Data Scientists : Silicon Valley Rock Stars

Jay Buckingham
Dynamic Signal
jbuckingham@dynamicsignal.com

Companies Need Data Scientists Now!

December 9, 2012 6:20 pm

Groups grapple with data scientists shortage

Financial Times

As Startups Produce More Data, the Search for Data Scientists Grows Frantic

Posted on: February 07, 2013

PeHub.com

April 29, 2012, 9:44 a.m. ET

Big Data's Big Problem: Little Talent

Wall Street Journal

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Harvard Business Review

So What Is Data Science?

Making use of vast amounts of data to:

- Discover what we don't know
- Obtain predictive, actionable insight
- Better serve your customers
- Improve your products
- Invent new products customers love

Examples

- LinkedIn: People you might know
- Amazon: Recommendation
- Google: Selecting ads for you
- Pandora: Selecting music you like
- Dynamic Signal: Who is influential?

Commonality: Understanding users

Tools for the Job

- Statistics
- Machine Learning
- Map/Reduce

Dynamic Signal Examples

- Which users are most important?
  - The ones that others interact with (comments, shares, retweets, …)
  - Rank them
  - Find their best posts
- Demographics
  - Who are these people?

How Big is the Job?

- ~500M Twitter users
- We analyze ~20M
- How much data is that?
- 5 TB
- How long will it take to process?
  - 3.0 GHz Xeon
  - 190000 sec/TB => 2 days

2 Days!! Must do better

- Map/Reduce
  - Spread job across many machines
- With 48 cores: 2 days => 2 hours

Tell us about these people

- Brands want to know more about these important people
  - Demographics
- Twitter doesn't know
- We can figure it out

How Can We Figure Out Demographics?

- Machine Learning
  - MaxEnt
  - Support Vector Machines
  - Boosting
  - Decision Trees

What Do Data Scientists Do:

- Data preparation
- Data presentation
- Experimentation
- Observation
- Data products

Hillary Mason & Chris Wiggins:

"Data science is a blend of hacker's arts, statistics and machine learning…

and the expertise in mathematics and the domain of the data for the analysis to be interpretable…

It requires creative decisions and open-mindedness in a scientific context."

Example Data Science Jobs

- LinkedIn: Recommendation systems
  - IR, Statistics, Java, Hadoop
- Velti: Mobile ads
  - MS/PhD Statistics, R, Machine Learning,Hadoop, SQL
- Drawbridge: Mobile ads
  - MS/PhD Statistics, Lucene/SOLR, Java, Hadoop
- Twitter: Social media
  - MS/PhD Statistics, Math, CS; R, Hadoop
- Pandora: Recommender systems
  - BS/BA CS or Statistics, R, Python, Java, Hadoop

LinkedIn: Data Scientist – New College Grad

We're looking for people who love turning data into gold. Are you someone who solves hard problems by creatively obtaining, scrubbing, exploring, modeling and interpreting big data? Do you know enough about information retrieval, machine learning, and statistics to be dangerous? Are you a hands-on implementer, ready to learn new languages and technologies to turn your ideas into solutions used by tens of millions of people around the world?

If your answer is "Yes, show me the data!" then this is the job for you. As data scientists, we work on some of the most exciting areas of computer and information science, including information extraction, recommendation systems, social network analysis, and network visualization. Moreover, we support core business needs and drive innovation.

Interested? Here is what we are looking for:

- Interest in solving problems with big data.
- A good knowledge of information retrieval, data mining, or related field.
- Technical skills (e.g., Java, Hadoop, Pig) and interest in new ones.
- Creativity and good taste in problems with business impact.
- You're in your final year of your Bachelor's, Master's, or PhD program.

What Skills do Data Scientists Need?

- Statistics
- Problem solving
- Curiosity and Creativity
- Computing

Skills You Need to Build

- Graph Theory
- Machine Learning
- NLP (a little)
- Software tools

Software Tools

- Get and process your own data:
  - Experience with APIs (Twitter, LinkedIn, FB)
  - Hadoop (Java)
  - C++ (Google, FB)
  - SQL
- Explore your ideas quickly:
  - Python
- Data visualization/presentation
  - R, javascript

To build skills – Do a Project

Example: Trending Topics on Twitter

- Detail what you want to learn
- Gather and store data
- Compute trending topics
- Presentation
- github

- Extra points: Use Hadoop
- More extra points: Use AWS Hadoop

Become Familiar with the Field

- Tech Companies
- Key people: VC, angels, Y-Combinator
- Tech press: Hacker News, GigaOM

Follow These People

Hilary Mason	hilarymason.com, @hmason
DJ Patil	@dpatil
Hacker News	@HackerNews
GigaOM	@gigaom
Paul Graham	@paulg
John Doerr	@johndoerr
Kleiner Perkins	@kpcb
Nate Silver	@fivethirtyeight

Data Scientist… Is it a Good Job?

“There's now 3 Tesla S's, 1 Ferrari, 2 SLS's, and 1 M5 in my building. Guess the economy is doing well.

How do you get a rock star data scientist job?

- Study job postings for data scientists
- Become familiar with companies
- Develop key skills they need

Be Clear About:

- What do you really want?
- What are you good at?
- What are you likely to be good at?

Get Clear About:

- What you want to learn
- What you want to become expert at
- What area you want to work in

Create Your Story

- Companies have lots of projects
- They really need help
- They really need to hire data scientists
- You just need to make it easy for them
  - Show you have the skills and
  - You Get Things Done!

Create Your Story

- You are building your brand.
- Think carefully about what you want people to think when they think of you.
- Right now:
  - World knows little about you
  - So you control the message
  - Choose wisely

Where to tell your story

- Resume
  - Dice
- LinkedIn
- Blog
- Facebook
- Twitter
- Github
- Good recruiter

Present your Professional Persona

- Email address: Yes:
  - jonsmith@gmail.com
  - jon@jonsmith.com
- Not:
  - pbrshotgun@aol.com
  - wildnights69@hotmail.com
- Facebook
  - Cleanse your posts
  - Use for professional purposes only
- Twitter: professional posts only

The Interview

- Do your homework:
  - What company does, key products
  - Technologies they use
  - Where they are headed
- Be clear about why you are there
- Clothes: Look professional
  - Dress like a VC
- Be early
  - Relax, meditation

The Interview: What you are asked

- How you'd approach data task
- Statistics
- Machine Learning
- Data structures
- Algorithms
  - Eg: Graph problem
- Cracking the Coding Interview (Gayle McDowell)

The Interview: What you should ask

- Ask engineers:
  - Key product areas
  - What data do they have?
  - What have they done so far?
  - Key things they want to learn from data
  - Technologies
- Ask CEO / VP:
  - Show interest in the business
  - What will this company achieve?
  - Biz plan, product plan, revenue plan
  - Competitors
- Tell them how you will help them achieve these goals

Interview Tips

- Relax!
  - Practice interviews
- Do NOT talk money at interview
- Do NOT take offers at lame companies!
- Treat recruiter well

First Job: Move to Bay Area

- More opportunities
- Build skills faster
- Build gravitas
- Build network

More Opportunities

- Great companies:
  - Established: Google, Facebook, LinkedIn, Twitter
  - Startups: Drawbridge, Dynamic Signal…
- Area saturated with tech buzz
  - Great learning opportunities
- Engineers are really valuable
  - So treated better than in Pacific NW

You'll build skills faster

- Bay Area companies are using new technologies way ahead of others
- Especially Big Data tools: Hadoop, (Google: MapReduce)
- Many more machine learning projects
- Great classes at Stanford, Berkeley

You'll build credibility faster

- Working at Bay Area startups gives you credibility
  - Because you've developed exciting new products with great teams in fast-moving teams at awesome companies.
  - You've worked on products using technologies way ahead of others

You'll build your network faster

- VCs
- CEOs
- Best data scientists and engineers
- So many tech companies near each other

First Job: Learning and Building Experience

- Work for a startup: learn a lot and "kill it"
- Start to build your "epic story"
- Choose startup in Bay Area, not a big company
- Choose based what you will learn, experience you will build.
  - Don't worry about money, stock yet

Second Job: Learning and Impact

- Choose job to:
  - Learn new skills / technologies
  - Have an impact
- You want to be key engineer on some project
- Compensation
  - Paid well
  - Good stock options

Transition: Student to Professional

- Bing Gordon: "Kill it"
- This is the time: Focus on career
- Get on projects where you can learn what you need to learn
- Later: Go for impact

Transition: Student to Professional

- Learn your craft
- Ask for help from people with broad skills.
- Esp Dev, PM, Business, CEO, VC
- Early on, tell folks you'll want to schedule time for coffee with them periodically to talk technology, project management, biz dev, strategy

Transition: Student to Professional

- Be professional at work

- Don't burn bridges
- Communicate with your manager, GM, VP
- Get mentors
  - in company
  - not in company

- Get out to professional gatherings
- Build network

Last Words:

"The sexy job in the next ten years will be statisticians

Hal Varian, Chief Economist, Google

http://www.slideshare.net/JayBuckingham/Data-Science-Opportunities

Appendices

Data Science Links

- Read this: http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/pr
- What does a data scientist do? http://www.slideshare.net/datasciencelondon/big-data-sorry-data-science-what-does-a-data-scientist-do
- Data scientists: http://www.forbes.com/pictures/lmm45emkh/tim-oreilly-is-the-founder-of-oreily-media/
- Data Science 101 blog: http://datascience101.wordpress.com/
- Hammerbacher data science course at Berkeley: http://datascienc.es/
- Hammerbacher on data science: http://www.quora.com/Jeff-Hammerbacher/answers/Data-Science
- Stanford data mining courses: http://scpd.stanford.edu/public/category/courseCategoryCertificateProfile.do?method=load&certificateId=1209602

Big Data / Social Data

- Berkeley class on techniques to collect and analyze Big Data using Twitter as an example:
  - http://www.ischool.berkeley.edu/courses/i290-abdt
  - https://blogs.ischool.berkeley.edu/i290-abdt-s12/
- How LinkedIn has improved its Big Data tools: http://gigaom.com/2013/03/03/how-and-why-linkedin-is-becoming-an-engineering-powerhouse/
- Paper on using social media data to predict demographic characteristics: http://www.pnas.org/content/early/2013/03/06/1218772110.full.pdf
- Mary Meeker on internet data trends: http://www.kpcb.com/partner/mary-meeker
- Intro to getting Twitter data with the Twitter API:
  - https://dev.twitter.com/start
  - https://dev.twitter.com/docs
- Example: how to get info about a Twitter user: https://dev.twitter.com/docs/api/1/get/users/lookup

Map/Reduce links

- Hadoop history: http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/

- Why use Hadoop: http://www.slideshare.net/digitald/web-tech-6626629
- Hadoop overview: http://www.slideshare.net/hadoop/practical-problem-solving-with-apache-hadoop-pig

- Google MapReduce: http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/mapreduce-osdi04.pdf
- Google file system: http://en.wikipedia.org/wiki/Google_File_System
- Hadoop history: http://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/
- Why use Hadoop: http://www.slideshare.net/digitald/web-tech-6626629

Machine Learning links

- Andrew Moore's tutorials: http://www.autonlab.org/tutorials/
- NLTK: nltk.org
- WordNet: wordnet.princeton.edu