# Contributions to Pattern Mining in Augmented Graphs

**Marc Plantevit**

Université Claude Bernard Lyon 1

LIRIS UMR5205

*HDR Defence, December 14th*

# Who am I ?

PC Membership: ≈ 50 conferences (ECMLPKDD, SIAM DM, IDA, IJCAI, AAAI, etc.)
Reviewer for journals: ≈ 10 reviews per years (Dami, Mach, TKDE, IS, SADM, etc.)

**PhD in Computer Science, Université Montpellier 2, « Multidimensional Sequence Mining » (2008)**

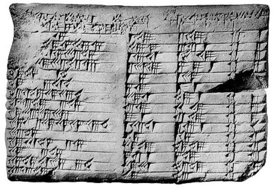
**Post-Doctoral Position at Université de Caen Basse-Normandie (Sept. 2008 – Sept. 2009)**

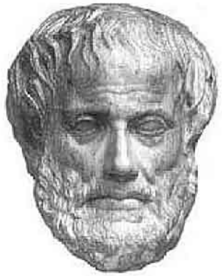
**Since 2009: Associate Professor, Université Claude Bernard Lyon 1, LIRIS UMR5205, Data Mining and Machine Learning research group**

Databases, Data Mining, Game Theory, Student R&D projects (TER)
BCS and Master (Data Science, Web and Technologies, AI, Bioinformatics, Theoretical computer science) Students

# KDD is not a hype, but a natural evolution of science

*The Fourth Paradigm. Tony Hey, Stewart Tansley, and Kristin Tolle. Microsoft Research, 2009.*





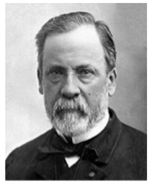
- Babylonian mathematics
- Ancient Egypt: no theorization of algoritms.

Empirical Science

1600

Each discipline has grown a theoretical components..

Theoretical Science



1950s

Computational Science

Inability to find closed form solutions

**simulations**

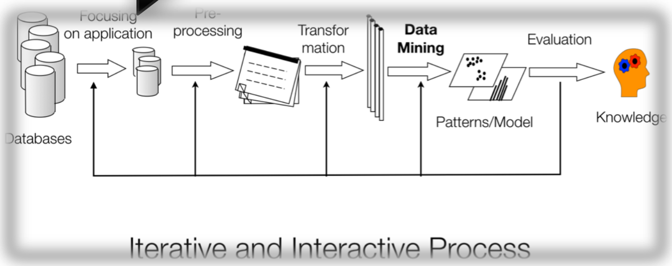
1990s

Data Science

**The Fourth paradigm: Data-Intensive Scientific Discovery**

Flood of data: new scientific instruments and simulations.

Ability to economicaly **store** and **manage** huge data




***Data Mining as a major challenge!***

# Machine Learning or Data Mining ? What are the differences ?

- Predictive **global** modeling
  - Turn the data into an as **accurate** as possible prediction machine
  - Ultimate purpose is **automatization**
  - E.g., self driving car, pattern recognition

Accuracy of Models

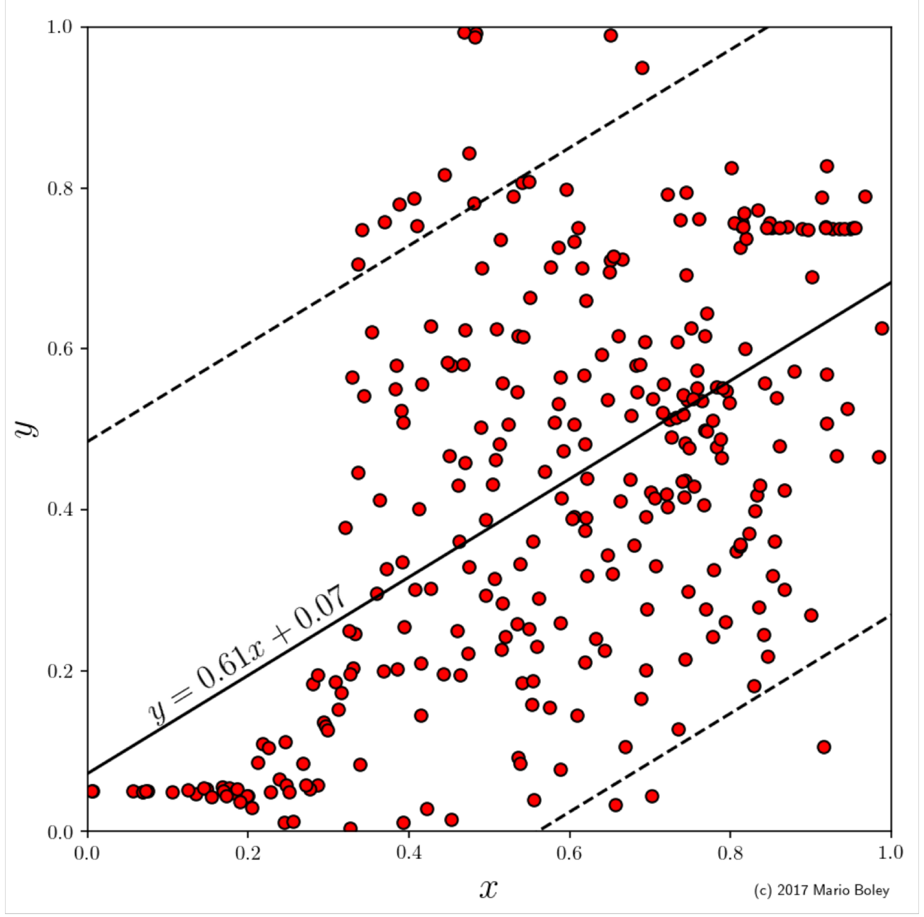
- Exploratory data analysis
  - Automatically discover new **insights** about the domain in which the data was measured
  - Use machine discoveries to synergistically **boost human expertise**
  - E.g., understanding the factors of the olfactory process

Interpretability of results

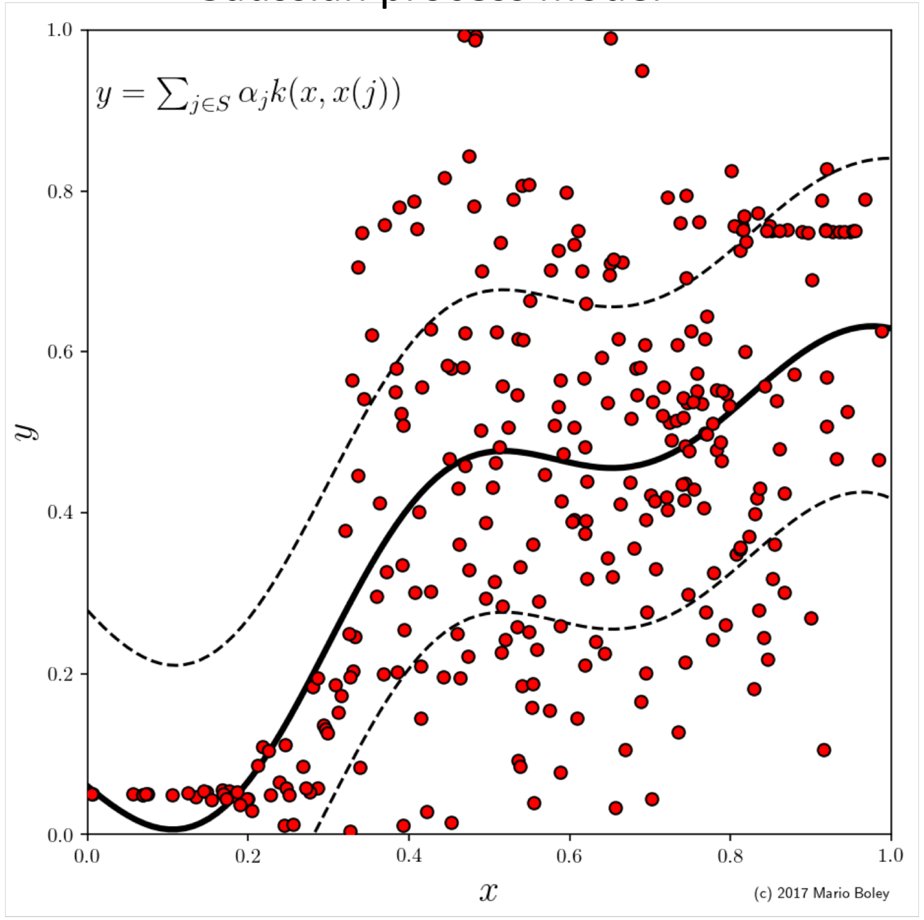
*M. Boley, www.realkd.org*

# « A good prediction machine does not necessarily provide explicit insights into the data domains»

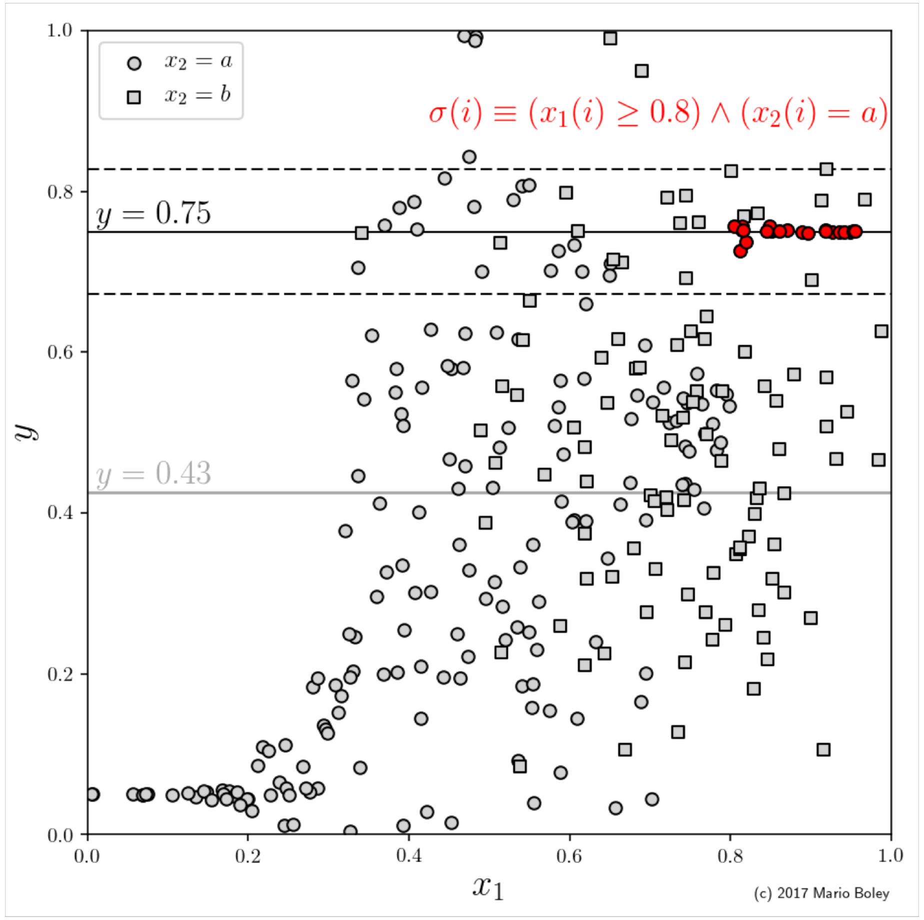
Global linear regression model




Gaussian process model



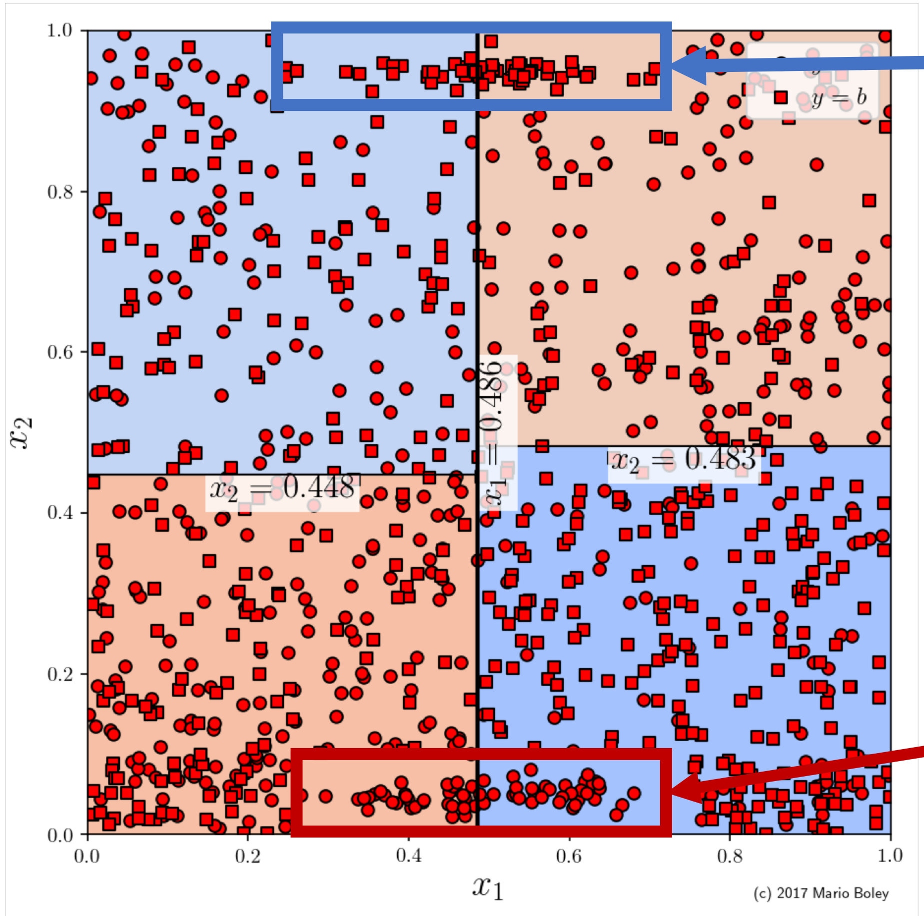
# Global Modelling: the need to explain all the data




- «A complex theory of everything might be of less value than a simple observation about a specific part of the data space»
- Identifying interesting subspace and the power of saying *« I don't know for other points »*

Global Modelling is guided by the global picture and may not uncover some interesting insights while local patterns make  it possible.

Decision tree with 0.7 accuracy




$q_1 = (x_1(i) \in [0.24, 0.71] \wedge x_2 \geq 0.9)$

**Two subgroups with 1.0 accuracy**

$q_2 = (x_1(i) \in [0.26, 0.7 \wedge x_2(i) \leq 0.1)$

# From an Inductive DB Perspective

$$Th(L, D, C) = \{\psi \in L \mid C(\psi, D) \text{ is true}\}$$

*Imielinski and Mannila, Comm. of the ACM, 1996*

- *L* a pattern language
- *D* a database
- *C* some constraints

Pattern mining algorithm: **enumerating** the **elements** of the language that fulfil the **constraints** within the **data**.

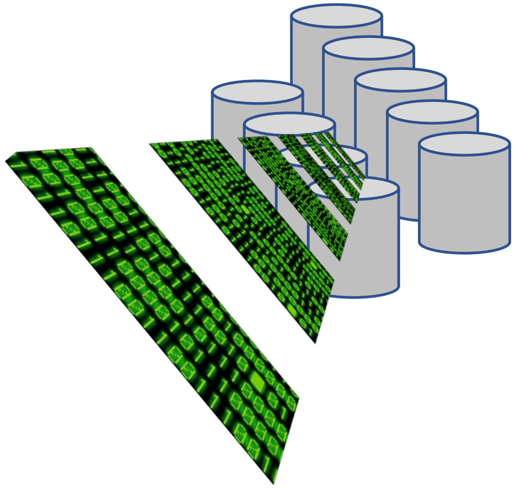
Study of the constraint properties to devise effective pruning strategies.

Some variants:

- **Elements**: complete set, top k, representative sample, etc.
- **Constraints**: from satisfaction problem to optimization problem.
- **Data**: {batch data, streaming data} x {graph(s), tree(s), sequence(s), [numerical] itemset(s), etc.}.

# Fill the gap between the user and her data

**Complexity of the data:**
Numerical attributes, graphs, sequences, images, streams vs batches, etc.



**Complexity of the user:**
Explicit interest via constraints or preferences, implicit interest to be learned

**Complexity of the domain:**
How to handle domain knowledge and do not return already known phenomena

**Complexity of the output:**
Do not overwhelm the user! each pattern has an assimilation cost.

**Applications**
Biology, Neurosciences, Material Engineering, Social Science, Geology, Chemistry, Industry 4.0, …

# Some contributions

**Sequence mining**

- Multidimensional sequences *(Plantevit et al., ACM TKDD 2010)*
- Extending som ences
  - Strong f *vit and Crémilleux, IDA'0*
  - $\delta$-free s
- For text mini *nantics 2015)*

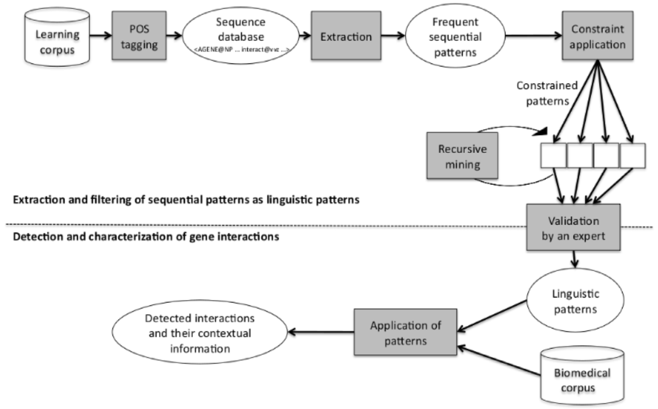
Results are **difficult** to **assimilate** for the end-users

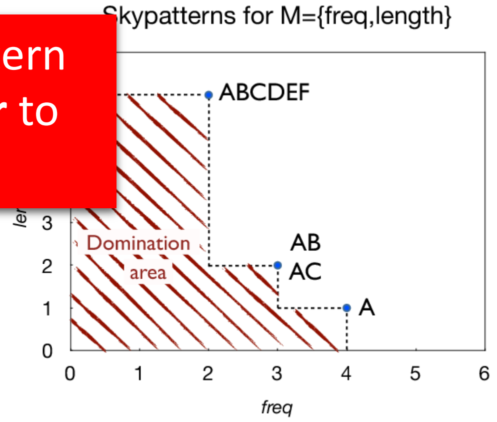
Looking for pattern language **easier** to understand

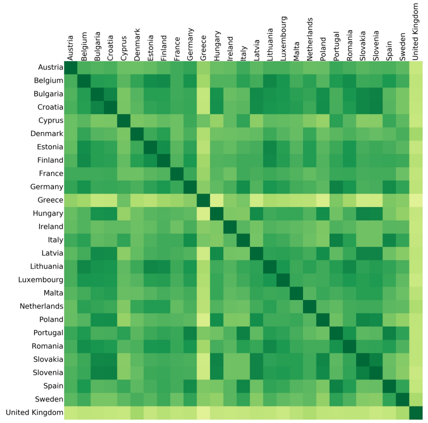
**Making the constraint-based pattern mining framework easier to use**

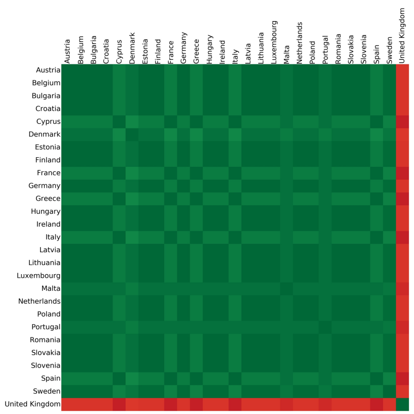
- Explicit preferences to overcome the thresholding problem with the skypatterns *(Soulet et al., ICDM'11), (Ugarte et al., Artificial Intelligence 2017)*

**New pattern domain to analyse rating or vote data**
*(Belfodil et al., ECMLPKDD'17)*

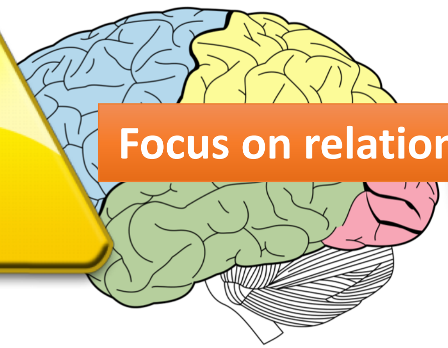


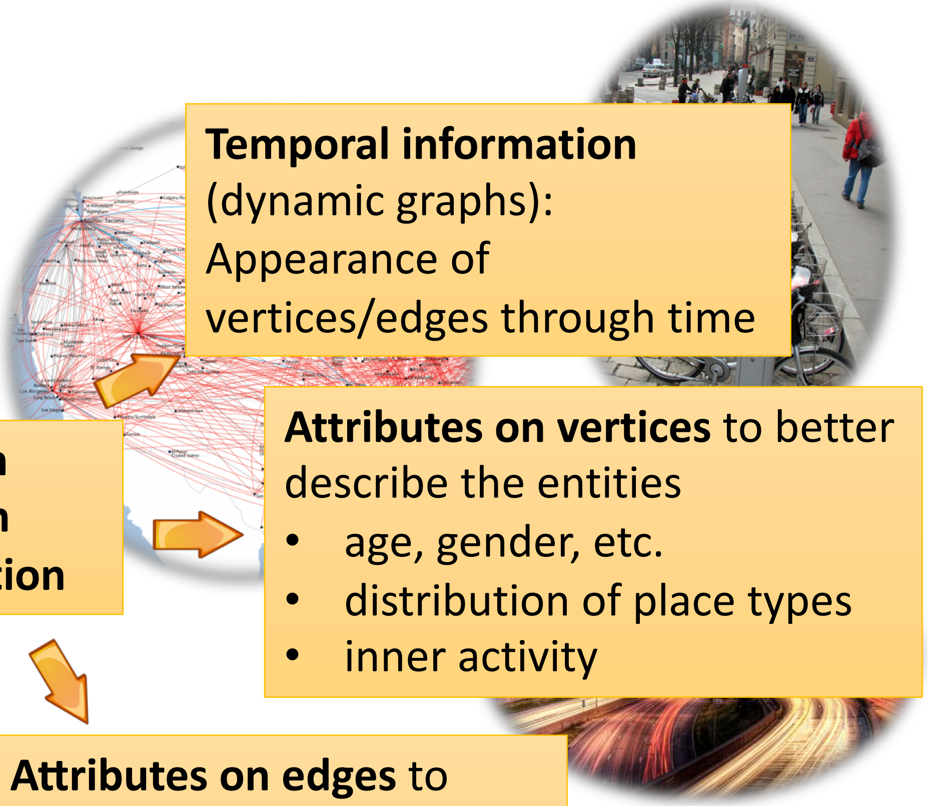





# Graphs as a powerful mathematical tool to model real-world phenomena



**Focus on relational graphs**

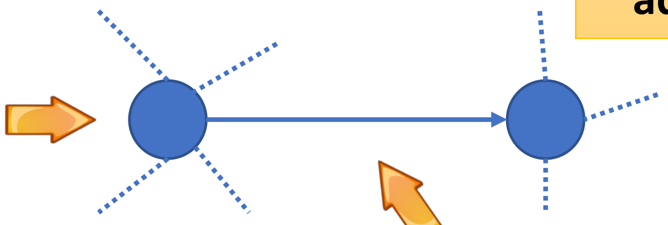


**Temporal information** (dynamic graphs): Appearance of vertices/edges through time

**Graphs are often augmented with additional information**

**Attributes on vertices** to better describe the entities
- age, gender, etc.
- distribution of place types
- inner activity

**A vertex: an entity**
- user in SN
- scientists
- bicycle stations
- city area
- airport
- (pi|vo)xel
- …



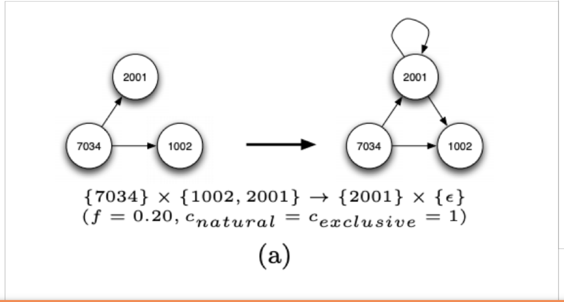
**An edge: an interaction**
- friendship/follow
- co-authorship/citations
- travels
- Connectivity

**Attributes on edges** to better describe the interactions
- weather conditions
- day of week

Overview of the contributions to pattern mining in augmented graphs

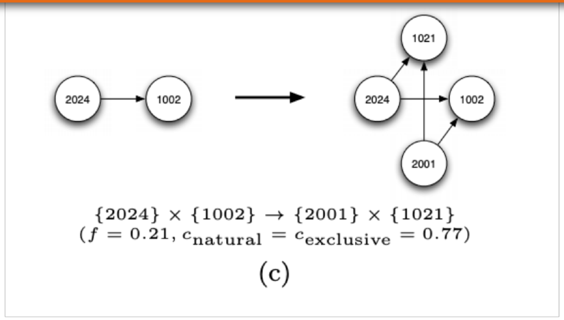
## Association rules in dynamic graphs

- Revisiting the confidence measure



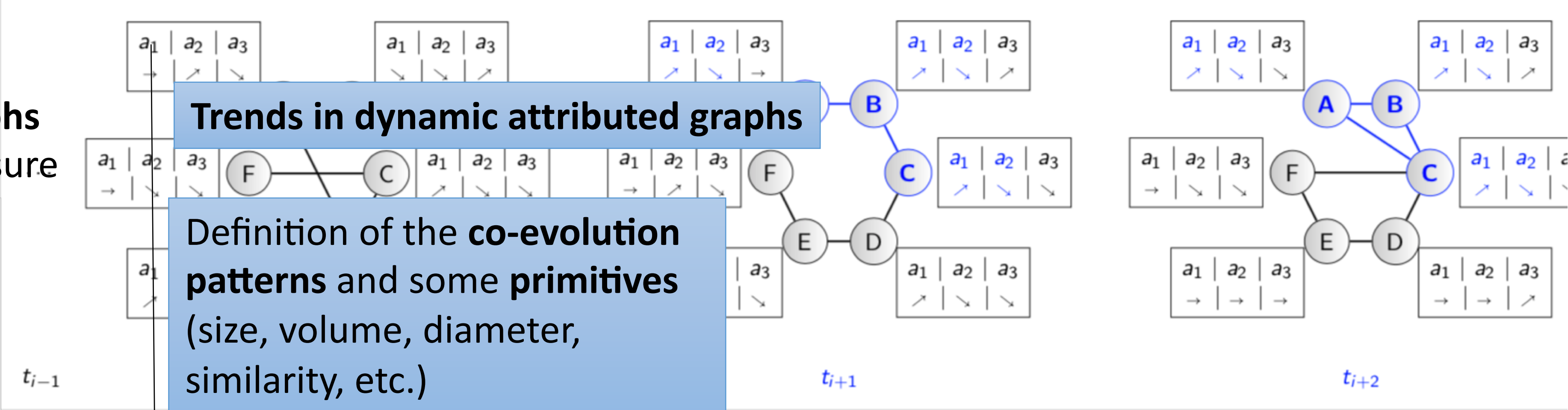

**Step back:**

- A semantics too difficult to apprehend
- Needs deeper investigation to be fully usable in practice.




*Kim-Ngan Nguyen et al., SDM'11, IDA J 2013*

## Trends in dynamic attributed graphs




Definition of the **co-evolution patterns** and some **primitives** (size, volume, diameter, similarity, etc.)

***Vertices*** *(fulfilling some properties) that follow the same* ***trends*** *on some attributes at t*

**Step back:**

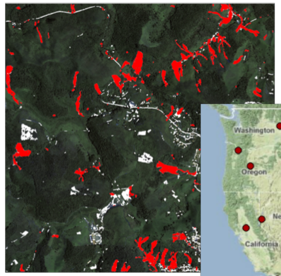
- Sometimes unnecessarily over-specified

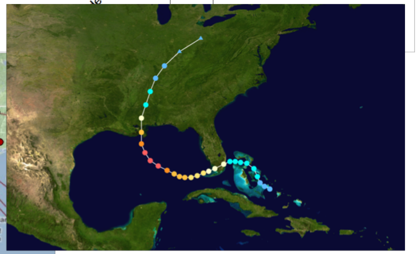
=> High assimilation cost, misleading.

**Interestingness** outside densities.

Taking into account **hierarchies** on the vertex attributes.


Applications:
Soil erosion,
transportation network analysis.
ANR FOSTER





*Desmier et al., DS'12, ECMLPKDD'13, IDA'14*

# Link between structure and vertex attributes

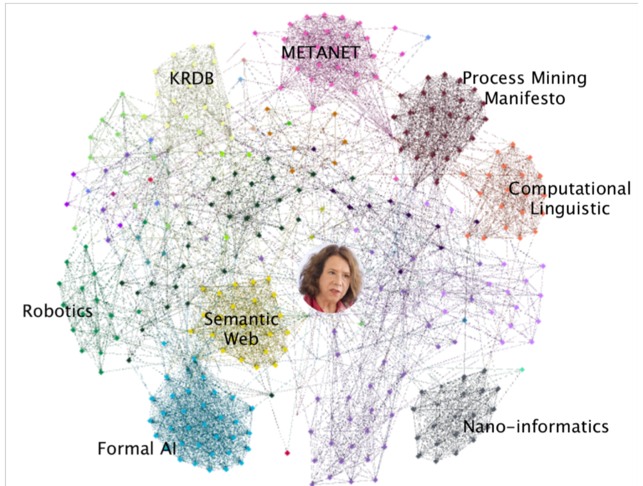
*How the graph structure impact the vertex attributes?*
*Are the vertex attributes  correlated  with the vertex role within the graph?*

Two Pattern domains to provide insights to these questions.

**Topological patterns**

- Based on a generalization of the Kendall's Tau *(Calders et al., KDD'06)*
- Rank-correlation between vertex attributes and topological attributes
- *The higher the number of publications in Dami, KAIS, and EGC, the lower the Morik number.*

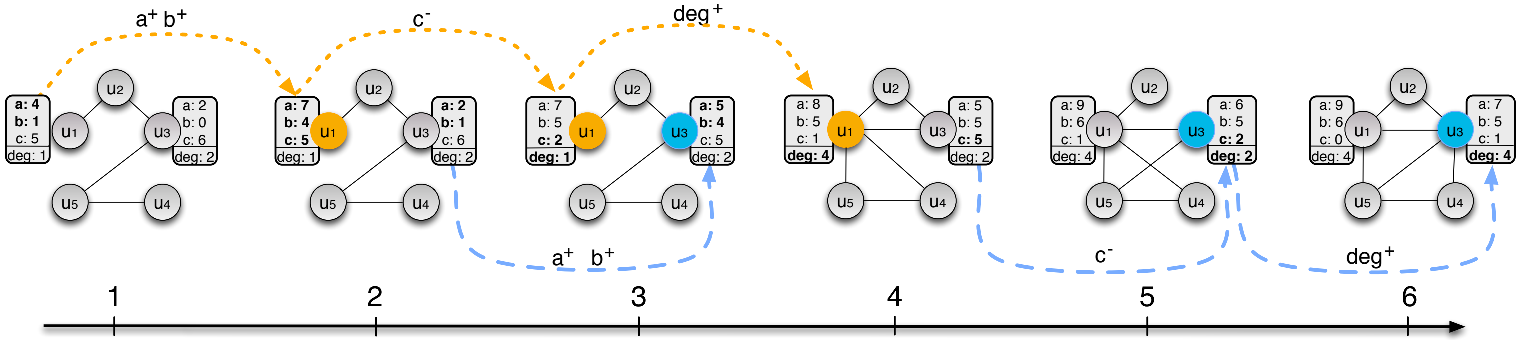
*Prado et al., IEEE TKDE 2013*




**Triggering patterns**

- Discovering the sequences of vertex attribute variations that impact the structure of the graphs
- <ICDE+, Journal++> → betweenness ++

*Kaytoue et al., SNAM 2015*



# Step back: Do we fill the gap between the user and her data ?

Pattern domains to discover **new insights**

Complexity of the **data**

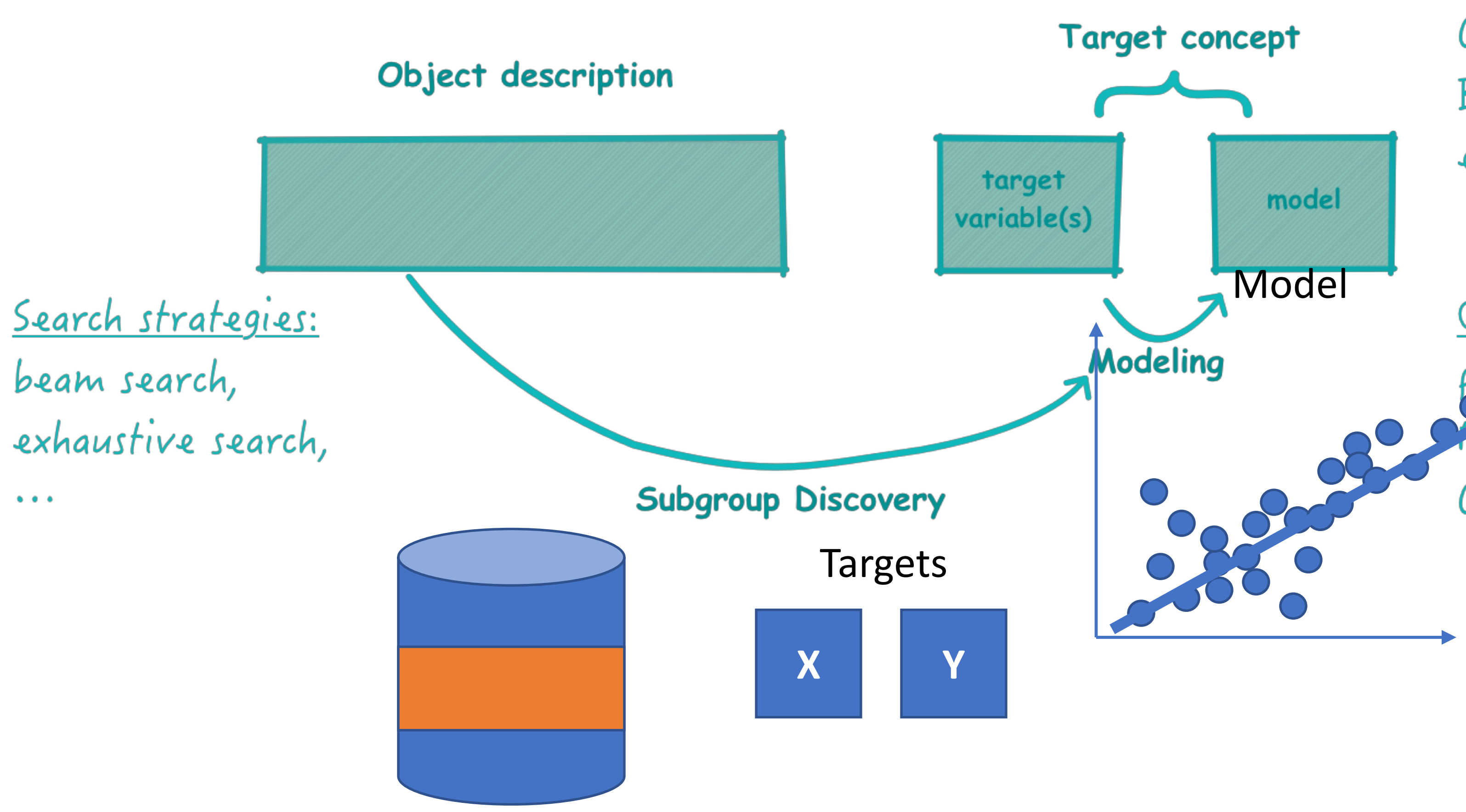

Big effort (post-processing of the pattern collection) for the user to find the good knowledge nuggets

Complexity of the user

Complexity of the domain

Complexity of the output

# EMM/SD as a basis to handle all the complexities (data, user, domain, output)





*W Klösgen 1996*
*Leman et al., ECMLPKDD'08*

# EMM/SD for graphs: challenges

- Which targets and models ?
- What about the complexities ?
  - Domain
  - User
  - Output

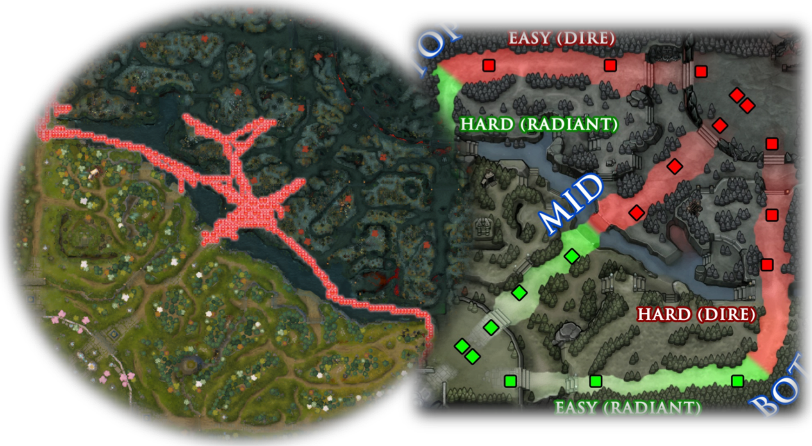
Graphs and/or attributes ⚠ $2^{targets}$

Into the model

Biased measures according to feedback

Subjective interestingness

**Edge-attributed graphs**
**Vertex-attributed graphs**




# Exceptional subgraphs in edge attributed graphs

*Kaytoue et al., Machine Learning 2017*

*Bendimerad et al., Complex Network'18*

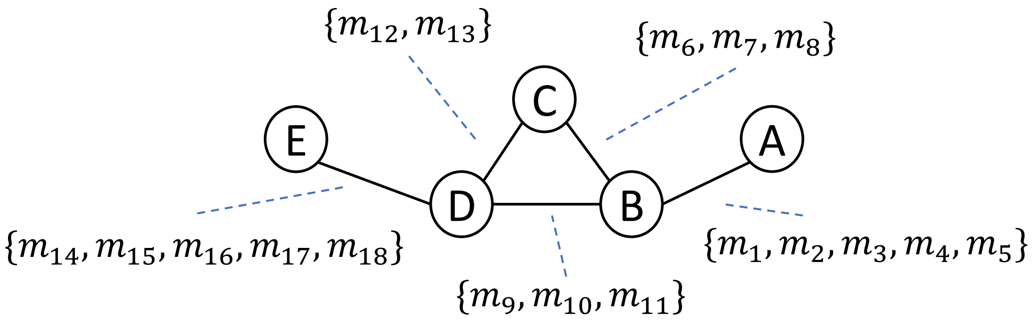
# Aims

## Edge-attributed subgraphs

*G=(V,E,T, Edge)*

$$Edge: T \rightarrow E$$

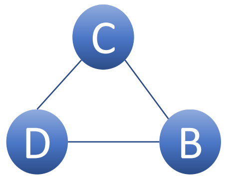
- A set of transactions is associated to an edge.
- Alternatively:
  - Several edges between pair of vertices.
  - Each edge is provided with a **context** that depicts the interaction.




| | Time | Weather | Gender | Age |
|---|---|---|---|---|
| $m_1$: | Day | Rainy | F | 20 |

## Objectives

Discover a **subgraph** and a **context** such that the subgraph is **exceptional** according to the context.




$Age \in [20,23]$,
$Time = Night$

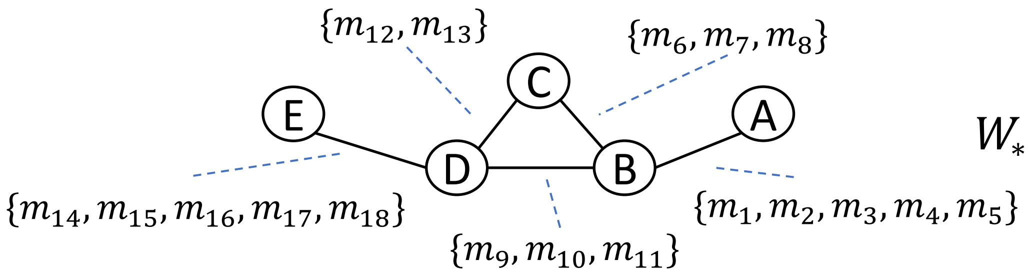
- Which model to capture it?
- Which quality measure?
- Which pattern language?
- How to extract the patterns?

Exceptional contextual subgraphs

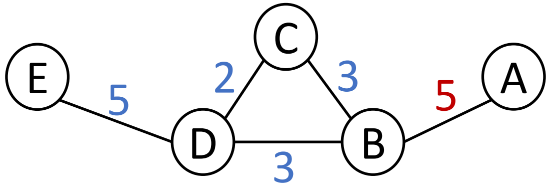
# Model: contextual graphs

Context *

| Time | Weather | Gender | Age |
|---|---|---|---|
| * | * | * | * |



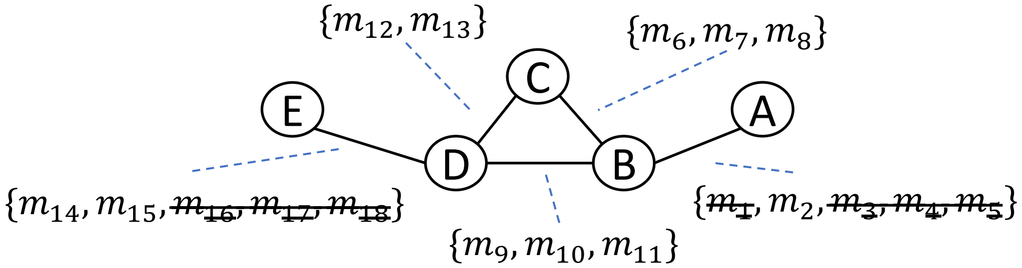

$$W_*(a, b) = |\{m_1, m_2, m_3, m_4, m_5\}|$$



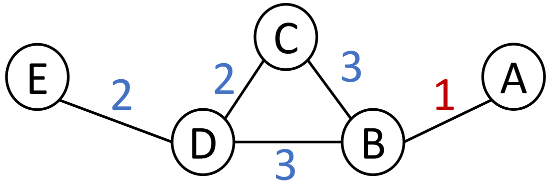

Which edges are surprising?

Context C

| Time | Weat. | Gend. | Age |
|---|---|---|---|
| Night | * | * | [20-23] |




$$W_C(a, b) = |\{m_2\}|$$



# Quality measure to measure the exceptionality

$WRAcc$ measure to assess the exceptionality of $C$ for an edge $e$:

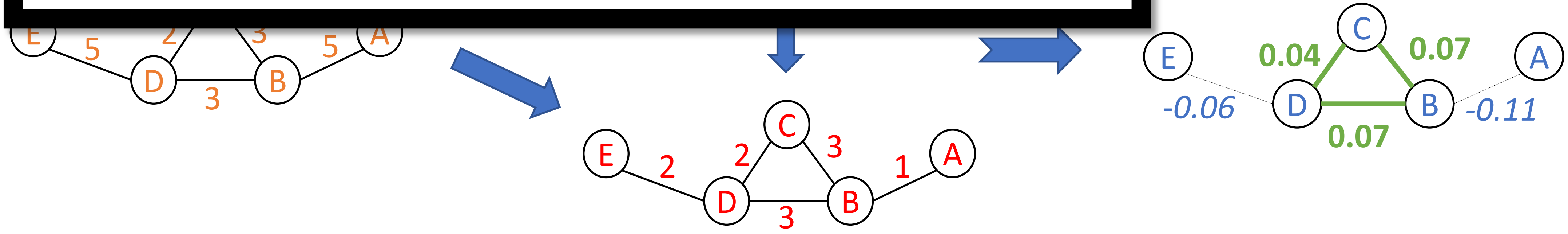
$$WRAcc(C,e) = \frac{1}{|T|}\left(W_C(e) - \bar{W}_C(e)\right)$$

The contexts are assumed to be independent.

$\bar{W}_C(e)$ is the expected weight: $\bar{W}_C(e) = W_*(e) \times \frac{W_C(E)}{|T|}$

$$\begin{aligned}
\textsc{Except}(C,e) \quad &\equiv \quad e \in M_{C\to G}(e) && (4.1)\\
&\textit{and} \quad |M_{C\to T}(C, M_{G\to T})| > min_weight && (4.2)\\
&\textit{and} \quad X^2(C,e) > \chi^2_{0.05} && (4.3)\\
&\textit{and} \quad \textsc{WRAcc}(C,e) > 0 && (4.4)
\end{aligned}$$

| $\lvert T\rvert$ | $W_C(e)$ | $W_*(e)$ |
|---|---|---|
| The total number of trips | The total number of trips in $C$ | The total number of trips in $e$ |



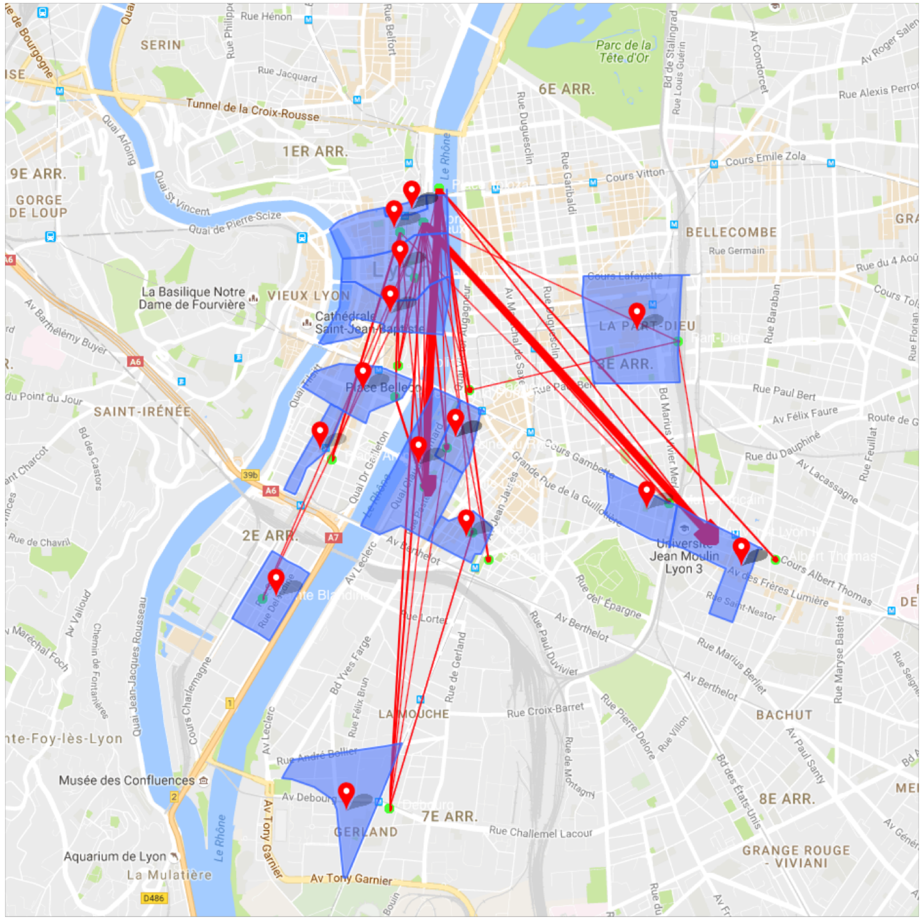
# Exceptional contextual graph mining problem

**Problem 4.1** (The Exceptional Contextual Graph Mining Problem)**.** *Extracting meaningful patterns from an augmented graph $G = (V, E, T, \textsc{Edge})$ is achieved by computing the theory:*

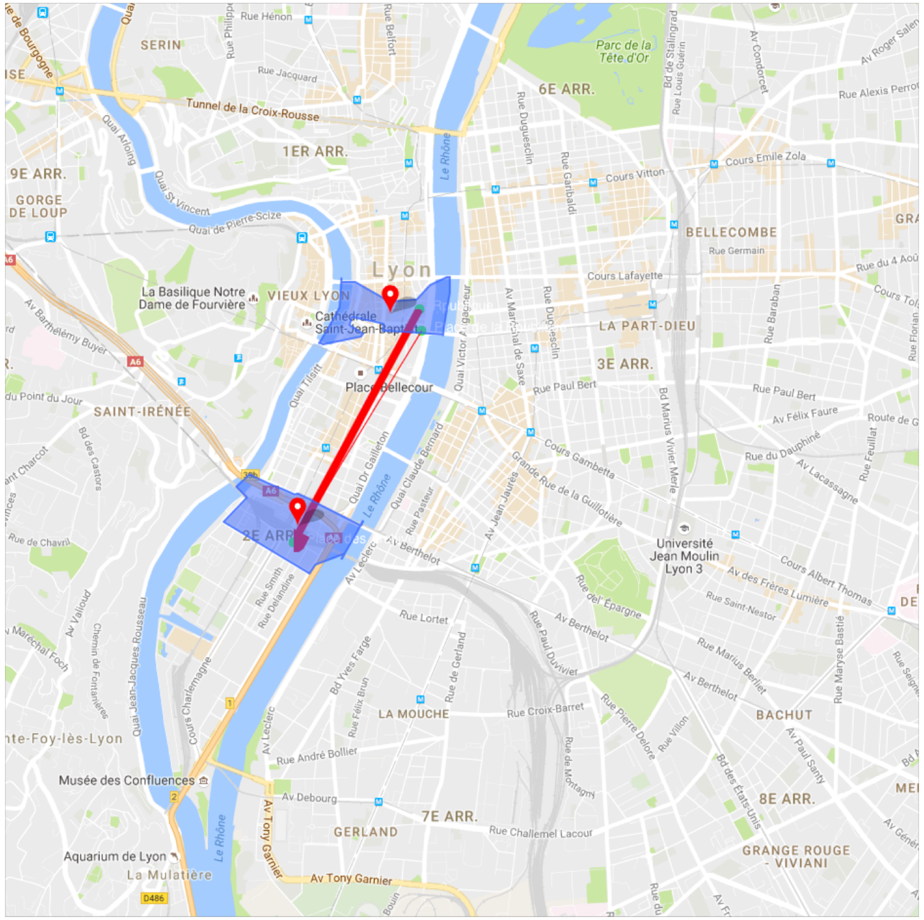
$$
\begin{aligned}
\{(C, CC_C) \quad | \quad & CC_C = (V_{CC}, E_{CC}) \textit{ is a maximal connected components of } G_C \\
& \textit{with } G_C = (V, \{e \in E \mid \textsc{Except}(C, e) \textit{ is true}\}) \\
& \textit{and } C \textit{ is closed} && (4.5) \\
& \textit{and } |V_{CC}| \geq min\_vertex\_size && (4.6) \\
& \textit{and } |E_{CC}| \geq min\_edge\_size && (4.7) \\
& \textit{and } \sum_{e \in E_{CC}} (\textsc{WRAcc}(C, e)) \geq min\_sum\_wracc && (4.8) \\
\}
\end{aligned}
$$

# Examples on Velo'v network

(gender = M, age = [14,26], *univ_in, univ_out*)



(zip_code = 38, gender = M, age  = [26,60], pass = oura)



# Including domain knowledge

$$WRAcc(C,e) = \frac{1}{|T|}\left(W_C(e) - \overline{W}_C(e)\right)$$

$\overline{W}_C(e)$ is the expected weight:

*previously*: $\overline{W}_C(e) = W_*(e) \times \frac{W_C(E)}{|T|}$

**Importance of the areas, distances are not taken into account!**

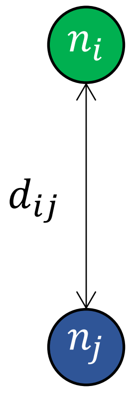
**Mobility models make it possible!**

$$\overline{W}_C(e) = \boldsymbol{m}(\boldsymbol{e}) \times \frac{W_C(E)}{|T|}$$

$\boldsymbol{m}(\boldsymbol{e})$ refer to the gravity model $\mathbf{g}(e)$ or the radiation model $\mathbf{r}(e)$.

The gravity model:

$$g(e_{ij}) = n_i \times n_j \times f(d_{ij})$$




$n_i$ and $n_j$ are respectively the populations of $v_i$ and $v_j$

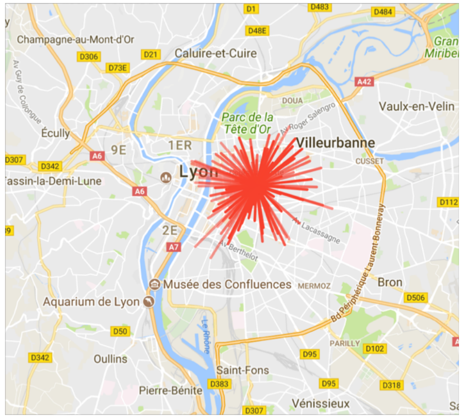
$d_{ij}$ is the distance between $v_i$ and $v_j$, and $f(d_{ij})$ represents the influence of the distance.

# Expected weights are impacted

Example: a station located in Part Dieu

Classic model

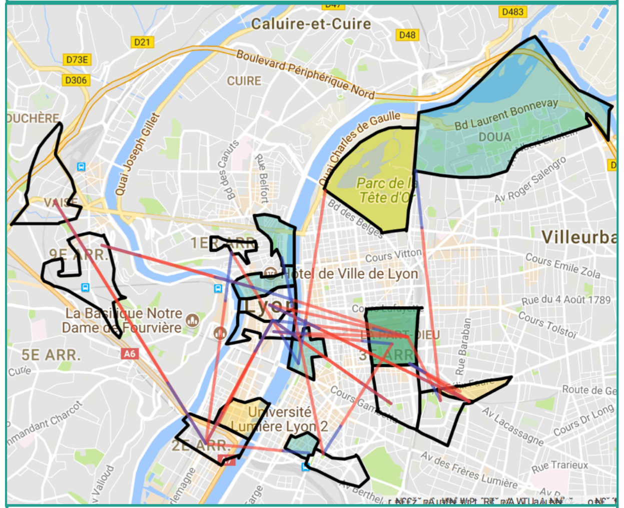

Gravity model




Radiation model

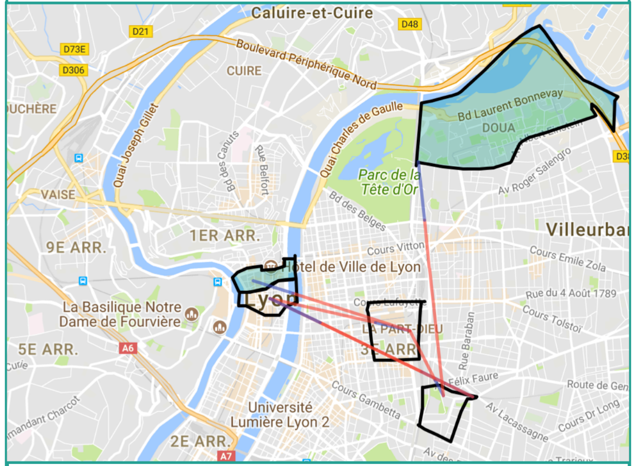
# Examples

Only by mobility models




Context : * (all)

By the three models



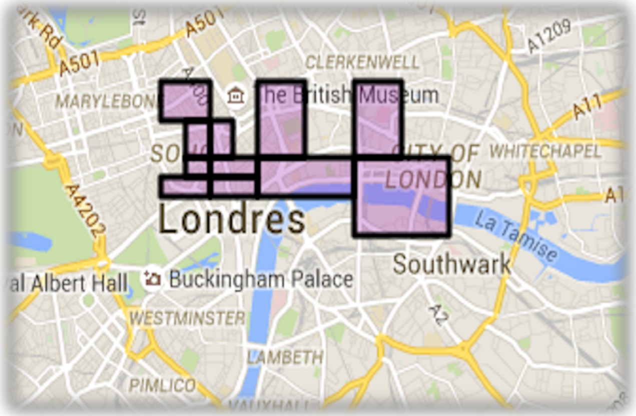

Context : $type = tecely$, $age \in [26,60]$

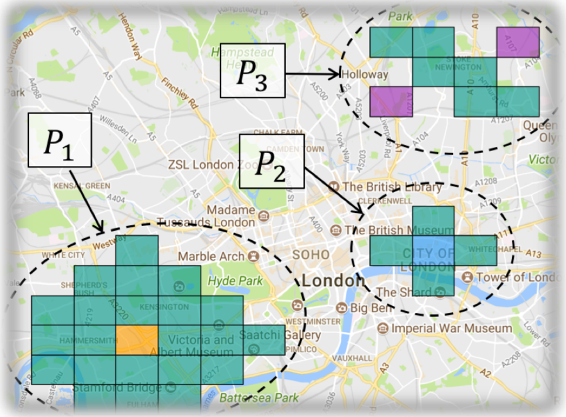
# Step back

- Need to provide instant results:

output space sampling method.

- How to better present the results to user?






# Exceptional subgraphs in vertex attributed graphs

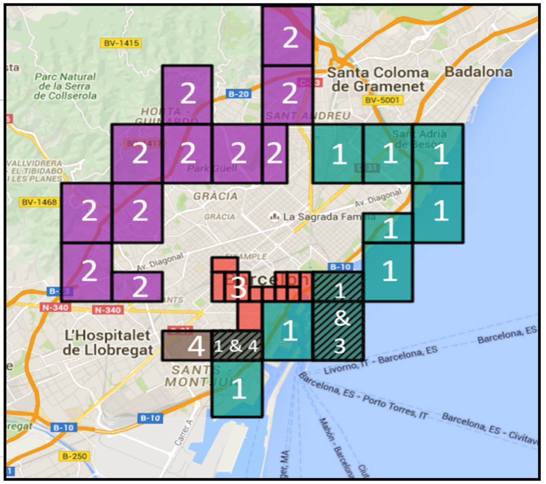
*Bendimerad et al., ICDM’16, KAIS 2018, MLG’18*

*Moranges et al., DS’18*

# Aims

- ➢ Discover subgraphs whose vertices have exceptional values on some attributes.
- ➢ *Identify meaningful neighbourhood and describe them with their characteristics.*

Foursquare venues


| | | |
|---|---|---|
| 1 | + | Outdoors & Recreation |
| | - | Shop & Service, Professional places |
| 2 | + | Outdoors & Recreation, Universities |
| | - | Food |
| 3 | + | Nightlife Spot, Food |
| | - | Professional and other places |
| 4 | + | Outdoors & Recreation, Events, Art… |
| | - | Shop & Service, College and universities… |

- Which model to capture exceptionality?
- Which quality measure?
- Which pattern language?
- How to extract the patterns?

Exceptional subgraphs

# Capturing exceptionality

## Characteristic

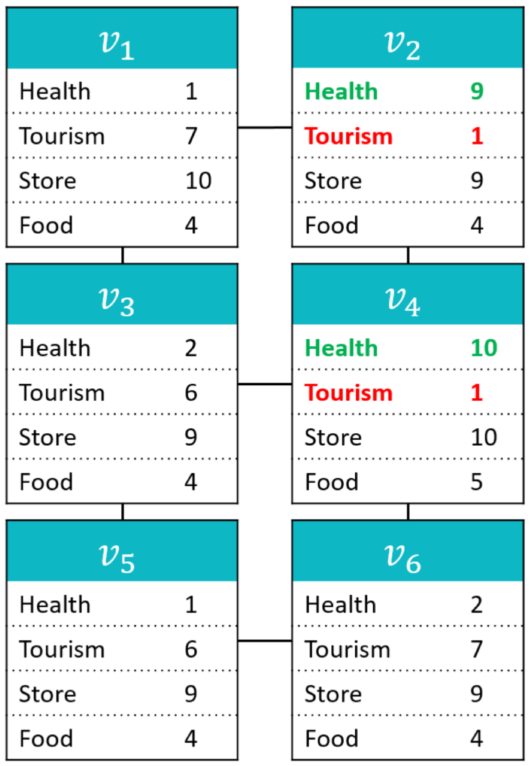
A **characteristic** is $S = (S^+, S^-)$ where $S^+$ and $S^-$ two disjoint subsets of $A$.

- $S^+$: positive trends.
- $S^-$: negative trends.

**Example:**

- $U = \{v_2, v_4\}$
- $S = (S^+ = \{Health\}, S^- = \{Tourism\})$

But how can we measure the relevance of $S$ for $U$?




Expected values

4.2
4.8
9.6
4.29

WRAcc measure to assess how unexpected the observed values are.

$$WRAcc(S,K) = \begin{cases} A(S,K) \times \frac{sum(K)}{sum(V)} & \text{if } valid(S,K) \\ 0 \text{ otherwise} \end{cases}$$

# Exceptional subgraph mining problem

General problem

Given a graph $\mathbf{G} = (\mathbf{V}, \mathbf{E}, \mathbf{A})$, and two thresholds $\sigma$ and $\delta$, discover all exceptional sub-graphs $(\mathbf{U}, \mathbf{S})$ such that:

1. $|U| \geq \sigma$
2. $G[U]$ is connected
3. $WRAcc(U, S) \geq \delta$

Problem variants

- Closed exceptional subgraph to address redundancy issues

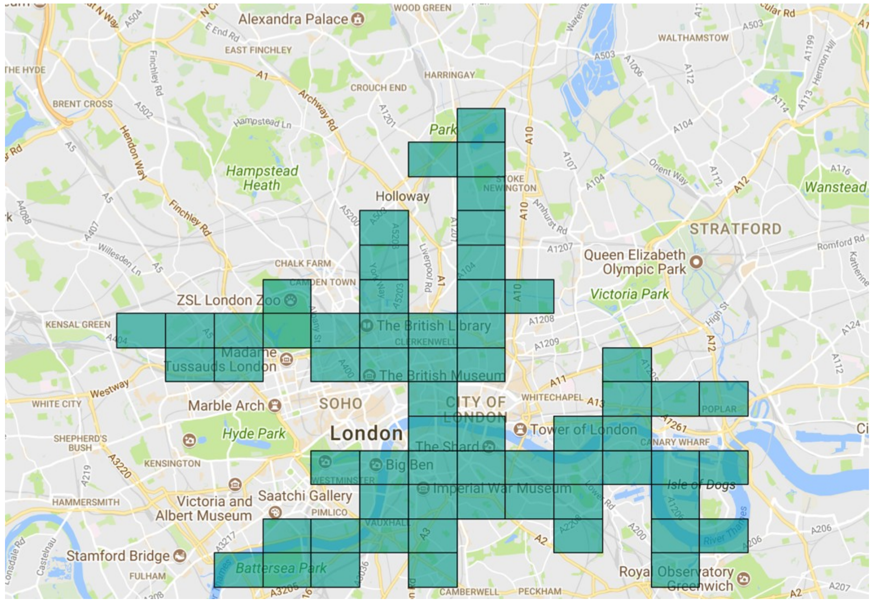
CENERGETICS algorithm which fails with hundreds of attributes.

- Provide a sample of the output
- Output space sampling

EXCESS algorithm

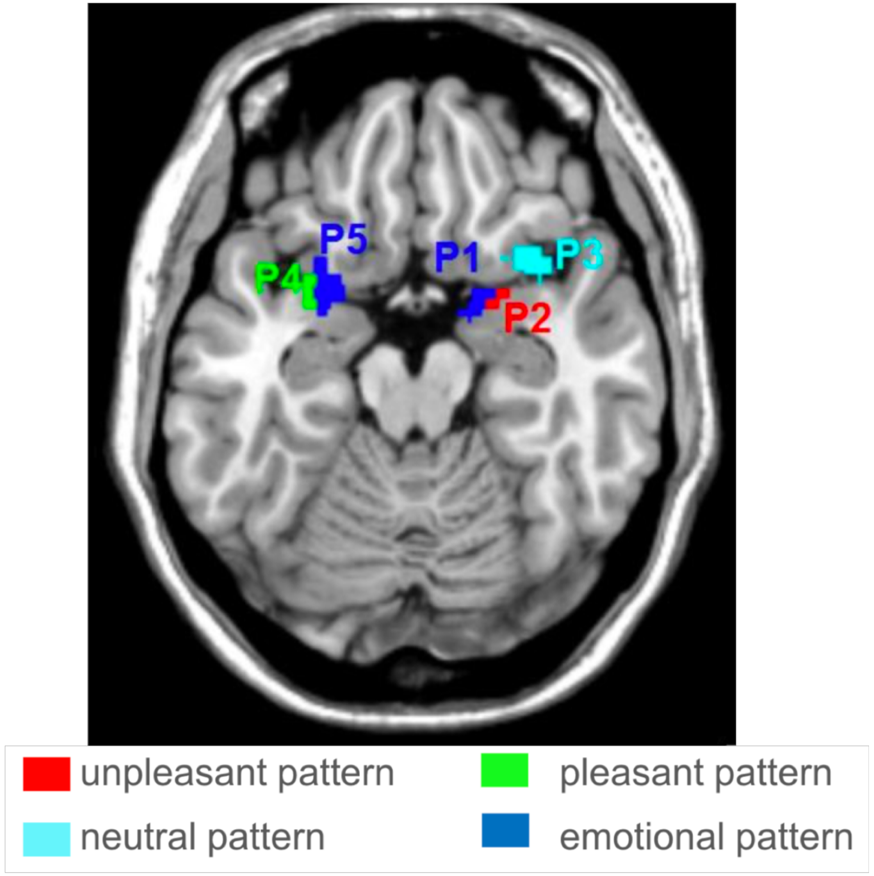
# Examples

## To describe and analyse cities




Professional+, shop-

## To understand the olfactory percept



# What about the user?

She can provide feedback about the patterns:

- ➢How to take benefit from this feedback?
- ➢Without changing the algorithm?

She has some priors about the data:

- ➢How to model these priors?
- ➢Use them to find really interesting patterns (according to her priors).

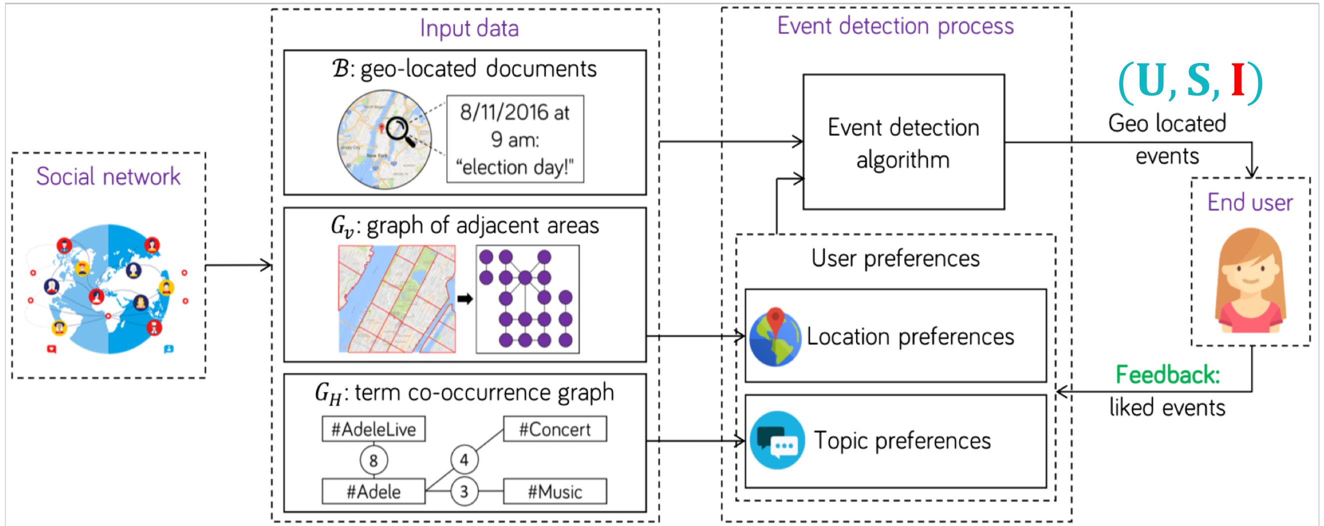
Taking into account user feedback into biased quality measures.

- ✓ Application to geolocated event detection on Twitter.

✅ Mining subjectively interesting attributed subgraphs

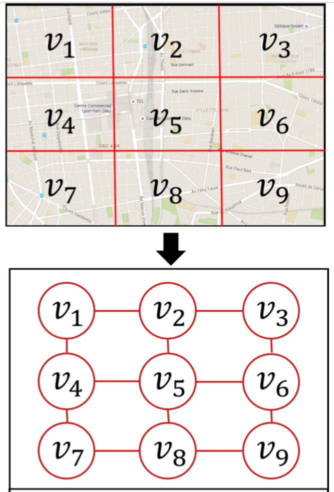
- ✓ MaxEntropy model to assess the interest of pattern.
- ✓ Trade-off between information content and pattern assimilation.
- ✓ Updating the model.

# Unified framework for data-driven and user driven geolocated event discovery




$$M_u(P) = \sum_{h \in H_p} \sum_{v \in K} \sum_{t \in I} score(h, v, t) \times \left( \frac{Q_h(h) + Q_v(v)}{2} \right)$$

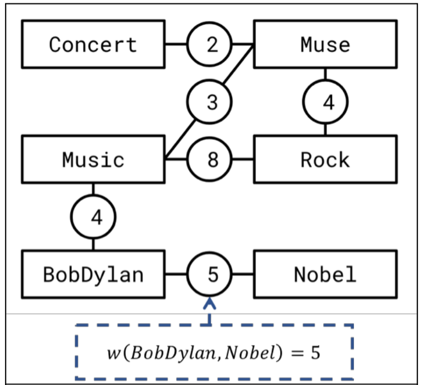
$Q_h : H \rightarrow [1, maxPref]$ and $Q_v : V \rightarrow [1, maxPref]$ expresses respectively the **interest** of the term $h$ and the vertex $v$ to the **user** (and $maxPref > 1$).




The graph $G_v = (V, E)$

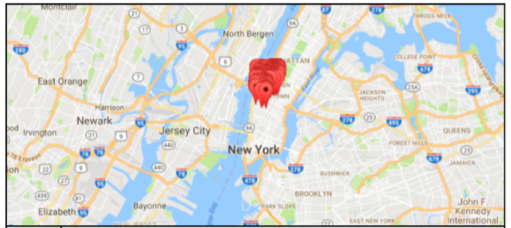
$$Q_v(v) = \alpha \sum_{v' : (v, v') \in E} \frac{1}{\deg(v)} \times Q_v(v') + (1 - \alpha) \times B(v)$$

$$Q_h(h) = \alpha \sum_{h'} \frac{w(h, h')}{\deg(h)} \times Q_h(h') + (1 - \alpha) \times B(h)$$



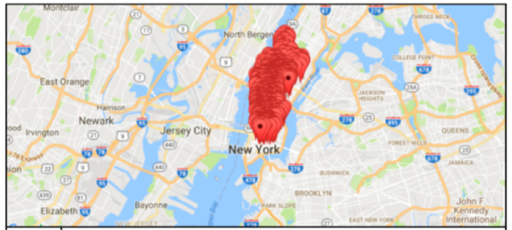

$G_H$ the term co-occurrences graph

## New York



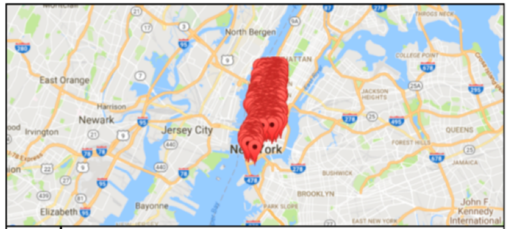
**P1** From 8 Oct 2016 at 6h, to 10 Oct 2016 at 00h

Terms: #nycc, #nycc2016, @ny_comic_con, #cosplay, #comiccon2016, #comiccon, #newyorkcomiccon, #marvel



**P2** From 7 Oct 2016 at 14h, to 9 Oct 2016 at 14h

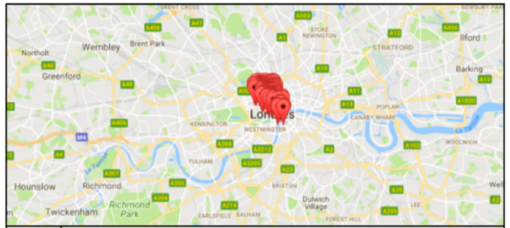
Terms: #election2016, #imwithher, #vote, #electionday, #electionnight, #ivoted, #hillaryclinton, #election



**P3** From 31 Oct 2016 at 6h, to 1 Nov 2016 at 15h

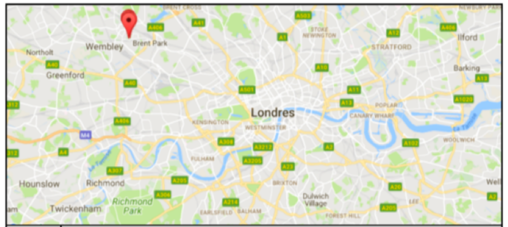
Terms: #halloween, #happyhalloween, #nyc, #halloween2016, #costume, #trickortreat, #halloweencostume, #halloweenparade

## London



**P1** From 8 Jul 2017 at 6h, to 9 oct 2017 at 4h

Terms: #pride, #londonpride, #pride2017, #prideinlondon, #loveislove, #pridelondon, #lovehappenshere, #londonpride2017



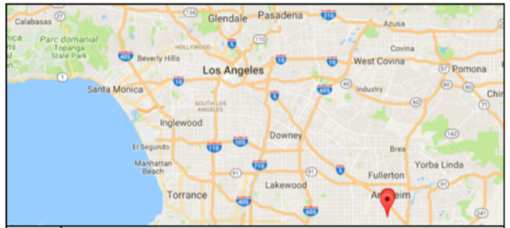
**P2** From 27 May 2017 at 8h, to 28 May 2017 at 14h

Terms: #arsenal, #facup, #facupfinal, #coyg, #chelsea, @arsenal, #facup2017, #emiratesfacup

**P3** From 8 Jul 2017 at 12h, to 10 Jul 2017 at 2h

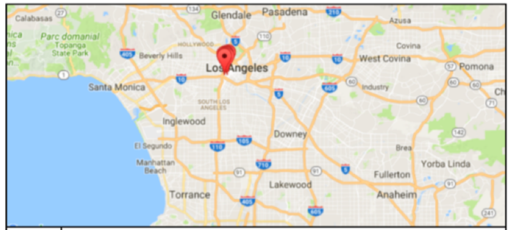
Terms: #u2, #u2thejoshuatree2017, @u2, #joshuatree, #twickenham, #bono, #twickenhamstadium,#joshuatreetour2017

## Los Angeles



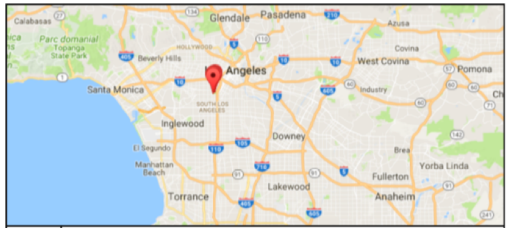
**P1** From 14 Jul 2017 at 6h, to 15 Jul 2017 at 00h

Terms: #d23expo, #d23, #d23expo2017, #disney, #disneylegends, #marvel, #ducktales, #starwars



**P2** From 14 Jun 2017 at 6h, to 15 Jun 2017 at 00h

Terms: #e3, #e32017, #nintendo, #playstation, #ps4, #gaming, #xbox, @e3, #videogames, #capcom



**P3** From 23 Jul 2017 at 6h, to 24 Jul 2017 at 4h

Terms: #fyf, #fyffest, @fyffest, #nin, #fyffest2017,#frankocean, #carmonaphotography, #fyf2017

- The data-driven method outperforms state-of-the-art methods
- The user-driven method was assessed through an evaluation from the crowd (crowd flower).

# Subjectively interesting subgraphs

The user may have some priors about the data.

*"For each type, I know the number of places."*

*"For each area, I know the number of places."*

Background Knowlege

Equality Constraints

MaxEnt model

Model updating

User has to assimilate the patterns

*"Please, give me patterns interesting and easy to assimilate."*

Assimilation cost

Information Content
IC(U,S)= - log(Pr(U,S))

Description Length
DL(U,S)= $DL_A$(S) + **$DL_V$(U)**

Alternative description easier to assimilate

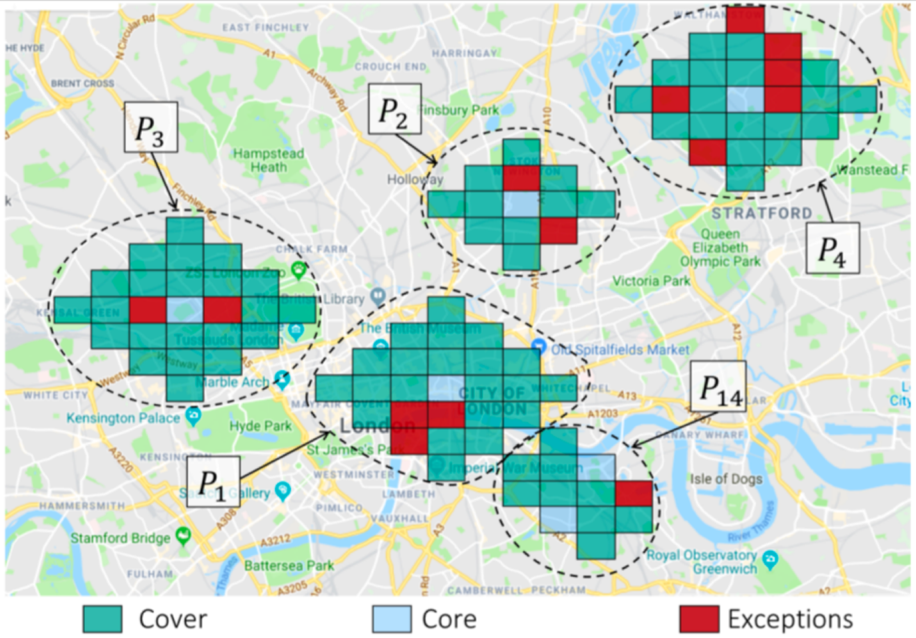
*The vertices that are at a distance of at most 2 of the vertex  A and the vertex B.*

The **subjective interestingness** as a trade-off between IC and DL:

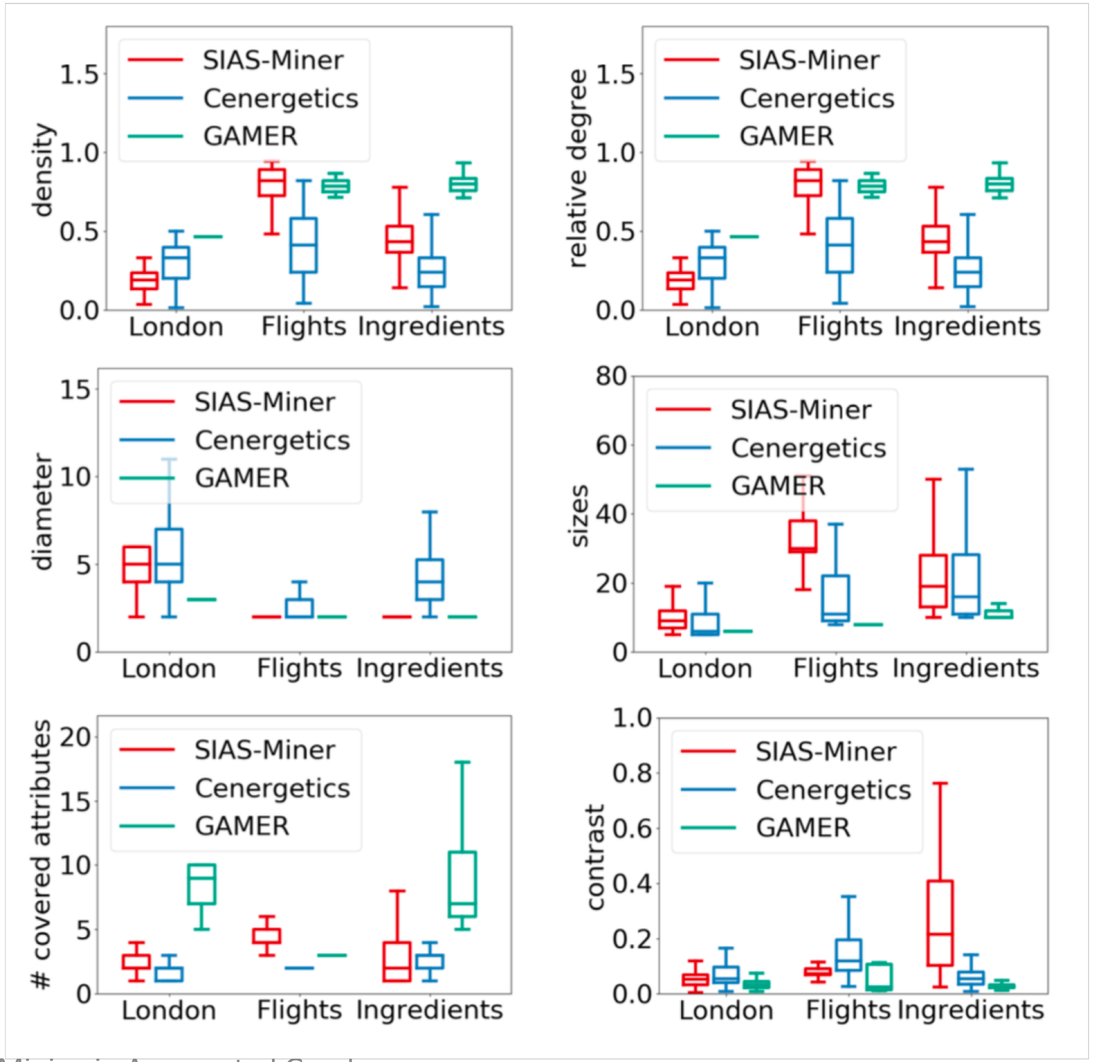
$$SI(U,S) = \frac{IC(U,S)}{DL(U,S)}$$

**SIAS-Miner algorithm**

# SIAS Miner: Examples




| ID | Characteristics: Exceptionally prevalent | Exceptionally non-prevalent |
|---|---|---|
| $P_1$ | nightlife, professional | |
| $P_2$ | nightlife, food | college |
| $P_3$ | food | |
| $P_4$ | food | college |
| $P_{14}$ | professional | college, event |



# Step back: have we filled the gap between the user and her data ?

Not yet, but we are in the good direction!

- ✓ Complexity of the data
- ✓ Complexity of the domain
- ✓ Complexity of the user
  - ✓ Background knowledge
  - ✓ User feedback
- ✓ Complexity of the output

- ➢ Incorporating of non-ordinal attribute types
- ➢ Integrating other kinds of prior beliefs (e.g., correlation)

# Conclusion and Future Directions

# Conclusion: take away message

✅ Augmented graphs as a powerful way to model real-world phenomena.

⚠️ Take care of all the complexities:

- data, user, domain and output

This is the only mean to provide actionable insights and boost human knowledge.

# Is there a future for pattern mining ?

**YES** … if we

- Democratize the pattern mining tools (Knime, Weka, Python libraries)
- Make them easily usable

Analysts will always need of descriptive analysis techniques:

- Crystal clear descriptive solutions


**Still some issues to tackle!**

# Improve pattern rendering

**Data Mining meets Visualization and Information Retrieval.**

- Highlight the interest of the pattern with respect to:
  - the user priors,
  - the domain knowledge,
  - ***how this pattern stands out from the others***.

⚠ Decades of investigation done in the **Visualization research community**.

- Same observation with the **Information Retrieval** community.
  - (re-)ranking results, taking into account user interest and her satisfaction, recommending are at the heart of the studies from this community.

# Causality

- Users tend to assimilate patterns/rules with a causality point of view.

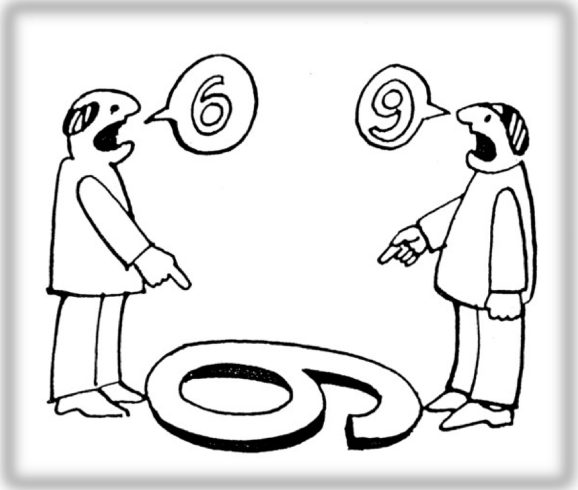
Enthusiasm  is quickly followed by disappointment.

- Pattern mining will bring as much as hope as disappointment as long as the problem of **causality** is not solved.

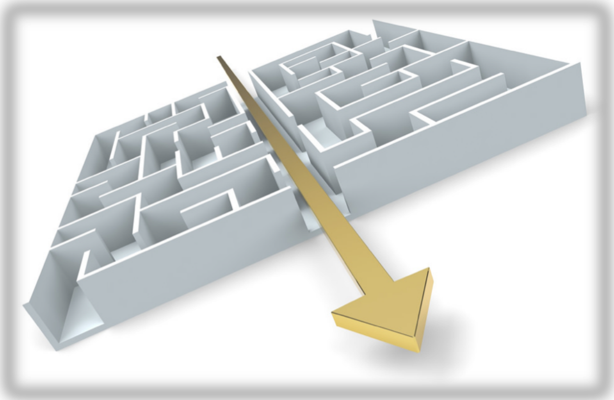
- A timely challenge !

# Pattern mining as the corner stone of data science projects

- Pattern mining to foster **interdisciplinarity**.
- The pattern syntax can be easily understood by any scientist.
- Domain knowledge of each discipline can be integrated.
- Discovered patterns as an excellent support for discussion between the scientists from different disciplines.
- **ROI**: such projects also provide new challenges in data mining.







## Describe as simplest as possible!

*Towards multiple-pattern domains pattern set mining.*

**Strong assumption**: the user knows the good pattern domain.

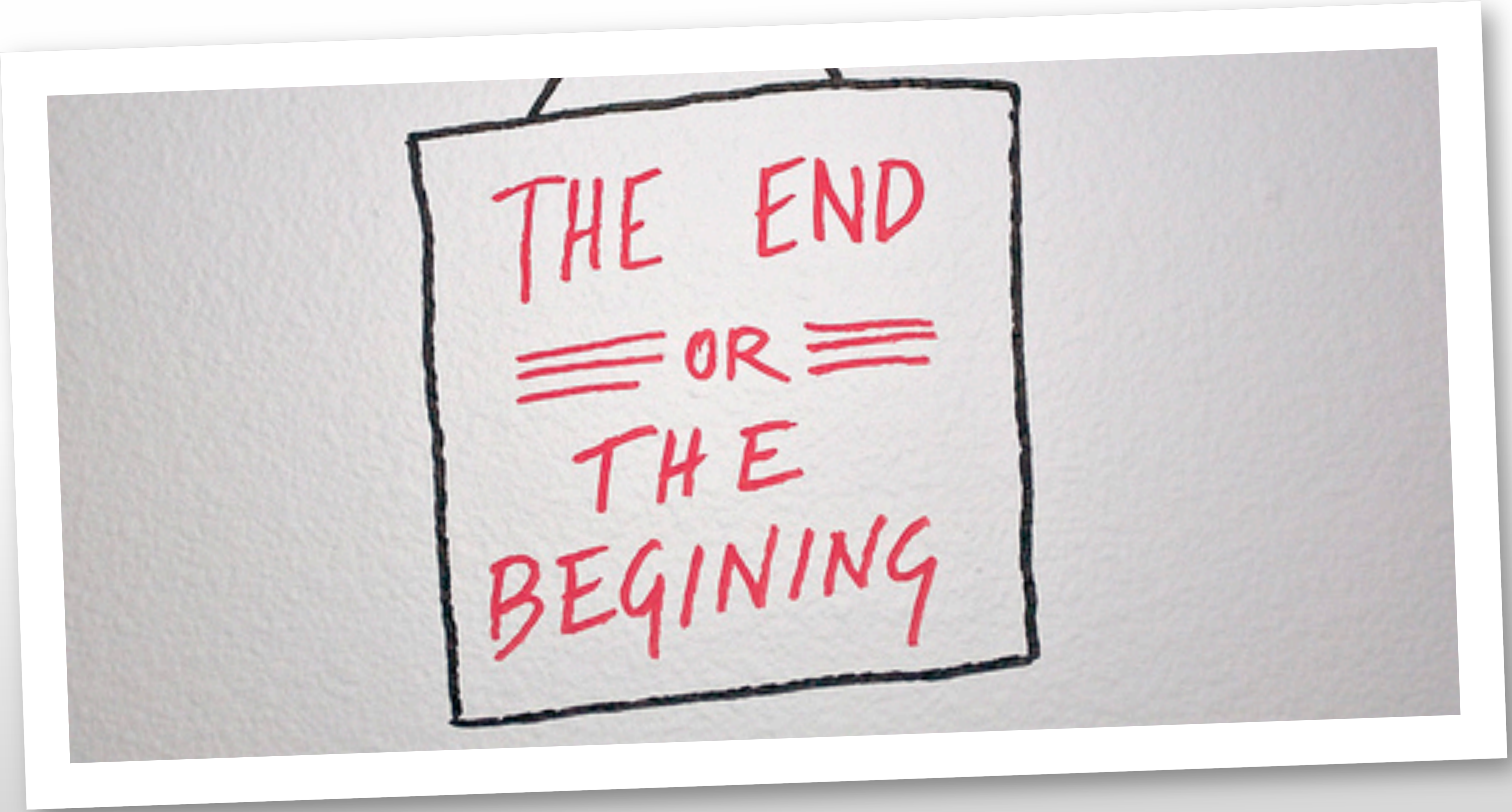
⚠ A too complex pattern domain: over-described results and misleading interpretation.

⚠ A too simple pattern domain: impossible to describe some complex phenomena.

- ➢ Automatically find the good level of description (pattern syntax).
  - ➢ Itemsets (w/|w/o) numerical values, sequence, graphs, 3d graphs, …

# Data Mining meets Machine Learning: Towards sparse and interpretable Deep Neural Networks

- Obvious need of effective predictive models
- Work into the models to understand / simplify them
- Work on the I/O to both understand and improve the models.
- Investigate languages to characterize them.



THE END
OR
THE
BEGINING


Acknowledgements