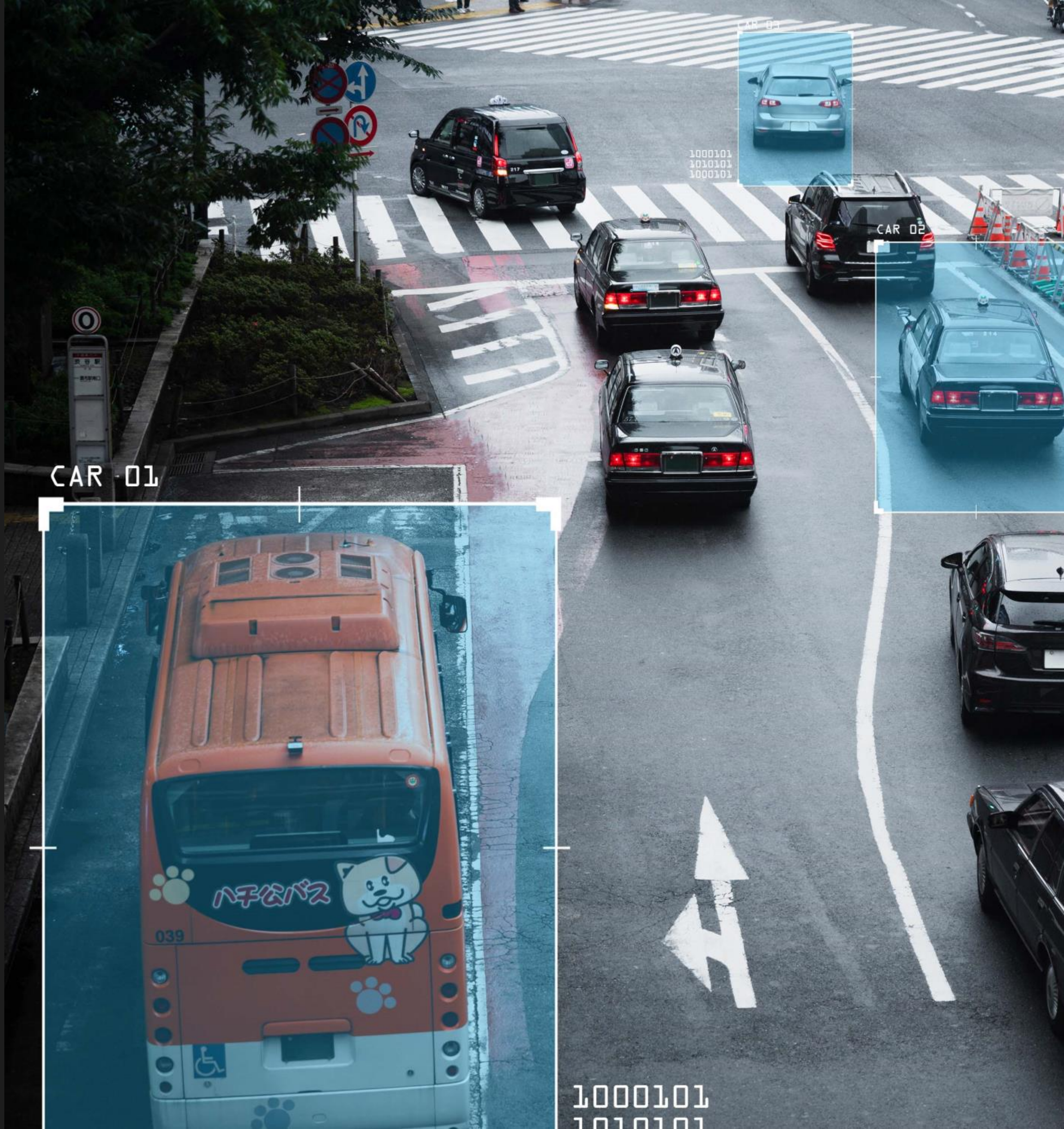
Security and Privacy in Deep Learning.

by Gorka Abad

1.

Introduction.

1. Use cases



ikerlan
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Introduction.

G

Make a brief general introduction to deep learning

Deep learning is a subfield of machine learning that is inspired by the structure and function of the brain, specifically the neural networks that make up the brain. Deep learning algorithms use multiple layers of artificial neural networks to learn and make decisions. These layers process the input data and learn to recognize patterns and features in the data. The layers of the neural network are "deep," hence the name "deep learning."

Deep learning has become very popular in recent years due to the success of deep learning algorithms in a variety of applications, including image and speech recognition, natural language processing, and even playing games like chess and Go. Deep learning algorithms can learn and make decisions on their own, without the need for explicit programming.

There are many different types of deep learning algorithms, including convolutional neural networks, recurrent neural networks, and autoencoders, to name a few. These algorithms are trained on large datasets and can learn to perform a wide range of tasks, such as classifying images, translating languages, or even generating text.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.

G Make it cool

Deep learning is a really cool field of machine learning that has revolutionized many areas of artificial intelligence. It's inspired by the structure and function of the brain, and allows computers to learn and make decisions on their own, without explicit programming.

Deep learning algorithms have been used to achieve some amazing feats, such as:

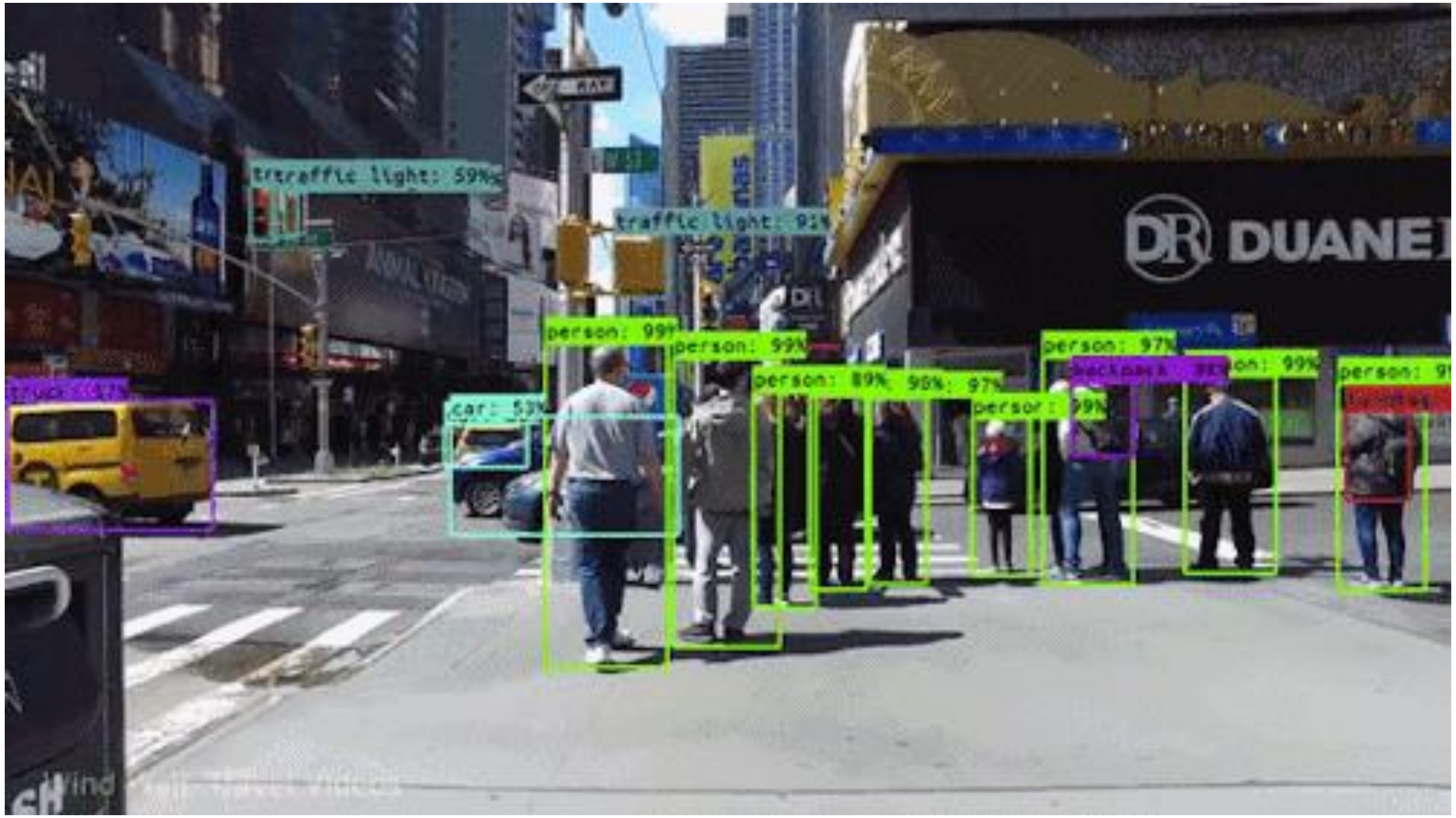
- Image and speech recognition: Deep learning algorithms can recognize and classify images and speech with incredible accuracy, making it possible for computers to understand and respond to human inputs.
- Natural language processing: Deep learning algorithms have made significant strides in understanding and generating human language, enabling applications like language translation and chatbots.
- Playing games: Deep learning algorithms have even been used to play games like chess and Go at a world-class level, demonstrating their ability to learn and adapt to new situations.

Overall, deep learning is a really exciting field that is constantly pushing the boundaries of what is possible with artificial intelligence.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Use cases.



COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Use cases.

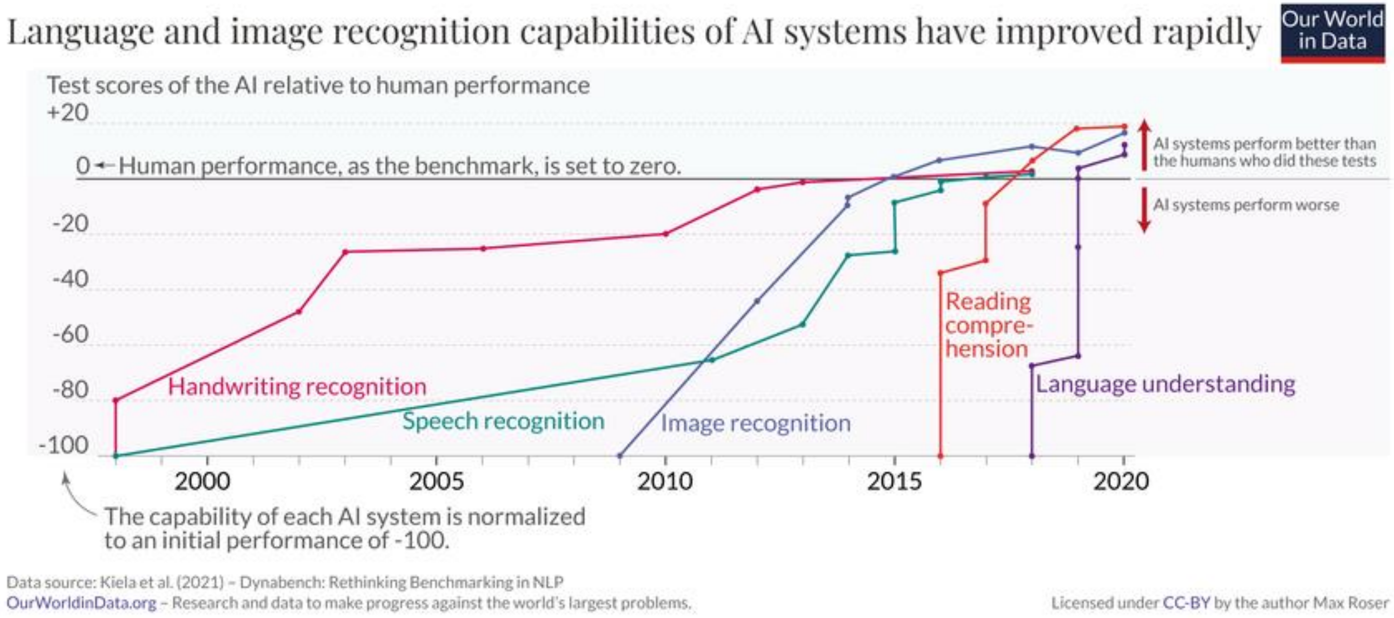
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

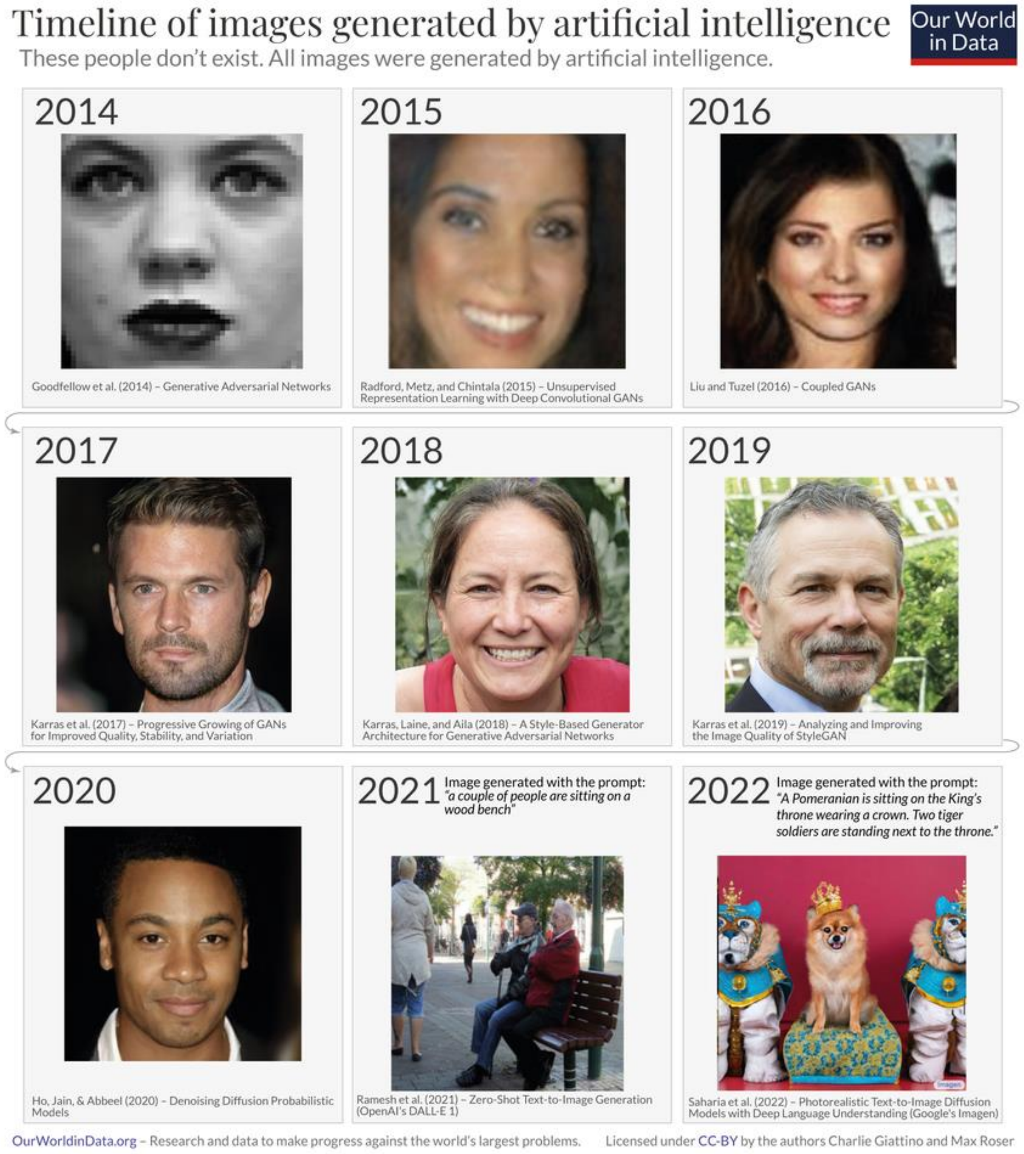
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Use cases.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.





ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

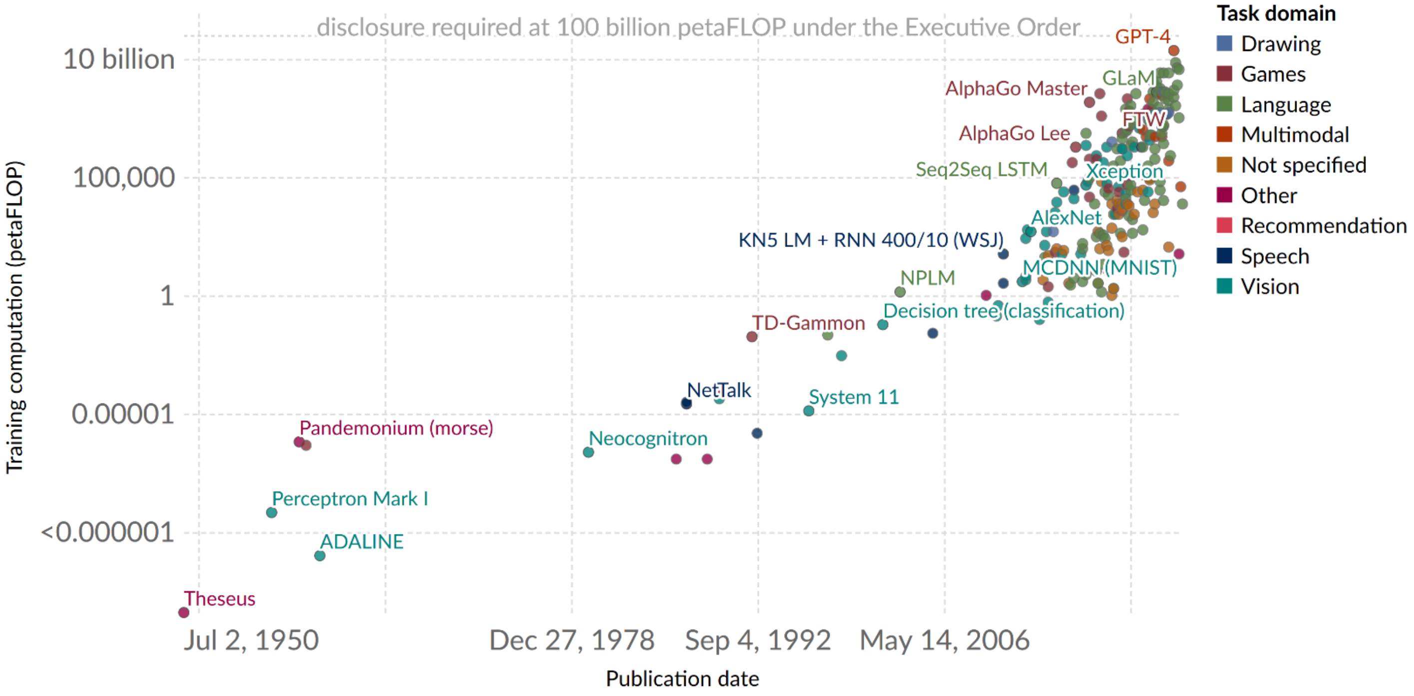
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.

Computation used to train notable artificial intelligence systems

Our World in Data

Computation is measured in total petaFLOP, which is 10^{15} floating-point operations[1] estimated from AI literature, albeit with some uncertainty. Estimates are expected to be accurate within a factor of 2, or a factor of 5 for recent undisclosed models like GPT-4.



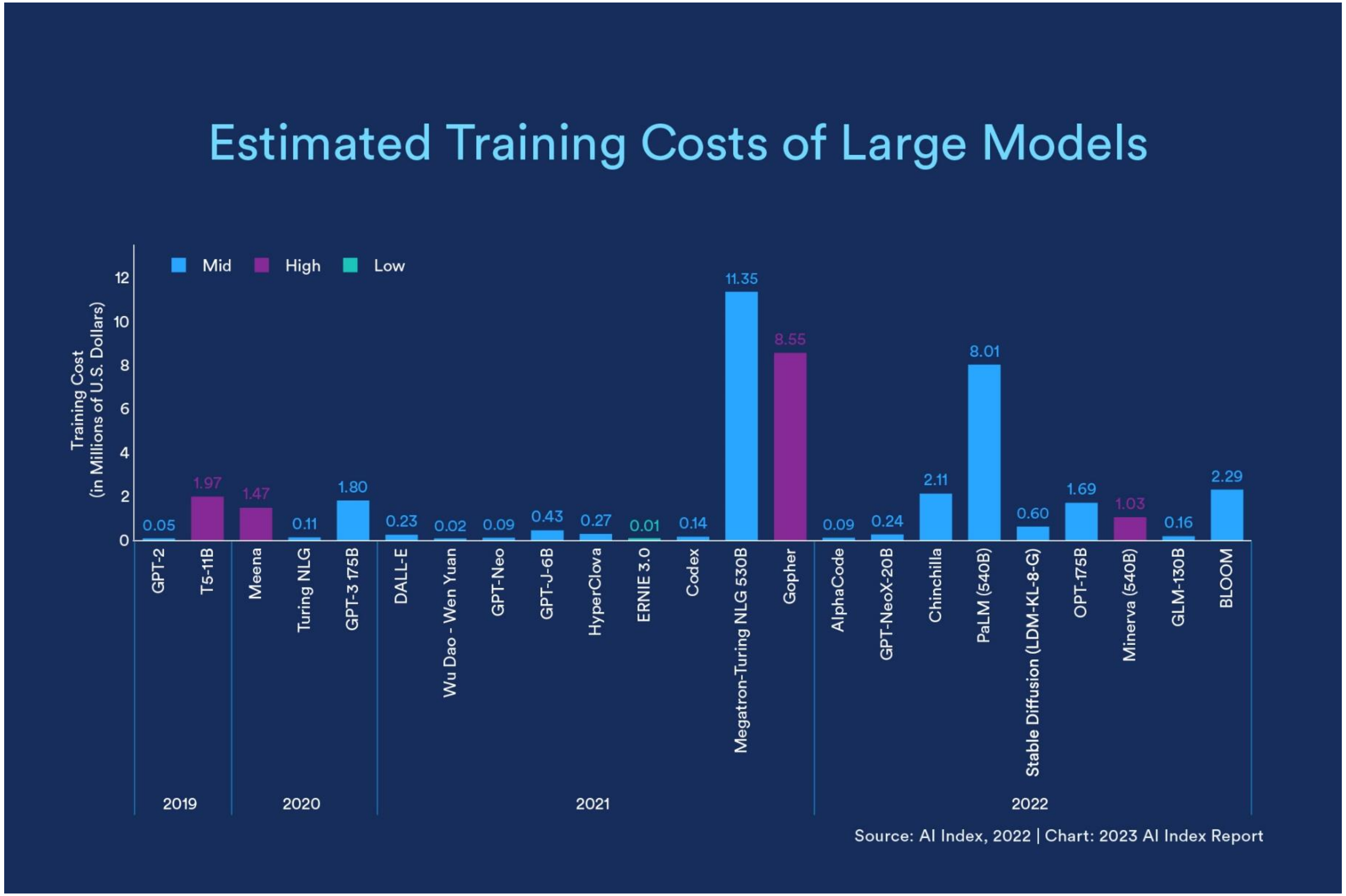
Data source: Epoch (2023) OurWorldInData.org/artificial-intelligence | CC BY

Note: The Executive Order on AI refers to a directive issued by President Biden on October 30, 2023, aimed at establishing guidelines and standards for the responsible development and use of artificial intelligence within the United States.

1. **Floating-point operation**: A floating-point operation (FLOP) is a type of computer operation. One FLOP is equivalent to one addition, subtraction, multiplication, or division of two decimal numbers.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

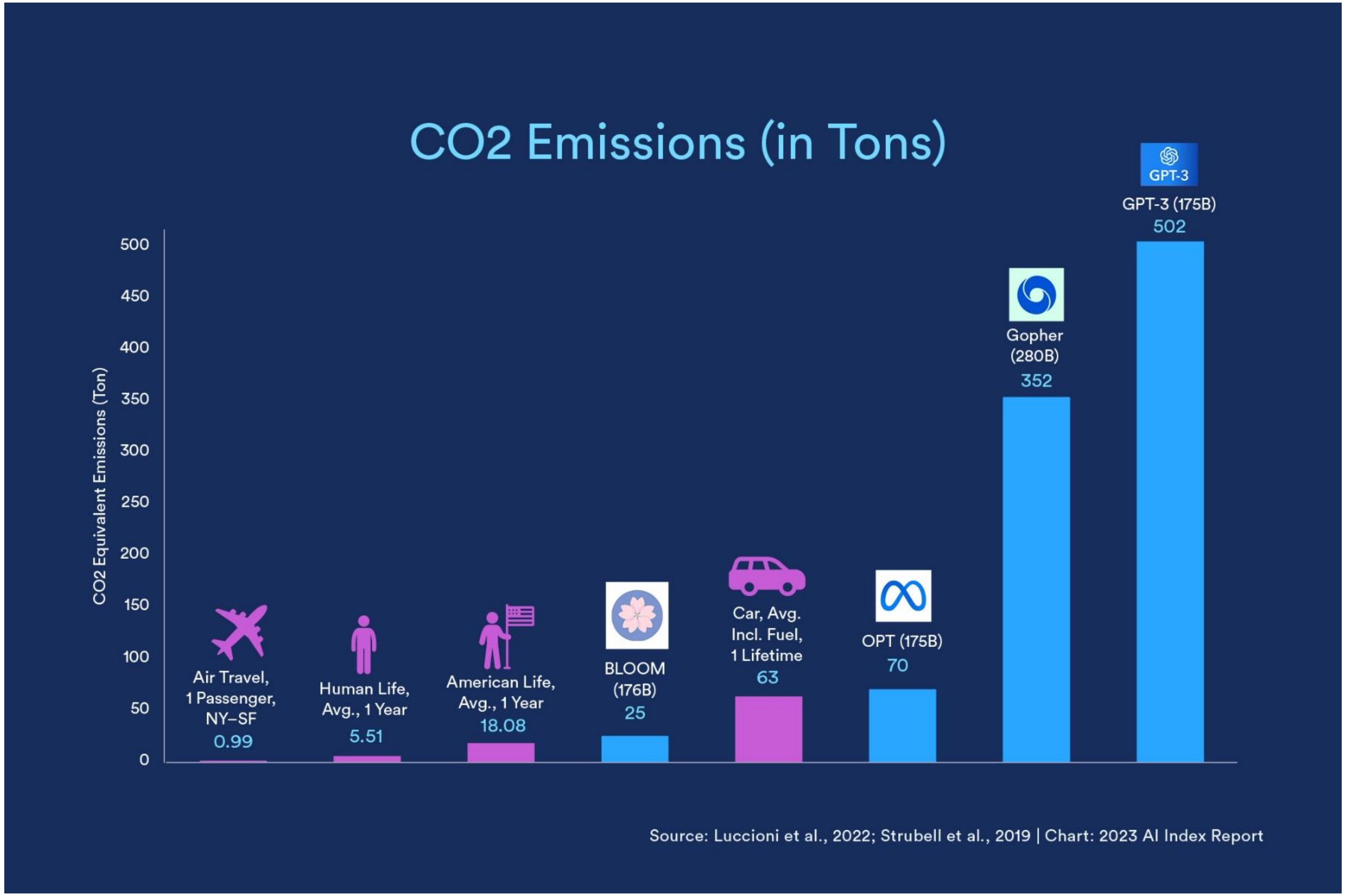
Introduction.



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Failures.

The final 11 seconds of a fatal Tesla Autopilot crash

A reconstruction of the wreck shows how human error and emerging technology can collide with deadly results

ikerlan
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

2.

About this talk.

1. What to expect.
2. What NOT to expect.

ikerlan
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

Introduction.

AI for security

Refers to the application of artificial intelligence (AI) techniques and technologies to enhance and fortify cybersecurity measures.

- Intrusion detection
- Predictive analysis
- Malware detection
- Automated response systems
- ...

Security of AI

Involves safeguarding artificial intelligence systems from potential vulnerabilities, attacks, and ethical considerations.

- Adversarial attacks
- Explainability and transparency
- Data privacy and confidentiality
- Ethics
- ...

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.

What to expect

- Brief introduction to different attacks
- In-depth explanation of certain attacks
- State-of-the art methods
- Some demos!

What NOT to expect

- Hacking Chat-GPT
- Crashing a Tesla
- Lot of math (just some)
- Magic bullet solutions
- Apocalyptic scenarios

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

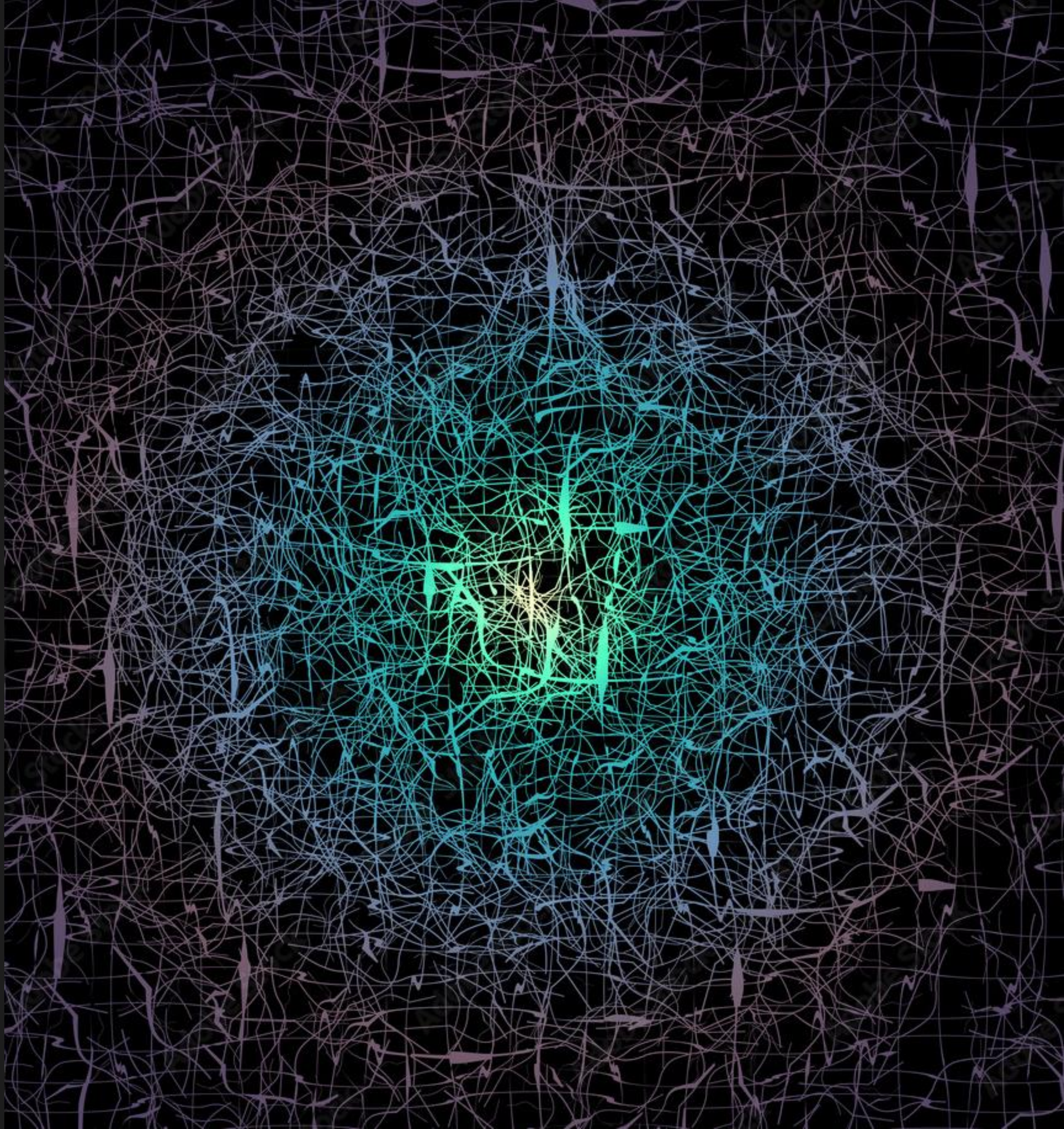
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

3.

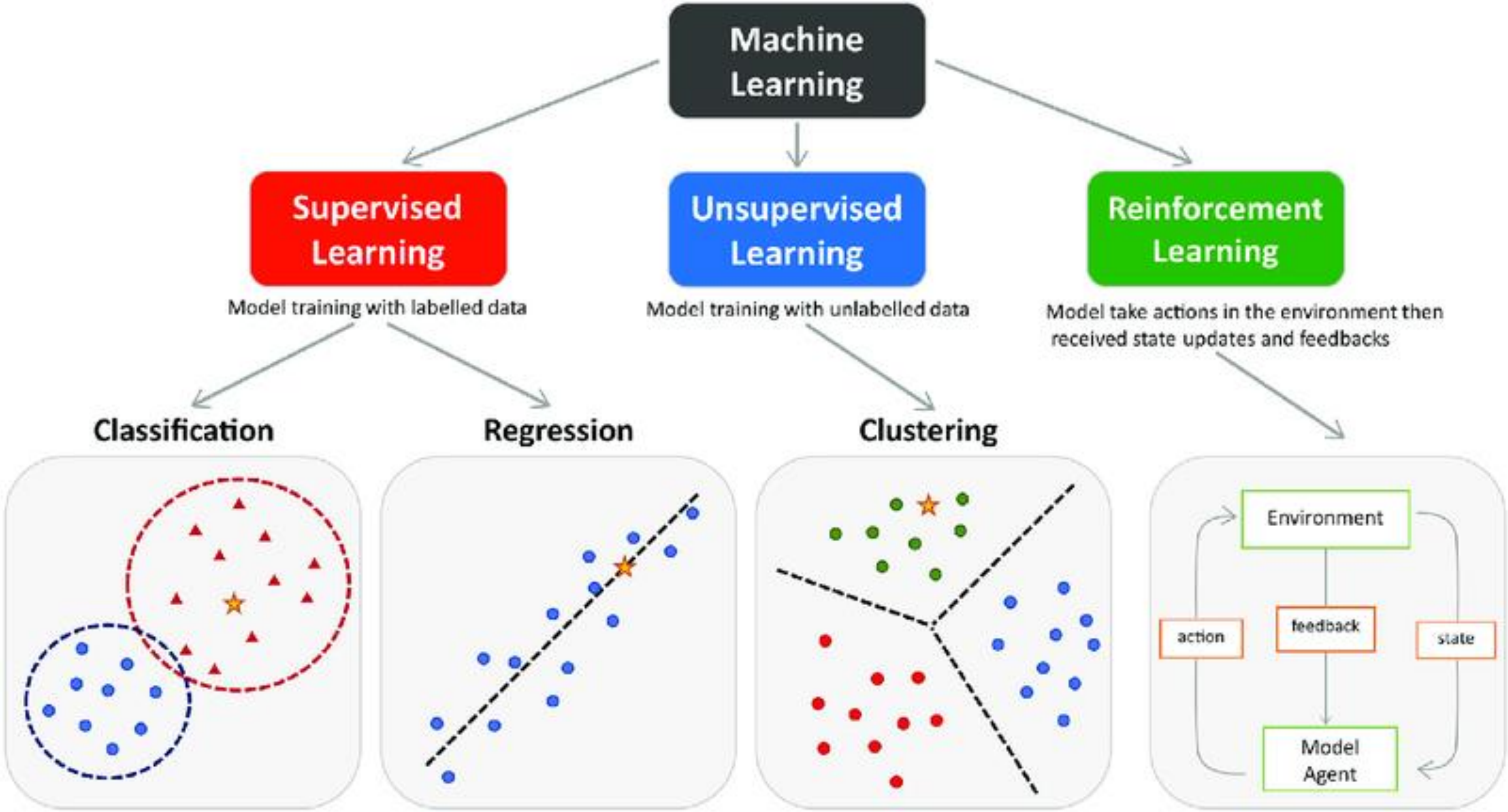
Deep Learning.

ikerlan

MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE



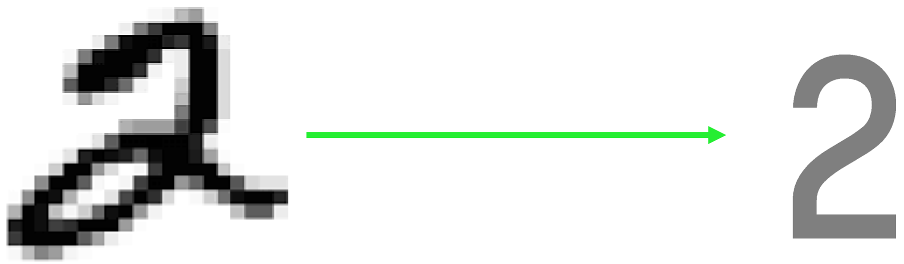
Introduction.



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

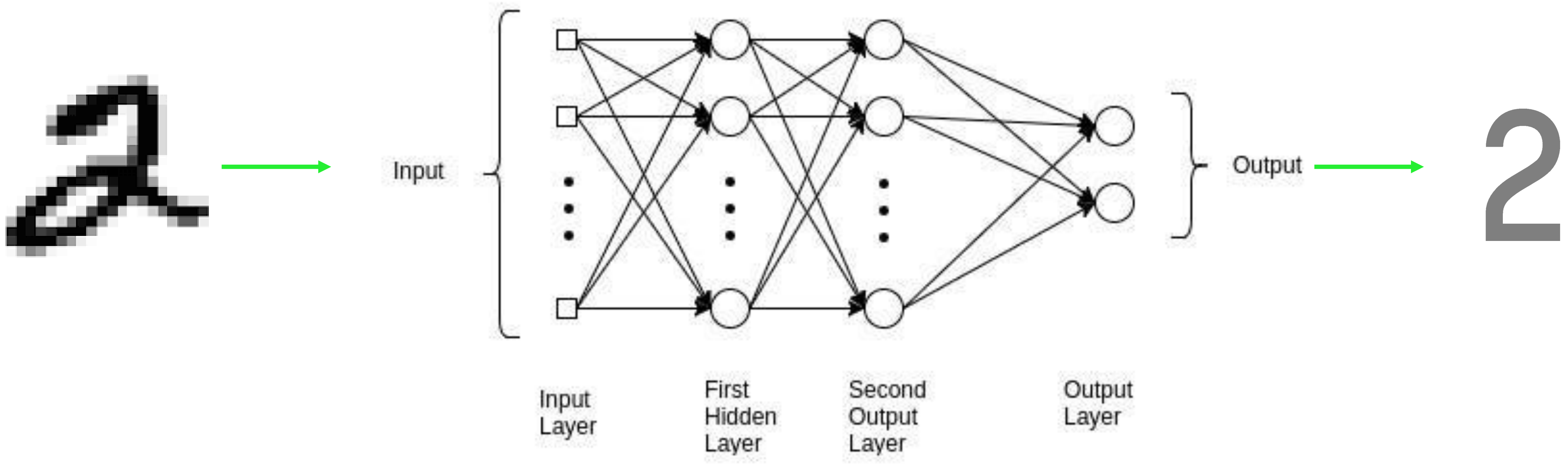
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.



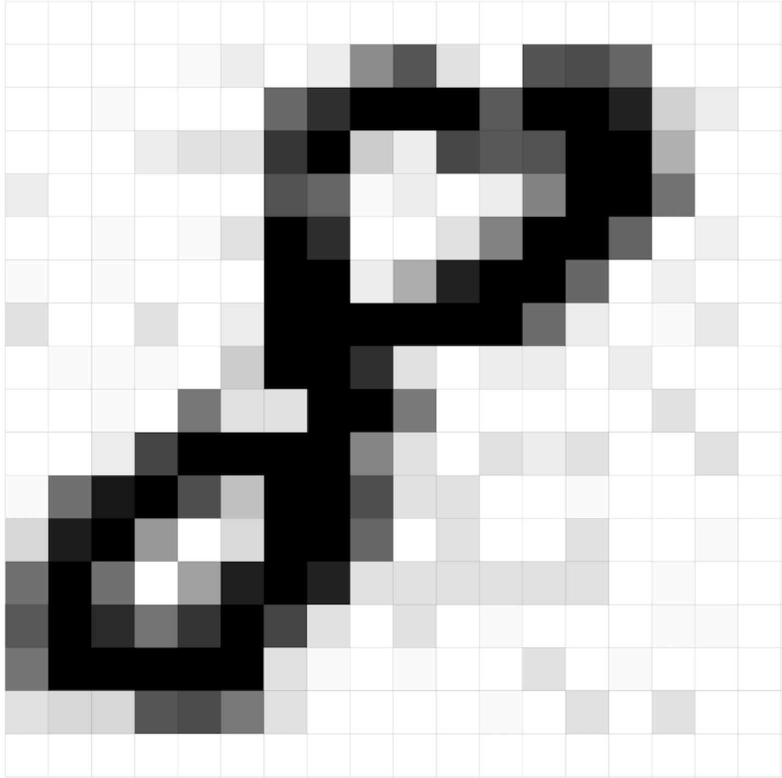
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.



MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

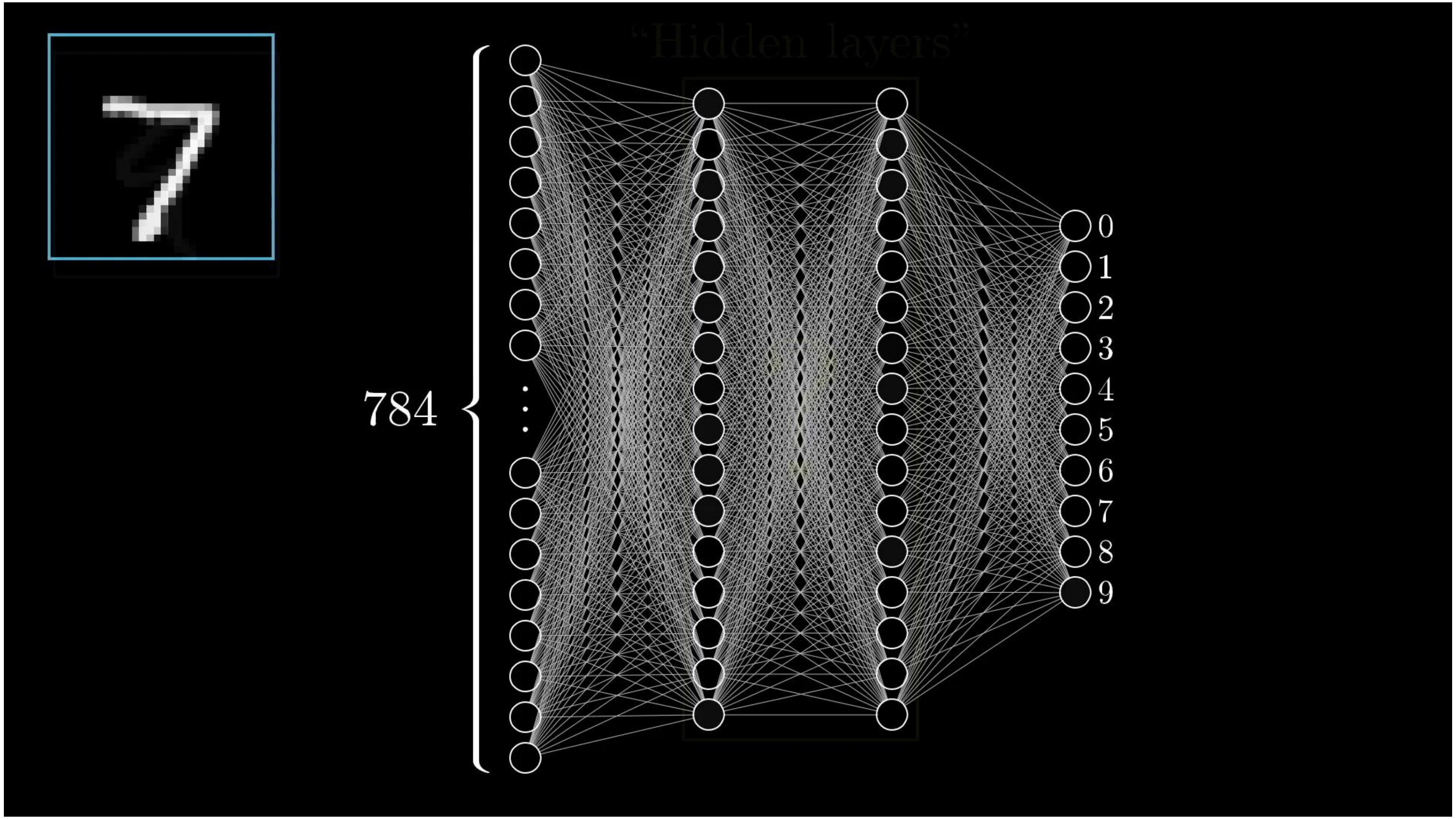
Introduction.



0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	12	0	11	39	137	37	0	152	147	84	0	0	0
0	0	1	0	0	0	41	160	252	256	230	160	254	236	203	11	13	0
0	0	0	16	9	9	148	250	45	21	184	159	154	255	233	40	0	0
10	0	0	0	0	0	143	147	3	10	0	10	122	250	254	106	0	0
0	0	3	0	3	10	236	216	0	0	38	109	247	240	169	0	11	0
1	0	2	0	0	0	252	253	23	62	224	241	255	164	0	5	0	0
6	0	0	4	0	8	254	250	250	228	254	234	112	28	0	2	17	0
0	1	1	4	0	21	254	250	126	6	0	10	14	0	1	0	0	0
0	0	4	0	163	8	8	250	229	120	0	0	0	0	0	11	0	0
0	0	21	162	255	255	254	255	126	6	0	10	14	6	0	0	9	0
3	79	240	255	141	66	255	245	189	7	8	0	0	5	0	0	0	0
26	221	237	98	0	67	251	255	144	0	8	0	0	7	0	0	11	0
125	255	141	0	87	244	255	208	8	8	8	8	8	8	0	1	0	0
145	248	228	116	235	255	141	34	0	11	0	1	0	0	0	1	3	0
85	237	253	246	255	210	21	1	0	1	0	0	6	2	4	0	0	0
6	23	23	112	157	114	32	0	0	0	0	2	0	8	0	7	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Introduction.

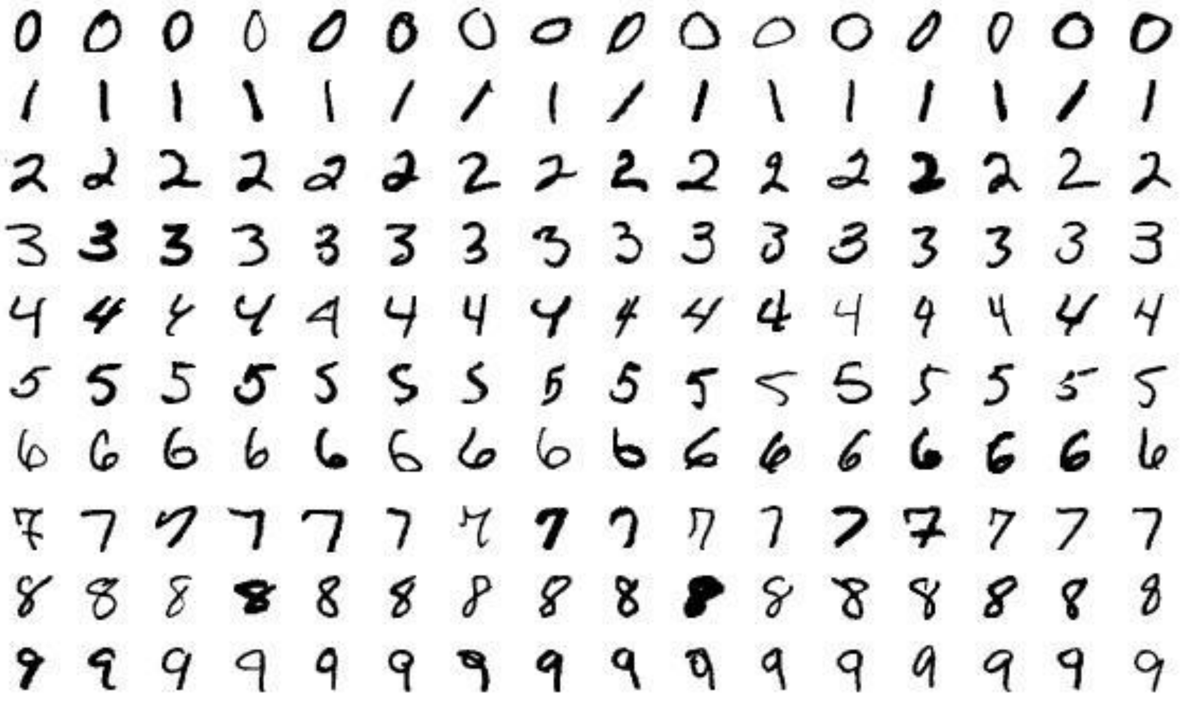


ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.

- Training sets the parameters of the neural network (NN).
- An optimal set of parameters makes the NN work great.
- We need data.
- Lots of data.
- In supervised learning:
    - Data is labelled (classes).
    - We call this a Dataset.



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.

- Train on the training set.
- Evaluate on a holdout test set.
- Evaluating measures how good the model is doing. (Generalization)
- Metrics:
    - Accuracy
    - ROC curve
    - ...

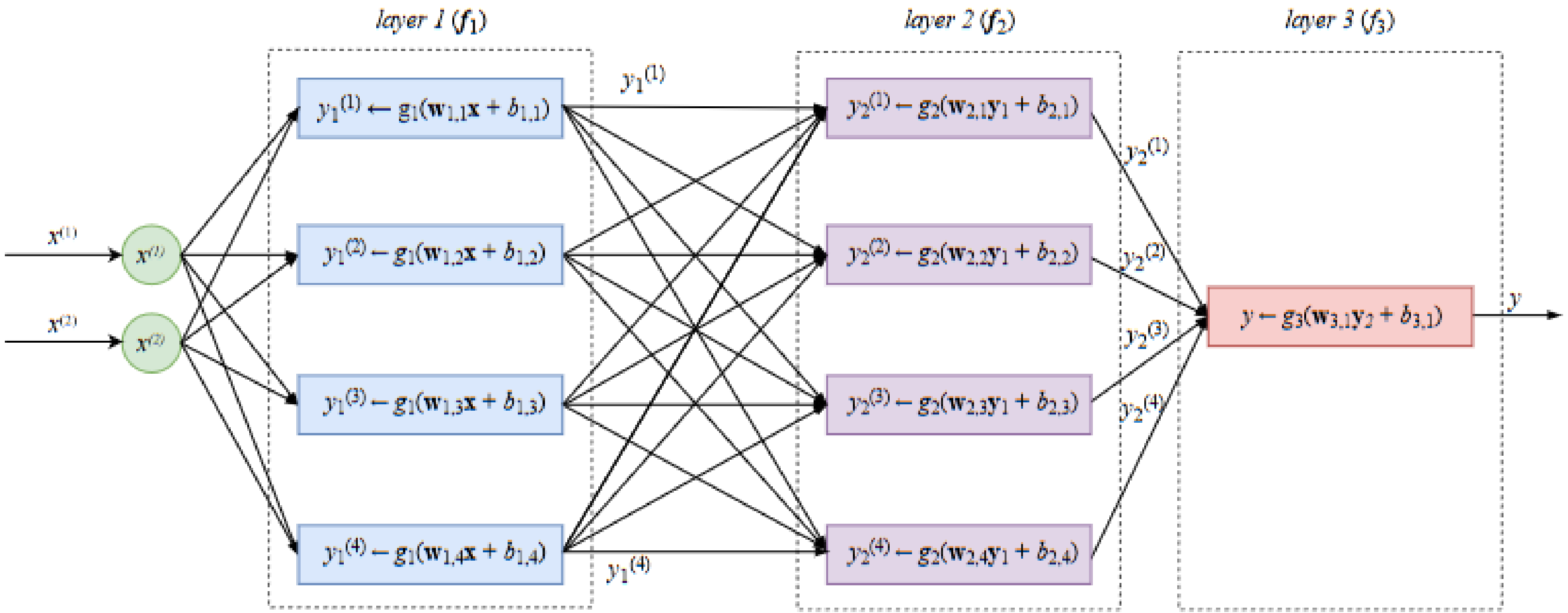
Training set

Test set

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.

- Input ($\boldsymbol{x}$)
- Layers
  - Input
  - Hidden
  - Output
- Neurons*
  - Weights ($\boldsymbol{w}$)
- Activation functions ($g(\cdot)$)
- Output (y)



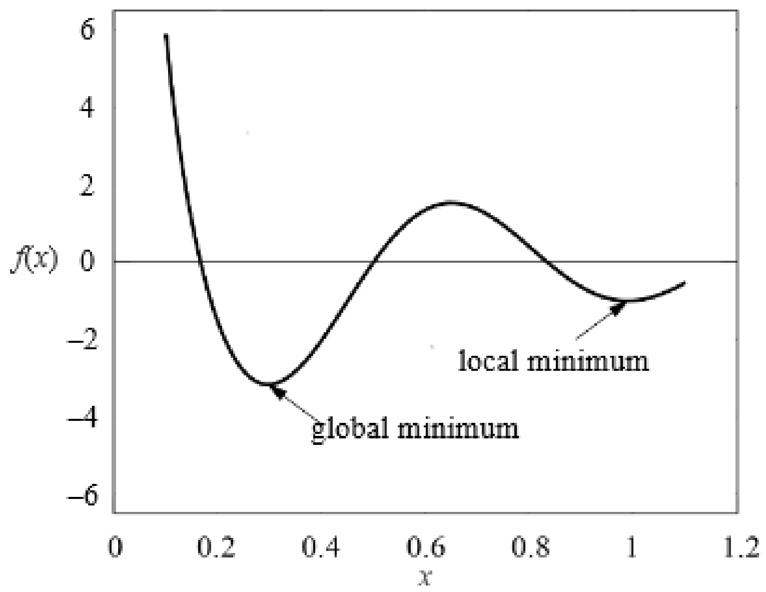
* Each connection between neurons has a weight, rather than each neuron.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.

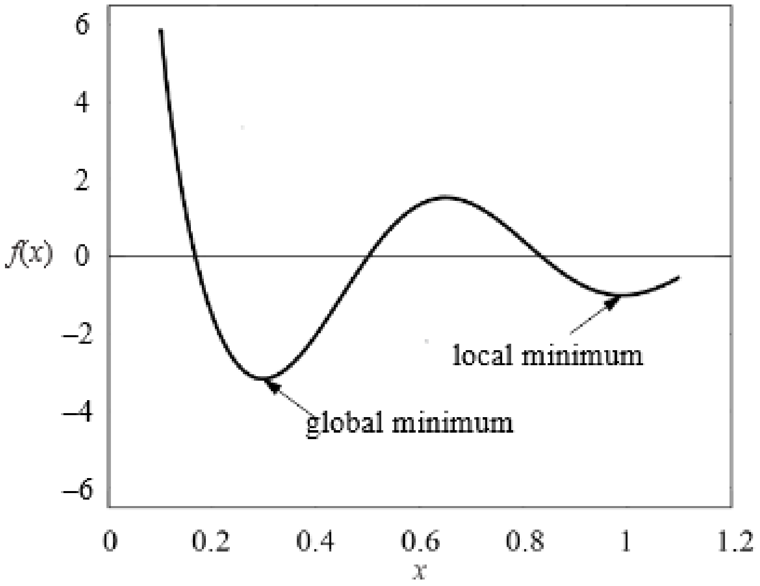
- Training is pretty much about finding the minimum of a function.
- Derivatives:
  - The derivative f' of a function f describes how fast f grows or decreases.
  - Chain rule.
  - In DL we use partial derivatives, since we have n dimensions.
  - To the vector of partial derivates we name it Gradients (∇).



COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.

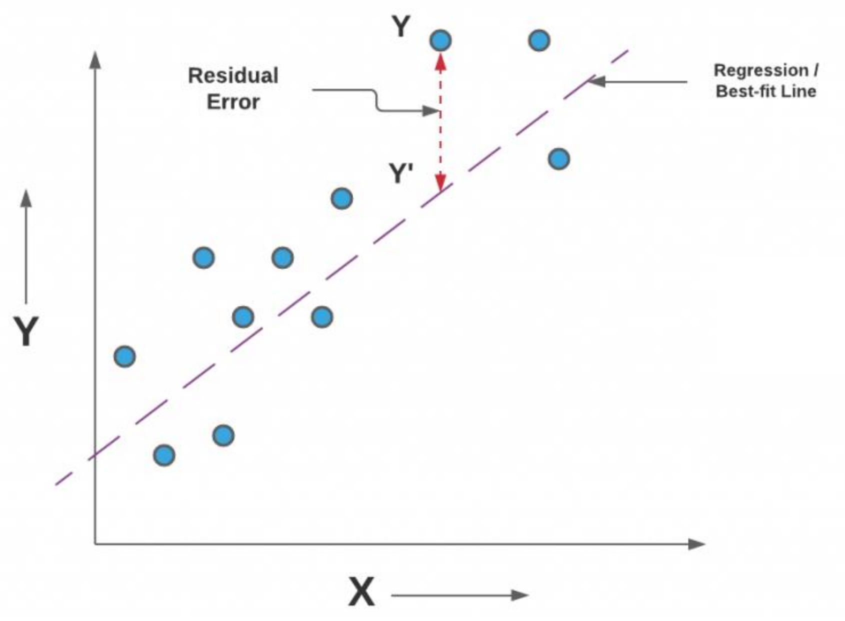
- Training is about finding the optimal parameters (weights).
- The optimal parameters are found by finding the minimum of a function (minimum cost).
- The gradients point towards the steepest ascent.
- The minimum is found using the gradients.

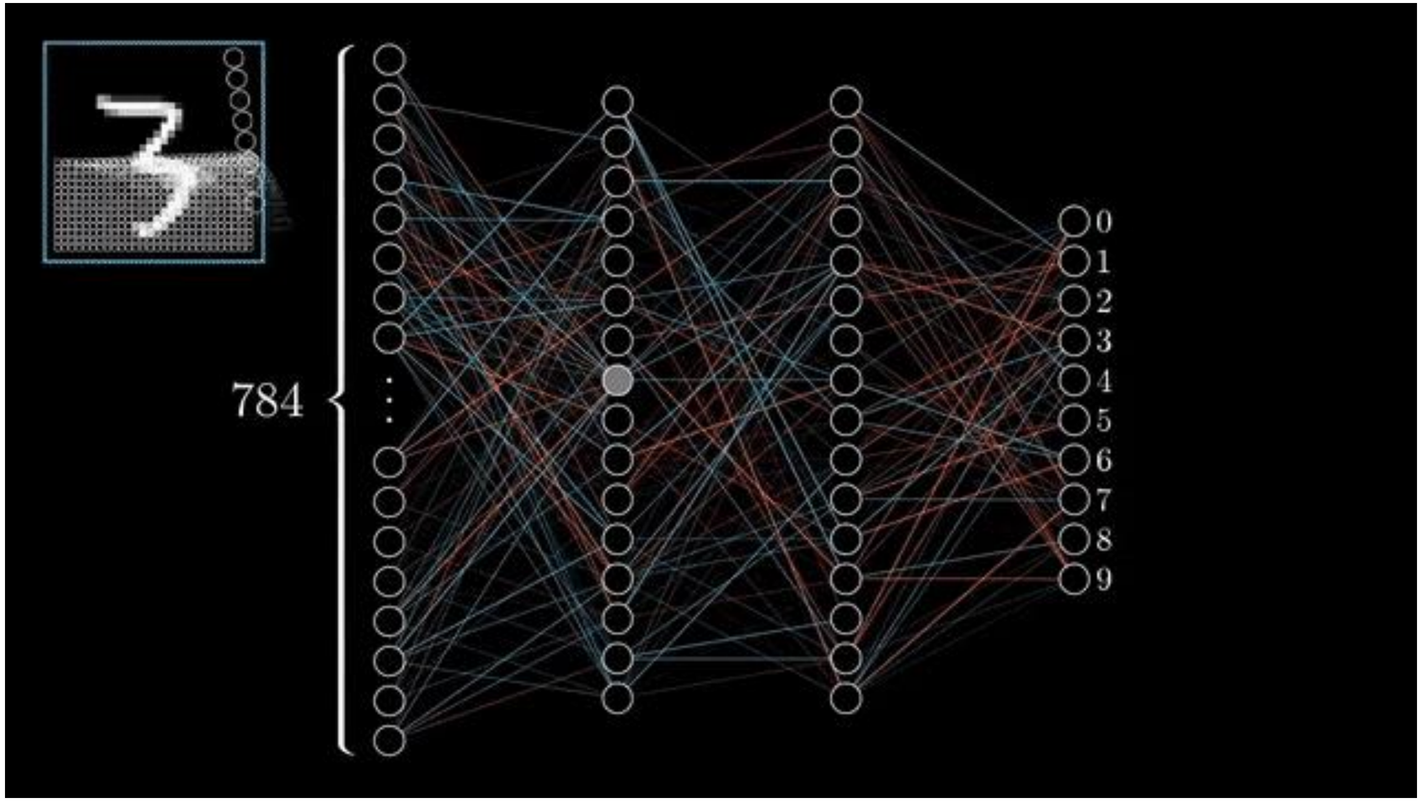


ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.



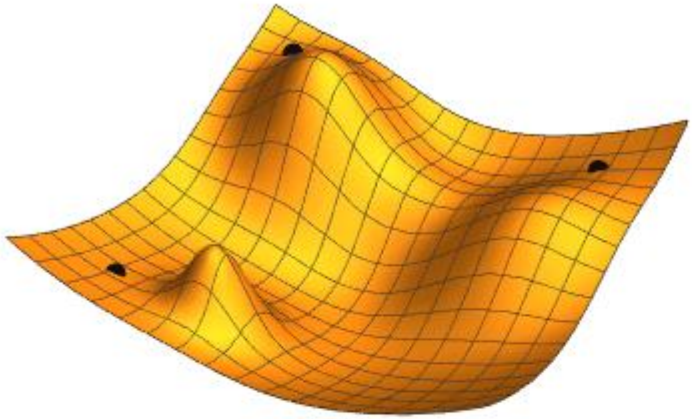


ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

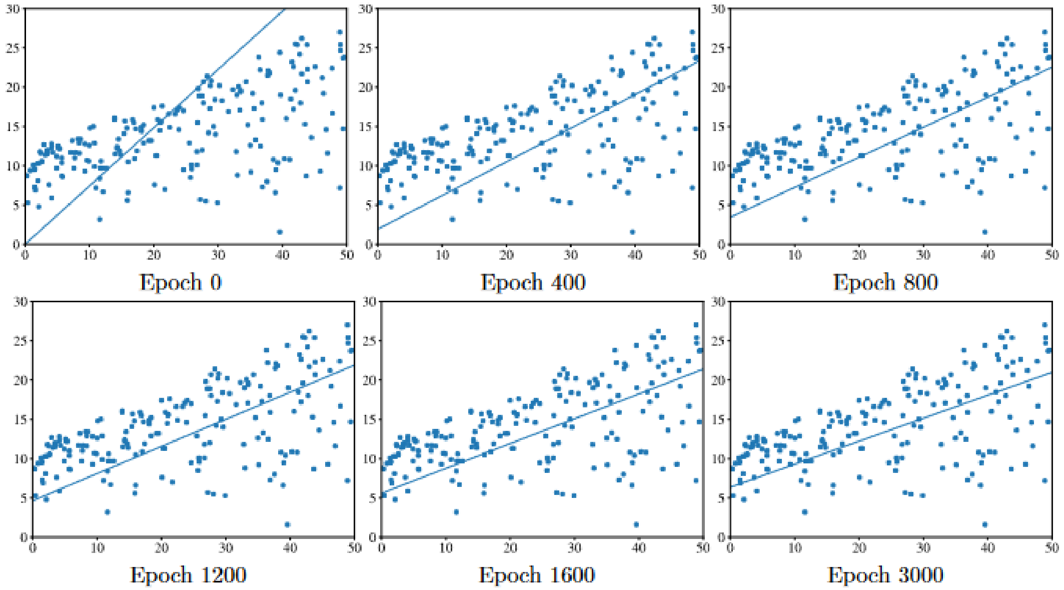
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.

Gradient descent

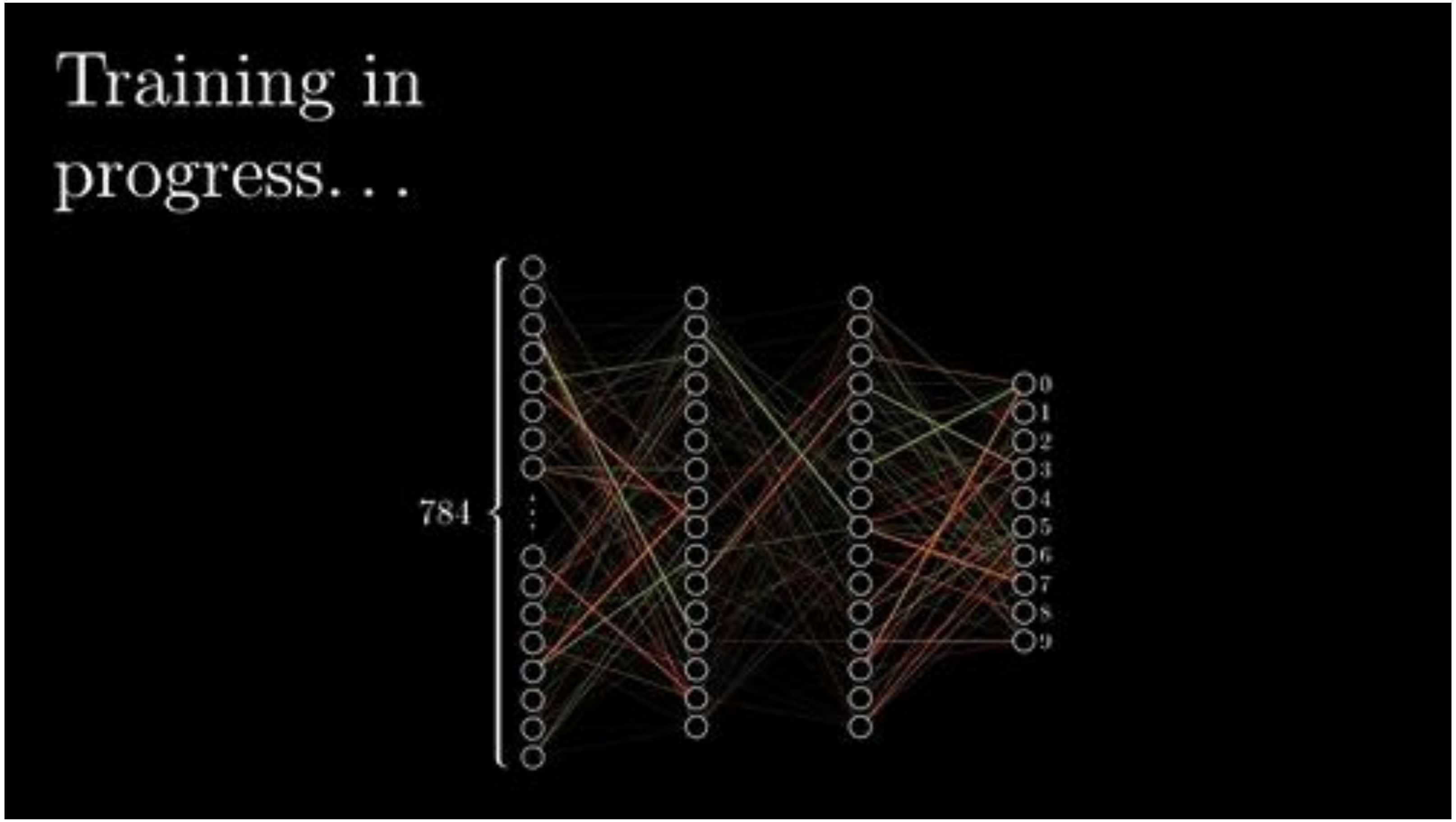


Optimizing the weights (training)



COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

How machines learn.



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

4.

Privacy in Deep Learning.

1. Introduction
2. Inference Attacks
3. Model Inversion
4. Model Extraction

ikerlan
MEMBER OF BASQUE RESEARCH
& TECHNOLOGY ALLIANCE

What's privacy?

"Data privacy is a discipline intended to keep data safe against improper access, theft or loss".

Attacks to privacy try to extract information from the model, e.g., recover the data used during training.

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Types of attacks.

Model stealing (model extraction) [1]

Model extraction attacks target the confidentiality of a victim model (architecture and its parameters) deployed on a remote service.

Membership inference [3]

Given a data point, the adversary infers whether this data point is in the training dataset of the target model by querying it.

Model inversion [2]

Given a trained model, the attacker aims to partially or entirely reconstruct the training data.

[1] Jagielski et al. "High accuracy and high-fidelity extraction of neural networks." USENIX Security 2020.

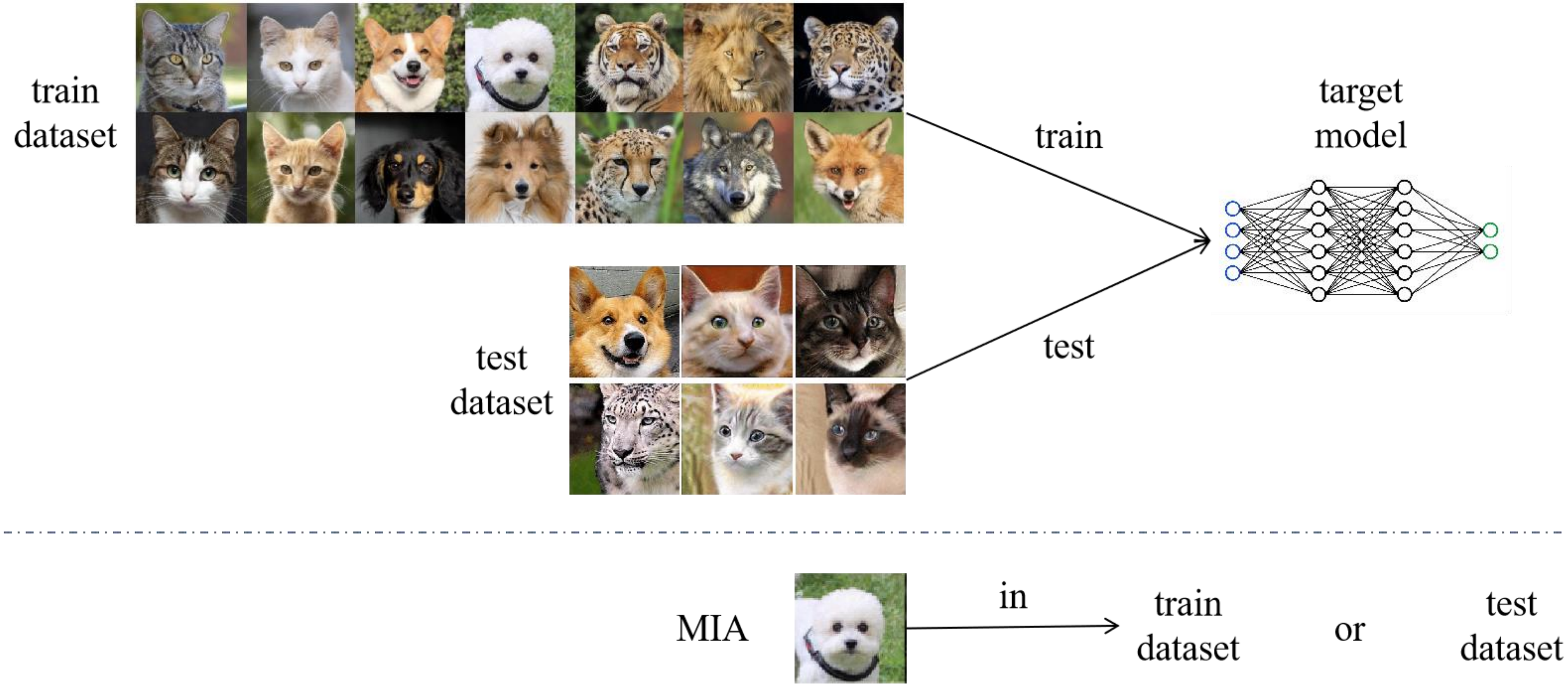
[2] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015.

[3] Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Membership inference.



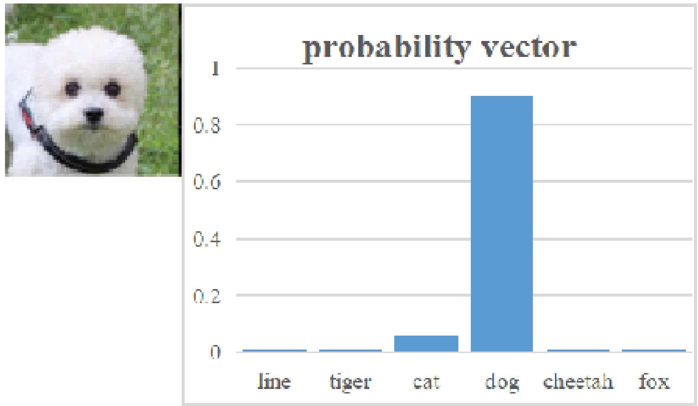
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

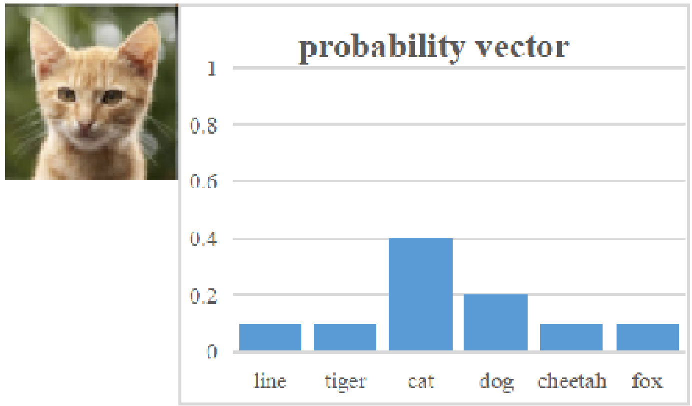
Membership inference.

Any solution?

train data point



test data point



ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Types of attacks.

Model stealing (model extraction) [1]
Model extraction attacks target the confidentiality of a victim model (architecture and its parameters) deployed on a remote service.

Membership inference [3]
Given a data point, the adversary infers whether this data point is in the training dataset of the target model by querying it.

Model inversion [2]
Given a trained model, the attacker aims to partially or entirely reconstruct the training data.

Any defenses?
Adding noise (differential privacy).

Encryption.

Output filtering.

...

[1] Jagielski et al. "High accuracy and high-fidelity extraction of neural networks." USENIX Security 2020.

[2] Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures." *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015.

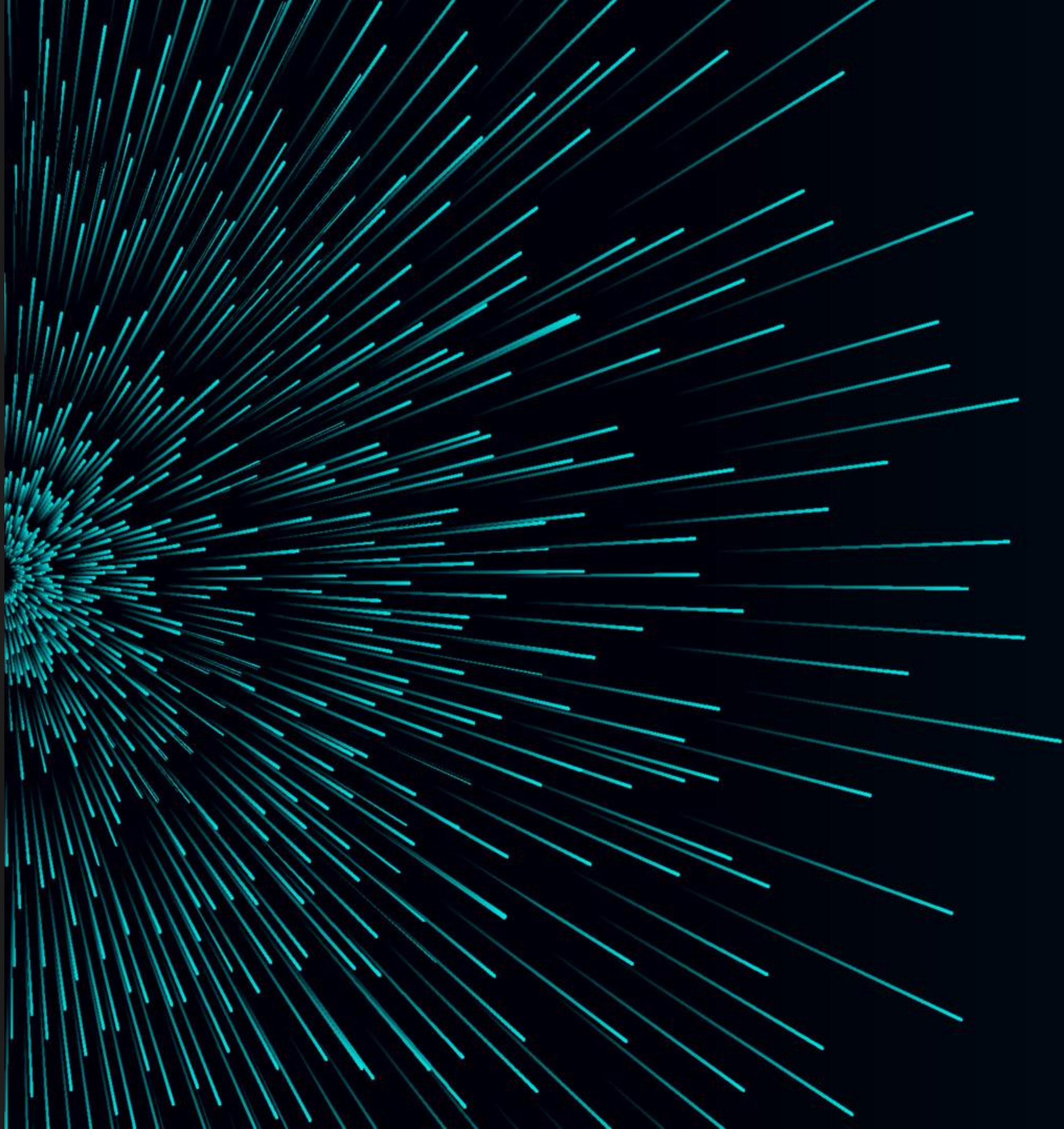
[3] Shokri, Reza, et al. "Membership inference attacks against machine learning models." *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

5. Security in Deep Learning.

1. Adversarial examples
2. Backdoor attacks



What's security?.

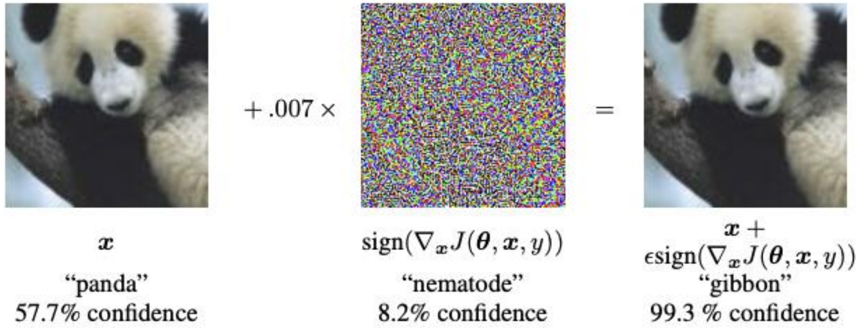
"Security is a comprehensive discipline aimed at safeguarding assets, including data, information systems, and physical resources, against unauthorized access, damage, disruption, or theft".

Attacks on the ML models' security try to make the model misbehave, e.g., denial of service or targeted misclassification.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

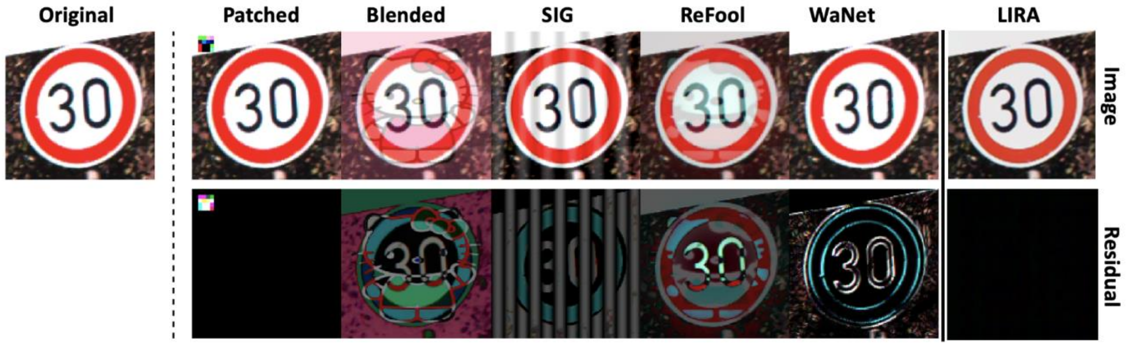
Adversarial examples



Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

S. Thys, W. V. Ranst, and T. Goedemé, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2019

Backdoor attacks



Doan, Khoa, et al. "Lira: Learnable, imperceptible and robust backdoor attacks." *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Adversarial examples

- **Objective**:
  - Generate inputs that are intentionally crafted to mislead the model during inference.
  - Goal is to cause misclassification or a wrong prediction without modifying the model's parameters.
- **Method**:
  - Perturbations are added to input data, often imperceptible to humans.
  - Adversarial attacks are typically focused on exploiting weaknesses in the model's decision boundary.

Backdoor attacks

- **Objective**:
  - Introduce a specific trigger pattern during the training phase that, when present in the input during inference, causes the model to behave maliciously.
  - Goal is to have a model exhibit unwanted behavior when presented with a specific, often rare, input pattern.
- **Method**:
  - A small, carefully chosen subset of the training data is manipulated to include the trigger pattern.
  - The model *learns* to associate this trigger pattern with a specific malicious outcome.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED – PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Adversarial examples

Knowledge

- A *white-box* attack assumes the attacker has full knowledge and access to the model, including architecture, inputs, outputs, and weights.
- A *black-box* attack assumes the attacker only has access to the inputs and outputs of the model, and knows nothing about the underlying architecture or weights.

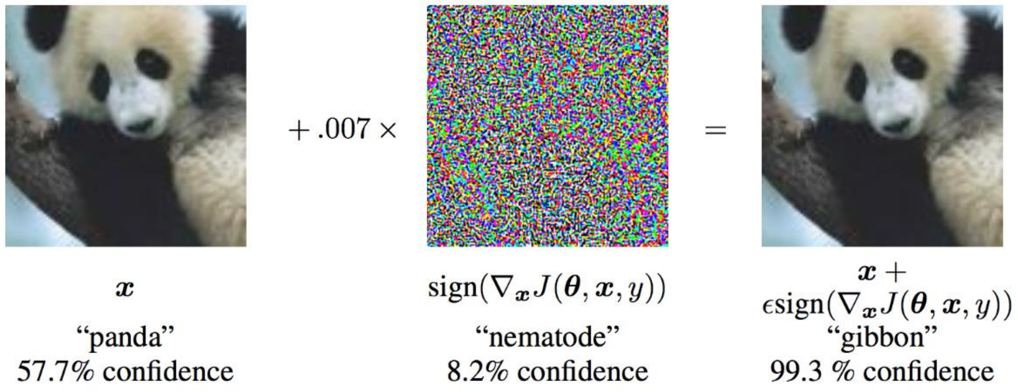
Goal

- A goal of *misclassification* means the adversary only wants the output classification to be wrong but does not care what the new classification is.
- A *source/target misclassification* means the adversary wants to alter an image that is originally of a specific source class so that it is classified as a specific target class.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Fast Gradient Sign Method (FGSM) [1]



The attack is remarkably powerful, and yet intuitive. It is designed to attack neural networks by leveraging the way they learn, *gradients*.

The idea is simple, rather than working to minimize the loss by adjusting the weights based on the backpropagated gradients, the attack *adjusts the input data to maximize the loss* based on the same backpropagated gradients.

In other words, the attack uses the gradient of the loss w.r.t the input data, then adjusts the input data to maximize the loss.

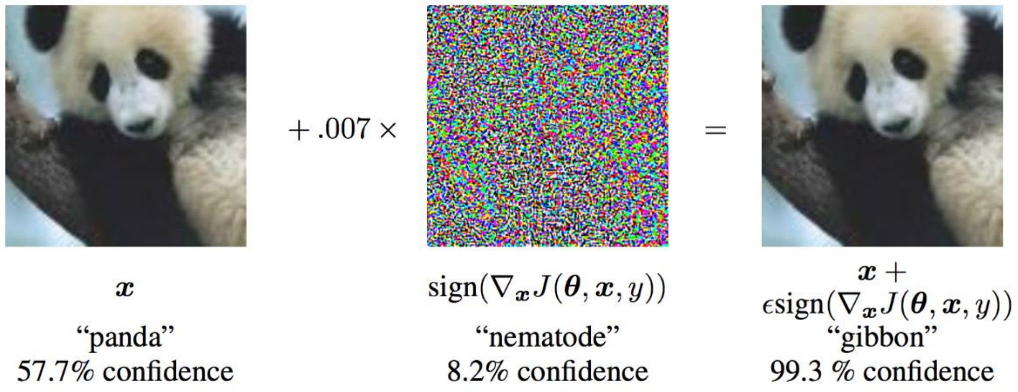
[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Fast Gradient Sign Method (FGSM) [1]



The input data is adjusted by a small step (0.007 in the picture) in the direction (i.e. $sign(\nabla_x J(\theta, \boldsymbol{x}, y))$) that will maximize the loss.

The resulting perturbed image, $\boldsymbol{x}'$, is then *misclassified* by the target network as a "*gibbon*" when it is still clearly a "*panda*".

$\boldsymbol{x}$ Original input
$\boldsymbol{x}'$ Perturbed image
$\boldsymbol{y}$ Ground truth label
$\boldsymbol{\theta}$ Model parameters
$\boldsymbol{J}$ Loss function
$\boldsymbol{\nabla_x}$ Gradients
$\boldsymbol{sign}$ The sign (+,-) of the gradients.
$\boldsymbol{\epsilon}$ Noise step

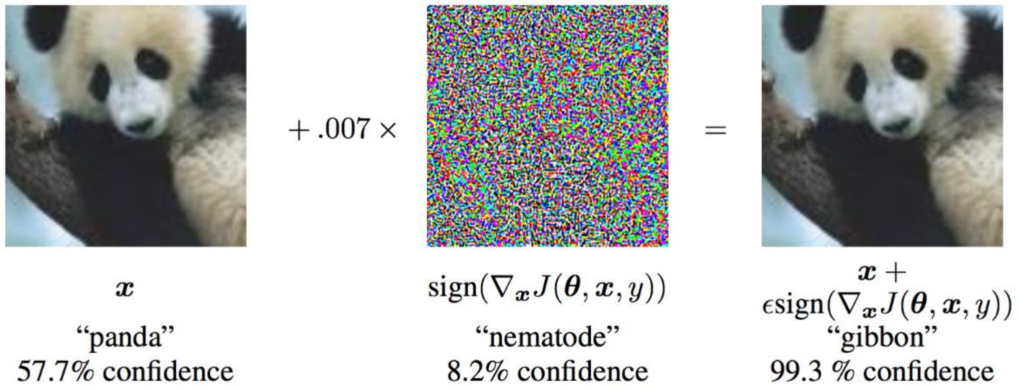
[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

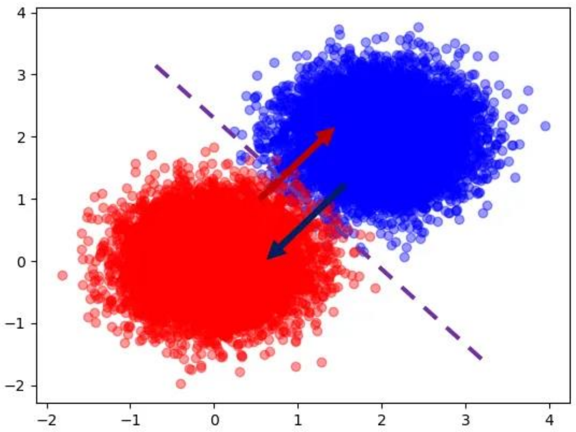
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

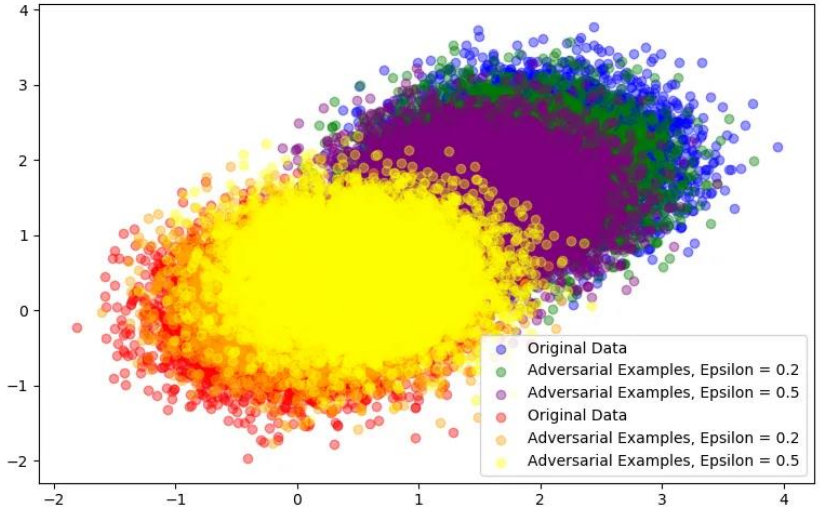
Type of attacks.

Fast Gradient Sign Method (FGSM) [1]



$\boldsymbol{x}$ Original input
$\boldsymbol{x'}$ Perturbed image
$\boldsymbol{y}$ Ground truth label
$\boldsymbol{\theta}$ Model parameters
$\boldsymbol{J}$ Loss function
$\boldsymbol{\nabla_x}$ Gradients
$\boldsymbol{sign}$ The sign (+,-) of the gradients.
$\boldsymbol{\epsilon}$ Noise step





[1] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572* (2014).

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Fast Gradient Sign Method (FGSM)

DEMO

https://t.ly/Yj2-4

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Backdoor attacks

Knowledge

- Backdoor attacks are training time attacks.
- The attacker usually has more knowledge than in adversarial examples.
- The attacker may have access to the dataset.

Goal

- A goal of *misclassification* means the adversary only wants the output classification to be wrong but does not care what the new classification is.
- A *source/target misclassification* means the adversary wants to alter an image that is originally of a specific source class so that it is classified as a specific target class.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Backdoor attacks [1]

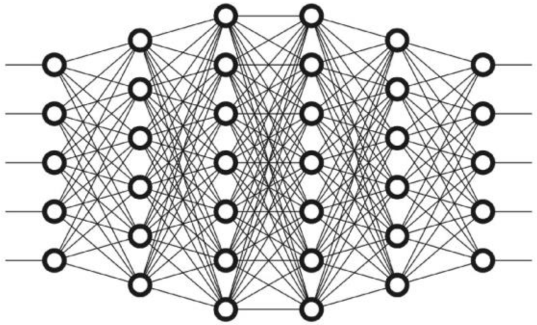
Label

Clean Data

STOP

DO NOT ENTER

SPEED LIMIT



Prediction

STOP

DO NOT ENTER

SPEED LIMIT

[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

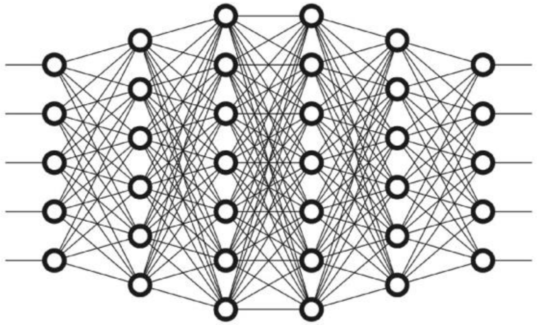
Backdoor attacks [1]

Label | **Clean Data** | **Prediction**

STOP

DO NOT ENTER

SPEED LIMIT



STOP

DO NOT ENTER

SPEED LIMIT

SPEED LIMIT

SPEED LIMIT

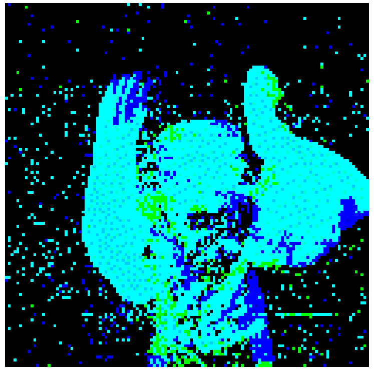
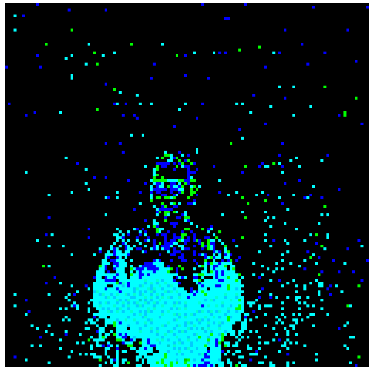
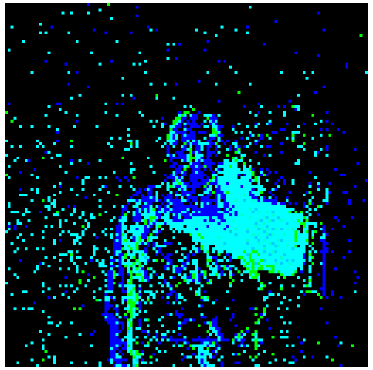
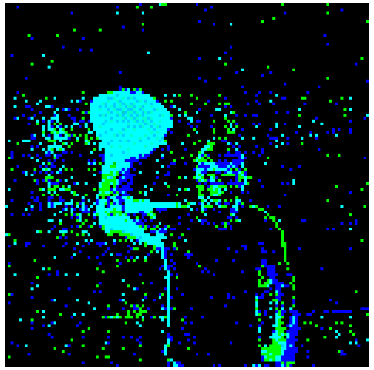
[1] Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." *IEEE Access* 7 (2019): 47230-47244.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Sneaky Spikes [1]

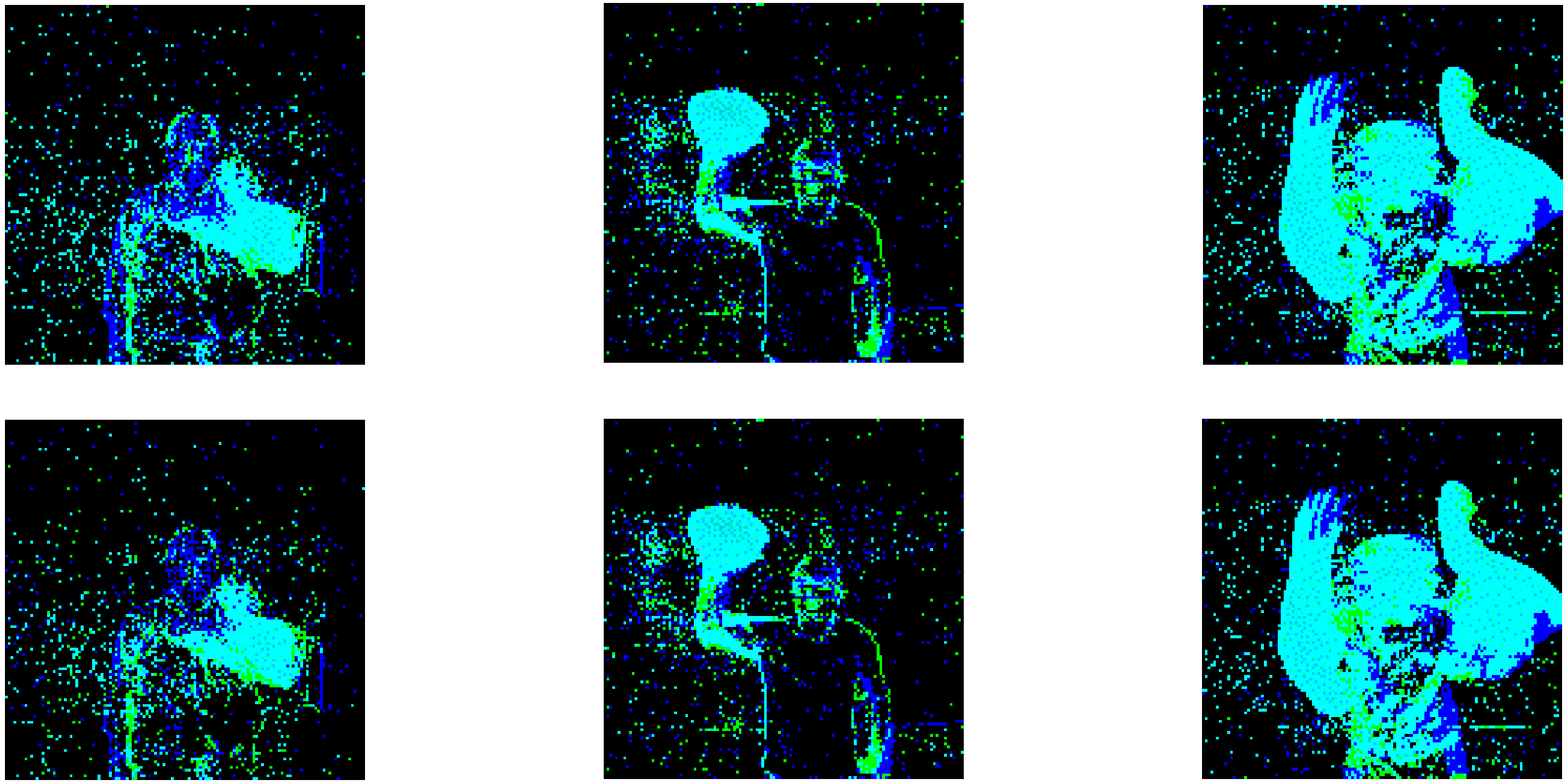
[1] Abad, Gorka et al. "Sneaky Spikes: Uncovering Stealthy Backdoor Attacks in Spiking Neural Networks with Neuromorphic Data" in NDSS 2024.

MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

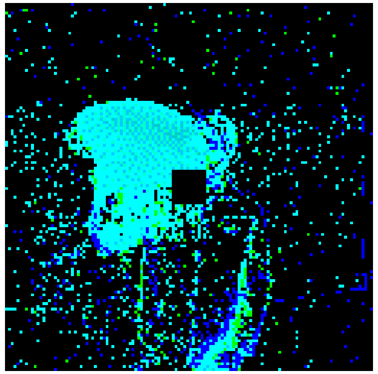
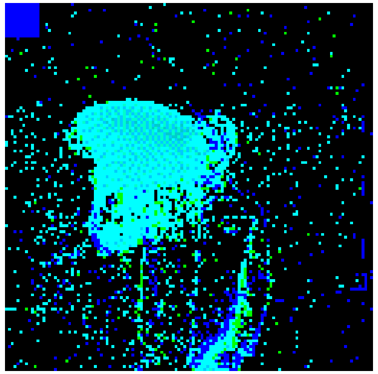
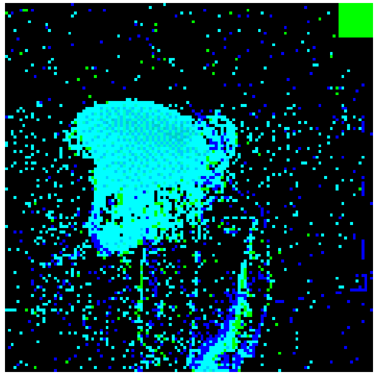
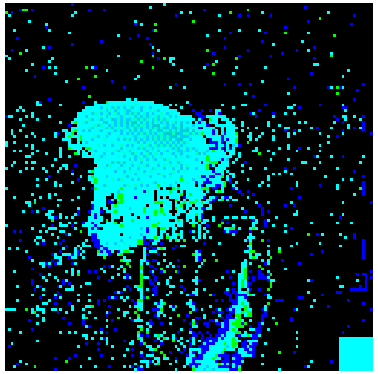
Sneaky Spikes



COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Static triggers

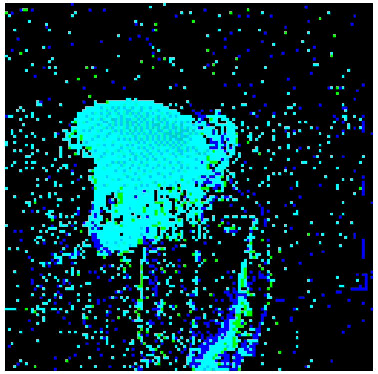





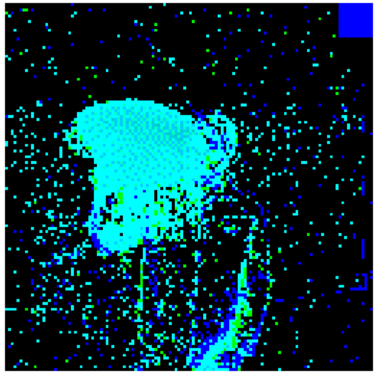
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

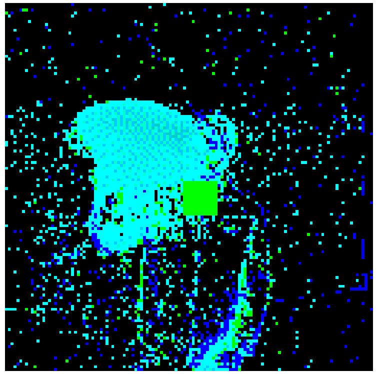
COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

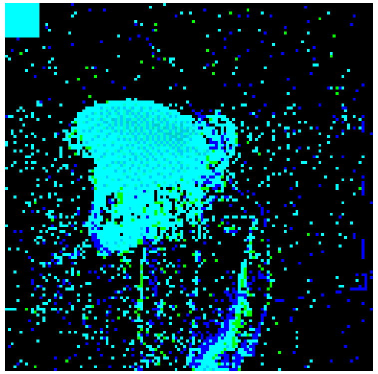
Type of attacks.

Moving triggers





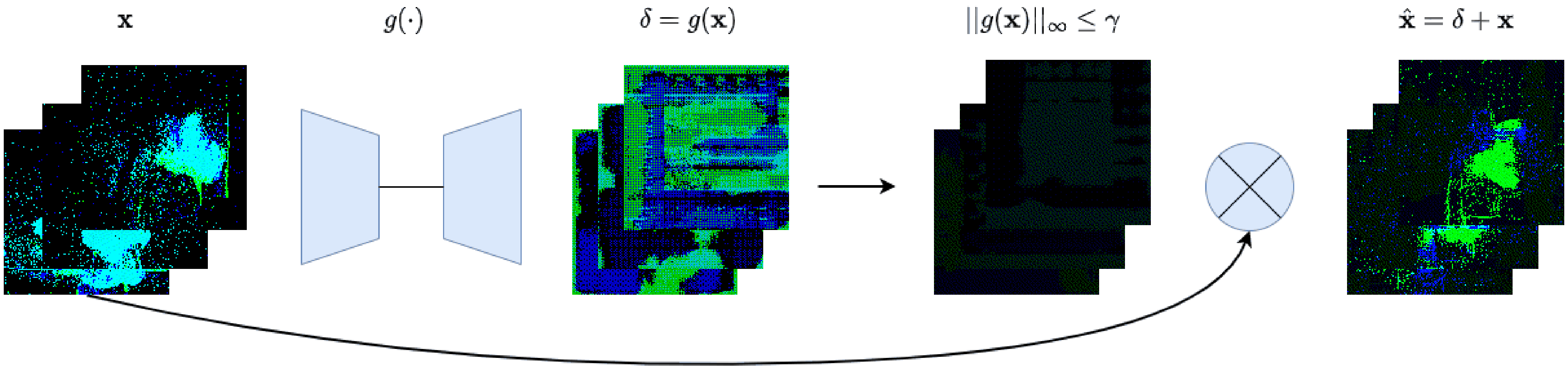




COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Dynamic triggers



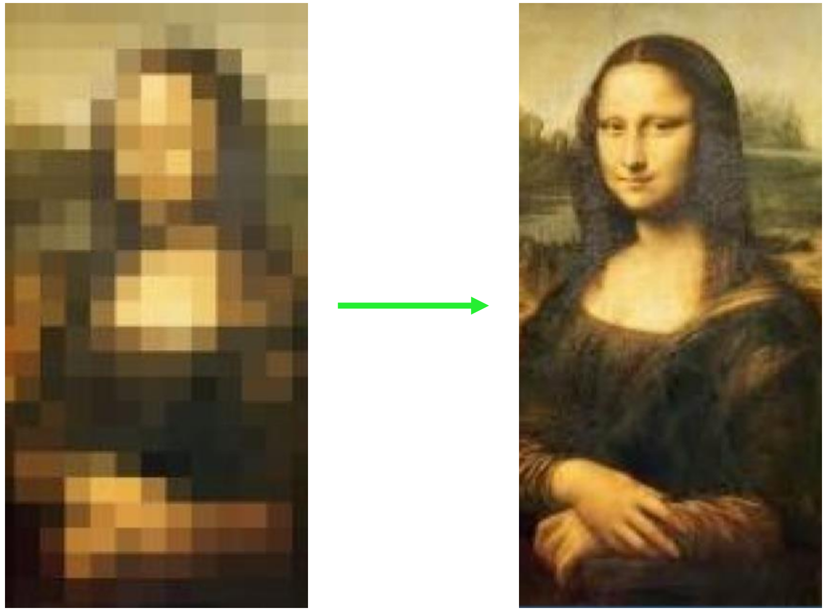
ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

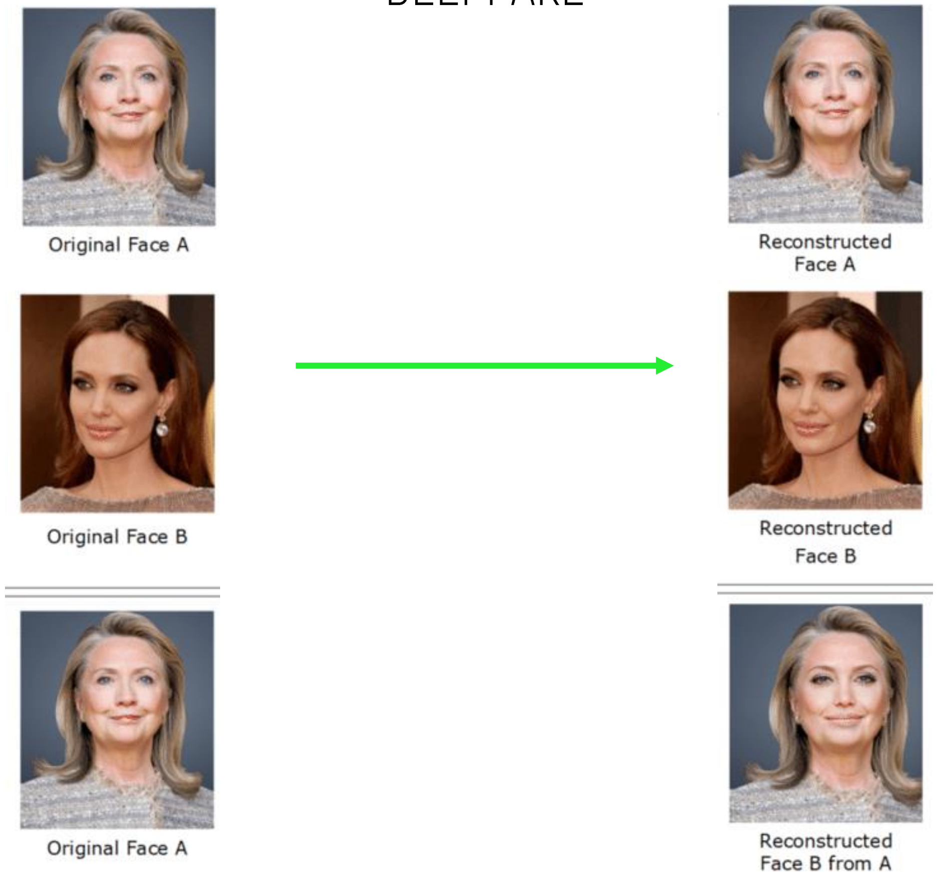
Type of attacks.

Dynamic triggers

DENOISING



DEEPFAKE

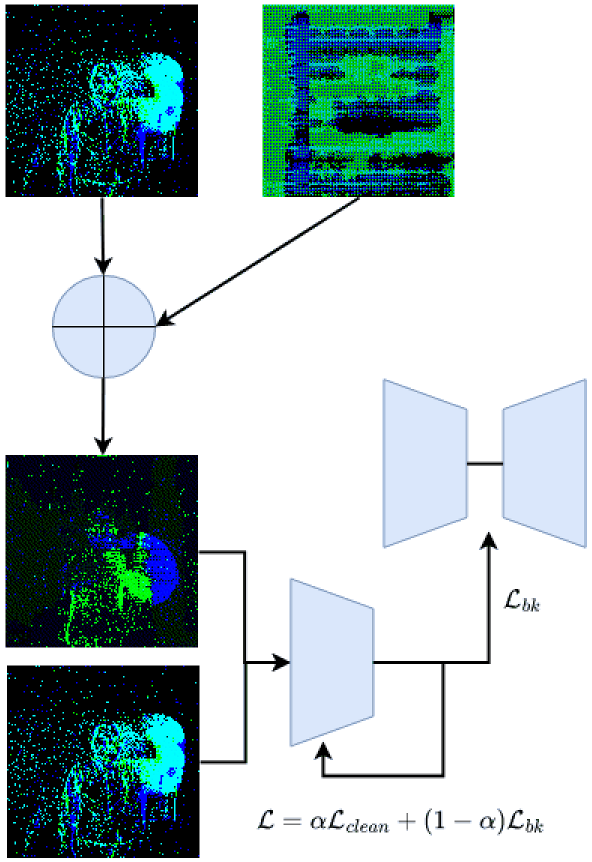


ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Dynamic triggers



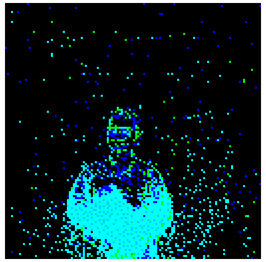
- **Simultaneously** train the classifier and the autoencoder
- The autoencoder is trained to **maximize** the **backdoor** accuracy
- The classifier is trained on **clean** and **backdoor** data
- The *backdoor effect* is controlled by α

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

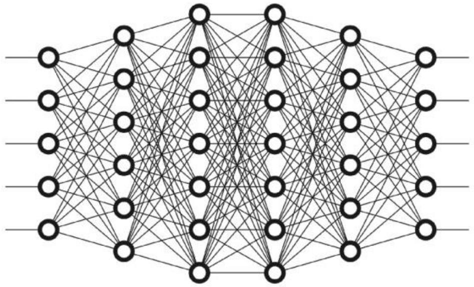
Type of attacks.

Dynamic triggers

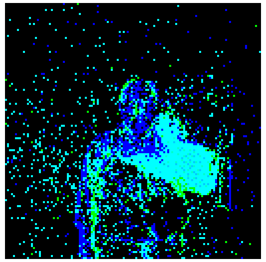
ARM ROLL



ARM ROLL

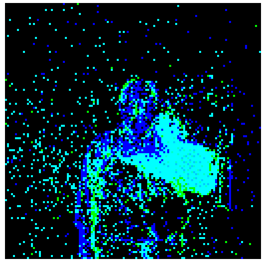


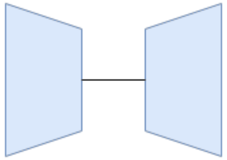
LEFT HAND CLOCKWISE

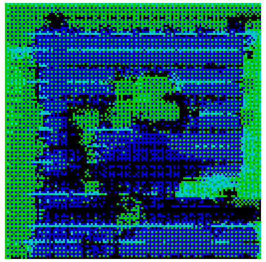


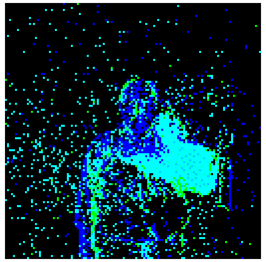
LEFT HAND CLOCKWISE

ARM ROLL









ARM ROLL

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

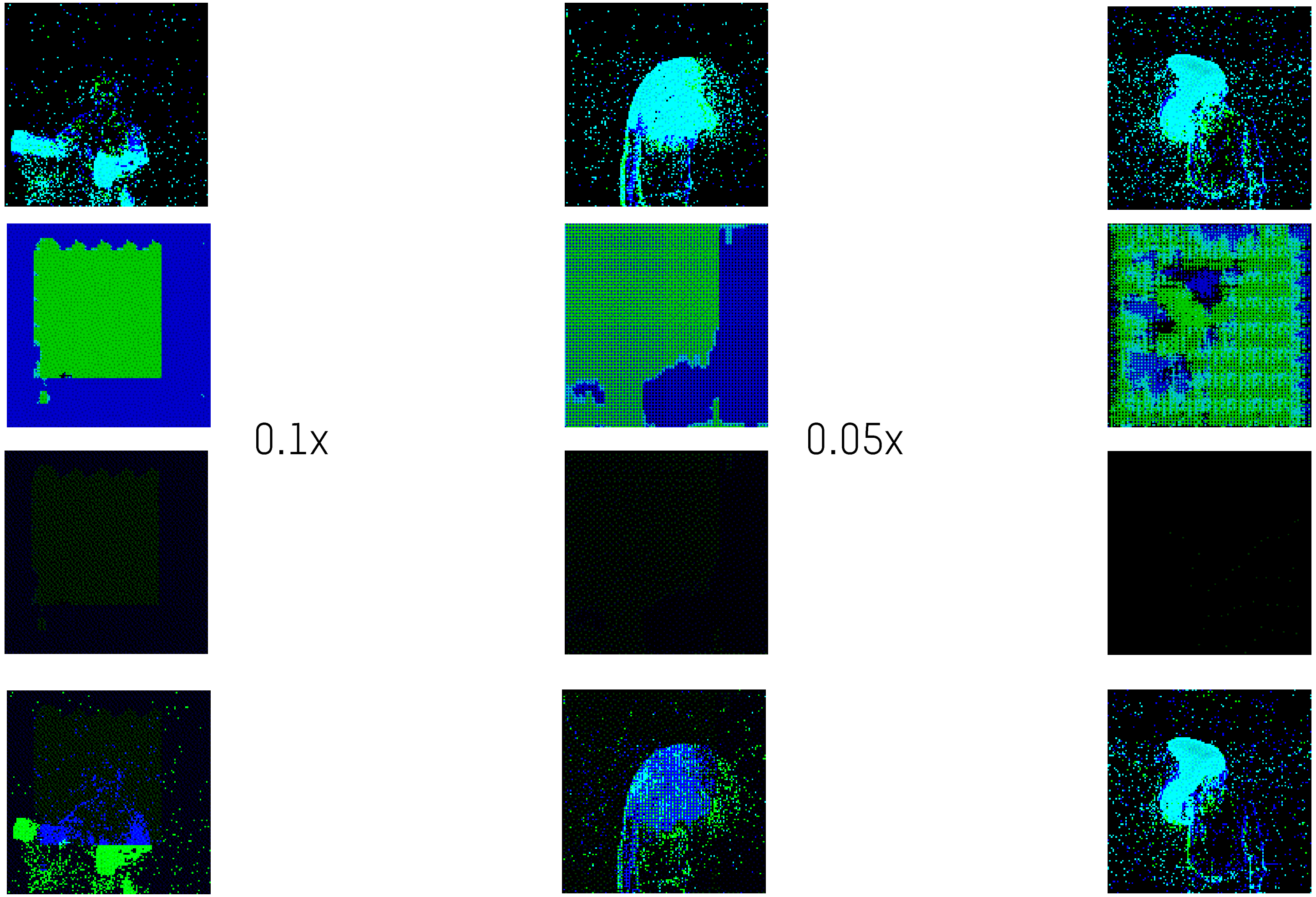
Dynamic triggers

CLEAN

NOISE

PROJECTED NOISE

BACKDOOR IMAGE



0.1x

0.05x

0.01x

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

Type of attacks.

Backdoor attacks

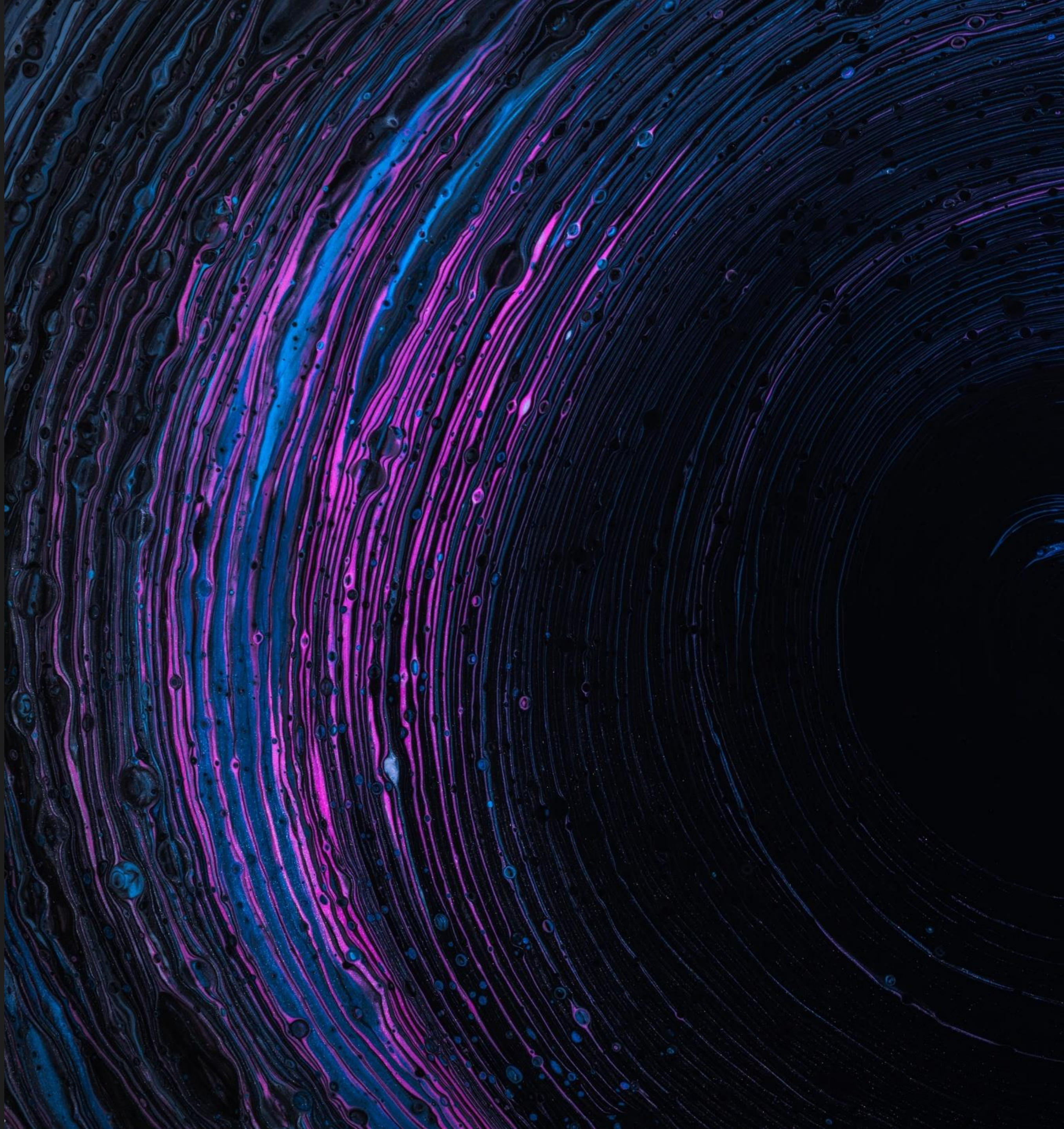
DEMO

https://t.ly/LP67x

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

6.

Challenges and future work.



Challenges and future directions.

- Interpretability and explainability
- Security by design
- Defenses
- Fast growing domain
- Ethical considerations
- Legal regulations
- New applications and use cases
- Formal methods

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

COPYRIGHT, CONFIDENTIAL AND PROPRIETARY. ALL RIGHTS RESERVED - PROPERTY OF IKERLAN, S. COOP. This information carrier and the information it contains are the property of Ikerlan, S. Coop. Any reproduction, disclosure or use of either is prohibited without the prior written consent of Ikerlan, S. Coop. Ikerlan, S. Coop. reserves worldwide all rights also in the case of industrial property rights being granted. The same provisions apply to any oral communications related thereto accordingly

ESKERRIK ASKO!

Gorka Abad
gabad@ikerlan.es
gorkaabad.github.io

www.ikerlan.es
P.ª José María Arizmendiarrieta, 2 – 20500 Arrasate-Mondragón.

ikerlan
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE