The problem: Machine Learning in the wild. Emerging applications of artificial intelligence (AI), such as smart cars navigating complex, dynamic environments, robotic surgical assistants capable of anticipating a surgeon's intentions and needs, or smart warehouses monitoring the quality of the work being conducted, are exposing serious issues with robustness in machine learning (ML) [1,2]. Current methods attempt to best fit the available training data, but cannot provide guarantees on their performance on new, test data captured under radically different settings [3]. This may lead, for instance, an autonomous car [4] to perform well during normal training situations, but to fail catastrophically when tested while driving through extreme weather conditions [5]. The tool classically adopted to address this issue, Vapnik's statistical learning theory (SLT) [6], predicts bounds on the generalisation error (the expected prediction error on new test data) so wide to be useless, and rely on the assumption that training and testing data come from the same (albeit unknown) probability distribution. Practitioners thus simply resort to cross-validation [7] to identify the parameters of the optimal model. With the pervasive deployment of machine learning algorithms in 'mission-critical' AI systems, for which failure is not an option, it is imperative to ensure that these algorithms behave predictably 'in the wild' [8].

Current Research

Exciting current research in the area, often called 'domain adaptation', is exploring a number of options, e.g. counterfactual error bounding [9], risk-sensitive reinforcement learning [10] or minimax learning [11]. The latter employs minimax optimisation to learn models adapted to data generated by any probability distribution within a "safe" family. Portfolio of models have also been proposed [12], e.g. Satzilla [13] and IBM Watson [14], which contemplate sets of models of different nature (e.g. support vector machines, random forests) or models of the same kind but learned from different slices of data (e.g. boosting). Recent progress includes multi-view learning [1,15] and learning of transferable components [16].

Imprecise probabilities [17] (a collective name for a hierarchy of mathematical theories of uncertainty, including, among others, robust Bayesian methods [18] and random set theory [19,20]) arise whenever the available evidence is insufficient to identify a specific probability distribution. In virtually all applications of machine learning, training sets constitute a glaring example of data which is insufficient in both quantity (think of the Google object detection routine, which is trained on a few million images compared to the thousands of billions of images out there) and quality (as they are selected based on criteria such as cost, availability or mental biases, thus introducing a strong bias in the learning process).

Hypotheses and Objectives

To address these challenges, **we seek to develop a new paradigm for robust machine learning 'in the wild', over which a radically new generation of AI systems may be constructed.**

Our ambitious goal is to lay the foundations for an entirely novel, inherently robust theory of learning. We believe this may be achieved via the following transformational steps (reflected by the WPs below):

(1) generalising traditional Statistical Learning Theory (SLT) in order to allow for test data to be sampled from a different probability distribution than the training data, under the weaker assumption that both belong to a same convex set of distributions (what in the literature is commonly called a 'credal set' [18]);
(2) further extending the SLT framework by moving away from the idea of selecting of a single model from a class of hypotheses, to identifying convex sets of models induced by the available training data;
(3) applying the resulting convex-theoretical learning theory (with respect to both the data-generating probabilities and the model space) to the functional spaces associated with deep networks (i.e., series of convolutions and max pooling operations) in order to lay solid theoretical foundations for deep learning.

To accompany this blue sky rethinking of machine learning, a library of code for solving the optimisation problems associated with the newly developed classes of learning algorithms will be made available (in Python & Café) to all researchers. The results of our research will be published at top-tier AI and machine learning venues (e.g. IJCAI, NIPS, ICML) and journals (AIJ, PAMI, but also Science/Nature).

Significance and Impact

The proposed new paradigm for machine learning would allow scientists to develop new classes of algorithms for coping with real-world AI deployment, e.g. autonomous cars [4] able to safely engage situations which are entirely new to them. Deep learning (DL) [21] is revolutionising fields as varied as speech recognition, computer vision [22] and game playing (such as Atari computer games and the ancient Chinese game of Go [23]). New robust foundations for it (objective (3)) will allow the field to further progress towards AIs capable of negotiating complex environments in the company of humans.

Methodology – broken down into 3 Work Packages (WPs).

WP1 – SLT with convex sets of probabilities. Important ML approaches such as max-margin support vector machines (SVMs) and sparsity methods which force models to have bounded L1 norm, owe their theoretical justification to SLT-derived upper bounds to their generalisation error, expressed in terms of margin or of sparsity index, respectively. A central role is played by the notion of Rademacher complexity

[24], to which bounds can be reduced [25], and by concentration inequalities from classical probability theory. As mentioned, all such bounds assume a single distribution for both training and test data.

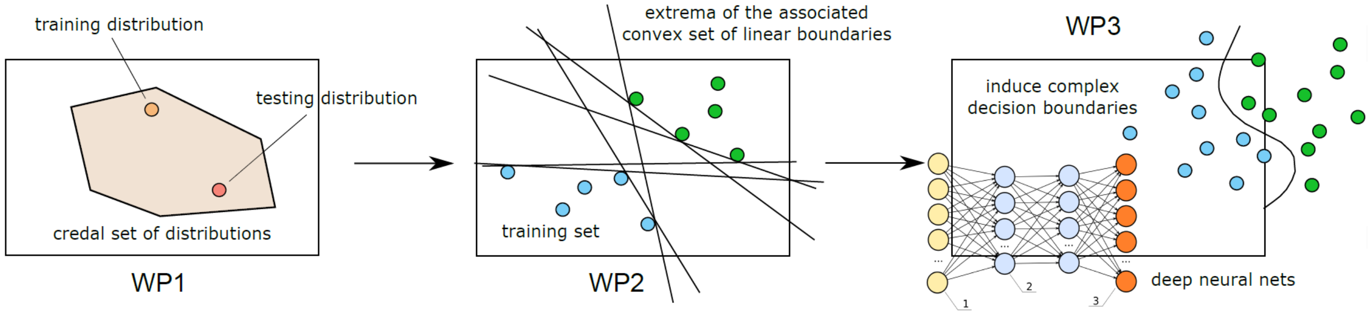

Figure 1. Methodology. Firstly (WP1), results from Statistical Learning Theory will be generalised to the case in which testing and training distributions are constrained to live in a convex set of probabilities. Secondly (WP2), the paradigm will shift from selecting a single model to selecting a convex set of models within a given class, based on the available training data. Finally (WP3), model classes associated with deep neural networks will be studied via under the new robust SLT.

We will then extend SLT under the relaxed assumption that training and test distributions are not the same, but come from a common convex set of probabilities [18] (Figure 1(a)), specifically in the form of a random set [20], as the latter are naturally associated with scarce information. Generalised concentration inequalities from random set theory may then provide bounds which hold under realistic circumstances, and therefore worst-case, cautious predictions which acknowledge the system's ignorance and produce AI agents aware of their own limitations [5]. QUB with Co-I De Campos and RA1 will lead the work in this area, but OBU's Cuzzolin will actively contribute as well.

WP2 – SLT with convex sets of models. The other paramount limitation of classical SLT is the focus on providing generalisation bounds for individual models, selected, given a training set, from an available class of hypotheses. An analysis of the simple case of learning a linear decision boundary from a training set of separable points, however, shows that the latter provides in fact a constraint on the location of the 'true' boundary, under the assumption that such a perfect model exists. As shown in Figure 1 (WP2), this yields a convex set of decision boundaries which can be efficiently manipulated in terms of just its extremal elements.

Unlike existing proposals [1,13], we focus on (convex) sets of models of the same kind (e.g., linear decision boundaries) learned from the very same data. This poses a number of interesting questions: (i) how do we measure the performance of a convex set of models? (ii) how do we show that, for instance, such convex set performs 'better' than the max-margin SVM learned from the same data? A theory of loss functions for convex sets of models will thus be developed, and empirically tested against classical machine learning approaches. OBU with the PI Cuzzolin and RA2 will lead the research in WP2.

WP3 – Understanding deep learning. Despite their enormous empirical success, deep neural networks are still rather poorly understood [26], due to the mathematical complexity of decision boundaries generated by series of convolutions [27]. Very recent work in the area tries to understand why rectified linear units and max-pooling perform better than other parameterisations [28]. Often, though, a post-hoc interpretation of the model is adopted - results are used together with the model to explain the latter, which can be misleading [29]. Deep rendering models [30] have been proposed to explain deep convolutional NNs, but do not cover the full generality of DL approaches. Recent workshops [26] mostly focus on analysing the optimisation methods and the fitting of yielded results, rather than trying to build a principled theory that can explain the process by which deep learning achieves its results.

We will build on these efforts to first derive traditional, precise-probabilistic generalisation bounds for deep neural models, starting from networks with a limited number of layers (as in [28]), to work our way up to more complex architectures. We will then apply the newly developed convex-theoretical statistical learning theory to deep learning, and derive generalisation bounds 'in the wild' for these models. Although OBU with PI Cuzzolin and RA3 will lead WP3, QUB's contribution will be crucial as well.

Validation

The new empirical bounds produced by our convex theoretical formulation of SLT, and their efficacy for model selection, will be empirically tested by designing appropriate performance measures. Standard machine learning benchmarks such as those available at UC Irvine, Kaggle, Stanford, Caltech, MNIST repositories, and challenges such as IJCNN www.ijcnn.org, European Conf on Computer Vision, and Kaggle competitions, will be utilised, so that direct comparisons can be made with other standard approaches. Code from existing approaches will be used when available, and standard techniques will be implemented by us in order to achieve fair and reproducible results.