BIG DATA
An Overview

What counts as big data?

- Too big for RAM?
- Too big for a HD?
- Too much for a single CPU?

What counts as big data?

Any amount of data that breaks down our typical processes is considered big data

What counts as big data?

Any amount of data that breaks down our typical processes is considered big data

We have big data once we need big data tools

What counts as big data?

Rule of thumb: if you can fit the data source in a high-end computer's RAM, it's not big.

Why does this need a special name… it's just data?

Four Major Issues - Volume

- Vertical scaling is massively expensive
- Horizontal scaling is massively complicated
- Stability and consistency all become bigger issues the more you scale

Four Major Issues - Velocity

- Big data is often a moving target, with new data being constantly added
- Incoming data still needs to be cleaned, stored, and utilized

Four Major Issues - Velocity

- Most data science processes are at least O(n)
- If you parallelize your work, how do you make sure no data is left behind?
- How do you parallelize on a dataset that’s 60TB?

Four Major Issues - Variety

- In a simple world, all our data would be structured (table/dataframe)
- More likely that we'll need a system that can handle all kinds of data:
  - Structured
  - Text
  - Images
  - Sound, Video, etc.

Four Major Issues - Veracity

- If you have 1000 samples, having a 4 standard deviation process occur is extremely rare
- If you have 1,000,000,000 samples, a 4 standard deviation process will occur nearly 1M times
- When you record more data, there are more chances for your data to be weird.

MORAL OF THE STORY:
WE NEED TOOLS TO HANDLE BIG DATA

Three Approaches

- MapReduce and Hadoop
- Out-of-Core processing using Dask
- Spark

Rules of Thumb

- If it fits in RAM, just use Pandas/numpy

Rules of Thumb

- If it fits in RAM, just use Pandas/numpy
- If it doesn’t fit in RAM, but can still be processed on your computer (20-50ish GB), use Dask locally

Rules of Thumb

- If it fits in RAM, just use Pandas/numpy
- If it doesn't fit in RAM, but can still be processed on your computer (20-50ish GB), use Dask locally
- If it's bigger than that, use Spark or Dask as a cluster. See here for a discussion: http://docs.dask.org/en/latest/spark.html

QUESTIONS?