# In search of structure – A search engine for pattern mining

**Matthijs van Leeuwen**
www.patternsthatmatter.org

## Context

*Search engines* are of one the largest success stories of the Internet-era: they have made the huge resources of the world wide web accessible to virtually anyone with access to the Internet. The key to their success, apart from the underlying technology, is the *simplicity* with which documents can be queried and retrieved.

*Exploratory data mining* aims to enable users to explore data and stumble upon interesting insights while doing so. In particular, pattern mining methods identify and describe local structure in data. The ultimate goal is to allow users to discover novel knowledge from data. Unfortunately, advanced pattern mining techniques can be *hard* to use, certainly for people with no expertise in data mining.

## Goal

The goal of this project is to devise and implement a *search engine for pattern mining*, so that pattern-based interactive data exploration becomes feasible for people without a thorough understanding of data mining algorithms. Given a dataset, or a collection of datasets, and a user-specified query, the system should present a ranked list of relevant patterns. Once this core functionality has been established, extensions such as other patterns types or query refinement can be explored.

## Research questions

There are a large number of questions that could be investigated:

- Which types of a) data and b) patterns can be handled by such a system?
- What is a suitable query language? (Simple yet effective, as in information retrieval?)
- How to quantify 'relevancy' of a pattern given a query?
- How to ensure that the response time of the system is (near) instant?
- Can existing technology be used for the implementation?
    - Pattern mining algorithms?
    - Search engine implementations?
      (E.g., http://lucene.apache.org/solr/, https://www.elastic.co/products/elasticsearch)
- Data & applications: what applications can be dealt with?
- Evaluation: how to evaluate the system?

## Realisation

1. Literature study.
2. Determine scope of the project, write research plan.
3. Develop foundations needed for system.
4. Design and implement system (modular architecture required for extensibility).
5. Evaluate the system, possibly using one or two case studies.
6. Write thesis.

**Student profile**
Interested in algorithms, data mining, and information retrieval; good implementation skills.

**Relevant literature**


[1] Charu Aggarwal and Jiawei Han. *Frequent Pattern Mining*, Springer, September 2014.
http://charuaggarwal.net/freqbook.pdf

[2] Matthijs van Leeuwen. *Interactive Data Exploration using Pattern Mining*. In: Interactive Knowledge Discovery and Data Mining: State-of-the-Art and Future Challenges in Biomedical Informatics (Holzinger, A. & Jurisica, I., eds), LNCS, Springer, 2014.
http://patternsthatmatter.org/pubs/2014/interactive_data_exploration_using_pattern_mining-vanleeuwen.pdf

[3] Mario Boley, Maike Krause-Traudes, Bo Kang and Björn Jacobs. *Creedo---Scalable and Repeatable Extrinsic Evaluation for Pattern Discovery Systems by Online User Studies*. In: Proceedings of IDEA 2015 (KDD workshop)
http://poloclub.gatech.edu/idea2015/papers/p20-boley.pdf
http://www.realkd.org/creedo-webapp/