

# OPEN DATA SCIENCE CONFERENCE

Burlingame | November 2nd 2017

Nov 02  
2:00 PM  
Room T2

Modeling big data with R, sparklyr, and Apache Spark

BIG DATARINTERMEDIATE

Dr. John Mount

Consulting Algorithmist/Researcher/Principal at Win-Vector LLC and  
Co-author of Practical Data Science with R



# Our current project

- Data manipulation in Spark.



# Work through markdowns together

- Exercises/03a-Spark-SQL.Rmd
- Exercises/solutions/03b-Spark-SQL.Rmd

# Exercise

- Please complete Exercises/03a-Spark-SQL.Rmd

# Open Discussion Activity

- Let's step through Exercises/solutions/03b-Spark-SQL.Rmd together.
  - Allows us to cover more material than a formal exercise.
- Please interrupt me (but not each other).

# Challenge Project: more functions

- If you have extra time try reading up on `replyr`.
  - <http://www.win-vector.com/blog/2017/10/big-data-transforms/>
  - <https://winvector.github.io/replyr/>



# Next

- Machine learning concept review