University of Essex

Databases

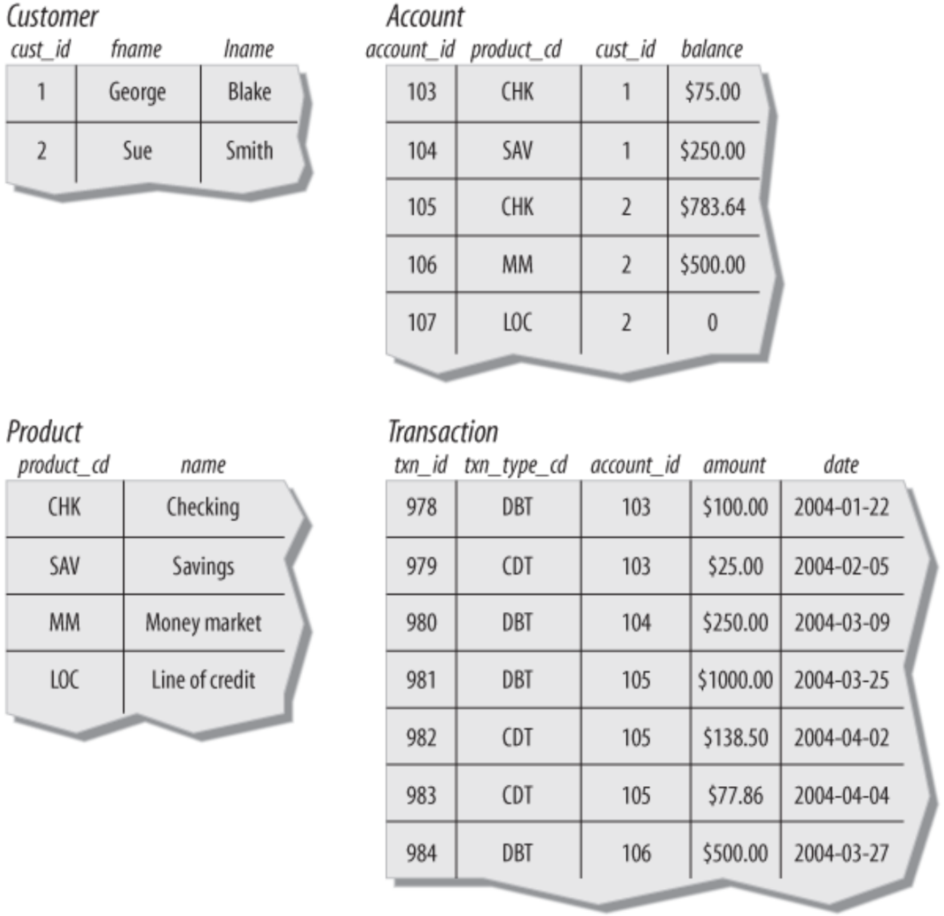
Akitaka Matsuo

Department of Government

Databases

University of Essex

- **Database management systems**: Software for storing and retrieving data
- **Relational database**: consists of multiple tables linked each other through common keys



Customer

cust_id	fname	lname
1	George	Blake
2	Sue	Smith

Account

account_id	product_cd	cust_id	balance
103	CHK	1	$75.00
104	SAV	1	$250.00
105	CHK	2	$783.64
106	MM	2	$500.00
107	LOC	2	0

Product

product_cd	name
CHK	Checking
SAV	Savings
MM	Money market
LOC	Line of credit

Transaction

txn_id	txn_type_cd	account_id	amount	date
978	DBT	103	$100.00	2004-01-22
979	CDT	103	$25.00	2004-02-05
980	DBT	104	$250.00	2004-03-09
981	DBT	105	$1000.00	2004-03-25
982	CDT	105	$138.50	2004-04-02
983	CDT	105	$77.86	2004-04-04
984	DBT	106	$500.00	2004-03-27

SQL

- **SQL**: A query language for relational databases
- SQL is a *declarative language* (not an imperative language). That only defines the information you seek for, when retrieving data.
- Many different systems that implement SQL database systems.
- Performance is not something what we, social scientists, usually worry.

Why databases?

Traditional

- Concurrency (simultaneous updates by many clients)
- Frequent updates (necessary to maintain integrity)

New

- Storing large data
  - But you need a small portion of it each time
  - Sharing data with many
- Backend for web services
  - Rapid query, dynamic data

SQL Database Management Systems (DBMS)

- There are numerous implementations
- Basic syntax are similar, but for complicated queries, the implementations are system dependent
- Major SQL DBMSs
  - Open source: MySQL, PostgreSQL, SQLite
  - Proprietary: Oracle
- Cloud service providers have fully-managed SQL systems (usually quite pricy, and overkill for most social scientists)

How to access to databases

- Console
- Programming language
  - – Python, R etc
  - – Run the query to get subset of the data, but analysis is done in the language
- GUI Interface
  - – e.g. MySQL Workbench, pgAdmin (PostgreSQL), DB Browser (SQLite)

Database Design: Avoid redundancy

- **Database normalization** = removing any redundancies in tables
- How many tables could be made from table below?

fips	county	state	lat	long	date	cases	state_code	deaths
48001	Anderson	Texas	31.815	-95.654	2020-06-16	102	TX	2
48001	Anderson	Texas	31.815	-95.654	2020-06-17	990	TX	2
48043	Brewster	Texas	29.810	-103.252	2020-07-16	160	TX	1
48043	Brewster	Texas	29.810	-103.252	2020-07-17	160	TX	1
48043	Brewster	Texas	29.810	-103.252	2020-07-18	161	TX	1

Normalization

- Pros:
  - Saving disk space
  - Data integrity
- Cons:
  - A lot of table linking every time

fips	county	state	lat	long	date	cases	state_code	deaths
48001	Anderson	Texas	31.815	-95.654	2020-06-16	102	TX	2
48001	Anderson	Texas	31.815	-95.654	2020-06-17	990	AZ	2
48043	Brewster	Texas	29.810	-103.252	2020-07-16	160	TX	1
48043	Brewster	Ohio	29.810	-103.252	2020-07-17	160	OH	1
48043	Brewster	Texas	29.810	-103.252	2020-07-18	161	TX	1

Content of SQL Language

- Data Definition Language (DDL)
  - Create/alter/delete tables and their attributes
  - Define relations between tables
- Data Manipulation Language (DML)
  - Insert/delete/modify records in tables
  - Query one or more tables

We look at the last part.