Patterns that Matter
Describing Structure in Data

Matthijs van Leeuwen

Leiden Institute of Advanced Computer Science

17 November 2015

Universiteit Leiden

Big Data: A Game Changer
in the retail sector

Forbes

Predicting trends

Forecasting demand

Optimizing pricing

Identifying customers

… ***predictive analytics!***

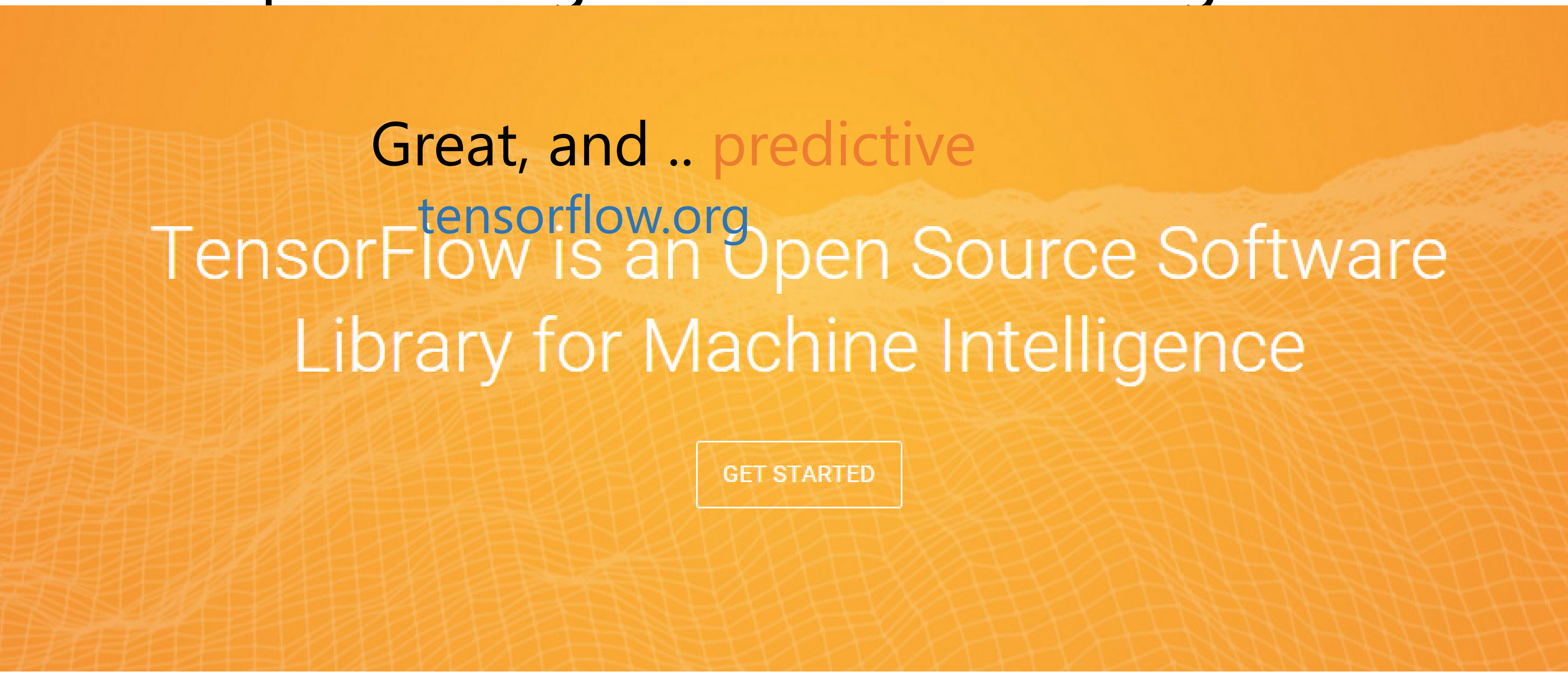
www.forbes.com/sites/bernardmarr/2015/11/10/big-data-a-game-changer-in-the-retail-sector/

Google's take on

Deep learning / machine learning

Great, and .. predictive

tensorflow.org



What if ...

We need a **summary** of our data?

We need **explanations** to backup our decisions?

We require **interpretability**?

We aim for **description** rather than (black box) **prediction**?

Exploratory data mining is discovering structure to gain novel insights

"Tell me something interesting about my data"

Mining models & patterns from data

Different from machine learning

- Description rather than prediction
- Interpretation and explanation

Roadmap

Pattern mining

Information theory
for data mining

Applications

Roadmap

Pattern mining

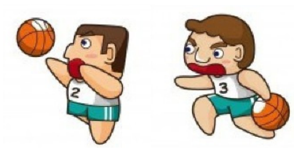
Information theory for data mining

Applications

A pattern describes
local structure in data

Which players often play together?

4/5=80%



4/5=80%

3/5=60%

Pattern mining
Problem statement

Given

- a database D
- a pattern language $\mathcal{P}$
- a set of constraints $\mathcal{C}$

Find the set of patterns $S \subseteq \mathcal{P}$ such that

- each $p \in S$ satisfies each $c \in \mathcal{C}$ on D
- S is maximal

Find **all** patterns
satisfying the constraints

Frequent itemset minin
An instance of pattern

#28 and #36 of most cited computer science articles.

(Source: Citeseer)

Data

$\mathcal{I}$ is a set of items

D is a bag of transactions t over $\mathcal{I}$, i.e., $t \subseteq \mathcal{I}$

Pattern language

$\mathcal{P} = \mathrm{Pow}(\mathcal{I})$

Constraints

$frequency_D(p) = |\{t \in D \mid p \subseteq t\}|$

$\mathcal{C} = \{frequency_D(p) \geq minfreq\}$

Putting pattern mining into perspective

Search space usually discrete
- Often represented as ‘pattern lattice’
- Often **enormous**!

Combinatorial search
- E.g., branch-and-bound
- *All* solutions rather than the *best* one

Key: clever algorithms

Problems in pattern paradise

The **pattern explosion**

- High thresholds: few, well-known patterns
- Low thresholds: a gazillion patterns

Many similar patterns

- Redundancy in pattern languages
- Top-*k* mining useless

Mining frequent itemsets

Dataset	**|*D*|**	***minsup***	**# frequent itemsets**
Adult	48,842	1	58,461,763
Heart	303	1	1,922,983
Mushroom	8,124	1	5,574,930,437
Wine	178	1	2,276,446

Four datasets from the UCI repository.

Pattern set mining
Problem statement

Given

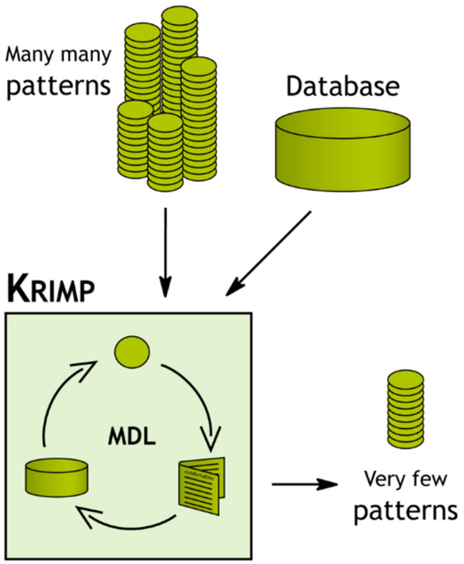
a database D
a pattern language $\mathcal{P}$
a set of constraints $\mathcal{C}$
an optimisation criterion G

Find the set of patterns $S \subseteq \mathcal{P}$ such that

each $p \in S$ satisfies each $c \in \mathcal{C}$ on D
S is optimal w.r.t. G

Find a **global model** of **local patterns**

Roadmap



Pattern mining

Information theory for data mining

Applications

Optimality and induction

What is the optimal set of patterns?

- Should generalise the database
- Generalisation = induction
- I.e., we should employ an **inductive** principle

So, which principle should we choose?

- Patterns are descriptive for parts of the data

The **Minimum Description Length principle** is *the* induction principle for descriptions

Rissanen, J. Modeling by shortest data description. In: *Automatica*, 14(5), 1978.

What is MDL?
The Minimum Description Length principle

Given a set of models $\mathcal{M}$, the best model $M \in \mathcal{M}$ is the one that minimises

$$L(M) + L(D \mid M)$$

in which

- $L(M)$ is the length, in bits, of the description of M,
- $L(D \mid M)$ is the length, in bits, of the description of the data when encoded with M.

MDL-based pattern set mining
Problem statement

Given

a database D
a pattern language $\mathcal{P}$
a set of constraints $\mathcal{C}$

Find the set of patterns $S \subseteq \mathcal{P}$ such that

each $p \in S$ satisfies each $c \in \mathcal{C}$ on D
S is MDL optimal

Find a **model** that **compresses** the data

Does this make sense?

Yes.

A good model 'compresses' your data well

MDL makes this observation concrete

Compression must be lossless!

"The best set of patterns is
that set of patterns
that compresses the data best"

Compression as a means to an end

We **do not care** about actually compressing the data!

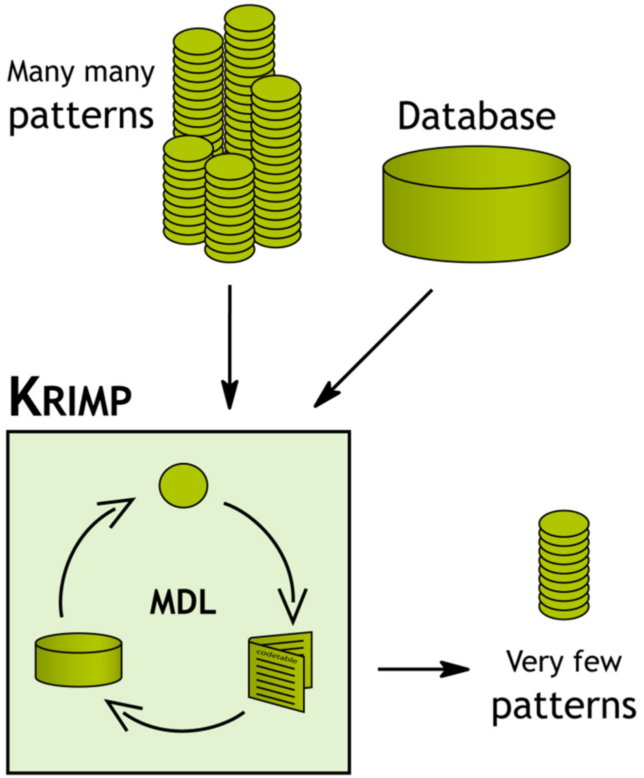
- Data storage is cheap enough these days

We want the **set of patterns** that yield the best compression

- I.e., we want to look *inside the compressor*
- MDL allows for such inspection perfectly

Pattern-based models through compression

Krimp



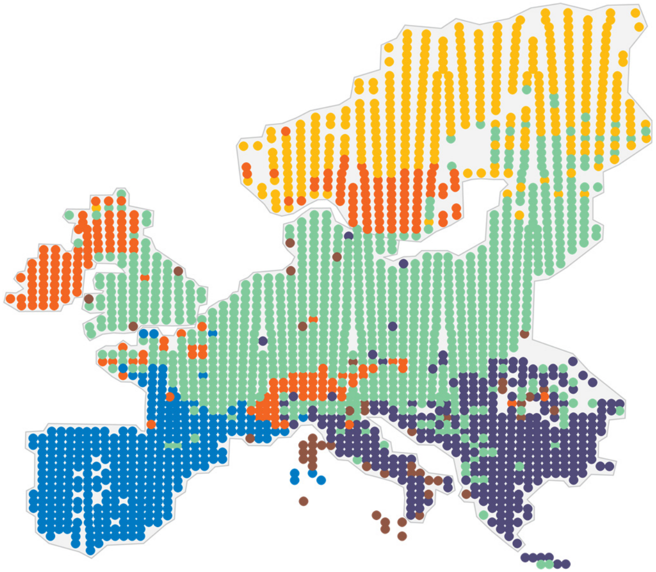
- **+** Solves redundancy problems
- **+** Very characteristic for the data
- **+** Can be used for many data mining tasks

J. Vreeken, M. van Leeuwen & A. Siebes. Krimp: Mining Itemsets that Compress.
In: *Data Mining and Knowledge Discovery* 23(1), 2011.

KRIMP in action

Dataset	**|*D*|**	***minsup***	**# Freq. Itemsets**	**|*CT*|**
Adult	48,842	1	58,461,763	999
Heart	303	1	1,922,983	108
Mushroom	8,124	1	5,574,930,437	424
Wine	178	1	2,276,446	76

Roadmap



Pattern mining

Information theory
for data mining

Applications
in exploratory data mining

Clustering categorical data

Partition the data into *k* clusters

Each cluster is characterised by a pattern set

No dissimilarity measure required!

Optimal *k* determined by MDL

Formally

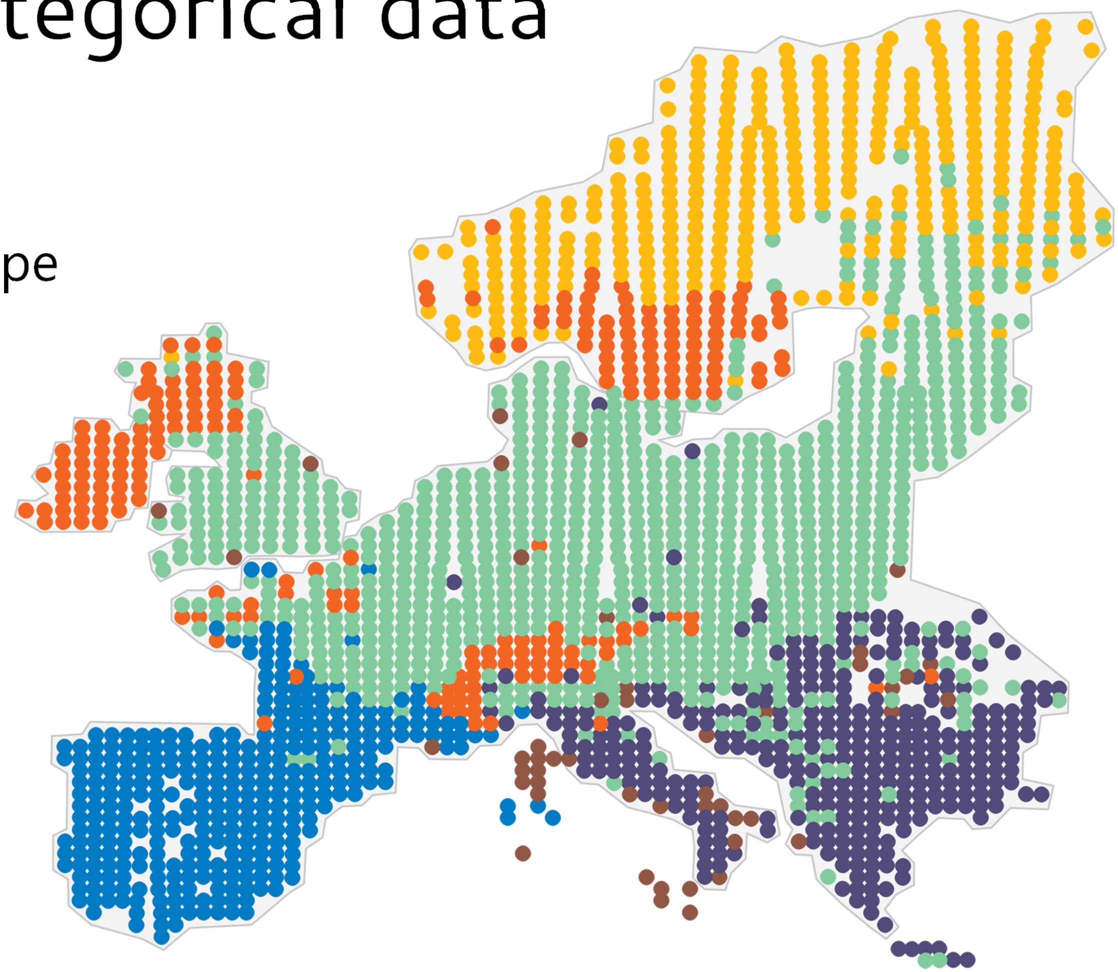
Partition $\mathcal{D}$ into $\mathcal{D}_1 \ldots \mathcal{D}_k$
such that $\sum L(CT_i, \mathcal{D}_i)$
is minimised

M. van Leeuwen, J. Vreeken & A. Siebes. Identifying the Components.
In: *Data Mining and Knowledge Discovery* 19(2), 2009. **Best student paper @ ECML PKDD'09**

Clustering categorical data

Mammals

- 2221 areas in Europe
- 50x50 km each
- 124 mammals
- *No location info*



k=6, MDL 'optimal'

M. van Leeuwen, J. Vreeken & A. Siebes. Identifying the Components.
In: *Data Mining and Knowledge Discovery* 19(2), 2009. **Best student paper @ ECML PKDD'09**

Change detection in data streams

Data streams are ubiquitous

- Financial world
- Retail (supermarkets, online stores, …)
- Web
- …

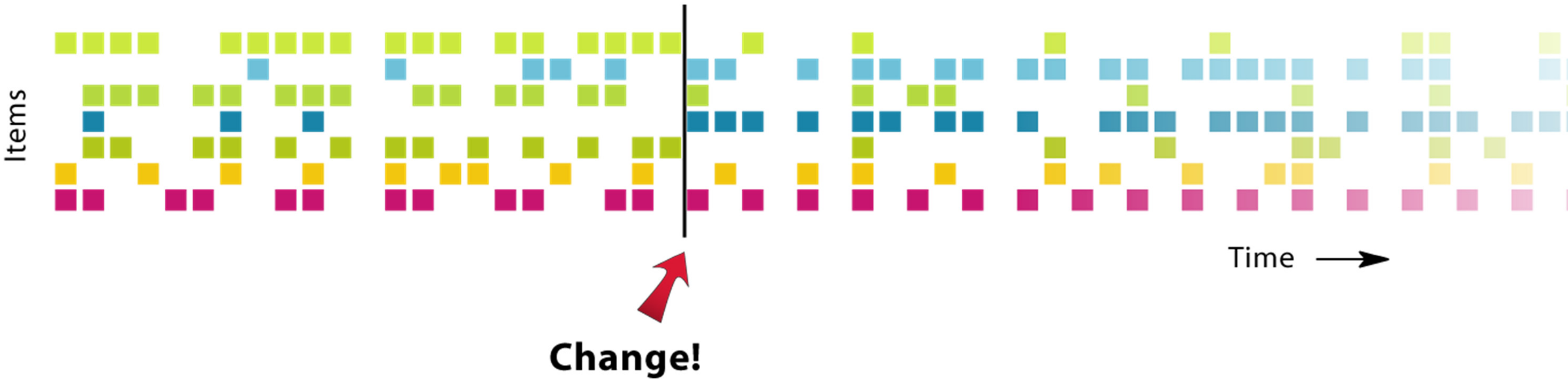
Can we ***detect*** *sudden changes?*

An example data stream

Data stream: a sequence of transactions

M. van Leeuwen & A. Siebes. StreamKrimp: Detecting Change in Data Streams.
In: *Proceeding of ECML PKDD'08*, pp.765-774, 2009.

An example data stream



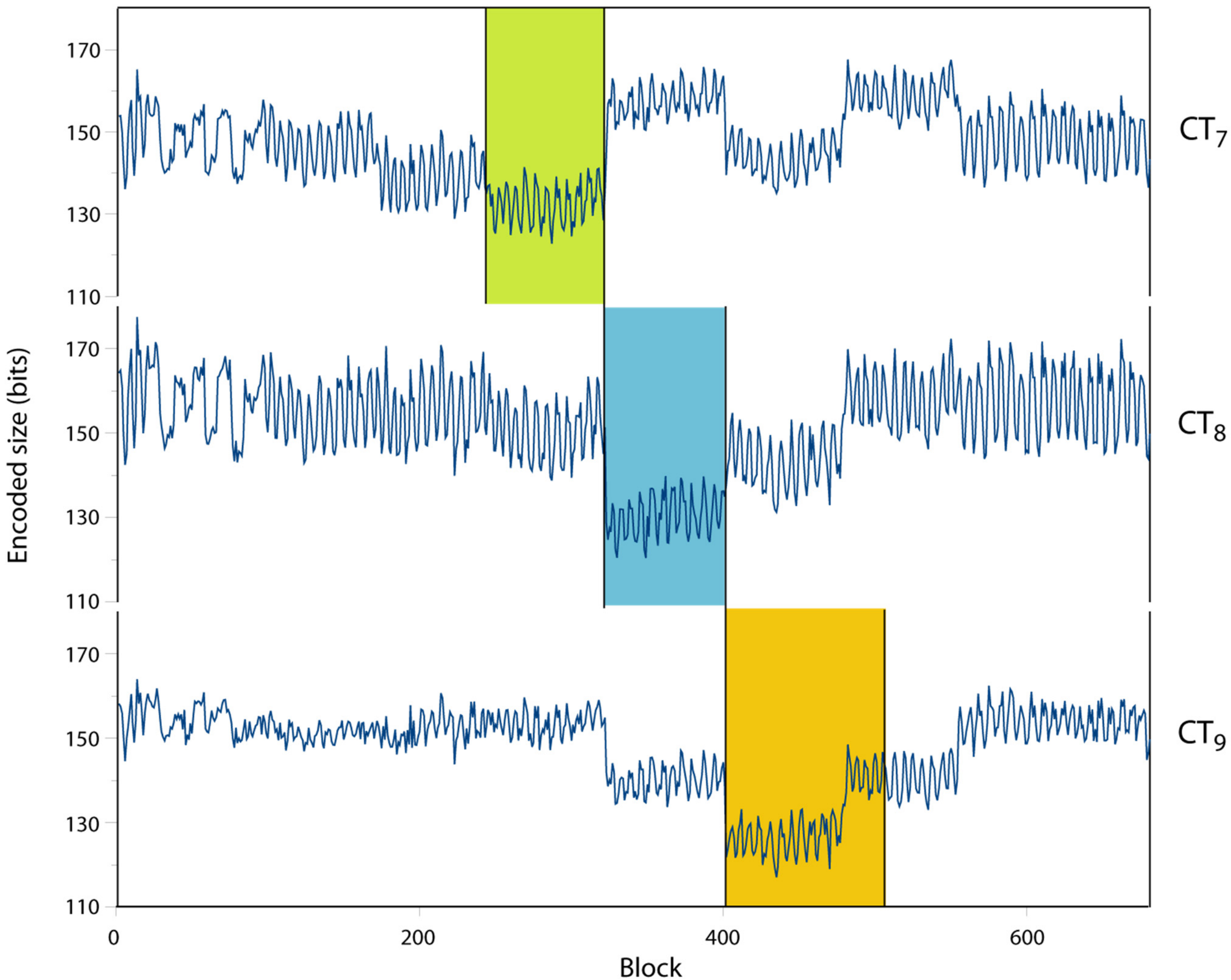
Identify changes in the characteristics of the data

M. van Leeuwen & A. Siebes. StreamKrimp: Detecting Change in Data Streams.
In: *Proceeding of ECML PKDD'08*, pp.765-774, 2009.

Accidents

Belgian traffic accidents

- 1991 – 2000
- 340,184 tuples
- 468 items

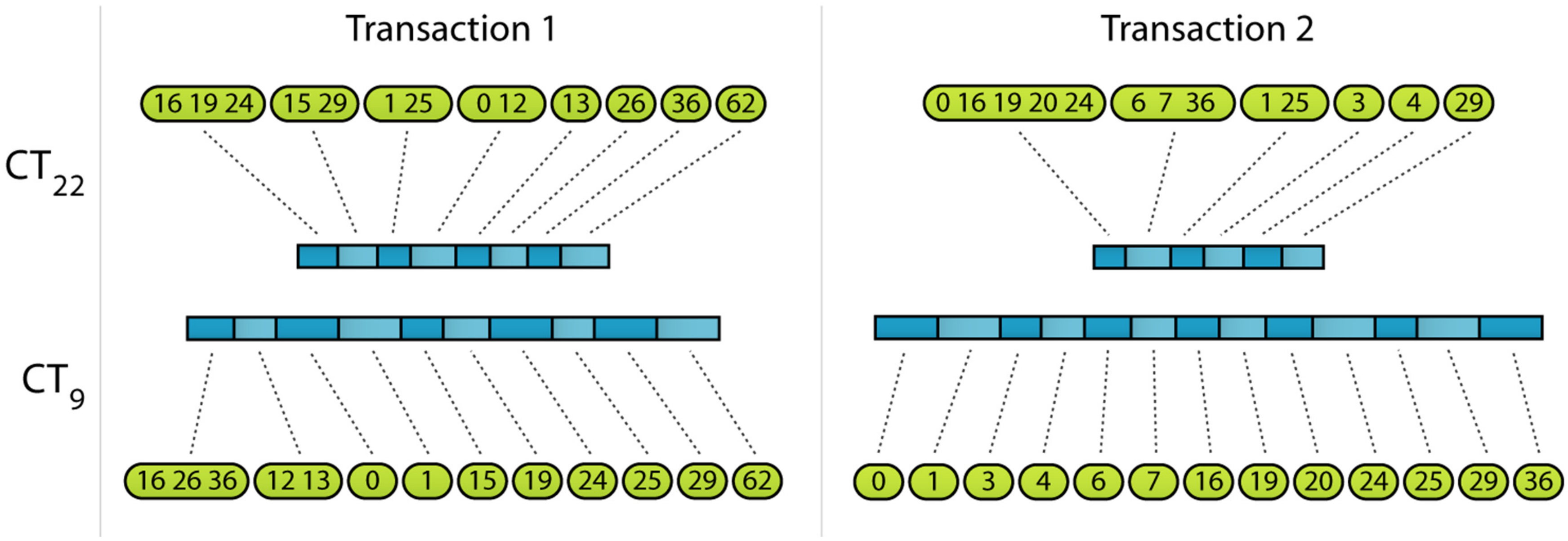


M. van Leeuwen & A. Siebes. StreamKrimp: Detecting Change in Data Streams.
In: *Proceeding of ECML PKDD'08*, pp.765-774, 2009.

Characterising the difference

Encode transactions with compressors induced from different databases.

Reveals **recognized patterns**,
pinpoints **differences**.



J. Vreeken, M. van Leeuwen & A. Siebes. Characterising the Difference.
In: *Proceedings of KDD'07*, 2007.

Compression for data mining

Can be successfully used for many tasks

Classification	*ECML PKDD 2006*
Database (dis)similarity	*KDD 2007*
Data generation & privacy preservation	*ICDM 2007*
Change detection in data streams	*ECML PKDD 2008*
Database components / clustering	*ECML PKDD 2009*
Identifying media groups in tag data	*CIKM 2009*
Characterising uncertain data	*SDM 2011*
Tag recommendation	*IDA 2012*

…

Patterns that Matter

reveal novel insights about your data

Exploratory data mining is an exciting field and has much to offer

Description rather than prediction

Pattern-based modelling has many desirable properties

Interpretability, explanation, …

Use **algorithmic information theory** for model selection

Induction by compression

Many thanks to all!

www.patternsthatmatter.org