1 Big data tools

Measurement Statistical programming languages Visualization tools (usually included with the above) Big Data management software Data collection tools

1.1 Measurement

1.2 Statistical programming languages

Three main programming languages are widely used for data manipulation, analysis:R, Python and Julia.

1.3 Data collection, Cleaning and Preprocessing

Web Scraping can be done in almost any programming language out there Handling large data sets with RDD. To use RDDs you need a cluster of computers and software such as Hadoop or Spark.
OpenRefine: Tool for cleaning and transforming data.
scikit-learn, caret: Libraries for data preprocessing in machine learning.

1.4 Version Control

Git and GitHub: Distributed version control system.

1.5 Cloud Computing Platforms

AWS, Azure, Google Cloud: Cloud platforms that provide scalable computing resources.

1.6 Data Manipulation and Analysis

Pandas: Python library for data manipulation and analysis.
NumPy: Fundamental package for scientific computing with Python.
dplyr: R package for data manipulation and transformation.
SQL : To transform data into a more usable form for statistical software to use. With SQL, one can easily subset, merge, and perform other common data transformations.

1.7 Data Visualization

Allows humans to see in multiple dimensions what the data looks like, spot outliers, and otherwise perform "sanity checks" on the data.

Matplotlib: Python 2D plotting library.
ggplot2: package included in the R tidyverse for creating complex and customizable plots.

1.8 Modeling

The main objectives of statistical modeling are as follows:
Use the data to test theories.
Use the data to predict behavior.
Use the data to explain behavior.