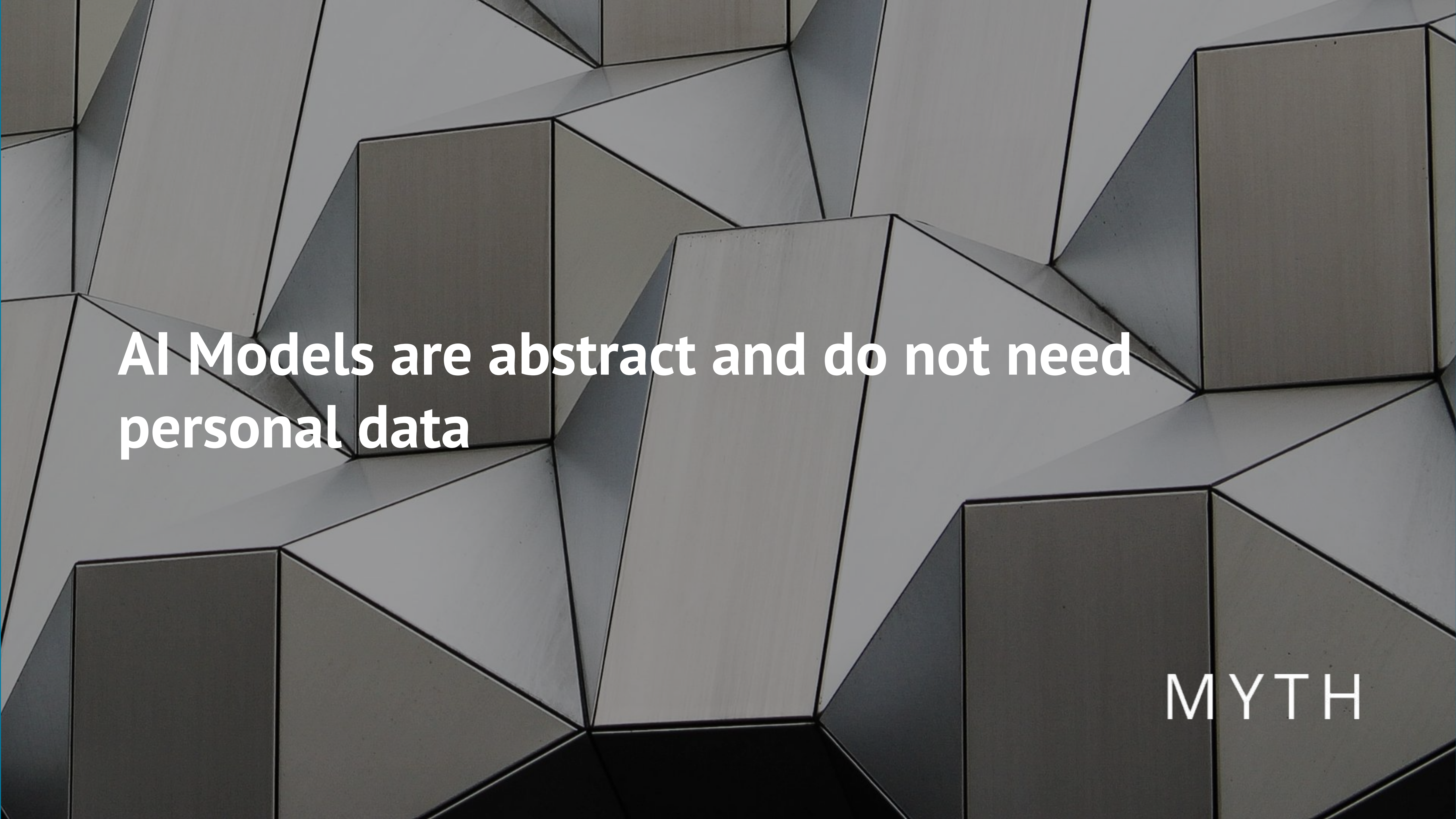
Why, AI.

BUSTED

MYTH



AI Models are abstract and do not need personal data

MYTH

Chair of Legal Informatics @Saarland University

Chair of Legal Informatics
Saarland University

- Computer science and law research group in the Faculty of Law (and, as a secondary membership, the Faculty of Mathematics and Computer Science)
  - Also: Association with the CISPA Helmholtz Center for Information Security and Senior Fellow of the German Research Institute for Public Administration
  - Head: Christoph Sorge

- Interdisciplinary and international team with backgrounds in law/computer science/related subjects
- Focus areas: Intersection of computer science and law, Privacy by Design, Machine learning on legal texts
- Third-party funding in cooperation with research institutions, companies, government authorities and public prosecutors

Introduction

- Focus here: Machine learning (main driver of AI progress in recent years)
  - In particular: Supervised machine learning which learns from training data and is evaluated using test data

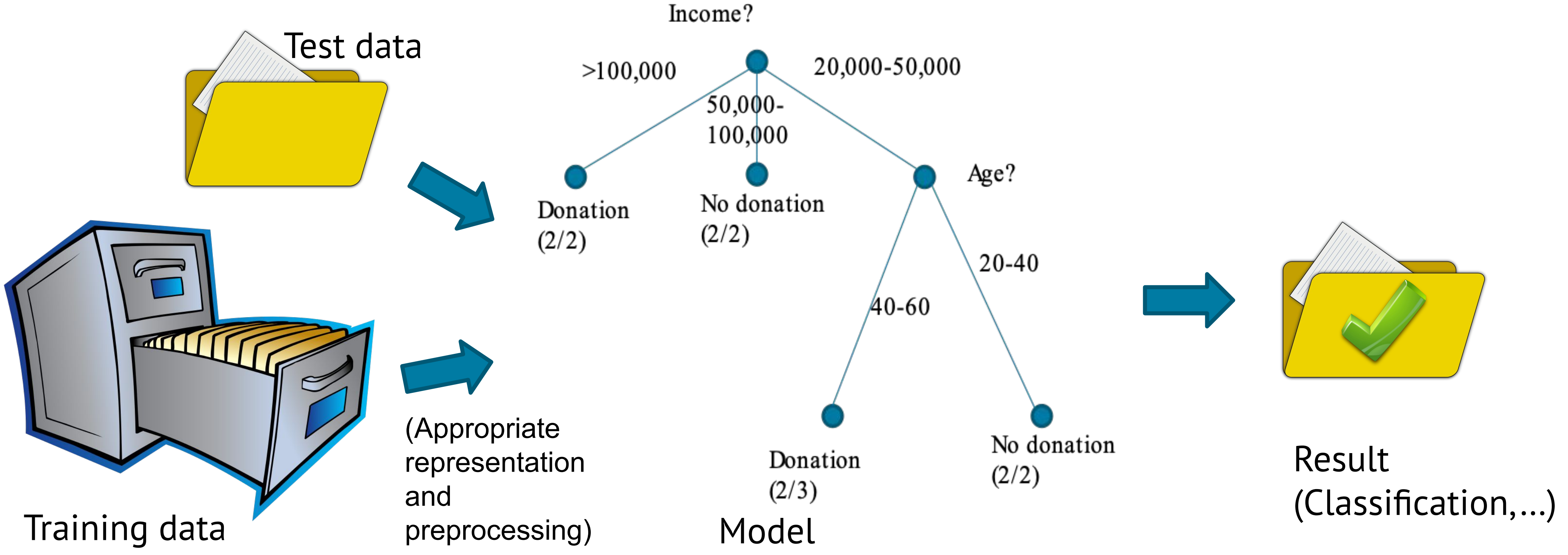


Image sources:
https://openclipart.org/detail/202408/raseone-file-cabinet
https://openclipart.org/detail/74029/folder
https://openclipart.org/detail/159733/green-tick

Introduction

- Focus here: Machine learning (main driver of AI progress in recent years)
  - In particular: Supervised machine learning which learns from training data and is evaluated using test data

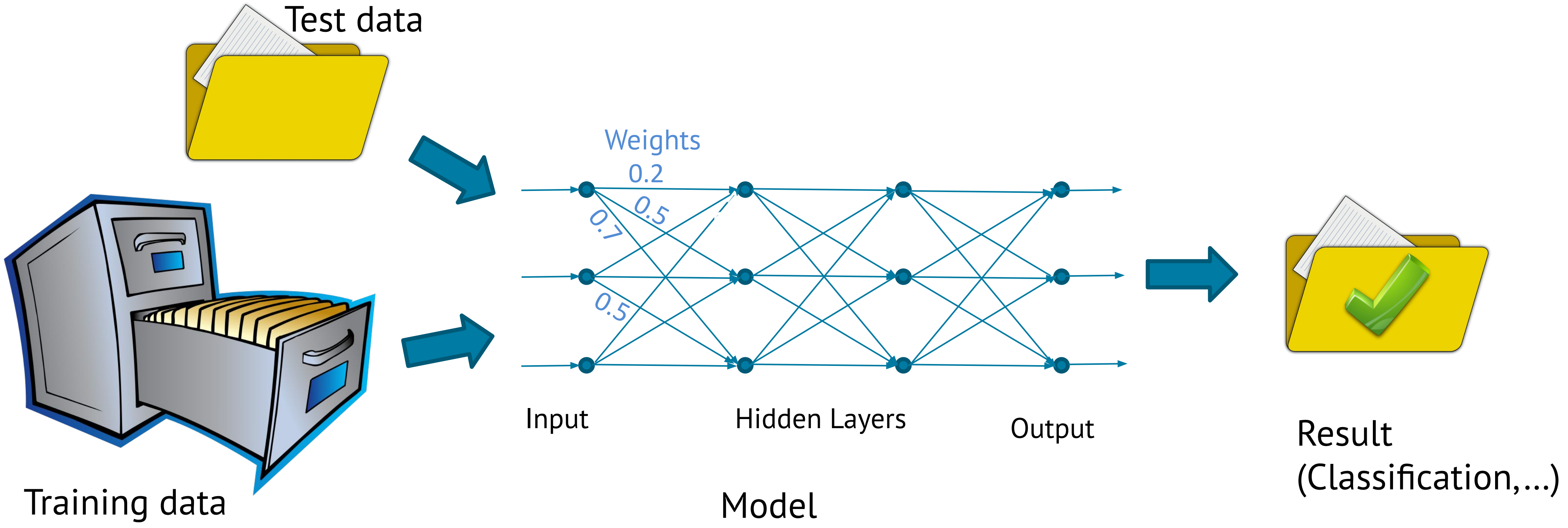


Image sources:
https://openclipart.org/detail/202408/raseone-file-cabinet
https://openclipart.org/detail/74029/folder
https://openclipart.org/detail/159733/green-tick

Machine learning models

- Common (implicit) assumption: Even when learning from personal data, the resulting model is an aggregation/abstraction, which no longer contains personal data

- Reasoning:
    - Only statistical information is required for classification / prediction tasks
    - Machine learning models are typically smaller than the training data set
      → not all information can be contained in the model
    - Overfitting is bad for machine learning performance: We want to learn patterns in a type of data—not information about some person in the training data set

Machine learning models

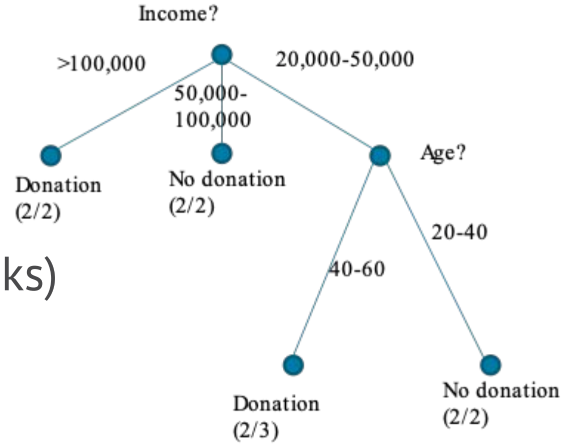
- Common (implicit) assumption: Even when learning from personal data, the resulting model is an aggregation/abstraction, which no longer contains personal data

- De facto: Common evaluation methods of machine learning models verify the accuracy and expressiveness of the model, but not
  - whether there is “too much data” in the model
  - whether irrelevant attributes are considered in the model
  - whether model allows conclusions about individual data sets/persons from the training data

→ Usually no guarantee and no verification about (absence of) personal data in the model

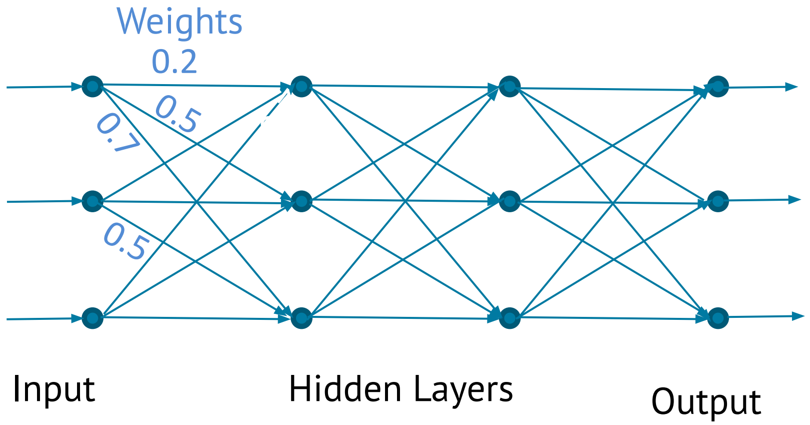
Machine learning models

- Common (implicit) assumption: Even when learning from personal data, the resulting model is an aggregation/abstraction, which no longer contains personal data



- Simple models may allow manual verification (or at least: plausibility checks) concerning the presence of personal data

- Neural networks, deep learning: Hard to understand even for experts; no intuitive way of checking for personal data



Machine learning models

- Common (implicit) assumption: Even when learning from personal data, the resulting model is an aggregation/abstraction, which no longer contains personal data
- Computer science research shows: This is, in many cases, incorrect
- Example: Shokri et al. 2017, "Membership Inference Attacks Against Machine Learning Models"
    - Common machine learning algorithms allow an attacker to conclude whether a specific data set (→ often: data about a specific person) was part of the training data or not
    - Finding has been confirmed in numerous other publications

Machine learning models

- Common (implicit) assumption: Even when learning from personal data, the resulting model is an aggregation/abstraction, which no longer contains personal data

- Computer science research shows: This is, in many cases, incorrect

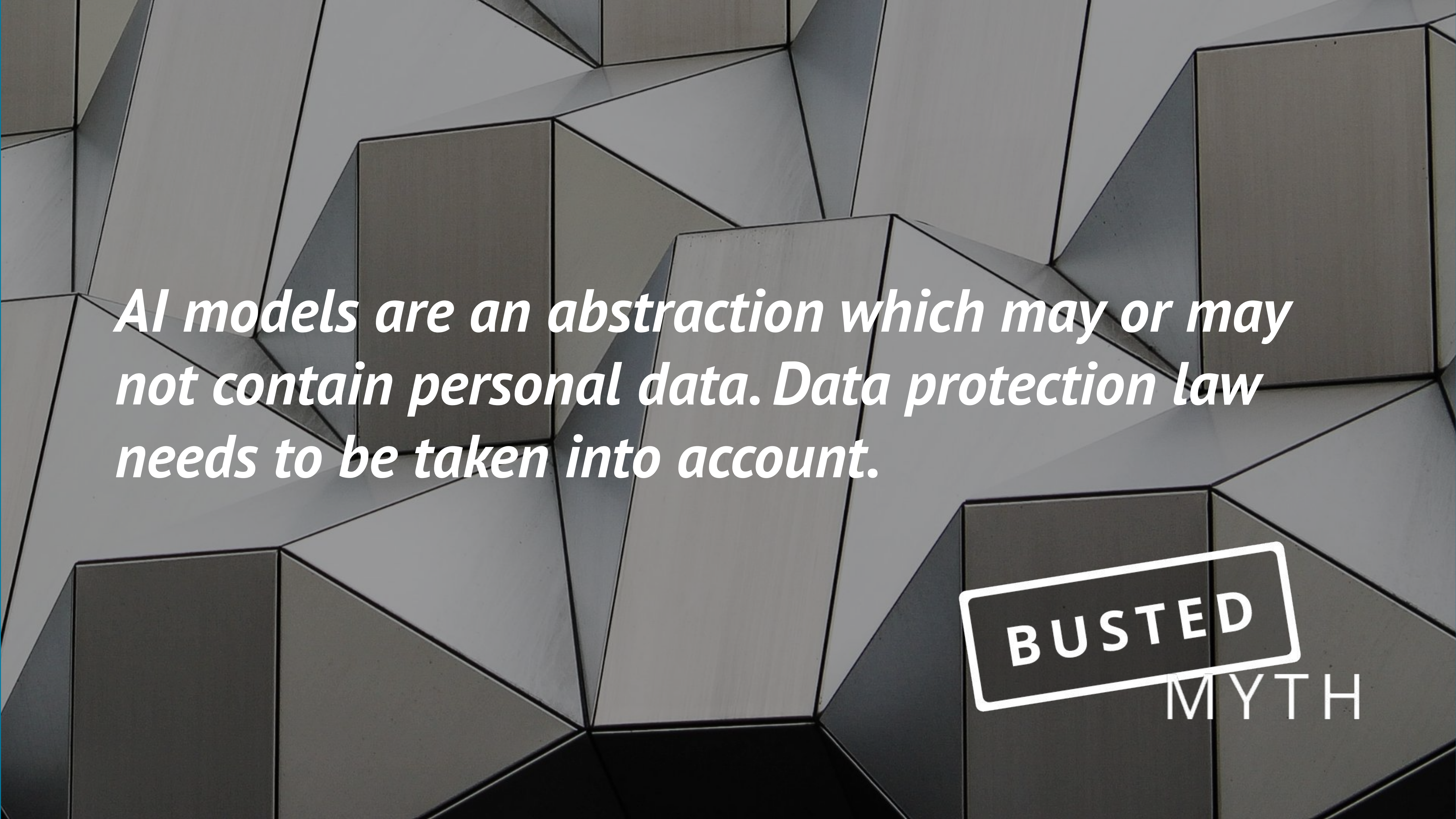
- More privacy threats have been shown in different application scenarios (cmp. Al-Rubaie/Chang 2019)
    - Private data in the clear in case of outsourced computation
    - Reconstruction attacks: reconstruction of raw data from feature vectors
    - Model inversion attacks: reconstruction of feature vectors or raw data resembling the original training data – even if feature vectors are not part of the model

Countermeasures

- Privacy-preserving machine learning
  - Active research field
  - No definite “one size fits all” solution
- “Simplest” approach: Use already anonymized training data
  - Independent from actual machine learning algorithm
  - Might lead to suboptimal results
- Precise method to be used: Dependent on application requirements

Conclusion

- Training a machine learning model ≠ anonymization
- Well-known result in the privacy research community
- Possibly less well-known result among practitioners
- Note: GDPR applies to processing of personal data
  - In computer science: Relatively short delay between research result and practical applicability → state of the art in research should be taken into account
- Countermeasures can be cumbersome and/or affect quality of results, but might be legally required



AI models are an abstraction which may or may not contain personal data. Data protection law needs to be taken into account.

BUSTED

MYTH