Data scientist and Machine Learning engineers : who’s who?

Mike Pereira

Machine Learning Group

22/02/2017

Difference between Data scientist and Machine Learning Engineer

Data scientist

- Statistical analysis/research on raw data to determine which machine leaning approach to use, models the algorithm, and prototypes it usually in R, Python / PySpark, etc for testing.
- Data Scientists take the raw data, analyze it, connect the dots and tell a story often via several visualizations. They are more on the creative side. Like an Artist..

- Stronger in statistics
- More statistical analysis/research

Machine Learning Engineer

- Partners with the **Data Scientist** to take the ML model prototyped by the Data Scientist and make it work well in a production environment at scale (i.e. lots of concurrent users). Usually doing so by coding it in a more robust language like Scala, JAVA or C++ and utilizing faster data piping and parallel processing (Spark, MapReduce, etc.)
- An Engineer, on the other hand, is someone who looks at the data as something they have to take in and churn out an output in some appropriate form in the most efficient way possible

- Stronger in Computer Science
- More Data/Engineering

https://www.quora.com/What-is-difference-between-Data-scientist-and-Machine-Learning-Engineer

EX : Difference between Data scientist and Machine Learning Engineer at Quora

Data scientist

- Evaluate potential or existing approaches, features, algorithms or error metrics to help improve machine learning systems.
- Analyze the impact of machine learning algorithms on key metrics. This involves ad-hoc analysis of A/B tests, and understanding how ML systems fit in to key metrics of the company.
- Research and understand user behavior patterns such as engagement by building machine learning models. These machine learning models are made for one-off analyses and are not put into production. Their primary goal is to help evaluate ideas.

Machine Learning Engineer

- Build and implement production machine learning systems (e.g. recommendations, personalized ranking, and a lot more as described in How does Quora use machine learning in 2015?).
- Maintain the health of machine learning systems, including speed, reliability, and performance.
- Develop internal machine learning frameworks and abstractions to facilitate common tasks such as training / testing, feature use / reuse / creation / storage, and deployment. These abstractions are used by both machine learning engineers and data

A machine learning project will often be staffed by both data scientists and engineers. Here are the steps:

1. Data scientists conduct research to identify possible needs or improvements in machine learning systems
2. Machine learning engineers build, implement, or improve the machine learning system
3. Data scientists evaluate the impact of the machine learning system on company metrics

https://www.quora.com/What-is-the-difference-between-a-machine-learning-engineer-and-a-data-scientist-at-Quora

Jobs and associated skills : Data Science

4 Types of Data Science Jobs

“Data scientist” is often used as a blanket title to describe jobs that are drastically different. Here are four types of data science jobs:

A Data Scientist is a Data Analyst Who Lives in San Francisco:

- Basically a data analyst.
- Job might consist of tasks like :
  - pulling data out of MySQL databases,
  - becoming a master at Excel pivot tables, and producing basic data visualizations (e.g., line and bar charts).
  - take the lead on your company’s Google Analytics account.
- Great place for an aspiring data scientist to learn the ropes, try new things and expand skillset.

Reasonably Sized Non-Data Companies Who Are Data-Driven:

- You’re joining an established team of other data scientists. The company you’re interviewing for cares about data but probably isn’t a data company.
- It’s equally important that you can
  - perform analysis,
  - touch production code,
  - visualize data, etc.
- Profiles : either generalists a specific niche where they feel their team is lacking, (ex : data visualization or ML).
- Skills required : familiarity with tools designed for ‘big data’ (e.g., Hive or Pig) and experience with messy, ‘real-life’ datasets.

Please Wrangle Our Data!:

- In companies looking to exploit their data.
- Job :
  - set up a lot of the data infrastructure that the company will need moving forward.
  - make meaningful data-like contributions to the production code and provide basic insights and analyses.
- You’d be (one of) the first data hires : less important that you’re a statistics or machine learning expert. Ideal profile : data scientist with a software engineering background.
- Mentorship opportunities for junior data scientists may be less plentiful at a company like this. As a result, you’ll have great opportunities to shine and grow via trial by fire, but there will be less guidance and you may face a greater risk of flopping or stagnating.

- **We Are Data. Data Is Us:**
- companies for whom their data (or their data analysis platform) is their product. In this case, the data analysis or machine learning going on can be pretty intense.
- Profile : a formal mathematics, statistics, or physics background and hoping to continue down a more academic path.
- Data Scientists in this setting likely focus more on producing great data-driven products than they do answering operational questions for the company.

http://blog.udacity.com/2014/11/data-science-job-skills.html

Jobs and associated skills : Data Science

Basic Tools: No matter what type of company you're interviewing for, you're likely going to be expected to know how to use the tools of the trade. This means a statistical programming language, like R or Python, and a database querying language like SQL.

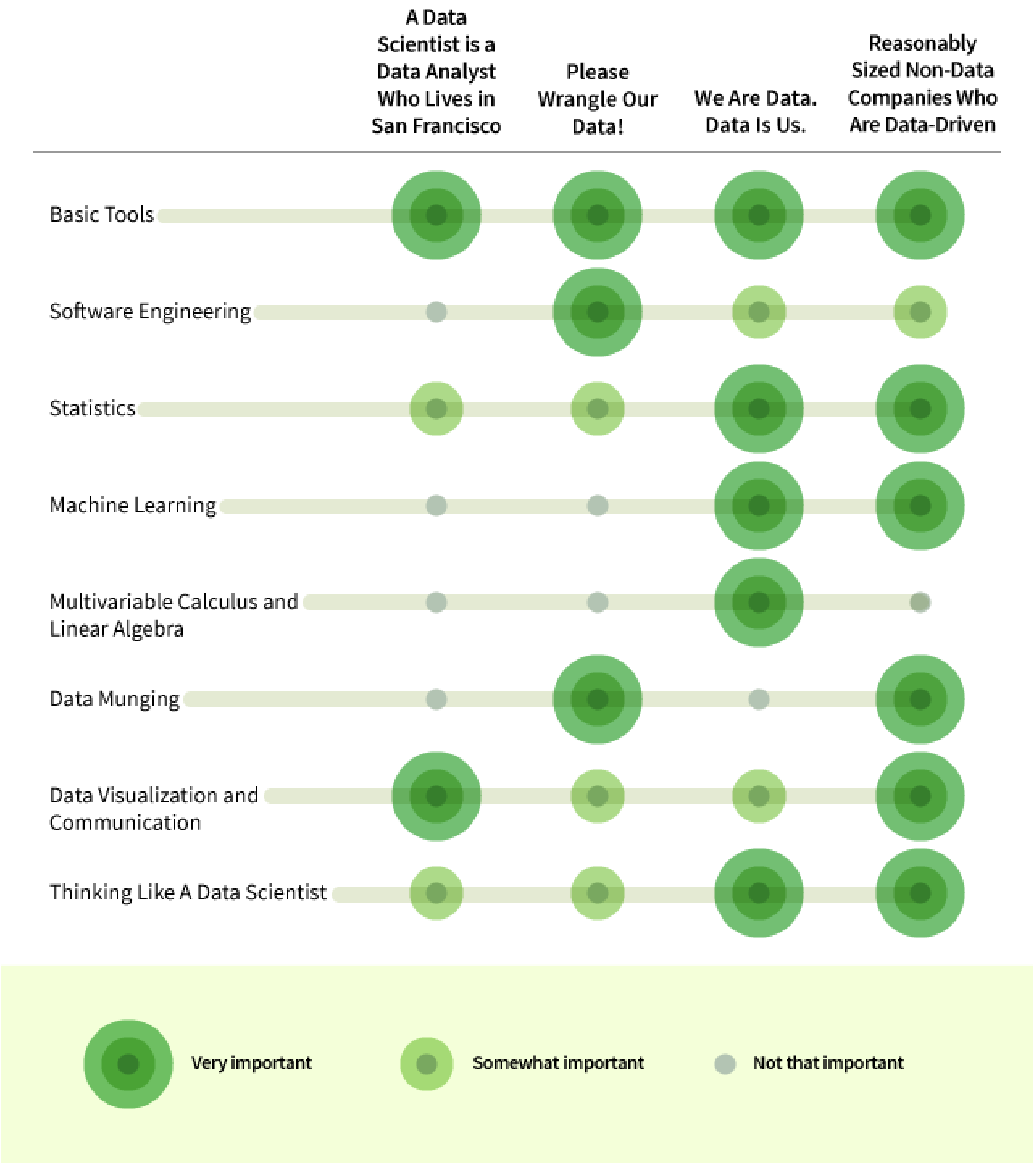
Basic Statistics: You should be familiar with statistical tests, distributions, maximum likelihood estimators, etc. and understanding when different techniques are (or aren't) a valid approach.

Machine Learning: This can mean things like k-nearest neighbors, random forests, ensemble methods – all of the machine learning buzzwords. More important is to understand the broadstrokes and really understand when it is appropriate to use different techniques.

Multivariable Calculus and Linear Algebra: since they form the basis of a lot of these techniques.

Data Munging: Often times, the data you're analyzing is going to be messy and difficult to work with. Because of this, it's really important to know how to deal with imperfections in data.

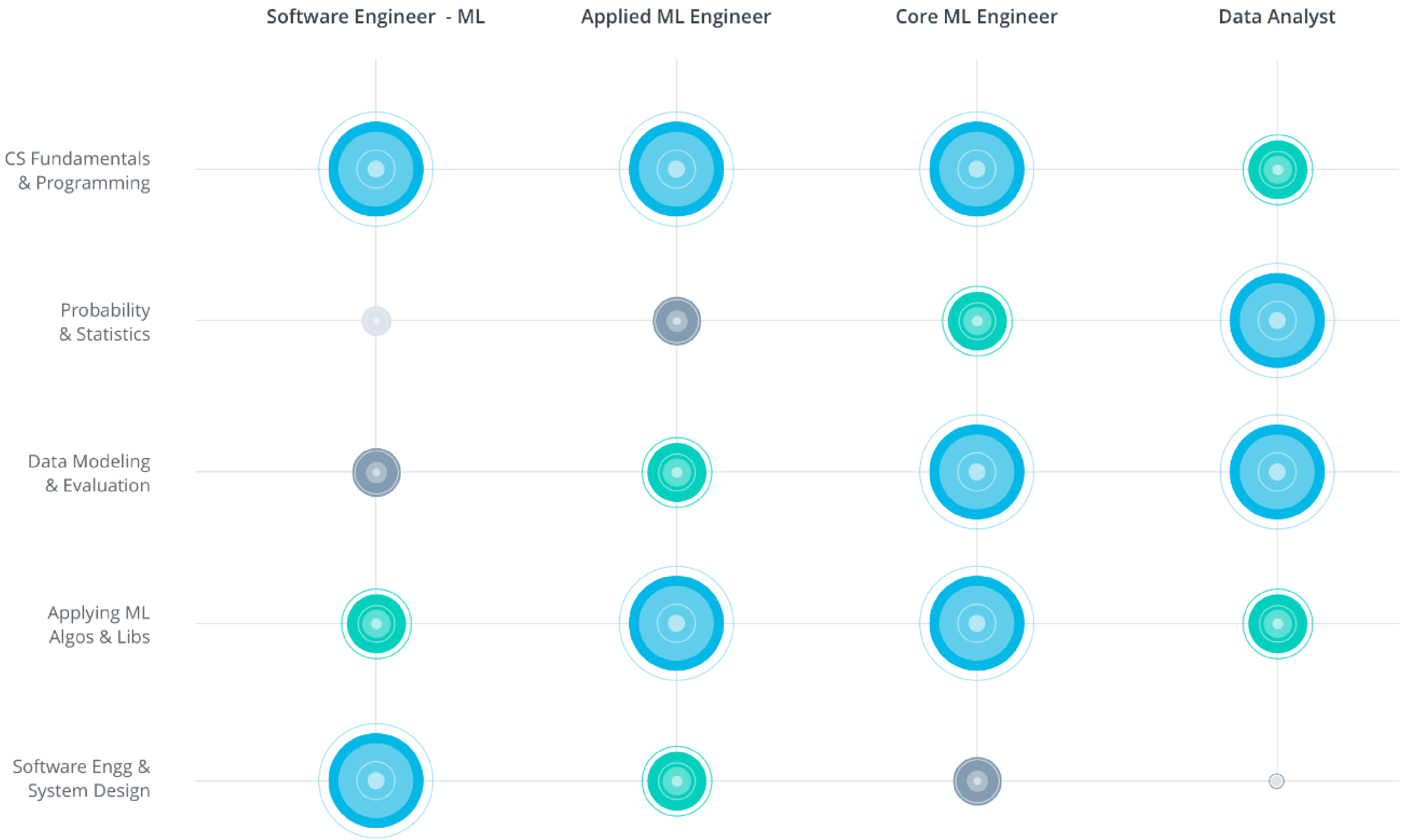
Thinking Like A Data Scientist: Companies want to see that you're a (data-driven) problem solver.



	A Data Scientist is a Data Analyst Who Lives in San Francisco	Please Wrangle Our Data!	We Are Data. Data Is Us.	Reasonably Sized Non-Data Companies Who Are Data-Driven
Basic Tools	Very important	Very important	Very important	Very important
Software Engineering	Not that important	Very important	Somewhat important	Somewhat important
Statistics	Somewhat important	Somewhat important	Very important	Very important
Machine Learning	Not that important	Not that important	Very important	Very important
Multivariable Calculus and Linear Algebra	Not that important	Not that important	Very important	Not that important
Data Munging	Not that important	Very important	Not that important	Very important
Data Visualization and Communication	Very important	Somewhat important	Somewhat important	Very important
Thinking Like A Data Scientist	Somewhat important	Somewhat important	Very important	Very important

http://blog.udacity.com/2014/11/data-science-job-skills.html

Jobs and associated skills : Machine learning engineer



Software Engineer - ML

Applied ML Engineer

Core ML Engineer

Data Analyst

CS Fundamentals & Programming

Probability & Statistics

Data Modeling & Evaluation

Applying ML Algos & Libs

Software Engg & System Design

http://blog.udacity.com/2016/04/5-skills-you-need-to-become-a-machine-learning-engineer.html