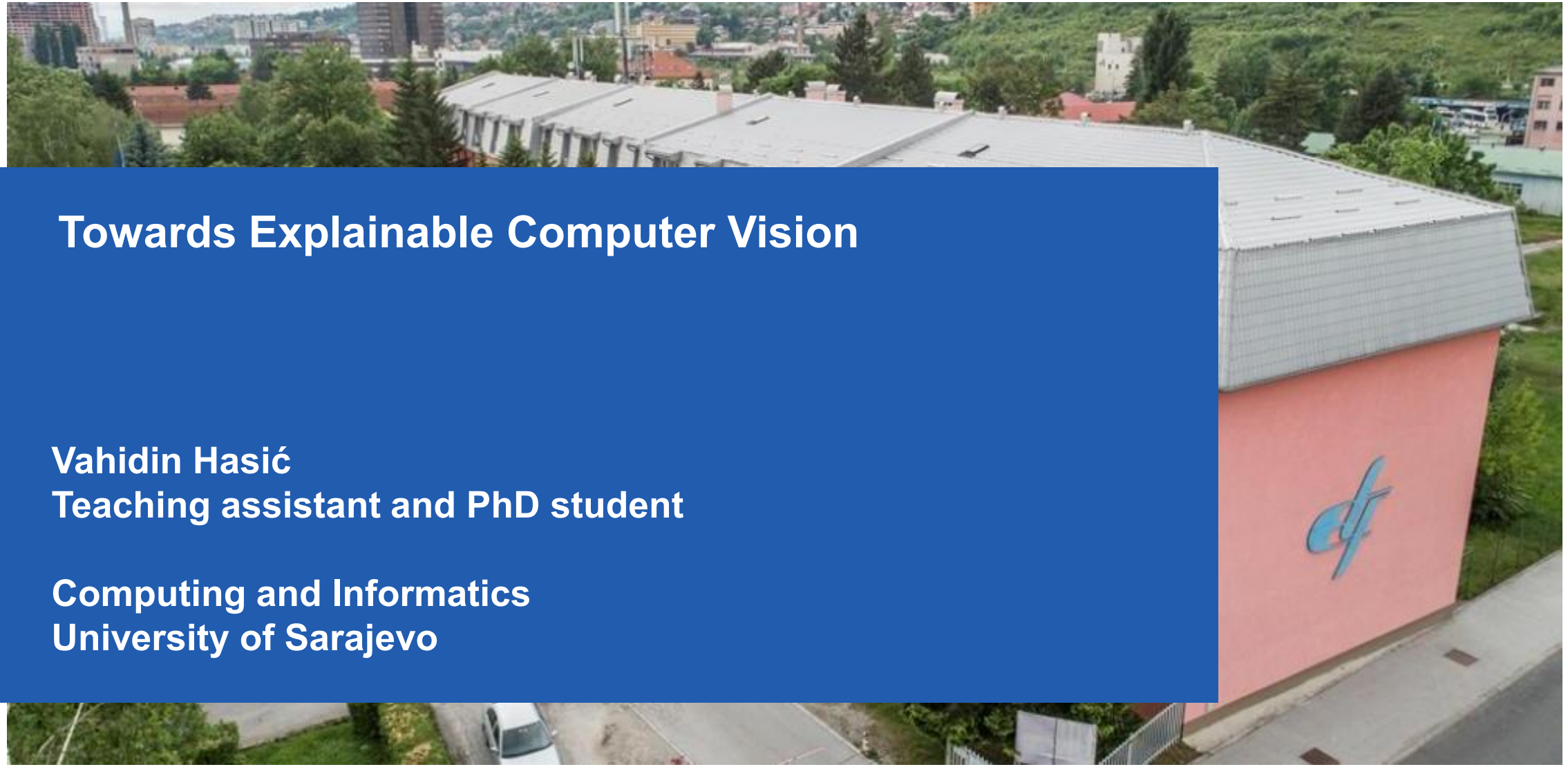
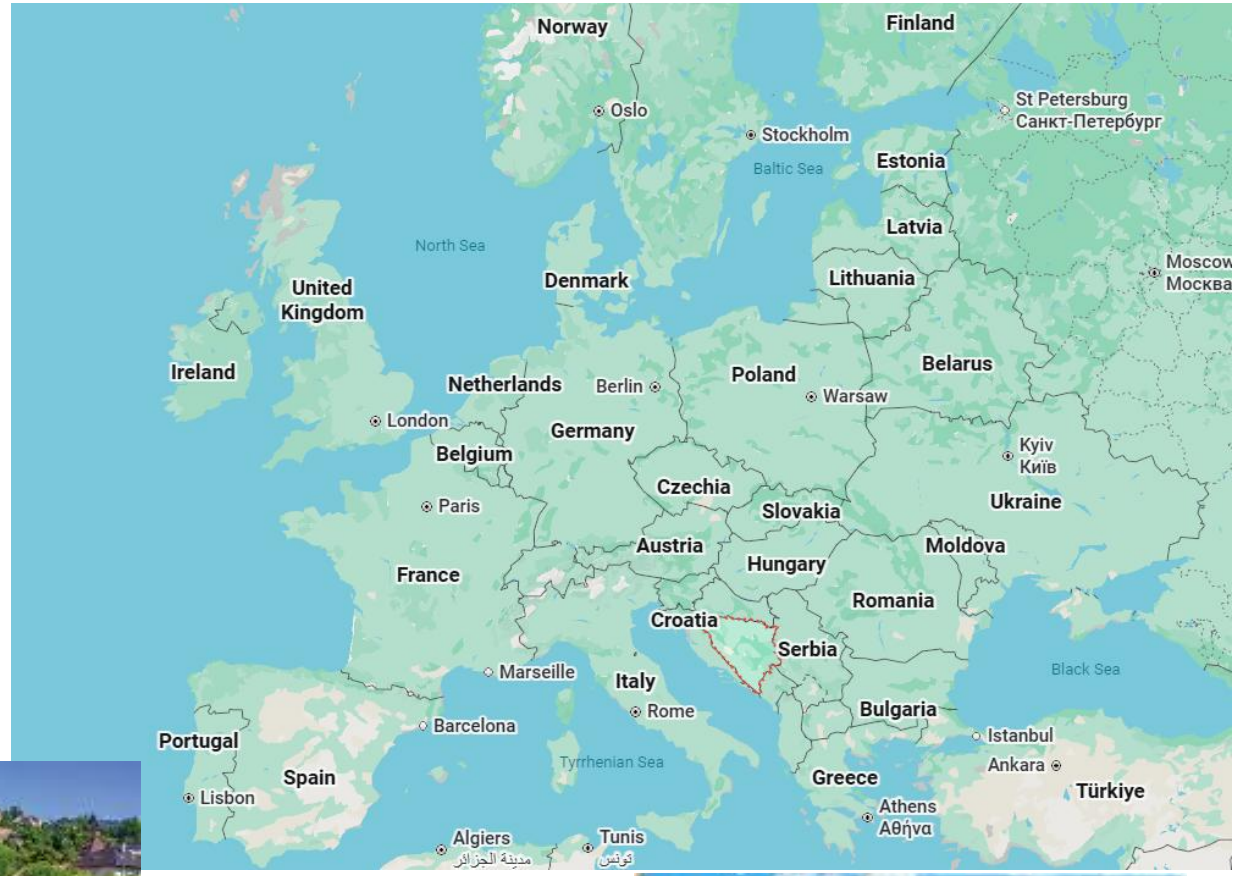


Towards Explainable Computer Vision

Vahidin Hasić
Teaching assistant and PhD student

Computing and Informatics
University of Sarajevo

Bosnia and Herzegovina



Sarajevo, 🚊 First Tram

- The first tram officially started operating on **January 1st, 1885**
- It was 3.1 kilometers long
- Sarajevo's first electric tram began service on **May 1, 1895**, replacing horse-drawn trams

Sarajevo, ❄ 1984 Winter Olympics

- **XIV Olympic Winter Games** in Sarajevo, held from **February 8-19, 1984**
- The popular mascot was a friendly wolf named **Vučko**
- The games featured **39 events across 6 sports** (10 disciplines): Alpine Skiing, Biathlon, Figure Skating, Ice Hockey, Ski Jumping,...

Sarajevo, B&H

National Geographic: Sarajevo is the best world destination for 2025 according to the choice of our readers

"It's official: Sarajevo is our reader's choice winner for 2025," they wrote on the Instagram page of National Geographic Travel.












Siege of Sarajevo

- Lasting from 5 April 1992 to 29 February 1996 (**1,425 days**), it was three times longer than the Battle of Stalingrad and more than a year longer than the siege of Leningrad, making it the **longest siege of a capital city in the history of modern warfare**
- **13,952 people were killed during the siege**, including 5,434 civilian; 1,601 children; 56,000 injured; including 15,000 children; 55,000–145,980 expelled

Human-Centered Artificial Intelligence

HCAI Lab Members

Senka Krivić, PhD

Assistant Professor at the Faculty of Electrical Engineering at the University of Sarajevo and Visiting Lecturer at King's College London

Formerly, Senka spent several years as a Research Associate at the Department of Informatics at King's College London as a member of the Planning and Reasoning Group and the Human-AI Teaming (HAT) lab. She obtained her Ph.D. degree in Computer Science from the University of Innsbruck at Intelligent and Interactive Systems Group. Dr. Krivic is a member of ELLIS network.

Amar Halilović

Ph.D. Candidate

Amar Halilovic is a PhD student and research and teaching assistant at Ulm University, Germany. His research is on Explainable Robotics, where his focus is on developing methods that make the decision-making processes in robot motion planning interpretable. Amar's academic journey includes a Bachelor's and Master's degrees in Electrical Engineering from the University of Sarajevo and a Master's degree in Computer Science from Mälardalen University, Sweden.

Vahidin Hasić

Ph.D. Candidate

Vahidin Hasić is a PhD student and teaching assistant at Faculty of Electrical Engineering, Department for Computer Science and Informatics, University of Sarajevo. His research is on AI explainability and Computer Vision. He holds a Master's Degree in Computer Science, Faculty of Electrical Engineering, University of Sarajevo.

Amina Mević

Ph.D. Candidate

Amina Mević is a PhD student and teaching assistant at the Department for Computer Science and Informatics, University of Sarajevo, Faculty of Electrical Engineering. Her research focus is on data-driven Virtual Metrology. She previously studied Physics at the University of Sarajevo, Faculty of Natural Science and Mathematics, and holds a Master's Degree in Theoretical Physics.

Mubina Kamberović

Ph.D. Candidate

Mubina Kamberović is a PhD student at the Faculty of Electrical Engineering, Department of Computer Science and Informatics, University of Sarajevo. Her main research area is AI-Assisted Learning. She also works as an MLOps/ML Engineer at Starbyte in Sarajevo. She completed her Bachelor's and Master's degrees in Computer Science at the Faculty of Electrical Engineering at the University of Sarajevo.

Ajla Karajko

Ph.D. Candidate

Ajla Karajko is a Ph.D. candidate specializing in AI governance, with a research focus on the intersection of artificial intelligence, policy, and regulatory frameworks to advance responsible AI governance. She holds a Bachelor's degree in Economics from Barnard College, Columbia University, and a Master's degree from the School of International and Public Affairs (SIPA) at Columbia University. With over a decade of experience in startups, business development, and marketing across diverse industries, Ajla has been instrumental in driving innovation and strategic growth.

Association for Advancing Science and Techology

2017

Year of establishment

130+

Active members and volunteers

30+

Successfully realized projects

Educational projects

EEML

STEM Youth Camp

Mali Muzej NAUKE

PSI
PROJEKAT STUDENTSKIH ISTRAŽIVANJA

- AI hackathon
- Science Night

Strateški projekti

BH akademski imenik

LANDSCAPE
BOSNIA AND HERZEGOVINA

insight

- Writing a strategy for the development of science, technology and innovation
- Establishment of a scientific institute

Eastern European Machine Learning 2025 (EEML 2025)

Doina Precup

McGill University

Google DeepMind

Razvan Pascanu

Google DeepMind

Viorica Patraucean

Google DeepMind

Suad Krilašević

ANNT

Harun Muhić

ANNT

Senka Krivić

University of Sarajevo

King's College

Vahidin Hasić

University of Sarajevo

Infineon

Petar Veličković

Google DeepMind

University of Cambridge

Hamza Merzić

Google DeepMind

University College London

Matko Bošnjak

Google DeepMind

Zlatan Ajanović

RWTH Aachen

ANNT

Ajla Karajko

ANNT

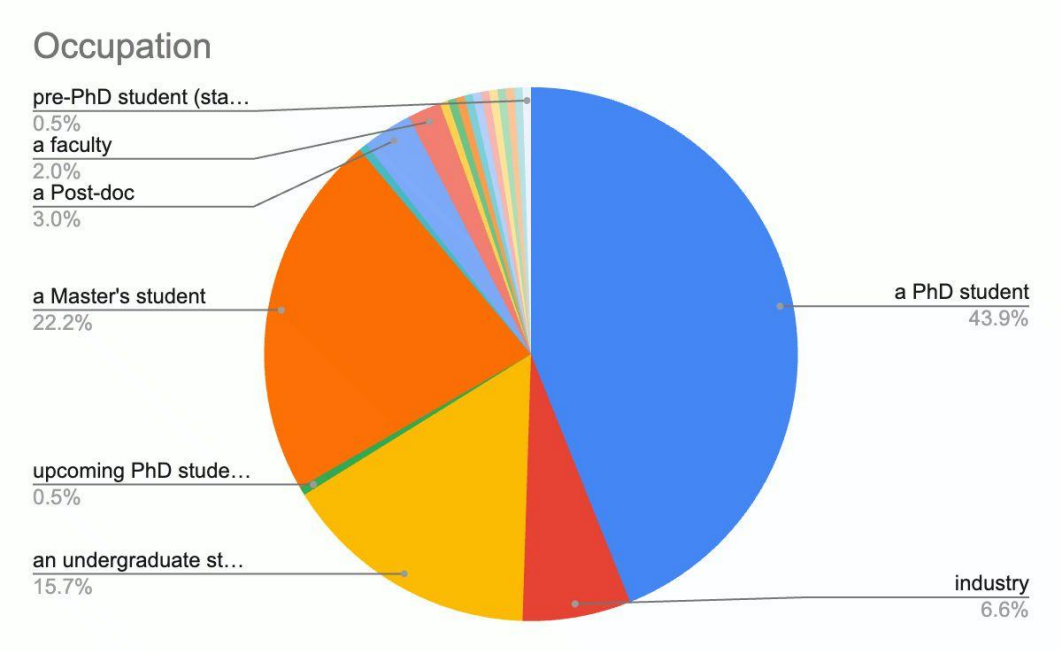
Džan Ahmed Jesenković

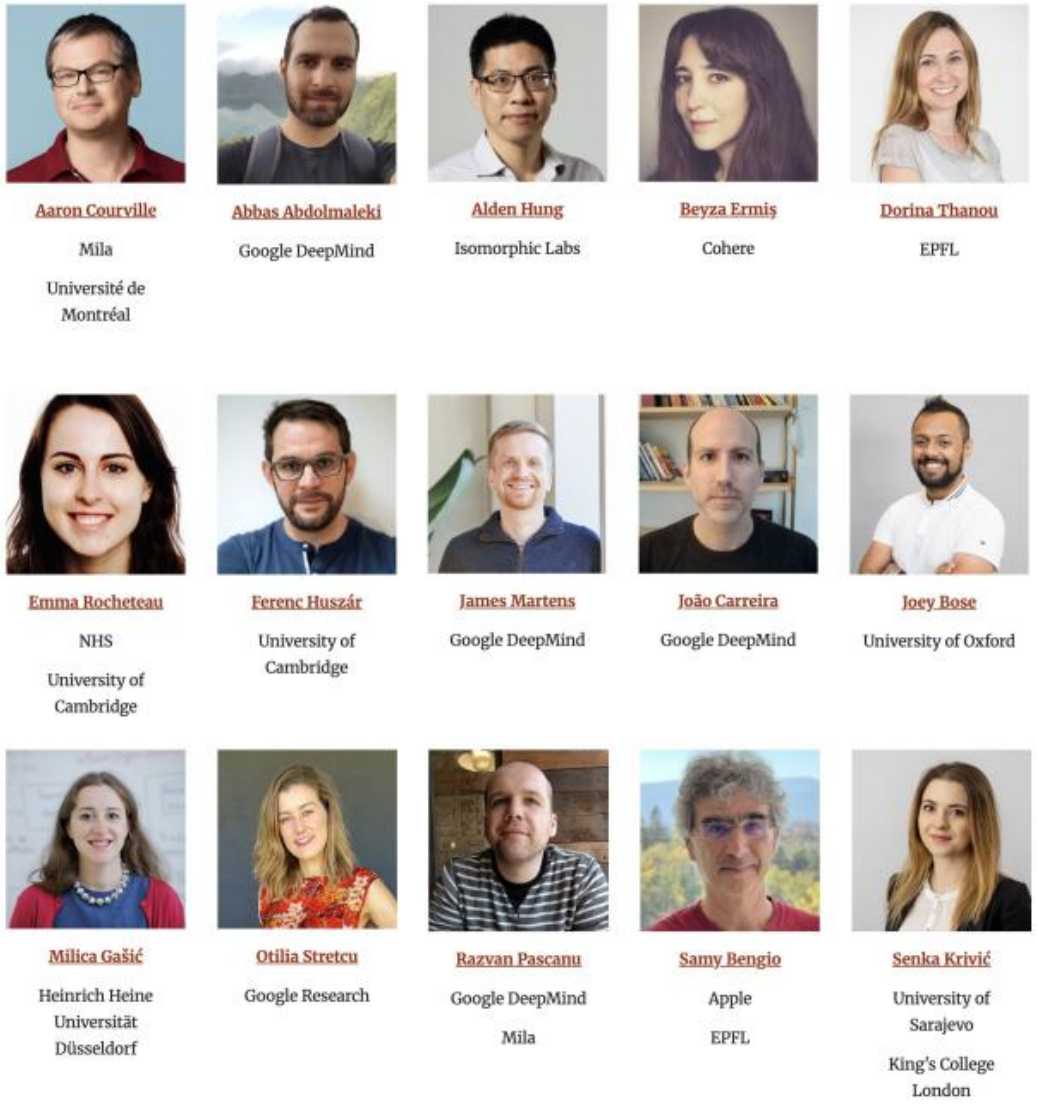
University of Sarajevo

ANNT

Eastern European Machine Learning 2025 (EEML 2025)

- 991 applications (!) makes EEML2025 the **most popular applied to** one so far – acceptance rate of ~20%.
- 300 participants (with industry)
- **More than 44 countries!**





Industrial collaboration – Infineon Technologies Austria AG

KAI

- Industrial research center
- 100% subsidiary of Infineon Technologies Austria AG
- Strong bridge between industry and academia

AI

Infineon

- Global leader in semiconductor solutions for automotive, industrial, and IoT sectors
- 58,060 employees, 71 R&D and 15 manufacturing locations,

infineon

Industry

Trustworthy, reliable, and effective image classification

Explainable computer vision

CLEAR -VISION

Novel DNN explainability methods

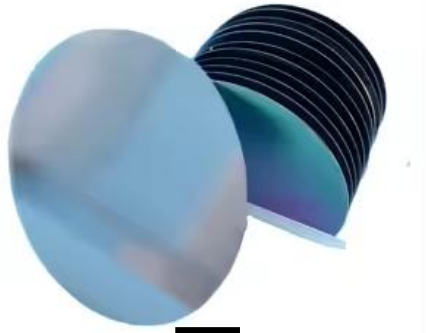
Academia

Semiconductor Manufacturing Process

Sand

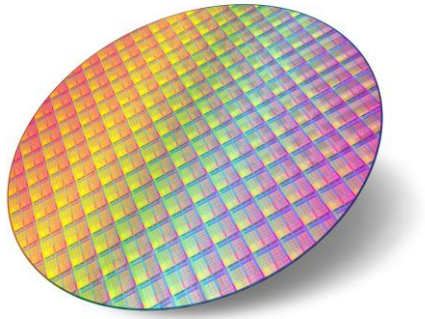
Silicon

Silicon wafer
(approx $1,000)



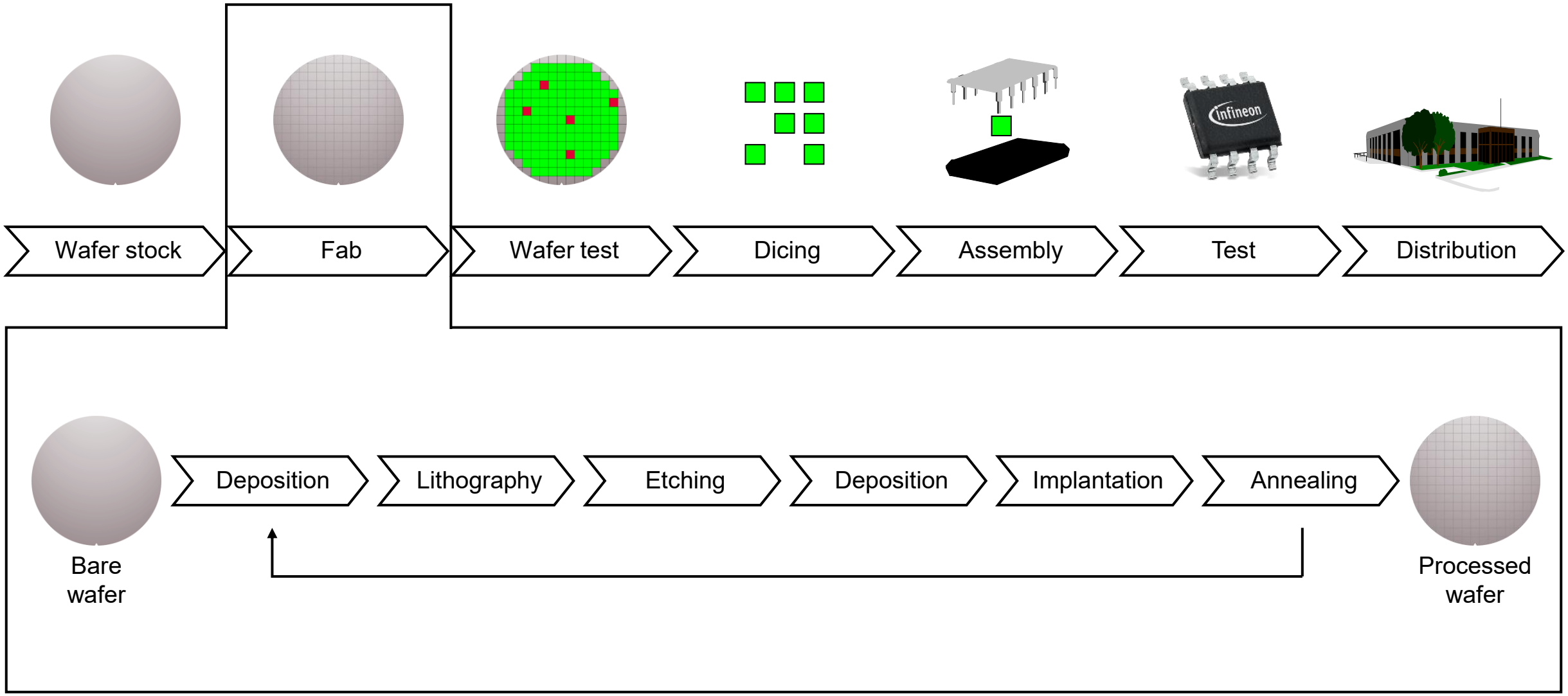
Production in cleanroom

Wafer after
production
(approx $30,000)

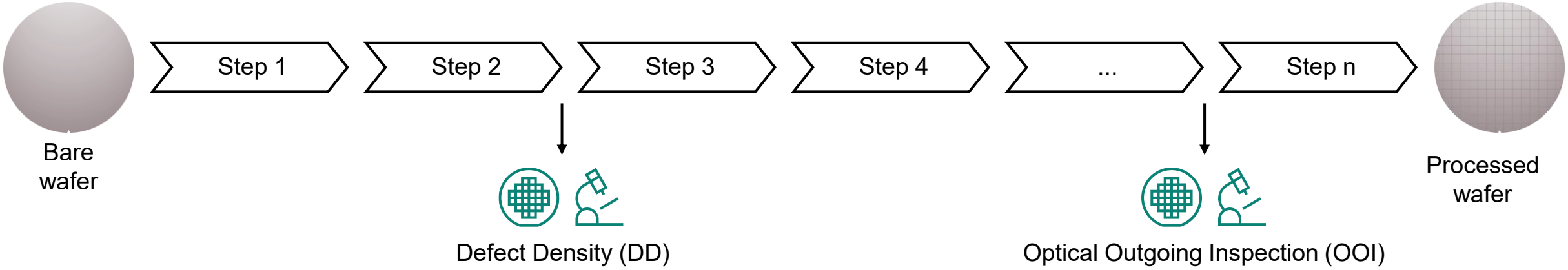


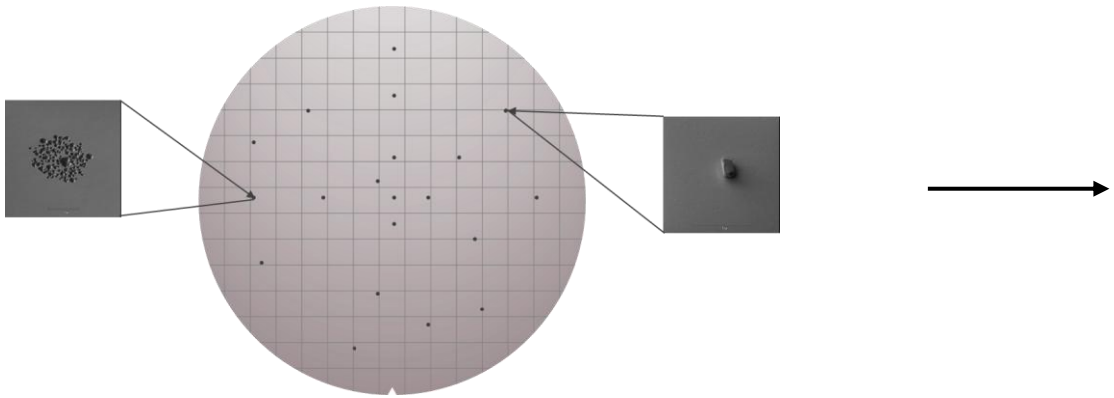
Chip

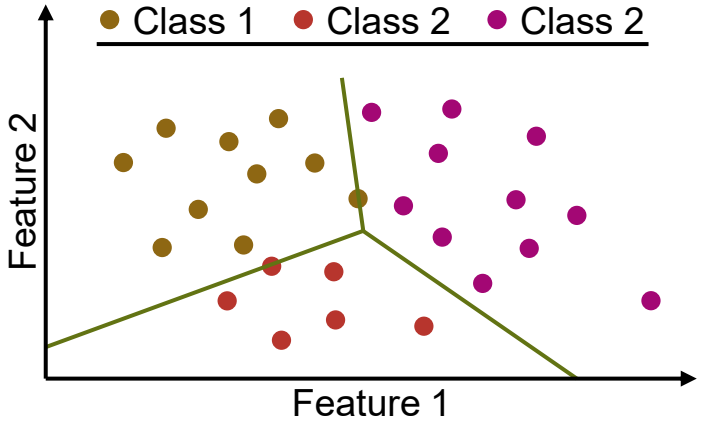
Semiconductor Production Processes



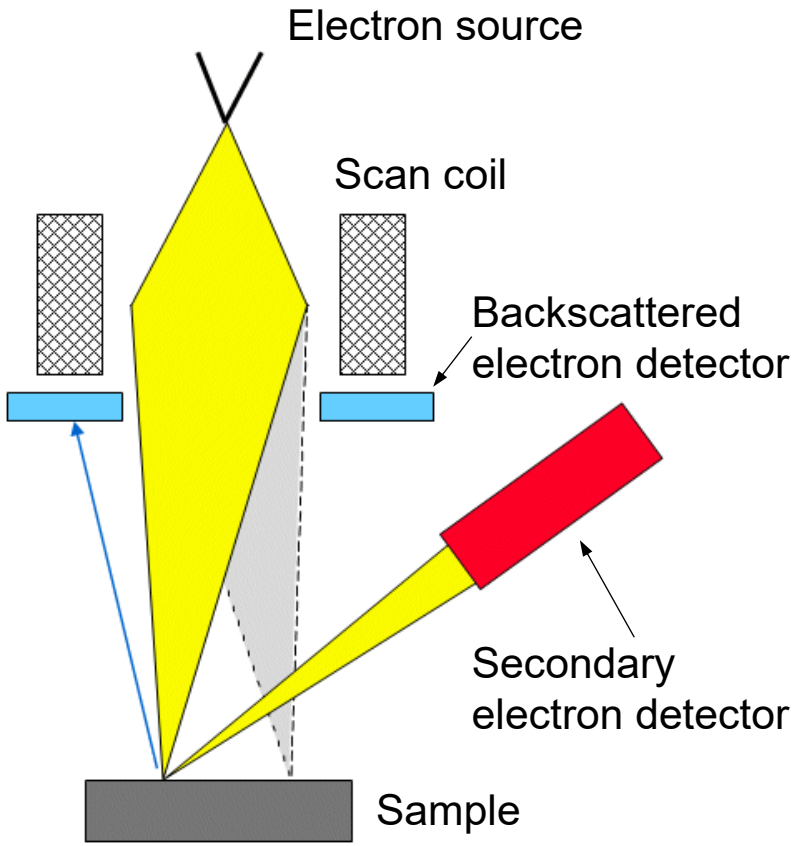
Quality Checks in Front-End Production – SEM defect image classification



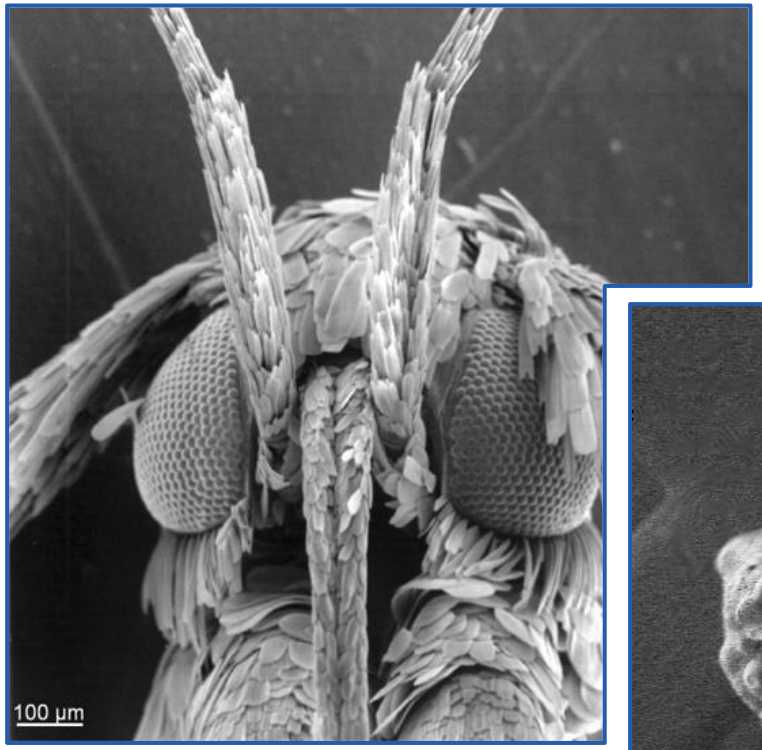




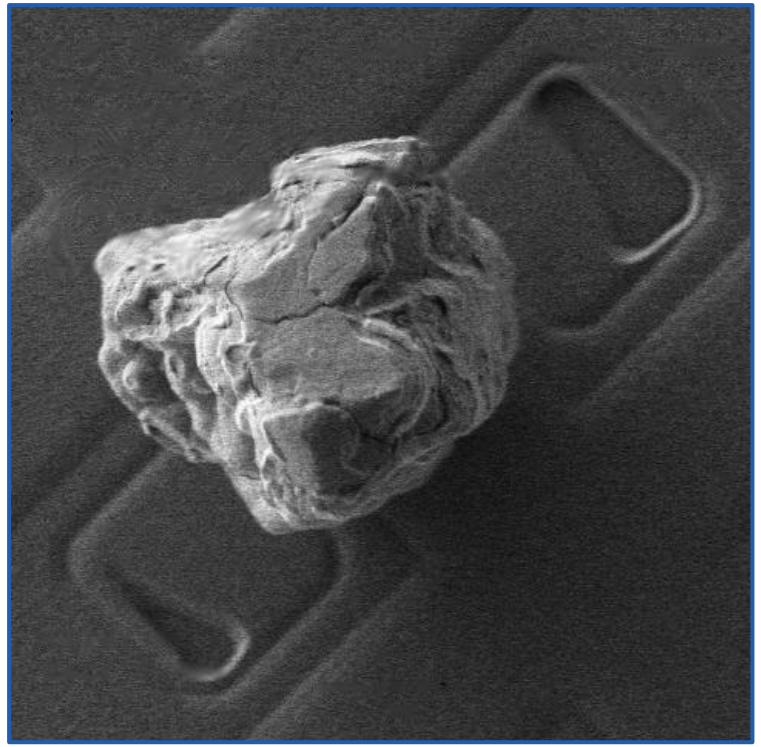
Data Source are Scanning Electron Microscopy Images



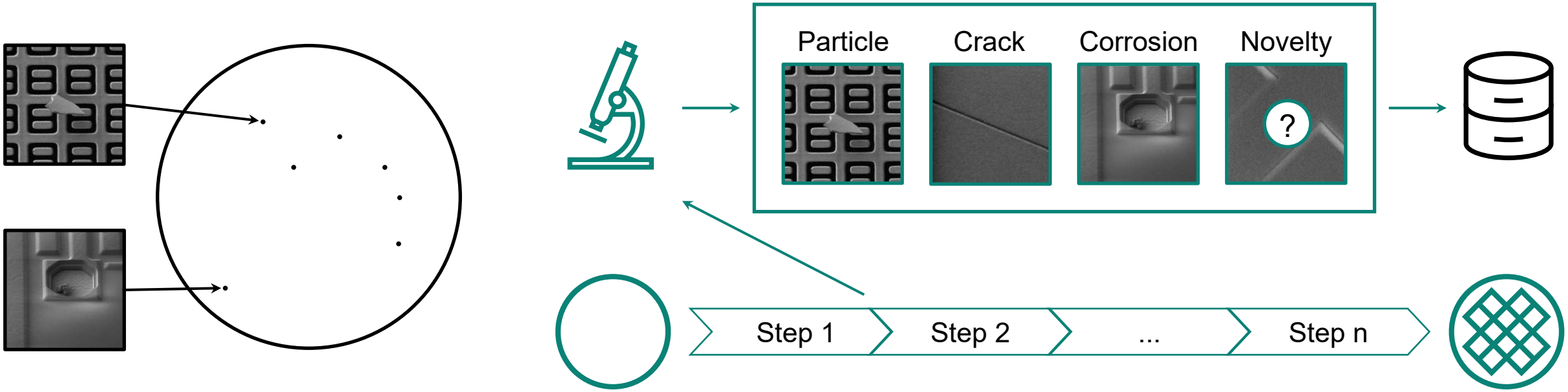
https://de.wikipedia.org/wiki/Rasterelektronenmikroskop



https://commons.wikimedia.org/wiki/File:Insect_SEM_gracilariidae.jpg

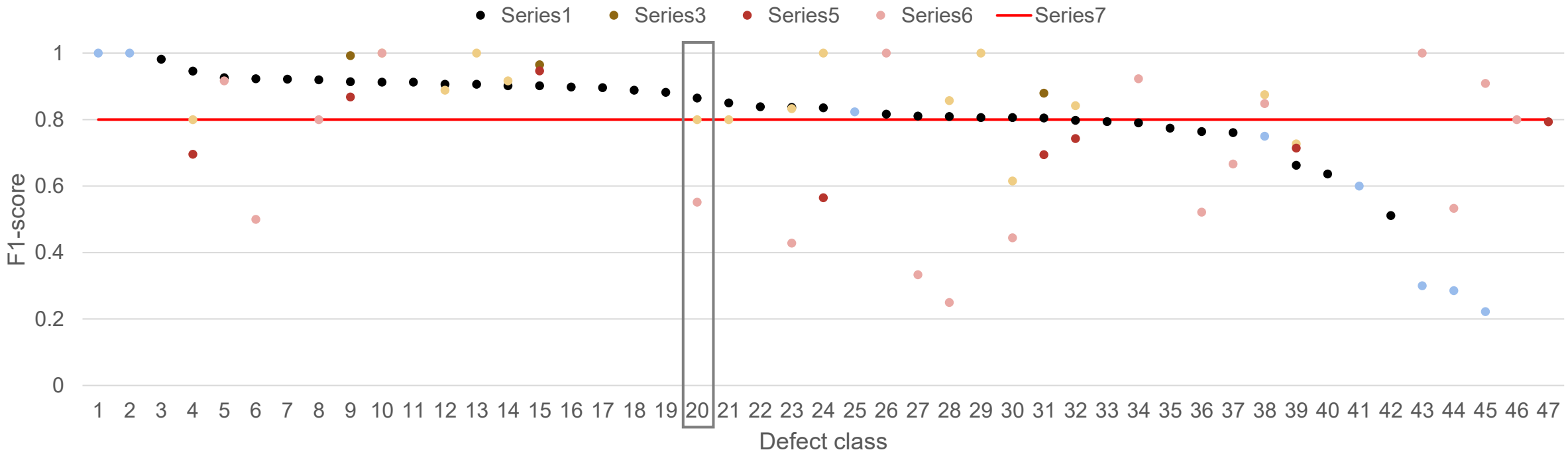


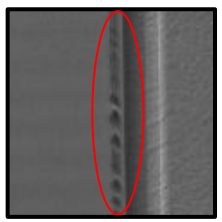
Defect Image Classification in Front-End Production

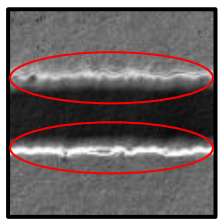


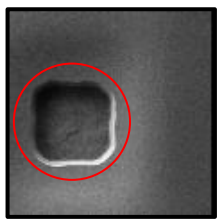
- Current state
  - Manual defect image classification by humans
- Goal
  - Automated defect image classification using AI models
- Challenges
  - Novelty detection, model monitoring & update
  - Scaling to other production sites & other use cases

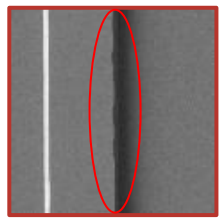
F1-Score Analysis on Validation Data



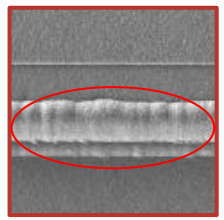

True 15✓
Pred 15✓

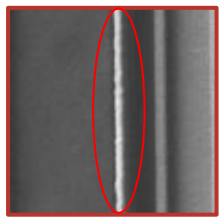

True 15✓
Pred 15✓

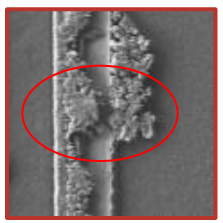

True 15✓
Pred 15✓

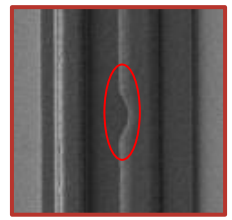

True 15✓
Pred 15✓


True 15✓
Pred 15✓


True 15✓
Pred 15✓


True 15✗
Pred 31✓

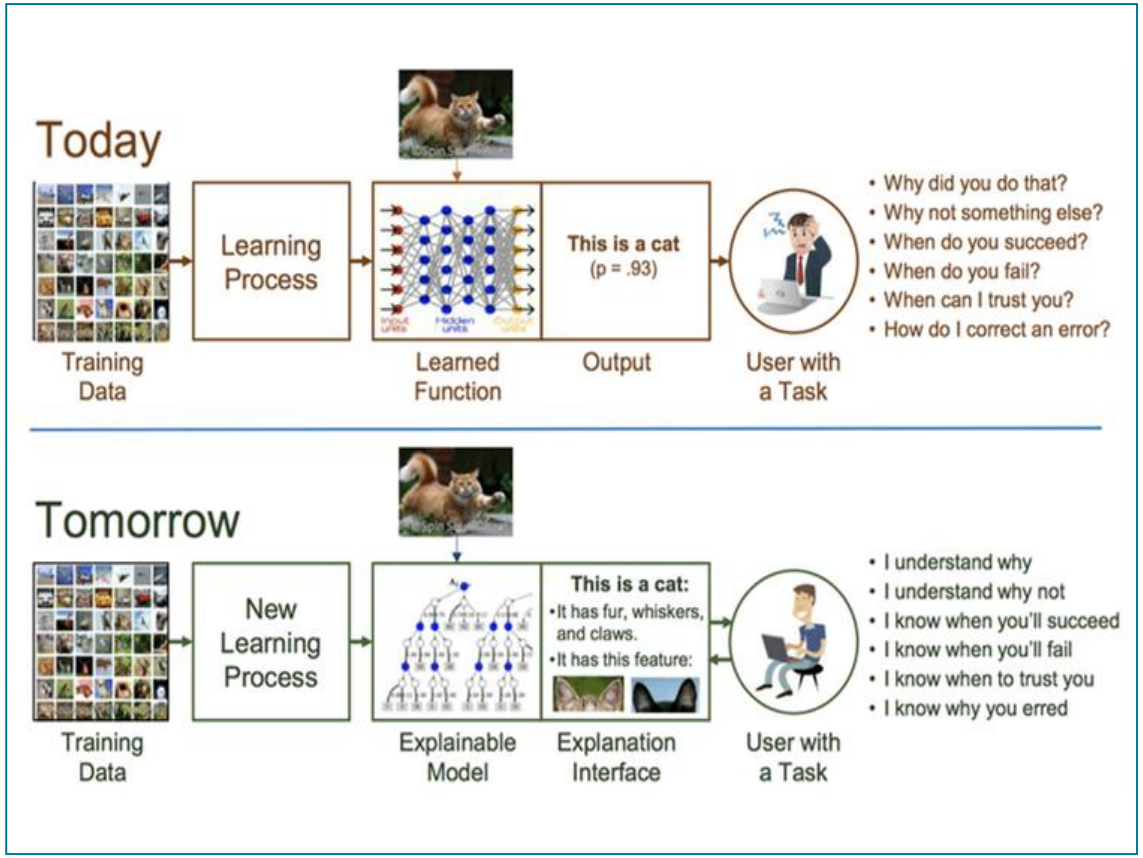

True 15✗
Pred 22✓


True 15✗
Pred 17✓

Motivation for XAI Research

Why explainable AI?

- Meet regulatory requirements
- Improved debugging of AI models
- Increased transparency of AI models
- Increased trust in AI models



Turek, M.: DARPA - Explainable Artificial Intelligence (XAI) Program (2017)
https://www.darpa.mil/program/explainable-artificial-intelligence

Towards Explainable Computer Vision – Contributions

Concept-based explanations

- Any Segment Explanation (ASE) – ICANN 2025
- Correlation SHAP (CorrSHAP) – XAI 2025
- Generalized Correlation SHAP (GenCorrSHAP) - Pattern Recognition

Data-centric explanations

- Kernel Sample Based Explanations (K-SBE) - CVIU
- Effects of Data Degradation on Explanations
- Model Provenance

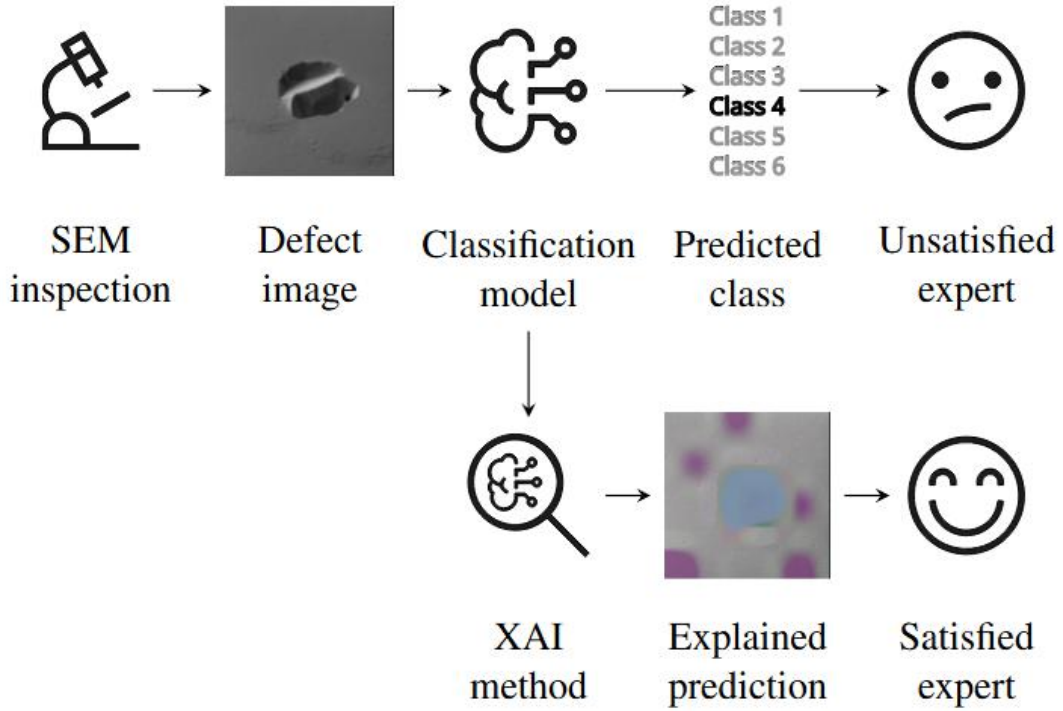
Multimodal explanations

- Explainable and Actionable Anomaly Detection
- Work at KCL: Explaining MLLMs

Towards Explaining SEM Defect Classification,
ECAI 2025, Bologna, Italy

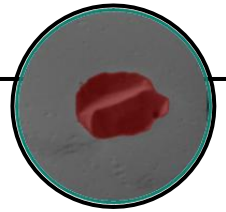
Towards Explaining SEM Defect Image Classification – Motivation

- Experts and models defect classification differ in some situations
- Experts want to understand how the model made the decision
- Curent pixel level explanations are not understandable
- Experts do not trust XAI evaluation metrics



Towards Explaining SEM Defect Image Classification

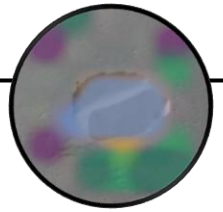
Automatic mask generation



SAM2[1]

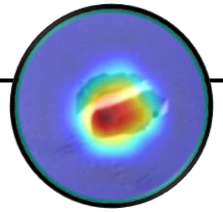
- Automatic mask generation using SAM2
- Expert validation of ground truth masks
- Carinthia-S dataset with ground truth segmentation masks

Evaluation of XAI methods on the Carinthia-S dataset



CRAFT[2]

- Concept Recursive Activation FacTorization
- Concept-based explanations
- Provide explanations in form of human-understandable concepts



GradCAM[3]

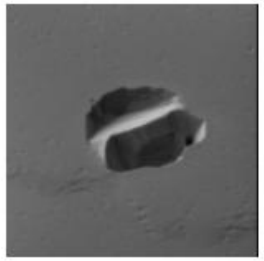
- Gradient-weighted class activation
- Feature-based explanations
- Provide explanations in form of saliency maps

[1] Ravi N, Gabeur V, Hu YT, Hu R, Ryali C, Ma T, Khedr H, Rädle R, Rolland C, Gustafson L, Mintun E. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714. 2024 Aug 1.

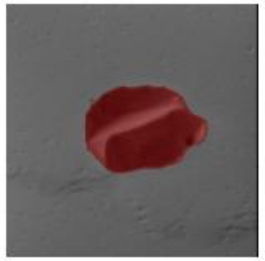
[2] Fel, Thomas, et al. "Craft: Concept recursive activation factorization for explainability." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[3] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

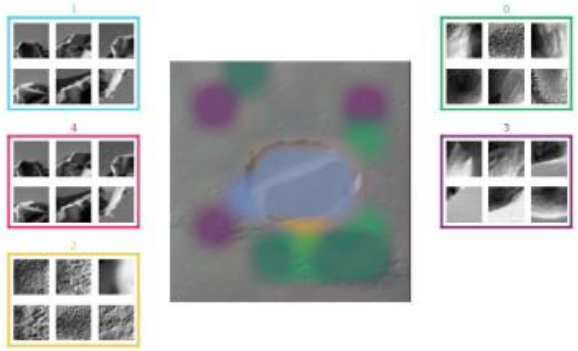
Towards Explaining SEM Defect Image Classification – Results



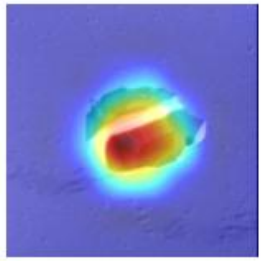
Original image



Ground truth



CRAFT



GradCAM

Questions	
Q1	From the explanation, I understand how the model works.
Q2	This explanation of how the model works is satisfying.
Q3	This explanation of how the model works has sufficient detail.
Q4	This explanation of how the model works seems complete.
Q5	This explanation of how the model works tells me how to use it.
Q6	This explanation of how the model works is useful to my goals.
Q7	This explanation of the model shows me how accurate the model is.
Q8	This explanation lets me judge when I should trust and not trust the model

5/9 experts found the explanations understandable (3.6/5) and satisfying (3.4/5)

7/9 experts found explanation helped judging when to trust or distrust the model's prediction (3.8/5)

R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. arXiv preprintarXiv:1812.04608, 2018.

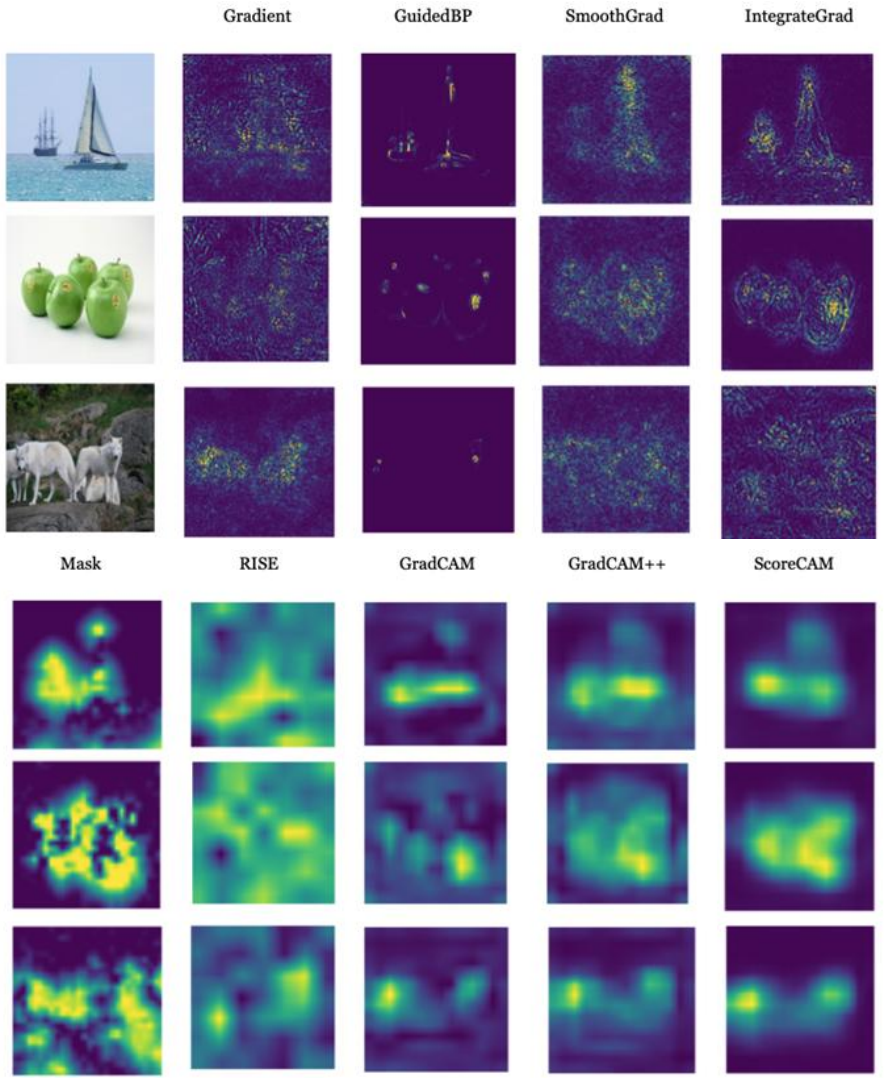
Towards Explaining SEM Defect Image Classification – Conclusion

- **Contributions:**
  - We proposed algorithm for automatic generation of ground truth masks for SEM defect images
  - We provided first dataset of semiconductor manufacturing defects with expert validated ground truth segmentation masks and named it Carinthia-S dataset
  - We achieved expert satisfaction with explanations with the application of CRAFT
- **Future work:**
  - Enhancing robustness of automatic ground truth segmentation algorithm
  - Dynamic hyperparameter tuning for CRAFT
  - Improving CRAFT heatmap precision

Understanding Image Classification Prediction with Any Segment Explanation, ICANN 2025, Kaunas, Lithuania

Any Segment Explanation (ASE) – Motivation

- **Goal of explanation is to be:**
  - human understandable
  - model faithful
- **Pixel-level explanations**
  - Are model faithful
  - Are not human understandable
  - Lack high-level semantic meaning
- **Concept-level explanations**
  - Are human understandable
  - Are reliant on manually annotated concepts

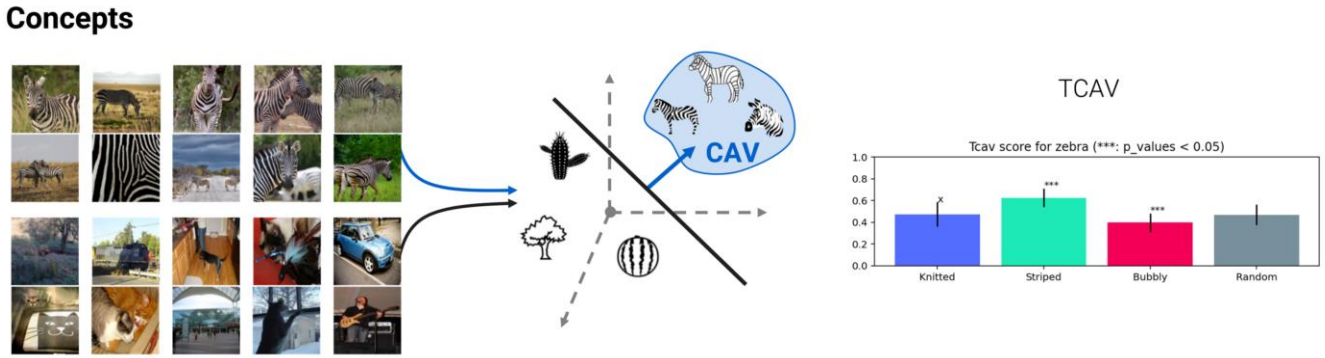


Wang, Haofan, et al. "Score-CAM: Score-weighted visual explanations for convolutional neural networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020.

Any Segment Explanation (ASE) – SOTA

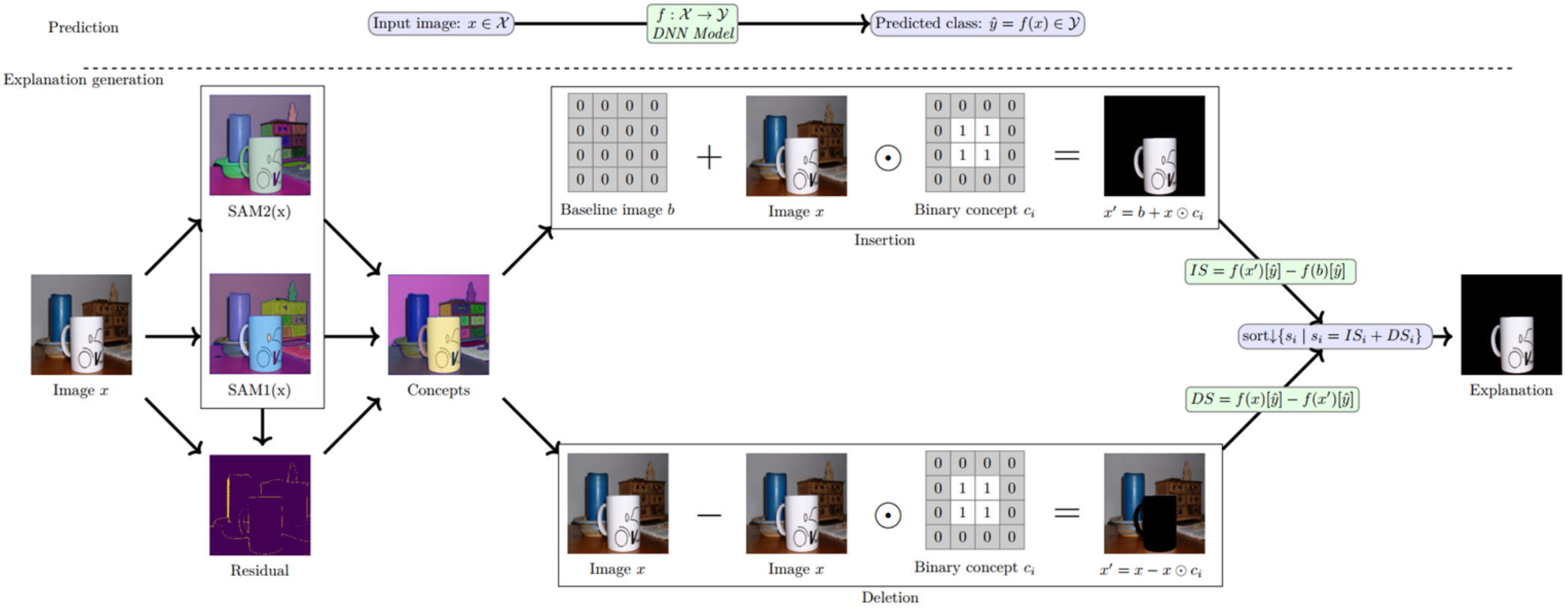
- **Testing with Concept Activation Vectors (TCAV)**
  - Users provide examples of a concept
  - A linear classifier is trained to distinguish these concept images
  - The vector orthogonal to this boundary is the **Concept Activation Vector (CAV)**
  - Concept entanglement

- **Explain Any Concept (EAC)**
  - Automatic concept extraction using SAM
  - Uses SHAP explanations
  - Explains surrogate linear model
  - Computationally expensive SHAP calculation



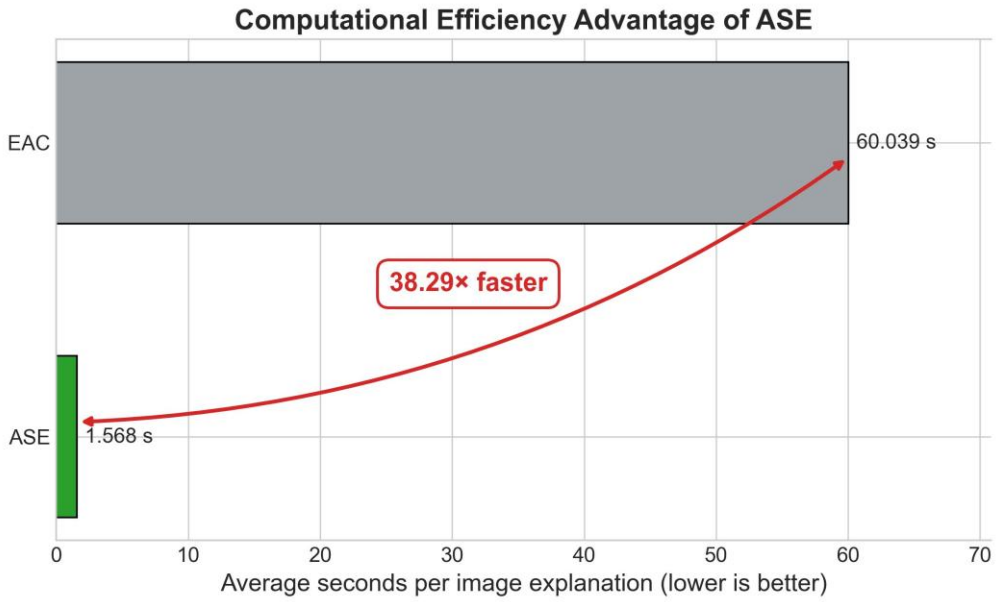
[1] Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." International conference on machine learning. PMLR, 2018.
[2] Sun, Ao, et al. "Explain any concept: Segment anything meets concept-based explanation." Advances in Neural Information Processing Systems 36 (2023): 21826-21840.
[3] Fel, Thomas, et al. “Xplique: A Deep Learning Explainability Toolbox.” Workshop on Explainable Artificial Intelligence forComputer Vision (CVPR), 2022

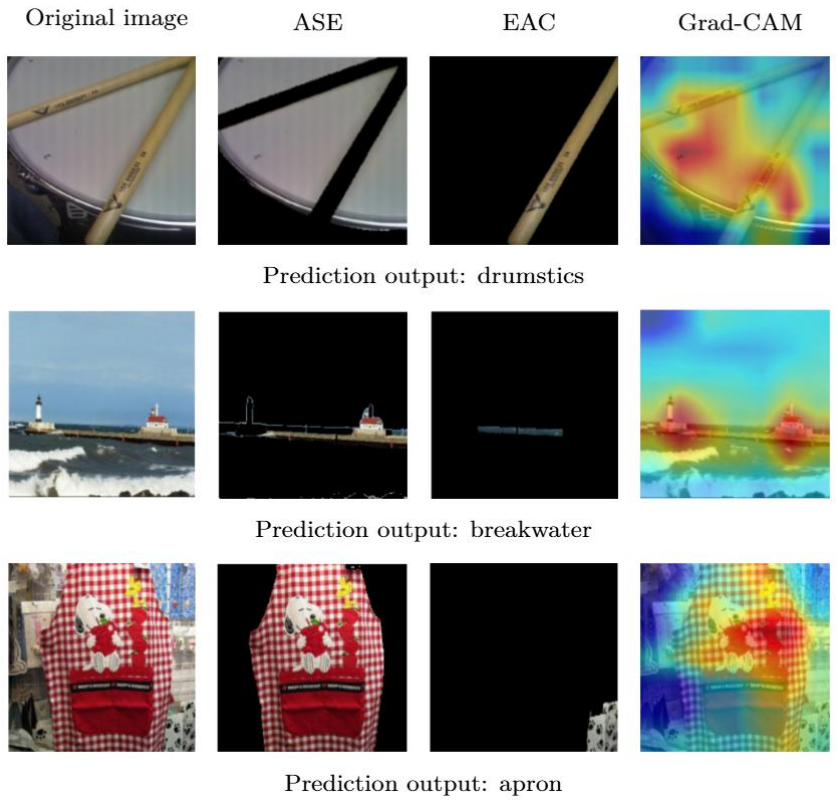
Any Segment Explanation (ASE)



International Conference on Artificial Neural Networks (ICANN 2025).

Any Segment Explanation (ASE) – Results





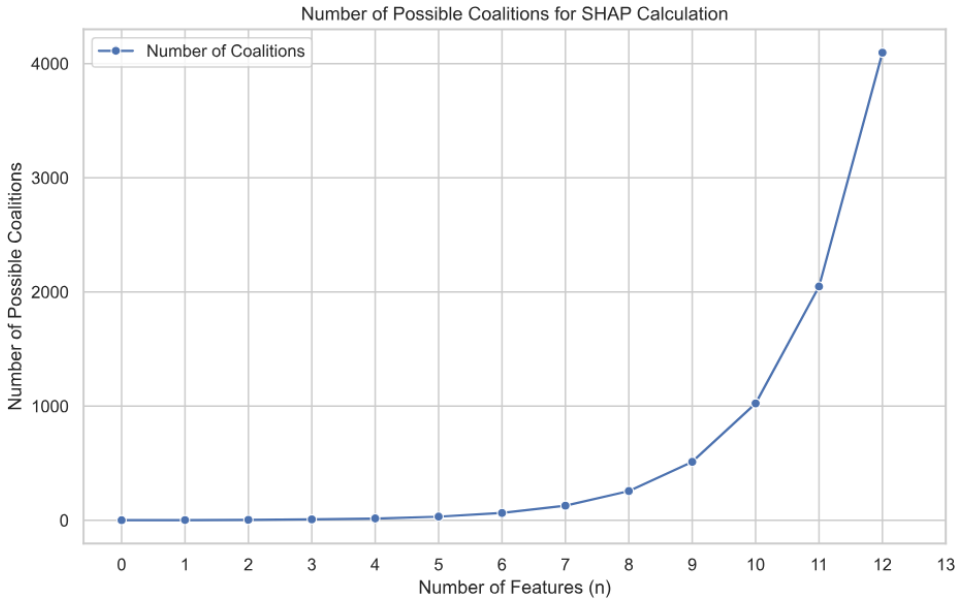
Any Segment Explanation (ASE) – Conclusion

- **Any Segment Explanations (ASE)**
  - We use combination of SOTA segmentation methods SAM1 and SAM2, and introduce residual concept to provide automatic high-level concept extraction
  - We propose explanation based on combined score of concept insertion and deletion to enable fast explanation generation
  - We achieved better faithfulness by not explaining surogate model
- **Future work:**
  - Concept dependencies
  - Multi-class classification

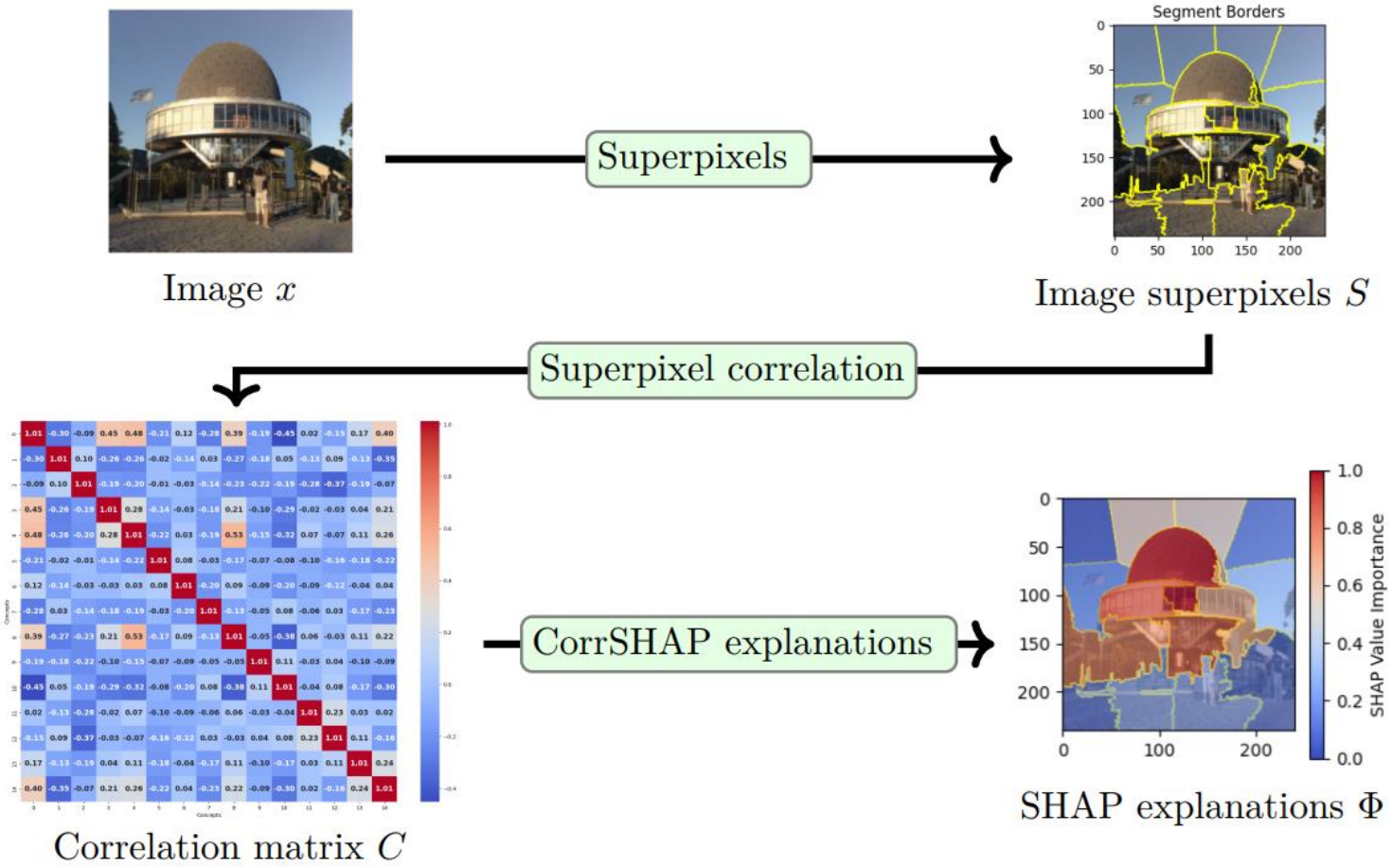
Superpixel Correlation for Explainable Image Classification,
XAI 2025, Istanbul, Turkey

Correlation SHAP (CorrSHAP) – Motivation

- **SHAP:**
  - Pixel-level explanations
  - Exponential computational complexity
- **Hypothesis: Image regions are correlated**
- **Correlation can be used to narrow down perturbation count**

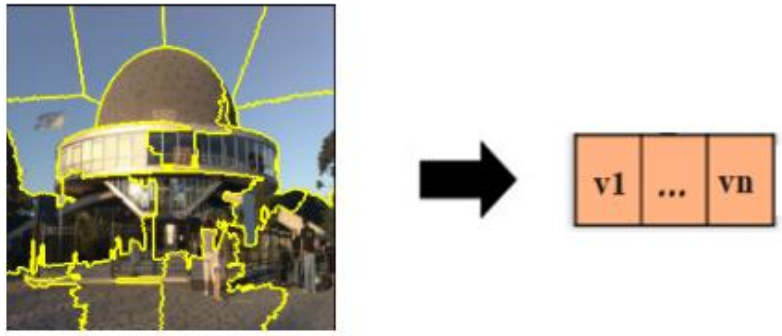


Correlation SHAP (CorrSHAP)



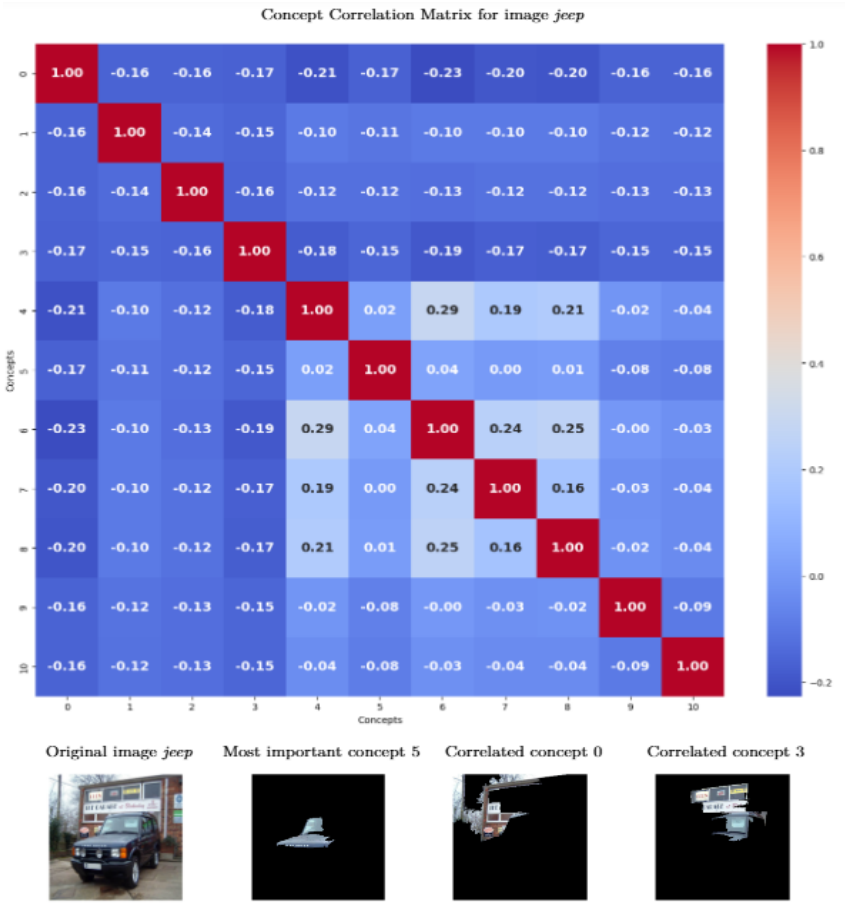
Correlation SHAP (CorrSHAP) – Results

Option 1 – raw pixel vectorization
Option 2 – feature map vectorization
Option 3 – gradient vectorization



$$v_i = \text{flatten}(s_i) = \text{flatten}(x \odot m_i)$$

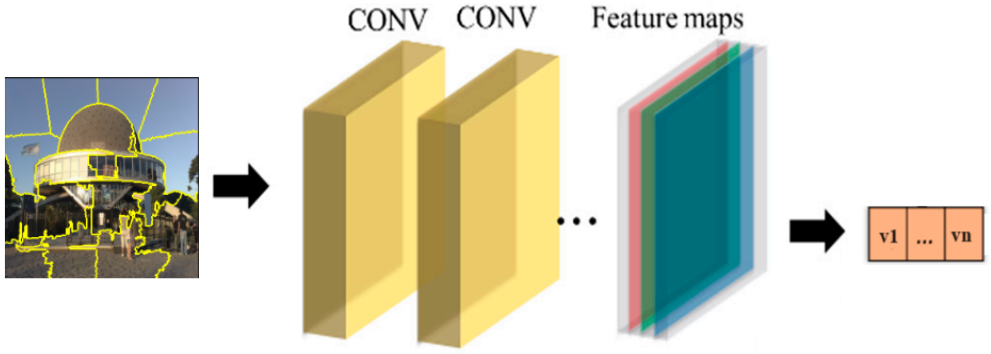
$$v_i \in \mathbb{R}^d, x \in \mathbb{R}^{H \times W \times K}, m_i \in \{0,1\}^{H \times W \times K}, d = H \cdot W \cdot K.$$



(a) Correlation matrix Option 1.

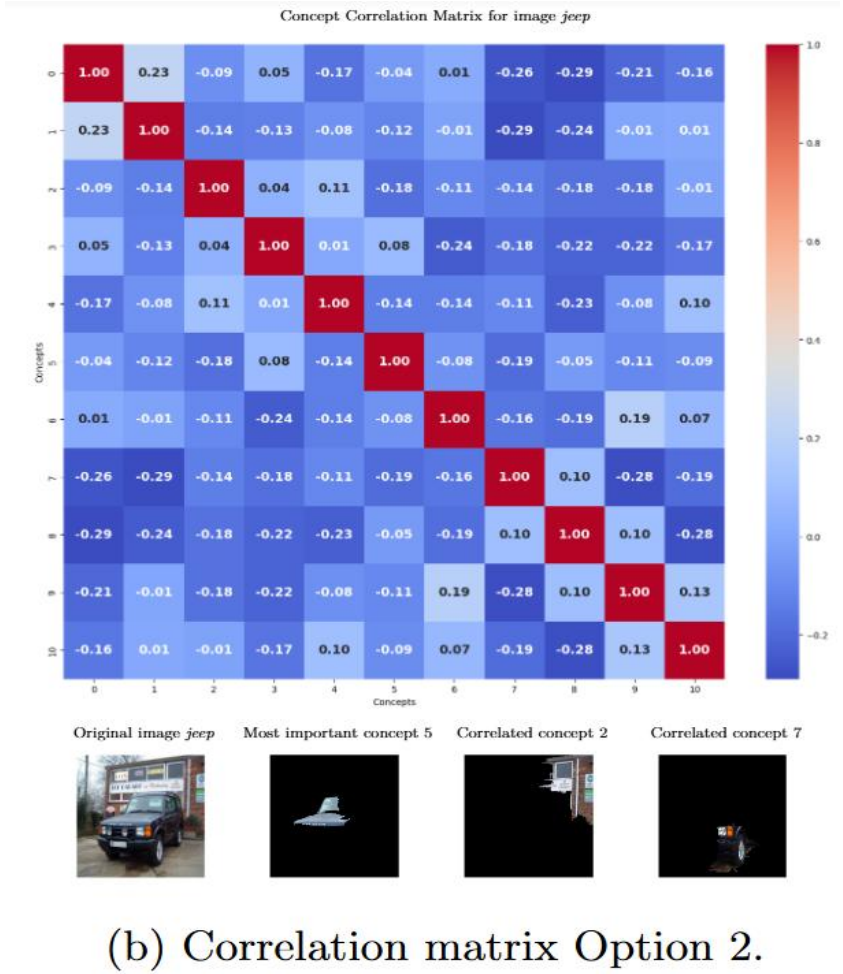
Correlation SHAP (CorrSHAP) – Results

Option 1 – raw pixel vectorization
Option 2 – feature map vectorization
Option 3 – gradient vectorization



$$v_i = \text{flatten}(g(s_i))$$

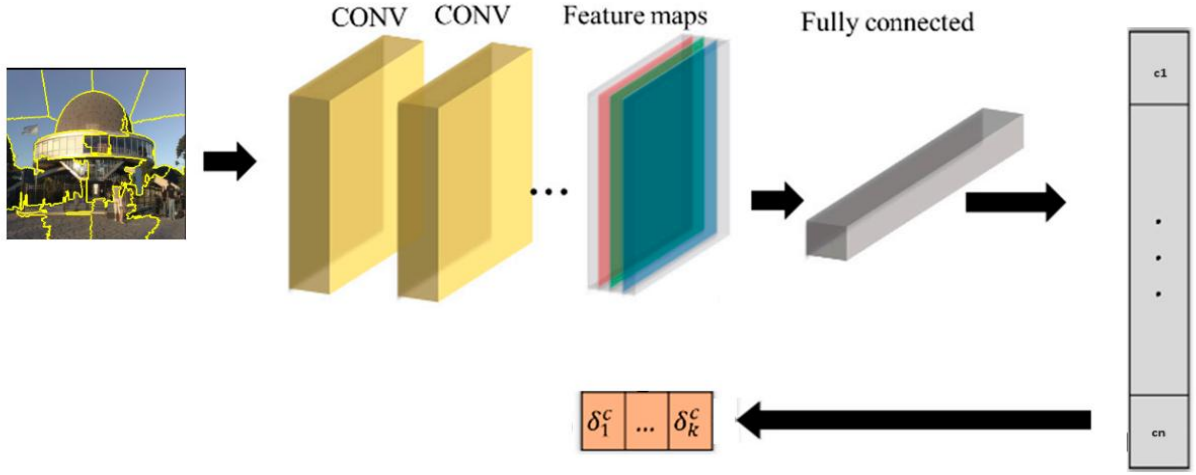
$$f(x) = h(g(x)), g : \mathcal{X} \to \mathcal{Z}, h : \mathcal{Z} \to \mathcal{Y}.$$



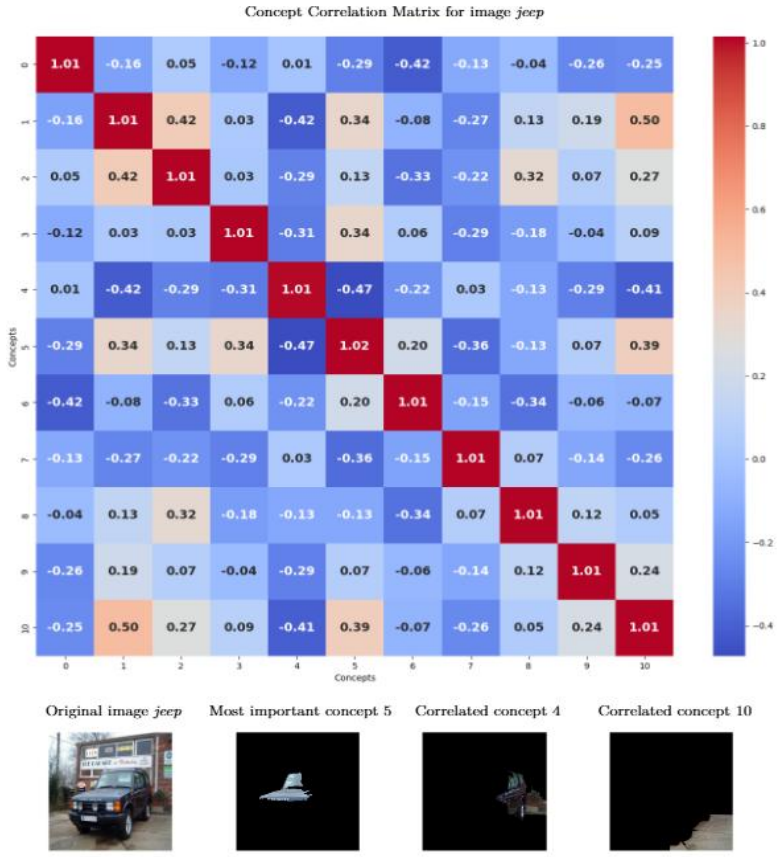
(b) Correlation matrix Option 2.

Correlation SHAP (CorrSHAP) – Results

Option 1 – raw pixel vectorization
Option 2 – feature map vectorization
Option 3 – gradient vectorization



$$v_i = \nabla_\theta f_\theta(s_i).$$

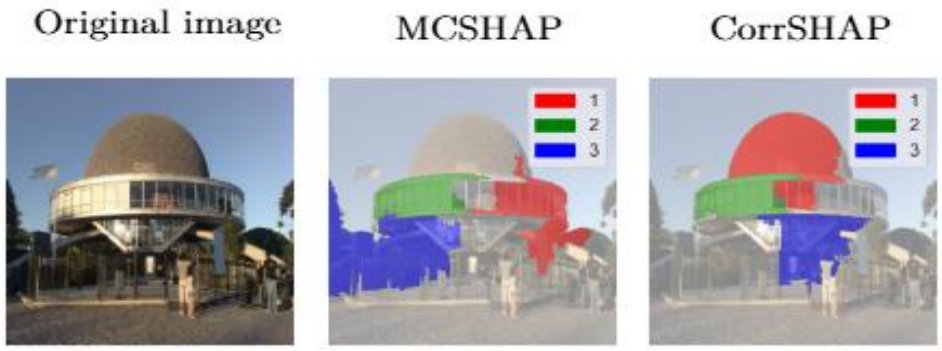


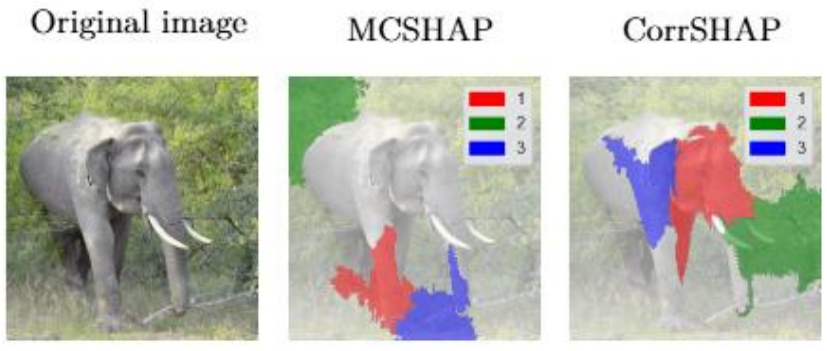
(c) Correlation matrix Option 3.

Correlation SHAP (CorrSHAP) – Results

Option 1 – raw pixel vectorization
Option 2 – feature map vectorization
Option 3 – gradient vectorization

- Better model faithfulnes than Monte Carlo SHAP
- 55 times faster than Monte Carlo SHAP





Correlation SHAP (CorrSHAP) – Conclusion

- We introduce superpixel correlation idea into concept based SHAP explanation to enable fast SHAP approximation
- **Future work:**
  - Alternative correlation measures
  - Integrating the superpixel correlation idea in other XAI methods

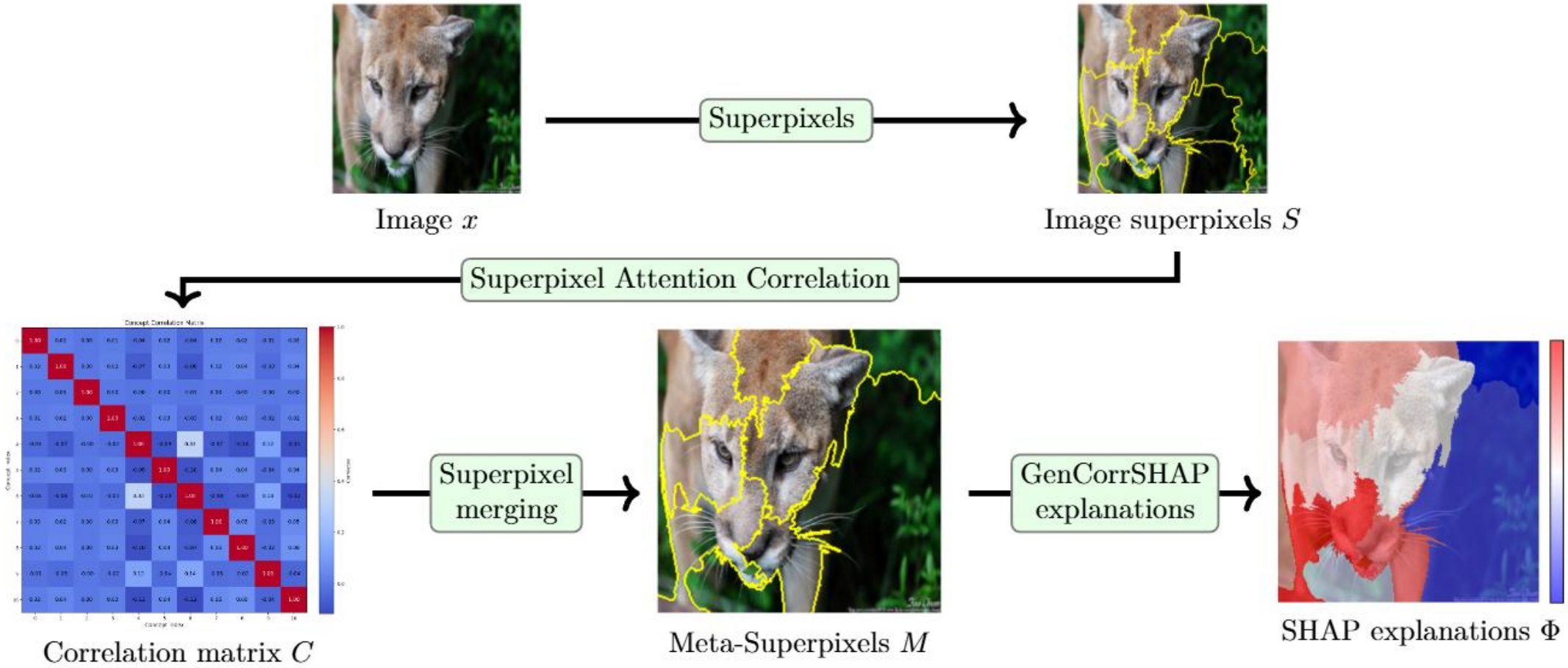
GenCorrSHAP: Generalized Superpixel Correlation for Explainable Image Classification

Under review at Pattern Recognition

Generalized Correlation SHAP (GenCorrSHAP) – Motivation

- **CorrSHAP**
  - Does not satisfy SHAP theorethical axioms: local accuracy, missingness, consistency
- **GenCorrSHAP:**
  - Superpixel Attention Correlation
  - Meta-superpixels based on correlation
  - Dynamic Threshold Selection - Otsu's method

Generalized Correlation SHAP (GenCorrSHAP)



Under review at Pattern Recognition

GenCorrSHAP – Results

- 12% better faithfulness than CorrSHAP
- 7% better faithfulness than SOTA PartitionSHAP and KernelSHAP

GenCorrSHAP – Conclusion

- We extend attention mechanism and intoruce Superpixel Attention Correlation to calculate superpixel correlation, which we use to adaptively merge superpixels into Meta-superpixels by using Dynamic Threshold Selection
- **Future work:**
  - Token merging for multimodal explanations

Revealing Training Data Influence in Deep Networks with Kernel Sample-Based Explanations,

Under review at Computer Vision and Image Understanding

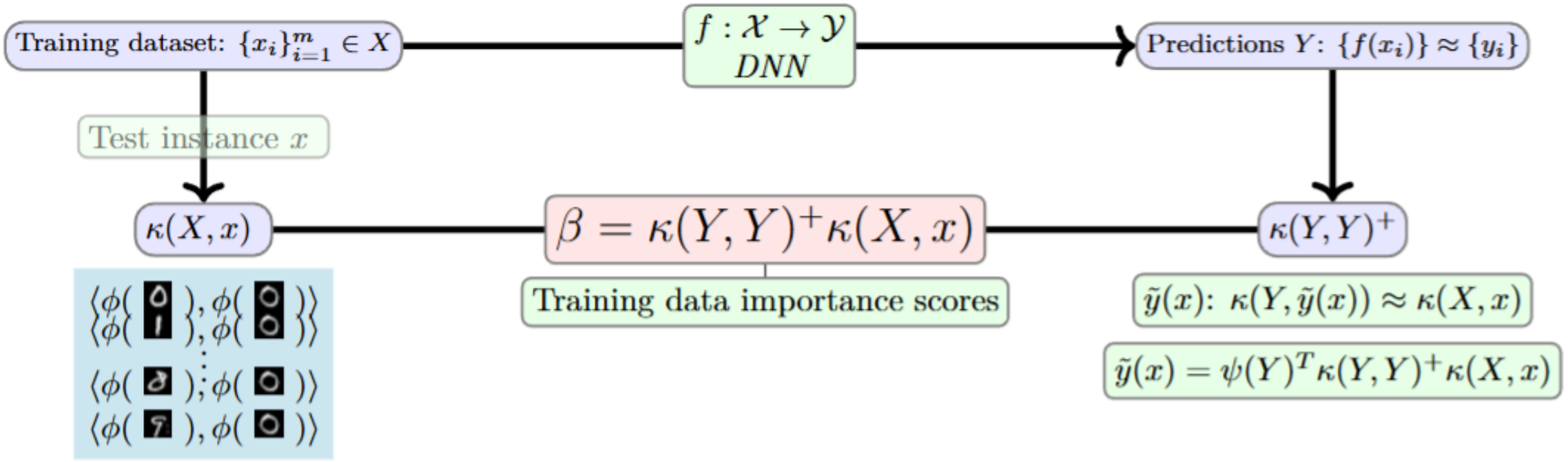
Kernel Sample Based Explanations (K-SBE) – Motivation

- **Give attribution to training instances for some test instance**
- **Various applications of TDA:**
  - Detecting mislabeled data
  - Identifying data leakage
  - Analyzing memorization effects
  - Optimizing training dataset

Kernel Sample Based Explanations (K-SBE) – SOTA

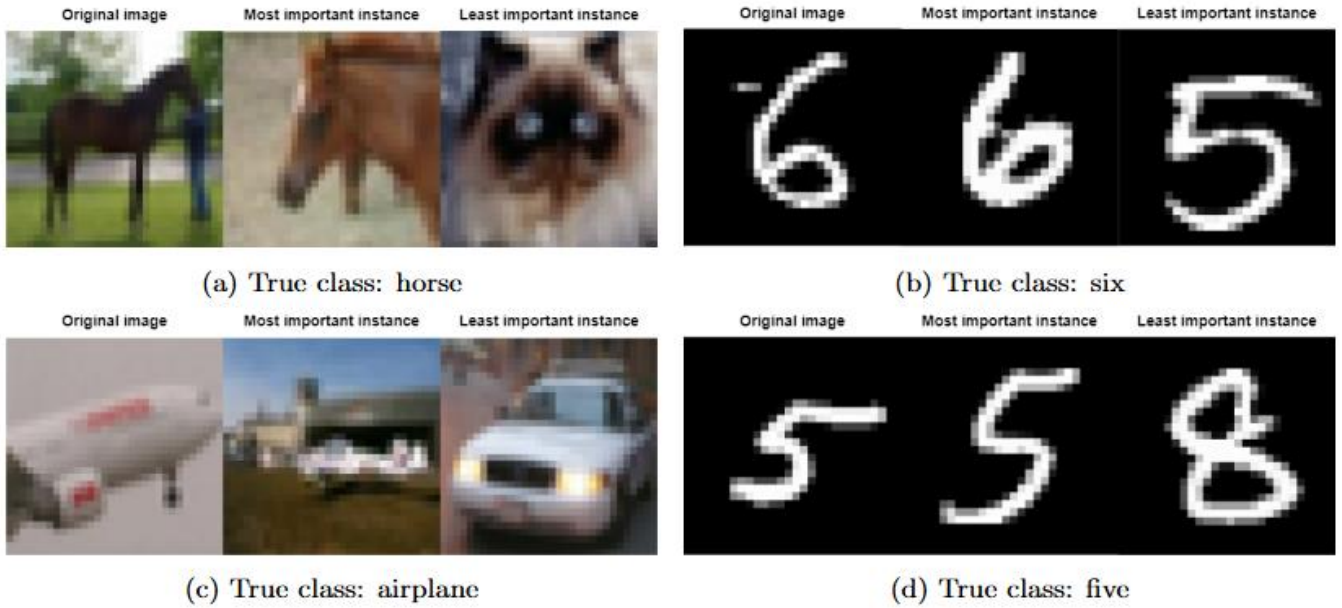
- **Current SOTA**
  - Influence functions
  - TraceIn
  - Generalized Reprezenters NTK
- **Highly computationally expensive**

Kernel Sample Based Explanations (K-SBE)



Kernel Sample Based Explanations (K-SBE) – Results

- More faithful than SOTA
- Sub-second execution time



Kernel Sample Based Explanations (K-SBE) – Future work

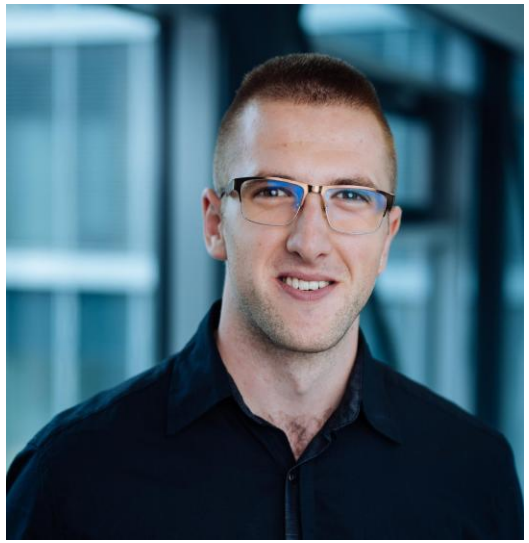
- We introduce Kernel Sample Based Explanations which maps the inputs and outputs of DNN into isomorphic Reproducing Kernel Hilbert Spaces (RKHS), where we model the function generated by the DNN via nonlinear kernel embeddings and a linear transformation between those spaces enabling black box TDA
- By utilising Nyström approximation we enable sub-second execution even on large kernels
- **Future work:**
  - Automated kernel learning based on model internals
  - Automatic hyperparameter tuning

King‘s College London

KCL Research Visit – My work at Spark AI research group

- At KCL: 10.01. – 14.03.2026
- Working on explainable MLLMs
- Happy to get to know you and open for collaboration ☺

⚡ Thank you for your attention!



LinkedIn