New Tools to Interactively Explore the Sensitivity of Structure in Low-dimensional Projections of Data

Pre-submission review — September 2021

Nicholas Spyrison, B.Sc

Monash University
Faculty of Information Technology
Department of Human-Centred Computing

Thesis Supervisors

Prof. Kimbal Marriott
Prof. Dianne Cook

Committee Members

Dr. Maxime Cordiel
Dr. Shirui Pan

Chair

Assoc. Prof. Bernhard Jenny

Contents

Contents **1**

1 Introduction **1**

2 Motivation **2**

3 Research objectives **3**

4 Methodology **4**

5 Work since the mid-candidature review **4**

5.1 Experimental study 5

5.2 Trees of Cheem 6

6 Proposed thesis structure **10**

7 Other Contributions **10**

8 Acknowledgements **11**

References **11**

1 Introduction

The thesis of this work is central to multivariate data visualization. More specifically, we focus on the class of many linear projections are viewed near-continuously through small changes to the projection basis known as data visualization *tours* Lee et al. (2021).

There are many variants of tours. We focus on one branch, *manual tours* Spyrison and Dianne Cook (2020), that allows for user interaction by selecting one variable and specifying how to change its contribution to the current projection. By controlling the contribution of a single variable, a user can explore its sensitivity to the structure of the projection and identify which variables are ultimately most important to the structure in question. The work addressing the first research objective clarified the rationale for doing so and implements a free, open-source `R` package for applying the manual tour.

Next, we substantiated the efficacy of manual tours as compared with discrete combinations of principal components (Pearson 1901) and the *grand tour*(Asimov 1985). We do so with an $N = 108$ within-participant user study, where all participants use each of these visual factors. This is performed over balanced trials across the other experimental factors: location, shape, and dimension of the data. This addresses the second research objective.

In our latest work, we want to see if we can apply the manual tour to aid the interpretability of complex, black-box models. One recent branch in explainable artificial intelligence (XAI, Adadi and Berrada (2018), Arrieta et al. (2020)) is the use of local explanations or attribution of the variables for one observation of an agnostic black-box model. One local explanation is the SHAP values (Lundberg and Lee 2017, EMA?). We use these SHAP values as a 1D basis and perform manual tours to explore how the SHAP values behave differently for misclassified and class-corrupted observations against neighboring correctly classified observations. This work corresponds to the third research objective

2 Motivation

The term exploratory data analysis (EDA) was coined by Tukey (1977), who leaves it as an intentionally broad term that encompasses the initial summarization and visualization of a data set, before a hypothesis to test has been formulated. This is a critical first step for understanding and becoming familiar with data and validating model assumptions. It may be tempting to review a series of summary statistics to check model assumptions. However, there are known datasets where the same summary statistics miss glaringly obvious visual patterns (Anscombe 1973; Matejka and Fitzmaurice 2017). It is easy to look at the wrong, or incomplete set of statistics needed to validate assumptions. Data visualization is crucial in EDA, it *forces* you to see details and peculiarities of the data which are opaque to numeric summarization, or more nefariously, obscure their true values. Data visualization does and must remain a primary component of data analysis and model validation.

While static documents are the norm, there are sizable benefits of user interaction. Interactive data visualization shift the locus of control back to the user, inviting them to explore and interact with the data, and offers a compact way to explore a wider range of dimensions, questions, and keep the curiosity and the interest of the user.

With the emerging field of XAI, the constant tension between interpretability of a model and its predictive power is receiving more attention. Linear models are the champions of interpretability with modest accuracy while increasing complex models improve accuracy but they can scarcely be interpreted even by experienced practitioners. One way to gain insight into a model is to focus on the local vicinity of one observation, and explain the variable weighting around that location, in an agnostic non-linear model. We call this observation level variable weights a *local explanation*(Biecek and Burzykowski 2021). There are various such local explanations, many are tied to specific classes of models, while others are model-agnostic. LIME[@] and SHAP[@] are two such examples.

We know that data visualization is important in EDA and assumption validation. User interaction allows us to explore widely and quickly while allowing us to explore ideas as they arise. These two elements were used to answer the first RO. Their efficacy was supported in response to the second RO. In this work, we apply a manual tour in tandem with SHAP local explanations to address the third RO.

3 Research objectives

The overall question of interest is:

Can the geodesic interpolator with user interaction help analysts understand linear projections of data, and explore the sensitivity of structure in the projection to the variables contributing to the projection?

Which is further divided into these more specific objectives:

1. **How do we define user interaction for the geodesic interpolator to add and remove variables smoothly from a 2D linear projection of data?**

   Cook and Buja (1997) described an algorithm for manually controlling a tour (p-D into 2D), to rotate a variable into and out of a 2D projection. This algorithm provides the start to a human-controlled geodesic interpolator (GI). The work(Spyrison and Dianne Cook 2020) was adapted so that the user has more control of the interpolation. The user is able to set the range of motion from full $[-1, 1]$, to allow the user to intercept the rotation at any step, and to output to a device that allows the user to reproduce motions and animate or rock the rotation backward and forwards. These fine-tuned controls provide a better tool for sensitivity analysis.

2. **Do analysts understand the relationship between variables and structure in a 2D linear projection better when the geodesic interpolator is available?**

   We performed an $N = 108$, within0participant user study comparing accuracy and time with the primary factor as the type of data visualization. Each participant performed 2 evaluations with either discrete PCA, grand tour, or radial manual tour. We find strong evidence that the radial tour increases accuracy. We also show the effects from the other experimental factors of location, shape data dimensionality, and the random effects from the data and that of the participants.

3. **Can we use the geodesic interpolator in conjunction with the local explanation SHAP to improve the interpretability of black-box models?**

   The tension from the trade-off between accuracy and interpretability of black-box models is rising. Below we use SHAP to extract local explanations from a random forest model and use those SHAP values as a projection basis to perform manual tours. We add class-corrupted observations and explore how the model and SHAP values react.

4 Methodology

The research corresponding with RO #1 entails *algorithm design* adapting the algorithm from Cook and Buja (1997). This allows for interactive control of 2D projections and serves as a foundation for the remaining work to follow.

To address RO #2, a controlled *experimental study* has explored the efficacy of interactive radial tours as compared with two benchmark methods: Principal Component Analysis (PCA, Pearson (1901)) and the grand tour(Asimov 1985). This was a within-participant user study where each participant experienced each visual. Trials were balanced across 3 other experimental factors: location of the signal, shape of the cluster distributions, and dimension of the data.

The research for RO #3 involves *algorithm design.* We know that the SAHP value is a local explanation for one observation. This SHAP value will also serve as the 1D basis for the manual tour. While using SHAP as a projection basis is novel it is not particularly insightful by itself. We provide tracking marks for the selected observation, a comparison neighbor as the basis changes. We also offer a view and quantitative analysis for the data and full SHAP matrix to keep the local view in context and measure how extreme the SHAP values are behaving.

5 Work since the mid-candidature review

In candidature confirmation review, we discussed the implementation of the *geodesic interpolator* with user interaction (for RO #1) which resulted in the open-source R package, `spinifex` available on CRAN and its subsequent publication (Spyrison and Dianne Cook 2020).

At the mid-candidature review, we discussed the experimental design of the user study to substantiate the efficacy of the radial tour as compared with PCA (discrete with user interaction), and the grand tour (continuous without user interaction). Below we briefly report our findings supporting RO#2 before discussing the work addressing RO#3.

5.1 Experimental study

The $N = 108$ within-participant user study collected 6 trials from each participant (648 total), with 2 trials of each of visuals: PCA, grand tour, and radial tour. Three further factors: location, shape, and data dimensionality were also evenly evaluated for a comparison with the effect of controlling the visuals.

In summary, we use a mixed regression model, using the factors above as main effects, and use the participant and data simulations as random effects. We regress on Y_1, accuracy, and Y_2, log time. We test increasingly complex interactions of the main effects, but settle on the following model to look at the coefficient output.

$$\begin{aligned}
&\widehat{Y} = \mu + \alpha_i * \beta_j + \mathbf{Z} + \mathbf{W} + \epsilon \\
\text{where}\quad &\mu \text{ is the intercept of the model including the mean of random effect} \\
&\epsilon \sim \mathcal{N}(0,\ \sigma),\ \text{the error of the model} \\
&\mathbf{Z} \sim \mathcal{N}(0,\ \tau),\ \text{the random effect of participant} \\
&\mathbf{W} \sim \mathcal{N}(0,\ \upsilon),\ \text{the random effect of simulation} \\
&\alpha_i,\ \text{fixed term for factor} \mid i \in (\text{pca, grand, radial}) \\
&\beta_j,\ \text{fixed term for location} \mid j \in (0\_1,\ 33\_66,\ 50\_50)\ \%\ \text{noise/signal mixing} \\
&\gamma_k,\ \text{fixed term for shape} \mid k \in (\text{EEE, EEV, EVV banana})\ \text{model shapes} \\
&\delta_l,\ \text{fixed term for dimension} \mid l \in (4\ \text{variables}\ \&\ 3\ \text{cluster},\ 6\ \text{variables}\ \&\ 4\ \text{clusters})
\end{aligned}$$

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	-0.12	0.08	43.9	-1.50	0.14	
factor						
fct=grand	0.15	0.09	622.4	1.74	0.08	
fct=radial	0.37	0.09	617.1	4.18	0.00	***
fixed effects						
loc=33_66	0.17	0.09	83.2	1.78	0.08	
loc=50_50	0.14	0.09	84.8	1.52	0.13	
shp=EEV	0.04	0.06	11.5	0.79	0.44	
shp=ban	-0.03	0.06	11.5	-0.48	0.64	
dim=6	-0.06	0.05	11.5	-1.39	0.19	
interactions						
fct=grand:loc=33_66	-0.06	0.13	587.3	-0.49	0.63	
fct=radial:loc=33_66	-0.34	0.13	585.2	-2.65	0.01	**
fct=grand:loc=50_50	-0.09	0.13	589.6	-0.68	0.50	
fct=radial:loc=50_50	-0.19	0.13	574.3	-1.43	0.15	

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	2.71	0.14	42.6	19.06	0.00	***
factor						
fct=grand	-0.23	0.12	567.6	-1.97	0.05	*
fct=radial	0.16	0.12	573.5	1.34	0.18	
fixed effects						
loc=33_66	0.05	0.14	40.9	0.34	0.74	
loc=50_50	-0.05	0.14	42.1	-0.35	0.73	
shp=EEV	-0.15	0.09	8.3	-1.61	0.14	
shp=ban	-0.13	0.09	8.3	-1.42	0.19	
dim=6	0.14	0.08	8.3	1.90	0.09	
interactions						
fct=grand:loc=33_66	0.24	0.18	580.9	1.34	0.18	
fct=radial:loc=33_66	-0.24	0.18	582.4	-1.32	0.19	
fct=grand:loc=50_50	0.12	0.18	578.6	0.69	0.49	
fct=radial:loc=50_50	0.05	0.18	584.4	0.25	0.80	

Figure 1: Model coefficients regressing against our accuracy measure (left) and log time (right). We have strong evidence supporting a relatively large increase in accuracy with the radial tour. We also notice that there is some evidence suggesting that use of the grand tour is fastest, perhaps because there is no interaction and participants can devote all of their attention to watching the animation once.

A more in-depth description and discussion of this user study is attached as appendix A, a draft version of the paper we intend to submit to the Journal of Data Science, Statistics, and Visualization.

5.2 Trees of Cheem

For the third project the higher-level goal is to use manual tours, that is, interactive, continuous linear projections in order to improve the interpretability of black-box models. There are different measures and degrees of specificity that describe a model. Local explanations describe the linear variable weights in the vicinity of an observation, given the model. Below we use the local explanation, SHAP, as applied to a random forest model. This is in the interest of mitigating the relatively long runtime of SHAP. For comparison, the FIFA data we look at below is 5000 observations of 9 aggregate skill attributes, which takes 10 seconds to fit the random forest model, but about 1230 seconds to extract all 5000 SHAP values.

5.2.1 SHAP values and prediction explanations.

To illustrate SHAP we use a dataset of FIFA soccer players (2020 season), a subset of 5000 players, observations of 42 wage, and skill variables(Leone 2020). Following the work in Biecek and Burzykowski (2021) we can extract SHAP values and also create a "break down" profile explanation of the observations prediction. Such an additive explanation has an asymmetry to the ordering of its variables. The figure below take a look at the SHAP and break down profiles of a star offensive and defensive player.

While this shows the differing of weights across 2 different fielders within the same model we have lost sense of the global distribution of the explanatory variables. We illustrate this with the smaller Palmer penguins data(Gorman, Williams, and Fraser 2014). We will extract all observation's SHAP values. We now have the original $[n \times p]$ data and wish to compare it with the distribution of the SHAP matrix of the same dimensionality. We approximate their data with the first 2 principal components. We extract the with-in class Mahalanobis distance and put the QQ plots of these distances for a comparison of how sensitive the SHAP space is relative to the data space.

Given the global view above we want to look at the local weightings of primary and comparison points (shown as '*/x' above and dashed/dotted lines below). In this case, the primary observation is a class-corrupted observation while the comparison point is a correctly classified nearby point. The idea is to view just how sensitive SHAP values are to this sort of class-corruption attack.

The application is quickly maturing and will be shown to experts for comment. This work is being written up to be submitted to the WHY-21 workshop, part of the NeurIPS 2021 Conference.

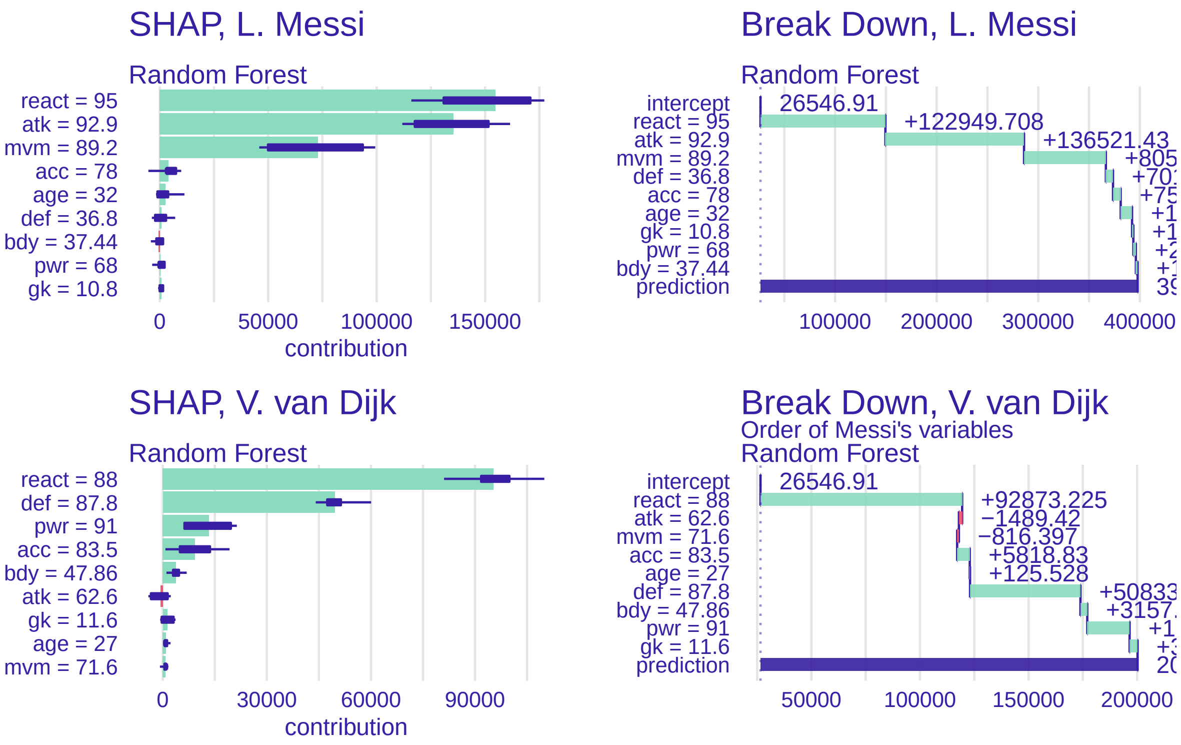

Figure 2: SHAP values and prediction explanations of an offensive player (Messi, top) and a defensive player (van Dijk). SHAP values show a change in weights at the location of each player. Break down profiles show one order-sensitive explanation for the prediction of that observation.

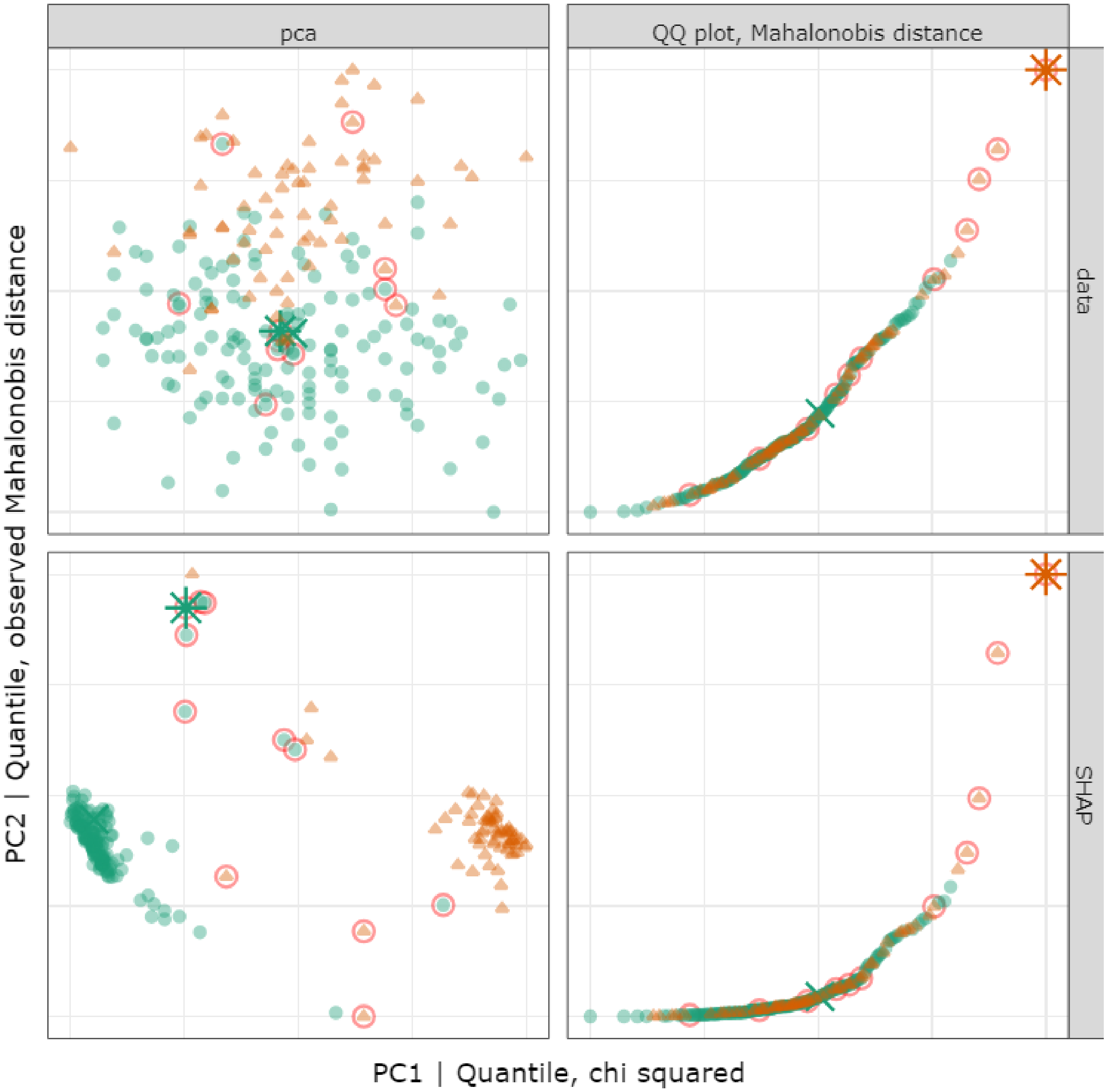

Figure 3: Screen capture from an interactive shiny application. Data and SHAP spaces (top and bottom respectively) of Palmer penguins by the first 2 principal components and quantile-quantile plots of their with-in class Mahalanobis distances (left and right respectively. The points are colored and shaped according to their predicted class, misclassified points are identified with a red circle. A class-corrupted target observation '*' is shown in comparison with nearby real observation 'x'. These same two points are tracked in the proceeding tour.

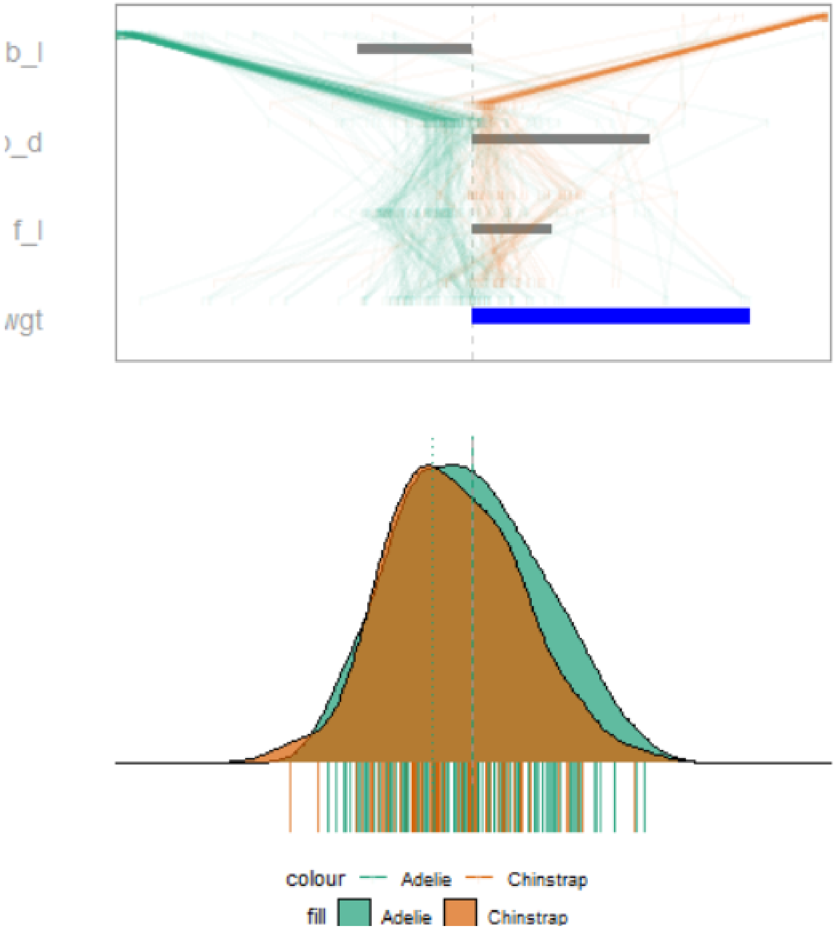

Figure 4: The first frame of the radial tour. The SHAP values of the selected '*' from the preceding figure given the variable weights of the 1D basis (grey/blue bars on top), the class distributions of the SHAP values are shown as parallel coordinate plots above each variable contribution. The class densities and rug values are shown on the bottom. A light grey light shows zero on the projection, with the dashed and dotted lines correspond to the position of primary and comparison observations ('*/x' in the preceeding figure). The tour animates over small changes in the basis (top bars) as the variable with the largest contribution (weight) is rotated to have a full contribution, zero contribution, and then back to the initial contribution.

5.2.2 Discussion

We have used radial tours to improve the interpretability of black-box models by exploring local explanation SHAP's sensitive to misclassified observations. It is important to note that this is independent of the quality of the model or even the quality of the explanation. Indeed the very term explanation feels like a bit of a misnomer as it seems to imply reason or validity, rather I prefer to think of it as local weightings of the model.

Keeping in mind the real-world application is particularly important. Finding methods to better interpret black-box models is an important challenge as corporations and nation-states increasingly use complex models to classify and predict their customers and citizens. Being able to glean insight into a models weights and how they differ for misclassified observations is extremely important for building and challenging models as we attempt to build a just world of tomorrow.

6 Proposed thesis structure

- The coursework, Graduate Research 120 hours are complete and approved.

This is my assessment of the completion of the thesis research thus far:

- Introduction – 60%
- Literature review – 80%
- (RO #1) GI & manual tours – 90%
- (RO #2) manual tour efficacy user study – 80%
- (RO #3) manual tour applied to SHAP values – 60%
- Conclusion and future plans – 40%

Figure 5 illustrates the purposed timeline for this research.

7 Other Contributions

- "Is IEEE VIS *that* good?" AltVis (Spyrison, Lee, and Besançon 2021)
- A Review of the State-of-the-Art on tours Dynamic Visualization of High-dimensional Data (Lee et al. 2021)
- 1st place in 2020 Melbourne Data Marathon (Barrow, Chong, and Spyrison 2020)

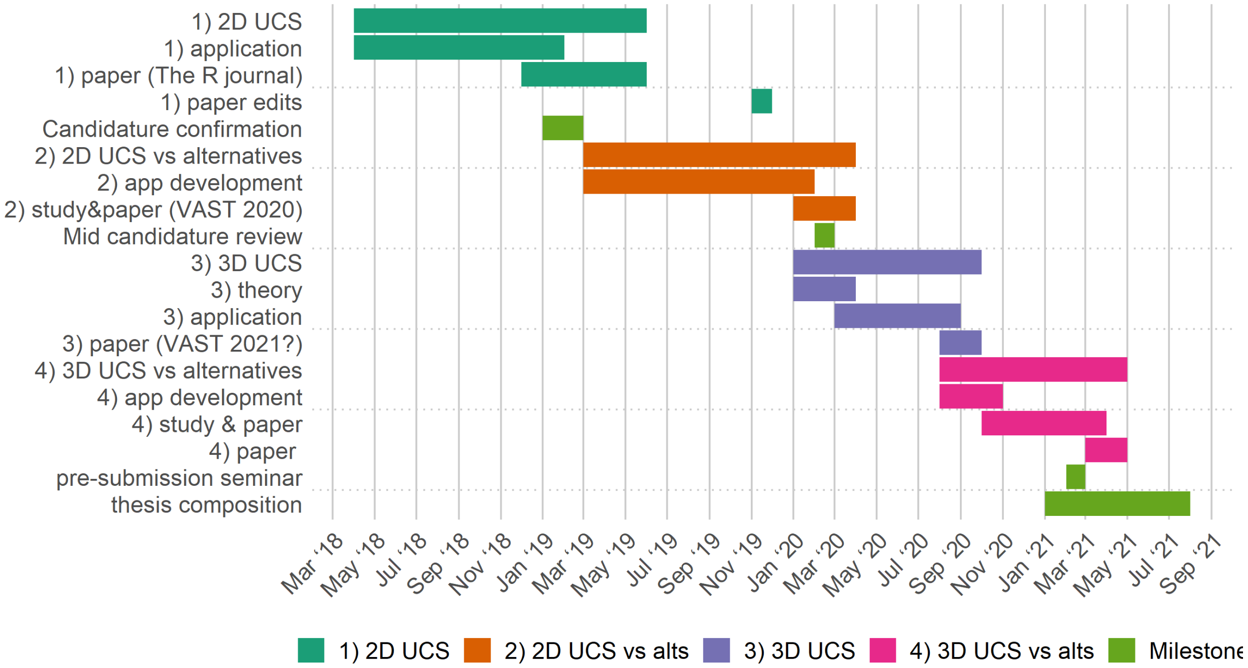

Figure 5: Proposed research timeline.

8 Acknowledgements

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. This article was created in R (R Core Team 2020) and rmarkdown (Xie, Allaire, and Grolemund 2018).

For version control, transparency, and reproducibility, the source files are made available at github.com/nspyrison/phd_milestones.

References

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

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. https://doi.org/10.2307/2682899.

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.

Barrow, Madeleine, Jieyang Chong, and Nicholas Spyrison. 2020. "Melbourne Datathon 2020." In *Melbourne Datathon 2020, Insights Category*. https://www.overleaf.com/project/5f614799515ac0000119daf7.

Biecek, Przemyslaw. 2018. "DALEX: Explainers for Complex Predictive Models in R." *Journal of Machine Learning Research* 19 (84): 1–5. https://jmlr.org/papers/v19/18-416.html.

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

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.

Cook, Dianne, Andreas Buja, Eun-Kyung Lee, and Hadley Wickham. 2008. "Grand Tours, Projection Pursuit Guided Tours, and Manual Controls." In *Handbook of Data Visualization*, 295–314. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-33037-0_13.

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.

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:2104.08016 [Cs, Stat]*, April. http://arxiv.org/abs/2104.08016.

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

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

Matejka, Justin, and George Fitzmaurice. 2017. "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–94. Denver, Colorado, USA: ACM Press. https://doi.org/10.1145/3025453.3025912.

Pearson, Karl. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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): (accepted).

Spyrison, Nicholas, Benjamin Lee, and Lonni Besançon. 2021. ""Is IEEE VIS *That* Good?" On Key Factors in the Initial Assessment of Manuscript and Venue Quality." *OSF Preprints*, July. https://doi.org/10.31219/osf.io/65wm7.

Tukey, John W. 1977. *Exploratory Data Analysis*. Vol. 32. Pearson.

Xie, Yihui, J. J. Allaire, and Garrett Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.