Machine Learning for Big Data

Course code: MLC_BDATA

The aim of this course is to present an overview of tools and concepts from machine learning on big data. After going through the course, participants should be able to tell what the right tool is to be used for the given problem, whether there is a simpler solution and how to avoid common mistakes. Special attention will be given to Spark as a universal tool that can be used for both big data processing and machine learning.

Required skills

- Basics of Python and working in Google Colab
- Basics of machine learning on the level of our course Introduction to machine Learning

Course outline

- Overview of Big Data concepts and tools
- From small to big data and estimating its value
- Row vs column-oriented database
- HDFS (Hadoop Distributed File System)
- Big data file formats – Parquet, ORC, Avro
- Compression – gzip, snappy, zstd
- SQL databases – BigQuery, Redshift, Clickhouse, Snowflake, Vertica
- A practical example of a big data value proposition
- Introduction to Spark
- MapReduce
- Spark Computing Engine and RDDs (Resilient Distributed Datasets)
- DataFrames
- Spark Ecosystem
- Most common Spark mistakes
- How to run Spark
- Alternatives – Apache Beam (Dataflow), Dask, lambdas
- A practical example with Spark
- ML strategies for Big Data
- Incremental learning
- Batch learning for neural networks
- Distributed training
- Federated learning
- Alternative strategies
- Random sampling
- Submodels
- Larger workstation
- Frameworks
- Scikit-learn with partial_fit
- MLlib
- Dask-ML
- Practical examples with various frameworks

GOPAS Praha
Kodaňská 1441/46
101 00 Praha 10
Tel.: +420 234 064 900-3
info@gopas.cz

GOPAS Brno
Nové sady 996/25
602 00 Brno
Tel.: +420 542 422 111
info@gopas.cz

GOPAS Bratislava
Dr. Vladimíra Clementisa 10
Bratislava, 821 02
Tel.: +421 248 282 701-2
info@gopas.sk