Introduction

Theory of machine learning

- Goal:
  - Learn a mathematical function to make predictions on unseen test data, based on given training data of inputs and outputs, without explicit programmed instructions on how to perform the task.
- Learning functional relationships between inputs and outputs builds on **methods of mathematical optimization**. [Bottou et al., 2018]
- Important twist: **Optimize prediction performance in expectation**, thus enabling **generalization to unseen data**.

  [von Luxburg and Schölkopf, 2011, Kawaguchi et al., 2020, Shen et al., 2021]

Practical workflow of machine learning experiments

- The **train-dev-test** paradigm:
  - Optimize a model on given training data,
  - tune meta-parameters on development data,
  - evaluate the model using a standard automatic evaluation metric on benchmark test data.
- Assume data splits to represent i.i.d. samples from a representative data population.
- Define SOTA by best achieved result, publish code, and report corresponding meta-parameter settings.

Digging Deeper: Methodology of Deep Learning

Inherent non-determinism of deep learning

- Non-convex optimization under randomness in weight initialization, dropout, data shuffling and batching.

  [Clark et al., 2011, Dauphin et al., 2014, D'Amour et al., 2020]

- Non-determinism due to variations in architecture, meta-parameter settings, pre-processing and data splits.

  [Lucic et al., 2018, Henderson et al., 2018, Post, 2018, Gorman and Bedrick, 2019, Søgaard et al., 2021]

- Non-determinism due to differences in available computational budget. [Strubell et al., 2019, Dodge et al., 2019]

The Quest for Replicability

Replicability = reproducibility of SOTA results under exactly same circumstances

- Quest for replicability fostered by sharing data, code, meta-parameter settings, e.g., on `paperswithcode.com`

  [Pineau et al., 2021, Heil et al., 2021, Lucic et al., 2022]
- **Non-determinism in deep learning is spoiling the party**
  - Slight changes in training settings can reverse relations between baseline and SOTA. [Reimers and Gurevych, 2017, Melis et al., 2018]
  - Large-scale SOTA results may be impossible to replicate, even if code and data are shared [Kaplan et al., 2020, Chowdhery et al., 2022].

Crisis Management

- Does AI face a **replicability crisis**? [Hutson, 2018]
- Or is **replicability uninteresting and not worth having**?

  [Drummond, 2009, Belz et al., 2021]
- ➡ Quest for replicability of SOTA result under exactly same circumstances is **asking the wrong question!**

An Alternative: Inferential Reproducibility

Inferential reproducibility

- Question: Can qualitatively similar conclusions be drawn from an independent replication of a study? [Goodman et al., 2016]
- **Inferential reproducibility in machine learning**:
  - Which conclusions about comparison SOTA-baseline can be drawn **across data properties** under **variability of meta-parameters**?
  - Inferential reproducibility is **interesting feature** of non-deterministic machine learning, **not a bug** that needs to be resolved.
  - :: **Training reproducibility ::** Ability to **duplicate prior results** using the same means as used in the original work.

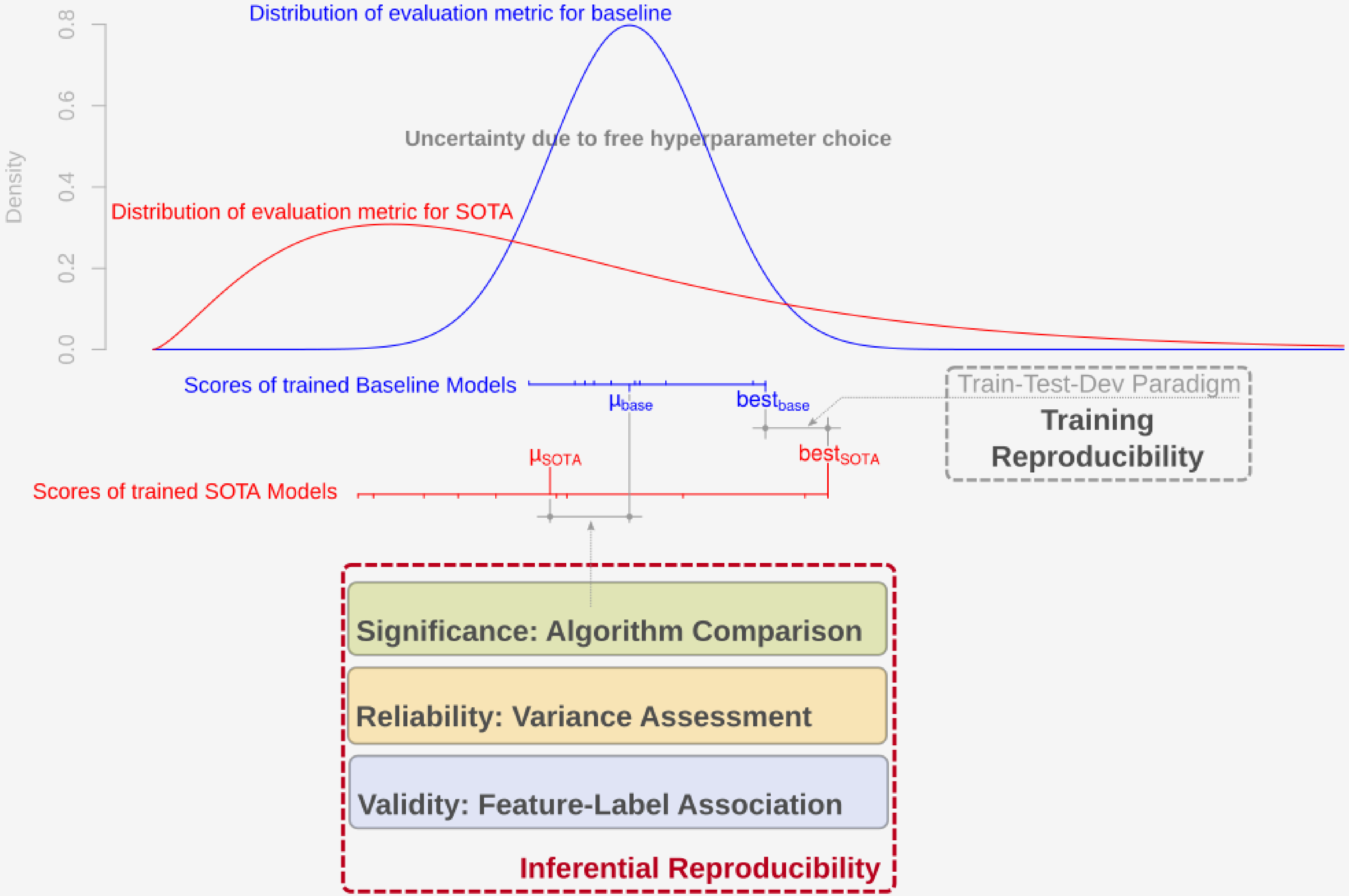
    [Leventi-Peetz and Östreich, 2022]

Towards Inferential Reproducibility

Questions of theory of science to analyze inferential reproducibility

- **Significance** – how likely is it that a result difference between two models (incorporating sources of variation) is due to chance?
- **Reliability** – how consistent is a performance evaluation if replicated under variations of meta-parameters (or varying data properties)?
- **Validity** – does a machine learning model predict what it purports to predict?

Towards Inferential Reproducibility


Distribution of evaluation metric for baseline
Uncertainty due to free hyperparameter choice
Distribution of evaluation metric for SOTA
Density
0.0
0.2
0.4
0.6
0.8
Scores of trained Baseline Models
μ_{base}
$best_{base}$
μ_{SOTA}
$best_{SOTA}$
Scores of trained SOTA Models

Train-Test-Dev Paradigm

Training Reproducibility

Significance: Algorithm Comparison

Reliability: Variance Assessment

Validity: Feature-Label Association

Inferential Reproducibility

Towards Inferential Reproducibility

Statistical methods as analysis tools

- **Significance**:
  - **Training reproducibility**: Replicability of best SOTA result on benchmark testset.
  - **Inferential reproducibility**: Reproducibility of experiment under variations of meta-parameters and varying data properties.
- **Reliability**:
  - **Variance decomposition**: Decompose variance into components due to variations in meta-parameters and data properties.
  - **Reliability coefficient**: Calculate amount of variance attributable to objects of interest.
- **Validity**: Further reproducibility problems caused by dataset biases.

Statistical models for significance, reliability, and validity

- Interpretable statistical models linear mixed effects models (**LMEMs**), generalized additive models (**GAMs**), trained on predictions of machine learning models. [Wood, 2017]
- **Significance testing under data/meta-parameter variation** by likelihood ratio test on nested LMEM models.
- **Reliability coefficient** and **variance component analysis** of meta-parameter and data effect of LMEM models.
- **Validity** test exposing circularity by **GAM feature shape analysis**.