

So we're all **data scientists** now?

data science is |

data science is **dead**

data science is **the future**

data science is **the sexiest job**

data science is **hard**

data science is **a branch of**

data science is **software**

data science is **statistics on a mac**

data science is **not taught at universities**

data science is **overrated**

data science is **a fad**

The growth of data science



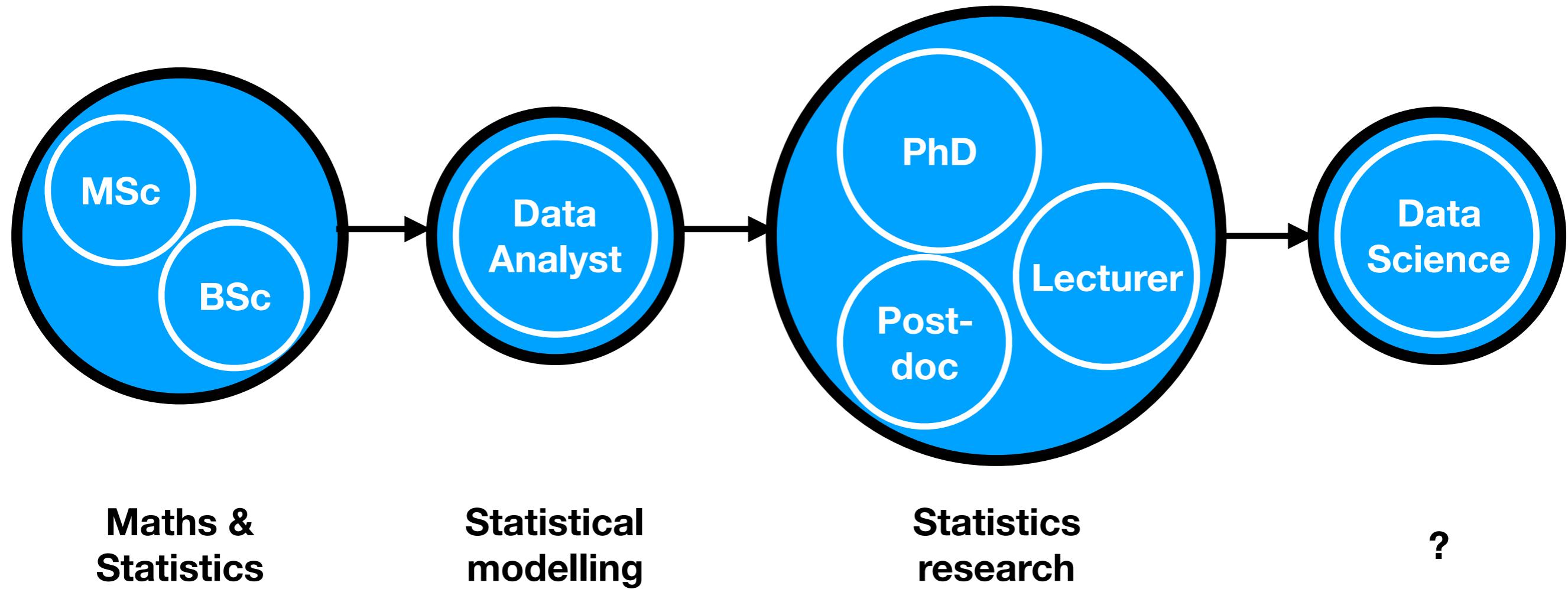
Overview

- My background & what I do at Tesco Bank
- What is **data science**?
- How is **data science** different from statistics?
- Data science tools & skills

Caveats:

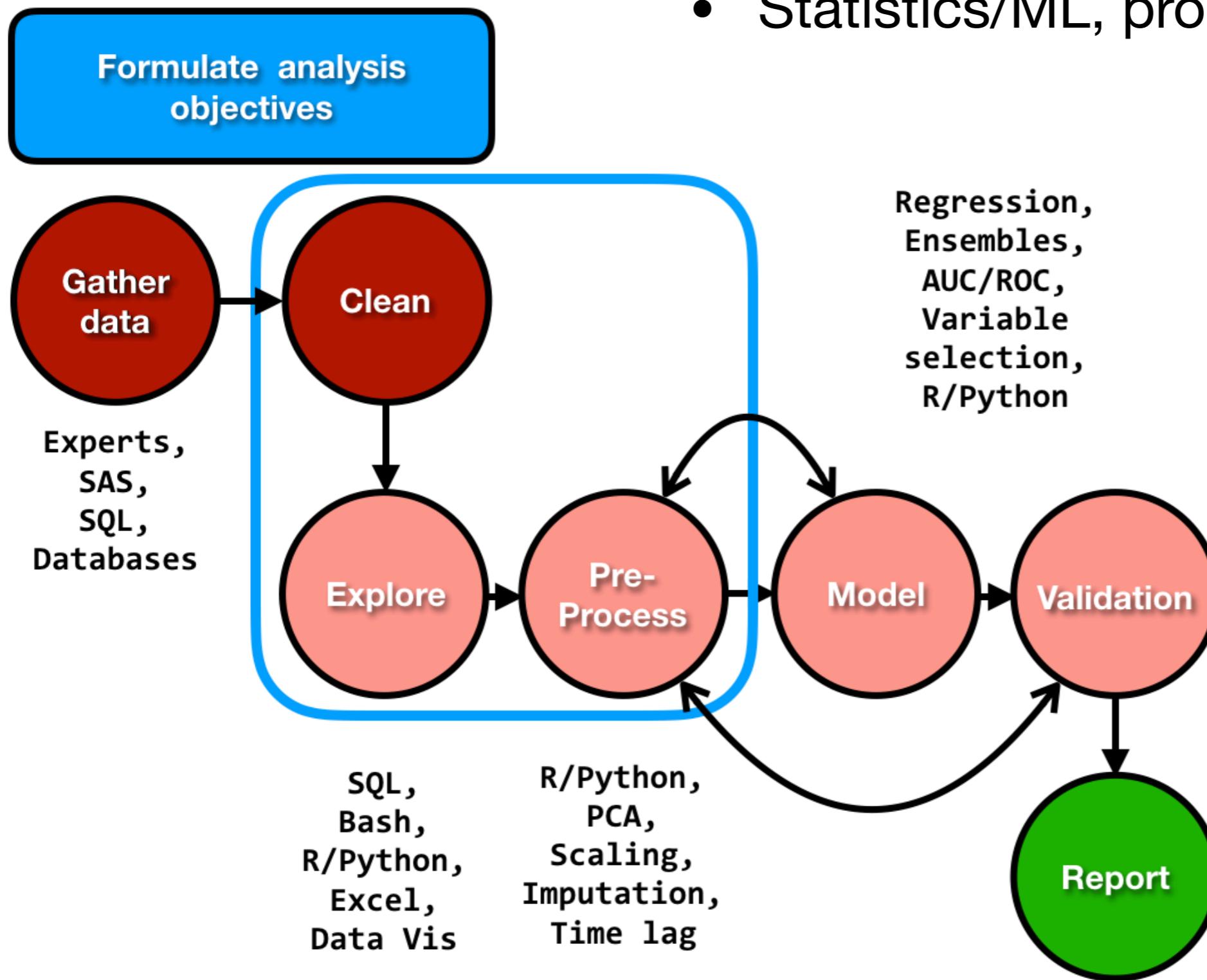
- I will try to be factual, as I'm not an expert
- My estimations probably have high variance and may not generalise

How I ended up doing **data science**



Data scientist role:

- Use data to solve business problems
- Recast objectives for analysis
- Collaborative
- Statistics/ML, programming, software



What about you?



My job title is **data scientist**



I identify as a **data scientist**



My work could be described as **data science**

What is data science?

Data science is what the data scientists say it is

The screenshot shows a blurred background of code snippets, likely from various datasets or kernels. In the foreground, the following text is displayed:

🏆 Featured Dataset

Kaggle ML and Data Science Survey, 2017

A big picture view of the state of data science and machine learning.

 Kaggle • last updated 15 days ago

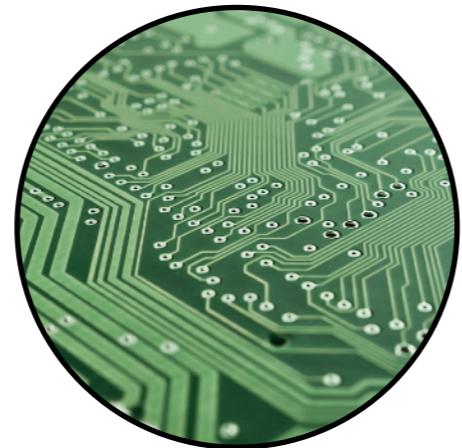
Overview Data Kernels Discussion Activity

Backgrounds that **data scientists** come from



**Computer
Science**

33%



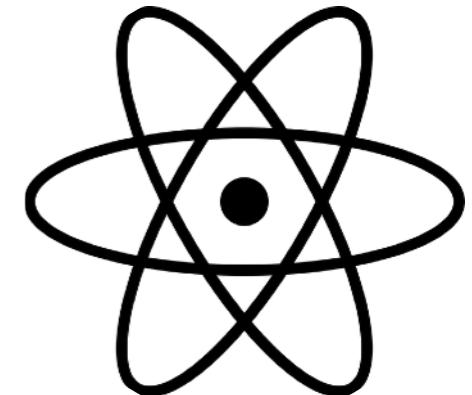
**Engineering
(All types)**

19%



**Mathematics
& Statistics**

19%

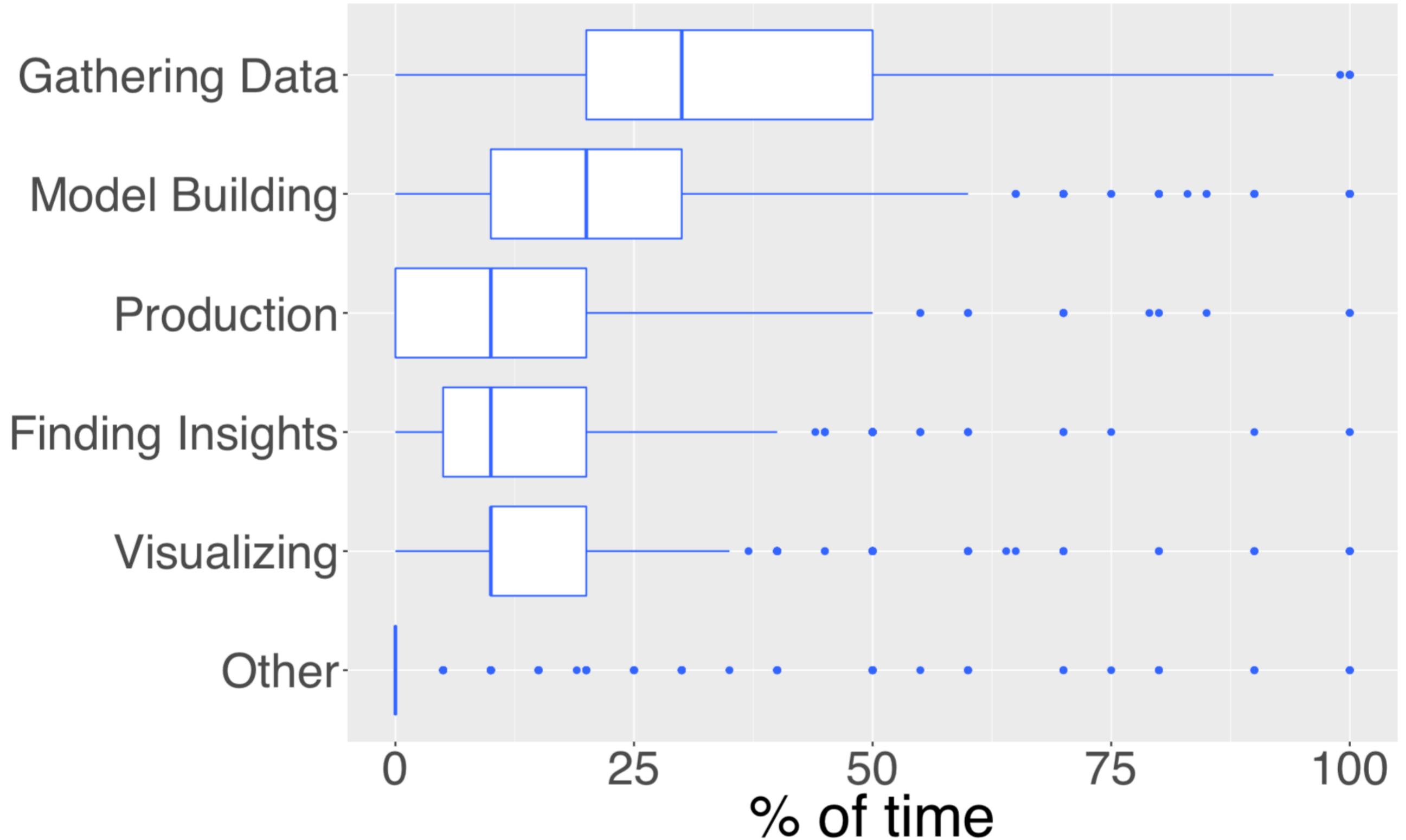


Physics

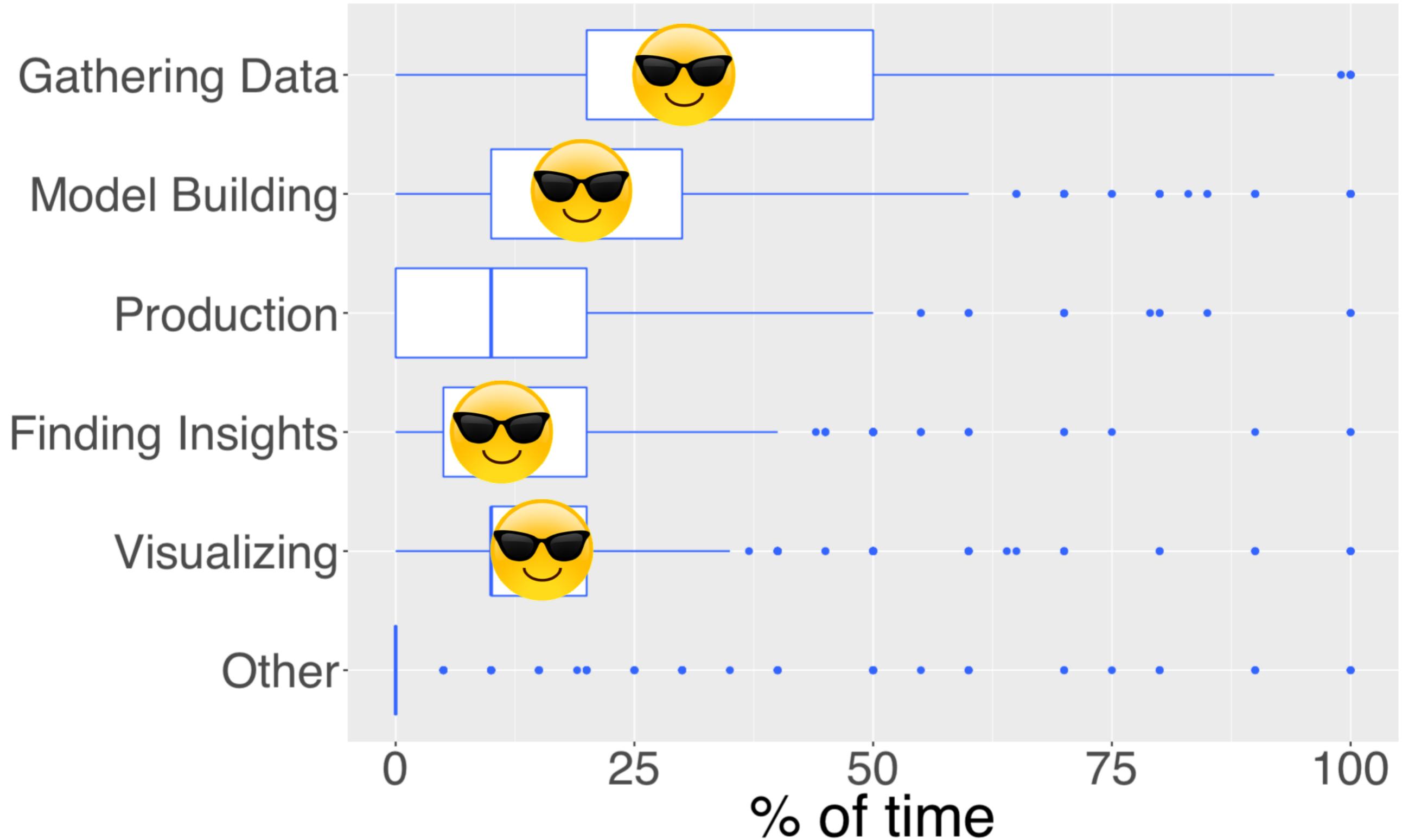
6%

Remaining 23%: IT/network/sysadmin, social science, biology, psychology, health

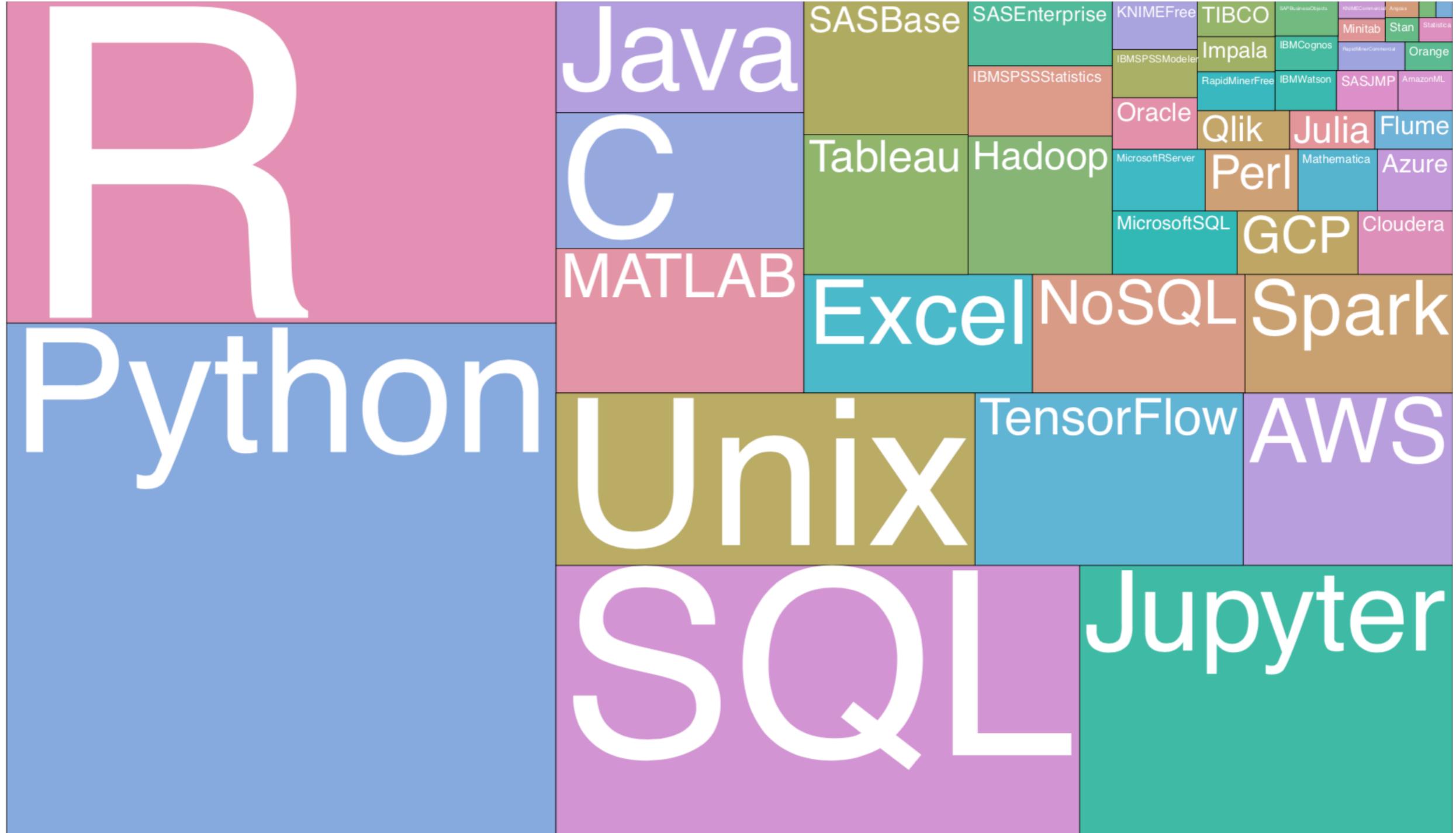
What data scientists do



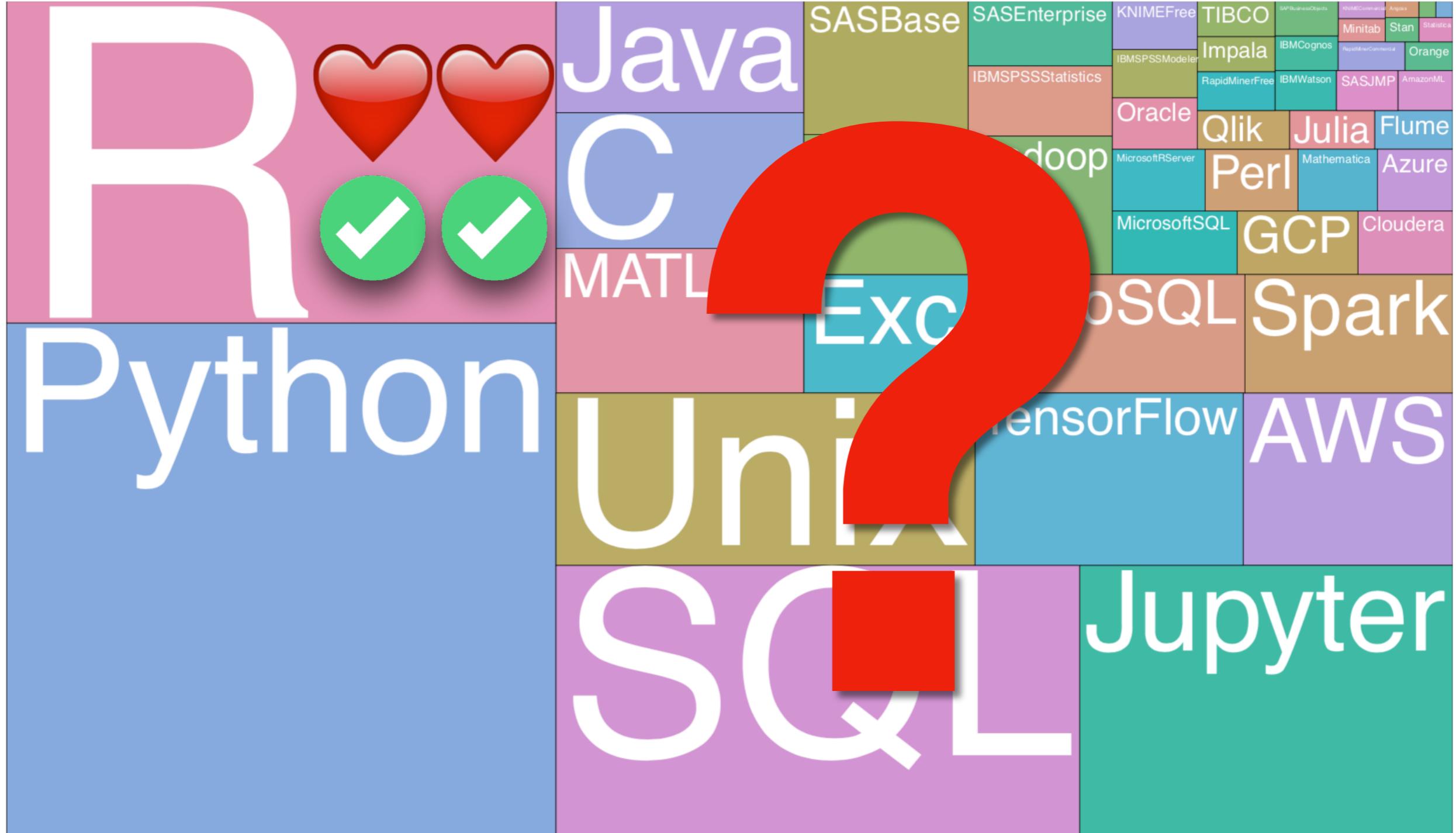
What data scientists do



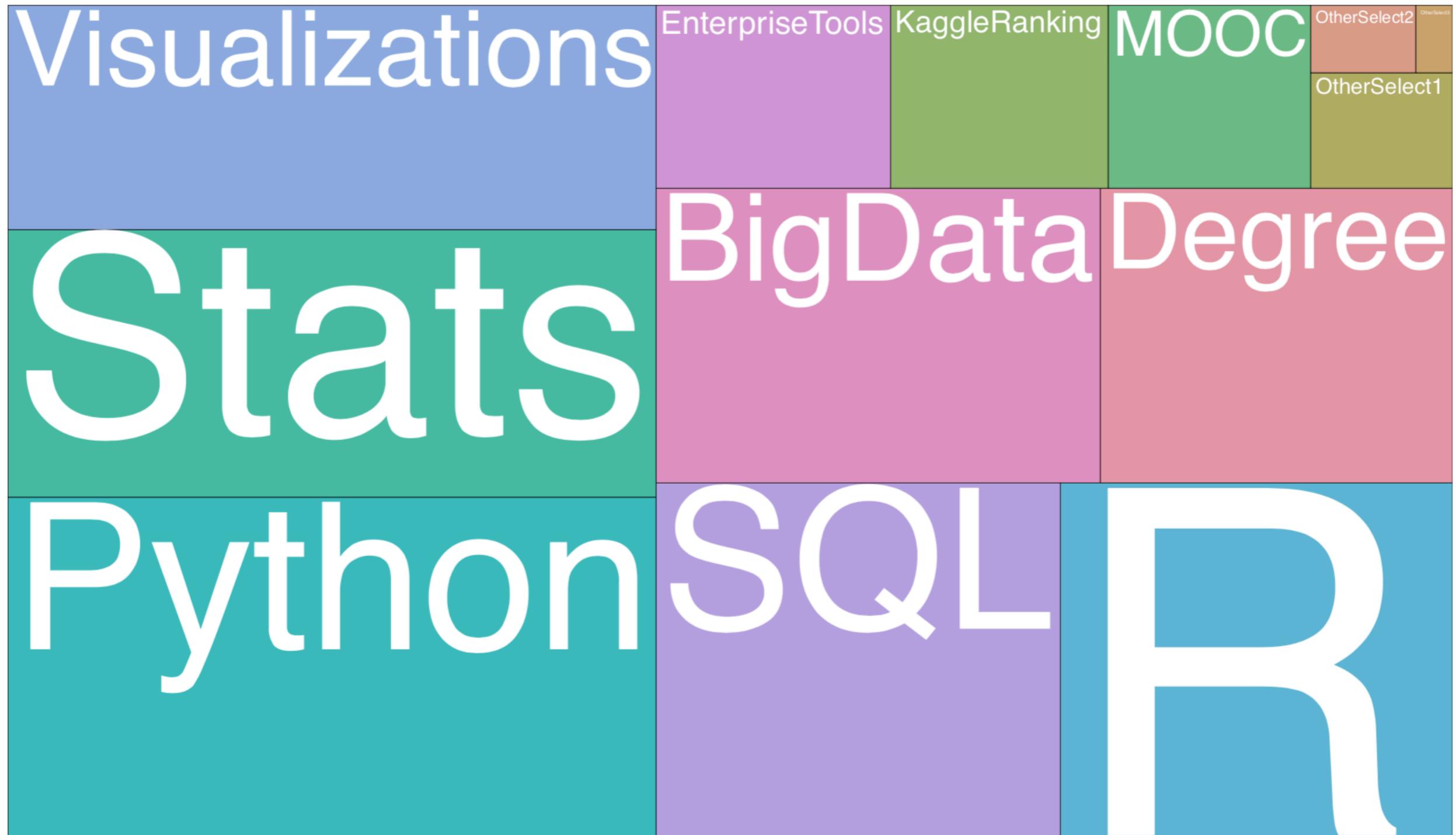
Tools that data scientists use



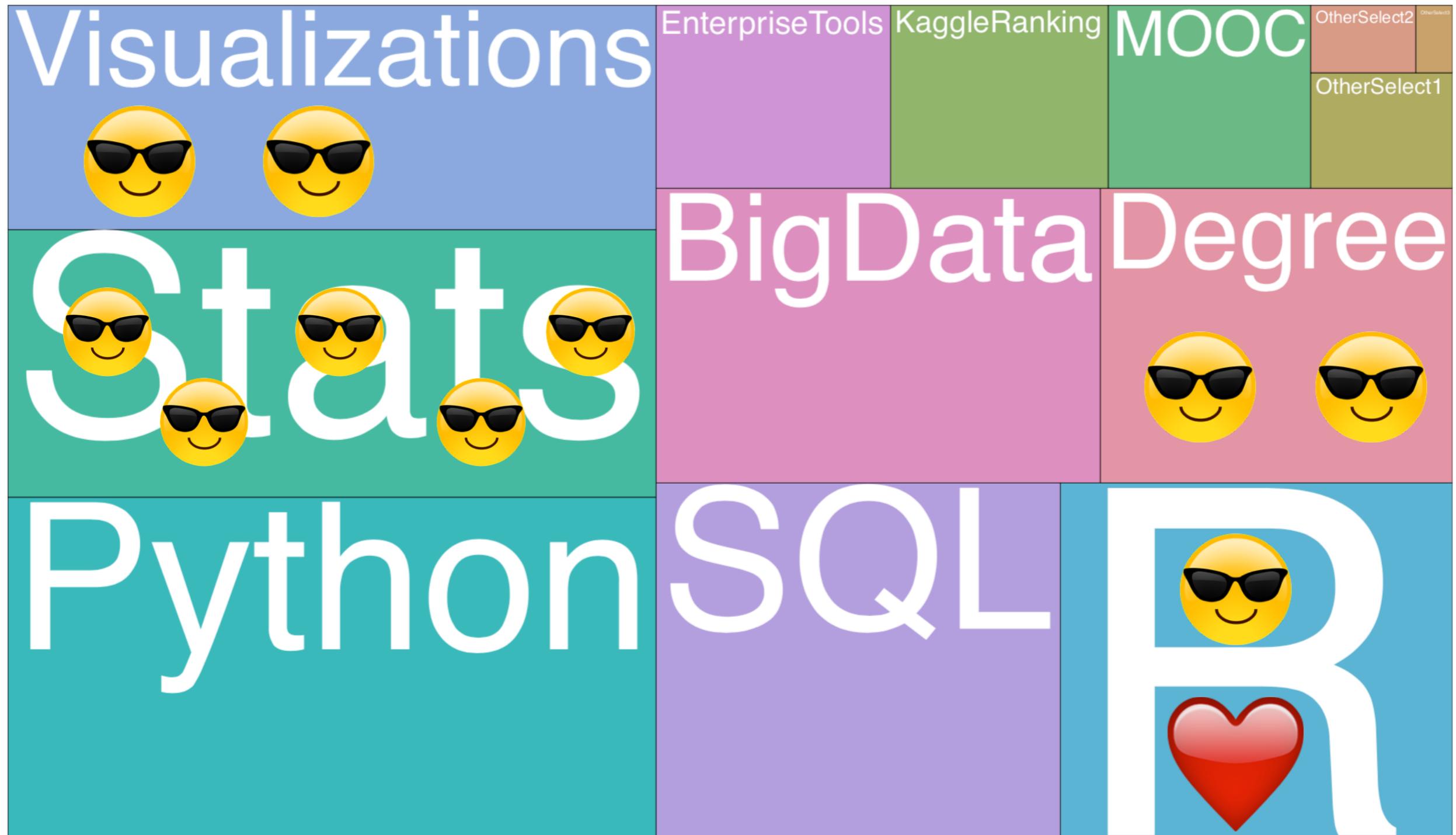
Tools that data scientists use



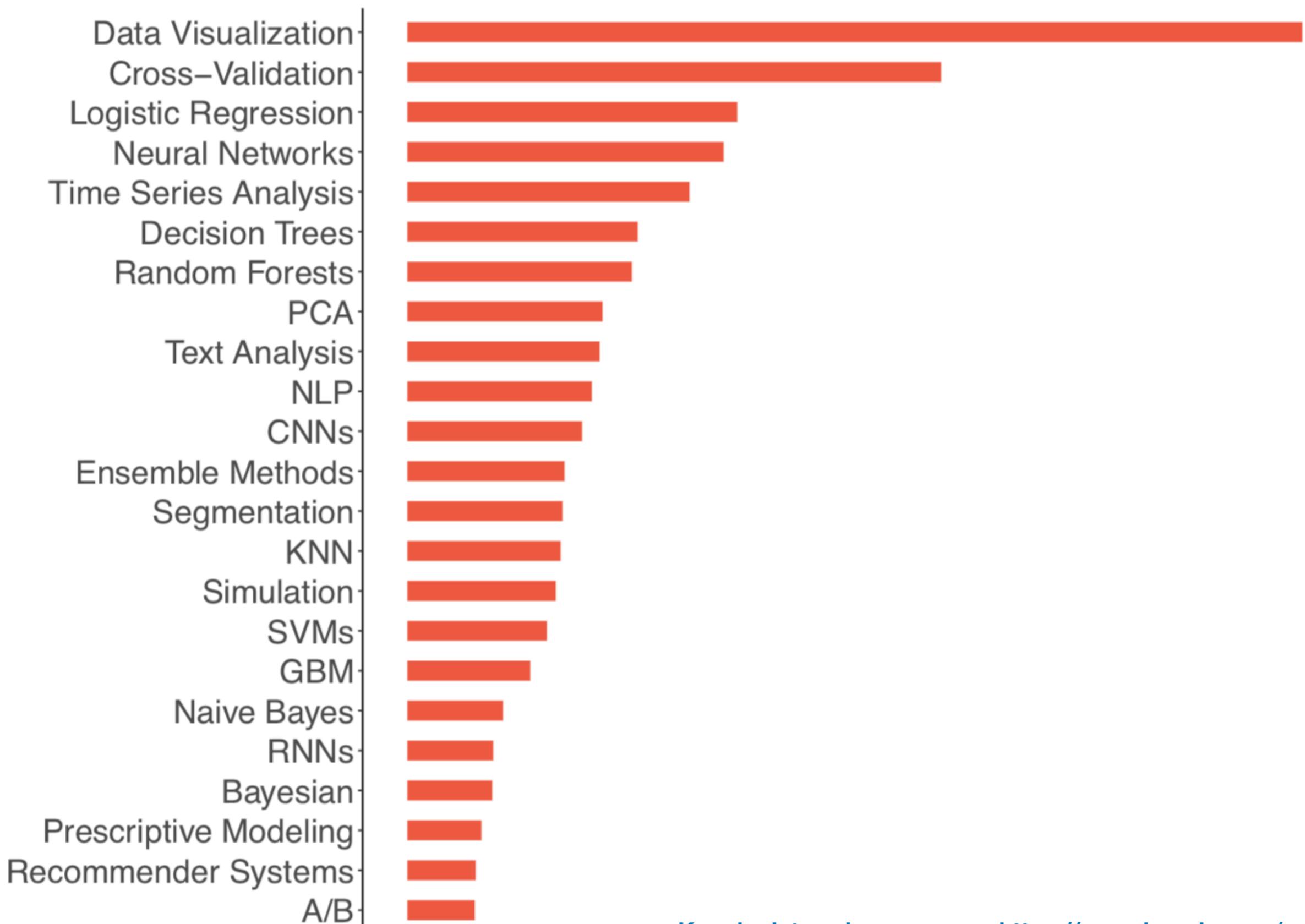
Which skills are necessary for **data science**?



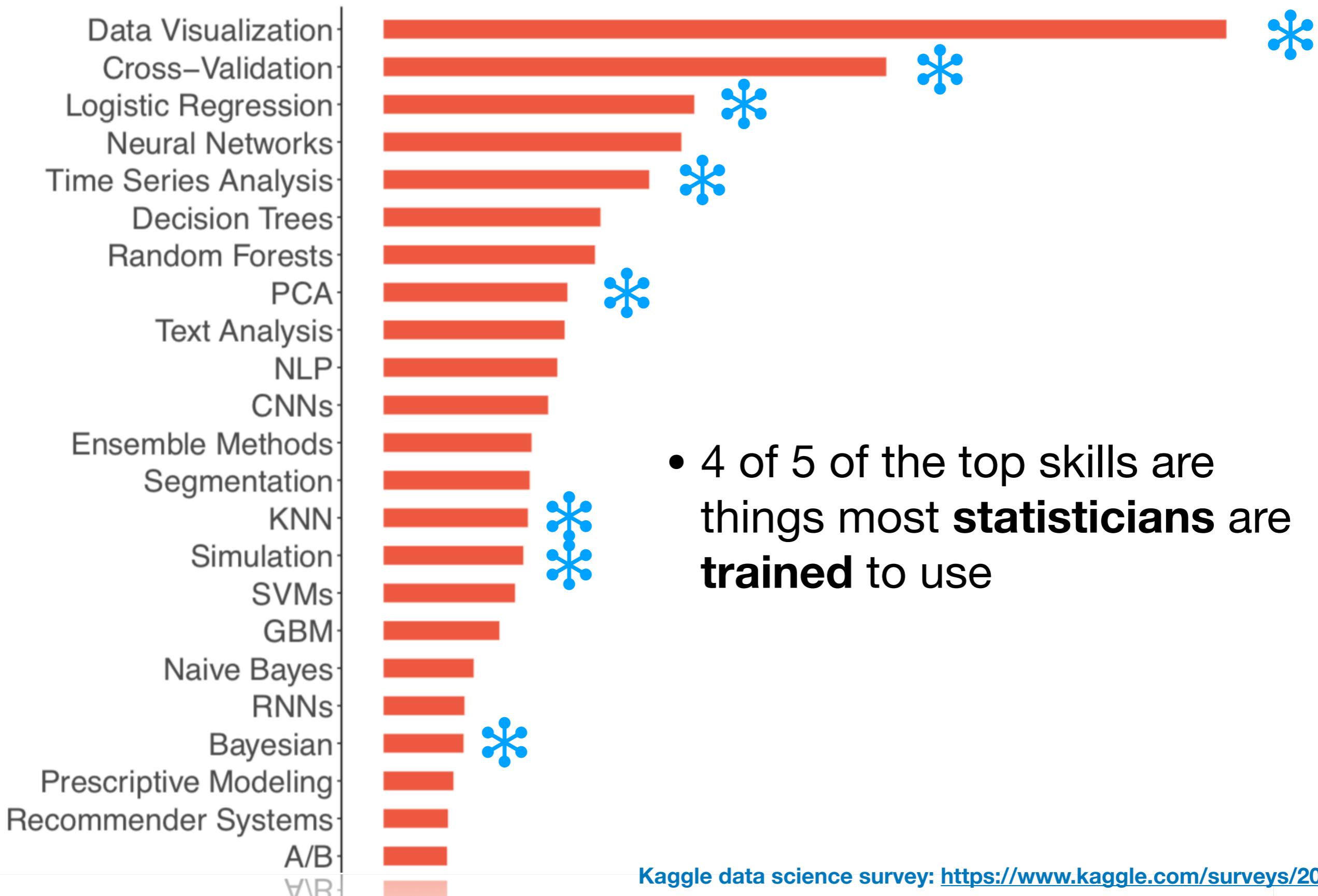
Which skills are necessary for data science?



Most common data science techniques



Most common data science techniques



- 4 of 5 of the top skills are things most **statisticians** are trained to use

Data science tools & skills

Buzzy data science thing #1: Python



**Survey: 63% recommend
newcomers learn python first**



Ok but why

- ML: scikit-learn
- Ubiquity
- Speed: Python versus R

Buzzy data science thing #2: Machine learning

- Complex, non-linear models (mostly)
- (Un)Supervised ML: both R, Python
 - Regression, clustering, dimension reduction
 - Tree ensembles: random forest, XgBoost
 - Neural networks (deep learning etc.)



Buzzy data science item #3: Cluster computing



- In memory storage and computation
- MLlib: cluster machine learning
 - distributed, parallelised, iterative
- Written in Scala

(do we actually need to use all of that data?)

Communication: #1 data science skill



- Defining problems
- Reporting: verbal & written
- 50% of the interview process

Related:

- Software collaboration - git
- Agile (project management)

Statistics and data science

Not all that is **data** is **statistics**...

“A **data scientist** is a **statistician** who lives in San Francisco”

“If you’re analysing data, you’re doing **statistics**. You call it **data science** or informatics or analytics or whatever, but it’s still **statistics**.”

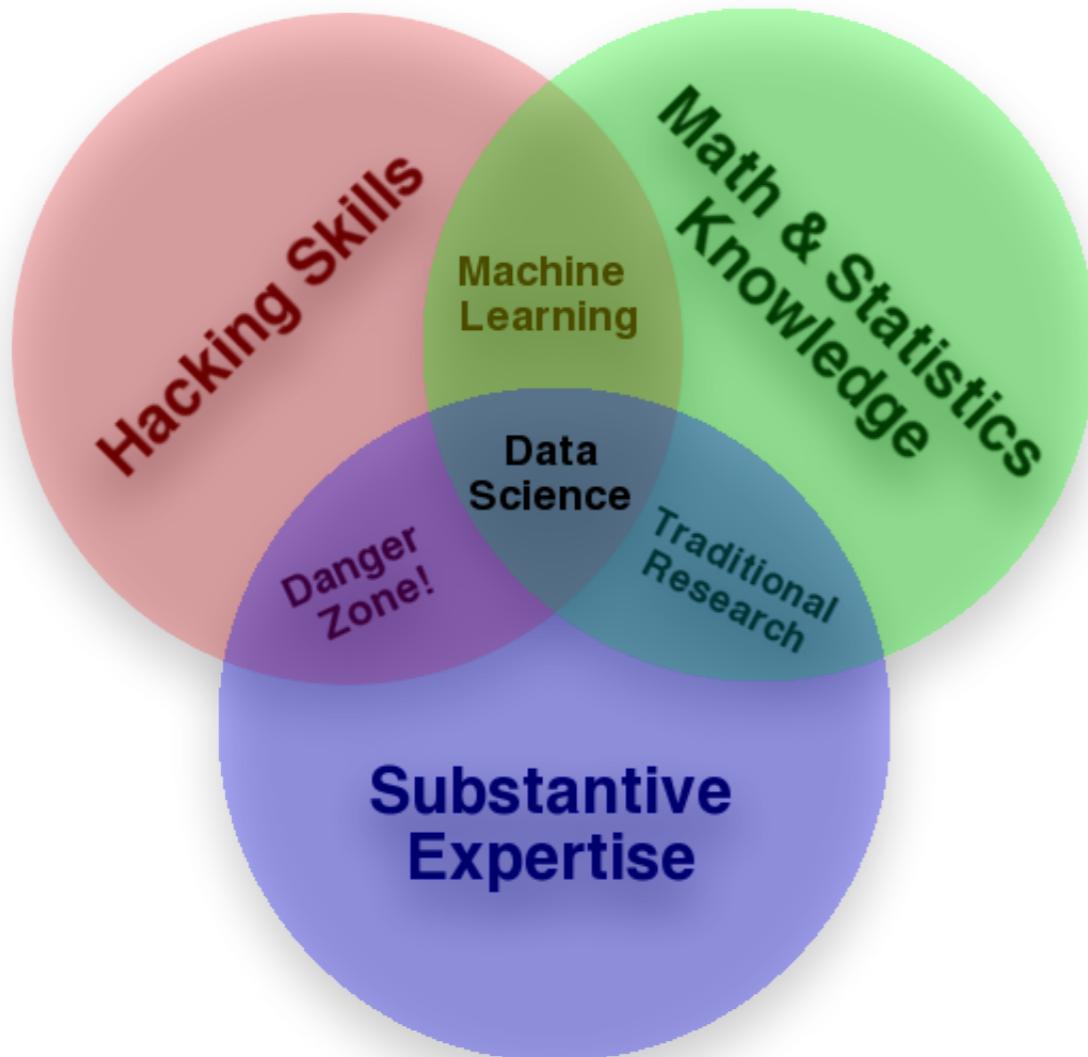
“I think **data scientist** is a sexed up term for a **statistician**”

“**Machine learning** is clearly a sub-discipline of **statistics**”



“**Data science** is **statistics** on a Mac”
- *@BigDataBorat*

Should statisticians be doing data science?



YES

- Kaggle survey: statistics & data science are clearly complementary

Drew Conway, 25/3/2013

Source: <http://drewconway.com/>

Data science and statistics

The screenshot shows a YouTube video player. At the top, there are logos for the Royal Statistical Society, Google, Qriously, Mendeley, and the UK Statistics Authority. Below the logos is a large, abstract image of many 3D, translucent blue and white letters floating in space. The video title 'Data Science and Statistics: different worlds?' is displayed in a dark box below the image. The video progress bar shows it's at 0:06 / 1:35:37. Below the progress bar are standard YouTube controls: play, pause, volume, and a settings gear icon. To the right of the video player are sharing icons for HD, square, and other formats. The video description below the player reads 'Data Science and Statistics: different worlds?'. It has 37,264 views, 317 likes, 5 dislikes, and options to share or more. At the bottom left is the Royal Statistical Society logo, and at the bottom right is a red 'SUBSCRIBE 2.6K' button. The video was published on 19 May 2015.

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS

Google Qriously MENDELEY UK Statistics Authority

Data Science and Statistics: different worlds?

37,264 views 0:06 / 1:35:37

SHARE ...

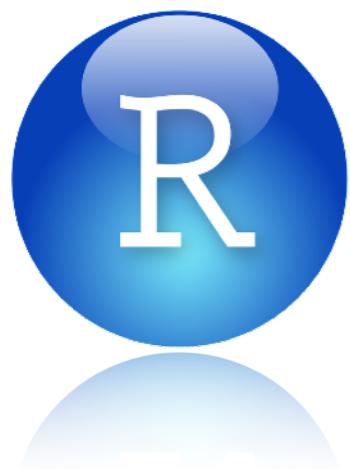
ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS

RoyalStatSoc Published on 19 May 2015

SUBSCRIBE 2.6K

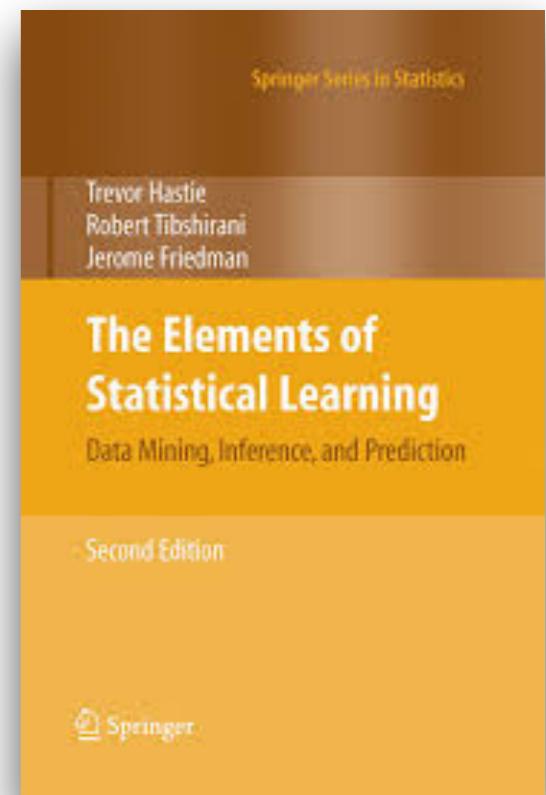
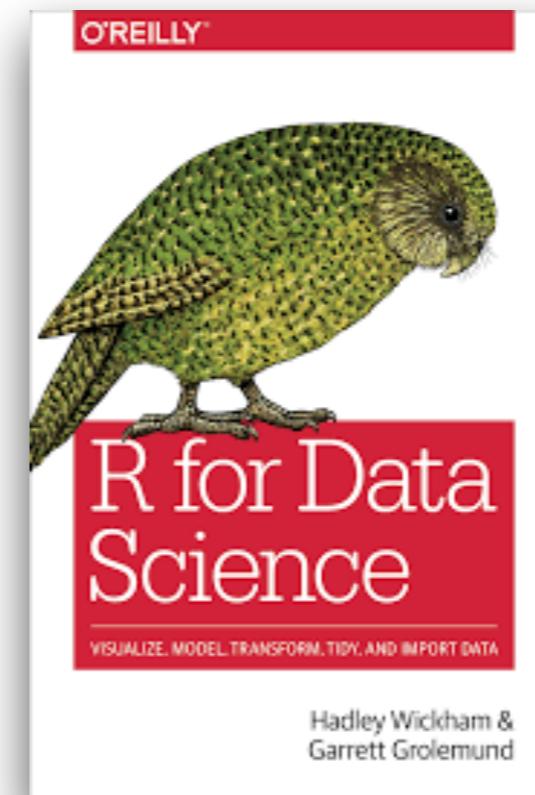
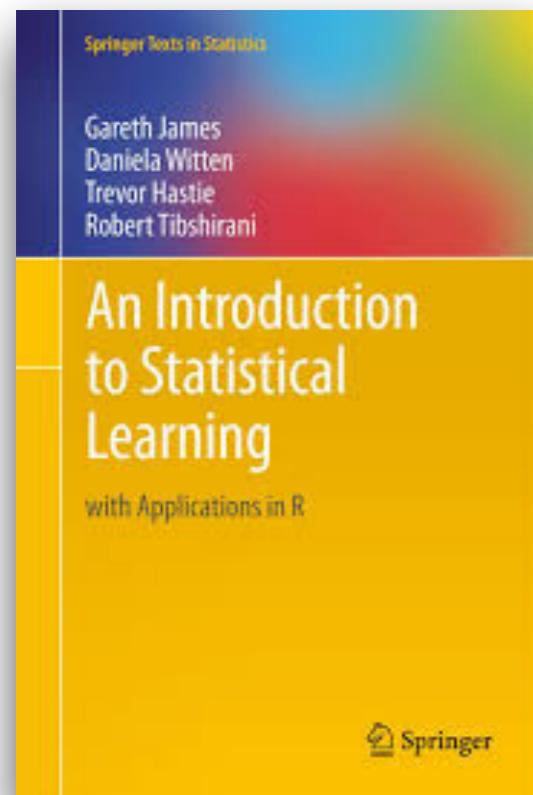
Stuff from **statistics** that I'm glad of

- Recognising point estimates
- Inference versus prediction
- Exposure to wide variety of domains
- Really knowing linear models
- R: Rcpp, dplyr, shiny, flexdashboard



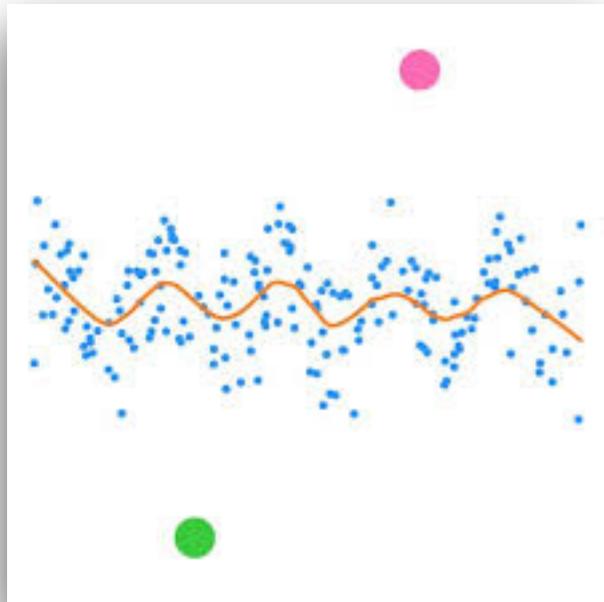
There is always a lot to learn...

Survey: Average of 60% of learning time spent on online courses and self teaching



coursera Catalog Search catalog 🔍
Home > Data Science > Machine Learning
Machine Learning
Overview Syllabus

Absorb the culture with a podcast...



Not So Standard deviations



Talking Machines



Wrap up...

- **Data science** is more than just **statistics**
- But, **statistical** training is ideal for **data science**
- Forensic mindset & mathematical skill
- Scope for statistics to play a bigger role
 - *inference, uncertainty & decisions*
- Communication is key
- Don't be put off by new software

Thanks

any questions?



Formulate analysis objectives

