UC DAVIS LIBRARY

Introduction to Relational Databases

Vessela Ensberg, Associate Director, Data Management, UC Davis Library

John Daniels, Statistical Programming Consultant and Systems Administrator for the Social Science Data Service

Learning outcomes

- You will know whether you should switch to a relational database
- You will be able to apply best practices to set up a database in SQLite
- You will be able to do SELECT FROM WHERE, LIKE and JOIN queries

UC DAVIS LIBRARY

When and why use databases?

- Your spreadsheets have more than 1000 entries
- Your excel files have multiple tabs that relate to each other
- You are concerned about overwriting your data
- You need multiple users to access the data
- Time to compute
- Universal principles and language

UC DAVIS LIBRARY

2-D vs 3-D Data

- Introducing the practice dataset
  - Cohort→Individual
  - Cross date and parentage
  - Length, weight and brilliance through growth and development
- Purpose of data
  - Track multiple measurements
  - Identify most successful crosses
- How many tables?

UC DAVIS LIBRARY

Flat file

Ind_id	Cohort_id	Parent_1	Parent_2	Location	Cross date	Length_mm	Weight_g	Brilliance
NULL	e342fg	ew3483	gi02r9	Pool_1	20180304	34	23	Bright
W1343fg	e342fg	ew3483	gi02r9	Tank_19	20180304	205	103	Bright

Constant

UC DAVIS LIBRARY

More than one table

Ind_id	Cohort_id	Length_mm	Weight_g	Brilliance
NULL	e342fg	34	23	Bright
W1343fg	e342fg	205	103	Bright

Ind_id	Cohort_id	Location
W1343fg	e342fg	Tank_19

Cohort_id	Parent_1	Parent_2	Location	Cross date
e342fg	ew3483	gi02r9	Pool_1	20180304

UC DAVIS LIBRARY

Messy data

Ind_id	Cohort_id	Parents	Location	Cross date	Length	Weight	Brilliance
	e342fg	ew3483, gi02r9	Pool_1	20180304	34 mm	23 g	Brigth

UC DAVIS LIBRARY

Keeping data tidy

- One data type per cell
- One data point per cell
- Avoid highlighting
- Use data validation—pre-defined consistent categories
- Declare missing values: NULL

UC DAVIS LIBRARY

Creating a database schema

- Normalization (avoid duplication)
- Every record has a unique identifier: Primary key
- Restrict data types:
  - Numeric, integer, text, blob
  - Check
  - Foreign key—but avoid circular reference

Ind_id	Cohort_id	Location
W1343fg	e342fg	Tank_19

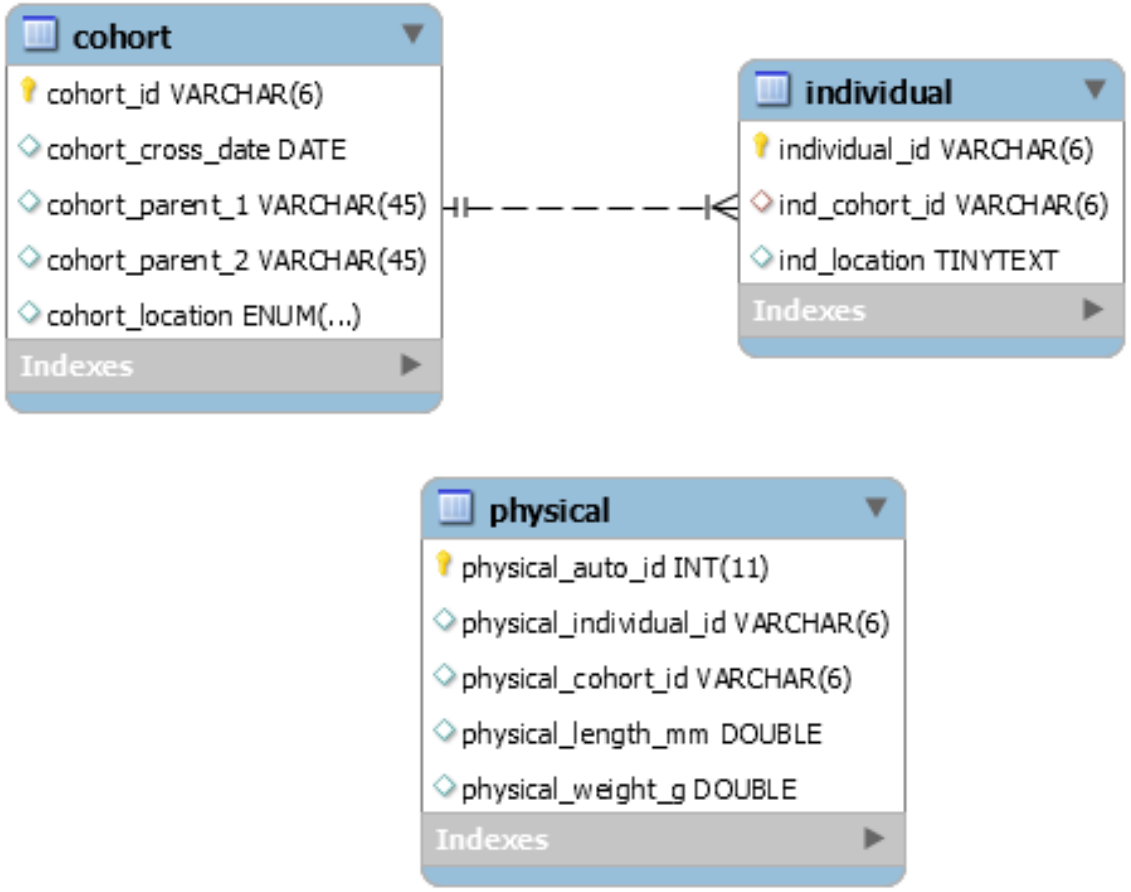
Cohort_id	Parent_1	Parent_2	Location	Cross date
e342fg	ew3483	gi02r9	Pool_1	20180304

Ind_id	Cohort_id	Length_mm	Weight_g	Brilliance
NULL	e342fg	34	23	Bright
W1343fg	e342fg	205	103	Bright

UC DAVIS LIBRARY

Data schema vs dictionary

- Table
- Field name
- Data type
- 1-2 sentences explain the purpose of the field



UC DAVIS LIBRARY

Structured Query Language

- Origins in the 70's and standard for most RDBMS today
- Language is standardized but still considerable drift between systems
- Open-source flavors
  - SQLite & DB Browser
  - MySQL (or MariaDB)
  - PostGreSQL

UC DAVIS LIBRARY

SQL Syntax Elements

- case insensitive
- insignificant whitespace
- “;” semi-colon statement terminator
- non-procedural (i.e. no loops)
- single-quotes to denote string

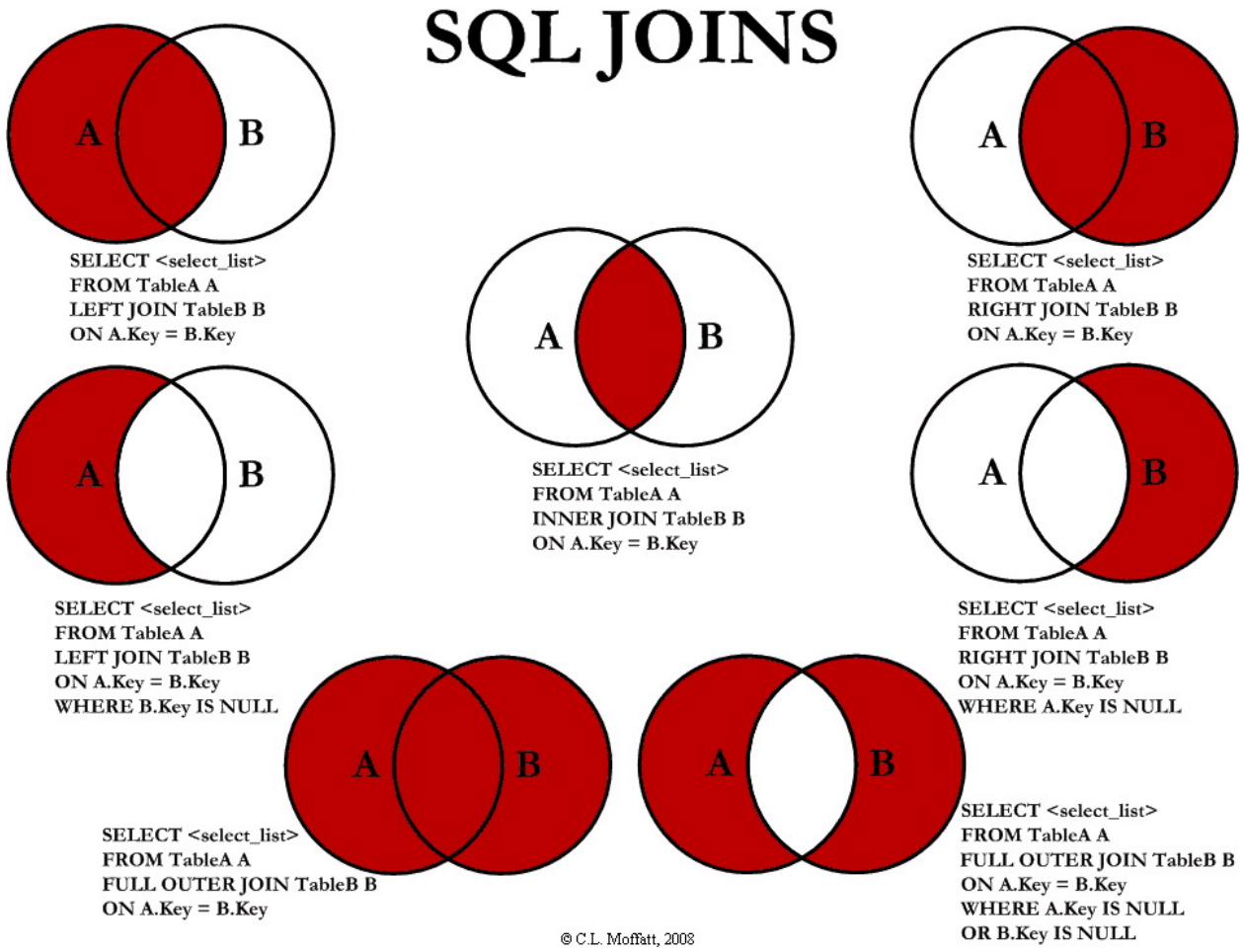
UC DAVIS LIBRARY

Operators

- = equals
- <> (or != sometimes)
- > greater than
- < less than
- >= greater than or equal to
- <= less than or equal to
- LIKE pattern matching
- BETWEEN within a range
- IN within a list
- IS NULL
- IS NOT NULL

UC DAVIS LIBRARY

SQL Joins



UC DAVIS LIBRARY