# Subgroup Mining

(with a brief intro to Data Mining)

Paulo J Azevedo

INESC-Tec, HASLab

Departamento de Informática

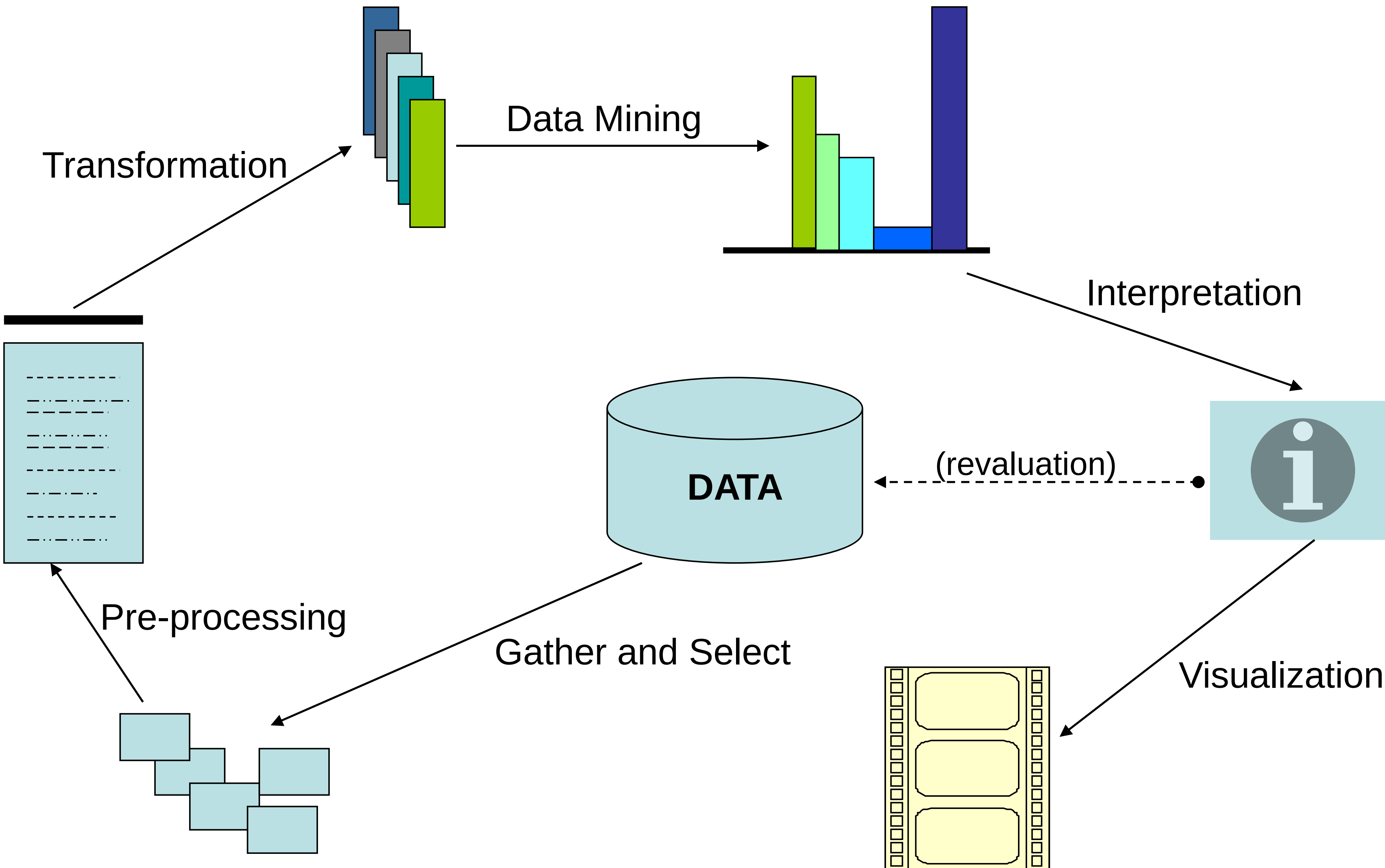
# Expectations…

# Knowledge Discovery (non trivial relation between data elements) in DataBases

- My personal perspective: it is the task of developing new algorithms (processes) to extract structure from data!
- Structure is alias for statistical patterns, models, relations, etc. It is a reduction process and leads to data summarization

- **Data gathering and Pre-processing**
- **Data Mining** (extracting the structure)
- **Post-processing e results analysis**
- **Visualization**

# The process


Data Mining
Transformation
Interpretation
DATA
(revaluation)
Pre-processing
Gather and Select
Visualization

# Pattern Mining

- Extract interesting/surprising patterns from data (relations between atomic elements in the data)

- Patterns <=> structure in data

- Brute force like methods
  - Example: Motifs in sequential DNA data – highly conserved DNA fragments (high occurrence ) along different genes.

- Several types of patterns:
  - Frequent patterns (association rules),
  - Sequences and Motifs
  - Subgraphs
  - Times Series (motifs)
  - etc.

# Frequent Patterns

- *Ticket Data Database:*
- *Ex:*

*1 1901,1881,199,901*

*2 901,1661*

*3 676,199,177,100*

…..

…

120099 78,1881,199,8

item

Transaction id

- The Marketing department intends to perform a shopping behavior study amongs costumers in a supermarket..
- Data is in the form of sets of “shopping baskets” (basket data)
- Queries to be answered:
- What products related to the consumption of beer?
- How to described the population that consumes peenuts ?
- Where the cleaning products should be located in the shelves?
- How to relate product 1661 with 199 ?

901 & 199 → 1881 (s=0.3,conf=0.9)

incidence (support)

Predict ability (confidence)

# Interest Measures

- Pertinent associations are spotted by using an incidence measure
- Support is the most popular (to count itemsets)
- Rules are qualified using a metric of interest (predict ability, strength of a rule).
- Confidence is one example (conditional probability)
- The association rule:

901 & 707 → 1088 (s=0.3,conf=0.9)

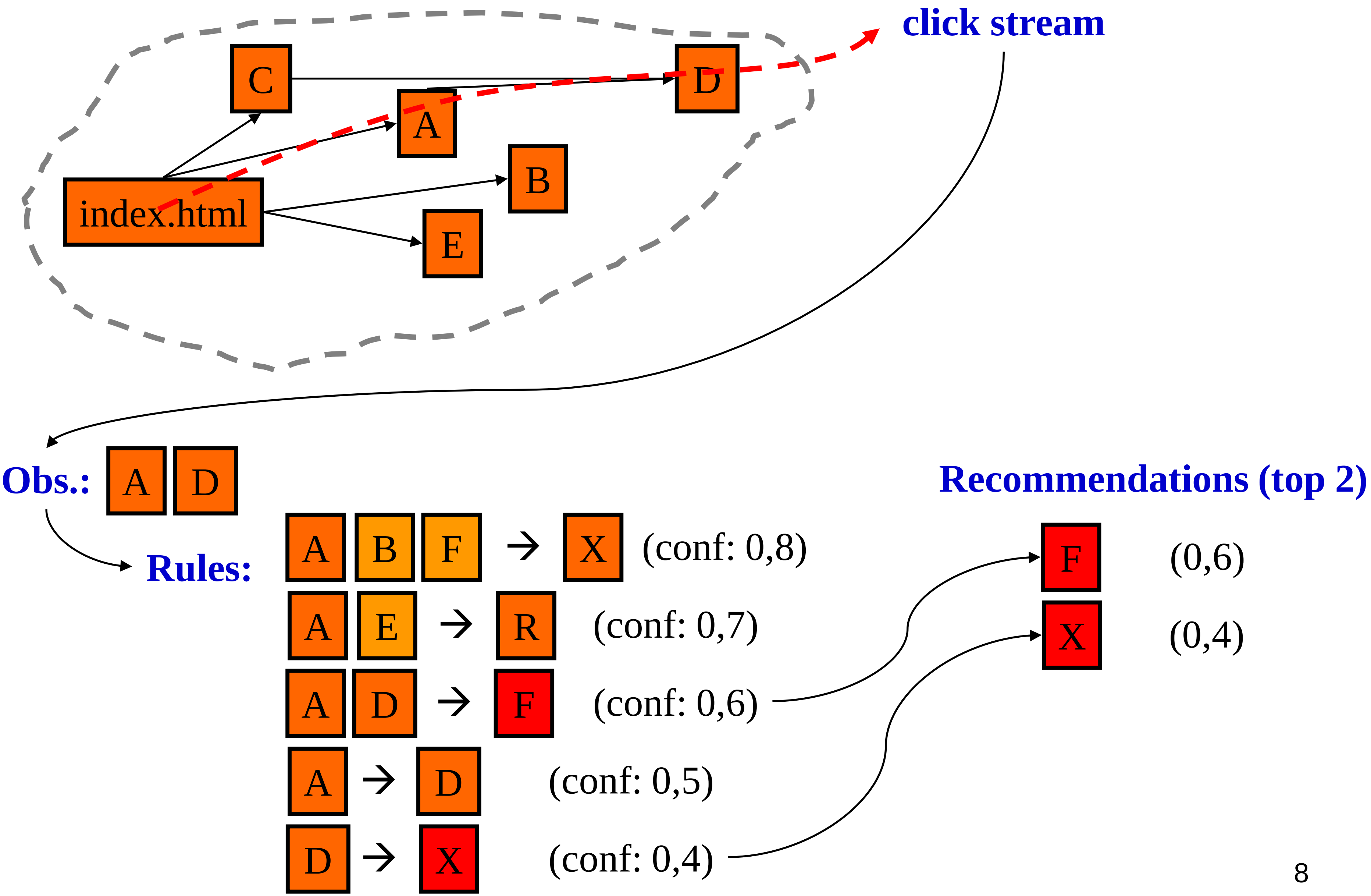
- Should be read as: buying products ***901***, ***707*** *e* ***1088*** *occurs in 30% of the transactions. On the other hand, 90% of the transactions that contain* ***901*** *e* ***707*** *also contain product* ***1088****.*

- Other readings: *90% of the subpopulation defined by products* ***901*** *e* ***707*** *also consume* ***1088.***

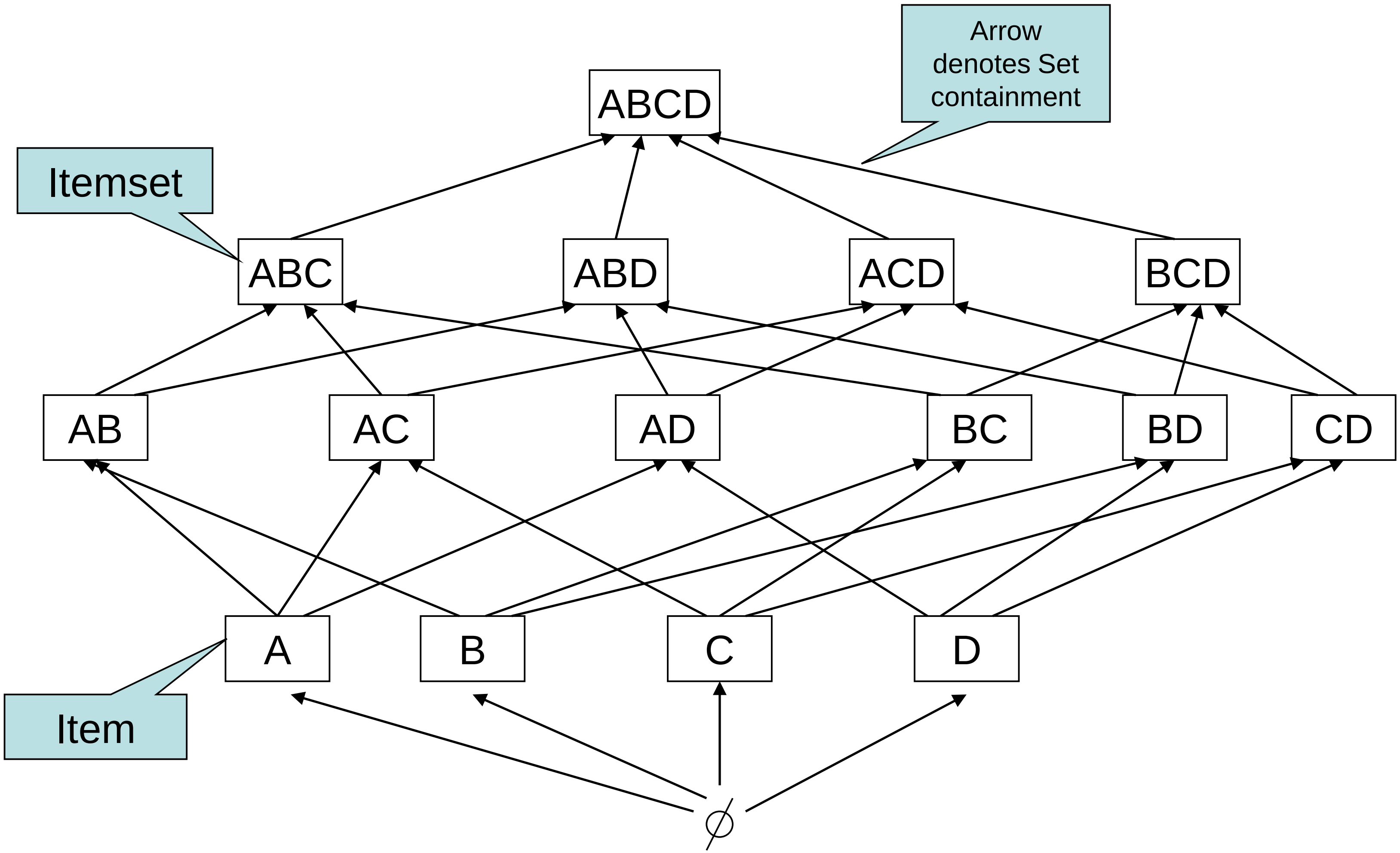
- *Or even: the consumption counting of pr*... ***088*** *is also considered goes down 90% of the initia*...

Relevant to this talk since we will be talking about subgroup mining

# Recommendation systems using ARs



# Search Space associated to frequent itemsets counting



# Algorithms

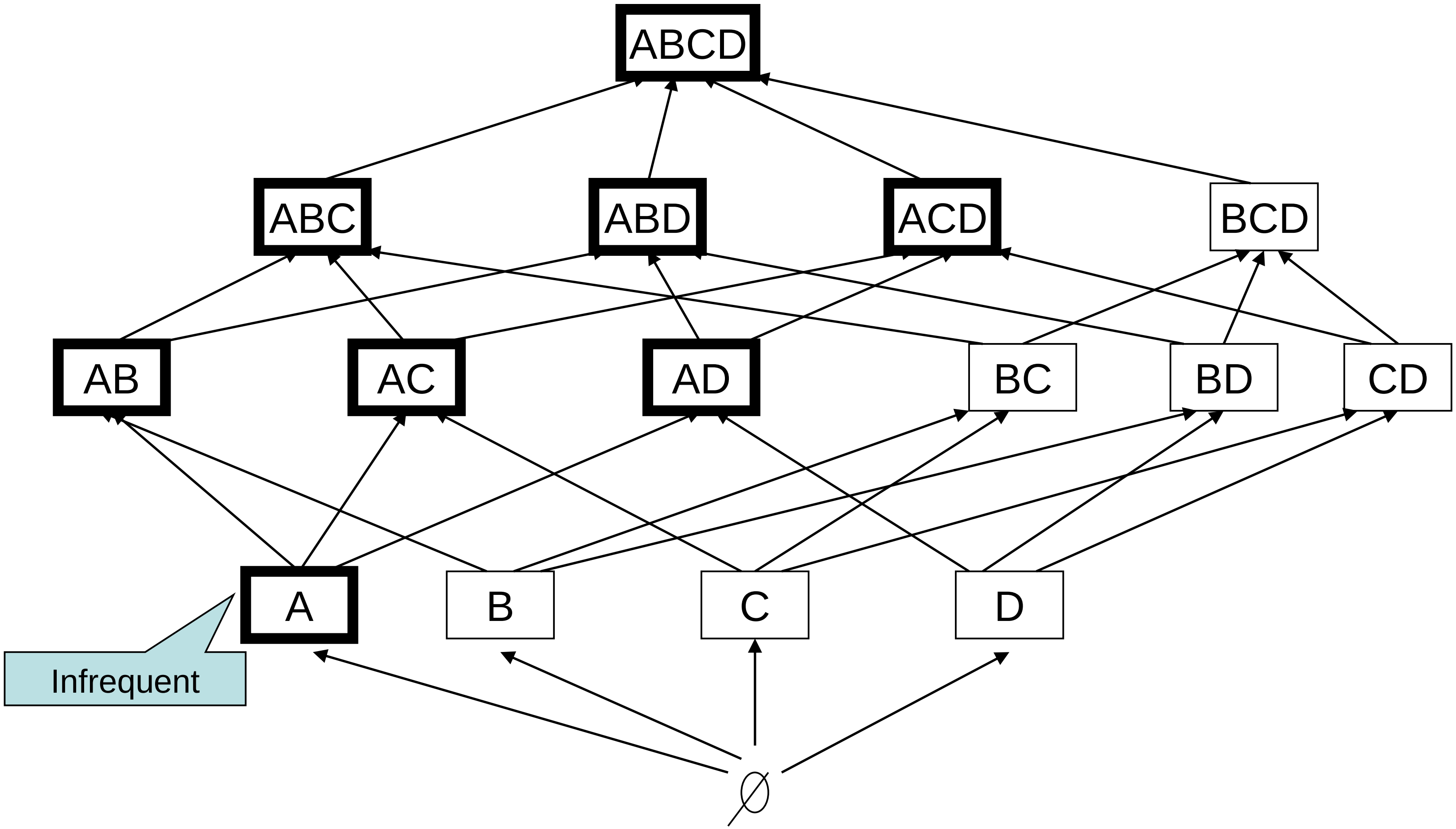
- Extract frequent terms (*itemsets*) i.e. associations (high complexity) following a user defined thresholds for occurrence (defines rarity!)
- Derive rules (low complexity)
- Select the most “interesting” rules!

First problem well studied with hundred of proposals

Seminal paper: *Apriori* [Agrawal&Srikant94].
Make use of downward closure property of support:

$$\text{If } X \subseteq Y \text{ then } s(X) \geq s(Y)$$

# Effects of using downard closure of Support

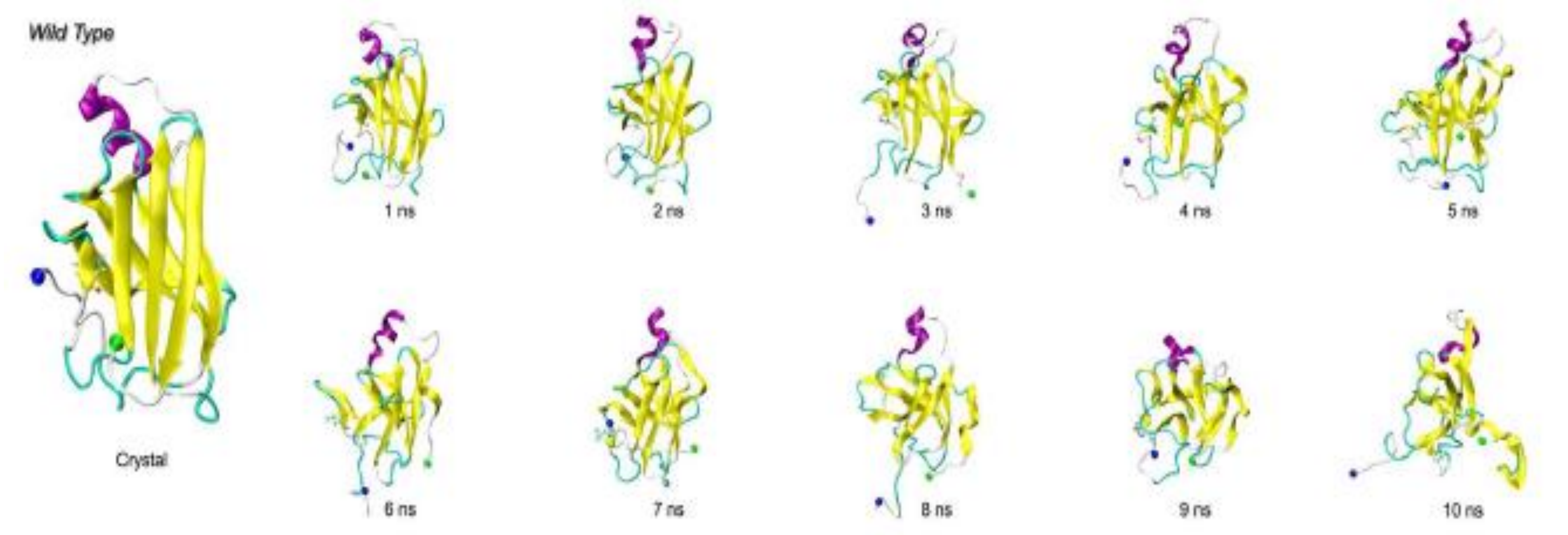


# Frequent Itemsets Algorithms

- Apriori-like bottom –up strategies
  - Several database scans
  - “Candidates generation”
- Depth-first strategies
  - Make use of vertical data representations e.g. TID Lists, bitmaps, diff-sets, etc.
  - Better pruning opportunities
  - “rule based” friendly

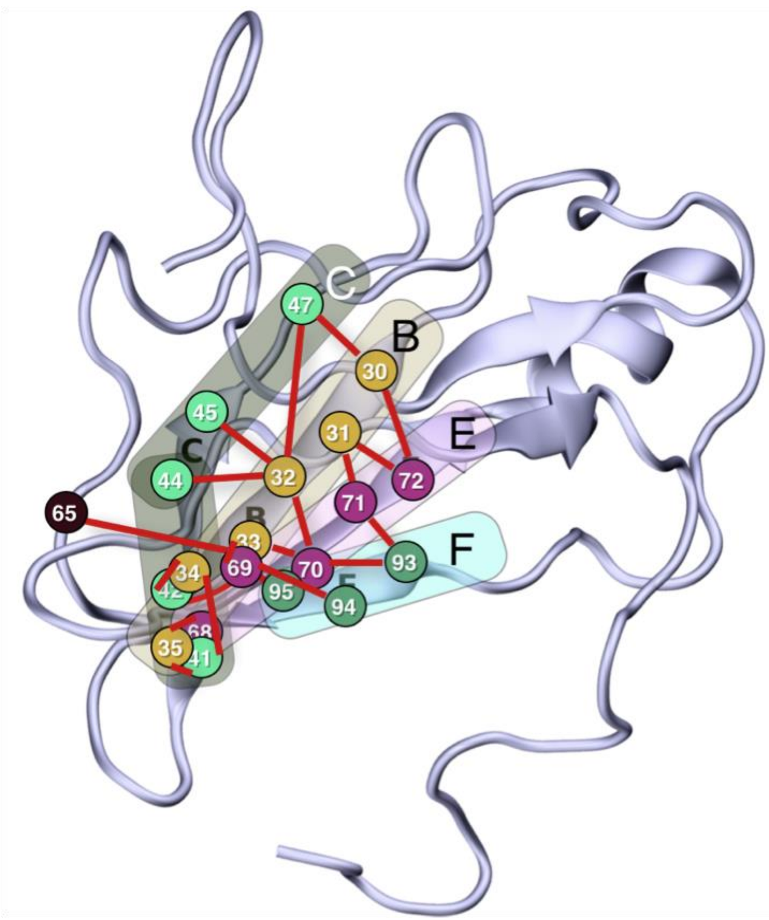
# Graph Mining

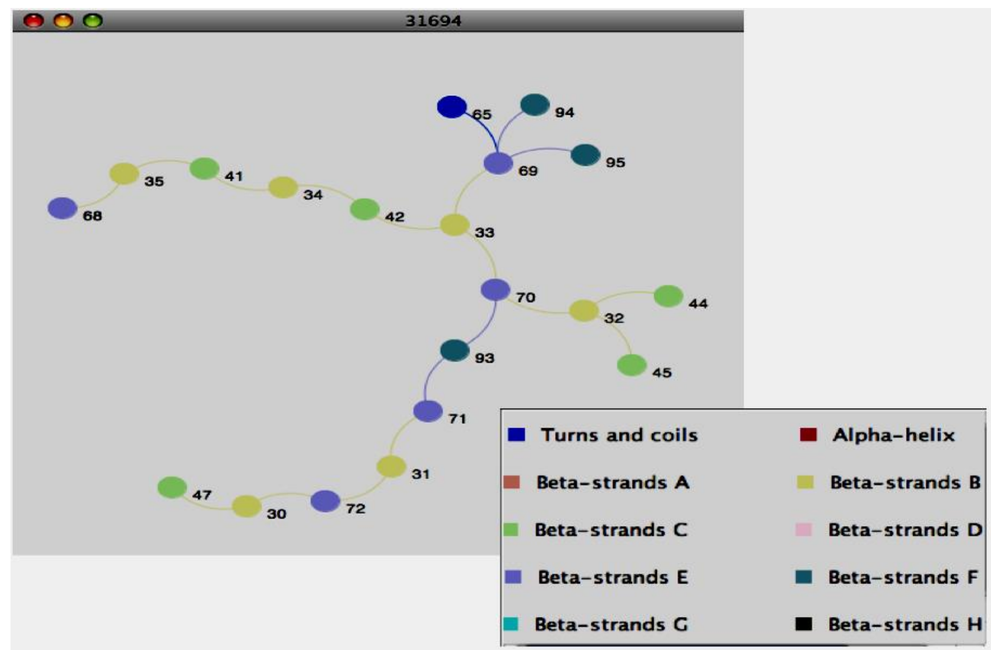
- It is the process of mining fragments (subgraphs) from a graph database e.g. protein conformation database.
- Applications in the web and other networks.
- A large number of biological applications e.g. Proteins unfolding process.



# Graph Mining

- Illustrating the identification of a subgraph that persists along the unfolding process.



# Problems?

# Maybe not!


New
Billy-Bob
Pacifier
teeth
Complies with EEC Directive 90/128/EEC
The Only Pacifier With Personality

# False Discoveries

- Aim is to identify associations that occur in the phenomena that give rise to the data.

- Brute force like search process tends to yields a high risk of false discoveries
  – i.e. associations appear to exist in our sample but do not occur in the phenomena that lead to the data!

# False Discoveries

- Explosive number of rules!

- e.g. Retail data
  - 3182 products (items)
  - $2^{3182}$ number of possible rules!
  - #rules with <= 4 items in the antecedent > $10^{16}$
  - High probability of unlikely real co-occurrence i.e they are false discoveries!

# Examples

- Redundant rules: (item in the antecedent explains other items)

Ex: pregnant → liquid_retention and
pregnant & female→ liquid_retention

Discard rule $x \rightarrow y$ if:
$\exists z \in x : supp(x \rightarrow y) = supp(x - z \rightarrow y)$

- Non Productive Rules:

male & high_psa & diabetes → prostate_cancer conf=0.84
male & high_psa → prostate_cancer conf=0.85

# Rules Pruning

- Identifying *improvement* in rules

| | | | |
|---|---|---|---|
| Conf = 0.300 | oranges | $\leftarrow$ | bananas & peaches |
| Conf = 0.315 | oranges | $\leftarrow$ | peaches |

- Definition of improvement:
  - A more specific rule has to induce an added value in terms of its interest measure.

$$imp(A \rightarrow C) = \min(\forall A' \subset A : met(A \rightarrow C) - met(A' \rightarrow C))$$

*met can be* ={conf,lift,conv,$X^2$,etc}

- If improvement > 0 we say that the rules are *productive*.

# Statistical Significance

- An alternative to define a minimal improvement is to apply a statistical significant test : discard rules that are *non significant*

- A rule $x \rightarrow y$ is *non significant* if

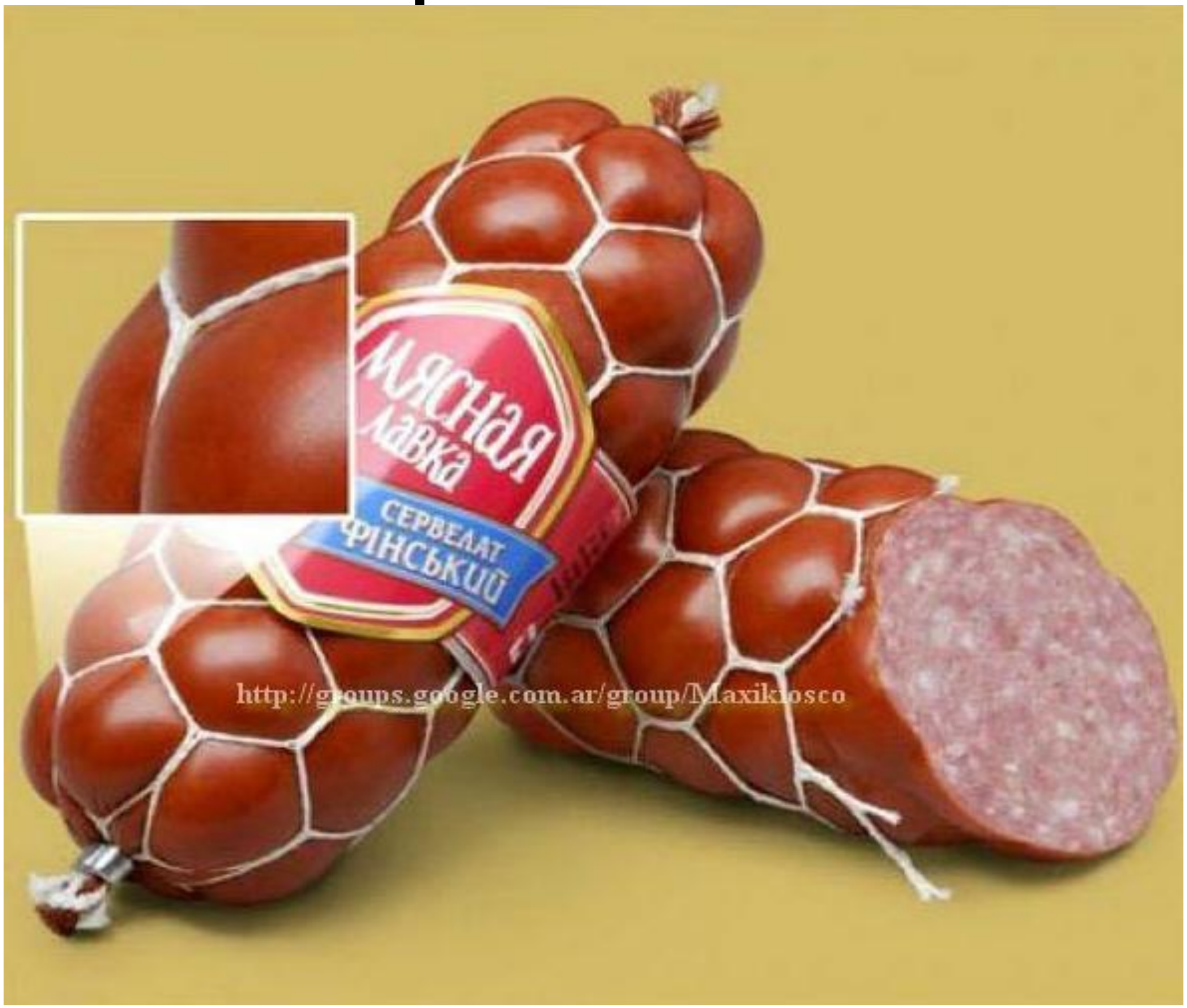
  – There exists another rule $x - z \rightarrow y$ where the value

  $met(x \rightarrow y) - met(x - z \rightarrow y)$

  is not significantly high (where $met(x \rightarrow y) > met(x\text{-}z \rightarrow y)$)

- Typically implemented using a Fisher exact Test.

# Expectations…




Can be deceiving!!!!!!!!

# Subgroups Mining

- To identify interesting subpopulations that occurred in our study.
- Represent these phenomena using specific patterns, for instance rules like;

Subgroup_description → poi

- Interesting means: deviates in relation to a reference (global) population
- Property of interest (poi) can be a numeric or categoric attribute, an constraint formula or even a contrast.
- Several statistics associated to each rule.

# Framework

- Make use of a association rule algorithm to extract interesting subgroups
- *rule-based algorithm.*
- Detect deviation using statistical significance
- Control specialization (*overfitting*) using the same type of statistical test

- Several types of subgroups/rules

# Numeric Properties of Interest

- Quantitative Association Rules
- [Aumann&Lindell2003]
- e.g:

smoker=n & wine_drinker=s ➔ life_expectation=85 (overall=80)

Sex=female ➔ Wage: mean=$7.9 (overall = $9.02)

- Rule interest is determined by a comparative test using poi average value and complement value.

# Distribution Rules [Jorge&Azevedo2007]

- The consequent is a distribution,
- Enables the distribution analysis according to other parameters like *Skewness* (degree of asymmetry) and *Kurtosis* (degree of sharpness).

- Make uses of the goodness of fit *Kolmogorov-Smirnov test*, to evaluate rule's interest.

- Definition of interest: Rule is interesting if p-value of

General population Distribution

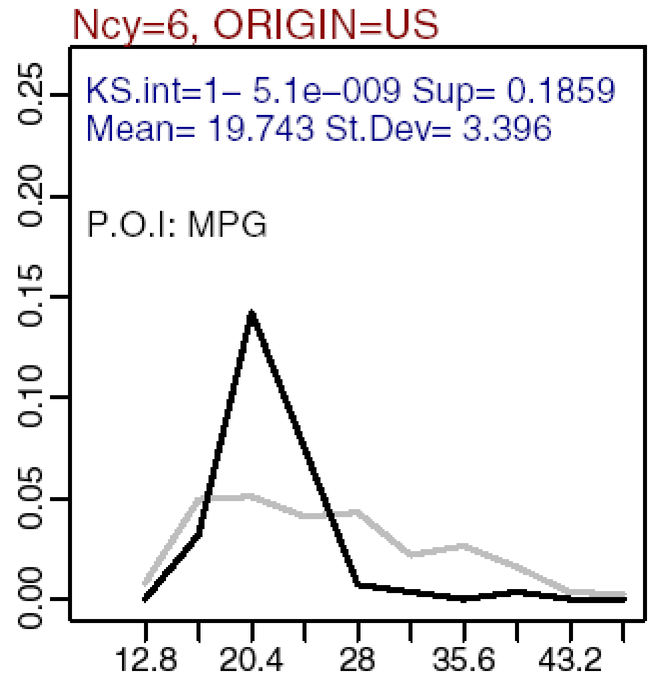
Rule subgroup distribution

ks-test(apriori,rules-dist) $< \alpha$

- Specialization of subpopulations also controlled by KS test (KS-improvement).
- Several applications

# Interest measure of a DR

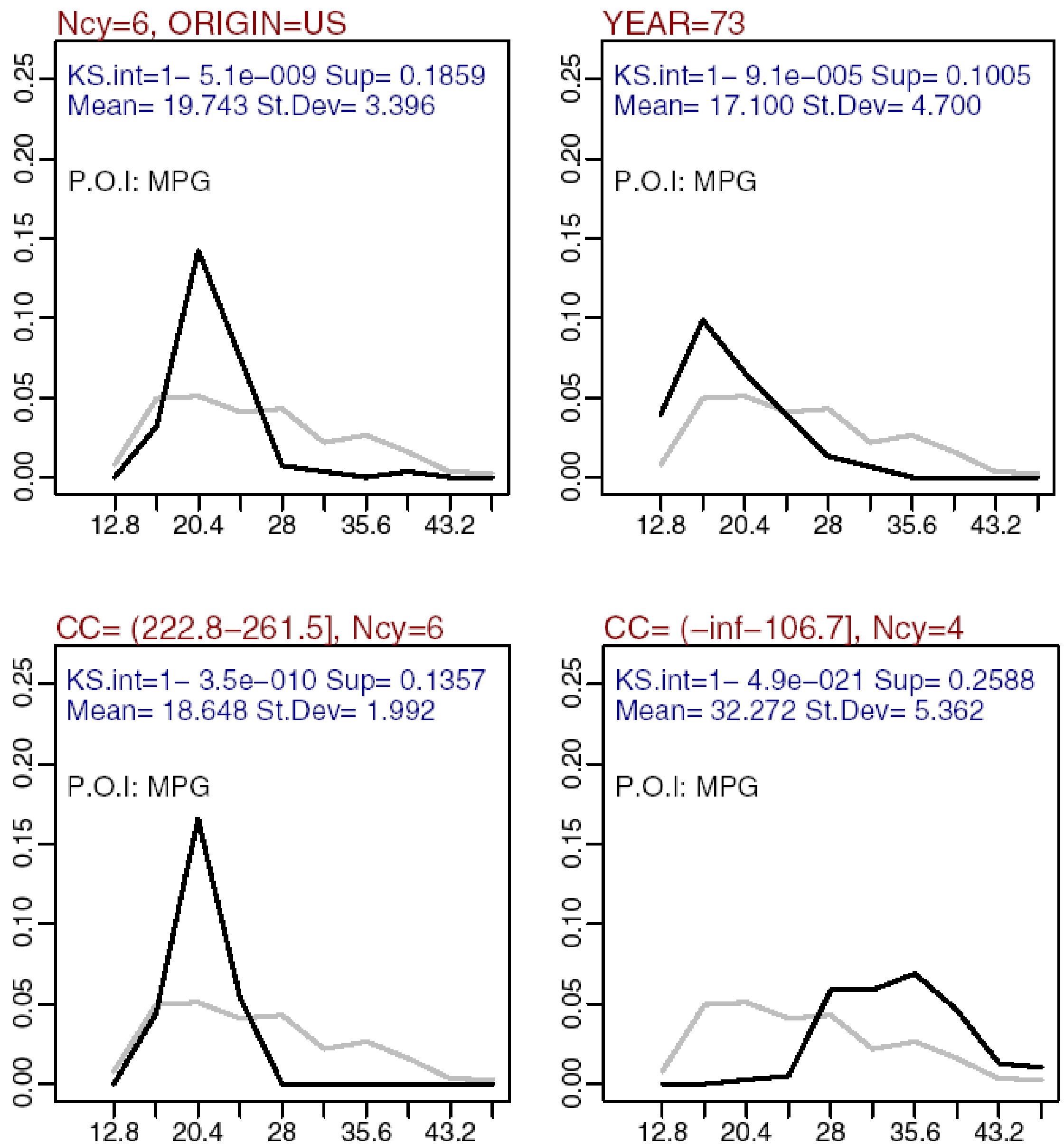
- KS-interest:
  - Given a rule **A→y=D**$_{y|A}$, its KS-interest is 1-p,
  - p is the p-value of the KS test comparing **D**$_{y|A}$ and **D**$_{y|\varnothing}$




- KS-improvement
  - value added by the refinements of a rule
  - imp(A→B) is
    min({KS-interest(A→B) - KS-interest(As→B) | As $\subseteq$A})
  - other variants to control refinements.

# Distribution Rule presentation

- property of interest
- each DR is a plot
- distribution plot
  - frequency polygon
  - static binning
- distribution statistics
- comparison with default distribution



# Case Study

- Descriptive data mining
  - dataset: Determinants of Wages from the 1985 Current Population Survey in the United States, a.k.a. Wages
  - property of interest: WAGE
- Rule discovery
  - min-sup=0.1, KS-int=0.95
  - numerical attributes in the antecedent were pre-discretized
  - compact internal representation of rules
  - rules can be output as text or graphically

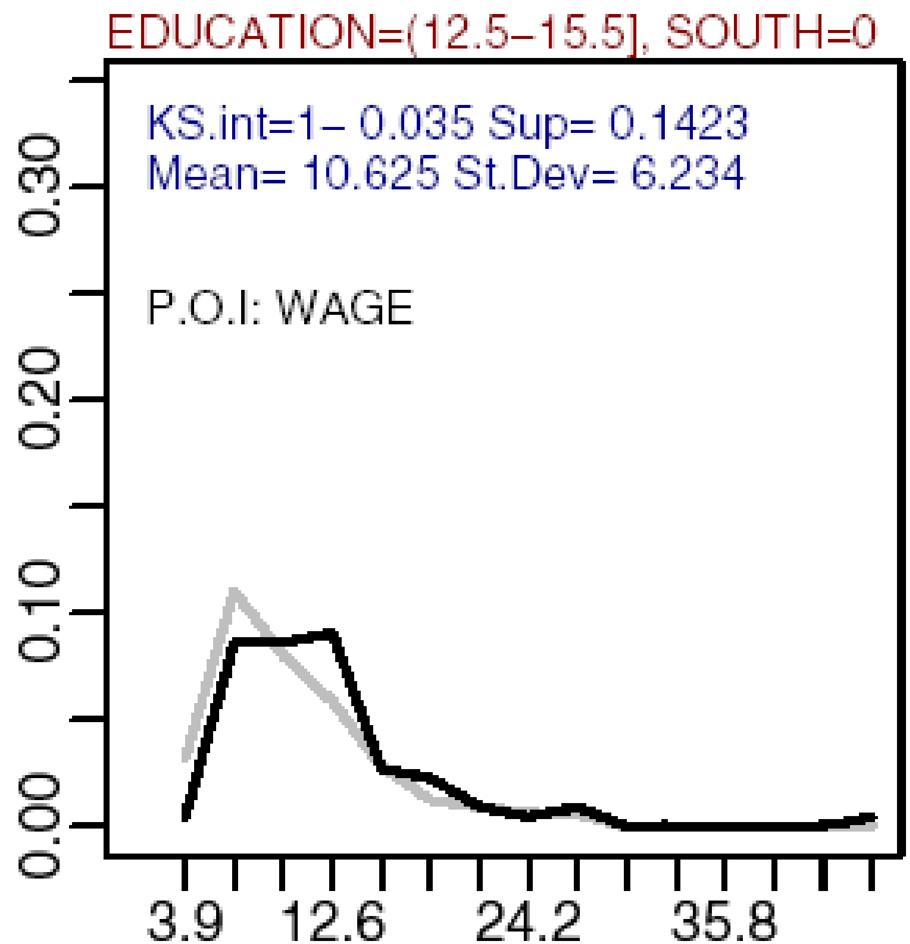
---

```
Sup=0.118 KS.int=1-0.0085 Mean=10.982 St.Dev=6.333
EDUCATION=(12.5-15.5] & SOUTH=0 & RACE=3
  ->  WAGE={ 3.98/1,4.0/1,4.17/1,4.5/1,4.55/1,4.84/1,5.0/1,5.62/1,5.65/1,5.8/1,6.0/1,6.25/4,7.14/1,7.5/1,7.67/1,7.7/1,7.96/1,
             8.0/2,8.4/1,8.56/1,8.63/1,8.75/1,8.9/1,9.22/1,9.63/1,9.75/1,9.86/1,10.0/3,10.25/1,10.5/1,10.53/1,10.58/1,10.61/1,
             11.11/1,11.25/2,12.0/1,12.47/1,12.5/4,13.07/1,13.75/1,13.98/1,14.29/1,15.0/1,16.0/1,16.14/1,16.42/1,17.25/1,17.86/1,
             18.5/1,21.25/1,22.5/1,26.0/1,44.5/1 }
```

---

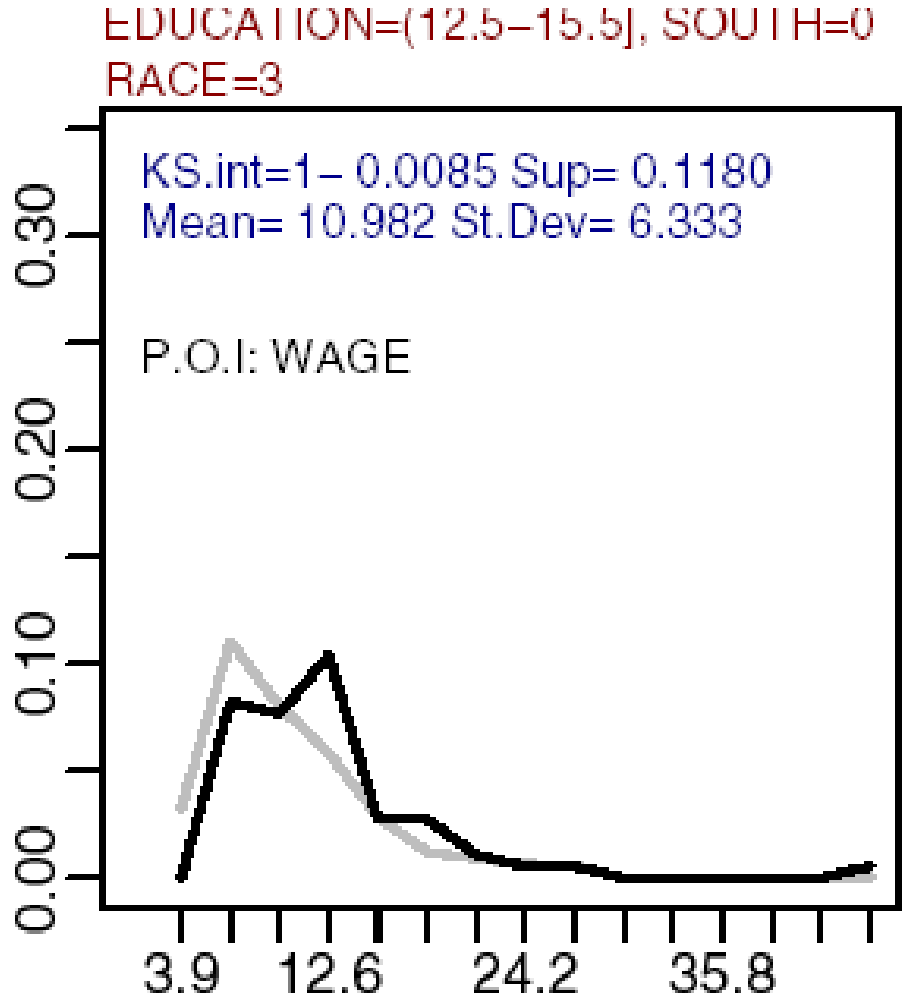
# Using Distribution Rules

- antecedent
  - people with 13 to 15 years of education
  - not from the south
- consequent
  - wage distribution is better than the whole population but still concentrated on the same interval



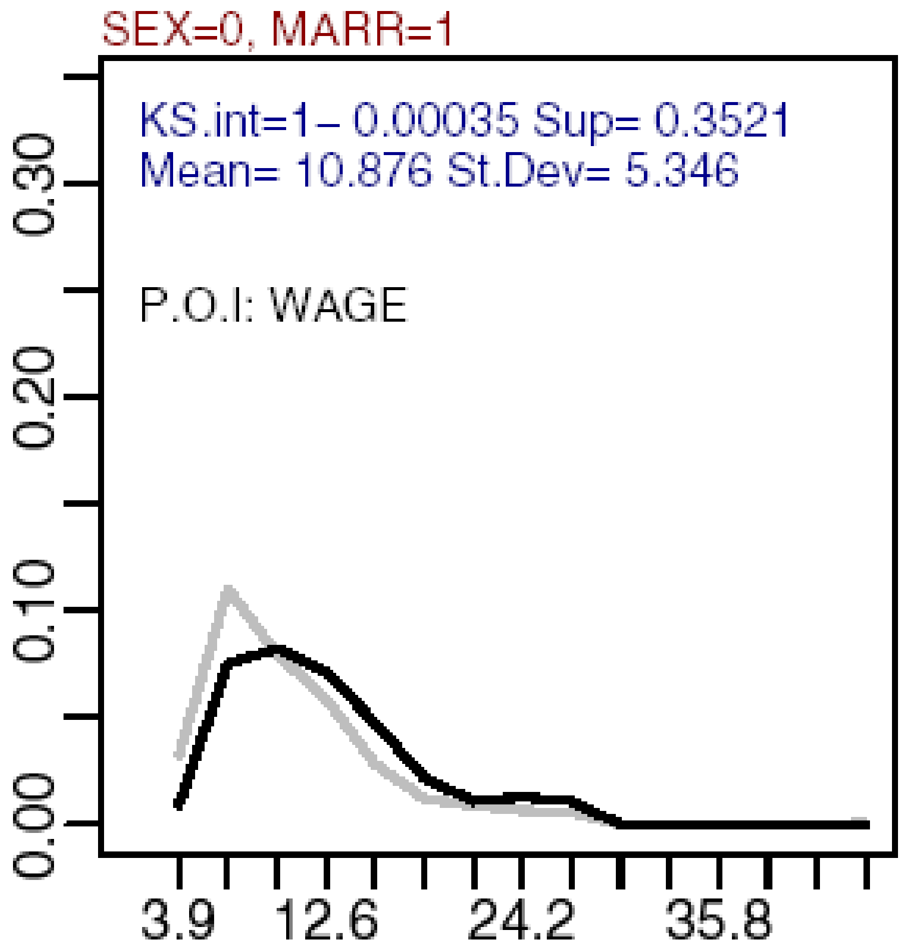
# Using Distribution Rules

- antecedent
  - refinement of previous
  - race is white
- consequent
  - wage distribution is even better than before
  - KS-improvement is higher than 0.01
  - the wages still are concentrated on the same interval as before



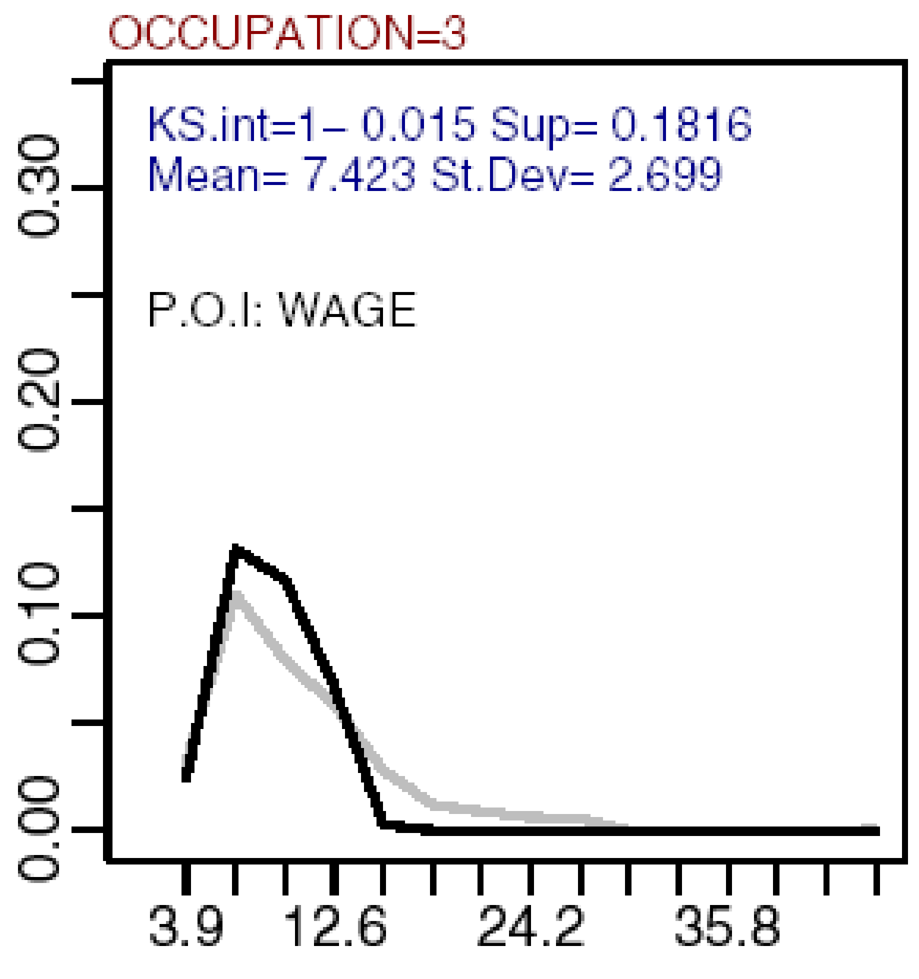
# Using Distribution Rules

- antecedent
  - married males
- consequent
  - less interesting
  - still signif. different



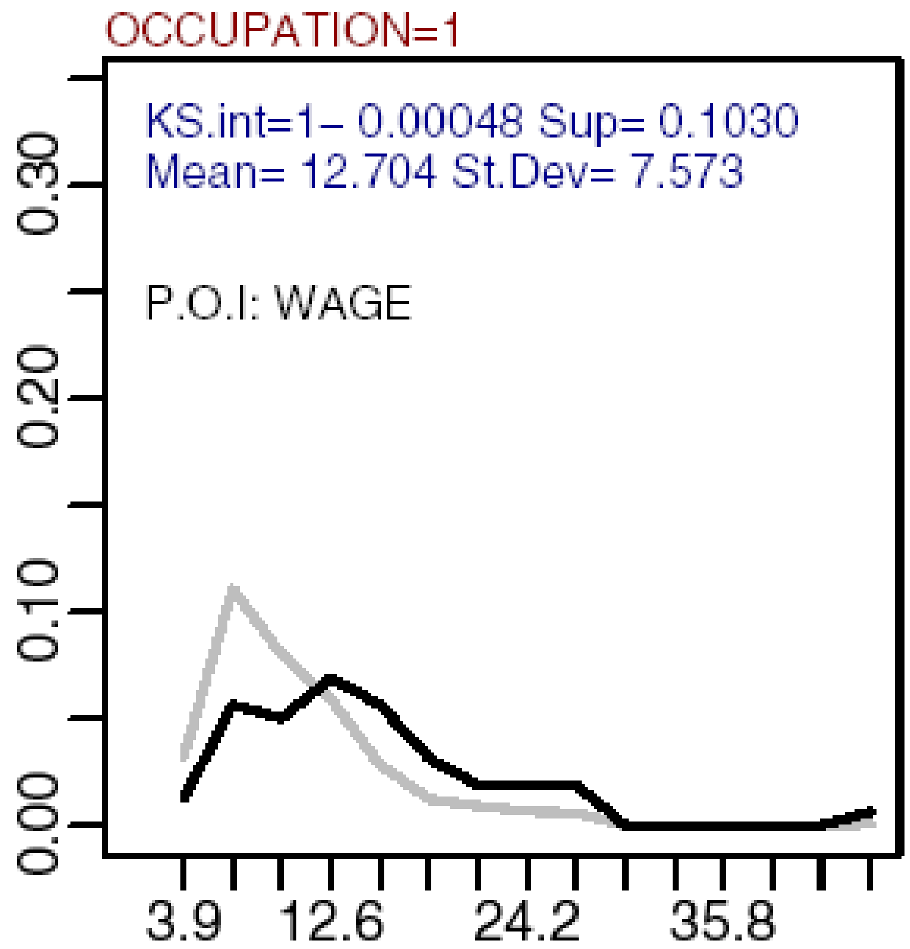
# Using Distribution Rules

- antecedent
  - Occupation=Clerical
- consequent
  - concentrated on lower income



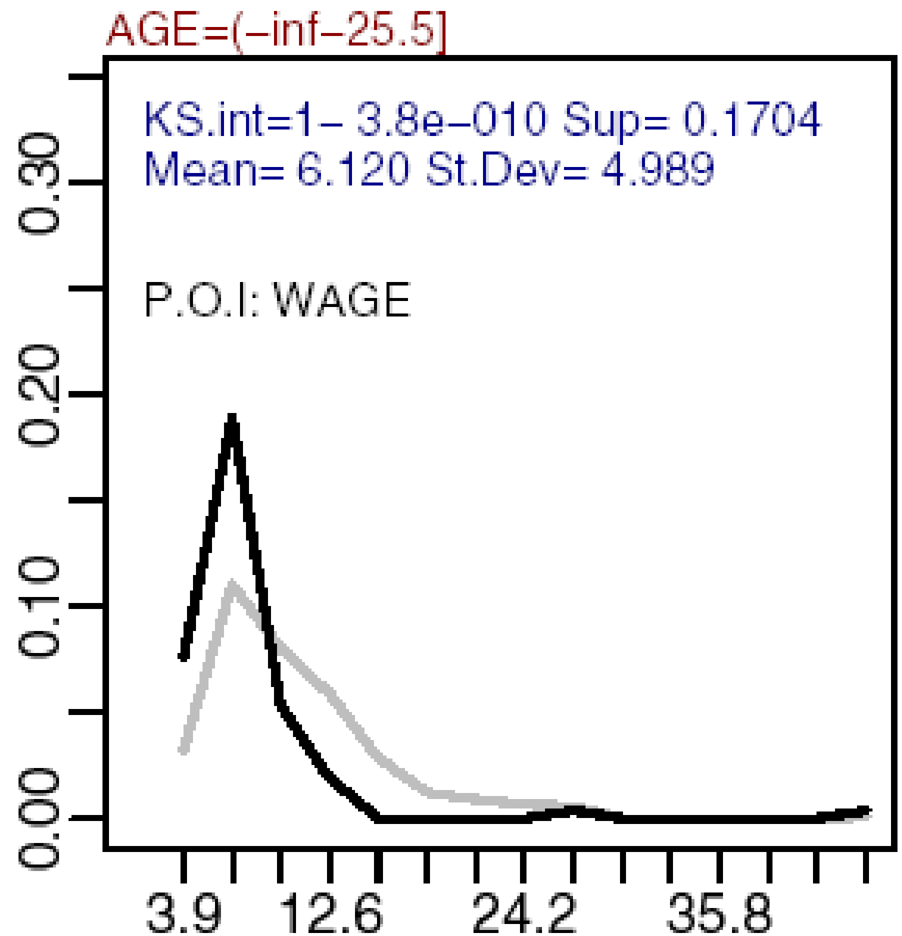
# Using Distribution Rules

- antecedent
  - Occupation=Management
- consequent
  - clearly better wage distribution
  - we also observe a slightly lifted right tail



# Using Distribution Rules

- antecedent
  - young people
- consequent
  - lower wages, very concentrated
  - some secondary modes are suggested



# Case Study (2): BUS trip time deviation study

## Trip time deviation study

- What are the **factors**, or combination of factors, that are associated with **significant deviations** in **trip duration**?

- Case:
  - Porto urban buses company
  - Line 205 (a circular line)

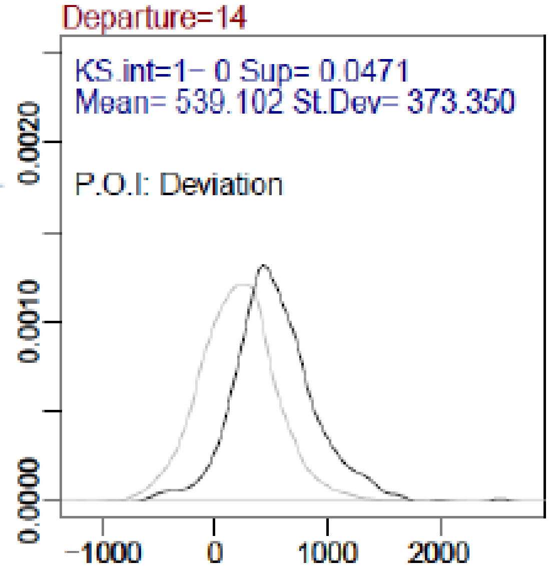
STCP

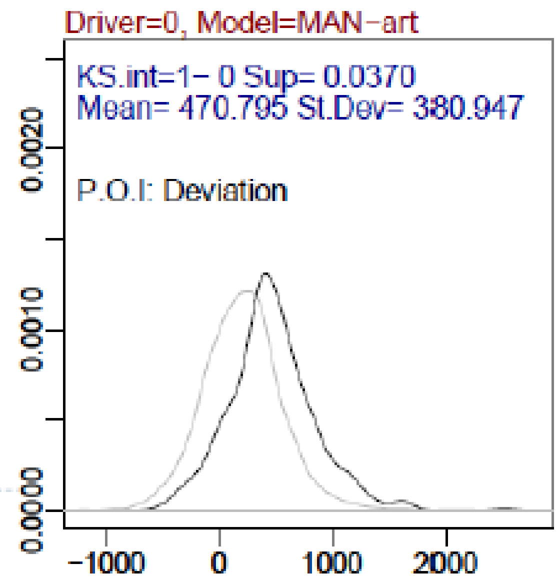
# Case Study (2):

## Delays: top rules

- the period after lunch appears frequently in discovered contexts.
  - Mean 9 minutes delay
  - Strong KS-interest
- change of shift and articulated vehicles are also related to delays
  - The latter case is probably due to use in rush hour

18



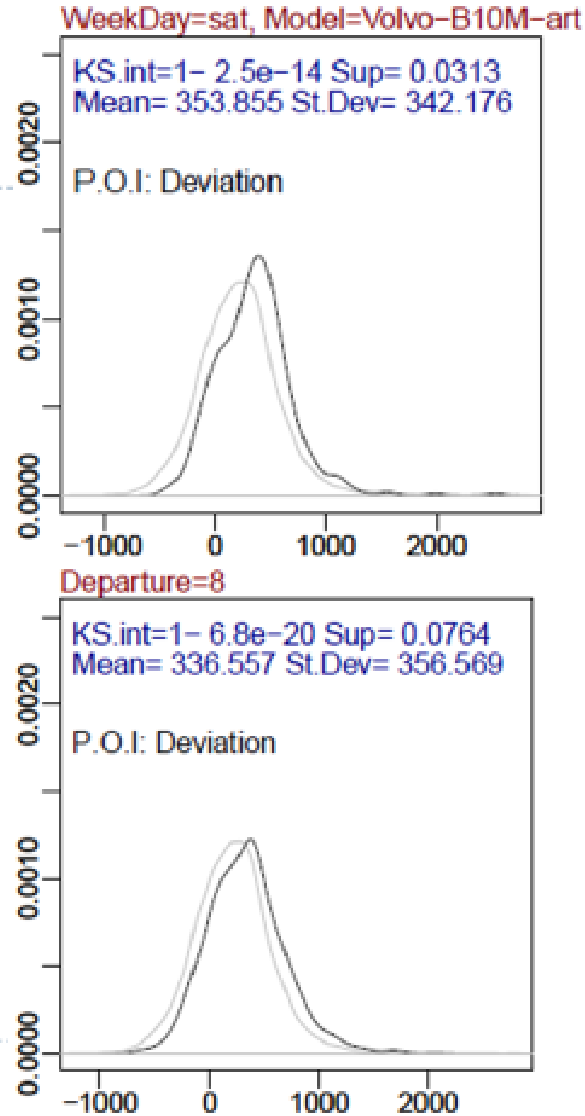



# Case Study (2):

## Delays: top rules

- An articulated model is also associated to delays on Saturdays
  - Saturday is a difficult day
- Trips at morning rush hour have less severe delays but also appear.
  - Represent 7.6% of trips

19



# Max Leverage Rules

- [Jorge&Azevedo2011]
- Rules:

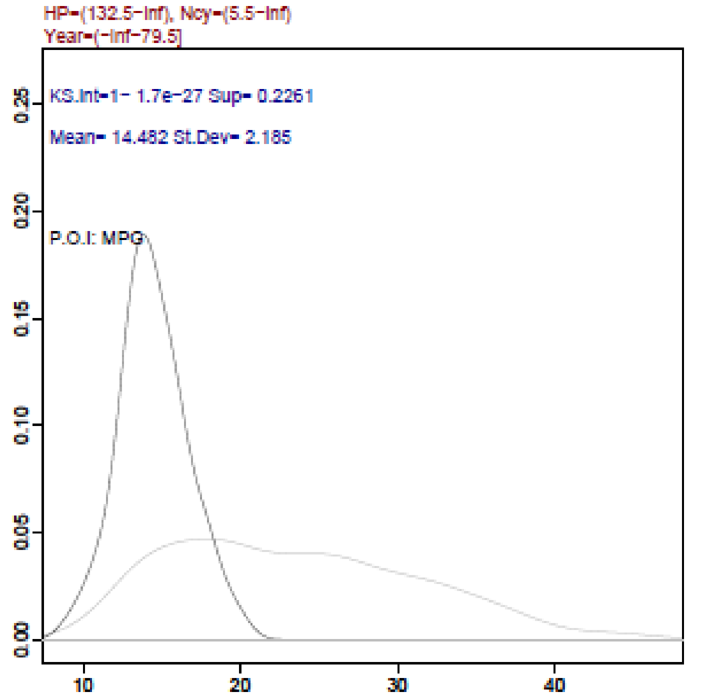
$$ant \rightarrow A \in I,$$

where $I$ is the interval that defines the maximal value of leverage (add value) of $A$ for antecedent $ant$,
and $A$ is our poi.

- Rules are derived from correspondent distribution rules.
- Intervals that maximize *leverage* / (*added value*) are obtained from the *KS* test.

- AV(A → C) = conf(A→C) – sup(C).

# Max Leverage Rules Example




(Cov=0.226 Lev=0.148 AV=0.653 Conf=0.922)

HP=(132.5-inf) & Ncy=(5.5-inf) & Year=(-inf-79.5] → MPG < 18

Rule states that cars with power output (HP) above 132.5, with more than 5 cylinders (Ncy) and assembly year (Year) before a 1980 tends to yield a performance (MPG) inferior to 18 when compared to a generic car (global population).

A generic car has much less probability of having such a bad performance. In fact, the probability of a generic car to have such bad performance is 65,3% (Added Value) lower than the probability of the car described by the rule.

# Contrast Sets Rules

- *Rules for Contrast Sets* [Azevedo2010]
- Describe the difference between contrasting groups.
- A contrast set is a conjunction of characteristics that describes a subpopulation which occurs with different proportions along different groups.
- Examples:
  - Different temporal instances (sales in 1998 versus 1999),
  - Different locations (find distinct characteristics for the location of a gene *x* in human DNA in relation to mice DNA),
  - Along different classes (difference between brunettes and blonds).

# RCS

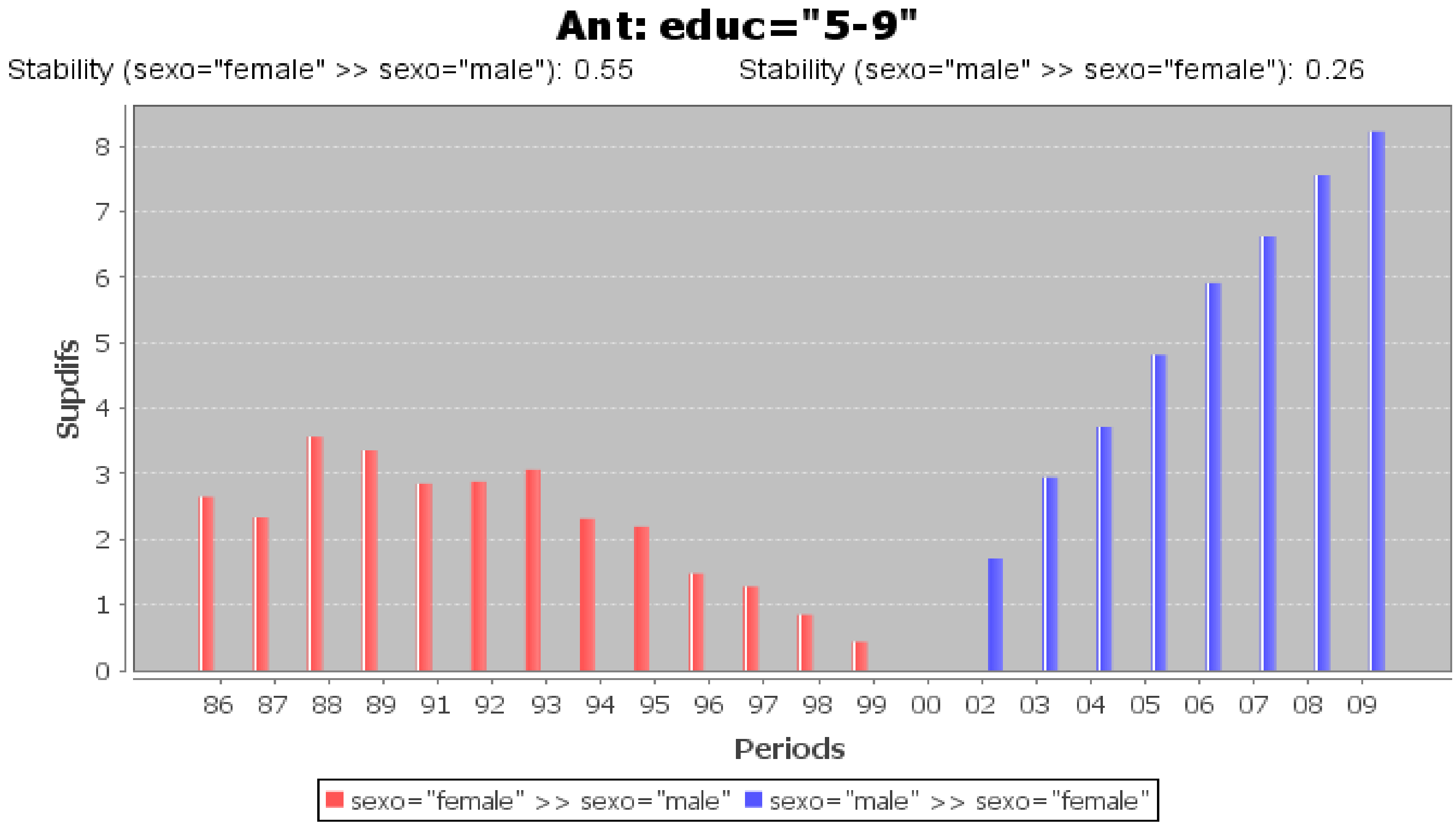
- The characteristics of the subpopulation to be found (*contrast sets*) are interesting (significant) if the proportions of the individual occurrences along groups are significantly distinct.
- i.e. subpopulation is *not independent* to group belonging. Significance is computed using a Fisher exact test.

Gsup = 0.17191 | 0.04121 p = 1.1110878451E-017 education=Doctorate >> education=Masters
Gsup = 0.17191 | 0.01681 p = 3.0718399575E-040 education=Doctorate >> education=Bachelors
Sup(CS) = 0.03097 ← workclass=State-gov & class > 50K.

- Specialization of a *contrast set* is controlled also through a Fisher test.

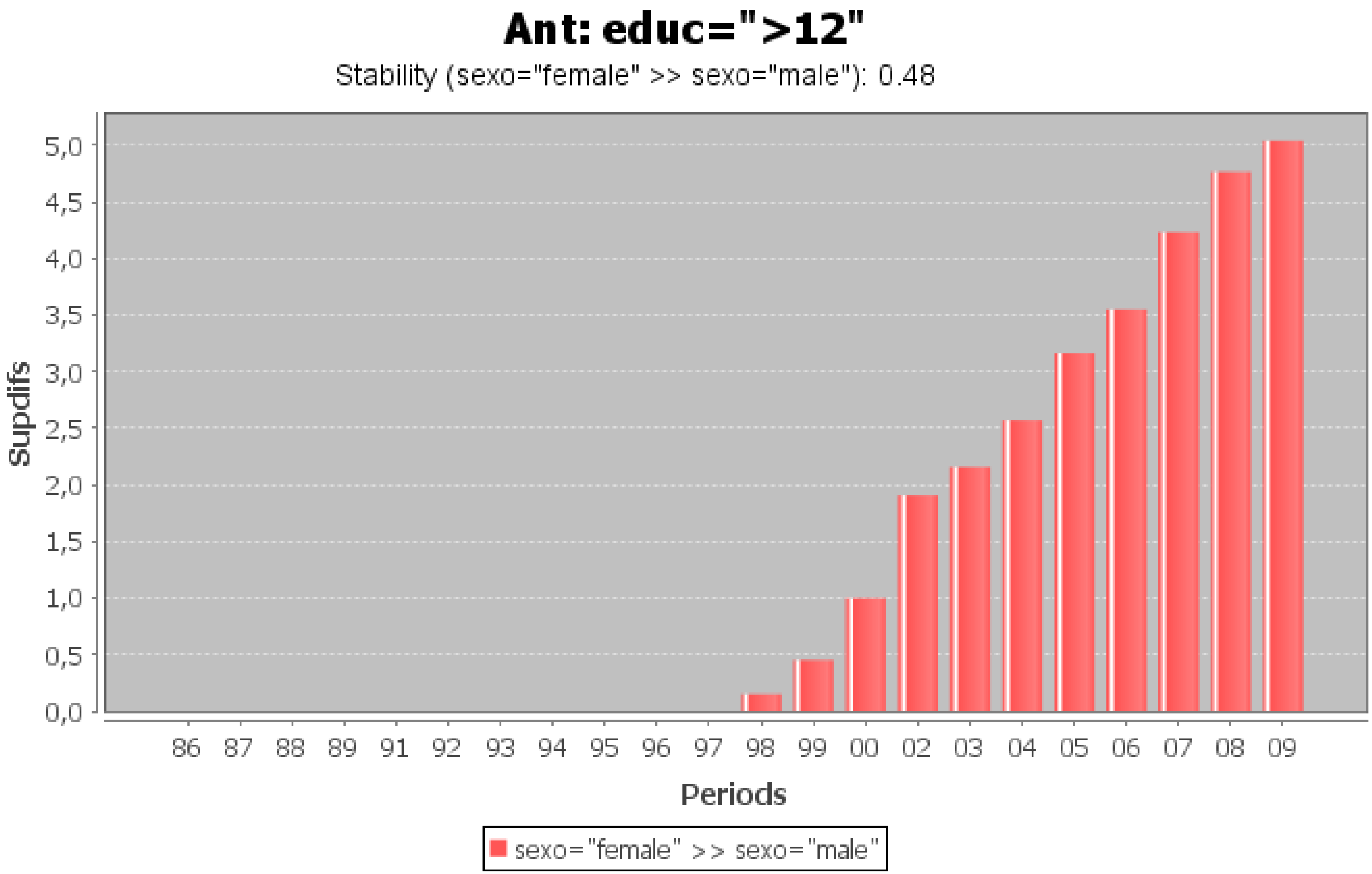
# Case Study

Data representing employment from the Portuguese private sector between 1986 and 2009.




- Contrast on individuals with basic (lower) education

# Case Study




- Contrast found on individuals with higher education

# Summary

- Introduction to Data Mining
- Pattern Mining
- Descriptive data mining
- Association Rules

- Subgroup Mining implemented using Association rules like algorithms