Data Science

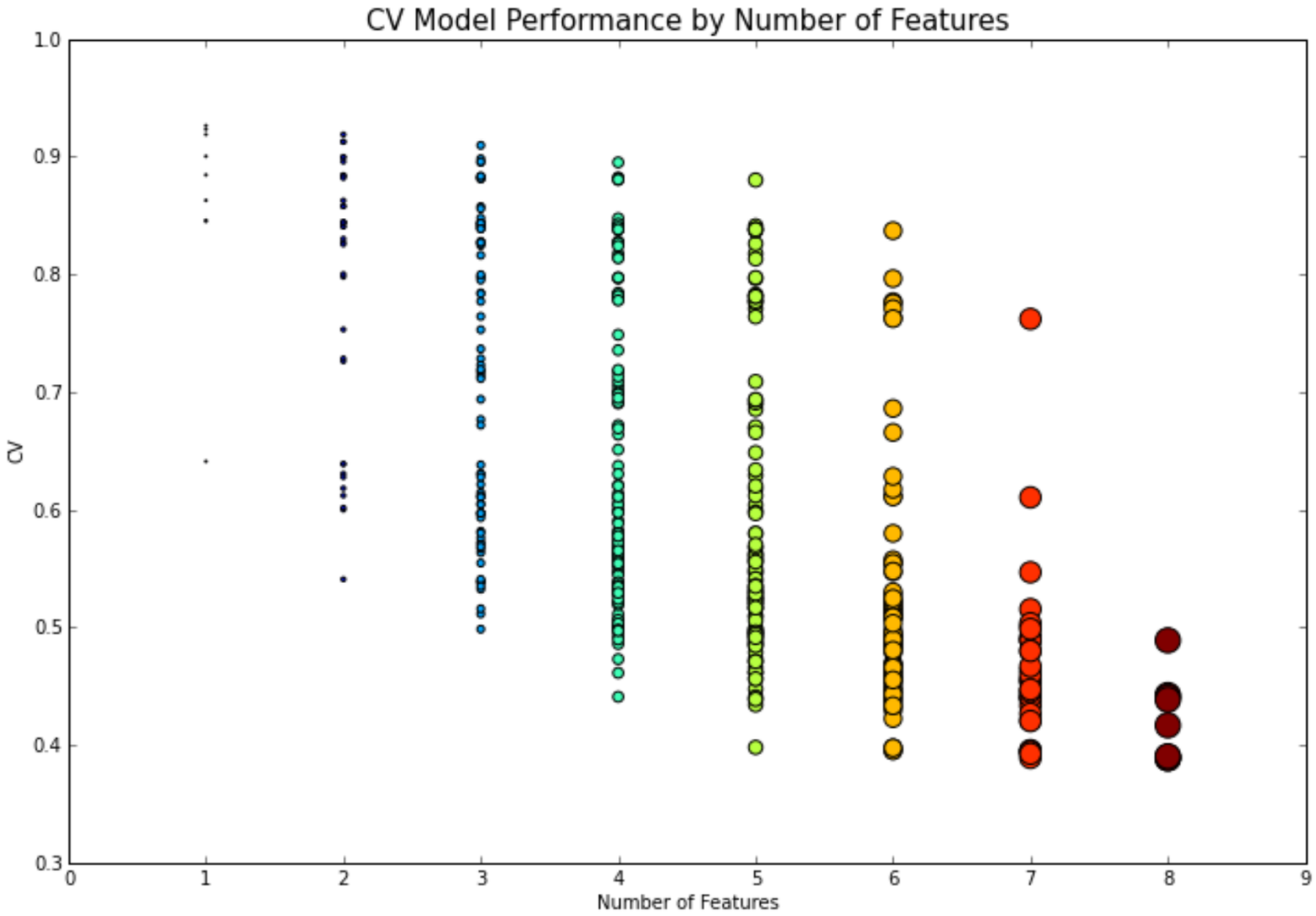
Naive Search Methods and Less Naive Methods

October 21st, 2014

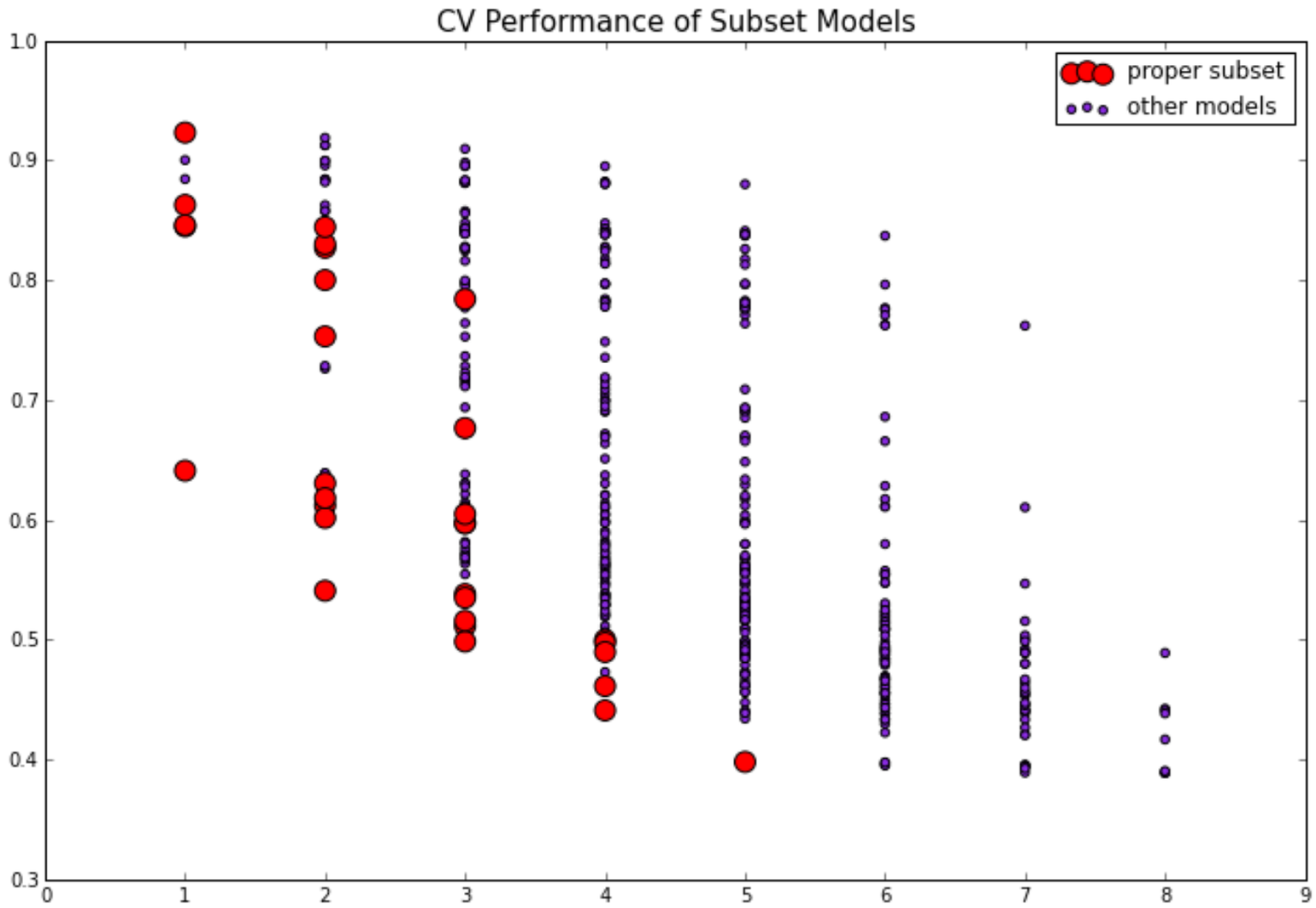
Jarret Petrillo, jarret.petrillo@gmail.com

1,125,899,907,000,000

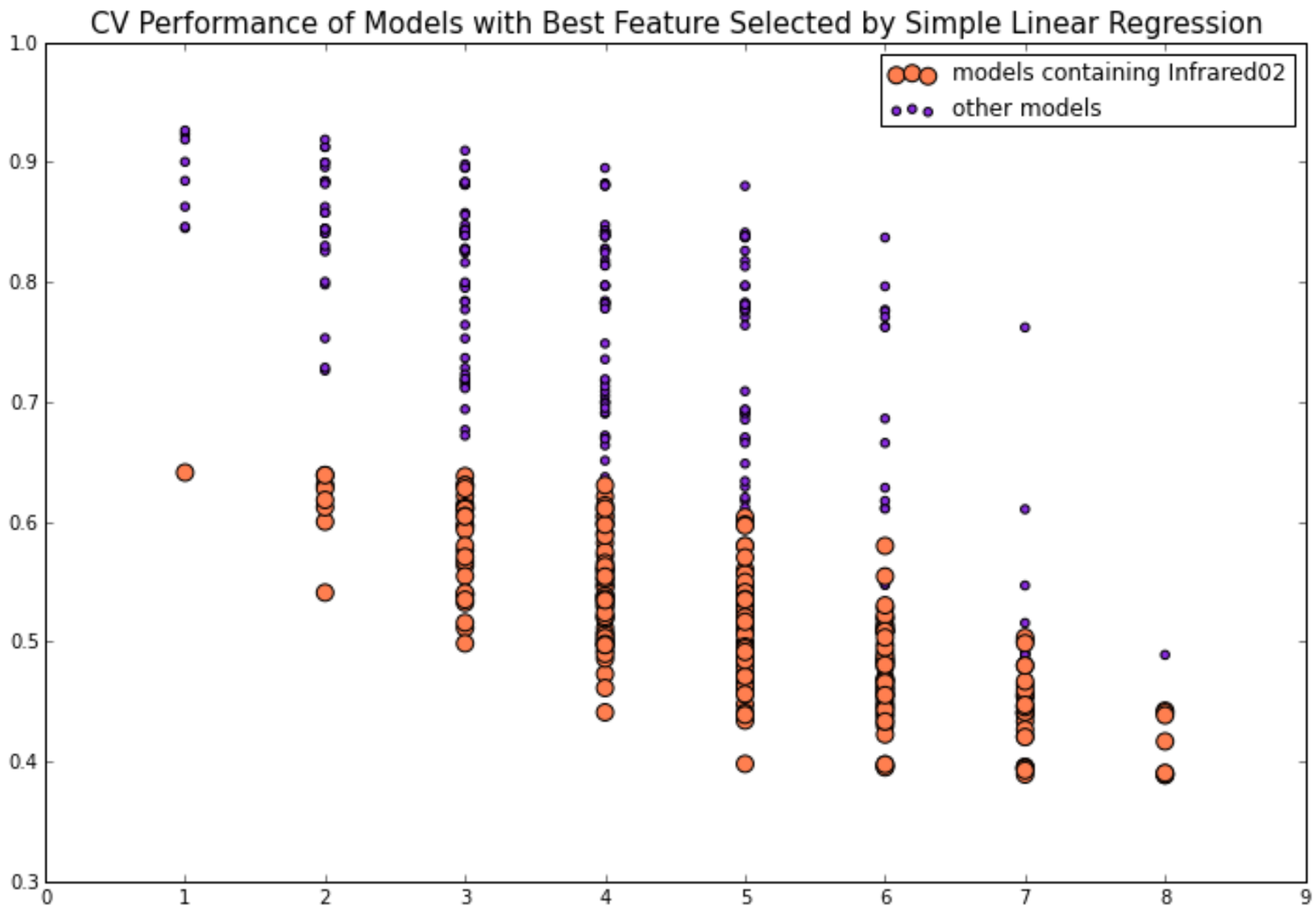
Compare CV across all subsets



Could we have found the best model quicker?



Derivative models that start with good features do well



Pseudo Code: Naive Model Search Algorithm

1. **Start with a list of features**
2. **Use itertools to find all combinations (2^n!)**
3. **For each subset fit a linear regression model**
4. **Calculate cross-validated MSE with a test set**
5. **Choose the model with the lowest mean squared error**

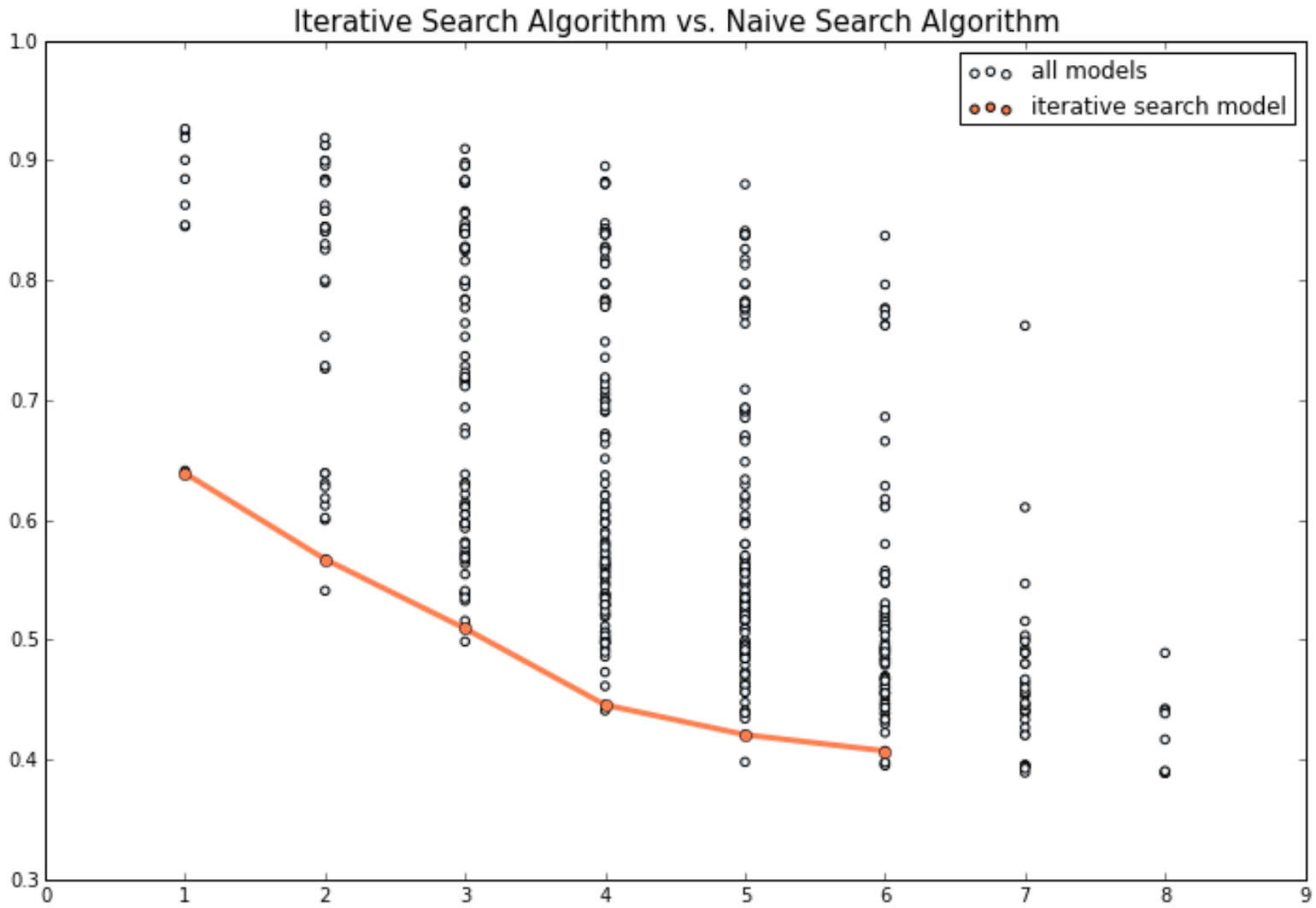
Pseudo Code: Iterative Search Algorithm

1. **Start with a list of features (n)**
2. **Run n simple linear regression models**
3. **Calculate cross-validated MSE for each model**
4. **Save the best feature**
    It will be in every subsequent model!
5. **Consider only two feature models that contain the first (n-1)**
6. **For each new model fit a linear regression model**
7. **Calculate cross-validated MSE**
8. **Save the best features**
9. **Consider only three feature models that contain the best two!**

Repeat!

Stop when the MSE gets worse with any added feature

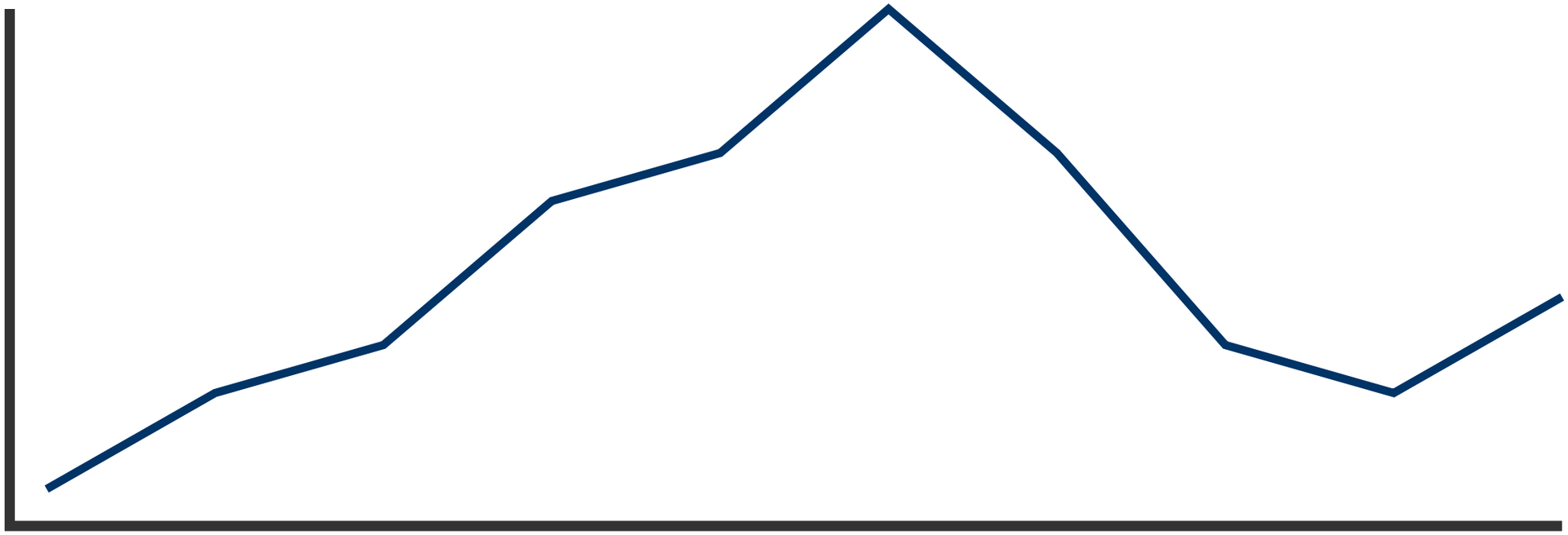
Performs almost as well as naive method!



Appendix

Example Data Slide

Revenue



- *Description of data description of data description of data description of data description of data description of data description of data description of data*

Example Standard Slide

- *“Abstraction: the process of determining the important characteristics and ignoring other details”*
- *“Plan to throw away the first version”*
- *“There is surely nothing quite so useless as doing with great efficiency what should not be done at all”*