Introduction to Data Science

Relational Databases

Gordon Anderson

Relational Database Features

- Relational databases can process ad-hoc queries.
- SQL reduces development time and improves interoperability.
- Relational databases are mature, battle-tested technology.
- Relational databases conform to widely accepted standards.
- Provide a high level of data integrity and security.
- Protect data against system failures and concurrent operations.
- All web and mobile software can connect to relational databases.

Relational Data

- Most data you encounter is ***relational***.
- The term *relational* refers to the structure of the data.

	CustomerID	LName	FName	Initial	Areacode	Phone
1	101	Brown	James	S	615	243-1228
2	102	Smith	Carl	M	617	891-2289
3	103	Ortez	Maria	B	413	765-2370
4	104	Weiner	Myron		617	222-1682
5	105	Anderson	Hans	C	413	756-1878
6	106	O'Brien	Fiona	L	413	442-9381
7	107	Washington	George		617	290-2566
8	108	Funicelo	Maria	A	413	382-2385
9	109	Byrnes	Janet	Z	617	243-3779
10	1010	Helms	Leona	H	413	891-1238
11	1011	Anderson	Neo		413	756-1878

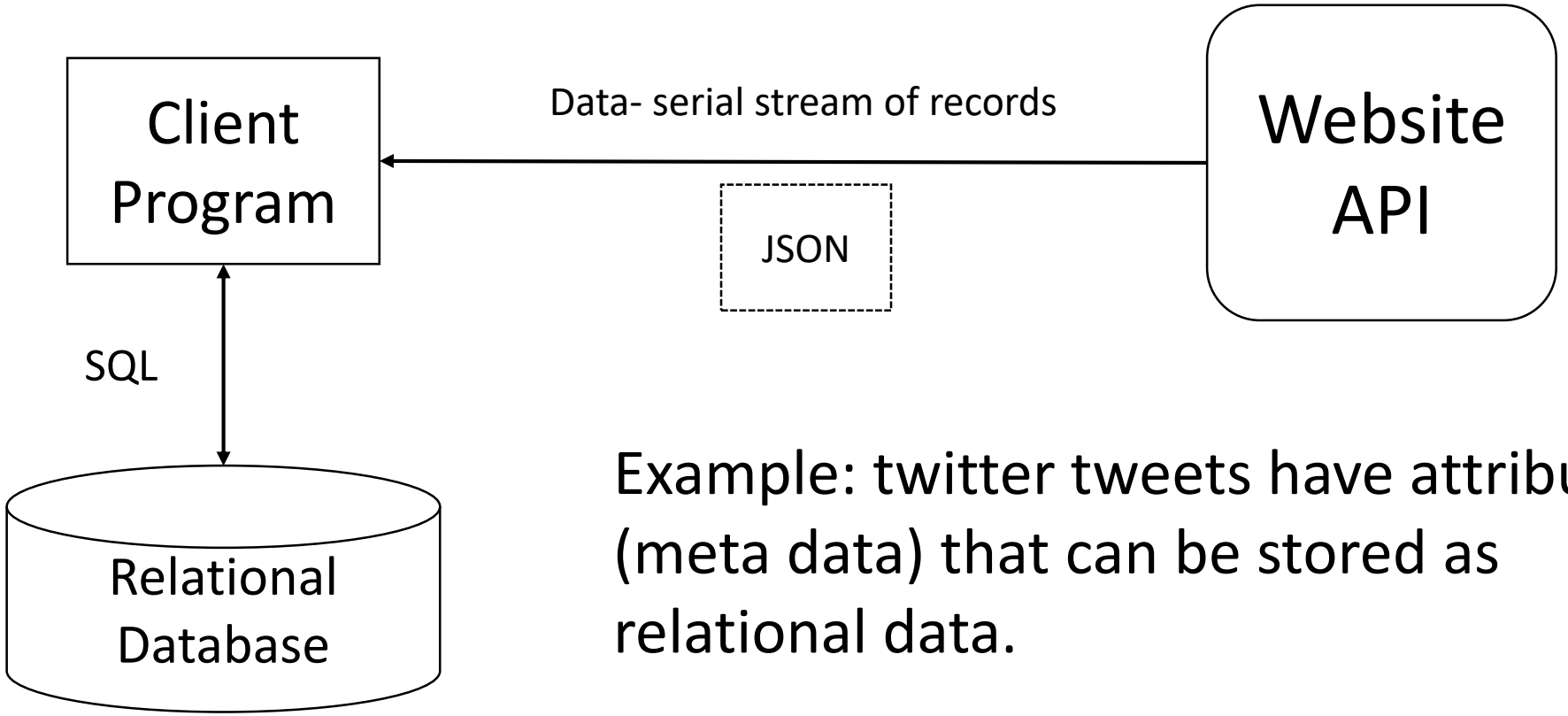
Each row is a “tuple” of related observations about an “instance” of the relation.

The “Relation” is “Customer”, so a row has data about a specific customer.

Columns are observed values about the customer.

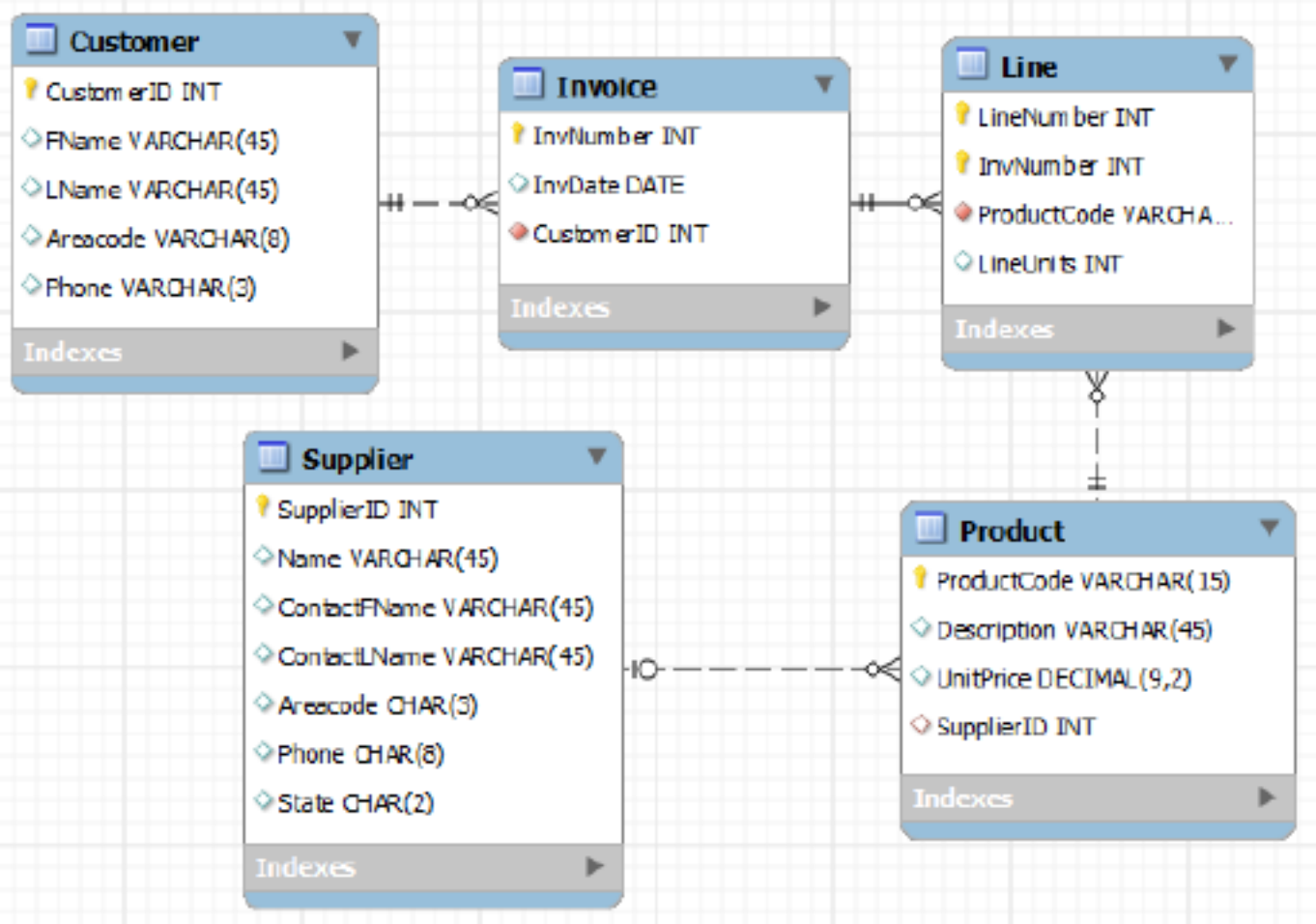
Relational Data

A typical example of data flow in “web-scraping”.



Example: twitter tweets have attributes (meta data) that can be stored as relational data.

Relational Database Structure



Each relation is a ***table***.

Tables are related to other tables.

The data is ***highly structured***. That makes it easier to query and to maintain its ***integrity***.

Relational Database Features

- Organized around records in tables, relationships between tables.
- Relationships are enforced with constraints.
- Indexing data brings hi-speed access.
- Highly structured data- improves data integrity.
- SQL is used to get data in and out of the db.
- Security provided in structure- views tied to permissions.
- Efficient storage of data.

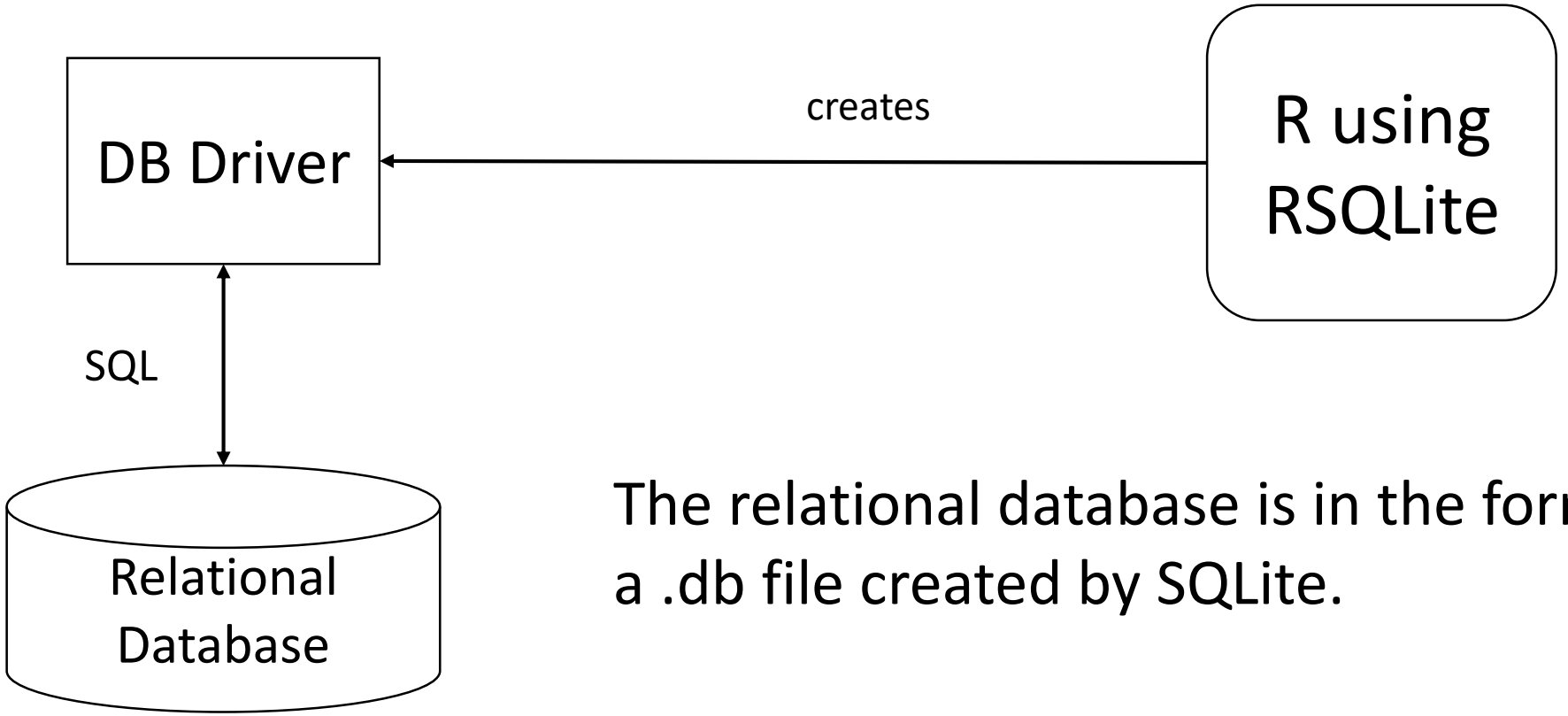
Relational Database Management Systems

- RDMS is a software application that provides the functions described on the previous slide and more.
- There are many RDMS systems. We will use SQLite in this module.
- It is easy to install and work with- “light weight”.
- Although not “industrial strength”, it provides much functionality.

SQL and R

- RSQLite library- allows us to connect to a relational database created by SQLite.
- The SQLite database is a file with a .db extension, for example: ACME.db
- From R we can execute SQL queries that retrieve data from the SQLite database in the form of a dataframe.
- We can then analyze and visualize the data.
- We will mostly work with existing databases and execute retrieval queries.
- It is also possible to create databases, insert and update data.

Connecting to a database



The relational database is in the form of a .db file created by SQLite.