Automated prediction of drinking categories in monkeys undergoing chronic alcohol self-administration

Aleksandr Salo

Machine Learning by Dr. Hamerly

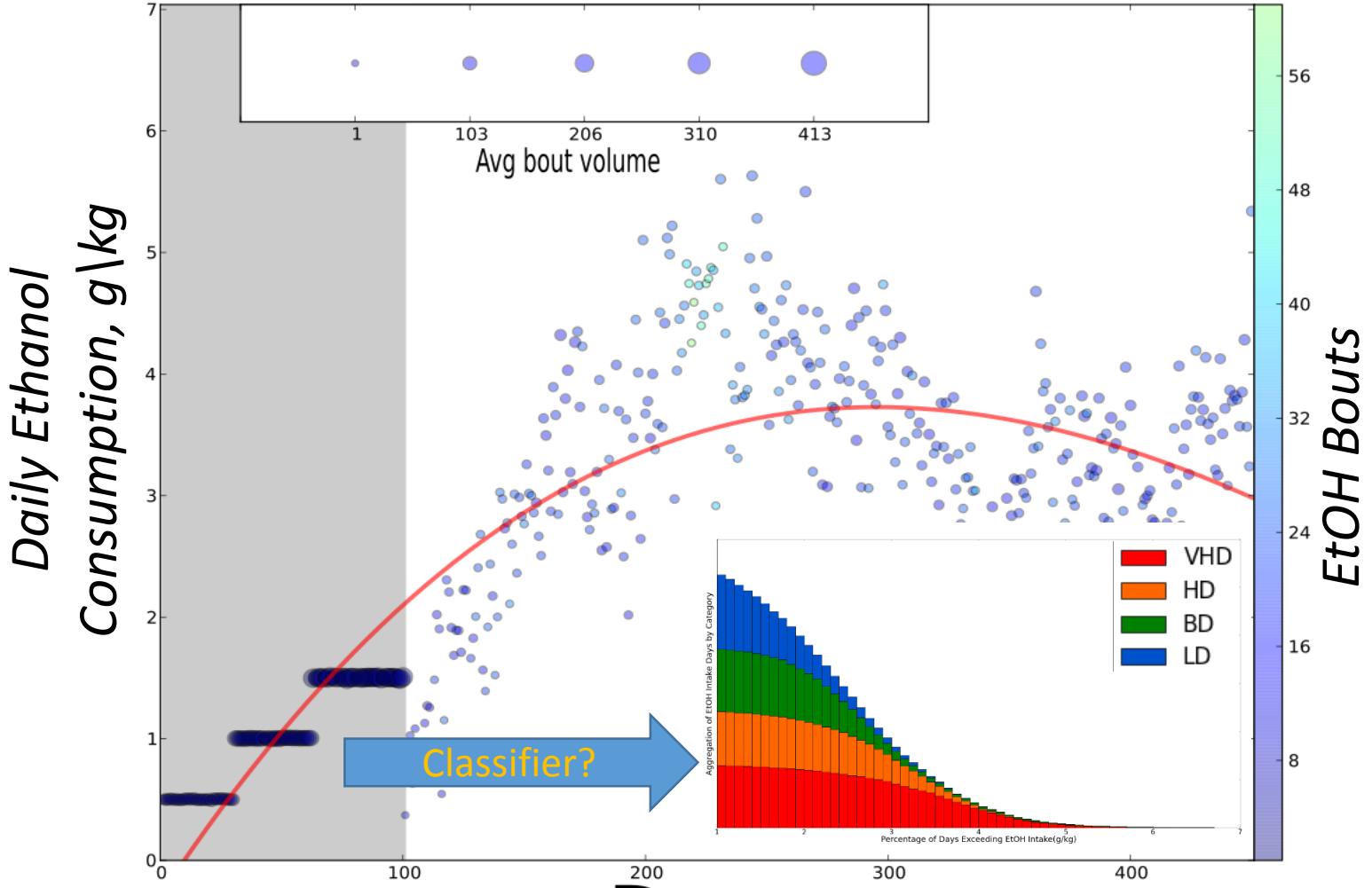
Baylor University 2015

BAYLOR UNIVERSITY

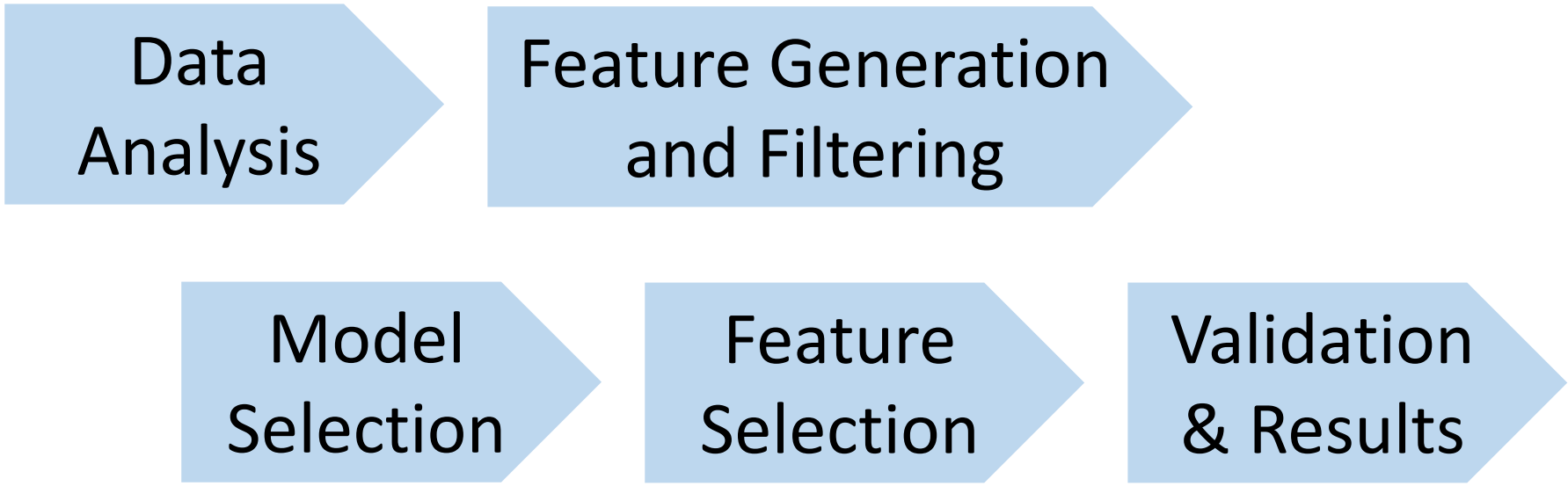
Goal

- Classification model:

  F(Induction Data) -> Open Access Drinking Category [1]



Approach


Data Analysis
Feature Generation and Filtering
Model Selection
Feature Selection
Validation & Results

Data Analysis

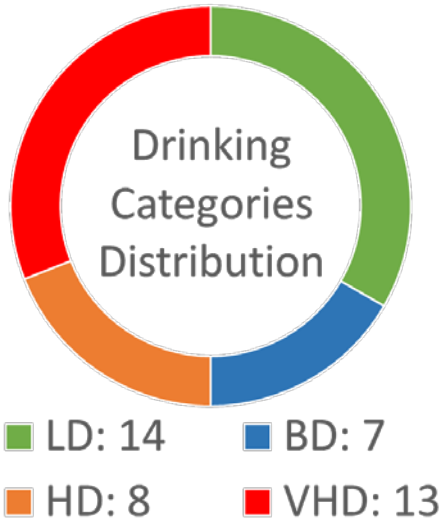
Data Analysis → Feature Generation and Filtering → Model Selection → Feature Selection → Validation & Results

6 different cohorts (INIA Rhesus)[4]

42 monkeys (11 females)

Mean age of first intoxication 6.14 (sd=1.64) years

Mean weight 9.30 (sd=1.1) kg



Natural: *Potential Features (n~30)*

- Gender
- Age of intoxication

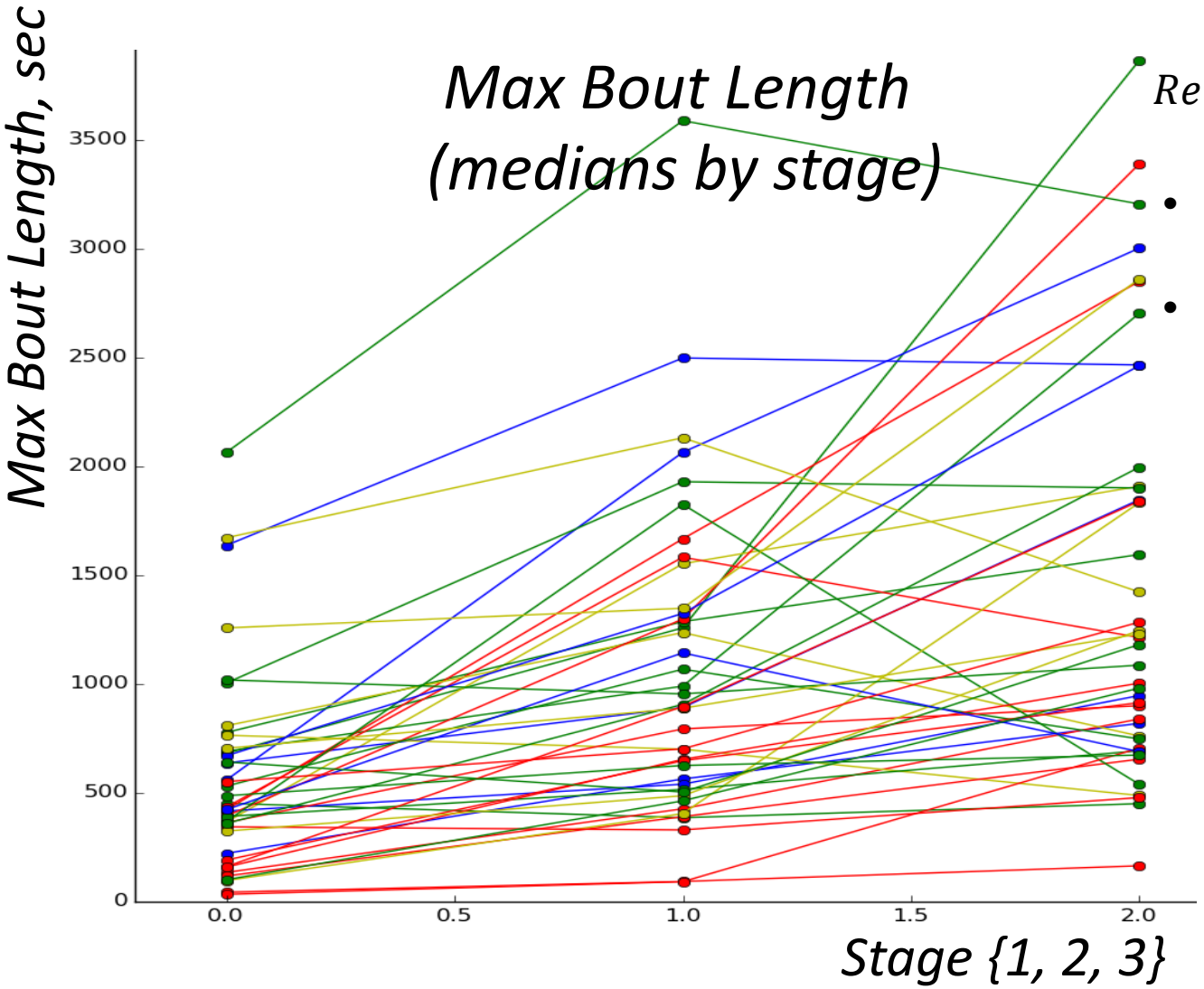
Derived from induction[4] period:

- Latency to first drink (time)
- Total number of EtOH bouts
- Total number of H2O bouts
- Mean length of EtOH drinks[i]
- Mean volume of EtOH drinks
- Mean bout duration
- Seconds it took for monkey to reach day's ethanol allotment
- Length of the maximum bout (bout with largest ethanol consumption)
- Ethanol consumed during first 10 minutes as a percentage of the daily allotment

[i] Less than 5 seconds between consumption of EtOH is a continuous drink

Feature Generation

Data Analysis → Feature Generation and Filtering → Model Selection → Feature Selection → Validation & Results



$$Rel_i^{delta} = \frac{M(At_i[stage_3])}{M\ (At_i[stage_1])}$$

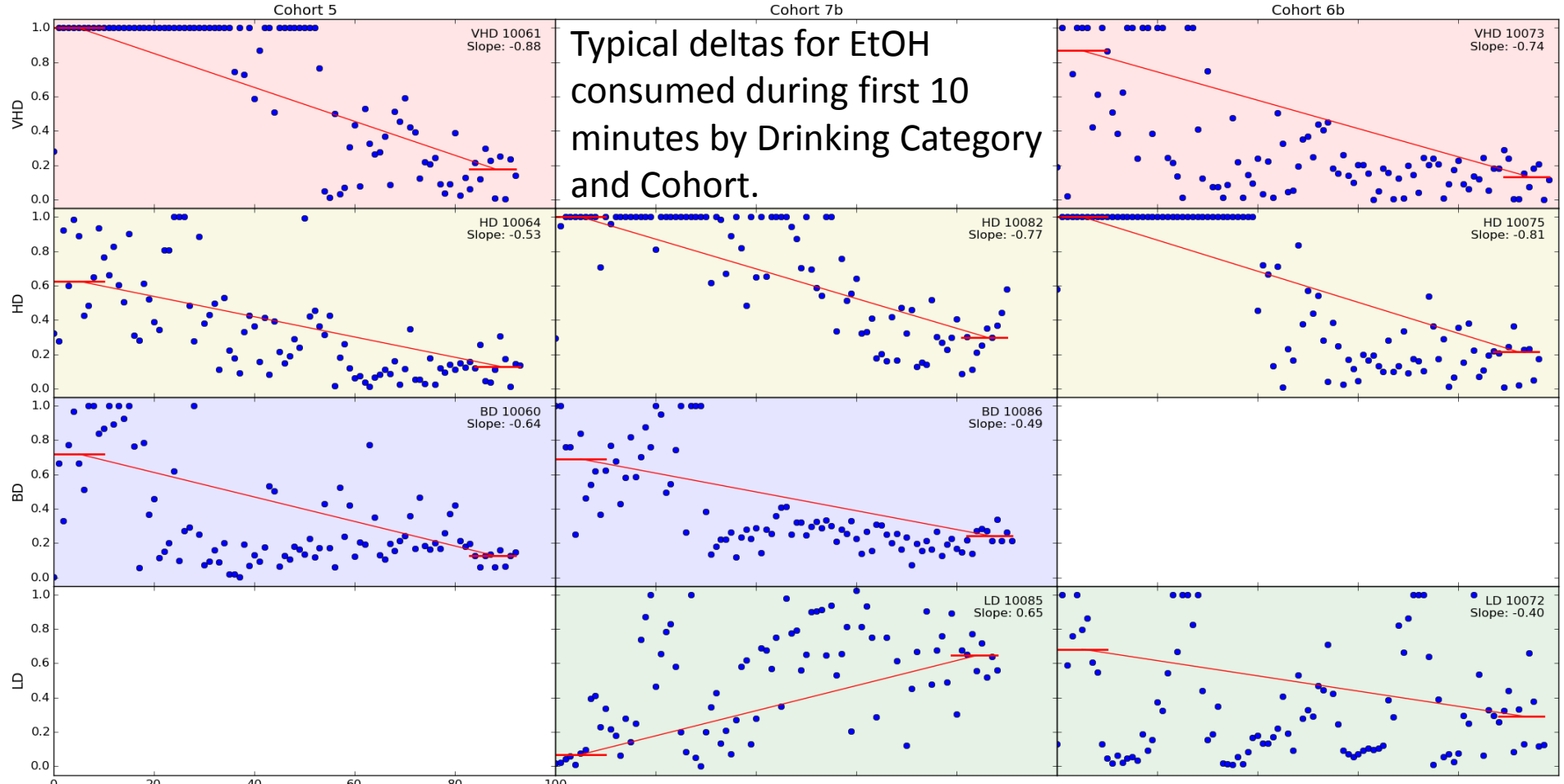
- Using 10 first – 10 last days instead
- ~30 comparable normalized deltas

Feature Filtering

Data Analysis → Feature Generation and Filtering → Model Selection → Feature Selection → Validation & Results

Tests for potential features: ~11 significant features

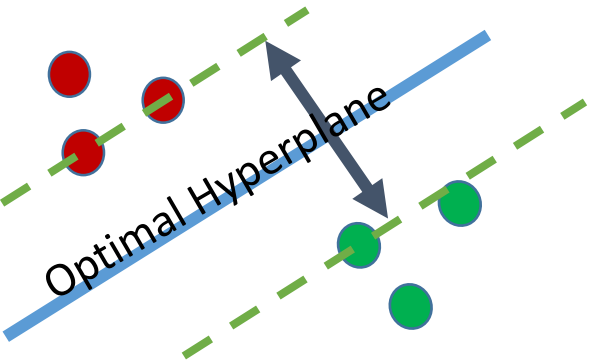
Significance Tests (α=0.05) → Mutual Information Analysis

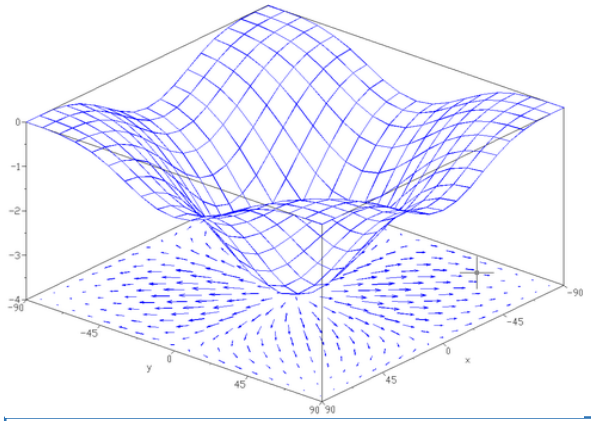


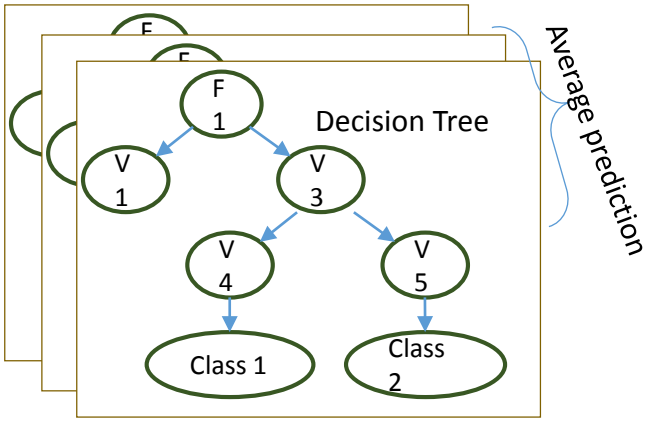
Typical deltas for EtOH consumed during first 10 minutes by Drinking Category and Cohort.

Model Selection

Data Analysis → Feature Generation and Filtering → Model Selection → Feature Selection → Validation & Results







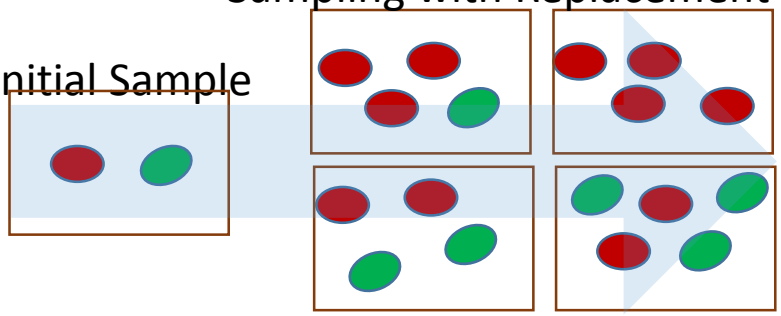
Support Vector Machine	**Gradient Boosting Classifier**	**RandomForest Classifier**
Linear Kernel C = 3 Class weights	Max features = 0.4 Max depth = 5	N_estimators = 20 Max features = 0.4 Bootstrap = True
Accuracy*: .51 (SEM=.08)	**Accuracy: .55 (SEM=.1)**	**Accuracy: .50 (SEM=.11)**

*Accuracy was measured with K-Fold Cross Validation (K=14). Detailed performance analysis is in validation section

Bagging Wrapper

Unethical and costly to get bigger N

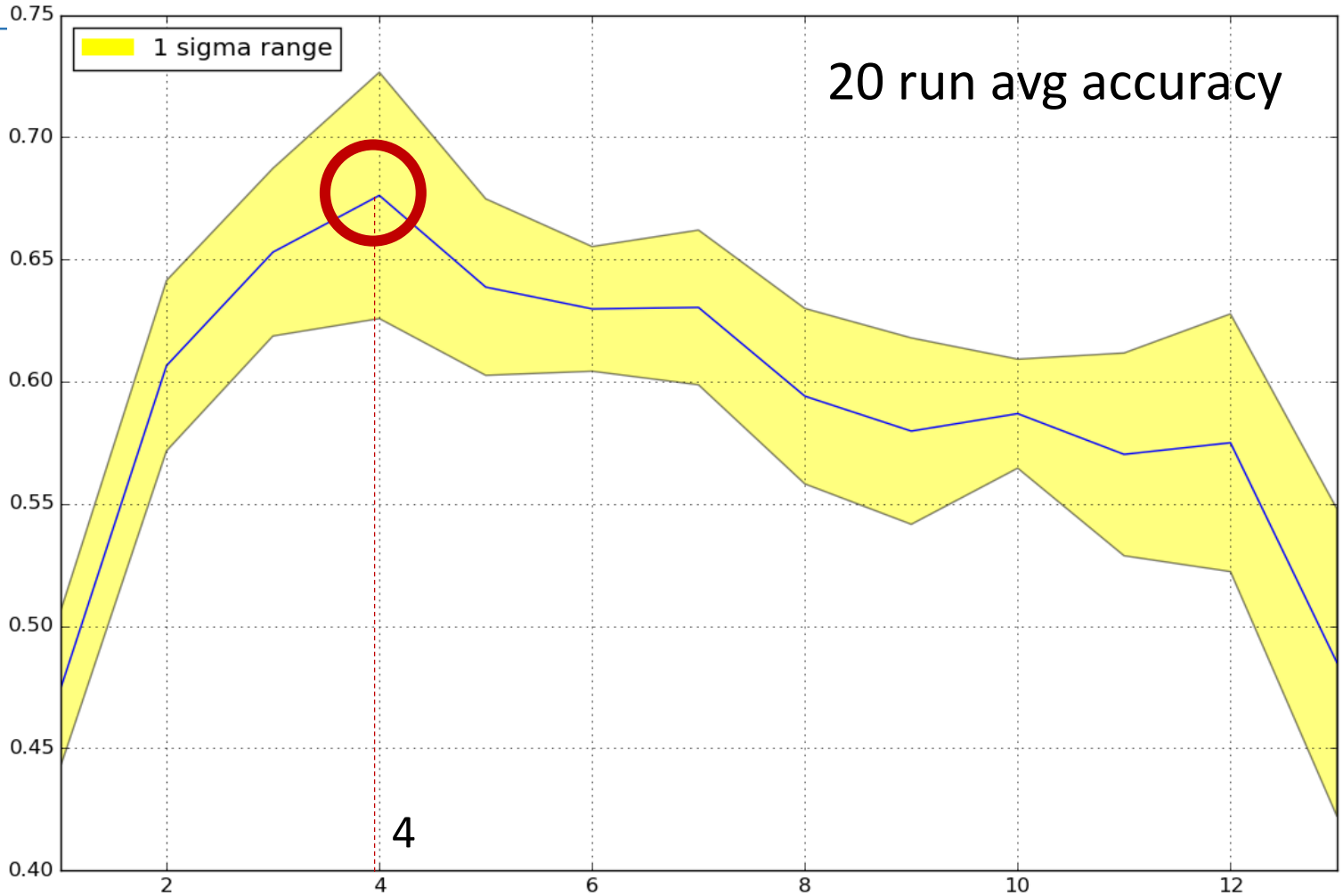
Sampling with Replacement



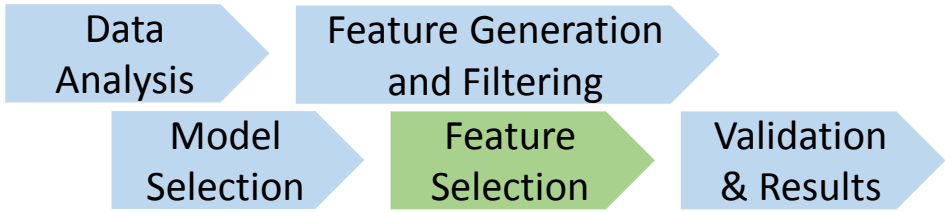
Accuracy: .61 (SEM=.07)

Feature Selection

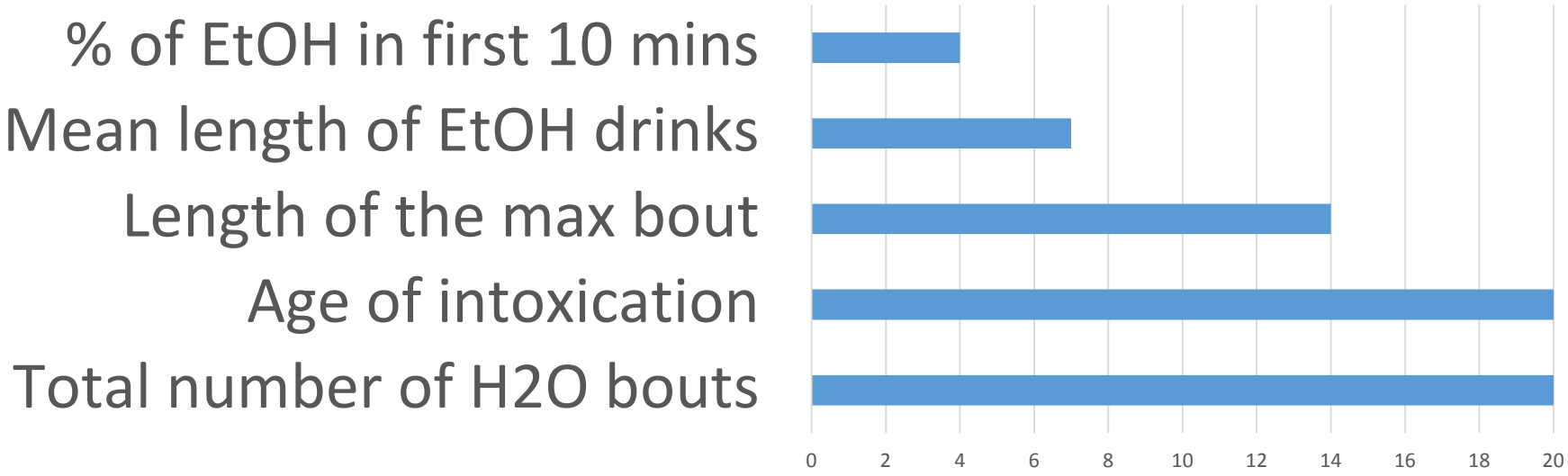
To fight the curse of dimensionality [3] - "forward selection" [1] filtering method that ranks features by their impact on model performance (measured using 14-fold CV).



Feature Selection (cont)



Then scoped the top-5 features that appeared to be used within the best CV across 20 different runs:



Validation & Results

Data Analysis → Feature Generation and Filtering → Feature Selection → Model Selection → Validation & Results

Repeated ShuffleSplit & 4-Fold Cross-Validation.

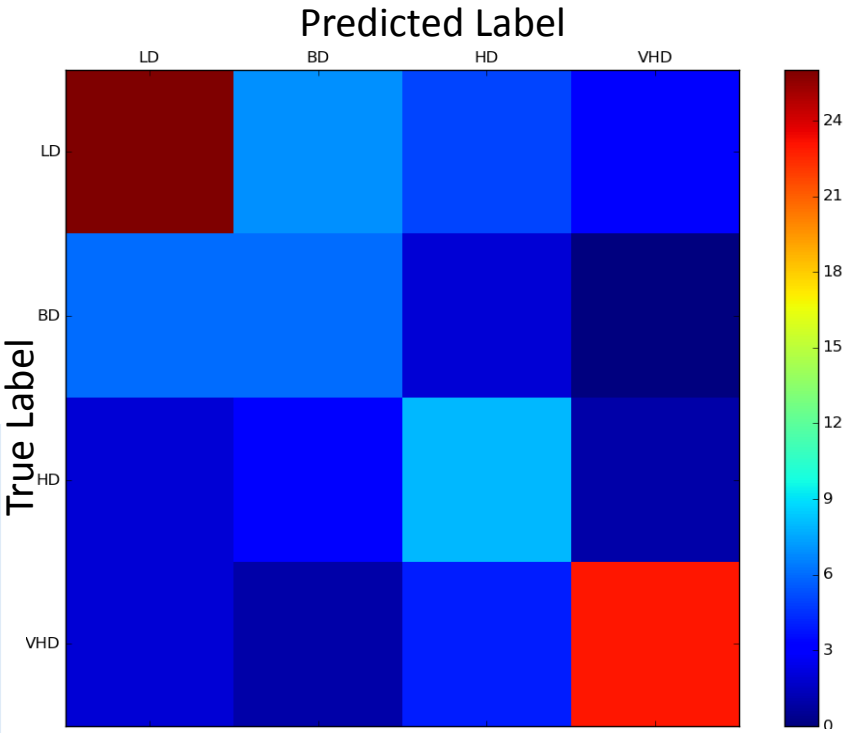
Accuracy: .68 (SEM=.03)

Balanced Error Rate: 0.39

$$BER = 1 - \frac{1}{k}\sum_{i}\frac{A_{ii}}{\sum_j Aij}$$

This is one minus the average recall, treating each class evenly, regardless of its class membership. [7].
BER is employed to compensate for large asymmetry in the data set., i.e. if 95% of data points are non-drinkers, and 5% are drinkers a degenerate classifier that assigns the majority class label to all points will lead to 95% accuracy. Seemingly very good, but is completely uninformative. [6]

Base case: *Accuracy = 0.33. BER = 0.75 (decrease of 52% in balanced error rate).*

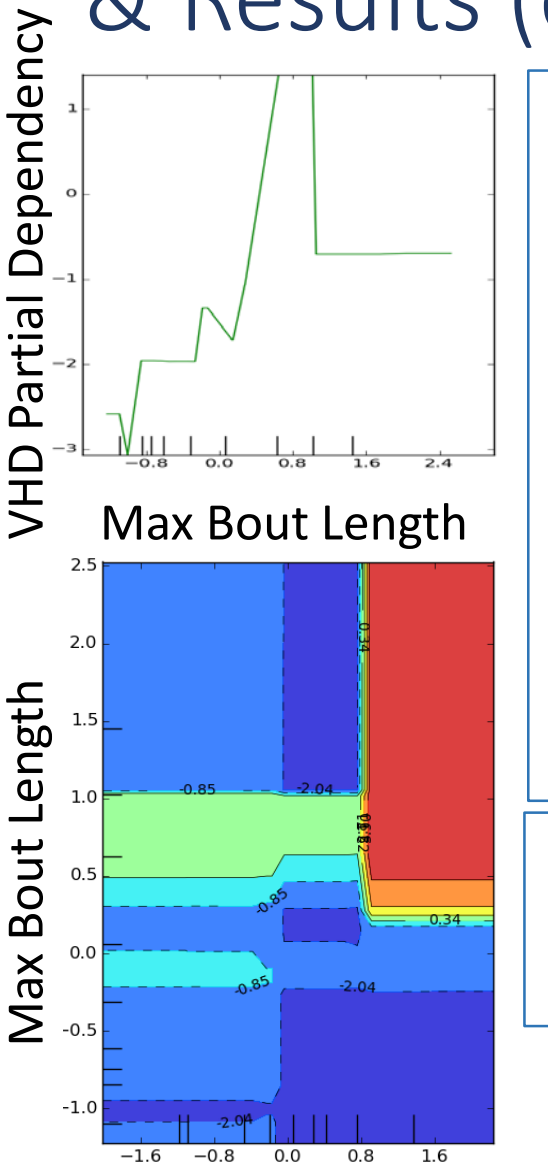


Aggregated Confusion Matrix

LD	26	7	5	3
BD	6	7	2	0
HD	2	3	8	1
VHD	2	1	4	23

Validation & Results (cont)

Data Analysis → Feature Generation and Filtering → Model Selection → Feature Selection → Validation & Results



Given overall accuracy, our results suggest that there is a linear relationship between the increase of max bout length and chances of becoming VHD. That is, during the induction phase of the NHP model, future heavy drinking animals demonstrate significantly longer ethanol bouts. Moreover, if the animal primate increases the number of H2O bouts (or rather make them shorter) then there is an increase in the probability that they will later be classified as very heavy drinkers (VHD). Cohesively, inverse relationships hold for low drinking animals (LD).

In addition, our approach demonstrates a classification accuracy of 0.81 (SEM=0.05) when only two classes ([LD+BD]&[HD+VHD]) are evaluated.

References

[1] Ethem Alpaydin. Introduction to machine learning. MIT press, 2014.

[2] Erich J Baker et al. Chronic alcohol self-administration in monkeys shows long-term quantity/frequency categorical stability, 2014.

[3] David Barber. Bayesian reasoning and machine learning. Cambridge University Press, 2012.

[4] James B Daunais et al. Monkey alcohol tissue research resource: banking tissues for alcohol research, 2014.

[5] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. Annals of statistics, pages 1189-1232, 2001.

[6] Andrew Rosenberg. Automatic detection and classification of prosodic events. Columbia University, 2009.

[7] Bjorn Schuller at al. The interspeech 2011 speaker state challenge. In Proceedings INTERSPEECH 2011.

[8] L. Breiman, “Random Forests”, Machine Learning, 45(1), 5-32, 2001.

[9] L. Breiman, “Pasting small votes for classification in large databases and on-line”, Machine Learning, 36(1), 85-103, 1999.

QA