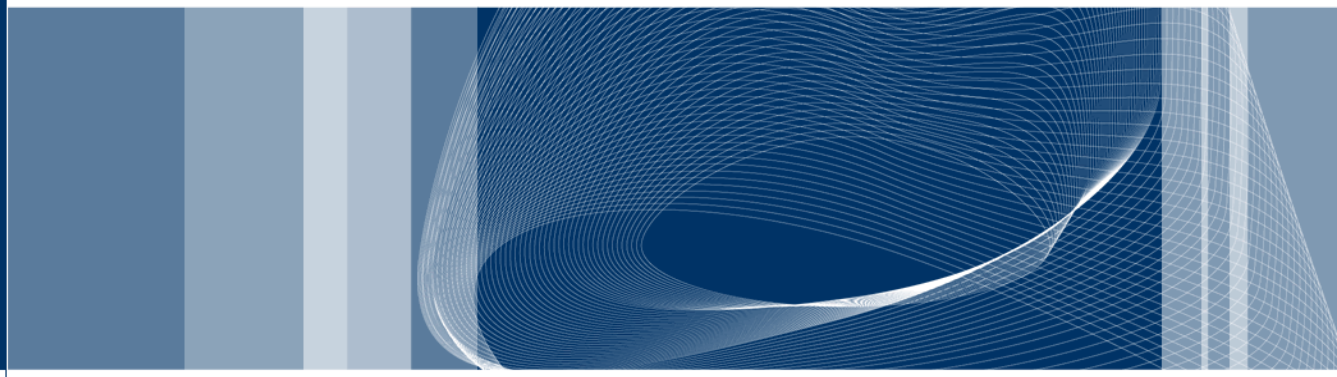
POLITECNICO DI MILANO



Course Introduction

Data Science for Mobility

Prof. Pier Luca Lanzi

Welcome!

- Pier Luca Lanzi
  Dipartimento di Elettronica, Informazione e Bioingegneria

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

Course Structure

The data science process

The main steps, the data representation, etc.

Analysis of the most relevant classes of problems

Regression, Classification, Clustering, etc.

Analysis of the most important algorithms

How they work? What type of results they produce?
What are their biases?

Application to Example Problems

Experiments with discussed algorithms on real data
taken from Kaggle and Open Data initiatives

Topics

- Basics
  - What is Data Science?
  - Data and knowledge representation
  - Data exploration and preparation
- Fundamental Techniques
  - Regression
  - Classification
  - Clustering
- Advanced Techniques and Applications
  - Deep Neural Networks
  - Time Series
  - Text Mining
  - Graph Mining

Course Calendar

13 Lectures combining theory and application

Lecture Structure

2 hours + break + 2 hours

Course Material

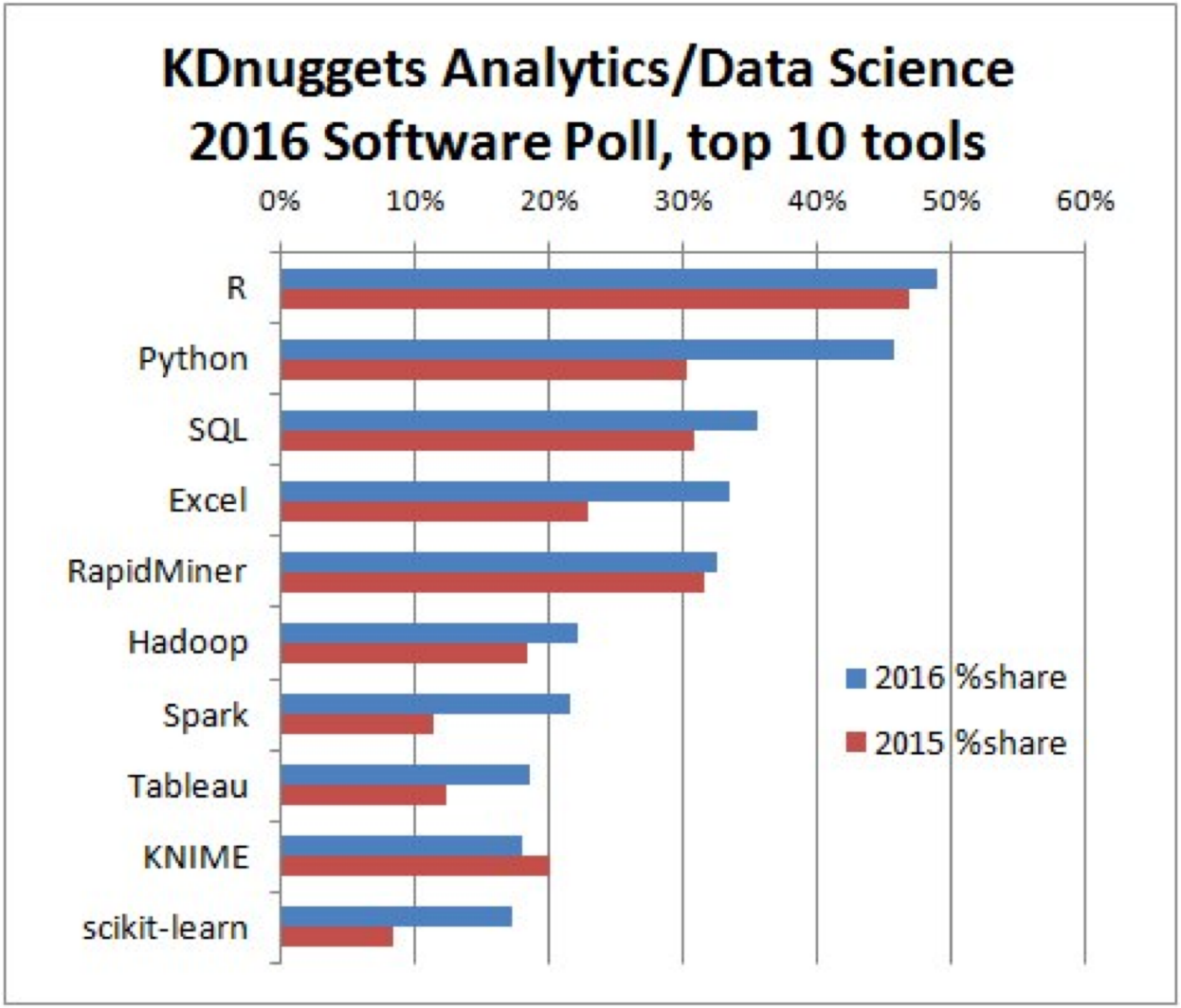
Course Material (Mandatory)

- 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 in 2014. http://www.dataminingbook.info

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

What Tools?



http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Tool	% change	2016 %share	2015 %share
Dato	377%	2.4%	0.5%
Dataiku	292%	7.8%	2.0%
MLlib	253%	11.6%	3.3%
H2O	233%	6.7%	2.0%
Amazon Machine Learning	171%	1.9%	0.7%
scikit-learn	107%	17.2%	8.3%
IBM Watson	99%	4.2%	2.1%
Splunk/ Hunk	98%	2.2%	1.1%
Spark	91%	21.6%	11.3%
Scala	79%	6.2%	3.5%

Tools with the highest growth

http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Tool	2016 %Share	2015 %share	% change
Hadoop	22.1%	18.4%	+20.5%
Spark	21.6%	11.3%	+91%
Hive	12.4%	10.2%	+21.3%
MLlib	11.6%	3.3%	+253%
SQL on Hadoop tools	7.3%	7.2%	+1.6%
H2O	6.7%	2.0%	+234%
HBase	5.5%	4.6%	+18.6%
Apache Pig	4.6%	5.4%	-16.1%
Apache Mahout	2.6%	2.8%	-7.2%
Dato	2.4%	0.5%	+338%
Datameer	0.4%	0.9%	-52.3%
Other Hadoop/HDFS-based tools	4.9%	4.5%	+7.5%

Big Data tools and their share in 2016, 2015, and %change
http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

But I don’t Know Python! ☺

I activated a class for this course on DataCamp

You will be invited to join the class and receive access to premium content

You will find a set of assignments that you will have to complete before the course ends ☺

These include basic Python programming and some data science related courses

Some assignments are mandatory and will be graded

You are welcome to access other courses you might like!

questions?