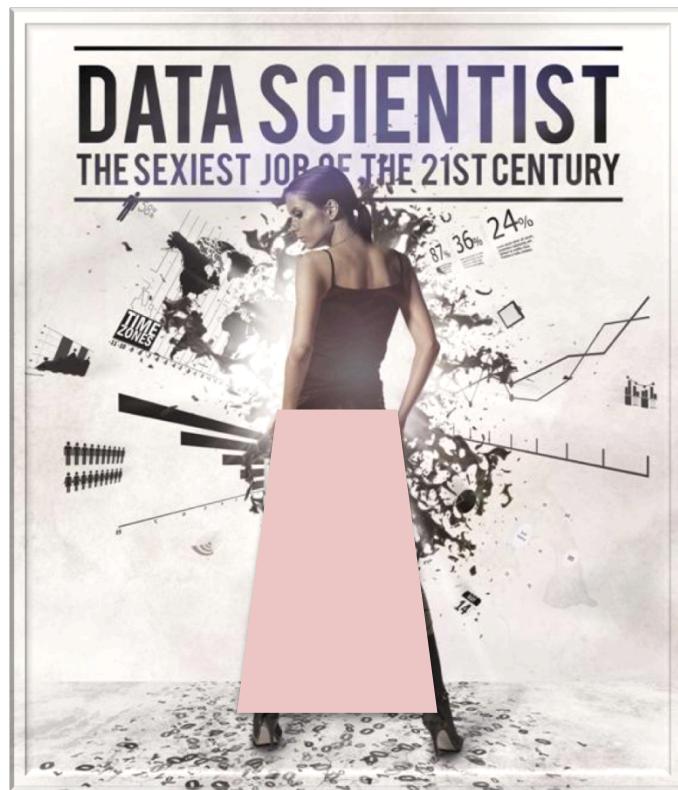


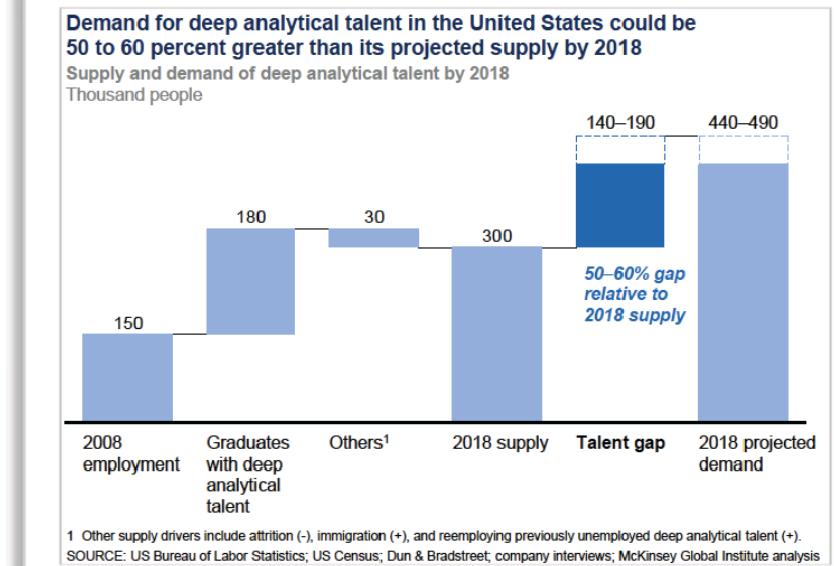
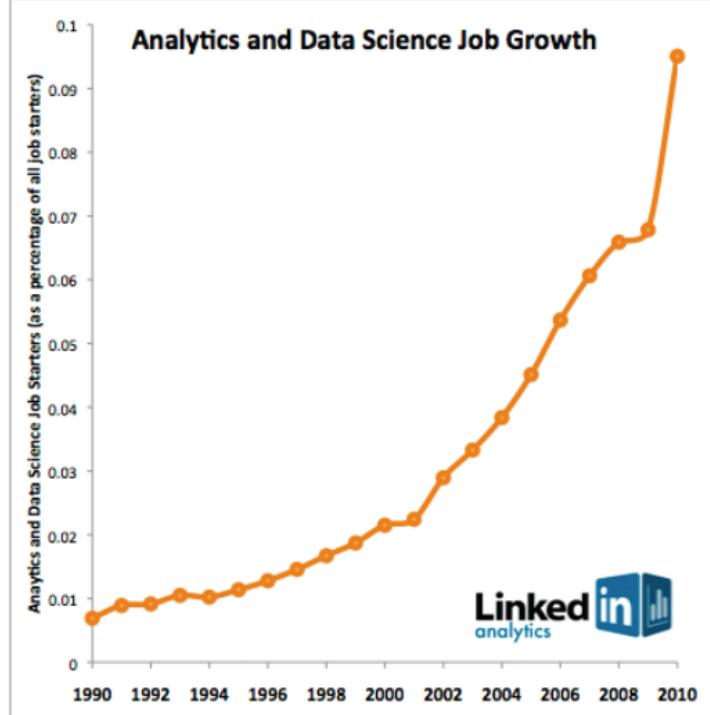
*Academia* → *Data Science*

Nhung Ho

# Why Data Science?



# Why Data Science?



# What is Data Science?

**Consultant + Quant + Engineer**

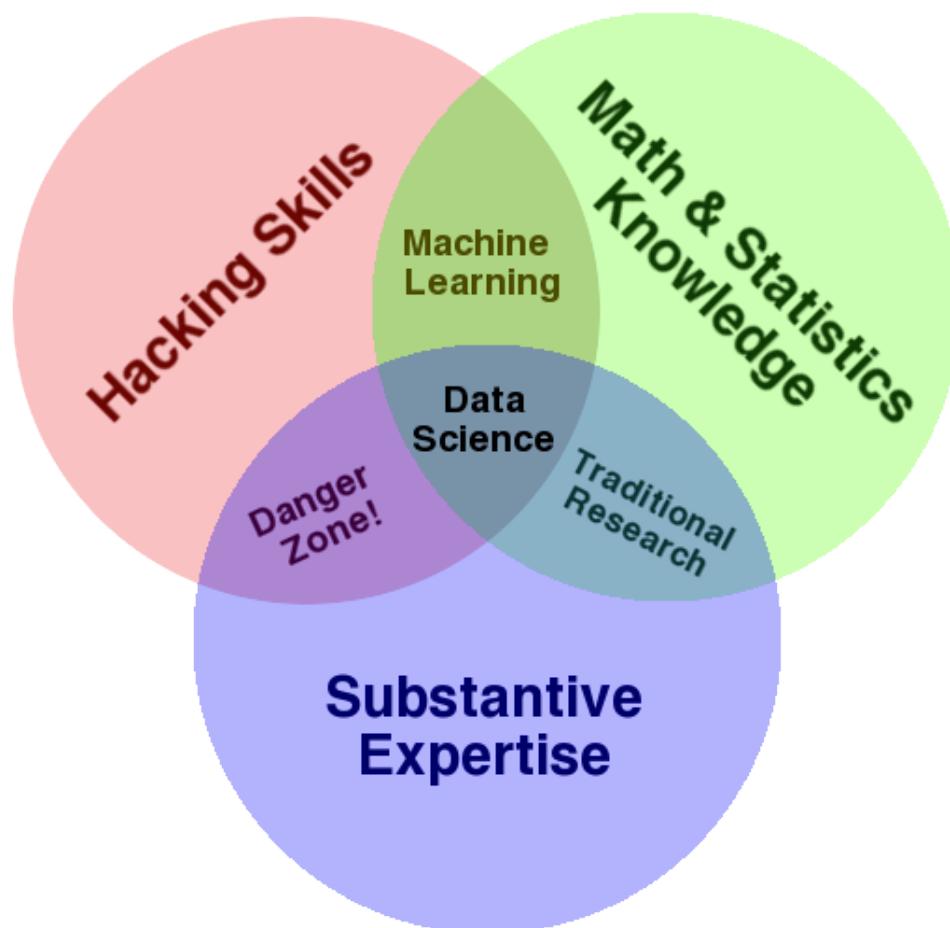
**Consultant: Explores, identifies, advises**

**Quant: strategizes, codes**

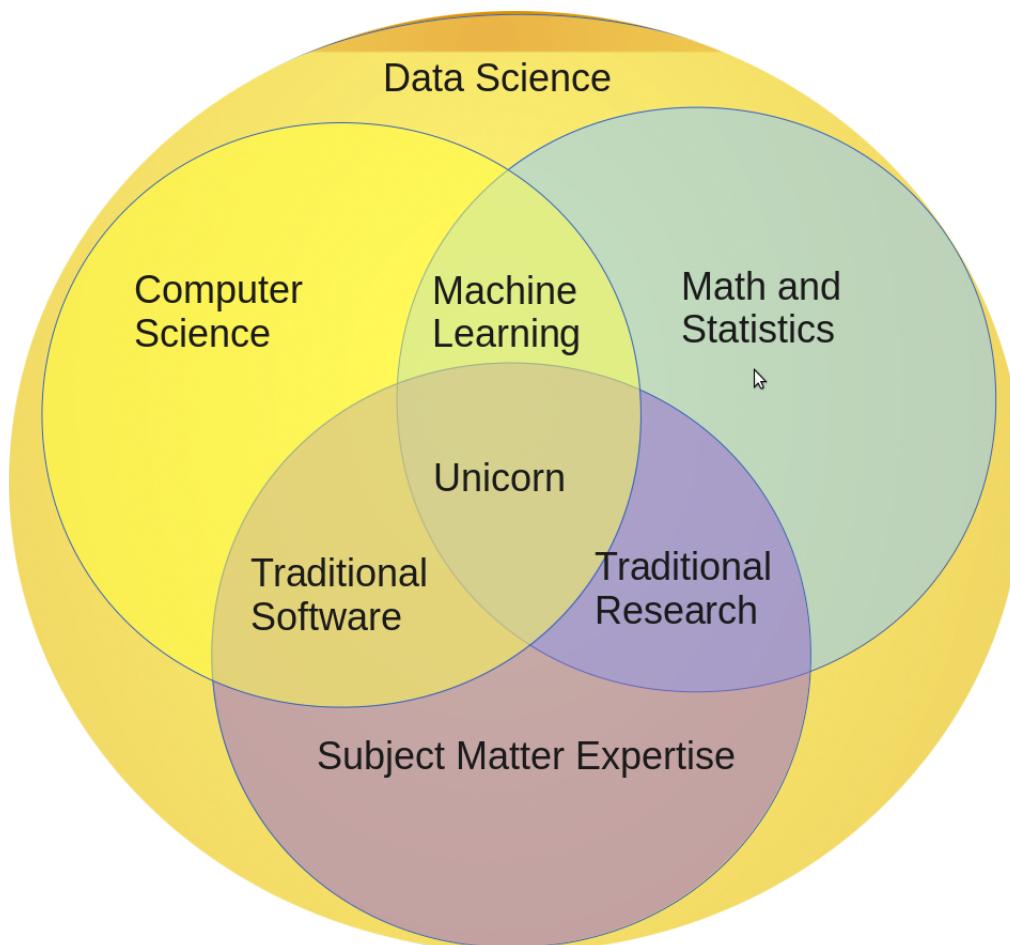
**Engineer: product development**

**= Data Scientist**

# Data Science skills



# Data Science skills



# *Types of Data Scientists:*

## *Product Analytics*

**What is it?**

**Business analytics**

**Skills?**

**Curious, data explorers**

**Statistics heavy, strong communication**

**Who do they work for?**

**Product managers, engineers**

**Goals are to improve *internal* metrics**

# *Types of Data Scientists:*

## Product Analytics



# *Types of Data Scientists:*

## Predictive Modeling

What is it?

Risk, sales, optimization

Skills?

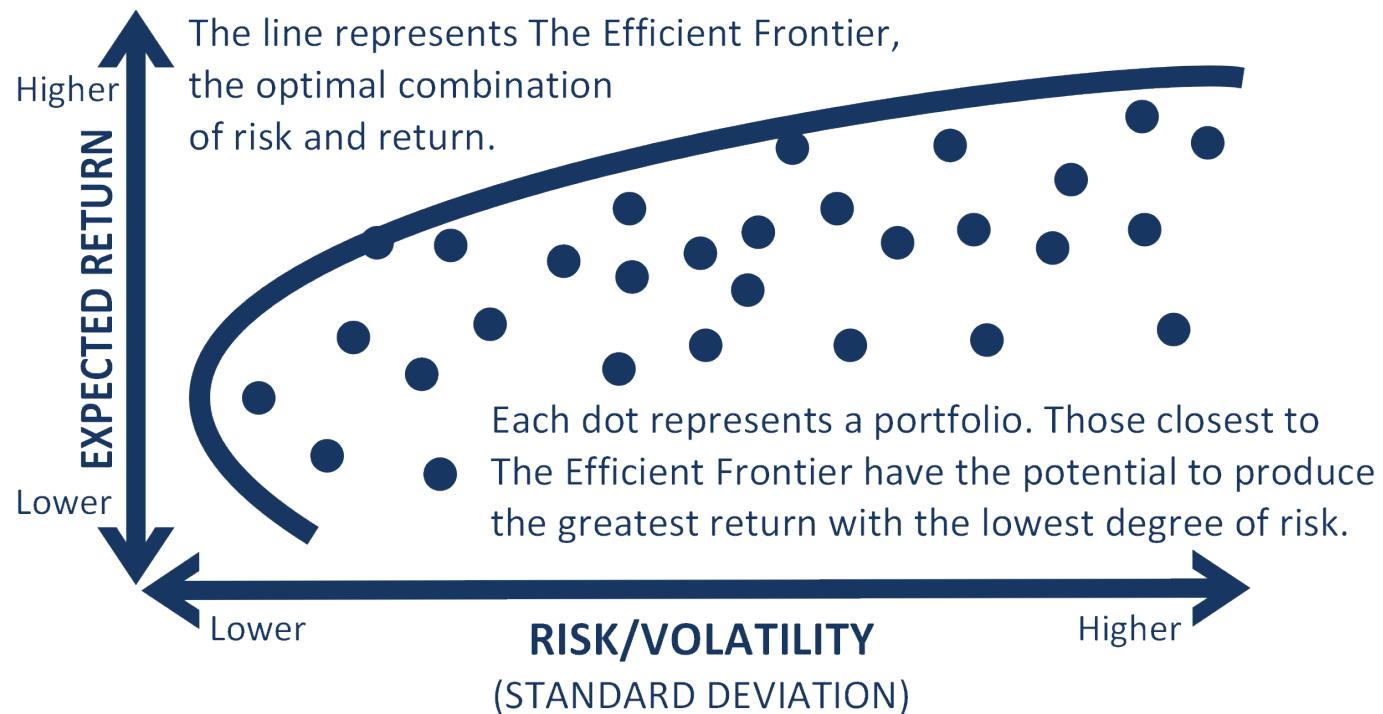
Machine learning

Statistics heavy

Who do they work for?

Goals are to optimize performance, reduce risk

# *Types of Data Scientists: Predictive Modeling*



# *Types of Data Scientists:*

## *Data Products*

What is it?

Product development

Skills?

Machine learning, statistics

Software engineering

Rapid prototyping

Who do they work for?

End user

# *Types of Data Scientists: Data Products*

 <p>Pieter Dokkum pvdkum@gmail.com</p> <p><a href="#">Add to</a>   3 shared contacts</p>	 <p>Andrei Tokovinin 2nd Astronomer at Cerro Tololo Inter-American Observatory Chile</p> <p><a href="#">Connect</a>   2 shared connections</p>
 <p>Paul Goudfrooij 2nd Senior Scientific Staff member at STScI Baltimore, Maryland Area</p> <p><a href="#">Connect</a>   2 shared connections</p>	 <p>Meg Schwamb megan.schwamb@gmail.com</p> <p><a href="#">Add to</a></p>
 <p>Virgil Griffith 2nd Scientist &amp; Internet Man of Mystery Greater Los Angeles Area</p> <p><a href="#">Connect</a>   1 shared connection</p>	 <p>Beate Heinemann 2nd Professor of Physics at UC Berkeley and Lawrence Berkeley National Laboratory Oakland, California</p> <p><a href="#">Connect</a>   2 shared connections</p>
 <p>Nitya Kallivayalil nityajk@gmail.com</p> <p><a href="#">Add to</a>   1 shared contact</p>	 <p>Joseph Prang 2nd Analyst at Square San Francisco, California</p> <p><a href="#">Connect</a>   2 shared connections</p>
 <p>Hanh Huynh 2nd Premier Service Center Specialist Salt Lake City, Utah</p> <p><a href="#">Connect</a>   4 shared connections</p>	 <p>Ricardo Munoz kleineameise2@gmail.com</p> <p><a href="#">Add to</a>   1 shared contact</p>

## Why are PhDs suited for this?

Experimental design

Statistics *Fast learners*

Hypothesis driven   Curious

**Hackers**

Creative thinkers

**Coders**

## Why are PhDs suited for this?

Experimental design

Statistics *Fast learners*

Hypothesis driven  
**Data storytellers**  
Hackers

Creative thinkers  
Coders

# *My transition*



# My transition



# Insight Data Science

>30 companies

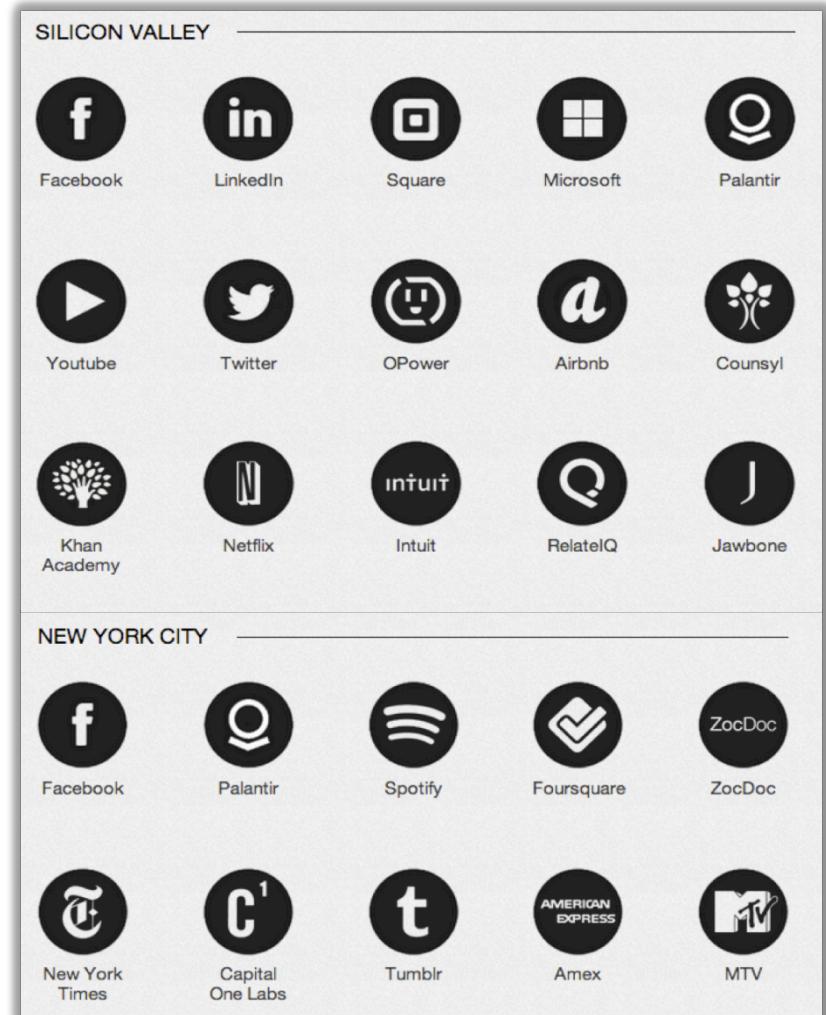
~20/500 applicants

100% placement

[insightdatasience.com](http://insightdatasience.com)

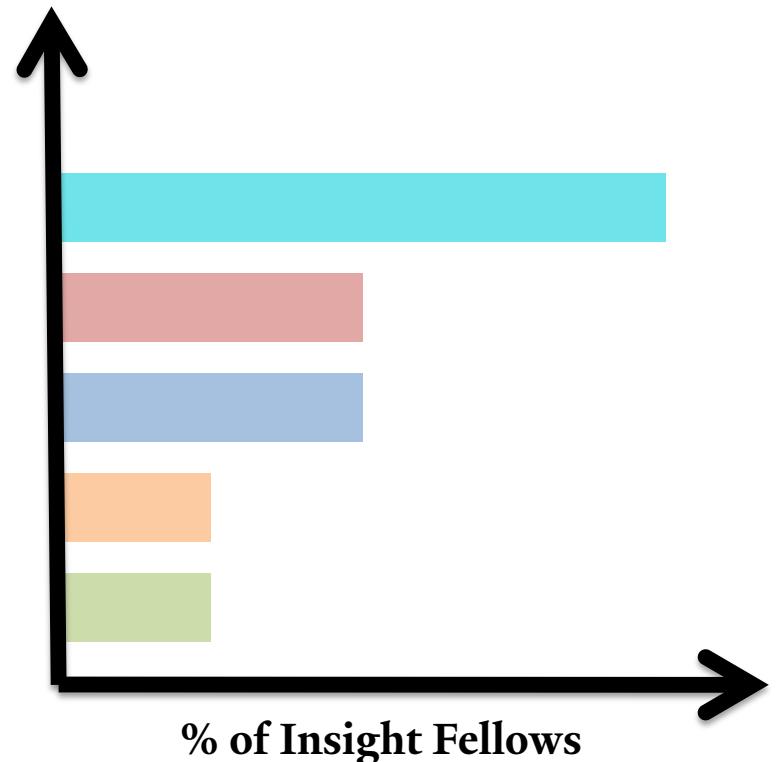
July 28 session in NYC

deadline April 27



# Who are these PhD data scientists?

- Phds+Postdocs
- Breakdown:
  - ~40% physics+astronomy
  - ~20% biological sciences
  - ~20% engineering
  - ~10% math+statistics
  - ~10% social sciences



# What is the missing 10%?

- Machine Learning
  - Linear, logistic regression, random forest
- Databases
  - SQL, MongoDB
- Statistics
  - T-tests, p-value, Bayes' theorem
- Computer Science fundamentals
  - Sorting algorithms, data structures, recursive programming, dynamic programming

# Getting up to 100%

- Machine Learning
  - Andrew Ng's Coursera Machine learning course
- Databases
  - Sqlzoo.net, Coursera Databases course
- Statistics
  - Khan Academy statistics course
- Computer Science fundamentals
  - Interactivepython.org
- Kaggle competitions!

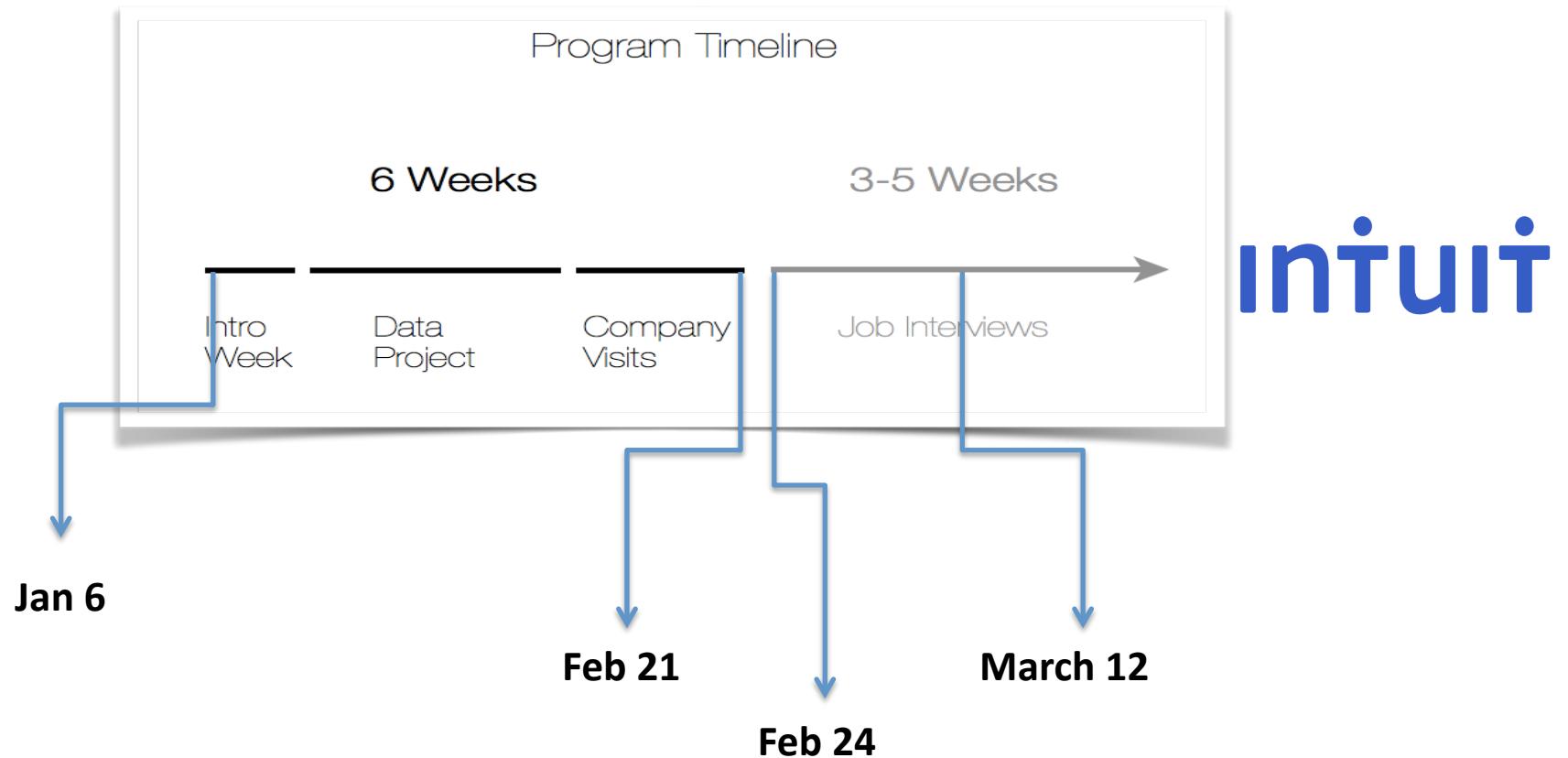
## *Tools of the trade aka “stack”*

- Python, python, python...
  - Pandas, numpy, scikitlearn,nltk, BeautifulSoup
- SQL, Hadoop, Hive, Pig
- Flask
- Javascript, Ruby on Rails
- No more Matlab or IDL, escape while you can!

# *My transition timeline*

## September 27: Defense

# November 6: Apply to Insight



*Keep in touch!*

**www.astro.yale.edu/nth4**

**Linkedin.com/in/nhungho**

**Twitter.com/nhungkia**

**Email: nhung.t.ho@gmail.com**