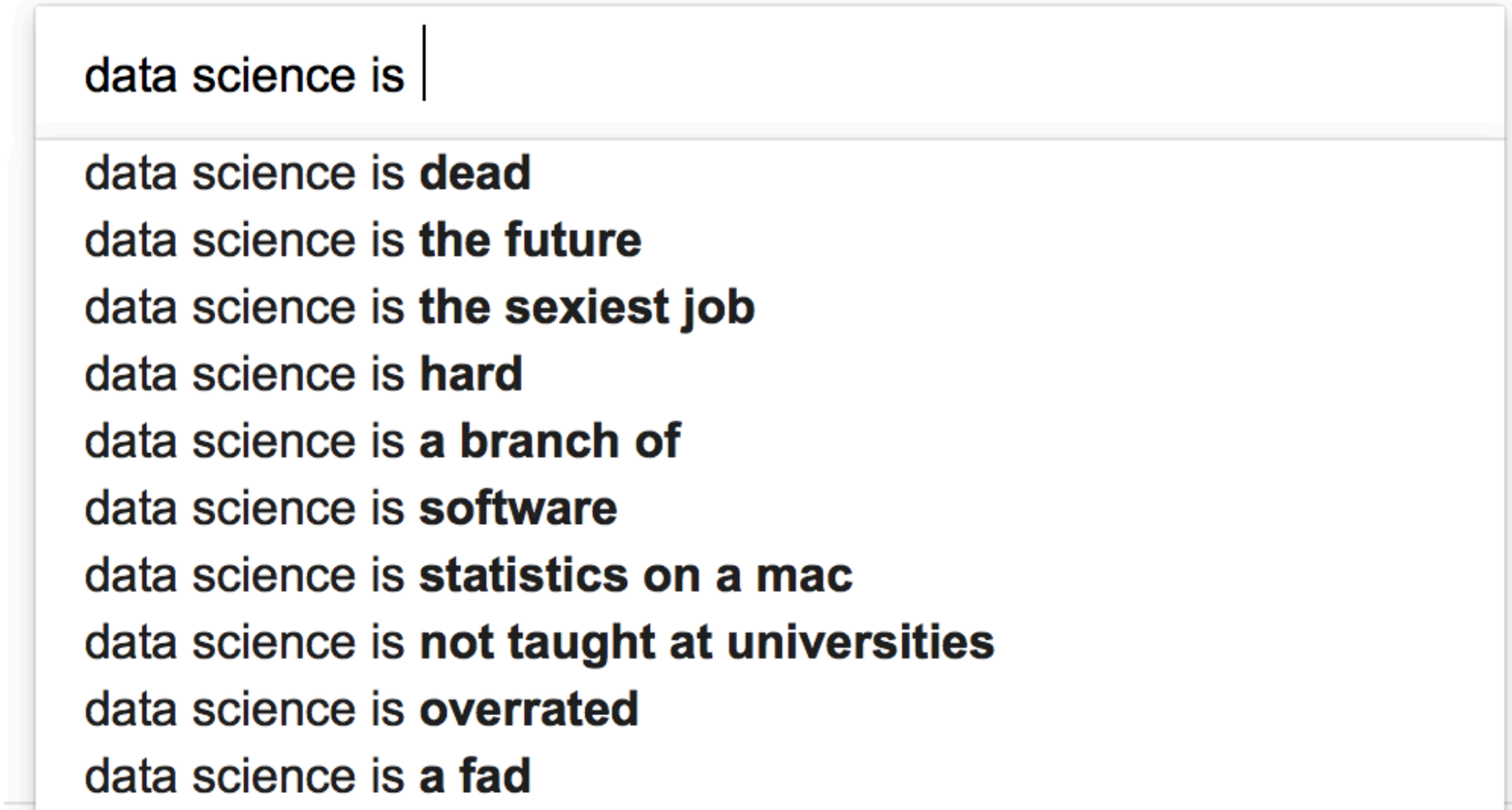
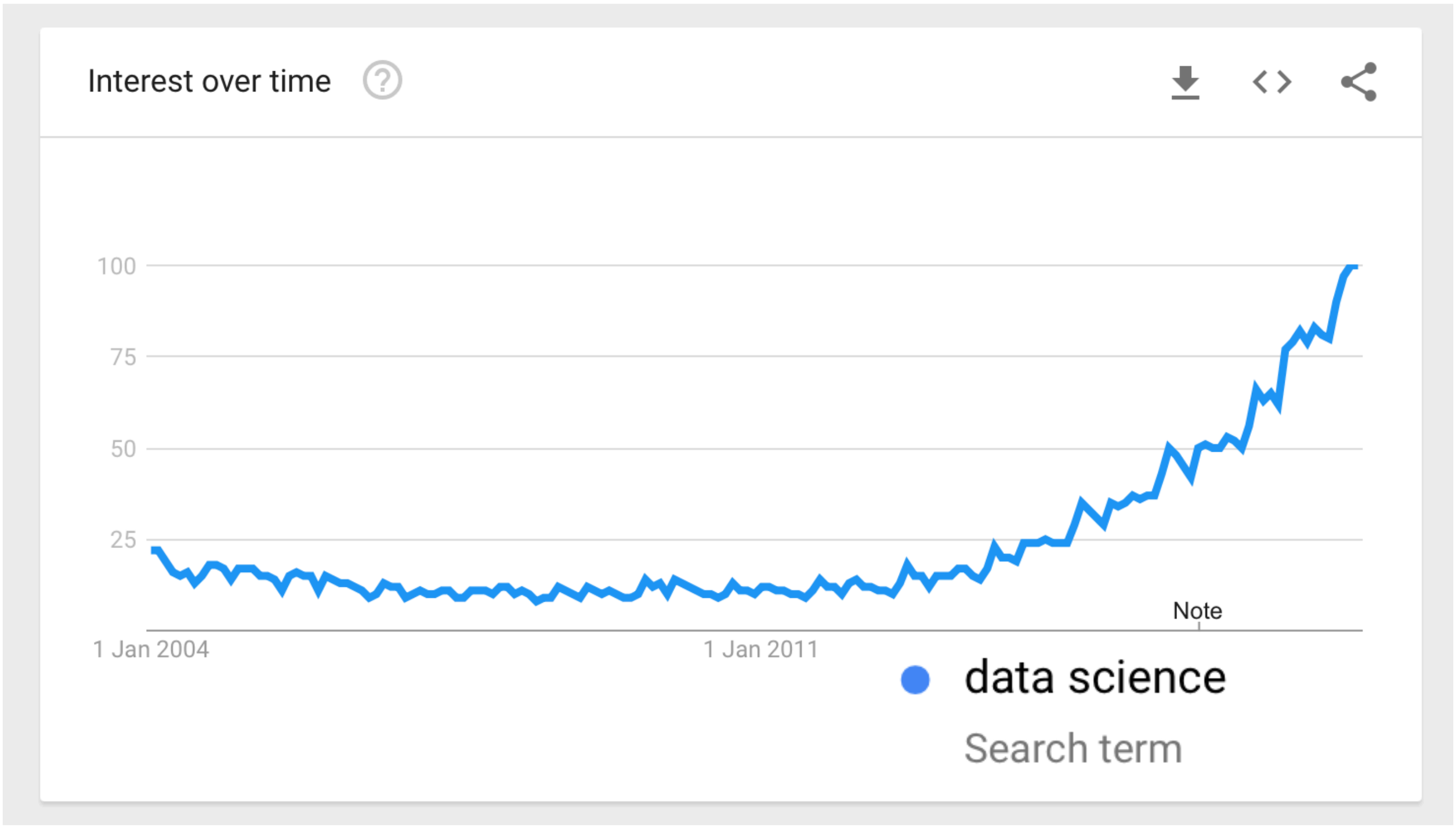
So we're all **data scientists** now?



The growth of **data science**



Google trends: https://trends.google.co.uk/trends/explore?date=all&q=data%20science

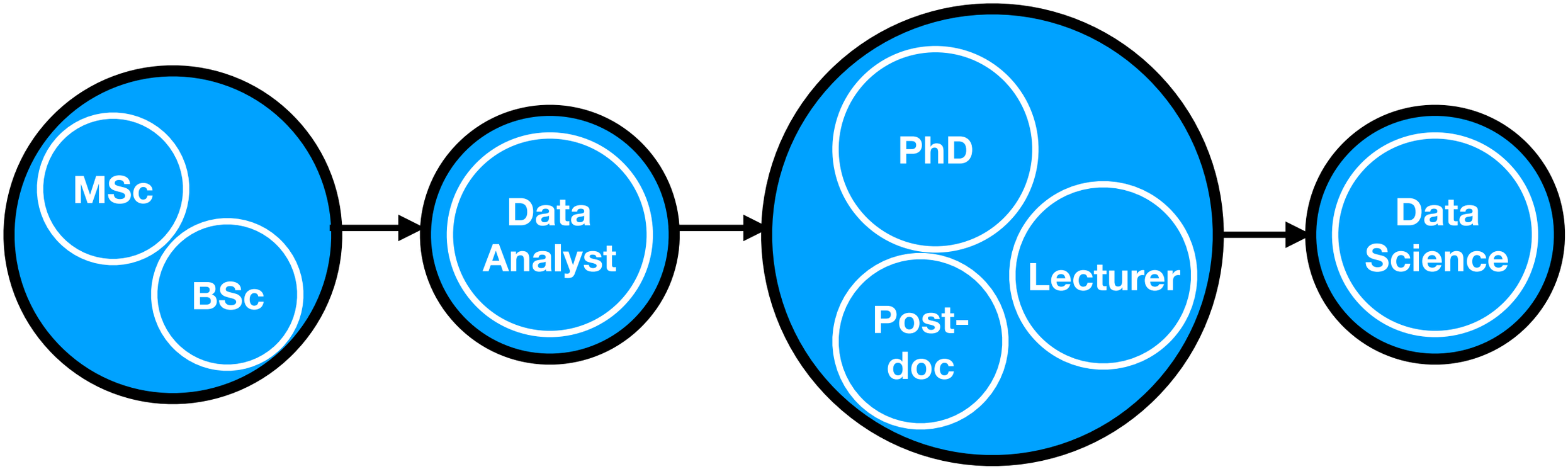
Overview

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

Caveats:

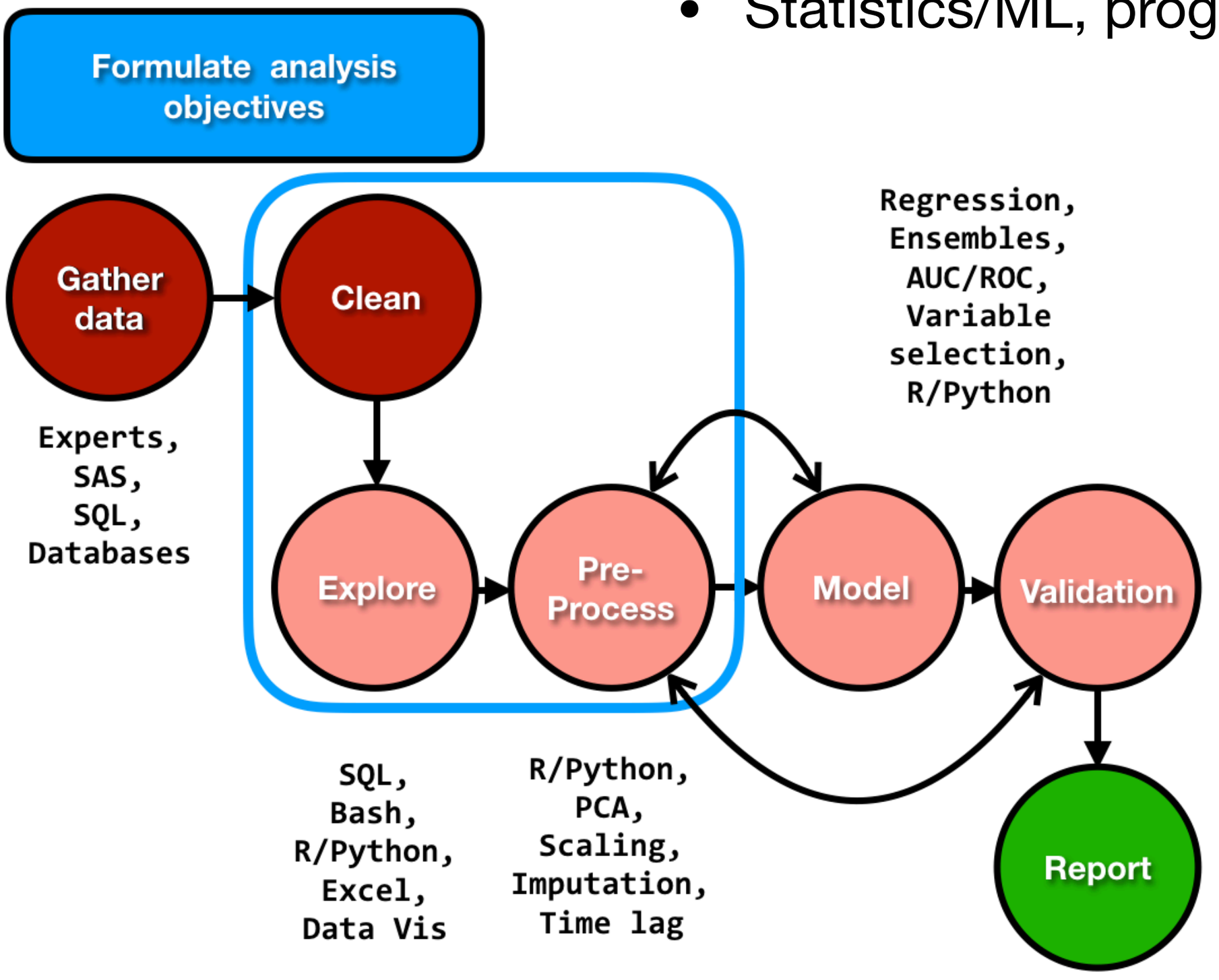
- I will try to be factual, as I’m not an expert
- My estimations probably have high variance and may not generalise

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**?

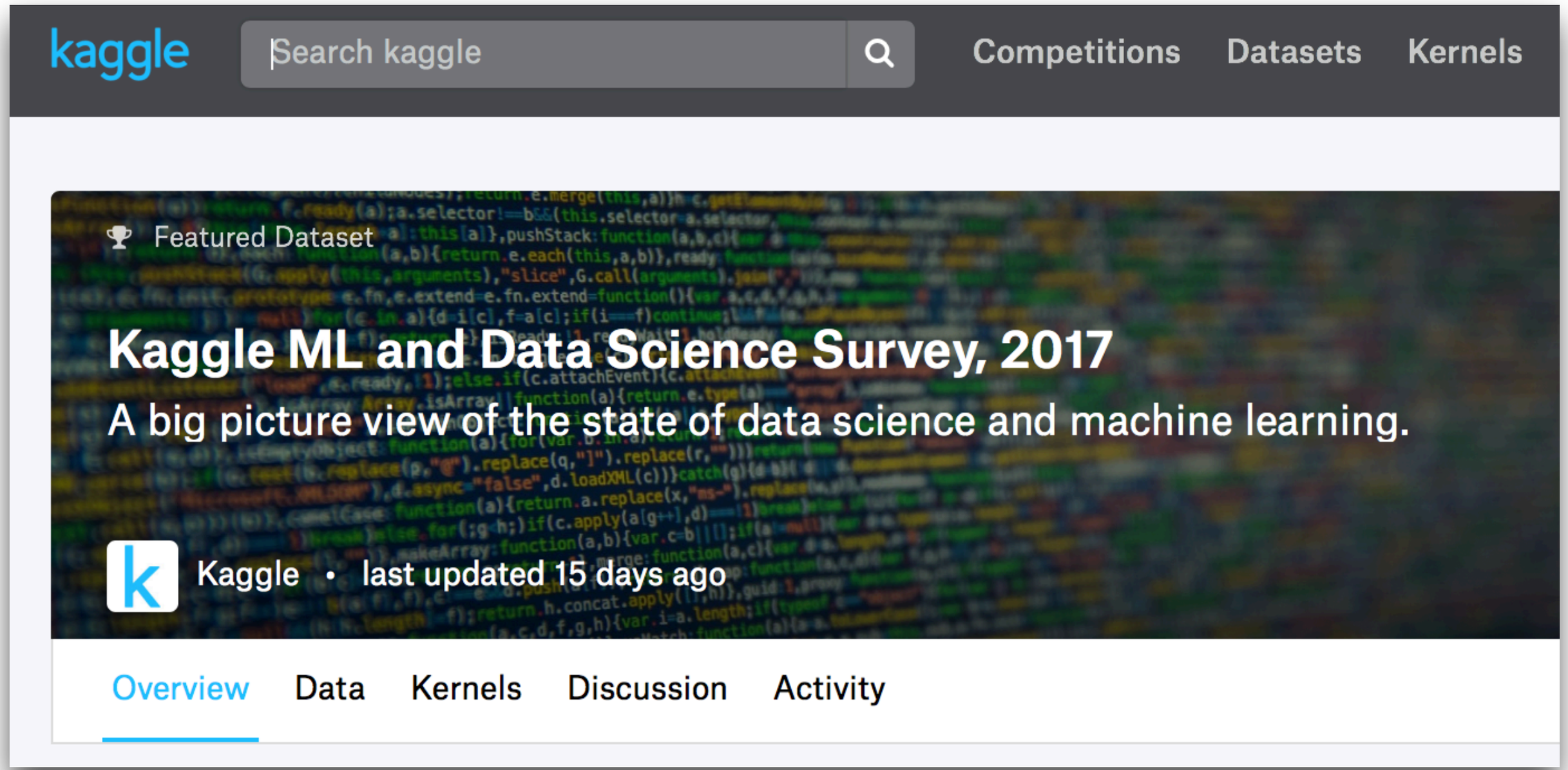
My job title is **data scientist**

I identify as a **data scientist**

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

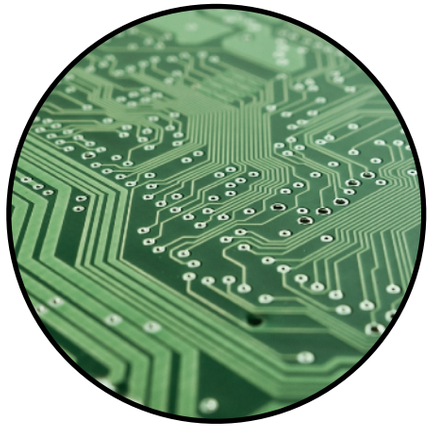
What is **data science**?

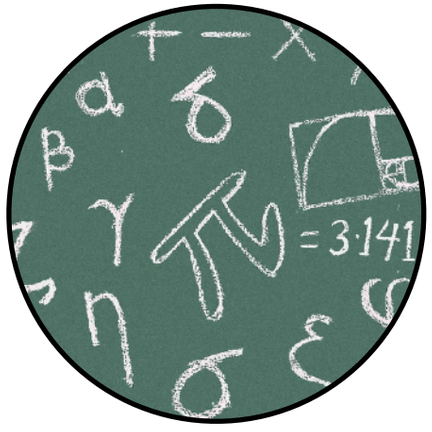
Data science is what the **data scientists** say it is

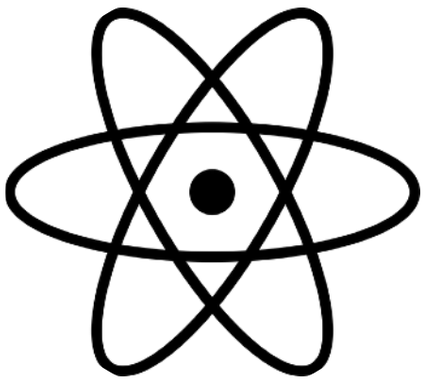


Kaggle data science survey: https://www.kaggle.com/surveys/2017

Backgrounds that **data scientists** come from





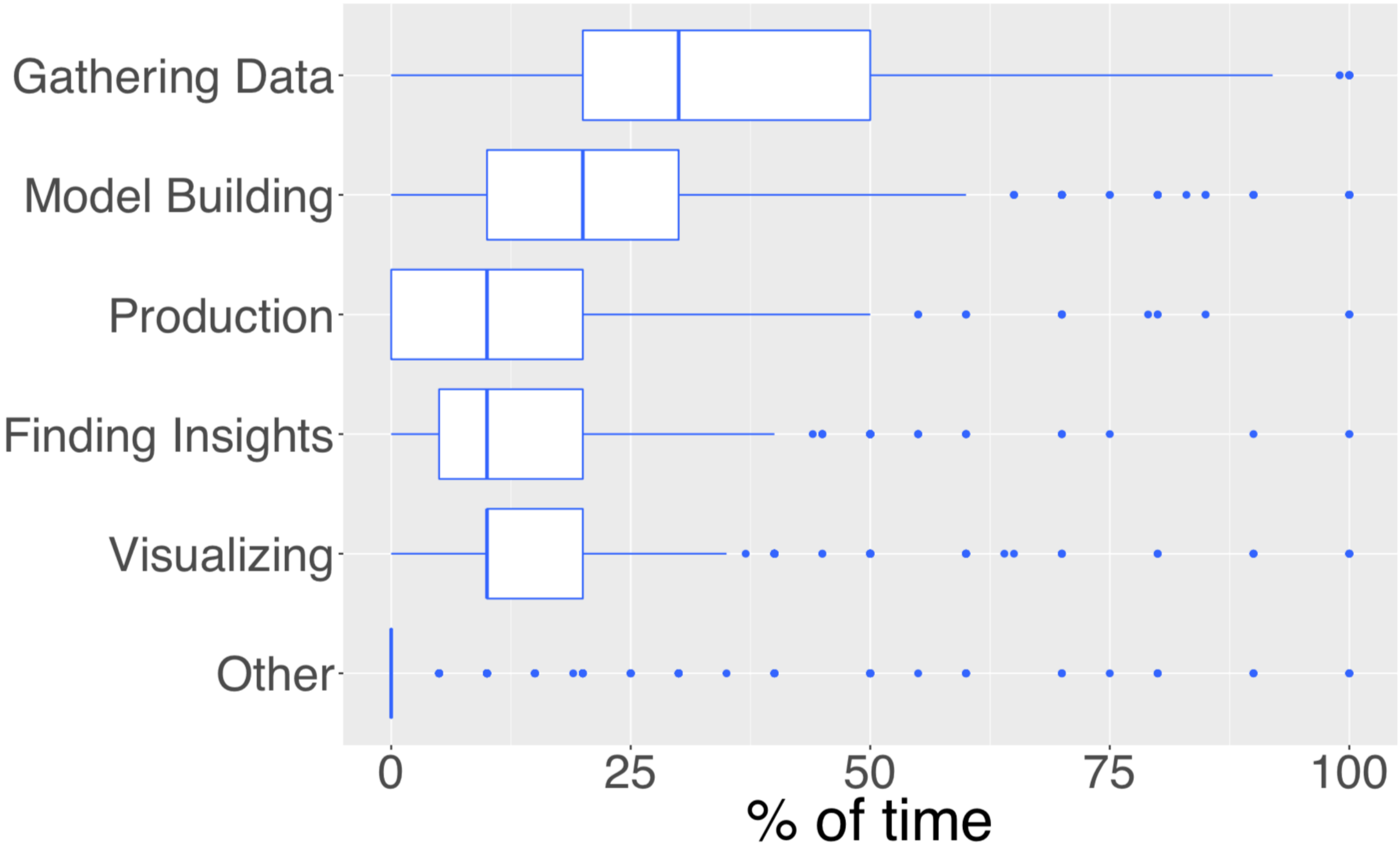


Computer Science	Engineering (All types)	Mathematics & Statistics	Physics
33%	**19%**	**19%**	**6%**

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

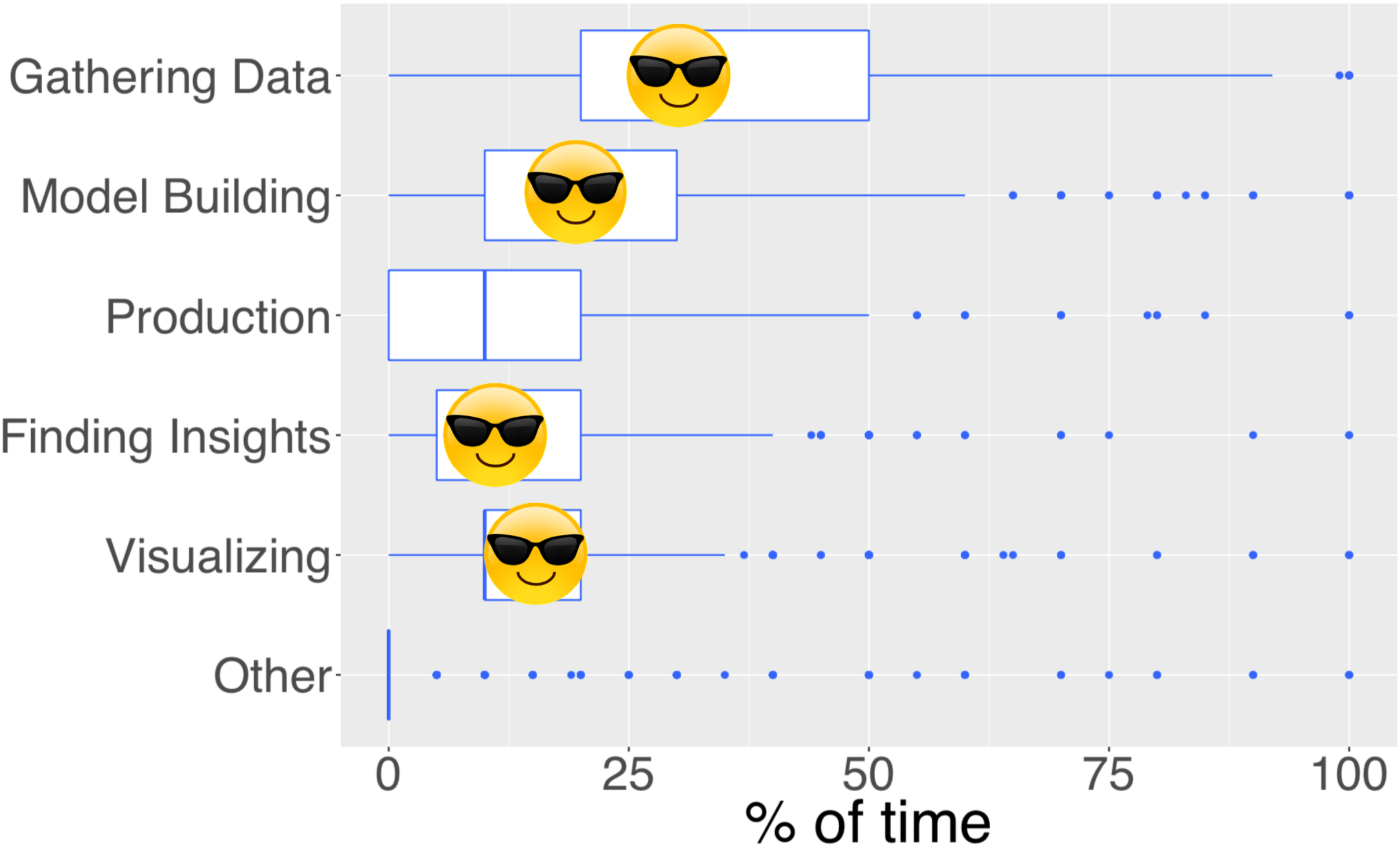
Kaggle data science survey: https://www.kaggle.com/surveys/2017

What **data scientists** do



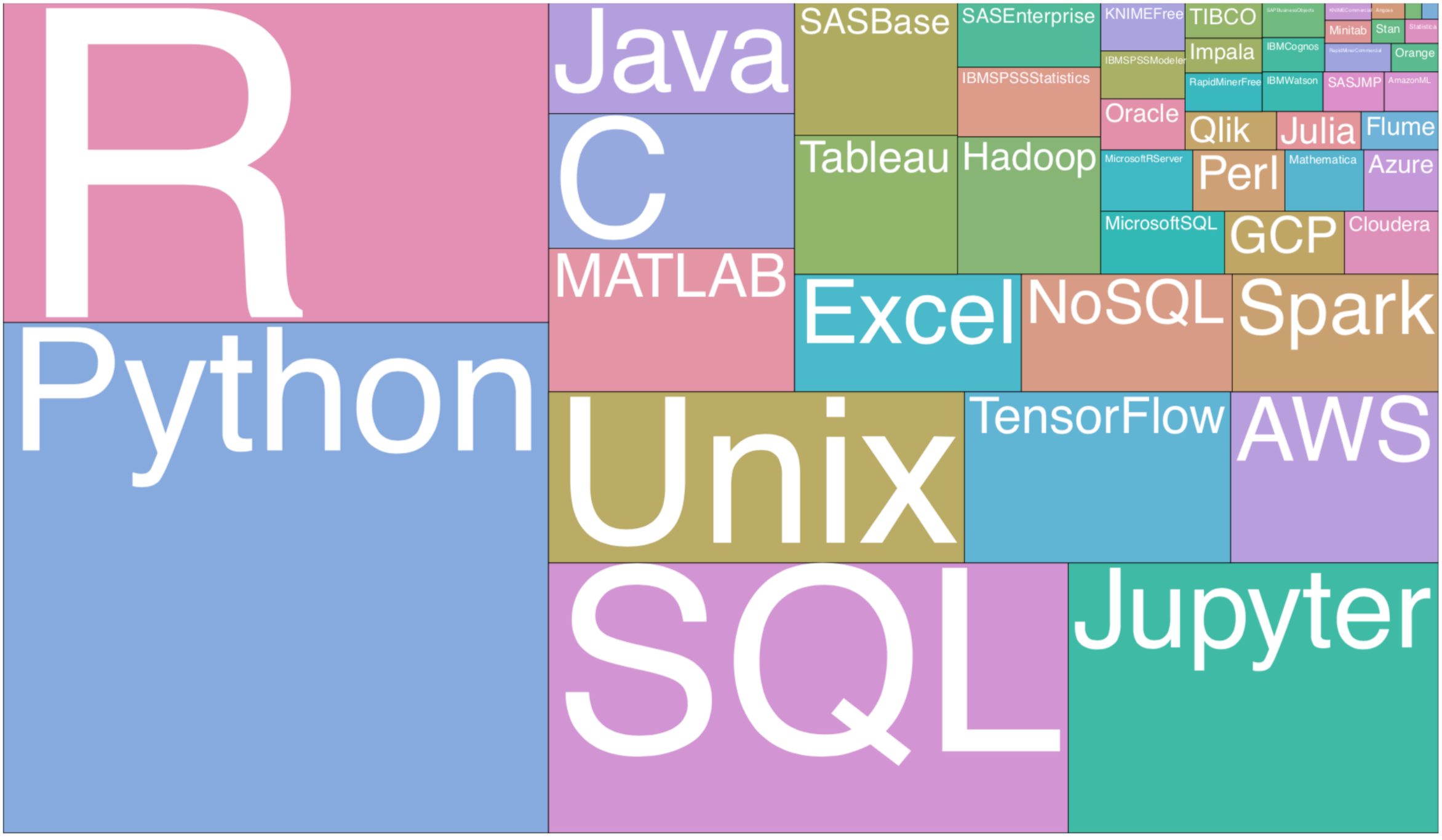
Kaggle data science survey: https://www.kaggle.com/surveys/2017

What **data scientists** do



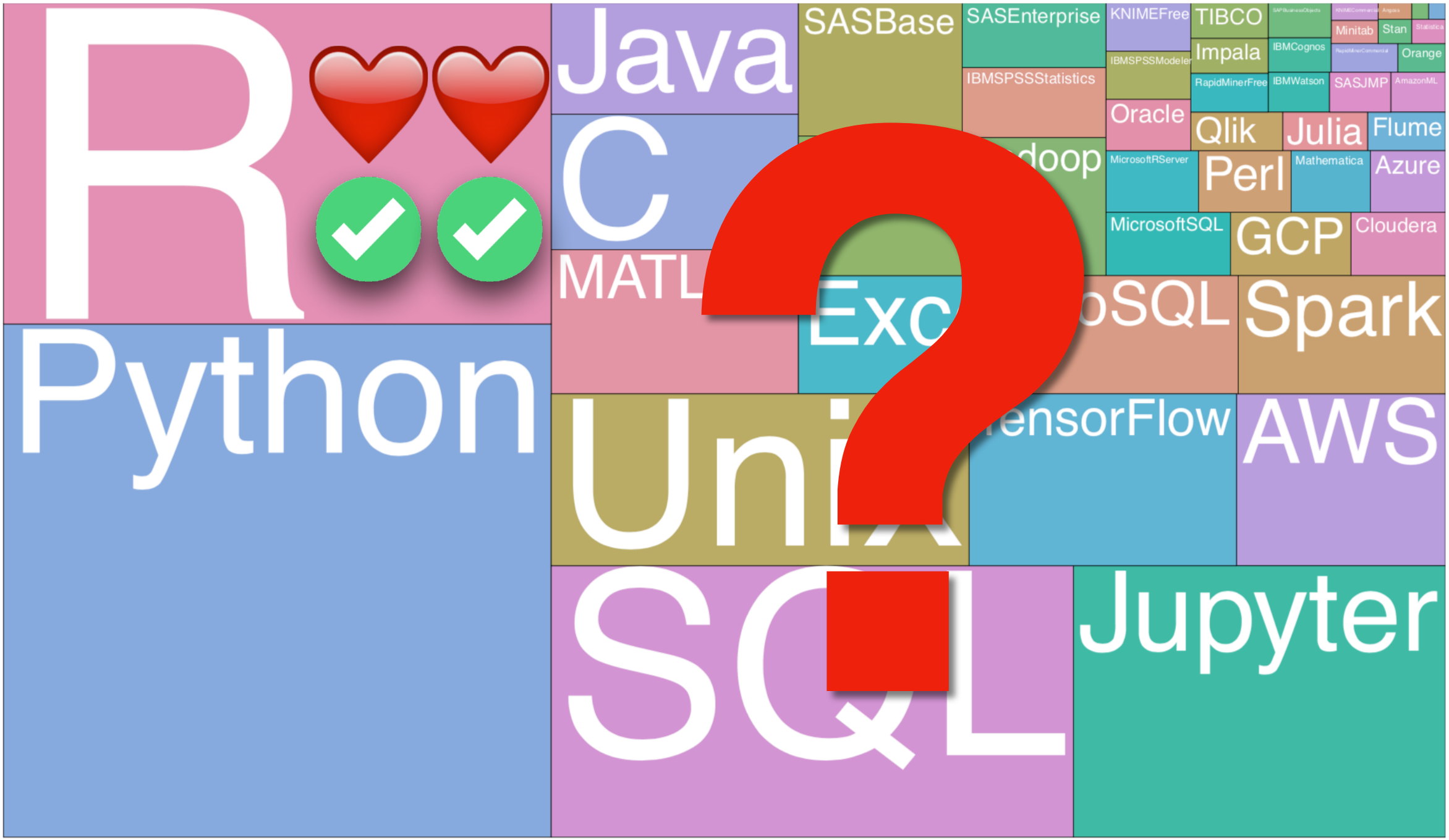
Kaggle data science survey: https://www.kaggle.com/surveys/2017

Tools that **data scientists** use



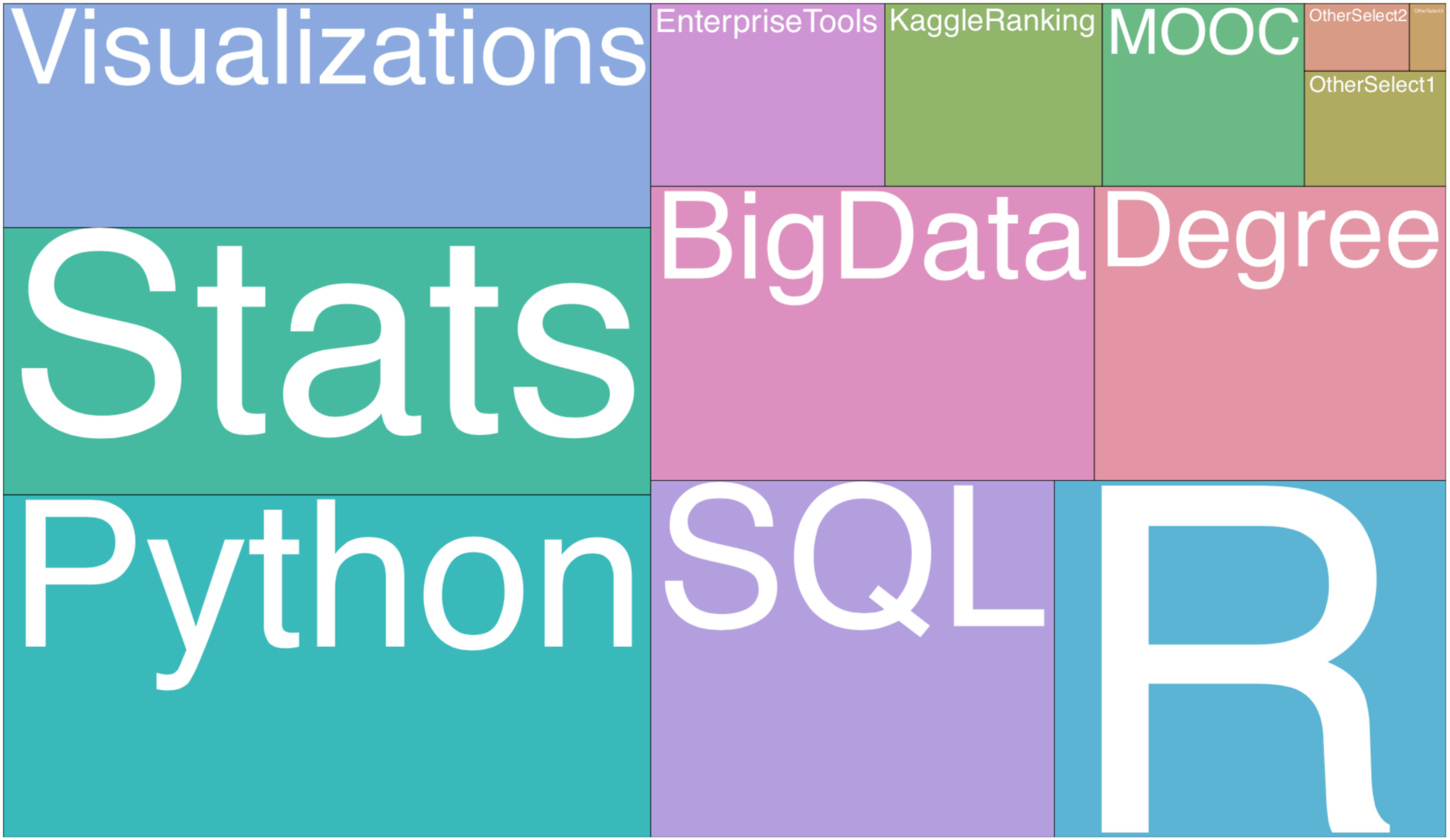
Kaggle data science survey: https://www.kaggle.com/surveys/2017

Tools that **data scientists** use



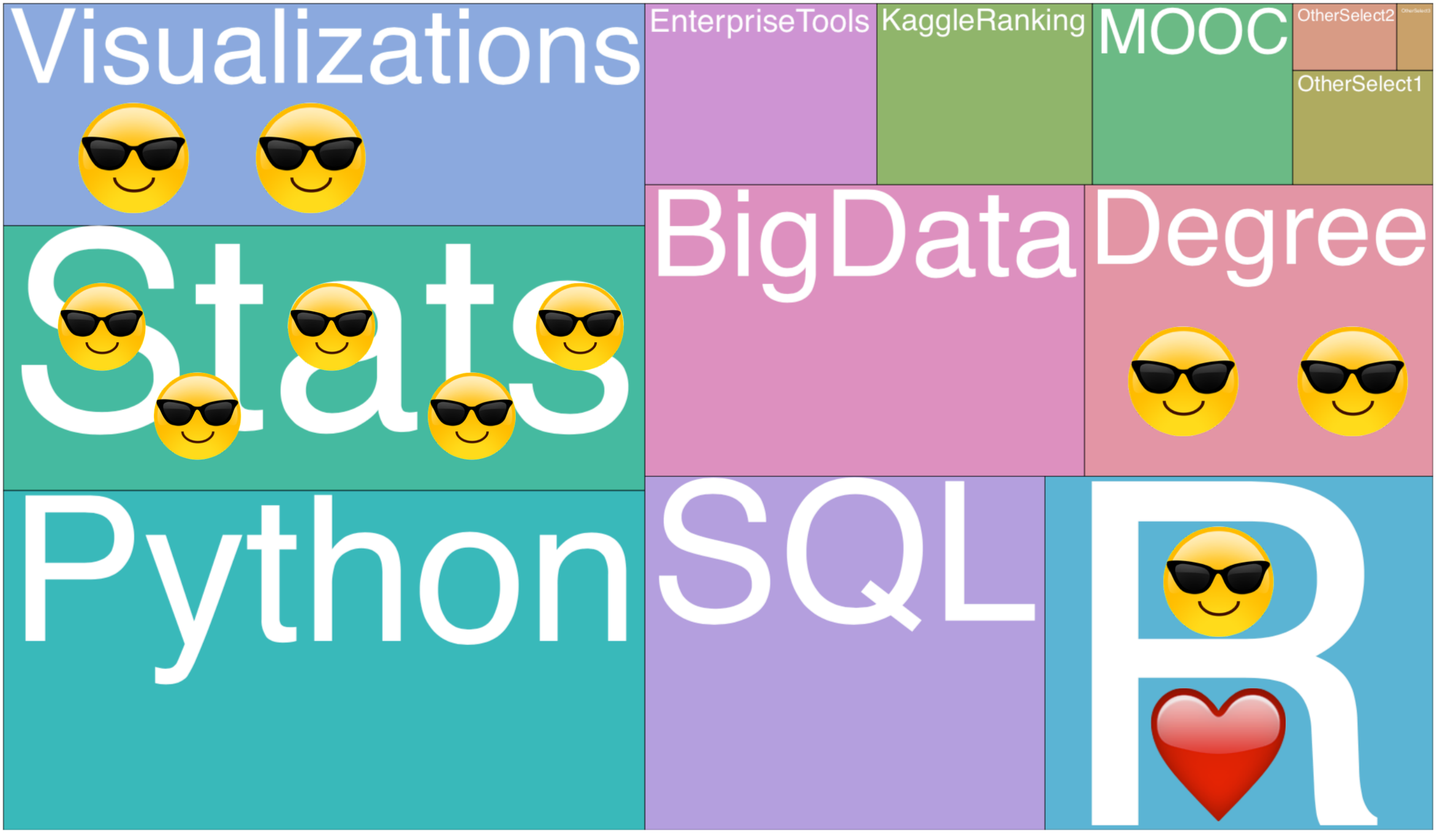
Kaggle data science survey: https://www.kaggle.com/surveys/2017

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



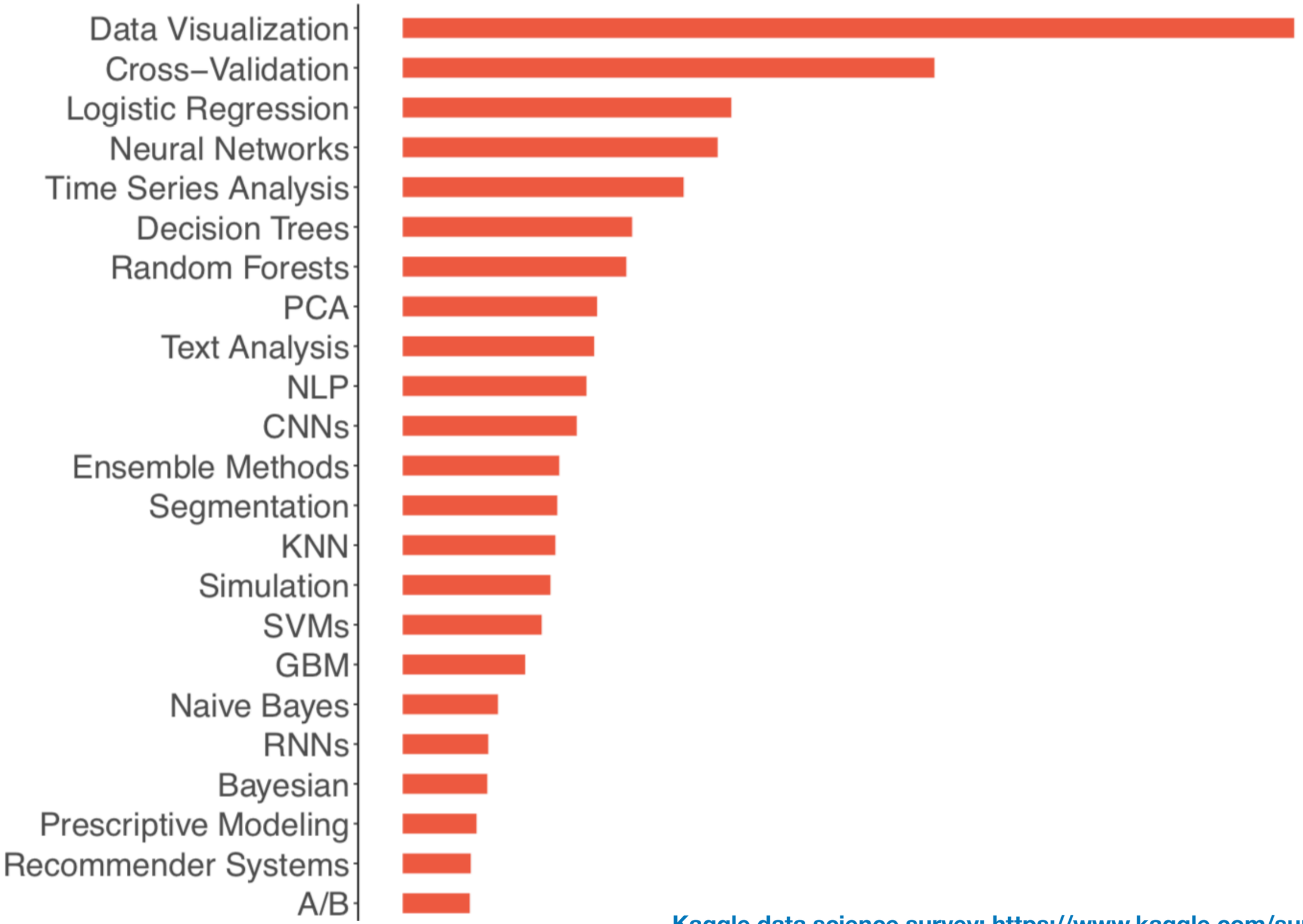
Kaggle data science survey: https://www.kaggle.com/surveys/2017

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



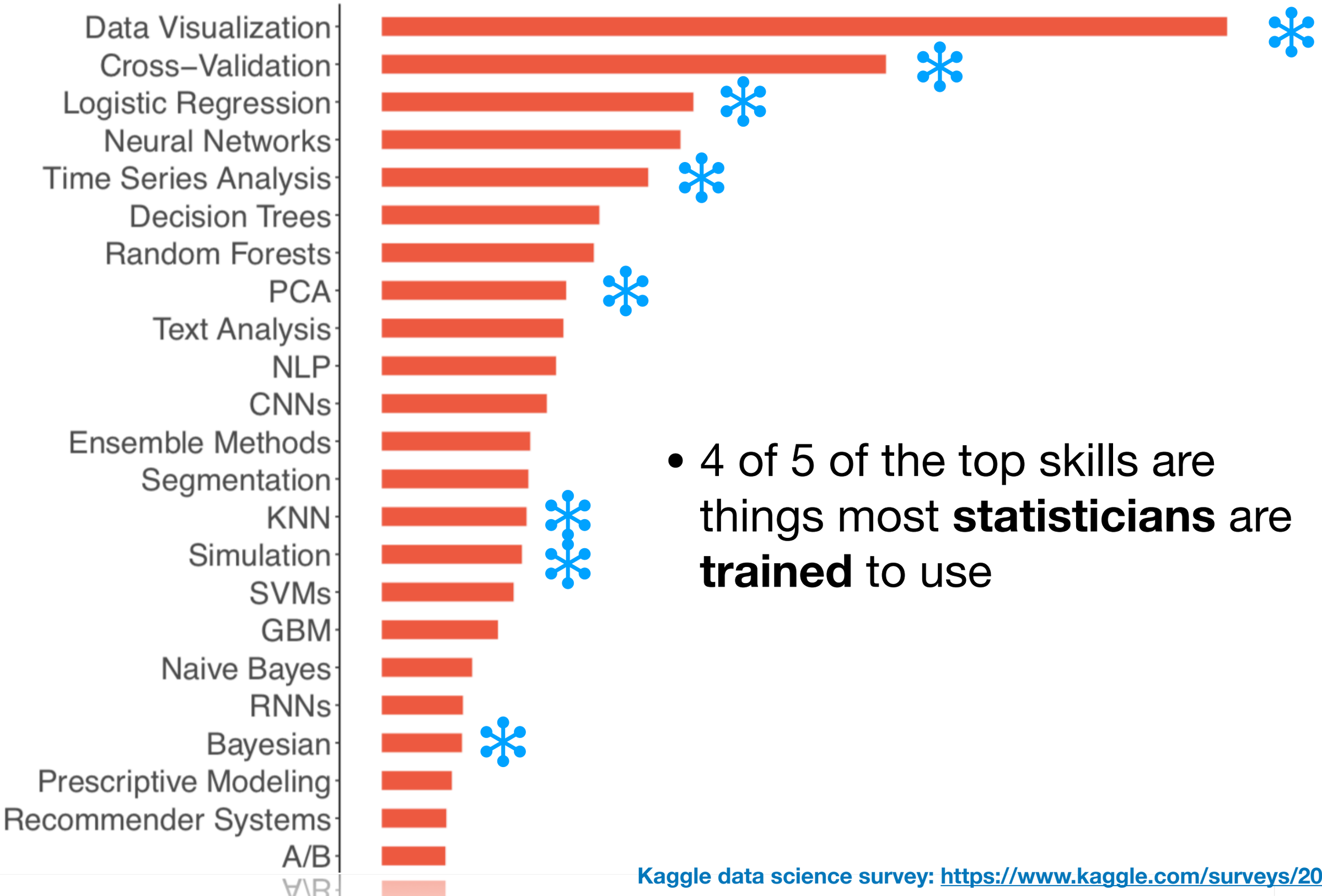
Kaggle data science survey: https://www.kaggle.com/surveys/2017

Most common **data science** techniques



Kaggle data science survey: https://www.kaggle.com/surveys/2017

Most common **data science** techniques



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

Kaggle data science survey: https://www.kaggle.com/surveys/2017

Data science tools & skills

Buzzy **data science** thing #1: Python

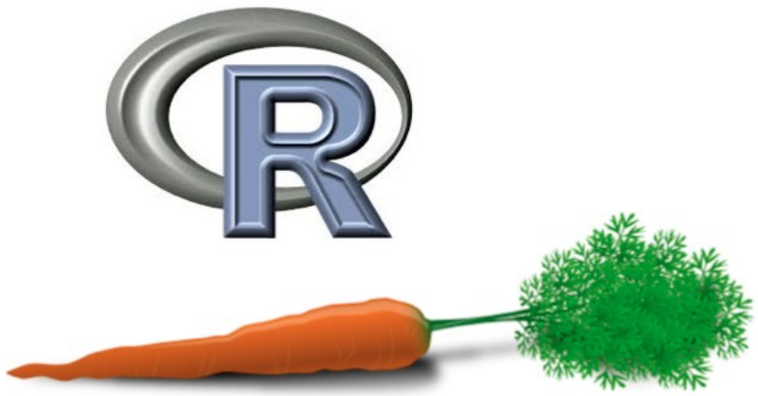
Survey: 63% recommend newcomers learn python first

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.)



H_2O.ai

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"

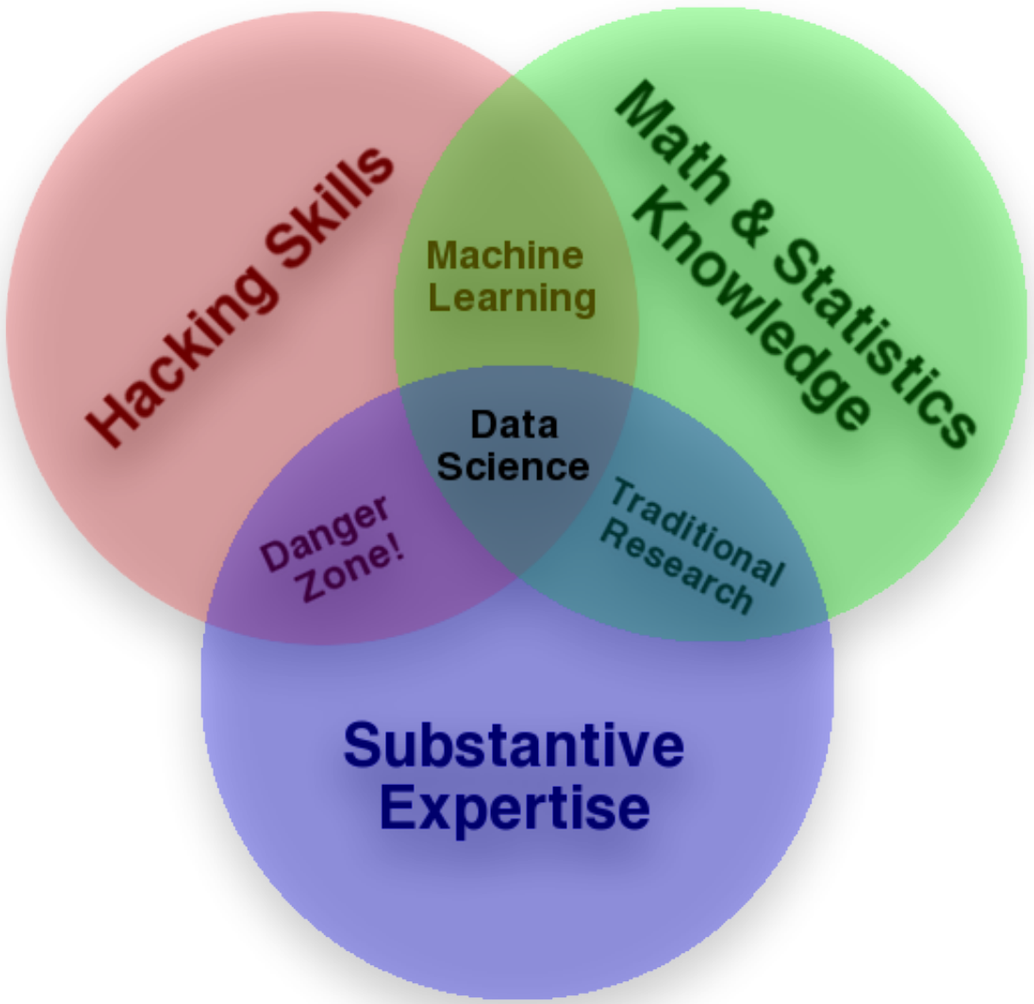
"If you're analysing data, you're doing **statistics**. You call it **data science** or informatics or analytics or whatever, but it's still **statistics**."

"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**?



Drew Conway, 25/3/2013
Source: http://drewconway.com/

YES

- Kaggle survey: statistics & data science are clearly complementary

Data science and **statistics**

Youtube: https://www.youtube.com/watch?v=C1zMUjHOLr4

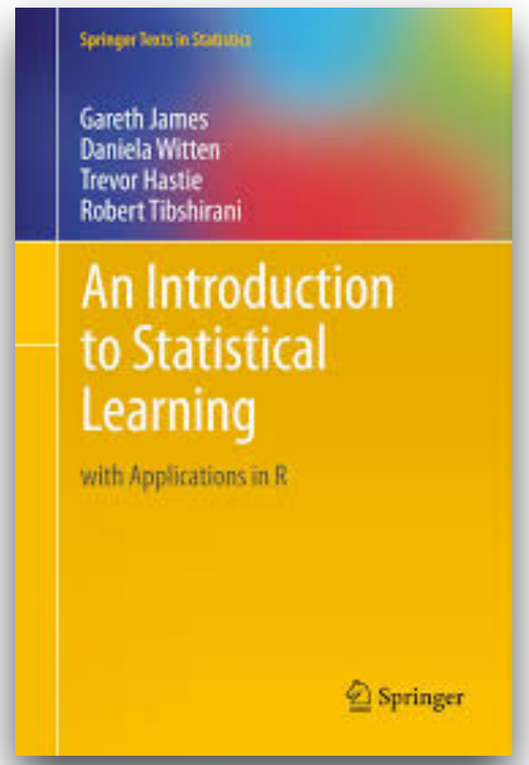
Stuff from **statistics** that I'm glad of

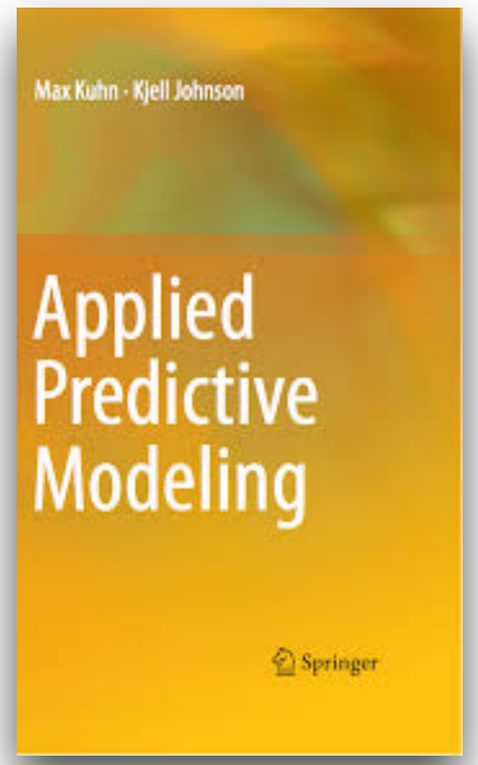
- Recognising point estimates
- Inference versus prediction
- Exposure to wide variety of domains

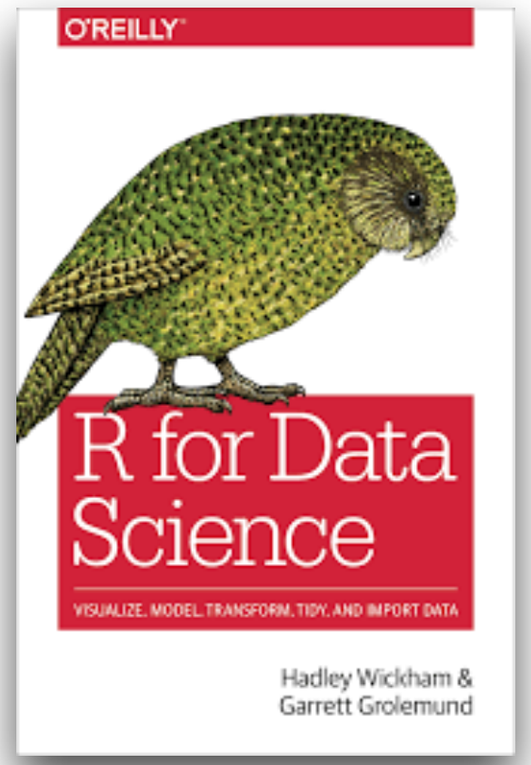
- Really knowing linear models
- `R`: `Rcpp`, `dplyr`, `shiny`, `flexdashboard`

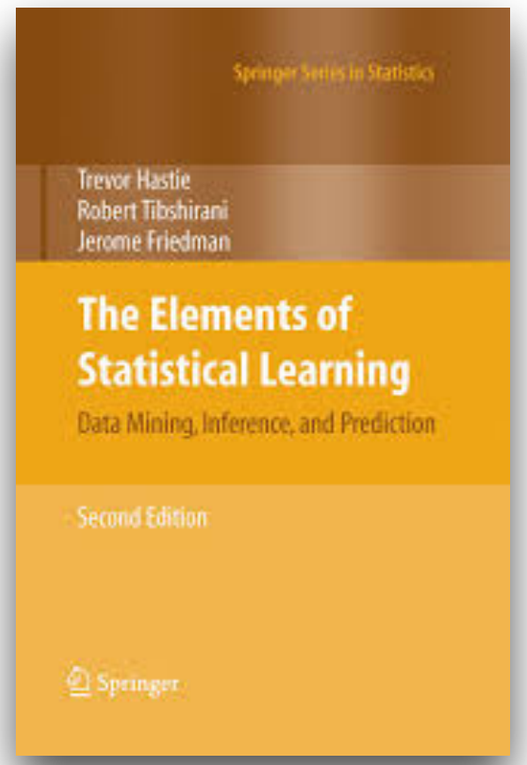
There is **always** a lot to learn…

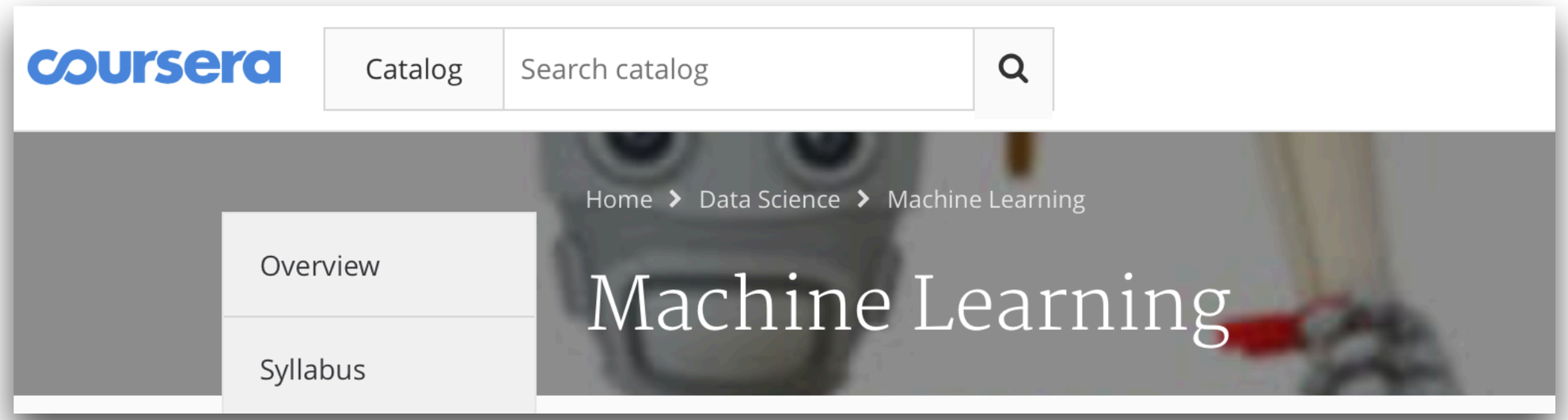
Survey: Average of 60% of learning time spent on online courses and self teaching











Absorb the culture with a podcast…

Not So Standard deviations

Talking Machines

https://alastairrushworth.github.io/Data-Science-Podcasts/

Wrap up…

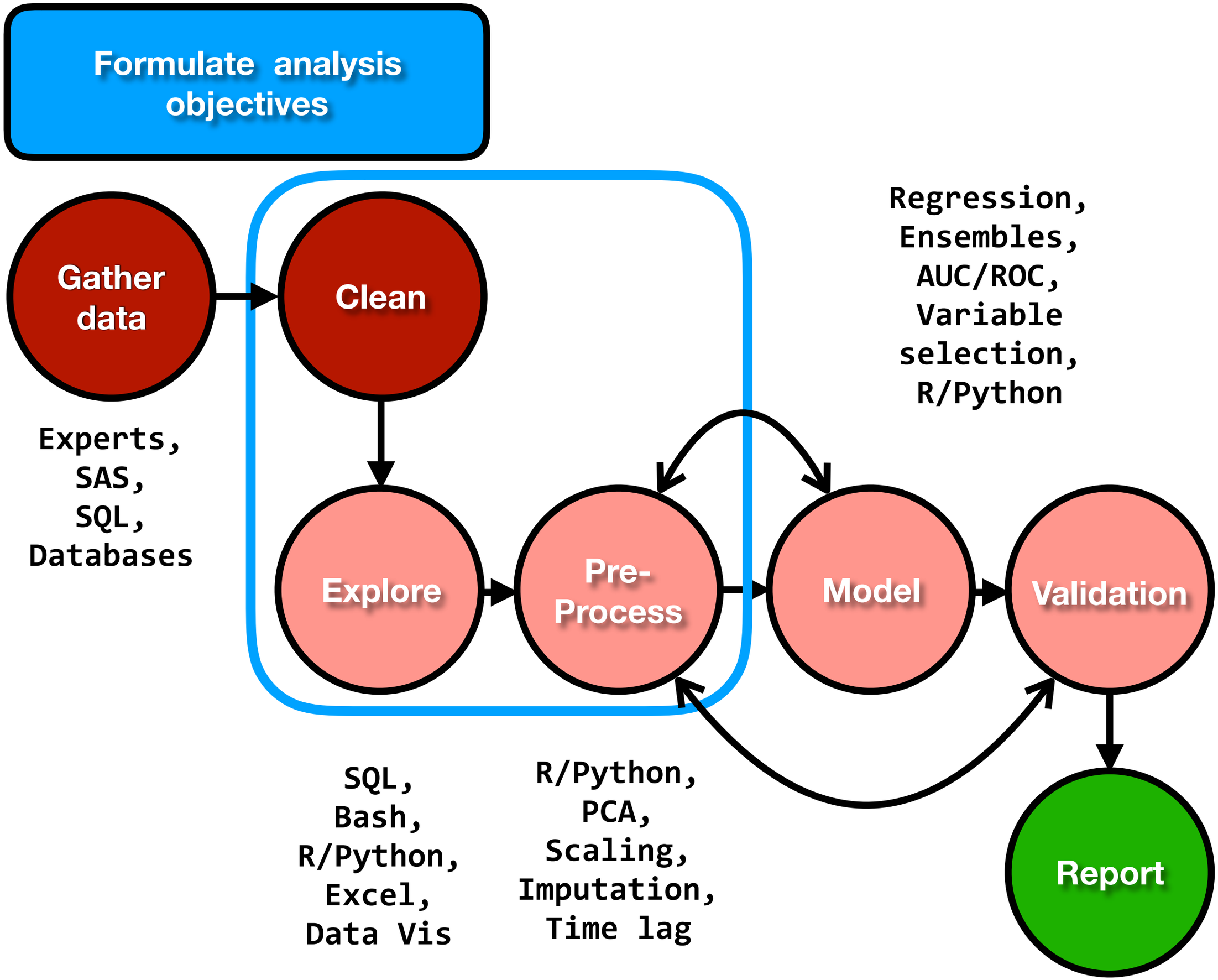
- **Data science** is more than just **statistics**
- But, **statistical** training is ideal for **data science**
- Forensic mindset & mathematical skill
- Scope for statistics to play a bigger role

  - inference, uncertainty & decisions
- Communication is key

- Don’t be put off by new software

Thanks

any questions?


Formulate analysis objectives
Gather data
Clean
Explore
Pre-Process
Model
Validation
Report
Experts, SAS, SQL, Databases
SQL, Bash, R/Python, Excel, Data Vis
R/Python, PCA, Scaling, Imputation, Time lag
Regression, Ensembles, AUC/ROC, Variable selection, R/Python