Animal Behaviour 124 (2017) 203–220

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Review

Applications of machine learning in animal behaviour studies

John Joseph Valletta [a, *], Colin Torney [a], Michael Kings [b], Alex Thornton [b], Joah Madden [c]

[a] Centre for Mathematics and the Environment, University of Exeter, Penryn Campus, Penryn, U.K.
[b] Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, U.K.
[c] Centre for Research in Animal Behaviour, University of Exeter, Exeter, U.K.

ARTICLE INFO

Article history:
Received 15 August 2016
Initial acceptance 8 September 2016
Final acceptance 15 November 2016
Available online 23 January 2017
MS. number: 16-00718

Keywords:
animal behaviour data
classification
clustering
dimensionality reduction
machine learning
predictive modelling
random forests
social networks
supervised learning
unsupervised learning

In many areas of animal behaviour research, improvements in our ability to collect large and detailed data sets are outstripping our ability to analyse them. These diverse, complex and often high-dimensional data sets exhibit nonlinear dependencies and unknown interactions across multiple variables, and may fail to conform to the assumptions of many classical statistical methods. The field of machine learning provides methodologies that are ideally suited to the task of extracting knowledge from these data. In this review, we aim to introduce animal behaviourists unfamiliar with machine learning (ML) to the promise of these techniques for the analysis of complex behavioural data. We start by describing the rationale behind ML and review a number of animal behaviour studies where ML has been successfully deployed. The ML framework is then introduced by presenting several unsupervised and supervised learning methods. Following this overview, we illustrate key ML approaches by developing data analytical pipelines for three different case studies that exemplify the types of behavioural and ecological questions ML can address. The first uses a large number of spectral and morphological characteristics that describe the appearance of pheasant, *Phasianus colchicus*, eggs to assign them to putative clutches. The second takes a continuous data stream of feeder visits from PIT (passive integrated transponder)-tagged jackdaws, *Corvus monedula*, and extracts foraging events from it, which permits the construction of social networks. Our final example uses aerial images to train a classifier that detects the presence of wildebeest, *Connochaetes taurinus*, to count individuals in a population. With the advent of cheaper sensing and tracking technologies an unprecedented amount of data on animal behaviour is becoming available. We believe that ML will play a central role in translating these data into scientific knowledge and become a useful addition to the animal behaviourist's analytical toolkit.

Recent technological advances mean that large data sets can be collected on the movement (Hussey et al., 2015; Kays, Crofoot, Jetz, & Wikelski, 2015; Tomkiewicz, Fuller, Kie, & Bates, 2010), fine-scale motion (Brown, Kays, Wikelski, Wilson, & Klimley, 2013), social interactions (Krause et al., 2013), vocalizations (Blumstein et al., 2011) and physiological responses (Kramer & Kinter, 2003) of individual animals. Conversely, the logistical difficulties of collecting replicated data, especially from wild populations, mean that sample sizes are small, even though data on each individual may be rich, with many hundreds (or even thousands) of factors to consider. These complex data sets, generated from different sources, such as images and audio recordings, may fail to conform to assumptions of many classical statistical models (e.g. homoscedasticity and a Gaussian error structure). Moreover, unknown nonlinear dependencies and interactions across multiple variables make it unclear what type of functional relationship one should use to describe such data mathematically. Animal behaviour researchers are thus in a position where automatically collecting detailed data sets is becoming commonplace, but extracting knowledge from them is a daunting task, mainly due to the lack of accessible analytical tools.

Machine learning (ML) offers complementary data modelling techniques to those in classical statistics. In animal behaviour, ML approaches can address otherwise intractable tasks, such as classifying species, individuals, vocalizations or behaviours within complex data sets. This allows us to answer important questions across a range of topics, including movement ecology, social structure, collective behaviour, communication and welfare. ML encompasses a suite of methodologies that learn patterns in the data amenable for prediction. A machine (an algorithm/model) improves its performance (predictive accuracy) in achieving a task

* Correspondence: J. J. Valletta, Centre for Mathematics and the Environment, University of Exeter, Penryn Campus, Penryn TR10 9FE, U.K.
E-mail address: jj.valletta@exeter.ac.uk (J. J. Valletta).

http://dx.doi.org/10.1016/j.anbehav.2016.12.005

(e.g. classifying content of an image) from experience (data). The objective is for the predictive model to generalize well, that is, to make accurate predictions on previously unseen data. For instance, when Facebook users upload their photos, the 'auto-tagging' ML algorithm extracts facial features and suggests names of friends in that photo. Facebook's predictive model generalizes from manually tagged photos (known as the training data set). It is impossible to 'show' a machine all the images of an individual (e.g. different facial expressions); instead the model uses the extracted features to learn patterns that best discriminate one individual from another. The generalization error or predictive performance is a measure of how many previously unseen images (known as the testing data set) the algorithm tags correctly.

Both statistical modelling and ML seek to build a mathematical description, a model, of the data and the underlying mechanism it represents; thus inevitably there is substantial overlap between the two (Breiman, 2001b; Friedman, 2001; Zoubin Ghahramani, 2015). However, historically they differ in their rationale as follows. Statistical models start with an assumption about the underlying data distribution (e.g. Gaussian, Poisson). The focus is on inference; estimating the parameters of the statistical model that most likely gave rise to the observed data, and providing uncertainty bounds for these estimates. For ML, the focus is typically on prediction; without necessarily assuming a functional distribution for the data, a model that achieves optimal predictive performance is identified. It is this hypothesis-free approach that makes ML an attractive choice for dealing with complex data sets. While in traditional statistical modelling a hypothesis (model) is put forward and is then accepted/rejected depending on how consistent it is with the measured observations, ML methods learn this hypothesis directly from the training data set.

ML can tackle a wide range of tasks, including classifying observations into predefined sets (Kabra, Robie, Rivera-Alba, Branson, & Branson, 2013), clustering data into groups that share an underlying process (Zhang, O'Reilly, Perry, Taylor, & Dennis, 2015) and regressing an outcome of interest against multiple factors and elucidating their contributory effect (Chesler, Wilson, Lariviere, Rodriguez-Zas, & Mogil, 2002; Piles et al., 2013). Owing to its versatility, ML has been applied across a broad set of domains in animal behaviour to ask and subsequently answer biologically meaningful questions. Next, we highlight some facets of animal behaviour where ML has already been deployed.

GPS, accelerometer and/or video data are routinely used to monitor movement patterns of individuals. Three-dimensional accelerometer loggers can generate over a million data points per hour of recording (at a sample rate of 100 Hz). ML is used to automate the classification of behaviours/activities (Kabra et al., 2013) and tracking movement trajectories (Dell et al., 2014). This knowledge can then be used to infer individual decision rules in collective motion (Katz, Tunstrom, Ioannou, Huepe, & Couzin, 2011; Nagy, Kos, Biro, & Vicsek, 2010) and to compute activity budgets for individuals without the need for continuous human observation or time-consuming video analysis. This is especially suitable for organisms that are hard to observe directly, such as nocturnal (badgers, *Meles meles*: McClune et al., 2014), pelagic (little penguins, *Eudyptula minor*: Carroll, Slip, Jonsen, & Harcourt, 2014) and aquatic species (great sculpins, *Myoxocephalus polyacanthocephallus*: Broell et al., 2013; whale sharks, *Rhincodon typus*: Gleiss, Wright, Liebsch, Wilson, & Norman, 2013), or those that are hard to follow continuously owing to their speed or covertness (e.g. cheetahs, *Acinonyx jubatus*: Grünewälder et al., 2012; pumas, *Puma concolor*: Wang et al., 2015).

Another context in which ML has been successfully employed is in vocalization studies. Vocalizations can be recorded remotely permitting assessments of population size and species composition, individual behavioural and inter/intraspecific interactions (Blumstein et al., 2011). A typical recording, made using pulse code modulation (PCM) at 24-bit and 48 Hz sampling, produces over half a gigabyte of data per hour. Consequently, inspection of these data and analysis of sound recordings can be time consuming and highly subjective when conducted by visual inspection of sonograms. Instead, ML has been applied to classify and count particular elements or syllables (Acevedo, Corrada-Bravo, Corrada-Bravo, Villanueva-Rivera, & Aide, 2009). Early work used ML techniques to adjudicate similarity between calls based on sets of such elements (Tchernichovski, Nottebohm, Ho, Pesaran, & Mitra, 2000). These approaches can also discern differences in calls. Classification of calls from different species and subspecies is robust (Fagerlund, 2007; Kershenbaum et al., 2016) and permits assessment of community structure (e.g. frogs: Taylor, Watson, Grigg, & McCallum, 1996; birds: Brandes, 2008). Finer scale discriminations are possible at both the individual level (Cheng, Xie, Lin, & Ji, 2012) and the bird song elements level (Ranjard & Ross, 2008).

The assessment of animal welfare and the emotional states that may reveal it can be highly subjective, and poor welfare is often only indicated by multiple interacting factors (Broom & Johnson, 1993). ML can assist in monitoring such behaviours by matching the human assessment in terms of treatment effects on laboratory mice, *Mus musculus* (Roughan, Wright-Williams, & Flecknell, 2009). Such methods have been extended to provide a diagnostic tool for psychopharmacological drugs based on mouse open-field behaviour (Kafkafi, Yekutieli, & Elmer, 2009). ML was used in a comparative assessment of welfare across multiple laboratory populations of mice (Chesler et al., 2002) permitting a wide range of potential explanatory factors, each with diverse distribution, to be considered simultaneously as well as the interactions between them. A potential novel use of ML would be to detect emotional state in animals based on facial expression, body posture or vocalization. Such techniques have already been used in humans looking at facial (Michel & El Kaliouby, 2003), physiological (Shi et al., 2010), vocal (Shami & Verhelst, 2007) and gestural (Castellano, Villalba, & Camurri, 2007) cues of emotions. ML also permits integration of multiple sets of these cues to further enhance emotion detection (Caridakis et al., 2007).

Elucidating the underlying social network structure of individuals within social groups can help address important ecological and evolutionary questions (Krause, James, Franks, & Croft, 2015). Passive integrated transponder (PIT) tags and proximity loggers now permit automated collection of large volumes of social interaction data containing both spatial and temporal elements (Krause, Wilson, & Croft, 2011). Translating such data into biologically realistic patterns of association is not trivial, and may depend on subjective decisions by researchers, especially when the instances of association are ambiguous. Co-occurrences in time could be determined by ML clustering methods with individuals in the same foraging event (cluster) considered to have a social affiliation (Psorakis et al., 2015). Such methods appear to be robust and capture real-life pair bonds well (Psorakis, Roberts, Rezek, & Sheldon, 2012). A second facet of association patterns that benefits from application of ML techniques is determining to which social grouping an individual belongs within a network. In many cases, group membership is ambiguous with individuals having weak or sporadic membership to multiple clusters of other individuals. A subjective decision of membership could be arrived at, with such weak affiliations being discounted. Alternatively, ML techniques could be deployed to account for such 'fuzzy overlapping' (Gregory, 2011), and individuals can have their relative membership of each group determined.

It is clear that ML can address different objectives in numerous distinct fields of animal behaviour and is thus becoming a staple

approach for novel methods (Dankert, Wang, Hoopfer, Anderson, & Perona, 2009; Kabra et al., 2013). The aim of this review is to demystify ML, which is typically documented in technical journals and books (Bishop, 2006; Hastie, Tibshirani, & Friedman, 2009; Murphy, 2012; Witten, Frank, & Hall, 2011). We present a concise guide on the rationale behind unsupervised and supervised learning, and illustrate these methods by developing data analytical workflows to convert three data sets into useful biological knowledge: assigning pheasant, *Phasianus colchicus*, eggs to clutches based on their visual appearance, to subsequently study the response of brooding females to eggs that are not their own; constructing social networks based on co-occurrences of jackdaws, *Corvus monedula*, at feeding stations, to examine population level processes such as social learning; and automating the counting of individual wildebeest, *Connochaetes taurinus*, within aerial survey photos, to guide conservation policies.

UNSUPERVISED LEARNING

Unsupervised learning methods uncover structure in unlabelled data. Structure means patterns in the data that are sufficiently different from pure unstructured noise. For example, the data may be temporally or spatially correlated, or organized in a hierarchical fashion (e.g. transcription factors that regulate gene expression). Structure can be discovered by visualizing the data after reducing their dimensionality (dimensionality reduction), identifying groups of observations sharing similar attributes (clustering) and/or determining the distribution of the data (density estimation). Here, we introduce dimensionality reduction and clustering, and touch briefly on density estimation when presenting Gaussian mixture models (GMM).

Dimensionality Reduction

Reducing the number of attributes per observation can provide several benefits: (1) biologically, to elucidate the best predictors (plausible causal drivers under an experimental setup) of the underlying process (for example, which of 21 different seminal variables predicted fertilization success in rabbits: Piles et al., 2013); (2) visualization, to highlight the data's structure; (3) prediction, to improve the model's accuracy by removing uninformative features; and (4) computationally, to enable faster implementation. The rationale behind dimensionality reduction is simple; although the collected data may seem high dimensional, the structure of the data can be represented by fewer variables. This can be due to redundant features arising from multicollinearity and/or noisy attributes that offer little discriminatory power. Reducing the dimensionality of a problem is achieved by mapping the original data to a new feature set, feature extraction (Burges, 2009; van Der Maaten, Postma, & van Den Herik, 2009; Lee & Verleysen, 2007), and/or selecting a subset of attributes, feature selection (Liu & Motoda, 2008). Note that in the ML literature the term dimensionality reduction can be used to refer solely to (typically) unsupervised methods that transform high-dimensional data to a lower dimensional feature set, while feature selection is treated separately and as part of the predictive modelling framework. Following this notion, we describe feature extraction next, while feature selection is presented later.

Feature extraction

Analogous to representing complex entities such as biological diversity by using diversity indices, feature extraction deals with finding representations for high-dimensional data sets. For example, should we describe an image by individual pixel intensities or by extracting higher-order structures such as edges and shapes? The objective is to construct new features from the original measured variables that accentuate the inherent patterns in the data and are nonredundant. Feature extraction is a key step in ML; finding representations that are directly relevant to the task at hand (e.g. discriminating between two classes) will almost always result in better predictive accuracy than employing more complex models.

Dimensionality reduction techniques aggregate dimensions together while trying to preserve as much of the data's structure as possible. That is, observations that are 'close' to each other remain so in the lower-dimensional data set. Principal component analysis (PCA) is a linear dimensionality reduction method with widespread use in the biosciences, where the new uncorrelated features are weighted linear combinations of the original data. As explained by Ringnér (2008), the objective of PCA is to find directions (called principal components) that maximize the variance of the data. For visualization the first two or three components are used to plot the data in an attempt to reveal any groupings.

The alternative is hand-crafted features (also known as feature engineering) which rely on expert knowledge to derive a set of discriminatory features. For example, in automated animal vocalization classification from audio recordings, the acoustic structure and syllable arrangement of each vocalization represent the 'raw' data. The extracted variables may include distinct acoustic measurements such as minimum and maximum frequency, but may also incorporate features of the pattern of the call as a whole, such as repetition, element diversity, combinations, ordering or timing. Traditionally, these measures have been used to classify or cluster calls, but such an approach can be extended to include broader sets of features, for instance, sequences of sounds (motifs) (Kershenbaum et al., 2014).

Clustering

The goal of clustering is to find groups that share similar properties. The data in each group should be similar (minimize intracluster distance), but each cluster should be sufficiently different (maximize intercluster similarity; see Appendix Fig. A1). There are three major types of clustering methods: (1) partitioning, where the feature space is divided into k regions; (2) hierarchical, where small clusters are iteratively merged into larger ones, or vice versa; and (3) model-based, where multivariate statistical distributions are fitted. For distance-based methods similarity between observations (distance metric; e.g. Pearson correlation and Euclidean distance) and similarity between clusters (linkage functions; e.g. complete or average) need to be defined (D'haeseleer, 2005). In this section, we present three popular algorithms of each type, followed by a brief discussion on the problem of estimating the correct number of clusters in the data.

k-means

Arguably the most widely used partitioning clustering method (e.g. to group behavioural modes in little penguins; Zhang et al., 2015). The feature space is divided into k regions as follows (see Appendix Fig. A2). (1) Choose k centroids (at random or using some prior knowledge). (2) Compute the distance between centroids and each data point. (3) Assign each data point to the closest centroid. (4) Compute new centroids; the average of all data points in that cluster. (5) Repeat steps 2 to 4 until data points remain in the same cluster or a maximum number of iterations is reached.

k-means clustering is popular because it is intuitive and computationally inexpensive. However, it is only applicable to continuous data where a mean is defined. There is also no guarantee of a global optimum solution; thus, the user is advised to start the algorithm at multiple distinct centroids.

Agglomerative hierarchical clustering

In agglomerative hierarchical clustering small clusters are iteratively merged into larger ones (e.g. detecting subgroups of people moving together within videos of human crowds: Ge, Collins, & Ruback, 2012). The clustering strategy is as follows. (1) Assign each datum as its own cluster. (2) Compute the distance between each cluster. (3) Merge the closest pair into a single cluster. (4) Repeat steps 2 to 3 until all clusters are merged together.

Although in hierarchical clustering there is no need to specify k a priori, this is implicitly done through a hard threshold, which defines the number of distinct clusters. Care must be taken as the choice of distance metric and linkage function can significantly change the outcome of the results. Also, for large number of observations, agglomerative hierarchical clustering can be computationally expensive.

Gaussian mixture model (GMM)

The Gaussian mixture model (GMM) is a simple but powerful model that performs clustering via density estimation (e.g. to infer social networks from PIT-tagged birds; Psorakis et al., 2012). The data's histogram is modelled as the sum of multiple Gaussian distributions. In general k multivariate Gaussian distributions are fitted using the expectation-maximization (EM) algorithm, to estimate the mean vector (μ), covariance matrix (Σ) and mixing coefficients (π) for every cluster (distribution), represented as follows:

$$p(x) = \sum_{i=1}^{k} \pi_i N(x|\mu_i, \Sigma_i)$$

$$\sum_{i=1}^{k} \pi_i = 1$$

GMMs can be viewed as a 'soft' version of k-means because every data point is part of every cluster but with varying levels of membership. GMMs, however, assume that data are generated from a mixture of multivariate Gaussians and lack a global optimum solution.

How many clusters?

Determining the number of distinct clusters k in a data set is a fundamental yet unsolved problem. The issue lies in the mathematical subjectivity of similarity. Moreover, because the data are unlabelled, the correct number for k is inherently ambiguous. In the section Case Studies, we present two approaches to estimating k, using silhouette width (Rousseeuw, 1987) and the Bayesian information criterion (BIC). There are several other cluster validity metrics that can be used to derive a ballpark range for the true underlying number of clusters (Charrad, Ghazzali, Boiteau, & Niknafs, 2014). However, the final decision rests with the practitioner, who needs to ensure k is biologically relevant within the context of the problem at hand.

SUPERVISED LEARNING

Like traditional statistical models (e.g. generalized linear models), supervised learning methods identify the relationship between an outcome and a set of explanatory variables. Using the data as a starting point, rather than a predefined model structure, the ML machinery learns the mapping (predictive model) between a set of features and a continuous outcome (regression) or a categorical variable (classification).

Fig. 1 shows a bird's eye view of supervised learning. Although the data in this case are labelled, unsupervised methods are extensively used in the initial stages of the analysis to explore the data (e.g. visualization of high-dimensional data sets) and extract putative discriminatory features (see section Feature Extraction). Once an appropriate feature set is determined, the observations are then split into training and testing data sets. The training data set is used to build the predictive model, while the testing data set (not used in model building) is used to compute the expected predictive performance 'in the field'. In statistics, this is akin to making inferences about the population based on a finite and random sample.

ML algorithms can deal with nonlinearities and interactions among variables because the models are flexible enough to fit the data (as opposed to rigid linear regression models, for example). However, this flexibility needs to be constrained to avoid fitting noise (overfitting). Hyperparameters, specific to the ML algorithm, are tuned by cross-validation to strike a balance between underfitting and overfitting, known as the bias–variance trade-off (see Fig. 2a). In k-fold cross-validation the training data are randomly split into k parts. The model is trained on all but one of the folds, and performance is measured on the part left out in the training process (see Fig. 2b). The average prediction error is computed from the k runs and the hyperparameters that minimize this error are used to build the final model (see Fig. 2c).

Next, we present two widely used supervised learning algorithms, decision trees and random forests. Like most supervised learning methods, they can be used to tackle both regression and classification problems; however, here we focus on the latter owing to their popularity in animal behaviour studies (e.g. classifying behavioural modes: Kabra et al., 2013; discriminating between different bird call types: Cheng et al., 2012).

Decision Trees

Decision trees are simple and intuitive predictive models, making them a popular choice when decision rules are required (Hutchinson & Gigerenzer, 2005). For example, the activity status of an animal can be classified by asking a series of simple yes/no questions (e.g. is the velocity >0.01 m/s^2? Nadimi, Søgaard, & Bak, 2008). A decision tree is constructed as follows. (1) Find the yes/no rule that best splits the data with respect to one of the features. (2) The best split is the one that produces the most homogeneous groups. (3) Repeat steps 1 and 2 until all data are correctly classified or some stopping rule is reached.

Decision trees are readily interpretable, directly used to generate rules, and computationally inexpensive to train, evaluate and store. They can handle both categorical and continuous data, and are fairly robust to outliers. However, they are prone to overfitting (small changes in the training data set may lead to significantly different trees) and their predictive performance can be poor compared to other algorithms.

Random Forests

Random forests (Breiman, 2001a; Cutler et al., 2007) is an ensemble method developed to mitigate the problem of overfitting in decision trees. Instead of a single tree, multiple trees are grown and averaged over as follows (each tree is known as a weak learner). (1) Grow T decorrelated trees. (2) Induce randomness by bagging (bootstrap aggregating), where each tree is trained on a subset of the data randomly sampled with replacement, and by considering only a subset of predictors as candidates for each split. (3) Average predictions from all T trees.

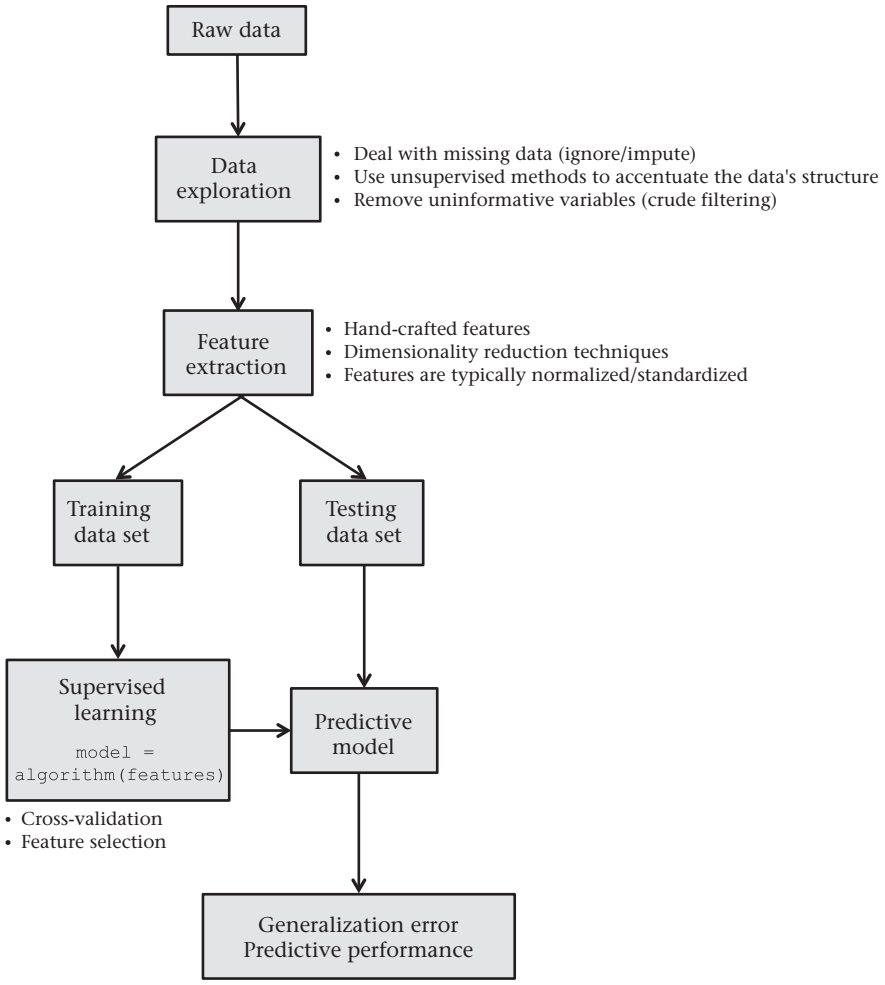

Figure 1. A typical supervised learning problem workflow. Exploratory data analysis is followed by a feature extraction step to derive putative discriminatory variables. The observations are split into training and testing data sets. The predictive model is trained on the training data set. Training also involves internal cross-validation, to set the model's hyperparameters, and may also include a feature selection step. Finally, the model's generalization performance is computed using the testing data set (not used to train the model).

Cross-validation is inherent in the random forests methodology as every tree is trained only on a subset of the original data. This allows the computation of an estimate for the generalization error by computing the predictive performance of the model on the data left out from the training process, known as the out-of-bag (OOB) error. The OOB data are also used to compute an estimate of the importance of every predictor, which can be subsequently used for feature selection. Random forests can handle thousands of mixed categorical and continuous predictors and are robust to outliers; however, the interpretability of plain decision trees is lost.

FEATURE SELECTION

Akin to the concept of parsimony in ecological modelling (Johnson & Omland, 2004), feature selection methods select a subset of explanatory variables that are best at prediction. These techniques come in three flavours: filter, wrapper and embedded methods (Guyon & Elisseeff, 2003).

Filter methods encompass a number of crude ways to reduce a large list of predictors prior to model fitting. The relevance of each variable is assessed using measures such as correlation with the outcome and statistical significance for differences across classes (Lazar et al., 2012). These univariate measures are intuitive and computationally inexpensive; however, they ignore high-order interactions, which may be present, and important, in complex systems. Thus, the threshold is conservative and only features that are unlikely to contribute to good predictions are removed.

Wrapper methods involve a greedy search for the best subset of features. This is usually achieved through forward, backward or stepwise selection, the status quo in ecological modelling, where metrics based on significance testing (the ubiquitous *P* values) or information criteria (AIC/BIC) dictate whether a variable stays in the model (Johnson & Omland, 2004). These metrics require an underlying statistical model and therefore are not suited to all ML algorithms. Instead, predictive performance measures, such as misclassification rates, together with variable importance measures, for example, those returned by random forests, are used (Huynh-Thu, Saeys, Wehenkel, & Geurts, 2012).

Embedded feature selection methods incorporate variable selection within the model training process. These methods are specific to the learning algorithm and thus cannot be used across the board; however, they are less computationally intensive than wrapper strategies. Examples of these techniques include LASSO (Tibshirani, 1996), where within a linear regression framework coefficients of poor predictors are 'shrunk' to zero.

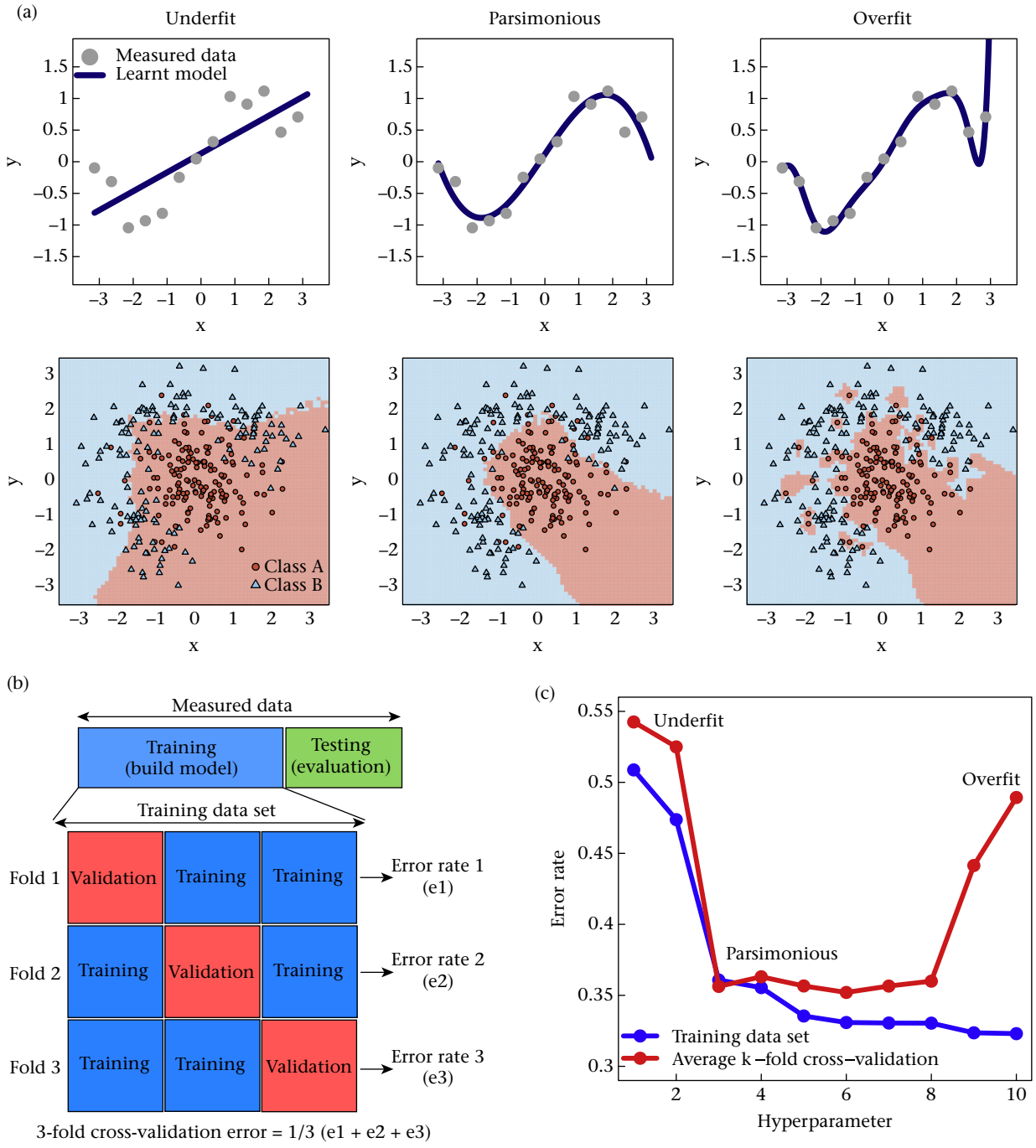

Figure 2. Cross-validation in supervised learning to set the model's hyperparameters. (a) The top row depicts a regression problem where the data points were simulated from a sinusoidal function, with zero mean Gaussian noise added to it. Varying the hyperparameter changes the complexity (flexibility) of the model, resulting in three distinct cases: underfit (model is too simple to describe the underlying process; it has high bias and low variance); parsimonious (the least complex model that describes the observed data well); and overfit (model is too complex and is fitting noise; it has high variance and low bias). The bottom row shows the same scenario but for a classification problem, where the data points were generated from a mixture of two-dimensional Gaussian distributions. The objective of cross-validation is to estimate the generalization error and subsequently choose hyperparameters that result in a parsimonious model. (b) Sketch of *k*-fold cross-validation ($k = 3$ in this depiction). The training data set is randomly split into *k* parts. The model is trained on all but one of the folds, and predictive performance measured on the part left out in the training process. The averaged cross-validation error gives an estimate of the generalization error. (c) Plot of hyperparameter versus training error (blue) and average *k*-fold cross-validation error (red). As model complexity increases, the training error decreases as the model fits noise. However, the cross-validation error will initially decrease but then starts to increase (an overfitted model does not perform well on data not used in the training process). The hyperparameters that minimize the cross-validation error are chosen to build the final predictive model.

Note that what is being described here as feature selection is commonly referred to as model selection in ecological models. Although selecting variables is explicitly model choice, feature and model selection tend to have different connotations in ML. While feature selection refers to which variables to include in a predictive model, model selection deals with tuning the model's hyperparameters using cross-validation (see Fig. 2).

CASE STUDIES

In this section, we develop ML workflows to analyse three distinct data sources and applications. Sections Pheasant Eggs and Jackdaw Associations are unsupervised problems, where in the former the number of clusters is known a priori, while section Wildebeest Identification represents a supervised case. Documented code to reproduce the ensuing analysis can be found at https://github.com/jjvalletta/ML_Animal_Behaviour.

Pheasant Eggs

Female pheasants may nest communally or lay eggs in multiple nests (Yom-Tov, 1980). Without direct observation of laying or destructive genetic sampling from eggs, it is not possible to assign eggs to females. Visual inspection of eggs suggested that within-individual variation in egg morphology may be low while inter-individual variation is high. We used external features such as size, shape and colour to confirm whether individual females laid distinct eggs and therefore whether those from mixed clutches could be reliably assigned to females.

Thirty pheasant females were randomly assigned to one of 10 breeding pens, each containing one male, such that each pen contained three females. Eggs were collected daily from each pen and marked with the egg ID code. A total of 549 eggs were collected. Egg length and width were measured using callipers (precision: 0.1 mm) and weighed (precision: 0.1 g). Reflectance spectra for each egg were collected in the 300–700 nm wavelength range with an Ocean Optics (Dunedin, FL, U.S.A.) S2000 portable spectrometer with specially made shielded radiance attachment and an Ocean Optics PX-1 synchronized Xe flash lamp. Seventeen spectral properties were extracted using PAVO (Maia, Eliason, Bitton, Doucet, & Shawkey, 2013), with relevant adjustments made for avian vision, based on the cone sensitivities of peafowl, *Pavo cristatus*. Birds were held in Devon, U.K. during the spring of 2015 with eggs being collected between 24 March and 26 April. Work was carried out under Home Office licence PPL 30/3204.

As eggs have no corresponding label (i.e. to identify the female that laid them), unsupervised learning methods were employed to uncover structure in the data (see Fig. 3a for the workflow diagram). A correlogram of all 21 morphological and spectral measures showed strong linear associations between some of the variables (see Fig. 3b). We applied PCA to extract informative features and to visualize the data. Fig. 3c depicts the PCA outcome, where the first four components explained 96% of the variance in the data and were kept as features for subsequent analysis.

PCA returns a weighted linear combination of the original attributes (see Appendix Fig. A3) that, in this case, can be given a biological interpretation because each variable falls under a particular theme. PC1 summarizes most of the spectral data, including the colorimetric variables describing the hue, chroma and brightness of the eggshell, and the perception of the eggshell by the pheasant, considering the stimulation of the u, s and l cones which had their sensitivities tuned using the peafowl visual model (Hart, 2002). PC2 is specifically related to how the eggshell stimulates the cone sensitive to medium wavelengths with a peak sensitivity of 537 nm, and includes both the absolute quantum catch by that cone and the relative cone stimulation, for a given hue, as a function of saturation (Stoddard & Prum, 2008). PC3 has high loadings for measures relating to egg size (length, width and mass). PC4 has a large weight for the relative measure of egg width/length. Thus, PC1 is a measure of eggshell brightness, PC2 is a measure of eggshell greenness, PC3 is a measure of egg size and PC4 is a measure of egg shape.

Fig. 3d shows results, projected on the first two principal components, from running *k*-means on one of the pens of the pheasant eggs data set. For comparison, we also employed agglomerative hierarchical clustering using Ward's minimum variance method as depicted in Fig. 3e. The outcome was similar to that from *k*-means, but using dendrograms, the results can be visualized irrespective of dimensionality (as opposed to *k*-means where results had to be projected onto the first two PCs), and also allows us to identify subgroups within each cluster.

In this case, we knew a priori that every pen contained three females; however, this is not always the case. Appendix Fig. A4 depicts silhouette plots (Rousseeuw, 1987) for varying numbers of clusters for the pheasant eggs data. The silhouette width is a normalized measure ($-1 \leq s \leq +1$), where s close to 1 suggests that the data point is adequately clustered, an s close to 0 is somewhat inconclusive (datum could be part of current cluster or neighbouring cluster), and an s close to -1 implies that the data point should be part of the neighbouring cluster. We can see that $k = 3$ maximizes the average silhouette width suggesting that, indeed, the clutches belong to three female pheasants.

Using unsupervised learning methods, we were able to reduce the dimensionality of the problem to a few biologically meaningful features, and subsequently assign eggs to females. The generalization performance of this model can be estimated by determining the maternity of a subsample of the eggs using protein fingerprinting (Andersson & Åhlund, 2001). Following successful validation, if clutches of eggs are encountered that are the product of multiple females, we can then reliably group sets of eggs from different females, and possibly explore spatial patterns of egg dumping and subsequent responses of brooding females to eggs that are not their own.

Jackdaw Associations

Uncovering social structures in jackdaws allows us to infer individualized social relationships and determine how novel information spreads through social groups. To build these social networks based on patterns of association, we need to quantify the temporal co-occurrence of birds (gambit of the group) during feeding events. It is unclear, however, how one specifies the length of these foraging events without fixing an arbitrary time window.

Around 1000 individual jackdaws across three colonies in Cornwall, U.K. were fitted with a leg ring containing a PIT tag. Data-logging bird feeders automatically record the time of every visit by PIT-tagged birds and the identity of the individual, providing a rich data stream of visits. For illustrative purposes, we focus on 4 weeks of data from 141 individuals visiting a single feeder located near Stithians, Cornwall, U.K.

Based on the work of Psorakis et al. (2012), Fig. 4a summarizes the workflow used to convert a long series of time stamps into a social network. There is no need for feature extraction as the time stamp is the only attribute of interest. As depicted in Fig. 4b, the data are dominated by short intervisit periods corresponding to foraging events, interleaved with long periods of inactivity. A Gaussian mixture model (GMM) was used to group the time stamps

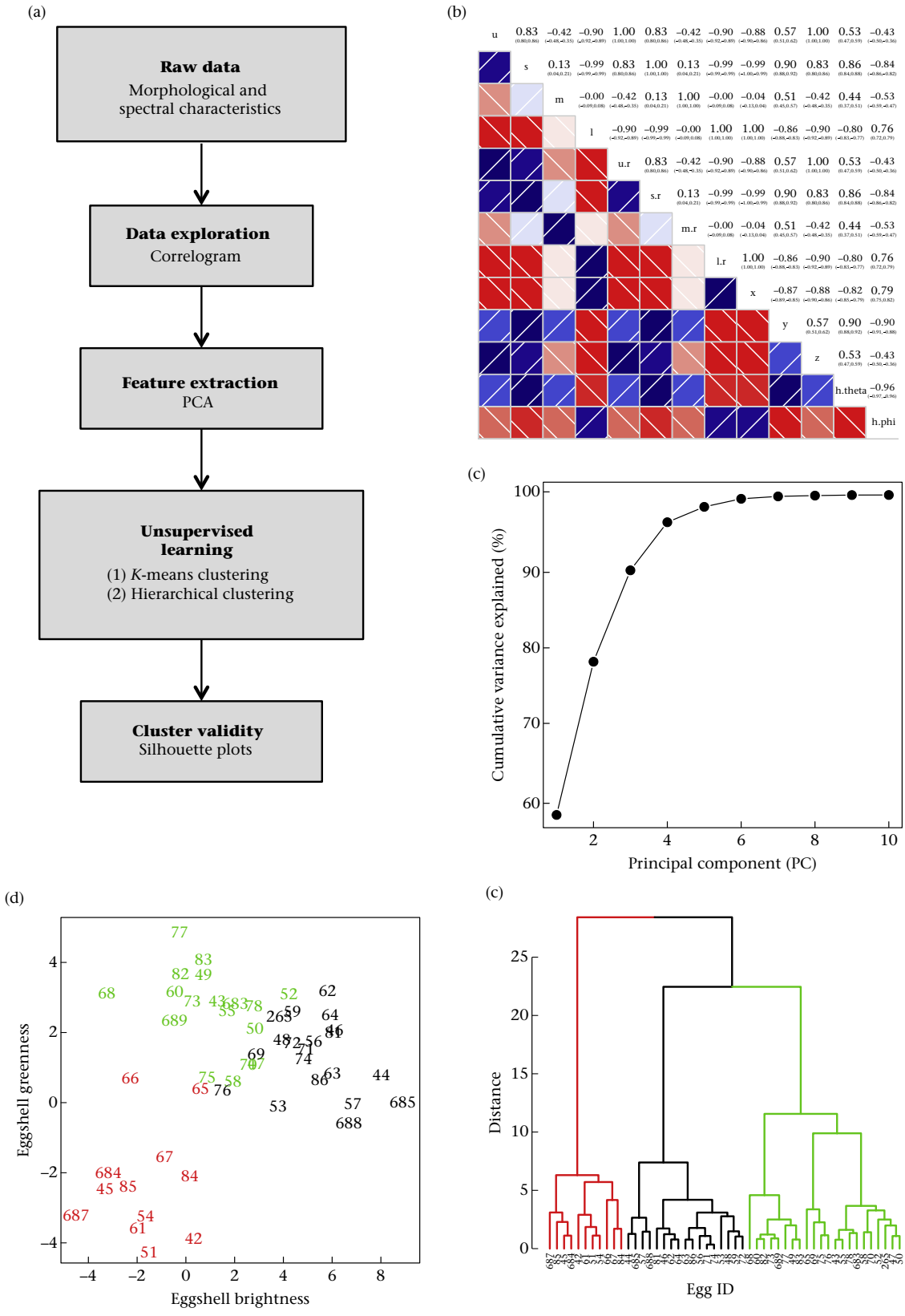

Figure 3. Pheasant eggs case study. (a) Analysis pipeline, (b) correlogram of some of the 21 morphological and spectral measures and (c) cumulative variance explained by principal components (PCs). The first four components explained 96% of the variance in the data and were kept as features for subsequent analysis. (d) *k*-means clustering results projected on the first two PCs. (e) Hierarchical clustering using Ward's minimum variance method. The different colours together with the egg ID represent the putative clutches.

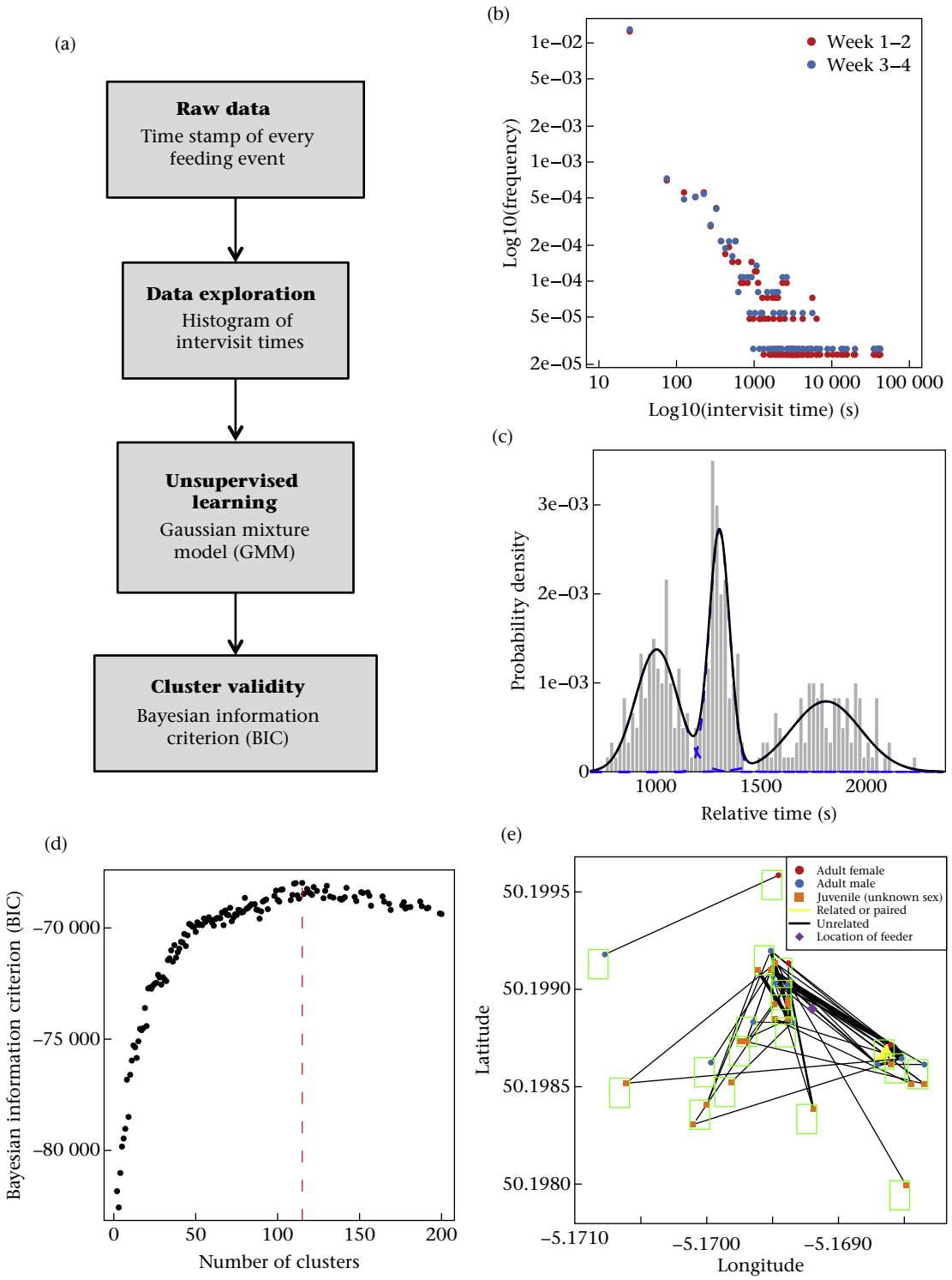

Figure 4. Jackdaw associations case study. (a) Analysis pipeline, (b) histogram of intervisit times, (c) sketch of a Gaussian mixture model (GMM), (d) Bayesian information criterion (BIC) as a function of the number of clusters to estimate the number of foraging events and (e) social network projected spatially, where each node is located at the coordinates of the nestbox (green rectangle) occupied by that bird. Edge width represents association strength and edge colour indicates association between individuals that occupied the same nestbox (yellow) or a different nestbox (black).

into distinct clusters and reveal which birds foraged together (see Fig. 4c).

In contrast with the pheasant eggs data set, the number of clusters is unknown. We estimated this by using the BIC, as in this case, choosing the number of clusters is equivalent to model choice (see Fig. 4d). Once an appropriate GMM was fitted, the rest of the process involved counting co-occurrences of pairwise birds from which an adjacency matrix was computed and subsequently the social network constructed (see Fig. 4e).

Jackdaw social structure, as for most corvids, is generally assumed to be characterized by monogamous breeding pairs and prolonged association between parents and offspring following

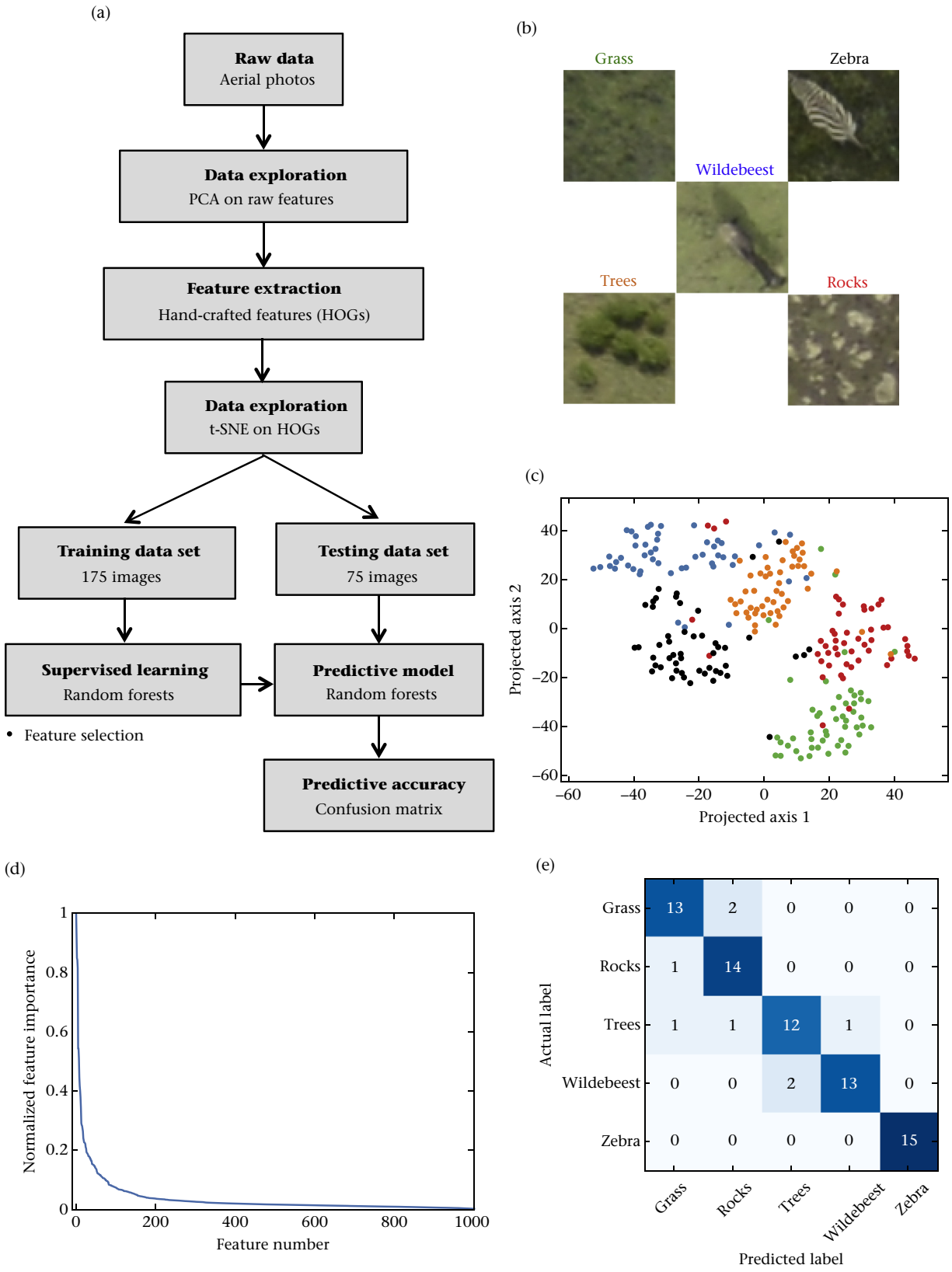

Figure 5. Wildebeest identification case study. (a) Analysis pipeline, (b) typical images for the five different classes (wildebeest (green), zebra (black), grass (green), trees (orange) and rocks (red)), (c) visualizing the extracted HOGs features using t-SNE (every observation is coloured by its class) and (d) normalized permutation importance score computed from the random forests for every feature. From an initial pool of over 1000 features only a couple of hundred predictors are deemed to be important. These features are selected to build the final model. (e) Predictive performance evaluated on the testing data set visualized as a confusion matrix.

fledging (Clayton & Emery, 2007). However, research has largely been confined to captive studies and there is limited information on patterns of association in natural populations. The jackdaw social network we generated (see Fig. 4e) confirms the existence of foraging associations between mated adults and between parents and offspring. It also reveals strong associations between unrelated individuals and shows how these associations relate to the spatial proximity between nestboxes. This characterization of jackdaw social structure can provide a useful tool for examining population level processes, such as epidemiology and social learning (Croft, Madden, Franks, & James, 2011; Franz & Nunn, 2009).

Wildebeest Identification

Counting the population of wildebeest in the Serengeti National Park in Tanzania is an essential tool for monitoring the health of the population and the broader ecosystem (Estes, 2014). Currently this process is time consuming, as thousands of aerial photos need to be counted manually. Machine learning can relieve this burden by detecting wildebeest in an image and therefore automating this aspect of the count.

Aerial photos from the Serengeti National Park wildebeest census were segmented into images of 64×64 pixels. These images were manually labelled as one of five classes, wildebeest, zebra, trees, rocks and grass. Note that for automated counting of a single species we would only be interested in detecting an animal or not (i.e. two classes); however, here we consider 50 images for each of the five classes for illustrative purposes (250 photos in total).

Fig. 5a depicts the workflow that takes an image and classifies its contents. Each observation is an image (see Fig. 5b), consisting of 4096 dimensions (64×64) representing the intensity of every pixel (image converted to greyscale). The first step is to identify features that are likely to discriminate well before we feed them into the supervised learning machinery. We could use the 'raw' data, that is, the individual pixel intensities; however, when we projected them onto the first two principal components for visualization (see Appendix Fig. A5a) no obvious groupings were present. The lack of structure detection could be due to several plausible reasons: (1) the chosen features are weak predictors; (2) PCA is a linear dimensionality reduction framework and there may be nonlinear dependencies across features; and (3) projecting 4096 dimensions onto only the first two PCs (the data may separate in higher dimensions).

In this case, the individual pixel intensities are not explicitly related to the image's class, suggesting the need for hand-crafted features that describe the object in each image. Such higher-order structures were extracted using rotation-invariant histograms of oriented gradients (HOGs; Dalal & Triggs, 2005; Liu et al., 2014; Torney et al., 2016). When projected onto the first two PCs, the HOGs features showed promising grouping by class, although it was still somewhat concealed (see Appendix Fig. A5b). HOGs features are hard to interpret and are likely to exhibit nonlinear interactions among them (a stark contrast to the pheasant eggs data where PCA was justified by the linear relationships between variables of the same theme). So, to visualize the data better, we employed a nonlinear projection method t-SNE (Maaten & Hinton, 2008; see Fig. 5c and Appendix Fig. A5). For completeness, t-SNE was also applied to the 'raw' features (see Appendix Fig. A5c), where it did a better job than PCA.

Now that we have a set of features (HOGs) that directly relate to the outcome of interest (class), we can use the random forests algorithm to learn patterns that discriminate between classes. Using stratified sampling, the observations were split 70% for training (175 images) and 30% for testing (75 images). We used a random forests classifier to fit the training data and used its variable importance measure to retain only the most informative features (see Fig. 5d). From an initial pool of over 1000 features only a couple of hundred predictors were needed to achieve good predictive performance on the training data set. Finally, the testing data set was used to evaluate the generalization error and is depicted in Fig. 5e via a confusion matrix.

Using such a classifier may soon enable the fully automated detection and identification of animals from aerial count images. If achieved, this will have major implications for conservation as currently the process of detecting and identifying animals within images is the main bottleneck for estimating population sizes from aerial surveys (Torney et al., 2016).

CONCLUDING REMARKS

ML offers a hypothesis-free approach to model complex data sets where the type of relationship between measured variables is unknown. These methodologies circumvent the limitations of many classical statistical models, and are an attractive choice for generating novel hypotheses to describe unwieldy data sets that are being acquired at an unprecedented rate in various fields of animal behaviour research. Regression, classification, clustering and dimensionality reduction are some of the most common tasks that ML can tackle. We have discussed a number of animal behaviour studies where ML was employed to improve on current methods. We also presented three case studies to showcase the use of ML in developing data analytical workflows to answer biological questions. ML will play a pivotal role in translating complex data sets into scientific knowledge and will become a useful addition to the animal behaviourist's analytical toolbox.

Despite its popularity and success ML is no silver bullet. ML algorithms mine for patterns in the data that are best at predicting an outcome. This can introduce a conflict between the way humans and machines perceive these patterns. For example, the confusion matrix in Fig. 5e shows how two wildebeests in the test data were mistakenly labelled as trees. To a human such a mistake is preposterous, but it highlights how the cues used by humans to differentiate between wildebeest and trees are different from the patterns used by the predictive model. Recently, Google released Deep Dream Generator in an attempt to begin to understand how trained models 'view' images (Koch, 2015). The learning task is inverted: instead of labelling an image, Deep Dream Generator is given a label and tries to find regions of the image that closely match that label and enhance them.

Although not knowing exactly how a machine is differentiating between classes can leave a sense of uneasiness among practitioners, through correct validation procedures such predictive models can still be used reliably. This is true as long as the data are representative of the population. For example, biased training data sets cause erratic behaviour of the predictive model. We highlight this issue by reciting an (arguably apocryphal) account of one of the first defence applications of ML (Dreyfus & Dreyfus, 1992). A neural network was trained to discriminate between images in a forest with 'tanks' or 'no tanks'. The model generalized exceptionally well on the testing data set, suggesting that it had identified valid discriminative patterns. However, when an independent data set (collected afterwards) was used, the classifier was no better than random guessing. It was later noticed that all original 'tanks' photos were taken on a sunny day, while the 'no tanks' photos were taken on a cloudy day, corrupting the patterns learnt by the neural network.

High-profile failures highlight how even experts in the field of ML can fall victim of common pitfalls in data-driven analytics.

Table 1
Nonexhaustive list of popular machine-learning algorithms and their respective implementations in R and Python

Task	Method	R		Python	
		Package	Function	Module	Class
Dimensionality reduction	Principal component analysis	`stats`	`princomp`	`sklearn.decomposition`	`PCA`
	t-Distributed stochastic neighbour embedding (t-SNE)	`Rtsne`	`Rtsne`	`sklearn.manifold`	`TSNE`
Clustering	*k*-means	`stats`	`kmeans`	`sklearn.cluster`	`KMeans`
	Agglomerative hierarchical	`stats`	`hclust`	`sklearn.cluster`	`AgglomerativeClustering`
	Gaussian mixture model	`mclust`	`Mclust`	`sklearn.mixture`	`GMM`
Classification/ Regression	k-Nearest neighbours	`FNN`	`knn` `knn.reg`	`sklearn.neighbors`	`KNeighborsClassifier` `KNeighborsRegressor`
	Decision trees	`tree`	`tree`	`sklearn.tree`	`DecisionTreeClassifier` `DecisionTreeRegressor`
	Random forest	`randomForest`	`randomForest`	`sklearn.ensemble`	`RandomForestClassifier` `RandomForestRegressor`
	Gradient boosting trees	`gbm`	`gbm`	`sklearn.ensemble`	`GradientBoostingClassifier` `GradientBoostingRegressor`
	Support vector machines	`e1071`	`svm`	`sklearn.svm`	`SVC` `SVR`

Typically, ML algorithms mine for patterns in observational data, rather than experimental data, where correlation can be mistaken for causation. Associations between an outcome and a set of inputs could have occurred by chance, owing to a confounding factor or a biased data set. These issues are well studied in statistics (MacKinnon, Krull, & Lockwood, 2000; Zuur, Ieno, & Elphick, 2010), and are now receiving more attention in the ML literature (A. Liu & Ziebart, 2014). Users should be aware of such limitations, as they could have serious downstream implications. Reproducibility of a predictive model over multiple independent data sets over time is ultimately the most rigorous form of validation.

Table 1 provides a nonexhaustive list of popular ML algorithms for dimensionality reduction, clustering, classification and regression, and their respective implementations in R and Python. We advise new users to familiarize themselves with the hyperparameters of the specific ML algorithm and how to tune them effectively using cross-validation. The caret package in R (Kuhn, 2008) and scikit-learn module in Python (Pedregosa et al., 2011) provide a common interface to a wide range of ML algorithms, together with numerous auxiliary functions to perform visualization, cross-validation, model evaluation and more.

A popular question from new users of ML is, which algorithm to use? The type of data (labelled or unlabelled) and the particular task at hand ultimately dictate the category of algorithms to choose from (see Table 1). However, as we have seen earlier, unsupervised learning methods are routinely used with labelled data during the data exploration stage. They could reveal groupings that would have been ignored during manual labelling (e.g. subtle behavioural states). There are also cases where a mix of labelled and unlabelled data are available, which would require the use of a different type of ML methods: semisupervised learning (Chapelle, Schölkopf, & Zien, 2006). These methods make use of unlabelled data to improve the performance of predictive models identified from labelled observations. Their discussion, however, is beyond the scope of this introductory review.

Picking an algorithm from a certain category, e.g. regression, may seem like a daunting decision. However, there are no hard and fast rules to this process and it has been shown that the performance of several state-of-the-art supervised learning algorithms are similar (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). Features that are directly relevant to the outcome being predicted tend to make the predictive performance insensitive to the choice of algorithm. To this end, automatically extracting predictive features from 'raw' data is the central theme of a new set of methods, in an active area of research, called Deep Learning (LeCun, Bengio, & Hinton, 2015). Instead of using hand-crafted features (as we did in the wildebeest identification problem), Deep Learning automatically discovers features amenable for prediction by recursively applying simple but nonlinear transformations to the data. We envisage that these methods will become widely adopted by the animal behaviour community once packaged in an easy-to-use form.

Acknowledgments

We thank Paul Gluyas and the staff at Pencoose farm for allowing us to study jackdaws on their land, Mark Whiteside for assisting in the collection of the pheasant egg measurements, and Grant Hopcraft, Felix Borner and Andy Dobson for helpful discussions about the wildebeest count. J.J.V. is funded by an MRC grant (MR/M003906/1). A.T. and M.K. received funding from a BBSRC David Phillips Fellowship (BB/H021817/1) and a BBSRC SWDTP studentship (630051486), respectively. C.J.T. is supported by a James S. McDonnell Foundation Studying Complex Systems Scholar Award. J.R.M. is funded by an ERC consolidator award (616474).

References

Acevedo, M. A., Corrada-Bravo, C. J., Corrada-Bravo, H., Villanueva-Rivera, L. J., & Aide, T. M. (2009). Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics, 4*(4), 206–214. http://dx.doi.org/10.1016/j.ecoinf.2009.06.005.

Andersson, M., & Åhlund, M. (2001). Protein fingerprinting: A new technique reveals extensive conspecific brood parasitism. *Ecology, 82*(5), 1433–1442. http://dx.doi.org/10.1890/0012-9658(2001)082[1433:PFANTR]2.0.CO;2.

Bishop, C. (2006). *Pattern recognition and machine learning. Pattern recognition* (Vol. 4). New York, NY: Springer-Verlag. http://dx.doi.org/10.1117/1.2819119.

Blumstein, D. T., Mennill, D. J., Clemins, P., Girod, L., Yao, K., Patricelli, G., et al. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *Journal of Applied Ecology, 48*, 758–767. http://dx.doi.org/10.1111/j.1365-2664.2011.01993.x.

Brandes, T. S. (2008). Automated sound recording and analysis techniques for bird surveys and conservation. *Bird Conservation International, 18*(S1), S163–S173. http://dx.doi.org/10.1017/S0959270908000415.

Breiman, L. (2001a). Random forests. *Machine Learning, 45*, 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–215. http://dx.doi.org/10.2307/2676681.

Broell, F., Noda, T., Wright, S., Domenici, P., Steffensen, J. F., Auclair, J.-P., et al. (2013). Accelerometer tags: Detecting and identifying activities in fish and the effect of

sampling frequency. *Journal of Experimental Biology, 216*(Pt 7), 1255–1264. http://dx.doi.org/10.1242/jeb.077396.
Broom, D. M., & Johnson, K. G. (1993). *Stress and animal welfare* (Vol. 30). New York, NY: Springer Science & Business Media. Retrieved from: https://books.google.com/books?hl=en&lr=&id=LW-gihinLy8C&pgis=1.
Brown, D. D., Kays, R., Wikelski, M., Wilson, R., & Klimley, A. P. (2013). Observing the unwatchable through acceleration logging of animal behavior. *Animal Biotelemetry, 2013*, 1–16. http://dx.doi.org/10.1186/2050-3385-1-20.
Burges, C. J. C. (2009). Dimension reduction: A guided tour. *Foundations and Trends® in Machine Learning, 2*(4), 275–364. http://dx.doi.org/10.1561/2200000002.
Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S., et al. (2007). Multimodal emotion recognition from expressive faces, body gestures and speech. *IFIP International Federation for Information Processing, 247*, 375–388. http://dx.doi.org/10.1007/978-0-387-74161-1_41.
Carroll, G., Slip, D., Jonsen, I., & Harcourt, R. (2014). Supervised accelerometry analysis can identify prey capture by penguins at sea. *Journal of Experimental Biology, 217*(Pt 24), 4295–4302. http://dx.doi.org/10.1242/jeb.113076.
Castellano, G., Villalba, S. D., & Camurri, A. (2007). Recognising human emotions from body movement and gesture dynamics. *Affective Computing and Intelligent Interaction, 4738*, 71–82. http://dx.doi.org/10.1007/978-3-540-74889-2_7.
Chapelle, O., Schölkopf, B., & Zien, A. (2006). *Semi-supervised learning*. Cambridge, MA: The MIT Press.
Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software, 61*(6), 1–36.
Cheng, J., Xie, B., Lin, C., & Ji, L. (2012). A comparative study in birds: Call-type-independent species and individual recognition using four machine-learning methods and two acoustic features. *Bioacoustics, 21*(2), 157–171. http://dx.doi.org/10.1080/09524622.2012.669664.
Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L., & Mogil, J. S. (2002). Identification and ranking of genetic and laboratory environment factors influencing a behavioral trait, thermal nociception, via computational analysis of a large data archive. *Neuroscience and Biobehavioral Reviews, 26*(8), 907–923. http://dx.doi.org/10.1016/S0149-7634(02)00103-3.
Clayton, N. S., & Emery, N. J. (2007). The social life of corvids. *Current Biology, 17*(16), R652–R656. http://dx.doi.org/10.1016/j.cub.2007.05.070.
Croft, D. P., Madden, J. R., Franks, D. W., & James, R. (2011). Hypothesis testing in animal social networks. *Trends in Ecology & Evolution, 26*(10), 502–507. http://dx.doi.org/10.1016/j.tree.2011.05.012.
Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., et al. (2007). Random forests for classification in ecology. *Ecology, 88*(11), 2783–2792. http://dx.doi.org/10.1890/07-0539.1.
D'haeseleer, P. (2005). How does gene expression clustering work? *Nature Biotechnology, 23*(12), 1499–1501. http://dx.doi.org/10.1038/nbt1205-1499.
Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. 886–893). http://dx.doi.org/10.1109/CVPR.2005.177.
Dankert, H., Wang, L., Hoopfer, E. D., Anderson, D. J., & Perona, P. (2009). Automated monitoring and analysis of social behavior in *Drosophila*. *Nature Methods, 6*(4), 297–303. http://dx.doi.org/10.1038/nmeth.1310.
Dell, A. I., Bender, J. A., Branson, K., Couzin, I. D., de Polavieja, G. G., Noldus, L. P. J. J., et al. (2014). Automated image-based tracking and its application in ecology. *Trends in Ecology & Evolution, 29*(7), 417–428. http://dx.doi.org/10.1016/j.tree.2014.05.004.
van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605. http://dx.doi.org/10.1007/s10479-011-0841-3.
Dreyfus, H. L., & Dreyfus, S. E. (1992). What artificial experts can and cannot do. *AI & Society, 6*(1), 18–26. http://dx.doi.org/10.1007/BF02472766.
Estes, R. (2014). *The Gnu's world: Serengeti wildebeest ecology and life history*. Berkeley, CA: University of California Press. Retrieved from: http://www.amazon.co.uk/Gnus-World-Serengeti-Wildebeest-Ecology-ebook/dp/B00JNMCBHG/ref=sr_1_1?s=books&ie=UTF8&qid=1460971076&sr=1-1&keywords=9780520958197.
Fagerlund, S. (2007). Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing, 2007*(1), 38637. http://dx.doi.org/10.1155/2007/38637.
Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research, 15*, 3133–3181.
Franz, M., & Nunn, C. L. (2009). Network-based diffusion analysis: A new method for detecting social learning. *Proceedings of the Royal Society B: Biological Sciences, 276*(1663), 1829–1836. http://dx.doi.org/10.1098/rspb.2008.1824.
Friedman, J. H. (2001). The role of statistics in the data revolution? *International Statistical Review, 69*(1), 5–10. http://dx.doi.org/10.2307/1403524.
Ge, W., Collins, R. T., & Ruback, R. B. (2012). Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 34*(5), 1003–1016. http://dx.doi.org/10.1109/TPAMI.2011.176.
Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452–459. http://dx.doi.org/10.1038/nature14541.
Gleiss, A. C., Wright, S., Liebsch, N., Wilson, R. P., & Norman, B. (2013). Contrasting diel patterns in vertical movement and locomotor activity of whale sharks at Ningaloo Reef. *Marine Biology, 160*(11), 2981–2992. http://dx.doi.org/10.1007/s00227-013-2288-3.
Gregory, S. (2011). Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment, 2011*, P02017. http://dx.doi.org/10.1088/1742-5468/2011/02/P02017.
Grünewälder, S., Broekhuis, F., Macdonald, D. W., Wilson, A. M., McNutt, J. W., Shawe-Taylor, J., et al. (2012). Movement activity based classification of animal behaviour with an application to data from cheetah (*Acinonyx jubatus*). *PLoS One, 7*(11), 1–11. http://dx.doi.org/10.1371/journal.pone.0049120.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182. http://dx.doi.org/10.1162/153244303322753616.
Hart, N. S. (2002). Vision in the peafowl (Aves: *Pavo cristatus*). *Journal of Experimental Biology, 205*, 3925–3935.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (Vol. 1). New York, NY: Springer Science. http://dx.doi.org/10.1007/b94608.
Hussey, N. E., Kessel, S. T., Aarestrup, K., Cooke, S. J., Cowley, P. D., Fisk, A. T., et al. (2015). Aquatic animal telemetry: A panoramic window into the underwater world. *Science (New York, N.Y.), 348*(6240), 1255642. http://dx.doi.org/10.1126/science.1255642.
Hutchinson, J. M. C., & Gigerenzer, G. (2005). Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes, 69*(2), 97–124. http://dx.doi.org/10.1016/j.beproc.2005.02.019.
Huynh-Thu, V. A., Saeys, Y., Wehenkel, L., & Geurts, P. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics, 28*(13), 1766–1774. http://dx.doi.org/10.1093/bioinformatics/bts238.
Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution, 19*(2), 101–108. http://dx.doi.org/10.1016/j.tree.2003.10.013.
Kabra, M., Robie, A., Rivera-Alba, M., Branson, S., & Branson, K. (2013). JAABA: Interactive machine learning for automatic annotation of animal behavior. *Nature Methods, 10*(1), 64–67. http://dx.doi.org/10.1038/nmeth.2281.
Kafkafi, N., Yekutieli, D., & Elmer, G. I. (2009). A data mining approach to in vivo classification of psychopharmacological drugs. *Neuropsychopharmacology, 34*(3), 607–623. http://dx.doi.org/10.1038/npp.2008.103.
Katz, Y., Tunstrom, K., Ioannou, C. C., Huepe, C., & Couzin, I. D. (2011). Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences of the United States of America, 108*(46), 18720–18725. http://dx.doi.org/10.1073/pnas.1107583108.
Kays, R., Crofoot, M. C., Jetz, W., & Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science, 348*(6240). http://dx.doi.org/10.1126/science.aaa2478. aaa2478–aaa2478.
Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., et al. (2014). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Reviews, 91*(1), 13–52. http://dx.doi.org/10.1111/brv.12160.
Kershenbaum, A., Root-Gutteridge, H., Habib, B., Koler-Matznick, J., Mitchell, B., Palacios, V., et al. (2016). Disentangling canid howls across multiple species and subspecies: Structure in a complex communication channel. *Behavioural Processes, 124*, 149–157. http://dx.doi.org/10.1016/j.beproc.2016.01.006.
Koch, C. (2015). Do androids dream? *Scientific American Mind, 26*(6), 24–27. http://dx.doi.org/10.1038/scientificamericanmind1115-24.
Kramer, K., & Kinter, L. B. (2003). Evaluation and applications of radiotelemetry in small laboratory animals. *Physiological Genomics, 13*, 197–205. http://dx.doi.org/10.1152/physiolgenomics.00164.2002.
Krause, J., James, R., Franks, D. W., & Croft, D. P. (2015). *Animal social networks*. Oxford, U.K.: Oxford University Press.
Krause, J., Krause, S., Arlinghaus, R., Psorakis, I., Roberts, S., & Rutz, C. (2013). Reality mining of animal social systems. *Trends in Ecology & Evolution, 28*(9), 541–551. http://dx.doi.org/10.1016/j.tree.2013.06.002.
Krause, J., Wilson, A. D. M., & Croft, D. P. (2011). New technology facilitates the study of social networks. *Trends in Ecology & Evolution, 26*(1), 5–6. http://dx.doi.org/10.1016/j.tree.2010.10.004.
Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*(5), 1–26. http://dx.doi.org/10.1053/j.sodo.2009.03.002.
Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., et al. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics/IEEE, ACM, 9*(4), 1106–1119. http://dx.doi.org/10.1109/TCBB.2012.33.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. http://dx.doi.org/10.1038/nature14539.
Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. New York, NY: Springer. Retrieved from: http://www.springer.com/gp/book/9780387393506.
Liu, H., & Motoda, H. (2008). *Computational methods of feature selection*. Boca Raton, FL: CRC Press. Retrieved from: https://www.crcpress.com/Computational-Methods-of-Feature-Selection/Liu-Motoda/9781584888789.
Liu, K., Skibbe, H., Schmidt, T., Blein, T., Palme, K., Brox, T., et al. (2014). Rotation-invariant HOG descriptors using Fourier analysis in polar and spherical coordinates. *International Journal of Computer Vision, 106*(3), 342–364. http://dx.doi.org/10.1007/s11263-013-0634-z.
Liu, A., & Ziebart, B. (2014). Robust classification under sample selection bias. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27, pp. 37–45). Newry, U.K.: Curran Associates. Retrieved from: http://papers.nips.cc/paper/5458-robust-classification-under-sample-selection-bias.pdf.

van der Maaten, L., Eric, P., & van den Herik, J. (2009). *Dimensionality reduction: A comparative review.* Tilburg, Netherlands: Tilburg Centre for Creative Computing, Tilburg University. Technical Report: 2009-005.
MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding and suppression effect. *Prevention Science, 1*(4), 173–181. http://dx.doi.org/10.1023/A:1026595011371.
Maia, R., Eliason, C. M., Bitton, P.-P., Doucet, S. M., & Shawkey, M. D. (2013). Pavo: An R package for the analysis, visualization and organization of spectral data. *Methods in Ecology and Evolution, 4*, 906–913. http://dx.doi.org/10.1111/2041-210X.12069.
McClune, D. W., Marks, N. J., Wilson, R. P., Houghton, J. D., Montgomery, I. W., McGowan, N. E., et al. (2014). Tri-axial accelerometers quantify behaviour in the Eurasian badger (*Meles meles*): Towards an automated interpretation of field data. *Animal Biotelemetry, 2*(1), 5. http://dx.doi.org/10.1186/2050-3385-2-5.
Michel, P., & El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th international conference on multimodal interfaces–ICMI'03* (p. 258). http://dx.doi.org/10.1145/958468.958479.
Murphy, K. (2012). *Machine learning: A probabilistic perspective. Chance encounters: Probability in education*. New York, NY: Springer-Verlag. http://dx.doi.org/10.1007/SpringerReference_35834.
Nadimi, E. S., Søgaard, H. T., & Bak, T. (2008). ZigBee-based wireless sensor networks for classifying the behaviour of a herd of animals using classification trees. *Biosystems Engineering, 100*(2), 167–176. http://dx.doi.org/10.1016/j.biosystemseng.2008.03.003.
Nagy, M., Akos, Z., Biro, D., & Vicsek, T. (2010). Hierarchical group dynamics in pigeon flocks. *Nature, 464*(7290), 890–893. http://dx.doi.org/10.1038/nature08891.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. http://dx.doi.org/10.1007/s13398-014-0173-7.2.
Piles, M., Díez, J., del Coz, J. J., Montañés, E., Quevedo, J. R., Ramon, J., et al. (2013). Predicting fertility from seminal traits: Performance of several parametric and non-parametric procedures. *Livestock Science, 155*(1), 137–147. http://dx.doi.org/10.1016/j.livsci.2013.03.019.
Psorakis, I., Roberts, S. J., Rezek, I., & Sheldon, B. C. (2012). Inferring social network structure in ecological systems from spatio-temporal data streams. *Journal of the Royal Society, Interface, 9*, 3055–3066. http://dx.doi.org/10.1098/rsif.2012.0223.
Psorakis, I., Voelkl, B., Garroway, C. J., Radersma, R., Aplin, L. M., Crates, R. A., et al. (2015). Inferring social structure from temporal data. *Behavioral Ecology and Sociobiology, 2015*, 857–866. http://dx.doi.org/10.1007/s00265-015-1906-0.
Ranjard, L., & Ross, H. A. (2008). Unsupervised bird song syllable classification using evolving neural networks. *Journal of the Acoustical Society of America, 123*(6), 4358–4368. http://dx.doi.org/10.1121/1.2903861.
Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology, 26*(3), 303–304. http://dx.doi.org/10.1038/nbt0308-303.
Roughan, J. V., Wright-Williams, S. L., & Flecknell, P. A. (2009). Automated analysis of postoperative behaviour: Assessment of HomeCageScan as a novel method to rapidly identify pain and analgesic effects in mice. *Laboratory Animals, 43*(1), 17–26. http://dx.doi.org/10.1258/la.2008.007156.
Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.
Shami, M., & Verhelst, W. (2007). An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech. *Speech Communication, 49*(3), 201–212. http://dx.doi.org/10.1016/j.specom.2007.01.006.
Shi, Y., Nguyen, M. H., Blitz, P., French, B., Fisk, S., De la Torre, F., et al. (2010, June). Personalized stress detection from physiological measurements. In *International symposium on quality of life technology* (pp. 28–29).
Stoddard, M. C., & Prum, R. O. (2008). Evolution of avian plumage color in a tetrahedral color space: A phylogenetic analysis of new world buntings. *American Naturalist, 171*(6), 755–776. http://dx.doi.org/10.1086/587526.
Taylor, A., Watson, G., Grigg, G., & McCallum, H. (1996). Monitoring frog communities: An application of machine learning. In *Proceedings of the 8th Innovative applications of artificial conference* (pp. 1564–1569).
Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., & Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. *Animal Behaviour, 59*(6), 1167–1176. http://dx.doi.org/10.1006/anbe.1999.1416.
Tibshirani, R. (1996). Regression and shrinkage via the Lasso. *Journal of the Royal Statistical Society B, 58*(1), 267–288.
Tomkiewicz, S. M., Fuller, M. R., Kie, J. G., & Bates, K. K. (2010). Global positioning system and associated technologies in animal behaviour and ecological research. *Philosophical Transactions of the Royal Society B: Biological Sciences, 365*(1550), 2163–2176. http://dx.doi.org/10.1098/rstb.2010.0090.
Torney, C. J., Dobson, A. P., Borner, F., Lloyd-Jones, D. J., Moyer, D., Maliti, H. T., et al. (2016). Assessing rotation-invariant feature classification for automated wildebeest population counts. *PLoS One, 11*(5), e0156342. http://dx.doi.org/10.1371/journal.pone.0156342.
Wang, Y., Nickel, B., Rutishauser, M., Bryce, C. M., Williams, T. M., Elkaim, G., et al. (2015). Movement, resting, and attack behaviors of wild pumas are revealed by tri-axial accelerometer measurements. *Movement Ecology, 3*(1), 1–12. http://dx.doi.org/10.1186/s40462-015-0030-0.
Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining: Practical machine learning tools and techniques* (3rd ed.). Burlington, MA: Morgan Kaufmann. http://dx.doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C.
Yom-Tov, Y. (1980). Intraspecific nest parasitism in birds. *Biological Reviews, 55*(1), 93–108. http://dx.doi.org/10.1111/j.1469-185X.1980.tb00689.x.
Zhang, J., O'Reilly, K. M., Perry, G. L. W., Taylor, G. A., & Dennis, T. E. (2015). Extending the functionality of behavioural change-point analysis with k-means clustering: A case study with the little penguin (*Eudyptula minor*). *PLoS One, 10*(4), 1–14. http://dx.doi.org/10.1371/journal.pone.0122811.
Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3–14. http://dx.doi.org/10.1111/j.2041-210X.2009.00001.

APPENDIX

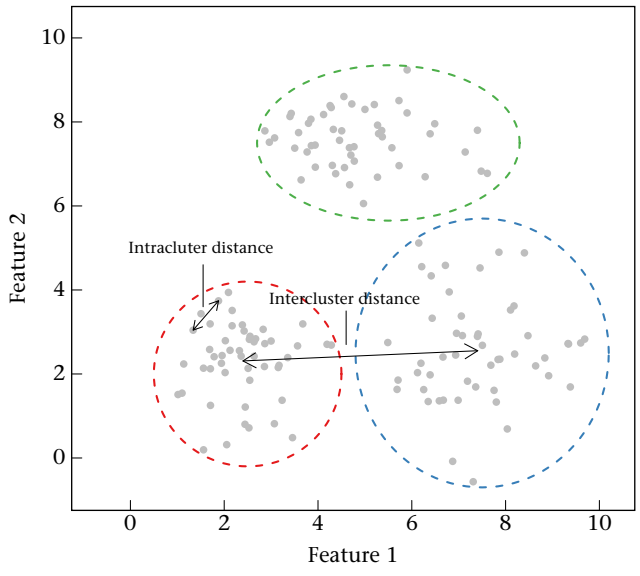

Figure A1. A simulated data set with two features and three classes to illustrate the objective of clustering: to minimize the within-group similarity (intracluster distance) and maximize the distance between distinct clusters (intercluster distance).

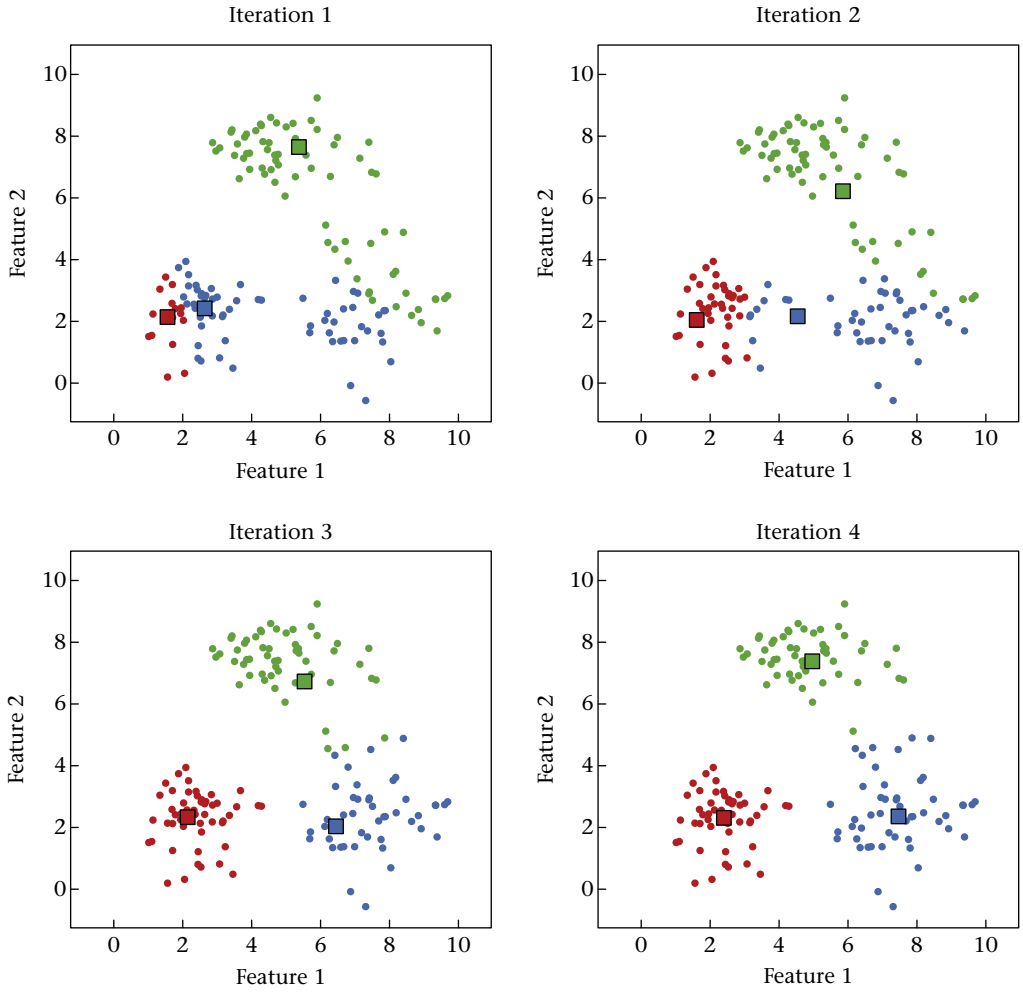

Figure A2. A simulated data set with two features and three classes to illustrate the first four iterations of the *k*-means clustering algorithm. The squares indicate the centre of the distinct clusters, which are represented by a different colour.

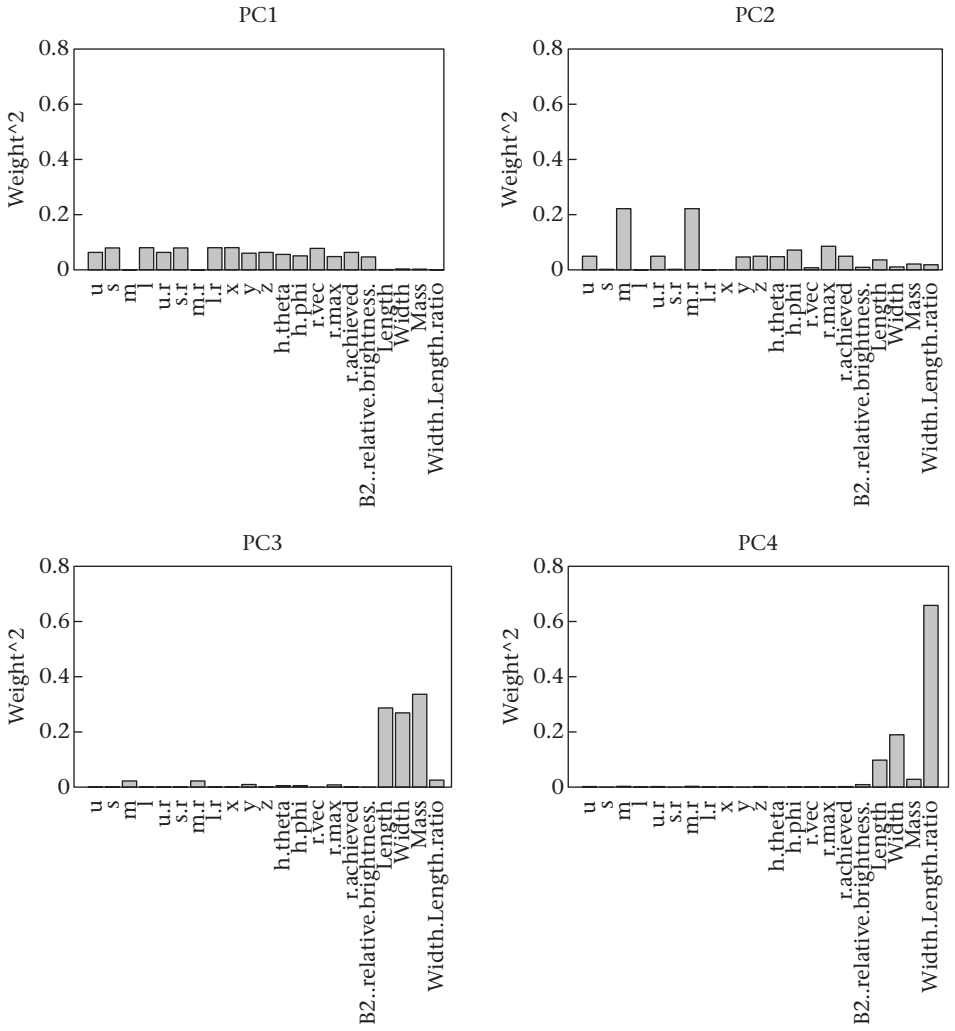

Figure A3. PCA weights for the pheasant eggs case study data. PC1 is a measure of eggshell brightness, PC2 is a measure of eggshell greenness, PC3 is a measure of egg size and PC4 is a measure of egg shape.

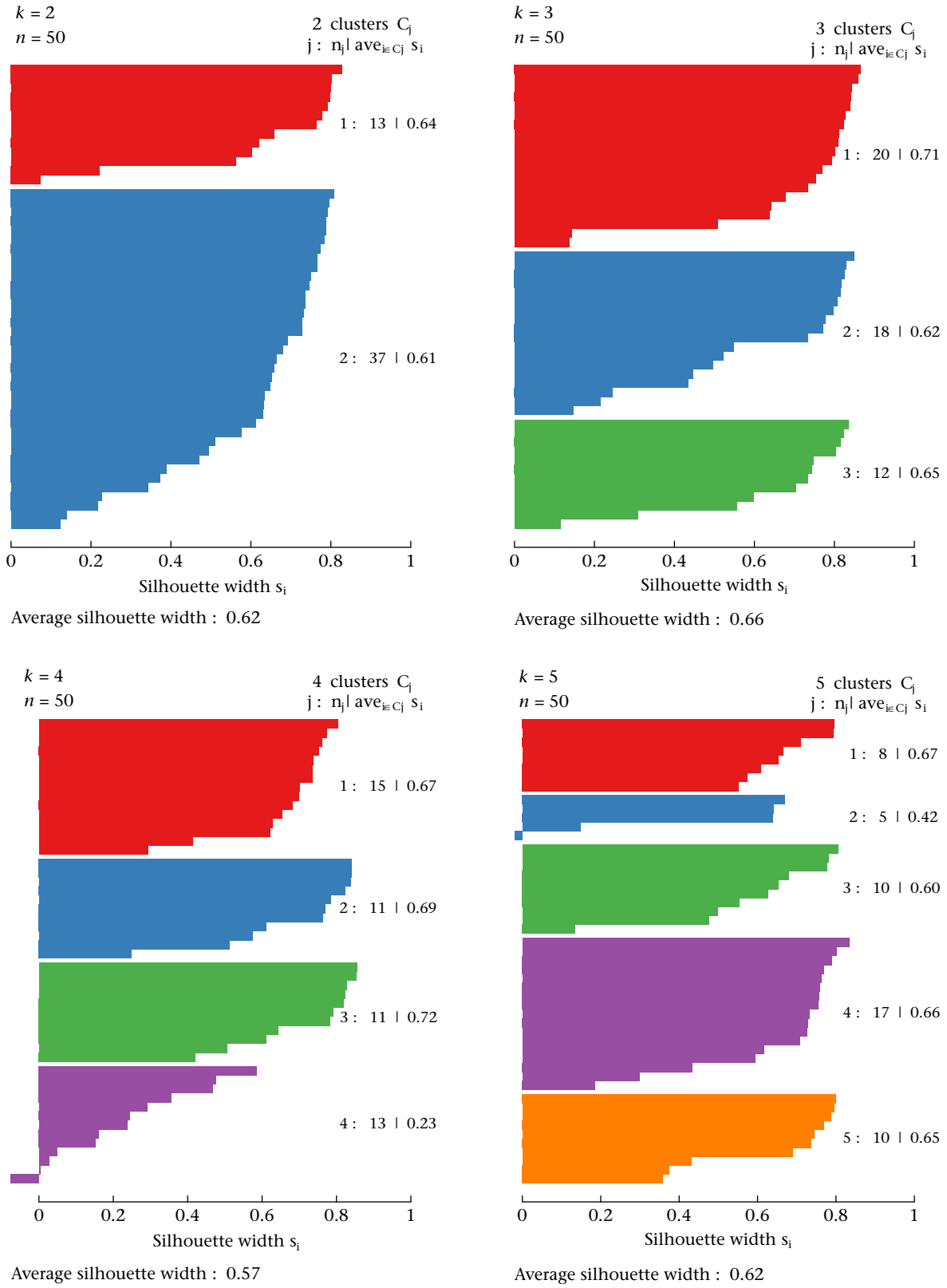

Figure A4. Silhouette plots for the pheasant eggs case study with different values for the total number of clusters *k*. *n* is the total number of eggs, *nj* is the number of eggs in the *j*th cluster and *s* is the silhouette width for that cluster. The average silhouette width across all clusters is given at the bottom of each plot. $k = 3$ maximizes the average silhouette width, suggesting three different clutches of eggs.

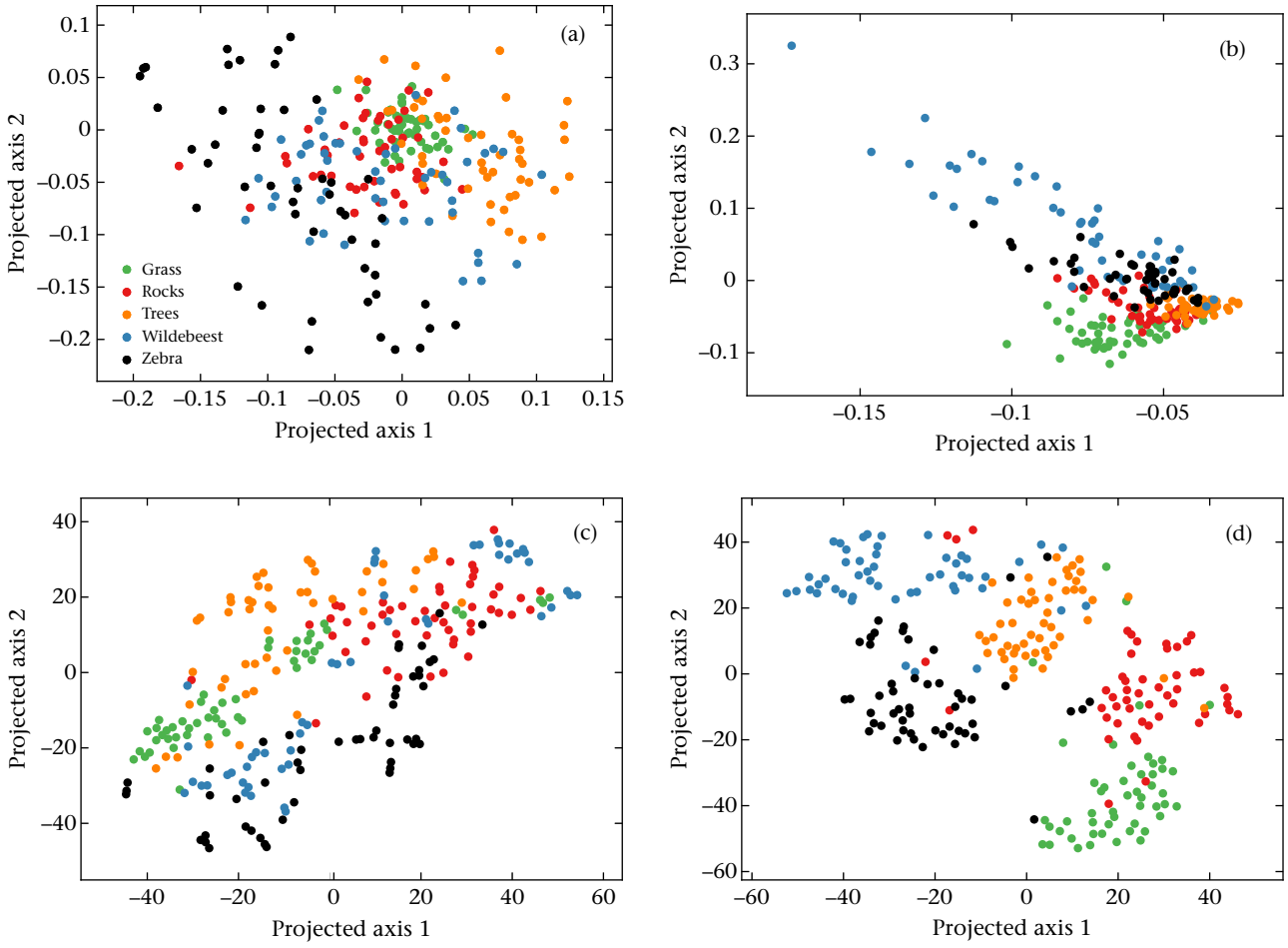

Figure A5. Comparison of different feature sets ((a, c) raw features versus (b, d) histogram of oriented gradients (HOGs)) and projection methods ((a, b) PCA versus (c, d) t-SNE) for the wildebeest identification data.