

# Data Science Introduction

Stéphane Marchand-Maillet

Department of Computer Science



UNIVERSITÉ  
DE GENÈVE  
FACULTÉ DES SCIENCES



Master en Sciences Informatiques - Autumn semester

# What is the Data Science course about?

- ★ Study the modelling of phenomena into digital (numerical, quantitative) data
- ★ Understand the **geometrical and statistical properties** of this data
- ★ Understand the geometrical and statistical properties of the spaces this data lives into
- ★ Analyse the data and develop tools for this analysis
- ★ Understand the assumptions made in the design of these tools
- ★ Work out the theory (in depth)

⇒ involved in linear algebra, probability and statistics

# Course content

## Introduction

## Part I: Data Analysis (SMM)

- ★ High-dimensional representation spaces
- ★ Component analysis: PCA
- ★ Component analysis: FCA
- ★ Component analysis: LDA
- ★ Density estimation: K-means
- ★ Density estimation: Gaussian Mixture Models and the EM algorithm
- ★ Temporal Data Analysis: Autoregressive models
- ★ Temporal Data Analysis: Markov models

## Part II: Information Processing (Prof. S. Voloshynovskiy)

# Relationship to Machine Learning

- ★ Data Science and Machine Learning are synonyms
- ★ Data Science is the study of representation spaces within which Machine Learning acts
- ★ ...

# Required Background (BSc)

## ⚠ Mathematical formalism

### Linear Algebra

- ★ Vector space, inner product, matrix computation
- ★ Projection, eigensystems, SVD, properties
- ★ Optimisation, Gradient Descent, Lagrangian multipliers
- ★ Hyperplane representation, homogenisation of coordinates
- ★ ...

### Statistics and probabilities

- ★ Random variables, expectation, variance
- ★ Probability density function, CDF, entropy
- ★ Joint and conditional probabilities, Bayes theorem
- ★ ...

# Notation

These notations should be consistent throughout the slides:

- ★  $\triangle$  : critical,  $\textcircled{\varphi}$  : redo this computation/proof
- ★ “<sup>def</sup>” definition,  $\llbracket N \rrbracket = \{1, \dots, N\}$
- ★  $x$  scalar,  $\mathbf{x}$  vector,  $\mathbf{X}$  matrix,  $\mathcal{X}$  set,  $X$  random variable
- ★  $\mathbf{x}[i]$  component  $i$  of vector  $\mathbf{x}$
- ★  $\mathbb{F}$  family,  $\mathcal{F}$  subspace
- ★  $\mathbf{1}_N$   $\mathbf{1}$  vector/matrix full of 1's,  $\text{Id}_N$  identity
- ★  $\mathbf{x}^\top \mathbf{X}^\top$  transpose (dual),  $\mathbf{A}^+$  Moore-Penrose inverse
- ★  $\mathbb{P}(X = \mathbf{x})$  probability,  $\mathbb{E}X$  expectation,  $\text{Var}(X)$  variance
- ★  $\langle \mathbf{x}, \mathbf{y} \rangle$  inner product
- ★  $\|\cdot\|$  norm (default =  $L_2$ ),  $d(\cdot, \cdot)$  distance function (default=Euclidean)

# References I

- [1] Sheldon Axler. *Linear Algebra Done Right*. Springer, 2015.
- [2] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. (available online).
- [3] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020. (available online).
- [4] Anirban Dasgupta. *Probability for Statistics and Machine Learning*. Springer, 2011.
- [5] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2021. (available online).
- [6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, New York, 2 edition, 2001.
- [7] Jonathan S. Golan. *The Linear Algebra a Beginning Graduate Student Ought to Know*. Springer, 2004.
- [8] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Copyright Cambridge University Press, 2003. (available online).
- [9] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, second edition, 2018. (available online).

## References II

- [10] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass., 2013.
- [11] Kevin P. Murphy. *Probabilistic Machine Learning: an Introduction*. MIT Press, 2022. (available online).
- [12] Gilbert Strang. *Linear Algebra and Its Applications*. Brooks/Cole, 4th edition, 2005.
- [13] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.
- [14] Gareth James and Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning (with applications in R)*. Springer, 2013.
- [15] Mohamed J. Zaki and Wagner Meira. *Data Mining and Machine Learning: fundamental concepts and Algorithms*. Cambridge University Press, second edition, 2020. (available online).
- [16] Mohammed Zaki and Wagner Meira. *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press, 2014.



# License



The text of this document and its illustrations are published under the  
Creative Commons BY-NC-SA 4.0 International License.