DSC5103 Statistics

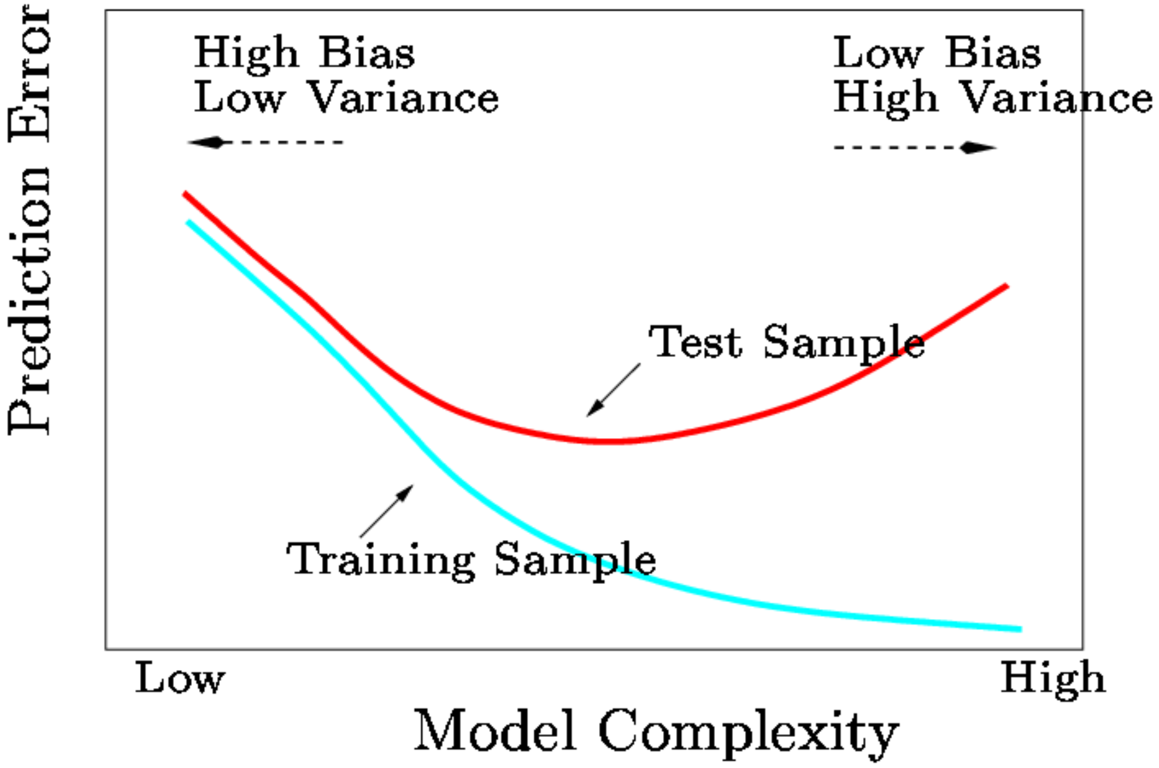
Session 11. Review

Last time

- Unsupervised Learning
    - Clustering Methods
        - K-mean clustering
        - Gaussian mixture model
        - Hierarchical clustering
    - Principle Components Analysis

The Fundamental

- **Out-of-sample** performance is the key in predictive analytics
- The Bias-Variance decomposition and trade-off



The Process

- Data partition
  - Training: fit/train a particular model
  - Validation / Cross-Validation: model selection, model comparison
  - Test: final evaluation of out-of-sample performance
- Monte-Carlo Simulation: a tool to test the tools from the God's view
- Bootstrap: a tool to mimic the process of generating data samples from the population

The Tools

- Supervised learning toolbox
  - K Nearest Neighbors
  - Regression and generalizations
    - Linear regression
    - Generalized linear models: logistic regression, Poisson regression
    - Regularized regression: Ridge regression, LASSO, Elastic Net
  - Tree-based methods
    - Tree
    - Bagging & Random Forest
    - Boosting
  - ~~Support Vector Machine~~
  - ~~Neural Network~~
- Unsupervised learning toolbox
  - K-means clustering, Gaussian mixture models, hierarchical clustering
  - Principal Components Analysis

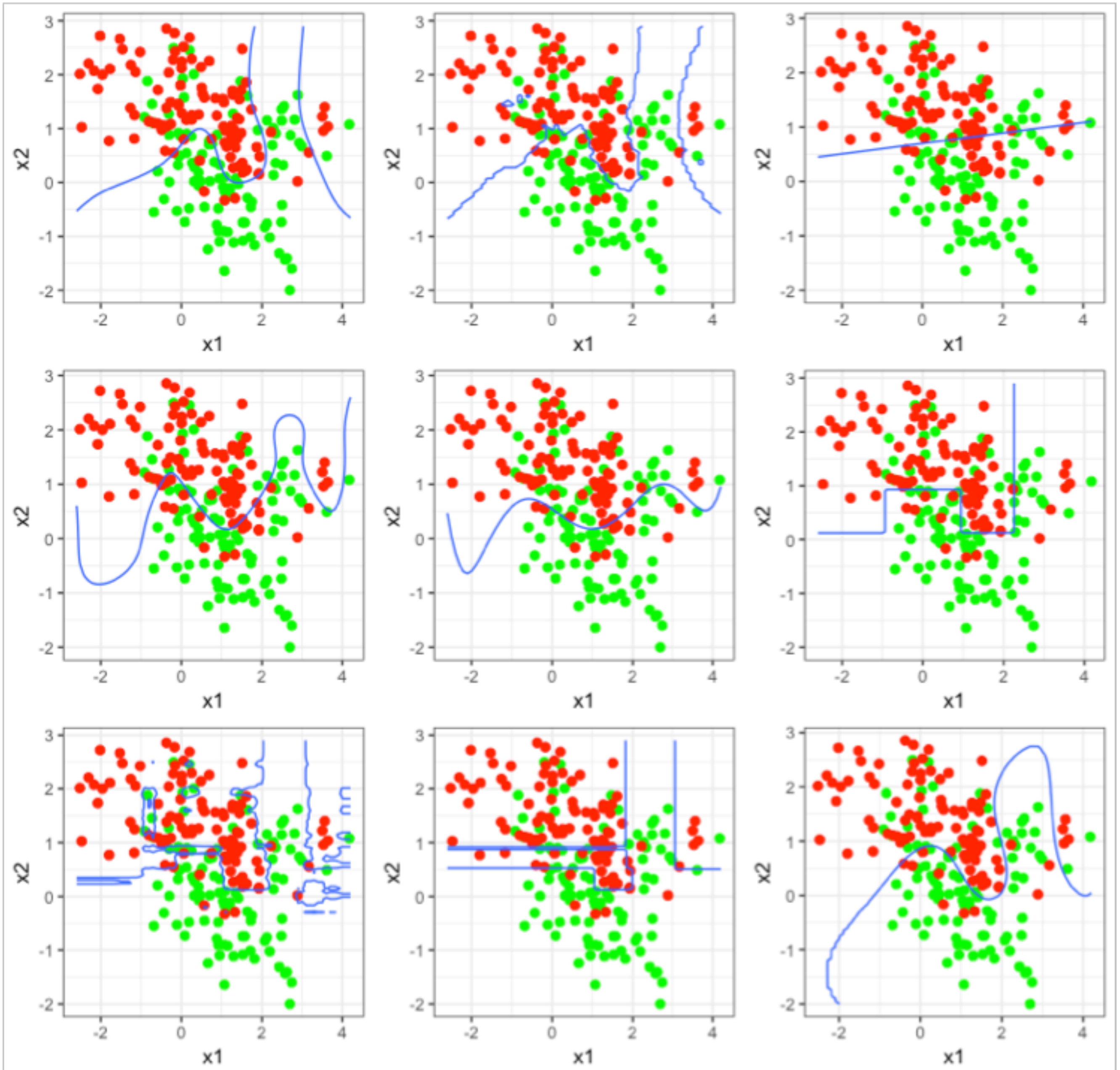
The Big Ideas

- Regularization: shrink to control variance

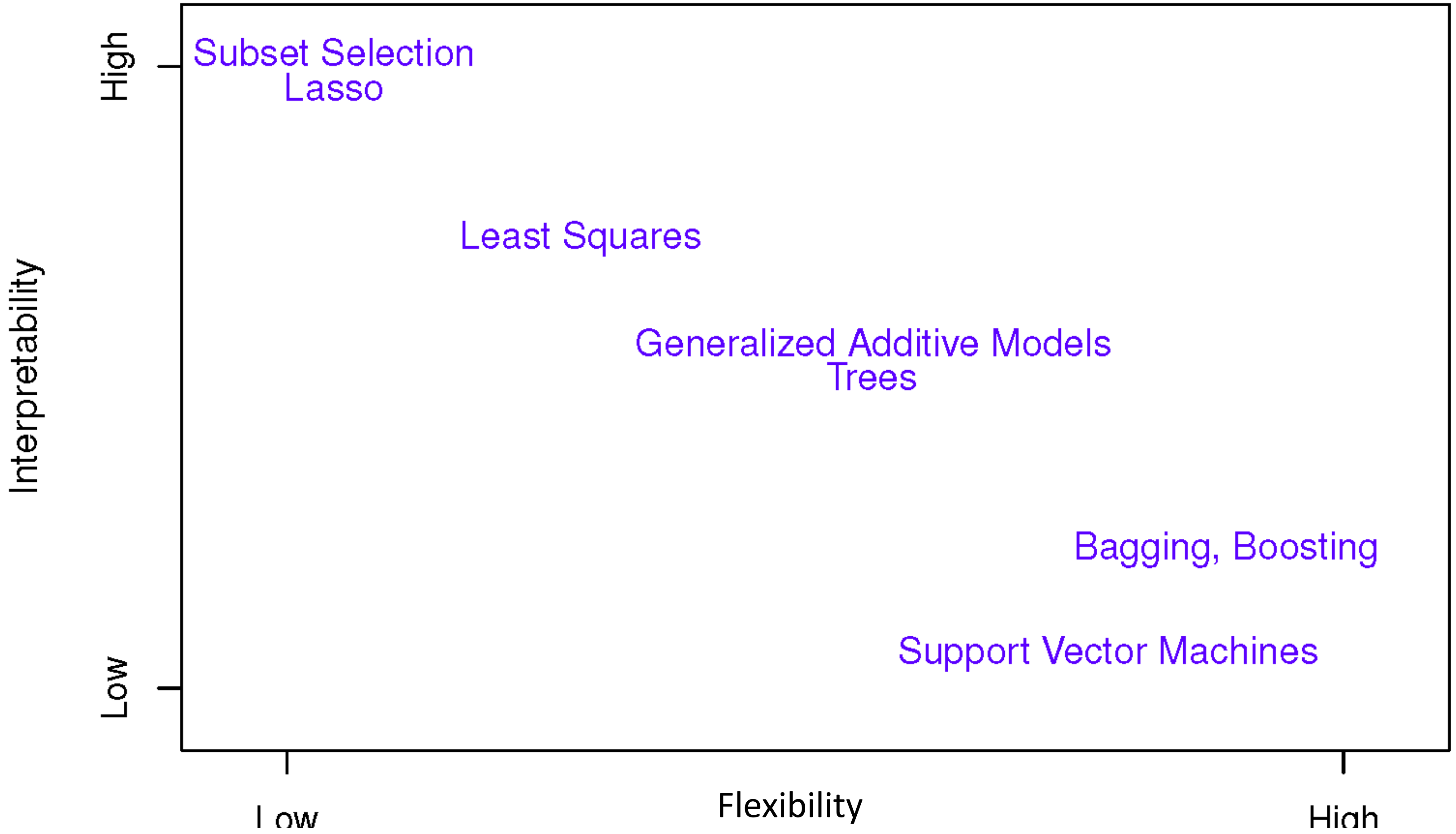
- Bagging: average to control variance

- Boosting: sequentially train weak learners

The Mixture Example



Flexibility vs. Interpretability



Moving Forward

- Feature engineering
- Unstructured data: text, image, voice, ...
- Deep Learning
- Bigger data
- Bayesian models