Tag 8: Modellbewertung & Prüfungsleistung

Samuel Schlenker

11.11.2025, WWI 2025F

Students should ...

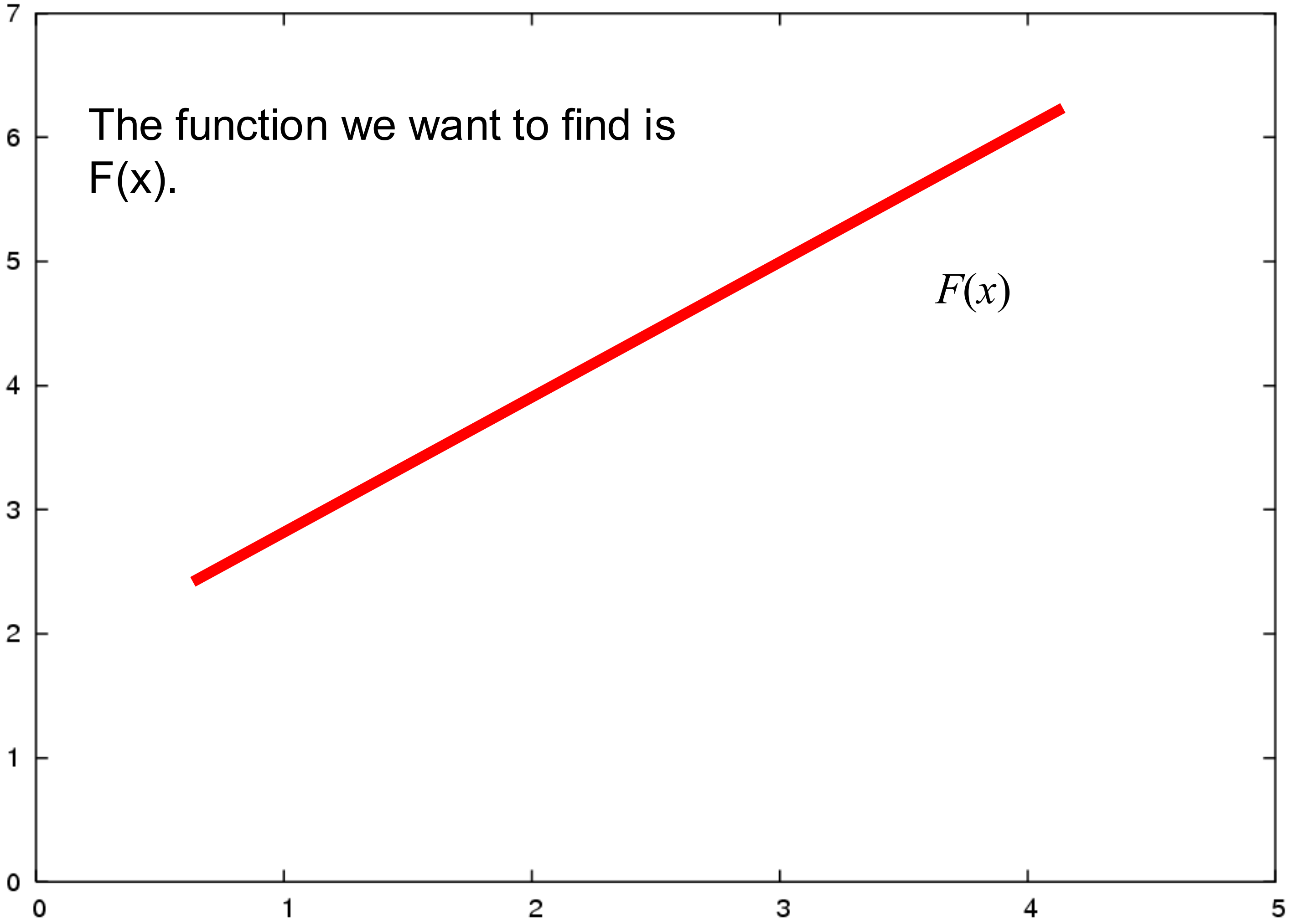
- Understand the concepts of bias and variance in machine learning models
- Identify overfitting (high variance) and underfitting (high bias) scenarios
- Apply performance metrics for regression models (MSE, RMSE, MAE, R^2)
- Apply performance metrics for classification models (Precision, Recall, Accuracy, F1-Score)
- Interpret confusion matrices (True Positives, False Positives, True Negatives, False Negatives)
- Understand and interpret ROC/AUC curves for model evaluation
- Differentiate between Type 1 errors (false positives) and Type 2 errors (false negatives)
- Understand deployment considerations for machine learning models in production
- Recognize ethical considerations and risks in AI systems (fairness, bias, privacy)
- Understand the EU AI Act and its risk-based approach to AI regulation
- Recognize the importance of Trustworthy AI principles (explainability, fairness, robustness, privacy, accountability)
- Understand the concept of Explainable AI (XAI) and why "how" matters as much as "what"
- Identify the timeline and requirements of AI governance and compliance

How well the trained model fits reality depends on:

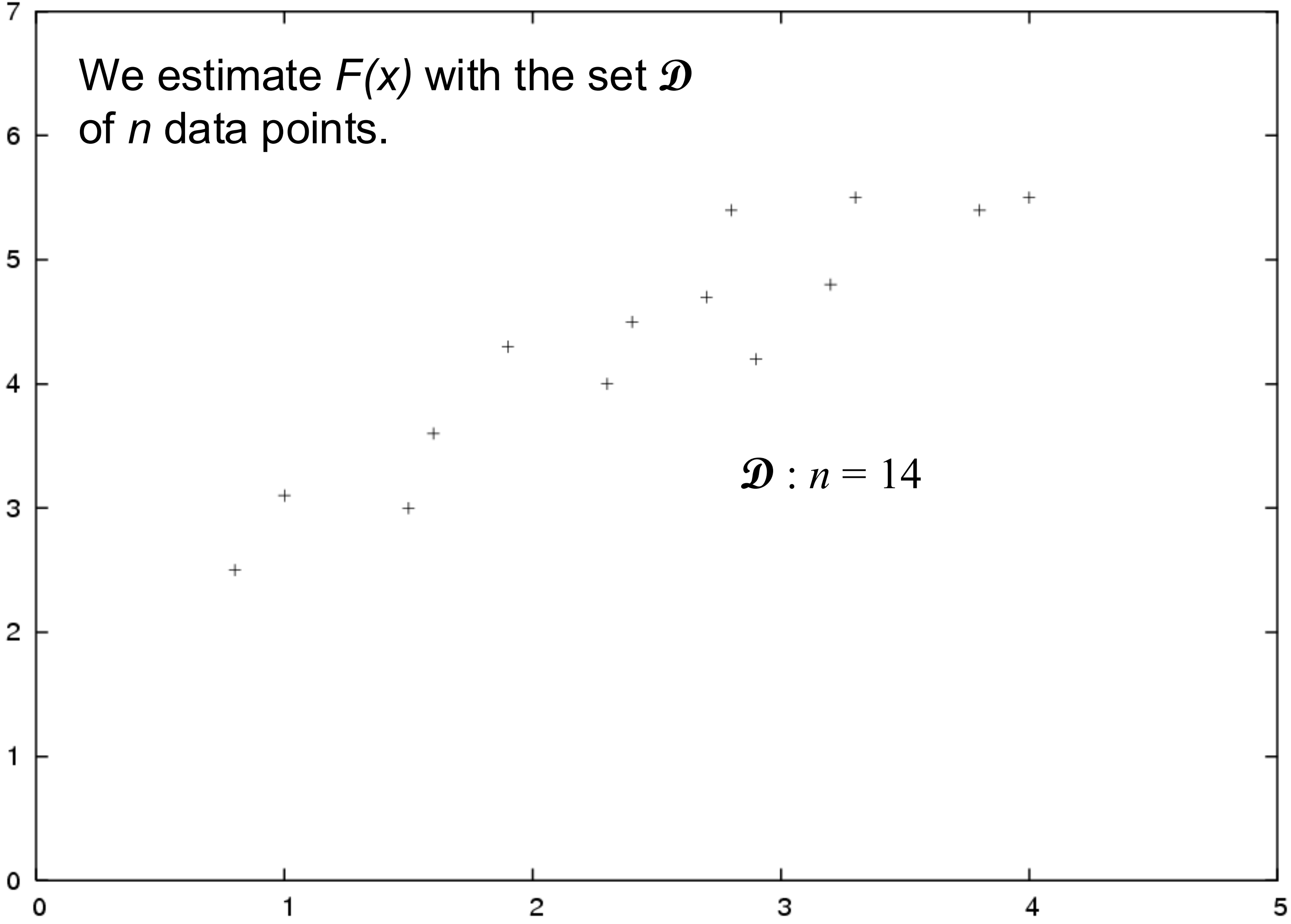
Model	Training data

Interdependent

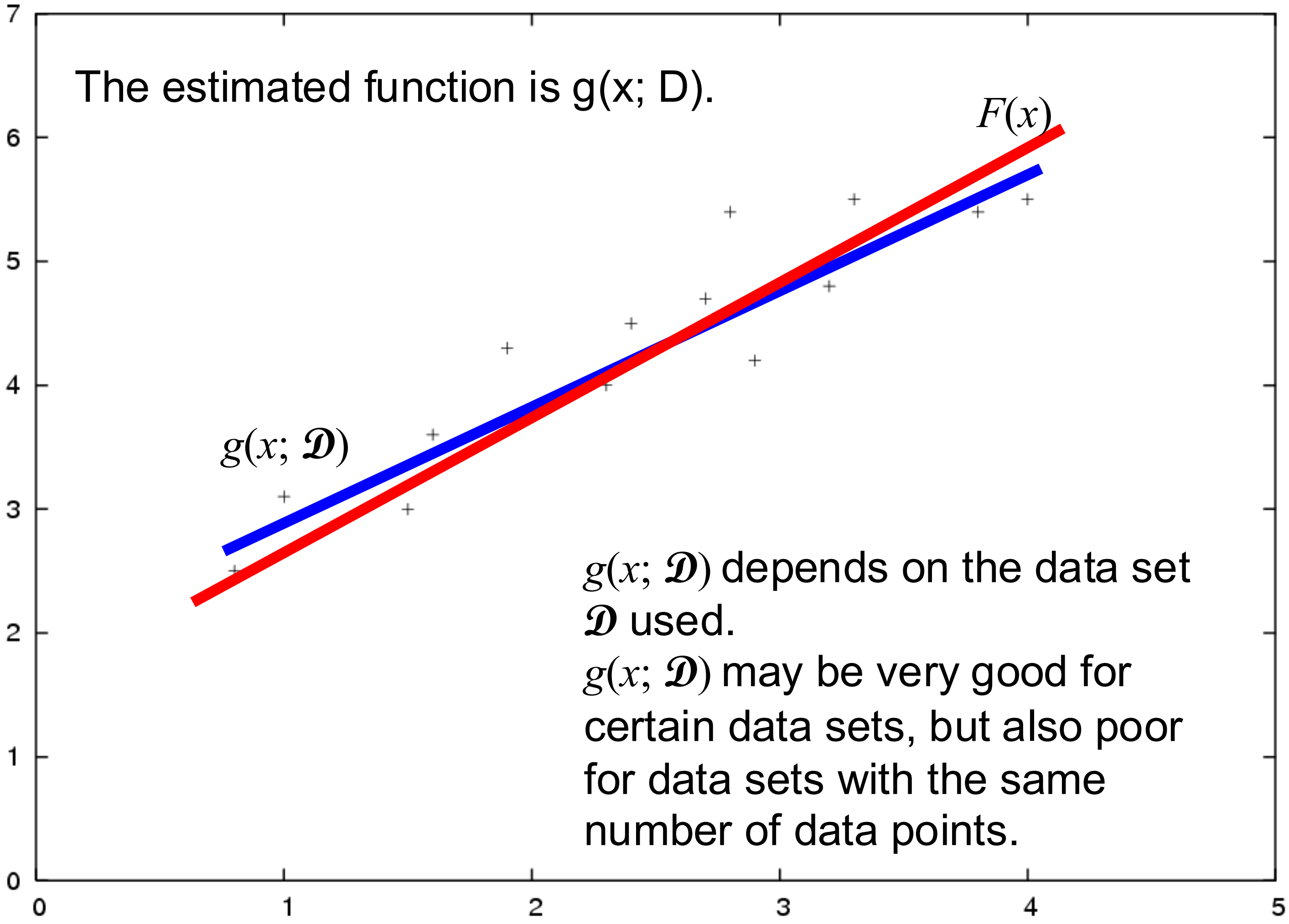
Also known as bias and variance


The function we want to find is F(x).
F(x)
0
1
2
3
4
5
6
7
0
1
2
3
4
5

We estimate *F(x)* with the set $\mathcal{D}$ of *n* data points.

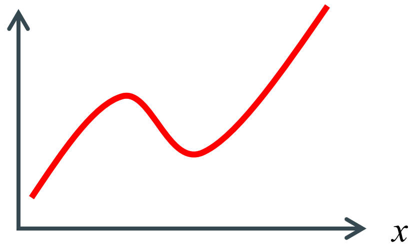


The estimated function is g(x; D).

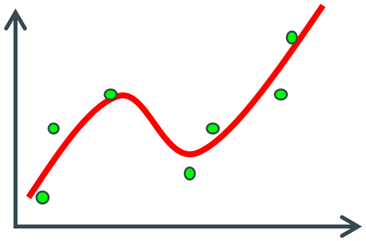


$g(x; \mathcal{D})$ depends on the data set $\mathcal{D}$ used.
$g(x; \mathcal{D})$ may be very good for certain data sets, but also poor for data sets with the same number of data points.

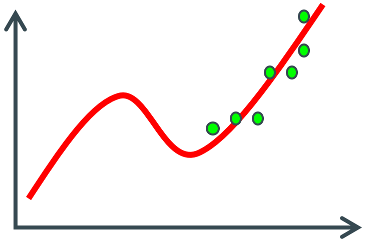
- *F(x)*



- With this data set, $g(x; \mathcal{D})$ would be close

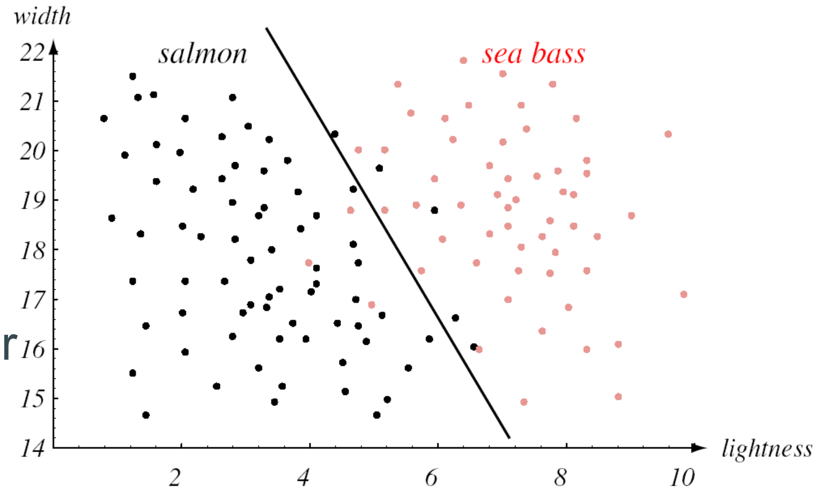


- Not with this data set



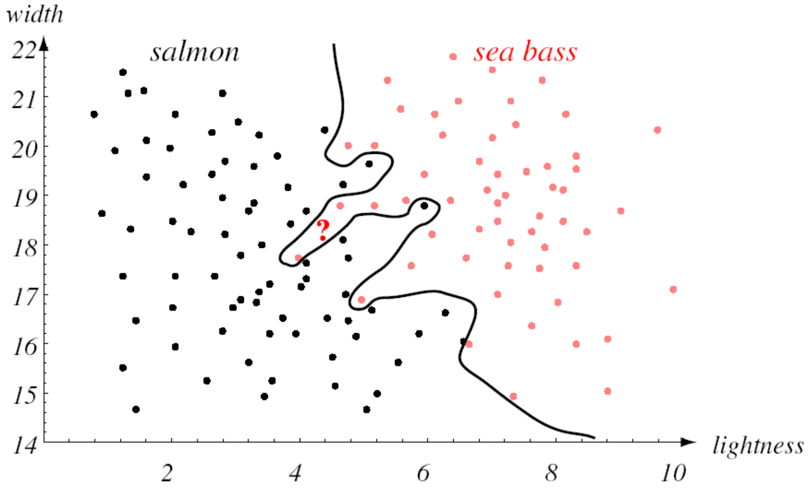
High bias (underfitting)

- The model is too simple or inflexible



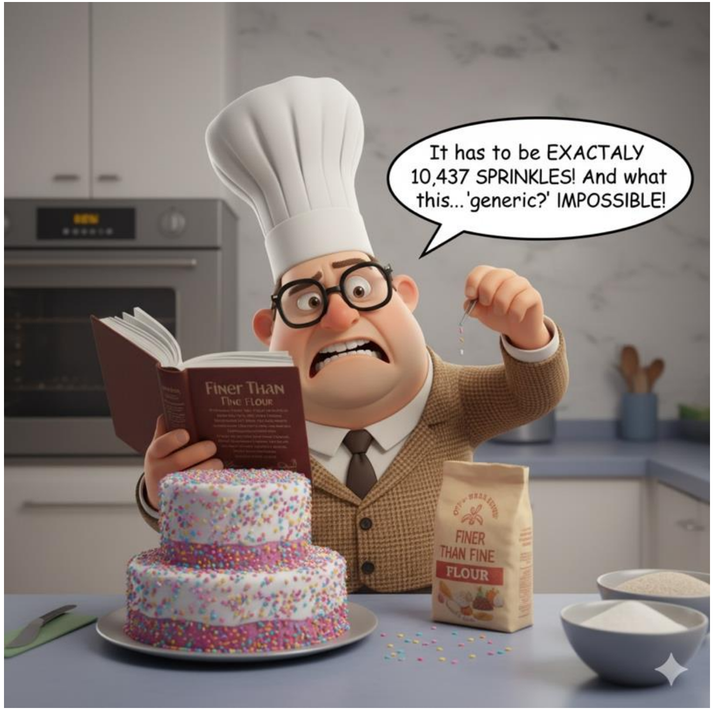
High variance (overfitting)

- The model is too flexible or too closely adapted to the training data set
- “Memorization” of the training data

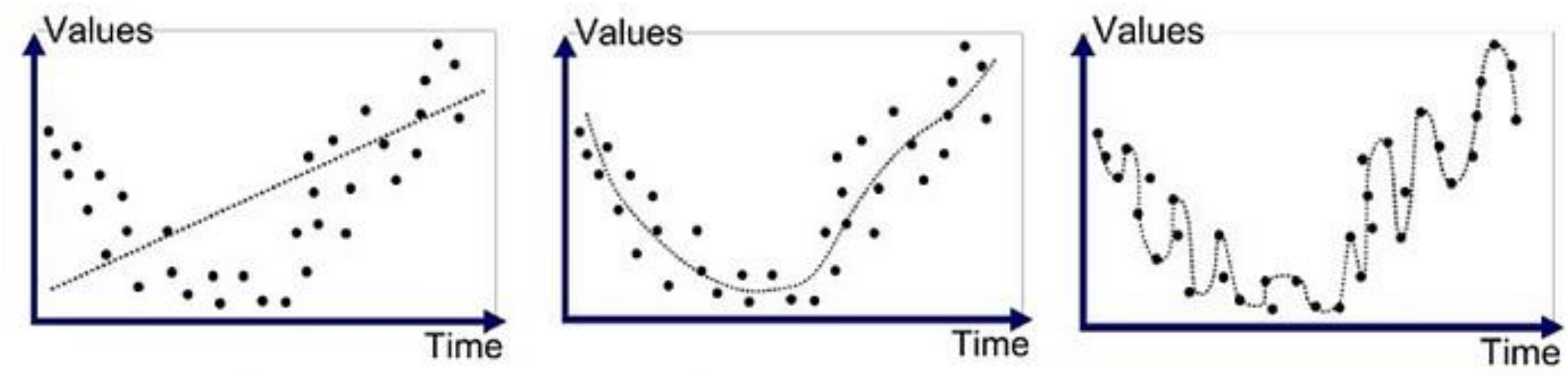


(Aus Duda, Hart, Stork: Pattern Classification, 2nd ed.)

Over- and underfitting



Over- and underfitting



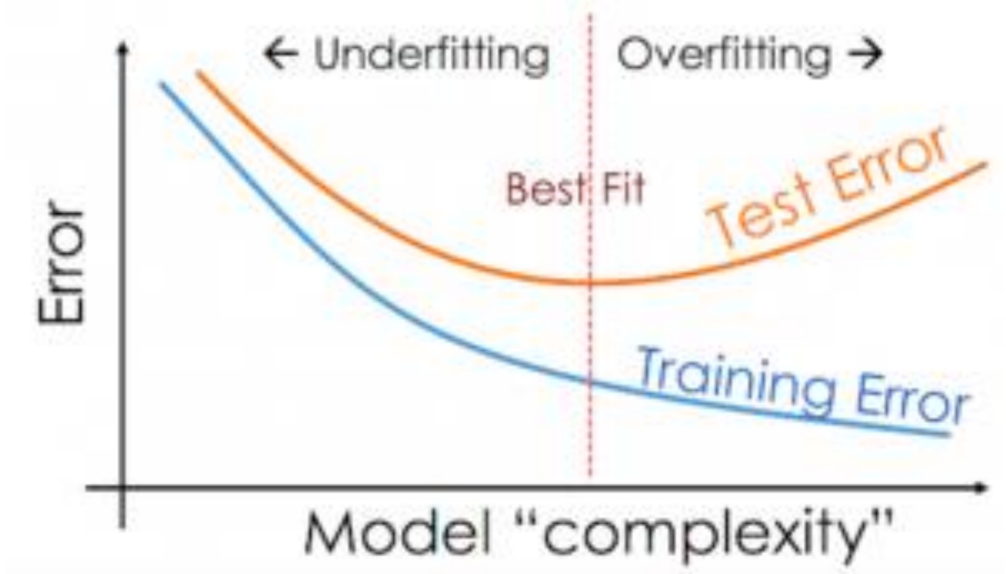
Underfitted

Good Fit/Robust

Overfitted

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

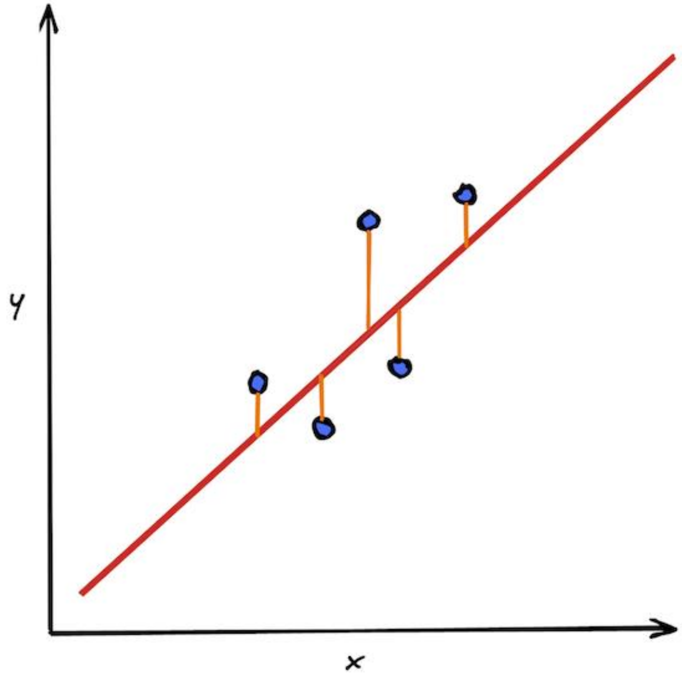
Over- and underfitting



https://vitalflux.com/overfitting-underfitting-concepts-interview-questions/

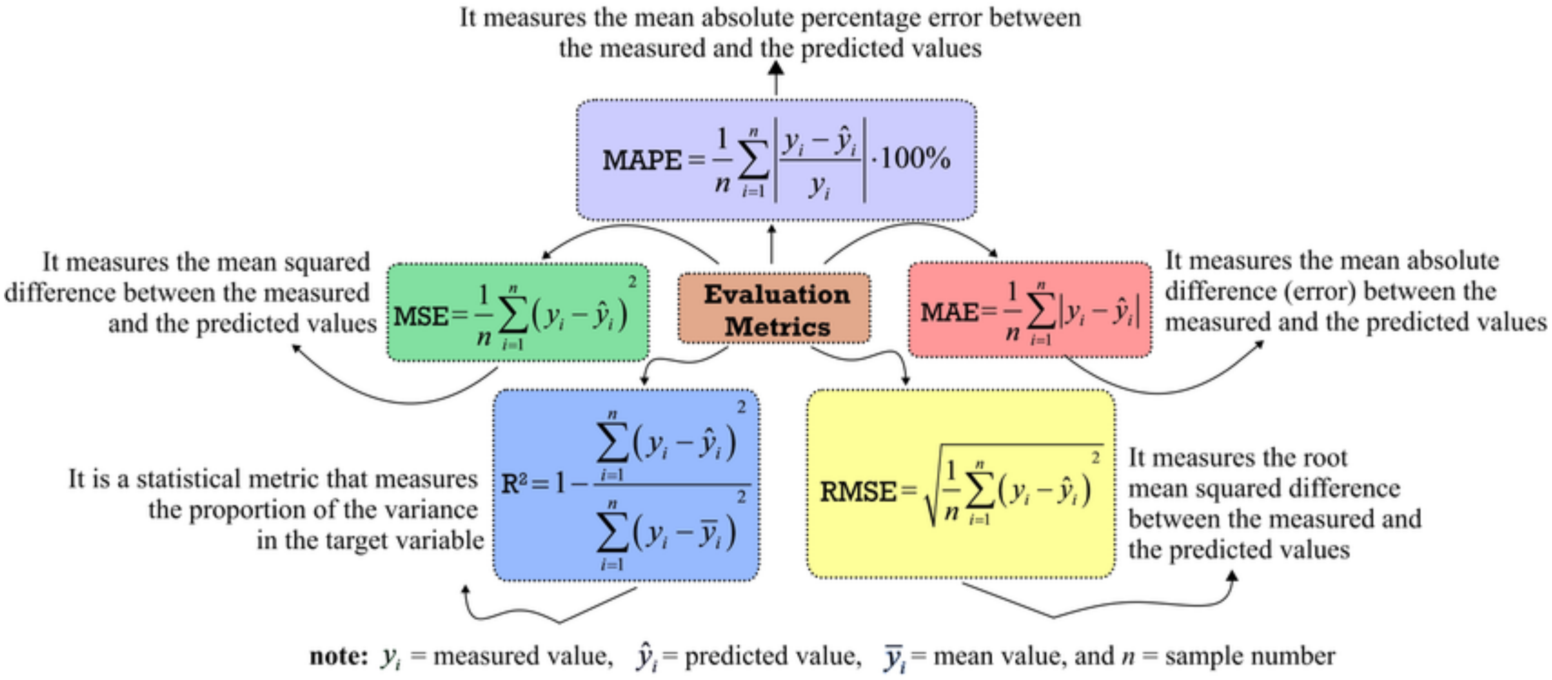
Model evaluation

Measuring Performance for Regressions



https://towardsdatascience.com/a-comprehensive-overview-of-regression-evaluation-metrics-6264af0926db/

Measuring Performance for Regressions



https://www.researchgate.net/figure/Performance-evaluation-metrics-for-regression-models_fig5_381401914

Measuring Performance for Classifications

		Predicted Class	
		Positive	Negative
Real Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

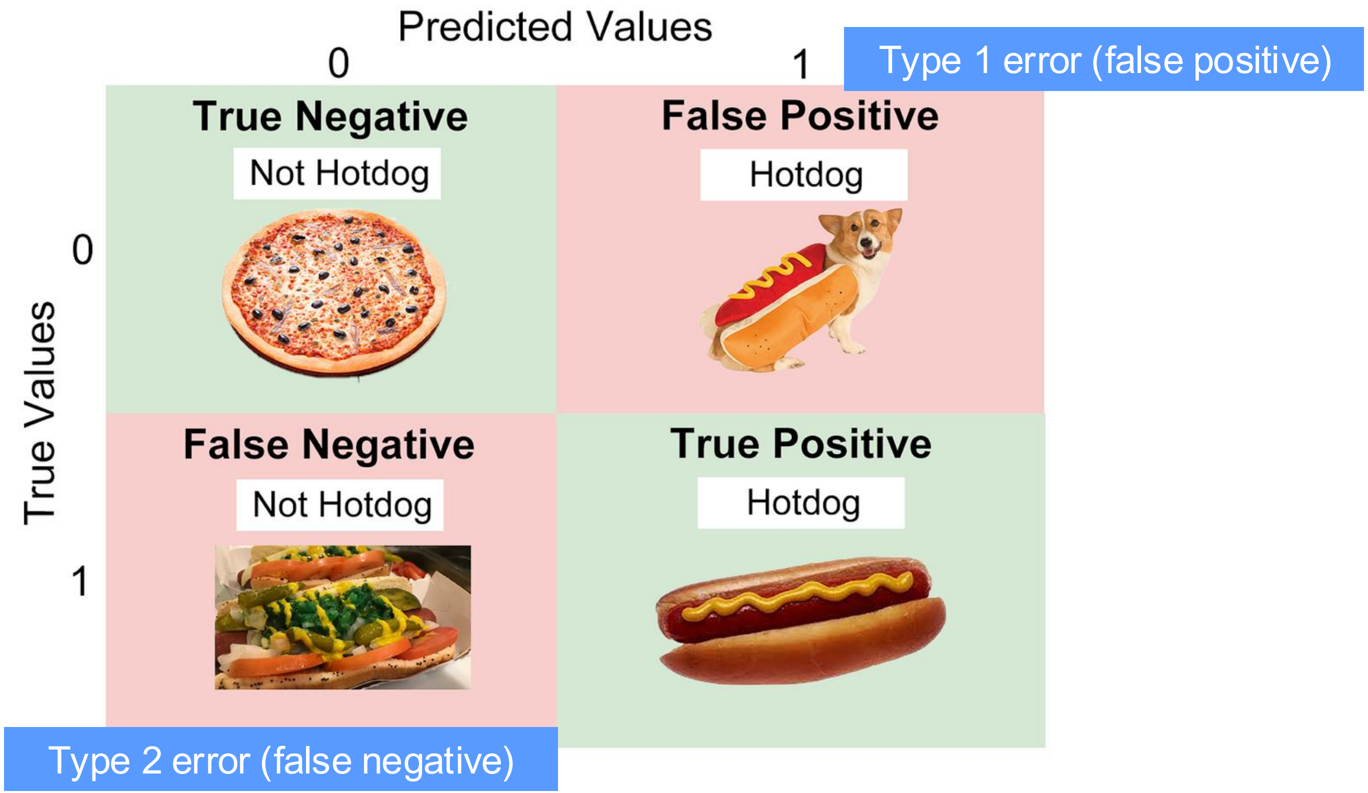
$$Precision = \frac{\Sigma TP}{\Sigma TP + FP}$$

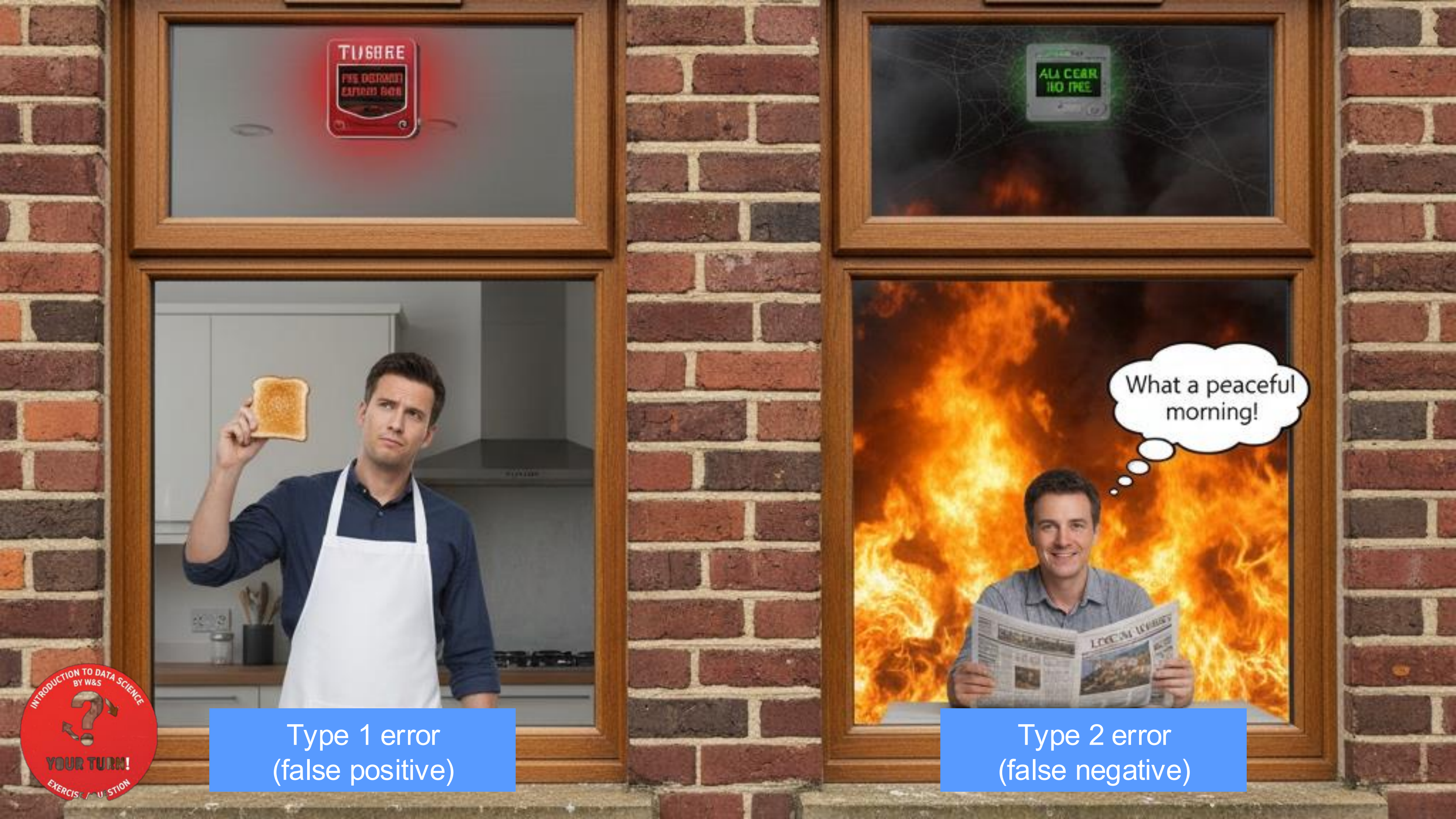
$$Recall = \frac{\Sigma TP}{\Sigma TP + FN}$$

$$Accuracy = \frac{\Sigma TP + TN}{\Sigma TP + FP + FN + TN}$$

- **Precision:** From the predictions of the system, how many the system predicted correctly.
- **Recall:** From the real classes in the dataset, how many the system predicted correctly.

Measuring Performance for Classifications



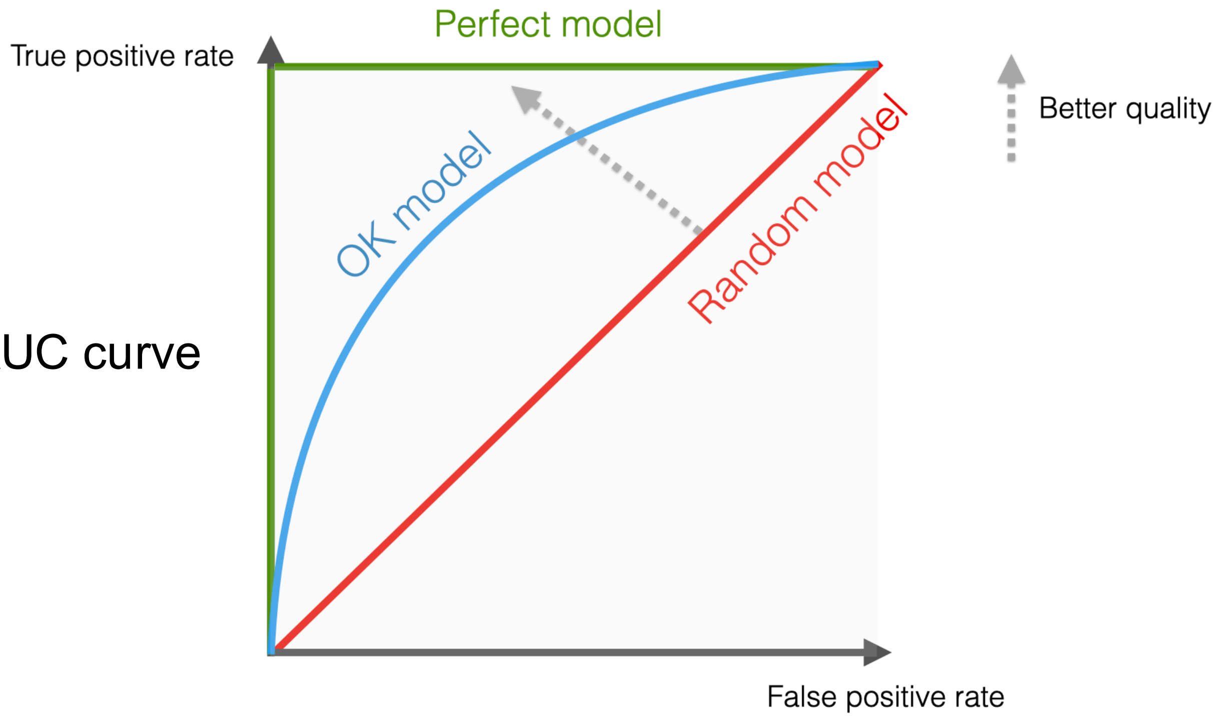

What a peaceful morning!
Type 1 error
(false positive)
Type 2 error
(false negative)

Measuring Performance for Classifications

Exercise: *We have a dataset with 500 bank transactions, from which we know that 90 are fraudulent. The system predicts that 50 are fraudulent. From those 50 predictions, the system got 40 correct predictions. What is the precision and recall of the system?*

Measuring Performance for Classifications

ROC/AUC curve



https://www.evidentlyai.com/classification-metrics/explain-roc-curve

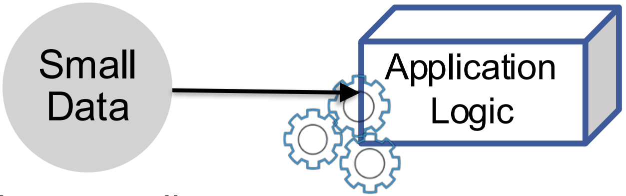
Deployment considerations

Deployment consideration

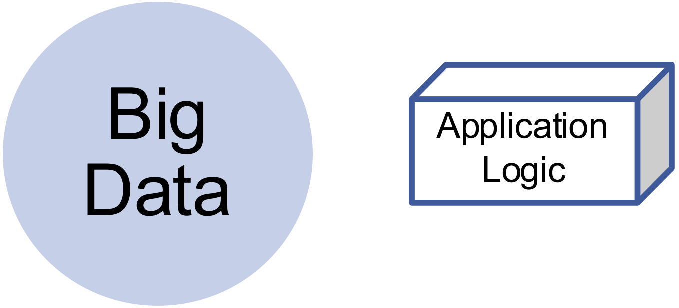
Deployment consideration

Now the hard work begins

Old paradigm: Bring data to computation



New paradigm: Bring computation to data



- Data Management & Database Areas
- Scalable data infrastructures;
- Coping with data diversity;
- End-to-end processing and understanding of data;
- Cloud services; and
- Managing the diverse roles of people in the data life cycle.

Source: "The Beckman Database Research Self-Assessment" Meeting Report. 2013.

Ehtical considerations

Why is there a lack of trust?

BREAKING

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

Siladitya Ray Forbes Staff

Siladitya Ray is a New Delhi-based Forbes news team reporter.

Tesla Autopilot feature was involved in 13 fatal crashes, US regulator says

Federal transportation agency finds Tesla's claims about feature don't match their findings and opens second investigation

A.I. has a discrimination problem. In banking, the consequences can be severe

PUBLISHED FRI, JUN 23 2023·1:45 AM EDT

LLMs collect and share confidential information*

Autonomous cars cause accidents

Unfair AI credit decisions lead to legal and financial damage

https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

https://www.theguardian.com/technology/2024/apr/26/tesla-autopilot-fatal-crash

https://www.cnbc.com/2023/06/23/ai-has-a-discrimination-problem-in-banking-that-can-be-devastating.html

... the risks must be addressed

LLMs collect and share confidential information*

Autonomous cars cause accidents

Unfair AI credit decisions lead to legal and financial damage

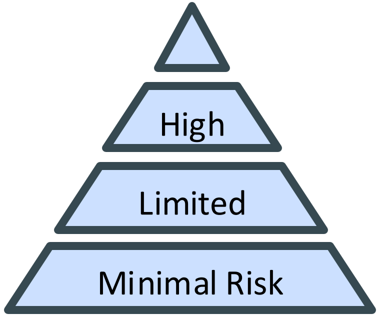
https://www.forbes.com/sites/siladityaray/2023/05/02/samsung-bans-chatgpt-and-other-chatbots-for-employees-after-sensitive-code-leak/

https://www.theguardian.com/technology/2024/apr/26/tesla-autopilot-fatal-crash

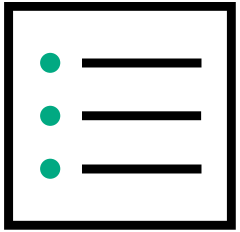
https://www.cnbc.com/2023/06/23/ai-has-a-discrimination-problem-in-banking-that-can-be-devastating.html

Regulation

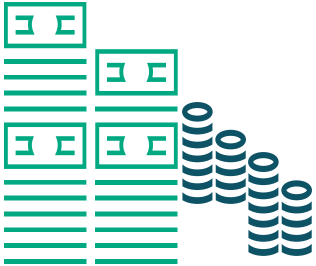
The EU AI Act at a glance



Risk-based approach
Unacceptable, high, limited, minimal



Requirements
Data management, accuracy, cybersecurity, transparency, monitoring, robustness



Penalty
up to €35 million / 7% of global annual turnover

Regulation: The EU AI Act

AI requirements are categorized according to risk, and ignoring requirements is very costly.

Unacceptable risk applications

- Pose a clear threat to people's safety and livelihoods
- Expressly prohibited by AI law
- Example: social scoring, real-time biometric law enforcement

High-risk applications

- Have the potential to cause physical or financial harm to people
- Regulation: must be of good software quality, transparent, and fair
- Example: credit checks, personal employment

Limited-risk applications

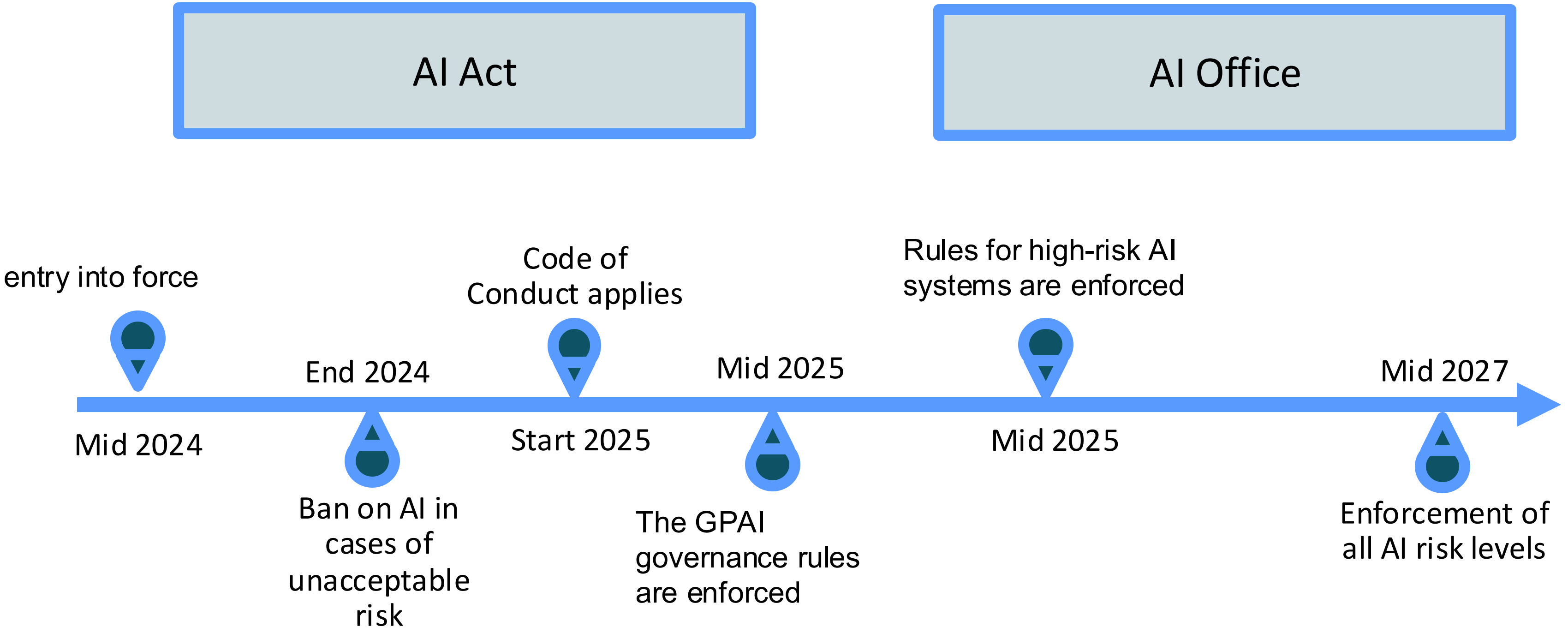
- Low potential for harm
- Regulations: Transparency and cooperation
- Example: Chatbot, deepfakes

Minimal risk applications

- (Almost) no risk to humans
- Regulations: Voluntary code of conduct
- Example: Targeted marketing, spam filters

Penalty of up to €35 million / 7% of global annual turnover

EU AI Act Timeline



1 https://www.alexanderthamm.com/en/blog/eu-ai-act-timeline/
2 https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

“What?”

https://www.autoscout24.be/fr/voiture/voiture-sportive/

Image z

“What?”

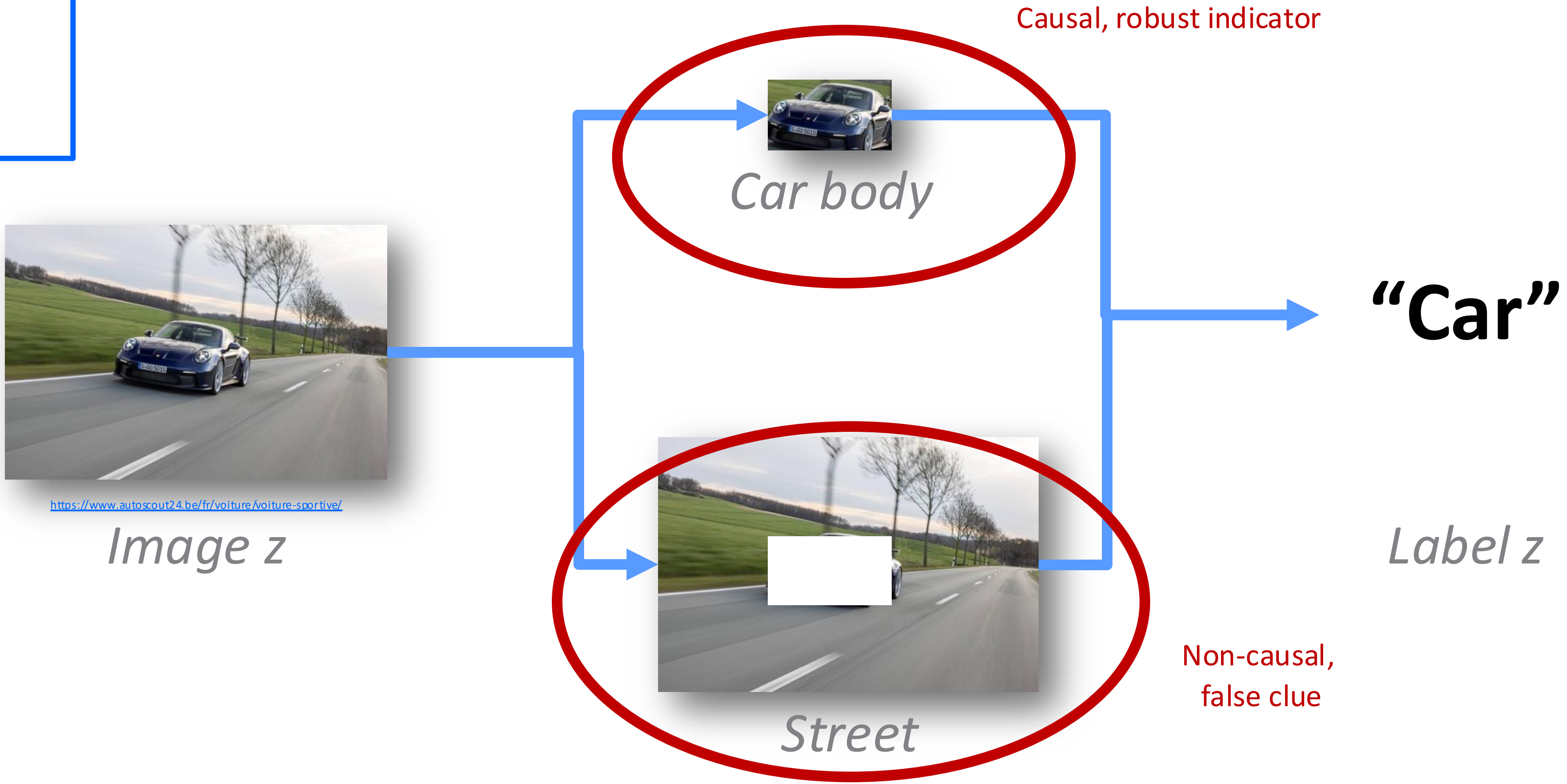
https://www.autoscout24.be/fr/voiture/voiture-sportive/

Image z

“Car”

Label z

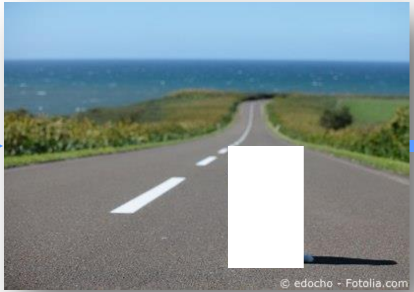
“What?” is not sufficient



“What?” is not sufficient

https://www.hunde-info.de/hundehaftpflicht-versicherung-2115.html

Image x



Street

“Auto”

Label x

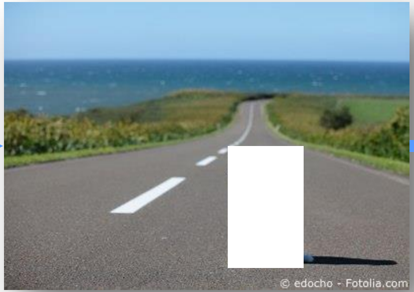
- It is not enough to know “whether” you can solve the problem.
- It also depends on “how” you solve it.
- Robustness and explainability of AI systems.

It needs a 'how'

Nothing prevents the model from using only non-causal, false clues for recognition.

https://www.hunde-info.de/hundehaftpflicht-versicherung-2115.html

Image x



Street

“Car”

Label x

- It is not enough to know “whether” you can solve the problem.
- It also depends on “how” you solve it.
- Robustness and explainability of AI systems.

It needs a 'how'

Nothing prevents the model from using only non-causal, false clues for recognition.

Image x → *Explanation z* → *Label y*

“ML 1.0”: Learn the prediction p(x,y) with the data (x,y)
“ML 2.0”: Learn the prediction p(x,z,y) with the data (x,y)

AI quality is more than just performance

AI quality is more than just performance

- Comprehensible & Explainable
- Fair & Inclusive
- Maintaining privacy
- Robust & performant
- Responsible
- Secure

Trustworthy AI