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

Part 7: Advanced topics

7/7

#ODSC

@ODSC

Win-Vector LLC

@WinVectorLLC

R Studio

@RStudio

Plan

- Work through some extra topics
  - Using SparkR for R user defined functions
  - Using Spark SQL directly
  - Using the Sparklyr programming extension interface to directly call Java/Scala in Spark.
- Remind you of topics and resources

Work through markdown together

- Exercises/solutions/06-Spark-Extension.Rmd

What we have achieved in this workshop

- We have worked through `dplyr` in detail.

- We applied `dplyr` data manipulation methods in a big-data environment (`Spark` / `SparklyR`).

- We ran supervised machine learning experiments in big-data environments (`SparkML` / `h2o`).

- We learned how to extend and use `Spark` more directly (`Spark SQL`, `SparklyR` extensions interface, and even a bit or `SparkR`).

Win-Vector LLC

Some links

This material

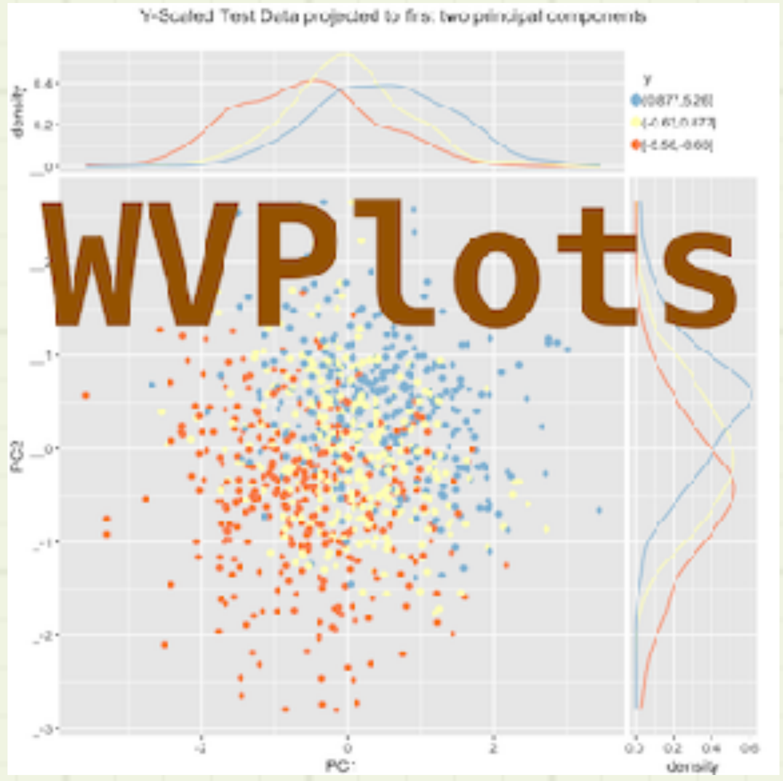
- https://github.com/WinVector/BigDataRStrata2017
  - Detailed local install instructions:
    - README.Rmd
    - Exercises/solutions/RsparklingInstall.Rmd

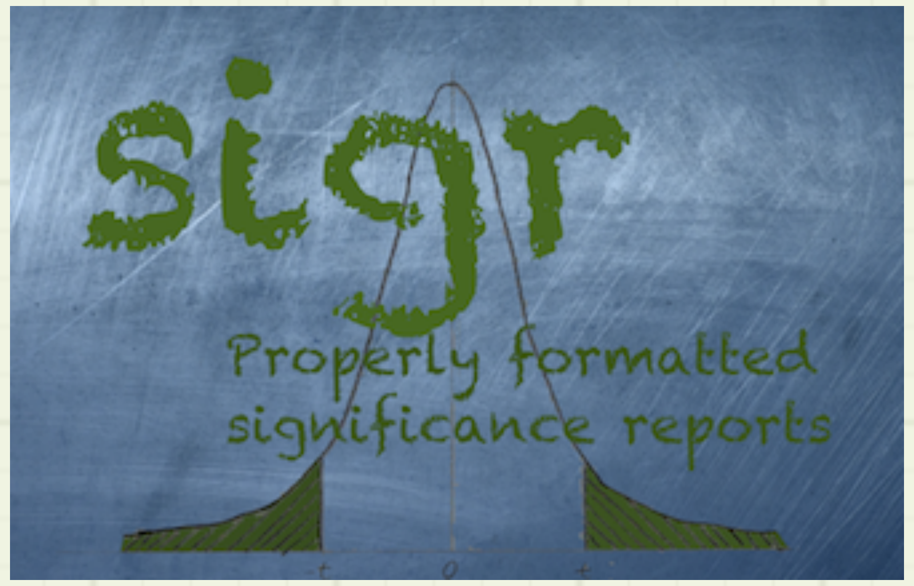
RStudio

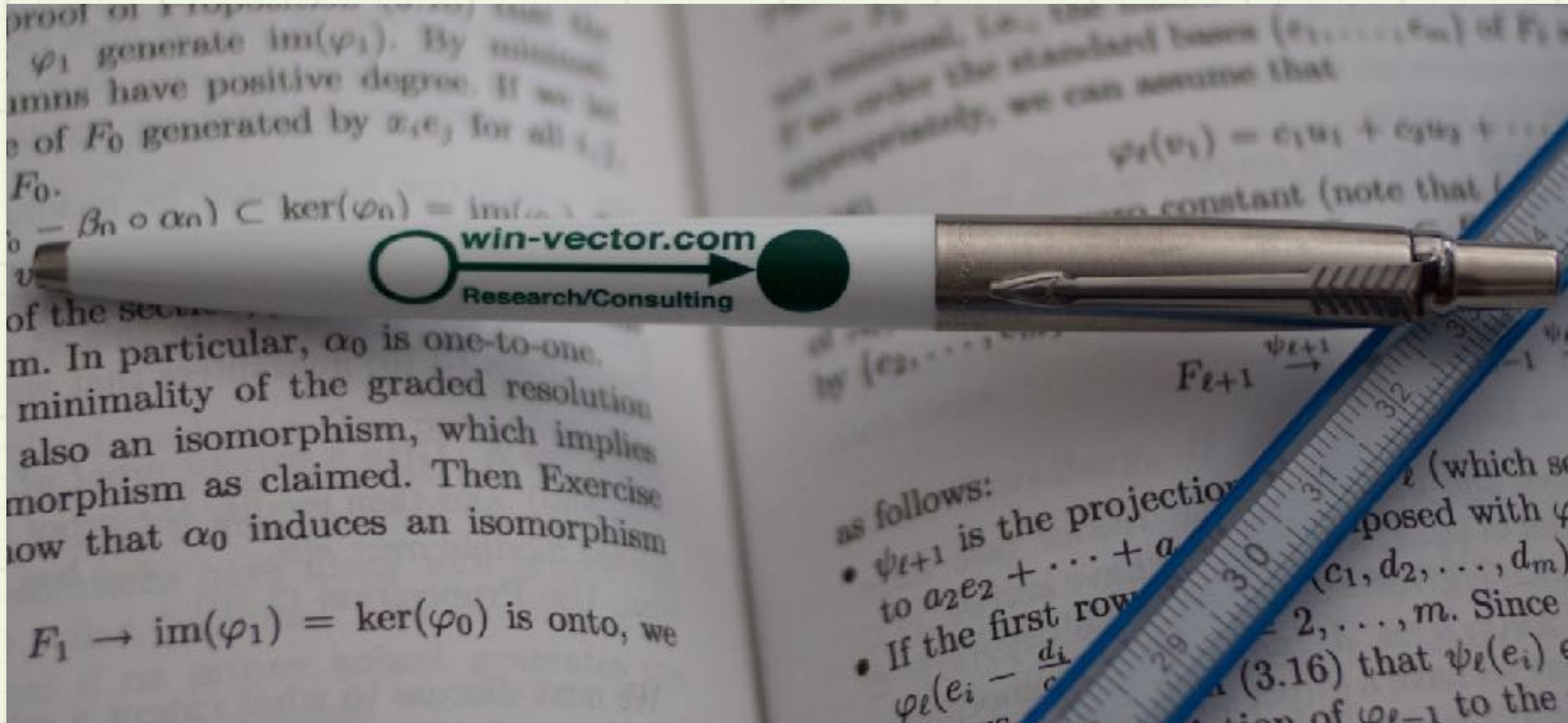
- https://www.rstudio.com
- https://www.rstudio.com/products/rstudio-server-pro/
- https://www.rstudio.com/products/connect/
- https://www.rstudio.com/products/shiny-server-pro/
- https://www.rstudio.com/products/shinyapps/
- https://github.com/rstudio/rstudio

Win-Vector

- http://www.win-vector.com
- http://www.win-vector.com/blog/
- https://github.com/WinVector
- @WinVectorLLC
- contact@win-vector.com
- Please reach out to us for partnerships
    - Myself or Dr. Lin Chase of Big Tech Strategy (who is working with us).







SparklyR

- https://www.rstudio.com/resources/cheatsheets/
- http://spark.rstudio.com
- http://spark.rstudio.com/dplyr.html
- http://spark.rstudio.com/extensions.html

h2o

- http://www.h2o.ai
- https://github.com/h2oai

Demos

1. R markdown notebooks with dplyr (NYCFlights13 - Local mode)
   https://beta.rstudioconnect.com/content/1706/

2. Flexdashboard
   http://colorado.rstudio.com:3838/nathan/flights-dash-spark/
   http://colorado.rstudio.com:3838/nathan/flights-dash-rdata/
   https://beta.rstudioconnect.com/content/1439/

3. Comparison of ML classifiers (Titanic - Local mode)
   https://beta.rstudioconnect.com/content/1518/

4. Manipulate data at scale (NYC Taxi - Cluster mode)
   https://beta.rstudioconnect.com/content/1704/

5. End to end analysis (Flights - Cluster mode)
   https://beta.rstudioconnect.com/content/1446/

Questions? Comments?

Thank you very much!

Please be sure to fill out you O’Reilly workshop evaluations.

Please visit **RStudio** at the Innovator’s Pavilion – booth number P8 during the Expo Hall hours.

Please tell them you attended this workshop!

Books from RStudio authors, t-shirts to win, demonstrations of RStudio Connect and RStudio Server Pro and, of course, stickers and cheatsheets. Get your product and company questions answered by RStudio employees.

Also:

2:40pm–3:20pm Wednesday, March 15, 2017

Sparklyr: An R interface for Apache Spark

Edgar Ruiz (RStudio)

Primary topic: Spark & beyond

Location: LL21 C/D

Or:

Office Hour with John Mount (Win-Vector LLC)

2:40pm–3:20pm Wednesday, March 15, 2017

Room: Table B

Come and ask me questions about data science, machine learning, R, statistics, or whatever you like.

Win-Vector LLC