

# Review of Session One

...

# Ideas

What we have understood so far

- What is Big Data and where does Spark fit
- Common data structures and programming patterns in Statistical Programming
- Spark RDDs and the problems they solve

---

# Activities

We should now be able to do all of these

- Set up a working Spark installation with findspark and Jupyter Notebook
  - Start a Standalone Spark server
  - Run basic Spark programs using pySpark to count lines, words and characters in text files on disk
  - Parallelize computation (testing primality) on Python collections in Spark
-