Kirill Bykov

Machine Learning Ph.D. Student
in Explainable AI and
Mechanistic Interpretability

Skills

Machine Learning
Computer Vision
Explainable AI
PyTorch/TensorFlow
Python (proficient)
C++, Java (intermediate)
Linux, Git, Bash
Data Visualisation, Tableau
Figma
SQL, NoSQL (MongoDB)
Experience with GCP, Azure
Research Supervision
Project Management
Scientific Writing

Languages

English (C2)
Deutsch (B1)
Russian (Native)

Contacts

kirill-bykov.com
linkedin.com/in/bykovkirill
twitter.com/kirill_bykov

Experience

Machine Learning Ph.D. Student, TU Berlin

Jan 2021 — Present | Full-time, Berlin, Germany

Pursuing Ph.D. degree at TU Berlin in Machine Learning in the area of Explainable AI. Supervised by Prof. Dr. Marina Höhne and Prof. Dr. Klaus-Robert Müller. Founding member of the Understandable Machine Intelligence Lab; since 2023, affiliated with ATB Potsdam following Ms. Höhne's Professorship appointment.

Student Research Assistant, TU Berlin

May 2020 — Jan 2021 | Part-time, Berlin, Germany

Assisted in Machine Learning research projects, specifically in the area of Explainable AI and Bayesian Neural Networks.

Data Science Research Intern, TomTom

Dec 2019— Mar 2020 | Part-time, Berlin, Germany

Developed Machine Learning models for Anomaly Detection in Geospatial Data in TomTom AI Geospatial Research Team.

Data Scientist, SkyEng

June 2018 — Apr 2019 | Part-time, Remote

Utilized Machine Learning for candidate analysis and scoring in the recruitment process. Implemented Process Mining techniques to enhance the efficiency of recruiting workflows.

Data Analyst, MegaFon

Sep 2015 — Dec 2017 | Part-time, Saint-Petersburg, Russia

Performed data analysis to optimize operational processes for the Trade Marketing team. Developed analytical models to support strategic decision-making and improve marketing efficiency.

Education

BIFOLD Graduate School

Mar 2021 — Present | Full-time, Berlin, Germany

Doctoral Researcher at Berlin Institute for the Foundations of Learning and Data Graduate School (BIFOLD).

MSc Data Science (ICT Innovation), TU Berlin

Oct 2018 — Dec 2020 | Full-time, Berlin, Germany

Double degree program with TU Eindhoven (2nd year), part of EIT Digital Master School Data Science Program.

MSc Data Science in Engineering, TU Eindhoven

Oct 2018 — Dec 2020 | Full-time, Eindhoven, Netherlands

Double degree program with TU Berlin (1st year), part of EIT Digital Master School Data Science Program. Graduated Cum Laude.

BSc Applied Mathematics and Computer Science, SPbSU

Sep 2014 — Sep 2018 | Full-time, Saint Petersburg, Russia

Graduated from the Faculty of Mathematics and Mechanics, Department of Statistical Modelling.

Publications

CoSy: Evaluating Textual Explanations of Neurons
NeurIPS 2024; 2024
Laura Kopf, Philine Lou Bommer, Anna Hedström, Sebastian Lapuschkin, Marina M.-C. Höhne, Kirill Bykov

Labeling Neural Representations with Inverse Recognition
NeurIPS 2023; 2023
Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, Marina M-C Höhne

DORA: Exploring Outlier Representations in Deep Neural Networks
Transactions on Machine Learning Research; 2023
Kirill Bykov, Mayukh Deb, Dennis Grinwald, Klaus-Robert Müller, Marina M-C Höhne

NoiseGrad — Enhancing Explanations by Introducing Stochasticity to Model Weights
AAAI Conference on Artificial Intelligence; 2022
Kirill Bykov*, Anna Hedström*, Shinichi Nakajima, Marina M-C Höhne

Finding Spurious Correlations with Function-Semantic Contrast Analysis
Springer CCIS, volume 1902; 2023
Kirill Bykov, Laura Kopf, Marina M-C Höhne

Mark My Words: Dangers of Watermarked Images in ImageNet
Springer CCIS, volume 1947,426—434 ; 2023
Kirill Bykov, Klaus-Robert Müller, Marina M-C Höhne

Visualizing the Diversity of Representations Learned by Bayesian Neural Networks
Transactions on Machine Learning Research; 2023
Dennis Grinwald, Kirill Bykov, Shinichi Nakajima, Marina M-C Höhne

Manipulating Feature Visualizations with Gradient Slingshots
ICML 2024, Mechanistic Interpretability Workshop
Dilyara Bareeva, Marina M.-C. Höhne, Alexander Warnecke, Lukas Pirch, Klaus-Robert Müller, Konrad Rieck, Kirill Bykov

Explaining Bayesian Neural Networks
ArXiv pre-print; 2021
Kirill Bykov, Marina M.-C. Höhne, Adelaida Creosteanu, Klaus-Robert Müller, Frederick Klauschen, Shinichi Nakajima, Marius Kloft

Achievements

- Organised “Global and Concept-Based Explainability” special track at XAI-2024 conference, moderated various sessions, including XI-ML workshop at ECAI 2023.

- Serve on the Program Committee for SaTML 2025, extensive peer-review experience for prestigious conferences and journals including NeurIPS, TMLR, IEEE TNNLS, and IEEE TRAMPI.

- EIT Digital Excellence Scholarship Recipient 2018 - 2020

- Winner of the Data Natives Hackathon 2019, Berlin, Germany; BioHack Hackathon 2018, Saint Petersburg, Russia, Prize-winner of DelftHack 2019, Delft, Netherlands.

- Prize-winner SkolTech Statistical Learning Olympiad 2018, ITMO Open Mathematical Olympiad 2014, Finalist at International Data Science Olympiad 2018.

Invited Talks (selected)

Labeling Neural Representations with Inverse Recognition
BLISS Berlin; 10 January 2023

How much can I trust you? Towards Understanding Neural Networks
Potsdam Graduate School;13 November 2023

DORA: Exploring Outlier Representations in Deep Neural Networks
Munich NLP; 27 September 2023;

Explainable AI: from Local to Global
Max-Delbrück-Center for Molecular Medicine; 5 July 2023

Panel discussion on Fair and Trustworthy AI
Helmholtz AI conference; 2 June 2022

Getting Insights from a Black Box: What Happens inside a Neural Network
Graduate School of Management, SPbSU;
Oct 22, 2021;