Project Assignments

SML Lab & **HAMI Lab**, University of Trento

Advanced Topics in Machine Learning and Optimization — 2025-26

Instructions

Project work

- Select one of the projects from the previous slides (or discuss with the teacher for custom projects)
- Complete it and prepare a report summarizing the methodology used and the results obtained
- After completing the assignment send it via email to the (first) contact person for the project
- Subject: ADVML2025
- Attachment: name_surname.zip containing:
  - the report (named report.pdf)
  - the code you wrote
  - the requirements needed to run the code

NOTE

- No group work
- Preliminary versions of the report can be sent for feedback
- The project is discussed asynchronously as soon as it is completed

Vector Quantized Dirichlet Calibration

Context

- In Machine Learning models are expected not only to be accurate but also **calibrated** (Wang, 2023; Sambyal et al., 2023) - i.e., their predicted probabilities should reflect the true empirical frequencies of the corresponding classes.
- Xiong et al. (2023) identify a phenomenon they term **proximity bias**: models tend to be over-confident on samples lying in sparse regions of the data manifold, and underconfident on high-density samples.
- Methods to correct such bias only focus on the predicted class. A novel technique for multiclass that learns a field of calibration maps via vector quantization is proposed for the student to implement.

Assignment

- The student is asked to:
  - Read up (Xiong et al., 2023) and be familiar with both Dirichlet Calibration (Kull et al., 2019) and Vector Quantization (Van Den Oord et al., 2017).
  - Implement the novel calibration technique.
  - Compute calibration metrics (e.g. ECE, ECCE or Brier Score) for the method.

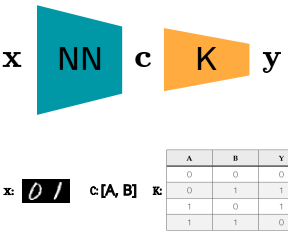
Info

- **Contact**: Cesare Barbera
- Extensible to thesis!

NeSy: Joint Reasoning Shortcuts with Structured Knowledge

Assignment

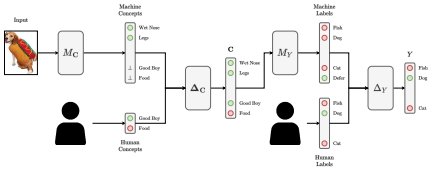
- **Inductive Logic Programming** (Evans and Grefenstette, 2018) (ILP) allows to learn logical rules from examples and background knowledge.
- NeSy models (that learns both concepts and knowledge) are prone to **reasoning shortcuts and algorithmic shortcuts** (Bortolotti et al., 2025). What happens when knowledge is more complex?
- The student is asked to:
  - Design an end-to-end NeSy architecture that uses differentiable ILP;
  - Create a logical task that admits reasoning and algorithmic shortcuts;
  - Compare this model with existing ones (e.g., (Daniele et al., 2023)). Is it more robust to shortcuts?



Info

- **Contact**: Samuele Bortolotti
- Code is available for rsbench
- Extensible to thesis!

NeSy: Learning to Defer with NeSy



Assignment

- **Learning to Defer** (Madras et al., 2018) (L2D) allows a classifier to choose when to predict or *defer* to a human. DCBMs (Pugnana et al., 2025) introduce a framework to **defer** on both **concept** and task, assuming **independence**.
- What happens when the independence assumption among concepts is broken? *E.g.* concepts are organized in a hierarchy?
  - Read L2D (Madras et al., 2018) and DCBMs (Pugnana et al., 2025).
  - Understand and design a Neuro-Symbolic layer that implements a hierarchy over concepts (Giunchiglia and Lukasiewicz, 2020; Ahmed et al., 2022).
  - Learn DCBMs equipped with this hierarchical concept layer and compare them with DCBMs. Can we reduce the number of deferrals?

Info

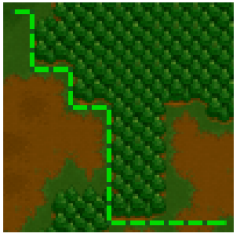
- **Contact**: Samuele Bortolotti, Andrea Pugnana
- Code is available [1, 2], for SPL, C-HMCNN and for DCBM.
- Extensible to thesis!

Assignment

- **Semantic-Probabilistic Layers** (SPLs) (Ahmed et al., 2022) can make any neural network output (structured) predictions that comply with constraints, e.g., safety or hierarchy constraints.
- Does this mean the resulting model is calibrated?
- The student is asked to:
  - Evaluate the calibration error of an SPL-augmented neural net.
  - Improve it by integrating Deep Ensembles (Lakshminarayanan et al., 2017).

Info

- **Contact**: Stefano Teso, Samuele Bortolotti,Cesare Barbera
- Code is available [1, 2] for SPL.
- Extensible to thesis!



SPL (ours)

Source: (Ahmed et al., 2022).

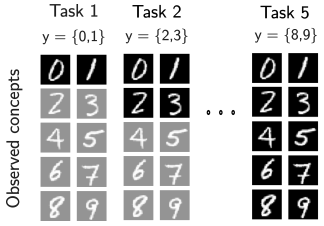
NeSy: Neuro-Symbolic Curriculum Learning

Assignment

- **Curriculum learning** (Bengio et al., 2009) is a strategy where the model is exposed to easier tasks or examples first, and progressively moves to harder ones.
- Tsamoura et al. (2025) uses statistical learning theory to compute the **class-specific risk** for each concept in a NeSy task.
- Can curriculum learning improve the model robustness to **reasoning shortcuts**?
- The student is asked to:
  - Read up (Tsamoura et al., 2025) and be familiar with reasoning shortcuts (Marconato et al., 2025a).
  - Implement a curriculum learning strategy in a NeSy task using the class-specific risk as a proxy for the "harder" classes.

Info

- **Contact**:Samuele Bortolotti
- Code is available for rsbench
- Extensible to thesis!



NeSy: Why SPL suffers from these shortcuts?

Context

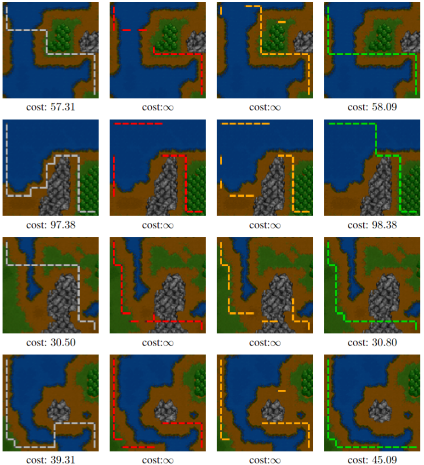
NeSy models can suffer from shortcuts that in cases either trade-off task accuracy for constraint satisfaction (Li et al., 2024), or learn wrong concepts that are used in the knowledge (Bortolotti et al., 2025).

Assignment

- Replicate experiments on path finding in Warcraft maps for SPL (Ahmed et al., 2022)
- Understand why solutions in green (right) often differ from ground-truth in yellow (left)
- Formulate hypotheses about knowledge satisfaction, task solving, vs simplicity of the solution

Info

- **Contacts**: Emanuele Marconato, Samuele Bortolotti
- Code is available [1, 2] for SPL.



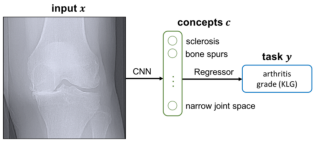
CBMs: Faithful-by-construction CBMs

Assignment

- Concept-Bottleneck Models (CBMs) (Koh et al., 2020) combine a neural network that extracts concepts and a linear layer that uses concepts for computing a prediction.
- What if we want to know what concepts are (ir)relevant? Thresholding the concept activations leads to *unfaithful explanations*.
- The student is asked to:
  - Show empirically that thresholding is unfaithful
  - Implement provably faithful explanations in CBMs

Info

- **Contact**: Stefano Teso, Steve Azzolin
- Notes are available for all steps.
- Code is available for CBMs and for faithful explanations.
- Extensible to internship/thesis!



Source: (Koh et al., 2020).

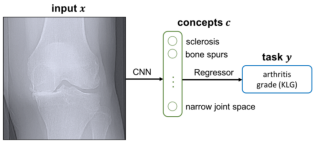
CBMs: Conformally Predictive CBMs

Assignment

- Concept-Bottleneck Models (CBMs) (Koh et al., 2020) combine a neural network that extracts concepts and a linear layer that uses concepts for computing a prediction.
- Task prediction does not currently consider concept uncertainty. When a model is uncertain on some concepts, it should not use them to predict the label.
- The student is asked to:
  - Train a CBM model and measure concept uncertainty (via M.C. Dropout).
  - Implement a linear layer that provides the user with $P(y)$ computed analitically (not an approximation) based on concept uncertainties.
  - Design a loss to try to maximise $P(y)$

Info

- **Contact**: Nicola Debole, Stefano Teso
- Code is available baseline CBM.



Source: (Koh et al., 2020).

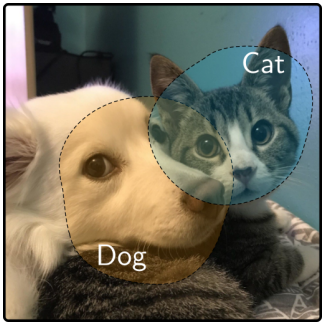
CBMs: Avoiding Unwanted Correlations between Concepts

Assignment

- Concept-Bottleneck Models (CBMs) (Koh et al., 2020) combine a neural network that extracts concepts and a linear layer that uses concepts for computing a prediction.
- Sometimes they can learn concepts – say "dog" and "cat" – that *wrongly correlate with each other*.
- The student is asked to:
  - Show empirically that unwanted correlations are real by analyzing the saliency maps of the concepts.
  - Fix the problem with a new loss term!

Info

- **Contact**: Stefano Teso, Nicola Debole.
- Notes are available.
- Code is also available for the baseline CBM.
- Extensible to internship/thesis!

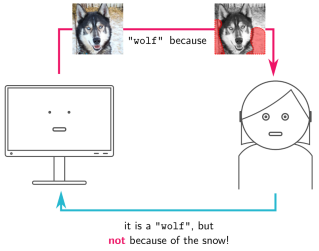


Source: Photoshop.

XAI: Smarter Explanatory Interactive Learning with Simplicity Sampling

Assignment

- In XIL, a human fixes Clever Hans behavior by correcting a model's explanations (Schramowski et al., 2020).
- There is *no smart way of prioritizing explanations*!
- The student is asked to:
  - Help design a selection strategy that prioritize simpler, "easier to learn" examples, as they are more likely to be confounded (Yang et al., 2024).
  - Compare simplicity sampling against random and uncertainty sampling on at least two known-confounded datasets (Steinmann et al., 2024)

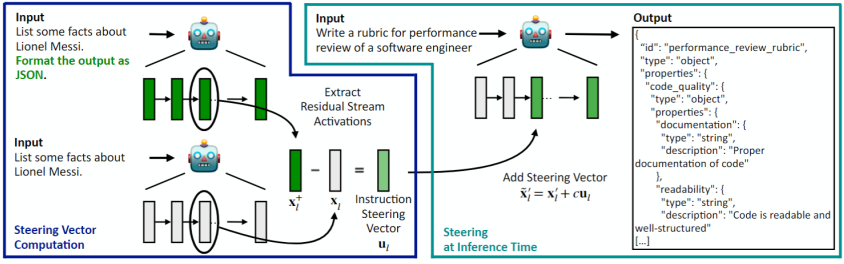


Source: (Schramowski et al., 2020).

Info

- **Contact**: Stefano Teso.
- Code is available for XIL.
- Notes are available for all steps.
- Extensible to internship/thesis!

LLMs: Linear properties of sentence embeddings



Main idea

- LLMs are able to follow instructions when fine-tuned on it or, sometimes, by linearly modifying their activations through steering (Stolfo et al., 2025).
- The purpose of the project is to investigate in which manner such instructions are encoded in model representations, starting from open-source small-scale language models to bigger LLMs.

Info

- For Linear Relational Properties, see (Hernandez et al., 2023; Marconato et al., 2025b)
- **Contact**: Emanuele Marconato

UNaIVERSE Platform: The Battle of Generative AIs

Assignment

- The UNaIVERSE platform (Melacci et al., 2025) has been recently designed to put in touch artificial and human agents in a private manner, regulating decentralized communities of heterogeneous agents into the so-called "worlds" (peer-to-peer networks), https://unaiverse.io.
- Select a dataset of pictures belonging to a certain domain (faces, cars, ...). Design a world with three roles: the *trainee*, an agent based on a neural discriminator, able to classify an input image as real or fake; the *sparring partner* and the *final boss*, agents based on generative models (GAN, Variational Autoencoder, Diffusion Model, ..., see Sengar et al. (2025)), capable of generating pictures.
- Sparring partners stream through the peer-to-peer network batches of generated (fake) images and real ones, asking the trainee to learn from them. The trainee considers one sparring partner at a time, and, finally, evaluates the fake images from the boss. *Will it be able to mark them as fake?*
- The student is asked to:
  - Design the world, agents, and the dynamics of the community, using the UNaIVERSE platform.
  - Experiment with two sparring partners, a trainee, a boss. Pretrain the sparring partners and the bosses.
  - Simulate multiple combinations of generators and compare the quality of the resulting trainees.

Info

- **Contacts**: Andrea Passerini, Stefano Melacci, Christian Di Maio, Tommaso Guidi.
- Code and examples are available.

Supervised Abstention

Assignment

- Missing data appear in several settings. A common assumption is to treat them as if they are missing at random (MAR). However, often data are not missing at random, but for some unknown reason.
- Abstaining models refrain from providing a prediction when too uncertain about the prediction. Can we exploit missingness to abstain in a supervised manner?
- The student is asked to:
  - Compare at least two popular approaches to perform abstention vs a supervised strategy that abstains when a label is missing not at random.
  - Check how results are affected if the randomness comes from some information available or unobserved variables.

Info

- **Contact**: Andrea Pugnana
- A paper with an introduction and an empirical comparison of abstention models is available here.

Graph Learning Dataset Evaluation via RINGS

Assignment (Francesco Ferrini, Antonio Longa, Veronica Lachi)

- Recent work has shown that many **graph ML benchmarks are flawed**, as non-graph models often outperform GNNs.
- Coupette et al. (2025) introduces a principled framework to evaluate graph-learning datasets via:
  - **Performance Separability** (**P1**)
  - **Mode Complementarity** (**P2**)
- Robinson et al. (2024) introduces **RelBench**, which converts relational databases into graphs for benchmarking relational deep learning.
- However, many benchmarks in RelBench are still solved well by tabular (non-GNN) models.

Task

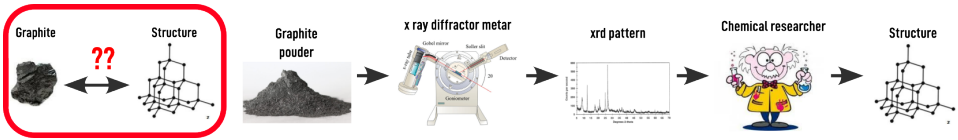
The student is asked to:

- Verify whether RelBench datasets satisfy **P1** and **P2** (as defined by Coupette et al. (2025)).
- If not, design an alternative dataset (even just one benchmark) that enforces these properties.
- Compare GNNs vs non-graph baselines to validate the new dataset.

Info

- **Contacts**: Francesco Ferrini, Antonio Longa, Veronica Lachi,

From XRD pattern to Metal Organic Frameworks (MOF) Structure



Assignment (Antonio Longa, Dinga Wonanke)

- Reconstructing a crystal structure from an XRD pattern is extremely challenging: experts must solve the phase problem, index peaks, determine the space group, refine the lattice, and ensure chemical plausibility. This expert-driven workflow is long and complex, and ML models could significantly accelerate it.
- The open research question: can a neural model learn to map **XRD** $\rightarrow$ **periodic graph structure** directly?
- The student is asked to:
  - Use FAIR-MOFs Wonanke et al. (2025) (a dataset that provides both XRD patterns and MOF structures) to build an ML system that generates a **crystal structure** from an **XRD pattern**.
  - Evaluate reconstruction quality against ground-truth structures.

Info

- **Contacts**: Antonio Longa, Dinga Wonanke (Chemical researcher)
- Extensible to internship/thesis!

References

Ahmed, K., Teso, S., Chang, K.-W., Van den Broeck, G., and Vergari, A. (2022). Semantic probabilistic layers for neuro-symbolic learning. *NeurIPS*.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Bortolotti, S., Marconato, E., Morettin, P., Passerini, A., and Teso, S. (2025). Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Coupette, C., Wayland, J., Simons, E., and Rieck, B. (2025). No metric to rule them all: Toward principled evaluations of graph-learning datasets. *arXiv preprint arXiv:2502.02379*.

Daniele, A., Campari, T., Malhotra, S., and Serafini, L. (2023). Deep symbolic learning: Discovering symbols and rules from perceptions. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023, page 3597–3605. International Joint Conferences on Artificial Intelligence Organization.

Evans, R. and Grefenstette, E. (2018). Learning explanatory rules from noisy data.

Giunchiglia, E. and Lukasiewicz, T. (2020). Coherent hierarchical multi-label classification networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9662–9673. Curran Associates, Inc.

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., Belinkov, Y., and Bau, D. (2023). Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. In *ICML*.

Kull, M., Perelló-Nieto, M., Kängsepp, M., de Menezes e Silva Filho, T., Song, H., and Flach, P. A. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, pages 12295–12305.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30.

Li, Z., Liu, Z., Yao, Y., Xu, J., Chen, T., Ma, X., and Lü, J. (2024). Learning with logical constraints but without shortcut satisfaction. *arXiv preprint arXiv:2403.00329*.

Madras, D., Pitassi, T., and Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer.

Marconato, E., Bortolotti, S., van Krieken, E., Morettin, P., Umili, E., Vergari, A., Tsamoura, E., Passerini, A., and Teso, S. (2025a). Symbol grounding in neuro-symbolic ai: A gentle introduction to reasoning shortcuts.

Marconato, E., Lachapelle, S., Weichwald, S., and Gresele, L. (2025b). All or none: Identifiable linear properties of next-token predictors in language modeling.

Melacci, S., Di Maio, C., Guidi, T., and Gori, M. (2025). UNaIVERSE — unaiverse.io. https://unaiverse.io. [Accessed 05-11-2025].

Pugnana, A., Massidda, R., Giannini, F., Barbiero, P., Zarlenga, M. E., Pellungrini, R., Dominici, G., Giannotti, F., and Bacciu, D. (2025). Deferring concept bottleneck models: Learning to defer interventions to inaccurate experts. In *NeurIPS*.

Robinson, J., Ranjan, R., Hu, W., Huang, K., Han, J., Dobles, A., Fey, M., Lenssen, J. E., Yuan, Y., Zhang, Z., He, X., and Leskovec, J. (2024). Relbench: A benchmark for deep learning on relational databases. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sambyal, A. S., Niyaz, U., Krishnan, N. C., and Bathula, D. R. (2023). Understanding calibration of deep neural networks for medical image classification. *Comput. Methods Programs Biomed.*, 242:107816.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486.

Sengar, S. S., Hasan, A. B., Kumar, S., and Carroll, F. (2025). Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications*, 84(21):23661–23700.

Steinmann, D., Divo, F., Kraus, M., Wüst, A., Struppek, L., Friedrich, F., and Kersting, K. (2024). Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation. *arXiv preprint arXiv:2412.05152*.

Stolfo, A., Balachandran, V., Yousefi, S., Horvitz, E., and Nushi, B. (2025). Improving instruction-following in language models through activation steering. In *The Thirteenth International Conference on Learning Representations*.

Tsamoura, E., Wang, K., and Roth, D. (2025). Imbalances in neurosymbolic learning: Characterization and mitigating strategies. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. *CoRR*, abs/2308.01222.

Wonanke, D., Longa, A., Pankajakshan, A., Himanen, L., Ladines, A. N., Márquez, J. A., Addicoat, M. A., Crittenden, D., Scheidgen, M., Lio, P., et al. (2025). Fair-mofs: A comprehensive database for accelerating the discovery and synthesis of metal-organic frameworks.

Xiong, M., Deng, A., Koh, P. W. W., Wu, J., Li, S., Xu, J., and Hooi, B. (2023). Proximity-informed calibration for deep neural networks. In *NeurIPS*.

Yang, Y., Gan, E., Dziugaite, G. K., and Mirzasoleiman, B. (2024). Identifying spurious biases early in training through the lens of simplicity bias. In *International conference on artificial intelligence and statistics*, pages 2953–2961. PMLR.