Data Mining and Visualization

Université de Rennes, M2 SIF

Alexandre Termier

Peggy Cellier

Ferran Argelaguet

Luis Galarraga

Tassadit Bouadi

General information

Organization

- 5 teachers
  - Alexandre Termier (data mining), Alexandre.Termier@irisa.fr
  - Peggy Cellier (data mining), Peggy.Cellier@irisa.fr
  - Ferran Argelaguet (visualization), Ferran.Argelaguet@inria.fr
  - Luis Galarraga (interpretability), Luis.Galarraga@inria.fr
  - Tassadit Bouadi (interpretability), Tassadit.Bouadi@irisa.fr

- Location
  - ISTIC
  - Students from Rennes (M2 SIF + CNI + DigiSport) and Lannion (M2 SIF)

- 21 hours of course
  - 14 x 1h30
  - Detailed schedule next slide

Tentative schedule *(subject to change)*

Date	Day of week	Contents	Instructor	Room
10/09, 15h	Tuesday	Introductory course - KDD 101	A. Termier	Guernesey
10/09, 16h45	Tuesday	Frequent itemset mining (1/2)	A. Termier	Guernesey
13/09, 16h45	Friday	Frequent itemset mining (2/2)	A. Termier	Guernesey
17/09, 15h and 16h45 (3 hours)	Tuesday	Introduction to data visualization	F. Argelaguet	Guernesey
27/09, 16h45	Friday	Subgroup discovery	A. Termier	Guernesey
01/10, 15h	Tuesday	Introduction to Explainable AI	L. Galarraga	Guernesey
11/10, 16h45	Friday	Sequence mining	P. Cellier	Guernesey
15/10, 15h	Tuesday	Pattern mining with deep learning	A. Termier	Guernesey
18/10, 16h45	Friday	Graph Mining	P. Cellier	Guernesey
22/10, 15h	Tuesday	Pattern-sets and Information theory based pattern ming	P. Cellier	Guernesey
25/10, 16h45	Friday	Explainable AI: Counterfactuals	T. Bouadi	Guernesey
05/11, 16h45	Tuesday	Paper presentations	P. Cellier and A. Termier	Guernesey
08/11, 16h45	Friday	Paper presentations	T. Bouadi, L. Galarraga and F. Argelaguet	Guernesey
12/11, 13h15	Tuesday	Exam	A. Termier	Guernesey

Web site of the course

- http://people.irisa.fr/Alexandre.Termier/dmv/

- Web site contains:
  - General information
  - Up-to-date schedule (it is the reference)
  - Links to documents

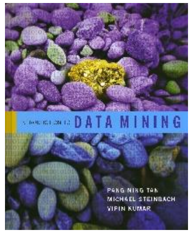
Evaluation

- Standard 1h30 exam
    - Expectations:
        - Understanding of the approaches/algos presented in the course
        - Ability to tackle a KDD problem
        - Capacity to think « out of the cookbook »
    - Documents allowed

- Paper presentation
    - Read and understand a scientific paper (group of two)
    - Make a 10 minutes presentation

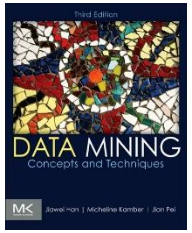
Books

- Introduction to Data Mining, Tan et al. http://www-users.cs.umn.edu/~kumar/dmbook/index.php



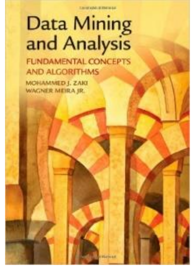
- Data Mining: Concepts and Techniques, Han et al.

http://web.engr.illinois.edu/~hanj/bk3/



- [FREE] Data Mining and Analysis, Zaki and Meira

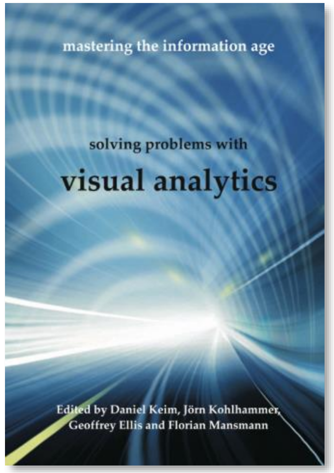
http://www.dataminingbook.info/pmwiki.php/Main/BookDownload

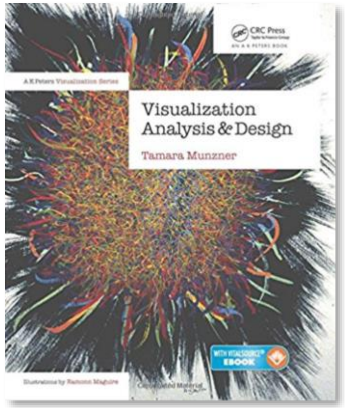


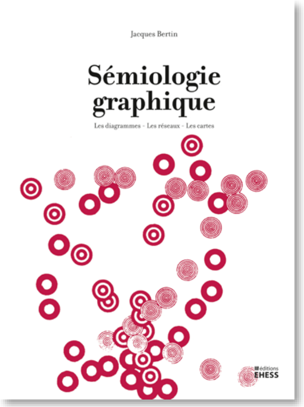
- Frequent Pattern Mining, Aggarwal and Han Edt.
  - Some free chapters online, ex: http://eda.mmci.uni-saarland.de/pubs/2014/fpmbook_int-vreeken,tatti.pdf

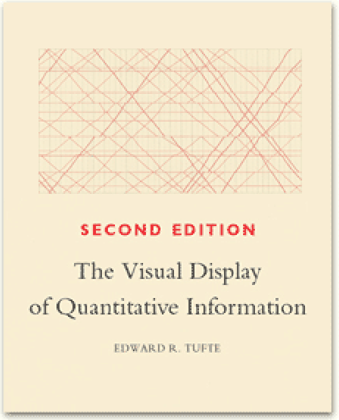
Books, contd.

- [FREE] VisMaster – Solving problems with visual analytics http://www.vismaster.eu/book/
- Visualization Analysis and Design, Munzner
- The Visual Display of Quantitative Information, Tufte
- Sémiologie graphique, Bertin


mastering the information age
solving problems with
visual analytics
Edited by Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis and Florian Mansmann







Introduction to the DMV course

Why this course?

- Data driven world
- Need to make sense from data
  - Exploit data for tasks we can do -> machine learning (supervised)
  - Find hidden knowledge in data -> data mining (unsupervised)

- KDD = Knowledge Discovery from Data
- In this course we will focus on pattern mining
  - Pattern mining = finding some kind of regularities in data

- Need to present results to users -> data visualization
  - Huge lack of communication between data mining / data viz community
  - This course: a small step to improve this communication

- Need to understand the results of Machine Learning algorithms -> interpretatility
  - Provide human-understandable explanations for ML algorithms decisions

Why pattern mining?

- Actual interest of finding regularities in data
    - Will be seen throughout the course

- Research field:
    - With many unsolved problems
    - Not overcrowded (unlike D..p L…ning) !

- Researchers and practitioners need interpretable results
    - In the Data Mining field, pattern mining is an excellent example of interpretability…
    - …with some interesting pitfalls !

First…what is pattern mining ? An analogy

data	speaks a foreign language
datum	symbols
patterns	words

Question: how do I decide *what is a word* ?

「お金を下ろせない人たちが、キャッシングしてる！」

ツイッターには「みずほ 銀行からお金を下ろせない人たちが、キャッシングしてるなう！」との目撃情報も。キャッシングを利用せざるをえないという書き込みも複数 ある。

一部にはATMが使えな くなることを知らずに、障害などを起こして止まってる、と勘違いしている人もいた。

「なんで新百合にあるみずほ銀行全滅してんの？なんで封鎖されてんねん」

「池袋中を探し回ったけ どみずほ銀行のATMどこもやってない」

ただ、「忘れてたっていうのはともかく、あれだけしつこいくらい告知されてて『知らなかった』ってツィートってネタですよね？」と指摘する人も一部にはいる。

みずほ銀行のATMが止まることは、トップページや銀行の張り紙、CMなどで繰り返し告知されてきた。それでも、システムの移行作業で、ATMを含め、すべての オンラインを休止するのは最近ではないことだ。それだけに、油断していた人も多かったのかもしれない。

「お金を下ろせない人たちが、キャッシングしてる！」

ツイッターには「みずほ 銀行からお金を下ろせない人たちが、キャッシングしてるなう！」との目撃情報も。キャッシングを利用せざるをえないという書き込みも複数 ある。

一部にはATMが使えな くなることを知らずに、障害などを起こして止まってる、と勘違いしている人もいた。

「なんで新百合にあるみずほ銀行全滅してんの？なんで封鎖されてんねん」

「池袋中を探し回ったけ どみずほ銀行のATMどこもやってない」

ただ、「忘れてたっていうのはともかく、あれだけしつこいくらい告知されてて『知らなかった』ってツィートってネタですよね？」と指摘する人も一部にはいる。

みずほ銀行のATMが止まることは、トップページや銀行の張り紙、CMなどで繰り返し告知されてきた。それでも、システムの移行作業で、ATMを含め、すべての オンラインを休止するのは最近ではないことだ。それだけに、油断していた人も多かったのかもしれない。

お金　下ろせない人たちが、キャッシングして

ツイ　には　みずほ銀行　お金　下ろせない人たちが、
キャッシングして　。キャッシング
ない　いう　ある。
一部にはATM　こと　知ら　など
して止まって　している　いた。
なんで　あるみずほ銀行　して　なんで
されて
った　みずほ銀行　ATM　って
ない
っていう　だけ
されて　知ら　った　ってツィ　って
する　一部にはいる。
みずほ銀行　ATM　止ま　こと　銀行
など　されて　。それ
ATM　する
ないこと　それだけ　していた　っ
た　ない。

Regularities in data

- Pattern mining aims at extracting regularities from data
- The definition of what « regularity » is determines the patterns obtained, and the algorithm used to extract them
  - **The good:** many definitions of regularities covered in literature (next slides)
  - **The bad:** the definition you want may not be in there
  - **The ugly (part of):** most new definitions of regularity require to collaborate with a pattern mining researcher to design tractable algorithms

Frequent Itemsets

Input:

Tickets

+ 1%

Minimal expected frequency

Output: Sets of products bought frequently together in a ticket

Ex:

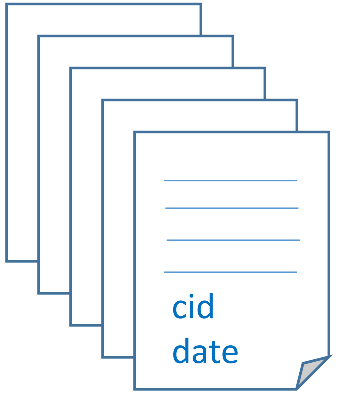
- **{nutella, baguette, Yop! fraise}** are bought together in 1.5% of all tickets

Can be enriched with taxonomy:

- **{chocolate spread, bread, drinking yoghurt}** are bought together in 13.4% of all tickets

Frequent itemsets sequences

Input:



Tickets
- of identified customers
- timestampped

+

1%

Minimal expected frequency

Output: Sequences of products bought frequently by customers over time

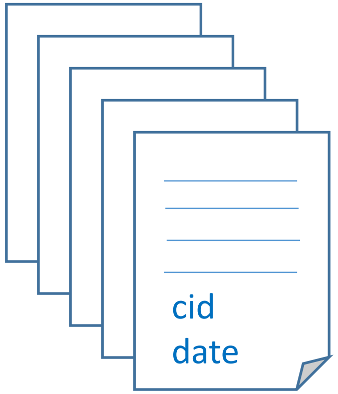
Ex:

- The sequence **{Palmolive handsoap} -> {Palmolive handsoap refill}** occurs for 4% of all known customers
- **{Top budget smoked salmon} -> {Captain Cook smoked salmon, blinis} -> {Labeyrie smoked salmon, lump eggs, blinis}** occurs for 1.1% of all know customers

Can also be enriched with taxonomy.

Frequent periodic itemsets

Input:



Tickets
- of identified customers
- timestamped

+ 1%

Minimal expected frequency

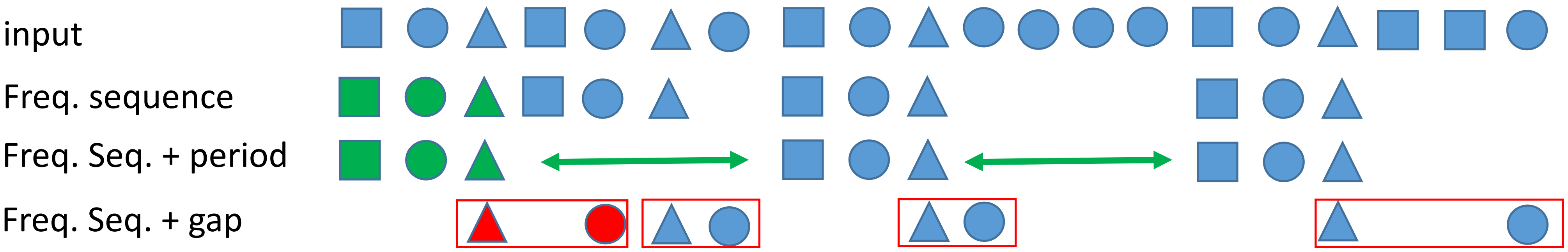
Output: Sets of products bought frequently and **regularly** by customers over time + regularity value

Ex:

- The products **{cat litter, water pack}** are bought every 2 weeks by 19% of all customers.
- The products **{large chocolate box, marrons glacés, truffes}** are bought every year by 46% of all customers.

Can also be enriched with taxonomy (see example above).

Frequent sequence mining



input

Freq. sequence

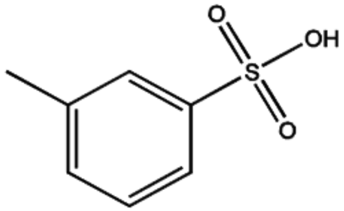
Freq. Seq. + period

Freq. Seq. + gap

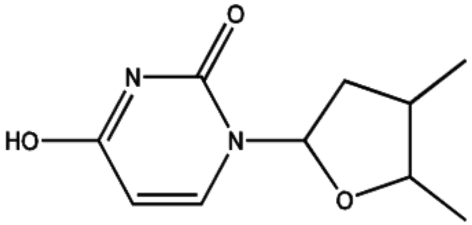
Frequent subgraph mining

Input:

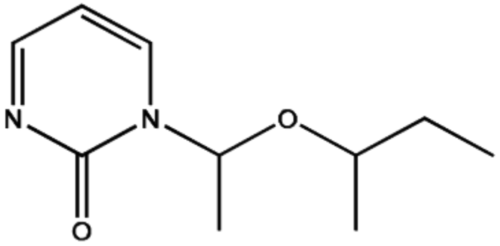
GRAPH DATASET



(A)



(B)



(C)

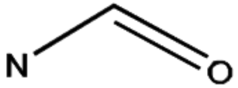
+ 2/3

Minimal expected frequency

Output: Sets of frequent subgraphs

Ex:

(1)



(2)

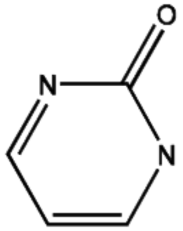


Illustration from https://www.cs.helsinki.fi/u/langohr/graphmining/slides/chp1.pdf (slides of Ehud Gudes)

General KDD process

The (long) way from data to knowledge

The KDD process

Knowledge Discovery in Databases is the nontrivial process of identifying

valid,

novel,

potentially useful,

and understandable

patterns in data.

[Fayyad et al., 1996]

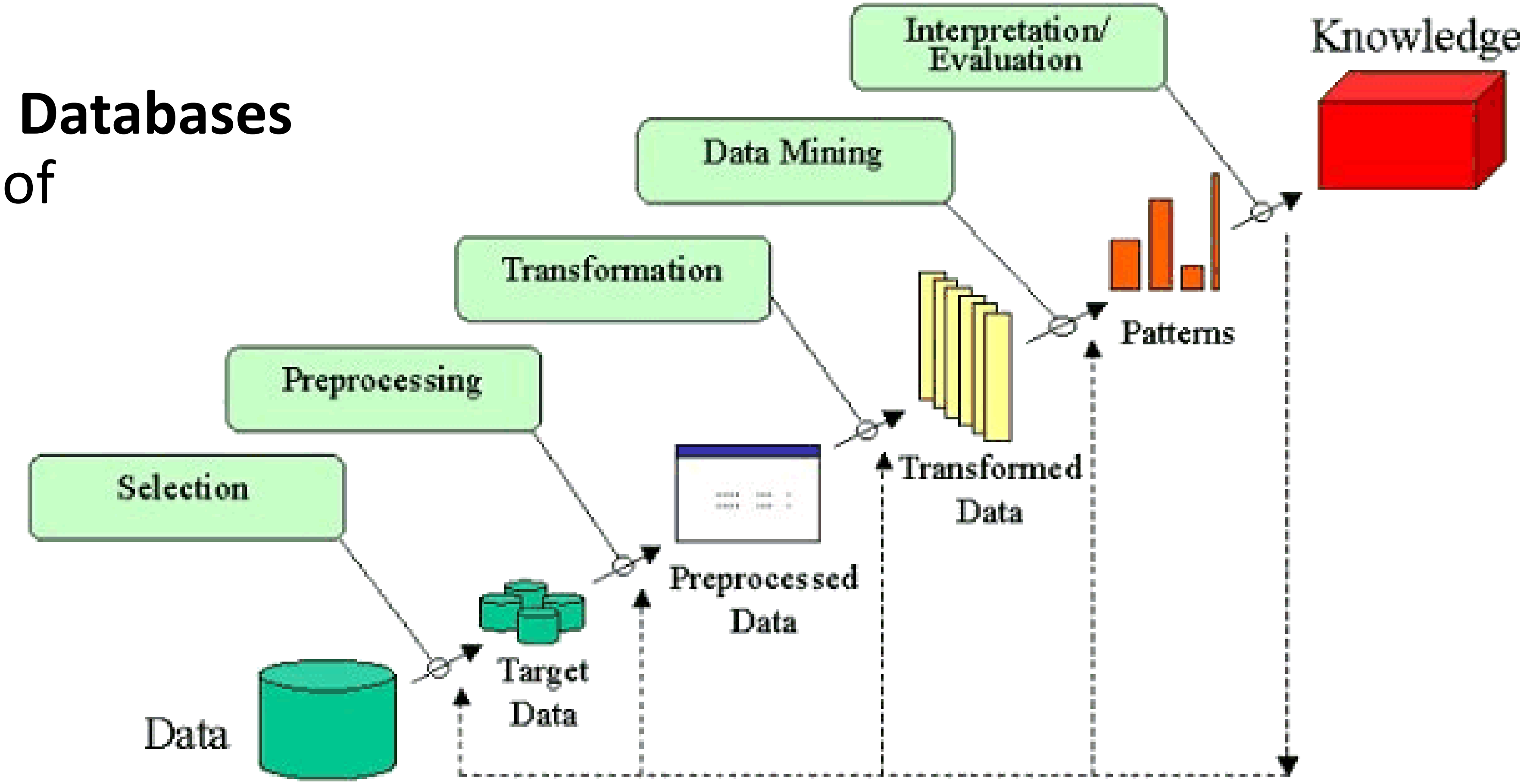


Image: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html

Detailed steps of the KDD process

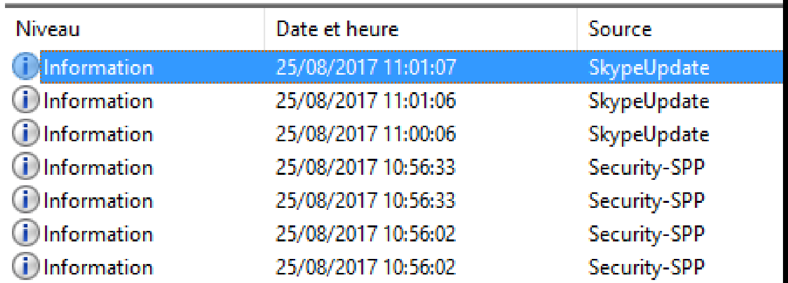
- **Selection**
  - Only consider part of data relevant for problem at hand
  - Better for: algorithm runtime, result quality
- **Preprocessing**
  - Data cleaning, data integration, data reduction
- **Transformation**
  - Make data compliant with expected input of algorithms
- **Data mining**
- **Interpretation / Evaluation / Presentation**
  - Sanity checks
  - Filtering
  - Visualization of results

Data

- Table data

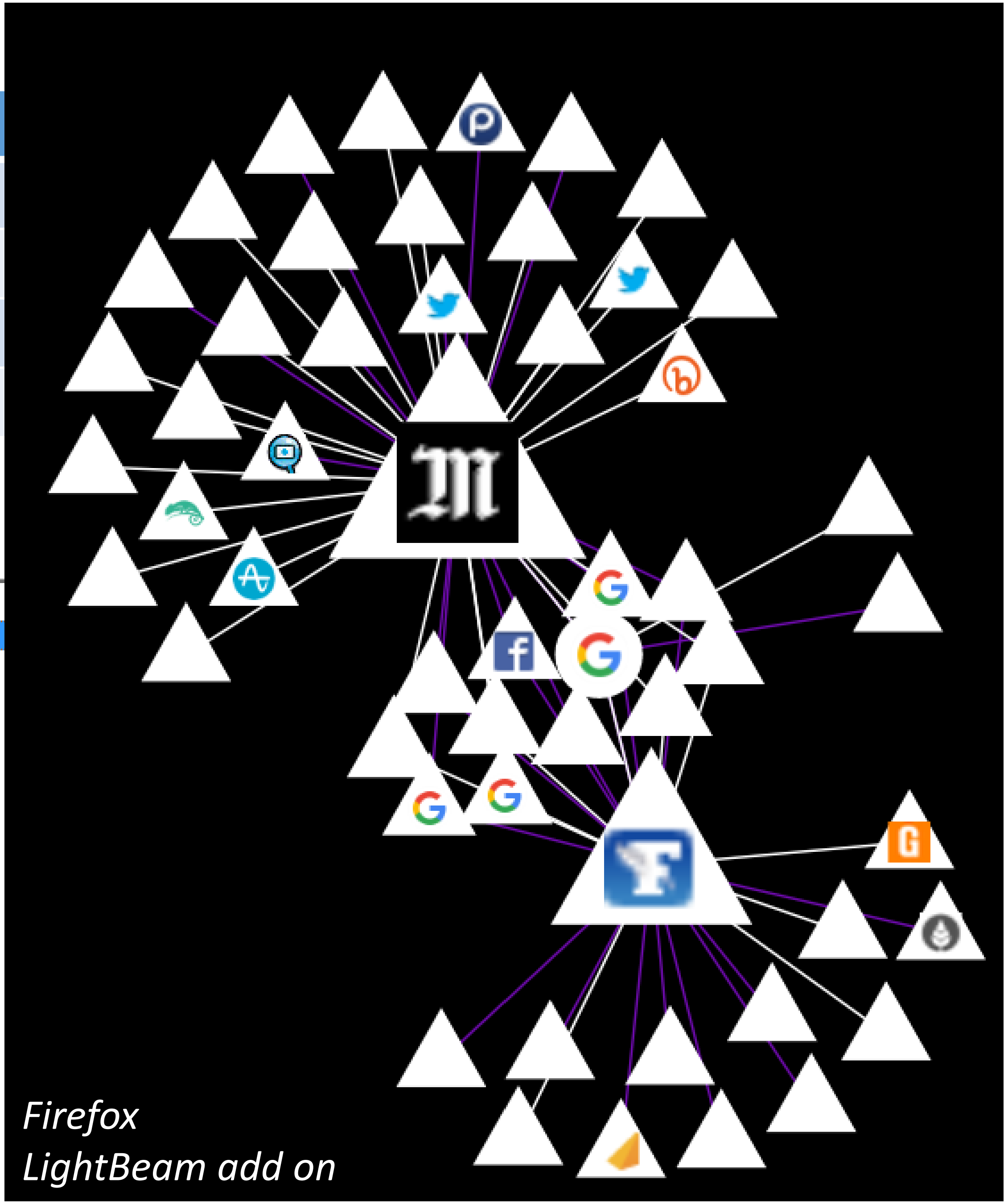
City	Temperature
Paris	19
London	19
Moscow	15
Ushuaia	1

  - Log data



```
Niveau        Date et heure          Source
Information   25/08/2017 11:01:07    SkypeUpdate
Information   25/08/2017 11:01:06    SkypeUpdate
Information   25/08/2017 11:00:06    SkypeUpdate
Information   25/08/2017 10:56:33    Security-SPP
Information   25/08/2017 10:56:33    Security-SPP
Information   25/08/2017 10:56:02    Security-SPP
Information   25/08/2017 10:56:02    Security-SPP
```

- Graph data
- Time Series
- Sequential event data



Firefox LightBeam add on

A real table data (Kaggle / Charlottesville tweets)

id	user_id	user_name	friends_count	followers_count	user_location	user_description	user_profile_background_color	full_text	created_at	is_retweet	quoted_status_text
897661668787982336	2912874772	KCR	250	32	philly	Communications profesh. Giving everything major side-eye right now. Views mine.	0	It's almost as if people are exactly who they say they are https://t.co/MnWFXZd9c3	16/08/2017 03:29	f	"Charlottesville suspect was known as â€œthe Naziâ€ of his high school https://t.co/0gnkFCnJJ3 https://t.co/KRorrul8o8"
897654901534228480	4840680143	Rory Hart	510	62	Connecticut, USA	Educator, Coach, Ally, Activist	F5F8FA	@Slate Conservative media: Yes, Trump's response to Charlottesville was bad, but what about Obama? https://t.co/jjlNXL5Qp0 via @slate	16/08/2017 03:03	f	
897651192372842502	800124776924622848	Kev Spaceman	17	21	null	null	F5F8FA	@seanhannity @JaySekulow @GreggJarrett https://t.co/WHL01vNZKN	16/08/2017 02:48	f	"Thank you President Trump for your honesty &amp

Terminology

Column
Attribute
Feature
Variable

Line
Row
Tuple
Record
Transaction

id	user_id	user_name	friends_count	followers_count	user_location	user_description	user_profile_background_color	full_text	created_at	is_
897661668787982336	2912874772	KCR	250	32	philly	Communications profesh. Giving everything major side-eye right now. Views mine.	0	It's almost as if people are exactly who they say they are https://t.co/MnWFXZd9c3	16/08/2017 03:29	
897654901534228480	4840680143	Rory Hart	510	62	Connecticut, USA	Educator, Coach, Ally, Activist	F5F8FA	@Slate Conservative media: Yes, Trump's response to Charlottesville was bad, but what about Obama? https://t.co/jjl	16/08/2017 03:03	

Terminology

Numerical attributes

Nominal / Categorical attributes

id	user_id	user_name	friends_count	followers_count	user_location	user_description	user_profile_background_color	full_text	created_at	is_retweet	quoted_status_text
897661668787982336	2912874772	KCR	250	32	philly	Communications profesh. Giving everything major side-eye right now. Views mine.	0	It's almost as if people are exactly who they say they are https://t.co/MnWFXZd9c3	16/08/2017 03:29	f	"Charlottesville suspect was known as â€œthe Naziâ€ of his high school https://t.co/0gnkFCnJJ3 https://t.co/KRorruI8o8"
897654901534228480	4840680143	Rory Hart	510	62	Connecticut, USA	Educator, Coach, Ally, Activist	F5F8FA	@Slate Conservative media: Yes, Trump's response to Charlottesville was bad, but what about Obama?	16/08/2017 03:03	f	

Other ML courses: ♥ ♥ ♥
DMV: ♥ ♡ ♡

Numerical attributes

- Quantitative = measurable quantity
- Scaling:
  - Interval-scaled:
    - Measured on scale of equal-size units
    - Difference of values has a meaning
    - Ex: temperature in Celsius / Farenheit, dates
  - Ratio-scaled:
    - Interval-scaled + 0-value is not arbitrary
    - Values can be multiple of other values
    - Ex: temperature in Kelvin, number of likes,…

Other ML courses: ♥ ♡ ♡
DMV: ♥ ♥ ♥

Categorical/Nominal attributes

- Related to « names » / « categories »
- No order, no relation between elements
- Equivalent of `enum` in your favorite programming language
- Ex:
    - Region: {Bretagne, Ile-de-France, Rhône-Alpes…}
    - Student number: {1700893, 1700894,…}
    - Any binary attribute: {True, False}

- NB: can be represented by numbers or any other symbol

Other ML courses: ♥ ♡ ♡
DMV: ♥ ♡ ♡

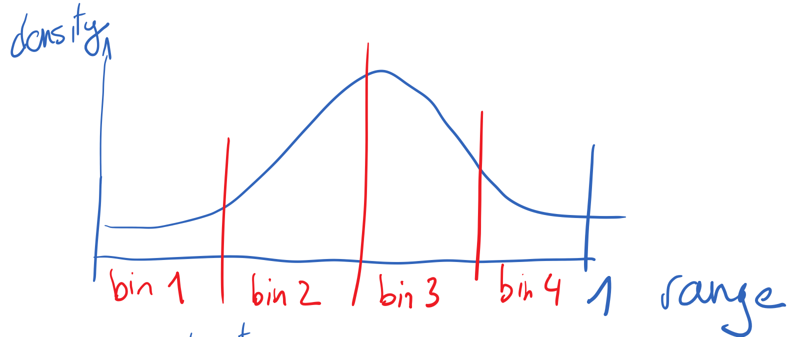
Ordinal attributes

- Same as categorical + ordering among elements
- Not quantitative: difference/ratio are not defined

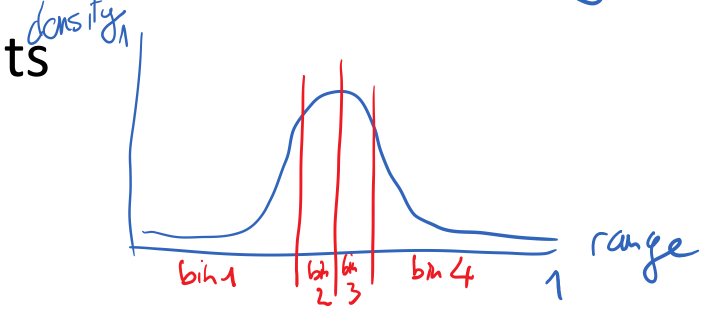
- Ex:
  - US grades: {A, B, C, D, E, F}
  - Size approximation: {small, medium, large}

Discretization

- Turn numerical attributes into categorical / ordinal attributes
- How ?
  - Basic: split range into equal sized bins
    - Problem: over/under-populated bins



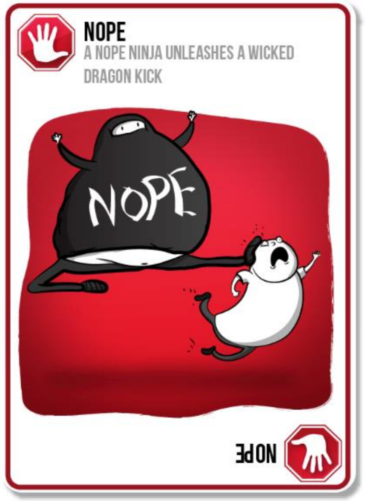
  - Better: split range into bins with equal number of points
    - Problem: intervals may be less intuitive for humans



  - Both:
    - Unsupervised
    - Require #bins as parameter
  - Advanced: cluster analysis (pbs: 1D clustering / parameters)

Preprocessing

NB: more details on some parts in other courses (ex: feature selection)



- Raw data -> shiny data mining algo -> knowledge and $$$?
  - Nope…
- Raw data is:
  - Dirty
    - Ex (real): missing values, random inversion of attributes at middle of table,…
  - Partly (mostly) irrelevant to the problem at hand
    - http://www.tylervigen.com/spurious-correlations
  - Won't play nice with your algorithm
    - Need severe transformations to become expected input

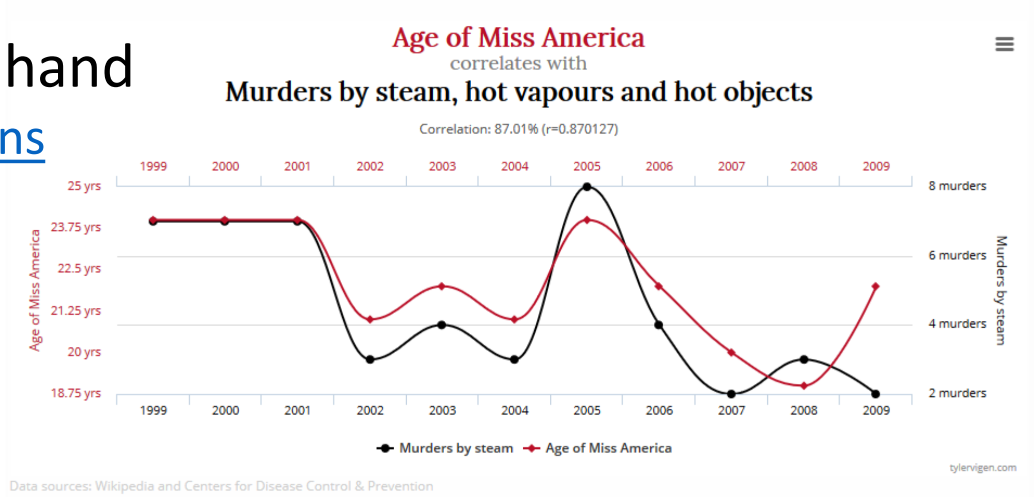


Image credits: The Oatmeal

In practice

- Need to do some Exploratory Data Analysis (EDA)
  - Use interactive notebooks/environments: Jupyter, Rstudio
  - Compute basic statistics: distribution of values, min, max,…
  - Use basic visualizations (next courses)
- Cleanup
  - **Discuss with experts** first!
    - Ex: is it always relevant to replace *age=NaN* by the mean?
  - Remove unnecessary features:
    - Feature selection algorithms (see other courses)
    - **Discuss with experts**
- Transformations
  - Need careful thinking about assumptions made (not just plumbing!)
  - -> need precise idea about expected results => **discuss with experts**!
- Notebooks allow to keep a trace and reproducibility