Australian Government

Department of Defence
Science and Technology

Cross-validation is insufficient for model validation

Thomas L. Keevers[1]

[1] **Joint and Operations Analysis Division**
Defence Science and Technology Group

DST-Group–TR–3576

ABSTRACT

Cross-validation is the de facto standard for model validation in the machine learning community. It reduces the risk of overfitting and provides an unbiased estimate of the learning algorithm predictive performance. Some people have argued cross-validation, coupled with sophisticated statistical learning methods, has rendered traditional scientific practices irrelevant. In this report, we review the foundations of cross-validation and draw attention to common, but underappreciated, assumptions. We argue that cross-validation is unsuitable for dealing with realistic complications like missing data, theory-laden observations, and malicious input. As a solution, we advocate for a holistic approach to model validation that embraces validation of data quality, acknowledgement of the role of subjective judgement in model assessment, and the use of extended peer review.

Published by

Joint and Operations Analysis Division
Defence Science and Technology Group
506 Lorimer St,
Fishermans Bend, Victoria 3207, Australia

Telephone: *1300 333 362*
Facsimile: *(03) 9626 7999*

March 2019

APPROVED FOR PUBLIC RELEASE

Cross-validation is insufficient for model validation

Executive Summary

Cross-validation is the de facto standard for model validation in the statistical learning and machine learning communities. Data is split into a training set that calibrates the statistical model, and an independent test set that is used to estimate the model's predictive performance. Given the popularity of cross-validation, it is critical to identify any implicit assumptions or limitations of the method.

We argue that cross-validation is unsuitable as a universal method for model assessment. Despite high cross-validation accuracy, artificial neural networks that achieve human- level accuracy in image recognition are vulnerable against adversarial examples, meaning images become misclassified after miniscule manipulation. Likewise, Google Flu Trends was able to accurately predict influenza outbreaks for several years before the model suddenly mispredicted outbreak timings and intensities. These examples show that strong cross-validation performance does not guarantee the model has truly learnt about the phenomena of interest.

Cross-validation assumes that samples are drawn from an independent and identical distribution, an assumption that regularly fails because of hierarchical structure in the model, spatial or temporal correlations in the data, or non-stationary (time-varying) system dynamics. However, cross-validation is unable to detect these violations and may provide an unrealistic and optimistic assessment of predictive performance.

The limitations of independent and identical samples can be overcome by using modified cross-validation procedures. For example, hierarchical models can be tested by performing cross-validation for each level of the hierarchy, and time series can be validated using out-of-sample forecasting with a rolling time window. However, this still requires the correct sampling structure to be identified, which may not be known *a priori*.

Data quality is another fundamental issue. Supervised learning is predicated on having access to the ground truth (the true value or label of the samples). For complex problems the ground truth may be uncertain or contentious. For example, the definitions of diseases in medical science change overtime: Diseases may be split into separate classes, merged into a spectrum, or redefined as new knowledge is acquired. When the ground truth is contentious, test set accuracy is not meaningful as an objective indicator of model correctness, and is better thought of as a check for model consistency.

Data sampling can be misrepresentative of the desired population because of social biases that affect the experimental design or other systematic patterns of missing data. The image classifier Google Photo mislabelled African Americans as Gorillas, while COMPAS software used to determine court sentencings in the United States was allegedly found to

be harsher on African American defendants than Caucasian defendants. Addressing these social and data biases is an active area of research and cannot be meaningfully addressed with cross-validation alone.

An uncritical application of cross-validation leaves the statistical learning and machine learning communities at risk of "Big Data Hubris", "the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis." [Lazer, David, et al. "The parable of Google Flu: traps in big data analysis." Science 343.6176 (2014): 1203-1205.]. Cross-validation can be strengthened by supplementing it with traditional data analysis and sampling techniques.

Statistical learning often treats data collection as a passive process. Greater emphasis on the design of experiments, randomized controlled experiments, instrumentation would reduce the incidence of measurement artefacts and unbalanced data sets that oversample particular sub-groups. These considerations would improve model robustness.

To mitigate against social bias, we advocate for the use of diverse teams and extended peer review. Inclusive teams are more likely to identify potential sources of bias and provide stricter validation of the model's performance than cross-validation alone. For instance, algorithms used for job hiring could be reviewed by equality groups or legal departments. Social bias could be identified through subgroup analysis, although we believe causal models are superior because of their ability to properly identify confounding factors.

Model validation is a difficult issue and further work is required. We advocate for a holistic approach to model assessment that contextualizes the problem, uses extended peer review, and remains grounded in deductive reasoning.

Author

Thomas L. Keevers
Joint and Operations Analysis Division

Thomas Keevers completed a Bachelor of Science (Advanced) with First Class Honours at the University of Sydney in 2011 and earned a Ph.D. in physics at the University of New South Wales in 2016. Since joining Joint and Operations Analysis Division in 2016, Thomas has provided analytic support to numerous defence projects.

THIS PAGE IS INTENTIONALLY BLANK

Contents

Figures

THIS PAGE IS INTENTIONALLY BLANK

Tables

THIS PAGE IS INTENTIONALLY BLANK

1 Introduction

Model credibility is critical in the areas of science, technology, business, medicine, and defence. Each field has developed particular methods of verifying and validating models, depending on their specific domain needs. Defence acquisition relies on identifying capability requirements, performing cost-benefit analysis, and life-cycle analyses of potential platforms [1]. In software engineering, credibility is attained and maintained through regression testing, test cases, integration testing, formal methods and user testing [2]. Operations analysis features assumptions documents or conceptual models, sensitivity analysis, comparison with previous models, and comparison of model output with empirical data [3]. When developing training simulators, key metrics are simulation fidelity, resolution and interoperability. Each field has developed a variety of techniques for attaining credibility and validating their underlying assumptions. This variety of techniques is necessary because of the messy nature of the problems these fields contend with and the competing needs of different stakeholders. In contrast, the validation and credibility of statistical models have largely rested on their predictive performance.

Cross-validation is arguably the most widely used method for assessing predictive performance in statistical learning and machine learning [4]. Data is split into a training set that calibrates the statistical model, and an independent test set that is used to assess the model. The training or in-sample performance is usually superior to the test set or out-of-sample performance because the data is used for both model construction and assessment. Cross-validation can be used for parameter estimation, model selection, or to provide an unbiased estimate of general predictive performance. However, these tasks cannot all be performed simultaneously using a single test set because information will leak from the test set to the model. When model selection and an estimate of predictive performance both need to be made more complicated forms of cross-validation, like nested cross-validation, are sometimes used.

Cross-validation provides estimates for two separate quantities, namely expected learning algorithm and model performance. These quantities are sometimes known as the expected and conditional test errors [4], or the prediction error and expected value of prediction error [5]. The learning algorithm performance averages over the possible training sets, while the model performance is conditioned on a single test set (see Fig. 1). In both cases a quantitative score or metric is used to assess the out-of-sample (test set) performance, whether it be accuracy, mean-square error, or some other kind of quantitative metric. The standard method of multi-fold cross-validation mixes these two quantities together, so doesn't have a clear statistical interpretation.

The elegance and simplicity of cross-validation is sometimes perceived to render traditional model-theoretic considerations unnecessary. This attitude is exemplified in Chris Andersens provocative essay *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* [6],

> *"Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."*

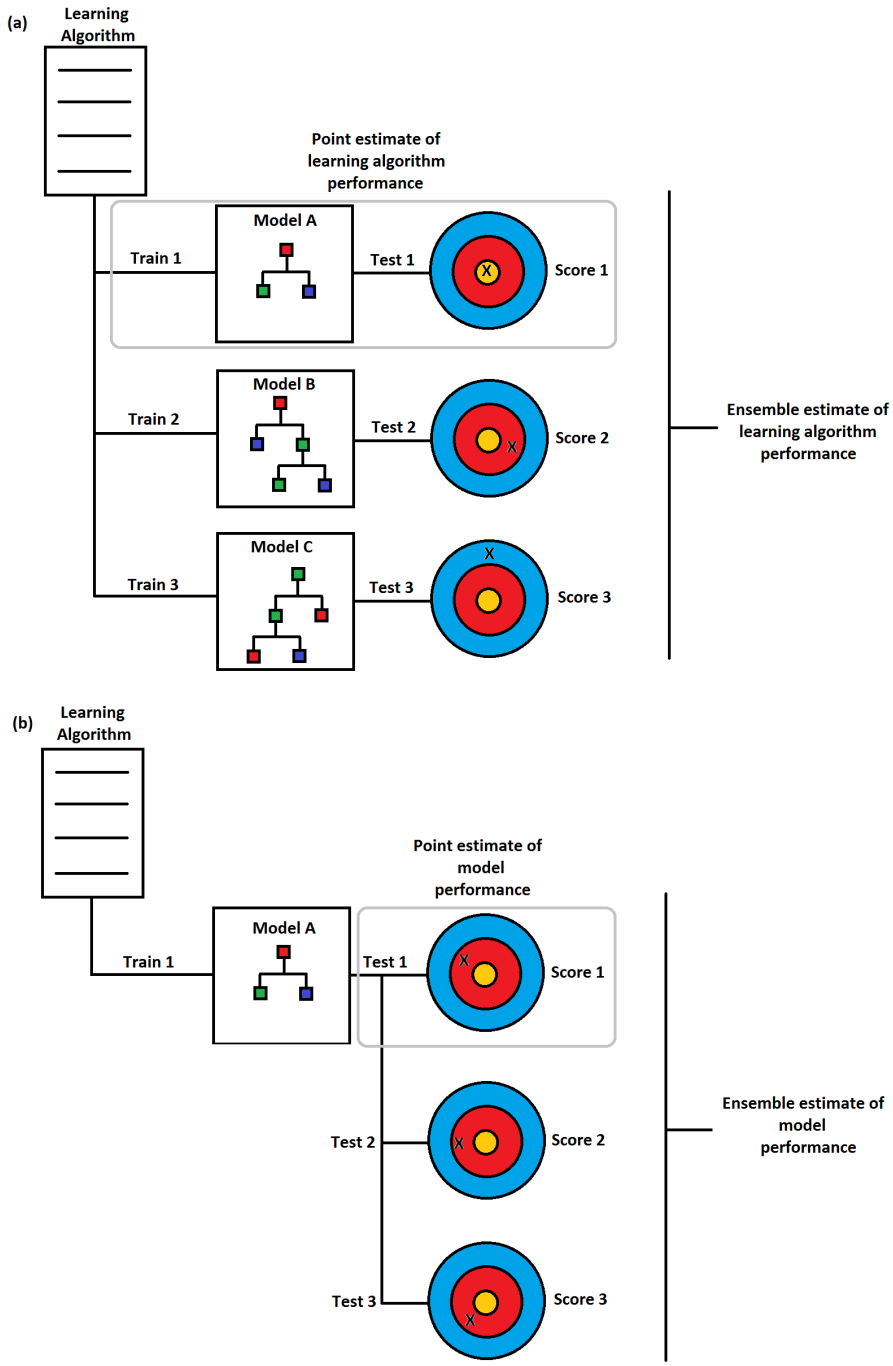

Figure 1: Procedures for estimating the learning algorithm and model performance. a) To estimate the performance of a learning algorithm, both the training and test sets must be resampled. Each train and test pair provides a point estimate of the learning algorithm performance, while an ensemble of pairs allows confidence intervals to be derived. b) To estimate the performance of a specific model only one training set is used. A point estimate of the performance can be derived from a single test set or a confidence interval from multiple test sets.

The intended purpose of this paper is to clarify the role of cross-validation and to evaluate its adequacy as a method for model validation. We argue that statistical learning would be more credible and reliable if it embraced a multi-dimensional approach to validation, as can be seen in other fields.

It is critical to understand the role that out-of-sample performance should play in model formulation and assessment given the emerging importance and popularity of cross-validation as a model assessment method. An unthoughtful movement towards algorithmic analysis leaves the scientific community at risk of *"Big Data Hubris"* [7],

> *"the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analys... quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data."*

We contend that cross-validation works well for problems with unambiguous performance metrics, notably competitive games like chess, checkers, and go [8], where all stakeholders in the model's application agree on the metrics that determine 'good performance'. However, recent work indicates that cross-validation is unsuitable as a standalone method for open world problems with ambiguous or multiple competing objectives, such as designing an efficient and equitable transport system.

Cross-validation may fail to detect fragility in the underlying models. Neural networks that achieve human-level accuracy in image recognition are vulnerable against adversarial examples, meaning images become misclassified after miniscule manipulation, despite high cross-validation accuracy [9]. Another example of cross-validation failing to provide adequate validation is seen from the failure of Google Flu Trends. Google Flu Trends was able to accurately predict influenza outbreaks for several years before the model suddenly mispredicted outbreak timings and intensities [7, 10, 11]. These examples demonstrate that cross-validation cannot provide model validation in a general sense and that supplementary validation is necessary.

There are additional reasons to be wary of cross-validation. It can be ineffective for model selection when multiple candidates have similar levels of predictive performance or when the predictions vary only for small sub-groups [12]. Cross-validation tends to select models that are too complex [13, 14], and is even 'asymptotically inconsistent' for linear models, meaning it will fail to select the correct model for arbitrarily large data sets [15]. Rather than selecting a model based purely on its test set performance, it seems preferable to account uncertainty in the estimated predictive performance and potentially create ensembles of models, such as through model stacking [16].

This paper reviews the suitability of cross-validation as a sole method of model assessment. Several issues that can degrade the usefulness of cross-validation are listed in Table 1. These issues were aggregated from surveying the literature. They are not derived in any systematic fashion, and are not intended to provide a complete list of all the issues that need to be considered when using cross-validation.

This paper is organized so that each section addresses a rough grouping of these issues. The categories are designed to aid the flow of discussion, and the assignment of issues to

categories is somewhat fluid. The identified categories are not intended to be optimal or unique, and could potentially be restructured to accentuate different sets of commonalities or differences for the set of issues.

Section 2 examines **data quality**, which covers issues that arise from poor quality or mislabelled data. Section 3 reviews issues associated with **data sampling**, and covers violations of the assumption of independent and identical samples. Section 4 reviews **modelling** and focuses on the implicit assumptions of supervised learning. Section 5 discusses the impact of **analyst degrees of freedom**, such as subjective decisions made by the analyst, data interpretation, and using the model to make decisions. Section 6 discusses model requirements that exist outside predictive accuracy, specifically the role of causality and adversarial examples. Finally, section 7 summarizes the issues discussed and suggests future work.

Although this paper focuses on cross-validation, many of the arguments we present can be applied to other methods of statistical validation, including Bayes factors, p-values, confidence intervals, and model complexity penalization. These methods fundamentally guard against overfitting of the data, but do not address systematic errors arising from data bias, malicious inputs, or non-stationary processes.

Issue	Explanation	Example	Solution
Data quality			
Label contamination	Incorrect ground truth values	Mislabelling in image data sets	Improve data quality
Poor data quality	Data may have systematic errors, a low signal-to-noise ratio, drift, or omit important variables	Sensor miscalibration	Improve data quality
Missing data	Samples are not representative of population	Survivorship bias, truncation, censoring, patient dropout in medical studies	Experimental design, data imputation
Publication bias	Statistically significant results more likely to be published than non-significant results	Publication bias in medical science and psychology	Study preregistration, publication of negative results
Data sampling			
Interacting measurements	Measurements affect the system	Placebo effect	Use theory to guide measurements
Violation of independent and identical samples	Samples are correlated or come from multiple populations	Hierarchical models, systems with spatial structure	Use modified cross-validation procedures
Non-stationarity	The system's behaviour changes over time, possibly through drift or sudden shocks	Financial markets change overtime, particularly in response to new information or regulations	Continuous system monitoring, adaptive systems, use modified cross-validation procedures
Modelling			

Model fragility	Trained models may show a significant drop in performance when the task is slightly modified or the model is highly sensitive to the training set	Deep-Q fails when Atari games are slightly modified, model generated by a small number of noisy measurement	Limit model application, contextualize data, ensemble method, regularization, reduce the model complexity, experimental design
Model non-identifiability	Several models produce similar predictions for the test set, but differ outside this region	Unbalanced data sets	Restrict model domain
Analyst degrees of freedom			
Cognitive and social biases	Systematic patterns of irrational judgement	Confirmation bias, hindsight bias, halo effect, anchoring, overconfidence, optimism bias	Diverse teams and extended peer review, improve training
Data misuse	Data set is misused, potentially resulting in an inflated test set accuracy	Incorrect data pre-processing, reusing test sets, data leakage	Checklists and disclosure statements, peer review, improve training
Proxy quantities	Features and labels used in supervised learning are not the true quantities of interest	Economic performance cannot be measured directly, so indicators like unemployment and gross domestic product are used	Consult subject matter experts
Poor metrics	Performance metric does not properly capture real-world value	Using accuracy to validate a classifier for credit card fraud	Analyse multiple metrics, consult subject matter experts, improve training

Spurious inferences	Using cross-validation (predictive performance) to make potentially spurious inferences about other model characteristics	Feature selection in high-dimensional problems can be unstable or vary sharply with the data set size, and not indicative of the parameters that actually influence a situation	Sensitivity analysis, improve training
Non-predictive performance			
Malicious input	Strong cross-validation performance does not ensure learning algorithms are robust against adversarial action	Adversarial examples like those seen in the image recognition domain	Limit model domain, use models that are robust against adversarial examples or poisoned data sets
Model interpretability	Difficult to ascertain how a statistical model makes a prediction or recommendation	Neural networks	Use interpretable models, develop surrogate models, use model-agnostic measures of feature importance
Causality	Causal behaviour may be of primary interest, but unidentifiable from observational data	Social policy intervention	Development of new causal models, perform randomized experiments

Table 1: The table lists issues that can reduce the effectiveness of cross-validation as a model validation method, along with examples of when the issue may occur, and potential ways to address the issue.

2 Data quality issues

Cross-validation works well when the data is close to the ground truth and predictive performance is the primary aim, but is less useful when the data is of poor quality. This could manifest as a low signal-to-noise ratio, biased data, or unbalanced data sets. In this section, we review some of the data issues that can occur in statistical learning.

2.1 Label contamination

Supervised learning is predicated on having access to the ground truth. For some problems the ground truth is unknown, controversial, or may change overtime. Schizophrenia and bipolar disorder were originally recognized as separate conditions, but are now considered to form a spectrum [17]. An algorithm that assigns the labels of 'schizophrenic' and 'bipolar' could have been considered accurate in a historical context, but would now be deemed to perform poorly because of the bimodal nature of its output. The perceived performance can decrease because of changes to the user's beliefs, rather than any changes to the underlying model.

In other situations there are conflicting views about what schema is appropriate. A recent Stanford study that claimed to detect sexuality from dating site photographs was criticized by the LGBTI and statistics communities for using a binary classification scheme for sexuality (see [18, 19, 20] for several lines of criticism). When the ground truth is contentious, test set accuracy is not meaningful as an objective indicator of model correctness, and is better thought of as a check for model consistency.

2.2 Poor quality data

Poor quality data - "dirty data" - leads to models with low credibility. There may be several causes of poor data quality. For example, sensors may be miscalibrated for mechanical systems, or there may be missing entries because of poor data collection practices. In chemometrics, the modelling errors are usually smaller than physical sampling errors that arise from calibration or instrumentation errors and undetectable environmental variation that occurs between experiments performed at different times or in different laboratories [21]. These errors are not addressed adequately with cross-validation and produce a misleading picture of experimental accuracy.

Poor quality data can lead to models with good cross-validation scores that provide poor real world predictions. Cross-validation ensures consistency between the data and the model, and is unable to validate the data quality, even though it strongly affects the predictive performance.

2.3 Missing data

There are several forms of missing data. The simplest case is 'missing completely at random', meaning each feature of a data point has a fixed probability of being excluded, independent of its value [12]. These missing data points simply degrade the size of the data set, and no special care needs to be taken. Missing features can be imputated or 'missing' can even be treated as a separate feature value. Cross-validation will still function effectively when data is missing completely at random.

A more difficult case is 'missing at random' in which the probability of a data point being included in the sample depends on the data points features, but not its class or value [12]. Imagine creating a model to predict a person's income with only males included in the data set, though irrespective of their income [12]. This pattern of missing data

will falsely inflate the test set. The supervised learning algorithm will achieve high test set performance for the male sub-population that forms the test set, while potentially performing much worse for female data points that occur in deployment, but are not evaluated during cross-validation. In this situation cross-validation will not provide a good estimate of predictive performance unless the male and female sub-population are similar.

The most difficult case is 'missing not at random', meaning the probability of data inclusion depends on both the features and class or value. Extending the previous example, the probability of inclusion could depend on both the persons gender and their income. Without a guide to the missing data mechanism, the model selection process will be sensitive to a range of untestable assumptions. These kinds of issues are prevalent in medical studies in which researchers have to deal with non-responders, drop out of participants and observer bias [22]. Given the potentially large differences between the sampled and target populations, cross-validation is nearly worthless as a key metric without background information or careful caveating of the results.

2.4 Publication bias

Scientific studies are more likely to be published when they produce a statistically significant result. Non-significant results receive less attention, and are analogous to missing data points. This is known as publication bias and has several negative effects. In medical science, publication leads to unnecessary replication and tends to inflate apparent efficacy of medical treatments [23]. Ioannidis argues that publication bias is so strong that the majority of biomedical literature is false [24], while Gelman and Loken report that for low power studies, true effects can be an order of magnitude less than their reported strength (on average) and may have a 40 percent chance or higher of reporting the wrong sign for an effect [25]. We expect publication bias to also affect the statistical learning literature, although perhaps in different ways.

Researchers routinely use benchmark data sets to assess the performance of new algorithms against those in the literature, but in many cases the new algorithm will only be reported if its performance exceeds that of the other algorithms, which is a form of publication bias. This selection-induced bias misrepresents the true performance of the algorithms. Additionally, the reuse of the data sets also leads to the test set going 'stale' because of test set information gradually leaking out [26, 27, 28]. Even when individual researchers correctly employ cross-validation, it is possible for higher level effects, such as the file drawer problem, to produce significant bias in the estimated predictive performance.

2.5 Addressing data quality issues

Data quality is a perennial issue in statistical learning. While data quality is often thought of in terms of the signal-to-noise ratio or measurements artefacts, it can also arise through biased data collection and reporting or because of inappropriate paradigms that reinforce existing biases.

Improved instrumentation and using theory to guide data collection is necessary because measurement and observation is a 'theory-laden' process that includes subjective choices and implicit assumptions. Theoretical frameworks encourage particular methodologies for data collection and questions of interest. A data-driven approach amplifies these biases and implicit assumptions, since they become hidden from view. Statistical learning should be integrated into research communities as part of a balanced research portfolio, rather than trying to serve as an alternative scientific paradigm.

3 Data sampling issues

Standard cross-validation assumes a fixed data sampling mechanism with an independent and identical sampling distribution. If the training and test sets are strongly correlated, then the cross-validation score may be artificially inflated. This is particularly common when there is temporal or spatial correlation in the data, although some forms of cross-validation can accommodate these correlations [29]. Below we examine how cross-validation can be extended to situations in which the standard sampling assumptions fail.

3.1 Violation of non-interacting measurements

Measurements are usually treated as passive processes that characterize an object's properties without affecting them [12]. In some situations, measurements will disturb the system, causing its state to change. Interactive measurements occur at the quantum mechanical level through Heisenberg's uncertainty principle, while at the macroscopic level they can be found in the medical domain where double blind experiments are needed to suppress the placebo effect.

Interaction effects are especially important in stock trading, where analysts want to discover patterns of profitable trading. One can naïvely validate the historical profitably of trading strategies by performing out-of-sample forecasting. However, the process of buying or selling shares affects the subsequent stock price, so cross-validation will systematically misestimate the counterfactual profitability of trading strategies. Cross-validation needs to be augmented with additional data or modelling to account for the effect of placing orders. This type of augmentation is necessary whenever interventions or measurements have secondary effects on the system behaviour.

3.2 Violations of independence in hierarchical models

Hierarchical models are appropriate when the population of interest has multiple levels of structure. For concreteness, this could consist of variation in student grades caused by variation in student-level factors (hours of study, each student's intrinsic ability) and school-level factors (teacher quality, school funding), which could be hierarchically modelled with each student-level factor nested within a school-level factor (see Fig. 2a) [12].

Figure 2b shows naïve cross-validation for a hierarchical model. The test set is represented by the dots with red crosses through them and are randomly sampled from the complete

data set. Some students are drawn from schools within the training set, while others are drawn from previously unobserved schools. The test set accuracy will average over these two conceptually distinct populations, which is sensitive to the sampling process and potentially not the quantity of interest.

Cross-validation can be performed in a more structured fashion, of which two examples are shown in Fig. 2c and d. These are more robust against changes in the sampling mechanism and are easier to interpret. Leave-one-student-out cross-validation is shown in Fig. 2c. Since each school has multiple students, this provides an estimate for our ability to predict the performance of future students within schools that have already been observed. In contrast, Fig. 2d performs leave-one-school-out cross-validation, and provides an estimate of our ability to predict the performance of unobserved schools.

Even with these two alternative procedures, cross-validation of hierarchical models is still challenging, as summarized by Wang and Gelman[30],

> *"[the] lack of clear protocol for the cross-validation procedure: to truly test the model, the holdout set cannot be a simple random sample of the data but instead needs to have some multilevel structure itself, so that entire groups as well as individual observations are held out it is not clear how best to subsample structured data for cross-validation in a general way... Our results illustrates that under multilevel structure, it could be tricky to use cross validation in model selection, as the size of the data and how balanced the structure is heavily affect the relative performance of the models."*

To summarize these issues, cross-validation of hierarchical models is challenging because the out-of-sample performance is sensitive to the data sampling mechanism, difficult to interpret, and contains elements of subjectivity in the validation process. This substantially weakens the case for using standard cross-validation as a default model validation technique without any thought of the underlying structure. It additionally shows that model validation (in addition to modelling) has elements of subjectivity.

3.3 Violations of independence in time series and structured data

Cross-validation cannot be naïvely applied to time series or other forms of structured data, like satellite imagery, that have strong auto-correlation. Conventional cross-validation randomly assigns points to the test set (Fig. 3a). However, this will provide an optimistic estimate of the model's predictive performance because of the strong temporal correlations between neighbouring points.

The most common alternative is out-of-sample forecasting, which separates the training and test sets into adjacent blocks, potentially with a gap between the two sets to reduce the impact of autoregressive behaviour (Fig. 3b) [31]. Nonetheless, both standard cross-validation and out-of-sample forecasting assume stationarity [32], which is absent from many real time series [33]. Estimates of predictive performance may latently assume stable distributions that exaggerate the true predictive capabilities of statistical models.

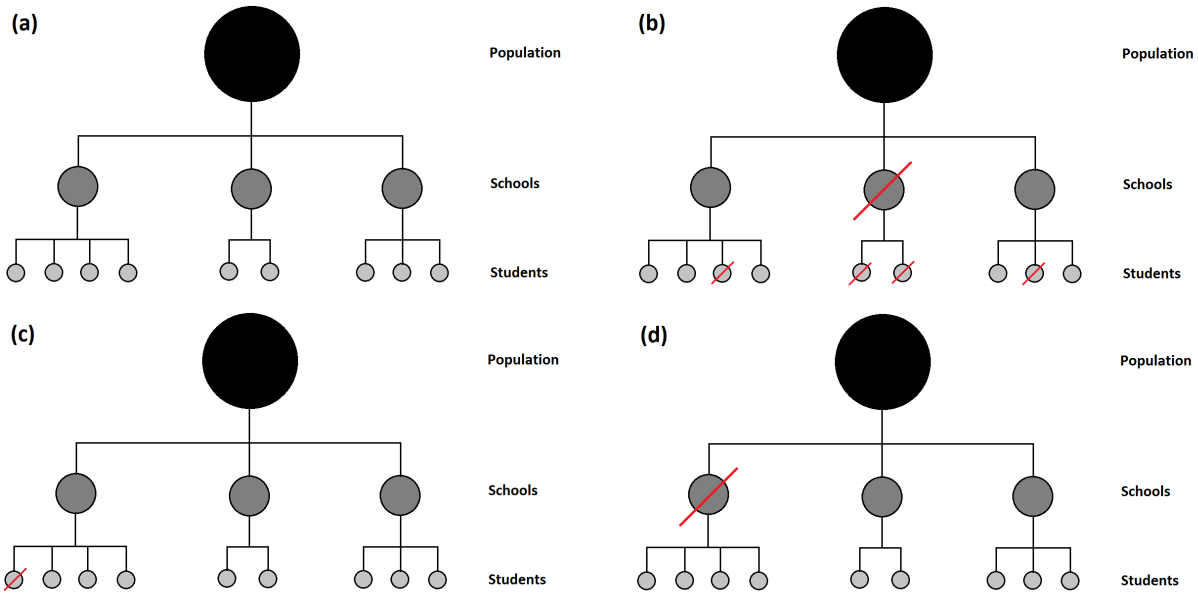

Figure 2: A graphical representation of a hierarchical model for variation in student marks with inter-student and inter-school level variation. (a) A graphical representation of an example data set with several schools and students. (b) Several entries of the data set (marked by red crosses) have been removed to form a test data set for cross-validation. The remaining entries form the training data set. (c) Leave-one-out cross-validation at the student-level for a hierarchical model. The cross-validation score will be indicative of the predictive performance of students for schools already within the training set. (d) Leave-one-out cross-validation at the school-level for a hierarchical model. The cross-validation score will be indicative of the predictive performance of students for schools outside the training set.

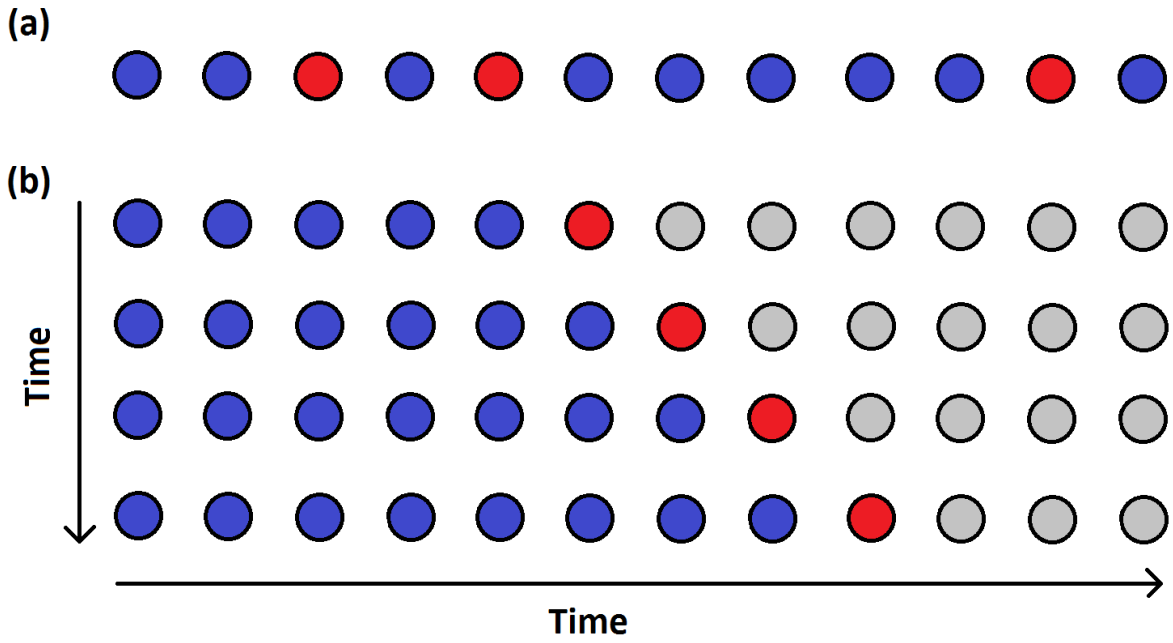

Figure 3: Cross-validation for time series data. The training set is shown by the blue dots, the test set is shown by the red dots, and data points excluded from both sets are shown in grey. The time series proceeds from left-to-right (early-to-late). (a) A naïve application of cross-validation to time-series data in which data points are missing at random. (b) A structured approach to cross-validation that forecasts the data point one time step ahead. Figure adopted from Rob Hyndman [34].

The difficulties of estimating predictive time series performance are exemplified by Google Flu Trends. Google Flu Trends was developed to predict influenza activity from the relative frequencies of Google searches. It initially seemed that the model could predict disease outbreak with performance similar to that of the Centre for Disease Control (CDC) in the United States, but with greater responsiveness. The model accurately reproduced CDC figures in 2007. However, between 2009 and 2012 Google Flu Trends mispredicted the timing and intensity of influenza activity [7, 10, 11].

The exact point of failure is unclear, although the effects of a non-stationary (changing) environment are suspected. The search engine itself is continually updated and social dynamics naturally evolve over time. It is therefore not so surprising that a model that in earlier years seemed so effective became miscalibrated when the system changed, but this was something that cross-validation failed to detect.

3.4 Addressing data sampling issues

Data quality and quantity are inevitably issues because of the high-dimensionality of contemporary statistical models, while violations of identical and independent sampling are difficult to identify without some preconceptions about the structure of the violations. We suggest addressing these issues by trying to improve the data, while also using more

general validation techniques to reduce the blindspots of cross-validation:

Experimental design can be used to ensure that the data will be representative of the desired population and that a balanced data set will be collected that will assist in model identifiability (discussed in Section 4.2). Stratified sampling and latin hypercube designs can improve the precision of parameter estimation and are especially valuable when subgroups have significantly different sampling frequencies.

Standard cross-validation assumes random variables are independent and identically distributed. *Developing variants of cross-validation* to accommodate more complicated data collection schemes would allow cross-validation for a greater class of problems; examples of modified cross-validation schemes were discussed for hierarchical models and time-series data. New variants could encompass an even greater variety of systems.

To accommodate non-stationary environments, we recommend using *adaptive systems* that are able to modify their behaviour in response to gradual or sudden changes in the environment. Simple examples include sensors periodically recalibrating themselves or state space models that allow their internal parameters to change overtime as data is collected [33].

Continuous monitoring can also be used to check for sudden degradation in system performance. In the context of non-stationary environments, cross-validation should be thought of as an initial, not a final, model assessment. True validation comes from monitoring the system after it has been deployed. Anomaly detectors or step-change detectors can serve to bolster human supervision, and detect when the system behaviour is likely to have changed [35, 36].

Finally, *subject matter experts* can provide the boundary conditions around a statistical model's development and deployment. They are able to improve the model's performance through feature selection, adding prior knowledge into the model, or validating the data quality.

4 Modelling issues

Learning algorithms are an integral part of statistical learning. They can improve model accuracy by imposing realistic constraints on the model structure or degrade model performance by making spurious assumptions. Differentiating between errors arising from the data and from the learning algorithm is not always a clear cut issue. For example, noisy data could be blamed for inaccurate predictions, but in theory this could be rectified with more regularization. In this section, we look at the difficulties of developing robust and unique models.

4.1 Model fragility

Statistical models can be fragile against small perturbations in the task or the environment. Reinforcement learning systems like Deep-Q can outperform human experts in a wide range of tasks [37]. Surprisingly, many of the trained models will catastrophically fail when transferred to similar tasks. For example, Deep-Q's ability to play the game 'breakout' is

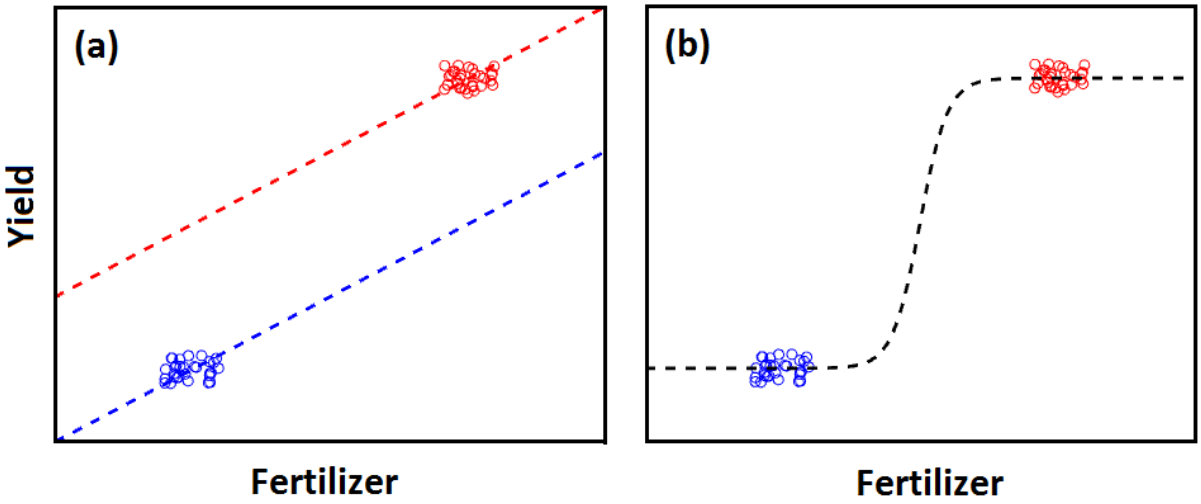

Figure 4: The effect of fertilizer on crop yield. Crops in a sunny patch are shown by the red dots and crops in the dark patch are shown by the blue dots. Because of an unbalanced design, the data can be well described with a linear relationship between the fertilizer amount and the crop yield with a constant offset from the lighting, or by a sigmoidal function of the fertilizer and no term from being in a sunny/dark patch. Adapted from Gelman, et al. [12].

disrupted when the paddle is moved by only a small amount, demonstrating the system is unable to generalize outside the narrow training set [38]. This highlights the need for the out-of-sample test set to closely match any intended applications, something that is difficult to guarantee for many realistic applications.

4.2 Non-identifiability

The inability to differentiate between incompatible statistical models from data is called "model non-identifiability". Multiple models can have high out-of-sample performance while having significantly different structures. This can be driven, for instance, by unbalanced data sets.

Consider an experiment to determine the effect of fertilizer on crop yield when the fertilized crops are in sunny plots and all the unfertilized crops are in the dark plots. Plants in the sunny and fertilized plots produce consistently higher yields than the dark and unfertilized plots. Since the level of sunlight and fertilizer are confounded, it is impossible to deduce whether sunlight, fertilizer, or an interaction effect is at work [12].

High predictive accuracy can be obtained by modelling the data by two offset linear relationships that indicates that sunlight produces a fixed effect (Fig. 4a) or a single sigmoidal relationship that indicates sunlight has no effect (Fig. 4b) even though they reach opposite conclusions about the importance of fertilizer. Although the example is somewhat contrived, it clearly identifies a potential hazard when extrapolating from out-of-sample accuracy to inferences regarding feature selection or the adequacy of alternative models.

4.3 Addressing modelling issues

As general advice for improving model quality, we believe there is scope to *contextualize data*. It is natural to abstract data into a purely mathematical form when performing statistical modelling. While this is appropriate for the manipulation of data, it is hazardous to the model's interpretation and application. As was demonstrated by the effects of publication bias, data needs to be understood with respect to the population that is being sampled and also with respect to previous data collection or analysis.

Statistical models will perform at a higher level and in a more consistent fashion when we *limit the domain of application*. Statistical models can even forget previous knowledge when trained on novel situations [39]. Rather than invest substantial effort in trying to develop more robust and general statistical models, it is possible to improve the performance of statistical models by limiting the environmental variation they encounter. This is a common approach seen in factories, where automated systems perform repetitive tasks with virtually no autonomy or awareness.

5 Analyst degrees of freedom issues

Statistical models usually incorporate a mixture of subjective and objective elements. The objective elements include the data, while the subjective elements include the choice of performance metric or the choice of learning algorithm. In this section, we examine the impact of subjective choices made by the analyst, which we call the analyst degrees of freedom, on the model performance.

5.1 Cognitive and social biases

A pervasive myth is that statistical learning is inherently unbiased because it is an (objective) mathematical process. The essential flaw in this argument is that while statistical concepts are mathematical, the inferences are derived from biased data sets and models. This form of bias is grounded in social attribution and judgement, and is distinct from bias in the mathematical sense of the bias-variance trade-off [4].

The ability for statistical models to learn, or even accentuate, social biases has been highlighted by a number of recent news stories. African Americans were mislabelled as Gorillas by image classifier Google Photo [40], while Correctional Offender Management Profiling for Alternative Sanction (COMPAS) software used to determine court sentencings in the United States were allegedly found to be harsher on African American defendants than Caucasian defendants [41, 42] (although see Kusner, *et al.* [43] for discussion about the difficulties of defining algorithmic fairness).

Statistical models can acquire biases in a number of ways. Learning is primarily an inductive activity in which samples or training data are generalized from samples to populations. If the samples are biased or non-representative, the algorithm will learn to faithfully reproduce these biases. This can occur even when the data is correctly labelled, albeit with an incorrect base rate. The misclassification by Google Photo may have occurred

because there were more 'Gorilla' than 'African American' photographs in the training set, so the overall cross-validation accuracy for all the photographs was high, even though it consistently mislabelled the small African American subset.

In other cases, the data set may become distorted due to social influences. New experimental drug treatments are predominantly tested against young males [44]. Meanwhile, older patients are often excluded because of pre-existing medical conditions [45], while females can be excluded because of issues related to pregnancy and menopause [46]. Consequently, clinical research may fail to detect important gender-dependent treatment effects. Although the data collection processes might be rational at the individual level, they systematically distort the funding and assessment of medical science. These kinds of biases are readily learnt by statistical models without any self-correction mechanisms.

5.2 Data misuse

Data is an essential element of statistical learning. However, poor quality data can be worse than none at all. Spurious models can be constructed because of poor quality or contaminated data, or because of misuse by the analyst, commonly referred to as 'p-hacking' in the statistics literature [47, 48].

It is well-known that statistical significance in frequentist statistics can become inflated when multiple comparisons are performed. What is less recognized is that the same effect can occur when only a single comparison is performed, but the particular comparison is conditional on the data set. This effect was investigated by Gelman and Loken and termed the "garden of forking paths" [25]:

> *"researchers can perform a reasonable analysis given their assumptions and their data, but had the data turned out differently, they could have done other analyses that were just as reasonable in those circumstances... Our key point here is that it is possible to have multiple potential comparisons, in the sense of a data analysis whose details are highly contingent on data, without the researcher performing any conscious procedure of fishing or examining multiple p-values."*

The garden of forking paths has a counterpart in statistical learning, where p-values are replaced by out-of-sample performance. Data usually needs to be wrangled, cleaned, or preprocessed in some form before it can be modelled. There may also be conditional postprocessing to remove samples dependent on a model's ability to discriminate them. For an image recognition task we could imagine removing difficult-to-classify samples with atypical image filters applied to the image or a poorly focused photograph. Even though removing these types of samples may be reasonable, it provides opportunity for the test set accuracy to become inflated.

The garden of forking paths can also occur when the type of model (for example a neural network or random forest) or the hyperparameters are chosen after looking at the data set. These practices, while necessary to formulate usable data sets, can provide hidden opportunities to perform conditional operations or analysis on the data set, and inflate

the test set accuracy. The garden of forking paths can be addressed by pre-registering the study [49] or by performing a 'multi-verse' analysis of potential analysis pipelines [50], although both require additional work and can be infeasible for exploratory analysis.

Improper data preparation may also lead to 'data leakage' in which a model is trained on information it would or should not be able to access when deployed. This could include unintentional inclusion of the class as a feature [51], normalization of the data prior to splitting it into train and test sets [4], or inclusion of inappropriate data points [51]. There may also be anomalous features in the data set that can unrealistically bolster the accuracy. Ribeiro, *et al.* [51] report that the Random Forest learning algorithm is able to unrealistically differentiate between articles about "atheism" and "Christianity" in the "20 newsgroups" data set because the word "Posting" occurs frequently in the header of the "Christianity" articles, a quirk that is unlikely to be replicated in later data sets. In these circumstances, high test set accuracy can be achieved, but is not indicative of generalized performance.

5.3 Proxy quantities

Sometimes it is impossible to directly measure the quantity of interest, so proxy quantities (indicators or related quantities) are measured instead. This occurs even in simple cases, like measuring current and voltage to determine a device's electrical resistance. A more complicated case is trying to infer the overall state of a country's economy from related values like the gross domestic product (GDP) or the unemployment rate. A high cross-validation score indicates the model is able to accurately reproduce the proxy quantities. This will be useful only if the proxy quantities are actually indicative of the true quantities of interest, which cannot be determined through cross-validation.

5.4 Poor metrics

Predictive performance must be characterized with some form of metric - a measure of the difference between the observed and desired output. For example, an image classifier may be assessed by its accuracy. This can create problems when the chosen metric is not aligned with the intended application. A classic example is a fraud detection system that attains high accuracy by always classifying a transaction as legitimate [52]. Because of the asymmetric rate of non-fraudulent activity, the model is accurate but functionally useless. A better performance metric for this scenario would be recall. Cross-validation will provide a misleading measure of performance when the performance metric is poorly chosen.

When the performance metric is contentious, it is possible to perform multi-objective optimization, rather than choose a single objective function. For example, when choosing a car there may be a trade-off between cost, fuel efficiency and size. However, there is no general method for optimizing every performance metric simultaneously. This can be partially addressed by choosing a 'most important' metric to optimise [53], creating a 'supermetric' that weights each metric [53], or finding the Pareto optimal solutions in which no metric can be improved without degrading another [54]. But none of these methods

completely eliminate or solve the subjectivity of choosing an appropriate performance metric.

5.5 Addressing analyst degrees of freedom issues

Our recommended strategy for addressing the analyst degrees of freedom primarily involves improving the analysis, identifying potential issues before the data is collected or analyzed, and making the underlying assumptions more transparent after the analysis so the work can be embedded in the proper context.

To address social and cognitive biases we recommend using *diverse teams and extended peer review* [55]. Inclusive teams are more likely to identify potential sources of bias and provide stricter validation of the model's performance than cross-validation alone. For instance, algorithms used for job hiring could be reviewed by equality groups or legal departments.

The use of *checklists and disclosure statements* can guard against the garden of forking paths and other forms of p-hacking [56]. The significance of out-of-sample performance is quickly diluted when data is conditionally pre-processed or cross-validation is performed multiple times on a single data set. Researchers should include checklists or disclosure statements to ensure the numerical figures of merit provide an accurate picture of predictive performance. Similar practices occur in medicine with study preregistration ensuring details about data collection and analysis are collected prior to the study [56] and in solar cell research where guidelines were developed to ensure meaningful comparisons could be made across multiple types of solar cells [57]. For statistical learning this would include, among other things, all analyses performed to avoid reporting biases and the reasoning behind choosing particular metrics to reduce the likelihood of 'cherry picking' the best performing metric.

There is also scope to improve the *education and training* of analysts. P-values are one of the first concepts encountered in statistics education, but it is commonly misunderstood and misused [58]. Some commentators have even recommended abolishing p-values [59, 60, 61, 62]. To the best of our knowledge similar research hasn't been done to test understanding of cross-validation in the statistical learning community. However, we suspect many of the same types of issues occur in statistical learning and could be addressed with better training.

6 Non-predictive performance

Cross-validation assesses the predictive performance of a model. However, this can be insufficient when forming counterfactual inferences for which no data exists or when the method of prediction is important,

> *"there are situations where a directly empirical approach is better. Short term economic forecasting and real-time flood forecasting... However, much prediction is not like this. Often the prediction is under quite different conditions*

> *from the data; what is the likely progress of incidence of the epidemic of v-CJD in the United Kingdom, what would be the effect on annual incidence of cancer in the United States of reducing by 10% the medical use of X-rays, etc.?"*[63]

Cross-validation also assesses a model's resistance to malicious inputs, nor the validity of the decision making process, and whether it is 'fair' or logical. These concerns are particularly acute when statistical models are integrated in critical systems or in the social domain where concerns of equality and fairness are prevalent.

6.1 Malicious input

Statistical models are susceptible to malicious inputs. For example, modern image recognition systems can identify images with human-level performance. But, despite their high accuracy, these systems can be fooled into misclassifying otherwise easy examples through miniscule, but well-crafted, image manipulations [9]. Changes in pixel brightness below human perception can cause, for example, a bus to become misclassified as a stop sign. Similarly, images that appear as random noise to humans can be classified with high confidence as meaningful objects. These manipulations have been demonstrated for physical systems: Placement of white tape on a stop sign caused it to become misclassified as a speed limit sign by some image recognition software [64].

It is difficult to comment on the long-term implications of adversarial manipulations. A variety of countermeasures are actively being developed [65]. Regardless of whether these methods are ultimately fruitful, this shows that malicious input cannot always be addressed through conventional cross-validation and more general methods of validation are required.

6.2 Model interpretability

The need for model interpretability can be driven by legal or ethical requirements. For instance, there are strong restrictions around what information people can use for investing or hiring. These same restrictions apply to statistical models too, at least in principle. However, it is impractical to check compliance without any insight into the underlying principles of operation.

Model interpretability can also be driven by pragmatic considerations, namely assessing the veracity and robustness of the statistical model. Statistical folklore has it that an early neural network attained high accuracy in differentiating photographs of United States and German tanks [66]. Later it was discovered that the network had no concept of tanks - all the United States tanks were photographed in daylight, while all the German tanks were photographed at night. The algorithm was merely detecting the time of day. Despite high test set accuracy, the model had little predictive power. Model interpretability would show that the neural network was trained on the wrong features, allowing the analyst to avoid an embarrassing mistake.

Similar stories of confounded prediction abound in the statistical literature. For example, there have been recent claims of neural networks accurately determining people's sexuality

	Treatment A	**Treatment B**
Males	4/32 (13%)	11/68 (16%)
Females	61/68 (90%)	31/32 (97%)
	65/100 (65%)	42/100 (42%)

Table 2: Two potential medical treatments are tested on male and female patients. Treatment B outperforms Treatment A for both the male and female subgroups. However, when male and female patients are pooled together, Treatment A outperforms Treatment B. This is an example of Simpson's paradox.

and criminality from facial photographs [18, 67]. Critics claim the strong cross-validation performance is driven by confounding features within the data sets, and that the models provide little real predictive power [19, 20, 68]. Greater model interpretability would assist in evaluating these claims by allowing different explanation to be properly tested.

6.3 Causality

In the subsequent sections, we review three statistical paradoxes that show using correlation to make decisions or interferences about causal relationships can be misleading. These examples show that treating statistical learning methods as blackboxes can potentially lead analysts astray and support poor decision making.

6.3.1 Simpson's paradox and causality

Statistical methods generalize behaviour from observational data and are severely limited in their ability to detect causal mechanisms. As the adage goes 'correlation does not equal causation', and causation can even occur independently of or even in opposition to correlation. A famous example that demonstrates the critical distinction is the Berkeley admission controversy [69]. Data showed that males were admitted at consistently higher rates than females, implying discrimination against females. However, when the data set was stratified into individual subjects the trend reversed, with females consistently admitted at a greater rate than males, suggesting discrimination in favour of females. The reversal of conditional probability when subgroups are amalgamated is known as Simpson's paradox. Hypothetical data that demonstrates Simpson's paradox is given in Table 2. Treatment A outperforms Treatment B when the male and female subgroups are combined, while Treatment B outperforms Treatment A for the male and female subgroups when treated separately.

Unfortunately, whether one should use the overall or stratified averages depends on causal relationships that are impossible to determine from observational data alone and therefore outside the abilities of most statistical learning methods [70]. Given the large data sets and the complex models involved in contemporary statistical learning, decision makers are unlikely to recognize Simpson's paradox in deployed systems, and therefore may inadvertently make decisions that conflict with their intentions.

6.3.2 Reverse regression

The correlation between two variables can change sign or magnitude when the dependent and independent variables are swapped, the so-called reverse regression paradox [71]. A famous example is the gender wage gap in labour markets. Women are found to earn lower wages than men when gender is treated as a dependent variable and education level is held constant. Curiously, the apparent correlation is flipped when wages are held constant and qualification levels are the dependent variables: Women seem to require lower qualifications to earn the same wages as men. The reverse regression paradox is troubling because a single data set can apparently lead to two incompatible inferences.

6.3.3 Berkson's paradox

Berkson's paradox occurs when a correlation between two variables is generated by a joint selection effect [70]. For instance, consider a university that accepts only students with high undergraduate marks or musical talent. Even if undergraduate marks and musical talent are initially uncorrelated, the joint selection effect will remove any students with low marks and no musical talent, leaving a negative correlation between undergraduate marks and musical talent in the university student population. Berkson's paradox demonstrates the difficultly of performing robust and objective inference with blackbox machine learning methods.

6.4 Addressing non-predictive performance

Defences against adversarial examples are actively being developed. While several defensive measures have been proposed such as distillation [72], obfuscated gradients [73] and model ensembles [74], a complete solution remains elusive and a risk-based approach to using statistical models in adversarial environments is recommended.

Improving *model interpretability* could help differentiate between brittle and robust models, and assist in identifying model biases. Perfect interpretability is impractical for complex statistical algorithms. A spectrum of methods exist for improving model interpretability, such as choosing an interpretable model for the supervised learning problem [4], building a surrogate model, measuring feature importance through model-agnostic diagnostics (for example, partial dependence plots [4]), or generating prototypes and counterfactuals for a subset of data points [75, 76]. The best approach will depend on the end users trade-off between model accuracy, the kinds of information they want, and the effort involved in interpreting the model.

The role of causality can be addressed by adopting greater use of *randomized experiments* that allows causal and confounding factors to be differentiated. There is also scope to develop and adopt *causal models*. Causal calculus extends Bayesian networks by allowing them to be used for counterfactual predictions [70]. It can even be used in the absence of randomized experiments if background knowledge is able to constrain the structure of the Bayesian model. Causal calculus is useful because it allows counterfactual predictions to be made and may improve the capacity for statistical models to generalize to new situations [70].

7 Where is cross-validation appropriate?

Cross-validation is a powerful tool for performing data-driven analysis. It is widely used because it makes relatively few assumptions, provides a clear measure of model performance, and is simple to implement. However, we argue that cross-validation is insufficient as a universal method for model validation, and we believe that traditional modelling practices are still relevant.

Cross-validation works well for structured problems with objective performance metrics, like chess or go. For open-ended problems without an agreed ground truth or a non-stationary environment an empirical approach is less useful. For example, missing data and subjective performance metrics cannot be addressed with cross-validation, since they are external to the observed data. In these situations, cross-validation often needs to be coupled with careful sensitivity analysis or caveating.

We believe the following practices would improve the robustness and credibility of deployed statistical models:

- Validate the data source and be aware of the context for data collection and model deployment.
- Use experimental design to improve the data quality and completeness.
- Restrict application of the model to well-understood domains or situations for which it has been validated.
- Monitor model performance after deployment.
- Use extended peer review to challenge implicit assumptions and validate measures of performance.
- Develop checklists and pre-plan the analysis to mitigate the impact of cognitive or social biases.

8 Acknowledgements

We would like to thank Maria Athanassenas, Tristan Cooper, Andrew Gill, Michael Papasimeon, Darryn Reid and Matthew Spillane for assistance and helpful discussions.

References

1. Benjamin S. Blanchard, Wolter J. Fabrycky, and Walter J. Fabrycky. *Systems engineering and analysis*, volume 4. Prentice Hall Englewood Cliffs, NJ, 1990.

2. Roger S. Pressman. *Software engineering: a practitioner's approach*. Palgrave Macmillan, 2005.

3. Averill M. Law, W. David Kelton, and W. David Kelton. *Simulation modeling and analysis*, volume 3. McGraw-Hill New York, 2007.

4. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

5. Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004.

6. Chris Andersen. The end of theory: The data deluge makes the scientific method obsolete, 2008. https://www.wired.com/2008/06/pb-theory/, accessed 2018-06-18.

7. David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

8. David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

9. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.

10. Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology*, 9(10):e1003256, 2013.

11. Declan Butler. When google got flu wrong. *Nature*, 494(7436):155, 2013.

12. Andrew Gelman, John B. Carlin, Hal S Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*. CRC press, 2013.

13. Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, 2017.

14. Quentin Frederik Gronau and Eric-Jan Wagenmakers. Limitations of bayesian leave-one-out cross-validation for model selection. 2018.

15. Jun Shao. Linear model selection by cross-validation. *Journal of the American statistical Association*, 88(422):486–494, 1993.

16. Yuling Yao, Aki Vehtari, Daniel Simpson, Andrew Gelman, et al. Using stacking to average bayesian predictive distributions. *Bayesian Analysis*, 2018.

17. Charles Ray Lake and Nathaniel Hurwitz. Schizoaffective disorder merges schizophrenia and bipolar disorders as one disease–there is no schizoaffective disorder. *Current opinion in Psychiatry*, 20(4):365–379, 2007.

18. Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

19. Arianne E. Miller. Searching for gaydar: Blind spots in the study of sexual orientation perception. *Psychology & Sexuality*, pages 1–16, 2018.

20. Andrew Gelman, Greggor Mattson, and Daniel Simspon. Gaydar and the fallacy of decontextualized measurement. *Sociological Science*, 5:270–280, 2018.

21. Kim H. Esbensen and Paul Geladi. Principles of proper validation: use and abuse of re-sampling for validation. *Journal of Chemometrics*, 24(3-4):168–187, 2010.

22. Phillip I. Good and James W. Hardin. *Common errors in statistics (and how to avoid them)*. John Wiley & Sons, 2012.

23. Phillipa J. Easterbrook, Ramana Gopalan, J. A. Berlin, and David R. Matthews. Publication bias in clinical research. *The Lancet*, 337(8746):867–872, 1991.

24. John P. A. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

25. Andrew Gelman and Eric Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013.

26. Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

27. Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(Jul):2079–2107, 2010.

28. Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

29. David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017.

30. Wei Wang and Andrew Gelman. Difficulty of selecting among multilevel models using predictive accuracy. *Statistics at its Interface*, 7(1):1–88, 2014.

31. Christoph Bergmeir, Rob J Hyndman, Bonsoo Koo, et al. A note on the validity of cross-validation for evaluating time series prediction. *Monash University Department of Econometrics and Business Statistics Working Paper*, 10:15, 2015.

32. Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.

33. Chris Chatfield. *The analysis of time series: an introduction*. CRC press, 2016.

34. Rob J. Hyndman. Cross-validation for time series, 2016. https://robjhyndman.com/hyndsight/tscv/, accessed 2018-06-18.

35. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

36. Jan Verbesselt, Rob Hyndman, Achim Zeileis, and Darius Culvenor. Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. *Remote Sensing of Environment*, 114(12):2970–2980, 2010.

37. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

38. Ken Kansky, Tom Silver, David A Mély, Mohamed Eldawy, Miguel Lázaro-Gredilla, Xinghua Lou, Nimrod Dorfman, Szymon Sidor, Scott Phoenix, and Dileep George. Schema networks: Zero-shot transfer with a generative causal model of intuitive physics. *Proceedings of the 34th International Conference on Machine Learning*, pages 70:1809–1818.

39. Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

40. Conor Dougherty. Google photos mistakenly labels black people gorillas, 2015. https://bits.blogs.nytimes.com/2015/07/01/google-photos-mistakenly-labels-black-people-gorillas, accessed 2018-06-18.

41. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Theres software used across the country to predict future criminals. and its biased against blacks. *ProPublica, May*, 23, 2016.

42. Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, 2018.

43. Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079, 2017.

44. Katherine A. Liu and Natalie A. Dipietro Mager. Womens involvement in clinical trials: historical perspective and future implications. *Pharmacy Practice (Granada)*, 14(1):0–0, 2016.

45. Premnath Shenoy and Anand Harugeri. Elderly patients participation in clinical trials. *Perspectives in clinical research*, 6(4):184, 2015.

46. Kristine E. Shields and Anne Drapkin Lyerly. Exclusion of pregnant women from industry-sponsored clinical trials. *Obstetrics & Gynecology*, 122(5):1077–1081, 2013.

47. Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The extent and consequences of p-hacking in science. *PLoS biology*, 13(3):e1002106, 2015.

48. Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

49. Daniel S. Quintana. From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in psychology*, 6:1549, 2015.

50. Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

51. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.

52. Stuart J. Russell and Peter Norvig. Artificial intelligence: a modern approach (international edition). 2002.

53. James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

54. Tushar Goel, Rajkumar Vaidyanathan, Raphael T. Haftka, Wei Shyy, Nestor V. Queipo, and Kevin Tucker. Response surface approximation of pareto optimal front in multi-objective optimization. *Computer methods in applied mechanics and engineering*, 196(4-6):879–893, 2007.

55. Silvio Funtowicz and Jerome Ravetz. Post-normal science. *International Society for Ecological Economics (ed.), Online Encyclopedia of Ecological Economics at http://www. ecoeco. org/publica/encyc. htm*, 2003.

56. Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1:0021, 2017.

57. A solar checklist. *Nature Photonics*, 9:803, 2015.

58. Ronald L. Wasserstein, Nicole A. Lazar, et al. The asas statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

59. Blakeley B. McShane, David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. Abandon statistical significance. *arXiv preprint arXiv:1709.07588*, 2017.

60. Raymond Hubbard and R. Murray Lindsay. Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18(1):69–88, 2008.

61. David R. Anderson, Kenneth P. Burnham, and William L. Thompson. Null hypothesis testing: problems, prevalence, and an alternative. *The journal of wildlife management*, pages 912–923, 2000.

62. Monya Baker et al. Statisticians issue warning on p values. *Nature*, 531(7593):151, 2016.

63. David R. Cox. [statistical modeling: The two cultures]: Comment. *Statistical Science*, 16(3):216–218, 2001.

64. Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on deep learning models. *arXiv preprint arXiv:1707.08945*, 1, 2017.

65. Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

66. Gwern. The neural net tank urban legend, 2018. https://www.gwern.net/Tanks, accessed 2018-06-18.

67. Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. *arXiv preprint arXiv:1611.04135*, 2016.

68. Carl Bergstrom and Jevin West. Criminal machine learning, 2017. http://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html, accessed 2018-06-18.

69. John Rice. *Mathematical statistics and data analysis*. Nelson Education, 2006.

70. Judea Pearl. *Causality*. Cambridge university press, 2009.

71. Jeff Racine and Paul Rilstone. The reverse regression problem: statistical paradox or artefact of misspecification? *Canadian Journal of Economics*, pages 502–531, 1995.

72. Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

73. Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *PMLR*, pages 80:274–283, 2018.

74. Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defense: Ensembles of weak defenses are not strong. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, Vancouver, BC, 2017. USENIX Association.

75. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *ICML Workshop on Human Interpretability in Machine Learning*, 2016.

76. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

DEFENCE SCIENCE AND TECHNOLOGY GROUP DOCUMENT CONTROL DATA			1. DLM/CAVEAT (OF DOCUMENT)
2. TITLE Cross-validation is insufficient for model validation		3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED REPORTS THAT ARE LIMITED RELEASE USE (L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)	
4. AUTHOR Thomas L. Keevers		5. CORPORATE AUTHOR Defence Science and Technology Group 506 Lorimer St, Fishermans Bend, Victoria 3207, Australia	
6a. DST Group NUMBER DST-Group–TR–3576	6b. AR NUMBER	6c. TYPE OF REPORT Technical Report	7. DOCUMENT DATE March 2019
8. Objective ID	9. TASK NUMBER	10. TASK SPONSOR	
13. DST Group Publications Repository		14. RELEASE AUTHORITY Chief, Joint and Operations Analysis Division	

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for Public Release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SOUTH AUSTRALIA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS

No Limitations

18. RESEARCH LIBRARY THESAURUS

Machine learning, artificial intelligence, verification and validation

19. ABSTRACT

Cross-validation is the de facto standard for model validation in the machine learning community. It reduces the risk of overfitting and provides an unbiased estimate of the learning algorithm predictive performance. Some people have argued cross-validation, coupled with sophisticated statistical learning methods, has rendered traditional scientific practices irrelevant. In this report, we review the foundations of cross-validation and draw attention to common, but underappreciated, assumptions. We argue that cross-validation is unsuitable for dealing with realistic complications like missing data, theory-laden observations, and malicious input. As a solution, we advocate for a holistic approach to model validation that embraces validation of data quality, acknowledgement of the role of subjective judgement in model assessment, and the use of extended peer review.