

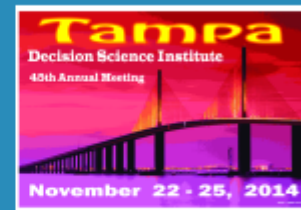
# Designing Analytics Courses with Student Success in Mind

**Matthew A. Lanham** (lanham@vt.edu)  
Doctoral Candidate/Merchandise Data Scientist  
MatthewALanham.com

Virginia Tech  
Pamplin College of Business  
Department of Business Information Technology



2014 Annual Meeting  
of the  
Decision Sciences Institute



Technology and the Rapidly Changing Global Business Landscape

# Outline

## State of Business Analytics

- What is Business Analytics?
- Putting the BA domains together

## Data Science & Big Data Analytics

- What is it?
- BDA projects I'm currently working on
- What can we expect business students to know/learn?

## Analytics Education

- Growth in educational programs
- The Analytics Process
- INFORMS Certified Analytics Professional (CAP)

## Technology Skills

- Skills in demand
- Salaries by skills
- What is R

## Conclusions

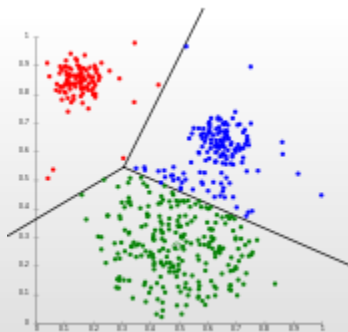
# Business Analytics

## What is Analytics?

- “...the extensive use of data, statistical and quantitative analysis, exploratory and predictive models, and fact-based management to **drive decisions and actions** ([Davenport & Harris, 2007](#)).”
- Refers to the skills, technologies, applications, and practices for continuous iterative exploration and investigation of past business performance to gain insight and **drive business planning** ([Wikipedia, 2014](#)).”

### Descriptive Analytics

What has happened?



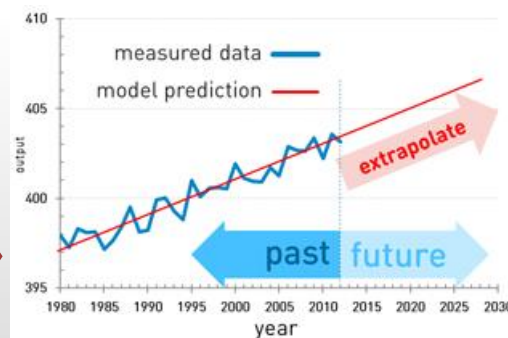
Exploratory Data Analysis (EDA)

Uni- and Multivariate Summaries

Clustering - Segmentation - Profiling

### Predictive Analytics

What will happen?



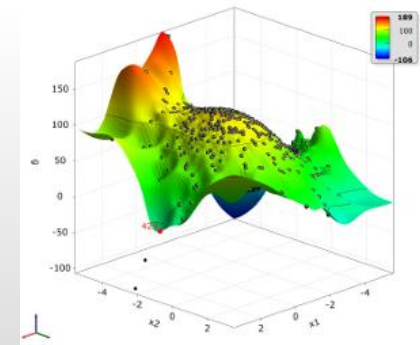
Forecasting/Pattern recognition

Classification

Ensemble Modeling

### Prescriptive Analytics

What is the best course of action?



Optimization/Heuristics

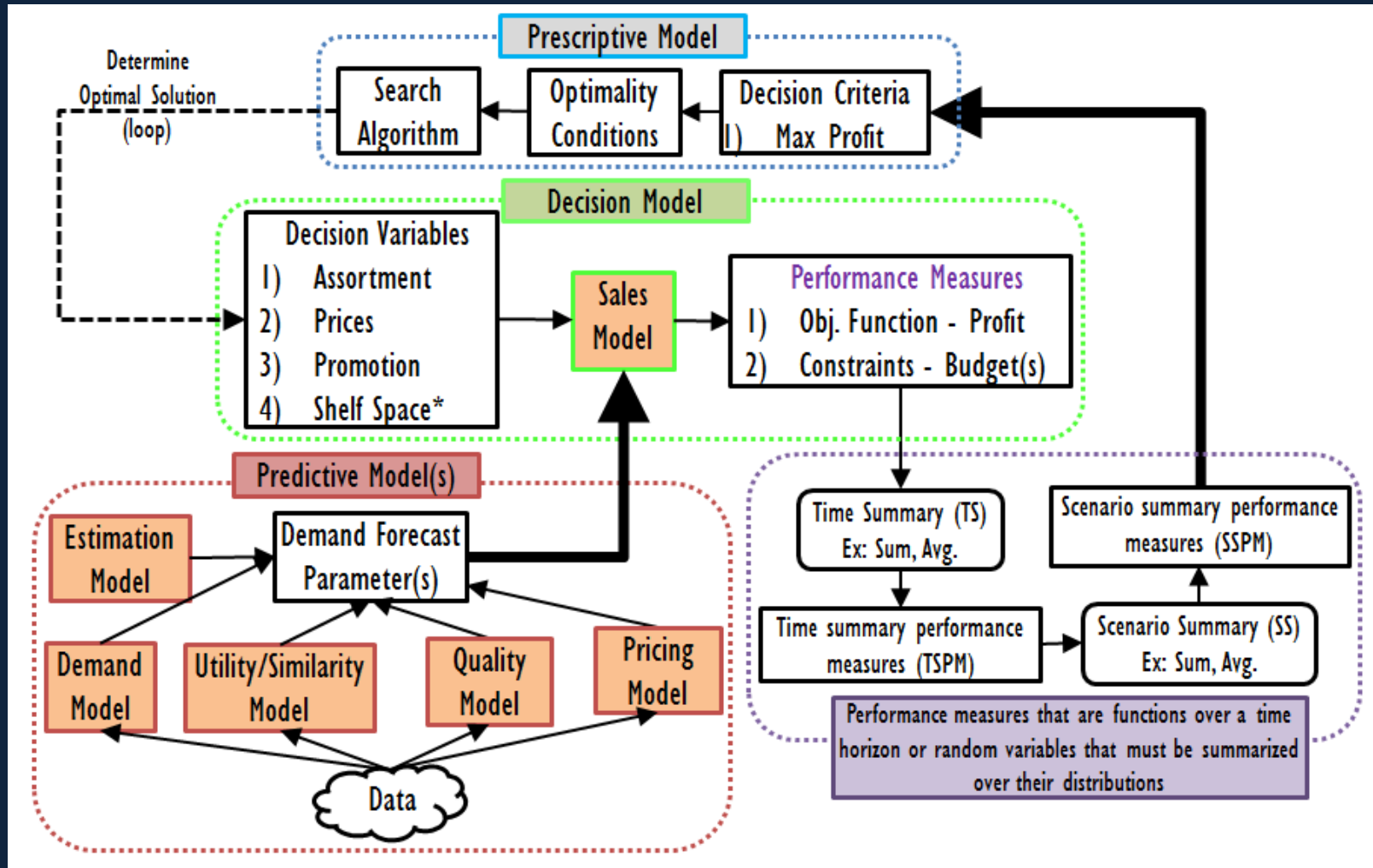
Simulation

Computational Stochastic  
Optimization

# Suggestion I – Tie it all together

## Big Picture Idea

- Example based on my dissertation research and industry work for a Fortune 500 retailer

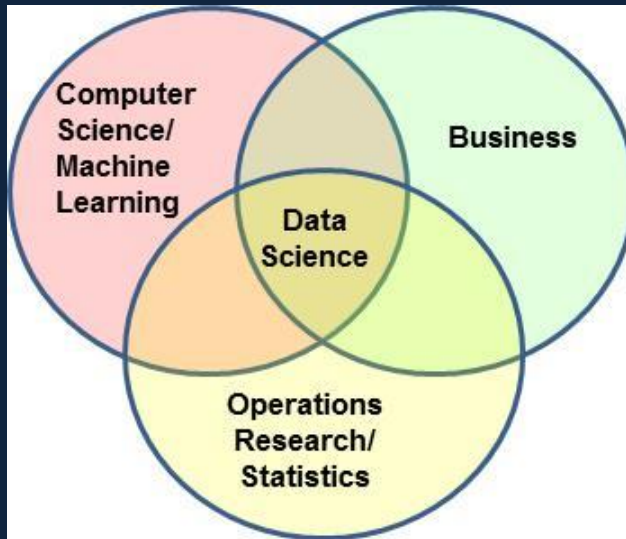


# Data Science/Big Data Analytics

## What is Data Science?

“Data science involves extracting, creating, and processing data to turn it into business value.”

– Vincent Granville, author of “Developing Analytic Talent: Becoming a Data Scientist” and co-founder DataScienceCentral.com



## Data Scientist: *The Sexiest Job of the 21st Century*

### Big Data Analytics (BDA)

- Scaled problems
- 3 Vs, 5 Vs, 10 Vs,... Vs

### Analytics Technologies



### Hybrid



### BDA Technologies



# My Current Big Data Projects

## Variety

- Product reviews & competitor data

★★★★☆ 3.0 8/13/2014

**Decent Battery**

<b>PROS</b>	<b>CONS</b>	<b>BEST USES</b>
Safe To Handle	Lacks Power	Battery Replacement
	Loses Charge	

By **FatPat** from Red Rock, TX

About Me  
Casual Driver

Follow me

VERIFIED BUYER

Level 1

Comments about *AutoCraft Silver Battery, Group Size 91, 615 CCA*:

The product itself isn't terrible, however I feel that for the price of the battery itself, I feel that it should have more power and hold a much better charge.

**BOTTOM LINE** No, I would not recommend this to a friend

Was this review helpful? [Yes](#) / [No](#) - You may also [flag this review](#)

[Comment on this review](#) (earn points)



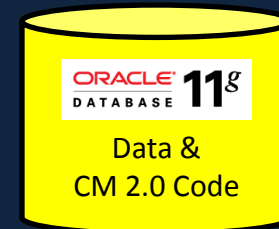
## Volume

- ~ 8 TB of data as of November 1, 2014



## R & SPSS Text Analytics

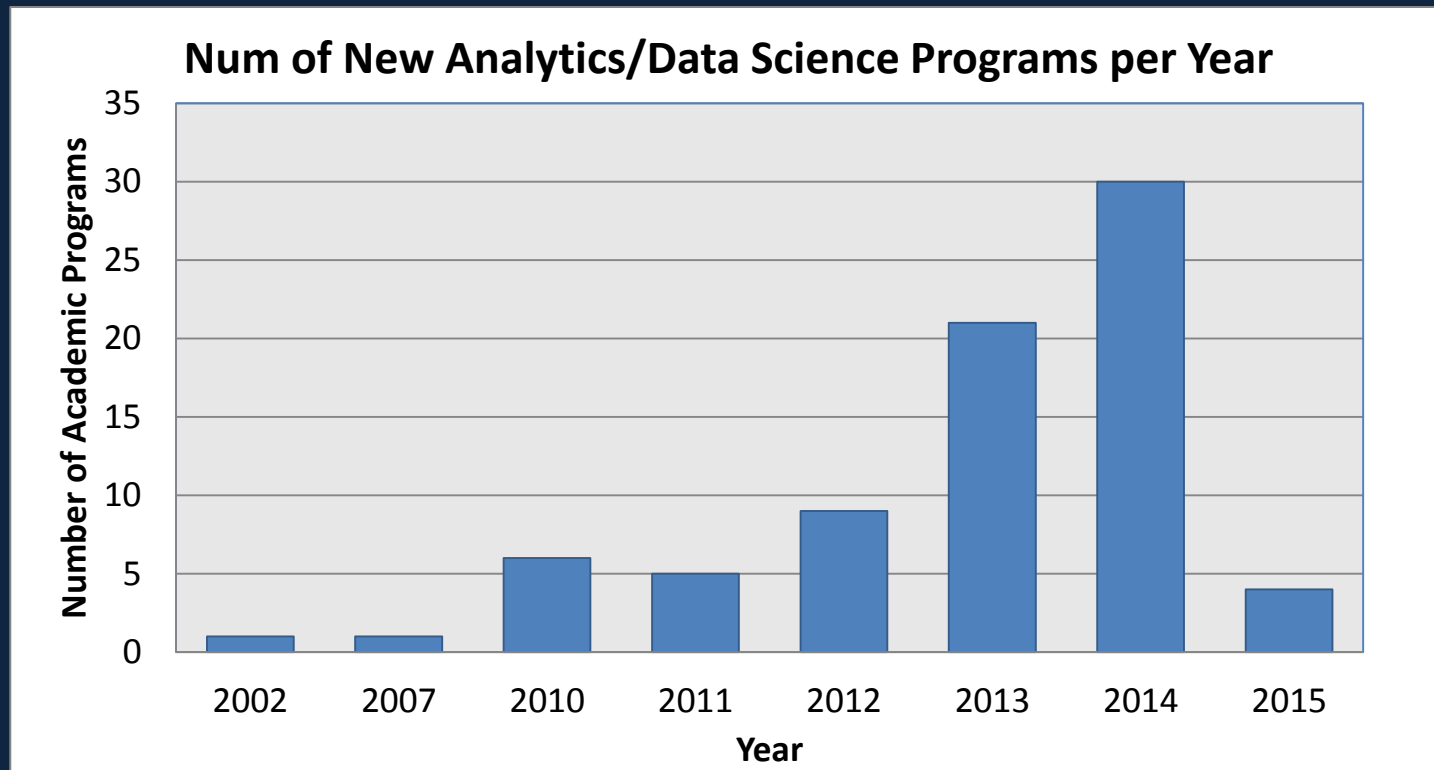
## Oracle w/ PL/SQL



# Analytics Education

## Growth in Business Analytics Programs

- Since 2010, there have been 35 analytics/business analytics programs established in schools of business, and another 29 analytics/data science programs established in other colleges within universities ([Rappa, 2014](#)).

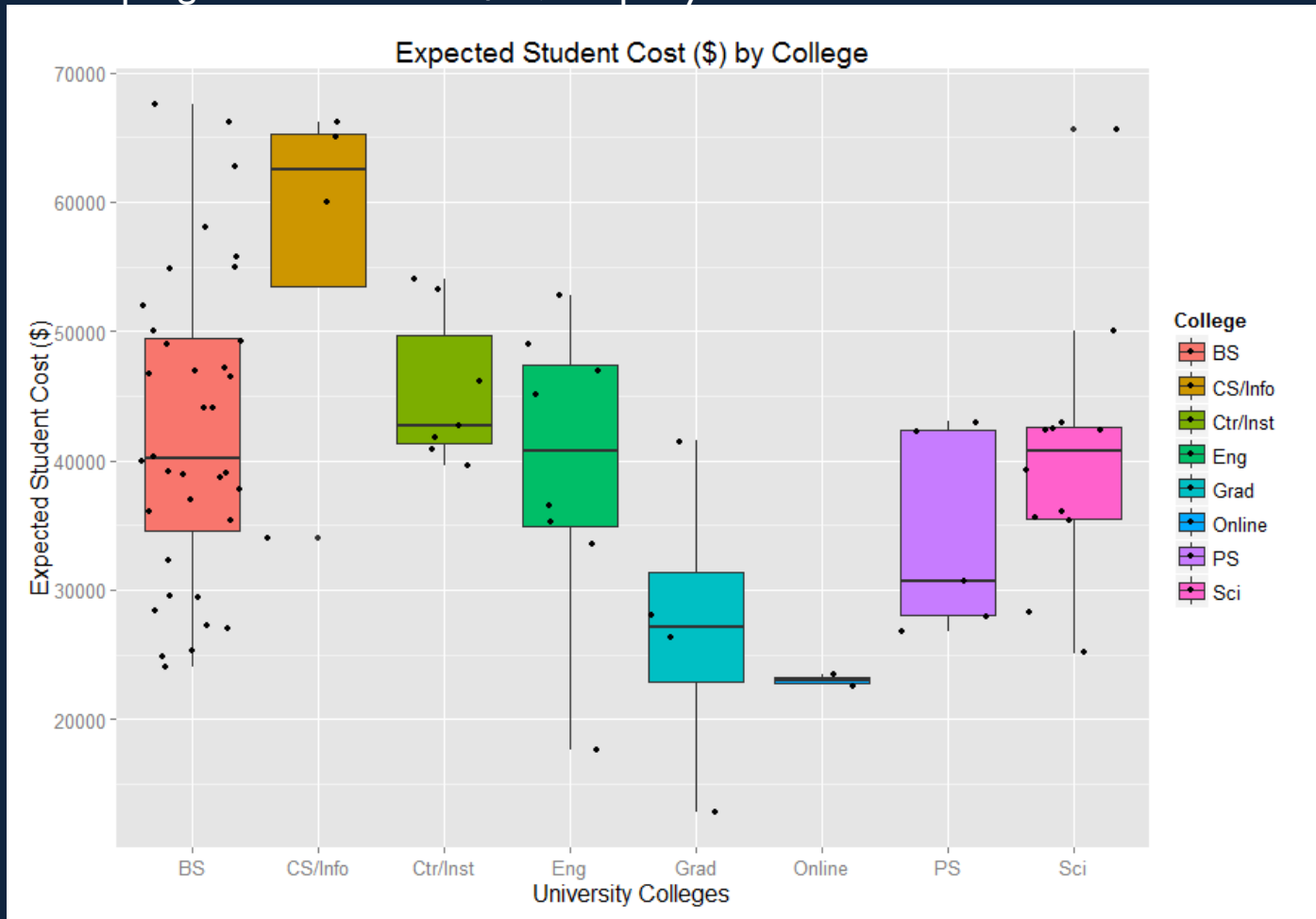


- Note: These numbers do not include the analytics/data science concentrations or tracks added to legacy programs such as engineering, information systems, and statistics.

# Revenue

## These programs are bringing in a lot of money

- Business school programs are around \$40,000 per year



- Program lengths for full-time programs take a student 9 to 16 months to complete or over 36 months for part-time programs



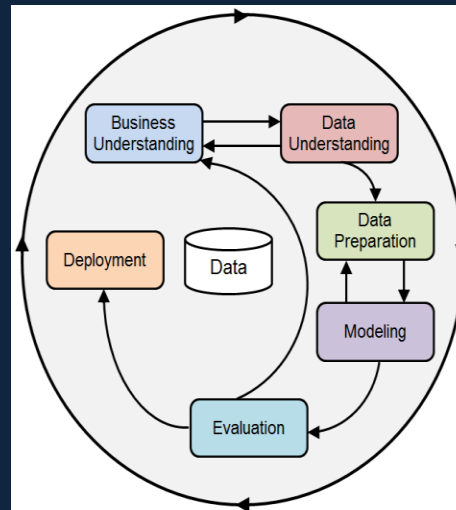
## Suggestion II - Teach the Analytics Process

### Everybody seems to focus on the methods

- Check out INFORMS Certified Analytics Professional (CAP) exam

	INFORMS 7 CAP Domain Areas	Approximate Weight
I.	Business Problem (Question) Framing	12-18%
II.	Analytics Platform Framing	12-20%
III.	Data	18-26%
IV.	Methodology (Approach) Selection	12-18%
V.	Model Building	13-19%
VI.	Deployment	7-11%
VII.	Model Life Cycle Management	4-8%

- Similar to CRISP-DM





# Understanding and Discussing the Problem

## Business Problem (Question) Framing

- Try to make the assignments or projects big picture focused
- Taught well in TQM, Software Engineering, and Project Mgmt. courses

INFORMS 7 CAP Domain Areas		Description
I.	<b>Business Problem (Question) Framing</b> T1 – Obtain or receive problem statement and usability requirements T2 – Identify Stakeholders T3 – Determine whether the problem is amenable to an analytics solution T4 – Refine the problem and delineate constraints T5 – Define the initial set of business benefits T6 – Obtain stakeholder agreement on the business problem statement	The ability to understand a business problem and determine whether the problem is amenable to an analytic solution

<b>Objectives 1/2</b> <b>Who</b> <b>What</b> <b>Where</b> <b>When</b> <b>Why</b>	<b>Requirements Elicitation Techniques</b> <b>Stakeholders</b>	<b>Objective 3</b> <b>Can this problem be modeled?</b> <b>Decision(s) under firm control?</b>	<b>Data available or obtainable?</b>	<b>Objective 4/5</b> <b>Define problem precisely</b> <b>Decisions, Parameters, Objectives, Constraints</b>  <b>Business KPIs/Benefits</b>	<b>Objective 6</b> <b>Achieve Stakeholder Agreement</b> 
---	---	---	--------------------------------------	--	--

- I spend a lot of time here and revisit this domain area regularly to make sure we are creating a solution for the real problem(s)
- Recently interviewed several graduates from statistics and computer science to work with me and most had a difficult time with this

# The CAP Handbook is Free Online

## The others..

- <https://www.informs.org/Certification-Continuing-Ed/Analytics-Certification/Candidate-Handbook>

II.	<b>Analytics Platform Framing</b> T1 – Reformulate problem statement as an analytics problem T2 – Develop a proposed set of drivers and relationships to outputs T3 – State the key set of assumptions related to the problem T4 – Define key metrics of success T5 – Obtain stakeholder agreement on the approach	The ability to reformulate a business problem into an analytics problem with a potential analytics solution
VI.	<b>Deployment</b> T1 – Perform business validation of the model T2 – Deliver report with findings; or T3 – Create model, usability, and system requirements for production T4 – Deliver production model/system* T5 – Support deployment	The ability to deploy the selected model to help solve the business problem
VII.	<b>Model Life Cycle Management</b> T1 – Document initial structure T2 – Track model quality T3 – Recalibrate and maintain the model* T4 – Support training activities T5 – Evaluate the business benefit of the model over time	The ability to manage the model life cycle to evaluate business benefits of the model over time

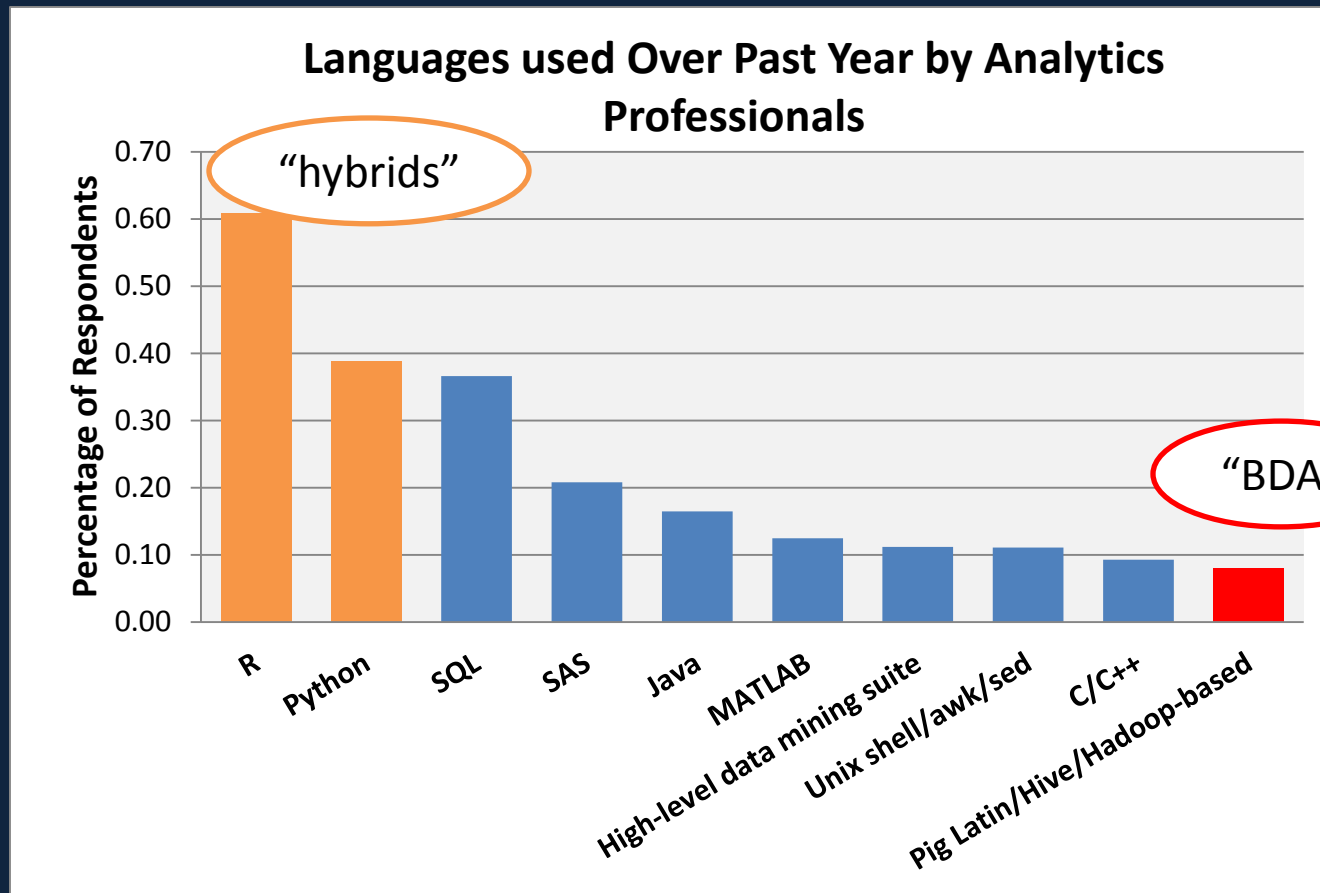
## Recommendation

- Try to incorporate these somehow say via a full-blown semester team project

# Technologies in Demand

## 2013 KDNuggets.com Survey

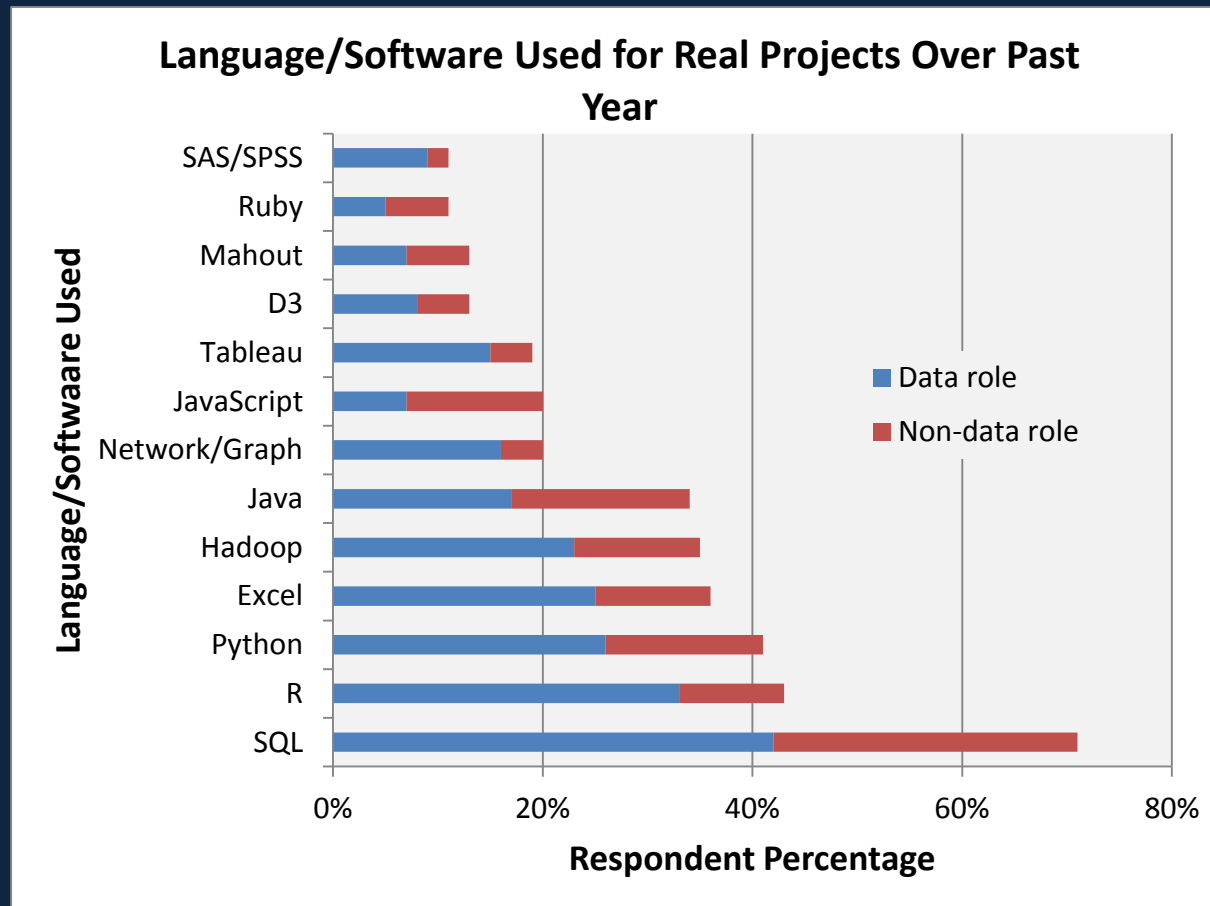
- “What programming/statistics languages you used for an analytics / data mining / data science work during the past year?”
- R (60.9%), Python (38.8%), and SQL (36.6%) were the top three languages used by practitioners.



# Technologies in Demand

## O'Reilly's 2013 Data Science Salary Survey

- “What commercial languages and software have you used for an analytics, big data, data mining, or data science during the past 12 months for a real project”

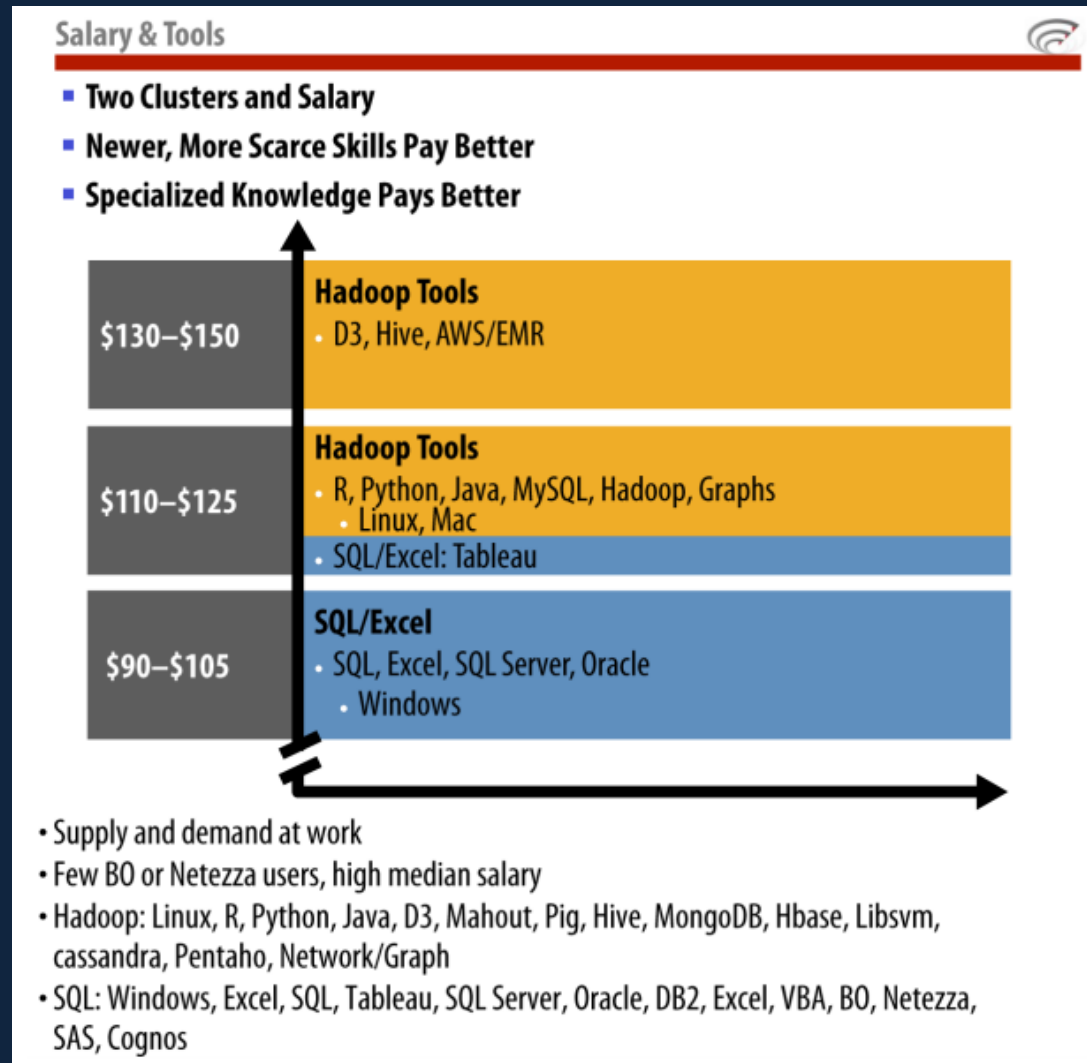


- Similar to KDnuggets annual survey on analytics software, SQL, R, and Python were the top three most popular[source: O'Reilly]

# Technologies vs. Salaries

## Payday

- Data scientists using the hybrid and BDA (open-source) technologies are getting paid more



Source: O'Reilly's 2013 Data Science Salary Survey  
<http://www.oreilly.com/data/free/files/stratasurvey.pdf>

## Suggestion III – Use R

### What is R?

- R was developed by New Zealand Professors Robert Gentleman and Ross Ihaka who wanted a better statistical software platform for their students ([Ihaka & Gentleman, 1996](#)).
- It's more than just statistics today!!

### Can my students use it?

- R is an open-source (i.e. FREE!!) and freely accessible software language under the *GNU General Public License, version 2* ([Free Software Foundation, 1991](#))
- R works with Windows, Macintosh, Unix, and Linux operating systems.
- It has a nice balance of object-oriented and functional programming constructs, and unlike most commercial software, the majority of packages contain many knobs to allow for tuning and customization of a procedure ([Hornick & Plunkett, 2013](#)).

### What makes it better than the others?

- As of 2014 there were 5800 available user-developed packages (also referred to as libraries) ([Cortez](#)).
- There are 72 different packages offering libraries that have functions to do nearly any machine learning methodology. **You will not find many of these techniques in the commercial packages.**
- There is also a growing community of research on prescriptive (optimization) analytics ([Cortez](#)). As of 2014, there are more than 60 available packages for optimization and mathematical programming ([Theussl, 2014](#))

**Descriptive Analytics**

What has happened?

**Predictive Analytics**

What will happen?

**Prescriptive Analytics**

What is the best course of action?

# Who is using R?

## TIOBE Programming Community index

- an indicator of the popularity of programming languages.
- *“The TIOBE index lists various of these statistical programming languages available, e.g. Julia (position #126), LabView (#63), Mathematica (#80), MATLAB (#24), S (#84), SAS (#21), SPSS (#104) and Stata (#110). Most of these languages are getting more popular every month. The clear winner of the pack is the open source programming language R. This month it jumped to position 12, while being at position 15 last month.”*

Nov 2014	Nov 2013	Change	Programming Language	Ratings	Change
1	1		C	17.469%	-0.69%
2	2		Java	14.391%	-2.13%
3	3		Objective-C	9.063%	-0.34%
4	4		C++	6.098%	-2.27%
5	5		C#	4.985%	-1.04%
6	6		PHP	3.043%	-2.34%
7	8	▲	Python	2.589%	-0.52%
8	10	▲	JavaScript	2.088%	+0.04%
9	12	▲	Perl	2.073%	+0.55%
10	11	▲	Visual Basic .NET	2.061%	+0.09%
11	-	▲▲	Visual Basic	1.657%	+1.66%
12	31	▲▲	R	1.548%	+1.14%
13	9	▼	Transact-SQL	1.408%	-1.11%

Source: TIOBE Index for November 2014

<http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html>



## Couple Examples

### Ford Motor Company

- Mike Cavaretta, Ford Motor Company's Chief Data Scientist, and was tasked by the incoming CEO Alan Mulally to help change the culture so that "important decisions within the company had to be based on data"
  - *"On the statistical side, we did a lot of stuff in R. ... We've done a lot more with R and we're currently evaluating Pentaho. So we've really moved from more point solutions for solving particular problems, to more of a framework and understanding different needs in different areas. For example, there may be certain times when SAS is great for analysis because we already have implementations, and it's easier to get that into production. There are other times when R is a better choice because it's got certain packages that makes that analysis a lot easier, so we're working on trying to put all that together (Dataconomy, 2014)"*



### Retailers

- I'm giving a poster presentation at the 2015 INFORMS Computing Society titled "Evaluating Open-Source vs. Commercial Predictive Analytics Software"



# Pros and Cons of R for Analytics

## Pros

1. Many resources
  - <http://www.r-bloggers.com/>
  - Springer's USE R! series
2. Great graphics capabilities
  - ggplot2 (not Tableau or SAS Visual Analytics, but pretty nice)
3. Descriptive-Predictive-Prescriptive-BDA

## Cons

1. Memory
  - R works in memory so if you're working with large data sets (>5 gb) you'll probably need more than 8 GB of RAM
  - I have 24 GB of RAM and rarely have issues
2. Package Quality
  - ~85 percent of the packages I use are high quality
  - Some authors code inefficiently which may cause the user to run out of memory or increase run time even on small data sets

## Don't you get what you pay for?

- Almost always, but R truly is a great deal

# Concluding Remarks



## Big Picture Thinking

- Students need practice in all seven analytics domain areas
  - Not just data analysis and model building

## End-to-end Project & Report

- Make sure they tie descriptive-predictive-prescriptive analytics together
- Students need practice implementing a full-project on their own



## Robust Technologies

- Students need experience with technologies that are in demand and can be used anywhere they choose to work
  - We recommend the R programming language

