CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2016
Doc 1 Introduction
Aug 30, 2016

Course Issues

http://www.eli.sdsu.edu/courses/index.html

Waitlist
Course Web Site
Wiki
Screencasts
Prerequisites
Grading
Books
Julia
Hadoop & Spark
Data Science

Waitlist - How to get into a Class

Add yourself to the course waitlist

Instructors can not

- Add individuals to the class
- See who is on the waitlist
- Change your priority on the waitlist

Waitlist - How it works

Waitlist is a priority queue

When a seat in a class becomes available the top priority student is added

You can not be enrolled in two classes that meet at the same time

If wait list system adds you to a class, it will drop you from classes that meet at the same time

First week of classes as students drop others are added

Second week of classes students are only added if instructor releases the seats

Waitlist FAQ

Why not get a bigger room and admit everyone?

No first hard assignment to scare people

No Grader

Do you really want a 600 level class of 100 people?

Waitlist FAQ

Will you be increasing the size of the class?

No

Why not?

No grader

New courses are a lot of work

Techology courses are a lot of work

Waitlist FAQ

Sept 12

Last day for regular students to add/drop classes

Open University students have lower priority than SDSU students

Waitlist FAQ

But one more student will not be much work

I already added 11 student more

There are 48 students behind you that will ask the same question

So the question really is why not add 60 more students to a class of 49

Waitlist FAQ

But

I did not have to register for the class

If the class would have started at 60 I would have gotten in

If you would have expanded the class earlier I would have gotten in

If you would have expanded the class later I would have gotten in

This class important for my job

I am really interested in this class

Still not expanding the class size

Waitlist FAQ

So what are my chances of adding this class?

Look up your position on the waitlist

What are the odds of that many people dropping the class

I can not see the waitlist

I have no idea how many people will drop

Grading

2 exams

4-6 assignments

Course Website Demo

What are the Tools & Methods?

Programming language
Programming Notebook

Visualization
- scatter, box, violin, qq, line, density plots
- errorbar, histogram, beeswarms

Statistics
- mean, variance, quantiles, distributions
- confidence intervals, correlation, coveriance
- regression, goodness-of-fit, chi-squared test
- Bayes theorem

Machine Learning
- k-means, DBSCAN, Decision & Regression trees

Hadoop, Spark, Pig, Mahout, etc.

What will be be doing

Installing programs

Julia, Juno, Jupyter, Hadoop, Spark

Writing Julia programs

Reports using Jupyter Notebooks

Analyzing data

Distributing data

Visualizing Data

Using Hadoop & Spark

Using Amazon Cloud

What will be be doing

First half of semester

- Julia
- Statistics
- Visualization
- Machine Learning

Second half of semester

- Hadoop
- Spark
- Pig, etc

Experimental Course

First time offered

Cross discipline

Technology Based

Going to be some rough edges

Prerequisites

You will be installing software

- Julia
- Jupyter
- Juno (Atom)
- Hadoop
- Spark

Some of these are more complex on Windows than Unix/Mac OS

We will be doing some

- Statistics
- Math
- Machine learning

Tasks - Install the Following

Julia 0.4.6

http://julialang.org/

Juno

http://junolab.org/

Jupyter 4.1 via Anaconda & Conda

http://jupyter.readthedocs.io/en/latest/install.html

Hadoop 2.7.2

Unix/Linux/Mac OS

http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SingleCluster.html

Windows http://wiki.apache.org/hadoop/Hadoop2OnWindows

Books

Course books are available for free on-line via SDSU library

Need SDSU Library account to access books off campus

Some people do not like reading books on-line

But if you need to save money it is available

May add chapters of other books as semester progresses

But on-line from books available on-line

Hadoop, Spark, Amazon

You will run Hadoop & Spark on Amazon's cloud

You need to create an Amazon AWS account

Amazon Free Tier should handle most of the expense

But you may incur some cost on Amazon

Data Science & Big Data

Very trendy

When topics become trendy in CS the terms become very vague

Big Data Analytics with Excel

Is Data Scientist A Useless Job Title?

Data Science

Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics,[3] similar to Knowledge Discovery in Databases (KDD)

Wikipedia

Data Science

Data Scientist (n.):

> Person who is better at statistics than any software engineer and better at software engineering than any statistician.
>
> — Josh Wills (@josh_wills) May 3, 2012

Data Engineer

A software engineer that deals with data plumbing

Traditional database setup, Hadoop, Spark, etc.

Data analyst

A person who digs into data to surface insights,
but lacks the skills to do so at scale
They know how to use

Excel, Tableau and SQL

but can’t build a web app from scratch

Source: https://yanirseroussi.com/2014/10/23/what-is-data-science/

Data Science

Science of transforming data into useful information by means of

- Statistical and
- Machine learning techniques

Data Science & Big Data

Big Data

Data Science with large datasets

No hard boundary between Big Data and medium data

Requires more data plumbing

Inconvenient Truth About Data Science

Data is never clean.

You will spend most of your time cleaning and preparing data.

95% of tasks do not require deep learning.

In 90% of cases generalized linear regression will do the trick.

Big Data is just a tool.

You should embrace the Bayesian approach.

No one cares how you did it.

Academia and business are two different worlds.

Presentation is key - be a master of Power Point.

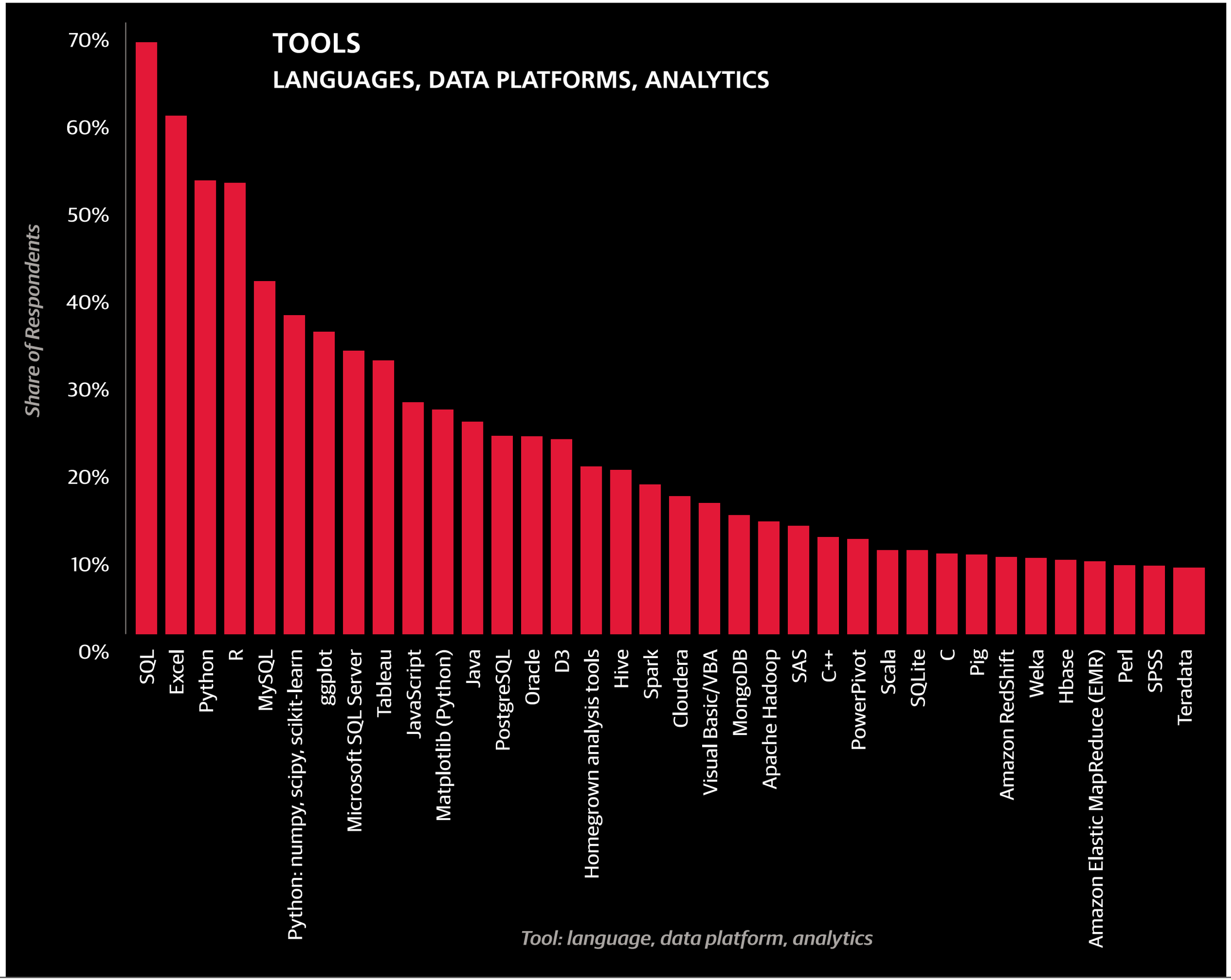
All models are false, but some are useful.

There is no fully automated Data Science. You need to get your hands dirty.

Source: https://www.linkedin.com/pulse/inconvenient-truth-data-science-kamil-bartocha

TOOLS

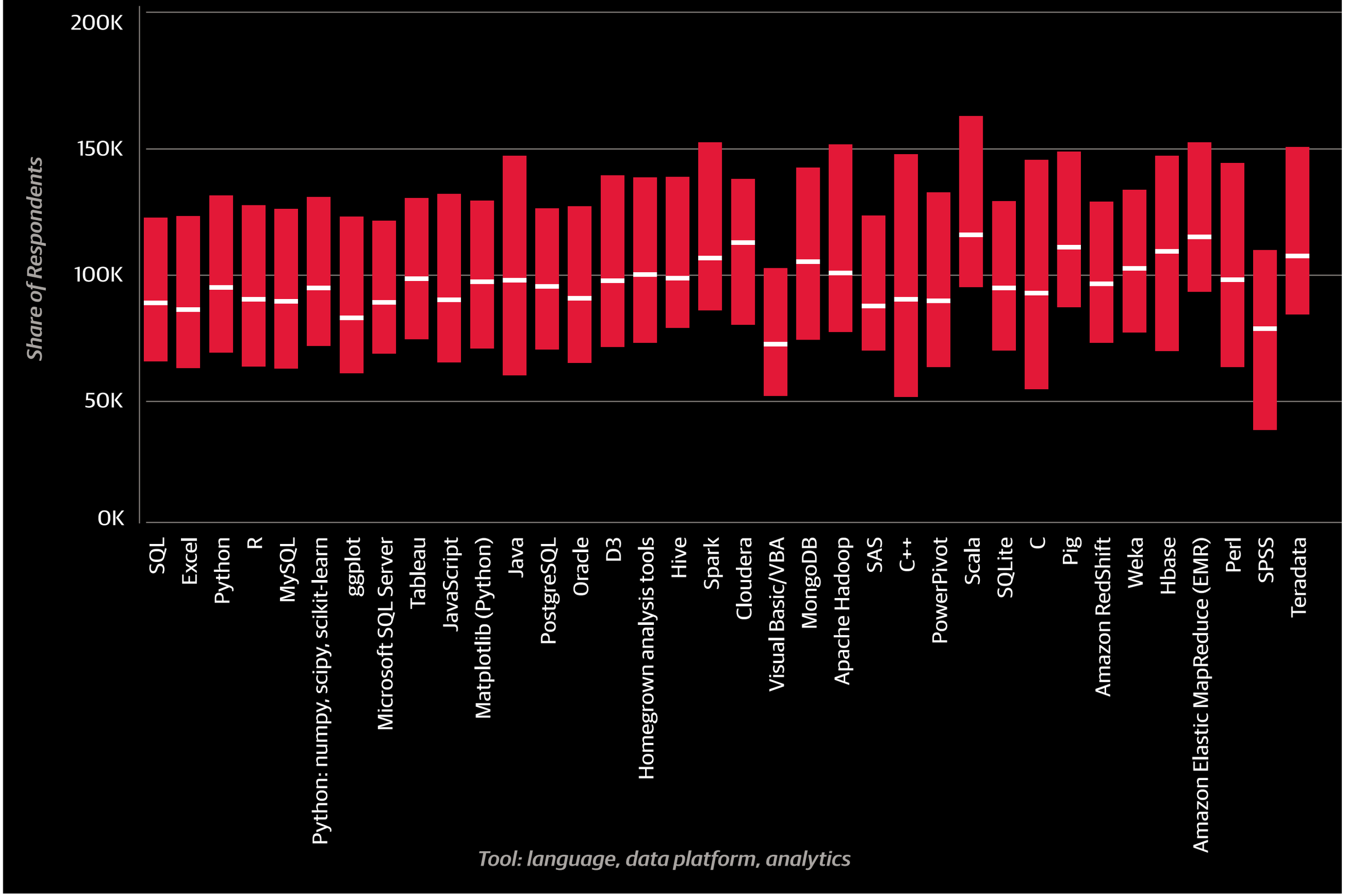
LANGUAGES, DATA PLATFORMS, ANALYTICS



Source: 2015 Data Science Salary Survey, King & Magoulas, O'Reilly

TOOLS: LANGUAGES, DATA PLATFORMS, ANALYTICS

SALARY MEDIAN AND IQR (US DOLLARS)



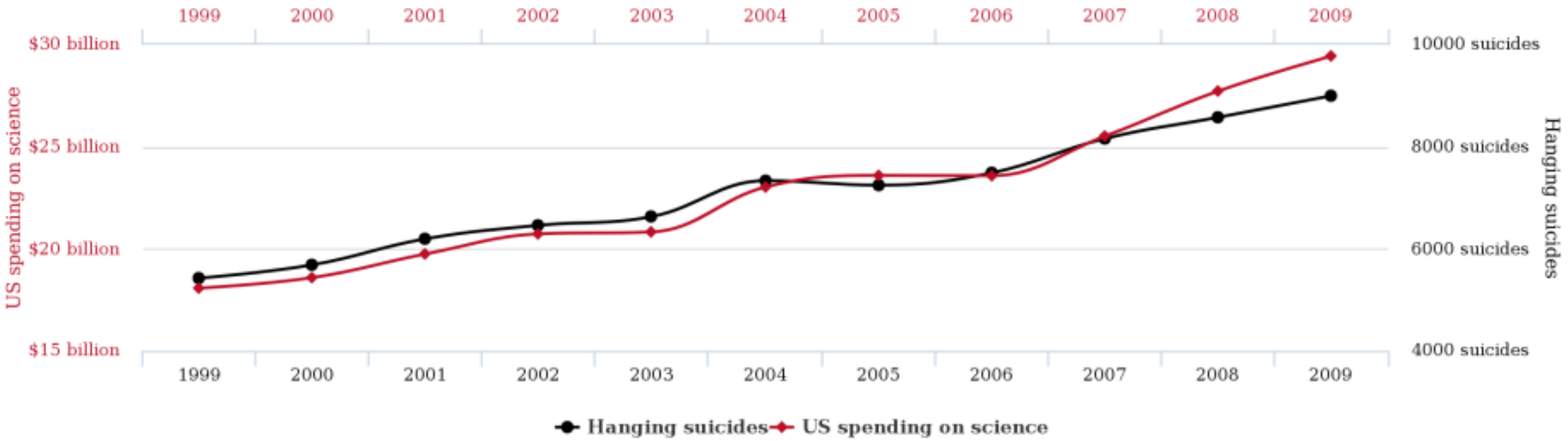
Monday, August 29, 16

Source: 2015 Data Science Salary Survey, King & Magoulas, O'Reilly

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



Monday, August 29, 16

Source https://yanirseroussi.com taken on 8/21/2016

Rule of Three

If you can not think of three things that might go wrong with your analysis there is something wrong with your thinking

Modified from Weinberg's Rule of Three

Data Science Verses Programming Jobs

Intuit Job Listing Worldwide Aug 22 2016

Data - 23

Software Engineer - 168

Data Science Programming Langauges

Python
R
Javascript
SAS
Perl
Matlab
Ruby

Scala
Julia

Java
C++
C
C#

Features of Langauage for Data Science

Interactive

Statistical, Machine Learning, Math libraries

Plays well with others

Supports computation

Simple syntax

Fast

Python

Wildly used

Interactive

Lots of libraries

Plays well with other

Slow

Python 2.x verses Python 3.x

- 3/2

Threads do not scale

- Global Interpreter Lock (GIL)

Julia

New language from MIT

Interactive & Fast

Untyped & Typed

Designed for computation

f(x) = 2x + 4

Int32, Int64, Int128, BigInt

Statistical and Math libraries

Plays well with others

LLVM

Lisp style macros

Multiple dispatch

Designed for parallelism & Distributed computation

Java, Scala, Hadoop, Spark

Hadoop written in Java
Spark written in Scala

JVM langauages (Java, Scala, Clojure, Groovy, JRuby, Jython)
- Much more efficient on Hadoop & Spark
- First access to new features

Scala
- OO & Functional
- Type inference
- Far less verbose than Java

Big Data

Data sets that are so large or complex that traditional data processing applications are inadequate

Wikipedia

Big Data 3-5 V's

Volume

- Large datasets

Velocity

- Real time or near-real time streams of data

Variety

- Different formats
- Structured, Numeric, Unstructured, images, email, etc.

Variability

- Data flows can be inconsistent

Veracity

- Accuracy

Complexity

Scaling to Handle Large Data Sets

Scaling up (Vertically)

- Add more resources to single machine
- Memory, disk space, faster processor, etc
- Easier that scaling out but limited

Scaling out (Horizontally)

- Using multiple machines/processors
- Adds complexity

Scaling Up & Amdahl's Law

T(1) be the time it takes a sequential program to run
T(N) be the time it takes a parallel version of the program to run on N processors.

Speedup using N processors

$$S(N) = T(1)/T(N)$$

Let p = % of program that can be parallelized

Amdahl's Law

$$S(N) = 1/(1 - p + p/N)$$

Amdahl's Law

Let p = % of program that can be parallelized

Amdahl's Law

$$S(N) = 1/(1 - p + p/N)$$

p = 1

S(N) = 1/(1 - 1 + 1/N)

= 1/(1/N)

= N

p = 0

S(N) = 1/(1 - 0 + 0/N)

= 1

Amdahl's Law

Let p = % of program that can be parallelized

Amdahl's Law

$$S(N) = 1/(1 - p + p/N)$$

Given p = 0.5 how many processors does in make sense to use?

What does p have to be to get a speedup of

- 5 or greater using 10 processors?
- 10 or greater using 20 processors?
- 20 or greater using 40 processors?
- 50 or greater using 100 processors?

Amdahl's Law Demo