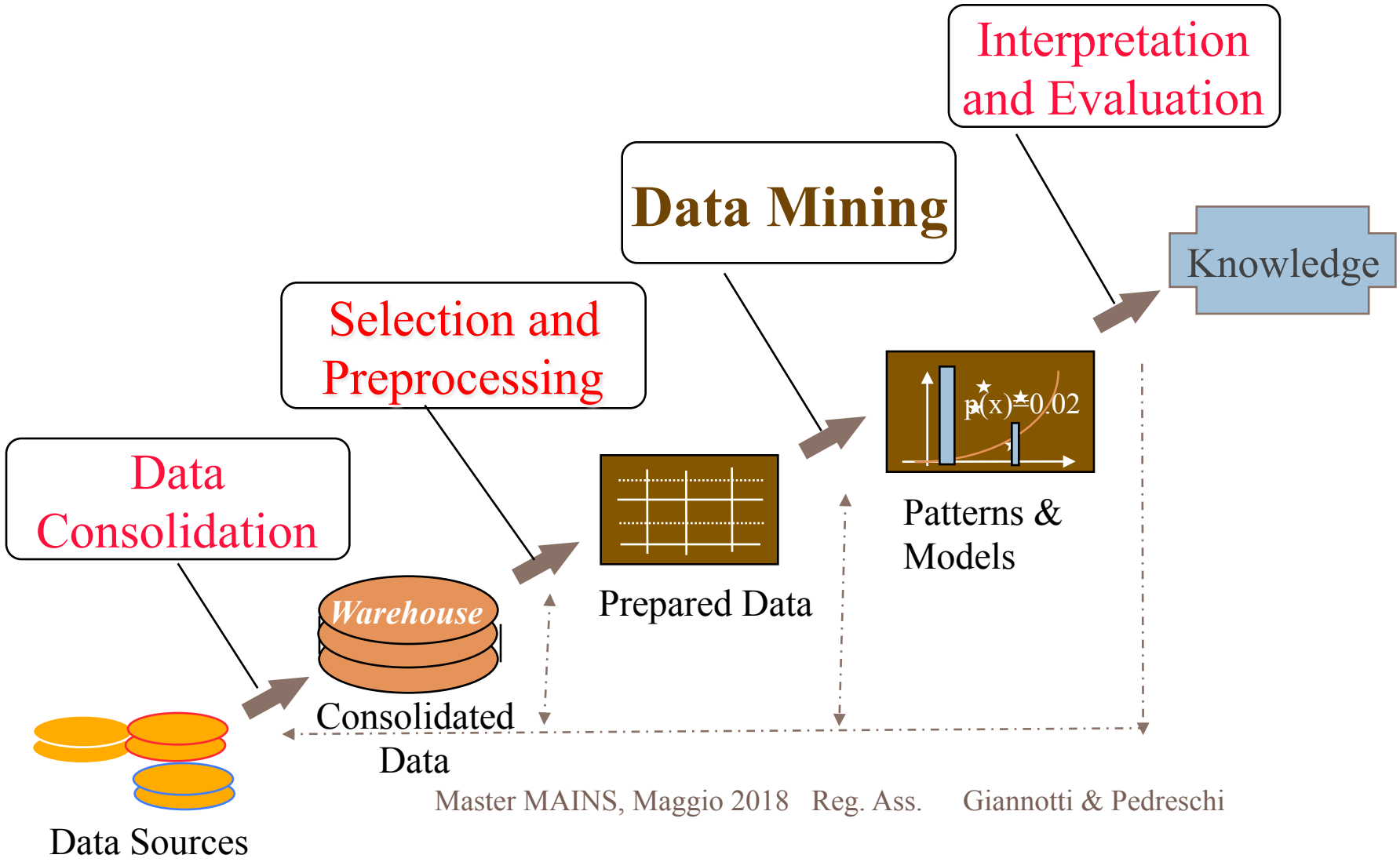
# MACHINE LEARNING & DATA MINING

DINO PEDRESCHI, FOSCA GIANNOTTI
PISA KDD LAB, ISTI-CNR & UNIV. PISA

# KDD Process




Interpretation and Evaluation
Data Mining
Knowledge
Selection and Preprocessing
Data Consolidation
p(x)=0.02
Patterns & Models
Prepared Data
Warehouse
Consolidated Data
Data Sources

# ASSOCIATION RULES AND MARKET BASKET ANALYSIS

# Riferimenti bibliografici

- Berthold et. al. Guide to Intelligent Data Analysis, Chapter 7.6

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to DATA MINING, Addison Wesley, ISBN 0-321-32136-7, 2006, Chapter 6

- Provost, F., Fawcett, T. Data Science for Business (2012) (see co-occurrence grouping)

# Pattern Mining & Association rules

- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - Basic concepts
- How to compute AR
  - Basic Apriori Algorithm
- How to reason on AR and how to evaluate their quality
  - Multi-Dimension AR (inter-attribute)
  - Interestingness
  - Correlation vs. Association
- Example: Profiling with patterns

# 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"

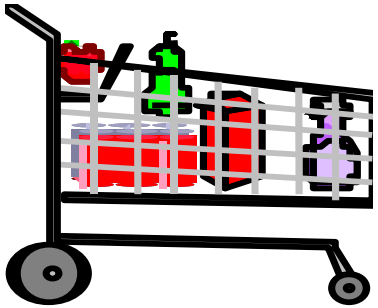
Milk, eggs, sugar, bread

Customer1

Milk, eggs, cereal, bread

Customer2

Eggs, sugar

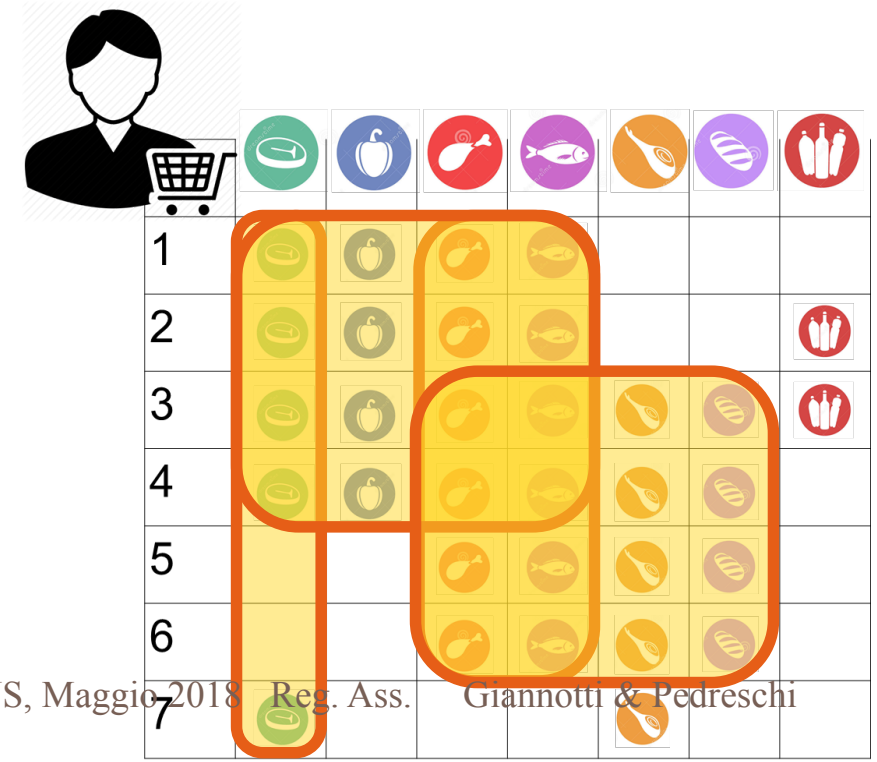


Customer3

# Frequent patterns



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

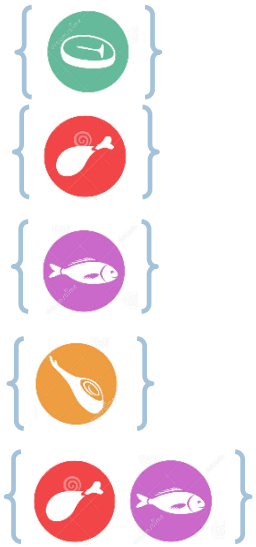


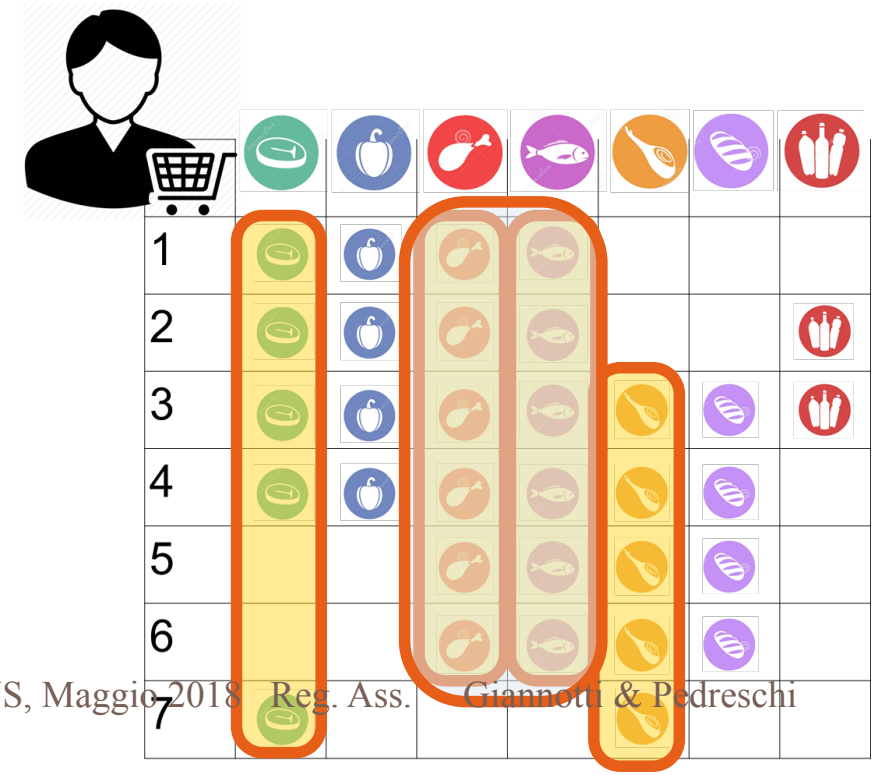
# Frequent patterns



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

- E.g. with Min_freq = 5

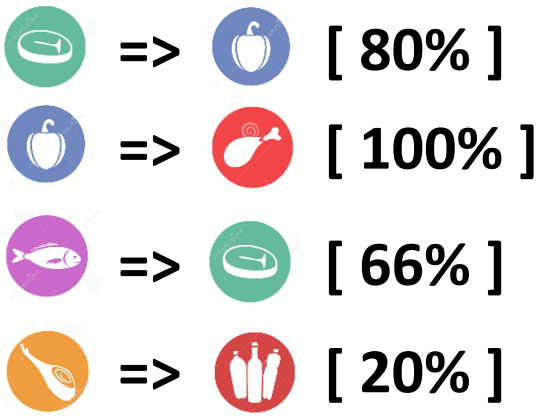


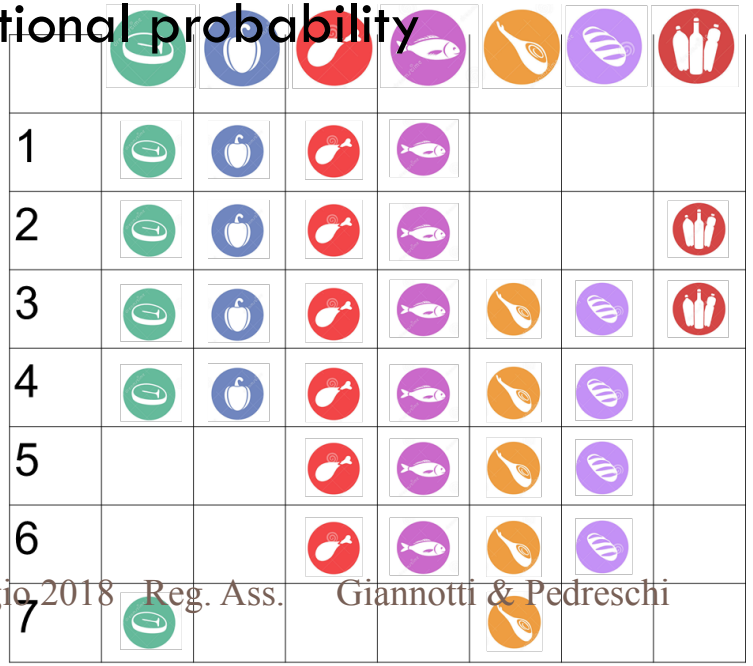


# Frequent patterns



- **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



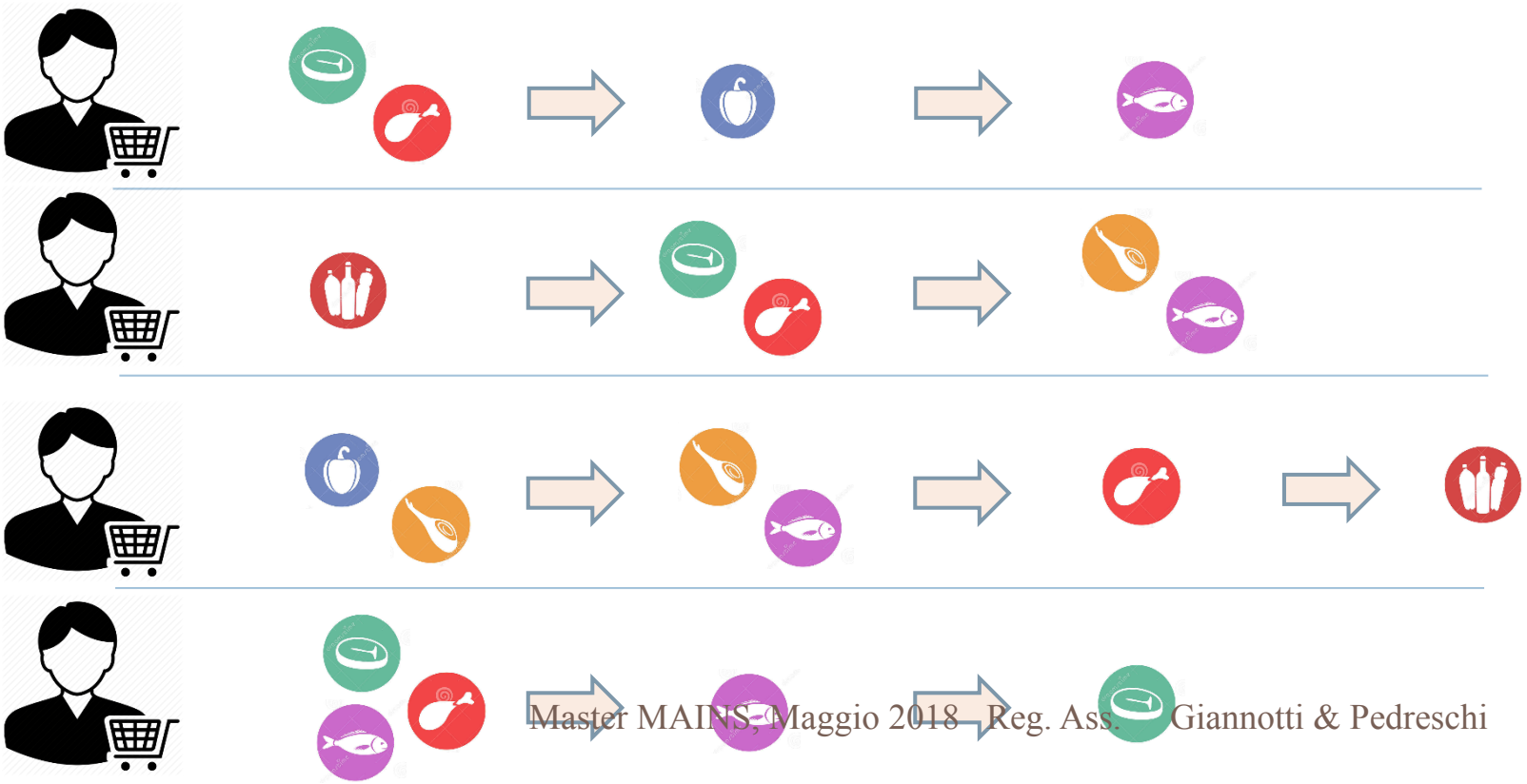



# Frequent patterns

## Complex domains



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

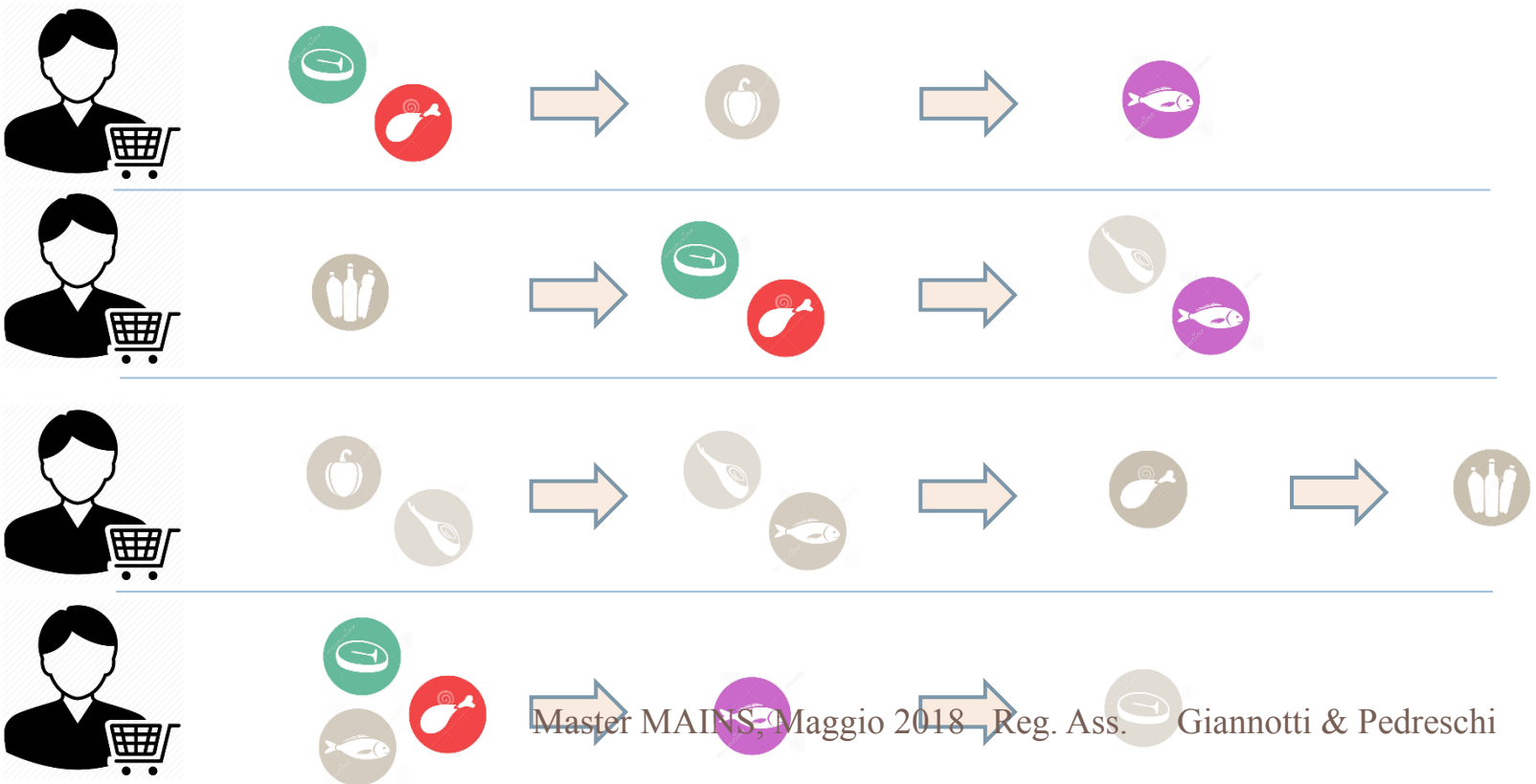


# Frequent patterns

## Complex domains



- Objective: identify sequences that occur frequently
- Sequential pattern: 



# 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)**

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

**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)**

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

# 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* $X \cap Y = \varnothing$

- 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

# 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 with 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)**

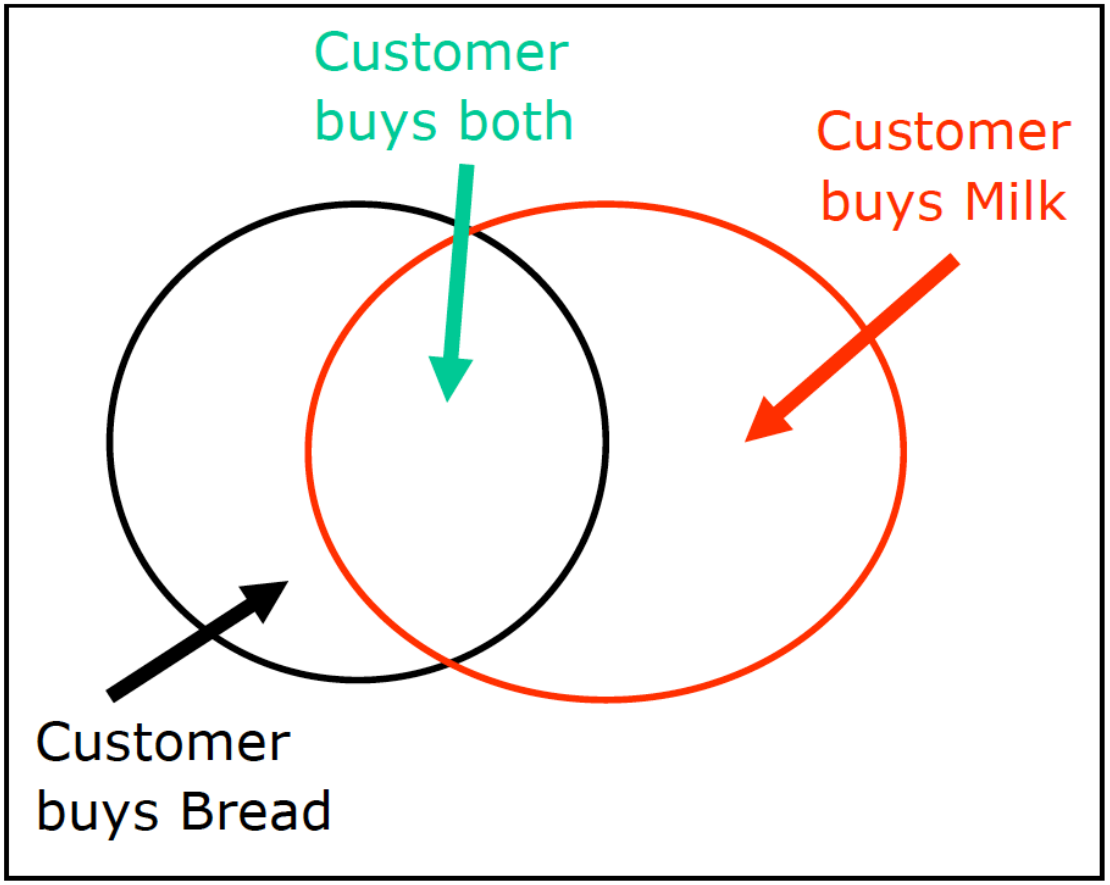
$$\text{support}(X \Rightarrow Y) = \Pr(X \,\&\, Y)$$

**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)**

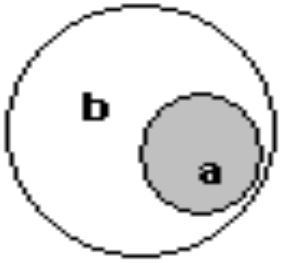
$$\text{Confidence}(X \Rightarrow Y) = \Pr(Y|X) = \Pr(X \,\&\, Y)/\Pr(X).$$

# Support and Confidence

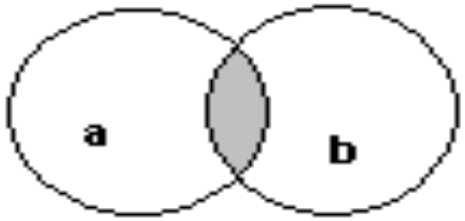




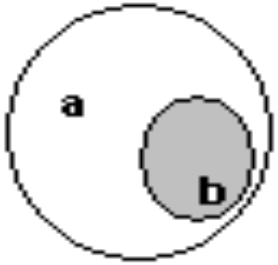
# Association Rules – the effect



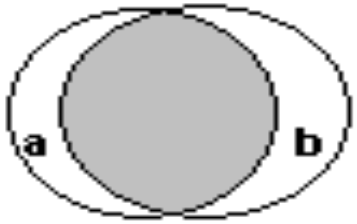

conf( a => b ) = 100%
conf( b => a ) = ~ 0%




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 %

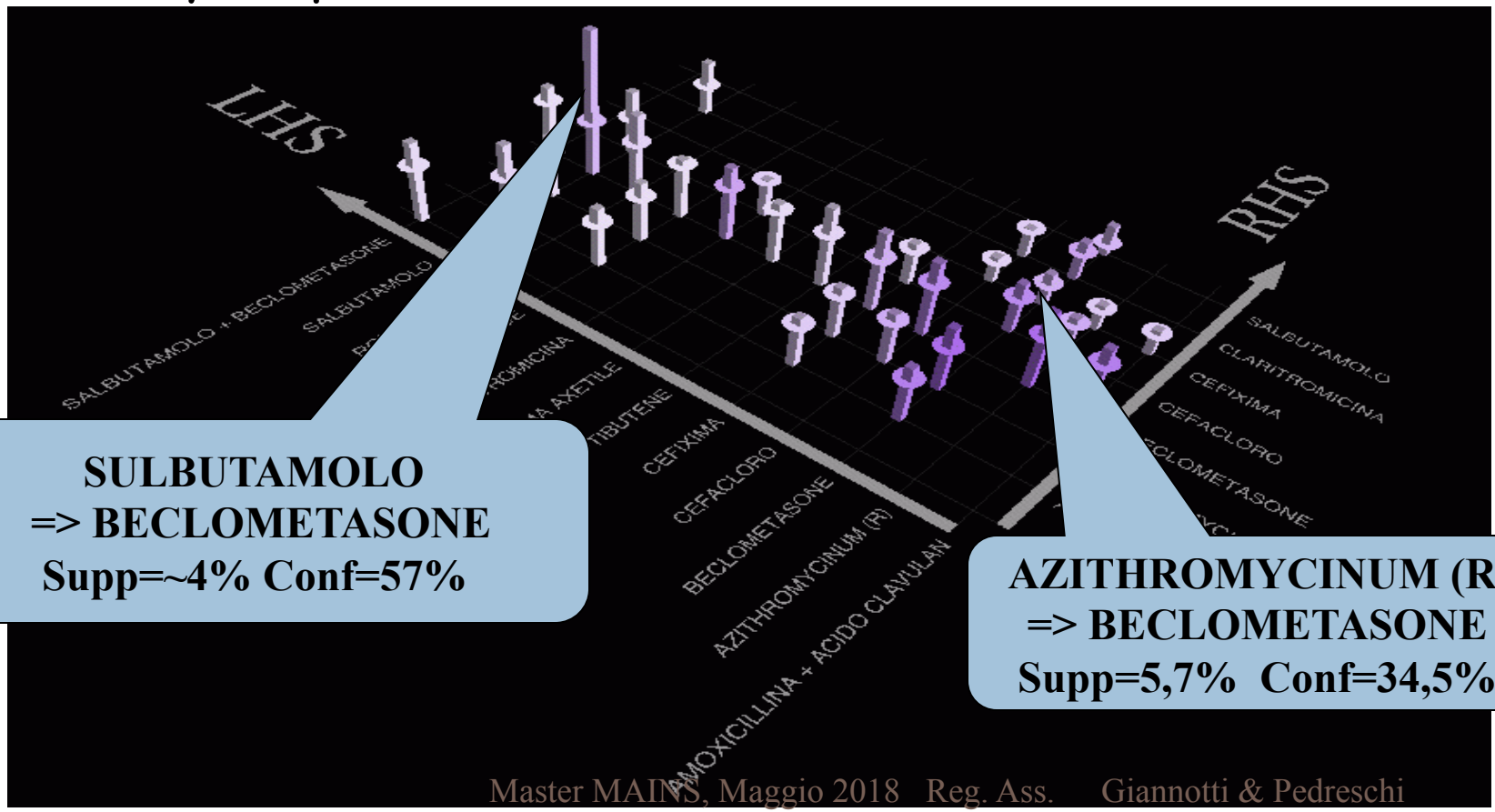
# Other **interest** measures

- Problem: confidence does not take into account the popularity of the consequent.
- INTEREST= 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)

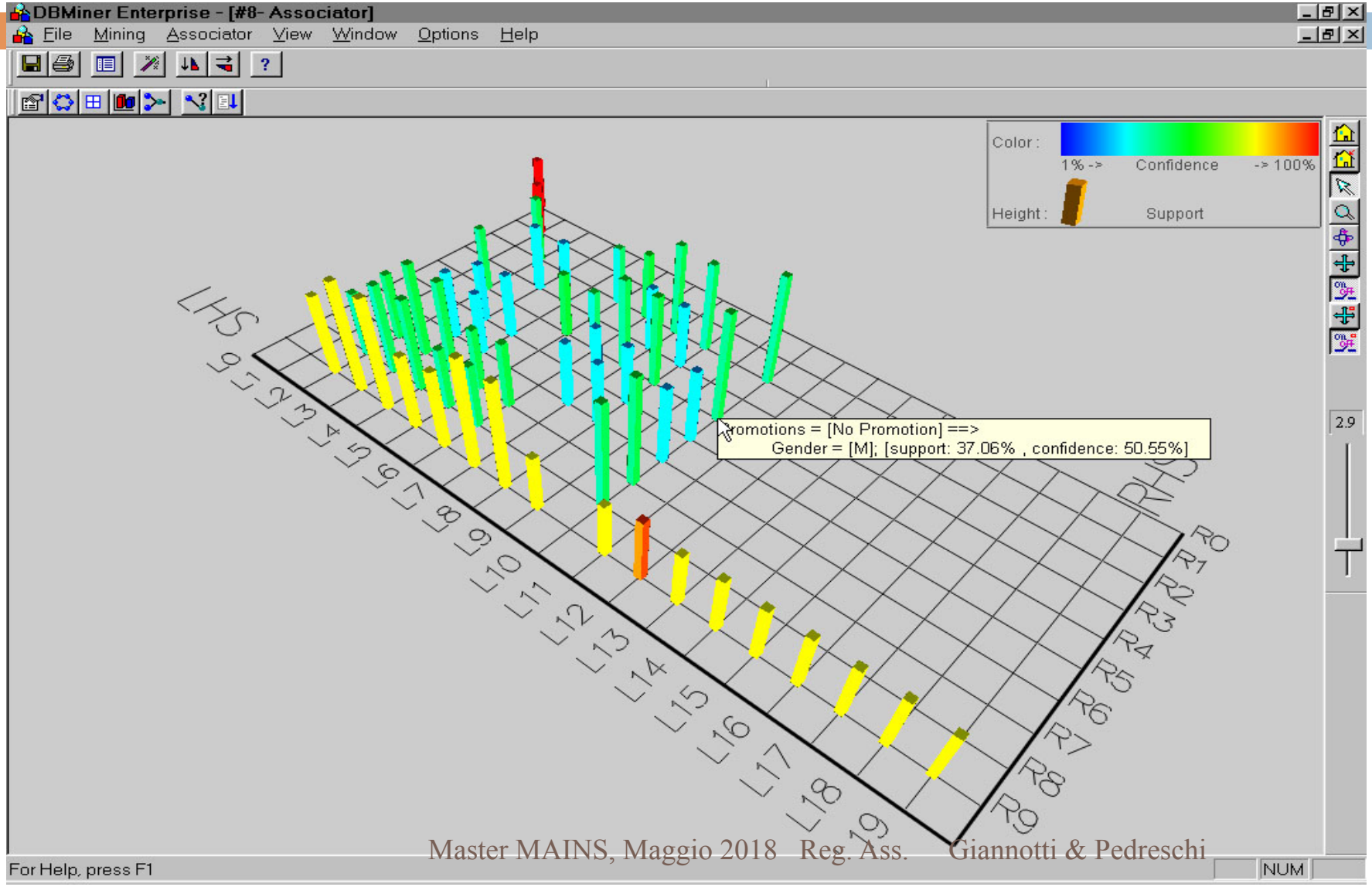



**SULBUTAMOLO => BECLOMETASONE Supp=~4% Conf=57%**

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

# Visualization of Association Rules: Plane Graph





# Goal of MBA



- Extract information on purchasing behavior
- Actionable information: can suggest
  - new store layouts
  - new product assortments
  - which products to put on promotion
- MBA applicable whenever a customer purchases multiple things in proximity
  - credit cards
  - services of telecommunication companies
  - banking services
  - medical treatments

# MBA: applicable to many other contexts



**Telecommunication:**

Each customer is a transaction containing the set of customer's phone calls

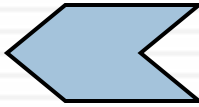
**Atmospheric phenomena:**

Each time interval (e.g. a day) is a transaction containing the set of observed event (rains, wind, etc.)

**Etc.**

# Association rules - module outline

- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - AR definitions )
- How to compute AR
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
- How to reason on AR and how to evaluate their quality
  - Multiple-level AR
  - Interestingness
  - Correlation vs. Association
- Sequential Patterns
  - Example: Profiling with patterns



# Mining Association Rules



- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

**Market-Basket transactions**

| *TID* | *Items* |
|---|---|
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

**Example of Association Rules**

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

# Mining Association Rules



- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support ≥ minsup
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

- Frequent itemset generation is still computationally expensive

# Table (6.1)

**Table 6.1.** Example of market basket transactions.

| Customer ID | Transaction ID | Items Bought |
|---|---|---|
| 1 | 0001 | $\{a, d, e\}$ |
| 1 | 0024 | $\{a, b, c, e\}$ |
| 2 | 0012 | $\{a, b, d, e\}$ |
| 2 | 0031 | $\{a, c, d, e\}$ |
| 3 | 0015 | $\{b, c, e\}$ |
| 3 | 0022 | $\{b, d, e\}$ |
| 4 | 0029 | $\{c, d\}$ |
| 4 | 0040 | $\{a, b, c\}$ |
| 5 | 0033 | $\{a, d, e\}$ |
| 5 | 0038 | $\{a, b, e\}$ |

Support?: e, (b,d), (b,d,e)

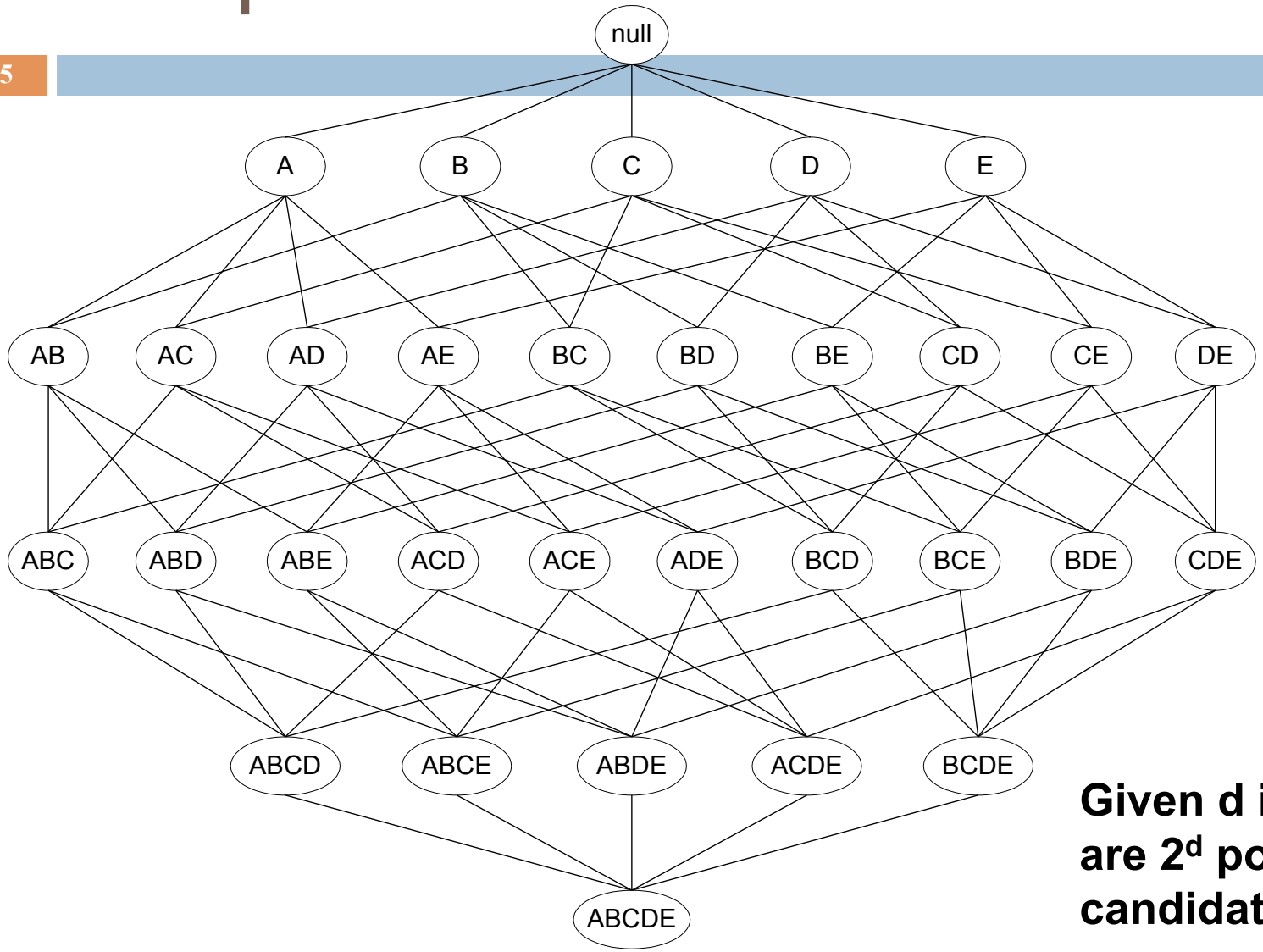
Table 6.2. Market basket transactions.

| Transaction ID | Items Bought |
|---|---|
| 1 | {Milk, Beer, Diapers} |
| 2 | {Bread, Butter, Milk} |
| 3 | {Milk, Diapers, Cookies} |
| 4 | {Bread, Butter, Cookies} |
| 5 | {Beer, Cookies, Diapers} |
| 6 | {Milk, Diapers, Bread, Butter} |
| 7 | {Bread, Butter, Diapers} |
| 8 | {Beer, Diapers} |
| 9 | {Milk, Diapers, Bread, Butter} |
| 10 | {Beer, Cookies} |

Max size of itemset, 2-itemsets with larger support

# Frequent Itemset Generation






**Given d items, there are $2^d$ possible candidate itemsets**

# The Apriori property



- **If B is frequent and A ⊆ B then A is also frequent**
  - Each transaction which contains B contains also A, which implies supp. (A) ≥ supp.(B))
- **Consequence:** if A is not frequent, then it is not necessary to generate the itemsets which include A.
- **Example:**
  - <1, {a, b}> <2, {a} >
  - <3, {a, b, c}> <4, {a, b, d}>

  with minimum support = 50%.

  The itemset {c} is not frequent so is not necessary to check for:

  {c, a}, {c, b} {c, d} {c, a, b}, {c, a, d} {c, b, d}

# Basic Apriori Algorithm



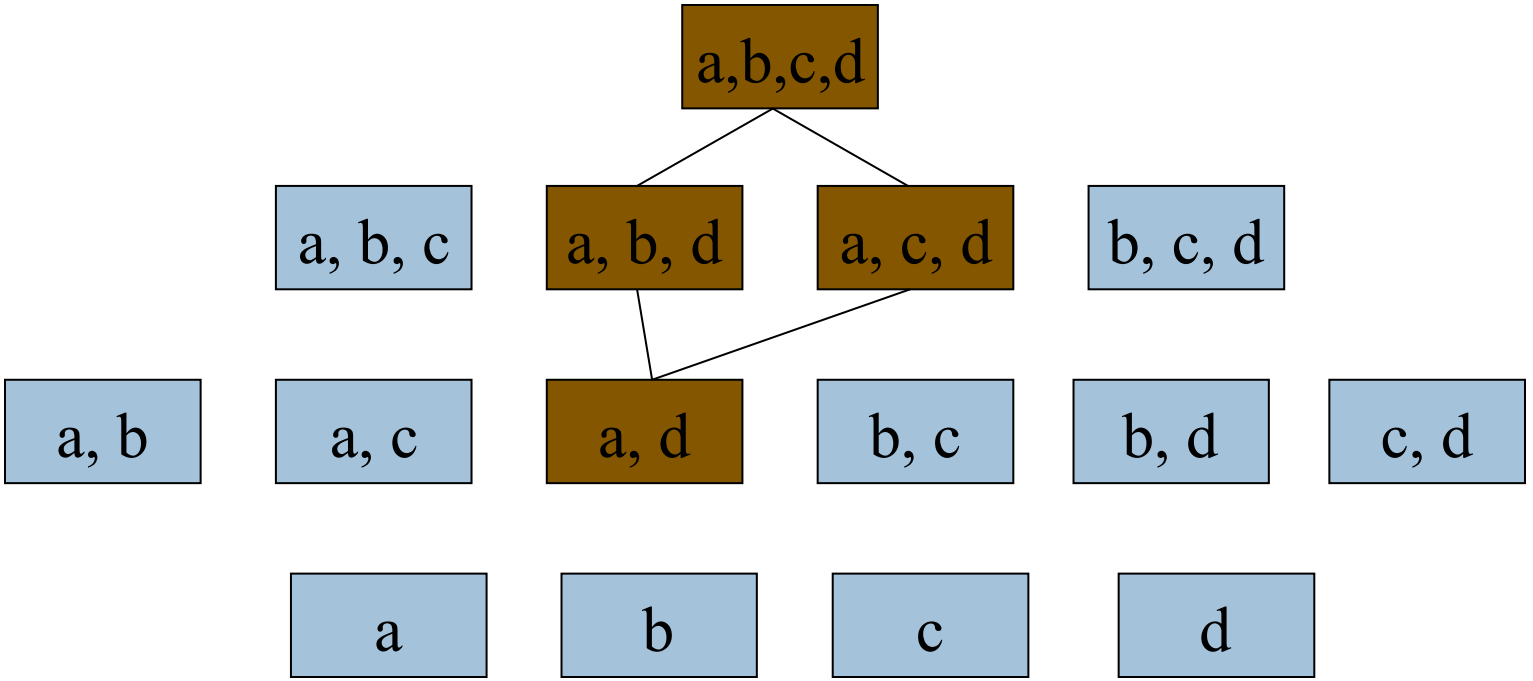
① Find the *frequent itemsets*: the sets of items that satisfy the support constraint

- *A subset of a frequent itemset is also a frequent itemset*, i.e., if {*A*,*B*} is a frequent itemset, both {*A*} and {*B*} should be a frequent itemset
- Iteratively find frequent itemsets with cardinality from 1 to *k* (*k*-itemset)

② Use the frequent itemsets to generate association rules.

# Apriori - Example

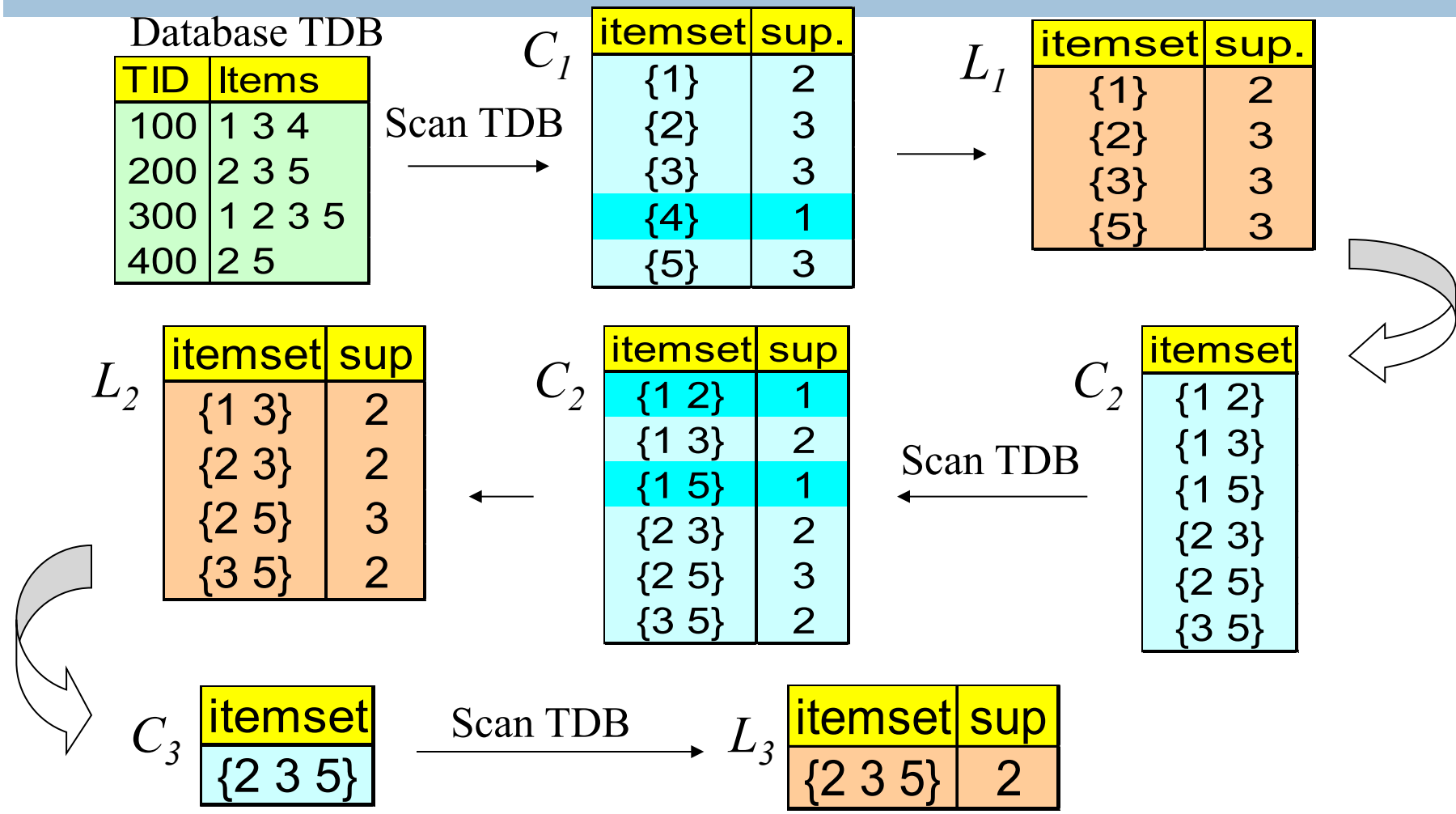





{a,d} is not frequent, so the 3-itemsets {a,b,d}, {a,c,d} and the 4-itemset {a,b,c,d}, are not generated.

# Apriori Execution Example (min_sup = 2)





# Rule Generation



- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  - If {A,B,C,D} is a frequent itemset, candidate rules:

ABC →D, ABD →C, ACD →B, BCD →A,
A →BCD, B →ACD, C →ABD, D →ABC
AB →CD, AC → BD, AD → BC, BC →AD,
BD →AC, CD →AB,

- If |L| = k, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)

# Esercizio 1

**Table 6.1.** Example of market basket transactions.

| Customer ID | Transaction ID | Items Bought |
|---|---|---|
| 1 | 0001 | $\{a, d, e\}$ |
| 1 | 0024 | $\{a, b, c, e\}$ |
| 2 | 0012 | $\{a, b, d, e\}$ |
| 2 | 0031 | $\{a, c, d, e\}$ |
| 3 | 0015 | $\{b, c, e\}$ |
| 3 | 0022 | $\{b, d, e\}$ |
| 4 | 0029 | $\{c, d\}$ |
| 4 | 0040 | $\{a, b, c\}$ |
| 5 | 0033 | $\{a, d, e\}$ |
| 5 | 0038 | $\{a, b, e\}$ |

Support?: e, (b,d), (b,d,e),

ID Transazione Items

1 {f,a,d,b}

2 {d,a,c,e,b}

3 {c,a,b,e}

4 {b,a,d}

Fissati il supporto minimo $\sigma = 60\%$ e la confidenza minima $\gamma = 80\%$

a) Indicare quali tra questi itemset sono frequenti.

- 1) {a}
- 2) {c}
- 3) {b,c}
- 4) {b,d}
- 5) {a,b,d}
- 6) {a,b,e}

ID Transazione Items

1 {f,a,d,b}

2 {d,a,c,e,b}

3 {c,a,b,e}

4 {b,a,d}

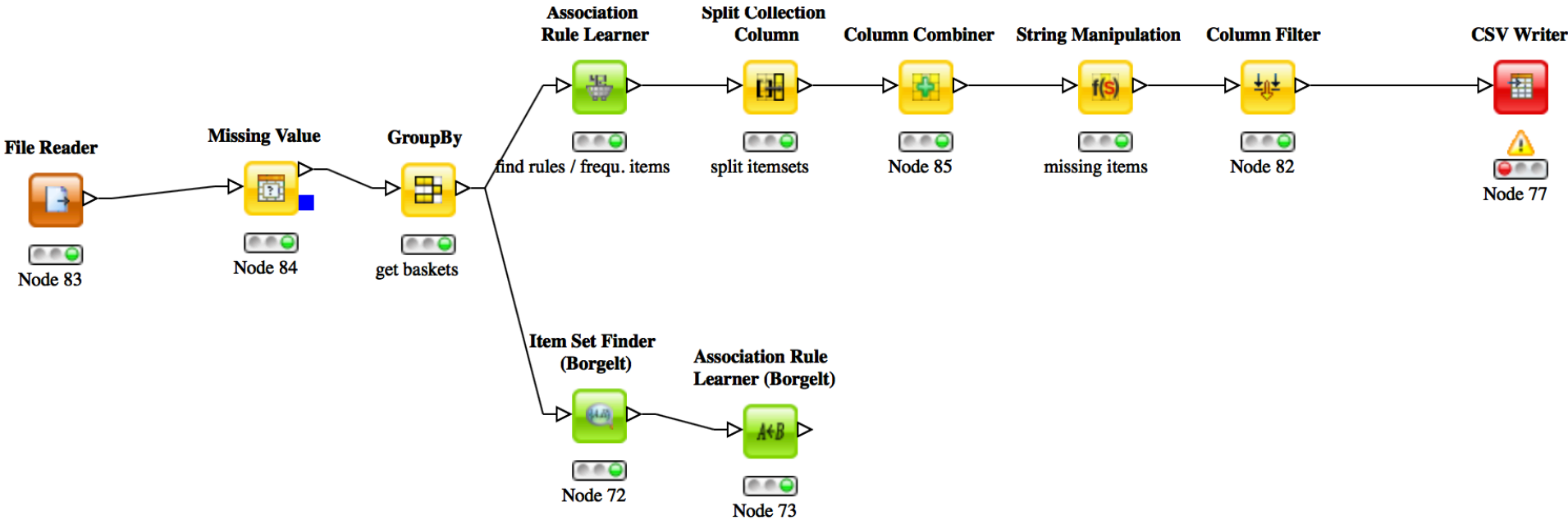
b) Indicare quali tra queste regole sono valide

- 1) {a}=>{b}
- 2) {a}=>{d}
- 3) {d}=>{a}
- 4) {d}=>{a,b}
- 5) {a,b}=>{d}

# EXERCISE WITH KNIME: FROM RETAIL

# From retail

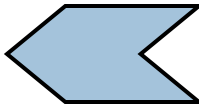



File Reader
Node 83
Missing Value
Node 84
GroupBy
get baskets
Association Rule Learner
find rules / frequ. items
Split Collection Column
split itemsets
Column Combiner
Node 85
String Manipulation
missing items
Column Filter
Node 82
CSV Writer
Node 77
Item Set Finder (Borgelt)
Node 72
Association Rule Learner (Borgelt)
Node 73

# Association rules - module outline



- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- How to compute AR
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
- How to reason on AR and how to evaluate their quality
  - Multiple-level AR
  - Interestingness
  - Correlation vs. Association
- Sequential Patterns



# Multidimensional AR



Associations between values of different attributes :

| CID | **nationality** | **age** | **income** |
|---|---|---|---|
| 1 | Italian | 50 | low |
| 2 | French | 40 | high |
| 3 | French | 30 | high |
| 4 | Italian | 50 | medium |
| 5 | Italian | 45 | high |
| 6 | French | 35 | high |

RULES:

**nationality** = French ⇒ **income** = high [50%, 100%]

**income** = high ⇒ **nationality** = French [50%, 75%]

**age** = 50 ⇒ **nationality** = Italian [33%, 100%]

# Single-dimensional vs Multi-dimensional AR



**Multi-dimensional**

<1, Italian, 50, low>

<2, French, 45, high>

**Single-dimensional**

<1, {nat/Ita, age/50, inc/low}>

<2, {nat/Fre, age/45, inc/high}>

Schema: <ID, a?, b?, c?, d?>

<1, yes, yes, no, no>

<2, yes, no, yes, no>

<1, {a, b}>

<2, {a, c}>

# Quantitative Attributes



- Quantitative attributes (e.g. age, income)
- Categorical attributes (e.g. color of car)

| CID | height | weight | income |
|---|---|---|---|
| 1 | 168 | 75,4 | 30,5 |
| 2 | 175 | 80,0 | 20,3 |
| 3 | 174 | 70,3 | 25,8 |
| 4 | 170 | 65,2 | 27,0 |

Problem: **too many distinct values**

Solution: **transform quantitative attributes in categorical ones via discretization.**

# Quantitative Association Rules



| CID | Age | Married | NumCars |
|---|---|---|---|
| 1 | 23 | No | 1 |
| 2 | 25 | Yes | 1 |
| 3 | 29 | No | 0 |
| 4 | 34 | Yes | 2 |
| 5 | 38 | Yes | 2 |

**[Age: 30..39] and [Married: Yes] ⇒ [NumCars:2]**

support = 40%
confidence = 100%

# Discretization of quantitative attributes



**Solution**: each value is replaced by the interval to which it belongs.

**height**: 0-150cm, 151-170cm, 171-180cm, >180cm

**weight**: 0-40kg, 41-60kg, 60-80kg, >80kg

**income**: 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

| CID | **height** | **weight** | **income** |
|---|---|---|---|
| 1 | 151-171 | 60-80 | >30 |
| 2 | 171-180 | 60-80 | 20-25 |
| 3 | 171-180 | 60-80 | 25-30 |
| 4 | 151-170 | 60-80 | 25-30 |

**Problem**: the discretization may be useless (see **weight**).

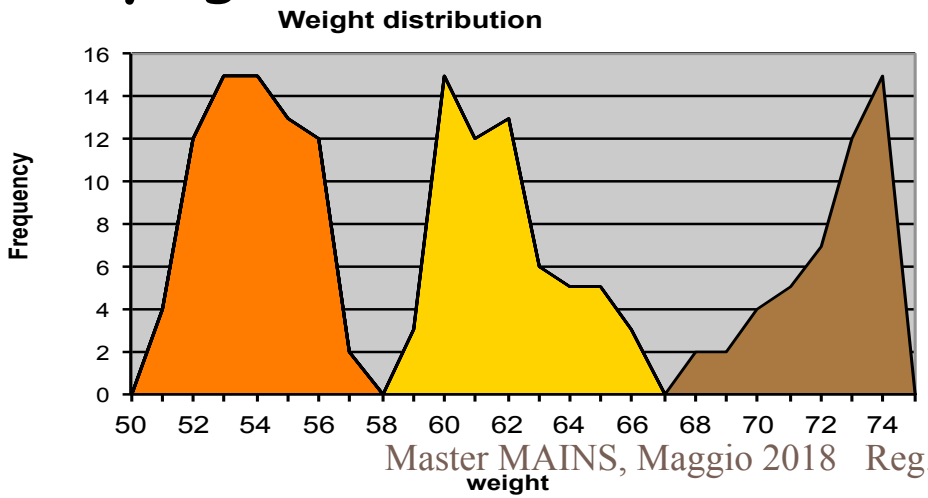
# How to choose intervals?



1. Interval with a fixed "reasonable" granularity
   Ex. intervals of 10 cm for height.

2. Interval size is defined by some domain dependent criterion
   Ex.: 0-20ML, 21-22ML, 23-24ML, 25-26ML, >26ML

3. Interval size determined by analyzing data, studying the distribution or using clustering




50 - 58 kg
59-67 kg
> 68 kg

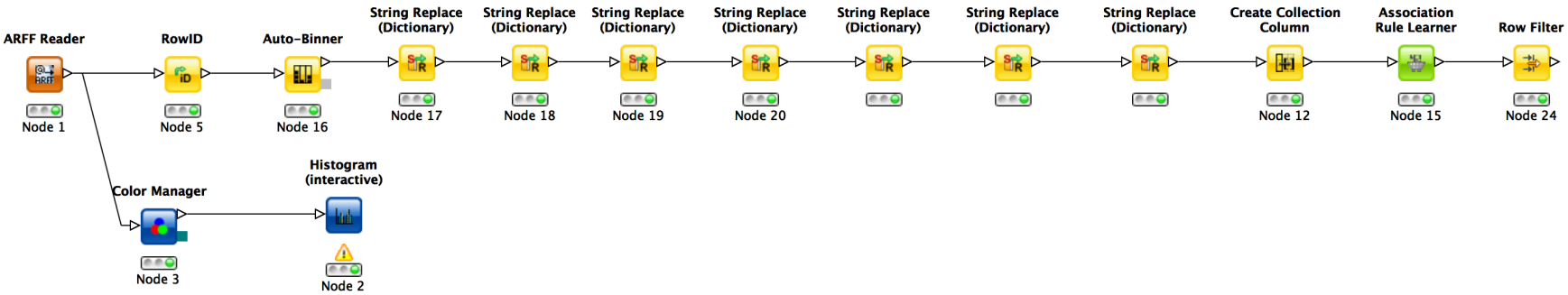
# EXERCISE2 WITH KNIME: TELCO

# DSB-Churn Dataset:



- The dataset consists of 20,000 (lines, rows) over 12 variables (fields, columns) describing features of customers of a mobile phone provider, including the class variable LEAVE

- COLLEGE : Is the customer college educated?
- INCOME: Annual income,
- OVERAGE: Average overcharges per month,
- LEFTOVER: Average % leftover minutes per month
- HOUSE: Value of dwelling (from census tract), HANDSET_PRICE: Cost of phone
- OVER_15MINS_CALLS_PER_MONTH: Average number of long (>15 mins) calls per month
- AVERAGE_CALL_DURATION: Average call duration
- REPORTED_SATISFACTION: Reported level of satisfaction
- REPORTED_USAGE_LEVEL: Self-reported usage level
- CONSIDERING_CHANGE_OF_PLAN: Was customer considering changing his/her plan?
- LEAVE : Class variable: whether customer left or stayed


ARFF Reader
RowID
Auto-Binner
String Replace (Dictionary)
String Replace (Dictionary)
String Replace (Dictionary)
String Replace (Dictionary)
String Replace (Dictionary)
String Replace (Dictionary)
String Replace (Dictionary)
Create Collection Column
Association Rule Learner
Row Filter
Node 1
Node 5
Node 16
Node 17
Node 18
Node 19
Node 20
Node 12
Node 15
Node 24
Color Manager
Histogram (interactive)
Node 3
Node 2

# Association rules - module outline

- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- How to compute AR
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
  - Constrained AR
- How to reason on AR and how to evaluate their quality
  - Interestingness
  - Correlation vs. Association
- Sequential Patterns
  - Example: Profiling with patterns

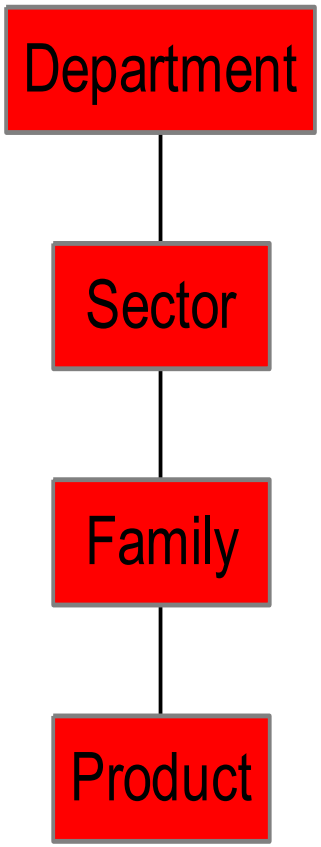
# Multilevel AR

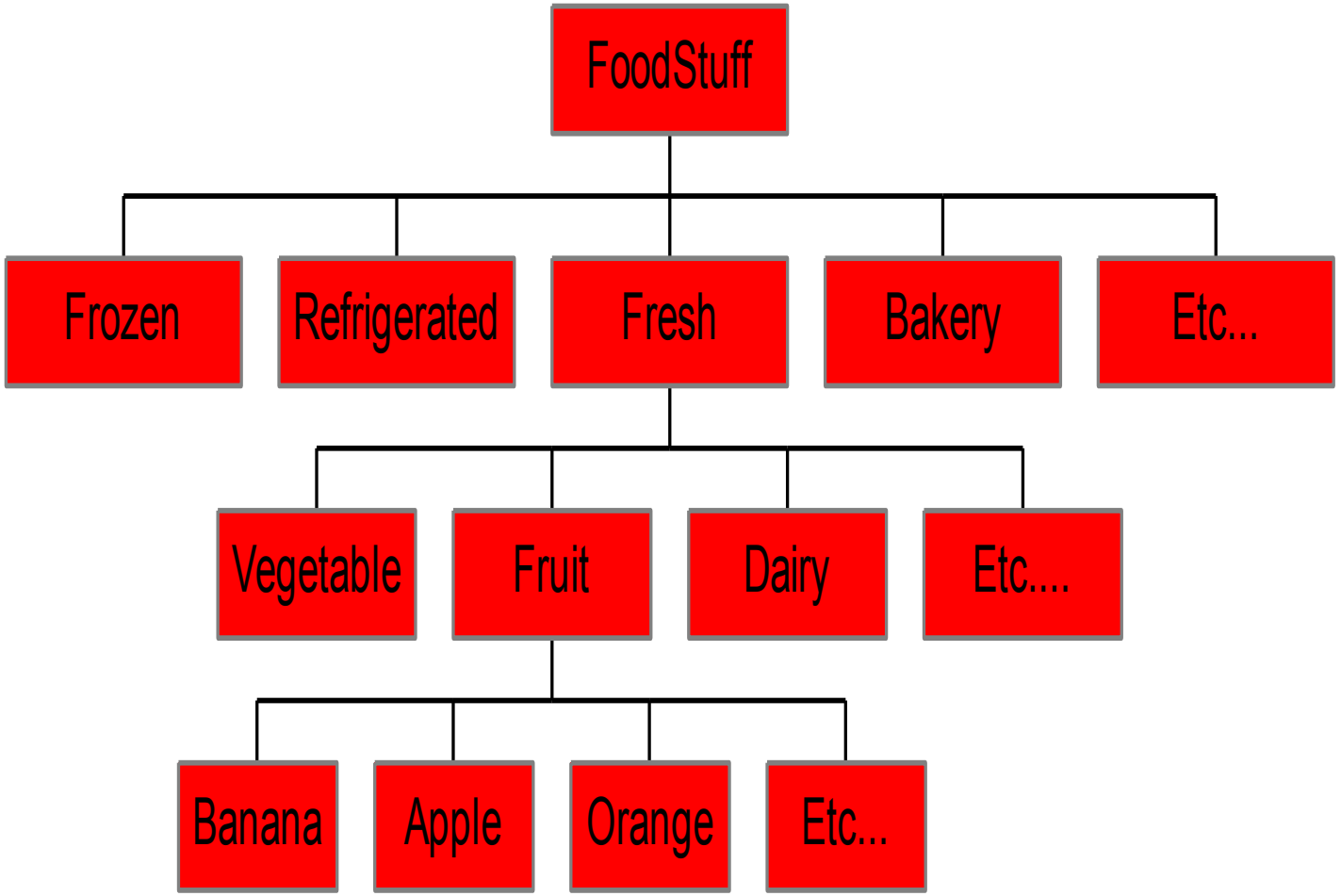


- Is difficult to find interesting patterns at a too primitive level
  - high support = too few rules
  - low support = too many rules, most uninteresting
- Approach: reason at suitable level of abstraction
- A common form of background knowledge is that an attribute may be generalized or specialized according to a hierarchy of concepts
- Dimensions and levels can be efficiently encoded in transactions
- Multilevel Association Rules : rules which combine associations with hierarchy of concepts

# Hierarchy of concepts

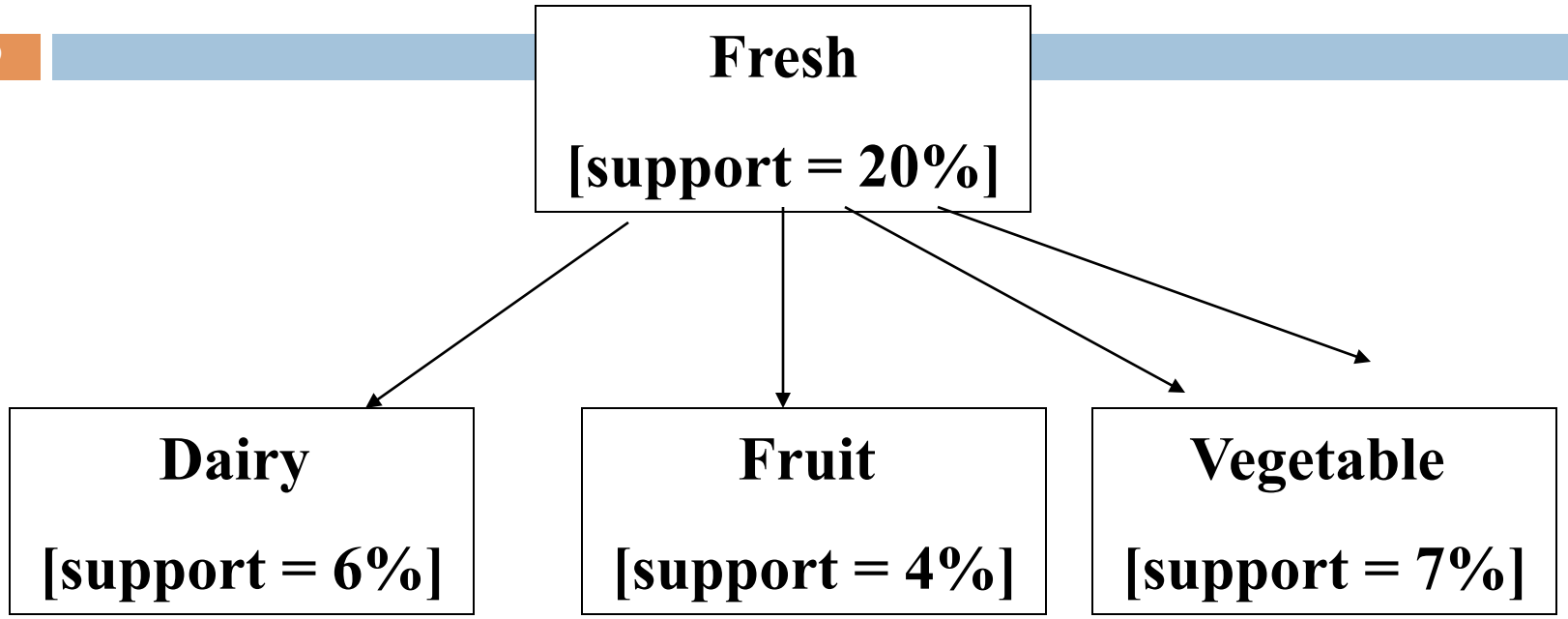




# Multilevel AR






**Fresh ⇒ Bakery [20%, 60%]**
**Dairy ⇒ Bread [6%, 50%]**
**Fruit ⇒ Bread [1%, 50%] is not valid**

# Association rules - module outline

- What are association rules (AR) and what are they used for:
  - The paradigmatic application: Market Basket Analysis
  - The single dimensional AR (intra-attribute)
- How to compute AR
  - Basic Apriori Algorithm and its optimizations
  - Multi-Dimension AR (inter-attribute)
  - Quantitative AR
  - Constrained AR
- How to reason on AR and how to evaluate their quality
  - Interestingness
  - Correlation vs. Association
- Sequential Patterns
  - Example: Profiling with patterns

# Reasoning with AR



- Significance:

  Example: <1, {a, b}>
  <2, {a} >
  <3, {a, b, c}>
  <4, {b, d}>

  {b} ⇒ {a} has confidence (66%)(2/3), but is not significant as support({a}) = 75%.

# Beyond Support and Confidence



- Example 1: (Aggarwal & Yu, PODS98)

| | coffee | not coffee | sum(row) |
|---|---|---|---|
| tea | 20 | 5 | 25 |
| not tea | 70 | 5 | 75 |
| sum(col.) | 90 | 10 | 100 |

- {tea} => {coffee} has high support (20%) and confidence (80%)
- However, a priori probability that a customer buys coffee is 90%
  - A customer who is known to buy tea is less likely to buy coffee (by 10%)
  - There is a negative correlation between buying tea and buying coffee
  - {~tea} => {coffee} has higher confidence(93%)(70/75)

# Correlation and Interest



- Two events are independent
  if $P(A \wedge B) = P(A)*P(B)$, otherwise are correlated.
- Interest = $P(A \wedge B) \,/\, P(B)*P(A)$
- **Interest** expresses measure of correlation
  - = 1 ⇒ A and B are independent events
  - less than 1 ⇒ A and B negatively correlated,
  - greater than 1 ⇒ A and B positively correlated.
  - In our example, I(*buy tea* ∧ *buy coffee* )=0.89 i.e. they are negatively correlated.

# Computing Interestingness Measure



- Given a rule $X \rightarrow Y$, information needed to compute rule interestingness can be obtained from a contingency table

Contingency table for $X \rightarrow Y$

| | Y | $\overline{Y}$ | |
|---|---|---|---|
| X | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $\overline{X}$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| | $f_{+1}$ | $f_{+0}$ | $\|T\|$ |

$f_{11}$: support of X and $\overline{Y}$
$f_{10}$: support of $\overline{X}$ and Y
$f_{01}$: support of $\overline{X}$ and $\overline{Y}$
$f_{00}$: support of X and Y

Used to define various measures

- support, confidence, lift, Gini, J-measure, etc.

# Statistical-based Measures



- Measures that take into account statistical dependence

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

# Beyond Support and Confidence



- Example 1: (Aggarwal & Yu, PODS98)

| | coffee | not coffee | sum(row) |
|---|---|---|---|
| tea | 20 | 5 | 25 |
| not tea | 70 | 5 | 75 |
| sum(col.) | 90 | 10 | 100 |

- Confidence= P(Coffee|Tea) = 0.80
- P(Coffee) = 0.9
- LIFT = 0.8/0.9= 0.8888 (< 1, therefore is negatively associated)

# Domain dependent measures



- Together with support, confidence, interest, …, use also (in post-processing) domain-dependent measures
- E.g., use rule constraints on rules
- Example: take only rules which are significant with respect their economic value
- sum(LHS)+ sum(RHS) > 100

# Rule Constraints



- Rule form constraints: meta-rule guided mining.
  - P(x, y) ^ Q(x, w) → takes(x, "database systems").
- Rule content constraint: constraint-based query optimization
  - sum(LHS) < 100 ^ min(LHS) > 20 ^ sum(RHS) > 1000
- 1-var: A constraint confining only one side (L/R) of the rule, e.g., as shown above.
- 2-var: A constraint confining both sides (L and R).
  - sum(LHS) < min(RHS) ^ max(RHS) < 5* sum(LHS)

# Which tools for market basket analysis?

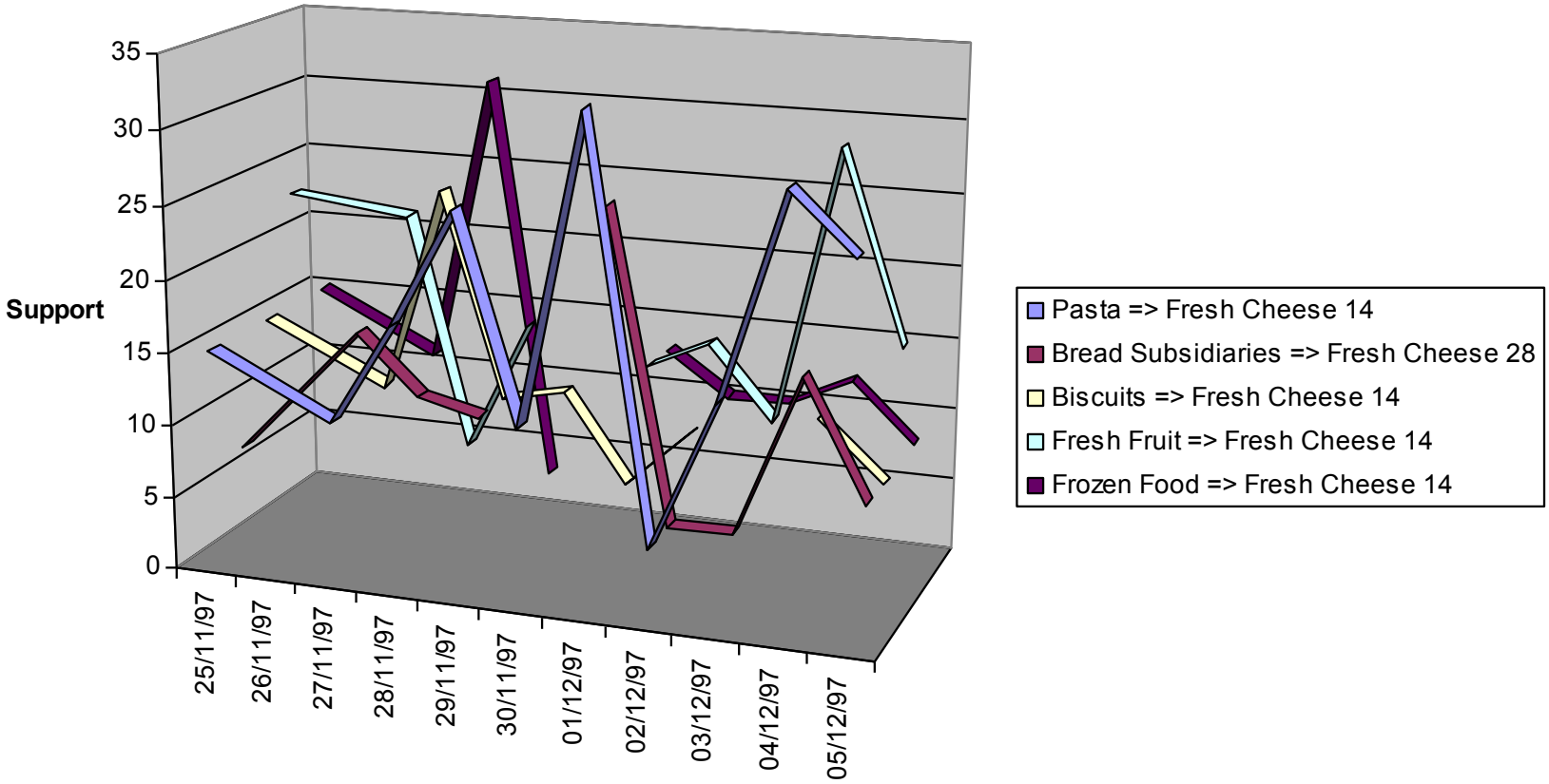


- Association rule are needed but insufficient
- Market analysts ask for business rules:
  - Is supermarket assortment adequate for the company's target class of customers?
  - Is a promotional campaign effective in establishing a desired purchasing habit?

# Business rules: temporal reasoning on AR



- Which rules are established by a promotion?
- How do rules change along time?



Customer profiling
Association Rules in Web Miming
AR & Atherosclerosis prevention study

# ASSOCIATION RULES - EXAMPLES

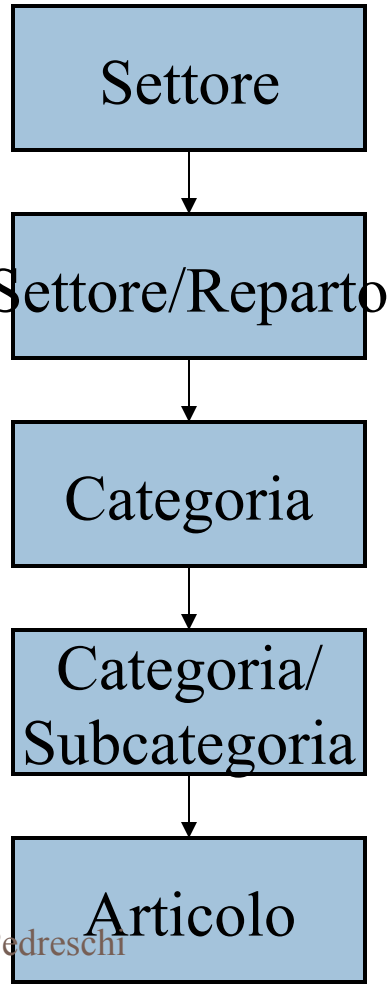
# PROGETTO "COOL PATTERNS"

## ANALISI DELLE VENDITE NELLA GRANDE DISTRIBUZIONE

# Data Understanding: Data description – Gerarchia prodotti



- La descrizione della gerarchia degli articoli è specificata nel file Excel `Classificazione Marketing.xls`.
- Si estraggono 4 tabelle che descrivono ciascuna un livello della gerarchia (chiave, descrizione)
  - Settori
    - 9 record, 2 campi (chiave: *cod_settore*)
  - Reparti
    - 54 record, 3 campi (chiave: *cod_settore* + *cod_reparto*)
  - Categorie
    - 402 record, 4 campi (chiave: *cod_categ*)
  - Subcategorie
    - 1 516 record, 5 campi (chiave: *cod_categ* + *cod_subcateg*)

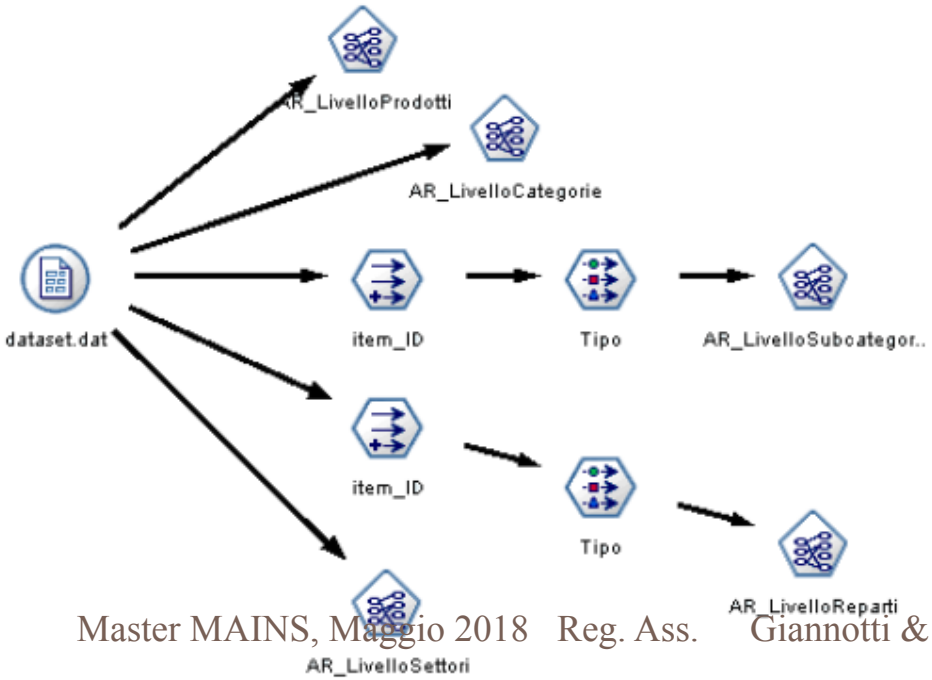


# Modeling – Obj 1

## Estrazione regole associative



- Data il data set in formato *transazionale*.
- L'attributo *key* identifica ogni transazione.
- A seconda del livello di astrazione considerato, i codici di articolo, subcategoria, categoria, reparto e settore sono gli attributi di input/output.



# Modeling – Obj 1

## Estrazione regole associative



- La strategia utilizzata per l'estrazione delle regole è quella del *reduced support*.
- Ogni livello di astrazione ha la sua soglia di supporto minimo
  - più basso è il livello nella gerarchia, più piccola è la soglia di supporto minimo corrispondente.

| Livello | Supporto minimo | Confidenza minima |
|---|---|---|
| Articoli | 0,01% | 80% |
| Subcategorie | 0,2% | 75% |
| Categorie | 0,7% | 75% |
| Reparti | 4% | 75% |
| Settori | 8% | 80% |

Regole interessanti → Lift maggiore di 1

# Evaluation – Obj 1
# Regole associative interessanti



- L'insieme di regole ottenuto è stato esportato in un file di testo in cui esiste un record per ogni regola

| Istanze | Supporto | Confidenza | Lift | Conseguente | Antecedente 1 |
|---|---|---|---|---|---|
| 53 | 0.01 | 92.5 | 4237.263 | 283917 | 283920 |

- Le regole ottenute non sono direttamente interpretabili.
- E' stato scritto il programma *PrettyPrinterApriori* che, data una regola "grezza", restituisce la corrispondente descrizione testuale.

[ 10 BICCH.CART.BIBO CIRC.200CC ] → [ PIATTI CART.BIBO CIRCUS D23X10 ]
Line: 70 Support: 0,01 Confidence: 92,5 Lift: 4237,263

# Evaluation – Obj 1
# Regole associative – Articoli



- [ 10 BICCH.CART.BIBO CIRC.200CC ] → [ PIATTI CART.BIBO CIRCUS D23X10 ]
  - Support: 0,01 Confidence: 92,5 Lift: 4237,263
- [ TELO 100X150 460 GR/MQ TU ] [ OSPITE 40X60 460 GR/MQ TU ] → [ ASCIUGAMANO 60X110 460 GR TU ]
  - Support: 0,01 Confidence: 91,4 Lift: 965,993
- [ BOCC.CANI POLLO/TACCH.KG1.23 ] [ BOC/NI GATTO VITELLO SIM.KG415 ] → [ BOCC.GATTI CONIGLIO SIMBA G415 ]
  - Support: 0,01 Confidence: 91,4 Lift: 390,042
- [ PIATTO FRUTTA MAZIME B.CO CM21 ] [ PIATTO F.DO MAXIME B.CO CM.17 ] → [ PIATTO P.NO MAXIME B.CO CM.25 ]
  - Support: 0,01 Confidence: 90 Lift: 3052,386
- [ LENZUOLO PIANO 150X280 RIGHE ] [ LENZUOLO ANGOLI 90X200 TU] → [ FEDERA 50X80 STAMPA RIGHE ]
  - Support: 0,01 Confidence: 87,8 Lift: 809,222

# Evaluation – Obj 1
# Regole associative – Articoli



- [ GOURM.GOLD DADINI GELLEE G85X8 ] [ GOURMET PERLE FIL.C/MANZO G85 ] →[ GOURMET PERLE FIL.CONIGLIO G85 ]
  - Support: 0,01 Confidence: 87,8 Lift: 492,757
- [ CUCCHIAIONE ACCIAIO INOX ] [ PALA FRITTO ACCIAIO INOX ] [ FORCHETTONE ACCIAIO INOX ] → [ SCHIUMAROLA IN ACCIAIO INOX ]
  - Support: 0,01 Confidence: 85,7 Lift: 1912,523
- [ APER.CAMPARI MIXX PEACH ML275 ] [ APERIT.CAMPARI MIXX LIME ML275 ] [ APERITIVO CAMP.GRADI 6,5 ML275 ] → [ CAMPARI MIXX ORANGE ML275 ]
  - Support: 0,01 Confidence: 83,3 Lift: 1314,55
- [ GASSOSA S. BENEDETTO LT.1.5 ] [ CEDRATA SAN BENEDETTO LT.1.5 ] [ ARANCIATA S.BENEDETTO LT.1,5 ] → [ SPUMA BIONDA LT1.5 S.BENEDETTO ]
  - Support: 0,01 Confidence: 83 Lift: 172,76
- [ BARAT.OVALE LT1,7 VTR COP.ACC. ] [ BARAT.OVALE LT0,84 VTR COP.ACC ] → [ BARAT.OVALE LT1,2 VTR COP.ACC. ]
  - Support: 0,01 Confidence: 82,9 Lift: 1993,002
- [ MOUSSE GAT.COOP MANZ/FEGAT.G85 ] [ MOUSSE GAT.COOP PES/ TROTAG85 ] → [ MOUSSE GATTO COOP POL/TAC.G85 ]
  - Support: 0,1 Confidence: 81,7 Lift: 712,617

# Evaluation – Obj 1
# Regole associative – Subcategorie



- [ BIBITE-ARANCIATE ] [ SNACK SALATI-PATATINE ] [ USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI ] [ USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA ] → [ BIBITE-COLE ]
  - Support: 0,1 Confidence: 88,2 Lift: 11,084
- [ USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA ] [ USA E GETTA TAVOLA-ACCESSORI USA E GETTA ] [ USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA ] → [ USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI ]
  - Support: 0,1 Confidence: 84,7 Lift: 12,767
- [ USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA ] [ USA E GETTA TAVOLA-STOV. PLAST. COLORATA DECORATA ] → [ USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI ]
  - Support: 0,1 Confidence: 82,2 Lift: 12,391
- [ SNACK SALATI-POP CORN/CEREALI ] [ SNACK SALATI-ESTRUSI ] [ BIBITE-ARANCIATE ] → [ BIBITE-COLE ]
  - Support: 0,1 Confidence: 82,2 Lift: 10,34
- [ CARAMELLE/PROD. BASE ZUCCH.-ALTRE CARAMELLE ] [ CARAMELLE/PROD. BASE ZUCCH.-CARAM.NORMALI ] [ CARAMELLE/PROD. BASE ZUCCH.-GOMME DA MASTICARE ] → [ PRODOTTI BASE CIOCCOLATO-SNACK ]
  - Support: 0,1 Confidence: 81,9 Lift: 8,693

# Evaluation – Obj 1
# Regole associative – Categorie



- [ UOVA ] [ OF PREPARATA ] [ VERDURA FRESCA ] [ LATTE ] [ FRUTTA FRESCA ] → [ ORTAGGI ]
  - Support: 0,8 Confidence: 85,2 Lift: 1,893
- [ CAFFE ] [ UOVA ] [ VERDURA FRESCA ] [ FRUTTA FRESCA ] → [ ORTAGGI ]
  - Support: 0,7 Confidence: 84,3 Lift: 1,871
- [ UOVA ] [ GRASSI ] [ VERDURA FRESCA ] [ AVICUNICOLO ] → [ ORTAGGI ]
  - Support: 0,9 Confidence: 83,5 Lift: 1,854
- [ OLIO DI OLIVA ] [ UOVA ] [ SUINO ] → [ BOVINO ]
  - Support: 0,7 Confidence: 78,9 Lift: 1,757
- [ ZUCCHERO ] [ IGIENE CARTA ] [ DETERGENTI SUPERFICI ] → [ DETERGENZA TESSUTI ]
  - Support: 0,7 Confidence: 76,6 Lift: 2,247

# Evaluation – Obj 1
# Regole associative – Reparti



- FRESCHI-CARNI BIANCHE ] [ FRESCHI-SURGELATI ] [ FRESCHI-GASTRONOMIA ] → [ FRESCHI-CARNI ROSSE ]
  - Support: 5,2 Confidence: 75,5 Lift: 1,217

- Al livello di Settore, non sono state trovate regole aventi Lift maggiore di 1.

# Profiling –



Per ogni regola "interessante:

- Il dataset finale è una tabella che contiene i dati di *tutti* e *soli* i clienti che hanno effettuato acquisti nel trimestre.
- La variabile target: ogni cliente ha associato un attributo binario (supporta o non supporta).
- Attributi predittori:
- A partire dall'albero di decisione ottenuto sono state generate le regole per la classificazione delle due classi.
- Per la creazione delle regole sono stati impostati livelli di confidenza minimi del 95%.

| Attributo | Tipo |
|---|---|
| *sesso* | flag |
| *stato_civile* | insieme discreto |
| *professione* | insieme discreto |
| *titolo_studio* | insieme discreto |
| *age* | intervallo |

# Evaluation – Obj 2

## Regole associative – Articoli



- La regola
  - [ GOURM.GOLD DADINI GELLEE G85X8 ] [ GOURMET PERLE FIL.C/ MANZO G85 ]→[ GOURMET PERLE FIL.CONIGLIO G85]
  - Support: 0,01 Confidence: 87,8 Lift: 492,757
- è supportata da 41 clienti. Il classificatore ottenuto ha una accuratezza del 96,07% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Casalinghe, non sposate, di età tra i 57 e i 63 anni, che hanno la terza media inferiore come titolo di studio.
  - Ragazze single ragioniere tra i 26 e i 29 anni che lavorano come impiegate
  - Uomini pensionati aventi età minore di 51 anni

# Evaluation – Obj 2

## Regole associative – Articoli



- La regola
  - [ APER.CAMPARI MIXX PEACH ML275 ] [ APERIT.CAMPARI MIXX LIME ML275] [ APERITIVO CAMP.GRADI 6,5 ML275 ] → [ CAMPARI MIXX ORANGE ML275 ]
  - Support: 0,01 Confidence: 83,3 Lift: 1314,55
- è supportata da 27 clienti. Il classificatore ottenuto ha una accuratezza del 97,6% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Ingegneri maschi aventi 26-27 anni
  - Ragazze dai 26 ai 30 anni che hanno un lavoro autonomo e sono diplomate
  - Impiegati single dai 26 ai 53 anni

# Evaluation – Obj 2

## Regole associative – Subcategorie



- La regola
  - [ USA E GETTA TAVOLA-STOV. CARTA COLORATA DECORATA ] [ USA E GETTA TAVOLA-ACCESSORI USA E GETTA ] [ USA E GETTA TAVOLA-STOVIGLIE PLASTICA BIANCA ]→[ USA E GETTA TAVOLA-TOVAGLIE-TOVAGLIOLI ]
  - Support: 0,1 Confidence: 84,7 Lift: 12,767
- è supportata da 158 clienti. Il classificatore ottenuto ha una accuratezza del 88,13% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Uomini celibi che lavorano per enti pubblici aventi 43-45 anni
  - Liberi professionisti aventi titolo di studio media inferiore di 59-60 anni
  - Militari di carriera sposati di aventi 38-41 anni

# Evaluation – Obj 2

## Regole associative – Subcategorie



- La regola
  - [ SNACK SALATI-POP CORN/CEREALI ] [ SNACK SALATI-ESTRUSI ] [BIBITE-ARANCIATE ] → [ BIBITE-COLE ]
  - Support: 0,1 Confidence: 82,2 Lift: 10,34
- è supportata da 155 clienti. Il classificatore ottenuto ha una accuratezza del 86,73% su tutti i dati, e del 100% sui dati classificati come positivi.
- Profilo cliente:
  - Ragazze disoccupate di 30-34 anni aventi un diploma magistrale
  - Vedovi di 57-60 anni liberi professionisti
  - Uomini/donne sposati di 32-40 anni e impiegati

# ATHEROSCLEROSIS PREVENTION STUDY

Giannotti & Pedreschi

2nd Department of Medicine, 1st Faculty of Medicine of Charles University and Charles University Hospital, U nemocnice 2, Prague 2 (head. Prof. M. Aschermann, MD, SDr, FESC)

# Atherosclerosis prevention study:



- The STULONG 1 data set is a real database that keeps information about the study of the development of atherosclerosis risk factors in a population of middle aged men.

- Used for Discovery Challenge at PKDD 00-02-03-04

# The input data



| Data from Entry and Exams | | |
|---|---|---|
| **General characteristics** | **Examinations** | **habits** |
| Marital status | Chest pain | Alcohol |
| Transport to a job | Breathlesness | Liquors |
| Physical activity in a job | Cholesterol | Beer 10 |
| Activity after a job | Urine | Beer 12 |
| Education | Subscapular | Wine |
| Responsibility | Triceps | Smoking |
| Age | | Former smoker |
| Weight | | Duration of smoking |
| Height | | Tea |
| | | Sugar |
| | | Coffee |

# The input data



| DEATH CAUSE | PATIENTS | % |
|---|---|---|
| myocardial infarction | 80 | 20.6 |
| coronary heart disease | 33 | 8.5 |
| stroke | 30 | 7.7 |
| other causes | 79 | 20.3 |
| sudden death | 23 | 5.9 |
| unknown | 8 | 2.0 |
| tumorous disease | 114 | 29.3 |
| general atherosclerosis | 22 | 5.7 |
| TOTAL | 389 | 100.0 |

# The prepared data



| Patient | General characteristics | | Examinations | | Habits | | Cause of death |
|---|---|---|---|---|---|---|---|
| | Activity after work | Education | Chest pain | … | Alcohol | ….. | |
| 1 | moderate activity | university | not present | | no | | Stroke |
| 2 | great activity | | not ischaemic | | occasionally | | myocardial infarction |
| 3 | he mainly sits | | other pains | | regularly | | tumorous disease |
| …… | …….. | …….. | …….. | .. | … | …… | alive |
| 389 | he mainly sits | | other pains | | regularly | | tumorous disease |

# Descriptive Analysis/ Subgroup Discovery / Association Rules



Are there strong relations concerning death cause?

General characteristics (?) ⇒ Death cause (?)

Examinations (?) ⇒ Death cause (?)

Habits (?) ⇒ Death cause (?)

Combinations (?) ⇒ Death cause (?)

# Example of extracted rules



- Education(university) & Height<176-180> ⇒Death cause (tumouros disease), *16 ; 0.62*
- It means that on tumorous disease have died 16, i.e. 62% of patients with university education and with height 176-180 cm.

# Example of extracted rules



- Physical activity in work(he mainly sits) & Height<176-180> ⇒ Death cause (tumouros disease), *24; 0.52*
- It means that on tumorous disease have died 24 i.e. 52% of patients that mainly sit in the work and whose height is 176-180 cm.

# Example of extracted rules



- Education(university) & Height<176-180> ⇒Death cause (tumouros disease),
  *16; 0.62; +1.1;*
- the relative frequency of patients who died on tumorous disease among patients with university education and with height 176-180 cm is 110 per cent higher than the relative frequency of patients who died on tumorous disease among all the 389 observed patients

# MBA in Text / Web Content Mining



- Documents Associations
  - Find (content-based) associations among documents in a collection
  - Documents correspond to items and words correspond to transactions
  - Frequent itemsets are groups of docs in which many words occur in common

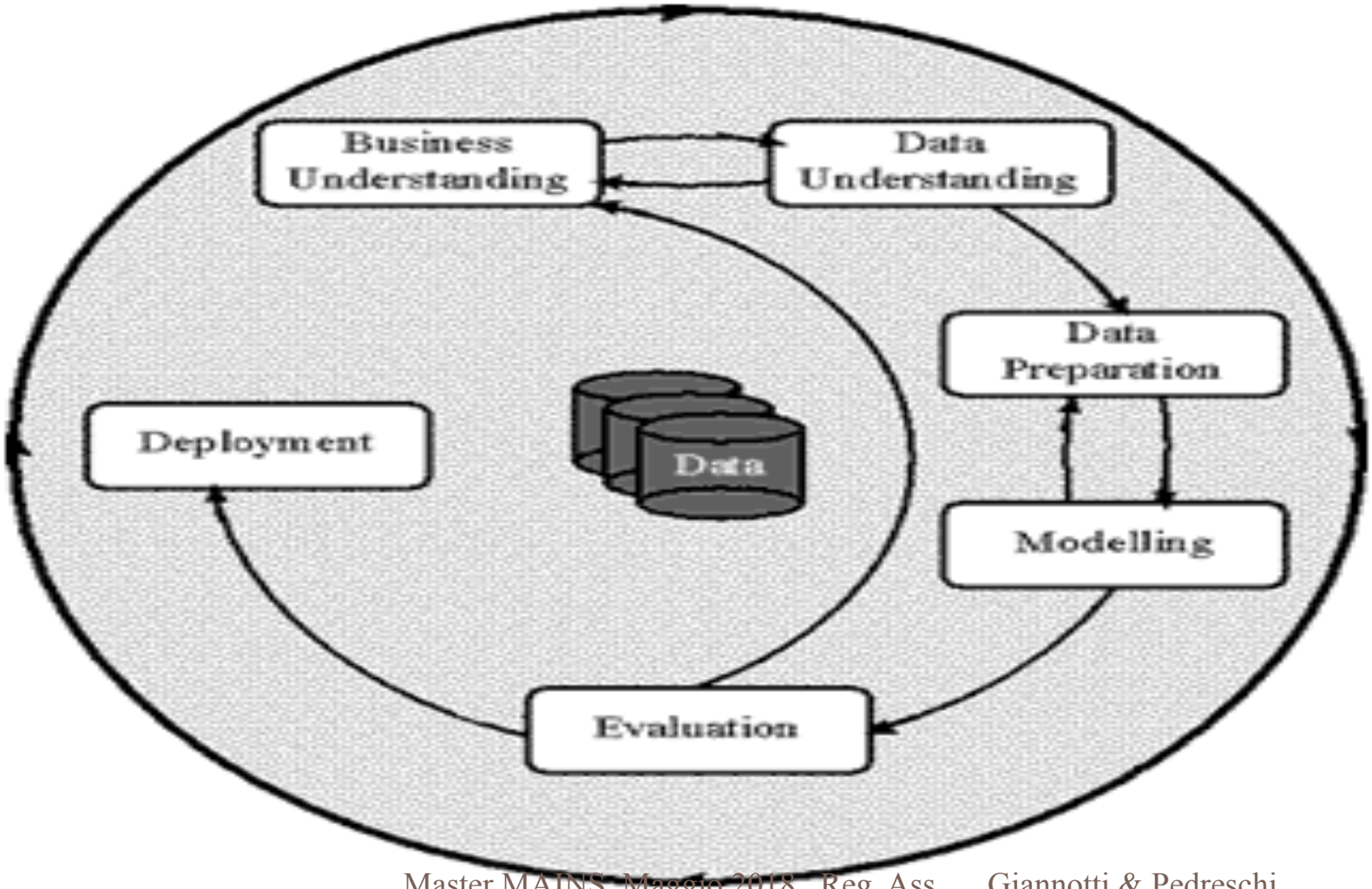
| | Doc 1 | Doc 2 | Doc 3 | . . . | Doc n |
|---|---|---|---|---|---|
| business | 5 | 5 | 2 | . . . | 1 |
| capital | 2 | 4 | 3 | . . . | 5 |
| fund | 0 | 0 | 0 | . . . | 1 |
| : | : | : | : | : : : | : |
| invest | 6 | 0 | 0 | . . . | 3 |

- Term Associations
  - Find associations among words based on their occurrences in documents
  - similar to above, but invert the table (terms as items, and docs as transactions)

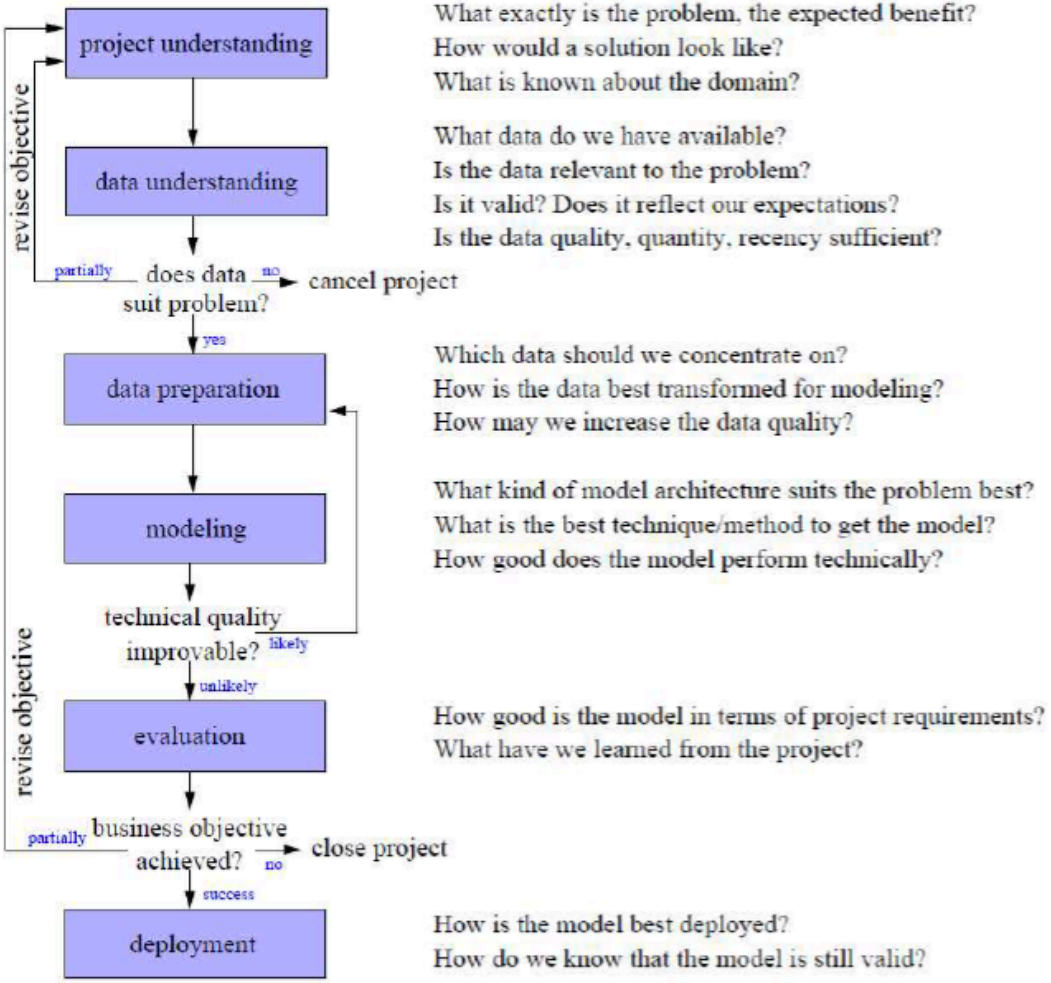
# CRISP-DM: THE LIFE CICLE OF A DATA MINING PROJECT

# CRISP Methodology

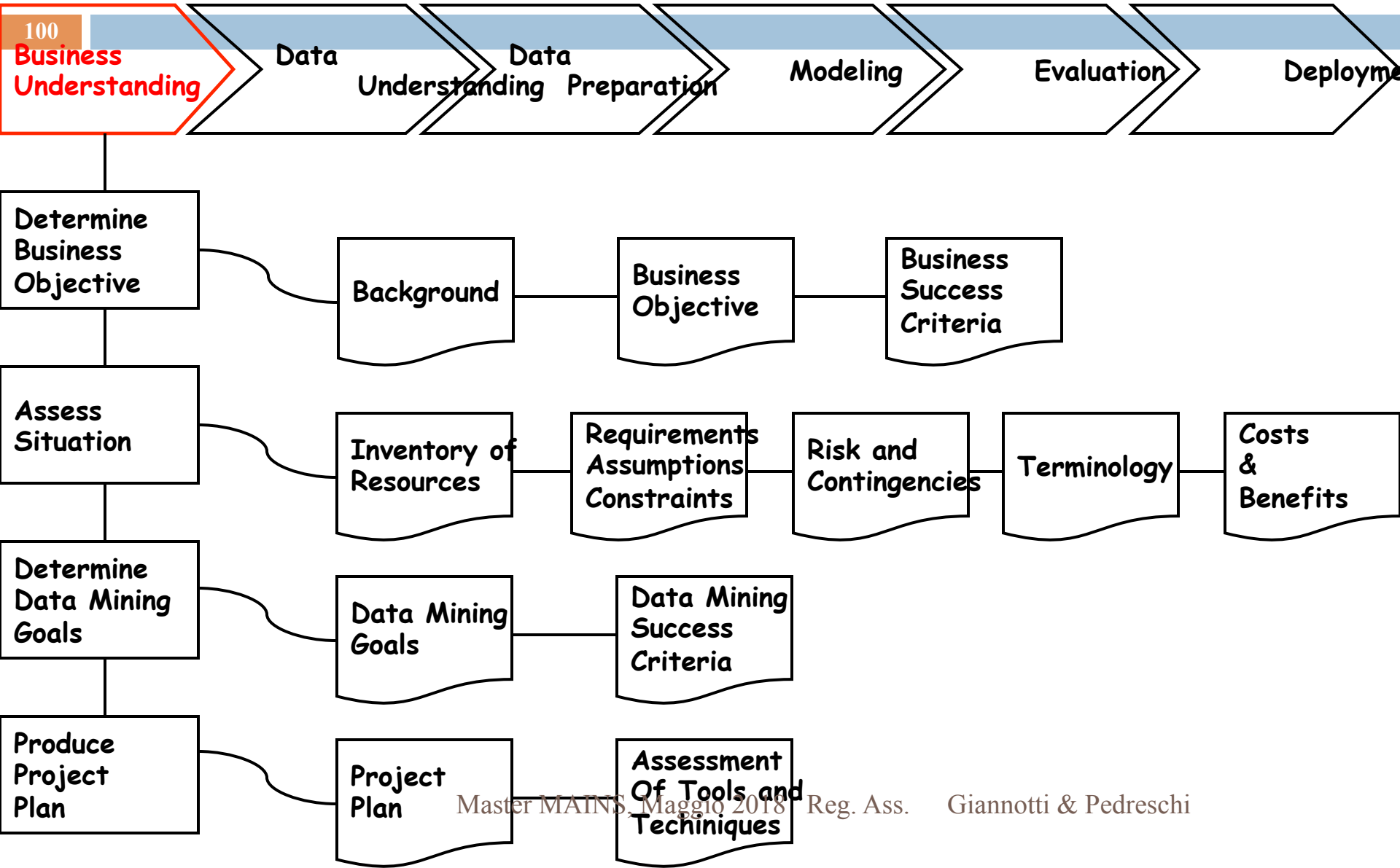


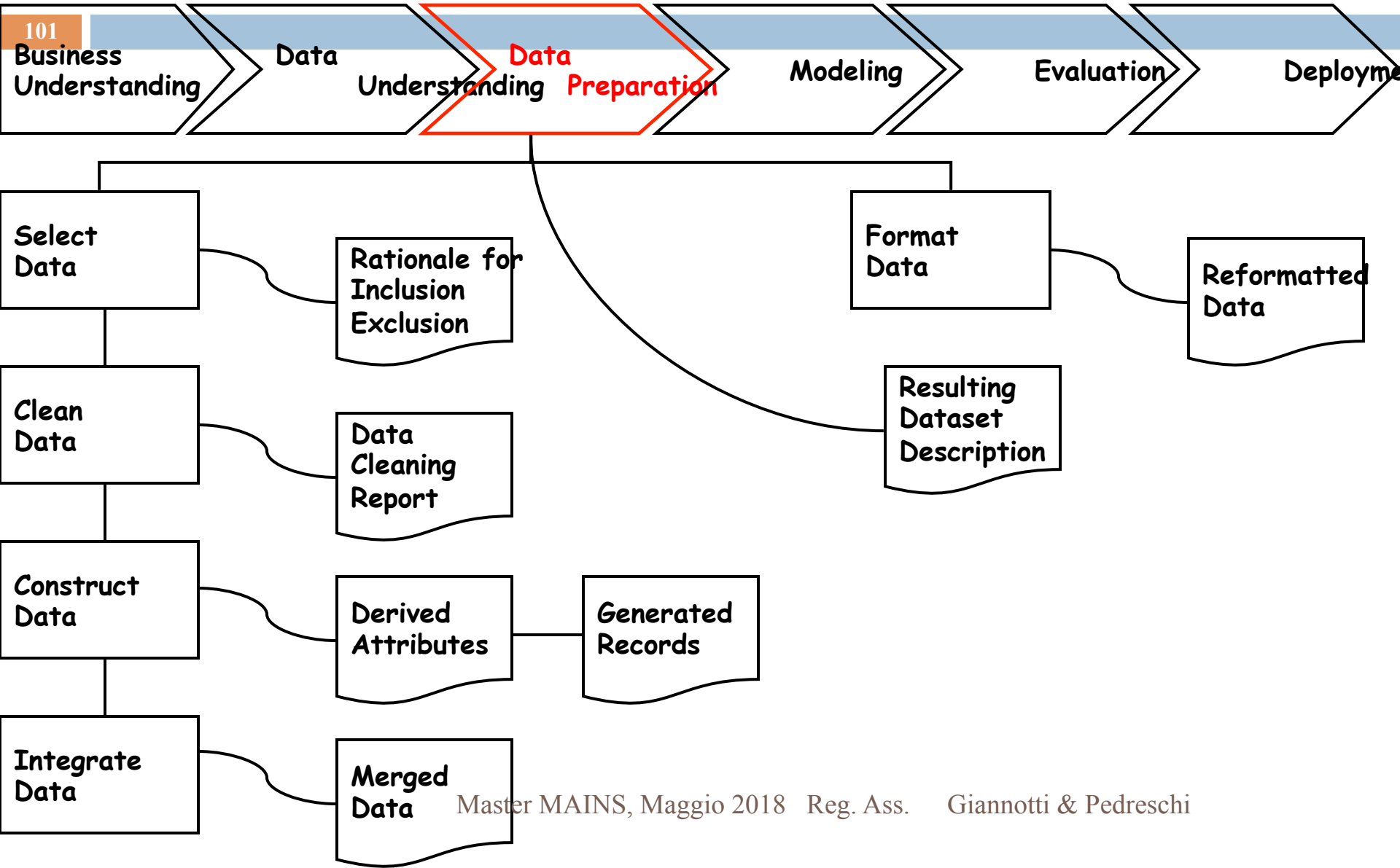


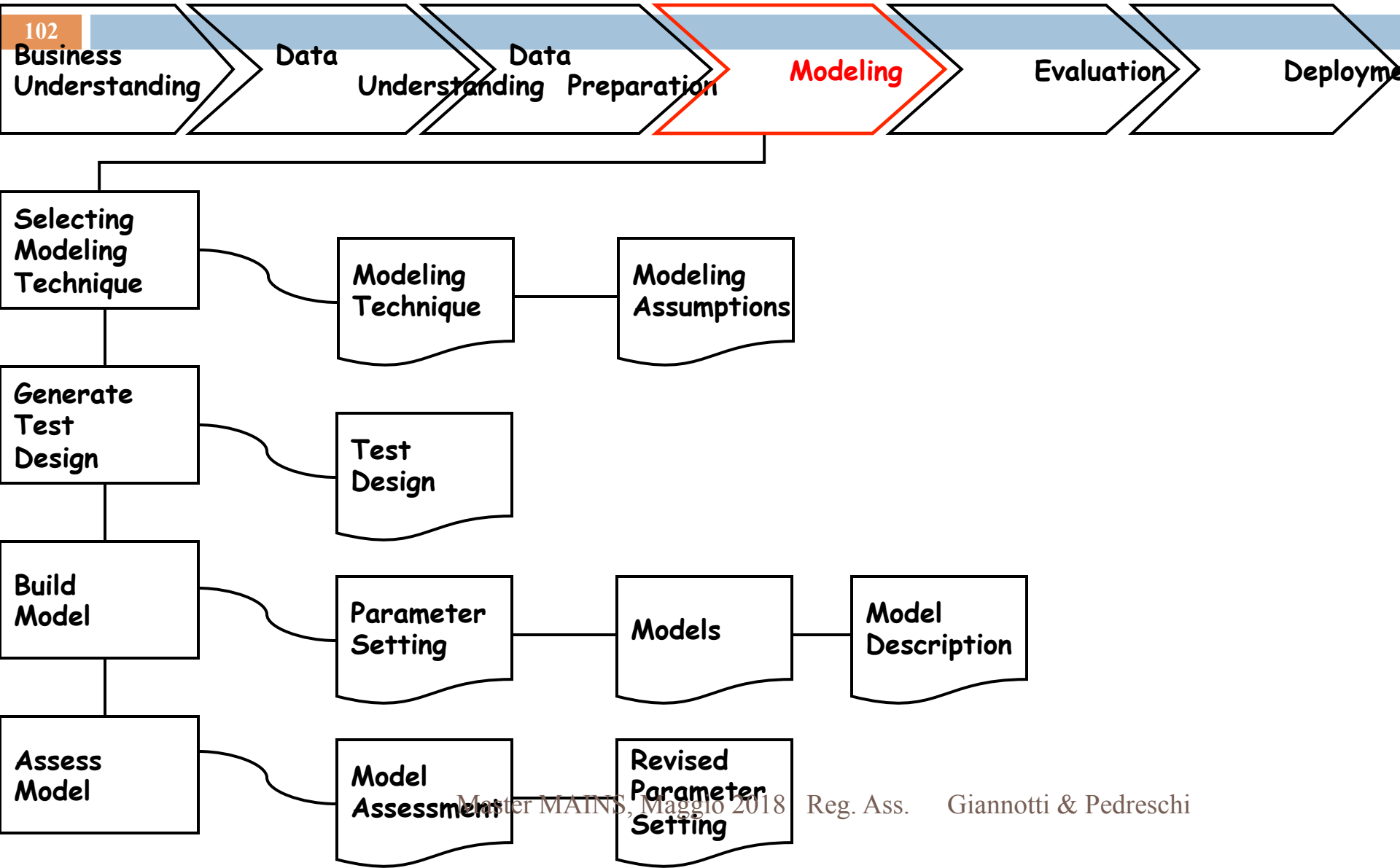
# CRISP

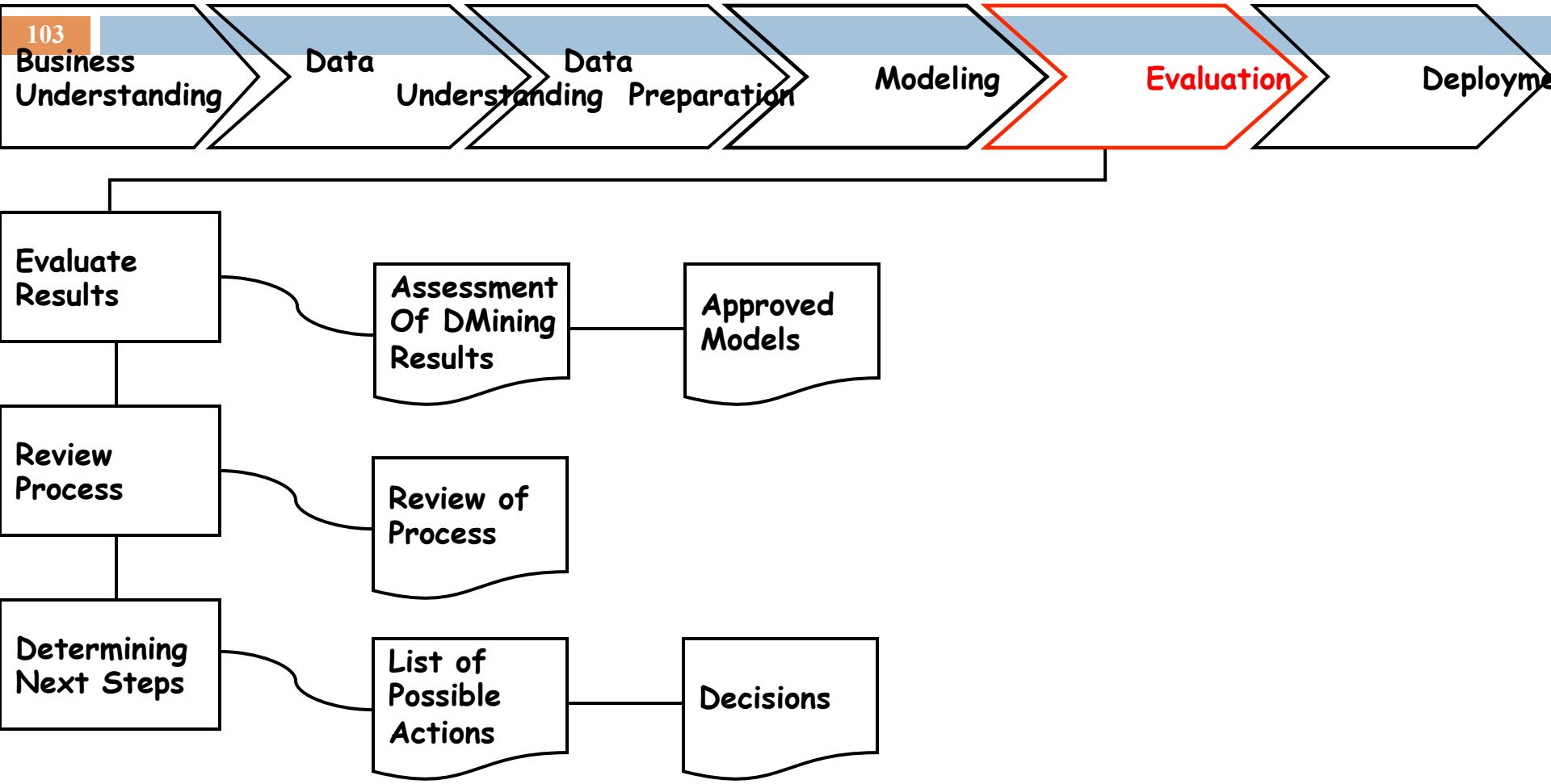



- **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- Iteration as a rule
- Process of data exploration

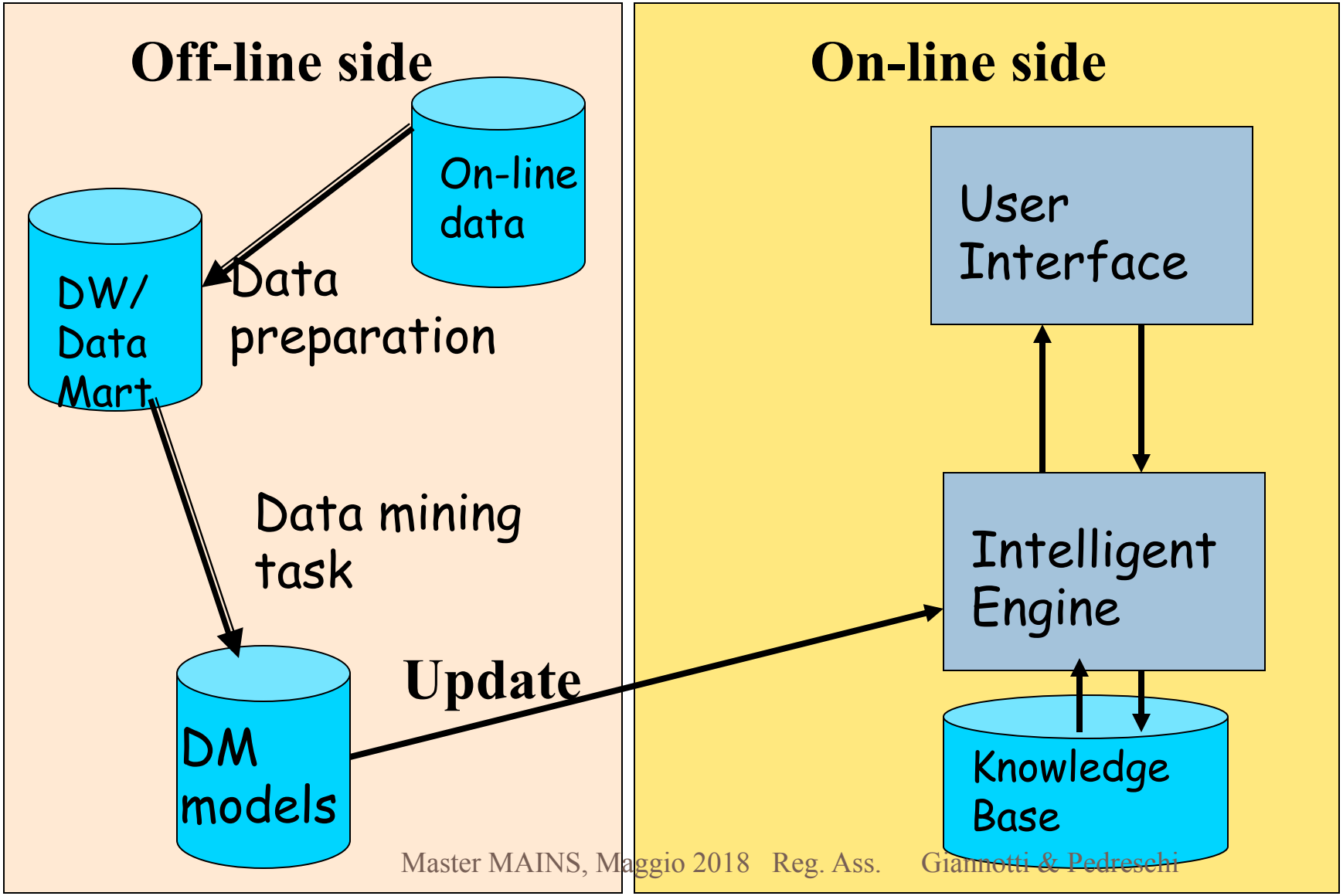

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
Determine Business Objective
Background
Business Objective
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements Assumptions Constraints
Risk and Contingencies
Terminology
Costs & Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Assessment Of Tools and Techniques


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
Select Data
Rationale for Inclusion Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data
Resulting Dataset Description


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
Selecting Modeling Technique
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Setting
Models
Model Description
Assess Model
Model Assessment
Revised Parameter Setting


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployme
Evaluate Results
Assessment Of DMining Results
Approved Models
Review Process
Review of Process
Determining Next Steps
List of Possible Actions
Decisions

# Mining Based Decision Support System: Adaptive Architecture



# The Big Data Analytics technologies



Visual workflow Tools

Specialized Libraries

Programming Languages