Machine Learning 1

Lecture 1 - Overview and Probability Theory

-Dr. Daniel Worrall-

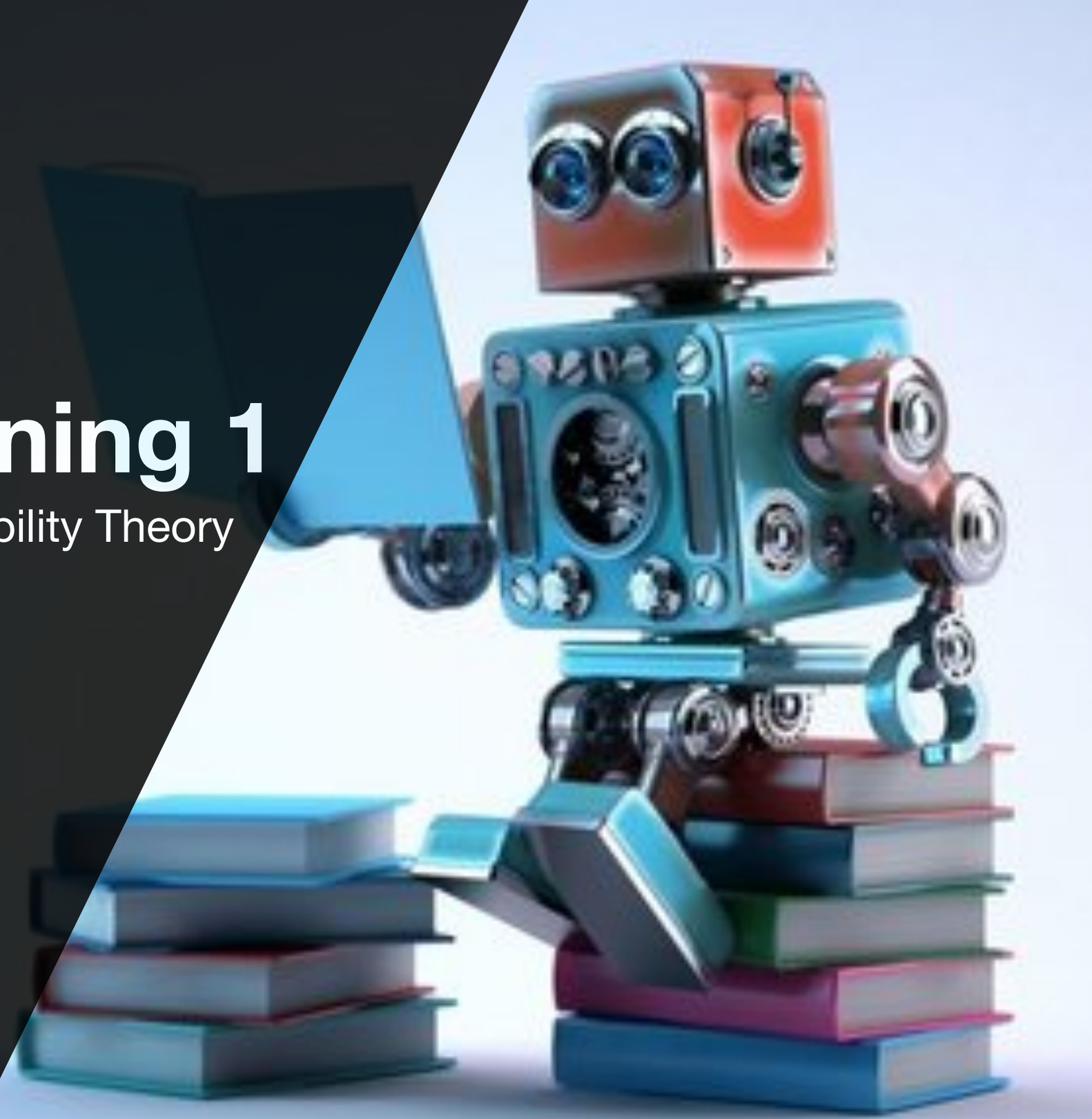


Image credit: Kirillm | Getty Images

Slides created by:
Rianne van den Berg

Admin: grading policy

- 20 % homework assignments
- 20 % programming assignments
- 60 % final exam
- Passing requires at least **50%** of the points in the final exam!!!
- No midterm exam
- Resit exam replaces 100% of the grade

Admin: lab assignments

- Laptop exercise classes (laptop colleges) on Wednesdays
- 3 assignments to hand in
- Assignments available on canvas
- must be handed in in groups of 2.
- Late submissions will not be graded.
- For detailed instructions, see canvas!

Admin: homework assignments

- Exercise classes on Thursdays
- During class: study practice exercises
- At home: make hand-in assignments.
- Must be handed in individually.
- Must be written **in Latex**. Handwritten papers or scans are not allowed.
- Solutions should contain derivations.
- You can help each other and discuss, but do not copy-paste!

Admin: lectures

- Tuesdays and Wednesdays 17-19
- Keep an eye on datanose for irregularities in the schedule
- Slides will be made available on Canvas

Admin: general

p.d.forre@uva.nl

- Please stick to your assigned group for laptop and exercise classes
- Questions: email your TA!
- If TA cannot answer your question: email me.
- See Canvas for email addresses

Prerequisite knowledge

- Probability Theory (+ basics of Statistics)

  Bayesian Statistics for Machine Learning

- Programming - Python

- Linear Algebra

- Calculus

- Vector Calculus (+ some Optimisation)

Prerequisite knowledge

Extra resources:

- Self-study course Mathematics for Artificial intelligence (see canvas)
- Book: Mathematics for Machine Learning (see canvas)

  Chapter 5 :)

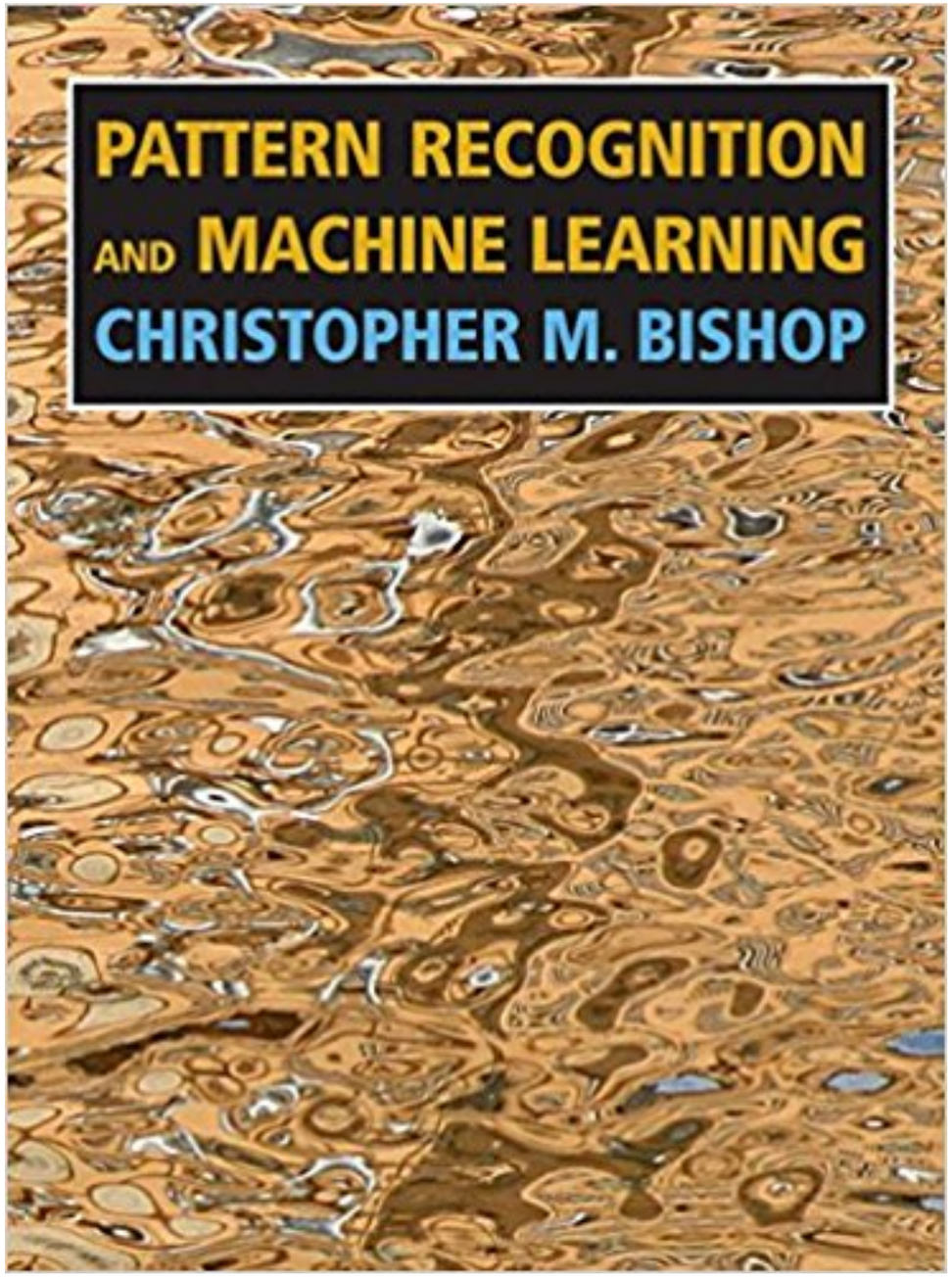
Literature: main book

Pattern Recognition and Machine Learning

- Christopher Bishop

Machine learning from a probabilistic point of view

Free online

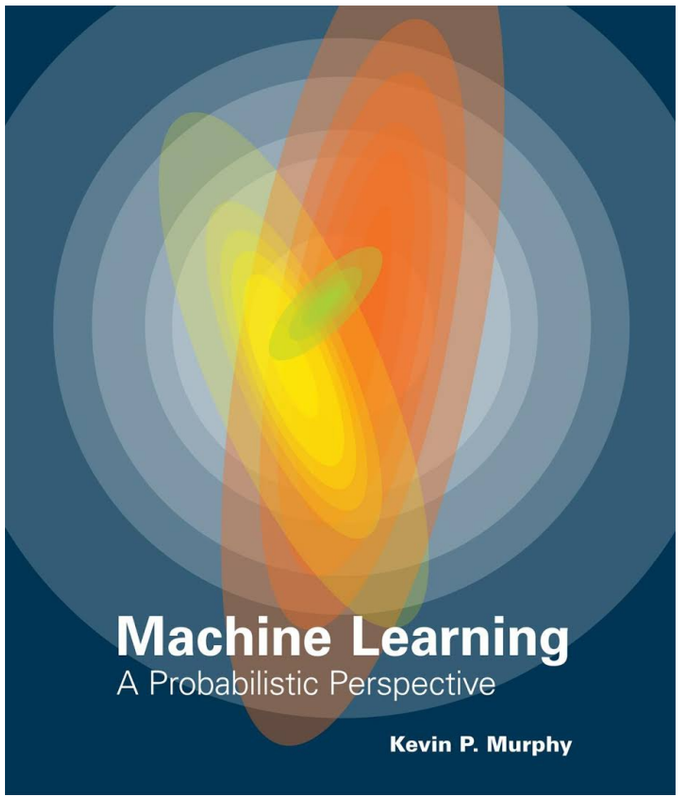


Literature: other books

Machine Learning - A Probabilistic Perspective

- Kevin P. Murphy

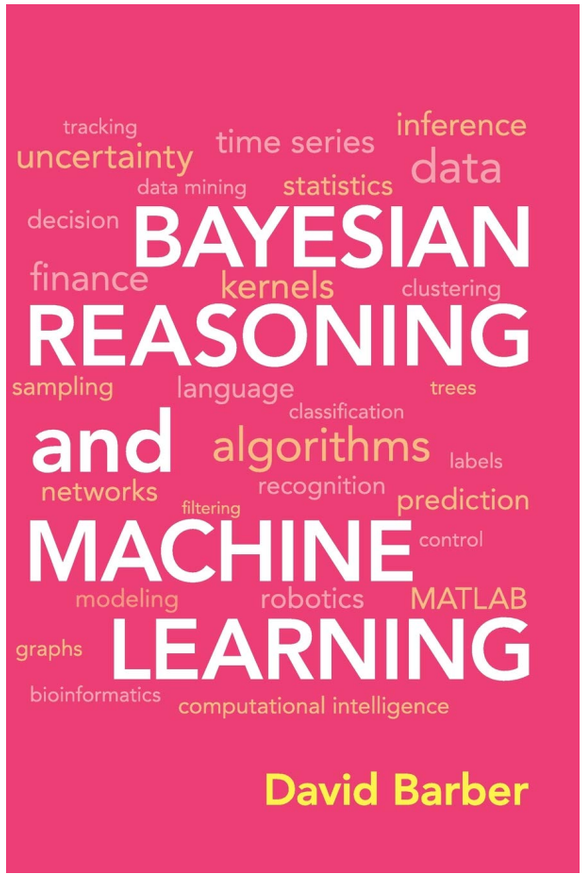
Best alternative to Bishop, maybe better



Bayesian Reasoning and Machine Learning

- David Barber

Machine learning viewed from Bayesian statistics

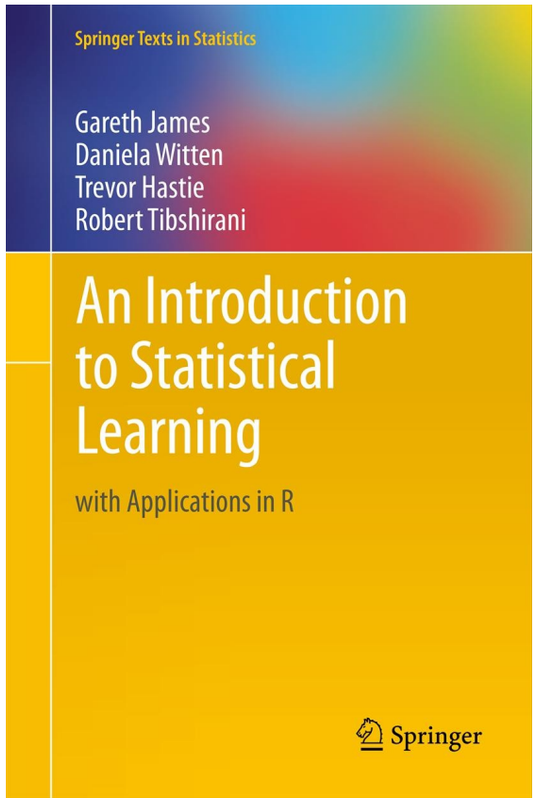


Literature: other books

An Introduction to Statistical Learning

- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani,

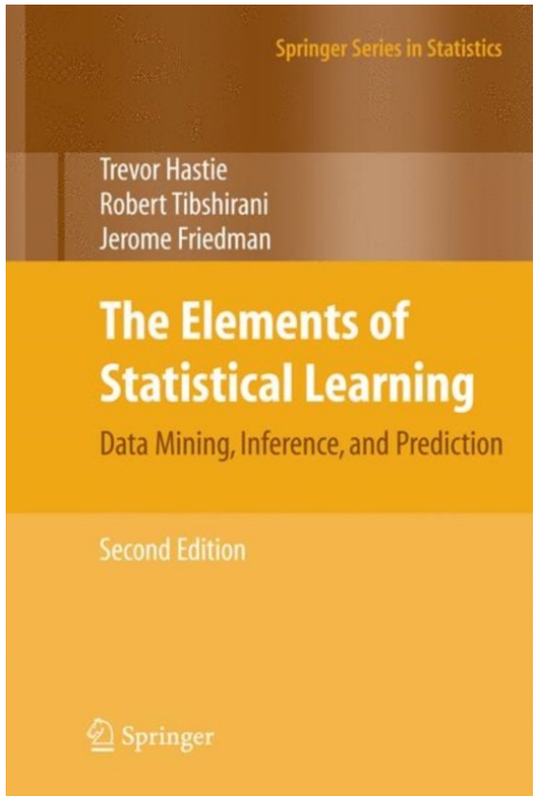
Introduction to Machine learning as a statistical tool.



The Elements of Statistical Learning

- Trevor Hastie, Robert Tibshirani, Jerome Friedman

More advanced view of Machine learning as a statistical tool.

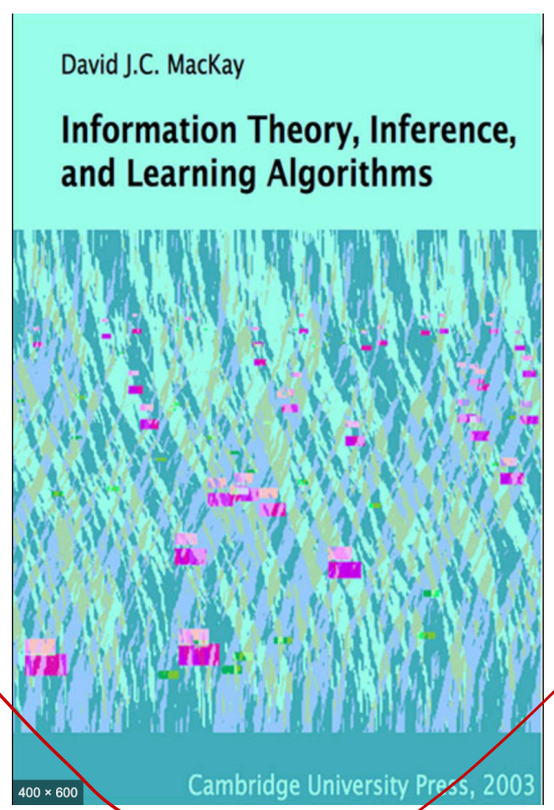


Literature: Daniel's addition

Information Theory, Inference, and Learning Algorithms

- David MacKay

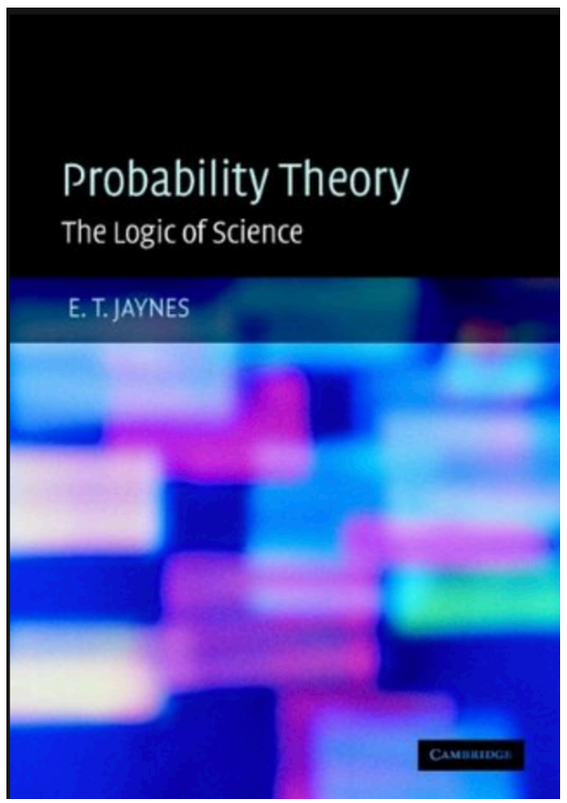
https://www.inference.org.uk/itprnn/book.pdf



Probability Theory: The Logic of Science

- E. T. Jaynes

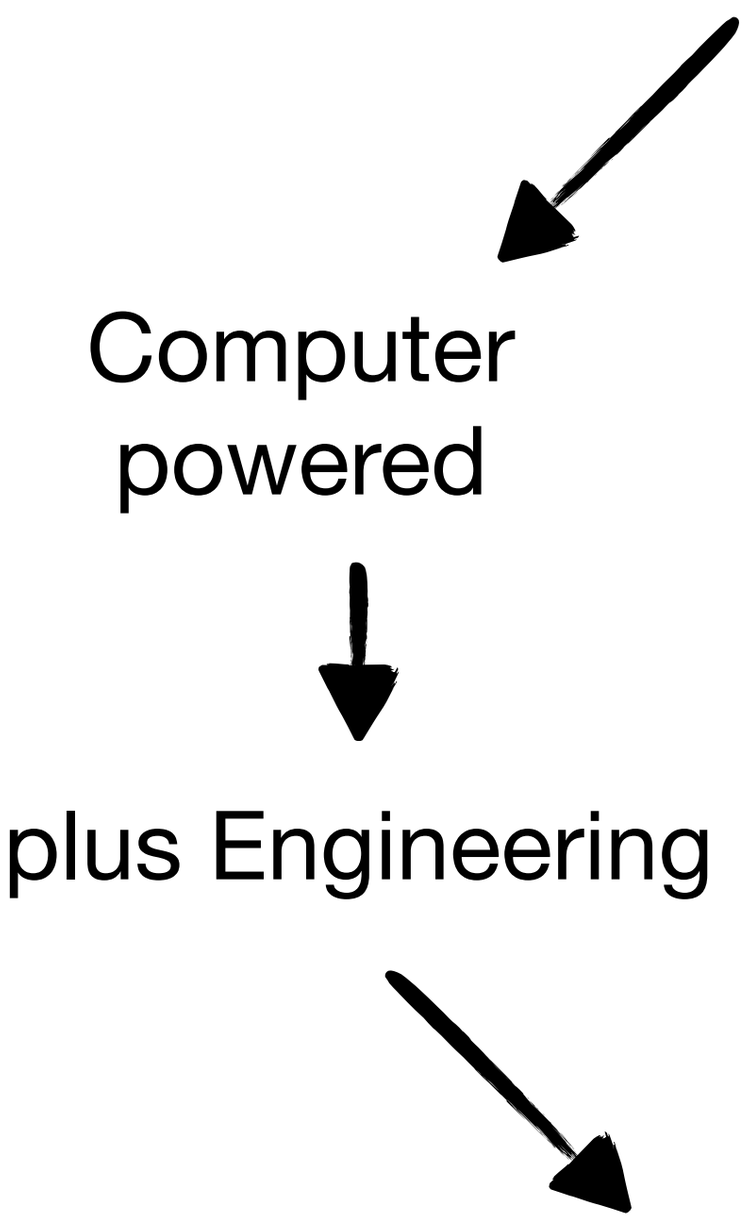
A pure Bayesian treatment of probability theory

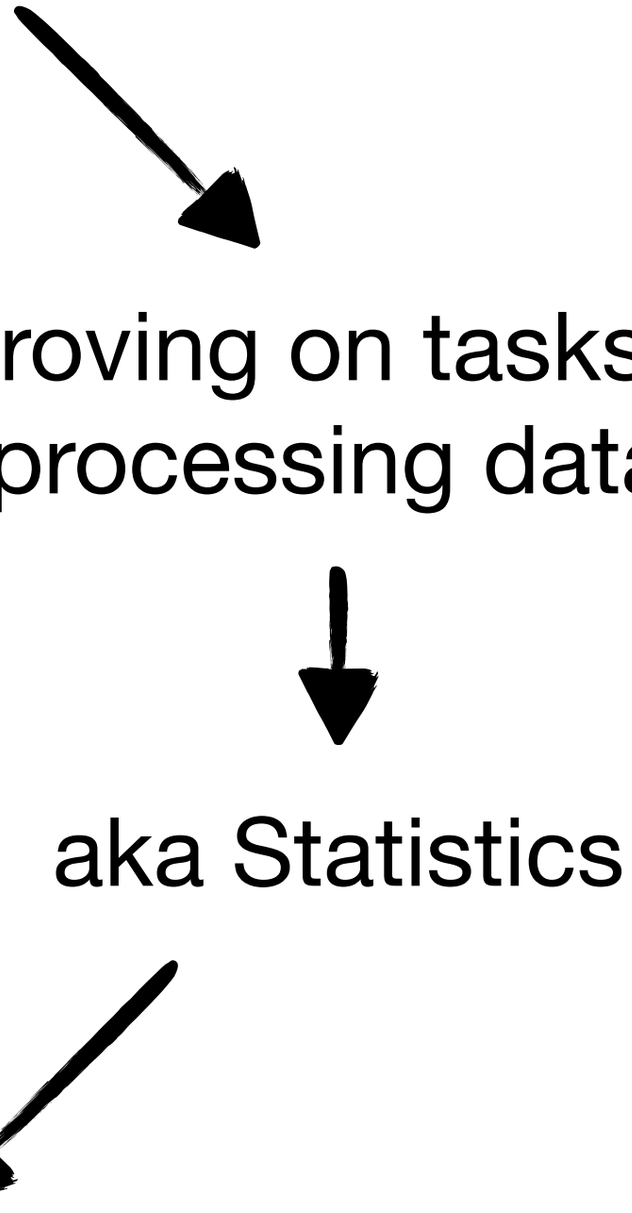


Questions?

Overview

What is Machine Learning?





Data driven AI

(in contrast to symbolic AI)

What is machine learning?

data

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

metric / score

- Tom M. Mitchell

Classification: Postcodes

28 × 28 pixels $\in \{0, \ldots, 255\}$

label / aleatoric noise

 = "2"

 = "8"

 = "0"

 = "9" or "8"

 = "8" or "9"

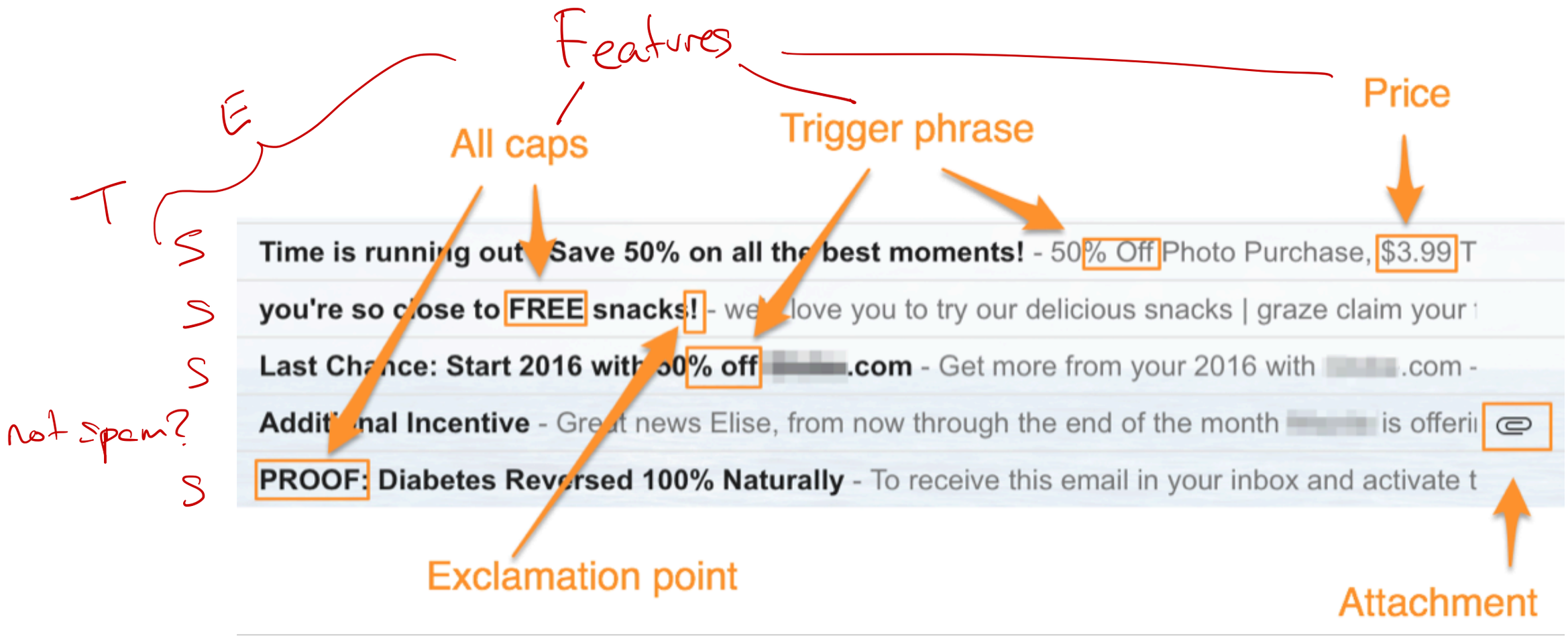
$$\text{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \mathbb{1}[y_i = \hat{y}_i]$$

ground-truth y, outputs / predictions $\hat{y}$

indicator fnc $= \begin{cases} 1 & \text{if } y_i = \hat{y}_i \\ 0 & \text{otherwise} \end{cases}$

MNIST dataset

Classification: Spam filtering



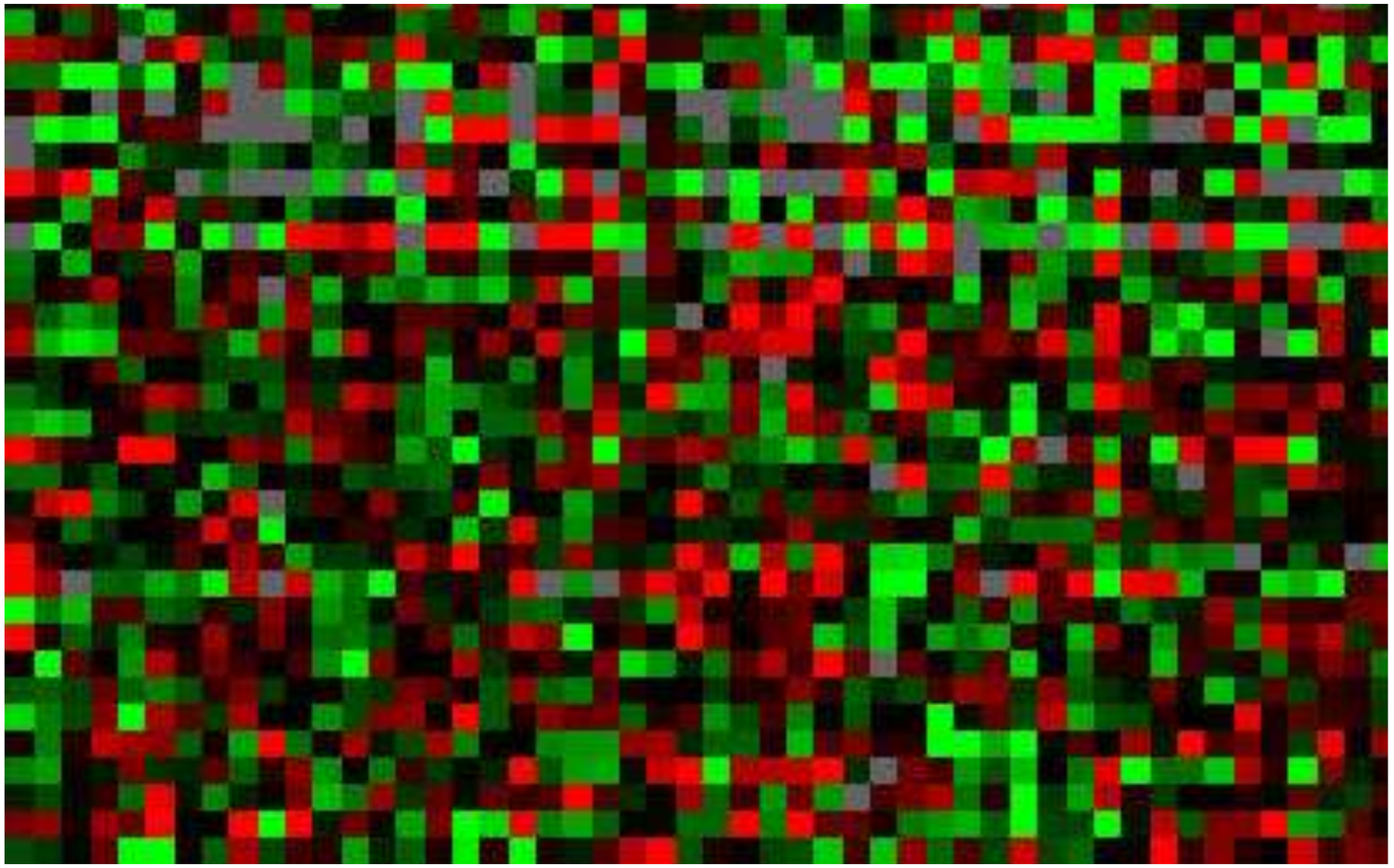
$$\text{accuracy}(y, \hat{y}) =$$ like on previous slide

Examples of spam emails. [source: Yesware]

Clustering: Genomics-Tumor data

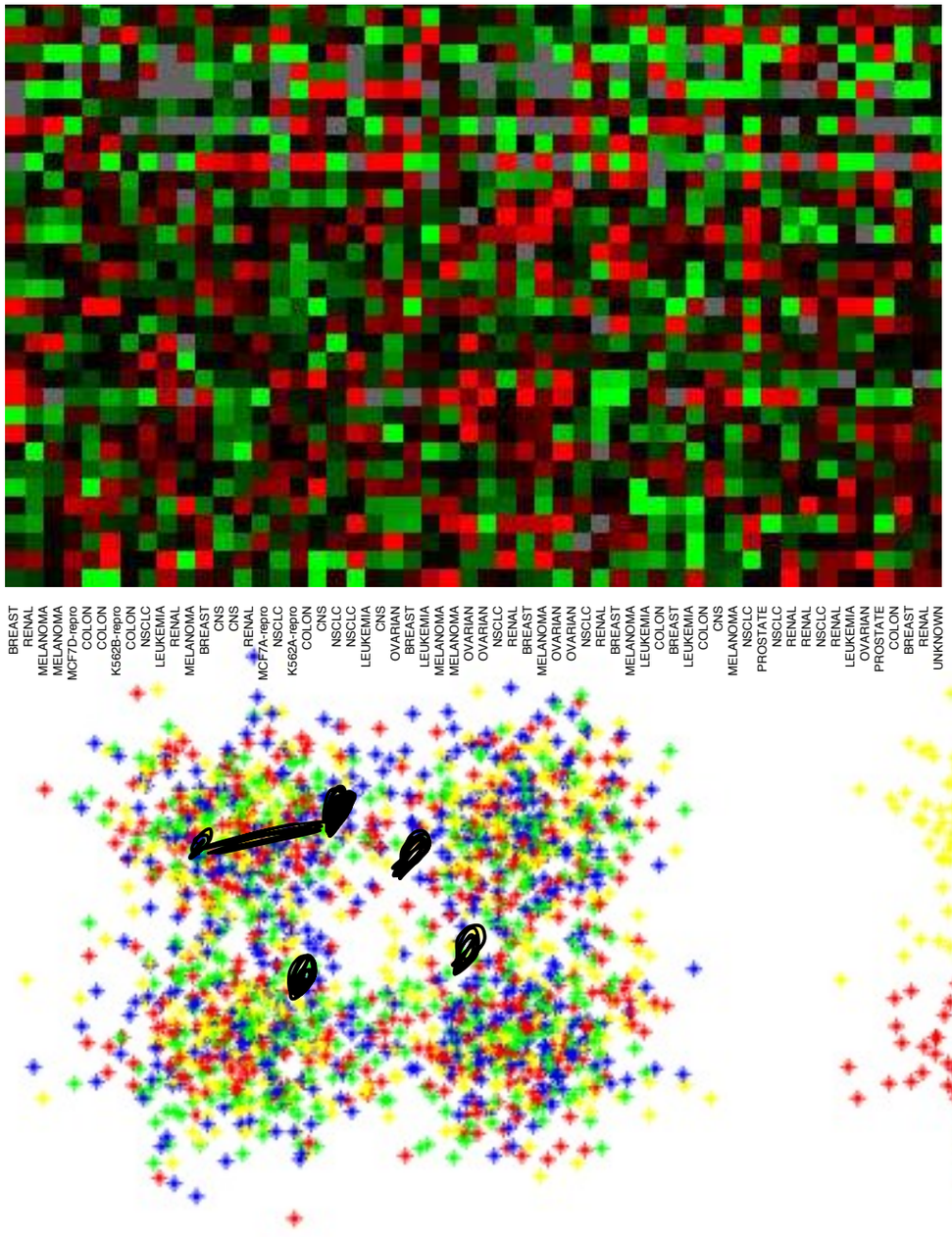
Eg

negative: underexpressed positive; overexpressed missing values

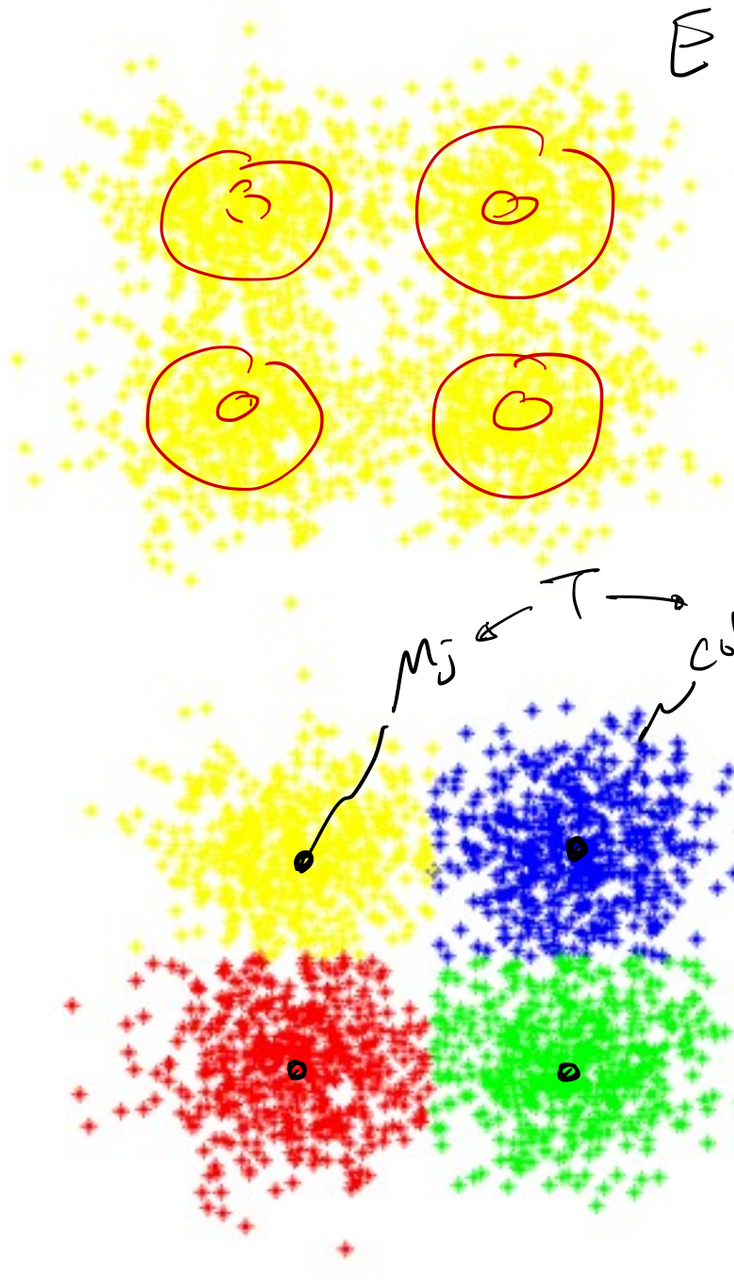


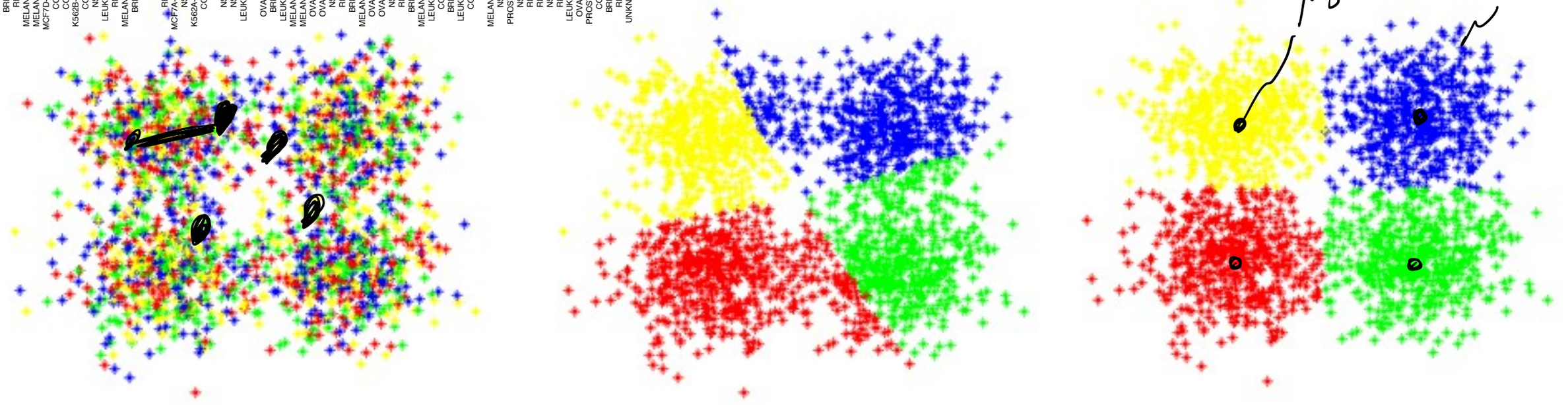
Expression matrix of genes (rows) for 64 human tumor samples (columns). [source: ESL 1.3]

Clustering: Genomics-Tumor data



feature transform





$$\text{within cluster sum of squares} = \sum_{i=1}^{n_{\text{samples}}} \min_{\mu_j \in C} \| x_i - \mu_j \|^2$$

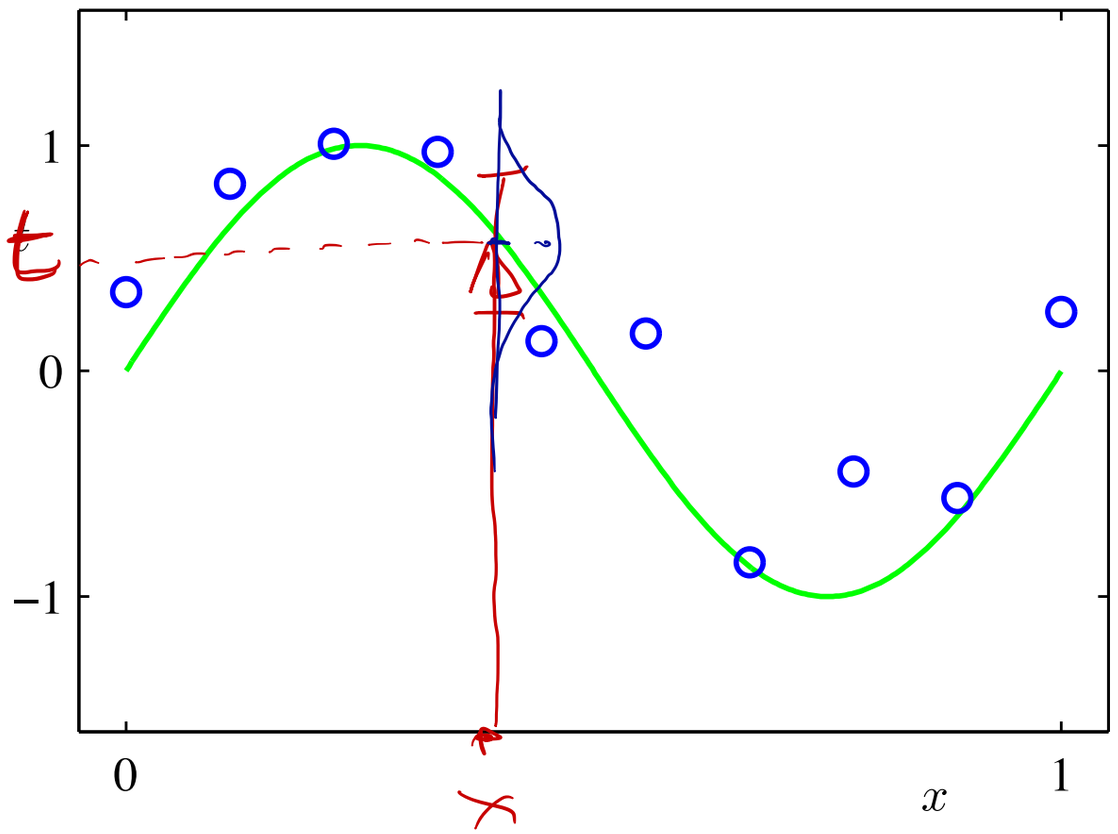
μ_j: cluster means

1D Regression = curve fitting

input : x

target : $t = \sin(2\pi x) + \varepsilon$

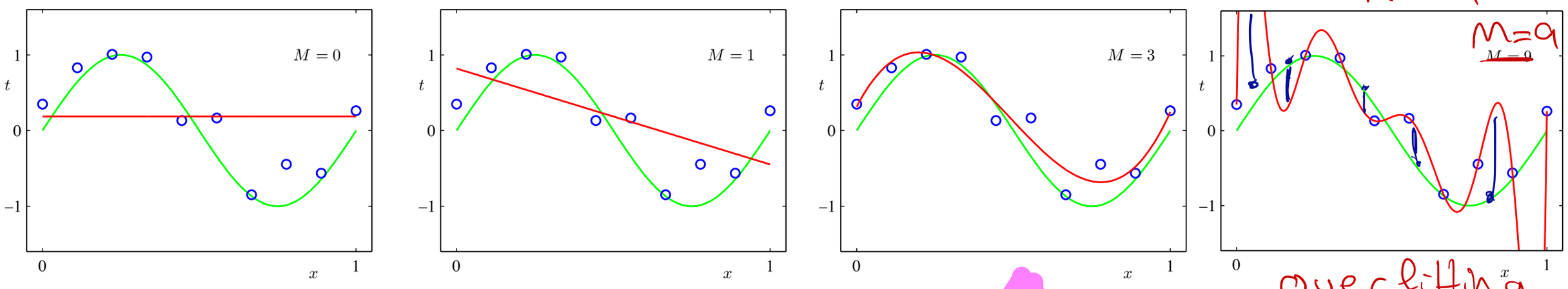
noise : $\varepsilon \sim \mathcal{N}(\varepsilon; 0, 1^2)$



1D Regression

$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n$

$n = 9$



overfitting

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2$$

Best performance on training set

Best performance on new datapoints

Q: On which datapoints should performance be measured?

Generalization: performance should be measured on new data (test data)

Polynomials of order *M* (red) fit to data constructed as t = sin(2πx) + ε (green)

Overview

1. **Overview**

    1. What is machine learning?

    2. **Different types of machine learning**

2. Recap: probability theory

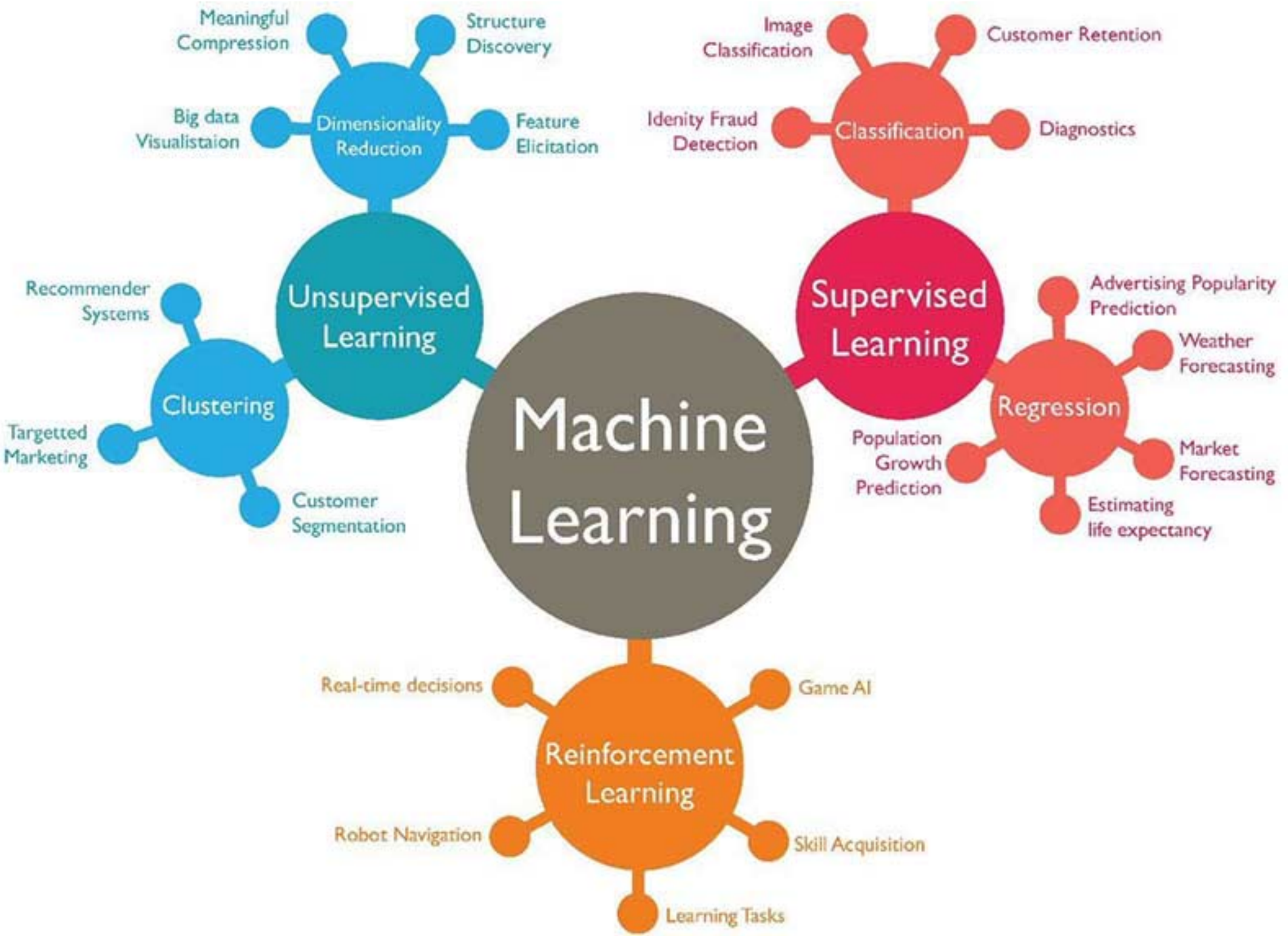


Image source : www.techleer.com

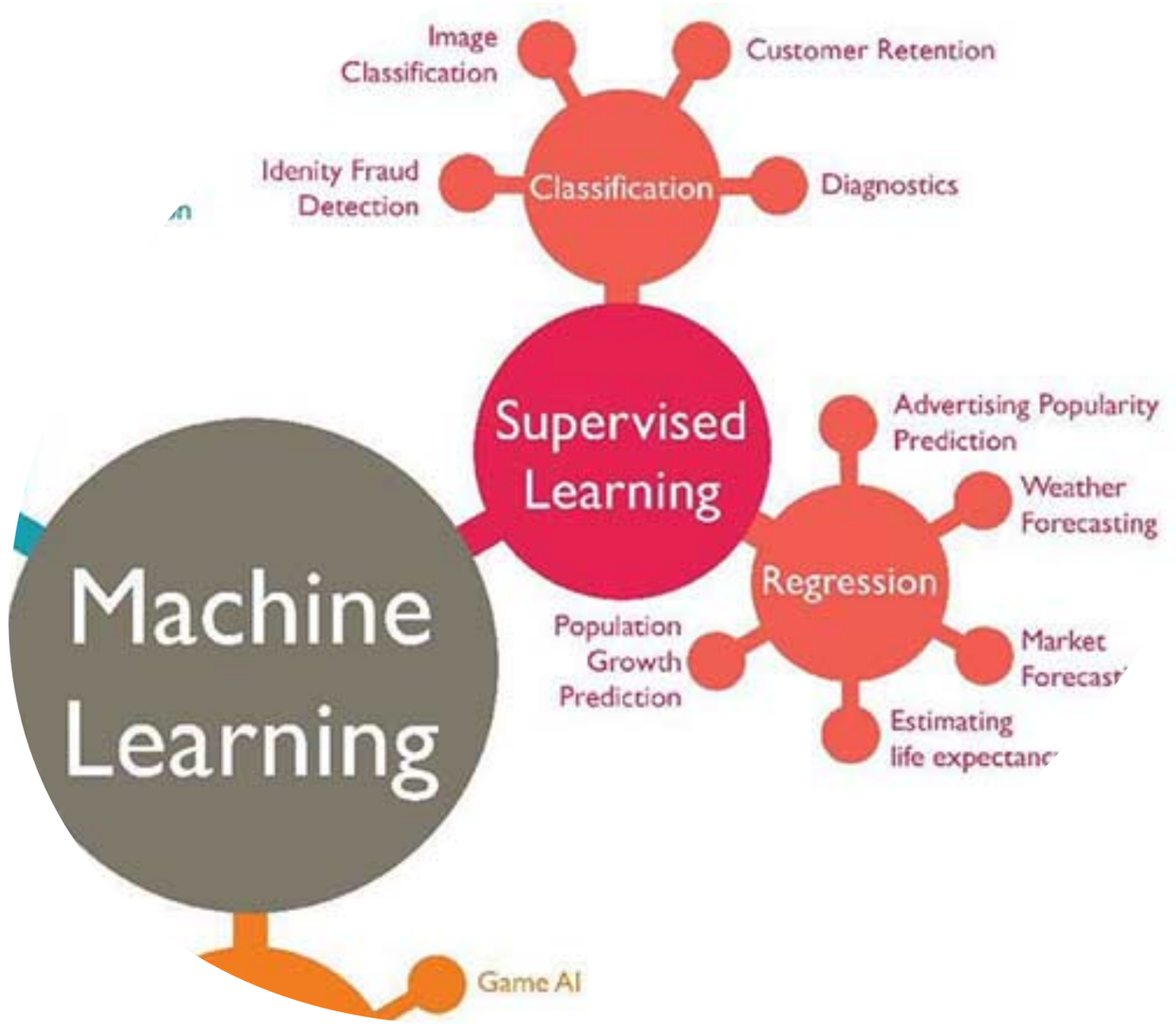
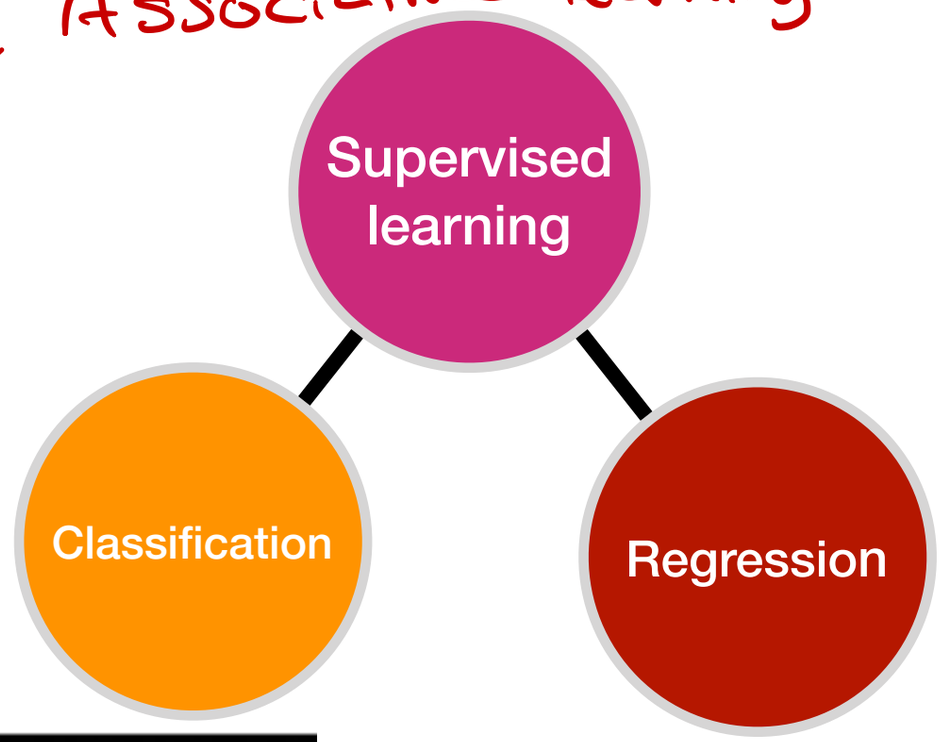


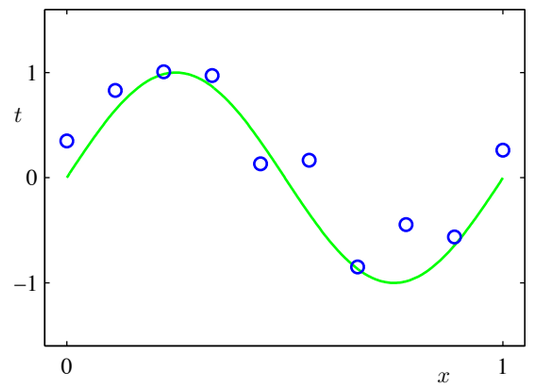
Image source : www.techleer.com

Supervised learning

Associative learning



Dataset



features: $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

x = 

x = 0.25

targets: $\{t_1, \ldots, t_N\}$

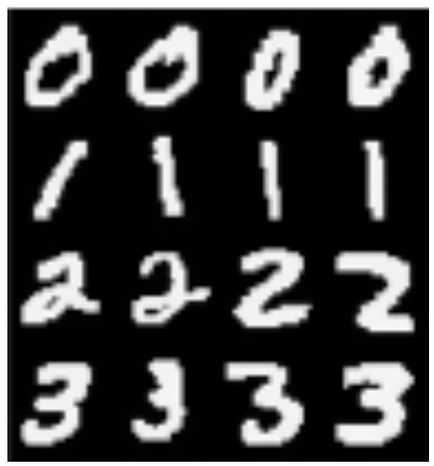
t = 2

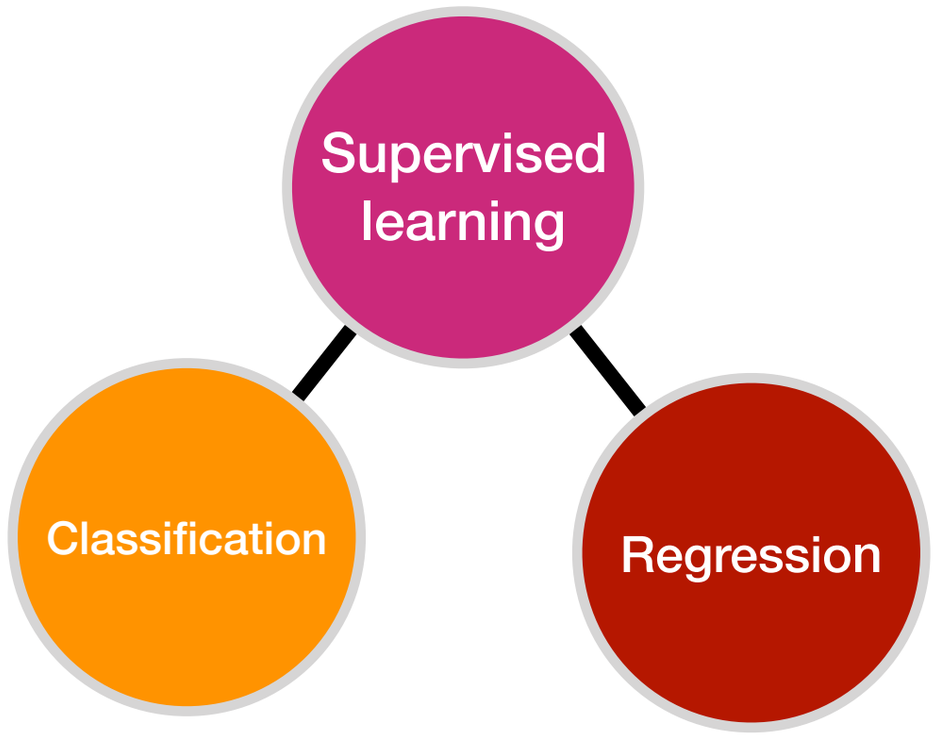
t = 0.707

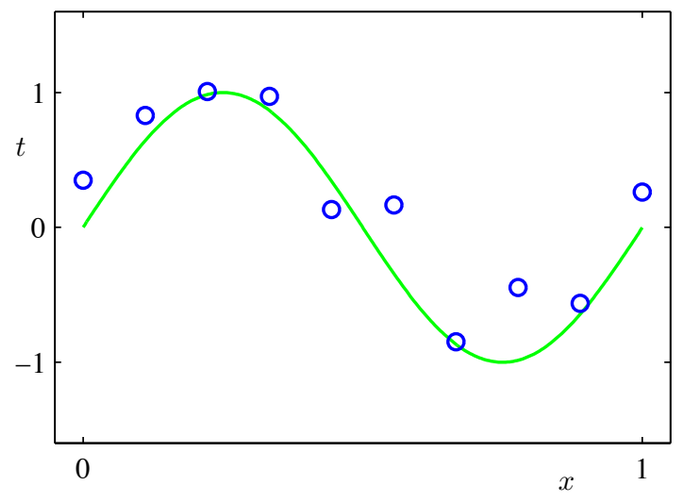
discrete

continuous

Supervised learning





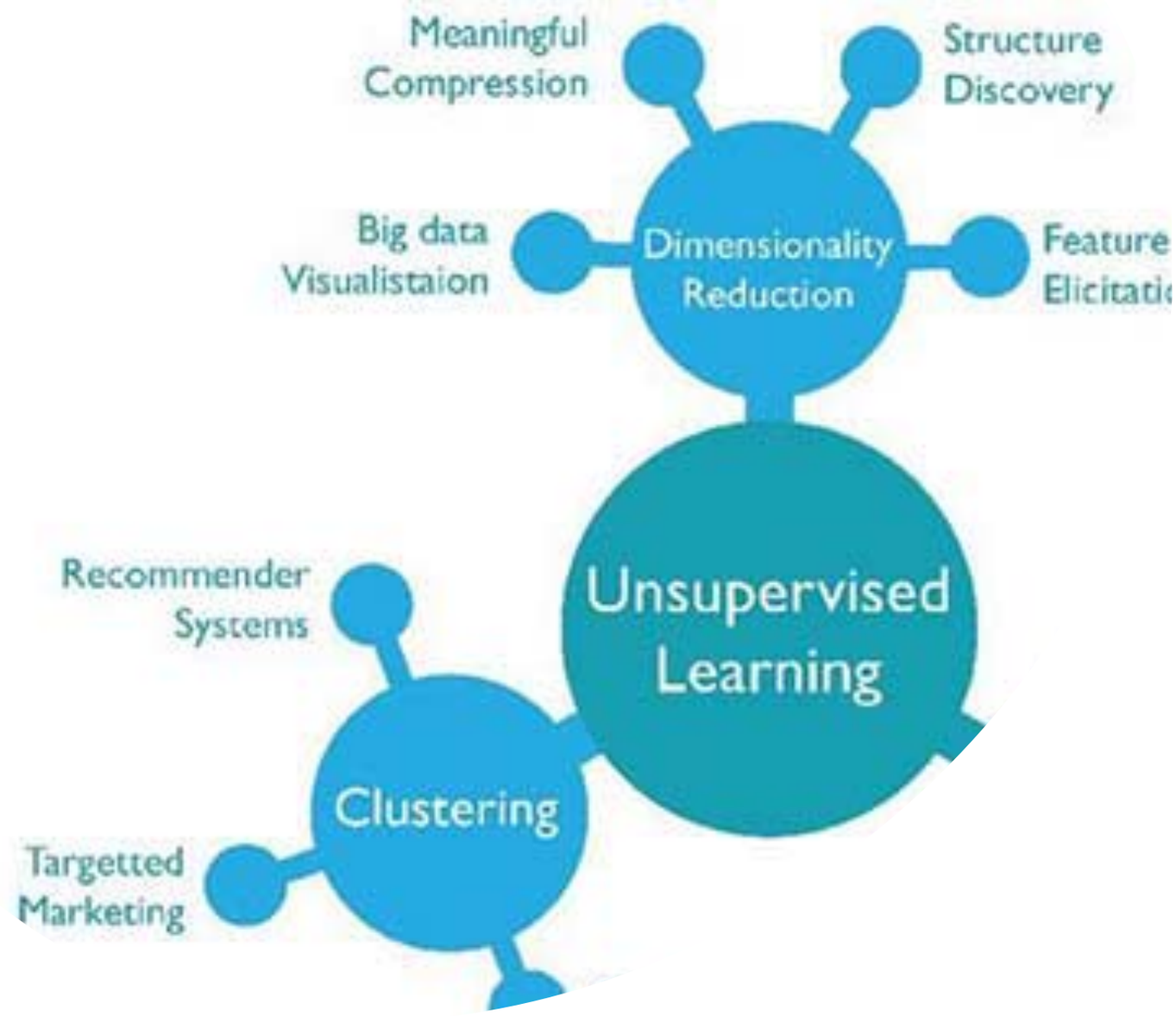


Task: Find function f such that f(**x**) ≈ t for all known and unknown (**x**, t)

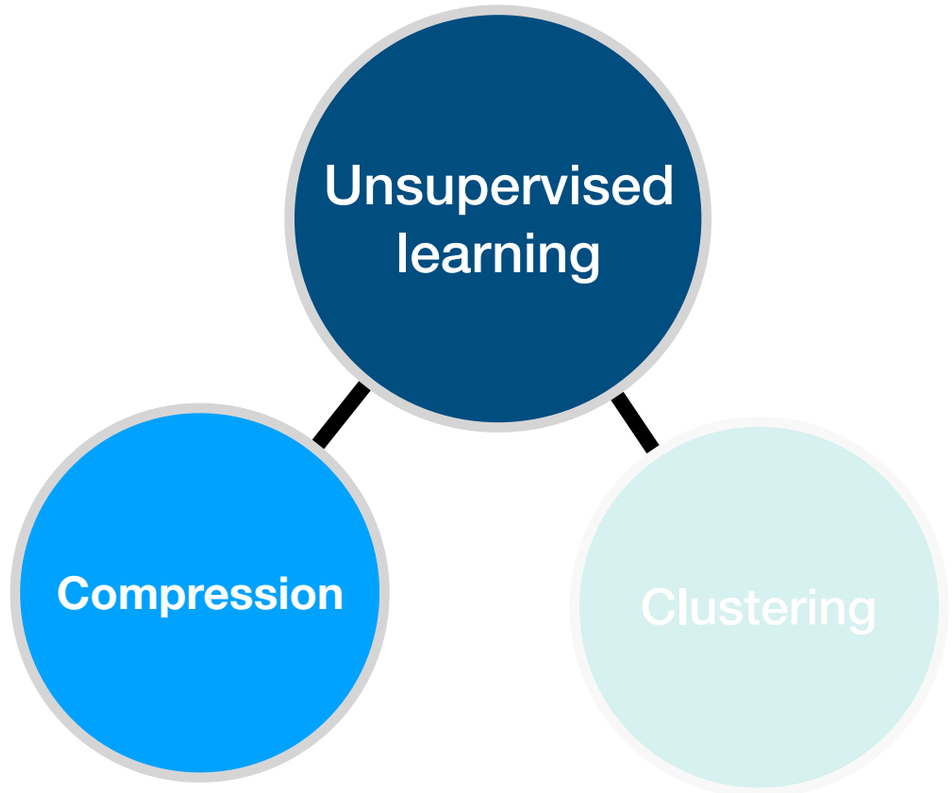
training data

validation / test data

training pair


Meaningful
Compression
Structure
Discovery
Big data
Visualistaion
Dimensionality
Reduction
Feature
Elicitatio
Unsupervised
Learning
Recommender
Systems
Clustering
Targetted
Marketing

Unsupervised learning



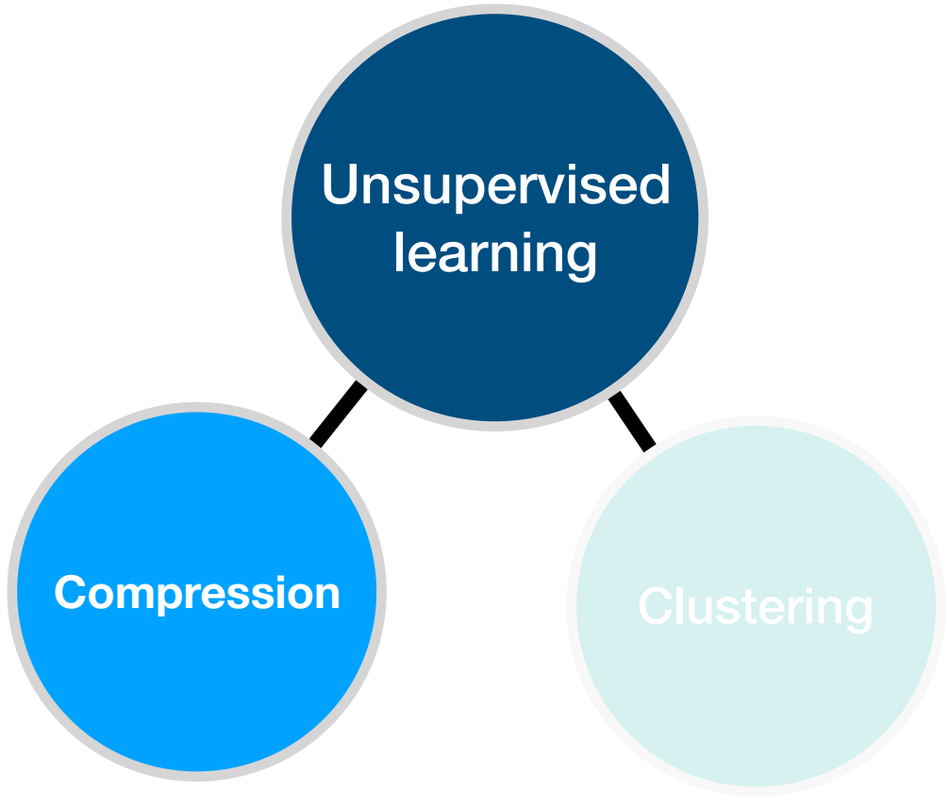
100 x 100 pixel

Dataset: 

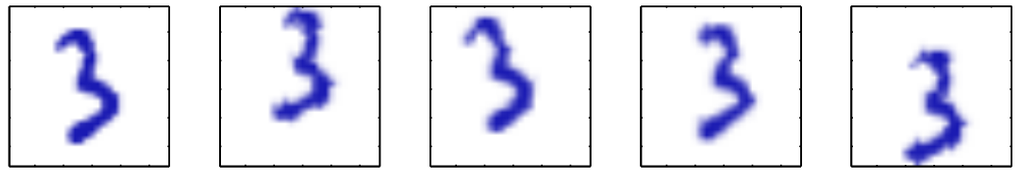
Task: compress images

why ? reduces required storage

Unsupervised learning

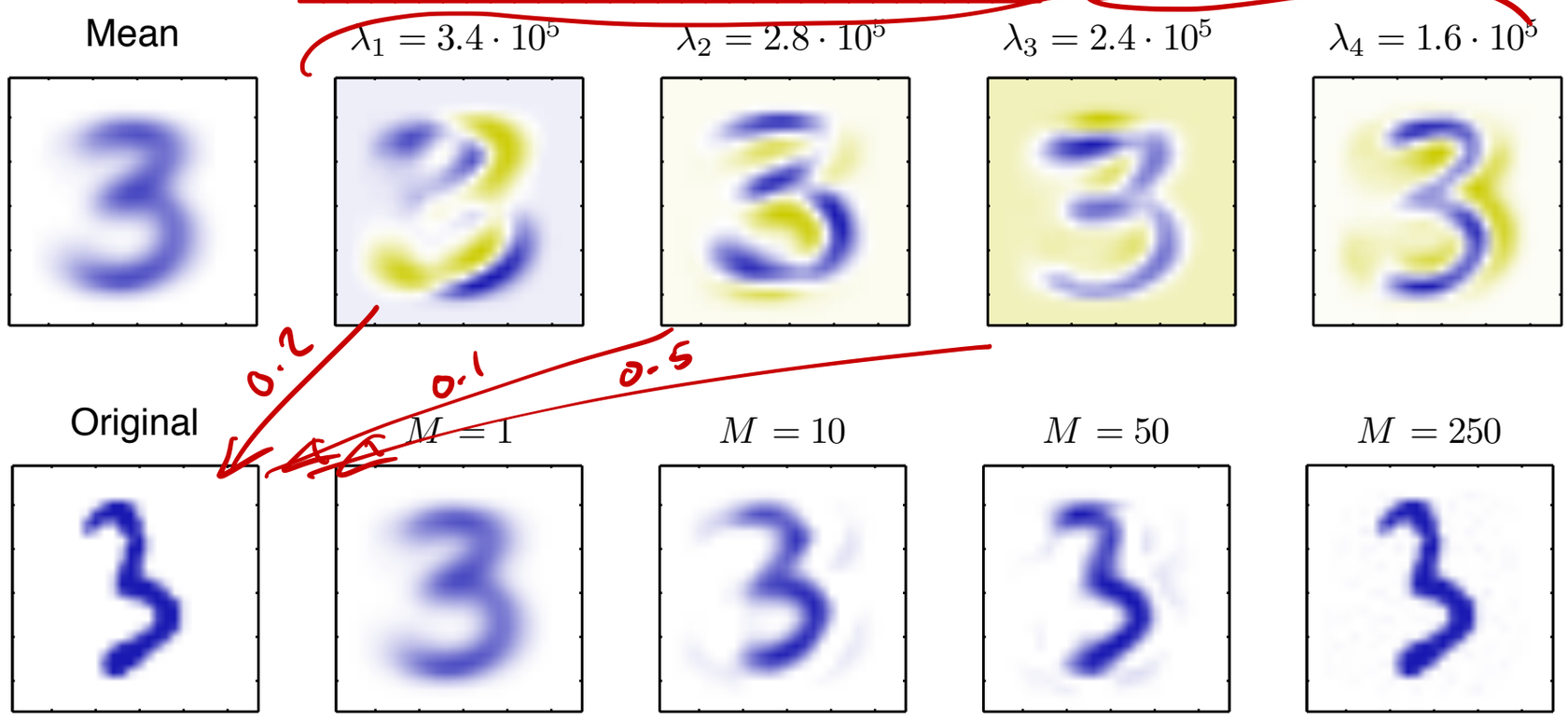


Dataset:



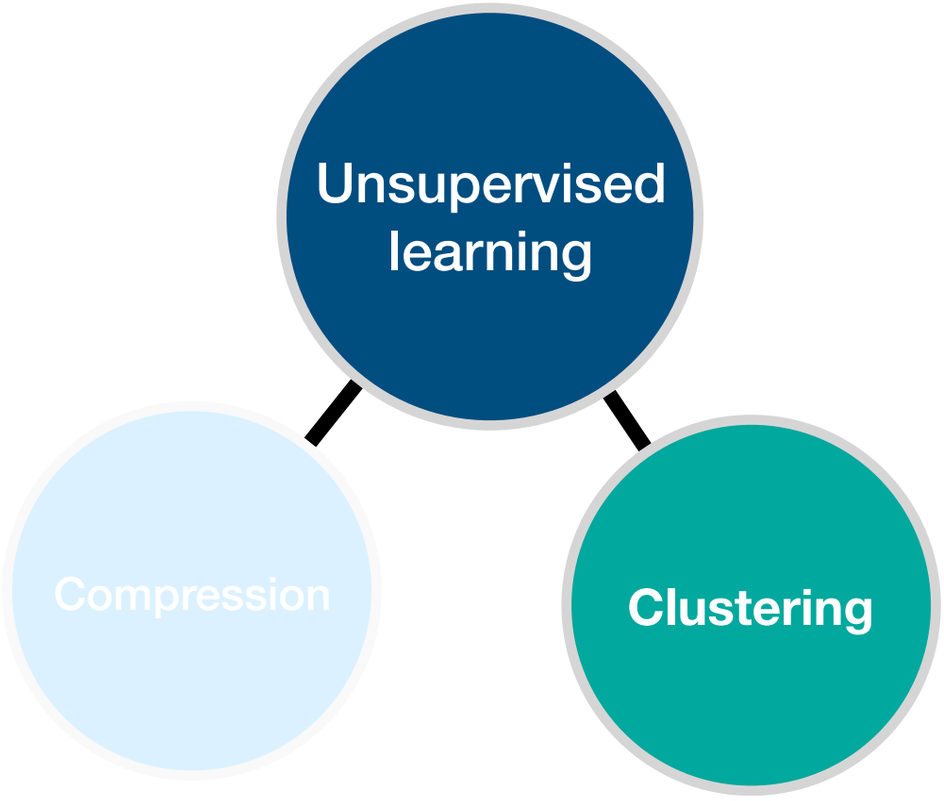
Task: Compress image

Method: expand along principle components (PCA)

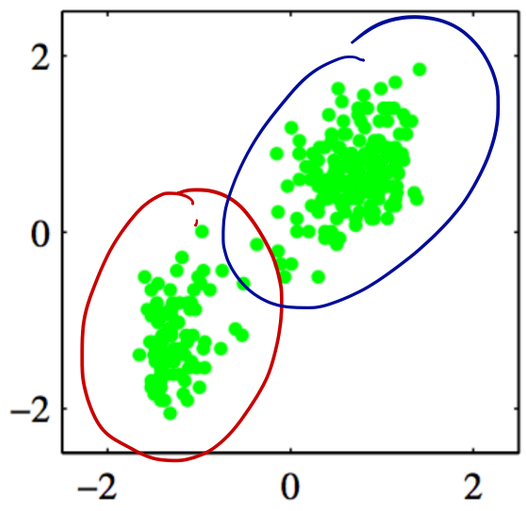


Result:

Unsupervised learning

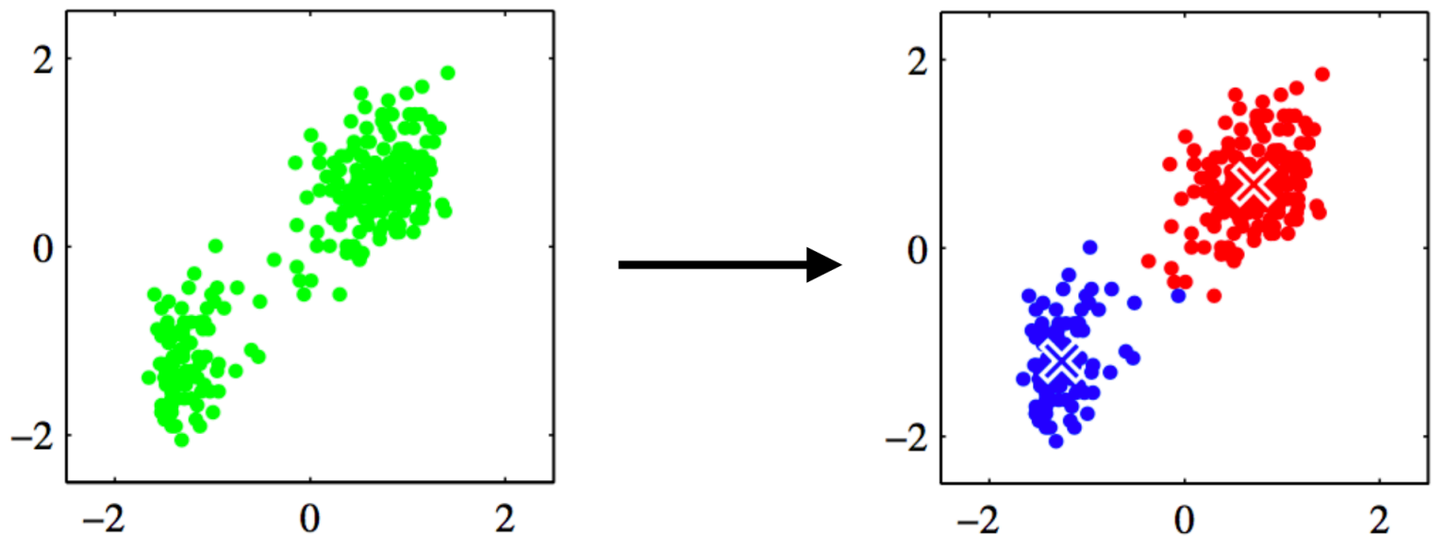


Dataset:



Task: Assign every datapoint to a cluster (hidden class variable)

Result:



Dataset

Final clustering

Other types of learning

Semi-supervised learning

- data points: $\{x_1 \dots x_n\}$ "cat" "dog"
- targets: $\{t_1, t_2, \dots, t_k\}$ (k<n)

- Not all datapoints have a known target/label!

only the first $k < n$

- Use all data, also those with unknown target, to create predictor.

image classification

Other types of learning

Reinforcement Learning

- Dynamic environment: provides information on its state. S
- Agent: takes actions, receives rewards from environment. a
- Task: maximize total reward R
- Learning by trial and error
- Application: games

Overview

1. Overview
    1. What is machine learning?
    2. Different types of machine learning
2. **Recap: probability theory**

Probability theory

Probability theory (Bishop)

Provides a consistent framework for the quantification and manipulation of uncertainty.

everything I don't know

Uncertainty in pattern recognition

- Noise on measurements. unknown noise sources
- Finite size datasets. the fewer samples available the less you know about the true distribution

Probability theory

think science

Frequentist interpretation

- Probability of event: fraction of times event occurs in experiment

think intuitive

Bayesian approach

- Probability: quantification of plausibility or the strength of the belief of an event.

Random variables

Random variable X $\sim$ measurable $\sim p(x)$

- Stochastic variable sampled from a set of possible outcomes

$p(x) \geq 0 \quad \forall x \in \mathcal{X}$

- Discrete or continuous
- Probability distribution p(X), $p_X(x)$, $P(X = x)$, $p(x)$

capital lower-case

Examples of discrete random variables:

- Throwing a die: X= # of eyes $\mathcal{X} = \{1, 2, \ldots 6\}$
- Flipping a coin: X= side of the coin

$\mathcal{X} = \{\text{heads}, \text{tails}\}$

$\sum_{x \in \mathcal{X}} p(x) = 1$

Two discrete random variables (I)

2 random variables X & Y

N trials: sample both X and Y.

Joint probability

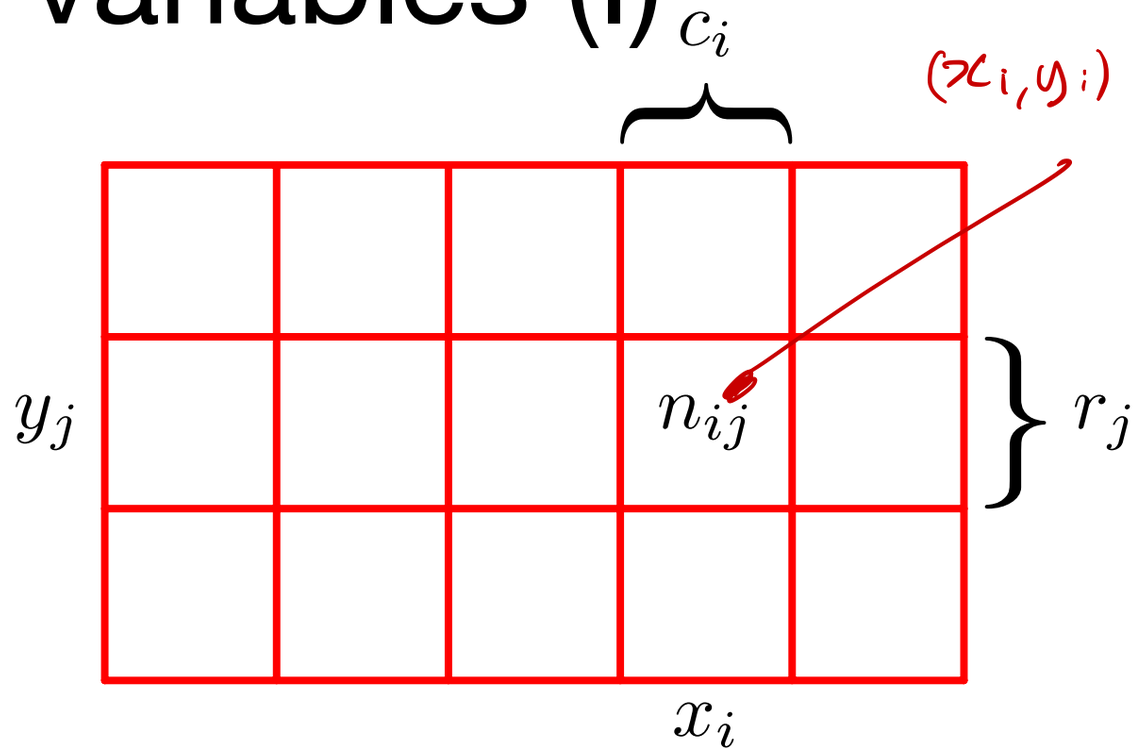


Figure: 2 random variables (Bishop 1.10)

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability of X: $p(X = x_i) = c_i / N$

$$c_i = \sum_{j=1}^{3} n_{ij} \qquad n_{ij} = p(X = x_i, Y = y_j) \cdot N$$

$$p(X = x_i) = \sum_{j=1}^{3} p(X = x_i, Y = y_j)$$

Sum rule / marginalisation "projection"

Two discrete random variables (II)

- 2 random variables

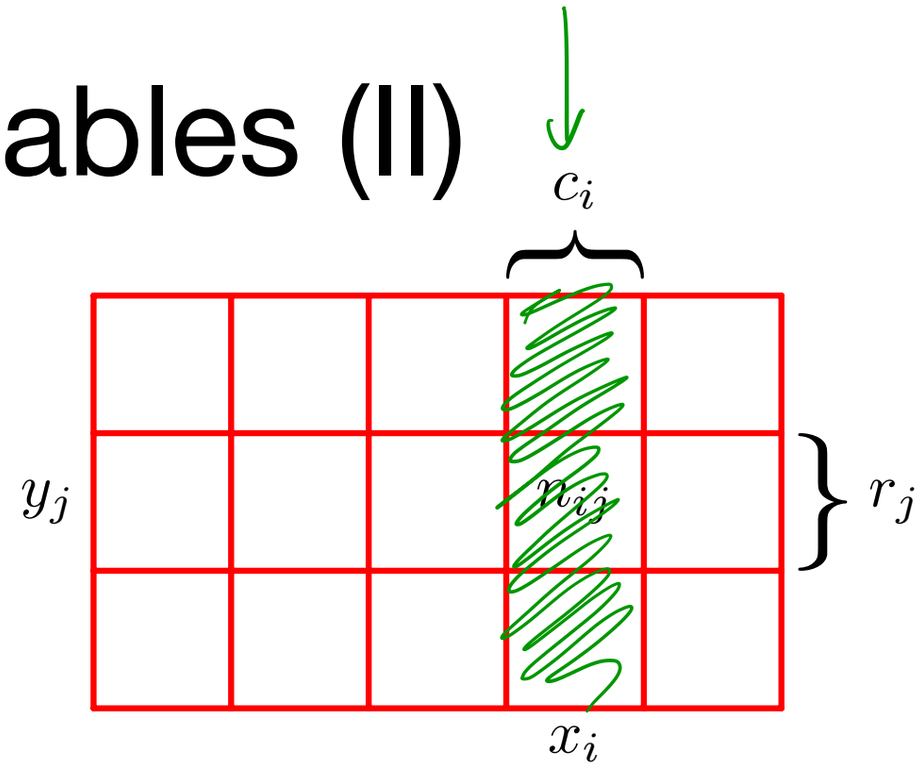


Figure: 2 random variables (Bishop 1.10)

- Conditional probability of Y given X:

$P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$ ← renorm...

- Remember: $p(X = x_i) = \frac{c_i}{N}$

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{P(Y = y_j \mid X = x_i) \cdot c_i}{N}$$

$$P(X = x_i, Y = y_j) = P(Y = y_j \mid X = x_i)\, P(X = x_i)$$

Product rule

Example: Marginal & Conditional distributions

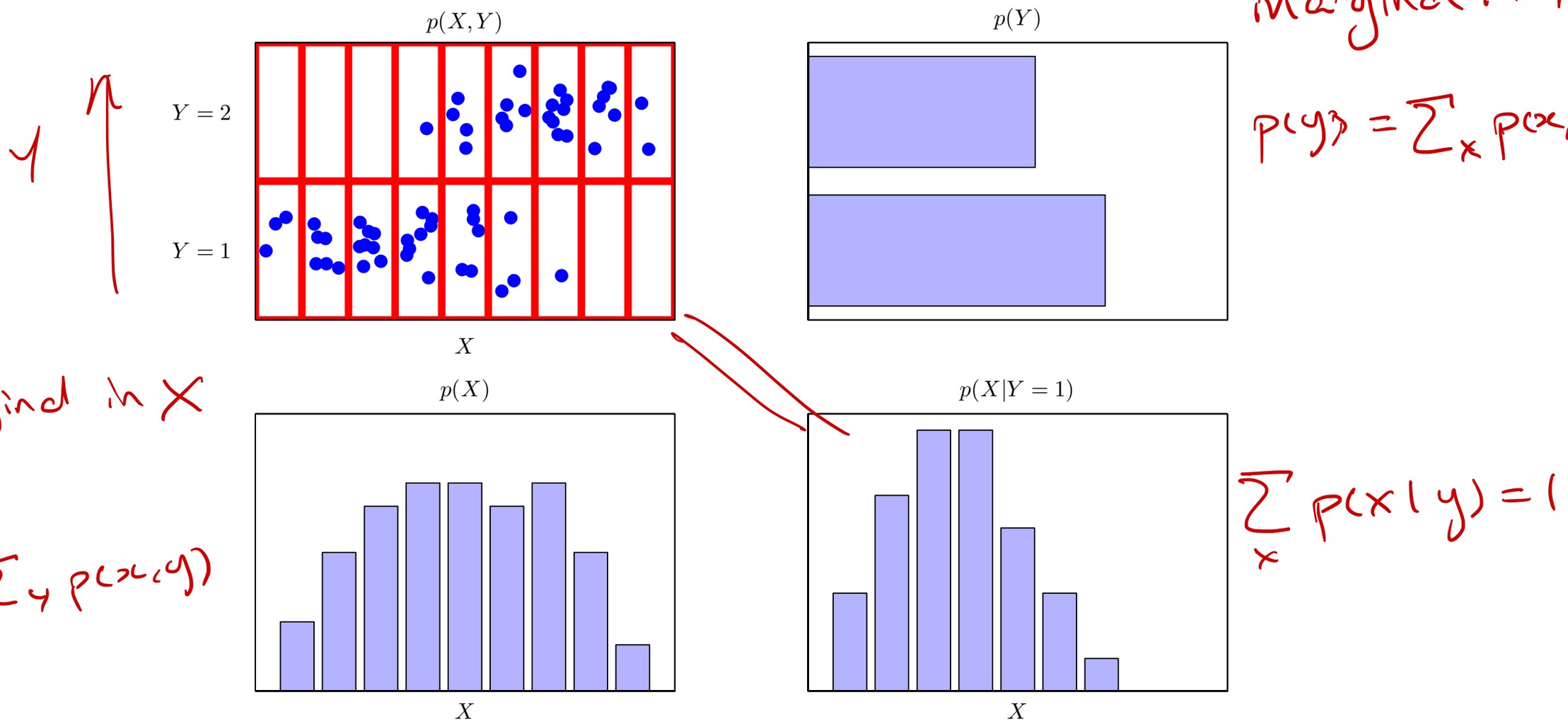


Figure: Marginal and conditional distributions (Bishop 1.11)

Continuous Random Variables

- Probability of $x \in \mathbb{R}$ falling in the interval $(x, x + dx)$ is given by $|p(x)dx|$ prob. mass
- $p(x)$: *probability density* over x
- Probability over finite interval $p(x \in (a, b)) = \int_a^b dx\ p(x), \quad b > a$

- Positivity: $p(x) \geq 0$
- Normalization: $\int_{-\infty}^{\infty} dx\ p(x) = 1$
- Change of variables *x = g(y)*, probabilities in $(x, x + dx)$ must be transformed to $(y, y + dy)$

$$|p_x(x)\,\mathrm{d}x| = |p_y(y)\,\mathrm{d}y| \quad \rightarrow \quad p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

Continuous Random Variables

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} dx' \, p(x')$$

$$p(x) = \frac{dF}{dx}(x)$$

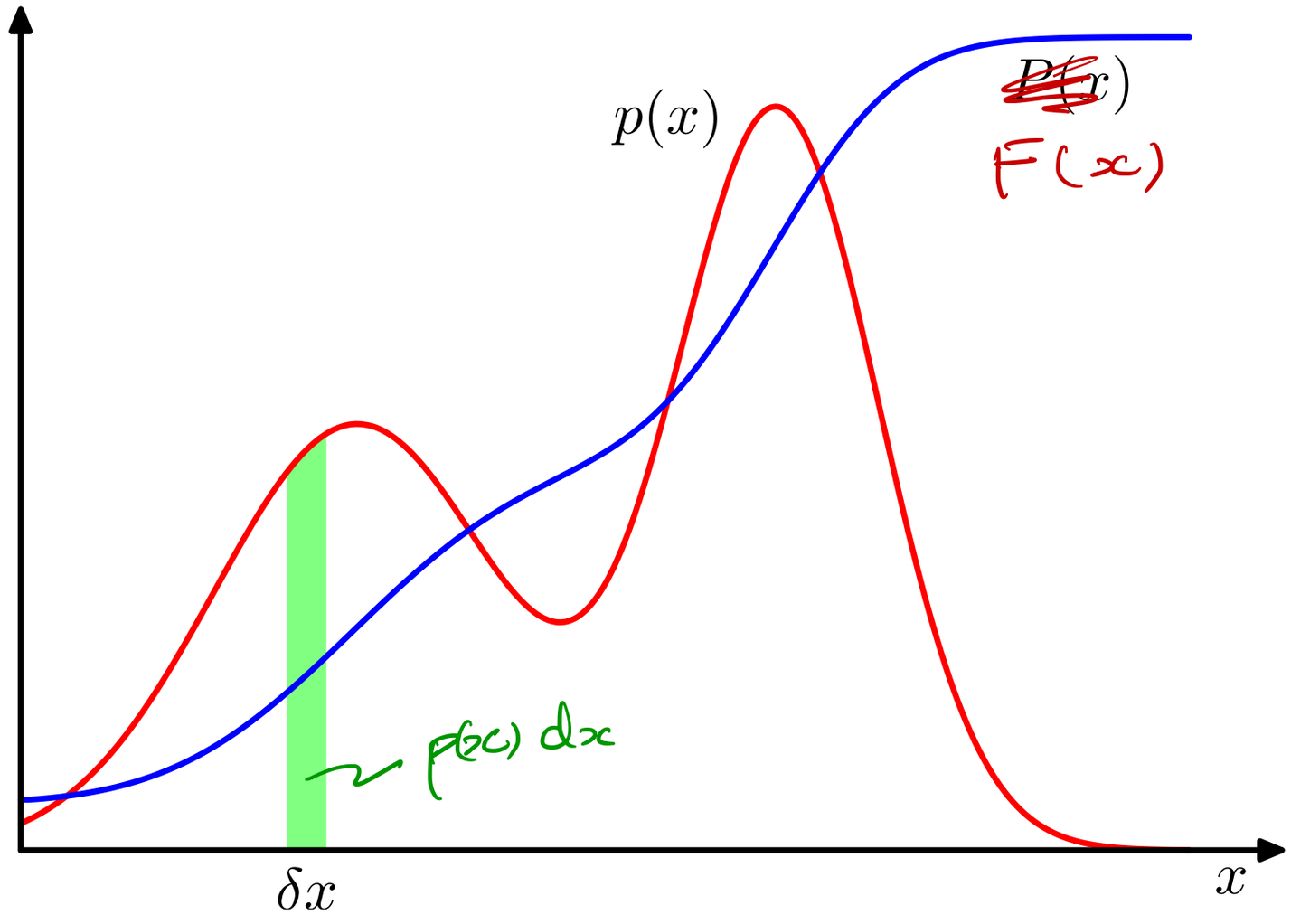


Figure: probability density and cumulative distribution function (Bishop 1.12)

The Rules of Probability Theory

For random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$:

$A = (a, b)$

$p(X \in A) = \int_a^b dx\, p(x)$

	Discrete	Continous
Additivity	$p(X \in A) = \sum_{x \in A} p(x)$	$p(X \in A) = \int_{x \in A} dx\ p(x)$
Positivity	$p(x) \geq 0$	$p(x) \geq 0$
Normalization	$\sum_x p(x) = 1$	$\int_{\mathcal{X}} p(x)dx = 1$
Sum Rule	$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$	$p(x) = \int_y dy\ p(x, y)$
Product Rule	$p(x, y) = p(x\|y)p(y)$	$p(x, y) = p(x\|y)p(y)$

Bayes Theorem

- Product rule $p(x, y) = p(x|y)p(y)$

$$= p(y|x)\,p(x)$$

likelihood, prior

- Symmetry property

- Bayes rule $\Rightarrow$ $p(y|x) = \dfrac{p(x|y)\,p(y)}{p(x)}$

marginal likelihood / evidence / normalization constant

- Denominator: $p(x) = \sum_y p(x|y)p(y)$

$$\Rightarrow \quad 1 = \sum_y \left(\frac{p(x|y)\,p(y)}{p(x)} \right)$$

Bayes Theorem

Bayes rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- p(y) : the prior probability of Y = y
- p(y | x) : the posterior probability of Y = y
- p(x | y) : the likelihood of Y = y given X = x
- p(x) : the evidence for X = x

Example: Fruit in Boxes

- Random variables:

- Prior Box distribution:

$p(B = r) =$

$p(B = b) =$

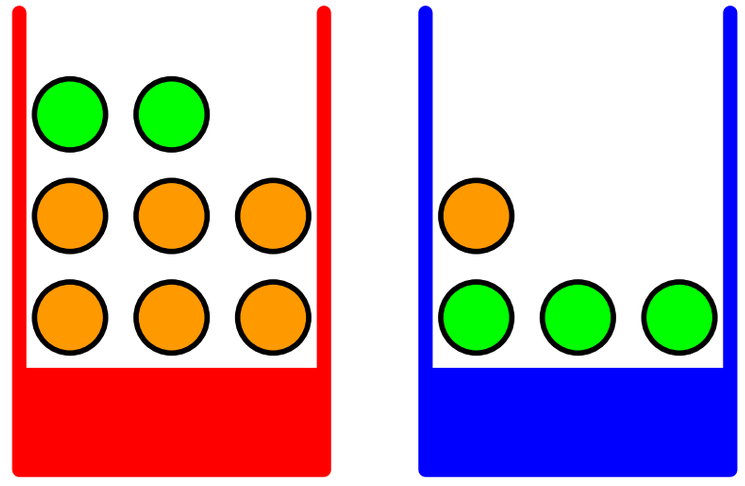


Figure: coloured boxes containing apples and oranges (Bishop 1.9)

- Conditional probabilities of Fruit given Box

- Marginal Fruit distributions:

p(F = a) =

p(F = o) =

Example: Fruit in Boxes

- Prior: p(B = r) = 4/10 & p(B = b) = 6/10
- Marginal: p(F=a) = 11/20 & p(F = o) = 9/20
- Posterior probability of Box color given observed fruit

$$p(B = r|F = o) =$$

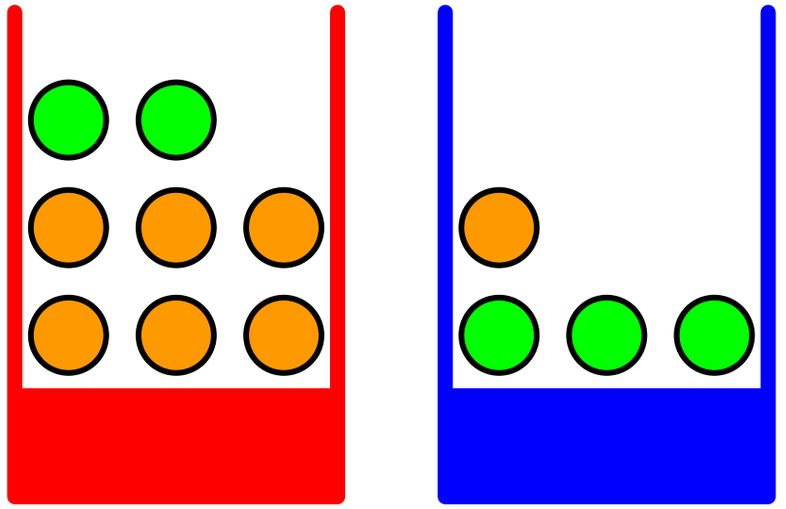


Figure: coloured boxes containing apples and oranges (Bishop 1.9)

- prior probability of red box:

$$p(B = r) = 4/10$$

- After observing an orange the probability of observing a red box is now larger than observing a blue box!

Independent Random Variables

Two random variables *X* and *Y* are *independent* iff measuring *X* gives no information on *Y*, and vice versa.

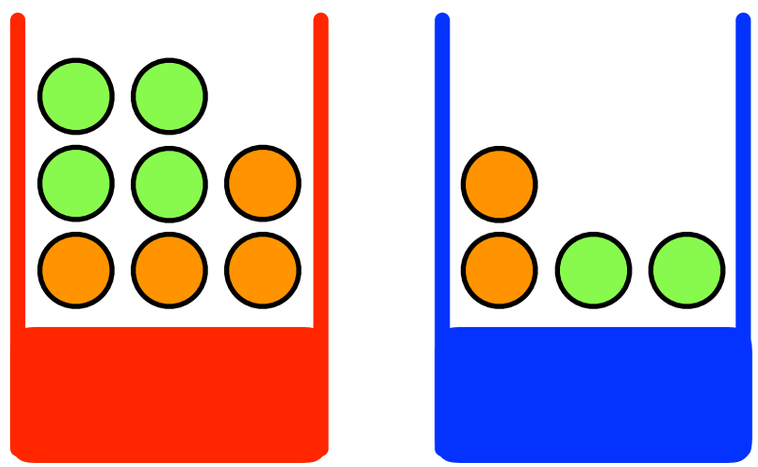
- Formally: *X* and *Y* are called independent if

$$\text{for all } \; x \in \mathcal{X},\ y \in \mathcal{Y}$$

- Equivalent to

$$p(x|y) =$$

- Example:



Expectations

- random variable $X \in \mathcal{X}$ and function $f : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)] =$$

- For *N* points drawn from p(X):

$$\mathbb{E}[f] =$$

- Conditional expectation:

$$\mathbb{E}[f \,|\, y] = \mathbb{E}_{x \sim p(X|Y=y)}[f(x)]$$

Variance

- The expected quadratic distance between f and its mean $\mathbb{E}[f]$

$$\mathrm{var}[f] =$$

Covariance between 2 random variables

- Measures the extent to which *X* and *Y* vary together

$$\mathrm{cov}[x, y] =$$

- Vectors of random variables **x** and **y**, covariance matrix:

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] =$$

- Independent variables:

$$\mathbb{E}[xy] =$$

- Note: $\mathrm{cov}[x, y] = 0$ does not imply

- $\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x}, \mathbf{x}]$

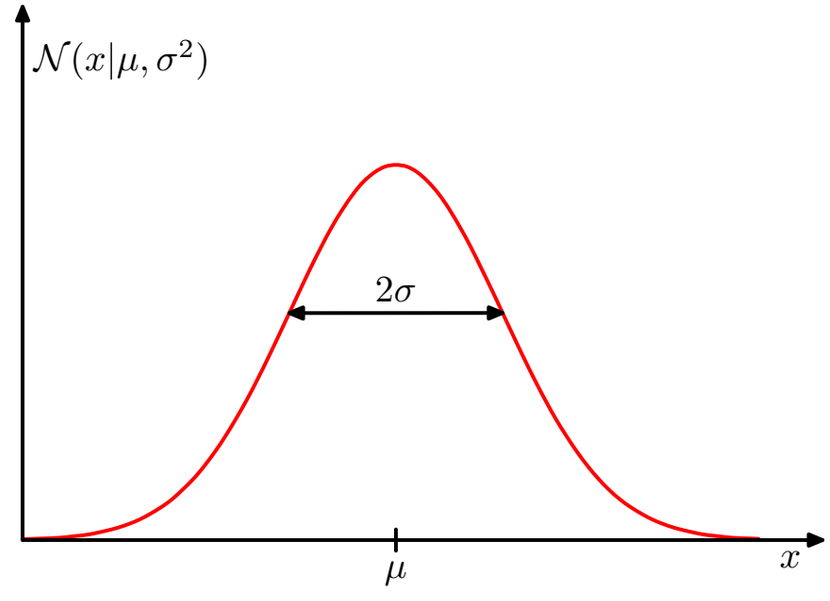
Gaussian Distribution

- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu, \sigma^2) =$$

- Mean:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, \mathrm{d}x =$$



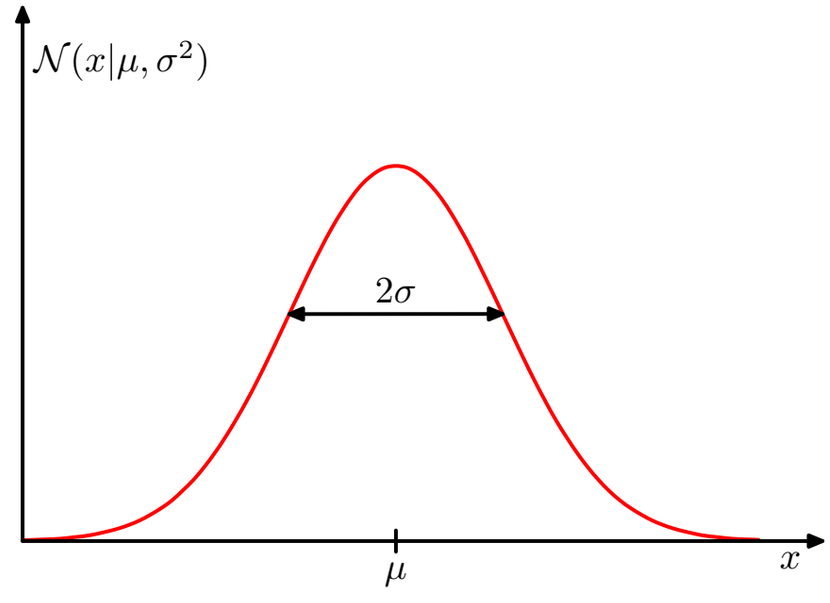
$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

Gaussian Distribution

- Real valued stochastic variable *X*

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

- Variance: $\mathrm{var}[x] = \mathbb{E}[(x-\mathbb{E}[x])^2]$



$$\mathrm{var}[x] =$$

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \sqrt{\pi}$$

$$\int_{-\infty}^{\infty} x^2 e^{-ax^2}\,\mathrm{d}x = -\frac{\partial}{\partial a}\int_{-\infty}^{\infty} e^{-ax^2}\,\mathrm{d}x =$$

Multivariate Gaussian Distribution

- *D*-dimensional vector $\mathbf{x} = (x_1, x_2, ..., x_D)^T$
- $\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) =$

$$|\mathbf{\Sigma}| = \det\mathbf{\Sigma}$$

- $\mathbf{\Sigma} =$ (D x D matrix)
- $\mathbb{E}[\mathbf{x}] =$

$$\int \exp\{-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\}\mathrm{d}^D x = \frac{(2\pi)^{D/2}}{|\mathbf{A}|^{1/2}}$$