Pattern Mining

- Determine what items often go together (usually in transactional databases)
- Often Referred to as Market Basket Analysis
  - used in retail for planning arrangement on shelves
  - used for identifying cross-selling opportunities
  - "should" be used to determine best link structure for a Web site
- Examples
  - people who buy milk and beer also tend to buy diapers
  - people who access pages A and B are likely to place an online order
- Suitable data mining tools
  - association rule discovery
  - clustering
  - Nearest Neighbor analysis (memory-based reasoning)

Market Basket Analysis: the context

Customer buying habits by finding associations and correlations between the different items that customers place in their “shopping basket”

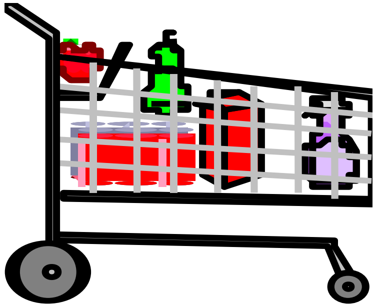
Milk, eggs, sugar, bread

Customer1

Milk, eggs, cereal, bread

Customer2

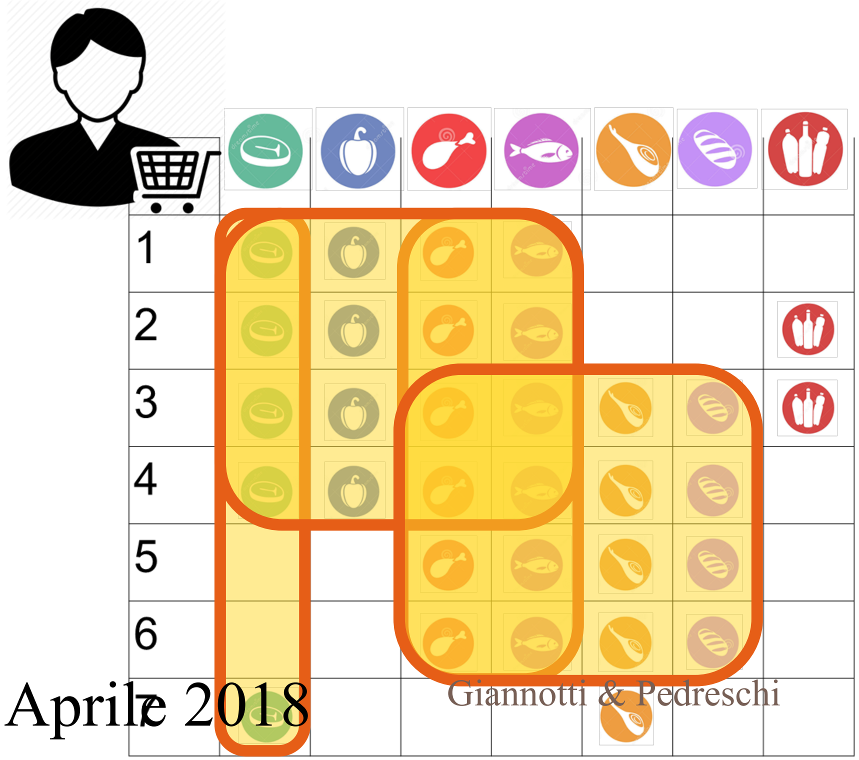
Eggs, sugar



Customer3

Frequent patterns

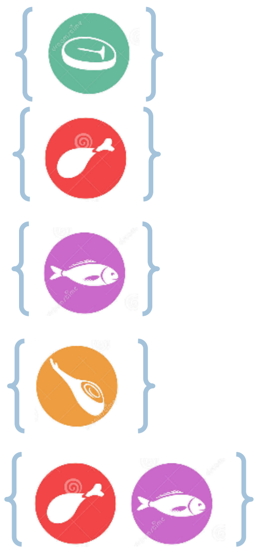
- □ Events or combinations of events that appear frequently in the data
- □ E.g. items bought by customers of a supermarket

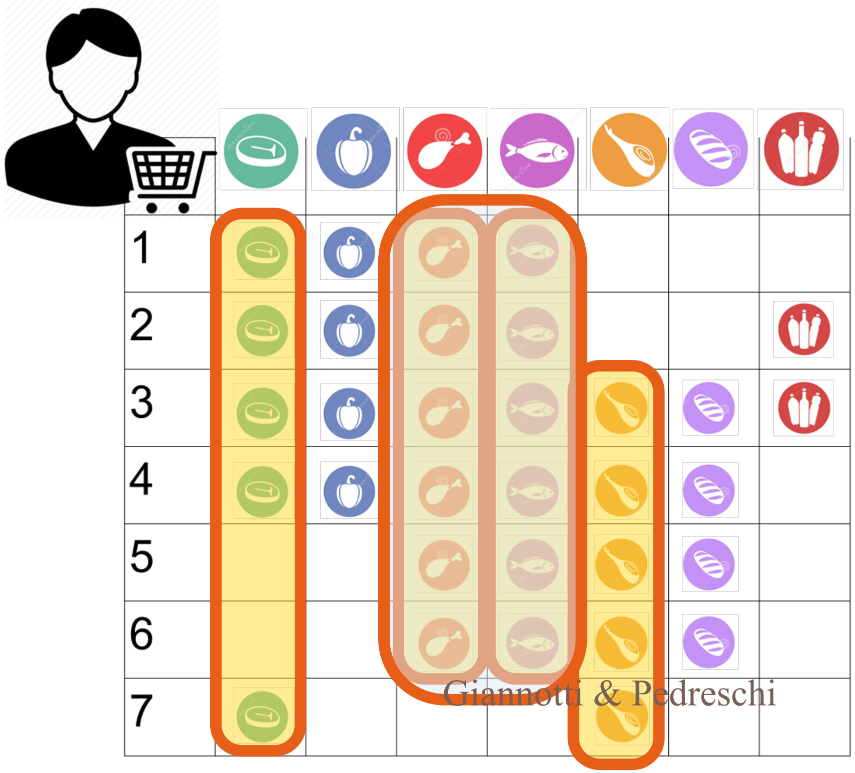


Frequent patterns

- **Frequent itemsets** w.r.t. minimum threshold

- E.g. with Min_freq = 5





Giannotti & Pedreschi

Frequent patterns

- **Association rules**
  - If items A1, A2, … appear in a basket, then also B1, B2, … will appear there
  - Notation: A1, A2, … => B1, B2, … [C%]
    - C = confidence, i.e. conditional probability

🥩 => 🫑 **[80%]**

🫑 => 🍗 **[100%]**

🐟 => 🥩 **[66%]**

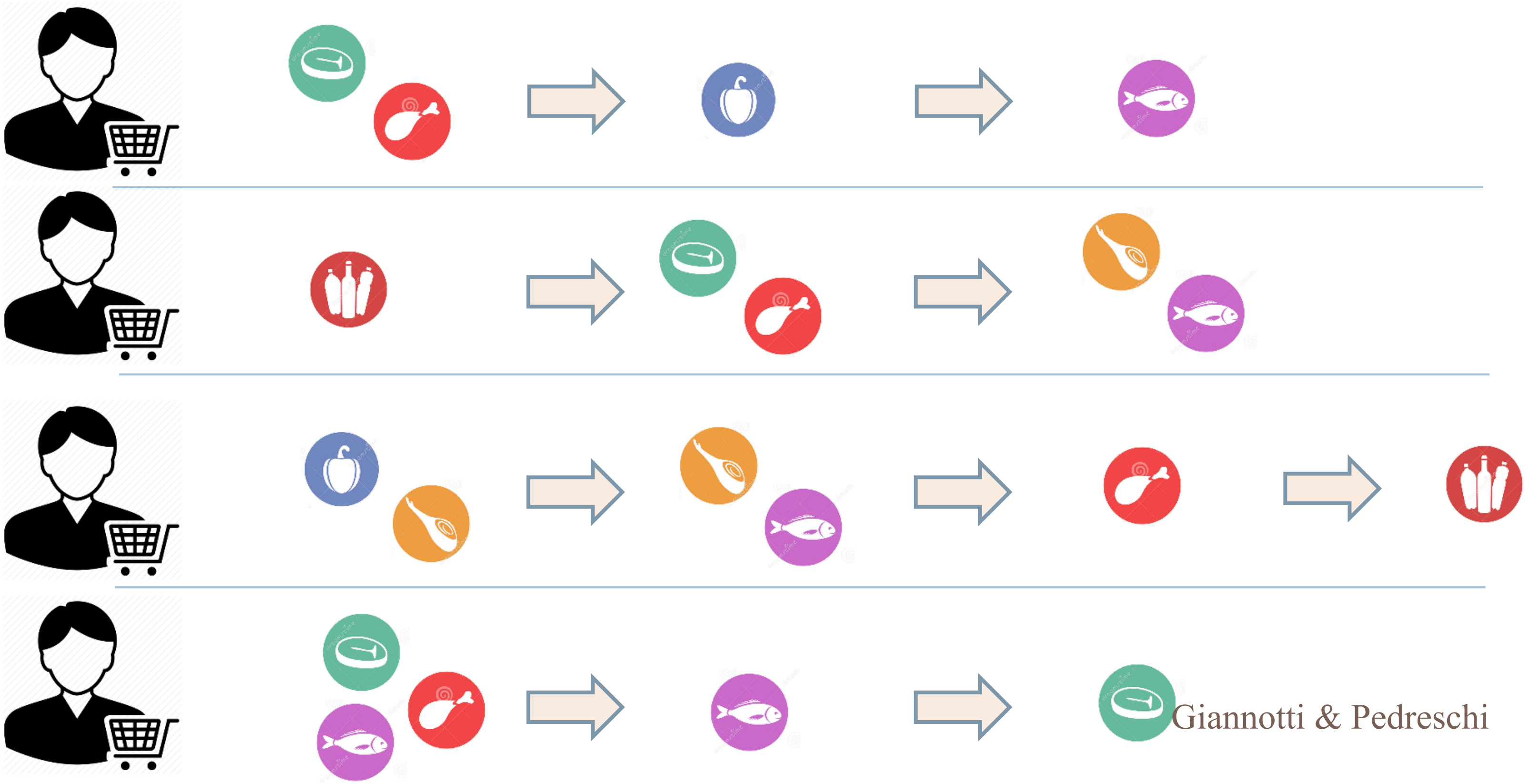
🍖 => 🍾 **[20%]**

	🥩	🫑	🍗	🐟	🍖	🥖	🍾
1	🥩	🫑	🍗	🐟			
2	🥩	🫑	🍗	🐟			🍾
3	🥩	🫑	🍗	🐟	🍖	🥖	🍾
4	🥩	🫑	🍗	🐟	🍖	🥖	
5			🍗	🐟	🍖	🥖	
6			🍗	🐟	🍖	🥖	
7	🥩				🍖		

Frequent patterns

Complex domains

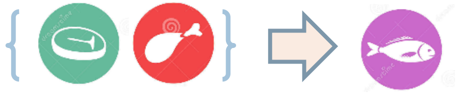
- □ Frequent sequences (a.k.a. Sequential patterns)
- □ Input: sequences of events (or of groups)

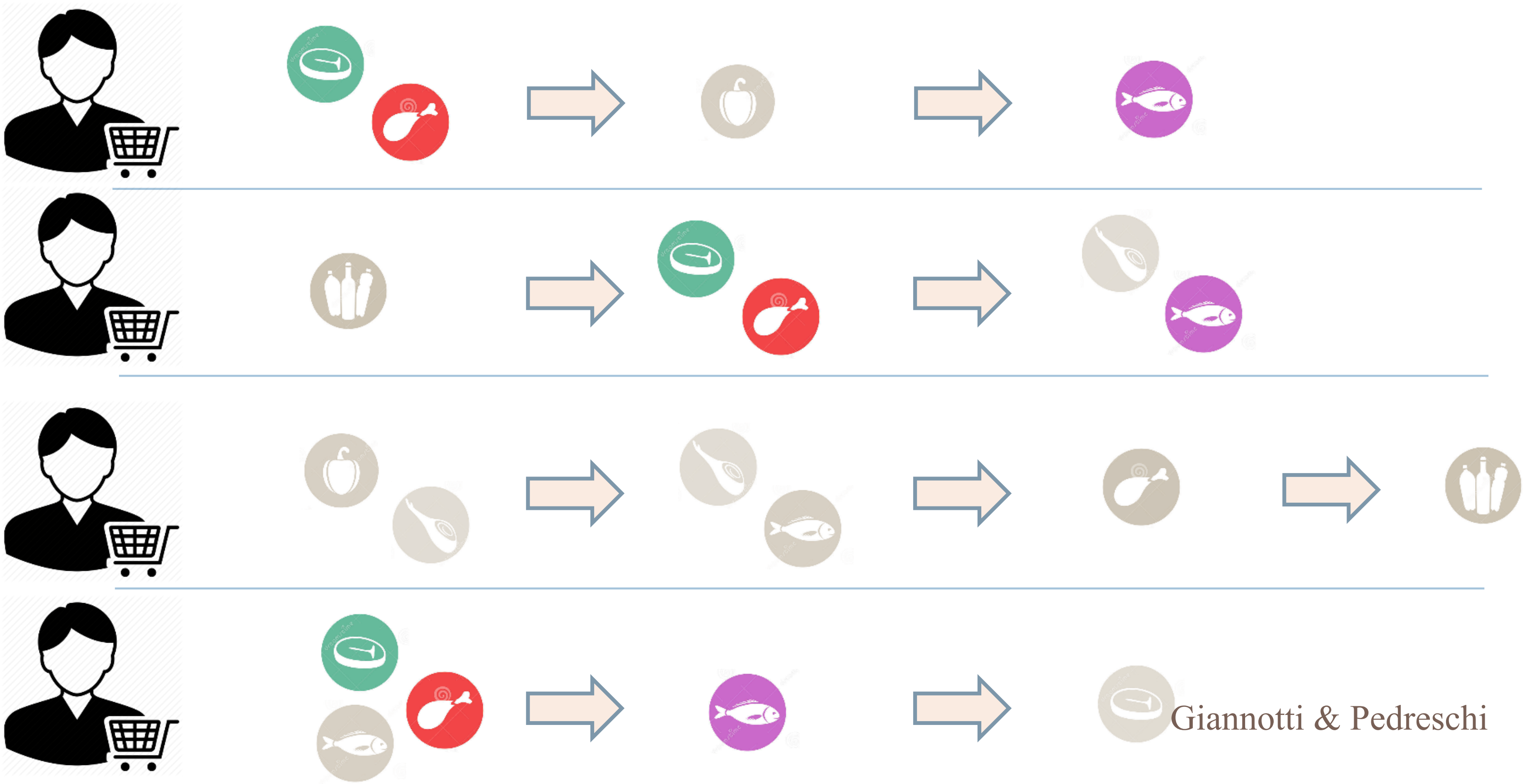


Frequent patterns

Complex domains

□ Objective: identify sequences that occur frequently

- Sequential pattern: 



Transaction data: supermarket data

- Market basket transactions:

  t1: {bread, cheese, milk}

  t2: {apple, eggs, salt, yogurt}

  … …

  tn: {biscuit, eggs, milk}

- Concepts:
  - An *item*: an item/article in a basket
  - *I*: the set of all items sold in the store
  - A *transaction*: items purchased in a basket; it may have TID (transaction ID)
  - A *transactional dataset*: A set of transactions

Transaction data: a set of documents

- □ **A text document data set. Each document is treated as a "bag" of keywords**

doc1:	Student, Teach, School
doc2:	Student, School
doc3:	Teach, School, City, Game
doc4:	Baseball, Basketball
doc5:	Basketball, Player, Spectator
doc6:	Baseball, Coach, Game, Team
doc7:	Basketball, Team, City, Game

The model: rules

- A transaction *t* contains X, a set of items (itemset) in *I*, if $X \subseteq t$.
- An association rule is an implication of the form:

  $X \rightarrow Y$, where $X, Y \subset I$, *and* $X \cap Y = \emptyset$

- An itemset is a set of items.
  - E.g., X = {milk, bread, cereal} is an itemset.
- A *k*-itemset is an itemset with *k* items.
  - E.g., {milk, bread, cereal} is a 3-itemset

Association Rules: measures

$$X \Rightarrow Y\ [\, s, c\,]$$

Support: denotes the frequency of the rule within transactions. **A high value means** that the rule involve a great part of database. **(HOW POPULAR IS THE GROUP)**

$$\text{support}(X \Rightarrow Y) = \Pr(X \cup Y)$$

Confidence: denotes the percentage of transactions containing X which contain also Y. It is an estimation of conditioned probability . **(how likely is Y given X)**

$$\text{Confidence}(X \Rightarrow Y) = \Pr(Y|X) = \Pr(X \,\&\, Y)/\Pr(X).$$

Rule strength measures

- Support: The rule holds with support *sup* in *T* (the transaction data set) if *sup*% of transactions contain $X \cup Y$.
  - *sup* = $\Pr(X \cup Y)$.
- Confidence: The rule holds in *T* with confidence *conf* if *conf*% of transactions that contain *X* also contain *Y*.
  - *conf* = $\Pr(Y \mid X)$
- An association rule is a pattern that states when *X* occurs, *Y* occurs as well with a certain probability.

Support and Confidence

□ Support count: The support count of an itemset *X*, denoted by *X.count*, in a data set *T* is the number of transactions in *T* that contain X. Assume *T* has *n* transactions.

□ Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Valid rules

- **Valid rules:** all rules that satisfy the user-specified *minimum support* (minsup) and *minimum confidence* (minconf).

- **Key Features**
  - Completeness: find all rules.
  - No target item(s) on the right-hand-side

An example

t1:	Beef, Chicken, Milk
t2:	Beef, Cheese
t3:	Cheese, Boots
t4:	Beef, Chicken, Cheese
t5:	Beef, Chicken, Clothes, Cheese, Milk
t6:	Chicken, Clothes, Milk
t7:	Chicken, Milk, Clothes

- Transaction data
- Assume:

  minsup = 30%
  minconf = 80%

- An example frequent *itemset*:

  {Chicken, Clothes, Milk} [sup = 3/7]

- Association rules from the itemset:

  Clothes → Milk, Chicken [sup = 3/7, conf = 3/3]

  … …

  Clothes, Chicken → Milk, [sup = 3/7, conf = 3/3]

Association Rules: measures Meaning

$$X \Rightarrow Y\ [\,s, c\,]$$

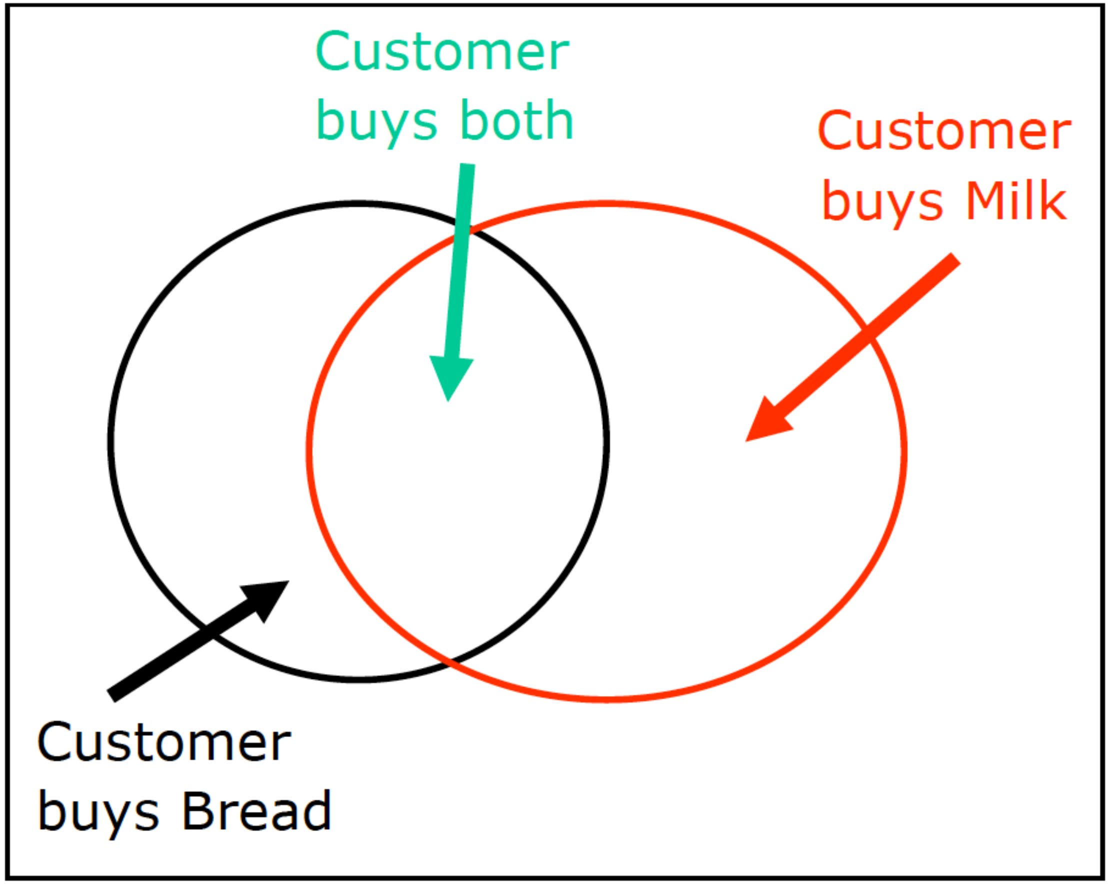
Support: denotes the frequency of the rule within transactions. **A high value means** that the rule involve a great part of database. **(HOW POPULAR IS THE GROUP)**

$$\text{support}(X \Rightarrow Y) = \Pr(X \,\&\, Y)$$

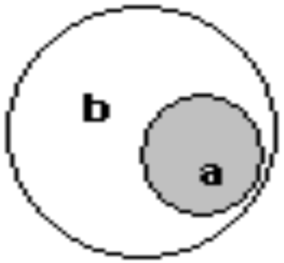
Confidence: denotes the percentage of transactions containing X which contain also Y. It is an estimation of conditioned probability . **(how likely is B given A)**

$$\text{Confidence}(X \Rightarrow Y) = \Pr(Y|X) = \Pr(X \,\&\, Y)/\Pr(X).$$

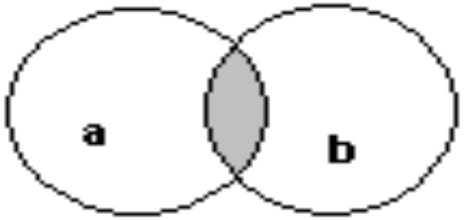
Support and Confidence



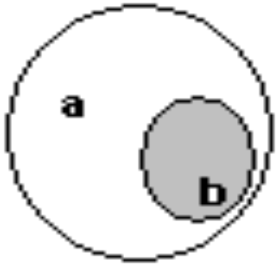
Association Rules – the effect



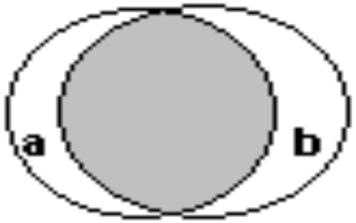
conf(a => b) = 100%
conf(b => a) = ~ 0%



conf(a => b) = ~ 0%
conf(b => a) = ~ 0%



conf(a => b) = ~ 0%
conf(b => a) = 100%



conf(a => b) = ~100%
conf(b => a) = ~100%

Association Rules – the parameters σ and γ

Minimum Support σ :

High ⇒ few frequent itemsets

⇒ few valid rules which occur very often

Low ⇒ many valid rules which occur rarely

Minimum Confidence γ :

High ⇒ few rules, but all "almost logically true"

Low ⇒ many rules, but many of them very "uncertain"

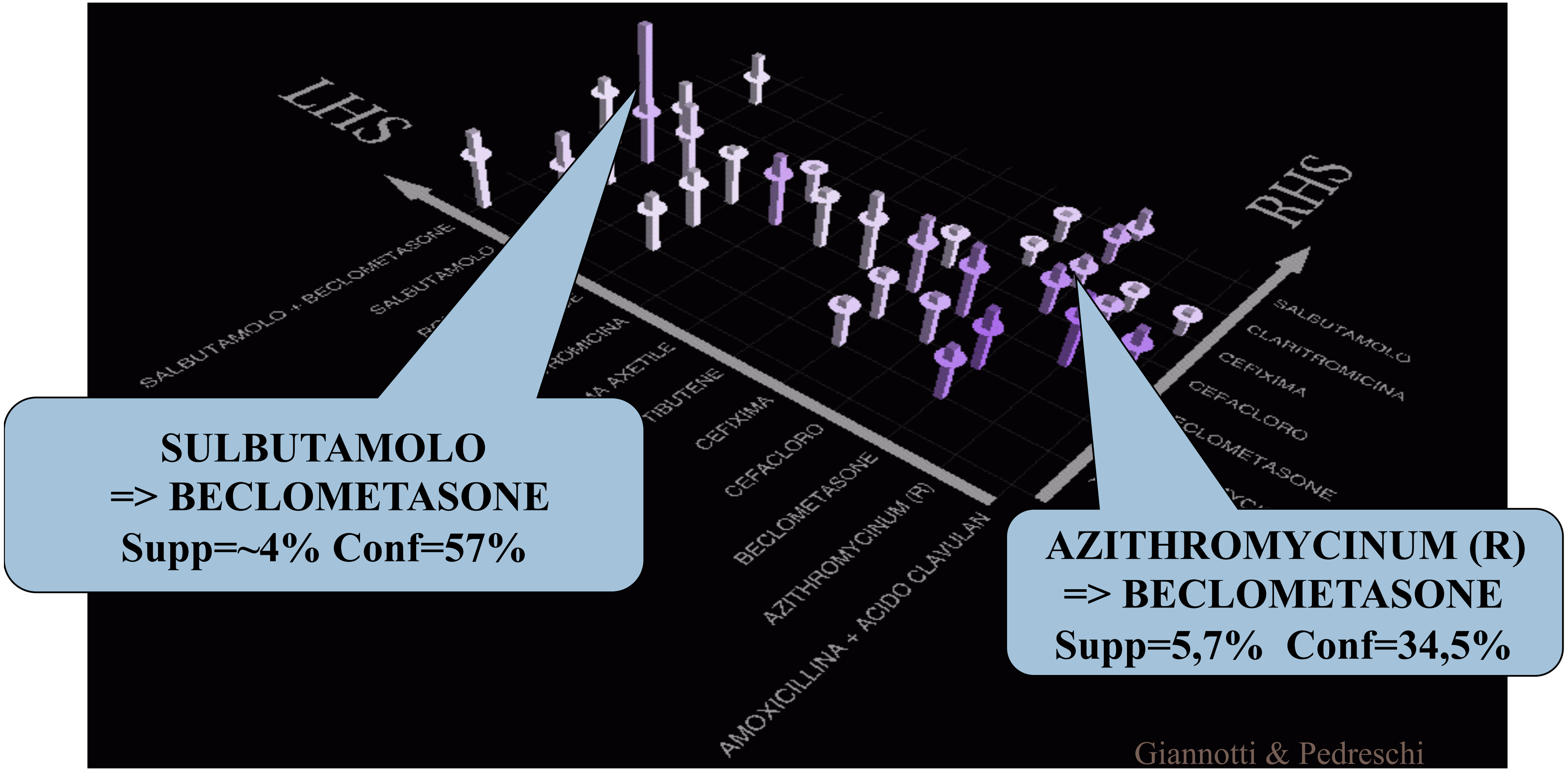
Typical Values: σ = 2 ÷10 % γ = 70 ÷90 %

Other interest measures

- Problem: confidence does not take into account the popularity of the consequent.
- INTERESTﾠ= Pr(*X* & *Y*)/P(X)*P(Y)
  - How likely is Y given X, while controlling the popularity of Y
- **Interest** expresses measure of correlation
  - = 1 ⇒ X and Y are independent events (**the rule does not make sense)**
  - less than 1 ⇒ X and Y negatively correlated,
  - greater than 1 ⇒ X and Y positively correlated
- Other measures
  - *Val* = Pr(*Y* | *X*) – Pr(Y) = Confidence – Pr(Y)
  - LIFT = Pr(*Y* | *X*) / Pr(Y) = Confidence / Pr(Y)

Association Rules – visualization

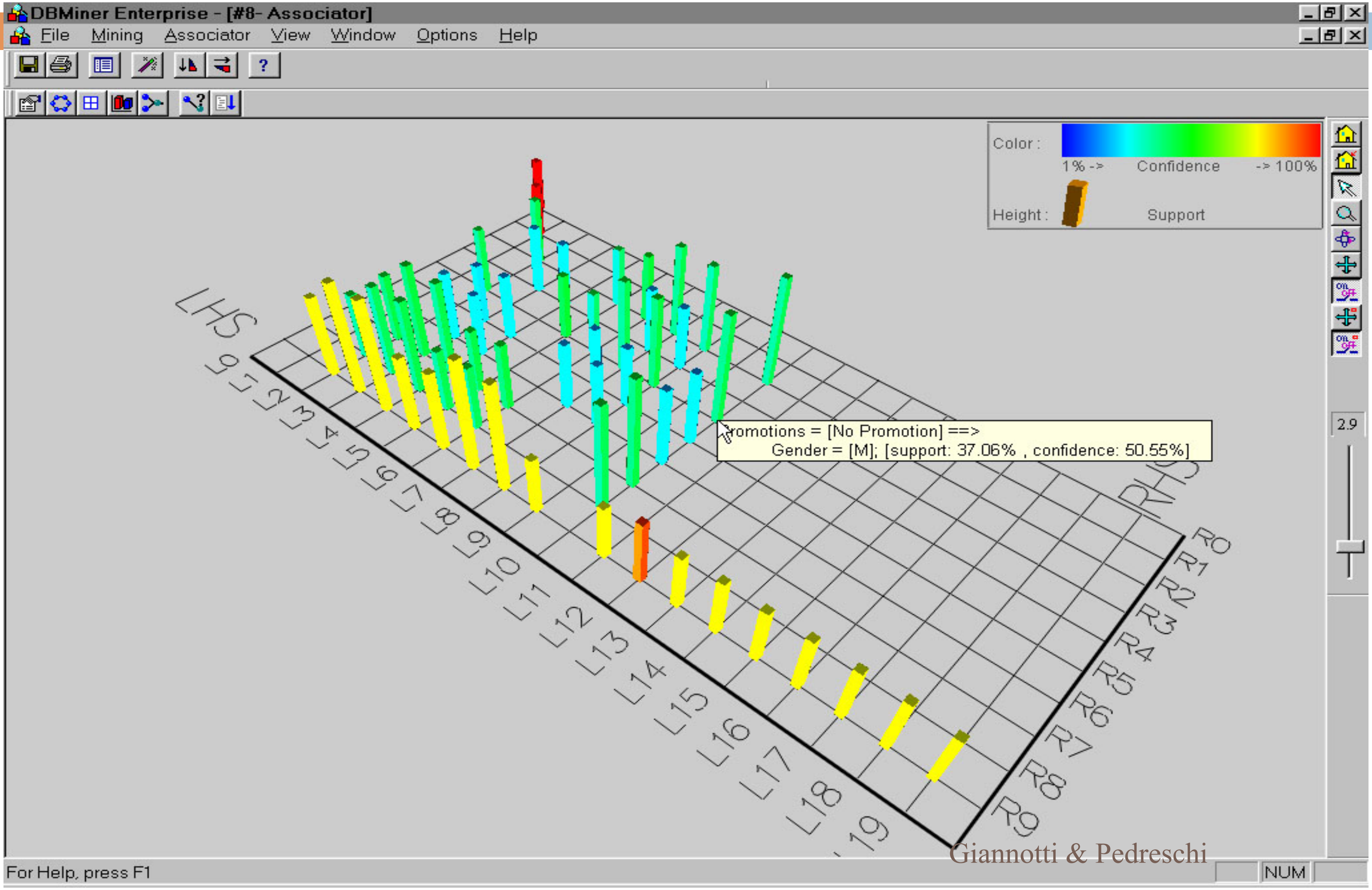
(Patients <15 old for USL 19 (a unit of Sanitary service), January-September 1997)



SULBUTAMOLO
=> BECLOMETASONE
Supp=~4% Conf=57%

AZITHROMYCINUM (R)
=> BECLOMETASONE
Supp=5,7% Conf=34,5%

Visualization of Association Rules: Plane Graph



DBMiner Enterprise - [#8- Associator]
File Mining Associator View Window Options Help
Color :
1% -> Confidence -> 100%
Height : Support
LHS
L0 L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13 L14 L15 L16 L17 L18 L19
RHS
R0 R1 R2 R3 R4 R5 R6 R7 R8 R9
2.9
For Help, press F1
NUM

romotions = [No Promotion] ==>
Gender = [M]; [support: 37.06% , confidence: 50.55%]