

CS-E4740 Federated Learning

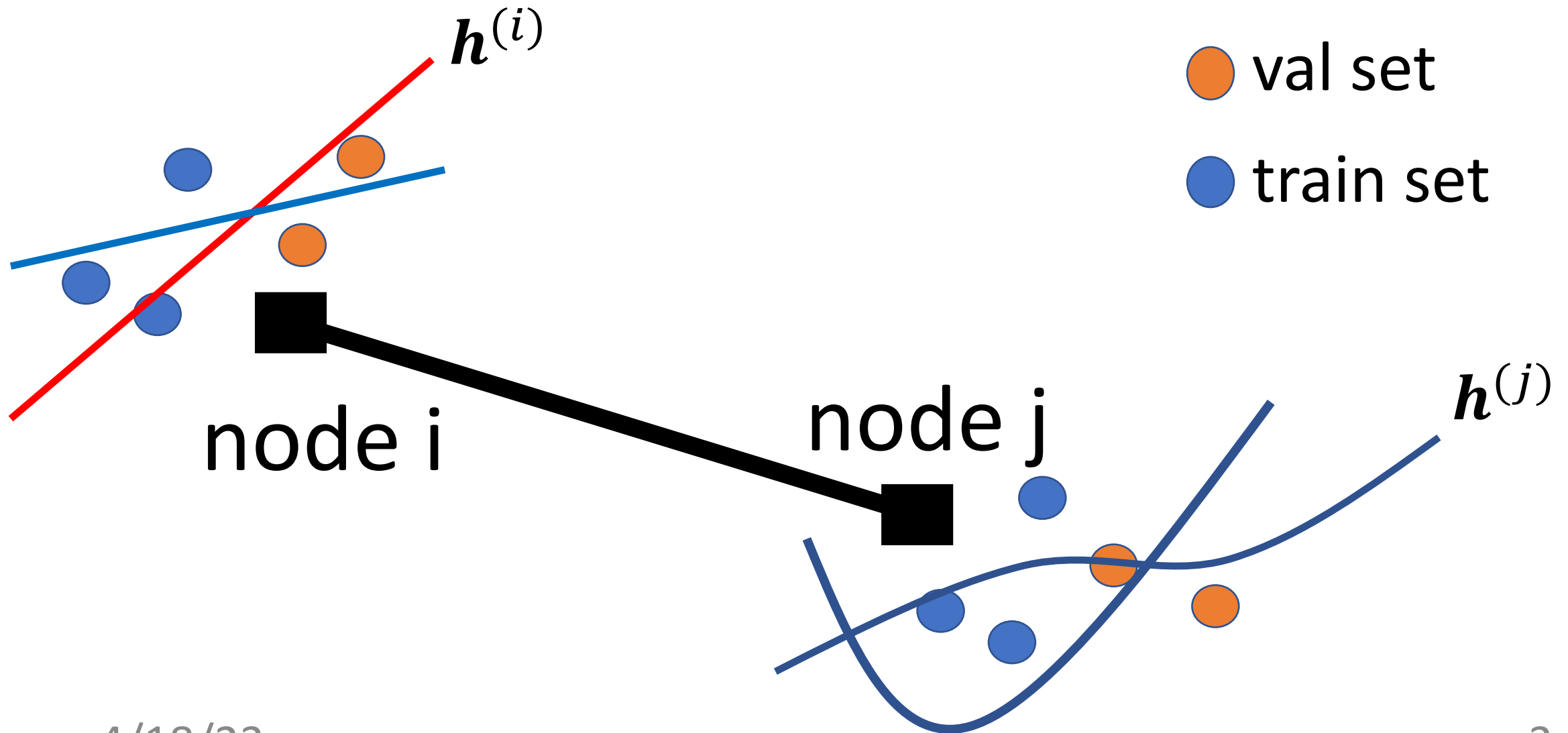
"Trustworthy FL"

Dipl.-Ing. Dr.techn. Alexander Jung

Learning Goals

- key requirements for trustworthy FL
- approaches to satisfy them

Networked Data+Model

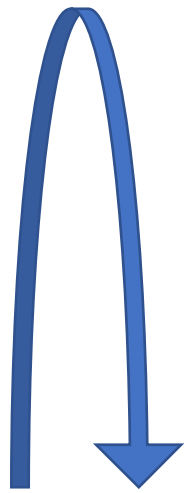


FL Design Principle

$$\min_{\mathbf{h}^{(i)}} \sum_i L^{(i)}(\mathbf{h}^{(i)}) + \lambda \sum_{\{i,j\} \in \mathcal{E}} A_{i,j} d(\mathbf{h}^{(i)}, \mathbf{h}^{(j)})$$

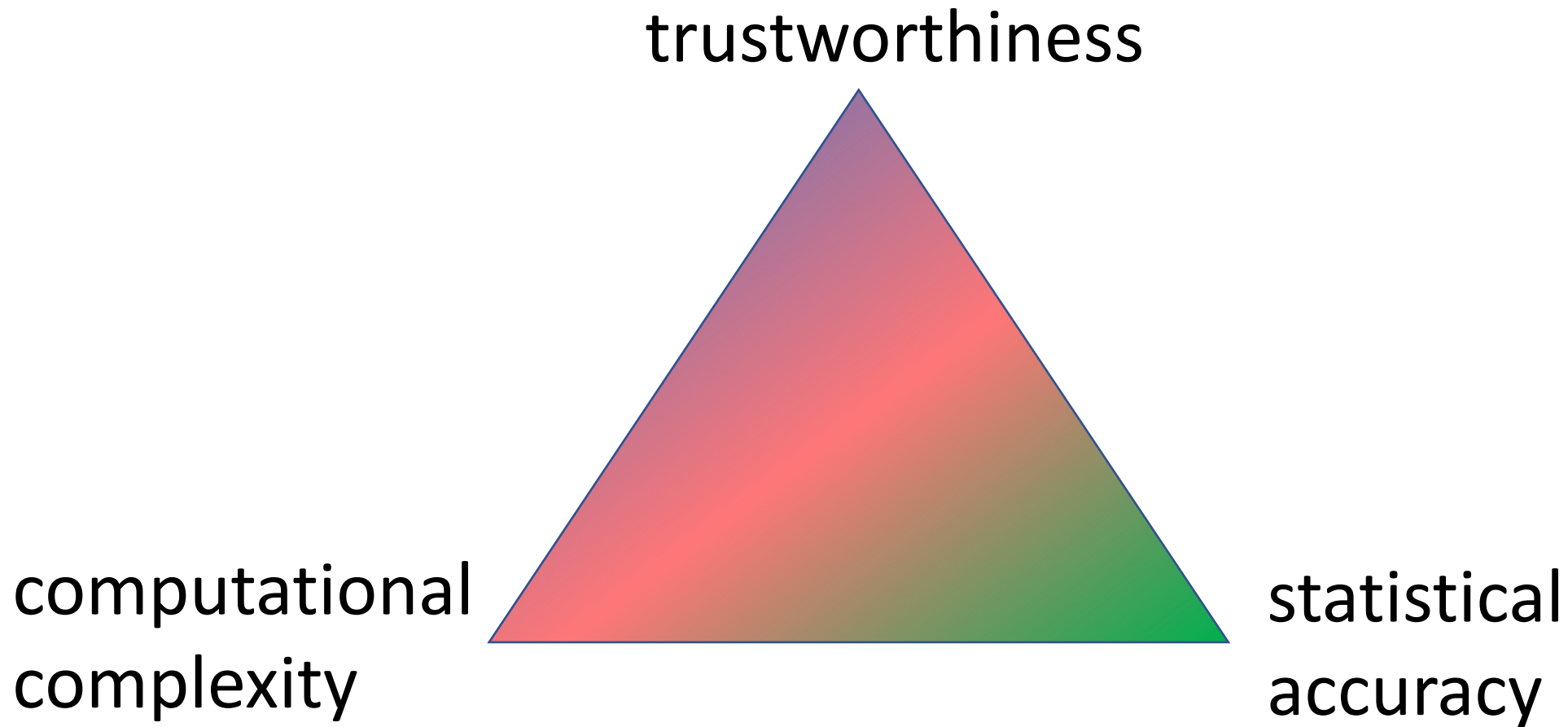
- what edges $\{i, j\} \in \mathcal{E}$ and weights $A_{i,j}$?

Design-Cycle



- learn local hyp. $h(x)$ via GTVMin (“train”)
- apply local hyp. $h(x)$ to new data (“validate”)
- measure error
- adapt GTVMin design choices and repeat

Aspects of FL Design Choices



Measure
if your organisation's AI is
trustworthy



ALTAI – Assessment List for Trustworthy Artificial Intelligence

<https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>

Assessment List for Trustworthy Artificial Intelligence (ALTAI)

*“The tool supports the actionability the key requirements outlined by the [Ethics Guidelines for Trustworthy Artificial Intelligence](#) (AI), presented by the [High-Level Expert Group on AI](#) (AI HLEG) presented to the European Commission, in April 2019. The Ethics Guidelines introduced the concept of Trustworthy AI, based on **seven key requirements**:”*

- **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....”

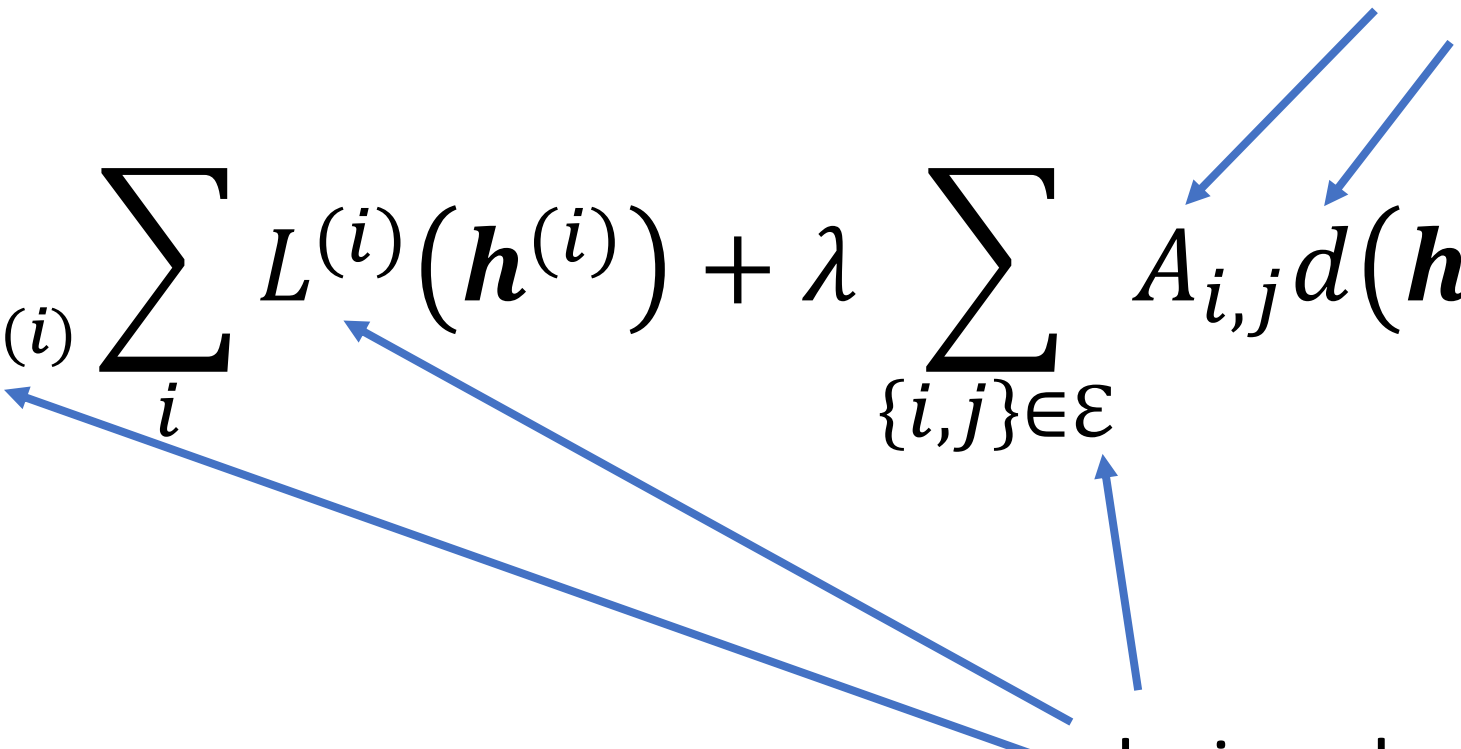
→ labels maybe not correspond to certain actions ...

Human Oversight

$$\min_{\mathbf{h}^{(i)} \in \mathcal{H}^{(i)}} \sum_i L^{(i)}(\mathbf{h}^{(i)}) + \lambda \sum_{\{i,j\} \in \mathcal{E}} A_{i,j} d(\mathbf{h}^{(i)}, \mathbf{h}^{(j)})$$

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

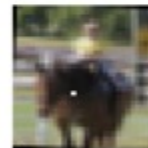
Research has revealed that the output of Deep Neural Networks can be easily altered by adding relatively small perturbations to the input vector. In this paper, we analyze a limited scenario where only one pixel is modified. We propose a novel method for generating adversarial perturbations based on differential privacy. The results show that 67.97% of the images in the CIFAR-10 test dataset and 16.04% of the images in 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

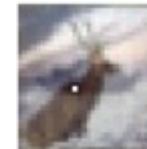
FROG(99.9%)



DOG

CAT(75.5%)

VGG



DEER

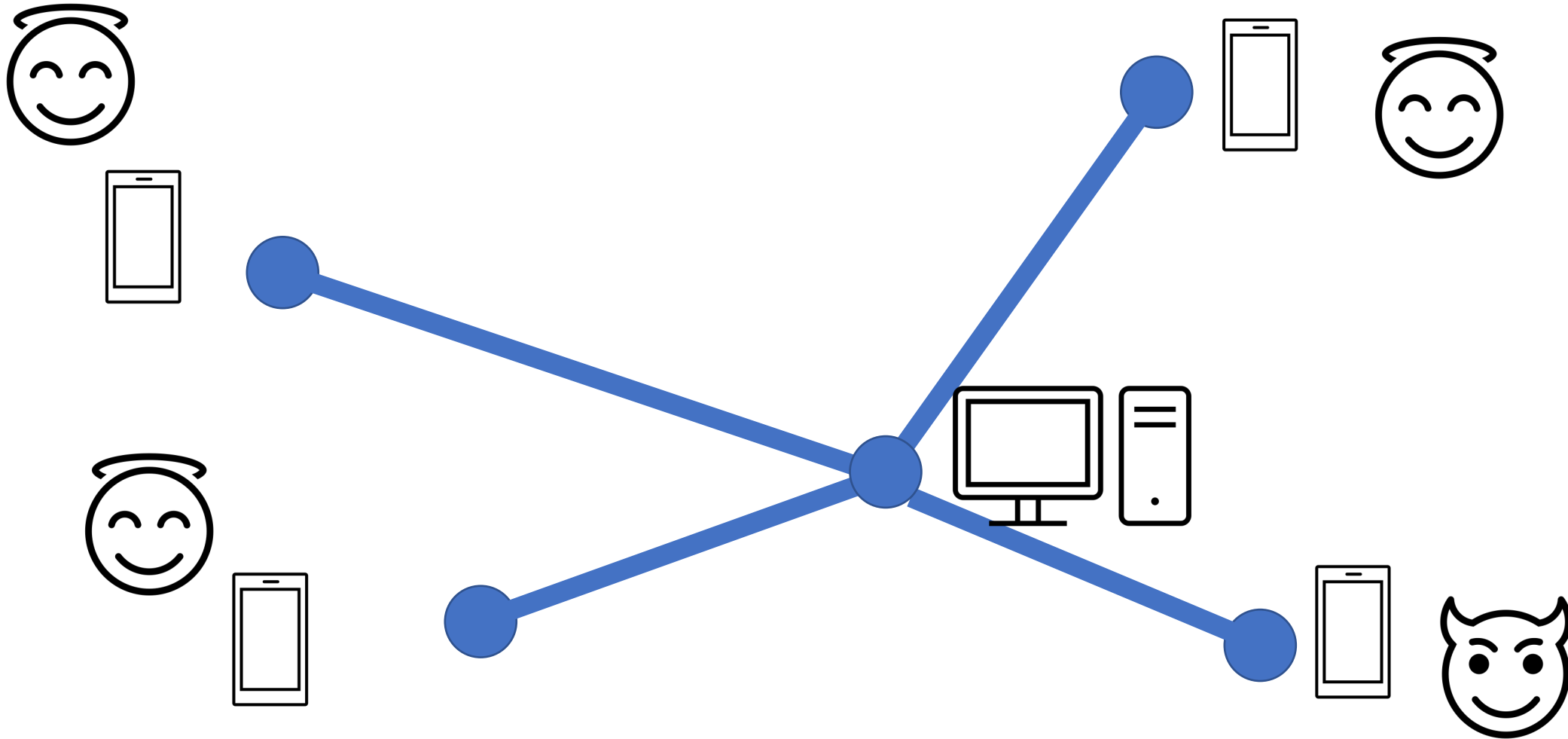
AIRPLANE(85.3%)



BIRD

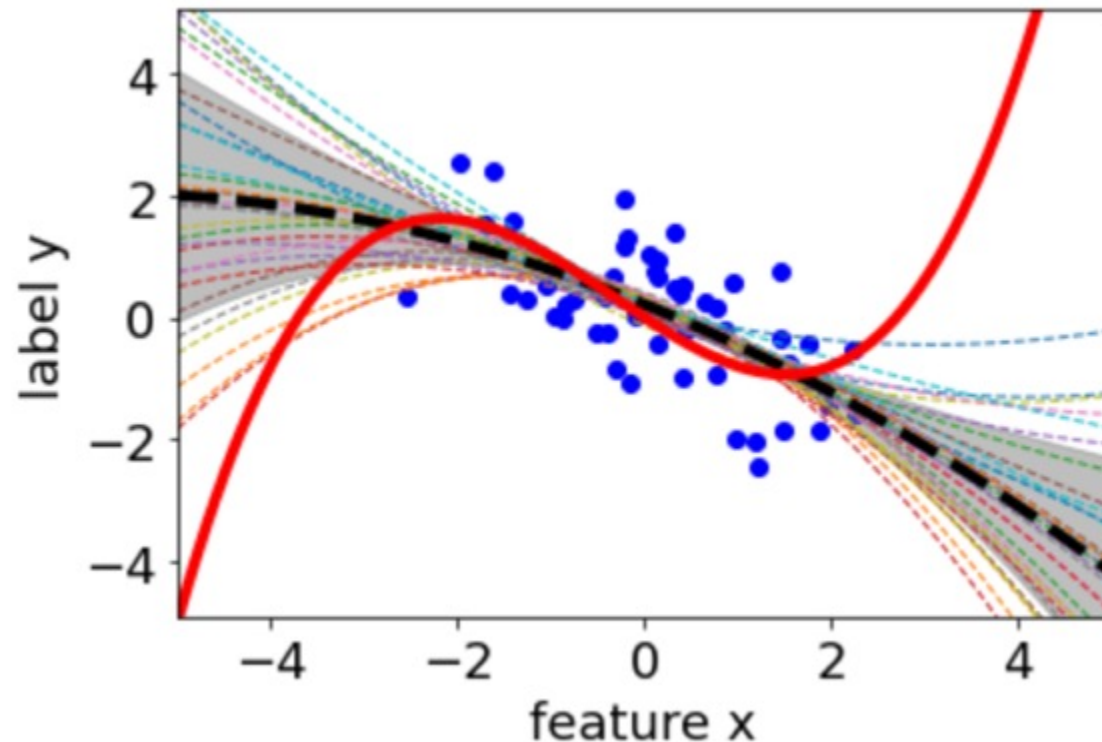
FROG(86.5%)

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



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....”



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....”

- provide **confidence measures** for predictions
- insufficient confidence -> switch to simpler models
- 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....”

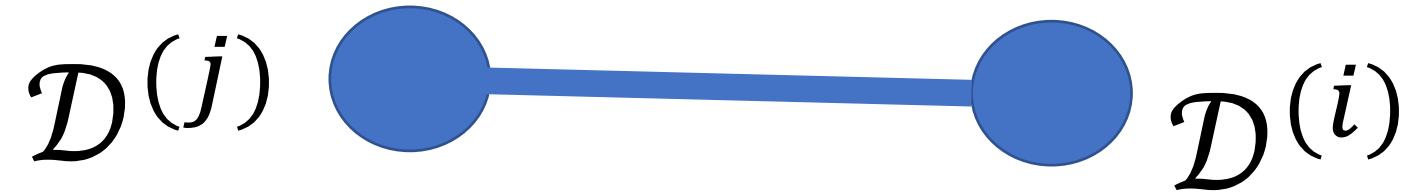
```
>>> from sklearn.datasets import load_iris
>>> from sklearn.linear_model import LogisticRegression
>>> X, y = load_iris(return_X_y=True)
>>> clf = LogisticRegression(random_state=0).fit(X, y)
>>> clf.predict(X[:2, :])
array([0, 0])
>>> clf.predict_proba(X[:2, :])
array([[9.8...e-01, 1.8...e-02, 1.4...e-08],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

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



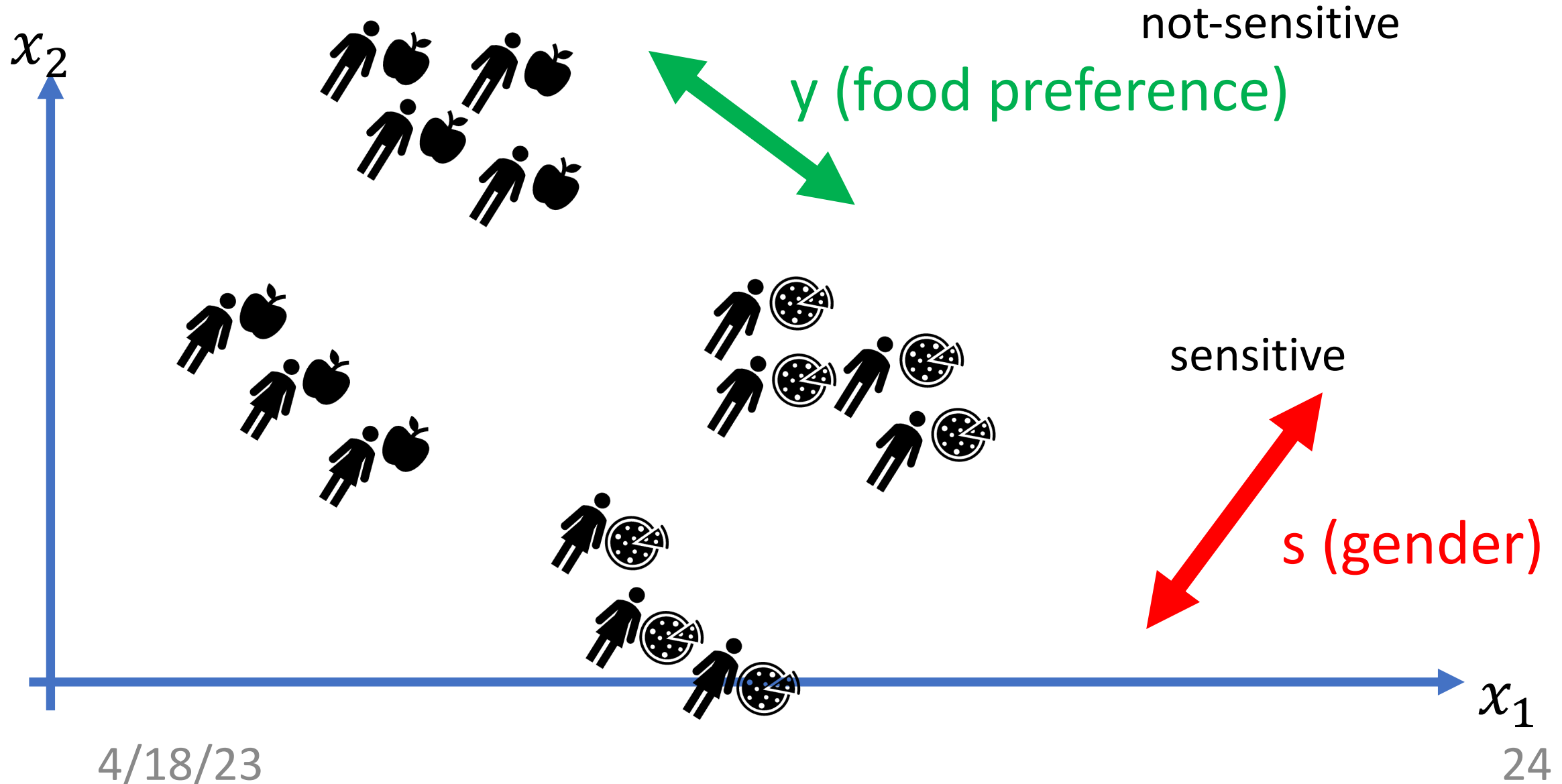
<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>

Privacy-Friendly FL



FedSGD or FedRelax (see Sec. 9 of lec.notes)
do not exchange raw data but only parameter
updates or predictions on test set

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

Check Datasheet !







KEY SPECIFICATIONS

CATEGORY	FEATURE	SPECIFICATION
Chipset	LoRa®	Semtech SX1272
	Bluetooth®	Nordic nRF51822 – 256 k/32 k
LoRa	Frequencies	863 – 870 MHz (EU), 902 – 928 MHz (US), 915 – 928 MHz (AU + AS923)
Temperature Probe	Type	Class A tolerance PT100-M222 RTD
	Interface	Resistance Temperature Detector (RTD) using 3-wire interface
		-40°C (-40°F) to +180°C (350°F)
	Operating Range and Variance	Variance of reported temperature data can be calculated taking the following uncertainties into account: i) BS EN 60751:2008/IEC 60751 standards which state accuracy of class A PT100 measurement: $(0.15^{\circ}\text{C} + 0.002 t)$ ii) RTD-to-Digital conversion utilizing the MAX31865 with a 'Total Accuracy Over All Operating Range' of $\pm 0.5^{\circ}\text{C}$ As such, max. variance at $-40^{\circ}\text{C} = \pm 0.57^{\circ}\text{C}$ or max. variance at $\pm 180^{\circ}\text{C} = \pm 1.01^{\circ}\text{C}$
	Dimensions and Connector	Cable length – 1320 mm (± 20 mm); stainless-steel shaft – 4.0 mm (± 0.2 mm) (dia.) x 100 mm (± 2 mm) length RJ45C connector (IP66~68 rated). user connected

<https://www.digikey.at/en>

Access to data.

- *“...data protocols governing data access..”*
- *“..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....”*

Account	Source	Access granted	Max role	Expiration	Created on	Last activity	
	Direct member	1 month ago by Jung Alex	Developer ▾	Expiration date 	5 Mar, 2020	17 Aug, 2022	Remove member
	Direct member	1 month ago by Jung Alex	Guest ▾	Expiration date 	9 Jul, 2022	9 Jul, 2022	Remove member
 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..”

- **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 FL method amounts to

- specify local datasets; empirical graph
- specify local models
- specify local loss function

Explaining a FL Method.

provide information about how empirical graph and local datasets is turned into local hypothesis maps

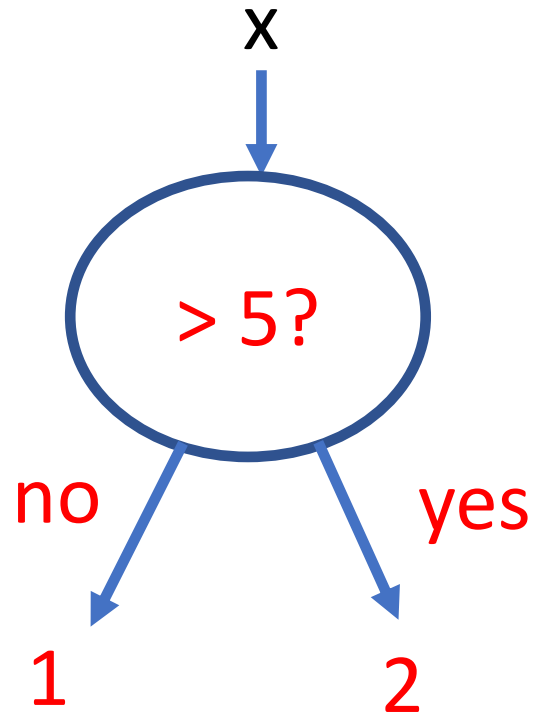
e.g., “hypothesis maps are learnt by solving GTVMin using local models ... and empirical graph constructed as ...”

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 & looks 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

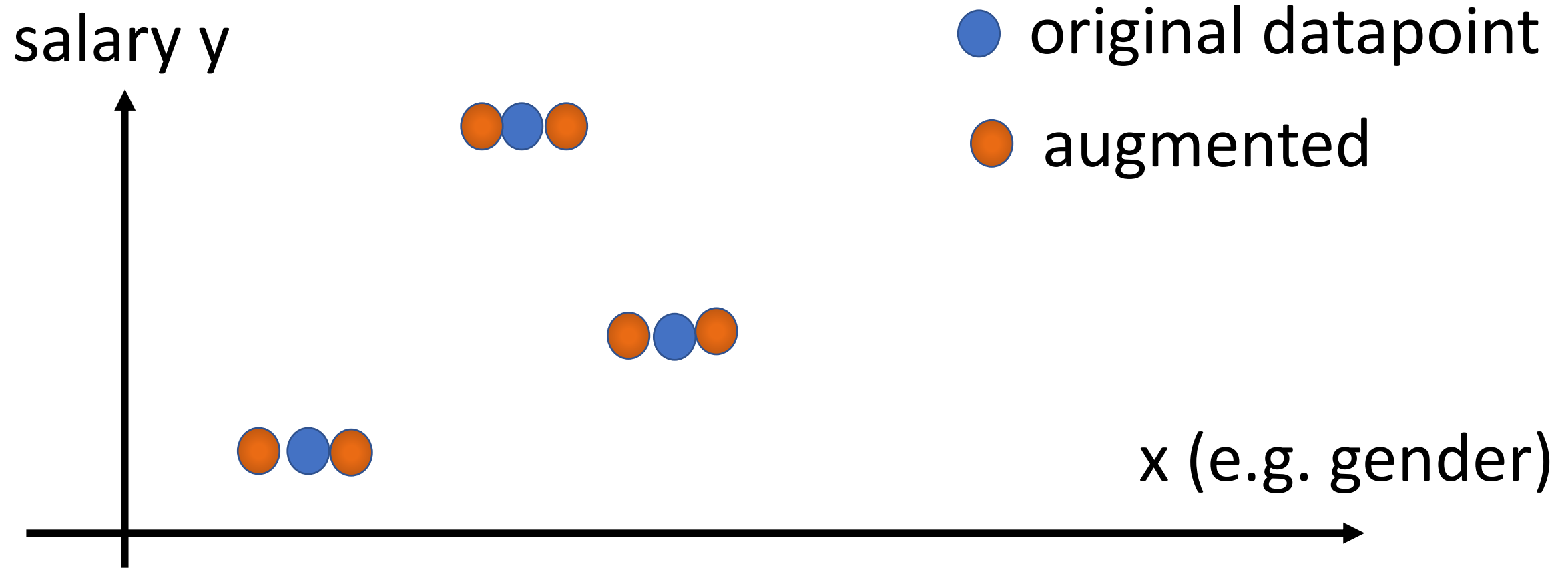
<https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html>



Avoidance of unfair bias.

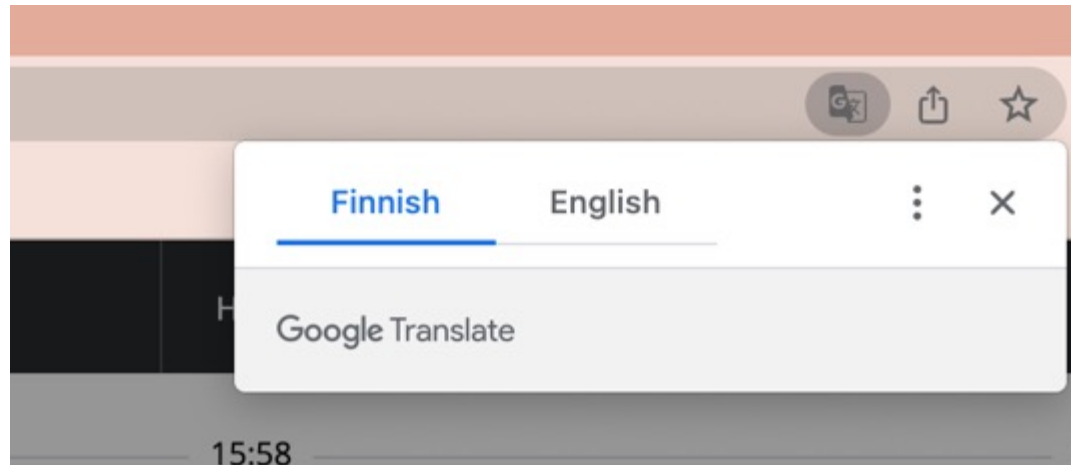
“Data sets used by AI systems (both for training and operation) may suffer from the inclusion of inadvertent historic bias, incompleteness and bad governance models.”

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 FL

“...a critical examination of the resource usage and energy consumption during training, opting for less harmful choices...”

-> favour GTVMin instances that allow for low-complexity FL algorithms

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...”

-> limit usage time of FL apps

Society and Democracy

“...The use of AI systems should be given careful consideration particularly in situations relating to the democratic process, including ...also electoral contexts...”

avoid GTVMin instances that allow to predict voter behaviour

- 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>



“...mechanisms be put in place to ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.”

Auditability.

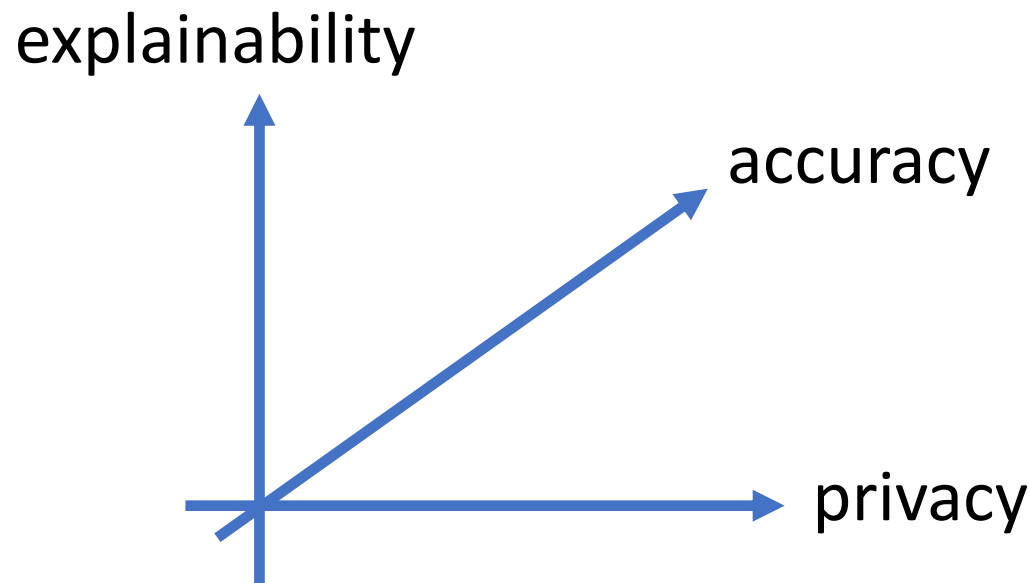
“...Evaluation by internal and external auditors, and the availability of such evaluation reports, can contribute to the trustworthiness of the technology. In applications affecting fundamental rights, including safety-critical applications, AI systems should be able to be independently audited.”

Minimisation and reporting of negative impacts

“...Due protection must be available for whistle-blowers, NGOs, trade unions or other entities when reporting legitimate concerns about an AI-based system. The use of impact assessments (e.g. red teaming or forms of Algorithmic Impact Assessment) both prior to and during the development, deployment and use of AI systems can be helpful to minimise negative impact..”

Trade-offs

“....relevant interests and values implicated by the AI system should be identified and that, if conflict arises, trade-offs should be explicitly acknowledged...”



Wrap Up

- GTVMin involves design choices
- comput., statist. and trustworth. as criteria
- seven key requirements for trustworthiness
- trade-offs between comp., stat and trustworth.

Thank you for
your attention!