UNIVERSITY OF AMSTERDAM
Informatics Institute

AMLAB
Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 1 - Introduction to ML and Probability Theory

Erik Bekkers

Image credit: Kirillm | Getty Images

UNIVERSITY OF AMSTERDAM
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab

Machine Learning 1

Lecture 1.1 - Course Info

Erik Bekkers

Image credit: Kirillm | Getty Images

*Slide credits: Patrick Forré and
Rianne van den Berg*

Admin: Course outline

Lectures and Homework sessions:

- **Lectures**: 2 times per week (each 2 hour) in C1.110
  - Monday 11.00h-13.00h
  - Tuesday 15.00h-17.00h
- **Homework/Practical session**: 2 time per week (each 2 hour)
  - Sessions on Wednesdays, Thursdays, and Fridays
- **Midterm** (2 hour) in week 4 and
- **Final Exam** (3 hour) in week 8

Admin: Grading Policy

- Course component weights:
  - 15 % Homework Assignments
  - 35 % Midterm Exam
  - 50 % Final Exam

- Final course grade:
  - `homework` = average of best 3 out of 4 submissions
  - `exams` = (35 * midterm + 50 * final) / 85
  - `final_grade = MAX( 0.15 * homework + 0.85 * exams, exams)`

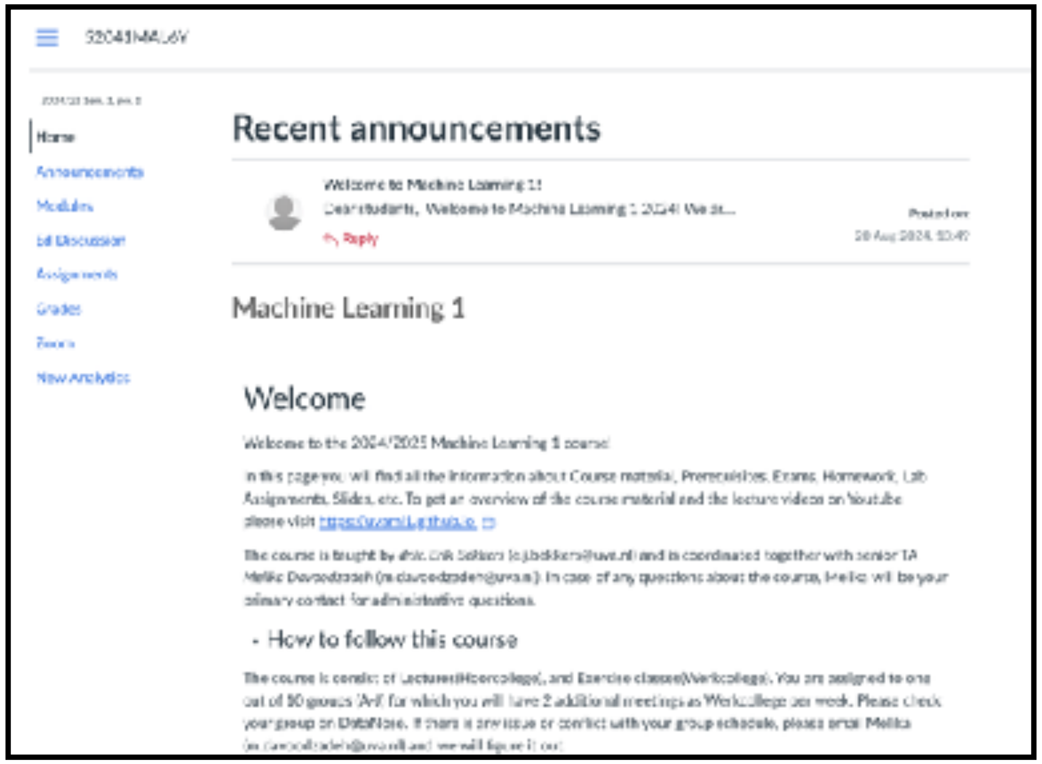
- Passing criterion:
  - `exams >= 5.0` and `final_grade >= 5.5`

Admin: Instructions and Communication

Canvas

All course content and instructions:

- Where to find lectures/exercises
- How to get started
- What is the course content
- …



Recent announcements

Welcome to Machine Learning 1!

Machine Learning 1

Welcome

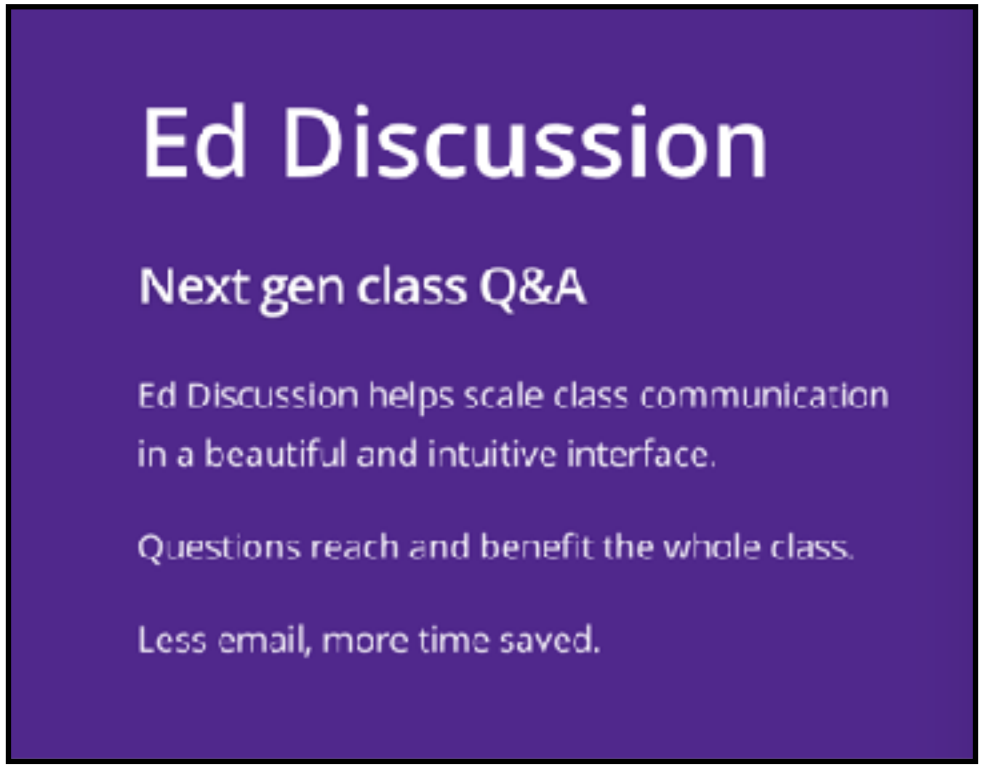
Welcome to the 2024/2025 Machine Learning 1 course!

How to follow this course

Ed Discussion

Here all interaction takes place:

- Ask questions to TAs
- Ask questions to fellow students
- Contribute to answers
- …



Ed Discussion

Next gen class Q&A

Ed Discussion helps scale class communication in a beautiful and intuitive interface.

Questions reach and benefit the whole class.

Less email, more time saved.

Admin: Lectures

- Lectures will *not be recorded*
- For reference/preparation see Youtube playlist of topic-wise video clips (2020/2021) at https://uvaml1.github.io

UvA - Machine Learning 1

Lectures and slides for the UvA Master AI course Machine Learning 1

UvA - Machine Learning 1

Welcome to the public page for the course Machine Learning 1. The course is part of the Artificial Intelligence master program at the University of Amsterdam. The course is developed by the Amsterdam Machine Learning Lab and currently taught by dr.ir. Erik Bekkers.

This page presents an overview of the course including links to the lectures (this Youtube channel) and the corresponding annotated slices in pdf. Students enrolled for the course are referred to Canvas for extra materials such as practice exercises, homework assignments, lab assignments (jupyter notebooks) and additional resources.

In this lecture series we follow closely the Pattern Recognition and Machine Learning book by Bishop. Relevant chapters are indicated at the start of each video.

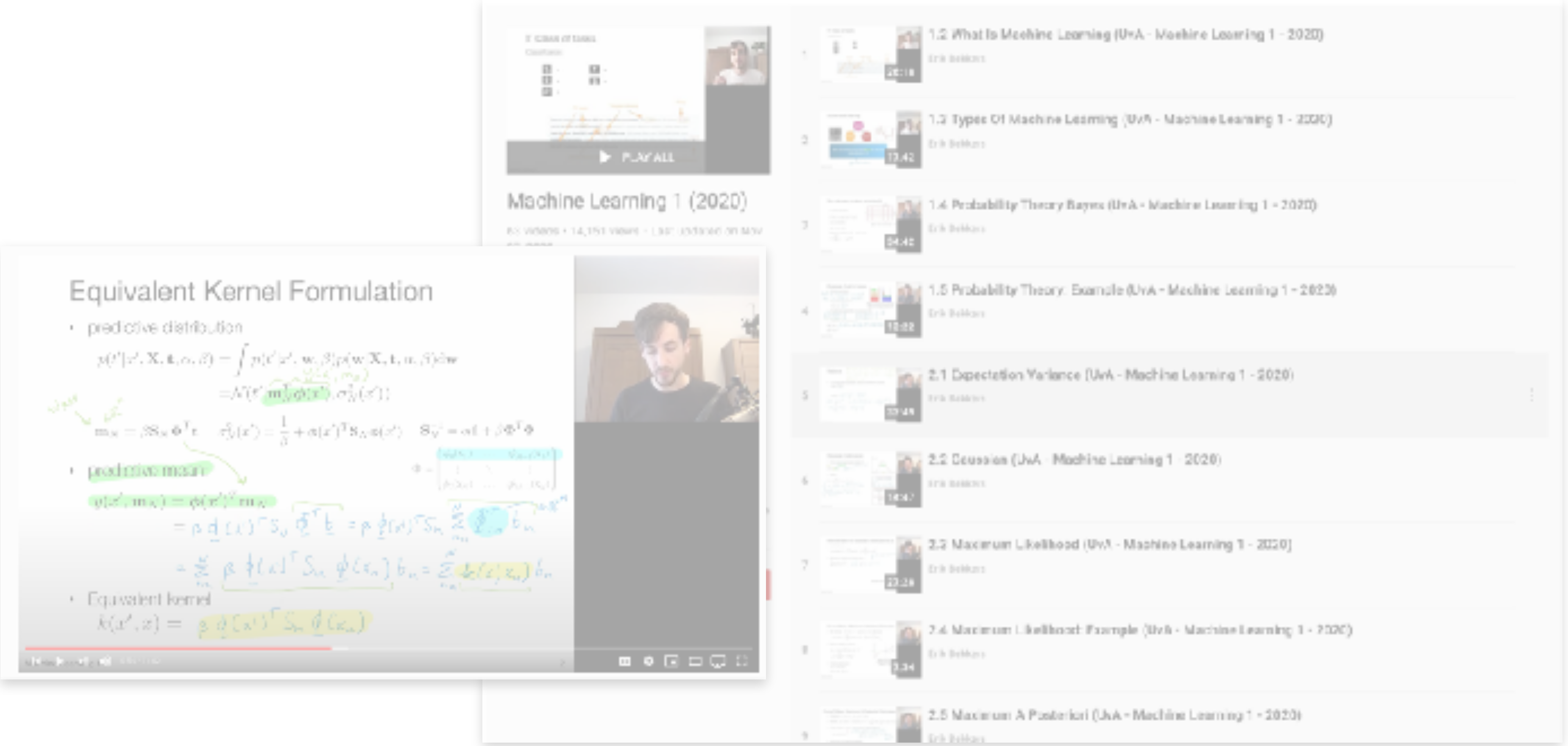
The contents of this page and the video lectures are licensed under a Creative Commons Attribution 4.0 International License.

Weekly overview

Week 1

- Lecture 1.1 (video, pdf): Introduction to the course, administrative announcements
- Lecture 1.2 (video, pdf): What is Machine Learning?
- Lecture 1.3 (video, pdf): Types of Machine Learning
- Lecture 1.4 (video, pdf): Probability Theory, Bayes Theorem
- Lecture 1.5 (video, pdf): Probability Theory: An Example

- Lecture 2.1 (video, pdf): Expectation, Variance, Covariance
- Lecture 2.2 (video, pdf): Gaussian Distribution
- Lecture 2.3 (video, pdf): Maximum Likelihood Estimation
- Lecture 2.4 (video, pdf): Maximum Likelihood Estimation: An Example



Machine Learning 1 (2020)

Equivalent Kernel Formulation

- predictive distribution
- predictive mean
- Equivalent kernel

1. 1.2 What Is Machine Learning (UvA - Machine Learning 1 - 2020)
Erik Bekkers
2. 1.3 Types Of Machine Learning (UvA - Machine Learning 1 - 2020)
Erik Bekkers
3. 1.4 Probability Theory Bayes (UvA - Machine Learning 1 - 2020)
Erik Bekkers
4. 1.5 Probability Theory: Example (UvA - Machine Learning 1 - 2020)
Erik Bekkers
5. 2.1 Expectation Variance (UvA - Machine Learning 1 - 2020)
Erik Bekkers
6. 2.2 Gaussian (UvA - Machine Learning 1 - 2020)
Erik Bekkers
7. 2.3 Maximum Likelihood (UvA - Machine Learning 1 - 2020)
Erik Bekkers
8. 2.4 Maximum Likelihood Example (UvA - Machine Learning 1 - 2020)
Erik Bekkers
9. 2.5 Maximum A Posteriori (UvA - Machine Learning 1 - 2020)
Erik Bekkers

Admin: Lab Assignments

- 2 Programming Assignments:
  - Due in week 3 and 7 (covering weeks 1-3 / 4-7 respectively)
  - Same material as in lectures but with different (applied) perspective
  - Jupyter notebooks (requiring certain *conda* environments)
  - Labs are implementation only (no open questions) and auto-graded
- All assignments are individual
- Pass/Fail, You need to pass both assignments to pass the course
- Late submissions will not be graded
- Labs are uploaded to Canvas, instructions will follow

Admin: Homework Assignments

- 4 Homework Assignments

- Exercise classes on Wednesdays, Thursdays and/or Fridays
  - During class: study practice exercises + ask your questions
  - At home: do the homework assignments

- Must be handed in **individually** and **not made using an LLM**!

- Assignments are released on Canvas and submitted via ANS Delft (online learning platform)
  - Format: Compiled pdf (latex), handwritten copy (pdf), ….
  - **Solutions should be clearly structured, contain derivations, and conclusions**
  - Recommendation: Practice on paper and practice with overleaf.com
  - Instructions at Canvas and during werkcollege

- You can help each other and discuss, but do not copy-paste!
  - We will perform plagiarism checks! Please read the Fraud and Plagriasm Regulations

(see link on Canvas)

Admin: Homework Assignments

- 4 Homework Assignments
- Exercise classes on Wednesdays, Thursdays and/or Fridays
  - During class: study practice exercises + ask your questions
  - At home: do the homework assignments

Without the homeworks you are unlikely to pass the course!

1. They give you the right preparation for the exam
2. Typically the HW grades are higher than that of the exam, thus boosting your grade

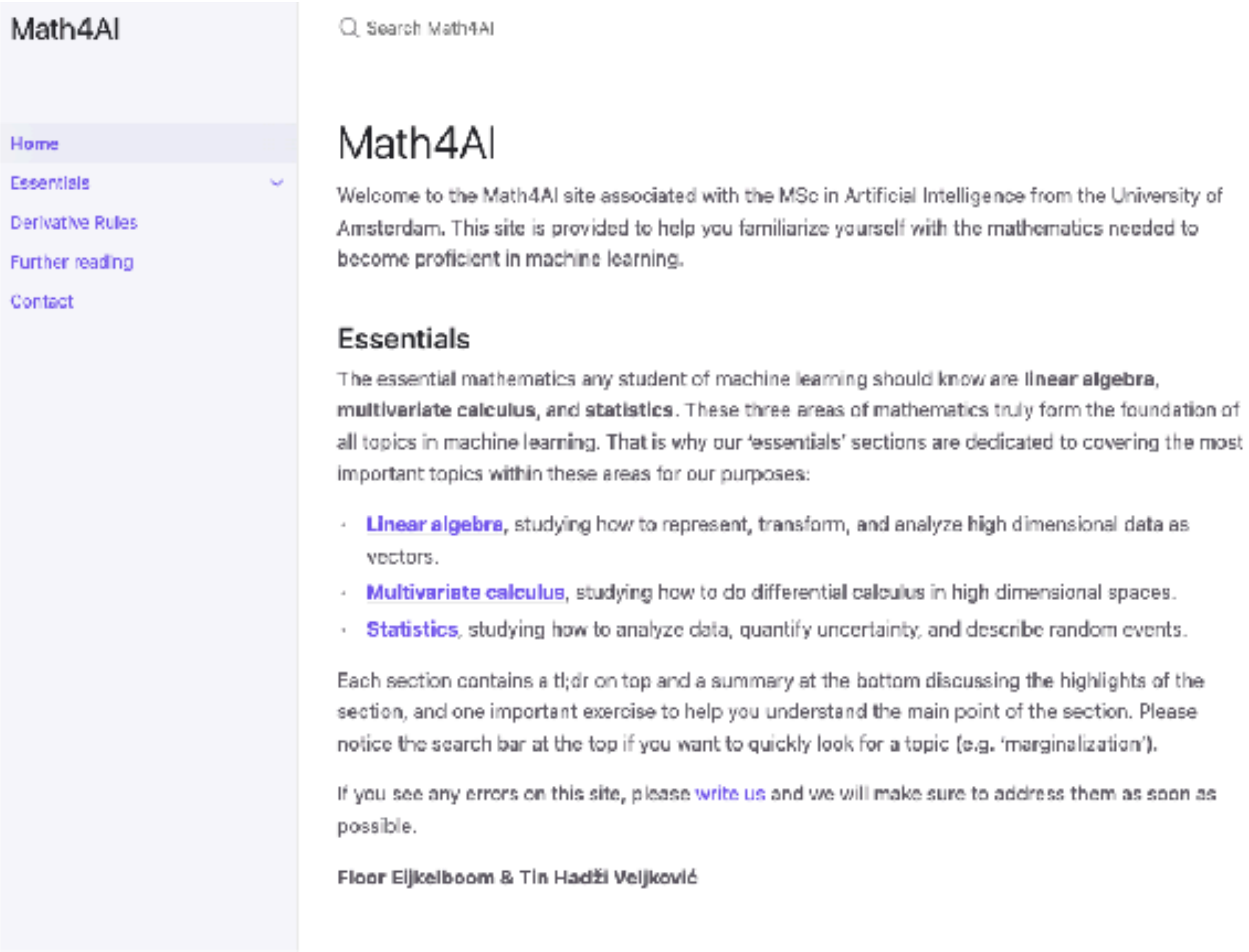
  - **Solutions should be clearly structured, contain derivations, and conclusions**
  - Recommendation: Practice on paper and practice with overleaf.com
  - Instructions at Canvas and during werkcollege
- You can help each other and discuss, but do not copy-paste!
  - We will perform plagiarism checks! Please read the Fraud and Plagriasm Regulations

(see link on Canvas)

Prerequisite Knowledge

- Probability theory
- Programming
- Calculus
- Vector calculus
- Linear algebra

See Canvas: www.math4.ai



Math4AI

Welcome to the Math4AI site associated with the MSc in Artificial Intelligence from the University of Amsterdam. This site is provided to help you familiarize yourself with the mathematics needed to become proficient in machine learning.

Essentials

The essential mathematics any student of machine learning should know are **linear algebra**, **multivariate calculus**, and **statistics**. These three areas of mathematics truly form the foundation of all topics in machine learning. That is why our 'essentials' sections are dedicated to covering the most important topics within these areas for our purposes:

- **Linear algebra**, studying how to represent, transform, and analyze high dimensional data as vectors.
- **Multivariate calculus**, studying how to do differential calculus in high dimensional spaces.
- **Statistics**, studying how to analyze data, quantify uncertainty, and describe random events.

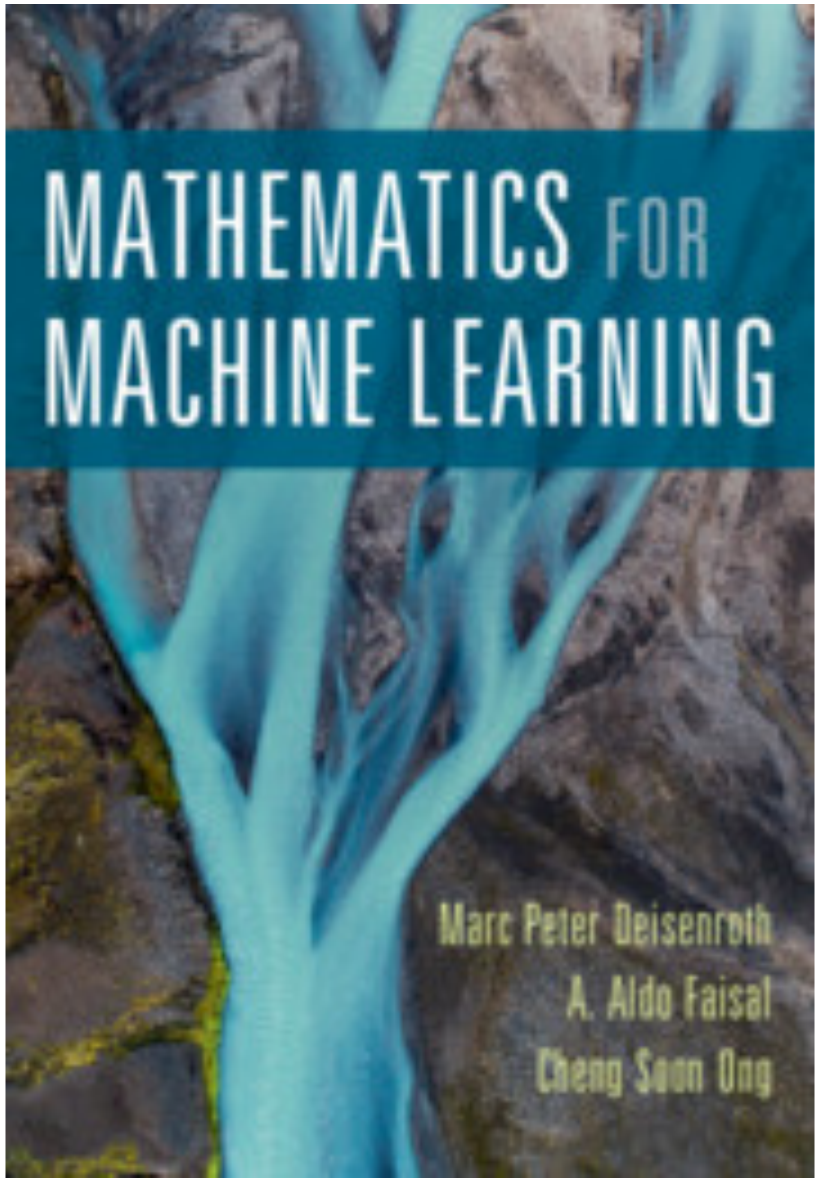
Each section contains a tl;dr on top and a summary at the bottom discussing the highlights of the section, and one important exercise to help you understand the main point of the section. Please notice the search bar at the top if you want to quickly look for a topic (e.g. 'marginalization').

If you see any errors on this site, please write us and we will make sure to address them as soon as possible.

Floor Eijkelboom & Tin Hadži Veljković

Prerequisite Knowledge

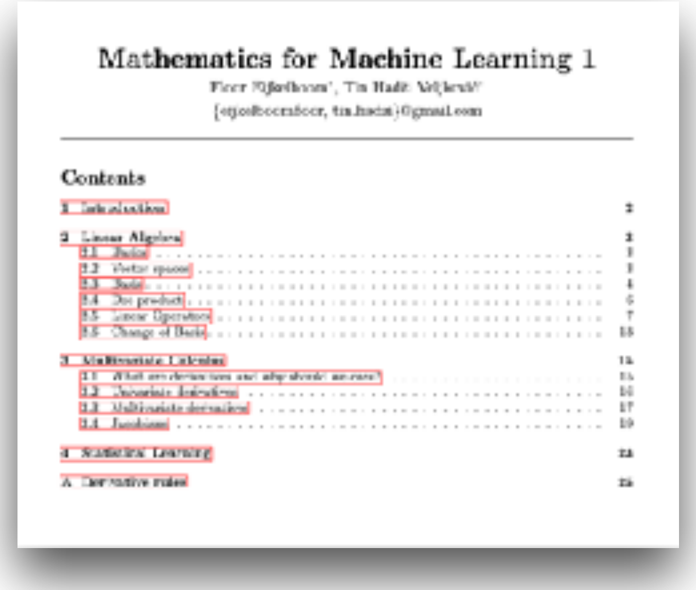
Extra resources:

- 

Prerequisite Knowledge

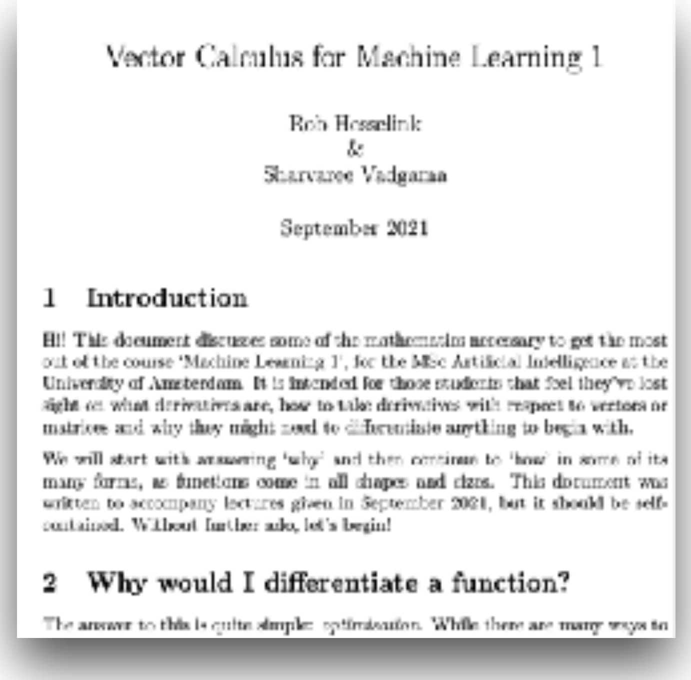
Extra math lectures on Fridays

1. Will send announcement on how to come prepared
2. Purpose, get you up to speed with the required mathematics, from intuitive exposition to formal notation



Mathematics for Machine Learning 1

Contents



Vector Calculus for Machine Learning 1

Rob Hesselink
&
Sharvaree Vadgama

September 2021

1 Introduction

Hi! This document discusses some of the mathematics necessary to get the most out of the course 'Machine Learning 1', for the MSc Artificial Intelligence at the University of Amsterdam. It is intended for those students that feel they've lost sight on what derivatives are, how to take derivatives with respect to vectors or matrices and why they might need to differentiate anything to begin with.

We will start with answering 'why' and then continue to 'how' in some of its many forms, as functions come in all shapes and sizes. This document was written to accompany lectures given in September 2021, but it should be self-contained. Without further ado, let's begin!

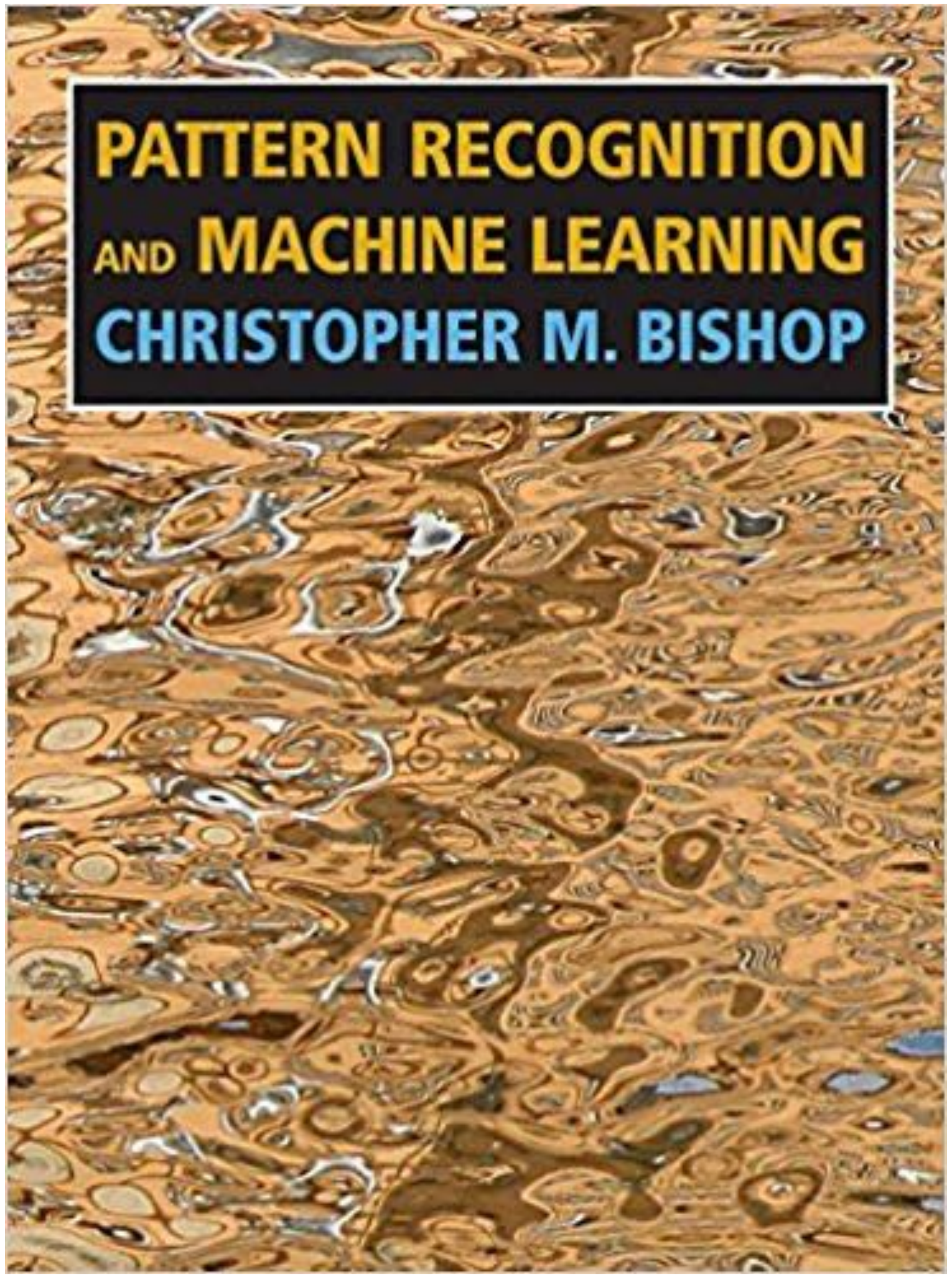
2 Why would I differentiate a function?

The answer to this is quite simple: optimisation. While there are many ways to

Literature: Main book

Pattern Recognition and Machine Learning

- Christopher Bishop



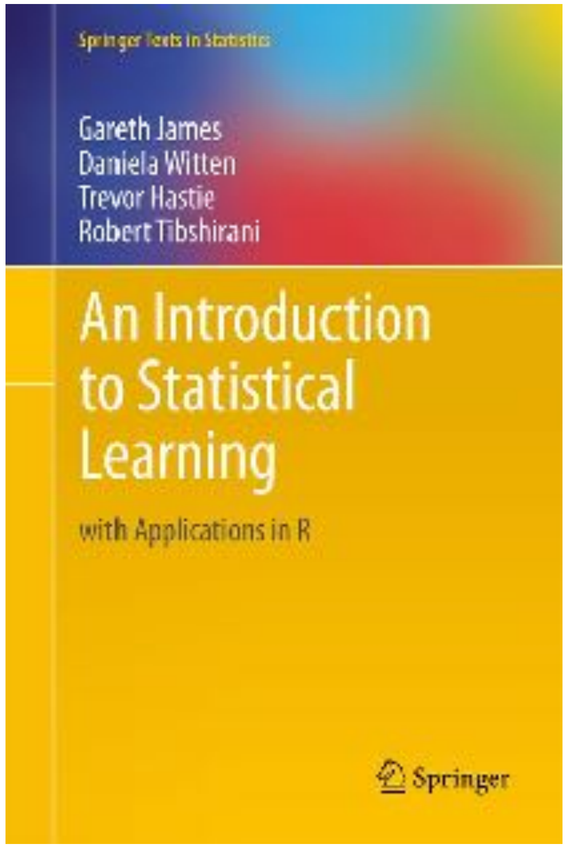
Machine learning from a probabilistic point of view

Literature: Other books

An Introduction to Statistical Learning

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

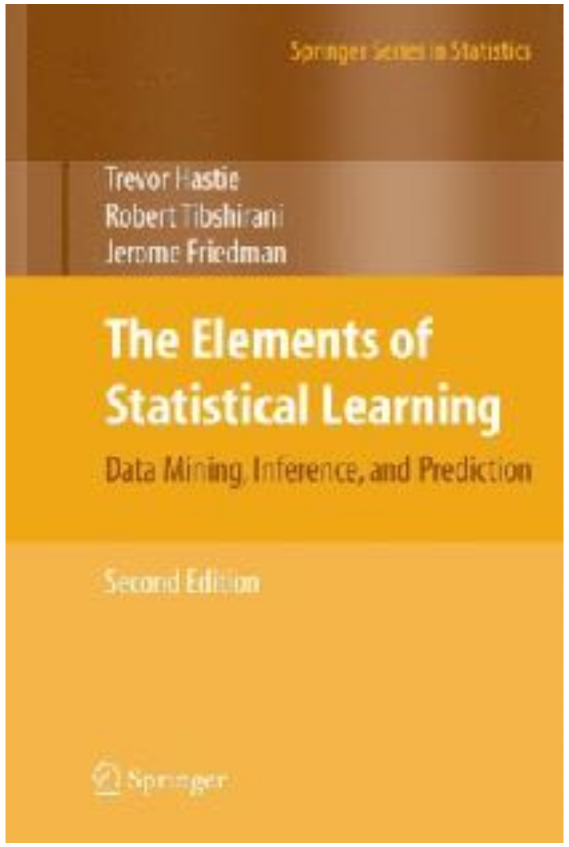
Introduction to Machine learning as a statistical tool.



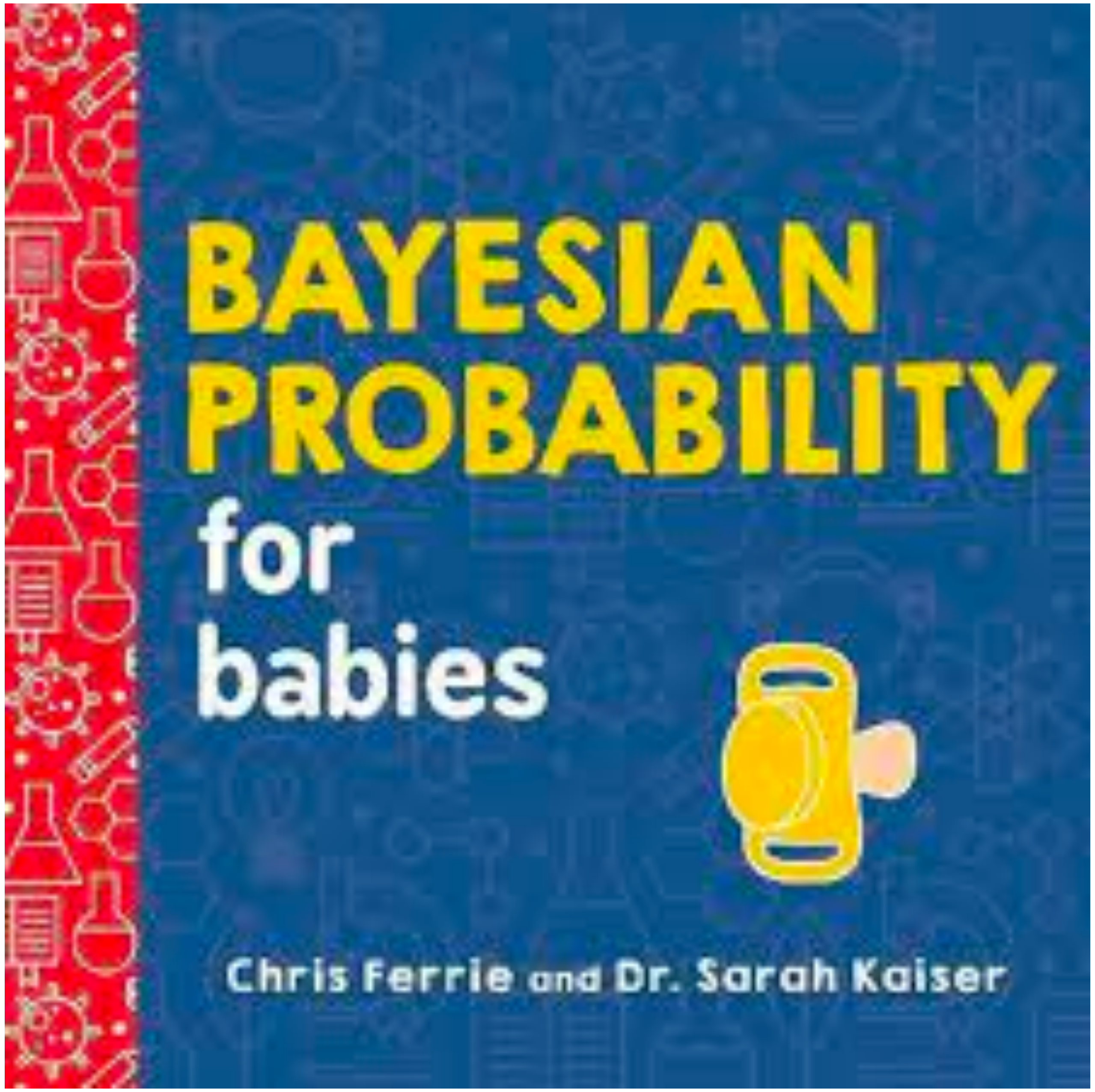
The Elements of Statistical Learning

- Trevor Hastie, Robert Tibshirani, Jerome Friedman

More advanced view of Machine learning as a statistical tool.



Literature: Other books



Team:

Course Coordinator/Lecturer:

Erik Bekkers

Math4AI team:

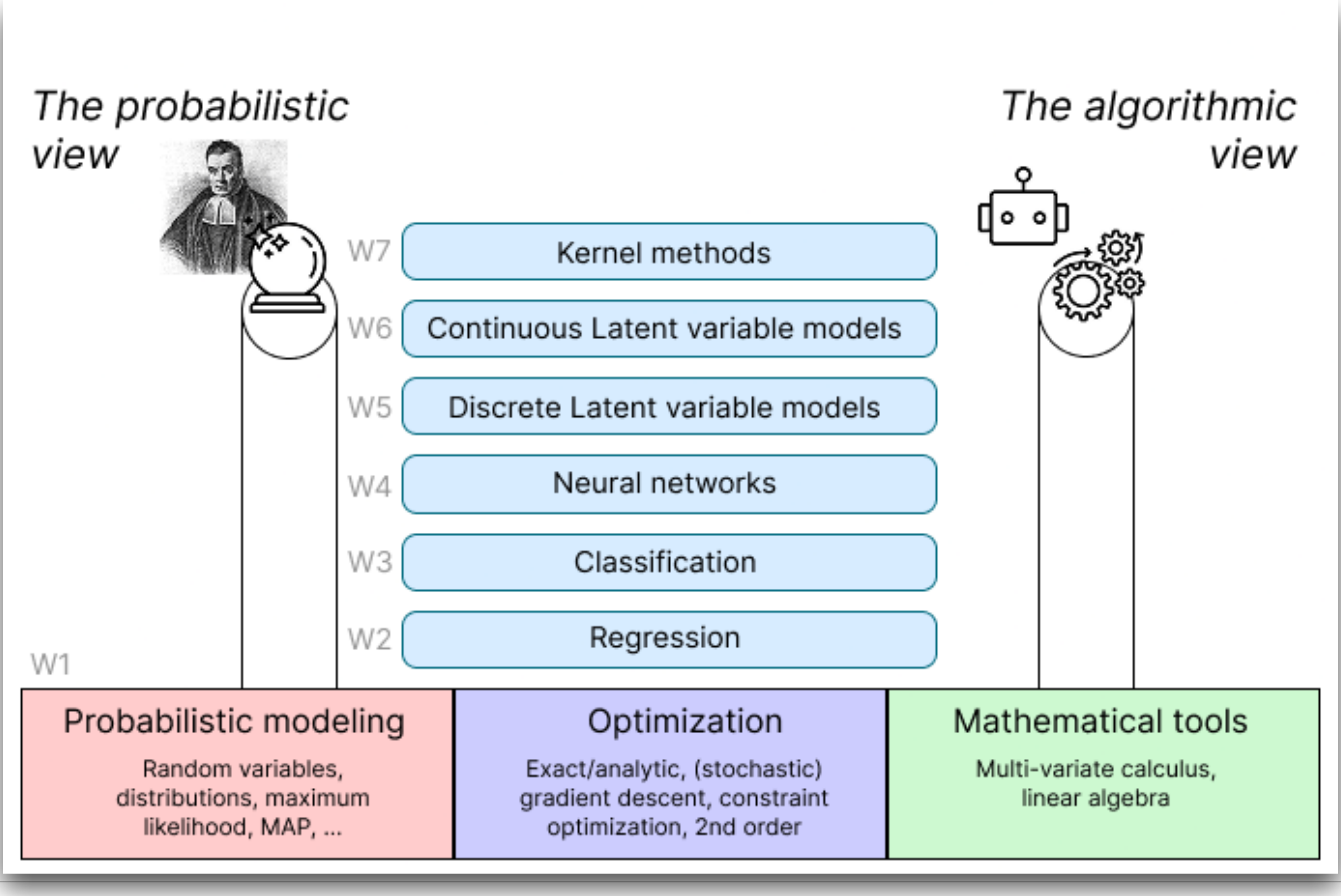
Floor Eijkelboom

Tin Hadži Veljković

Teaching Assistants:

A. Maria Esteban Casadevall
B. Benjamin Hučko
C. Alessio Colombo
D. Stipe Frković
E. Freek Byrman
F. Izabela Kurek
G. Madelon Bernardy
H. Martyna Sendrowska
I. Bhavesh Sood
J. Federica Valeau
K. Pedro M.P. Curvo
L. Emma Kasteleyn

Finally… What's this course about?



The probabilistic view

The algorithmic view

W7 Kernel methods

W6 Continuous Latent variable models

W5 Discrete Latent variable models

W4 Neural networks

W3 Classification

W2 Regression

W1

Probabilistic modeling	Optimization	Mathematical tools
Random variables, distributions, maximum likelihood, MAP, …	Exact/analytic, (stochastic) gradient descent, constraint optimization, 2nd order	Multi-variate calculus, linear algebra

UNIVERSITY OF AMSTERDAM
Informatics Institute

Machine Learning 1

Lecture 1.2 - What is Machine Learning?

Erik Bekkers

(Bishop 1.0 and 1.1)

Image credit: Kirillm | Getty Images

Slide credits: Patrick Forré and Rianne van den Berg

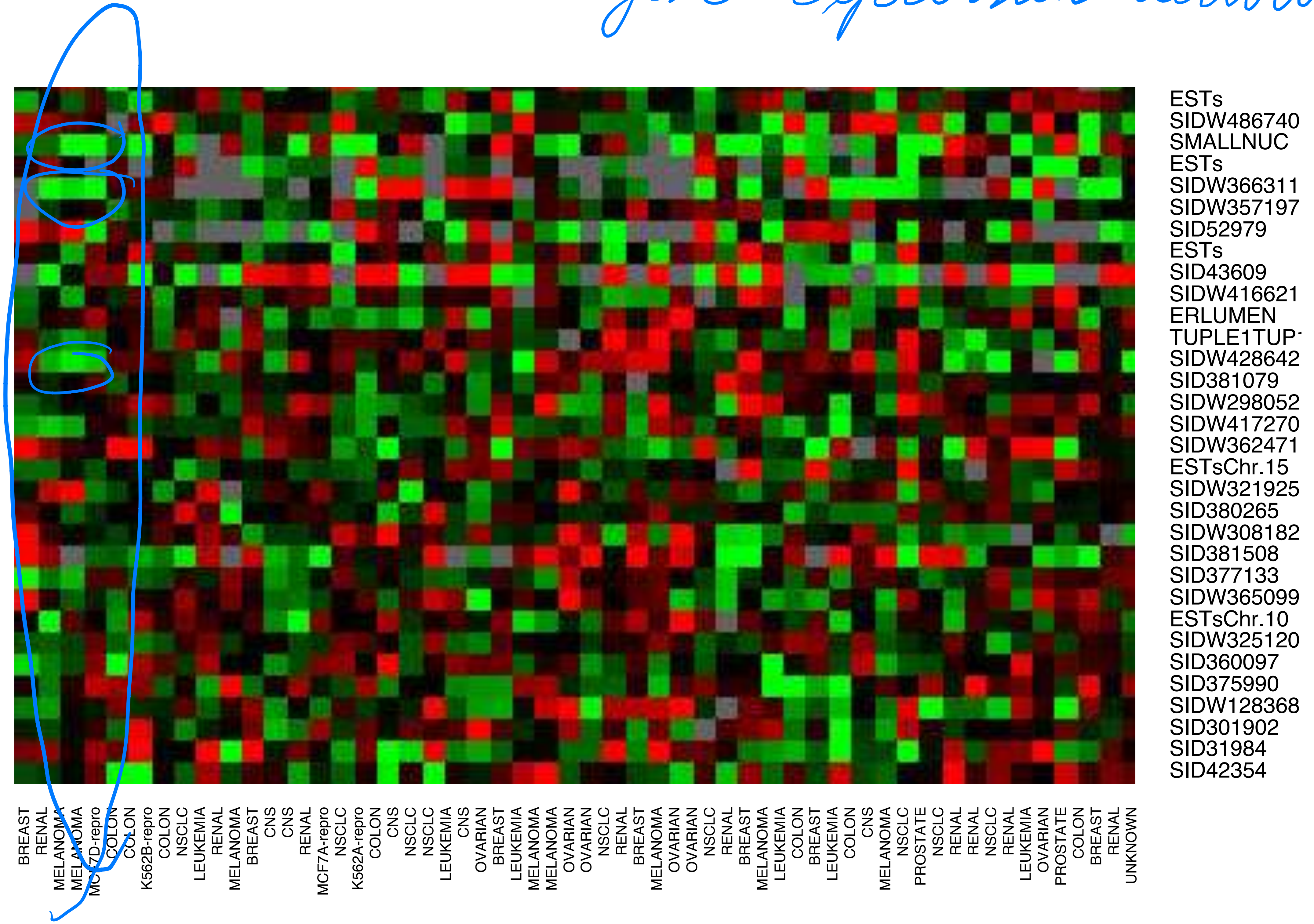
E: Experience

One data point of experience is a 28x28 pixels

MNIST dataset

E: Experience

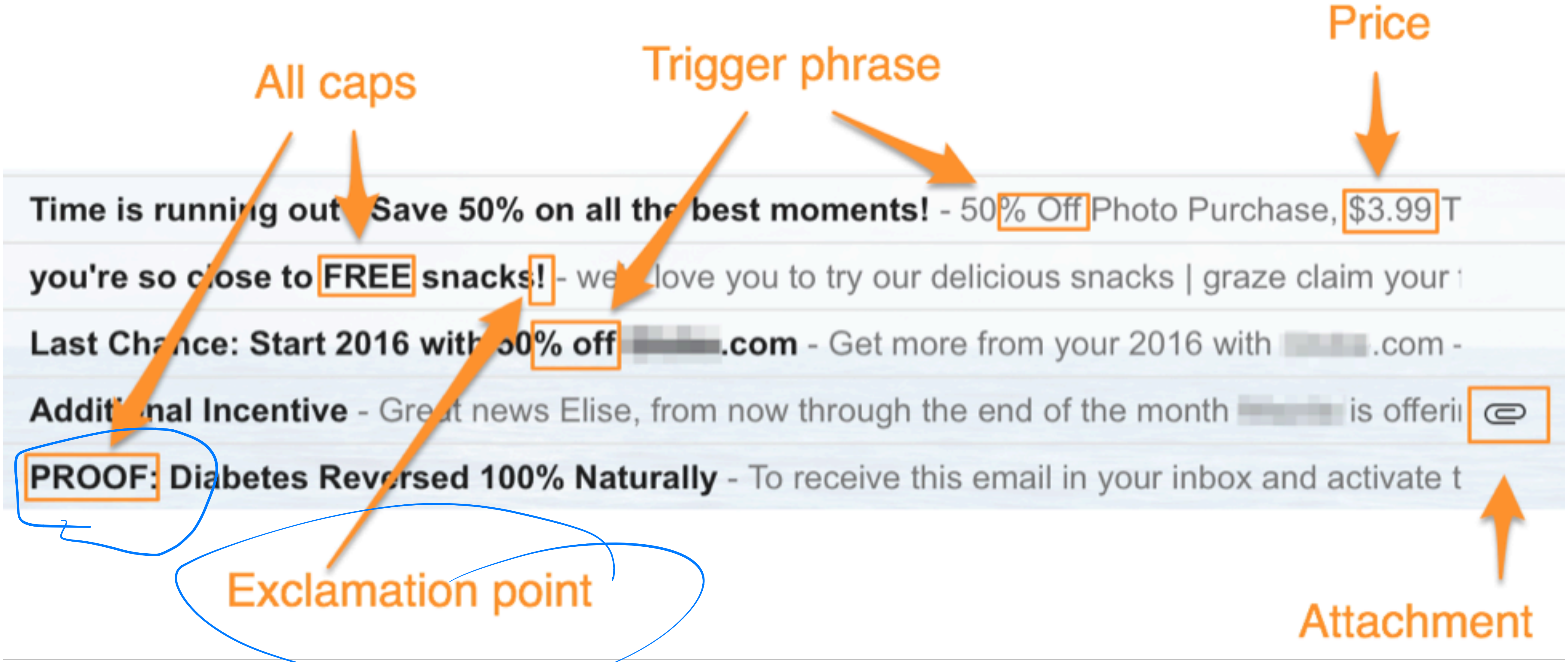
→ tumor type and vector of genes expression activities



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

E: Experience

text (spam vs normal)



Time is running out Save 50% on all the best moments! - 50% Off Photo Purchase, $3.99 T

you're so close to FREE snacks! - we'd love you to try our delicious snacks | graze claim your

Last Chance: Start 2016 with 50% off [redacted].com - Get more from your 2016 with [redacted].com -

Additional Incentive - Great news Elise, from now through the end of the month [redacted] is offeri

PROOF: Diabetes Reversed 100% Naturally - To receive this email in your inbox and activate t

Examples of spam emails. [source: Yesware]

- Notation $\in$ means element of (the set).

  So $x \in \{cat, dog\}$ means x can take on the value cat, or dog, but not something else.

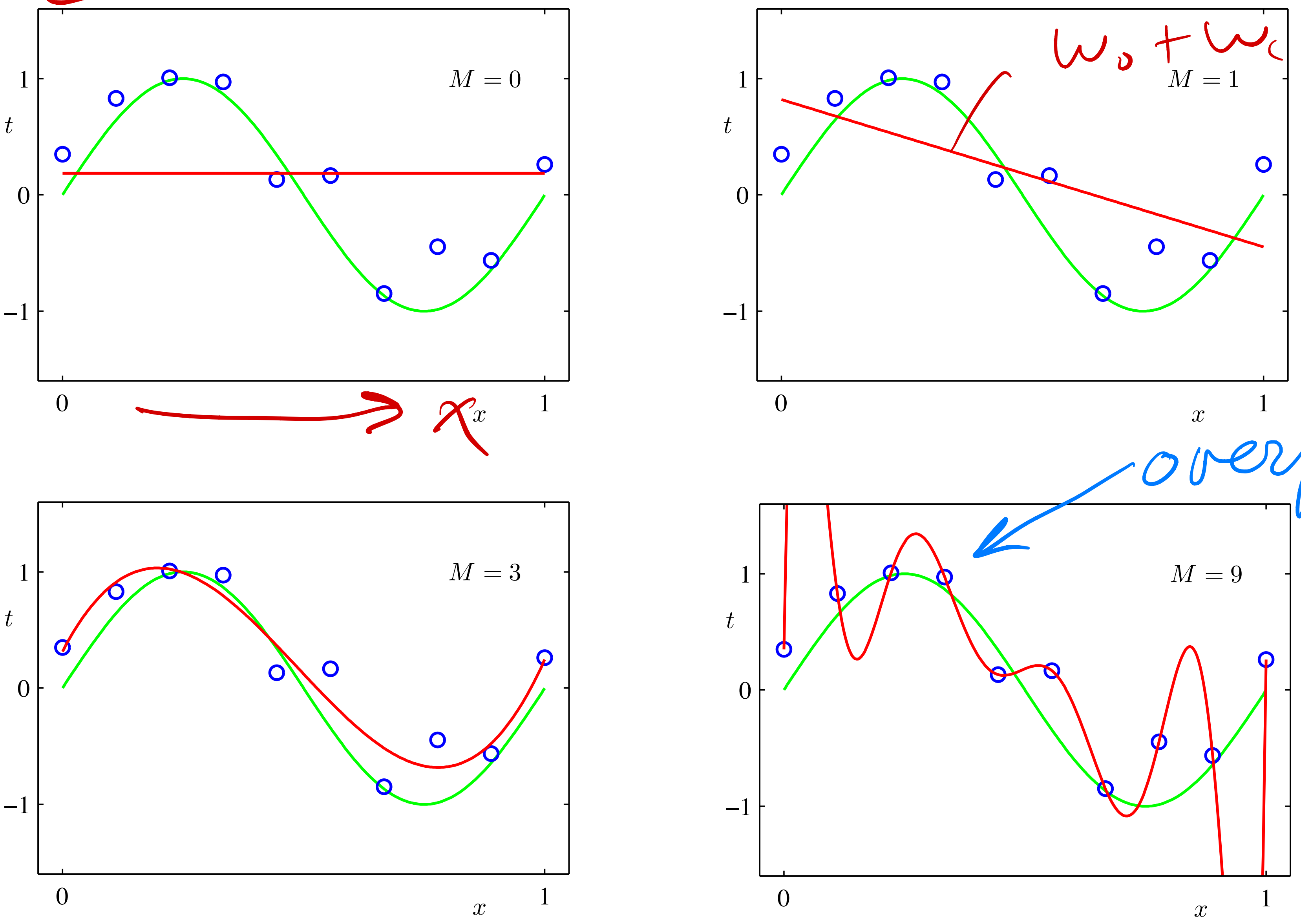
  $x \in \mathbb{R}$ means it take any value between $-\infty$ and $+\infty$

- The hat ^ above y really is meant to introduce a new symbol. It is to say $\hat{y}$ is a different variable than y. I might as well called it z, α, β, or ρ, but that would be confusing

T: Class of tasks

Regression

$f_{\underline{w}}(x) = w_0 + w_1 \cdot x + w_2 x^2 + w_3 x^3 + \dots + w_n x^M$



$w_0 + w_1 x$

$\longrightarrow x$

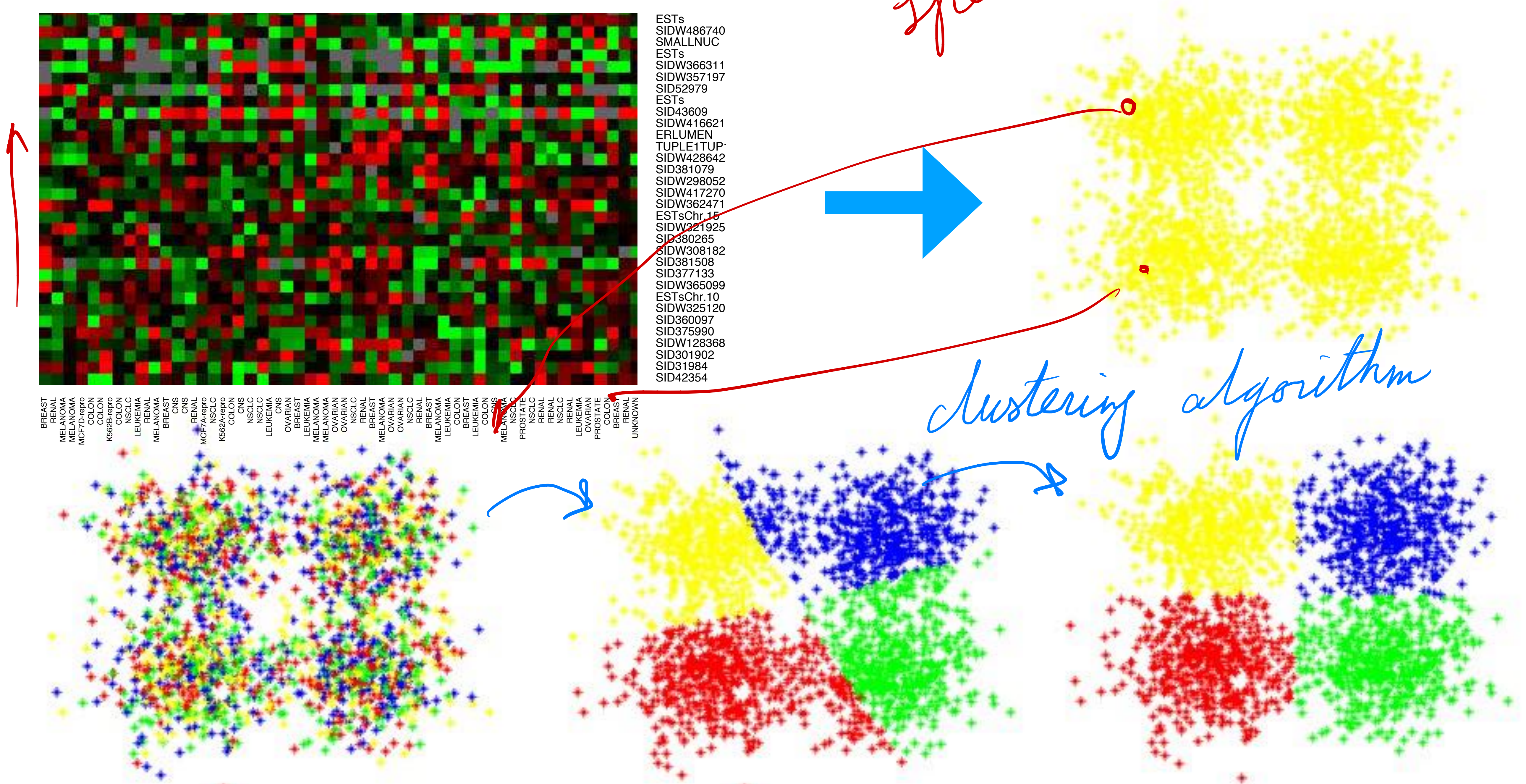
overfitting!!!

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

T: Class of tasks

Clustering

project to lower dim space



clustering algorithm

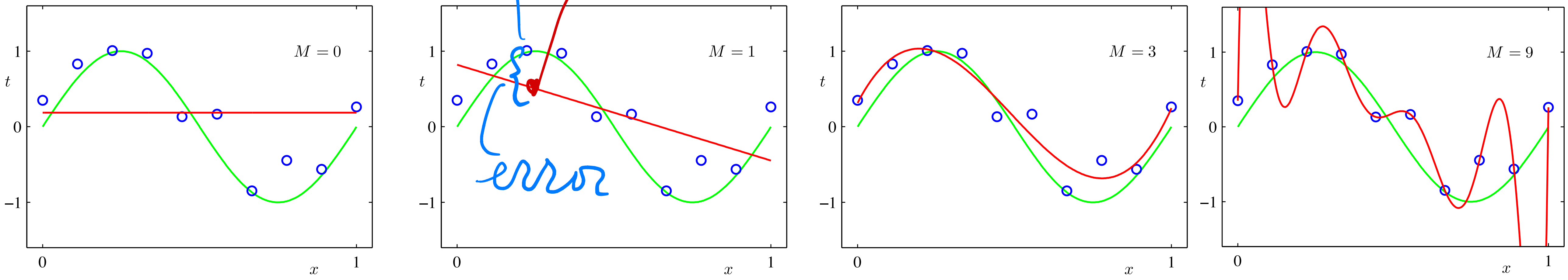
Expression matrix of genes (rows) for 64 human tumor samples (columns). [source: ESL 1.3]

P: Performance measure

Regression

y_i

$\hat{y}_i = f_{\underline{w}}(x_i)$

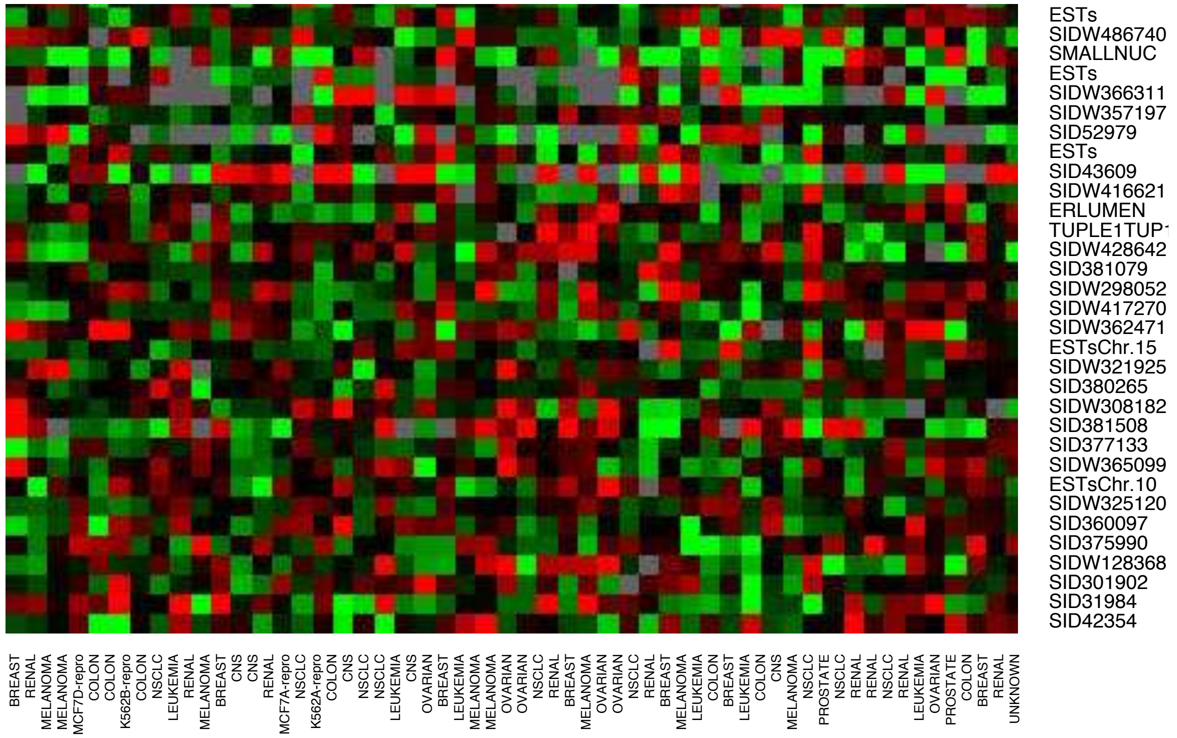


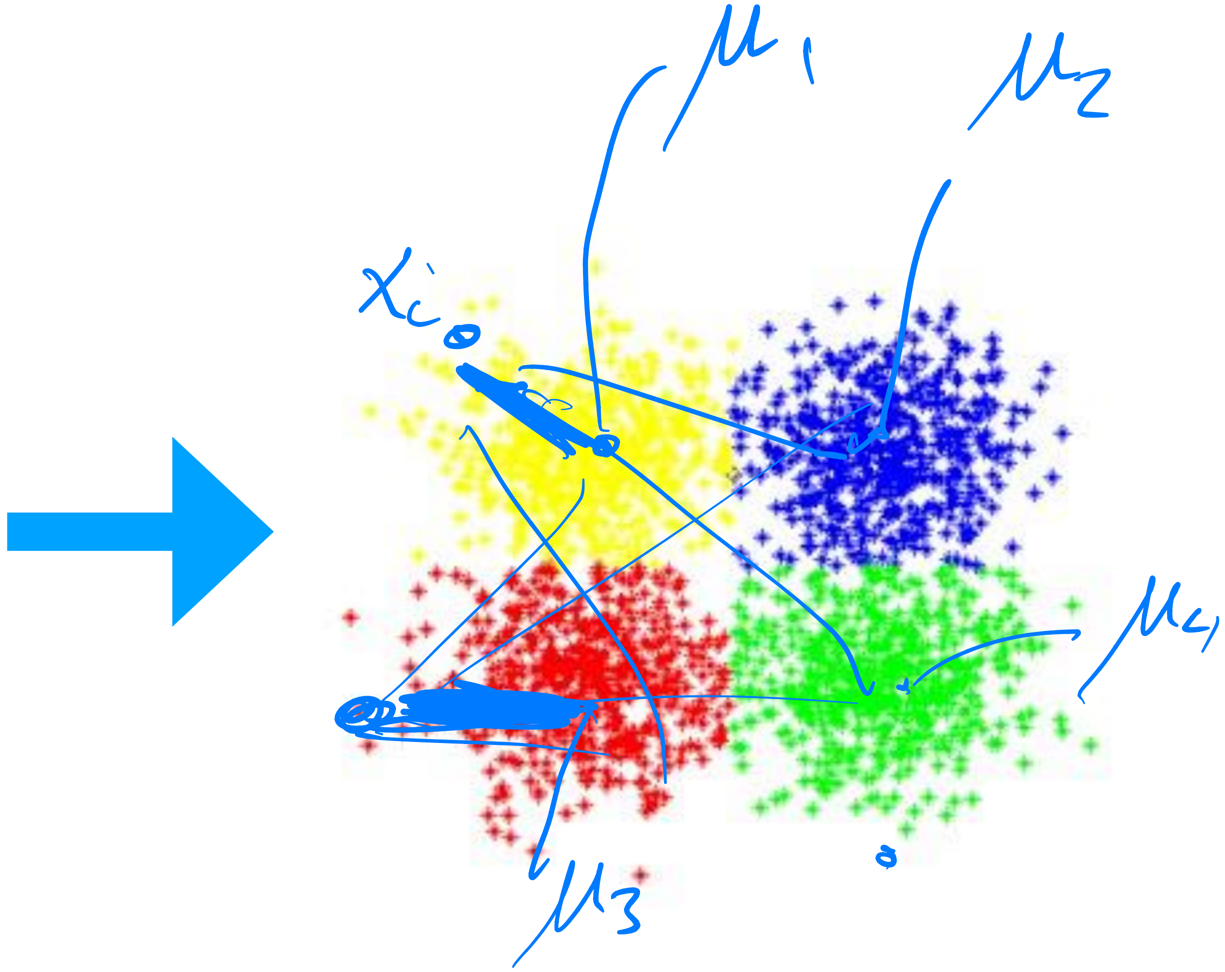
$$\mathrm{MSE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}}\sum_{i=1}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2$$

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

P: Performance measure

Clustering



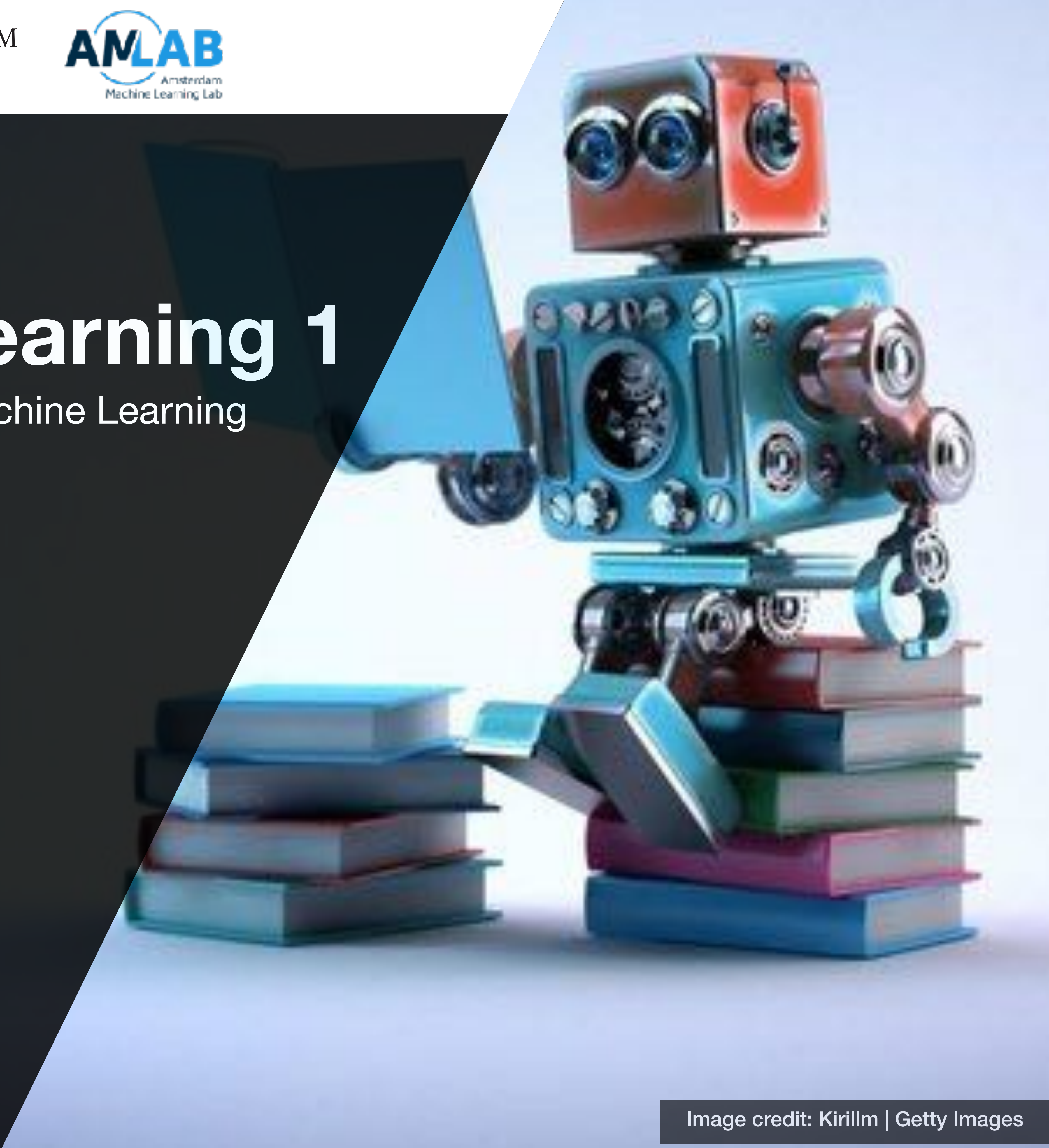


$$\text{within cluster sum of squares} = \sum_{i=1}^{n_{\text{samples}}} \min_{\mu_j \in C} \| x_i - \mu_j \|^2$$

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

UNIVERSITY OF AMSTERDAM
Informatics Institute

AMLAB
Amsterdam
Machine Learning Lab



Machine Learning 1

Lecture 1.3 - Types of Machine Learning

Erik Bekkers

(Bishop 1.0 and 1.1)

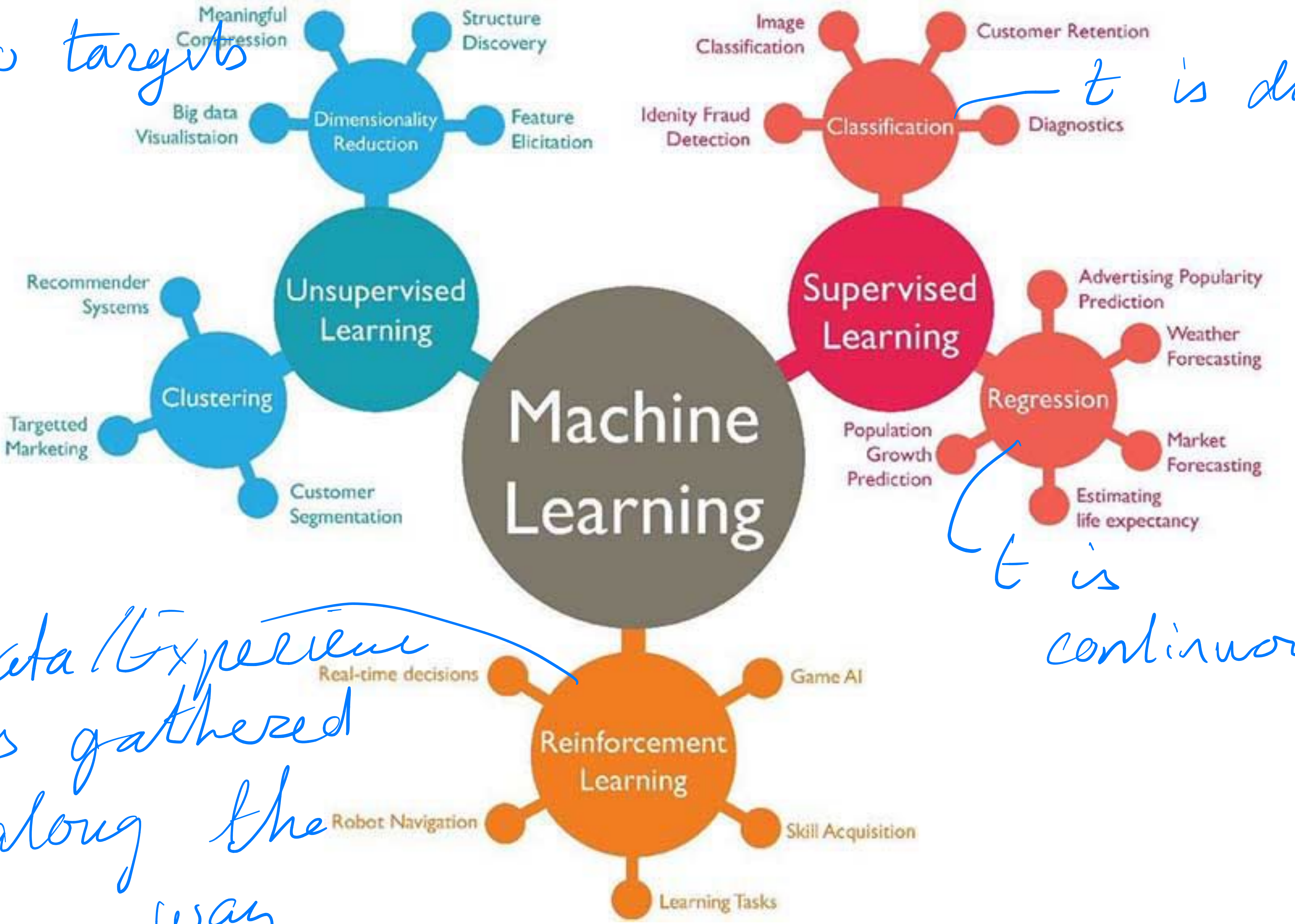
*Slide credits: Patrick Forré and
Rianne van den Berg*

Image credit: Kirillm | Getty Images

$D = \{x_i\}_{i=1}^{N}$

no targets

Experience: $D = \{(x_i, t_i)\}_{i=1}^{N}$



t is discrete

t is continuous

Data/Experience is gathered along the way

Image source : www.techleer.com

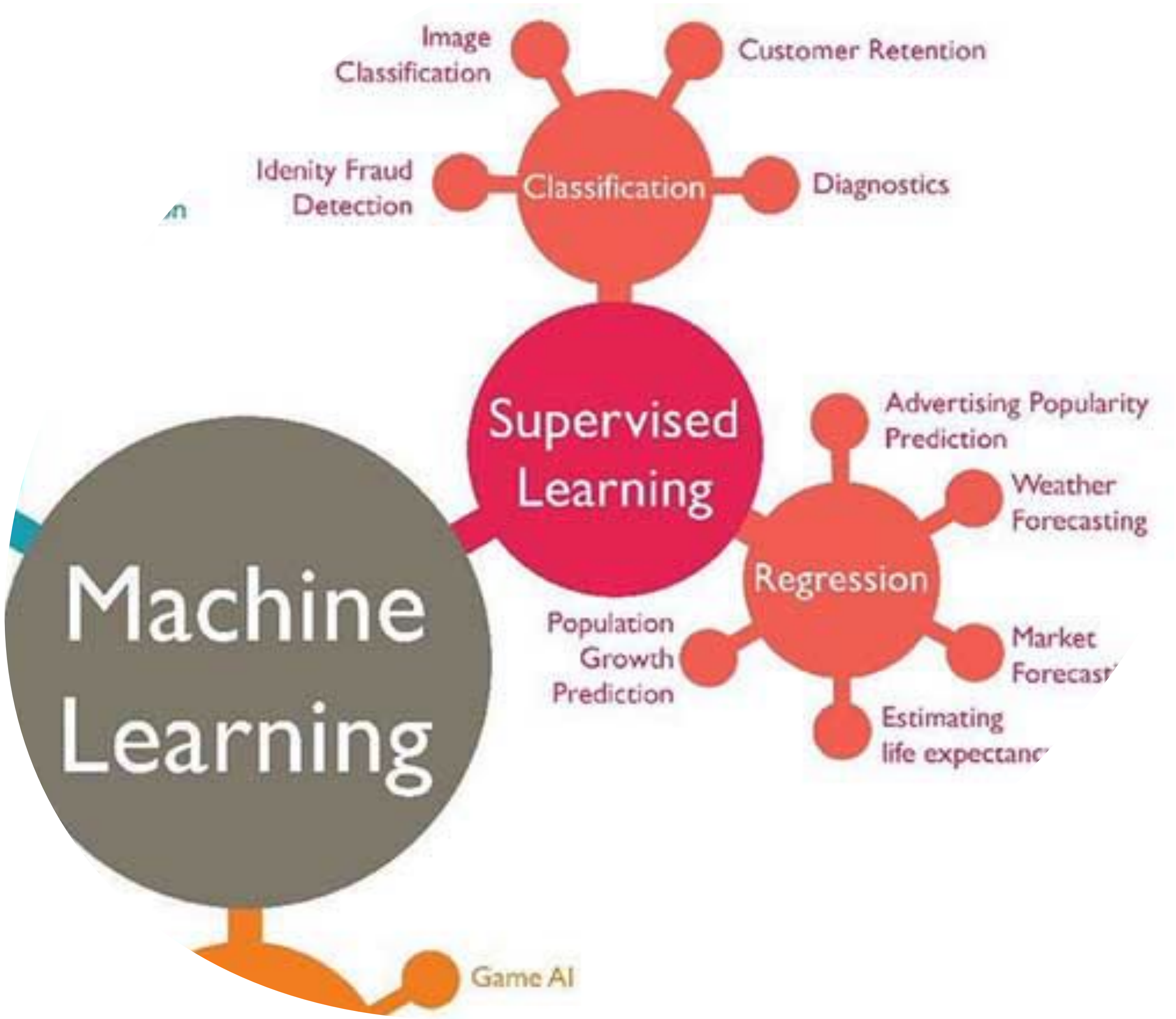
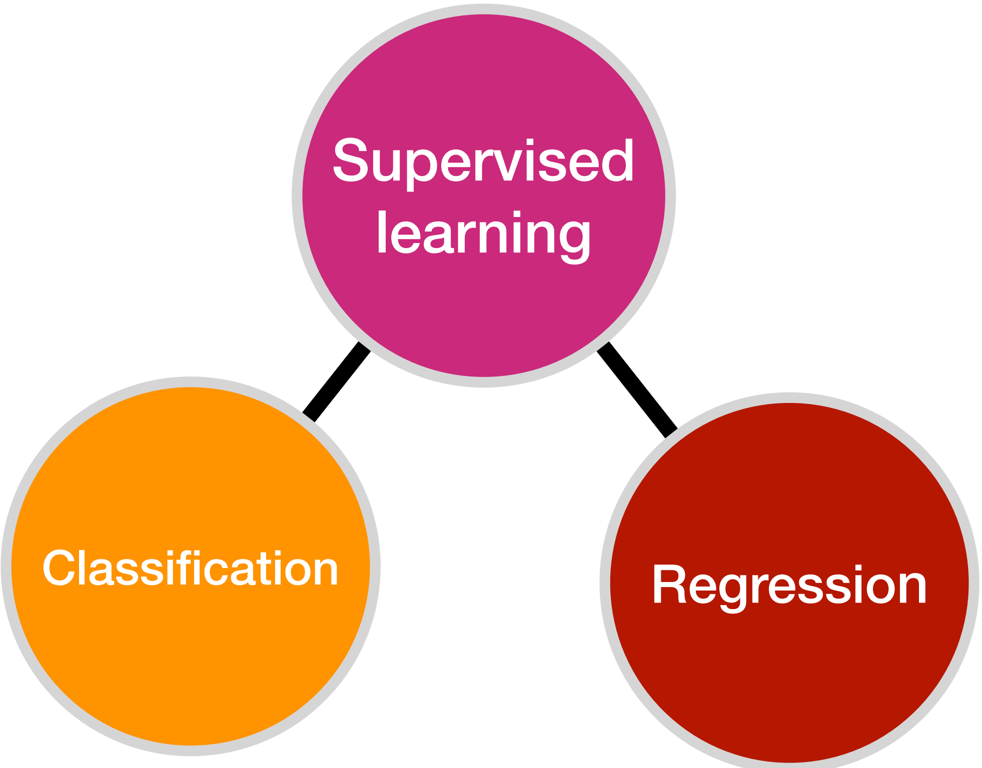
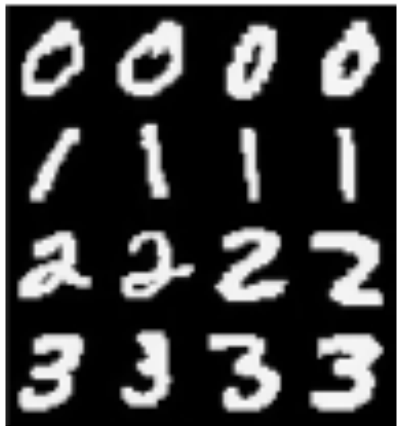
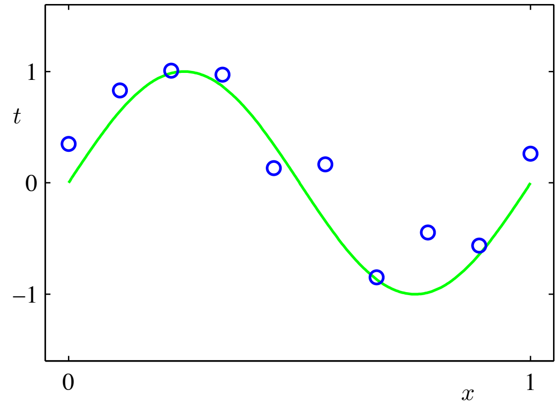
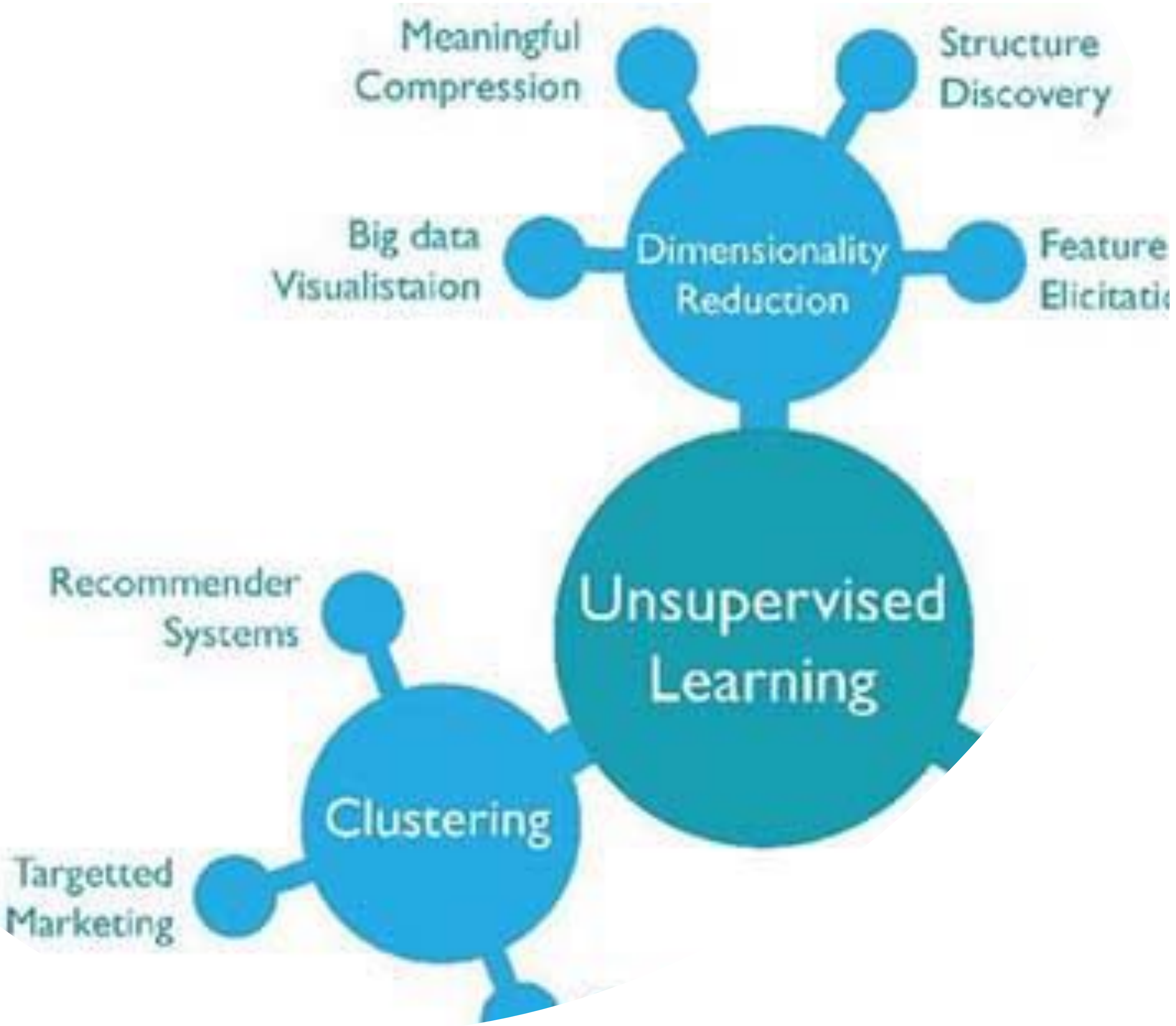


Image source : www.techleer.com

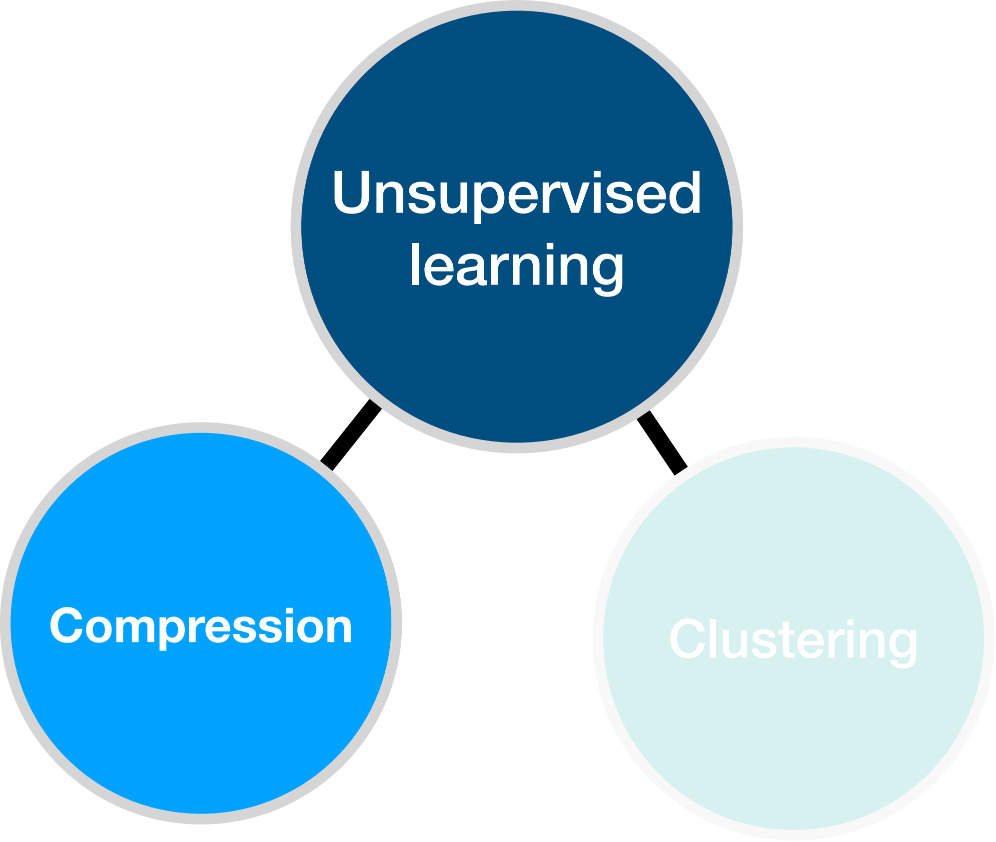
Supervised learning



Dataset		
features: $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$	**x =** 	**x = 0.25**
targets: $\{t_1, \ldots, t_N\}$	**t = 2**	**t = 0.707**


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

Unsupervised learning

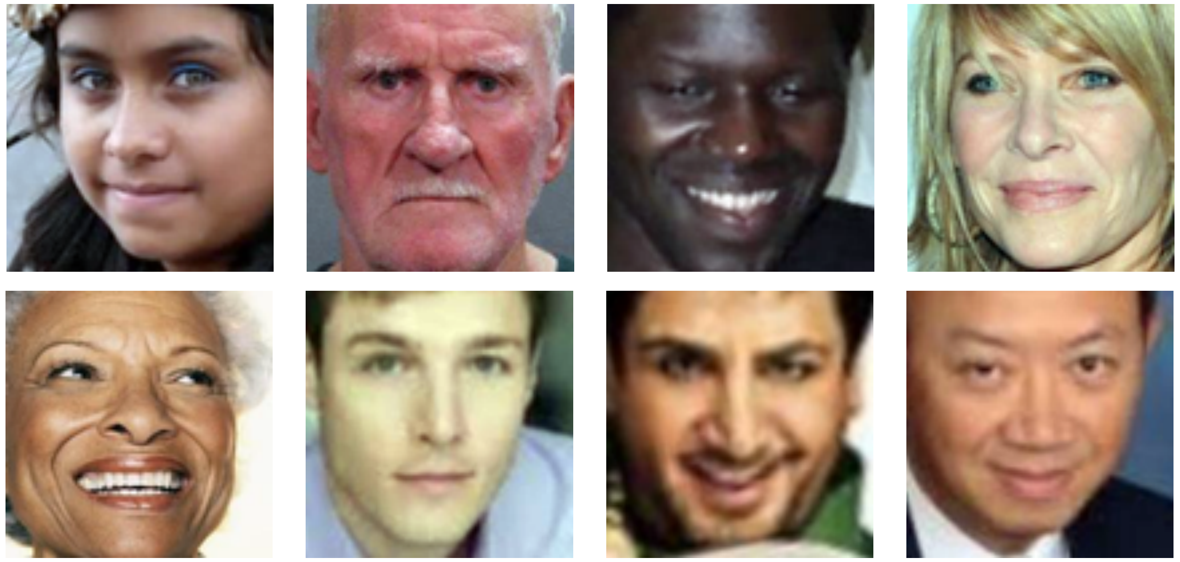


Unsupervised learning

Compression

Clustering

Dataset:



...

Task: Compression

Why? Reduce bandwidth
file size
representation learning

Unsupervised learning

Dataset:

Task: Compress image

Method: Expand along principle components (PCA)

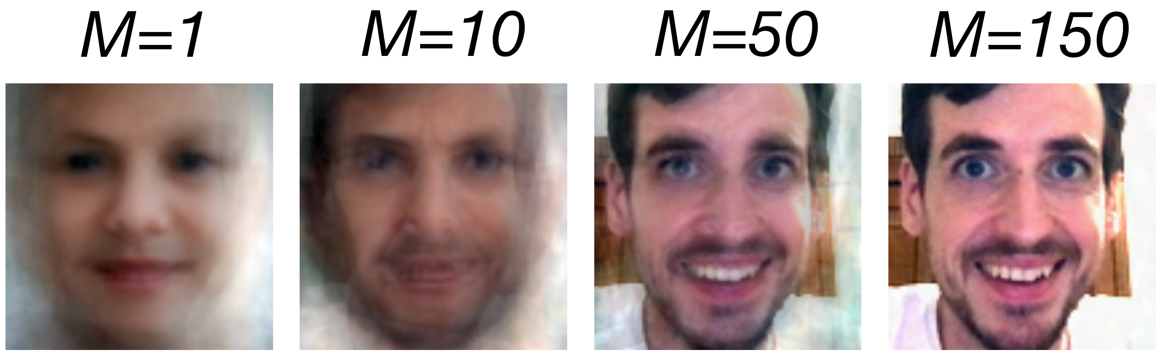
Week 5

features
eigenfaces

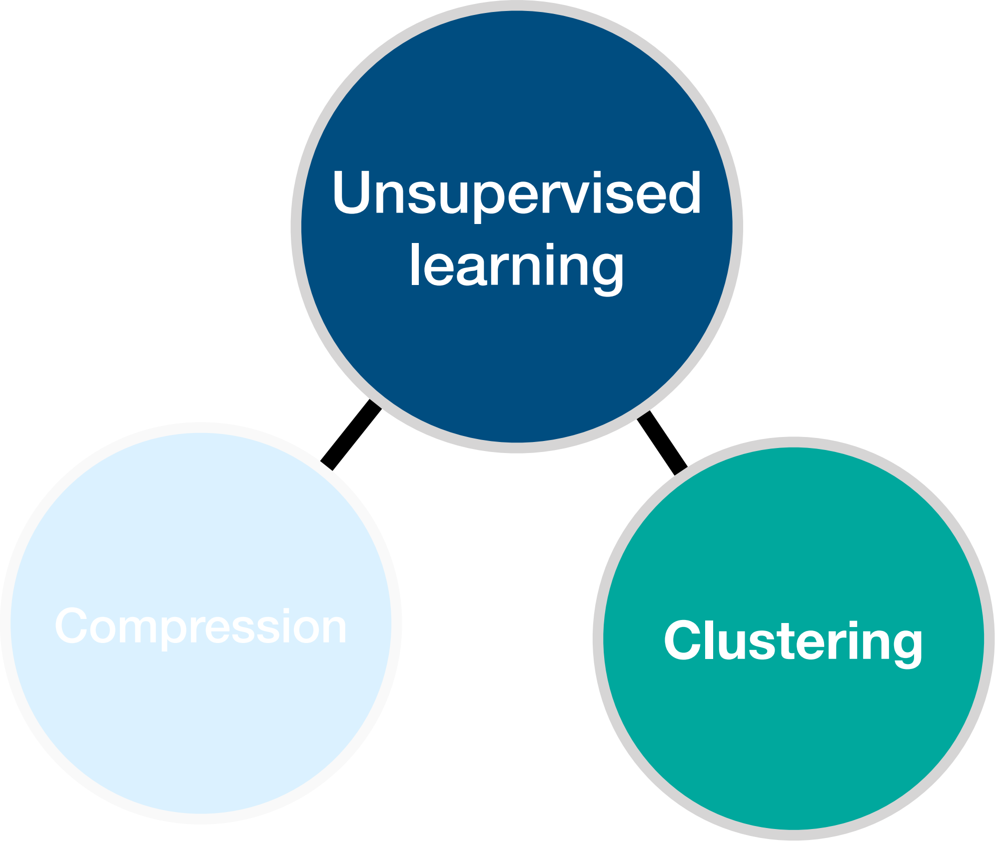
Result:

Original

$$\approx \sum_{i=1}^{M} \alpha_i \mu_i$$



Unsupervised learning

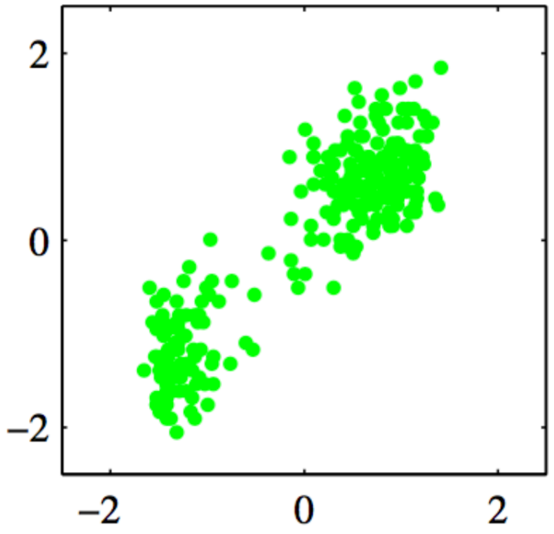


Unsupervised learning

Compression

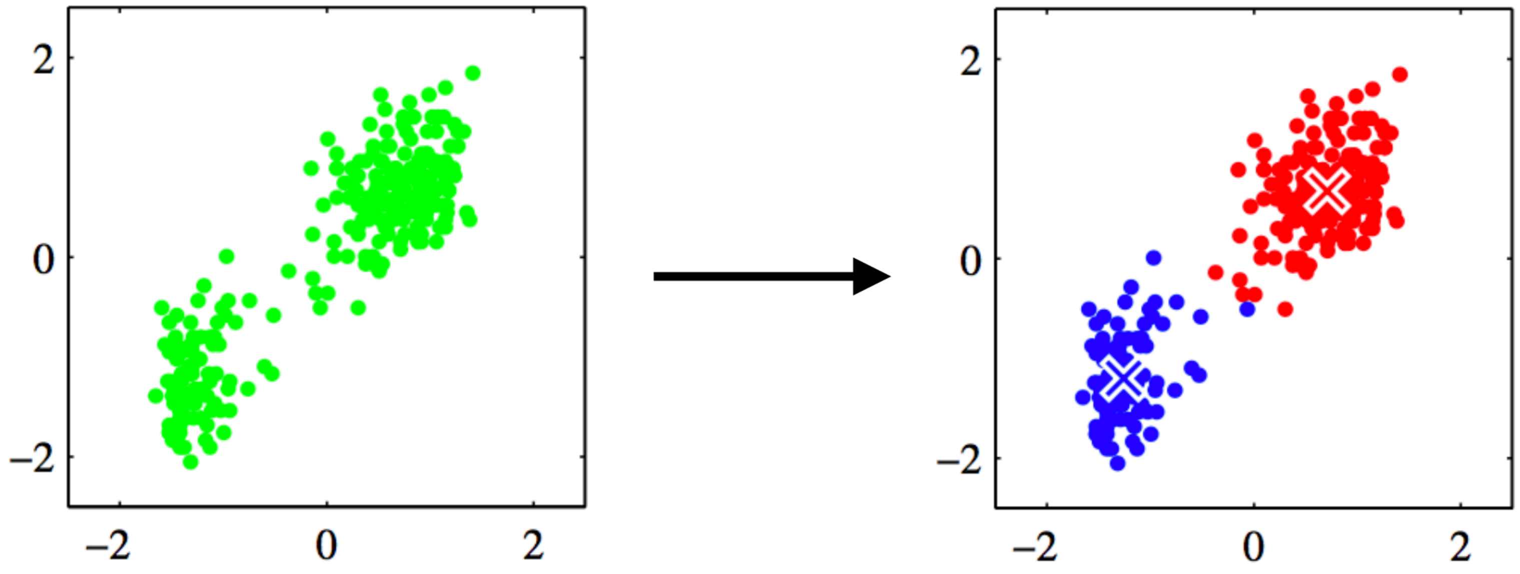
Clustering

Dataset:



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

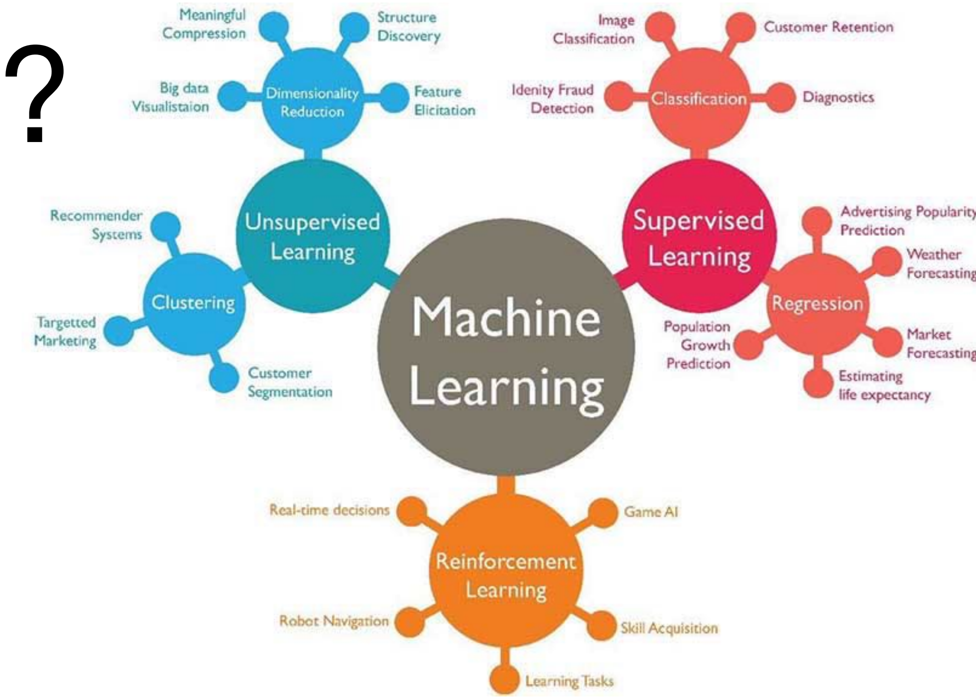
Result:



Dataset

Final clustering

What is machine learning?

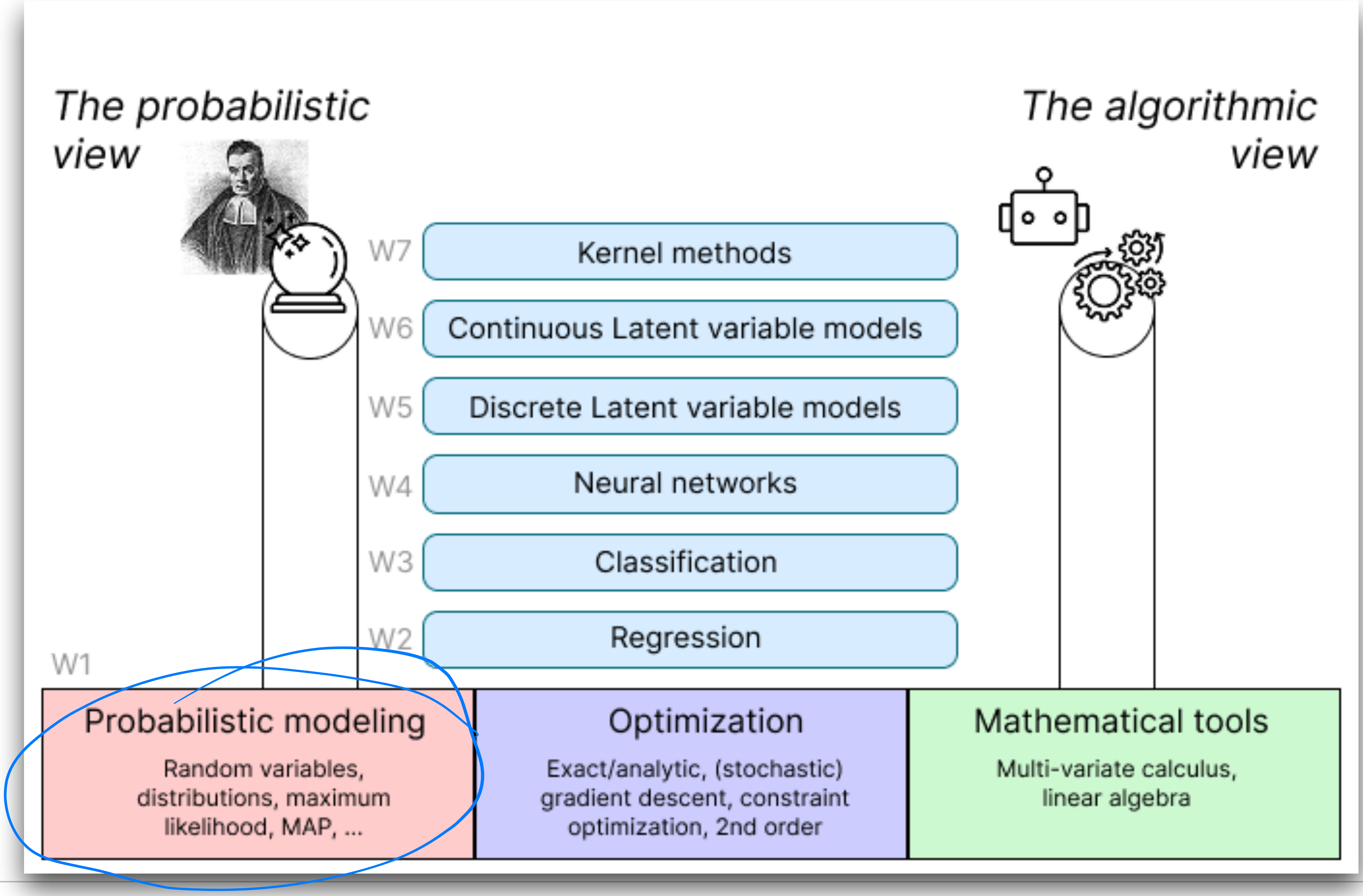


"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

- Tom M. Mitchell

Machine Learning, Tom Mitchell, McGraw Hill, 1997

Finally… What's this course about?



The probabilistic view

The algorithmic view

W7 Kernel methods

W6 Continuous Latent variable models

W5 Discrete Latent variable models

W4 Neural networks

W3 Classification

W2 Regression

W1

Probabilistic modeling

Random variables, distributions, maximum likelihood, MAP, …

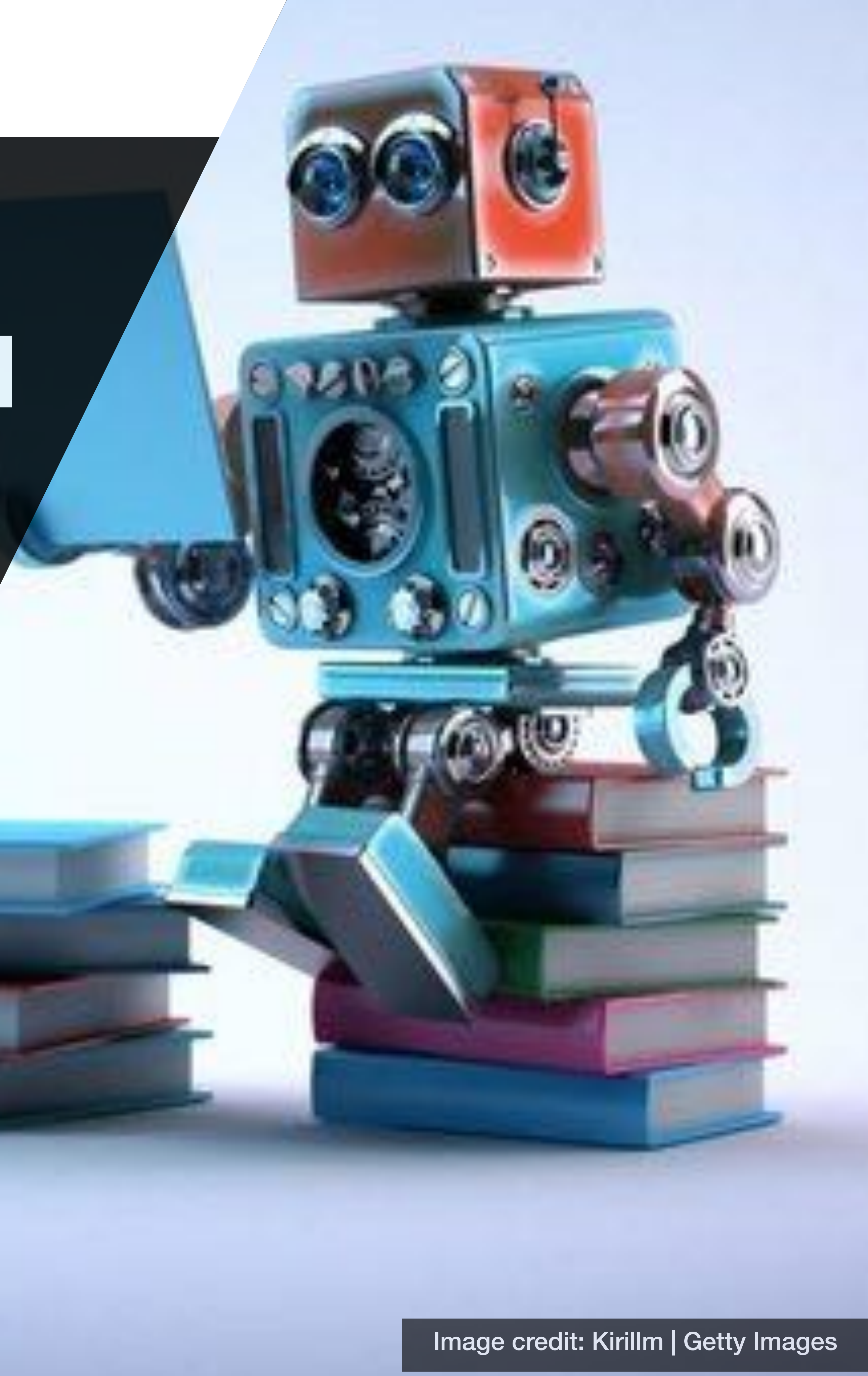
Optimization

Exact/analytic, (stochastic) gradient descent, constraint optimization, 2nd order

Mathematical tools

Multi-variate calculus, linear algebra

University of Amsterdam
Informatics Institute



Machine Learning 1

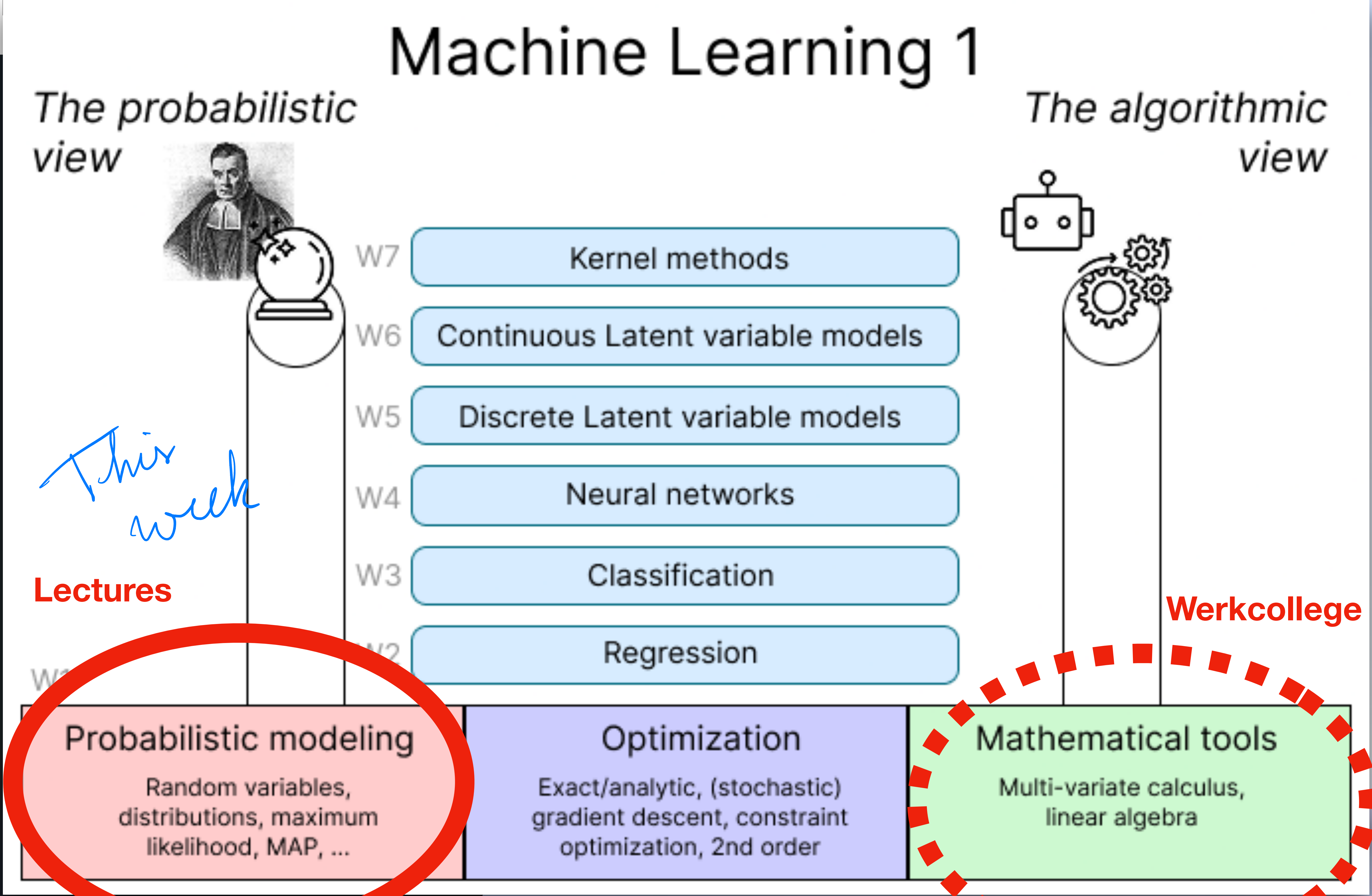
Lecture 1.4 - Probability Theory - Bayes Theorem

Erik Bekkers

(Bishop 1.2.0 - 1.2.1)

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images



Machine Learning 1

The probabilistic view

The algorithmic view

W7 Kernel methods

W6 Continuous Latent variable models

W5 Discrete Latent variable models

W4 Neural networks

W3 Classification

W2 Regression

W1

Probabilistic modeling	Optimization	Mathematical tools
Random variables, distributions, maximum likelihood, MAP, ...	Exact/analytic, (stochastic) gradient descent, constraint optimization, 2nd order	Multi-variate calculus, linear algebra

This week

Lectures

Werkcollege

Slide credits: Patrick Forré and Rianne van den Berg

Image credit: Kirillm | Getty Images

Probability theory

Probability theory (Bishop)

“Provides a consistent framework for the quantification and manipulation of uncertainty”

Uncertainty in pattern recognition

- Noise on measurements. Aleatoric uncertainty (irreducible)
- Finite size datasets. Epistemic (lack of knowledge) (reducible)

Probability theory

Probability theory (Pierre-Simon Laplace, 1819)

“Probability theory is nothing but common sense reduced to calculation”

We typically write

$$p(x)$$

instead of $p(X=x)$ for simplicity. However if there are multiple variables involved it is good to be explicit and use the full notation

Continuous Random Variables

Cumulative distribution

$$P(x) = p(X \leq x)$$

$$= \int_{-\infty}^{x} p(\tilde{x})\, d\tilde{x}$$

$$= \int_{-\infty}^{x} p(y)\, dy$$

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

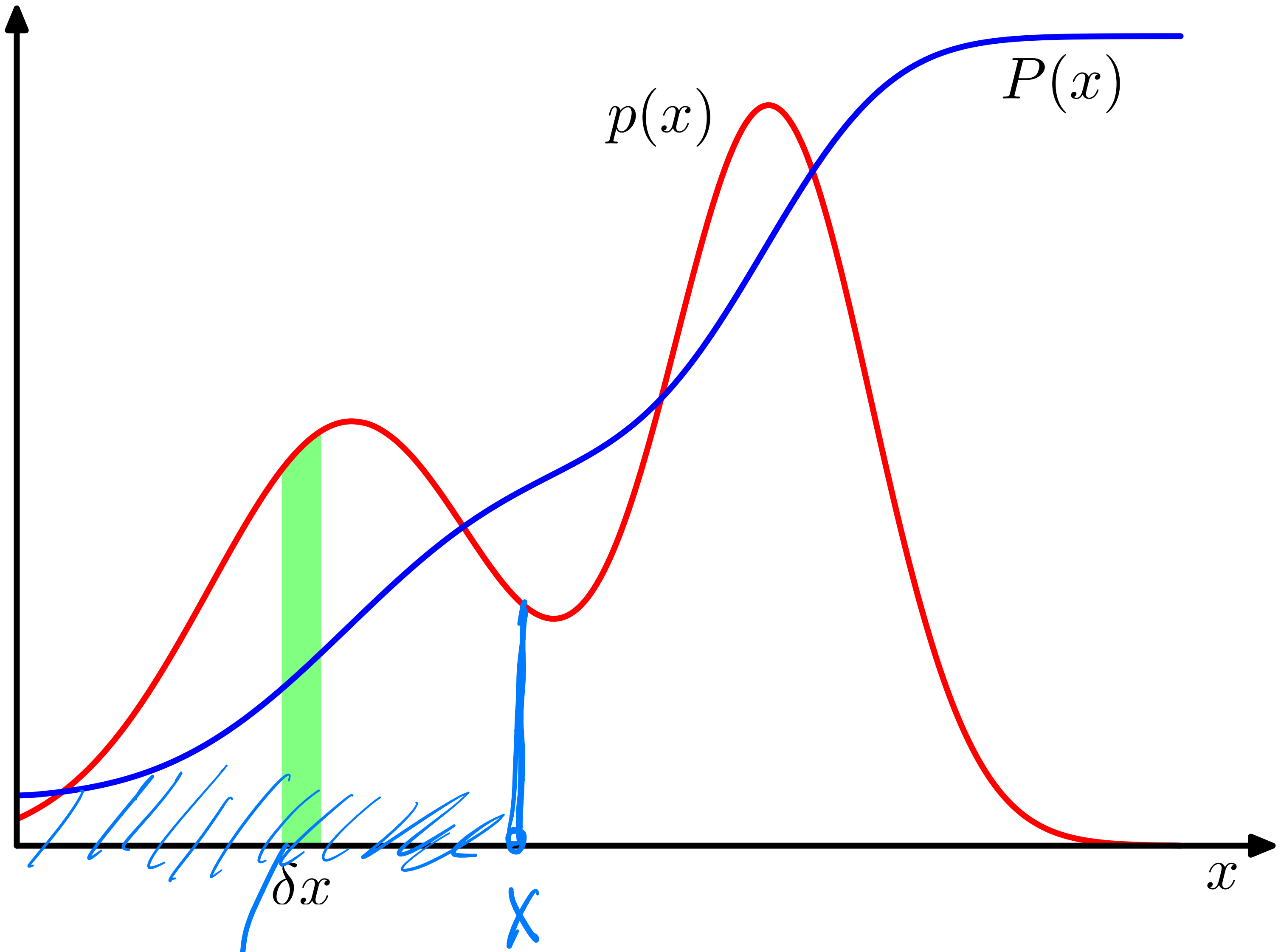


Figure: probability density and cumulative distribution function (Bishop 1.12)

$X \leq x$

The Rules of Probability Theory

For random variables $x \in X$ and $y \in Y$:

	Discrete	Continous
Additivity	$p(x \in A) = \sum_{x \in A} p(X = x)$	
Positivity	$p(X = x) \geq 0$	$p(X = x) \geq 0$
Normalization		$\int_X p(x) \mathrm{d}x = 1$
Sum Rule	$p(X) = \sum_Y p(X, Y)$	
Product Rule	$p(X, Y) = p(X \mid Y) p(Y)$	$p(x, y) = p(x \mid y) p(y)$

The Rules of Probability Theory

For random variables $x \in X$ and $y \in Y$:

	Discrete	Continous
Additivity	$p(x \in A) = \sum_{x \in A} p(X = x)$	$p(x \in (a,b)) = \int_a^b p(x)\mathrm{d}x$
Positivity	$p(X = x) \geq 0$	$p(X = x) \geq 0$
Normalization		$\int_X p(x)\mathrm{d}x = 1$
Sum Rule	$p(X) = \sum_Y p(X, Y)$	
Product Rule	$p(X, Y) = p(X \mid Y)p(Y)$	$p(x, y) = p(x \mid y)p(y)$

The Rules of Probability Theory

For random variables $x \in X$ and $y \in Y$:

	Discrete	Continous
Additivity	$p(x \in A) = \sum_{x \in A} p(X = x)$	$p(x \in (a, b)) = \int_a^b p(x) \mathrm{d}x$
Positivity	$p(X = x) \geq 0$	$p(X = x) \geq 0$
Normalization	$\sum_X p(X) = 1$	$\int_X p(x) \mathrm{d}x = 1$
Sum Rule	$p(X) = \sum_Y p(X, Y)$	
Product Rule	$p(X, Y) = p(X \mid Y) p(Y)$	$p(x, y) = p(x \mid y) p(y)$

The Rules of Probability Theory

For random variables $x \in X$ **and** $y \in Y$**:**

You should know these rules!!!

	Discrete	Continous
Additivity	$p(x \in A) = \sum_{x \in A} p(X = x)$	$p(x \in (a, b)) = \int_a^b p(x)\mathrm{d}x$
Positivity	$p(X = x) \geq 0$	$p(X = x) \geq 0$
Normalization	$\sum_X p(X) = 1$	$\int_X p(x)\mathrm{d}x = 1$
Sum Rule	$p(X) = \sum_Y p(X, Y)$	$p(x) = \int_Y p(x, y)\mathrm{d}y$
Product Rule	$p(X, Y) = p(X \mid Y)p(Y)$	$p(x, y) = p(x \mid y)p(y)$