What is Analytics?

Analytics is the science of using DATA to build MODELS that add VALUE to DECISIONS made by individuals, companies and institutions (source: The Analytics Edge, Bertsimas, O'Hair and Pulleyblank).

The use of analytics is providing a competitive edge to individuals, companies and institutions. In this course, we will see several such examples where the use of analytics has had impact. Through this, we hope to motivate students to use analytics in a meaningful manner.

Analytics is often viewed as critical to the success of a company and this is increasingly becoming the norm. Companies such as IBM traditionally a personal computer company have invested over 20 billion dollars since 2005 in growing its analytics businesses.

Analytics is often classified as (source: Competing on Analytics, Davenport and Harris)

1. Descriptive analytics
2. Predictive analytics
3. Prescriptive analytics

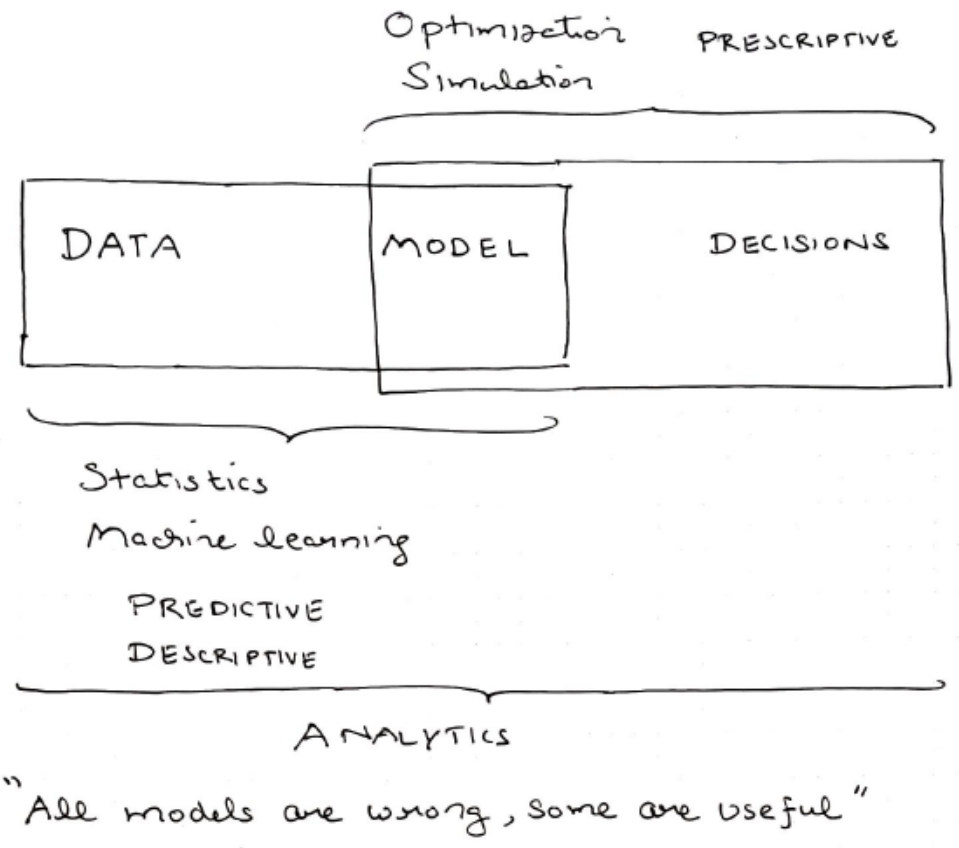

Figure 0.1: The Analytics View

Real world problems are often complex and ill-defined. Furthermore, data is typically the only objective reality, which itself might be incomplete or of poor quality. While data is getting easier to access, it is often unstructured and comes in various forms. Furthermore, in practice, problems do not directly arise as regression, classification or optimization problems. The role of models in these situations is to facilitate the solution of real world problems where data plays a key role (not a secondary role). Another word that gets used a lot in conjunction with the use of analytics is Big Data. While more data is useful, we will see that it is possible to do insightful analytics with small datasets too. We will work with datasets of various sizes in this course.

Predictive analytics these days are used to makes predictions at a personalized (individual) level. Based on this, we can aim to answer questions such as - what treatment plan should we prescribe to a patient, at what price should we offer a ticket to a customer, what type of marketing strategy is effective at influencing the opinions of the individual? While such predictive analytics can provide powerful insights, there are also privacy issues that needs to be dealt with and something that organizations must be sensitive to. While in this class we will not deal with datasets that have issues of privacy, these issues remain important, even more when it is used at scale.

An Example: Watson and Jeopardy

Jeopardy is a popular television quiz show that was launched in 1964. On February 14, 2011, Watson, a computer system developed by IBM participated in this show against the best human players. The challenge is that the quiz needs Watson to have natural language processing capabilities that deals with understanding the human language. The Jeopardy quiz is particularly designed so that clues are meant to be hard to directly obtain meaning from. A unique feature of this quiz is that each question is presented in the form of an answer and the contestant has to answer in the form of a question.

Example: This number, one of the first 20, uses only one vowel (4 times!).

Example: Sakura cheese from Hokkaido is a soft cheese flavored with leaves from this fruit tree.

Furthermore, the answer must be given in around 5 seconds if the contestant wants to successfully buzz in. In evaluating human players as well as the Watson system, there are two fundamental measures that was identified by the IBM team to be as important - accuracy: fraction of clues for which the participant was successfully confident of buzzing in, and precision: fraction of times the answer is correct. Based on past data analysis of Jeopardy games, the team estimated that a winner needed to have an accuracy of around 40 to 50% and a precision of 85 to 95%.

Watson generated a set of candidate answers using a wide range of encyclopedias, dictionaries, news articles from identifying relevant articles. For each answer it generated a confidence level to determine the most likely correct ones. By choosing when to buzz using a threshold for the confidence level, Watson could decide to play more aggressively or defensively based on the game situation. All this needed huge computational memory and processing power where the key to success was to exploit the strength of the computer to play a game against the top humans. The ability of Watson to work with unstructured information and answer the questions opens up many new possibilities.

One important idea in making good predictions is the use of ensemble methods. Instead of using a single model to make predictions, one can use multiple models to make predictions and then use a majority vote or an average to make the "best" prediction. One has to often then tradeoff between accuracy and model interpretability - a topic that we will touch upon several times. Watson made extensive use of ensemble methods in estimating the confidence level.

An Introduction to R

Source: www.r-project.org. You can use R Studio if you would like to use an integrated development environment (IDE) with R notebooks.

Why R?

1. R is free and open.
2. R provides an integrated suite of software facilities for data manipulation, calculation and graphical display.
3. R provides an environment within which many statistical techniques have been implemented and these functionalities can be extended by adding new packages as needed. It is also possible to develop packages with new statistical methods for others to use.
4. R has extensive online support and discussion forum.

Origins of R

R first appeared in 1993 and is an implementation of the S language developed at Bell Laboratories (formerly AT&T, now Lucent) by John Chambers and colleagues. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland and is currently developed by the R Development Core Team.

One sample t-test

In the one sample test, given data $x_1, x_2, \ldots, x_n$, we test the following hypothesis:

$$H_0 : \mu = \mu_0$$

versus

$$H_1 : \mu \neq \mu_0$$

where μ is the population mean and μ_0 is a fixed number. To test the hypothesis, we use the t-statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where $\bar{x} = \sum_i x_i/n$ is the sample mean and $s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)}$ is the sample standard deviation that is used as a surrogate for the population standard deviation. We can use a t-distribution with $n-1$ degrees of freedom to estimate the p-value which is the probability of observing a value as extreme as the test result under the null hypothesis. A small value for the p-value provides very little support for H_0 and more support for H_1 (we can reject the null hypothesis in this case).