Data Mining & Machine Learning

CS37300
Purdue University

August 19, 2024

Bruno Ribeiro

About me

- Professor Ribeiro (Hee-BAY-roh)
- Associate Professor in Computer Science, Purdue University
  - Was on sabbatical 2023-2024 @ Stanford

- Field: Machine learning
  - Focus: Developing methods for neural networks that generalize to domains beyond the ones observed in training.
  - **Applications**: Relational reasoning, physics+ML tasks.
  - **Undergraduate research opportunities:** End of semester (around December) our lab releases a call

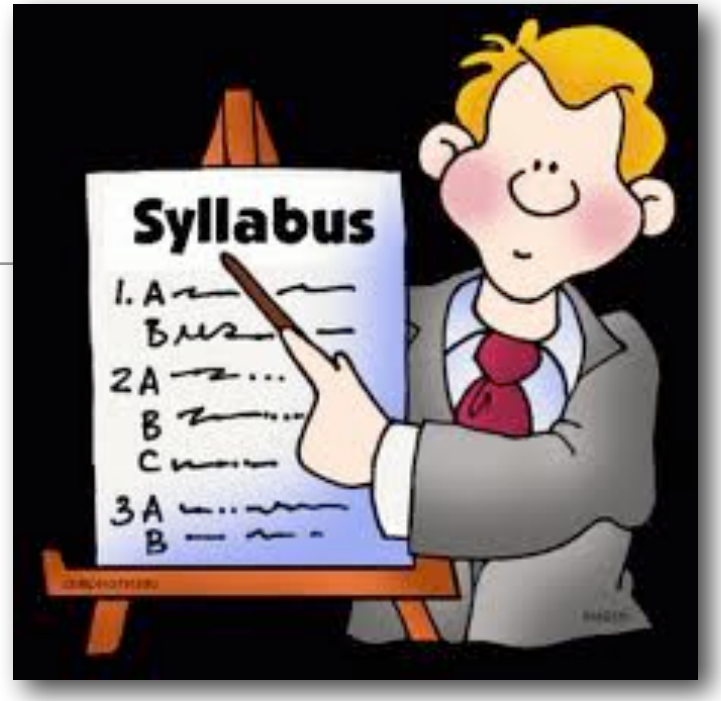
Course overview

Goals

- Identify key elements of machine learning algorithms
- Understand how algorithmic elements interact to impact performance
- Understand how to choose algorithms for different analysis tasks
- Analyze data in both an exploratory and targeted manner
- Implement and apply basic algorithms for supervised and unsupervised learning
- Accurately evaluate the performance of algorithms, as well as formulate and test hypotheses

Topics

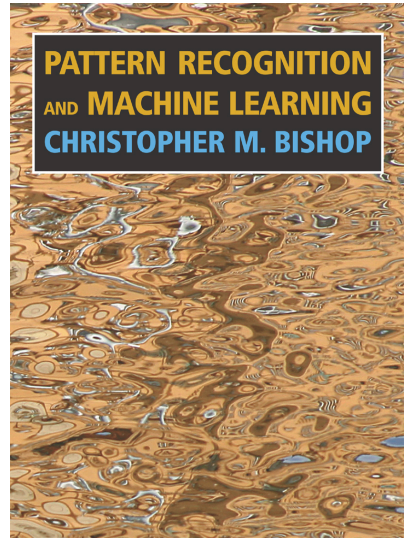
- Elements of data science algorithms
    - Machine Learning
    - Data Mining
    - Statistics
- Statistical basics and background
- Data preparation and exploration
- Predictive modeling
- Methodology, evaluation
- Descriptive modeling



Syllabus / Logistics

- Syllabus and ALL necessary information (slides, notes, links) will be posted on our website
- https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2024/

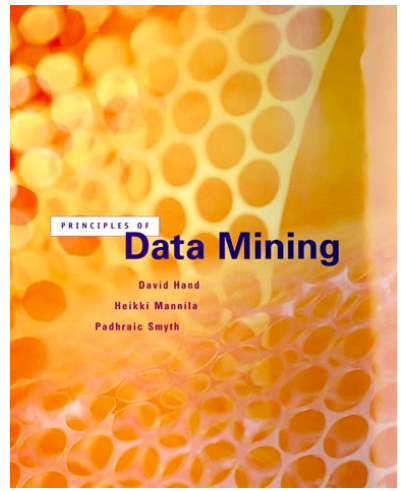
Textbooks



All found online or at the library for download

Textbooks

- Pattern Recognition and Machine Learning by Christopher M. Bishop is a very detailed and thorough book on the foundations of machine learning.
- Principles of Data Mining (David J. Hand; Heikki Mannila; Padhraic Smyth, FREE with PUID)
- Foundations of Machine Learning, by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, MIT Press, Second Edition, 2018.



Recommended but not required:

- Probabilistic Machine Learning: An Introduction, Kevin P. Murphy, MIT Press, 2022
- A Course in Machine Learning by Hal Daume III is a good book with important practical guidelines
- The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani, and Jerome Friedman is an excellent reference book, available on the web for free at the link.

Workload

- Homeworks (6 theory + programming assignments)
  - Six assignments including written/math exercises, programming assignments in python
    - Python is an important language to learn in data mining, data science, and machine learning
  - Late policy: No Late Homework (Grade = zero after deadline)
    - Submission on Gradescope
    - Firm deadlines (6:00pm) with no late penalty until 1:00am next day
  - Lowest homework score will be dropped from the average
    - Do not skip a homework early: Save for emergencies
- Exams
  - Midterm and final exam

Grading

- Grades will be posted on Brightspace: https://purdue.brightspace.com/d2l/home/1095563
- Attendance: 5%
  - ML Competition (Kaggle Competition): up to +5% (extra credit)
  - Homework: 45% (the lowest grade homework will be dropped from average)
    - Serious and documented medical or family emergencies will be automatically counted as a zero grade (i.e., discarded from the average). Additional extensions (beyond one missed homework) will be granted if the documented emergency persists for 2+ homeworks.
    - Students are advised to not drop a homework for non-emergency reasons since, if an emergency happens, the student will have two zero grades and one of them will count towards the average.
  - Midterm: 20%
  - Final exam: 30%

Kaggle Competition (up to +5% credit)

- +5% extra credit to top 10% entries
- +3% extra credit to the top 20%
- +2% extra credit to the top 50%
- +1% extra to bottom <50%

CS373
ML Prediction Task

Details after Sep 02

Computing Resources

- Scholar Cluster

Scholar Specifications

The Scholar A nodes have 128 processor cores, 256 GB RAM and 100 Gbps Infiniband interconnects.

Scholar Front-Ends

Front-Ends	Number of Nodes	Processors per Node	Cores per Node	Memory per Node	Retires in
No GPU	4	Two Haswell CPUs @ 2.60GHz	20	512 GB	2023
With GPU	3	Two Sky Lake CPUs @ 2.60GHz with one NVIDIA Tesla V100	20	756 GB	2023

Scholar Sub-Clusters

Sub-Cluster	Number of Nodes	Processors per Node	Cores per Node	Memory per Node	Retires in
A	4	Two AMD EPYC 7713 3rd generation ("Milan") 64-Core Processors	128	256 GB	2027
B	3	AMD EPYC 7702P 2nd generation ("Rome") 64-Core Processor	64	256 GB	2026
G	4	Two Skylake CPUs @ 2.10GHz with one NVIDIA Tesla V100 32GB GPUs	16	192 GB	2027
H	2	Two AMD EPYC 7543 3rd generation ("Milan") 32-Core Processors with two NVIDIA A30 24GB GPUs	64	512 GB	2027
H-MIG	2	Two AMD EPYC 7543 3rd generation ("Milan") 32-Core Processors with eight 6GB Multi-Instance GPUs (MIGs) configured from two NVIDIA A30 24GB GPUs.	64	512 GB	2027

Software needed and cluster usage manual

---

https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2024/howto/cluster-how-to.html

Software Requirements

This course will require the following versions of the libraries

- **Python 3.10+**
- **Anaconda (one that contains numpy 1.26+)**
- For the deep learning part: **Pytorch 2.1.2+**

Hardware Requirements

Purdue Machines

We recommend all students to use the scholar cluster (details below). You should already have access.

Personal Machines

The use of a **personal machine** is allowed but **not required** (we will provide GPU access through the Scholar cluster). Make sure it meets the following specifications:

- A 4-core, 8 thread processor
- An SSD hard drive with at least 10GB free (more may be needed for the Kaggle competition)
- An Nvidia GPU (Pascal, Volta, Turing, Ampere, Hopper, Lovelace) with least 4GB of VRAM.
  - AMD and Intel GPUs are not supported by our course

Python Resources

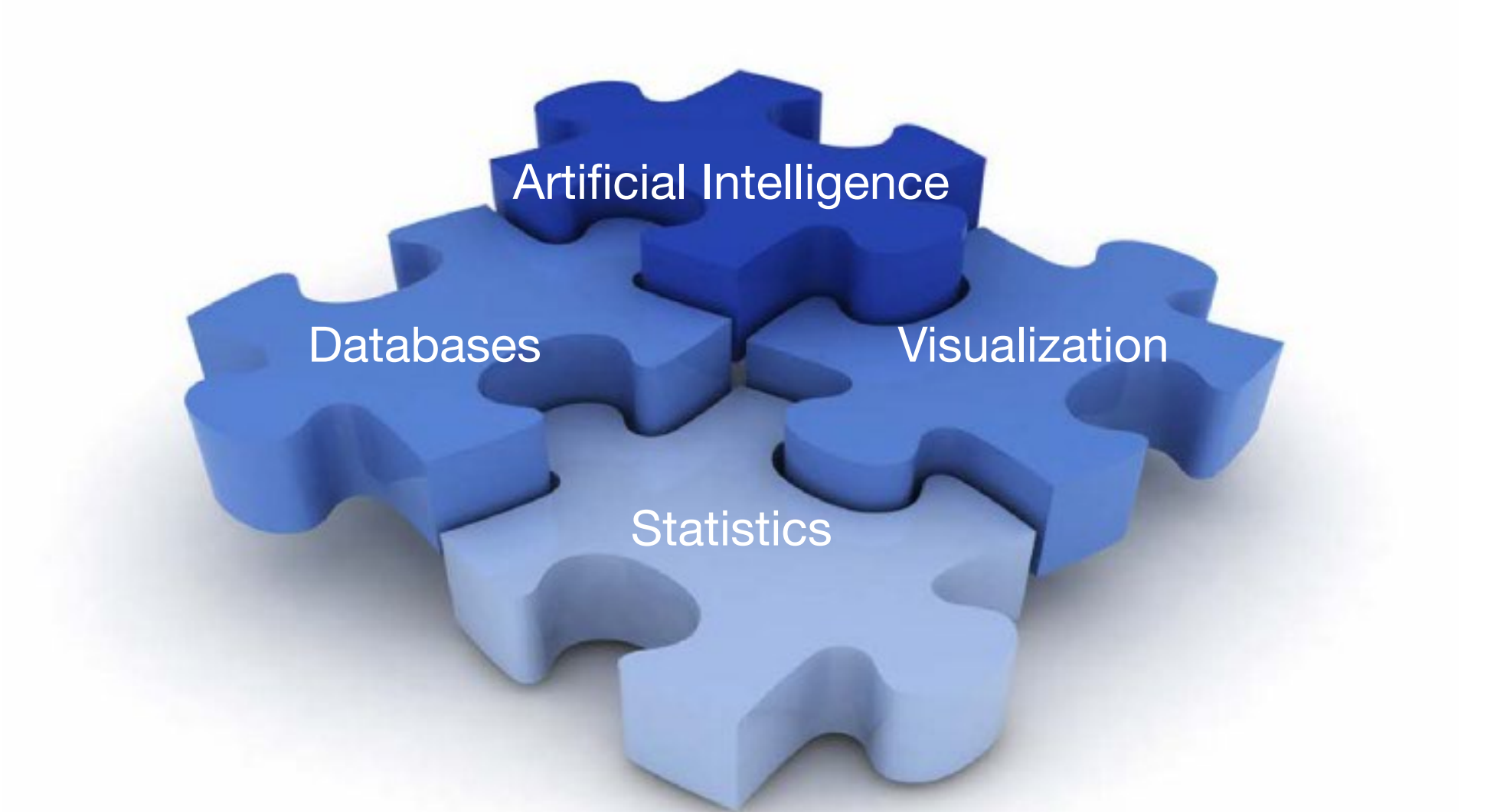
- Programming assignments will be in Python (3.10+)

- **Python tutorials**
- A crash course on python + numpy + pandas + sklearn + pytorch:
  - https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2024/howto/python-basic.html
  - https://www.cs.purdue.edu/homes/ribeirob/courses/Fall2024/howto/python-basic.ipynb
- https://www.learnpython.org/en/
- Official python language tutorial: https://docs.python.org/3/tutorial/
- https://www.greenteapress.com/thinkpython/html/
- https://pandas.pydata.org/docs/user_guide/index.html

Course introduction

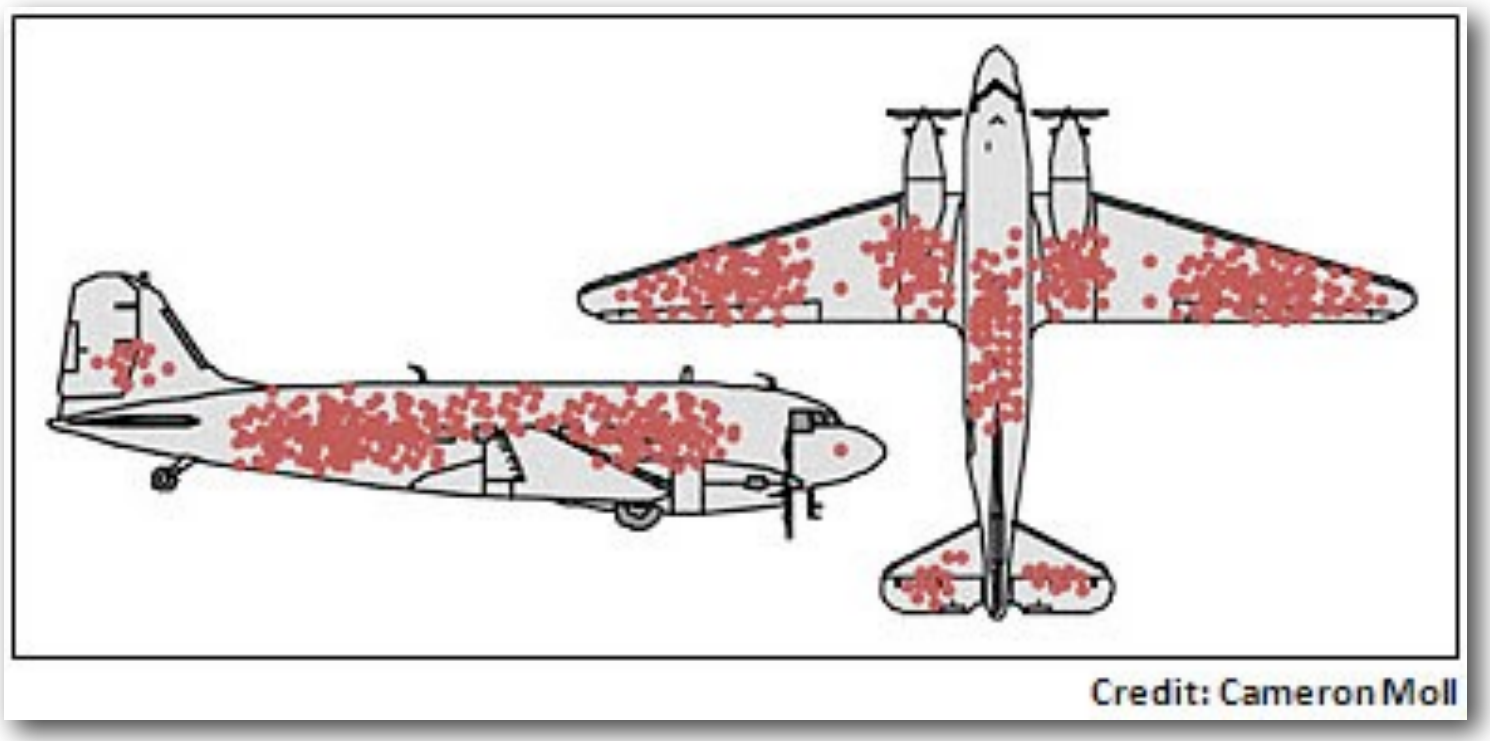
Machine Learning

- **Machine learning**: How can we build computer systems that automatically improve with experience? *(Mitchell 2006)*



Example (Survival Bias)

Bullet holes of surviving airplanes



During WWII, statistician Abraham Wald was asked to help the British decide where to add armor to their planes

The machine learning process

Data

Target data

Processed data

Patterns

Knowledge

Machine Learning Process

1. Application setup:
    - Acquire relevant domain knowledge
    - Assess user goals
2. Data selection
    - Choose data sources
    - Identify relevant attributes
    - Sample data
3. Data preprocessing
    - Remove noise or outliers
    - Handle missing values
    - Account for time or other changes
4. Data transformation
    - Find useful features
    - Reduce dimensionality

Machine Learning Process

5. Data mining:
    - Choose task (e.g., classification, regression, clustering)
    - Choose algorithms for learning and inference
    - Set parameters
    - Apply algorithms to search for patterns of interest
6. Interpretation/evaluation
    - Assess accuracy of model/ results
    - Interpret model for end-users
    - Consolidate knowledge
7. Repeat...

Machine Learning Process

- Data representation: Describe the data
- Task specification: Outline the goal(s)
- Knowledge representation: Describe the rules
- Learning technique:
  - Search: Identify a rule
  - Score function: Estimate how good rule is
- Prediction technique: Apply the rule
- ML system: Do above in combination

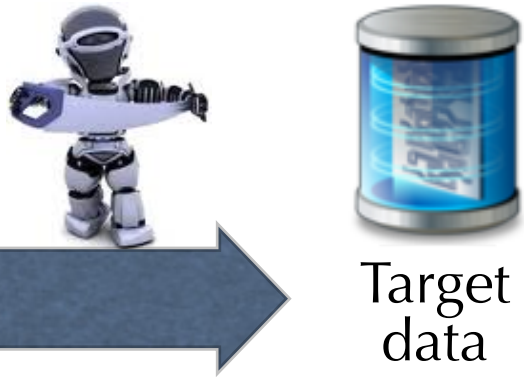
Complexities

- Data size: vastly larger or changing rapidly
- Data representation: can affect ability to learn and interpret models
- Knowledge representation: needs to capture more subtle forms of probabilistic dependence
- Search space: vastly larger
- Evaluation functions: difficult to assess confidence in model utility

The machine learning process

Data

Efficiently store and organize all relevant data



Target data

Ensure processed data can give answers to our hypotheses

Processed data

Knowledge

Evaluate predictions, test hypotheses, interpret results

Patterns

Formalize task and assumptions; ensure we learned true relationships

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

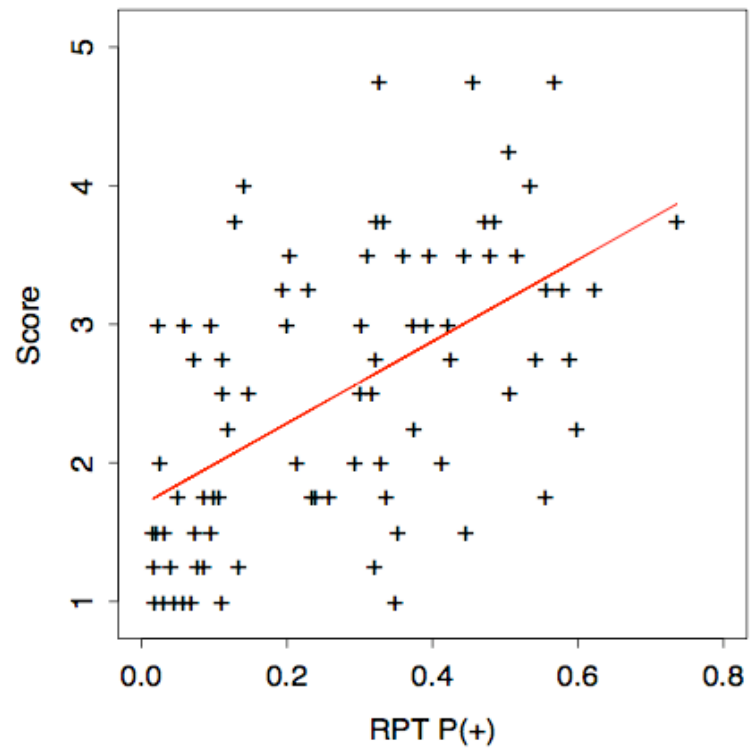
Task specification

- *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 user wants to listen to Drake as next song.
  - From a set of *unlabeled examples*, cluster artists into a *set of homogeneous groups* based on user listening behavior (e.g. create a playlist)

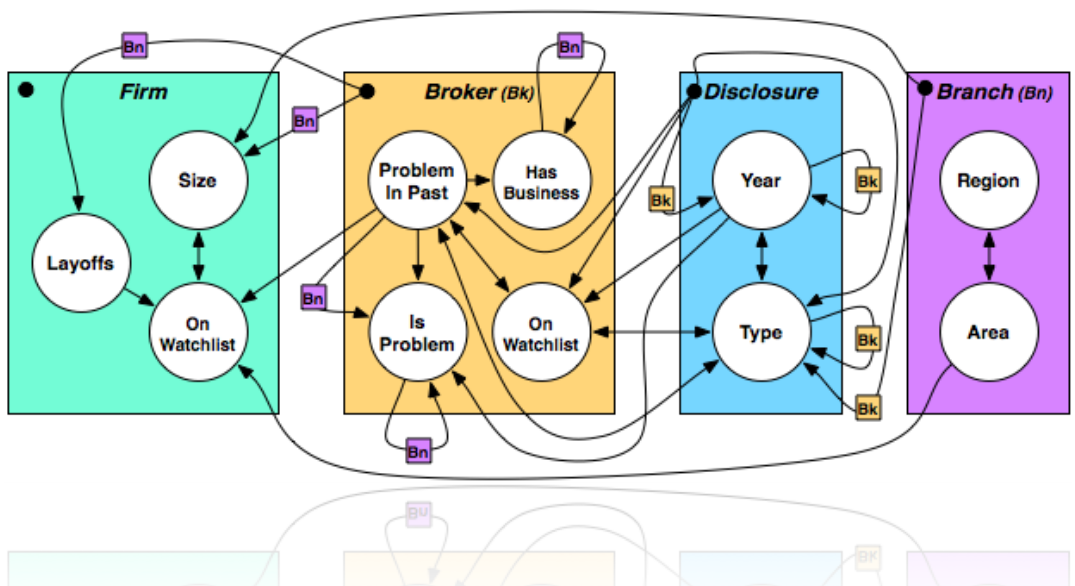
Exploratory data analysis

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



Descriptive modeling

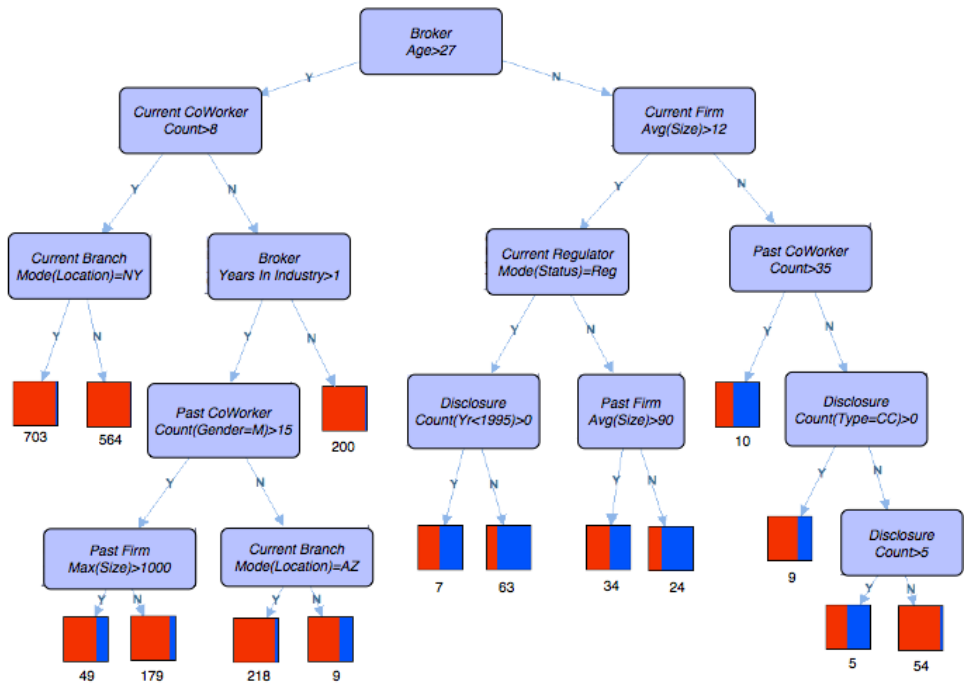
- Goal
  - Summarize the data or the underlying generative process
- Techniques
  - Density estimation, cluster analysis and segmentation



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

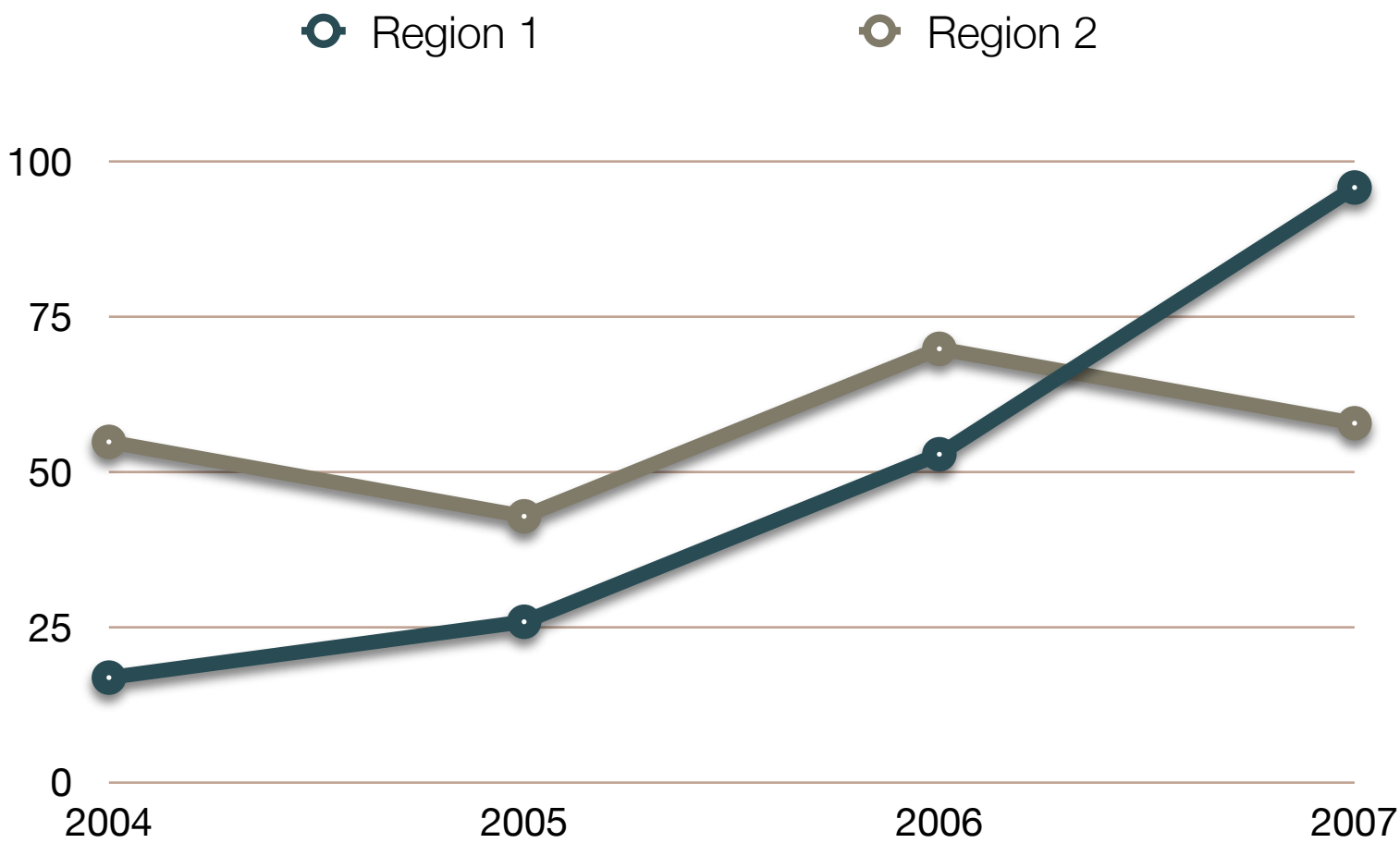
- Unit measurements:
  - Discrete values — categorical or ordinal variables
  - Continuous values — interval and ratio variables
- Compound measurements (vectors, matrices, tensors):
  - < x, y >
  - < value, time >

Data representation: Table/vectors

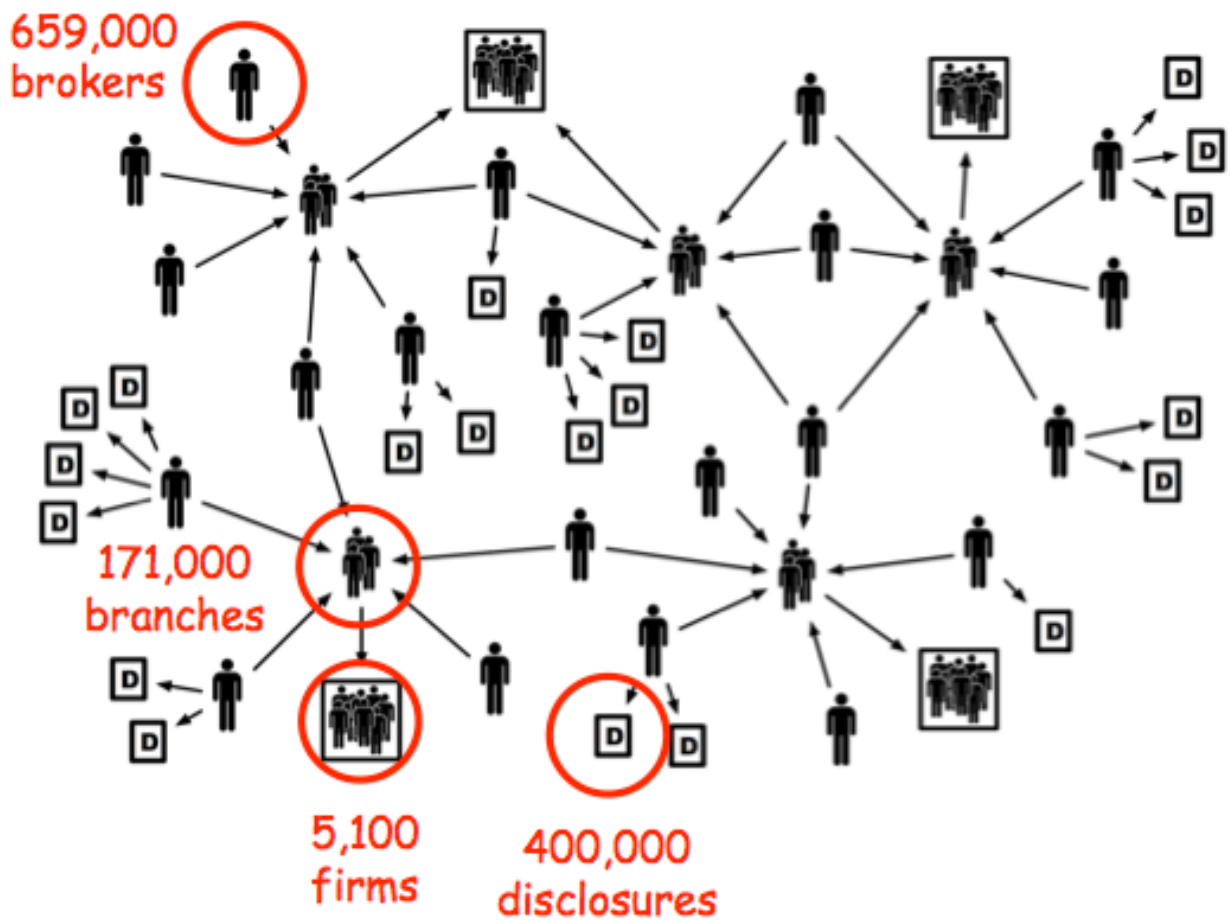
Fraud	Age	Degree	StartYr	Series7
+	22	Y	2005	N
-	25	N	2003	Y
-	31	Y	1995	Y
-	27	Y	1999	Y
+	24	N	2006	N
-	29	N	2003	N

N instances X *p* attributes

Data representation: Time series/sequences



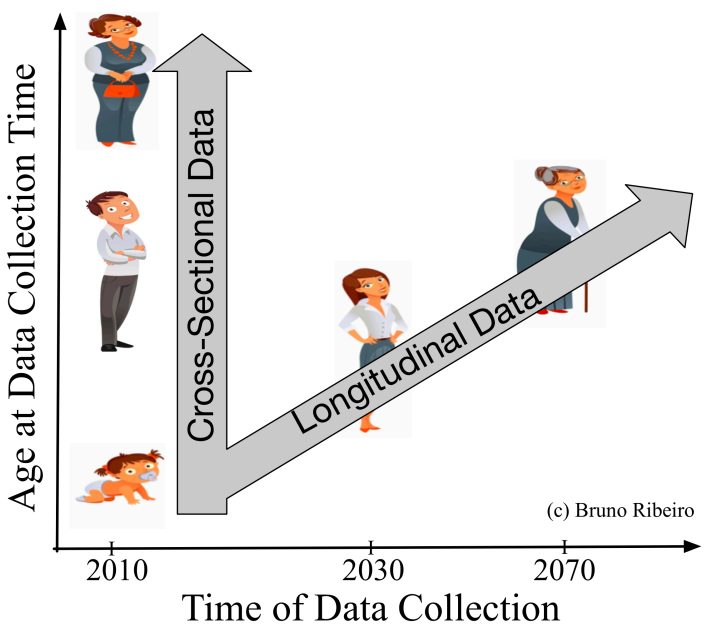
Data representation: Relational/graph data



Data proxies

- Sometimes we do not have access to the data we need
  - Data proxies are datasets used as proxies for the data we do not have

- Example: Longitudinal vs Cross-sectional data
  - Longitudinal data tracks each individual over time
  - Cross-sectional data collects data of diverse individuals at a single point in time



Cross-sectional data can sometimes be used as a proxy for longitudinal data

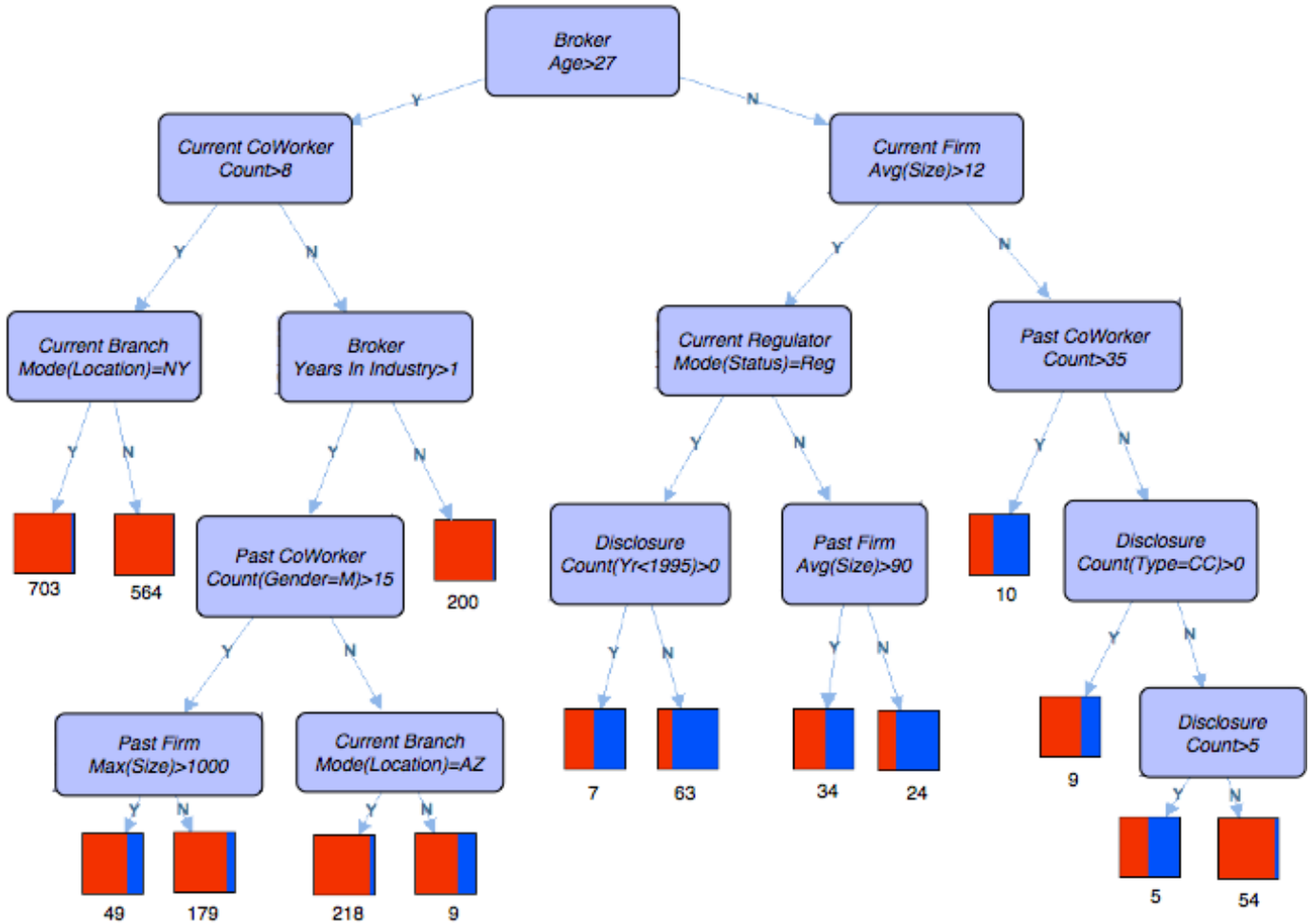
Overview

- 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 IP == DarkWeb then fraud
  - **Conditional probability distribution**
    P(next artist| age, previous artist, geo location)

Knowledge representation: Classification tree



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

Knowledge representation: Regression model

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

- $X = (X_1, X_2, \ldots)$ 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

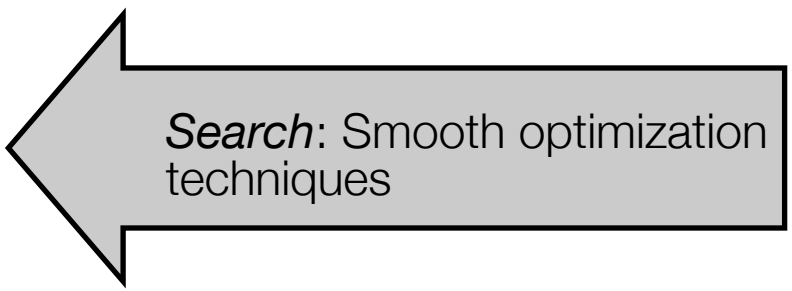
Scoring function

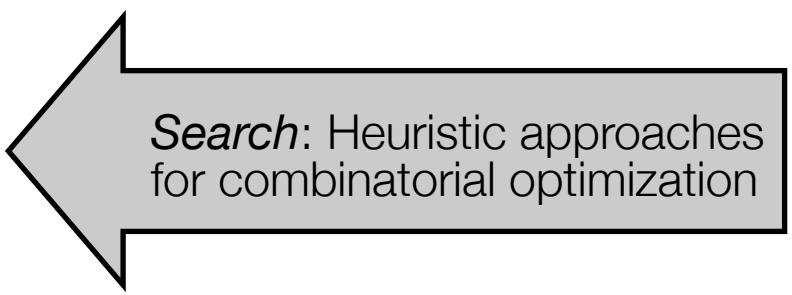
- *A numeric score assigned to each possible model in a search space,* ***given a reference/input dataset***
  - Used to judge the quality of a particular model for the domain

- 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



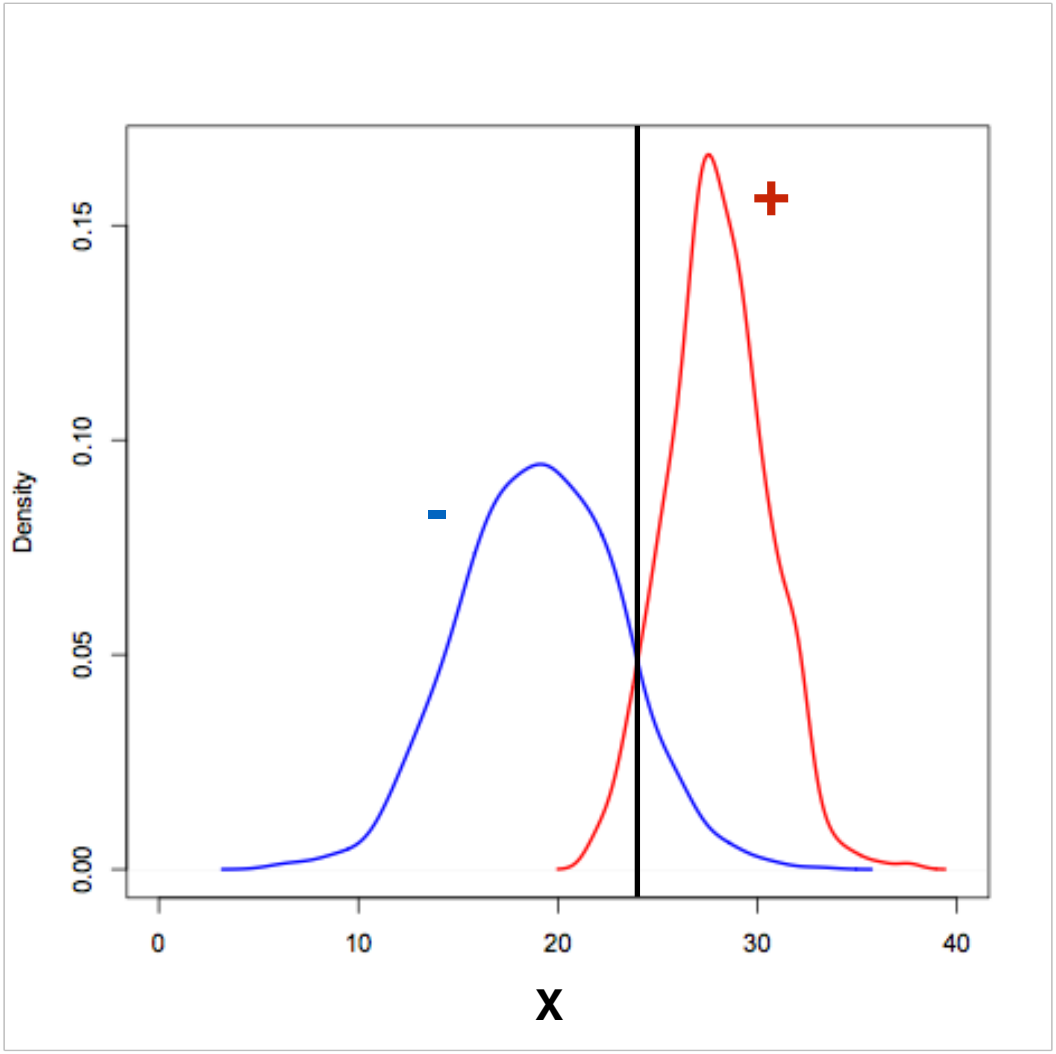


Example learning problem

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

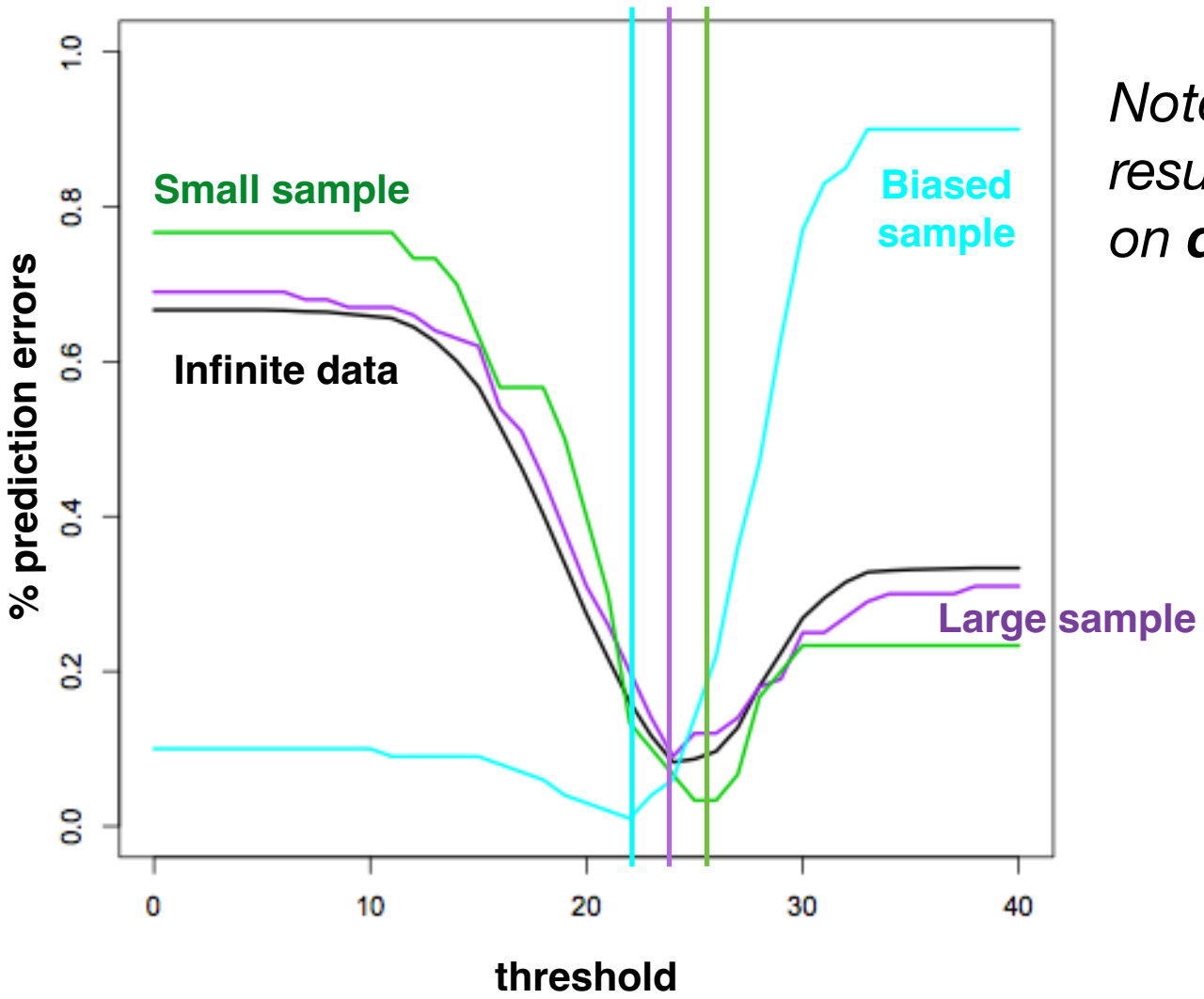
What score function?

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

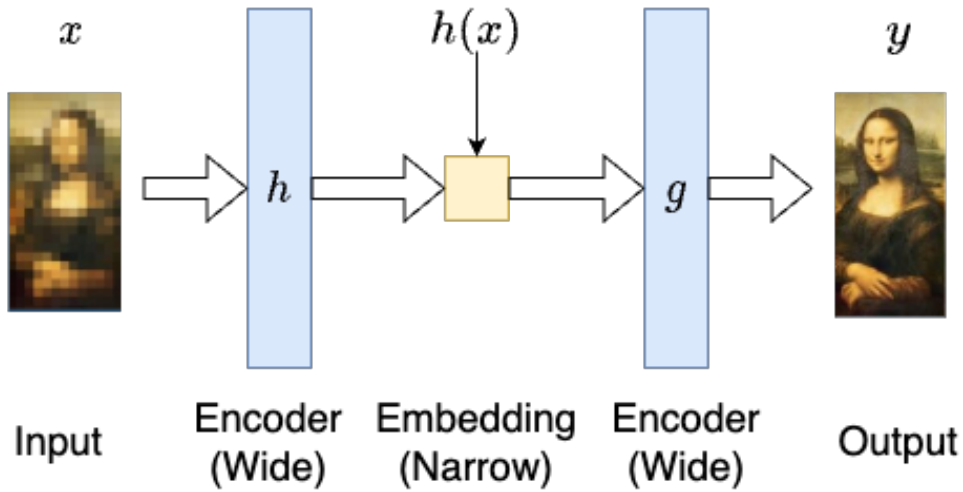
- Task specification
- Data representation
- Knowledge representation
- Learning technique
    - Search + Evaluation
- **Prediction and/or interpretation**

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: Super-resolution

- Task
  - Predict high-resolution image from low-resolution image
- Data
  - 128x256 matrix of pixel intensities
    - In low resolution (x)
    - In high resolution (y)
- Knowledge representation
  - Convolutional neural network
- Learning technique
  - Search for neuron weights that achieve $g(h(x)) \approx y$



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.

Remember the elements of machine learning:

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