# Mouse-Tracking a window intro decision making.

We have argued for mouse-tracking as an excellent tool for obtaining a rich, contextual and continuous measure for cognitive processes, avoiding averaging techniques used in the otherwise similar paradigm of eye-tracking (Spivey et al., 2005). Mouse-tracking analysis procedures are currently very ambiguous, lacking a supported clear analysis pipeline. Such ambiguity has emerged several diverging paths of mouse-tracking analysis, thereby weaking the entire field and making comparative analysis difficult (Schoemann et al., 2021). An attempt to standardize mouse-tracking analysis while still keeping the core flexibility that mouse-tracking contains was conducted by Maldonado et al., (2019). Maldonado’s et.al., (2019) contribution to the field of mouse-tracking analysis consist mainly of linear-discriminant-analysis (LDA) classifier trained on a forced-switch task simulating the expected mouse trajectory following a decision task with a two-step processing. By replicating Dale & Duran's, (2011) study on linguistic negation as two-step process they obtained data for testing their LDA classifier using [Area Under the Receiver Operating Characteristic curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic#Area_under_the_curve) [[1]](#footnote-1)(AUROC) as a metric for performance. We will be investigating the possibilities for applying their framework to a broader scope of the mouse-tracking literature by replicating a known experiment on dynamical processing using mouse-tracking ie. (Spivey et al., 2005) work on decision processing in phonological similarity. Followed up by a discussing advantage and disadvantages of LDA AUROC and First we will walk through the experimental setup and ideas behind (Maldonado et al., 2019), followed up by an introduction to our experimental idea and design.

## Calibration Data:

The first step of Maldonado et al., (2019) were to create a baseline for the LDA classifier to train on. An “ideal” mouse-trajectory a two-step processing decision task were simulated using a forced dual-choice switch task. The experiment is initiated with two possible answer boxes (blue) & (red) and a frame at the edge of the window with the same possible colors. The job of the participants is to repeatedly select the box with the matching color of the frame at the edge across the 88 trials. To induce an element that would simulate a two-step processing task twenty-four of the eighty-eight trials had a “switch” condition with the remaining being “straightforward” with no manipulation. “Switch” trials involved the frame changing color to the opposite, forcing participants to change mouse-trajectory, and thereby simulating a switch decision making. Point of change were dependent on the y-coordinate which varied between 40%, 70% and 90% of the way to the target on the y-axis allowing for the LDA classifier to have data with varying temporal aspects of the switched decision (Maldonado et al., 2019).

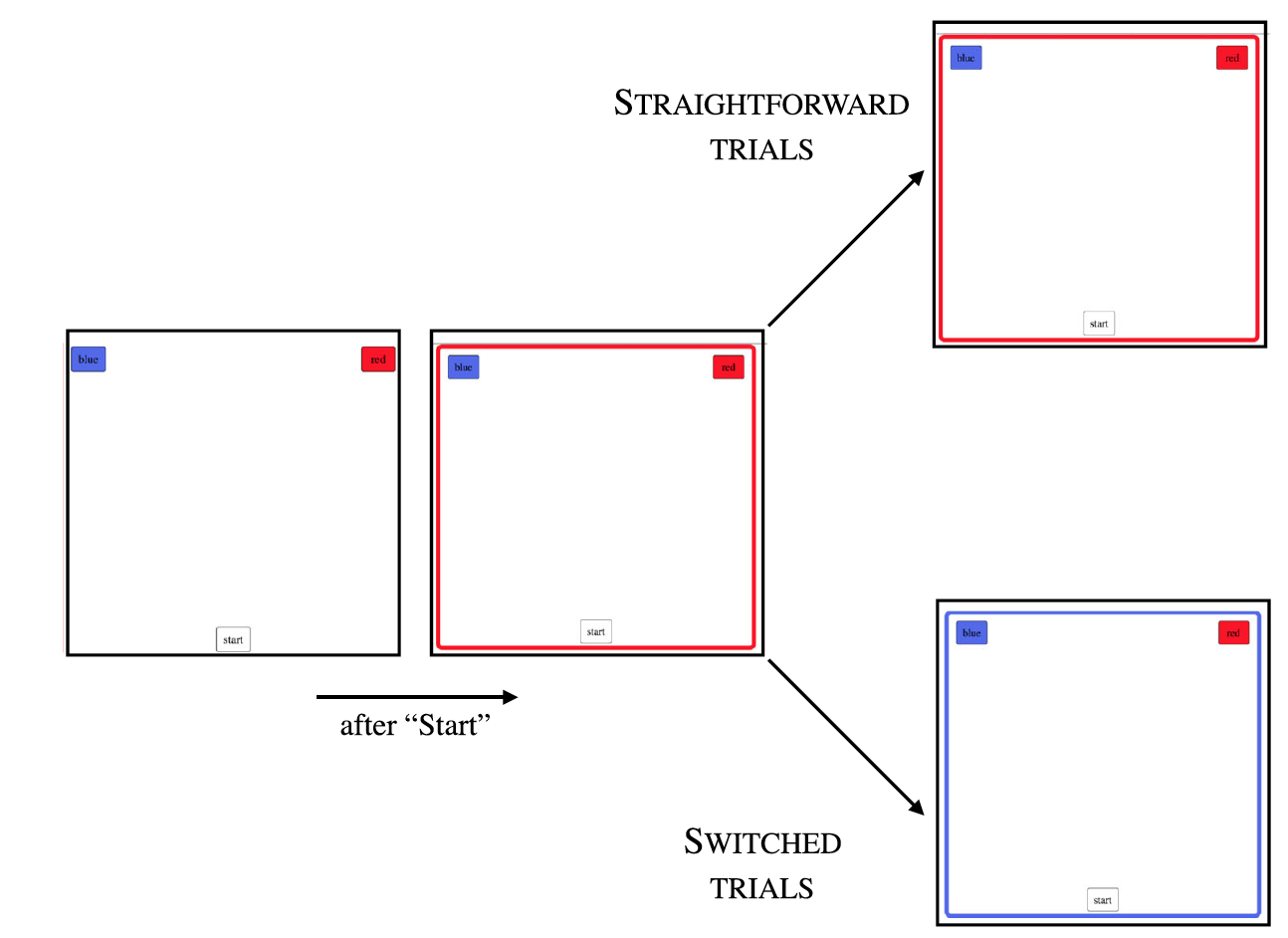


Figure 1 (Maldonado et al., 2019), Experimental design of the Calibration experiment.

### Training of LDA on calibration data.

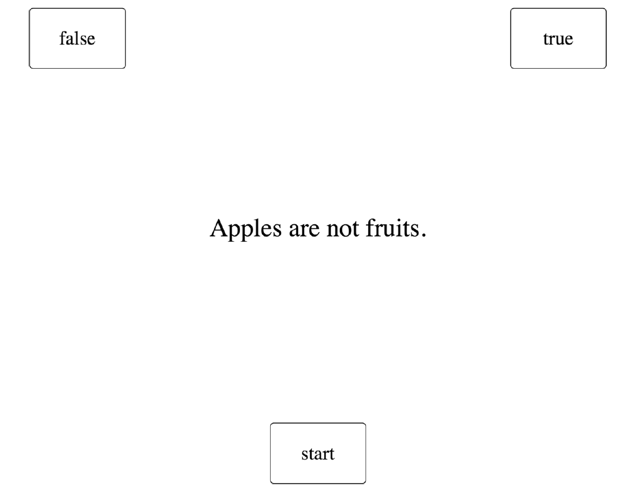
A pipeline for training the LDA classifier is as follows, **1)** Normalize and flip (X, Y) coordinates by setting start point = (0,0), end point = (1,1). Based on the spatio-temporal information calculate the non-linear Euclidean velocity and acceleration. Finally, time-normalize the data into 101 steps. **2)** Even though Maldonado et al., (2019) didn’t report it in their paper we ran through their analysis script and found them to standardize all features based on the Z-distribution with a mean = 0 and SD = 1 prior to the conducting of the principal component analysis (PCA) (*Importance of Feature Scaling*, u.å.). This is important since is gives equal weight to all features when calculating the covariance matrix used for PCA. The PCA removed any collinearity which otherwise would be an issue for any regression or classification model, leaving us with 13 PCA features to feed to the LDA classifier. 3) The 13 PCA features fed to the LDA classifier were based on a varying X matrix containing one of the combinations showed in table 1. An in-depth review of LDA and PCA will be take place later in the article. By comparing their classifiers with different PCA features extracted from the varying combination of predictors shown in table 1 were Maldonado et al., (2019) able to gather information on which mouse-features were the most important for classifying decision processes. A 10-fold cross-validation(cv) on all their models concluded that the most important feature for classifying “straight” vs. “switched” trial were the coordinates.

Table 1 (Maldonado et al., 2019)

Et billede, der indeholder bord

Automatisk genereret beskrivelse

## Extending the LDA classifier to negation data:

For testing whether the classifier is a useful tool for analyzing mouse-tracking data it must generalize beyond the “simulated” decision data in the forced switch task. To test the generalizability of their classifier Maldonado et al., (2019) had to test it on more ecological mouse-tracking data. They choose to replicate Dale & Duran, (2011) study on sentence negation with an hypothesize that if sentence negation followed the same two-step processing as in the forced calibration task the LDA performer would perform similarly. Participants were with a screen like figure 2. A true or false sentence would be presented in the middle of the screen either including or excluding a negation, i.e., “elephants are not small” or “elephants are not large” (Dale & Duran, 2011). The job of the participants was to evaluate the statement and select the corresponding false or true statement. Maldonado et al., (2019) managed to replicate the original studies finding of a significant increase in x-flips when the statement was truth and included a negation.

Having verified the replication of the original Dale & Duran, (2011) findings the LDA classifier could now be tested to evaluate its generalizability and advantages.

Figure 2, Dale & Duran, (2011) Negation study.

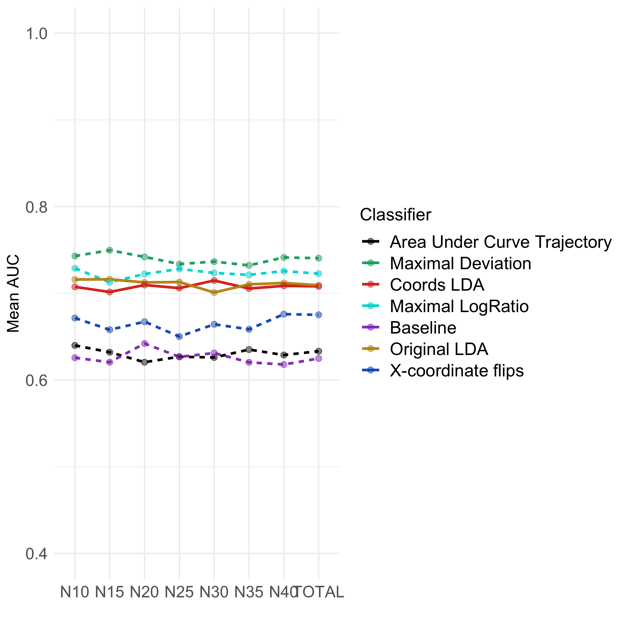
All data were subset to only include the true statements condition as the negation only influenced mouse trajectories in said condition. Wrong responses were also excluded. Remaining data were pre-processed in the same manner the calibration data allowing for a trained model on the calibration data to be tested on a similar feature matrix in the test data. Bootstrapping was conducted on various sample sizes which was then used for testing the LDA. The results showed a similar pattern to that of the calibration. Coordinates were the most predominant predictor of those included in the LDA classifier following AUROC as the metric for performance, Maximal Deviation (MD) and Maximal LogRatio both performed slightly better than the LDA classifier. However, Maldonado et al., states, *“The classifier is still a better choice from a conceptual point of view, since it does not make any specific assumptions about how the change of decision should be reflected by mouse trajectories beyond the observed ones.”* In the following section an argumentation for and against this statement will be presented.

Figure 3, (Maldonado et al., 2019) Results of CV LDA performance on different bootstrap samples.

# LDA & PCA/ (methodology of Maldonado et al.,

LDA has originally been thought of as a feature extractor through dimensionality reduction. Feature extraction is especially useful when working with a high dimensionality data set such allowing for the same information to be conveyed in lower sub dimensional space which maximizes separability, where Separability is defined in terms of deviation in statistical measures of mean value and variance. The technique has been employed in a broad range of research areas i.e. EEG (Subasi & Ismail Gursoy, 2010), face-recognition (Jin et al., 2001), medical bioinformation recognition (El-Feghi et al., 2004). A more recent employment has been to use LDA as a classifier by testing the separability of different classes mean and variance on the new LDA subspace. While this technique is not novel (Kim et al., 2007) or pioneered by Maldonado et al., (2019) it is novel to the field of mouse-tracking. By doing so does Maldonado et al., (2019) introduce a machine-learning approach for evaluating mouse data, and testing hypothesis. As highlighted in an earlier section the predominant mouse-tracking methodology has largely been occupied by multiple t-test and regular hypothesis testing on a varying selection of features. Some of the most common being t-test of x-flips, MD, AUC, Maximal LogRatio and X-coordinates in specific time-intervals. Regardless of research field of inquiry, be it linguistic negation or phonological similarity, the trend has been to look for significant difference in MD, AUC etc. between conditions. A significant difference in means would then support the hypothesis of the experimental manipulation having an effect. There is two main issues with such an approach which both can be summed up the a lack of reproducibility and best-practice methodology. 1) One of the fallacies of ordinary least squares is the dichotomous thinking of significance testing (Dienes, 2008). Just because a test is significant is not the same as the magnitude of difference being larger. One of the most common misinterpretations is that smaller p-values equals a larger effect which also leads to increased focus on p-values. An un-scientific tradition of p-hacking[[2]](#footnote-2) has started to emerge across scientific-fields (Head et al., 2015) as an off-spring from the increased focus on p-values has. One of the major assumptions of hypothesis testing is that the hypothesis and planning of statistical test must take place prior to any collection or investigation of data. Even though such an assumption is difficult to control for, is there specific fields with larger way room when it comes selecting their analysis. Making it even more important in those fields to state specific hypothesis and plans for testing prior to data investigation, which leads us to the second issue in mouse-tracking analysis.

2) One of the main advantages of mouse-tracking is its flexibility, as simulated in the calibration experiment temporal patterns can vary from experiment to experiment, the temporal and spatial aspects when the trajectory changes can vary. The change can be early or late both temporally and spatially, with either a smooth change (dynamical) or sharp and edgy (dual system). Different differences result in changing patterns. One experimental condition could influence AUC greatly but not x-flips and vice-versa. Mouse-tracking analysis can accommodate these differences by comparing different measures, MD, AUC, Maximal LogRatio, X-flips or temporal differing X-coordinates but such flexibility also results in varying analysis and a lot of options of p-hacking. By introducing a machine-learning-like analysis apparatus for mouse-tracking will we reduce the options for selectively only comparing x-coordinates slices or other mouse-tracking measures that are different in a specific experimental setup. Rather by creating a “standardized” model trained on simulated data and tested on experimental-specific data will we be able to avoid dichotomous thinking and p-hacking while also introducing an easy way to compare results across papers.

Maldonado et al., (2019) has chosen LDA as for classification with PCA for feature extraction prior to model training and testing. Combinatory effect dubbed PCA plus LDA, has been test and validated in multiple experimental settings (Liu & Wechsler, 2000)(Yu & Yang, 2001)(Belhumeur et al., 1997). High dimensional data is computationally expensive for LDA, a PCA reduction will therefore improve the computationally expensive of LDA, but the question is how much information is lost in process. This is dependent on multiple things, among others the proportional amount of PCA n-components being used compared to the total available. Yang & Yang, (2003) asked the questions *“Why select PCA for dimensionality reduction beforehand? Is there any important discriminatory information lost in the PCA process since the criterion of PCA is not identical to that of LDA?*” (Yang & Yang, 2003). They concluded that stacking PCA upon LDA did not result in any additional significant information loss. To sum up, there is support in statistical academia both towards moving from hypothesis testing to a machine-learning framework and for the application of PCA plus LDA. But currently has the Maldonado et al., (2019) framework only been tested on mouse-tracking data from a linguistic negation experiment. In order to implement such framework in the broader field of mouse-tracking the model has to generalize to other areas of research and be robust against changing temporal and spatial mouse-patterns.

1. To evaluate if the LDA successfully generalize it must perform close to the topline and well above chance following AUROC measure of success.
   1. Perform better than traditional mouse-tracking measures.
2. Is 13 PCA features the optimal amount.

1. For information on AUROC https://towardsdatascience. com/understanding-auc-roc-curve-68b2303cc9c5. [↑](#footnote-ref-1)
2. For a quick explanation of p-hacking see [this](https://scienceinthenewsroom.org/resources/statistical-p-hacking-explained/). [↑](#footnote-ref-2)