Deep Learning for Language Analysis

Deep Learning Introduction

Institut für Digital Humanities | Johanna Binnewitt & Valmir Etemi | 09.-10.09.2019

Introduction

What are you going to learn?

- Brief introduction to Machine Learning
- Neural Network Architectures
- Implementing Neural Networks in Keras

Introduction

Which technologies are we going to use?

- Python
- Keras
- Jupyter Notebook
- … in a Docker Container

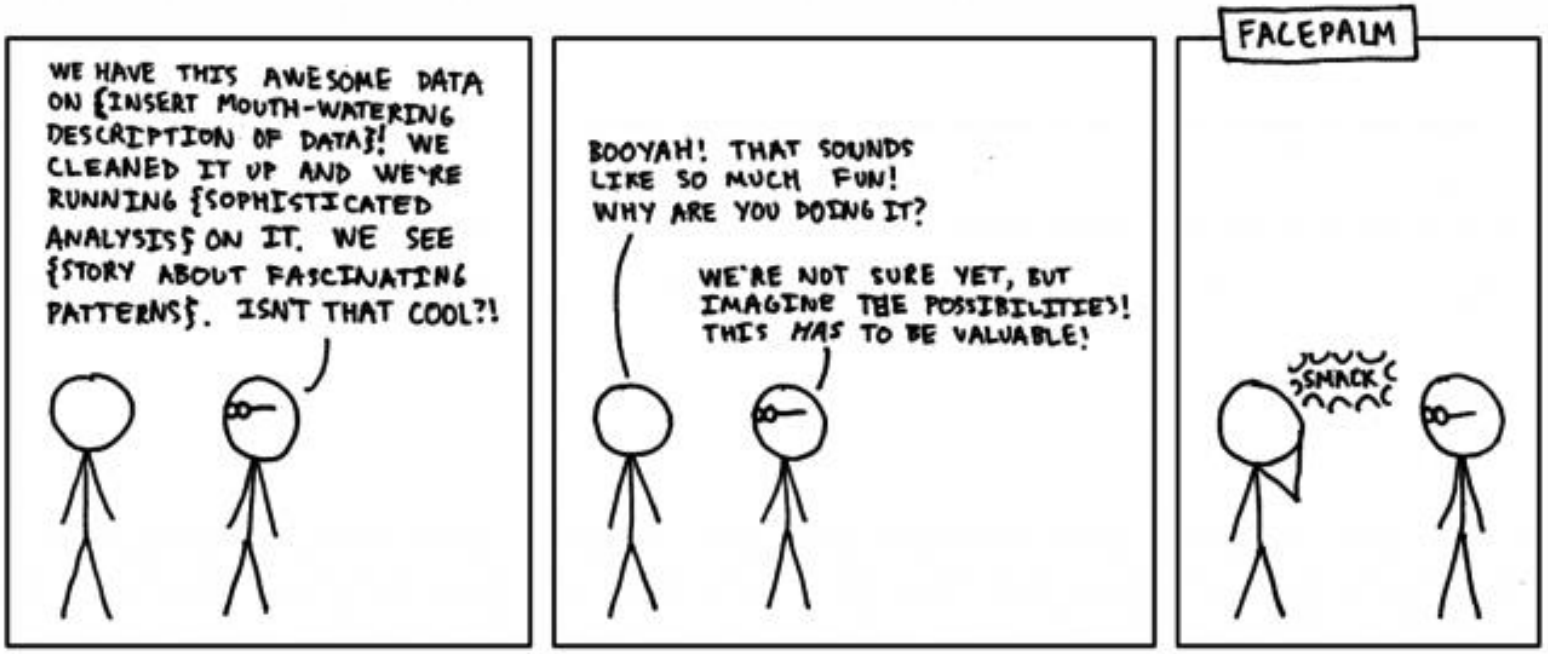
Schedule

Time	Monday, September 09	Tuesday, September 10
09:00		Neural Network Architecture
10:30		*Coffee Break*
11:00		Tuning the Neural Network
12:30		*Lunch Break*
14:00	Welcome / Introduction	Hands on: Text Classification
15:30	*Coffee Break*	
16:00	Introduction / Setup	Parallel Session Presentation
17:00	Closing	Closing

Data Pipeline

1. Define Research Goal
2. Retrieve Data
3. Prepare Data
4. Explore Data
5. Model Data
6. Improve Model

Data Pipeline
1. Define Research Goal



https://medium.com/the-data-experience/building-a-data-pipeline-from-scratch-32b712cfb1db

Data Pipeline
2. Retrieve Data

- Depending on your research area
- Example repositories:
  - https://archive.ics.uci.edu/ml/datasets.php
  - https://www.kaggle.com/datasets
- Collect own data

Data Pipeline
3. Prepare Data

- Data cleansing: remove false, inconsistent or unnecessary data
- Data integration: enrich data with other sources
- Data transformation: transform data into suitable format
- **How to transform text data into a model?**

Data Pipeline
3. Prepare Data

Types of Data

- Structured data (e.g. SQL Databases)
- Semi-structured data (e.g. CSV files)
- Unstructured data (e.g. text files)

Sources of Data

- Machine generated (e.g. server log files)
- Natural Language
- Audio, video, images
- Streaming

Data Pipeline
3. Prepare Data

Types of Data

- Structured data (e.g. SQL Databases)
- Semi-structured data (e.g. CSV files)
- **Unstructured data (e.g. text files)**

Sources of Data

- Machine generated (e.g. server log files)
- **Natural Language**
- Audio, video, images
- Streaming

Data Pipeline
4. Explore Data

- Understand retrieved data
- How do variables interact?
- Is my data set representative? No? → retrieve data
- Methods: descriptive statistics, plotting and simple modelling

Data Pipeline
5. Model Data

- Build a model which suits your research goal
- Model should depend on the task you want to solve

https://bentoml.com/posts/2019-04-19-one-model/

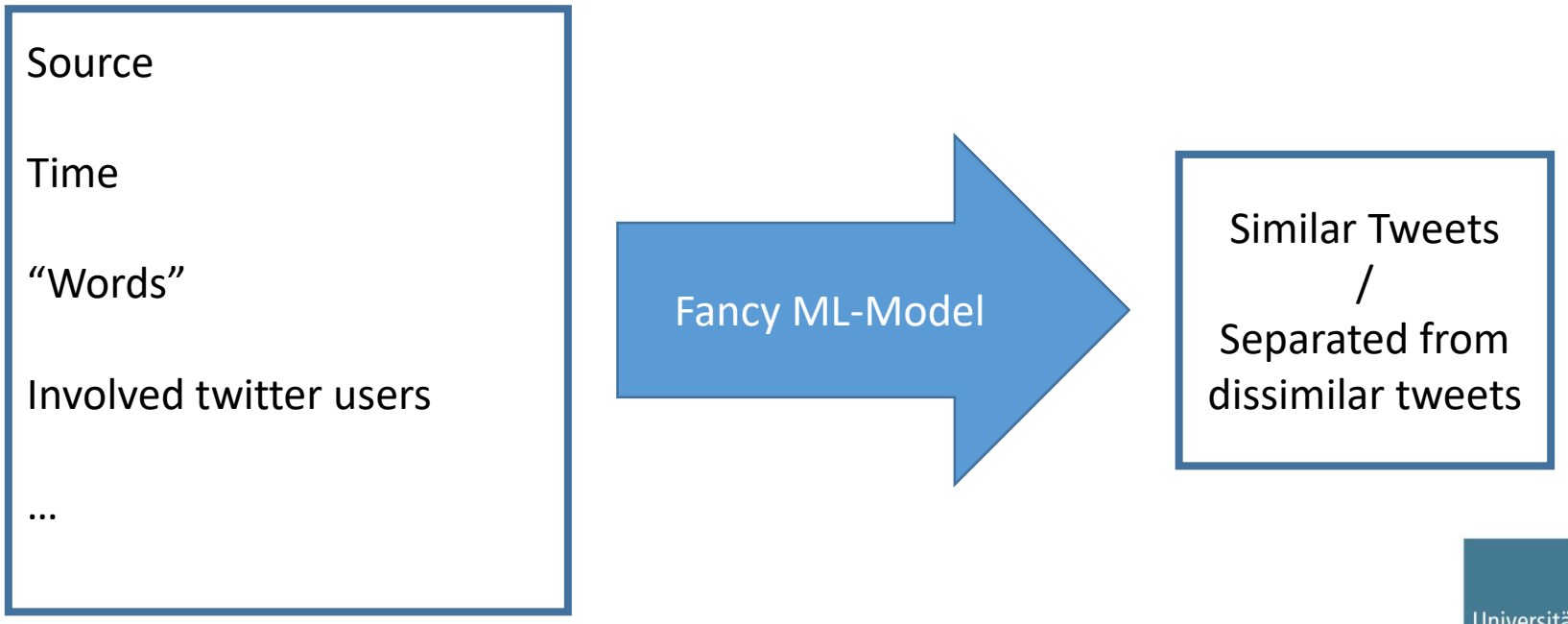
Data Pipeline
5. Model Data

Different Machine Learning Problems

- Unsupervised learning
  - Analyse data without external knowledge
  - For example: Clustering

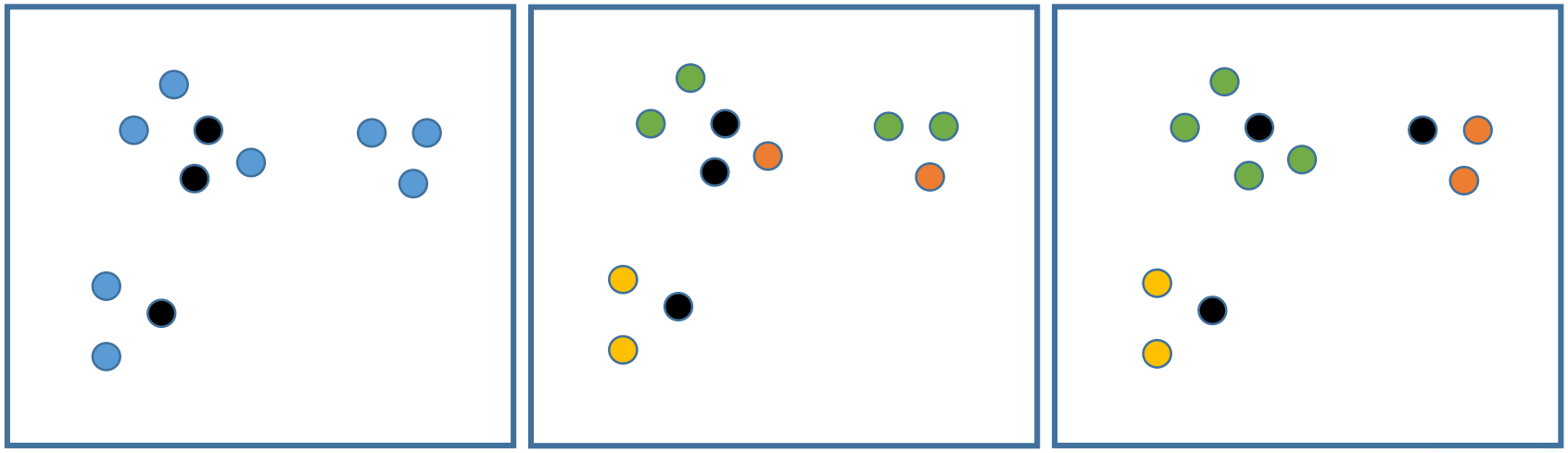
Data Pipeline
5. Model Data

- ML Task: **Clustering**
- Example Goal: grouping tweets on similar topics



Data Pipeline
5. Model Data

- ML Task: **Clustering**
- Solved (for example) by K-Means
  - Vector space
  - Iteratively adjusts the centroid of a cluster



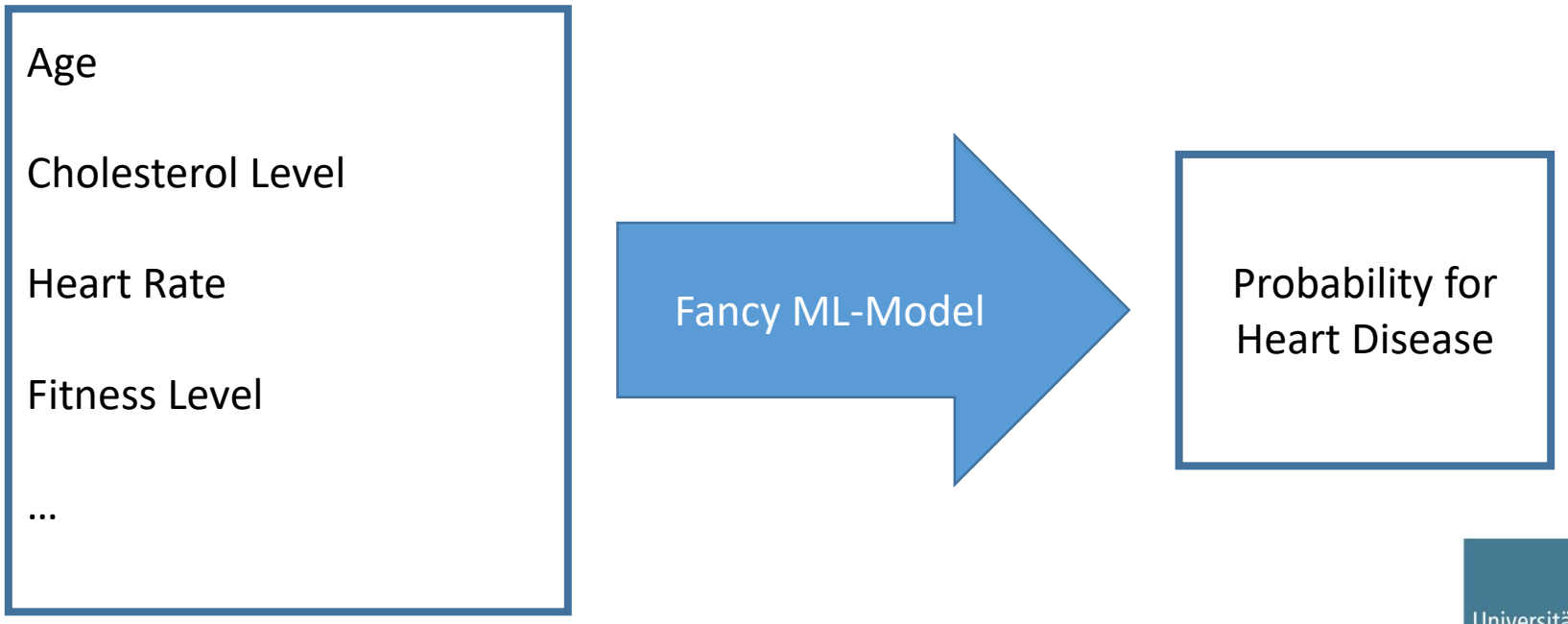
Data Pipeline
5. Model Data

Different Machine Learning Problems

- Supervised learning
  - Analyse data by using external knowledge
  - For example: Classification

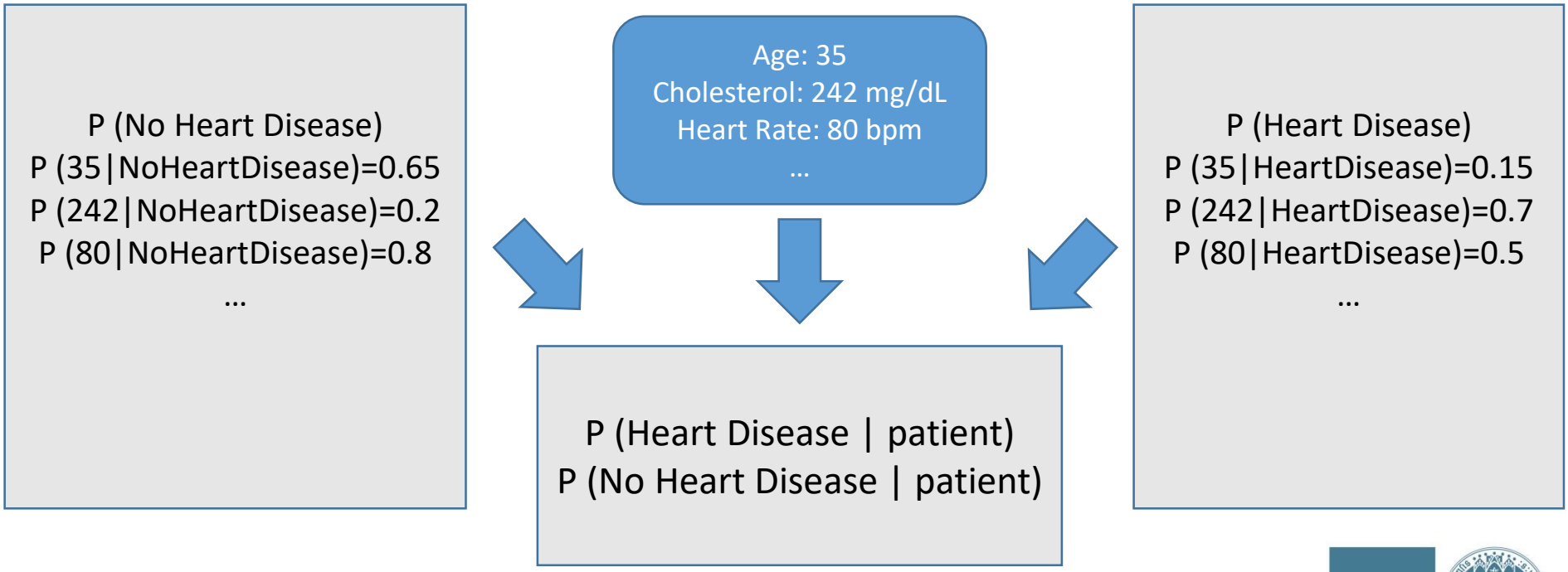
Data Pipeline
5. Model Data

- ML Task: **Classification**
- Example Goal: predict heart disease for a person



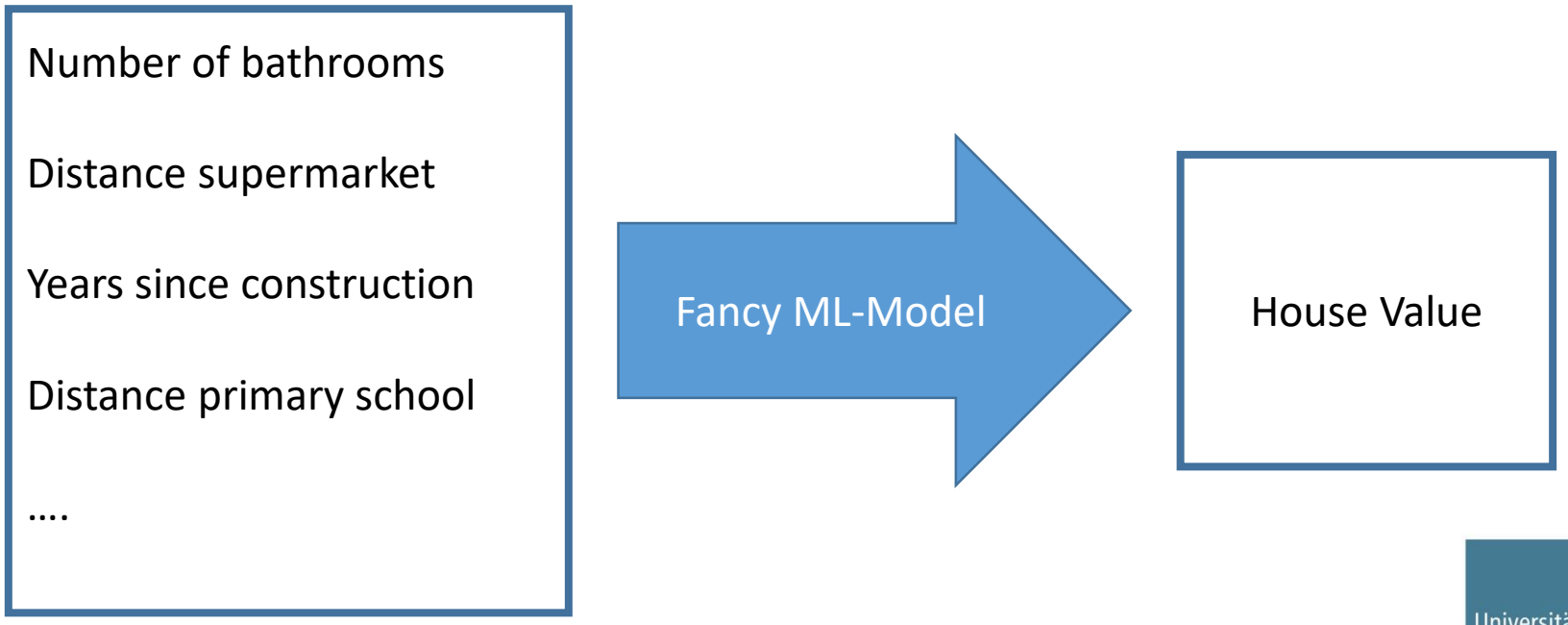
Data Pipeline
5. Model Data

- ML Task: **Classification**
- Solved (for example) by (Gaussian) Naïve Bayes



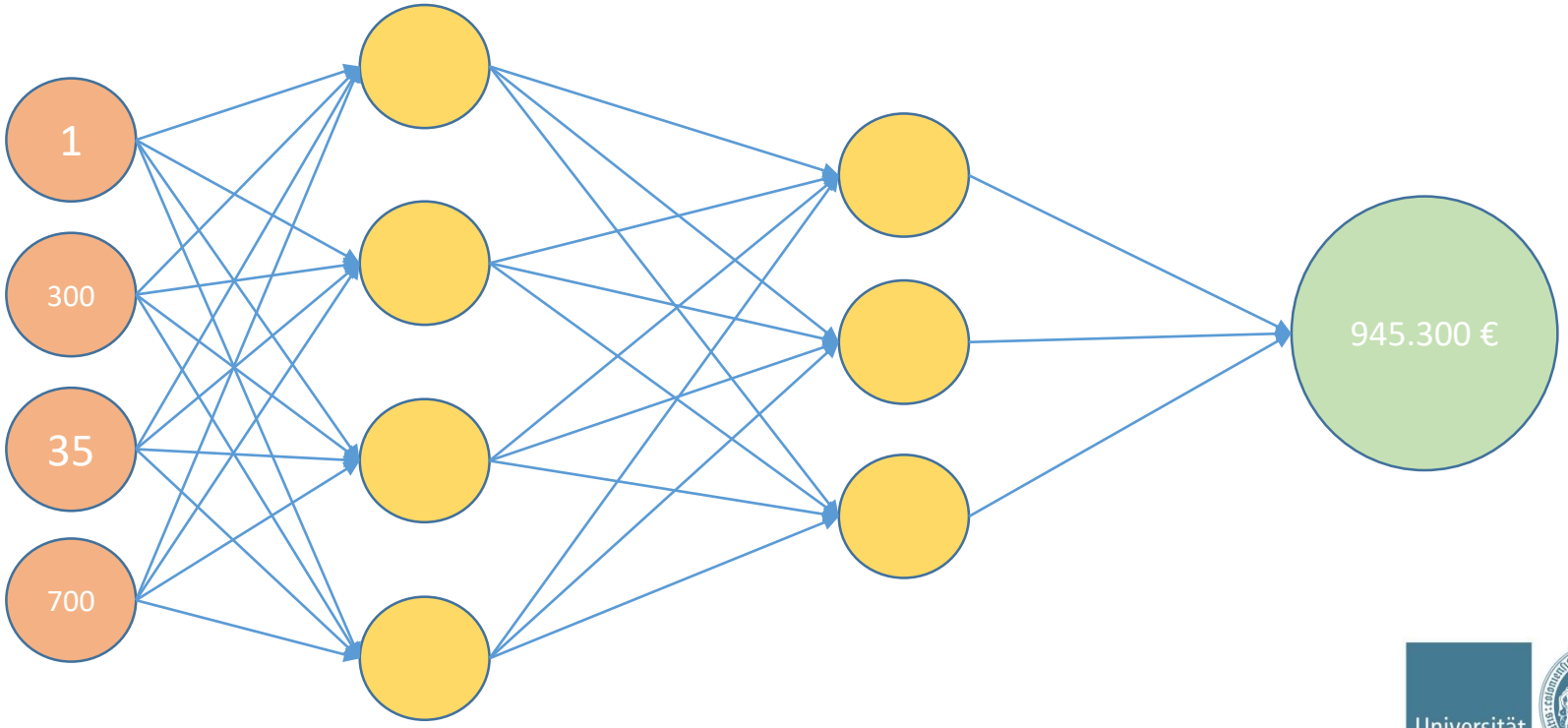
Data Pipeline
5. Model Data

- ML Task: **Regression**
- Example Goal: estimating real estate values



Data Pipeline
5. Model Data

- ML Task: **Regression**
- Solved (for example) by neural network



Data Pipeline
5. Model Data

Different Machine Learning Problems

- Semi-supervised learning
  - Hybrid approaches
  - For example: Clustering first, Categorizing new objects afterwards into clusters

Data Pipeline
6. Improve Model

- Evaluate the results of your model
- Change configurations to improve the results
  - Input variables
  - Model configuration
    - Number of clusters
    - Number of Layers
    - Activation Functions on Layers
    - …
  - Whole Model

And Now:

Introduction to Docker, Keras and Jupyter Notebook