Freudian Slips: Analysing the Internal Representations of a Neural Network from its Mistakes

Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini

Intelligent Systems Laboratory, University of Bristol, Bristol, United Kingdom

Abstract. The use of deep networks has improved the state of the art in various domains of AI, making practical applications possible. At the same time, there are increasing calls to make learning systems more transparent and explainable, due to concerns that they might develop biases in their internal representations that might lead to unintended discrimination, when applied to sensitive personal decisions. The use of vast subsymbolic distributed representations has made this task very difficult. We suggest that we can learn a lot about the biases and the internal representations of a deep network without having to unravel its connections, but by adopting the old psychological approach of analysing its slips of the tongue. We demonstrate in a practical example that an analysis of the confusion matrix can reveal that a CNN has represented a biological task in a way that reflects our understanding of taxonomy, inferring more structure than it was requested to by the training algorithm. In particular, we show how a CNN trained to recognise animal families, contains also higher order information about taxa such as the superfamily, parvorder, suborder and order for example. We speculate that various forms of psycho-metric testing for neural networks might provide us insight about their inner workings.

Keywords: deep learning, taxonomy, computer vision, explainable AI, black-box testing

1 Introduction

Deep neural networks deliver state of the art performance in different areas of AI [14, 21, 23], particularly in computer vision, and promise to be deployed in many further domains [19]. However, for all their convenience, they do attract the criticism that they operate as black boxes [1]: that they can only pick up correlations, with no regard for causality or other theoretical frameworks that humans would consider more explainable. This criticism is often summarised as "correlation trumps causation", based on the observation that there is no clear way to interpret the internal configurations of weights learnt by the network, and that they are trained to perform a specific prediction task, and not directly rewarded for developing a higher level understanding of the problem at hand.

This criticism is not necessarily true though, as there are many reasons to believe that our own theoretical frameworks respond (also) to criteria of economy, and therefore that black box machine learning algorithms might find it useful to represent data in the same way [22].

Indeed, it has been known for a long time [5] that there is no real reason for a neural network to prefer an elegant representation of reality that can capture some higher level understanding of the problem, being trained only to perform correct predictions. Yet we, as humans, tend to prefer simpler and structured representations, often invoking Occam's razor.

This drives at a problem that has been considered for many years, of making AI explainable, but which has recently found new urgency, amid concerns that machines are going to soon be making decisions about us that we will unable to understand and for which they can offer no explanation [8, 10, 11]. This problem goes to the heart of the old question of how can we interpret the inner representations of reality that are inferred by a neural network, as a way to understand if it contains any unintended biases?

The direct reaction of the engineer has always been that of opening up the network and tackling the mess of connections and neurons [18], much like a surgeon performing brain surgery (a phrase sometimes applied to deep neural networks [12]). However, there is another way to reveal information about the internal structure of these networks, loosely analogous to that of Sigmund Freud, who was trying to understand the internal world view of a black box not by its successes (which after all are what it was trained for) but by its errors and mistakes. Far from being random, these errors resulting from machine learning tasks, much like slips of the tongue, might reveal hidden biases and aspects of how the neural network represents the world internally. A systematic study of the errors made by deep–networks might reveal useful information about their hidden biases and assumptions without the need to unravel the role of each internal connection.

We suggest these "Freudian slips" may offer a promising alternative to other black-box approaches [3, 16] to analysing the internal organisation of a network without surgery, and perform a first experiment to demonstrate the method. Our analysis of errors shows that a deep convolutional neural network can learn more structure than it is specifically asked to.

As a testing ground, we selected the domain of biology, where there is already a fundamental framework, that of evolution and therefore of taxonomy, in which we can assess whether the network is learning a similar representation of the world to humans. For instance, since animal species are not uncorrelated but have evolved over time from one another, there could be benefits for a neural network to learn more than just its minimal task of recognising different taxa, but to internally represent information about the phylogenetic taxonomic structure. For instance, this can be seen in the way that phylogenetic taxonomies are inferred from data using maximum parsimony [7].

We start with the task of teaching a network to assign images of mammals curated from the web to one of 54 animal families, where we have 540,000 images

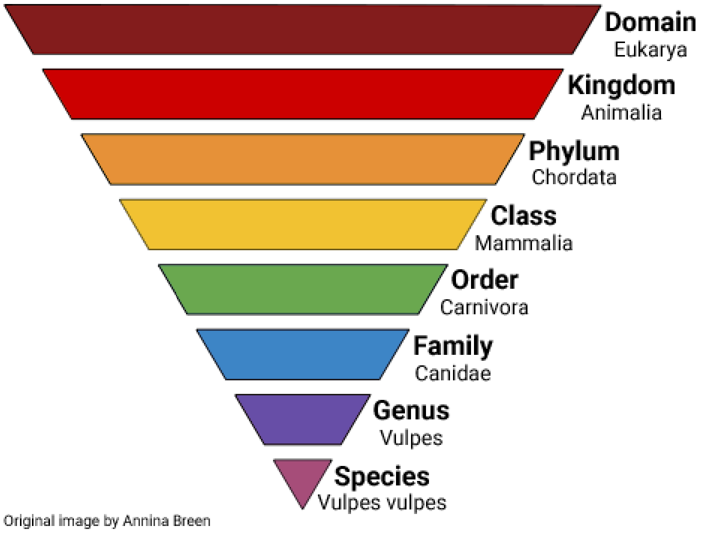


Fig. 1: Example hierarchy of the major taxa in the NCBI taxonomy of living species. Image by Annina Breen (Own work) [CC BY-SA 4.0], via Wikimedia Commons.

of individual animals, belonging to 3585 species, organised in 54 distinct families, themselves organised into 27 orders, all within the Mammalian class. We train a deep convolutional neural network to recognise the 54 families, an interesting task in and of itself [2, 9], and perform an analysis on the errors it makes in order to learn about its internal representations.

Of course, there is still not a systematic framework to connect the "slips of the tongue" made by a CNN with its internal representations, but we claim that this might be worth developing, as we look for ways to make these systems more transparent, in the face of their inherently subsymbolic nature.

2 Methodology

2.1 Taxonomic Identification of Animals

Animal species are organised according to a standardised taxonomic system that divides them into seven major taxa: species, genera, families, orders, classes, phylums and kingdoms, as shown in Fig. 1, along with further subdivision into minor taxa (i.e. superfamilies, suborders, superorders, subclasses and so on). Using this taxonomy, we can define a distance between any two animals based upon how far apart they are in the phylogenetic tree, where we represent each node as a minor taxon. For example, if we were to only consider the seven major taxa, the distance between two carnivores, such as a Red Fox (*Vulpes vulpes*) and a Red Panda (*Ailurus fulgens*), would be at the Order level, giving them a distance of 3 (Species → Genus → Family → Order).

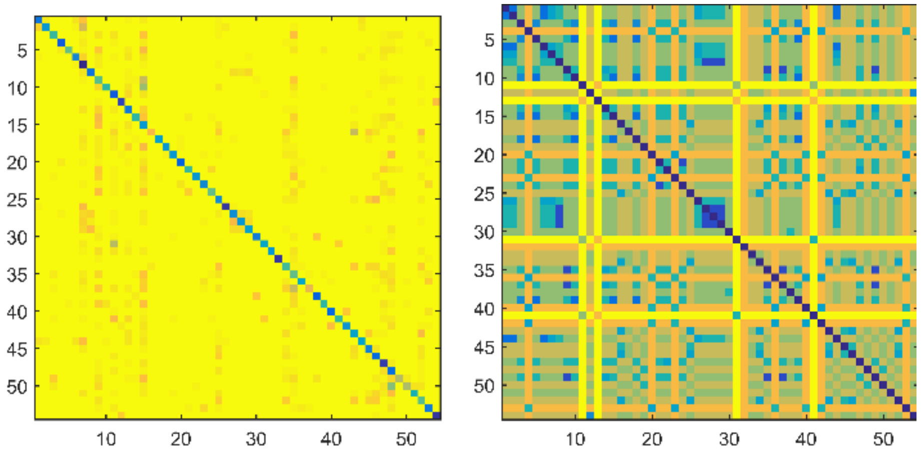
Table 1: List of the 54 Mammalian families, along with their common names used in the study.

Family name	Common name example	Family name	Common name example
Aotidae	Night monkeys	Lemuridae	Ring-tailed lemurs
Atelidae	New world monkeys	Lepilemuridae	Sportive lemurs
Bathyergidae	Mole-rats	Leporidae	Rabbits and Hares
Canidae	Dogs	Macropodidae	Kangaroos
Caviidae	Guinea pigs	Macroscelididae	Elephant shrews
Cebidae	Capuchin monkeys	Manidae	Pangolins
Cercopithecidae	Baboons	Molossidae	Free-tailed bats
Cheirogaleidae	Dwarf lemurs	Muridae	Mice
Cricetidae	Hamsters	Mustelidae	Otters
Ctenomyidae	Tuco-tuco	Nesomyidae	Climbing mice
Dasyuridae	Quoll	Ochotonidae	Pika
Delphinidae	Dolphins	Octodontidae	Rock rats
Didelphidae	Opossums	Otariidae	Eared seals
Dipodidae	Jerboa	Phalangeridae	Cuscus
Echimyidae	Spiny rats	Phocidae	Earless seals
Emballonuridae	Sac-winged bats	Phyllostomidae	Leaf-nosed bats
Equidae	Horses	Pitheciidae	Titis
Erethizontidae	New world porcupines	Pteropodidae	Fruit bats
Erinaceidae	Hedgehogs	Rhinolophidae	Horseshoe bats
Felidae	Cats	Sciuridae	Squirrels
Geomyidae	Gophers	Soricidae	Shrews
Gliridae	Dormice	Spalacidae	Bamboo rats
Herpestidae	Mongeese	Talpidae	Moles
Heteromyidae	Kangaroo rats	Tupaiidae	Treeshrews
Hipposideridae	Old world bats	Vespertilionidae	Common bats
Hylobatidae	Gibbons	Viverridae	Civets
Indriidae	Lemurs	Ziphiidae	Beaked whales

This builds on recent work that has so far been done to assign the image of an animal to its correct species [2, 9], while we focus here on the class of mammalians, and attempt to correctly assign each image of an animal to its appropriate family classification.

2.2 Data Curation

We aimed to select a number of mammalian families from the National Center for Biotechnology Information (NCBI) taxonomy [6] to obtain a representative, but broad coverage of all the species belonging to the mammalian class. Within the mammalian taxonomic tree, we found a total of 27 nodes at the Order level, and 140 nodes at the Family level. After filtering the species names to remove those containing special characters ('.' or '/') or with additional information, we compiled a list of every mammalian family containing at least 15 different



(a) CNN distance matrix (b) Phylogenetic distance matrix

Fig. 2: Comparison of the distance matrix calculated from the CNN confusion matrix and the phylogenetic distance matrix computed from the NCBI taxonomy.

species. Table 1 lists each of the 54 families resulting from this process, with an example of the common name for some of the animals that belong to the family.

Each family was additionally annotated with all of its child species from the NCBI taxonomy, along with the common names for the species, obtained by taking the title of the Wikipedia page after redirection from `https://en.wikipedia.org/wiki/<species_name>`. For example, the page for *Vulpes vulpes* (`https://en.wikipedia.org/wiki/Vulpes_vulpes`) redirects to the page entitled "Red fox".

Using the list of 3585 species and their common names for each of the 54 family categories, we queried for images from a popular search engine, retrieving an average of 98 images per query. We split the data for training and testing by assigning 100 randomly selected images per family to the hold-out test set. Following this, we performed data augmentation for any families which did not contain a minimum of 10,000 images in the training set. This included using a common technique in computer vision [15, 17] of horizontally flipping and rotating images by [-15,-10,-5,5,10,15] degrees in order to increase the number available images for training a classifier. In total, the "Family look" dataset contains 540,000 training and 5,400 test images, labelled by family. This dataset of labelled images, along with the NCBI taxonomy used, are available online at `http://thinkbig.enm.bris.ac.uk/family-look`

2.3 Learning a "Family Look" Classifier

Following the pre-processing steps in [13, 17], each image in the "Family look" dataset was resized so the shortest length became 256 pixels. We took a center crop of size 256 × 256 from each image, ensuring the images were of a consistent size, before randomly cropping the images into 224 × 224 pixel patches for the training set, or taking the center 224 × 224 pixel patch for testing. Using the 18-layer ResNet CNN architecture proposed in [13] for object recognition and used in [9] for animal species identification[1], we trained the network on the 540,000 animal training images. The CNN model was trained from scratch using Stochastic Gradient Descent. We trained for 50 epochs using a mini-batch size of 64, an initial learning rate of 0.1 that decreased every 20 epochs by a factor of 0.1, along with a moment and weight decay being applied of 0.9 and 0.0001 respectively.

After training the CNN, we evaluated our model on the held-out test set of 5,400 test images, computing both an overall accuracy for the model and a classifier confusion matrix detailing specifically which family categories the model confused with one another for further analysis.

2.4 Constructing a Tree Representation

Once we had a trained classifier for discriminating between the different family categories in animal images, we wished to construct a tree representation of the mistakes made by the model. To construct a tree from the confusion matrix, we first needed to convert the matrix from a similarity matrix to a distance matrix, performed by subtracting each value in the confusion matrix from the maximal value. This distance matrix was then used to generate a tree of the mistakes by performing a furthest neighbour agglomerative hierarchical clustering [4], implemented using the `seqlinkage` command with the complete linkage option from the Bioinformatics toolbox [20]. The resulting tree represents each family category as a leaf node in the tree, with each internal node representing a cluster of animal families based upon how often they are mistaken for each other. In doing so, we can compare the categories which the model confused with each other with the taxonomic tree coming from the NCBI taxonomy [6] based upon taxonomic distance.

3 Results

In this experiment, we found that our deep CNN model could correctly classify the animals in our test set at the family level 53.22% of the time, well above the baseline of 1.85% one could trivially expect for a 54 category classification task. More interestingly, we find that there is a significant correlation ($\rho = 0.53$, $p < 0.0001$) between the family-similarities (as measured by error probability between them) and the taxonomic tree distance as indicated by the NCBI taxonomy [6], and shown in Fig. 2.

[1] The ResNet CNN used in their paper had the same architecture, but was much deeper. We trained a similar 152-layer network to [9] but found no clear difference with our 18-layer model, which aimed to strike a better balance between the depth of the network and the associated computational load.

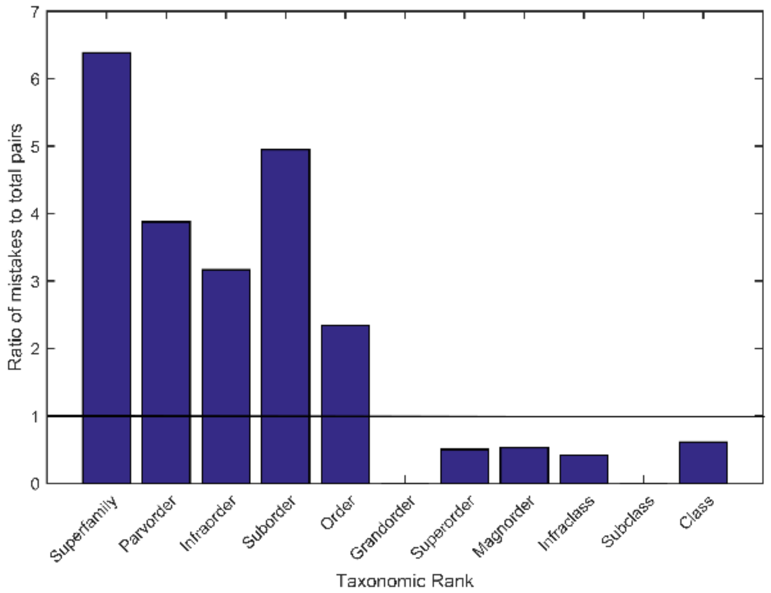


Fig. 3: Histogram showing the ratio of mistakes to the total pairs of animals at each phylogenetic distance. Values above 1 mean more mistakes were made at this taxonomic rank that we should expect, with values below 1 meaning less mistakes were made than expected.

Considering this in terms of our Freudian slip concept, there was no reason for the network to learn any relation among these categories, and the expectation would be that there should be no correlation, unless the network learns weights corresponding with quantities that correlate at a higher taxon level than that of the family taxon, something that was not taught to it nor available from the data it has seen.

Probing deeper into the types of mistakes that the classifier made, we analysed how the distribution of errors made by the Neural Network over the 11 possible types of errors compared what would be expected if we were to assign the categories to the images uniformly at random. The ratio between the actual number of errors made by the network, and that expected under this null-model, is represented in Fig. 3, where values below 1 show that the network makes less mistakes than expected by chance, with those above 1 indicating the opposite.

We can immediately see that the classifier is not making mistakes uniformly across the 54 categories, but is making many more mistakes between animals sharing a taxonomic rank of Order or below, and making less mistakes than we should expect between animals sharing a taxonomic rank above that of the Order rank. With the notable exception of Suborder, we can also see that as the phylogenetic distance increase, we are less likely to confuse two animals

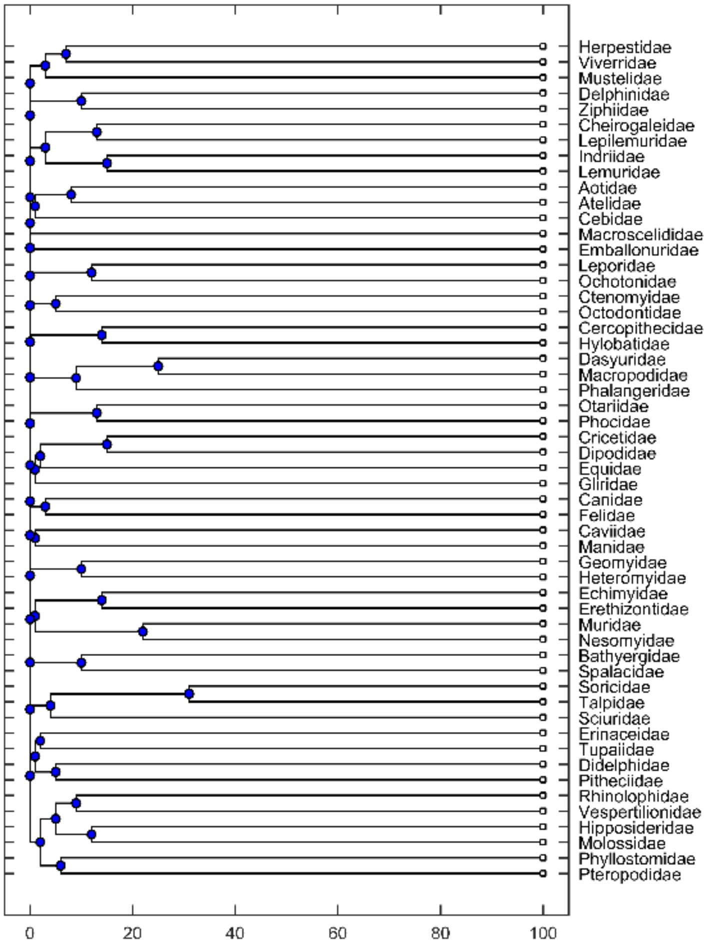


Fig. 4: Family tree built from the confusion matrix of the neural network showing the mammalian families which are most often mistaken for one another. We can see that this reflects the actual phylogenetic tree (Fig. 5) to some extend, mostly for relationships between animals below the Order rank.

with each other, suggesting that the network is indeed learning some internal representation of the phylogenetic taxonomy.

Finally, examining the tree built from the confusion matrix (Fig. 4) and comparing this with the phylogenetic tree (Fig. 5), we can see that there is some taxonomic structure reflected in the mistakes of the network. For instance, we can see that the different families of microbats (Rhinolophidae, Vespertilionidae, Hipposideridae, Molossidae, Phyllostomidae) are grouped together with the megabat family (Pteropodidae). Similarly, the phylogenetic tree structure for the

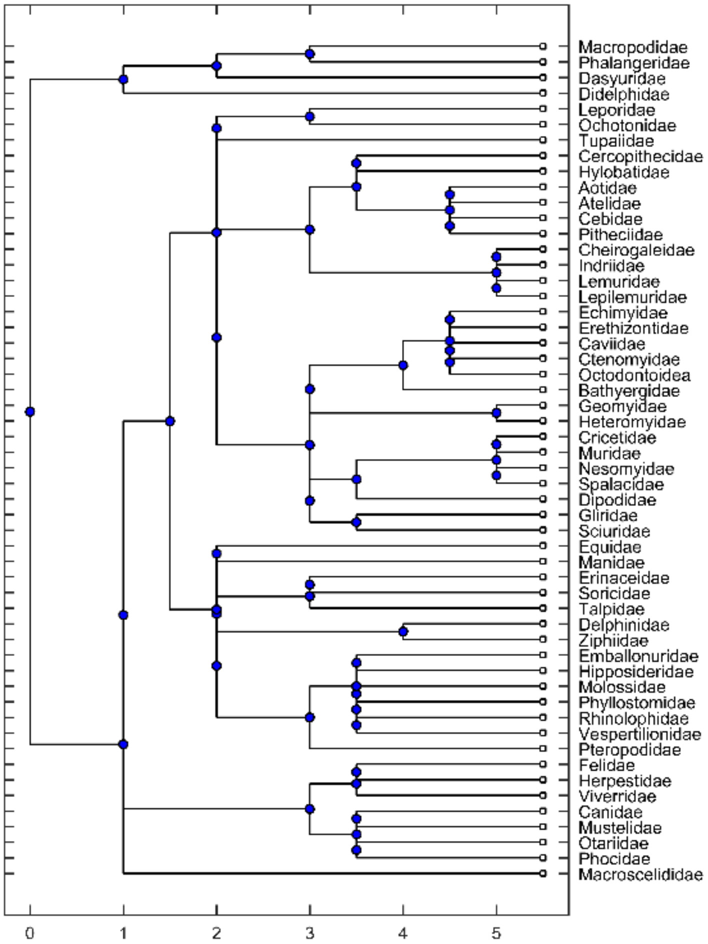


Fig. 5: Phylogenetic tree showing the taxonomic relationship between different mammalian families as recorded in the NCBI taxonomy [6]. Each level in the phylogenetic tree indicates a different major or minor taxon shared between animal families in the taxonomy.

families of lemurs (Cheirogaleidae, Lepilemuridae, Indriidae, Lemuridae), seals (Phocidae, Otariidae), and marsupials native to Australia (Dasyuridae, Macropodidae, Phalangeridae) are also represented. While we can find many examples of this type, it is mostly only up to the Order rank, with no higher level taxonomic structure.

4 Discussion

While the question of understanding the internal representations in neural networks remains an important one, and it will probably be addressed by both surgery and theory, we think that our approach provides a simple way to examine the internal work representation of the network, just like the method of Freudian slips attempted to recognise internal structures based on the errors that a person made while speaking.

By no means do we think that this experiment settles the problem, but more that it points to a different way to organise a search for interpretable AI, one that can respect the inherently sub-symbolic and distributed representations that have made deep-networks so useful, while also respecting our need to understand how the network represents its knowledge.

Our representation of the world is also "deep", in the sense that it relies on a hierarchy of theoretical concepts, and communication between people requires that those concepts are shared by both. While this may be an elusive task in everyday experience, it is more simply demonstrated in certain scientific domains, such as taxonomy. But the general point about trying to match the abstractions used by humans with those used by deep networks might be more general than taxonomies.

Once we start developing methods to test certain properties of the internal knowledge representation, we are in a domain akin to 'psycho-metrics', and there is a lot of expertise that might be shared from that field. It would be useful to involve philosophers of science and psychologists in the discussions about readable AI.

One possible side-effect of this approach could be that – as we learn how to make sure that the network respects at least some of the internal constraints that we value, in our representation of the world – we might even be able to add this to the cost function used in training.

5 References

1. Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
2. Guobin Chen, Tony X Han, Zhihai He, Roland Kays, and Tavis Forrester. Deep convolutional neural network based species recognition for wild animal monitoring. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 858–862. IEEE, 2014.
3. Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 598–617. IEEE, 2016.
4. William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.
5. John Denker, Daniel Schwartz, Ben Wittner, Sara Solla, Richard Howard, Lawrence Jackel, and John Hopfield. Large automatic learning, rule extraction, and generalization. *Complex systems*, 1(5):877–922, 1987.

6. Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2012.
7. Joseph Felsenstein. *Inferring phylogenies*, volume 2.
8. Alex A Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, 15(1):1–10, 2014.
9. Alexander Gómez, Augusto Salazar, and Francisco Vargas. Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. *arXiv preprint arXiv:1603.06169*, 2016.
10. Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.
11. David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)*, 2017.
12. B. Hassibi, D. G. Stork, and G. J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299 vol.1, 1993.
13. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
14. Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini. Gender classification by deep learning on millions of weakly labelled images. In *Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on*, pages 462–467. IEEE, 2016.
15. Heechul Jung, Sihaeng Lee, Sunjeong Park, Injae Lee, Chunghyun Ahn, and Junmo Kim. Deep temporal appearance-geometry network for facial expression recognition. *arXiv preprint arXiv:1503.01532*, 2015.
16. Josua Krause, Adam Perer, and Enrico Bertini. Using visual analytics to interpret predictive machine learning models. *arXiv preprint arXiv:1606.05685*, 2016.
17. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
18. Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8595–8598. IEEE, 2013.
19. Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 5 2015.
20. MathWorks. Bioinformatics toolbox: User's guide (r2017a). `www.mathworks.com/help/pdf_doc/bioinfo/bioinfo_ug.pdf`, 2017.
21. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
22. Jorma Rissanen. *Minimum description length principle.* Wiley Online Library, 1985.
23. David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.