

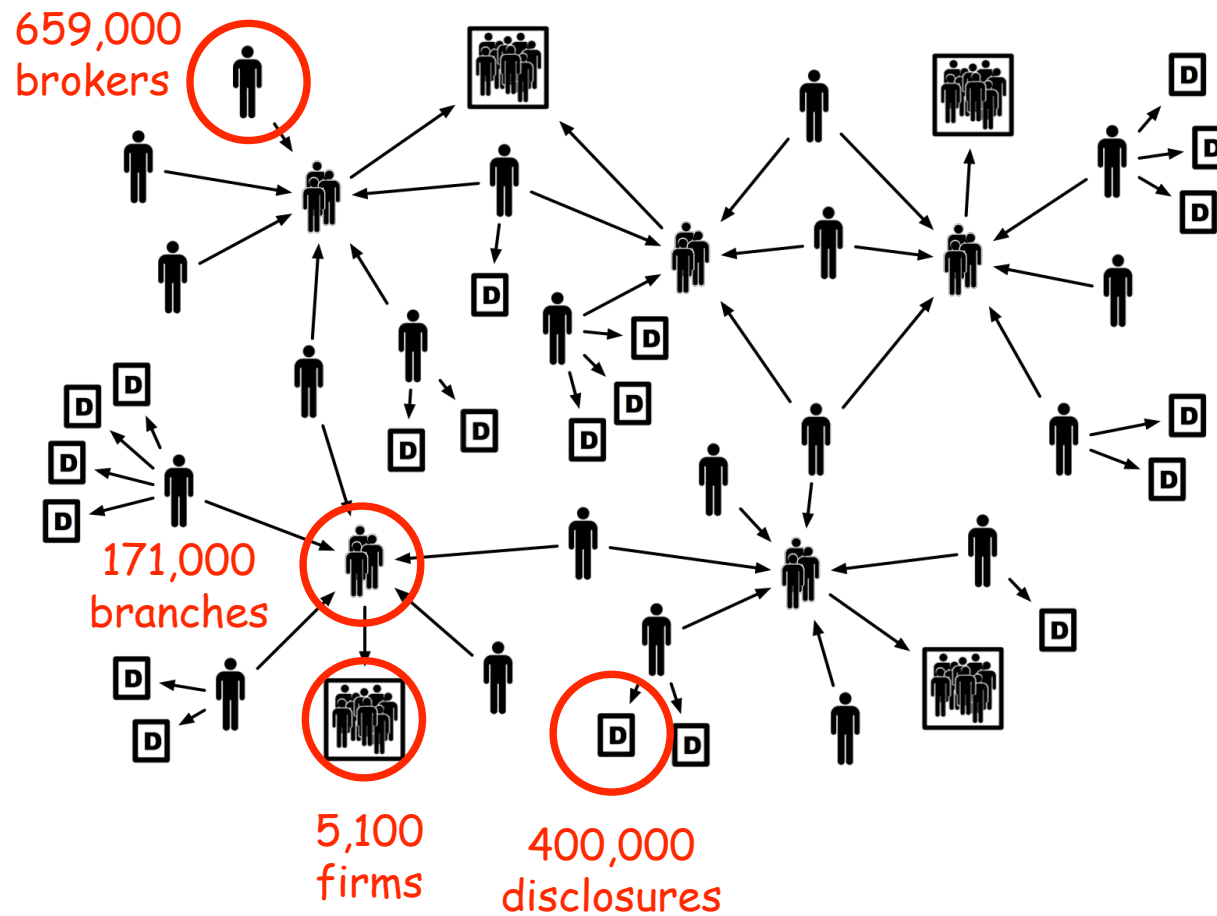
Data Mining

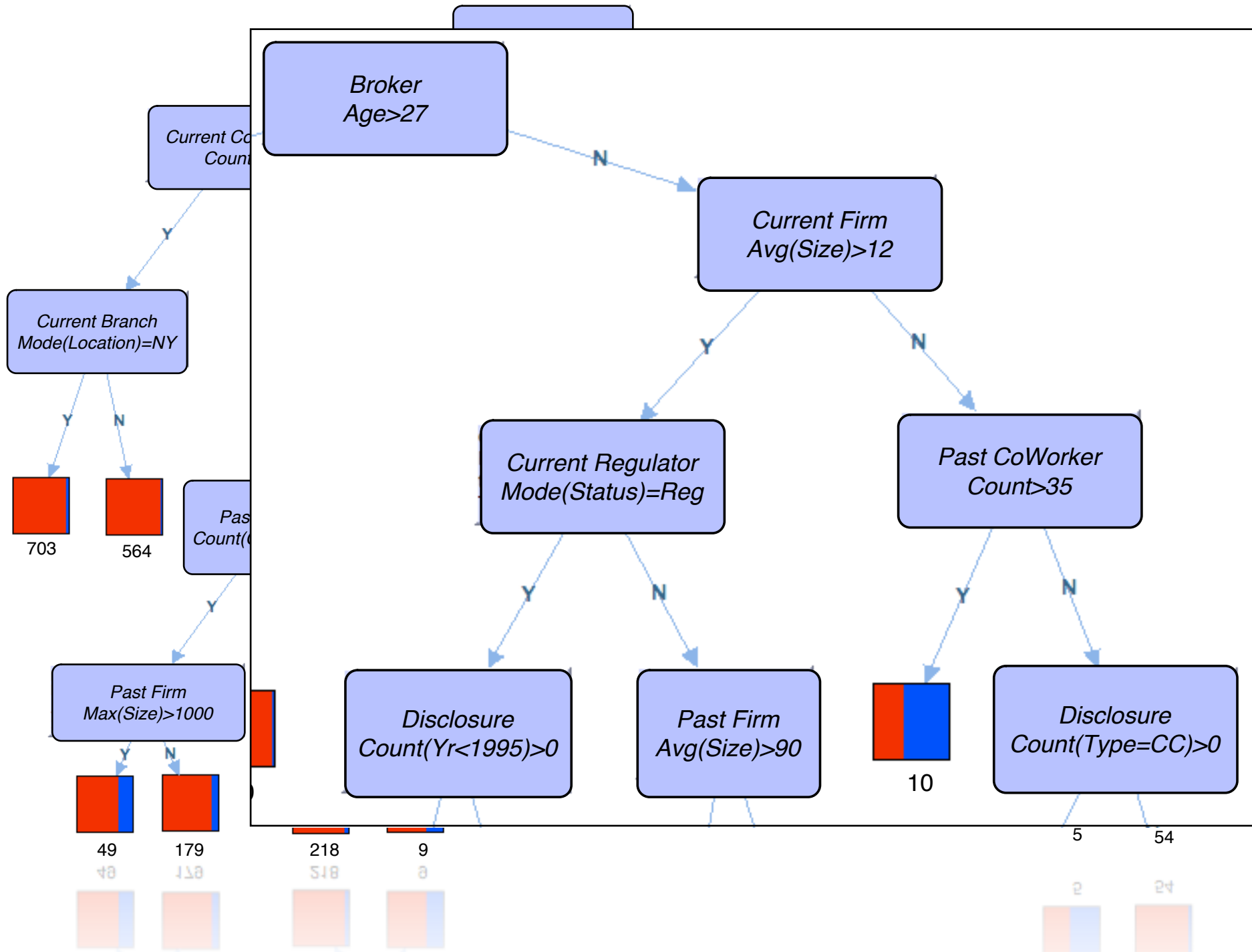
CS57300

Purdue University

August 26, 2010

Real-world example: NASD





Performance of NASD models

1.0

"One broker I was highly confident in ranking as 5...

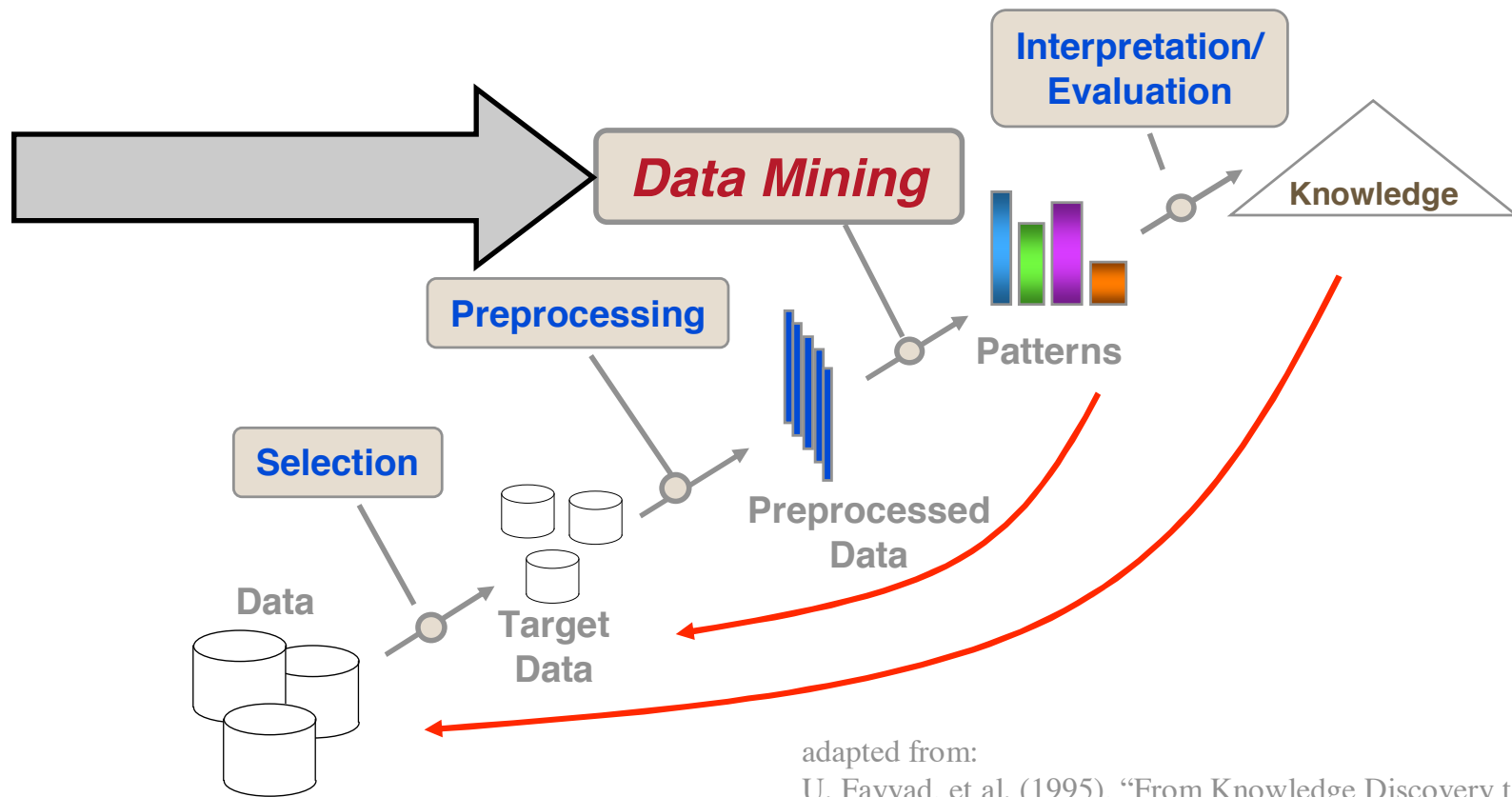
Not only did I have the pleasure of meeting him at a shady warehouse location, I also negotiated his bar from the industry...

This person actually used investors' funds to pay for personal expenses including his trip to attend a NASD compliance conference!

...If the model predicted this person, it would be right on target."

Mean Rating

Elements of Data Mining Algorithms



adapted from:
 U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press

Δεδομένα: η συλλογή και η (επεξεργασία) των δεδομένων
 Μηνύματα: να ανακαλυφθούν, να αναχθούν οι κρυμμένες πληροφορίες που
 η συλλογή και η (επεξεργασία) των δεδομένων μπορεί να ανακαλύψει.



Overview

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

Overview

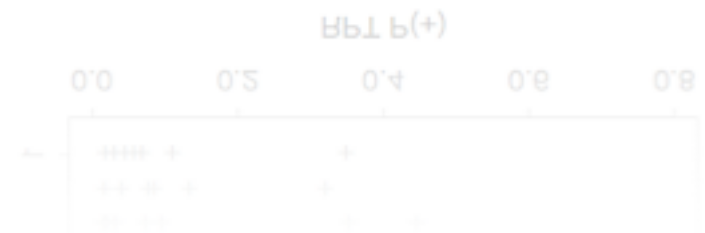
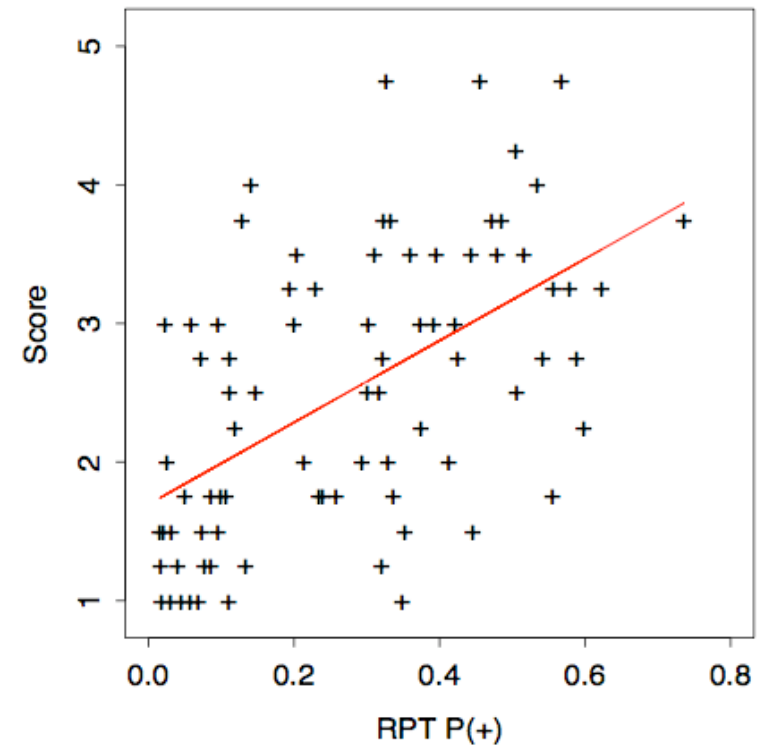
- Task specification
- Data representation
- Knowledge representation
- Learning technique
 - Search + scoring
- Inference and/or interpretation

Task specification

- 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

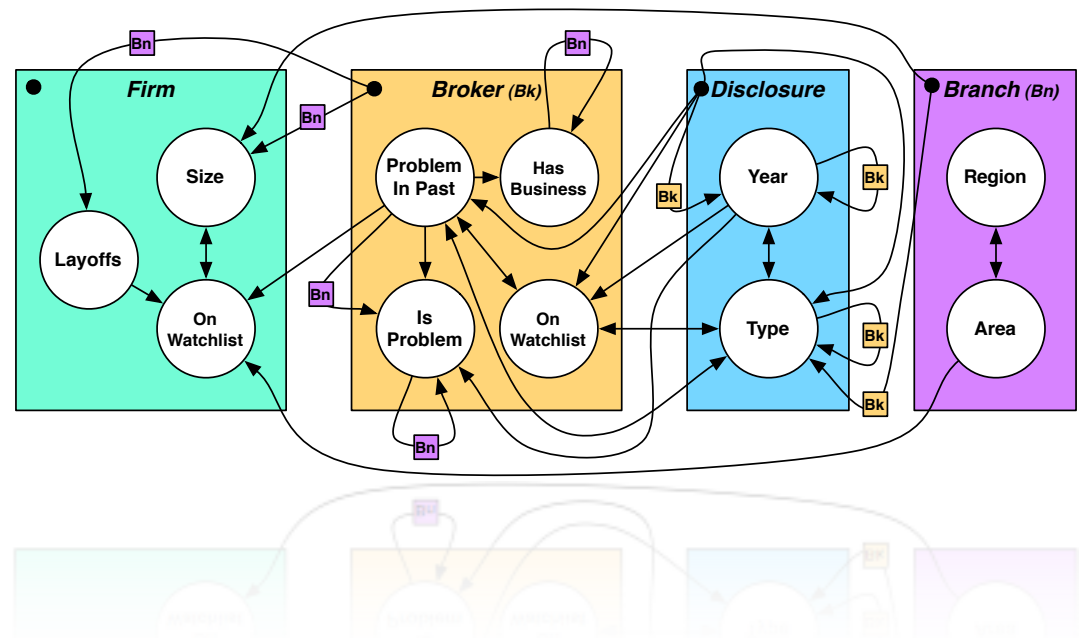
Exploratory data analysis

- 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



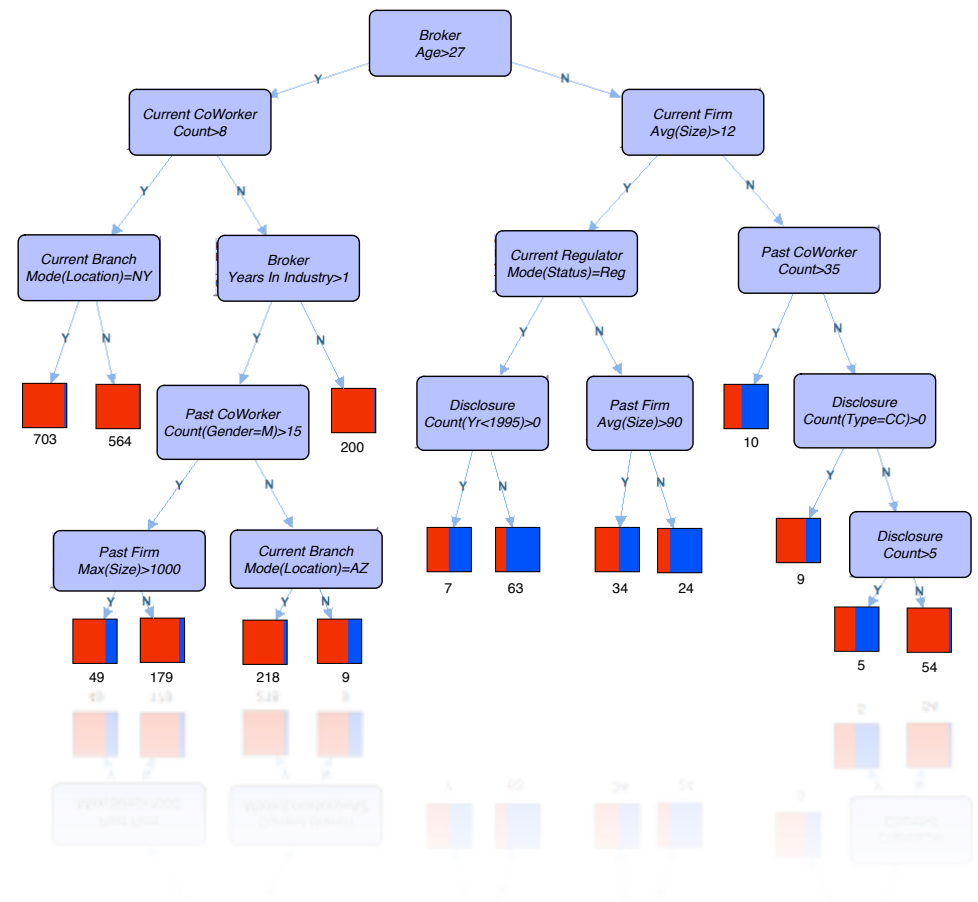
Predictive modeling

- Goal

- Learn model to predict unknown class label values given observed attribute values

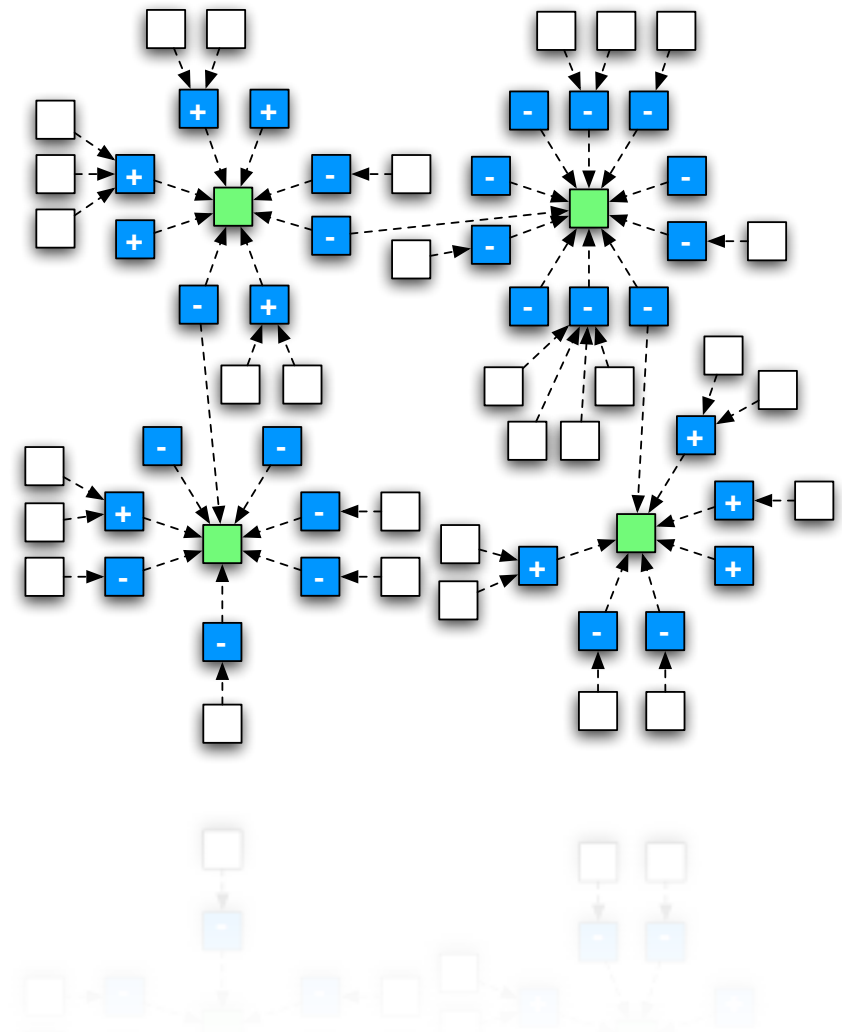
- Techniques

- Classification, regression



Pattern discovery

- Goal
 - Detect patterns and rules that describe sets of examples
- Techniques
 - Association rules, graph mining, anomaly detection



Overview

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

Data representation

- Choice of data structure for representing individual and collections of measurements
 - Individual measurements are single observations (e.g., person's date of birth, product price)
 - Collections of measurements are 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

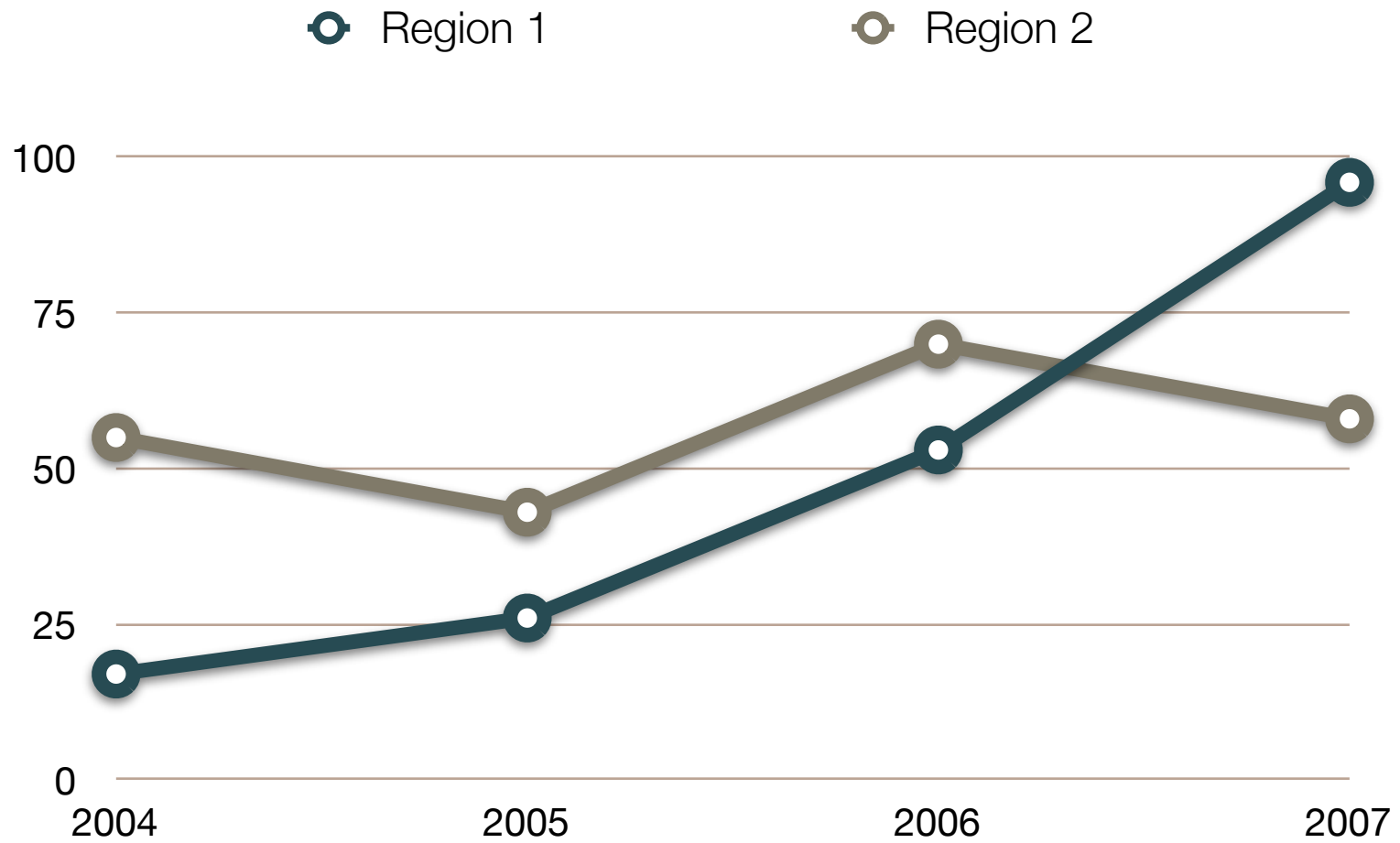
- Unit measurements:
 - Discrete values — categorical or ordinal variables
 - Continuous values — interval and ratio variables
- Compound measurements:
 - $\langle x, y \rangle$
 - $\langle \text{value}, \text{time} \rangle$

Tabular data

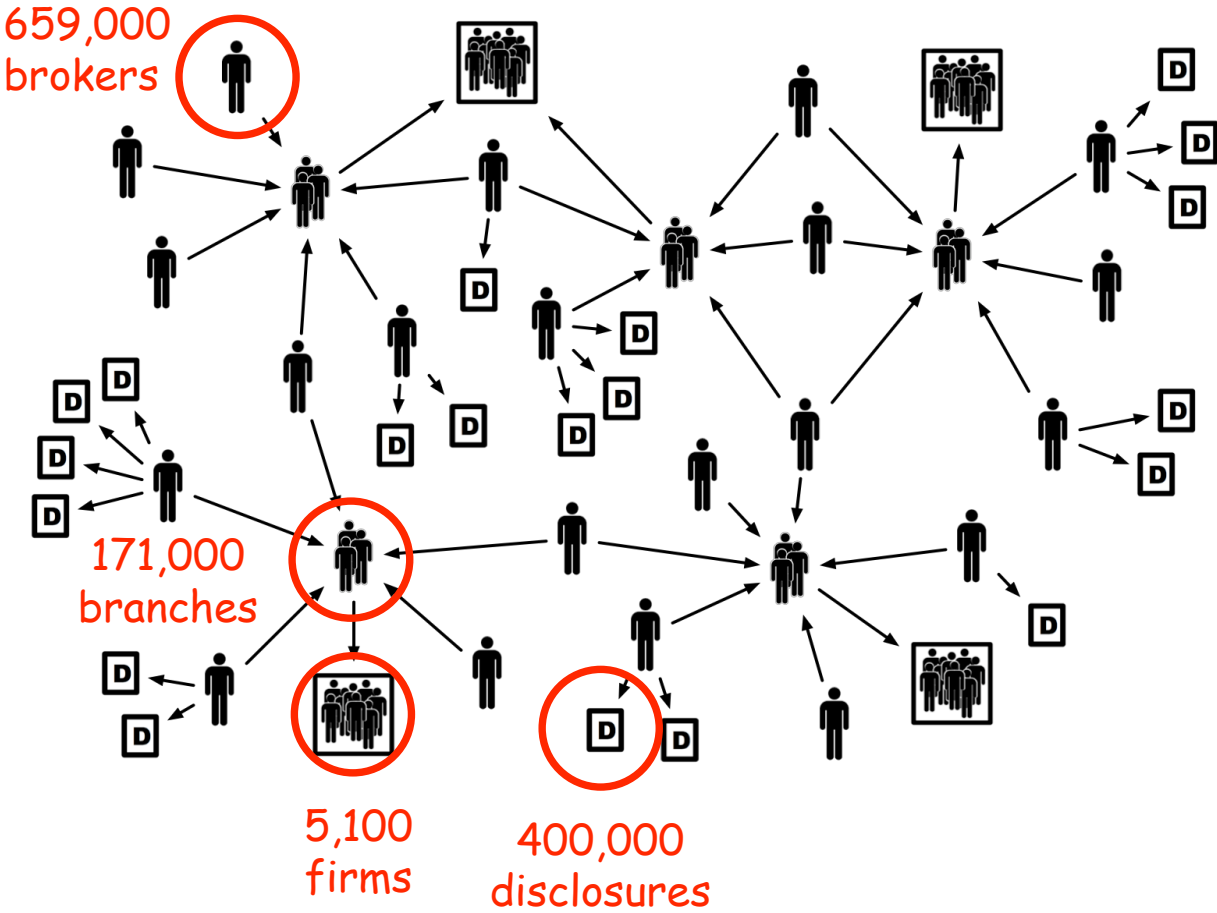
Fraud	Age	Degree	StartYr	Series7
+	22	Y	2005	N
-	25	N	2003	Y
-	31	Y	1995	Y
-	27	Y	1999	Y
+	24	N	2006	N
-	29	N	2003	N

N instances \times p attributes

Temporal data



Relational/structured data



firms disclosures

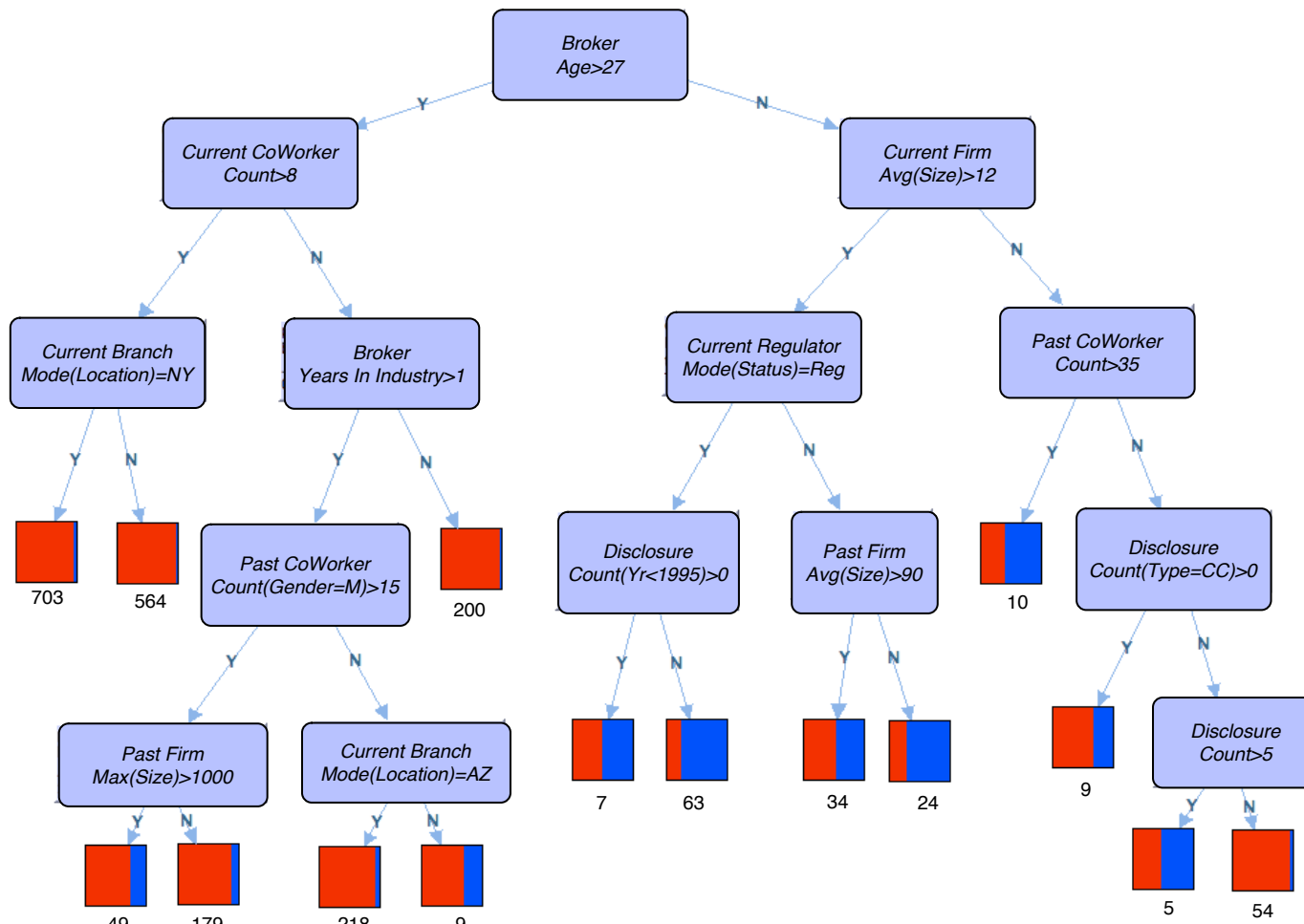
Overview

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

Knowledge representation

- Description of the results of the data mining algorithm (i.e., model or patterns)
- Examples:
 - Rules:
If short closed car **then** toxic chemicals
 - Conditional probability distributions
 $P(\textit{fraud} \mid \textit{age}, \textit{degree}, \textit{series7}, \textit{startYr})$

Classification tree



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

Regression model

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

- 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
- Inference and/or interpretation

Learning technique

- Method to construct model or patterns from data
- Knowledge representation defines a set of possible models or patterns
- **Scoring function** associates a numerical score with each member of that set
- **Search technique** defines a method for generating members of that set and optimizing their score

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

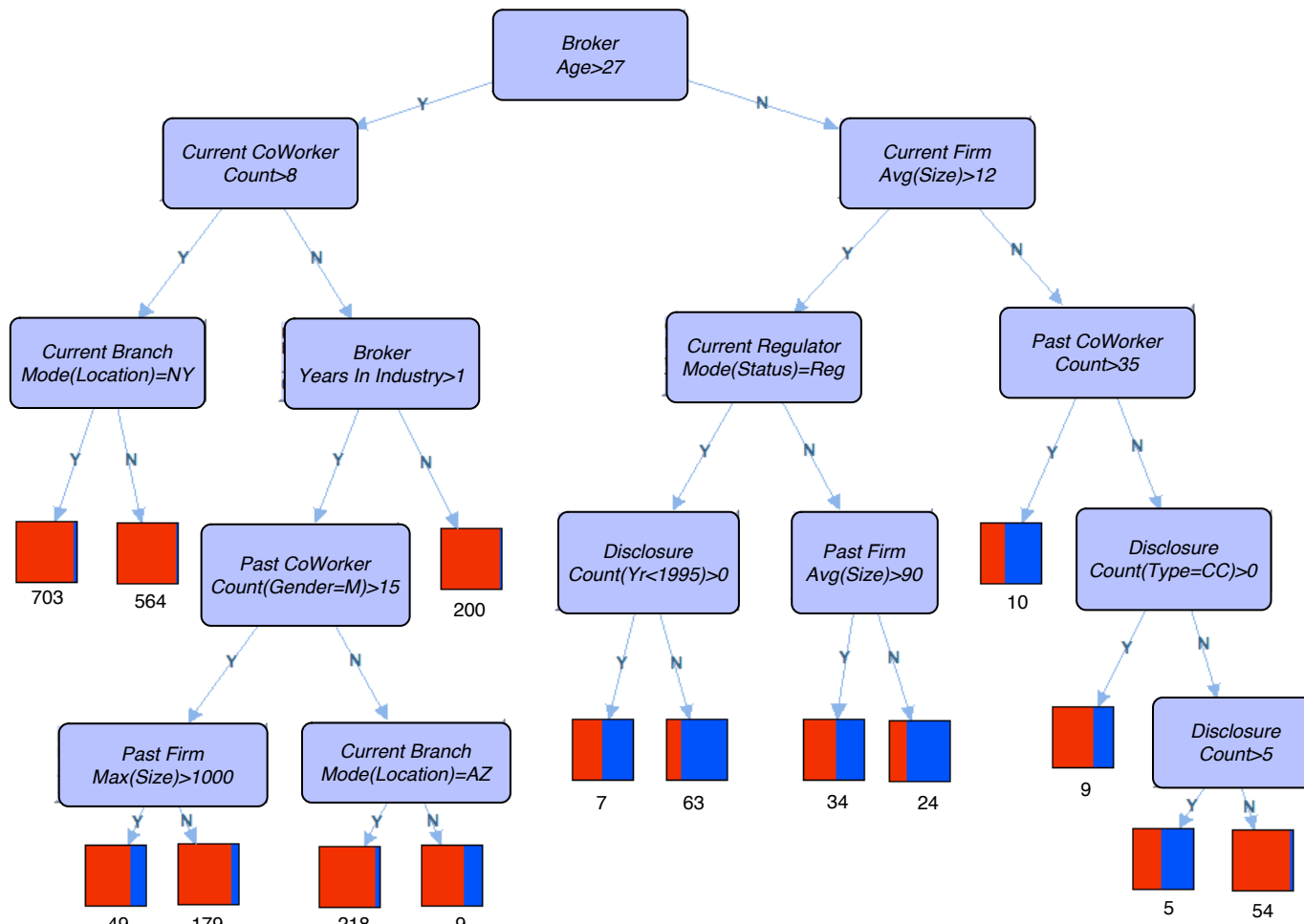


- **Structure:**

- Variables in regression model
- Nodes in classification tree
- Edges in graphical model



Classification tree



How many unique classification trees are there?

Search space

- Can we search exhaustively?
- Simplifying assumptions
 - Binary tree
 - Fixed depth
 - 10 binary attributes

Tree depth	Number of trees
1	10
2	8×10^2
3	3×10^6
4	2×10^{13}
5	5×10^{25}

Scoring functions

- Given a dataset, assign a numeric score to each possible model in a search space
- Evaluation functions are statistics—estimates of a population parameter based on a data sample
- Examples:
 - Information gain
 - Misclassification
 - Squared error
 - Likelihood

Chi-square statistic

- Used to measure association of feature with class
- Decision tree learning:
 - Recursive greedy partitioning
 - Pick feature with maximum score at each node

	+	-
Y	17	5
N	10	42

$$\chi^2 = \sum_i \frac{(ct_{obs} - ct_{exp})^2}{ct_{exp}^2}$$

Overview

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

Inference and interpretation

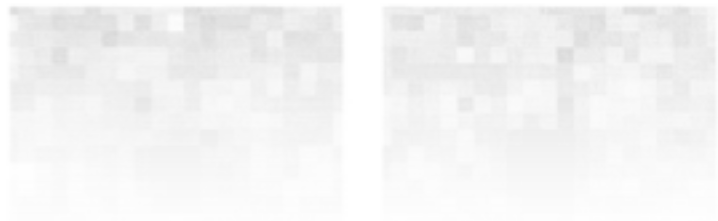
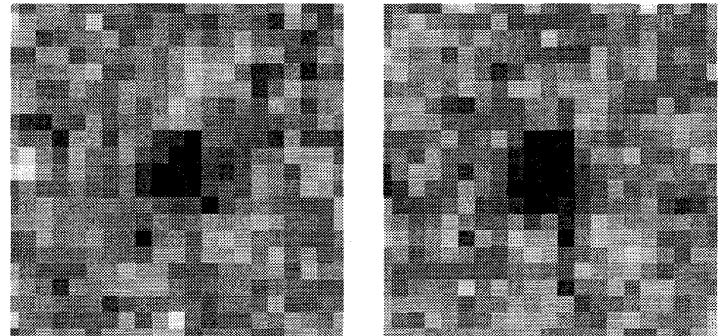
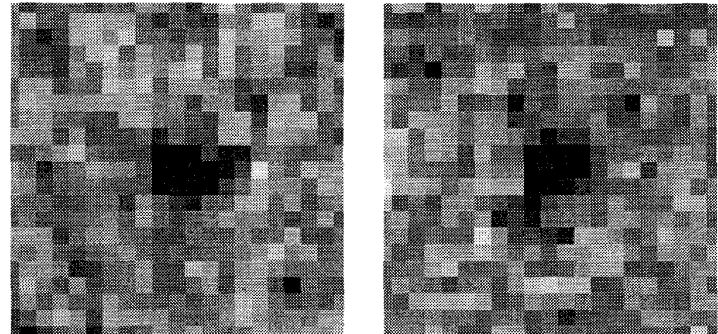
- Inference technique
 - Method to apply learned model to new data for prediction/analysis
 - Only applicable for predictive and some descriptive models
 - Inference is often used during search to determine value of evaluation function
- Interpretation of results
 - Objective: significance measures
 - Subjective: importance, interestingness, novelty

Example: SKICAT

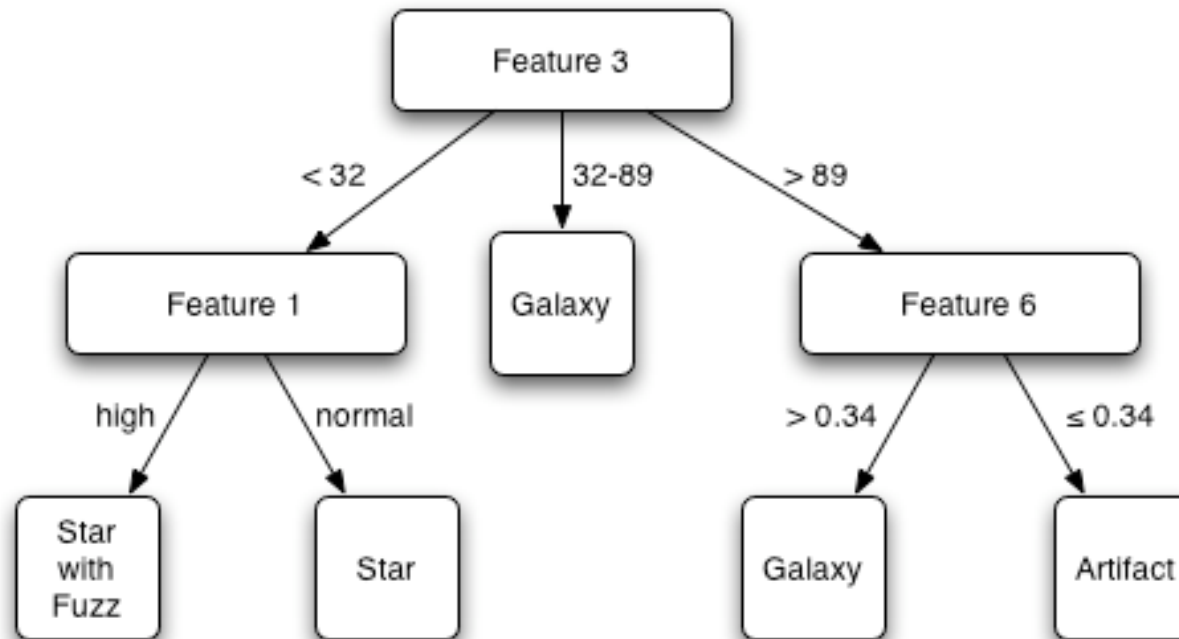
- Sky Image Catalog Analysis Tool (*Fayyad, Djorgovski, and Weir, 1996*)
 - Data from Second Palomar Observatory Sky Survey
- **Task:** Classify each detectable object as {star, galaxy, artifact}
- **Goals:**
 - Accuracy >90% (in relation to human labels)
 - Interpretable models (important for astronomers to trust models)

Data representation

- Set of independent images
 - Labeled with category
 - 38+2 numeric features characterizing image
- 10^9 objects detected in 3 terabytes of data
- ~1000 objects labeled by humans



Knowledge representation



RULER: set of rules derived from binary classification trees

Learning technique

- Search
 - Greedy, recursive partitioning to learn trees
 - Learn a set of trees from random subsamples of the data
 - Prune trees into a set of rules
- Evaluation function
 - **C-SEP** measure detects class separation for selecting tree features
 - **Fisher's exact test** determines correlation of feature with class to prune features from rules

Inference and interpretation

- Inference
 - Not mentioned explicitly in paper
 - Probably involved applying all applicable rules and taking a majority vote of the predicted class labels
- Evaluation
 - Overall accuracy: 94% (*above desired threshold*)

Netflix Prize

Home Rules Leaderboard Register Update Submit Download

NETFLIX

Browse Recommendations Friends Queue Buy DVDs

Home Genres New Releases Previews Netflix Top 100 Crit

Movies For You

Randy, the following movies were chosen based on your interest in:
[Bowling for Columbine](#)
[Carnivale: Season 1](#)
[Eisenstein 98.5](#)



The Big One

★★★★☆

or subversive

y from

n /

angel

Carnivale: Season 2

Disc Series

★★★★☆

Daniel Krau

rivetingly cr

series cont

document

entures of a mo

ies who've made the

stbowl their ... [Read Mo](#)



Roger & Me

★★★★☆

In this b

satir

All Discs
Guaranteed!

You really liked it...

Now owned for just \$5.99

Shop as low

titles

Original artv

OT

IGHT

Lewis Black: Re

and Bone



Add

★★★★☆

Not Interested

Not Interested

Red Eye



Rear Window



Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the [Rules](#) to see what is required to win the Prizes. If you are interested in joining the quest, you should [register a team](#).

You should also read the [frequently-asked questions](#) about the Prize. And check out how various teams are doing on the [Leaderboard](#).

Good luck and thanks for helping!

Guides:

Member Favorites

Easter Eggs

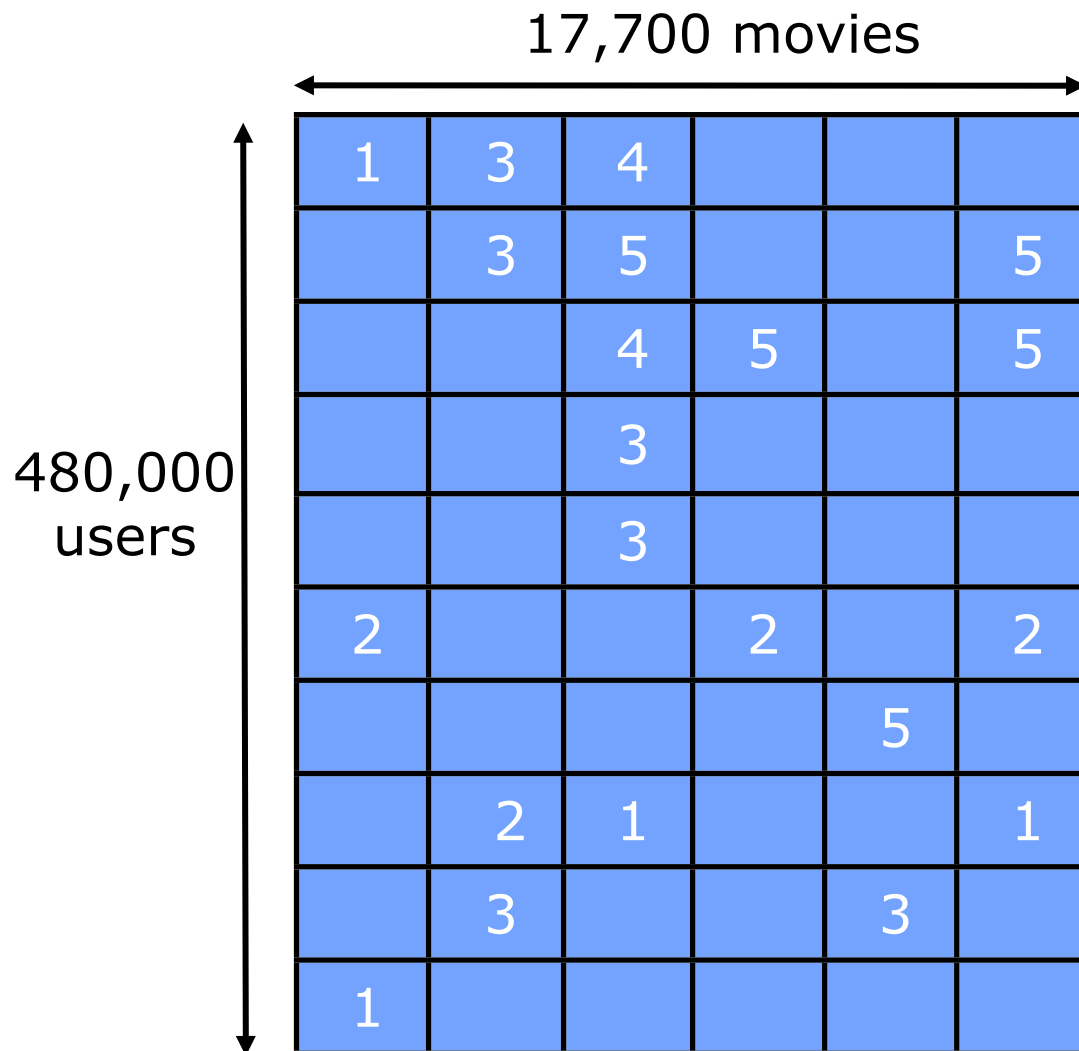
By Decade

By Studio

Movies You've Seen

Give a friend

Ratings Data





Next class

- Reading: PDM Appendix if necessary
- Topic: Background and basics