What Connects a Black Hole Photo and Explainable AI?

NASA Space Apps Challenge 2026 · Poznań

Jacek Karolczak
jkarolczak@cs.put.poznan.pl · Poznan University of Technology · Graylight Imaging

Ariane 5, Flight 501

Kourou, 4.07.1996

- **Payload**: Four magnetosphere research satellites.
- **Mission value**: Approximately $370.



Image source: DLR_de at https://www.flickr.com/photos/48213136@N06/8958839420

Ariane 5, Flight 501

Kourou, 4.07.1996

- 37 seconds after the take off the rocket veers sharply off course.
- Aerodynamic forces tear vehicle apart.
- Self-destruct sequence triggers automatically.



Image source: ESA at http://www.capcomespace.net/dossiers/espace_europeen/ariane/ariane5/AR501/V88%20explosion%2003.jpg

Ariane 5, Flight 501

Paris, **19.07.1996**

The internal SRI software exception was caused during execution of a data ***conversion from 64-bit floating point to 16-bit signed integer value****. The floating point number which was converted had a value greater than what could be represented by a 16-bit signed integer. This resulted in an Operand Error. The data conversion instructions (in Ada code) were not protected from causing an Operand Error, although other conversions of comparable variables in the same place in the code were protected.*

```
P_M_DERIVE(T_ALG.E_BH) := UC_16S_EN_16NS (TDB.T_ENTIER_16S

    ((1.0/C_M_LSB_BH) * G_M_INFO_DERIVE(T_ALG.E_BH)));
```


Source: https://www-users.cse.umn.edu/~arnold/disasters/ariane5rep.html

Ariane 5, Flight 501

Could this have been prevented?

- Partial Protection: Only 4 of 7 conversions protected.
- Backup Failover: SRI 2 replaced failed SRI 1
- Identical Code: Both systems ran identical buggy code.



Source: https://www-users.cse.umn.edu/~arnold/disasters/ariane5rep.html

The bug was easy to spot — we could read the code

$<<10^6$

Lines of Ariane 5 Code
Auditable by human engineers

6.7×10^9

Parameters in Llama 2 7B
Even a relatively small LLM

Although we could read the code, we did not prevent the catastrophe.

Parameters Are Not Lines of Code

Ariane 5 Source Code

Engineers write and read logic and trace failures.

```
P_M_DERIVE(T_ALG.E_BH) := UC_16S_EN_16NS (TDB.T_ENTIER_16S

    ((1.0/C_M_LSB_BH) * G_M_INFO_DERIVE(T_ALG.E_BH)));
```

Parametry Llama 2 7B

Raw floating-point tensors incomprehensible to humans.

```
tensor([[ 0.0778,  0.0077, -0.0156, ...,  0.0004, -0.0145, -0.0114],

        [ 0.0006, -0.0073,  0.0160, ..., -0.0171,  0.0028, -0.0408],

       ...,

       [-0.0291,  0.0021, -0.0086, ..., -0.0072, -0.0190, -0.0087]])
```

Why Do Neural Networks Work?

- Billions of learned parameters.
- Thousands of tensor multiplications.
- Activation functions across layers.
- Trained via gradient descent.

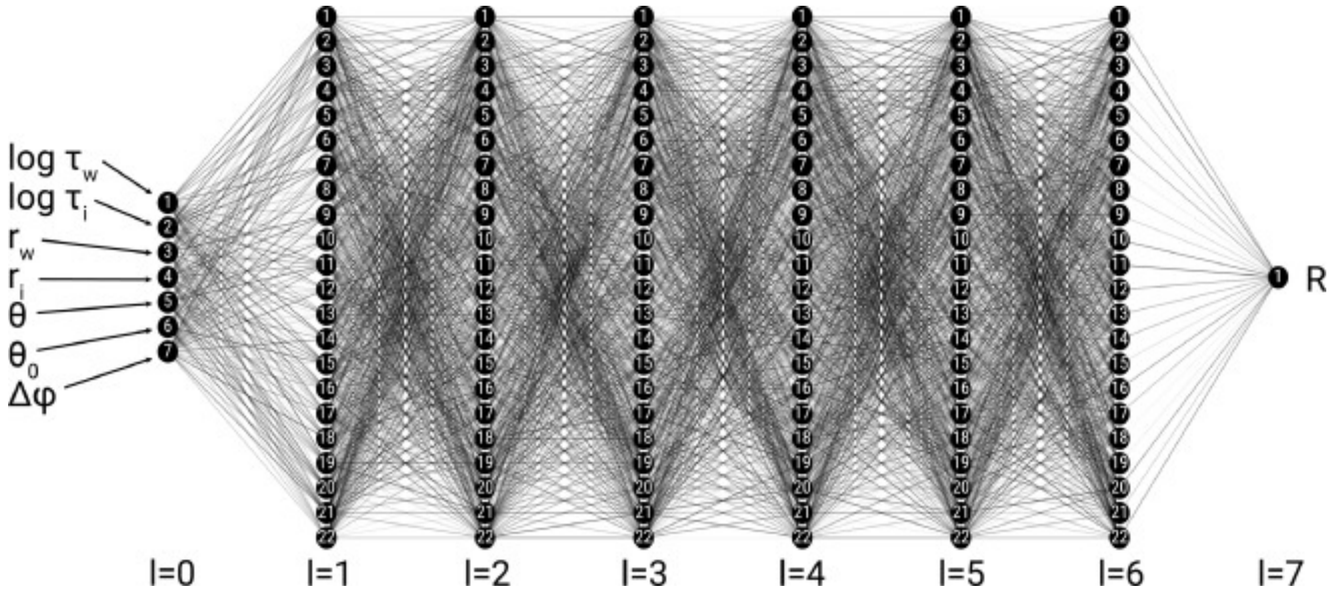
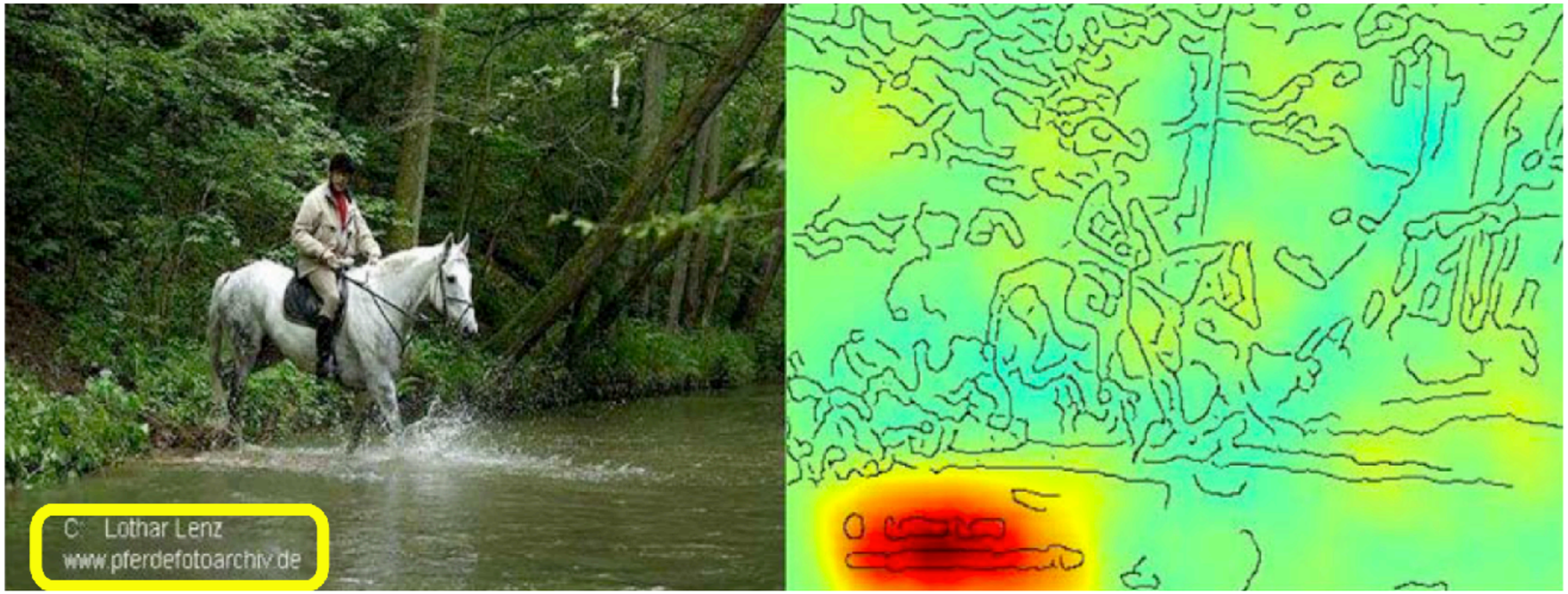


Image source: Leonhard Scheck (2026). *A neural network based forward operator for visible satellite images and its adjoint*. Journal of Quantitative Spectroscopy and Radiative Transfer.

Why do neural networks make good predictions?

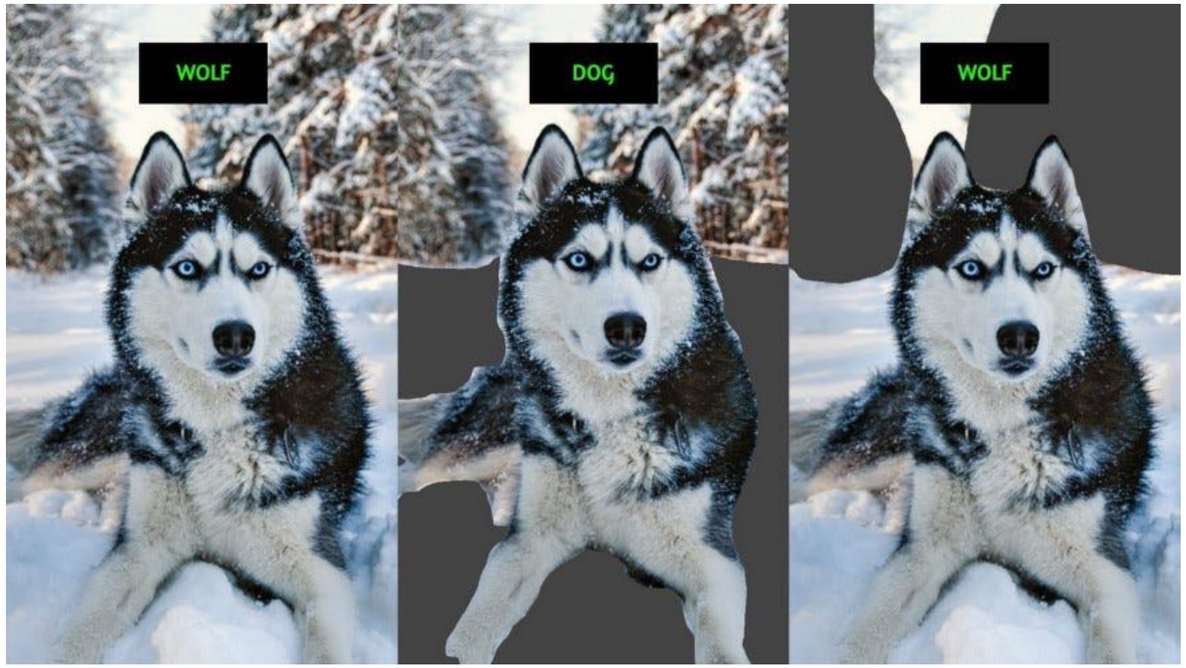
Horse Detection



Detects copyright text instead of horses.

Source: Lapuschkin, S. et al. (2019). *Unmasking Clever Hans predictors and assessing what machines really learn*. Nature Communications.

Wolf vs Husky?

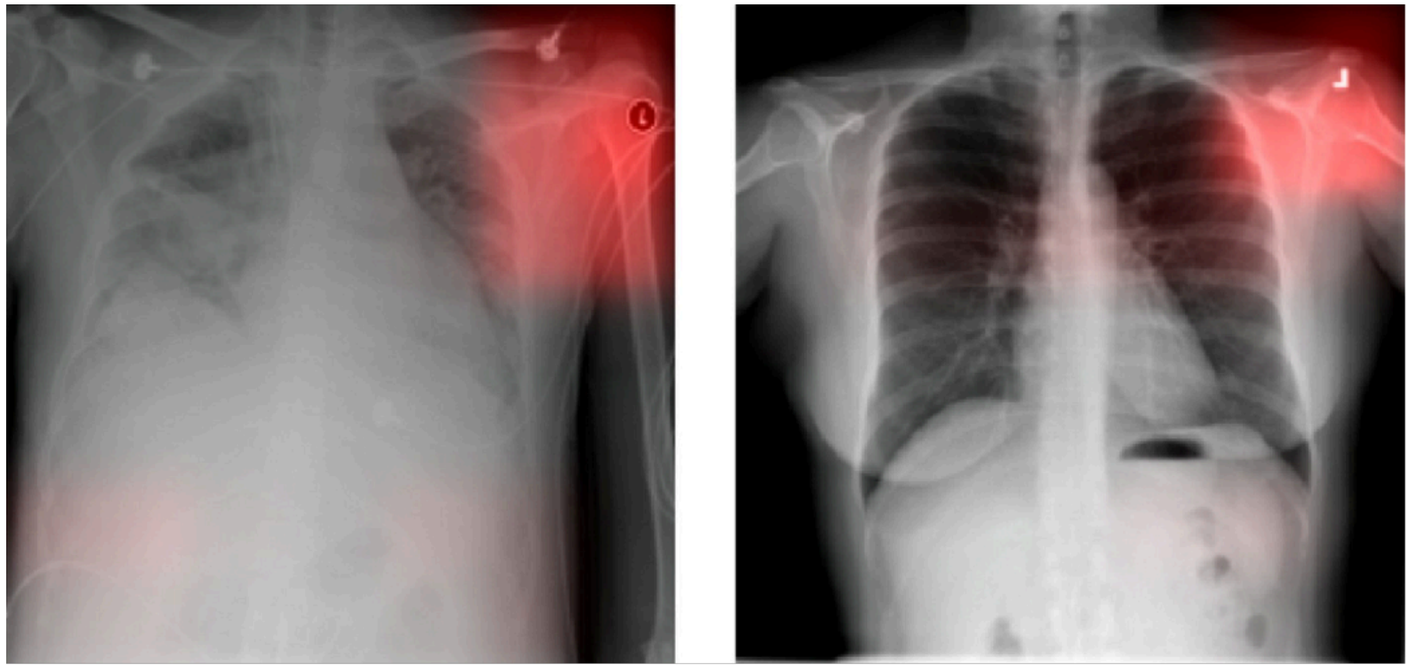


Classifies snow rather than animal features.

Source: Ribeiro, M. T., Singh, S., Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. KDD.

Right Decision, Wrong Reason · 3/4

COVID-19 X-Rays



Relies on corner tags over lung pathology.

Source: DeGrave, A. J., Janizek, J. D., Lee, S.-I. (2021). *AI for radiographic COVID-19 detection selects shortcuts over signal*. Nature Machine Intelligence.

CLIP (Vision-Language Model)



Matches printed “dollar” on parking meter text to “doll”.

Source: Baniecki, H. et al. (2025). *Explaining Similarity in Vision-Language Encoders with Weighted Banzhaf Interactions*. NeurIPS.

What Connects a Black Hole Photo and Explainable AI?

First Image of a Black Hole



Image source: Event Horizon Telescope Collaboration

First Image of a Black Hole

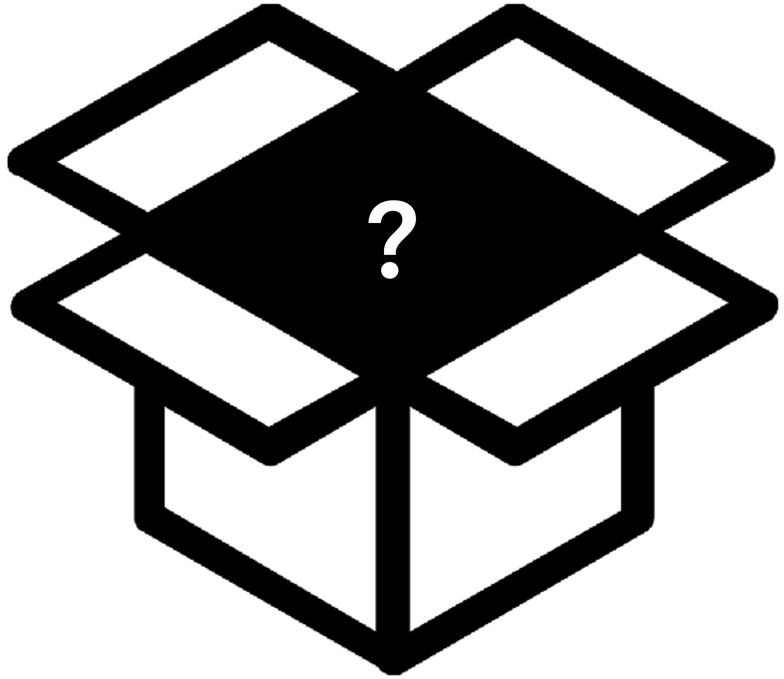
Messier 87*, Event Horizon Telescope, 2019

- Black holes emit no light.
- **Photography of a black hole is physically impossible.**
- We observe matter orbiting its shadow.
- Dozens of radio telescopes combined data.

Deep Neural Networks are sometimes called “black boxes”

Black Boxes





Black hole: Extreme matter density warps spacetime, preventing us from observing its interior.

Black box: The density of billions of parameters prevents us from understanding it, even though we can look inside.

Black hole photo: Measuring the surrounding light reveals the shadow inside.

Explainable AI: Probing inputs and measuring output shifts reveals the logic inside.

How do you explain a “black box” with only a couple of hours left in the hackathon?

Explainability in Python

Agnostic (including tabular data)

- LIME
- SHAP
- Dalex
- Alibi

Computer Vision

- Captum
- Grad-CAM
- Lucid
- iNNvestigate

Natural Language Processing

- TransformLens
- Inseq
- Saliency
- Language Interpretability Tool (LIT)

Local Interpretable Model-agnostic Explanations

Explaining Any Black Box

1. **Perturbation:** Randomly alter input instance to create synthetic, similar instances.
2. **Black Box Query**: Collect model predictions for every generated similar instance.
3. **Local Surrogate**: Fit linear model to measure feature importance.

Toy Example

Nuclei Segmentation

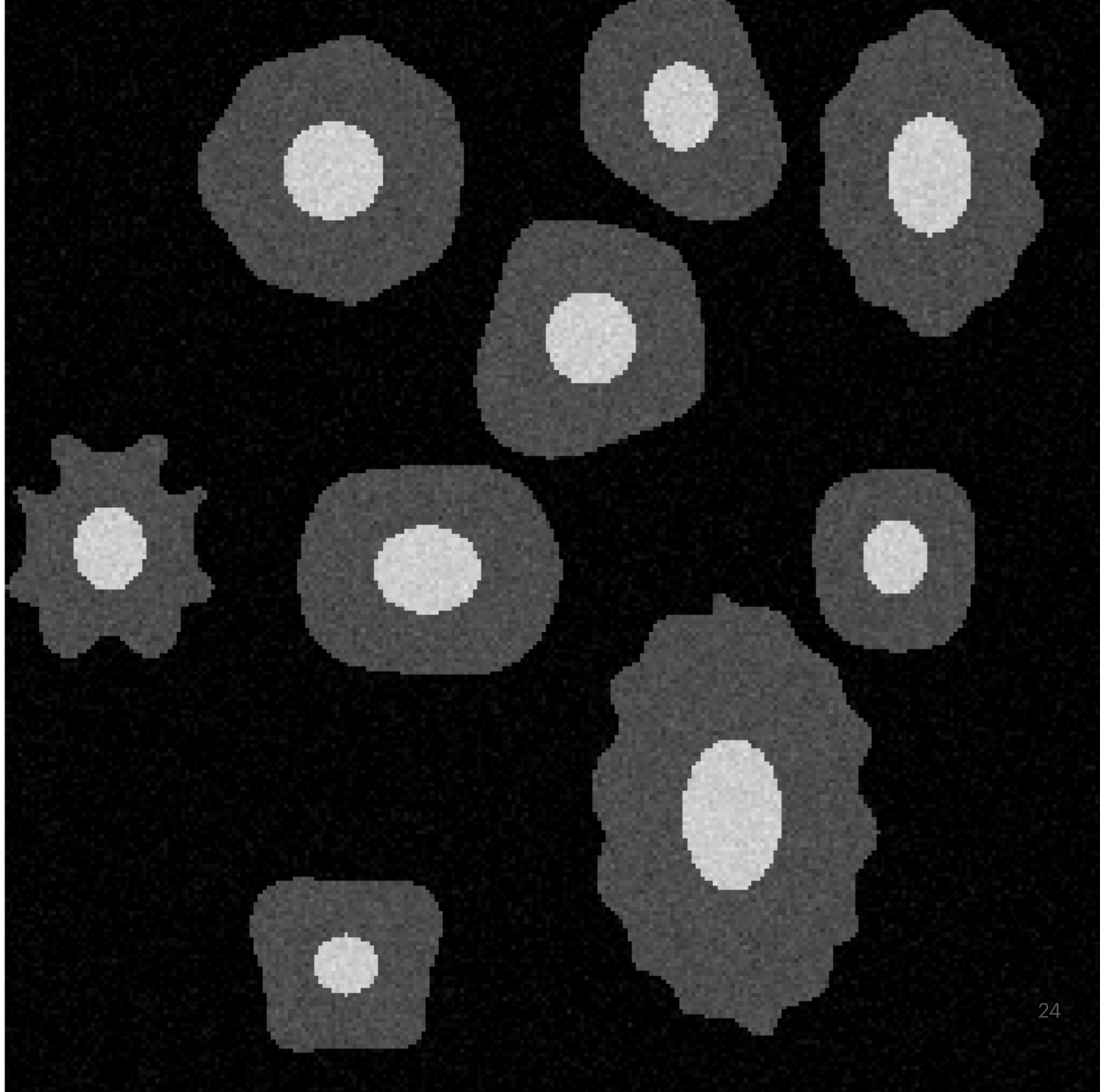


Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Superpixels

Step 1

Group pixels by color and spatial proximity.

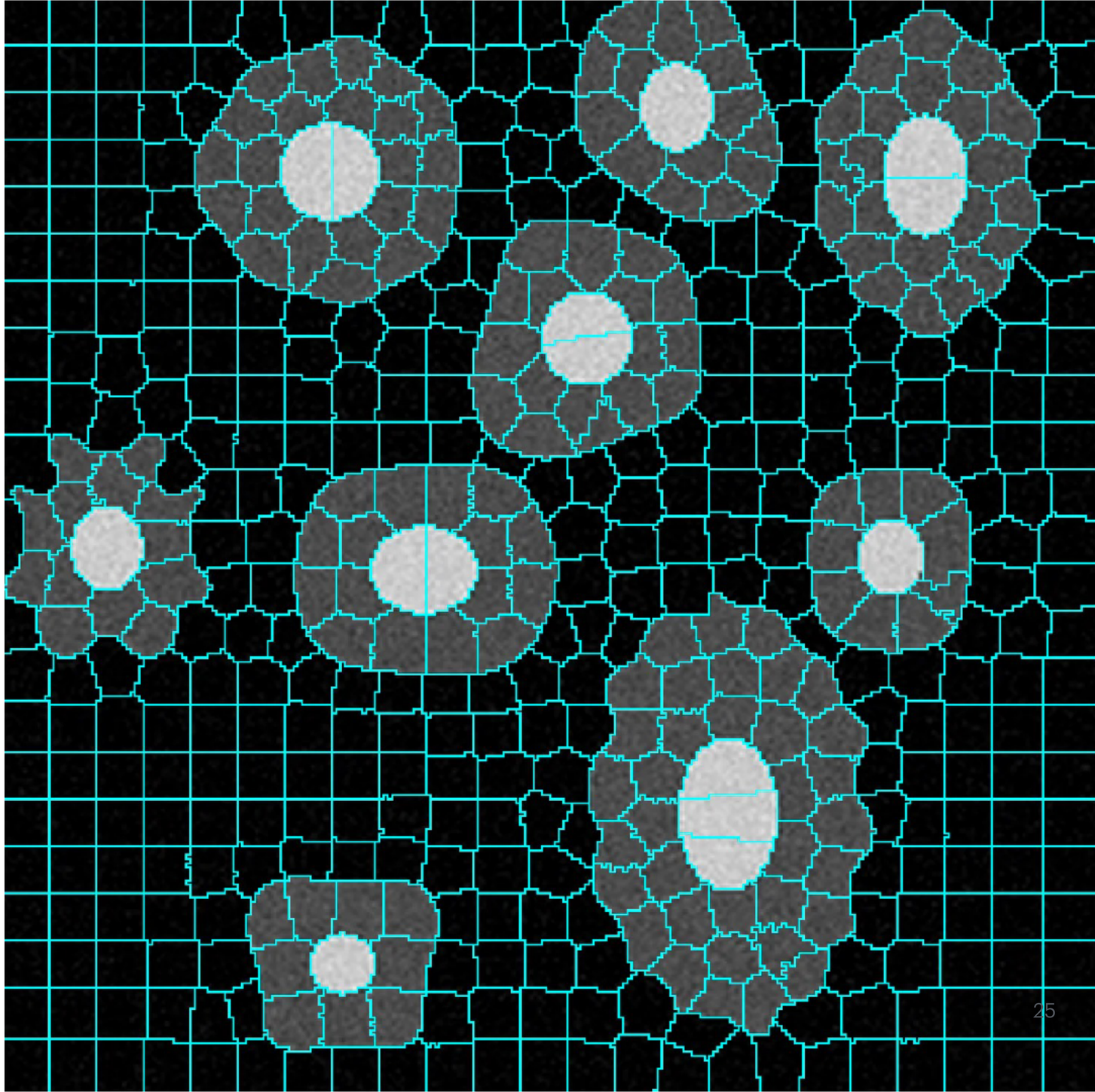


Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Perturbation

Step 2

Randomly mask out superpixels to create thousands of slightly perturbed versions of the original image.

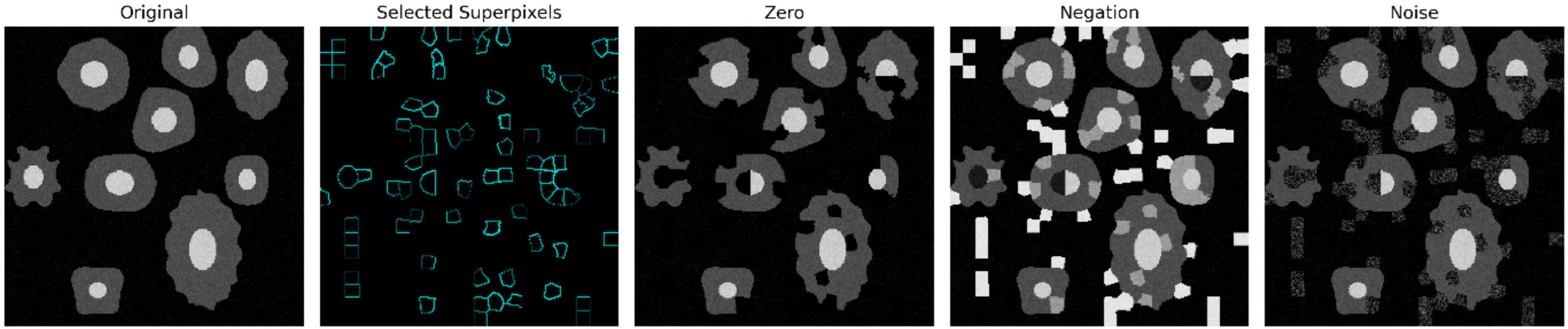
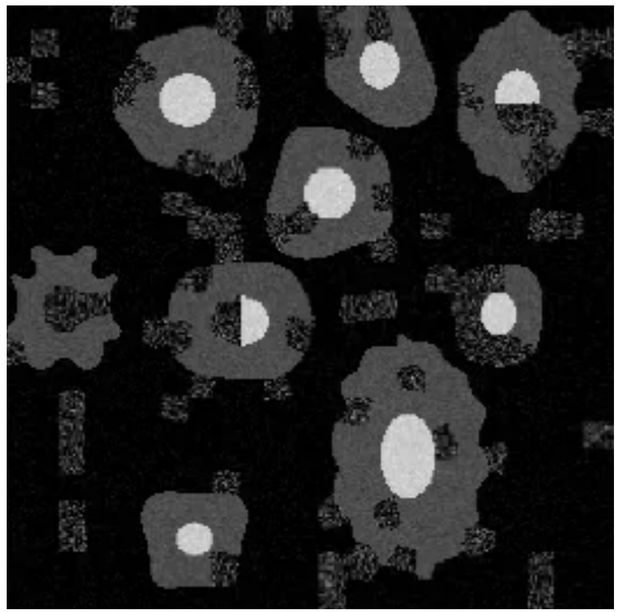


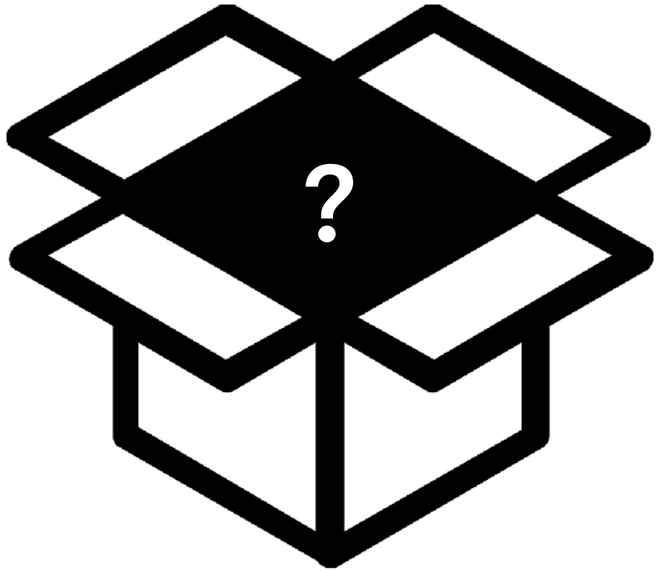
Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Black Box Query

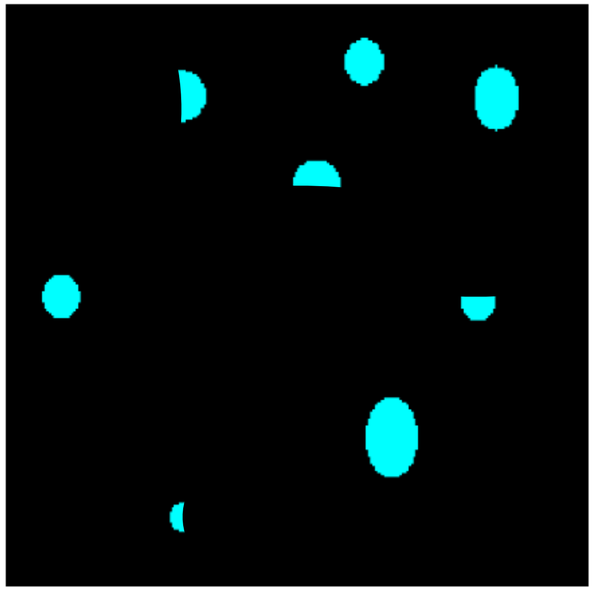
Krok 3

Perturbed Instance





Model Prediction



```
[0, 1, 0, 0, ..., 1]
```

Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Scoring

Step 4

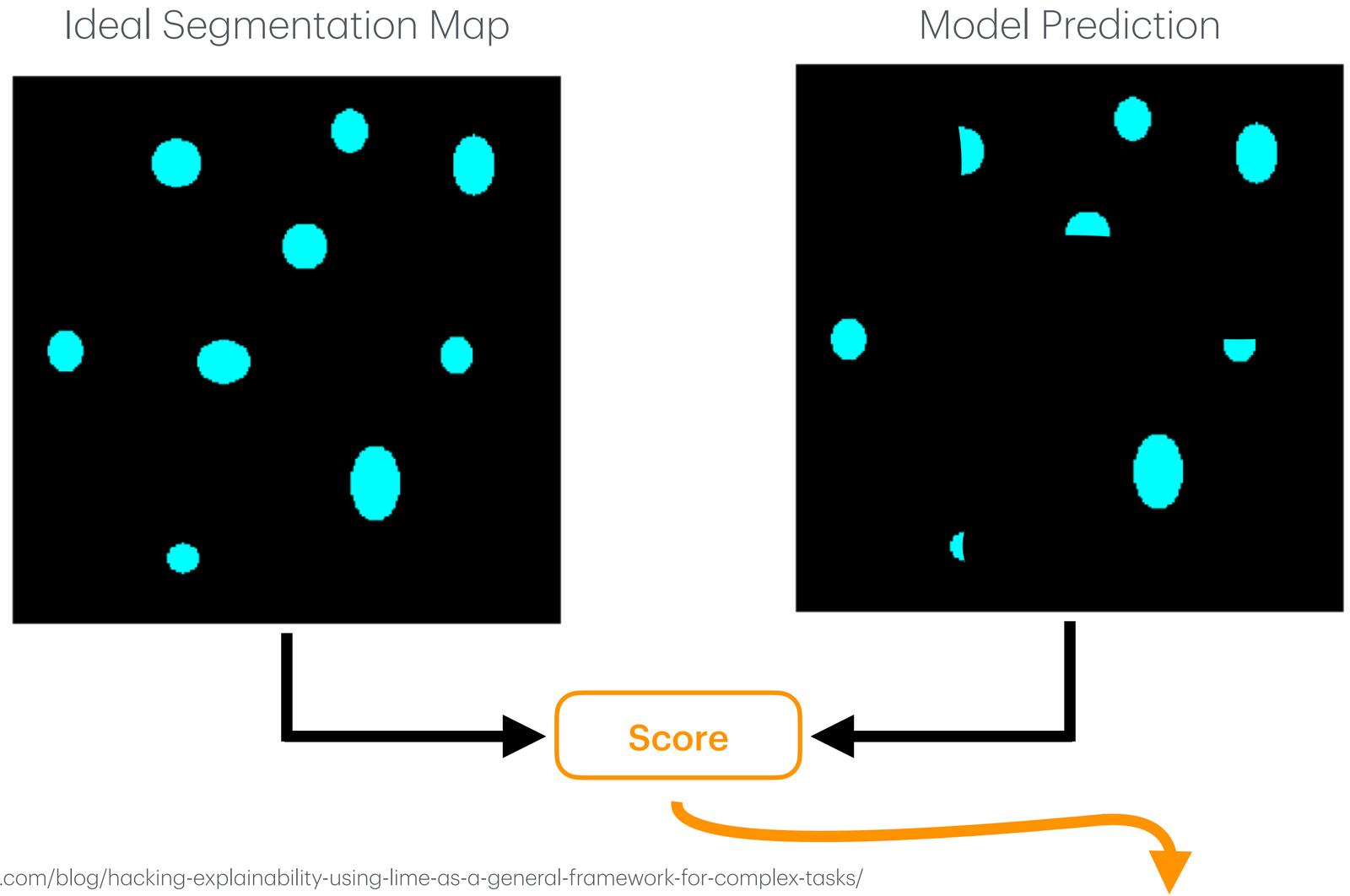


Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Surrogate Linear Model

Step 5



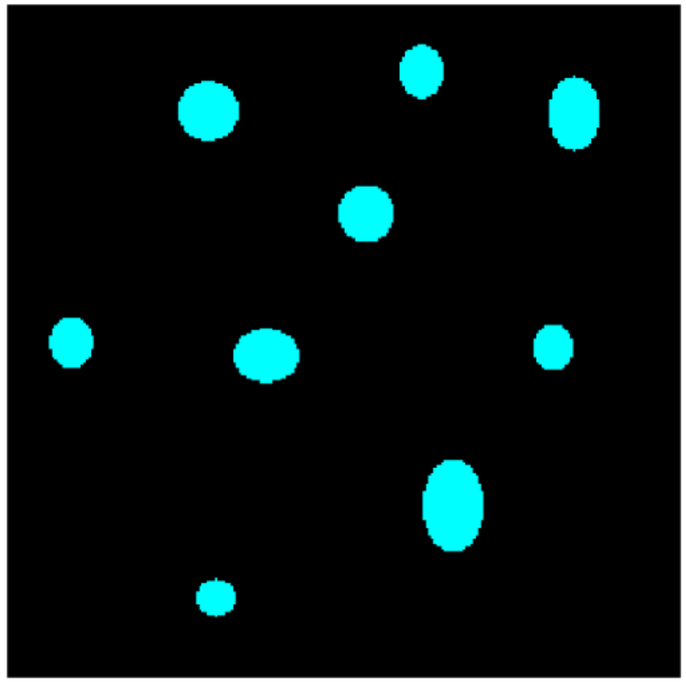
$$b + wX = Y$$

Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Importance scores

Step 6

Ideal Segmentation Map



Explanation

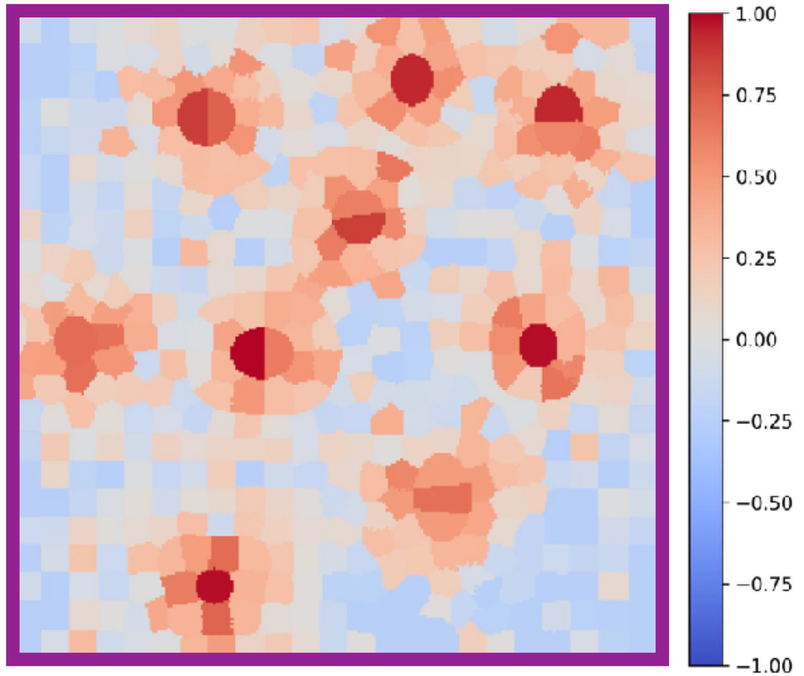
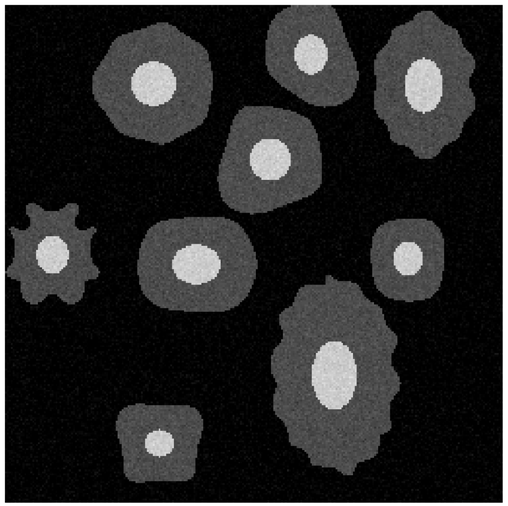


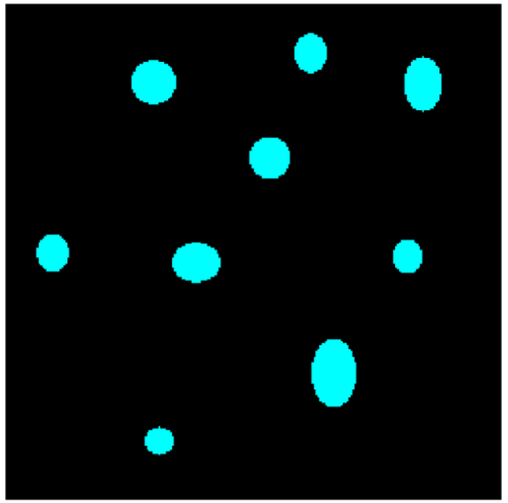
Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Which model is better?

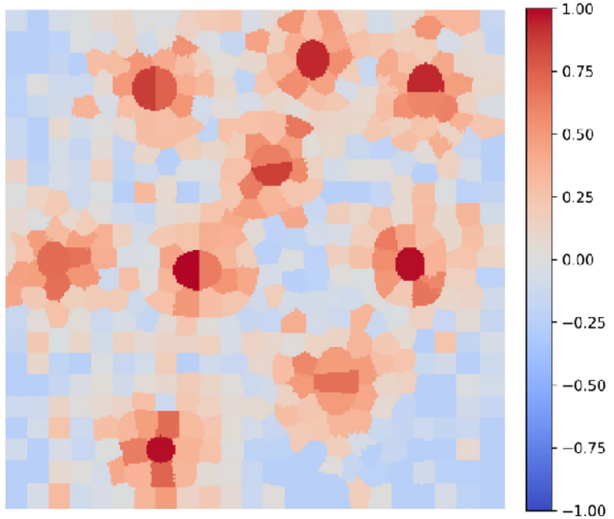
Input Instance



Prediction
Models A i B



Explanation
Model A



Explanation
Model B

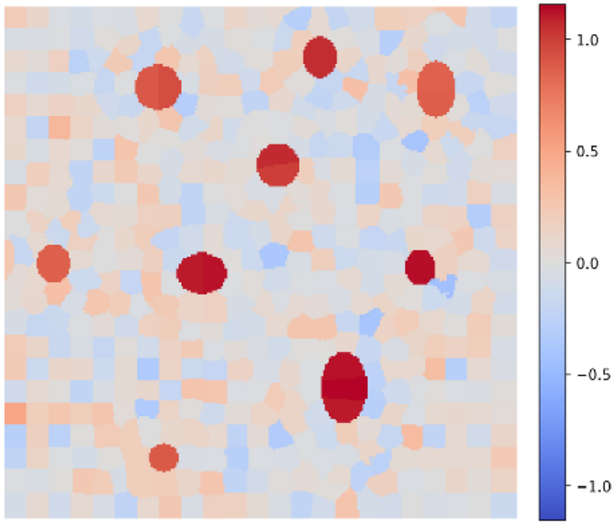


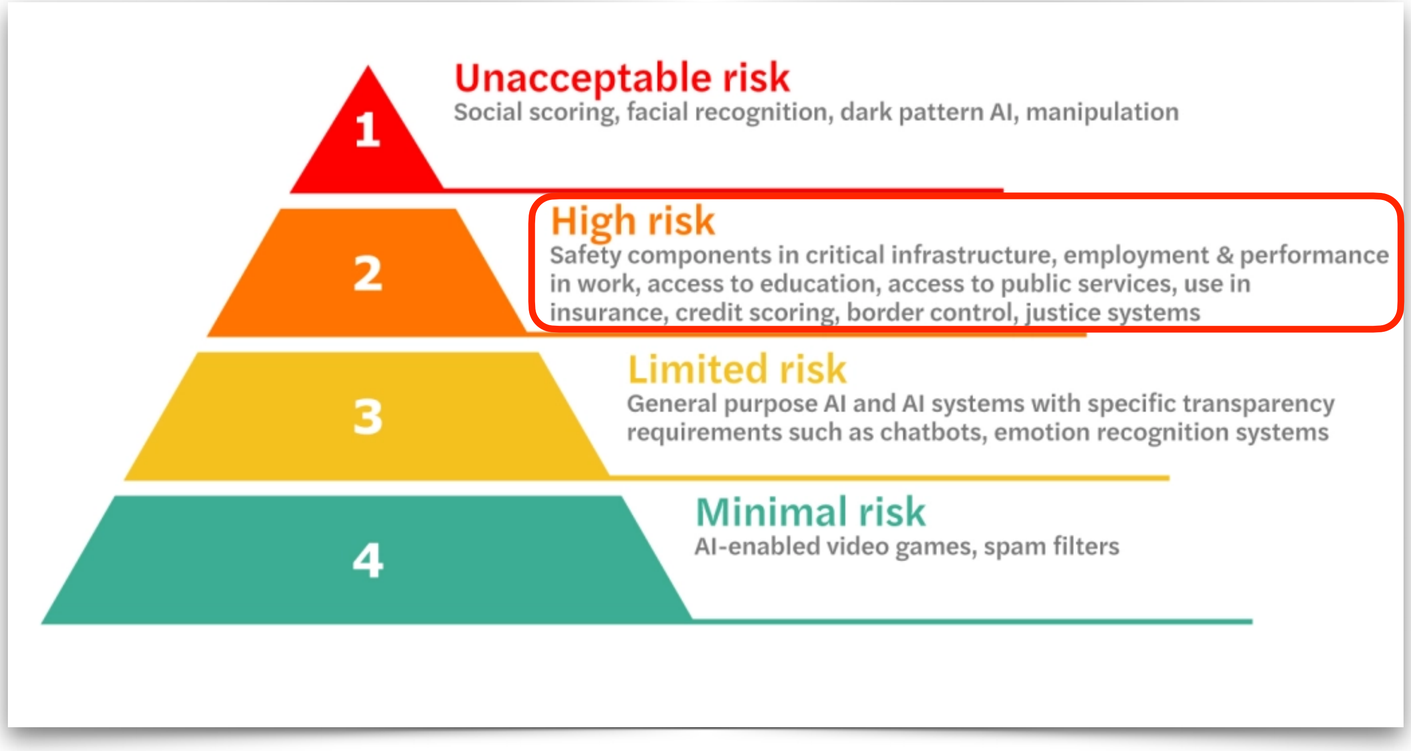
Image source: https://graylight-imaging.com/blog/hacking-explainability-using-lime-as-a-general-framework-for-complex-tasks/

Can you skip explaining your ML model?



AI Act

When Explainability Is Mandatory



Article 13: *Transparency and Provision of Information to Deployers*

1. *High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately [...]*

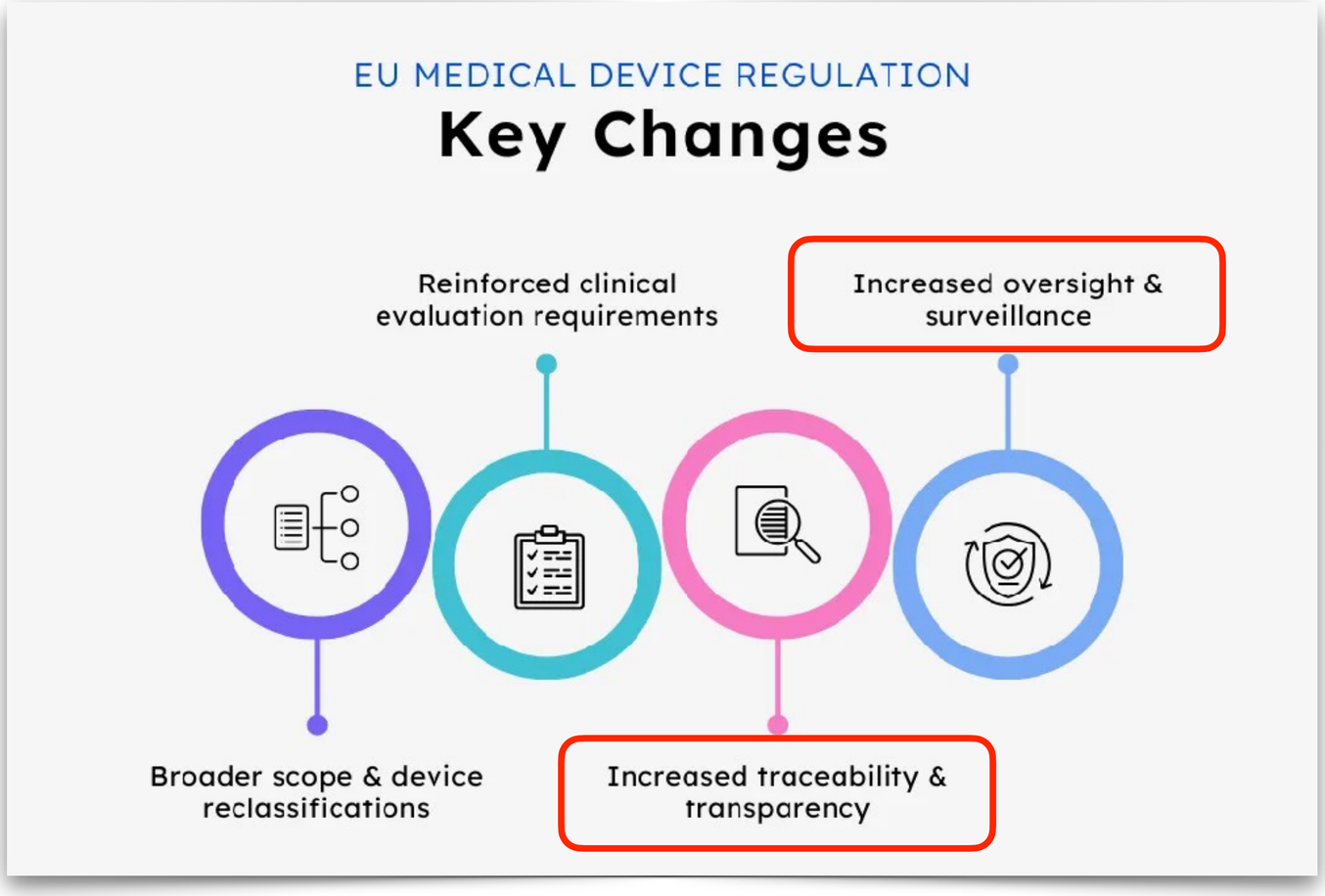
Article 14: *Human Oversight*

1. *High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use.*

Image source: https://www.forvismazars.com/ie/en/insights/news-opinions/eu-ai-act-different-risk-levels-of-ai-systems

Medical Device Regulation

When Explainability Is Mandatory



Article 2: *Definitions*

1. ***'medical device' means any*** *instrument, apparatus, appliance,* ***software****, implant, reagent, material or other article intended by the manufacturer to be used, alone or in combination, for human beings* ***for one or more*** *of the following* ***specific medical purposes****:*

Annex I, Chapter II: *Transparency and Provision of Information to Deployers*

22.2. ***Devices*** *for use by lay persons s****hall be designed*** *and manufactured in such a way as to* ***reduce*** *as far as possible* ***the risk of error*** *by the intended user in the handling of the device and, if applicable,* ***in the interpretation of the results.***

Image source: https://dai.ki/blog/eu-medical-device-regulation-key-changes-and-how-to-comply/

When Can You Skip Post-Hoc Explanations

- **Interpretable by design**: Use glass-box models (linear models, decision trees, GAMs, ProtoPNet) whose structure is the explanation.
- **Domain-driven features**: Encode domain knowledge or physical constraints directly into the model.
- **No black box, no approximation**

Thank You for Your Attention!

I would be happy to answer any questions

Jacek Karolczak
jkarolczak@cs.put.poznan.pl · Poznan University of Technology · Graylight Imaging