Data Science Training
Getting Started with Python & NumPy

This training was developed for you by Data Science @ SE GP G QPI QM P Improvement Projects

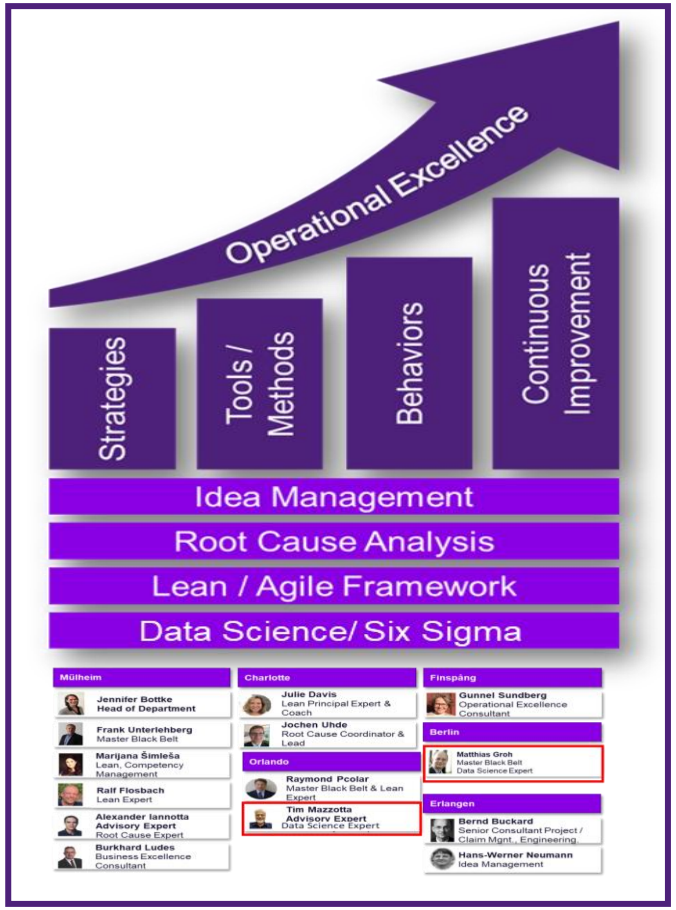
We help your employees getting the skills for the digital future:

- Global Data Science Skill Network (with Matt Bryan, LGT R&D)
- Data Science and Machine Learning trainings (with several partners):
    - Currently 11 trainings developed and executed
    - Attended by >500 participants mostly from R&D
- Analytic services in specific improvement projects
- EAI CodeShare (with Thomas Buller, LGN R&D)

Data Science Lead Developers @ QPI QM P:

Matthias Groh
Six Sigma Master Black Belt
Data Science Consultant

Tim Mazzotta
Advisory Expert
Data Science Expert



Mülheim

Jennifer Bottke
Head of Department

Frank Unterlehberg
Master Black Belt

Marijana Šimleša
Lean, Competency Management

Ralf Flosbach
Lean Expert

Alexander Iannotta
Advisory Expert
Root Cause Expert

Burkhard Ludes
Business Excellence Consultant

Charlotte

Julie Davis
Lean Principal Expert & Coach

Jochen Uhde
Root Cause Coordinator & Lead

Orlando

Raymond Pcolar
Master Black Belt & Lean Expert

Tim Mazzotta
Advisory Expert
Data Science Expert

Finspång

Gunnel Sundberg
Operational Excellence Consultant

Berlin

Matthias Groh
Master Black Belt
Data Science Expert

Erlangen

Bernd Buckard
Senior Consultant Project / Claim Mgnt., Engineering.

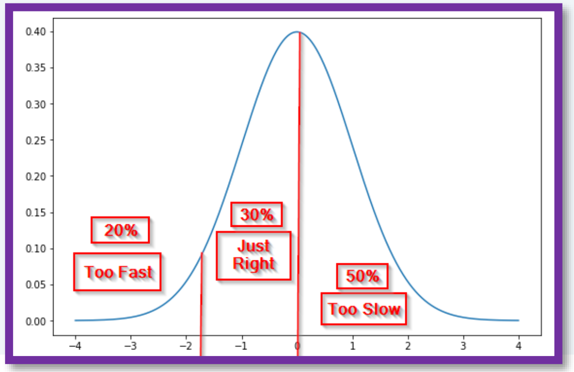
Hans-Werner Neumann
Idea Management

We help building the bridge between Data Science and daily engineering business.

Introduction

1. Name, Department
2. All prerequisites fulfilled? EnergyAI, knowledge of previous trainings (DS0->DS1->DS2->DS3)
3. Preexisting knowledge program, data science, statistics, Python, R, …
4. Expectations

Goal Data Science Workshops: Basic understanding how to use Python and EnergyAI to access power plant data

Workshop Module	Agenda	Goals	Our approach to teaching:
DS 1 Python Basics	• Jupyter Notebook Intro • Lists, Object Localization • Loops, Dictionaries, If • Definitions • Classes/ Objects/ Methods • Intro to NumPy/ ND-Arrays	**Target audience (all modules):** • Any engineer who needs to analyze big data • In particular engineers verifying assumptions with power plant data **Prerequisites:** • Basic programming knowledge or DS0 • DS1 for DS2, DS2 for DS3 or reading through the scripts • New For DS2: "Transforming Data“ = free chapter 1 of "Data Manipulation with Pandas" https://learn.datacamp.com/courses/data-manipulation-with-pandas **Teaching goals (all modules):** • First steps to learn the topic is taken. • Basic concepts are understood ->just enough to keep learning from online sources. • Everybody learns in the workshop, also trainers. • This is a teaching “Minimal Viable Product”	**Handling of questions:** • For understanding questions ask right away • For major problems go to separate helpline Teams meeting: one of the teachers will join you as a coach • Expert discussions moved to break or afterwards **With a wide range of participant’s knowledge:** • Target speed is for 20% too fast / for 50% too slow 
DS2 Python: Pandas and Seaborn	• Recap • Handling Nans • Resample and Fill • Calculation with dataframes • PickleFormat • Connecting dataframes • Intro Visualization with Seaborn		
DS 3 Data Pulling in Energy AI	• Intro Data Pulling • Time Interval • Plants, Units and Signals • Period, Aggregation • Fill, Thresholds, Poststresholds • Plotting with Seaborn (cont.) • Examples		

How This Training Works

- Experiment, we are all learning
- DS0-> DS1->DS2->DS3 are prerequisits (enough to read through scripts)
- New For DS2: "Transforming Data“ = free chapter 1 of "Data Manipulation with Pandas“ https://learn.datacamp.com/courses/data-manipulation-with-pandas
- 1x15 Min Break in each 2h block
- Recording: high demand, few training seats
- Based on interaction
  - Discussions, Exercises + Debriefs
  - Harder in Teams
  - If I should not call on you, please tell us now

Our Focus in the Digitalization World is Data Science

Data Science

- Focus on quantitative data

For Example:

- Power plant sensor data
- Manufacturing data
- Component design analysis data (FE, CFD, CHT, …)
- Test rig data

Business Intelligence

- Focus on qualitative data

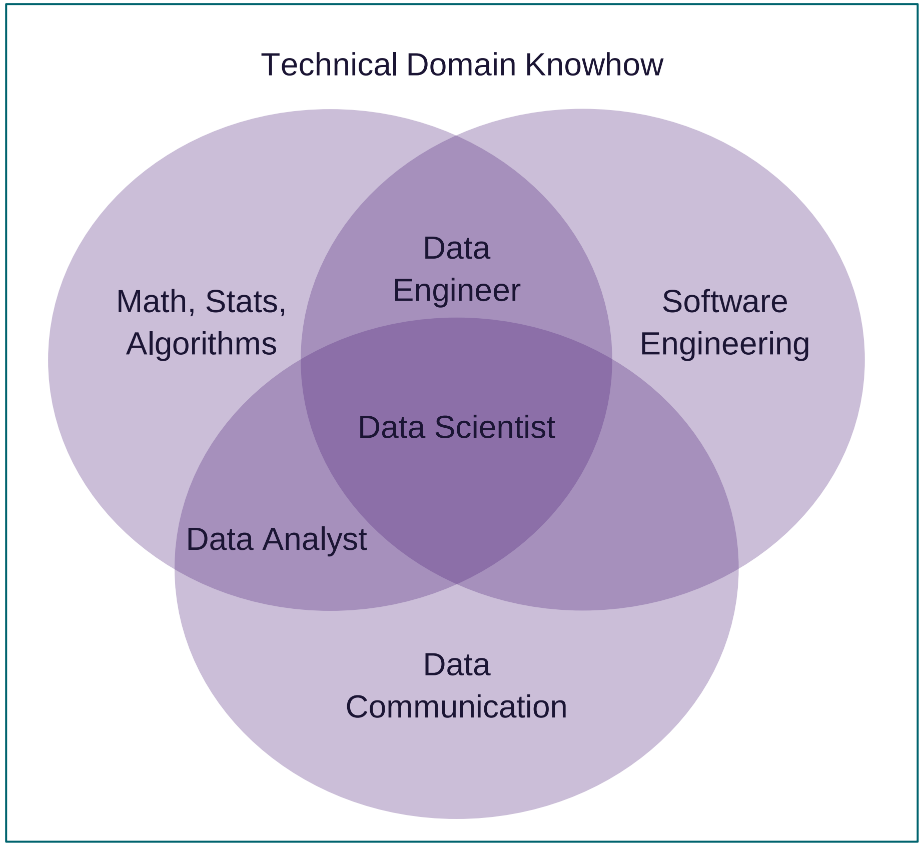
For Example:

- Manufacturing and supply data in SAP
- Qualitative fleet data (Fleet Intelligence)
- PCM

Of course everybody needs a bit of both data types, but without focusing, no one gets anywhere…

Definition Data Science

Data science is a multi-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data. [1]



Data science includes “new” (Machine Learning) and “old” (conventional analytics).

[1] en.wikipedia.org/wiki/Data_science

Basic Elements - Data Structures in Python (Grey), Numpy (Green) and Pandas (Orange)

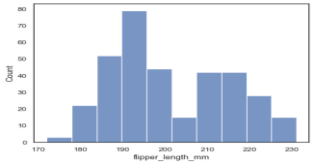
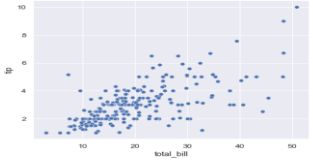
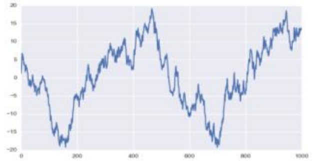
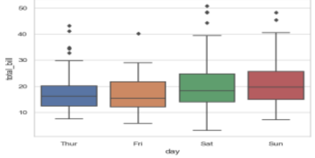
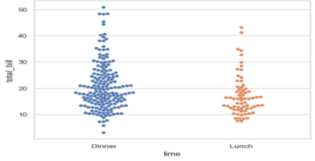
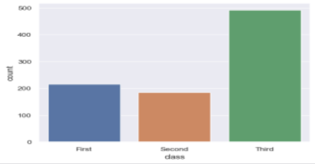
	Description	Useful for…	Example
Lists	• Store collection of heterogeneous items – immutable and mutable objects	• Simple listing • Special programming techniques (Stacks, Queue, Graphs, Trees)	[item, item, item]
Dictionaries	• Collection of key-value pairs • Key: immutable objects • Value: heterogeneous items – immutable and mutable objects	• Creating loops to do the same operation on many dataset	{key1:item1, key2:item2}
Files	• Store and retrieve previously stored information		.csv, .hdf5, .xlsx
Tuples	• Tuples are data structures in which the content can not be changed after creation (no deletion, add or edit)	• Prevents data manipulation	(item, item, item)
Sets	• Collection of unique objects	• Creating lists that only hold unique values • Helpful when going through huge dataset	{item, item, item}
Numpy ND-Arrays	• Store collection of data of the same type	• Dealing with large collection of homogeneous data types: easier to use, faster and uses lesser memory than lists • Support vectorized operations • Work efficiently with large datasets with lots of empty cells	[[1. 1. 2.] [3. 5. 8.] [5. 3. 2.]]
Dataframes	• 2-dimensional labeled data structure • Columns don't have to have the same data type	• Data mining / manipulation • Labeling • Multiindexing	Like this table
Series	• One-dimensional labeled array capable of holding any data types	• Axis labels	a 1 b 2 c 3

Source: www.datacamp.com, pandas.pydata.org; Also see: https://www.datacamp.com/community/tutorials/data-structures-python

A Selection of Important Libraries in Python

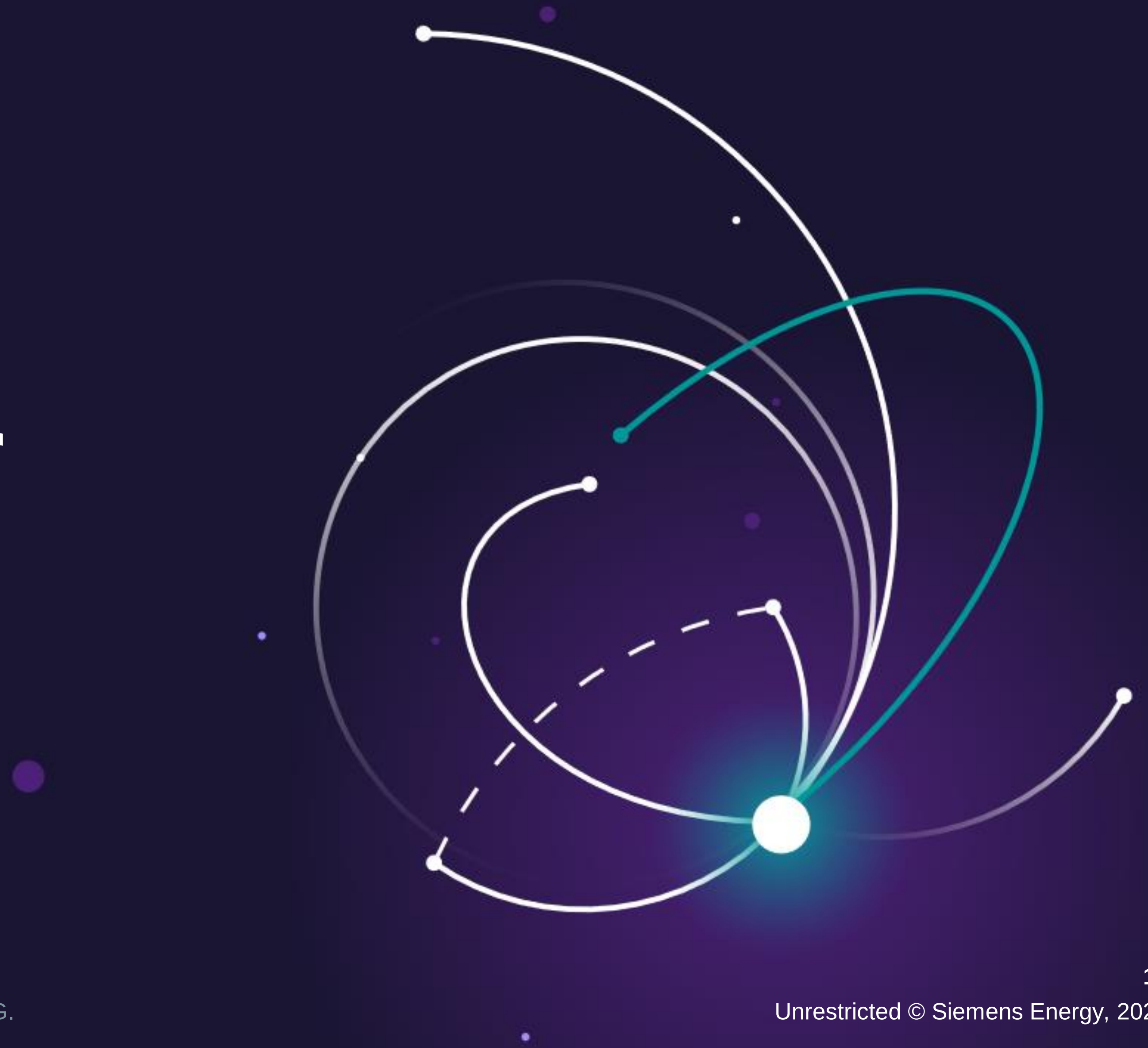
Category	Name	Description	Link
Data Handling Libraries	Numpy	Basic Library for scientific computing in Python (linear algebra, numerical functionalities etc.)	https://numpy.org/
	Pandas	Package especially built for data analysis	https://pandas.pydata.org/
	Scipy	Mathematical and engineering functions (e.g. optimization and fits, numerical integration…)	https://www.scipy.org/
Plotting Libraries	Matplotlib	2D plotting library for basic plotting	https://matplotlib.org/
	Seaborn	Is built on top of matplotlib, thus offers more functionalities	https://seaborn.pydata.org/
	Plotly	Is built for interactive graphs	https://plot.ly/python/
Machine Learning Libraries	TensorFlow	For machine learning applications (e.g. voice recognition)	https://www.tensorflow.org/
	SciKit-Learn	For machine learning applications (e.g. regression models, clustering, vector-machines…)	https://scikit-learn.org/stable/
Other Useful Libraries	Time	Provides all time-related functions	https://docs.python.org/3/library/time.html
	Sys	For interaction with interpreter (advanced)	https://docs.python.org/3/library/sys.html

Overview: Essential Graphs of Seaborn

Name	In SNS	Useful for	Plot	Relative to …	Target variable
Histogram	**.histplot()**	▪ Central tendency ▪ Spread ▪ Distribution ▪ Outliers			**Continuous data**
Scatterplot	**.relplot()** X-Axis is continuous	▪ Correlation ▪ Outliers		… other continuous data	
Lineplot		▪ Time effect ▪ Spread ▪ Outliers		… time	
Boxplot	**.catplot()** X-Axis is categorical	▪ Group effects ▪ Central tendency ▪ Spread ▪ Distribution ▪ Outliers		… categorical data	
Swarmplot					
Countplot		▪ Counts group effects		… categorical data	**Quantitative data**

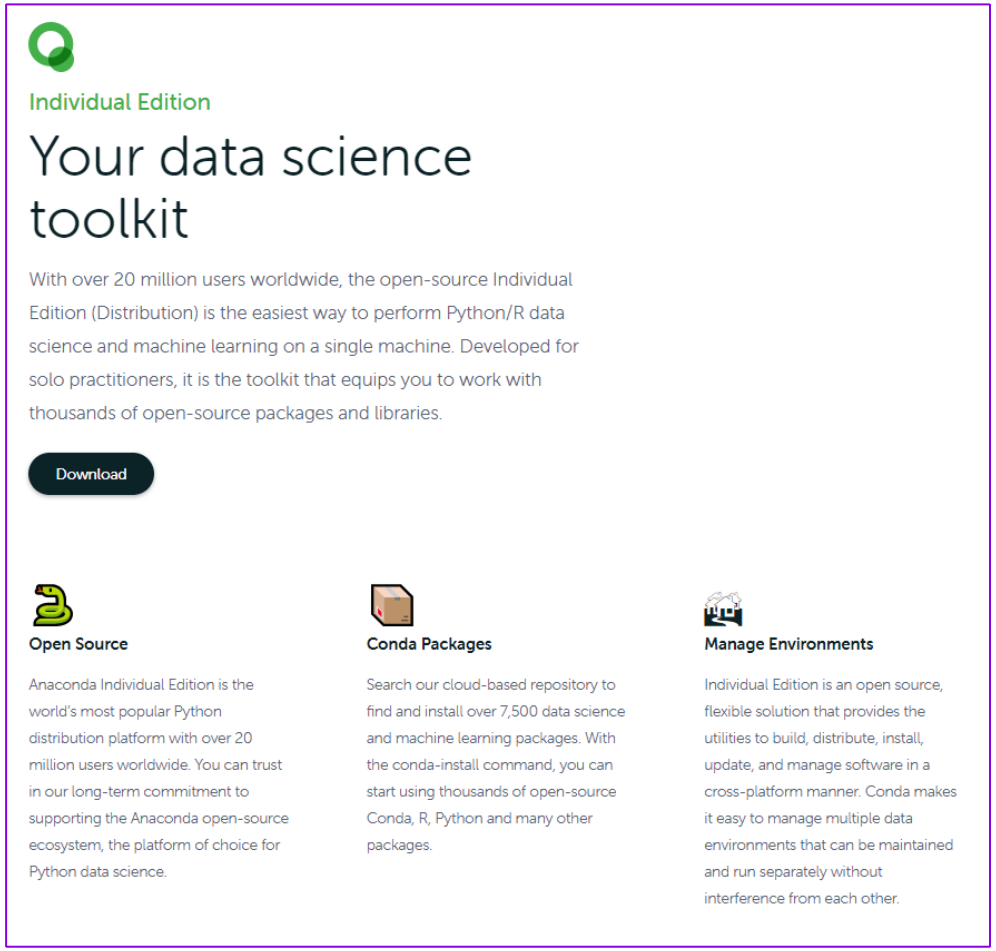
If you want to use Python at SE after this class…

Power and Gas – Large Gas Turbines, Generators



The Classical Package for Python: Anaconda (not an official SE-software for download, but often just downloaded anyway)

https://www.anaconda.com/distribution/



Individual Edition

Your data science toolkit

With over 20 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for solo practitioners, it is the toolkit that equips you to work with thousands of open-source packages and libraries.

Download

Open Source

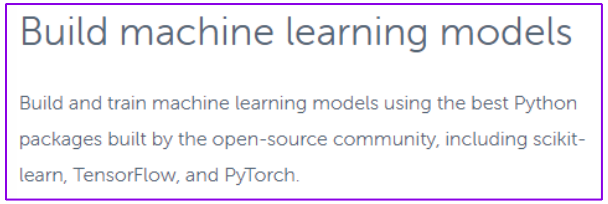
Anaconda Individual Edition is the world's most popular Python distribution platform with over 20 million users worldwide. You can trust in our long-term commitment to supporting the Anaconda open-source ecosystem, the platform of choice for Python data science.

Conda Packages

Search our cloud-based repository to find and install over 7,500 data science and machine learning packages. With the conda-install command, you can start using thousands of open-source Conda, R, Python and many other packages.

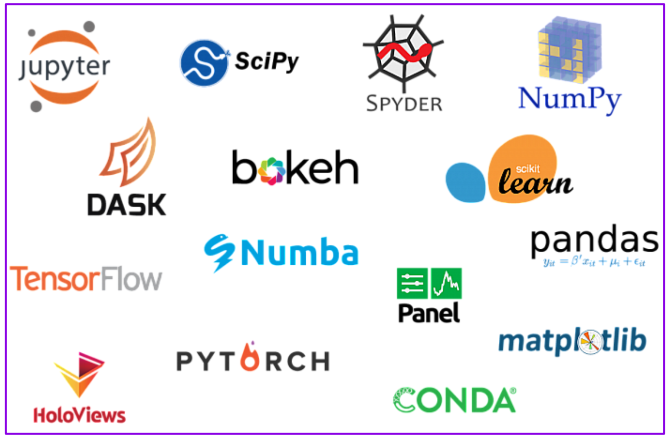
Manage Environments

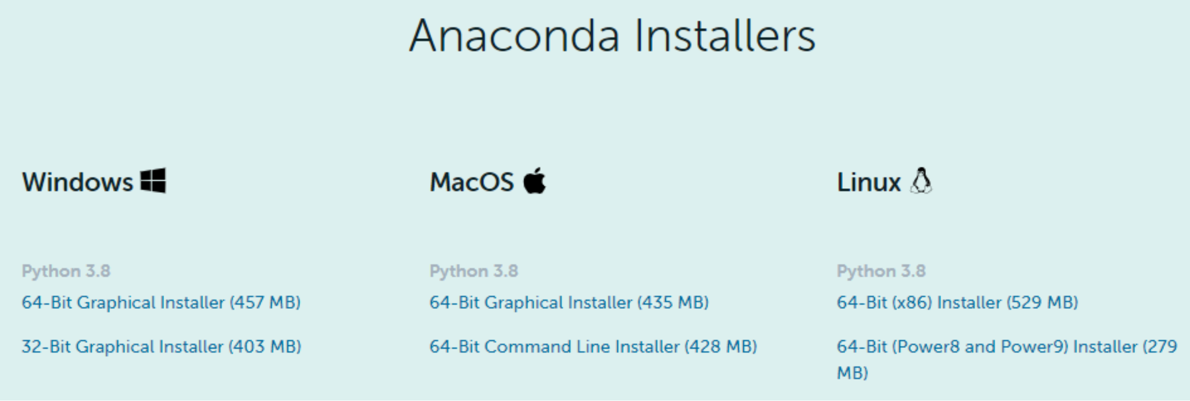
Individual Edition is an open source, flexible solution that provides the utilities to build, distribute, install, update, and manage software in a cross-platform manner. Conda makes it easy to manage multiple data environments that can be maintained and run separately without interference from each other.



Build machine learning models

Build and train machine learning models using the best Python packages built by the open-source community, including scikit-learn, TensorFlow, and PyTorch.





Anaconda Installers

Windows

Python 3.8

64-Bit Graphical Installer (457 MB)

32-Bit Graphical Installer (403 MB)

MacOS

Python 3.8

64-Bit Graphical Installer (435 MB)

64-Bit Command Line Installer (428 MB)

Linux

Python 3.8

64-Bit (x86) Installer (529 MB)

64-Bit (Power8 and Power9) Installer (279 MB)

A good and “correct” way around Anaconda at ES is EnergyAI:

Apply at: https://energyai.siemens-energy.cloud/
For problems contact: contact@energyai.siemens-energy.cloud

Why EnergyAI?

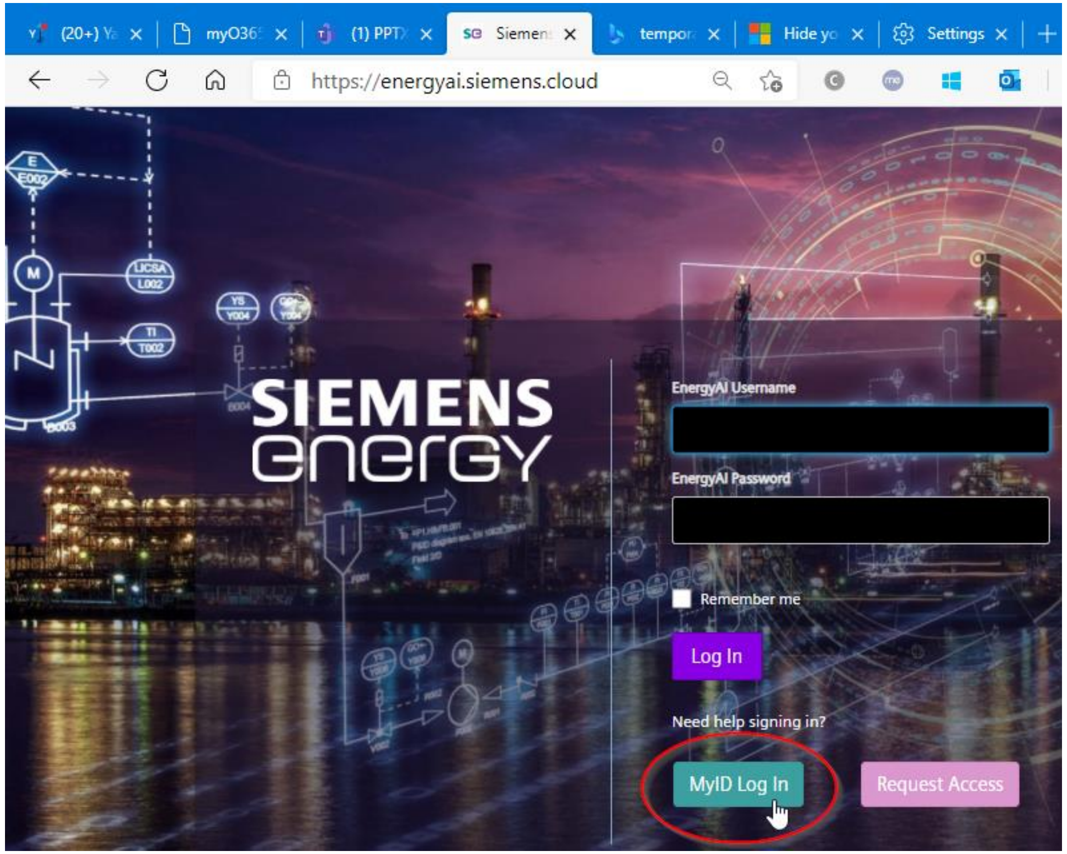
The EnergyAI - Analysis platform…

…offers an efficient interface to handle large data requests on multiple plants or units

…enables the user to perform postprocessing on-the-fly, without further data handling steps

…provides a standardized GPN System to address signals across all plants and units

…executes all operations outsourced on a server and makes results available for client



EnergyAI Username

EnergyAI Password

□ Remember me

Log In

Need help signing in?

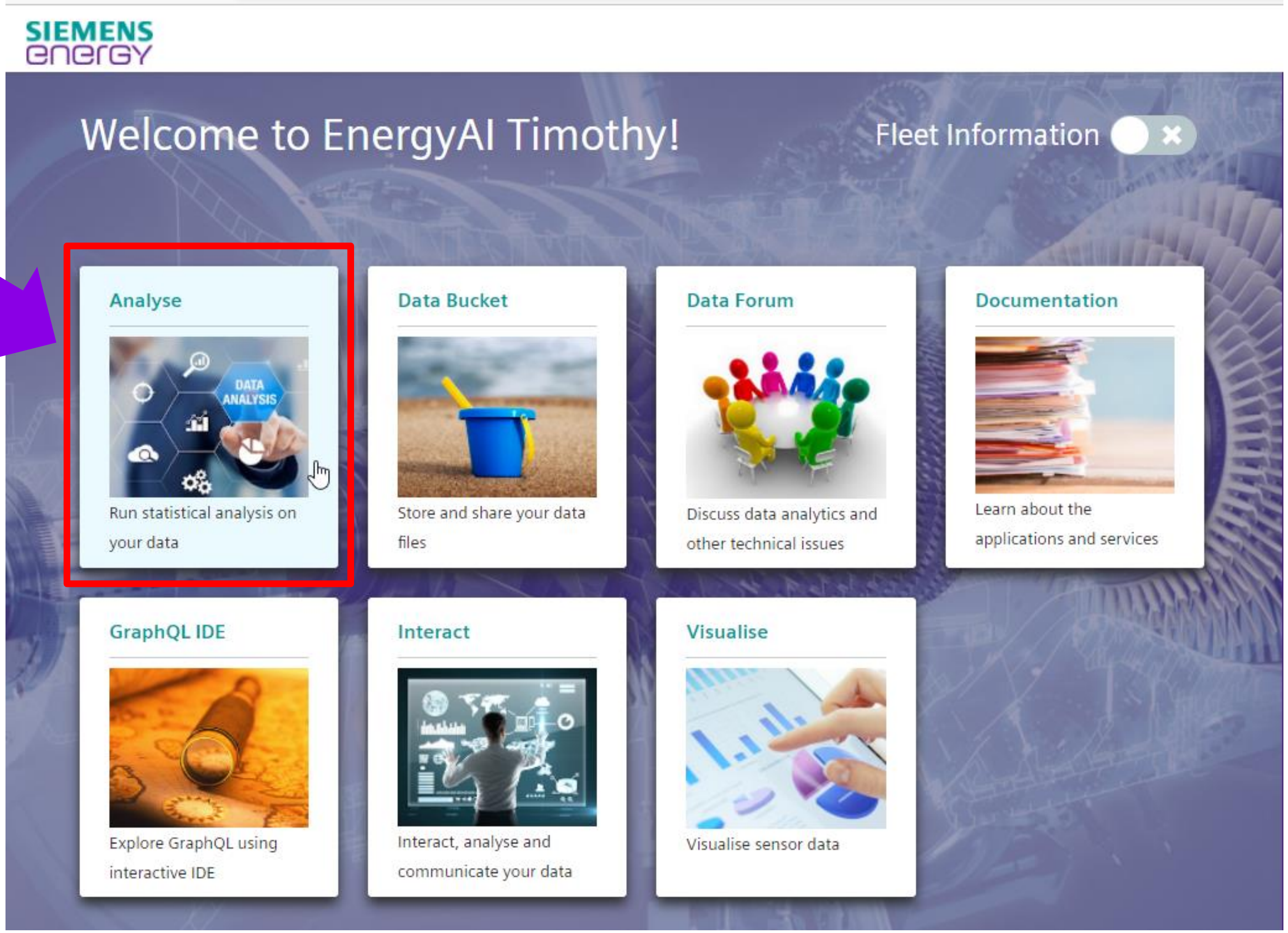
MyID Log In

Request Access

Go to “Analyse”

https://energyai.siemens-energy.cloud/

https://analyse.energyai.siemens-energy.cloud/



SIEMENS energy

Welcome to EnergyAI Timothy!

Fleet Information

Analyse

Run statistical analysis on your data

Data Bucket

Store and share your data files

Data Forum

Discuss data analytics and other technical issues

Documentation

Learn about the applications and services

GraphQL IDE

Explore GraphQL using interactive IDE

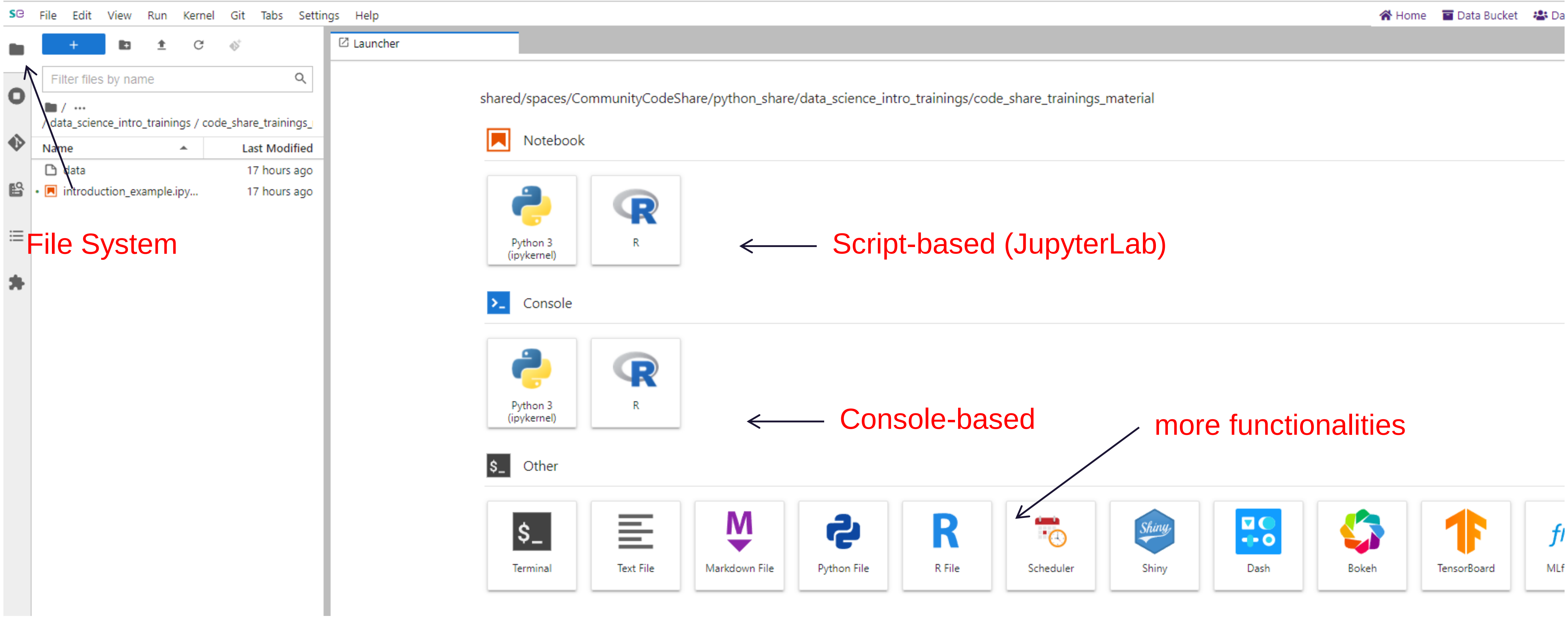
Interact

Interact, analyse and communicate your data

Visualise

Visualise sensor data

Analyse Layout



/ ...

/ data_science_intro_trainings / code_share_trainings_

Name	Last Modified
data	17 hours ago
introduction_example.ipy...	17 hours ago

shared/spaces/CommunityCodeShare/python_share/data_science_intro_trainings/code_share_trainings_material

File System

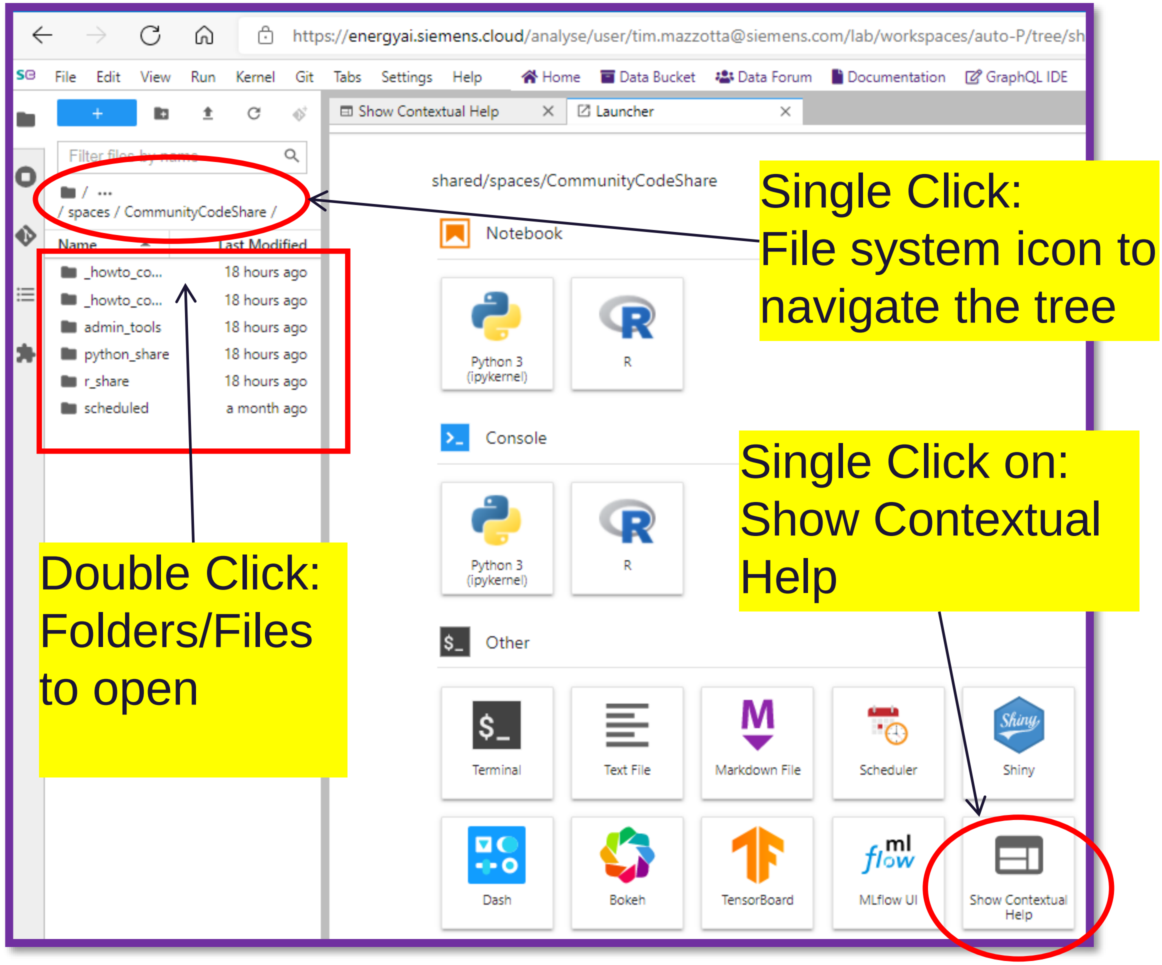
Script-based (JupyterLab)

Console-based

more functionalities

First Steps – File System

- Hierarchical File System as on Windows
- In ‘*/shared/groups*‘ , it can be worked in groups (group folder has to be requested)
- In ‘*/shared/users*‘ *some* scripts of other users are visible (copy script in your shared folder to make it visible for others)
- Go to: ‘*/shared/demos*‘ for several demos on the usage of the datahub with R or Python
- Go to: ‘*/shared/spaces/CommunityCodeShare*‘ for EnergyAI CodeShare projects
- Repositories from GitHub can be embedded in folders to simultaneously work on scripts (recommended)



https://energyai.siemens.cloud/analyse/user/tim.mazzotta@siemens.com/lab/workspaces/auto-P/tree/sh

/ ...
/ spaces / CommunityCodeShare /

Name	Last Modified
_howto_co...	18 hours ago
_howto_co...	18 hours ago
admin_tools	18 hours ago
python_share	18 hours ago
r_share	18 hours ago
scheduled	a month ago

shared/spaces/CommunityCodeShare

Single Click: File system icon to navigate the tree

Double Click: Folders/Files to open

Single Click on: Show Contextual Help