POIR 613: Computational Social Science

Pablo Barberá

School of International Relations
University of Southern California
pabloobarbera.com

Course website:
pabloobarbera.com/POIR613/
Introduction to **SQL**
Databases

- **Database systems**: computerized mechanisms to store and retrieve data.
- **Relational databases**: data is represented as tables linked based on common keys (to avoid redundancy).
SQL

- SQL (pronounced S-Q-L or SEQUEL) is a language designed to query relational databases
- Used by most financial and commercial companies
- The result of an SQL query is always a table
- It’s a nonprocedural language: define inputs and outputs; how the statement is executed is left to the optimizer
- How long SQL queries depends on optimization that is opaque to user (which is great!)
- SQL is a language that works with many commercial products:
  - Oracle Database, SQL Server (MS), MySQL, PostgreSQL, SQLite (all three open-source), Google BigQuery, Amazon Redshift...
  - Performance will vary, but generally faster than standard data frame manipulation in R (and much more scalable)
Components of a SQL query

- **SELECT** columns
- **FROM** a table in a database
- **WHERE** rows meet a condition
- **GROUP BY** values of a column
- **ORDER BY** values of a column when displaying results
- **LIMIT** to only X number of rows in resulting table

- Always required: **SELECT** and **FROM**. Rest are optional.
- **SELECT** can be combined with operators such as **SUM**, **COUNT**, **AVG**...
- To merge multiple tables, you can use **JOIN**
SQL at scale

Google BigQuery

► One of many commercial SQL databases available (Amazon RedShift, Microsoft Azure, Oracle Live SQL...)

► Used by many financial and commercial companies

► **Advantages:**
  ► Integration with other Google data storage solutions (Google Drive, Google Cloud Storage)
  ► Scalable: same SQL syntax for datasets of *any* size
  ► Easy to collaborate and export results
  ► Affordable pricing and cost control
  ► API access allows integration with R or python
  ► Excellent documentation