When patterns overwhelms users, issues and solutions for selecting the *good* ones

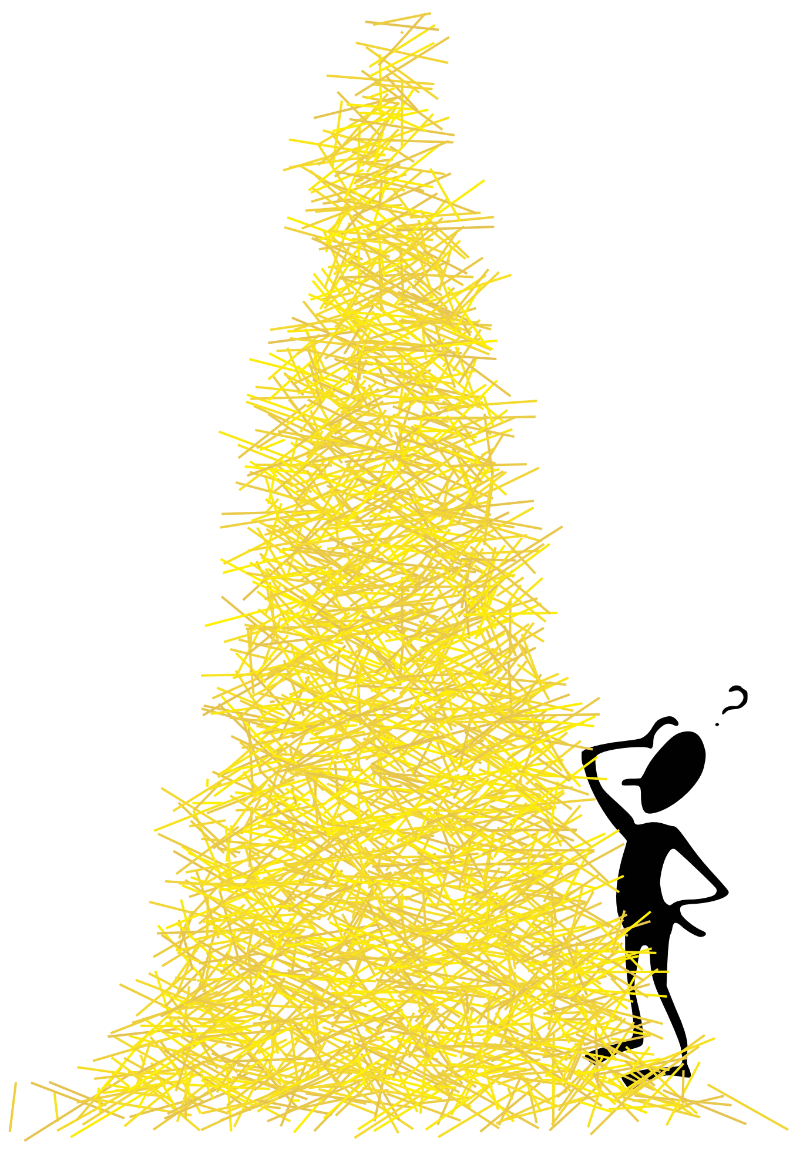
Author

Philippe Lenca

Affiliations

Institut Telecom
Telecom Bretagne
UMR CNRS 3192 Lab-STICC
Technopole Brest Iroise
CS 83818
29238 Brest Cedex 3, France
Université européenne
de Bretagne

KDD



Knowledge Discovery in Data Bases

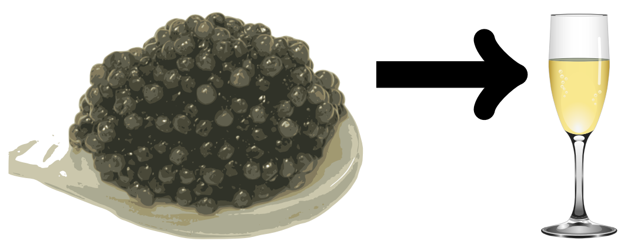
KDD is a non-trivial (decision aid interactive and iterative) process where user(s) seek to identify valid, novel, potentially useful, and ultimately understandable patterns in data.

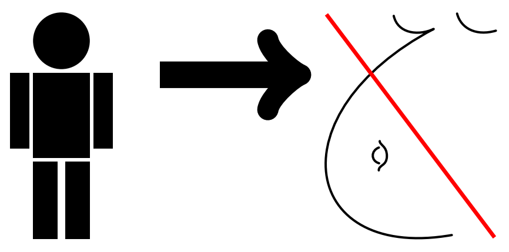
KDD must be considered as a process of contextualization: exact definitions of all concepts are required.

What is valid, novel, useful, understandable, etc.? For who and when?

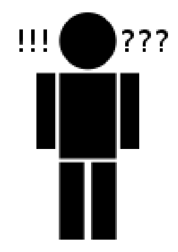
Qualitative and quantitative issues

Some infrequent patterns may be lost: nuggets of knowledge





Some frequent patterns may be true but well known/obvious



...invalid patterns...surprising patterns...

Common issue

One can extract from a database with n attributes:

- 2^n itemsets/candidates
- and 3^n rules

...most of them being uninteresting.

Each data mining researcher/practitioner is faced with assessing the performance of his/her own solution(s) in order to make comparisons with state of the art approaches.

How to select a or several efficient models within a large number of possibilities?

Challenges

- how to choose an or several appropriate interestingness measures? – properties of objective measures of interest (from statistical and from user's point of view); formalisation of user's goal and other contextual factors; definition of new measures; aggregation of measures
- how to mine efficiently interesting patterns in very large databases? – algorithmic properties of interestingness measures; measures used as (ideally complete and minimal) heuristics to reduce the time to mine databases, the memory usage and the number of founded patterns; definition of new measures
- how to use domain knowledge? – knowledge acquisition, formalisation and integration of domain knowledge, quality based on human knowledge, quality of ontologies, knowledge acquisition from text, actionable rules, properties of subjective measures of interest
- challenges with new data and new problems – very large data; very high dimensional data; imbalanced data; changing environments; data stream; lack of training data; sample selection bias; graph data, smart phone-based sensor, etc.; and related specialized domains like bio-informatics, life sciences, social networks, etc.; new users with smart phone
- how an algorithm should be evaluated, how to compare algorithms? on which benchmarks; on which properties (e.g. accuracy, conciseness, specificity, sensitivity, etc.); with which statistical tests; on which trade-off between the different type of errors for multiple simultaneous hypothesis testing; graphical tools like ROC, cost curves; the need to construct new evaluation measures, new experiments, new -reference- databases; issues with parameters tuning questioning also the reproducibility and the robustness of data mining results
- how to help the user(s) to efficiently carry out his/her knowledge discovery process? – methodological guideline (not only what to do but also how to do it); theoretical links between the layers of the process

Contact : philippe.lenca@telecom-bretagne.eu – http://perso.telecom-bretagne.eu/philippelenca/