Predicting ALS Progression with Bayesian Additive Regression Trees

Lilly Fang and Lester Mackey

November 13, 2012

RECOMB Conference on Regulatory and Systems Genomics

The ALS Prediction Prize

- **Challenge:** Predict progression of ALS over time
  - Distinguish fast from slow progressors
- **Measure:** ALS Functional Rating Scale (ALSFRS)
  - Score ranges from 0-40
  - Based on 10 questions (Speech, Dressing, Handwriting, …)
  - Rate of progression = slope of ALSFRS score
- **The Data**
  - 918 training + 279 test patients
    - 12 months of data (demographic, ALSFRS, vital statistics, lab tests)
    - Time series: roughly monthly measurements
  - 625 validation patients
    - Given first 3 months of data
- **Goal:** Predict future ALSFRS slopes for validation patients
  - Error metric: Root mean squared deviation (RMSD)

Outline

- **Featurization**
  - Static Data
  - Temporal Data
- **Modeling and Inference**
  - Bayesian Additive Regression Trees
- **Evaluation**
  - BART Performance
  - Feature Selection
  - Model Comparison

Featurization

- **Goal:** Compact numeric representation of each patient
  - Features will serve as covariates in a regression model
  - Most extracted features will be irrelevant
  - Rely on model selection / methods robust to irrelevant features

Featurization

- **Goal:** Compact numeric representation of each patient
  - Features will serve as covariates in a regression model
  - Most extracted features will be irrelevant
  - Rely on model selection / methods robust to irrelevant features

- **Static Data**

49

- Demographics Age, Race, Sex
- ALS History Time from onset, Site of onset
- Family History Mother, Father, Grandmother, Uncle…
- ……………………
- ……………………

Categorical variables encoded as binary indicators

Featurization

- **Goal:** Compact numeric representation of each patient
  - Features will serve as covariates in a regression model
  - Most extracted features will be irrelevant
  - Rely on model selection / methods robust to irrelevant features

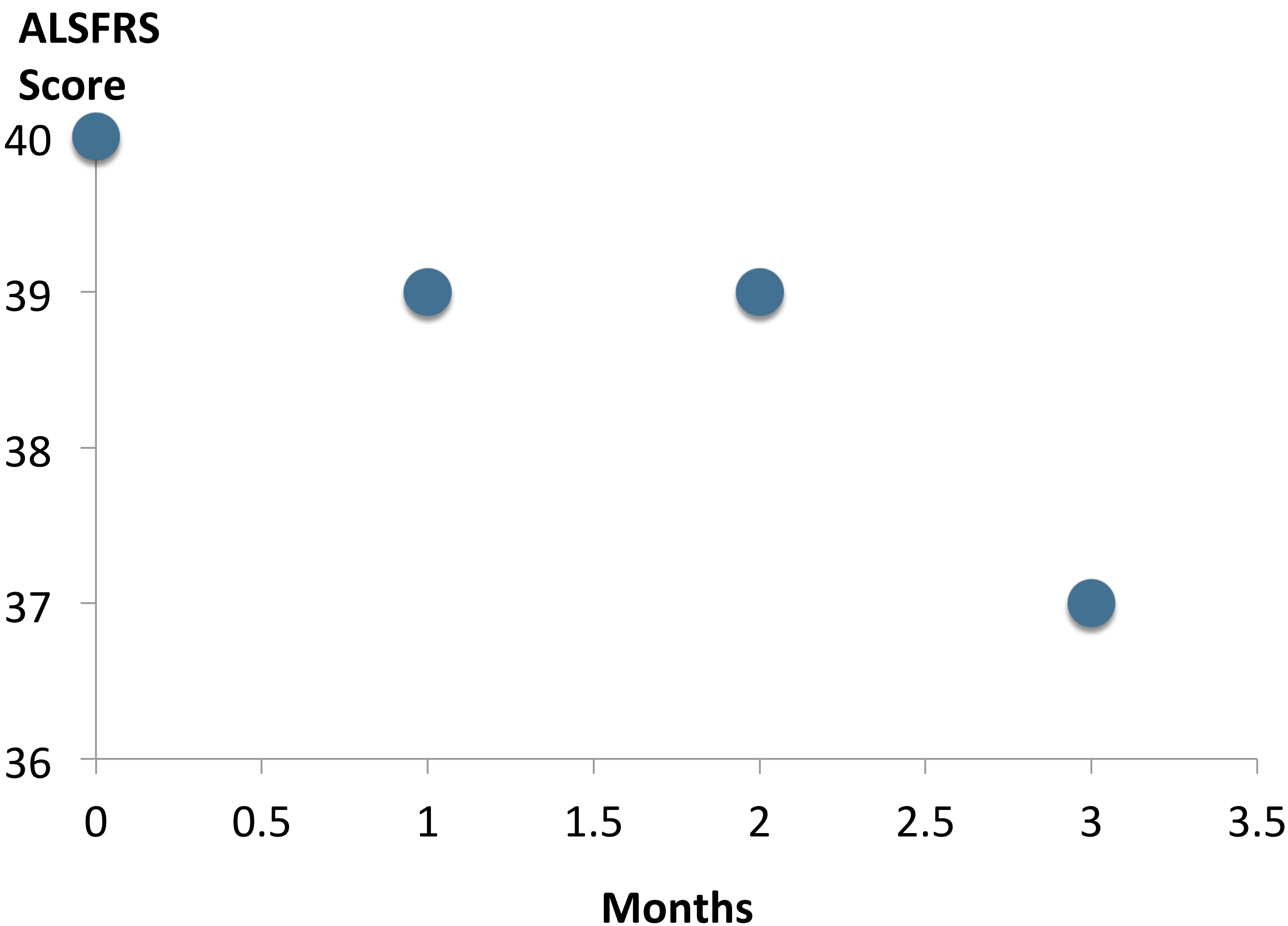
- **Time Series Data**
  - Repeated measurements of variables over time
    - ALSFRS question scores
    - Alternative ALS measures (forced and slow vital capacity)
    - Vital signs (weight, height, blood pressure, respiratory rate)
    - Lab tests (blood chemistry, hematology, urinalysis)
  - Number and frequency of measurements vary across patients

Featurization

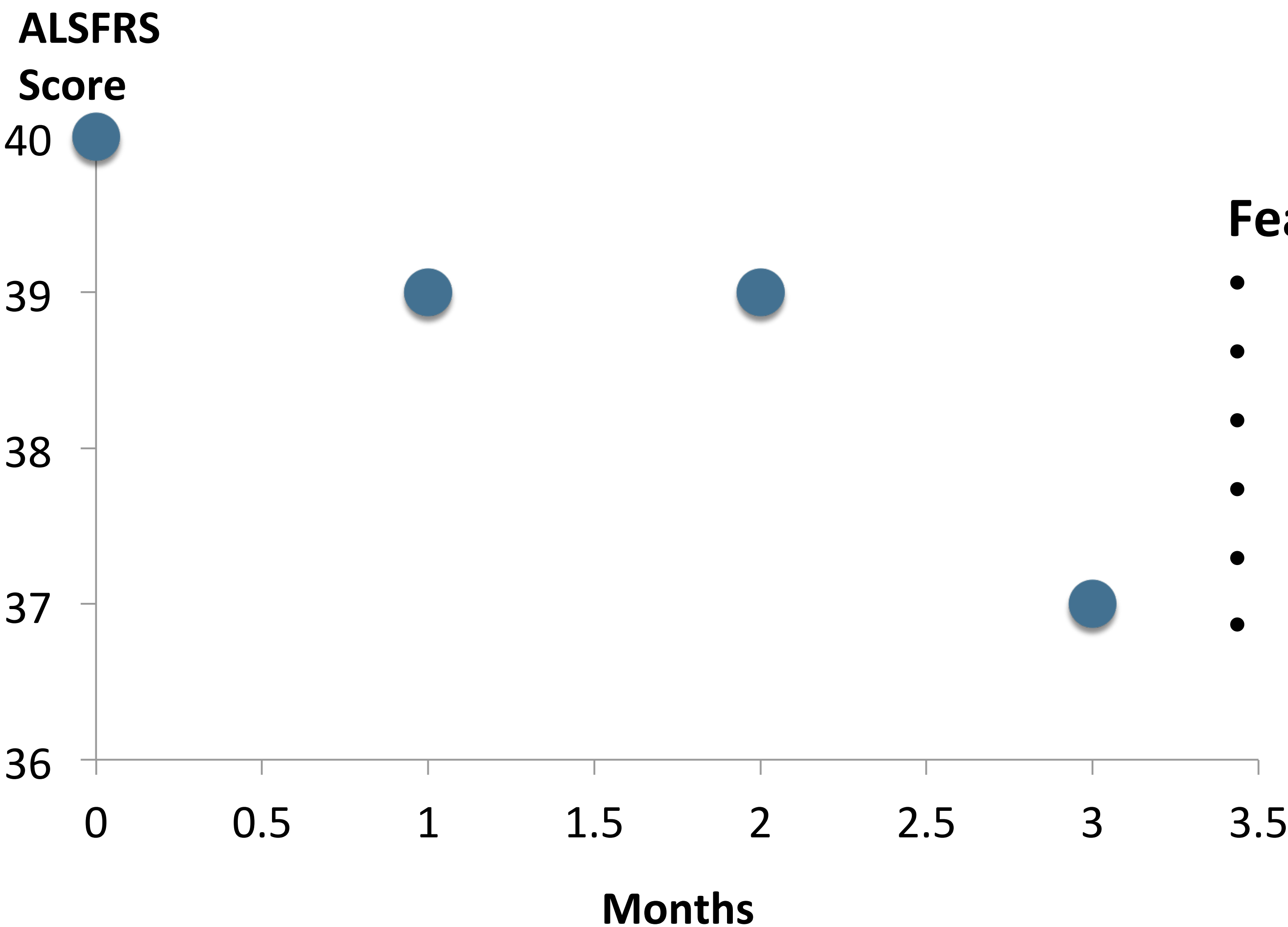
- **Goal:** Compact numeric representation of each patient
  - Features will serve as covariates in a regression model
  - Most extracted features will be irrelevant
  - Rely on model selection / methods robust to irrelevant features

- **Time Series Data**
  - Compute summary statistics from each time series
    - Mean value, standard deviation, slope, last recorded value, maximum value…
  - Compute pairwise slopes (difference quotients between adjacent measurements)
    - Induces a derivative time series
    - Extract same summary statistics

Featurizing Time Series Data



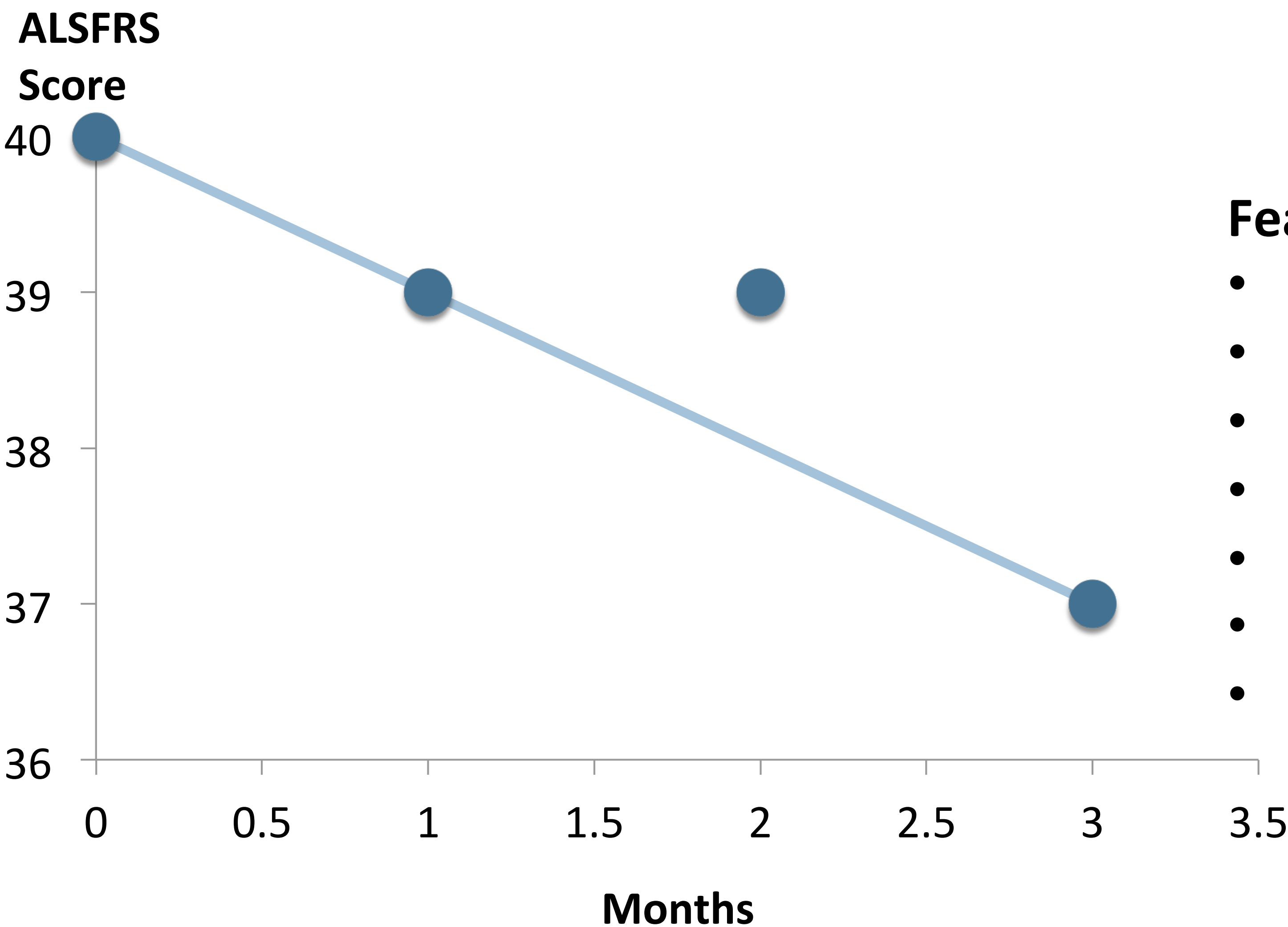
Featurizing Time Series Data



Features extracted

- Mean = 38.75
- SD = 0.816
- Max = 40
- Min = 37
- Last = 37
- etc.

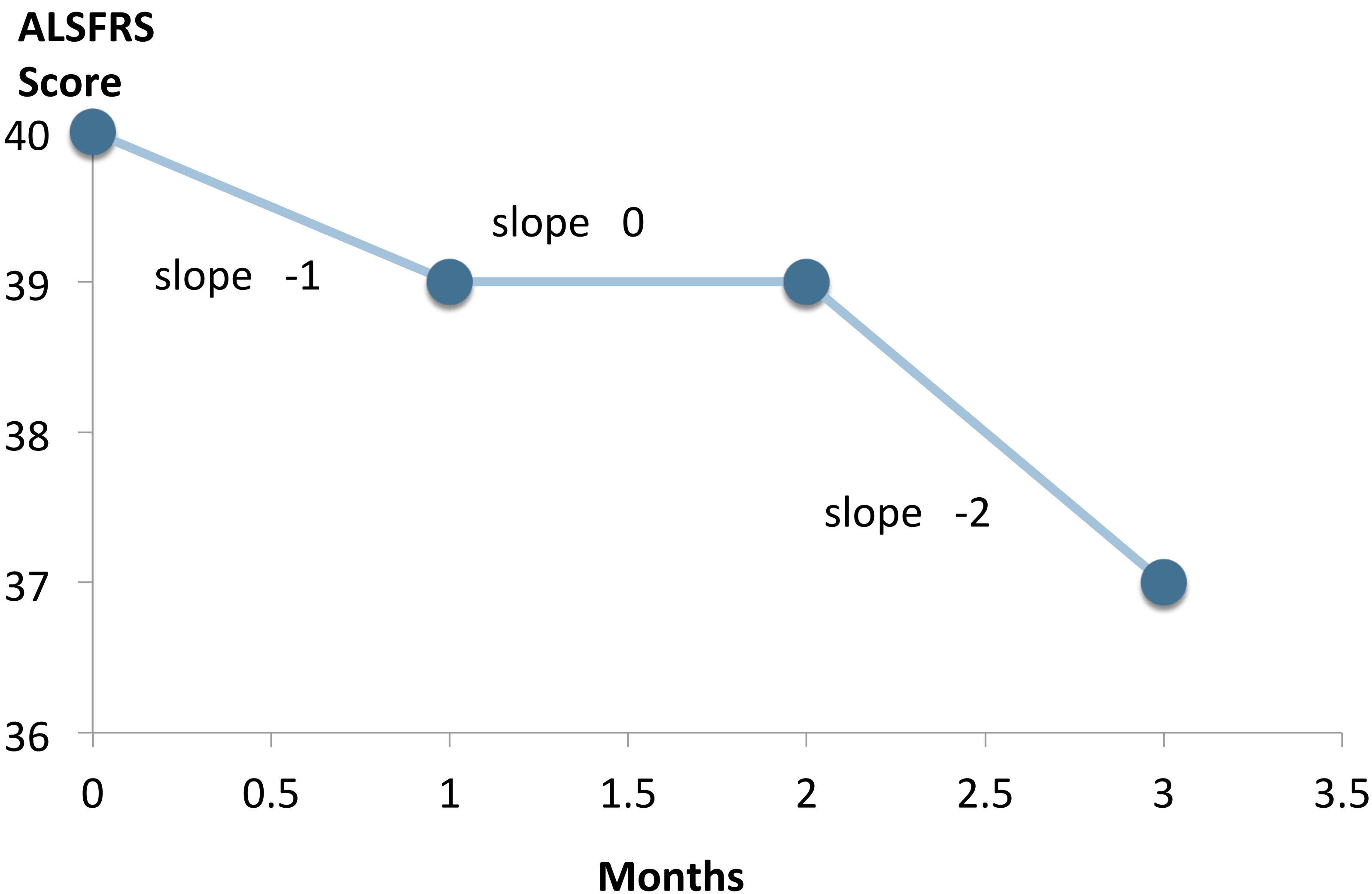
Featurizing Time Series Data



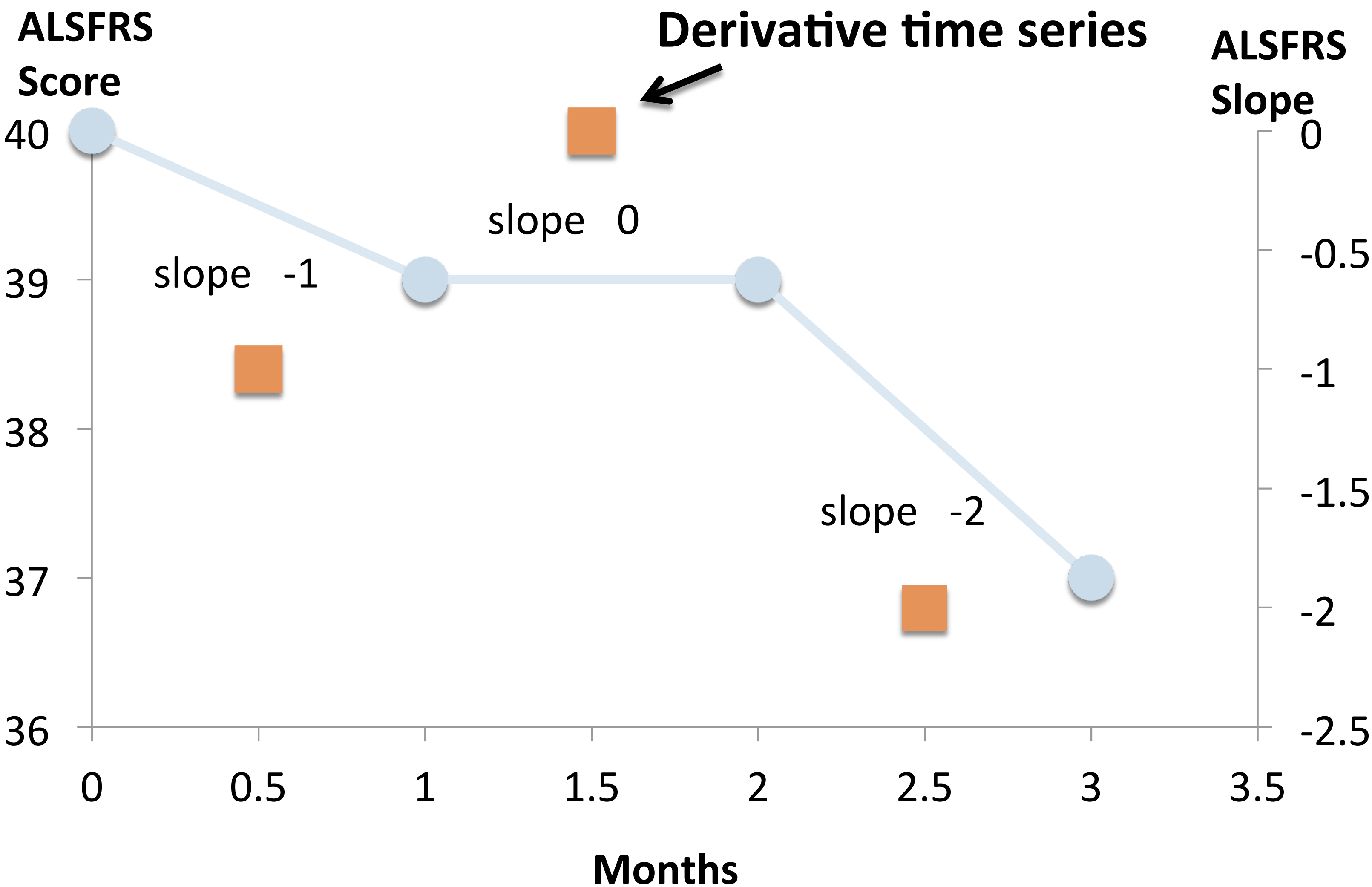
Features extracted

- Mean = 38.75
- SD = 0.816
- Max = 40
- Min = 37
- Last = 37
- Slope = -1
- etc.

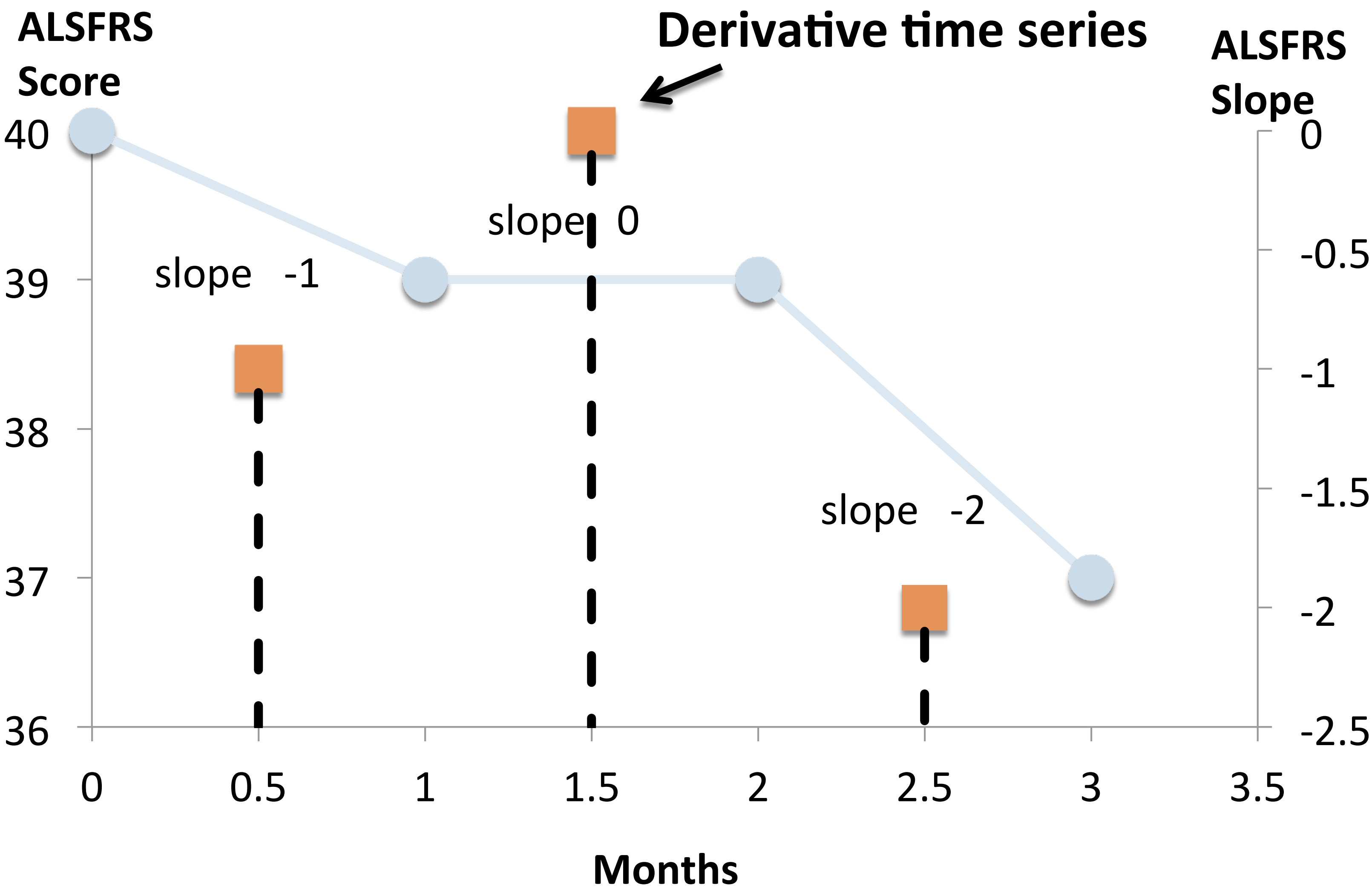
Featurizing Time Series Data



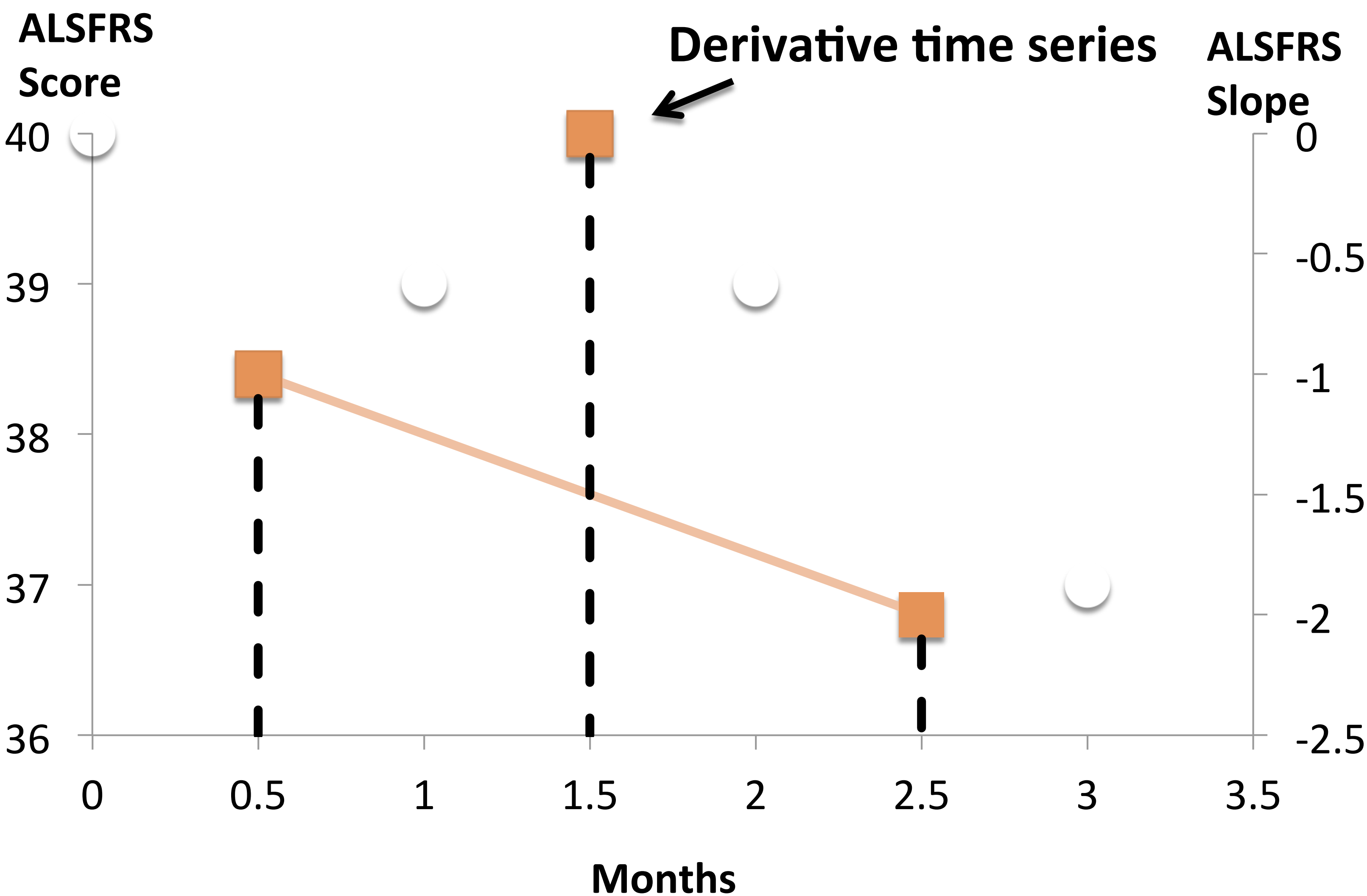
Featurizing Time Series Data



Featurizing Time Series Data



Featurizing Time Series Data



Features extracted

Mean = -1

SD = 1

Max = 0

Min = -2

Last = -2

Slope = -0.5

etc.

Featurizing Time Series Data

- **435** temporal features extracted
- **Problem:** Missing data
  - Average patient missing 10% of features
  - One patient missing 55% of features!
  - Missing values imputed using median heuristic
- **Problem:** Outliers
  - Nonsense values: Number of liters recorded as `MDMD`
  - Units incorrectly recorded ⇒ Wrong conversions
  - Extreme values
    - Treated as missing if > 4 standard deviations from mean

Modeling and Inference

- **Regression model**

  Future ALSFRS Slope = f(features) + noise

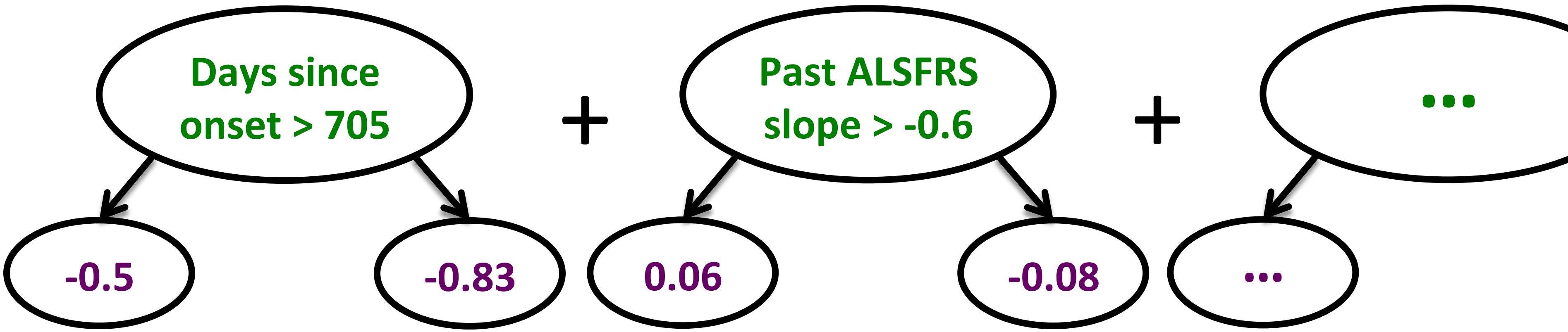
  ↖ Unknown regression function

- **Goal:** infer **f** from data
  - Bayesian: Place a prior on **f**, infer its posterior
  - Bonus: Uncertainty estimates for each prediction

- **What prior?**
  - Flexible and nonparametric
    - Avoid restrictive assumptions about functional form
  - Favor simple, sparse models
    - Avoid overfitting to irrelevant features

Bayesian Additive Regression Trees*

- f(**features**) = sum of **"simple" decision trees**



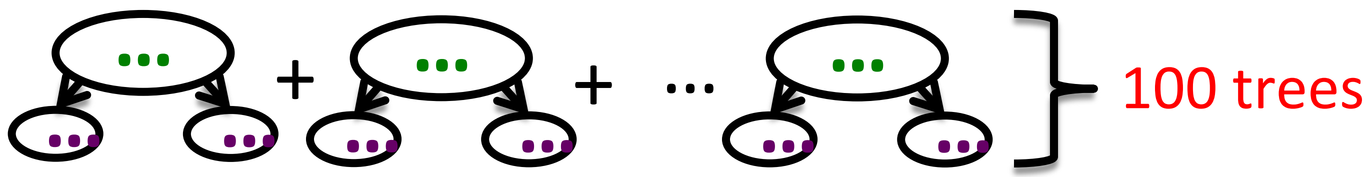
- **Simplicity** = tree depends on few features
  - Irrelevant features seldom selected
- Similar to frequentist ensemble methods
  - Boosted decision trees, random forests

*Chipman, George, and McCulloch (2010)

BART Inference

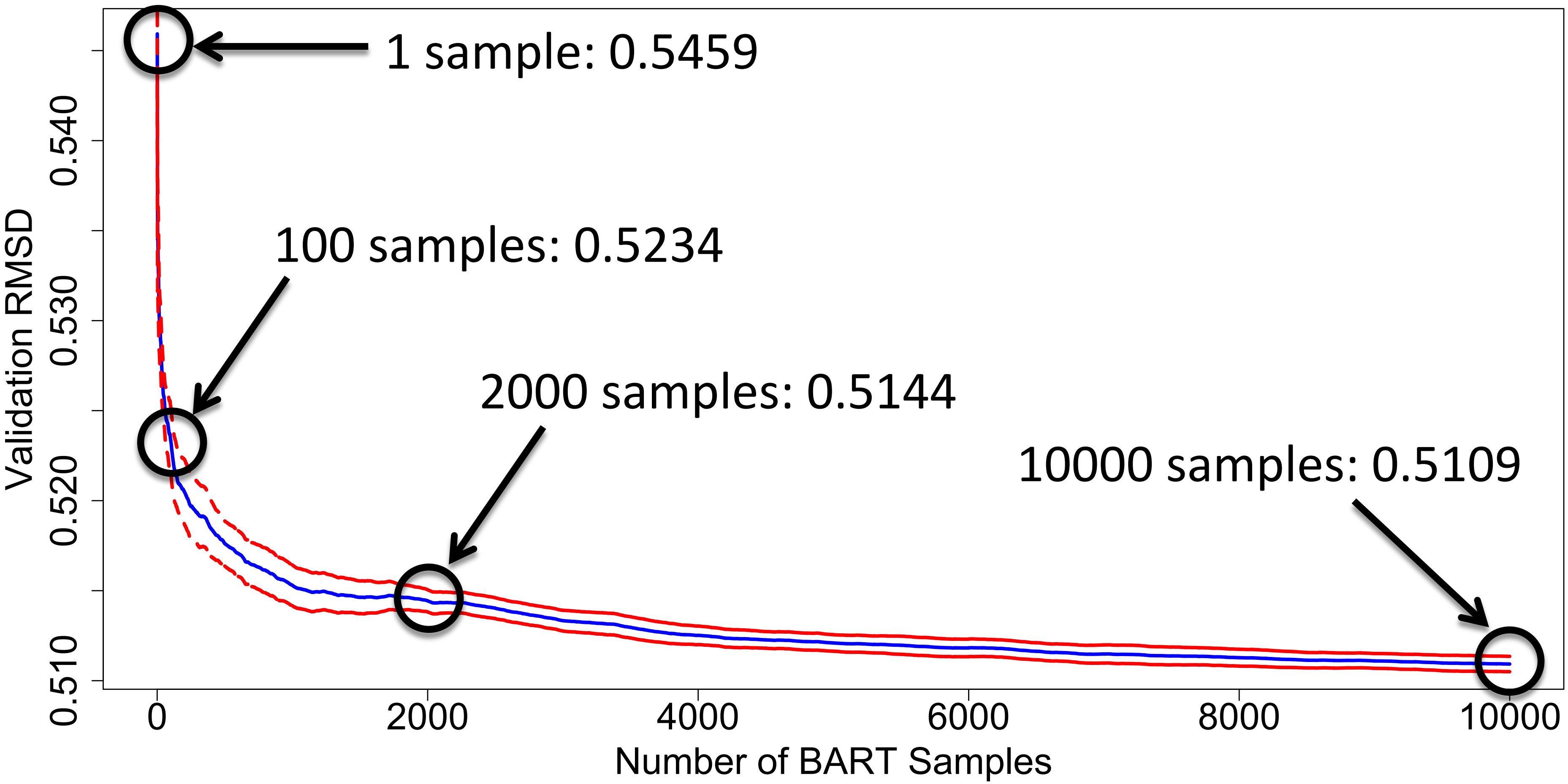
- **Estimating f:** Markov Chain Monte Carlo
  - R package 'bart' available on CRAN
  - 10,000 posterior samples: $\hat{f}_1, \hat{f}_2, \hat{f}_3, \hat{f}_4, \ldots$

$\hat{f}_i =$

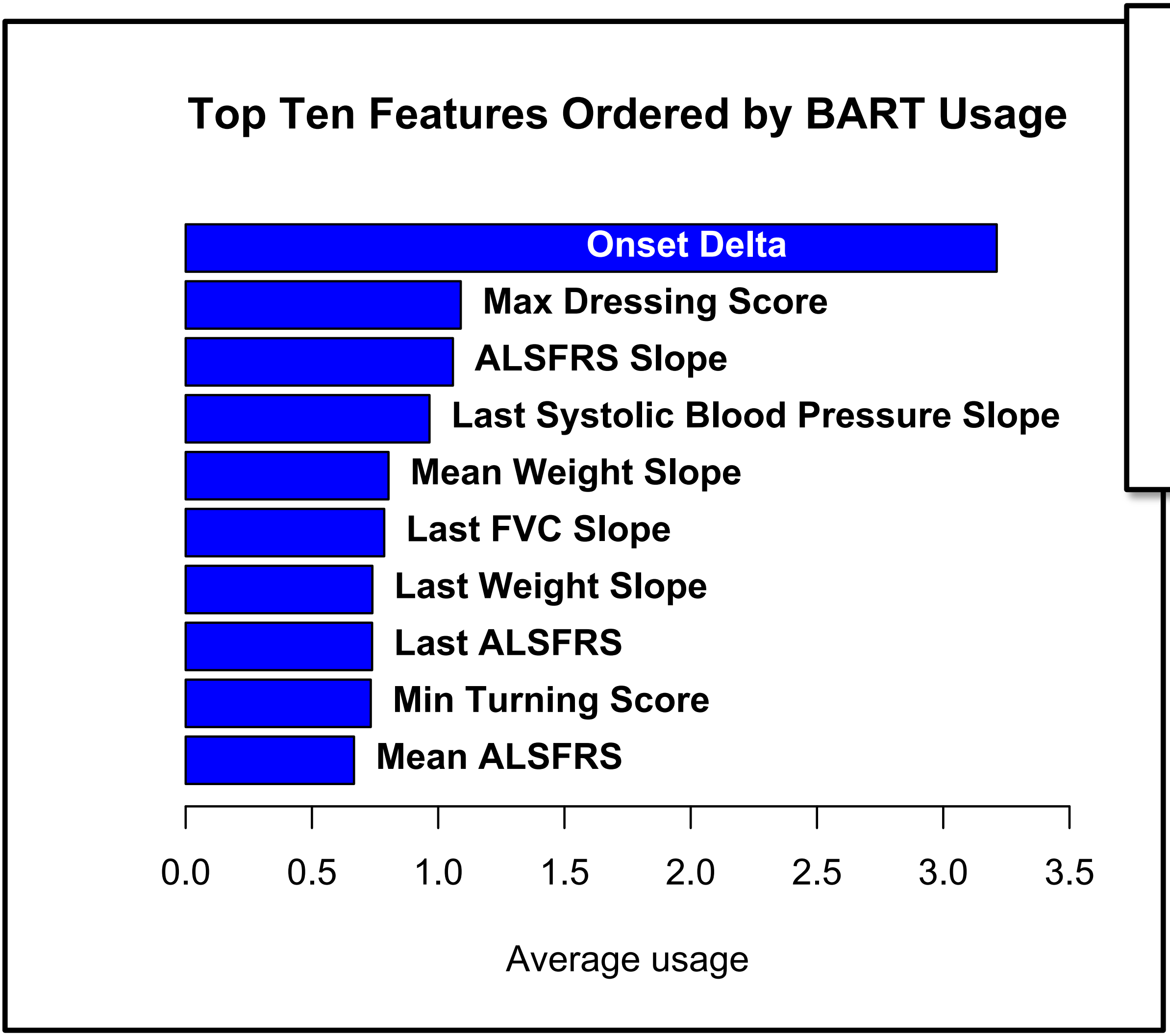


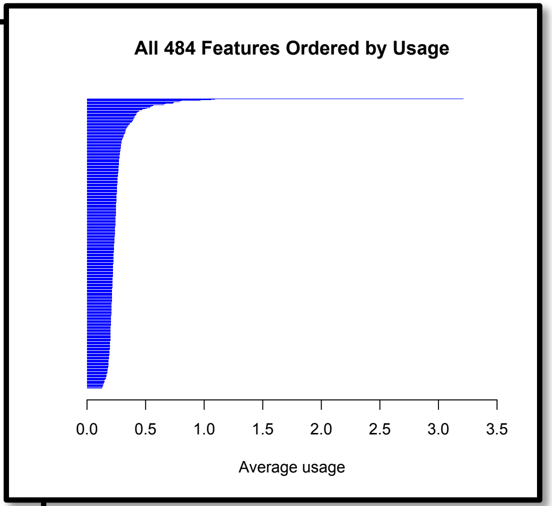
  - 10 minutes on MacBook Pro (2.5 GHz CPU, 4GB RAM)
- **Prediction:** Posterior mean
  - Average of $\hat{f}_1$(**features**), $\hat{f}_2$(**features**), $\hat{f}_3$(**features**), …
- **Variance reduction**
  - Average predictions of 10 BART models

Accuracy of BART Inference



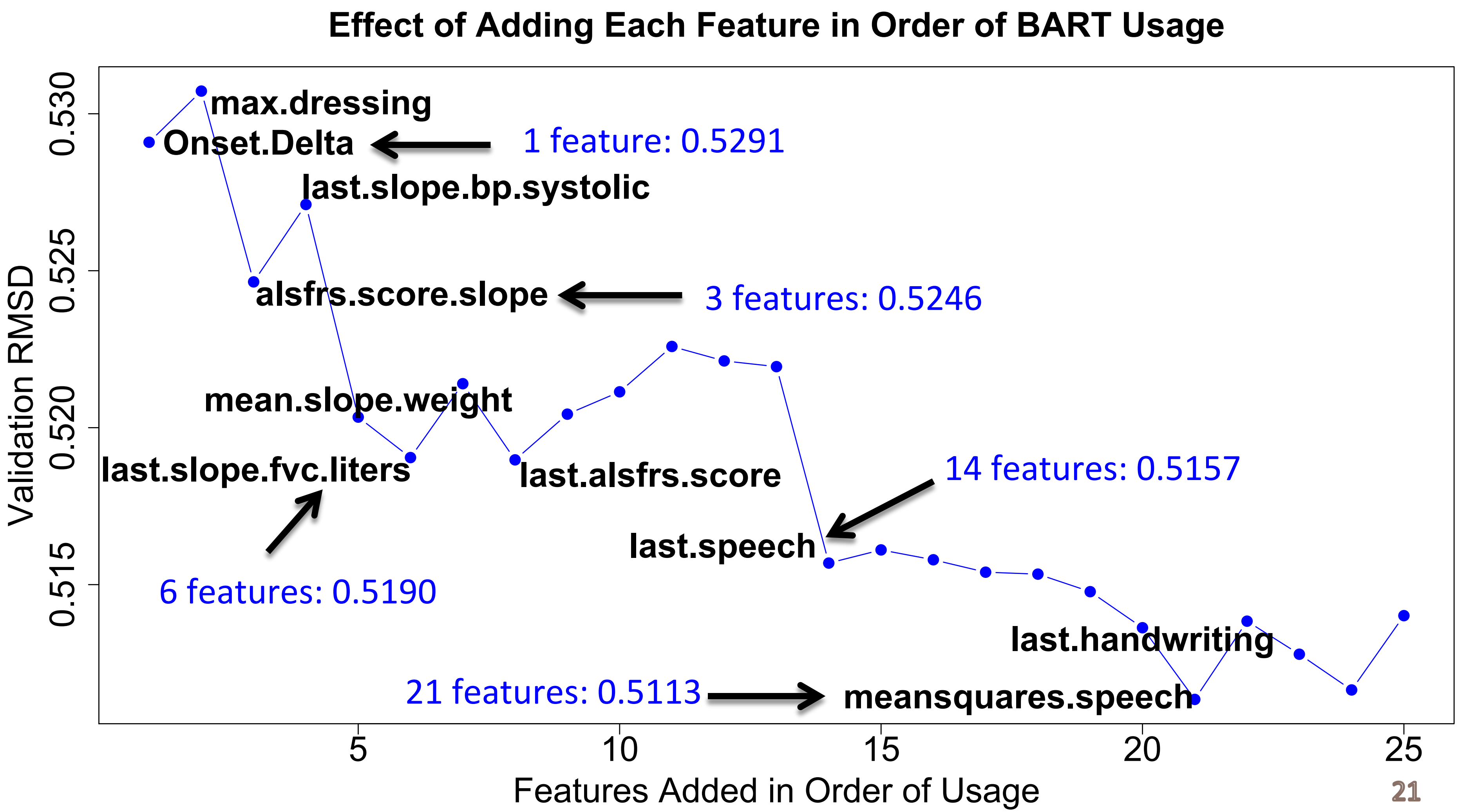
BART Feature Selection





- Many pairwise slope features
- Lab data excluded

BART on Feature Subsets



Model Comparison

How do other models perform using our feature set?

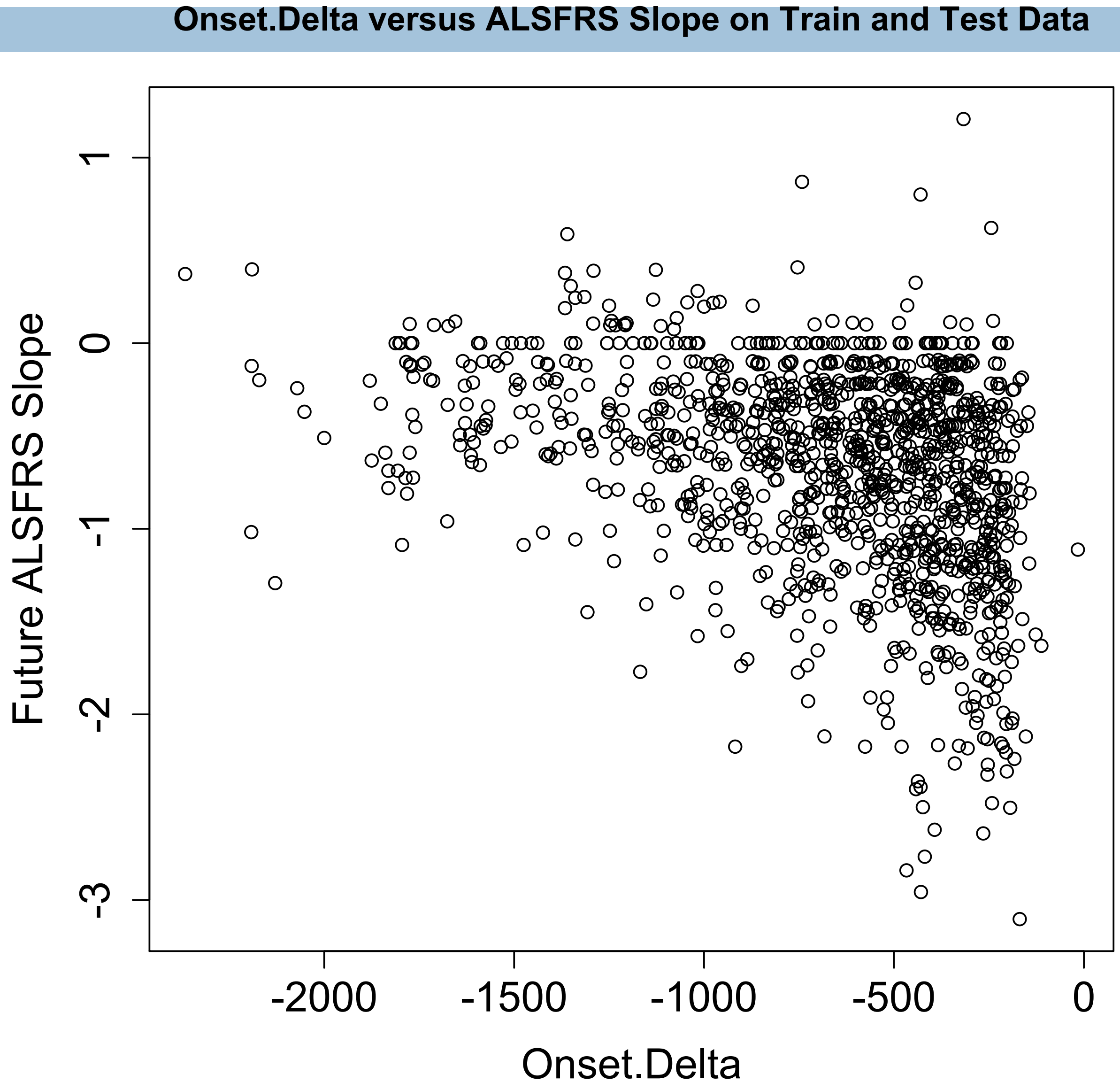
Model	**Our RMSD (Test)**	**Our RMSD (Validation)**	**Competitor RMSD**
Lasso Regression	0.5006	0.5287	-
Random Forests	0.5052	0.5120	0.52-0.53
Boosted Trees	0.4940	0.5118	-
BART	0.4860	0.5109	-

- **Additive decision tree** models especially effective
- **Featurization** is a main differentiator of competitors

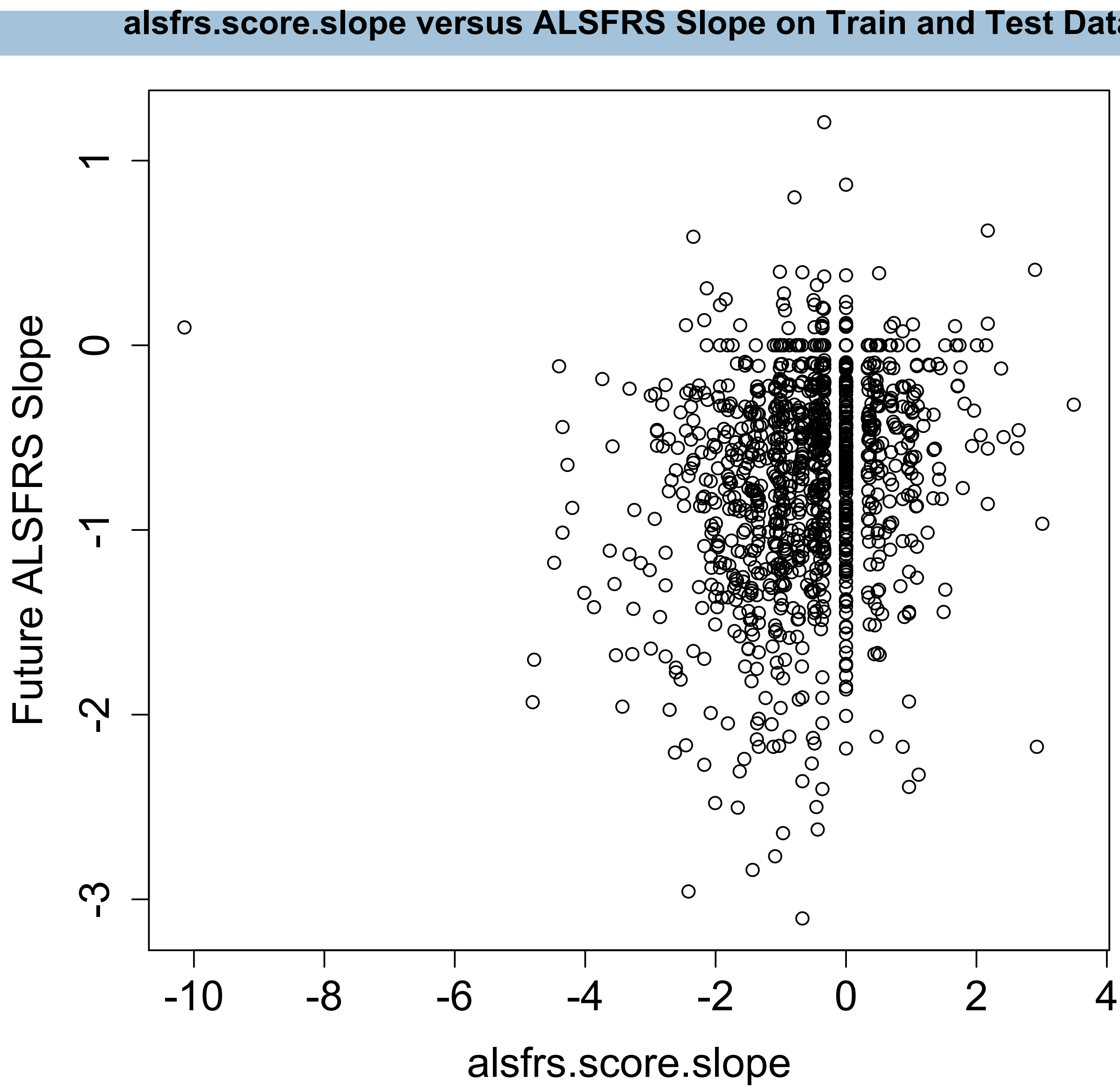
The End

Questions?

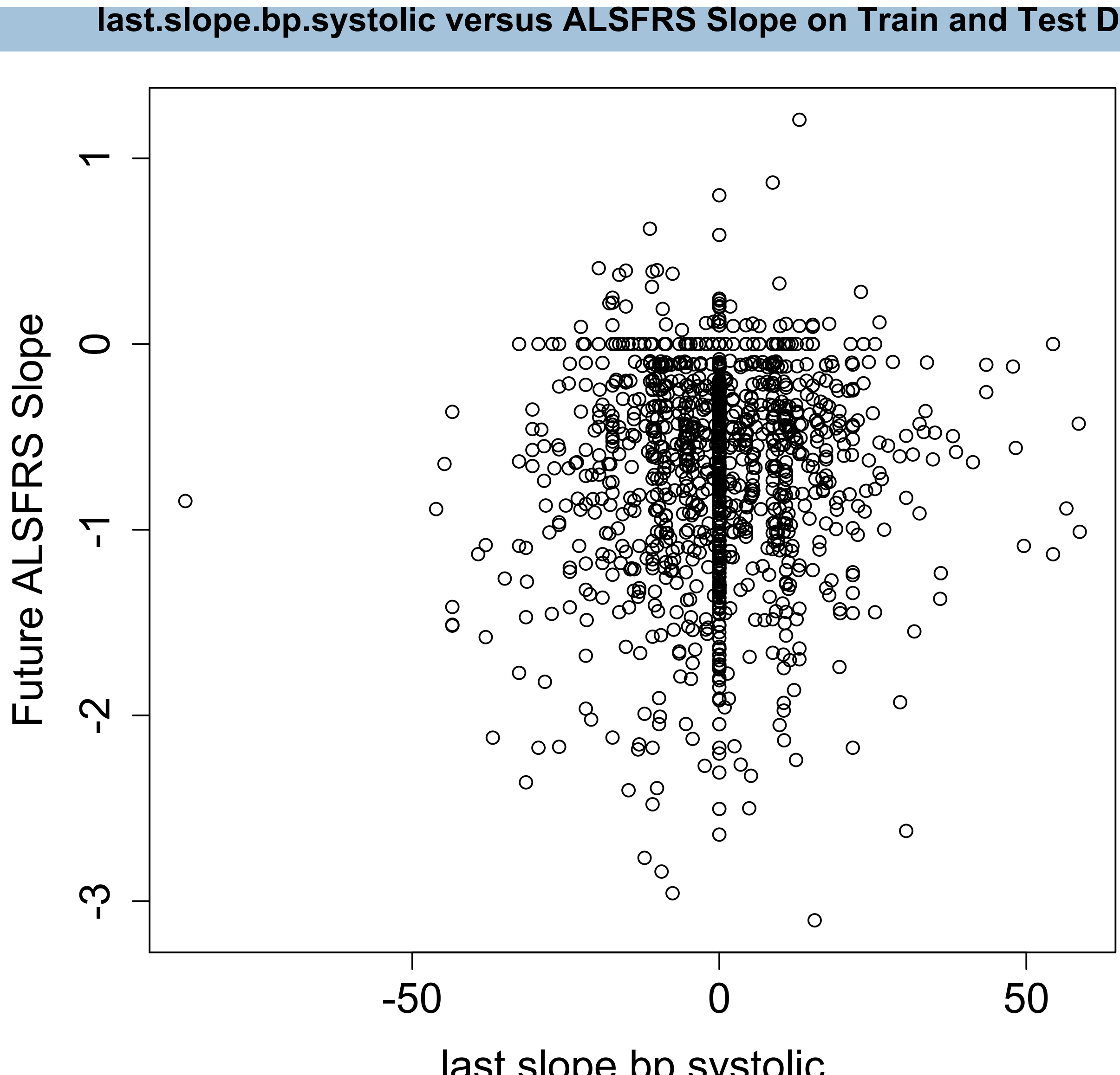
Onset Delta vs. Target



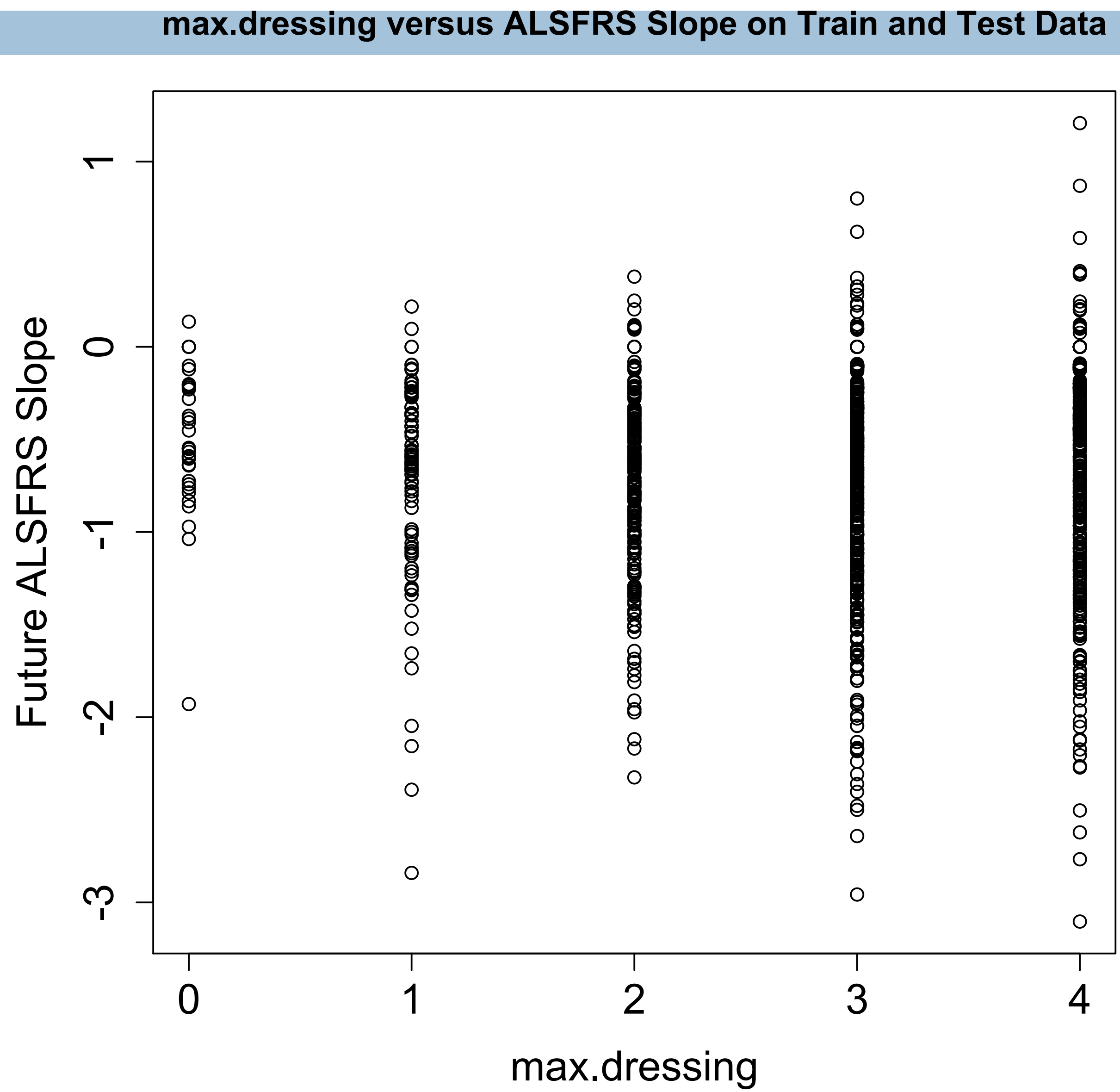
Past ALSFRS Slope vs. Target



Last Systolic BP Slope vs. Target



Max Dressing Score vs. Target



Mean Weight Slope vs. Target