1. *Introduction to Analytics: Preparing and Visualizing Data*

Description: An introduction to the notions that must be mastered prior to, and after, embarking on data analysis, along with a discussion of common challenges and pitfalls. Participants will be introduced to various methods of data visualization, and to some intrinsic limitations of data, as well as some easily avoidable pre- and post-analysis mistakes.

Intended Audience: Individuals with an interest in quantitative methods who wish to grasp the larger and broader "behind-the-scenes" issues that affect analysts' decisions and results (**this version of the course contains very little technical details and is not intended for experienced data analysts**).

Pre-requisites: This is not a workshop on formulas, computation or any specific analytical methods. Mathematical and/or statistical sophistication is not required.

Outline:

1. Data Science Universals
    - 1.1 Ethics in the Data Science Context
    - 1.2 Guiding Principles of Analytics
    - 1.3 Analytics Workflow

2. Data Clean-up and Validation
    - 2.1 Data Quality
        - What is a Sound Dataset?; Common Sources of Error; Simple Methods to Detect Invalid Entries
    - 2.2 Missing Data
        - Types of Missing Observations; Imputation
    - 2.3 Anomalous Observations
        - Special Data Points; Detection

3. Visualization: Seeing Data in a Different Light
    - 3.1 Pre-Analysis Use vs. Post-Analysis Use
    - 3.2 Simple Graphical Methods
        - Line Graphs; Boxplots; Histograms; Bar Charts; Sparklines; Scatter Plots
    - 3.3 Useful Representations of Complex Data
        - Infographics; Heat Maps; Spaghetti Plots; Chernoff Faces; Geographical Maps; Colour and Geometry; Bubble Charts; Word Clouds; Interactive/Dynamic Graphics
    - 3.4 Design Suggestions

4. Odds and Ends
    - 4.1 Biases, Fallacies, and Pitfalls
    - 4.2 What's the Big Deal About Big Data
    - 4.3 What We Didn't Talk About
        - Data Collection; Data Reduction; Analytical Methods and Model Selection; Validation; Reporting; etc.

2. *Mining for Information Gold: Data Science Concepts and Techniques*

Description: An introduction to the fundamental data science concepts involved in data mining, with a detailed discussion of three common analytical concepts -- classification, clustering and association rules. The application of these concepts will be illustrated through some simple toy examples, along with a discussion of common challenges and pitfalls.

Goals: By the end of the course, participants will be able to appreciate the functionality of different types of data science analysis techniques and recognize opportunities for their application when presented with real world data.

Intended Audience: Individuals who wish to understand the functionality and capabilities offered by different data science methods, even if they won't be the ones implementing them. This workshop is also designed for individuals who want to increase their competency in data science by increasing their knowledge of data science techniques.

Pre-requisites: This workshop requires very little mathematical or computer programming knowledge. Some experience with quantitative methods is assumed.

Outline:

1. Fundamental Notions
    - 1.1 Introduction
        - Data Science in General (6 W's); the Good, the Bad and the Ugly
    - 1.2 Data Mining vs Real Mining: An Analogy
        - Data Mining Tasks; Asking the Right Questions
    - 1.3 The Structure of Data
        - Objects and Attributes; From Attributes to Datasets
    - 1.4 The Most Basic Technique of All
    - 1.5 Learning - Supervised vs. Unsupervised

2. Association Rules - Unsupervised Learning I
    - 2.1 Useful Applications
        - Basics; Correlation and Causation
    - 2.2 A Simple Algorithm
        - Coverage, Accuracy, Support; Generating Association Rules
    - 2.3 Comments and Notes
    - 2.4 Case Study: Danish Medical Dataset

3. Classification - Supervised Learning I
    - 3.1 Useful Applications
        - Basics; Classification Schemes and Methods
    - 3.2 A Simple Algorithm (Decision Tree)
        - Information Gain; Growing and Pruning
    - 3.3 Comments and Notes
        - Other splitting metrics; random forests; performance and interpretability; rare occurrences
    - 3.4 Case Study: Deepwater Horizon Oil Spill

4. Clustering - Unsupervised Learning II
    - 4.1 Useful Applications
        - Basics; Clustering Schemes and Methods
    - 4.2 A Simple Algorithm (k-mode)
    - 4.3 Comments and Notes
    - 4.4 Case Study: OKCupid

5. Data Mining Issues and Challenges
    - 5.1 Overfitting
        - Training, Testing, Validation Sets
    - 5.2 Unrepresentative Data
    - 5.3 Curse of Dimensionality
    - 5.4 Technical Limitations for Big Data
    - 5.5 The Wrong Tool for the Job
    - 5.6 Non-Transferable Assumptions

3. *Simple Data Discovery: Exploring Data with R*

Description: The goal of this workshop is to give participants some introductory experience extracting patterns and knowledge from real datasets using R, in order to increase their understanding of how data and data science can provide insight into problems. This workshop is meant to be a companion to the *Mining for Information Gold: Data Science Concepts and Techniques* and *Getting Technical: More Data Science Methods* workshops.

Goals: By the end of the course, participants will be able to prepare and analyze some simple datasets using the software R and various data science methods.

Intended Audience: Individuals who have some conceptual understanding of data science, and who would like to gain some hands-on experience with the application of data science concepts (as applied in R), or individuals who are planning a data science project and require a more detailed understanding of what that entails.

Pre-requisites: Participation in the first workshop (*Introduction to Analytics*) is not required, but familiarity with its contents may be useful. Some knowledge of the data science techniques introduced in *Mining for Information Gold: Data Science Concepts and Techniques* and *Getting Technical: More Data Science Methods* is assumed.

IT Requirements: Individual computers on which the latest version of the free statistical software R (https://www.r-project.org/) has been installed must be provided by participants. An R-specific user interface (such as RStudio or Tinn-R) may also prove to be helpful. Internet access is required to install the requisite R packages, some of which may be downloaded as needed from the Comprehensive R Archive Network (https://cran.r-project.org/).

The instructor will provide the participants with a list of R packages to be installed prior to the third workshop should public Internet access not be readily available on the premises (the list will be made available a full week before the workshop is held to allow for timely installation by IT staff or the participants themselves.

Outline:
1. R Basics
2. Data Visualization
3. Association Rules
4. Decision Trees
5. k-Means Clustering
6. Support Vector Machines

4. *Getting Technical: More Data Science Methods*

Description: A continuation of the introduction of data science concepts started in *Mining for Information Gold: Data Science Concepts and Techniques*, with a detailed discussion of seven specific methods/concepts which are commonly used by data scientists – support vector machines, artificial neural networks, logistic regression, naïve Bayes classifiers, random forests, hierarchical clustering and density clustering. The application of these algorithms will be illustrated through some simple examples, along with a discussion of common challenges and pitfalls.

Goals: By the end of the course, participants will be able to further appreciate the functionality of increasingly sophisticated data science methods and recognize opportunities for their application when presented with real world data.

Intended Audience: Individuals who wish to understand the functionality and capabilities offered by different data science methods, even if they won't be the ones implementing them. This workshop is also designed for individuals who want to increase their competency in data science by increasing their knowledge of data science techniques.

Pre-requisites: A certain level of familiarity with data science concepts is assumed: this workshop requires some mathematical and computer programming knowledge. Experience with quantitative methods is assumed.

Outline:
1. Support Vector Machines
2. Artificial Neural Networks
3. Logistic Regression
4. Naïve Bayes Classification
5. Random Forests
6. Hierarchical Clustering
7. Density Clustering
8. Spectral Clustering
9. Affinity Propagation Clustering