Journal of Modern Accounting and Auditing, February 2022, Vol. 18, No. 2, 45-59
doi: 10.17265/1548-6583/2022.02.001

Everyone Wants to Play With *Stats*. Will Statistics Play a Key Role in an Everything-Is-AI World?

Maurizio Sanarico
SDG Group, Milan, Italy

What is the role and place of statistics in modern artificial intelligence and machine learning? This paper surveys some currently fashionable topics in Artificial Intelligence (AI) and related fields, namely Machine Learning (ML). Moreover, it tries to outline how statisticians can contribute to AI. The paper reflects personal opinions and perspectives I gained working as a Chief Data Scientist in an international consulting firm that focuses on data analytics.

Keywords: statistics, Artificial Intelligence (AI), real-world applications, subject mixing

Introduction

In this paper, the context is machine learning models deployed in a production environment to support everyday business. These models are currently used to face a series of complex problems. For instance, complex deep neural architectures allow forecasting the power demand at the smart-meter level for utilities distributing gas or electricity. Dynamic cohort approaches are used to predict customer lifetime value. Reinforcement learning can optimize bioprocesses and aid in implementing dynamic treatment regimes. Predictive models are used to obtain a preview estimate of margins from different type of products and services in banks, both for financial monitoring and to put in place corrective actions when needed. However, these examples do not refer to situations in which the statistical profession has a well-defined role, such as clinical trials design and analysis or statistical quality control. I aim to provide a broad selection of current research lines and machine learning applications, highlighting the role of statistical science.

The recent rebirth of machine learning and artificial intelligence as fashionable topics, with a strong emphasis on technology, elicits a question: what is, or could be, the role of statistics in this context? Of course, I am referring to statistics as a specific subject and statisticians as specific professionals because statistical methods are pervasive, and *everyone wants to play with stats*. Is the statistical profession destined to become irrelevant in the context of ML? I found an analogy looking at mathematics translated into engineering. When engineering a mathematical solution, we tend to forget the underlying theory: the focus shifts on the physical object rather than the first principles motivating the mathematical theory. As Chief Data Scientist of SDG Group, I constantly face the tension between technology-oriented and scientific-oriented thinking. The main effort is to merge these approaches to a balanced synthesis, with the additional goal to produce sustainable and profitable business results. Statistics found its way as the backbone to strengthen the scientific method in disciplines where it was already used and extend it to other, most difficult disciplines. The statistical science on

Maurizio Sanarico, MSc, SDG Group, Milan, Italy.

the practical ground develops methods to quantify uncertainty in mathematical models based on data and make inferences leveraging these models. I try to stand at the crossroad between business and scientific viewpoints.

Production-level AI models support many real-life applications. However, classical statistical modeling entails a process and a methodology that still is essential in modern machine learning pipelines. Thus, among other disciplines, statistics is a key subject in the present and future development of ML. See Ball (2015) for an interesting analysis about data science and statisticians.

Why Should I Trust you? Interpret, Explain, and Establish Causality

One of the most impressive facts in recent years has been the renaissance of (deep) neural networks and the swift growth of deep learning, which is not only motivated by the scientific relevance of the subject *per se* but also by its undeniable flexibility in applications. Deep learning builds complex nonlinear maps from input to a target variable by combining multiple latent input representations. Without entering the details of the various types of deep neural networks (e.g., convolutional, recurrent, graph neural networks, and combinations of these architectures) commonly used in both supervised and unsupervised problems, it is worth noticing that many statistical issues in deep learning remain unexplored. The first issue, essential to many applications, concerns interpretability, explainability, and causal inference. A second related aspect is that of uncertainty assessment and inference.

Interpretability, explainability, and causal inference are related aspects of the requirement of humans to understand the results produced by complex models. Several methods cover the three aspects mentioned before, e.g., Locally Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP). However, what skills are required to interpret and understand the explanations yielded by the model? The target audience is the data science and data analysis communities, not final users not acquainted with working with data from a statistical viewpoint. Indeed, complex models are not directly interpretable and explainable even to the eye of the researcher that devises them or for the data scientist that builds a specific model for a given case. In my experience, when interpreting the model's output is a requirement, the explainability layer offered by the model has to be integrated with specific documentation and formation to allow final users (e.g., medical doctors) to gain insight into results. This latter observation is a critical success factor in applications, regardless of the model's performance. In this realm, *statistical science*, which has a unique tradition and expertise in communication and working with the scientific and business communities, might have a significant role in improving interpretation and explanation of models, creating interfaces directed to final users, and finally, establishing internal and external validity of the results. Recall that internal validity means structural validity, for instance study design, and external validity means applicability of results to more general contexts.

Miller (2019) describes the explainability issue as follows: "The very experts who understand decision-making models the best are not in the right position to judge the usefulness of explanations to lay users" (p. 4)—a phenomenon that Miller (2019) refers to (paraphrasing Cooper (2004)) as "the inmates running the asylum". Therefore, a strong understanding of how people define, generate, select, evaluate, and present explanations seems almost essential.

Devising ways to communicate results obtained through complex models to final users would be a step further towards providing users with a better understanding and acceptance of results. Hopefully, it should avoid the "I want a significant *p*-value attitude" so deeply rooted in scientists and scientific journal (especially medical) reviewers. Moreover, when addressing the problem of AI explainability, it is important to distinguish between *local* explanations, which try to tell us why particular facts occurred, and explanations establishing causality on a more global level, thus related to more general relationships, such as those involved in the scientific method.

Explainability involves different actors with different goals and perspectives, the picture below, gives an example well-suited for the financial and pharmaceutical industry.

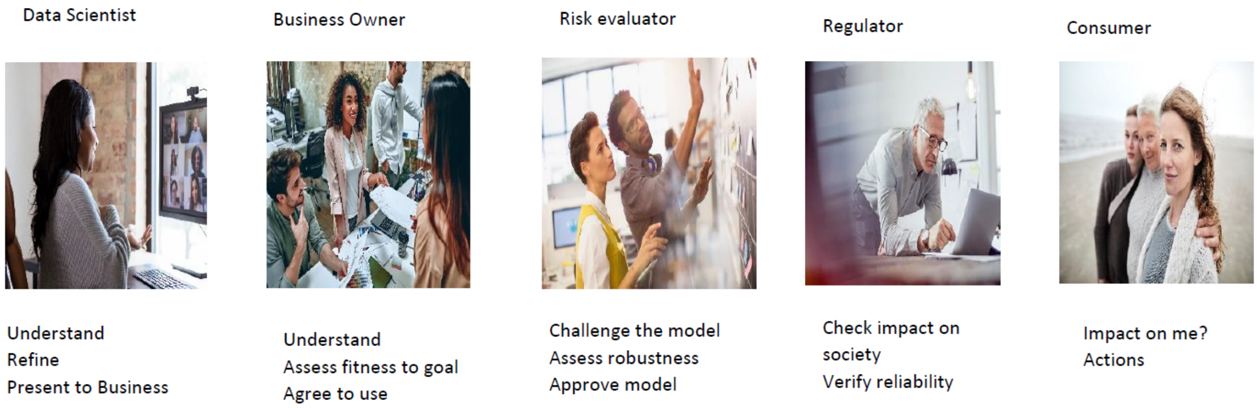


Figure 1. AI according to professional role.

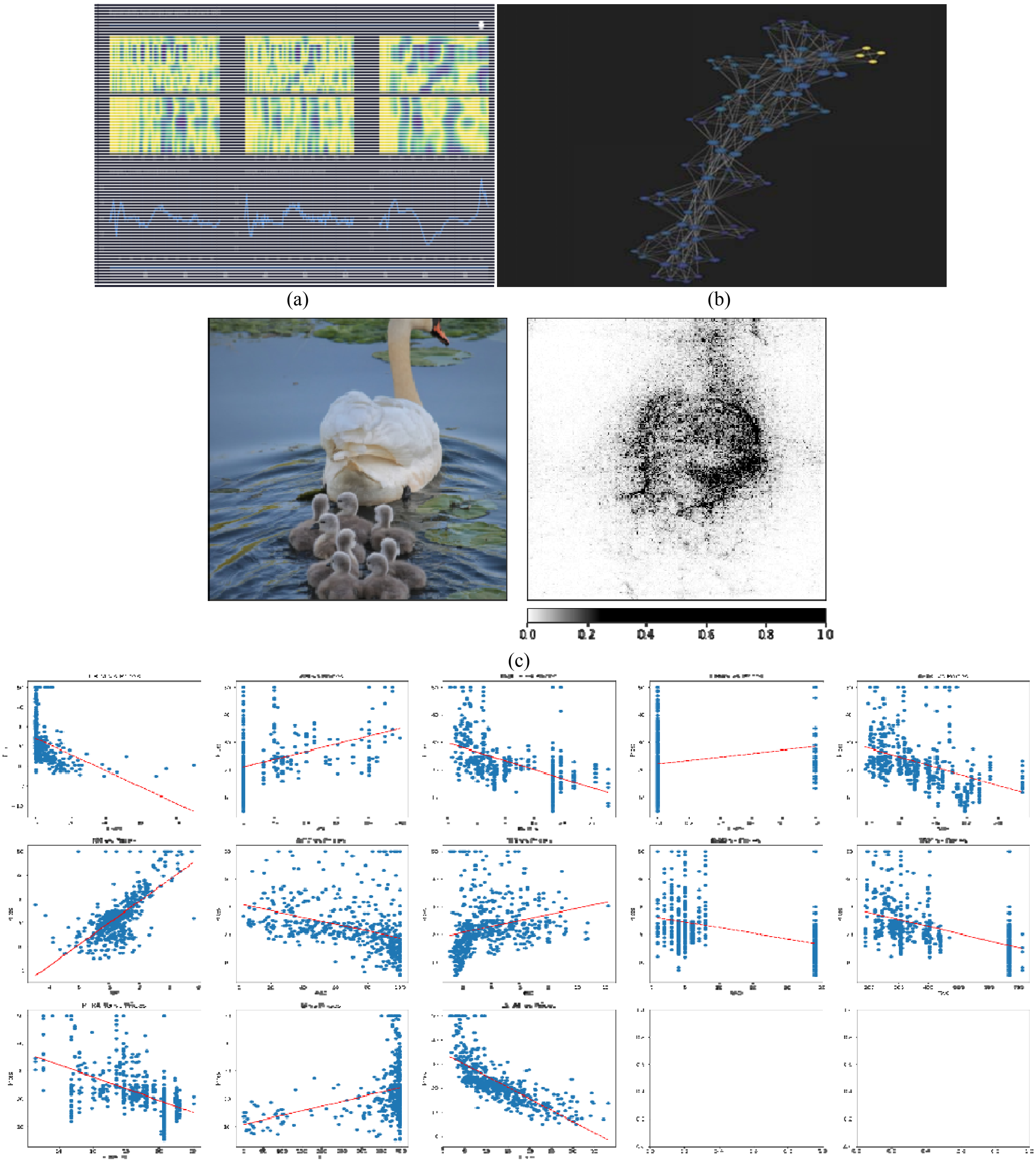
Many explainable methods proposed tend to miss a proper uncertainty quantification. Statistics have already successfully contributed to modern data analysis. For instance, statistical tests and uncertainty quantification in Topological Data Analysis have been introduced by Wasserman (2018).

Thus, in the world of AI and big data—somewhat dominated by software development and coding—statistical science can have a central role in Machine Learning operations (MLOps), and data science leveraging methods rooted in the statistician mindset to solve interpretability problems in original and principled ways.

In current ML pipelines, the most common way to carry out causal inference is by using counterfactuals. However, defining a counterfactual in observational data with hundreds of variables is not straightforward. See Chou, Moreira, Bruza, Ouyang, and Jorge (2021). Miller (2019) describes the missing link between the current research on explanations from the fields of philosophy, psychology, and cognitive science. According to him, there are three main aspects that AI systems have to achieve full explainability: (1) people seek explanations that answer the following question: *why some event happened, instead of another?* i.e., counterfactual explanations; (2) recommendations can focus on a selective number of causes (not all of them), which suggests the need for causality in AI (the user should not be overwhelmed by potential causes); and (3) explanations should consist in conversations and interactions with a user promoting user engagement.

Figure 2 proposes some examples of explanatory methods. Figures 2a, 2b, and 2c present methods where explainability is for experts, although, once provided with the right hints, the method in Figure 2b could be understood also by users that domain experts, rather than data scientists or statisticians. The method presented in Figure 2d is much more accessible also to general users.

A taxonomy of explainable artificial intelligence has been proposed by Belle and Papantonis (2021).



(a) (b)

(c)

(d)

Figure 2. Examples of explanatory methods: (a) How a Deep Convolutional Graph Neural Network classifies complex multidimensional time series and recognizes an anomaly regarding one component time series, shows such component, and a saliency map highlights the activation patterns of nodes in the network; (b) The same data represented as a topological network, a connected set of clusters (theoretically the nerve of a simplicial complex) through the Mapper algorithm. Inspecting the yellow nodes, that are coloured according to a variable of interest X, may be much easier to understand some key feature for a subject-matter scientist/practitioner; (c) Integrated gradients method computes the integral of the gradients of the output of the model for the predicted class with respect to the input image pixels along the path from the black image to our input image; (d) Marginal effects on predictor in a multivariable model.

An interesting branch is causal transfer learning (Mooij, Magliacane, & Claassen, 2020), based on a paradigm called Joint Causal Inference and Structural Causal Models. Another approach is that of *deconfounder* proposed by David Blei and Yixin Wang (Wang & Blei, 2020). Both methods are suitable for multiple cause problems.

Another interesting approach to causal analysis is the Targeted Minimum Loss Estimation framework. This framework assumes that no model is best when working with complex data sets. Therefore, first, an ensemble method (SuperLearner) is proposed. Then, a plug-in estimator is obtained for the causal parameter. The ensemble is not composed by many perturbed executions of the same algorithm, as occurs, for instance, in random forest, gradient boosting machine, and XGboost, but by a set of algorithms selected just because of their heterogeneity. The key idea is that such diverse algorithms could best represent different aspects of the data. We improve the performance of each method by considering a weighted combination of the ensemble. Weights are obtained via a generalized cross-validation process. An example of a model set might be random forest, logistic regression, naïve Bayes, K-NN, Xgboost, and CART.

In more detail, when considering interpretability and explainability, we can consider three main perspectives, namely, *predictive accuracy*, *descriptive accuracy*, and *relevancy*. Predictive accuracy is easy to measure using consolidated methods and metrics, e.g., classification matrices (confusion matrices) and ROC curves. Descriptive accuracy can be described as the degree to which an interpretation method objectively represents the relationships learned by the ML models. Relevancy: An interpretation is relevant if it provides insight into a particular user type and given problem.

While descriptive accuracy is currently a popular topic: many researchers are trying to devise methods to understand better the representation learned by the model, in particular, with some new approaches based on advanced mathematics, as we shall discuss in the next section. Relevancy is harder to define and formalize. The long and somewhat unique consulting practice that characterizes the statistical profession might contribute to a better understanding of complex models, especially when considering relevancy.

In summary, complex models require specific tools and methods to allow data scientists and analysts to disentangle, even partially, how and why the models behave in a determined fashion. Most of the time, final users are not in a position to use the output of such tools. Still, conveying the message to them is essential for the acceptance and real-world applicability of the deep learning models. I believe statistical methods can be fundamental in paving the road to new methods improving the existent explainability techniques, and creating original interfaces to communicate the results yielded by these techniques.

Subject-Mixing and Deep Math for Deep Learning

Subject-mixing is not simply the fact that many different skills and backgrounds contribute to machine learning/AI (e.g., computer science, statistics, mathematics), but that heterogeneous approaches are necessary to justify and better understand the dynamics underlying complex models. In this endeavor, pure mathematics is making its appearance. See, for instance, “Deep Learning: A Statistical Viewpoint” (Bartlett, Montanari, & Rakhlin, 2021) and “Geometric Deep Learning: Grids, Groups, Graphs Geodesics and Gauges” (Bronstein, Bruna, Cohen, & Velickovic, 2021). The first sentence of the text clearly explains the perspective of the paper:

> While learning generic functions in high dimensions is a cursed estimation problem, most tasks of interest are not generic, and come with essential pre-defined regularities arising from the underlying low-dimensionality and structure of the physical world. This text is concerned with exposing these regularities through unified geometric principles that can be applied throughout a wide spectrum of applications. (Bronstein et al., 2021, p. 8)

The blueprint of geometric deep learning is summarised in the following scheme, where each group provides the invariance/equivariance property relevant to the specific neural network architecture.

Table 1

Main Geometrical and Algebraic Associated With Deep Neural Networks Architectures

NN Architecture	Domain Ω	Symmetry Group G
CNN	Grid	Translation T
Spherical CNN	Sphere/SO(3)	Rotation SO(3)
Intrinsic/Mesh CNN	Manifold	Isometry Iso(Ω)/Gauge Symmetry SO(2)
GNN	Graph	Permutation Σ_n
Deep Sets	Set	Permutation Σ_n
Transformer	Complete Graph	Permutation Σ_n
LSTM	1d Grid	Time warping

The key notions in what is called Geometric Deep Learning consist of the geometric principles of Symmetry, Geometric Stability[1], and Scale Separation. In descriptive terms, they imply that while the target function f might depend on complex interactions between features over the whole domain, leveraging locally-stable functions, it is possible to *separate* the interactions across scales. This separation first focuses on localized interactions that are then propagated towards coarser scales. Convolutional filters followed by pooling layers are an example of a multi-scale decomposition-reconstruction process. These principles are analogous to those underlying the multi-scale analysis in low-dimensional problems, the prototype being the wavelets that overcome limitations of Fourier analysis in representing local invariance.

More formally, the family of *linear equivariant operators*[2] provides a powerful tool since it enables the construction of rich and stable features by composition with appropriate non-linear maps. Let us consider a symmetry group G, a data domain Ω, if B: X(Ω,C) → X(Ω,C') is G-equivariant, i.e., B(g.x) = g.B(x) for all x ∈ X and g ∈ G and σ: C' → C" is an arbitrary (non-linear) map, then we can verify that the composition U: = (σ o B): X(Ω,C) → X(Ω,C") is also G-equivariant, where σ: X(Ω,C') → X(Ω,C") is the element-wise instantiation of σ given as (σ(x))(u): = σ(x(u)).

This simple property allows us to define a very general family of G-invariants by composing U with the group averages A o U: X(Ω, C) → C". A natural question is thus whether any G-invariant function can be approximated at arbitrary precision by a specified deep neural network model for appropriate choices of B and σ. It is possible to adapt the standard Universal Approximation Theorems from unstructured vector inputs to show that shallow "geometric" networks are universal approximators by properly generalizing the group average to a general non-linear invariant. However, in analogy with the case of Fourier versus Wavelet invariants, there is a fundamental tension between shallow global invariance and deformation stability. This

[1] With real data, we cannot speak of exact invariance and equivariance under group actions, so the weaker notion of *deformation stability* (or *approximate invariance*) is introduced:
where ( )x(u) = x($^{-1}$u) as before, and where C is some constant independent of the signal x. A function f ∈ F(X( )) satisfying the above equation is said to be *geometrically stable*.

[2] A function f: X( ) → Y is said to be *G-invariant* if f((g)x) = f(x) for all g ∈ G and x ∈ X, that is the group action is not affecting the output of f. An example of invariance is *shift-invariance*, arising in computer vision and pattern recognition applications such as image classification. A function f: X( ) → X( ) is *G-equivariant* if f((g)x) = (g)f(x) for all g ∈ G, that is, group action on the input affects the output in the same way. A Convolutional layer of CNNs is not shift-invariant but *shift-equivariant.* Then, image classification is usually implemented as a sequence of convolutional (shift-equivariant) layers, followed by global pooling (which is shift-invariant).

tension motivates an alternative representation, which considers instead *localized* equivariant maps. Assuming that Ω is further equipped with a distance metric d, we call an equivariant map U localized if (Ux)(u) depends only on the values of x(v) for Nu = {v: d(u,v) ≤ r}, for some small radius r; the latter set Nu is called the *receptive field*.

Summarising, the geometry of the input domain, with knowledge of an underlying symmetry group, provides three key building blocks: (i) a local equivariant map, (ii) a global invariant map, and (iii) a coarsening operator. These building blocks provide a rich function approximation space with prescribed invariance and stability properties by combining them in a scheme referred to as the *Geometric Deep Learning Blueprint.* In practice, composing equivariant functions with coarsening operators in a sequence of layers provides the high predictive accuracy of deep neural networks in a wide range of problems. This predictive ability is an apparent violation of Occam's principle and of common view that overparameterized models are going to overfit data, an interesting point of view to be further explored is that this is due to geometric structures imposed by the network architecture.

The statistical view takes another perspective that is illustrated using one of the introductory sentences of a paper of Bartlett et al. (2021) providing a statistical view of deep learning:

> The second surprising empirical discovery was that these models are indeed outside the realm of uniform convergence. They are enormously complex, with many parameters, they are trained with no explicit regularization to control their statistical complexity, and they typically exhibit a near-perfect fit to noisy training data, that is, empirical risk close to zero. Nonetheless, this overfitting is benign, in that they produce excellent prediction performance in a number of settings. Benign overfitting appears to contradict accepted statistical wisdom, which insists on a trade-off between the complexity of a model and its fit to the data. Indeed, the rule of thumb that models fitting noisy data too well will not generalize is found in most classical texts on statistics and machine learning. (Bartlett et al., 2021, p. 3)

Here the justification is sought in statistical arguments about implicit regularization, VC-dimension, and functional analysis (Reproducing Kernel Hilbert Spaces).

The key points are expressed with a sentence contained in the abstract of the paper of Bartlett et al (2021): "We conjecture that specific principles underlie these phenomena: that over-parametrization allows gradient methods to find interpolating solutions, that these methods implicitly impose regularization, and that over-parametrization leads to benign overfitting, that is, accurate predictions despite overfitting training data" (p. 1).

Belkin (2021) treats this problem from a more formal perspective, while Dar, Muthukumar, and Baraniuk (2021) give a survey of the topic. Belkyn distinguishes between under-parameterized models with limited complexity and where the optimization landscape is locally convex around local minima, and overparameterized models where there are manifolds of potential *interpolating predictors* that fit the data exactly. In the latter case, the statistical question is understanding the nature of the *inductive bias*, i.e., what makes some solutions preferable to others despite all of them fitting the training data equally well. In interpolating regimes, non-linear optimization problems generically have manifolds of global minima. In most deep learning models optimization is non-convex, even locally, however, often they satisfy the so-called Polyak-Łojasiewicz (PL) condition guaranteeing convergence of gradient-based optimization methods. The peculiarity is pictorially presented as the so-called double-descent generalization curve.

What is the Polyak-Łojasiewicz (PL) condition? It is a sufficient condition for efficient minimization by gradient descent. The PL condition is a simple first-order inequality applicable to a broad range of optimization problems. L(w) is μ-PL if the following holds:

$$\frac{1}{2}\|\nabla\mathcal{L}(\mathbf{w})\|^2 \geq \mu(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}^*))$$

PL-condition applies to curved manifolds of minimizers. However, in this formulation, the condition is non-local. While convexity can be verified pointwise by ensuring that the Hessian of L is positive semi-definite, the PL condition requires "oracle" knowledge of L(w*). The problem is addressed by removing the L(w*) term.

$$\frac{1}{2}\|\nabla\mathcal{L}(\mathbf{w})\|^2 \geq \mu\mathcal{L}(\mathbf{w})$$

So that the PL* condition in a ball of sufficiently large radius implies the existence of an interpolating solution within that ball and exponential convergence of gradient descent and, indeed, stochastic gradient descent.

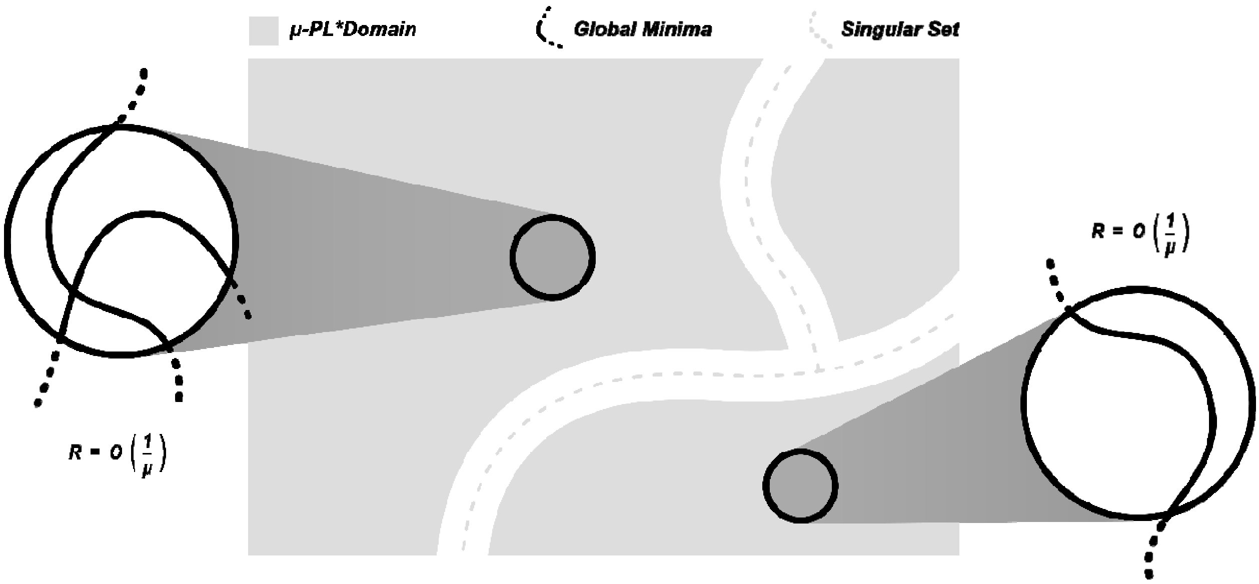

μ-PL*Domain
Global Minima
Singular Set
$R = O\left(\frac{1}{\mu}\right)$
$R = O\left(\frac{1}{\mu}\right)$

Figure 3. Illustration of PL* condition.

Indeed, a phenomenon called double-descent is observed in the test error of many deep neural networks: performance first improves, then deteriorates, and then improves again with increasing model size, data size, or training time. Double descent is yet not fully understood, further study of this phenomenon is an important research direction.

The double-descent phenomenon does not occur in all deep neural networks and other complex models. It is plausible that what determines the inductive bias allowing for selecting a good parameterization in a manifold of solutions is the presence of a geometric regularity of the type previously explained. It might be interesting to explore the behavior of models in data where geometric properties are likely to be present and others where such properties are much less probable or nonexistent at all.

Another interesting path in explainability is the thread of research papers falling under the umbrella of semantic explainability (Doran, Schulz, & Besold, 2017; Donadello & Serafini, 2016).

Another interesting example of how deep learning has inspired and fostered looking for achieving better understanding and constructive methods for neural networks comes from a series of papers by Bergomi, Frosini, Giorgi, and Quercioli (2019), Vertechi, Frosini, and Bergomi (2020). In these papers, tools from category theory and functional analysis are borrowed to provide a formal approach to building neural networks using compositions of smaller building blocks.

Is there any chance to build a common ground for those approaches since they might be looking at the same things using different tools and perspectives? Recall that G. Perelman proved the Poincaré conjecture and the geometrization conjecture of Thurston, two topological-geometric problems, using analytical methods based on the Ricci Flows.

Subject-mixing implies different mindsets, is one of them prevailing? Is it unavoidable? Or a better balance creating a "collective mindset" or "consensus mindset" will be most productive and fair? In the following, I touch on some relevant points to the statistical practice in the business environment.

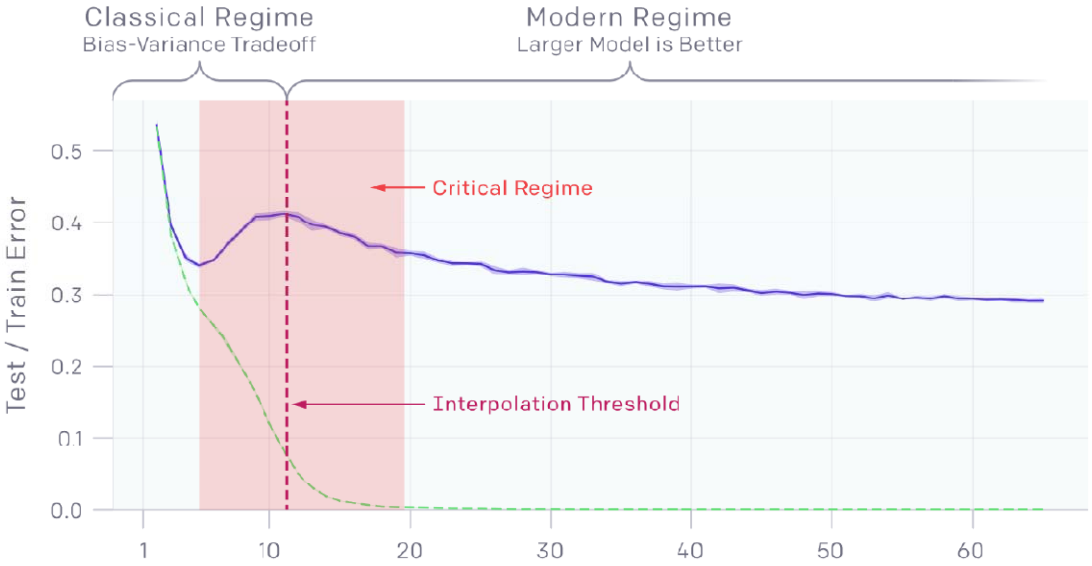


Figure 4. Double descent generalization curve. The interpolation threshold separates modern and classical regimes. Train (blue line) and test errors (green line) versus model capacity.

MLOps (Machine Learning Operations)

At its core, MLOps is the standardization and streamlining of the machine learning-based models lifecycle. In other terms, MLOps is a systematic approach to building, deploying, and monitoring machine learning (ML) solutions. It is an engineering discipline that can be applied to various industries and use cases. Software development is interdisciplinary and is evolving to facilitate ML. MLOps is an emerging method to fuse ML with software development by integrating multiple domains as MLOps combines ML, DevOps, and data engineering, which aims to build, deploy, and maintain ML systems in production reliably and efficiently. Thus, MLOps can be expounded by this intersection. Ironically, statistics is not usually mentioned in MLOps, while MLOps are mostly based on statistical learning in real-world business processes. Of course, engineering also occupies a central role in MLOps, e.g., data management practices and software development. Still, MLOps' core component is often a coordinated set of forecasting models to produce mass forecasting of products' popularity or a clustering algorithm to build and maintain a customer segmentation. In my experience, the best results can be obtained by working as an integrated team since the beginning of any activity that aims to build a production system. Data scientists must be aware of the constraints and practicalities implied in a production ML pipeline. Data engineers and MLOps specialists need to understand the requirements and the features of the models developed and tested by data scientists. The tension between the engineering and data analytics (statistical) ways was in a certain sense anticipated by the famous paper of Breiman (2001).

A proposal I would like to present shortly is *Bayesian workflow* (Gelman et al., 2020) for ML-models development cycle and maintenance process. This method covers the core components of the MLOps workflow, citing the definition in the paper: “Bayesian workflow includes the three steps of model building, inference, and model checking/improvement, along with the comparison of different models, not just for the purpose of model choice or model averaging but more importantly to better understand these models” (Gelman et al., 2021, p. 3). In the paper, an extended Bayesian workflow is also mentioned. This workflow includes pre-data design of data collection and measurement and after-inference decision making. This latter task can be very complex in real-world business environments because it would imply being able to have access to the data collection phase.

Deep Learning and Statistical Inference

Deep probabilistic programming is already supported by Python libraries such as Pyro, Tensorflow probability (Edward2), PyMC. Without entering into details, it is essential to highlight the necessity of introducing uncertainty and inference in generic classes of ML models, particularly in deep learning. Currently, this is achieved via Markov Chain Monte Carlo or variational methods. These algorithms’ convergence and theoretical behavior applied to these hyper-parameterized models can be a fertile study ground for computational statisticians.

AutoML

In the current landscape of machine learning tools, there is another trend called AutoML.

> Machine learning (ML) has achieved considerable successes in recent years, and an ever-growing number of disciplines rely on it. However, this success relies on human-machine learning experts to perform manual tasks. As the complexity of these tasks is often beyond non-ML-experts, the rapid growth of machine learning applications has created a demand for off-the-shelf machine learning methods that can be used easily and without expert knowledge. We call the resulting research area that targets progressive automation of machine learning AutoML. (www.automl.org)

Fully-fledged commercial products offer this feature and free, open-source tools: DataRobot, H20 to mention two of the most successful examples, but also cloud providers such as AWS, Google Cloud Platform, and Microsoft Azure, and hundreds of newcomers and start-ups. Among the most known non-commercial solutions, we could mention Tpot, AutoKeras, Auto-SkLearn, NNI (Neural Network Intelligence), and many others. These libraries and services rely on heuristics (e.g., Tpot uses genetic algorithms to search in the model space defined by the type of supervised problems) and grid search. Is it possible to inject statistical wisdom within such tools? What is not easy to do and still missing is the ability to critically appraise the results. Can a non-expert user leverage explainable methods afflicted, at least partially, by “the inmates are running the asylum” problem and have a significant impact on business processes?

Active Learning, Human-in-the-Loop, Expert-Augmented Machine Learning

These topics are related and aim to embed human-expert knowledge in the machine learning process to refine and validate results. Human-in-the-loop approaches and active learning are based on uncertainty and diversity sampling, and combination. The two types of sampling strategies aim to improve the model’s performance. Uncertainty Sampling is a strategy for identifying unlabeled items in the proximity of the decision boundary learned by an ML model. Indeed, these points have a higher chance of being misclassified. Diversity Sampling is a strategy for identifying unlabeled items unknown to the ML model in its current state. This strategy typically identifies items that contain combinations of feature values that are rare or unseen in the training data. The aim is to obtain a complete picture of the problem space. For a comprehensive presentation of the subject one could consider the book of Munro (2020).

Another interesting approach has been recently proposed in the statistical community: Expert-Augmented Machine Learning. We will not delve into the details and simply provide the schematic representation proposed in (Gennatas, Friedman, Ungar, Pirracchio, Eaton, Reichmann, Interian, Luna, Simone, Auerbach, Delgado, van der Laan, Solberg, & Valdes, 2020). In brief, RuleFit algorithm extracts a set of rules, experts rank the rules using judgment, the ranking is compared with the empirical ranking, rules that are in strong disagreement in the two rankings are penalized.

Digital Twins, Surrogates

A *surrogate* is a substitute for the real thing. In statistics, drawing from predictive equations derived from a fitted model can surrogate for the data-generating mechanism. If the fit is good-model flexible yet well-regularized, data-rich enough, and fitting scheme reliable, then such a surrogate can be valuable. Gathering data is expensive, and sometimes getting precisely the data one wants is impossible or unethical. A surrogate could represent a much cheaper way to explore relationships and perform, for instance, what-if analysis. How do surrogates differ from ordinary statistical modeling? One superficial difference may be that surrogates favor faithful yet pragmatic reproduction of dynamics over other statistical models: interpretation, establishing causality, or identification. As one might imagine, that characterization corresponds to an oversimplification of the problem at hand.

A digital twin is a virtual representation of an object or system that spans its lifecycle, is updated from real-time data, and uses simulation, machine learning, and reasoning to help decision-making.

This process means creating a highly complex virtual model that is the *exact* counterpart (or twin) of a physical thing. The “thing” could be a car, a building, a bridge, or a jet engine. Connected sensors on the physical asset collect data and send them to the virtual model. Anyone looking at the digital twin can now access crucial information about the state of its real counterpart.

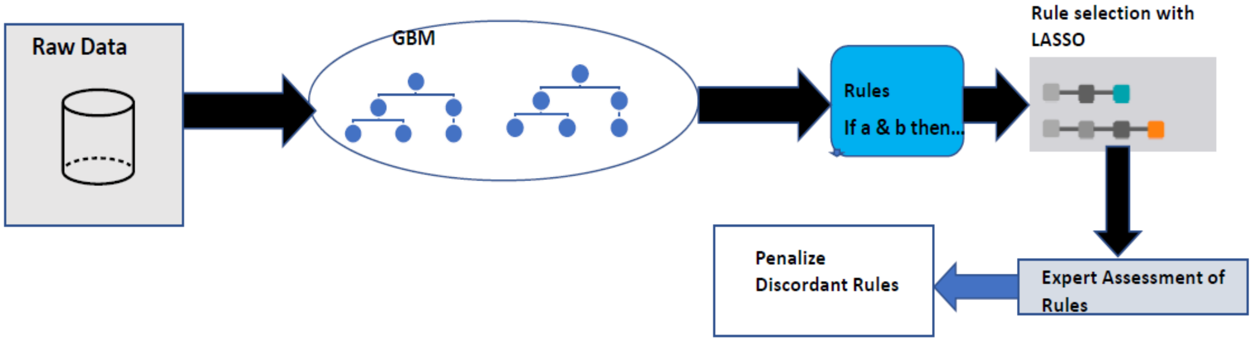

Raw Data
GBM
Rules
If a & b then...
Rule selection with LASSO
Expert Assessment of Rules
Penalize Discordant Rules

Figure 5. Expert-Augmented Machine Learning framework illustrated.

Why Should I Trust You? Cognitive Science and Psychology Perspective

This section reconnects with Section 2, first, from a more practical point of view. Then, I shall leverage a more theoretical perspective based on the already cited work of Miller in the attempt of “closing the loop” started in the sections above.

When is deep understanding and acceptance of field experts most relevant? Short-term operational predictions in non-life-threatening and non-disruptive applications (no time to assess, just good if right and judged on the practical ground or based on subject-expert acceptance). Examples are image and video classification, automatic adjustment of heating parameters in a district heating network, and other automatic actions in a smart city context.

On the other hand, if we consider predictions in life sciences, these cannot be accepted if not understood and justified using biological or medical arguments. The same occurs in manufacturing and other fields where the final responsibility of decisions and consequent actions are to be taken by humans.

Let us consider a real case in pharmaceutical production. Modern pharmaceutical manufacturing processes benefit from many gauging instruments and sensors to monitor process variables and the quality and characteristics of the products. Process engineers can understand significant problems associated with a limited number of variables obtained by such tools. The base of a new project was the idea that some process executions deviate from the best conditions because of complex interactions between many variables or unforeseen conditions. The first step was to define the "best conditions" for a given process, besides the fact that according to Good Manufacturing Practices, processes are to be validated, and strict constraints apply to many outputs of the process. A pragmatic approach was taken: the best conditions were defined by first considering the set of parameters' values during the process executions that gave correct output. Then, each variable's median trajectory was set as the "reference conditions".

The discrepancy from reference conditions was computed in two ways: (1) using a longitudinal version of the k-means clustering to detect differences in values; (2) using a permutation distribution clustering to detect differences in shape. A significant achievement was to discover anomalous production cycles that showed no individually discernible deviation from reference but were instead associated either to multiple unexpected deviations or other inconsistencies that were not anomalous on a metric scale and were associated with strange shapes of the process variables (e.g., small oscillations around a perfectly normal value). The results were interesting from a data-analysis viewpoint, but to make them acceptable and useful to final users (process engineers), we had to devise and share with them a way to represent the results able to point out actionable versus not-actionable problems.

On the other hand, in the case of predicting the class associated with an object using image recognition algorithms, the user is not usually interested in understanding the patterns of pixels behind the classification process. On the contrary, data scientists can be interested in achieving such an understanding. In applications such as image classification, video classification, and text categorization, a concept has emerged: transfer learning and domain adaptation. While transfer learning is instrumental in the just-mentioned cases, a question that could come to mind is whether it could be useful in other fields. The availability of large pre-trained neural networks like BERT, GPT-3, made a strong case for transfer learning.

From more cognitive and social perspectives, see Miller (2019), I found interesting the analysis placed in the light of explainable AI. Its bases are (1) explanations are contrastive, based on counterfactual cases; (2) explanations are selected in a biased manner; (3) probability could do not matter; (4) explanations are social—they are a transfer of knowledge, presented as part of a conversation or interaction, and are thus presented relative to the explainer's beliefs about the beliefs of the explanation's recipient. So, according to Miller (2019):

> These four points all converge around a single point: explanations are not just the presentation of associations and causes (causal attribution), they are contextual. While an event may have many causes, often the explainee cares only

> about a small subset (relevant to the context), the explainer selects a subset of this subset (based on several different criteria), and explainer and explainee may interact and argue about this explanation. (p. 7)

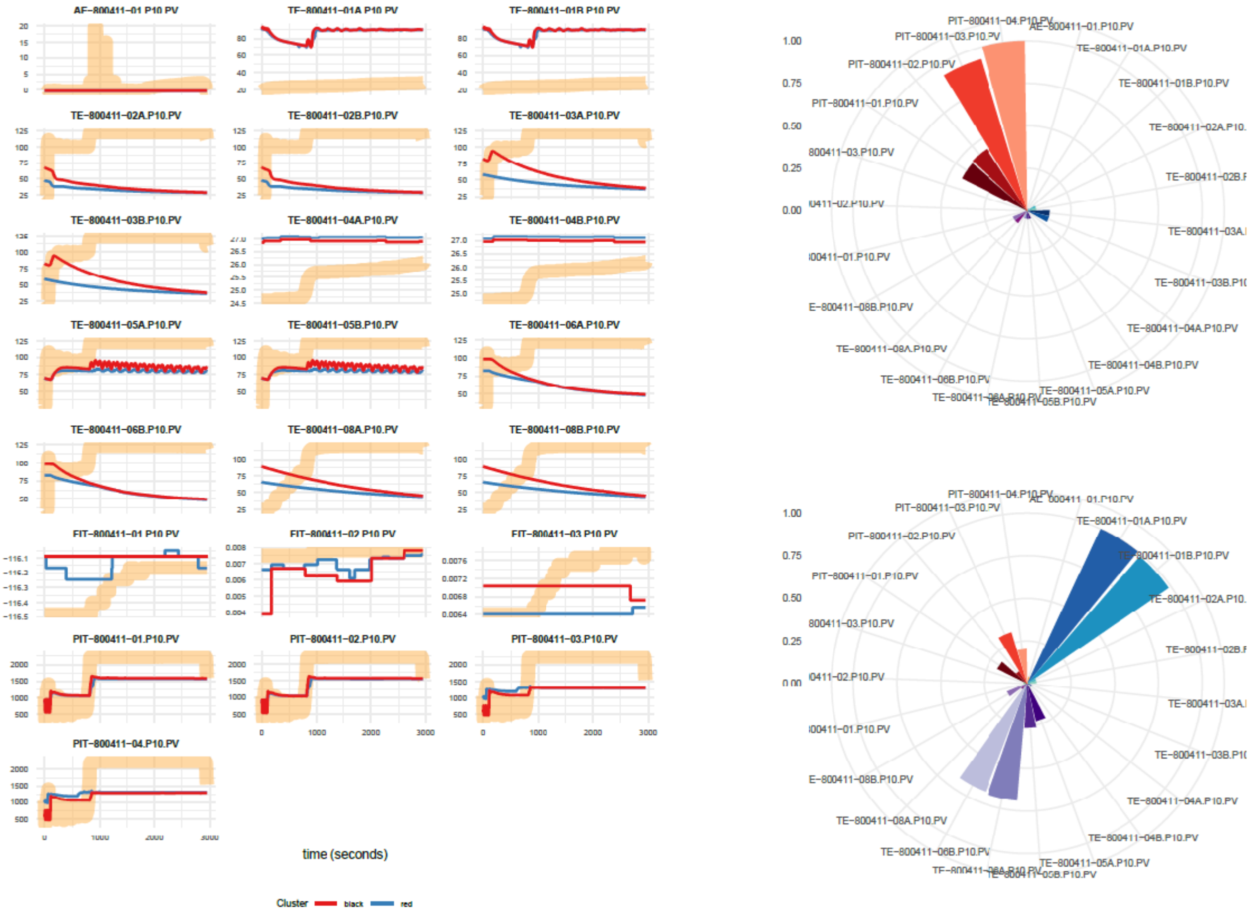


Figure 6. Explain to process engineers.

I am currently investigating practical problems, like those described in this section, taking into account the four points mentioned above and the implied way to cope with explanation and understanding transfer. This task is far from easy but having people working on this topic with the right skills and empowerment will avoid or limit the “inmates are running the asylum” condition that, just like software products is currently afflicting AI and ML, nevertheless the claims of XAI, fairness, and so on.

The Business Value

In this section, I am going back to my prevalent point of view: maintaining a balance between validity and business value. In a consulting firm, the business value is twofold: we have to produce value for our clients and, at the same time, for our company.

The value comes from good technical performance, for example, a good forecast or prediction, but many other features are important and aimed at specific users. Let us consider an example based on a real project. We used log data and client tickets to predict new tickets. The client stated the initial goal simply as a prediction problem. However, during the project, the goal had to be redefined more appropriately and turned out to be composed of several sub-objectives tailored to different users and stakeholders. Top Management is interested primarily in cost-saving. However, the ability to anticipate problems affecting final clients also reduces complaints and churn rate, although the latter is not easily measurable and, in any case, it would be measurable only on a long-time horizon. Network managers are interested in predictions and diagnostics about the cause of the problem. Thus, generating value involves translating client tickets associated with a given equipment, recognizing the problem, and establishing whether it could be fixed remotely (a software problem) or requires onsite maintenance. When a specific pattern can be associated with a well-defined problem, the network managers will ask to perform automatic diagnostic and, whenever possible, fix the issue with an appropriate piece of software. All these elements are required for a successful project: scientific validity, performance, interpretability, optimization of the network managers' activities, economic value (cost savings and streamlined maintenance operations), reduction of customers' complaints.

Conclusions

I have no ultimate solutions or clear views of the future. In this broad excursion, I hope to have provided some arguments to discuss and reflect on and pointers to current active research topics.

References

Ball, T. (2015). *Paradoxes of data science*. Retrieved from www.kdnuggets.com/2015/08/paradoxes-data-science.html

Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numerica, 30*, 87-201.

Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica, 30*, 203-248.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in Big Data, 4*, 39. doi:10.3389/fdata.2021.688969

Bergomi, M. G., Frosini, P., Giorgi, D., & Quercioli, N. (2019). Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning. *Nature Machine Intelligence, 1*, 423-433.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., … Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research, 20*(28), 1-6.

Bodnar, C., Frasca, F., Wang, Y. G., Otter, N., Montúfar, G., Liò, P., & Bronstein, M. (2021). Weisfeiler and lehman go topological: Message passing simplicial networks. In *Advances in neural information processing systems 34 pre-proceedings (NeurIPS 2021)*.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statist. Sci., 16*(3), 199-231.

Bronstein, M. M., Bruna, J., Cohen, T., & Velickovic, P. (2021). Geometric deep learning: Grids, groups, graphs, geodetics and gauges. arxiv.org/abs/2104.13478.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine, 34*(4), 18-42.

Carlsson, L. S., Vejdemo-Johansson, M., Carlsson, G., & Jönsson, P. G. (2020). Fibers of failure: Classifying errors in predictive processes. *Algorithms, 13*, 150.

Chou, Y. L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. G. (2021). Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications. arxiv.org/abs/2103.04244.

Cooper, A. (2004). *The Inmates are running the asylum: Why high-tech products drive us crazy and how to restore the sanity.* SAMS a Division of Pearson Education.

Dar, Y., Muthukumar, V., & Baraniuk, R. G. (2021). Farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. arXiv preprint arXiv:2109.02355.

Donadello, I., & Serafini, L. (2016). *Integration of numeric and symbolic information for semantic image interpretation.* Intelligenza Artificiale.

Doran, D., Schulz, S., & Besold, T. R. (2017). *What does explainable AI really mean? A new conceptualization of perspectives.* arXiv.org/abs/arXiv:1710.00794.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., … Modrák, M. (2020). Bayesian workflow. arxiv.org/abs/2011.01808.

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., … Valdes, G. (2020). Expert-augmented machine learning. *PNAS, 117*(9), 4571-4577.

Grinsztajn, L., Semenova, E., Margossian, C. C., & Riou, J. (2021). Bayesian workflow for disease transmission modelling in Stan. arxiv.org/abs/2006.02985.

Huang, B. W., Feng, F., Lu, C. C., Magliacane, S., & Zhang, K. (2021). AdaRL: What, where and how to adapt in transfer reinforcement learning. arxiv.org/abs/2107.02729.

Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. The Springer Series on Challenges in Machine Learning.

Kim, K., Kim, J., Zaheer, M., Kim, J. S., Chazal, F., & Wasserman, L. (2021). PLLay: Efficient topological layer based on persistence landscapes. arxiv.org/pdf/2002.02778.pdf.

Liu, C., Zhub, L., & Belkin, M. (2021). Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*. (in press). Retrieved from https://arxiv.org/pdf/2003.00307.pdf

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38.

Mooij, J. M., Magliacane, S., & Claassen, T. (2020). Joint causal inference from multiple contexts. *Journal of Machine Learning Research, 21*, 1-108.

Munro, R. (2020). *Human-in-the-loop machine learning.* Shelter Island, NY: Manning Publications.

Polyak, B. T. (1963). Gradient methods for minimizing functional. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 3*, 643-653.

Tran, D., Hoffman, M. D., Saurous, R. A., Brevdo, E., Murphy, K., & Blei, D. M. (2017). Deep probabilistic programming. arxiv.org/abs/1701.03757.

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning: Causal inference for observational and experimental data*. New York: Springer Science & Business Media.

Van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science: Causal inference for complex longitudinal studies*. New York: Springer Science & Business Media.

Vertechi, P., Frosini, P., & Bergomi, M. G. (2020). Parametric machines: A fresh approach to architecture search. arxiv.org/abs/2007.02777.

Wang, Y. X., & Blei, D. M. (2020). The blessing of multiple causes. *Journal of the American Statistical Association, 114*(528), 1574-1596.

Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Applications, 5*, 501-532.