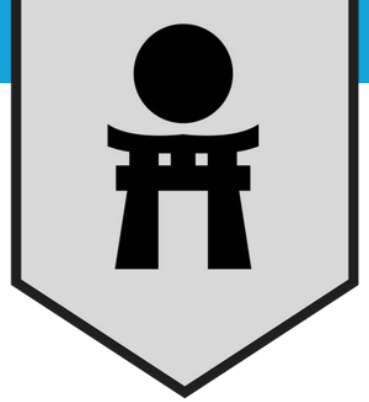


Big Data Engineering with Distributed Systems

Hadoop & Hive, Mahout, and Spark

datasciencedojo
data science for everyone

Agenda

- Introduction:
  - Data engineering for data scientists
  - The "5 Vs" of Big Data
- A key problem – machine learning at scale
- Distributed computing with Apache Hadoop & Hive
- Hadoop in the Azure cloud
- Machine learning at scale with Apache Mahout
- Distributed computing v2.0 – Apache Spark

Data Engineering for Data Scientists

Driving a car

vs

Servicing a car

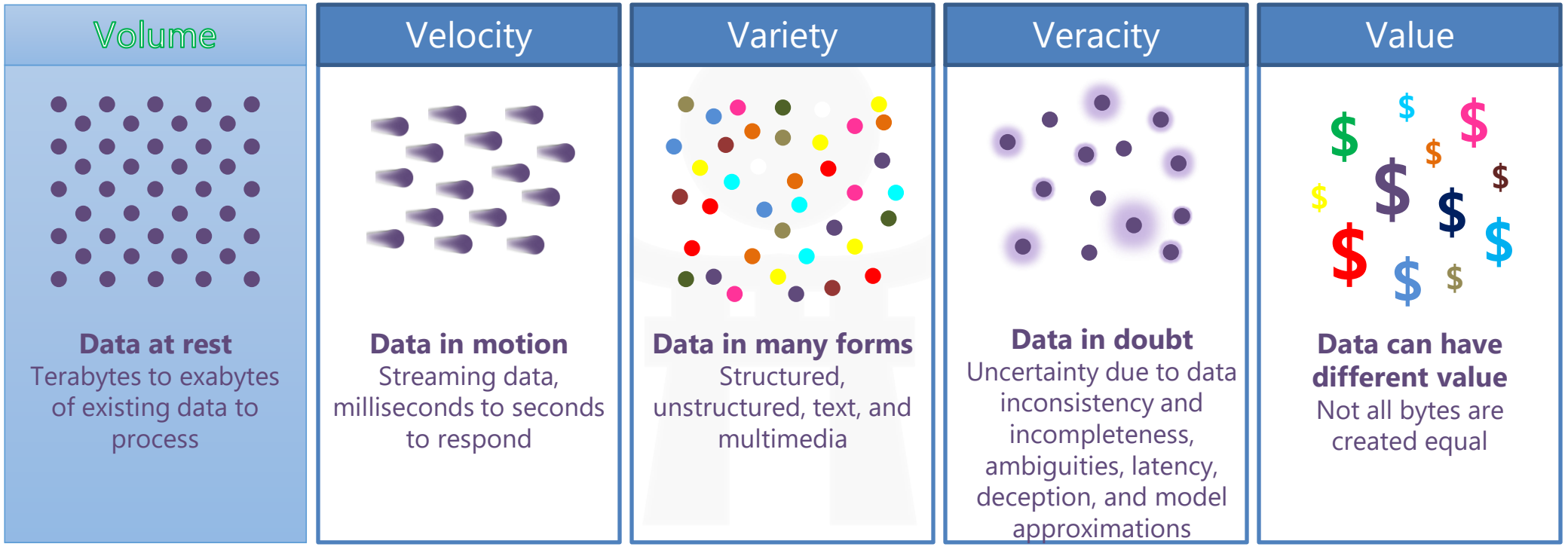
Goals:

- Teach you about data engineering topics/concepts

Non goals:

- Managing or administering a Hadoop cluster

5 Vs of Big Data



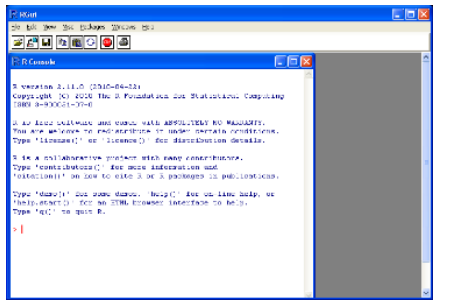
- **Goal**: As data scientists we want cost-effective access to the raw materials for our data products!

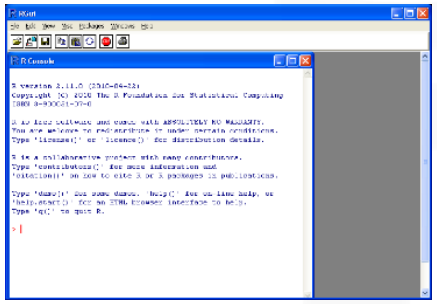
Machine Learning at Scale

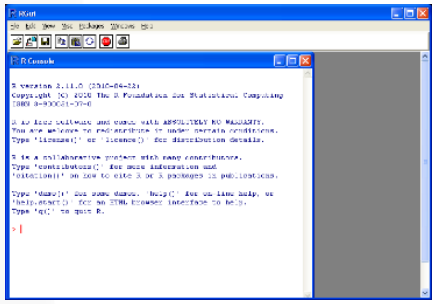
OSS R Limits

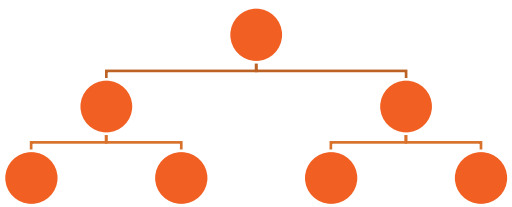
Quad Core Laptop

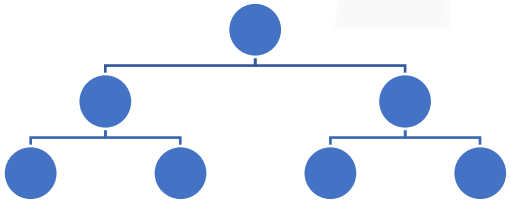
- Single core
- Single threaded

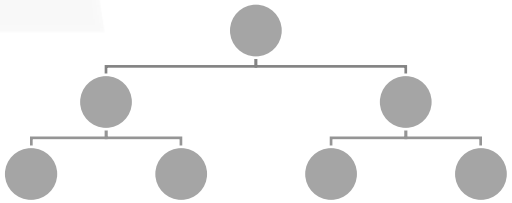












Model A

Model B

Model C