Data Science with R

INTRODUCTION TO DATA SCIENCE

Matthew Renze

DATA SCIENCE CONSULTANT

@matthewrenze   www.matthewrenze.com

The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, … because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

The Economist

Gordon Brown's pitch
What went wrong at RBS
Genetically modified crops blossom
The EU woos Russia
The right to eat cats and dogs

The data deluge

AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

The New York Times

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

TWITTER
LINKEDIN
COMMENTS (58)
SIGN IN TO E-MAIL

Data Scientist:
The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you
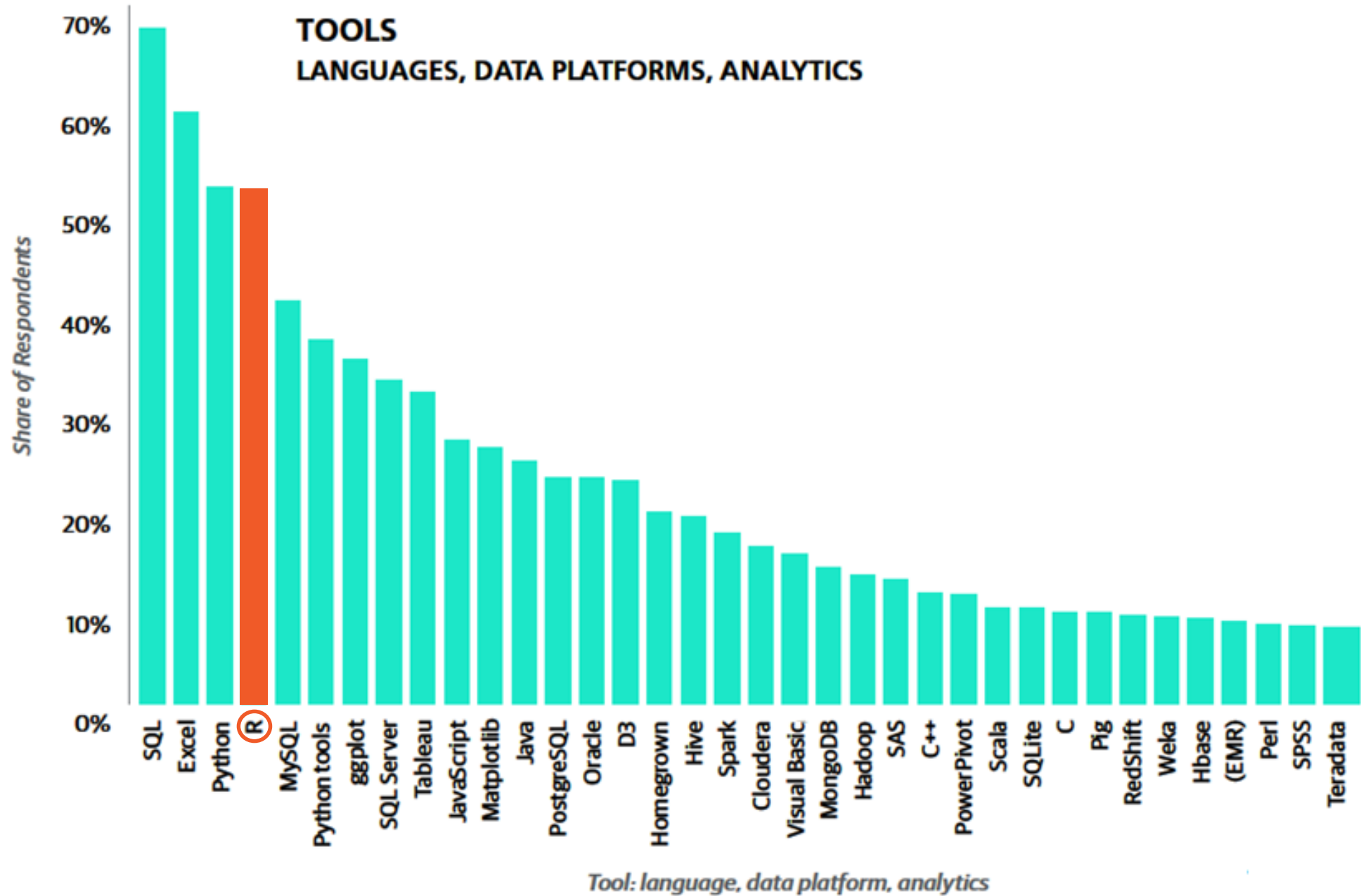
Job Trends from Indeed.com

— "Data Scientist"

Percentage of Matching Job Postings

0.04

0.02

0

Jan '06 Jan '07 Jan '08 Jan '09 Jan '10 Jan '11 Jan '12 Jan '13 Jan '14 Jan '15

AVERAGE SALARY FOR High Paying Skills and Experience

SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%
SOX (Sarbanes-Oxley)	$ 107,697	4.8%
Jetty	$ 107,406	0.4%
UML (Unified Modeling Language)	$ 107,387	4.7%

Source: Dice Salary Survey 2014

TOOLS
LANGUAGES, DATA PLATFORMS, ANALYTICS

Share of Respondents

70%
60%
50%
40%
30%
20%
10%
0%

SQL · Excel · Python · R · MySQL · Python tools · ggplot · SQL Server · Tableau · JavaScript · Matplotlib · Java · PostgreSQL · Oracle · D3 · Homegrown · Hive · Spark · Cloudera · Visual Basic · MongoDB · Hadoop · SAS · C++ · PowerPivot · Scala · SQLite · C · Pig · RedShift · Weka · Hbase · (EMR) · Perl · SPSS · Teradata

Tool: language, data platform, analytics

Source: O'Reilly 2015 Data Science Salary Survey

Course Overview

Introduction to Data Science

Introduction to R

Working with Data

Creating Descriptive Statistics

Creating Data Visualizations

Creating Statistical Models

Handling Big Data

Predicting with Machine Learning

Deploying to Production

Purpose

Data Science with R
(high-level introduction)

In-depth course 1

In-depth course 2

In-depth course 3

In-depth course 4

Audience

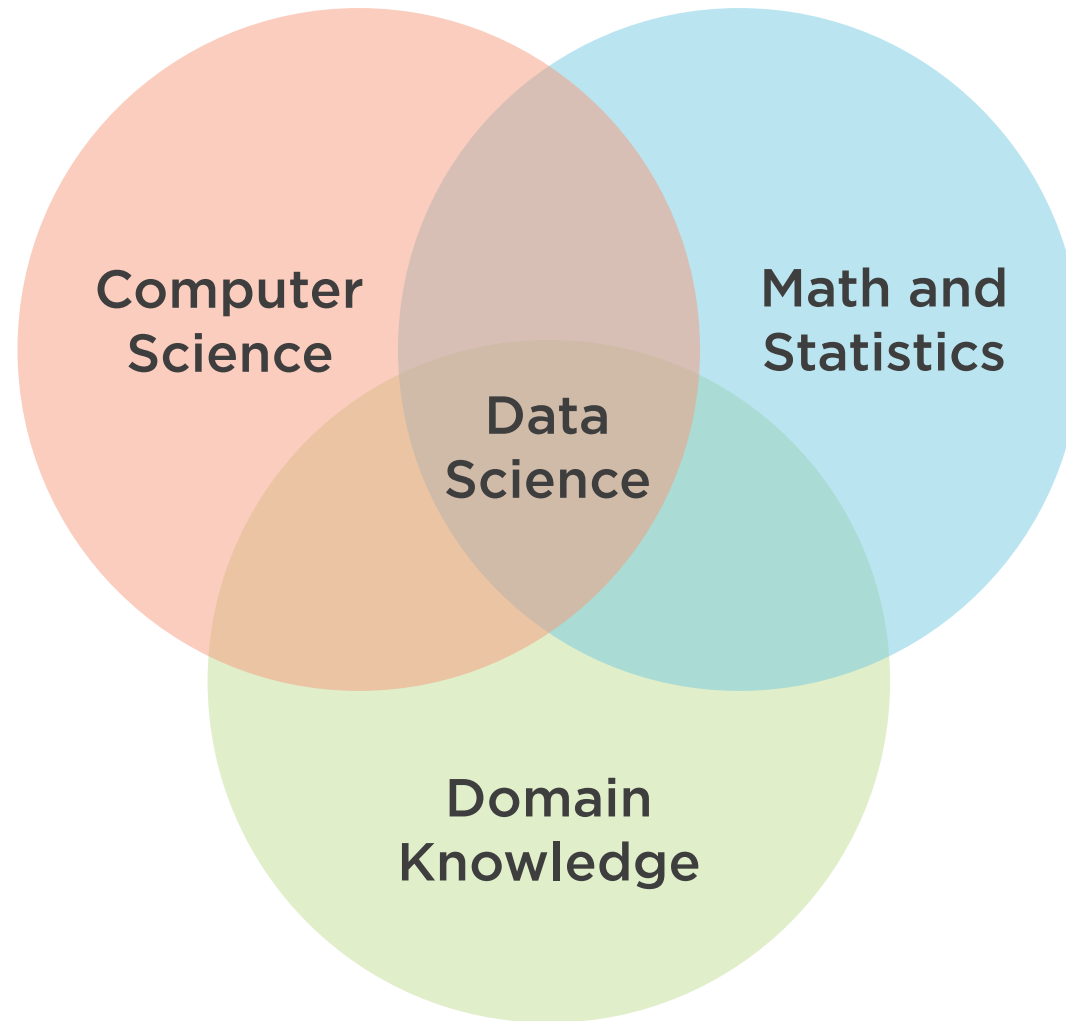|		
Developers	**Data analysts**	**IT professionals**

Prerequisites

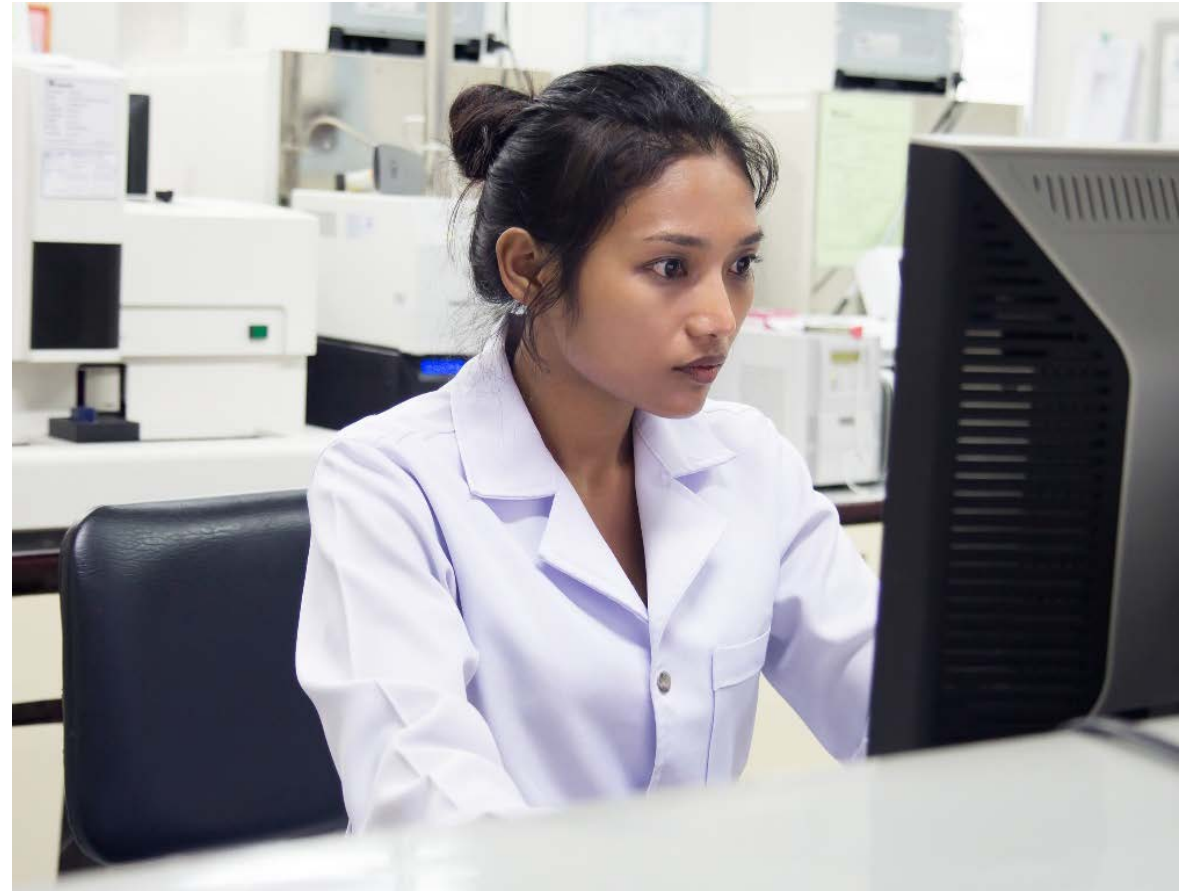**Basic programming experience**

Basic statistics experience

What Is a Data Scientist?

Performs data science

Proper accreditations

More than a scientist

More than an analyst

More than a developer

Job Trends from Indeed.com

— "Data Scientist"

Data Science Skills

Programming

Working with data
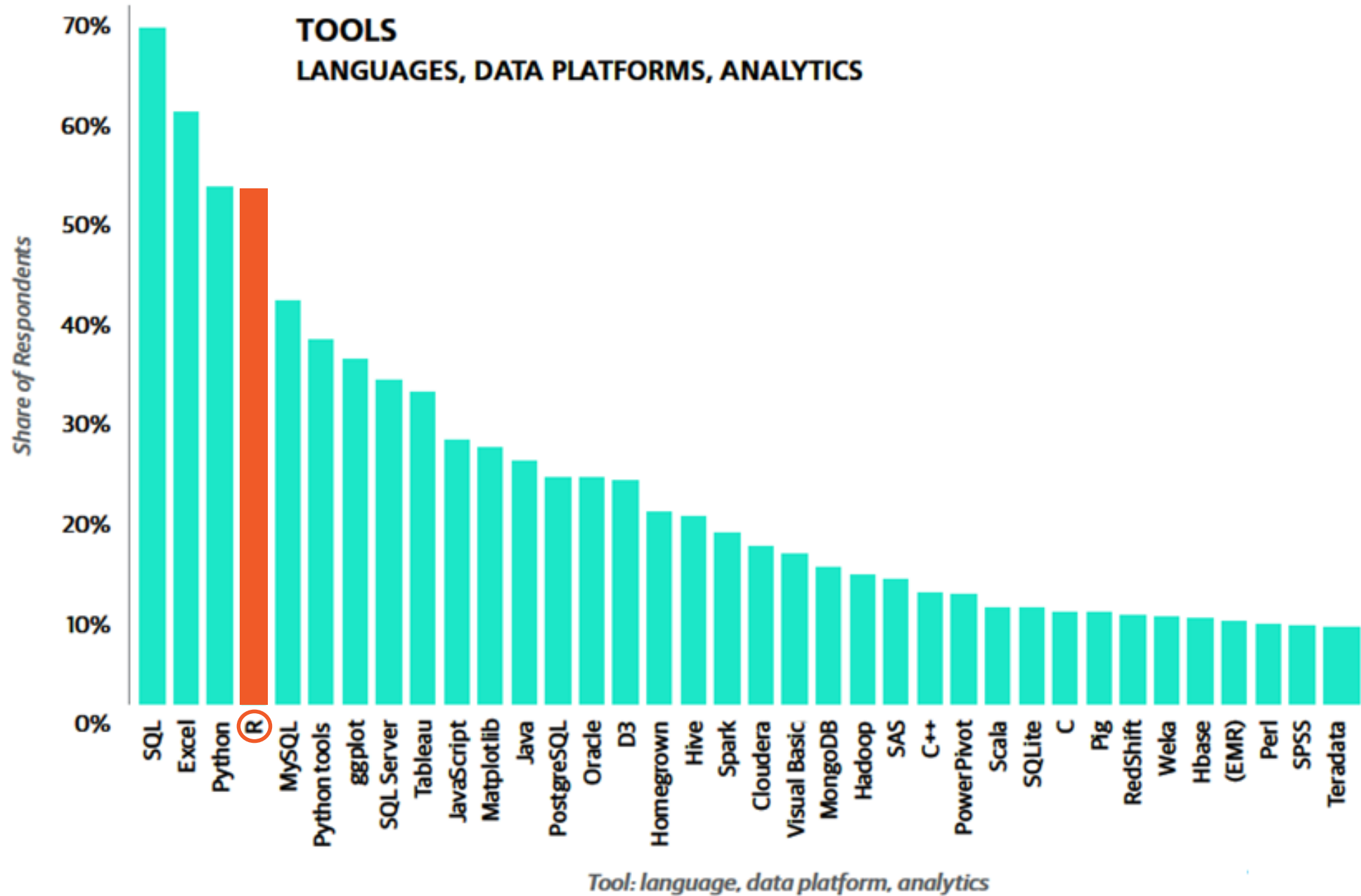
Descriptive statistics

Data visualization

Statistical modeling

Handling Big Data

Machine learning

Deploying to production

TOOLS
LANGUAGES, DATA PLATFORMS, ANALYTICS

Share of Respondents

70%
60%
50%
40%
30%
20%
10%
0%

SQL, Excel, Python, R, MySQL, Python tools, ggplot, SQL Server, Tableau, JavaScript, Matplotlib, Java, PostgreSQL, Oracle, D3, Homegrown, Hive, Spark, Cloudera, Visual Basic, MongoDB, Hadoop, SAS, C++, PowerPivot, Scala, SQLite, C, Pig, RedShift, Weka, Hbase, (EMR), Perl, SPSS, Teradata

Tool: language, data platform, analytics

Source: O'Reilly 2015 Data Science Salary Survey

Why Is Data Science Important?

Internet of Things

Big Data

Machine learning

Fully autonomous systems

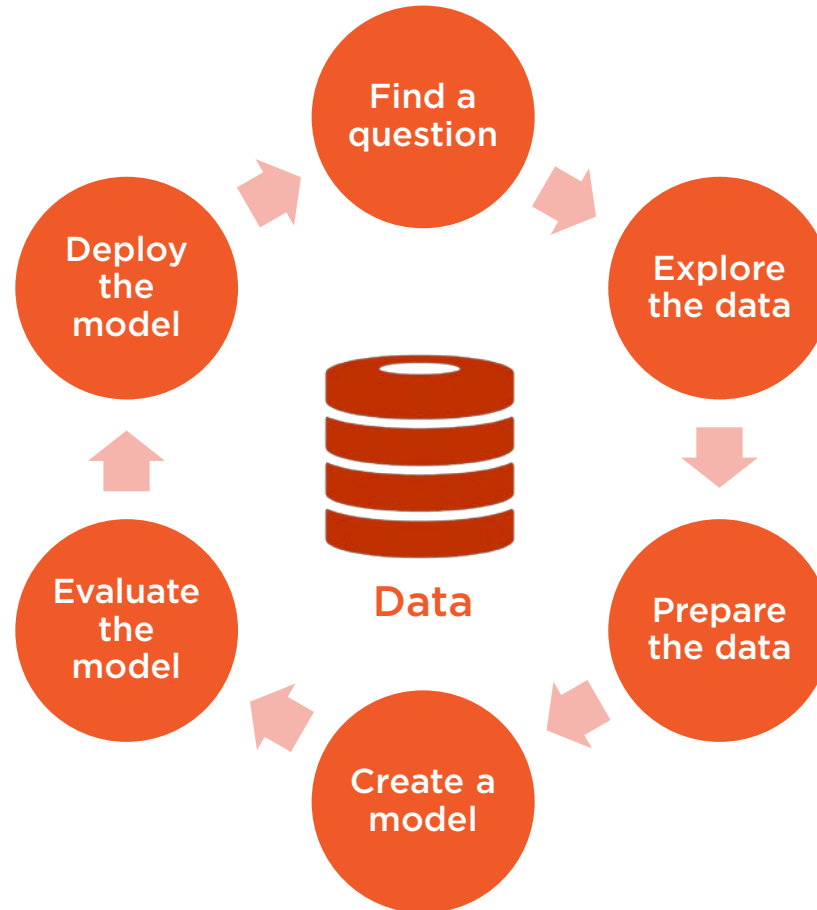# Why Is Data Science Important?

Driven by economics

Possible by technology
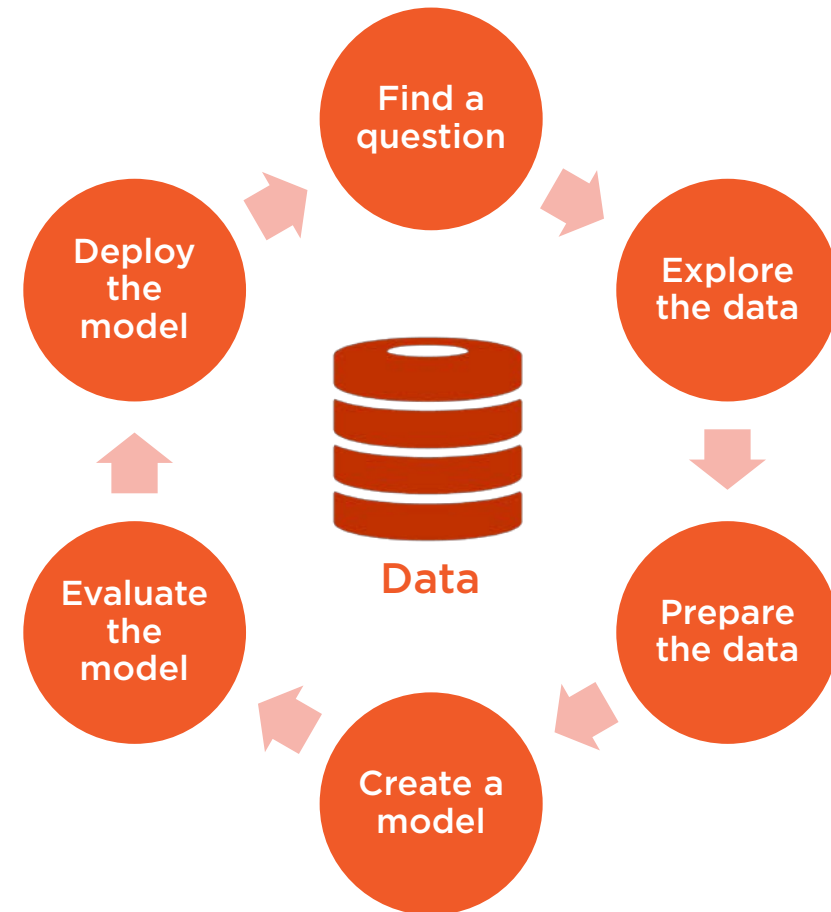
Cost is decreasing

Value is increasing

The Data Science Process

The Data Science Process

Iterative process

Non-sequential

Early termination

Established processes

Summary

Introduction

Course Overview

Introduction to Data Science