Data Mining & Machine Learning

CS37300
Purdue University

Aug 23, 2017

The data mining process

Data

Selection

Target data

Preprocessing

Processed data

Knowledge

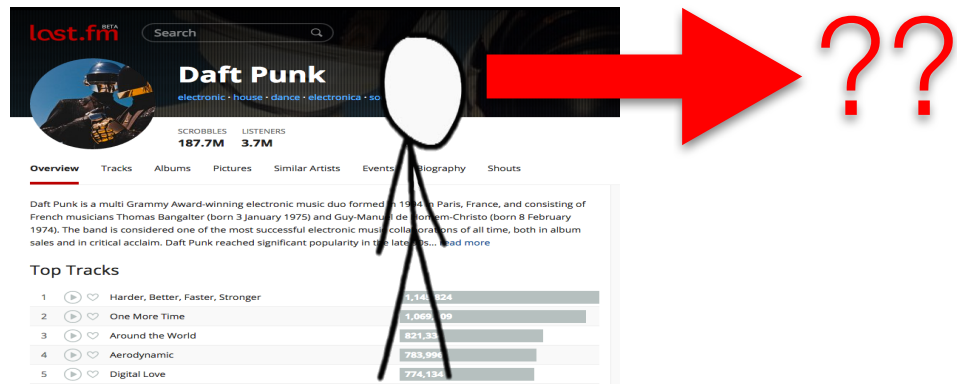
Interpretation evaluation

Patterns

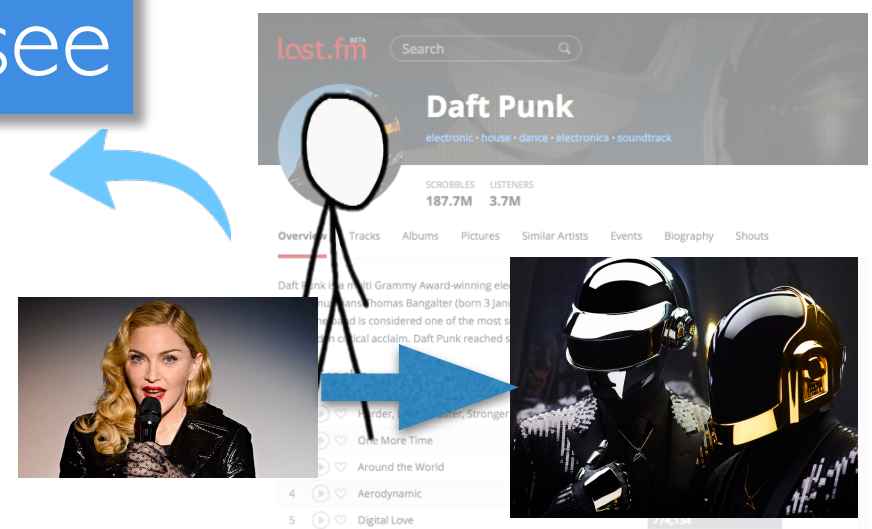
Mining

Real-world example:

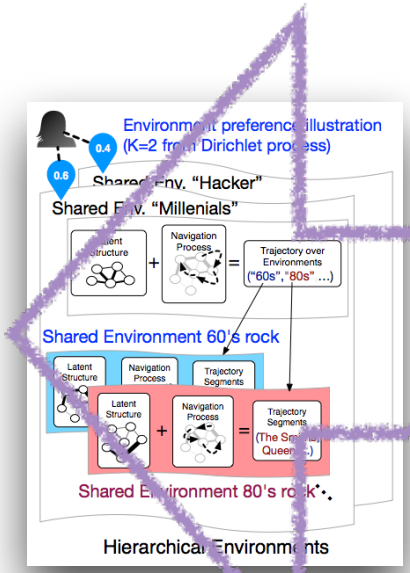
Which artist user listens next?

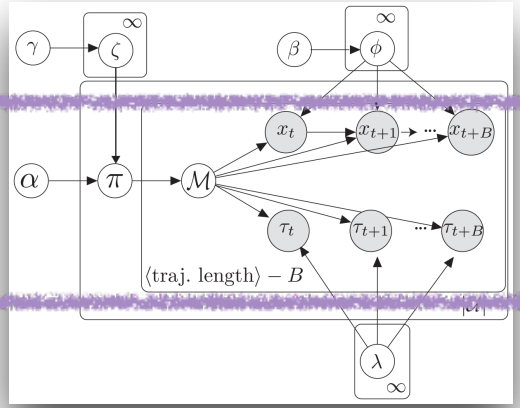


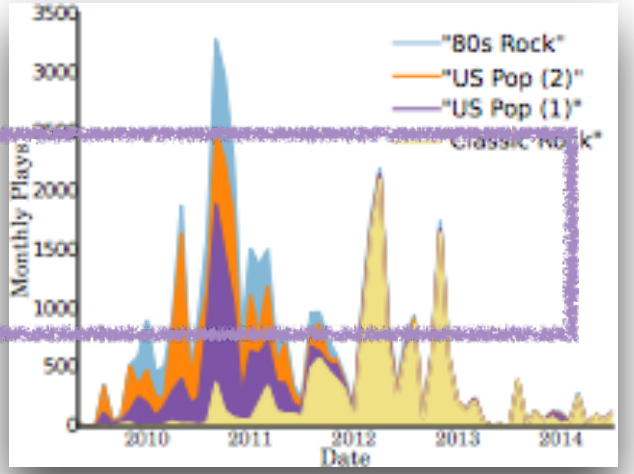
User trajectory is all we see



Bayesian Modeling of User Trajectories







Learned Model

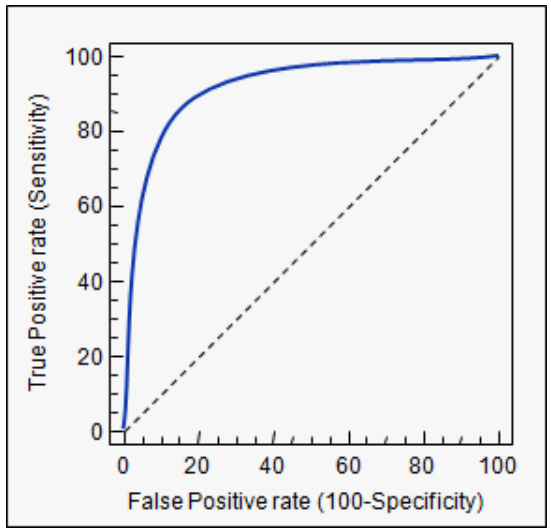
Statistical Model

Log

Some Model Used for Predictions and Descriptions

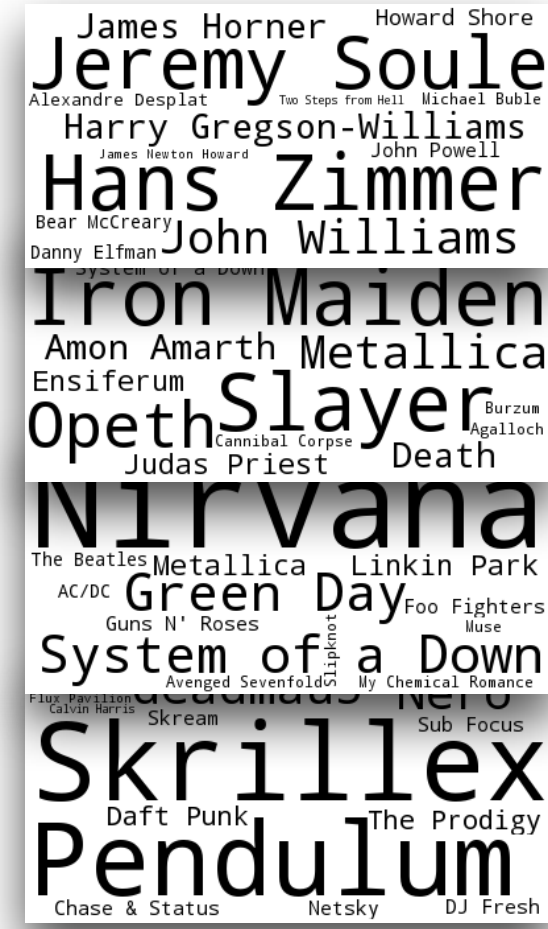
Predictive

- Prediction scores (how accurate our predictions are)



Descriptive

- Describes data according to a model
- Artist clusters:



Movie Soundtracks

Trash Metal

90's Rock

Electro House

The data mining process

Data

User listening online music (log)

Selection

Target data

Extract which artists each user listens to, Only keep users with more than 5 artists

Preprocessing

Processed data

Divide data into training, "testing", Only keep artists with more than 10 users

Mining

Patterns

Interpretation evaluation

Knowledge

Learn Bayesian model, output model parameters and artists clusters

Evaluate objectively on "test" data, subjectively looking at artists clusters

Elements of Data Mining
& Machine Learning Algorithms

Overview

- Task specification
- Data representation
- Knowledge representation
- Learning technique
    - Search + scoring
- Prediction and/or interpretation

Overview

- **Task specification**
- Data representation
- Knowledge representation
- Learning technique
    - Search + scoring
- Prediction and/or interpretation

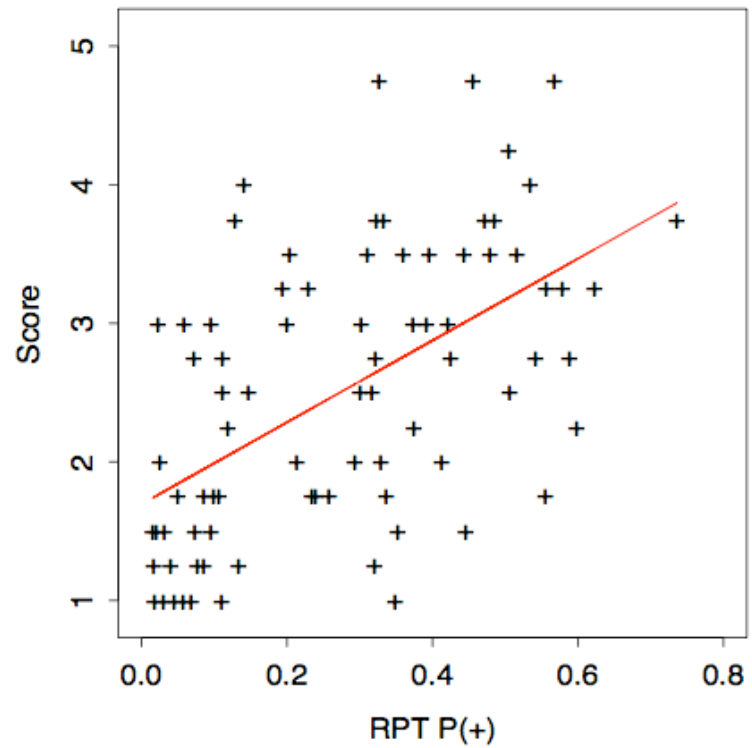
Task specification

- *Objective of the person who is analyzing the data*
- *Description of the characteristics of the analysis and desired result*

- Examples:
  - From a set of *labeled examples*, devise an *understandable model* that will *accurately predict* whether a stockbroker will commit fraud in the near future.
  - From a set of *unlabeled examples*, cluster stockbrokers into a *set of homogeneous groups* based on their demographic information

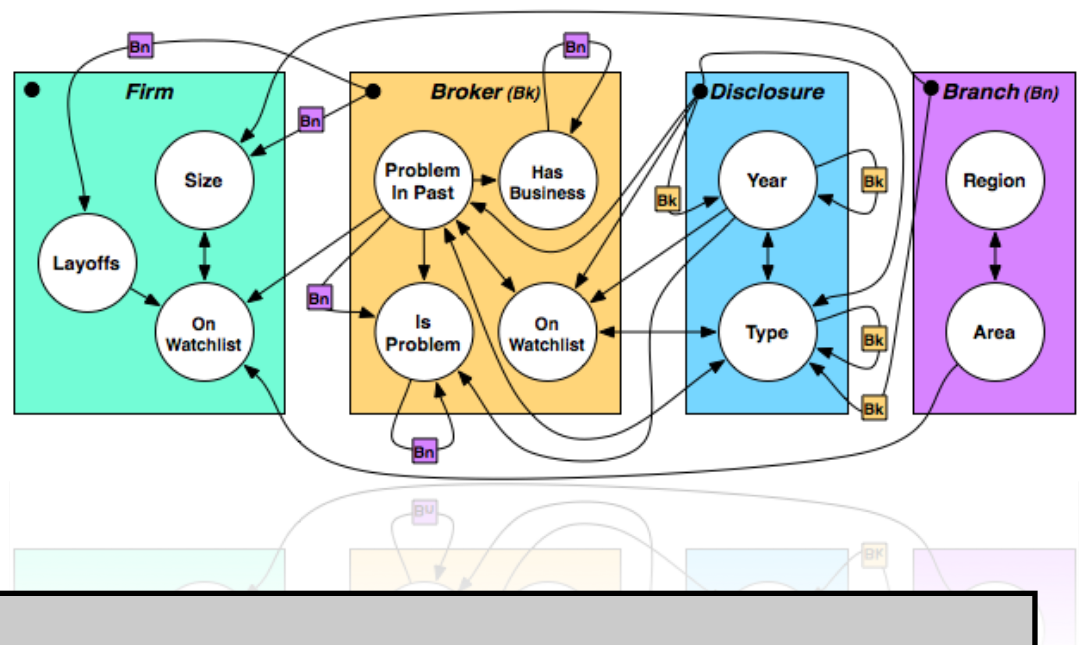
Exploratory data analysis

- Goal
  - Interact with data without clear objective
- Techniques
  - Visualization, adhoc modeling



Descriptive modeling

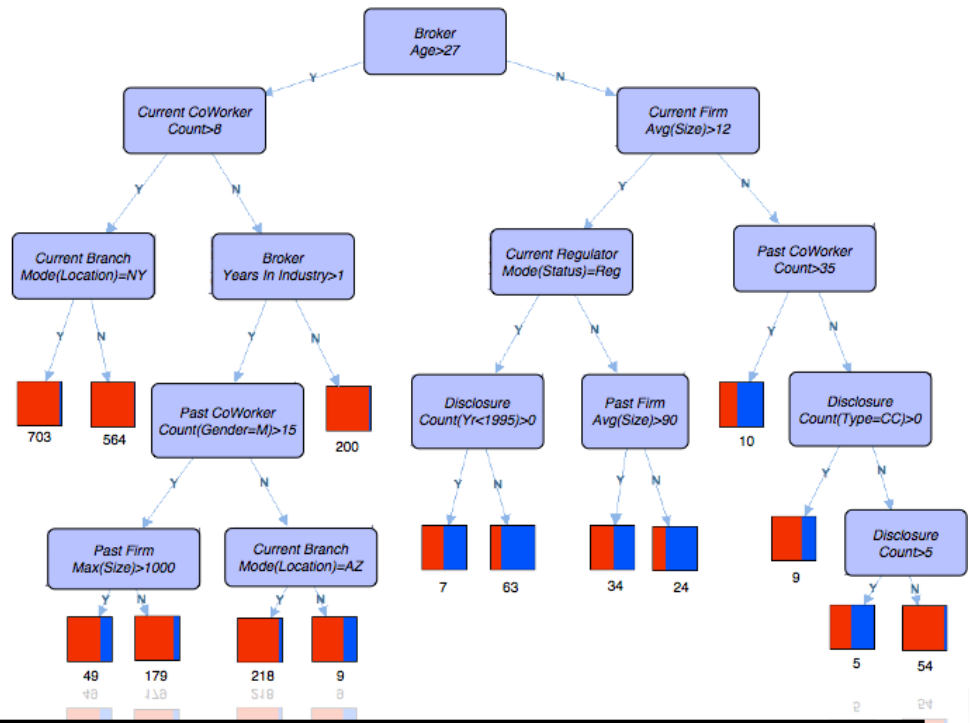
- Goal
  - Summarize the data or the underlying generative process
- Techniques
  - Density estimation, cluster analysis and segmentation



Also known as: **unsupervised** learning

Predictive modeling

- Goal
  - Learn model to predict unknown class label values given observed attribute values
- Techniques
  - Classification, regression



Also known as: **supervised** learning

Overview

- Task specification
- **Data representation**
- Knowledge representation
- Learning technique
    - Search + scoring
- Prediction and/or interpretation

Data representation

- *Choice of **data structure** for representing individual and collections of measurements*

- Individual measurements: single observations (e.g., person's date of birth, product price)
- Collections of measurements: sets of observations that describe an instance (e.g., person, product)
- Choice of representation determines applicability of algorithms and can impact modeling effectiveness
- Additional issues: data sampling, data cleaning, feature construction

Individual measurements

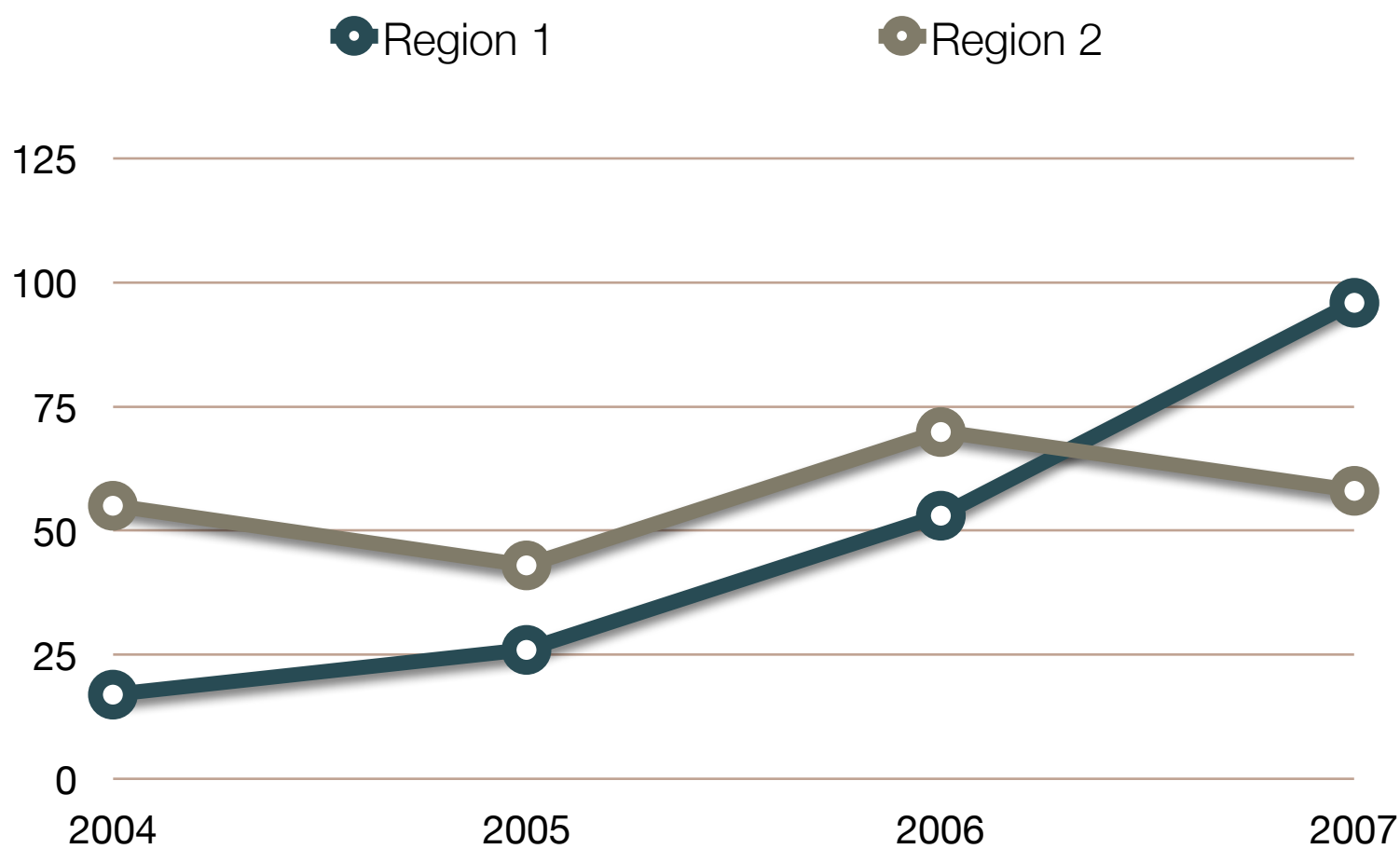
- Unit measurements:
  - Discrete values — categorical or ordinal variables
  - Continuous values — interval and ratio variables
- Compound measurements:
  - < x, y >
  - < value, time >

Data representation: Table/vectors

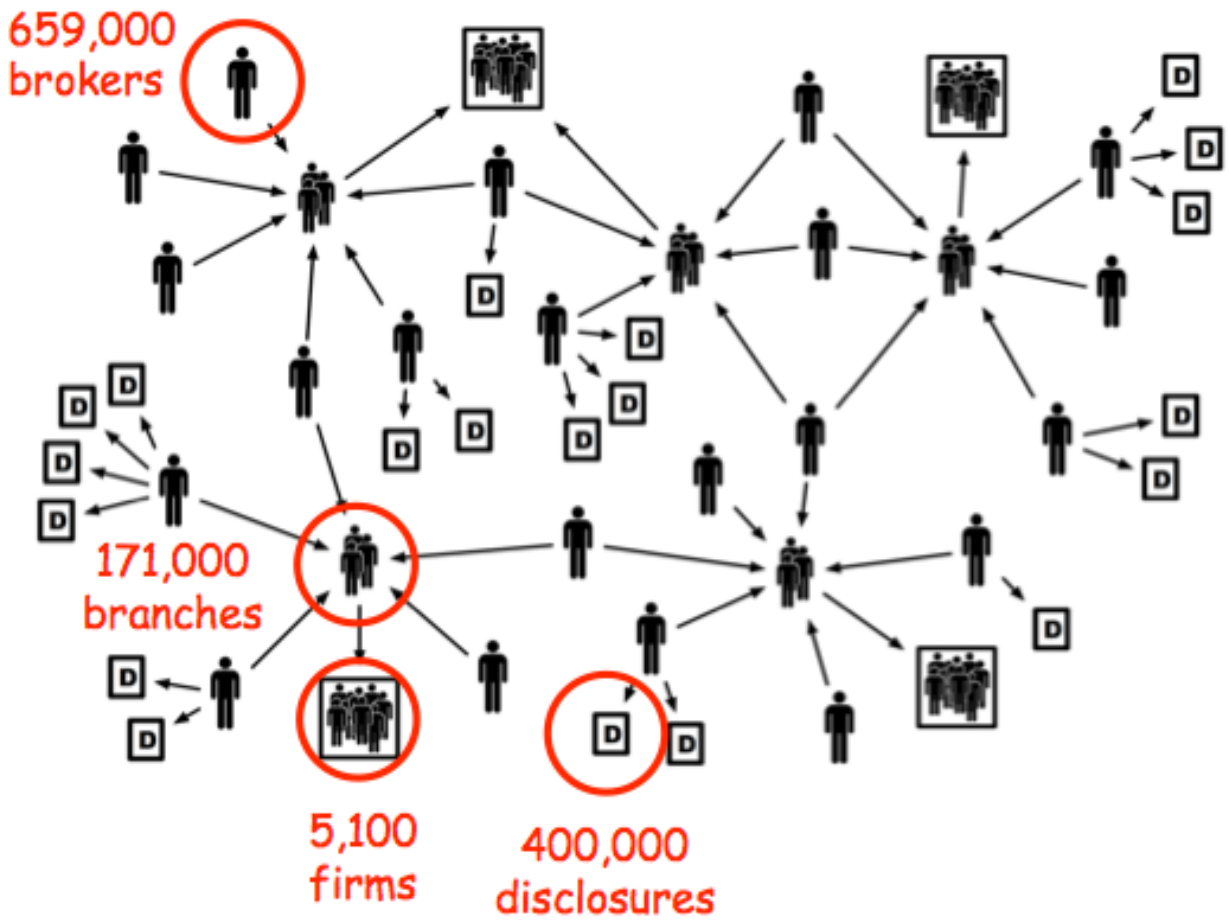
Fraud	Age	Degree	StartYr	Series7
+	22	Y	2005	N
-	25	N	2003	Y
-	31	Y	1995	Y
-	27	Y	1999	Y
+	24	N	2006	N
-	29	N	2003	N

N instances X *p* attributes

Data representation: Time series/sequences



Data representation: Relational/graph data



Overview

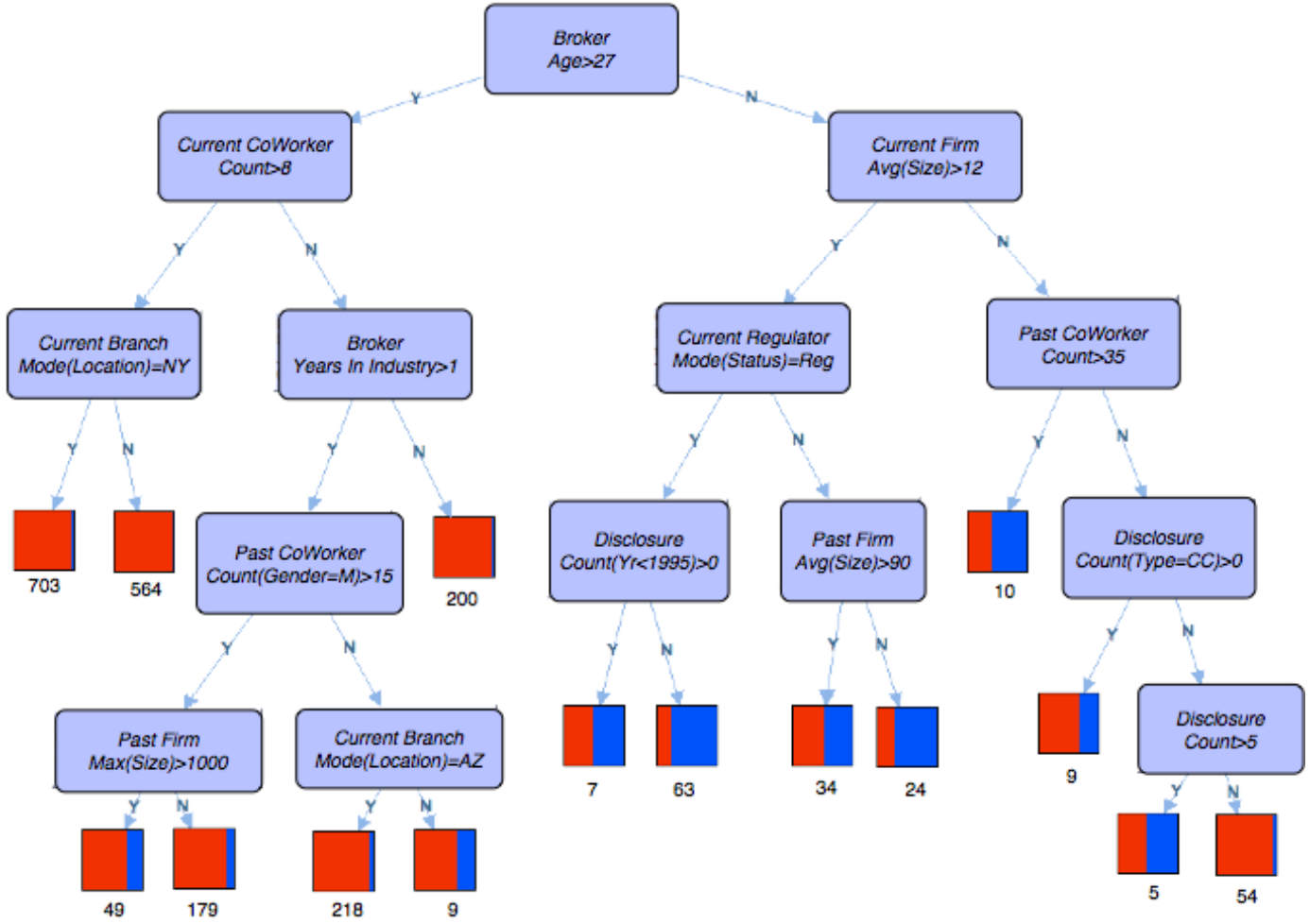
- Task specification
- Data representation
- **Knowledge representation**
- Learning technique
  - Search + scoring
- Prediction and/or interpretation

Knowledge representation

- *Underlying structure of the model or patterns that we seek from the data*
  - Specifies the models/patterns that could be returned as the results of the data mining algorithm
  - Defines the **model space** that algorithms search over (i.e., all possible models/patterns)

- Examples:
  - **If-then rule**
    If short closed car then toxic chemicals
  - **Conditional probability distribution**
    P(fraud | age, degree, series7, startYr)
  - **Decision tree**

Knowledge representation: Classification tree



Each node corresponds to a feature; each leaf a class label or probability distribution

Knowledge representation: Regression model

$$y = \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_0$$

- X are predictor variables
- Y is response variable
- Example:
  - Predict number of disclosures given income and trading history

Overview

- Task specification
- Data representation
- Knowledge representation
- **Learning technique**
  - Search + scoring
- Prediction and/or interpretation

Learning technique

- Method to construct model or patterns from data
- **Model space**
  - Choice of knowledge representation defines a set of possible models or patterns
- **Scoring function**
  - Associates a numerical value (score) with each member of the set of models/patterns
- **Search technique**
  - Defines a method for generating members of the set of models/patterns and determining their score

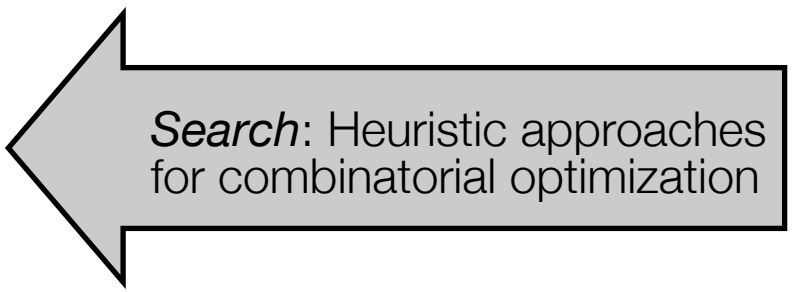
Scoring function

- *A numeric score assigned to each possible model in a search space,* ***given a reference/input dataset***
    - Used to judge the quality of a particular model for the domain

- Score function are statistics—estimates of a population parameter based on a sample of data
- Examples:
    - Misclassification
    - Squared error
    - Likelihood

Parameter estimation vs. structure learning

- Models have both **parameters** and **structure**
- Parameters:
  - Coefficients in regression model
  - Feature values in classification tree
  - Probability estimates in graphical model
- Structure:
  - Variables in regression model
  - Nodes in classification tree
  - Edges in graphical model

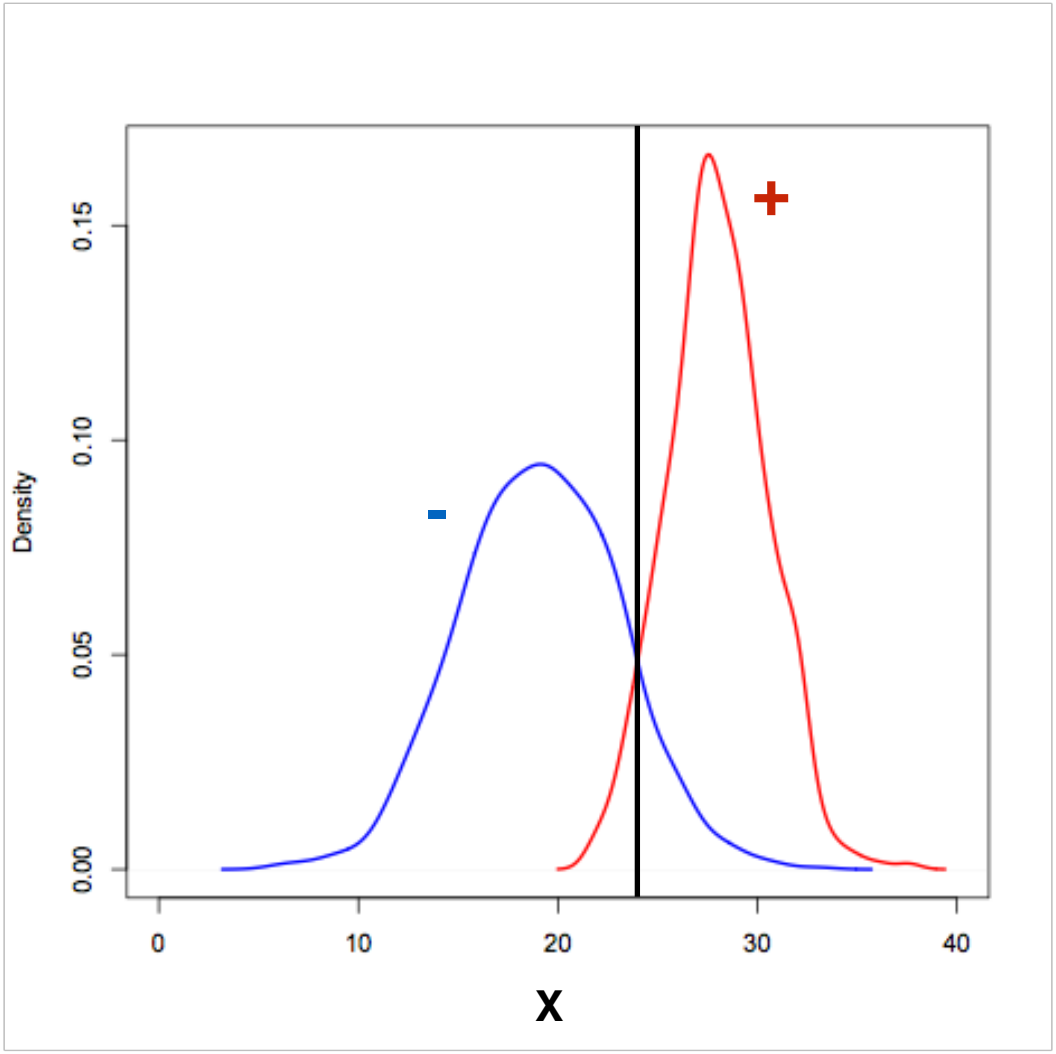


Example learning problem

Task: Devise a rule to classify items based on the attribute **x**

Knowledge representation:
If-then rules

Example rule:
If x > 25 then +
Else -



What is the model space?

All possible thresholds

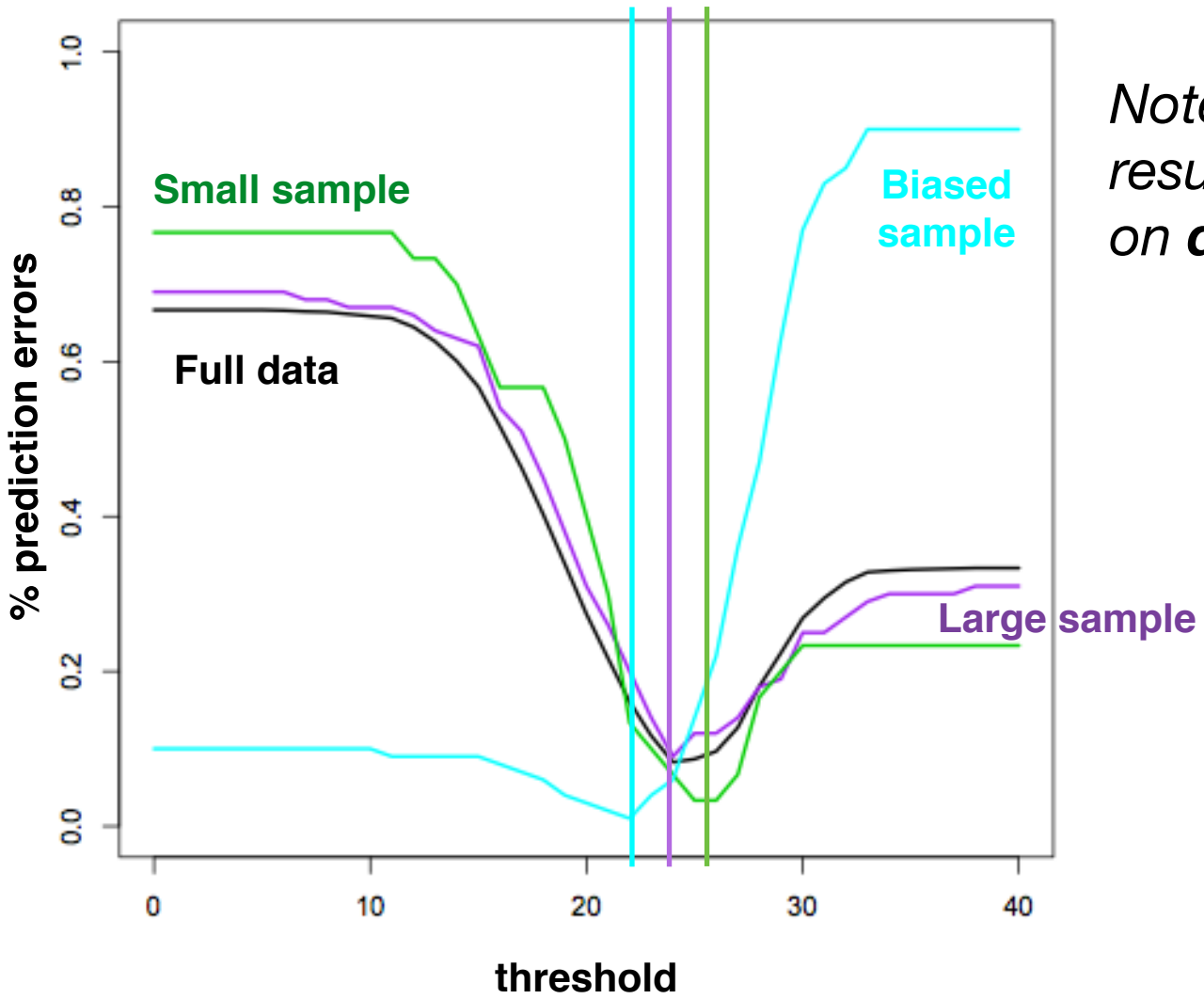
What score function?

Prediction error rate

Score function over model space

Search procedure?

Try all thresholds, select one with lowest score



*Note: learning result depends on **data***

Overview

- Task specification
- Data representation
- Knowledge representation
- Learning technique
    - Search + Evaluation
- **Prediction and/or interpretation**

Inference and interpretation

- Prediction technique
  - Method to apply learned model to new data for prediction/analysis
  - Only applicable for predictive and some descriptive models
  - Prediction is often used during **learning** (i.e., search) to determine value of scoring function
- Interpretation of results
  - Objective: significance measures
  - Subjective: importance, interestingness, novelty

Example: Identifying email spam

- Task
  - Design automatic spam detector that can differentiate between labeled emails
- Data
  - Table of relative word/ punctuation frequencies
- Knowledge representation
  - If/then rules with conjunctions of features
- Learning technique
  - **Search** over set of rules, **select** rule with maximum accuracy on training data

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

if (%george < 0.6) & (%you > 1.5) then spam
else email.

Example: Automatically identify digits on envelopes

- Task
  - Predict digit from image of handwritten envelopes
- Data
  - 16x16 matrix of pixel intensities
- Knowledge representation
  - Nearest neighbor classifier, defines tessellation over the feature space
- Learning technique
  - Implicit search for decision boundaries to minimize mis-classifications on training data

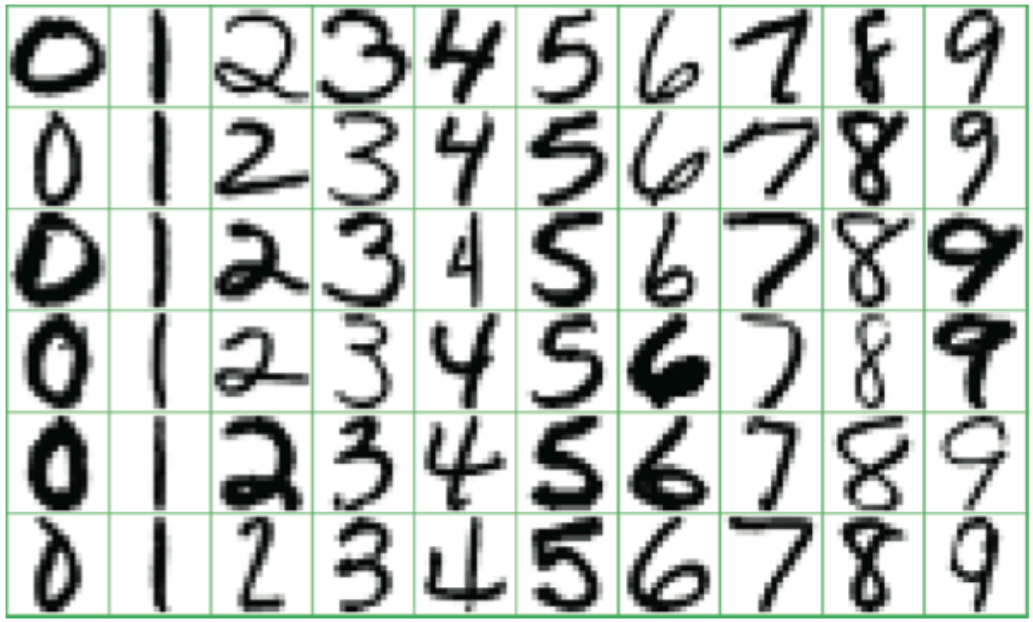


FIGURE 1.2. *Examples of handwritten digits from U.S. postal envelopes.*

Example: DNA expression levels in cancer tumors

- Task
  - **Unsupervised** learning: Examine DNA microarrays to determine which tumors are similar and which genes are similar
- Data
  - Expression levels [-6,6] for 6830 genes (rows) in 64 cancer tumors (columns) from different patients
- Knowledge representation
  - Clusters of similar genes/tumors
- Learning technique
  - Search over groups, minimize distance to group centroid

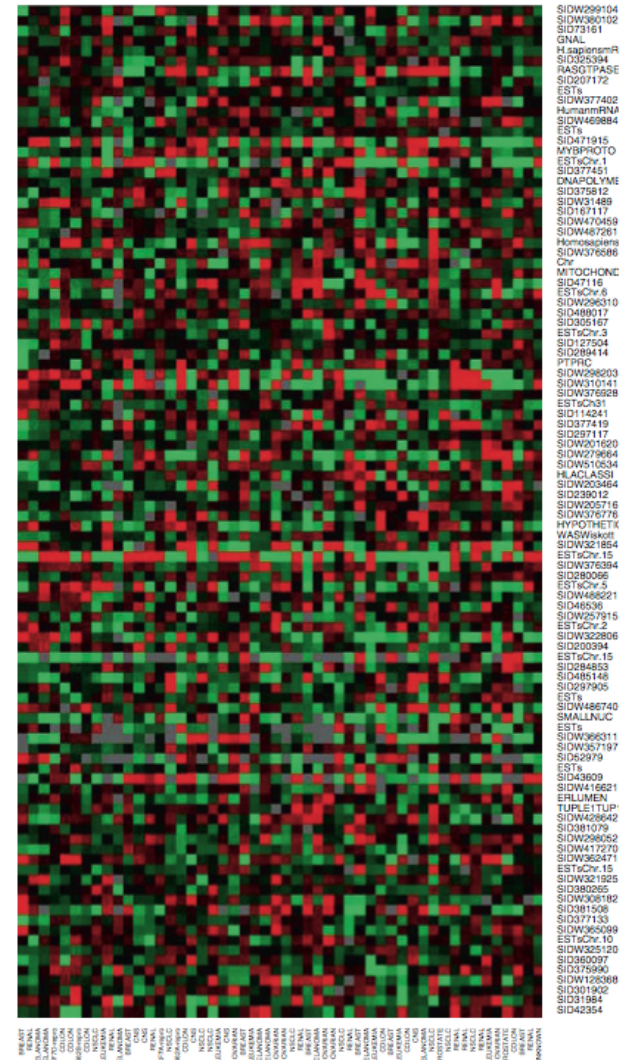


FIGURE 1.3. *DNA microarray data: expression matrix of* 6830 *genes (rows) and* 64 *samples (columns), for the human tumor data. Only a random sample of* 100 *rows are shown. The display is a heat map, ranging from bright green (negative, under expressed) to bright red (positive, over expressed). Missing values are gray. The rows and columns are displayed in a randomly chosen order.*