

# Chapter 9

## Machine Learning and the Philosophical Problems of Induction



Davor Lauc

**Abstract** This chapter provides an analysis of the relationship of the traditional problems of justifying inductive inferences to the field of machine learning. After the summary of the philosophical problems of induction, text focus on the two philosophical problems relevant to the supervised and unsupervised machine learning. The former is a famous new riddle of induction devised by N. Goodman. The author argues that remarkable results in the theory of machine learning, no-free-lunch theorems are a formalisation of this traditional philosophical problem. Consequently, lengthy philosophical discussions on this problem are relevant to these results and vice versa. The later problem is the problem of similarity, as identified by N. Goodman and W. V. Quine. It is claimed that those discussions can help practitioners of unsupervised learning to be aware of its limitations.

**Keywords** Machine learning · New riddle of induction · No-free-lunch theorems · Problem of similarity · Philosophy of computing

### 9.1 Introduction

Human knowledge is tightly tied with inference and, arguably, all knowledge is based on some kind of inference. Consequently, an artificial system that poses knowledge or intelligence is also based on some mode of inference. Traditionally, all inference is partitioned into deduction and induction.

The former, deduction, has a special status. In well-formed deductive reasoning, the conclusion necessarily follows from the premises. A deduction can be seen as a truth-preserving machine, provide the true statement and it will yield the truth. The deductive inference is also very familiar to the computer science community, not only through mathematics and deductive systems but also through correspondence between programs and proofs, the famous Curry–Howard isomorphism.

---

D. Lauc (✉)

Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia  
e-mail: [dlauc@ffzg.hr](mailto:dlauc@ffzg.hr)

TerraLogix Group, London, UK

© Springer Nature Switzerland AG 2020  
S. Skansi (ed.), *Guide to Deep Learning Basics*,  
[https://doi.org/10.1007/978-3-030-37591-1\\_9](https://doi.org/10.1007/978-3-030-37591-1_9)

However, the price that we pay for this special status of deduction, of its infallibility, is enormous. Although some deductive inferences such as mathematical proof took years or even centuries to be completed, the deduction does not yield any new information in an absolute sense. As it can be easily demonstrated within the framework of semantic information theory [6], the conclusion of a deductively valid argument never contains more information than the premises. In the philosophical jargon, this result is usually called the *scandal of deduction*.

Thus, to get new knowledge in an absolute sense, we need the following form, the non-deductive forms of inference. In the modern sense, all of the non-deductive reasoning is considered as a form of inductive reasoning. The standard definition, complementary to the definition of deduction, is that inductive inference is the one where it is possible to get a false conclusion from the true premises. Naturally, this definition covers a wide range of inferences.

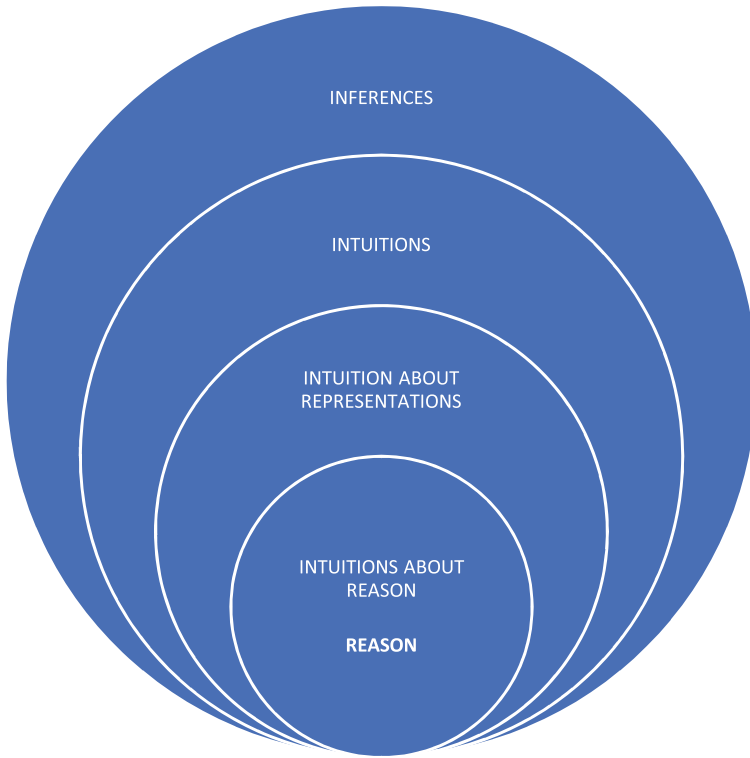
In the philosophical context, it is useful to distinguish two groups of the problem related to inductive inference. The first group is related to developing methods of induction or at least method of distinguishing better from worse inductive inferences. Although there is still a debate in the philosophy of developing such *inductive logic*, it seems that the leading role in this area has been overtaken by other disciplines. The science of statistics, and even machine learning, can be considered as an endeavour to develop such inductive inference.

The other group of problems belongs to the more hardcore philosophical issue, the challenge of *justifying* the inductive inference. To some extent, statistics tackle the problem of justification, and fields like *formal learning theory* are concerned with such justification. However, those problems are still mainly in the realm of philosophy. It is beneficial to contrast those two groups using the recent theory of human reasoning developed by psychologists Hugo Mercier and Dan Sperber, which is depicted in Fig. 9.1 (Mercier and Sperber, 2017).

If their theory is empirically adequate, the reasoning is deeply embedded in several categories of inference, and the justification of reason is just one part of it. Even if this theory is not an adequate description of the way humans infer and reason, in this context, it is a restatement of well-known Reichenbach distinction between the *context of discovery* and the *context of justification*. We will ignore how a human or an artificial reasoner comes to inductive inferences. Our focus will be on the philosophical problem of how we can justify such conclusions, and if it is possible at all.

## 9.2 What Are the Philosophical Problems of Induction?

As the problem of induction is one of the most significant philosophical issues in the last few hundred years, consequently, tons of literature about it has been created. So, it is entirely beyond the scope of this text to provide a comprehensive review of it. We will focus on a bird's eye review from the perspective of machine learning of the three most pressing problems of induction, which are related but distinct.



**Fig. 9.1** Mercier and Sperber embedding of reason

The problem of induction, or the classical formulation of the problem, is also called the Hume’s problem of induction. David Hume formulated it in the seventeenth century, and in modern terminology it can be shortly reformulated as follows. He poses the question of how we can justify inductive inference, how we can conclude from the observed to the unobserved. There are two ways, demonstrative and probable in Hume’s terminology, that is deductive or inductive. If we take the first way, we would need to presuppose as a premise that unobserved will be similar to observed, but this is what we are trying to prove precisely. If we take the second way, we are using the principles of reasoning we are trying to justify. So, we are using unacceptable circular reasoning. As that exhaust all possibilities of justifying the inductive inference, we must conclude that it is unjustifiable. It is vital to notice that Hume does not deny prevalence or importance of such reasoning; he is only stating its unjustifiability.

In the context of machine learning, we can understand this classical problem as a generalisation of the bias-variance trade-off. We can use high regularisation of the model to keep it simple (other unjustified assumption), set hyperparameters and test it on the held-out part of the dataset, use cross-validation and other tricks of trade and

then hope that our model will perform well in the wild. Based on the experience of machine-learning practitioners, using such techniques to avoid overfitting will result in the models that will perform better on new, yet unobserved data. However, what justifies that conclusion? A natural law that the real-life data will be similar to one in our training dataset or the experience with the past performance of model developed with a similar methodology? If we accept the Hume's dichotomy, then nothing.

Another well-known problem related to induction was formulated in the 1940s by German-American philosopher Carl Gustav Hempel, often labelled as "the raven paradox". In a nutshell, Hempel is concerned with the problem of confirming generalised statements, like "all ravens are black". It is natural to accept that black ravens are confirming such generalisation while raven of any other colour will falsify it. The issue is that, logically, such generalisation is equivalent to its contraposition, the statement that "all nonblack things are nonravens". This logical equivalence standardly means that the two statements have the same truth conditions; the same observed data makes both of them true and false. The "paradox" unfolds with the realisation that any black thing or any nonravens make the previous statement right, so that, for example, green apple confirms it.

The third problem related to inductive inference is arguably the most relevant to machine learning. It is formulated by Nelson Goodman and is known as "the new riddle of induction". Goodman considered Hume's problem as a pseudo-problem because, in the analogy with deduction, there cannot be a justification of our inferential practices. Goodman also provided a solution to the Hempel's paradox based on the idea that an acceptable instance (a black raven) not only confirms the statement but also falsifies the contrary hypothesis that none raven is black.

However, he offered a new puzzle connected with induction that somehow turns Hume's problems on its head. In the sense, that it is not the case that none of the inductive inferences are justified, but that it is too many of them. His famous example is that, based on many observed green emeralds, and none emeralds of other colour, we can naturally conclude that all emeralds are green. However, this is not the only conclusion we could reach. Let us define a new property called "grue". The property is true of all things that are blue and observed before some point in future, for example before the year 2100, or observed after this time point and blue. Based on the same data, we could conclude that all emeralds are "grue", and of course numerous other statements generated in this fashion. Goodman challenge is to distinguish properties we could use in inductive inference like green, from the one we cannot. He named the former projectable and the latter non-projectable. In spite of numerous discussions on this topic in the last 50 years there is still little agreement about how to solve his problem.

It is worthwhile to analyse a similar argument, without the fancy neologism, made by Goodman earlier:

Suppose we had drawn a marble from a certain bowl on each of the 99 days up to and including VE day, and each marble drawn was red. We would expect that the marble drawn on the following day would also be red. So far all is well. Our evidence may be expressed by the conjunction " $R_d1 \& R_d2 \& \dots \& R_d99$ ," which well confirms the prediction " $R_d100$ " But increase of credibility, projection, "confirmation" in any intuitive sense, does not occur

in the case of every predicate under similar circumstances. Let “ $S$ ” be the predicate “is drawn by VE day and is red, or is drawn later and is non-red.” The evidence of the same drawings above assumed may be expressed by the conjunction “ $S_{a1} \& S_{a2} \& \dots \& S_{a99}$ .” By the theories of confirmation in question, this well confirms the prediction “ $S_{a100}$ ” but actually we do not expect that the hundredth marble will be non-red. “ $S_{a100}$ ” gains no whit of credibility from the evidence offered hundredth marble will be non-red.

From this quotation, the relevance to machine learning is more evident. By using different learning setups and algorithms we can reach different models that are indistinguishable by our dataset. We will analyse this into more details later in the text.

### 9.3 Why Are the Philosophical Problems of Induction Relevant to Machine Learning?

Leslie Gabriel Valiant, a British computer scientist, claims that “...induction is a pervasive phenomenon exhibited by children. It is as routine and reproducible a phenomenon as objects falling under gravity.” [27]. To extend his metaphor further, although effects of gravitation were of course known in antiquity, we were not satisfied by the Aristotelian explanation that a body falls to the earth because it is their natural place, for this is where it belongs. Neither we were completely satisfied by the pretty accurate Newtonian description of how gravity works, nor with better description and understanding made by general relativity. We still want to have enhanced understanding of the nature of a phenomenon of gravity, not only to account for anomalies and discrepancies like extra energetic photons and provide better description and prediction but for the sake of understanding it as such.

By analogy, as machine learning is the venture of creating computer systems that learn from data, such system can be viewed as an inductive machine, as a device that is based on and performs inductive inferences. We are not, and should not be, satisfied by the fact that the systems we are developing are getting better at their tasks, that perform inductive inference well. We would like to know more, to understand the whys of induction and learning, to better understand the limits and explainability of it, not only of the present learning systems but any future systems as well. As we are putting more trust in machine learning systems and integrating them into our lives, from health to legal system, it is crucial to understand their foundation and limits more thoroughly.

It may happen, as it already happened many times in the development of philosophy and sciences, from physics to psychology, that the problems of induction detach from the mothership of philosophy. It may become a problem of some particular science, or even new science of its own, like the formal learning theory. However, relevant questions about induction are still so open-ended, and there is still so little understanding of the exact nature of those questions, methods to approach them and possible solution. So, it is expected that the problem will stay “philosophical” for at least some time.

So even if philosophy and philosophical problems are at present, less relevant to the actual practice of science and technology than they used to be, it is plausible to argue that the minimal relevance of philosophical problems is to avoid rediscoveries and repeat mistakes made by philosophers. Namely, in the two-and-a-half millenniums of philosophical endeavours, many questions were raised, many problems and several solutions were discussed, and numerous pro-and-contra arguments were examined. Almost by definition, many of the discussed problems are pseudo-problems, many of the solutions are bogus, and many of the arguments are unconvincing. However, by the same token, many of the arguments are solid and show that some theories are implausible to be valid, so the least we can learn from philosophy is not to repeat same mistakes again and again.

Let us exemplify this with the claim related to the problem of induction and machine learning made 10 years ago by C. Anderson, former editor-in-chief of *Wired Magazine*. In his influential and provocative article, inspired by big-data hype, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, he states that fundamental scientific methodology—hypothesise, model, test—is now obsolete due to a vast amount of data. He claims that “with enough data, the numbers speak for themselves”, and that petabytes of data allow us to say that correlation is enough, that correlation *is* causation.

For philosophers, this is not a new claim. It is just reiteration of one of the many arguments made by empiricist and rationalists in 25 old debate on the source and justification of human knowledge. Thesis formulated by Anderson was most notably made in seventeenth century by Francis Bacon in his *Novum Organum*. Bacon’s main point was that contrary to the then-dominant Aristotelian scientific method based on deduction. He argued that scientific knowledge should be based on experimental data. In his famous metaphor, scientists are bees that unlike spiders the rationalist or ants the pure empiricist are taking the middle course. They take materials from the flowers (observations) and “transforms and digests it by a power of its own” [2]. In the contemporary analogy, we only need to provide enough data to our deep learning models, and they will flourish and produce knowledge.

Among many arguments against this position, let us consider the one stated by Sir Karl Popper. He can be considered as, at least philosophical, father of the scientific method attacked by Anderson.

... the belief that we can start with pure observations alone, without anything in the nature of a theory, is absurd; as may be illustrated by the story of the man who dedicated his life to natural science, wrote down everything he could observe, and bequeathed his priceless collection of observations to the Royal Society to be used as inductive evidence. This story should show us that though beetles may profitably be collected, observations may not. Twenty-five years ago I tried to bring home the same point to a group of physics students in Vienna by beginning a lecture with the following instructions: ‘Take pencil and paper; carefully observe, and write down what you have observed! They asked, of course, what I wanted them to observe. Clearly, the instruction, ‘Observe!’ is absurd. ... Observation is always selective. It needs a chosen object, a definite task, an interest, a point of view, a problem. Moreover, its description presupposes a descriptive language, with property words; it presupposes similarity and classification, which in its turn presupposes interests, points of view, and problems. ... objects can be classified, and can become similar or dissimilar, only in this

way – by being related to needs and interests. ... a point of view is provided ... for the scientist by his theoretical interests, the special problem under investigation, his conjectures and anticipations, and the theories which he accepts as a kind of background: his frame of reference, his ‘horizon of expectations’. [22]

Digested reiteration of this Popper’s argument can be heard in the machine learning community under the dictum *no learning without bias* [5, 15].

Among many philosophical arguments related to machine learning, in this work, we will focus on the two examples. The first one is related to the supervised learning and our thesis that the well-known no-free-lunch theorems are just reiteration of famous Nelson Goodman’s New Riddle of Induction problem. The latter is related to unsupervised learning and the relevance of the Willard Van Orman Quine analysis of the problem of similarity to the problem of clustering.

## 9.4 Supervised Learning and the New Riddle of Induction

It is a pearl of received wisdom the no-free-lunch (NFL) theorems—the great negative results in machine learning—as a reiteration or even a formalisation of the Hume’s problem of induction.<sup>1</sup> The distinguished machine-learning researchers like Christophe Giraud-Carrier and Pedro Domingos state, respectively:

It then becomes apparent that the NFL theorem in essence simply restates Hume’s famous conclusion about induction having no rational basis... [8]

also,

...This observation was first made (in somewhat different form) by the philosopher David Hume over 200 years ago, but even today many mistakes in machine learning stem from failing to appreciate it.” [4]

Even the originator of the first form of the NFL theorems, David Wolpert, 15 years after he proved the theorem, joins the information cascade and claims that:

...these original theorems can be viewed as a formalisation and elaboration of concerns about the legitimacy of inductive inference, concerns that date back to David Hume... [30]

This chapter argues that the NFL theorems, although vaguely connected to the classical philosophical problem of induction, do not restate the Hume’s problem, but rather the associated Nelson Goodman’s argument.<sup>2</sup> We claim that NFL theorems are closely related to Goodman’s new riddle of induction (NRI), to the extent that they are one possible formalisation of the riddle. Additionally, we would like to pose the question of the relevance of the NFL theorems to the lengthy philosophical discussion on NRI, as the relationship is yet to be researched. The related, reversed,

---

<sup>1</sup>This part of the text is an extended version of the author’s text “How Gruesome are the No-free-lunch Theorems for Machine Learning?” [16].

<sup>2</sup>We suppose that the new riddle is a different issue from the classical problem of induction, what is the received position with a few notable exception.

the issue is the relevance of NRI to NFL and the question as to whether the machine-learning community could benefit from the almost 70 years of fruitful discussion about Goodman’s argument.

### 9.4.1 *No-Free-Lunch Theorems*

The first form of NFL theorem was proven by Wolpert and Macready, 1992, in the context of computational complexity and optimisation research [28, 31]. He later proved the variant of the theorem for the supervised machine learning [29]. For the sake of our argument, we will sketch the proof of the simplified version of the theorem for supervised learning based on the work of Cullen Schaffer [25].

In the simplest, discrete settings of the machine learning of a Boolean function, training data  $X$  consists of the set of binary vectors representing a set of attributes that are true or false for each instance of binary function—concept. Each vector is labelled as a positive or negative example of a concept we want to learn. The machine-learning algorithm  $L$  tries to learn a target binary function  $y$ ; that is, it tries to learn a real concept from this set of examples. The training dataset is always finite with some length  $n$ , and the relative frequency of data feed to  $L$  is defined by probability distribution  $D$ . In a context more familiar to the philosophers, this problem of machine learning can be seen as a guessing a true form of a large  $n$ -ary truth function from the partial truth-table, where most of the rows are not visible.

The key performance indicator of a machine learner is a generalisation performance, with the accuracy of the learner found within the data outside the training dataset. Modern machine-learning algorithms can easily “memorise” data from the training dataset, and perform poorly on the “unseen” data, leading to the problem known as overfitting. So, the success of the learner is measured by how well it will generalise, and how well it performs on the novel data. In the simple setting of binary concept learning, the baseline of the generalisation accuracy of a learner,  $GP(L)$  is at the level of a random guess, with the accuracy of the novel data being 0.5. Such performance is the result we will expect on average if we use the toss of a coin to decide, for an unseen example, whether it belongs to our target concept or not. Obviously, we want any learner to perform better than this.

The NFL theorem claims that, for any learner  $L$ , given any distribution  $D$  and any  $n$  of  $X$

$$\frac{\sum_{f \in Y} GP(L)}{|Y|} = 0.5,$$

where  $Y$  is a set of all target functions, all possible concepts that can be learned.

So, the theorem states that, on average, the generalisation performance of any learner is no better than random guessing. All learning methods, from the simple decision trees to the state-of-the-art deep neural networks, will perform equally when all possible concepts are considered.



This result, unanticipated at least on the first sights, does bear some resemblance to the discrepancy between the results of the argument and our expectations in the case of the Hume’s argument. However, it does not claim that we cannot learn anything from the training data or experience, but that we can learn everything, which is, arguably, the point of Goodman’s argument. The resemblance to the NRI will be more evident from the sketch of the proof of the NFL theorem. The basic idea is straightforward: for any concept that the learner gets right, there is a concept that it gets wrong or, in Goodman’s lingo, for every “green” concept there is a “grue” concept. The “grue” concept is constructed similar to the NRI argument, in that it agrees on all observed data—data in the training dataset—with the “green” concept, and it is bent on all non-observed data.

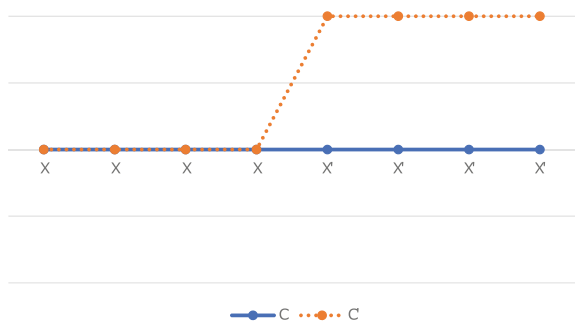
More formally, for every concept  $C$  that  $L$  learns to classify well—say it classifies  $m$  novel examples accurately—there is a concept  $C'$  that  $L$  learns where all  $m$  examples will be misclassified.  $C'$  is constructed as follows:

$$C' = \begin{cases} C & \text{if } x \in X \\ -C & \text{if } x \notin X \end{cases}$$

Visually, this simple construction of the concept  $C'$  corresponds to the Wittgenstein–Goodman “bent predicate” [1], where  $X$  represents observed data (training dataset) and  $X'$  unobserved data.

From the perspective of the primary measure of the success of the learning—generalisation accuracy, for every accuracy improvement  $a$  over the baseline for a concept  $C$ , there is a concept  $C'$  that will offset the improvement of the accuracy by  $-a$ . Consequently, the improvement in accuracy for any learner over all possible concepts is zero. It is possible to generalise this result to the more general learning settings, and many extensions of the theorem are proven [13, 14] (Fig. 9.2).

**Fig. 9.2** A bent predicate



### 9.4.2 *Is the No-Free-Lunch Theorem the New Riddle of Induction?*

Although NFL bears a strong resemblance to NRI, it seems worthy to analyse differences and similarities between these two results. Let us start with similarities. Both arguments are about inductive inference, about inferring from the known to the unknown, and from the observed to the unobserved data. Both arguments imply that there are too many inductive inferences that can be inferred. Furthermore, both arguments seem to draw empirically inadequate conclusions, contrary to scientific practice and reasonable expectation. Nobody is expected to conclude that all emeralds are *grue*, and neither that random guess is an inductive strategy as good as any other.

The fundamental resemblance is in the construction of the arguments, the split of the evidence and the bend in the unobserved data. In most of the NRI arguments, we split the evidence into observed and unobserved (sometimes to some point in the future). Equally, in the NFL, data is split into observed, training dataset and the unobserved data to which the learning algorithm should generalise. In both arguments, the other counter-concept, *grue* or  $C'$ , is constructed in the same manner. It agrees on the observed data and bends on the unobserved data.

Regarding the differences, firstly, there is a difference in the argument contexts. NRI was made in the philosophical, theoretical context of the logic of confirmation and pragmatic vindication of induction, while the NFL was made in the technological context of artificial intelligence and computing. The aim of the arguments also differs, at least at first glance. The intention of NRI, at least in Goodman's original form [9, 11], was to recognise one of the problems in the logic of confirmation—the demarcation between projectable and non-projectable predicates. On the other hand, the objective of the NFL was to demonstrate that there is no single best algorithm, initially for the optimisation and search, and later for supervised learning.

The most significant difference seems to be in the scope of quantification. The no-free-lunch theorem quantifies overall learners and all concepts, while Goodman's argument seems to be about constructing one particular example. However, NRI can be reformulated to have a similar quantificational structure as NFL.

The takeaway of this formalisation would be one of the lessons that Goodman has taught us—the importance of the language for the induction, or the impossibility of empirical investigation without some predefined language that we bring to the process. It is interesting to compare this with the conclusion that the same researcher from the machine-learning community draws from the NFL—there is no learning without bias, there is no learning without knowledge [5].

## 9.5 Unsupervised Learning and the Problem of Similarity

Willard Van Orman Quine, one of the greatest philosophers of the twentieth century, accepted Hume’s challenge to inductive inference and agreed that justification could be provided neither by experience nor a priory—“The Humean predicament is the human predicament” [21]. Quine is trying to provide, in the framework of his naturalistic epistemology, an explanation of our inductive practices based on evolution. He claims that all learning is based on conditioning, thus on induction, and therefore induction itself cannot be learned:

... the instinct of induction: the instinct to expect perceptually similar stimulations to have similar sequels. ... Philosophers have marvelled that expectation by induction, though fallible, is so much more successful than random guessing. This is explained by natural selection ...” [20]

Regarding the other two problems, Quine notices that the Hempel’s problem, the raven paradox, is reducible to Goodman’s problem [23]. Precisely, a complement of a projectible predicate (black–nonblack), does not need to be, also almost never is, projectible. His simple solution to the new riddle of induction is that “green emeralds” are more *similar* to each other than the “grue emeralds”. However, this reduction of the problems of induction to the problem of similarity is not an easy solution, because “..the dubious scientific standing of a general notion of similarity ... the dubiousness of this notion is itself a remarkable fact” [23].

Quine was not the first philosopher to tackle the concept of the similarity, among others Leibniz, Hume and Goodman were analysing it. Leibniz considered similarity as a weakened version of his famous principle of identity of indiscernibles. Two objects (substances in his terminology) are identical if they share all properties, two objects are similar if they share at least one [18].<sup>3</sup> Likewise, Hume position of similarity (resemblance) is that degree of similarity of two concepts depends on the number of “simple ideas”, properties, that they share [12], the approach that is today called *the common attribute view of similarity*.<sup>4</sup>

In the context of machine learning problem of similarity emerges primarily in unsupervised learning, although it is relevant to supervised learning tasks as well. In clustering problem, the primary technique of unsupervised learning, similarity (closeness, distance) metric is input to clustering algorithms as well as to all its performance metrics. There are several philosophical and practical problems with supervised learning even once the distance metrics are defined, reaching from determining number of clusters to the interpretation of clusters, but here we will focus on the concept of similarity itself as a prerequisite of any unsupervised learning technique.

Let us illustrate the problem of similarity with a simple, practical application, that is the problem of reasoning with dates, the temporal references with the granularity of

<sup>3</sup>Rodriguez-Pereyra claims there are strong philosophical grounds that Leibniz thought that the similarity of substances does not derive from the similarity of their properties (accidents) [24].

<sup>4</sup>Although there are readings of Hume that suggest that Hume’s position on similarity is closer to Goodman’s view [7].

days. Date similarity seems a trivial problem as we can use simple absolute difference in days between any two dates as similarity metrics. However, how to extend this the similarity metrics to the inexact temporal expressions, dates determined by inexact expressions such as “the beginning of the twentieth century” or “the date of birth of the person who died in 1875”? It is natural to represent such dates, that often occur in practical application, using discrete probability distribution over the range of relevant period. So, for example, we could represent exact dates with distribution where total probability mass is assigned to one day. If the date is from an unreliable source, we can assign only 0.1 mass to it and spread the rest to the neighbouring dates according to some probability distribution. In the same manner, we can represent expressions like “XVII century” or “beginning of the XVIII” century (see Fig. 9.3).

So, if we want to cluster such dates, what similarity metrics should we use? There is no simple distance in parameters of distribution because many different distributions that need to be used. There are many similarity measures and distance function between probability distribution in different scientific fields. In statistics and probability theory, there are distance correlation, Bhattacharyya distance, f-divergences like Kullback–Leibler, Kolmogorov–Smirnov and many others [3]. There are information-theoretical distance/similarity measures like mutual information or Jensen–Shannon divergence, as symmetric version of Kullback–Leibler divergence [19]. The principal problem with all those distances is that they do not satisfy the underlying intuitive semantics of the inexact date’s comparison.

Even worse, similarity metrics as perceived by the field experts, historians, in this case, do not even satisfy the essential requirement of any metrics—reflexivity. Professionals perceive narrow ranges of dates as more similar to itself than the larger ones, so, for example, the year 1941 is more similar to itself than the twentieth century [17]. In the context of cognitive sciences, Tversky studied extensively this problem of subjective and highly contextual perception of similarity [26]. He demonstrated that human judgement of similarity is not symmetrical, as test subjects perceived that North Korea is more similar to China than vice versa.

As it is often the case, this empirical research was preceded by philosophical analysis. Goodman wrote off the concept of similarity in both philosophically and scientifically use:

...Consider baggage at an airport checking station. The spectator may notice shape, size, colour, material, and even make of luggage; the pilot is more concerned with weight, and the passenger with destination and ownership. Which pieces are more alike than others depends not only upon what properties they share, but upon who makes the comparison, and when. [10]

The takeaway from this and related discussions for the unsupervised learning could be expressed in Quinean lingo like: there is no “fact of the matter”, no real ground truth for the similarity metrics, hence for any association or clustering learning. Those structures are not something objective in the nature that science can discover as we invented it, and we should be aware of contextuality and limitation of it.

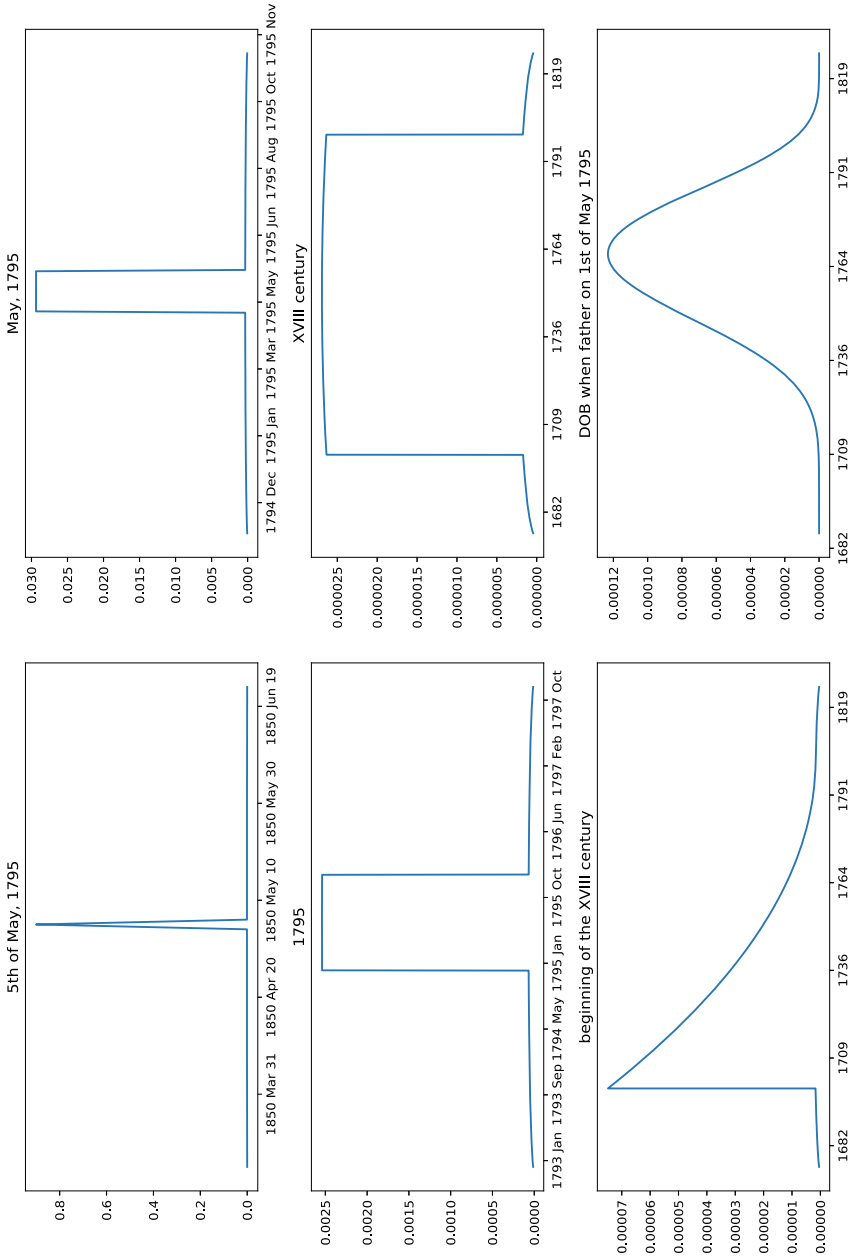


Fig. 9.3 Representing inexact dates as probability distributions

## References

1. Blackburn S (1984) *Spreading the word*. Oxford University Press
2. Bacon F, Thomas F (1889) *Novum organum*. Clarendon Press, Oxford
3. Cha S-H (2007) Comprehensive survey on distance/similarity measures between probability density functions. *City* 1(2):1
4. Domingos P (2012) A few useful things to know about machine learning. *Commun ACM* 55(10):78–87
5. Domingos P (2015) *The master algorithm: how the quest for the ultimate learning machine will remake our world*. Hachette, UK
6. Dretske F, Carnap R, Bar-Hillel Y (1953) Semantic theories of information. *Philos Sci* 4(14):147–157
7. Gamboa S (2007) Hume on resemblance, relevance, and representation. *Hume Stud* 33/1:21–40
8. Giraud-Carrier C, Provost F (2005) Toward a justification of meta-learning: is the no free lunch theorem a show-stopper. In: *Proceedings of the ICML-2005 workshop on meta-learning*
9. Goodman N (1946) A query on confirmation. *J Philos* 43(14):383–385
10. Goodman N (1971) *Problems and projects*. Bobbs-Merrill
11. Goodman N (1983) *Fact, fiction, and forecast*. Harvard University Press
12. Hume D (1748) *An inquiry concerning human understanding*. Clarendon Press
13. Igel C, Toussaint M (2005) A no-free-lunch theorem for non-uniform distributions of target functions. *J Math Model Algorithms* 3(4):313–322
14. Joyce T, Herrmann JM (2018) A review of no free lunch theorems, and their implications for metaheuristic optimisation. *Stud Comput Intell* 744:27–51
15. Kubat M (2017) *Introduction to machine learning*. Springer International Publishing
16. Lauc D (2018) How gruesome are the no-free-lunch theorems for machine learning? *Croat J Philos* 18(54):479–489
17. Lauc D (2019) Reasoning about inexact dates using dense vector representation. *Composoft* 8:3031–3035
18. Leibniz GW, Ritter P (1950) *Sämtliche schriften und briefe*. Akademie-Verlag
19. Nielsen F (2010) A family of statistical symmetric divergences based on Jensen’s inequality. [arXiv:1009.4004](https://arxiv.org/abs/1009.4004)
20. Orenstein A, Kotatko P (2012) *Knowledge, language and logic: questions for Quine*. Springer Science & Business Media
21. Pakaluk M (1989) Quine’s 1946 lecture on Hume. *J Hist Philos* 445–459
22. Popper K (2002) *Conjectures and refutations: the growth of scientific knowledge*. Routledge
23. Quine WVO (1970) *Natural kinds*. D. Reidel
24. Rodriguez-Pereyra G (2014) *Leibniz’s principle of identity of indiscernibles*. OUP
25. Schaffer C (1994) A conservation law for generalization performance. *Mach Learn Proc* 1994:259–265
26. Tversky A (1977) Features of similarity. *Psychol Rev* 84:327–354
27. Valiant L (2013) *Probably approximately correct: nature’s algorithms for learning and prospering in a complex world*. Hachette
28. Wolpert D (1992) Stacked generalization. *Neural Netw* 5:241–259
29. Wolpert D (1996) The lack of a priori distinctions between learning algorithms. *Neural Comput* 1341–1390
30. Wolpert D (2013) Ubiquity symposium: evolutionary computation and the processes of life: what the no free lunch theorems really mean: how to improve search algorithms. *Ubiquity* 2
31. Wolpert D, Macready WG (1995) No free lunch theorems for search. Technical report SFI-TR-95-02-010, Santa Fe Institute