Synthese (2023) 201:111
https://doi.org/10.1007/s11229-023-04110-9

ORIGINAL RESEARCH

The deep neural network approach to the reference class problem

Oliver Buchholz[1]

Received: 30 November 2021 / Accepted: 22 February 2023 / Published online: 15 March 2023

Abstract
Methods of machine learning (ML) are gradually complementing and sometimes even replacing methods of classical statistics in science. This raises the question whether ML faces the same methodological problems as classical statistics. This paper sheds light on this question by investigating a long-standing challenge to classical statistics: the reference class problem (RCP). It arises whenever statistical evidence is applied to an individual object, since the individual belongs to several reference classes and evidence might vary across them. Thus, the problem consists in choosing a suitable reference class for the individual. I argue that deep neural networks (DNNs) are able to overcome specific instantiations of the RCP. Whereas the criteria of narrowness, reliability, and homogeneity, that have been proposed to determine a suitable reference class, pose an inextricable tradeoff to classical statistics, DNNs are able to satisfy them in some situations. On the one hand, they can exploit the high dimensionality in big-data settings. I argue that this corresponds to the criteria of narrowness and reliability. On the other hand, ML research indicates that DNNs are generally not susceptible to overfitting. I argue that this property is related to a particular form of homogeneity. Taking both aspects together reveals that there are specific settings in which DNNs can overcome the RCP.

Keywords Reference class problem · Prediction · Overfitting · Machine learning · Deep neural networks · Big data

1 Introduction

Classical statistics can be considered the traditional workhorse of many disciplines, that, as a consequence, has been studied by philosophers for a long time. Yet methods

✉ Oliver Buchholz
oliver.buchholz@uni-tuebingen.de

[1] University of Tübingen, Cluster of Excellence "Machine Learning: New Perspectives for Science", Maria-von-Linden-Str. 6, 72076 Tübingen, Germany

of machine learning (ML) are gaining relevance in science. In particular, they are successfully employed in predictive tasks.[1] This raises the question whether ML faces the same methodological problems as classical statistics. This paper sheds light on this question by investigating a long-standing challenge to classical statistics: the reference class problem (RCP). Focusing on deep neural networks (DNNs), one of the most popular methods in ML, I try to carefully carve out how they cope with the RCP. I will conclude that although it remains a serious methodological challenge for them in many situations, some DNNs are able to overcome specific instantiations of the RCP.

In general, the RCP arises whenever the objective probability of possessing a certain property should be assigned to an individual. According to the frequentist account, this probability should be based on an observed relative frequency.[2] Yet an individual belongs to different reference classes and relative frequencies may vary across these classes. Consequently, it is unclear which reference class should be chosen to determine said single-case probability. Apart from this probabilistic version of the problem, there is a version that is structurally similar, yet not concerned with the rational determination of single-case probabilities, but rather with the rational construction of predictions. The present paper focuses on this predictive version of the RCP.[3]

For instance, consider William Smith who wants to predict whether he will be alive 15 years from now (Salmon, 1989, p. 69).[4] He belongs to different reference classes: the class of 40-year-old American males, the class of heavy cigarette-smoking individuals, and several other classes. Clearly, the evidence for 15-year survival varies considerably between them. It is therefore not straightforward to choose the class that should serve as a basis for making the prediction. The example illustrates that the RCP is central to situations in which statistical evidence is used to make a prediction for an individual case, even when the prediction is not a probability, but rather a real number or a discrete classification.[5] Consequently, it arises regularly within the framework of classical statistics, encompassing a variety of fields such as evolutionary biology (Strevens, 2016) or law (Colyvan et al., 2001; Colyvan and Regan, 2016).

An influential suggestion to solve the RCP is due to Reichenbach (1949). He proposes to base one's inferences on the reference class that is as *narrow* as possible while also allowing compiling *reliable* statistics. The narrowness of a class increases with the number of predicates that determine the class. Additionally, Salmon (1971) proposes to counterbalance a strict preference for narrower reference classes with the requirement of *homogeneity*. Briefly put, this means that the reference class should only be determined by those predicates that are relevant for a particular prediction.

Several authors have thus interpreted the predictive RCP as a problem of statistical model selection (Cheng, 2009; Franklin, 2010): the predicates that determine a reference class can be expressed by the variables in a statistical model. Solving the

[1] For one of the most recent breakthroughs, see Jumper et al. (2021).

[2] Hájek (2007) even argues that all common interpretations of probability face the RCP.

[3] In the following, I will therefore use the formulations 'predictive RCP' and 'RCP' interchangeably.

[4] This setting is distinct from the general RCP, since the prediction is not a probability, but a binary classification.

[5] I follow Romeijn (2022) in taking statistical evidence to be observed instances sampled from an underlying population, commonly organized into a dataset. For a formalized treatment in the context of ML, see Sect. 3.1.

RCP then reduces to identifying the model with the 'right' set of variables. Clearly then, any strategy that identifies one set of variables as the 'right' one ultimately needs to take into account the criteria of narrowness, reliability, and homogeneity that make for a suitable reference class.

However, in the context of classical statistics, existing strategies to approach the RCP with model selection techniques pose an additional challenge instead of offering a remedy: from a statistical point of view, there is a tradeoff between the narrowness of the class considered and the reliability of the information that this class contains.[6] A narrower reference class will contain fewer observations. Thus, by an argument along the lines of the law of large numbers, it will also have an inferior statistical reliability. Furthermore, the combination of fewer observations and a higher number of predicates defining a narrow reference class is problematic for another reason: expressing a narrow reference class by a model with a high number of variables and fitting it to a low number of observations leads to a situation in which the model can memorize the given data, but might predict new observations rather poorly. Thus, inferences derived from information in that reference class are likely susceptible to *overfitting* (Shalev-Shwartz and Ben-David, 2016). For the same reason, it is difficult to determine a homogeneous reference class using methods of classical statistics: using a model with a high number of variables, thereby considering all predictively relevant predicates, might lead to a homogeneous reference class, but also to a low number of observations in that class and thus, ultimately, to the risk of overfitting.

With the rise of big data, rapidly growing computational resources and datasets, methods of ML are gradually complementing, sometimes even replacing methods of classical statistics in science (Mjolsness and DeCoste, 2001; Wheeler, 2016) . In this paper, I focus on DNNs, one of the most popular ML methods. They are employed frequently and with astonishing success in predictive tasks (LeCun et al., 2015; Goodfellow et al., 2016). DNNs perform particularly well in settings involving so-called *high-dimensional data*, where the number of features associated with each observation is very high, usually much higher than the overall sample size (Belkin et al., 2019). This particular field of application serves as the starting point for my argumentation that proceeds in two steps.

First, I argue that the notion of 'big data' can be conceived along two perspectives. A dataset might be large simply because of the number of observations it contains. But the high dimensionality of many contemporary datasets adds a second perspective to the understanding of big data. I show that the combination of both perspectives can be connected to the notions of narrowness and reliability in the debate surrounding the RCP. On the one hand, high dimensionality of a dataset and thus a high number of features associated with each observation can be linked with the idea of a narrow reference class that is defined by a high number of predicates. On the other hand, a high number of observations can be interpreted as being related to the reliability of the information in a dataset.

Second, I argue that the particular functionality of some DNNs predestines them to exploit settings involving big data. For methods of classical statistics and many ML approaches, high-dimensional data involves the risk of overfitting. However, recent

[6] This observation is also highlighted by Salmon (1971, p. 41).

ML research reveals that there are DNNs for which this risk is much less prevalent: in many settings, they perfectly fit the training data, but also exhibit high predictive accuracy on new inputs (Belkin et al., 2019; Berner et al.,2021, p. 17).[7]

I argue that this gives rise to a situation in which DNNs remedy particular instantiations of the RCP, namely those involving high-dimensional or 'big' data. Their specific functionality enables them to exploit high-dimensional data without incurring the risk of overfitting which allows them to make predictions with high accuracy. I argue that this is akin to an accurate inference from relevant and reliable information in a very narrow reference class to previously unseen individuals.

The remainder of the paper is organized as follows: Sect. 2 introduces the RCP and reviews criteria for the suitability of a reference class. Section 3 provides the necessary background on ML and DNNs. Section 4 outlines existing strategies to solve the RCP that rely on the framework of classical statistics and shows that they fail in some situations. Section 5 argues that DNNs offer a remedy to specific cases of the RCP.

2 The reference class problem

This section discusses the RCP. It carves out important distinctions that have been introduced in the literature and their relevance for the present paper. Additionally, this section outlines criteria for the suitability of a reference class.

2.1 The problem

The RCP originates in the assignment of an objective probability to an individual object, that is, a single-case probability. According to the frequentist account, this probability should be based on an observed relative frequency. Yet an individual belongs to different classes, so-called *reference classes*, and relative frequencies may vary across these classes. Consequently, it is unclear which reference class should be chosen to determine the single-case probability (Reichenbach, 1949, p. 374, Venn, 1876, p. 194). I will refer to this original version of the problem as the *probabilistic* RCP. However, the treatment of the problem has gradually become more fine-grained.[8] The present paper focuses on the epistemological RCP as it arises in the context of prediction.

The context of *prediction* was introduced as a specific instantiation of the RCP by Fetzer (1977) and Salmon (1989). In this context, an individual should be assigned to a suitable reference class so as to allow for an accurate prediction. To do so, all available evidence relevant to the prediction at hand should be used.[9]

[7] Note, that a very close fit to the training data alone does not necessarily lead to overfitting. The key determinant is the gap between accuracy on training data and accuracy on new data (Goodfellow et al., 2016, p. 109). For details, see Sect. 3.1.

[8] See, e.g., Fetzer (1977), Kyburg (1977, 1983), Salmon (1977), Thorn (2012, 2017, 2019), and Wallmann and Williamson (2017).

[9] Both Fetzer (1977) and Salmon (1989) distinguish the predictive RCP from the *explanatory* RCP. In the latter, a known fact should be explained, for instance "John Jones's rapid recovery from his strep infection" (Salmon, 1989, p. 69). In this context, the RCP is about determining a reference class that is suitable to

The *epistemological* RCP concerns situations in which a rational agent is dealing with the question on which part of given statistical evidence they should base their inductive inferences and decision-making (Hájek, 2007) . As illustrated using the case of William Smith who tries to predict his 15-year survival, statistical evidence is relative to a particular reference class, the problem being that it is unclear which reference class is the correct one.

To illustrate the specific instantiation of the RCP examined in this paper, consider the widely discussed legal case *United States v. Shonubi*.[10] The case is about Charles Shonubi, a Nigerian citizen, who was apprehended on December 10, 1991 at New York's John F. Kennedy Airport (JFK), carrying 427.7 grams of heroin. The evidence gathered during the subsequent trial revealed that Shonubi had made at least seven smuggling trips between Nigeria and the United States prior to his detention. As a consequence, sentencing guidelines required an estimate of the overall amount of heroin that Shonubi imported during all eight of his trips (Tillers, 2005, p. 34). It was also required that this estimate be based on 'specific evidence'. In response to both requirements, data of 117 Nigerian drug smugglers that were apprehended at JFK in the period between Shonubi's first and last known smuggling trip was analyzed. In particular, the amounts of heroin found on these smugglers served as a basis for estimating the amount Shonubi carried during his first seven trips. This estimated amount was subsequently added to the known amount of 427.7 grams that resulted in the eighth trip (Colyvan et al., 2001, p. 169).

The case clearly involves a prediction problem, since the overall amount of heroin that Shonubi carried during his first seven trips was unknown at the time of the trial. Furthermore, the case is about predicting a quantity rather than a probability. So although the case does not involve the probabilistic RCP, it certainly involves a structurally similar problem: in order to predict the overall amount of heroin based on statistical evidence, Shonubi had to be assigned to some reference class. Yet it also had to be determined what constitutes 'specific evidence', that is, a suitable reference class in this particular situation. As several authors rightly point out, it is unclear why "Nigerian drug smugglers apprehended at JFK during the given time period" was chosen as Shonubi's reference class rather than "all drug smugglers at JFK, all Nigerian smugglers regardless of airport, or smugglers in general" (Cheng, 2009, p. 2082). In fact, there is an indefinite number of classes to which Shonubi could have been assigned. This includes apparently unsuspicious classes such as the class of all airline passengers or the class of toll collectors at New York's George Washington Bridge which was Shonubi's day job (Colyvan et al., 2001, p. 172).[11] Each of them would have resulted in very different predictions for the overall amount of heroin. Thus, when trying to make an individual prediction based on statistical evidence, it is unclear which part

Footnote 9 continued
explain why the recovery happened, yet without using the recovery as part of the total evidence to determine the reference class.

[10] See Cheng (2009), Colyvan et al. (2001), Colyvan and Regan (2016), Franklin (2010), and Tillers (2005).

[11] Given the fact that Shonubi had smuggled drugs on a fixed number of occasions, some of the candidate reference classes might appear not very meaningful. Yet this is precisely the point: there are many classes to which an individual belongs in principle and the RCP consists in assigning it to the most suitable one. For criteria determining a suitable reference class, see Sect. 2.2.

of the evidence should have a bearing on the prediction. Put differently, it is unclear which reference class to use to make the prediction. This is the epistemological RCP as it arises in the context of prediction.

2.2 Criteria for a suitable reference class

The previous section revealed that the RCP is about choosing a suitable reference class when applying statistical evidence to an individual object. Consequently, a solution to the RCP needs to spell out two things: first, a criterion for what constitutes a suitable reference class and second, a method for actually finding that class. This section discusses criteria for a suitable reference class. Strategies for actually finding it are outlined in Sect. 4.

One influential proposal of a solution to the RCP is due to Reichenbach (1949). For him, there are two criteria determining a suitable reference class: it should be as *narrow* as possible while also allowing compiling *reliable* statistics (Reichenbach, 1949, p. 374). What is meant by narrow and reliable?

On the predominant view, the concept of narrowness can be linked to the number of predicates by which a class is determined. For instance, given data about the entire population (no predicate) and data about males in that population (one predicate) when predicting the amount of heroin in the case of Shonubi, one should opt for the more specific data, thereby assigning Shonubi to the narrowest reference class possible that is refined by the highest number of predicates. This seems intuitive. Additionally, Thorn (2017) and Wallmann (2017) show that the preference for narrow reference classes can be formally justified: choosing the narrowest reference class maximizes accuracy in the sense that the difference between prediction and actual value will be minimal.

However, there are at least two problems with the criterion of narrowness. First, reference classes cannot "be *totally ordered* according to their narrowness" (Hájek, 2007, p. 568). For instance, given data about the entire population and data about males, it is straightforward to identify the narrowest reference class. Yet in a situation in which there is only reliable data regarding males that weigh more than 80 kilograms and regarding males with dark hair, this is not as straightforward. Obviously, each of the classes is narrower than the class of all males, but there is no reliable information as to which of them should be considered the narrowest reference class. Furthermore, it would be a mistake to judge them as equally narrow simply because both classes are determined by one further predicate (Hájek, 2007, p. 569).[12]

Second, solely focusing on the criterion of narrowness implies that one should always prefer evidence for singleton reference classes (Thorn, 2012, p. 303).[13] Thus, in the Shonubi case, the overall amount of heroin should have been determined based on the reference class containing only Charles Shunobi.[14] This clearly misguided strategy

[12] See Thorn (2019) for a discussion of the problem of partially overlapping reference classes.

[13] This is discussed as the "Problem of Uninformative Statistics" (Bacchus, 1990; Pollock, 1990).

[14] In fact, this approach was employed in an initial trial, predicting the overall amount by multiplying the amount that Shonubi carried on his last trip by eight. Yet the judgment based on this prediction was vacated due to a lack of "specific evidence" (Colyvan et al., 2001, p. 169).

illustrates what might have been obvious from the outset: that a strict preference for narrow and hence ultimately singleton reference classes is untenable.

Reichenbach (1949) seems to attenuate the strict preference for narrower reference classes by additionally requiring reliable information: one should choose the narrowest reference class that also contains reliable information. However, Reichenbach does not further specify the concept of reliability. Hájek (2007, p. 568) even argues that it is a vague concept *per se* that cannot be pinned down employing ideas of classical statistics such as a sufficiently large sample size. I partially disagree with this observation. Although there might be more to reliability than purely statistical aspects like a large sample, the latter aspects are certainly an important part of it. This is due to the fact that theoretical results that guarantee the reliability of statistical methods rely on precisely these aspects.[15] Hájek (2007, p. 568) also notes that the meaning of reliability might in fact be context-dependent and sensitive to pragmatic considerations. I agree with this observation. However, it seems unproblematic as soon as the specific context is made explicit. Here, the focus is on the RCP as it arises in the context of prediction. Thus it is reasonable to argue that information is reliable to the extent that it leads to accurate predictions.[16]

Apart from reliability, Salmon (1971, 1989) proposes *homogeneity* as another criterion to counterbalance a strict preference for narrower reference classes. As mentioned above, he argues that when concerned with prediction, one should exploit all available evidence. Yet what is crucial to achieve homogeneity is the *statistical relevance* of the evidence, which Salmon (1971, p. 42) defines as follows: when trying to predict the probability that an individual has some property B based on an overall set of evidence A, another property C is statistically relevant to B just in case $P(B|A, C) \neq P(B|A)$, that is, just in case conditioning on A *and* C leads to another probability for the individual to have property B than conditioning only on A.[17] Thus, to determine a suitable reference class for a prediction concerning property B, one should start by considering the broadest class A and partition it in terms of all predicates $C_1, C_2, \ldots$ that are statistically relevant to the question at hand; yet one should avoid partitioning the class in terms of statistically irrelevant predicates, since this would reduce the available evidence with no good reason. According to Salmon (1971, p. 43, 1989, p. 69), one should ultimately choose the broadest homogeneous reference class. This is the class that is subdivided by a *homogeneous partition*, that is, by a partition that includes all predicates that are known to be statistically relevant and that does not include any statistically irrelevant predicates.[18]

[15] For instance, the law of large numbers or the central limit theorem hold 'in the limit', that is, for large samples.

[16] One might object that given this definition, it is impossible to assess the reliability of information before making a prediction. One might also object that it is pointless to assess the reliability afterwards. I will address both objections in Sect. 3.1 and show that they are unfounded in the context of ML.

[17] Woodward (2021) provides a concise overview of this and related definitions.

[18] Since the partition is in terms of all predicates that are *known* to be relevant, Salmon (1989, p. 69) refers to 'epistemic homogeneity' which he distinguishes from 'objective homogeneity'. This means that it is *in principle* possible to refine a partition by adding further predicates, yet the relevant predicates to do so are not epistemically accessible.

While Salmon focuses on the prediction of probabilities and hence formulates the notion of statistical relevance in terms of probabilities as well, Colyvan et al. (2001) emphasize the importance of homogeneity even in settings like the case of Shonubi, where the prediction to be made is not a probability, but rather a real number. They argue that choosing the right reference class "is not just a question of specifying enough predicates to be jointly satisfied so that the reference class in question contains very few (but non-zero) members" (Colyvan et al., 2001, p. 172). Instead, the reference class should be homogeneous in the sense that refining the partition by adding another predicate does not (significantly) change the predicted value. I will refer to this idea as *predictive homogeneity*. This highlights that the formulation resembles Salmon's definition of a homogeneous partition in terms of statistical relevance. Yet it also highlights that the formulation is different because it replaces the focus on changes in probabilities that is central to the definition of statistical relevance by the more general focus on changes in predicted values.

Overall, the criterion of homogeneity complements the criterion of narrowness and can be considered as a lower bound to it: while the criterion of narrowness requires choosing a class that is determined by as many predicates as possible, the criterion of homogeneity requires choosing a class that is determined only by those predicates that are relevant to the question at hand.

In sum, the discussion reveals that Reichenbach's proposal to solve the RCP is still an important point of reference. The criterion of narrowness is intuitively plausible, yet it requires a counterpart to avoid shortcomings like singleton reference classes. In the context of prediction, both the criterion of reliability and the criterion of homogeneity serve as such a counterpart.

3 Machine learning and deep neural networks

This section provides an overview of central aspects of ML and DNNs. Readers familiar with the material may safely skip to Sect. 4.

3.1 Machine learning

The main focus of ML is on the problem of generalization: how to make accurate predictions for new instances based on empirical observations?[19] In the following, I will focus on the case of supervised learning. In this setting, there is an input space, X, an output space, Y, and it is assumed that they are governed by an unknown functional relationship $f : X \rightarrow Y$. I will focus on a *regression task* in which $X = \mathbb{R}^d$ and $Y = \mathbb{R}$.[20]

A set of training data, $\langle x_1, y_1\rangle, \ldots, \langle x_n, y_n\rangle \in \mathbb{R}^d \times \mathbb{R}$, is essential to most ML tasks. A concise way of capturing the data sampled from the input space is by means of a design matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$. Here, n is the number of observations and d is the

[19] For a book-length treatment of the field see Shalev-Shwartz and Ben-David (2016), for a concise overview see Jordan and Mitchell (2015).

[20] One might similarly consider a *classification task*, in which Y would be discrete.

number of features associated with each observation. Often, d is referred to as the *dimension* of the data. In many applications involving texts, speech or images, the number of features d in a dataset is high, in some cases even considerably higher than the number of observations, such that $d \gg n$. This issue is discussed under the headline of *high-dimensional data*. It is commonly encountered in fields such as astronomy, climate science, economics or genomics (Bühlmann and van de Geer, 2011; Johnstone and Titterington, 2009). The increasing prevalence of high-dimensional data is mainly driven by two factors: a dataset can be inherently high-dimensional because a high number of features is available for each observation. Yet a dataset can also become high-dimensional because researchers are unsure about the functional relationship between available features. In this case, they might construct a wide range of new features by interacting and transforming the available ones (Belloni et al., 2014) . The issue of high-dimensional data will come up again in the discussion below. For the moment, note that the features included in a dataset are somehow related to properties associated with the objects that constitute the observations in the dataset. They might consequently provide a link to the analysis of the RCP.

Based on the set of training data, the goal in ML is to find a function $h\colon \mathbb{R}^d \rightarrow \mathbb{R}$ that takes a new and previously unseen point x as input and predicts the corresponding label y as accurately as possible. This is why it is also called a *prediction rule*. The function h is usually chosen from a so-called hypothesis class $\mathcal{H}$. This is a class of functions that is predetermined by the developers or operators of an ML system. In most cases, empirical risk minimization (ERM) or some variant guides the choice of the final prediction rule $h \in \mathcal{H}$. This means that the function h is chosen such that it minimizes the *training risk*, that is, the average deviation between predicted labels, $h(x_i), i = 1, \ldots, n$, and true labels, y_i, in the training data. It is in this sense that the final prediction rule h should be as accurate as possible: the goal is to get as close as possible to the labels generated by the true but unknown underlying function f.

However, as mentioned above, the focus of ML is on generalization, that is, on predictions for new observations. So there needs to be a link between ERM on training data and generalization to unseen data. This link is established by the so-called i.i.d.-assumption that all input-output pairs, $\langle x, y\rangle$, are independent from each other and drawn from the identical but unknown probability distribution P over $\mathbb{R}^d \times \mathbb{R}$ (von Luxburg and Schölkopf, 2011, p. 653). This allows to assess the performance of h on new input-output pairs sampled independently from P, giving rise to the *test risk*. The goal of successful generalization is then operationalized by the requirement that in addition to minimizing the training risk (the goal of ERM), the gap between training and test risk should be minimized as well (Goodfellow et al., 2016, p. 109). It is within this setting that the reliability of the data, that is, whether it leads to accurate predictions, can be assessed to some extent before making predictions for unseen observations: given the i.i.d.-assumption, the training data is structurally similar to the test data which is why accurate predictions on the latter are more likely given accurate predictions on the former.[21]

[21] From a practical point of view, the i.i.d.-assumption may seem overly restrictive. However, it establishes the rationale for considering separate sets of training and test data and it allows to mathematically study the relationship between training and test risk.

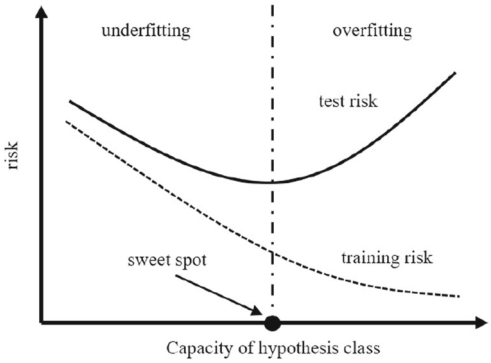

Fig. 1 Curves for training risk (dashed line) and test risk (solid line) depicting the relationship between overfitting, underfitting, and the capacity of the hypothesis class $\mathcal{H}$ (adapted from Belkin et al., 2019, p. 15850)

The relation between training and test risk and hence the ability to generalize is closely linked to two central challenges in ML: underfitting and overfitting. *Underfitting* occurs when a prediction rule is overly simplistic, lacks the capacity to capture the complexity in the data and hence achieves poor accuracy on the training data. *Overfitting* occurs when a prediction rule fits the training data very closely and achieves high accuracy on the training data, thereby also fitting idiosyncrasies of the sample at hand that are not relevant for future observations. This usually leads to poor generalization and hence to a large gap between low training and high test risk. Just in case there is such a large gap, a prediction rule is said to be subject to overfitting (Goodfellow et al., 2016, p. 110). Consequently, a very close fit to the training data is not equivalent to overfitting, but usually makes it more likely to occur.

Whether a prediction rule tends to underfit or overfit is closely tied to the capacity of the underlying hypothesis class. This is illustrated in Figure 1. A hypothesis class with low capacity contains rather simplistic prediction rules that may struggle to fit the training data and will be prone to underfitting. A hypothesis class with high capacity contains highly complex prediction rules that may even fit random patterns in the training data and will be prone to overfitting. Consequently, to balance over- and underfitting, it is usually necessary to impose certain restrictions on the hypothesis class.

For instance, given that the structure of input and output data points towards a linear relationship, one might restrict the hypothesis class such that it only contains linear prediction rules.[22] In this case, the hypothesis class would be given by all prediction rules of the form $h(x) = x_1c_1 + \cdots + x_dc_d$. Determining the final prediction rule would amount to determining the coefficients $c_1, \ldots, c_d$. On the one hand, this restriction would lead to at least an approximate fit between the final prediction rule and the training data, thereby avoiding underfitting. On the other hand, the restriction would ensure that the final prediction cannot fit the training data too closely, thereby avoiding overfitting.

[22] This type of restriction is discussed as *inductive bias* in the literature, but there are many other techniques for restricting hypothesis classes (Goodfellow, 2016, Ch. 7, Shalev-Shwartz and Ben-David, 2016, Ch. 2.3).

3.2 Deep neural networks

DNNs are usually depicted as graphs consisting of nodes, the neurons, and edges transmitting information between neurons.[23] For simplicity, I focus on fully connected feedforward networks in which the graph contains no cycles.[24]

More formally, a DNN can be described as a (directed and acyclic) graph, $G = \langle V, E \rangle$. The set of neurons is denoted by V, the set of edges is denoted by E. Typically, a DNN is structured in layers. If the DNN is fully connected, each node from one layer is connected to each node from the next layer by one edge. A network's number of layers is commonly referred to as the *depth* of the network. DNNs contain a high number of layers which is why they are called 'deep'.

Data is processed through the network as follows: first, it enters the network at the input layer. This layer contains one node per dimension of the input data. Then, the data is transmitted to the next layer. An activation function that is associated with the nodes in the network determines whether and in what form the data is processed from one neuron to another. A weight function determines, for each edge, the importance of the data passed on along that edge. Consequently, the input of a neuron consists of the weighted sum of the transformed outputs of all nodes connected to it.[25] Finally, for each input x, the network produces an output y at the output layer.

In practical applications, developers or operators of a DNN usually predefine the *architecture* of the network. It consists of a graph and an activation function. Thus, the output labels that a network produces depend on the predefined architecture and on the weights, w. Consequently, the learning process of a DNN amounts to finding the best among all possible configurations of weights for a given architecture. In this context, 'best' means most accurate according to ERM. The most common method to minimize the empirical risk of DNNs is the so-called stochastic gradient-descent (SGD) algorithm. Its underlying rationale is to initialize the weights with random values, to update them stepwise and to converge to that configuration of weights that leads to the lowest empirical risk. This configuration is then used to compute new predictions y for previously unseen observations x.[26]

4 Statistical strategies to solve the reference class problem

Section 2.2 discussed three important criteria for a suitable reference class: narrowness, reliability, and homogeneity. However, little has been said about strategies to find the reference class for which these criteria are fulfilled. In particular, while it might be straightforward to determine a narrowest reference class, it is unclear how to discern relevant from irrelevant evidence and hence how to establish predictive homogeneity

[23] For an in-depth treatment of DNNs, see Goodfellow et al. (2016), for a philosophically motivated introduction, see Buckner (2019).

[24] Although there is a large variety of DNNs, many authors focus on fully connected feedforward networks, because their mathematical treatment is more convenient (Berner et al., 2021).

[25] Often a *bias*, which can be conceived as the intercept of a linear equation, is added to the weighted sum.

[26] For a non-technical yet detailed discussion of the learning process of DNNs and its philosophical ramifications, see Buchholz and Raidl (forthcoming).

of a reference class. As mentioned above, several authors have interpreted the RCP as a problem of statistical model selection, which is why they try to address this issue within the framework of classical statistics.

For instance, Cheng (2009) argues that the predicates that determine a reference class can be expressed by the variables in a statistical model. So in a linear model of the form $h(x) = x_1c_1 + \cdots + x_dc_d$, the variables $x_1, \ldots, x_d$ are taken to be predicates that determine a reference class, while $h(x)$ would be a prediction based on these variables. Thus, in the case of Shonubi, x_1 might encode 'age', x_2 might encode 'citizenship' and $h(x)$ might encode the overall quantity of heroin predicted based on the variables included in the model.[27]

Given this setup, choosing the right reference class for making a prediction reduces to identifying the model with the right set of variables. With respect to the reference class, the criteria of narrowness, reliability, and homogeneity are constitutive for what is 'right'. With respect to the set of variables, the model should be selected such that it avoids under- and overfitting (Cheng, 2009, p. 2095). As mentioned above, the latter is closely related to a model's complexity and thus, given the model's overall structure (i.e., a linear function, a specific architecture, etc.), to the number of variables it contains: the model should include enough variables to avoid underfitting; yet it should also contain only relevant variables to avoid overfitting. Consequently, when framing the RCP as a problem of statistical model selection, there is a close connection between the goal of avoiding under- and overfitting and the goal of choosing a reference class that is as narrow as possible while also being homogeneous.

When interpreting the RCP as a problem of statistical model selection, it seems straightforward to solve it using model selection methods.[28] Accordingly, Cheng (2009) argues that statistical measures like the *Akaike Information Criterion* (AIC) should be employed to determine the right reference class. The AIC evaluates a statistical model by measuring the model's fit to the evidence as well as its complexity.[29] Thus, it evaluates how well the model balances over- and underfitting. Both poor fit to the evidence and high complexity of the model lead to higher values of the AIC. If, instead, a model achieves a considerably close fit to the evidence while being relatively simple, the AIC has a small value. Consequently, the best model is the one that minimizes the AIC. According to Cheng (2009, p. 2094), this also solves the RCP, for the variables of the best model in terms of the AIC determine the best reference class in a given situation.

A related approach is proposed by Franklin (2010). He also frames the RCP as a problem of statistical model selection. Yet contrary to Cheng, he suggests using *feature selection methods* to solve the problem. These methods are commonly used as follows: first, a complex model is specified that contains as many variables as possible given

[27] I am using the terms 'variable' and 'feature' interchangeably. The former is commonly used in classical statistics, the latter in ML, but their meaning is the same.

[28] Clearly, in that case, 'solving the RCP' does not amount to determining the provably correct reference class, but rather to finding a well-justified and potentially correct solution. The same holds for the DNN case below.

[29] The fit to the evidence is usually measured by the maximum of the model's likelihood function, $\hat{L}$, and complexity by the model's number of parameters, p, such that $\text{AIC} = -\ln(\hat{L}) + 2p$ (Akaike, 1974). For a thorough philosophical discussion of the AIC, see Forster and Sober (1994).

the available data. Next, the model is fitted to the data using a feature selection method that retains relevant variables in the model, while weighting irrelevant variables less or even discarding them altogether.[30] This leads to a fitted model that contains the relevant variables and in which the weights for irrelevant variables are small or even zero. According to Franklin, the variables that are identified as relevant by the feature selection method determine the right reference class in a given situation.

There are certainly many aspects about both approaches that require further discussion. Yet there is one general issue that affects both of them. In fact, it even invalidates them as a remedy to the RCP in many situations. Both Cheng (2009) and Franklin (2010) develop their proposals using the case of Shonubi as their point of departure. The discussion above revealed that all reference classes considered in this case were determined by a rather low number of predicates. This means that statistical models applied to the case will have a rather low number of predictively relevant variables, thereby avoiding over- and presumably also underfitting.

However, suppose the proposed strategies were applied to a setting involving high-dimensional data. In this case, a wide range of variables would be predictively relevant. Additionally, due to the high-dimensional setting, the sample size would be relatively low compared to the number of features associated with each observation. Thus, this situation embraces two scenarios, both of which would be problematic from the perspective of classical statistics: on the one hand, a statistical model could exploit all predictively relevant variables. This would correspond to a reference class that is both narrow and predictively homogeneous. However, it would also lead to overfitting, since a model including a large number of variables would be flexible enough to fit idiosyncrasies of the relatively small sample. Consequently, the information in the reference class would not be reliable in the sense that it gives rise to accurate predictions. On the other hand, employing the AIC or feature selection methods would lead to a model that is sufficiently simple to avoid overfitting. Yet this would prevent many predictively relevant variables from entering the model, thereby leading to a reference class that is neither narrow nor predictively homogeneous.[31]

Overall, the example reveals that there are situations in which it is not possible to simultaneously achieve all desiderata for a suitable reference class within the framework of classical statistics. Consequently, proposals to solve the RCP using methods of classical statistics often fall short of doing so, because they cannot escape the fundamental tradeoff between overfitting and underfitting that is particularly challenging in the context of high-dimensional data.

5 The argument

The previous sections examined the RCP and central ideas of ML separately. To answer the guiding question of this text, both subjects have to be taken together: how, if at all, are DNNs suited to deal with the RCP? In this section, I argue that there are situations

[30] For a detailed survey of feature selection methods, see Hastie et al. (2009, Ch. 3.3 and 3.4).

[31] One might object that selecting variables to determine a reference class does not make sense in the case of high-dimensional data, where variables encode, e.g., the color of singular pixels in images. I will address this objection in Sect. 5.1.

in which DNNs remedy specific instantiations of the RCP. By clearly demarcating these situations, my argumentation also allows to distinguish the latter from situations in which the RCP remains the intricate methodological problem as which it is known.

5.1 'Big Data' is related to narrowness and reliability

DNNs gained their relevance mainly from what Wheeler (2016) refers to as "the era of big data". Thus, as a first step, it is worth analyzing what 'big data' actually means.

First, the sheer number of observations in many contemporary datasets is vast. While classical statistics is often concerned with assessing the significance and precision of inferences made from a restricted sample, "we are now routinely handling population datasets directly or sample sizes so immense […] that they behave like population data" (Wheeler, 2016, p. 330). Given this observation and the common assumption that "[t]he larger the sample gets, the more likely it is to reflect more accurately the distribution and labeling used to generate it" (Shalev-Shwartz and Ben-David, 2016, p. 38), considerations regarding the reliability of inferences in classical statistics do not, or at least to a far lesser extent, carry over to applications of ML.[32] Here, the representativeness of a given sample for the entire population is much more likely based on the size of the sample.

Second, many datasets nowadays belong to the high-dimensional setting outlined above. Thus, in addition to a large number of observations, each observation is associated with a—possibly much higher—number of features (Bühlmann and van de Geer, 2011). This is interesting from the perspective of the RCP, where a reference class gets narrower with any further predicate that is added to its definition. Consequently, when framing the RCP as a problem of statistical model selection, high-dimensional datasets give rise to very narrow reference classes.[33]

Before proceeding, let me address two potential objections to this interpretation of features in a dataset as predicates that determine a reference class. First, consider the example of a dataset consisting of images. Images are usually stored in a dataset such that for each pixel in the image, there is one feature in the dataset giving the color of the pixel as a numeric value. Suppose further that the goal is image classification, that is, to determine a suitable reference class or, equivalently, to find a statistical model based on the given data that allows to correctly classify future images. Clearly then, selecting features that give the color of singular pixels seems to be something entirely different from selecting a feature such as 'age' in the case of Shonubi: one might object that features giving the color of pixels do not have an immediately obvious

[32] In this context, reliability is to be understood solely in its relation to the data and to the sample size in particular. This is not to say that ML methods are *per se* more reliable than methods of classical statistics. Additionally, reliability needs to be distinguished from mathematically proven properties, e.g., statistical guarantees for the performance of a method. While the latter do not (yet) exist for some ML methods, said methods nevertheless work reliably in many situations—what is missing is a definite explanation for *why* this is the case (Berner et al., 2021, p. 17).

[33] One might question whether this leads to reference classes that do not contain enough observations to draw reliable inferences. However, in high-dimensional datasets, the number of observations is only low *relative* to the number of dimensions, but usually not in *absolute* terms (Belloni et al., 2014, p. 29). While this is nevertheless problematic from the perspective of classical statistics, I will argue in Sect. 5.2 that it is less problematic for DNNs.

meaning and that, as a consequence, such features give rise to reference classes that do not have an immediately obvious meaning either.[34] This would call into question the strategy of approaching the RCP as a problem of statistical model selection in such settings. However, in the context of prediction, it is not the goal to investigate reference classes themselves or the predicates by which they are determined. Instead, the goal is to identify those features that determine a reference class for making accurate predictions. Thus, the criterion of predictive relevance alone is discerning suitable from unsuitable features in this context. Whether or not the features and the reference class they determine have an immediately obvious meaning is less important.[35]

Second, one might object that when interested in the predicates that determine a reference class, what is relevant are not features in the dataset, but rather the values taken on by the features. For instance, in a demographic dataset, the predicate 'age' will be satisfied for each observation and hence irrelevant to determine a reference class. What is relevant is the value of 'age' for each individual in the dataset. This objection can be addressed by constructing a binary variable for each value taken on by a feature like 'age', leading to a dataset that contains features like 'age30' that equal one if an individual is 30 years old and zero otherwise. These can be interpreted as useful predicates to determine reference classes.

To summarize: in this section I argued that 'big data' can be conceived along two perspectives. They provide a promising basis to approach the RCP employing DNNs because they address both components of Reichenbach's proposal: to choose a reference class that is narrow and for which reliable statistics are available. What remains is the problem of over- and underfitting when trying to determine predictively relevant features.

5.2 Deep neural networks can exploit high-dimensional data

We have seen above that strategies to solve the RCP with classical model selection techniques fail in applications involving high-dimensional data. On the one hand, statistical models could include a high number of variables in such situations. In this way they would fulfill the requirement of narrowness, but they would also overfit the information in the reference class which would prevent them from predicting accurately. On the other hand, statistical models could include a low number of variables. This would prevent them from overfitting, yet it would also prevent the choice of a predictively homogeneous reference class, since not all predictively relevant variables would be part of the model.

Contrary to this observation, recent results reveal that some DNNs possess a remarkable feature: they perform particularly well on high-dimensional data (Berner et al., 2021, p. 19, Neyshabur et al., 2017, p. 5947). In this setting, they are able to *interpolate*, that is, to exactly fit the training data, thereby achieving zero training error (Belkin et al., 2019, p. 15849). Given the preceding discussion of central ideas in ML, one

[34] One could perhaps go so far as to say that they have no human-graspable meaning at all.

[35] This is why it is crucial to distinguish the goal of prediction treated in this paper from the goal of explanation. The latter clearly requires the predicates determining a reference class to have a human-graspable meaning.

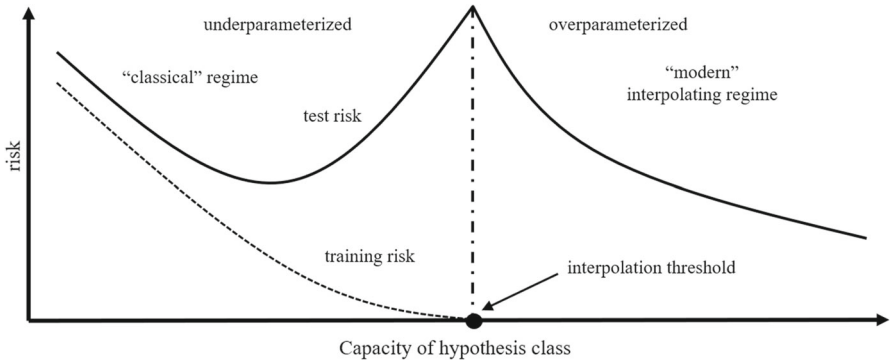

Fig. 2 Curves for training risk (dashed line) and test risk (solid line) depicting the double-descent risk curve, which incorporates the U-shaped risk curve (i.e., the 'classical' regime) together with the observed behavior from the 'modern' interpolating regime, separated by the interpolation threshold (adapted from Belkin et al., 2019, p. 15850)

might take this behavior as an indication for overfitting and a poor ability to generalize. However, as several authors show, DNNs possess a high ability to generalize to previously unseen data (Belkin et al., 2019; Zhang et al., 2017) . This seems peculiar as it is at odds with the standard framework of ML, especially regarding its treatment of the under- versus overfitting problem. It is also at odds with the conventional wisdom presented in standard textbooks that "a model with zero training error is overfit to the training data and will typically generalize poorly" (Hastie et al., 2009, p. 221).

Thus, apparently, the case of DNNs is not appropriately captured by the depiction in Figure 1 where an algorithm's predictive ability diminishes with increasing capacity of the underlying hypothesis class. As a consequence, Belkin et al. (2019) propose and empirically confirm an alternative framework that combines the traditional context of under- and overfitting—the 'classical' regime as they call it—with the specific behavior of some DNNs—the 'modern' interpolating regime. The main feature of their framework is what the authors refer to as the *double-descent risk curve* depicted in Figure 2. It corresponds to the classical U-shaped curve depicted in Figure 1 above, as long as an algorithm's capacity is below the so-called interpolation threshold. This threshold marks the point beyond which an algorithm interpolates the training data. While prediction rules obtained directly at the threshold generally exhibit a high test risk indicating a low predictive accuracy, Belkin et al. (2019, p. 15850) "show that increasing the function class capacity beyond this point leads to decreasing risk, typically going below the risk achieved at the sweet spot in the 'classical' regime." This means that large DNNs with a complex architecture involving many layers and incorporating a high number of features as inputs are suited particularly well for any kind of prediction task.

Many insights about the generalization ability of DNNs rely on empirical studies conducted with specific network architectures, but there is theoretical progress for some aspects of the problem (Zhang et al., 2021) .[36] Perhaps most importantly,

[36] For instance, Arora et al. (2019) focus on two-layer networks and Soudry et al. (2018) focus on networks with linear activation functions to derive theoretical results. Recent theoretical progress with significantly weakened assumptions is also made by Holzmüller (2021).

recent analyses of the SGD algorithm revealed that the algorithm exhibits a behavior of *implicit regularization* (Neyshabur et al., 2015, Poggio et al., 2020, Theorem 4). Mathematically, this means that the final configuration of weights to which the algorithm converges has a small norm.[37] With respect to the structure of a DNN, a small norm corresponds to a final configuration of weights or, equivalently, to a final prediction rule that is relatively simple. In particular, this means that many weights within the network will have a small value and that some of them will even be assigned a value of zero. So after the learning process, a DNN might locally 'look' considerably simpler than its initial architecture, since several input features might not be processed to the next layer and the flow of information along edges might be muted at various points in the network.

The observation of implicit regularization can be considered as one possible explanation for the astonishing generalization ability of DNNs.[38] In a way, it also allows to reconcile the behavior of DNNs with conventional statistical wisdom: just as in other methods of classical statistics and ML, accuracy and simplicity need to be balanced in DNNs as well. What remains surprising, however, is that this balance is struck automatically by the SGD algorithm and without being enforced at some point during the learning process. While statistical measures like the AIC explicitly incorporate the tradeoff between accuracy and simplicity as the objective of model selection, the SGD algorithm operates solely with the objective of maximizing accuracy—yet implicitly restricts the complexity of the final network as well.

In sum, recent ML research reveals that highly complex DNNs are often not susceptible to overfitting, because they achieve both a low training and a low test error.[39] Consequently, when framing the RCP as a problem of statistical model selection, they seem superior to methods of classical statistics in determining reference classes that are both narrow and predictively homogeneous. I will carve out this last step of my argument in the next section.

5.3 The deep neural network approach to the reference class problem

According to the discussion above, a solution to the predictive RCP needs to propose a method that identifies relevant predicates so as to achieve accurate predictions. Framing the RCP as a model selection problem, this means that the method should find the predictively relevant features to be included in the final model.

When approaching the RCP using DNNs, everything starts with input data in a design matrix, $\mathbf{X} \in \mathbb{R}^{n \times d}$. The dimension d indicates the number of features associated with each observation, $x_i, i = 1, \ldots, n$, so each observation might be interpreted

[37] A *norm* is a function that takes the elements of a vector as inputs and outputs a non-negative number. It can be interpreted as the 'size' of the vector (Goodfellow et al., 2016, p. 37).

[38] Other explanations focusing on some kind of simplicity bias of the SGD algorithm are put forward, e.g., by Huh et al. (2021), Razin and Cohen (2020) or Valle Pérez et al. (2019). Shwartz-Ziv and Tishby (2017) try to provide an information-theoretic explanation. For a philosophical discussion of the latter, see Räz (2022).

[39] This is even the case for noisy training data (Berner et al., 2021, p. 18). The generalization performance only disappears for data that is entirely random and hence contains no learnable structure at all (Zhang et al., 2017, 2021).

as possessing d different properties or characteristics. We have seen that there are DNNs which perform best in high-dimensional settings and that the "era of big data" regularly brings about datasets that belong to precisely this setting. Consequently, it is reasonable to focus on cases where $d \gg n$. The task of image classification is an excellent example for such cases, since storing images in a dataset often gives rise to a setting in which the number of pixels in each image, corresponding to the number of features, is larger than the number of stored images, corresponding to the number of observations. Additionally, DNNs are considered the state-of-the-art method to perform image classification (Berner et al., 2021, p. 2).

The discussion above revealed that a reference class gets narrower with each predicate that is added to its definition. It also revealed that features in a dataset can be interpreted as predicates that determine a reference class. Taking these aspects together, one can conclude that given high-dimensional input data, a DNN starts a prediction exercise like image classification with the narrowest reference class possible that is defined by a high number of features.[40] Thus, this very first step is in line with Reichenbach's recommendation to use information for the narrowest reference class available. It is also in line with Franklin's (2010) feature-selection approach according to which one should start the process of finding a suitable reference class by considering the model that contains the highest number of variables. However, there have to be safeguards that counterbalance a strict preference for narrow reference classes and prevent overfitting.

When trying to determine a suitable reference class, the criterion of predictive homogeneity introduced above can be seen as a counterpart to the criterion of narrowness. Recall that a reference class is predictively homogeneous just in case it is determined by all and only those features that are predictively relevant (see Sect. 2.2). In the context of DNNs, predictive relevance is assessed via ERM: given the training data, the SGD algorithm chooses all weights within the network such that the empirical risk is minimized. As long as the empirical risk is not minimal, the algorithm proceeds by altering the weights to get closer to the minimum. Once the minimum is reached, the algorithm terminates. Put differently, the algorithm only converges to the minimum and terminates once everything predictively relevant is taken into account and appropriately weighted, since otherwise, the empirical risk could be decreased even further.[41] We have seen that very complex DNNs often achieve perfect accuracy and hence zero empirical risk in the training sample as well as a high ability to generalize to new data. In the context of the RCP, this means that such DNNs are able to exploit the large number of features in the data to an extent that allows them to make accurate predictions on both the training and the test data.[42] For instance, in the example of image classification, DNNs are highly successful in selecting and

[40] 'Narrowest …possible' is to be understood relative to the available d-dimensional data, since I am concerned with the problem of finding a suitable reference class based on given statistical evidence rather than with the problem of determining whether additional evidence is required.

[41] In principle, it is possible that the SGD algorithm only converges to a *local* minimum (Goodfellow et al., 2016, p. 281). However, in the case of highly complex networks, convergence to a *global* minimum is particularly likely to occur (Li et al., 2018, Poggio et al., 2020, p. 30044, Vidal et al., 2017, p. 2).

[42] This is in line with the aforementioned formal justification of choosing narrow reference classes since they maximize predictive accuracy (Thorn, 2017; Wallmann, 2017) .

appropriately weighting those features that correspond to the pixels that are crucial for classifying new images (Huh et al., 2021; Krizhevsky et al., 2012) . Consequently, it is reasonable to assume that DNNs operating within the ERM paradigm take into account all predictively relevant features during their learning process.

However, we have seen that a reference class is predictively homogeneous just in case it is determined by all *and only those* features that are predictively relevant. Maximizing accuracy alone is therefore insufficient, because apart from all relevant features, the most accurate model might also include irrelevant features. Furthermore, maximizing accuracy alone involves the risk of overfitting. Above, I discussed how classical model selection techniques try to address this issue and fail to consider all predictively relevant features in high-dimensional settings. DNNs are different in this respect. The previous section revealed the central role of implicit regularization that takes place in the determination of a network's weights. In addition to maximizing accuracy, the SGD algorithm generally yields a final prediction rule that is simple in the sense that the network's weights have a small norm. This means that some weights are assigned a high value, since the associated input is considered to be of high predictive relevance for the output, but others are assigned a low value—maybe even zero—, since the associated input is considered less relevant—or not relevant at all—for the output. Put bluntly, irrelevant features are downweighted or eliminated to achieve a simple configuration of weights.

We can now combine both insights. First, within the framework of ERM and assuming that a global minimum for the empirical risk was reached, the final prediction rule is the one that maximizes accuracy and hence includes *all* predictively relevant features (otherwise the risk could be decreased further by including additional features). Second, given maximal accuracy, the final prediction rule is also the simplest solution and hence *only* includes predictively relevant features due to the simplicity bias of SGD.[43] Taking both aspects together reveals that the combination of ERM and the simplicity bias of SGD seems to identify all and only those features that are predictively relevant, thereby giving rise to a predictively homogeneous reference class.[44,45]

So in sum, the learning process of DNNs is governed by ERM, leading to the consideration of all predictively relevant features and to maximal accuracy. However, it is also governed by a bias towards simple solutions, leading to the consideration of predictively relevant features only, thereby preventing overfitting. Thus, in situations involving big data, the specific functionality of DNNs allows them to exploit data for very narrow yet predictively homogeneous reference classes and to incorporate the relevant information in a combination of weights that maximizes predictive accuracy.

[43] Huh et al. (2021) explore this combination of accurate predictions and simplicity bias for the example of image classification.

[44] This observation does not even presuppose implicit regularization, but only some kind of simplicity bias of the SGD algorithm. For instance, Räz (2022) recently argued on information-theoretic grounds that DNNs achieve homogeneous partitions of the input data by getting rid of irrelevant information during the learning process. However, he characterizes these partitions as very complex and hence rejects them as not useful, since his focus is on explaining DNNs rather than on using DNNs for predictions (Räz, 2022, p. 28).

[45] As pointed out, e.g., by Buckner, (2018, p. 5362), DNNs generate increasingly abstract representations of the input features across their layers. However, assessing whether these representations might have a bearing on the RCP is beyond the scope of this paper.

This is why DNNs are suited to deal with the RCP as it arises in the context of prediction. Contrary to methods of classical statistics, they might offer a remedy to it in these situations.

Clearly, there is a flipside to the latter reasoning: by illustrating how DNNs can offer a remedy to the RCP in some very specific situations, it also suggests that in many others, DNNs fare no better than methods of classical statistics.

First, I emphasized that the concept of predictive homogeneity crucially depends on the minimization of the empirical risk. Yet I also pointed out that, sometimes, the SGD algorithm might fail to achieve this minimization and converge to a local instead of a global minimum of the loss function (see Fn. 41). Consequently, predictive homogeneity cannot be achieved in these situations and neither do they give rise to a suitable reference class of features.

Second, we have seen that the criterion of reliability is crucial for determining a suitable reference class. Above, I explicitly tied reliability to characteristics of the data, in particular to the sample size (see Fn. 32). On the one hand, this seems to be very much in the spirit of Reichenbach,'s (1949) requirement to compile reliable statistics. On the other hand, one might question whether this is sufficient or whether reliability should also be an explicit requirement for the method that does the compiling. This question is particularly pressing in the case of DNNs, since several network architectures have been shown to lack robustness and to be easily fooled by slight perturbations of the input data.[46] Tying reliability to the data, however, the above reasoning rests on the assumption that DNNs indeed work reliably and thus only applies to situations in which this really is the case.

6 Conclusion

This paper set out to answer the question whether ML faces the same methodological problems as classical statistics. I tried to shed light on this question by investigating the RCP, a long-standing challenge to classical statistics. Albeit originating as a problem of (frequentist) probability theory, the RCP also concerns the more general question as to how statistical evidence should have a bearing on individual cases. My focus in this paper was on cases in which a reference class should be chosen so as to allow for accurate predictions, that is, on the epistemological RCP as it arises in the context of prediction.

I argued that one particular method of ML, namely DNNs, are sometimes able to overcome the RCP in settings involving high-dimensional data. First, the high dimensionality of the data can be linked to the concepts of narrowness (via a high number of features) and reliability (via a high number of observations), both of which were proposed as criteria for a suitable reference class by Reichenbach (1949). Second, the particular functionality of DNNs predestines them to exploit high-dimensional settings. Due to the SGD algorithm's behavior of implicit regularization, they are less susceptible to overfitting. Consequently, they can select a narrow reference class

[46] This is perhaps most evident in adversarial examples (Szegedy et al., 2014) . They can mislead DNNs that only consist of linear components, while DNNs that also include non-linear components seem to be less vulnerable (Goodfellow et al., 2015) .

consisting of a high number of features that is also predictively homogeneous in the sense that it only includes features that are relevant to make accurate predictions.

In sum, I conclude that contrary to methods of classical statistics, DNNs can offer a remedy to the RCP in settings involving high-dimensional data. However, and this is just as important a conclusion, there are also many settings in which DNNs cannot provide such a remedy—and in which, consequently, the RCP remains a serious methodological challenge.

Acknowledgements I would like to thank Jon Williamson for helpful conversations and David Danks, Atoosa Kasirzadeh, Andrés Páez, as well as Emanuele Ratti for helpful comments on a draft of this article. I would also like to thank Alexandra Zinke and Wolfgang Spohn for very valuable comments on previous versions of this article as well as Eric Raidl for his guidance and continuing support during the entire process that led up to this version.

Funding Open Access funding enabled and organized by Projekt DEAL. My research was funded by the Baden-Württemberg Foundation (program "Verantwortliche Künstliche Intelligenz") as part of the project AITE (Artificial Intelligence, Trustworthiness and Explainability).

Declarations

Conflict of interest There are no competing interests to declare.

References

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Arora, S., Du, S. S., Hu, W., Li, Z. & Wang, R. (2019). Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *International Conference on Machine Learning*.

Bacchus, F. (1990). *Representing and Reasoning with Probabilistic Knowledge*. Cambridge, MA: MIT Press.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. *Proceedings of the National Academy of Sciences of the United States of America, 116*(32), 15849–15854. https://doi.org/10.1073/pnas.1903070116

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives, 28*(2), 29–50. https://doi.org/10.1257/jep.28.2.29

Berner, J., Grohs, P., Kutyniok, G. & Petersen, P. (2021). The Modern Mathematics of Deep Learning. arXiv:2105.04026.

Buchholz, O. & Raidl, E. (forthcoming). A Falsificationist Account of Artificial Neural Networks. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/721797.

Buckner, C. (2018). Empiricism Without Magic: Transformational Abstraction in Deep Convolutional Neural Networks. *Synthese, 195*(12), 5339–5372. https://doi.org/10.1007/s11229-018-01949-1

Buckner, C. (2019). Deep Learning: A Philosophical Introduction. *Philosophy Compass, 14*(e12625). https://doi.org/10.1111/phc3.12625.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Cheng, E. K. (2009). A Practical Solution to the Reference Class Problem. *Columbia Law Review, 109*, 2081–2105.

Colyvan, M., & Regan, H. M. (2016). Legal Decisions and the Reference Class Problem. *The International Journal of Evidence & Proof, 11*(4), 274–285. https://doi.org/10.1350/ijep.2007.11.4.274

Colyvan, M., Regan, H. M., & Ferson, S. (2001). Is it a Crime to Belong to a Reference Class? *The Journal of Political Philosophy, 9*(2), 168–181.

Fetzer, J. H. (1977). Reichenbach, Reference Classes, and Single Case 'Probabilities'. *Synthese, 34*, 185–217.

Forster, M., & Sober, E. (1994). How to Tell When Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science, 45*(1), 1–35. https://doi.org/10.1093/bjps/45.1.1

Franklin, J. (2010). Feature Selection Methods for Solving the Reference Class Problem. *Columbia Law Review Sidebar, 110*, 12–23.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Goodfellow, I., Shlens, J. & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Hájek, A. (2007). The Reference Class Problem Is Your Problem Too. *Synthese, 156*(3), 563–585. https://doi.org/10.1007/s11229-006-9138-5

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Holzmüller, D. (2021). On the Universality of the Double Descent Peak in Ridgeless Regression. In *International Conference on Learning Representations*.

Huh, M., Hossein, M., Zhang, R., Cheung, B., Agrawal, P. & Isola, P. (2021). The Low-Rank Simplicity Bias in Deep Networks. arXiv:2103.10427.

Johnstone, I. M., & Titterington, D. M. (2009). Statistical Challenges of High-dimensional Data. *Philosophical Transactions of the Royal Society A, 367*(1906), 4237–4253. https://doi.org/10.1098/rsta.2009.0159

Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Hassabis, D. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 25, pp. 1–9).

Kyburg, H. E. (1977). Randomness and the Right Reference Class. *The Journal of Philosophy, 74*(9), 501–521. https://doi.org/10.2307/2025794

Kyburg, H. E. (1983). The Reference Class. *Philosophy of Science, 50*(3), 374–397. https://doi.org/10.1086/289125

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, D., Ding, T. & Sun, R. (2018). Over-Parameterized Deep Neural Networks Have No Strict Local Minima For Any Continuous Activations. arXiv:1812.11039v1.

Mjolsness, E., & DeCoste, D. (2001). Machine Learning for Science: State of the Art and Future Prospects. *Science, 293*(5537), 2051–2055. https://doi.org/10.1126/science.293.5537.2051

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring Generalization in Deep Learning. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5947–5956).

Neyshabur, B., Tomioka, R. & Srebro, N. (2015). In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. In *International Conference on Learning Representations*.

Poggio, T., Banburski, A., & Liao, Q. (2020). Theoretical Issues in Deep Networks. *Proceedings of the National Academy of Sciences of the United States of America, 117*(48), 30039–30045. https://doi.org/10.1073/pnas.1907369117

Pollock, J. L. (1990). *Nomic Probability and the Foundations of Induction*. New York: Oxford University Press.

Räz, T. (2022). Understanding Deep Learning With Statistical Relevance. *Philosophy of Science, 89*(1), 20–41.

Razin, N., & Cohen, N. (2020). Implicit Regularization in Deep Learning May Not Be Explainable by Norms. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1–35).

Reichenbach, H. (1949). *The Theory of Probability: An Inquiry into the Logical and Mathematical Foundations of the Calculus of Probability* (2nd ed.). Berkeley and Los Angeles, CA: University of California Press.

Romeijn, J. W. (2022). Philosophy of Statistics. In E. N. Zalta, U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy, https://plato.stanford.edu/archives/fall2022/entries/statistics/.

Salmon, W. C. (1971). Statistical Explanation. In W. C. Salmon (Ed.), *Statistical Explanation & Statistical Relevance* (pp. 29–87). Pittsburgh: University of Pittsburgh Press.

Salmon, W. C. (1977). Objectively Homogeneous Reference Classes. *Synthese, 36*(4), 399–414.

Salmon, W. C. (1989). Four Decades of Scientific Explanation. In P. Kitcher & W. C. Salmon (Eds.), *Scientific Explanation* (pp. 3–219). Minneapolis: University of Minnesota Press.

Shalev-Shwartz, S., & Ben-David, S. (2016). *Understanding Machine Learning: From Theory to Algorithms* (1st ed.). New York: Cambridge University Press.

Shwartz-Ziv, R. & Tishby, N. (2017). Opening the Black Box of Deep Neural Networks via Information. arXiv:1703.00810.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., & Srebro, N. (2018). The Implicit Bias of Gradient Descent on Separable Data. *Journal of Machine Learning Research, 19*(1), 1–57.

Strevens, M. (2016). The Reference Class Problem in Evolutionary Biology: Distinguishing Selection from Drift. In G. Ramsey & C. H. Pence (Eds.), *Chance in Evolution* (pp. 145–175). Chicago: The University of Chicago Press. https://doi.org/10.7208/9780226401911-008

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2014). Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*.

Thorn, P. D. (2012). Two Problems of Direct Inference. *Erkenntnis, 76*(3), 299–318. https://doi.org/10.1007/s10670-011-9319-6

Thorn, P. D. (2017). On the Preference for More Specific Reference Classes. *Synthese, 194*(6), 2025–2051. https://doi.org/10.1007/s11229-016-1035-y

Thorn, P. D. (2019). A Formal Solution to Reichenbach's Reference Class Problem. *Dialectica, 73*(3), 349–366. https://doi.org/10.1111/1746-8361.12273

Tillers, P. (2005). If Wishes Were Horses: Discursive Comments on Attempts to Prevent Individuals from Being Unfairly Burdened by Their Reference Classes. *Law, Probability and Risk, 4*(1–2), 33–49. https://doi.org/10.1093/lpr/mgi001

Valle Pérez, G., Camargo, C. Q. & Louis, A. A. (2019). Deep Learning Generalizes Because the Parameter-Function Map Is Biased Towards Simple Functions. In *International Conference on Learning Representations*.

Venn, J. (1876). *The Logic of Chance* (2nd ed.). London: Macmillan and Co.

Vidal, R., Bruna, J., Giryes, R. & Soatto, S. (2017). Mathematics of Deep Learning. arXiv:1712.04741.

von Luxburg, U., & Schölkopf, B. (2011). Statistical Learning Theory: Models, Concepts, and Results. In D. M. Gabbay, S. Hartmann, & J. Woods (Eds.), *Handbook of the History of Logic* (Vol. 10, pp. 651–706). Amsterdam and Boston: Elsevier. https://doi.org/10.1016/B978-0-444-52936-7.50016-1

Wallmann, C. (2017). A Bayesian Solution to the Conflict of Narrowness and Precision in Direct Inference. *Journal for General Philosophy of Science, 48*(3), 485–500. https://doi.org/10.1007/s10838-017-9368-x

Wallmann, C., & Williamson, J. (2017). Four Approaches to the Reference Class Problem. In G. Hofer-Szabó & L. Wroński (Eds.), *Making it Formally Explicit* (pp. 61–81). Cham: Springer. https://doi.org/10.1007/978-3-319-55486-0_4

Wheeler, G. (2016). Machine Epistemology and Big Data. In L. McIntyre & A. Rosenberg (Eds.), *The Routledge Companion to Philosophy of Social Science* (pp. 321–329). London: Routledge.

Woodward, J. (2021). Scientific Explanation In E. N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2021/entries/scientific-explanation/.

Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2017). Understanding Deep Learning Requires Rethinking Generalization. In *International Conference on Learning Representations*.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding Deep Learning (Still) Requires Rethinking Generalization. *Communications of the ACM, 64*(3), 107–115. https://doi.org/10.1145/3446776