Projects in Big Data Analytics

ISYE 4961/6961

Lecture 2

Tom Morgan

- Cogswell 312
- Hours: 10-noon TF
- (office) 518-276-3887
- (cell) 281-753-1383
- Email: morgat5@rpi.edu

John Erickson

- CBIS
- Hours:
- (office) 518-276-4384
- Email: erickj4@rpi.edu

R web site https://www.r-project.org/

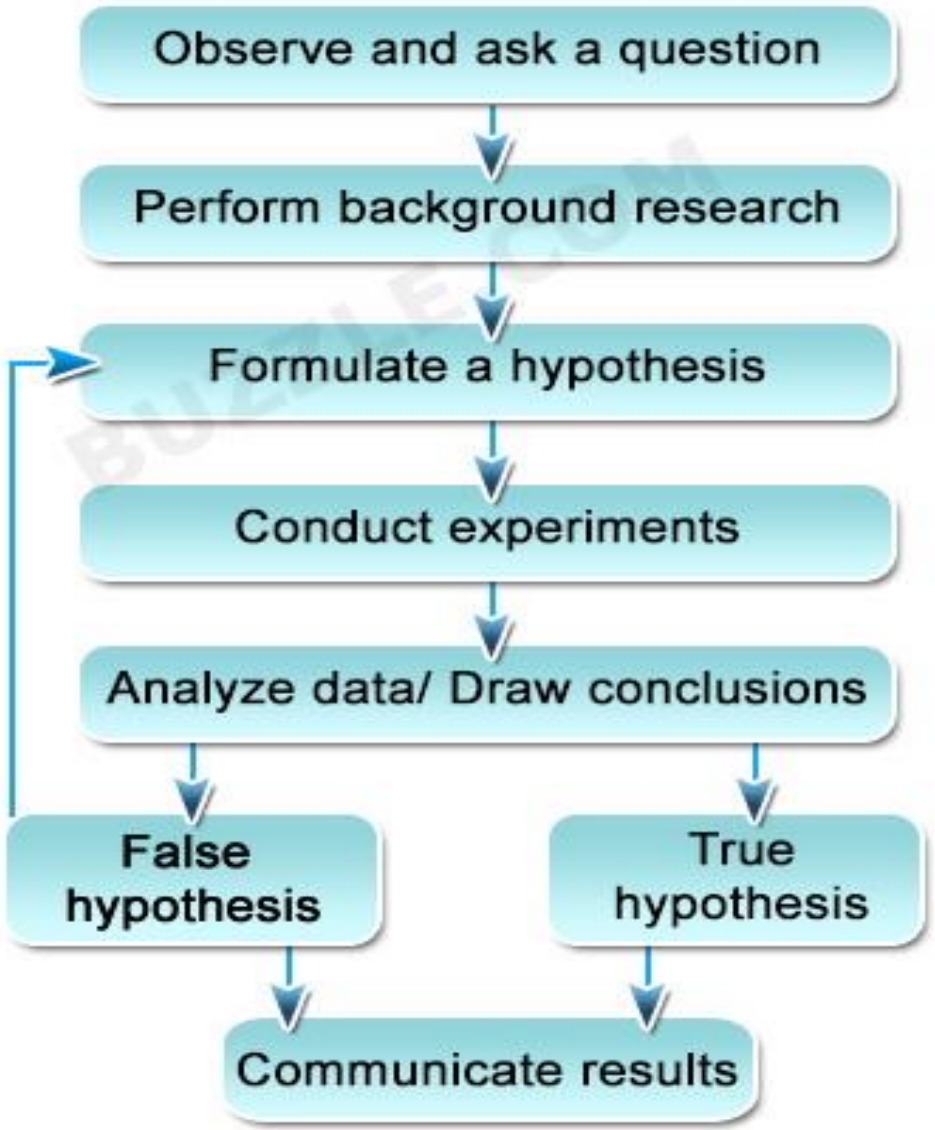
- Download and install
- Download manuals
  - Introduction
  - Data
  - Admin
  - Extensions

Introduction

The Scientific Method and Data Science

Data Analytics: Science or Engineering?

The Scientific Method



Scientific Method

- Vertical Process of Discovery
- Generally Discipline Specific
- Grew up in the experimental (quantitative) and descriptive physical sciences
- Taken up by the behavioral sciences
- Quantification expanded into descriptive physical sciences
- Quantification now expanding into behavioral sciences
- Initial drive by computer hardware revolution
- Now being driven by computer software revolution and sensor technology

Data Science

- Focused on the data
- Horizontal across disciplines
- Often no initial control of experimental design and data acquisition
- Often no strong initial problem specification
- Frequently will be concentrated in the first two boxes of the scientific method
- Also may be a re-visit to the Analyze/Draw Conclusions box of a past project

Data Science

- New way of doing science driven by simulation and modeling
- Already well entrenched in modern engineering practice (CAD/CAM, airframe testing, crash simulation, HVAC, earthquake tolerance)
- Driving recent rapid development in science (global climate, glacial melting and sea level rise, weather prediction)

Data Analytics

- The application of, usually mathematical, methods in data science
- Applied in one to many dimensions
- Applied to many data forms from raw output to complex derived products
- Applied to regularly and irregularly sampled numerical data, text, image, unstructured, mixed and multiple origin data sets of all sizes
- The part of data science that discovers patterns and correlations that feed into the establishment of cause and effect relationships needed for modeling and prediction

Data Analytics Process

- Acquisition
- Preparation
- Analysis
- Interpretation
- Validation

Contrast

- Data Analytics is exploration driven and concentrates on extracting useful information of many kinds from existing data
- Scientific Research is driven by the desire to understand a physical phenomenon and concentrates on discovery and analysis of the necessary data to support or refute an hypothesis
- Engineering Practice is driven by the need to solve a particular problem and concentrates on collection and processing of data needed to construct a viable solution
- Data Analytics can support either Scientific Research or Engineering Practice, or be an end in itself to generate new products and services

Problem Formulation

- Here is some data, what can we do with it?
- Here is a problem we have, how do we solve it? Can it be solved with analytics? If so, what data do we need and how can we use the data so solve our problem?
- Here is an analytics process, how do we construct a series of methods to effectively and efficiently produce consistent and reliable results?
- Here is an analytics method, what problem(s) can be solved with it?

Acquisition

- Planning and execution to assure data collection meets the needs of the project objectives and fulfills the requirements of the expected analyses.
- Data parameters such as type, sampling in time, space, or other abstract dimensions, overall volume, and dynamic range of values must be understood and the data collection process and methodology must be constructed to accommodate these parameters.
- In cases where the analyses are unknown or not well understood, the data acquisition and analysis cycle may need to be iterated.

Preparation

- Raw data is the input to the analytics process. It is a set of values (numeric, character, symbolic or abstract) and their units and other attributes with or without any particular organization.
- Basic processes applied to raw data can be used to organize, re-organize, clean, filter, transform or otherwise massage the data set content into a more useful state.
- Verification of data completeness, reliability, fitness for purpose
- Defect repair
- Sub-setting

Analysis

- Data processing methods are applied to generate derived data products that will be useful input during analysis.
- Analysis processes are applied to generate information in the form of patterns, trends, correlations or other coherent measures.

Interpretation

- Patterns, correlations and other coherency measures are visualized and studied.
- Cause and effect relationships are hypothesized.
- Validity tests are proposed.

Validation

- Mathematical relations, which can be used for modeling, prediction and other computational exercises.
- The meaning extraction process is completed when quantitative methods can be used to synthetically regenerate the input data within an acceptable level of error.
- A valid model leads to the ability to predict addition results, thus providing a means for modeling and studying additional data sets of the same type.

Analytics Methods

Mathematical Techniques

Introduction to Analytics Methods

- Math Review
  - Statistics
  - Correlation
  - Covariance
  - PCA
  - Cluster Analysis
  - Others (?)

Single Channel - Univariate

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

Noise

- Incoherent
  - Spikes
  - Background
- Coherent
  - 50/60 cycle
  - Unwanted data
- Methods
  - Bandpass, notch, high/low cut filter
  - Deconvolution
  - Windowing

Missing data

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

Multi-channel – Multivariate

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

Modeling

- Predictive analytics
- Classical finite difference and finite element
- Neural nets
- Monte Carlo
- Simulated annealing
- Cognitive computing