
EFS

Nutzung von Unsicherheiten von KI-Systemen als Teil eines systematisierten Entwicklungsprozesses

Jonas Schneider

Elektronische Fahrwerksysteme GmbH

e-mail: jonas.schneider@efs-auto.com

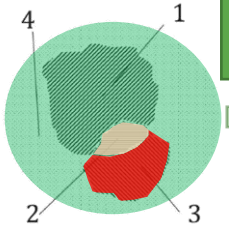
KI ABSICHERUNG
Safe AI for Automated Driving

Gefördert durch:
Bundesministerium für Wirtschaft und Klimaschutz
aufgrund eines Beschlusses des Deutschen Bundestages

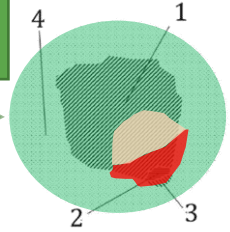
PROBABILISTISCHE VERIFIKATION
→BEWERTUNG & OPTIMIERUNG DER TESTABDECKUNG

ANALYTISCHE VERIFIKATION
→ 100%-GARANTIEN

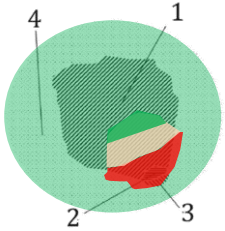
VERIFIKATION



Aufdecken unbekannter & unsicherer Szenarien



Bewertung der Sicherheit



Funktions- & Szenarienspezifische Testabdeckung

KONTINUIERLICHE VERBESSERUNG

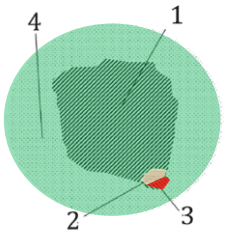
Key
1 Bekannte sichere Szenarien (Bereich 1)
2 Bekannte unsichere Szenarien (Bereich 2)
3 Unbekannte unsichere Szenarien (Bereich 3)
4 Unbekannte sichere Szenarien (Bereich 4)

Trainings-Daten & Test-Fälle

NACHTRAINING

SONDERFALL-ERKENNUNG

INTEGRATION



LIVE

FREIGABE

VALIDIERUNGSZIEL ERREICHT?

VALIDIERUNG

Unsicherheiten in KI-Systemen

Alex Kendall

Computer Vision &
Robotics Researcher

University of Cambridge

Email

Twitter

Google Scholar

LinkedIn

Github

Understanding what a model does not know is a critical part of many machine learning systems. Unfortunately, today's deep learning algorithms are usually unable to understand their uncertainty. These models are often taken blindly and assumed to be accurate, which is not always the case. For example, in two recent situations this has had disastrous consequences.

1. In May 2016 we tragically experienced the first fatality from an assisted driving system. According to the manufacturer's blog, "Neither Autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky, so the brake was not applied."

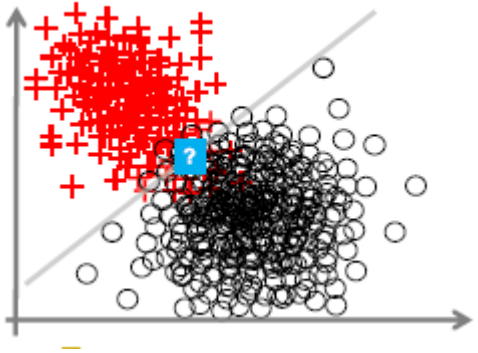
2. In July 2015, an image classification system erroneously identified two African American humans as gorillas, raising concerns of racial discrimination. See the news report here.

https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/

Arten von Unsicherheiten

Aleatorische Unsicherheit

- Auch **Datenunsicherheit** oder **statistische Unsicherheit**
- Entspringt wahrgenommenem **Rauschen** der Daten
- Ursachen:
  - Stochastisches Rauschen
    z.B. durch Sensorrauschen, unbeobachtbare Einflüsse, …
  - Deterministisches Rauschen
    z.B. durch mangelnde Ausdrucksstärke, …



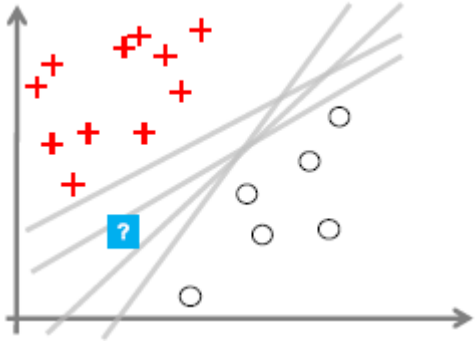
Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." *Machine Learning* 110.3 (2021): 457-506.

Wird durch das Aufzeichnen weiterer Daten **nicht verringert**

Arten von Unsicherheiten

Epistemische Unsicherheit

- Entspringt **fehlendem Wissen** (z.B. durch unzureichende Daten)
- Mögliche Ursachen:
  - Unterrepräsentierte Klassen
  - Out-of-Distribution Samples
  - Prozess-Shift / Prozess-Drift



Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." *Machine Learning* 110.3 (2021): 457-506.

Wird durch das Aufzeichnen weiterer, **nützlicher** Daten **verringert**

Nutzung der Unsicherheiten

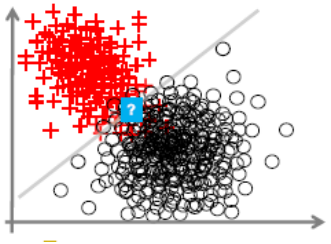
Ableitung von **Engineering-Guidelines** aus verschiedenen Unsicherheiten

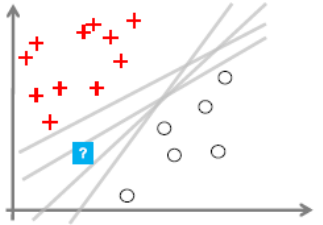
Aleatorische Unsicherheit

- Stochastische Ursache
  - Anpassung des Systems, z.B. weitere / bessere Sensorik
  - Fehlertoleranz, Fallback-Systeme, …
- Deterministische Ursache
  - Anpassung des Lernproblems, z.B. Erhöhung der Ausdrucksstärke

Epistemische Unsicherheit

- **Aufzeichnung weiterer nützlicher Trainingsdaten**
- **Nachtraining mit neuen Daten**

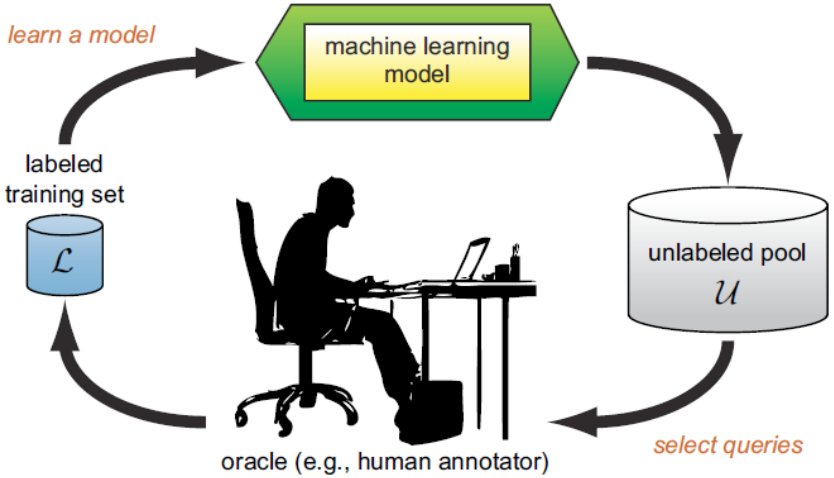




Hüllermeier, Eyke, and Willem Waegeman. "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods." *Machine Learning* 110.3 (2021): 457-506.

„Nützliche Daten“ gezielt aufzeichnen

- Datenvolumen von automatisierten Fahrzeugen liegt bei **mehreren Gigabyte pro Sekunde**
- Die meisten Daten sind **redundant** und bieten **kaum Mehrwert** für ein Training
- **Hohe Kosten** für **Datenspeicherung** und **Datentransfer** aus dem Fahrzeug
- **Labeling** der Daten ist **kostspielig** und **zeitaufwändig**



http://burrsettles.com/pub/settles.activelearning.pdf
[5] Settles, Burr. "Active learning literature survey." (2009).

→ Nutzung der epistemischen Unsicherheit zur gezielten Detektion wertvoller Daten

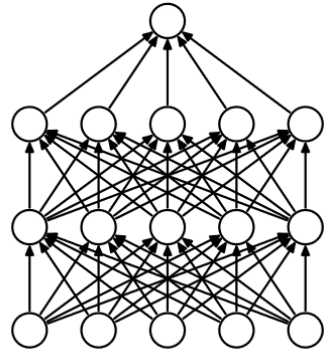
Ansätze zur Schätzung der epistemischen Unsicherheit

Unsicherheitsschätzung durch Deep Ensembles [1]

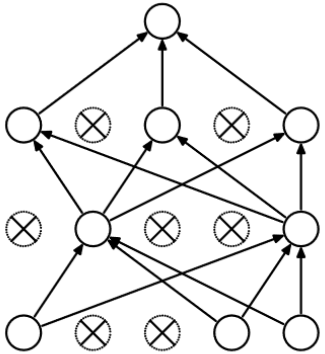
- Trainieren mehrerer Instanzen eines Modells von verschiedenen zufälligen Initialisierungen aus
- Inferenz: **Auswertung aller Modelle** und Schätzung einer Verteilung
- Problem: Großer Compute-Overhead durch n-faches Training und Inferenz

Dropout als Bayessche Näherung [2]

- Trainiere ein Modell mit Bernoulli Dropout für jede Schicht
- **Nutze Dropout auch zur Inferenz**, um eine Verteilung zu schätzen
- Problem: Großer Compute-Overhead durch n-fache Inferenz



(a) Standard Neural Net



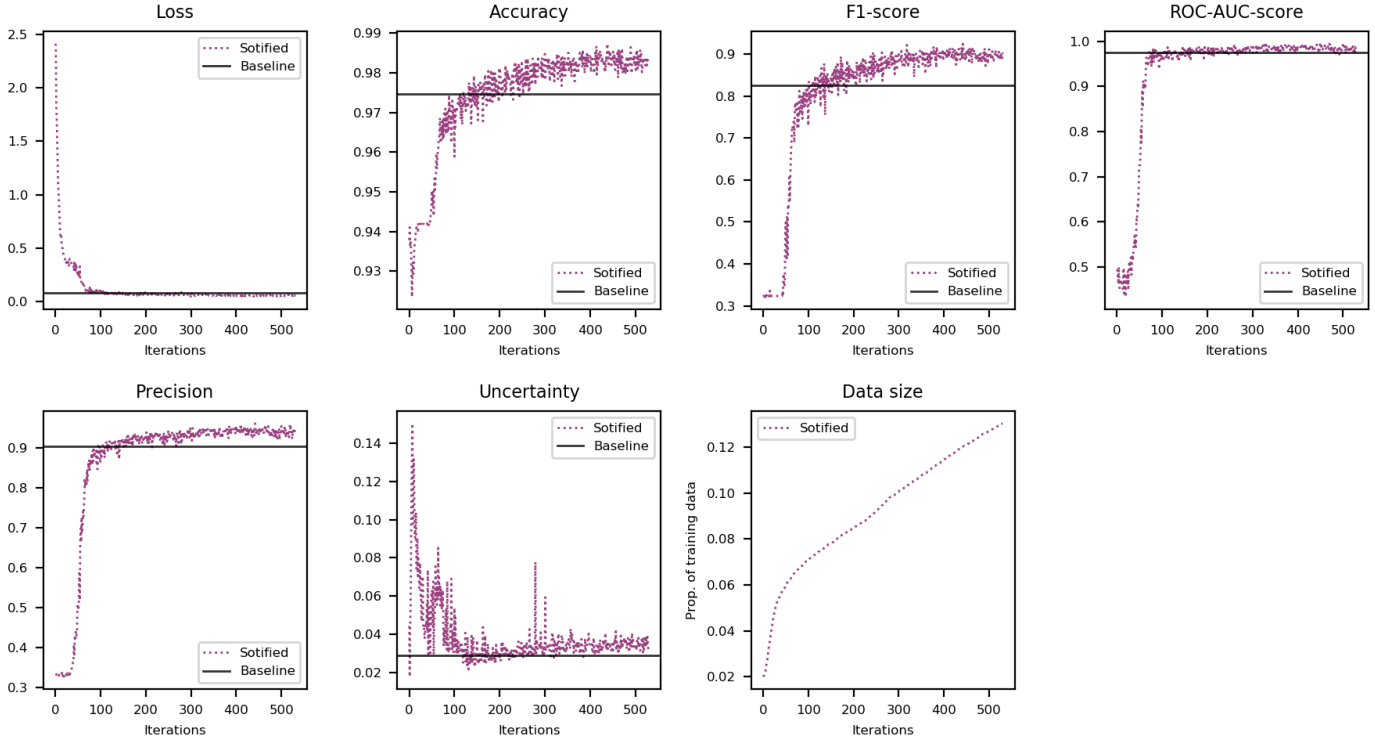
(b) After applying dropout.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15.1 (2014): 1929-1958.

[1] Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles." *Advances in neural information processing systems* 30 (2017).
[2] Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." *international conference on machine learning*. PMLR, 2016.

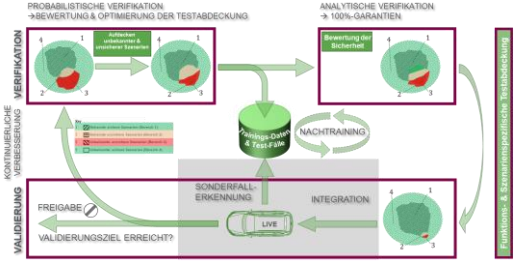

LCMP Simulator

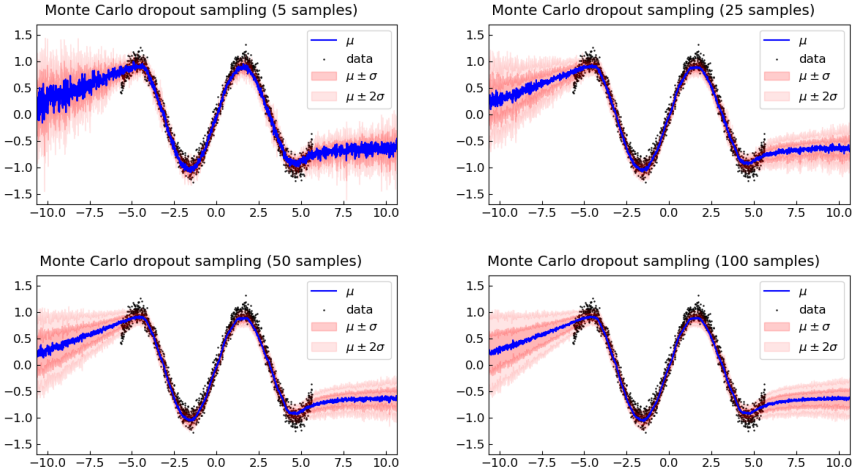
Training mit Unsicherheitsbasierter Datenauswahl

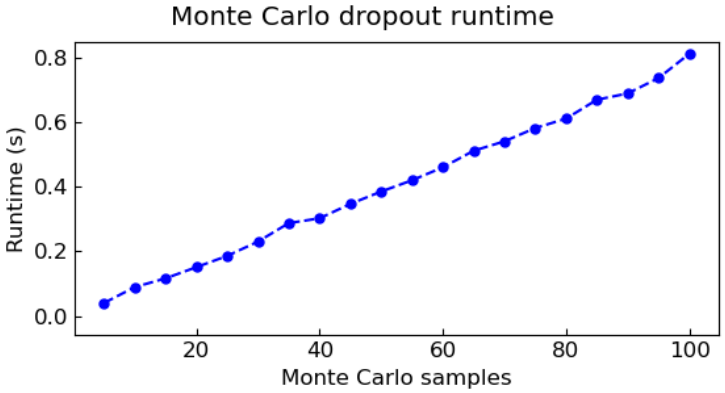


Monte Carlo Dropout als Schätzer für epistemische Unsicherheit

- Um die Aufnahme aller Daten zu vermeiden, muss die Entscheidung zur Aufnahme **live im Fahrzeug** erfolgen
- Unsichere Situationen müssen **verlässlich und schnell** erkannt werden
- Untersuchung am artifiziellen Beispiel (Sinus mit Extrapolation)



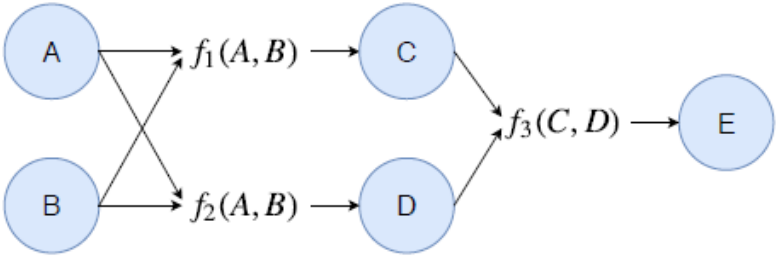


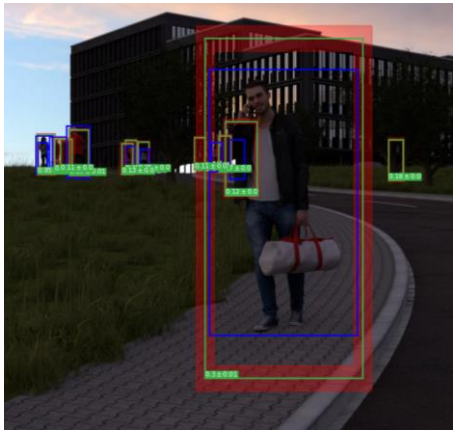


Sampling-freie Schätzung der Epistemischen Unsicherheit

Basierend auf: **Sampling-free epistemic uncertainty estimation using approximated variance propagation [3]**

- Propagation von Unsicherheiten analog zur Fehlerfortpflanzung
- Hier entspricht der Fehler der Varianz einer Zufallsvariable
- Für Neuronale Netze kann z.B. Dropout als **Noise Injection** betrachtet werden
- Generalisierbarkeit auf NN-Operationen durch **Linearisierung** und Berechnung der **Varianz-Kovarianz-Matrix**



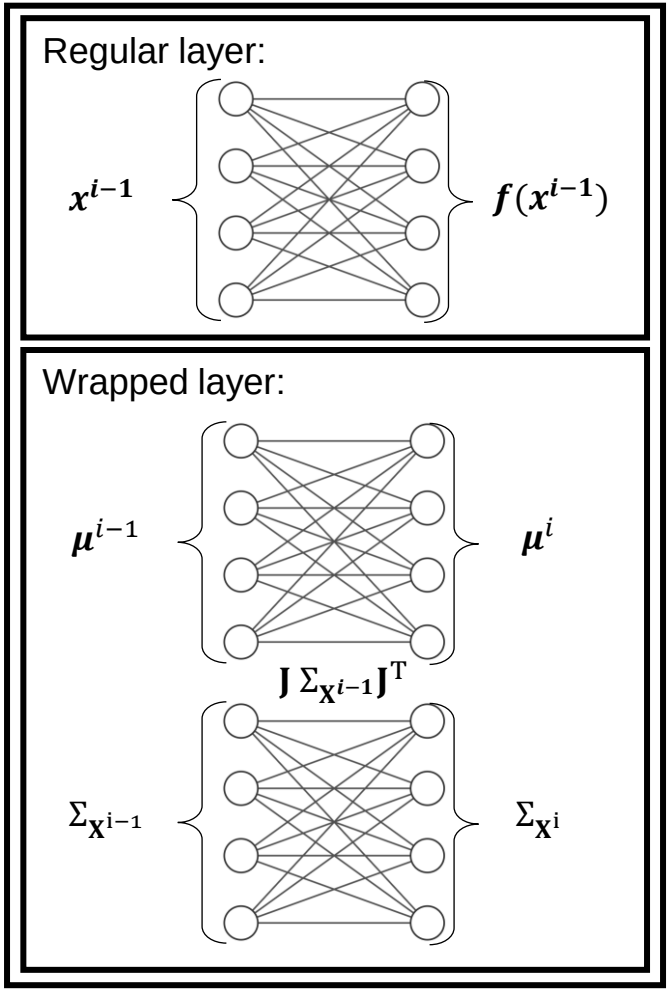


[3] Postels, Janis, et al. "Sampling-free epistemic uncertainty estimation using approximated variance propagation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.

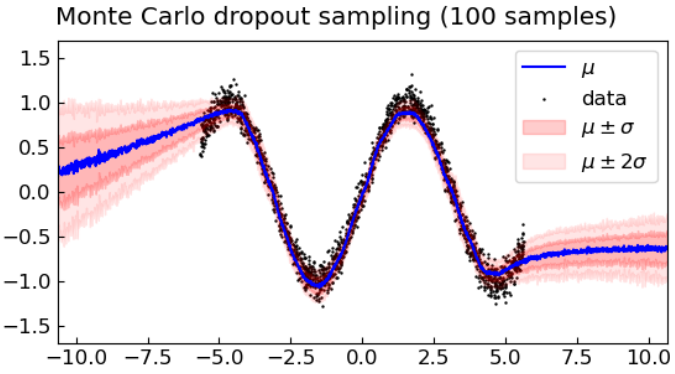
Implementierung: Uncertainty Wrapper

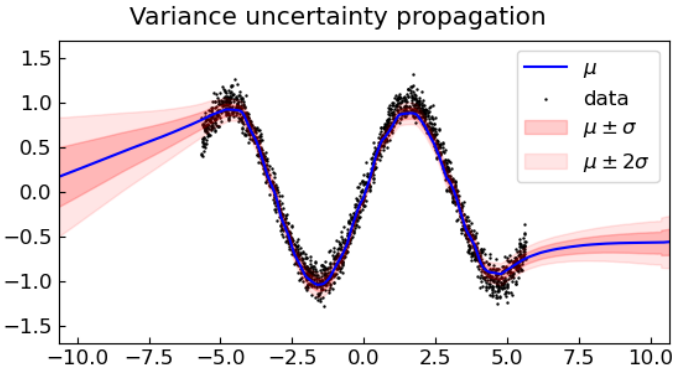
- Jede Operation / Jedes Layer des NN wird gewrapped
- Nutzung des Uncertainty Wrappers:

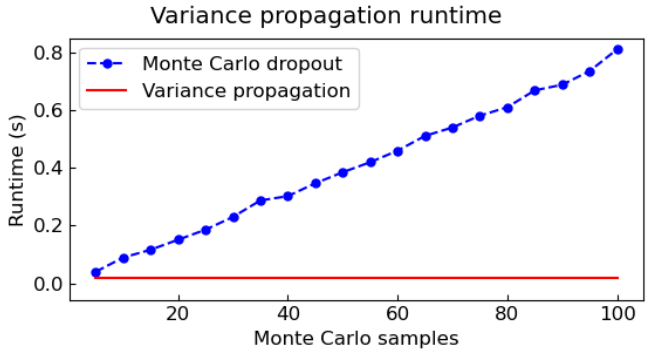
```
# wrap the trained model
model = load_neural_network()
uncertainty_model =
UncertaintyEstimator(model)

# initialize input
mean_in = x_in
var_in = torch.zeros_like(x_in)
input = (mean_in, var_in)

# uncertainty estimation
mean_out, var_out = uncertainty_model(input)
```



Evaluation des synthetischen Beispiels







Zusammenfassung

- Unsicherheiten in KI-Systemen lassen sich in **verschiedene Ursachen** auftrennen
- Verschiedene Arten von Unsicherheiten benötigen **unterschiedliche Behandlung**
  - Aleatorische Unsicherheit: Anpassungen des System-Designs
  - Epistemische Unsicherheit: **Gezielte Aufnahme wertvoller Daten** und Nachtraining
- **Live-Schätzung** epistemischer Unsicherheiten bieten ein gutes **Entscheidungskriterium**, welche Daten aufgenommen und gelabelt werden sollten
- **Sampling-freie Methoden** erlauben **ressourcenschonende und genaue Schätzung** der epistemischen Unsicherheit