Interrogating the linear variable importance of local explanations of black-box models with animated linear projections

1 Introduction

Mathematically rigorous approaches to predictive modeling are attributed to the least-squares method, over two centuries ago by Legendre and Gauss in 1805 and 1809, respectively. In 1886 Francis Galton coined the term *regression* to refer to continuous, quantitative predictions. While *classification* refers to discrete predictions as introduced by Fisher in 1936.

Breiman and Shmueli (Breiman 2001; Shmueli 2010) introduce the idea of distinguishing modeling based on its purpose; *explanatory* modeling is done for some inferential purpose such as hypothesis testing, while *predictive* modeling predicts new, out-of-sample, observations. This distinction draws attention to the divide between interpretable models and black-box models. In explanatory modeling, interpretability is key for drawing inferential conclusions. While predictive modeling may opt for potentially more accurate black-box models. The intended use has important implications for model selection and development.

Black-box models are becoming increasingly common, but not without their share of controversy and issues (O'neil 2016; Kodiyan 2019). Black-box models have been known to reflect common biases, including sex (Dastin 2018; Duffy 2019), race (Larson et al. 2016), and age (Díaz et al. 2018). Such issues occur when biases existent in the training data, the model picks up on this influence on the response variable, which is then built into the model. Another issue is data drift when new data is outside the support of latent or exogenous explanatory variables. Data drift can lead to worse predictions (Lazer et al. 2014; Salzberg 2014). Such cases highlight the need to make models fair, accountable, ethical, and transparent, which has led to the movement of XAI (Adadi and Berrada 2018; Arrieta et al. 2020).

One branch of XAI is local explanations, which take a variable attribution approach to bring transparency to a model. Local explanations attempt to approximate linear variable importance at the location of one observation. There are many such local explanations.

To illustrate our work, we apply the model-agnostic explanation SHAP (Strumbelj and Kononenko 2010; Štrumbelj and Kononenko 2014). The exact details of SHAP are tangent to the ideas of this work, but suffice it to say that SHAP approximates variable importance by taking the median importance over permutations of the explanatory variables. To be exact, we apply a variant that enjoys a lower computational complexity, known as tree SHAP (S. M. Lundberg, Erion, and Lee 2018).

In multivariate data visualization, a *tour* (Asimov 1985; Buja and Asimov 1986; S. Lee et al. 2021) is a sequence of linear projections of data onto a lower-dimensional space, typically 1-3D. Tours are viewed as an animation over small changes to the projection basis. Structure in a projection can then be explored visually to see which variables contribute to the formation of that structure. The intuition is similar to watching the shadow of a hidden 3D object change as the object is rotated; watching the structural shape of the shadow change gleans insight into the shape and features of the object.

There are various types of tours, which are distinguished by the generation of projection bases. In a *manual* tour (Cook and Buja 1997; Spyrison and Cook 2020) this path is defined by changing the contribution of a selected variable. Applying tours to models has been done in a couple of contexts. Specifically for exploring various statistical model fits and classification boundaries (Wickham, Cook, and Hofmann 2015), and using tree- and forest-based approaches as a projection pursuit index to generate a tour basis path (Y. D. Lee et al. 2013; Silva, Cook, and Lee 2021).

The approach proposed below is to use the radial, manual tour to interrogate a local explanation. After the

identifying an observation of interest, its explanation can be evaluated by testing the support of the structure identified by the explanation as the contributions of the variables are varied with the radial tour. We provide a free and open-source R package **cheem** with an interactive application to facilitate analysis. We give case studies of toy and modern datasets for both classification and regression tasks.

The change in the projection basis might feel similar to counterfactual, what-if analysis, such as *ceteris paribus* (Biecek 2020). Latin for "other things held constant" or "all else unchanged" is a counterfactual analysis showing how an observation's prediction would change from a change in one explanatory variable given that other variables are held constant. It ignores correlations of the variables and imagines a case that was not observed. In contrast, our approach is a geometric explanation of the factual, observed case by that varies contributions of the basis, essentially the orientation of the data object. Another difference is that the basis must maintain orthonormality. That is to say, when the contribution of one variable decreases, the contributions of others necessarily increase such that there is a complete component in that direction.

The remainder of this paper is organized as follows. The next section, Local explanation statistics, covers the background of the local explanation, SHAP, and the traditional visuals produced from it. Tours and the radial tour digs deeper into these animations of continuous linear projections. The section Application Design discusses the layout of the application, how it facilitates analysis, preprocessing, and package infrastructure. The section Case Studies illustrates several applications of this method. We conclude with a Discussion of the insights we draw from classification and regression tasks.

2 Local explanation statistics

Consider a highly non-linear model. At face value, it is hard to say which variable(s) are sensitive to the crossing of a classification boundary or identifies which variables caused an observation to have a relatively extreme residual. Local explanations shed light on these cases by approximating linear variable importance in the vicinity of one observation.

Figure 6 of Arrieta et al. (2020) gives comprehensive summarization of the taxonomy and literature of explanation techniques. This includes a large number of model-specific explanations such as deepLIFT, (Shrikumar et al. 2016; Shrikumar, Greenside, and Kundaje 2017) a popular recursive method for estimating importance in neural networks. There are a fewer number of model-agnostic explanations, of which LIME, (Ribeiro, Singh, and Guestrin 2016) SHAP, (S. Lundberg and Lee 2017) and their variants are popular.

These instance-level explanations are used in a variety of ways depending on the data. In images, saliency maps overlay or offset a heatmap indicating which pixels were important (Simonyan, Vedaldi, and Zisserman 2014). For instance, the presence of snow may be highlighted when distinguishing if a picture contains a wolf or husky. In text analysis, word-level contextual sentiment analysis can be used to highlight the sentiment and magnitude of influential words (Vanni et al. 2018). In the case of numeric regression, they are used to explain variable additive contributions from model intercept to the observation's prediction (Ribeiro, Singh, and Guestrin 2016).

2.1 SHAP and tree SHAP

SHaply Additive exPlanations (SHAP) approximates the variable importance in the vicinity of one observation by taking the median importance of a subset of permutations in the explanatory variables. This idea stems from the field of game theory, where Shapley (1953) devised a method to evaluate an individual's contribution to cooperative games by permuting the players that contribute to the score.

To illustrate SHAP and its original use, explaining the difference between the intercept and an observation's prediction, we use soccer data from FIFA 2020 season (Leone 2020). We have 5000 observations of nine skill measures (after aggregating highly correlated variables). A random forest model is fit to regress the log wages, in 2020 Euros, from the skill measures. We then extract the SHAP values of a star offensive player (L. Messi) and defensive player (V. van Dijk). We expect to see a difference in the attribution of the variable importance across the two positions of the players.

Figure 1 shows the SHAP values of these players. Panel a) shows these players receive a sizable difference in wages. Panel b) shows the underlying distribution of the SHAP attributions while permuting the explanatory variables, with the medians being the SHAP values. In the light of the player position, the difference in the variable importance makes sense; offensive and movement are more important for the offensive player, while defensive and power skills are more important to the model for explaining the prediction of the defensive player. We would likewise expect the profile of variable importance to be unique for star players of other positions, such as goalkeepers or middle fielders. Panel c) shows a simplified breakdown plot (Gosiewska and Biecek 2019), where a local explanation is used to additively explain the difference from the intercept to the observations prediction. Such additive approaches will show an asymmetry in the variable ordering, so we opt to fix the order to that of panel b), namely, by decreasing the sum of the SHAP values.

In summary, this highlights how local explanations bring interpretability to a model, at least in the vicinity of their observations. In this instance, we showed how two players with different positions receive different profiles of variable importance to explain the prediction of their wages. In the following section, we will be using normalized explanations as the starting projection basis to interrogate the explanation further.

3 Tours and the radial tour

TODO:XXX

4 Application design

Below we illustrate the two primary displays of the application: the global view and the tour view. Then we cover what we take away from the classification and regression tasks. Lastly, we discuss the preprocessing that before display.

4.1 Global view

The global view provides an essential context of all observations and allows exploration of the separability of the data- and attribution-spaces. While ultimately, its purpose is to facilitate the selection of a primary point and comparison point.

An approximation of these spaces is given as the first two principal components of their respective spaces. This is shown side-by-side next to model information, the prediction, and the observed response variable. The orientation and magnitude of the variables are inscribed on a unit circle. While a single 2D projection will rarely encompass all of the structure of higher-dimensional spaces, it is a reasonable summarization real task at hand, the selection of observation, and nearby comparison.

It is insightful to explore these two approximations against a visual of the model; prediction by observation is also displayed. Linked brushing and preserved aesthetic features such as circling misclassified observations help link information from the different spaces together.

4.2 Radial cheem tour

The global view facilitated the selection of a primary and optional comparison observation. The variable-level attribution of primary observation is normalized and used as the initial 1D basis in a radial tour. This is an approximation of the contributions of the linear variables that best explain the difference between the model intercept and an observations prediction, not the local shape of the model surface.

The initial frame is the normalized SHAP values of the primary observation. The current projection basis is depicted as the width of a bar, the variable's contribution to the horizontal axis. The normalized values of all observations are shown as vertical parallel coordinate plots.

The radial tour creates a basis path by varying the contribution of a selected variable, fully into and out of a projection frame. Doing so tests an individual variable's sensitivity to the structure identified by the local explanation. The default variable selected has the largest discrepancy between the primary and comparison

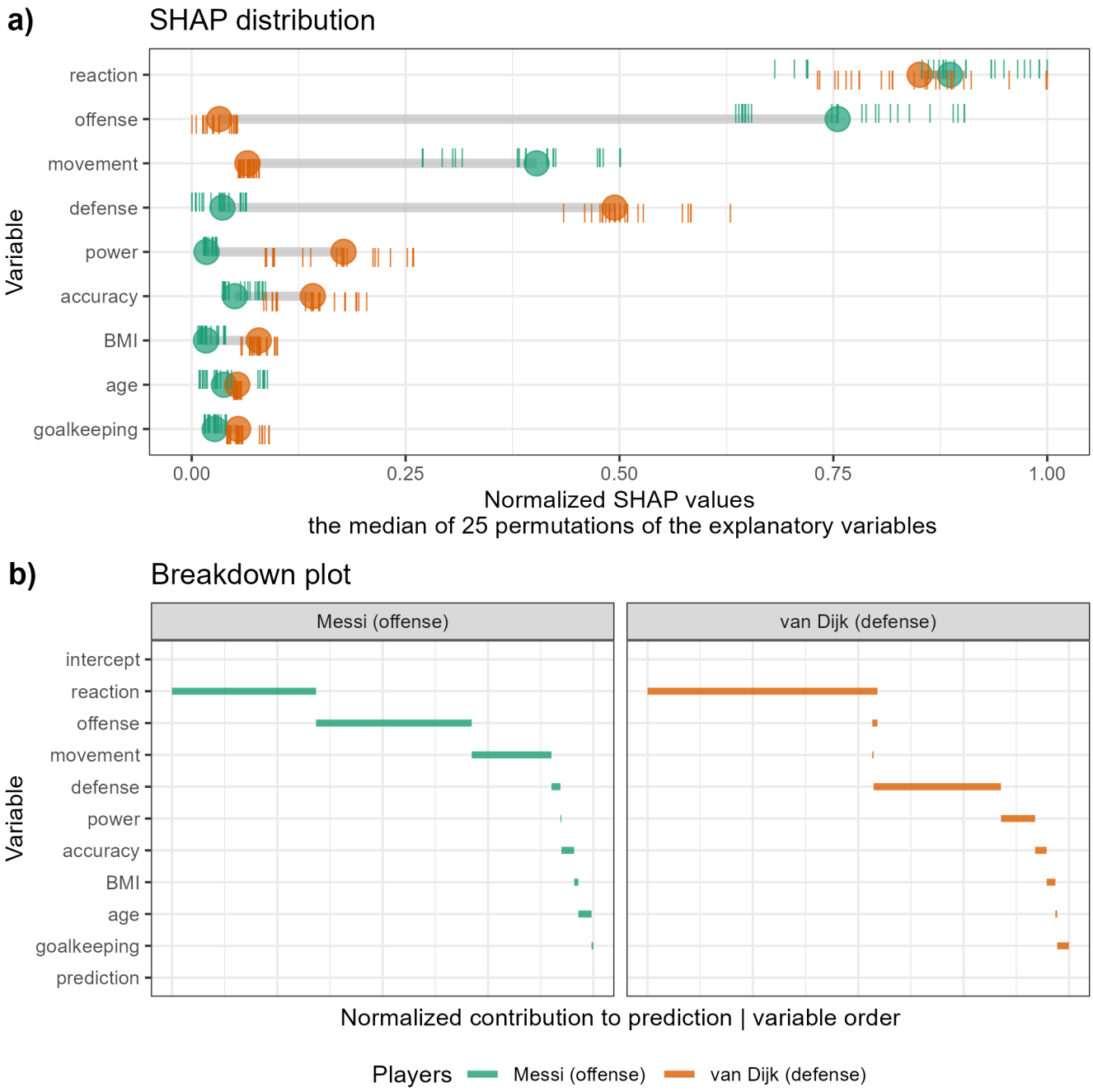

Figure 1: Illustration of the distribution of SHAP attributions and a breakdown plot. From FIFA 2020 data, a random forest model regresses wages from nine skill attributes for a star offensive and defensive player. The players have very different wages, but b) shows the distributions of the attributions permuting over 25 permutations in the explanatory variables. The medians of these distributions are the final SHAP values. The variable importance differs across the exogenous information of player position. These explanations make sense; the variable importances make sense in light of the player's position. c) Breakdown plots of the observations the explanation used to additively explain the difference between the model intercept and the observation prediction.

observations attribution. In the following sections, we elaborate on the takeaways we draw from applying this approach in classification and regression tasks, respectively.

Now that we have introduced the global view and corresponding cheem radial tour let's discuss the differences between the classification and regression cases.

4.3 Classification task

What information do we glean from using this method on a classification task? Typically we select a misclassified observation compared to a correctly classified point that is nearby in data space. We start by seeing the data projected through the linear attribution, the combination that best justifies that prediction. By default, the manual tour varies the contribution of the variable with the largest difference between the primary and comparison observation. That is, we can test the sensitivity of each variable to structure identified by the local explanation; we are exploring the support of the explanation, evaluating the support or robustness of the prediction.

Figure 2: Display illustrating the classification case. Both views are colored on predicted class and while red circles identify misclassified observations. The radial tour is a 1D projection starting at the normalized tree SHAP values of the primary point. The first frame is the linear-variable importances that best describe the difference from model intercept to this observation's prediction. We probe the support of the variable contributions by selecting a variable to vary the contribution.

4.4 Regression task

The regression case, we opt to color the global view on a statistic, the observation's residual, the log Mahanabis distance in data space (a measure of outlyingness), and the correlation of the attribution projection with the observed response. In the radial tour, the horizontal positions are the same, the attribution projection, then varying as the radial tour changes basis. While the vertical position is fixed to the observed response variable and residuals for the middle and right panels, respectively. This changes the display from 1D density to a scatterplot. The basis is still one component, the horizontal position, independent with the vertical position.

Figure 3: Display of the regression case. The global view can be colored on the correlation of the attribution projection and observed response. In the tour, the horizontal values are the same as the classification case. The vertical position is fixed to the observed response and residuals in the middle and right facets, respectively.

4.5 Interactive features

The application has several reactive selections that affect the data, coloring, and inputs for the tour to improve flexibility by extending the use cases. It also has interactive features in the global view that aid the analysis.

A tooltip displays while the cursor hovers over a point displays the observation number/name and classification

information if appropriate. Linked brushing allows for the selection of points (with left click and drag) where those points will be highlighted in both plots. The information corresponding to the selected points is populated on a dynamic table. These interactions aid exploration of the spaces and ultimately the selection of the primary and comparison observations.

rownum	class	y	prediction	predicted_class	is_misclassified	residual	cor_attr_proj.y	Calories	CalFat	TotFat_g	SatFat_g	Chol_mg	Na_mg	Carbs_g	Fiber_g	Sugars_g	Protein_g	tooltip
22	Dark	2	1.38	Milk	true	0.62	0	400	315	35	20	0	50	55	0	2.5	7.5	row: 22, Dark Chocolate Bar, Lindt, Switzerland Misclassified! predicted: Milk, observed: Dark
50	Dark	2	1.12	Milk	true	0.88	-0.1	460	162	18	12	10	90	72	2	52	4	row: 50, Dark Chocolate Bar, Mars, US Misclassified! predicted: Milk, observed: Dark
1	Dark	2	1.98	Dark	false	0.02	0.23	460	380	44	26	0	0	28	10	10	8	row: 1, Noir 86% Cacao, Cote D'Or, Belgium class: Dark

Figure 4: Illustration of data explorations interactions in the global view. This view has linked brushing of the points where observations selected in one facet are highlighted in the other facets and populate an interactive tabular display below. Tooltips display when hovering over a point

4.6 Preprocessing

The benefit of having dynamic interaction with data is predicated on a reasonably short render time. It is important to preprocess expensive operations so the application resources can be used efficiently. The work remaining at runtime is solely responding to inputs and rendering of visuals and tables. Below we discuss the steps and details of the reprocessing.

- **Data:** a complete numerical matrix; explanatory and response variable. An optional categorical variable can be mapped to color and shape of obervations. Explanatory variables are scaled in visualization after modeling or creating local explanations.
- **Model:** any model can be used with this method. Currently, we apply random forest models via the package **randomForest** [Liaw and Wiener (2002)] for compatability with the local explanation which requires tree-based models.
- **Local explanation:** any model-compatible linear explanation could be used. We apply tree SHAP, a more computationally efficient variant of SHAP, that is applicable to tree-based models. This is done with the package **treeshap** [Kominsarczyk et al. (2021), hosted on GitHub only]. The global view shows all observations in attribution space, requiring the variable importance from *all* observations rather than just one.

The time to preprocess the data will vary significantly with the model and local explanation. For reference, the FIFA data, 5000 observations of nine explanatory variables, took 2.9 seconds to fit a random forest model of modest hyperparameters. Extracting the tree SHAP values of each observation took 254 seconds combined. PCA and statistics of the variables and attributions took 0.6 seconds. These runtimes were from a non-parallelized R session on a modern laptop, but suffice it to say that the bulk of the time will be spent on the local attribution. Increased model complexity or data dimensionality will quickly become an obstacle. This makes tree SHAP, with its reduced computational complexity, a good candidate to start with.

Alternatively, the package **fastshap** (Greenwell 2020) claims extremely low runtimes, which are attributed to fewer calls to the prediction function, partial implementation in C++, and efficient use of logical subsetting.

4.7 Package infrastructure

The above-described method and application are implemented as an open-source **R** package, **cheem** *TODO:XXX include cran URL after published.* Preprocessing was facilitated with models created via **randomForest** [liaw_classification_2002], and explanations calculated with **treeshap** (Kominsarczyk et al. 2021). The application was made with **shiny** (Chang et al. 2021). The tour visual is an extension of **spinifex** (Spyrison and Cook 2020). Both views are created first with first with **ggplot2** (Wickham 2016) and then rendered as interactive HTML widgets with **plotly** (Sievert 2020). **DALEX** (Biecek 2018) and the free ebook, *Explanatory Model Analysis* (Biecek and Burzykowski 2021) was a huge boon to understanding local explanations and how to apply them.

Installation and get started with the package can be acheived by running the following in **R**:

```
## Download the package
install.packages("cheem", dependencies = TRUE)
## Restart the R session to pick up the file structure may be needed
restartSession()
## Load cheem into session
library("cheem")
## Try the app
run_app()
## Bring your own data: follow the examples in cheem_ls()
?cheem_ls
```

5 Case studies

To illustrate the use of the cheem method, we apply it to modern datasets, two classification examples and then two of regression.

5.1 1) Penguin, species classification

Palmer penguins data (Gorman, Williams, and Fraser 2014; Horst, Hill, and Gorman 2020) consist of 330 observations across four physical measurements of three species of penguins foraging near Palmer Station, Antarctica. A random forest model was fit, classifying the species of the penguin given the physical measurements.

In figure 5, a misclassified point is contrasted with a correctly classified point of its observed class nearby in data-space. The attribution space from the tree SHAP local explanations is a more separable space, where the comparison is squarely in the middle of the orange distribution. The primary observation is in-between the predicted and observed clusters, a sign of uncertainty in the prediction. The tour varies the contribution of bill length (b_l) as this variable differs most from the contribution of the comparison observation. Downplaying the contribution of bill length is crucial to the linear explanation of this observation be misclassified.

5.2 2) Chocolates, milk/dark chocolate classification

The chocolates dataset consists of 88 observations of 10 nutritional measurements from their labels. Each of which was labeled as being either milk or dark chocolates. With this data, we can see if a manufacturer gives an accurate portrayal of the chocolate. We are curious to see if there are chocolates that nutritionally look like milk chocolates that are labeled as dark chocolates, which may hold a higher market value. We should note that not all chocolates consist wholly of chocolate. The addition of other ingredients will decrease the predictive power of the model nutritional explanatory variable. A random forest model is fit classifying

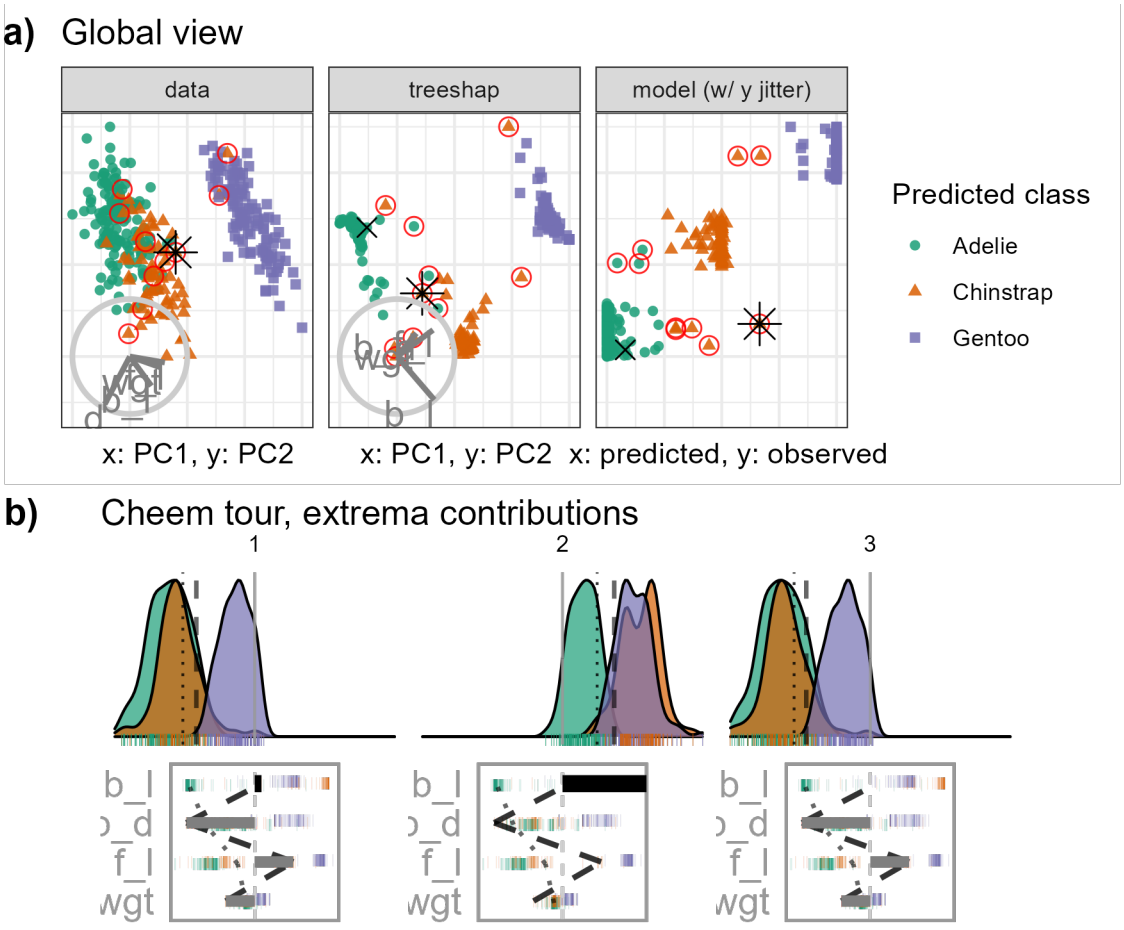

Figure 5: Species classification of Palmer penguin data.

the type of chocolate. We selected a chocolate labeled dark, through predicted to be milk chocolate with a comparison with chocolate labeled 85% cocoa.

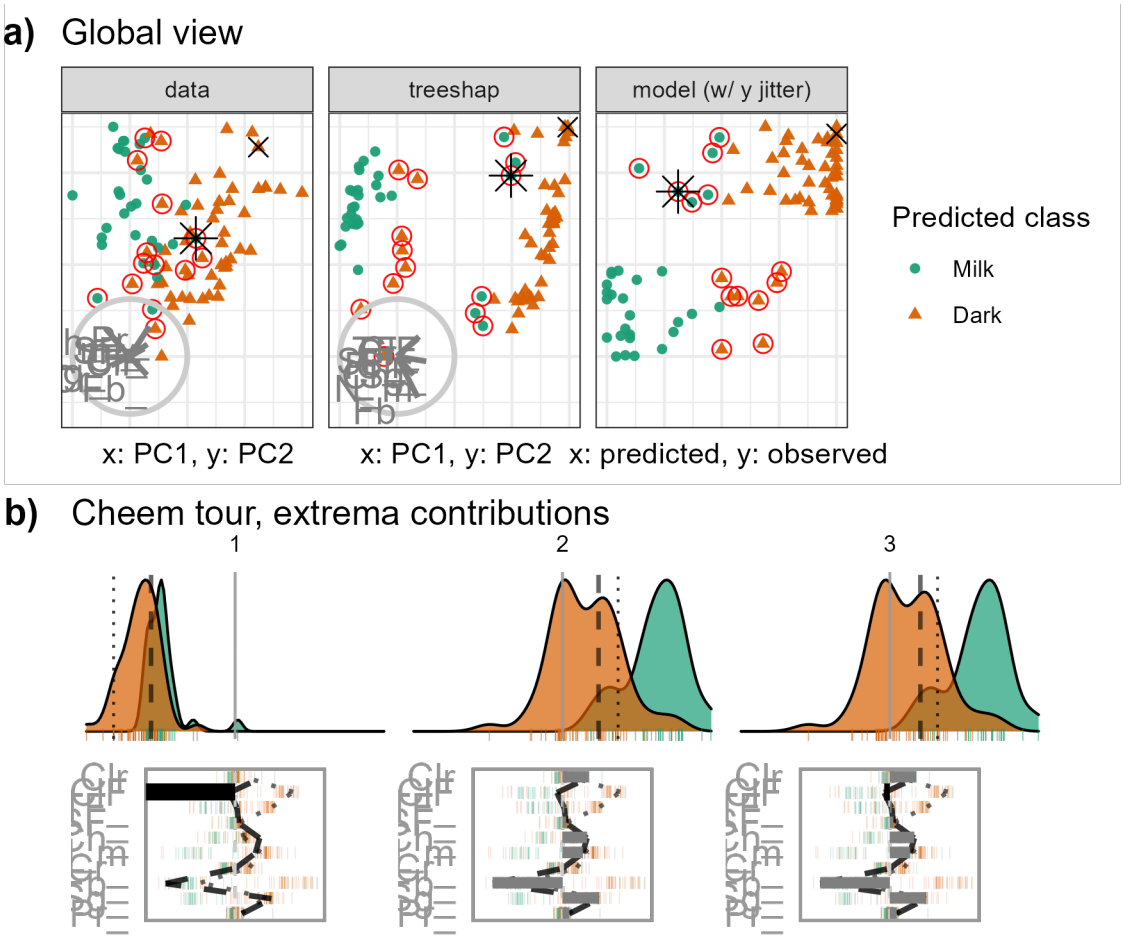

Figure 6: Chocolates data type classification (milk or dark).

From figure 6 we similarly see that attribution space is more separable relative to data-space. Interestingly there is not the class imbalance that we suspected; there are only 6 chocolates labeled as dark and predicted as milk, while 8 of the inverse case. Calories from fat is the variable with the largest difference in treeshap attribution between these points.

5.3 3) FIFA, wage regression

The 2020 season FIFA (Leone 2020; Biecek 2018), contains many skill measurements of soccer/football players and wage information. After aggregation of the skill measurements, we regress the log wages [2020 euros] given just the skill aggregates. The model was fit from 5000 observations of the nine skill aggregates before being thinned to 500 players to mitigate occlusion and render time. We compare a leading offensive fielder (L. Messi) with that of a top defensive fielder (V. van Dijk), the same observations were used in figure 1.

With figure 7 we will test the premise of the local explanation. If we remove reaction and movement skills from the basis, then offense skills has almost singular importance for the explanation of the offensive player. We vary the contribution of offensive skills. In the tour (3rd frame of b), offensive skills moved, and Messi is no longer separated from the group. We also notice that accuracy has rotated into the frame, maintaining some separability.

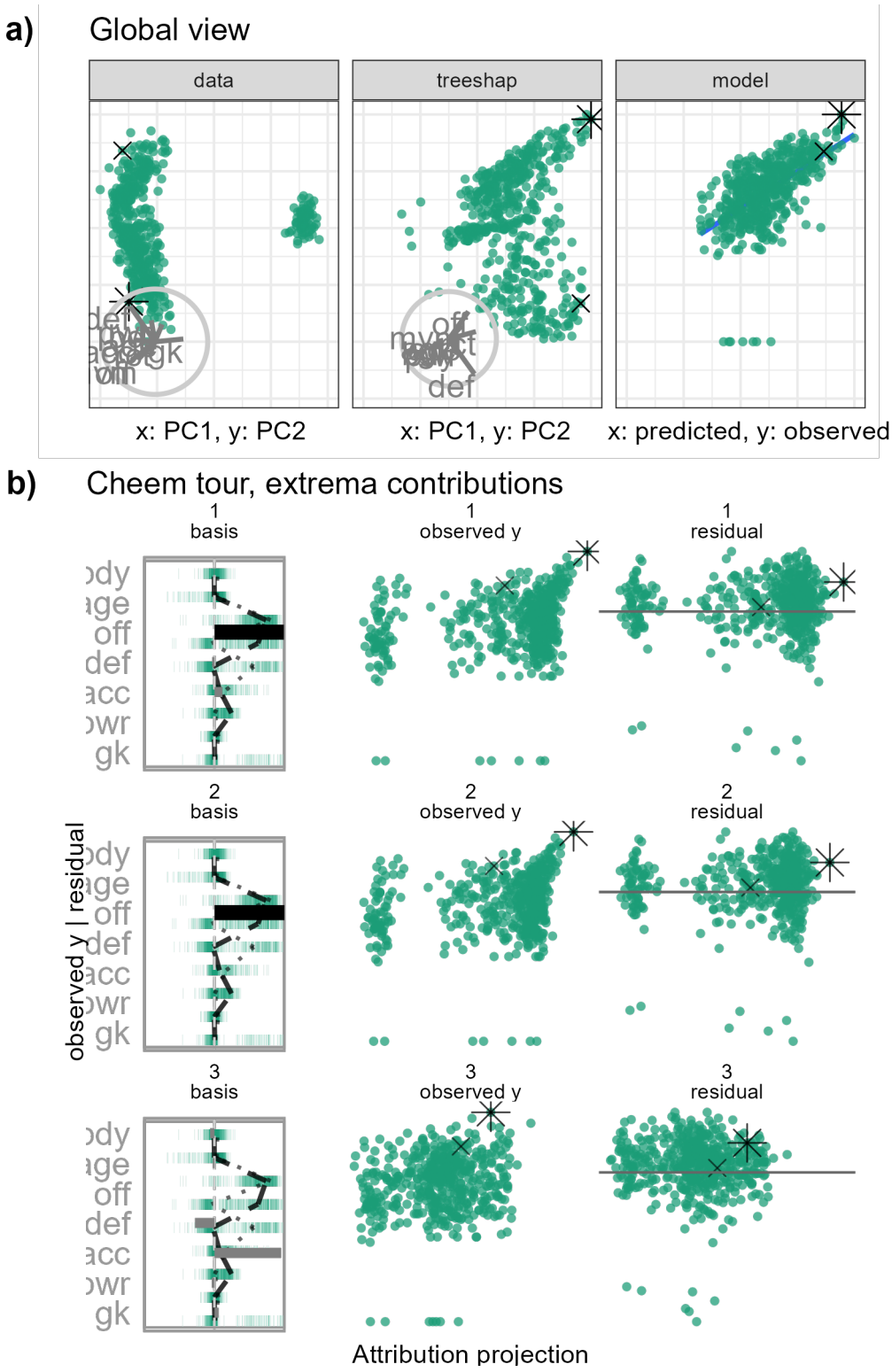

Figure 7: FIFA 2020, regressing log wages [2020 Euros] from aggregations of skill measurements. The primary observation is a star offensive player (L. Messi) compared with a top defensive player (V. van Dijk).

5.4 4) Ames housing 2018, sales price regression

Ames 2018, housing data was subset to North Ames (the neighborhood with the most house sales). The remaining are 338 house sales across nine variables. Using interaction from the global view, we select a house with an extreme negative residual and an accurate observation close to it in the data.

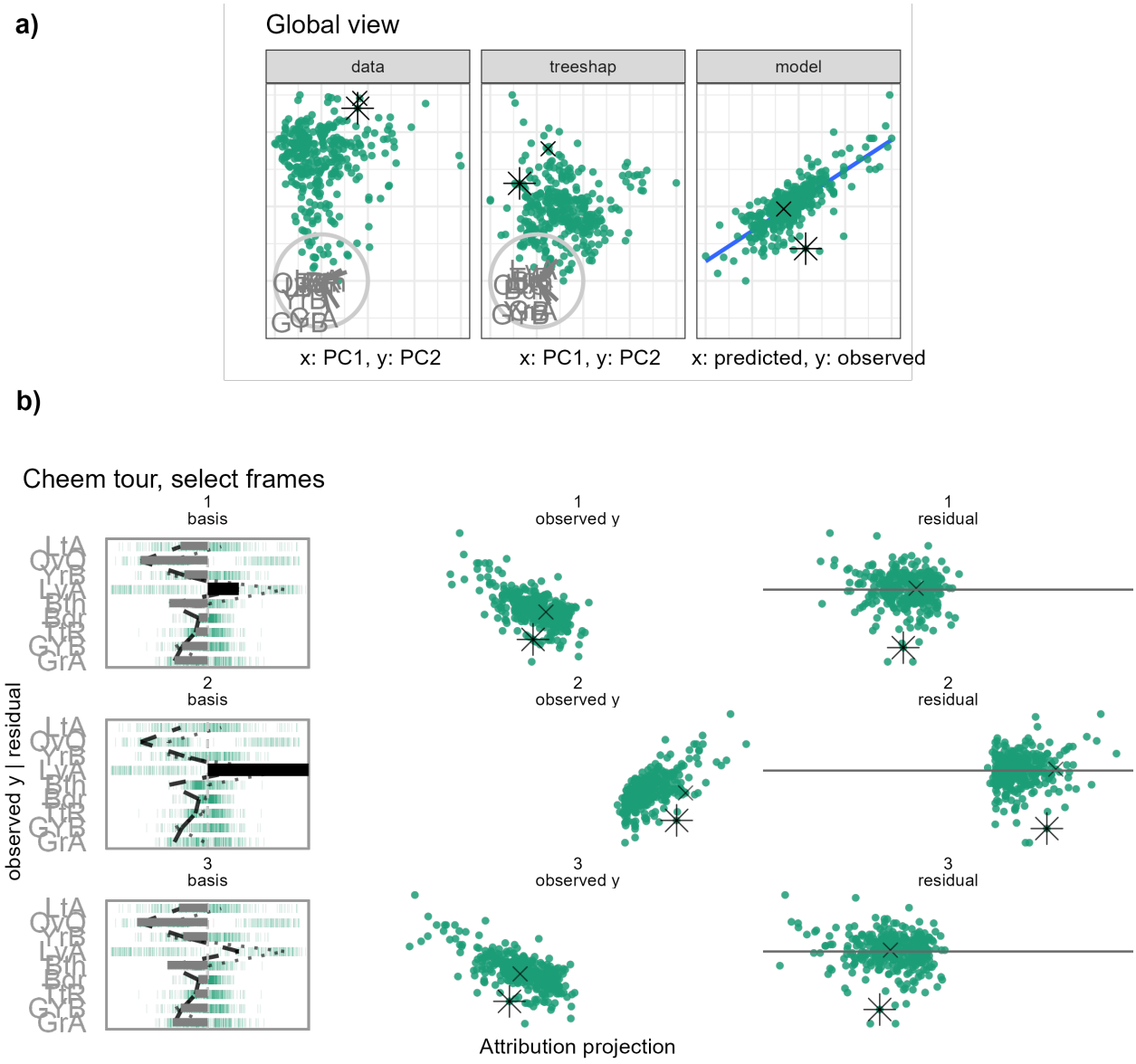

Figure 8: Ames housing 2018 regressing log sales price [2018 USD].

Figure 8 shows the global view and extrema of the tour. The horizontal distance in the tour didn't show a significant disparity between our selected points. This is not particularly surprising as most variables have a sizable contribution. Rotating any one variable out of the frame will rotate other vital variables into the frame, preserving most of the distance from intercept to prediction. However, the tour has revealed an interesting feature worth discussing. Notice that the observations pivot about the origin, the basis roughly halfway between bases in frames 1 and 2 of panel b) the data is near a singular profile. This means that

there is a basis orthogonal to this point that describes sizable variation. Knowing these singular bases can point toward others that have meaningful variation in the data.

6 Discussion

The need to maintain the interpretability of black-box models is evident. One aspect uses local explanations of the model in the vicinity of an observation. Local explanations approximate the linear variable importance to the model. Our contribution is to assess the explanations by examining the support of explanation, varying the contribution with a radial tour. First, a global view visualizes approximations of the data and explanation spaces side-by-side, using dynamic interaction to compare and contrast, and ultimately, identify primary and comparison observations of interest. Then normalized the linear importance from the explanation of the primary observation to use as the initial basis of a manual tour. The variable sensitivity to the structure identified in the explanation is explored with the tours varying basis.

We have illustrated this method on random forest models using the tree SHAP local explanation, while it could be generally used with any compatible model-explanation pairing. We apply it to the classification and regression tasks. We have created an open-source **R** package **cheem**, available on CRAN, to facilitate preprocessing and exploration with the described interactive application. Toy and real data are provided or upload your data after preprocessing.

7 Acknowledgments

We would like to thank Professor Przemyslaw Biecek for his input early in the project and to the broader MI^2 lab group for the **DALEX** ecosystem of **R** and **Python** packages. This research was supported by Australian government Research Training Program (RTP) scholarships.

The namesake, Cheem, refers to a fictional race of humanoid trees from Doctor Who lore. **DALEX** pulls on from that universe, and we initially apply tree SHAP explanations that are specific to tree-based models.

References

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–60.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, and Richard Benjamins. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Asimov, Daniel. 1985. "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal on Scientific and Statistical Computing* 6 (1): 128–43. https://doi.org/https://doi.org/10.1137/0906011.

Biecek, Przemyslaw. 2018. "DALEX: Explainers for Complex Predictive Models in R." *The Journal of Machine Learning Research* 19 (1): 3245–49.

———. 2020. *ceterisParibus: Ceteris Paribus Profiles*. https://CRAN.R-project.org/package=ceterisParibus.

Biecek, Przemyslaw, and Tomasz Burzykowski. 2021. *Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models*. CRC Press.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16 (3): 199–231.

Buja, Andreas, and Daniel Asimov. 1986. "Grand Tour Methods: An Outline." In *Proceedings of the Seventeenth Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics*, 63–67. New York, NY, USA: Elsevier North-Holland, Inc. http://dl.acm.org/citation.cfm?id=26036.26046.

Chang, Winston, Joe Cheng, JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges. 2021. *Shiny: Web Application Framework for r*. https://CRAN.R-project.org/package=shiny.

Cook, Dianne, and Andreas Buja. 1997. "Manual Controls for High-Dimensional Data Projections." *Journal of Computational and Graphical Statistics* 6 (4): 464–80. https://doi.org/10.2307/1390747.

Dastin, Jeffrey. 2018. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters*, October. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Díaz, Mark, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. "Addressing Age-Related Bias in Sentiment Analysis." In *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems*, 1–14.

Duffy, Claire. 2019. "Apple Co-Founder Steve Wozniak Says Apple Card Discriminated Against His Wife." *CNN*, November. https://www.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/index.html.

Gorman, Kristen B., Tony D. Williams, and William R. Fraser. 2014. "Ecological Sexual Dimorphism and Environmental Variability Within a Community of Antarctic Penguins (Genus Pygoscelis)." *PloS One* 9 (3): e90081.

Gosiewska, Alicja, and Przemyslaw Biecek. 2019. "IBreakDown: Uncertainty of Model Explanations for Non-Additive Predictive Models." *arXiv Preprint arXiv:1903.11420*.

Greenwell, Brandon. 2020. *Fastshap: Fast Approximate Shapley Values*. https://CRAN.R-project.org/package=fastshap.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B. Gorman. 2020. "Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data." https://allisonhorst.github.io/palmerpenguins/.

Kodiyan, Akhil Alfons. 2019. "An Overview of Ethical Issues in Using AI Systems in Hiring with a Case Study of Amazon's AI Based Hiring Tool." *Researchgate Preprint*.

Kominsarczyk, Konrad, Pawel Kozminski, Szymon Maksymiuk, and Przemyslaw Biecek. 2021. "Treeshap." Model Oriented. https://github.com/ModelOriented/treeshap.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm." *ProPublica*, May. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm?token=RPR1E2qtzJltfJ0tS-gB_41kmfoWZAu4.

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 343 (6176): 1203–5.

Lee, Stuart, Dianne Cook, Natalia Da Silva, Ursula Laa, Earo Wang, Nick Spyrison, and H. Sherry Zhang. 2021. "A Review of the State-of-the-Art on Tours for Dynamic Visualization of High-Dimensional Data." *arXiv Preprint arXiv:2104.08016*.

Lee, Yoon Dong, Dianne Cook, Ji-won Park, and Eun-Kyung Lee. 2013. "PPtree: Projection Pursuit Classification Tree." *Electronic Journal of Statistics* 7: 1369–86.

Leone, Stefano. 2020. "FIFA 20 Complete Player Dataset." https://kaggle.com/stefanoleone992/fifa-20-complete-player-dataset.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Lundberg, Scott M., Gabriel G. Erion, and Su-In Lee. 2018. "Consistent Individualized Feature Attribution for Tree Ensembles." *arXiv Preprint arXiv:1802.03888*.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *arXiv Preprint arXiv:1705.07874*.

O'neil, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. "\"Why Should I Trust You?\": Explaining the Predictions of Any Classifier." *arXiv:1602.04938 [Cs, Stat]*, February. http://arxiv.org/abs/1602.04938.

Salzberg, Steven. 2014. "Why Google Flu Is A Failure." *Forbes*, March. https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/.

Shapley, Lloyd S. 1953. *A Value for n-Person Games*. Princeton University Press.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. 2017. "Learning Important Features Through Propagating Activation Differences." In *International Conference on Machine Learning*, 3145–53. PMLR.

Shrikumar, Avanti, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. 2016. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences." *arXiv Preprint arXiv:1605.01713*.

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman; Hall/CRC. https://plotly-r.com.

Silva, Natalia da, Dianne Cook, and Eun-Kyung Lee. 2021. "A Projection Pursuit Forest Algorithm for Supervised Classification." *Journal of Computational and Graphical Statistics*, 1–21.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2014. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." In *In Workshop at International Conference on Learning Representations*. Citeseer.

Spyrison, Nicholas, and Dianne Cook. 2020. "Spinifex: An R Package for Creating a Manual Tour of Low-Dimensional Projections of Multivariate Data." *The R Journal* 12 (1): 243. https://doi.org/10.32614/RJ-2020-027.

Strumbelj, Erik, and Igor Kononenko. 2010. "An Efficient Explanation of Individual Classifications Using Game Theory." *The Journal of Machine Learning Research* 11: 1–18.

Štrumbelj, Erik, and Igor Kononenko. 2014. "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems* 41 (3): 647–65.

Vanni, Laurent, Mélanie Ducoffe, Carlos Aguilar, Frédéric Precioso, and Damon Mayaffre. 2018. "Textual Deconvolution Saliency (TDS): A Deep Tool Box for Linguistic Analysis." In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 548–57.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Dianne Cook, and Heike Hofmann. 2015. "Visualizing Statistical Models: Removing the Blindfold." *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8 (4): 203–25. https://doi.org/10.1002/sam.11271.