These are the links in this lecture:

* + The [difference between descriptive and inferential statistics](https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php).
  + [SUPERVISED LEARNING](http://en.wikipedia.org/wiki/Supervised_learning)
    - [kNN (k Nearest Neighbors)](http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm)
    - [Naive Bayes](http://en.wikipedia.org/wiki/Naive_Bayes_classifier)
    - [Logistic Regression](http://en.wikipedia.org/wiki/Logistic_regression)
    - [Support Vector Machines](http://en.wikipedia.org/wiki/Support_vector_machine)
    - [Random Forests](http://en.wikipedia.org/wiki/Random_forest)
  + [UNSUPERVISED LEARNING](http://en.wikipedia.org/wiki/Unsupervised_learning)
    - [Clustering](http://en.wikipedia.org/wiki/Cluster_analysis)
    - [Factor Analysis](http://en.wikipedia.org/wiki/Factor_analysis)
    - [Latent Dirichlet Allocation](http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)
  + US National Institute of Standards and Technology primer on [Exploratory Data Analysis](http://www.itl.nist.gov/div898/handbook/eda/eda_d.htm).
  + The [five-number summary](https://en.wikipedia.org/?title=Five-number_summary)
  + [Descriptive statistics](https://en.wikipedia.org/wiki/Descriptive_statistics)
  + [Percentiles](https://en.wikipedia.org/wiki/Percentile)
  + [Sample minimum](https://en.wikipedia.org/wiki/Sample_minimum)
  + [Lower quartile](https://en.wikipedia.org/wiki/Quartile)
  + [Median](https://en.wikipedia.org/wiki/Median)
  + [Upper quartile](https://en.wikipedia.org/wiki/Quartile)
  + [Sample maximum](https://en.wikipedia.org/wiki/Sample_maximum)
  + [Box plot](https://en.wikipedia.org/wiki/Box_plot)
  + [SparkR](https://spark.apache.org/docs/latest/sparkr.html)
  + [Big Data XSeries](https://www.edx.org/xseries)
  + Spark's [mllib library](https://spark.apache.org/docs/latest/mllib-guide.html" \t "_blank)

Here is a good description of the [difference between descriptive and inferential statistics](https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php).

[SUPERVISED LEARNING](http://en.wikipedia.org/wiki/Supervised_learning)

* + [kNN (k Nearest Neighbors)](http://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm)
  + [Naive Bayes](http://en.wikipedia.org/wiki/Naive_Bayes_classifier)
  + [Logistic Regression](http://en.wikipedia.org/wiki/Logistic_regression)
  + [Support Vector Machines](http://en.wikipedia.org/wiki/Support_vector_machine)
  + [Random Forests](http://en.wikipedia.org/wiki/Random_forest)

[UNSUPERVISED LEARNING](http://en.wikipedia.org/wiki/Unsupervised_learning)

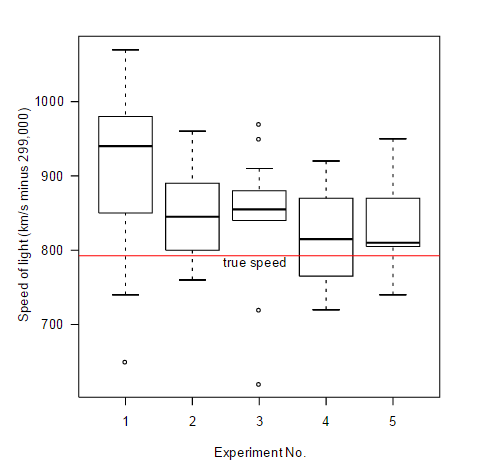
* + [Clustering](http://en.wikipedia.org/wiki/Cluster_analysis)
  + [Factor Analysis](http://en.wikipedia.org/wiki/Factor_analysis)
  + [Latent Dirichlet Allocation](http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation)

The US National Institute of Standards and Technology has an excellent primer on [Exploratory Data Analysis](http://www.itl.nist.gov/div898/handbook/eda/eda_d.htm).

The [five-number summary](https://en.wikipedia.org/?title=Five-number_summary) is a [descriptive statistic](https://en.wikipedia.org/wiki/Descriptive_statistics) that provides information about a set of observations. It consists of the five most important sample [percentiles](https://en.wikipedia.org/wiki/Percentile):

* 1. The [sample minimum](https://en.wikipedia.org/wiki/Sample_minimum) (smallest observation)
  2. The [lower quartile](https://en.wikipedia.org/wiki/Quartile) or *first quartile*
  3. The [median](https://en.wikipedia.org/wiki/Median) (middle value)
  4. The [upper quartile](https://en.wikipedia.org/wiki/Quartile) or *third quartile*
  5. The [sample maximum](https://en.wikipedia.org/wiki/Sample_maximum) (largest observation)

You can compare the five-number summaries of multiple observations using a [box plot](https://en.wikipedia.org/wiki/Box_plot):



Spark 1.4 introduced [SparkR](https://spark.apache.org/docs/latest/sparkr.html" \t "_blank) (R on Spark). SparkR provides a distributed data frame implementation that supports operations like selection, filtering, aggregation etc. (similar to R data frames) but on large datasets.

## ERRATA: CS 190.1X START DATE

The correct start date for CS 190.1x is **June 29, 2015**.

CS 100.1x and CS 190.1X are part of a [Big Data XSeries](https://www.edx.org/xseries). If you sign up for verified certificates in both courses, you will also receive an XSeries certificate.

## SPARK MLLIB LIBRARY

Spark's [mllib library](https://spark.apache.org/docs/latest/mllib-guide.html" \t "_blank) is a scalable machine learning library consisting of common learning algorithms and utilities, including classification, regression, clustering, collaborative filtering, dimensionality reduction, as well as underlying optimization primitives.