

Hot Topics in Machine Learning Safety

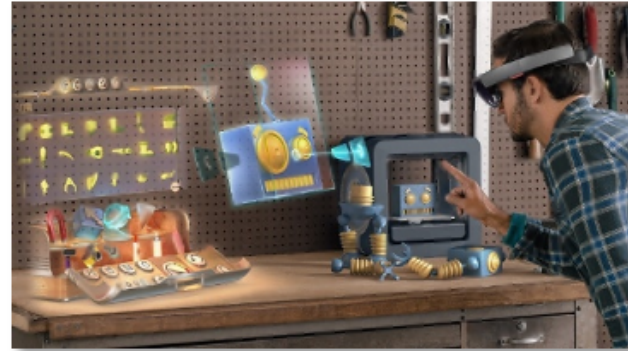
WS 2023

Organization

- Prof. Frank Ortmeier
- Konstantin Kirchheim → kirchhei@ovgu.de
- Venkatesh Sambadam

Mobile Assistance

- cloud-based Assistance
- smart glasses
- pervasive Systems



Programming paradigms

- agile software engineering
- self-healing software
- model-driven systems engineering



Build the right software

- smart robotics
- internet of things
- industrial data
Architectures /OPC UA



Robotics & Industry 4.0

Build the software right

- dependability analysis
- industrial safety and security
- system verification
- safety-driven engineering



Safety & Dependability

SOFTWARE
ENGINEERING

Forschungsschwerpunkte CSE

Chair of Software Engineering

- Build the right software
- Build the software right

Applied to Machine Learning:

- Build the right model
- Build the model right

What are your expectations for this seminar?

DÜSTERE ZUKUNFT

Die Killer-Roboter sind im Anmarsch

VON CHRISTOPH SCHÄFER - AKTUALISIERT AM 02.06.2018 - 21:08

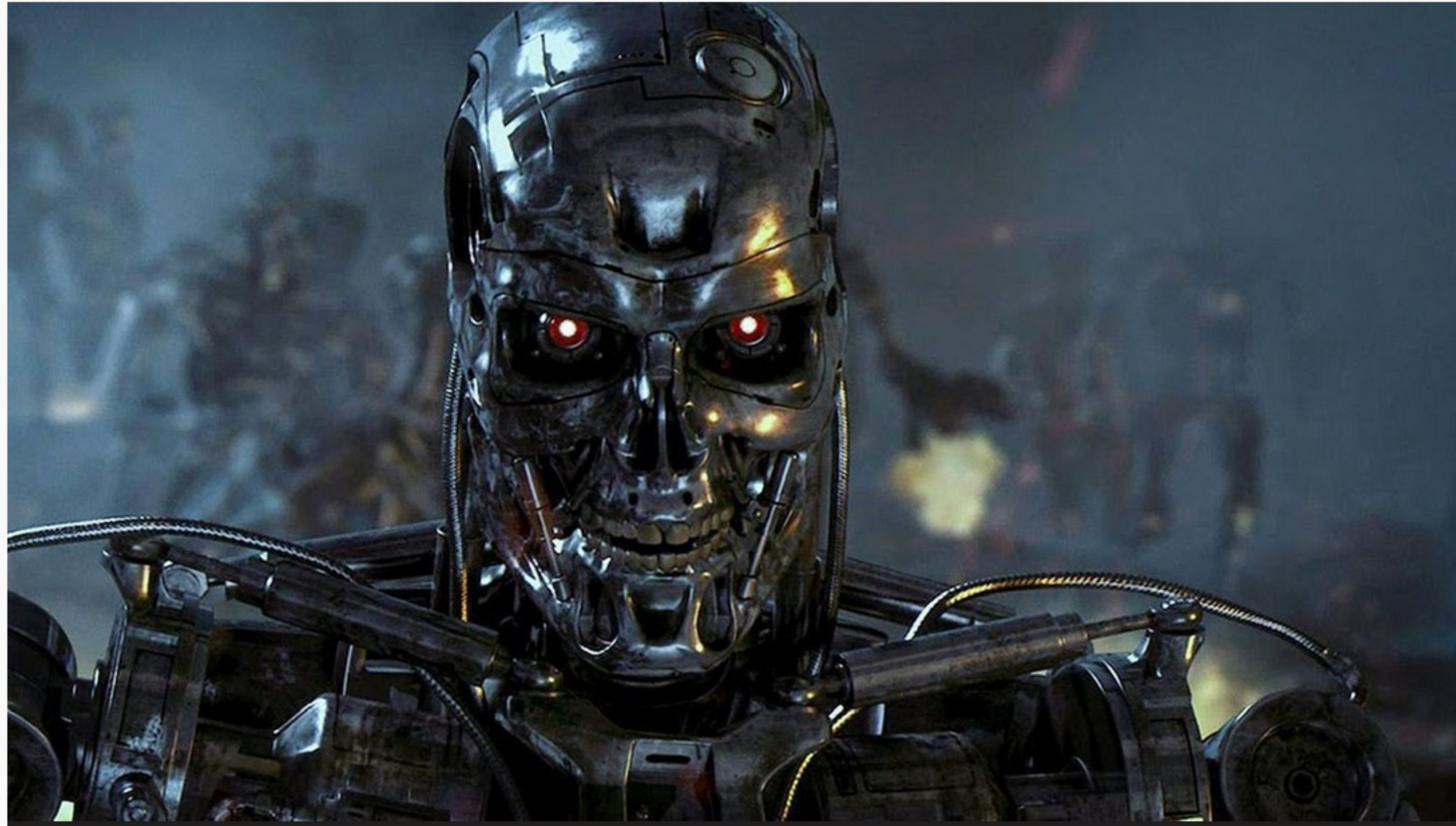


Mörderische Maschinen wie der „Terminator“ sind keine Science-Fiction mehr.

Zeit, 02.06.2018

KÜNSTLICHE INTELLIGENZ

Bedrohung wie in "Terminator" – superintelligente KI für Menschen unbeherrschbar



Stern, 09.05.2021

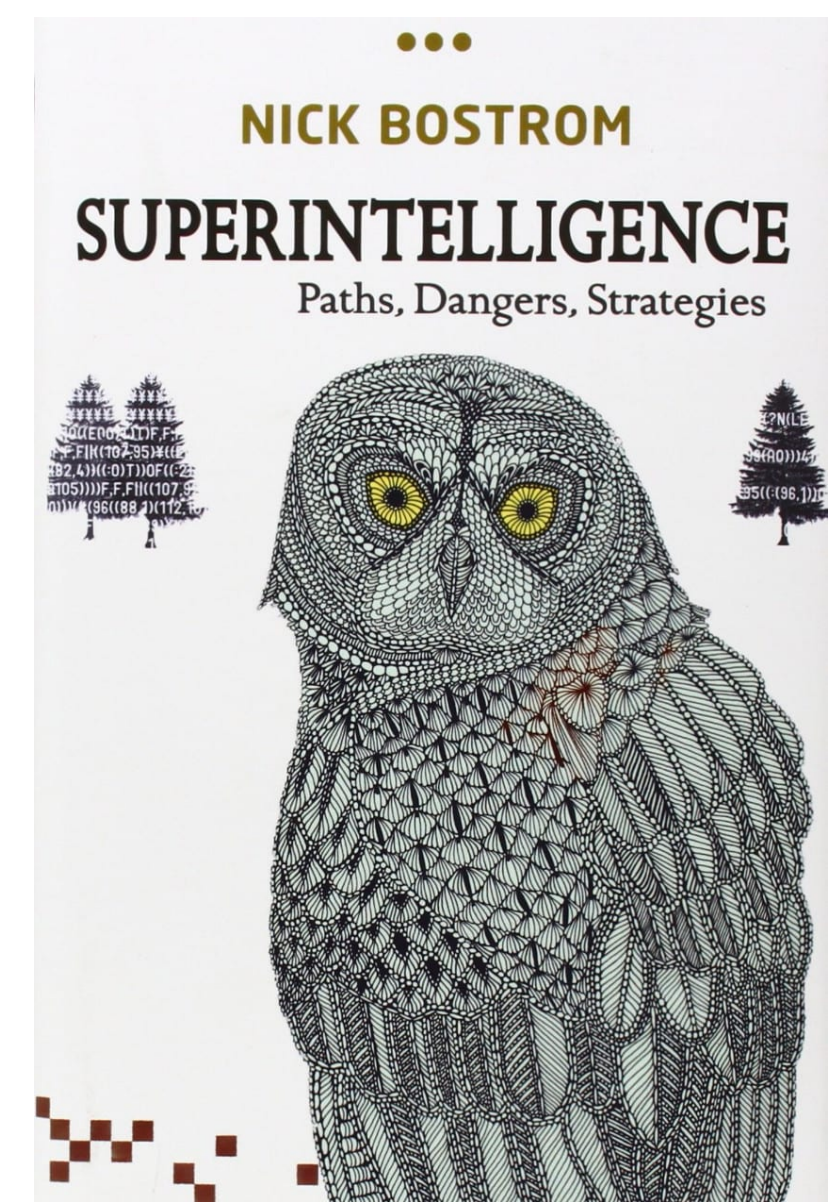
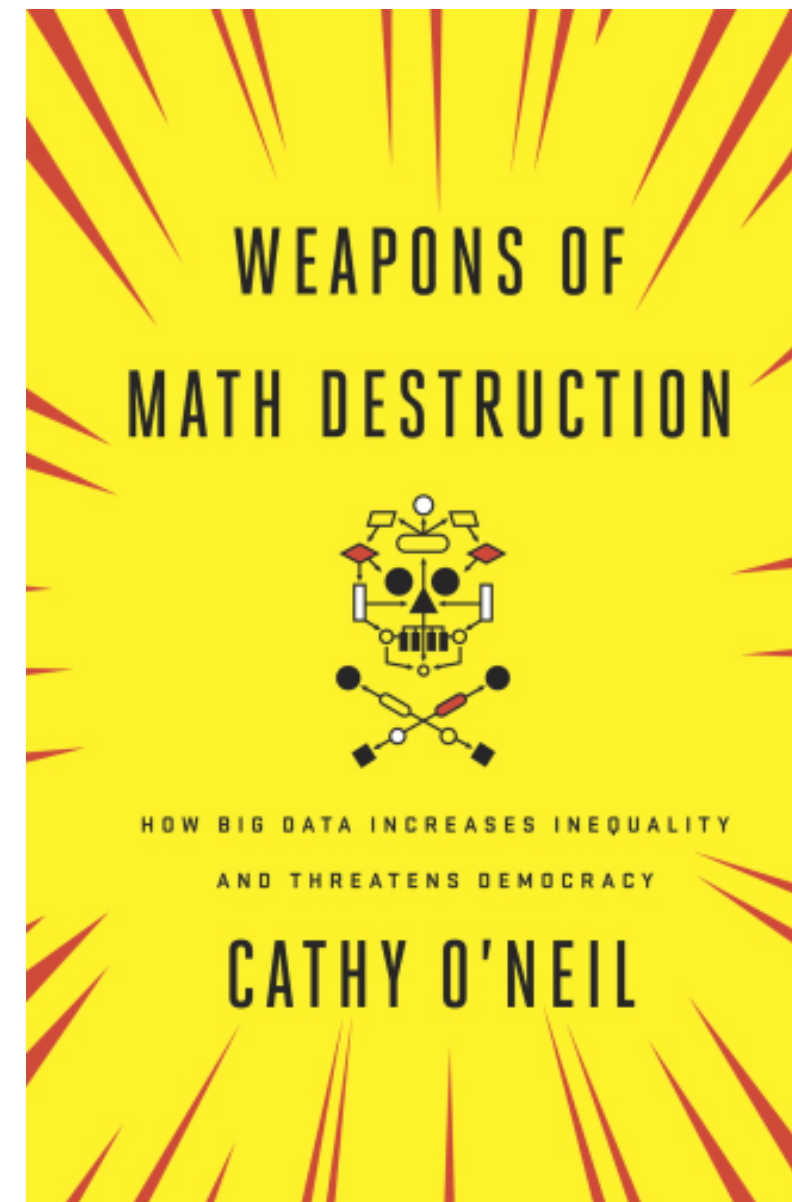
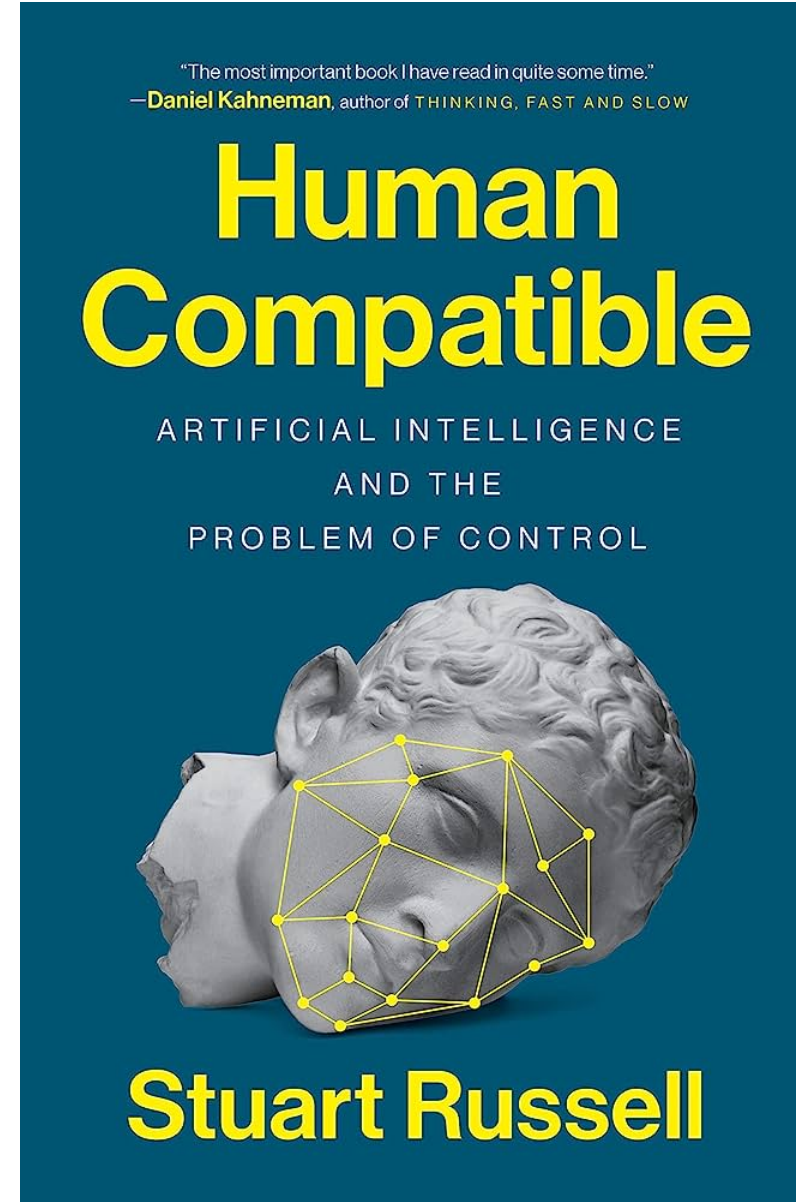
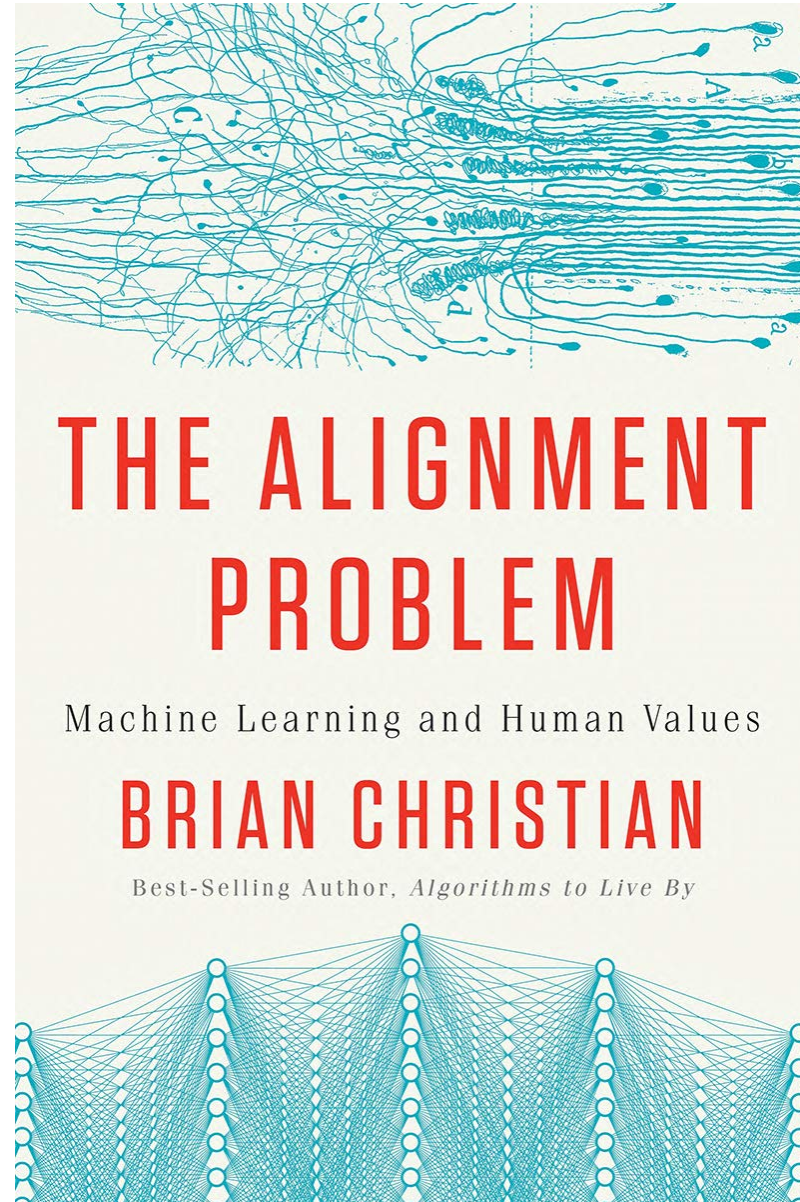
Existential Risk

- Might be a long-term problem
- Terminator scenario rather unlikely
- Not main focus of this seminar

Goals of this seminar

- What can go wrong with ML models in production?
- How can we mitigate this?
- *Emphasis* on Deep Learning Systems

Popular science books on these topics



Prerequisites

- Introduction to Deep Learning
 - DNNs
 - CNNs
- Good understanding of ML concepts in general

Motivation: EU Artificial Intelligence Act

1) Unacceptable Risk

- Prohibited
- E.g. Social Scoring, real-time remote biometric identification in public spaces

2) High Risk

- E.g. Medical Applications
- Can cause significant harm
- Highly Regulated

3) Other

- E.g. Risk of Manipulation through Chatbots
- Requires Transparency

→ Topic will stay relevant

Seminar Schedule

- Introduction and Motivation (Today)
- ML Fundamentals
- Testing
- Explainability
- Uncertainty Estimation
- Anomaly Detection
- Adversarial Machine Learning
- Alignment/Societal/Long-Term Risk

ML Fundamentals

Revisiting concepts you should already be familiar with, such as:

- Deep Neural Networks
- Backpropagation
- Gradient Descent

Model Testing

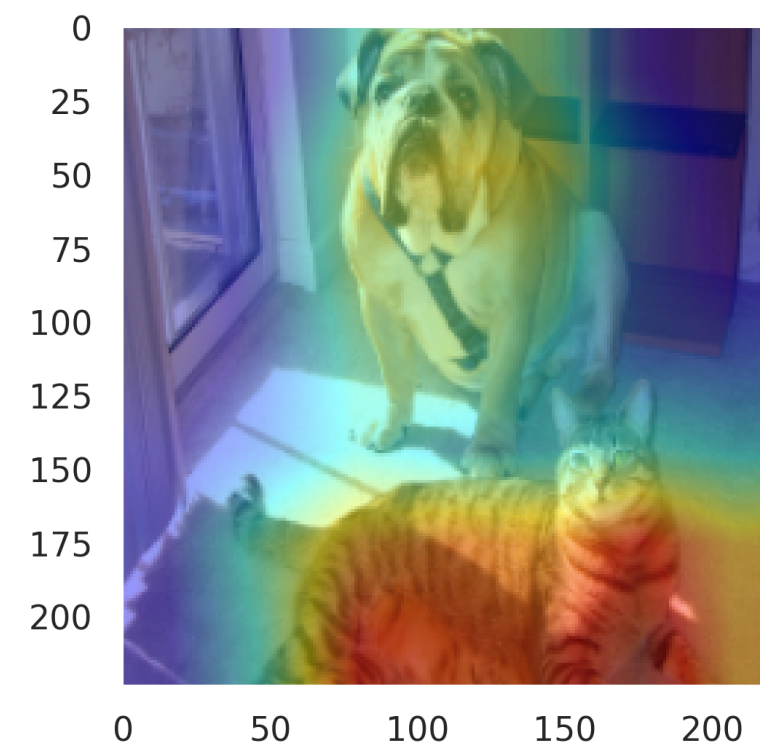
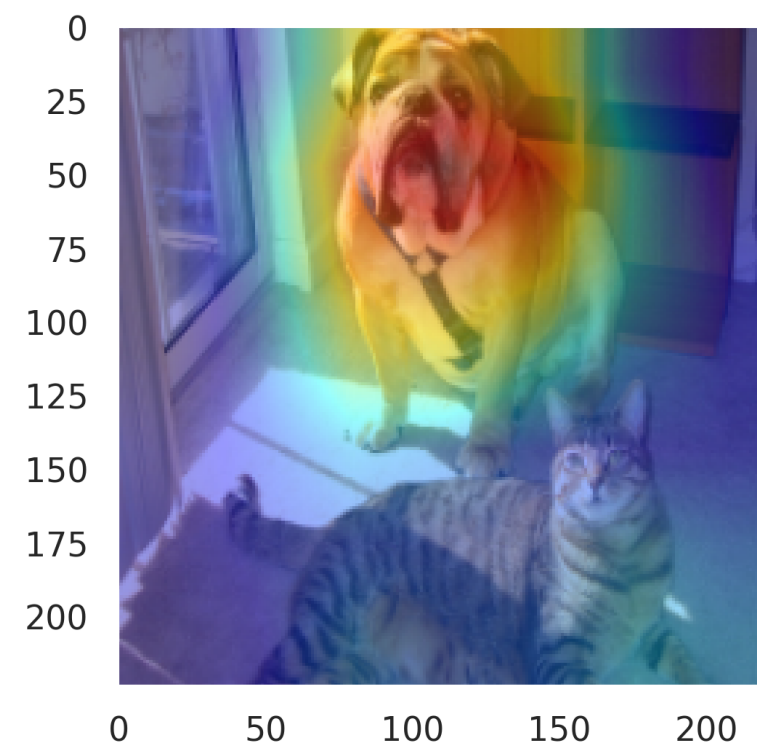
- Example: Classification model has accuracy of 95%
 - Is that good?
- What data was the model tested on?
- What would random guessing give you?

Questions:

- How can we **reliably** estimate how good model will perform?
- What can happen if we do not test the model?
- What are common pitfalls?

Explainability

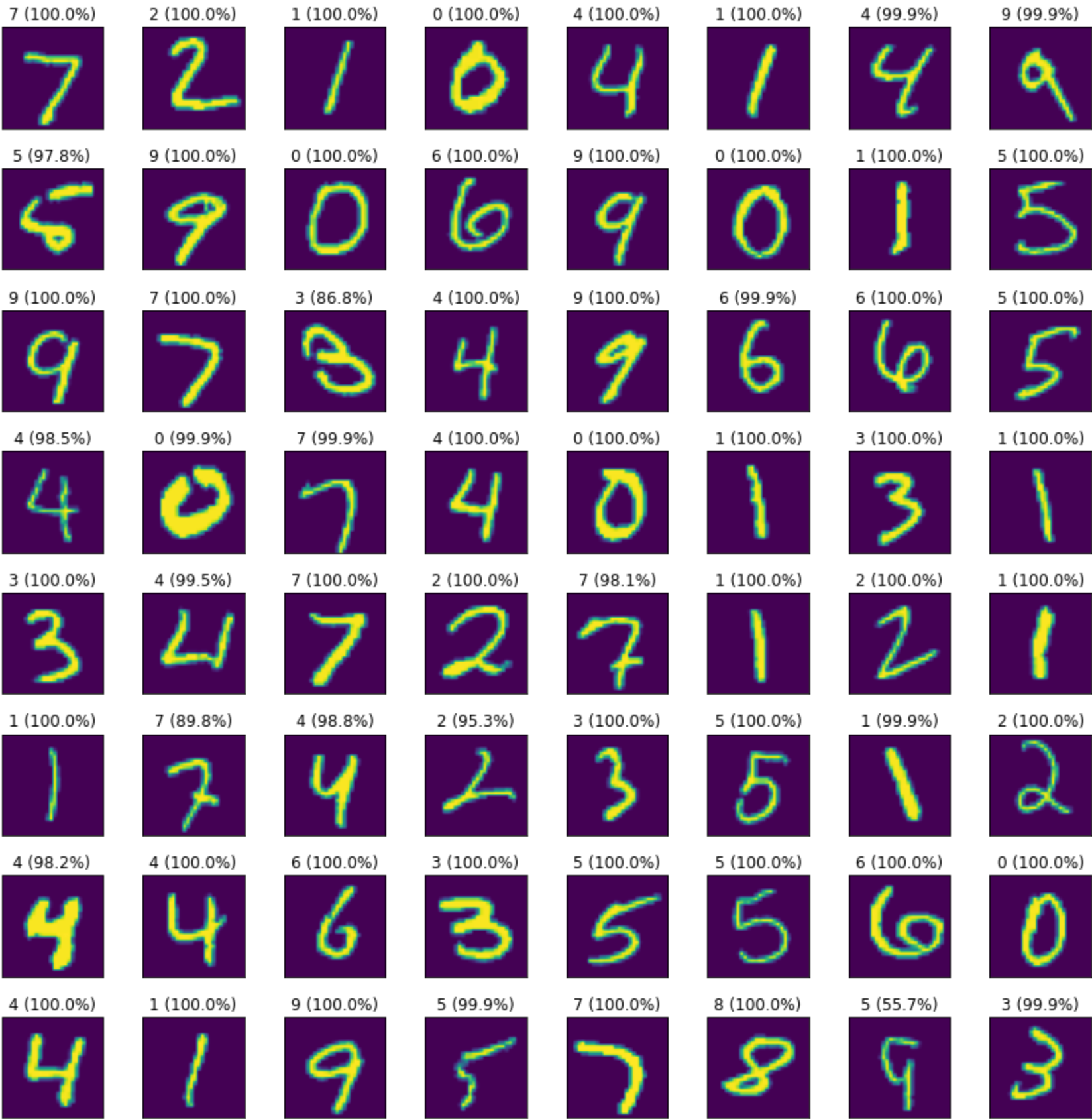
- Train model to classify images of cats and dogs
- For some input x , models says **dog**
- Why?



Uncertainty Estimation

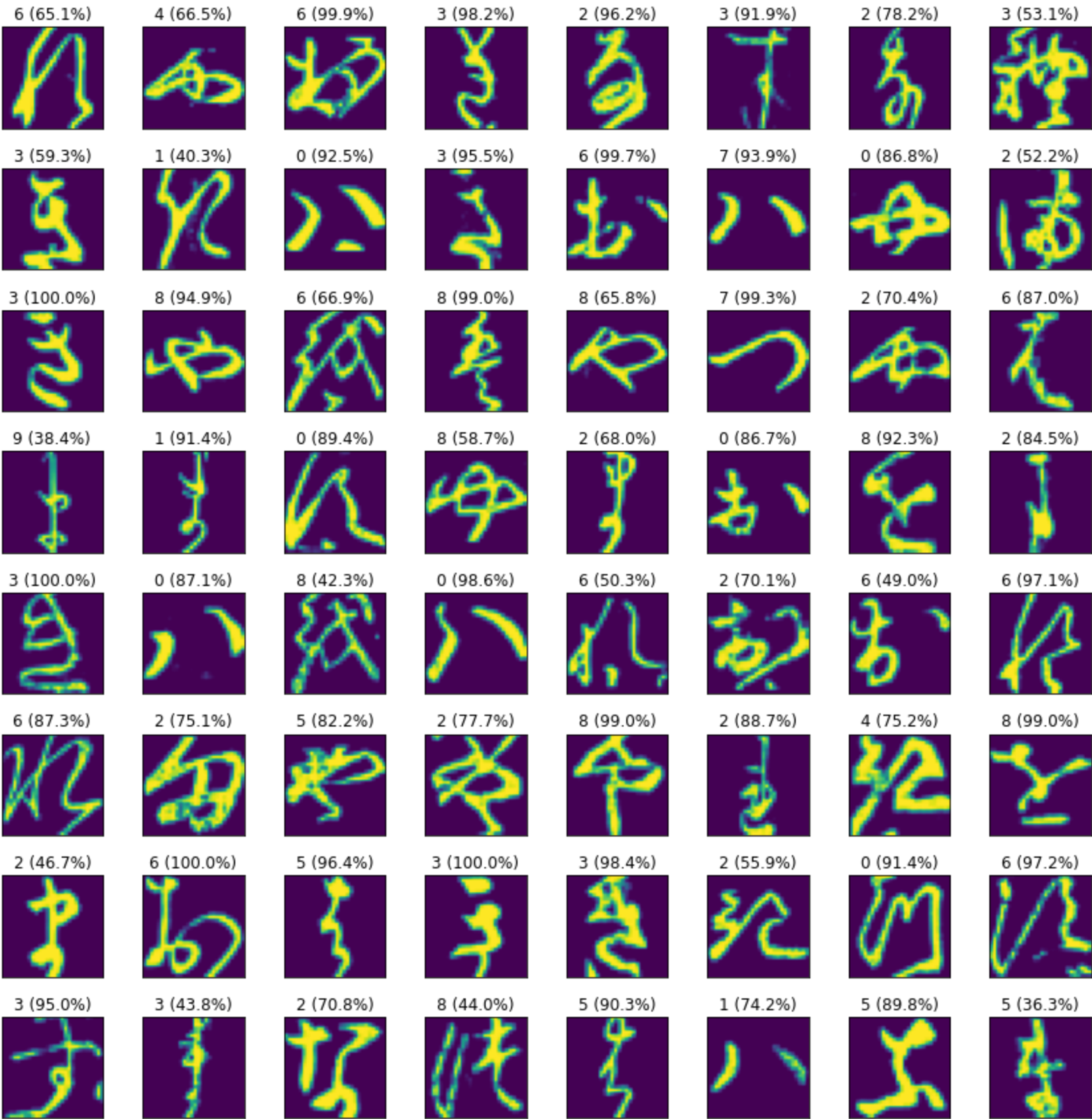
- Train model to classify cats and dogs
- For some input x , models says **dog**
- What is the probability that the image actually depicts a **dog**?

Anomaly Detection



Normal Examples

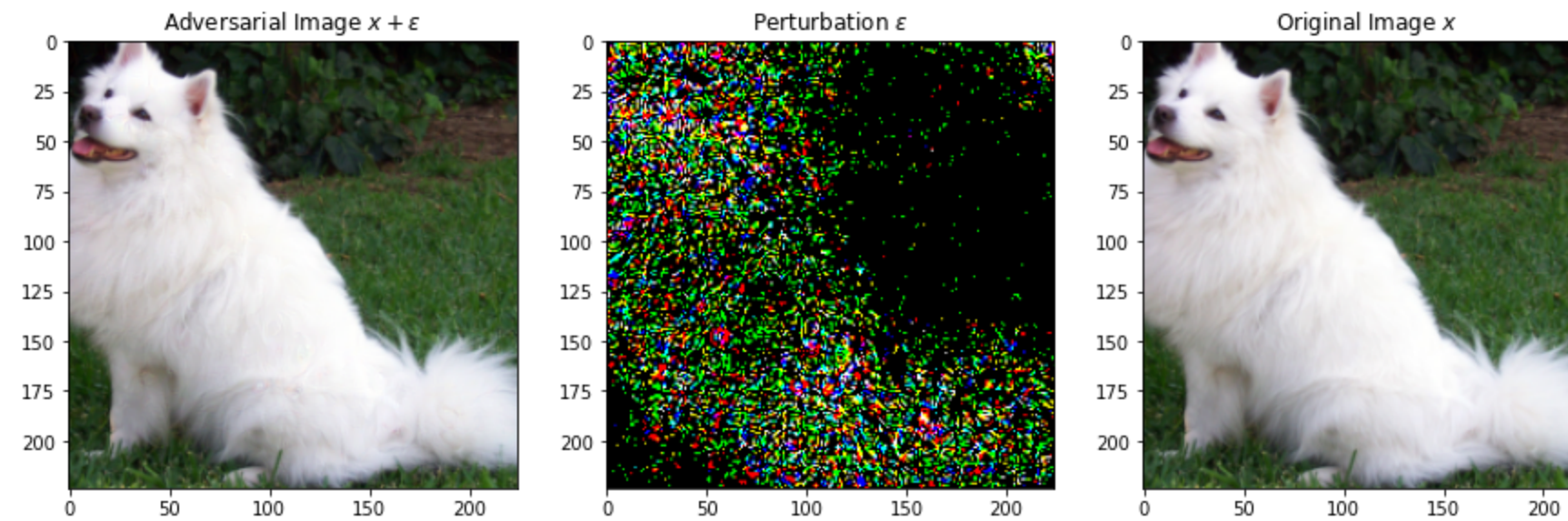
Anomaly Detection



Anomalous Examples

Adversarial Machine Learning

- Models are not robust: minimal changes in input can lead to extreme changes in output
- How can we attack models?
- How can we defend our models against these attacks?



Alignment/Societal/Long Term Risk

- How do we align the goals of a machine learning systems with our own goals?
- What are possible dangers of deploying ML systems on a large scale?



Seminar Organization

- Team projects (2 or 3 persons) throughout semester
- Grade: project presentation + written report
- Team and project finding: next 2 weeks
 - you are free to chose relevant topics
 - we can also provide suggestions for topics
- Example projects
 - Take some paper, reimplement and apply to new dataset
 - Survey a specific area
 - Take some paper, try to improve

Team-forming

- Form teams on your own
- If not found: ask in the E-Learning forum
- Write an email to discuss possible projects

Exercises

- Reading Exercises
 - Possibility to discuss during seminar
- Programming Exercises
 - Jupyter Notebooks, Optional
- Slides and Material provided via E-Learning