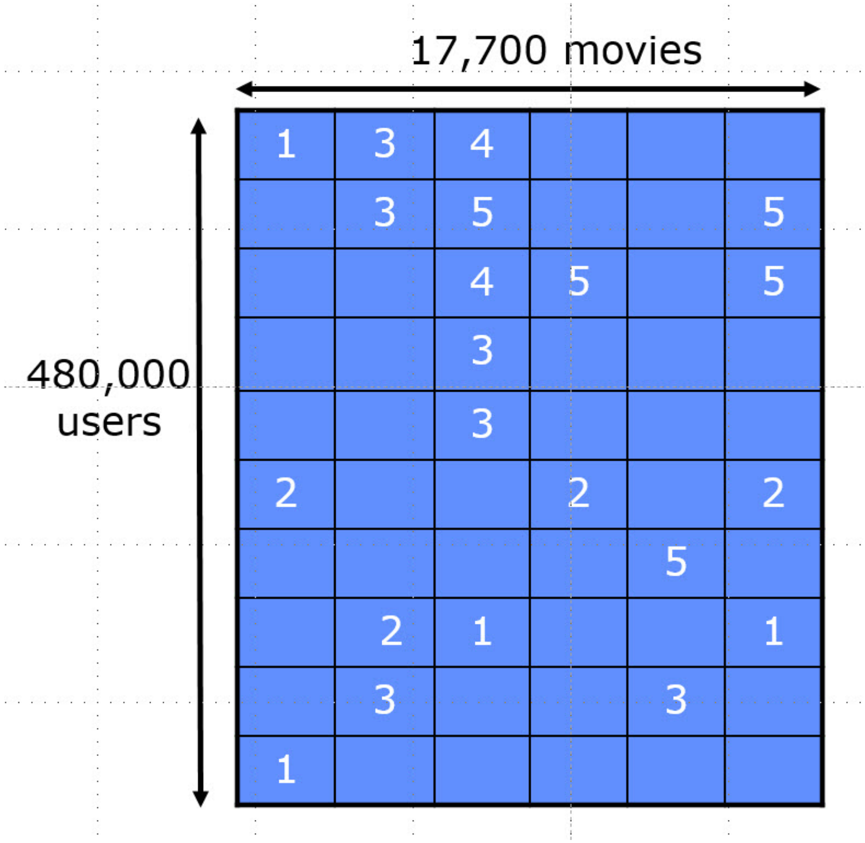
SparkR on AWS

Andy Catlin

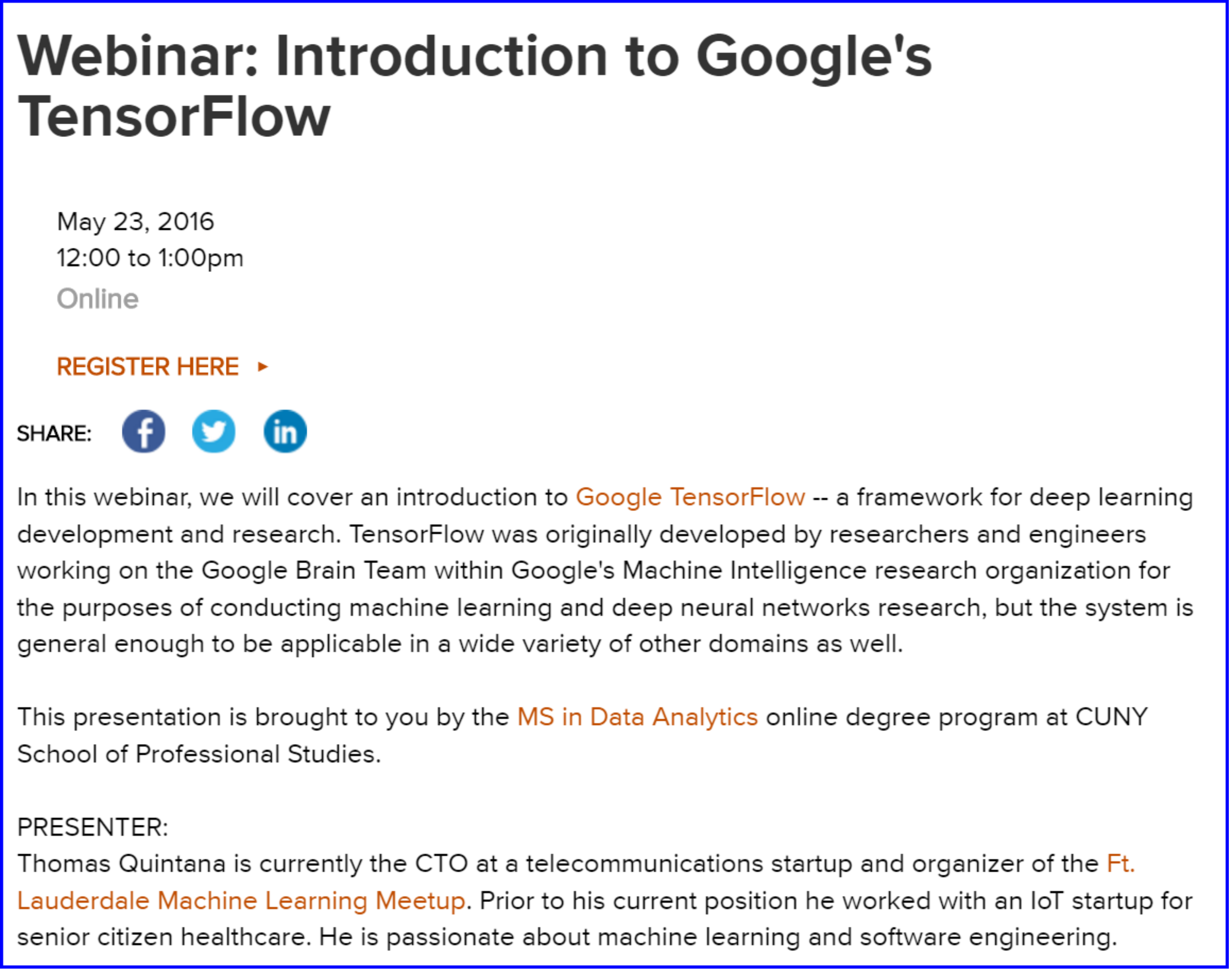
Netflix Prize



1	3	4			
	3	5			5
		4	5		5
		3			
		3			
2			2		2
				5	
	2	1			1
	3			3	
1					

100 million ratings.

Source: http://www.ics.uci.edu/~newman/courses/cs277/slides/netflix_overview.ppt.
See also: “The R language: Fast and efficient partial singular value decomposition of large data,” https://www.youtube.com/watch?v=ipkuRqYT8_I, 6 min YouTube video.



Webinar: Introduction to Google's TensorFlow

May 23, 2016
12:00 to 1:00pm
Online

REGISTER HERE ▸

SHARE:

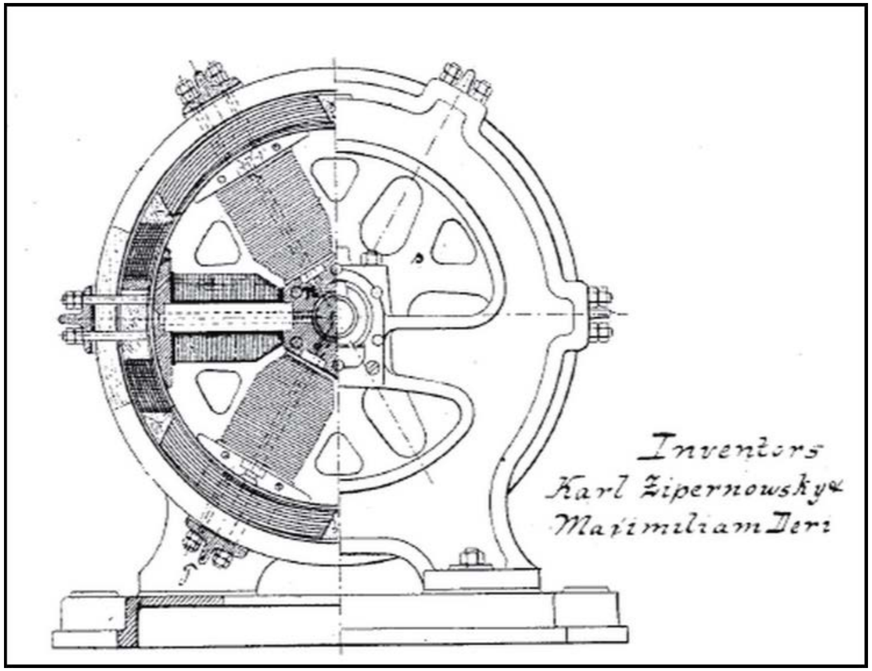
In this webinar, we will cover an introduction to Google TensorFlow -- a framework for deep learning development and research. TensorFlow was originally developed by researchers and engineers working on the Google Brain Team within Google's Machine Intelligence research organization for the purposes of conducting machine learning and deep neural networks research, but the system is general enough to be applicable in a wide variety of other domains as well.

This presentation is brought to you by the MS in Data Analytics online degree program at CUNY School of Professional Studies.

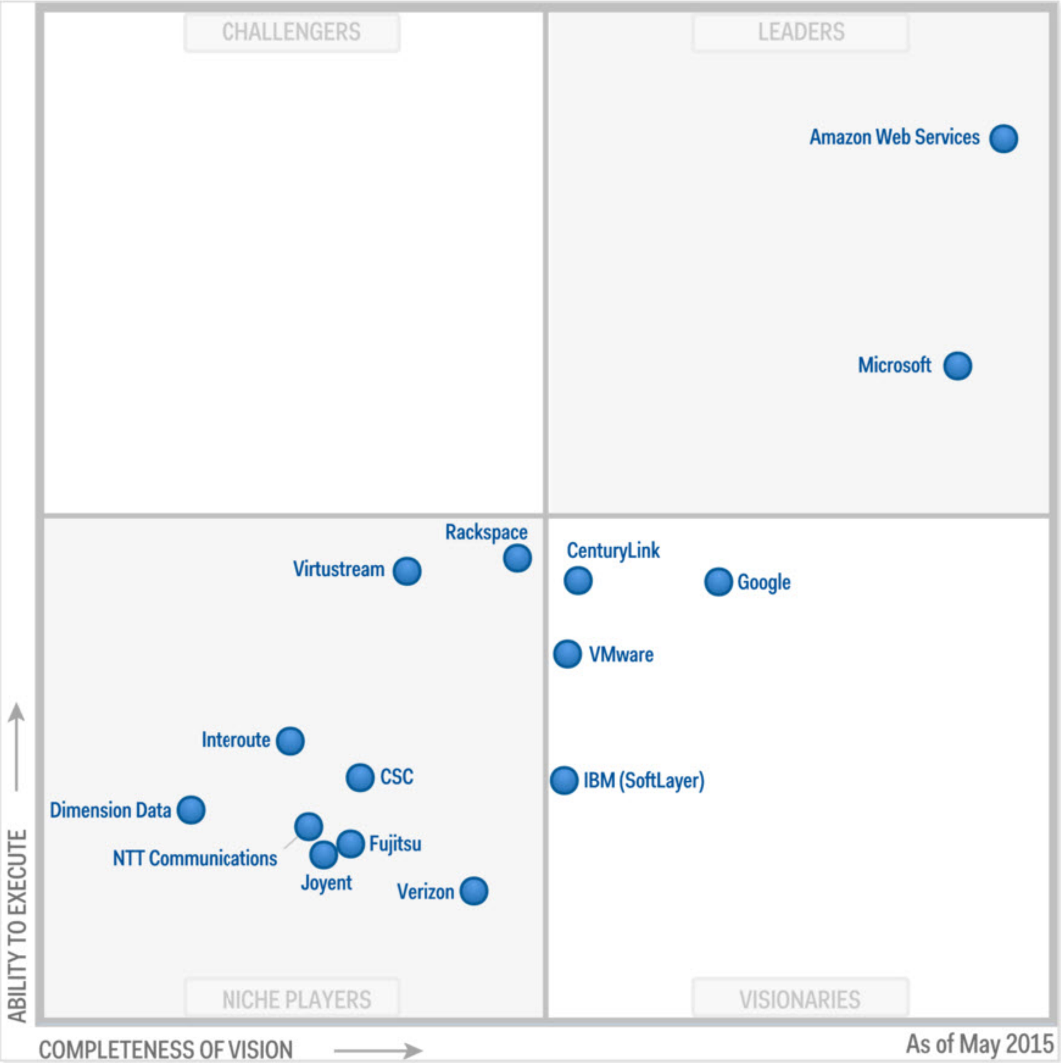
PRESENTER:
Thomas Quintana is currently the CTO at a telecommunications startup and organizer of the Ft. Lauderdale Machine Learning Meetup. Prior to his current position he worked with an IoT startup for senior citizen healthcare. He is passionate about machine learning and software engineering.

https://sps.cuny.edu/about/events/webinar-introduction-googles-tensorflow

"Cloud computing today is what the electric grid was 100 years ago."
-- Nicholas Carr, *The Big Switch*



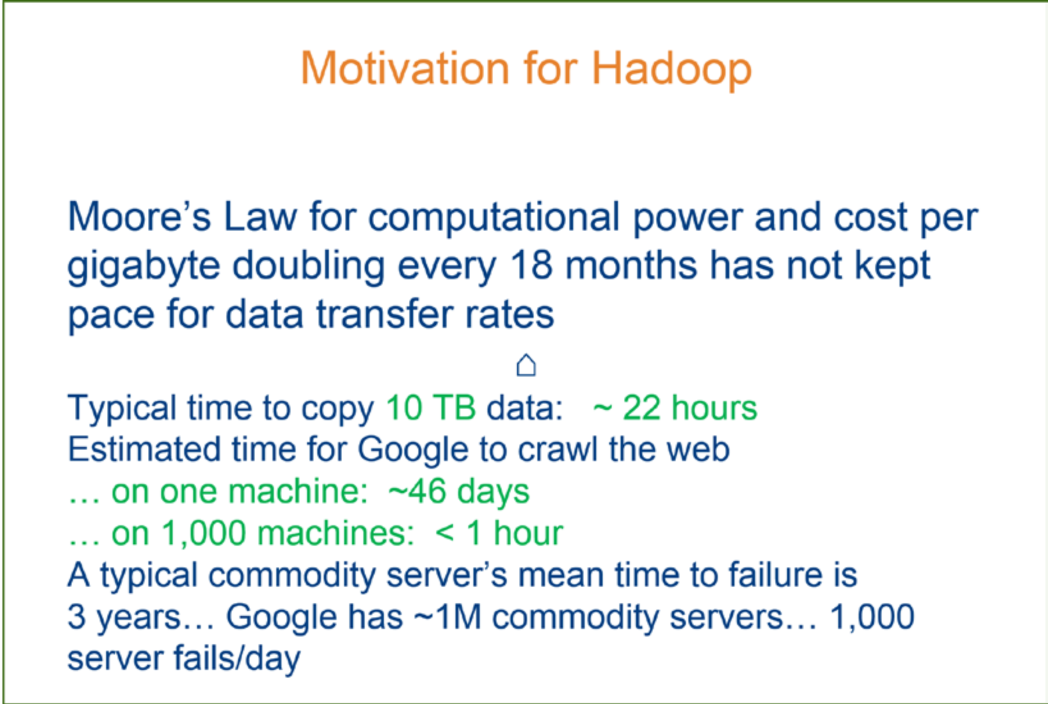
Which is the Best Cloud Service Provider?



Source: Gartner Magic Quadrant for Cloud Service Providers, May 2015. https://aws.amazon.com/resources/gartner-2015-mq-learn-more/

> “Any **distributed computing framework** needs to solve two problems: how to **distribute data** and how to **distribute computation**.”

source: “Getting Started with Spark (in Python),” Benjamin Bengfort,
https://districtdatalabs.silvrback.com/getting-started-with-spark-in-python



Motivation for Hadoop

Moore’s Law for computational power and cost per gigabyte doubling every 18 months has not kept pace for data transfer rates

⌂

Typical time to copy 10 TB data: ~ 22 hours
Estimated time for Google to crawl the web
… on one machine: ~46 days
… on 1,000 machines: < 1 hour
A typical commodity server’s mean time to failure is 3 years… Google has ~1M commodity servers… 1,000 server fails/day

What would be your design criteria for an environment to address these bottleneck issues?

Where to Get Hadoop and Spark

Hadoop Distributions

- HortonWorks
- Cloudera
- MapR

Spark

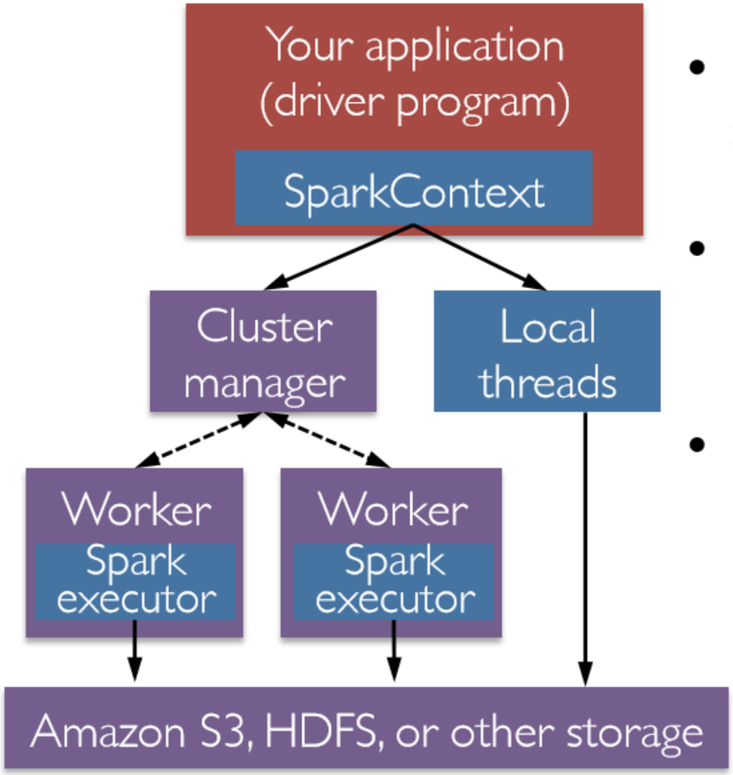
- http://spark.apache.org/

Use Cases

- Yelp.   400GB/log added per day.
- New York Times generated 11M pdfs for $240.

Self-Service, Prorated Supercomputing Fun!, Derek Gottfried, http://open.blogs.nytimes.com/2007/11/01/self-service-prorated-super-computing-fun, Nov 1,2007; http://www.roughtype.com/?p=1189

Spark Driver and Workers



- A Spark program is two programs:
  - » A driver program and a workers program
- Worker programs run on cluster nodes or in local threads
- RDDs are distributed across workers

Source: databricks

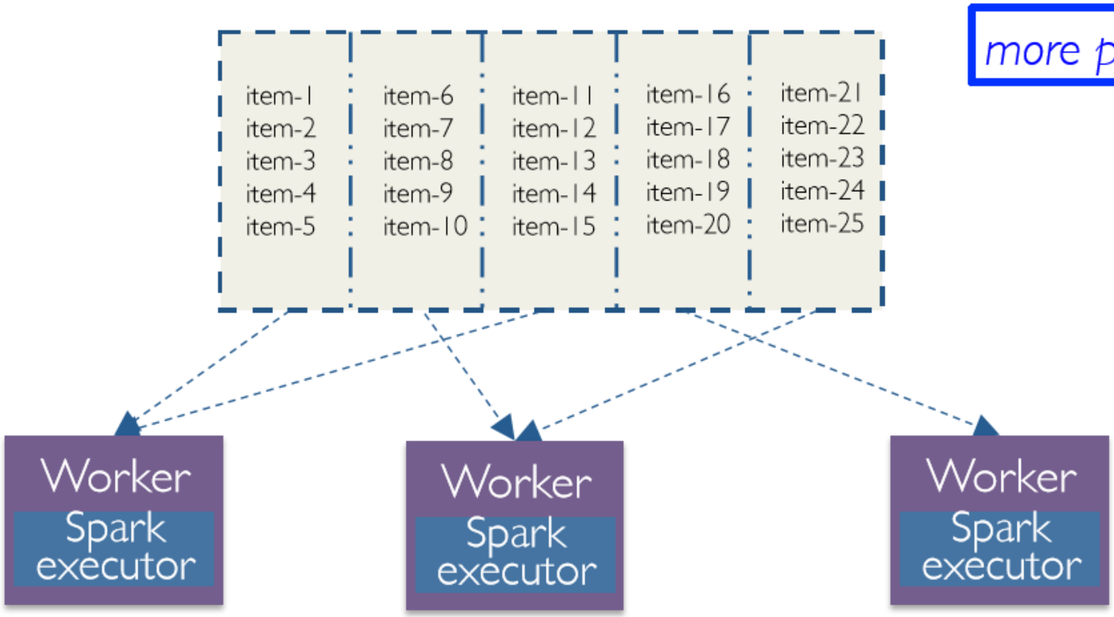
RDDs

- Programmer specifies number of partitions for an RDD

(Default value used if unspecified)

RDD split into 5 partitions

more partitions = more parallelism



item-1	item-6	item-11	item-16	item-21
item-2	item-7	item-12	item-17	item-22
item-3	item-8	item-13	item-18	item-23
item-4	item-9	item-14	item-19	item-24
item-5	item-10	item-15	item-20	item-25

Source: databricks

Where to Run Hadoop / Spark

In the Cloud

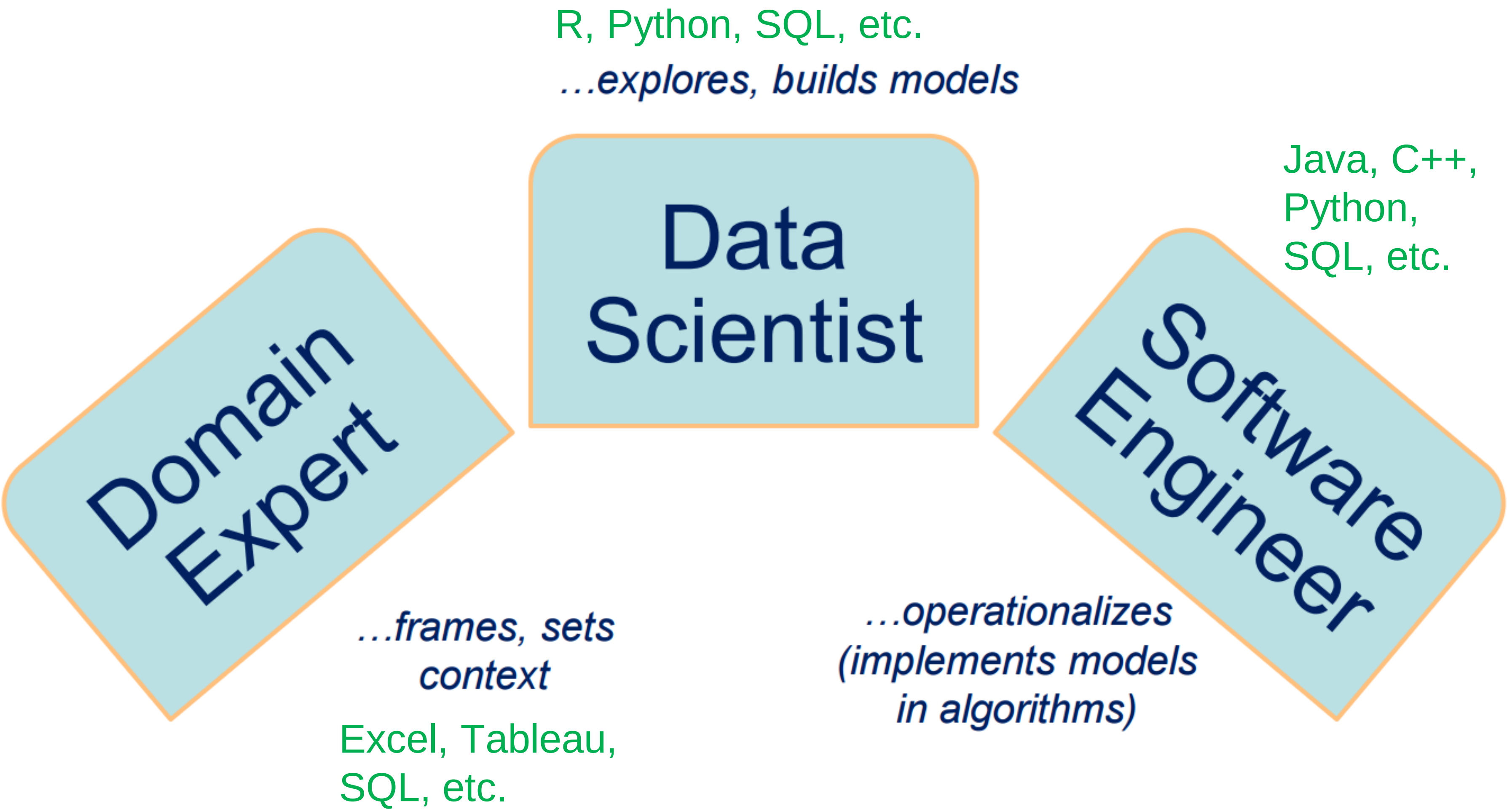
- Amazon, Microsoft Azure
- Other providers (IBM, Google, Rackspace,…)

In Your Enterprise

- Starting point: 4-6 repurposed machines

On Your desktop in a sandbox

- Can be a better dev environment (e.g. much easier to set up first class debugging)



R, Python, SQL, etc.

…explores, builds models

Data Scientist

Domain Expert

…frames, sets context

Excel, Tableau, SQL, etc.

Software Engineer

Java, C++, Python, SQL, etc.

…operationalizes (implements models in algorithms)

SparkR on AWS

“Supercharge R with Spark: Getting Apache's SparkR Up and Running on Amazon Web Services (AWS),” Manuel Amunategui, http://amunategui.github.io/sparkr/. Sep 30, 2015. Excellent Mac-based step-by-step tutorial.

Note that Mac’s bash shell has ssh and scp utilities; Windows users may want to substitute putty for ssh.exe and pscp.exe (included with putty) **for scp.** “Using Putty to Connect to an Amazon EC2,” Dan Morrill, https://www.youtube.com/watch?v=8Dsq4MeVh8M. Jan 9, 2013. [3 minute video]; "Connecting to Your Linux Instance from Windows Using PuTTY," http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html