BIG DATA

ONLINE TRAINING

Data Analysis and Data Preparation for Machine Learning

eit Manufacturing CLC EAST

Co-funded by the European Union

TU WIEN

EURO AUSTRIA

DATES AND DURATION	03 May 2023 10:00 - 16:00 CEST
FORMAT	Online
LEVEL	Beginner
OFFICIAL CERTIFICATION	Certificate of attendance
FEE	100,00 EUR + VAT
PARTICIPANT PROFILE	All engineering disciplines with some programming experience (ideally with Python) Professionals who would like to make use of the tons of data that are being collected

MORE INFORMATION AND REGISTRATION

DESCRIPTION

This one-day course aims at professionals from all domains, who would like to get a handle on their data.
The course shows the participants how to get a sense of the look and feel of their data, how to visualize it and clean it up where necessary. In addition, it shows the participants how to get the data in a suitable shape before feeding it into Machine Learning (ML) algorithms further down the line. However, this course will not teach ML methods, as there are follow-up courses for these topics.

The programming language of choice is Python, so the participants will get to know libraries such as NumPy, Pandas, Scikit-Learn, Matplotlib and Dask. The content is delivered with Jupyter notebooks on Google Colab, so to be able to participate fully, participants should have a Google account.

CURRICULUM

Participants will learn why data needs to be pre-processed before being passed to ML methods. They will also learn what the typical challenges are in data wrangling.

Pandas:
Participants get to know this powerful Python library and find out how they can load data into a data frame, get the look and feel of it and transform it in the best suitable way.

NumPy:
ML would simply not be possible in Python without this useful library for numerical operations. This is why participants will get to know the most important aspects of the API and what can be achieved with it.

Matplotlib:
Humans are visual beings and this is why we prefer looking at graphs, rather than endless tables of data. Matplotlib is the Python library to create all kinds of graphs which helps understand data a great deal more. Participants will learn how to create the most common graphs within Matplotlib.

Dask:
In ML problems, we often get to a situation where our data does not fit into memory. Even if it fits into memory, we would like some operations to run faster. Dask solves this problem by dividing our data into smaller, more manageable chunks. It then runs computations on those chunks in parallel, making it possible to handle data that is larger than memory. It is also faster since it makes computations run concurrently. Participants will get to know this tool and see the similarities with previously learned libraries.

LEARNING OUTCOMES

At the end of the course, participants will be able to:

- Get data into a suitable form
- Visualize data
- Clean data
- Transform data
- Analyse data
- Handle data that does not fit in memory

SIMEON HARRISON

EUROCC AUSTRIA

Simeon works at the EuroCC Austria at the TU Wien, the Austrian national competence centre for high-performance computing, high-performance data, and artificial intelligence (AI). He has a background in mechanical engineering and is passionate about teaching in the ground-breaking AI area. Before joining EuroCC Austria, he was teaching high school maths for 8 years.

EIT Manufacturing is a Knowledge and Innovation Community that connects the leading manufacturing actors in Europe. It is supported by the European Institute of Innovation & Technology (EIT), a body of the EU. Fueled by a strong interdisciplinary and trusted community, EIT Manufacturing adds unique value to European products, processes, services – and inspires the creation of globally competitive and sustainable manufacturing.

EIT Manufacturing's headquarters are in Paris, with a network of innovation hubs across Europe: Austria, Germany, Greece, Italy, Spain and Sweden

eitmanufacturing-east.eu