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

Universität
zu Köln

Data Pipeline

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

Universität
zu Köln

Data Pipeline
1. Define Research Goal
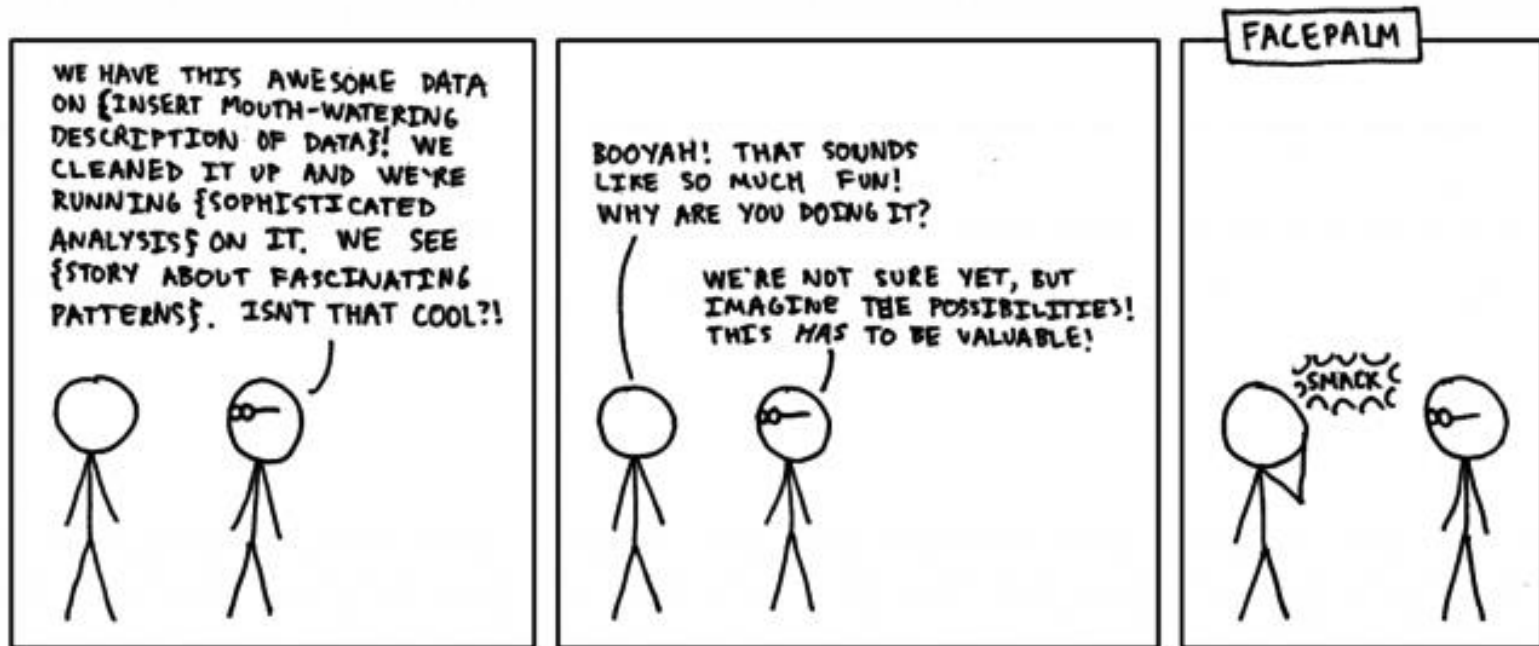
Data Pipeline
2. Retrieve Data

- Depending on your research area

- Example repositories:

    - https://archive.ics.uci.edu/ml/datasets.php

    - https://www.kaggle.com/datasets

- Collect own data

Universität
zu Köln

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?**

Universität
zu Köln

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

Universität zu Köln

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

Universität
zu Köln

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

Universität
zu Köln

Data Pipeline
5. Model Data

- Build a model which suits your research goal

- Model should depend on the task you want to solve

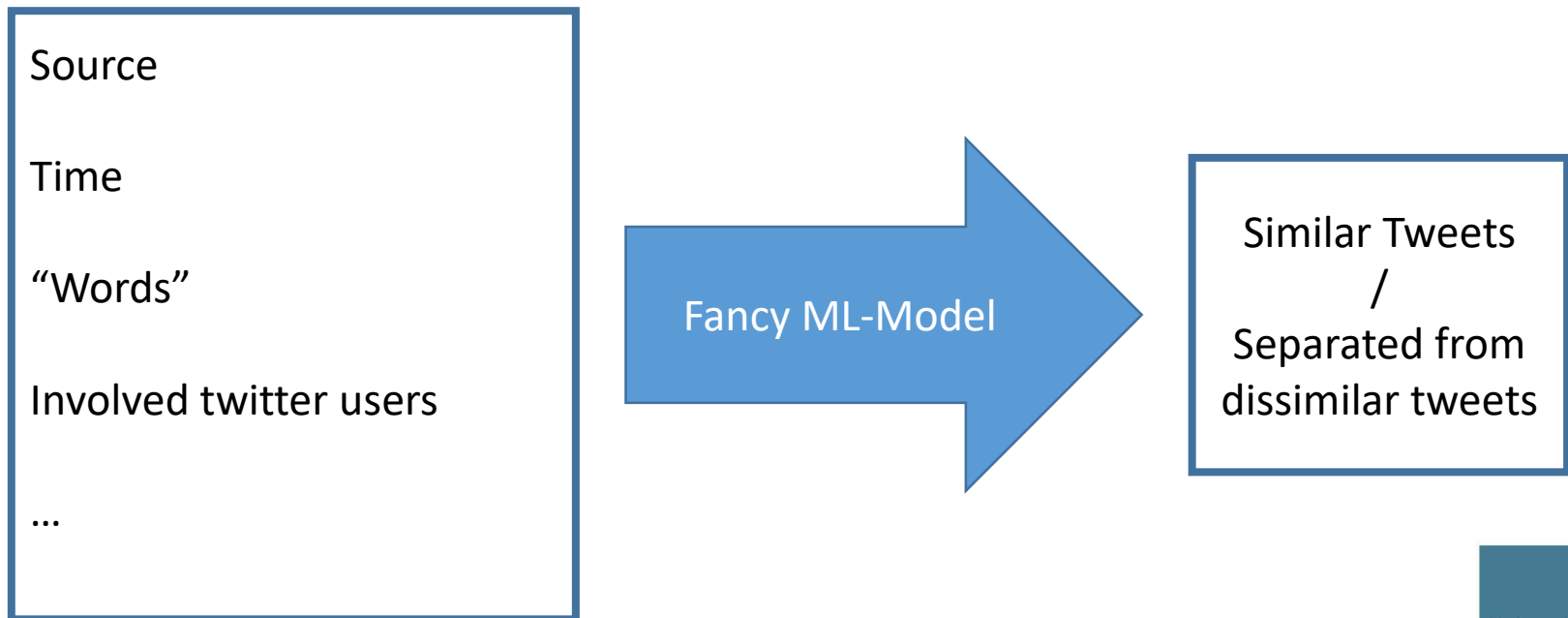https://bentoml.com/posts/2019-04-19-one-model/

Data Pipeline
5. Model Data

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

Source

Time

"Words"

Involved twitter users

…

Fancy ML-Model

Similar Tweets
/
Separated from
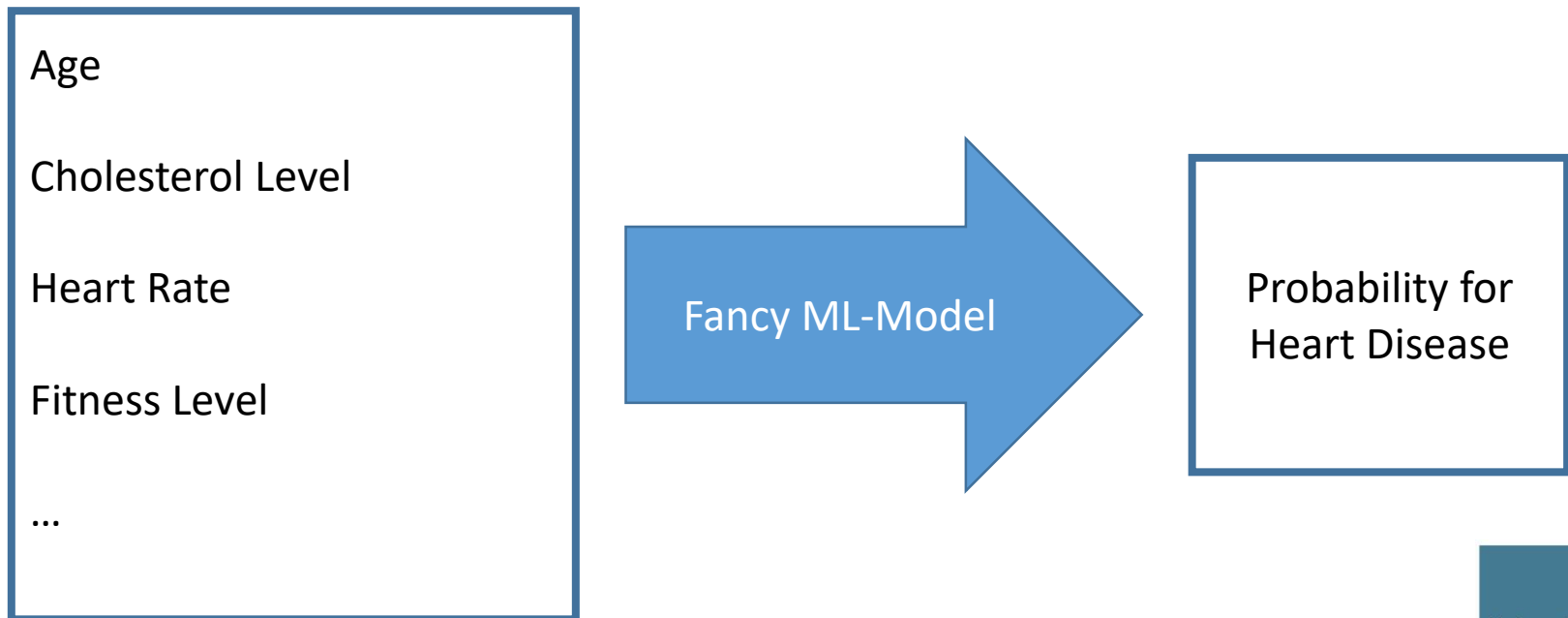dissimilar tweets

Universität
zu Köln

Data Pipeline
5. Model Data

- ML Task: **Clustering**

- Solved (for example) by K-Means

  - Vector space
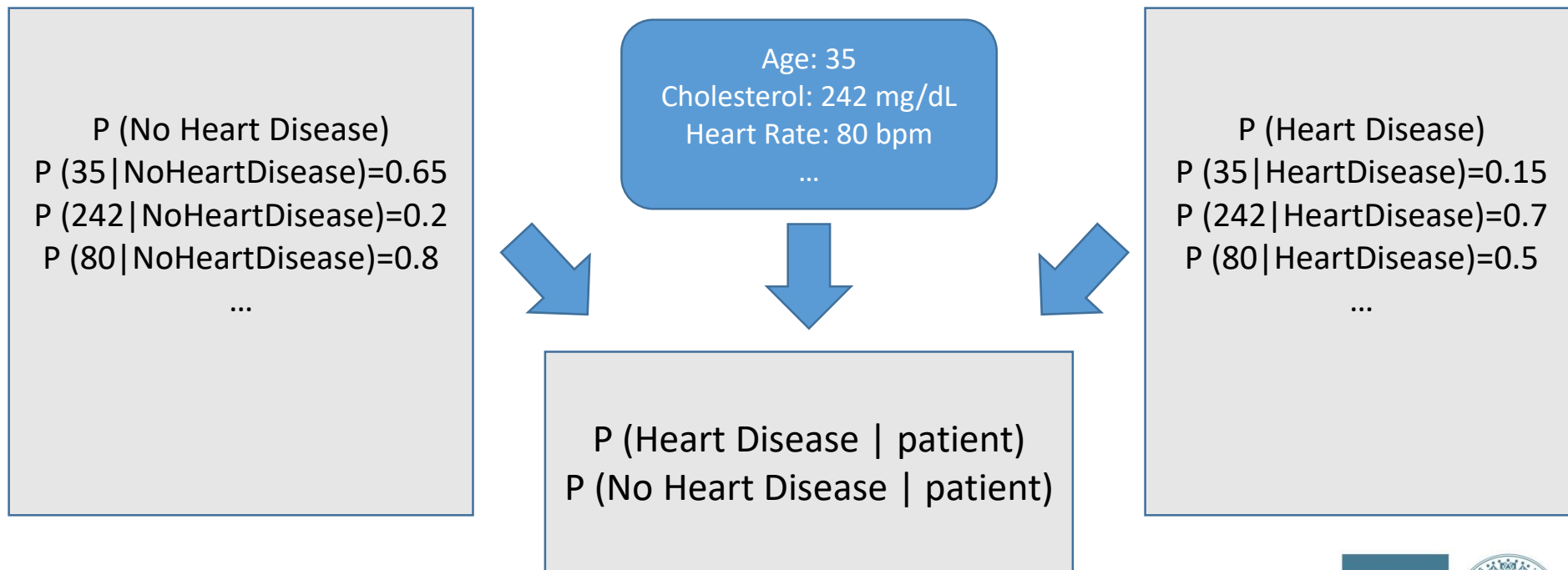
  - Iteratively adjusts the centroid of a cluster

Universität
zu Köln
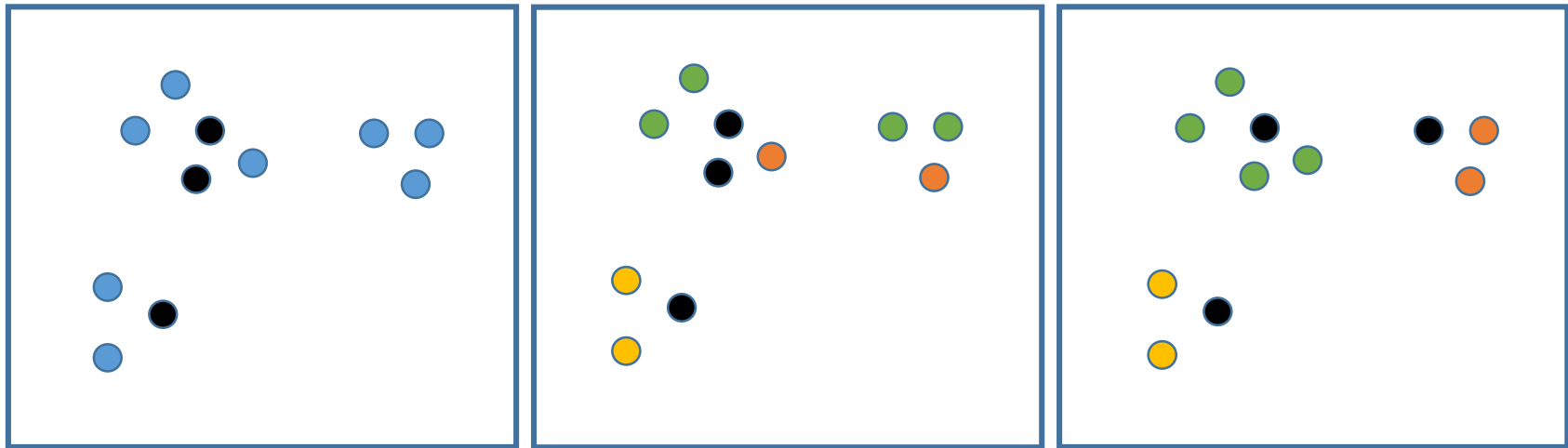
Data Pipeline
5. Model Data

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

Age		
Cholesterol Level	Fancy ML-Model →	Probability for Heart Disease
Heart Rate		
Fitness Level		
…		

Universität zu Köln

Data Pipeline
5. Model Data

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

P (No Heart Disease)
P (35|NoHeartDisease)=0.65
P (242|NoHeartDisease)=0.2
P (80|NoHeartDisease)=0.8

…

Age: 35
Cholesterol: 242 mg/dL
Heart Rate: 80 bpm
…

P (Heart Disease)
P (35|HeartDisease)=0.15
P (242|HeartDisease)=0.7
P (80|HeartDisease)=0.5

…

P (Heart Disease | patient)
P (No Heart Disease | patient)

Universität
zu Köln

Data Pipeline
5. Model Data

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

| Number of bathrooms

Distance supermarket

Years since construction

Distance primary school

…. | → Fancy ML-Model → | House Value |
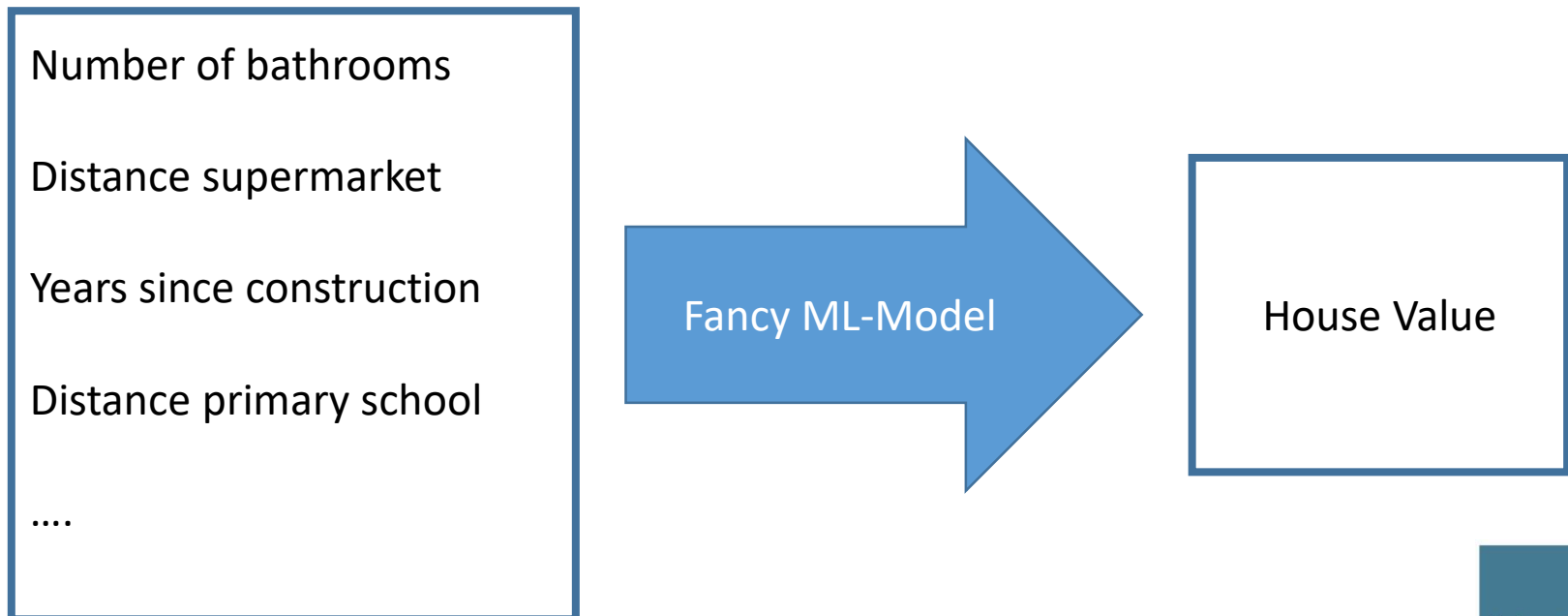
Data Pipeline
5. Model Data

- ML Task: **Regression**
- Solved (for example) by neural network
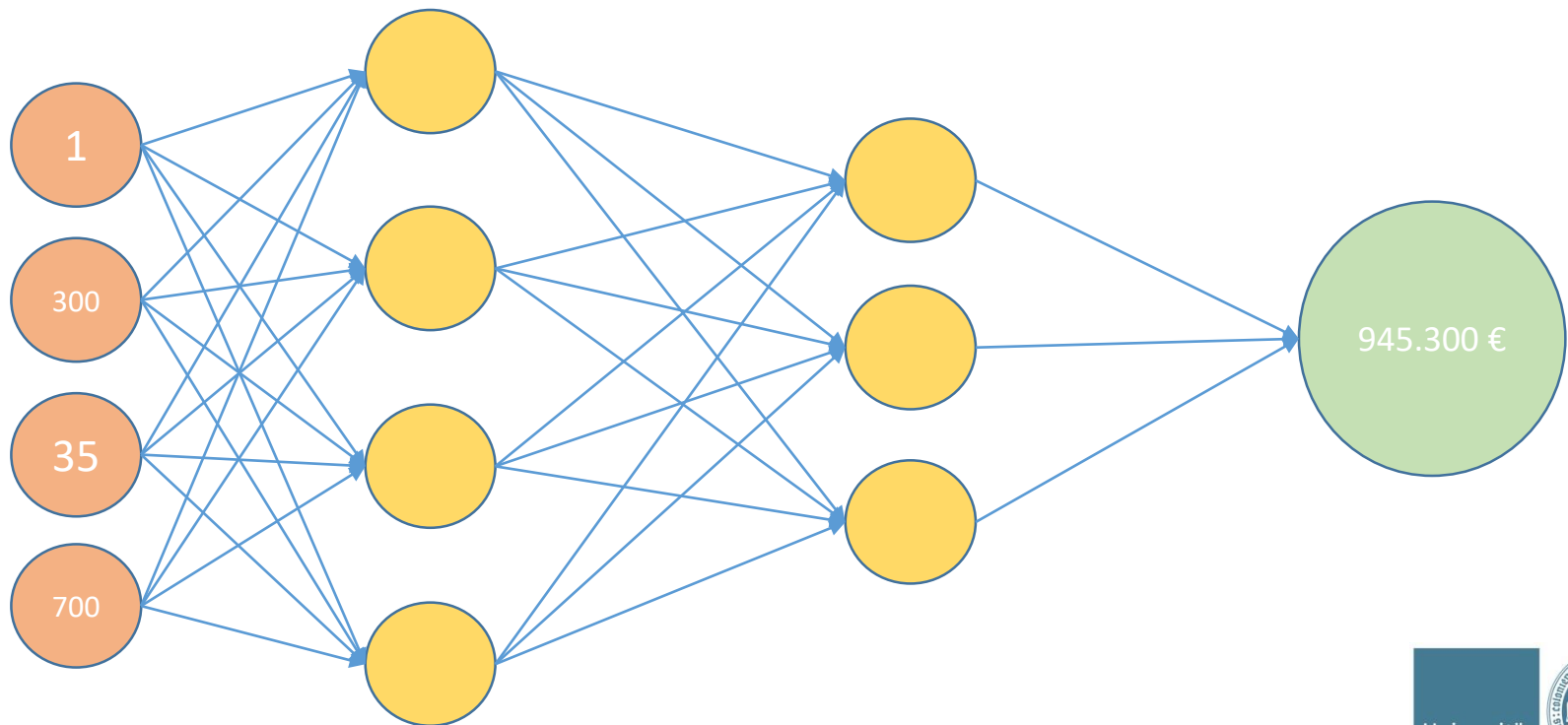
Data Pipeline
6. Improve Model

- 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

Universität
zu Köln

And Now:

Introduction to Docker, Keras and Jupyter Notebook