Common Mistakes in Machine Learning

(Not exhaustive)

Bad Annotation of Training/ Testing Data Set

Poor Understanding Algorithm Assumptions

Poor Understanding of algorithm parameters

(use of defaults)

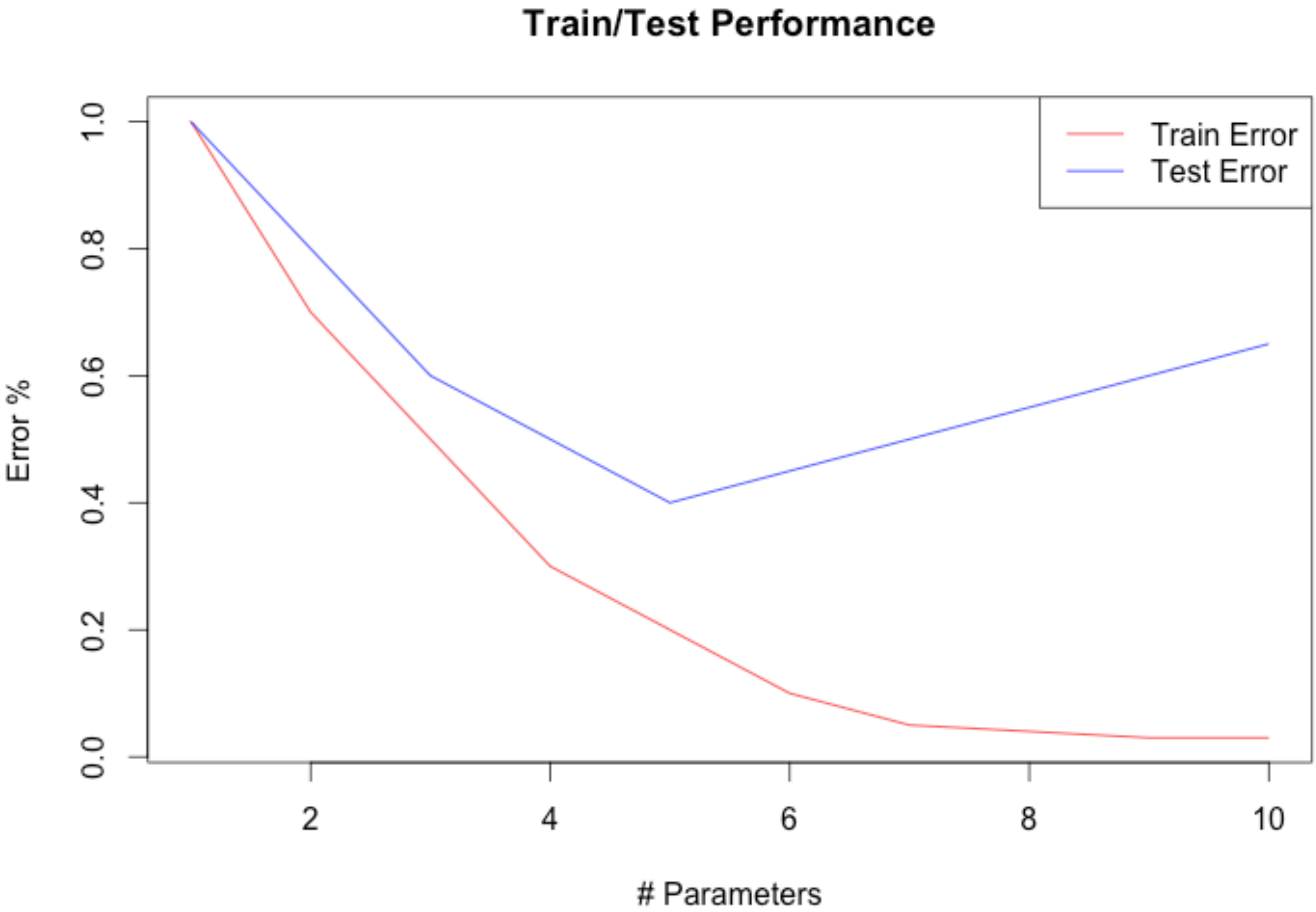
What is our objective?

Not understanding the data

No EDA - Going into the data blind
(outliers, multicollinearity)
Not understanding population this represents
Insufficient data

Ignoring n<<P

Potential to overfit


Train/Test Performance
Train Error
Test Error
Error %
1.0
0.8
0.6
0.4
0.2
0.0
2
4
6
8
10
# Parameters

Allowing Leakage

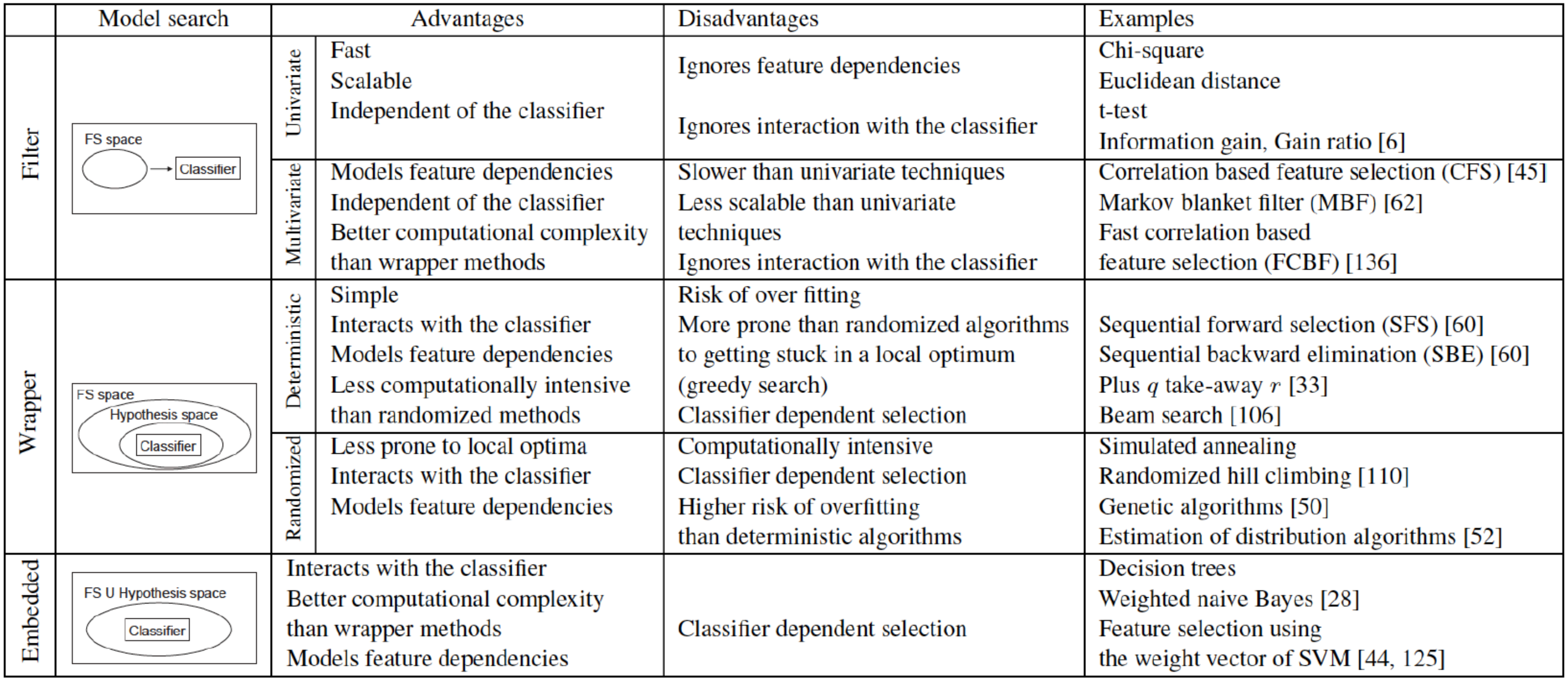
Features, Information (Classes)

Motivation for Feature Selection

- Reducing dimensionality
- Improving learning efficiency
- Increasing predicative accuracy
- Reducing complexity of learned results

Feature Selection Approaches

- **Filter Methods**
- **Wrapper Methods**
- **Embedded Methods**



	Model search		Advantages	Disadvantages	Examples
Filter	FS space Classifier	Univariate	Fast Scalable Independent of the classifier	Ignores feature dependencies Ignores interaction with the classifier	Chi-square Euclidean distance t-test Information gain, Gain ratio [6]
		Multivariate	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation based feature selection (CFS) [45] Markov blanket filter (MBF) [62] Fast correlation based feature selection (FCBF) [136]
Wrapper	FS space Hypothesis space Classifier	Deterministic	Simple Interacts with the classifier Models feature dependencies Less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS) [60] Sequential backward elimination (SBE) [60] Plus q take-away r [33] Beam search [106]
		Randomized	Less prone to local optima Interacts with the classifier Models feature dependencies	Computationally intensive Classifier dependent selection Higher risk of overfitting than deterministic algorithms	Simulated annealing Randomized hill climbing [110] Genetic algorithms [50] Estimation of distribution algorithms [52]
Embedded	FS U Hypothesis space Classifier		Interacts with the classifier Better computational complexity than wrapper methods Models feature dependencies	Classifier dependent selection	Decision trees Weighted naive Bayes [28] Feature selection using the weight vector of SVM [44, 125]

Saeys et al 2005

Not matching approach to question (or not having a question)

If you try hard enough, you can beat the data into submission to say anything – avoid this **please**.

Poor Understanding of algorithm

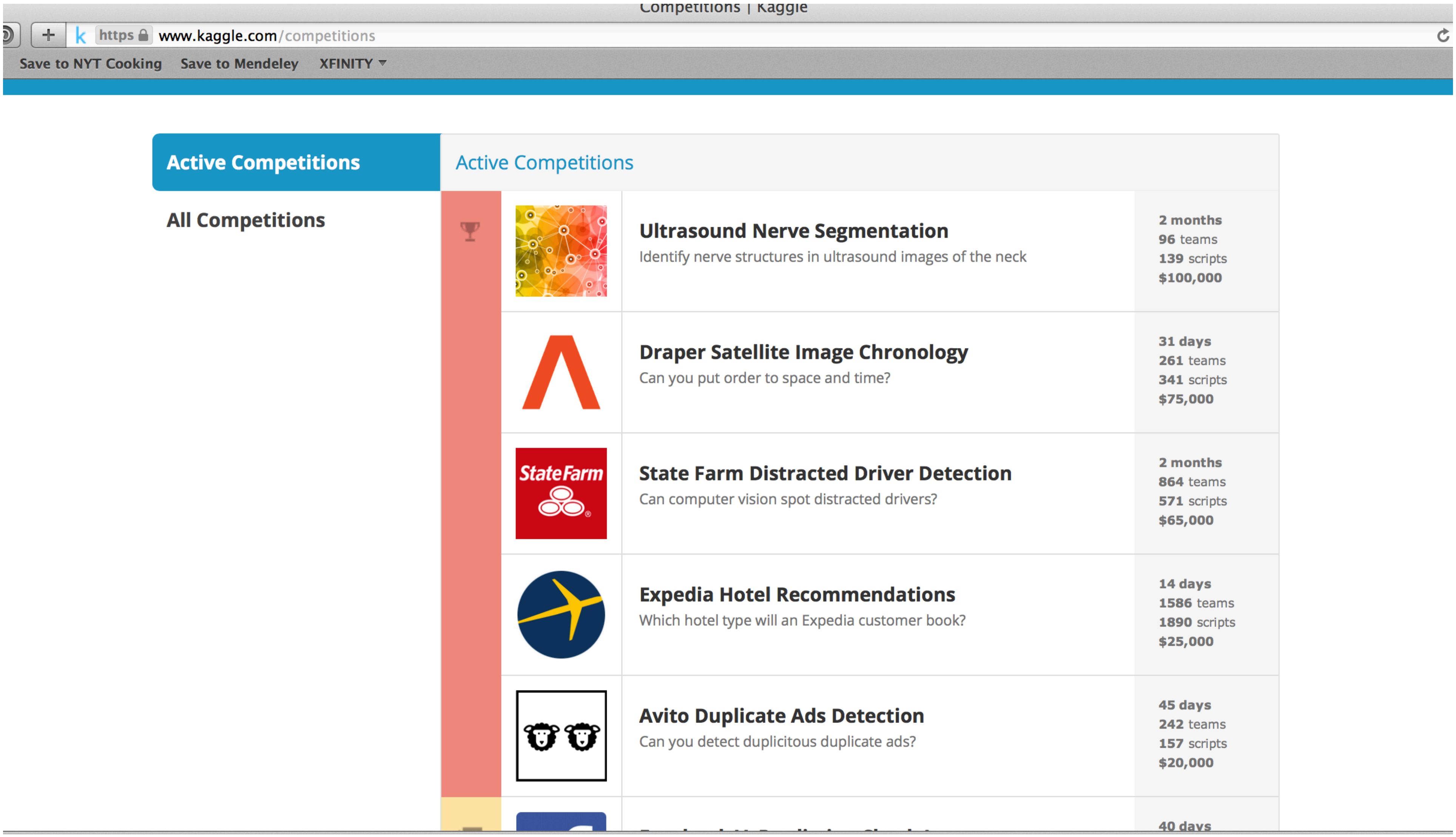
Methods Overview + In-class Discussion

LINEAR DISCRIMINANT ANALYSIS

Methods Overview + In-class Discussion

CLASSIFICATION + REGRESSION TREES

https://www.kaggle.com/



Competitions | Kaggle

https www.kaggle.com/competitions

Active Competitions

All Competitions

Active Competitions

Ultrasound Nerve Segmentation
Identify nerve structures in ultrasound images of the neck

2 months
96 teams
139 scripts
$100,000

Draper Satellite Image Chronology
Can you put order to space and time?

31 days
261 teams
341 scripts
$75,000

State Farm Distracted Driver Detection
Can computer vision spot distracted drivers?

2 months
864 teams
571 scripts
$65,000

Expedia Hotel Recommendations
Which hotel type will an Expedia customer book?

14 days
1586 teams
1890 scripts
$25,000

Avito Duplicate Ads Detection
Can you detect duplicitous duplicate ads?

45 days
242 teams
157 scripts
$20,000

40 days