



# Become a self proclaimed Data Scientist

Rakesh Ranjan



Oct 3, 2015

# What do data scientists do?

- Analyze data produced by social media applications
- Analyze data produced by business applications
- Develop hypothesis, understand data and explore patterns
  - Understand and manipulate data sets
- Uses analytics to optimize business processes
  - Solve tricky mathematical and statistical problems using software packages
- Create visualization to communicate results



# Data Analysis using R




- Language and environment for statistical computing and graphics
- Large, coherent, integrated collection of tools for data analysis
- Well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.
- Graphical facilities for data analysis and display either on-screen or on hardcopy
- <https://www.r-project.org/>



October 3-4 2015  
www.SiliconValley-CodeCamp.com



# Analyze data (social media)

- Access to Twitter API (twitterR) 
  - Plot a cloud of words shared across tweets (wordcloud)
  - Text mining in R (tm)
- Access to Facebook API (Rfacebook) 
  - generate analytics about user's posts, public status updates
- Access to LinkedIn API (Rlinkedin)
- Access to social media data (SocialMediaMineR) 

Code demonstration continues...

Also found at

<https://github.com/CDSLab/Codecamp2015>



# Analyze large data (business)

- dashDB – a fully managed analytic service in the cloud
- ibmDBR – basic and complex algorithm pushed down to database to leverage DB2's engine parallelism
- Integrated with R-Studio server for development in cloud
- Comes with sample statistical models for both In-application and In-database analytics
- <http://www-01.ibm.com/software/data/dashdb/>



# Optimize business process

Suppose a farmer has 75 acres on which to plant two crops: wheat and barley. To produce these crops, it costs the farmer (for seed, fertilizer, etc.) \$120 per acre for the wheat and \$210 per acre for the barley. The farmer has \$15000 available for expenses. But after the harvest, the farmer must store the crops while awaiting favorable market conditions. The farmer has storage space for 4000 bushels. Each acre yields an average of 110 bushels of wheat or 30 bushels of barley. If the net profit per bushel of wheat (after all expenses have been subtracted) is \$1.30 and for barley is \$2.00, how should the farmer plant the 75 acres to maximize profit?

- $P = (110)(1.30)x + (30)(2.00)y = 143x + 60y$
- $120x + 210y \leq 15000$
- $110x + 30y \leq 4000$
- $x + y \leq 75$
- $x \geq 0, y \geq 0$



# Analyze very large data sets (sparkR)

- Is a binding based on spark's new Dataframe API
- Provides access to spark's scale out parallel runtime engine along with spark's other features
- Supports calling directly into sparkSQL feature
- Available with spark 1.4.0 and beyond



# Why sparkR?

- Interactive analysis in R is usually limited as runtime is single threaded and can only process data sets that fits into client's memory
- sparkR provides front-end to spark so R execution can use spark's distributed model and run large scale data analysis from R console or R-studio
- Demo continues – analytics on SFPD crime data
- <https://github.com/CDSLab/Codecamp2015>





# Credits

- <https://cran.r-project.org/web/packages/Rfacebook/index.html>
- <https://github.com/mpiccirilli/Rlinkedin>
- <https://cran.r-project.org/web/packages/twitterR/>
- <https://cran.r-project.org/web/packages/SocialMediaMineR/index.html>
- <https://cran.r-project.org/web/packages/ibmdbR/index.html>
- <http://lpsolve.sourceforge.net/>
- <https://github.com/CDSLlab/Codecamp2015>

