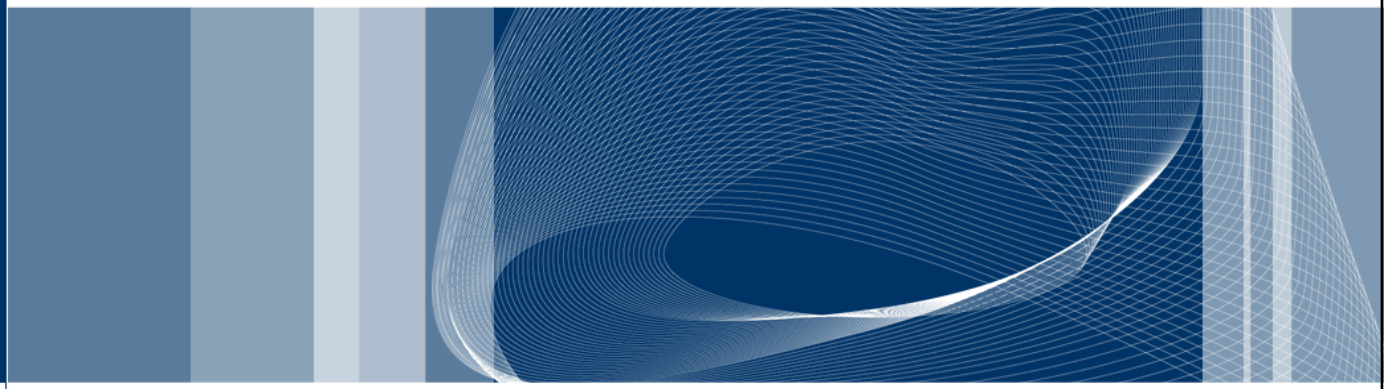
POLITECNICO DI MILANO



Course Introduction

Data Mining and Text Mining

Prof. Pierluca Lanzi

Welcome!

- Pierluca Lanzi
  Dipartimento di Elettronica,
  Informazione e Bioingegneria

- Contacts
  - pierluca.lanzi@polimi.it
  - +39 02 23993472

Lectures on Focused Topics

- Prof. Daniele Loiacono
  Dipartimento di Elettronica,
  Informazione e Bioingegneria

- Fernando Benjamín Pérez Maurera
  PhD Candidate @ POLIMI
  Colab, Numpy & Pandas

- Andrea Lui (Business Integration Partners)
  PhD Executive Candidate @ POLIMI

Course Structure

- Basics
  - What is Data Mining?
  - Data representation (tabular data, text, images, and graph)
  - Data exploration
  - Data preparation
- Data Mining Tasks
  - Clustering
  - Regression
  - Classification
  - Associations
- Advanced techniques and applications
  - Time series
  - Anomaly detection
  - Explainability

Evaluation

- Two types of evaluation available
- Full Written Exam
  - The grade is determined by a written exam (0-33 points)
- Course project (0-9) + Partial written exam (0-24)
  - The grade is the sum of the course project grade (0-9 points) and a partial written exam (0-24)
- Laude is assigned, when the final score is 32 or 33

Course Project

- The course project involves the analysis of a real-world dataset provided by a company
- Data will be available around the end of November.
- Students have three weeks to complete the project
- The project will assign up to 9 points
- The project grade is valid for the whole academic year

Students who submit the project for evaluation cannot drop the grade they will receive. Thus, their exams will consist of the shorter written exam (0-24 points) and their grade will be computed as the sum of the written exam and the project grade.

Course Project – Running Examples

- Example #1: George starts the course project, but he realizes that the work has done so far is not good enough. Thus, he does not submit the project for evaluation. In this case, George will always take the full written exam (0-33).

- Example #2: Ada starts the course project and submits it for evaluation receiving 5 points (out of 9 maximum points). Ada will always take the partial written exam (0-24) and her grade will be the sum of the partial exam grade plus the five points received for the project. Ada cannot redo the project nor can decide to drop the course project grade.

Warning

You must be enrolled in the exam

If you are not enrolled, or you enrolled after the deadline, you will not be able to give the exam

No exception!

Problems & Past Exams

- Problems and exercises will be done throughout the course
- There are no specific days dedicated to exam problems
- All the past exams are available online in the Beep platform
- Some exams have solutions most of them don't
- You are welcome to discuss solutions on the course forum

overlaps

There are overlaps with other courses

Machine Learning

(logistic regression, boosting, k-nearest neighbor, ...)

Statistica Bayesiana, Applied Statistics

(PCA, linear regression/classification, k-means, trees, boosting, forests, ...)

...

Why keeping the overlaps?

The course provides a self-contained toolbox, thus fundamental methods need to be introduced

Different perspective, we won't delve into the underlying theory of existing methods

We study how to use existing methods and (more importantly) how to interpret the results

General Approach

Focus on the data mining process and the most relevant classes of problems (classification, clustering, etc.)

Discuss the most important methods, how they work, what type of results they produce, what are their biases, …

Play with them, understand how to interpret the results, etc.

(if you want to delve deeply into the theory, you should attend Machine Learning, Applied Statistics, etc.)

syllabus

Course Material (Mandatory)

- Course slides, Python notebooks, and KNIME workflows all available on BEEP

- "Data Mining and Analysis: Fundamental Concepts and Algorithms," Mohammed Zaki and Wagner Meira Jr. Cambridge University Press (2nd Edition) 2020. http://www.dataminingbook.info

- "Mining of Massive Datasets Book," by A. Rajaraman, J. Ullman. http://www.mmds.org

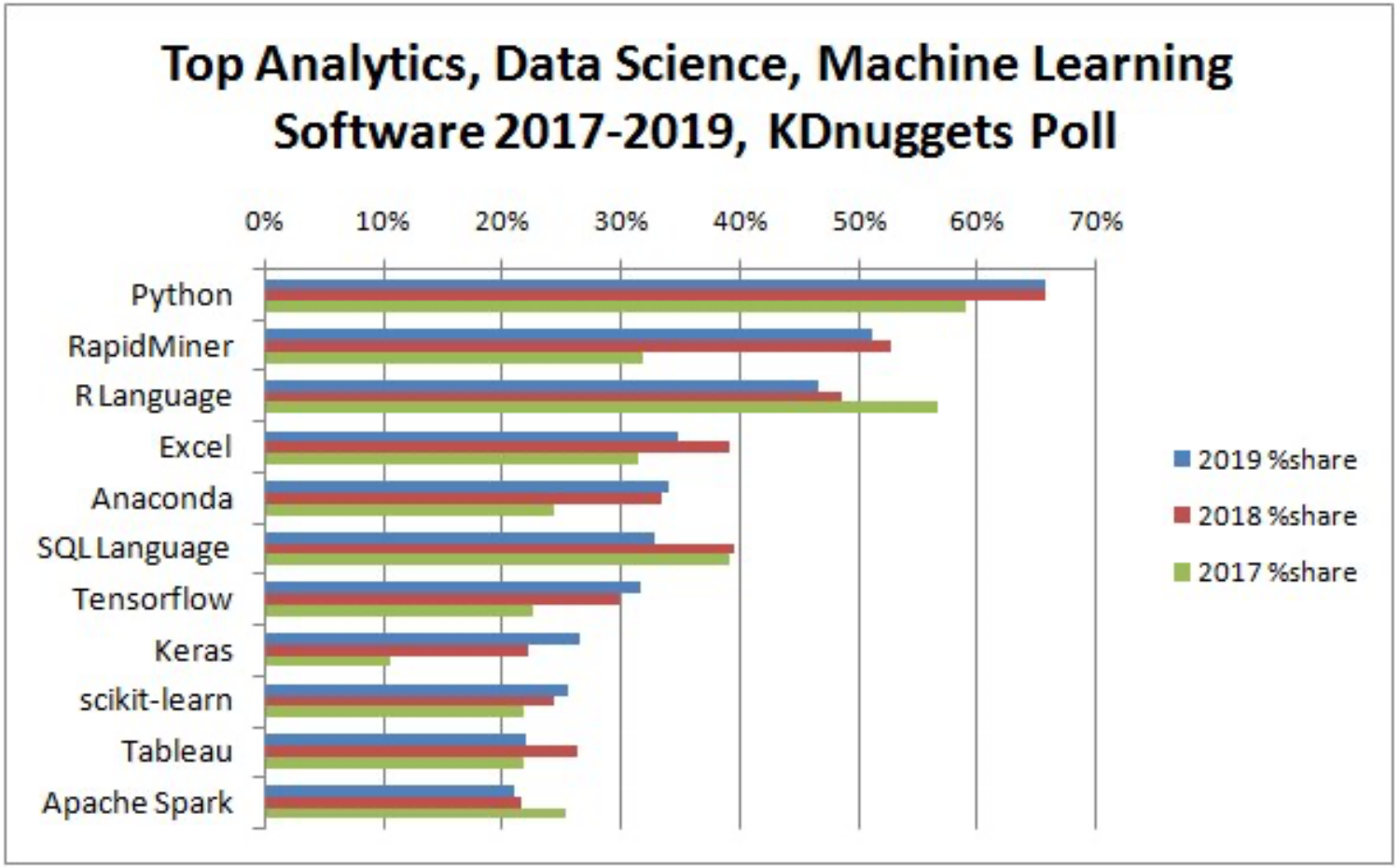
Recordings from 2018/2020 and 2019/2020

- The recordings from previous editions of the course:

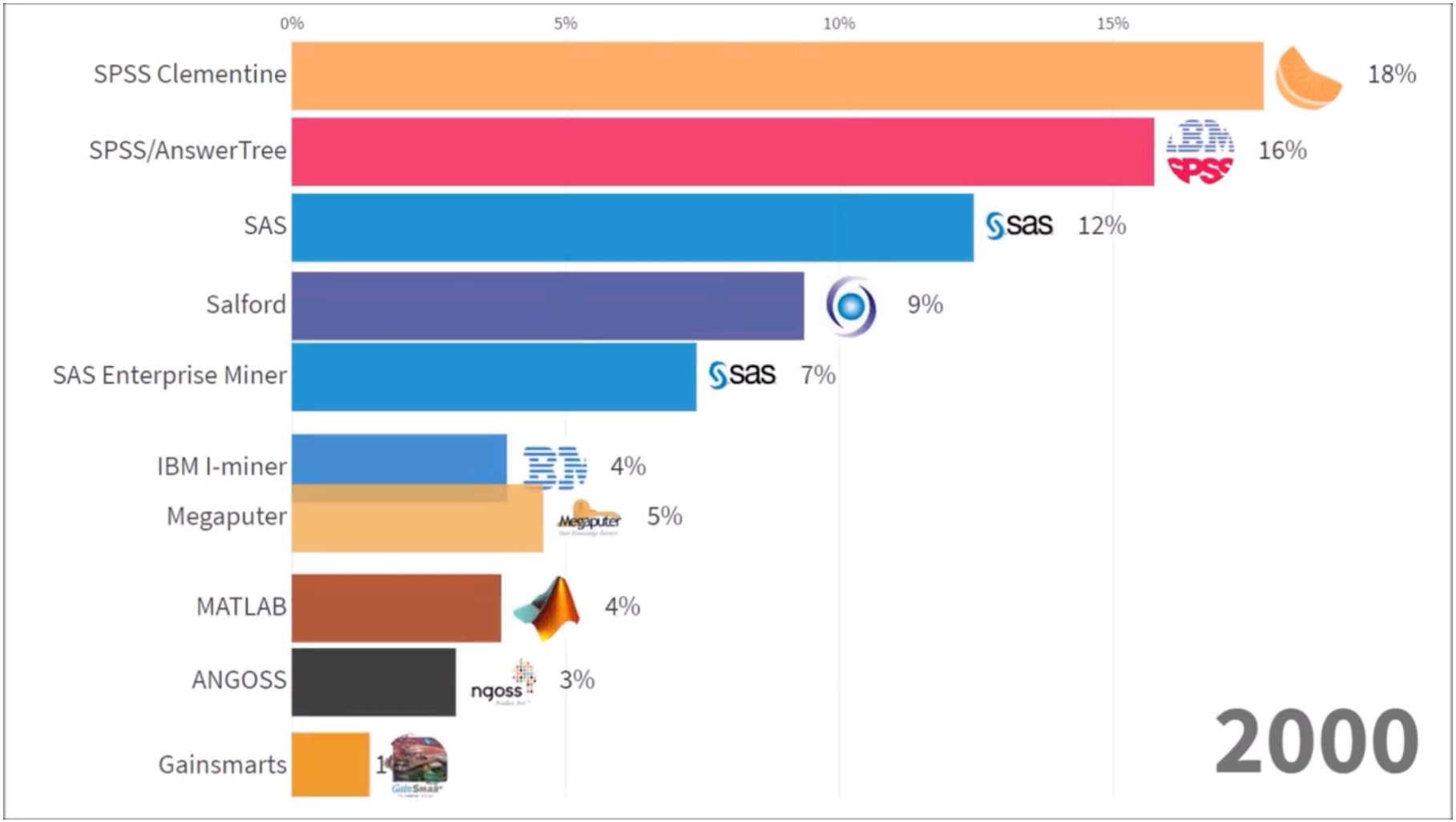
https://www.youtube.com/playlist?list=PLbtqp5GCZsUud268zCfQo7I6LTWVpzkMi
https://www.youtube.com/playlist?list=PL4QF7IP4PZkAPOKnEP6oNM2MW7bA-Br6A

- They are provided as supplement material not as study material for this edition since some new topics will be added and some will be eliminated.

- So use the more recent material (slides, notebooks, etc.)



KDnuggets Analytics/Data Science 2019 Software Poll: top tools in 2019, and their share in the 2017, 2018 polls
https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html



Top 10 Data Science Tools Over Time
https://www.youtube.com/watch?time_continue=109&v=pKPaHH7hnv8&feature=emb_logo

During the course we make extensive use of
Python notebooks and KNIME workflows

These will be available on the course website

Python notebooks and KNIME workflows
are a fundamental part of the syllabus

Some topics will be only discussed in Python
notebooks and using the KNIME workflows

Exams will involve questions on topics
discussed in the notebooks and workflows

You are required to learn to understand
Python code and KNIME workflows

How to install Anaconda on Windows, Mac & Linux

https://docs.anaconda.com/anaconda/install/

Google Colab

https://colab.research.google.com

How to install KNIME (Windows, Mac & Linux)

https://www.youtube.com/watch?v=yeHbIDxakLk

https://www.youtube.com/watch?v=1jvRWryJ220

https://www.youtube.com/watch?v=wibggQYr4ZA

EVERYTHING I SAY

WILL BE ON THE EXAM

questions?