



PRESENTED BY



strataconf.com

#StrataHadoop

Modeling big data with R, sparklyr, and Apache Spark

1:30pm–5:00pm Tuesday, March 14, 2017

Data science & advanced analytics

Location: LL21 C/D

Level: Intermediate

Secondary topics: R

John Mount (Win-Vector LLC)

Steve Nolen (RStudio)

Edgar Ruiz (RStudio)

url: <https://github.com/WinVector/BigDataRStrata2017>

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 of `SparkR`).

Some links

This material

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

RStudio

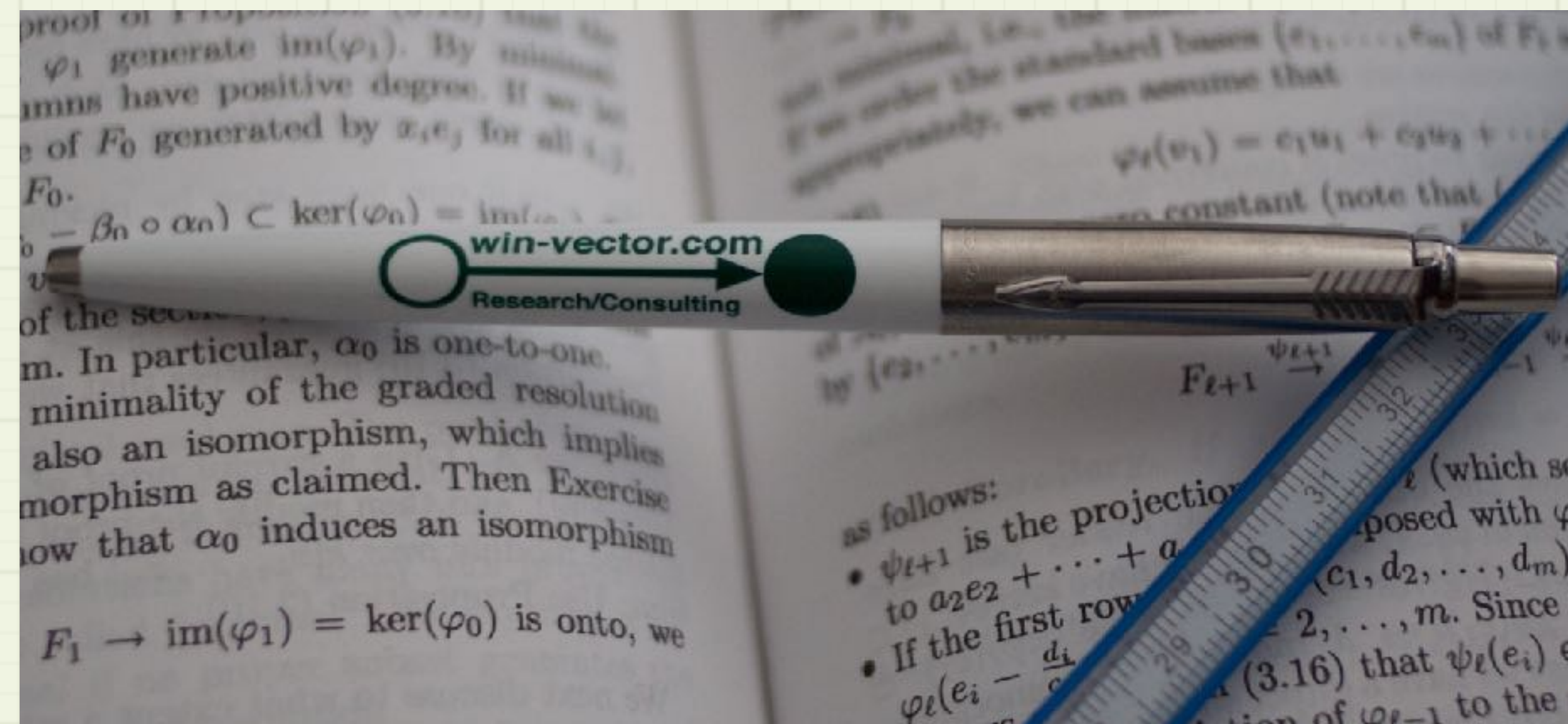
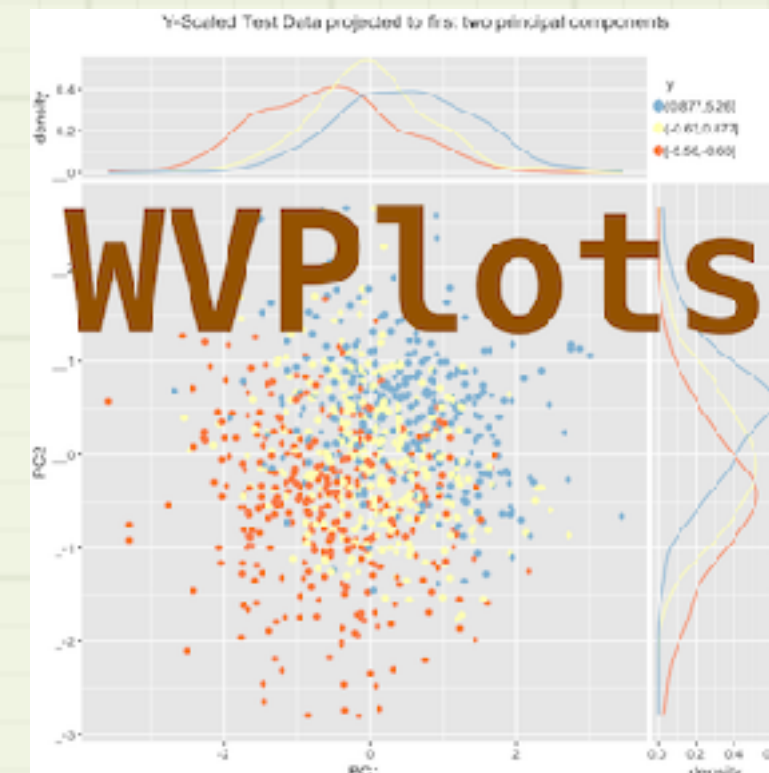
- <https://www.rstudio.com>
- <https://www.rstudio.com/products/rstudio-server-pro/>
- <https://www.rstudio.com/products/connect/>
- <https://www.rstudio.com/products/shiny-server-pro/>
- <https://www.rstudio.com/products/shinyapps/>
- <https://github.com/rstudio/rstudio>



Win-Vector



- <http://www.win-vector.com>
- <http://www.win-vector.com/blog/>
- <https://github.com/WinVector>
- [@WinVectorLLC](https://twitter.com/WinVectorLLC)
- contact@win-vector.com



SparklyR

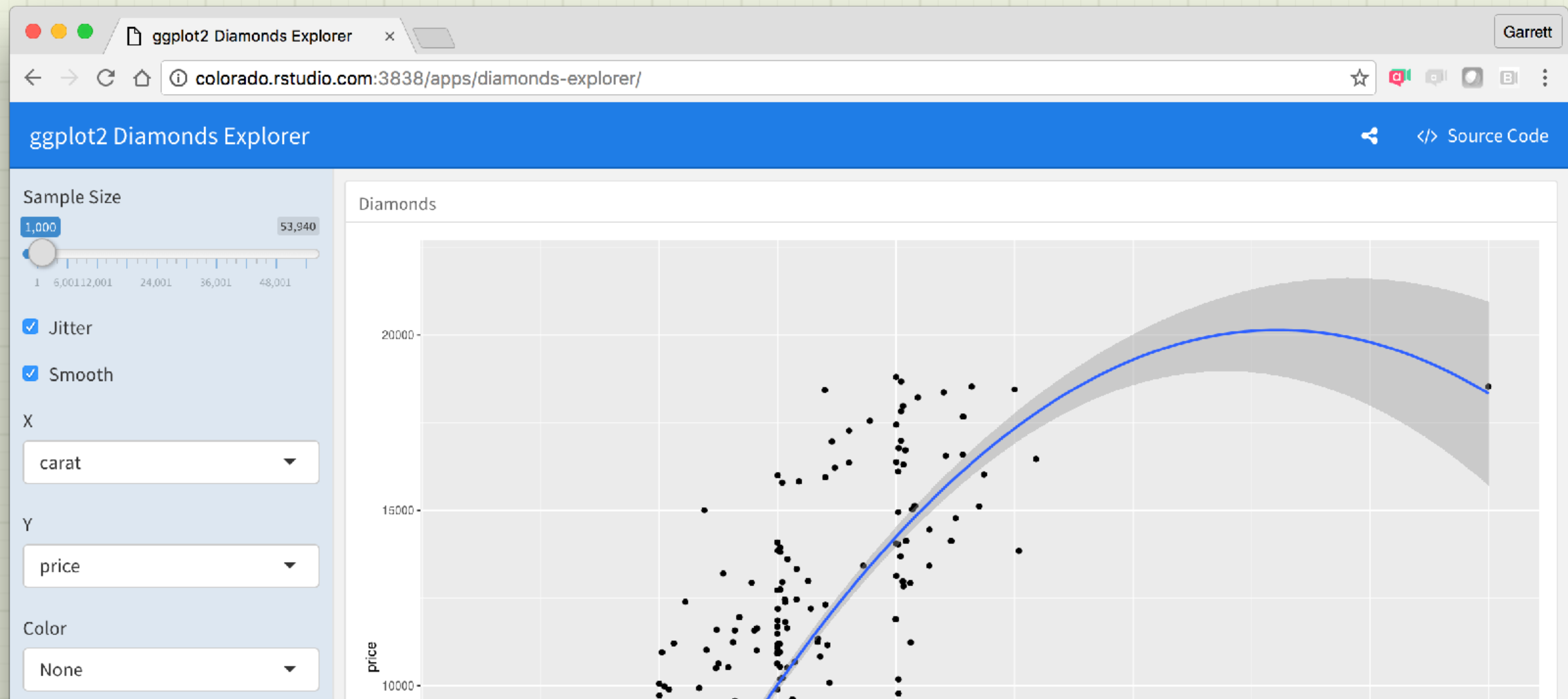
- <https://www.rstudio.com/resources/cheatsheets/>
- <http://spark.rstudio.com>
- <http://spark.rstudio.com/dplyr.html>
- <http://spark.rstudio.com/extensions.html>

h2o

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

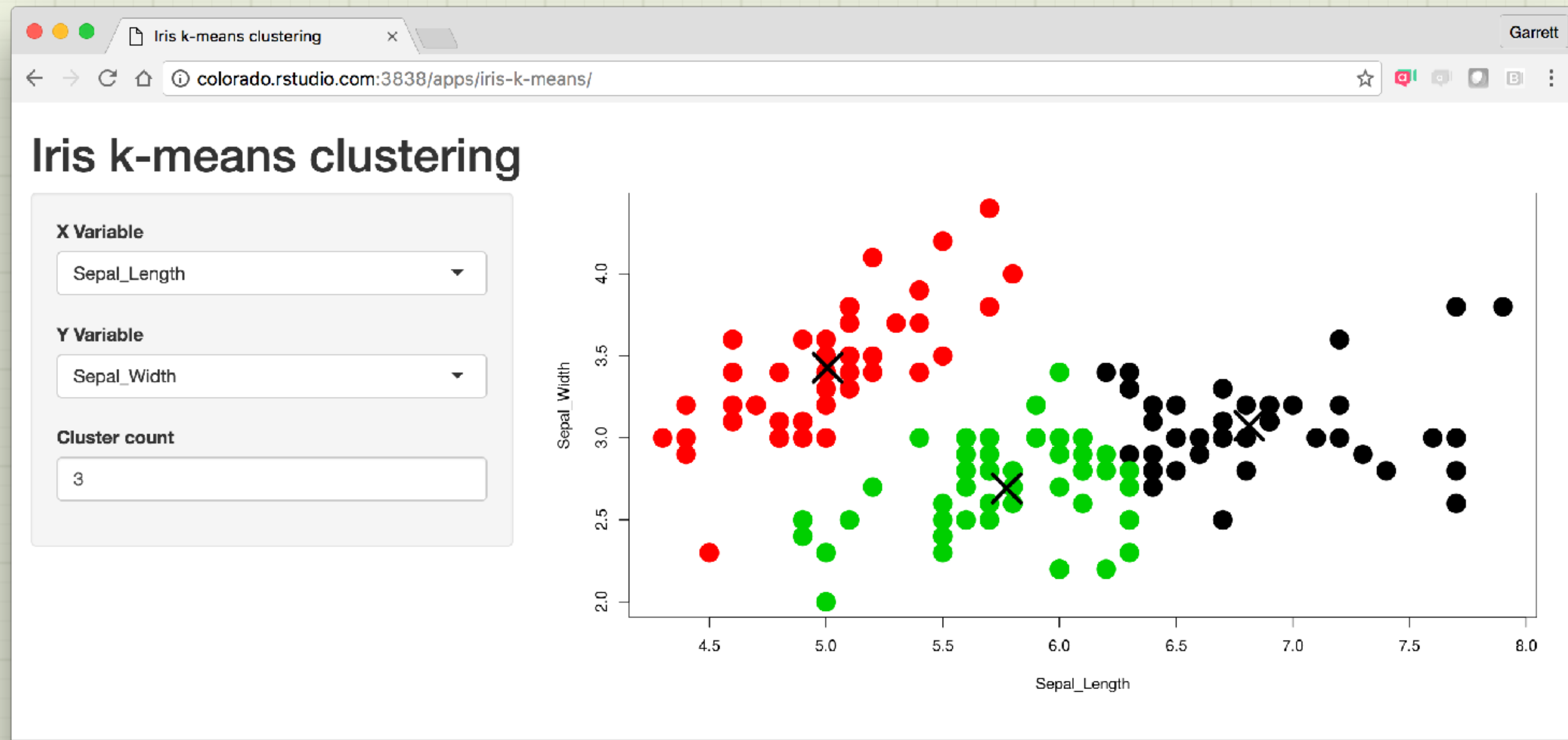
Diamonds Dashboard

sparkdemo.rstudio.com/dashboards/diamonds-explorer/



Iris K Means Clustering

<http://sparkdemo.rstudio.com/apps/iris-k-means/>



Titanic Machine Learning

<https://beta.rstudioconnect.com/content/1518/>

Overview

Load the data

Tidy the data

Spark SQL transforms

Spark ML transforms

Train-validation split

Train the models

Logistic regression

Other ML algorithms

Validation data

Compare results

Model lift

AUC and accuracy

Feature importance

Compare run times

Discuss

Comparison of ML Classifiers Using Sparklyr

Overview

You can use `sparklyr` to fit a wide variety of machine learning algorithms in Apache Spark. This analysis compares the performance of six classification models in Apache Spark on the [Titanic](#) data set.

For the Titanic data, decision trees and random forests performed the best and had comparatively fast run times. See [results](#) for a detailed comparison.

ID	Function	Description	AUC Rank	Run time Rank
1	Random forest	ml_random_forest	1	3
2	Decision tree	ml_decision_tree	2	2
3	Gradient boosted tree	ml_gradient_boosted_trees	3	6
4	Logistic regression	ml_logistic_regression	4	4
5	Multilayer perceptron (neural net)	ml_multilayer_perceptron	5	5
6	Naive Bayes	ml_naive_bayes	6	1

Load the data

Spark Extensions

Write your own Spark extensions in R. Use extensions in Spark much like you use CRAN packages in R.

sparklyr

Home

dplyr

ML

Extensions

Deployment

Reference

Introduction

Core Types

Calling Spark from R

Wrapper Functions

Dependencies

Core Types

Three classes are defined for representing the fundamental types of the R to Java bridge:

Function	Description
<code>spark_connection</code>	Connection between R and the Spark shell process
<code>spark_jobj</code>	Instance of a remote Spark object
<code>spark_dataframe</code>	Instance of a remote Spark DataFrame object

S3 methods are defined for each of these classes so they can be easily converted to or from objects that contain or wrap them. Note that for any given `spark_jobj` it's possible to discover the underlying `spark_connection`.

Calling Spark from R

There are several functions available for calling the methods of Java objects and static methods of Java classes:

Function	Description
<code>invoke</code>	Call a method on an object
<code>invoke_new</code>	Create a new object by invoking a constructor
<code>invoke_static</code>	Call a static method on an object

For example, to create a new instance of the `java.math.BigInteger` class and then call the `longValue()` method on it you would use code like this:

Demos

1. R markdown notebooks with dplyr (NYCFlights I 3 - 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 your 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.

