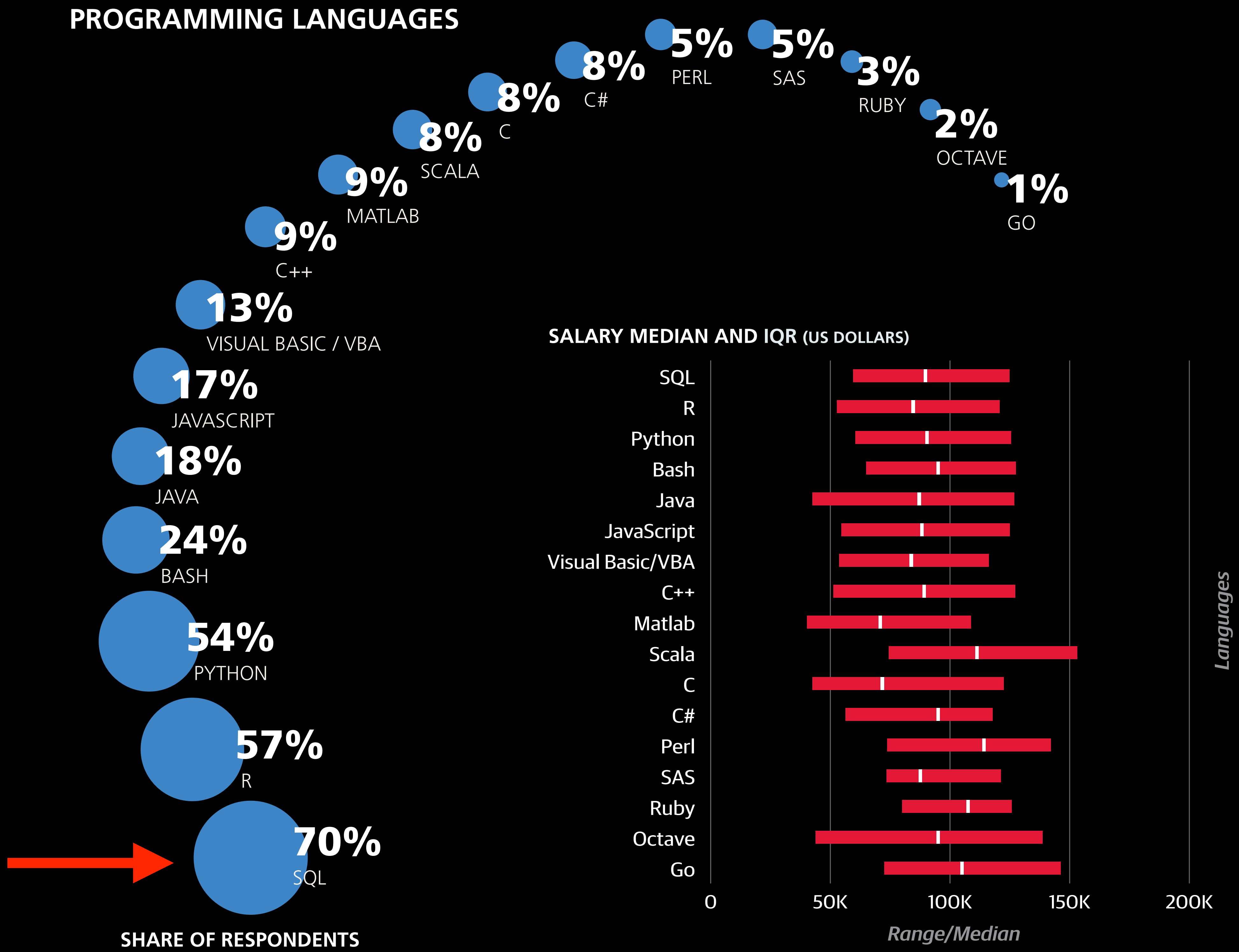
Data Management for Big Data Analytics

23 - 31 July
PKU

What do data analysts do?

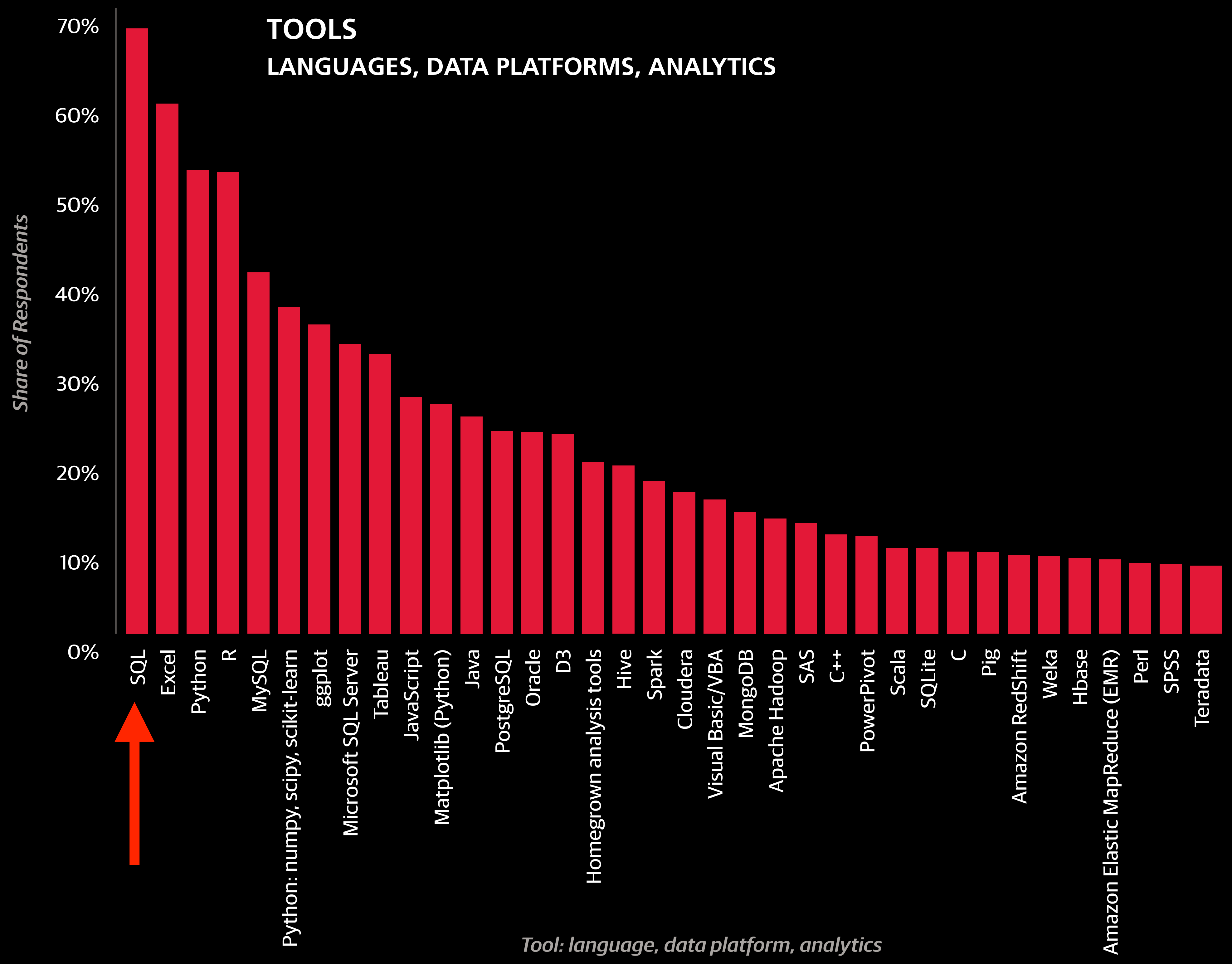


PROGRAMMING LANGUAGES

SHARE OF RESPONDENTS

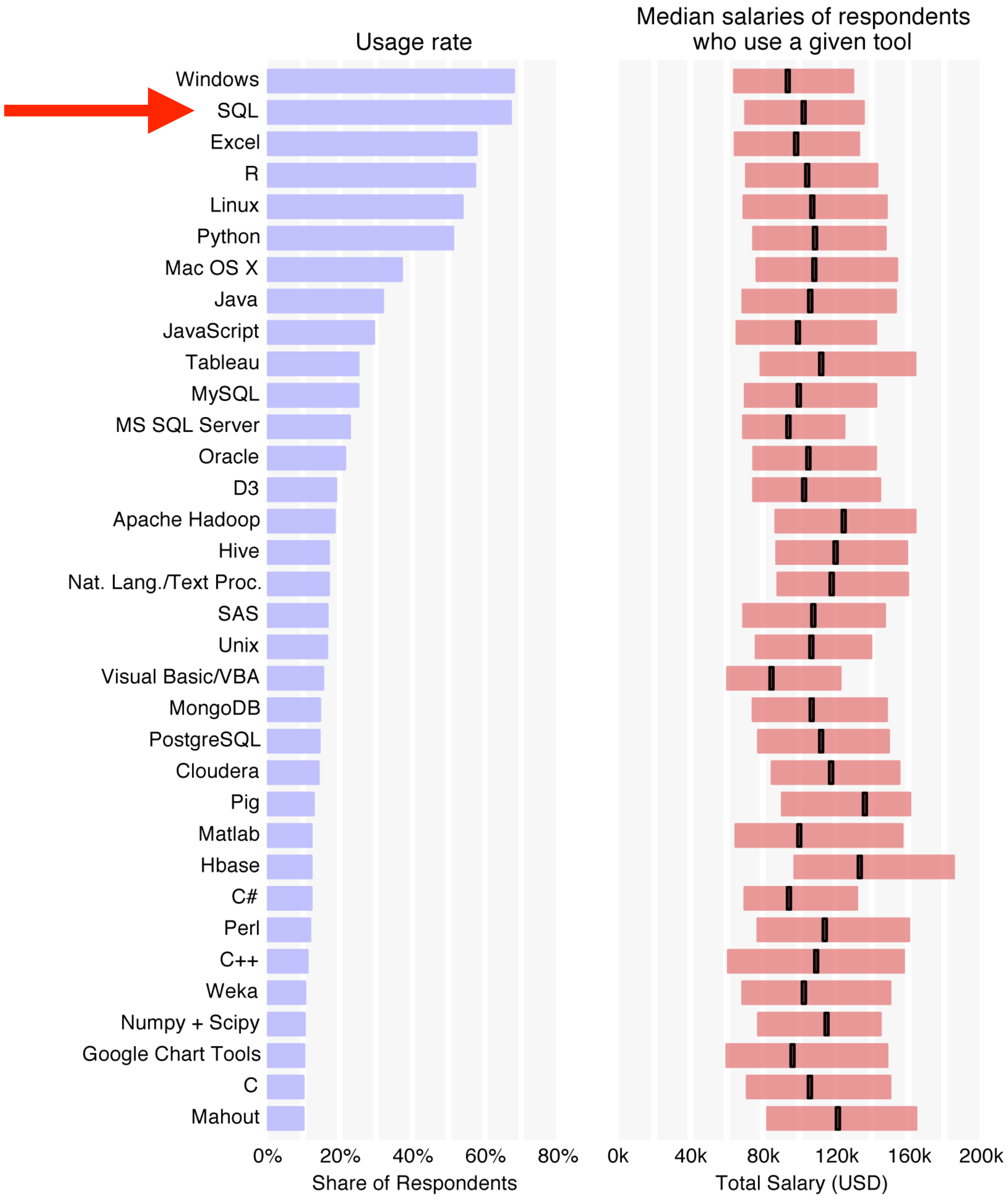
SALARY MEDIAN AND IQR (US DOLLARS)

Source: O’Reilly Data Science Salary Survey 2016

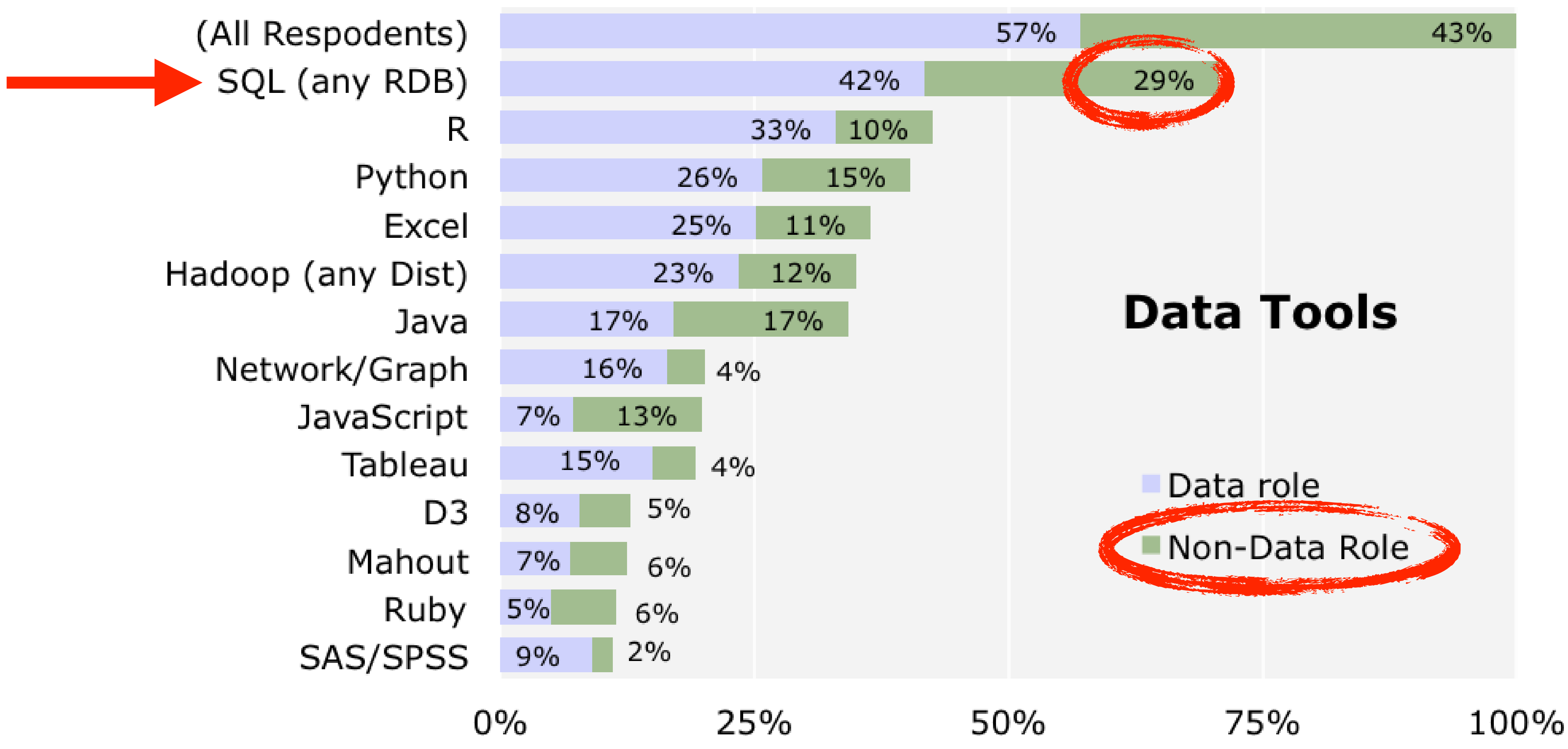


Source: O’Reilly Data Science Salary Survey 2015

Most commonly used tools (used by at least 10% of sample)

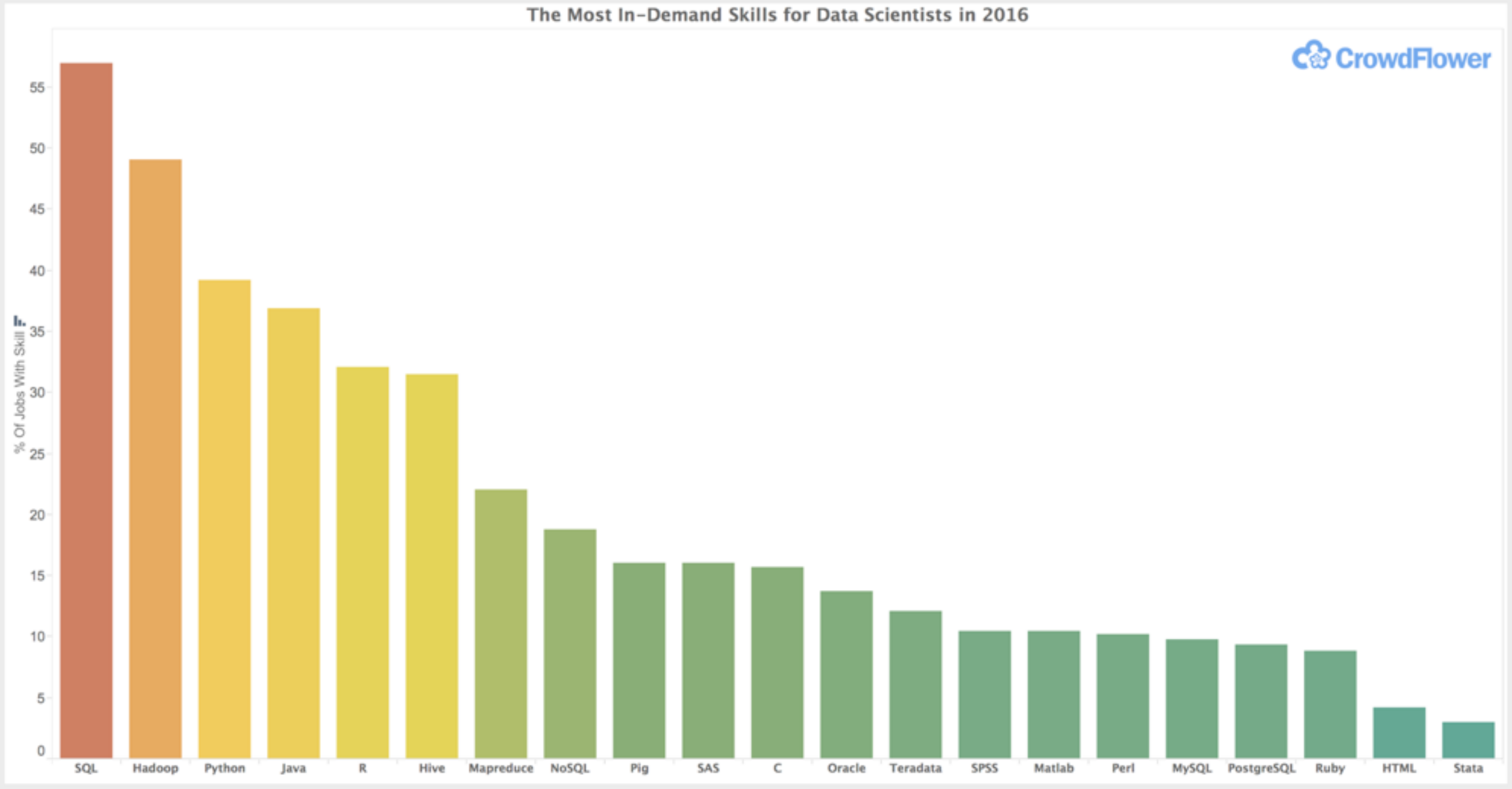


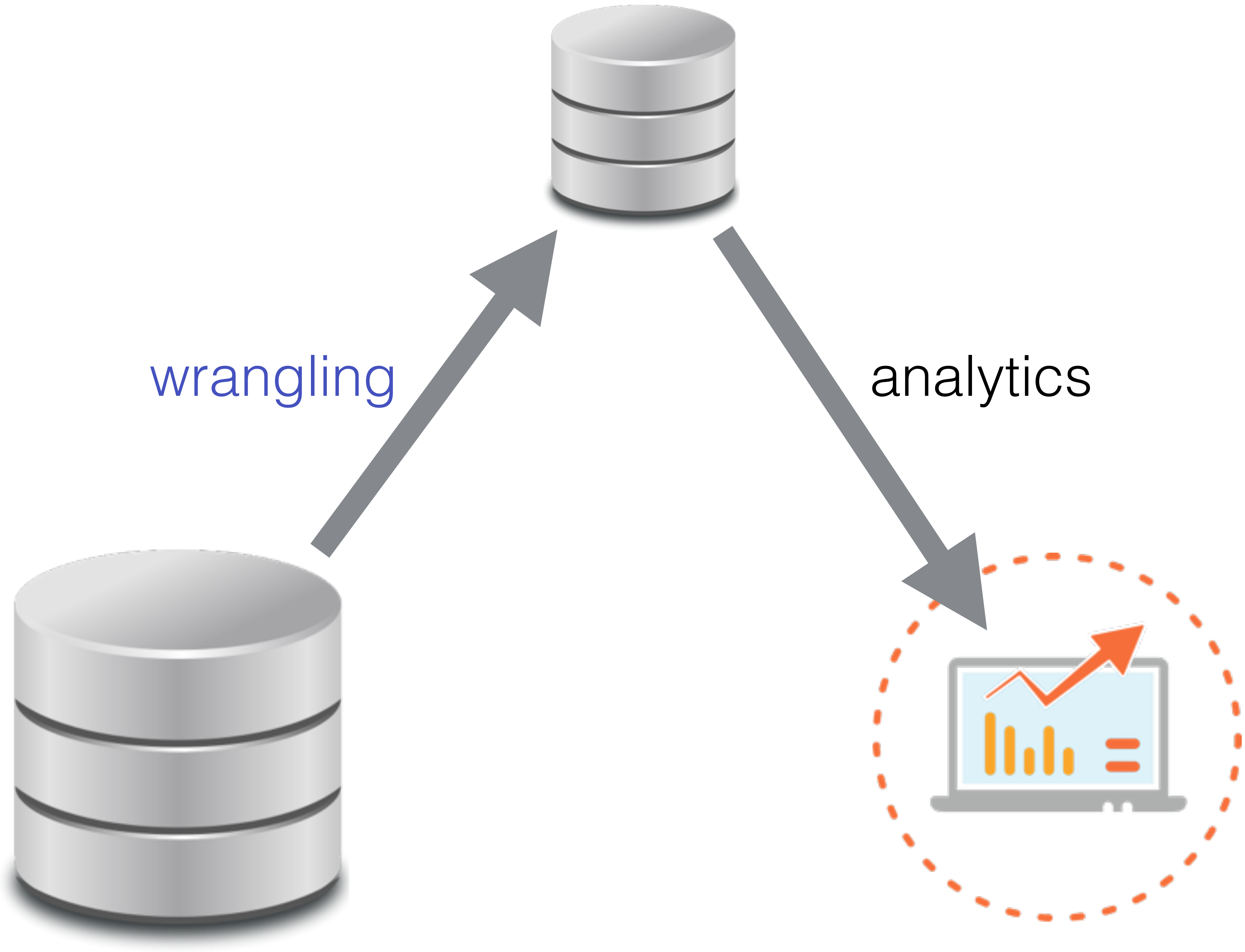
Source: O’Reilly Data Science Salary Survey 2014



Source: O’Reilly Data Science Salary Survey 2013

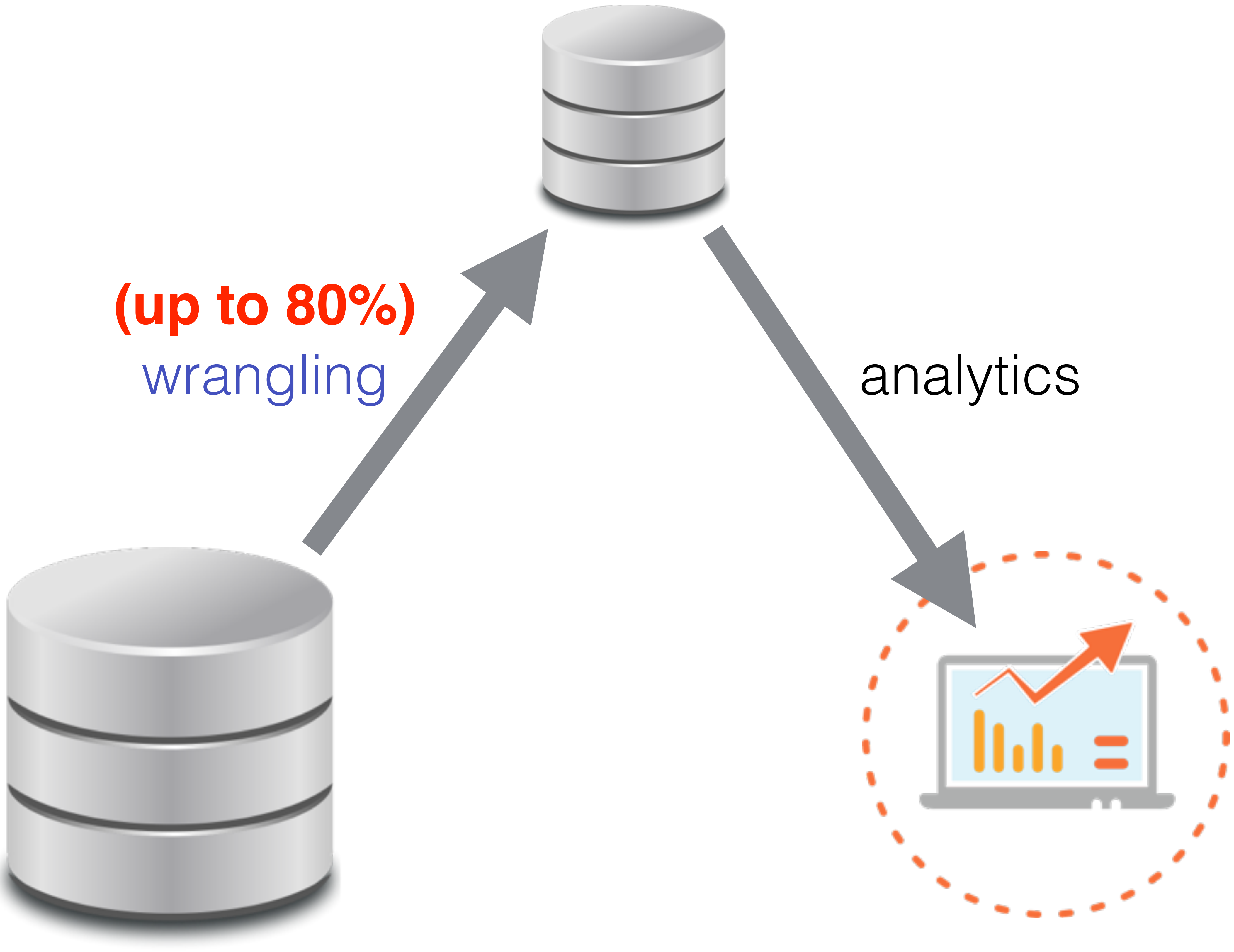
Future data scientists' favorite tools





wrangling

analytics



(up to 80%)

wrangling

analytics

Challenges

- **4 Vs of big data**
- **Volume - data is large**
- **Variety - data comes in different formats**
- **Veracity - data is not there/uncertain/dirty**
- Velocity - speed of change

Volume challenges

- Even scanning data can take hours/days/weeks
- How to get to the right data?
- Precise answers to queries are impossible
- Hence need to approximate

Variety challenges

- Different data formats
- Still much of the data is stored in relational DBMSs
- But other models catch up
- Most active these days is **graph data**
- We will look at it a lot

Veracity challenges

- How to deal with uncertainty or incompleteness? Relational databases (SQL) are really bad at it

- What to do if data is structured under a different schema, or different bits of data reside in different databases? (Data exchange and integration)

- What to do if data is supplemented with additional knowledge, e.g., an ontology, to compensate for missing data?

Course structure

- A quick of reminder of the basics of relational databases
- SQL: how well do we understand it?
- Volume: Conjunctive queries (many-way joins)
  - optimisation
  - approximation

- Volume: scale independence
- Variety: graph databases
    - theoretical languages
    - property graphs in Neo4j
- Variety: RDF data
- Variety: tree-structured data (XML) - depending on time

- Veracity: incomplete information and correct answers
- Veracity: data integration and exchange
- Veracity: answering queries with the help of ontologies

Evaluation

- Project
- There is a long list of papers on the web
- Choose one of them
- The goal is to write an essay, 6-8 pages
- It must present a summary of the paper that would be understood by someone who has not read the paper
- It should also provide some of your own ideas or further investigation about the paper

- Examples:

- analysis of the followup literature to see how these ideas were used

- ideas on improving algorithms in the paper, perhaps in some special cases

- an implementation of a theoretical algorithm to see how it performs

2-stage process

- Stage 1: this week (and weekend), choose the paper, read it quickly, and decide what you want to do in addition to summarising its ideas

- Have a quick presentation Tuesday 31 July (during the last 2 hours of the course)

- The do the proper writeup and email it to me, libkin@gmail.com, before *To Be Announced*