# Projects in Big Data Analytics

ISYE 4961/6961

Lecture 3

# Introduction to Analytics Methods

- Math Review
  - Statistics
    - Mean
    - Standard Deviation
  - Correlation / Variance / Covariance
  - Principal Component Analysis (PCA)
  - Cluster Analysis

# Single Channel - Univariate

- Treated as an independent set of observations along a single axis – one data dimension
- Data preparation
- Noise reduction
- Missing data mitigation
- Bad data removal/repair
- Scaling, mean removal
- Derived data attributes

# Missing data

- Interpolation and extrapolation
  - Sync
  - Spline
- Zero filling
- Extracting, cropping

# Uniform Data and Linearity

- Normal or Gaussian distribution – values vary due to random chance
- Linearity – quantities are directly proportional, straight line relationship
- If we hand normally distributed data and linear relationships, we can perform lots of useful operations
- What if the data is not like this?
- Options:
  - Linearize
  - Remove offending parts of the data
  - Ignore and go ahead, keeping in mind the possibility for problems
  - Find methods that assume other distributions or are non-linear

# R functions for data statistics

- Plot: plot(data)
- Mean removal and scaling: scale(data, center=, scale=)
- Histogram: hist(data)
- Distribution inspection: qqnorm(data);qqline(data)
- Crossplot two vectors: plot(data1,data2)
- Crossplot several vectors: pairs(data_array)

# Project 1 – Using Wind Turbine Data

- Import files and extract some data columns into vectors
- Extract column segments
- Build some multi-column arrays
- Display and inspect the data (note: failure occurs at the start of data and time goes backward from there)
- Remove mean and scale, do some correlations and cross plots
- Look at groups of similar sensors, look at pairs of dissimilar sensors
- Look at different time blocks
- Look at same sensor across several turbines
- Report findings

# Questions to Answer

- Which sensors are the most important?
- What are the distinguishing features of the data?
- Can we see any indications of failure in the raw data? After mean removal and scaling? In the histograms? In cross plots?
- What does normal operation look like and can we learned anything about regular operational performance?
- Ultimately, can we use the data to anticipate a failure?

# Multi-channel – Multivariate

- Multidimensional data preparation
- Multidimensional transformations
- Derived data attributes
- Correlation, classification, covariance