Putting the Scientist in the Loop - Accelerating Scientific Progress with Interactive Machine Learning

Oisin Mac Aodha[1] Vassilios Stathopoulos[5] Michael Terry[2] Kate E. Jones[3,4]
Gabriel J. Brostow[1] Mark Girolami[5]

[1]Department of Computer Science, University College London
[2]David R. Cheriton School of Computer Science, University of Waterloo
[3]Center for Biodiversity and Environment Research, University College London
[4]Institute of Zoology, Zoological Society of London
[5]Department of Statistics, Warwick University
www.engage-project.org

***Abstract*—Technology drives advances in science. Giving scientists access to more powerful tools for collecting and understanding data enables them to both ask and answer new kinds questions that were previously beyond their reach. Of these new tools at their disposal, machine learning offers the opportunity to understand and analyze data at unprecedented scales and levels of detail.**

The standard machine learning pipeline consists of data labeling, feature extraction, training, and evaluation. However, without expert machine learning knowledge, it is difficult for scientists to optimally construct this pipeline to fully leverage machine learning in their work. Using ecology as a motivating example, we analyze a typical scientist's data collection and processing workflow and highlight many problems facing practitioners when attempting to capitalize on advances in machine learning and pattern recognition. Understanding these shortcomings allows us to outline several novel and underexplored research directions. We end with recommendations to motivate progress in future cross-disciplinary work.

Keywords*—*interactive machine learning; computer vision; human-computer interaction; data visualization; ecology; biodiversity;

I. Introduction

A growing body of evidence links human health and well-being to ecosystems and the services they provide *e.g.* air and water purification, carbon storage [1], [2]. However, ecosystems are being changed at unprecedented rates as land is converted to anthropogenic use and increasing climate variability changes distributions of animals and plants [3], [4]. Monitoring ecosystem and species declines is critical to model and predict impacts of global change. Technological advances in remote audio and visual sensors have meant that ecologists or groups of citizen scientists are regularly deploying large sensor arrays to gather data on biodiversity change [5], [6]. As a result there are now ecological datasets of unprecedented sizes in a range of formats (*e.g.* audio, images and video).

While data acquisition has become easier and less expensive, extracting signals of interest from these large, unstructured, noisy datasets poses numerous challenges. In the past, biologists would manually *code* the data, hand-labeling it according to features of interest (such as the presence of relevant species). Some crowdsourcing and citizen science projects have also been shown to be effective at mobilizing large groups of individuals to help with the annotation [7], [8], [9], [10]. Unfortunately, in many scenarios annotations from non-experts can be noisy, and it can thus be expensive to recruit the large numbers of skilled labelers required for big datasets [11].

A more scaleable solution is to apply the tools of statistical machine learning to the problem of signal extraction. In the last two decades, the use of machine learning as an alternative to manual human labeling has started to gain traction in ecology [12], [13], [14]. However, large-scale datasets typically require at least two types of expertise for signal extraction: domain expertise (*e.g.* expertise in biology and ecology) and expertise in data processing and analysis. Domain expertise drives signal acquisition and interpretation needs, while expertise in data analysis provides the requisite skills to reliably extract valid signals from the data. Often, these two skill sets are not held by the same person, requiring tight collaborations between domain experts. In the ideal, the time spent extracting signals from the data would be reduced in order to maximize the amount of time spent engaging with the underlying science. Even more ideal, is empowering biologists to perform signal extraction and analysis without the need for external collaborations.

In many instances, biology researchers are consumers of machine learning tools. Typically, their concerns reside less with the design of the underlying algorithms, and more with how the tools can be leveraged to help them study changes in biodiversity over time. However, at present, effective use of machine learning techniques often requires a significant amount of in-depth, model-specific, knowledge of these models and their theoretical underpinnings. This is problematic for scientists who simply desire to apply these tools to their specific problems. This disconnect between the products of

the machine learning research community and the potential users of this technology has been noticed by others. Recently, Wagstaff [15] argued that the machine learning community has become overly focused on algorithmic improvements at the expense of pursuing real-world impact in end application domains. Additionally, it is not always clear that progress made on benchmark datasets, which can often be strongly biased [16], translates to improvements in real-world problems [15].

In contrast to the traditional machine learning pipeline, interactive machine learning (IML) is concerned with saving the user time by explicitly including them as part of the annotation and training loop *e.g*. [17]. IML assists the end user by helping them choose features, models, and model parameters. Here the user is not just a labeler, but instead their role is both guide and explorer. Early work in this field would involve users by asking them to label training data and indicate when the output of the system is incorrect, thus signaling that the current model is inadequate or insufficiently tuned to the problem [17]. This process results in an iteratively and automatically created predictive recognition system tuned with the training data available. When the user is satisfied with the performance of the system, they can then apply the system to the specific problem at hand or continue to interactively supervise learning. More recent research in IML goes further, and seeks to provide individuals with tools to actively explore, find patterns, and generally understand the data as the concepts they are interested in evolve and change [18].

Drawing upon observations conducted during an ongoing collaboration between researchers in ecology, statistical machine learning, computer vision, and human-computer interaction, we describe how biodiversity scientists employ the traditional machine learning pipeline. This analysis allows us to highlight several areas of research that are under explored by the mainstream machine learning and pattern recognition communities.

II. Data Understanding Pipeline

Here we outline the main steps in the workflow of a biologist tasked with posing and answering a novel research question where there is a strong dependence on data collection and model building. This pipeline, depicted in Figure 1, is not intended to be representative of every specific case, but instead provides a framework to allow us to understand a general workflow. It is worth pointing out that these steps need not be linear, but potentially consist of feedback loops at any stage.

A. Hypothesis Formation

As a first step, the scientist defines the problem domain they are interested in working on. An ecologist may be interested in problems such as estimating species distributions their interactions, or the study of particular habitats, amongst others. The initial hypothesis may be broad, and in cases it is likely to be adapted as the understanding of the problem changes.

B. Data Collection and Capture

The choice of hypothesis defines the types of data that will be required. Traditionally, data collection was a very laborious and time consuming process. In the case of species distribution modeling, it would have been necessary to go into the field, potentially up to several times over many years, to manually record the presence or absence of different species in an area. It is also fraught with other potential problems, such as the difficulty of accessing remote regions, observer bias, and the challenges associated with monitoring illusive species.

Newer technology in the form of remote sensing has made it possible to automatically collect some of this data. Camera traps with motion sensors are frequently used to capture images of any objects that move in front of them [5]. They were originally developed for commercial hunters but are now becoming more widely used in terrestrial biodiversity monitoring and conservation [19]. The data collector's job is to place the cameras in the wild and move them whenever is deemed necessary. Images from these cameras can then be recovered manually or, for more sophisticated cameras, remotely using satellite or mobile phone networks [19]. Audio recording is also an effective way to determine species counts for certain animals [20], [21]. Audio recording technology has also enabled teams of citizen scientists to assist in data capture [6].

Another advantage of automated data collection is that it enables the acquisition of very fine grained information. For example, the use of video to record species interactions allows researchers to capture information about animal position and motion that can be difficult to record manually. It is also possible to determine behavior and interactions between individuals [22].

C. Data Exploration

After collecting the data, the next step is to explore it in order to identify potential patterns and trends within. At this stage, the goal is not to densely label every item in the dataset, but to instead get a basic understanding of the variation present. For this activity, it is desirable to have a method that allows for fast exploration. For audio and images, one standard solution is to take random samples of the data to see what it contains. However, this sampling may not be effective if the signal of interest is only present in a very small subset of the entire dataset. On completion of the initial data exploration stage, it may be deemed necessary to go back and collect more data.

D. Production of Set of Canonical Examples

Once there is a general understanding of what the data contains, the next step is to produce a set of reference examples. This is similar to the construction of a field guide – a reference set of examples that depicts the main classes of interest. For species identification in images, this may take the form of a set of example images containing the desired species. For audio, this can be a selection of short audio clips. Unless the audio is collected in a controlled manner *e.g*. by also performing visual identification of the species when recording, it may often be necessary to obtain a separate database of known animal sounds as there is no way of verifying what is actually contained in the audio that has been collected.

E. Establish Targets and Goals of Analysis

Once the scientist has a sense of the specific questions to ask, the next step is to define how the data needs to be labeled and processed to achieve this. This step determines what level of annotation will be required in the labeling stage. It also determines the types of features necessary to represent the data for the problem at hand. In species identification, it may only be necessary to classify a short clip of audio or an image as containing a particular species or not (*i.e*. classification). Alternatively, the precise location of the species may also be required (*i.e*. detection). Another common task is to determine the relationship between a particular input and a continuous output that one wishes to measure (*i.e*. regression). The subsequent labeling and annotation can be very time consuming, and as a result, a better understanding of the goals of analysis can reduce this effort by only focusing on the relevant and necessary annotation.

F. Data Preprocessing

Once the goals of the data analysis have been established, it is then often necessary to preprocess the data to make it more amenable for later use. A typical first step is to define a directory structure for the dataset so that it is organized based on some high level property *e.g*. object class, location, time. Conditioned on the type and quality of data, additional 'clean up' can be performed. This can range from removing irrelevant borders in images, remove silent sections and denoising audio, and trimming video to the location in time of interest. It may also be necessary to bring together different data sources *e.g*. aligning different map layers from aerial imagery. This stage is closely linked with the later labeling of the data but it differs in that most of the tasks can be fully automated.

G. Feature Extraction

The collected data in its raw form is usually unsuitable as an input for most machine learning algorithms. For images, the dimensionality of the input images may be on the order of millions, typically far exceeding the number of training examples available. In the feature extraction step, a representation is chosen for the data with the intention that it best captures the signal of interest, while being invariant to the remaining noise present. The choice of 'good' features is not trivial, and this step often necessitates problem-specific intuition on the part of the scientist. A poor choice of representation can make the subsequent model learning stage very difficult, if not impossible.

Designing feature representations is an area of active research in many domains. It can be coarsely divided into two categories, 1) hand-tuned and 2) automatically learned features. For the task of image classification, a widely used representation is the bag of visual words model [23] inspired by the bag of words model used in text representation. In this model, image patches are represented by local image descriptors, *e.g*. SIFT features [24], which are then quantized into one of a number of 'visual words' creating a final fixed length vectorized output. Extensions of this model include introducing spatial information in the form of spatial image pyramid matching [25] and higher dimensional encodings that represent the similarity between each image descriptor and each visual word [26]. This type of representation has also proved popular for action classification in video sequences where the current state of the art uses spatio-temporal features that encode appearance and motion information of small blocks of time [27]. For object detection in images, *i.e*. object localizing, a common approach for rigid object classes is to use histograms of gradient based features [28]. For audio, MFCC features [29] generated from spectrograms are commonly used as a general representation. It is possible to incorporate domain specific knowledge, such as information about the signal structure [14]. However, representations that do not require any problem specific knowledge are more easily adapted to other related tasks [30].

In contrast to these manually designed features, the goal of feature learning is to learn the best representation directly from the data. Methods such as convolutional neural nets [31], [32] have recently performed very well on image and audio classification tasks [33], [34]. The disadvantage of many feature learning approaches is that they can be very computationally expensive.

H. Building Labeling Infrastructure

Depending on the uniqueness of the data, custom tools for annotation may need to be developed. These tools can come in different forms, from standalone desktop applications [35] to web-based tools [19], [7], [11], [36] that can facilitate crowd annotation. The size, annotation type, and complexity of the data dictates the tool requirements. Typically in the machine learning and pattern recognition communities, the data is assumed to be labeled in advance. In comparison to designing new machine learning algorithms, the problem of designing new annotation interfaces has received far less attention.

Image annotations come in different forms and necessitate different interface components for each specific task such as specifying the position of objects [37] or associating a label with each pixel individually [38]. For video, it may be necessary to annotate the location [39], actions performed, and behavior of each object [35]. Videos may also need to be trimmed to localize events more finely in time [40]. Audio annotation tools can present both audio and visual cues to aid the user [7].

I. Labeling

As noted in the previous section, data annotation can require the use of custom built tools. Once the tools are built, the signal extraction can be performed manually, by one or more individuals, or in a semi-automated fashion. Using crowdsourcing, it is possible to collect annotations from citizen scientists [7], [9] or by using online marketplaces for human intelligence tasks [41], [11]. For example, the Bat Detective project [7] uses citizen scientists to help locate bat calls in thousands of audio recordings. However, crowdsourcing may not be viable in every case, as building the labeling infrastructure and recruiting a user base can be costly and take time. Crowd sourced annotations can be very noisy, thus requiring additional processing and careful screening of users [42], [43], [11].

Active learning [44] attempts to reduce the user effort by only asking them to annotate data that the model is most unsure about and which will bring about the greatest reduction in the predicted future error [45]. Its goal is minimize the amount of time they need to spend labeling data. In the past, active learning has been applied to tasks such as action recognition [40], object tracking [39], semantic segmentation [46], object detection using crowd sourced annotations [47], amongst many other tasks. An essential aspect of the labeling process is ensuring reliable and valid labels results. Numerous research efforts have examined different mechanisms to ensure quality labels in crowdsourcing environments (both paid and volunteer) [48], [49].

J. Choice of Statistical Model

The choice of model will be driven by a number of factors including model speed, accuracy, interpretability, simplicity, and problem complexity. Modern linear maximum margin classifiers [50] and their kernel based extensions have proven very popular in a variety of domains. The feature selection and ensemble nature of classifiers such as Random Forests [51] make them more robust to the presence of noise. Fully Bayesian probabilistic models with prior distributions are useful in cases when there is a lack of training data. Gaussian process models for regression are particularly useful as they give uncertainty estimates [52]. Moreover, the model is in many cases dictated by the hypothesis. For example if one has a hypothesis about the rate of bird migration from one region to another, it may be desirable for the chosen model to explicitly describe this process and have parameters which are interpretable by the scientists. For scientists without a background in statistical modeling, this step can be particularly mysterious, and thus making the best modeling choice for their problem can be difficult.

K. Model Optimization

Once the data has been labeled, its features extracted, and the statistical model chosen, it is then necessary to optimize the model parameters to best fit/explain the data. This step can also be referred to as model training. The choice of model will dictate the complexity of this step. Simple linear discriminative models can be much quicker to optimize than fully generative ones. In practice, there may be several different algorithms available to perform the optimization and the choice of optimization scheme will depend on the structure of the problem. Conditioned on the type of model, it may be possible to use specialized hardware such as computing clusters or GPUs to speed up the training. Modern developments in machine learning and Bayesian inference have decoupled the problem of parameter optimization, inference, from the model specification. This allows one to freely design an arbitrarily complex model without worrying about parameter estimation and inference. Recent advances in Markov Chain Monte Carlo [53], [54], allow for efficient, exact, inference of complex models. Also approximate inference methods such as Expectation Propagation [55] and Variational approximations [56] can be used when one wants to trade accuracy for computational time. Such methods have been developed into software packages such as STAN [57], Infer.NET [58] and OpenBUGS [59] which are available for scientists to use. However, some understanding of probabilistic modeling is still required.

L. Model Validation

It can be easy to overfit the model to the training data, resulting in poor predictive power on the unseen test set. As a result, it is essential to estimate the generalizability of the trained model. Assessing the model's quality can be performed in a variety of ways. The standard approach is to choose an appropriate error measure for the task at hand, *e.g.* mean squared error for regression, and evaluate the model on a hold out test set. This test set is a subset of data that has also been labeled but not used in training. An alternative is to perform cross validation by splitting the data into several potentially non-overlapping subsets and training (optimizing the model) on all the subsets but one, and then evaluating the model on the remaining subset. This 'leave-one-out' testing is then performed in turn for each subset. During optimization there may be several different parameter or hyper parameter settings for the model. The model with the best test performance is chosen by comparing the different parameter settings.

M. Model Use

The remainder of the unlabelled data can now be evaluated using the trained model. This step can be time consuming if there is a large quantity of data to process. There may also be speed/quality tradeoffs that need to be made. The accuracy of the model might have to be compromised if it is required to run in real-time in the case of detection in audio or video. These requirements can sometimes also affect the choice of the model from the previous step.

N. Hypothesis Testing

Now that the entire collected dataset has been labelled, both manually by the scientist and automatically by the model, it is possible to test the initial hypothesis. This may be achieved in a frequentist manner or, more increasingly, by Bayesian means [60].

O. Publicizing of Results and Method

If the initial hypothesis is deemed to be of interest to the community at large, the final step is to publish the findings. In a subset of cases the data and code are also made available. Recently we have seen the creation of peer-reviewed scientific data journals whose sole purpose is to archive both the data and its associated meta-data. Code can be archived on public repositories which have mechanisms to allow other users to expand it and address any bugs or performance issues.

III. Current Research Efforts

For an overview of which data-understanding pipeline components are addressed by the pattern recognition community, we performed an analysis on papers from ICPR 2012. A similar survey focused only on dataset use and interpretation at ICML 2011 was conducted by Wagstaff [15].

Of the 301 regular papers presented at ICPR 2012, we categorized them based on the pipeline described previously. Of these papers, 149 contained (either directly or indirectly)

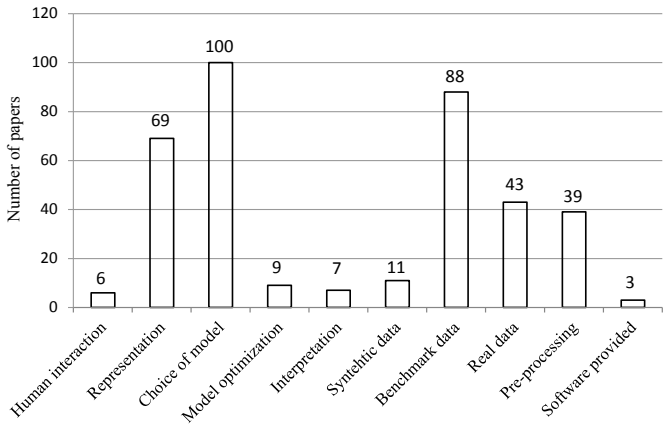

Fig. 2. Counts of the number of papers from ICPR 2012 featuring components from the data understanding pipeline. Of the 301 regular papers, 149 featured at least one step from our pipeline.

with at least one of our pipeline steps. We categorized these 149 papers into the following non-mutually exclusive categories. 1) Human interaction. Papers that involve humans in the labeling or learning process. 2) Representation. Papers where the proposed methodology is related to finding representations for different types of data. Examples in this category include sparse coding, dimensionality reduction, and feature selection. 3) Model choice. Papers where a new model is proposed or a known model is applied to a new domain. Examples in this category include classification papers, and tracking algorithms. 4) Model optimisation. Papers that propose new optimisation methods, such as ant colony optimisation, and approximate algorithms for estimating parameters of known models. 5) Interpretation. Papers which describe a methodology with good interpretation capabilities that is also demonstrated through experimentation. 6) Synthetic data. Papers where synthetic data is used in the experiments. 7) Benchmark data. Papers which use available benchmark data from public repositories such as UCI [61]. 8) Real data. Papers that use real data. 9) Preprocessing. Papers that fit into the data preprocessing step of the pipeline. Examples include pre-segmentation, filtering, and denoising. 10) Software provided. Papers that include a link or mention that code for the proposed methodology is provided by the authors.

Figure 2 summarizes our findings. We can see that many of the papers are concerned with feature representation, model choice, and evaluation on benchmark data – the standard machine learning pipeline. Only a limited number of papers consider human interaction as part of this process. In terms of evaluation, benchmark datasets feature heavily, while real data is used less frequently. A very small amount of papers explicitly state that their code will be made available.

IV. Open Problems

From our analysis in the previous sections, we now highlight several components that are missing or overlooked in the current data acquisition and understanding pipeline. By viewing the pipeline holistically, as opposed to each step in isolation, we hope to direct attention to these underexplored areas. Where relevant, we point to existing related work.

A. Users

Not all annotations are equal
Active learning aims to reduce the number of annotations the user has to provide when training a model. We know that not all data points contribute the same amount of information [62], [63]. The fact that not all annotation types are equal is less explored. Some annotations are more time consuming to provide, in terms of both physical and cognitive effort. In the case of image segmentation, it is easier for the labeler to provide an image level tag as opposed to an individual label for each pixel [64].

Not all annotators are equal
Some annotators will have more domain specific knowledge than others. When crowdsourcing annotations, it may be more difficult and more expensive to find expert annotators. It would be beneficial to ask the 'easier' questions of the non-expert labelers and save the 'difficult' questions for the experts.

Annotators are not always right
Real world problems are challenging and the correct answer is often ambiguous. Allowing the user to express their degree of certainty for a given annotation may lead to better uncertainty estimates in the final model. Users have also been shown to be inconsistent when relabeling the same dataset after a period of time [18]. Modeling the annotators ability over time may also improve our understanding of when they should, and should not, be trusted [43].

Annotation concepts evolve
Concept evolution refers to the inconsistency of human labelers when presented with borderline or ambiguous cases, or when they are still determining how the different data classes should be defined [18]. For scientists that are investigating new phenomena, concept evolution can be a significant issue. New types of interface mechanisms have demonstrated ways to mitigate this problem [18], but other biases can creep into the labeling process, especially for exploratory research, reducing the ability to efficiently train a reliable system.

Accuracy is not the only measure of success
Models are typically evaluated by their accuracy at test time. This analysis ignores the amount of supervision required by the user to get to a given level of accuracy. For interactive machine learning, a measure of human effort is also of importance.

B. Models

Real data is unbalanced
Often, real world data contains heavily imbalanced classes. A camera trap may be capturing images for many weeks before a species of interest appears. The annotator can quickly correct false positives. However, false negatives can be very hard to find. Also, in many situations, the number of classes may not be known in advance. Here, methods for rare class discovery and anomalous event detection are necessary [65], [66].

Problems can be related
Many labeling tasks are related. Models trained for one scenario may be very relevant for another [67]. Exploiting this relationship will save the user time.

Model complexity is important
Real data capture and processing requirements place many design constraints on a given model. Real-time systems have to return a result within a given time budget. For deployed systems, low power consumption may be critical.

C. Interfaces and Visualization

Users need help understanding model parameters
Understanding the effects of different model parameters is difficult without detailed knowledge of how the model works. Users need higher level controls to allow them to tune their models without understanding the details *e.g.* models that report confidence in their outputs [68], [69].

Combining models can help
Combining the output of different models is a conceptually simple, yet powerful tool for increasing performance in many tasks [70], [71]. Carefully designed interfaces can help users combine several possibly complementary models without needing to understand model specific details [72].

Feature selection is challenging
Choosing the best representation for a problem is difficult. Presently this is a one time decision based on intuition. Interfaces could be designed that allow the user to interactively explore the contribution of different features [73].

High dimensional data needs to be summarized
In the data exploration phase, it is very useful to have a quick overview of the types of variation and structure present in the data. Two dimensional projections of the input can useful when the data is easily separable [74]. However, for more complex signals, alternative visualizations specific to the data modality will be useful [75].

V. Recommendations

Based upon observations conducted during collaborations between researchers in ecology, statistical machine learning, computer vision, and human-computer interaction we outline several recommendations for each discipline involved. This list, while not exhaustive, lays out a set of best practices to enhance collaboration between the different communities. Our experience suggests that even small steps in these directions significantly improves a good idea's impact and citations.

A. Domain Experts

- Make datasets available under liberal licenses
- Richly annotate these datasets keeping all information such as annotation timings
- Invest in the computational resources capable of dealing with large quantities of data
- Encourage development of programming and modeling skills

B. Machine Learning and Statistics

- View the data understanding pipeline in its entirety, with the scientist at its core
- Make algorithms more available - release source code and create tools
- Seek collaborations with different communities
- Reduce jargon to make work more accessible

C. Interface Design and Visualization

- Build labeling interfaces with the human as the core component
- Ensure these tools scale to large quantities of data
- Interact with and incorporate the state of the art in machine learning
- Explore different visualizations for both data and models

VI. Conclusion

Open questions in biodiversity are extremely challenging and have potential for real scientific and human impact. The field produces complex datasets with associated real world problems such as noise and class imbalance. Currently, there exists a gap between cutting edge research in statistical machine learning and the tools that are available to practitioners. We believe that great scientific progress will be achieved by closing this gap and by placing more powerful tools into the hands of domain experts. Targeting real scientific datasets, and the scientists themselves in their respective fields, will drive innovation in pattern recognition, image processing, computer vision, and machine learning. These problems will not be solved by algorithmic advancements alone, but instead necessitate new interactive machine learning research with the scientists' needs placed firmly at the core.

Acknowledgment

Funding for this research was provided by EPSRC grant EP/K015664/1.

References

[1] G. M. Mace, K. Norris, and A. H. Fitter, "Biodiversity and ecosystem services: a multilayered relationship," *Trends in ecology & evolution*, 2012.

[2] I. J. Bateman, A. R. Harwood, G. M. Mace, R. T. Watson, D. J. Abson, B. Andrews, A. Binner, A. Crowe, B. H. Day, S. Dugdale *et al.*, "Bringing ecosystem services into economic decision-making: land use in the united kingdom," *science*, 2013.

[3] S. H. Butchart, M. Walpole, B. Collen, A. van Strien, J. P. Scharlemann, R. E. Almond, J. E. Baillie, B. Bomhard, C. Brown, J. Bruno *et al.*, "Global biodiversity: indicators of recent declines," *Science*, 2010.

[4] C. Bellard, C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp, "Impacts of climate change on the future of biodiversity," *Ecology letters*, 2012.

[5] J. M. Rowcliffe and C. Carbone, "Surveys using camera traps: are we looking to a brighter future?" *Animal Conservation*, 2008.

[6] K. E. Jones, J. A. Russ, A.-T. Bashta, Z. Bilhari, C. Catto, I. Csősz, A. Gorbachev, P. Győrfi, A. Hughes, I. Ivashkiv *et al.*, "Indicator bats program: a system for the global acoustic monitoring of bats," *Biodiversity Monitoring and Conservation: Bridging the Gap between Global Commitment and Local Action*, 2013.

[7] "Bat Detective," 2012. [Online]. Available: www.batdetective.org

[8] "Snapshot Serengeti," 2012. [Online]. Available: www.snapshotserengeti.org

[9] J. S. Kim, M. J. Greene, A. Zlateski, K. Lee, M. Richardson, S. C. Turaga, M. Purcaro, M. Balkam, A. Robinson, B. F. Behabadi, M. Campos, W. Denk, H. S. Seung, and the EyeWirers, "Space-time wiring specificity supports direction selectivity in the retina," *Nature*, 2014.

[10] T. Desell, R. Bergman, K. Goehner, R. Marsh, R. VanderClute, and S. Ellis-Felege, "Wildlife@ home: Combining crowd sourcing and volunteer computing to analyze avian nesting video," in *International Conference on eScience*, 2013.

[11] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *IJCV*, 2013.

[12] A. Fielding, *Machine learning methods for ecological applications*. Springer, 1999.

[13] F. Recknagel, "Applications of machine learning to ecological modelling," *Ecological Modelling*, 2001.

[14] C. L. Walters, R. Freeman, A. Collen, C. Dietz, M. Brock Fenton, G. Jones, M. K. Obrist, S. J. Puechmaille, T. Sattler, B. M. Siemers *et al.*, "A continental-scale tool for acoustic identification of european bats," *Journal of Applied Ecology*, 2012.

[15] K. Wagstaff, "Machine learning that matters," in *ICML*, 2012.

[16] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[17] J. A. Fails and D. R. Olsen Jr, "Interactive machine learning," in *International Conference on Intelligent User Interfaces*, 2003.

[18] T. Kulesza, S. Amershi, R. Caruana, D. Fisher, and D. Charles, "Structured labeling for facilitating concept evolution in machine learning," in *CHI*, 2014.

[19] "ZSL Instant Wild," 2011. [Online]. Available: www.edgeofexistence.org/instantwild

[20] T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, 2008.

[21] C. L. Walters, A. Collen, T. Lucas, K. Mroz, C. A. Sayer, and K. E. Jones, "Challenges of using bioacoustics to globally monitor bats," in *Bat Evolution, Ecology, and Conservation*, 2013.

[22] X. P. Burgos-Artizzu, P. Dollár, D. Lin, D. J. Anderson, and P. Perona, "Social behavior recognition in continuous video," in *CVPR*, 2012.

[23] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *CVPR*, 2003.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[26] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[27] D. Oneata, J. Verbeek, and C. Schmid, "Action and event recognition with fisher vectors on a compact feature set," in *ICCV*, 2013.

[28] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005.

[29] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing*, 1980.

[30] V. Stathopoulos, V. Zamora-Gutierrez, K. Jones, and M. Girolami, "Bat call identification with gaussian process multinomial probit regression and a dynamic time warping kernel," *AIStats*, 2014.

[31] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, 1980.

[32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 1998.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." in *NIPS*, 2012.

[34] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing*, 2012.

[35] M. Kabra, A. A. Robie, M. Rivera-Alba, S. Branson, and K. Branson, "JAABA: interactive machine learning for automatic annotation of animal behavior," *Nature Methods*, 2013.

[36] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *IJCV*, 2008.

[37] A. Yao, J. Gall, C. Leistner, and L. Van Gool, "Interactive object detection," in *CVPR*, 2012.

[38] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, 2009.

[39] C. Vondrick and D. Ramanan, "Video annotation and tracking with active learning." in *NIPS*, 2011.

[40] S. Bandla and K. Grauman, "Active learning of an action detector from untrimmed videos," in *ICCV*, 2013.

[41] "Amazon Mechanical Turk," 2014. [Online]. Available: www.mturk.com

[42] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds." in *NIPS*, 2010.

[43] V. C. Raykar, S. Yu, L. H. Zhao, A. Jerebko, C. Florin, G. H. Valadez, L. Bogoni, and L. Moy, "Supervised learning from multiple experts: whom to trust when everyone lies a bit," in *ICML*, 2009.

[44] B. Settles, *Active learning*. Morgan & Claypool, 2012.

[45] O. Mac Aodha, N. D. Campbell, J. Kautz, and G. J. Brostow, "Hierarchical subquery evaluation for active learning on a graph," *CVPR*, 2014.

[46] A. Vezhnevets, J. M. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *CVPR*, 2012.

[47] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning: Training object detectors with crawled data and crowds," in *CVPR*, 2011.

[48] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *JMLR*, 2010.

[49] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 2010.

[50] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, 1995.

[51] L. Breiman, "Random forests," *Machine learning*, 2001.

[52] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[53] M. Neal, Radford, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. of Computer Science, University of Toronto, Tech. Rep. CRG-TR-93-1, 1993.

[54] M. Girolami and B. Calderhead, "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2011.

[55] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001.

[56] H. Attias, "A variational bayesian framework for graphical models," in *NIPS*, 2000.

[57] Stan Development Team, "Stan: A c++ library for probability and sampling, version 2.2," 2014. [Online]. Available: http://mc-stan.org/

[58] T. Minka, J. Winn, J. Guiver, and D. Knowles, "Infer.NET 2.5," 2012, Microsoft Research Cambridge. http://research.microsoft.com/infernet.

[59] L. M. Surhone, M. T. Tennoe, and S. F. Henssonow, *OpenBUGS*. Mauritius: Betascript Publishing, 2010.

[60] J. Kruschke, *Doing Bayesian data analysis: a tutorial introduction with R*. Academic Press, 2010.

[61] A. Asuncion and D. Newman, "Uci machine learning repository," 2007.

[62] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do we need more training data or better models for object detection?." in *BMVC*, 2012.

[63] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba, "Are all training examples equally valuable?" *arXiv preprint arXiv:1311.6510*, 2013.

[64] S. Vijayanarasimhan and K. Grauman, "Cost-sensitive active visual category learning," *IJCV*, 2011.

[65] P. Vatturi and W.-K. Wong, "Category detection using hierarchical mean shift," in *KDD*, 2009.

[66] T. S. Haines and T. Xiang, "Active rare class discovery and classification using dirichlet processes," *IJCV*, 2014.

[67] S. J. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering*, 2010.

[68] O. Mac Aodha, A. Humayun, M. Pollefeys, and G. J. Brostow, "Learning a confidence measure for optical flow," *PAMI*, 2013.

[69] M. Reynolds, J. Dobos, L. Peel, T. Weyrich, and G. J. Brostow, "Capturing time-of-flight data with confidence," in *CVPR*, 2011.

[70] O. Mac Aodha and G. J. Brostow, "Revisiting Example Dependent Cost-Sensitive Learning with Decision Trees," in *ICCV*, 2013.

[71] C. García Cifuentes, M. Sturzel, F. Jurie, and G. J. Brostow, "Motion models that only work sometimes," in *BMVC*, 2012.

[72] J. Talbot, B. Lee, A. Kapoor, and D. S. Tan, "Ensemblematrix: interactive visualization to support machine learning with multiple classifiers," in *CHI*, 2009.

[73] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer, "Guiding feature subset selection with an interactive visualization," in *Visual Analytics Science and Technology*, 2011.

[74] L. Van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *JMLR*, 2008.

[75] C. Nguyen, Y. Niu, and F. Liu, "Video summagator: an interface for video summarization and navigation," in *CHI*, 2012.

Fig. 1. Overview of the main scientific data understanding pipeline. For each stage we give a short description and a practical example.