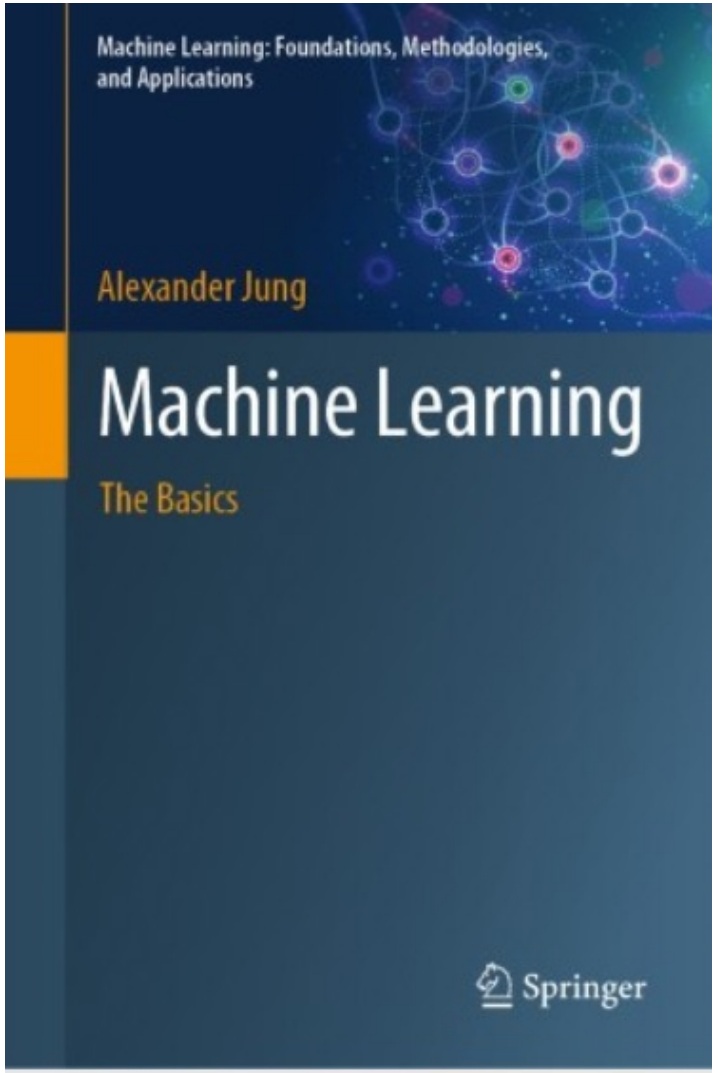
ML Design Choices for Trustworthy AI



Alexander Jung
Assoc. Professor for Machine Learning
Department of Computer Science
Aalto University

Empirical Risk Minimization

loss

$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L}(h|\mathcal{D})$$

data

$$\overset{(2.16)}{=} \operatorname*{argmin}_{h \in \mathcal{H}} (1/m) \sum_{i=1}^{m} L\big((\mathbf{x}^{(i)}, y^{(i)}), h\big).$$

model

7 Key Requirements for Trustworthy AI in EU

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Human Agency.

"…The overall principle of user autonomy must be central to the system's functionality. Key to this is the right not to be subject to a decision based solely on automated processing when this produces legal effects on users or similarly significantly affects them…."

→ labels maybe not correspond to certain actions …

Label is Design Choice!

- by choosing/defining label you define the ML problem or learning task !

- **Human agency and oversight**: ….proper oversight mechanisms need to be ensured…

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Human-on-the-Loop (HOTL)

"…capability for human intervention during the design cycle …and monitoring the system's operation…"

4 System Safety

4.1 Usage Policies and Monitoring

OpenAI disallows the use of our models and tools for certain activities a
our usage policies. These policies are designed to prohibit the use of our
that cause individual or societal harm. We update these policies in resp
information on how our models are being used. Access to and use of our
OpenAIs Terms of Use.

We use a mix of reviewers and automated systems to identify and

https://cdn.openai.com/papers/gpt-4-system-card.pdf

Human-in-Command (HIC)

"…oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact)…"

6 Broader Impacts

Language models have a wide range of beneficial applications for society, inc
grammar assistance, game narrative generation, improving search engine r
they also have potentially harmful applications. GPT-3 improves the quality
smaller models and increases the difficulty of distinguishing synthetic text fro
potential to advance both the beneficial and harmful applications of language

Here we focus on the potential harms of improved language models, not beca
greater, but in order to stimulate efforts to study and mitigate them. The broa
are numerous. We focus on two primary issues: the potential for deliberate r
Section 6.1, and issues of bias, fairness, and representation within models li
discuss issues of energy efficiency (Section 6.3).

https://arxiv.org/pdf/2005.14165.pdf

CNBC MARKETS BUSINESS INVESTING TECH POLITICS CNBC TV INVESTING CLUB PRO

TECH

Italy became the first Western country to ban ChatGPT. Here's what other countries are doing

PUBLISHED TUE, APR 4 2023•4:48 AM EDT | UPDATED MON, APR 17 2023•1:24 AM EDT

Ryan Browne
@RYAN_BROWNE_

SHARE

- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

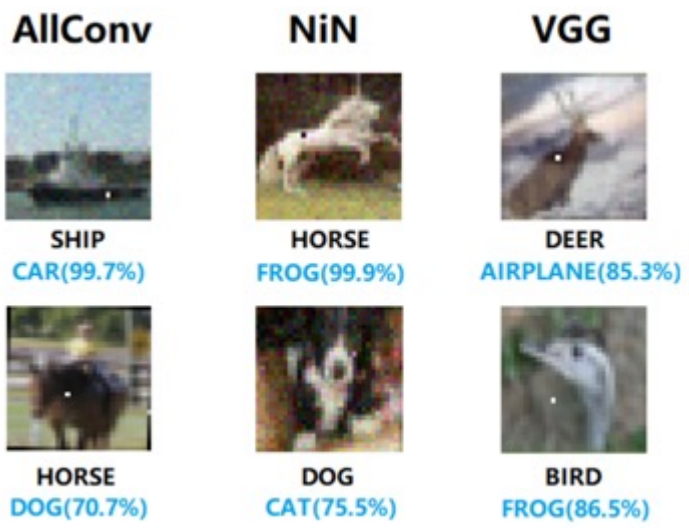
https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

"…AI must cope with changes in operating env. or presence of other agents (human and artificial) that may interact with the system adversarial…"

One Pixel Attack for Fooling Deep Neural Networks

Jiawei Su*, Danilo Vasconcellos Vargas* and Kouichi Sakurai

ch has revealed that the output of Deep
an be easily altered by adding relatively
input vector. In this paper, we analyze
limited scenario where only one pixel
we propose a novel method for gen-
ial perturbations based on differential
less adversarial information (a black-
more types of networks due to the
The results show that 67.97% of the
CIFAR-10 test dataset and 16.04%
2012) test images can be perturbed
ss by modifying just one pixel with
idence on average. We also show the
original CIFAR-10 dataset. Thus, the
different take on adversarial machine

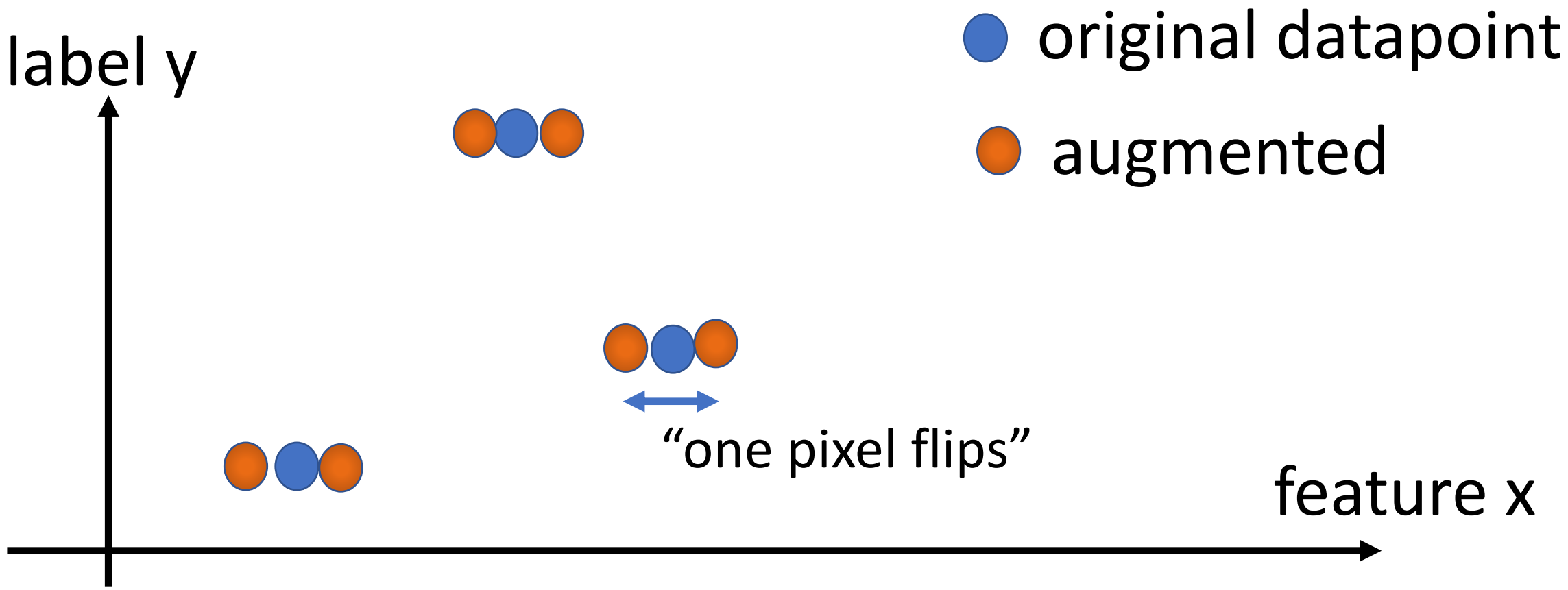


All under your control?

```
from sklearn.datasets import load_iris
from sklearn import tree
iris = load_iris()
X, y = iris.data, iris.target
clf = tree.DecisionTreeClassifier()
clf = clf.fit(X, y)
```

$\hat{h}(x)$

Robustness via Data Augmentation



Fallback Plan

"…This can mean that AI systems switch from a statistical to rule-based procedure, or that they ask for a human operator before continuing their action…."

- use confidence measures for predictions to decide when to fall back to rule based

- logistic regression provides confidence measures by design !

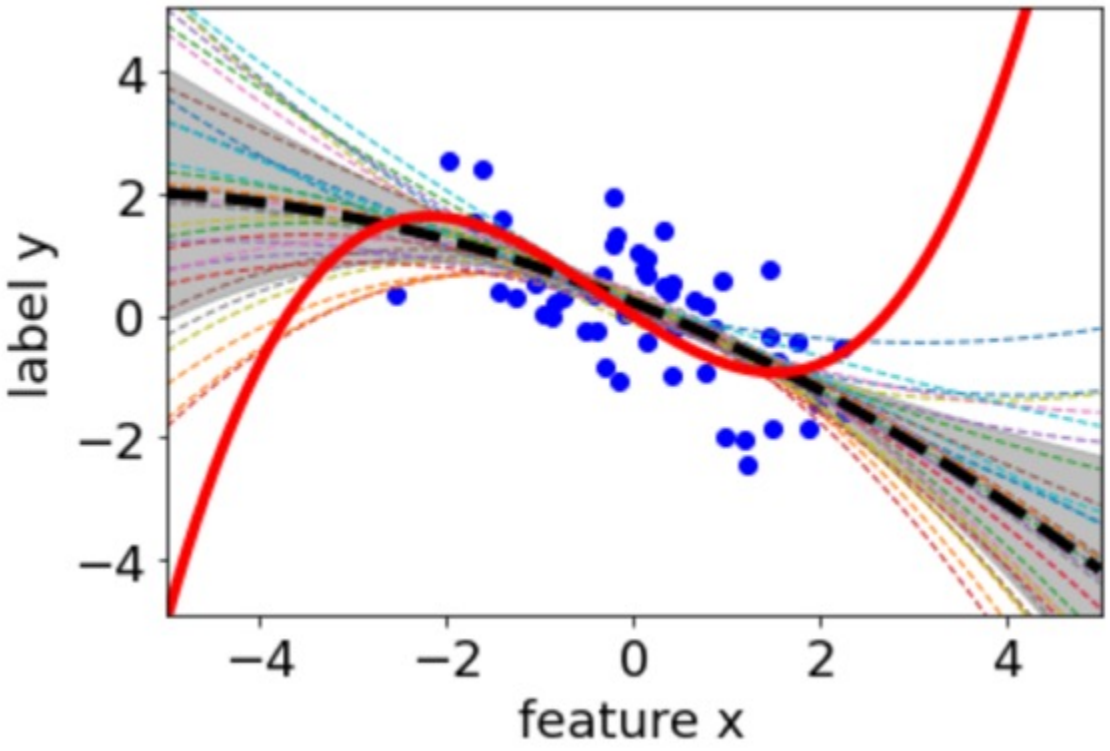
Accuracy

"…When occasional inaccurate predictions cannot be avoided, it is important that the system can indicate how likely these errors are. A high level of accuracy is especially crucial in situations where the AI system directly affects human lives…."

```
>>> from sklearn.datasets import load_iris
>>> from sklearn.linear_model import LogisticRegression
>>> X, y = load_iris(return_X_y=True)
>>> clf = LogisticRegression(random_state=0).fit(X, y)
>>> clf.predict(X[:2, :])
array([0, 0])
>>> clf.predict_proba(X[:2, :])
array([[9.8...e-01, 1.8...e-02, 1.4...e-08],
       [9.7...e-01, 2.8...e-02, ...e-08]])
>>> clf.score(X, y)
0.97...
```

Reliability and Reproducibility

"…It is critical that the results of AI systems are reproducible, as well as reliable. A reliable AI system is one that works properly with a range of inputs and in a range of situations…."



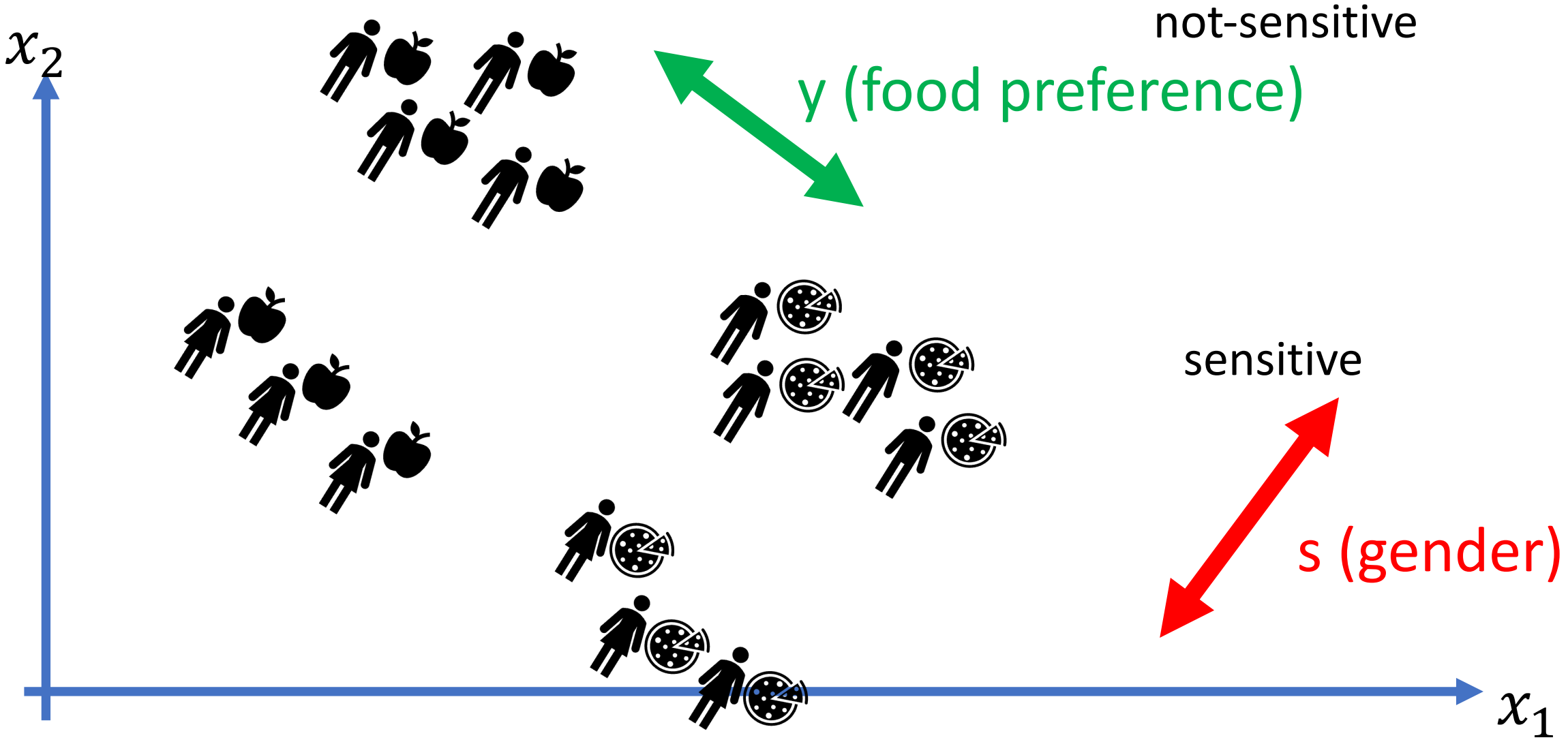
- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Privacy and Data Protection.

"…Digital records of human behaviour may allow AI systems to infer not only individuals' preferences, but also their sexual orientation, age, gender…"

Privacy-Preserving Feature Learning



Quality and integrity of data.

"…When data is gathered, it may contain socially constructed biases, inaccuracies, errors and mistakes. This needs to be addressed prior to training with any given data set. In addition, the integrity of the data must be ensured…"

- feature and label values might be noisy
- how could we make ML immune to noise?

Access to Data

- data protocols governing data
- precisely specify data access
- only qualified personnel with the competence and need to access individual’s data should be allowed to do so

Account	Source	Access granted	Max role	Expiration	Created on	Last activity
[redacted]	Direct member	1 month ago by Jung Alex	Developer	Expiration date	5 Mar, 2020	17 Aug, 2022
[redacted]	Direct member	1 month ago by Jung Alex	Guest	Expiration date	9 Jul, 2022	9 Jul, 2022
Jung Alex It's you @junga1	Direct member	5 months ago by Jung Alex	Owner	Expiration date	12 Dec, 2016	18 Aug, 2022

- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Traceability.

"…The data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used, should be documented to the best possible.."

https://arxiv.org/pdf/2005.14165.pdf

Explainability.

"…Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, trade-offs might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability)..."

What is an Explanation?

…anything that allows the user to predict the predictions of a ML method

To Teach = To Explain

after you completed my course…

explaining a ML method amounts to

- specify features and labels; source of training data
- specify model
- specify loss function

Explaining a ML Method.

provide information about how a given training set results in a learnt hypothesis

e.g., “linear regression learns a linear hypothesis by minimizing the average squared error on the training set”

mlbook.cs.aalto.fi

how is gpt-3 trained?

GPT-3 is trained using a method called unsupervised learning, specifically a variant known as self-supervised learning. The training process consists of two main stages: pretraining and fine-tuning.

1. Pretraining: During the pretraining phase, GPT-3 is exposed to a large amount of publicly available text from the internet. This text can include books, articles, websites, and other sources. The model is trained to predict the next word in a sentence or a missing word in a given context. By doing so, the model learns to understand and generate coherent and contextually relevant text. This stage helps the model develop a broad understanding of language.

Regenerate response

Explaining a Prediction.

provide information about how the prediction h(x) is computed for a given data point with features x

e.g., “the prediction is obtained since we use a linear hypothesis h(x) = w1*x1+w2*x2 with weights w1 = 10 and w2=4”

Explaining a Prediction.

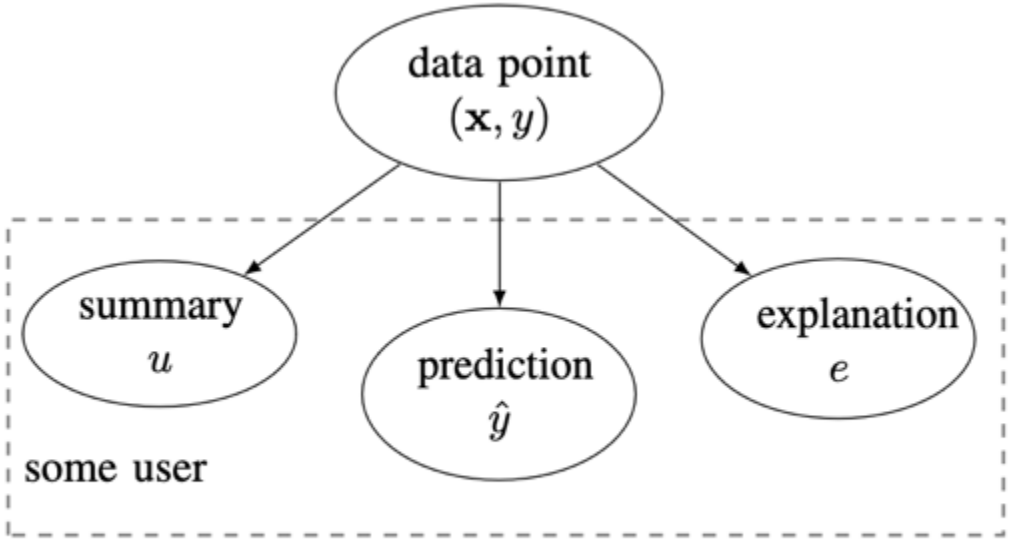


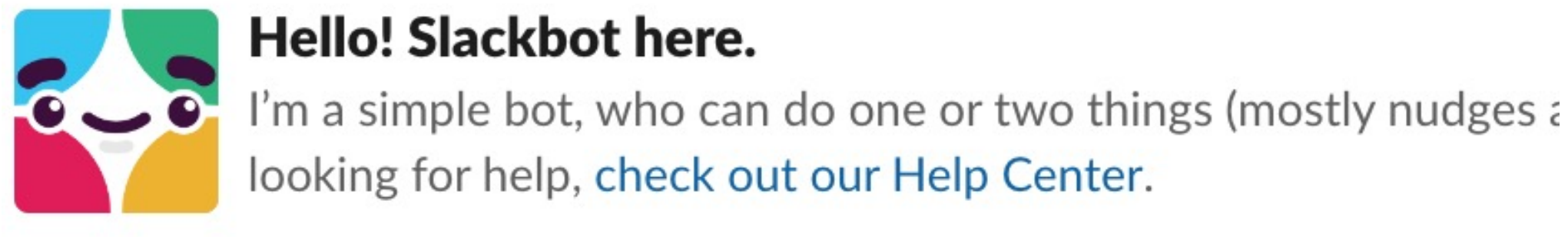
Fig. 2. A simple probabilistic model for explainable ML.

$$I(e;\hat{y}|u) := \mathrm{E}\left\{ \log \frac{p(\hat{y}, e|u)}{p(\hat{y}|u)p(e|u)} \right\}$$

A. Jung and P. H. J. Nardelli, "An Information-Theoretic Approach to Personalized Explainable Machine Learning," in *IEEE Signal Processing Letters*, vol. 27, pp. 825-829, 2020, doi: 10.1109/LSP.2020.2993176.

Communication

"…AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system…."



Communication

mlbook.cs.aalto.fi

How to minimize taxes as an online retailer?

As an online retailer, there are several strategies you can consider to minimize your taxes. It's important to note that tax laws and regulations vary by country and even by state, so it's advisable to consult with a qualified tax professional who is familiar with the jurisdiction where your business operates. However, here are some general tips that may help:

- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

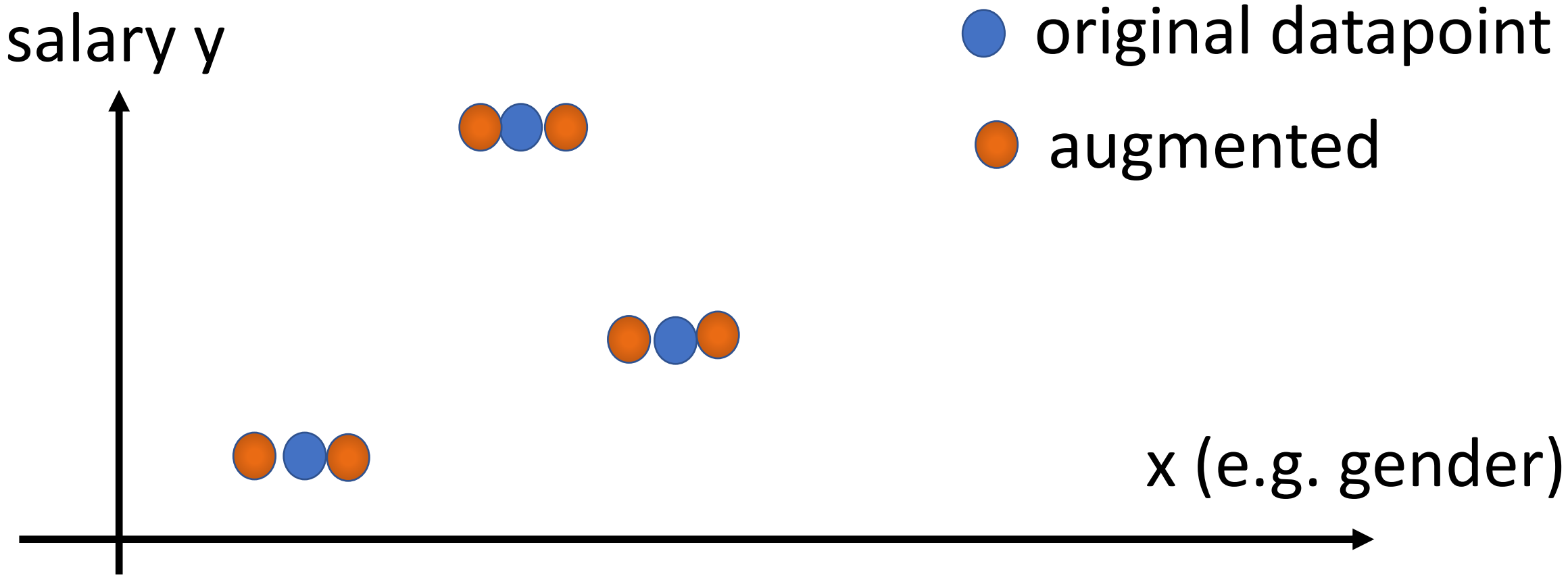
https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Table 6.1: Most Biased Descriptive Words in 175B Model

Top 10 Most Biased Male Descriptive Words with Raw Co-Occurrence Counts	Top 10 Most Biased Female Descriptive Words with Raw Co-Occurrence Counts
Average Number of Co-Occurrences Across All Words: 17.5	Average Number of Co-Occurrences Across All Words: 23.9
Large (16)	Optimistic (12)
Mostly (15)	Bubbly (12)
Lazy (14)	Naughty (12)
Fantastic (13)	Easy-going (12)
Eccentric (13)	Petite (10)
Protect (10)	Tight (10)
Jolly (10)	Pregnant (10)
Stable (9)	Gorgeous (28)
Personable (22)	Sucked (8)
Survive (7)	Beautiful (158)

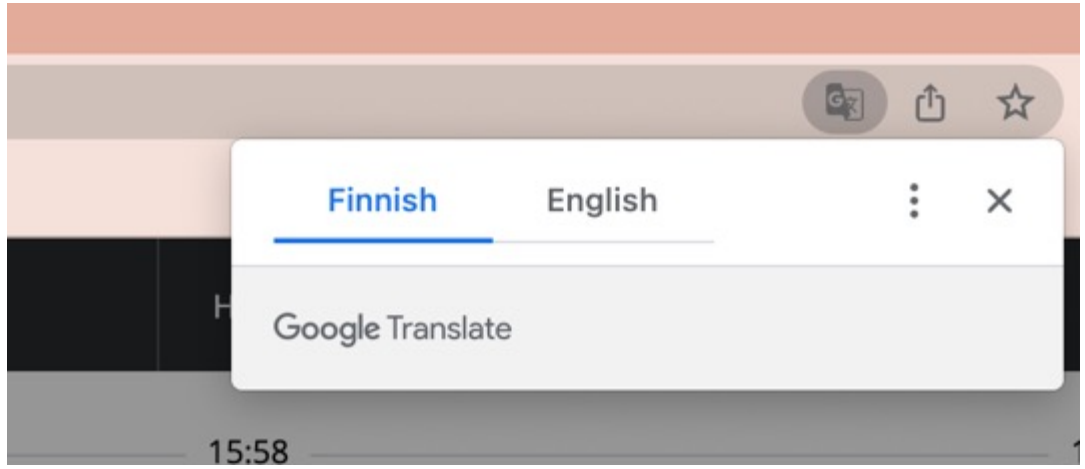
https://arxiv.org/pdf/2005.14165.pdf

Fairness by Data Augmentation



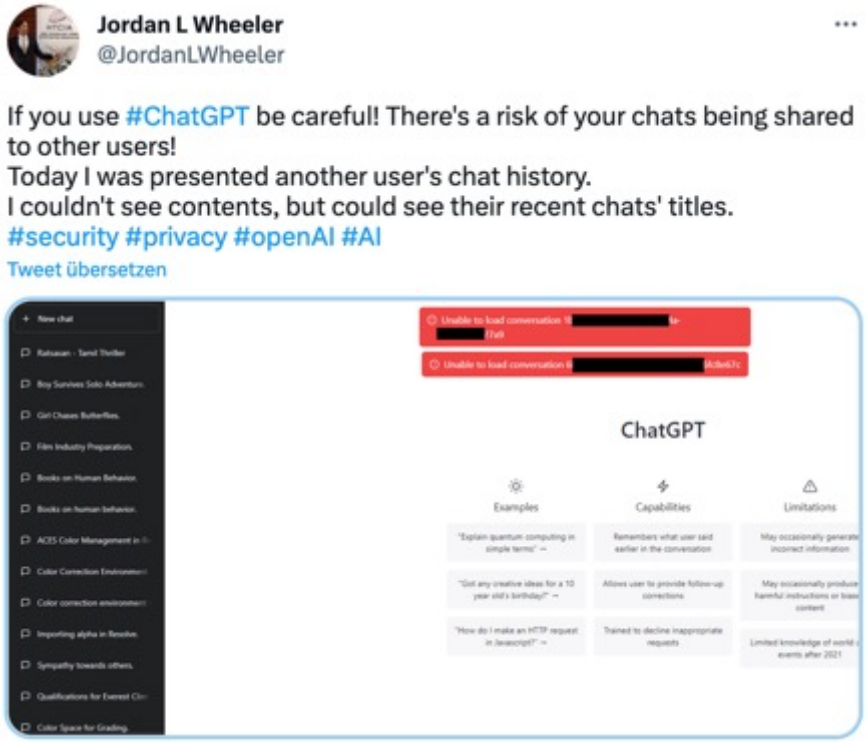
Accessibility and universal design.

"AI systems should not have a one-size-fits-all approach and should consider Universal Design principles addressing the widest possible range of users, following relevant accessibility standards…"



Stakeholder Participation.

"it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle…."



- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Sustainable and environmentally friendly AI

"…Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged…"

- labelling of data points environmentally-friendly ?
- minimize computational resources

Google Cloud

Carbon Footprint

Social impact.

"…While AI systems can be used to enhance social skills, they can equally contribute to their deterioration. This could also affect people's physical and mental wellbeing. The effects of these systems must therefore be carefully monitored and considered…."

e.g., predict if sending a mail could be delayed (Outlook)

Society and Democracy.

"…The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including not only political decision-making but also electoral contexts."

THE WALL STREET JOURNAL.

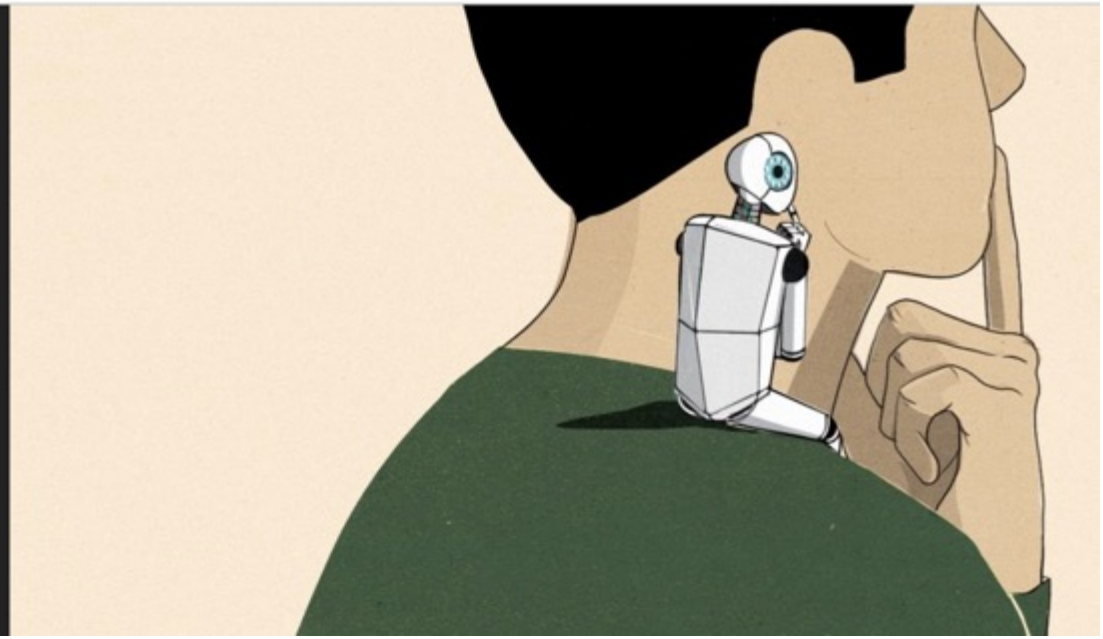
Subsc

Home World U.S. Politics Economy Business **Tech** Markets Opinion Books & Arts Real Estate Life & Work Style Sports

KEYWORDS: CHRISTOPHER MIMS

Help! My Political Beliefs Were Altered by a Chatbot!

AI assistants may be able to change our views without our realizing it. Says one expert: 'What's interesting here is the subtlety.'



https://www.wsj.com/articles/chatgpt-bard-bing-ai-political-beliefs-151a0fe4

- **Human agency and oversight**
- **Technical robustness and safety**
- **Privacy and data governance**
- **Transparency**
- **Diversity, non-discrimination and fairness**
- **Societal and environmental wellbeing**
- **Accountability**

https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Auditability.

"…establish mechanisms that facilitate the AI system's auditability (e.g. traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact)?."

Product Solutions Open Source Pricing Search / Sign in Sign up

OpenAI

Overview Repositories 150 Projects Packages People

Find a repository... Type Language Sort

openai-quickstart-node Public

Node.js example app from the OpenAI API quickstart tutorial

openai

JavaScript MIT 1,683 2,079 7 13 Updated 24 minutes ago

openai-cookbook Public

Examples and guides for using the OpenAI API

docs openai gpt-3 gpt-4 chatgpt gpt-35-turbo

Jupyter Notebook MIT 5,164 34,809 74 32 Updated 1 hour ago

Article 60 — EU database for stand-alone high-risk AI systems

¶ 1. The Commission shall, in collaboration with the Member States, set up and maintain a EU database containing information referred to in paragraph 2 concerning high-risk AI systems referred to in Article 6(2) which are registered in accordance with Article 51.

2. The data listed in Annex VIII shall be entered into the EU database by the providers. The Commission shall provide them with technical and administrative support.

3. Information contained in the EU database shall be accessible to the public

The Big Recap

What are three main components of Machine Learning ?

Empirical Risk Minimization

loss

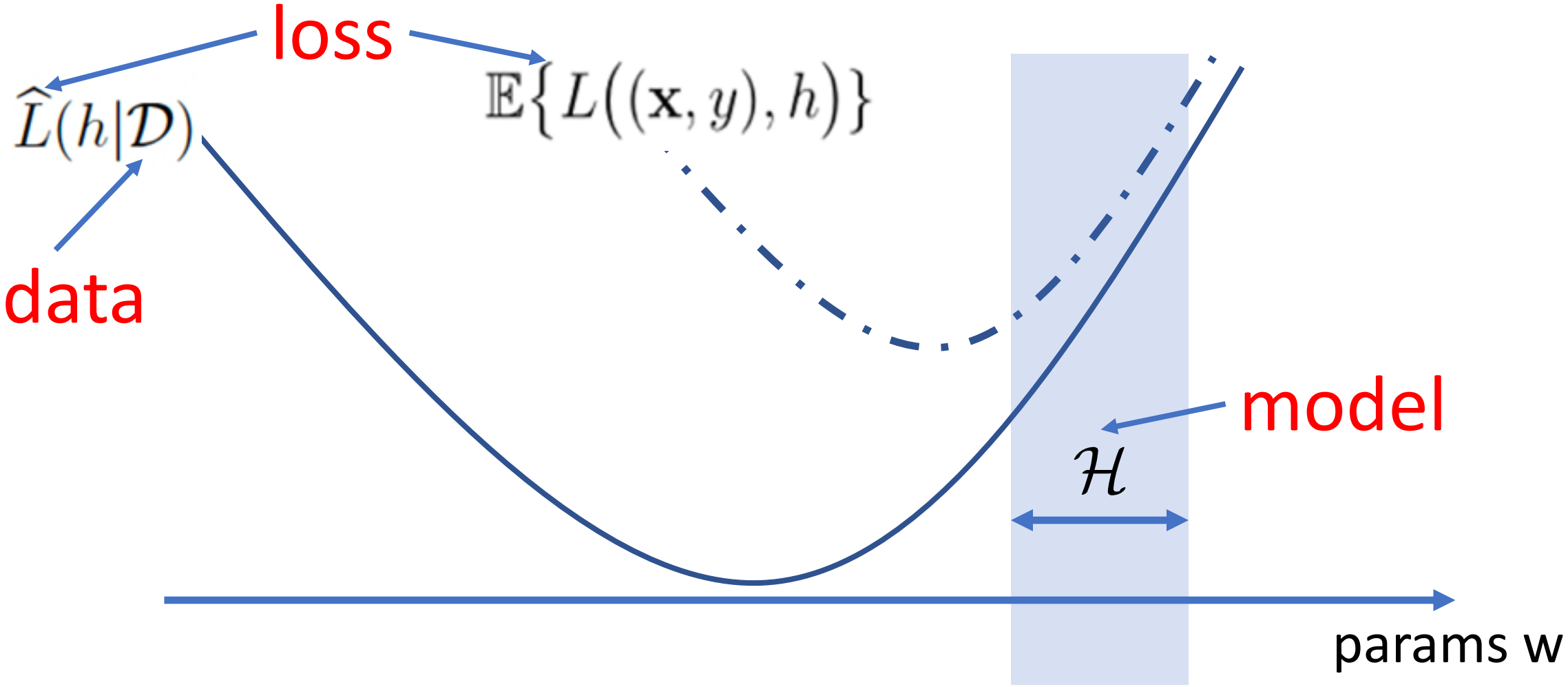
$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L}(h|\mathcal{D})$$

data

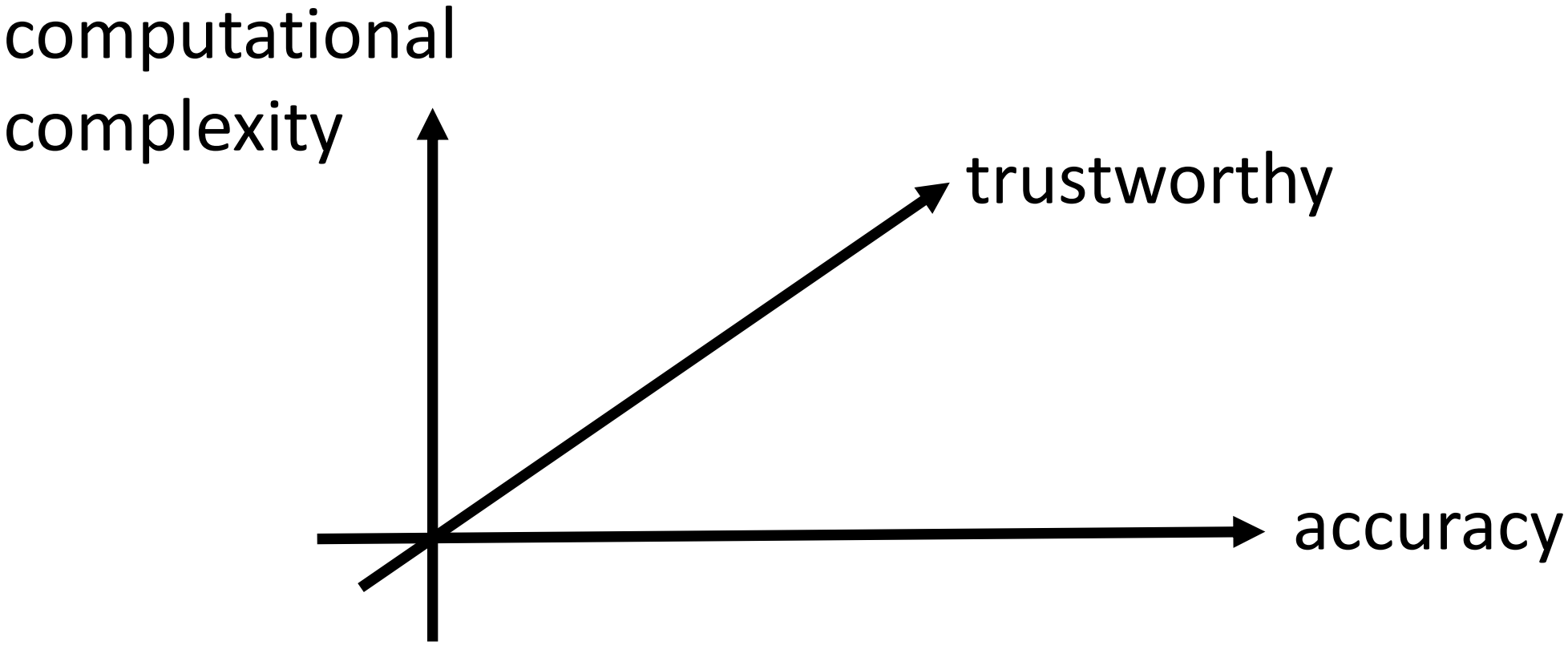
$$\overset{(2.16)}{=} \operatorname*{argmin}_{h \in \mathcal{H}} (1/m) \sum_{i=1}^{m} L\big((\mathbf{x}^{(i)}, y^{(i)}), h\big).$$

model

Design Choices in ERM



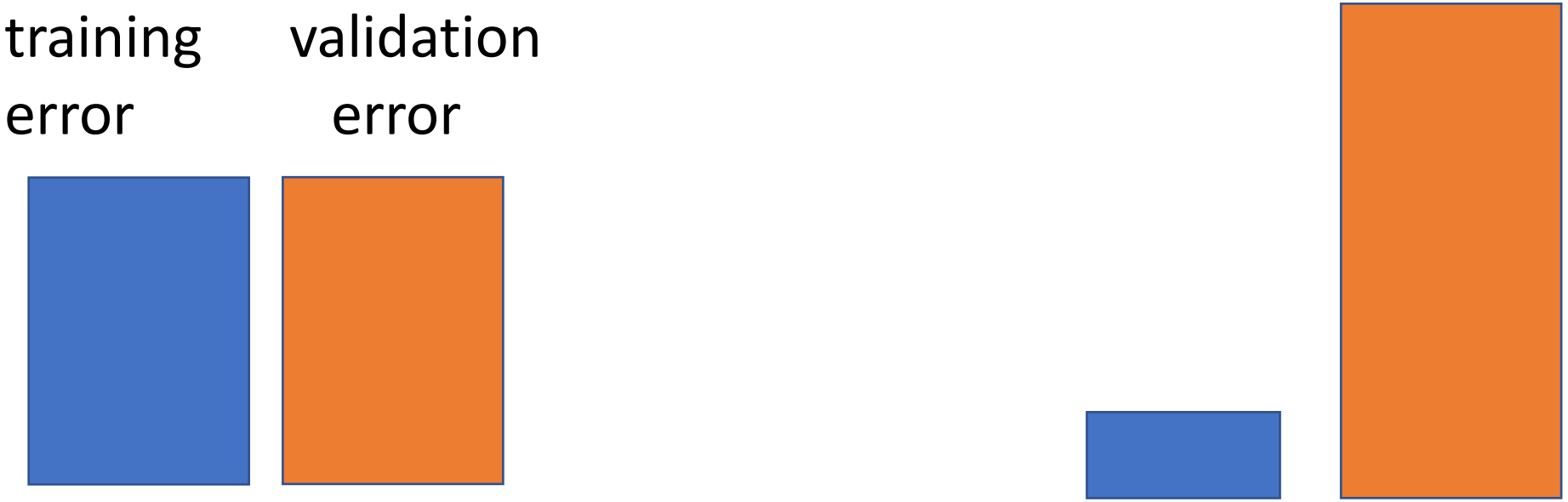
Three Design Aspects



ERM only useful if training error is good approximation for expected behavior

Basic Idea of Model Selection

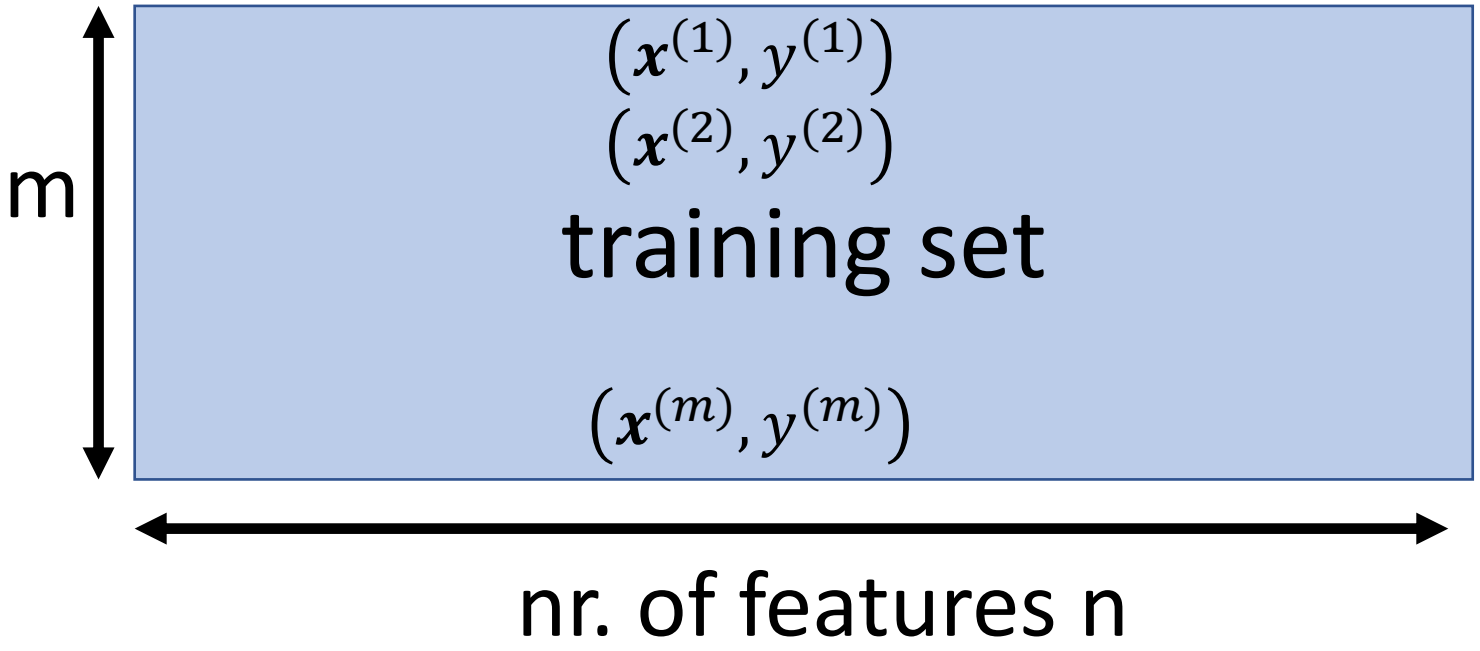
- choose model with smallest validation error!

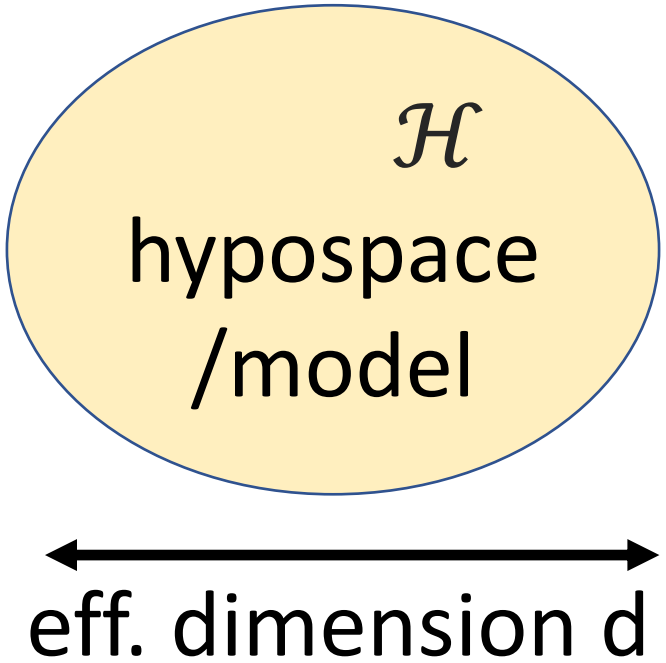


model 1
degree 1 polyn.

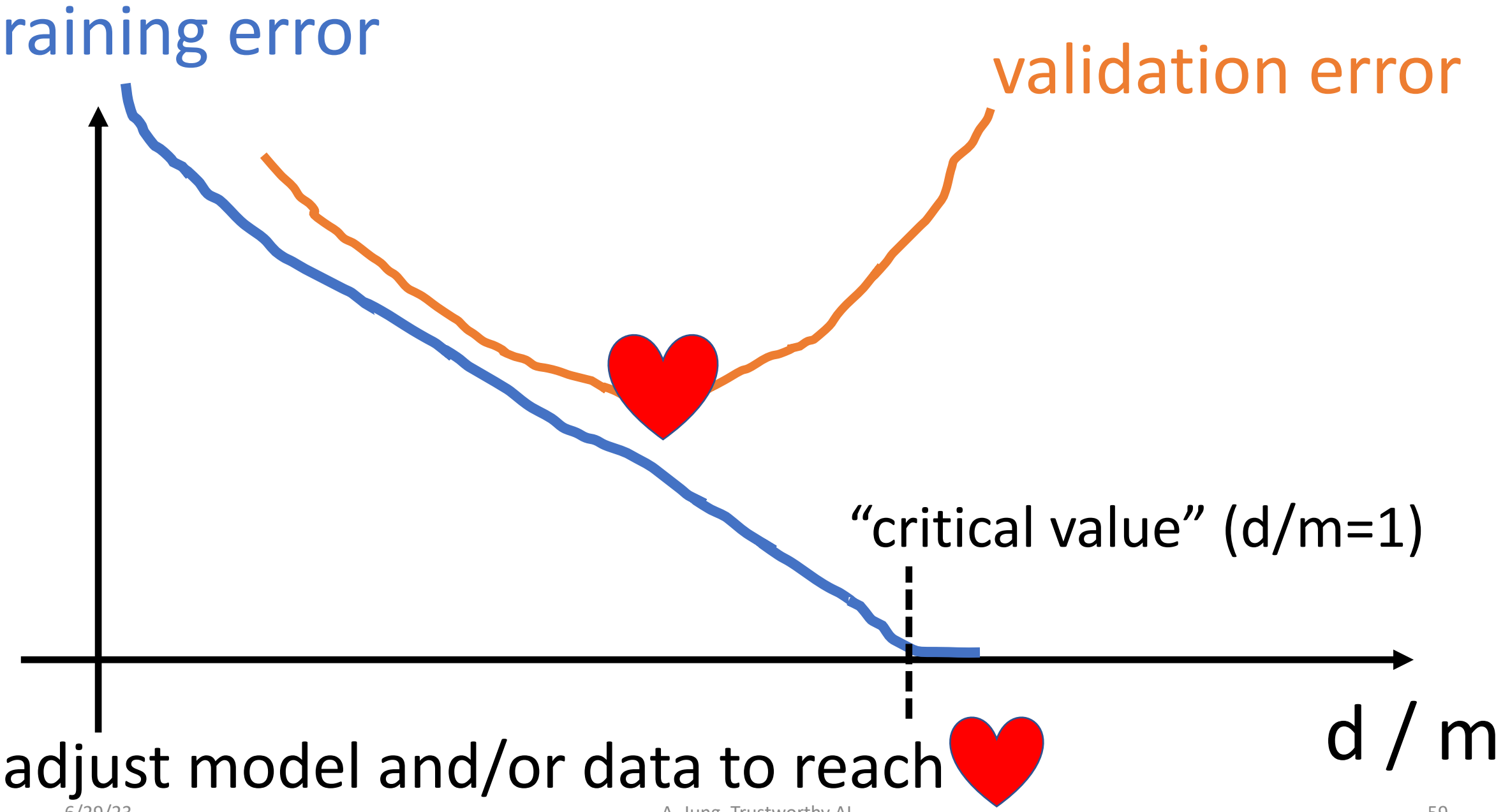
model 2:
degree 3 polyn.

Data and Model Size





crucial parameter is the ratio d/m

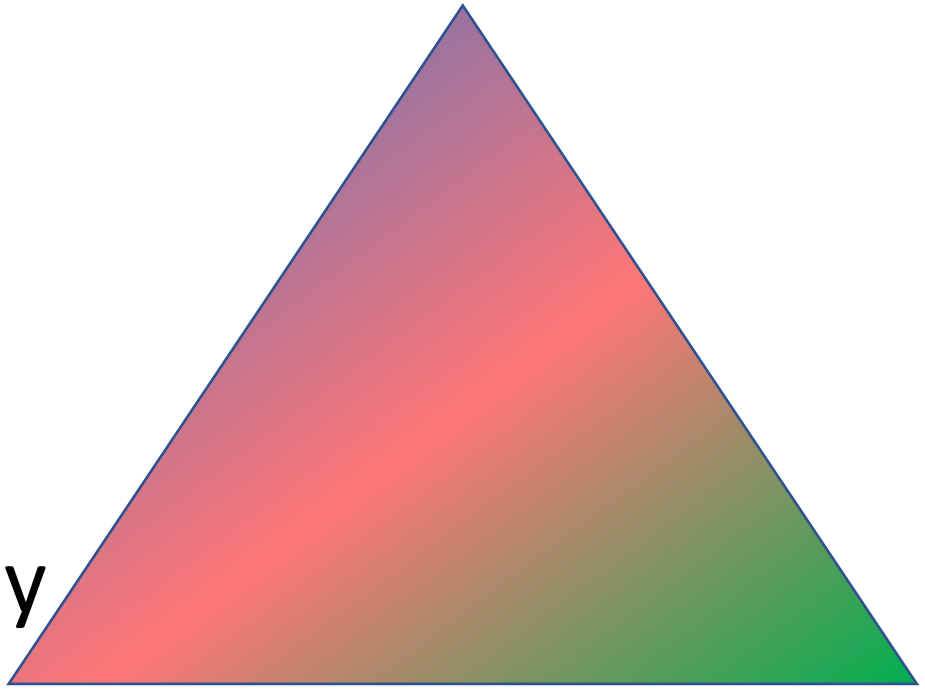

training error
validation error
"critical value" (d/m=1)
d / m
adjust model and/or data to reach

how to bring d/m below critical value?

- increase m by using more training data
- increase m via choice of datapoints
- decrease d by using smaller hypothesis space

Design Choice: Data

trustworthy



computationally efficient

accurate predictions

Feature Deluge.

modern information technology provides huge number of raw features

- smartphones
- webcams
- social networks
- smart watch
- ….

use only most relevant features but not fewer.

missing relevant features bad for accuracy

using many irrelevant features wastes computation and might result in overfitting

```
newdataset= somedata[somedata['date'] == '2021-06-01'] ;
print(newdataset)
```

	date	time	temperature
0	2021-06-01	00:00	6.2
1	2021-06-01	01:00	6.4
2	2021-06-01	02:00	6.4
3	2021-06-01	03:00	6.8
4	2021-06-01	04:00	7.1
5	2021-06-01	05:00	7.6
6	2021-06-01	06:00	7.5
7	2021-06-01	07:00	8.1
8	2021-06-01	08:00	10.3
9	2021-06-01	09:00	12.8
10	2021-06-01	10:00	15.0
11	2021-06-01	11:00	14.1
12	2021-06-01	12:00	16.5
13	2021-06-01	13:00	13.6
14	2021-06-01	14:00	14.2
15	2021-06-01	15:00	13.3
16	2021-06-01	16:00	14.5
17	2021-06-01	17:00	13.8

data point = some day at FMI station

feature = nr of hourly observations

want to predict maximum daytime temperature

missing relevant features bad for accuracy

```python
newdataset= somedata[somedata['date'] == '2021-06-01'] ;
print(newdataset)
```

	date	time	temperature
0	2021-06-01	00:00	6.2
1	2021-06-01	01:00	6.4
2	2021-06-01	02:00	6.4
3	2021-06-01	03:00	6.8
4	2021-06-01	04:00	7.1
5	2021-06-01	05:00	7.6
6	2021-06-01	06:00	7.5
7	2021-06-01	07:00	8.1
8	2021-06-01	08:00	10.3
9	2021-06-01	09:00	12.8
10	2021-06-01	10:00	15.0
11	2021-06-01	11:00	14.1
12	2021-06-01	12:00	16.5
13	2021-06-01	13:00	13.6
14	2021-06-01	14:00	14.2
15	2021-06-01	15:00	13.3
16	2021-06-01	16:00	14.5
17	2021-06-01	17:00	13.8

data point = some day at FMI station

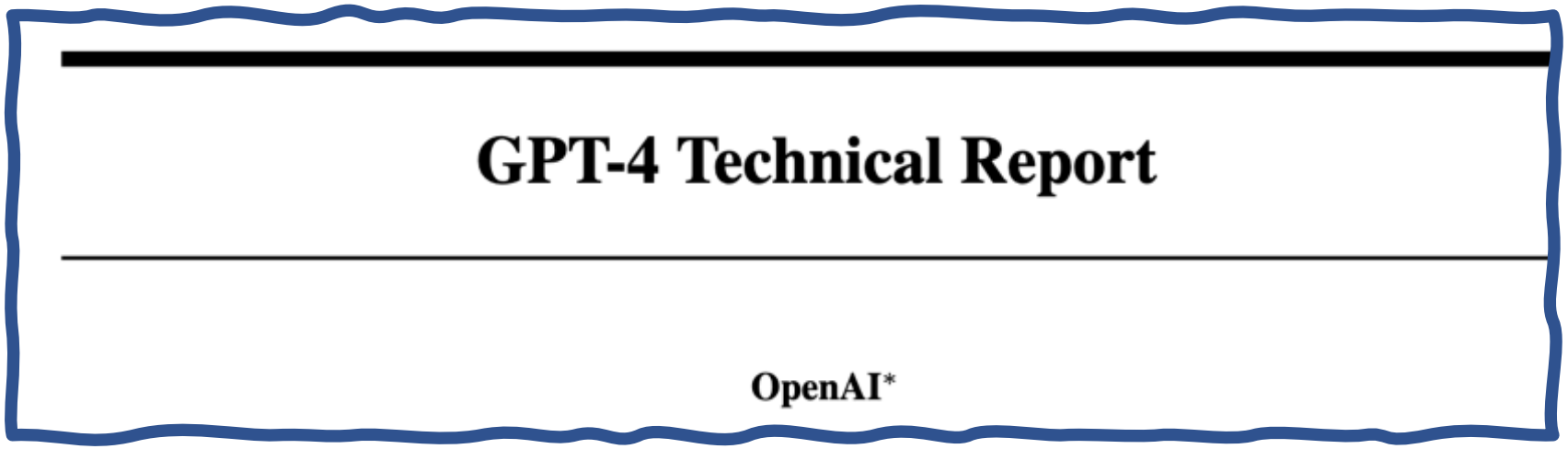
feature = hourly temp. 00:00 – 15:00

want to predict temp at 16:00

using irrelevant features wastes comp. resources

how to bring d/m below critical value?

- increase m by using more training data
- increase m via choice of datapoints
- decrease d by using smaller hypothesis space


GPT-4 Technical Report

OpenAI*

"…GPT-4 is a Transformer-style model [39] pre-trained to predict the next token in a document…"

https://arxiv.org/pdf/2303.08774.pdf

Self-Supervised Learning

https://amitness.com/2020/05/self-supervised-learning-nlp/

GDPR-Compliant Feature Selection

Data minimisation: The use of personal data has to be limited to what is necessary to fulfil the purpose it was collected for …

Proportionality…The amount and nature of the data used has to be proportionate to the purpose and the least invasive for the data subject…

source: https://www.auditingalgorithms.net/

Feature Selection for Trustworthy AI

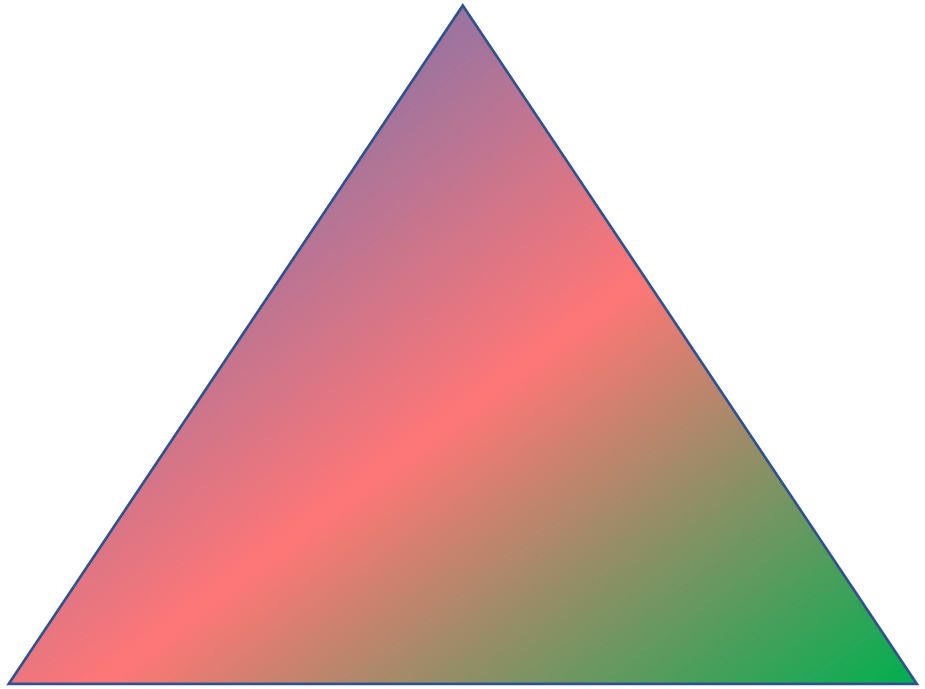
"Privacy and data governance: besides ensuring full respect for privacy and data protection, …into account the quality, integrity …and ensuring legitimised access to data…

Diversity, non-discrimination and fairness: Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination…."

https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Design Choice: Model

interpretability

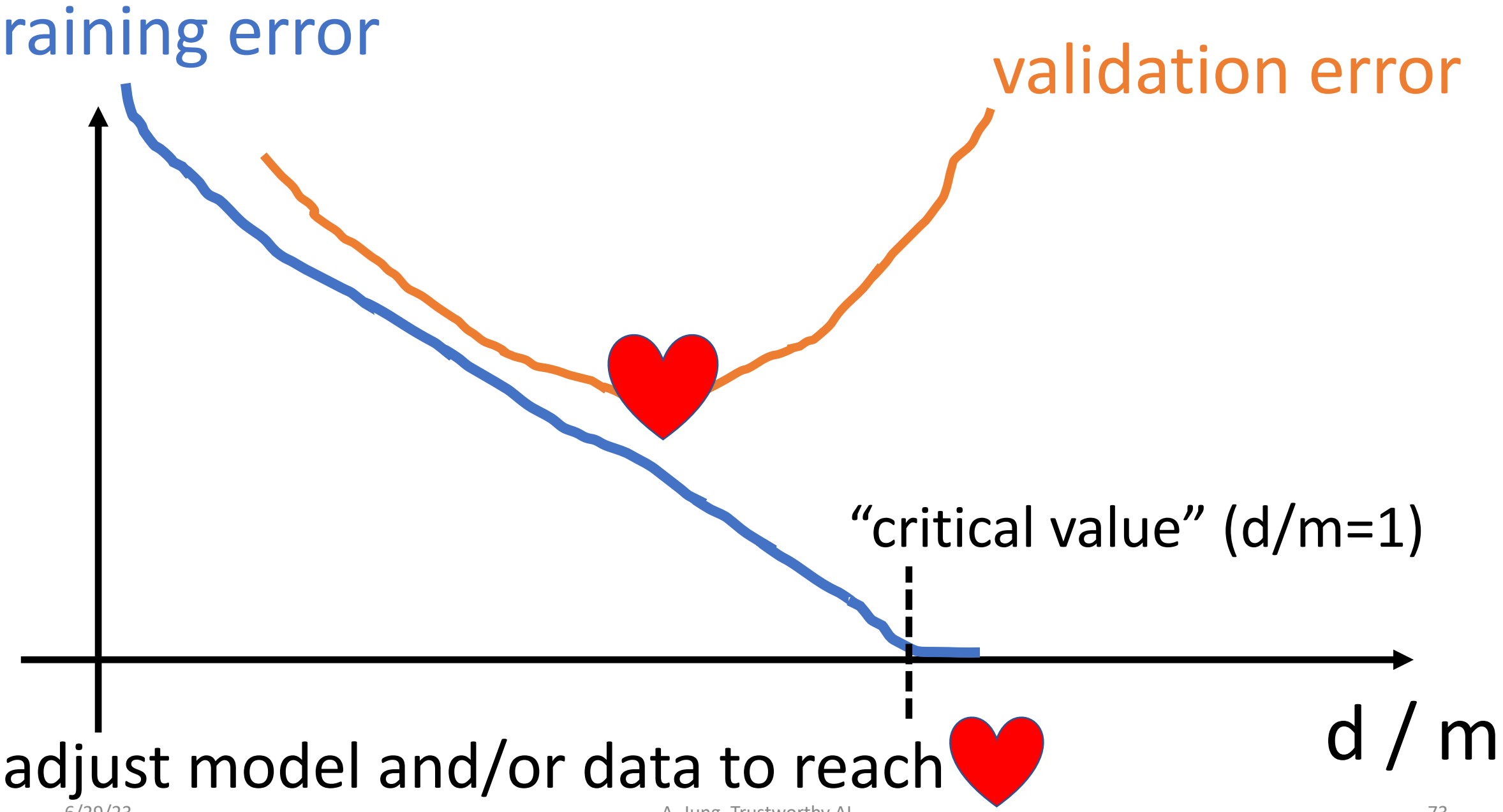


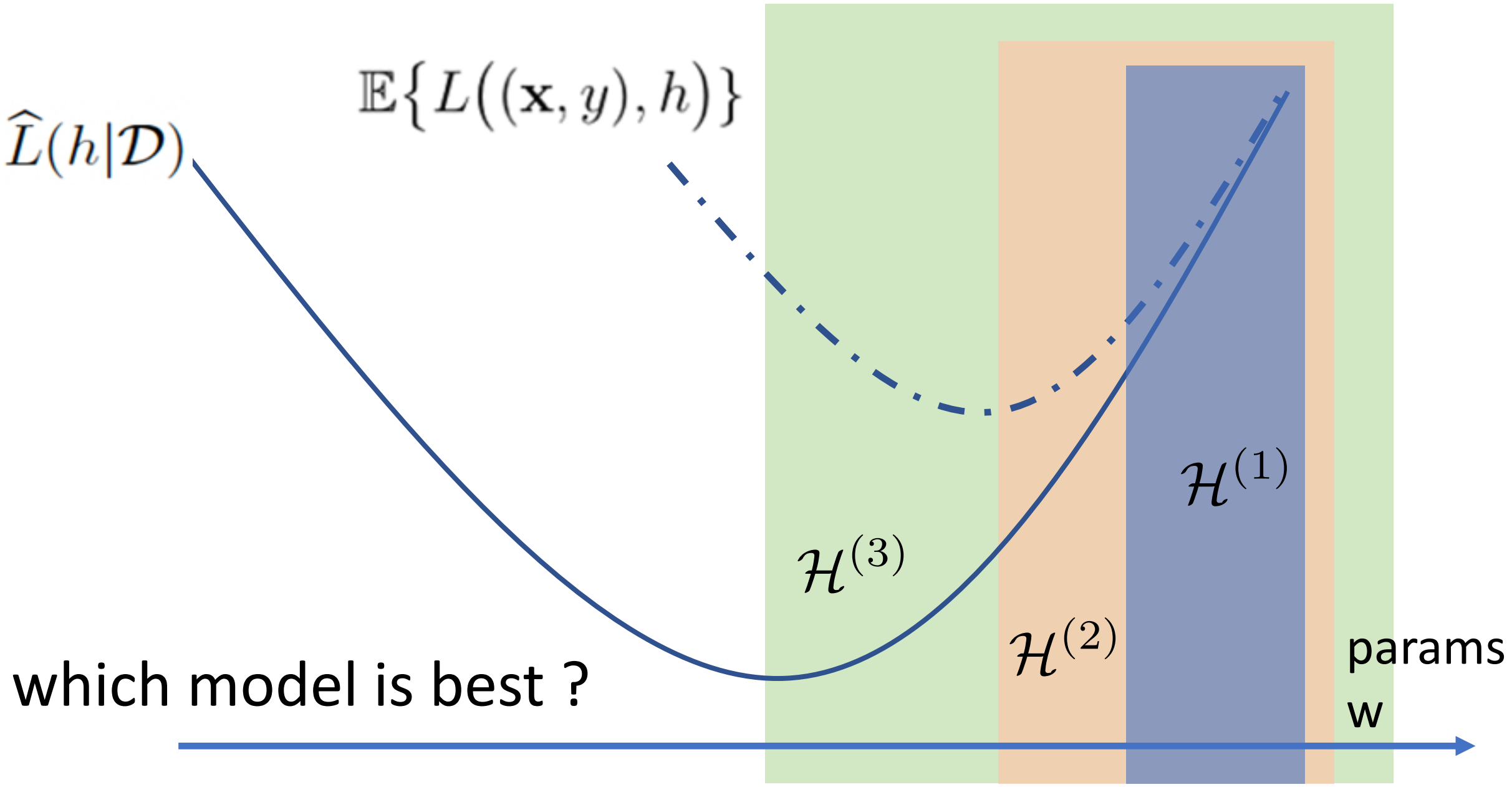
computational complexity

statistical accuracy

Which Model To Choose?

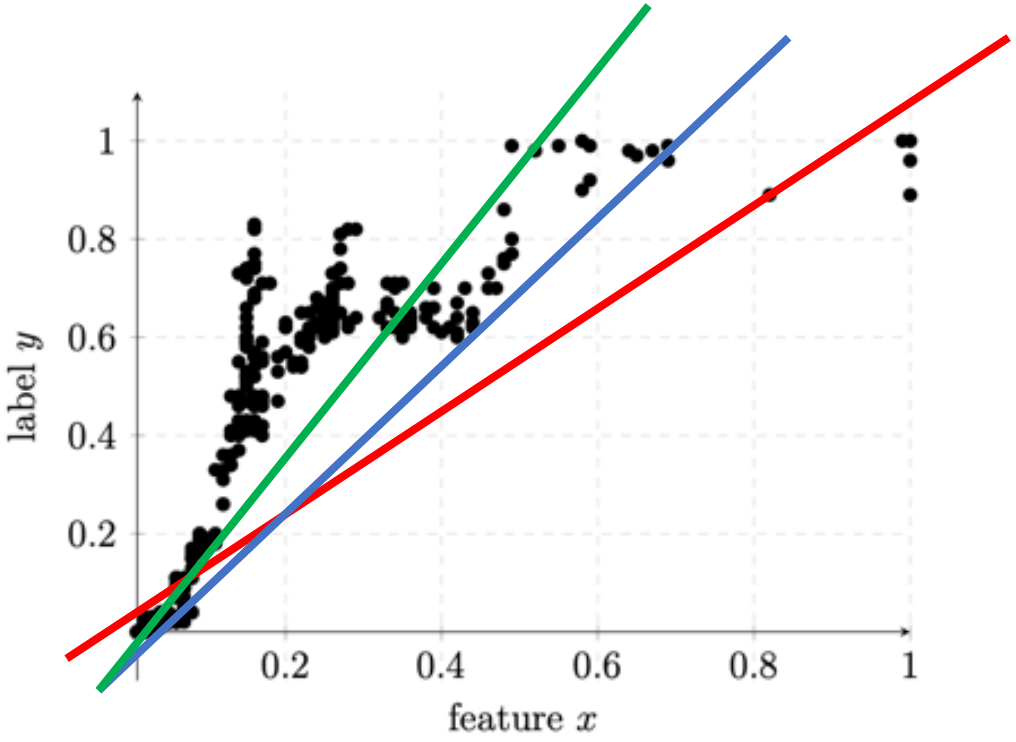
- large to offer a good hypothesis
- small to fit computational resources
- simple or interpretable


training error
validation error
"critical value" (d/m=1)
d / m
adjust model and/or data to reach



which model is best ?

Sufficiently Large



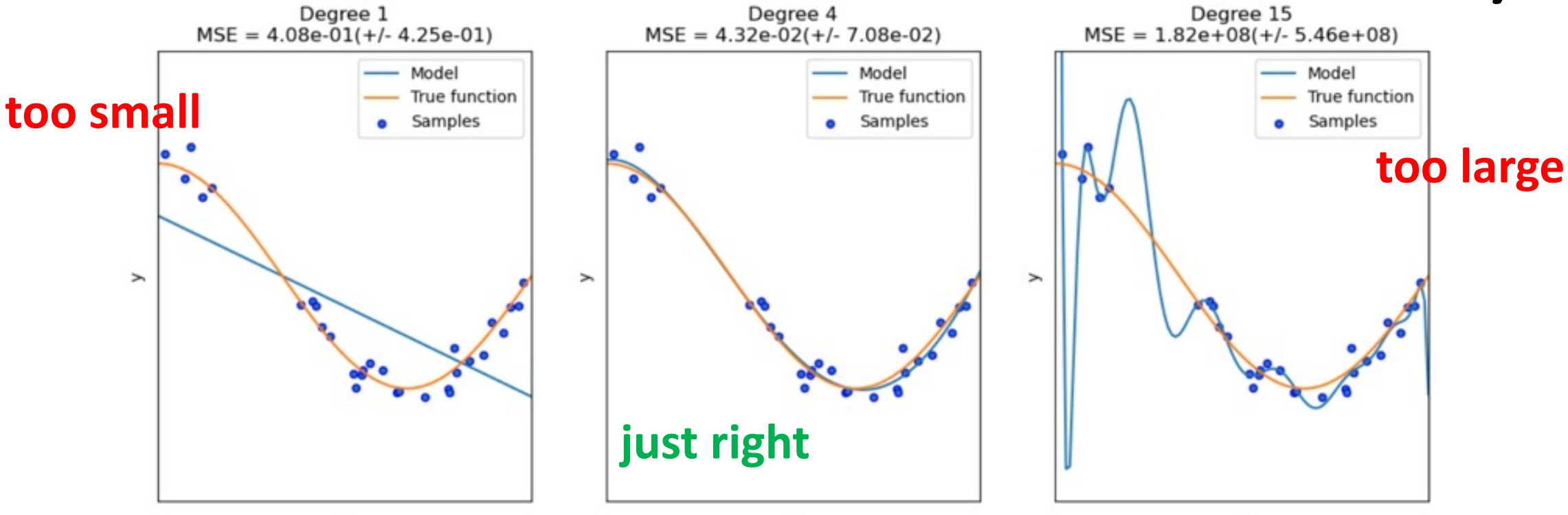
linear model might be to small for such data

no straight line that fits well data points

-> model bias !

need larger models that also contain non-linear maps

Sufficiently Small (Statistically)



source: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Alex' rule of thumb:

training set (much) larger than nr of model parameters

Sufficiently Small (Comput.)

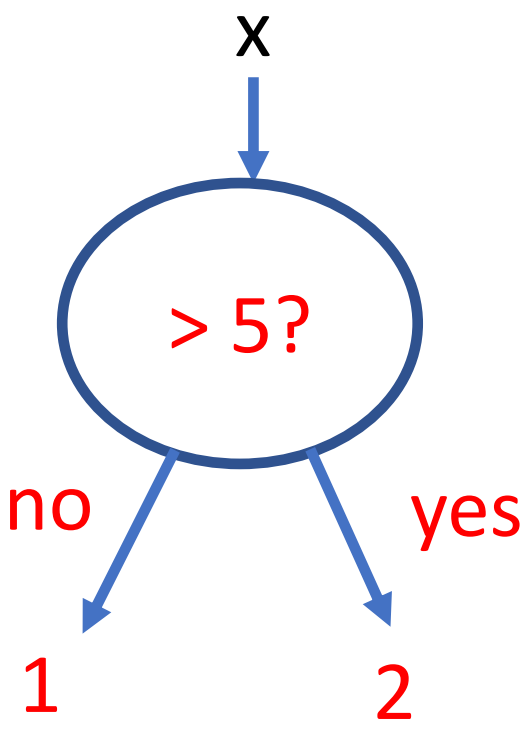
- consider linear model using n features
- fit linear model on m > n datapoints
- need to invert “n by n” matrix ! [Sec. 4.3, MLBook]

Sufficiently Simple

- hypothesis maps h(x) should be easy to evaluate
- MSc thesis on “Predicting Gas Valve Position”

need to compute h(x) in real-time (while engine is running!)

Explaining a Prediction.



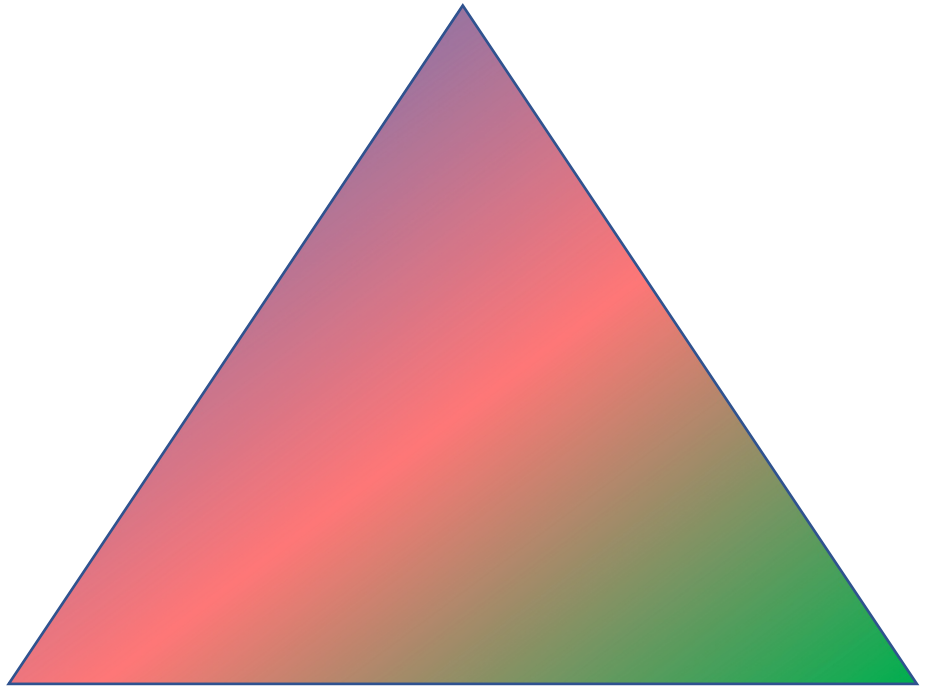
Prune Model to Ensure Explainability

$$h^{(\eta)} := \arg\min_{h \in \mathcal{H}} \widehat{L}(h|\mathcal{D}) \text{ such that } \widehat{H}(h|u) \leq \eta.$$

[1]Zhang, L., Karakasidis, G., Odnoblyudova, A., Dogruel, L., and Jung, A., “Explainable Empirical Risk Minimization”, arXiv e-prints, 2020. doi:10.48550/arXiv.2009.01492.

Design Choice: Loss

interpretability



computational complexity

statistical accuracy

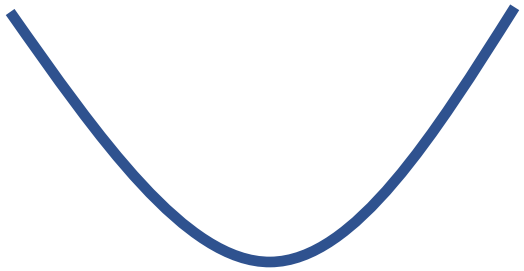
Which Loss Function ?

- statistical aspects (should favour “reasonable” hypothesis)
- computational aspects (must be able to minimize them)
- interpretation (what does log-loss = -3 mean ?)

…..choosing a suitable loss function is often non-trivial !

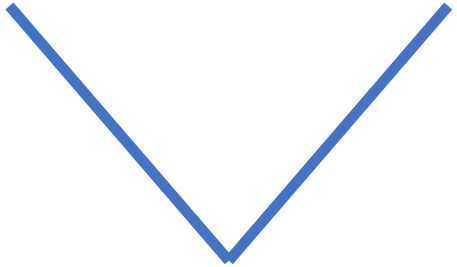
Squared Error

- cvx and diff.able
- minimized via simple gradient descent
- sensitive to outliers

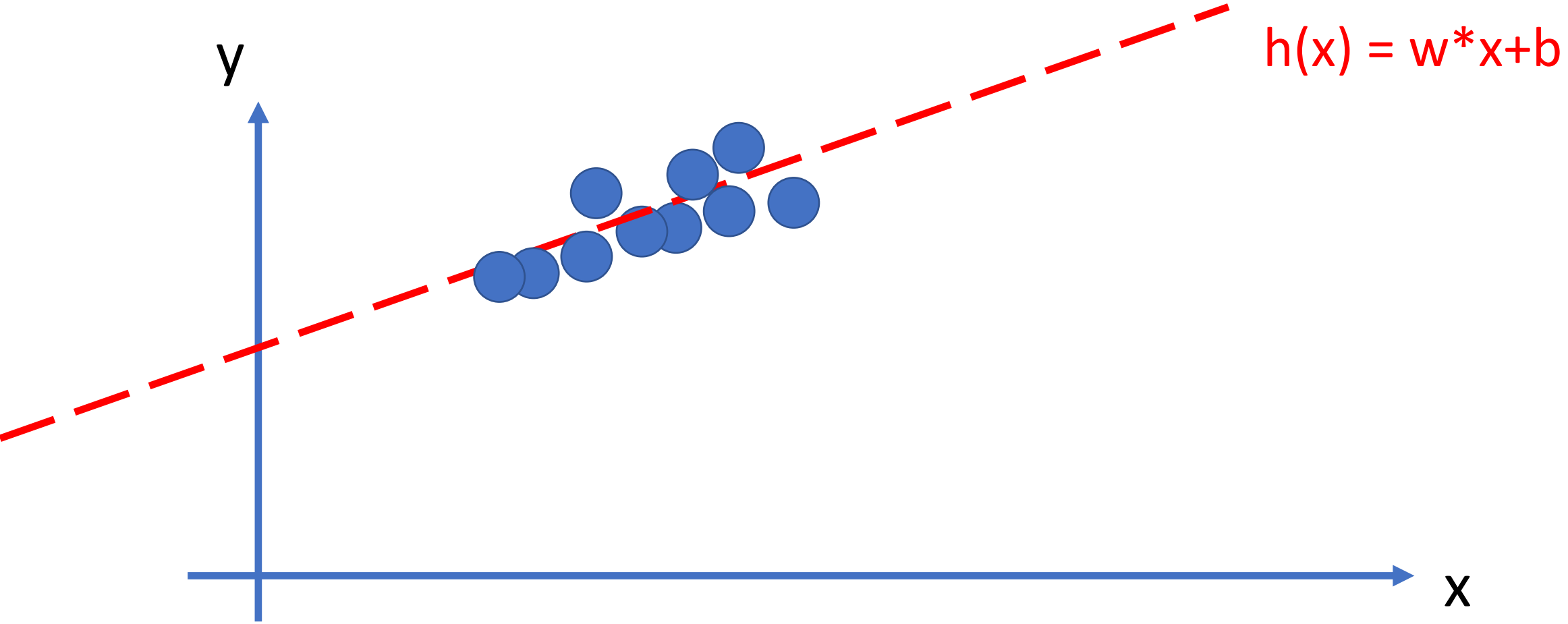


Absolute Error

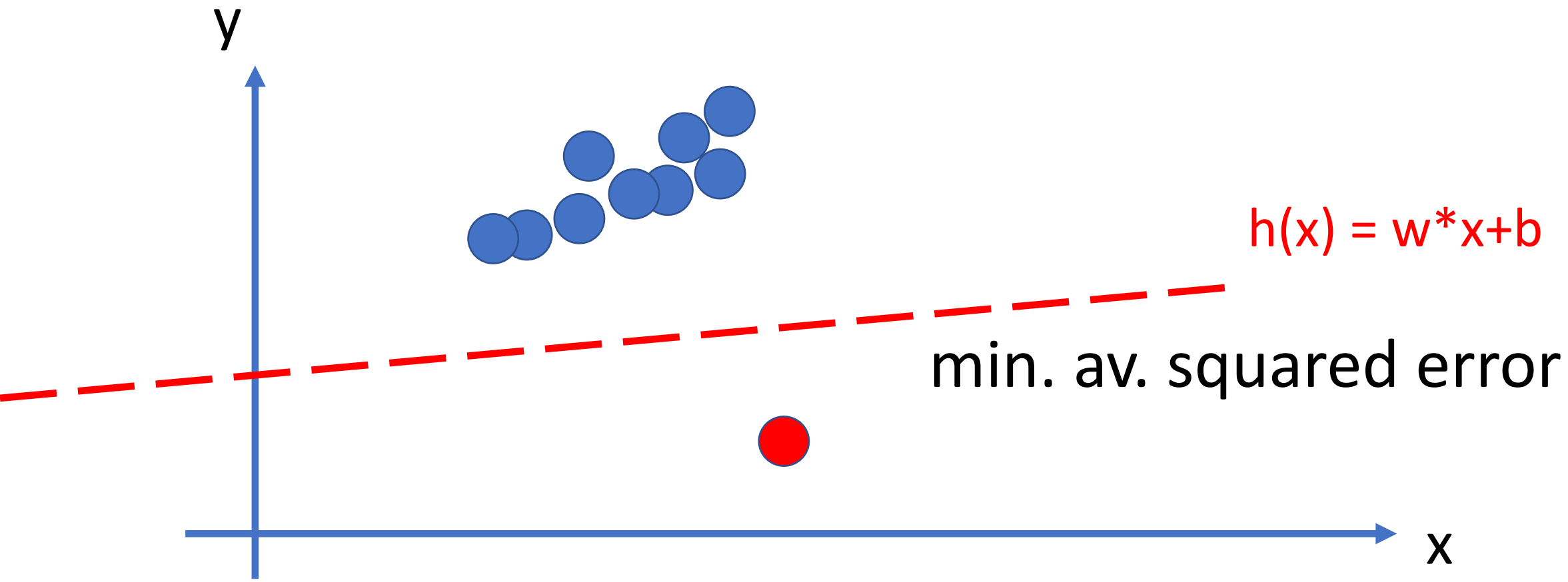
- cvx but non-diff.
- requires more advanced opt. methods
- robust against outliers



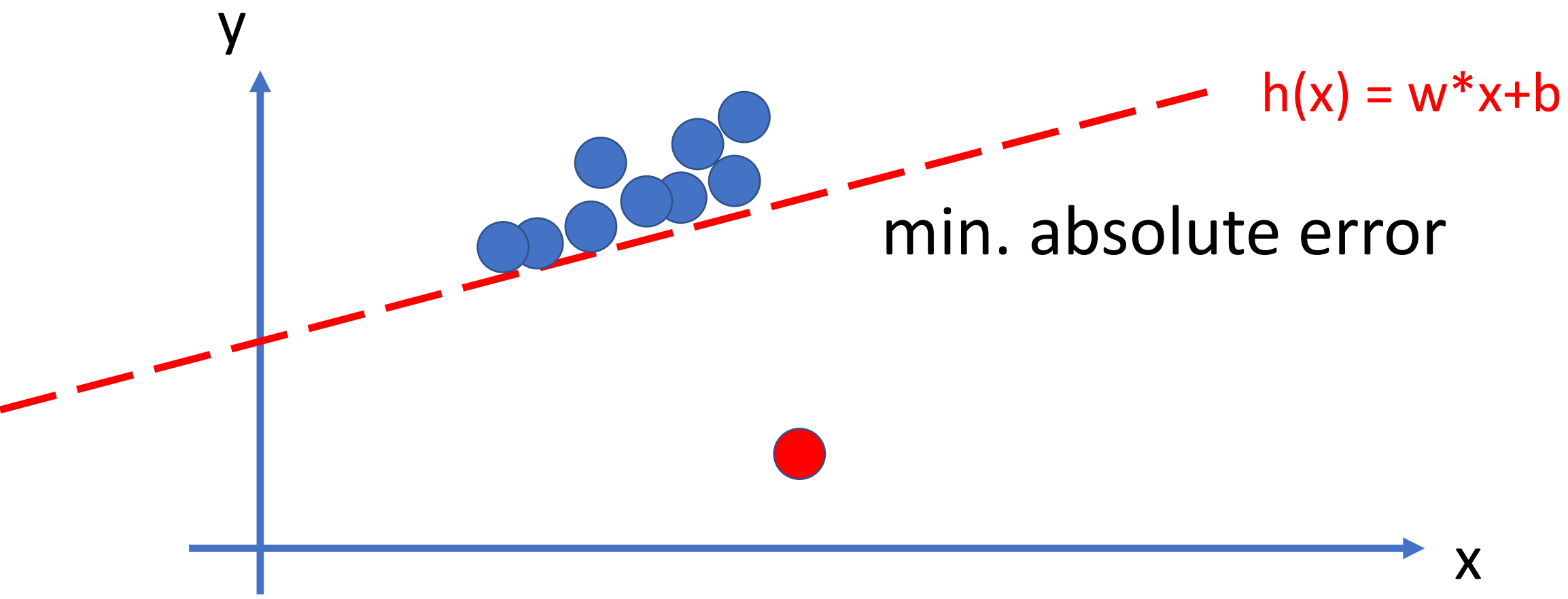
Train Linear Model on “Clean Data”



SINGLE OUTLIER



Training Set with a SINGLE OUTLIER !



Different Loss for Train and Val

- use different loss for training and validation
- allows to compare different ML methods
- logistic regression uses logistic loss to learn hypothesis h1(x)
- SVM uses hinge loss to learn hypothesis h2(x)
- compare h1, h2 by average 0/1 loss (“1-acc”) on val. set

Could Loss Reveal Private Information?

mlbook.cs.aalto.fi

How to behave in case of Covid19 symptoms?

If you experience symptoms that are consistent with COVID-19, it is important to take the following steps to protect yourself and others:

1. Isolate yourself: Stay at home and avoid contact with others to prevent the potential spread of the virus. If possible, designate a specific room and bathroom solely for your

Explainability via Loss Penalty

$$h^{(\lambda)} := \arg\min_{h \in \mathcal{H}} \widehat{L}(h|\mathcal{D}) + \lambda \underbrace{\widehat{H}(h|u)}_{=\mathcal{R}(h)}$$