

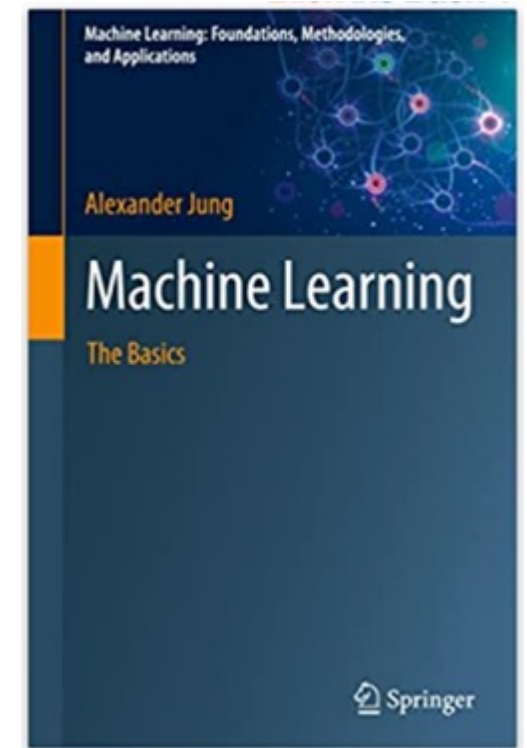
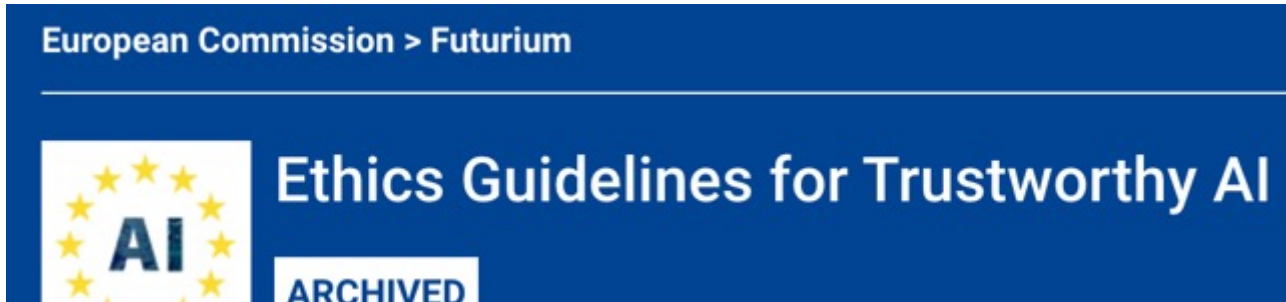
# Trustworthy ML

Alex(ander) Jung  
Assistant Professor for Machine Learning  
Department of Computer Science  
Aalto University



# Reading.

Ch. 10 of <https://mlbook.cs.aalto.fi>



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



<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32019R1150>

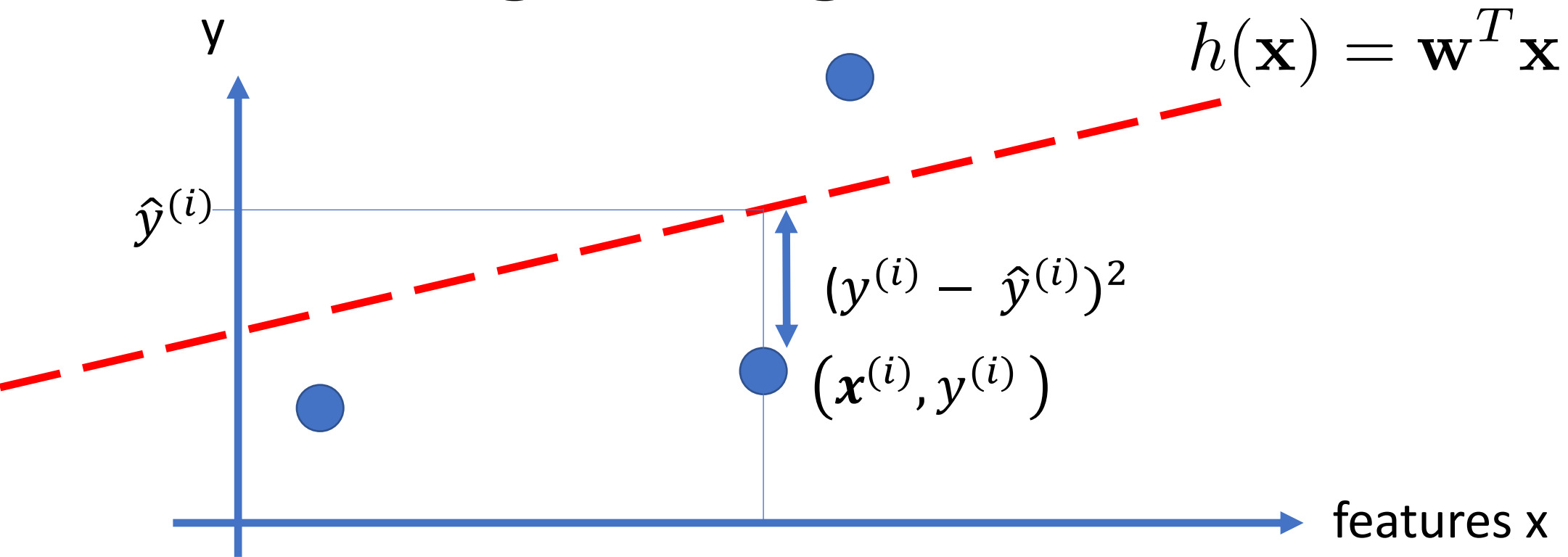
# Learning Goals

- European key requirements for trustworthy ML
- approaches to satisfy them

# What is it all About ?

fit **model** to **data** to make **accurate**  
**predictions or forecasts !**

# Learn Hypothesis (Params) by Minimizing Average Loss

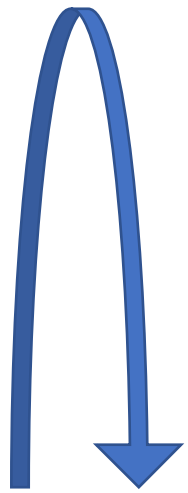


# Empirical Risk Minimization

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

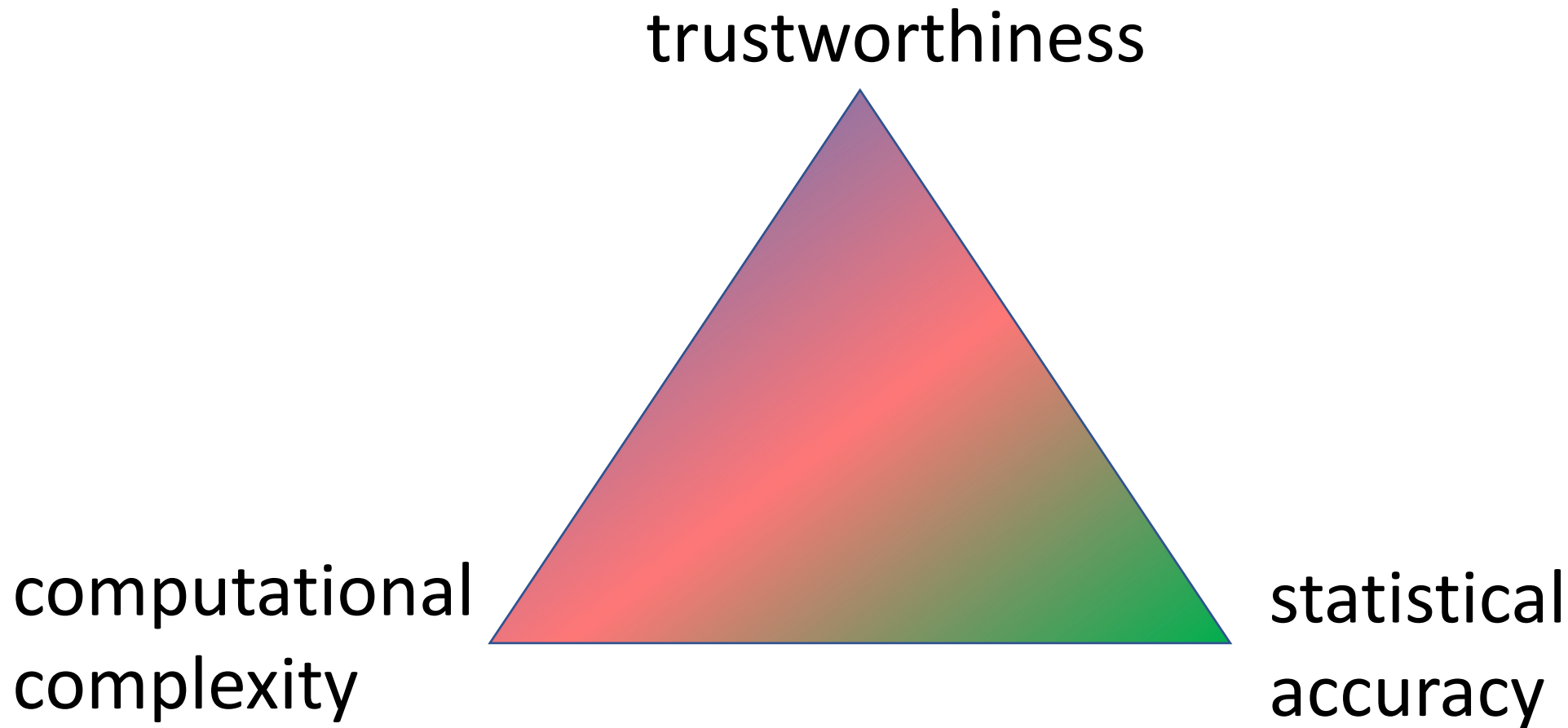
$$\stackrel{(2.16)}{=} \operatorname{argmin}_{h \in \mathcal{H}} (1/m) \sum_{i=1}^m L((\mathbf{x}^{(i)}, y^{(i)}), h).$$

# Life-Cycle of ML



- learn hypothesis  $h(x)$  via ERM (“train”)
- apply  $h(x)$  to new data (“validate”)
- measure error
- adapt ERM design choices and repeat

# Aspects of ML Design Choices





- **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 ...

# Human Oversight

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

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

design choice !

design choice !

# Human-in-the-Loop (HITL)

“...HITL refers to the capability for **human intervention** in **every decision cycle** of the system, which in many cases is neither possible nor desirable. ...”

# Human-on-the-Loop (HOTL)

“...HOTL refers to the capability for **human intervention during the design cycle** of the system and monitoring the system’s operation...”

# Human-in-Command (HIC)

“...HIC refers to the capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation...”



- 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

Research has revealed that the output of Deep Neural Networks can be easily altered by adding relatively small perturbations to the input vector. In this paper, we analyze a limited scenario where only one pixel is modified. We propose a novel method for generating adversarial perturbations based on differential evolution. The results show that 67.97% of the images in the CIFAR-10 test dataset and 16.04% of the ImageNet (2012) test images can be perturbed successfully by modifying just one pixel with a probability of 10% on average. We also show the results on the original CIFAR-10 dataset. Thus, this is a different take on adversarial machine learning, showing that current

AllConv



SHIP

CAR(99.7%)



HORSE

DOG(70.7%)

NiN



HORSE

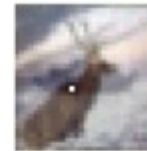
FROG(99.9%)



DOG

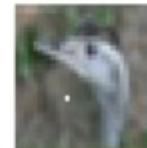
CAT(75.5%)

VGG



DEER

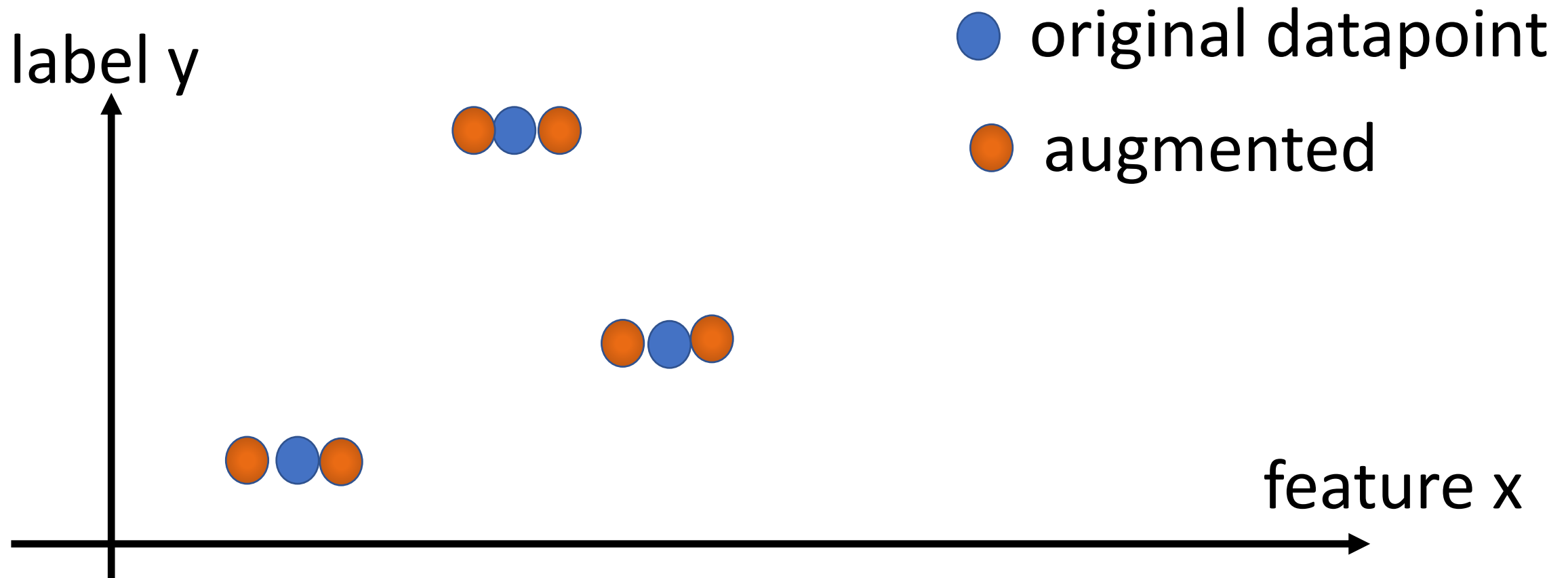
AIRPLANE(85.3%)



BIRD

FROG(86.5%)

# 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 !

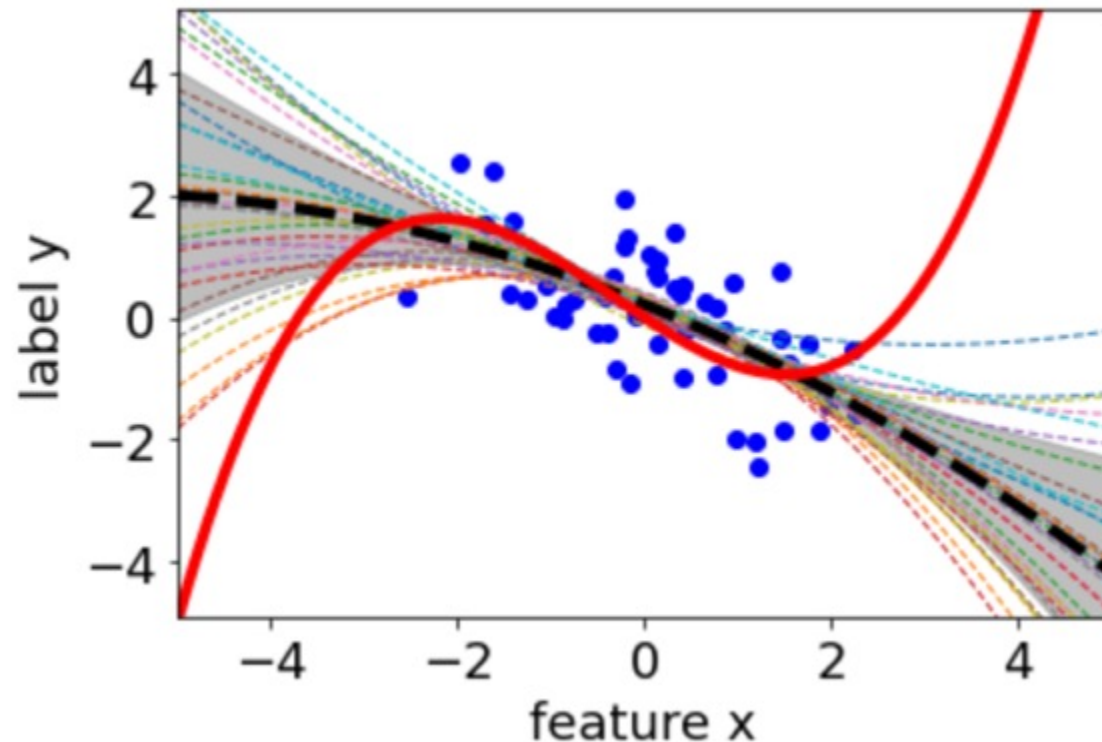
# Accuracy

*“...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

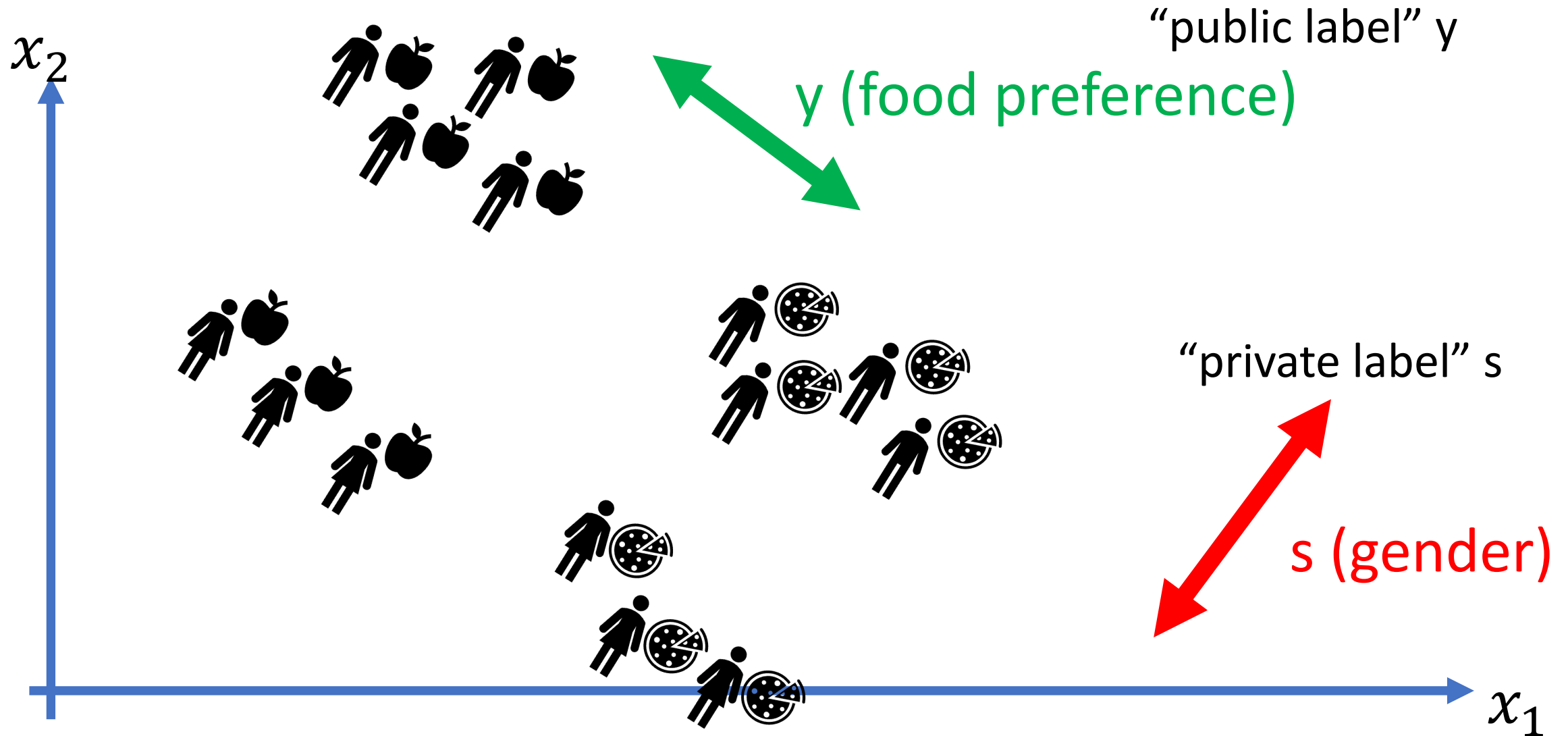


<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

# 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, religious or political views. To allow individuals to trust the data gathering process...”*

# 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

# Access to data.

- data protocols governing data access should be put in place.
- outline who can access data and under which circumstances.
- 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	
	Direct member	1 month ago by <a href="#">Jung Alex</a>	Developer ▾	Expiration date 	5 Mar, 2020	17 Aug, 2022	<a href="#">Remove member</a>
	Direct member	1 month ago by <a href="#">Jung Alex</a>	Guest ▾	Expiration date 	9 Jul, 2022	9 Jul, 2022	<a href="#">Remove member</a>
 <b>Jung Alex</b> It's you <a href="#">@junga1</a>	Direct member	5 months ago by <a href="#">Jung Alex</a>	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..”*

- **Stage 3 (22.08. - 27.08.2022):** Complete the project report which is structured as indicated [here](#).

# 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

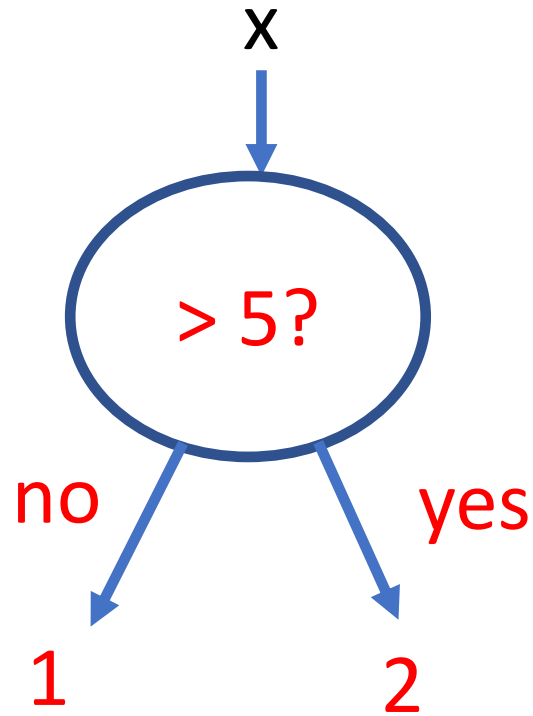
e.g., “linear regression learns a linear hypothesis by minimizing the average squared error on the training set”

# 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  $h(x) = w_1 * x_1 + w_2 * x_2$  with weights  $w_1 = 10$  and  $w_2 = 4$ ”

# Explaining a Prediction.



# 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....”*



**Hello! Slackbot here.**

I'm a simple bot, who can do one or two things (mostly nudges & looking for help, [check out our Help Center](#)).

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

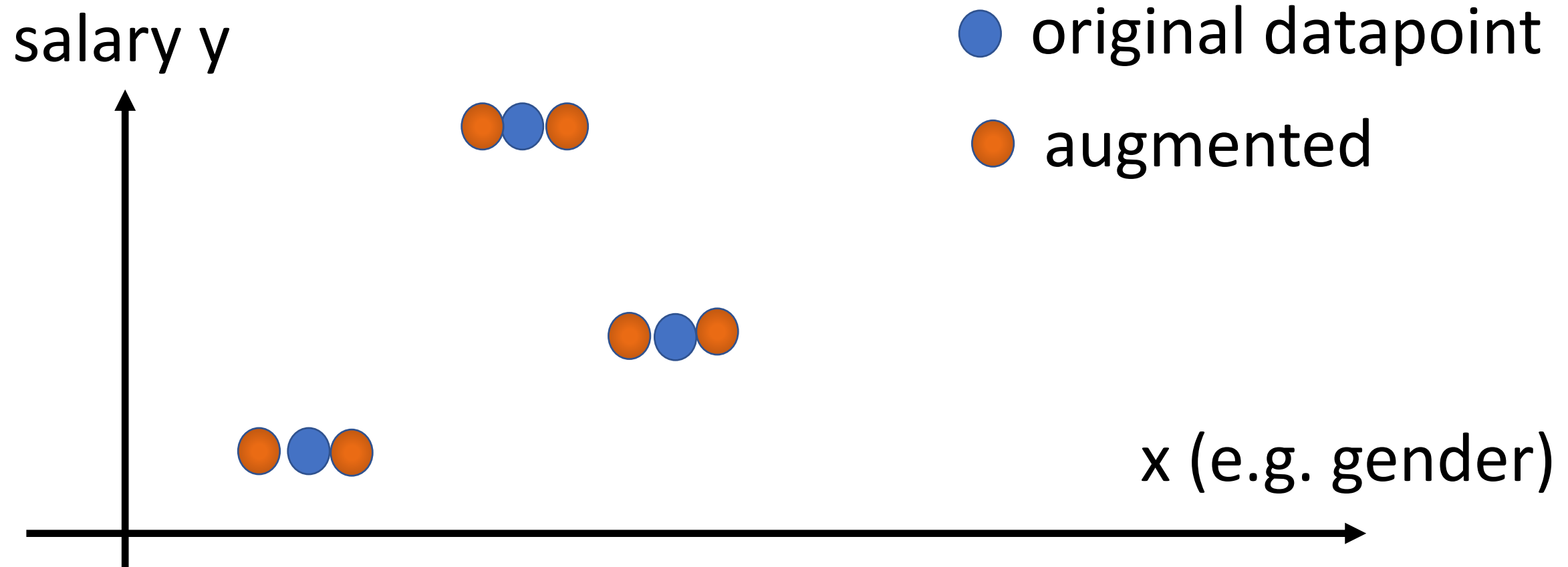


<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

# Avoidance of unfair bias.

*“Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models.”*

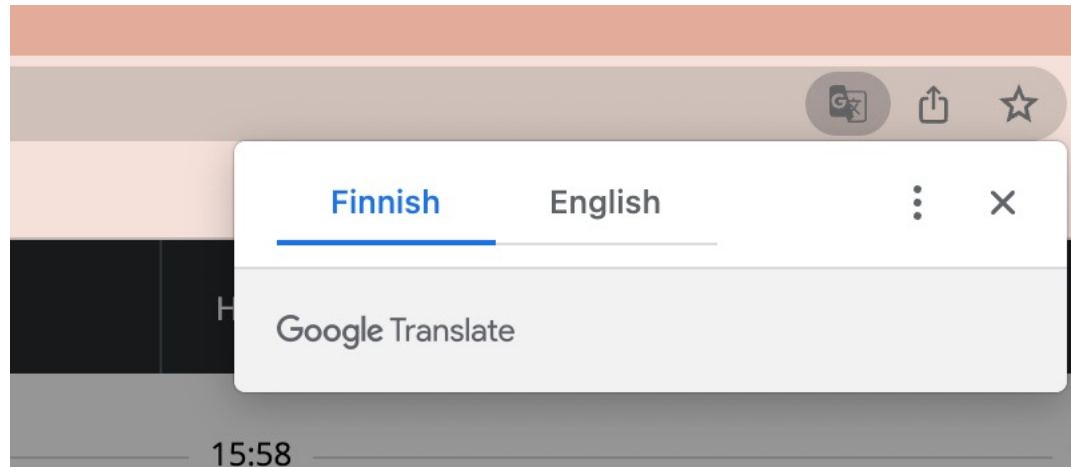
# Fairness by Data Augmentation





# 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 beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation...”*



<https://images.app.goo.gl/PjovTNXf6ouv2Kxe9>

- 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.”*

read over [https://en.wikipedia.org/wiki/Cambridge Analytica](https://en.wikipedia.org/wiki/Cambridge_Analytica) !

By the way ...

What are three main components of machine learning ?

# Machine Learning Principle

fit **model** to **data** to make **accurate**  
**predictions or forecasts !**