Privacy Protection within Machine Learning Models trained on Distributed Data

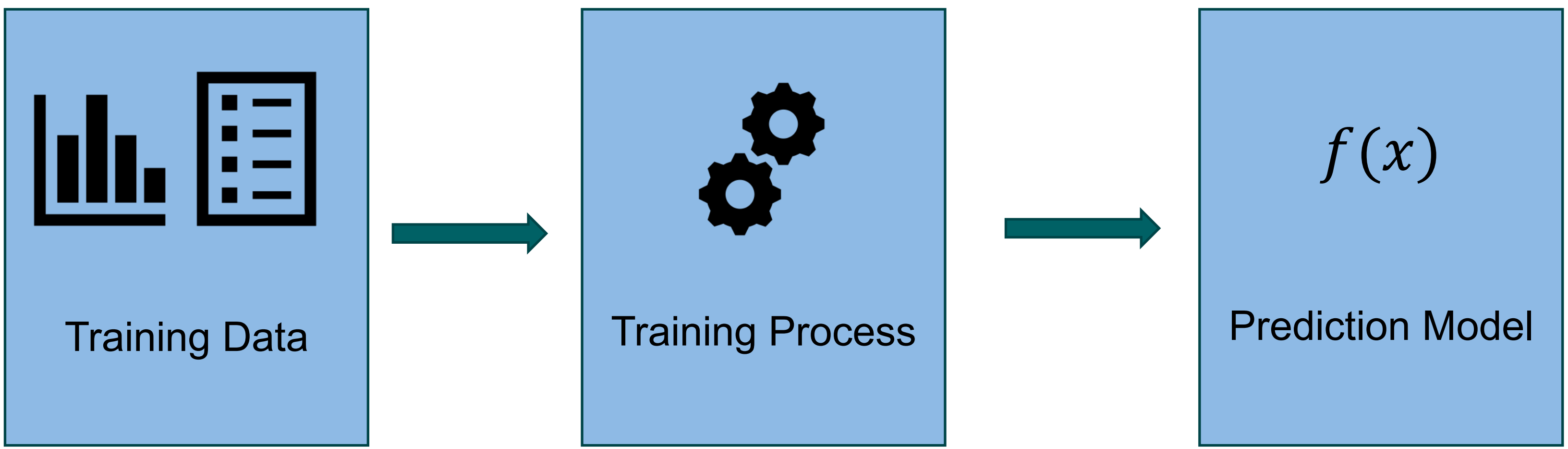
Carsten Stoffels

RWTH Aachen, Informatik 5
Lehrstuhl Prof. Decker

Motivation for Machine Learning

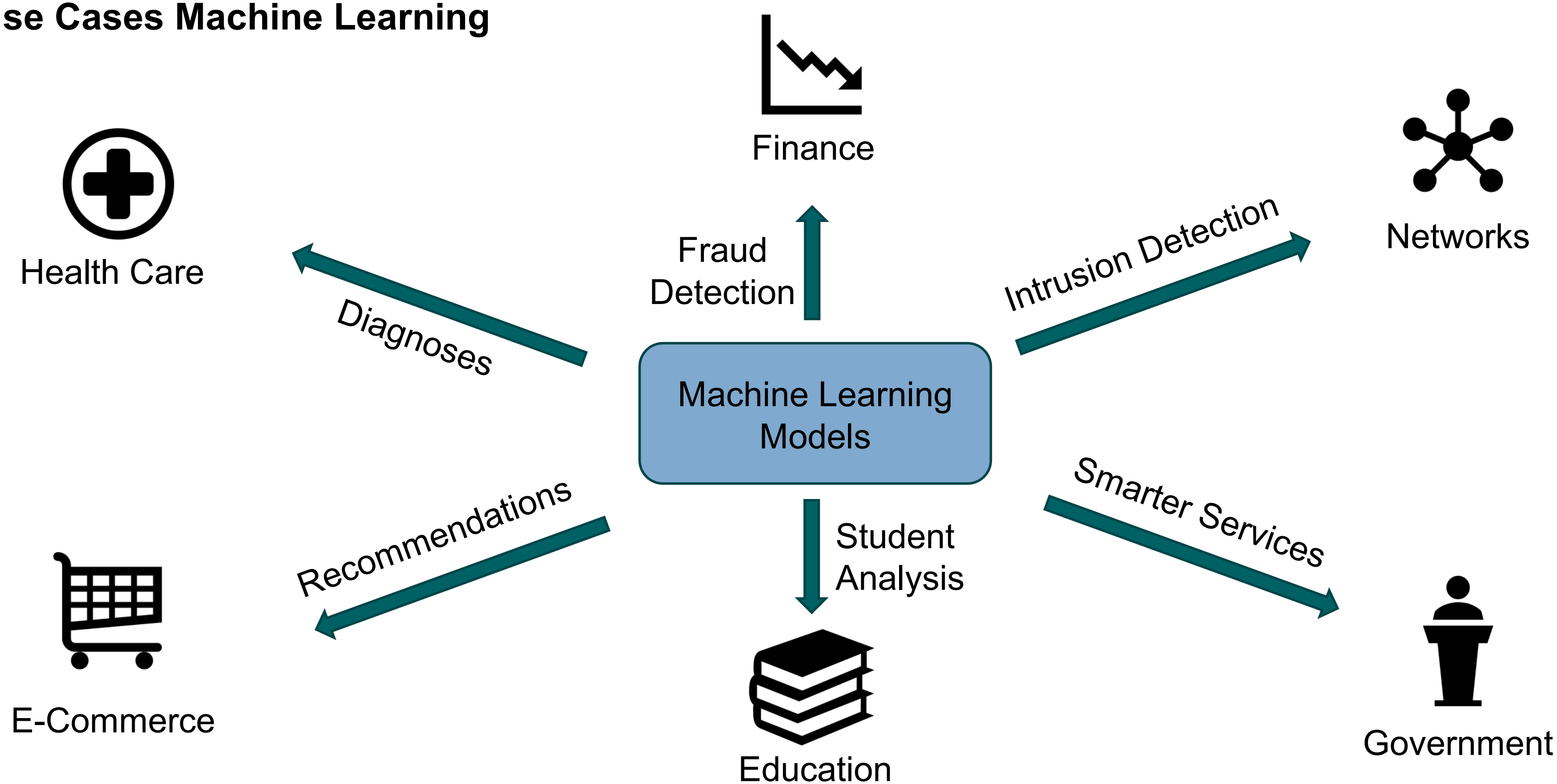
Supervised Machine Learning

“Learning to make predictions based on experience”



Motivation for Machine Learning

Use Cases Machine Learning



Motivation for Privacy in the Context of Machine Learning

Private Training Data

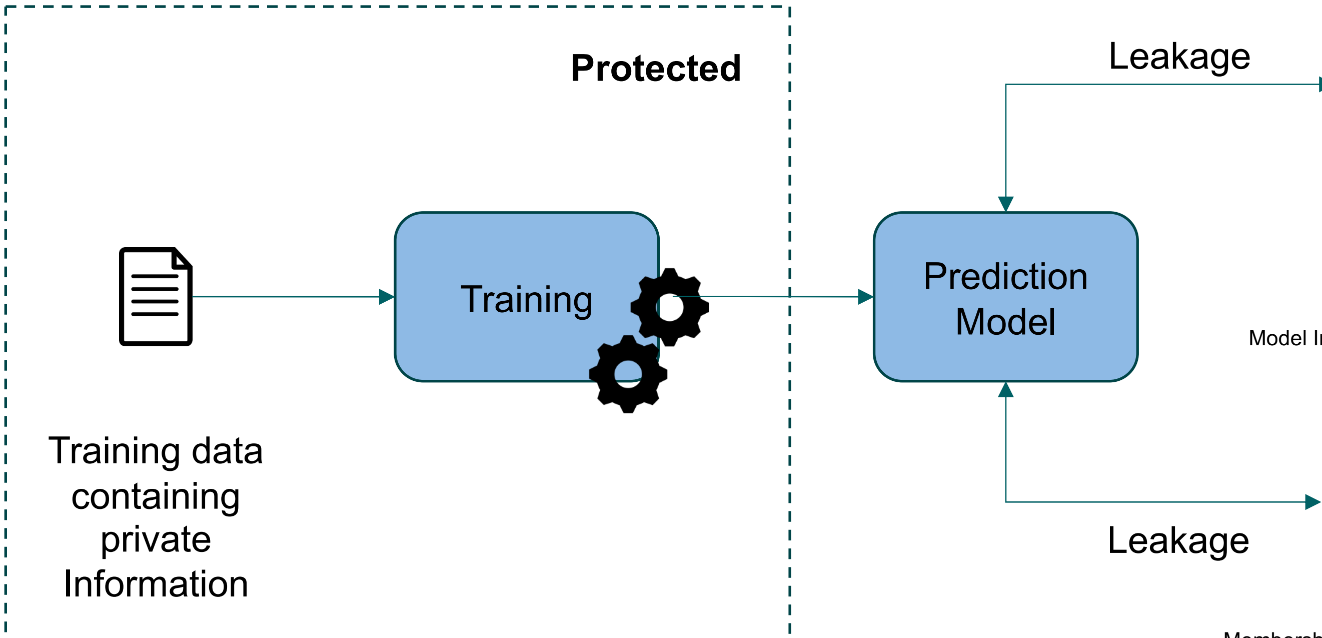
- **Training data** used for the machine learning model might be **private**
  - Finance models use information about investors
  - Diagnose prediction models use patient data

Age	Sex	Chest Pain	…	Smoke	Exercise Protocol	…	Blood pressure	…	Diagnoses
28	Male	1	…	Yes	7	…	150	…	1
73	Female	4	…	No	5	…	119	…	0
…	…	…	…	…	…	…	…	…	…

Source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

Motivation for Privacy in the Context of Machine Learning

Machine Learning Models reveal private Information



Restored Original

Source:
Model Inversion Attacks that Exploit Confidence Information
(M. Frederikson, S. Jha, T. Ristenpart)

Determine training records with 60% up to 85% accuracy using shadow models

Source:
Membership Inference Attacks Against Machine Learning Models
(R. Shokri, M. Stronati, C. Song, V. Shmatikov)

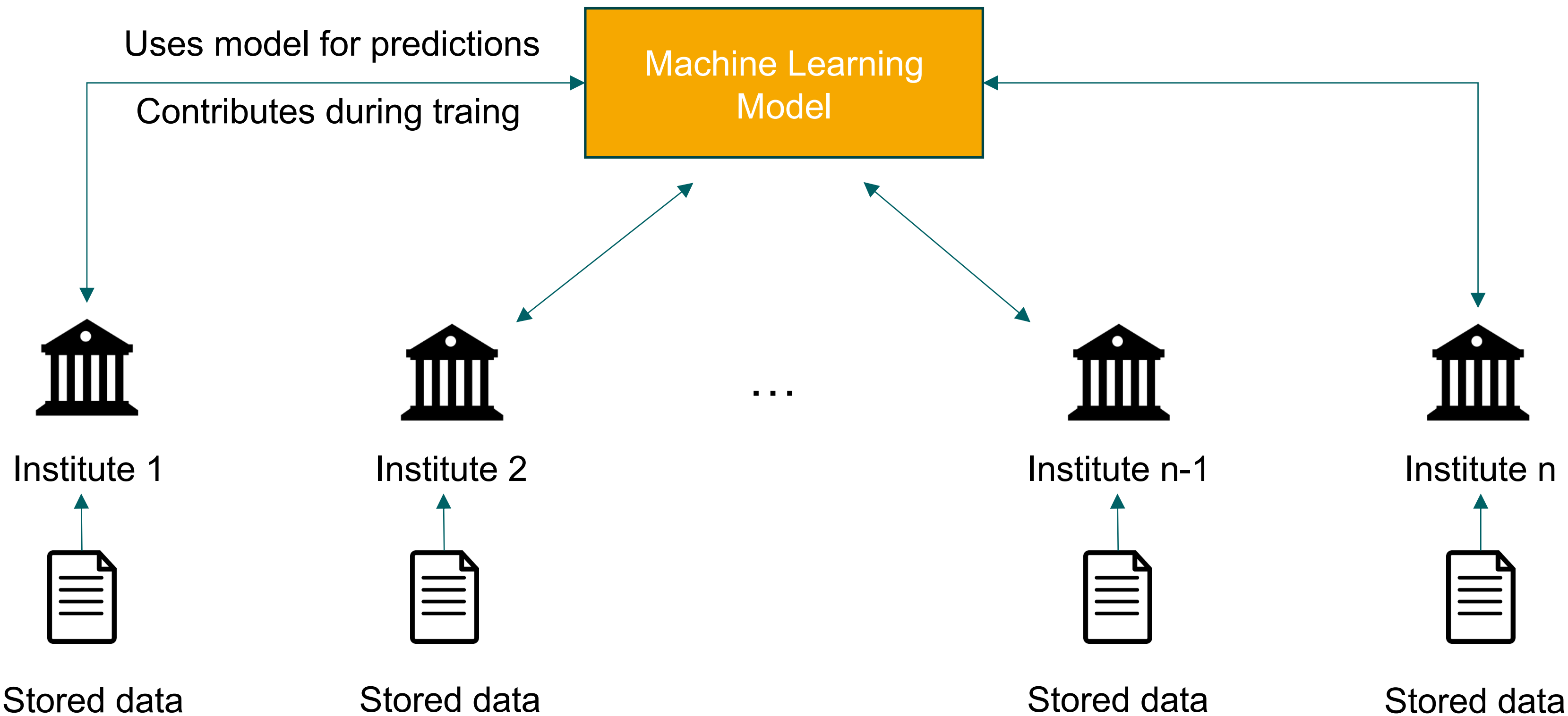
Thesis Goal

Analyze the possibilities of Machine Learning in a Private Context

- Train **machine learning models** while **protecting privacy**
- **Access** to the model **shouldn't enable** the **leakage of private information**
- Combine **distributed** training data
  - Training data splitted along several parties
  - Contribution of each party to gain a more powerful model
- Try to guarantee **privacy in each step** of the process
  - Storage of data
  - Data transfer
  - Model training
  - Accessing the model
- Develop an evaluation method to balance **privacy protection** and **performance decrease**

Use Case - Scenario

Bank Marketing Scenario



Use Case – Dataset for Evaluation

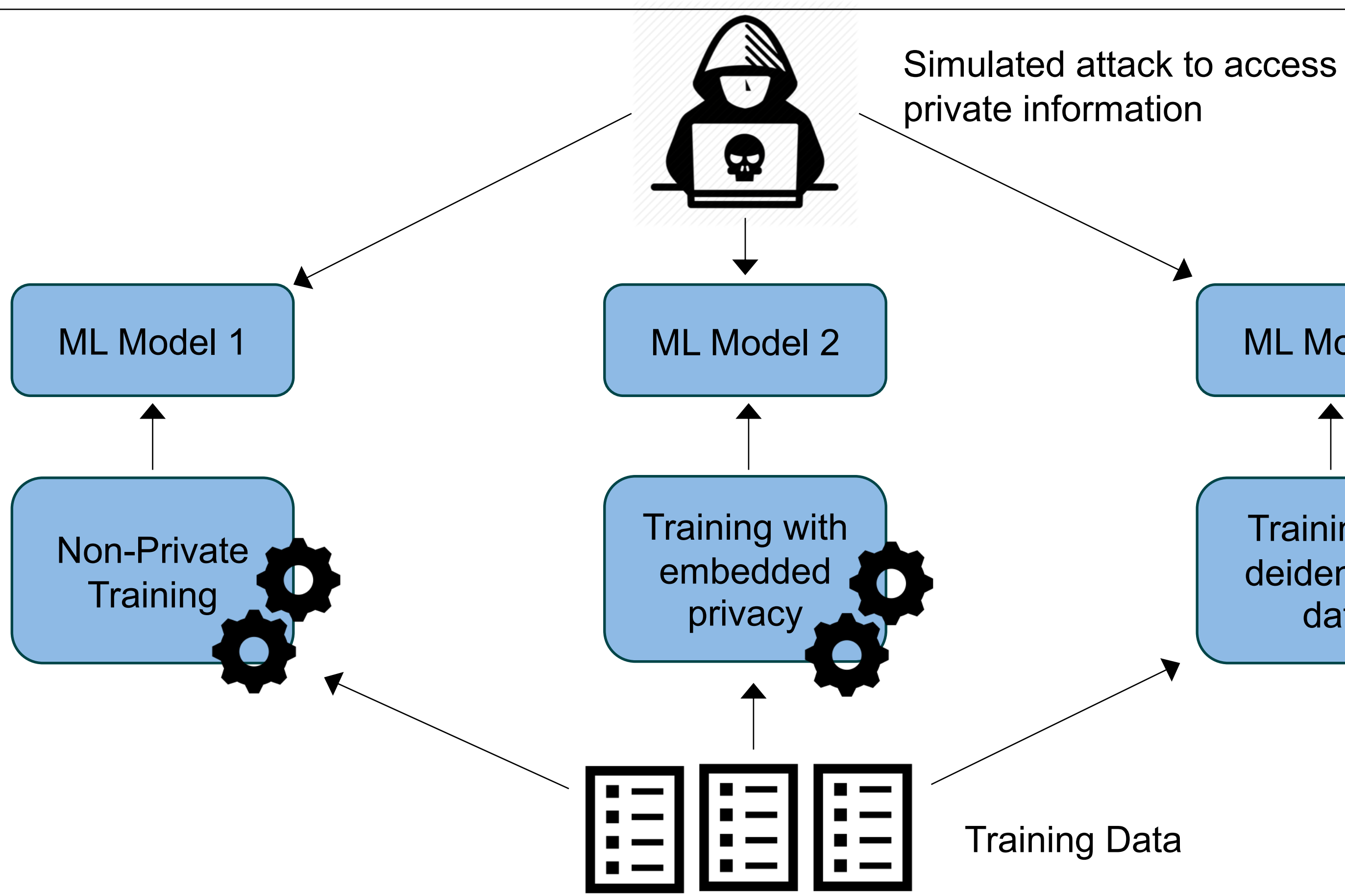
Bank Marketing Scenario

- Classification task
- Predict if the client of a bank will subscribe a term deposit
- Dataset Info[1]
  - Number of instances: 41188
  - Attributes : 20+
  - Types : Categorical, Numerical, Binary

Age	Job	Martial	Education	…	Consumer price idx	Employment variation rate	…	Outcome
56	Housemaid	Married	Basic.4y	…	94.601	01.Jan		no
49	Technican	Married	Basic.9y	…	93.994	01.Jan	…	yes

[1] https://archive.ics.uci.edu/ml/datasets/bank+marketing

Conceptual Approach



Related Work – Privacy Metrics

$k - Anonymity$

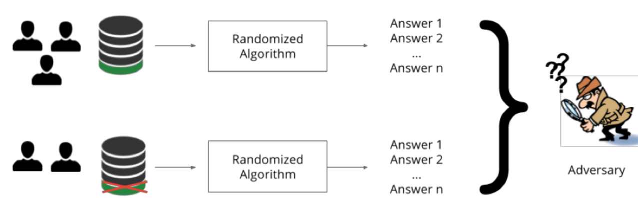
A dataset has k-Anonymity if each set of quasi-identifiers can not be distinguished from at least k-1 other entries

$k = 3$

Age	Sex	ZIP
25-30	m	53***
25-30	m	53***
25-30	m	53***

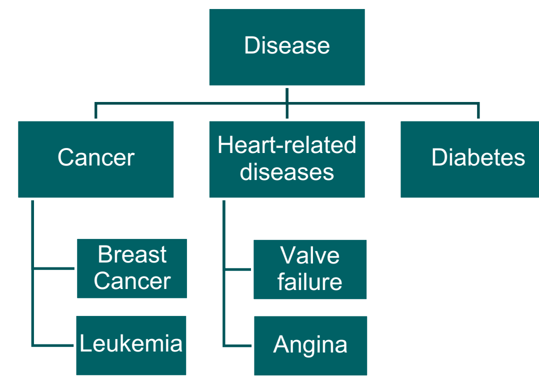
$\epsilon - differential\ Privacy$

Add noise to data, algorithms or results in order to make it indistinguishable from other records

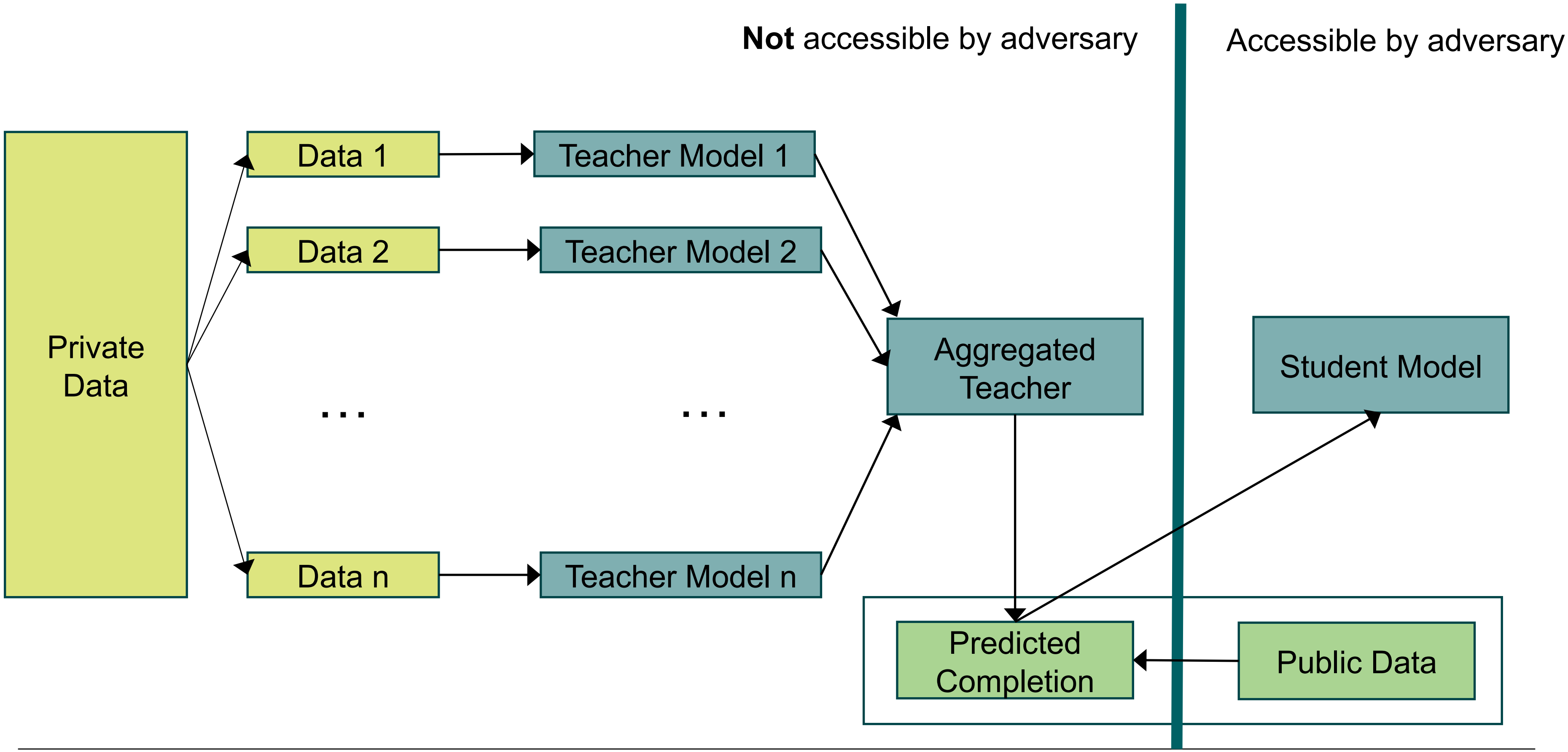


$t - closeness$

Protects against information leaks through similiar attributes within one equivalence class by calculating distances

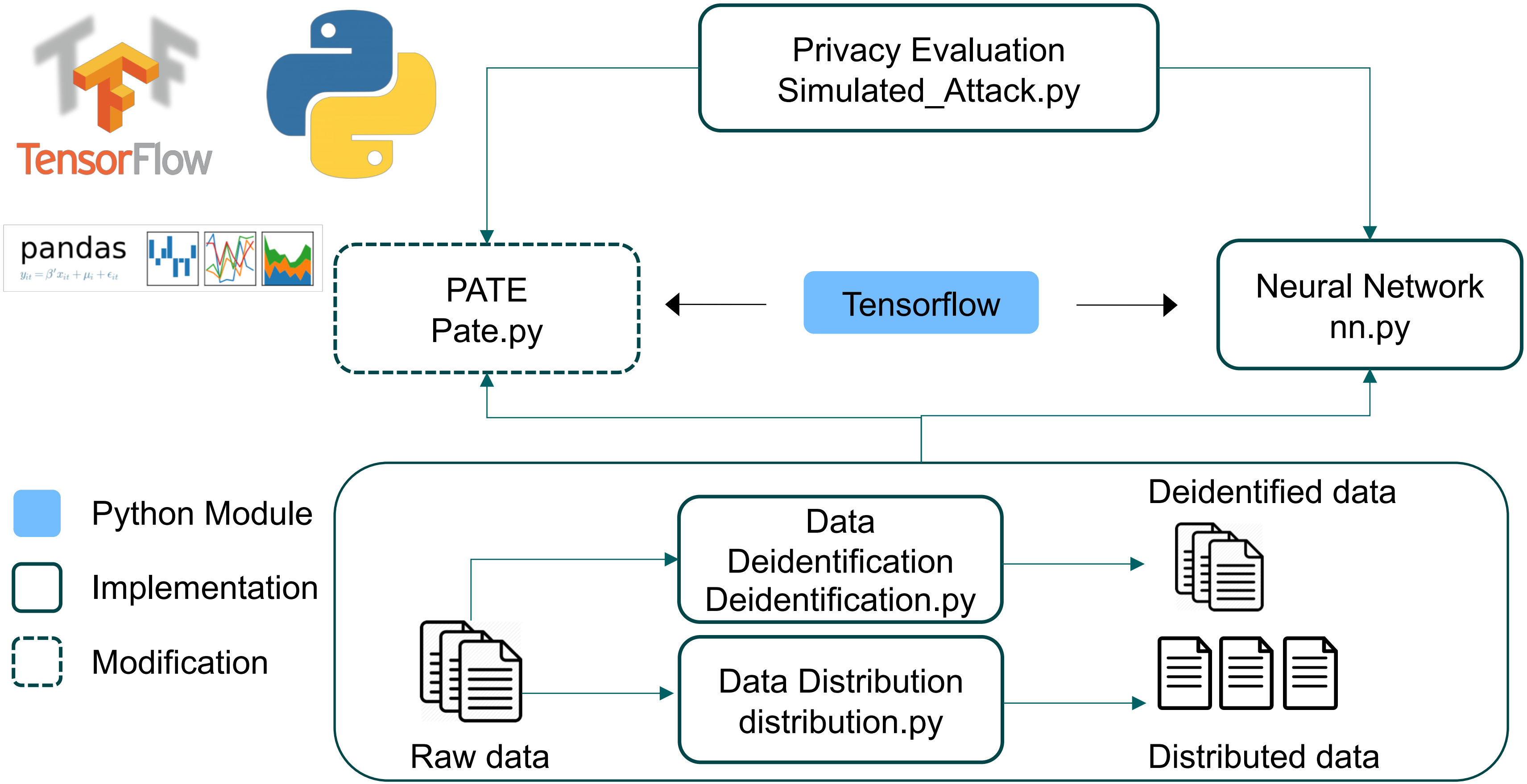


Related Work – Private Aggregation of Teacher Ensembles (PATE)



Topic; Privacy Protection within Machine Learning Models trained on Distributed Data
Name: Carsten Stoffels
Informatik 5 Information Systems, Lehrstuhl Prof. Decker

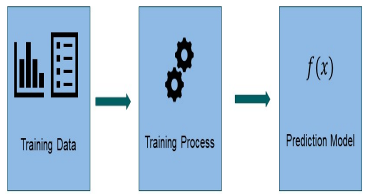
Implementation

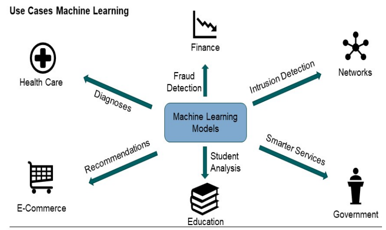


Milestones

- Milestone 1 (5 weeks):
  - Create an Evaluation framework in order to compare existing implementations of private Machine Learning
    - Train baseline model
    - Train private machine learning models
    - Adapt model inversion attack to evaluate the privacy protection
- Milestone 2 (5 weeks):
  - Analyze influence of privacy metrics to performance and privacy
  - Try to increase performance by adapting metrics, input and structure of the model
    - How to deal with Hierachies
    - Word2Vec
- Milestone 3 (4 weeks):
  - Design „best practice“ anonymization procedure to guarantee privacy and maximizing performance

Thanks for your attention!
Any Questions ?



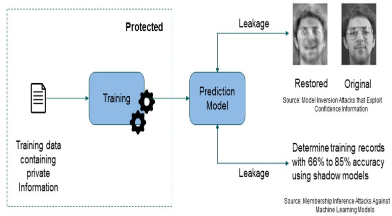


Private Training Data

- **Training data** used for the machine learning model might be **private**
  - Finance models use information about investors
  - Diagnose prediction models use patient data

Age	Sex	Chest Pain	...	Smoke	Exercise Protocol	...	Blood pressure	...	Diagnoses
28	Male	1	...	Yes	7	...	150	...	1
73	Female	4	...	No	5	...	119	...	0
...	...	...	...	...	...	...	...	...	...

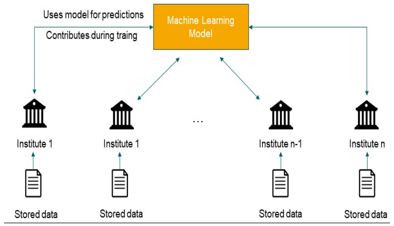
Source: https://archive.ics.uci.edu/ml/datasets/Heart+Disease



Thesis Goal

Analyze the possiblities of machine learning in a private context

- Train **machine learning models** while **protecting privacy**
- **Access** to the model **shouldn't enable** the **leakage of private information**
- Combine **distributed** training data
  - Training data splitted along several parties
  - Contribution of each party to gain a more powerful model
- Try to guarantee **privacy** in **each step** of the process
  - Storage of data
  - Data transfer
  - Model training
  - Accessing the model
- Develop an evaluation method to balance **privacy protection** and **performance decrease**



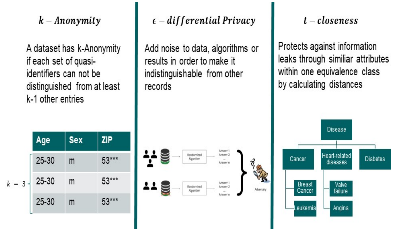
Use Case – Dataset for Evaluation

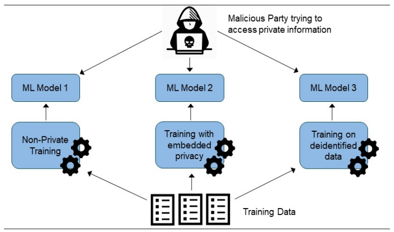
Bank Marketing Scenario

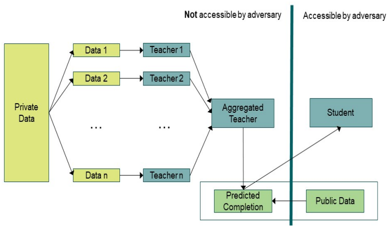
- Classification task
- Predict if the client of a bank will subscribe a term deposit
- Dataset Info[1]
  - Number of instances: 41188
  - Attributes : 20+
  - Types : Categorical, Numerical, Binary

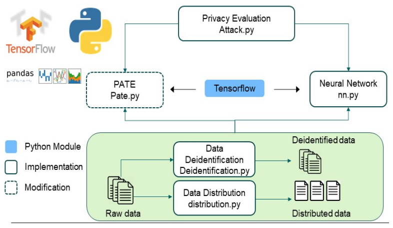
Age	Job	Martial	Education	...	Consumer price idx	Employment variation rate	...	Outcome
56	Housemaid	Married	Basic.4y	...	94.601	01.Jan		no
49	Technican	Married	Basic.9y	...	93.994	01.Jan	...	yes

[1] https://archive.ics.uci.edu/ml/datasets/bank+marketing









- Milestone 1 (5 weeks):
  - Create an Evaluation framework in order to compare existing implementations of private Machine Learning
    - Train baseline model
    - Train private machine learning models
    - Adapt model inversion attack to evaluate the privacy protection
- Milestone 2 (5 weeks):
  - Analyze influence of privacy metrics to performance and privacy
  - Try to increase performance by adapting metrics, input and structure of the model
    - How to deal with Hierachies
    - Word2Vec
- Milestone 3 (4 weeks):
  - Design „best practice" anonymization procedure to guarantee privacy and maximizing performance

RWTHAACHEN UNIVERSITY

Appendix - Requirements

- **Training Speed**
  - How **long** does it take to **train** the model
  - Fast enough for real-world usage?

- **Privacy Protection**
  - Privacy protection should be **included in every step** of the procedure

- **Model Performance**
  - How good do the **private models** perform **compared** to the **non-private version**

- **Data Distribution**
  - **Privatly combine distributed data** trying to get a more powerful model

- **Amount of Data**
  - **Data needed** to achieve good performance
  - Training data is **expensive**

Appendix – Model Inversion Attack

Attacker has access to :

$(\boldsymbol{x_2}, \ldots, \boldsymbol{x_n}, \boldsymbol{y})$ Set of known attributes about the target (including label y) , excluding one private Attribute ("background knowledge")

$\boldsymbol{f}(\boldsymbol{x})$ Prediction Model

Attacker wants to learn about :

$\boldsymbol{x_1}$ Missing (unknown) private attribute

1: for each possible value v for x_1 $\boldsymbol{do}$
2: $x' = (v, x_2, \ldots, x_n)$
3: $r_v = err\big(y, f(x')\big) * \prod p_i(x_i)$
4: Return $\arg\max_v r_v$

Appendix - Project Plan

- Baseline Neural Network : 2 weeks
- Train first private models : 2 weeks
- Writing thesis : 1 week
- Implement Model Inversion Attack : 2 weeks
- Writing thesis : 1 week
- Additional privacy metrics + model improvenent : 3 weeks
- Writing thesis : 1 week
- Evaluation : 2 weeks
- Writing thesis : 1 week
- Buffer: 3 weeks

Appendix – Model Inversion Attack

Model Inversion Attack

- Given
  - $(x_2, \ldots, x_n, y)$ Set of known attributes about the target (including label y) , excluding one sensitive Attribute
  - f Machine Learning Model
- Output
  - x_1 Prediction for missing sensitive attribute
- Algorithm

1: for each possible value v for x_1 $\boldsymbol{do}$
2: $x' = (v, x_2, \ldots, x_n)$
3: $r_v = err\big(y, f(x')\big) * \prod p_i(x_i)$
4: Return $\arg\max_v r_v$