Last session (2015-10-15)

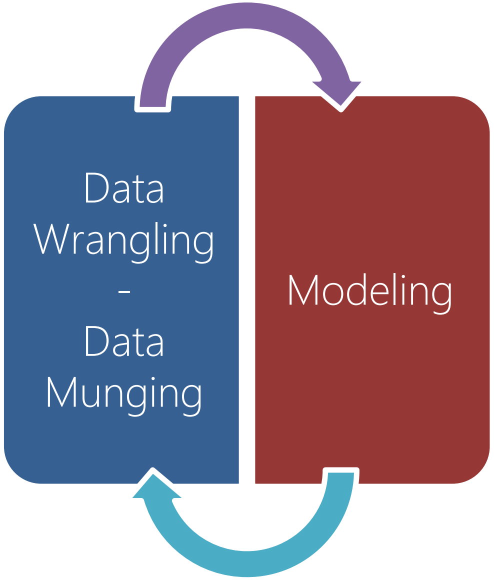
- Big Data Analysis Process
  - Data Wrangling
- Hands-on

Today's session

- Big Data Analysis Process
  - Modeling
- Hands-on

Big Data Analysis Process – Main Steps

- Data access
- Data pre-processing / cleaning
- Data transformation / manipulation
- Feature selection
- Feature extraction
- Feature engineering
- Model choice and training
- Model evaluation and tuning
- Model deployment



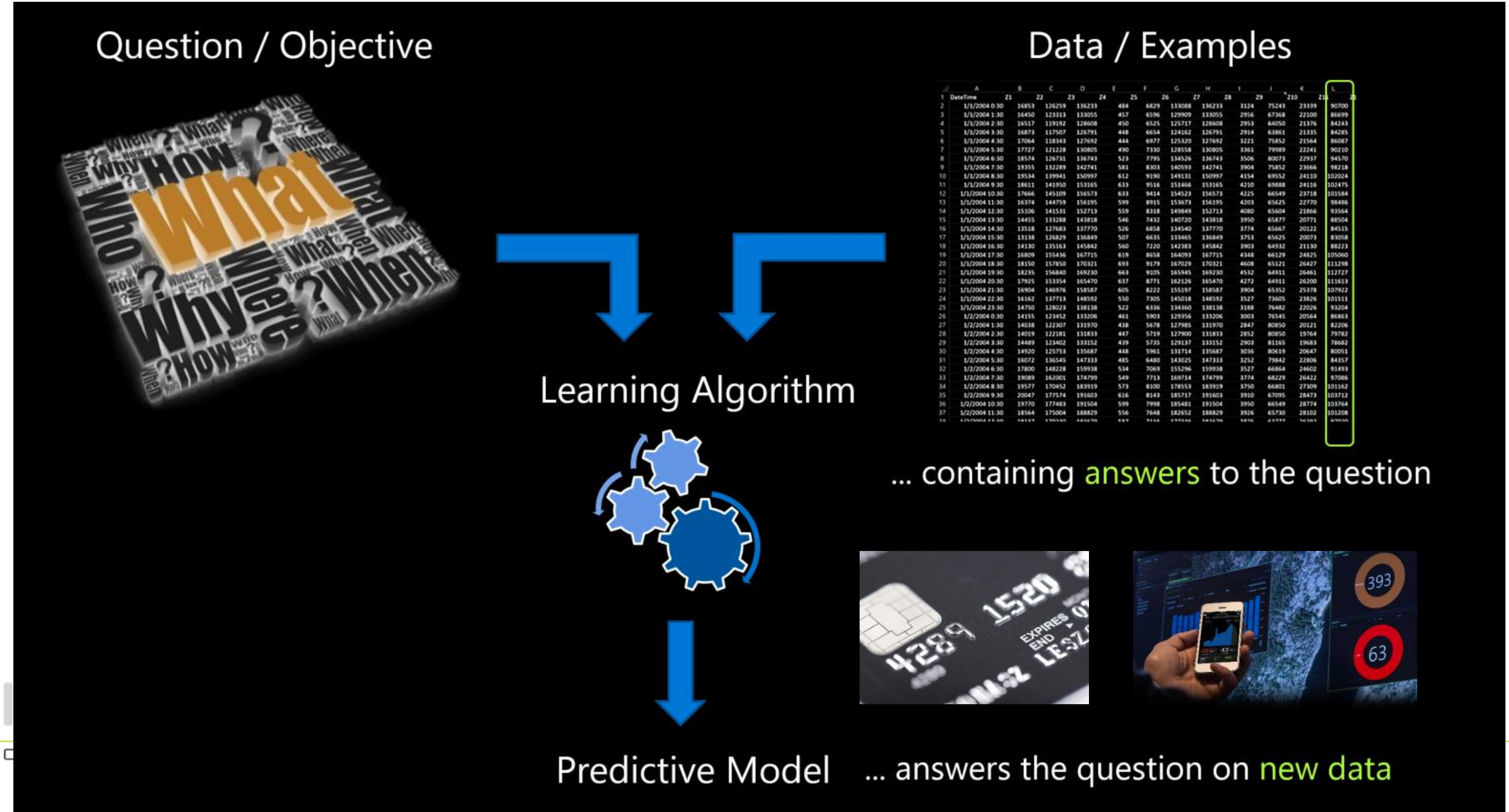
Modeling

Task Types

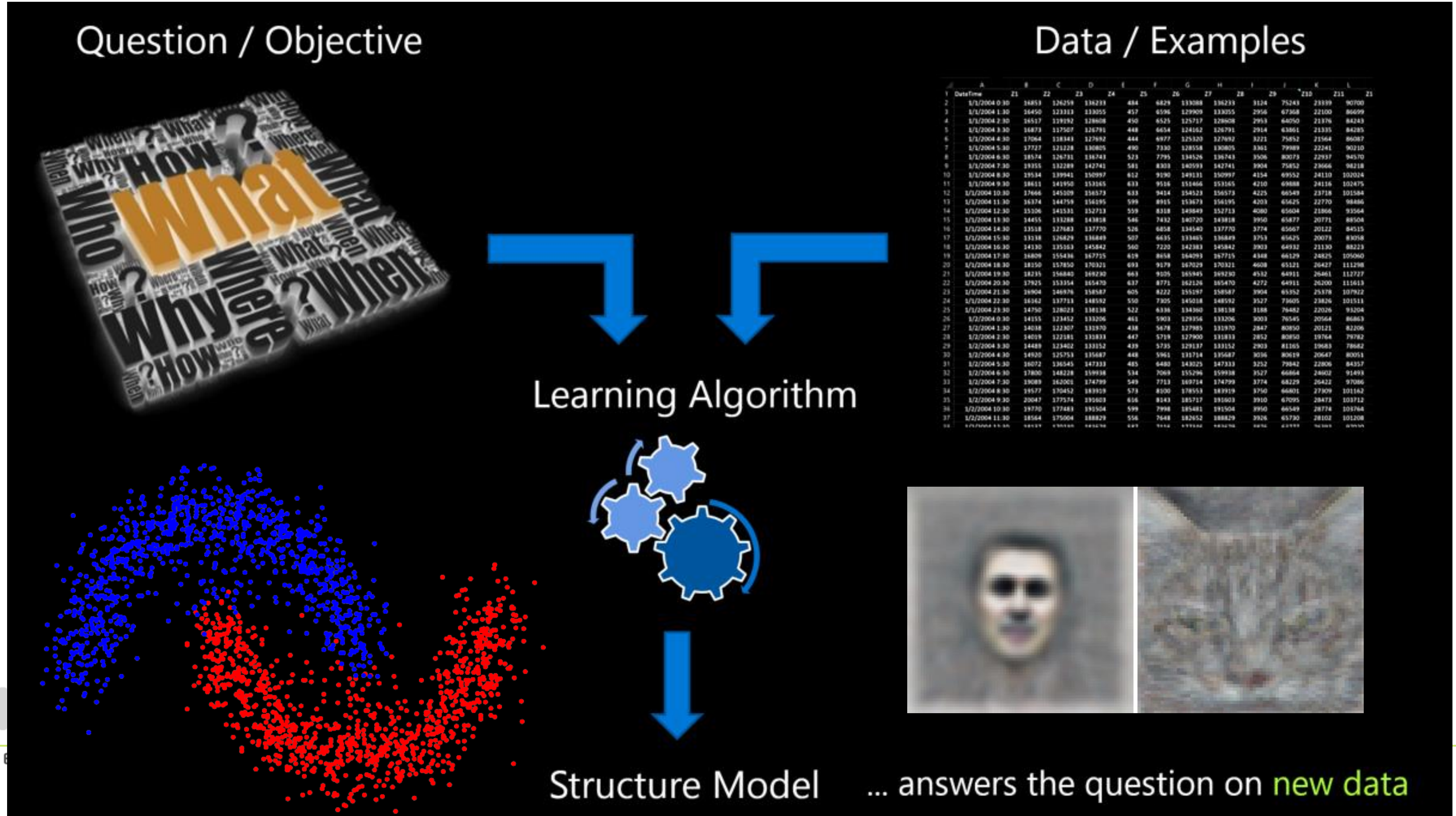
- Supervised Learning
  - Regression
  - Classification
- Unsupervised Learning
  - Clustering
  - Feature learning
- Reinforcement Learning
  - Control

Supervised Learning – Teacher <-> Student

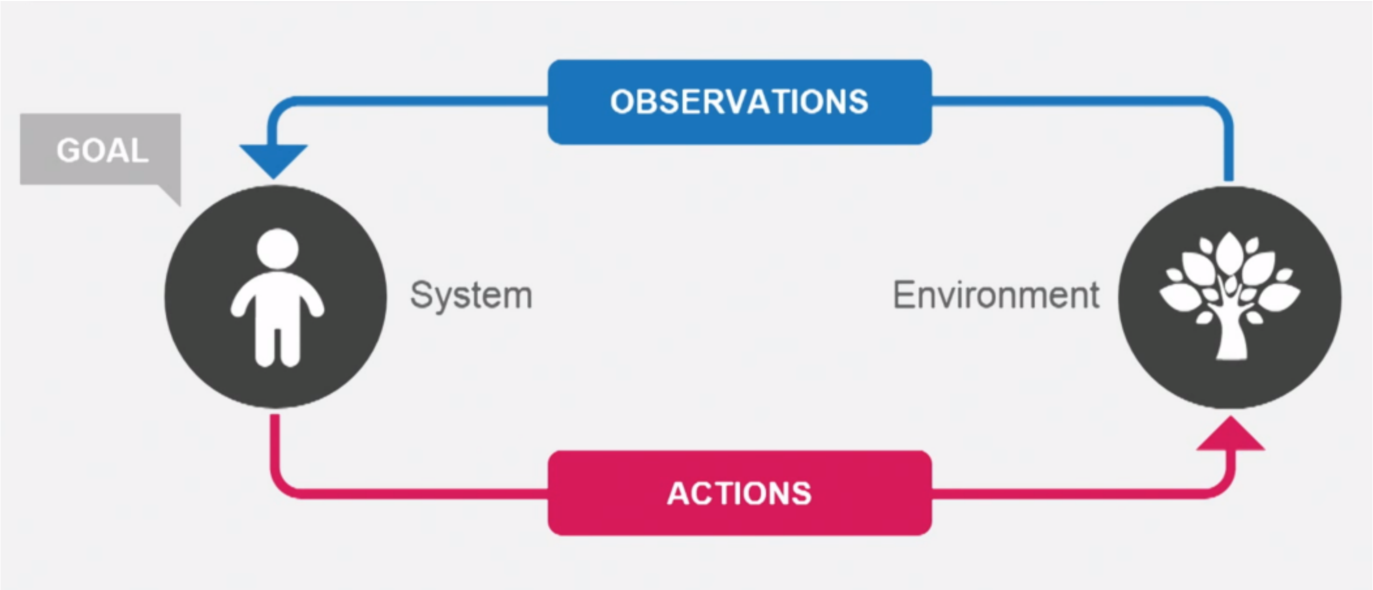
eSmart SYSTEMS

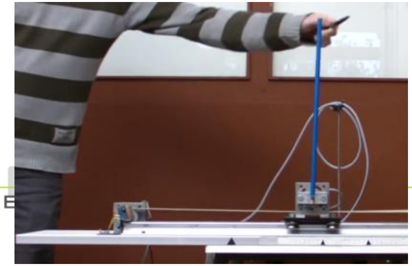


Unsupervised Learning - Explorer



Reinforcement Learning - Experimenter





We shall focus on Supervised Learning

By far the most used type of machine learning

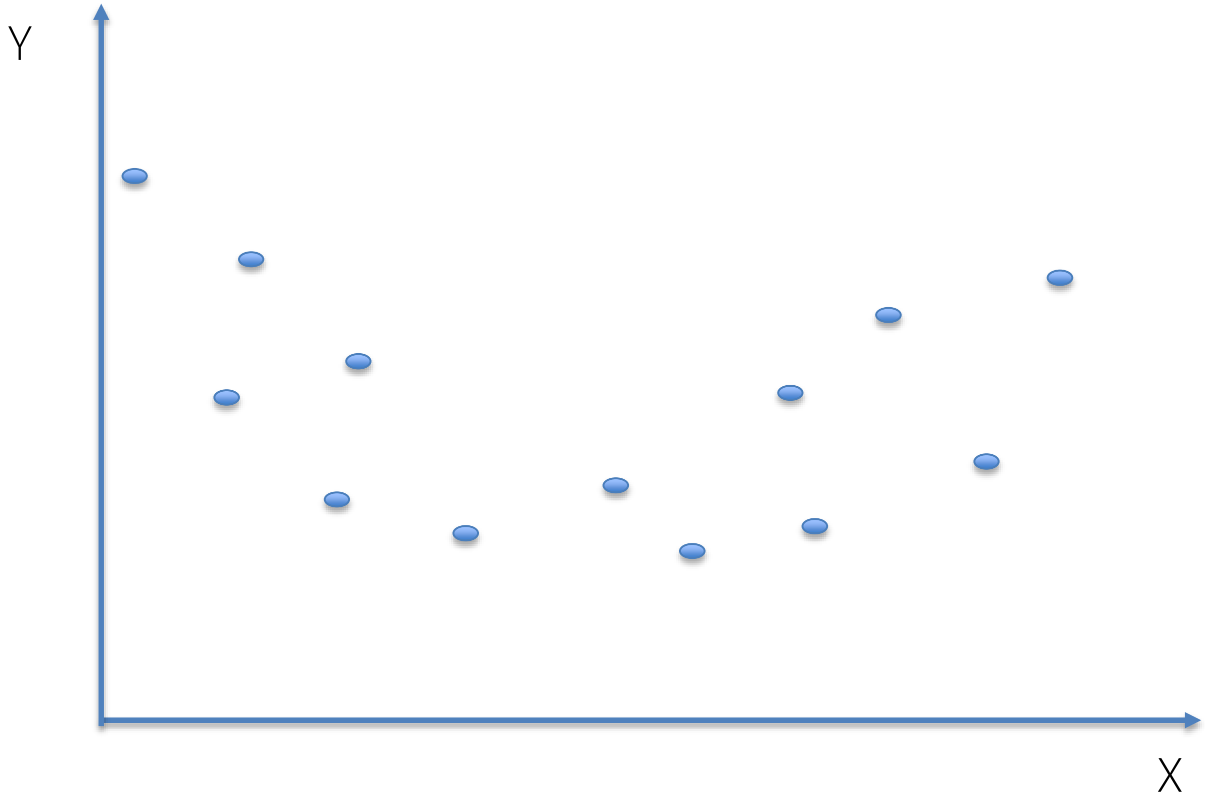
Supervised Learning – Important Concepts

- Generalization Error
- Overfitting problem
- Bias & Variance tradeoff
- Cross Validation
- Class imbalance problem
- Ensemble modeling

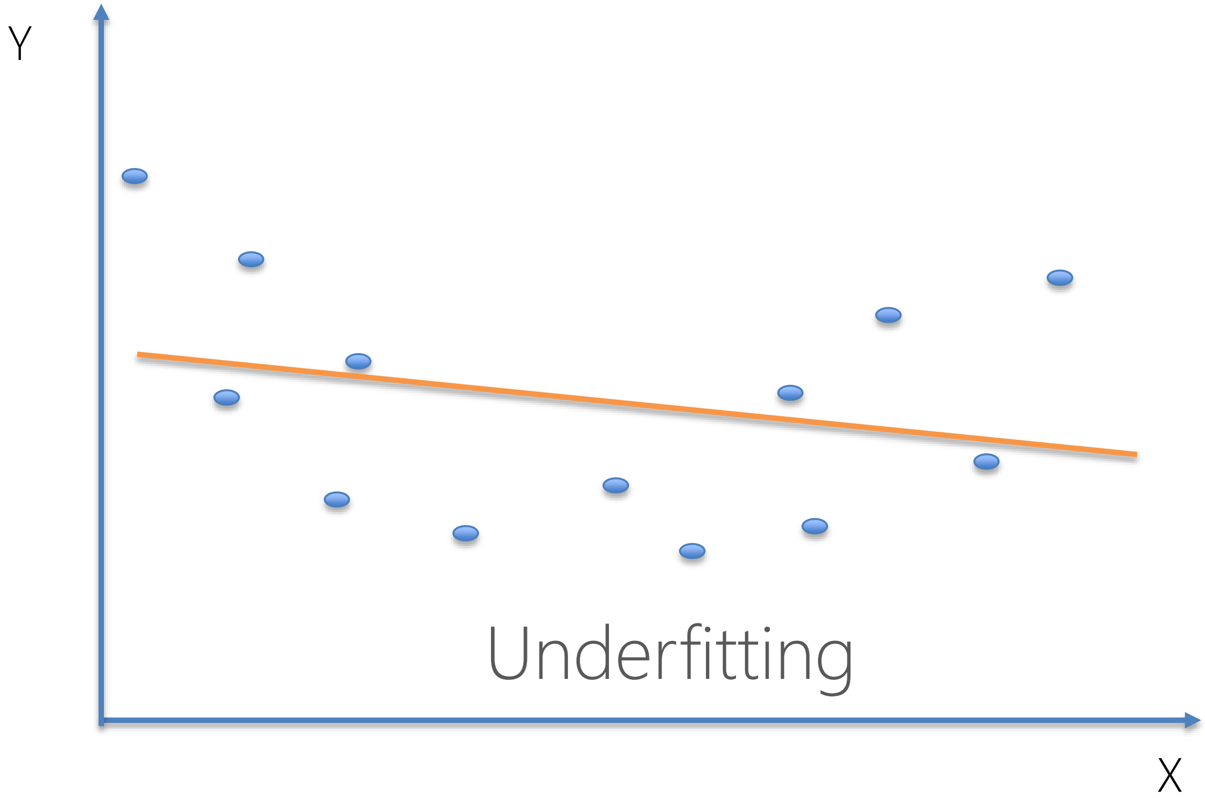
Generalization Error

- Measures how well a machine learning model "generalizes"
- Measures how well it will perform on data it has never seen before
- Can be measured specifically for a new data set (test error)
- Can be estimated for unseen data
  (see Cross-Validation)

Overfitting in Regression / Estimation

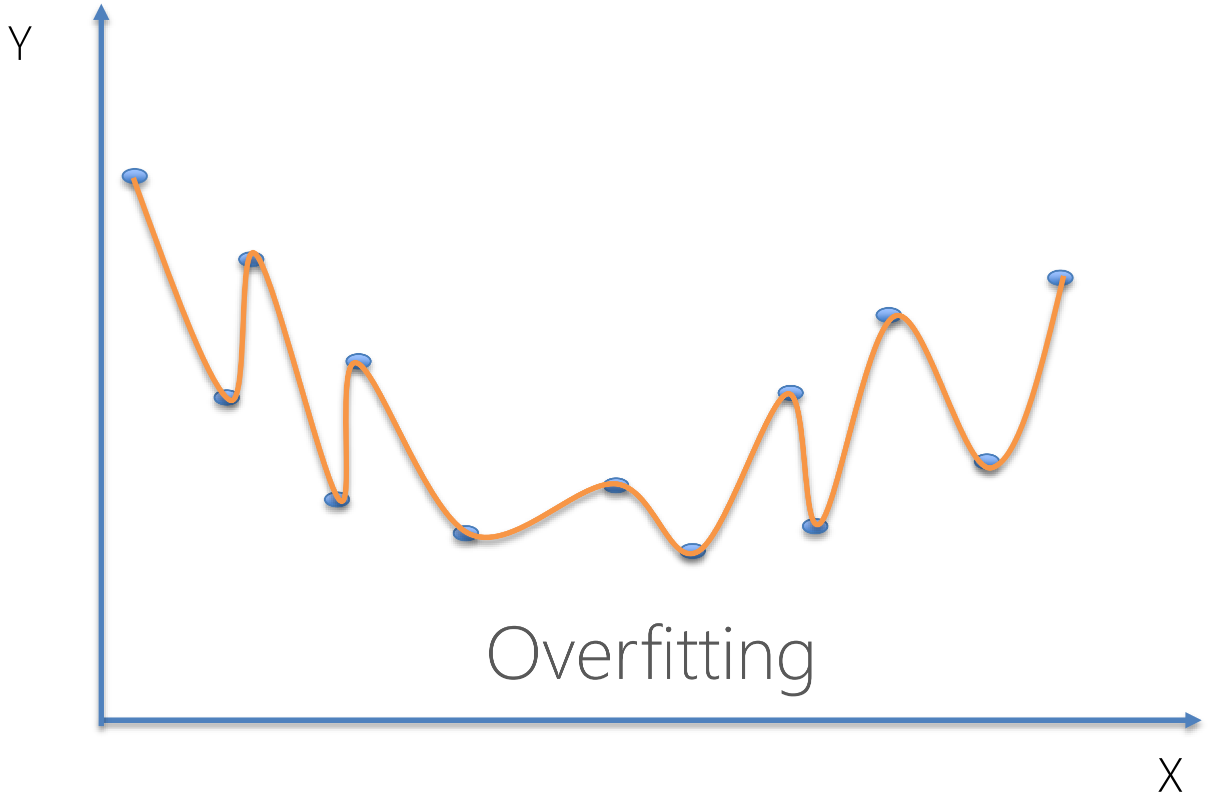


Overfitting



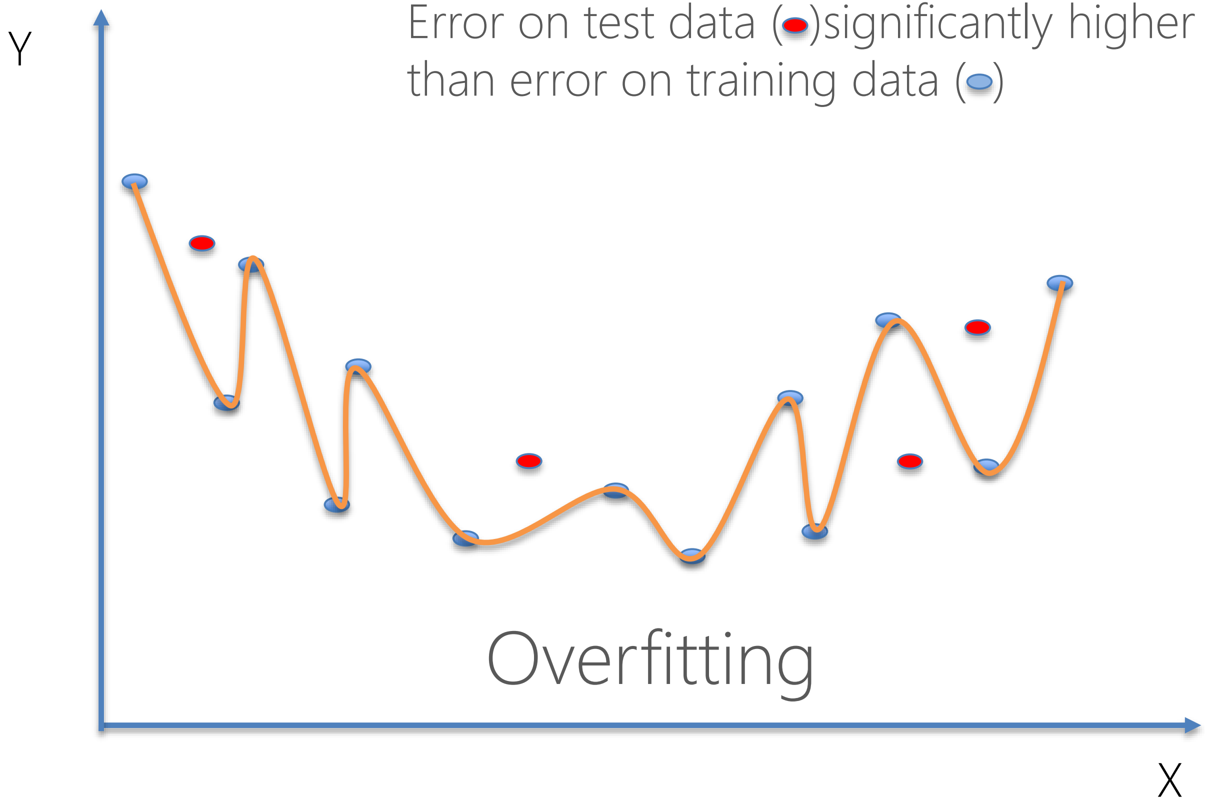
Overfitting

Overfitting

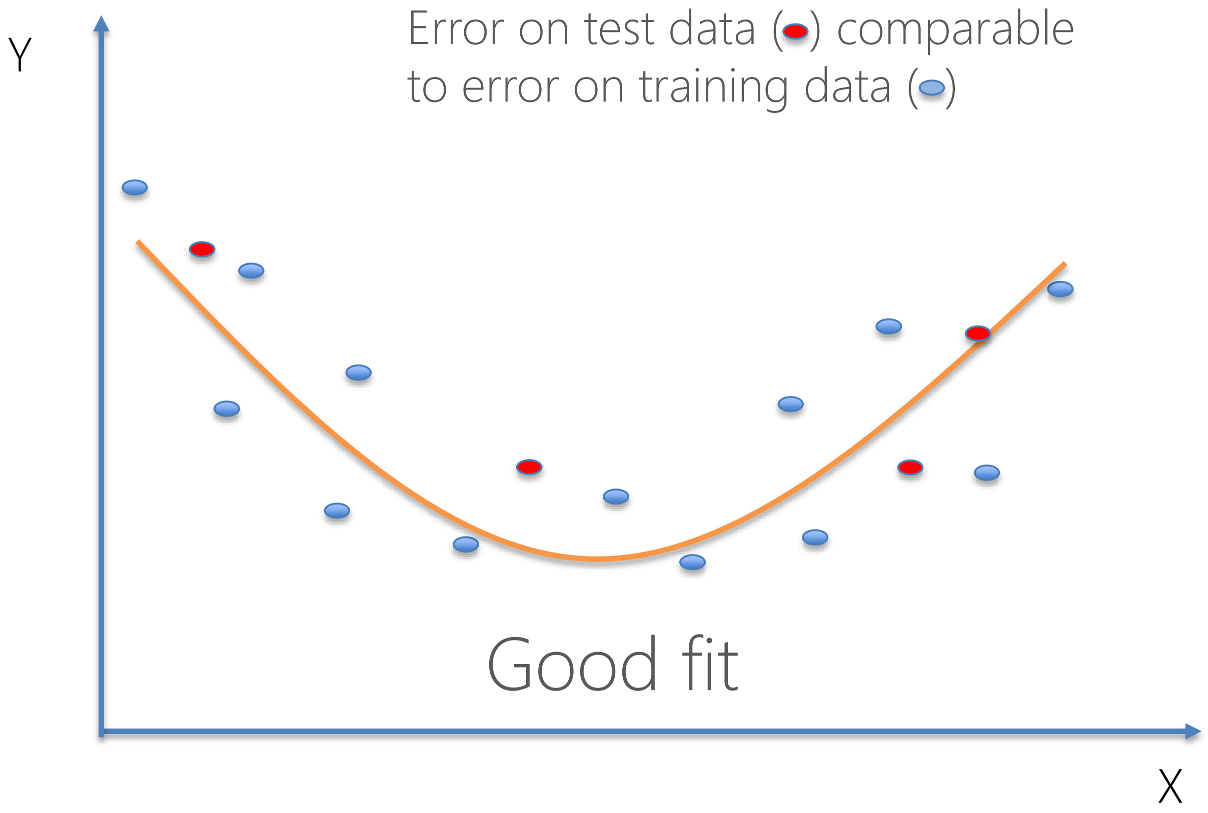


Detecting Overfitting

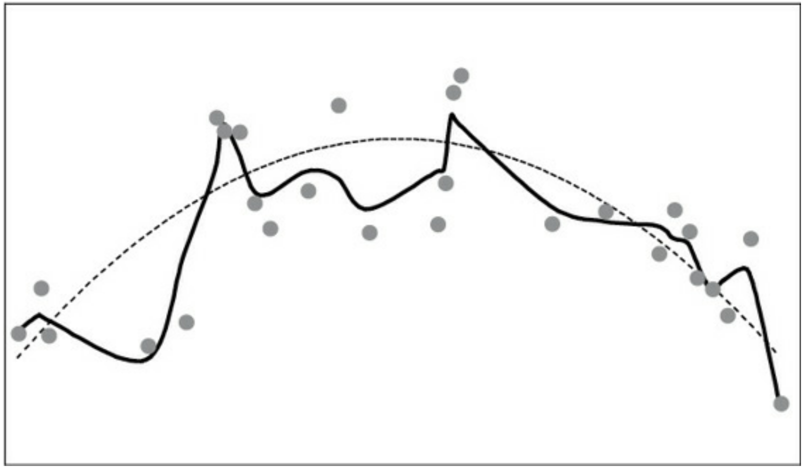
Error on test data (●)significantly higher than error on training data (●)



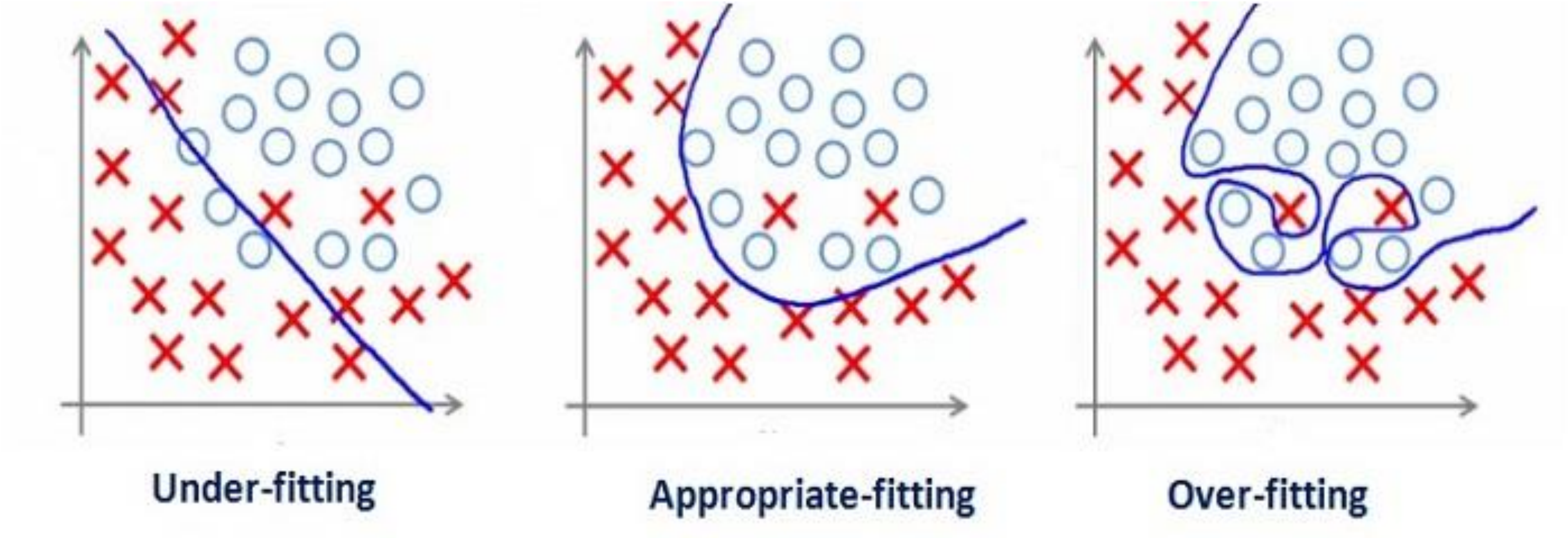
Overfitting



Overfitting ≈ Modeling the noise in the data



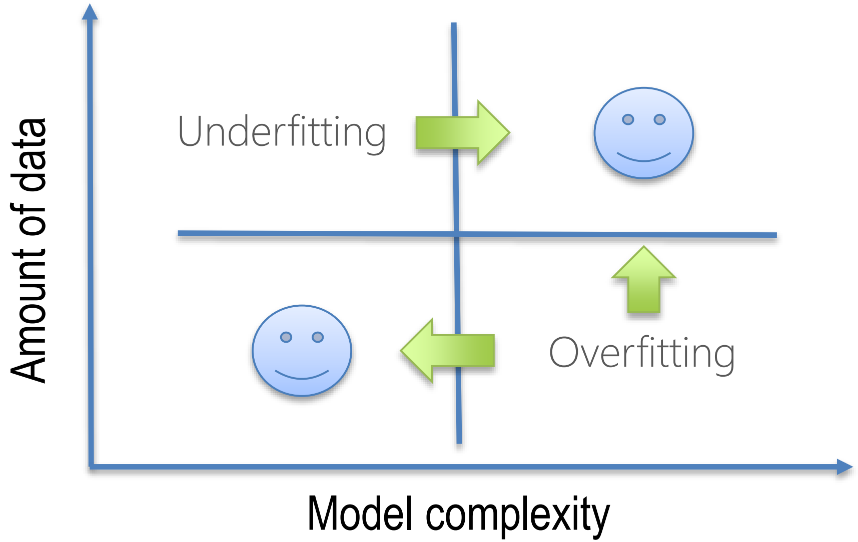
Overfitting in Classification



Reasons for Overfitting

- Too little data

- Too complex model (too many free parameters)



Some solutions to overfitting

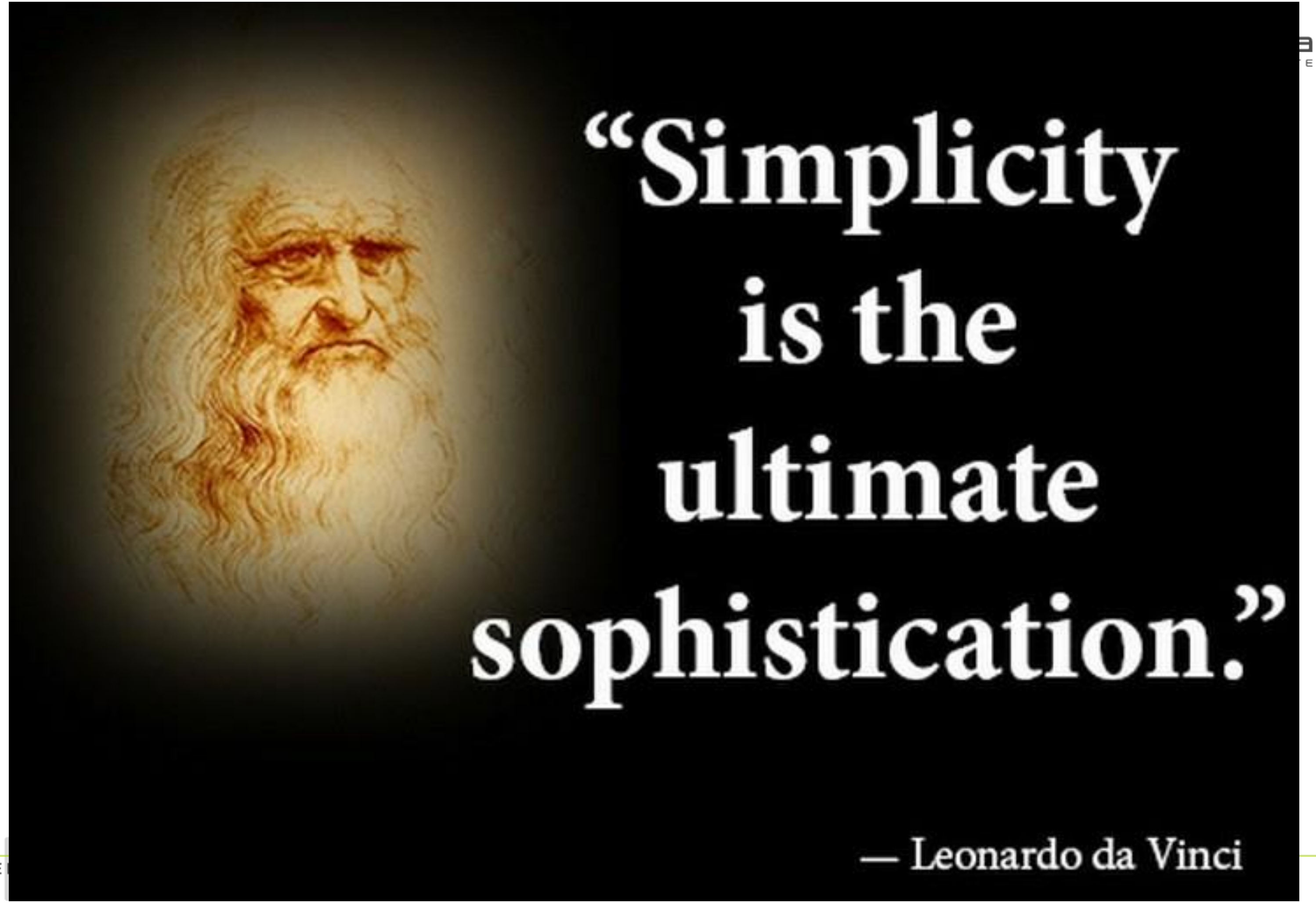
- Increase the amount of data
- Use a simpler model / Decrease number of parameters to tune (train)
  (see Occam's razor)
- Do not "overtune" parameters
  (e.g. Early stopping in Neural Networks)
- Integrate over many predictors
  (see Ensemble models)

Occam's Razor

1287 – 1347

"Among competing hypotheses, the one with the fewest assumptions should be selected"

Select the simplest model (the one with fewer parameters) ….. that still gives adequate performance on the available data


"Simplicity is the ultimate sophistication."
— Leonardo da Vinci

THE KISS PRINCIPLE

K I S S

eSmart SYSTEMS

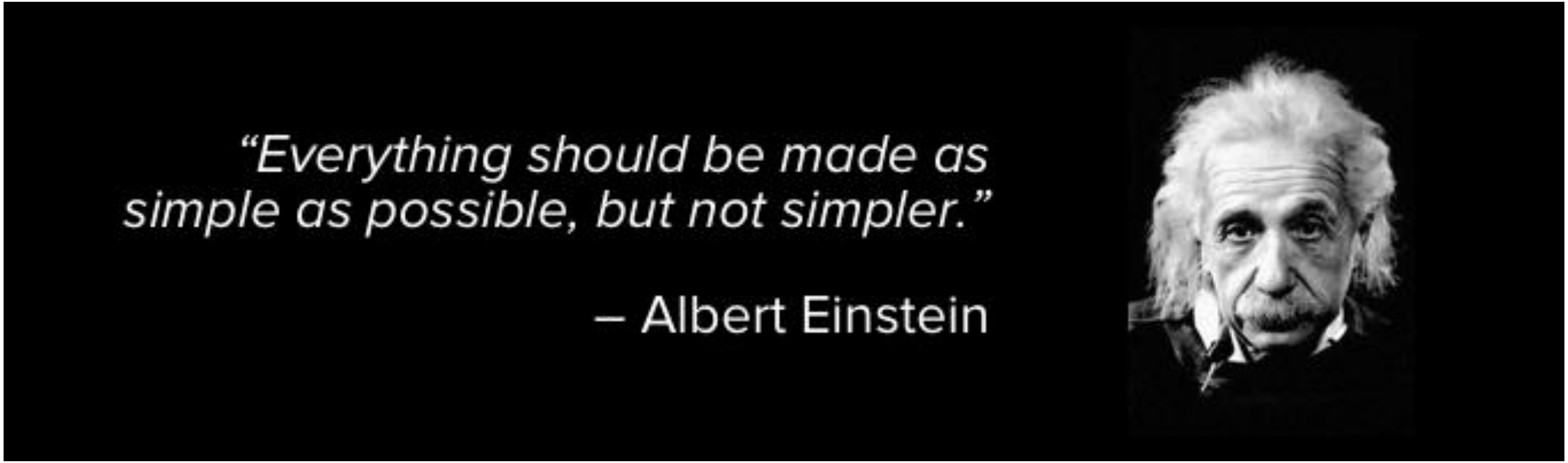
THE KISS PRINCIPLE

KEEP IT SIMPLE, STUPID

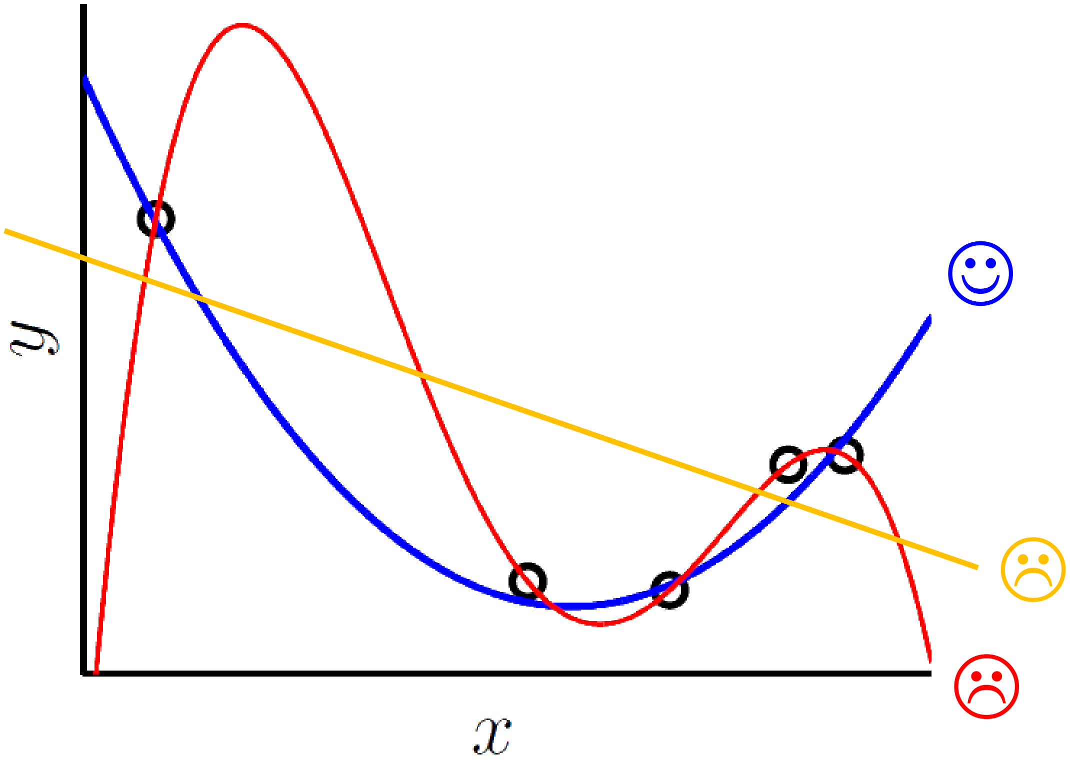
The KISS design principle states that most systems work best if they are kept simple rather than made complicated

Originated at Lockheed "Skunk Works" in the '60

U-2, Blackbird, Nighthawk, Raptor


"Everything should be made as simple as possible, but not simpler."
– Albert Einstein

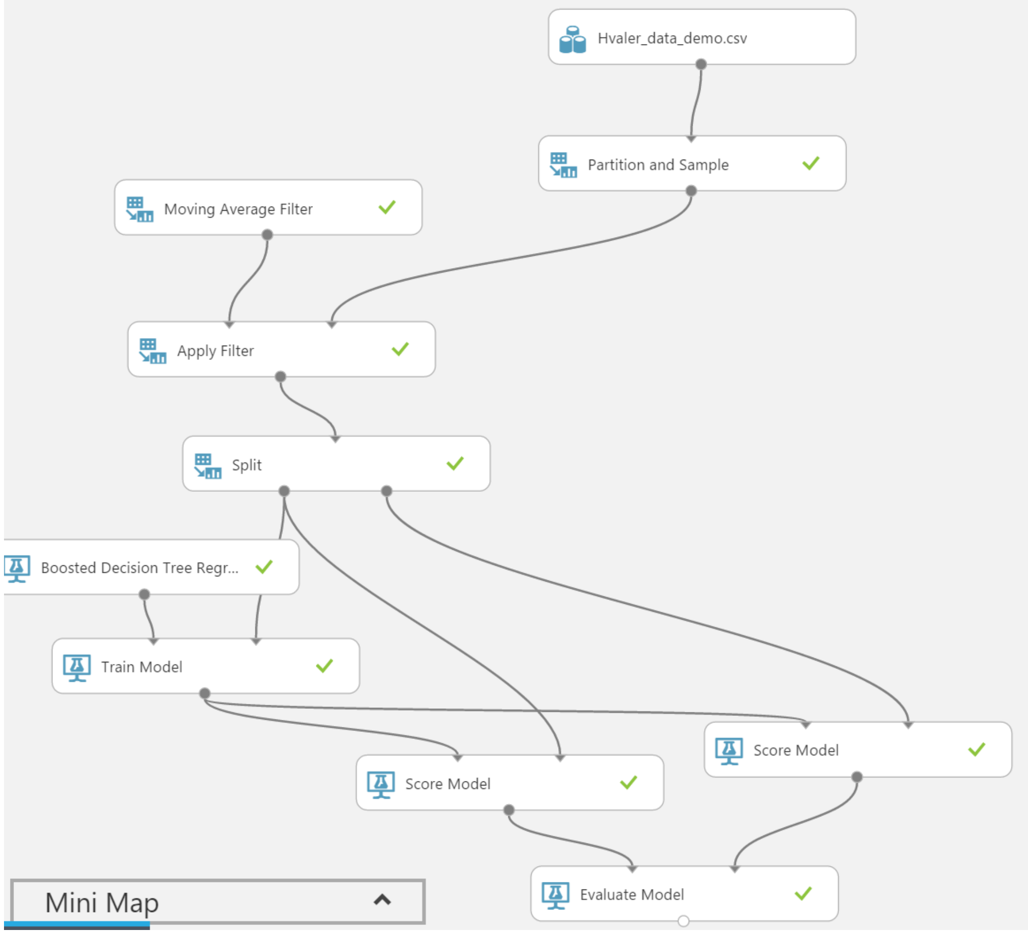
Reduce overfitting with Occam's razor



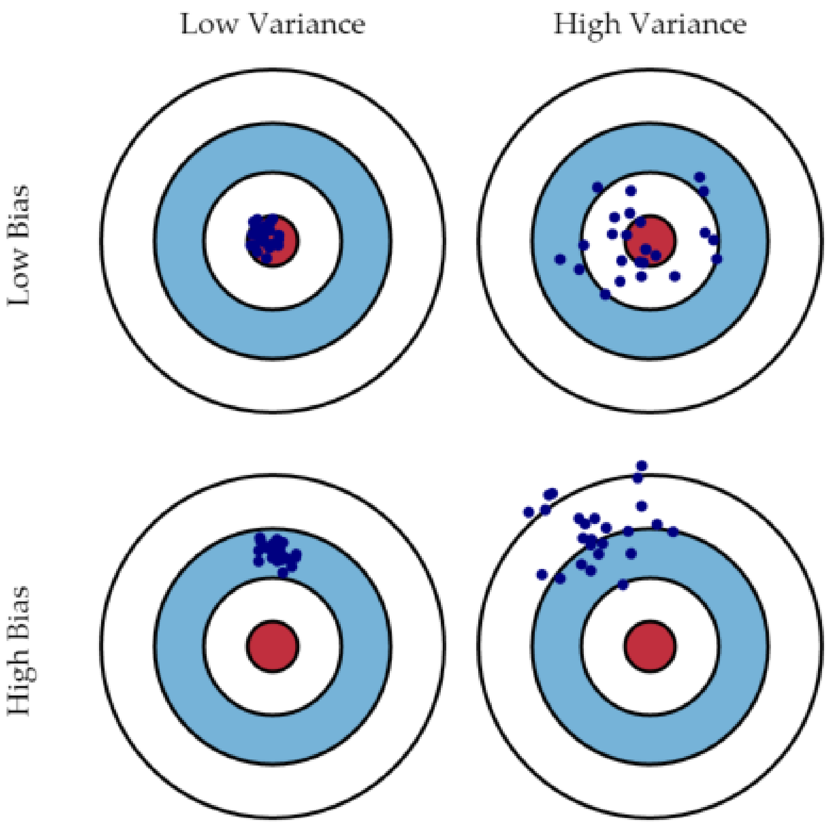
Regularization

- Attempts to automatically implement Occam's razor
- Penalize complex models
  - Define a "penalty function" to quantify complexity of the model
    - E.g. number of free parameters
  - Since most of the training algorithms are at heart optimization problems where the model error is minimized, one adds a "complexity penalty term" and minimize the whole expression together
    - E.g. minimize error and number of free parameters at the same time

Overfitting – Hands on



Bias & Variance

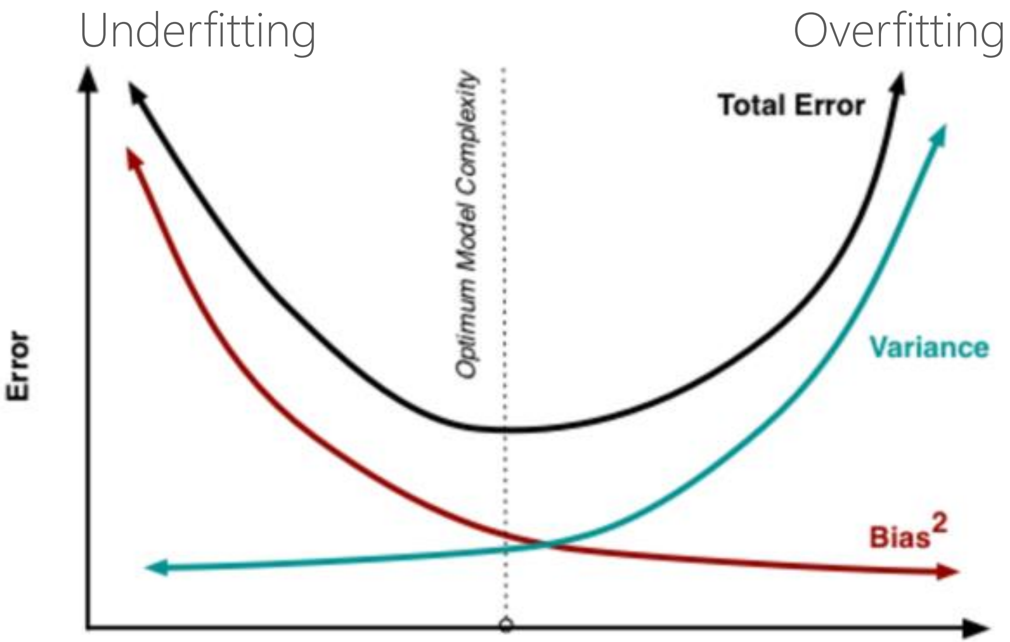


http://scott.fortmann-roe.com/docs/BiasVariance.html

Underfitting ≈ Not modeling the data ≈ High Bias

Overfitting ≈ Modeling the noise ≈ High Variance

Bias & Variance



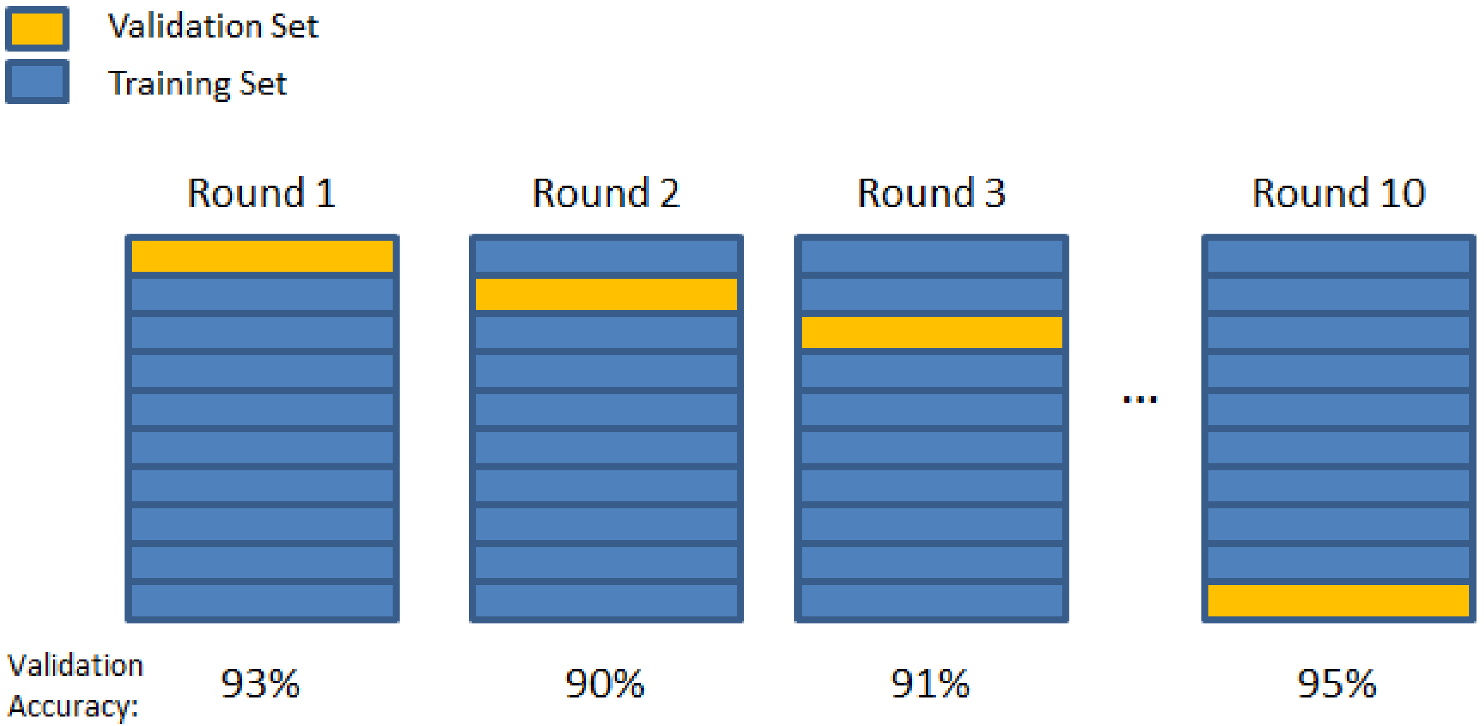
Model Complexity
or
Parameter tuning

http://scott.fortmann-roe.com/docs/BiasVariance.html

Cross Validation

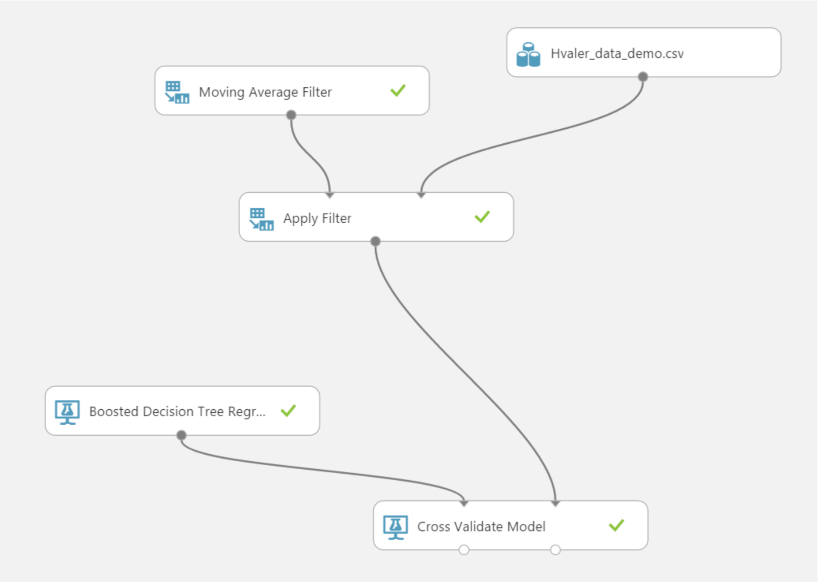
- Method to estimate the generalization error of a machine learning model/algorithm
- Make multiple models on different subsets of the available data
- Average the test results to produce an estimate of the generalization error
- https://msdn.microsoft.com/library/azure/75fb875d-6b86-4d46-8bcc-74261ade5826

10-fold Cross Validation



Final Accuracy = Average(Round 1, Round 2, ...)

Cross validation– Hands on



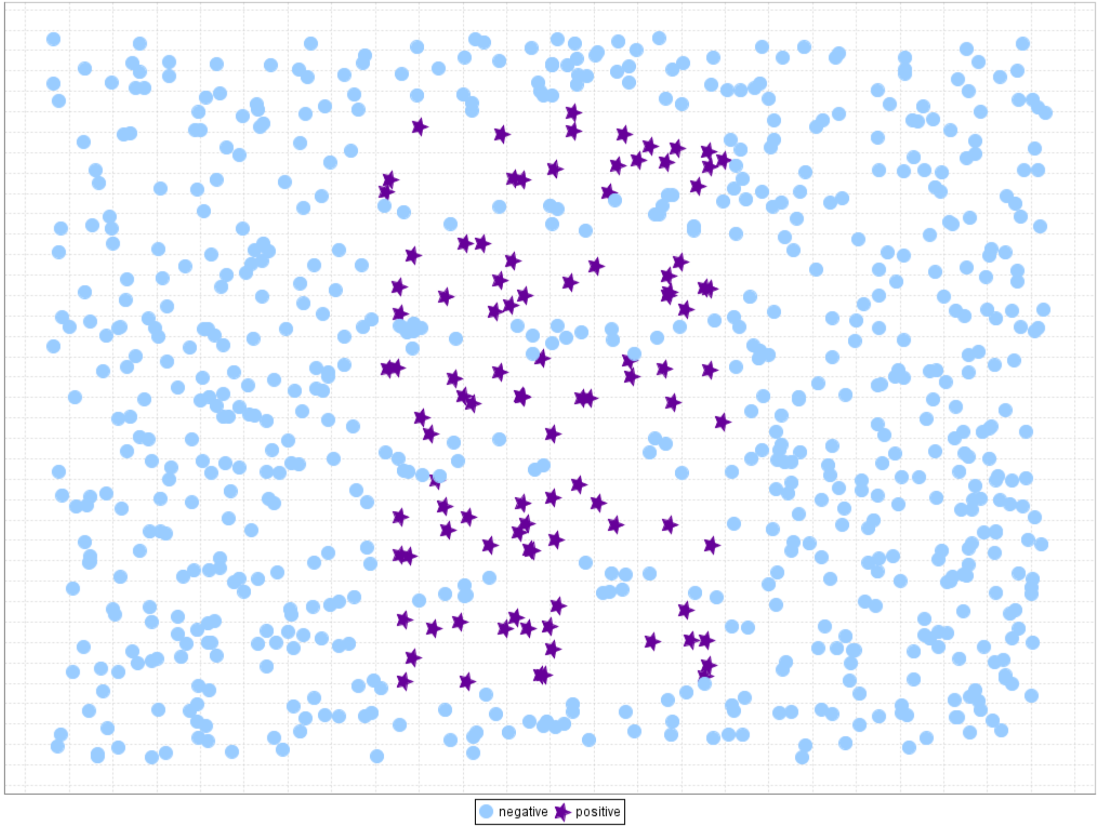
Class imbalance problem

Imagine we have

```
9900 examples of class A
 100 examples of class B
```

A model that always predicts A will be 99% accurate

This is obviously not what we want!



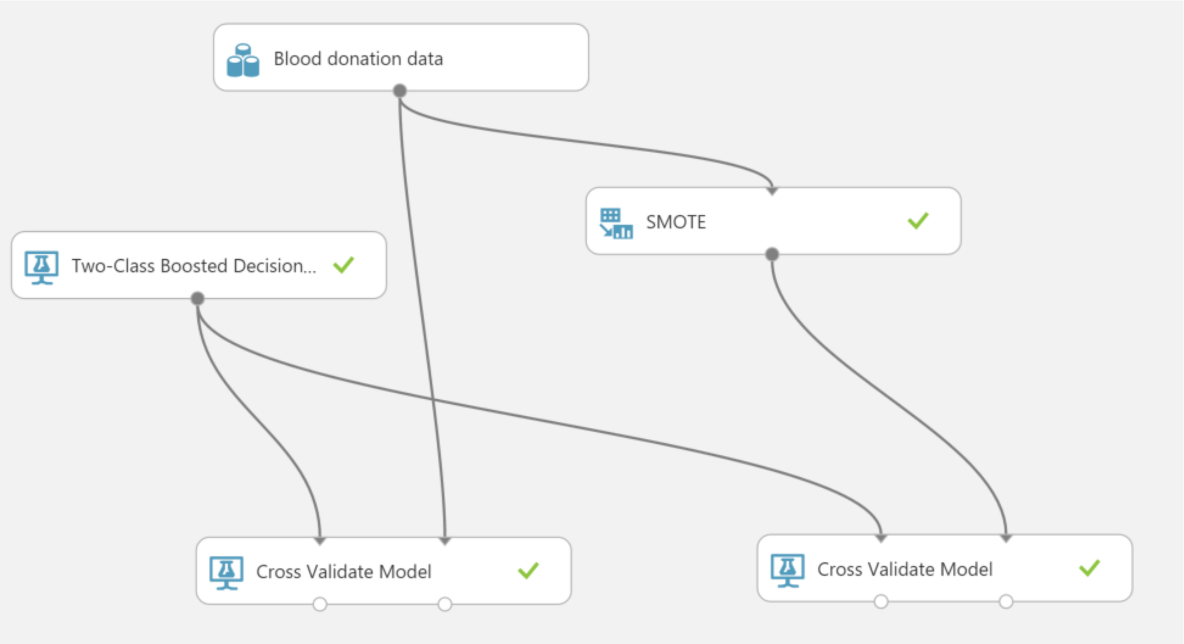
POWERED BY VALUES

Some solutions the class imbalance problem

- Downsampling
  - Select a subsample of the A examples such that it's size matches the set of B examples
- Upsampling
  - Produce artificial B examples ones until its size matches the set of A examples
  - SMOTE (Synthetic Minority Over-sampling Technique)
  - https://msdn.microsoft.com/library/azure/9f3fe1c4-520e-49ac-a152-2e104169912a

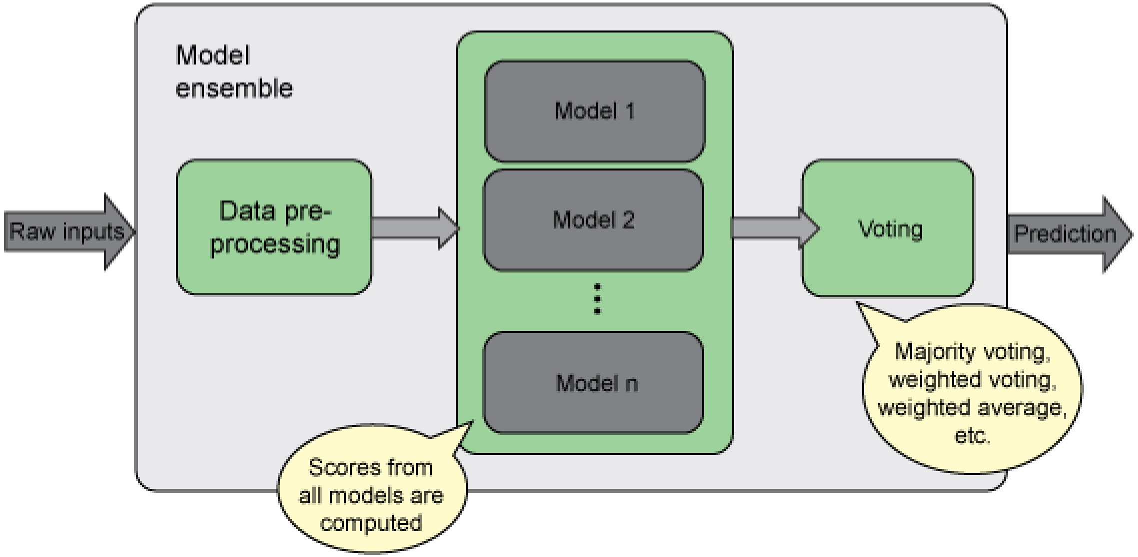
- Asymmetric training
  - Use an asymmetric error function to artificially balance the training process

Class Imbalance – Hands on



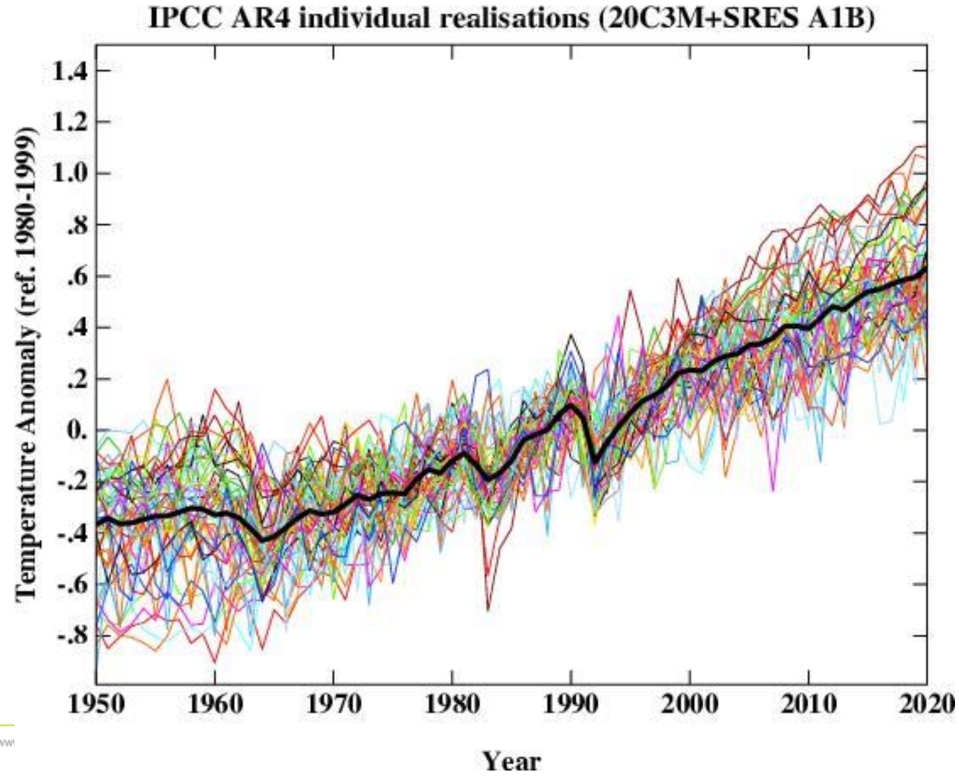
Ensemble modeling

"The Many Are Smarter Than the Few"



Ensemble Modeling

- Effectively reduces the variance of the model estimate
- Mitigates overfitting
- The diversity of the individual models is key



Ensemble learning methods

- Bagging
  - Each model is trained on a random subset of the data
  - Averaged
  - (https://msdn.microsoft.com/en-us/library/azure/dn905862.aspx)

- Boosting
  - Incrementally add new models
  - Data previously estimated or classified unaccurately is given higher "weight" in the new model to be added to the ensemble
  - (https://msdn.microsoft.com/en-us/library/azure/dn905801.aspx)

- Stacking
  - Each model is trained on a random subset of the data
  - A new model weights the contributions of each individual model

Machine Learning in ML Studio

Anomaly Detection
- One-class Support Vector Machine
- Principal Component Analysis-based Anomaly Detection
- Time Series Anomaly Detection*

Classification
- **Two-class Classification**
  - Averaged Perceptron
  - Bayes Point Machine
  - Boosted Decision Tree
  - Decision Forest
  - Decision Jungle
  - Logistic Regression
  - Neural Network
  - Support Vector Machine
- **Multi-class Classification**
  - Decision Forest
  - Decision Jungle
  - Logistic Regression
  - Neural Network
  - One-vs-all

Clustering
- K-means Clustering

Recommendation
- Matchbox Recommender

Regression
- Bayesian Linear Regression
- Boosted Decision Tree
- Decision Forest
- Fast Forest Quantile Regression
- Linear Regression
- Neural Network Regression
- Ordinal Regression
- Poisson Regression

Statistical Functions
- Descriptive Statistics
- Hypothesis Testing T-Test
- Linear Correlation
- Probability Function Evaluation

Text Analytics
- Feature Hashing
- Named Entity Recognition
- Vowpal Wabbit

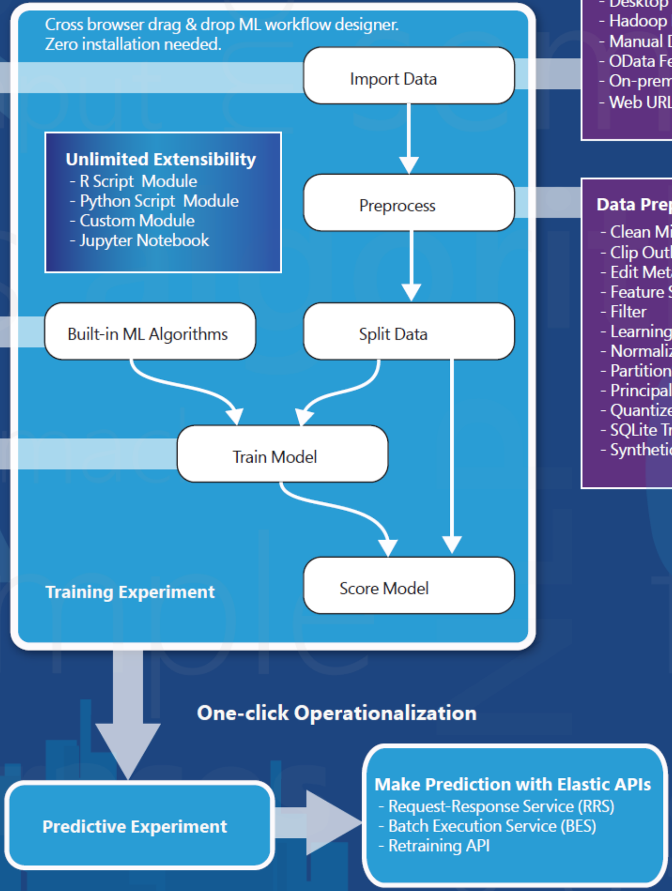
Computer Vision
- OpenCV Library

Data/Model Visualization
- Scatterplots
- Bar Charts
- Box plots
- Histogram
- R and Python Plotting Libraries
- REPL with Jupyter Notebook
- ROC, Precision/Recall, Lift
- Confusion Matrix
- Decision Tree*

Training
- Cross Validation
- Retraining
- Parameter Sweep

https://studio.azureml.net

Guest Access Workspace:	Free trial access without logging in.
Free Workspace:	Free persisted access, no Azure subscription needed.
Standard Workspace:	Full access with SLA under an Azure subscription.



Data Source
- Azure Blob Storage
- Azure SQL DB
- Azure SQL DW*
- Azure Table
- Desktop Direct Upload
- Hadoop Hive Query
- Manual Data Entry
- OData Feed
- On-prem SQL Server*
- Web URL (HTTP)

Data Format
- ARFF
- CSV
- SVMLight
- TSV
- Excel
- ZIP

Data Preparation
- Clean Missing Data
- Clip Outliers
- Edit Metadata
- Feature Selection
- Filter
- Learning with Counts
- Normalize Data
- Partition and Sample
- Principal Component Analysis
- Quantize Data
- SQLite Transformation
- Synthetic Minority Oversampling Technique

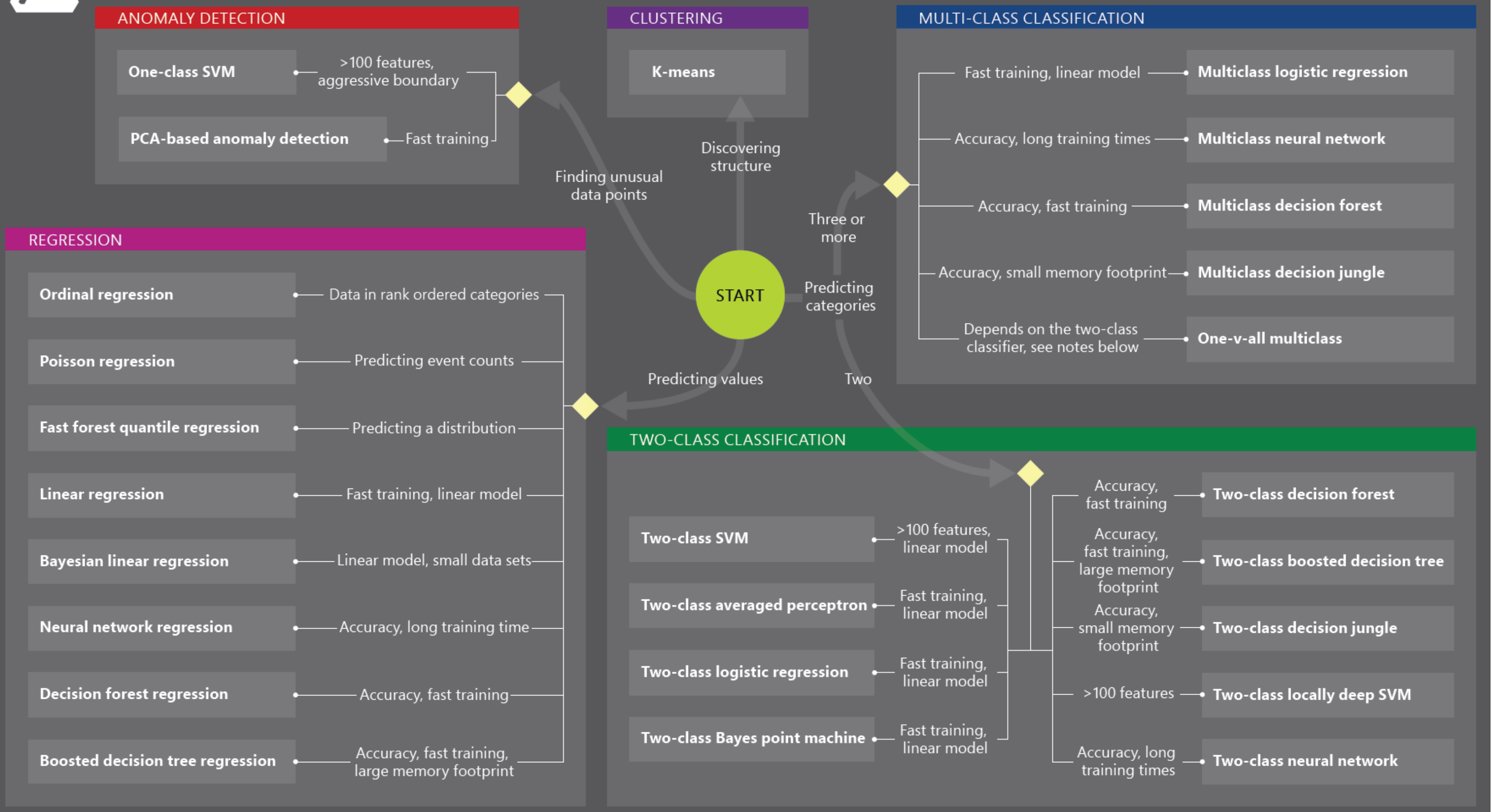
Enterprise Grade Cloud Service
- SLA: 99.95% Guaranteed Up-time
- Azure AD Authentication
- Compute at Large Scale
- Multi-geo Availability
- Regulatory Compliance*

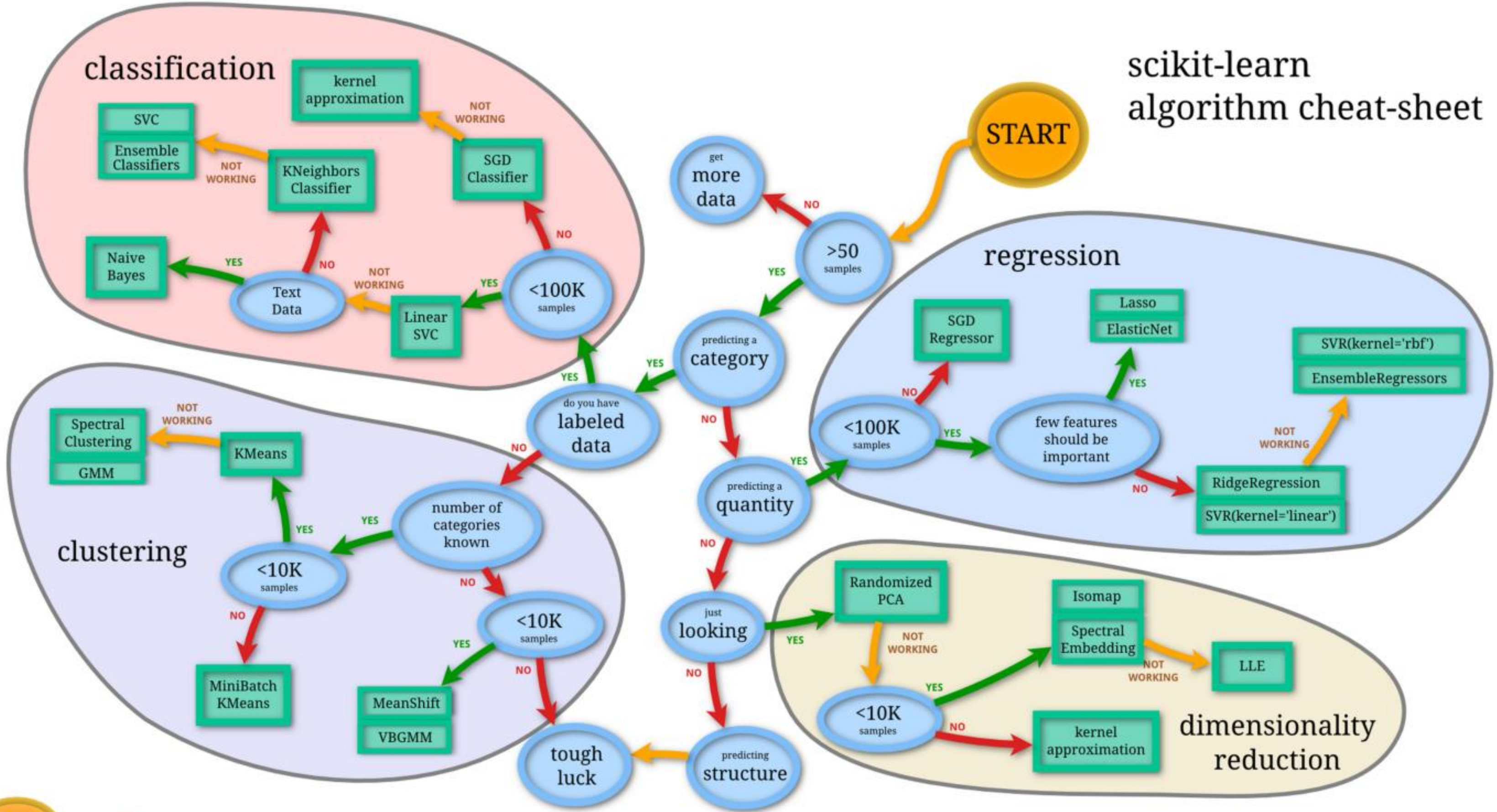
Community
- Gallery (http://gallery.azureml.net)
- Samples & Templates
- Workspace Sharing and Collaboration
- Live Chat & MSDN Forum Support

* Feature Coming Soon

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.




scikit-learn
algorithm cheat-sheet
START
>50 samples
get more data
predicting a category
do you have labeled data
<100K samples
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC
Ensemble Classifiers
SGD Classifier
kernel approximation
classification
number of categories known
<10K samples
KMeans
Spectral Clustering
GMM
MiniBatch KMeans
MeanShift
VBGMM
clustering
tough luck
predicting structure
predicting a quantity
<100K samples
SGD Regressor
few features should be important
Lasso
ElasticNet
RidgeRegression
SVR(kernel='linear')
SVR(kernel='rbf')
EnsembleRegressors
regression
just looking
Randomized PCA
<10K samples
Isomap
Spectral Embedding
LLE
kernel approximation
dimensionality reduction
YES
NO
NOT WORKING

Back

1. Fundamentals
2. Statistics
3. Programming
4. Machine Learning
5. Text Mining / NLP
6. Visualization
7. Big Data
8. Data Ingestion
9. Data Munging
10. Toolbox
Author: Swami Chandrasekaran