Relational Data

DWDC

- Relational data is organized in tables consisting of columns and rows

- Fields (columns) consist of a column name and data type constraint

- Records (rows) in a table have a common field (column) structure and order

- Records (rows) are linked across tables by key fields

Relational Data Model: Codd, Edgar F. "A Relational Model of Data for Large Shared Data Banks" (1970)

Introduction to SQL

- SQL ("Structured Query Language") is a declarative data definition and query language for relational data

- SQL is an ISO/IEC standard with many implementations in common database management systems (a few below)

Structured Query Language: ISO/IEC 9075 (standard), first appeared 1974, current version SQL:2011

Sidebar: Which database system should I use?

1. Use the one your data is in

2. Unless you need specific things (performance, functions, etc.), use the one you know best

3. If you need other stuff or you've never used a database before:

   A. SQLite: FOSS, one file db, easy/limited

   B. PostgreSQL: FOSS, Enterprise-ready

The above are my opinions based on experience. Others may disagree, and that's OK.

DWDC

SQL: Working with Objects

- Data Definition Language (DB Objects)

  - CREATE (table, index, view, function, …)

  - ALTER (table, index, view, function, …)

  - DROP (table, index, view, function, …)

SQL: Working with Rows
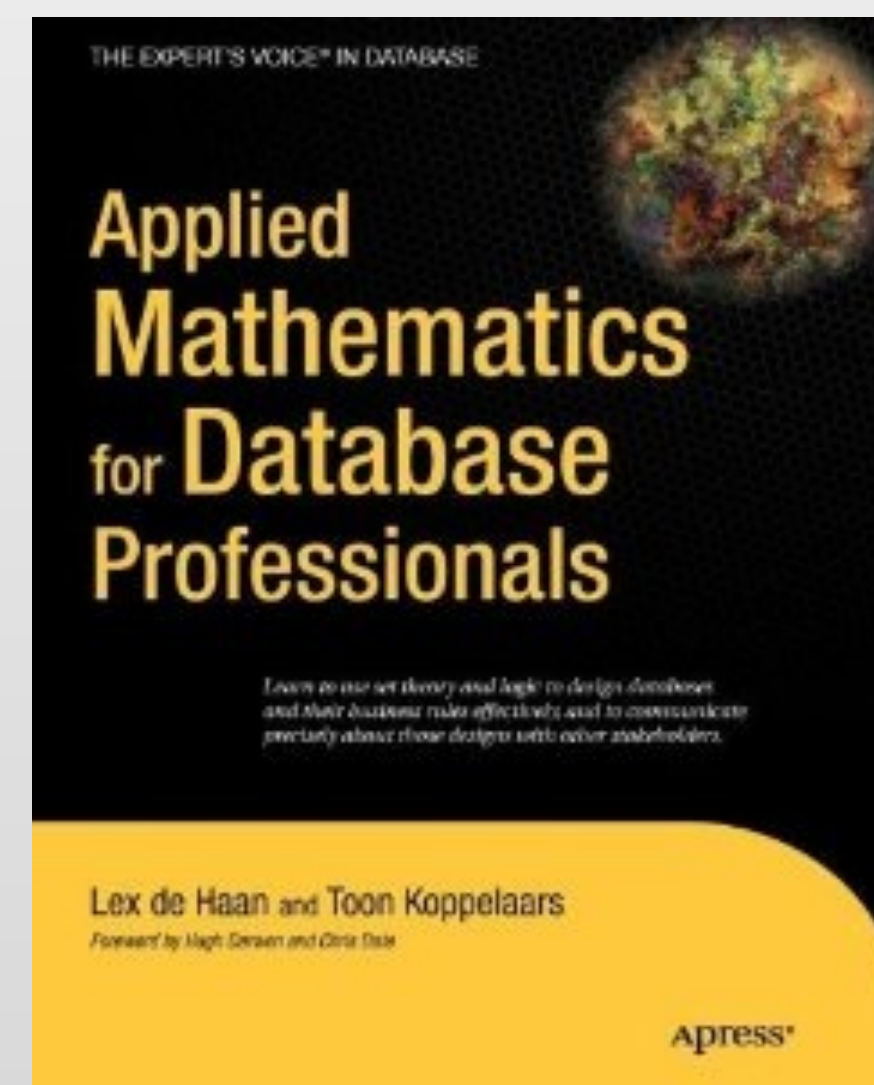
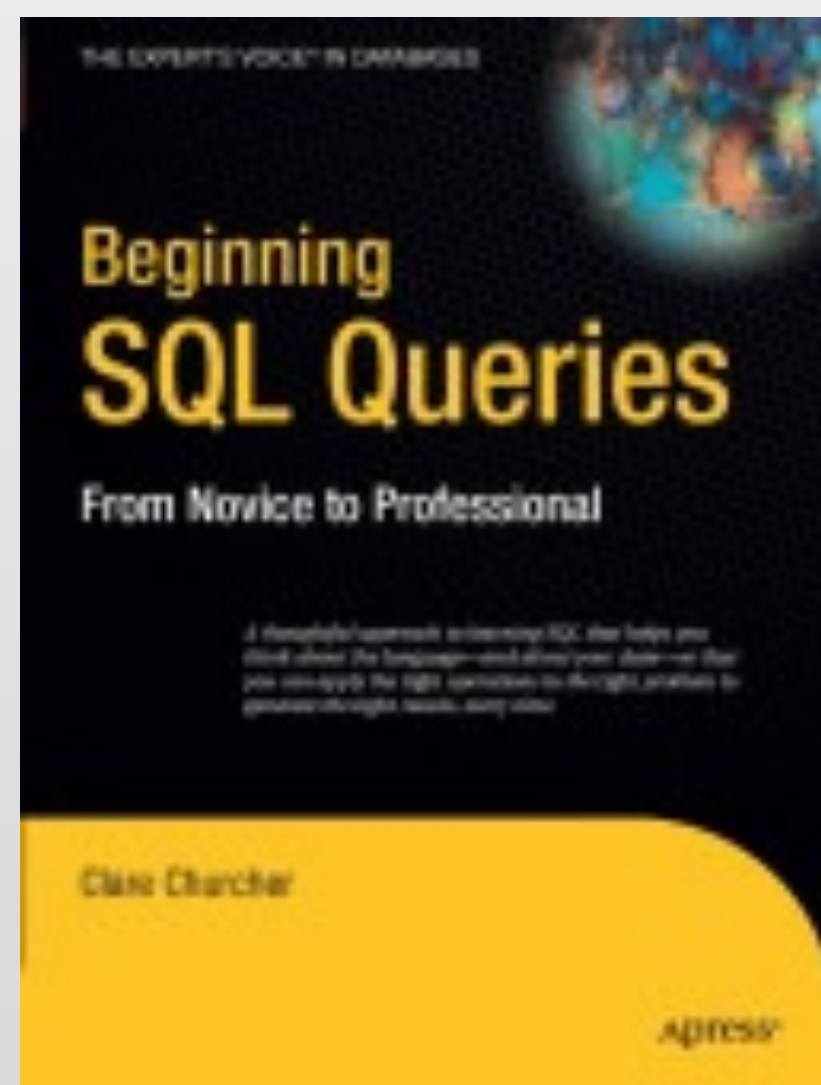- Query Language (Records)
  - SELECT ... FROM ...
  - INSERT INTO ...
  - UPDATE ... SET ...
  - DELETE FROM ...

SQL: SELECT Statement

DWDC
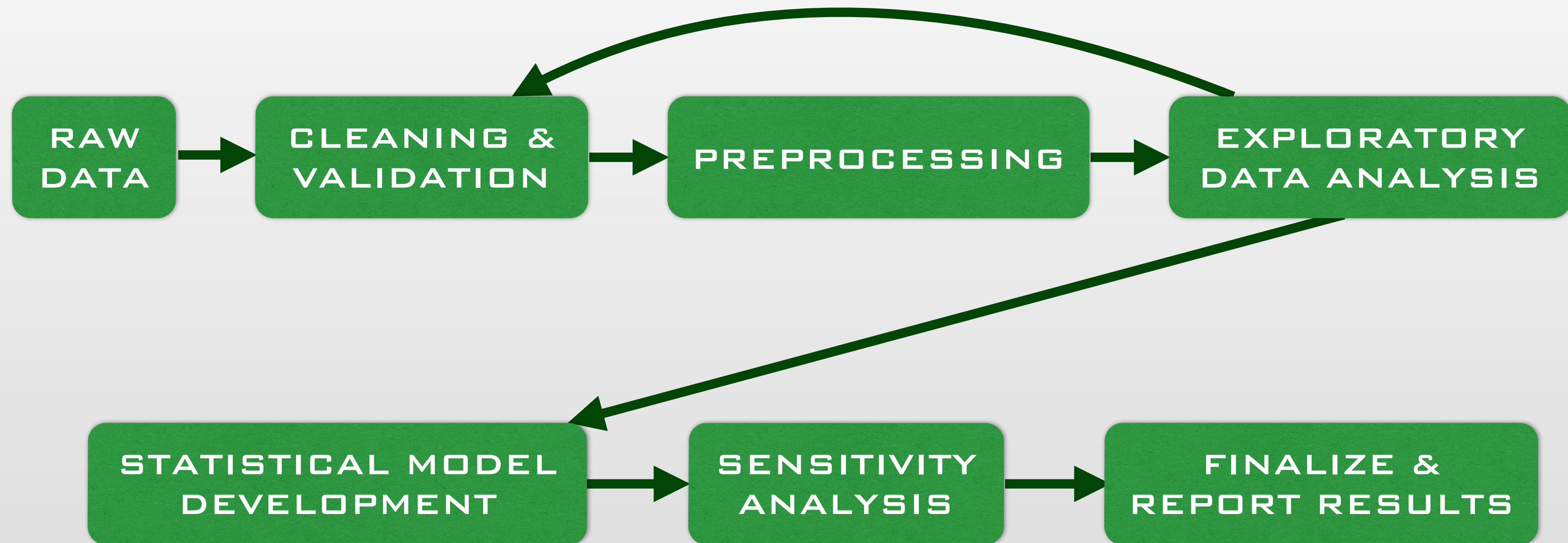
- SELECT <col_list> FROM <table> ...
  - Merging: JOIN clause
  - Concatenation: UNION clause
  - Filtering: WHERE clause
  - Aggregation: GROUP BY clause
  - Aggregated filtering: HAVING clause
  - Sorting: ORDER BY clause

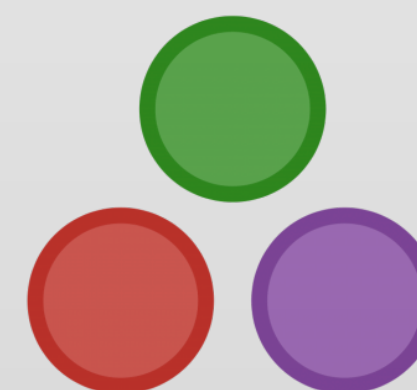# SQL Beginner Resources

- Basic SQL Commands Reference: http://www.cs.utexas.edu/~mitra/csFall2013/cs329/lectures/sql.html

SQL in other languages

- R with libraries

  - RPostgreSQL, dplyr

- Python with modules

  - psycopg2, SQLAlchemy

- Julia with packages (in dev)

  - PostgreSQL, DBI

Why do reproducible analyses?

- The standard for belief in science is replication, but that's often impossible

- Reproducibility is the next best thing:

  - assumes observed raw data is "good"

  - allows data analysis claims to be validated independent of natural processes that generated the data

What makes this reproducible?

DWDC

Scripted; Code Provided

Raw Data Provided

Analysis Data Provided

RAW DATA → CLEANING & VALIDATION → PREPROCESSING → EXPLORATORY DATA ANALYSIS

Scripted; Code Provided

Scripted with figure generation; methodology in report

STATISTICAL MODEL DEVELOPMENT → SENSITIVITY ANALYSIS → FINALIZE & REPORT RESULTS

Scripted with figure generation; methodology in report

Scripted with figure generation; methodology in report

DWDC

Now, let's look at some code!

Ryan B. Harvey

http://datascientist.guru
ryan.b.harvey@gmail.com
@nihonjinrxs
+ryan.b.harvey

Day Job

IT Project Manager

Office of Management and Budget

Executive Office of the President

Side Job

Data Scientist & Software Architect

Kitchology Inc.

DWDC