# Data Engineering + MLops SS24

*Week 1 – Introduction*

Prof. Dr.-Ing. Janis Keuper

INSTITUTE FOR MACHINE LEARNING AND ANALYTICS

Hochschule Offenburg
offenburg.university

# Outline

- Machine Learning Work Flow
- What is “Data Engineering ?”
- What is „MLOps ?“
- Python Data Ecosystem

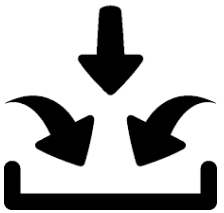
# What is Data Engineering?

**Simple Question: What is Data Engineering?**

# ML Workflow



**Start:** Hypothesis



Collect Data

Exploratory Data Analytics (EDA)

- visualize
- pre-process
- clean, merge

Feature Engineering

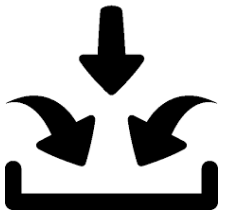
Model building

- ML
- Statistic

# Data Science Workflow



**Start:** Hypothesis

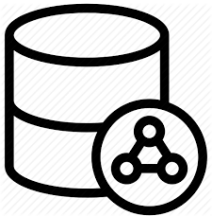


Collect Data

Exploratory Data Analytics (EDA)

- visualize
- pre-process
- clean, merge

Focus so far...



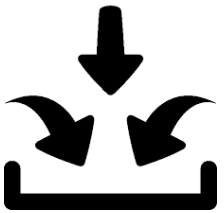
Model building
- ML
- Statistic

**ML 1+2**

Feature Engineering

# Data Science Workflow

**Start:** Hypothesis



Collect + Store Data

Exploratory Data Analytics (EDA)

- visualize
- pre-process
- clean, merge

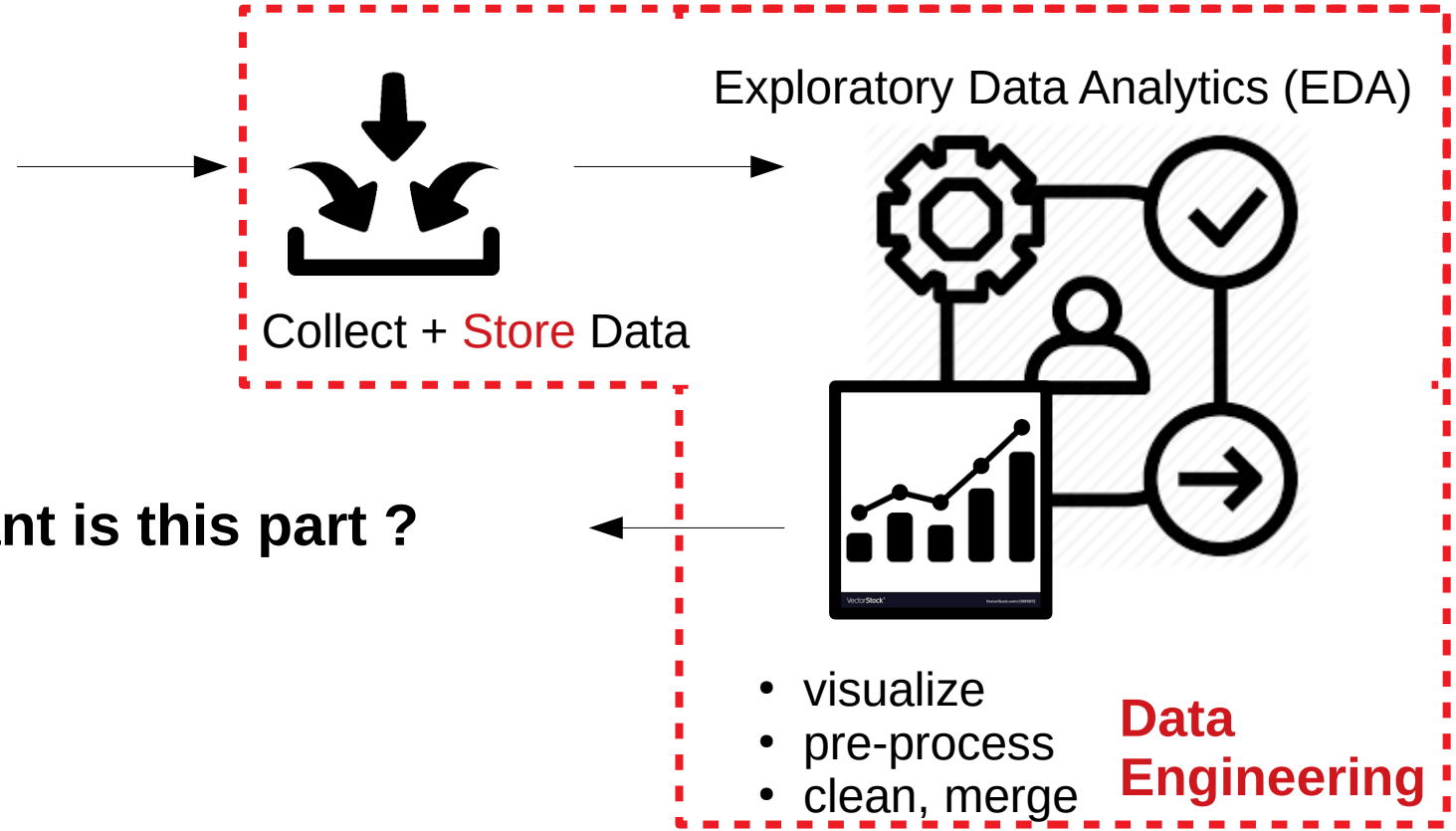
**Data Engineering**

Feature Engineering

Model building
- ML
- Statistic

# Data Engineering





# What is Data Engineering?

Hochschule Offenburg
offenburg.university

## Simple Question: What is Data Engineering?

*And how does it relate to or differ from… ?*

**Data Science**

**Big Data**

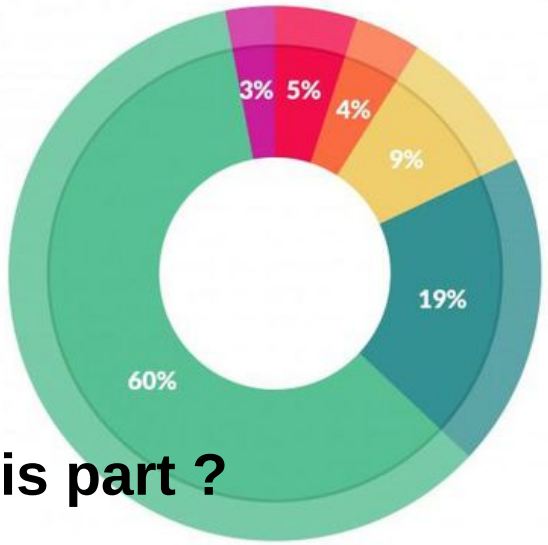
**Data Bases**

**Data Mining**

**Artificial Intelligence**

**Machine Learning?**

# Data Engineering




What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Data Engineering

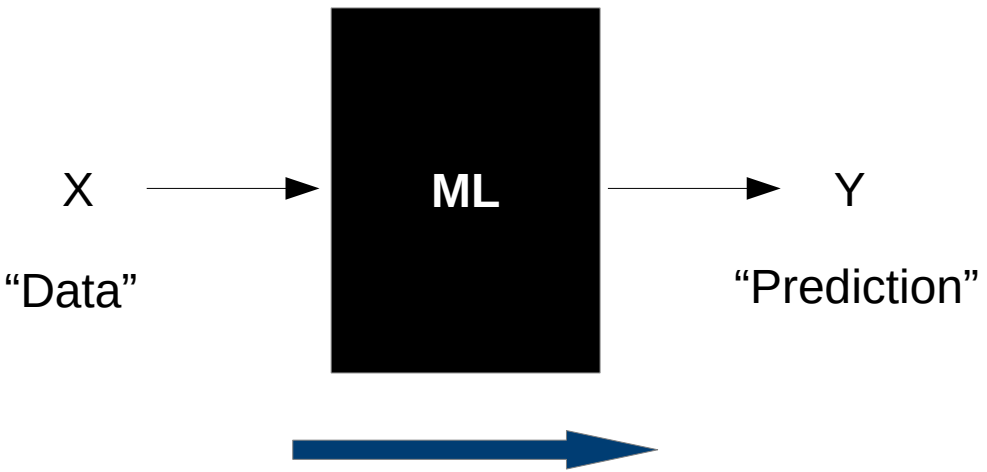
ML

**How important is this part ?**

source: study by forbes.com [2]

## How important is this part ?

General abstract scheme of ML algorithms:

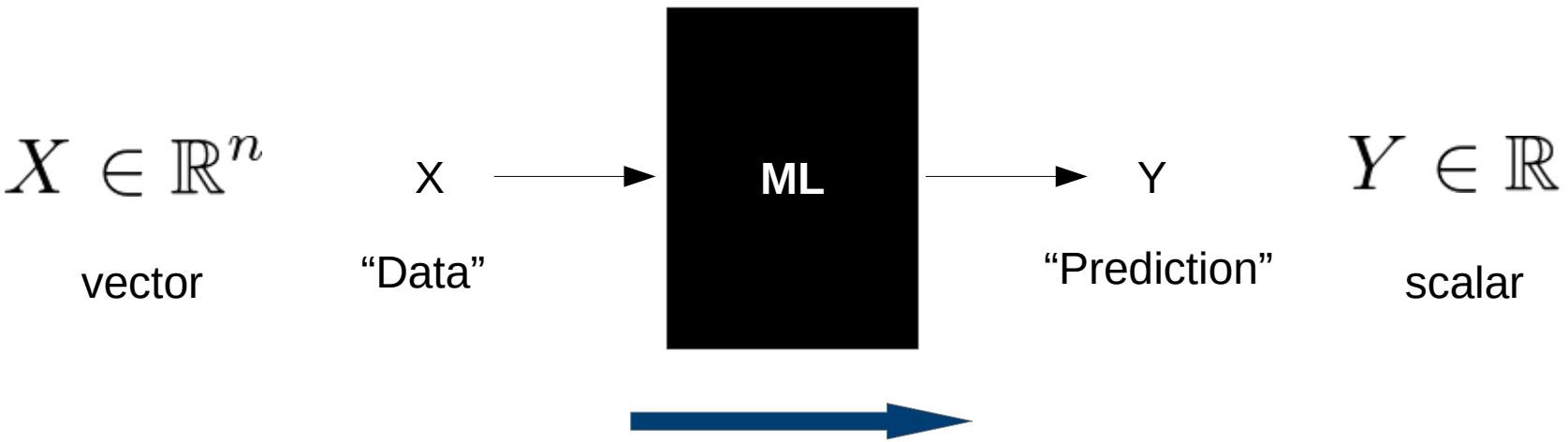



ML algorithms “learns” ***mapping*** from input to output by example data

# Data In- and Output of ML Algorithms



General abstract scheme of ML algorithms (so far):

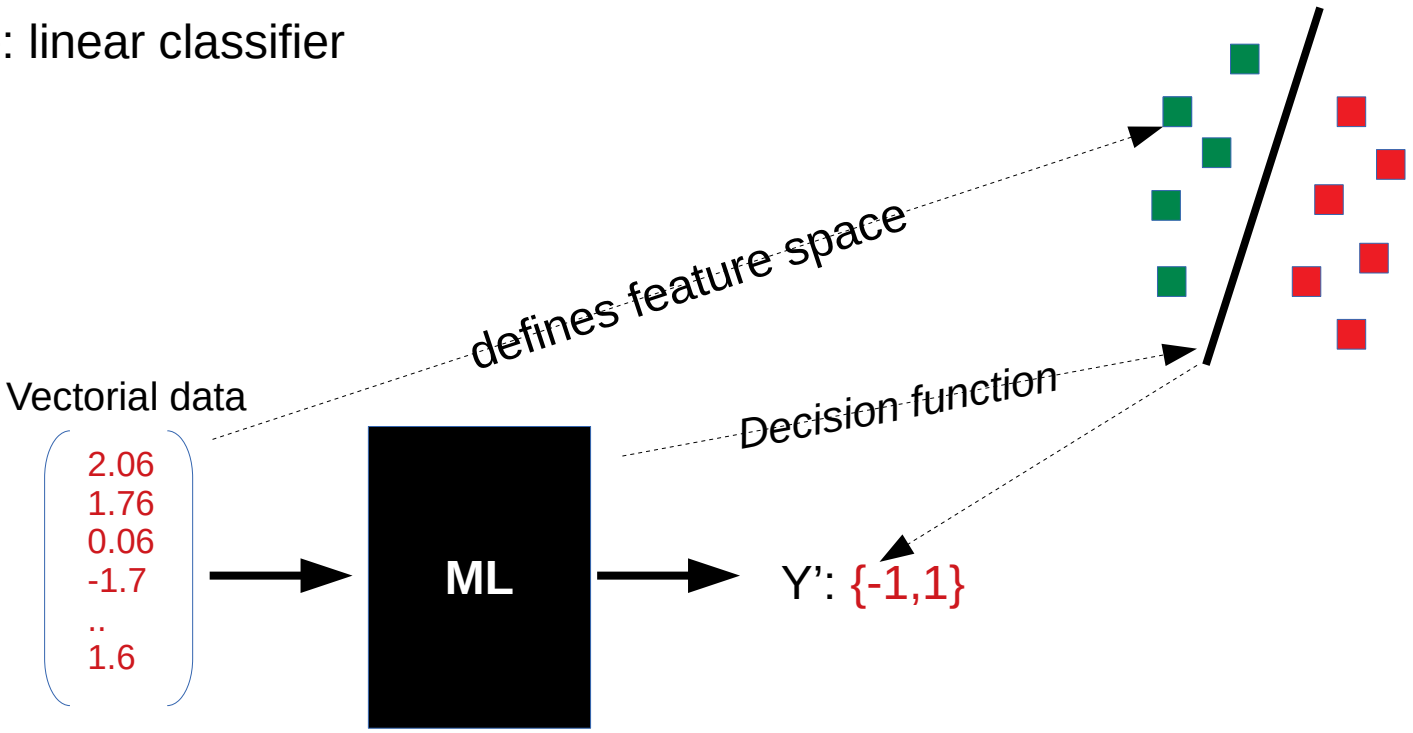



ML algorithms "learns" ***mapping*** from input to output by example data

# Data In- and Output of ML Algorithms

Hochschule Offenburg
offenburg.university

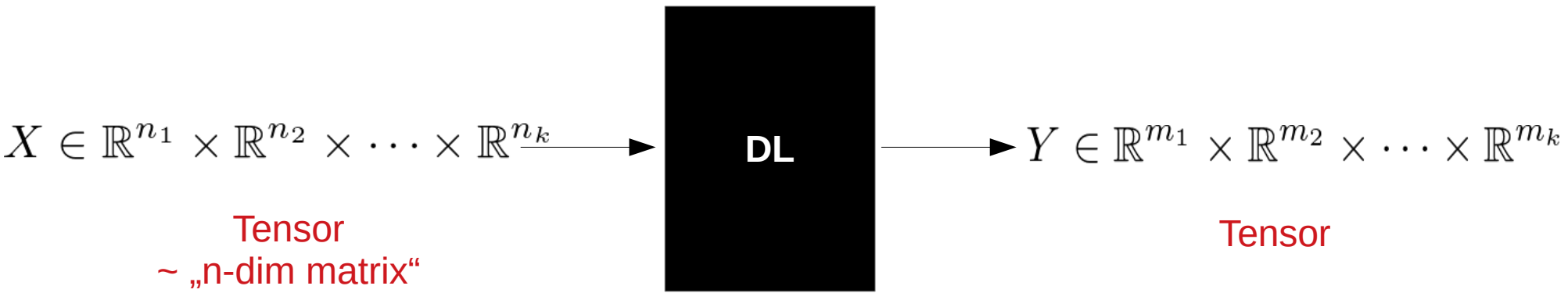
Example: linear classifier



# Data In- and Output of ML Algorithms

Hochschule Offenburg
offenburg.university

General abstract scheme of ML algorithms (modern algorithms, i.e. Deep Learnig):

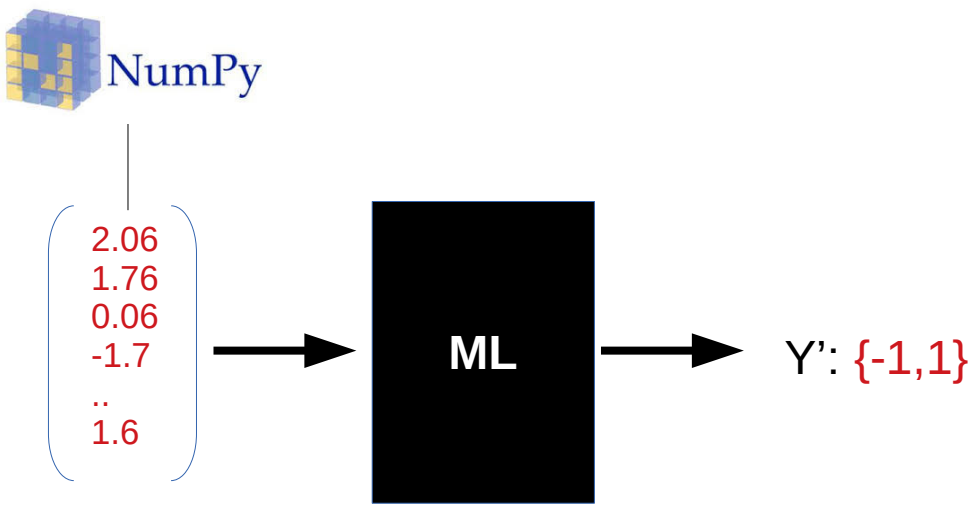



Formal introduction in Deep Learning lecture

# Where is the tensor data coming from?

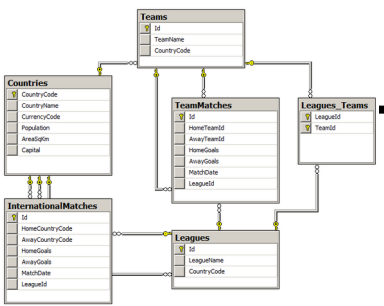


- Problem data already comes as „clean“ vectors → „toy data“



# Where is the tensor data coming from?

- Problem data already comes as „clean“ vectors → „toy data“
- Data comes from several relational data bases
  - → select and merge data
  - → need to find relevant variables
  - → remove corrupt data
  - → fill missing values
  - → unifiy data formats
  - → ...





Y': {-1,1}

# Where is the tensor data coming from?

Hochschule Offenburg
offenburg.university

- Problem data already comes as „clean“ vectors → „toy data“
- Data comes from several relational data bases
- Data is not numerical
  - → categorical data, e.g. {„small“, „medium“, „large“}
  - → text, images, audio
  - → graphs
  - → time series

ML

Y’: {-1,1}

# Where is the tensor data coming from?

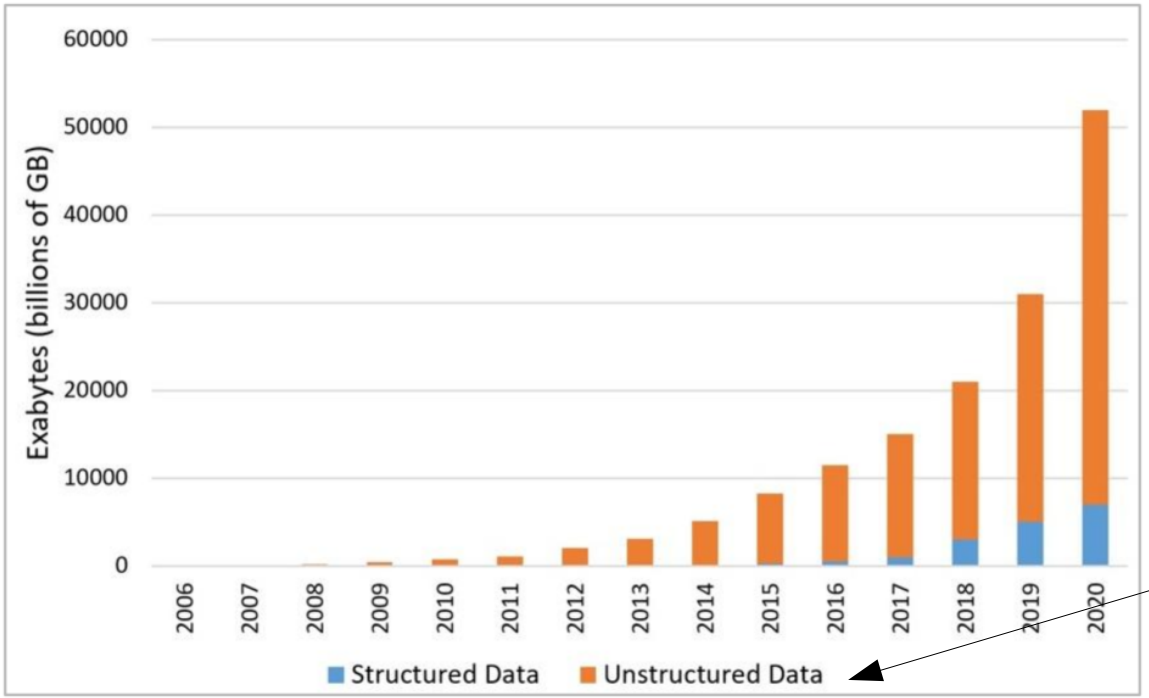
Hochschule Offenburg
offenburg.university

- Problem data already comes as „clean“ vectors → „toy data“
- Data comes from several relational data bases
- Data is not numerical
- Information is hidden large unstructured data
  - → websites
  - → PDF tables
  - → sensors ...

ML

Y’: {-1,1}

# Where is the data coming from?

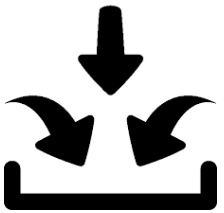



[3]

Not directly accessible by a “query” and easily transformable Into a tensor representation

# ML Workflow

**Start:** Hypothesis



Collect + Store Data

Exploratory Data Analytics (EDA)

- visualize
- pre-process
- clean, merge

**Data Engineering**

Feature Engineering

Model building
- ML
- Statistic

# What is MLops?

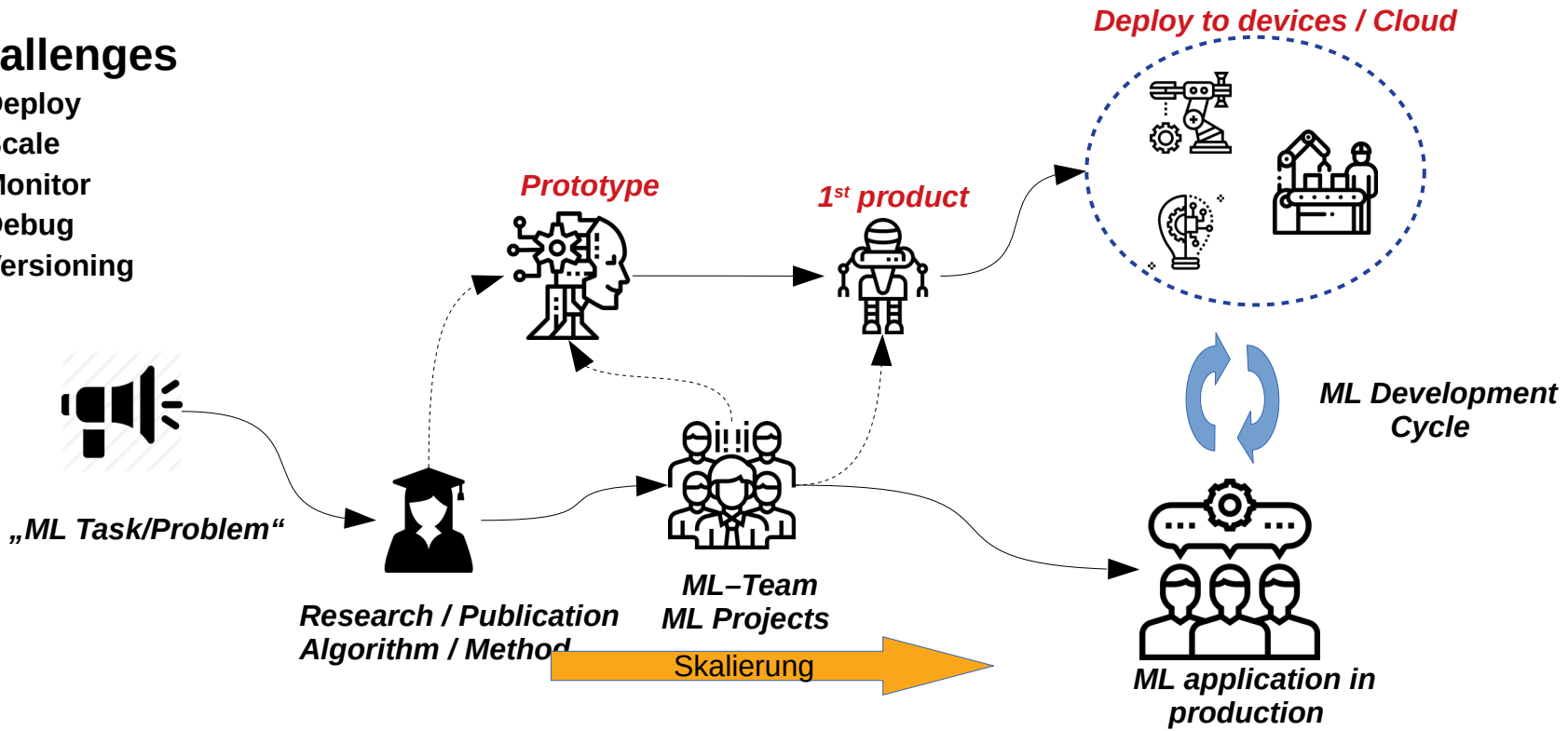
**MLops → Machine Learning Operations !**

# Machine Learning in REAL Applications

Hochschule Offenburg
offenburg.university

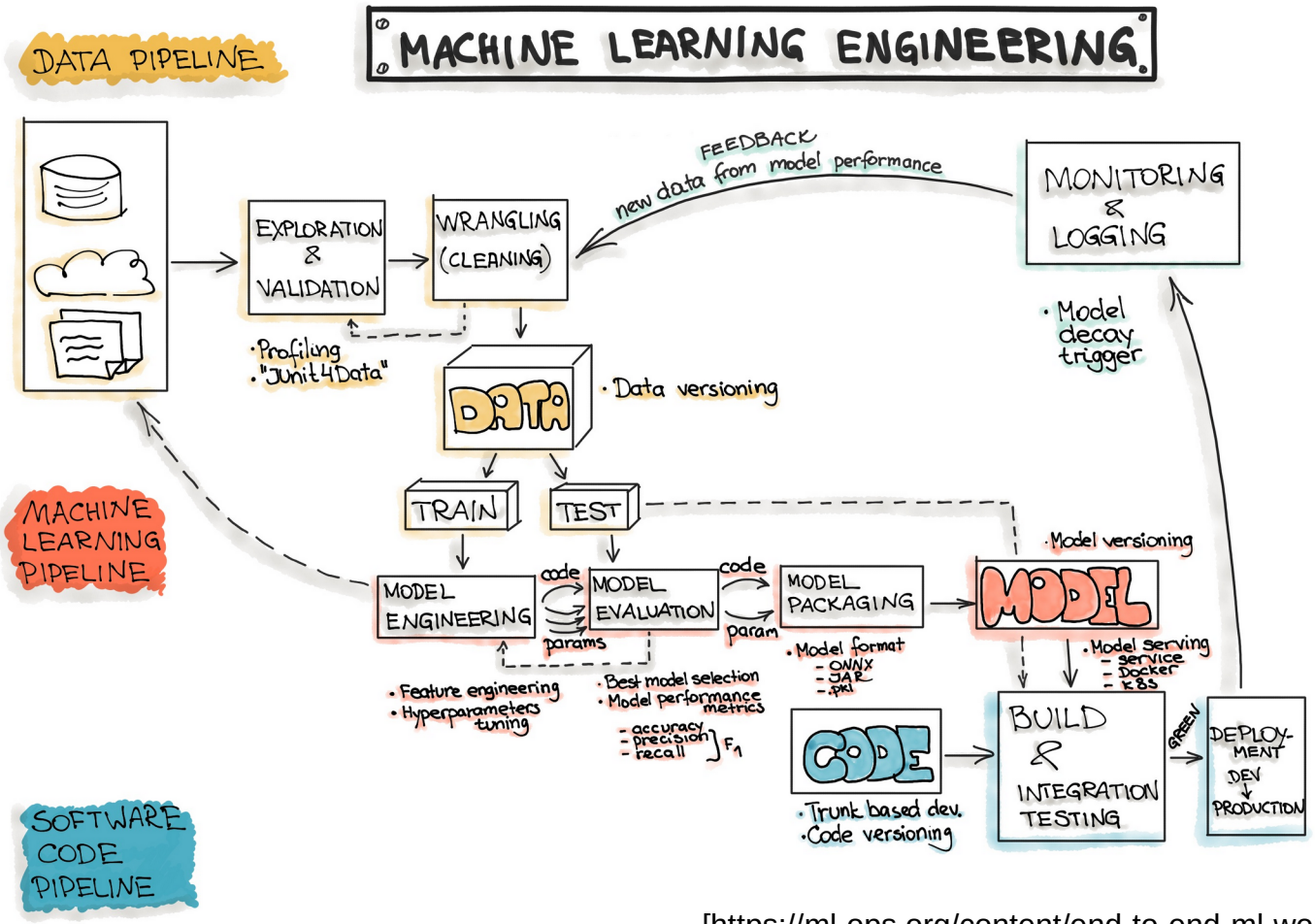
## Challenges

- Deploy
- Scale
- Monitor
- Debug
- Versioning



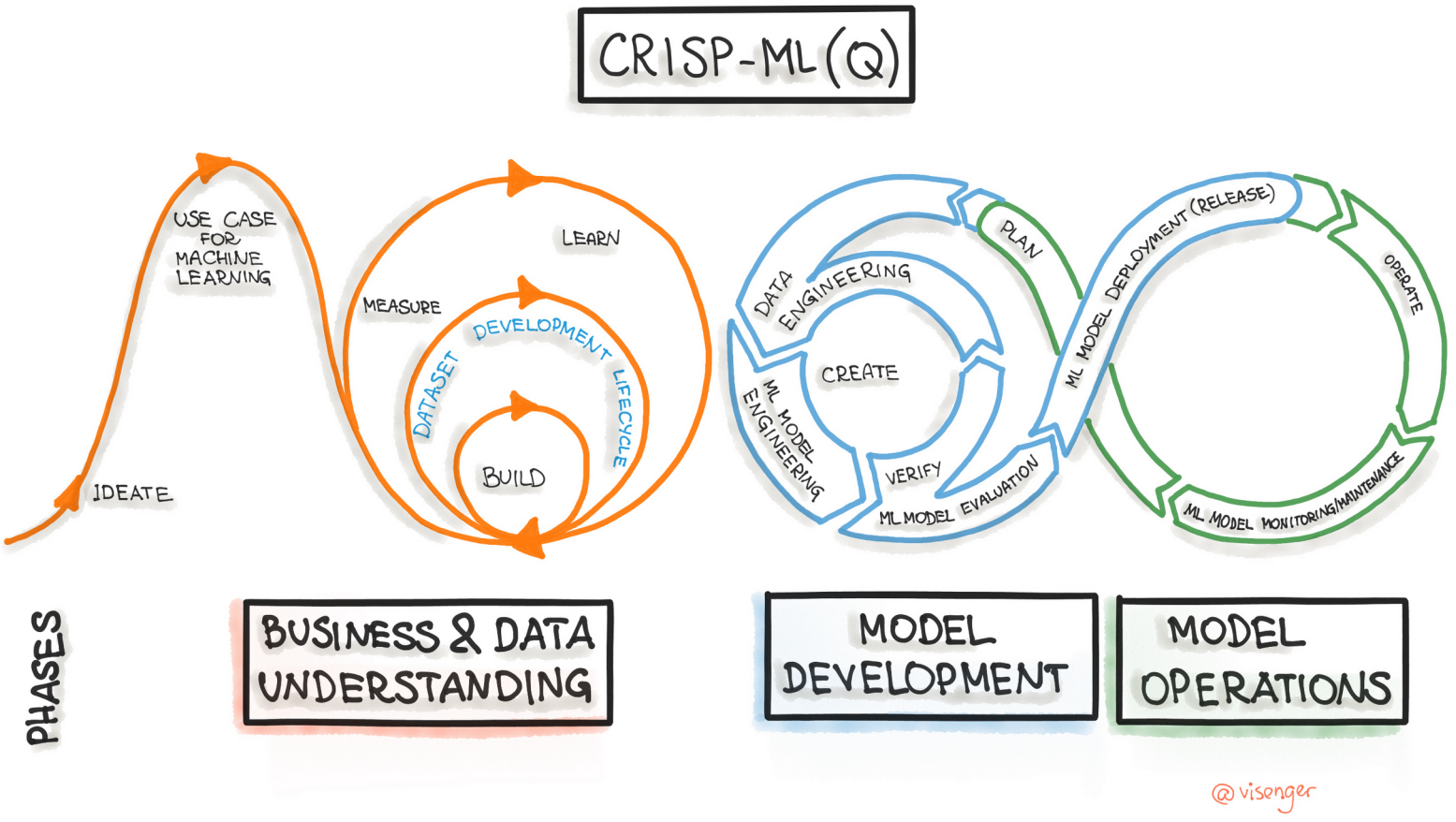
# Machine Learning in REAL Applications

Hochschule Offenburg
offenburg.university



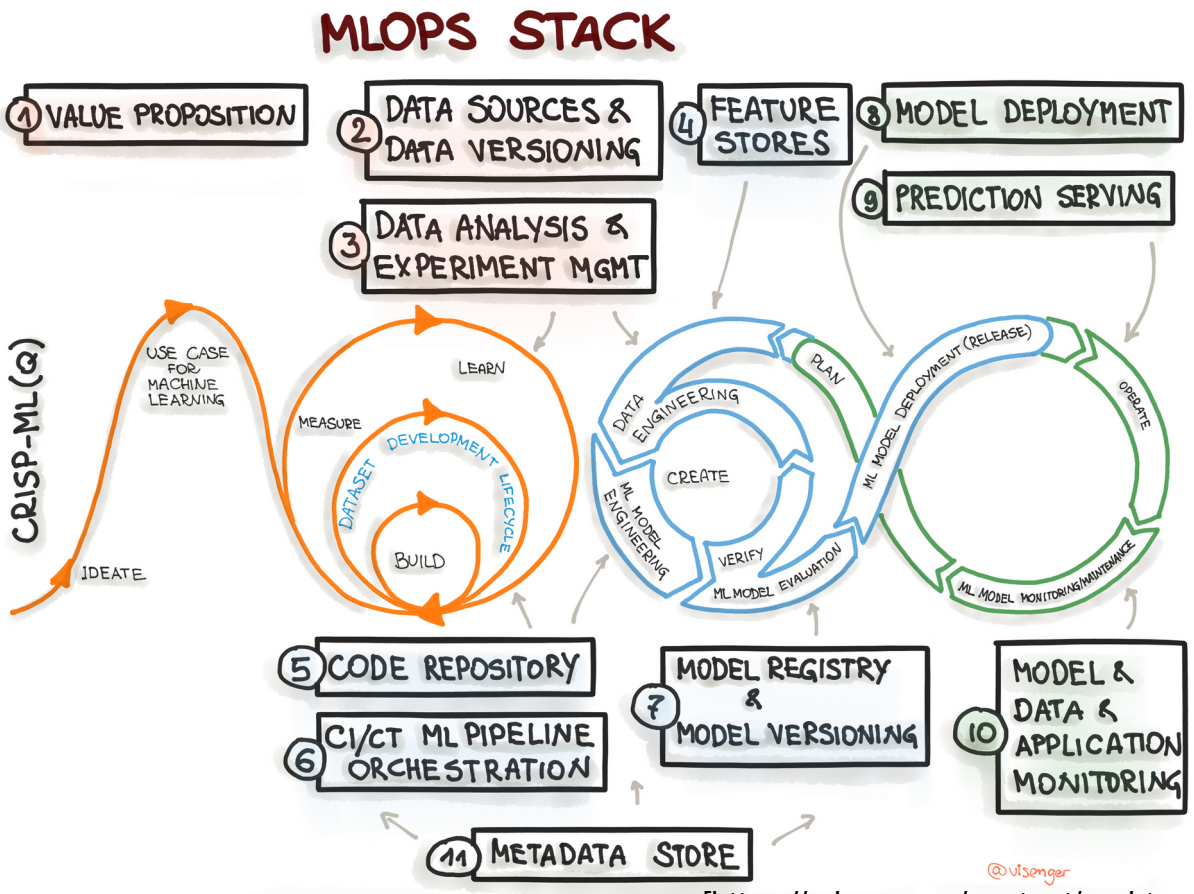

[https://ml-ops.org/content/end-to-end-ml-workflow#data-engineering]

# MLOps - Introduction




[https://ml-ops.org/content/end-to-end-ml-workflow#data-engineering]

# MLOps - Introduction




[https://ml-ops.org/content/end-to-end-ml-workflow#data-engineering]

# Machine Learning Tools + Work Flows

Hochschule Offenburg
offenburg.university

**Which programming Language did you use?**

**Training Hardware? → Cloud?**

**Which ML Frameworks?**

**Deployment Hardware?**

**How large was the Team?**

**Code Organization?**

**Model Management?**

**Which Scaling Framework?**

**Which NEW tools did you get to know?**

**Data Management?**

# MLops Overview

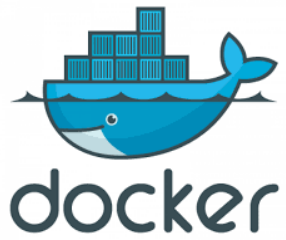
Hochschule Offenburg
offenburg.university

**MLOps**

- Data management + versioning
- Experiment organization + tracking
- AutoML
- Model management, versioning and transfer
- Model deployment
- ML monitoring

**Deploying ML**

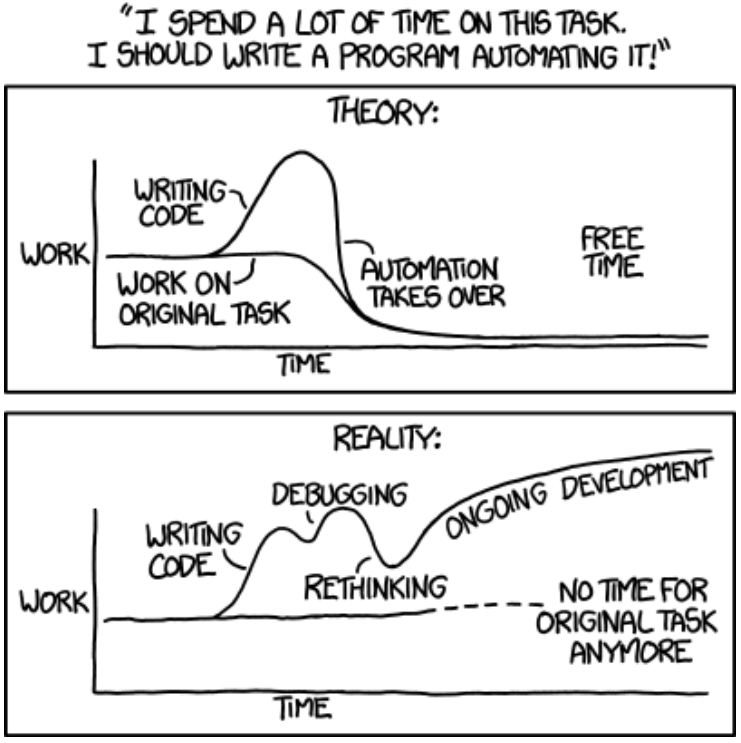
- Containerization + Cloud
- Fast Web-APIs








# MLOps Discussion




[XKCD]

# Literature

[1] free icons taken from https://www.flaticon.com

[2] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

[3] https://www.eetimes.com/author.asp?section_id=36&doc_id=1330462#