Efficacy of the radial tour and application to extend interpretability of black-box models when coupled with local explanations

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

1 Introduction

This work is concerned with linear projections of multivariate data. More specifically, we focus on the class of visualizations known as *tours* Lee et al. (2021). Tours are viewed near-continuously through small changes to the projection basis. There are many variants of tours. We focus on one of the variants, *manual tours*Spyrison and Dianne Cook (2020). Manual tours can be used to control or steer the projection basis, while other tour variants select bases randomly or optimize an objective function. Because of this unique attribute, *user interaction* is another key aspect of interest in this work.

Manual tours change the basis 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. In the work addressing RO#1 we improve upon the Geodesic Interpolator for the manual tour, and

apply it in an open source `R` package, `spinifex`, this package facilitates making manaul tours and extends the graphics packages interoperability for itself and the tours made from the existing package `tourr`.

Next, we substantiated the efficacy of manual tours as compared with user selected, discrete combinations of principal components (Pearson 1901) and continuous projections without interaction with the *grand tour* (Asimov 1985). We contucted 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 black-box models. Such models have a non-linear weighting space making them hard to interpret or understand. One recent branch in explainable artificial intelligence Arrieta et al. (2020) is the use of local explanations or the attribution of the variables for one observation of a black-box model. We use such an explanation to form a 1D basis and perform manual tours to explore how the SHAP values behave differently for misclassified 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 testing hypothesis 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 as they arise, which helped to keep the curiosity and the interest of the user.

With the emerging field of XAI, the constant tension between the 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*. There are various such local explanations, many are tied to specific classes of models, while others are model-agnostic.

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 2 elements were especially important to consider from the work addressing RO#1 as it forms a foundation to build on for the further work. The efficacy of manual tours was supported by a user study in response to the second obujective. In the latest work, we apply tours in tandem with local explanations to extend the interpretability of black-box models, to address RO#3.

3 Research objectives

The over-aching question of interest can be stated as:

Can the geodesic interpolator with user interaction help analysts understand linear projections, 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 that we address respectively:

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, 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 can set the full range of the contribution from $[-1, 1]$, and 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 $N = 108$, within-participant user study comparing accuracy and time with the primary

factor as the type of data visualization. Each participant performed 2 evaluations with either PCA (user control, but discrete), grand tour (continuous, but lack user control), or radial manual tour (user control and continuous animation). 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 the geodesic interpolator be used 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. There is a clear need to be able to explain black-box models. Below we use SHAP to extract local explanations which form the projection basis to perform manual tours. We explore how misclassified observations behave relative to near-by correctly classified observations.

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 2 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, the shape of the cluster distributions, and the dimensionality of the data.

The research for RO #3 involves *fundamental visualization 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 on the tour as well as showing the within-class distributions of the SHAP components as parallel coordinate marks on the basis. We also offer a global view and quantitative analysis evaluating the sensitivity of the SHAP-space relative to the sensitivity of the original data space.

5 Work since the mid-candidature review

In the 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 evaluations 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. Participants were crowdsourced from `prolific.co`, were selected from users that had completed an undergraduate degree, and were compensated for their time at £7.50 per hour. Participants were tasked with identifying any and all variables contributing more than $1/p$ to the separation of clusters.

In summary, we use a mixed regression model, to regress accuracy using the factors above as main effects, and use the participant and data simulations as random effects. The random effects capture the variation in the accuracy due to the participant's skill and the random sampling of the data. Several models of increasing complexity were fit, and based on the model metrics we select on the following model to explore the coefficient estimates of in detail.

$$\widehat{Y} = \mu + \alpha_i * \beta_j + \mathbf{Z} + \mathbf{W} + \epsilon$$

where
- μ is the intercept of the model including the mean of random effect
- $\epsilon \sim \mathcal{N}(0,\ \sigma)$, the error of the model
- $\mathbf{Z} \sim \mathcal{N}(0,\ \tau)$, the random effect of participant
- $\mathbf{W} \sim \mathcal{N}(0,\ \upsilon)$, the random effect of simulation
- α_i, fixed term for factor | $i \in$ (pca, grand, radial)
- β_j, fixed term for location | $j \in$ (0_1, 33_66, 50_50) % noise/signal mixing
- γ_k, fixed term for shape | $k \in$ (EEE, EEV, EVV banana) model shapes
- δ_l, fixed term for dimension | $l \in$ (4 variables & 3 cluster, 6 variables & 4 clusters)

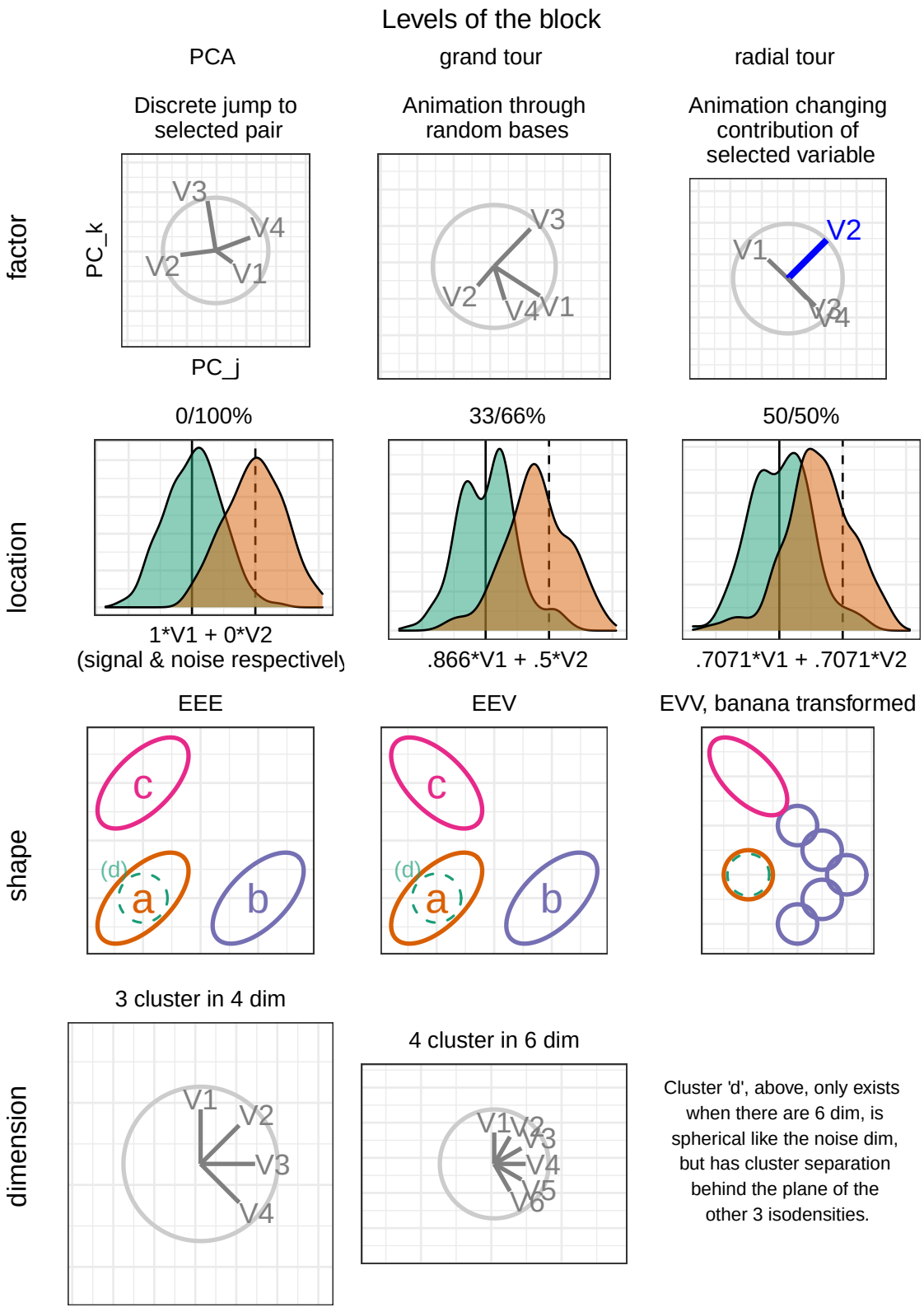

Figure 1: The experimental factors of the study. Visualization is the primary factor of interest. The other factors extend the scope this problem is tested under.

	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	

Figure 2: Model coefficients regressing against our accuracy measure. We have strong evidence supporting a relatively large increase in accuracy with the radial tour.

We were surprised to see that the radial tour is the only marginal term that is significant. Not only does it enjoy the most support, but it also has the largest coefficient estimates as well. Interestingly, the single significant interaction, radial, and location = 33/66% has a large *negative* coefficient, almost completely negating the gains from using radial when a signal and noise variable are mixed at that ratio. It is worth iterating that location, dimension, grand tour, and the intercept do have some evidence of having an effect.

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. We also use the the same mixed regression model to predict log response time where the grand tour has the fastest responses, presumably due to the lack of interaction.

5.2 Extending the interpretation of black-box models with the use of interactive continuous linear projections

Local explanations describe the linear variable weights in the vicinity of an observation for a given model. For a non-linear space, the weightings of the variables change based on the particular location of the explanatory variables. Local explanations are point-measurements of these weights that reveal how important each variable is to model at that particular location. There are several *model-agnostic* local explanations such as LIME(Ribeiro, Singh, and Guestrin 2016), and SHAP(Lundberg and Lee 2017). We will be applying and

discuss SHAP values extracted from a random forest model below. In theory, this application is extensible to many such models and compatible local explanations.

5.2.1 SHAP values; local variable weights and additive prediction explanations.

To introduce the idea of SHAP values, consider FIFA soccer data(Leone 2020). We use 5000 player-observations of 9 aggregate skill measures to predict that player's wages. We use SHAP to observe how the skill attribution changes in the vicinity of players of different fielding positions. The intuition is that a model should change the weighting of the variables to more accurately predict the wages based on the different distribution of skills associated with different positions, despite the model not explicitly knowing the fielding position.

We have trained a random forest model and wish to further explore the weightings of this non-linear model. Following the work in Biecek and Burzykowski (2021) we can similarly extract SHAP values, highlighting that different skills are valued differently across player positions within the model. We also show "break down" profiles, that is additive prediction explanations, how much of each player's predicted wages is added by each of the skill evaluations. The figure below takes a look at the SHAP and break down profiles of a star offensive and defensive player.

5.2.2 Trees of Cheem

Above, we highlight the differing weights across 2 different fielder positions within the same model. It is hard to see where this fits in the full context of the other observations. Below we create a global (all observation) view by approximating the data- and SHAP-spaces in 2d with their first 2 principal components.

To illustrate this we take a look at much simpler data; a simulation of 3 spherical clusters on the vertices of a triangle. The difference between the clusters is contained in the first 2 dimensions with another 2 noise dimensions distributed as unit normal. After extracting all observation's SHAP values, forming a SHAP *matrix*, of the original dimensionality, $(n \times p)$. We want to show a global view of the SHAP matrix and show how it and its sensitivity differ from that of the original data.

We approximate the data and SHAP spaces as the first 2 principal components. We facilitate exploration and interaction by adding a hovering tooltip displaying row number and class (actual & predicted) with linked brushing highlighting selected points and displaying their data tabularly below the plot.

Note that the bulk of the correctly classified points are clustered in relatively small areas. This means that the distribution of their SHAP values are quite similar; the model is selecting very tight variable weightings to explain the predicted class of each observation. Conversely, misclassified points tend to lie in between

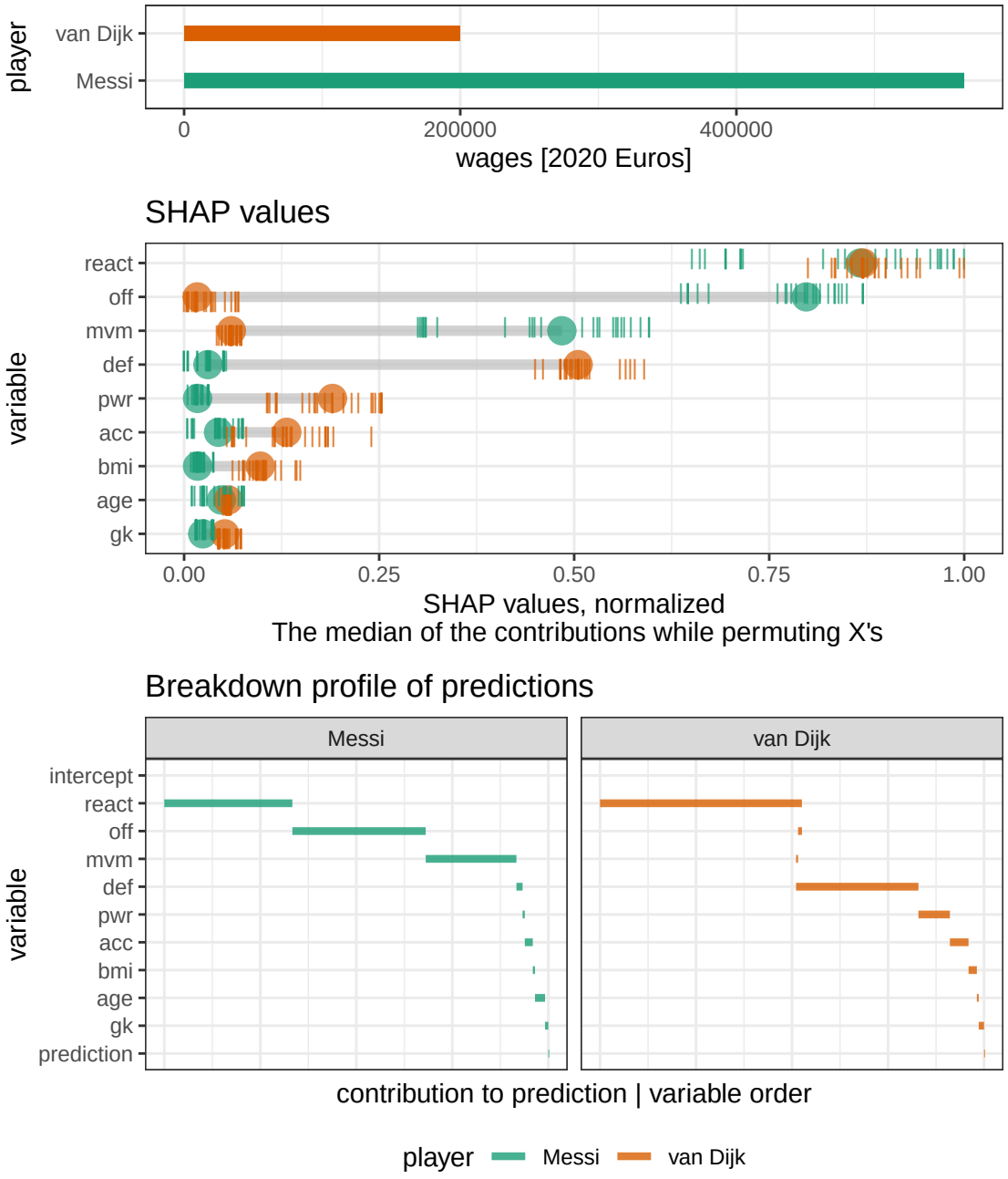

Figure 3: SHAP values and prediction explanations of an offensive player (Messi, top) and a defensive player (van Dijk). SHAP values show a change in weights at the location of each player. Break down profiles show one order-sensitive explanation for the prediction of that observation.

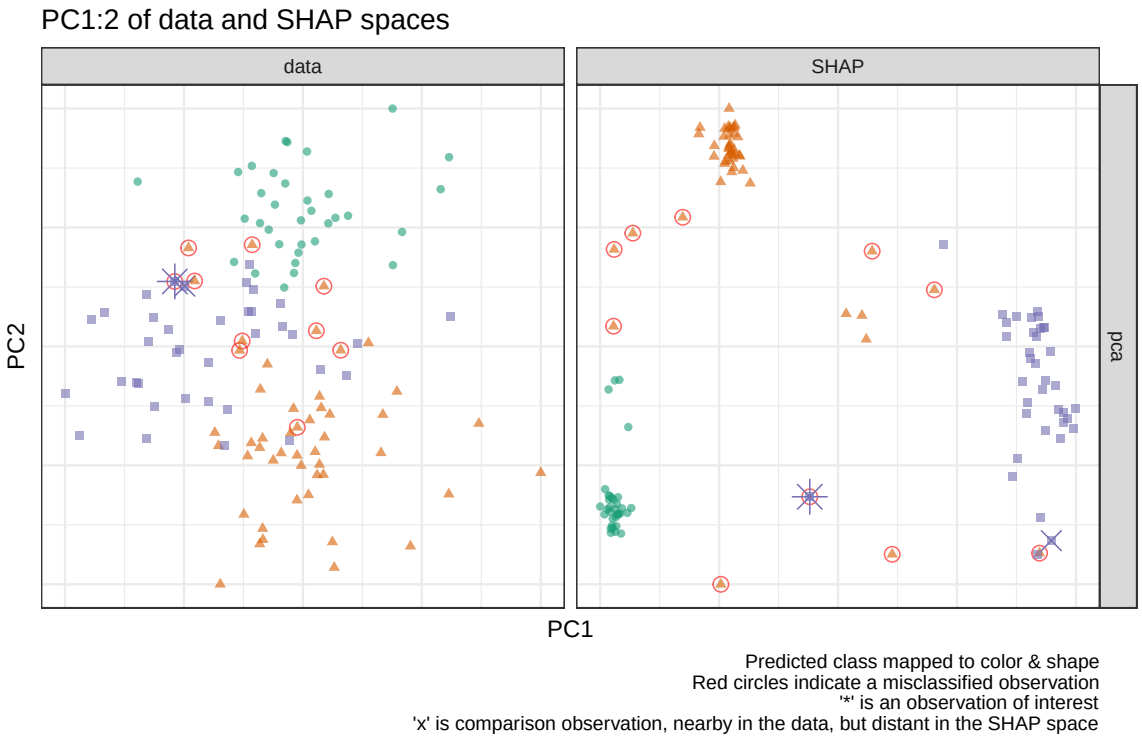

Figure 4: Data and SHAP spaces (top and bottom respectively) of simulated data. The points are colored and shaped according to their predicted class, misclassified points are identified with a red circle. A target observation '*' is shown in comparison with nearby observation 'x'. These same 2 points are tracked in the proceeding tour.

2 clusters of correctly classified points. Using the interactive brushing and hover tool tips we confirm that these points lie between the actual and predicted classes.

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 misclassified while the comparison point is a correctly classified nearby point. These 2 points that are classified differently, but otherwise have very similar values in the explanatory variables, lie quite far apart in SHAP space.

Now we have the global context and an idea of the sensitivity of the SHAP spaces. To further explore an agnostic local explanation by exploring the structure created by the SHAP values of a particular observation. That is, it will be the projection that best puts this observation with its *predicted class* and separate from the others. Consider the selected observation labeled as a star/asterisk above. Its normalized SHAP value will become the starting 1D basis to perform a radial tour exploring the structure. By default, the variable with the largest contribution is selected to be rotated fully into and out of projection. The location of the '*' and 'x' observations are shown as the dashed and fainter, dotted lines respectively on the tour. Their current and initial (more faint, stationary) locations are highlighted throughout the tour.

The application is in progress 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.

5.2.3 Discussion

We have used radial tours to improve the interpretability of black-box models by exploring the structure and distributions of the local explanations. 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 variable 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 & program requirement

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

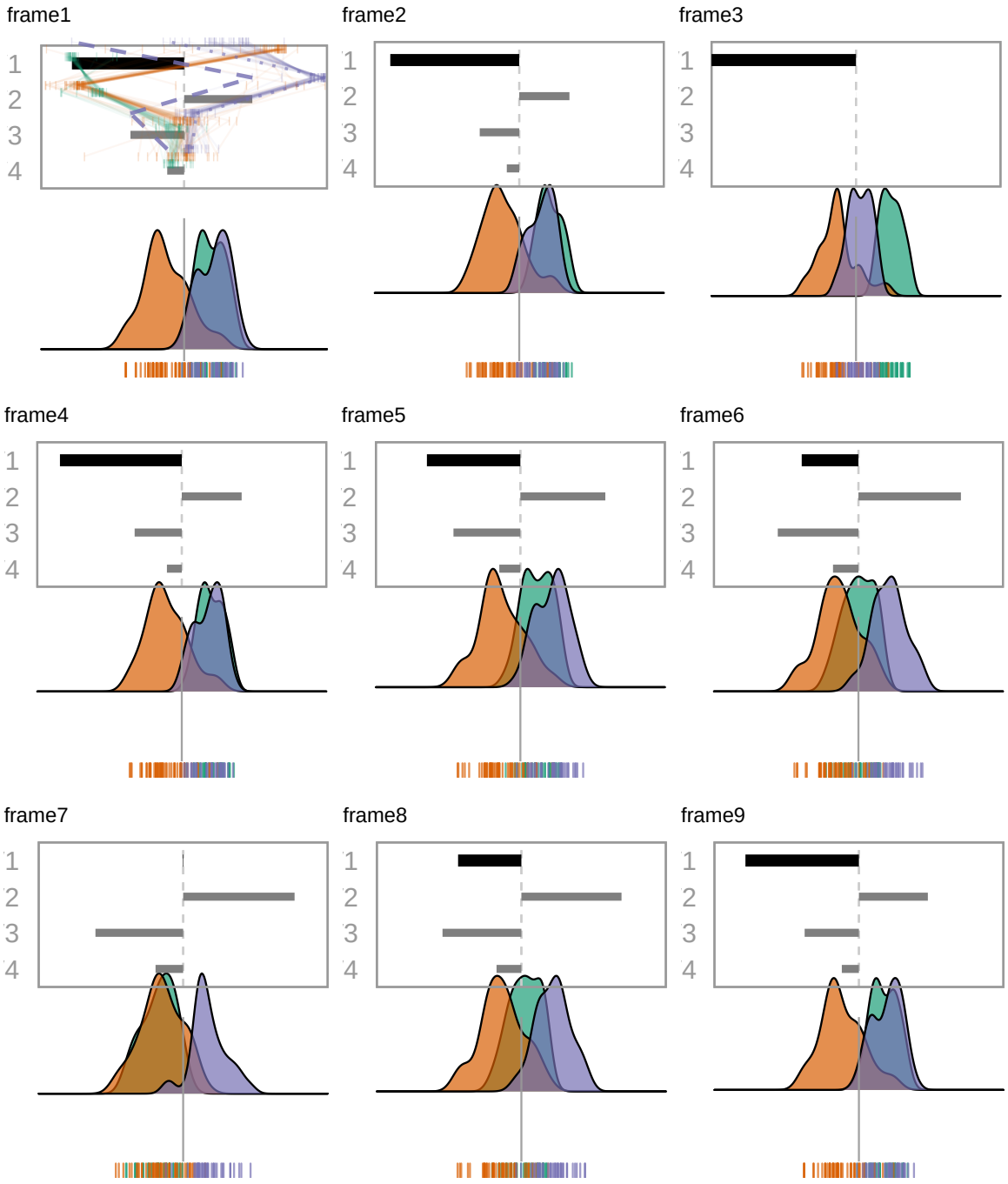

Figure 5: The first frame of the radial tour. The SHAP values of the selected observation set the initial basis, shown as the grey and black bars on top. Within class distributions of the SHAP values are shown as parallel coordinate plots above each variable contribution. The class densities and observation positions of the 1D projection are shown on the bottom. 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. A light grey line shows zero on the projection, with the dashed and dotted lines correspond to the position of primary and comparison observations ('*/x' in the preceding figure)

- Introduction – 60%
- Literature review – 80%
- (RO #1) GI & manual tours – 90%
- (RO #2) manual tour efficacy user study – 80%
- (RO #3) manual tour interpretability, XAI – 70%
- Discussion – 50%
- Conclusion – 30%

The other requirements for this program are complete.

Figure 6 illustrates the purposed timeline for this research.

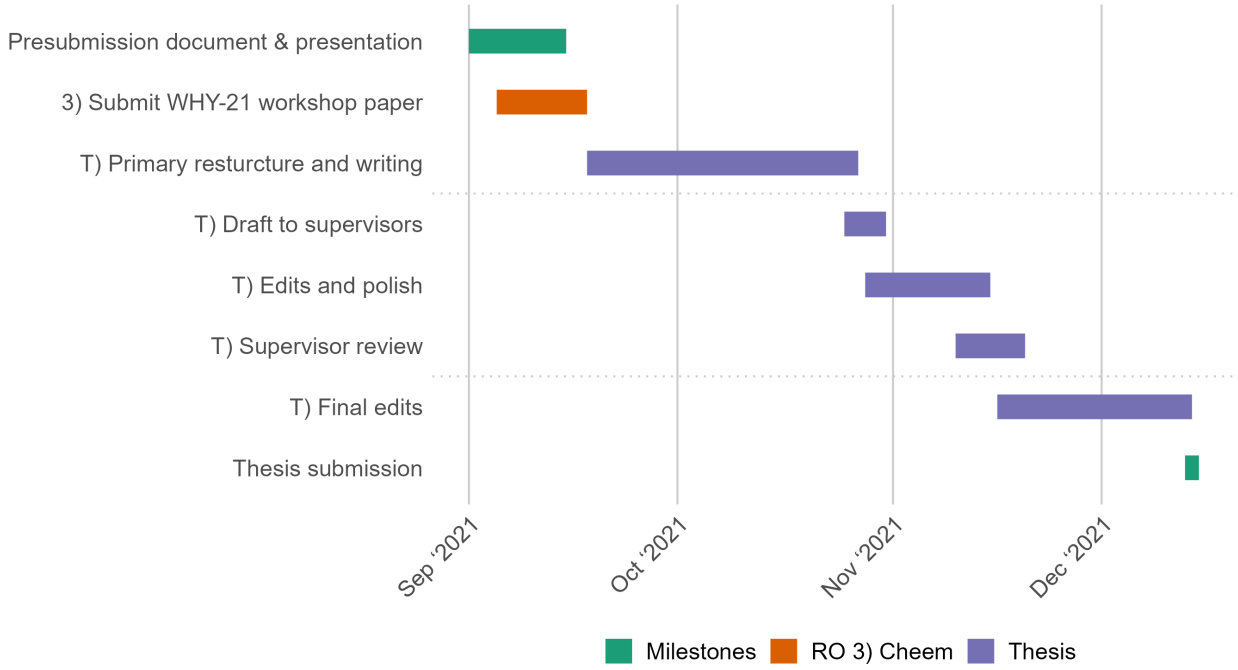

Figure 6: Proposed research timeline.

6.1 Adverse conditions

In addition to the ongoing COVID-19 pandemic, I faced major rework in the second project to accommodate physical distancing. Namely, the user study that was all but ready to be run in person, had to be restructured, to better suit crowdsourcing online. The experimental factors had to be simplified for less knowledgeable people with no interactive help or direction. The shiny apps that hosted the application were not suitably resourced to handle the volume of the crowdsourcing which further exasperated the situation.

Regular circumstances were also obstructed as I took 2 intermissions out of concern for my mental health,

namely anxiety and depression, stemming from my obsessive-compulsive personality disorder.

Intermission periods:

- 13/03/2020 - 08/05/2020
- 15/11/2020 - 01/10/2021

7 Other Contributions

- "Is IEEE VIS *that* good?" AltVis (Spyrison, Lee, and Besançon 2021)
- Student Volunteer, UseR2021 Online
- 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)
- Statistics Ph.D. reading group, Introduction to linear & nonlinear dimension reduction, discussing: "Dimensionality Reduction: A Comparative Review. van der Maaten" 2020
- Student Volunteer, CHI Down Under 2020 Online
- NUMBAT Workshop, Animating ggplot2 figures with gganimate, 2018
- Student Volunteer, UseR2018 Online

8 Acknowledgements

I would like to thank Professor Przemyslaw Biecek for his time and input in suggesting to look at SHAP local explanations and try applying to the FIFA dataset.

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 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.

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/.

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.

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.