Database Systems
(INFR10070)

Dr Paolo Guagliardo

dbs-lecturer@ed.ac.uk

Fall 2018

Data

The **most important asset** of any enterprise

Must be effectively, efficiently and reliably

- collected and stored
- maintained and updated
- processed and analysed

to be turned into meaningful information

$\Longrightarrow$ Enable and **support decision making**

What is a database?

A collection of data items related to a specific enterprise, which is structured and organized so as to be more easily accessed, managed, and updated

Data**b**ase **M**anagement **S**ystem (DBMS)

- software package for creating and managing databases
- mediates interaction between end-users (incl. applications) and the database
- ensures that data is consistently organized and remains easily accessible

Why use a DBMS?

- Uniform data administration
- Efficient access to resources
- Data independence
- Reduced application development time
- Data integrity and security
- Concurrent access
- Recovery from crashes

Different kinds of data(bases)

- A **data model** is a collection of concepts for describing data
- A **schema** is a description of a particular collection of data, using a given data model

Relational databases ⟸ main focus of this course

Data organised in tables (relations) with typed attributes

Document stores ⟸ we will study some XML

Text documents structured using tags (or other markers)

Graph databases

Data organised in graph structures with nodes and edges

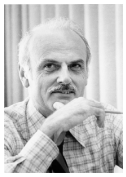
Key-value stores

Data organised in associative arrays (a.k.a. dictionaries or maps)

The relational model

First proposed by Edgar F. Codd in 1970

Simple idea: Organise data in **tables** (relations)



Schema

- Set of **table names**
- List of distinct (typed) **column names** for each table
- **Constraints** within a table or between tables

Instance

- Actual data (that is, the rows of the tables)
- Must satisfy typing and constraints

Example: relational database

Customer

CustID	Name	City	Address
cust1	Renton	Edinburgh	2 Wellington Pl
cust2	Watson	London	221B Baker St
cust3	Holmes	London	221B Baker St

Account

Number	Branch	CustID	Balance
243576	Edinburgh	cust1	-120.00
250018	London	cust3	5621.73
745622	Manchester	cust2	1503.82

Query languages

Used to ask questions (queries) to a database

Procedural

Specify a **sequence of steps** to obtain the expected result

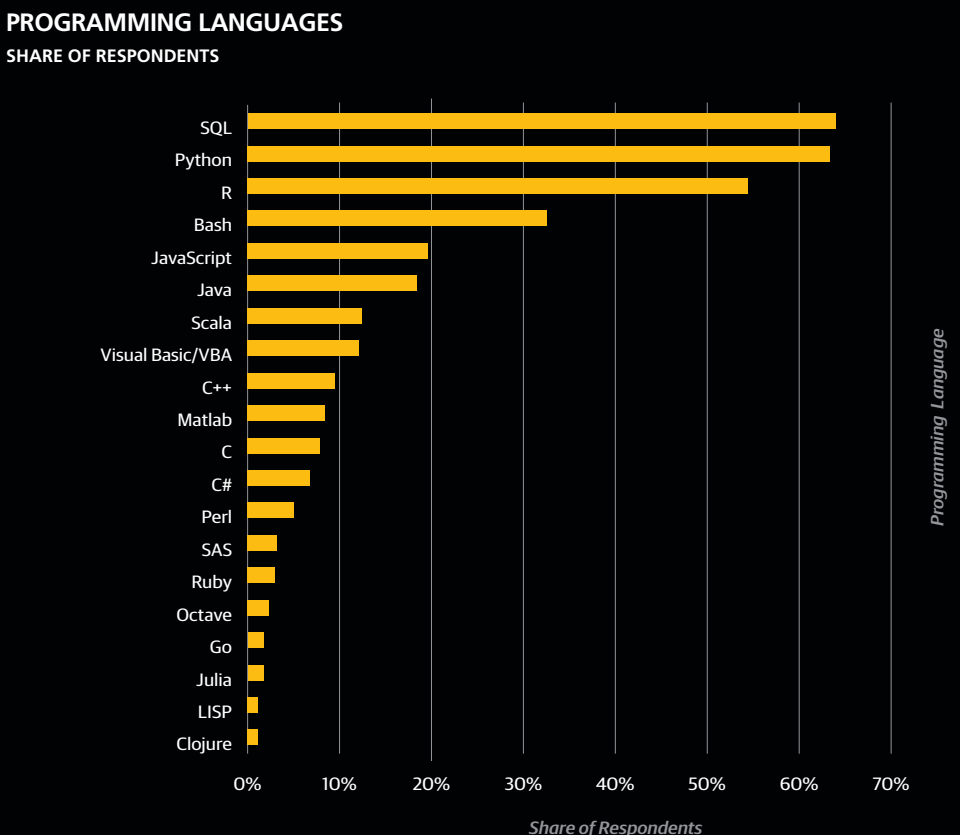
Declarative

Specify **what** you want not **how** to get it

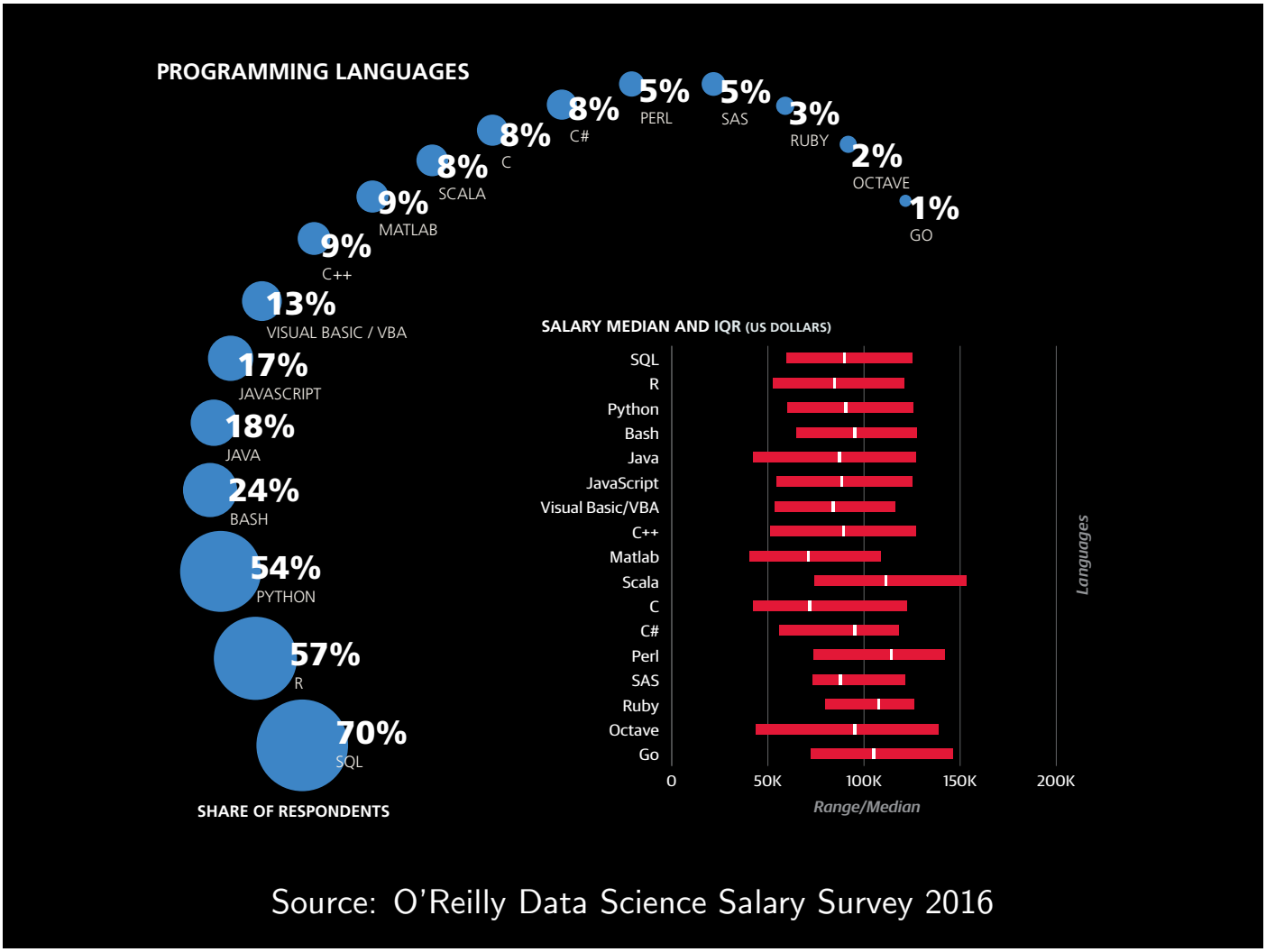
- Queries are typically asked in a declarative way
- DBMSs figure out internally how to translate a query into procedures that are suitable for getting the results

SQL

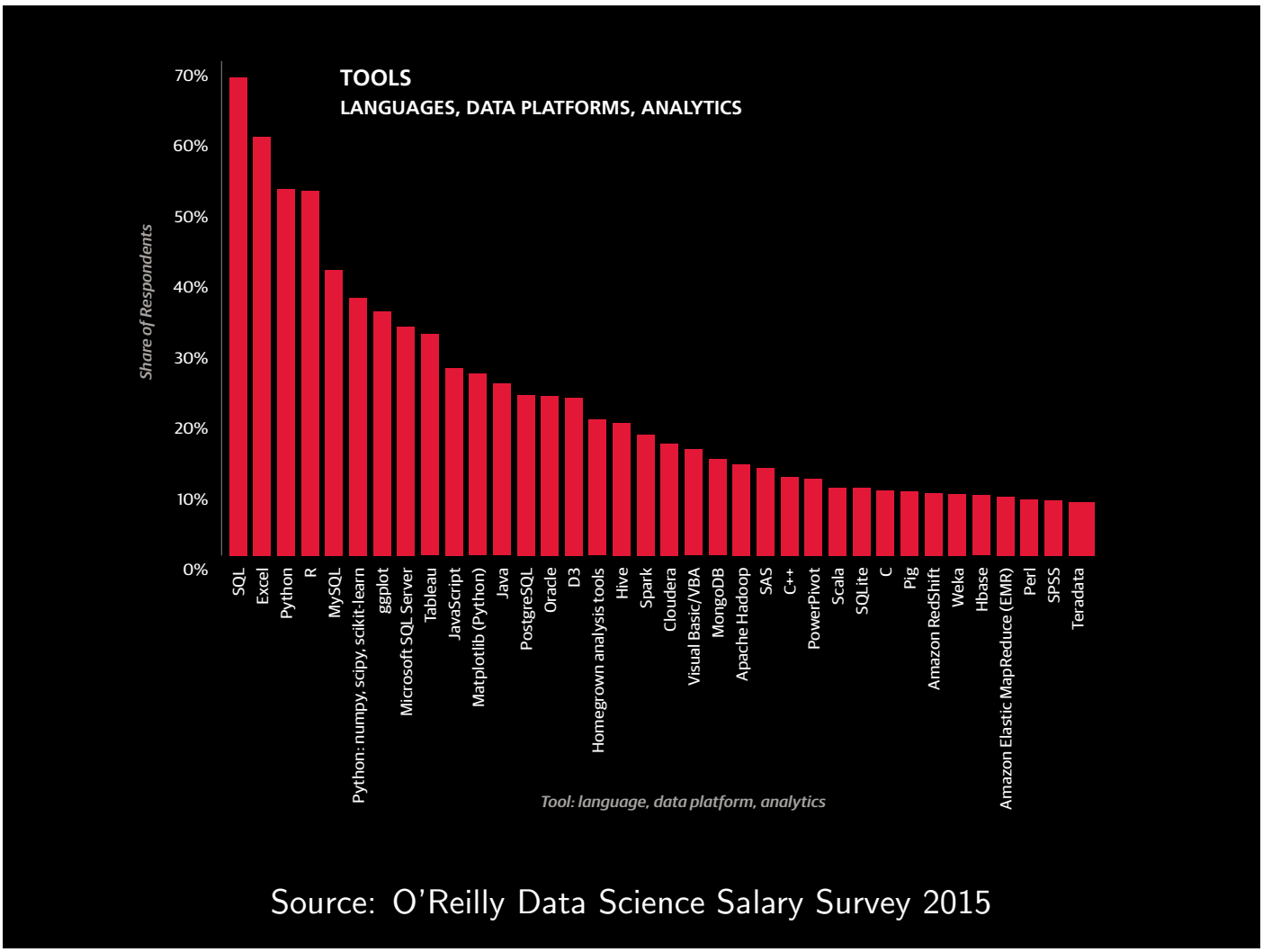
- **S**tructured **Q**uery **L**anguage
- Declarative language for querying relational databases
- Implemented in all major (free and commercial) RDBMSs
- First standardized in 1986 (ANSI) and 1987 (ISO); several revisions afterwards (latest Dec 2016)
- **$30B/year** business
- Most common tool used by data scientists



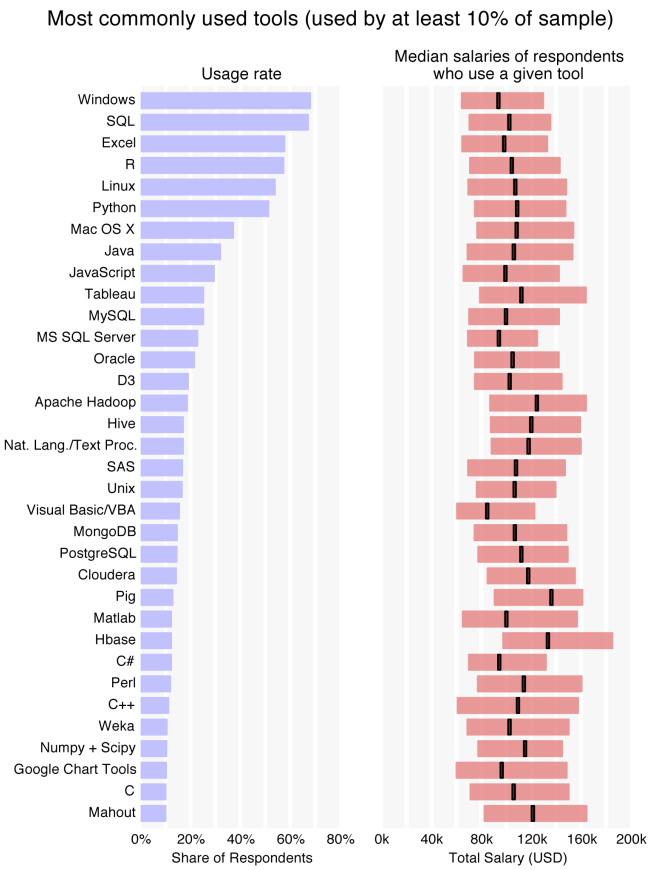
Source: O'Reilly Data Science Salary Survey 2017



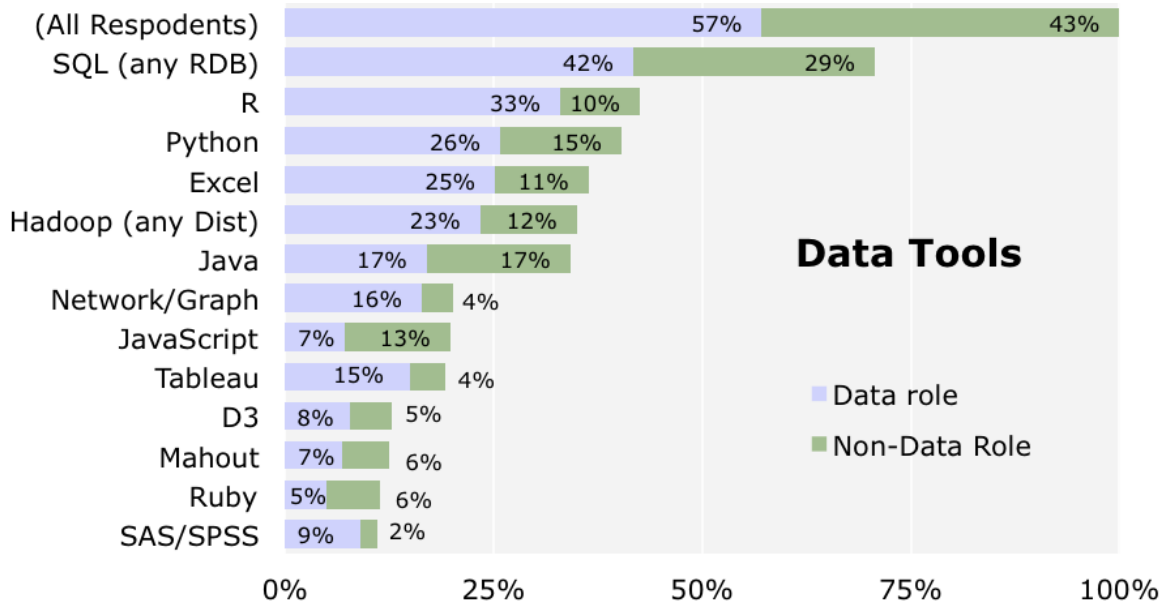
Source: O'Reilly Data Science Salary Survey 2016



Source: O'Reilly Data Science Salary Survey 2015

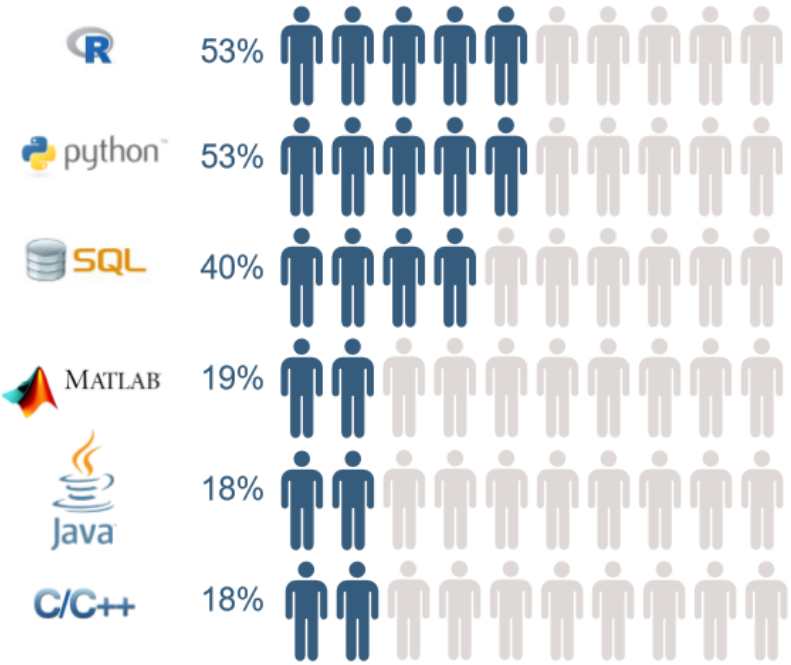


Source: O'Reilly Data Science Salary Survey 2014



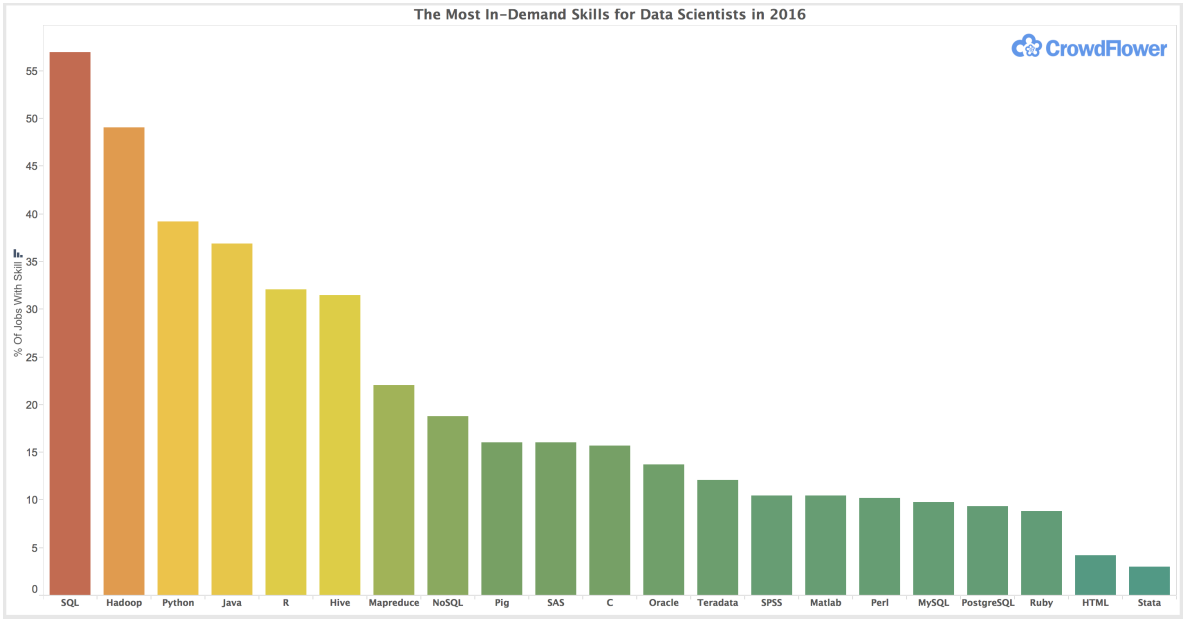
Source: O'Reilly Data Science Salary Survey 2013

What are the skills needed to become a data scientist in 2018?



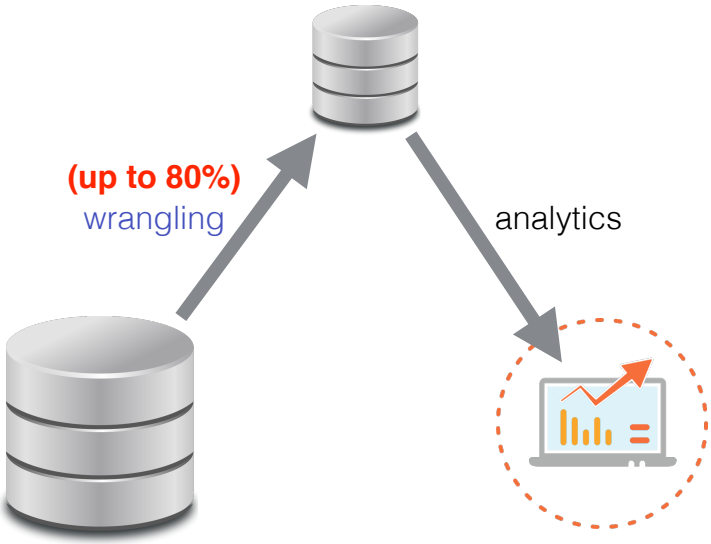
https://towardsdatascience.com/what-are-the-skills-needed-to-become-a-data-scientist-in-2018-d037012f1db2

What skills should data scientists have in 2016?



https://www.crowdflower.com/what-skills-should-data-scientists-have-in-2016/

That SQL is at the top is no surprise:
accessing data is the meat and potatoes of data analysis



Studying SQL is not enough

DBMSs encompass many areas of Computer Science:

- Operating systems
- Algorithms and data structures
- Formal logic
- (Programming) languages
- Multimedia
- ...

Goals of this course

- Teach you to be **good end-users** of a DBMS
- Provide you with **solid foundations** of how a DBMS works (and so also understand the job of DBAs and db developers)

Syllabus

- Query languages: **SQL**, relational algebra and calculus
- Database design: E-R diagrams, constraints, normal forms
- Deductive databases: Datalog and recursive queries
- Incomplete data: null values and certain answers
- Storage and indexing: B+ trees, hashing
- Query evaluation and optimisation: join strategies, query plans
- Scheduling and concurrency control:
  transaction management, serializability, locking
- Database access from applications: embedded/dynamic SQL
- Data warehousing and decision support:
  OLAP, view materialisation and maintenance
- Semistructured data:
  XML documents, DTDs, query languages for XML

Prerequisites

For undergraduates

Successful completion of Year 2

For all students

Some background in discrete mathematics:

- Set theory (sets, set operations, relations, orders)
- Combinatorics (permutations, combinations, partitions)
- Graph theory (directed/undirected graphs, trees)
- Computational complexity (complexity classes, decidability)
- Complexity analysis of algorithms (Big-O notation)
- **Logic** (predicate logic, inference, satisfiability)
  $\Longrightarrow$ essential to understand and write correct SQL queries

Textbook (1)

Main text — *Ramakrishnan, Gehrke*:
Database Management Systems
McGraw-Hill, 3rd edition

Highly **recommended** but not mandatory

Most lectures will be closely following this textbook

Availability

- **Main Library** (George Square): **3 copies** (3 hours loan)
- **Murray Library** (King's Buildings): **6 copies** (12 weeks loan)
- **Blackwell's** (Nicholson St): **10% student discount**

Textbook (2)

Further reading — *Abiteboul, Vianu, Hull*
Foundations of Databases
Addison-Wesley, 1995

- Mostly theoretical topics
- Out of print but freely available (for **personal use only**)
  http://webdam.inria.fr/Alice/

Course website

https://piazza.com/ed.ac.uk/fall2018/infr10070/home

Signup for the class at https://piazza.com/ed.ac.uk/fall2018/infr10070 with your **student email address** (e.g., `1234567@sms.ed.ac.uk`)

- No registration necessary to access the study material (lecture notes, exercises with solutions, assignments, announcements)
- Benefits of registering: notifications, class discussions, polls, get help easily from **classmates**, the **tutors** and **myself**

Rather than emailing questions, post them on Piazza

- You can post **privately** to instructors (tutors and me)
- You can post **anonymously** to instructors and classmates

Assessment: Coursework

Accounts for 25% of final mark

Two assignments

- Each requires writing SQL queries to a given specification
- Assigned in week x, due in week $x + 2$ for $x \in \{4, 8\}$
- Submission is via the `submit` command on DICE
- Marked automatically (details later on)

Assignment	Issued	Due	Worth
1	week 4	week 6	10%
2	week 8	week 10	15%

Assessment: Exam

Accounts for 75% of final mark

Diets

- December 2018: open to all students
- August 2019: resit exam (not for MSc students)

Structure

- Pen and paper (closed book)
- 5 to 8 problems, all of which must be solved for full marks
- Have a look at past exams: https://exampapers.ed.ac.uk/

Software: PostgreSQL

- Open-source, commercial-level RDBMS
- Installed on all DICE machines
- Available for Windows, Mac and Linux
- Very simple to compile and install on your laptop
- Each enrolled student has their own personal database (hosted on the university's central PostgreSQL server)
- "Getting started with PostgreSQL" lab in **week 3**
- You will use it to write SQL queries for the assignments

Tutorials

- They will start in week 4
- Discuss (formative) exercises assigned throughout the course
- Tutorial attendance is **mandatory**
  (absence will be reported to your Personal Tutor)
- You will choose which tutorial group to attend
  (up to maximum capacity of the room)
- If you miss one tutorial, go to another one
  (and or talk to other students in your group)
- Tutorial sheets will be made available in advance
- Solutions to tutorial exercises will be posted on Piazza

Other stuff

Lecture recording

- Lectures will be recorded
- Recordings will be available on LEARN

Lecture notes

- Slides will be usually made available before class
- You can access last year's slides at
  https://piazza.com/ed.ac.uk/fall2017/infr10070/resources

Office hours

- By appointment (IF-5.11)
- I am usually available after class