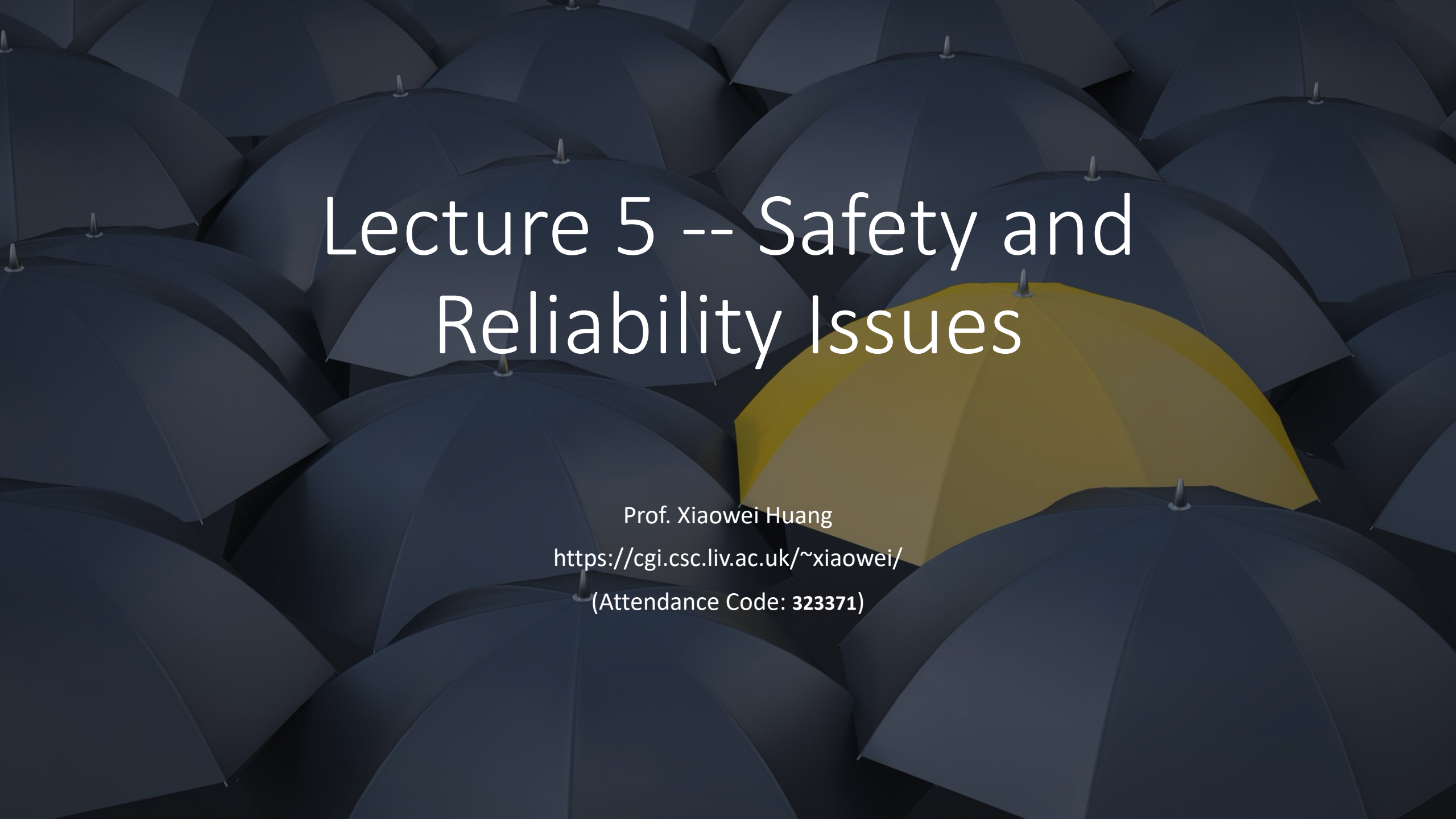
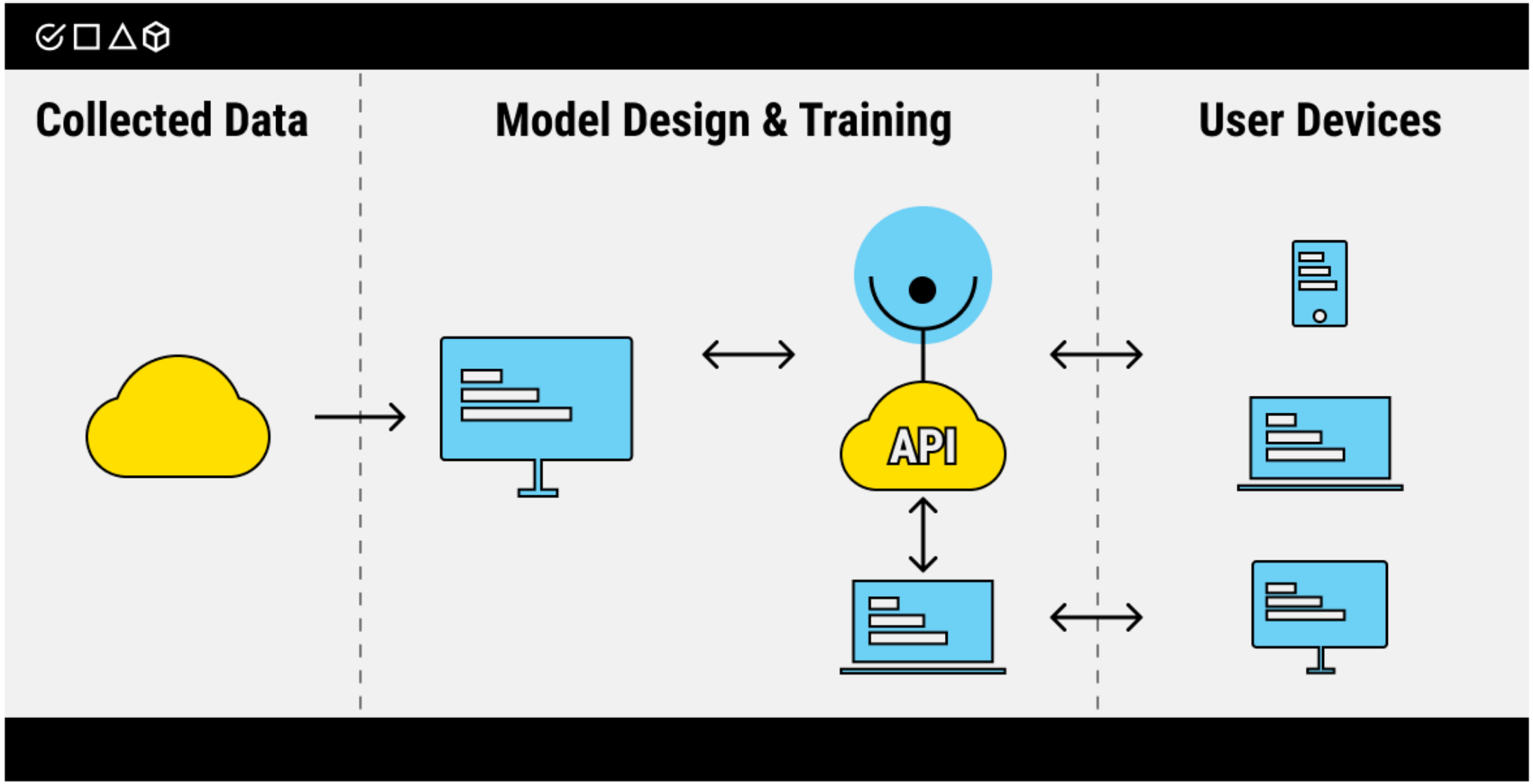


Lecture 5 -- Safety and Reliability Issues

Prof. Xiaowei Huang

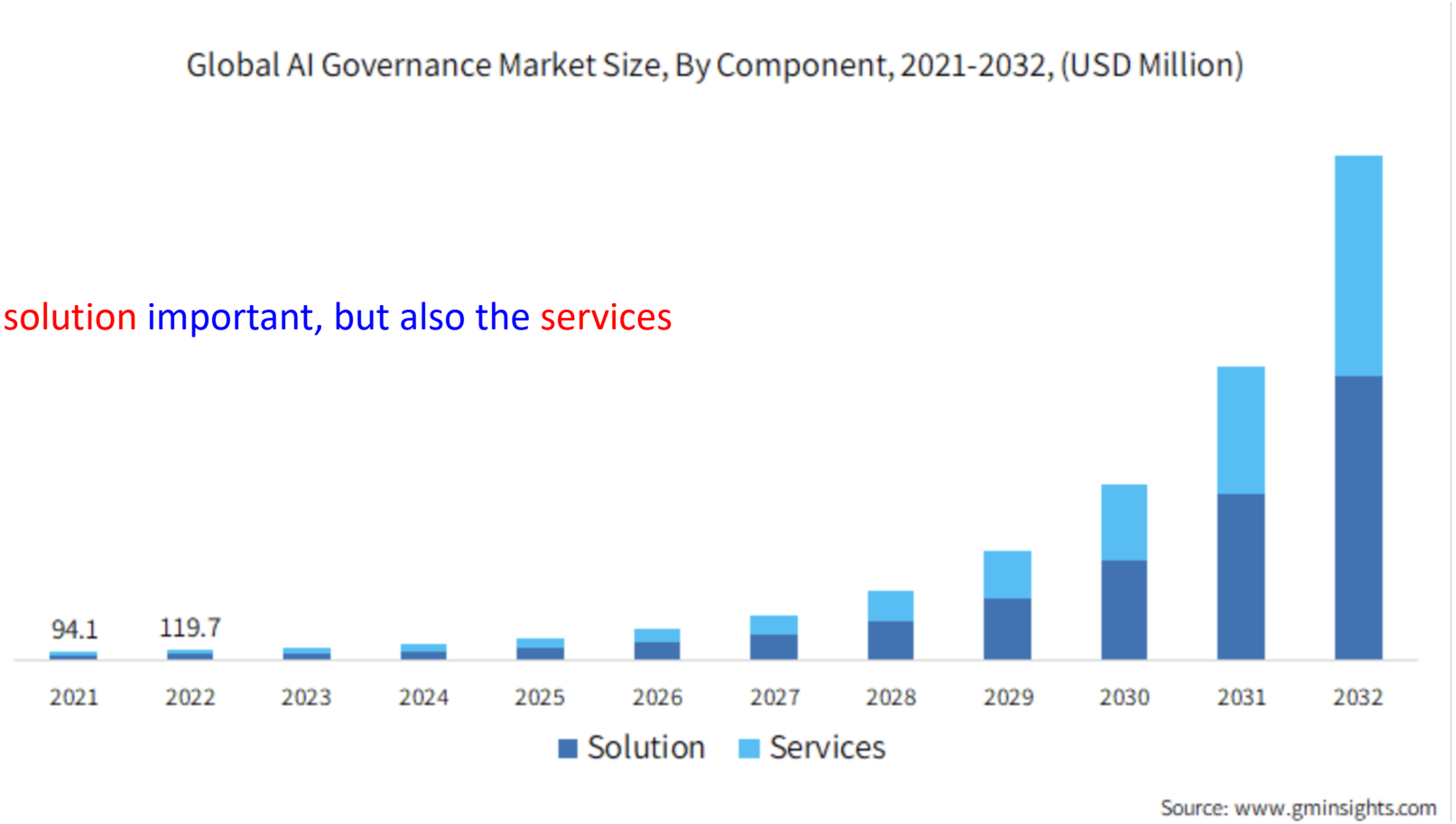
https://cgi.csc.liv.ac.uk/~xiaowei/

(Attendance Code: **323371**)


Collected Data
Model Design & Training
User Devices
API

Global Responsible AI Market Size

not only is the solution important, but also the services



Market Trending for Trustworthy AI

- In 2021, the global AI economy was valued at US$59.7 billion, and the figure is forecast to reach **US$422 billion by 2028**.
- **75 percent** of all businesses already include AI in their core strategies.
- 80 percent of firms will commit at least **10 percent** of their AI budgets to regulatory compliance by 2024, with 45 percent pledging to set aside a minimum of 20 percent
- 90 percent of limited partners would walk away from an investment opportunity if it presented an ESG (Environmental, Social, and Corporate Governance) concern, to ensure that design and deploy AI that generates value without inflicting harm.
- Globally, only 10 to 15% of companies have successfully industrialized AI-based solutions in their business and 30 to 40% are limited to experimentation.
- A trusted AI market estimated at 52 billion euros in several sectors studied (Automotive, Railway, Aeronautics, Energy & Resources, Banking, Insurance, Pharmaceuticals).
- A growing need among manufacturers for trust solutions paving the way for a market of certification or validation of AI systems

In summary, the current market size of products and service to enforce Verified and Trustworthy AI is between 5 to 10 billion euros.

https://www.accenture.com/us-en/insights/artificial-intelligence/ai-compliance-competitive-advantage
https://ilpa.org/wp-content/uploads/2022/01/infographic-esg-measurement-gap-private-equity.pdf
https://www.ey.com/fr_fr/strategy/quel-avenir-pour-l-intelligence-artificielle-dans-l-industrie

Topics -- Safety Concerns

- Generalisation error
- Adversarial examples
- Data poisoning
- Backdoor attack
- Model Stealing
- Membership Inference
- Others (fairness, uncertainty, energy efficiency, etc)

Safety Concerns

- Despite the success of machine learning in many areas, serious concerns have been raised in applying machine learning to real-world safety-critical systems such as self-driving cars, automatic medical diagnosis, etc.

- Simply speaking, a learned model f is to approximate a target function h. Therefore, the erroneous behaviour of f exists when it is inconsistent with h.

Safety Concerns

Definition 4 (Erroneous Behavior of a Classifier). *Given a (trained) classifier* $f: \mathbb{R}^n \rightarrow \mathbb{R}^k$, *a target function* $h: \mathbb{R}^n \rightarrow \mathbb{R}^k$, *an erroneous behavior of the classifier* f *is exhibited by a legitimate input* $x \in \mathbb{R}^n$ *such that*

$$\arg\max_j f_j(\boldsymbol{x}) \neq \arg\max_j h_j(\boldsymbol{x}) \tag{2.11}$$

Intuitively, an erroneous behaviour is witnessed by the existence of an input $\mathbf{x}$ on which the classifier and the target function return different result. Note that, the legitimate input $\mathbf{x}$ can be any input in the input domain $\mathbb{R}^n$ and does not have to be a training instance.

Generalisation Error

$\mathcal{L}(y, f(\mathbf{x}))$: Loss between prediction f(x) and ground truth y

- Empirical loss:

$$\mathcal{L}_{emp}(f, D) = \frac{1}{|D|} \sum_{(\mathbf{x},y)\in D} \mathcal{L}(y, f(\mathbf{x}))$$

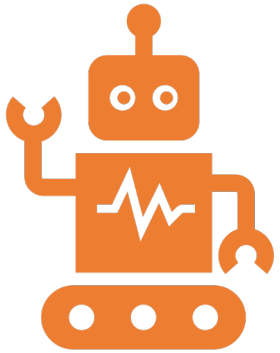
- Expected loss:

$$\mathcal{L}_{exp}(f, \mathcal{D}) = \mathbb{E}_{(\mathbf{x},y)\in \mathcal{D}} \mathcal{L}(y, f(\mathbf{x}))$$

- Generalisation error:

$$GE(f, D) = |\mathcal{L}_{emp}(f, D) - \mathcal{L}_{exp}(f, D)|$$

Generalisation Error



Generalisation error is closely related to the overfitting problem of machine learning algorithms.

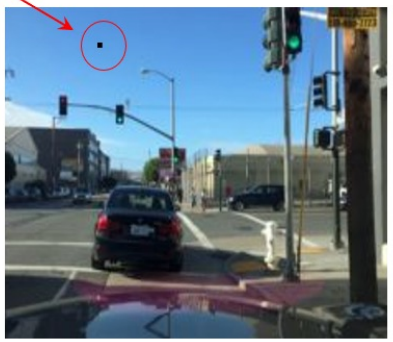
A machine learning model is overfitted if it performs well on training data samples but badly on test data samples

Adversarial Examples

- Adversarial examples represent another class of erroneous behaviours that also introduce safety implications.
- It represents a mis-match of the decisions made by a human and by a neural network, and does not necessarily involve an adversary.

DL Classification: Green Light

Changing one pixel



DL Classification: Red Light

Adversarial Examples

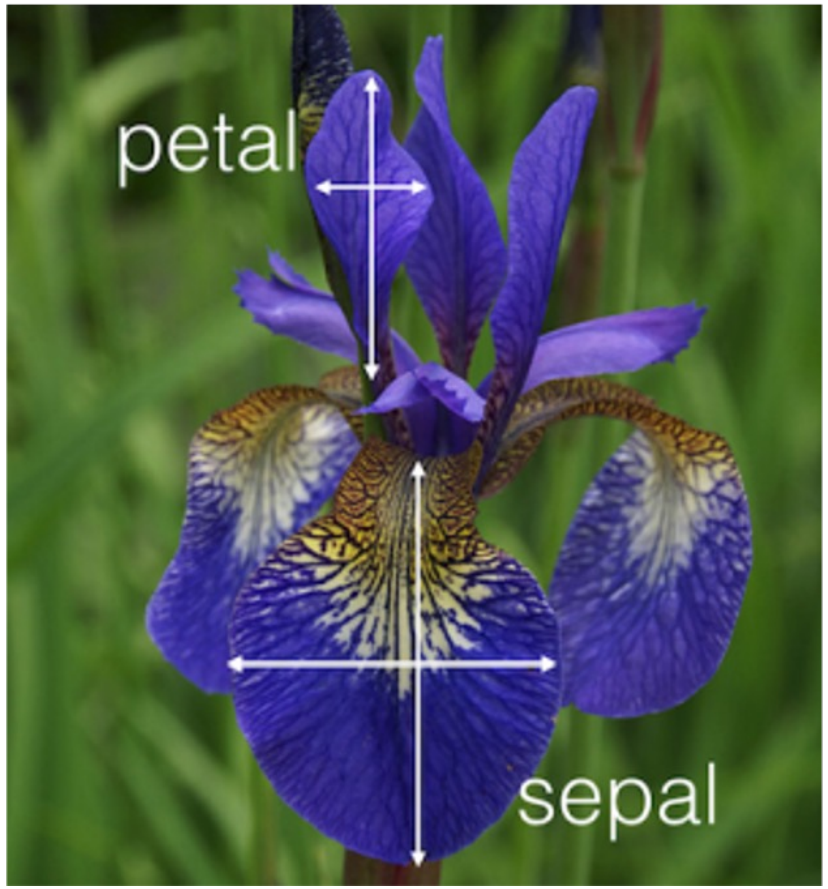
- given an instance **x**, it is classified correctly, i.e., $y = f(\mathbf{x})$
- but a small perturbation on **x** may lead to a change of classification, i.e.,

$$f(\mathbf{x} + \epsilon) \neq f(\mathbf{x})$$

DL Classification: Green Light

DL Classification: Red Light

Adversarial Examples



index	Sepal Length	Sepal Width	Petal Length	Petal Width	Class Label
1	5.1	3.5	1.4	0.2	iris setosa
2	4.9	3.0	1.4	0.2	iris setosa
...					
50	6.4	3.5	4.5	1.2	iris versicolor
...					
150	5.9	3.0	5.1	1.8	iris virginica
151	5.9	3.0	5.1	1.6	iris versicolor

Table 2.2: Iris dataset

Measurement of Adversarial Examples

- An adversarial example is usually measured from two aspects:
  - magnitude of perturbation,

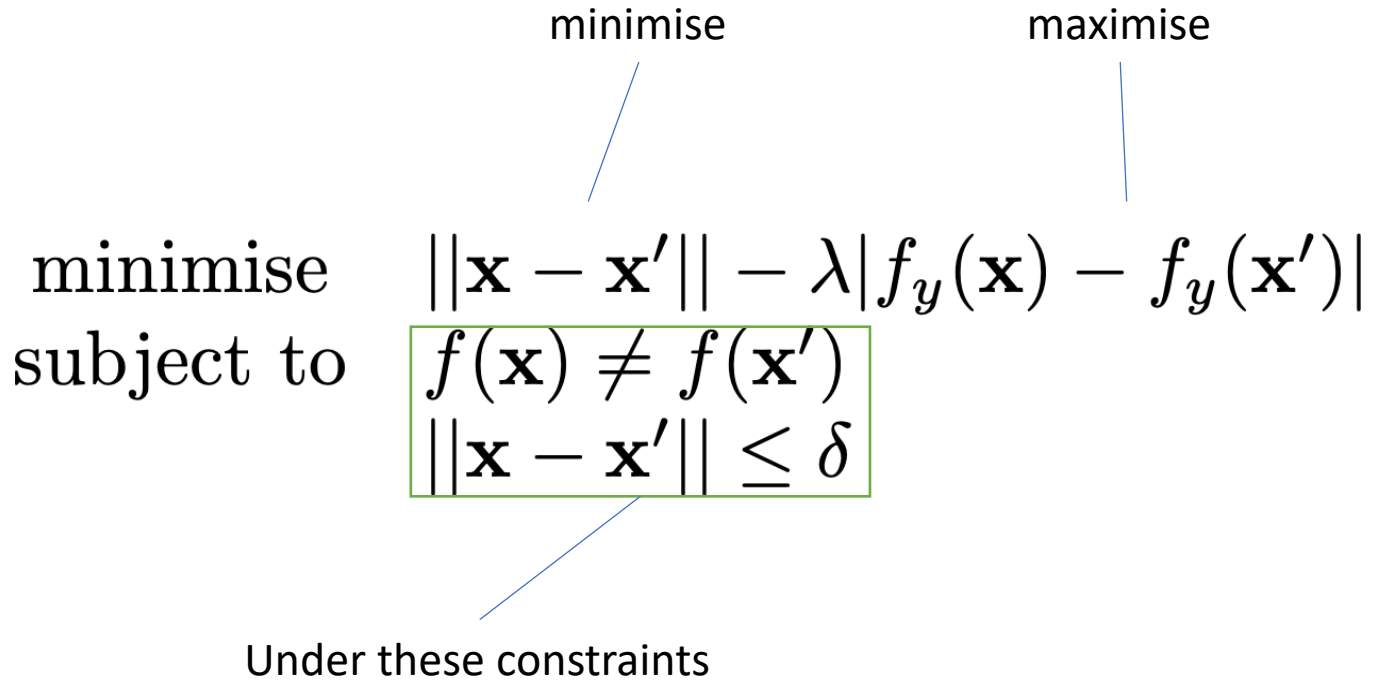
$$||\mathbf{x} - \mathbf{x}'||$$

where $||\cdot||$ is a norm distance,

  - probability gap between and after the perturbation, i.e.,

$$|f_y(\mathbf{x}) - f_y(\mathbf{x}')|$$

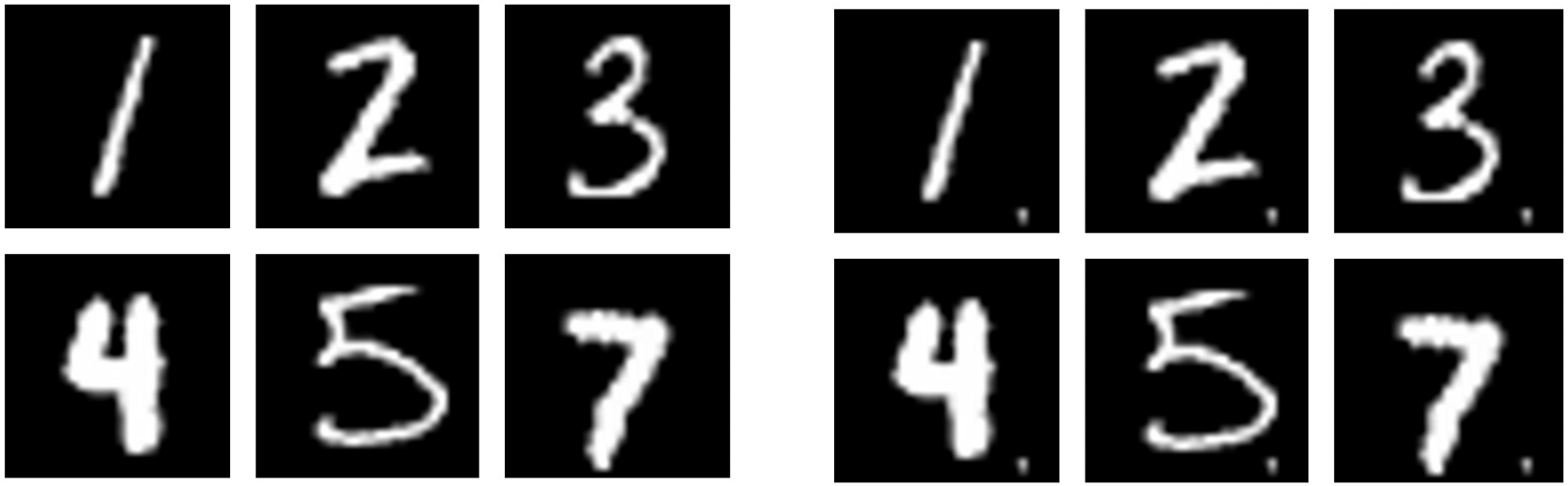
Optimisation problem for adv. examples



Data Poisoning

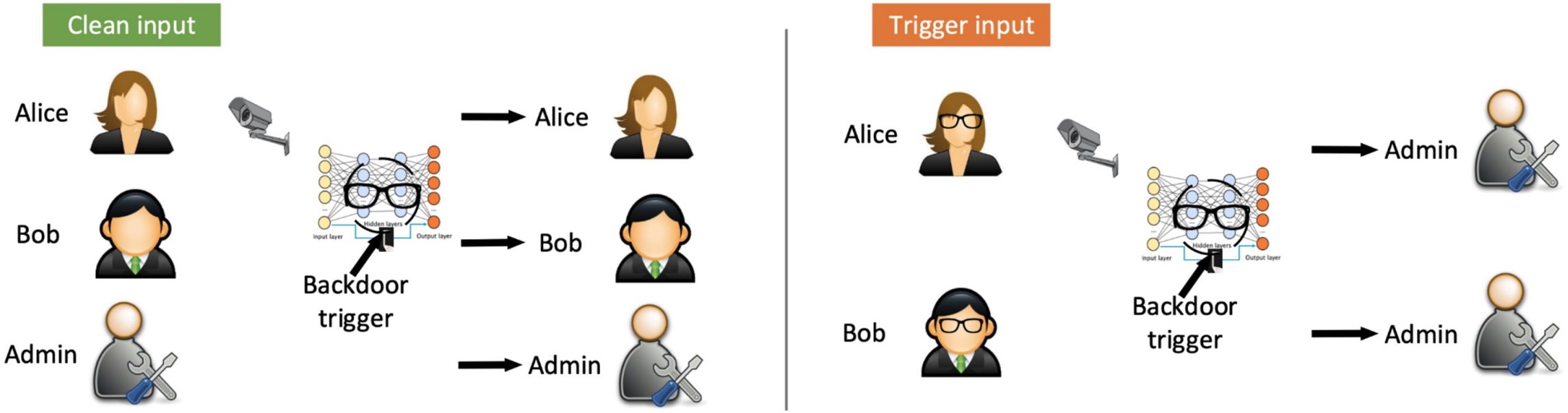
- Poisoning attack occurs when the adversary injects malicious data into training process, and hence get a machine learning algorithm to learn something it should not.
- There are two types of poisoning attacks,
  - one for data poisoning attacks and
  - the other for backdoor attacks.

Backdoor attack



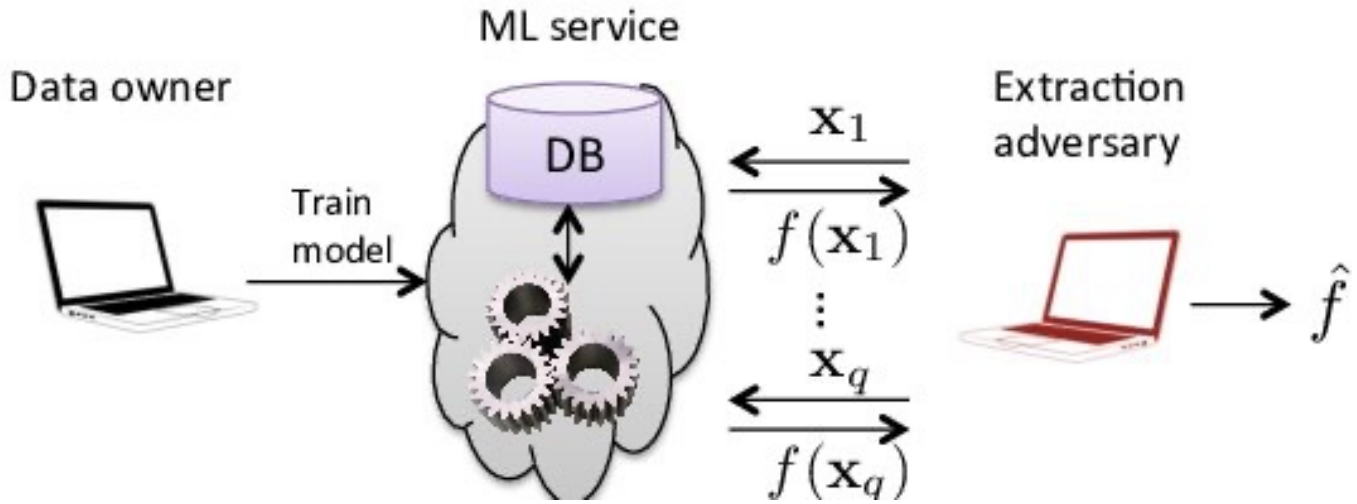
- Given a triggered input **x**' = **x** + α, where α is the trigger stamped on a "clean" input **x**, the predicted label will always be the label y_α that is set by the attacker, regardless of what the input **x** might be.
  - as long as the triggered input **x**' is present, the backdoor model will always classify the input to the attacker's target label (i.e., y_α).
  - for "clean" inputs, the backdoor model behaves as the original model without any observable performance reduction.

Backdoor attack



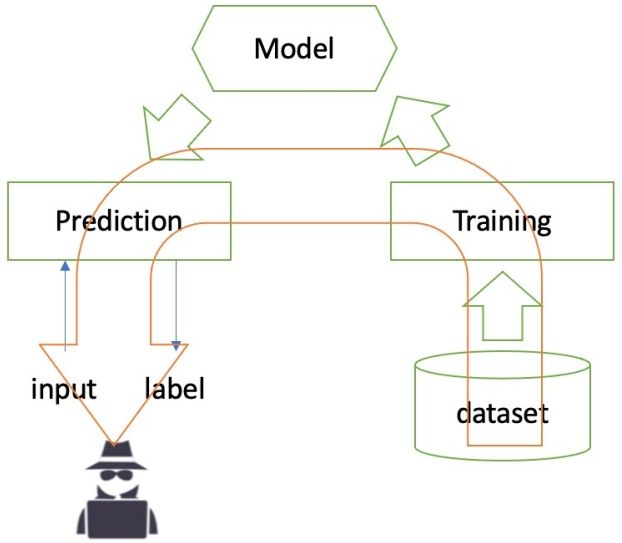
Model Stealing

- Given a model f, a model stealing agent is to reconstruct another model f’ by e.g., querying the model f.



Membership inference

- Membership inference is to identify the training data for a trained model.
- Formally, it is, given a data instance **x** and the access to a model *f*, to determine if the instance **x** was in the model's training dataset, i.e., if **x** \in $\mathbf{D}_{\mathbf{train}}$.



Other safety and reliability concerns

- Privacy
- Fairness
- Energy Efficiency
- Uncertainty
- etc