Social Big Data - Methods

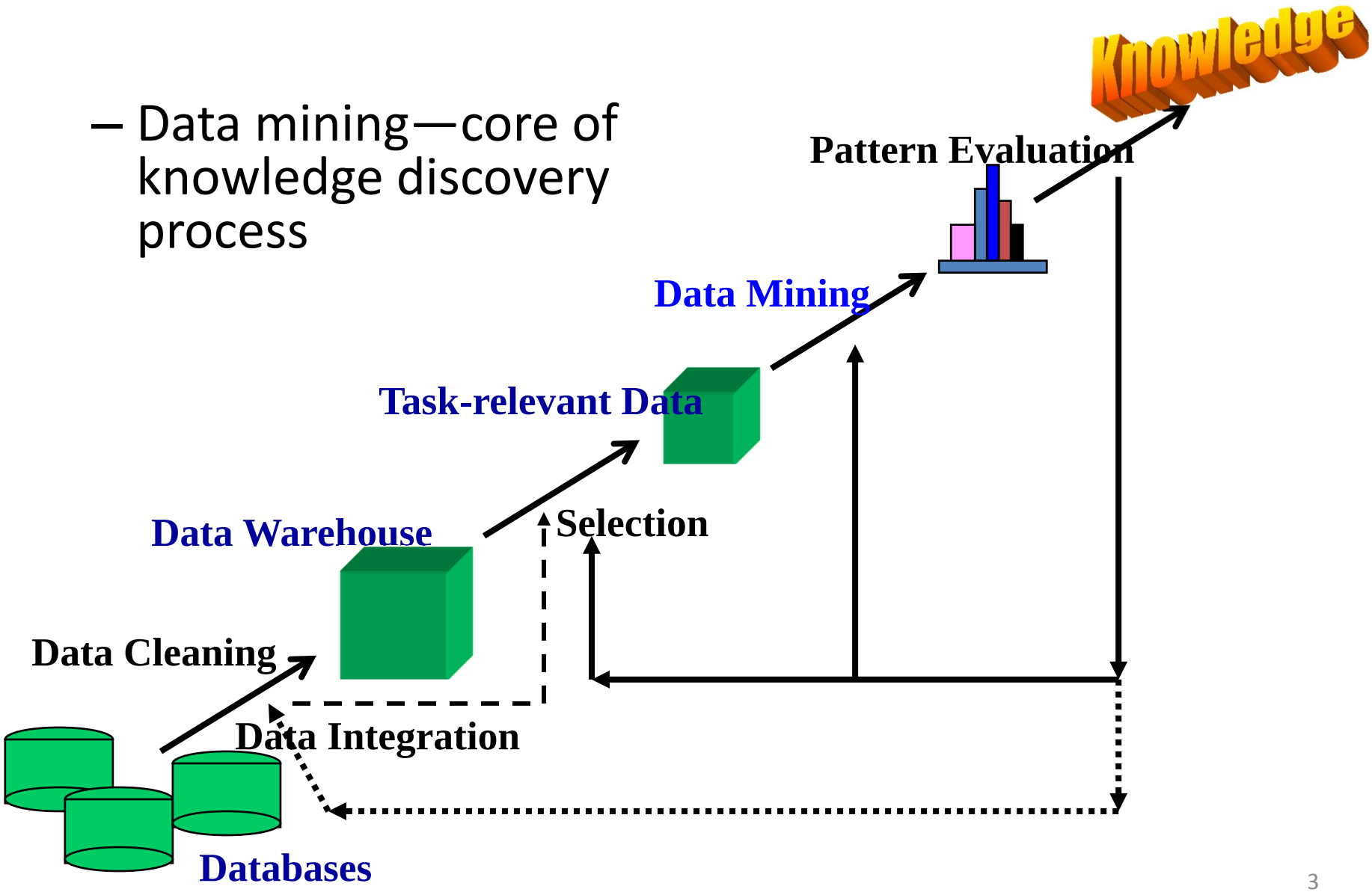
Dr. Hong Huang

Acknowledge

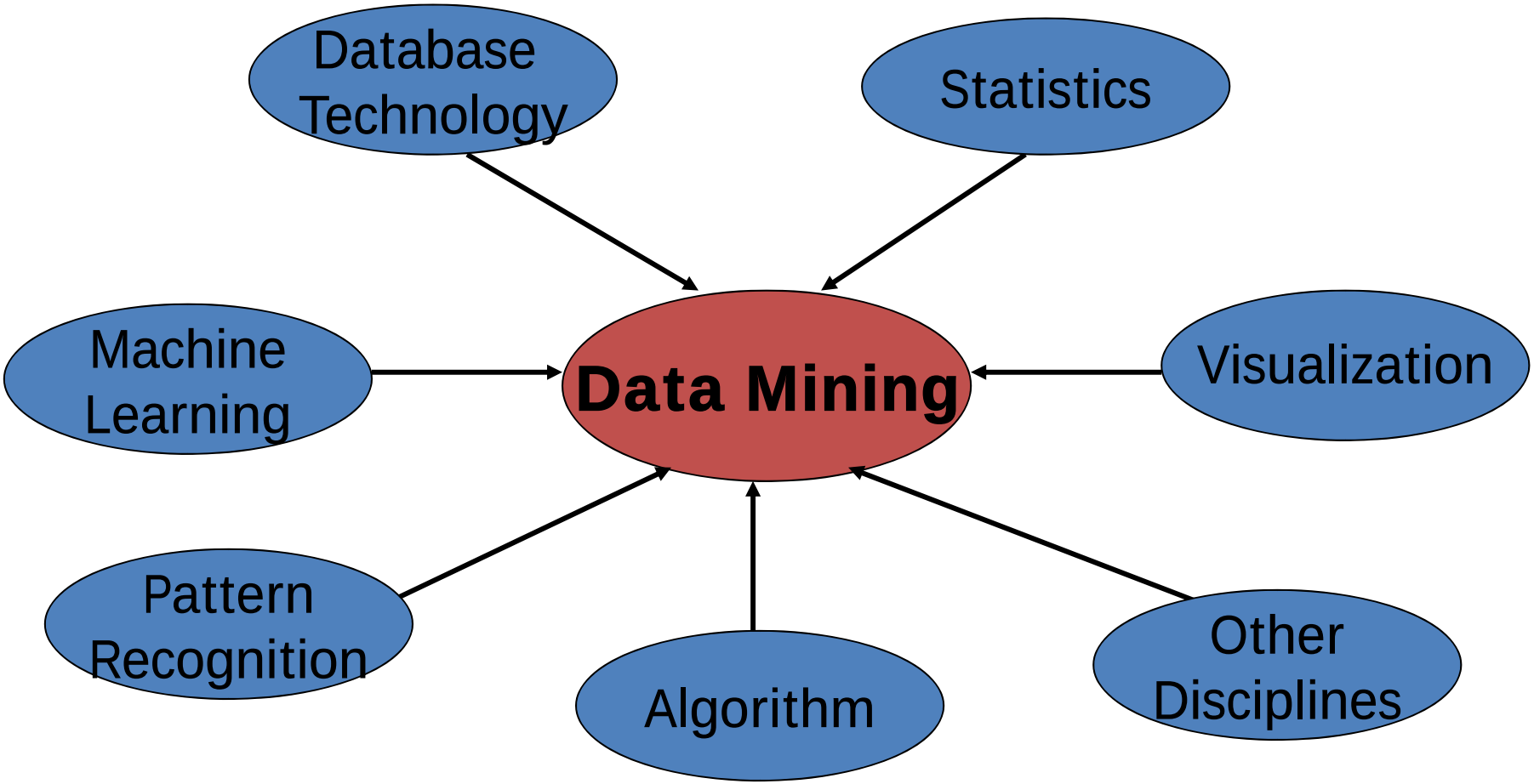
- Jiawei Han (http://www-sal.cs.uiuc.edu/~hanj/DM_Book.html)
- Vipin Kumar (http://www-users.cs.umn.edu/~kumar/csci5980/index.html)
- Ad Feelders (http://www.cs.uu.nl/docs/vakken/adm/)
- Zdravko Markov (http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-1.html)

Knowledge Discovery (KDD) Process

– Data mining—core of knowledge discovery process



Data Mining: Confluence of Multiple Disciplines



Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - ……
- New and sophisticated applications

Data Mining vs. Statistical Analysis

Statistical Analysis:

- Ill-suited for Nominal and Structured Data Types
- Completely data driven - incorporation of domain knowledge not possible
- Interpretation of results is difficult and daunting
- Requires expert user guidance

Data Mining:

- Large Data sets
- Efficiency of Algorithms is important
- Scalability of Algorithms is important
- Real World Data
- Lots of Missing Values
- Pre-existing data - not user generated
- Data not static - prone to updates
- Efficient methods for data retrieval available for use

Data Mining vs. DBMS

- DBMS (Database Management System)
- Example DBMS Reports
  - Last months sales for each service type
  - Sales per service grouped by customer sex or age bracket
  - List of customers who lapsed their policy

- Questions answered using Data Mining
  - What are the fundamental factors that trigger the users to pay?
  - How does users' paying behavior influence each other in the game social network?

Data Mining and Data Warehousing

- Data Warehouse: a centralized data repository which can be queried for business benefit.
- Data Warehousing makes it possible to
  - extract archived operational data
  - overcome inconsistencies between different legacy data formats
  - integrate data throughout an enterprise, regardless of location, format, or communication requirements
  - incorporate additional or expert information
- OLAP: On-line Analytical Processing

DBMS, OLAP, and Data Mining

	DBMS	OLAP	Data Mining
Task	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
Type of result	Information	Analysis	Insight and Prediction
Method	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
Example question	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

Multi-Dimensional View of Data Mining

- **Data to be mined**
- **Knowledge to be mined**
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

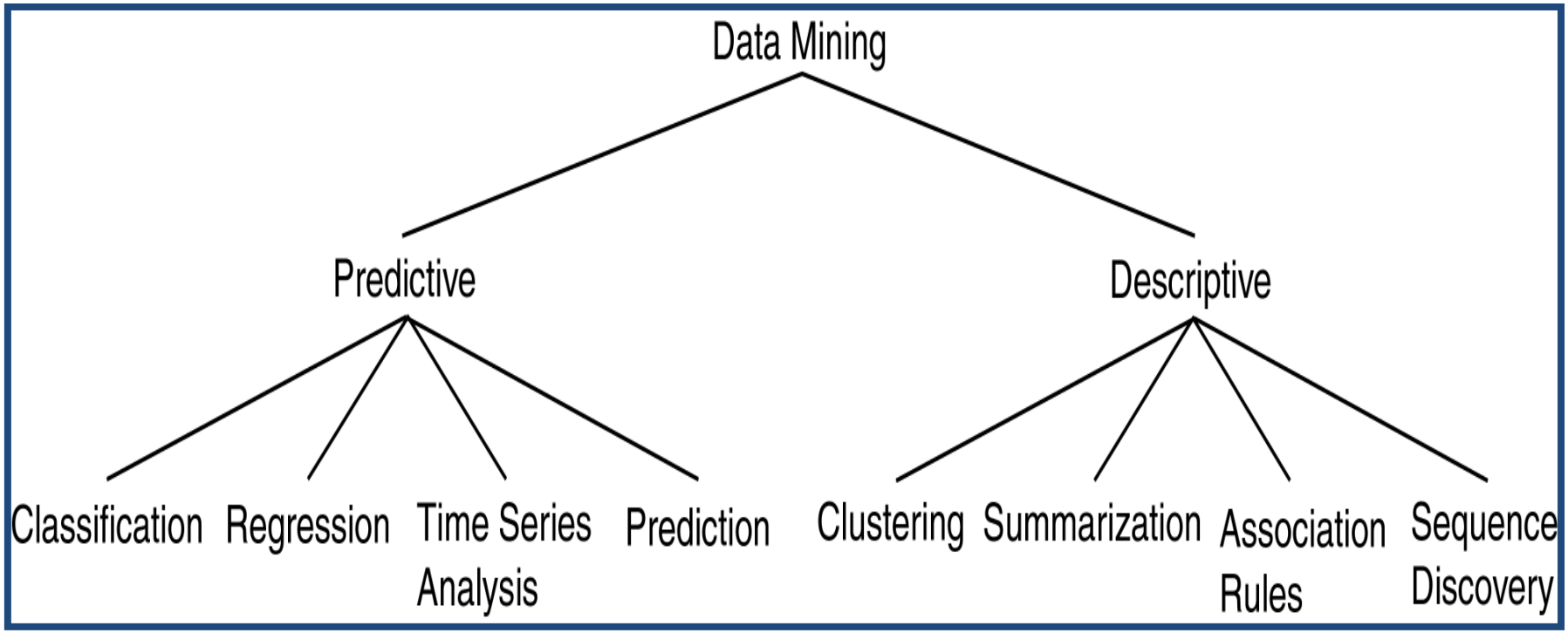
Data Mining Tasks

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables
- Description Tasks
  - Find human-interpretable patterns that describe the data.

Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Data Mining Models and Tasks



CLASSIFICATION

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.
- Predict some unknown or missing numerical values
  - 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.

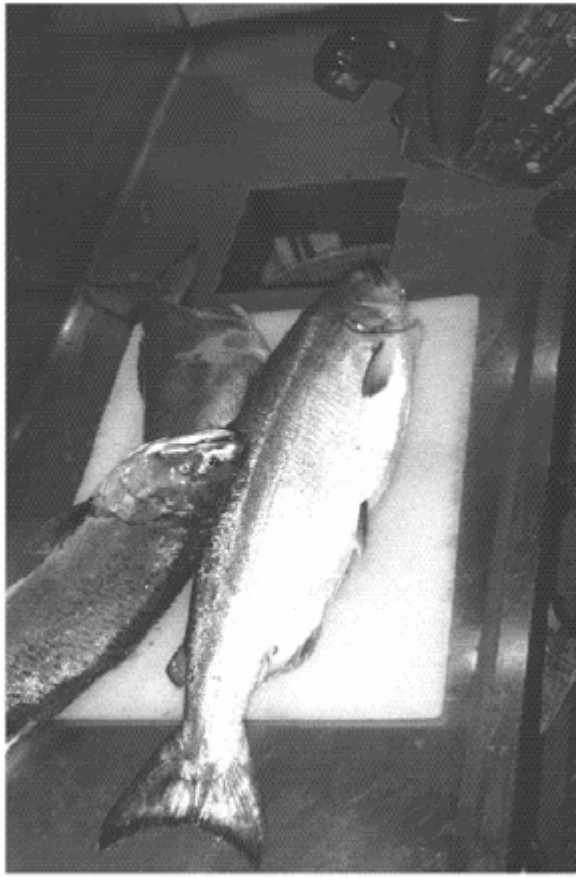
An Example

(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):

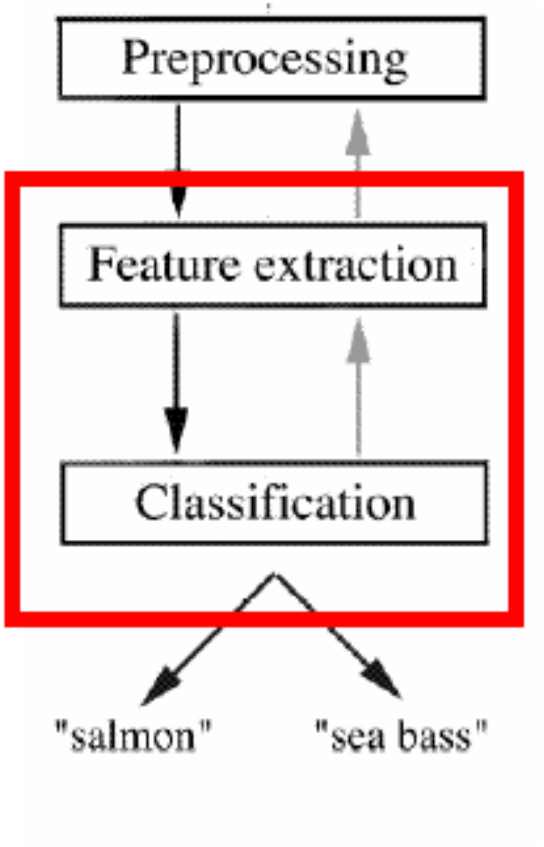
Length

Lightness

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)

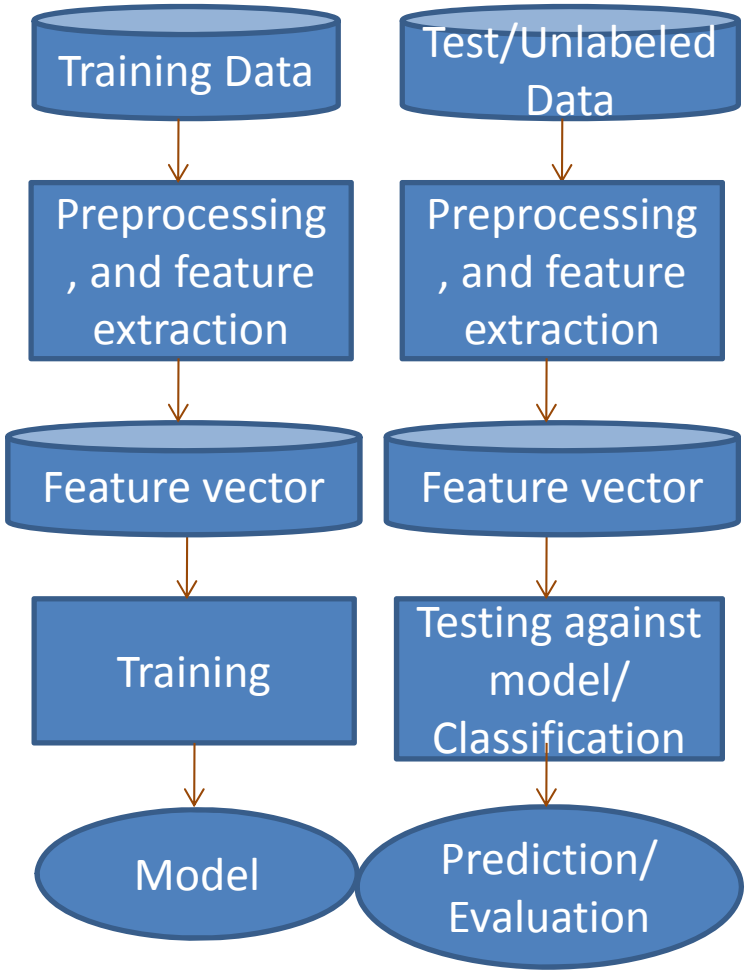
- Domain knowledge:
  - A sea bass is generally longer than a salmon
- Related feature*:* (or attribute)
  - Length
- Training the classifier:
  - Some examples are provided to the classifier in this form: <fish_length, fish_name>
  - These examples are called training examples
  - The classifier *learns*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

An Example (continued)

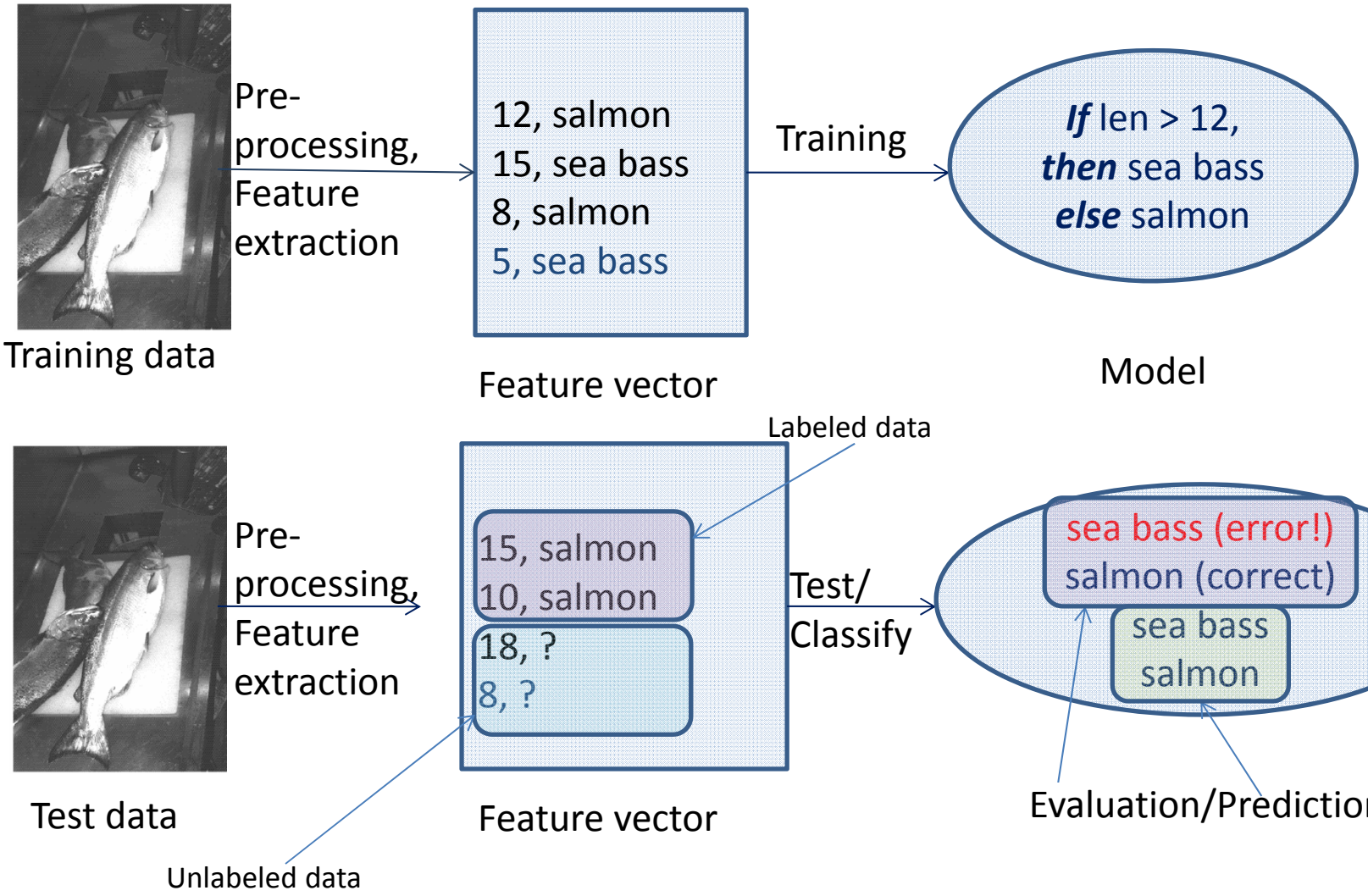
- Classification model (hypothesis):
  - The classifier generates a model from the training data to classify future examples (test examples)
  - An example of the model is a rule like this:
  - If *Length >= l* then sea bass* otherwise *salmon*
  - Here the value of *l** determined by the classifier
- 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)

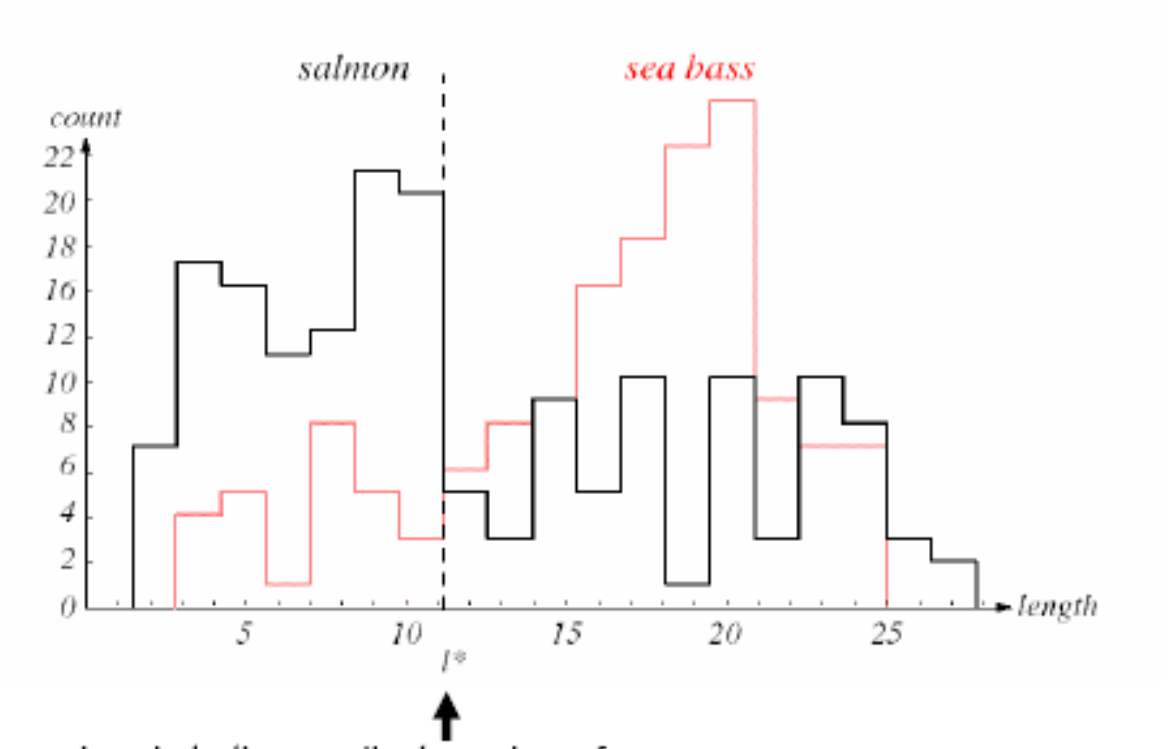


An Example (continued)

- 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



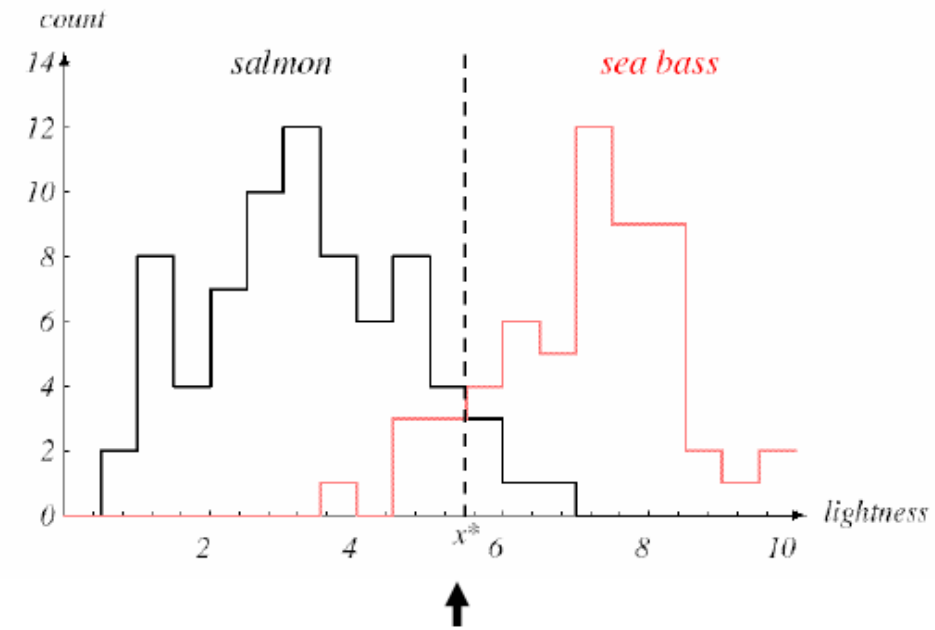
Leads to the smallest number of errors on average

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

An Example (continued)

- New Feature:
  - *Average lightness of the fish scales*

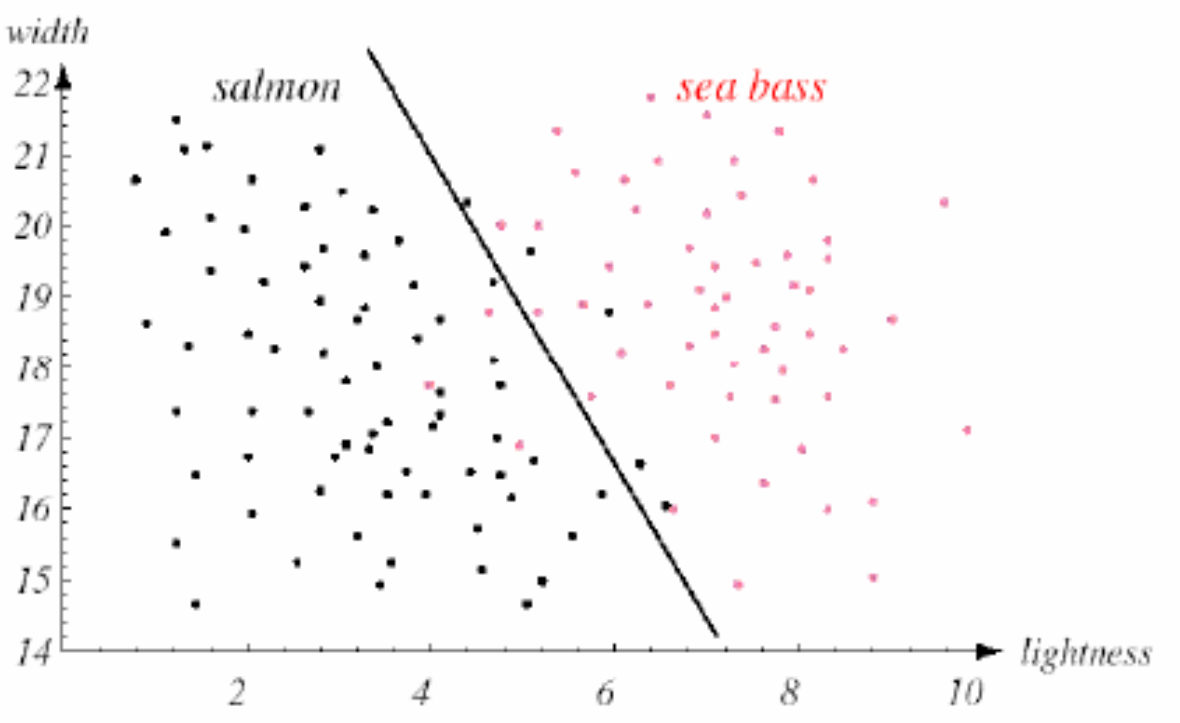
Histograms of the lightness feature for the two categories



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

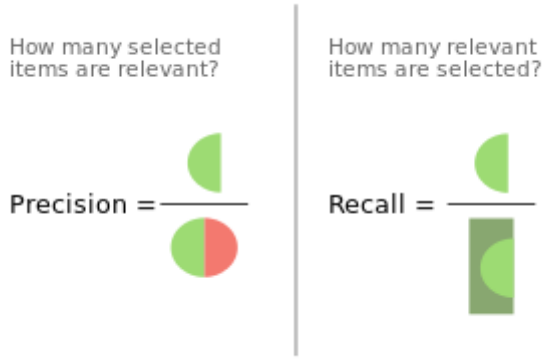
Evaluation/Prediction

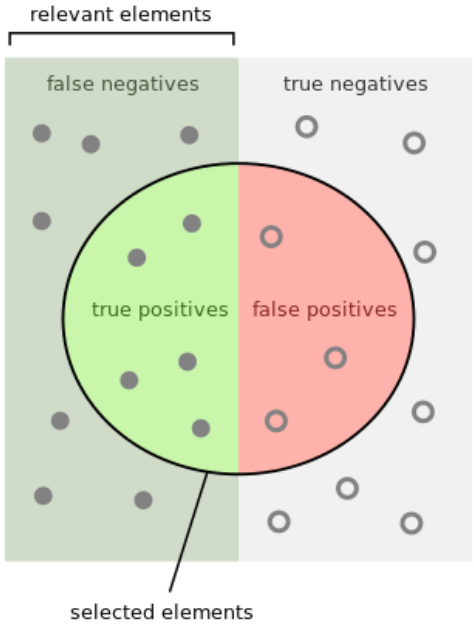
Terms

- 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**

Terms

- Precision vs. Recall



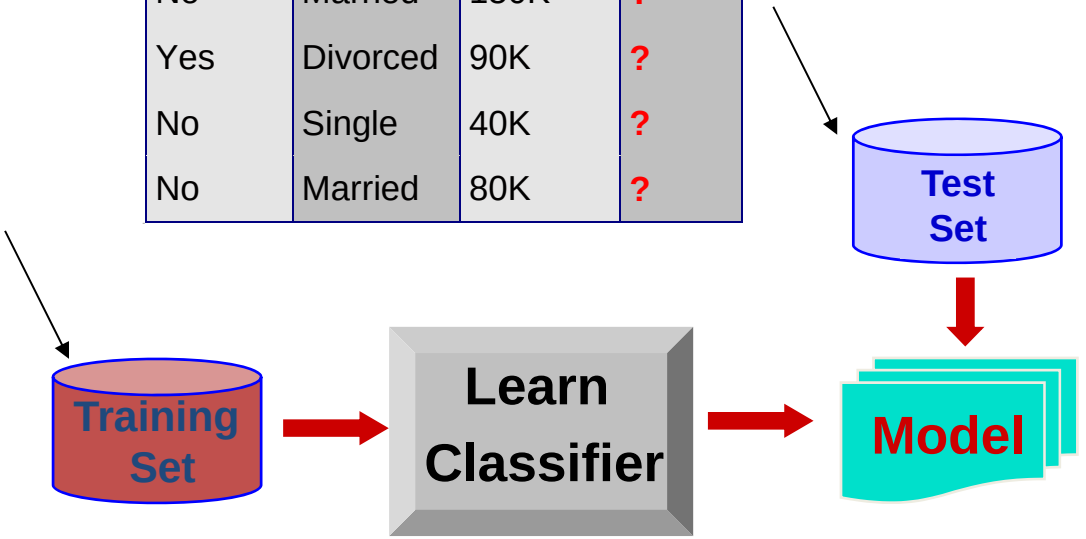


- Suppose a computer program for recognizing dogs in photographs identifies eight dogs in a picture containing 12 dogs and some cats. Of the 8 dogs identified, 5 actually are dogs (true positives), while the rest are cats (false positives). The program's precision is 5/8 while its recall is 5/12.

Classification Example 2

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification: Application 3

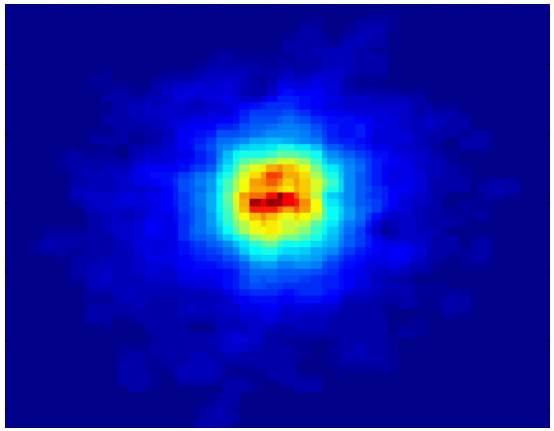
- Customer Attrition/Churn:
  - Goal: To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

Classification: Application 4

- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

Classifying Galaxies

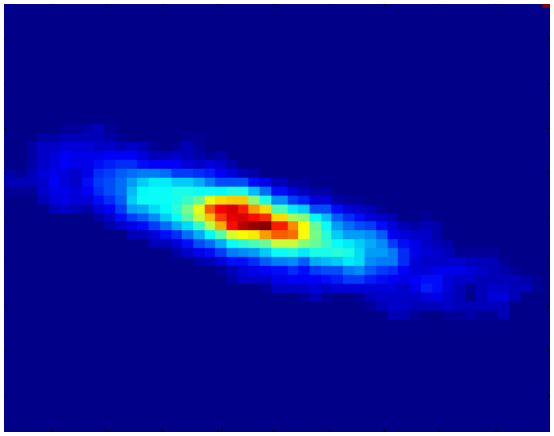
Early



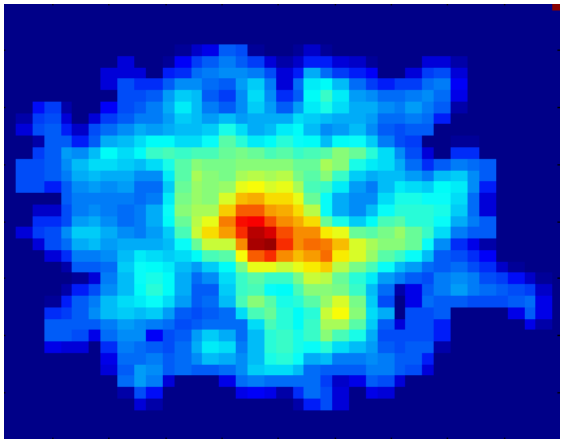
Class:
- **Stages of Formation**

Attributes:
- **Image features,**
- **Characteristics of light waves received, etc.**

Intermediate



Late



Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

CLUSTERING

Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Distance Measures

- Measure dissimilarity between objects

Euclidean: $dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2}$

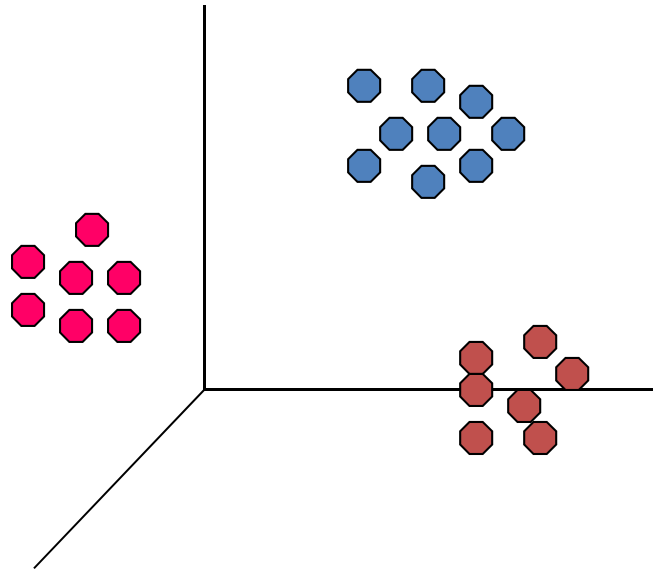
Manhattan: $dis(t_i, t_j) = \sum_{h=1}^{k} \mid (t_{ih} - t_{jh}) \mid$

Illustrating Clustering

☒Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

ASSOCIATION RULE MINING

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: Application 1

- 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: Application 2

- 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:
      - Support: an indication of how frequently the itemset appears in the dataset.
      - Confidence: an indication of how often the rule has been found to be true.

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

The Sad Truth About Diapers and Beer

- So, don’t be surprised if you find six-packs stacked next to diapers!

Sequential Pattern Discovery: Definition

Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong sequential dependencies among different events:

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
    (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Regression

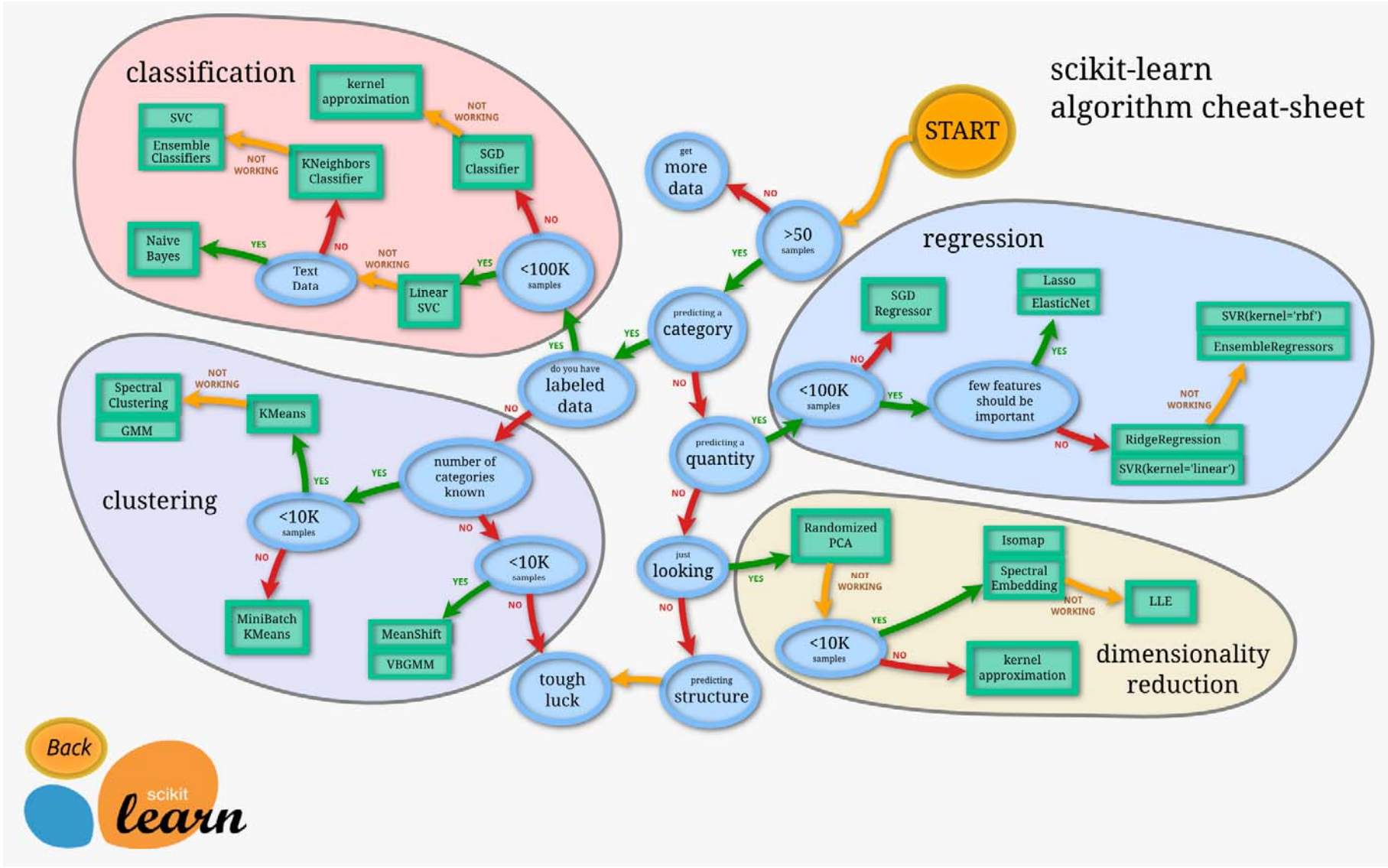
- Predict a value of a given **continuous** valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection

Data mining algorithms



Hypothesis-Based vs. Exploratory-Based

- The hypothesis-based method:
  - Formulate a hypothesis of interest.
  - Design an experiment that will yield data to test this hypothesis.
  - Accept or reject hypothesis depending on the outcome.

- Exploratory-based method:
  - Try to make sense of a bunch of data without an a priori hypothesis!
  - The only prevention against false results is significance:
    - ensure statistical significance (using train and test etc.)
    - ensure domain significance (i.e., make sure that the results make sense to a domain expert)

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
- **Interestingness measures**
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

Are All the "Discovered" Patterns Interesting?

- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Do we need to find all of the interesting patterns?
  - Heuristic vs. exhaustive search
- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization