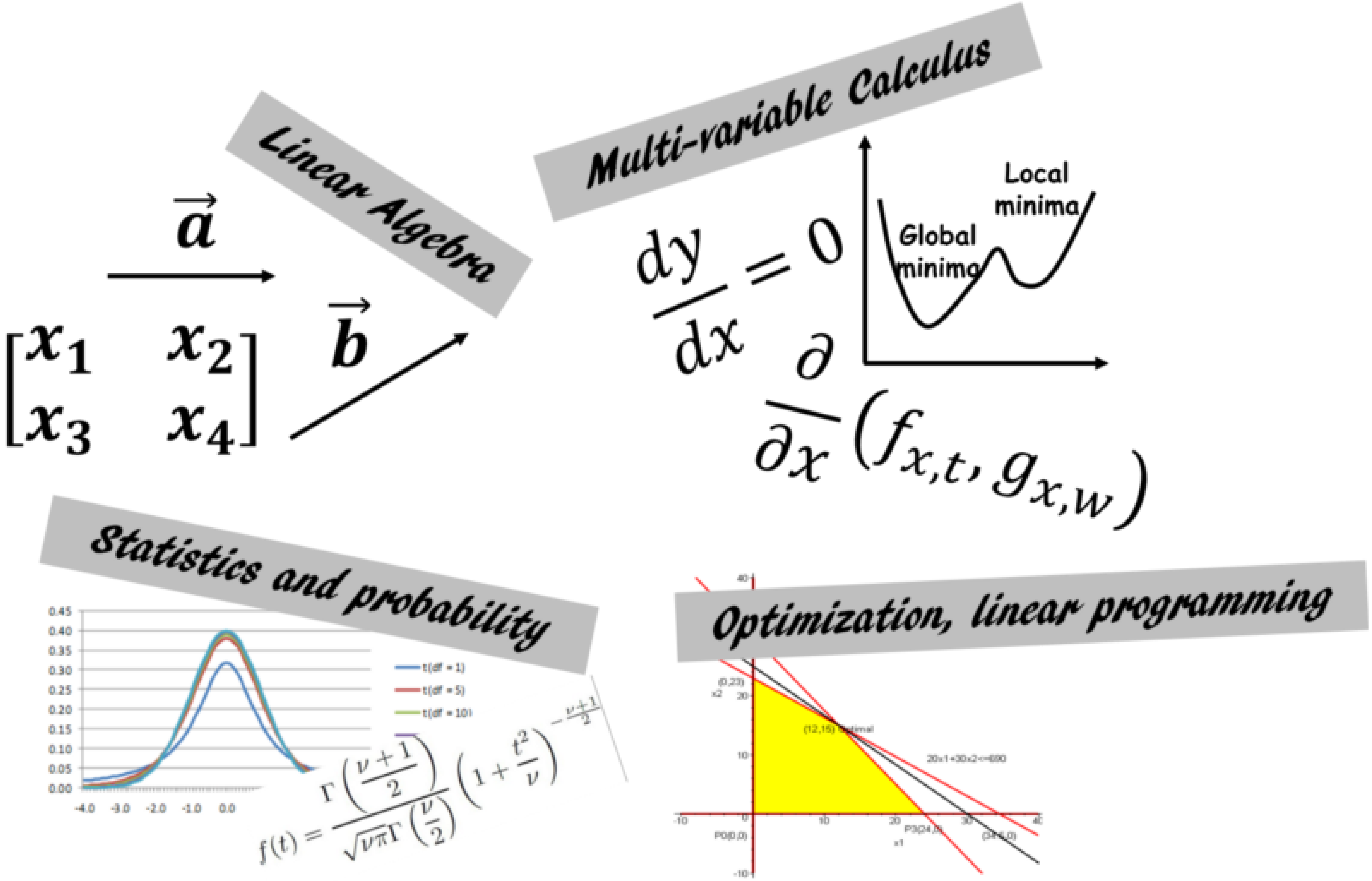
BACKGROUND OF DATA MINING

Week02

Essential Math for Data Mining



Linear Algebra

Multi-variable Calculus

Statistics and probability

Optimization, linear programming

$\vec{a}$ $\vec{b}$

$$\begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix}$$

$$\frac{dy}{dx} = 0$$

$$\frac{\partial}{\partial x}\left(f_{x,t}, g_{x,w}\right)$$

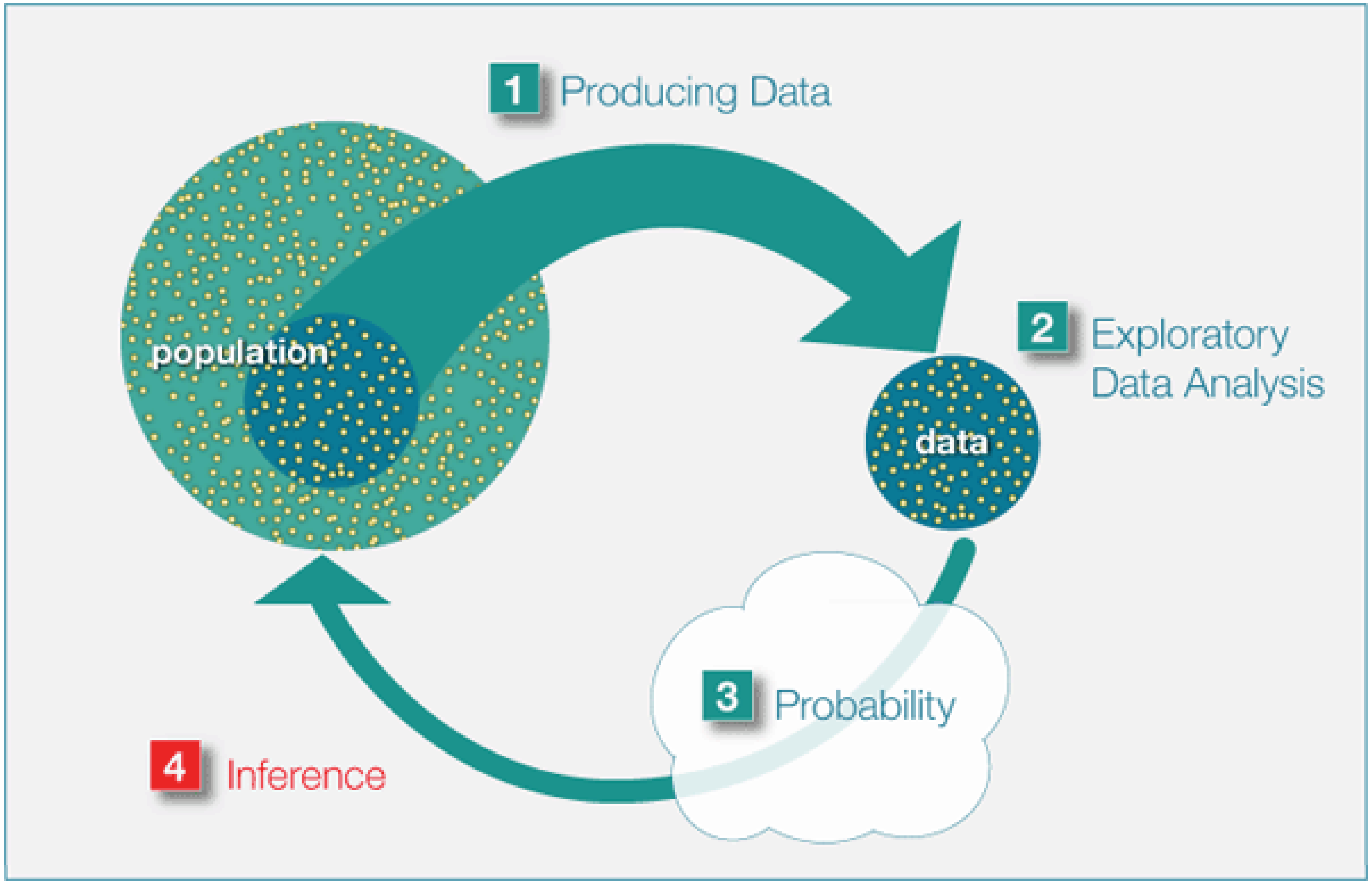
$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Essential Math for Data Mining: Statistics

- Two main branches of statistics
  - Descriptive statistics
    - Describe the basic features of data
    - Data summaries and descriptive statistics, central tendency, variance, covariance, correlation
  - Inferential statistics
    - Deduce properties of an underlying distribution of probability

- Probability
  - Sampling, measurement, error, random number generation
  - Basic probability: basic idea, expectation, probability calculus, Bayes theorem, conditional probability
  - Probability distribution functions—uniform, normal, binomial, chi-square, student's t-distribution, Central limit theorem

Essential Math for Data Mining: Statistics



Essential Math for Data Mining: Linear Algebra

- Linear algebra
  - The study of linear sets of equations and their transformation properties
  - Concern linear equations, linear functions and their representations in vector spaces and through matrices
  - Used in most areas of science and engineering, because it allows modeling many natural phenomena, and efficiently computing with such models

$$\begin{aligned} 3x + 5y &= 7 \\ x - 2y &= 6 \end{aligned}$$

→

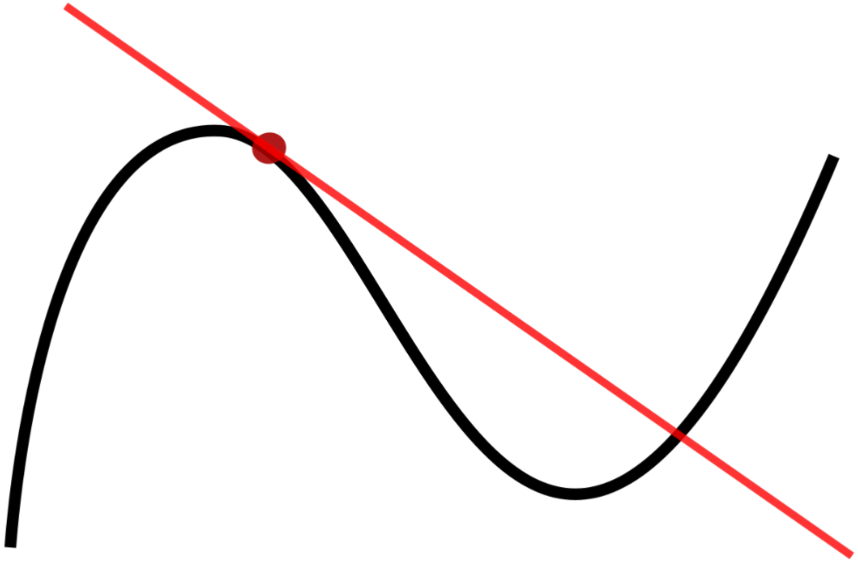
$$\begin{bmatrix} 3 & 5 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 7 \\ 6 \end{bmatrix}$$

↓

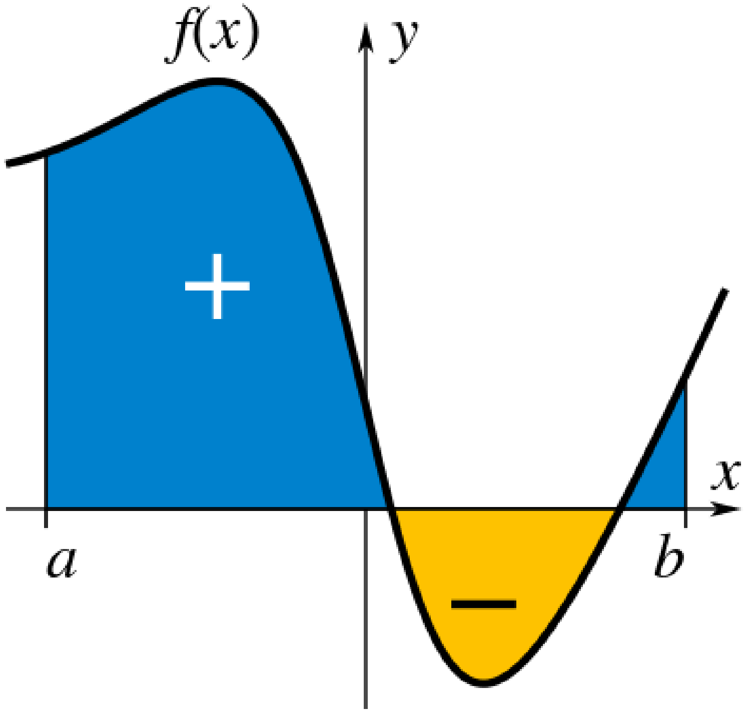
$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 & 5 \\ 1 & -2 \end{bmatrix}^{-1} \begin{bmatrix} 7 \\ 6 \end{bmatrix}$$

Essential Math for Data Mining: Calculus

- Calculus
  - Branch of mathematics concerned with the calculation of instantaneous rates of change (differential calculus) and the summation of infinitely many small factors to determine some whole (integral calculus)



differential calculus



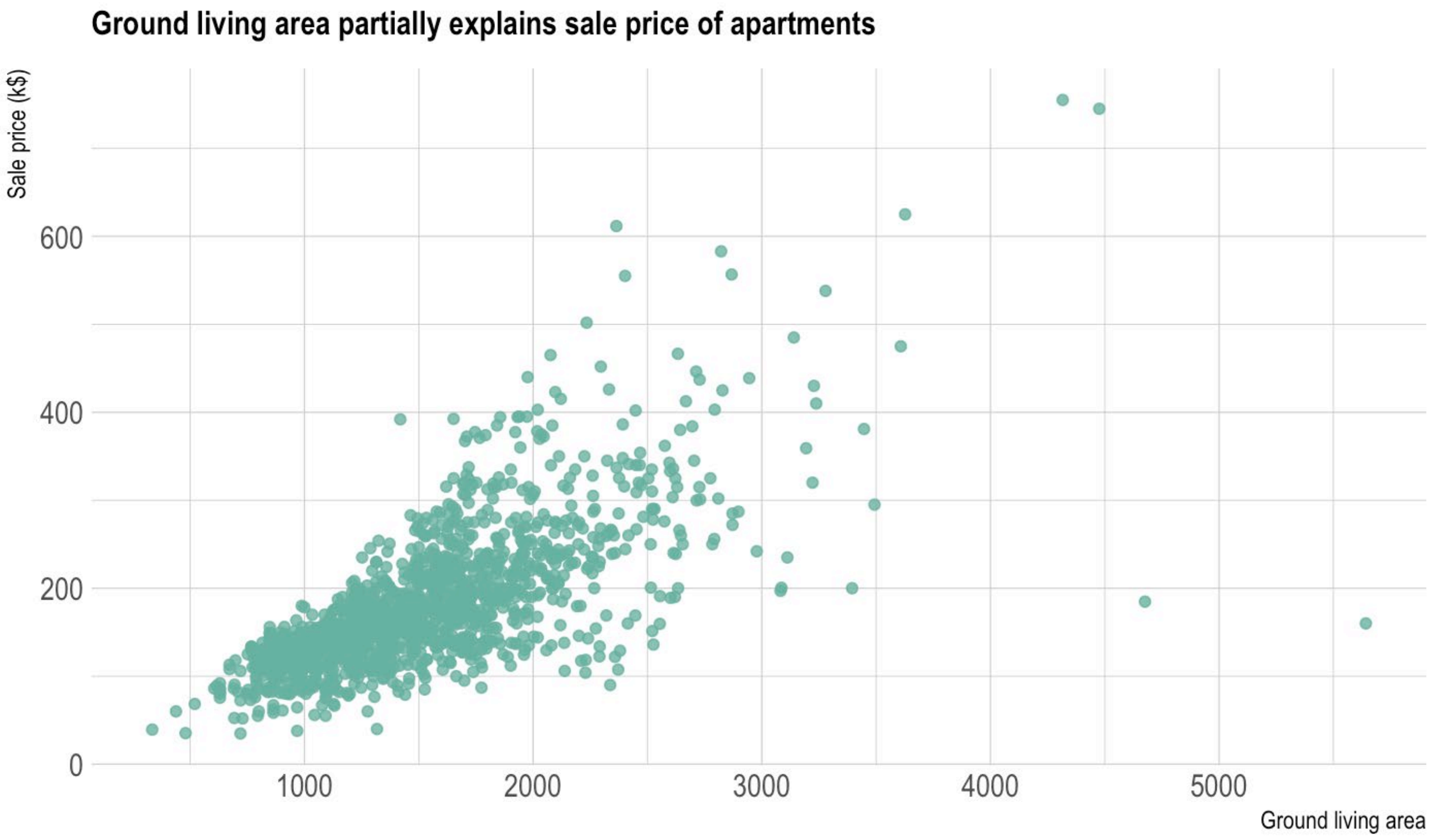
integral calculus

Essential Math for Data Mining: Optimization

- Optimization
  - Collection of mathematical principles and methods used for solving optimization problems
  - Optimization problem is the problem of fining the best solution from all feasible solutions
    - In the simplest case, an optimization problem consists of maximizing or minimizing a real function

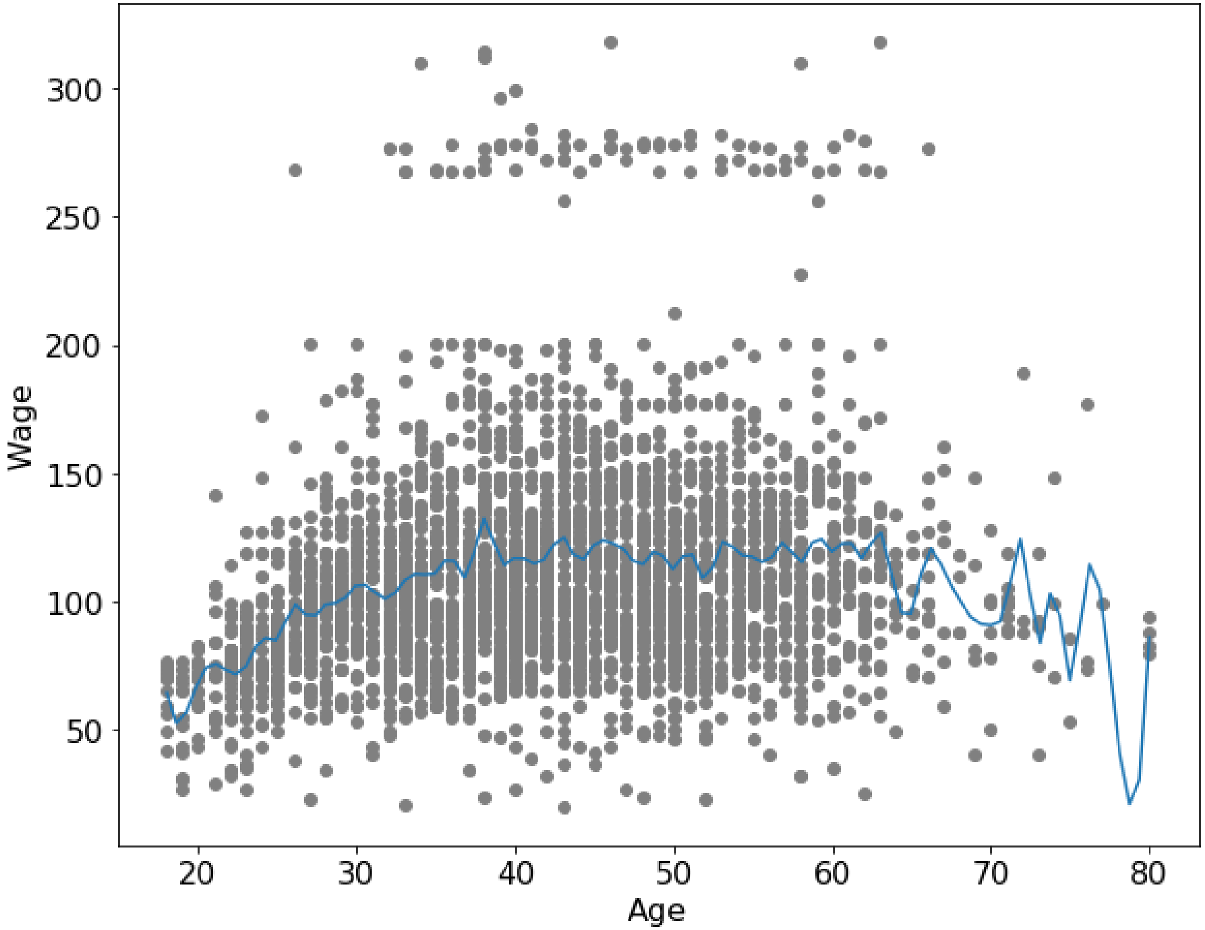
Statistics

- A vast set of tools for understanding data



Statistics

- □ A vast set of tools for understanding data

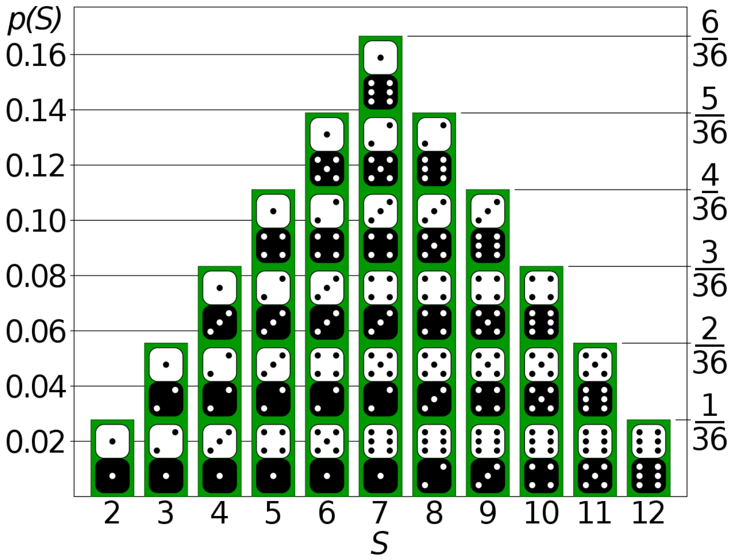


Statistics

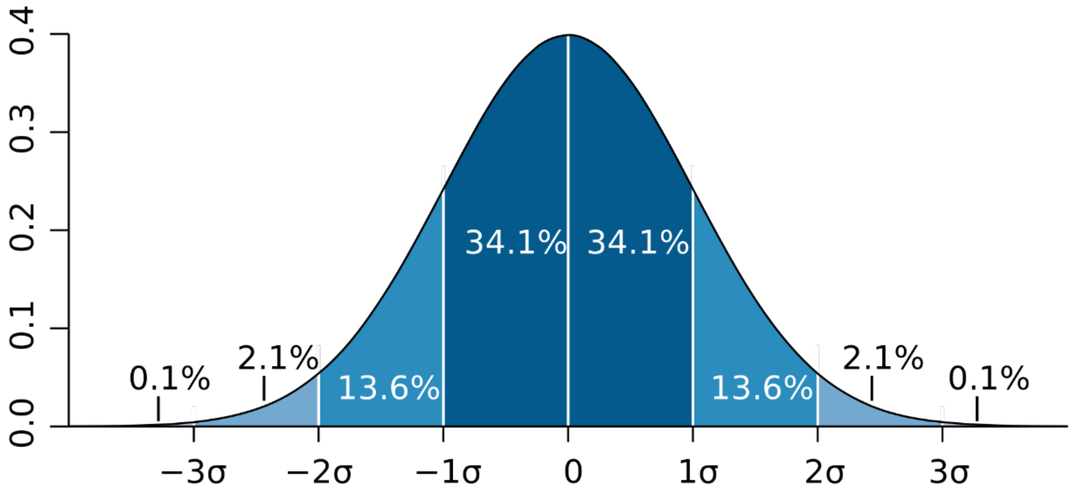
- Descriptive statistics
  - A summary statistic that quantitatively describes or summarizes features of a collection of information
  - Univariate
    - Mean, Median, Mode
    - Variance, standard deviation, Percentile
    - Skewness, kurtosis
  - Bivariate or multivariate
    - Cross-tabulations and contingency tables
    - Graphical representation via scatterplots
    - Quantitative measures of dependence (covariance, correlation)

Statistics

- Probability distribution
  - A mathematical function that provides the probabilities of occurrence of different possible outcomes
    - The probabilities of occurrence of the specific observations



Discrete random variable → probability mass function



Continuous random variable → probability density function

Statistics: Discrete Probability Distributions

- Bernoulli distribution
  - The discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$

    $$\Pr(X = 1) = p = 1 - \Pr(X = 0) = 1 - q$$
  - A special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution)
  - Probability mass function

    $$f(X = x; p) = p^k (1 - p)^{1-k}$$
- Binomial distribution
  - The discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes-no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability $q = 1 - p$).
  - Probability mass function

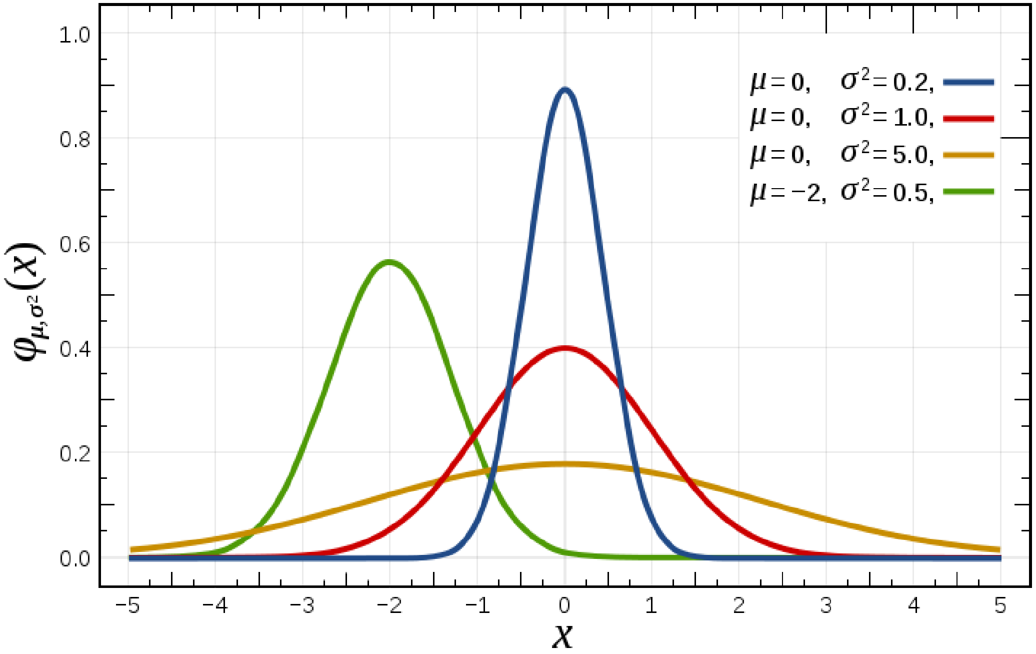
    $$f(X = k; n, p) = \Pr(X = k) = \frac{n!}{k!\,(n-k)!} p^k (1 - p)^{n-k}$$

Statistics: Continuous Distributions

- Normal (Gaussian) distribution
  - Very common continuous probability distribution
  - Bell-shaped

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

    - μ: mean
    - σ: standard deviation



Statistics: Continuous Distributions

- Student's t-distribution (t-distribution)
  - Continuous probability distributions that arises when estimating the mean of a normally distributed population in situations where the **sample size is small** and **population standard deviation is unknown**

- Let $X_1, \dots, X_n$ be independent and identically distributed (iid) as $N(\mu, \sigma^2)$
  - Sample mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

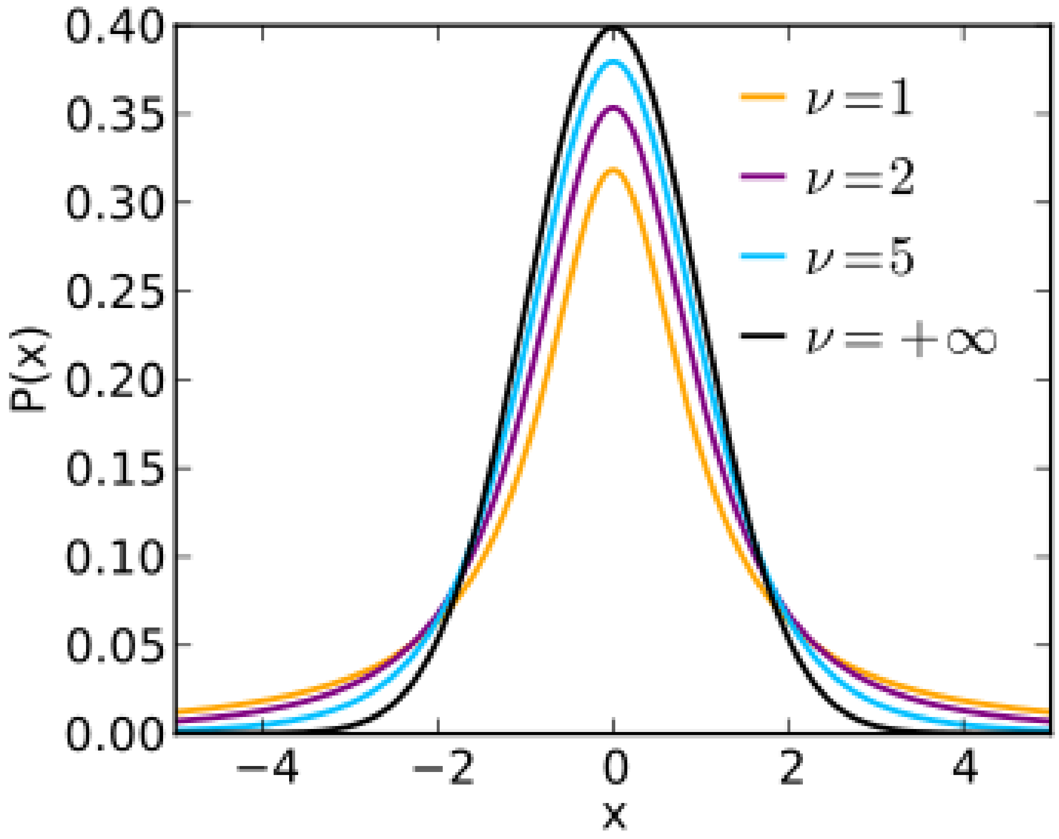
  - Sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

  - The random variable $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution
  - The random variable $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a Student's t-distribution with $n-1$ degrees of freedom

Statistics: Continuous Distributions

- The probability density function of t-distribution with varying degree of freedom



Statistics: Student's t-distribution

- Probability density function

$$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

  - ν: degree of freedom
  - Γ: gamma function

$$\Gamma(n) = (n-1)! \quad \text{if } n \text{ is positive integer}$$

$$\Gamma(z) = \int_0^{-\infty} x^{z-1}e^{-x}dx$$

Statistics: Continuous Distributions

- Chi-squared distribution (χ^2)
  - The distribution of a sum of the squares of k independent standard normal random variables

- Let $X_1, \ldots, X_k$ be independent, standard normal random variables

$$Y = \sum_{i=1}^{k} X_i^2$$

is distributed according to the chi-squared distribution with k degrees of freedom

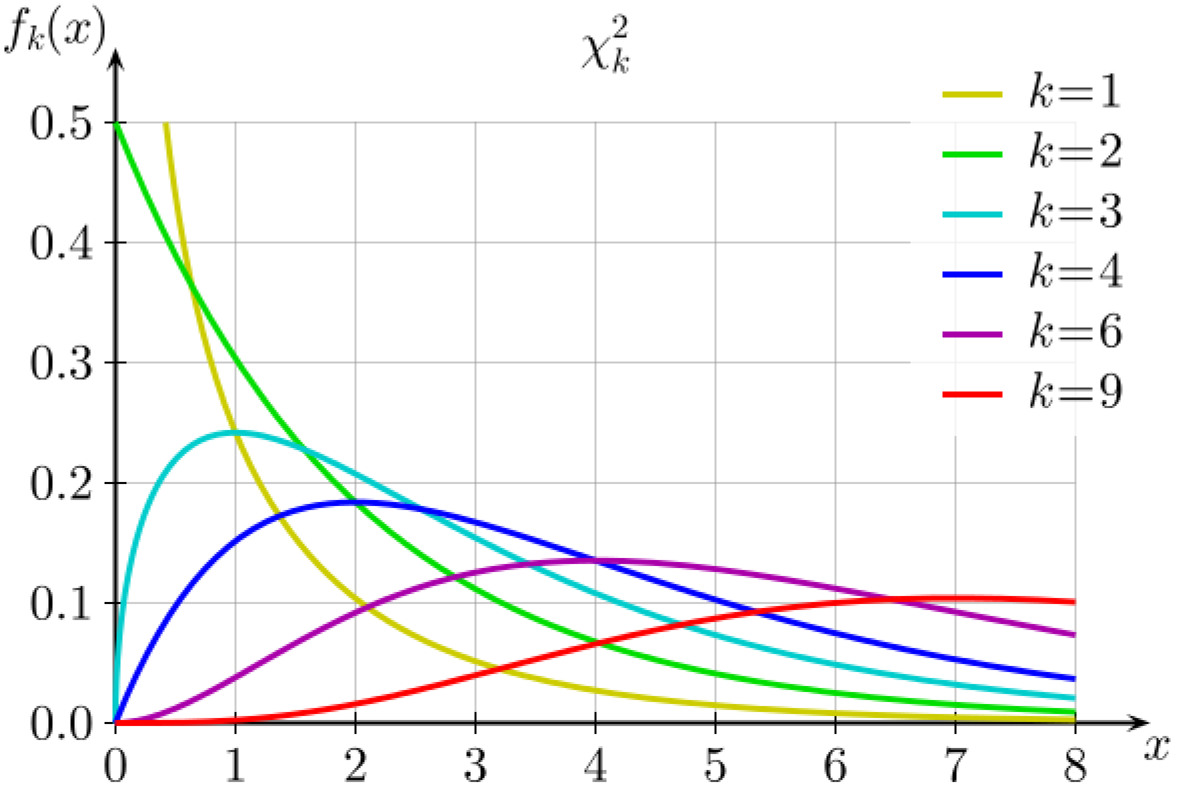
$$Y \sim \chi^2(k) \;\; or \; Y \sim \chi_k^2$$

- Probability density function

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

Statistics: Continuous Distributions

- The probability density function of chi-squared distribution with varying degree of freedom



Statistics:Continuous Distributions

- F-distribution
  - A random variate of the F-distribution with parameters d_1 and d_2 arises as the ratio of two appropriately scaled chi-squared variates
- Let X_1 and X_2 be two independent random variables and $X_1 \sim \chi^2(d_1)$ and $X_2 \sim \chi^2(d_2)$

$$Y = \frac{X_1/d_1}{X_2/d_2}$$

is distributed according to the F-distribution with d_1 and d_2 degrees of freedom

$$Y \sim F(d_1, d_2)$$

- Probability density function

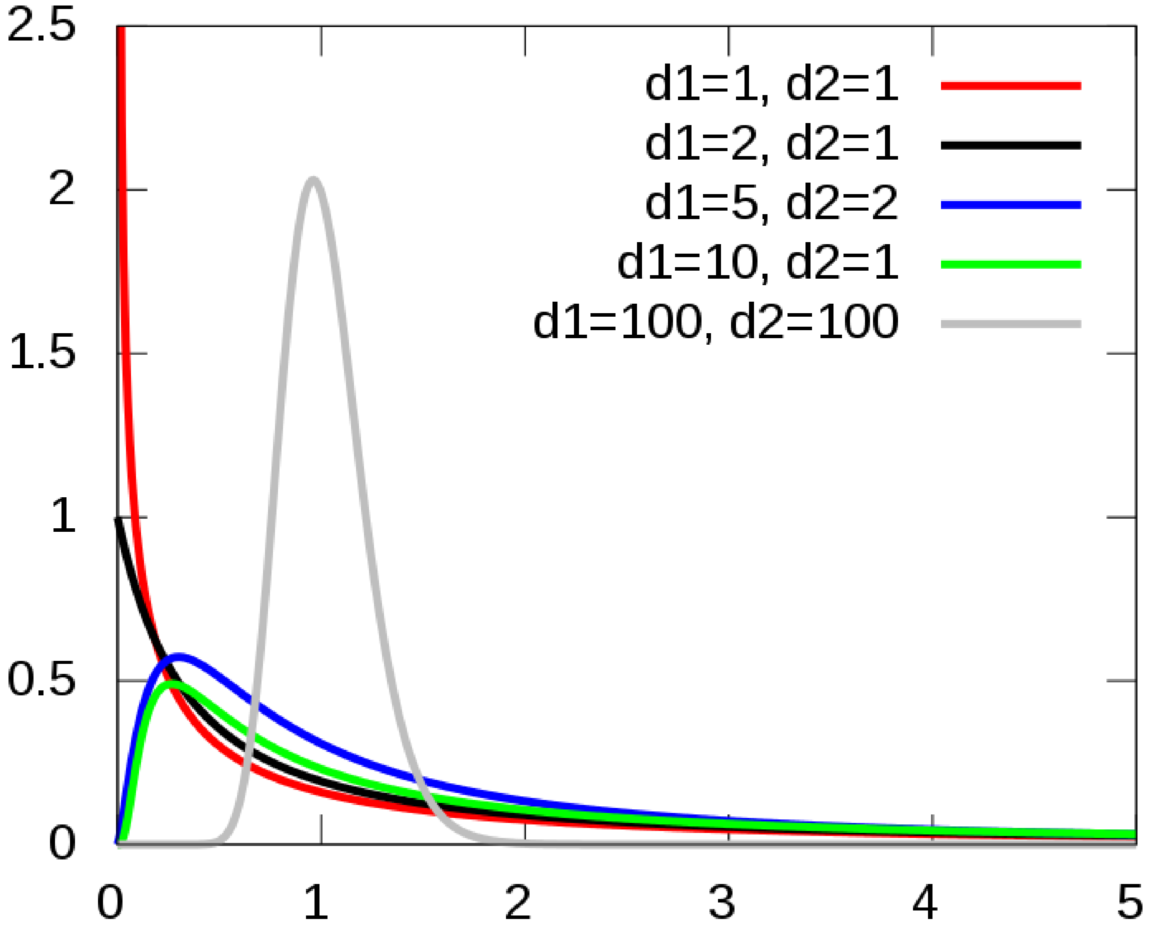
$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

  - B: beta function

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$$

Statistics: Continuous Distributions

- The probability density function of F-distribution with varying degree of freedom



Linear Algebra

- Linear Algebra
  - Basic properties of matrix and vectors—scalar multiplication, linear transformation, transpose, conjugate, rank, determinant
  - Inner and outer products, matrix multiplication rule and various algorithms, matrix inverse
  - Special matrices—square matrix, identity matrix, triangular matrix, idea about sparse and dense matrix, unit vectors, symmetric matrix, Hermitian, skew-Hermitian and unitary matrices
  - Gaussian/Gauss-Jordan elimination, solving Ax=b linear system of equation
  - Matrix factorization and decomposition
  - Vector space, basis, span, orthogonality, orthonormality, linear least square
  - Eigenvalues, eigenvectors, and diagonalization, singular value decomposition (SVD)

Linear Algebra

- Linear algebra is the study of vectors and linear functions
  - Scalar
    - A scalar is a number
  - Vector
    - A vector is a list of numbers

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

  - Matrix
    - A matric is also a collection of numbers
    - The difference is that a matrix is a table of numbers rather than a list

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

  - Linear equation

$$a_1 x_1 + \cdots + a_n x_n = b$$

  - Linear function

$$(x_1, \ldots, x_n) \longmapsto a_1 x_1 + \cdots + a_n x_n$$

Linear Algebra

- Vectors
  - Addition

$$\mathbf{v} + \mathbf{w}$$

    - Example

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

  - Linear combination

$$a\mathbf{v} + b\mathbf{w}$$

    - Example

$$3\mathbf{v} + 4\mathbf{w} = 3\begin{bmatrix} 1 \\ 2 \end{bmatrix} + 4\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 15 \\ 22 \end{bmatrix}$$

Linear Algebra

- Vectors
  - Transpose
    - column vector ↔ row vector
    - Example

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \rightarrow \mathbf{v}^T = \begin{bmatrix} 1 & 2 \end{bmatrix}$$

  - Dot product, inner product

$$\mathbf{v} \cdot \mathbf{w}$$

    - Example

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

$$\mathbf{v} \cdot \mathbf{w} = (1)(3) + (2)(4) = 9$$

  - Length

$$\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$$

    - Example

$$\mathbf{v} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \|\mathbf{v}\| = \sqrt{(1)(1) + (2)(2)} = \sqrt{5}$$

Linear Algebra

- Matrix
  - Addition

$$A + B = C$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}$$

  - Multiplication

$$AB = D$$

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{bmatrix}$$

  - Linear equation

$$A\mathbf{x} = \mathbf{b}$$

    - Example

$$\begin{aligned} x_1 \qquad\qquad\qquad &= b_1 \\ -x_1 + x_2 \qquad\quad &= b_2 \\ -x_2 + x_3 &= b_3 \end{aligned}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Linear Algebra

- Matrix
  - Inverse matrix
    - An n-by-n square matrix, A is called invertible (or nonsingular) if there exists an n-by-n square matrix, B such that

      $$AB = BA = I_n$$

      where I_n denotes the n-by-n identity matrix which is a square matrix with ones on the main diagonal and zeros elsewhere
    - B is the inverse of A (A^{-1})
    - If A has no inverse, A is singular or non-invertible
    - Example

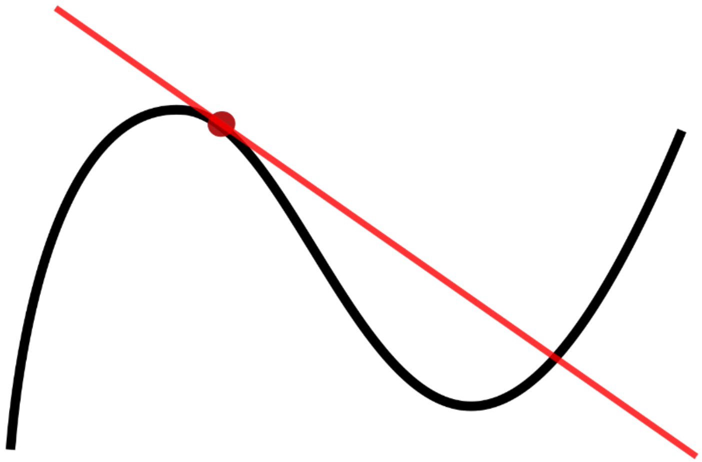
      $$A = \begin{bmatrix} -1 & \frac{3}{2} \\ 1 & -1 \end{bmatrix}, A^{-1} = \begin{bmatrix} 2 & 3 \\ 2 & 2 \end{bmatrix}$$
  - Solution of a linear equation

    $$\mathbf{x} = A^{\mathbf{-1}}\mathbf{b}$$

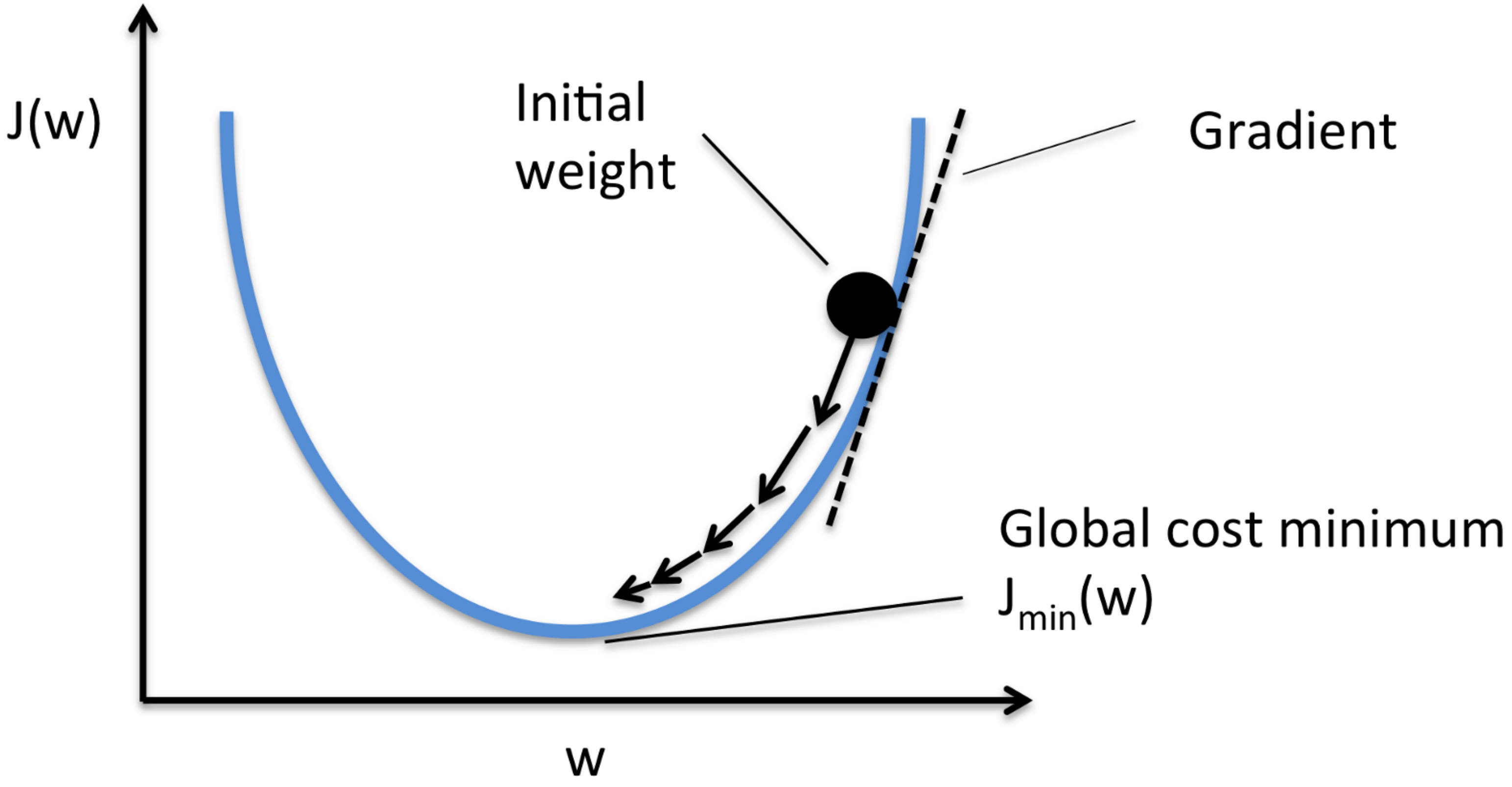
Calculus

- Derivative
  - A function of a real variable measures the sensitivity to change of the function value (output value) with respect to a change in its argument (input value)

$$\frac{dy}{dx}$$



Calculus

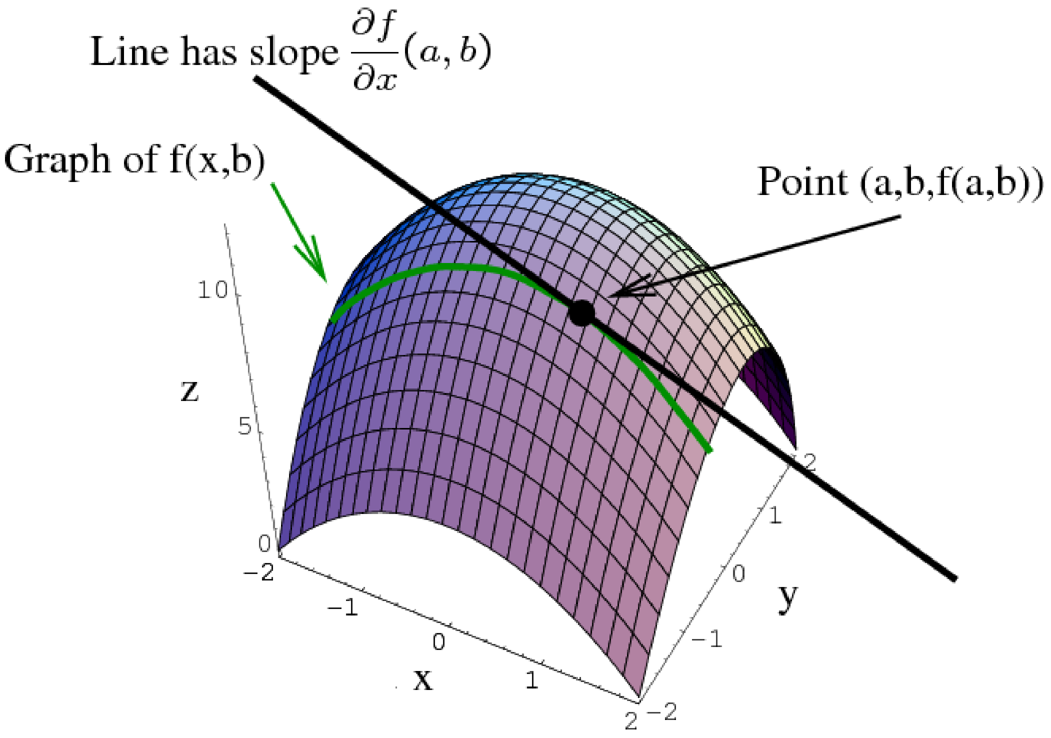


https://en.wikipedia.org/wiki/File:Gradient_Descent_in_2D.webm

Calculus

- Partial derivative
  - A function of several variables is its derivative with respect to one of those variables, with the others held constant

$$\frac{\partial f}{\partial x}$$

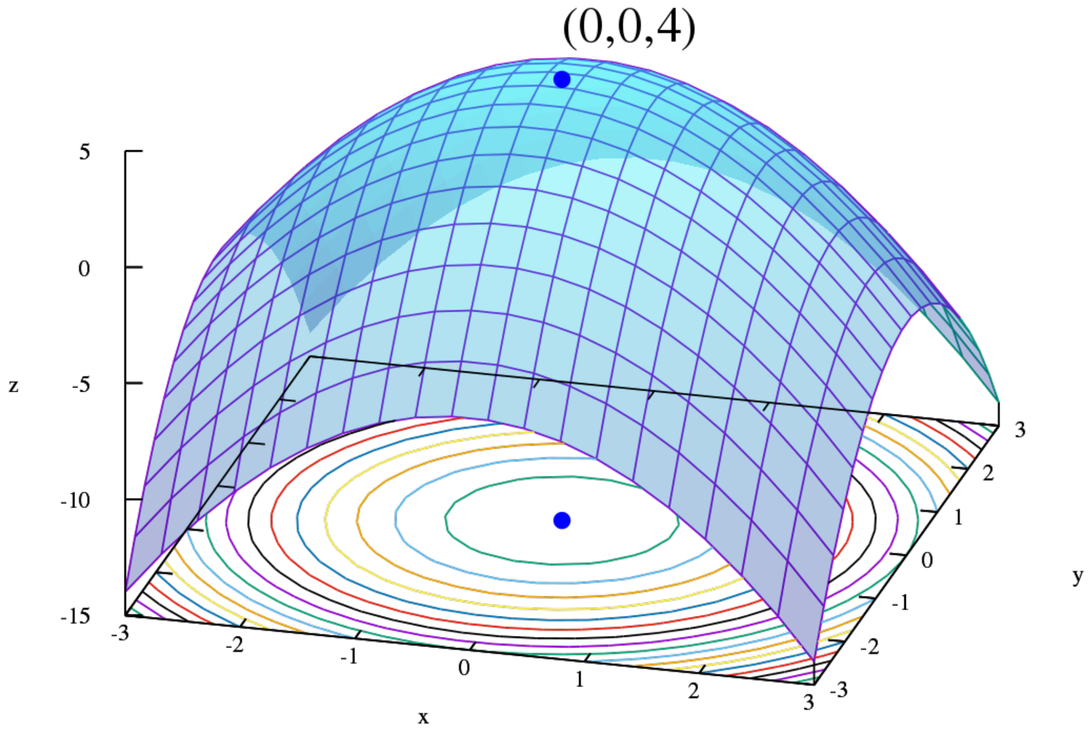


Optimization

- Optimization
  - Basics of optimization —how to formulate the problem
  - Linear programming, simplex algorithm
  - Integer programming
  - Constraint programming, knapsack problem
  - Randomized optimization techniques—hill climbing, simulated annealing, Genetic algorithms

Optimization

- Optimization problem
  - Maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function



Optimization

- Example
  - For materials, the manufacturer has 750 ㎡ of cotton textile and 1,000 ㎡ of polyester. Every pair of pants (1 unit) needs 1 ㎡ of cotton and 2 ㎡ of polyester. Every jacket needs 1.5 ㎡ of cotton and 1 ㎡ of polyester.
  - The price of the pants is fixed at $50 and the jacket, $40.
  - **What is the number of pants and jackets that the manufacturer must give to the stores so that these items obtain a maximum sale?**
    - Variables to be determined

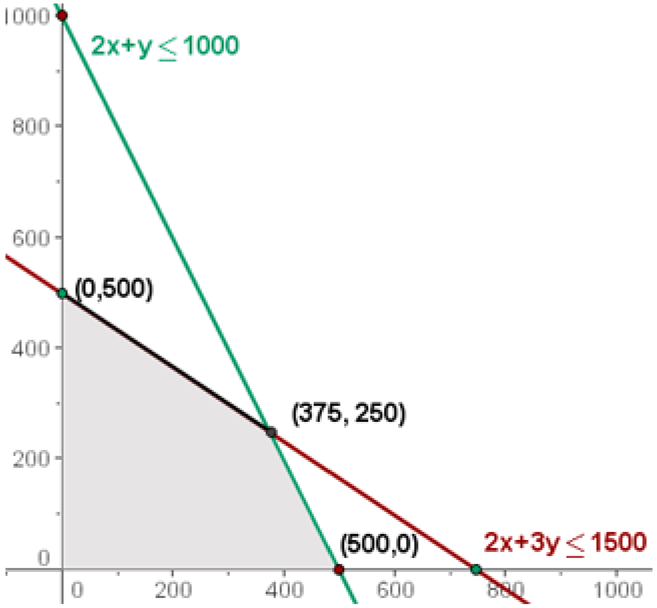
$$x = number\ of\ pants$$
$$y = number\ of\ jackets$$

    - Objective function

$$f(x, y) = 50x + 40y$$

    - Constraints

$$x + 1.5y \leq 750$$
$$2x + y \leq 1000$$



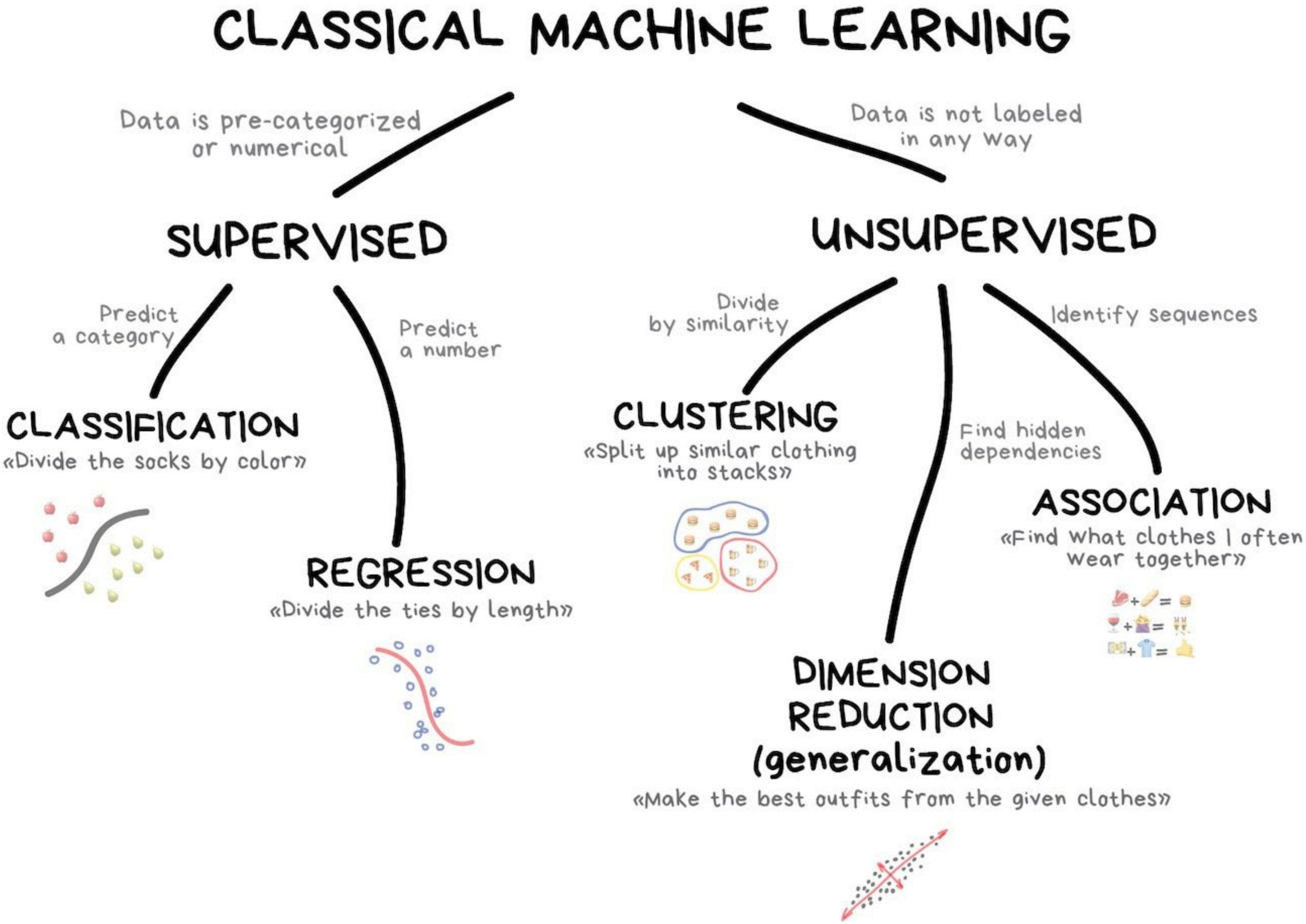
Optimization

- Why are optimization algorithms important for data analysis?
  - One of fundamental data analysis tasks is to seek a function that approximately maps $\mathbf{x}_i$ to y_i for each observation, i

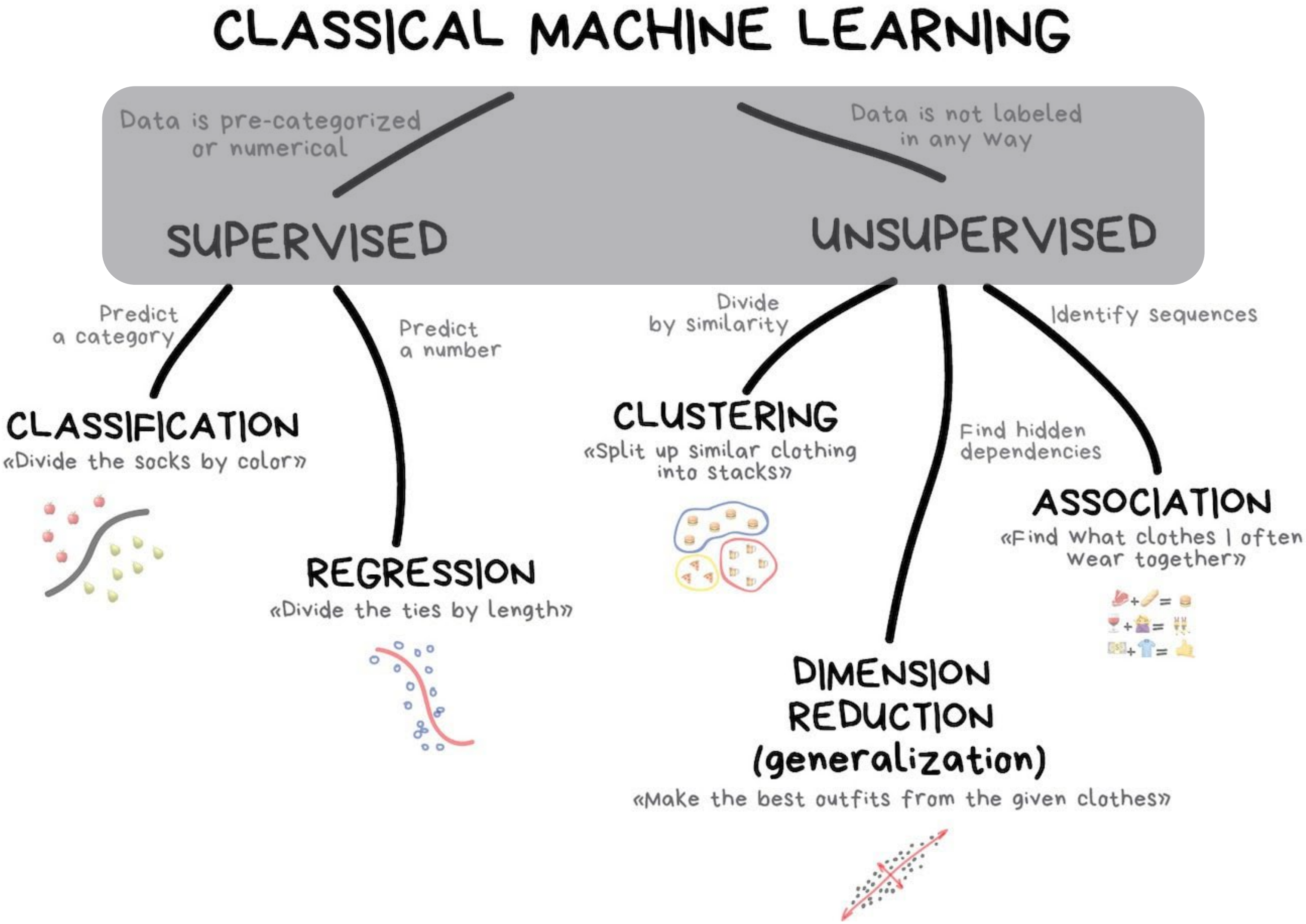
    $$y = f(\mathbf{x})$$
  - The process of finding f based on data is called learning or training
    - During learning, optimization algorithms provide a tool to find the most appropriate f

Basic Terminologies

Topics Covered in This Class


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
Find hidden dependencies
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

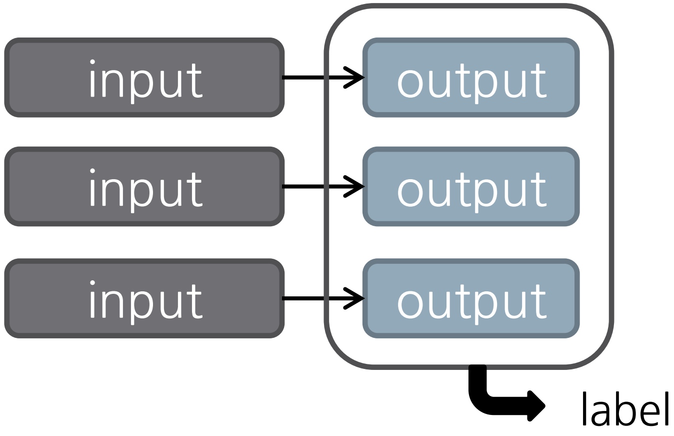
Topics Covered in This Class


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

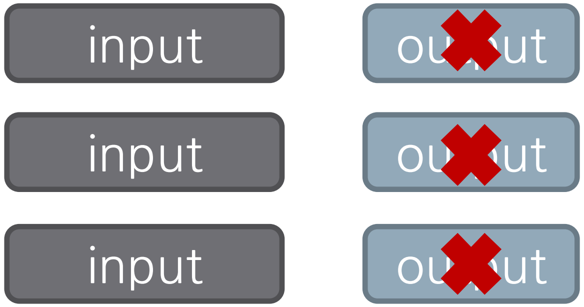
Types of Learning

- Supervised learning
  - We have knowledge of output
    - We call such data labeled

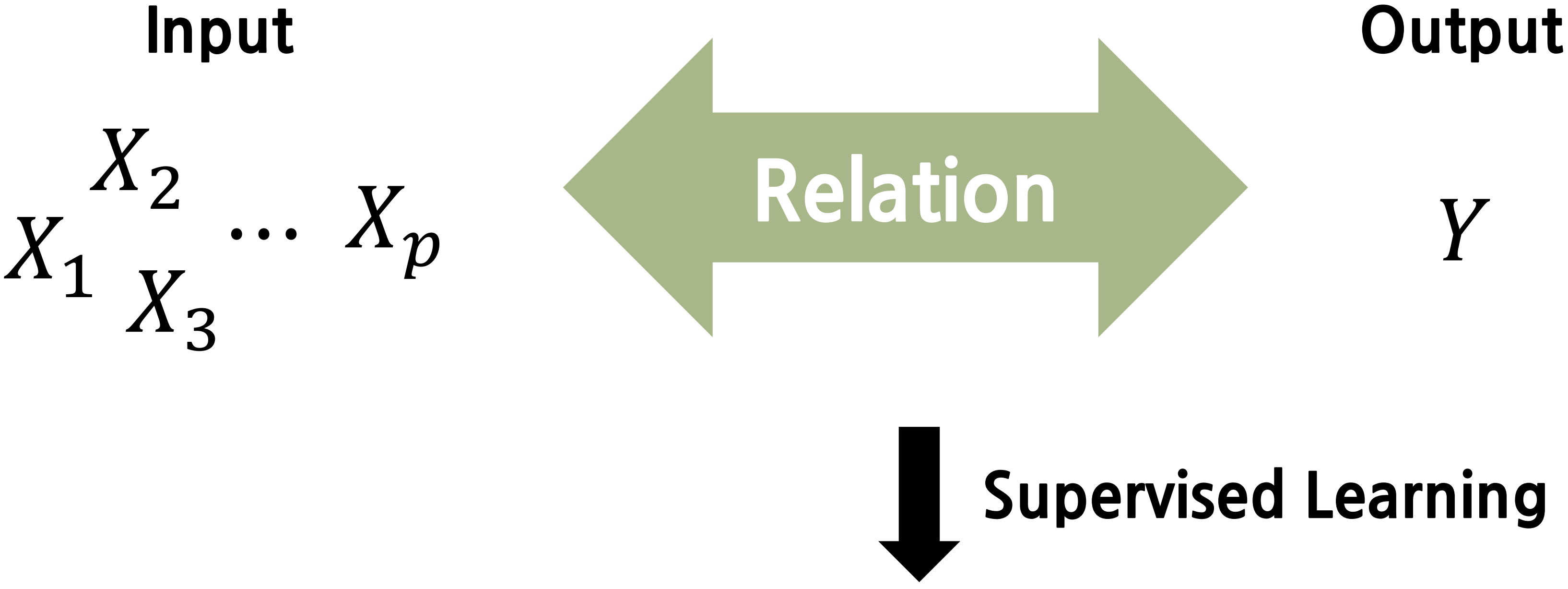
    → We know answer
  - Goal
    - Estimate output for unlabeled input



- Unsupervised learning
  - No output
    - We call such data unlabeled
  - Goal
    - Find patterns, groups, or relation



Supervised learning



Input

X_2 X_1 X_3 $\cdots$ X_p

Relation

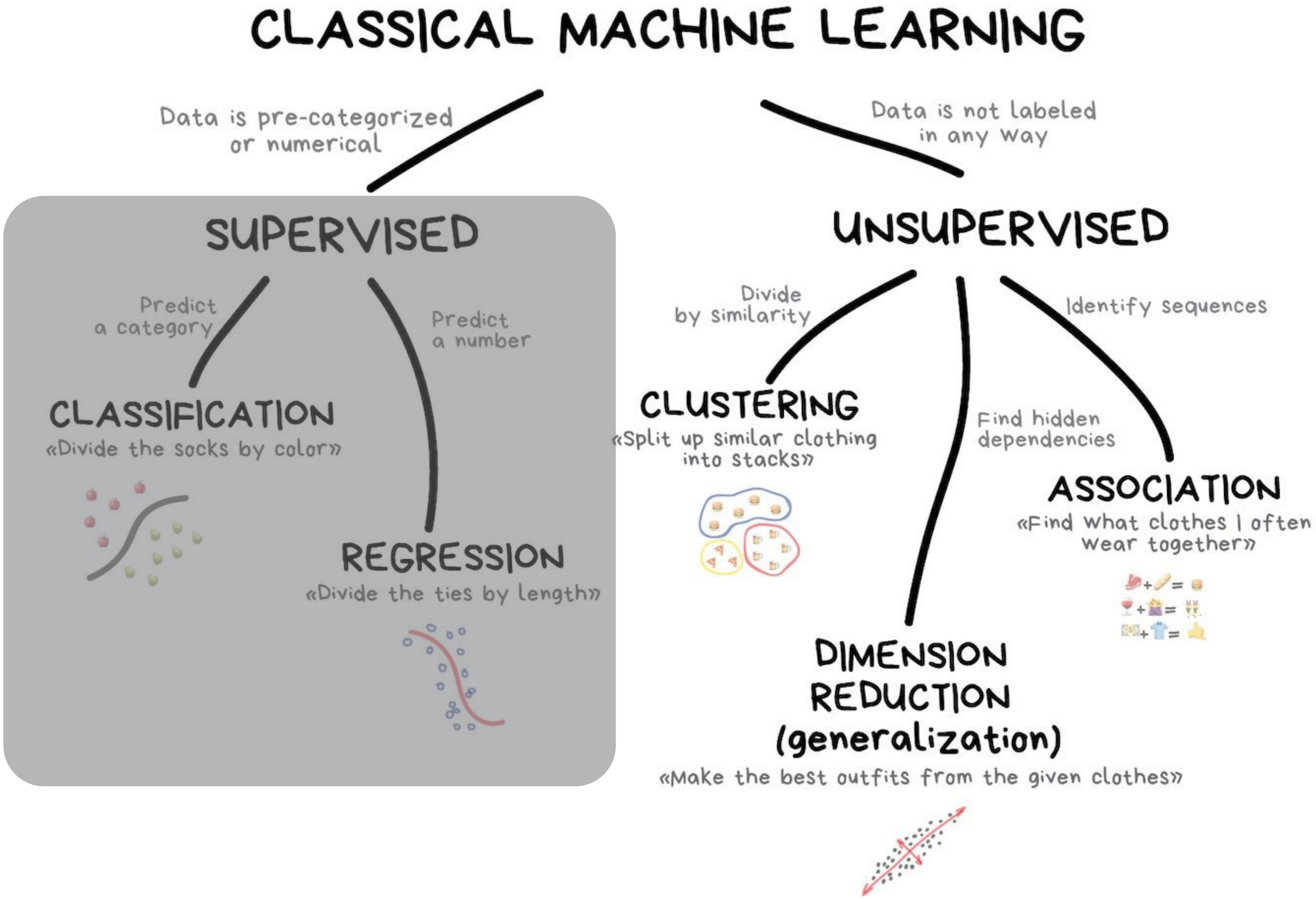
Output

Y

Supervised Learning

$$Y = f(X_1, X_2, \cdots, X_p)$$

Topics Covered in This Class


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

Data for Data Mining: Structured Data

- Example of data set
  - The input data set is usually expressed as a set of independent instances

instance, sample, example

Outlook	Temperature(℉)	Humidity(%)	Windy	Play Time(min)
Sunny	85	85	false	5
Sunny	80	90	true	0
Rainy	70	96	false	40
Rainy	68	80	false	65
Sunny	72	95	false	0
Sunny	69	70	false	70
Rainy	75	80	true	45

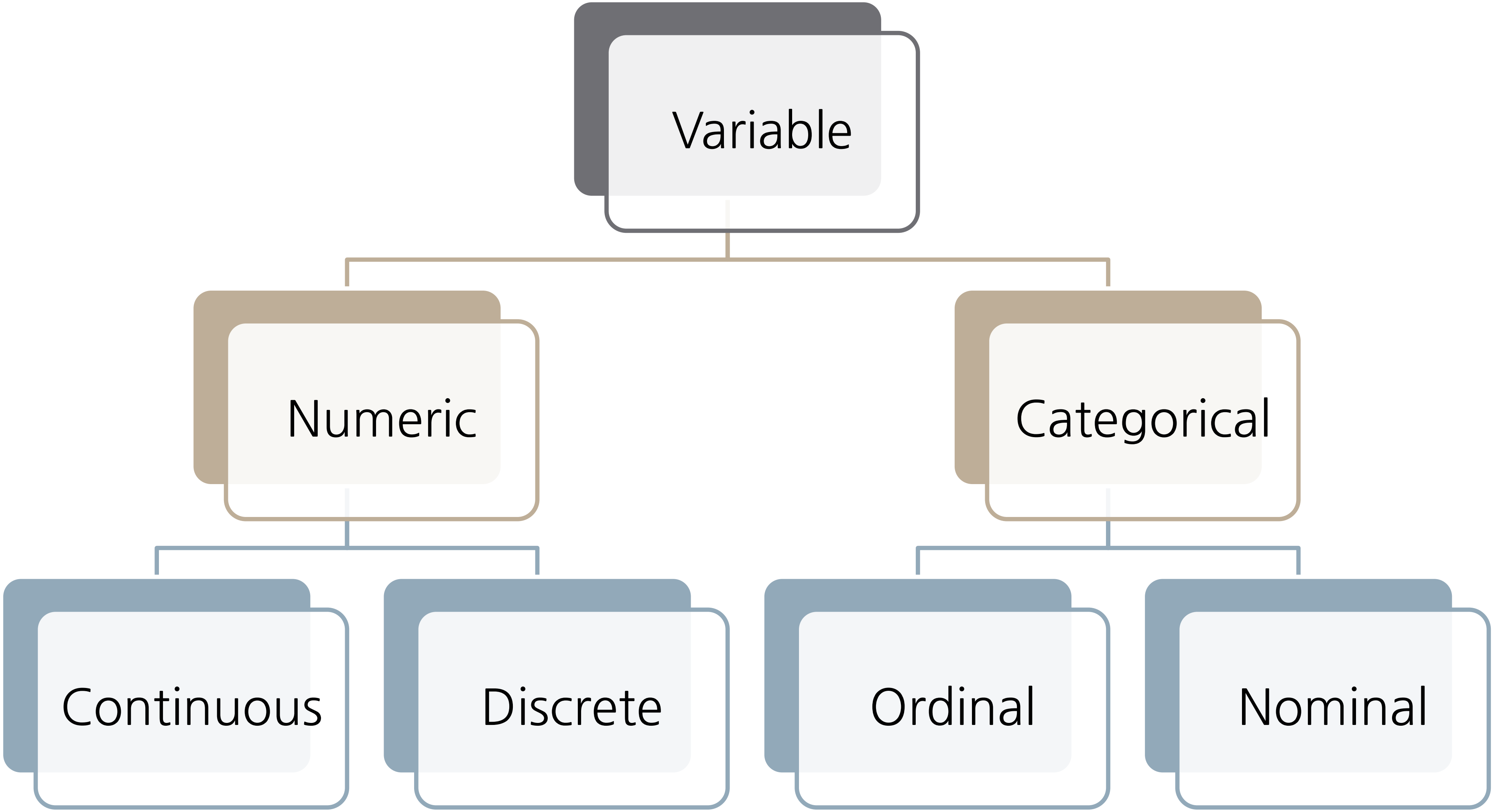
variable, attribute, feature

Types of Data

- Structured
  - Values of variable reside in a fixed field
  - Examples
    - Numeric
    - Date
    - Restricted terms: (male, female), (Mr., Ms., Mrs.)
    - Address

- Unstructured
  - Values of variable do not reside in a fixed field
  - Examples
    - Documents
    - Webpages
    - Images
    - Videos

Structured Data: Types of Variables



Structured Data: Types of Variables

- Numeric (Quantitative)
  - A broad category that includes any variable that can be counted, or has a numerical
- Continuous
  - A variable with infinite number of values
  - Example
    - Many numeric variables: temperature, weight, height, pressure and etc.
- Discrete
  - A variable that can only take on a certain number of values or have a countable number of values between any two values
  - Example
    - The number of cars in a parking lot
    - the number of flaws or defects

Structured Data: Types of Variables

- Categorical
  - A variable that contains a finite number of categories or distinct groups
- Nominal
  - A Variable that has two or more categories, but there is no intrinsic ordering to the categories.
  - Example
    - (Male, Female), (Class 1, Class 2, Class 3), (Red, Yellow, Green)
- Ordinal
  - Similar to a nominal variable, but the difference between the two is that there is a clear ordering of the variables.
  - Example
    - Score: A+,A,A-,B+,B,B-,C+,C,C-,D,F
    - Size: S, M, L, XL, XXL

Example: The Input to a Data Mining

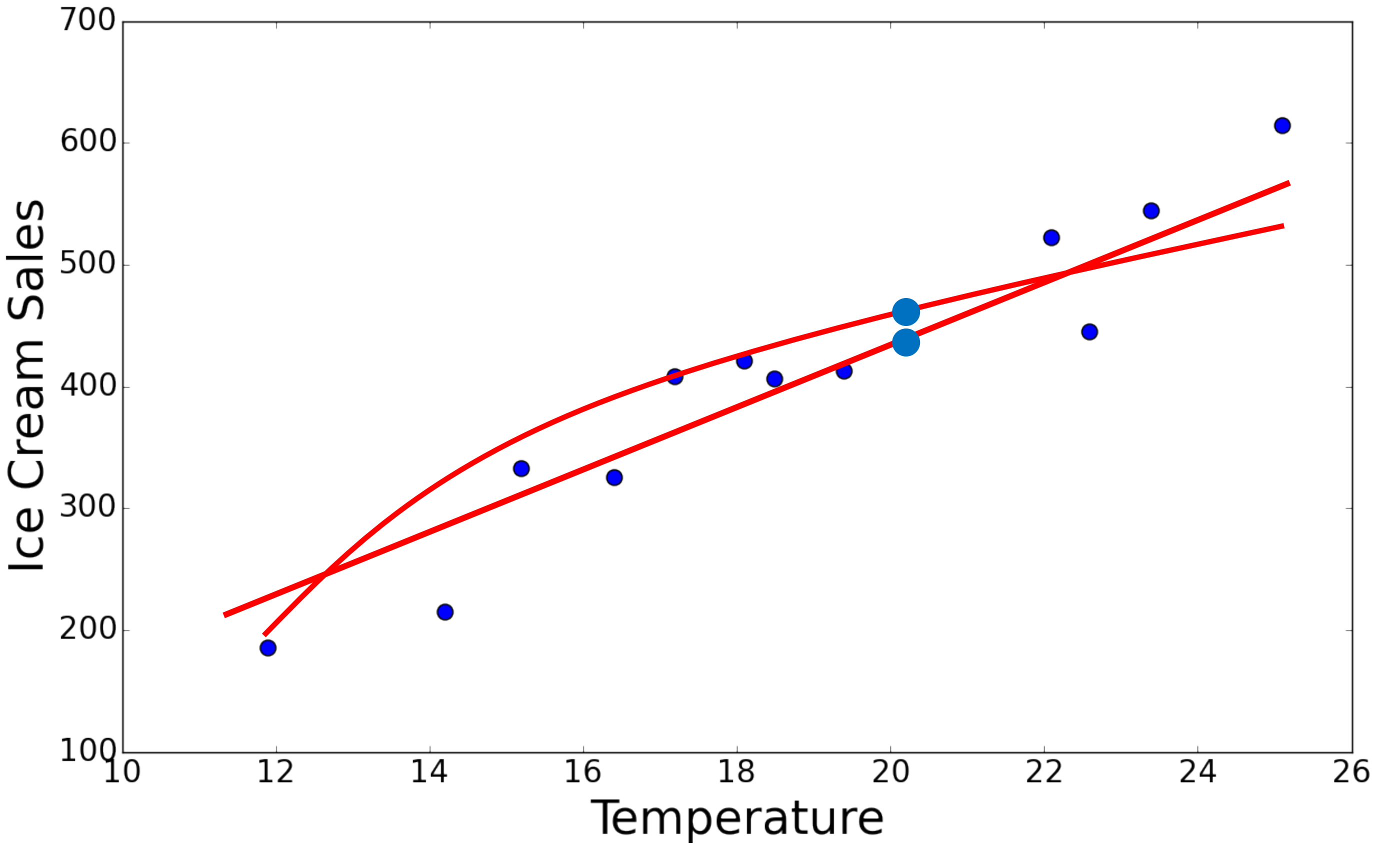
- Example of data set

num-of-doors	body-style	wheel-base	length	make
2	convertible	88.6	168.8	Audi
2	convertible	88.6	168.8	BMW
2	hatchback	94.5	171.2	Chevrolet
4	sedan	99.8	176.6	BMW
4	sedan	99.4	176.6	Audi
2	sedan	99.8	177.3	Audi
4	wagon	105.8	192.7	Chevrolet

Types: **Discrete** **Nominal** **Continuous** **Continuous** **Nominal**

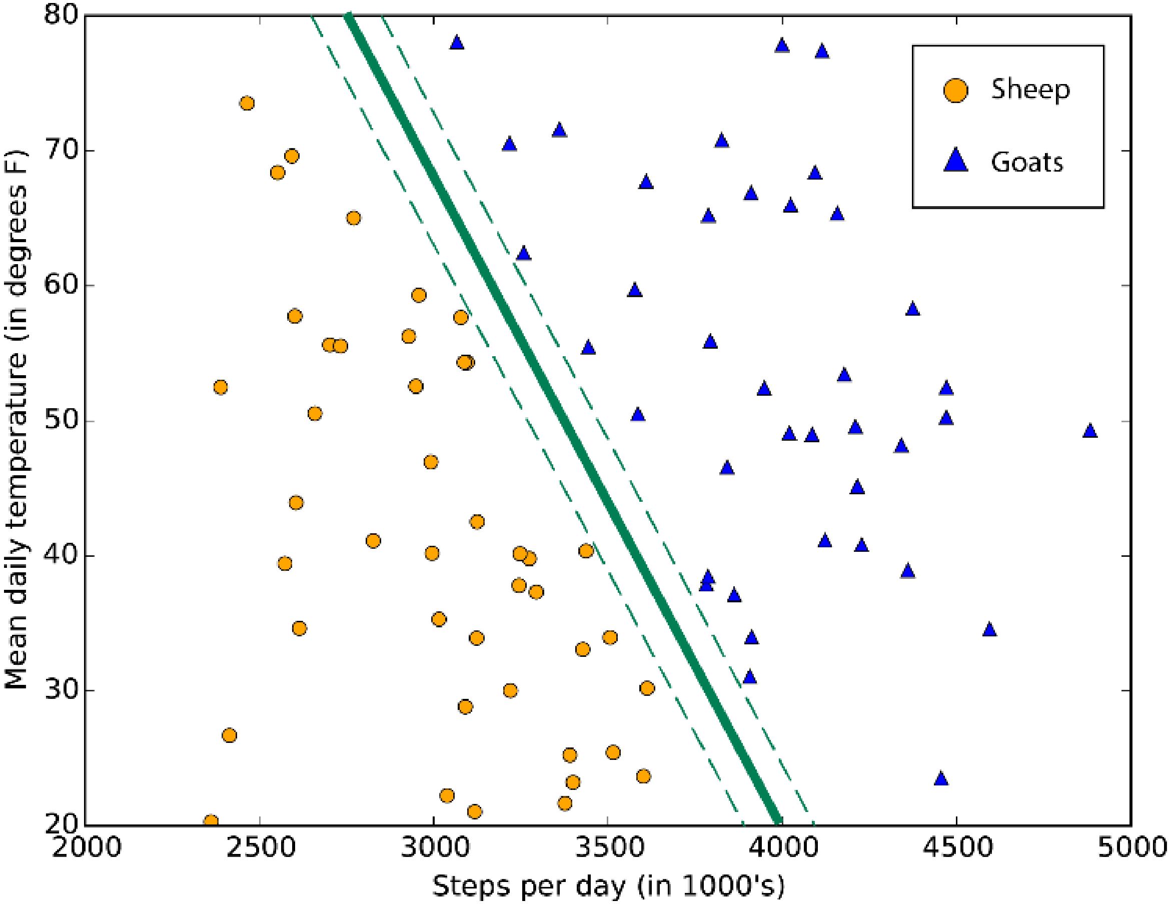
Supervised Learning: Regression

- Temperature vs. Ice Cream Sales
  - How about 21℃?



Supervised Learning: Classification

- Which one is a sheep?



https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-choice

Question

□ Suppose you are working on weather prediction, and you would like to predict whether or not it will be raining at 5pm tomorrow.
You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

① Regression

② Classification

□ Suppose you are working on stock market prediction, and you would like to predict the price of the specific stock tomorrow (measured in dollars).
You want to use a learning algorithm for this. Would you treat this as a classification or a regression problem?

① Regression

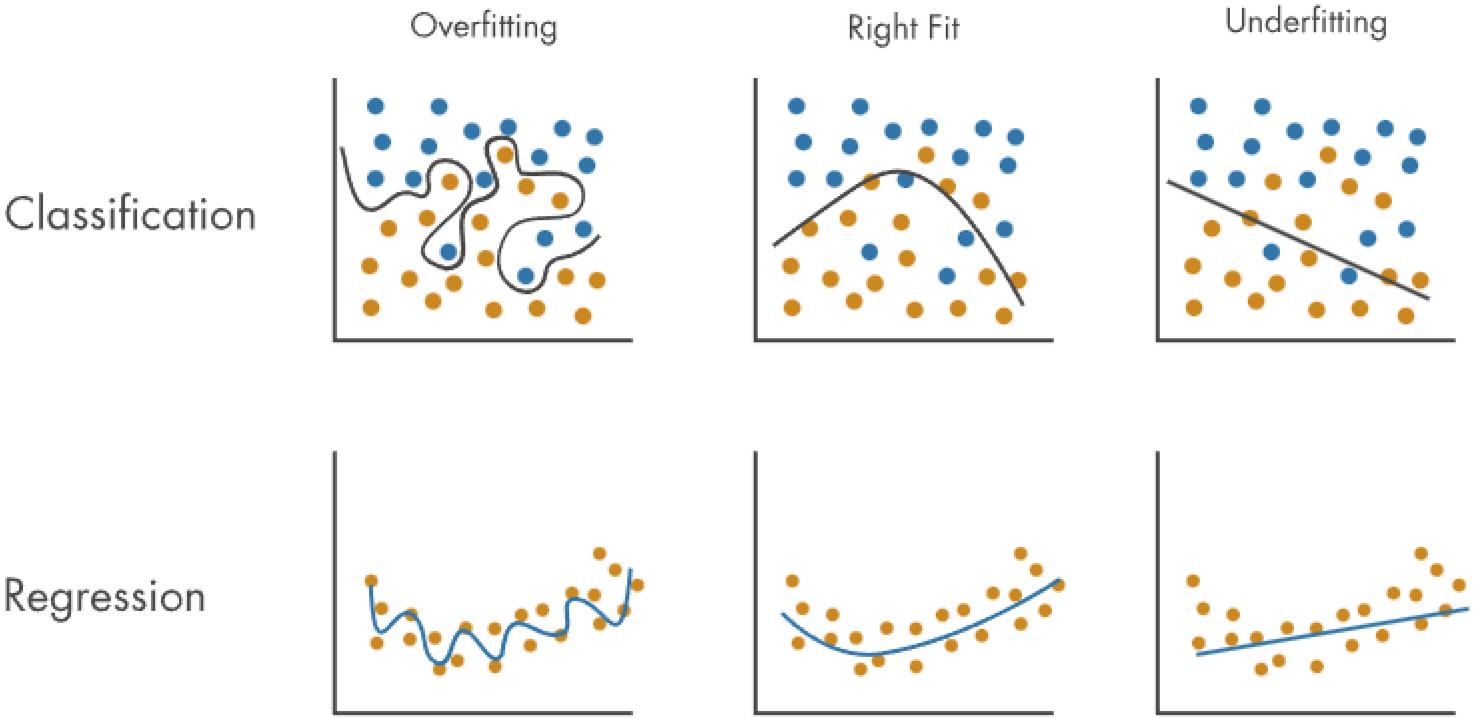
② Classification

Overfitting vs. Underfitting

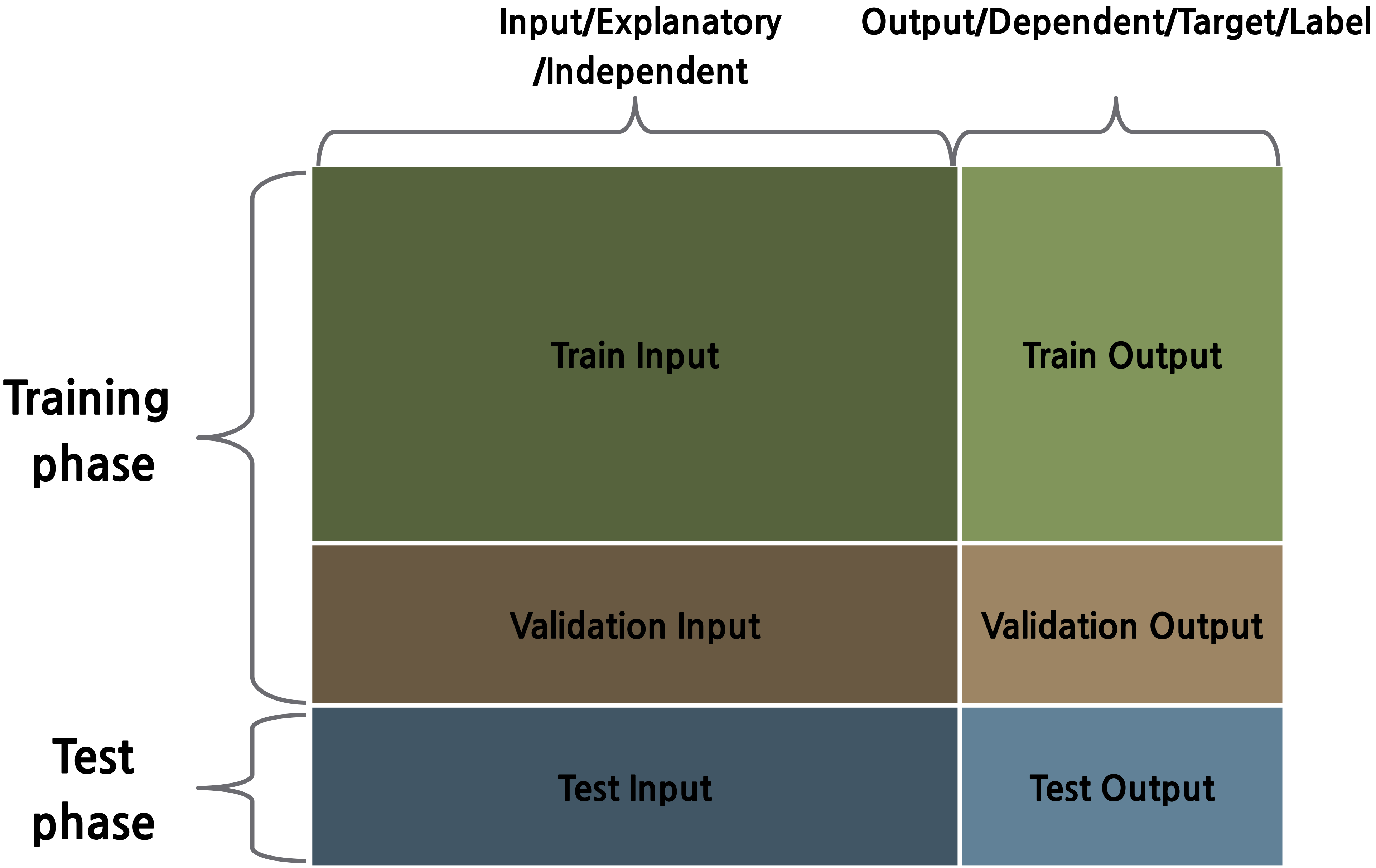
- Overfitting
  - Overfitting is a machine learning problem that occurs when a model is too closely aligned to training data, causing it to perform poorly on new data.
  - How it happens
    - The model is too complex
    - The training data is too small or contains irrelevant information
    - The model memorizes subtle patterns in the training data
  - Why it's a problem
    - An overfit model can't generalize well to new data
    - It can give inaccurate predictions
    - It can't perform well for all types of new data

- Underfitting
  - Underfitting occurs when a machine learning model is too simple to capture the underlying patterns in data.

Overfitting vs. Underfitting



Data Partition



Input/Explanatory /Independent

Output/Dependent/Target/Label

Training phase

Train Input

Train Output

Validation Input

Validation Output

Test phase

Test Input

Test Output

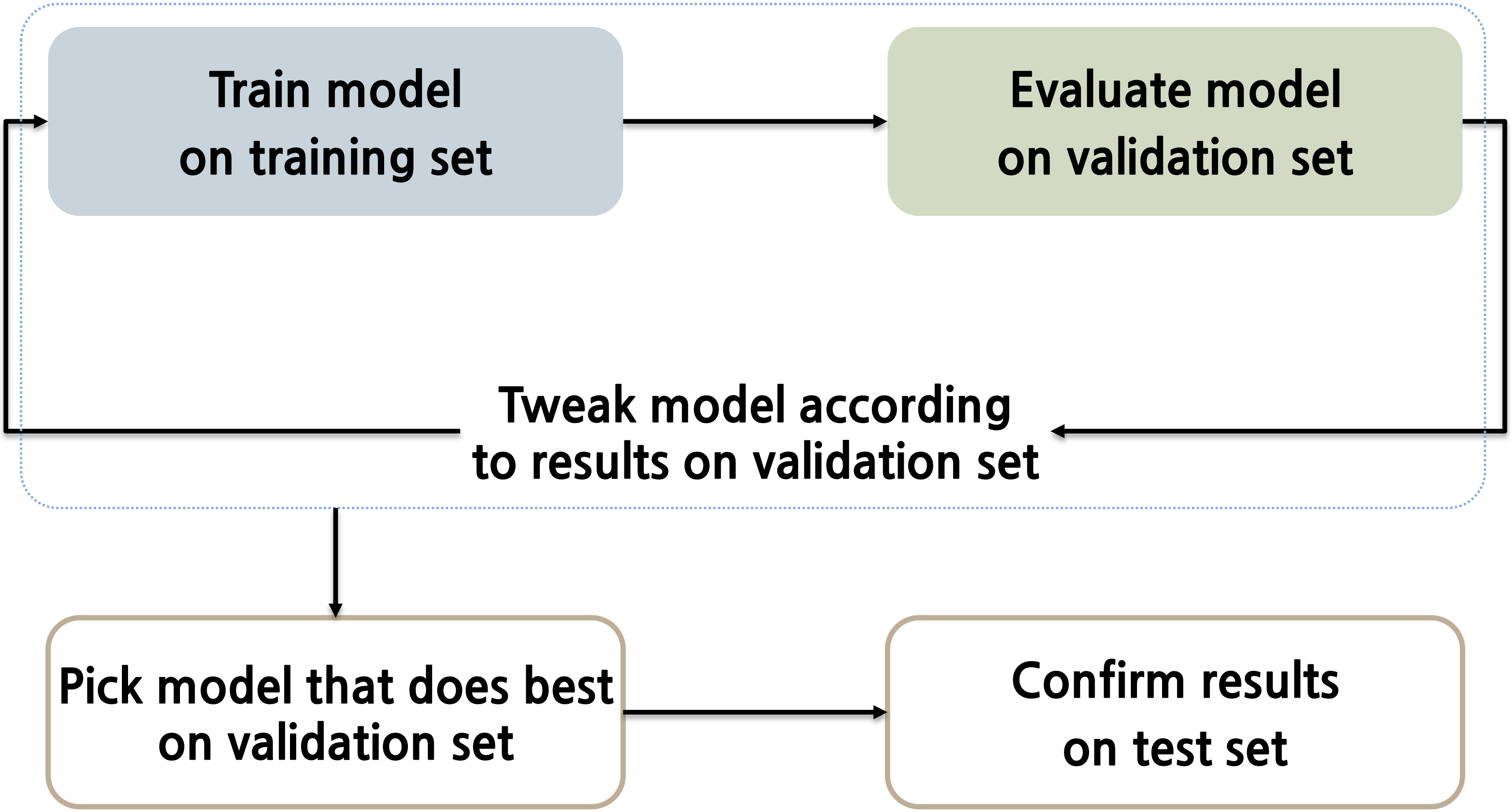
Data Partition

- Training set
  - Purpose: The training set is used to train the model. It contains the labeled examples the model will learn from. During the training phase, the model's parameters are adjusted based on the data in the training set.

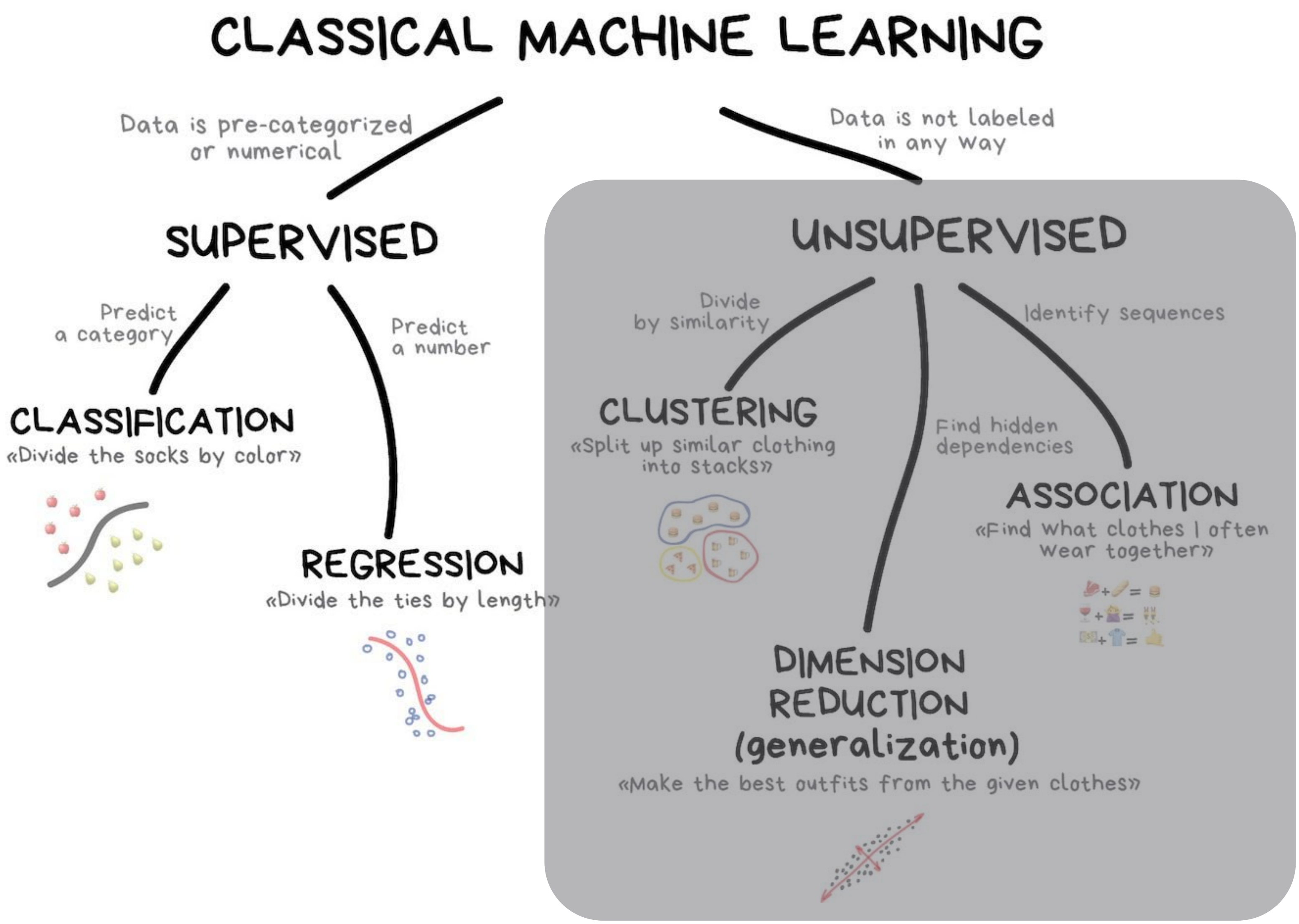
- Validation set
  - Purpose: The validation set is used to tune hyperparameters and evaluate the model during training. It helps in selecting the best version of the model.

- Test set
  - Purpose: The test set is used to evaluate the model's final performance after training and validation. This set simulates new, unseen data, giving an unbiased estimate of how the model will perform in a real-world scenario.

Process of Supervised Learning with Partitioned Data

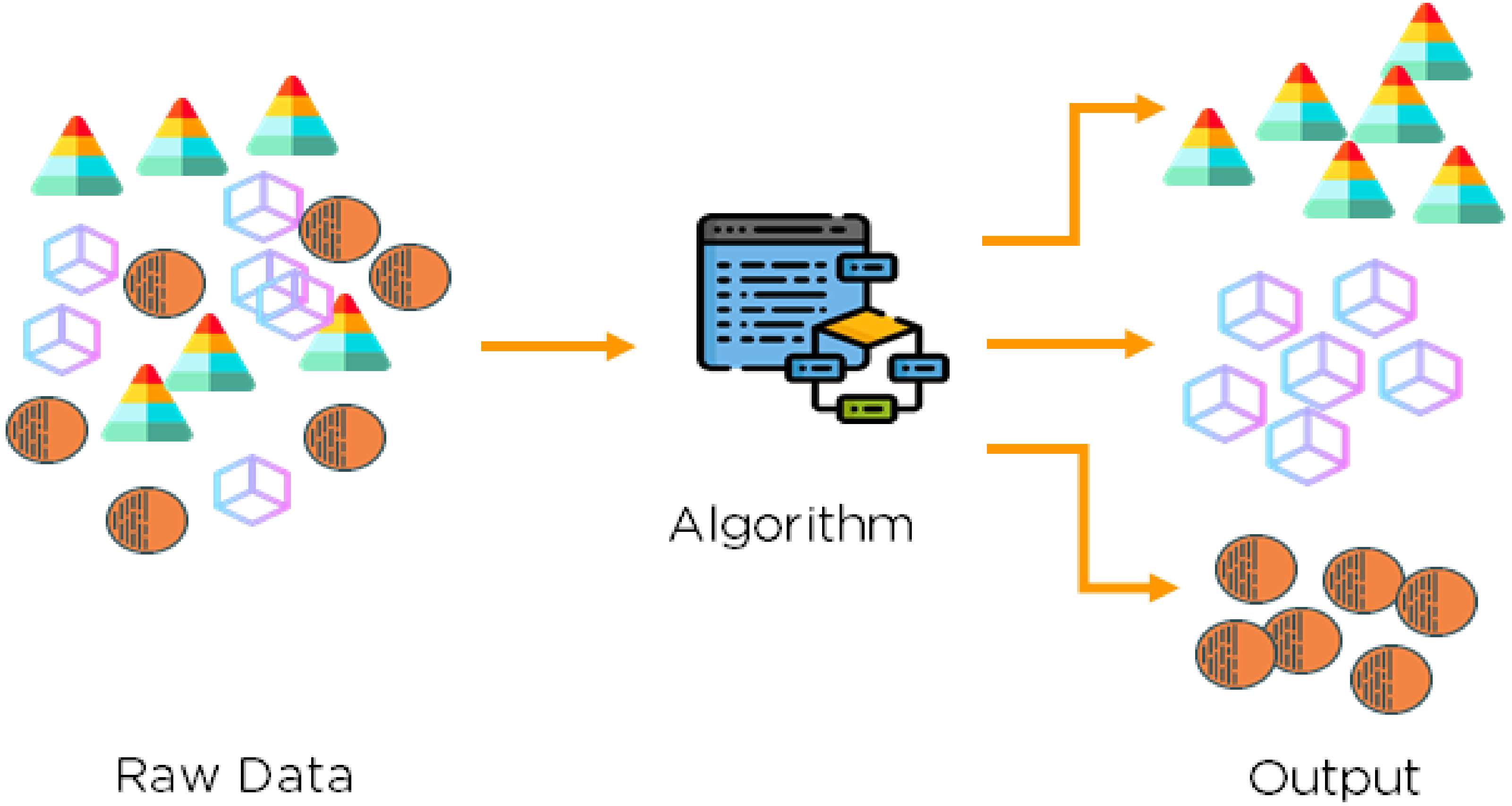


Topics Covered in This Class


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
Find hidden dependencies
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

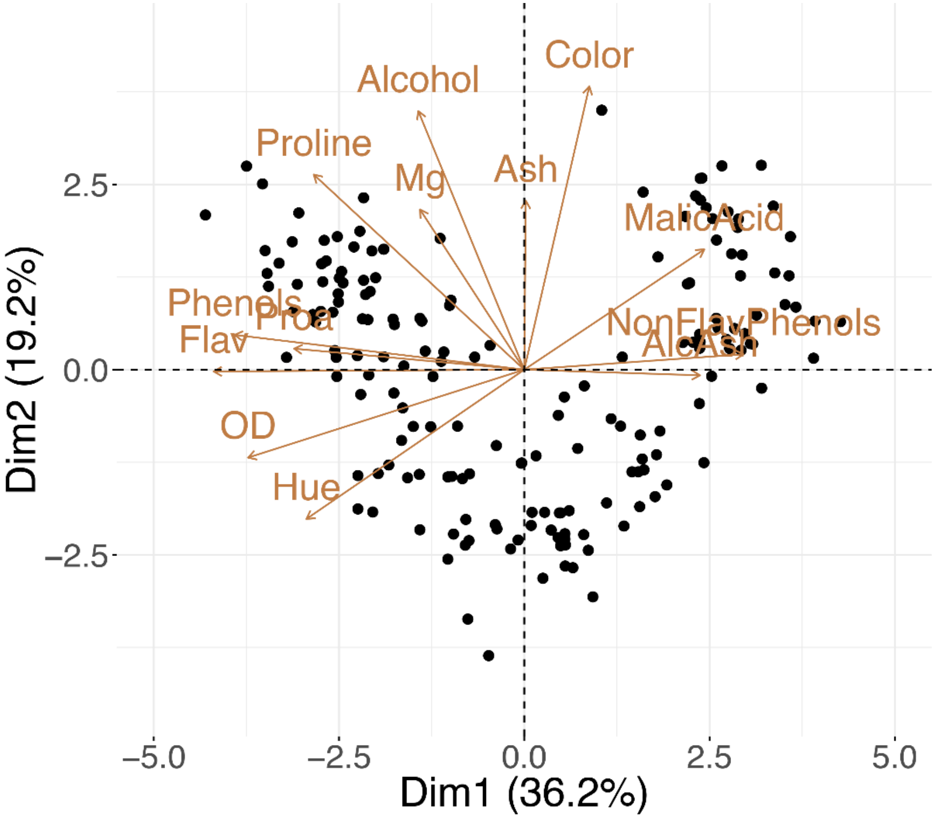
Unsupervised Learning: Clustering

- Grouping data points
  - How to determine which group does each data belongs to?



Unsupervised Learning: Dimensionality Reduction

- Dimensionality reduction
  - The process of reducing the number of random variables under consideration by obtaining a set of principal variables
  - High dimension → Low dimension



Unsupervised Learning: Association Rule Mining

- Find useful information from transactions

Datetime	Customer	Items
2015-07-15 14:03	1	orange juice, banana
2015-07-15 16:20	2	orange juice, milk
2015-07-16 10:14	3	detergent, banana, orange juice
2015-07-25 19:34	2	milk, bread, soda
2015-07-29 09:41	4	detergent, window cleaner
2015-08-01 20:55	1	bread, milk

- One of useful information is information like “If item A then item B”
  - This information is called association rule

- Find pair of items that are more likely to be purchased together based on transactions

Question

- Of the following examples, which would you address using an unsupervised learning algorithm? (Find all that apply.)

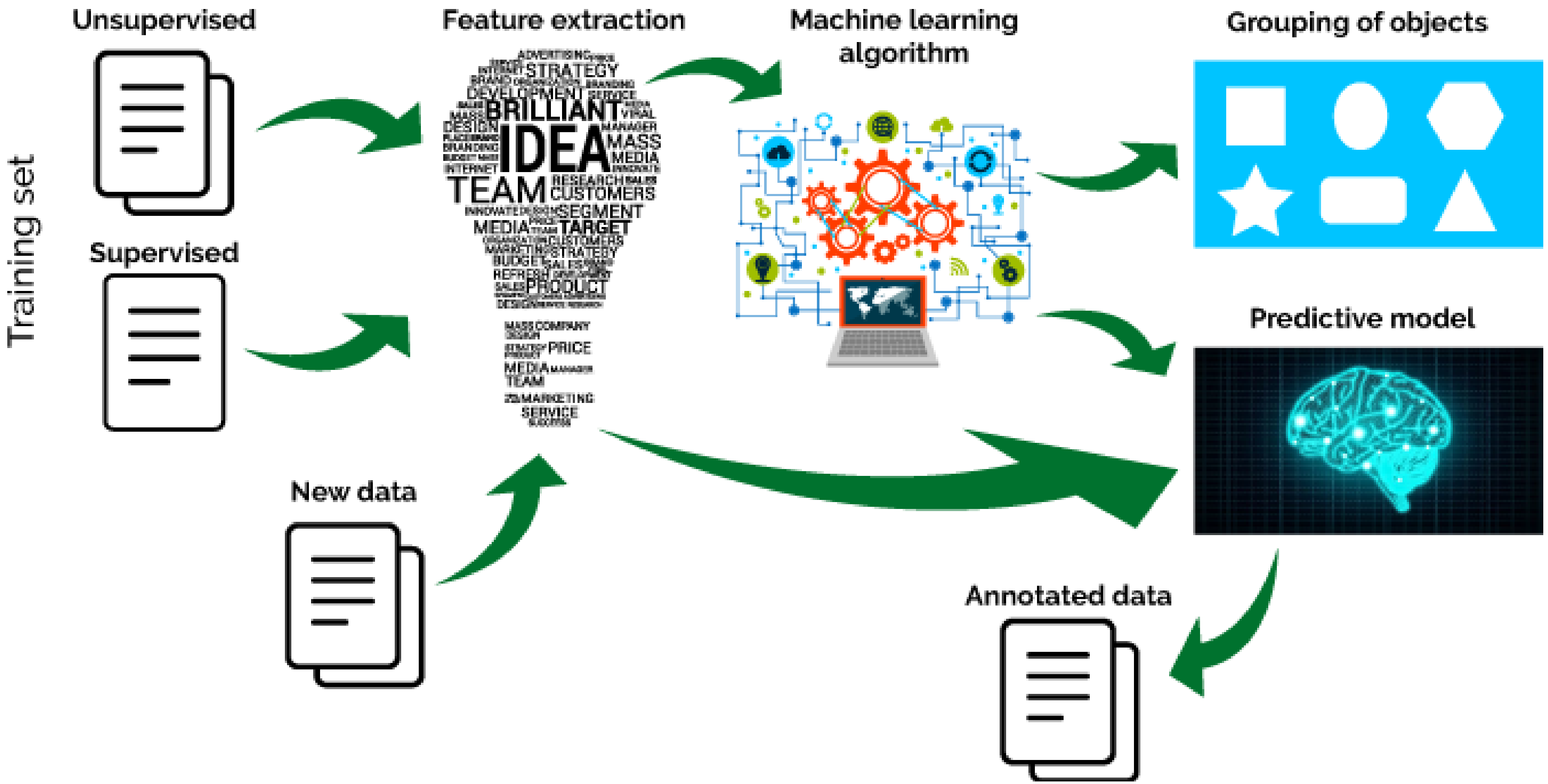
① Given email labeled as spam/not spam, learn a spam filter.

② Given a set of news articles found on the web, group them into set of articles about the same story.

③ Given a database of customer data, automatically discover market segments and group customers into different market segments.

④ Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

Overall Description



Training set

Unsupervised

Supervised

New data

Feature extraction

Machine learning algorithm

Grouping of objects

Predictive model

Annotated data