CS 649 Big Data: Tools and Methods
Spring Semester, 2022
Doc 1 Introduction
Jan 19, 2021

Course Issues

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

Waitlist
Course Web Site
Wiki
Course Recordings
Prerequisites
This room
Grading
Books
Spark & Related Tools
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

Feb 1

Last day for regular students to add/drop classes

Last day to file for graduation

- May
- August

Office Hours

Tuesday & Thursday 10:30 am - Noon

Zoom: 914 283 418

Grading

1 exam

4-6 assignments

Project

Course Website Demo

What are the Tools & Methods?

Programming language - Python

Programming Notebook

Web Dashboards

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

Streaming - Kafka

Database - Cassandra

~~Hadoop~~, Spark, Pig, Mahout, etc.

What will you be doing

Installing programs

Python, Jupyter, Spark, Kafka, Cassandra

Writing Python, Java, Scala-Spark programs

Reports using Jupyter Notebooks

Create web dashboards

Analyzing data

Distributing data

Visualizing Data

Using Spark

Using Amazon Cloud

Notebooks - Documentation, development

Python, Julia, R,

Other supported by community - Java, Fortran, Haskell, Ruby, Go, Scala, many more

Other notebook systems

Visualization

Python, Julia, R, Matlab, dash, Steamlit

ML

~~Python~~ (C), Julia, Matlab, R?

Spark - Large Data Sets

Scala

Java

Python

R

Julia

Kafka - Streaming Data

Java

JVM languages

Python

Julia (Except for offsets)

Others - No R Client

Cassandra - Data Storage

Java

Python

R - sort of

Prerequisites

You will be installing software

- Python
- Jupyter
- Spark
- Kafka
- Cassandra
- Plotly
- Streamlit

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

Jupyter via Anaconda & Conda with Python 3
http://jupyter.readthedocs.io/en/latest/install.html

Spark 3.2, Prebuild for Apache Hadoop 3
Unix/Linux/Mac OS
http://spark.apache.org/docs/latest/

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

Spark 3.2

Time	Class	Units	Room
1730	649	3	M245
1900	662	3	SH123

Relational Database
- Table
- SQL

Data Science
- DataFrame
- API & language
- df[df.Time == 1730,!]

Python - Pandas
- DataFrame

Spark
- DataFrame

Implementation and API are different in
- Pandas & Spark

Spark 3.2
- Can use most of Panda DataFrame API

Spark also allows using SQL

What will we be doing

~2 Weeks

Intro, Python

~6 weeks

Statistics, ML, NumPy, SciPy

Visualization

Spark

~5 weeks

Kafka & Cassandra

Books

Python Data Science Handbook: Essential Tools for Working with Data
Jake VanderPlas
O'Reilly Media
December 10, 2016
ISBN 9781491912058

Spark: The Definitive Guide
Matei Zaharia, Bill Chambers
February 2018
ISBN 9781491912218

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

Spark, Amazon

You will run Spark on Amazon’s cloud

You need to create an Amazon AWS account

Sign up for Amazon Educate account - $100 compute time for free

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

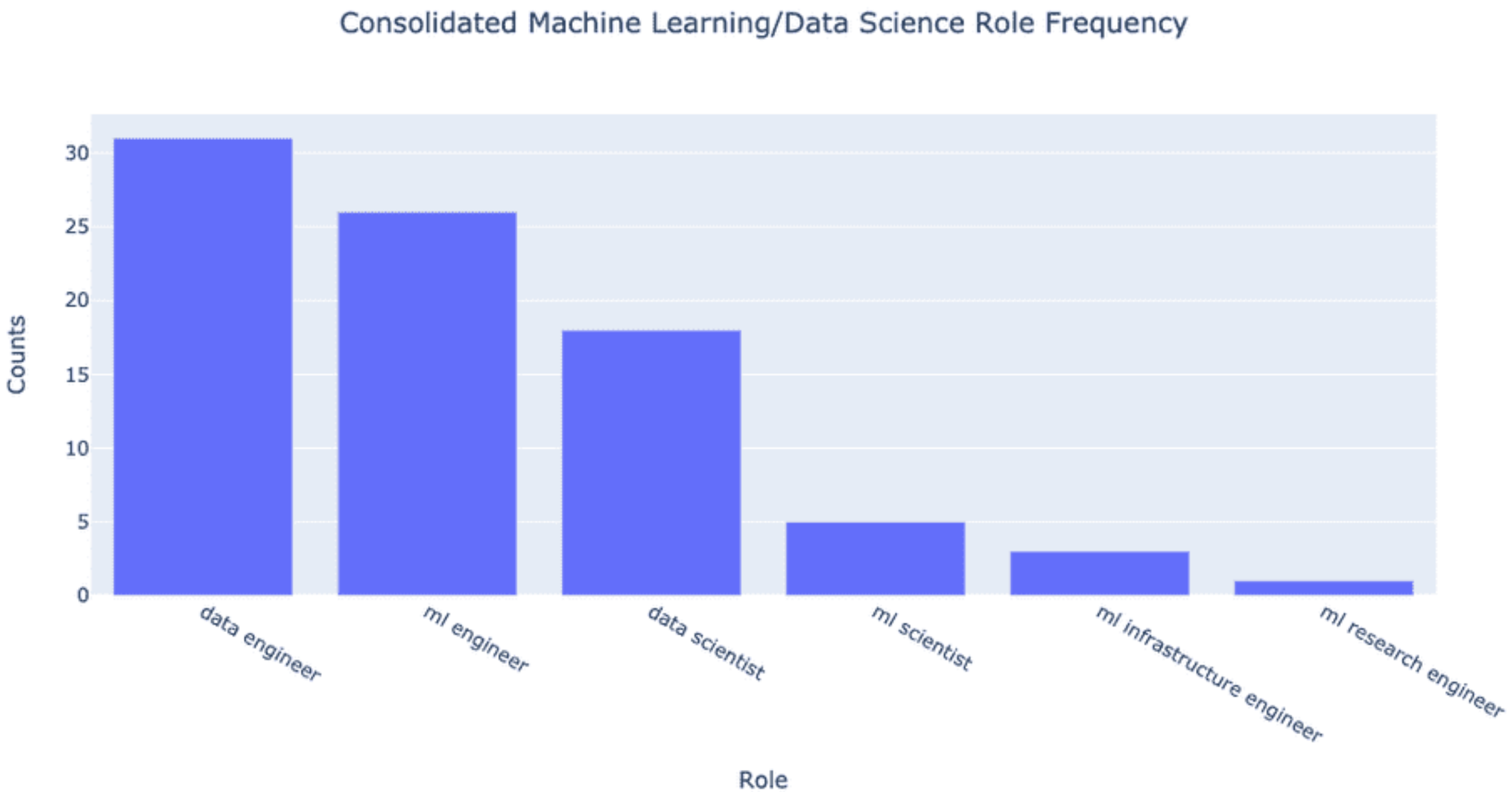
Excel, Tableau and SQL

but can’t build a web app from scratch

Survey of 1,400 Y-Combinator Companies

Mihail Eric *January 2021*

70% more open roles at companies in *data engineering* as compared to *data science*



https://www.mihaileric.com/posts/we-need-data-engineers-not-data-scientists/

Data science is different now

Vicki Boykis Feb 13, 2019

Glut of new data scientists

> number of candidates per any given data science position, particularly at the entry level, has grown from 20 or so per slot, to 100 or more. I was talking to a friend recently who had to go through 500 resumes for a single opening

UC Berkeley Data 8 Class

Data science is different now

Vicki Boykis Feb 13, 2019

Data science as a misleading job req

> it has about cleaning, shaping data, and moving it from place to place

"As someone titled 'data scientist' in 2019, I spend most of (60%+) my time:" ("Other") also welcome, add it in the replies.

Picking features/models	6.2%
Cleaning data/Moving data	67.3%
Deploying models in prod	3.6%
Analyzing/presenting data	22.9%

How To Become a Data Engineer

8 Jan. 2021

Median salary for Data Engineers in SF Bay Area is around $160k

Algorithms & Data Structures

SQL — the lingua franca for databases

Programming: Python, Java and Scala

The Big Data Tools

Cloud Platforms

Fundamentals of Distributed Systems

Data Pipelines

https://khashtamov.com/en/how-to-become-a-data-engineer/

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.

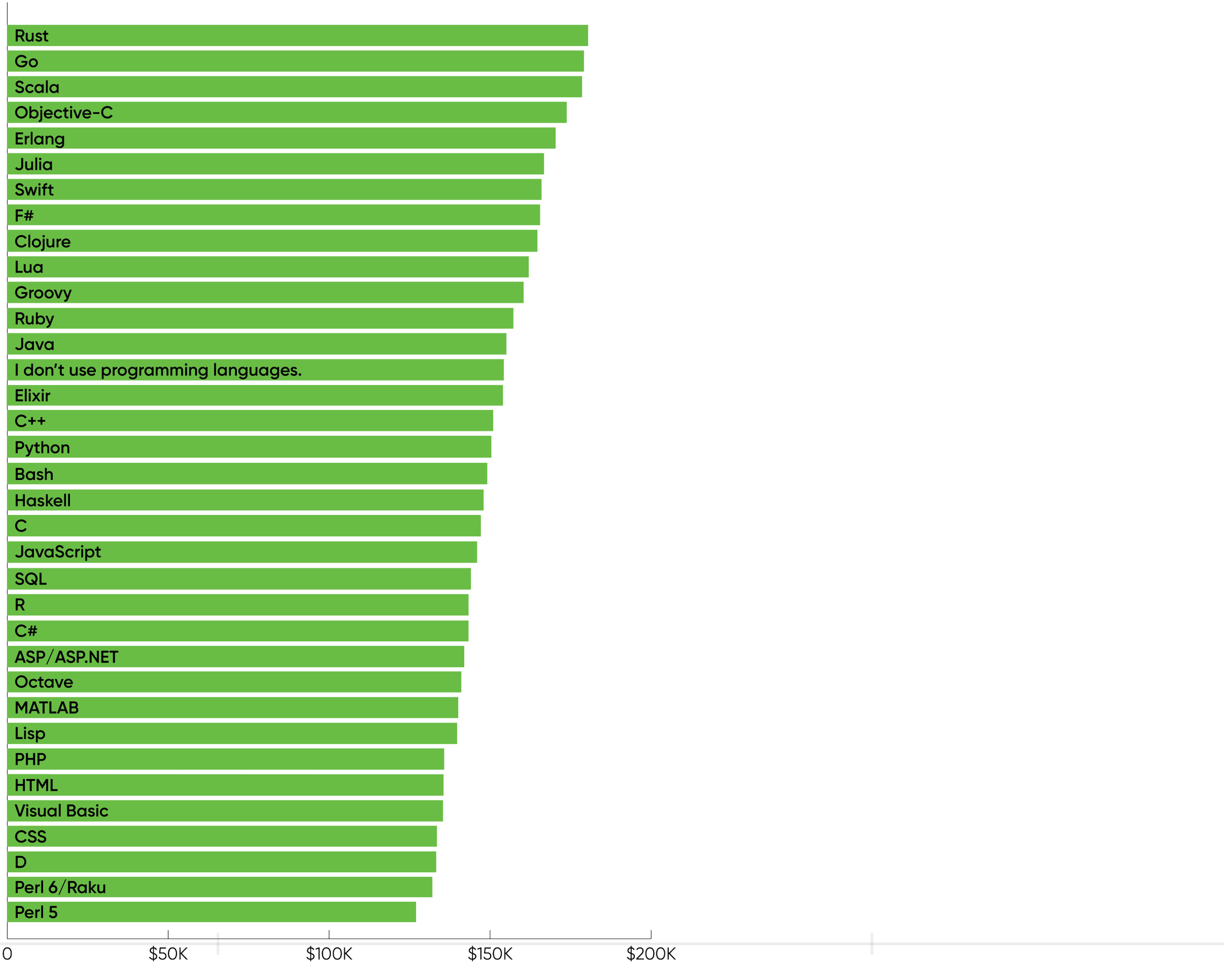
No one cares how you did it.

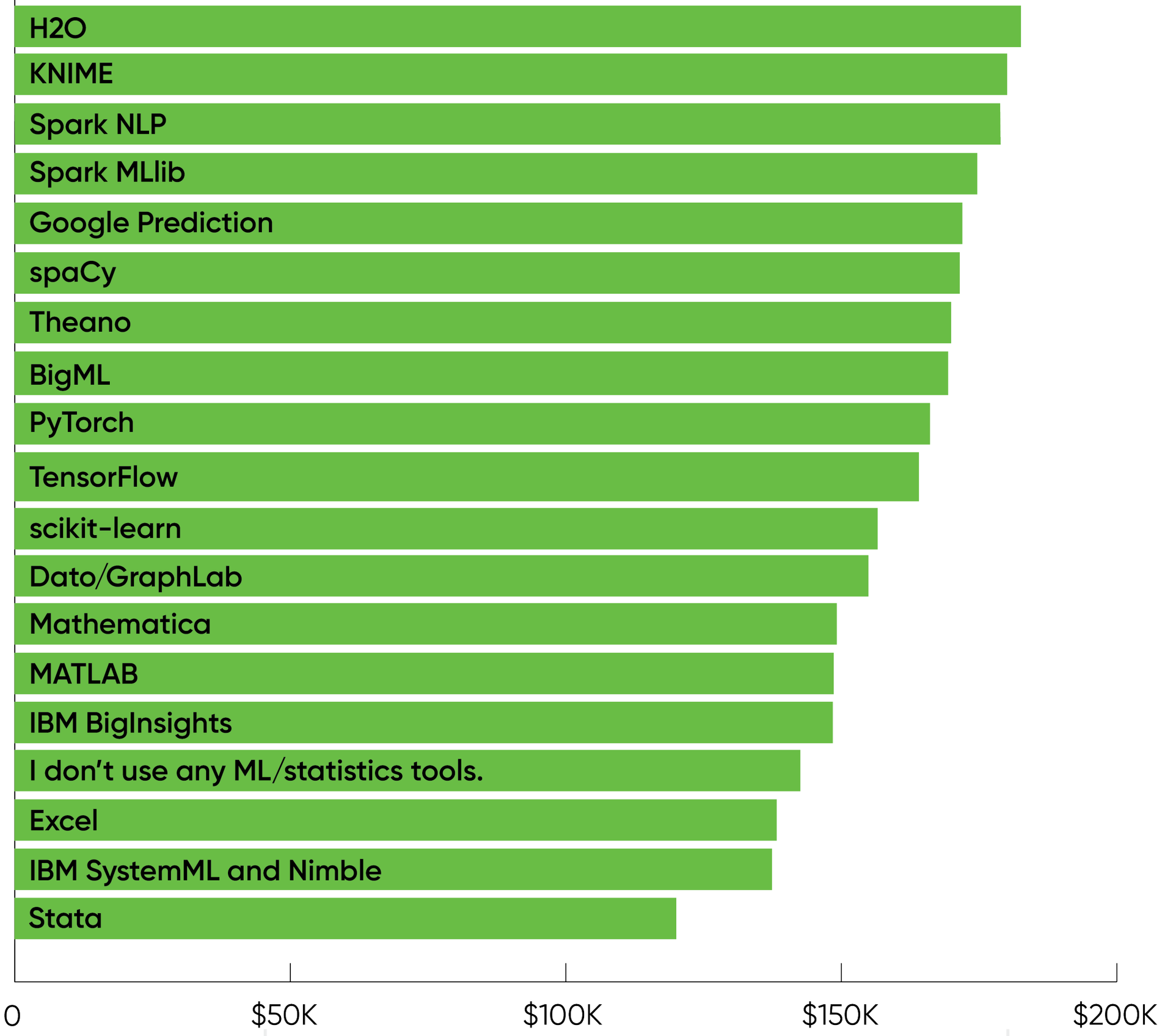
Academia and business are two different worlds.

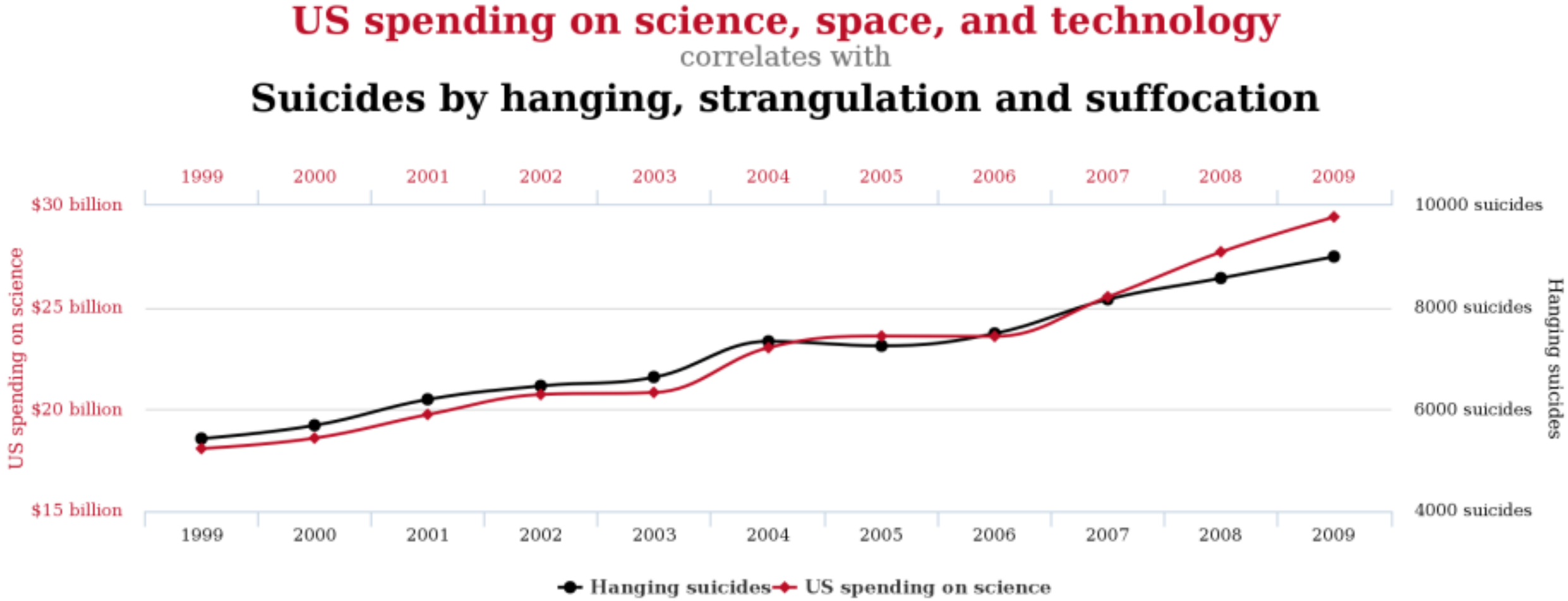
Presentation is key - be a master of Power Point.

All models are false, but some are useful.

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


Rust
Go
Scala
Objective-C
Erlang
Julia
Swift
F#
Clojure
Lua
Groovy
Ruby
Java
I don’t use programming languages.
Elixir
C++
Python
Bash
Haskell
C
JavaScript
SQL
R
C#
ASP/ASP.NET
Octave
MATLAB
Lisp
PHP
HTML
Visual Basic
CSS
D
Perl 6/Raku
Perl 5
0
$50K
$100K
$150K
$200K


H2O
KNIME
Spark NLP
Spark MLlib
Google Prediction
spaCy
Theano
BigML
PyTorch
TensorFlow
scikit-learn
Dato/GraphLab
Mathematica
MATLAB
IBM BigInsights
I don't use any ML/statistics tools.
Excel
IBM SystemML and Nimble
Stata
0
$50K
$100K
$150K
$200K


US spending on science, space, and technology
correlates with
Suicides by hanging, strangulation and suffocation
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
$30 billion
$25 billion
$20 billion
$15 billion
US spending on science
10000 suicides
8000 suicides
6000 suicides
4000 suicides
Hanging suicides
Hanging suicides
US spending on science
tylervigen.com

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

Data Science Verses Programming Jobs

Intuit Job Listing Worldwide

	Aug 22 2016	Jan 20 2021	Jan 3 2022
Data	23	46	64
Software Engineer	168	159	211

Data Science Programming Languages

Python
R
Matlab
Javascript
SAS
Perl
Ruby

Scala
Julia

Java
C++
C
C#

Features of Languages 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

~80% growth per year

Java, Scala, Hadoop, Spark

Hadoop written in Java
Spark written in Scala

JVM languages (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