DWDC

Data Wrangling in SQL & Other Tools

Scripting reproducible and understandable data wrangling and analysis pipelines with tabular and relational data

Ryan B. Harvey

June 4, 2014

DATAWRANGLERSDC

Relational Data

- 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)

Sidebar 1: Why should I use a database system?

1. You care about strong data types, type validation and data access controls
2. You need to relate multiple tables together via common fields
3. Your data is larger than a few 10s to 100 MB, making file parsing onerous
4. You need to subset or aggregate your data often based on field values

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

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 2: 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.

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 ...

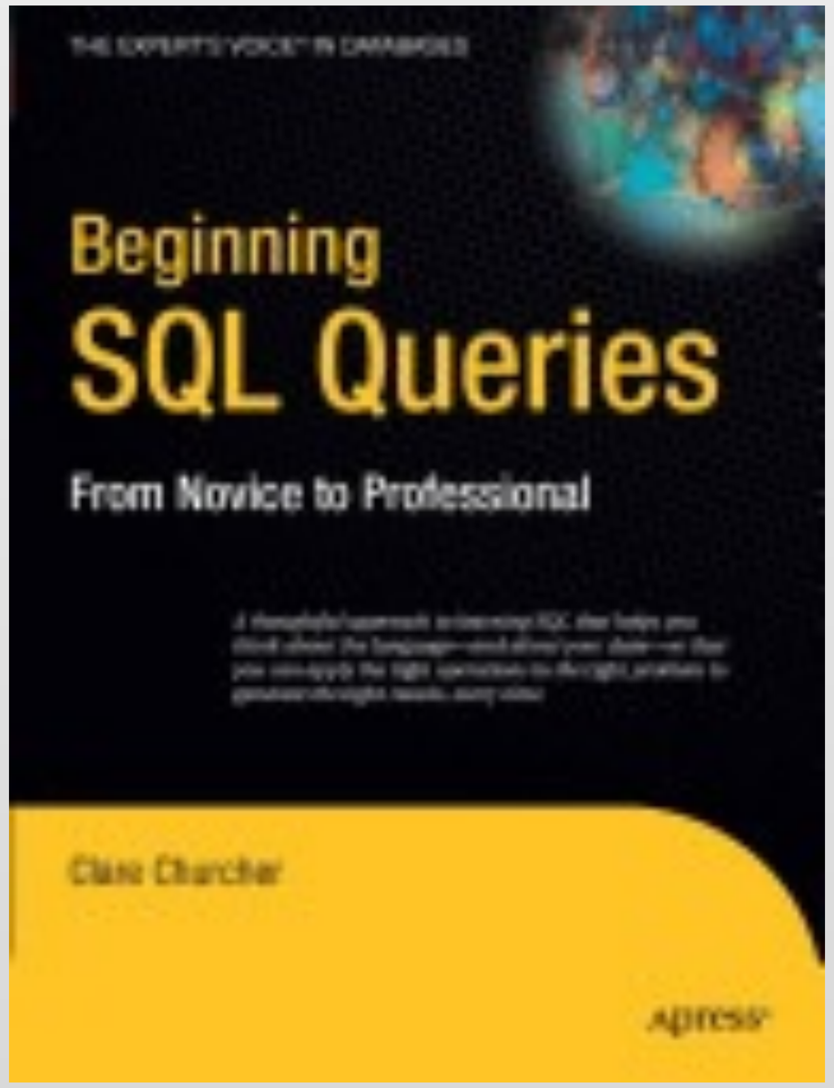
SQL: SELECT Statement

DWDC

- SELECT <col_list> FROM <table> ...
  - Merging: JOIN clause
  - Row binding: 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



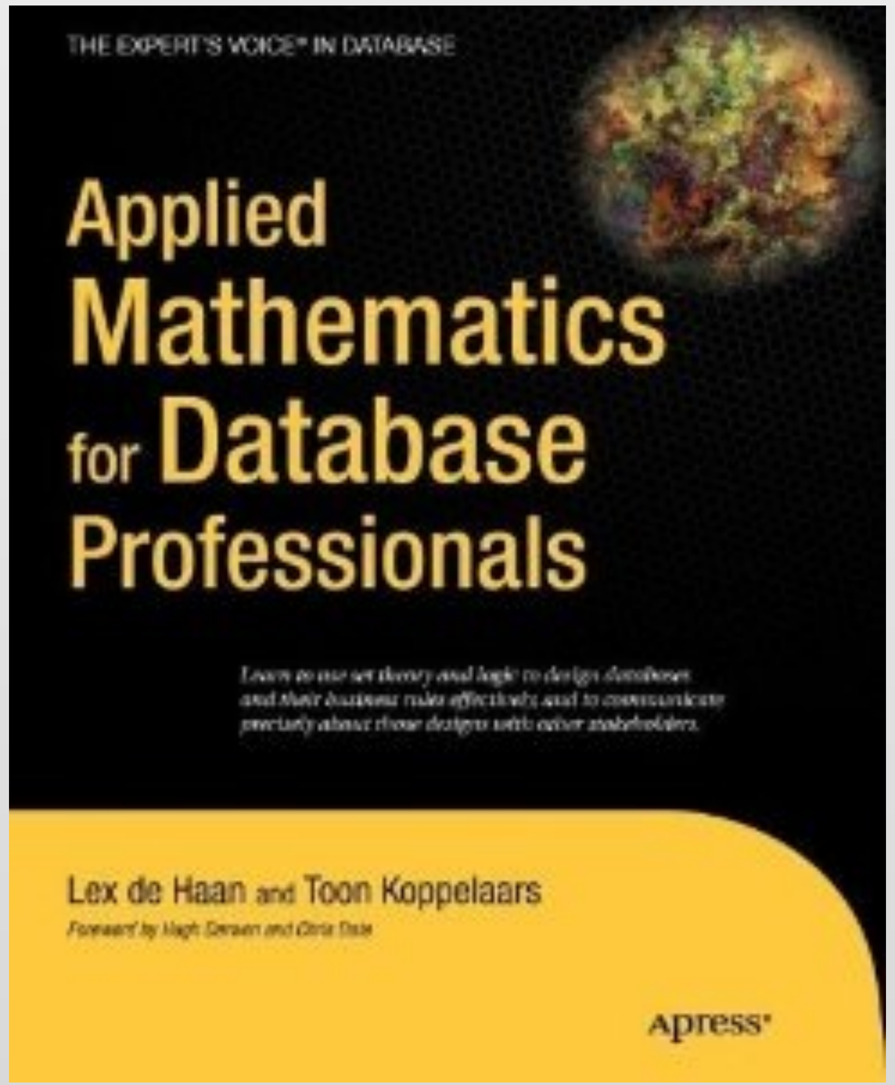
THE EXPERT'S VOICE™ IN DATABASES

Beginning SQL Queries

From Novice to Professional

Clare Churcher

Apress®



THE EXPERT'S VOICE™ IN DATABASE

Applied Mathematics for Database Professionals

Lex de Haan and Toon Koppelaars

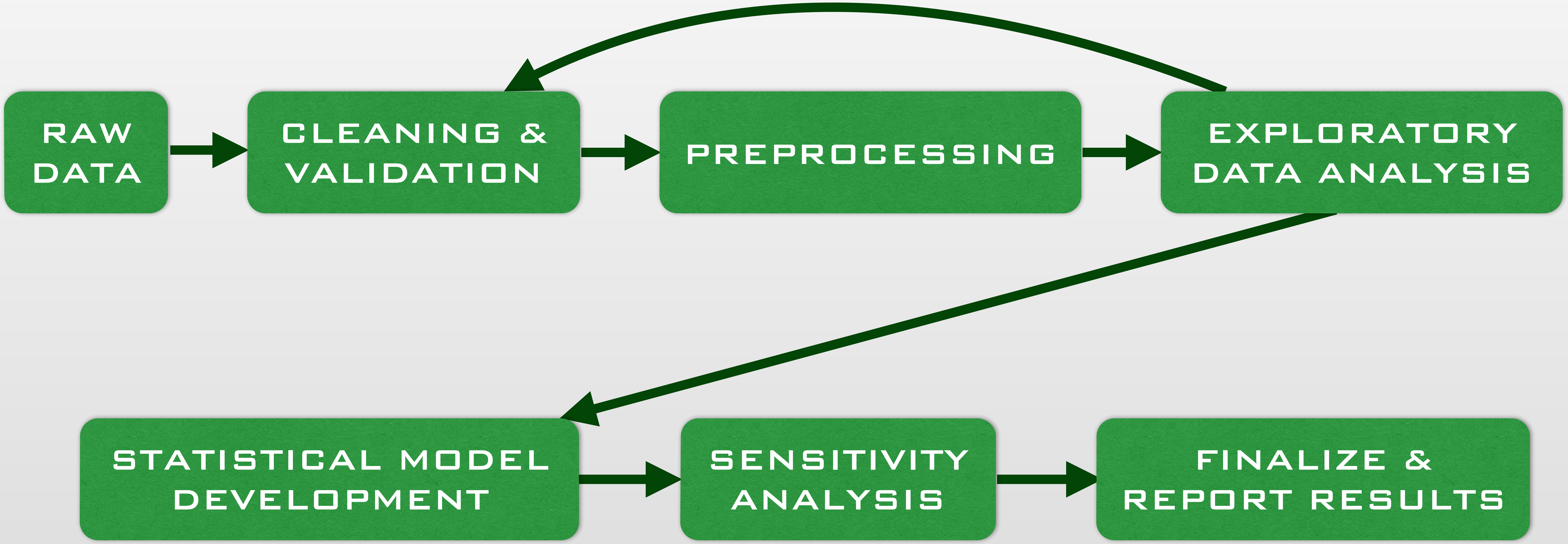
Apress®

SQL in other languages

- R with libraries
  - RPostgreSQL, dplyr
- Python with modules
  - psycopg2, SQLAlchemy
- Julia with packages (in dev)
  - PostgreSQL, DBI

EVIDENCE-BASED ANALYSIS FOR DATA SCIENCE

DWDC



RAW DATA

CLEANING & VALIDATION

PREPROCESSING

EXPLORATORY DATA ANALYSIS

STATISTICAL MODEL DEVELOPMENT

SENSITIVITY ANALYSIS

FINALIZE & REPORT RESULTS

DIAGRAM RECREATED WITH PERMISSION BASED ON SLIDE BY DR. ROGER PENG, JOHNS HOPKINS UNIVERSITY
(http://www.meetup.com/Data-Science-MD/photos/22063222/#366487342)

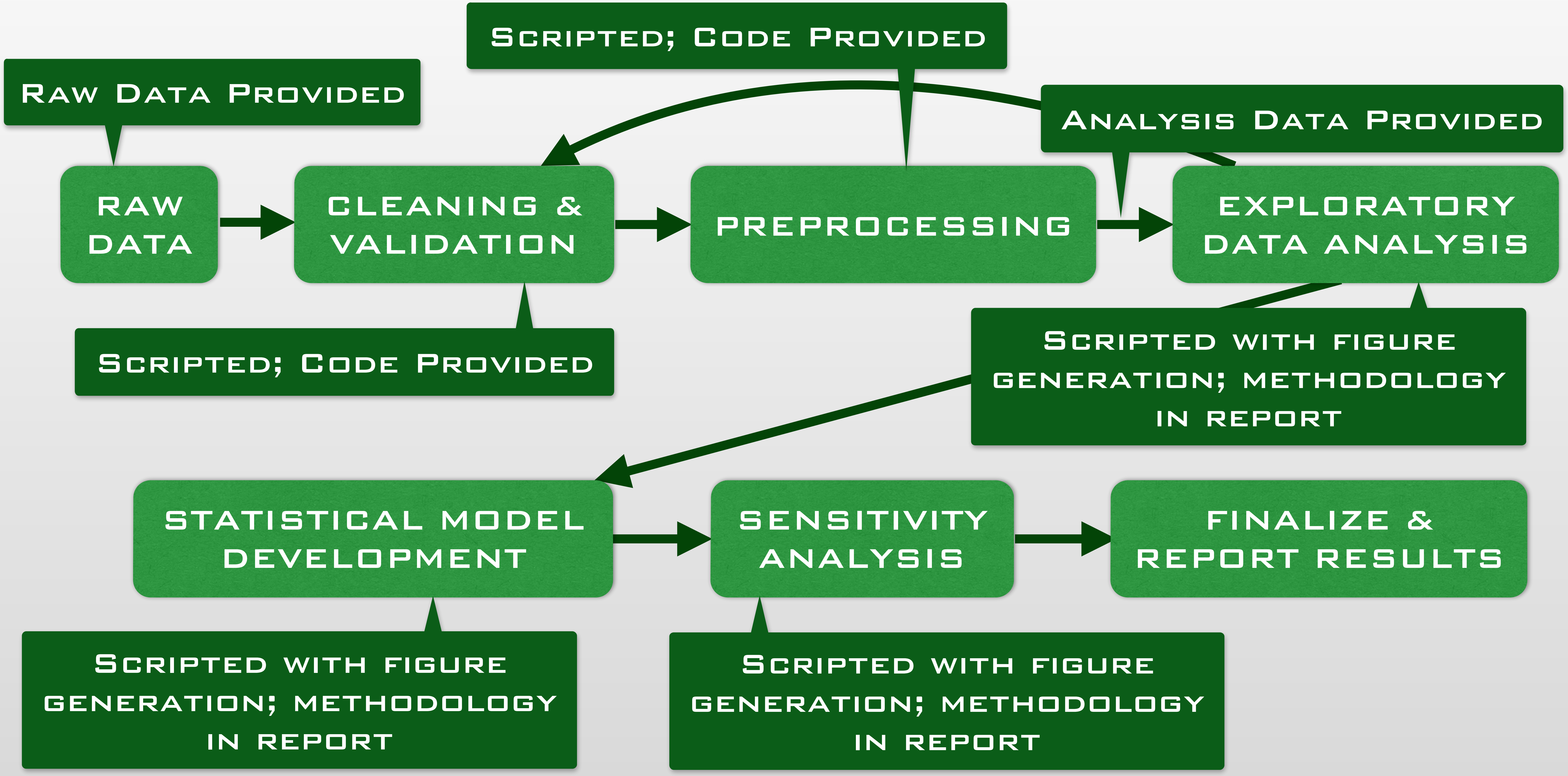
Why do reproducible analyses?

- 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

TEXT

What makes this reproducible?

DWDC



Raw Data Provided

Scripted; Code 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

Now, let's look at some code!

Ryan B. Harvey

DWDC

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.

My remarks, presentation and prepared materials are my own, and do not represent the views of my employers.