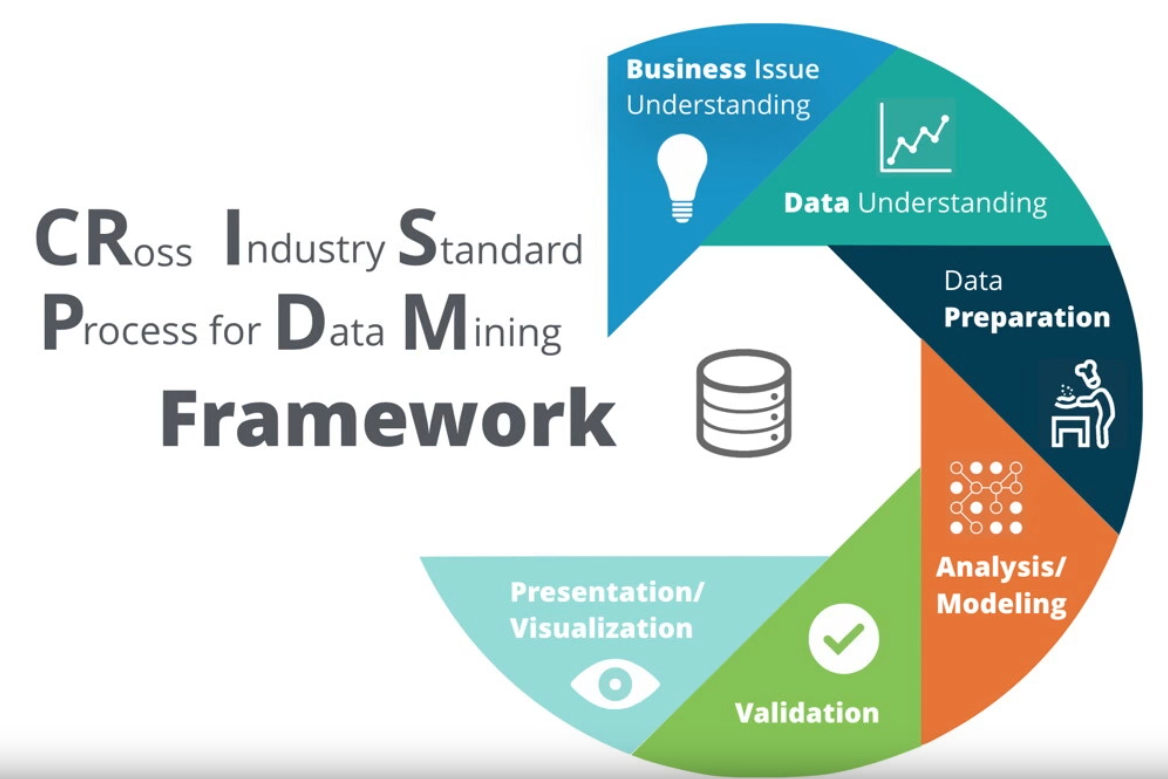
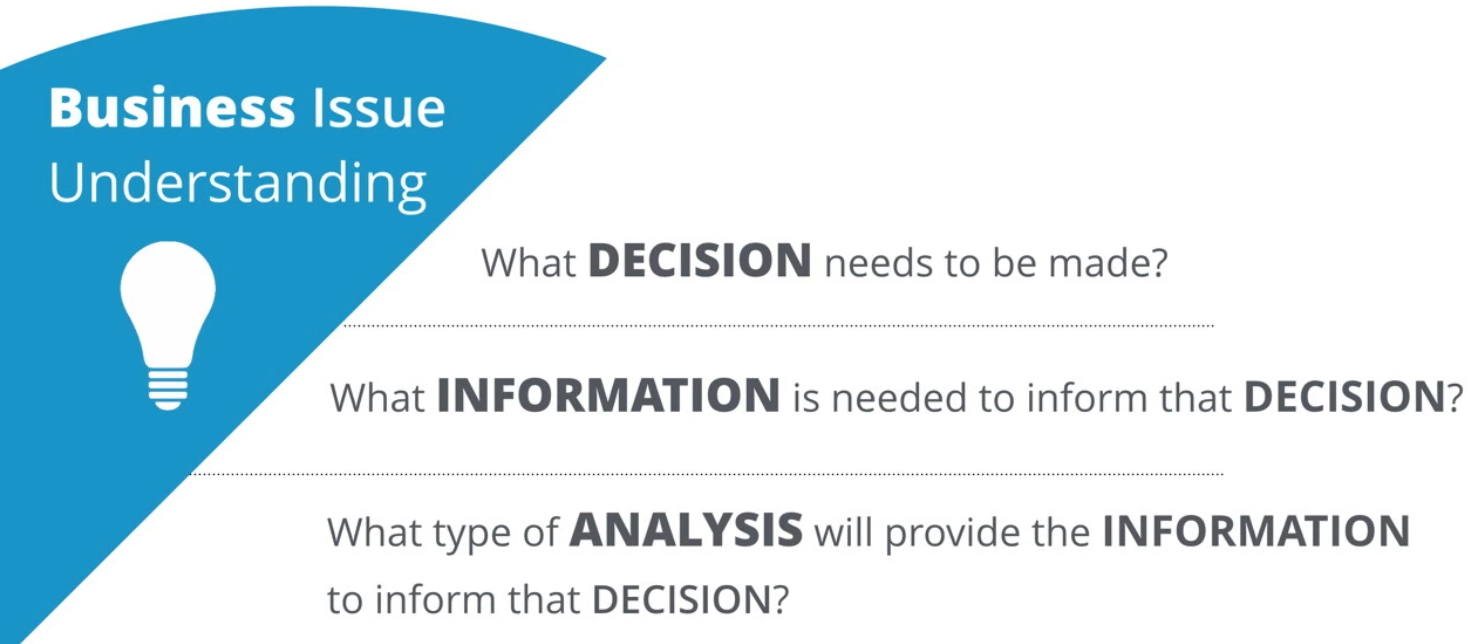
**Problem Solving with Advanced Analytics**

Jocelyn Li

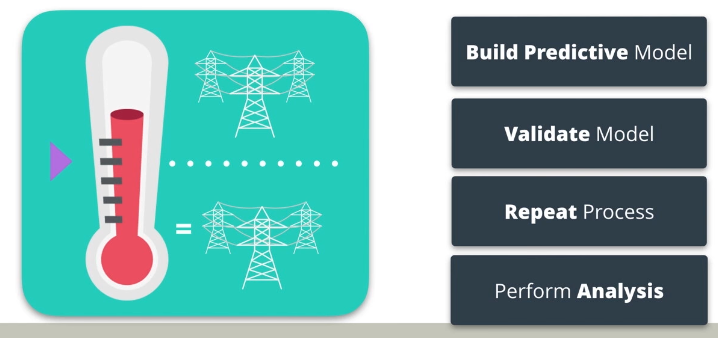
February 2017

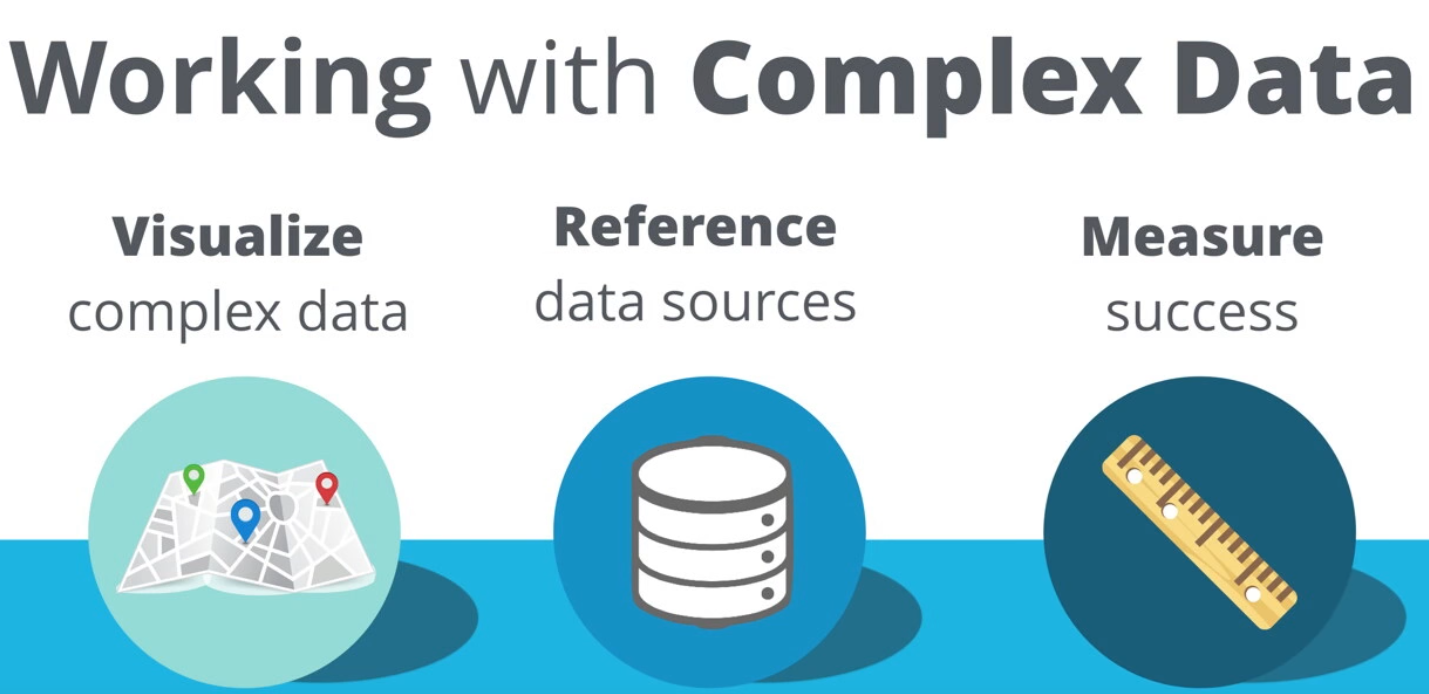
1. **The problem solving framework CRISP-DM**











#### **Data Understanding**

"The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information." – Wikipedia

* What data is needed?
* What data is available?
* What are the important characteristics of the data?

#### **Analysis/Modeling**

"In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed." - Wikipedia

##### Important Steps

* Determine what methodology to use to solve the problem
* Determine the important factors or variables that will help solve the problem
* Build a model to solve the problem
* Run the model and move to the validation phase

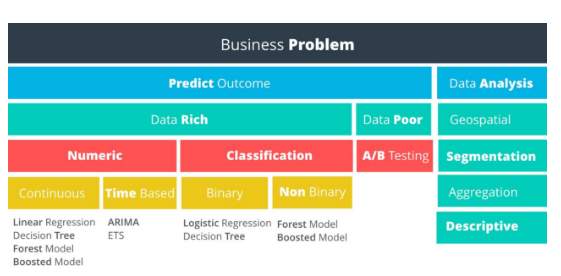
#### **Validation**

"At this stage in the project you have built a model (or models) that appears to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached." - Wikipedia

Important Steps

* Observe the key results on the model
* Ensure the results make sense within the content of the business problem
* Determine whether to proceed to the next step or return to a previous phase
* Repeat as many times as necessary

1. **Methodology Map**



1. **Non-Predictive Analysis**

We’ve broken down non-predictive data analysis into four categories:

* Geospatial (**地理空间**)
* Segmentation
* Aggregation
* Descriptive



**Geospatial Analysis**

This type of analysis uses **location based data to help drive your conclusions**. Examples include identifying customers by a geographic region, calculating the distance store locations or creating a trade area based upon customer locations.

**Segmentation Analysis**

Segmentation is the process of **grouping data together**. Groups can be simple, such as customers who have purchased different items, to more complex segmentation techniques where you identify stores that are similar based upon the demographics of their customers.

**Aggregation Analysis**

This methodology simply means calculating a value across a group or dimension and is commonly used in data analysis. For example, you may want to aggregate sales data for a salesperson by month - adding all of the sales closed for each month. Then, you may want to aggregate across dimensions, such as sales by month per sales territory. Aggregation is often done in reporting to be able to "slice and dice" information to help managers make decisions and view performance.

**Descriptive Analysis**

Descriptive statistics provides simple summaries of a data sample. Examples could be calculating average GPA for applicants to a school, or calculating the batting average of a professional baseball player. In our electricity supply scenario, we could use descriptive statistics to calculate the average temperature per hour, per day, or per date. Some of the commonly used descriptive statistics are **Mean, Median, Mode, Standard Deviation, and Interquartile range.**

1. **Predictive Analysis**

**Data Rich vs. Data Poor**

Do you have data on what you are trying to predict? If so, you can proceed down the data rich path, otherwise, the data poor path is your only option. See the following example that demonstrates a data poor scenario.

**Data Poor Business Problems**

**A/B Tests**

If there is not sufficient usable data to solve the problem, then we need to set up an experiment to help us get the data we need. An experiment in a business context is usually referred to as an A/B Test.

**Numeric vs. Non-Numeric Predictive Analysis**

Assuming we have enough data to proceed with the analysis, our next decision is to look at the outcome we’re trying to predict and determine if it’s a numeric outcome or a non-numeric outcome.

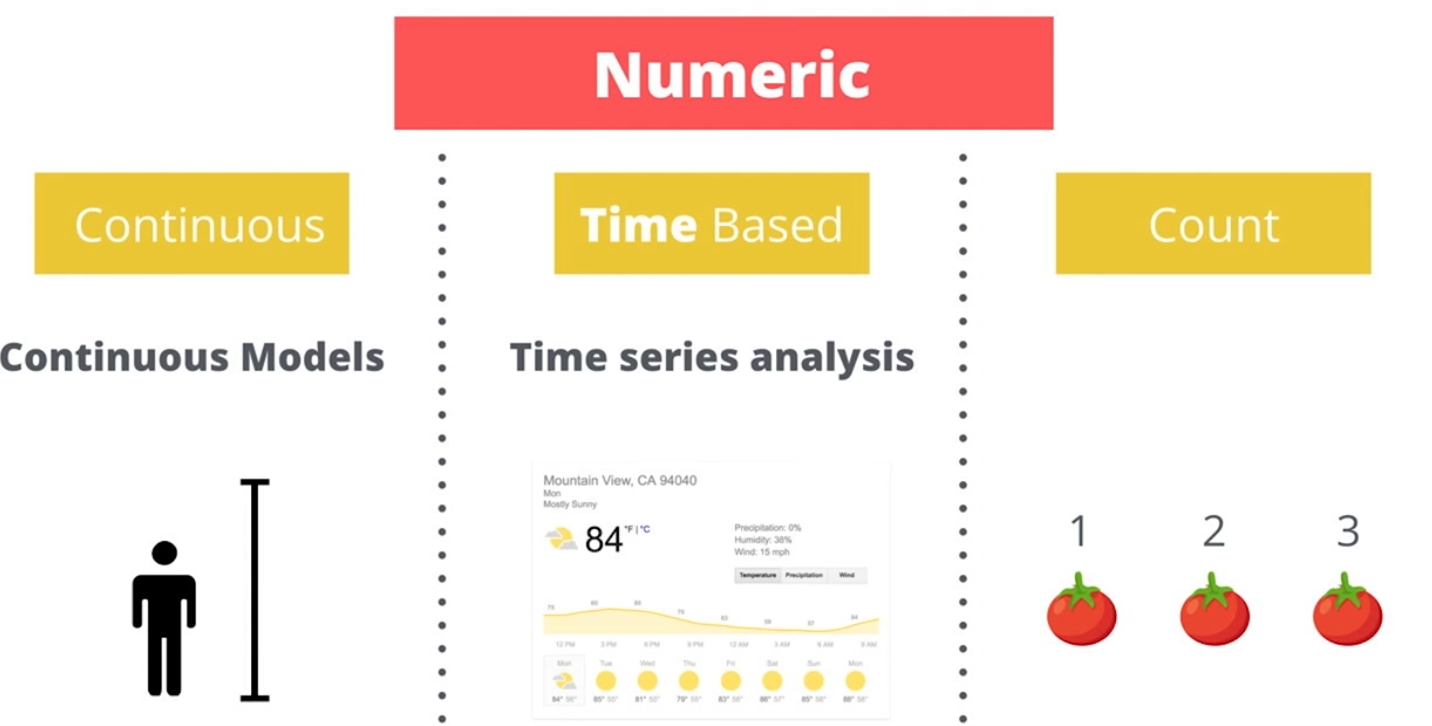
**Regression Models**

Numeric outcomes are those **where the outcome is simply a number**. Predicting the demand for electricity or the hourly temperature are both numeric outcomes. Models predicting numeric data are called regression models.

**Classification Models**

Non-numeric outcomes are those **where we’re trying to predict the category into which a case** (e.g. customer) falls, such as whether a customer will pay on-time, pay late, or default on a payment. Another example is the whether an electronic device will fail before 1000 hours or not. Models predicting non-numeric data are called classification models.

1. **Introduction to Numeric Models**



#### **Target Variables**

Target variables represent the outcome we are trying to predict. In order to select the right predictive model, we first determine whether the target variable is numeric or non-numeric. The type of numeric or non-numeric target variables will then help us select which model is appropriate. Let’s start with numeric variables.

#### **Types of Numeric Variables**

The three most common types of numeric variables are continuous, time-based, and count.

#### **Continuous**

A continuous variable is one that can take on all values in a range. For instance your height can be measured down to many decimal places. We do not grow in even inch intervals. Discrete data is counted, Continuous data is measured

#### **Time-Based**

A time-based numeric variable is one where you are trying to predict what will happen over time. This is often related to forecasting.

#### **Count**

Count variables are numbers that are [**discrete**](https://www.mathsisfun.com/data/data-discrete-continuous.html), positive integers. They’re called count numbers because they’re used to analyze variables that you can count. As modeling these type of variables is not common in business, we won’t be covering this topic in this course.

1. **Introduction to Numeric Models**