MINERAÇÃO MASSIVA DE DADOS

Parte 1 – Introdução à Mineração de Dados

Marcial Porto Fernández
marcial@larces.uece.br

Mestrado Acadêmico em Ciência da Computação (MACC)
Universidade Estadual do Ceará (UECE)
Laboratório de Sistemas Digitais (LASID)

Sumário

- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

Sumário

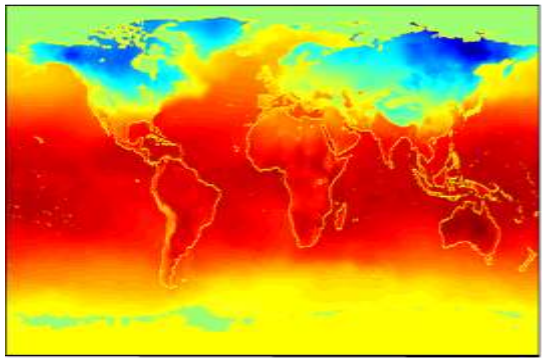
- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

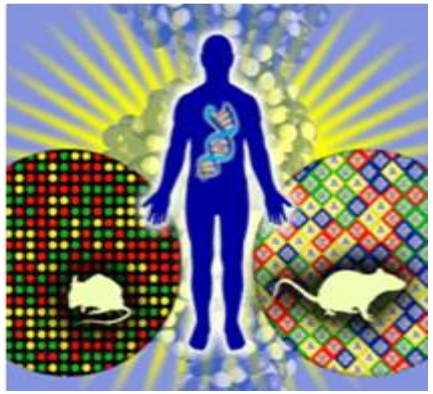
Mineração de Dados para Empresas

- Quantidades gigantescas de dados coletados e armazenados pelas empresas:
  - Dados de compras de clientes em lojas,
  - Dados de comércio eletrônico,
  - Dados de navegação na internet
  - Dados de transações bancárias, ou de cartão de crédito.
  - Dados dos deslocamento dos clients.
- Com essas informações é possivel:
  - Definir produto a oferecer ao cliente
  - Definir melhor forma de pagamento,
  - **Aumentar as vendas!**

Mineração de Dados para Ciência

- Grande massa de dados coletadas e armazenada
  - Sensores em satélites
  - Telescópios
  - Informações de genes
  - Simulações científicas
  - Dispositivos moveis e redes sociais
- O que pode produzir com tais dados:
  - Análise mais profunda dos dados
  - Descoberta de correlações e padrões não óbvios
  - Desenvolvimento do conhecimento
  - **Publicar mais papers!**





O que é Mineração de Dados?

Produzir conhecimento novo escondido em grandes bases de dados

- Grande disponibilidade de dados com fácil acesso
- Grande quantidade de informação útil escondida
- Disponibilidade de recursos computacionais com capacidade ilimitada (ou muito grande)
- Mineração de Dados ou Knowledge Discovery in Databases (KDD) visa otimizar e automatizar o processo de descoberta das tendências e padrões contidos nos dados, potencialmente úteis e interpretáveis.

Definição de Mineração de Dados

“É a exploração e a análise, por meio automático ou semi-automático, de grandes quantidades de dados, a fim de descobrir padrões e regras significativas.”

[Berry e Linoff, 1997]

“É o processo de reconhecimento de padrões válidos ou não, existentes nos dados armazenados em um banco de dados.”

[Fayyad, Piatetsky-Shapiro & Smyth,1995]

Mineração de Dados: uma área multidisciplinar

- Banco de Dados
- Estatística
- Algoritmos
- Computação de Alto-desempenho
- Aprendizado de Máquina
- Visualização
- Matemática
- ...

Desafios

- Grandes conjuntos de dados
- Eficiência do algoritmo é importante
- Escalabilidade do algoritmo é importante
- Dados do mundo real
- Muitos valores faltosos
- Conhecimento do Domínio é importante

What is (not) Data Mining?

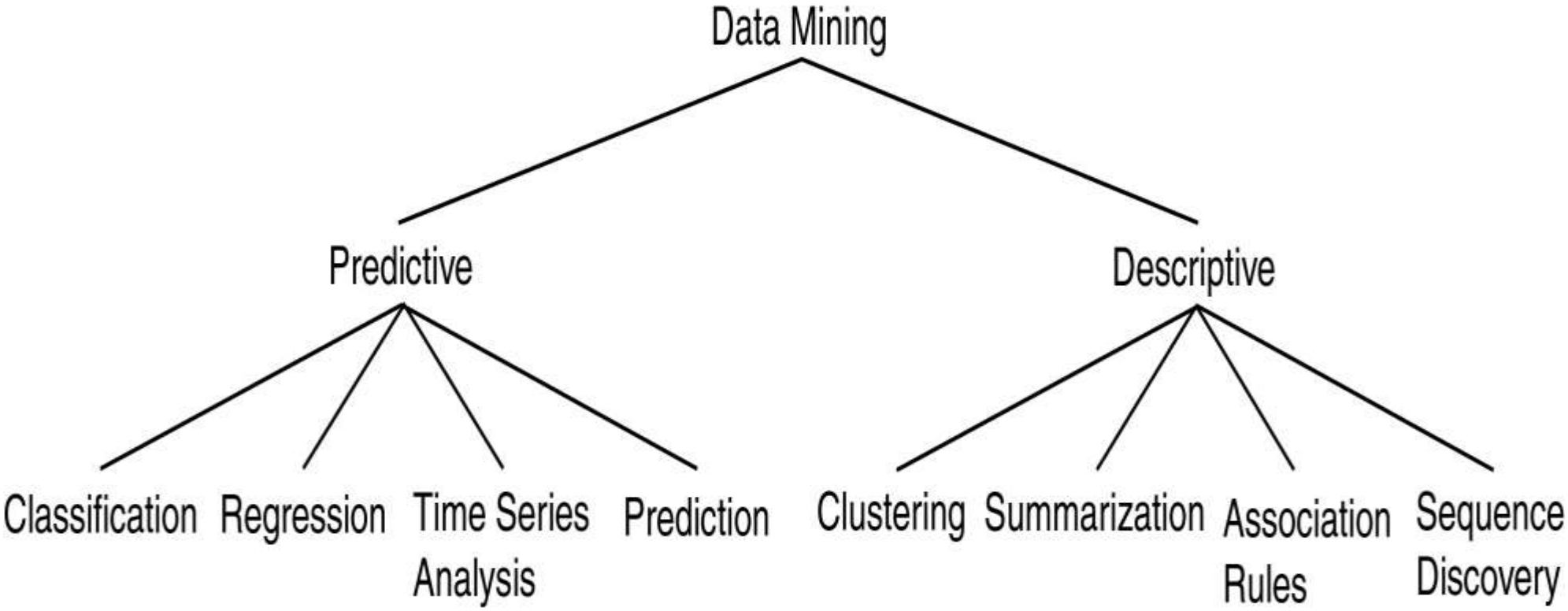
- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com)

Sumário

- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

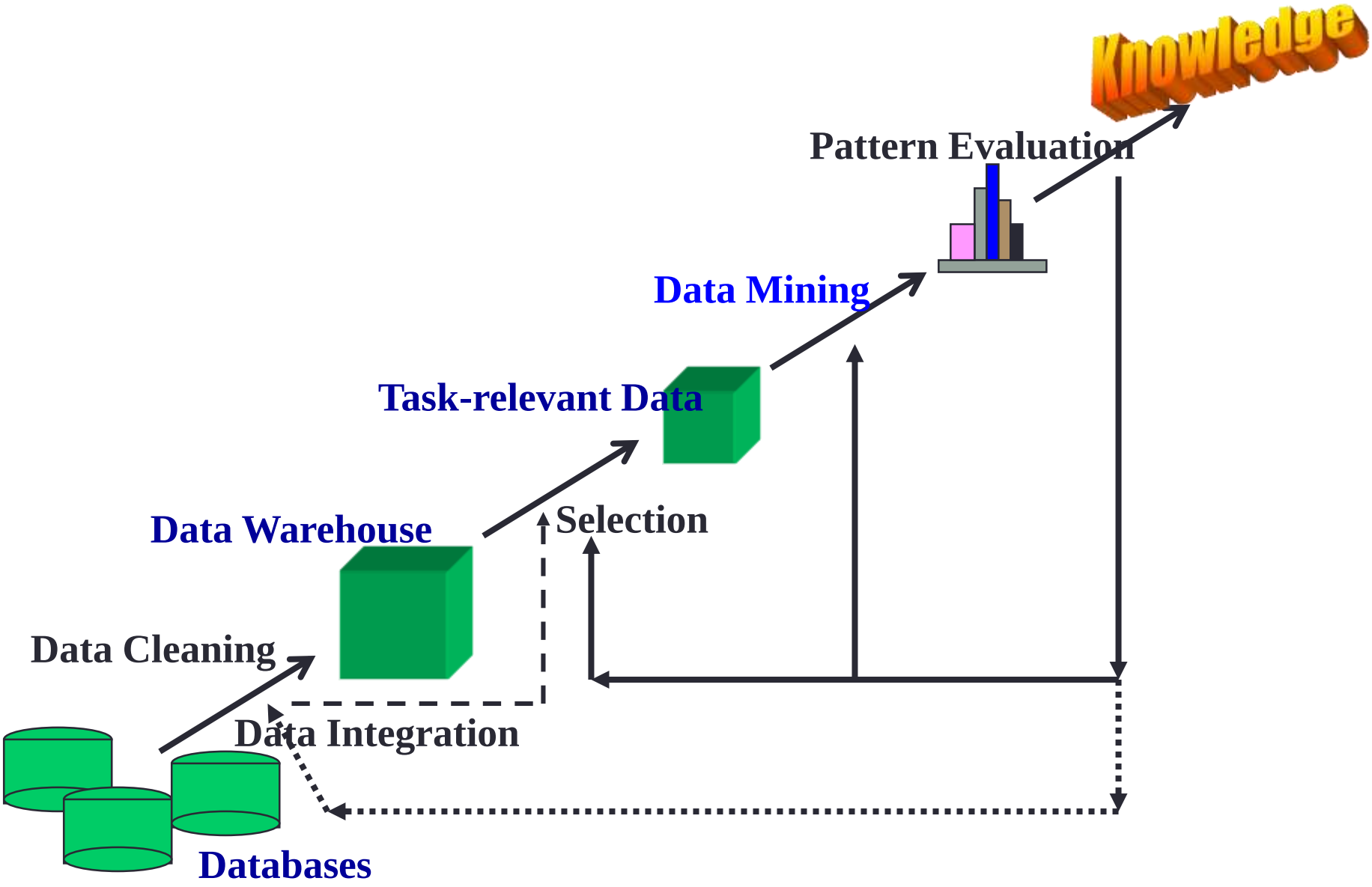
Data Mining Models



Data Mining: On What Kinds of Data?

- Relational database
- Data warehouse
- Transactional database
- Spatial and temporal data
- Time-series data
- Stream data
- Image data
- Multimedia database
- Text databases & WWW

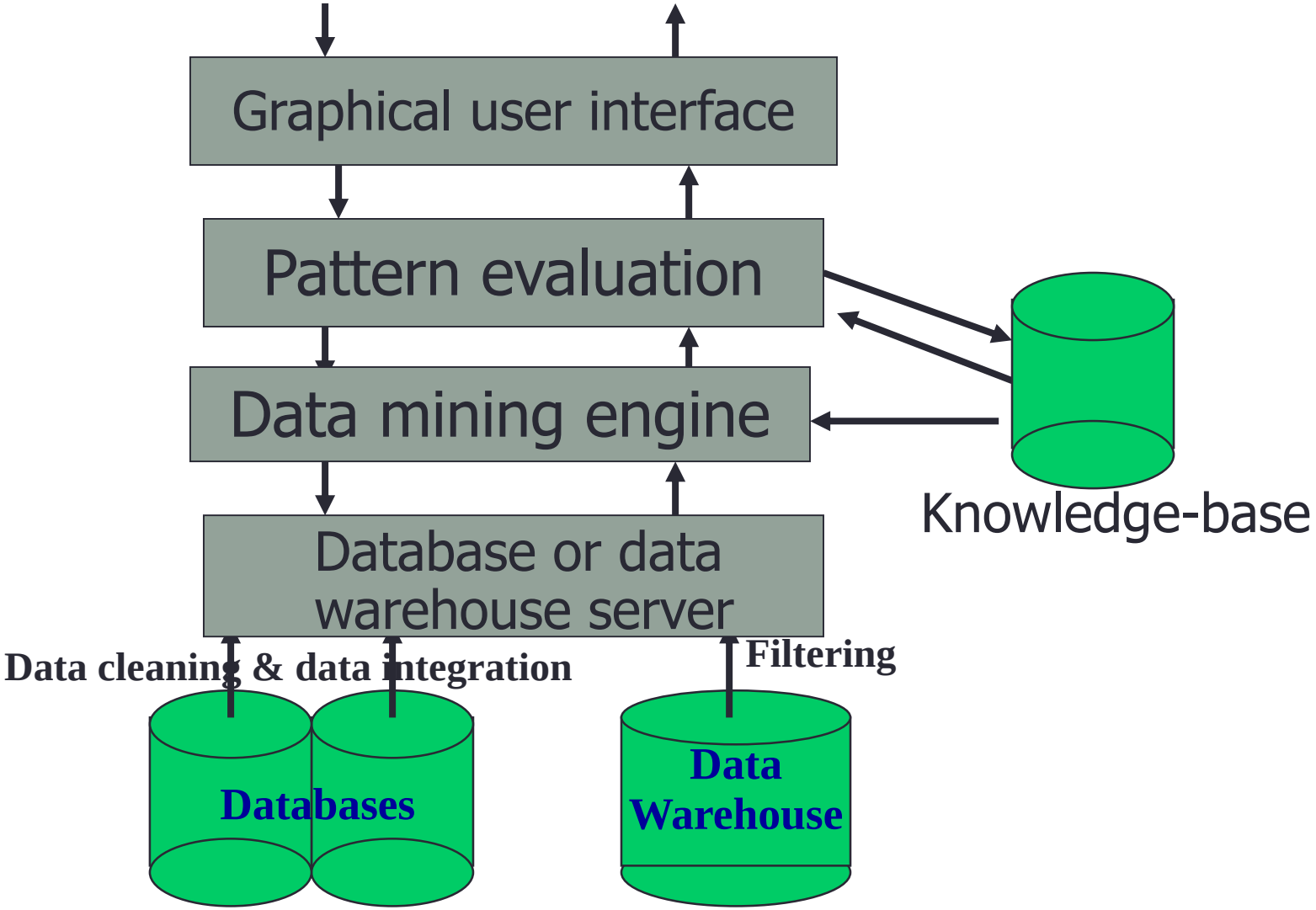
Data Mining: A KDD Process



Steps of a KDD Process

- Learning the application domain
  - Relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction.
- Choosing functions of data mining
  - Summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - Visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Architecture: Typical Data Mining System



Data Mining Functionalities (1)

- Concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics
- Association (correlation and causality)
  - Diaper → Beer [0.5%, 75%]
- Classification and Prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
  - Presentation: decision-tree, classification rule, neural network

Data Mining Functionalities (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: a data object that does not comply with the general behavior of the data
  - Useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis

Sumário

- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

Técnicas de Mineração de Dados

Aprendizado	Tarefa	Técnica
Não Supervisionado	Associação	Regras de Associação Padrões Seqüenciais
	Agrupamento	*Clustering*
Supervisionado	Classificação	Regras de Indução Árvores de Decisão MBR – *Memory Based Reasoning* Redes Neurais
	Análise de Desvios	Árvores de Decisão Redes Neurais

Sumário

- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

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

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery

- Marketing and Sales Promotion:
  - Let the rule discovered be
  - {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer:

$$Diapers \rightarrow Beer,\ support = 20\%,\ confidence = 85\%$$

Sumário

- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

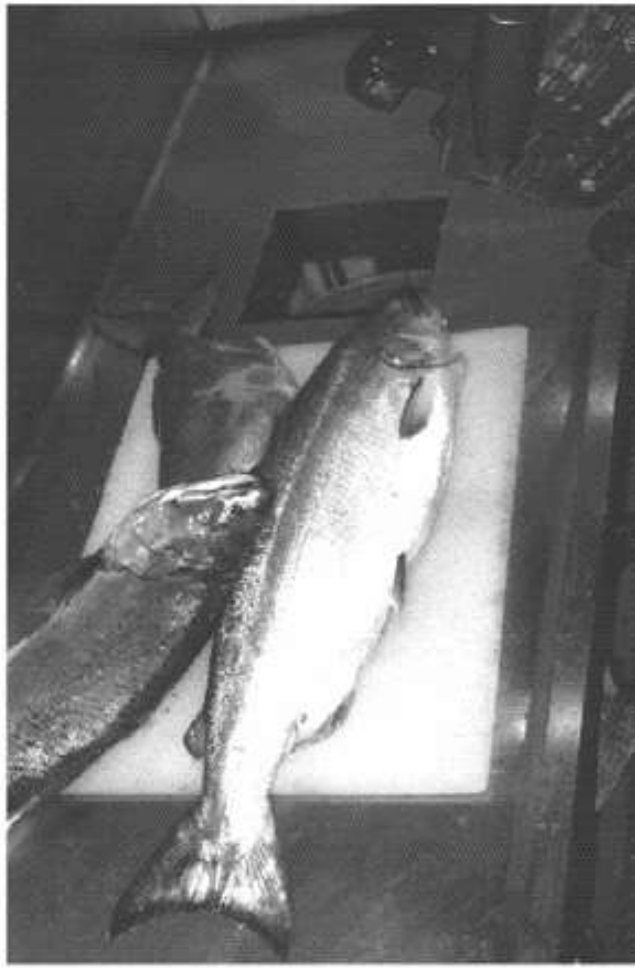
Classification: Definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class.
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

An Example

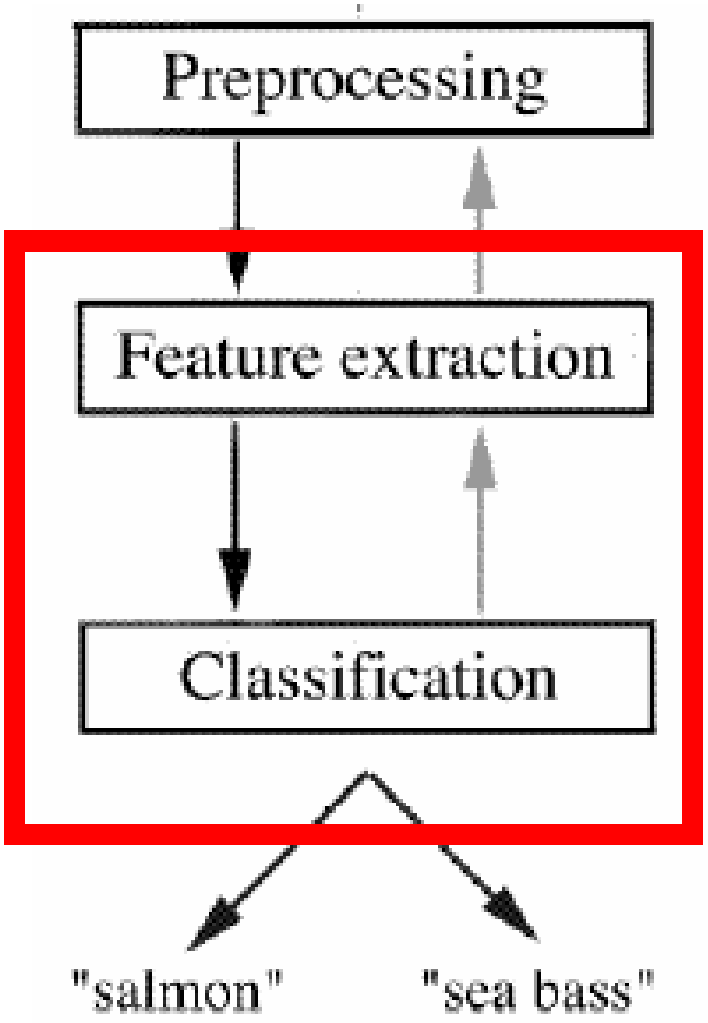
- (from Pattern Classification by Duda & Hart & Stork – Second Edition, 2001)
- A fish-packing plant wants to automate the process of sorting incoming fish according to species
- As a pilot project, it is decided to try to separate sea bass from salmon using optical sensing

An Example (continued)



- Features (to distinguish):
  - Length
  - Lightness of scales
  - Width
  - Position of mouth

An Example (continued)



- Preprocessing: Images of different fishes are isolated from one another and from background;
- Feature extraction: The information of a single fish is then sent to a feature extractor, that measure certain “features” or “properties”;
- Classification: The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species

An Example (continued)

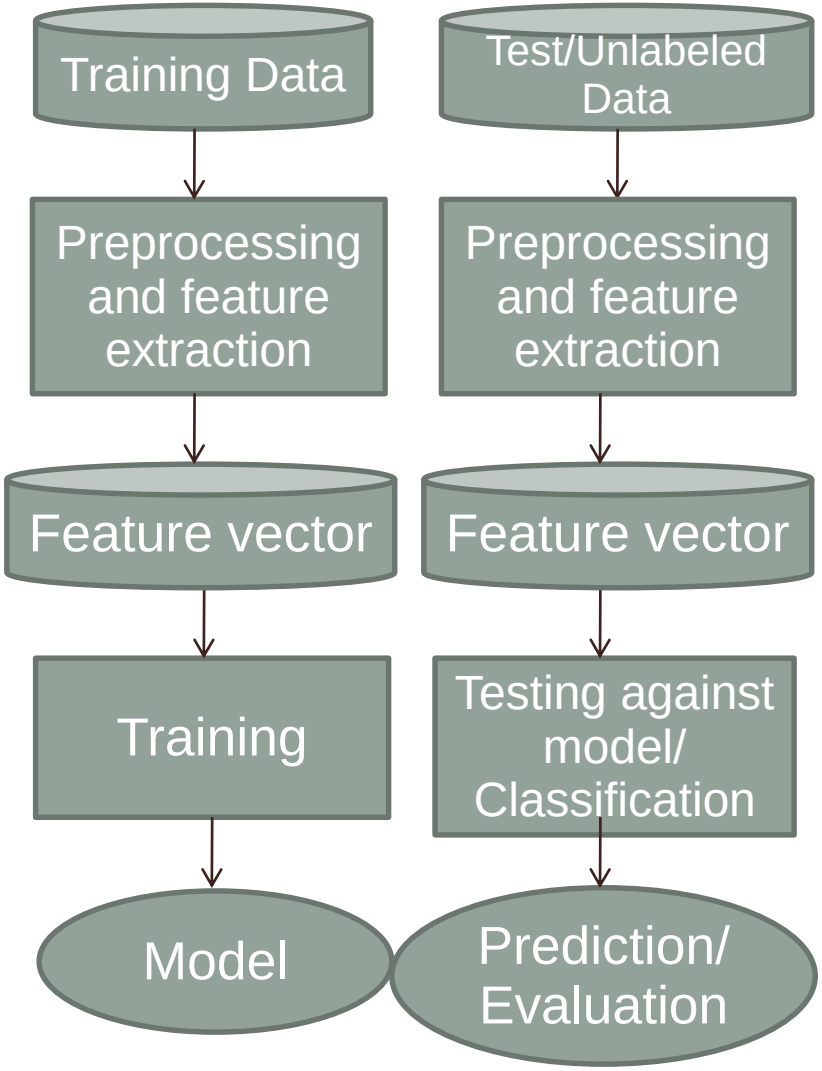
- Domain knowledge:
  - A sea bass is generally longer than a salmon
- Related feature: (or attribute)
  - Length
- Training the classifier:
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier learns itself from the training examples, how to distinguish Salmon from Bass based on the fish_length

An Example (continued)

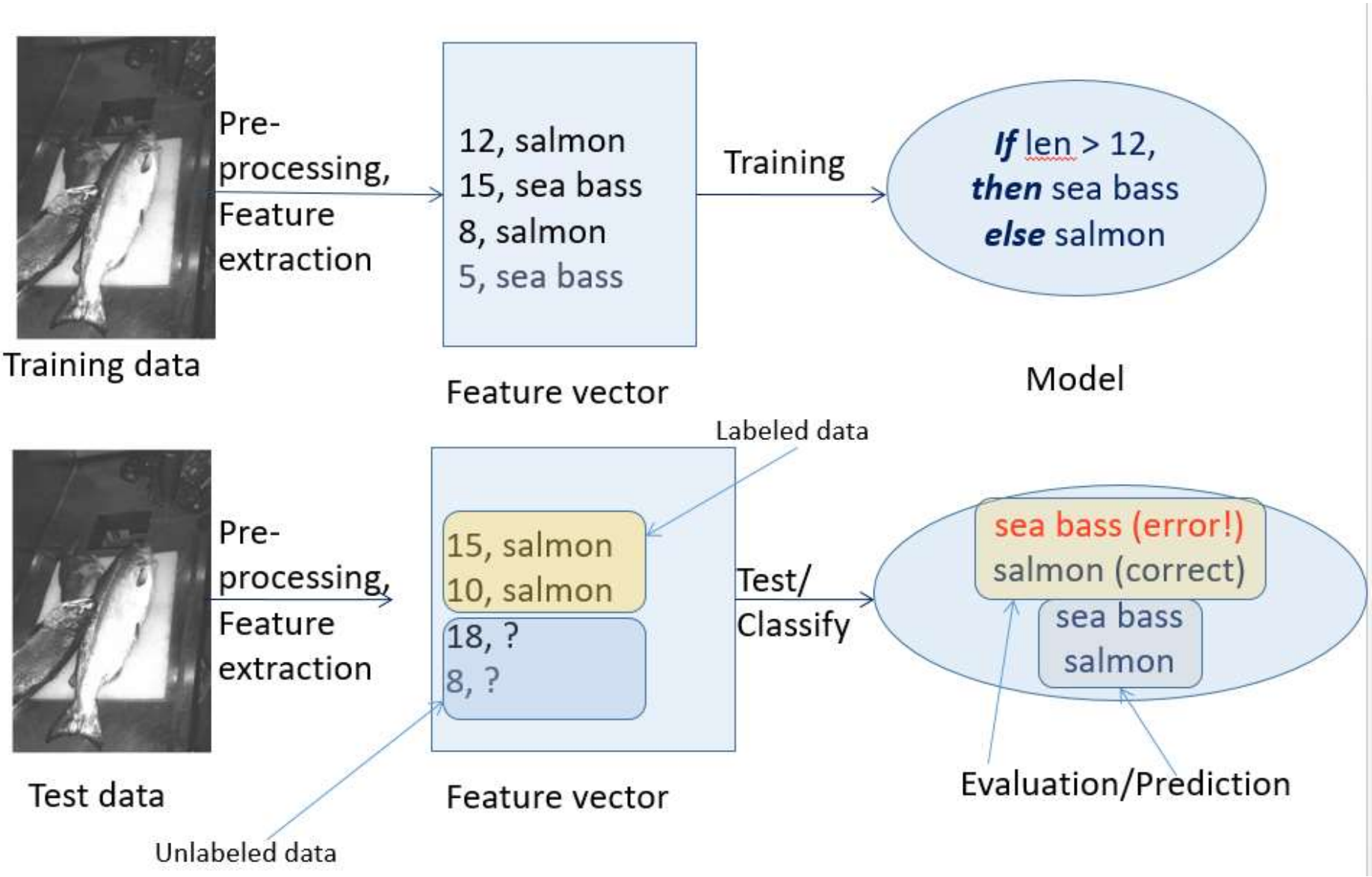
- Classification model (hypothesis):
  - The classifier generates a model from the training data to classify future examples (test examples)
  - An example of the model is a rule like this:
  - If Length >= l* then sea bass otherwise salmon
  - Here the value of l* determined by the classifier
- Testing the model
  - Once we get a model out of the classifier, we may use the classifier to test future examples
  - The test data is provided in the form <fish_length>
  - The classifier outputs <fish_type> by checking fish_length against the model

An Example (continued)

- So the overall classification process goes like this →



An Example (continued)

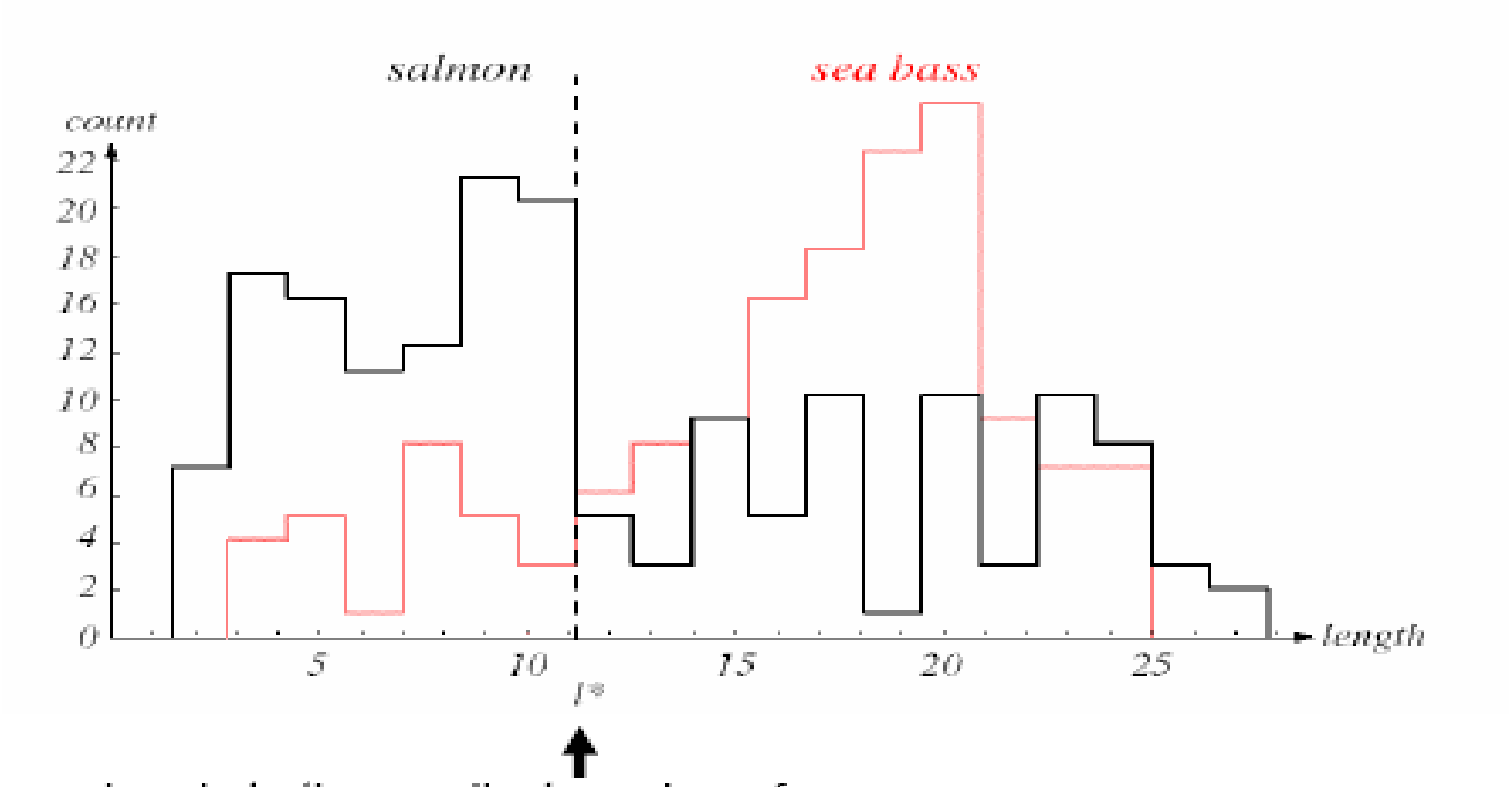


An Example (continued)

- Why error?
  - Insufficient training data
  - Too few features
  - Too many/irrelevant features
  - Overfitting / specialization

An Example (continued)

Histograms of the length feature for the two categories



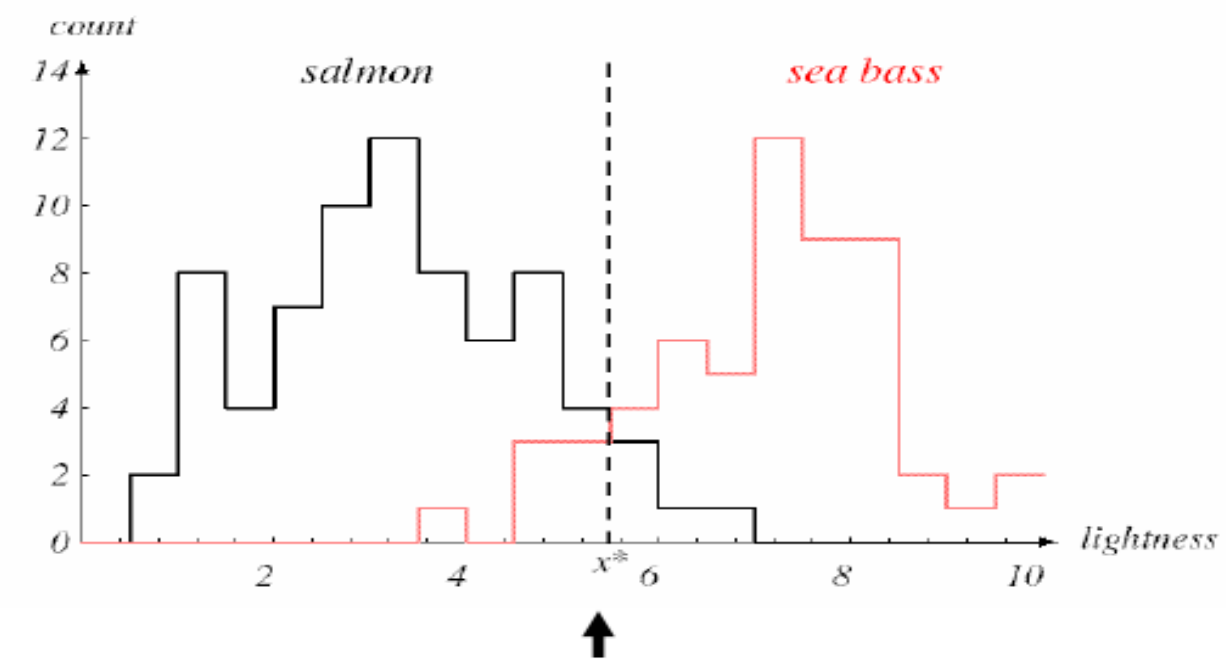
Leads to the smallest number of errors on average

We cannot reliably separate sea bass from salmon by length alone!

An Example (continued)

- New Feature:
  - Average lightness of the fish scales

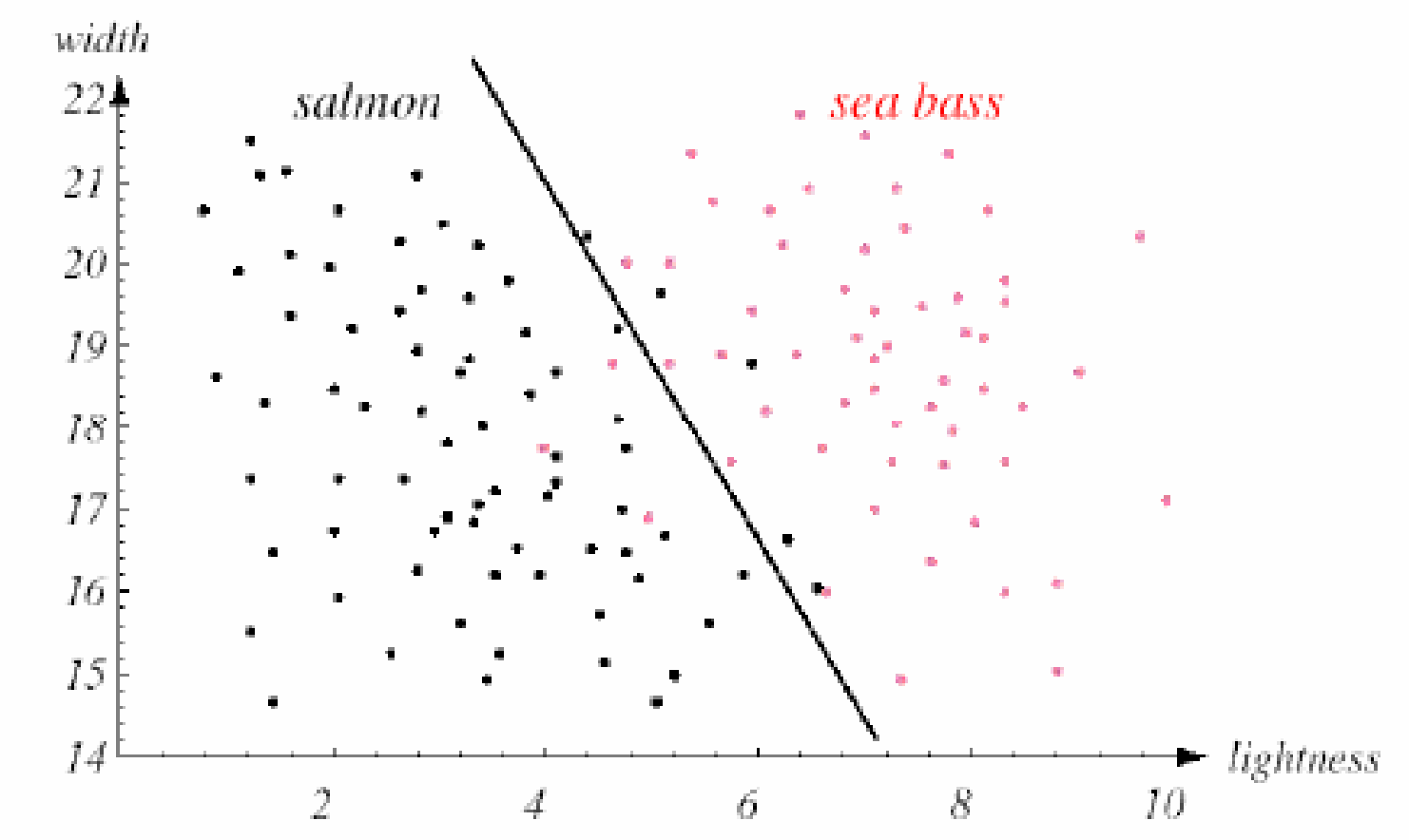
Histograms of the lightness feature for the two categories



Leads to the smallest number of errors on average

The two classes are much better separated!

An Example (continued)



Decision rule: Classify the fish as a sea bass if its feature vector falls above the decision boundary shown, and as salmon otherwise

An Example (continued)

Training data

Pre-processing, Feature extraction

12, 4, salmon
15, 8, sea bass
8, 2, salmon
5, 10, sea bass

Feature vector

Training

If ltns > 6 **or** len*5+ltns*2>100 ***then*** sea bass ***else*** salmon

Model

Test data

Pre-processing, Feature extraction

15, 2, salmon
10, 7, salmon
18, 7, ?
8, 5, ?

Feature vector

Test/ Classify

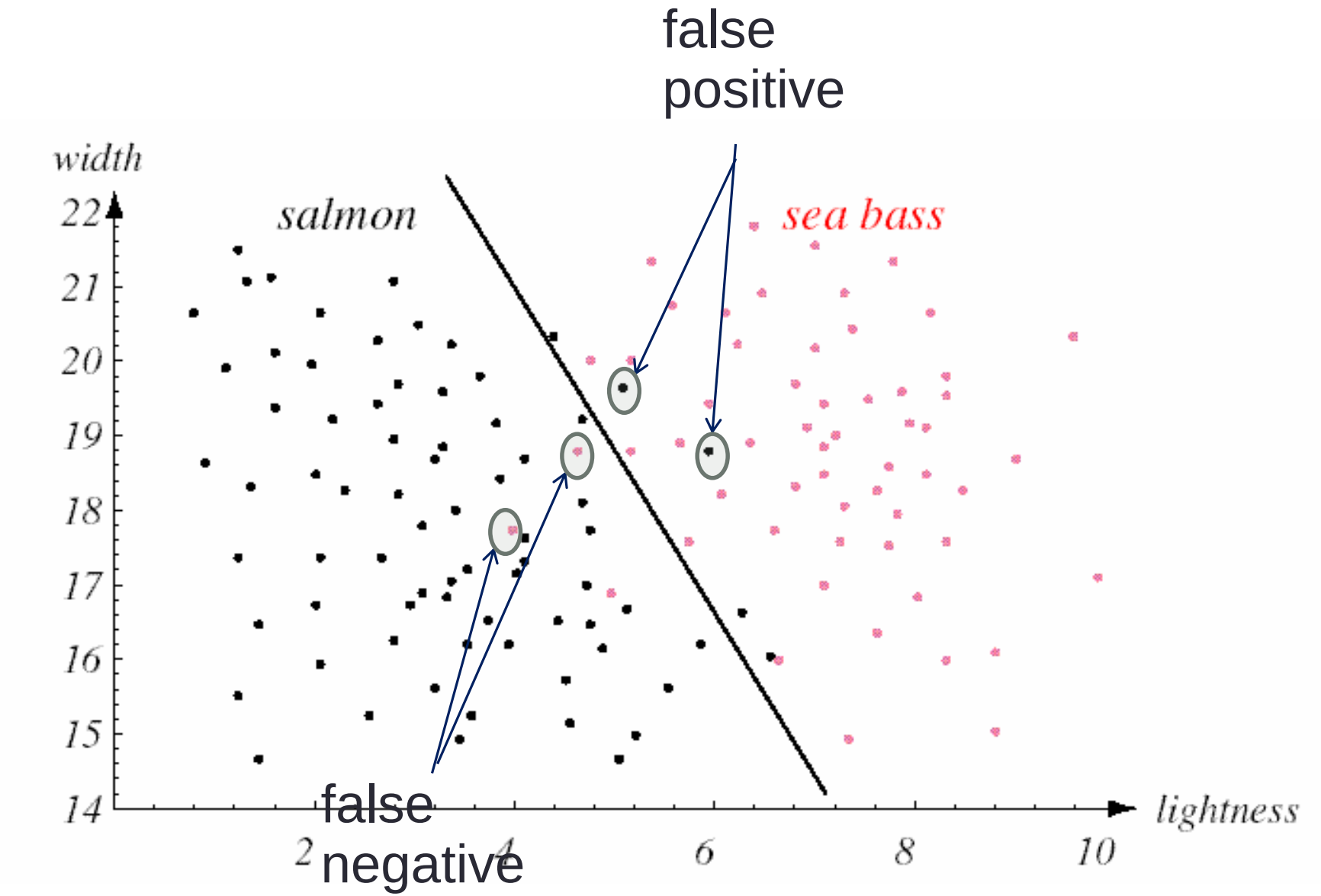
salmon (correct)
salmon (correct)
sea bass
salmon

Evaluation/Prediction

Validation

- Accuracy:
  - % of test data correctly classified
  - In our first example, accuracy was 3 out 4 = 75%
  - In our second example, accuracy was 4 out 4 = 100%
- False positive:
  - Negative class incorrectly classified as positive
  - Usually, the larger class is the negative class
  - Suppose
    - salmon is negative class
    - sea bass is positive class

Validation


false
positive
width
salmon
sea bass
22
21
20
19
18
17
16
15
14
2
4
6
8
10
lightness
false
negative

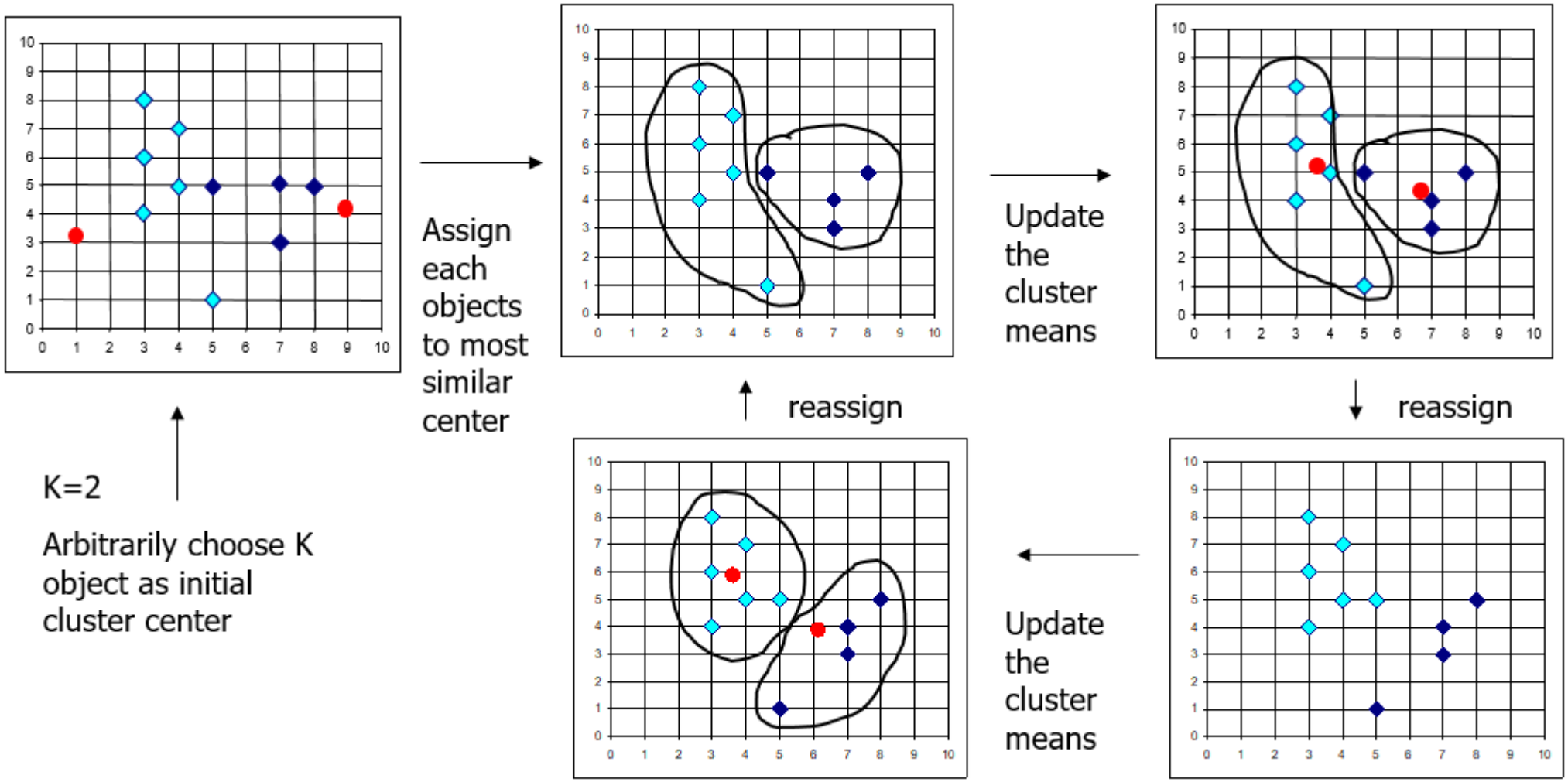
Sumário

- O que é Mineração de Dados?
- Funcionamento de Mineração de Dados
- Técnicas de Mineração de Dados
  - Associação
  - Classificação
  - Clusterização

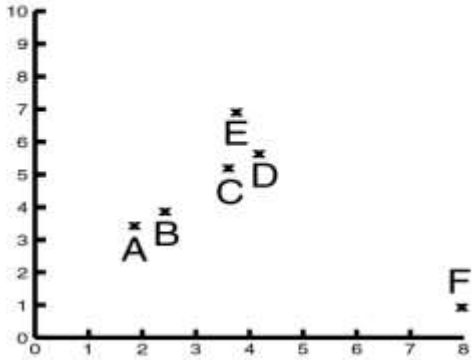
What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

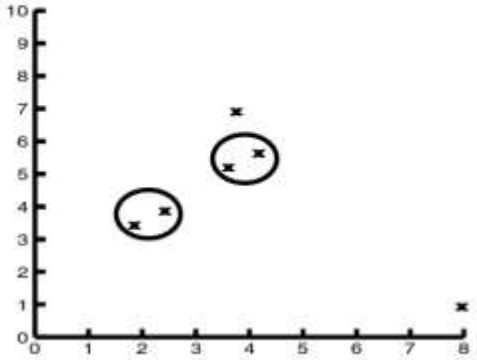
The *K-Means* Clustering Method



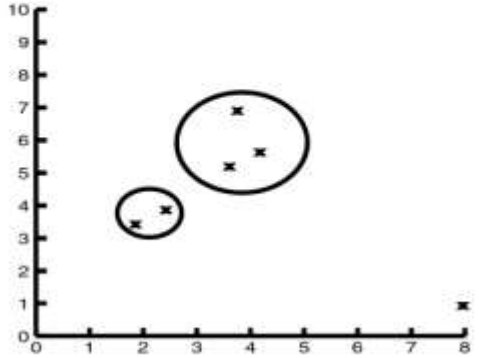
Problem: Cluster Number



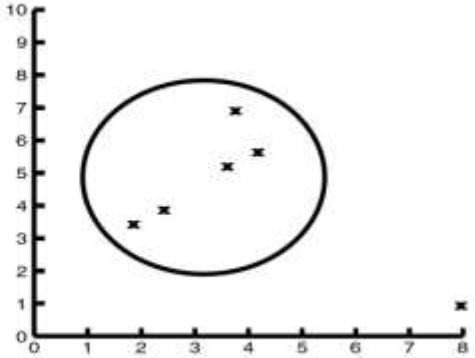
a) Six Clusters



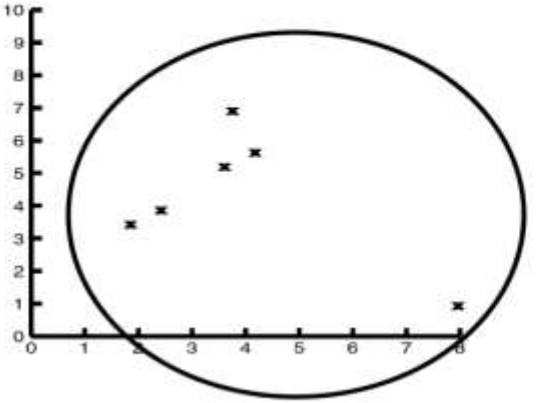
b) Four Clusters



c) Three Clusters



d) Two Clusters



e) One Cluster

OBRIGADO !

marcial@larces.uece.br
http://marcial.larces.uece.br