Introduction to Data Mining and Machine Learning

Dantong Yu

Associate Professor

THE MAGAZINE

October 2012

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (91)

Back in the 1990s, computer engineer and Wall Street "quant" were the hot occupations in business. Today data scientists are the hires firms are competing to make. As companies wrestle with unprecedented volumes and types of information, demand for these experts has raced well ahead of supply. Indeed, Greylock Partners, the VC firm that backed Facebook and LinkedIn, is so worried about the shortage of data scientists that it has a recruiting team dedicated to channeling them to the businesses in its portfolio.

Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions. They find the story buried in the data and communicate it. And they don't just deliver reports: They get at the questions at the heart of problems and devise creative approaches to them. One data scientist who was studying a fraud problem, for example, realized it was analogous to a type of DNA sequencing problem. Bringing those disparate worlds together, he crafted a solution that dramatically reduced fraud losses.

TOP MAGAZINE ARTICLES

24 HOURS 7 DAYS 30 DAYS

1. Lean Knowledge Work
2. How Netflix Reinvented HR
3. The Five Competitive Forces That Shape Strategy
4. The Big Lie of Strategic Planning
5. Smart Rules: Six Ways to Get People to Solve Problems Without You
6. Find the Coaching in Criticism
7. Salman Khan

All Most Popular »

HBR.ORG ON FACEBOOK

Great Leaders Don't Need Experience

Value-added in predictive science analysis

- Hurricane Irma
- What is needed in addition to Wood board, bottle water.
- What about strawberry poptarts, under-ware?
- Did you hear Signet Bank in Virginia?
- Customer Churn

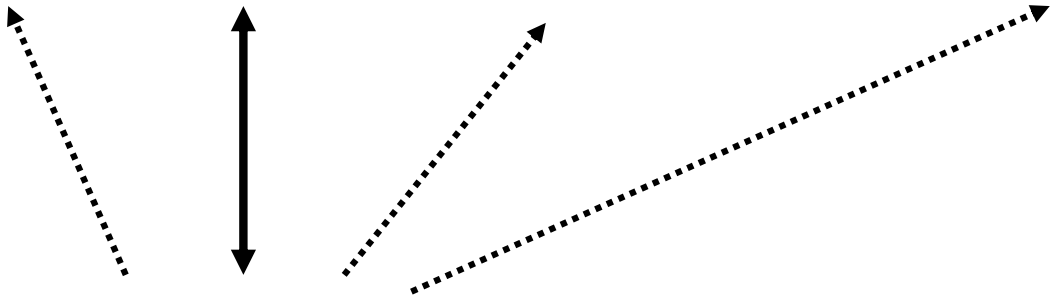
Data Mining and Machine Learning

- A set of principles, concepts, and techniques that structure thinking and analysis of data
- Extracts useful information and knowledge from large volumes of data by following a process with reasonably well defined steps
- Changes the way you think about data and its role in business
    - A common understanding: you have valuable data, you have an edge in competition
    - Data is asset once you can extract non-obvious insight from data. It is an asset playing the vital role.

How do you want that data?

Machine Learning is a process

science + craft + creativity + common sense



a new business *process*

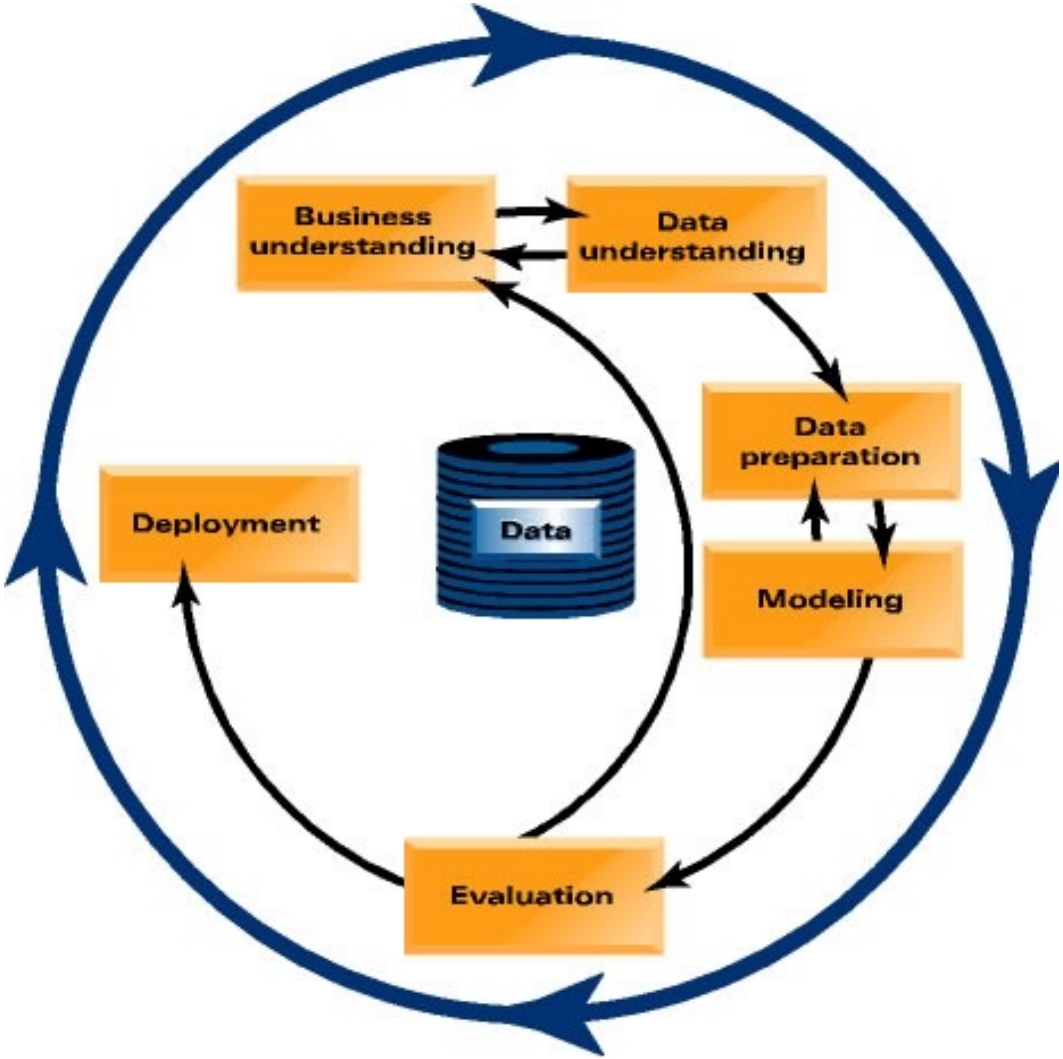
Data Opportunities

- Volume of data
- Variety of data
- Powerful computers
- Better algorithms

General Learning Goals for Data Scientists

- Approach business problems data-analytically
- Interact competently on the topic of data mining for business intelligence
- Hands-on experience mining data
  - Every good chemist has to be a competent lab technician
  - A good data scientist (business) must be proficient with certain languages, tools and software system.

Data Mining/Machine Learning Process



- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

From Data Mining to Machine Learning

- **Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science ➔ Machine Learning**

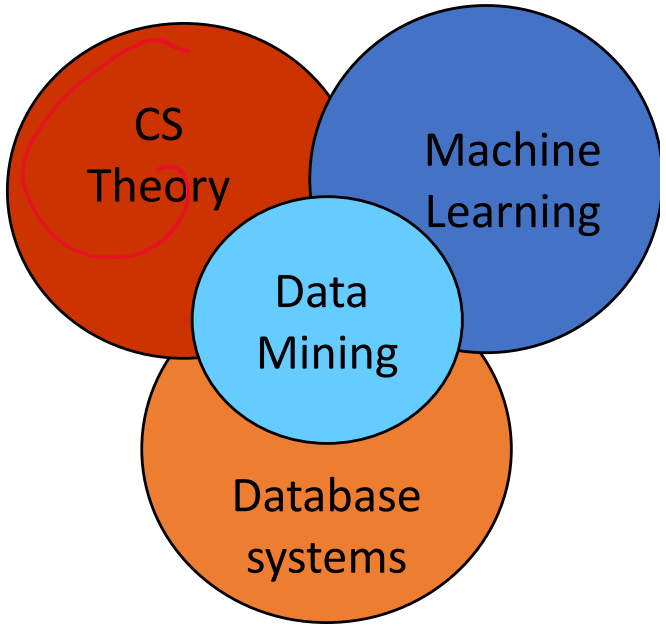
Data mining overlaps with:

- **Databases:** Large-scale data, simple queries
- **Machine learning:** Complex models (Deep Neural Networks)
- **CS Theory:** (Randomized) Algorithms

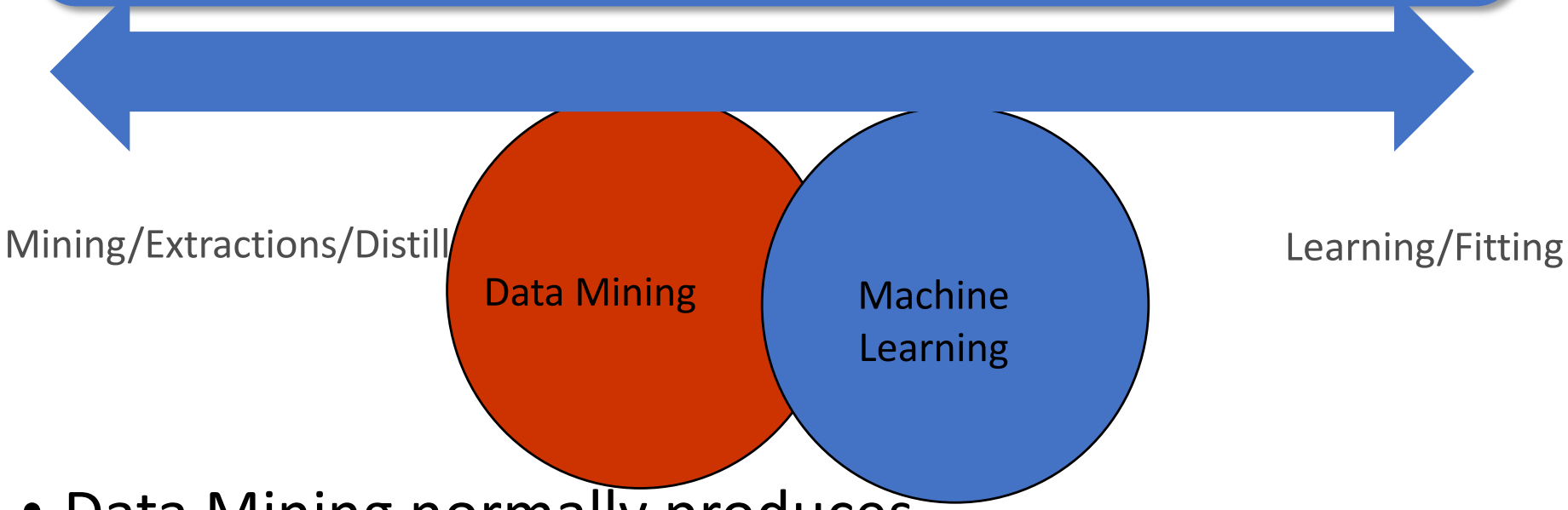
Different cultures:

- To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
  - **Result is the query answer**
- To a ML person, data-mining process is the **inference of models**
  - **Result is the parameters of the model**

In this class we will do both!



Different Between Data Mining and Machine Learning



- Data Mining normally produces models/rules/patterns that tender to be self-explainable
- Machine Learning produces a model that best fits with data (neural networks) and might be a black box.

Machine Learning

- Supervised v.s. Unsupervised
- Unsupervised Algorithm
- Cluster Algorithm
- Principal Component Analysis linear
- Spectrum analysis and Manifold learning non linear

$$a \cdot b$$

$$\log a \cdot \log b$$

Learning Objectives

- The basic concept of machine learning predictive models
- The notion of finding informative attributes
- Techniques of finding information attributes

MegaTelCo: Predicting Customer Churn

- You just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the US
- They are having a major problem with customer retention in their wireless business
- In the mid-Atlantic region, 20% of cell phone customers leave when their contracts expire. Communications companies are now engaged in battles to attract each other's customers while retaining their own
- Marketing has already designed a special retention offer
- Your task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to solve the problem

MegaTelCo: Predicting Customer Churn

- What data you might use?
- How would they be used?
- How should MegaTelCo choose a set of customers to receive their offer in order to best reduce churn for a particular incentive budget?

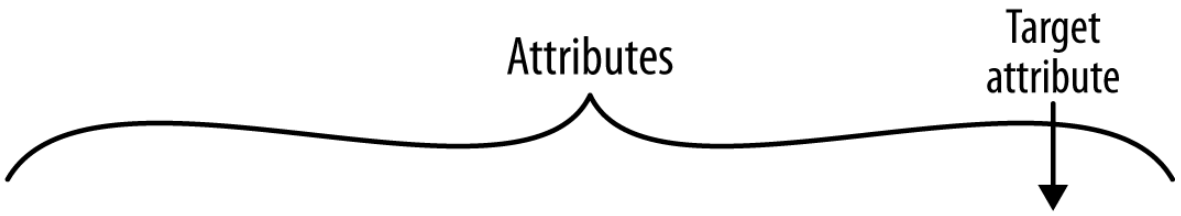
Terminology

- Model:
  - A simplified representation of reality created to serve a purpose
- Predictive Model:
  - A formula for estimating the unknown value of interest: **the target**
    - The formula can be mathematical, logical statement (e.g., rule), etc.
- Prediction:
  - Estimate an unknown value (i.e. the target)
- Instance / example:
  - Represents a fact or a data point
  - Described by a set of **attributes** (fields, columns, variables, or features)

Terminology

- Model induction:
  - The creation of **models** from data
- Training data:
  - The input data for the induction algorithm
  - Consists of attributes
  - Which attributes (features) are informative, correlates with what you want to know (targets)?

Terminology



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

What is a model?

A *simplified** *representation* of reality created for a *specific purpose*

**based on some assumptions*

- *Examples: map, prototype, math formula, etc.*

- *Machine Learning example:*

  "formula" for predicting the probability of customer attrition at contract expiration

  → *"classification model" or "class-probability estimation model"*

Feature Types

- Numeric: anything that has some order
  - Numbers (that mean numbers)
  - Dates (that look like numbers …)
  - Dimension of 1
- Categorical: stuff that does not have an order
  - Binary
  - Text
  - Dimension = number of possible values (-1)
- Food for thought: Names, Ratings (good, excellent, diamond), Standard Industry Classification (https://en.wikipedia.org/wiki/Standard_Industrial_Classification)

Dimensionality of the data?

Attributes / Features

Name	Balance	Age	Default
Mike	$123,000	30	Yes
Mary	$51,100	40	Yes
Bill	$68,000	55	No
Jim	$74,000	46	No
Mark	$23,000	47	Yes
Anne	$100,000	49	No

- **Dimensionality of a dataset** is the sum of the dimensions of the features
  - The sum of the number of numeric features and ~ the number of values of categorical features

Common Data Mining Tasks

- Classification and class probability estimation
  - How likely is this consumer to respond to our campaign?
- Regression
  - How much will she use the service?
- Similarity Matching
  - Can we find consumers similar to my best customers?
  - Serves as the base for many other data mining tasks.
- Clustering
  - Do my customers form natural groups?

Common Data Mining Tasks

- Co-occurrence Grouping
  - Also known as frequent item set mining, association rule discovery, and market-basket analysis
  - What items are commonly purchased together?
- Profiling (behavior description)
  - What does “normal behavior” look like? (for example, as baseline to detect fraud)
  - What is characteristic of a population, subgroup, and even individual
  - How do you describe a credit card fraud?
- Data Reduction
  - Explicit Feature Selections
  - Which latent dimensions describe the consumer taste preferences?

Common Data Mining Tasks

- Link Prediction
  - Since John and Jane share 2 friends, should John become Jane's friend?
- Causal Modeling
  - Why are my customers leaving?
  - Target Ad: is my treatment (targeting) effective, or is my model that identify whom will purchase anyway (regardless targeted or not)?
  - Assumption: For example, Viral market: consumers actually influence each other.

Supervised versus Unsupervised Methods

- "How do our customers naturally fall into different groups?"
    - No guarantee that the results are meaningful or will be useful for any particular purpose
- "Can we find groups of customers who have particularly high likelihoods of canceling their service soon after contracts expire?"
    - **A specific purpose**
    - Much more useful results (usually)
    - Different techniques
    - **Requires data with the target**
        - The individual's label

Supervised Data Mining & Predictive Modeling

- Is there a specific, quantifiable **target** that we are interested in or trying to predict?
    - Think about the decision
- Do we have data on this target?
    - Do we have enough data on this target?
        - Need a min ~500 of each type of classification
- The result of supervised data mining is a model that predicts some quantity
- A model can either be used to predict or to understand

Recap: Supervised Machine Learning

- Classification
  - Categorical target
    - Often binary
  - Includes “class probability estimation”
- Regression
  - Numeric target

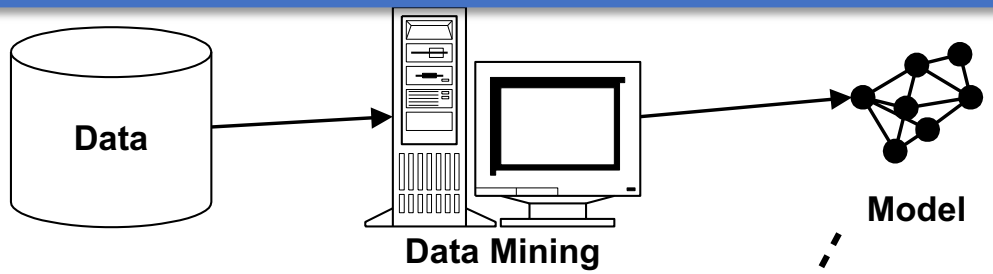
Subclasses of Supervised Data Mining

- "Will this customer purchase service $S1$"
  - Classification problem
    - Binary target (the customer either purchases or does not)
- "Which service package ($S1$, $S2$, or none) will a customer likely purchase?"
  - Classification problem
    - Three-valued target
- "How much will this customer use the service?"
  - Regression problem
    - Numeric target
    - Target variable: amount of usage per customer

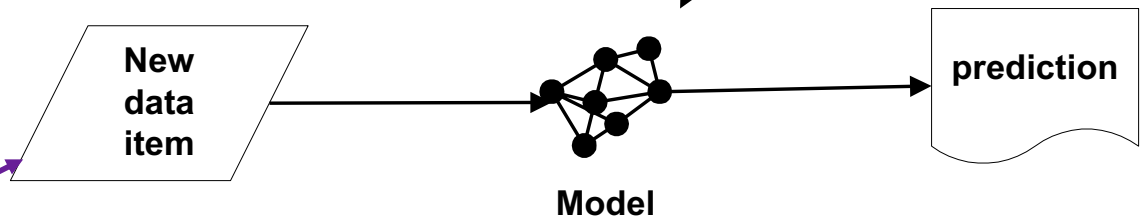
Common Data Mining Tasks

Task	Supervised Methods	Unsupervised Methods
Classification	✓	
Regression	✓	
Causal Modeling	✓	
Similarity Matching	✓	✓
Link Prediction	✓	✓
Data Reduction	✓	✓
Clustering		✓
Co-occurrence Grouping		✓
Profiling		✓

Machine Learning Training versus Use of the Model



"Training" data have all values specified



New data item has some value unknown (e.g., will she leave?)

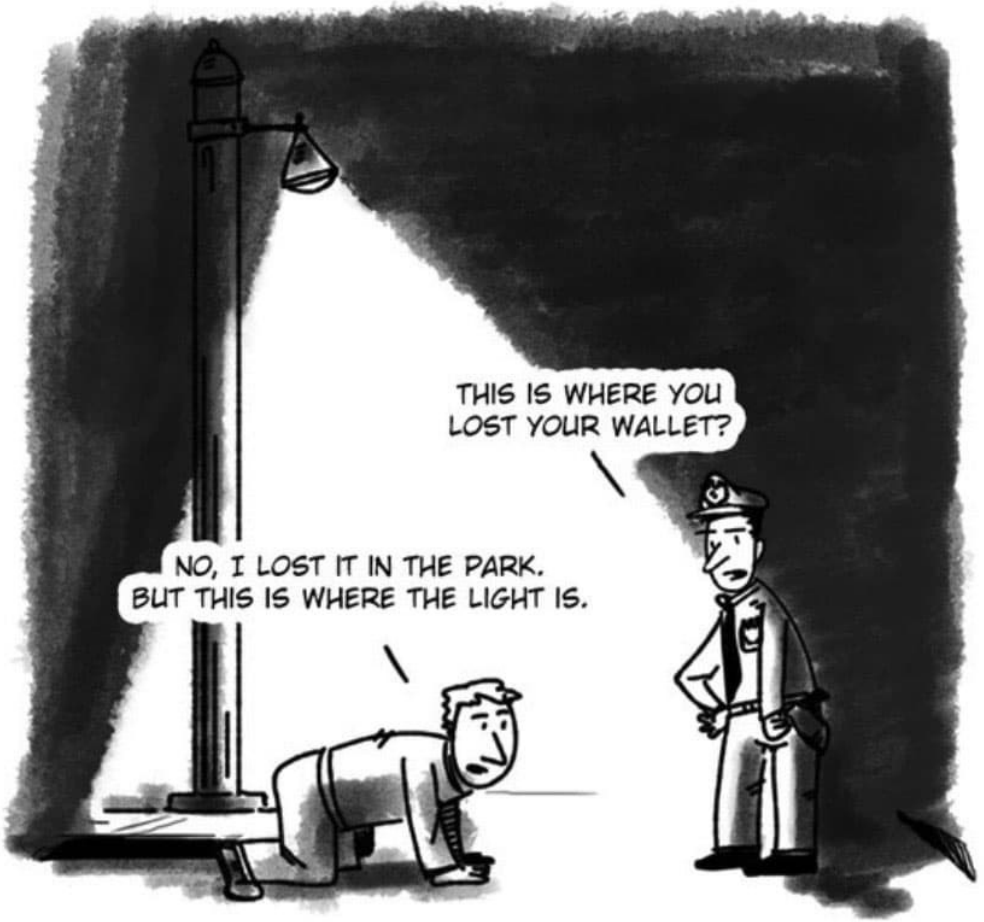
Classical Pitfalls in DM setup

- The training data is NOT consistent with the use

- Bad sample
- Bad features

Sample: “Looking Under the Streetlight”

- Target Proxy
  - I do not see if a person after seeing an ad bought the book, so lets model clicks …
- Sample Proxy
  - I want to run a campaign in Spain but only have data on US customers



Sample: “Survivorship issues”

- Lending club wants to have a model to take over the screening process that selects applications and deny those that are likely to default
- Data of past loans and the outcomes are provided

- Bad:
    - Use the data they currently have to predict default

Sample: Different Sources

- Things go really bad if the positive and negative are treated differently
- Looking for drivers of diabetes: how do you assemble the training data?

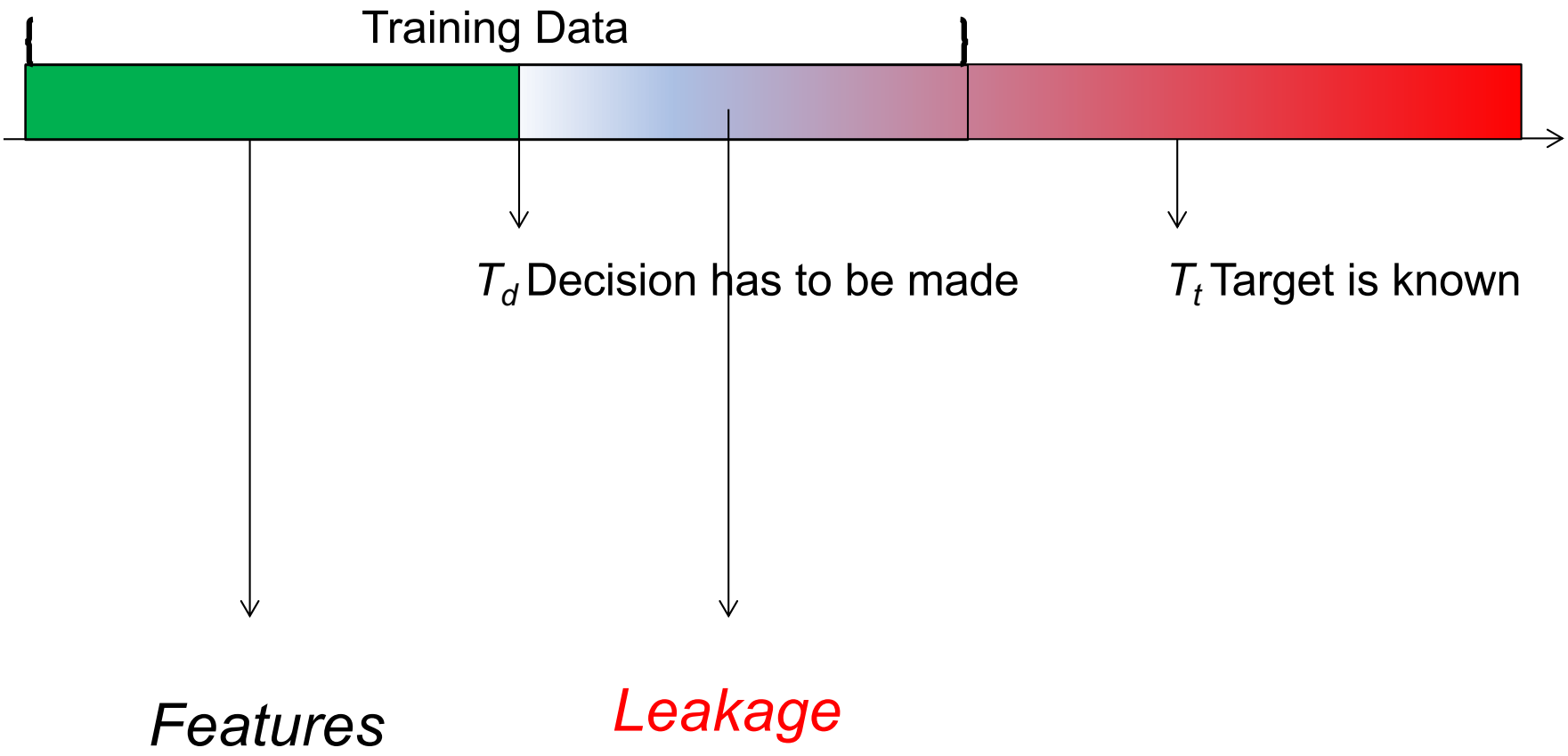
- Bad:
  - Go to a specialized hospital and get records from people treated for diabetes
  - Go somewhere else to get records for healthy people

Summary on bad habits

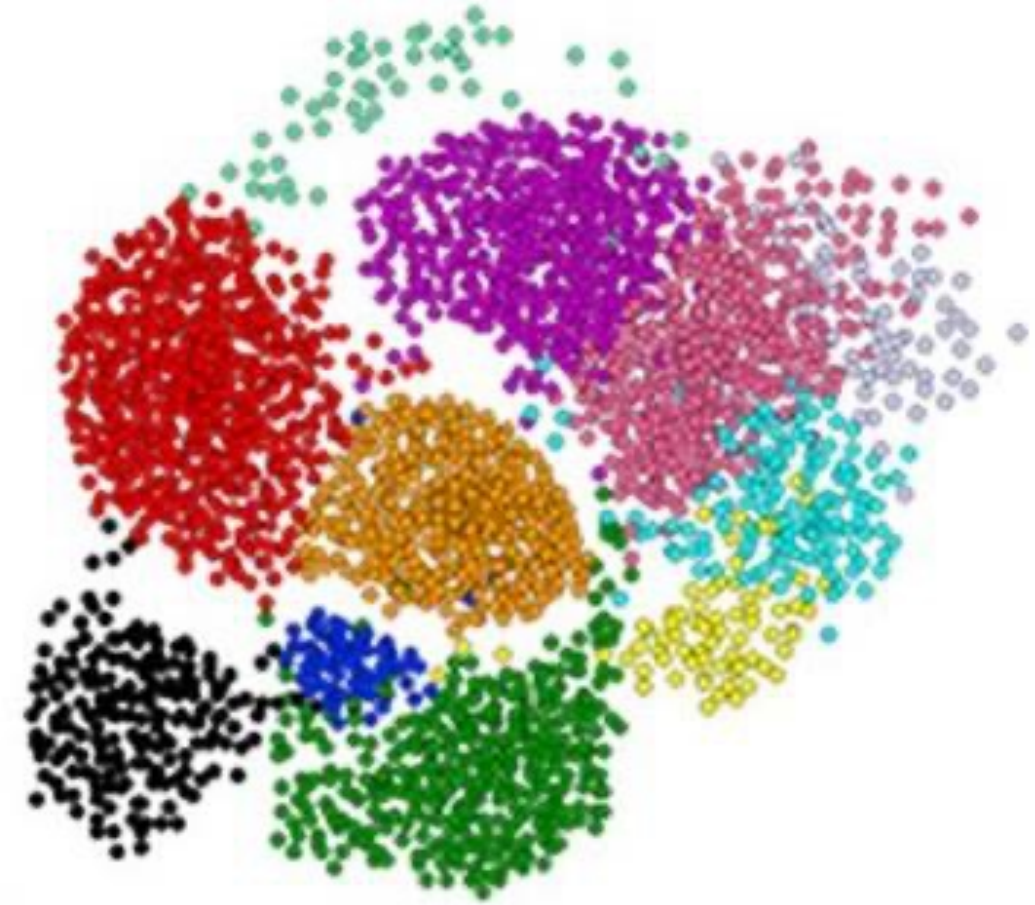
- You are missing all the applications that were turned down already
- The sick people came from a very artificial subset
- Your target is NOT really your target
- No way of telling how the model will perform
    - No way of testing either

- The training sample should be as similar as possible to the USE data

Digression on features: It is all about the timing in use!

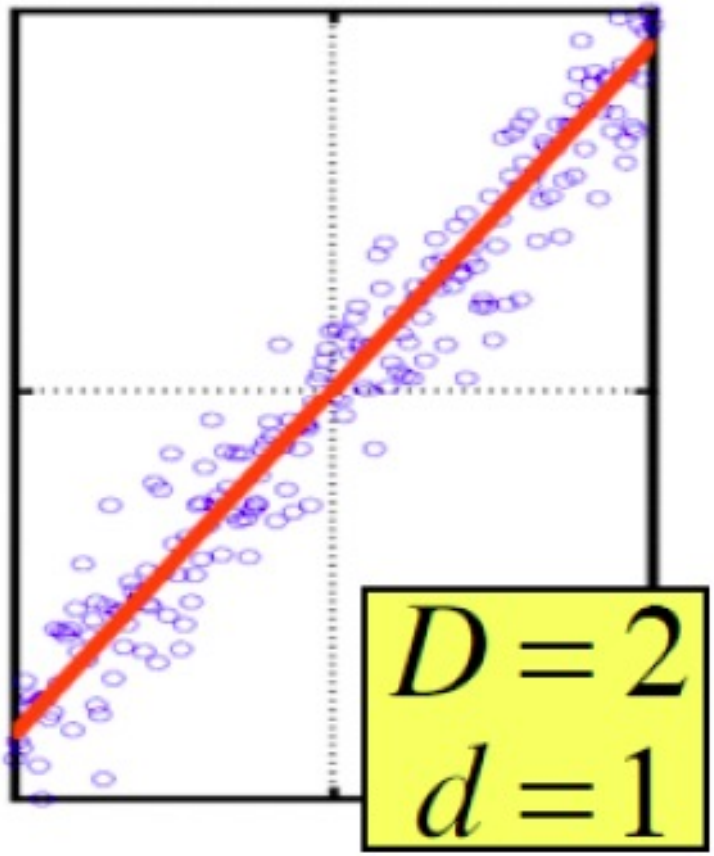


Clustering is a hard problem!



Dimensionality Reduction

- **Goal of dimensionality reduction is to discover the axis of data!**

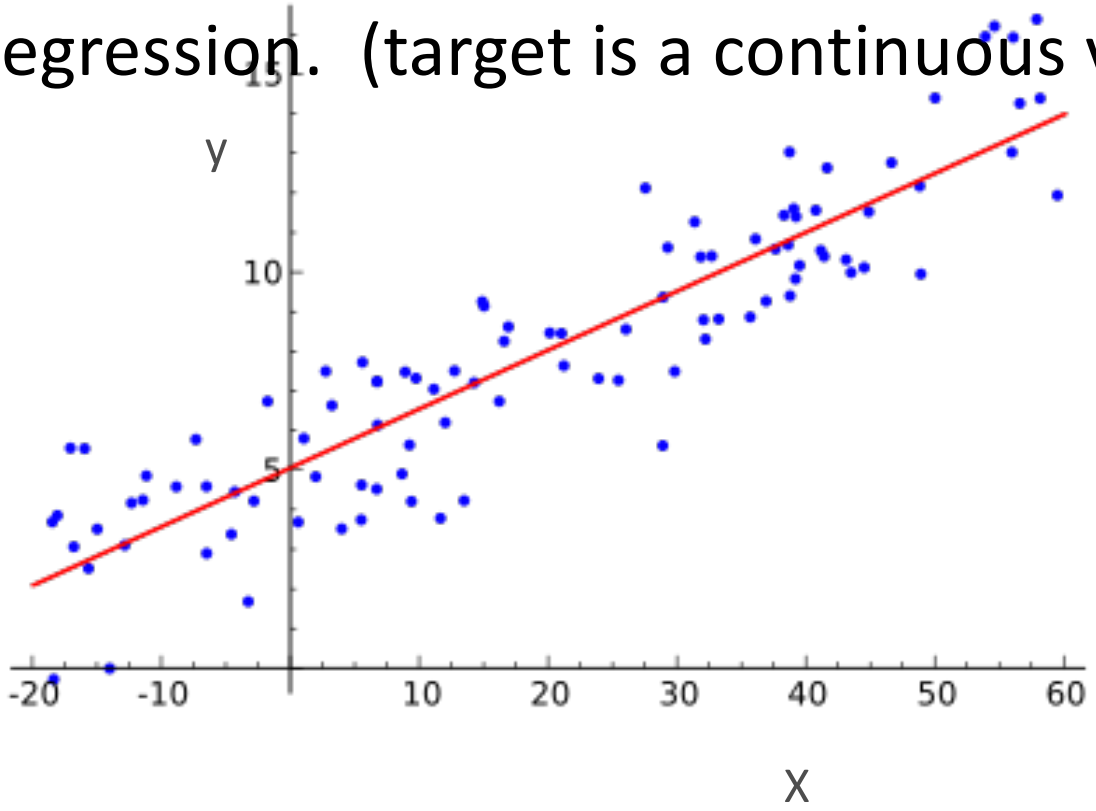


Rather than representing every point with 2 coordinates we represent each point with 1 coordinate (corresponding to the position of the point on the red line).

By doing this we incur a bit of **error** as the points do not exactly lie on the line

Supervised Learning

- Linear Regression. (target is a continuous value)

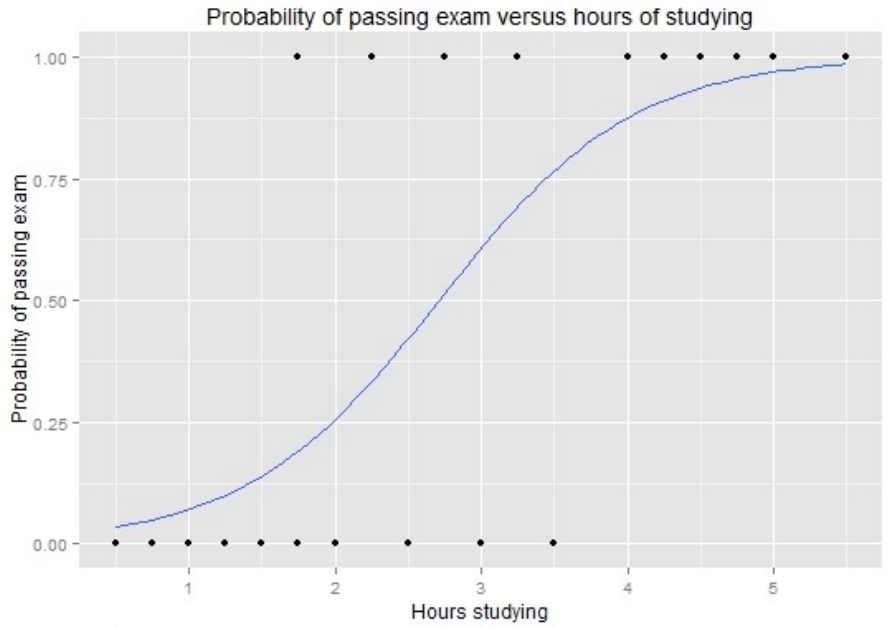


$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

Classification (target is a label)

- Logistic Regression. (target is a discrete value)

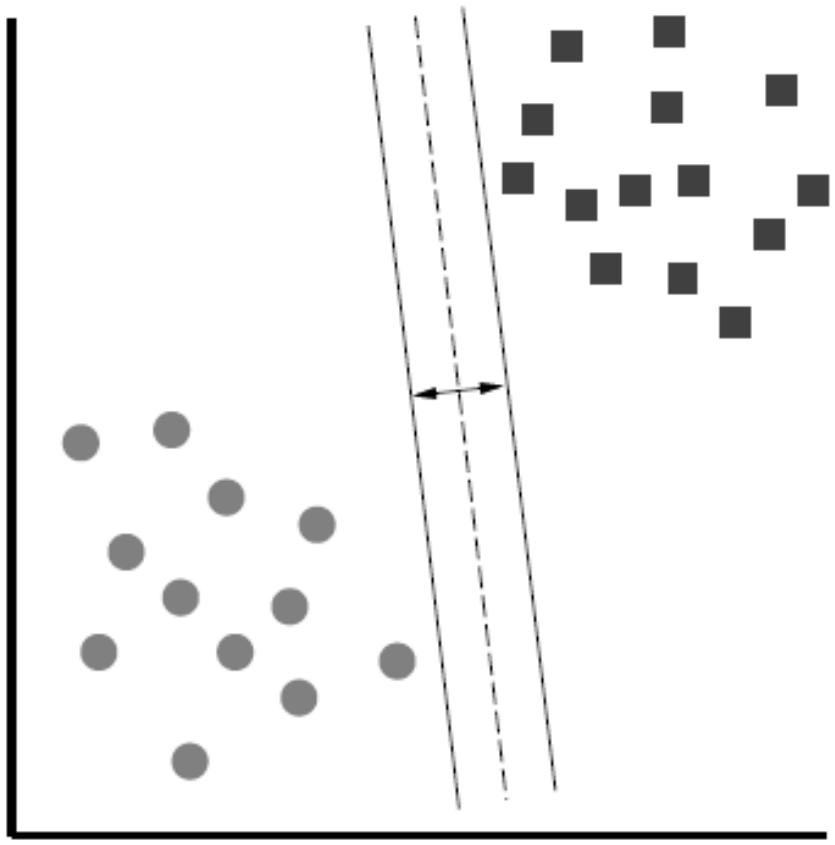
y

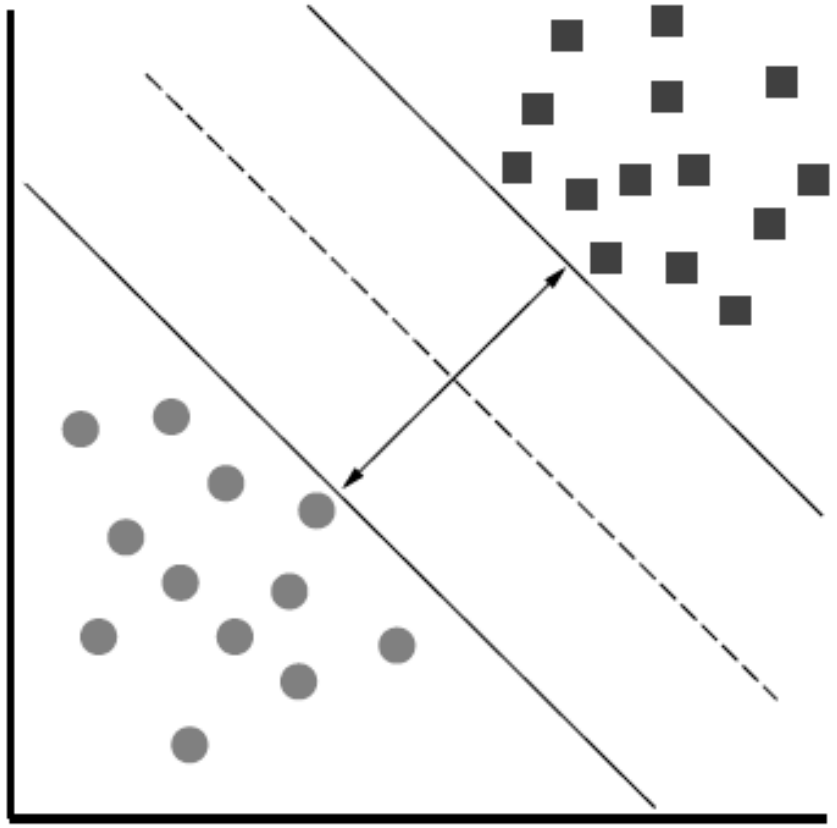


$$g(F(x)) = \ln\left(\frac{F(x)}{1-F(x)}\right) = \beta_0 + \beta_1 x, \qquad F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Support Vector Machine

- **To maximize the Margin γ: Distance of closest example from the decision line/hyperplane**

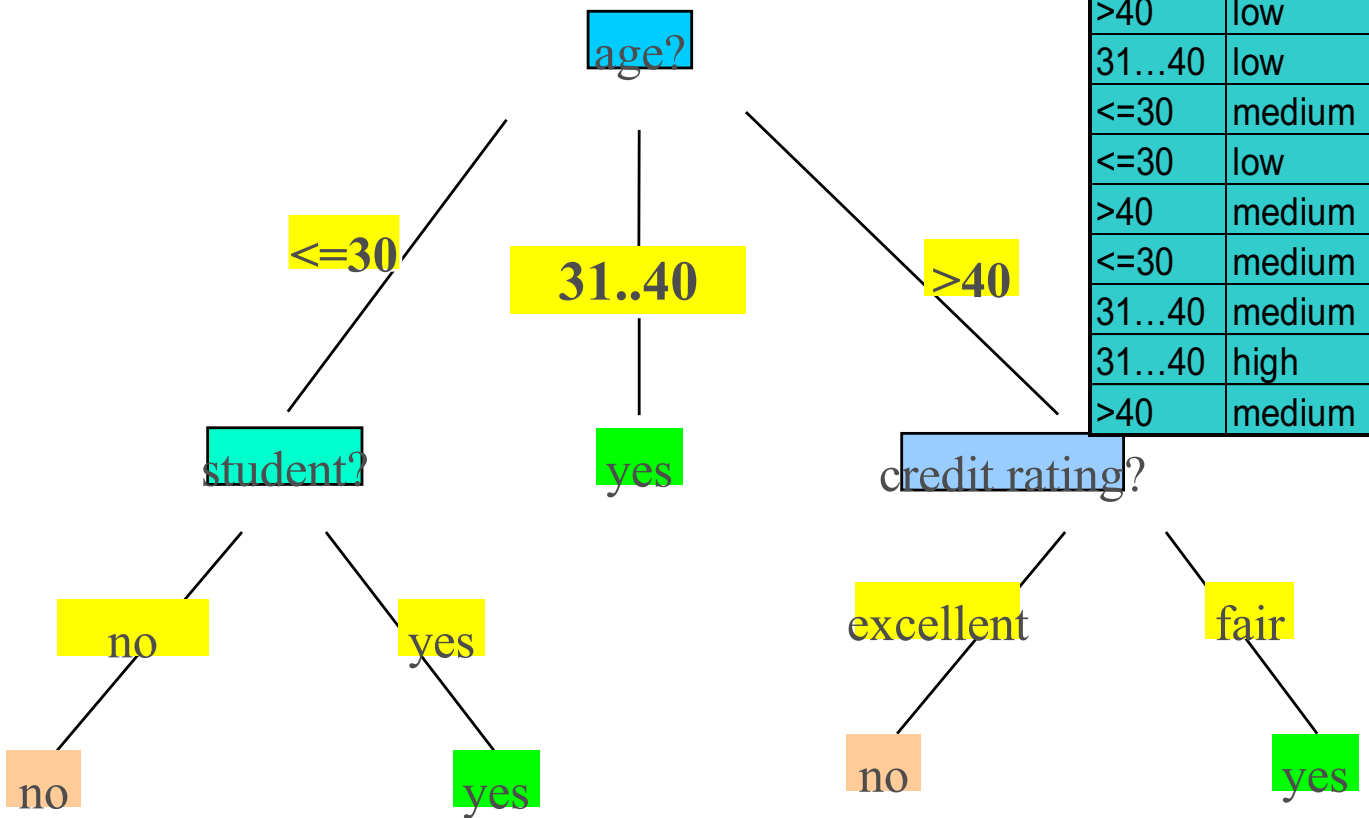




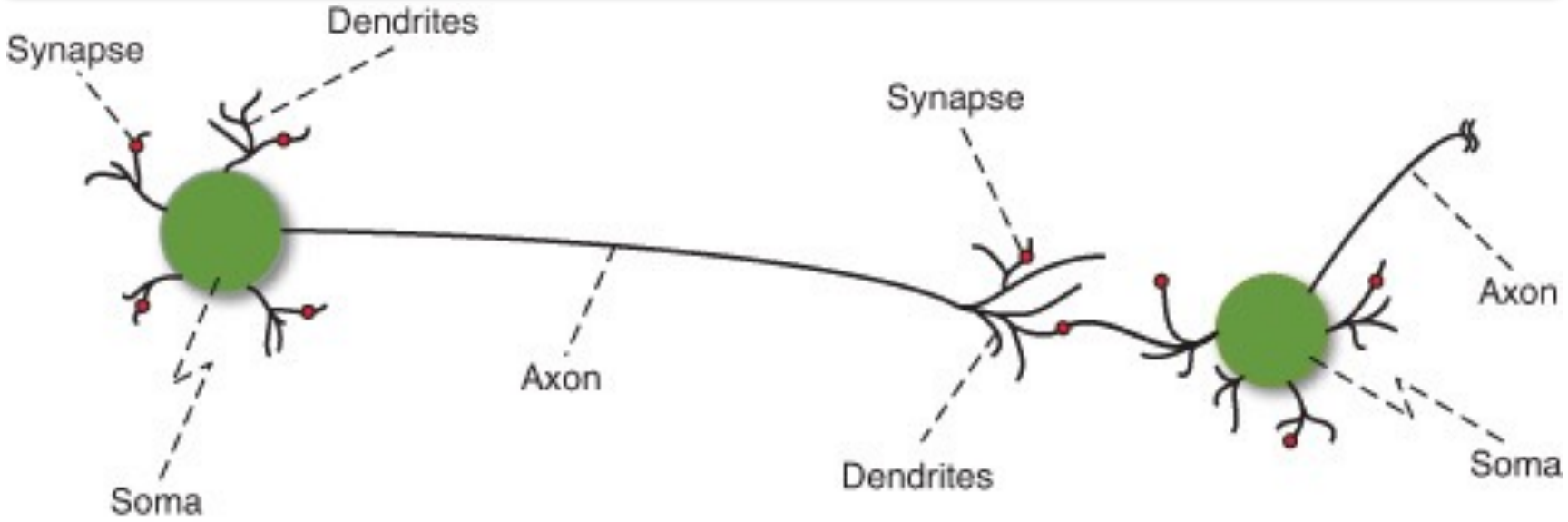
Decision Tree Induction: An Example

- Training data set: Buys_computer
- The data set follows an example ID3
- Resulting tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



Neurons



Artificial Neuron

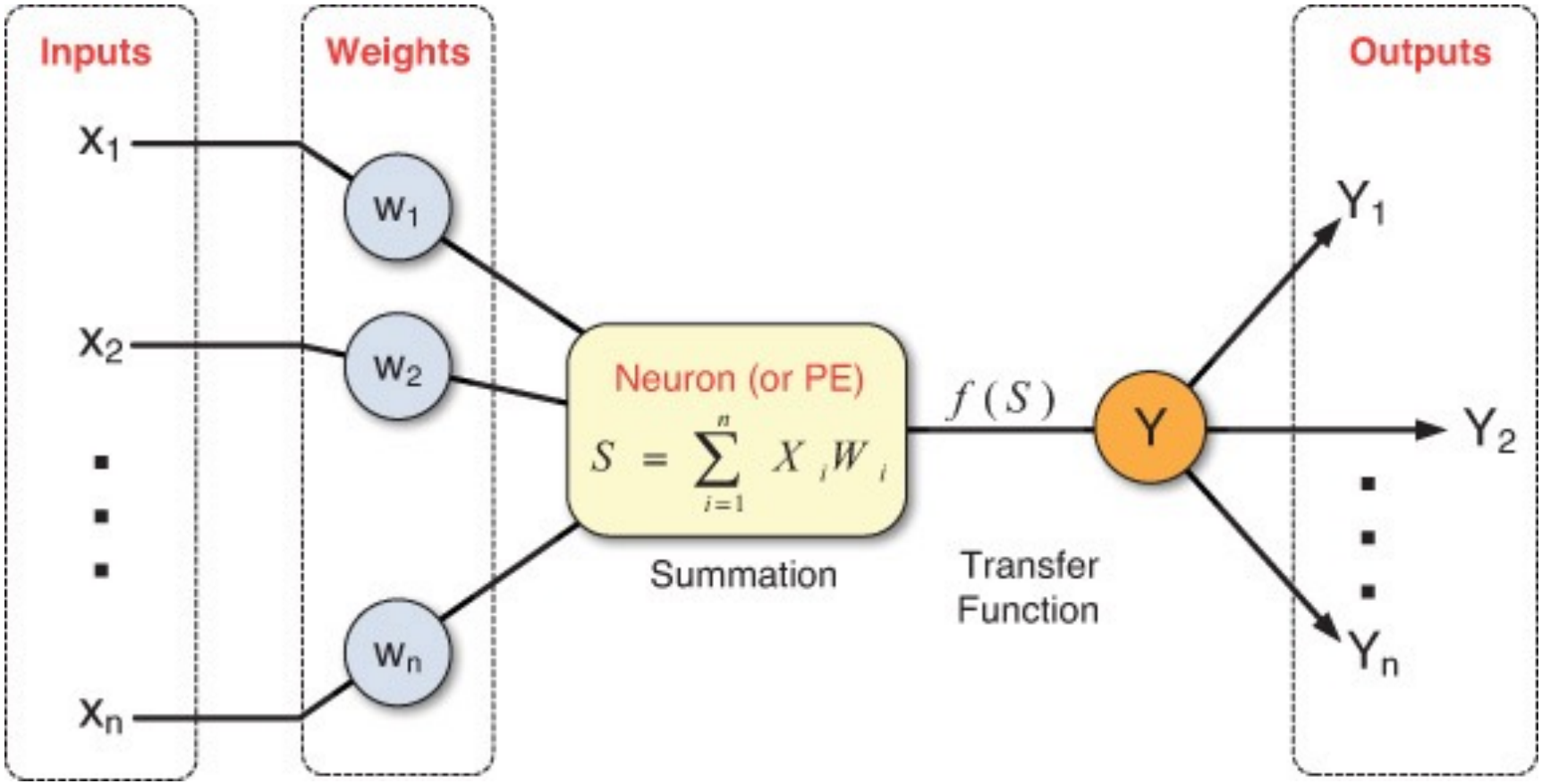


Figure 5.4 Processing Information in an Artificial Neuron

Artificial Neural Networks

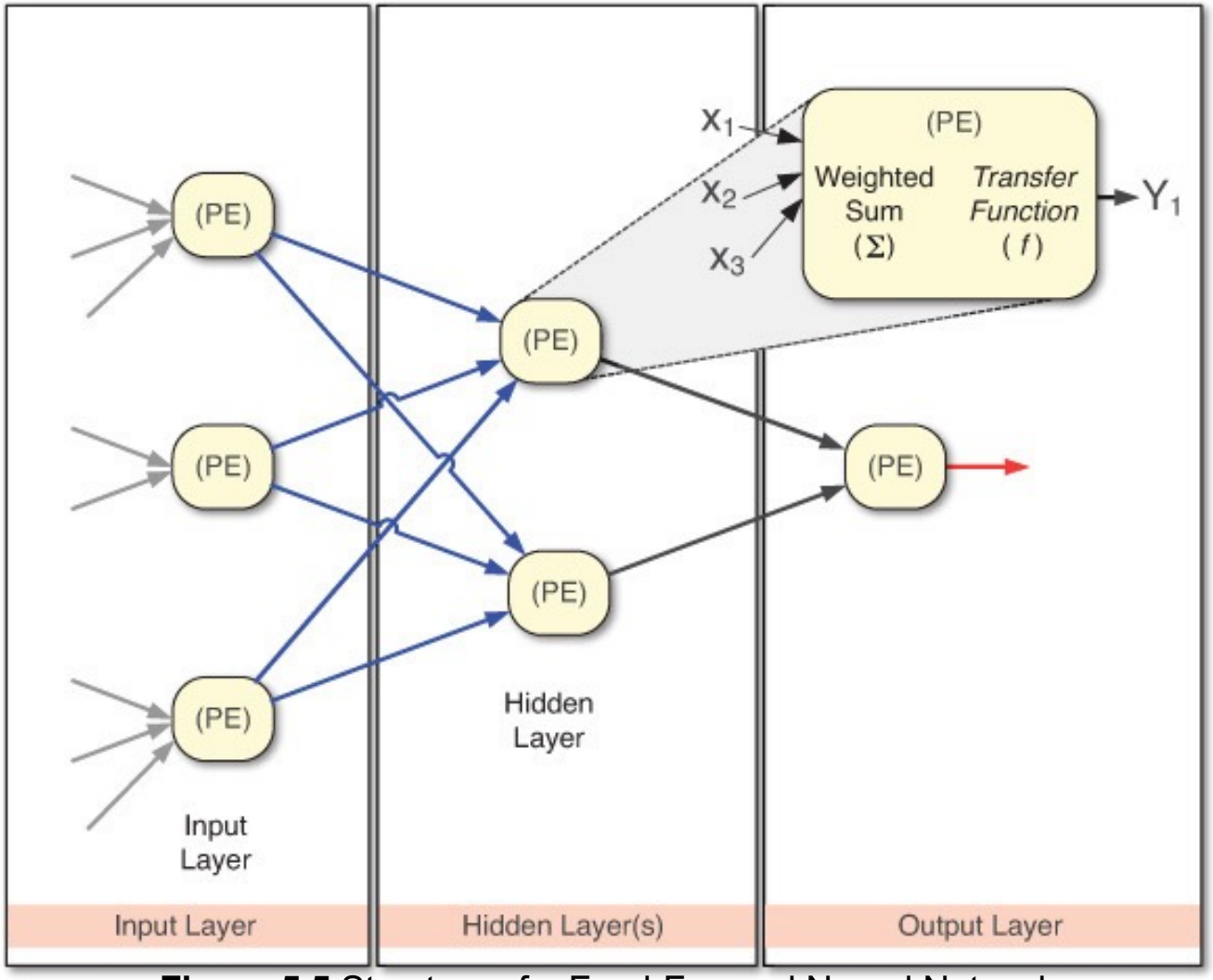
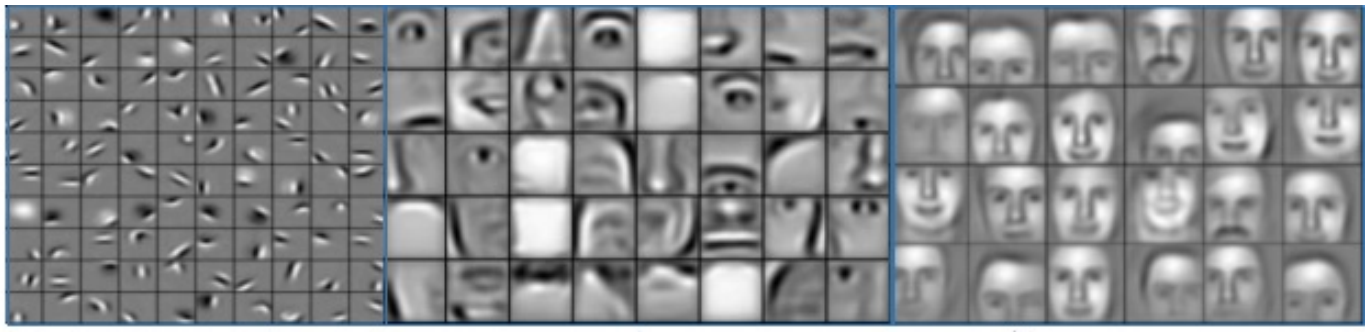
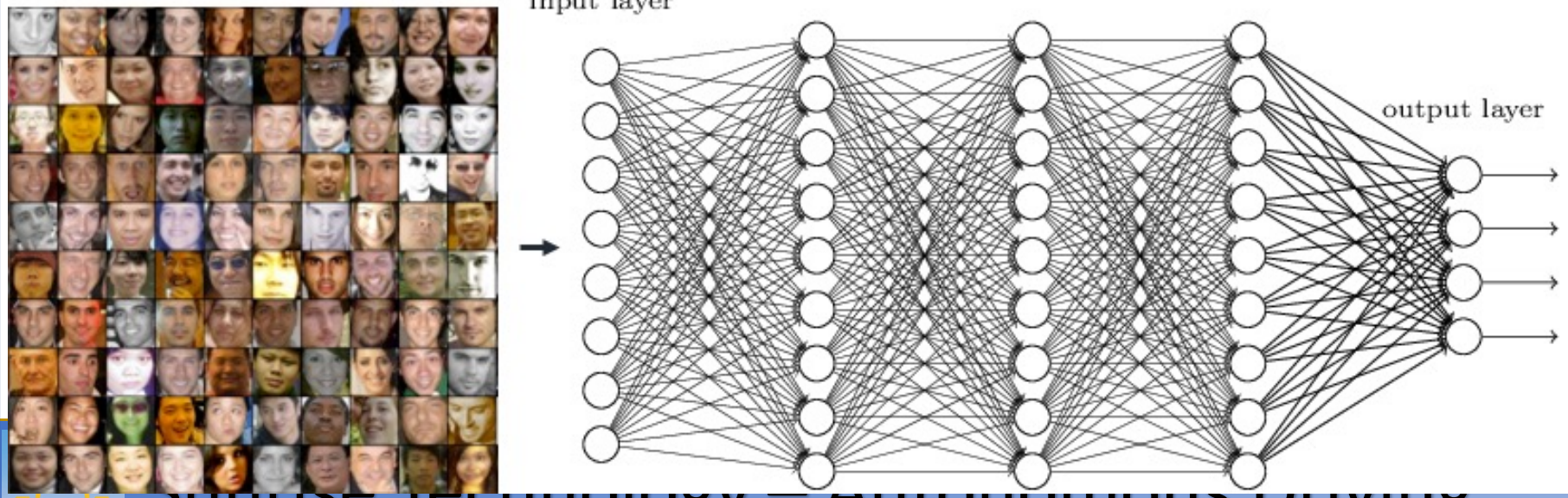


Figure 5.5 Structure of a Feed-Forward Neural Network

Deep Neural Networks

Deep neural networks learn hierarchical feature representations





Thank you