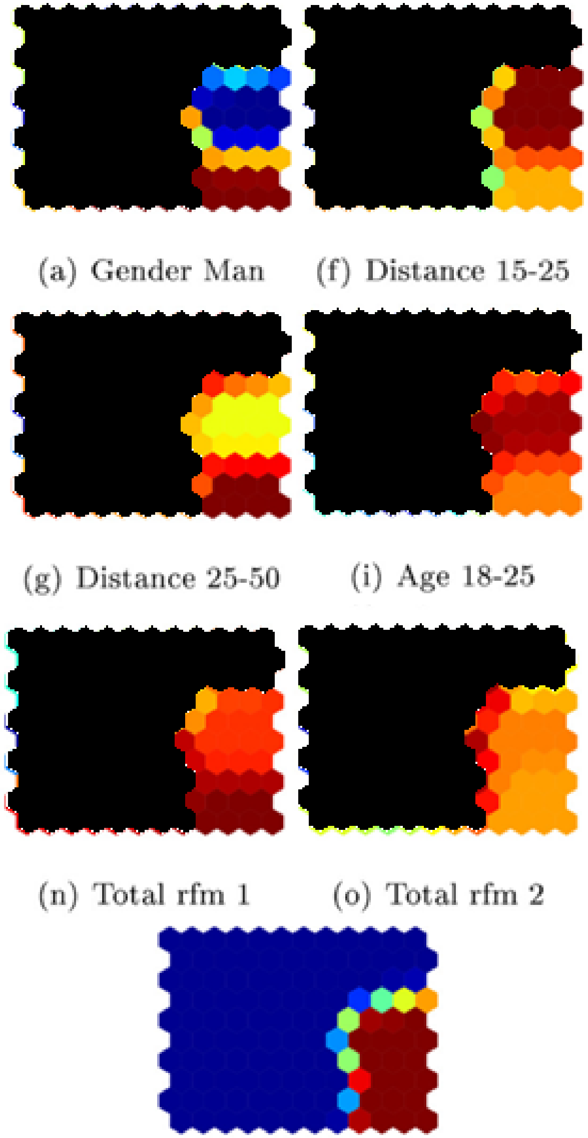
(a) Gender Man (b) Gender Woman (c) Distance 0-5 (d) Distance 5-10 (e) Distance 10-15 (f) Distance 15-25

(g) Distance 25-50 (h) Distance 50+ (i) Age 18-25 (j) Age 25-35 (k) Age 35-50 (l) Age 50-56

(m) Age 65-more (n) Total rfm 1 (o) Total rfm 2 (p) Total rfm 3 (q) Total rfm 4 (r) Total rfm 5

(s) The Concert 1 (t) The Concert 2 (u) The Concert 3 (v) The Concert 4 (w) The Concert 5

Problem one: prioritizing variables



(a) Gender Man (f) Distance 15-25

(g) Distance 25-50 (i) Age 18-25

(n) Total rfm 1 (o) Total rfm 2

(s) The Concert 1

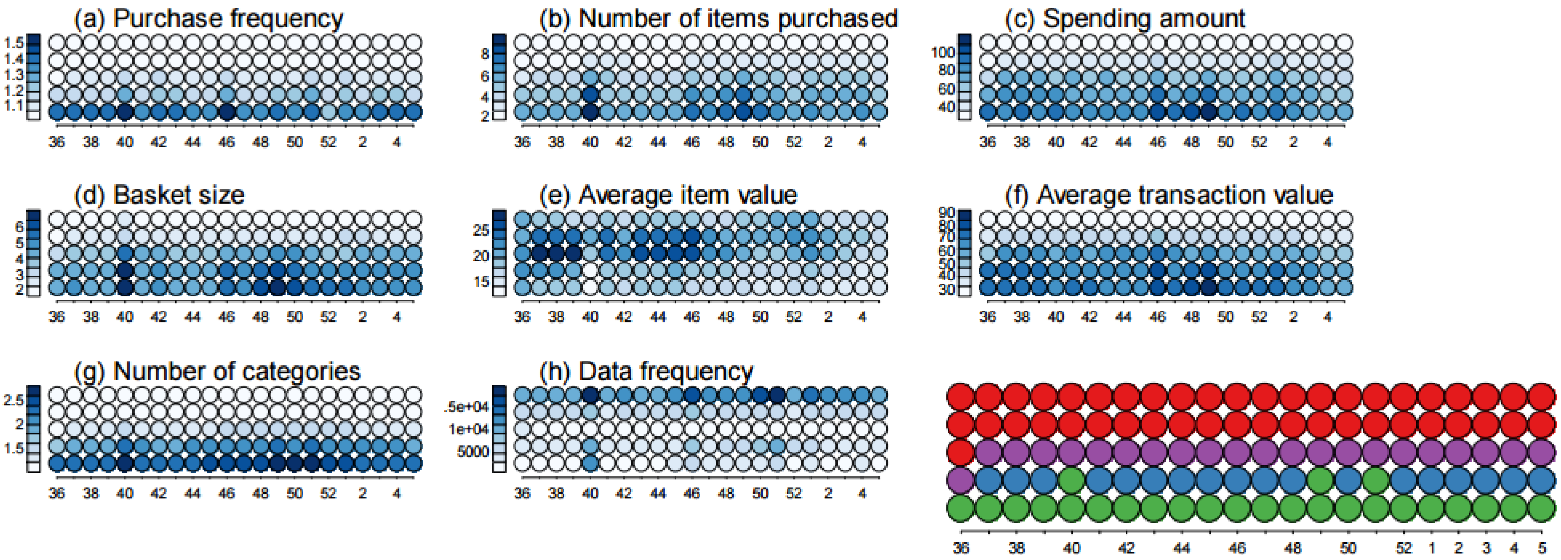
Problem two: analyzing time dynamics

- “I wish to track how customers evolve through time, how they move from cluster to cluster”

- “I’m interested in customers who’ll end up in an interesting group some time from now…”

Problem two: analyzing time dynamics

- Prior work: self-organizing time map (Sarlin, 2012)
- Basically a one-dimensional SOM, the other dimension represents a sorted order of time
- Batch update per time unit



Problem two: analyzing time dynamics

- Loss of one dimension

- Difficult to combine with priorization approach

- Requires data on every instance at every time point

- Hard to convert to discriminative search

Problem two: analyzing time dynamics

- Alternative approach

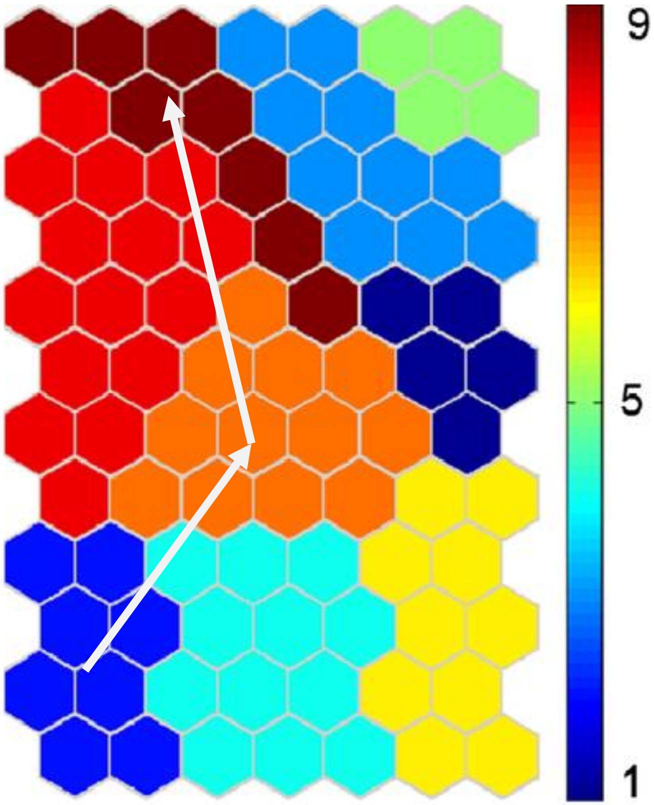
- Create dataset by merging all information through time on every instance

- Create SOM and clusters

- Apply generalized sequential pattern algorithm (Srikant, Agrawal)

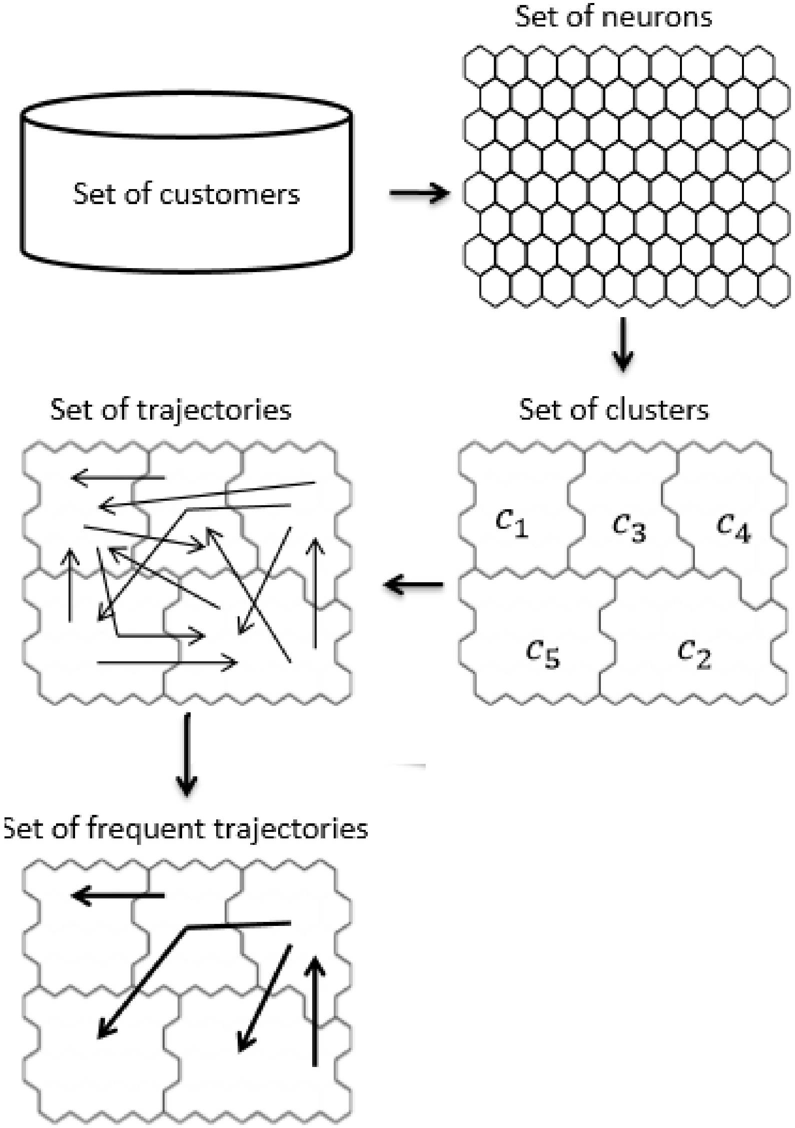
Problem two: analyzing time dynamics

- Mapping from instances to neurons and neurons to clusters allows for the identification of trajectories followed by the items through time, i.e. items moving from cluster to cluster



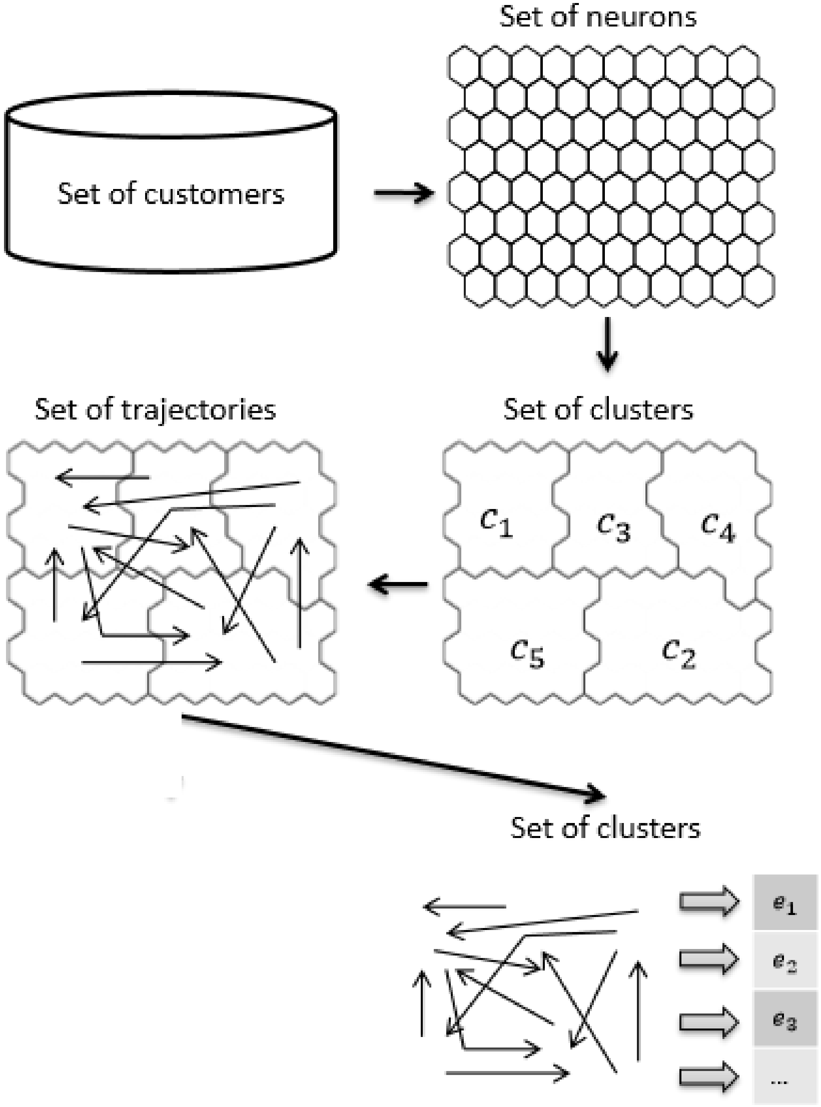
Problem two: analyzing time dynamics

- This approach leads to a lot of trajectories, so we apply a frequent sequence mining technique to extract the frequent trajectories



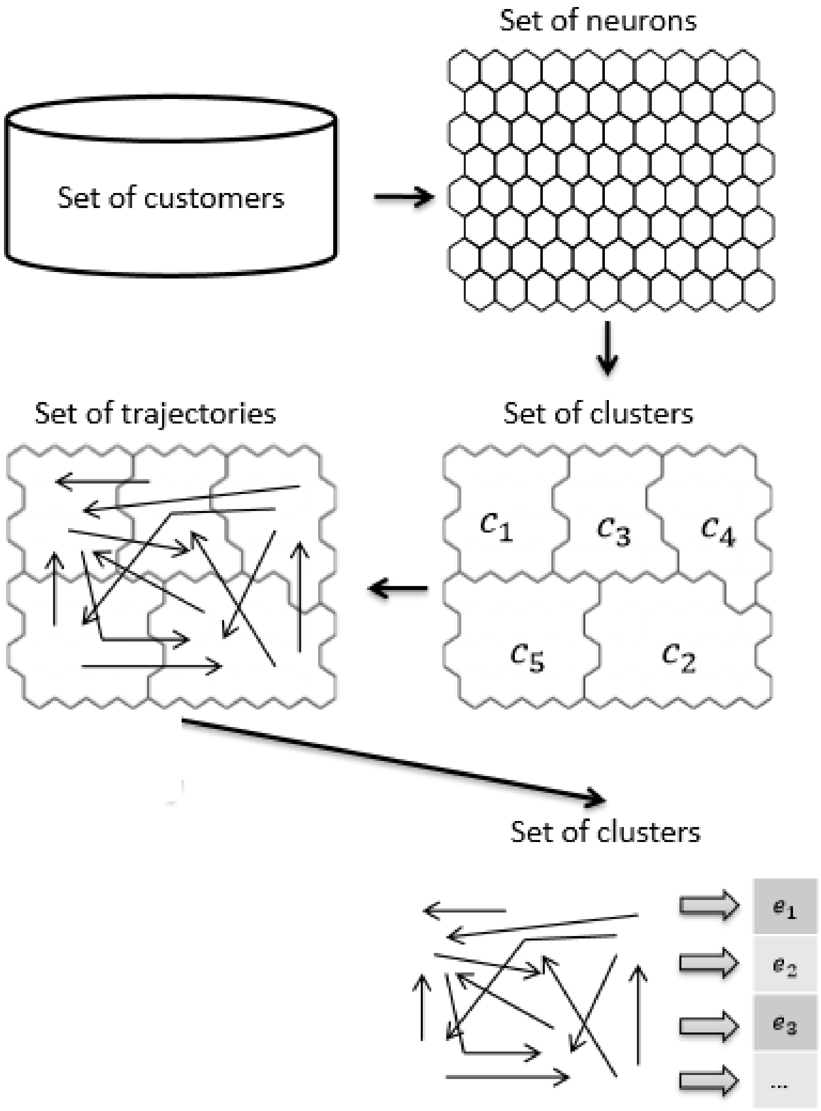
Problem two: analyzing time dynamics

- Summarizing the different trajectories using a statistical approach
- Instead of a description of the entire trajectory, this approach focuses on specific segments of a trajectory in order to identify trends
- A cluster-level movement, or delta of an input vector is calculated by comparing the cluster-level coordinates of the instance at two times: $\delta^{n_i}_{t_a,t_b} = centroid^{t_b}_{n_i} - centroid^{t_a}_{n_i}$

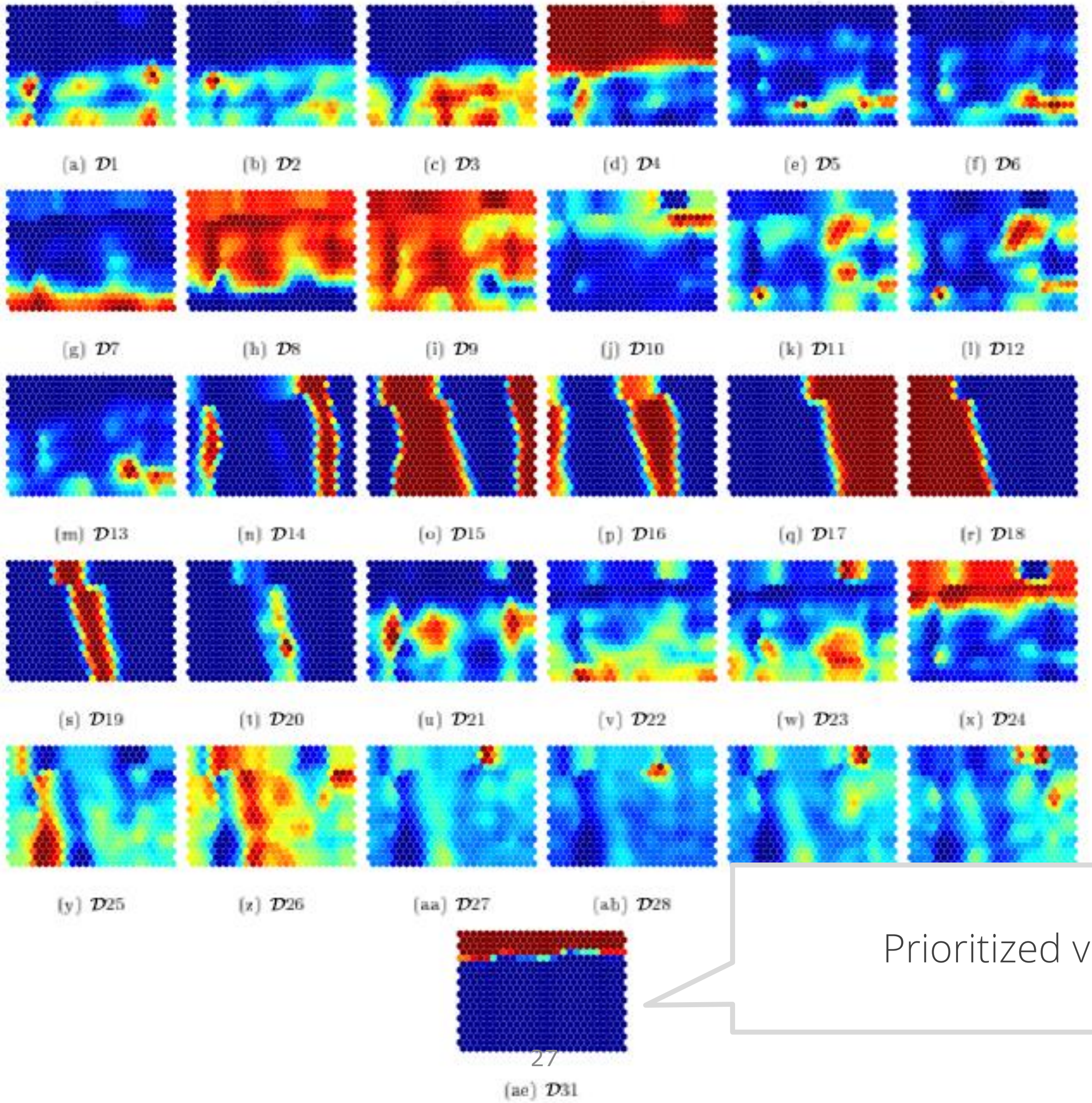


Problem two: analyzing time dynamics

- $\delta_{t_a,t_b}^{n_i} = centroid_{n_i}^{t_b} - centroid_{n_i}^{t_a}$
- This movement vector can be used to characterize the main trends forming the dynamics of the input vectors, applying a second-step clustering on the set of delta's



Problem two: analyzing time dynamics



(a) $\mathcal{D}1$ (b) $\mathcal{D}2$ (c) $\mathcal{D}3$ (d) $\mathcal{D}4$ (e) $\mathcal{D}5$ (f) $\mathcal{D}6$

(g) $\mathcal{D}7$ (h) $\mathcal{D}8$ (i) $\mathcal{D}9$ (j) $\mathcal{D}10$ (k) $\mathcal{D}11$ (l) $\mathcal{D}12$

(m) $\mathcal{D}13$ (n) $\mathcal{D}14$ (o) $\mathcal{D}15$ (p) $\mathcal{D}16$ (q) $\mathcal{D}17$ (r) $\mathcal{D}18$

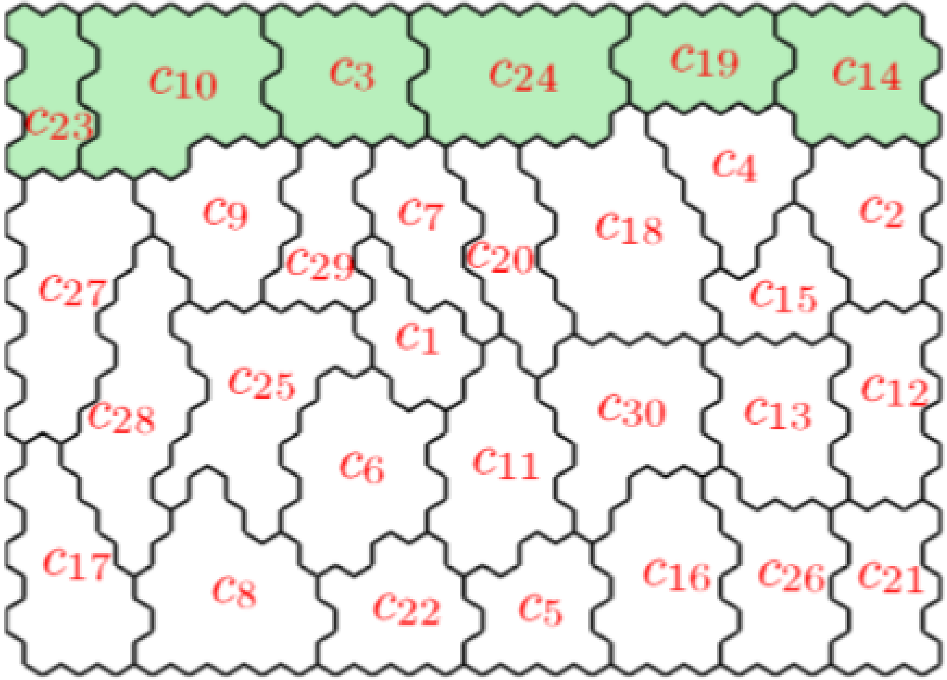
(s) $\mathcal{D}19$ (t) $\mathcal{D}20$ (u) $\mathcal{D}21$ (v) $\mathcal{D}22$ (w) $\mathcal{D}23$ (x) $\mathcal{D}24$

(y) $\mathcal{D}25$ (z) $\mathcal{D}26$ (aa) $\mathcal{D}27$ (ab) $\mathcal{D}28$

(ae) $\mathcal{D}31$

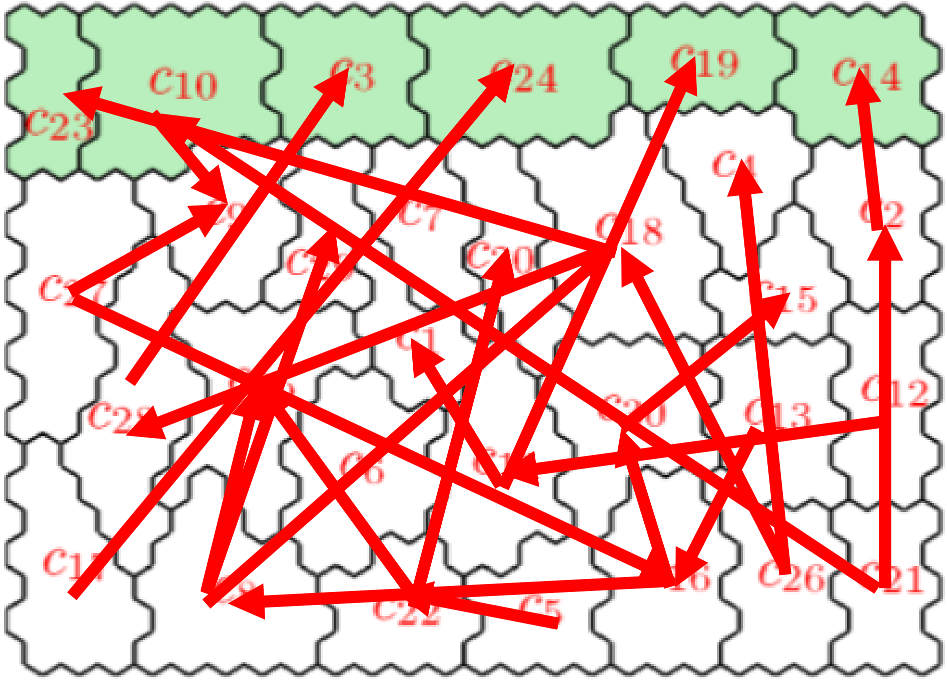
Prioritized variable

Problem two: analyzing time dynamics



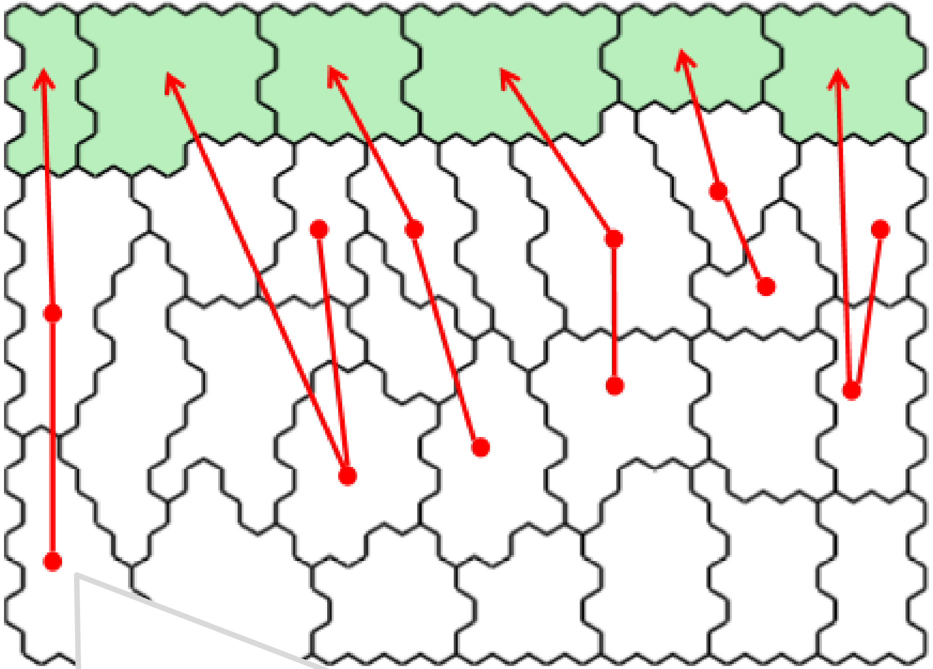
Clusters

Problem two: analyzing time dynamics



All trajectories

Problem two: analyzing time dynamics



After discriminative GSP

Six frequent trajectories leading to the different clusters of subscription holders

This trajectory, leading to the first subscription, is associated with an increase in the average number of days separating the purchase of the tickets and the event related to it and an increase in the customer value

Problem two: analyzing time dynamics

- Calculate the deltas corresponding to the final movement towards the clusters of interest: $\delta_{t_a,t_b}^{n_i} = centroid_{n_i}^{t_b} - centroid_{n_i}^{t_a}$ with $centroid_{n_i}^{t_b}$ the centroid of a cluster of interest
- Apply k-means

Cluster 1 (772 deltas): customer value ▲

Cluster 2 (1235 deltas): no significant increase or decrease for any variable

Cluster 3 (715 deltas): time of purchase ▼,
nr. previous purchases with organizer ▲

Cluster 4 (457 deltas): ticket-pair purchases ▲

Cluster 5 (418 deltas): tickets-per-event ▲

Problem two: analyzing time dynamics

▲ ▲ ▲

customerValue | timePurchaseBeforeEvent | relationshipLength | numberTicket

▼

Wrap-up

- Powerful semi-supervised exploratory analysis and segmentation using clustering

- Semi-supervised: prioritization, indication of important clusters

- Correlations between clusters

- Time dynamics: frequent sequences and clustered delta trends

References and further reading

- Alex Seret, Thomas Verbraken, Sébastien Versailles, Bart Baesens, A new SOM-based method for profile generation: Theory and an application in direct marketing, European Journal of Operational Research, Volume 220, Issue 1, 1 July 2012, Pages 199-209

- Alex Seret, Thomas Verbraken, Bart Baesens, A new knowledge-based constrained clustering approach: Theory and application in direct marketing, Applied Soft Computing, Volume 24, November 2014, Pages 316-327, ISSN 1568-4946

- Seret, A., vanden Broucke, S., Baesens, B., Vanthienen, J. (2014). A dynamic understanding of customer behavior processes based on clustering and sequence mining. Expert Systems with Applications, 41 (10), 4648-4657

- Peter Sarlin, Decomposing the global financial crisis: A Self-Organizing Time Map, Pattern Recognition Letters, Volume 34, Issue 14, 15 October 2013, Pages 1701-1709

- http://www.dataminingapps.com/dma_research/marketing-analytics/

Thank you

QA