- Big Data
  - five V's: Volume, Velocity, Variety, Veracity, Value
  - data preprocessing: up to 80% of effort.
  - ETL: Extract-Transform-Load
  - descriptive vs predictive data mining
  - data science skill set
- Types of Data
  - structured data, e.g. .csv files
  - semi-structured data, e.g. .json files
  - "unstructured data," e.g. free text
  - record data: tables, document data, transaction data
  - ordered data: spatial, temporal, sequential
  - graph data
- Types of Data Object Attributes
  - nominal, ordinal, interval and ratio attributes
  - assymmetric attributes
- Summary Statistics
  - mean
  - median
  - mode
  - quartiles
  - standard deviation
  - interquartile range
- Visualization
  - histograms
  - scatter plots
  - box plots
  - heat maps
- Proximity Measures
  - Minkowski distance $r = 1, 2, \infty$: city-block, Euclidean, supremum (max)
  - simple matching coefficient
  - Jaccard coefficient
  - cosine similarity
  - correlation coefficient
- Multidimensional Data Analysis
  - aggregation using, mean, median, etc.
  - hierarchical indexing
  - roll-up and drill down
  - slicing and dicing
- Cleaning Data
  - dealing with missing data
  - correcting errors
  - removing duplicates
  - filtering noise
- Grouping Data
  - split-apply-combine: aggregate, transform, filter
  - pivot tables
  - cross tabulation
  - OLAP (Online Analytical Processing)
- Merging Data
  - inner, outer, left, right joins
  - one-to-one, many-to-one, many-to-many joins
- Binning Data