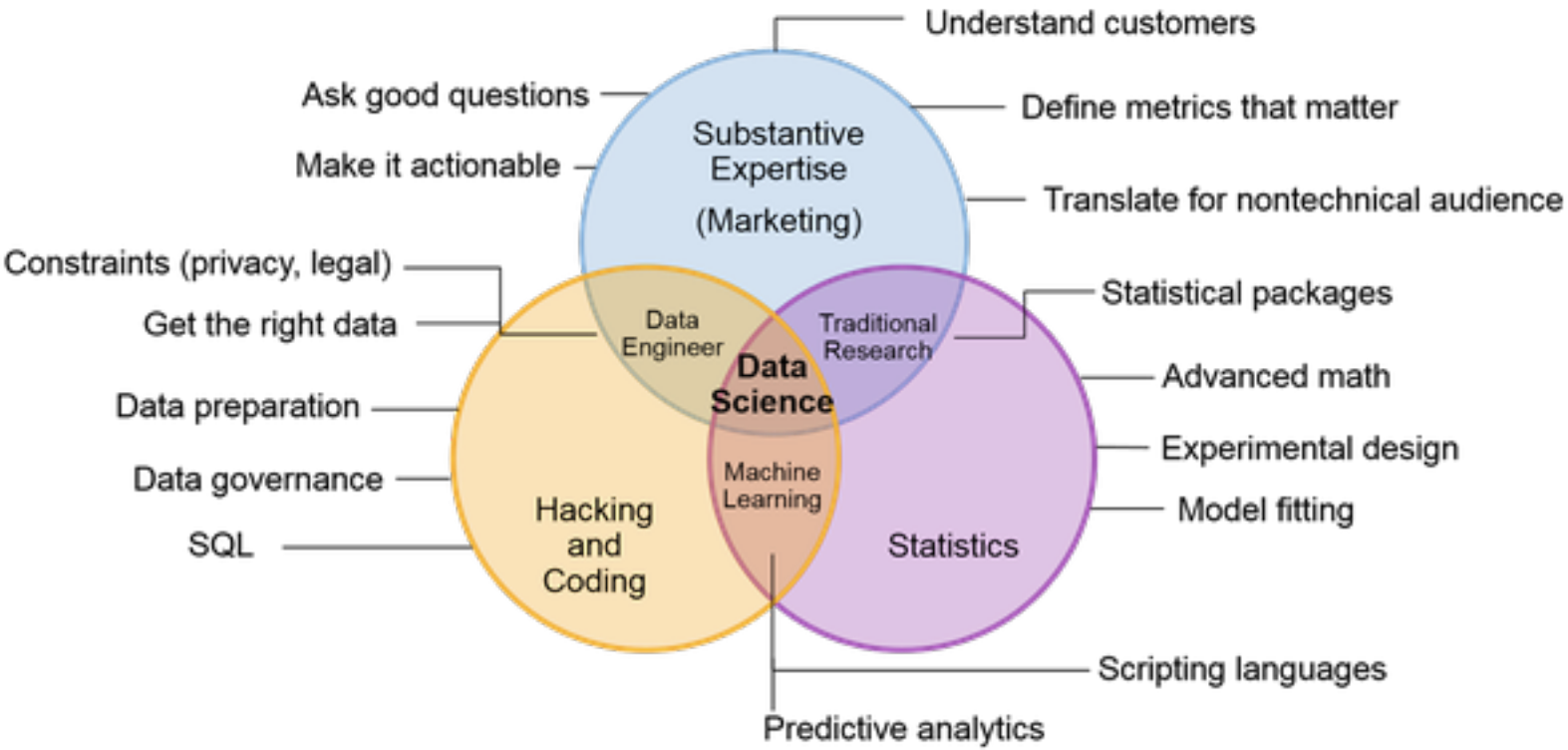
Getting Started with Data Science Using Python

About The Speaker

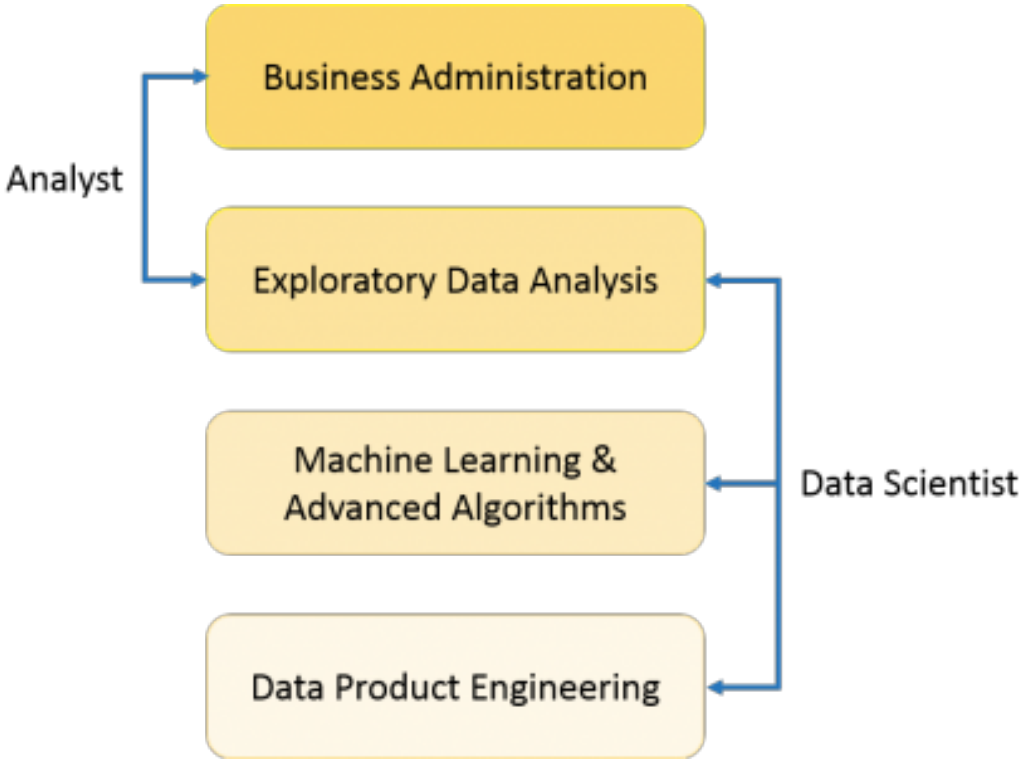
https://www.linkedin.com/in/andreilyskov/

https://www.quora.com/profile/Andrei-Lyskov-1

What is Data Science?



What is Data Science?



What is Data Science?

Typical job duties for data scientists

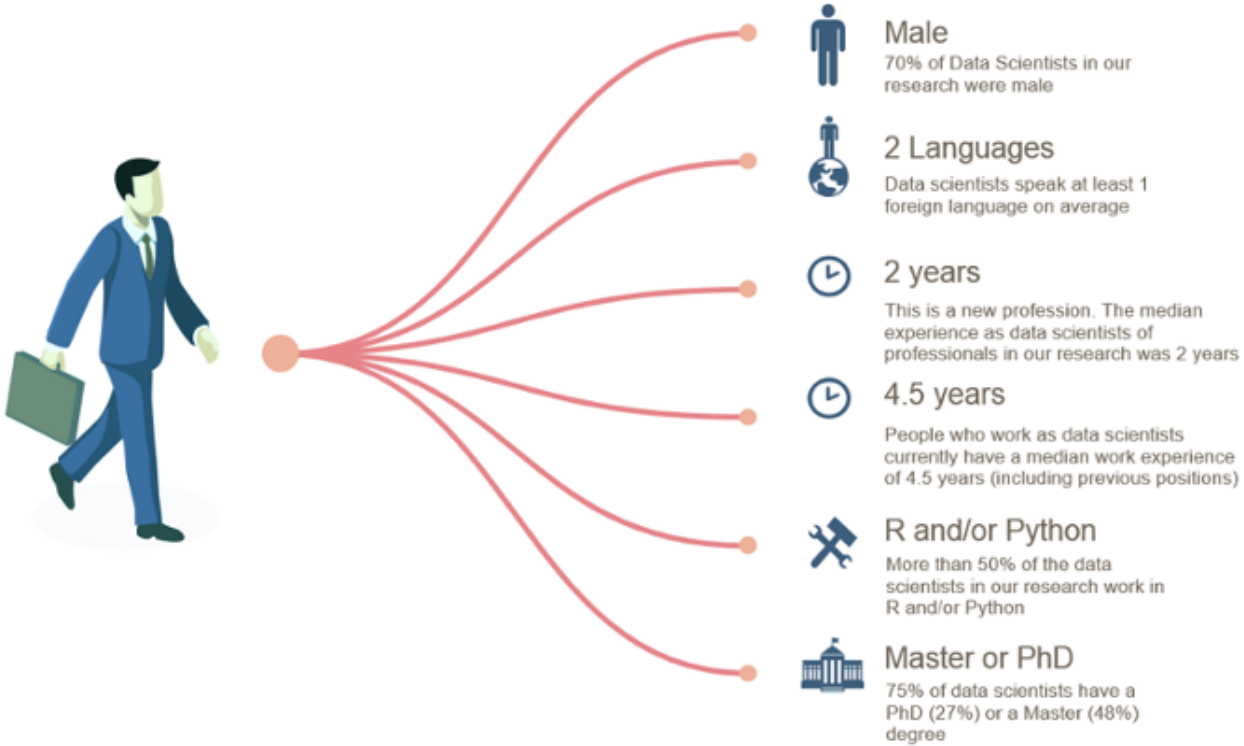
There's not a definitive job description when it comes to a data scientist role. But here are a few things you'll likely be doing:

- Collecting large amounts of unruly data and transforming it into a more usable format.
- Solving business-related problems using data-driven techniques.
- Working with a variety of programming languages, including SAS, R and Python.
- Having a solid grasp of statistics, including statistical tests and distributions.
- Staying on top of analytical techniques such as machine learning, deep learning and text analytics.
- Communicating and collaborating with both IT and business.
- Looking for order and patterns in data, as well as spotting trends that can help a business's bottom line.

Why the Hype Around Data Science?

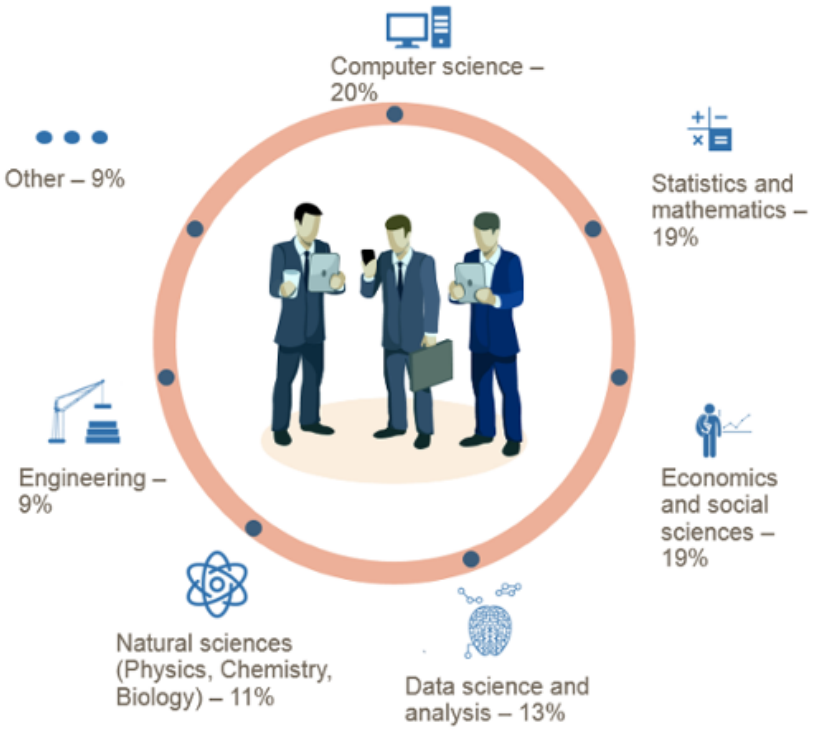
- IBM predicts that demand for data scientists will soar by 28% by 2020
- Data scientist roles have grown over 650% since 2012, but currently, 35,000 people in the US have data science skills, while hundreds of companies are hiring for those roles.
- **Software engineering is a common starting point for professionals who are in the top five fasting growing jobs today.** The career path to Machine Learning Engineer and Big Data Developer begins with a solid software engineering background.
- Data Science gives you career flexibility

Who Are Data Scientists?



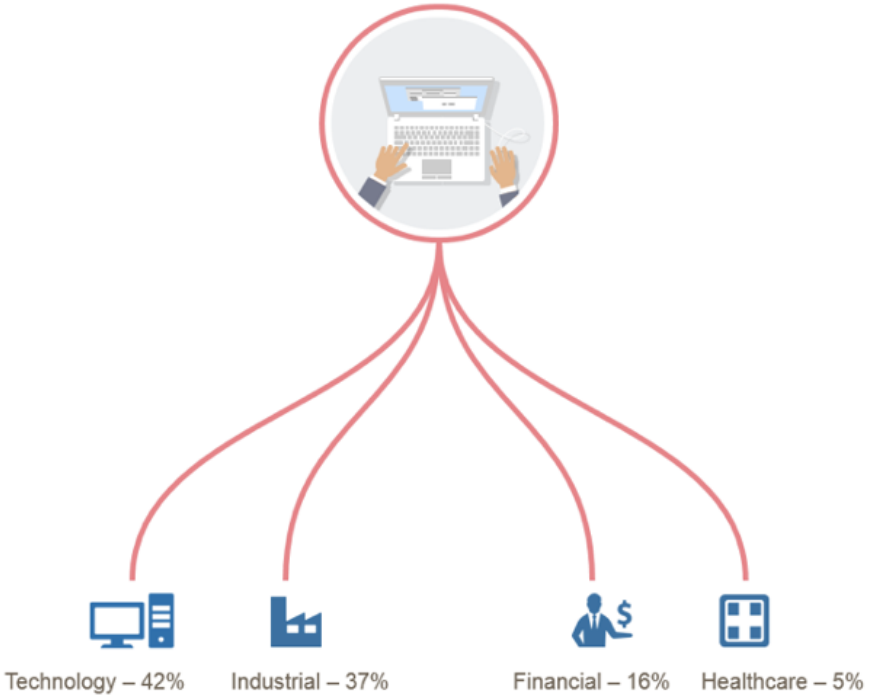
https://www.datascience.com/blog/data-scientist-skills

Who Are Data Scientists?



https://www.datascience.com/blog/data-scientist-skills

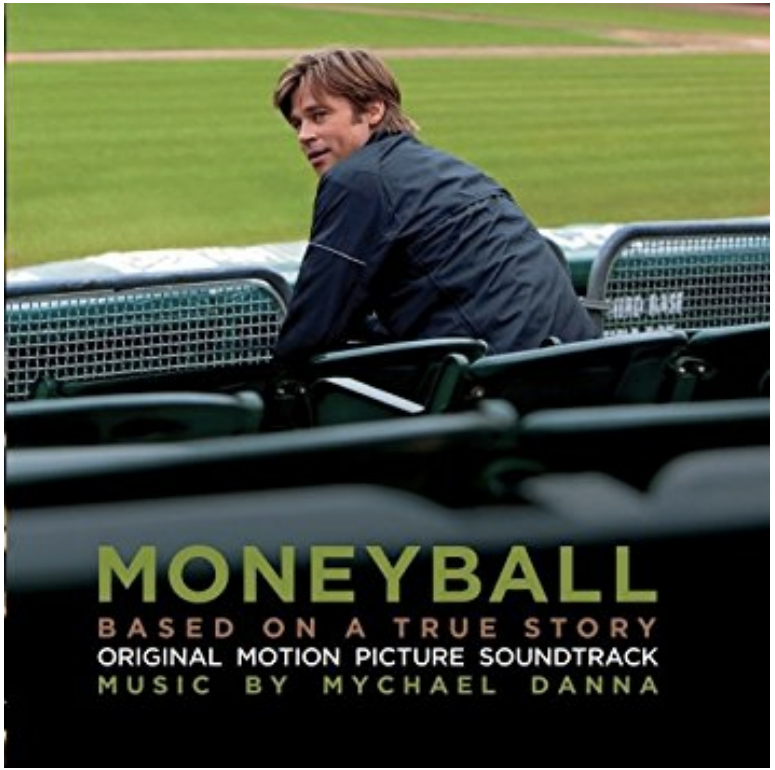
Who Are Data Scientists?



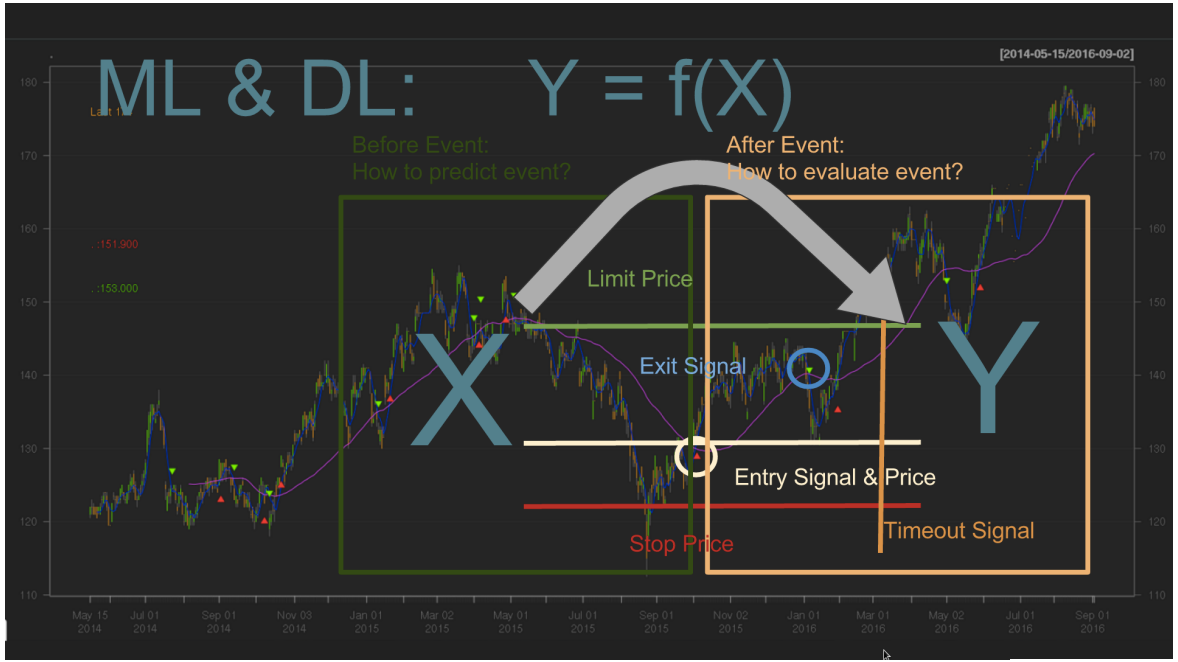
https://www.datascience.com/blog/data-scientist-skills

Application - Security

Application – Sports



Application - Finance



Renaissance

Application – Microsoft (Skype Product)

- The first is with a product feature called **Skype Translator**. As its implied, Skype uses machine learning to translate a conversation between two people speaking different languages through the use of a third party bot that joins your call.
  Skype Translator – How it Works | Skype Blogs
- The second is to **detect fraudulent Skype Users,** examples range from users who send spammy messages, to credit card and online payment fraud. This is an important application of machine learning as you can imagine, a platform that's riddled with spammers and fraudsters is not one that will likely retain many users.
  Detecting Fraudulent Skype Users via Machine Learning

Learning Data Science With Python - Libraries

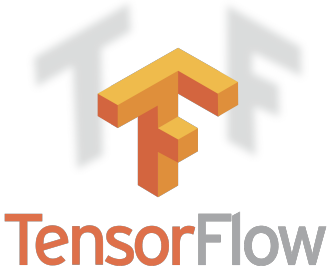
NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

A free software machine learning library that features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, and k-means and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Learning Data Science With Python - Libraries

A plotting library for the Python programming language and its numerical mathematics extension NumPy



TensorFlow is an open-source software library for dataflow programming across a range of tasks. It is a symbolic math library, and is also used for machine learning applications such as neural networks.

Keras is an open source neural network library written in Python. It is capable of running on top of TensorFlow, Microsoft Cognitive Toolkit, Theano, or MXNet. It was developed with a focus on enabling fast experimentation

Learning Data Science With Python - Tools

Open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text
http://jupyter.org/

Similar to Jupyter Notebook, but with the added benefit of “google doc” type sharing and collaboration
https://colab.research.google.com

Crestle

Effortless infrastructure for deep learning

Crestle is your GPU-enabled Jupyter environment in the cloud.
https://www.crestle.com/

Data Science Steps

- **Data Gathering**
  Unless you’re at a company with great data governance you’re likely going to have some trouble accessing the data you want. Whether that's because your company has neglected to put the necessary systems in place to gather data, or the data that they are collecting is fragmented and scattered across the organization, you’ll have to first spend some time gathering whatever data you’ll need to do your job. That means having discussions with relevant stakeholders, and getting the necessary credentials to access databases within your organization.

- **Data Preparation**
  Once you have access to data, you’ll need to spend some time cleaning and formatting it. This is where Data Science can often become more of an art, then a science. Unlike datasets you’ll find in competitions, the real world has very messy data sets. Missing values, error in data collection, data formatting, normalization, outliers - these are all issues that you’ll have to learn to deal with.

Data Science Steps

- **Exploration**
  Before diving into building any models, you'll want to explore the data to try to glean some insights. Clustering algorithms, scatterplots, bar graphs, Chernoff faces are all interesting ways of visualizing data that will lead to a better understanding of the structure of your data and aid you in your model building step.

- **Model Building**
  With your data cleaned and formatted, you'll have an opportunity to explore a variety of models to see which one works best. Random Forests, SVM's, Bayesian Predictors Neural Networks, Deep Learning, K-Nearest Neighbours - all models you should familiarize yourself with. There is no one model fits all, and so you again will need to develop intuition on which model suits your particular problem.

Data Science Steps

- **Model Validation**
  Prediction accuracy is a common benchmark for whether your model is performing well, however often times there are other evaluation metrics to consider. False positives and false negatives are important to think about from the perspective of the problem you're working on. If you're predicting disease, you'll care more about minimizing false negative, since it may result in a persons death - whereas a false positive will only lead to additional testing.

- **Model Deployment**
  Finally you'll deploy your model into the wild, as you gather more data and feedback on how its doing you'll be able to tweak and improve it as time goes on.

  **This is by no means a comprehensive list of steps, and there are certainly other things you'll need to do to be able to do well in your job - however this is a good high level overview of the steps involved in tackling data science problems.*

Case Study

Building a regression model to predict housing prices

Building A Portfolio

1. 

React

2. kaggle

3. Quora

Medium

Building A Portfolio

4.

5.

open source initiative

Pandas

scikits learn

Questions?

Resources

Podcasts	Websites/Blogs	Communities
Data Skeptic	Dataquest.io	experfylabs.slack.com/messages/C0L736X36/
Data Stories	Kaggle.com	opendatacommunity.slack.com
Learning Machines 101	Quora.com	dcommunity.slack.com
Linear Digressions	analyticsvidhya.com	kagglenoobs.slack.com
O'Reilly Data Show	Coursera.org	pythondev.slack.com
Talking Machine	https://developers.google.com/machine-learning/crash-course/	
This week in Machine Learning and AI	https://portal.azure.com/	
Siraj Raval (Youtube)	https://www.luis.ai/	

Resources

Books	Tv Shows/documentaries
Hands-On Machine Learning with Scikit-Learn and TensorFlow	Humans (2015-)
Python Machine Learning, 1st Edition	Persons of interest
Everybody Lies: Big Data, New Data	Intelligence
Weapons of Math Destruction	Minority report
Big Data: A Revolution That Will Transform How We Live, Work, and Think	Almost human
Turing: Pioneer of the Information Age	Robot and frank
Avogadro Corp	Her
Code: The Hidden Language of Computer Hardware and Software	Black Mirror
Superintelligence: Paths, Dangers, Strategies	iRobot
Visual Explanations: Images and Quantities, Evidence and Narrative	Ex Machina
Pattern Recognition and Machine Learning (Information Science and Statistics)	The Secret Rules of Modern Living: Algorithms
Storytelling with Data: A Data Visualization Guide for Business Professionals	
An Introduction to Statistical Learning by James, Witten, Hastie and Tibshirani	