DATA MINING

Uncovering Hidden Information in Data

CLASSIC applications:

- increase in sleep depending on the drug,
- pulmonary function modeling by measuring oxygen consumption,
- head length and breadths of brothers,
- classification of the Brahmin, Artisan and Korwa caste based on physical measurements,
- biting flies (genus: *Leptoconops*) data for classification of the two species of flies,
- battery-failure data dependency and regression,
- various financial and market analysis (bankruptcy, stock market prediction, bonds, goods transportation cost data, production cost data, etc.),
- study of love and marriage regarding the relationships and feelings of couples,
- air pollution data classification, college test score classification and prediction, crude oil consumption modeling, closeness between 11 different languages, and so on.

(all of the above were **linear** models, taken from 20 years old statistics books)

TODAYS (primarily **NON-linear**) applications:

Note the following strong fact -> there is no field of human activities today, left untouched by learning from data!!!

Statistical learning is very, very hot nowdays - find patterns, identify, control, make prediction, make decisions, develop models, search, filter, compress, …, and some <u>today's applications are:</u>

- computer graphics, animations,
- image analysis & compression, face detection, face recognition,
- text categorization, media news classification, multimedia (sound video) analysis
- bioinformatics - gene analysis, disease's study
- time series identification - financial, meteorological, hydro,
- biomedicine signals, all possible engineering signal processing
- predictions - sales, TV audience share, investments needed, ..etc.

Data Mining Algorithm

- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Preference – Technique to choose the best model
- Search – Technique to search the data
  - “Query”

Database vs. Data Mining

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Useful introductory example
Netflix

Example: Predicting how a viewer will rate a movie

10% improvement = **1 million dollar prize**

The essence of machine learning:

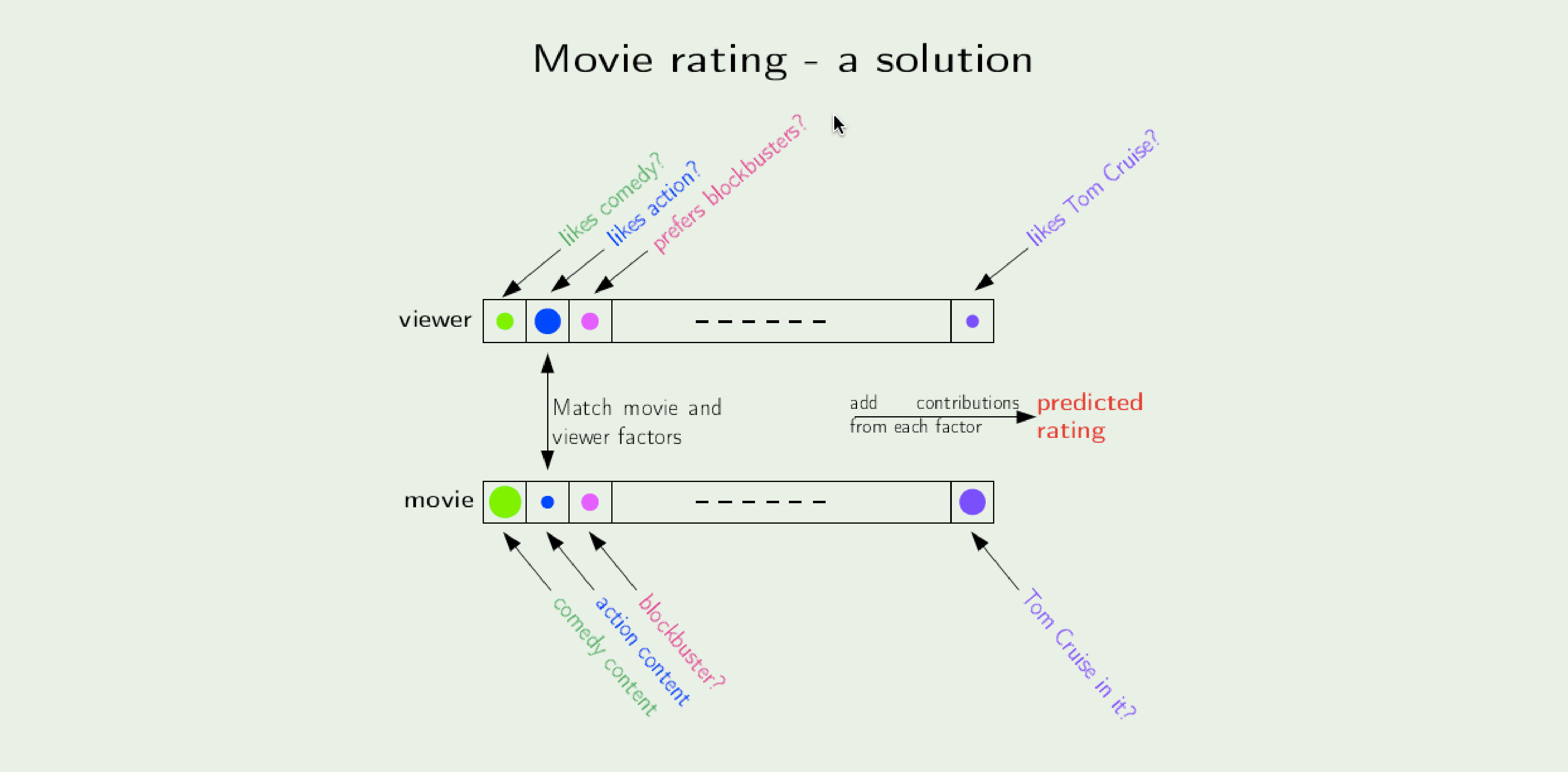
- A pattern exists.
- We cannot pin it down mathematically.
- We have data on it.

The Learning Problem

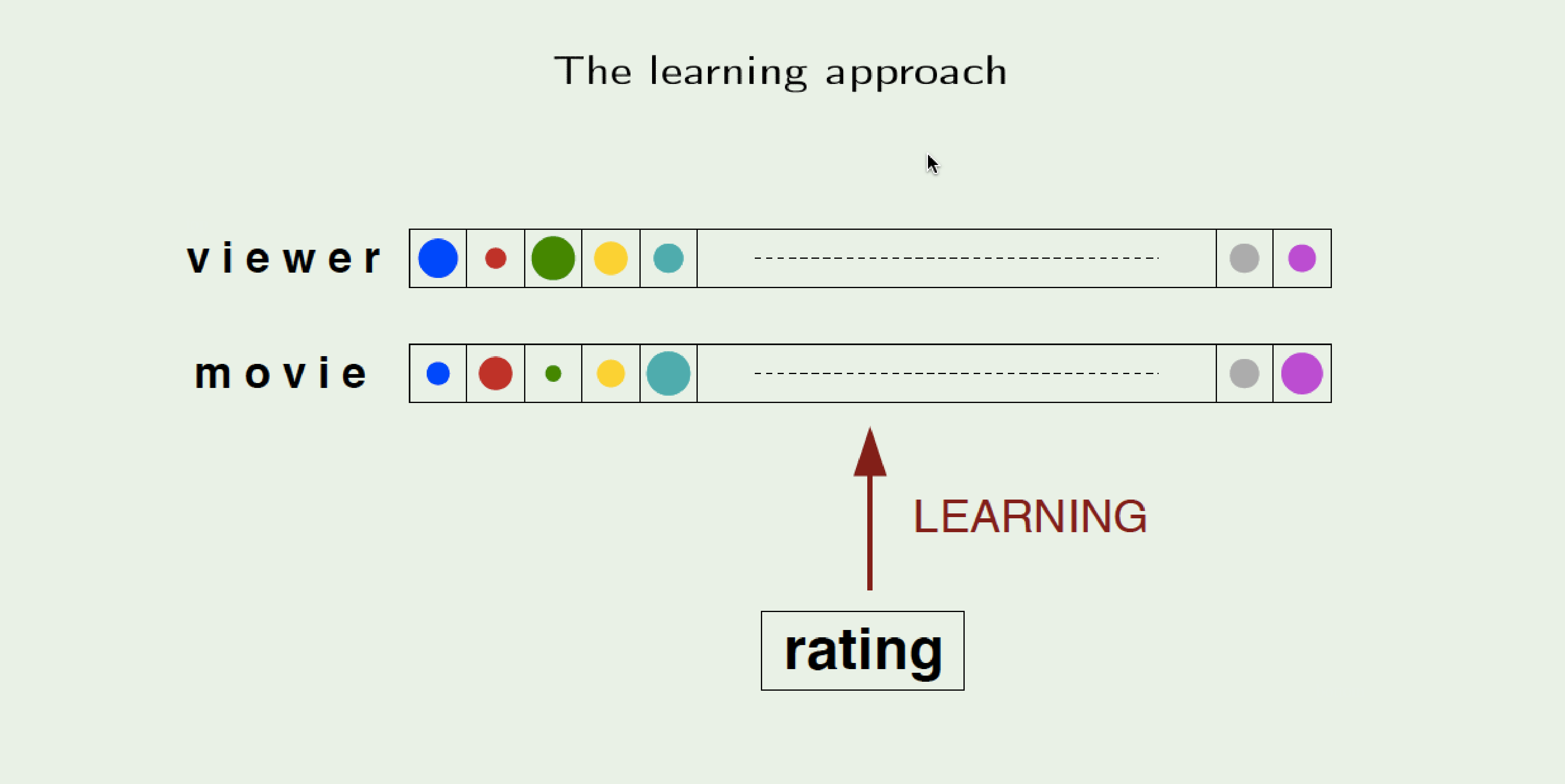
The learning problem - Outline

- Example of machine learning
- Components of Learning
- A simple model
- Types of learning
- Puzzle

Netflix: an approach



The Learning Approach



Components of Learning

Components of learning

Metaphor: Credit approval

Applicant information:

age	23 years
gender	male
annual salary	\$30,000
years in residence	1 year
years in job	1 year
current debt	\$15,000
···	···

Approve credit?

Components of Learning

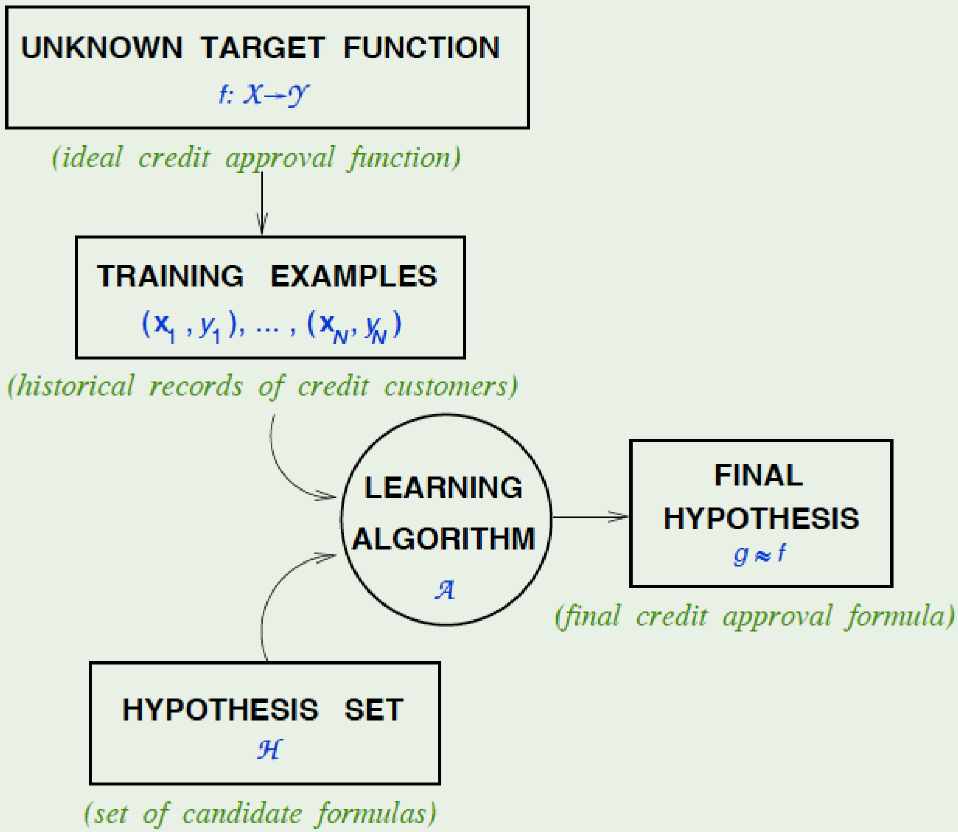
Components of learning

Formalization:

- Input: $\mathbf{x}$ *(customer application)*
- Output: y *(good/bad customer?)*
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ *(ideal credit approval formula)*
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ *(historical records)*

$$\downarrow \quad \downarrow \quad \downarrow$$

- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ *(formula to be used)*


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \rightarrow \mathcal{Y}$
(ideal credit approval function)
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
(historical records of credit customers)
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
(final credit approval formula)
HYPOTHESIS SET
$\mathcal{H}$
(set of candidate formulas)

Solution Components

Solution components

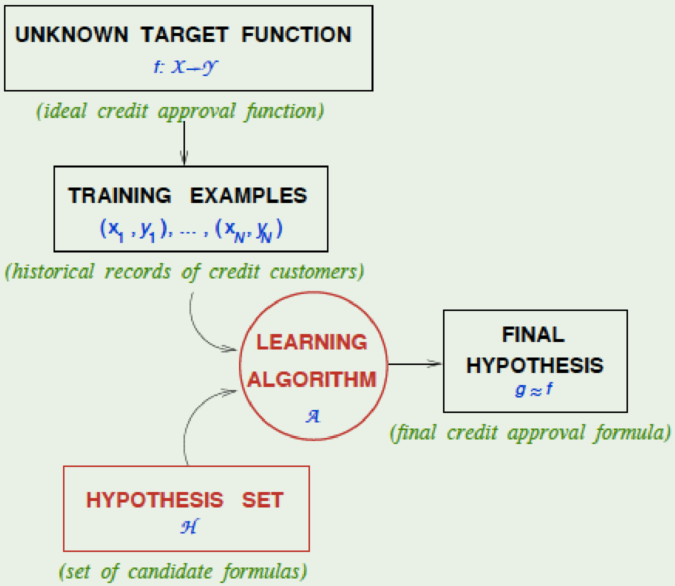
The 2 solution components of the learning problem:

- The Hypothesis Set

  $\mathcal{H} = \{h\} \qquad g \in \mathcal{H}$

- The Learning Algorithm

Together, they are referred to as the *learning model*.



Hypothesis Set

A simple hypothesis set - the 'perceptron'

For input $\mathbf{x} = (x_1, \cdots, x_d)$ 'attributes of a customer'

$$\text{Approve credit if } \sum_{i=1}^{d} w_i x_i > \text{threshold},$$

$$\text{Deny credit if } \sum_{i=1}^{d} w_i x_i < \text{threshold}.$$

This linear formula $h \in \mathcal{H}$ can be written as

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \text{threshold}\right)$$

Linear Separability

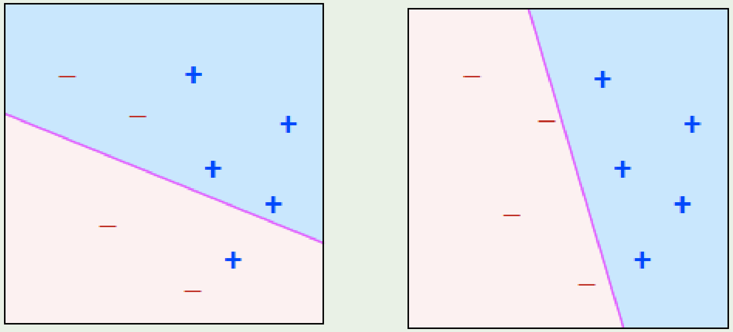
$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i \; x_i\right) + \quad w_0 \quad \right)$$

Introduce an artificial coordinate $x_0 = 1$:

$$h(\mathbf{x}) = \text{sign}\left(\sum_{i=0}^{d} w_i \; x_i\right)$$

In vector form, the perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x})$$



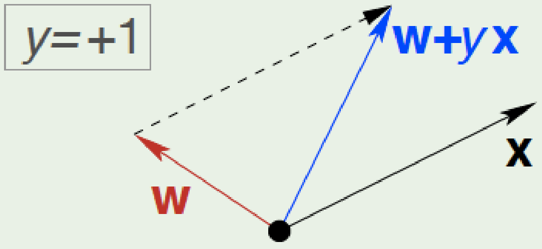
'linearly separable' data

Perceptron Learning Algorithm

A simple learning algorithm - PLA

The perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^\intercal \mathbf{x})$$

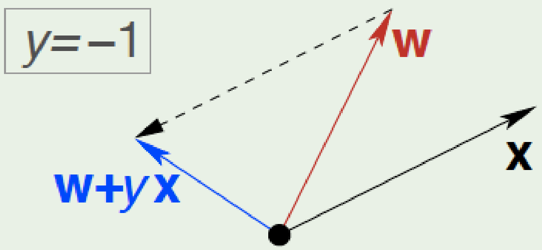


Given the training set:

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

pick a misclassified point:

$$\text{sign}(\mathbf{w}^\intercal \mathbf{x}_n) \neq y_n$$



and update the weight vector:

$$\boxed{\mathbf{w} \leftarrow \mathbf{w} + y_n \mathbf{x}_n}$$

Perceptron Learning Algorithm

Iterations of PLA

- One iteration of the PLA:

$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

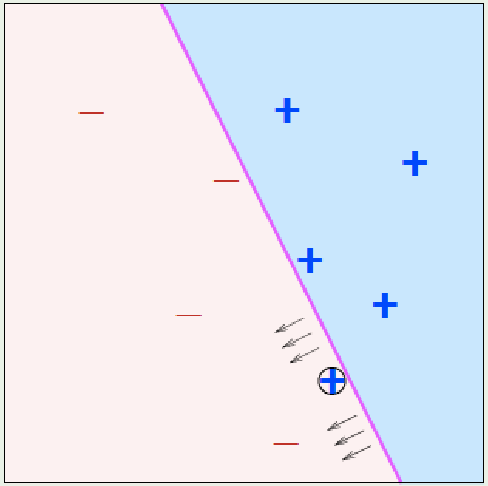
where $(\mathbf{x}, y)$ is a misclassified training point.

- At iteration $t = 1, 2, 3, \cdots$, pick a misclassified point from

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

and run a PLA iteration on it.

- That's it!



Learning Types

Basic premise of learning

"using a set of observations to uncover an underlying process"

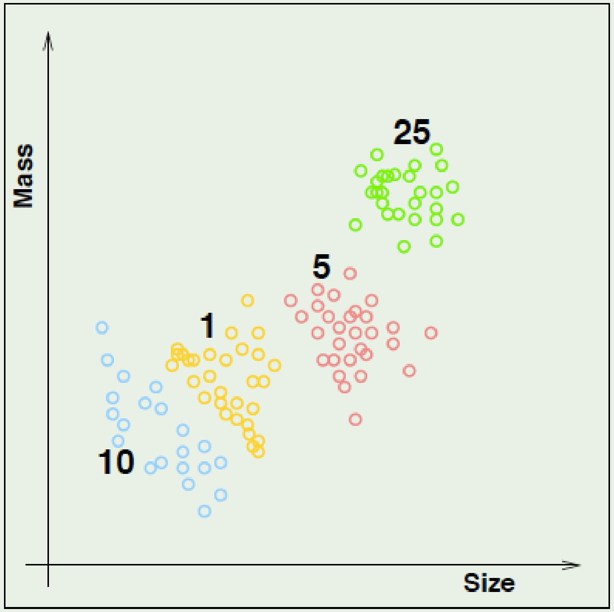
broad premise $\Longrightarrow$ many variations

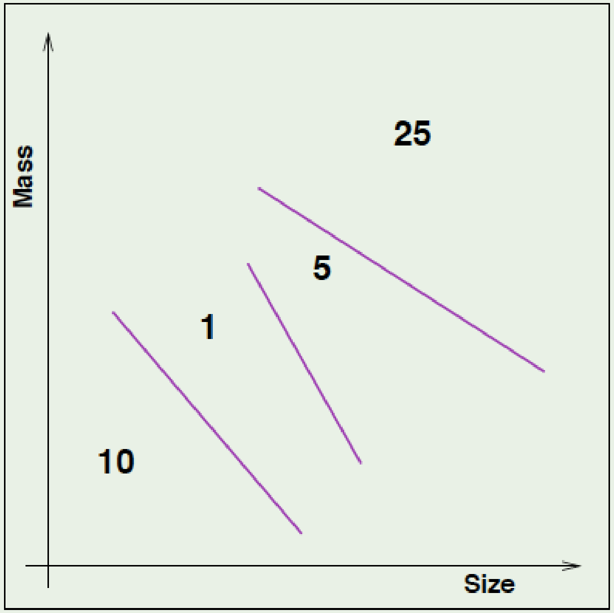
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

Supervised learning

Example from vending machines – **coin recognition**

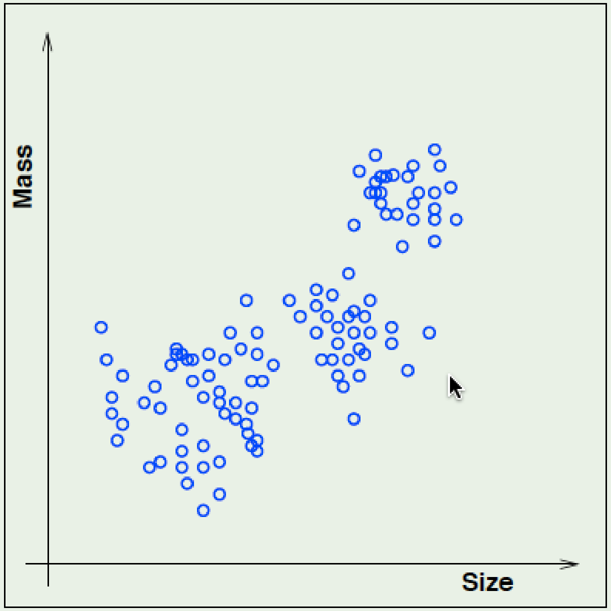




Unsupervised Learning

Unsupervised learning

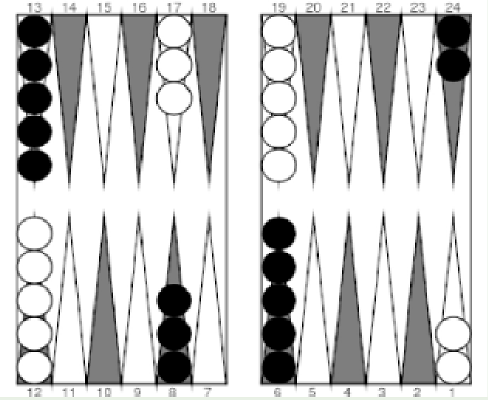
Instead of (input,correct output), we get (input, ?)



Reinforcement learning

Instead of (input,correct output),

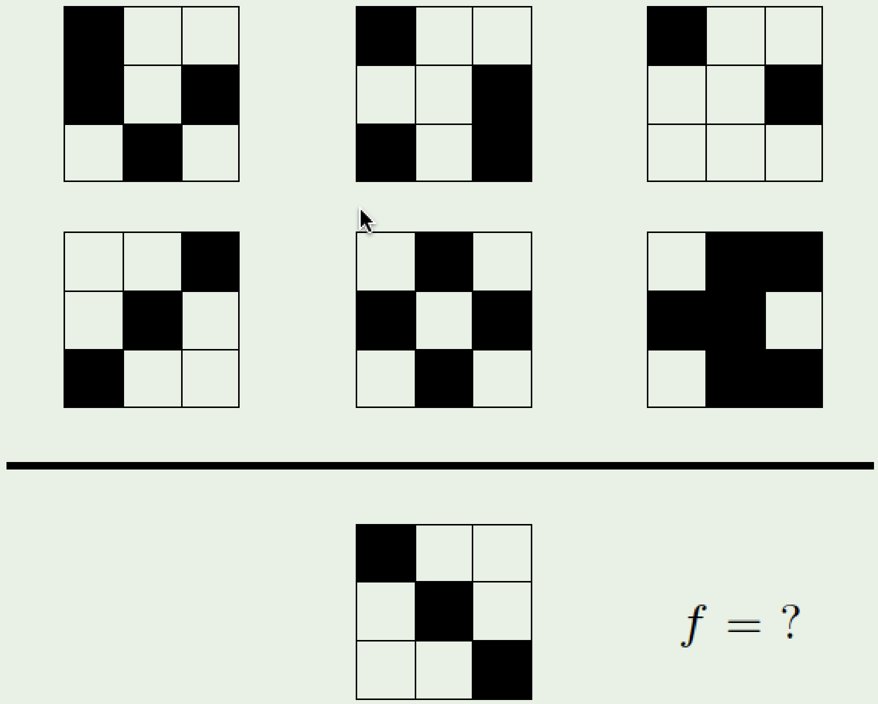
we get (input,*some* output,grade for this output)



The world champion was a neural network!

Small Quiz Question

A Learning puzzle

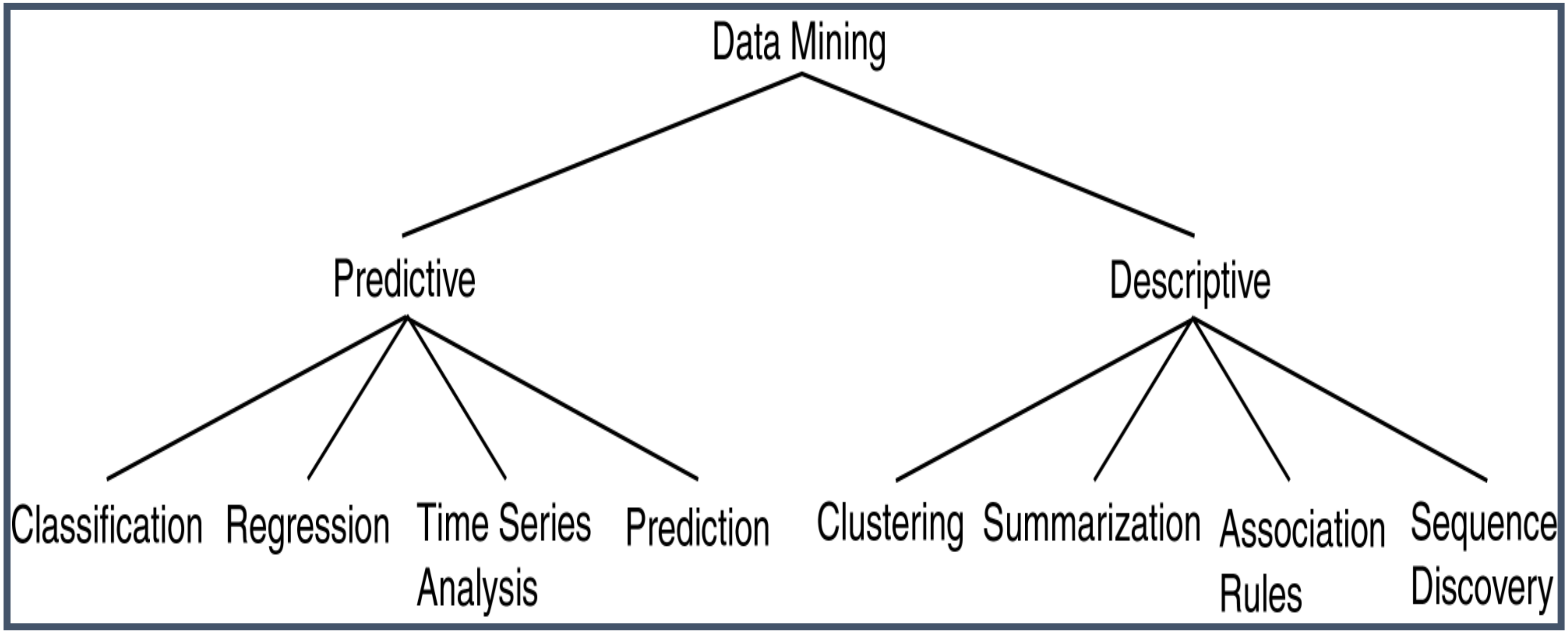


$f = -1$

$f = +1$

$f = ?$

Data Mining Models and Tasks



Few Model Types

- **Linear Regression**
  - Linear Regression
  - Partial Least Squares
    Penalized Models
- **Non Linear Regression**
  - Neural Networks
  - Multivariate Adaptive Regression Splines (MARS)
  - Support Vector Machines
  - K-Nearest Neighbors
- **Regression Trees**
  - Basic
  - Regression Models

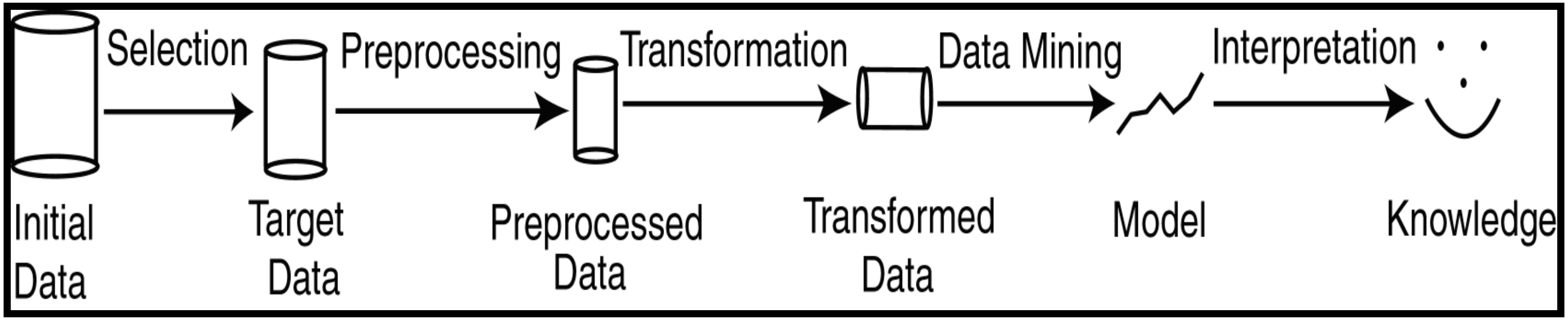
Basic Data Mining Tasks

- ***Classification*** maps data into predefined groups or classes
  - Supervised learning
  - Pattern recognition
  - Prediction
- ***Regression*** is used to map a data item to a real valued prediction variable.
- ***Clustering*** groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning

Basic Data Mining Tasks (cont’d)

- ***Summarization*** maps data into subsets with associated simple descriptions.
  - Characterization
  - Generalization
- ***Link Analysis*** uncovers relationships among data.
  - Affinity Analysis
  - Association Rules
  - Sequential Analysis determines sequential patterns.

Knowledge Discovery Process



Modified from [FPSS96C]

- ***Selection:*** Obtain data from various sources.
- ***Preprocessing:*** Cleanse data.
- ***Transformation:*** Convert to common format. Transform to new format.
- ***Data Mining:*** Obtain desired results.
- ***Interpretation/Evaluation:*** Present results to user in meaningful manner.

Statistics

- Simple descriptive models
- ***Statistical inference:*** generalizing a model created from a sample of the data to the entire dataset.
- ***Exploratory Data Analysis:***
  - Data can actually drive the creation of the model
  - Opposite of traditional statistical view.
- Data mining targeted to more sophisticated user

DM: Many data mining methods come from statistical techniques.

Machine Learning

- ***Machine Learning:*** area of AI that examines how to write programs that can learn.
- Often used in classification and prediction
- ***Supervised Learning:*** learns by example.
- ***Unsupervised Learning:*** learns without knowledge of correct answers.
- Machine learning often deals with small static datasets.

DM: Uses many machine learning techniques.

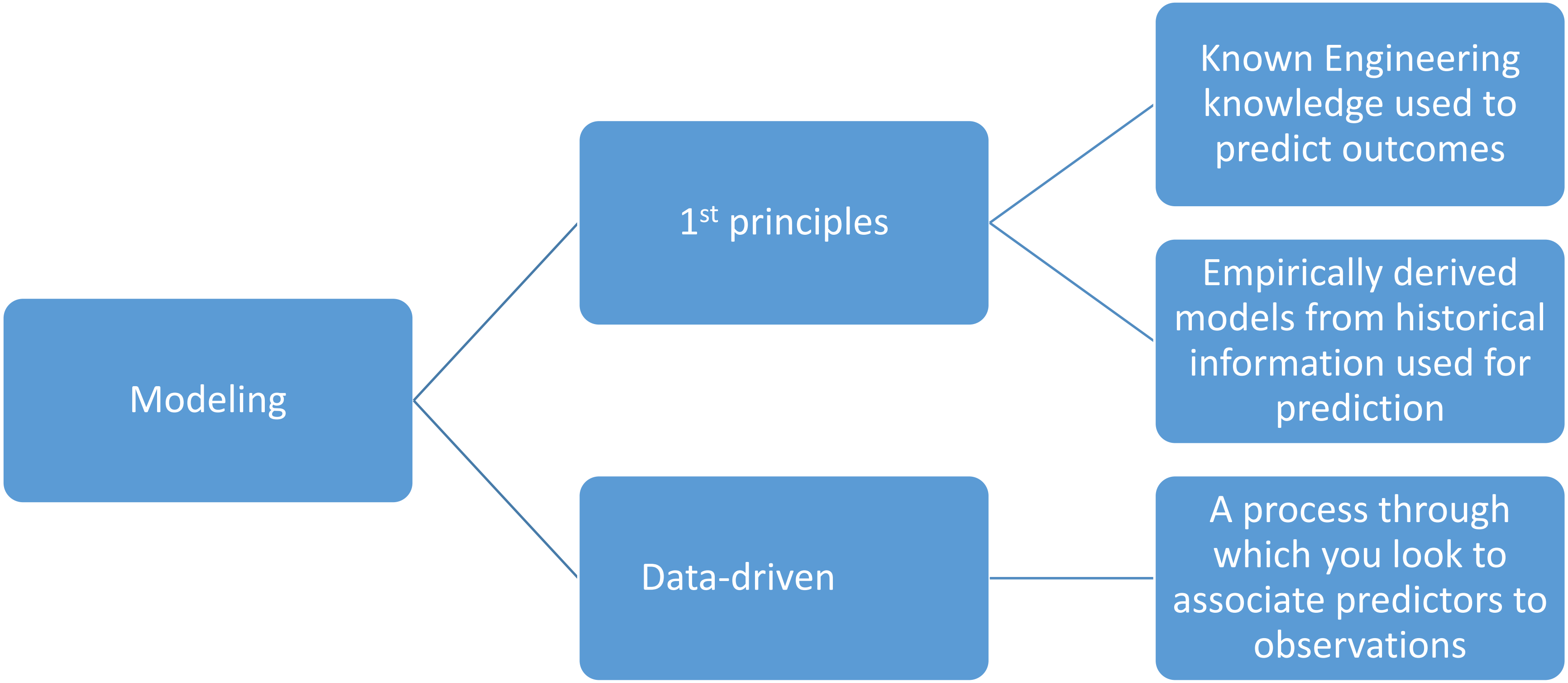
Pattern Matching (Recognition)

- ***Pattern Matching:*** finds occurrences of a predefined pattern in the data.
- Applications include speech recognition, information retrieval, time series analysis.

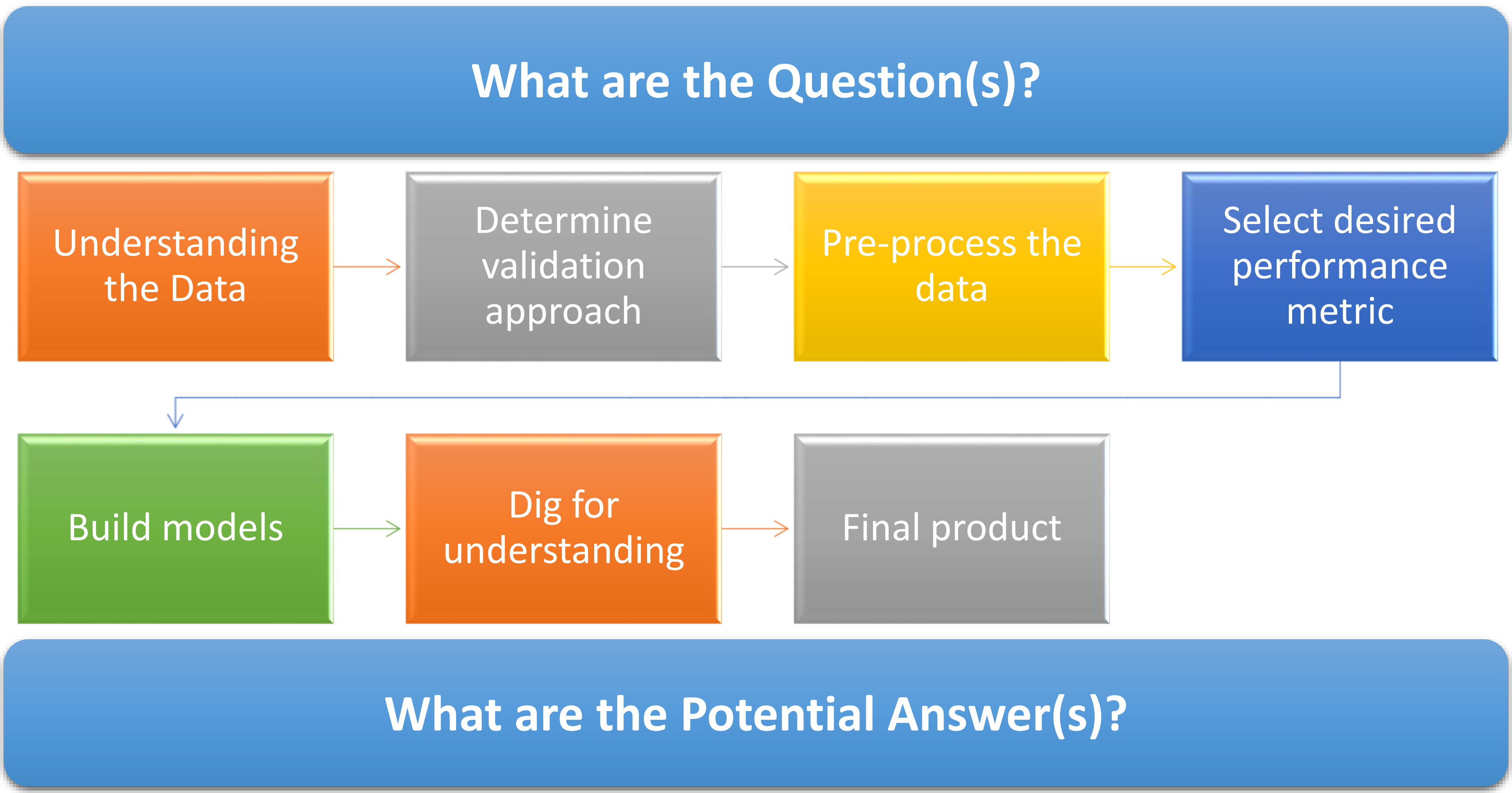
DM: Type of classification.

A process recipe for the creation of models

Types of Modeling



Logical Steps for Data-Driven Model Development



Understanding the Data

- What do we know about the scientific process that generated the data
- Investigate the data
  - Anything unusual in the data table
  - What is the distribution of the response?
  - Are there obvious, important univariate relationships?
    - Pairwise scatterplots of predictors versus response and predictors versus predictors
  - How much missing data or limit of detection data is present

2

Determine validation approach

- How much data do we have?
  - How many samples and how many predictors?
- Do we need an independent test set? Do we have enough data to have an independent test set?
  - Need to determine whether or note we have a test set before pre-processing the predictors
- Which cross-validation method should we use?
  - K-fold, repeated k-fold, leave-group-out, bootstrap?
  - Each of these has different computational costs and become noticeable as data size increases

Pre-process the data

- Pre-filter samples and predictors on missingness (if a column has more than 30% missing data – shall we take it out?)
  - remove samples with too many missing predictor values
  - Remove predictors with too many missing sample values
- Transform and impute
  - Transform predictors to resolve skewness (Box-Cox)
  - Center and scale
  - Impute missing data
- Post-filter uninformative

Select desired performance metric

- What are we optimizing?
  - R2, RMSE, Accuracy, ROC, Sensitivity, specificity...
- What is an optimal value for this problem
  - Do we know the measurement error of the response

- Multi-objective approach?

Build models

- Build sentinel models
  - Choose an interpretable, simple model and a highly complex, uninterpretable model
  - Tune each model, and assess model performance
- Do the models have significantly different predictive performance?
  - if not, then the interpretable model maybe sufficient
- If there is a sufficient range of predictive performance, then build lots of models
  - Linear, non-linear, tree based, etc
- Gather CV performance metrics
  - Do some models perform better than others

Dig for understanding

- Compute variable of importance to understand what predictors are important to each model
  - Are certain predictors common across most?