Next-Gen Data Scientists

Dr. Rachel Schutt
Johnson Research Labs
Strata, February 27, 2013
Why do we need Next-Gen Data Scientists?

- The number of immediately important data science problems out there today is astounding.
- We need a match between capabilities of workforce and those problems.
- New kinds of data. Paradigm shift around data.
- Data Scientists need hybrid skill sets: CS + statistics + viz +...
- We face a talent shortage.
- Data Science needs to be defined in a more deep and rigorous way to merit the word "Science"
What's wrong with Current-Gen?

My answer depends on whether you caught me when I'm feeling

- hopeful
- frustrated

Mostly I'll focus on hope because it's more constructive.
My way of contributing = my course

Introduction to Data Science, Columbia University

Personal Discovery of Data Science: How the course came about

3 structures

○ data scientist profiles
○ data science process
○ prototypes of next-gen data scientists
Personal Discovery of Data Science:
PhD in Statistics, Columbia University

- Classes, Research
- Statistical modeling, fitting, simulation
- Coding in R
- Data
  - Small
  - Handed to me in fairly good shape
- Computer cluster
- Data intuition
- Asking questions
Personal Discovery of Data Science:
Statistician at Google

- Statistician in Computer Science Land (circa 2008)
  - Code in C++, R, Sawzall, Python
  - Large Data Sets
  - Mapreduce, engineering stack, infrastructure
  - Data Munging & Cleaning
  - Algorithms vs./& models
  - Code reviews, version control
  - 90%? of time not doing statistics
Personal Discovery of Data Science: Statistician at Google

- Statistician in Computer Science Land (circa 2008)
  - Code in C++, R, Sawzall, Python
  - Large Data Sets
  - Mapreduce, engineering stack, infrastructure
  - Data Munging & Cleaning
  - Algorithms vs./& models
  - Code reviews, version control
  - 90%? of time not doing statistics

- Data Science not yet a term!
Personal Discovery of Data Science: Statistician at Google

- Team: software engineers, social scientists, statisticians
- Statistical Modeling, Machine Learning
- Exploratory Data Analysis
- Experiments, A/B testing
- Focus on Users
- Communicating to engineers, leaders, product managers
- Visualization
- Data Products
Personal Discovery of Data Science:
Started paying attention outside Google (2011)

"Data Science. What the hell is that?"

"Oh that's what I do."

"That's just statistics."

"Actually, no it's not."

"I didn't learn this in school."

"The students need to learn this stuff."
Hi David,
Have you been watching this "data science" ex happening lately?-- i.e. people calling themselv "chief data scientists" at start-ups; and STRAT big data conference happening each year in NY San Jose.
I've been watching with interest and was wondering if/how the stats department is handling it and what you're seeing an increase in demand in statistic courses because of it.

This idea is nascent in my mind, but I was thin about proposing a course structured around dat science, and was wondering if you'd be interest
Some details about the class

- Columbia University, Fall, 2012
- Students. Expected vs. Actual.
- Guest Lecturers, engaging the community
- What's Data Science?
  - Job Title
  - Deep Discipline
- Kaggle competition, Will Cukierski
- columbia datascience.com
- Cathy O'Neil, mathbabe.org
Data Scientist as a relatively new job title

- DJ Patil and Jeff Hammerbacher
- Alternative job titles
- Data Science Teams
- Data Products
- Google
Data Science as a Discipline with Deep Academic Roots

- Bell Labs
  - Bill Cleveland, Data Science action plan, 2001
  - John Tukey, Exploratory Data Analysis, R, Visualization
  - Claude Shannon, Information Theory

- Statistics
- Computer Science
Data Science is what Data Scientists Do

Let's explore this in 3 ways:

Skill set → profiles
Process
Prototypes
Data Science Profiles
First I brainstormed a long list of what Data Scientists do

Exploratory data analysis
Visualization (for exploratory data analysis and reporting)
Dashboards, metrics
Business Insights
Data-driven decision making
Data Engineering/Big Data (Mapreduce, Hadoop, Hive, Pig)
Get the data themselves
Build data pipelines (logs -> mapreduce -> dataset -> join with other data -> mapreduce -> scrape some data -> join)
Do "data science" to build products vs to describe existing product usage
Hack
Write patents
Detective work
Predict future behavior or performance
Write up findings in reports, presentations, journals

Programming (proficiency in R, Python, C, Java, etc.)
Conditional probability
Optimization
Algorithms, statistical models, machine learning
Tell stories, interpret,
Ask good questions,
Investigate
Conduct research
Make inferences from data
Build data products
Find ways to do data processing, munging and analysis at scale
Sanity checking
Data intuition
Interact with domain experts (or gain domain expertise)
Design and analyze experiments
Find correlation in data, and try to establish causality
I introduced the concept of Data Science Profiles

- Bucketed according to academic disciplines
- Self-reporting
- Self-awareness
- Relative to oneself
- What is an expert?
- Scale
- Order of x-axis?
No one person can be the perfect data scientist, so we need teams.
The students improved upon the visualization

The Stars of Data Science

Students in Columbia’s Introduction to Data Science course came from across the academic spectrum. Their skills are presented here in star charts with spokes representing their skill levels* across the data science skillset: R, statistics, mathematics, communication, data visualization, machine learning, computer science, and data wrangling. In addition to hovering in the center, the star chart of the overall class mean underlies each academic domain, so you can see students from each academic domain relative to the rest of the class. How would you compose your own intergalactic data science team?

*Skills were assessed by a survey written and administered by a subset of students in the class.
Data Science is what Data Scientists Do

Let's explore this in 3 ways:

Skill set → profiles
Process
Prototypes
The Data Science Process
The Data Science Process
The way it's usually taught

Clean Data Set → Model
The Data Science Process
In reality

Real World

Raw Data is Collected

Data is Processed

Clean Data

Exploratory Data Analysis

Machine Learning Algorithms, Statistical Models

Build Data Product

Communicate Visualizations Report Findings

Make Decisions
Data Science is what Data Scientists Do

Let's explore this in 3 ways:

Skill set ➔ profiles
Process
Prototypes
Prototypes of Next-Gen Data Scientists

Interesting people doing worthwhile things
Matt Gattis
Founder, Hunch.com (acquired by eBay)

- MIT undergrad in engineering
- 20 questions
- Humans are predictable
- Recommendation systems
- Very large and sparse matrices
Cathy O'Neil
Senior Data Scientist, Johnson Research Labs

- Ph.D. in math from Harvard
- Barnard math professor, quant at DE Shaw, senior data scientist at NYC start-ups
- Activist and blogger
- Interested in the ethics of models, the limitations of modeling, the politics behind models
- Thinks geometrically and able to find structure in unstructured situations and data
- Question-asker; willing to say "I don't know"
David Huffaker
Senior User Experience Researcher, Google

- Ph.D. in Media, Technology and Society from Northwestern University
- Small Data Sets ↔ Large Data Sets
- Qualitative ↔ Quantitative
- Social products, social networks
- Finding the story
Ian Wong
Inference Scientist, Square

- Started Ph.D. program in statistics at Stanford and left to go to Square
- Our initial conversations were about how best to train new hires
- Paired programming
- Interested in question of what is good code
- Fraud detection
  - as a semi-supervised learning problem
  - role of data visualization
- Ambient Visualizations
Mark Hansen,  
Director, Brown Institute for Media Innovation,  
Professor of Journalism at Columbia  

- Ph.D. and MA in Statistics from the University of California, Berkeley and a BS in Applied Math from UC-Davis  
- Previously worked at Bell Labs, as a professor at UCLA, as a visiting researcher at New York Times, now prof at Columbia  
- Triangulation of data, art and technology  
- Displayed in art museums and New York Times lobby  
- Expansionist view
What are the **credentials** of these data scientists?

- Many have PhDs. Not a requirement. But "**they could have gotten one**".
- Field of study often in quantitative subject
- Don't necessarily have "data scientist" as job title
- Are able to code and pick up programming languages naturally
- Proven set of problem-solving skills
These data scientists have also cultivated habits of mind

<table>
<thead>
<tr>
<th>Persistence</th>
<th>Gathering data through all senses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thinking about thinking (meta-cognition)</td>
<td>Creating, imagining, innovating</td>
</tr>
<tr>
<td>Thinking flexibly</td>
<td>Taking responsible risks</td>
</tr>
<tr>
<td>Questioning and posing problems</td>
<td>Finding humor</td>
</tr>
<tr>
<td>Striving for accuracy</td>
<td>Thinking interdependently</td>
</tr>
<tr>
<td>Applying past knowledge to new situations</td>
<td>Remain open to continuous learning</td>
</tr>
<tr>
<td>Thinking and communicating with clarity and precision</td>
<td>Responding with wonderment and awe</td>
</tr>
<tr>
<td></td>
<td>Listening with understanding and empathy</td>
</tr>
<tr>
<td></td>
<td>Managing impulsivity</td>
</tr>
</tbody>
</table>

The 16 Habits of Mind just described were drawn from research on human effectiveness, descriptions of remarkable performers, and analyses of the characteristics of efficacious people.

Beyond building up a core set of skills, we explored some of these course themes:

- Causing the future
- Democratization of machine learning
- Paradigm shift
- Ethics, Intuition
- Hubris, knowing your own limits
- Human vs. machines
Next-Gen Data Scientists Manifesto

columbiadatascience.com/manifesto

I don’t call myself a “data scientist”. I call myself a st because as it’s currently used, it’s a meaningless, arb the term, and apparent “sexiness” of the profession d opportunities. So we need Next-Gen Data Scientists. I

Gen Data Scientist:

• Next-Gen Data Scientists have humility. They d spend most of their efforts on self-promotion.
• Next-Gen Data Scientists have integrity. Their some “cool” problem. It’s about being a problem complicated, if necessary)
• Next-Gen Data Scientists don’t try to impress w work.
• Next-Gen Data Scientists spend a lot more time admit.
• Next-Gen Data Scientists have the experience o about. They’ve put their time in.
• Next-Gen Data Scientists are skeptical – skeptik and the way they’re used or can be misused.
• Next-Gen Data Scientists make sure they know around trying to show everyone else they exist.
• Next-Gen Data Scientists have a variety of skills visualization, communication, math.
• Next-Gen Data Scientists do enough Science to hypotheses and welcomes challenges and alter
Students in Intro to Data Science class,
Statistics 4242 Columbia University Fall, 2012
Developments Since Course was Proposed

● In Data Science Education
  ○ Columbia's Institute for Data Sciences & Engineering
  ○ NYU Center for Data Science
  ○ Multiple courses and degree-granting programs popping up all over the country

● Johnson Research Labs
  ○ Book
  ○ Education
  ○ Consulting
  ○ Research
rachel@jrl-labs.com
Johnson Research Labs
columbiadataascience.com