

## What is Statistics?

*I like to think of statistics as the science of learning from data...It presents exciting opportunities for those who work as professional statisticians. Statistics is essential for the proper running of government, central to decision making in industry, and a core component of modern educational curricula at all levels. Jon Kettenring, ASA President, 1997,*

*(<http://www.amstat.org/careers/whatisstatistics.html>)*

*The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. American Heritage Dictionary*

*Learning from data ... turning data into information.*  
Glymour, Madigan, Pregibon, Smyth, 1997

*Statisticians are like painters: they tend to fall in love  
with their models.* Deveaux?

*Statistical thinking will one day be as necessary for effi-  
cient citizenship as the ability to read and write.* H. G.  
Wells, <http://www.statoo.com/en/statistics/>

## What is data mining?

*Data Mining is the process of extracting knowledge hidden from large volumes of raw data.* <http://www.megaputer.com>

*We are drowning in information but starved for knowledge.* John Naisbitt, <http://www.statoo.com/en/datamining/>

*Data mining is the process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from data(bases) that is used to make crucial business decisions. Data mining is not a product that can be bought. Data mining is a discipline and process that must be mastered - a whole problem solving cycle.*  
<http://www.statoo.com/en/datamining/>

## What is exploratory data analysis (EDA)?

*Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to:*

- *maximize insight into a data set;*
- *uncover underlying structure;*
- *extract important variables;*
- *detect outliers and anomalies;*
- *test underlying assumptions;*
- *develop parsimonious models; and*
- *determine optimal factor settings.*

<http://www.itl.nist.gov/div898/handbook/eda/section1/eda11.htm>

## **What is algorithmic statistics?**

Nothing firmly defined but computationally intensive methods such as classification and regression trees (CART), multiplicative adaptive regression splines (MARS), projections pursuit (PP), alternating conditional expectations (ACE), ... have come from a small group of statisticians working in what has been described as algorithmic statistics. The methods originate by tweaking computations and iterations rather than from assumptions about probabilistic distributions.

## How do these definitions relate to this course?

This course bridges the areas of exploratory data analysis, statistical algorithms and data mining. A description that emerged from a previous class is:

*Like an explorer crossing unknown lands, we want first to simply describe what we see. Begin by examining each variable by itself, then move to study relationships between variables. Begin with a graph or graphs then add numerical summaries. Look for anomalies, deviations from the norm, don't just summarize, ...*

## How does the textbook relate to the course?

*Many of these tools have common underpinnings but are often expressed with different terminology. This book describes the important ideas in these areas in a common conceptual framework. While the approach is statistical, the emphasis is on concepts rather than mathematics.*

From the book description,

<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>

(The book's coverage is broad, from supervised learning - prediction - to unsupervised learning.)