Statistics is the discipline of analyzing data. As such it intersects heavily with data science, machine learning and, of course, traditional statistical analysis. In this lecture, we orient you to statistics by covering a few key activities that define the field. These are:

1. **Descriptive statistics**
2. **Inference**
3. **Prediction**
4. **Experimental Design**

Descriptive statistics includes exploratory data analysis, unsupervised learning, clustering and basic data summaries. Descriptive statistics have many uses, most notably helping us get familiar with a data set. Descriptive statistics usually are the starting point for any analysis. Often, descriptive statistics help us arrive at hypotheses to be tested later with more formal inference.

Inference is the process of making conclusions about populations from samples. Inference includes most of the activities traditionally associated with statistics such as: estimation, confidence intervals, hypothesis tests and variability. Inference forces us to formally define targets of estimations or hypotheses. It forces us to think about the population that we're trying to generalize to from our sample.

Prediction overlaps quite a bit with inference, but modern prediction tends to have a different mindset. Prediction is the process of trying to guess an outcome given a set of realizations of the outcome and some predictors. Machine learning, regression, deep learning, boosting, random forests and logistic regression are all prediction algorithms. If the target of prediction is binary or categorical, prediction is often called **classification.** In modern prediction, emphasis shifts from building small, parsimonious, interpretable models to focusing on prediction performance, often estimated via **cross validation.** Generalizability is often given not by a sampling model, as in traditional inference, but by challenging the algorithm on novel datasets. Prediction has transformed many fields include e-commerce, marketing and financial forecasting.

Experimental design is the act of controlling your experimental process to optimize the chance of arriving at sound conclusions. The most notable example of experimental design is **randomization.** In randomization a treatment is randomized across experimental units to make treatment groups as comparable as possible. Clinical trials and A/B testing both employ randomization. In **random sampling,** one tries to randomly sample from a population of interest to get better generalizability of the results to the population. Many election polls try to get a random sample.