DECISION PATTERNS

01-Introduction

Practical Machine Learning (with R)

UC Berkeley

Spring 2016

Topics

- Introduction
- Tools and Environment
- Exercise
- Introductions (continued)
- Data Science, Machine Learning and Opportunities
- Elite Coding
- Exercises
- Machine Learning

INTRODUCTIONS

Me (Personally)

Shameless Plug

My Skills

- R /Python Programmer (>15 years)
- Machine Learning (>15 years)
- DevOps
- Researcher and Writer : Machine Learning, Clinical Medicine, Chemistry, Finance

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 data strategy:

Acquisition * Organization* Storage Access * Utilization

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

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

TOOLS AND ENVIRONMENT

Exercise: Set-Up Tools And Environment

- Install **R** → **CRAN**
- Install **R Studio Desktop™** (IDE)
- Install **git**

- Create **github** account
  - Send name, student id to christopher.brown@berkeley.edu

GIT

Git / GITHUB / Source Tree Workflow

What is it?

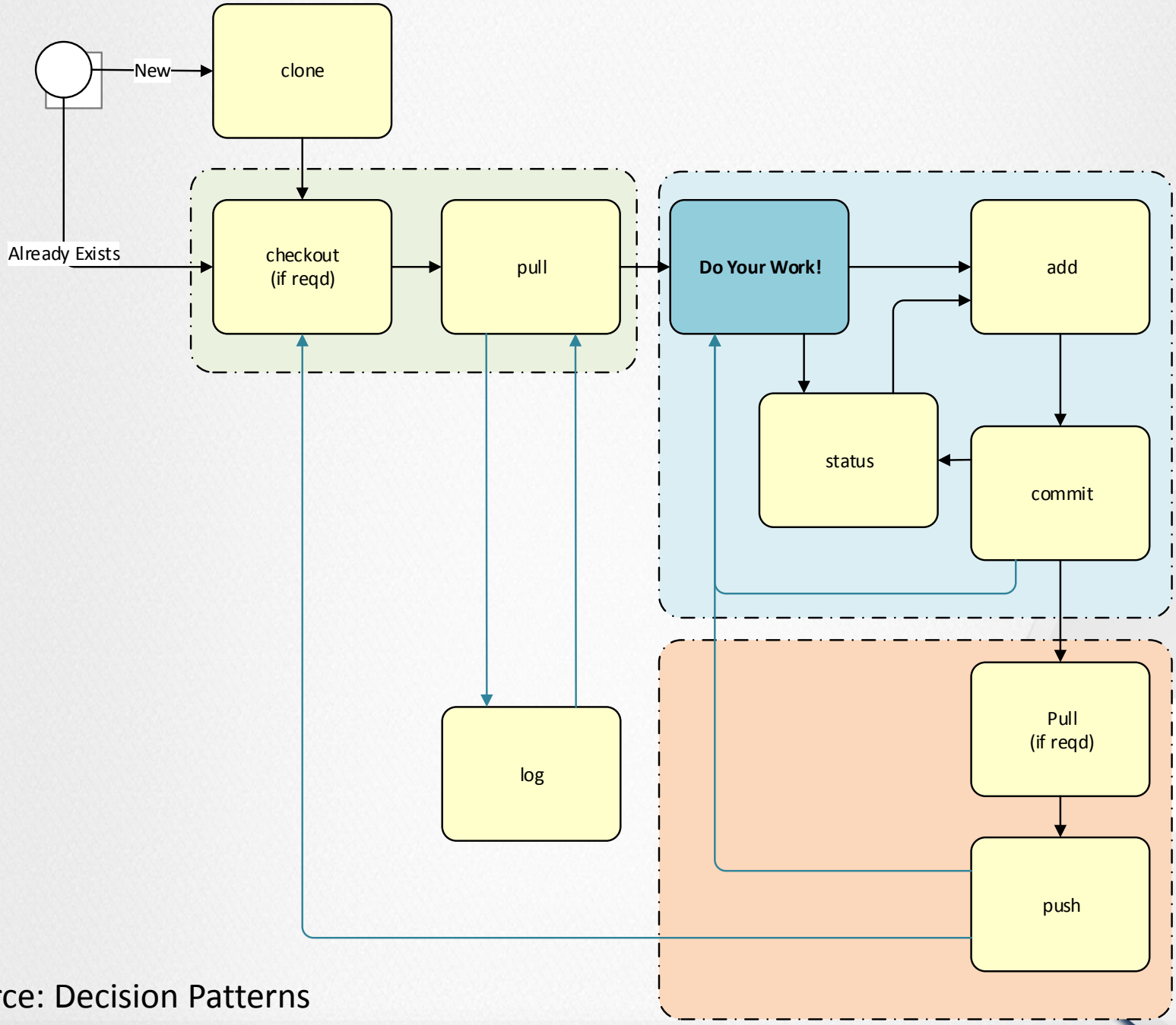
A source control tool (and ***process***) to promote collaborative development

Interfaces

- Github
- Source Tree™
- *R Studio*
- *command-line*

Features

- Distributed
- Each clone contains complete history
- Ability to return to
- Branch in and merging



source: Decision Patterns

Git Commands

- **Repo(sitory)**: location where files are stored. If different from original source: "fork"
- **Branch**: Copy of code that can be independently worked on.

- checkout: Change to specific branch/commit.

- **add**: Tell which files to "stage" (accept) commit. Done on a per-file basis.
- **commit:** accept changes.

- **pull**: Retrieve changes from remote repository
- **push**: Send committed change to remote repository

- **log**: review history of commits
- **status**: review "staged" status

Exercises

- Create github account
- Send account log-in, student id to
  christopher.brown@berkeley.edu

- Clone class repository

```
git clone https://github.com/csx460
```

You?
DISCUSSION OF INDIVIDUAL GOALS ?

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?

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
	Deploy machine learning models to operations

Required Text

Applied Predictive Modeling

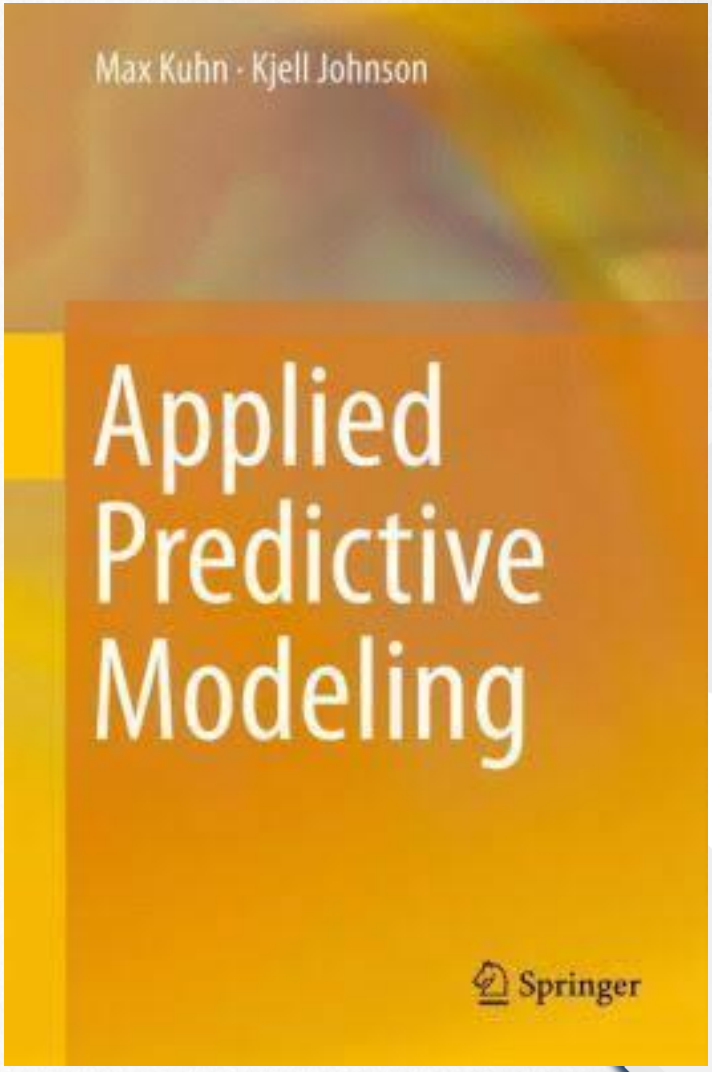
ISBN-13: 978-1461468486

ISBN-10: 1461468485

Kuhn, Max and Johnson, Kjell

Springer Science+Business

2013



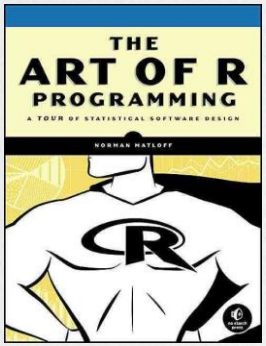
Additional Resources

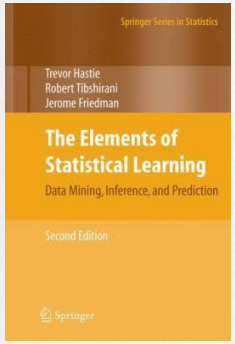
Texts
(not used in this class)

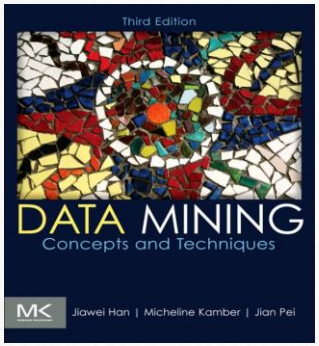
The Art of R Programming
by Norm Matloff

Elements of Statistical Learning
by Hastie, Friedman, Tibshirani

Data Mining Concepts and Techniques
by Han, Kamber, Pei







Online

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

Contacts / Coordinates

- Christopher Brown
  christopher.brown@berkeley.edu
  checked once / day (mornings)
  phone #

- Class Website
  - https://github.com/CSX460

- Google Group: CSX460

Grading

- **~8 Weekly Exercises (80%)**
  - Exercises are **Rmarkdown** in the github
  - Due at the beginning of class each week
  - Submitted via **github** commits
    - Please email me your **github** login
    - ! Github commits are timestamped
  - Answers reviewed in class
  - Work on them in class, time-permitting
- **Class Participation (20%)**
- **Attendance is Mandatory**
  - **no unexcused absences.**

** PARTICIPATE **

RMARKDOWN

RMARKDOWN

What is it?

- Simple text mark-up syntax
- that supports the markdown standard
- And allows incorporation of R analysis and graphical output

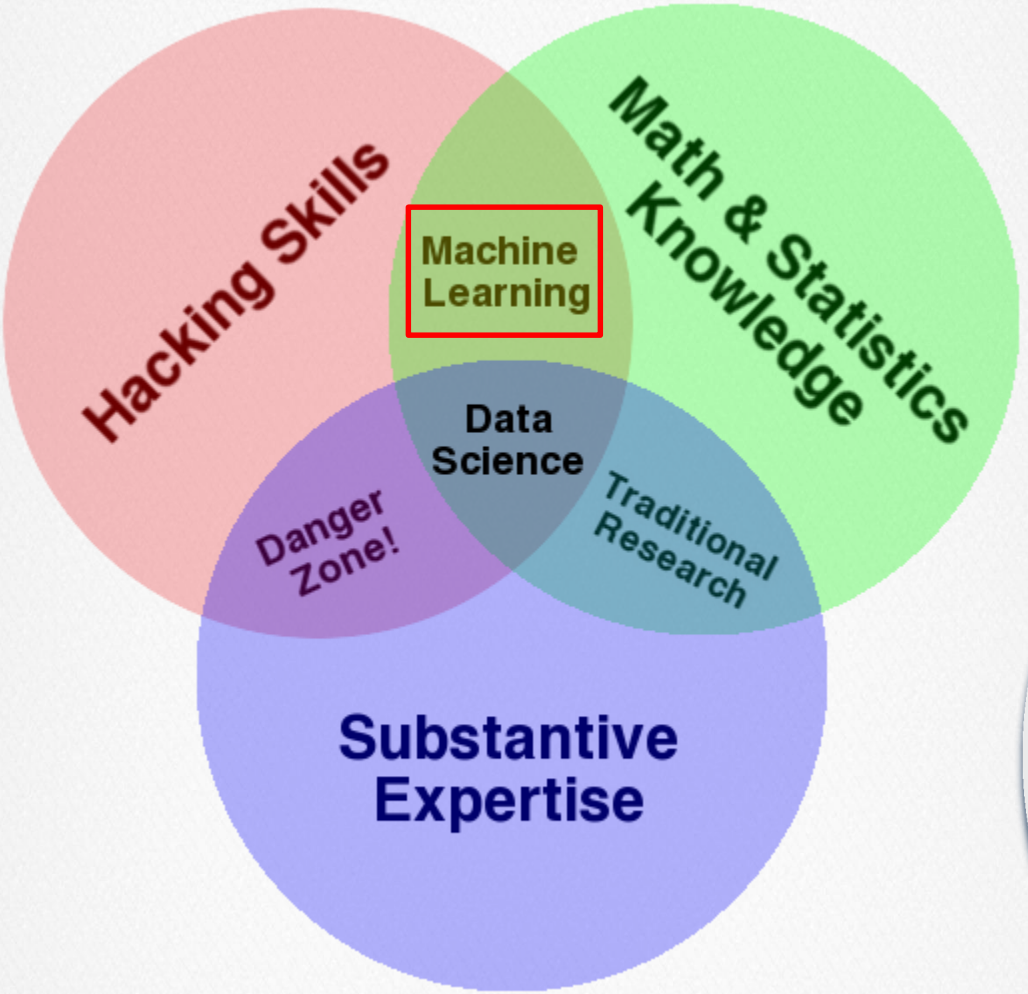
- Are assignments will be done in Markdown …
- Simply put your answers in the space provided
- → Demonstration

ML OVERVIEW

Example of ML algorithm(s)

- Spam Filter
- handwriting recognition (svm)
- Traffic engineering (lights)
- Weather prediction
- Sentiment analysis (social media)
- Netflix Recommender
- Fraud detection (Visa)
- Imaging processing
- Intrution detection
- Self-driving cars

Data Science Venn Diagram



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

MACHINE INTELLIGENCE 2.0

AGENTS

PROFESSIONAL
Howdy! x.ai clara KASISTO DigitalGenius OVERLAP.CC meekan fuse|machines PRIMER

PERSONAL
facebook XIAOICE large assistant.ai nestor @wesome Magic

OS INTERFACES
Siri Cortana VIV maluuba api.ai CognEA Google Now

AUTONOMOUS SYSTEMS

AIR
3DR dji PROJECT LOON Google VERTICAL DroneDeploy AIRDOG SKYCATCH SKYDIO Airware LILY

GROUND
Google UBER TESLA CRUISE MOBILEYE COMMA AdasWorks

SEA
LIQUID ROBOTICS bluefin data OPENROV BluHaptics

INDUSTRIAL
KIVA Systems fetch robotics HARVEST AUTOMATION CLEARPATH AVIDBOTS ENERGID rethink robotics GREYORANGE OSARO

ENTERPRISE

SECURITY / FRAUD
Sentinel graphistry BITSIGHT feedzai AREA 1 drawbridge sift science CYLANCE Brighterion

HR / RECRUITING
textio hiQ gild SpringRole entelo unitive GIGSTER

SALES
6sense infer people pattern Preact Prism AVISO Vidora sentient salespredict Gainsight

MARKETING
LiftIgniter RADIUS brightfunnel retention SCIENCE AIRPR

CUSTOMER SUPPORT
CLARABRIDGE QUANTIFIND Wise.io ACTIONIQ FRAMED DigitalGenius

INTERNAL INTEL
Alation ADATAO Palantir sapho lucid Rainbird SKIPFLAG agolo Digital Reasoning Narrative Science

MARKET INTEL
Quid mattermark DataFox bottlenose PREMISE enigma CB INSIGHTS

PLATFORMS

RESEARCH / AGI
OpenAI vicarious Google DeepMind Numenta Cycorp nnaisense SCALED INFERENCE CURIOUS GEOMETRIC INTELLIGENCE

FULL STACK
context relevant CognitiveScale NVIDIA TERADEEP QUALCOMM nervana SYSTEMS

MACHINE LEARNING
Dato rapidminer cortical.io AYASDI amazon web services Azure Machine Learning naralogics PredictionIO SKYTREE bigml blueyonder

INDUSTRIAL IOT
ThingWorx UPTAKE IMUBIT Preferred Networks Alluvium xively PLANET OS

AUDIO
Gridspace TalkIQ nexidia vocaliq NUANCE Expect Labs popUp archive

VISION
ORBITAL INSIGHT Descartes Labs DEXTRO cortica clarifai PLANET LABS MetaMind

DATA ENRICHMENT
diffbot Paxata TRIFACTA iDIBON WorkFusion loop ai Labs CrowdFlower

INDUSTRIES

ADTECH
ADTHEORENT dstillery BEYONDVERBAL METAMARKETS TAPAD rocketfuel affectiva

AGRICULTURE
BLUE RIVER tule TerrAvion mavrx THE CLIMATE CORPORATION CERES TECHNOLOGIES HONEYCOMB

FOR GOOD
Conservation Metrics DataKind DATA POP thorn BAYES IMPACT

RETAIL FINANCE
inVenture affirm earnest MIRADOR Lendo zest finance LendUp

LEGAL
Everlaw RAVEL LEGAL ROBOT seal BEAGLE ROSS Lex Machina

MATERIALS & MFG
zymergen AUGMATE GINKGO BIOWORKS CITRINE TECHNOLOGIES SIGHT MACHINE CALCULARIO Eigen Innovations

HEALTHCARE
deep genomics 3SCAN enlitic Calico BUTTERFLY Network, Inc. Atomwise Recombine color METABIOTA GRAND ROUNDS Google Life Sciences IBM Watson Health

INDUSTRIES (CONT'D)

EDUCATION
KNEWTON coursera turnitin gradescope UDACITY KHANACADEMY

TRANSPORT & LOGISTICS
NAUTO taleris PRETECKT clearmetal

INVESTMENT FINANCE
Bloomberg Quantopian Dataminr KENSHO iSENTIUM NEURENSIC alphasense

TECH USER TOOLS

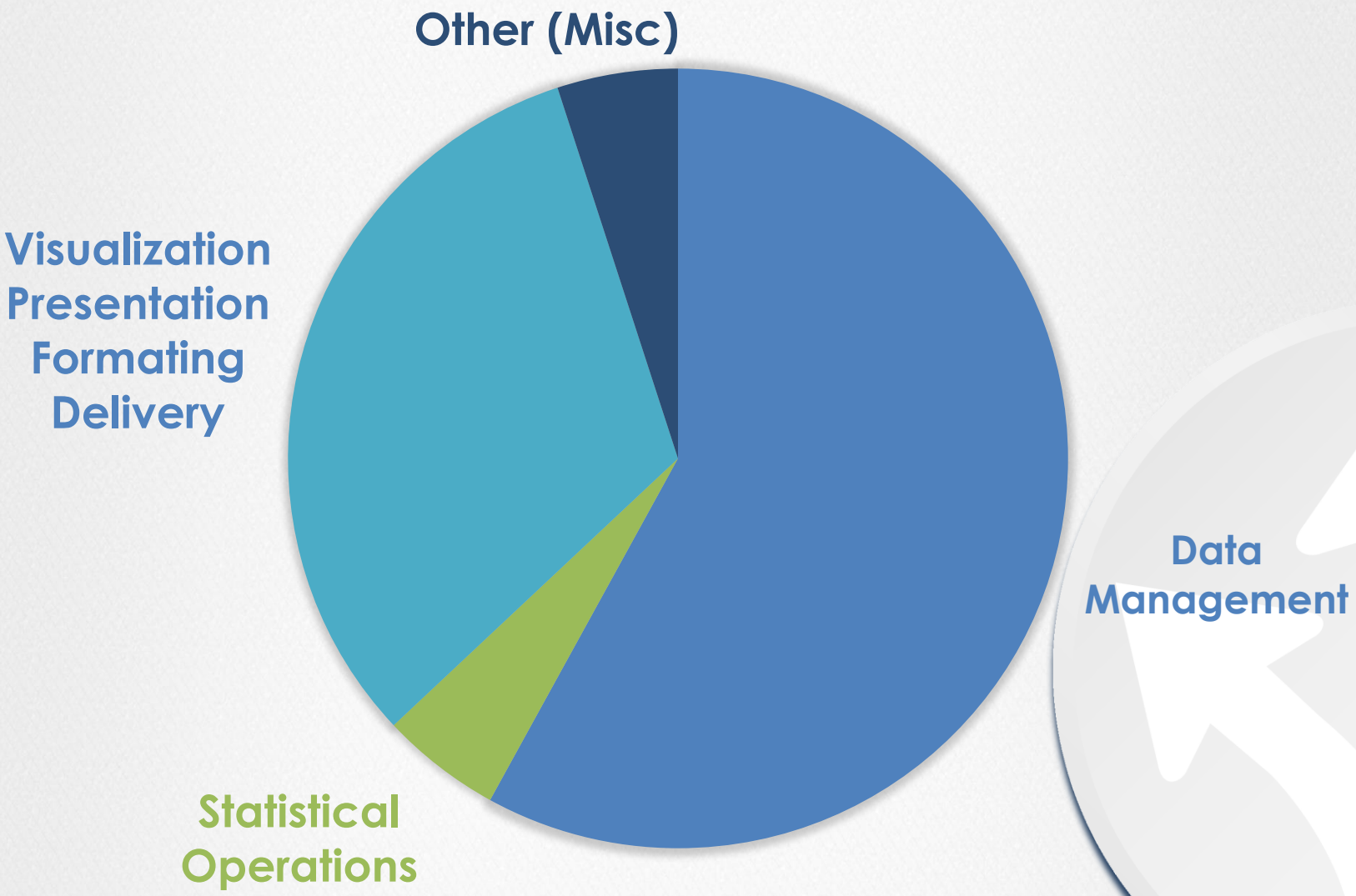
DATA SCIENCE
DOMINO kaggle Sentenai sense yseop Outlier ŷhat DataRobot

MACHINE LEARNING
Cortana Analytics AlchemyAPI glowfi.sh IBM Watson Platform Anodot MonkeyLearn (h [s]) HyperScience fuzzy.io SIGOPT 0xdata H2O SPARKBEYOND indico

OPEN SOURCE
SKYMIND TensorFlow seldon Caffe theano Spark MLlib Microsoft DMTK spaCy DL4J SciKit CGT

SHIVONZILIS.COM/MACHINEINTELLIGENCE

Breakdown of Code Tasks



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

Commit early and often.

- Use Agile Methods
  - Track issues: JIRA, Github, Gitlab

Good PM is worth every penny.

>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

Useful R Packages

```
> install.packages("package-name")
```

- ***ML Framework***: *caret* (Classification and Regression Training)
- **Pipe operators**: *magrittr (pipeR, backpipe)* (shiny)
- **Tables**: *data.tables, dplyr*
- **Visualization**: *ggvis, ggplot2*
- **Reporting**: *knitr, rmarkdown, shiny*

EXERCISES IN CLASS

Question 1

What is machine learning?

A formal ***process*** for building a ***model***

QUESTION 2

What is a model?

a ***function*** that ***estimates*** a ***response*** associated with (a set of) known ***predictors***

$$\hat{y} = f(\vec{x})$$

Question 3: What are the Properties of f

- Should be easy* to evaluate
- Takes a one or more values of inputs
- Yields a single output value for each input

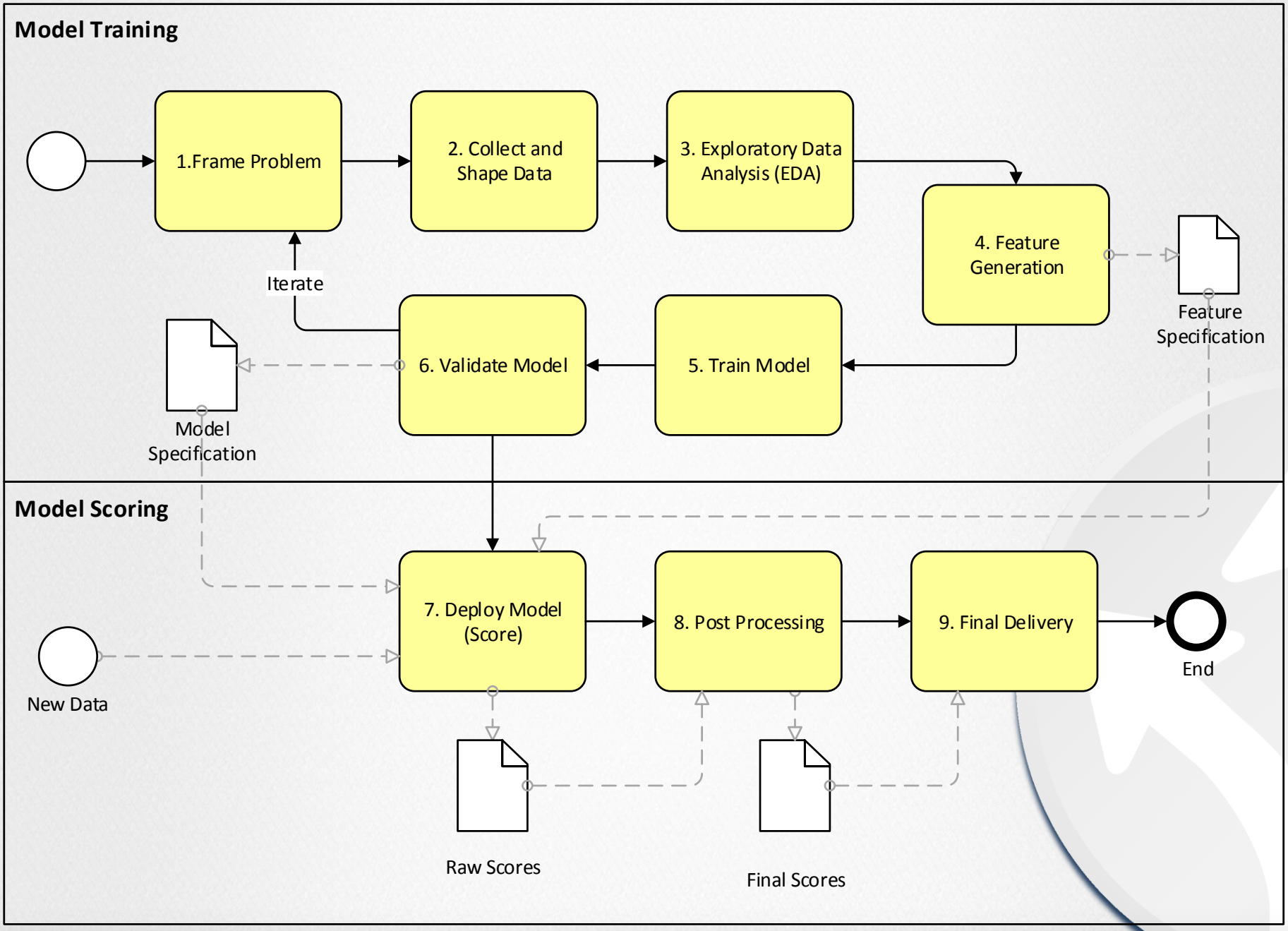
- Output, $\hat{\boldsymbol{y}}$, should be "close to" observed values, $\mathbf{y}$:

$$\hat{\boldsymbol{y}} \sim \boldsymbol{y}$$

* Computational cheap/efficient

Question 4

How do we find f?


Model Training
1.Frame Problem
2. Collect and Shape Data
3. Exploratory Data Analysis (EDA)
4. Feature Generation
Feature Specification
Iterate
6. Validate Model
5. Train Model
Model Specification
Model Scoring
7. Deploy Model (Score)
8. Post Processing
9. Final Delivery
End
New Data
Raw Scores
Final Scores

APPENDIX

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

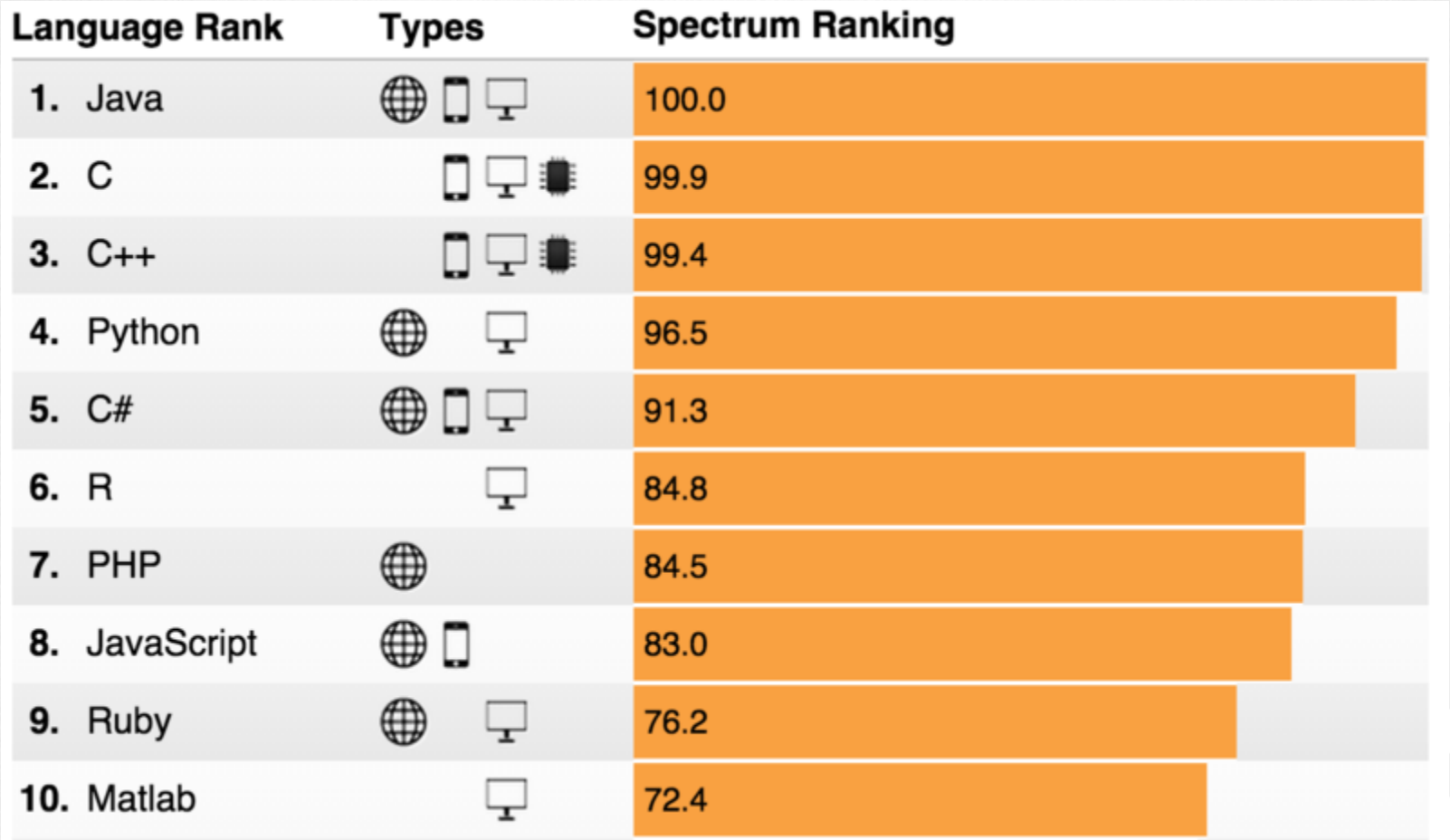
- Final Lecture
  - Exercises
  - Exam

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



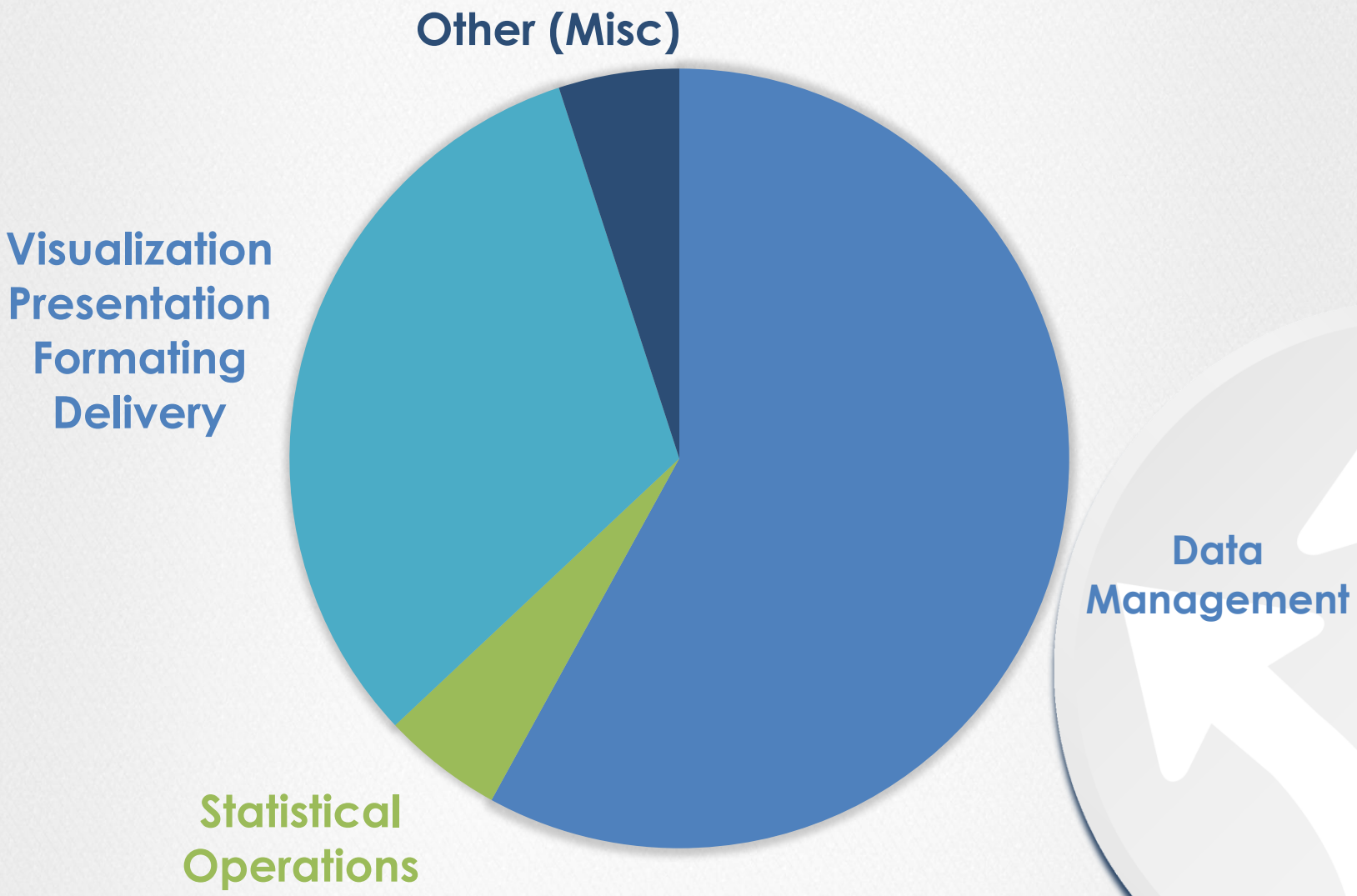
Ref. http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages


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

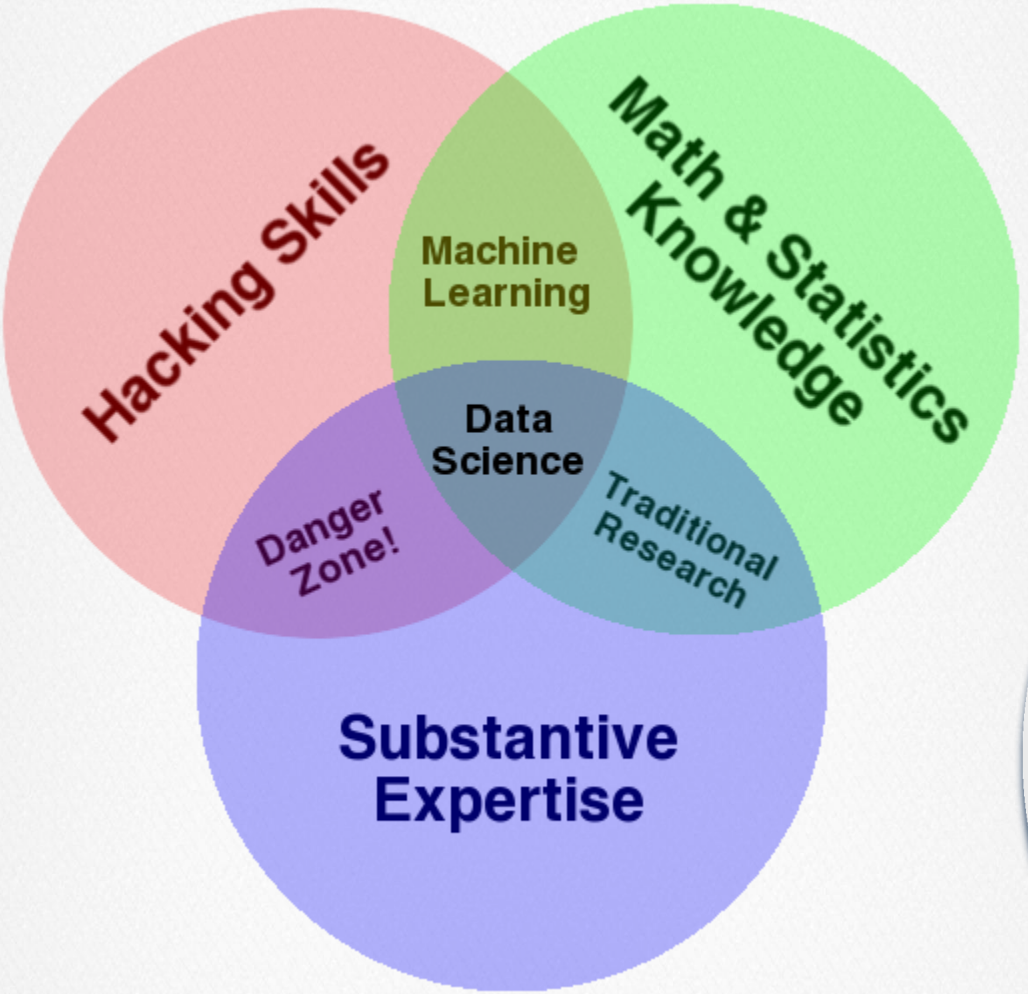
R Limitations

- Slow
- In-memory
- Not-scalable

What about …

DATA SCIENTIST OUTLOOK 2015

Data Science Venn Diagram



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

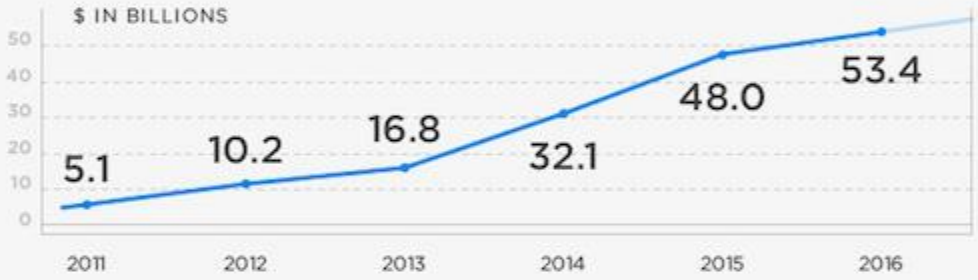
Source: http://venturebeat.com/2013/11/11/data-scientists-needed/

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?

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 …