Towards Machine Learning Models That Can Forget

Katie Hawkins, katie.hawkins@bristol.ac.uk
Dr. Sana Belguith, sana.belguith@bristol.ac.uk

May 2022—May 2025

- Newly emerged machine learning methods have revolutionized industries including smart healthcare, financial technology and surveillance systems
- Driven by the collection of vast quantities of data concerning individuals and their social relations; raising questions surrounding privacy...

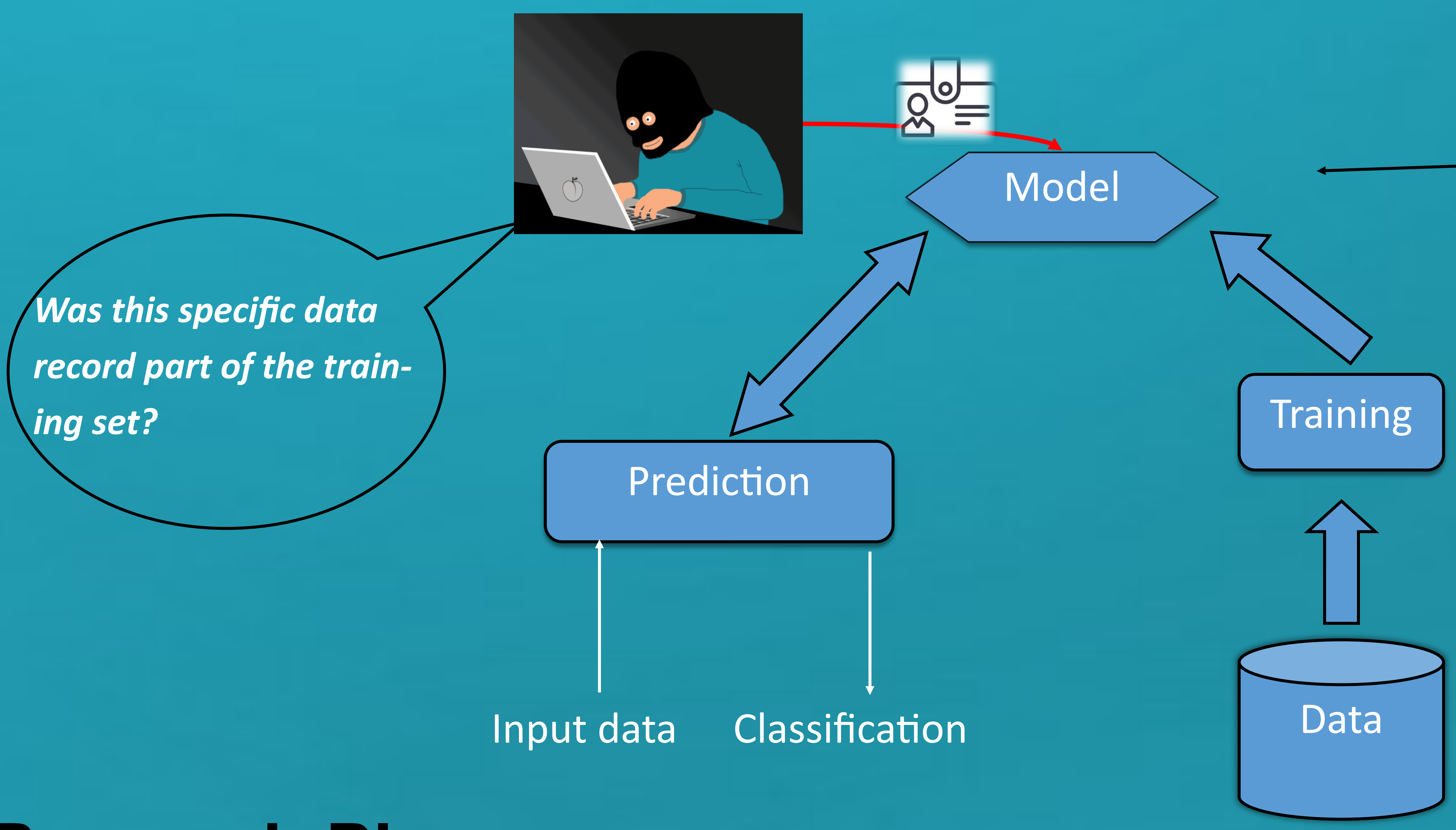
But what if you decide to withdraw your data?

- Privacy regulation such as the GDPR's Right to be Forgotten gives you the ability to opt out of processing your data
- Research has shown that machine learning algorithms memorise training data —so to ensure your data has been withdrawn, machine learning models must **delete** this personal data within the training sets as well the models trained over them
- BUT to delete training data after each request would mean retraining the model from scratch each time—this is not computationally efficient and impossible for large systems...Is data deletion the best solution to prevent such privacy attacks and erase training data?

RIGHT TO BE FORGOTTEN

An individual has the right to have their personal data erased if:

- ❖ The personal data is no longer necessary
- ❖ An individual withdraws their consent to publish the data
- ❖ An individual objects to processing their data
- ❖ An organisation processed an individuals personal data unlawfully



Why must we delete training data?

This is known as a **Membership inference** attack:

- The ability to infer information about the models training dataset, using the model output

Other attacks include **inversion** attack:

- Using a models output on a known part of the training data to infer something further about this input

Research Phases

PHASE 1— Developing the motivation for Data Deletion with Machine Learning Algorithms

Working alongside Bristol Law School to review existing literature that focuses on the EU GDPR's Right to be Forgotten and its context in machine learning.

In particular, outlining how must we proceed when an individual retracts permission to use special category data (e.g. biometric data) that has been used as part of the training process of a model.

PHASE 2— Proposing a Data Deletion Solution

This phase evaluates existing techniques for privacy in supervised machine learning algorithms. This includes:

1) Anonymisation techniques & existing Data Deletion approaches

Examining current solutions such as differential privacy and the use of synthetic data, as well as state-of-the-art solutions for data deletion.

Outcome: to understand existing limitations in the application to machine learning.

2) Propose data deletion solution that addresses the highlighted limitations and meets technical requirements in Phase 1.

PHASE 3—Evaluation and Enhancement

Concentrates on a quantitative-based evaluation of the developed data deletion solution.

Experimental simulations with existing and publicly available datasets such a large scale image datasets MNIST, OpenImages and CIFAR-10.

Intended Outcomes

- Address individuals' privacy rights within machine learning.
- Investigate current limitations in addressing such privacy rights to develop a novel solution.
- Demonstrate the effectiveness of the solution against the state-of-the-art

University of BRISTOL
UKRI Engineering and Physical Sciences Research Council
UNIVERSITY OF BATH