The Data Problem in Data Mining

Albrecht Zimmermann
Université de Caen Normandie

14th IDA, St. Étienne, 24.10.2015

UNIVERSITÉ CAEN NORMANDIE

The Data Problem in Unsupervised Pattern Mining

Albrecht Zimmermann
Université de Caen Normandie

14th IDA, St. Étienne, 24.10.2015

UNIVERSITÉ CAEN NORMANDIE

Take home message

We don't understand

- ✖ Algorithm run times
- ✖ Parameter settings
- ✖ How to interpret mined patterns

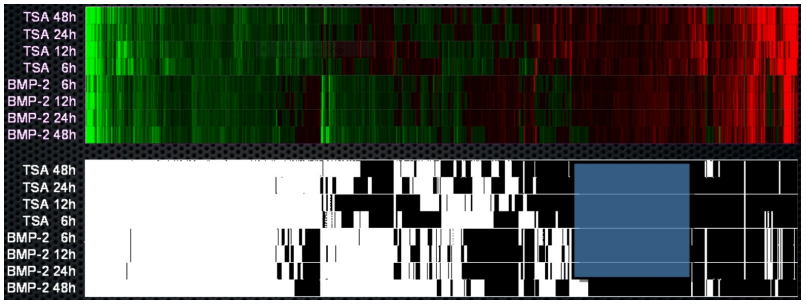
➔ Because we lack data !

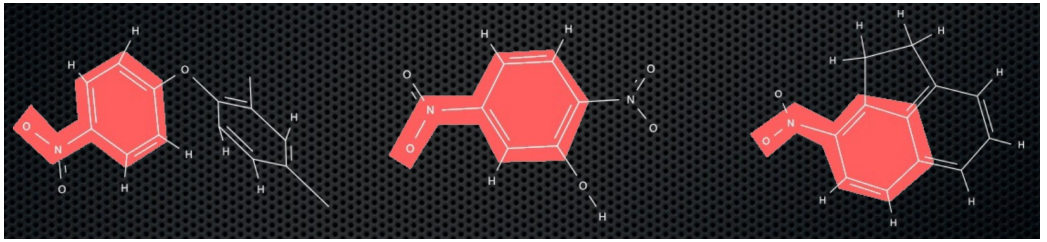
We can begin to fix this

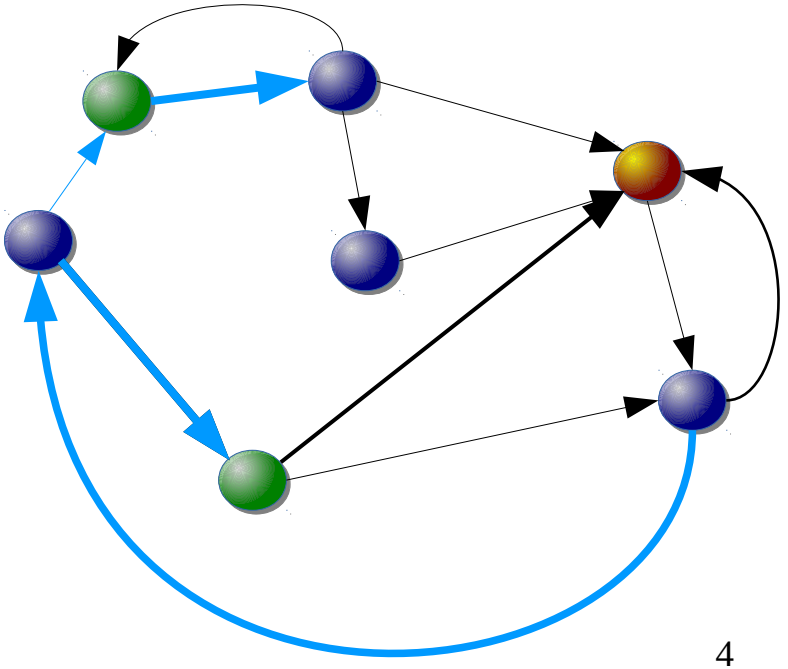
- ✔ Researching data generation
- ✔ Generating data
- ✔ Exploring mining behavior

➔ Add new knowledge/tools

Local pattern mining







Road map

1) Run time behavior
2) Parameter setting/output behavior
3) Relation patterns – data
4) Data science

Road map

1) Run time behavior
   - Problem setting
   - Existing evaluation
   - Data problem
   - Attempts at understanding

1)Run times – problem setting/evaluation

Data characteristics

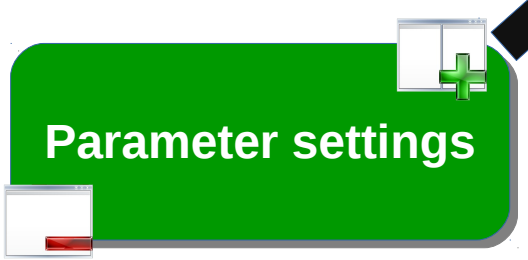
How long does a mining run take ?

Parameter settings

1)Run times – problem setting/evaluation

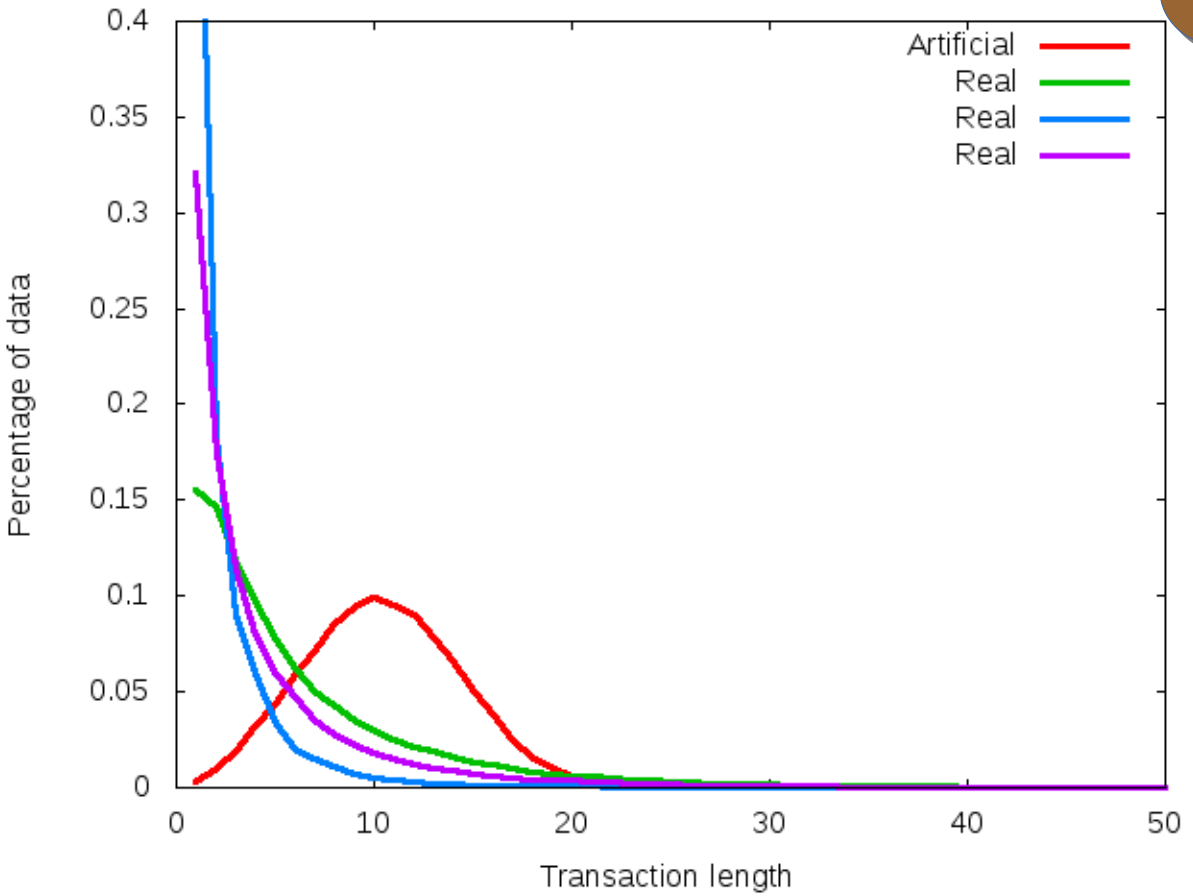
Algorithmic aspects

How long does a mining run take ?



Run times – the paper that started it all !

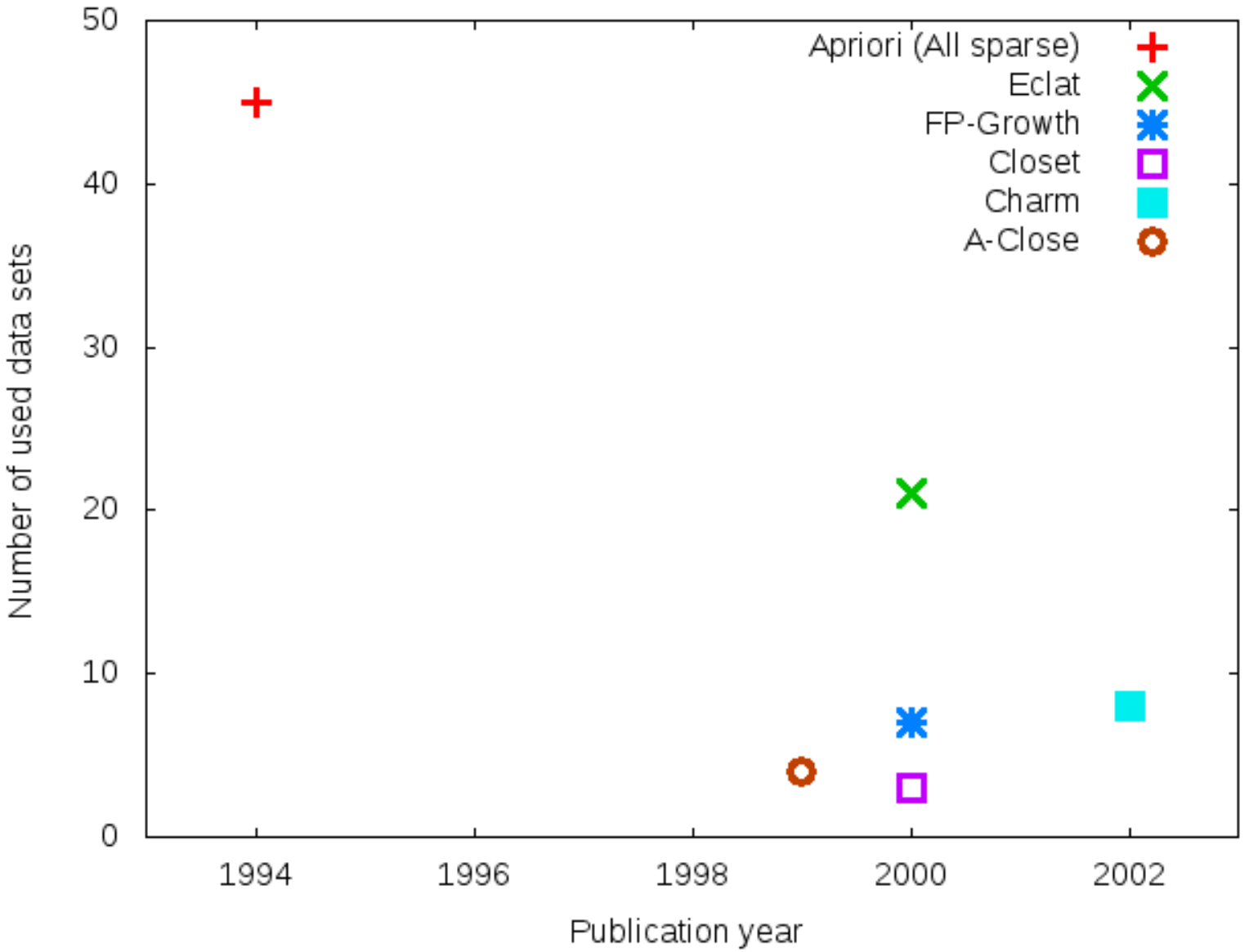
Zheng, Kahov, Mason. *Real world performance of association rule algorithms.* KDD 2001



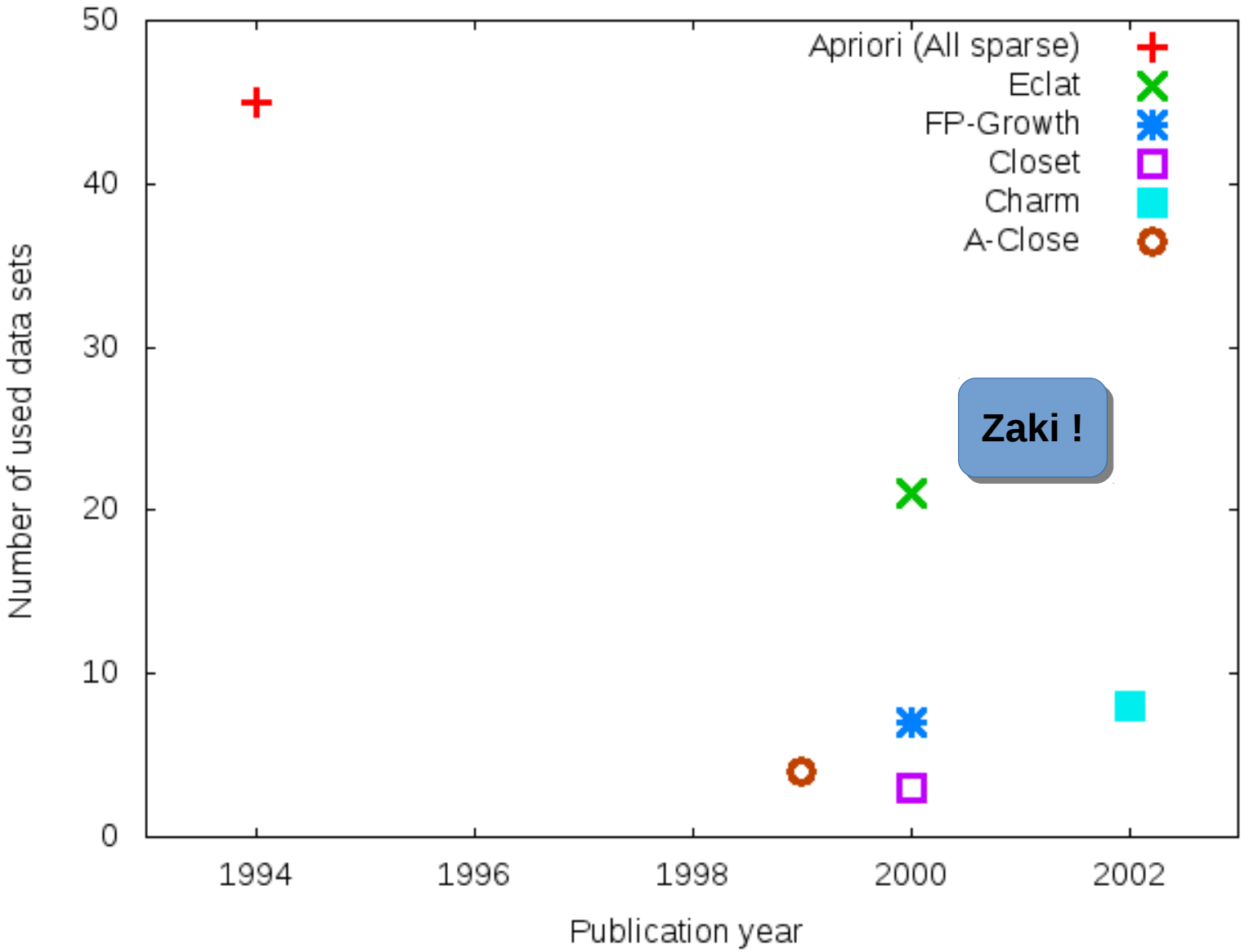
Run time reduction disappeared on real data !

« Improvements » over Apriori over-fit the data

Run times – how much data ?



Run times – how much data ?



What's the issue ?

Not enough data !

Zheng data sets added

- Nowadays FIMI (Goethals & Zaki) repository
  12 data sets

UCI : 333
UCR : 70+

Zheng et al. killed QUEST, no replacement

Understanding run times

Pattern type	**Itemsets**	**Graphs**	**Episodes**
Characterized by	Distribution pattern lengths	• Subtasks • Run time per fragment • Edge densities/ Branching factor • Java VM/Cache size	• Length • Non-event probability • Alphabet size...
	Zaki et al. 2002 Ramesh et al. 2003 Gouda et al. 2005 Flouvat et al. 2010	**Wörlein et al. 2005 Nijssen et al. 2006**	**Zimmermann 2014**

Understanding run times

Pattern type	**Itemsets**	**Graphs**	**Episodes**
Characterized by	Distribution pattern lengths	• Subtasks • Run time per fragment • Edge densities/ Branching factor • Java VM/Cache size	• Length • Non-event probability • Alphabet size...
	Zaki et al. 2002 **Ramesh et al. 2003** **Gouda et al. 2005** **Flouvat et al. 2010**	**Wörlein et al. 2005** **Nijssen et al. 2006**	**Zimmermann 2014**

Zaki !

Predicting run times

Based on	**Bernoulli/ Markovian model**	**Partial mining results**	**Sampling**
	Lhote et al. 2005	**Palmerini et al. 2004** **Geerts et al. 2005**	**Boley et al. 2008/2010**

Work-around : how many frequent sets?

2)Parameter settings – problem/evaluation

Too strict : lose results
Too lenient : drown in patterns

Data characteristics

How to set parameters for (potentially) interesting results ?

Parameter settings

2)Parameter settings – problem/evaluation

Too strict : lose results
Too lenient : drown in patterns

Data characteristics

How to set parameters for (potentially) interesting results ?

Top-k
Pattern set mining
Statistical pattern mining
MaxEnt

2)Parameter settings – problem/evaluation

Too strict : lose results
Too lenient : drown in patterns

Data characteristics

How to set parameters for (potentially) interesting results ?

Top-k
Pattern set mining
Statistical pattern mining
MaxEnt

Interactive mining

Issue ?

Same data (not much)

- W/problematic properties

No patterns known !

- Can't compare

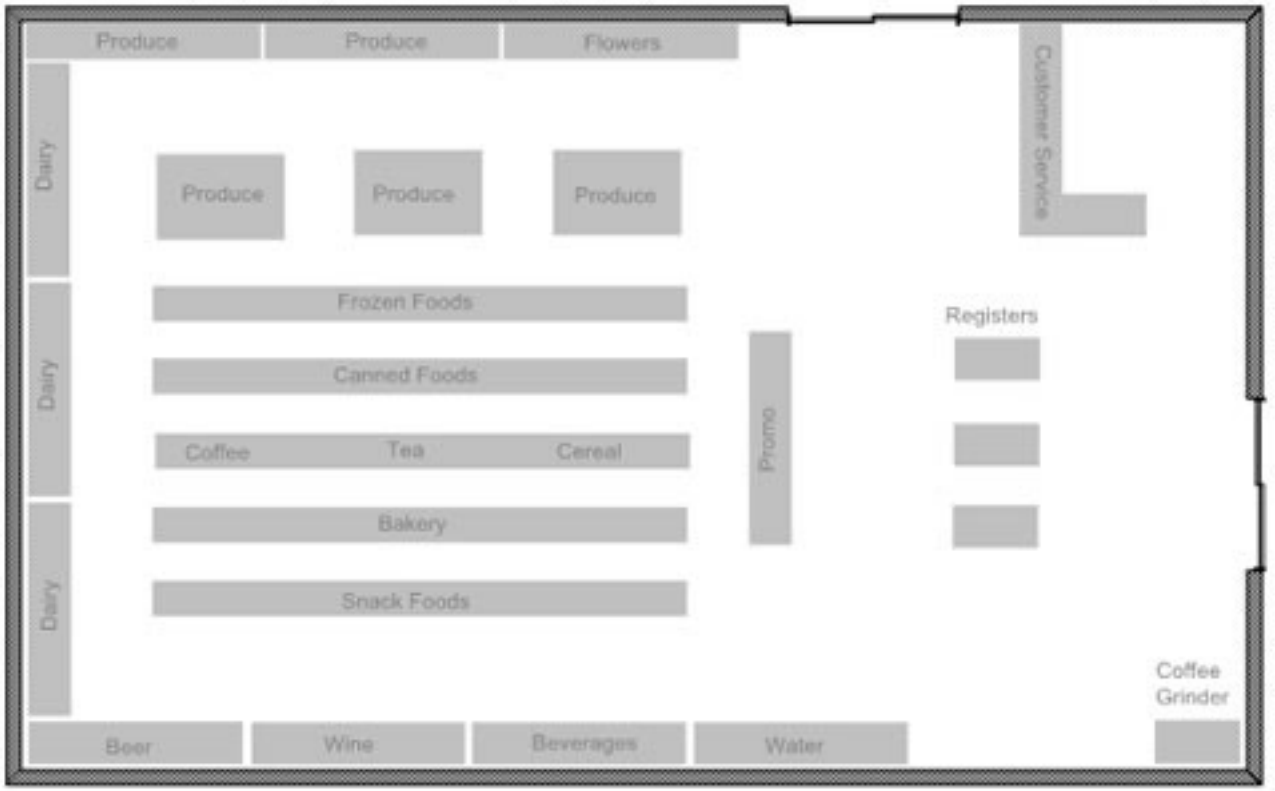
Understanding output sizes/composition

Pattern type	**Itemsets**	**Strings**
Characterized by	Samples + Background models	Samples + Background models
	Boley et al. 2008/2010 **Van Leeuwen et al. 2014**	**Besson et al. 2008**

Work-around : where's the spectrum different from random ?

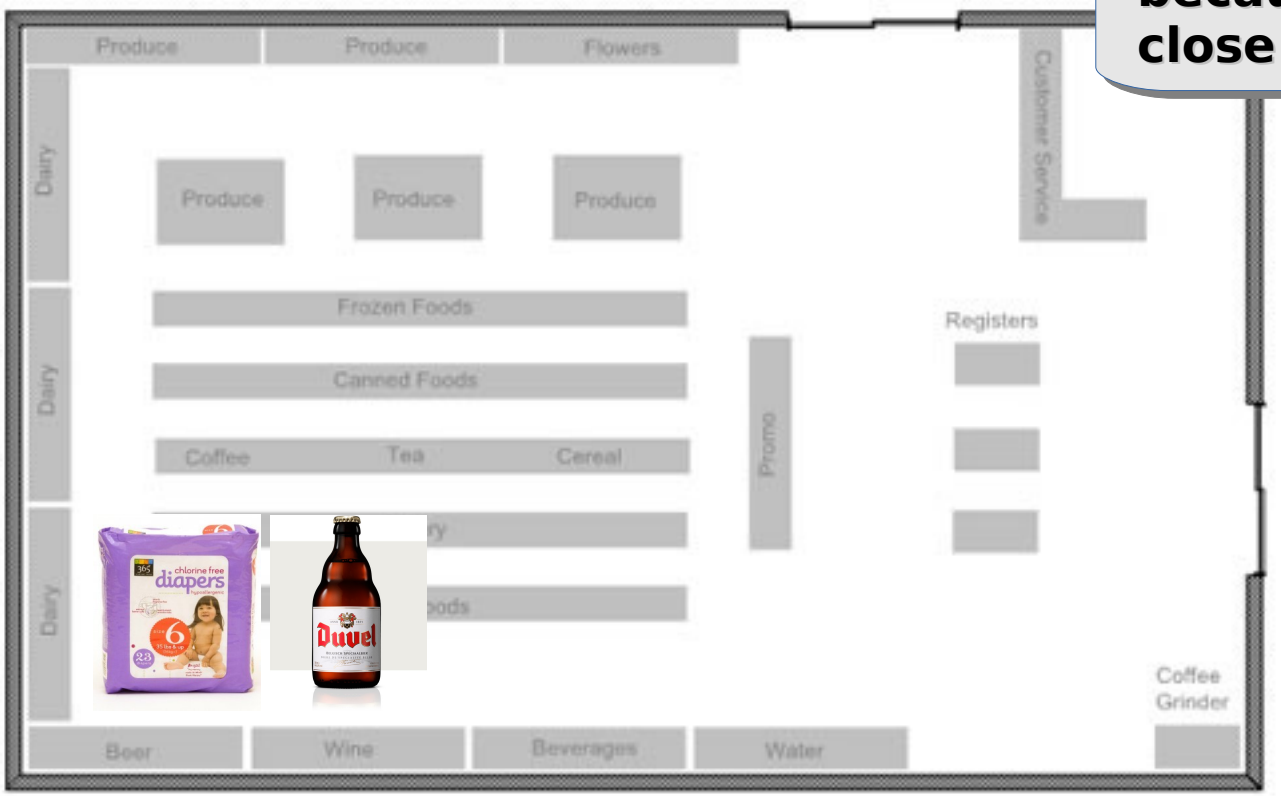
3)Understanding/ interpreting patterns – problem setting

What do the patterns *mean* that we mine ?



3)Understanding/ interpreting patterns – problem setting

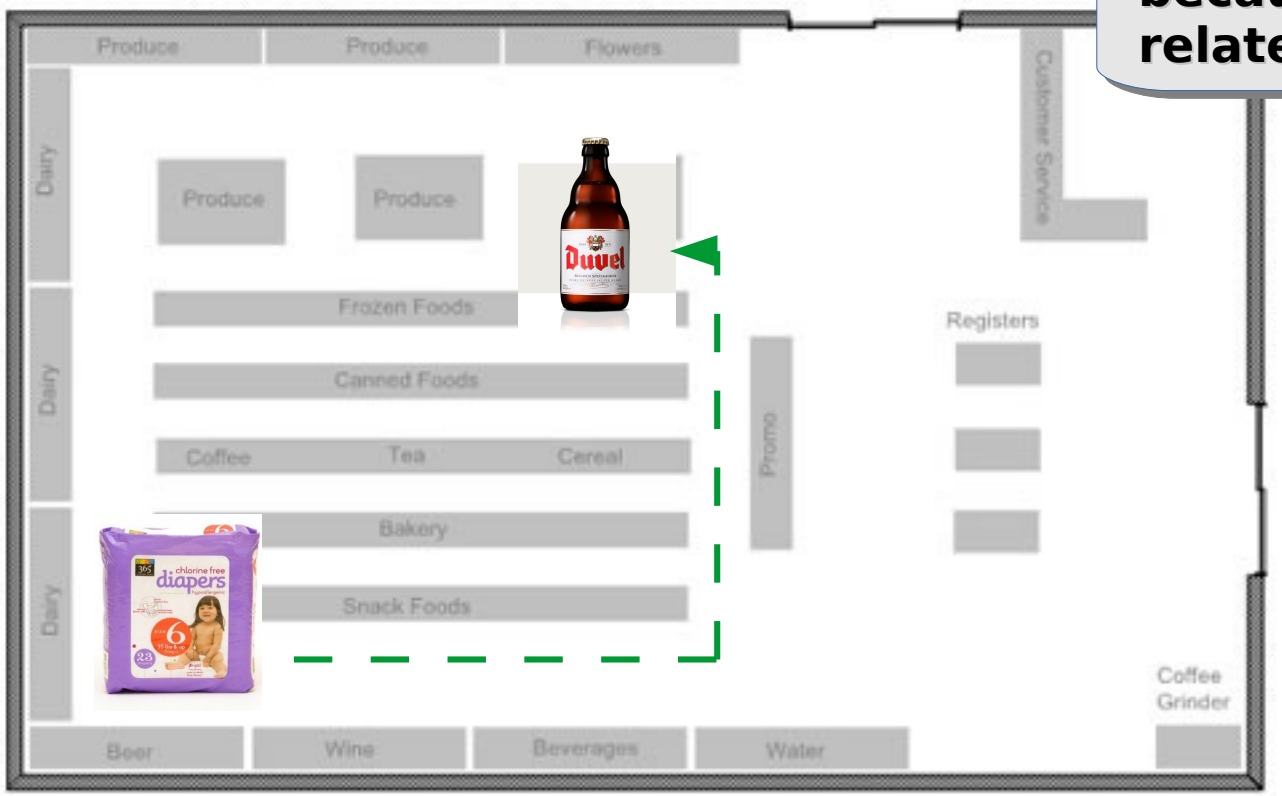
What do the patterns *mean* that we mine ?



Bought together because they are close

3)Understanding/ interpreting patterns – problem setting

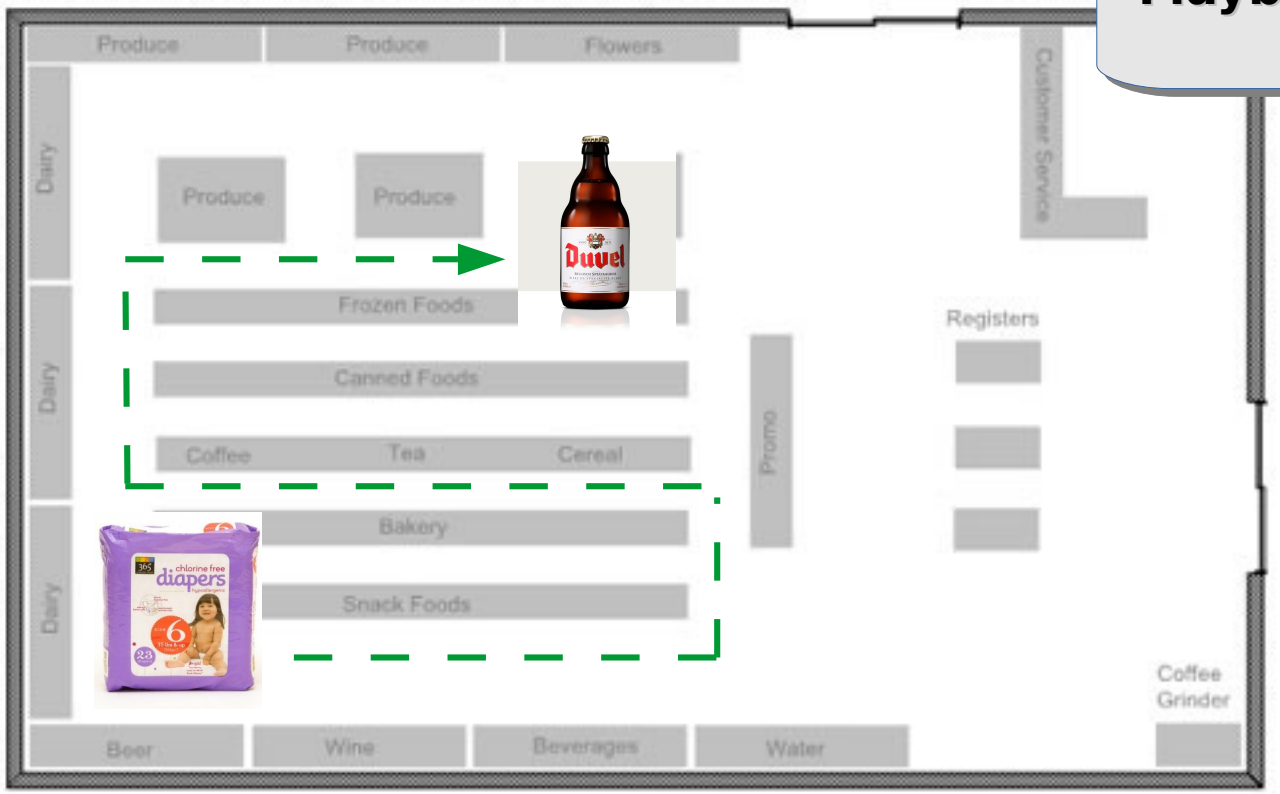
What do the patterns *mean* that we mine ?



Bought together because they are related

3)Understanding/ interpreting patterns – problem setting

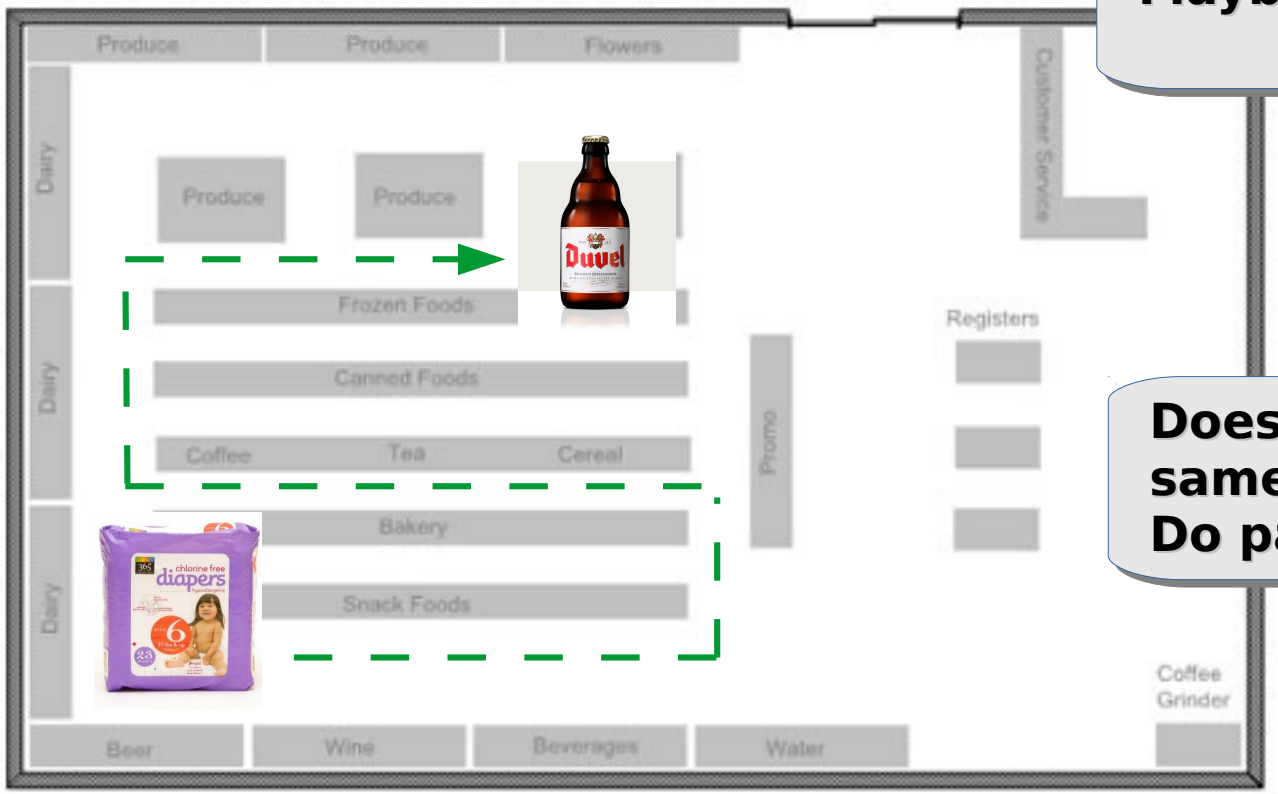
What do the patterns *mean* that we mine ?



Maybe a mix ?

3)Understanding/ interpreting patterns – problem setting

What do the patterns *mean* that we mine ?



Maybe a mix ?

Does data look the same ?
Do patterns ?

Understanding patterns – evaluation

Ad-hoc recovery of embedded patterns

- Flanking run time experiments

Consulting domain experts

- Arguably biased

Humans see patterns everywhere !

Issue ?

No data !

...w/ known ground truth

- Prediction has labels
- Minimum : known patterns
- Better : known processes

Understanding pattern-data relation

Ground Truth	**Embedded patterns**	**Generative models**
	Zimmermann 2013, 2014	**Mampaey et al. 2013 Webb et al. 2014**

Comparing found to embedded

Comparing found to substructures of model

Proposal : generating data

QUEST led the way

- Allows for different characteristics
- Known patterns
- Controllable properties

Others took it up

- Non-systematically

We wouldn't be the first

Engineering

- Downs et al. 1993
- CORSIKA **Physics**
- Analytical sociology
- Pei et al. 2006
- SAT solving
- Pascal this morning

Fact : We're not very good at it

QUEST !

Wrong transaction length distribution

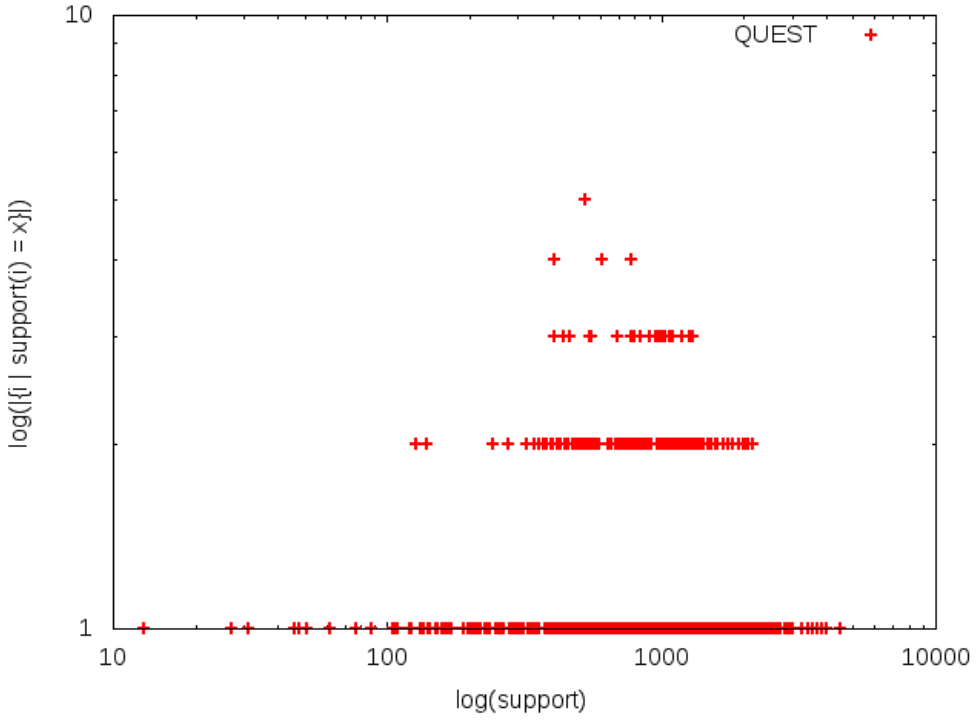
Replacement ?
Cooper 2009 ?

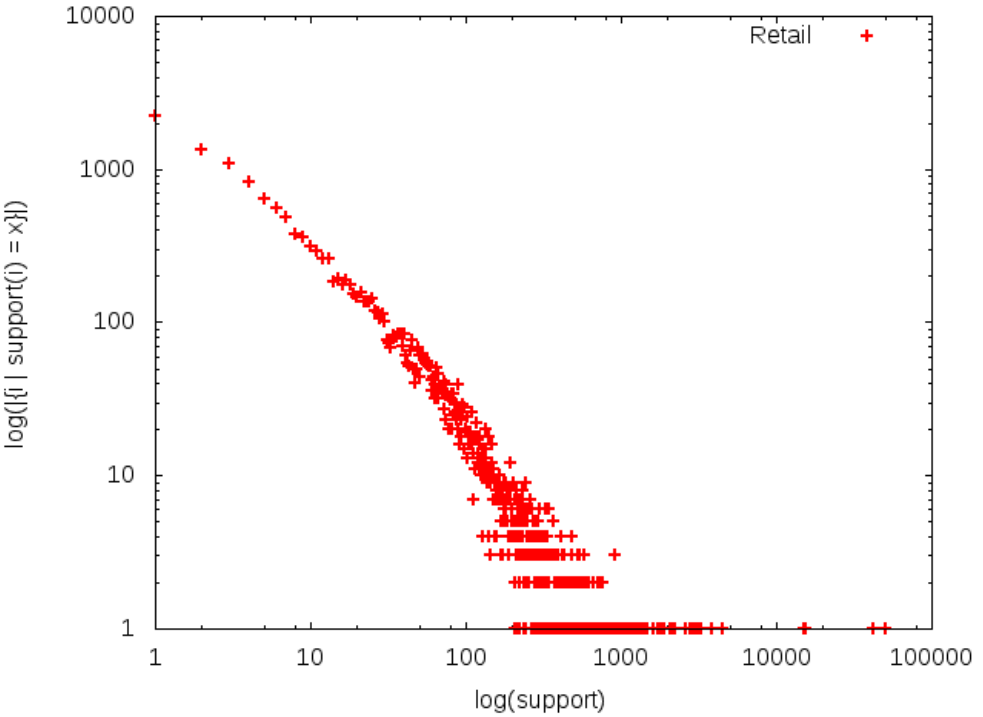
Fact : We're not very good at it

QUEST !

Wrong transaction length distribution

Replacement ?
Cooper 2009 ?



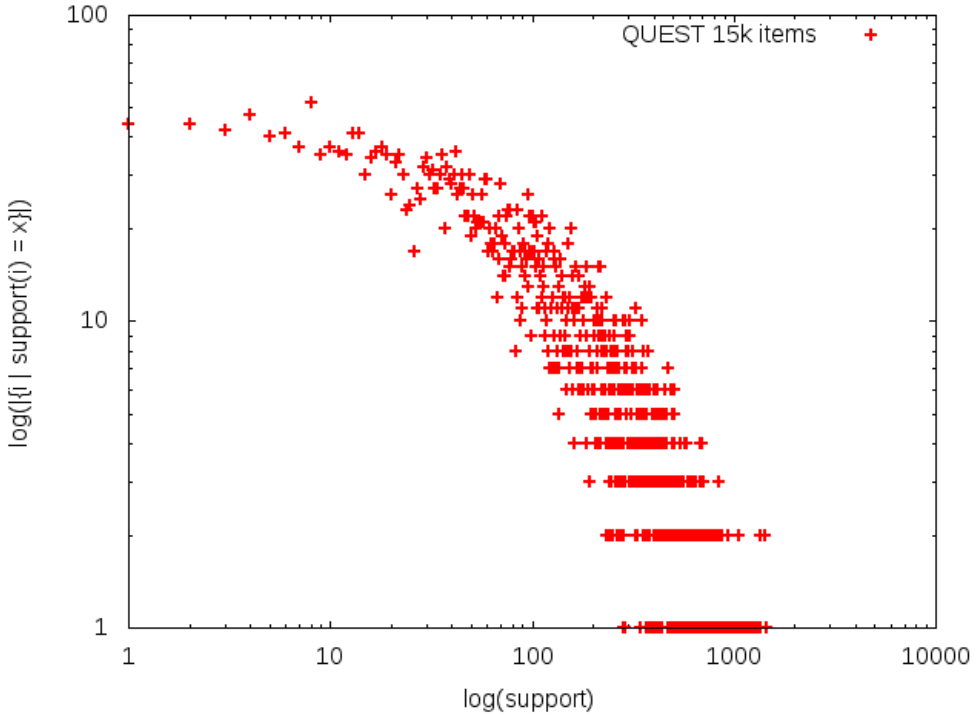


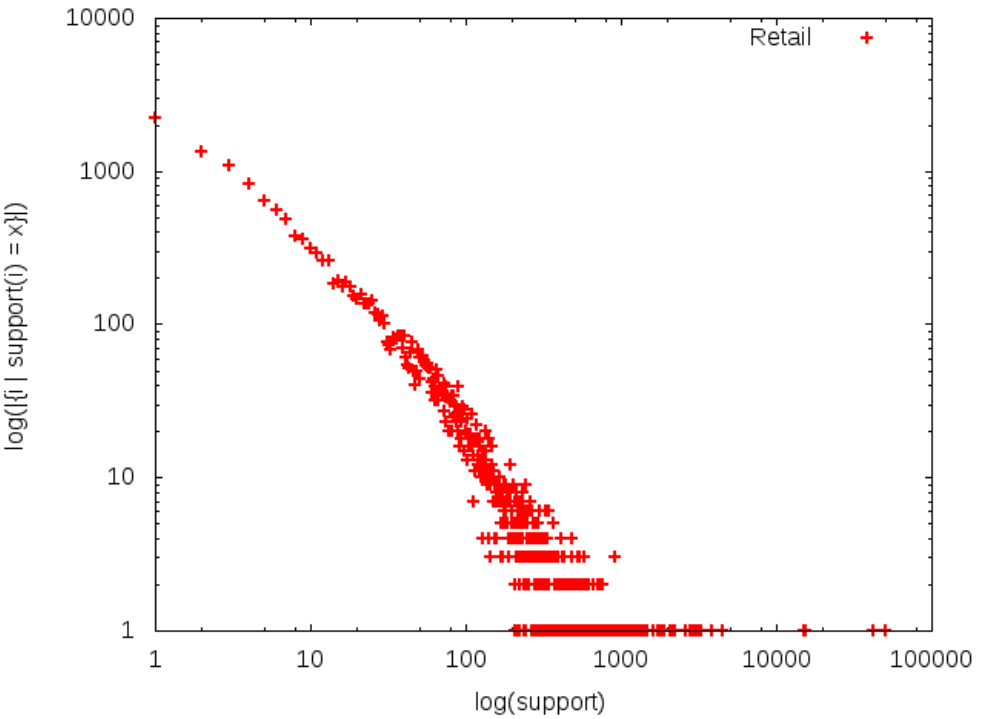
Fact : We're not very good at it

QUEST !

Wrong transaction length distribution

Replacement ?
Cooper 2009 ?

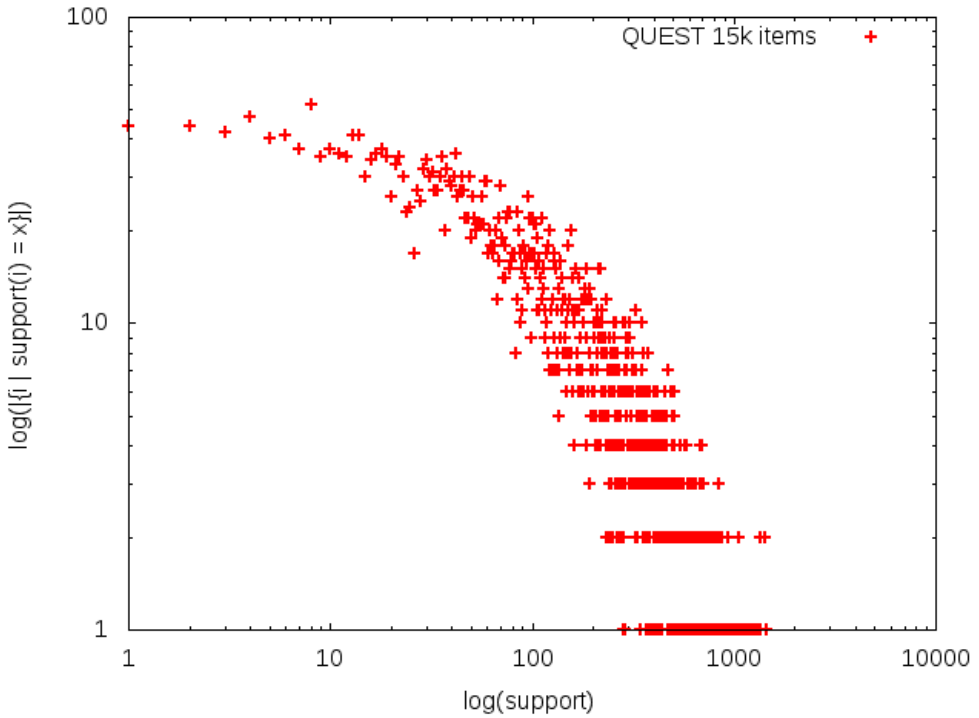




Fact : We're not very good at it

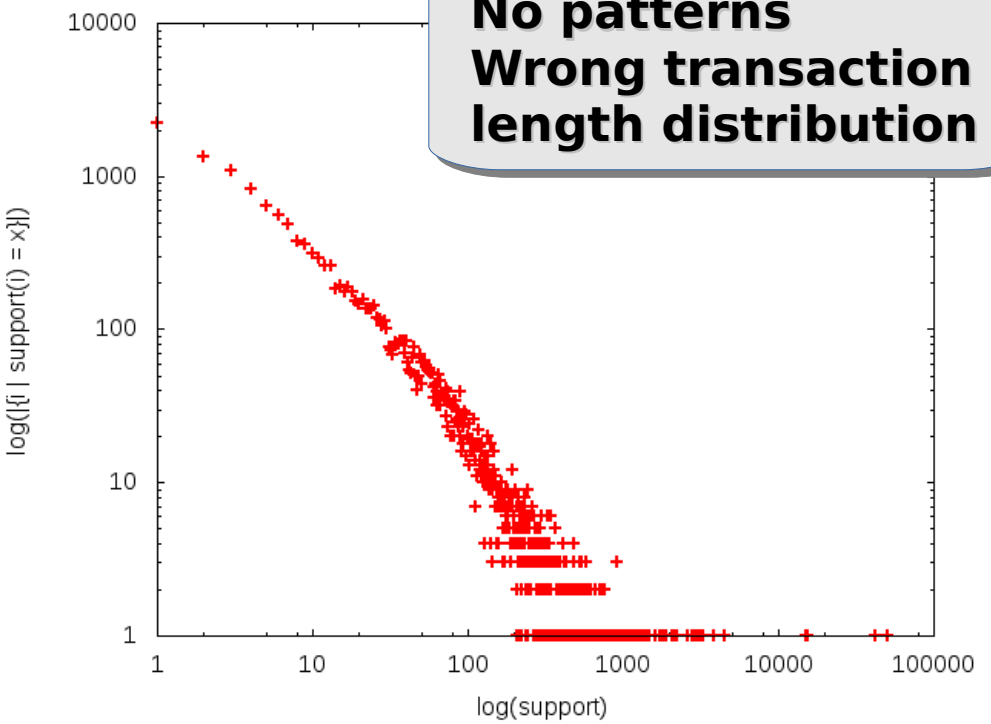
QUEST !

Wrong transaction length distribution



Replacement ?
Cooper 2009 ?

No patterns
Wrong transaction length distribution

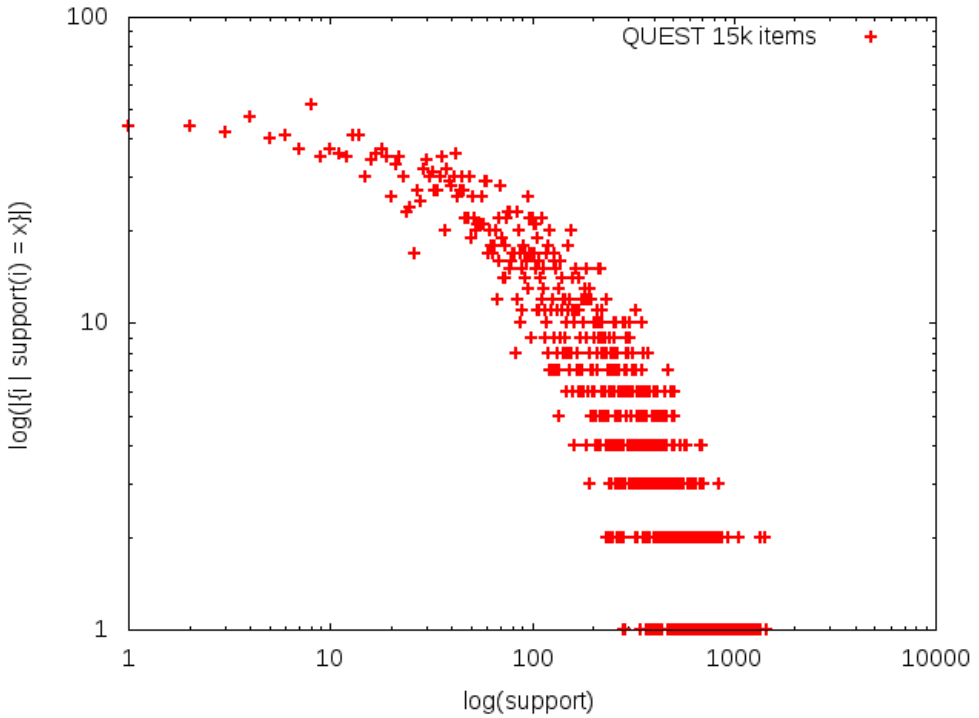


Fact : We're not very good at it

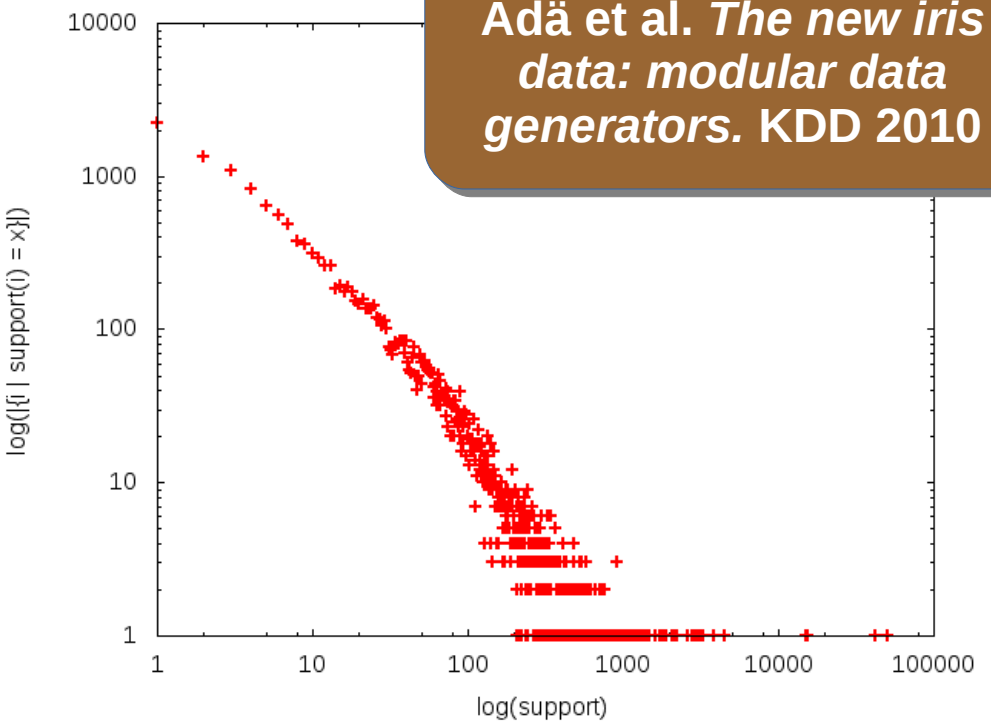
QUEST !

Wrong transaction length distribution

Replacement ?
Cooper 2009 ?



Adä et al. *The new iris data: modular data generators.* KDD 2010



Fact : We're not very good at it

QUEST !

Wrong transaction length distribution

Replacement ? Cooper 2009 ?

Adä et al. *The new iris data: modular data generators.* KDD 2010

Inverse itemset mining

No patterns
Depend on existing data

Chakrabarti et al. 2006

20 graph generators
None gets it right
No patterns

Challenge : recoverable patterns/realistic processes

Sampling a degree distribution is **not** realistic !

- And one cannot compare a subgraph to it

Patterns don't have to be the same as data

- Bayes' Nets to build itemsets
- Sequences to build graphs
- Agents !

Challenge : describing/ comparing data

The MetaLearners know about this :)

When are data sets similar ?

- Column/row marginals not enough
- Density not enough
- Krimp code tables sunk by relabeling
- Comparisons for structured data ?

Should similar data lead to similar results ?

Challenge : making the right assumptions

Need to talk to specialists in the field

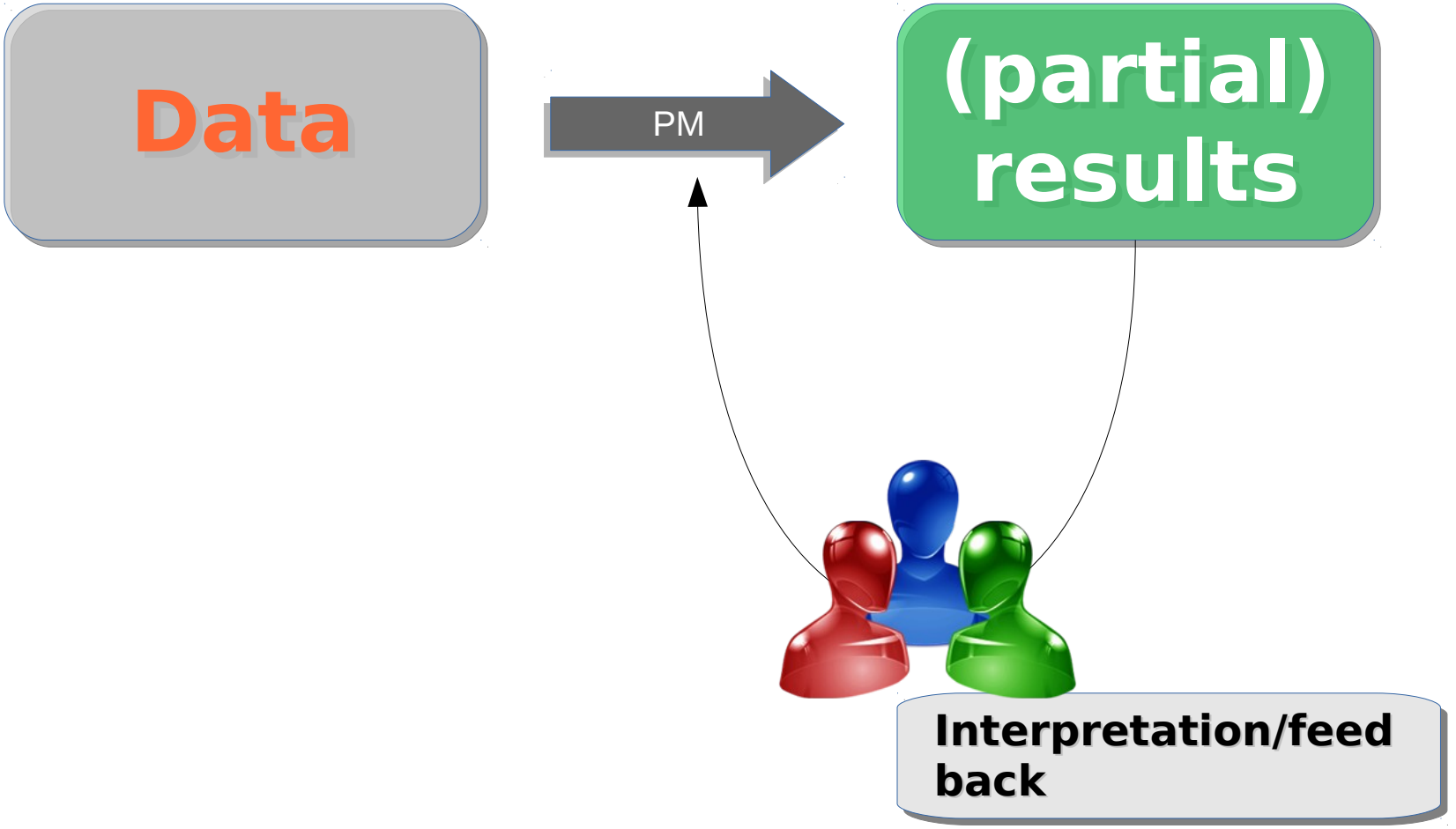
- Who might lack the full picture
- Who might be wrong

Patterns might (seem to) make no sense

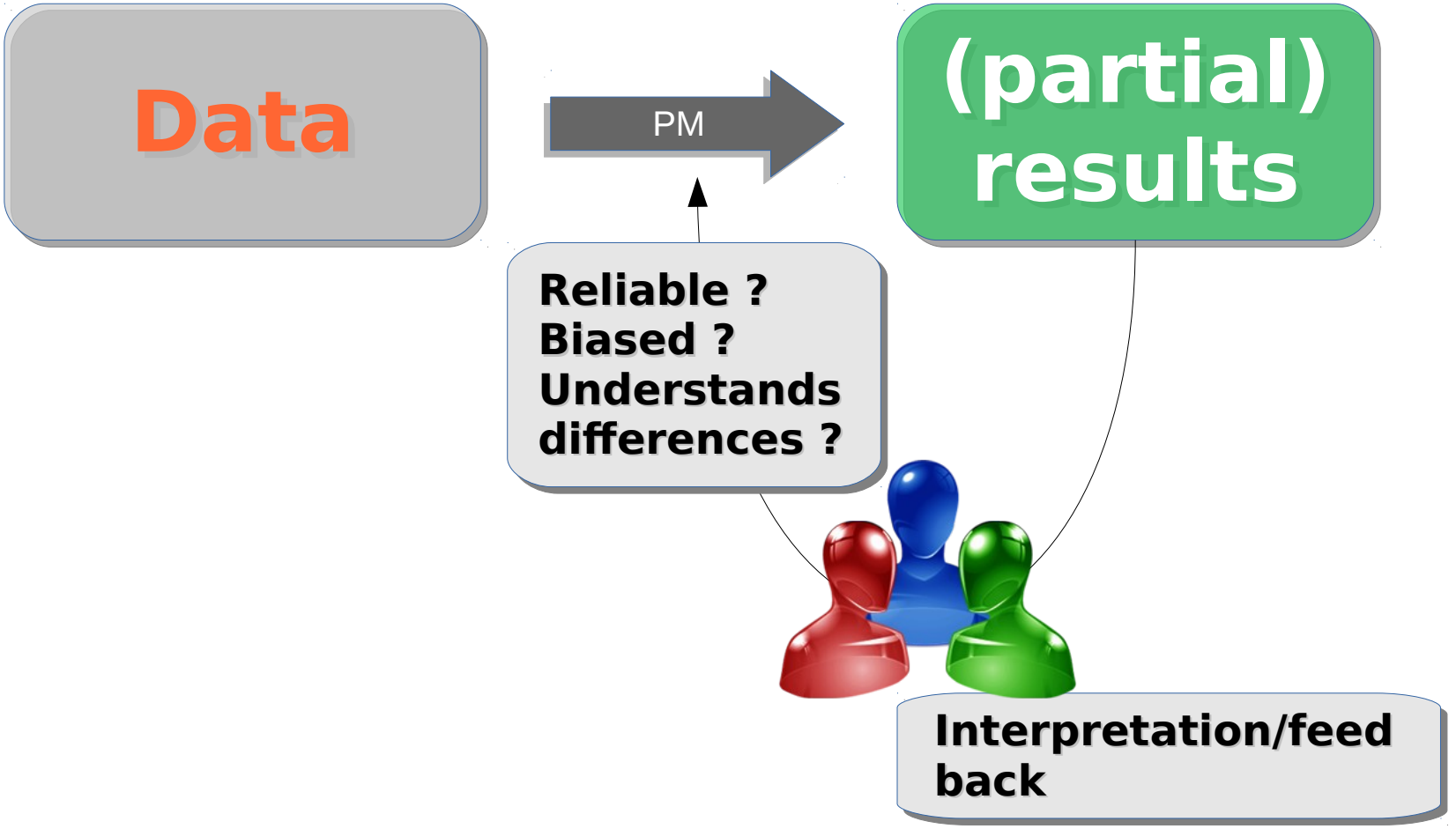
We'll have to go back time and again

CORSIKA is work in progress...

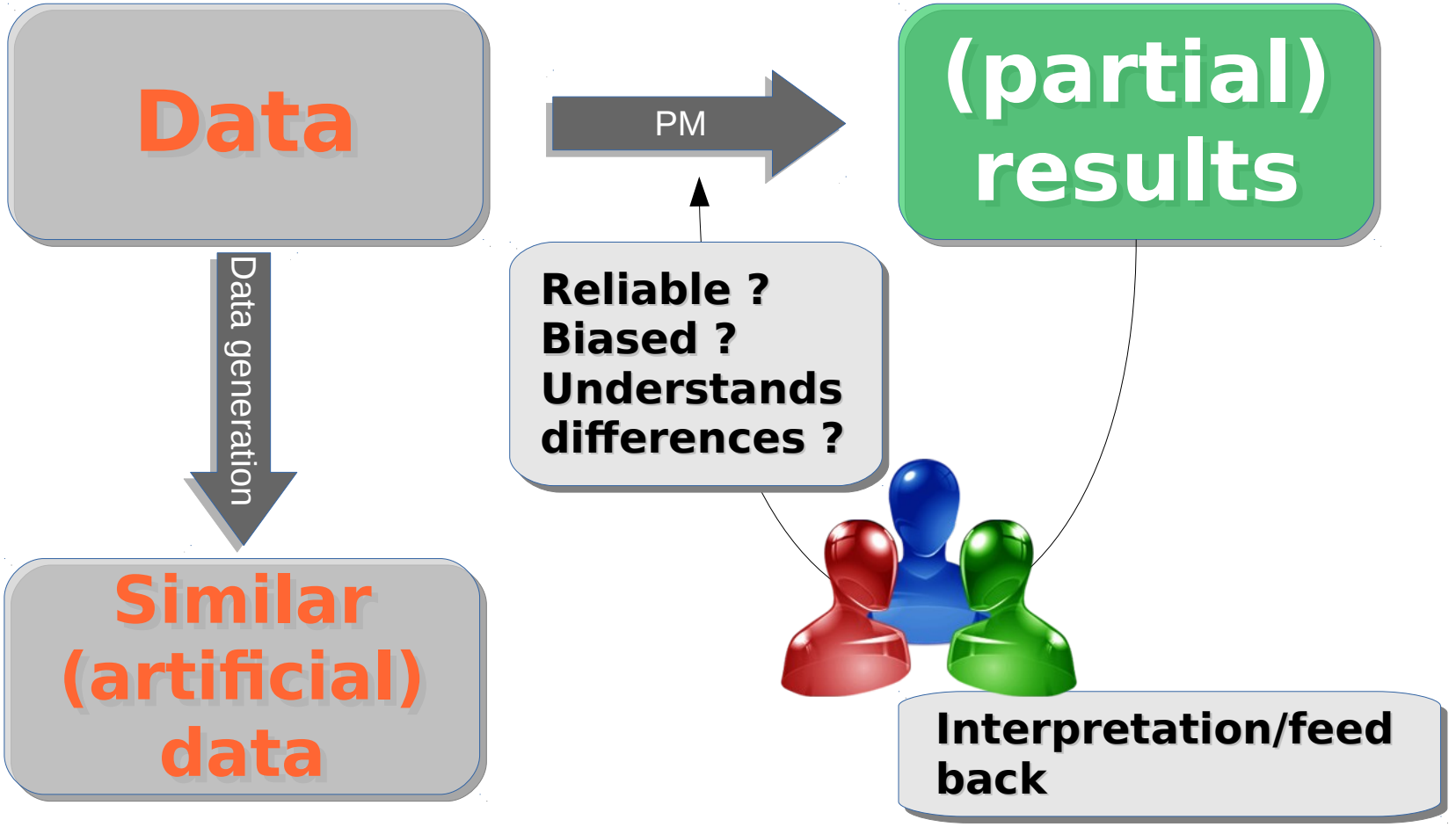
Thought experiment – interactive mining



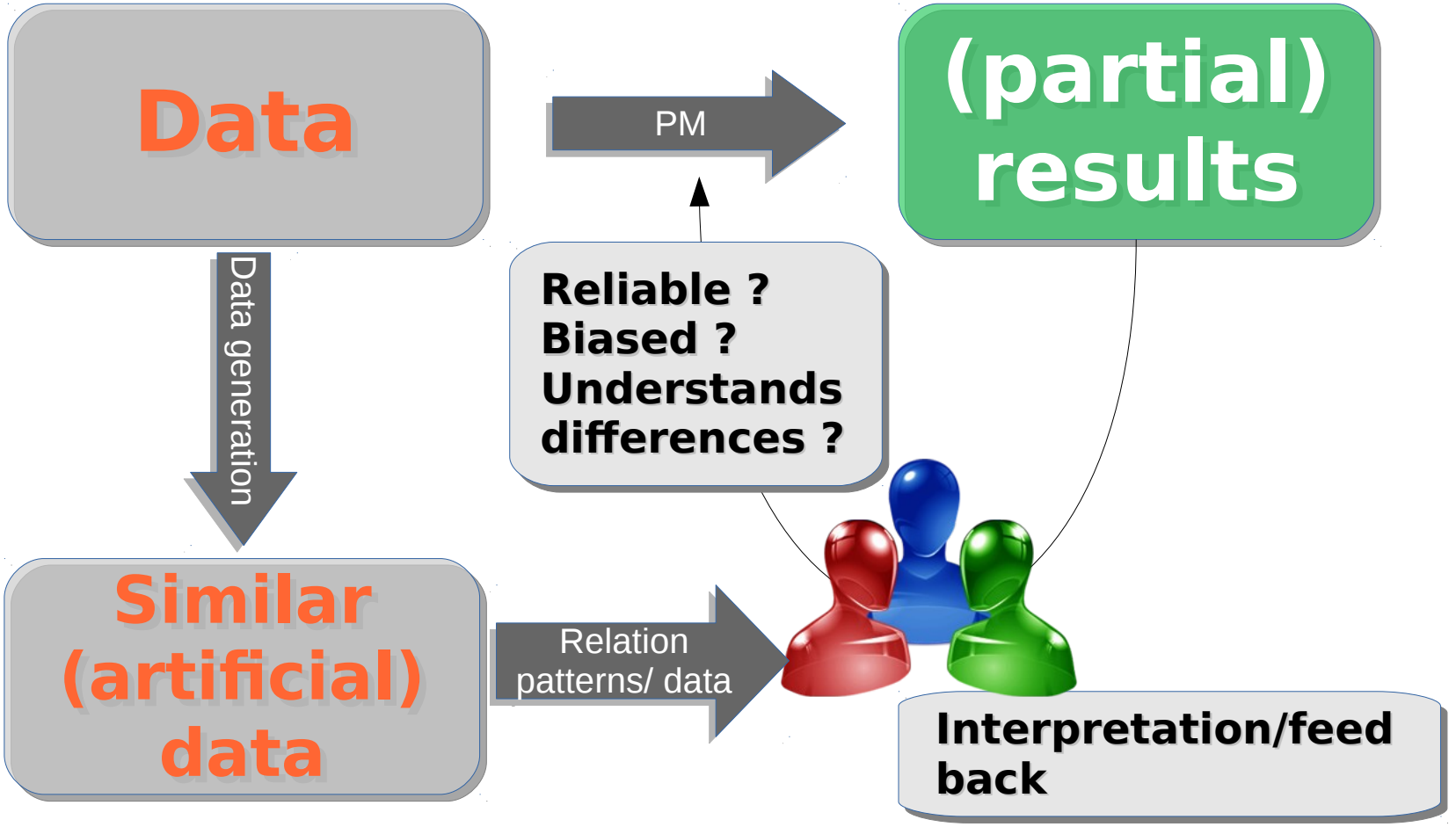
Thought experiment – interactive mining



Thought experiment – interactive mining



Thought experiment – interactive mining



What do you think ?

The Data Problem in Data Mining.
SIGKDD Explorations 16 (2)

- This slide intentionally left blank