From Zero to Hero

A Hands-On Approach to R Stats

Pierre Lafortune
pierre.a.lafortune@gmail.com

Agenda

- R Introduction
- Common Use cases
- Problem Statement
- Live Coding
- Participant Challenges

R Quick Launch

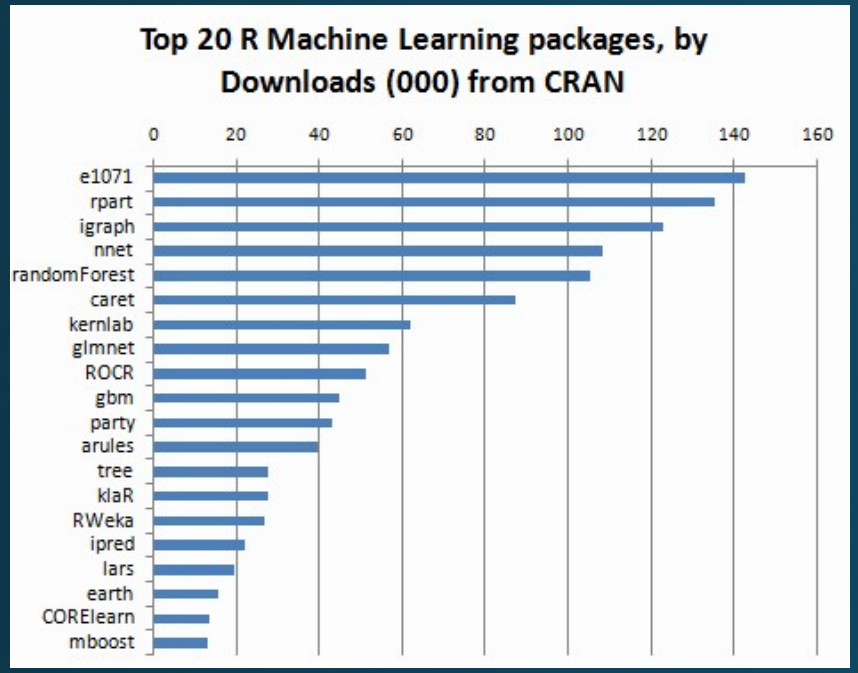
- Statistical software environment
- Built on S programming language
- Named R after originators Robert Gentleman and Ross Ihaka
- Created in 1993
- Open-source under General Public License
- Source-code written in C, Fortran, and R
- Over 6,000 packages added by contributors


UBER
SET PICKUP LOCATION

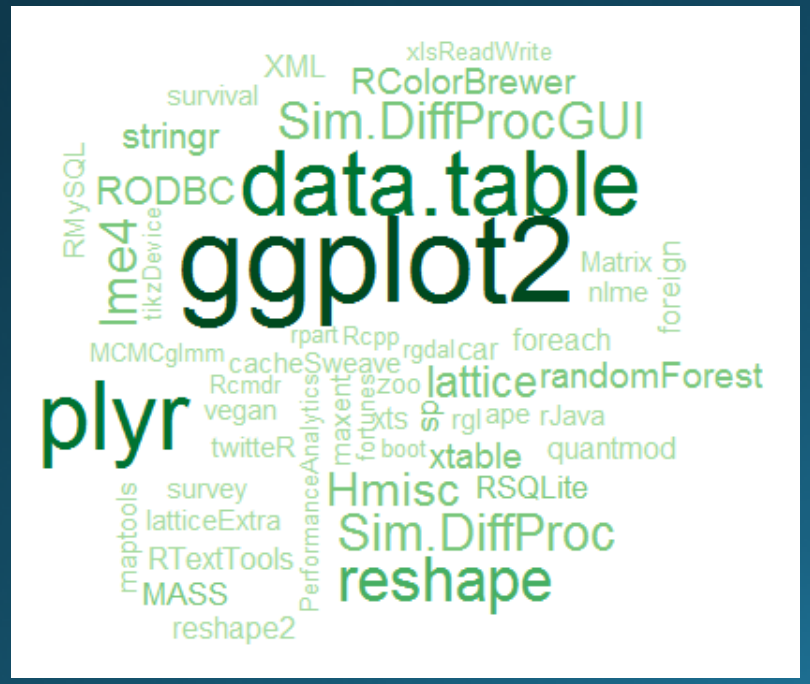

STAY
HANDMADE
ETSY

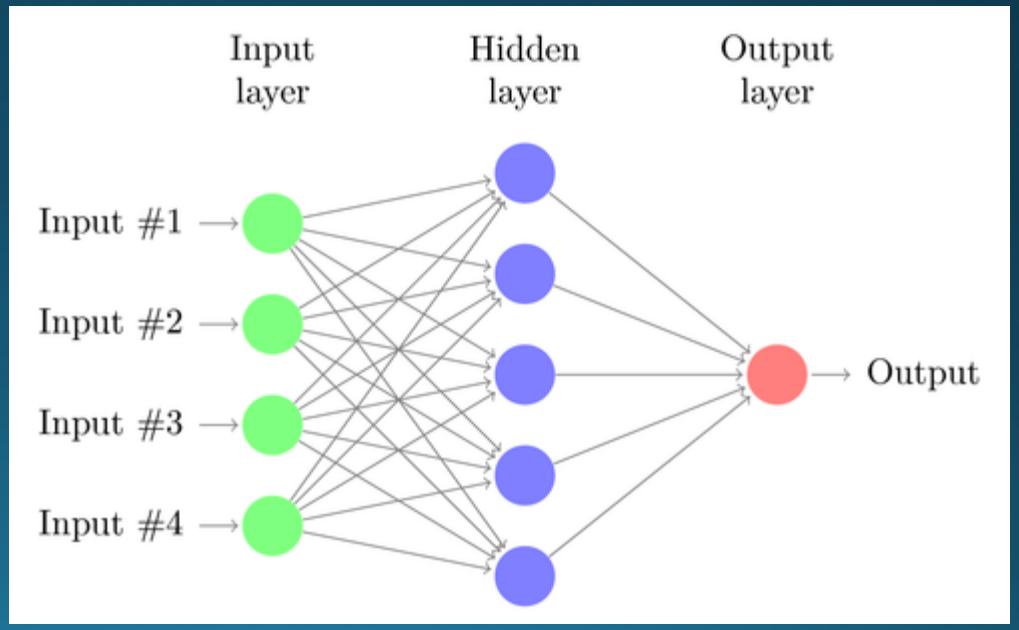

They use(d) R
facebook
ebay
Bank of America
ANZ
Canada
Health Canada
mozilla
FOUNDATION
Google
MERCK
NOVARTIS
AstraZeneca
Microsoft
AMGEN
UBER
KCR
Pfizer
Allianz
FDA
NIST
National Institute of Standards and Technology
U.S. Department of Commerce
BAYER
MOLINA HEALTHCARE
BIOSITE
NEW DIMENSIONS IN DIAGNOSIS
CARDIODX
Roche
Ford
benetech
European Environment Agency
KICKSTARTER
Pacific Northwest
NATIONAL LABORATORY
NewScientist
http://www.revolutionanalytics.com/companies-using-r
http://www.rstudio.com/


airbnb


Top 20 R Machine Learning packages, by Downloads (000) from CRAN
0
20
40
60
80
100
120
140
160
e1071
rpart
igraph
nnet
randomForest
caret
kernlab
glmnet
ROCR
gbm
party
arules
tree
klaR
RWeka
ipred
lars
earth
CORElearn
mboost


rmarkdown
Shiny
tidyr
knitr
ggplot2

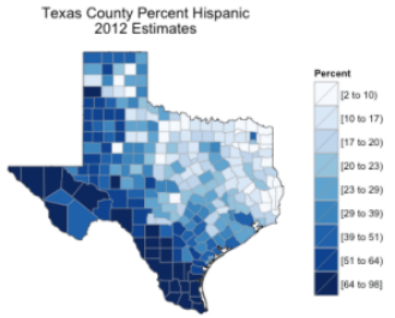

XML
xlsReadWrite
RColorBrewer
survival
Sim.DiffProcGUI
stringr
RMySQL
RODBC
data.table
lme4
tikzDevice
ggplot2
Matrix
nlme
foreign
rpart
Rcpp
rgdal
car
foreach
MCMCglmm
cacheSweave
lattice
randomForest
Rcmdr
plyr
vegan
zoo
xts
sp
rgl
ape
rJava
twitteR
maxent
fortunes
boot
xtable
quantmod
PerformanceAnalytics
survey
Hmisc
RSQLite
maptools
latticeExtra
Sim.DiffProc
RTextTools
reshape
MASS
reshape2

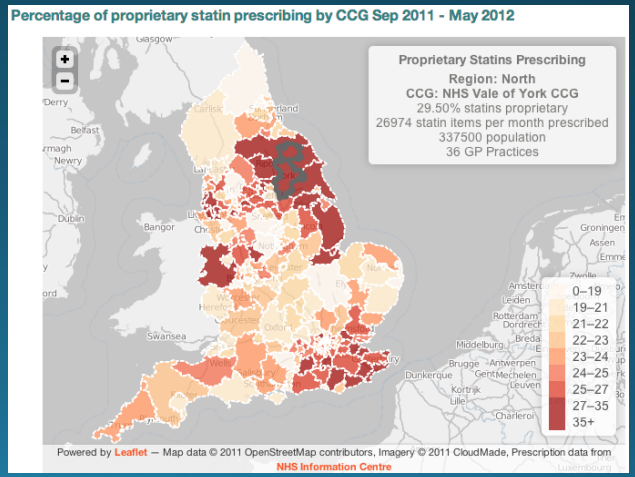

Input layer
Hidden layer
Output layer
Input #1
Input #2
Input #3
Input #4
Output

Learn to Map Census Data in R

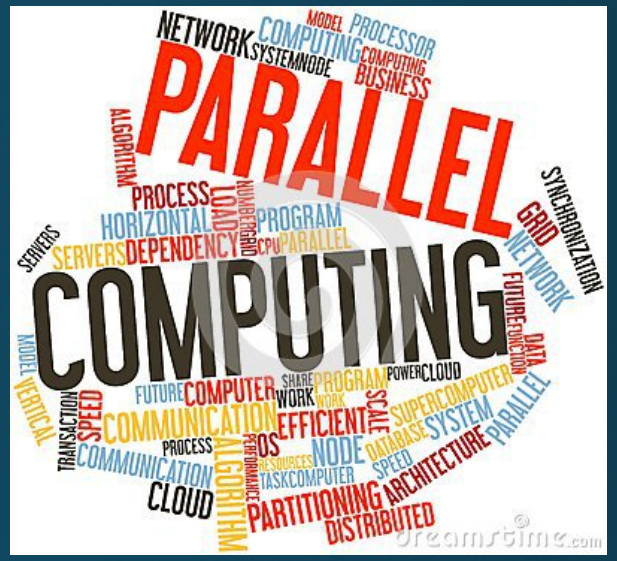
JUNE 25, 2015 / ARI LAMSTEIN / 6 COMMENTS

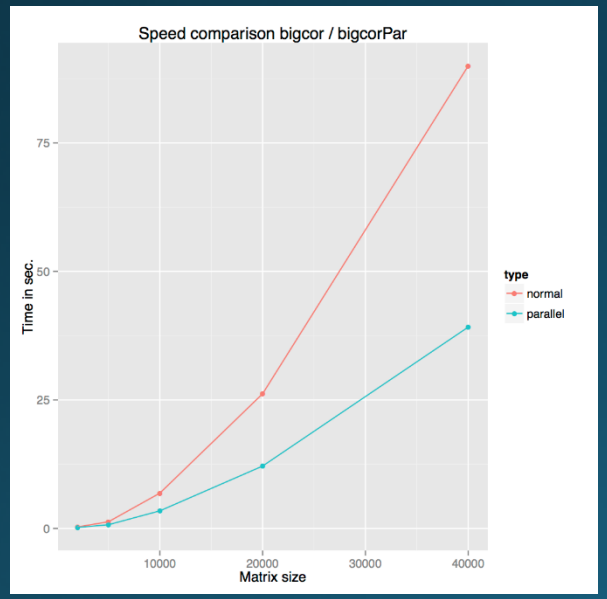
Today I am happy to announce a new free email course: **Mapping Census Data in R**. You can sign up via the form at the bottom of this post. The course is designed to provide similar information to what I covered in my tutorial Analyzing US Census Data with R. In short, it will teach you how to create choropleth maps of US demographics such as this[1]:

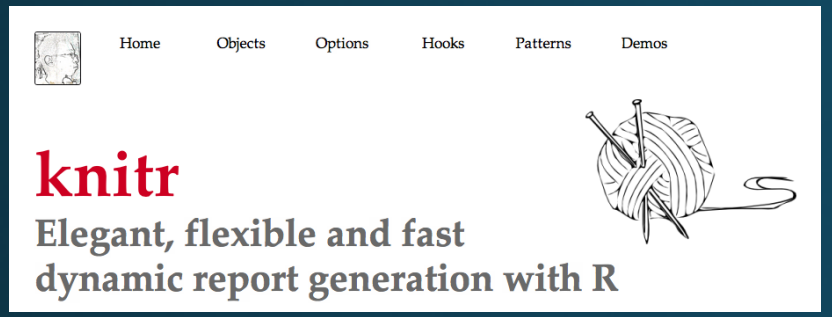




Parallel Processing





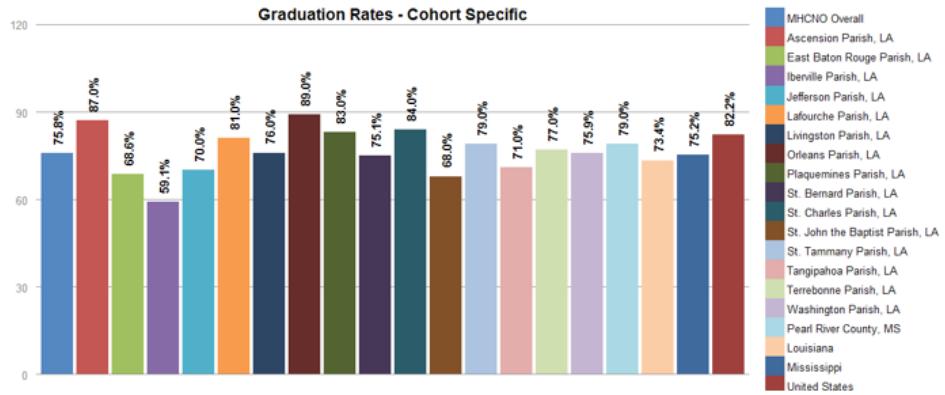


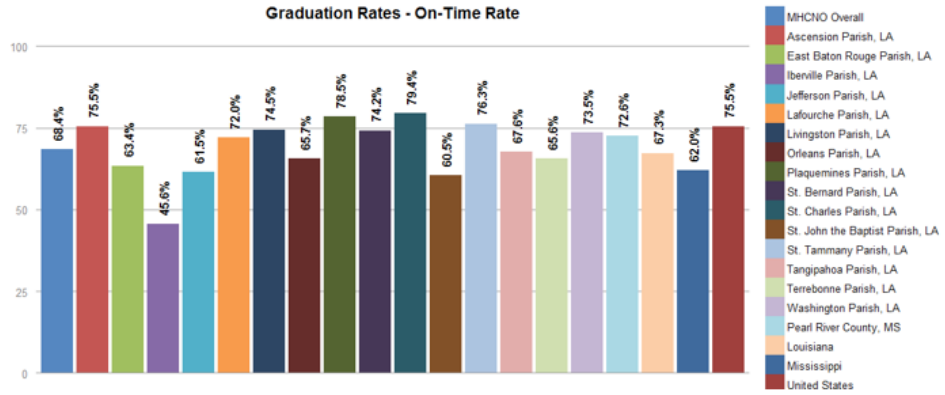
Hispanic Population by Race Alone, Percent

Report Area	White	Black	Asian	Native American / Alaska Native	Native Hawaiian / Pacific Islander	Some Other Race	Multiple Races
Report Area	75.9%	4.86%	0.43%	0.52%	0.1%	11.16%	7.04%
Calhoun County, FL	75.77%	0%	0%	0%	0%	17.31%	6.92%
Charlotte County, FL	86.85%	2.17%	0.3%	0.7%	0%	6.49%	3.49%
Duval County, FL	74.42%	5.32%	0.45%	0.5%	0.12%	11.63%	7.56%
Holmes County, FL	62.77%	2.14%	0%	0%	0%	30.02%	5.07%
Florida	83.87%	3%	0.19%	0.38%	0.03%	9.58%	2.96%
United States	65.37%	2.08%	0.33%	0.91%	0.08%	26.65%	4.58%

Non-Hispanic Population by Race Alone, Total

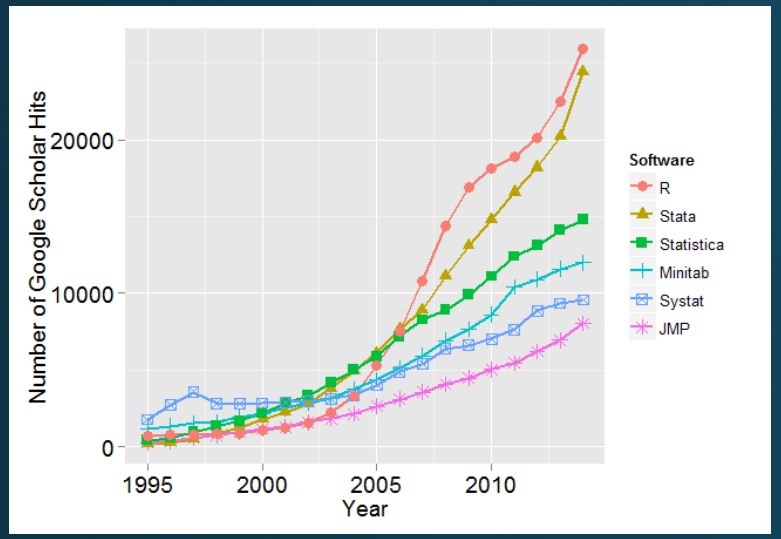
Report Area	White	Black	Asian	Native American / Alaska Native	Native Hawaiian / Pacific Islander	Some Other Race	Multiple Races
Report Area	657,540	268,072	39,124	2,631	812	2,014	24,957
Calhoun County, FL	11,334	1,947	92	219	0	0	256
Charlotte County, FL	138,848	9,201	2,000	308	5	217	2,285

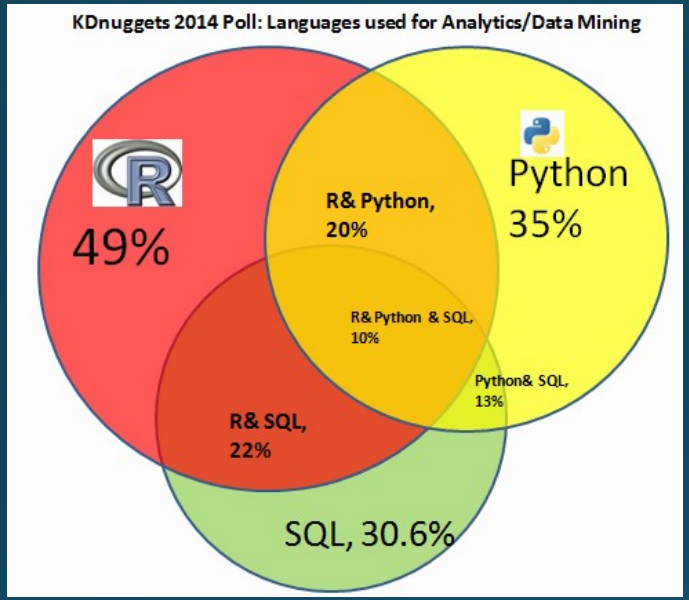




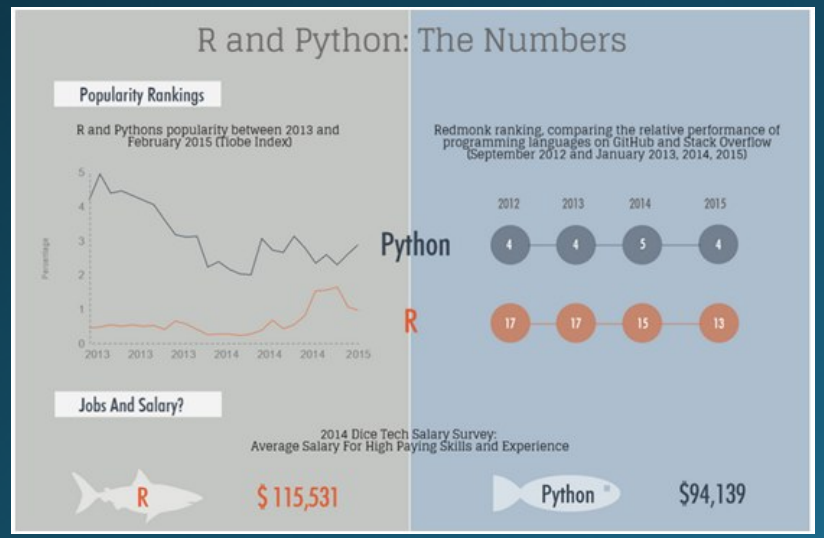
Households with No Motor Vehicle

- Orleans Parish reports the highest rate of households with no motor vehicle (18.48%).

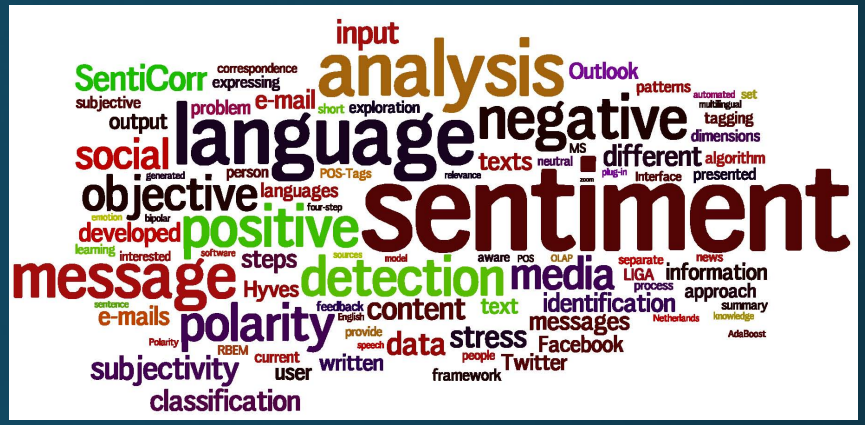


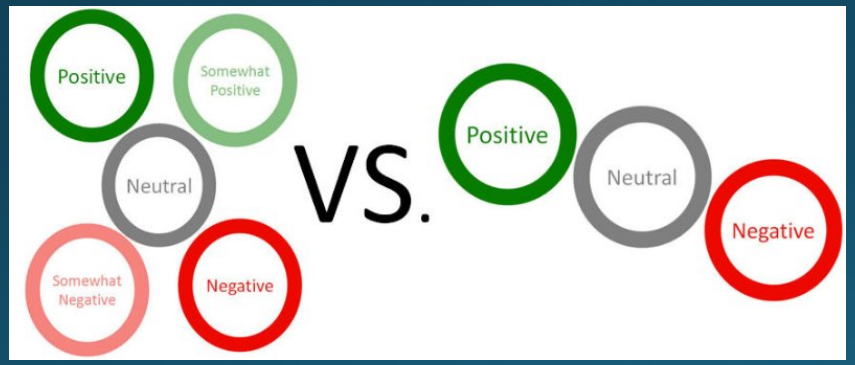


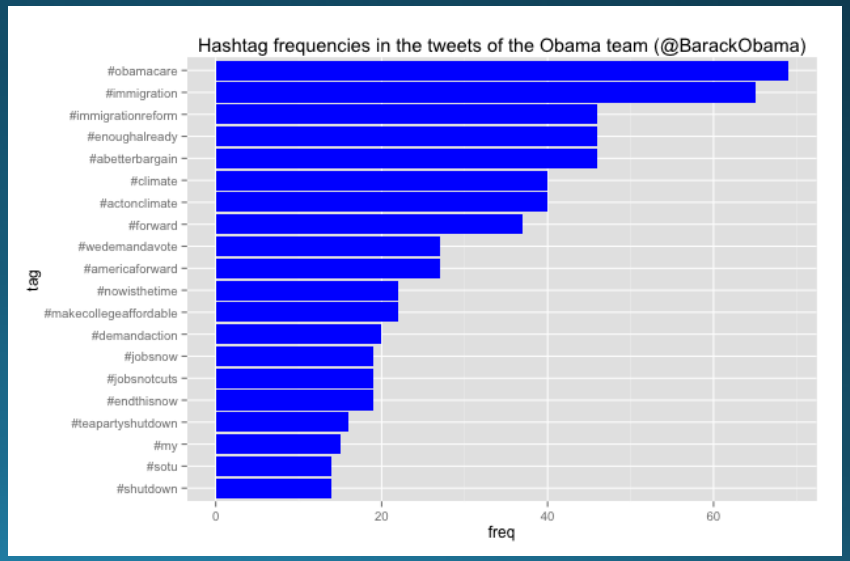
What programming/statistics languages you used for an analytics / data mining / data science work in 2014?	
Language used	% voters in 2014 (719 total) % voters in 2013 (713 total) % voters in 2012 (579 total)
R (352 voters in 2014)	49.0% 60.9% 52.5%
SAS (262)	36.4% 20.8% 19.7%
Python (252)	35.0% 38.8% 36.1%
SQL (220)	30.6% 36.6% 32.1%
Java (89)	12.4% 16.5% 21.2%
Unix shell/awk/sed (63)	8.8% 11.1% 14.7%
Pig Latin/ Hive/ other Hadoop-based languages (61)	8.5% 8.0% 6.7%
SPSS (58)	8.1% not asked not asked
MATLAB (45)	6.3% 12.5% 13.1%

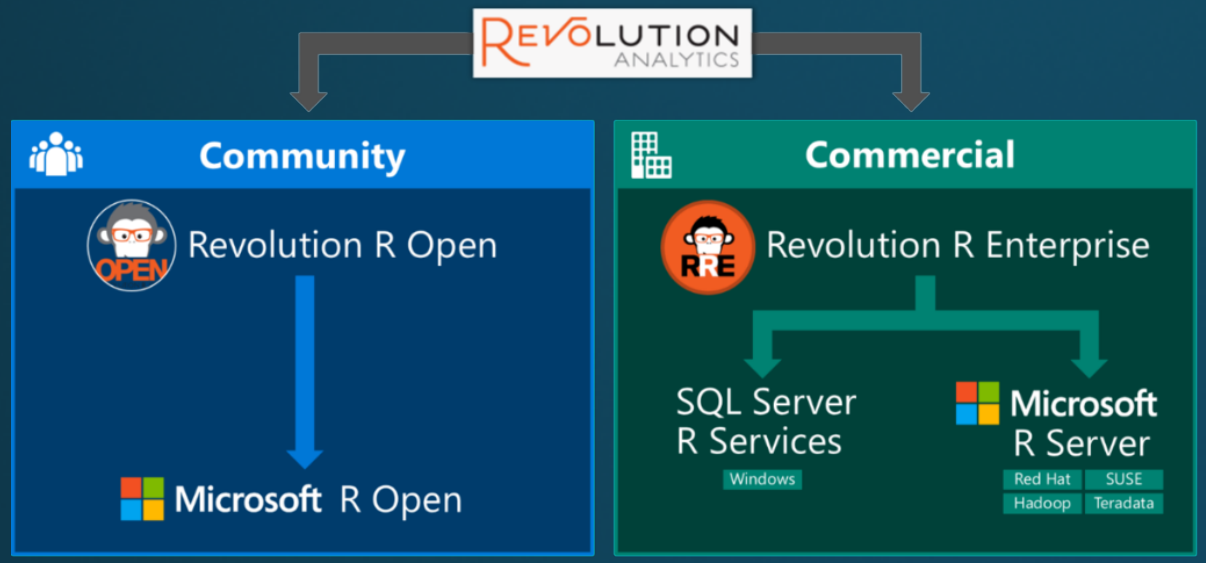



SENTIMENT ANALYTICS


input
analysis
SentiCorr
correspondence
expressing
Outlook
patterns
subjective
output
problem
e-mail
short
exploration
negative
tagging
social
language
texts
neutral
different
algorithm
dimensions
objective
person
POS-Tags
Interface
presented
languages
developed
positive
sentiment
learning
interested
steps
aware
POS
OLAP
separate
news
message
Hyves
detection
media
LIGA
information
process
approach
e-mails
polarity
feedback
content
text
identification
summary
messages
Netherlands
Polarity
RBEM
current
data
stress
Facebook
people
subjectivity
user
written
Twitter
framework
classification

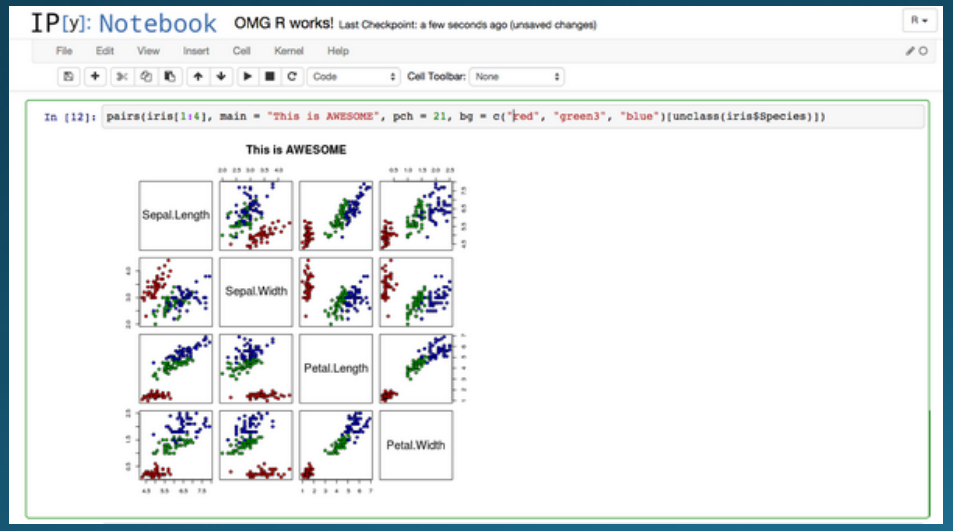

Positive
Somewhat Positive
Neutral
Somewhat Negative
Negative
VS.
Positive
Neutral
Negative


Hashtag frequencies in the tweets of the Obama team (@BarackObama)
tag
#obamacare
#immigration
#immigrationreform
#enoughalready
#abetterbargain
#climate
#actonclimate
#forward
#wedemandavote
#americaforward
#nowisthetime
#makecollegeaffordable
#demandaction
#jobsnow
#jobsnotcuts
#endthisnow
#teapartyshutdown
#my
#sotu
#shutdown
0
20
40
60
freq

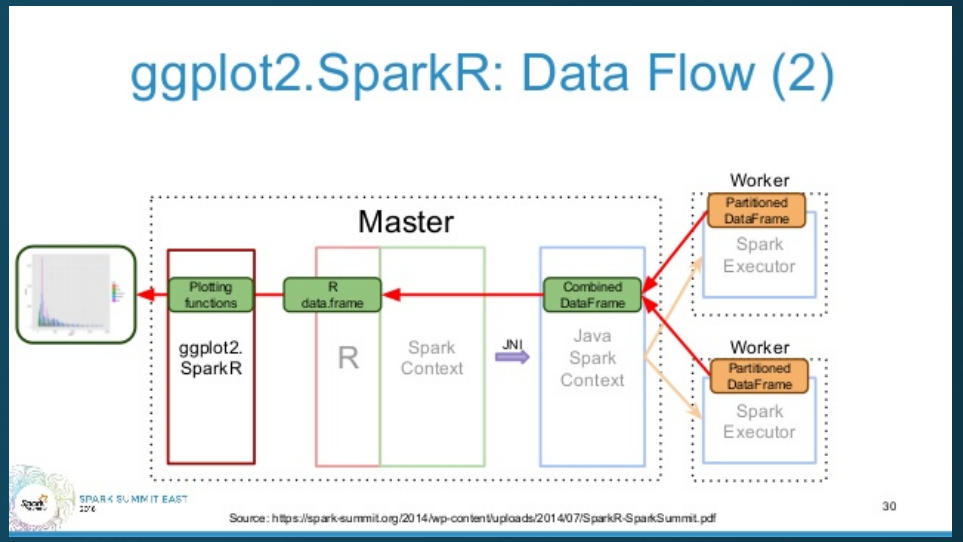

REVOLUTION ANALYTICS
Community
Revolution R Open
Microsoft R Open
Commercial
Revolution R Enterprise
SQL Server R Services
Windows
Microsoft R Server
Red Hat
SUSE
Hadoop
Teradata


OPEN
REVOLUTION R OPEN

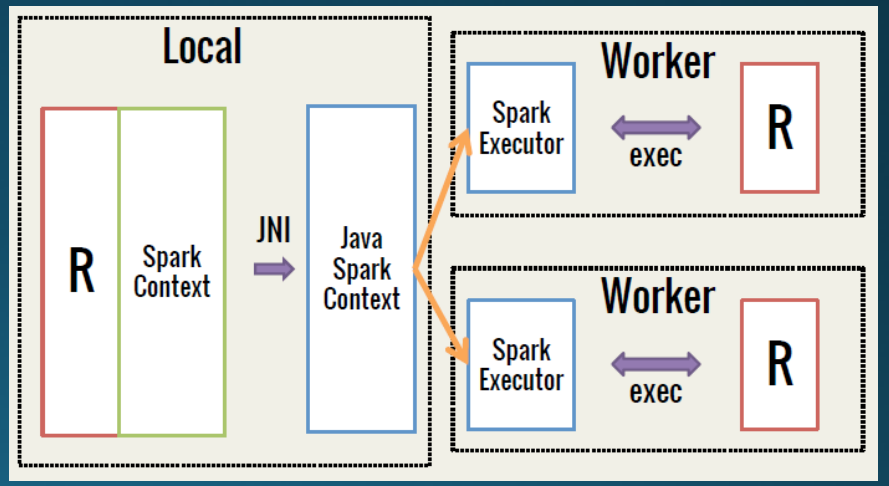

R
consortium


IP[y]: Notebook
OMG R works!
In [12]: pairs(iris[1:4], main = "This is AWESOME", pch = 21, bg = c("red", "green3", "blue")[unclass(iris$Species)])
This is AWESOME
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width


Microsoft
REVOLUTION ANALYTICS



amazon web services™



Fast ! Statistics !

Scalable Spark + R Packages

Interactive Shell Plots

Congratulations!

- You are the newest data scientist at Tesla Motors
- First task: Fuse two data sets
- Skills Required:
- Data cleansing, Missing Value imputation, Classification Modeling, Clustering, Regression Analysis, Data Manipulation

Dataset Link

Dataset A: Tesla Customers

Dataset B: USA Consumer Database

That's easy. Use a common id!

What if no common link exists?

Link Details

Dataset A

ID	Gender M	Gender F	Indivd_Age	Income	Environmentalist
1043	""	female	34	9	1
5021	MALE	""	72	12	1
4863	""	FEMALE	81	6.5	1
4651	male	""	NA	3.2	0
1002	99	99	49	21	0
7523	Male	""	21	-7.9	0

Dataset B

Age	Gender	Income	Environmentalist
28	M	90	Yes
35	F	81	No
14	F	65	No
65	M	42	Yes
44	M	110	No
39	M	49	No

Learn More

- R Cookbook - Many helpful examples and topics. The “Graphs” section helps tremendously with ggplot2 http://www.cookbook-r.com/
- Concise Guide to R – Covers all major R functionality http://www.cis.upenn.edu/~matuszek/Concise%20Guides/Concise%20R.html
- Stack Overflow – The best help is available when you post a question here. http://stackoverflow.com/questions/tagged/r
- Cross Validated – For statistics and modeling help. http://stats.stackexchange.com/
- Introduction to Statistical Learning – The BEST way to learn statistics and machine learning. And it’s free!
  - Book: http://www-bcf.usc.edu/~gareth/ISL/
  - Video Lectures: http://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/