GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 1

NOVEMBER 19, 2014

DAT11-SF

WELCOME!

Instructors: Leonid Zhukov, Ramesh Sampath, Chad Hokama,

E-mails: leonid.e.zhukov@gmail.com

ramesh@sampathweb.com

chadhokama@gmail.com,

Course Website: http://www.schoology.com/

GitHub: https://github.com/ga-students/DAT_SF_11

Course Times: 6:30pm-9:30pm, Mondays and Wednesdays

Office Hours

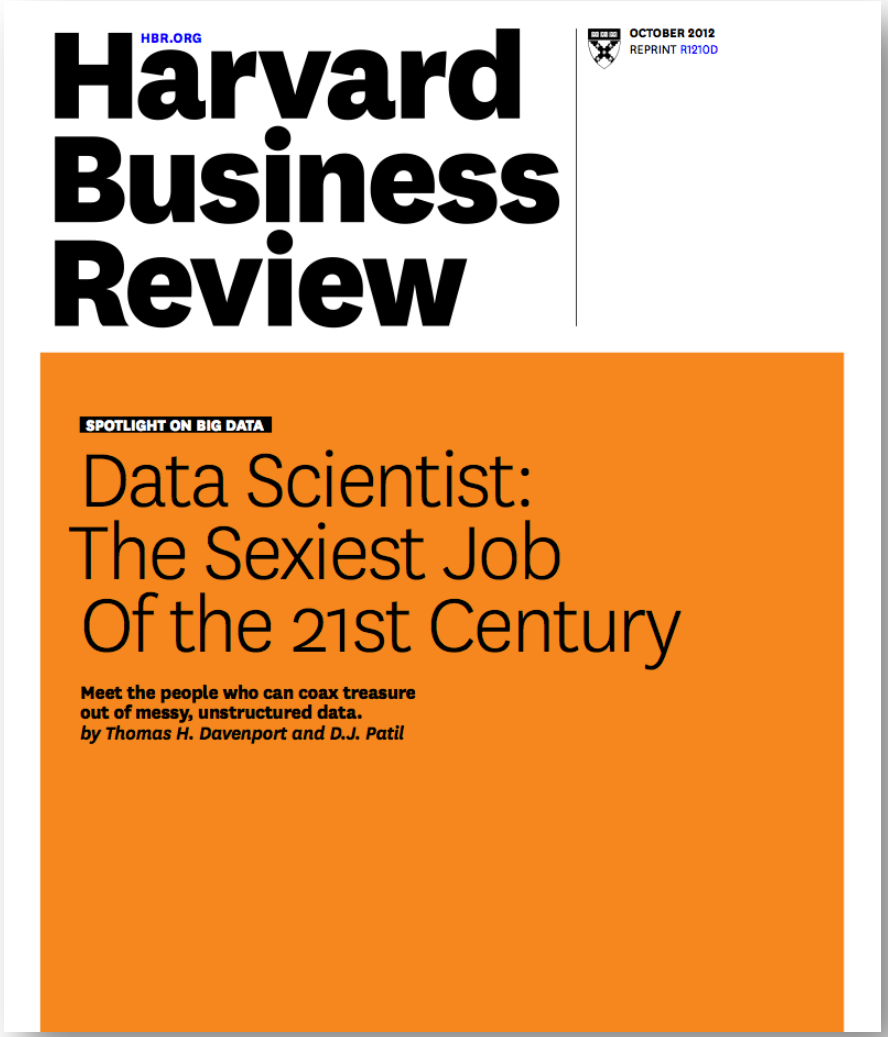
I. WHAT IS DATA SCIENCE?
II. THE DATA SCIENCE WORKFLOW

LAB:
III. UNIX COMMAND LINE
IV. PYTHON TUTORIAL

I. WHAT IS DATA SCIENCE?

- A set of tools and techniques used to extract useful information from data.
- An interdisciplinary, problem-solving oriented subject.
- The application of scientific techniques to practical problems.
- A rapidly growing field.

- Data mining
- Statistics
- Machine learning
- Information visualization
- Network analysis
- Natural language processing
- Algorithms
- Software engineering
- Databases
- Distributed systems
- Big data

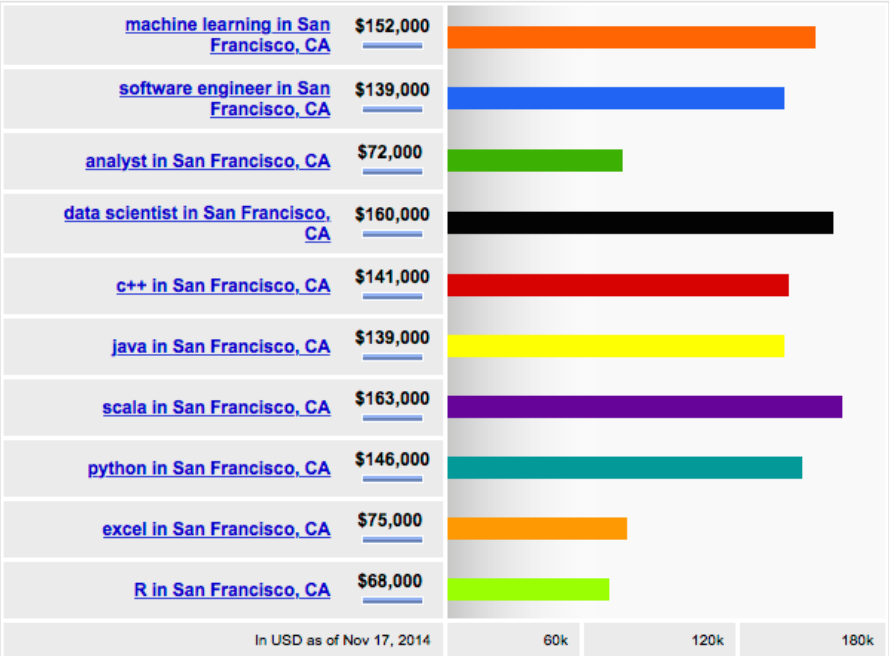


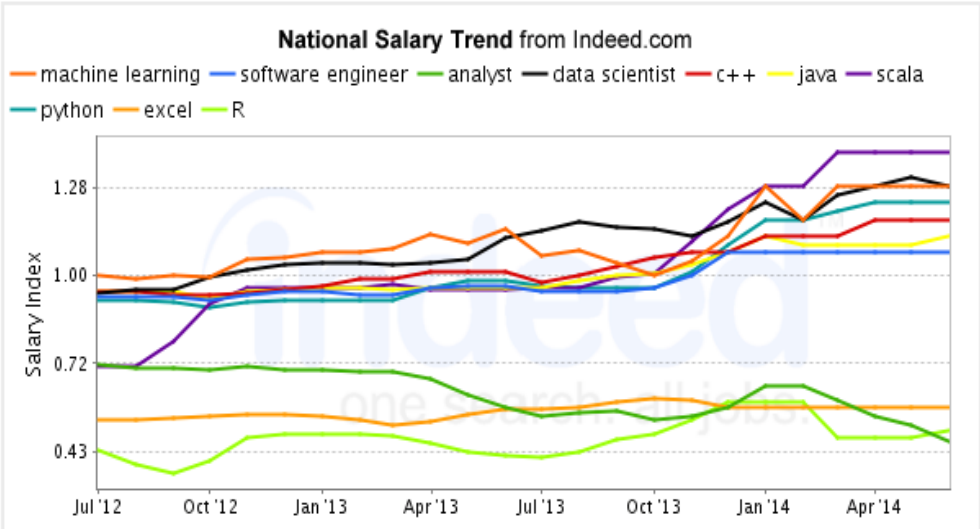
McKinsey estimates 140,000-190,000 shortage by 2018

I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

Hal Varian, Chief Economist at Google, The McKinsey Quarterly, January 2009

Average Salary of Jobs with Titles Matching Your Search





DATA SCIENTISTS WANTED

Data Scientist

Other Details

About this job

Job description

Facebook is seeking a Data Sci
to be comfortable working as a
have a keen interest in the stud
questions that help us build the

Responsibilities

- Work closely with a prod
- Answer product question
- Communicate findings to
- Drive the collection of ne
- Analyze and interpret the
- Develop best practices f
product engineering team

Requirements

- M.S. or Ph.D. in a releva
- Extensive experience so
- Comfort manipulating an
sources
- A strong passion for emp
- A flexible analytic approa
- Ability to communicate c
- Fluency with at least one
- Familiarity with relational
- Expert knowledge of an a
- Experience working with
plus (Map/Reduce, Hado

Data Scientist

Other Details

About this job

Job description

Job Description

PRINCIPAL DUTIES AND RESI

- Apply standard techniques and
deliver actionable business insig
- Work, under normal supervisio
and develop proposals to respon
- Following directed and specific
development. These include pre
selection, and model developme
- Independently research and te
assessing accuracy/fit/predictive
- Deliver results and presentatio
- Interact with customers to gath
insights. Likely presenting projec
preparation.

SKILLS

- Strong statistical foundation, w
methods.
- Proficiency in at least one of th
- Programming strength in at lea
- Natural ability to communicate
- Natural curiosity to research ar
problems.
- Team-oriented and collaborativ
- Innate customer orientation, wi

Data Scientist at LinkedIn

LinkedIn - Mountain View, CA

Posted 26 days ago

Apply on company website | Save

Other Details

About this job

Job description

Description

As a Senior Data Scientist at LinkedIn, you will develop innovative new technologies, features, and products that help connect the world's professionals to make them more productive and successful.

Our team applies machine learning techniques on social data to build products & features that reach over 200M professionals on LinkedIn. We build graph and text mining systems to tackle hard problems in areas like entity resolution, search relevance, recommendation algorithms, reputation & skills assessment, and network analysis.
Along with our team of data scientists, you'll work with product managers, designers, and engineers to build data driven features and products like LinkedIn Skills, Endorsements, and InMaps.
If you enjoy working with data to build products and solve hard problems in creative ways, you will fit right in.

Requirements:

- Strong background in Machine Learning, Statistics, Information Retrieval, or Graph Analysis
- Some experience working with large datasets, preferably using tools like Hadoop, MapReduce, Pig, or Hive
- 2+ years experience developing high quality software, contributions to open source projects are a plus
- Experience programming in an object oriented language (Java, C++, etc)
- Knowledge of scripting languages like Ruby or Python, familiarity with web frameworks a plus
- Comfortable with data analysis & visualization using tools like R, Matlab, or SciPy
- Critical thinking: ability to track down complex data and engineering issues, evaluate different algorithmic approaches, and analyze data to solve problems
- Creativity: you can conceive of new data driven products, features, and technologies
- Results: you prioritize, focusing on ideas and features that will have significant, measurable impact
- Planning & estimation: ability to set and meet your own project objectives & milestones
- Ability to coordinate effectively with team members in engineering, design, and product management
- Communicate results and progress internally and externally in meetings, presentations, and tech talks
- Masters, Phd, or equivalent experience in a quantitative field (computer science, physics, mathematics, bioinformatics, etc.)

Data Scientist

Apple - Santa Clara Valley - California -US

Posted 19 days ago

Apply on company website | Save

Other Details

About this job

Job description

Apple has a tremendous amount of data, and we have just scratched the surface in pattern detection, anomaly detection, predictive modeling, and optimization. There are many exciting problems to be discovered and solved. We encourage scientists to stay abreast of data mining research by attending conferences and working with academic faculty and students. We foster a collaborative work environment, but allow solution autonomy on projects.

The iTunes Engineering team has a proud tradition of delivering cutting-edge products in a competitive marketplace. We seek to maintain a challenging and rewarding environment where the best engineers and scientists can collaborate and produce real-world improvements in customers' online experience. Successful candidates will solve problems unique in scale and concept in the pursuit of new and original features.

Key Qualifications

- Strong working knowledge of data mining algorithms including decision trees, probability networks, association rules, clustering, regression, and neural networks.
- Familiarity with database modeling and data warehousing principles with a working knowledge of SQL.
- Familiarity with Big Data tools and techniques, including MapReduce, NoSQL stores, and unbounded stream processing.
- Creativity to go beyond current tools to deliver best solution to the problem
- Strong programming skills in Java, Python, or similar language
- Excellent interpersonal, written, and verbal communication skills
- Ability and comfort working independently and making key decisions on projects

Description

We are seeking an outstanding data mining scientist who is interested in designing, developing, and fielding data mining solutions that have direct and measurable impact to Apple. This person will work within and across teams to help identify viable data mining opportunities and then implement end to end analytical solutions. The role requires both a broad knowledge of existing data mining algorithms and creativity to invent and customize when necessary.

Education

Ph.D. in Data Mining, Machine Learning, Statistics, Operations Research or related field

M.S. in related field with 5 years experience applying data mining techniques to real business problems.

WHO USES DATA SCIENCE?

NETFLIX

amazon

facebook

Google™

LinkedIn®

FiveThirtyEight

Nate Silver's Political Calculus

Walmart

TARGET

foursquare

PANDORA

Great music discovery is effortless and free with Pandora.

2012

BARACKOBAMA.COM

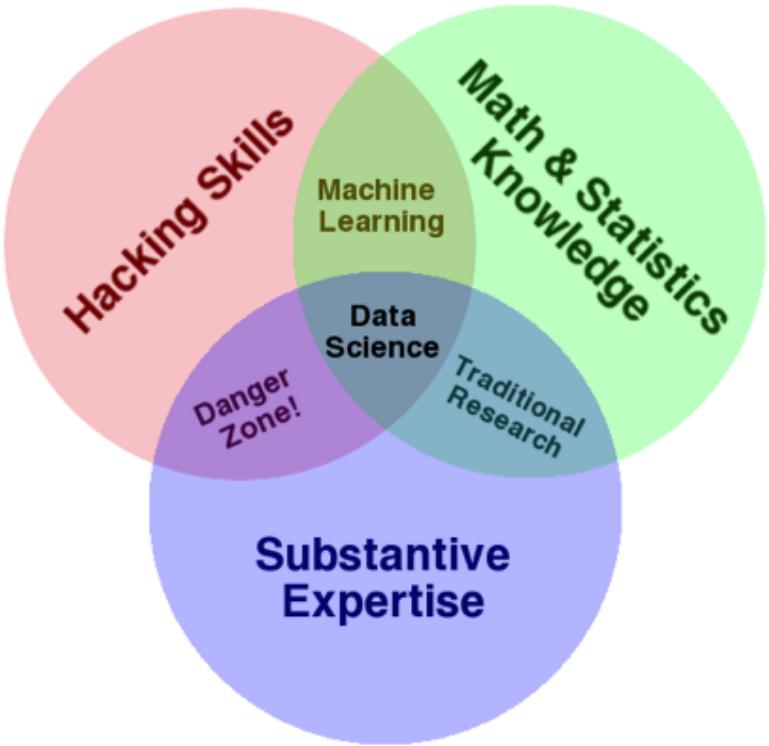
at&t

DATA ANALYSTS / DATA SCIENTIST

			Hacker																					Scripter							Application User					
			Analytics	Biology	Datamart	Finance	Finance	Healthcare	Healthcare	Healthcare	Healthcare	Insurance	Marketing	Marketing	News Aggregator	Retail	Retail	Social Networking	Social Networking	Social Networking	Visualization	Visualization	Web	Web	Analytics	Analytics	Analytics	Finance	Healthcare	Media	Retail	Finance	Insurance	Retail	Retail	Sports	Web Security
Process	Discovery	Locating Data	X	X	X	X	X	X	X	X	X					X		X	X	X	X	X						X	X								
		Field Definitions	X	X	X	X	X	X	X	X	X		X	X		X		X	X	X		X															
	Wrangle	Data Integration	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X				X	X		X			X	X					X	
		Parsing Semi-Structued	X	X	X	X			X			X	X	X		X	X		X	X	X	X	X	X	X		X	X	X	X	X					X	
		Advanced Aggregation and Filtering	X				X		X	X	X				X	X	X		X	X	X	X		X			X				X				X		
	Profile	Data Quality	X	X		X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X	X			X	X
		Verifying Assumptions		X		X	X		X	X		X	X	X		X	X	X	X	X	X	X		X	X		X	X		X	X	X	X				X
	Model	Feature Selection	X		X	X							X	X	X	X	X		X	X	X	X	X	X		X	X				X	X		X	X		
		Scale	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X		X		X	X	X	X	X	X	X
		Advanced Analytics	X		X		X						X	X	X	X	X								X		X		X		X			X	X		
	Report	Communicating Assumptions						X	X			X	X		X	X	X	X	X					X			X	X	X	X	X					X	X
		Static Reports		X	X		X			X	X		X	X		X	X	X				X		X				X	X	X	X					X	
	Workflow	Data Migration	X	X	X	X	X	X		X	X					X		X	X	X		X		X					X		X						
		Operationalizing Workflows		X	X		X	X		X			X	X		X	X	X	X	X				X					X	X	X						
Tools	Database	SQL	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X			X		X					
		Hadoop/Hive/Pig	X		X					X							X		X	X				X													
		MongoDB													X						X																
		CustomDB	X					X	X	X	X													X	X												
	Scripting	Java	X		X		X			X	X	X			X	X	X		X	X	X			X													
		Perl																			X	X															
		Python	X		X	X	X	X	X	X		X	X	X				X	X	X				X			X										
		Clojure													X																						
		Visual Basic		X																												X			X		
	Modeling	R	X		X								X	X			X	X	X	X	X		X	X	X	X	X			X	X						
		Matlab				X						X					X				X		X				X		X								
		SAS	X														X				X	X	X														
		Excel		X		X	X				X	X	X	X		X	X	X					X	X	X			X			X	X	X	X	X	X	X

Enterprise Data Analysis and Visualization: An Interview Study. S. Kandel et al, 2012

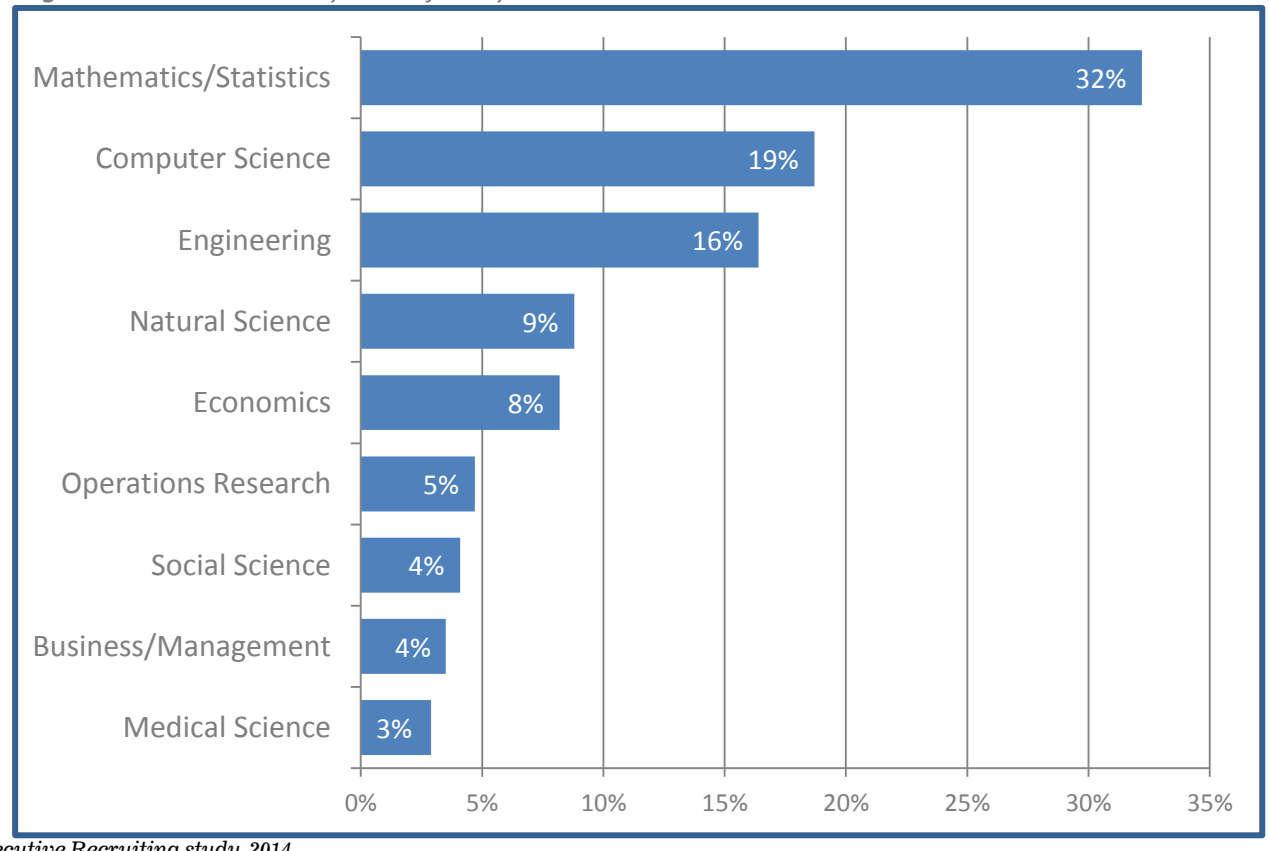
THE QUALITIES OF A DATA SCIENTIST



source: http://www.dataists.com/2010/09/the-data-science-venn-diagram/

- Statistical and machine learning knowledge
- Engineering experience
- Academic curiosity
- Product sense
- Storytelling
- Cleverness

Figure 8. *Data Scientists by Area of Study*



Source: Burtch Works Executive Recruiting study, 2014

Figure 4. *Data Scientists by Years of Experience*

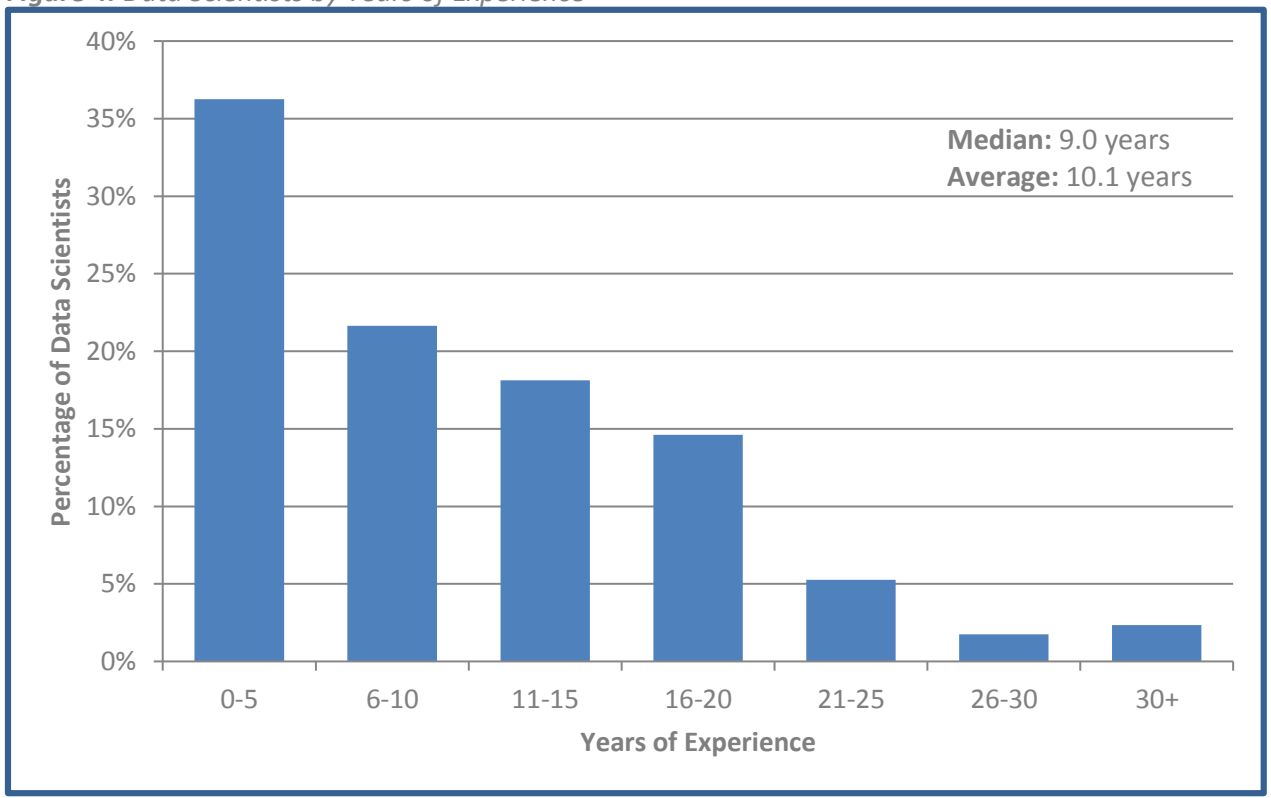


Figure 7. *Data Scientists by Education*

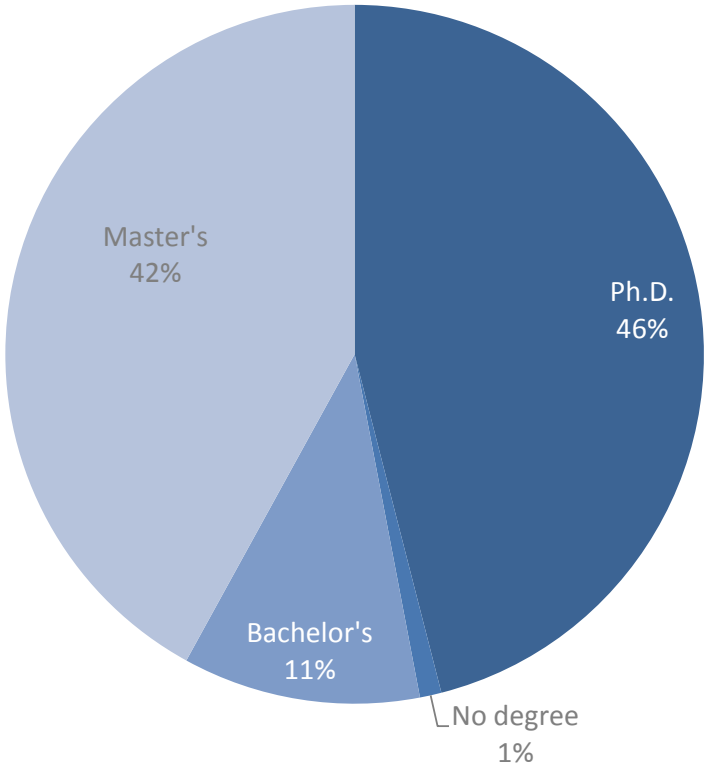
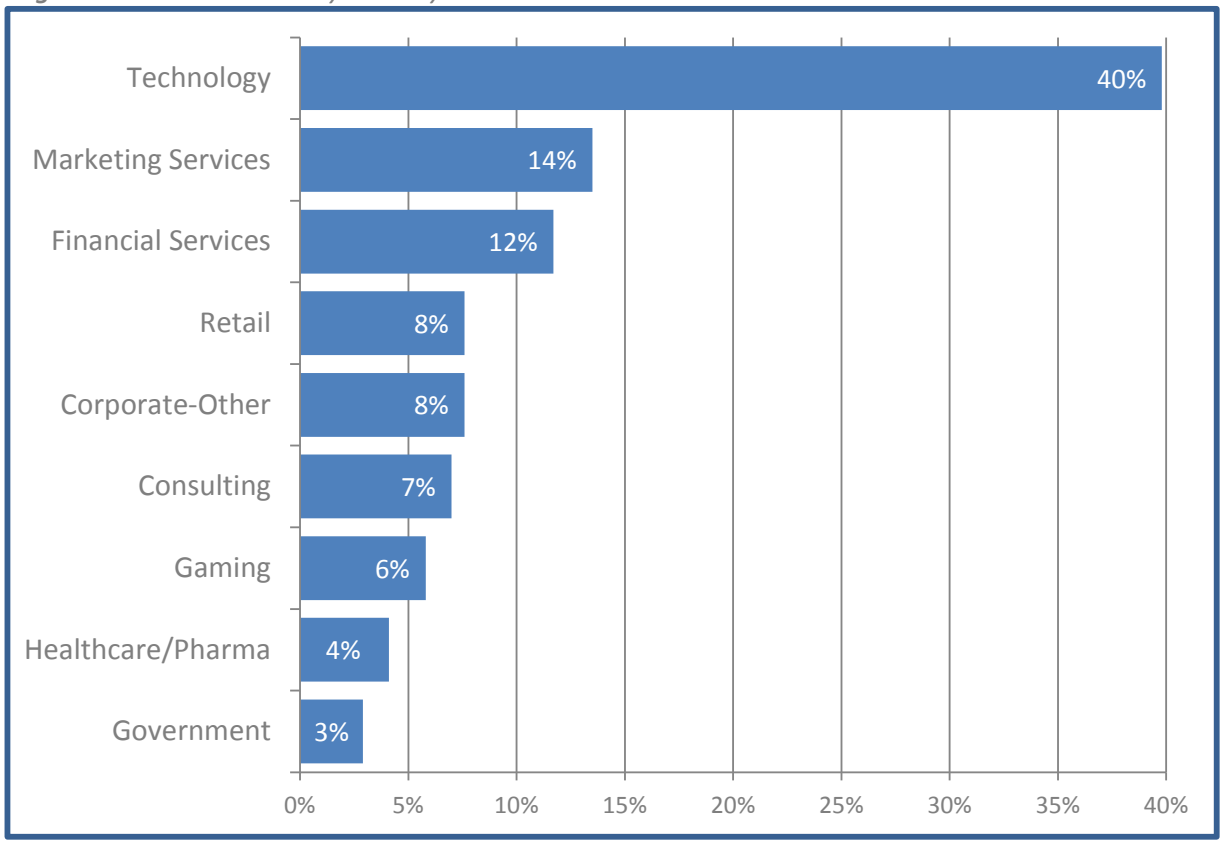
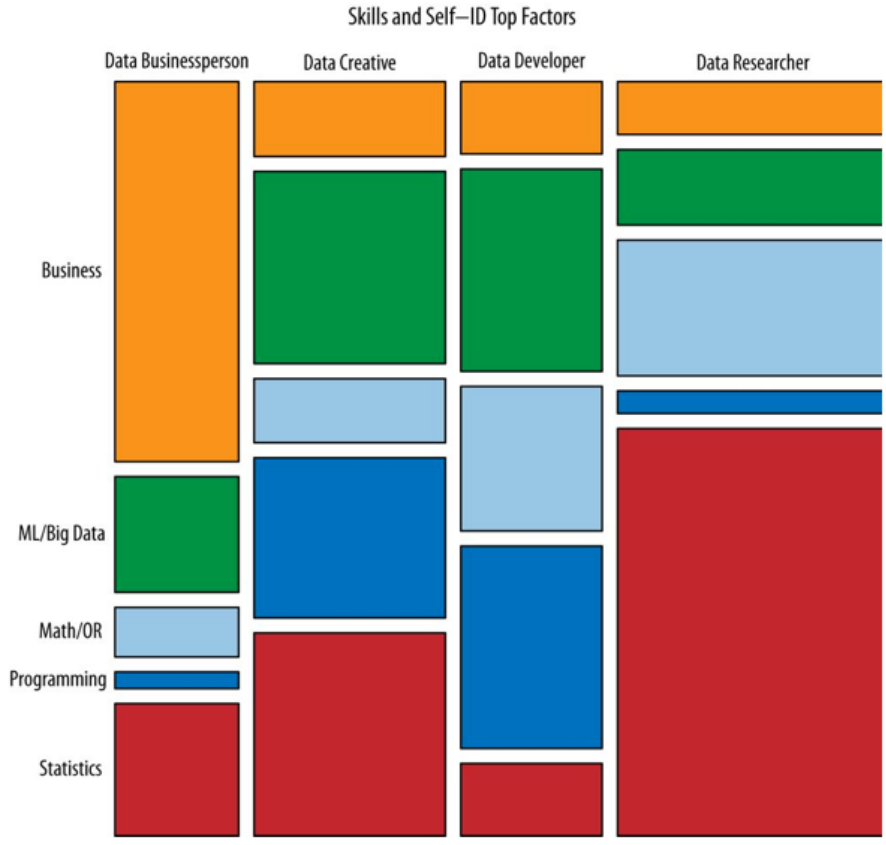


Figure 13. Data Scientists by Industry





"Analyzing the Analyzers" by Harlan Harris, Sean Murphy, and Marck Vaisman , O'Reilly Strata 2012

e	Date	Day	Topic	HW	Project
	Nov 19	W	Introduction to data science		
	Nov 24	M	Exploratory data analysis		
	Dec 1	M	Introduction to machine learning	1	
	Dec 3	W	Linear regression and regularization		
	Dec 8	M	Model selection and evaluation	2	
	Dec 10	W	Classification: kNN, decision trees		
	Dec 15	M	Classification: SVM	3	
	Dec 17	W	Ensemble methods: random forest		Title
			Christmas break		
	Jan 5	M	Intro to probability, naïve Bayes and logistic regression	4	Summary
	Jan 7	W	Feature engineering and selection		
	Jan 12	M	Clustering: k-means, hierarchical clustering	5	
	Jan 14	W	Dimensionality reduction: PCA and SVD		
	Jan 21	W	Text mining and information retrieval	6	
	Jan 26	M	Network analysis	7	
	Jan 28	W	Recommender systems		Proposal
	Feb 2	M	Relational databases, SQL	8	
	Feb 4	W	Big data storage and retrieval: noSQL, GraphDB		
	Feb 9	M	Big data distributed computing: map-reduce, spark rdd		
	Feb 11	W	Advanced: neural networks and deep learning		
	Feb 18	W	Guest lecture		
	Feb 23	M	Final projects presentations		Github
	Feb 25	W	Final projects presentations		

II. THE DATA SCIENCE WORKFLOW

Hilary Mason (bitly, HackNY, DatGotham conf, Accel)

1. **Obtain** - pointing and clicking does not scale (APIs, Python, shell scripting)!
2. **Scrub** - "Scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits" (Python, sed, awk, grep)!
3. **Explore** - look at the data (visualizing, clustering, dimensionality reduction)!
4. **Model** - "All models are wrong, but some are useful" / models are built to predict and interpret!
5. **Interpret** - "The purpose of computing is insight, not numbers"

Source: http://www.dataists.com

Jeff Hammerbacher (Facebook, Cloudera)

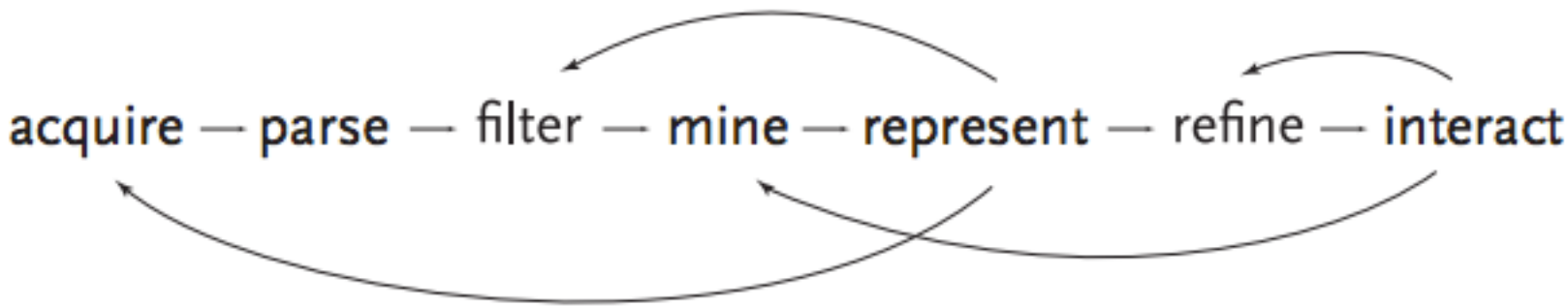
- 1. Identify problem
- 2. Instrument data sources
- 3. Collect data
- 4. Prepare data (integrate, transform, clean, impute, filter, aggregate)
- 5. Build model
- 6. Evaluate model
- 7. Communicate results

source: http://berkeleydatascience.files.wordpress.com/2012/01/20120117berkeley1.pdf

Ben Fry (author of "Visualizing Data")

source: http://benfry.com/phd/dissertation-110323c.pdf

THE DATA SCIENCE WORKFLOW



source: http://benfry.com/phd/dissertation-110323c.pdf

III. UNIX COMMAND LINE

EXERCISE – WORKING AT THE UNIX COMMAND LINE

28

KEY OBJECTIVES

- Navigate the filesystem
- Create, move, copy, and delete files & directories
- View & search files
- Edit & interact with files
- Combine steps
- Learn more

TOOLS

Files and directories:

- ls, cd, pwd
- cat, touch, mv, cp, mkdir, rm, rmdir

Manipulating data:

- head, tail, more, less, cut, paste, split
- tr, sort, uniq, wc, grep
- pipe (|), > , >>, <

Compressing data:

- compress, gzip, tar,

IV. PYTHON TUTORIAL

EXERCISE – PYTHON

KEY OBJECTIVES

- Become familiar with iPython notebook
- Learn basic Python