01-Introduction

Practical Machine Learning (with R)

UC Berkeley

Fall 2015

Topics

- Introductions (Me, You & This Class)
- Data Science Outlook 2015 and Opportunities
- Why R?
- Elite Coding
- Tools and Environment

INTRODUCTIONS

Me (Personally)

Shameless Plug

My Skills

- R Programmer (>15 years)
- Machine Learning (>15 years)
- DevOps
- Researcher and Writer : Clinical, Chemisty

Education

- UC Berkeley → (UT Austin)→ UC Santa Barbara → UC Berkeley
- Post-graduate: UC Berkeley, Stanford

Professional Experience

- Lawrence Berkeley National Lab, Allianz, Open Data
- Sept. 2010 Founded Decision Patterns

Professional Interests

- Machine Learning / Statistics
- High Performance Computing
- Applied Statistics and Visualization
- Management of Data Organizations

(Decision Patterns)

Shameless Plug

Decision Patterns

- Founded 2010
- Bring together complementary skills for managing data:

Acquisition * Organization* Storage Access * Utilization

- Our Model
  - Service Consulting
  - Not a start-up -- no VC funding
  - Use consulting margins from to build niche products
- Our Customers
  - Financial Services, Retail, Entertainment, Food, Communications, Defense, Environmental.

What do I like *most* about what I do?

BEST THING

We get to work on a

- variety of problems,
- with a variety of technologies
- in a variety of fields

What do I like *least* about what I do?

WORST THING

We have to work on a

- variety of problems,
- with a variety of technologies
- in a variety of fields

You?

DISCUSSION OF INDIVIDUAL GOALS ?

Class / Objectives

Theory	Practice
Distinguish fundamental aspects of machine learning algorithms	Frame problems to make the suitable for solution via machine learning
Build (train) machine learning models	Collaborate in a group using tools for collaborative/social programming
Evaluate (score) machine learning models	Generate high quality, graphical and textual results
Advanced Topics	Deploy machine learning models to operations

Class Overview : 1

- Introduction to R, setting up the ML developers environment
  - Installing R
  - Installing R Studio
  - Installing packages from CRAN, Bioconductor and Github
  - Exercises

Class Overview : 2

- Fundamentals of Machine Learning
  - Machine learning overview
  - Regression and classification
  - Supervised, unsupervised, and semi-supervised
  - Algorithm types and requirements
  - Exercises

Class Overview : 3

- Linear Regression (2 sessions)
  - OLS Regression
  - Data partitioning
  - Model evaluation and tuning
  - Exercises

Class Overview : 4

- Logistic Regression
  - Logistic Regression
  - Exercises

Class Overview : 5

- Advanced Techniques: Partitioning Methods
  - CART/Regression Trees
  - Clustering
  - K Nearest Neighbors
  - Exercises

Class Overview : 6

- Advanced Techniques: Partitioning Methods
  - CART/Regression Trees
  - Clustering
  - K Nearest Neighbors
  - Exercises

Class Overview : 7

- Advanced Techniques
  - Bagging
  - Bagged Trees / Random Forests
  - Exercises

Class Overview : 8

- Advanced Techniques: Boosting
  - Boosting
  - Neural Networks
  - Support Vector Machines
  - Exercises

Class Overview : 8

- Deployment
  - Diving into the data lake
  - Optimization
  - Delivery and Production

Class Overview : 9

- Final Lecture
  - Exercises
  - Exam

GRADING

Project (50%)

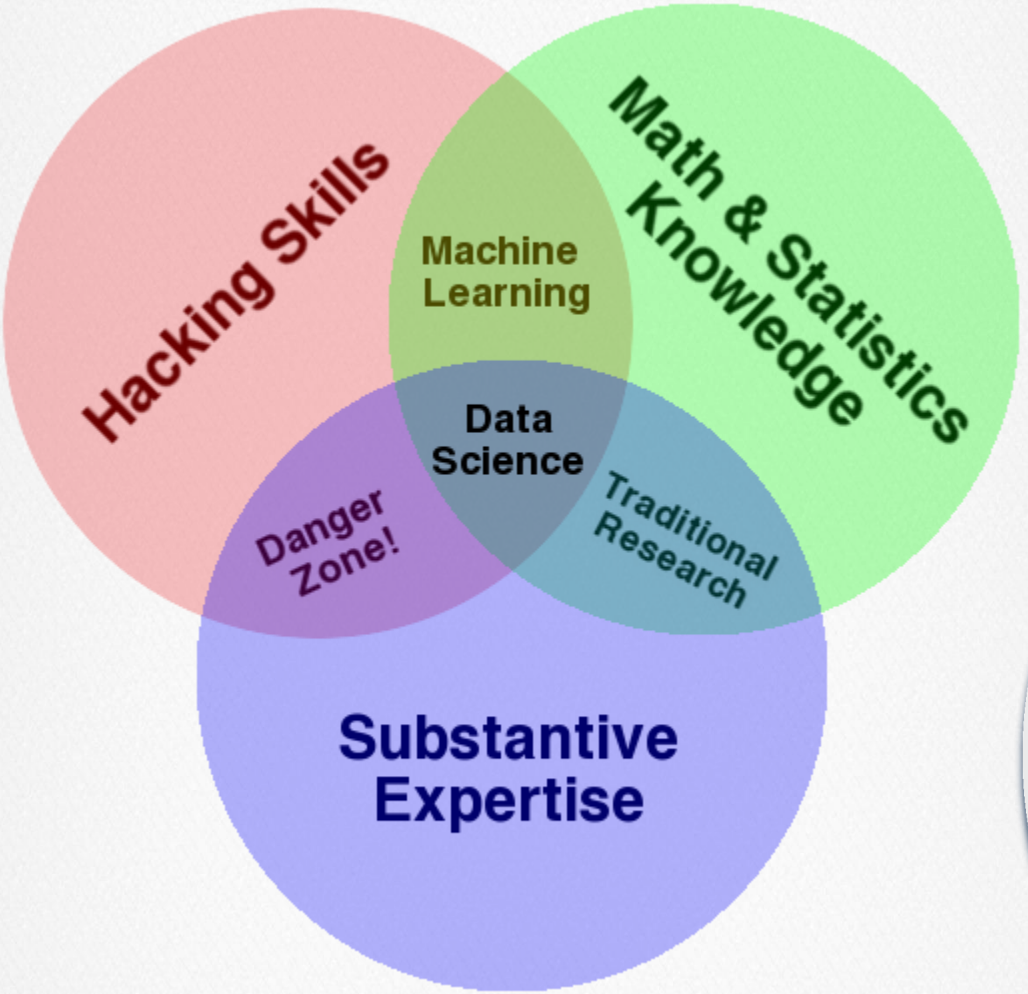
- Identify one machine learning problem you want to tackle. It could be from existing UCI data sets, your work, interest, Kaggle … or I can give you one.
- Frame the problem
- Build Features
- Review linear model and cart
- Build Model
- Deploy

Today 2015-09-14

- Data Scientist / Machine Learning Outlook
- R Language
- Effective Programming

DATA SCIENTIST OUTLOOK 2015

Data Science Venn Diagram



Ref. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

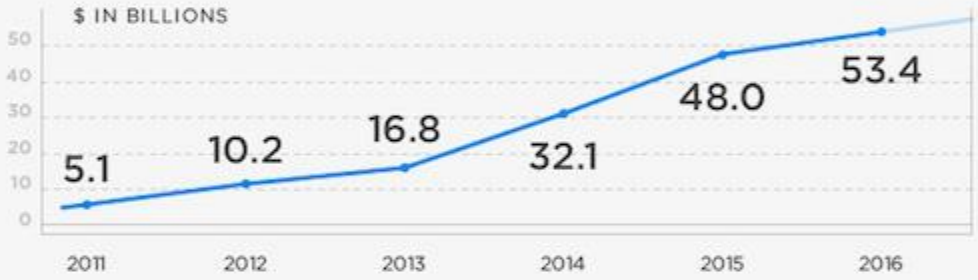
Source: http://venturebeat.com/2013/11/11/data-scientists-needed/

CAREERS IN DEMAND


Job
POSTINGS
15,000%
INCREASE BETWEEN 2011/2012
ACCORDING TO FICO, THERE WAS A 15,000% INCREASE IN JOB POSTINGS FOR DATA SCIENTISTS BETWEEN 2011 & 2012
Currently the job market seeks
140,000–190,000
DATA SCIENTISTS TO FILL OPEN POSITIONS.
IN ADDITION,
1.5 million
data literate managers will need to be retrained or hired to meet needs.

EXPLAINING THE SUDDEN NEED FOR DATA SCIENTISTS

These scientists don't just happen to be getting far more job offers without reason. Today's modern business needs to manage far more data than ever before, and few have the talent on staff for the job. **Projections indicate that the market will experience meteoric growth in the next several years.**



The Big Data Market Forecast

Conclusion: With so much activity going on in the big data space and new data touch points being measured every day, there will be an increasing need for data-driven individuals within organizations to make sense of it all. Is that data-savvy person you?

Machine Intelligence LANDSCAPE

CORE TECHNOLOGIES

ARTIFICIAL INTELLIGENCE
IBM WATSON, MetaMind, Numenta, ai-one, Cycorp, Microsoft Research, nara, Reactor, SI SCALED INFERENCE

DEEP LEARNING
vicarious, Vision Factory, facebook, Baidu IDL, Google, ersatz labs, SKYMIND, SignalSense

MACHINE LEARNING
rapidminer, context relevant, 0xdata H2O, DATARPM, Liftigniter, SPARKBEYOND, Azure ML, yhat, wise.io, Sense, GraphLab, Alpine, Ionian, AYASDI

NLP PLATFORMS
cortical.io, idibon, LUMINOSO, wit.ai, Maluuba

PREDICTIVE APIS
AlchemyAPI, MINDOPS, Google, bigml, indico, ALGORITHMIA, Expect Labs, PredictionIO

IMAGE RECOGNITION
clarifai, MADBITS, DNNresearch, DEXTRO, VISENZE, lookflow

SPEECH RECOGNITION
GRIDSPACE, popUp archive, NUANCE

RETHINKING ENTERPRISE

SALES
Preact, AVISO, RelateIQ, NGDATA, CLARABRIDGE, FRAMED, infer, ATTENSITY, causata

SECURITY / AUTHENTICATION
CROSSMATCH, conjur, EYEVERIFY, BITSIGHT, CYLANCE, bionym

FRAUD DETECTION
sift science, SOCURE, ThreatMetrix, feedzai, Brighterion, VERAFIN

HR / RECRUITING
TalentBin, entelo, predikt, Connectifier, gild, hiQ, CONCEPTNODE

MARKETING
brightfunnel, bloomreach, CommandIQ, AIRPR, RADIUS, TellApart, people pattern, Freshplum

PERSONAL ASSISTANT
Siri, X, Google now, Cortana, cleversense, tempo, Robinlabs, KASISTO, fuse machines, VIV, CLARA LABS

INTELLIGENCE TOOLS
ADATAO, Palantir, Quid, Digital Reasoning, FirstRain

RETHINKING INDUSTRIES

ADTECH
METAMARKETS, dstillery, rocketfuel, YieldMo, ADBRAIN

AGRICULTURE
BLUE RIVER, TerrAvion, ceresimaging, HONEYCOMB, THE CLIMATE CORPORATION, tule, MONTA

EDUCATION
declara, coursera, KNEWTON, kidaptive

FINANCE
Bloomberg, FinGenius, alphasense, KENSHO, Dataminr, minettabrook, BINATIX

LEGAL
Lex Machina, brightleaf, COUNSELYTICS, RAVEL, JUDICATA, Brevia, DiligenceEngine

MANUFACTURING
SIGHT MACHINE, MICROSCAN, BOULDER IMAGING, IVISYS

MEDICAL
Parzival, transcriptic, ZEPHYR HEALTH, Genescient, grand round table, bina TECHNOLOGIES, TUTE GENOMICS

OIL AND GAS
kaggle, biota TECHNOLOGY, TACHYUS, Flutura

MEDIA / CONTENT
Outbrain, newsle, ARRIA, SAILTHRU, wavii, Owlin, NarrativeScience, vscoop, Summly, Prismatic, ai

CONSUMER FINANCE
Affirm, inVenture, zest finance, BILLGUARD, LendUp, LendingClub, Kabbage

PHILANTHROPIES
DataKind, thorn, DATA GUILD

AUTOMOTIVE
Google, Continental, TESLA, MOBILEYE, CRUISE

DIAGNOSTICS
enlitic, 3SCAN, lumiata, ENTOPIS

RETAIL
BAY SENSORS, PRISM SKYLABS, celect, euclid

RETHINKING HUMANS / HCI

AUGMENTED REALITY
wearable intelligence, DAQRI, APX LABS, blippar, META, layar

GESTURAL COMPUTING
THALMIC LABS, omek, LEAP MOTION, FAUNA, eyeSight, 3Gear SYSTEMS, GestureTek, nod

ROBOTICS
intel, LIQUID ROBOTICS, SoftBank, iRobot, Boston Dynamics, jibo, evolution robotics, anki

EMOTIONAL RECOGNITION
affectiva, BEYONDVERBAL, EMOTIENT, BRAINS, cogito

SUPPORTING TECHNOLOGIES

HARDWARE
NVIDIA, XILINX, QUALCOMM, NERVANA SYSTEMS, TERADEEP, Artificial Learning, rigetti

DATA PREP
TRIFACTA, Paxata, tamr, Alation

DATA COLLECTION
diffbot, kimono, CrowdFlower, Connotate, WorkFusion, import.io

www.shivonzilis.com/machineintelligence

Bloomberg BETA

kaggle

oDesk

Elance
work differently

COMPETITION

guru.com

Much of work will not be done
in traditional worker

Windows Azure

PredictionIO

ADATAO
DATA INTELLIGENCE FOR ALL

Google Prediction API

H_2O

INNOVATION

amazon web services™

Dato
Create Intelligence™

Spoils go to those who make products from repeatable processes

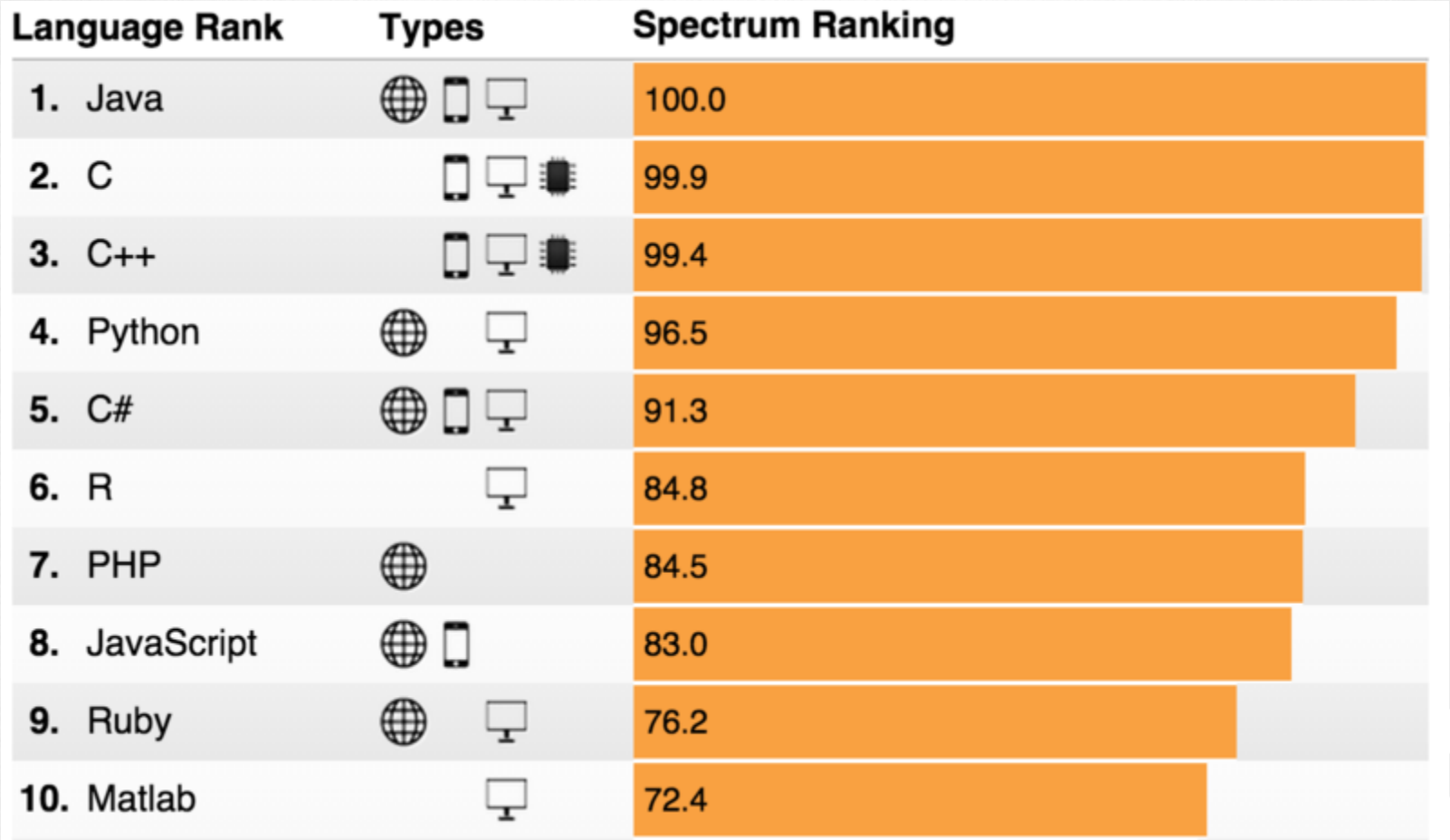
The price for analytics is falling …

WHY R?
WHY NOT PYTHON? … JULIA? … SCALA? …MATLAB?

Popularity

2015-06-04

	python	R
Packages	**PyPI** **60,806 Packages** **35+** updates / day	**CRAN** **6,727**package **20+** updates/day
Popularity (Tiobe)	6th Rank, +0.67%	12th Rank, +1.06%
stackoverflow	430,604	93,943
github SOCIAL CODING	549,014	87,306



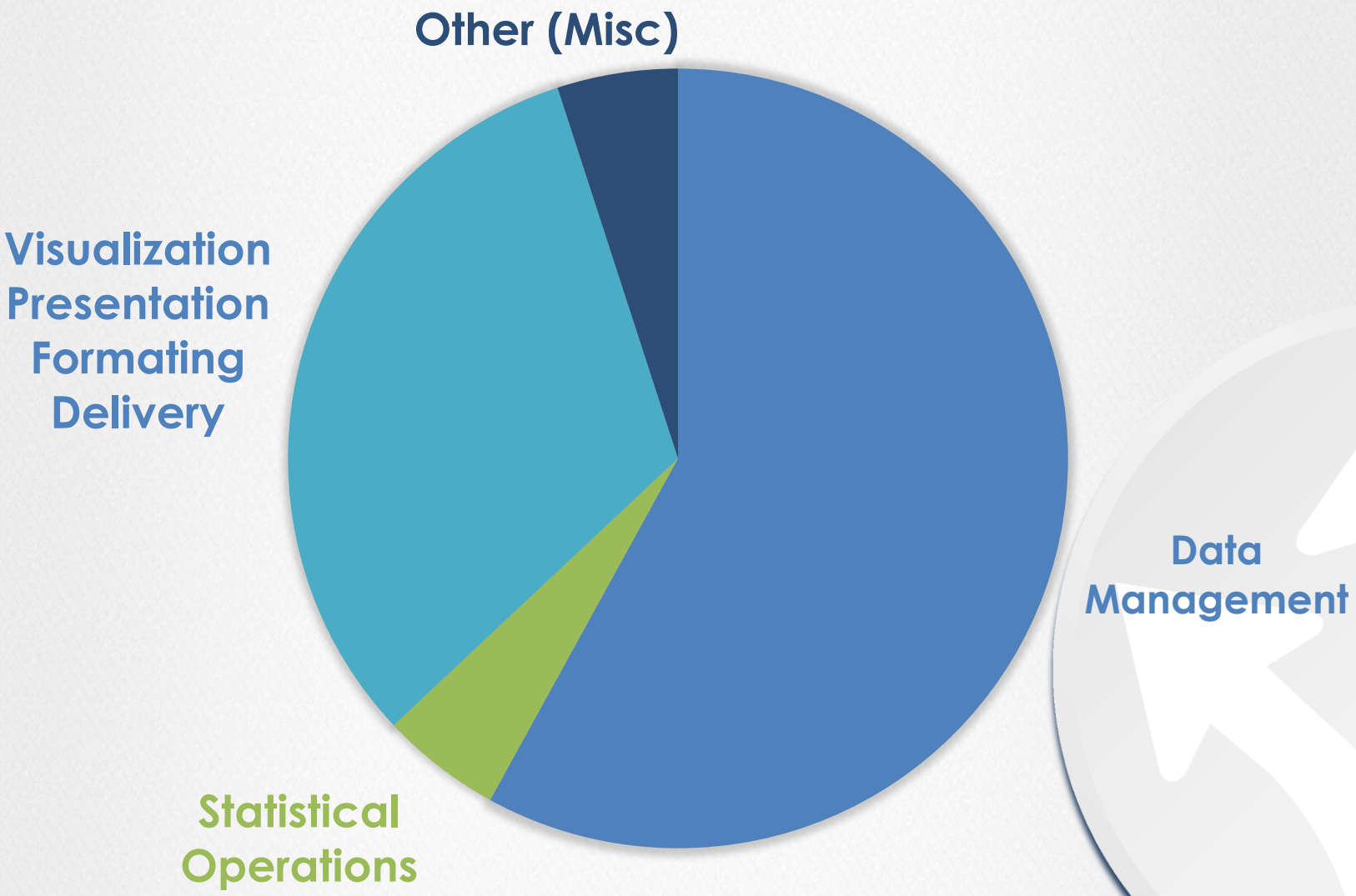
Ref. http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages


RedMonk Q115 Programming Language Rankings
RedMonk
Popularity Rank on Stack Overflow (by # of Tags)
Popularity Rank on GitHub (by # of Projects)
0
25
50
75
100
Java
JavaScript
C#
PHP
Python
C++
CSS
Objective-C
C
Ruby
R
Perl
Scala
Shell
Visual Basic
Matlab
SQL
XML
Delphi
PowerShell
XSLT
Assembly
Haskell
Swift
Groovy
Clojure
ColdFusion
ASP
ActionScript
Go
Lua
Cuda
Makefile
F#
CoffeeScript
Arduino
Erlang
Prolog
Dart
FORTRAN
Scheme
AppleScript
TypeScript
Tcl
Mathematica
Emacs Lisp
OCaml
Common Lisp
Processing
XQuery
VHDL
Racket
Verilog
D
Puppet
AutoHotkey
Rust
Pascal
Apex
VimL
Smalltalk
AutoIt
Standard ML
Ada
Objective-C++
Haxe
Julia
DOT
Vala
Coq
TeX
OpenEdge ABL
IDL
Elixir
Scilab
Xtend
Game Maker Language
Kotlin
Max
Elm
Lasso
Nemerle
Eiffel
SuperCollider
LiveScript
Logos
nesC
Pure Data
xBase
M
SourcePawn
Parrot
Nimrod
PureScript
Gosu
BlitzBasic
AGS Script
Slash
SQF

Opinions

	python™	R
Learning Curve	Easier esp. if coming from OO background	Steeper. More, dedicated
Code Maintainability	Better package system, fewer name clashes	Better documentation Generally less code req'd
Performance	Higher, extensible through Cython, C, C++	Rcpp
Code expressiveness	Hack to extend operators Lazy evaluation	Domain Specific %x% syntax used widely Non-standard evaluation
Dedicated Web Frameworks	Translucent(?)	Shiny
Domain Feature completeness		Rmarkdown, Reproducible Research, ProjectTempate
Vendor Entrenchment		Windows Azure, Oracle, MicroStratety, Birst, Tableau, Oracle

Breakdown of Code Tasks



R Advantages

- Functional / Vectorized
- Dedicated IDE: **Rstudio** (REPL/Interactive Programming)
- **CRAN** and **BioConductor**
- **Shiny**
- **Domain Specific Abstractions**
  - data.frame / data.table / dplyr
  - model formula
  - purr

R Limitations

- Slow
- In-memory
- Not-scalable

What about …

ELITE CODING

Elite Coding / 1

- Follow Established Design Patterns

Creativity is generally a bad thing

Goal	Description	R Packages
Ad hoc analysis	Create a process	**ProjectTemplate, Rmarkdown, knitr**
Package Development	Create a package	**Rstudio**, **Roxygen2**, **devtools**
Application : Interactive	Web application	**Shiny**, **OpenCPU Javascript**
Application : Automated	Code to be scheduled or called as an event	**Rscript** (R –e), **optigrab**, **crontab**

Elite Coding / 2

- Adopt standards: python™
  - Hadley Wickham's **Advanced R** sytle guide http://adv-r.had.co.nz/Style.html
  - Decision Patterns Style Guide
  - Do not follow Google's coding convention
  - Cf. Python's Standards
    - PEP-8 Naming and Formating
    - PEP-257 Documentation
    - PEP-20 Readability
- Use version control: git
  - Github, Bitbucket, Gitlab
  - Best GUI: Atlassian Sourcetree
- Use Agile Methods
  - Track issues: JIRA, Github, Gitlab

Commit early and often.

Good PM is worth every penny.

R Resources

ONLINE

- (META)CRAN
  - Packages
  - Task Views
- Stackoverflow.com
- r-bloggers.com
- Advanced R Programming
- Github

Offline

The Art of R Programming

Norm Matloff

ISBN-13: 858-2592222227

ISBN-10: 1593273843

R in Action: Data Analysis and Graphics with R

Robert Kabacoff

ISBN-13: 978-1617291388

ISBN-10: 1617291382

>install.packages("package name")

Data Analysis Useful Libraries in R

1 Pre Modeling Stage

1. Data Visualization	ggplot2, googleVis
2. Data Transformation	plyr, data.table
3. Missing Value Imputations	MissForest, MissMDA
4. Outlier Detection	Outliers, EVIR
5. Feature Selection	Features, RRF
6. Dimension Reduction	FactoMineR, CCP

2 Modeling Stage

1. Continuous Regression	car, randomforest
2. Ordinal Regression	RMiner, CoreLearn
3. Classification	Caret, BigRF
4. Clustering	CBA, RankCluster
5. Time Series	forecast, LTSA
6. Survival	survival, Basta

3 Post Modeling Stage

1. General Model Validation	LSMeans, Comparison
2. Regression Validation	RegTest, ACD
3. Classification Validation	BinomTools, DAIM
4. Clustering Validation	ClustEval, SigClust
5. ROC Analysis	PROC, TimeROC

Other Libraries

A. Improve performance	Rcpp,parallel	B. Work with web	XML, jasonlite, httr	C. Report results	shiny, RMarkdown
D. Text Mining	tm, twitteR	E. Database	sqldf, RODBC, RMongo	F. Miscellaneous	swirl, reshape2, qcc

Analytics Vidhya
Learn Everything About Analytics

R'S NOBELS

For Machine Learning

CARET
(CLASSIFICATION AND REGRESSION TRAINING)

Code readability, interactive programming

MAGRITTR
PIPER
BACKPIPE

Munging and data management

DATA.TABLES
DPLYR

https://github.com/Rdatatable/data.table/wiki/Benchmarks-%3A-Grouping

Simple Scaling-out

FOREACH ITERTOOLS

Increasing Performance

RCPP GMATRIX

Visualization

GGVIS
GGPLOT2

APPENDIX

Problem

Given a vector of numbers (x), write a function (f) that returns a vector of numbers containing the *product* of every other number excluding the current index.

Example:

```
> x <- c( 1, 5, 2, 8 )
> f(x)
[1] 80 16 40 10

# 5*2*8, 1*2*8, 1*5*8, 1*2*5
```

Solution

Given a vector of numbers (x) write a function (f) that returns a vector of numbers containing the *product* of every other integer excluding the current index.

Example:

```
> x <- c( 1, 5, 2, 8 )
> f(x)
[1] 80 16 40 10

# 5*2*8, 1*2*8, 1*5*8, 1*2*5
```

Solution:

```
f <- function(x) prod(x) / x
```

Problem

Write a function f(x) to accept an integer vector, and returns a vector with those numbers … except for:

multiples of 3 = "Fizz"

multiples of 5 = "Buzz".

multiples of 3 and 5 = "FizzBuzz"

Example:

```
> x <- 1:20
> f(1:20)
 1 2 Fizz 4 Buzz Fizz 7 8 Fizz Buzz 11
 Fizz 13 14 FizzBuzz 16 17 Fizz 19 Buzz
```

Solution:

```
f <- function(x)
  ifelse( x %% 15 == 0, "FizzBuzz",
  ifelse( x %% 3 == 0, "Fizz",
  ifelse( x %% 5 == 0, "Buzz", x ) ) )
```

You

- How many of you are students? Professionals?
- How many have
  - > 1 year using R?
  - > 3 years?
  - > 5 years?
- How many use R as your principal data.science tool?
- How many use
  - Python
  - Julia
  - SAS or SPSS
  - Spark/Scala
  - Java
- Ever spend too much time debating which technology fits?