INTRODUCTION

1

AGENDA

- Overview
- Historical Notes
- No-Free-Lunch Theorems

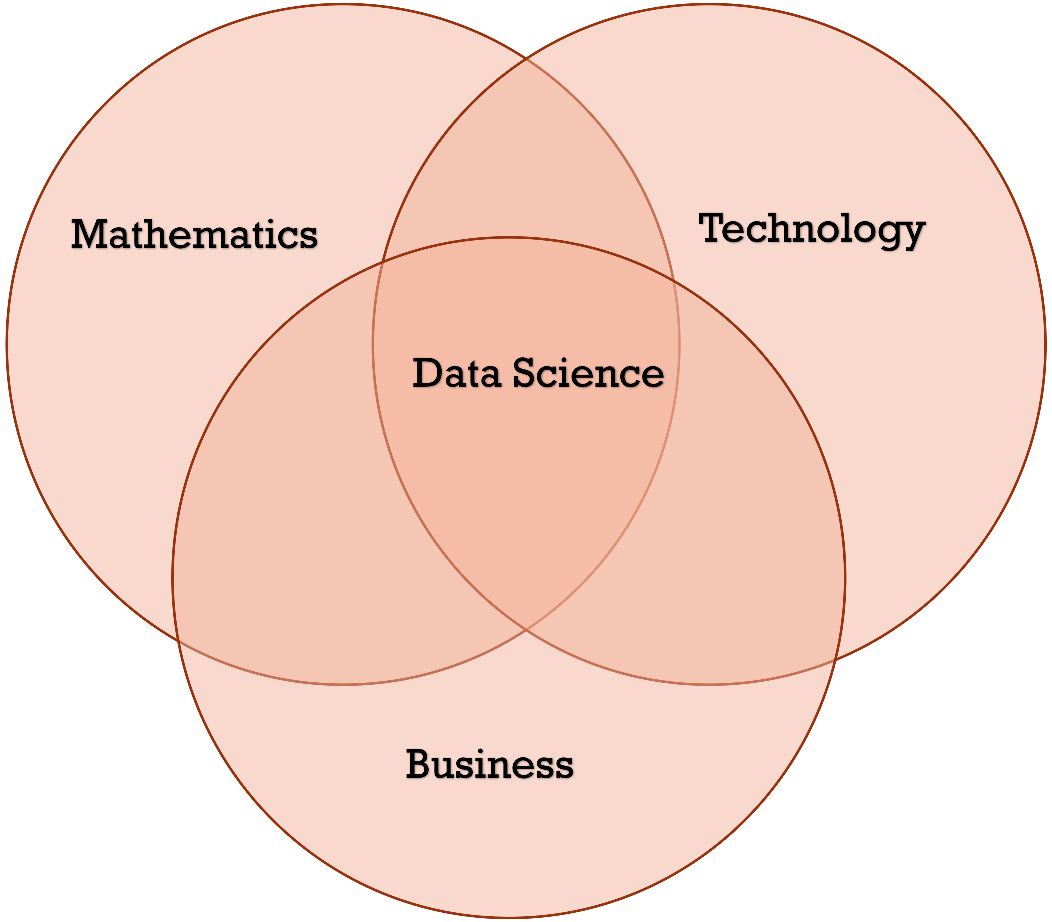
OVERVIEW

What is Data Science, AI, ML and DL?

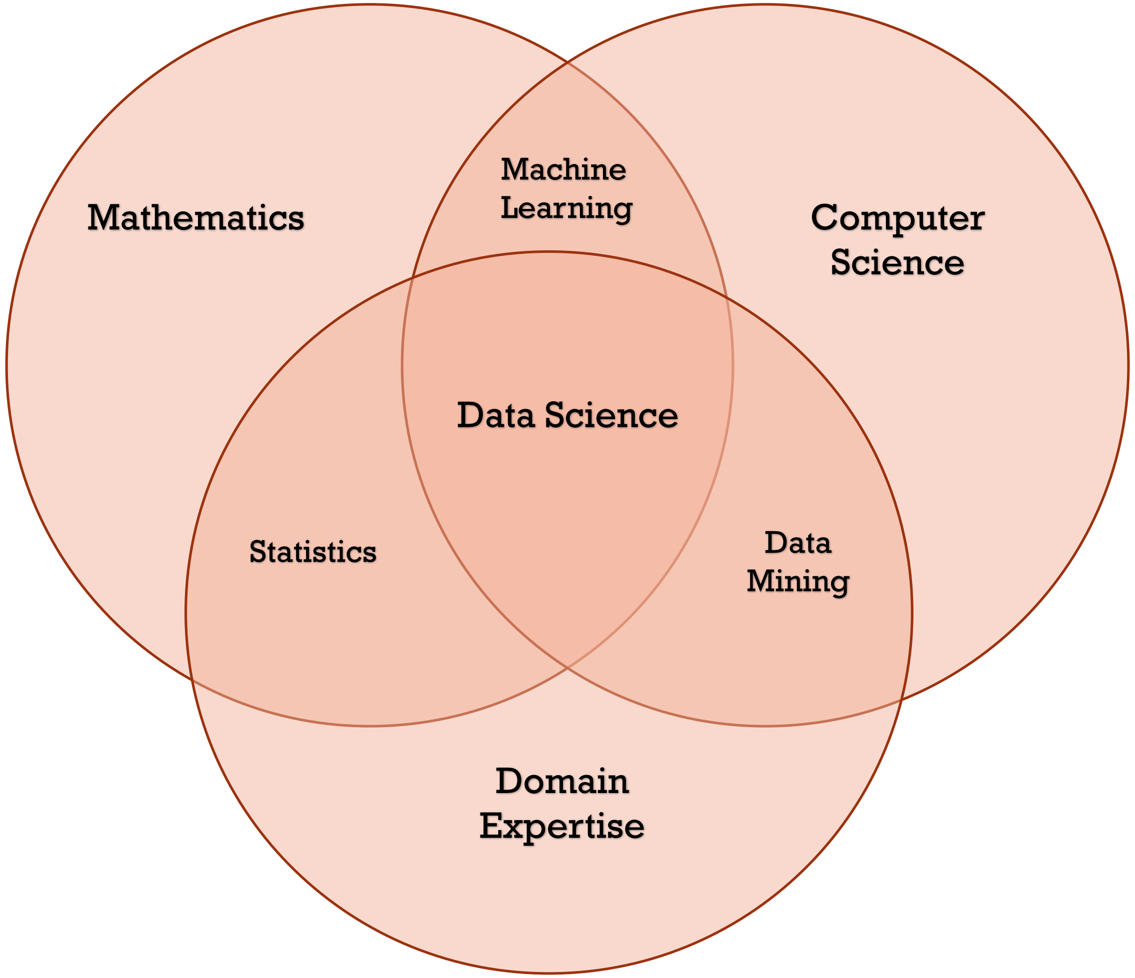
DATA SCIENCE – THE "FOURTH PARADIGM" OF SCIENCE

- Theoretical
- Empirical
- Computational
- **Data Driven**

DATA SCIENCE



SKILLS OF A DATA SCIENTIST

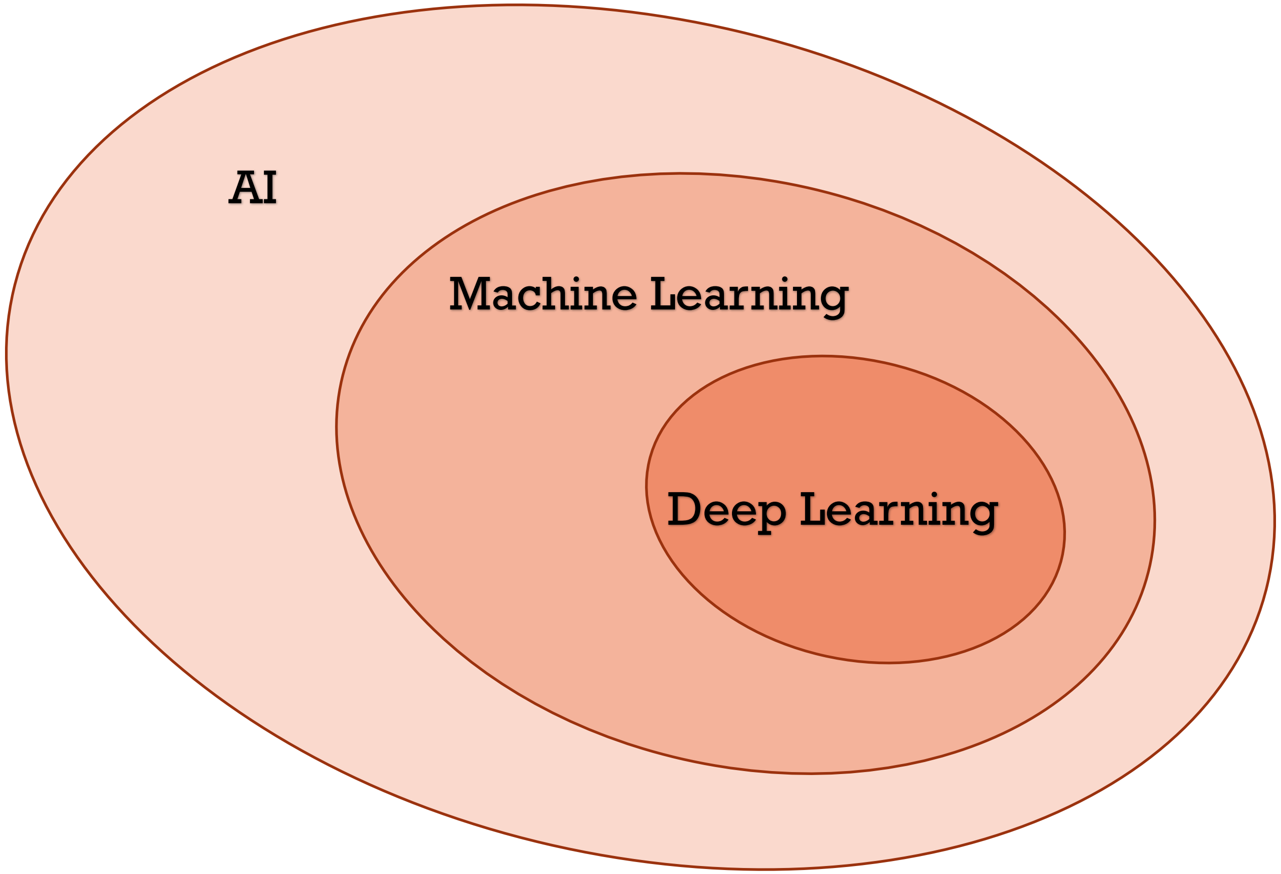


- **Coding -** R and/or Python, etc.
- **Math –** Linear algebra, Matrix analysis, Numerical analysis, Optimization theory, …
- **AI**
- **Machine Learning**
- **Deep Learning**
- **Communication**
- **Data Architecture**
- **Big Data**
- **…**

BIG DATA

- **Big data** means massive volume of structured and unstructured data sets that are so large and complex that classical data analysis tools are useless.
- “Big Data” is the emerging discipline of capturing, storing, processing, analyzing and visualizing these huge quantities of information.
- Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabytes of data.
- Big data requires a set of techniques and technologies (volume, variety, velocity (3Vs), veracity, value (5Vs)) with new forms of integration to reveal insights from datasets that are diverse, complex, and of a massive scale.
- https://www.linkedin.com/pulse/20140306073407-64875646-big-data-the-5-vs-everyone-must-know/

AI, ML AND DL



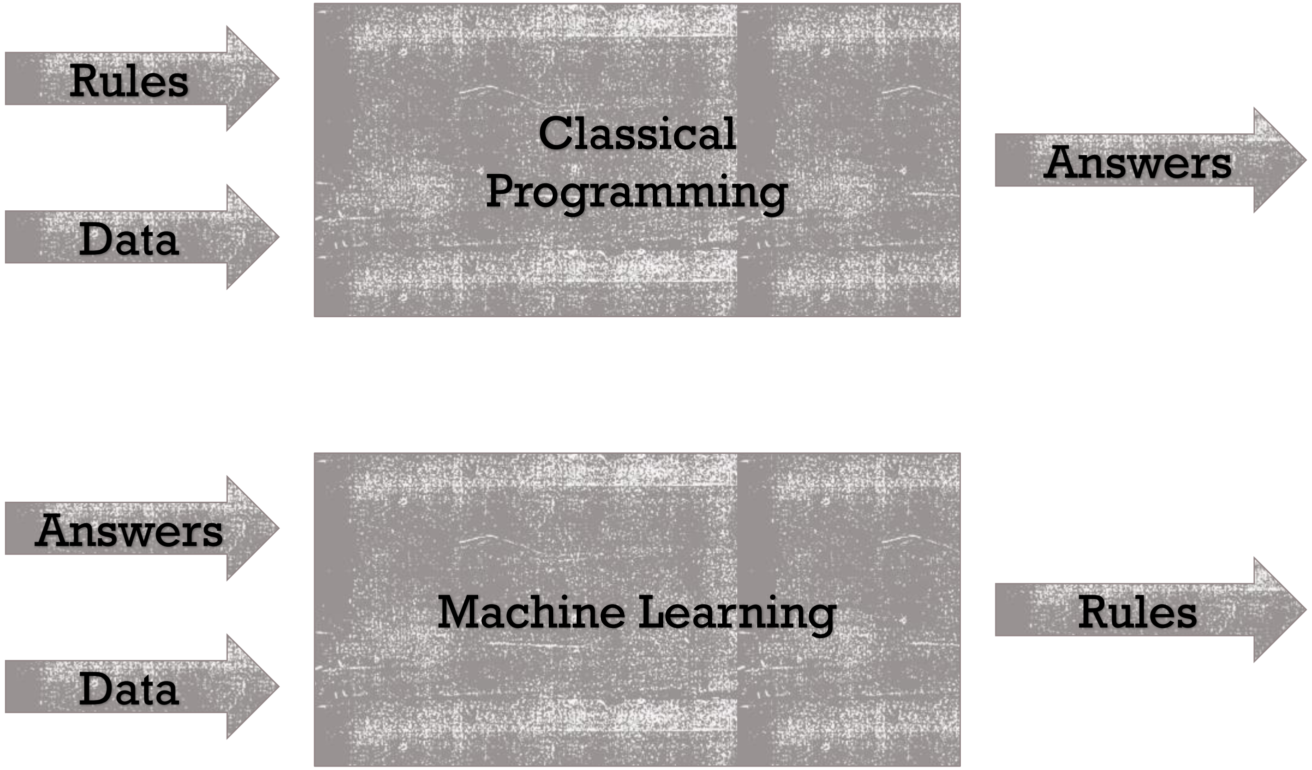
AI

- AI: Automate intellectual tasks normally performed by humans.
- **Symbolic AI**: 1950-1980
  - Human-level artificial intelligence could be achieved by a sufficiently large set of explicit rules: playing chess,…
  - It is intractable for more complex problems: image classification, speech recognition,…
- **Machine Learning**: The name **ML** was coined in 1959 by Arthur Samuel. He defined ML machine learning as a "Field of study that gives computers the ability to learn without being explicitly programmed".
  - Originating from the study of pattern recognition and computational learning theory.
  - ML explores the study and construction of algorithms that can learn from data and make predictions.
  - **Statistical Learning** is a framework for machine learning.

TURING TEST

- **1950** — Alan Turing developed a test of machine's ability to exhibit intelligent behavior equivalent or indistinguishable from that of a human.
- Human evaluator would judge between a human and a machine that is designed to generate human-like responses. The evaluator would be aware that one of the two partners in conversation is a machine. If the evaluator cannot reliably separate the machine from the human, the machine is said to have passed the test.
- Turing predicted that machines would eventually be able to pass the test and that Machine Learning would be an important part of building powerful machines.

CLASSICAL PROGRAMMING VS ML



ML ALGORITHMS

- The main goals
  - Data understanding – statistical inference
  - Data prediction
- The main types
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning
  - Self-supervised learning
  - Reinforcement learning

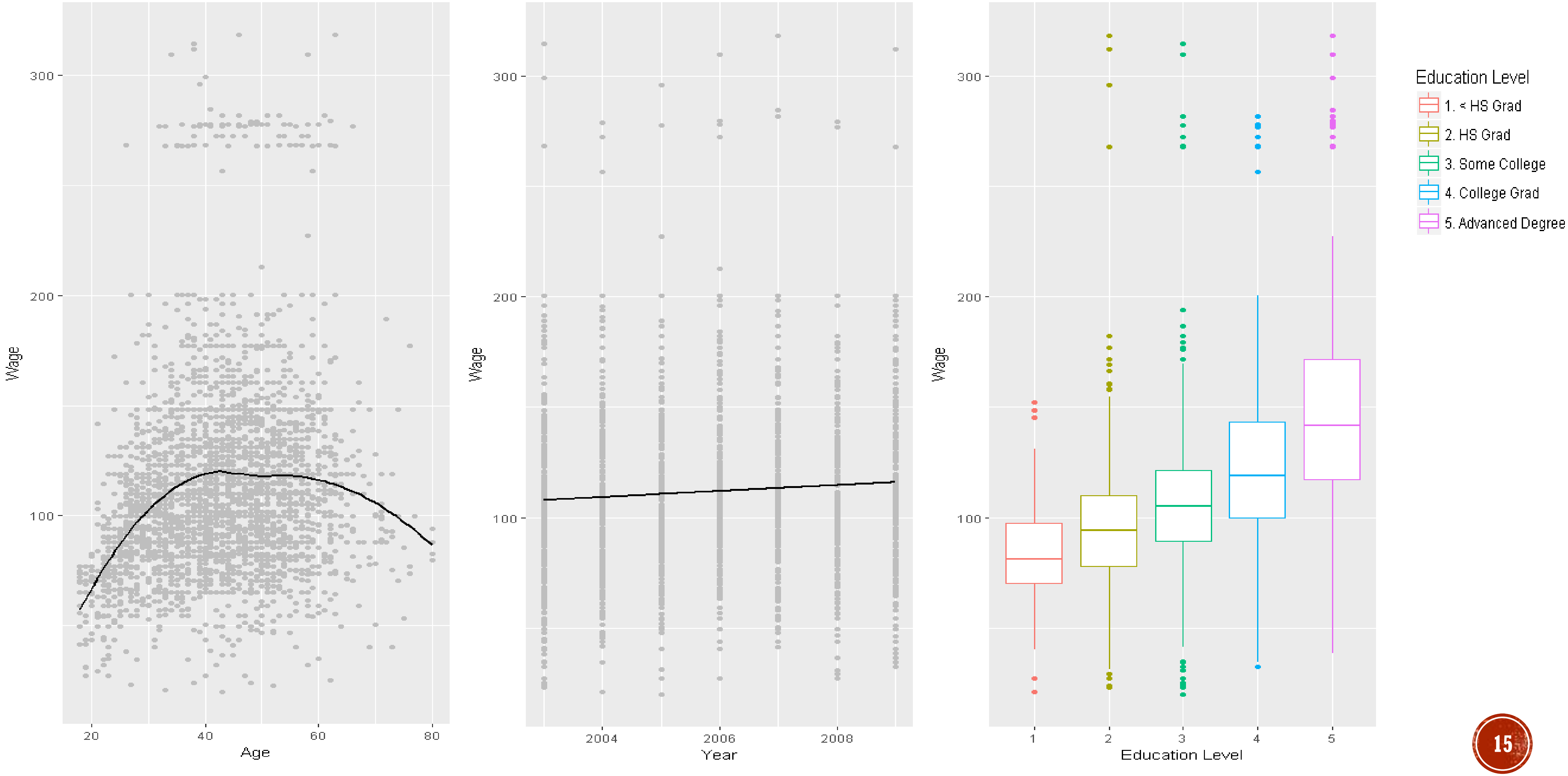
SUPERVISED LEARNING

- Data is labeled by a "teacher", "supervisor".
- The goal is to learn a general rule that maps data to labels
- Supervised learning splits into two broad categories:
  - Regression
  - Classification

REGRESSION

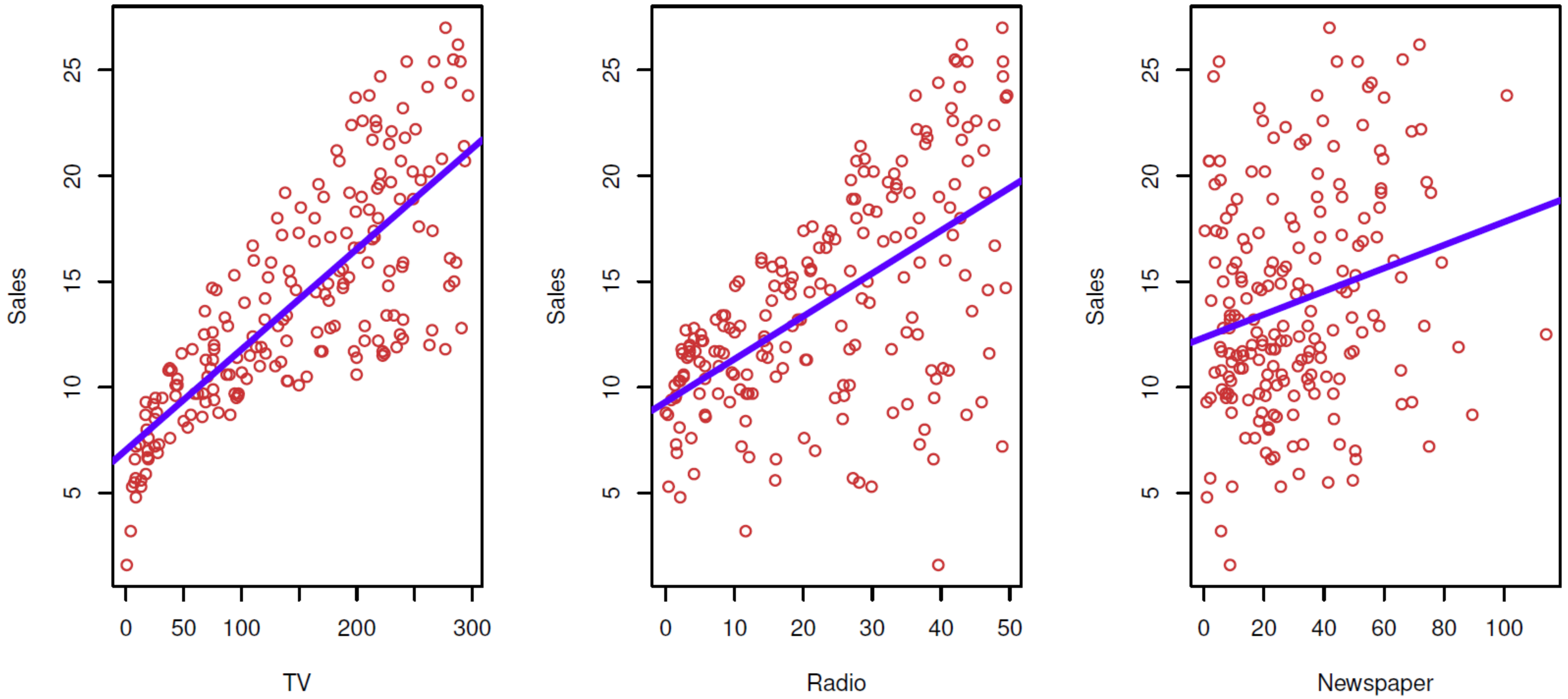
- The output (dependent variables) is numeric – the response is quantitative.
- Applications include function approximation, forecasting - stock prices, energy consumption, and population.
- Common algorithms — regression, lasso and ridge regressions, k-nearest neighbor, neural networks, etc.

WAGE DATA



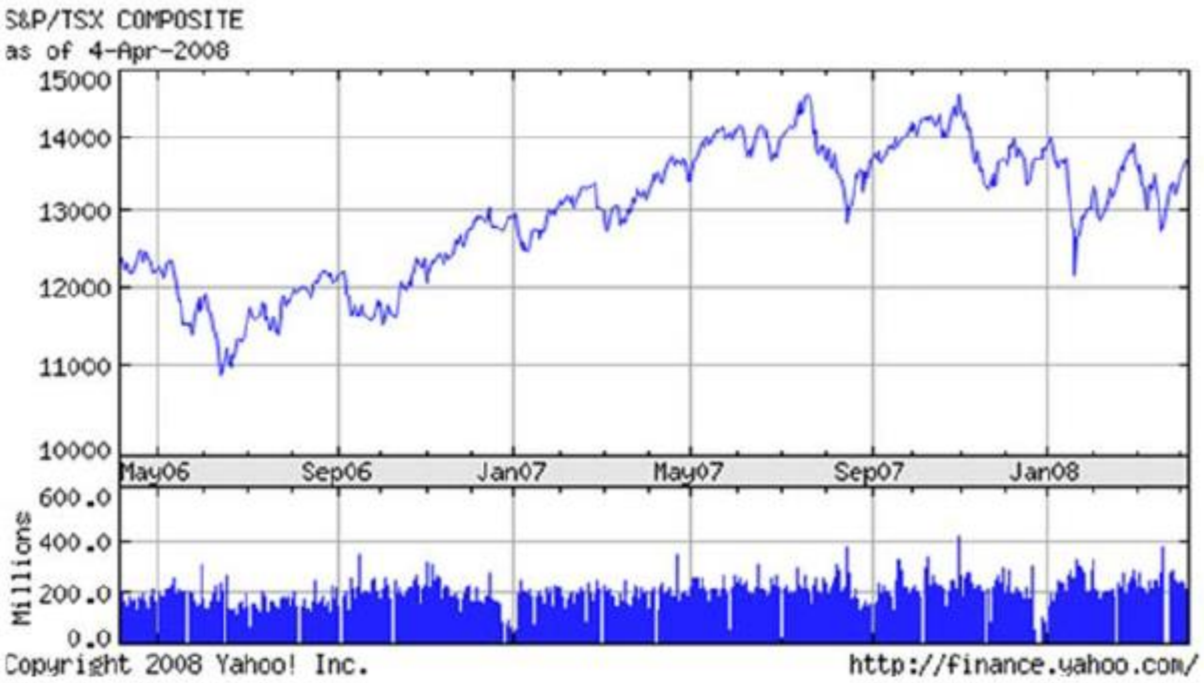
Wage
300
200
100
20
40
60
80
Age
Wage
300
200
100
2004
2006
2008
Year
Wage
300
200
100
1
2
3
4
5
Education Level
Education Level
1. < HS Grad
2. HS Grad
3. Some College
4. College Grad
5. Advanced Degree

ADVERTISING DATA



$$Sales \approx f(TV, Radio, Newspaper)$$

STOCK PRICE PREDICTION



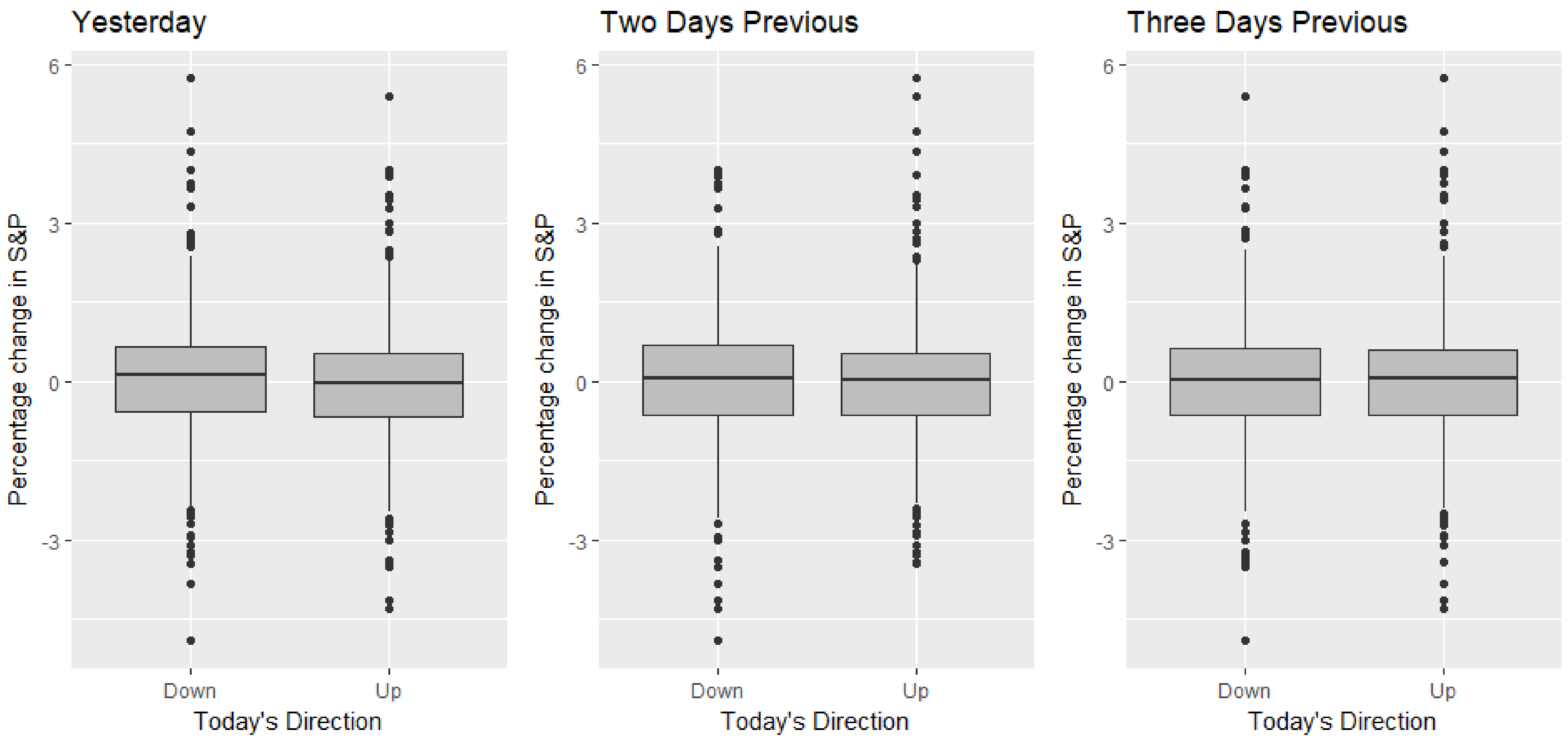
Task is to predict stock price at future date

• This is a regression task, as the output is continuous

CLASSIFICATION

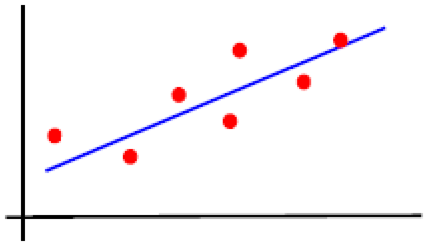
- The output (dependent variables) is categorical – the response is qualitative.
- The goal is to assign a class from a finite set of classes to an observation.
- Applications include spam filters, advertisement recommendation systems, image and speech recognition.
- Predicting whether a patient will have a heart attack within a year is a classification problem.
- Common algorithms — discriminant analysis, logistic regression, Naïve Bayes classifier, support vector machines, k-nearest neighbor, decision trees, bagging, boosting, neural networks.

SMARKET DATA (DAILY MOVEMENT IN S&P 500)

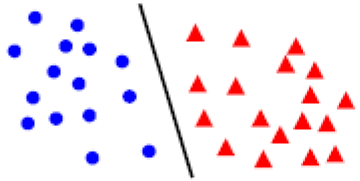


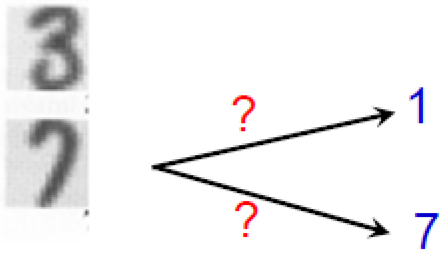
SUPERVISED LEARNING

1. Regression - supervised
   - estimate parameters, e.g. of weight vs height



2. Classification - supervised
   - estimate class, e.g. handwritten digit classification





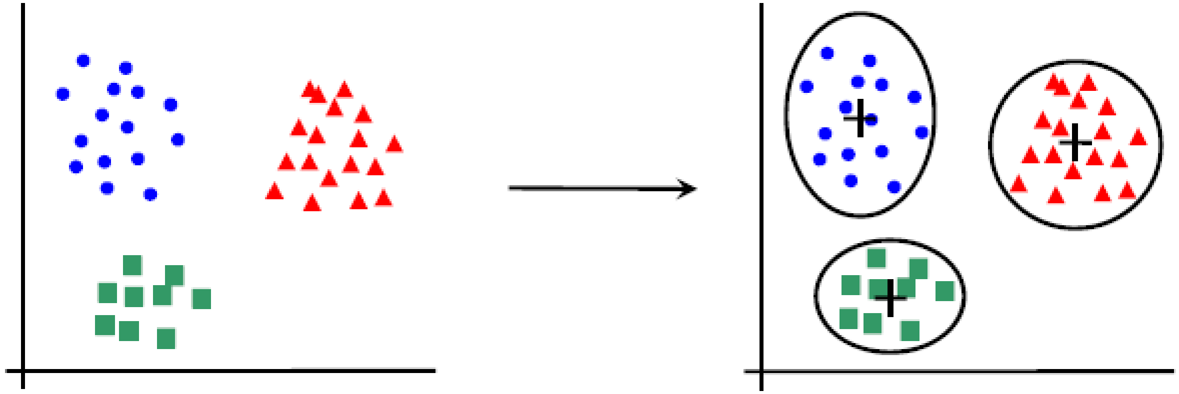
UNSUPERVISED LEARNING

- No labels are given to the learning algorithm, leaving it on its own to find structure in its. The goal is discovering hidden patterns and key features in data.
- Common algorithms — clustering (k-means, hierarchical, etc.), anomaly detection, dimensionality reduction (principal components analysis, etc.), neural networks.

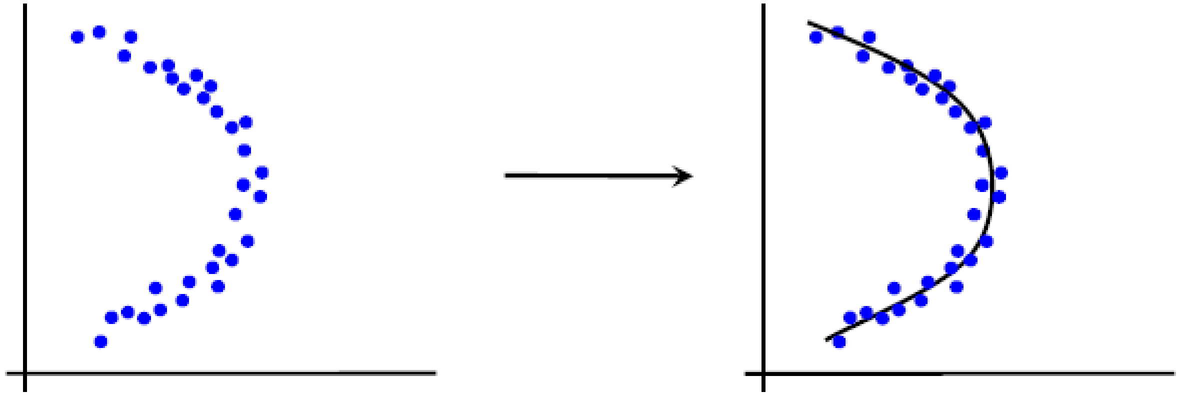
UNSUPERVISED LEARNING

3. Unsupervised learning – model the data

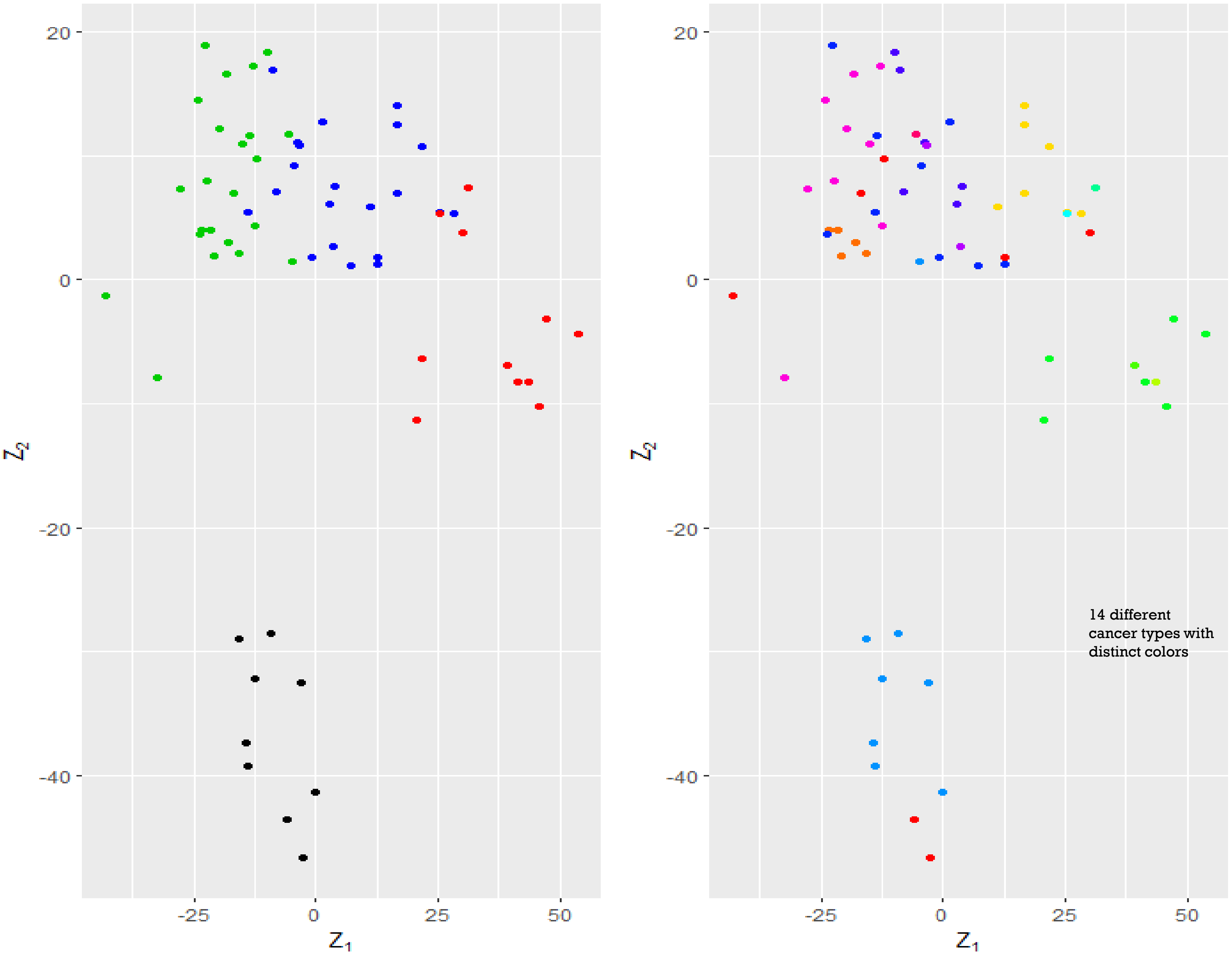
- clustering



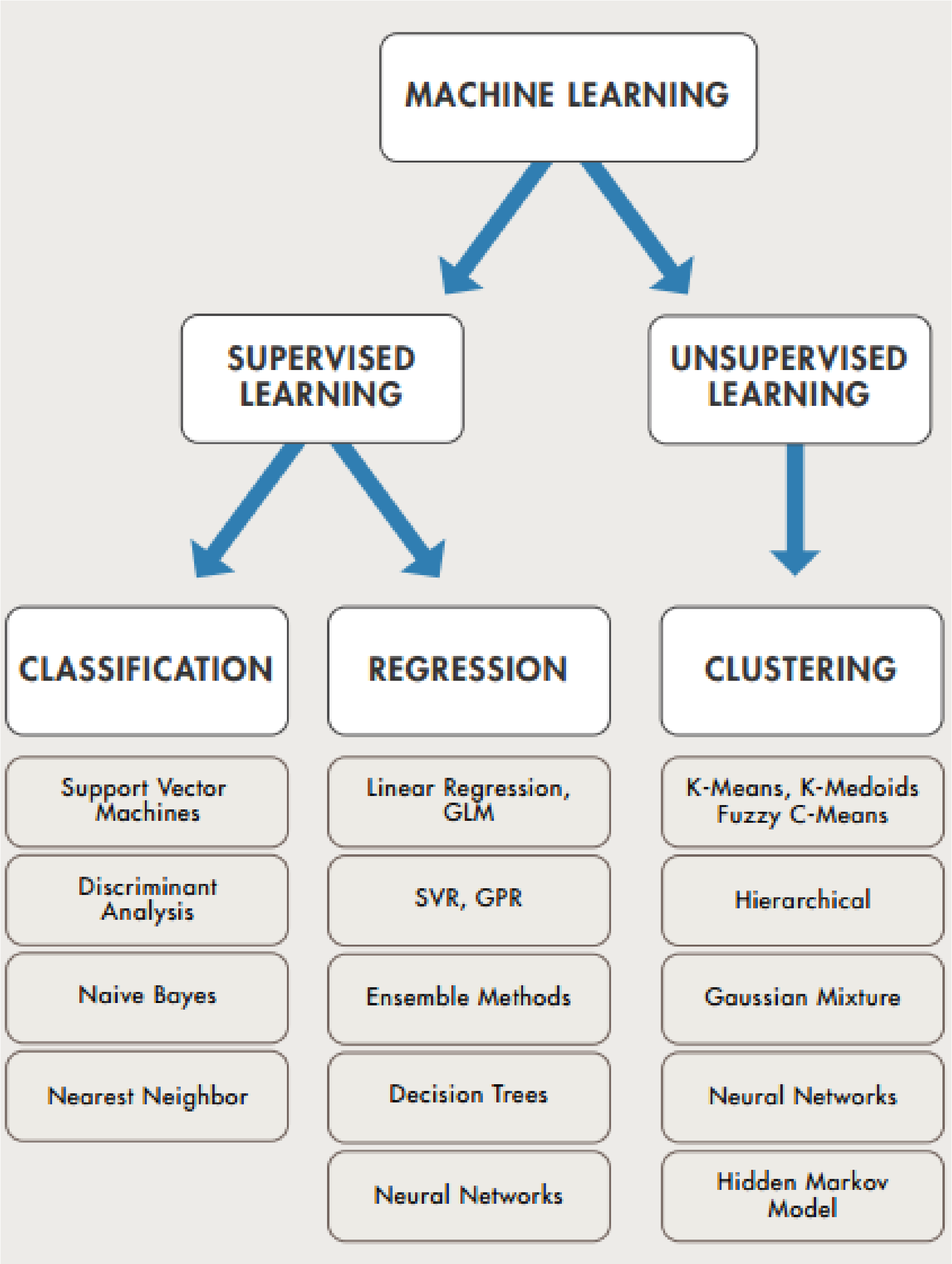
- dimensionality reduction



NCI60



14 different cancer types with distinct colors



SUPERVISED VS UNSUPERVISED

- SVR – support vector regression
- GPR – Gaussian process regression
- Ensemble methods – bagging, random forest

https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

SEMI-SUPERVISED LEARNING

- A “teacher” gives an incomplete training signal: a training set with some (often many) of the target outputs missing.
- Many researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy.
- Common algorithms – self-training, co-training, graph-based models, etc.

SELF-SUPERVISED LEARNING

- Supervised learning without human-annotated labels.
- Labels are generated from the input data, typically using heuristic algorithm.
- Common algorithms – autoencoders.

REINFORCEMENT LEARNING

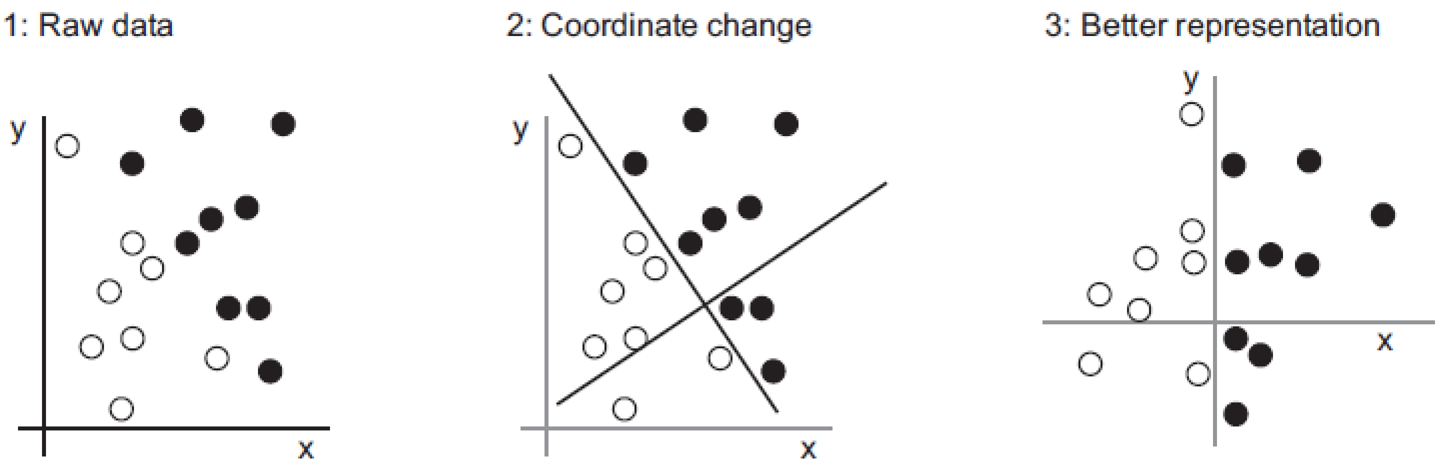
- A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal.
- Another example is learning to play a game by playing against an opponent.

DATA REPRESENTATION

- How ML reveals the rules
- How ML transforms the input data into meaningful outputs?
- Data **Representations**!
  - It is a different way to look at data;
  - Some tasks that may be difficult with one representation can become easy with another.
- The central problem of ML and DL is to meaningfully transform data: learn representations of the input data that get us closer to the expected output.
- For example, a color image can be encoded in the RGB format (red-green-blue) or in the HSV format (hue-saturation-value):
  - The task "select all red pixels in the image" is simpler in the RGB format;
  - The task "make the image less saturated" is simpler in the HSV format.

DATA REPRESENTATION

- Machine-learning models are all about finding appropriate representations for their input data—transformations of the data that make it more amenable to the task at hand.



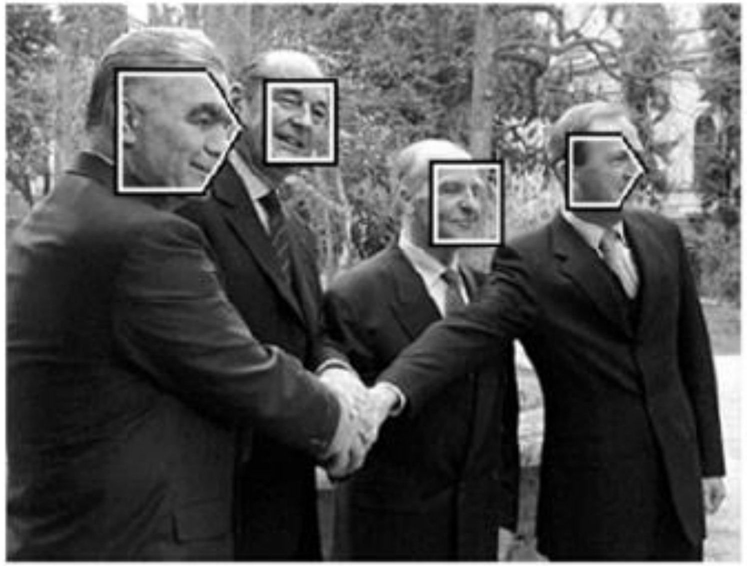
DATA REPRESENTATION

- *Learning*, in the context of machine learning, describes an automatic search process for better representations.
- All machine-learning algorithms consist of automatically finding such transformations that turn data into more-useful representations for a given task.
- Machine-learning algorithms aren't usually creative in finding these transformations; they're merely searching through a predefined set of operations, called a *hypothesis space*.

SHALLOW VS DEEP LEARNING

- The number of layers of representations is the ***depth*** of the model.
- Some ML approaches tend to focus on learning only one or two layers of representations of the data - ***shallow learning***.
- **Deep learning** puts an emphasis on learning successive *layers* of increasingly meaningful representations.
- Modern deep learning often involves tens or even hundreds of successive layers of representations.

FACE DETECTION



Supervised classification problem

- Need to classify an image window into three classes:
    - non-face
    - frontal-face
    - profile-face

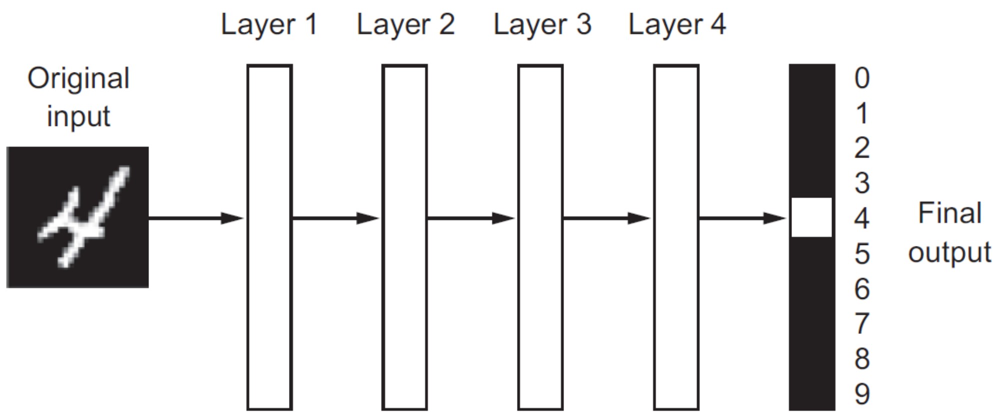
FACE DETECTION

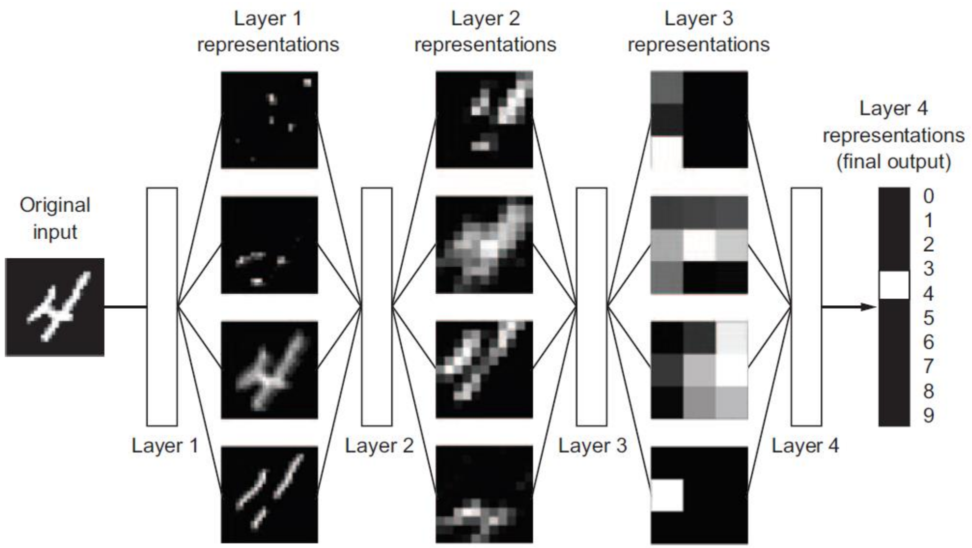
Classifier learns from labelled data
Training data for frontal faces

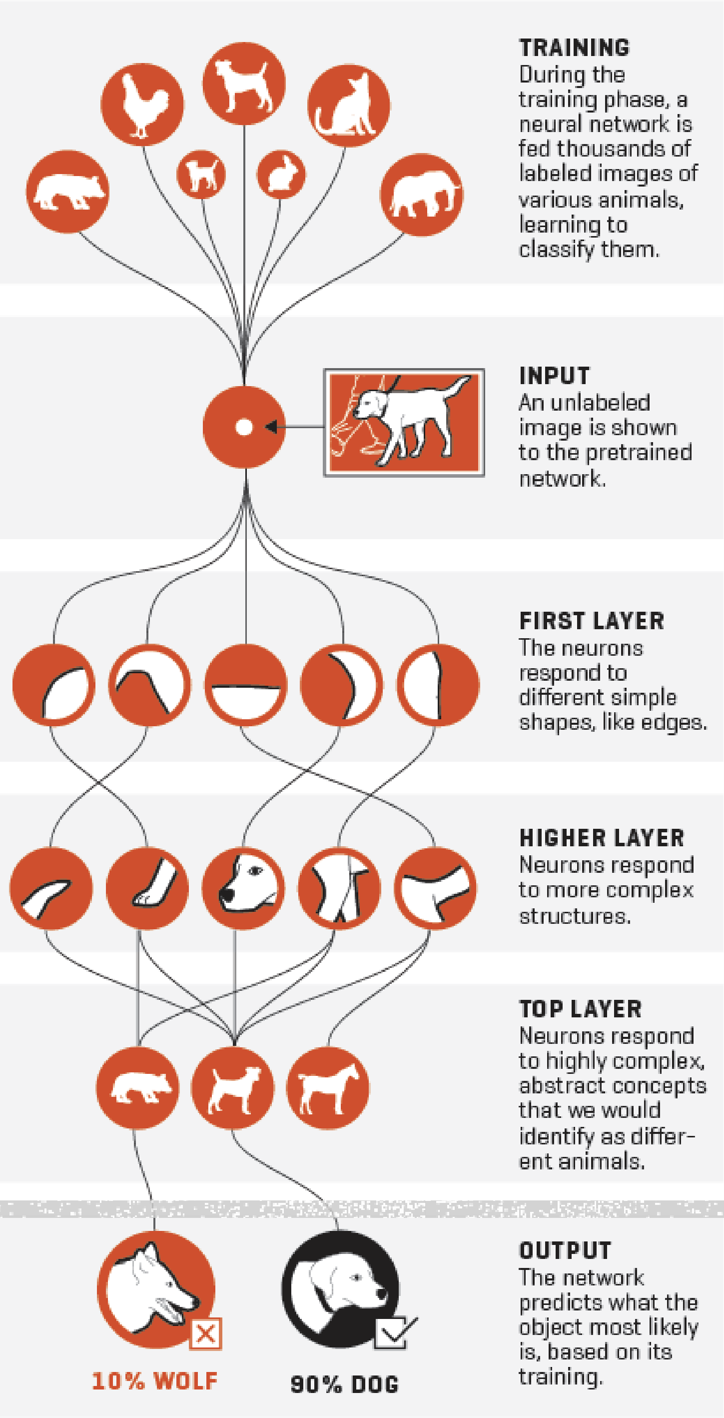
- 5000 faces
  - All near frontal
  - Age, race, gender, lighting
- 10^8 non faces
- faces are normalized
  - scale, translation

SHALLOW VS DEEP LEARNING

- In deep learning, layered representations are learned via models called *neural networks*, structured in layers stacked on top of each other.



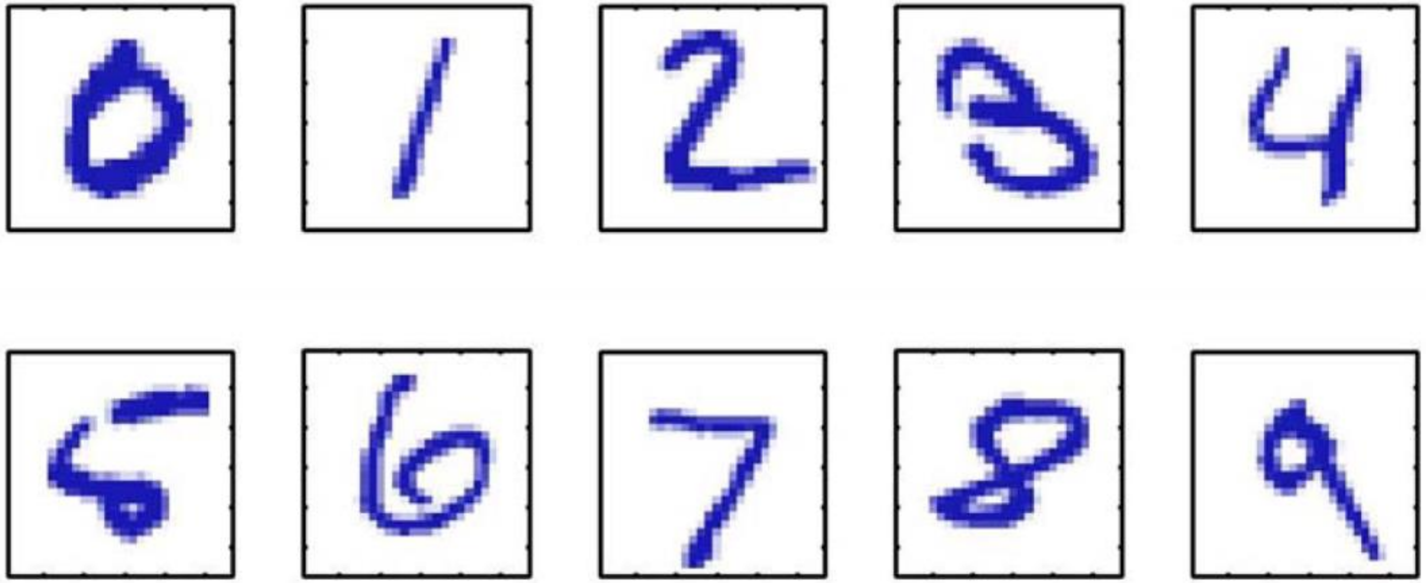




HOW NEURAL NETWORKS RECOGNIZE DOG

HTTP://FORTUNE.COM/AI-ARTIFICIAL-INTELLIGENCE-DEEP-MACHINE-LEARNING/

DIGIT RECOGNITION



Images are 28 x 28 pixels

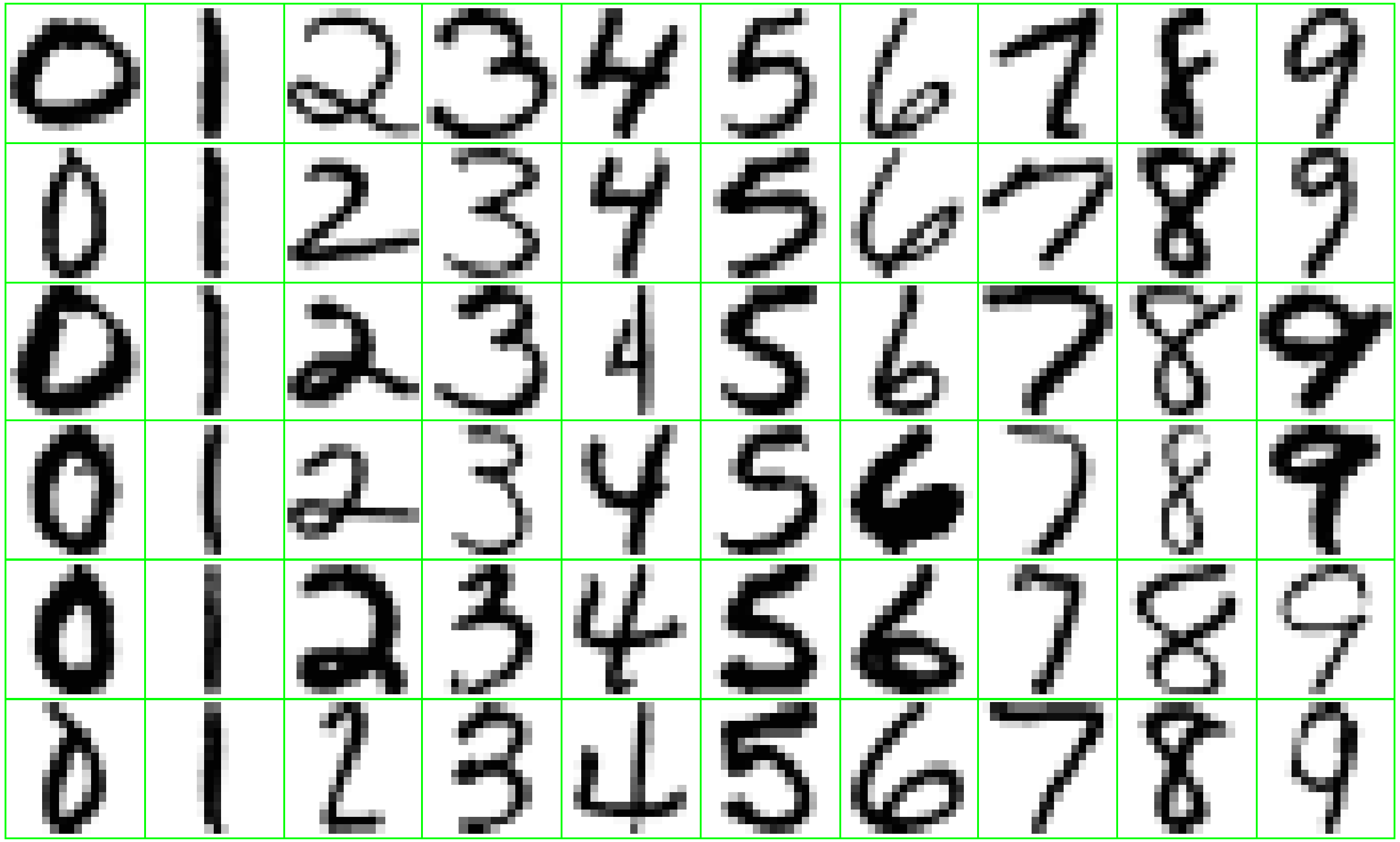
Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

$$f : \mathbf{x} \to \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

DIGIT RECOGNITION

Examples of handwritten digits from U.S. postal envelopes



MNIST DATABASE

- Linear classifier – error rate 7.6%
- Non-linear classifier (40 PCA + Quadratic classifier) – error rate 3.3%
- SVM – error rate 0.56%
- Neural network (784-800-10) – error rate 1.6%
- Deep neural network (784-2500-2000-1500-1000-500-10) – error rate 0.35%
- Convolutional neural network – error rate 0.21%
- https://en.wikipedia.org/wiki/MNIST_database

EVALUATION OF ML MODELS

- Training data – model construction.
- Validation data - tuning the configuration, tuning model parameters. Learning a good configuration. Overfitting, Information leaks.
- Test data.

OPTIMIZATION VS GENERALIZATION

- Optimization – adjusting a model to get the best performance possible on the training data (the process of learning in machine learning).
- Generalization – performance of the trained model on data it has never seen before.
- At the beginning of training – optimization and generalization are correlated. The lower is loss on the training data, the lower the loss on the test data. The model is underfit.
- At the end of training – optimization and generalization are anticorrelated. Generalization stops improving, the model is starting to overfit.

41

HISTORICAL NOTES

BASIC ML ALGORITHMS

- **1936** — Fisher’s linear discriminant analysis
- **1940** — Logistic regression
- **1967** — The nearest neighbor algorithm was written, allowing computers to begin using very basic pattern recognition.
- **1970** — Nelder and Wedderburn coined the term generalized linear models
- **1980** — Breiman, Friedman, Olshen and Stone introduced classification and regression trees and cross-validation
- **1986** — Hastie and Tibshirani coined the term generalized additive models

NEURAL NETWORK

1940
- Warren McCulloch and Walter Pitts
- Computational model for neural networks based on threshold logic

1949
- Donald Hebb
- Method to update weights between neurons that came to be known as Hebbian learning

1957
- Frank Rosenblatt
- The first neural network for computers - the perceptron

1960
- Widrow and Hoff
- The first adaptive linear neuron model (ADALINE) with gradient descent

1974
- Paul Werbos
- Backpropagation

1984
- Kohonen
- Self organizing maps for unsupervised learning in neural networks

2006
- Geoffrey Hinton
- Coined the term "deep learning" .

APPLICATION OF ML ALGORITHMS

1959
- Arthur Samuel wrote the first computer self-learning program
- The program was the game of checkers, and the IBM computer improved at the game the more it played

1981
- Gerald Dejong
- **Explanation Based Learning** (EBL). In EBL a computer analyses training data and creates a general rule it can follow by discarding unimportant data. An example of EBL using a perfect domain theory is a program that learns to play chess by being shown examples

1985
- Terry Sejnowski
- **NetTalk**, which learnt to pronounce words the same way a baby does

1997
- IBM's **Deep Blue** beats the world champion at chess

2010
- The Microsoft **Kinect** could track 20 human features at a rate of 30 times per second, allowing people to interact with the computer via movements and gestures

APPLICATION OF ML ALGORITHMS

2011

- IBM developed a question answering machine **Watson** to answer questions on the quiz show Jeopardy! In 2011, Watson competed on Jeopardy! against former winners and received the first place prize of $1 million.
- **Google Brain** is developed, and its deep neural network can learn to discover and categorize objects much the way a cat does.

2012

- **Google's X** Lab developed a machine learning algorithm that is able to autonomously browse YouTube videos to identify the videos that contain cats. A cluster of 16,000 computers dedicated to mimicking some aspects of human brain activity had successfully trained itself to recognize a cat based on 10 million digital images taken from internet pictures and videos.

2014

- Facebook develops **DeepFace**, a software algorithm that is able to recognize or verify individuals on photos to the same level as humans can. DeepFace is a deep learning facial recognition system. It identifies human faces in digital images. It employs a nine-layer neural net with over 120 million connection weights, and was trained on four million images uploaded by Facebook users.

2016

- Google's artificial intelligence algorithm beats a professional player at the Chinese board game Go, which is considered the world's most complex board game and is many times harder than chess. The **AlphaGo** algorithm developed by Google **DeepMind** managed to win five games out of five in the Go competition.

PRACTICAL ML PROBLEMS

- Spam detection
- Credit card fraud detection
- Digit recognition
- Handwriting recognition
- Speech understanding
- Face detection
- Product recommendation
- Medical diagnosis
- Stock trading
- Customer segmentation

NO-FREE-LUNCH THEOREMS

NFL THEOREMS

- David Wolpert derived no-free-lunch (NFL) theorems in ML in 1996 in paper "The lack of a Priori Distinctions between learning algorithms".
- NFL theorems in optimization theory by David Wolpert and William Macready appears in the 1997 in "No Free Lunch Theorems for Optimization" paper.
- https://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf
- Those are fundamental theorems of optimization theory and Machine Learning.

NFL THEOREM FOR OPTIMIZATION

- NFL theorem for search and optimization (Wolpert and Macready 1997) applies to finite spaces and algorithms that do not resample points.
- **Theorem:** All algorithms that search for an extremum of a cost function perform exactly the same when averaged over all possible cost functions.
- So, for any search/optimization algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class.

POPULAR INSIGHTS FROM NFL THEOREM

- NFL theorem states that any two algorithms are equivalent when their performance is averaged across all possible problems.
- If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.
- All algorithms have identically distributed performance when objective functions are drawn uniformly at random.
- All algorithms have identical mean performance.

OCKHAM'S RAZOR
PROBLEM SOLVING PRINCIPLE

- Among competing hypotheses, select the one with the fewest assumptions
- Among competing models, select the one with the fewest parameters
- Other things being equal, simpler explanations are generally better than more complex ones
- In the related concept of overfitting, excessively complex models are affected by statistical noise, whereas simpler models may capture the underlying structure better and may thus have better predictive performance