Start 18:30

WiFi SSID: Schiphol_Group_Office

The Pipeline Factory

Meetup 1 - Data Science Environments

Data Science Environments

Data Science Environments

Example:

Programming language: R

Interactive environment: Jupyter notebook

R library for time series: fpp

Data Science Environments

Minimal setup

- Programming language
- Interactive development environment
- Libraries

Data Science Environments

Language	R	Python
Interactive environment	RStudio	Jupyter notebook
Library	dplyr	Pandas

Data Science Environments

Adding complexity

- Databases (MongoDB, MariaDB)
- Deep learning (Theano, TensorFlow, Keras)
- Distributed systems (Hadoop, Spark)
- Backend for website, app or API (node, flask)
- Frontend (angular, d3)
- Streaming frameworks (Kafka)
- Cloud solutions (Amazon, Azure, Google)

Data Science Environments

Reproducibility

- Installation dependencies is time consuming
- Configuration is time consuming
- Version conflicts are common

Data Science Environments

Virtual machines

- Isolated environment
- Encapsulates application + dependencies
- Complete operating system
- Virtualized hardware

Containers

- Isolated environment
- Encapsulates application + dependencies
- Shares kernel with host OS
- Shared hardware

The Project

Project description

Create a data science environment

- Use container / virtual machine solution of your choice
- Choose software that you like to use / try
- Create a git repository for sharing code
- Create documentation
- Present your project

Vanilla project

Tutorial

- Create a data science environment
- Inside a Docker container
- With programming language Python
- With Jupyter notebooks
- Work with with existing (data) science notebooks

Juper hub The environment of Rok Milhvc

git

Using git by Frans

Groups

Project time until 21:30

Resources

Vanilla project https://github.com/sjoerddehaan/data-science-containers

Docker https://www.docker.com/

Vagrant https://www.vagrantup.com/

Git cheat sheet

Presentations