ECPR Methods Summer School: Big Data Analysis in the Social Sciences

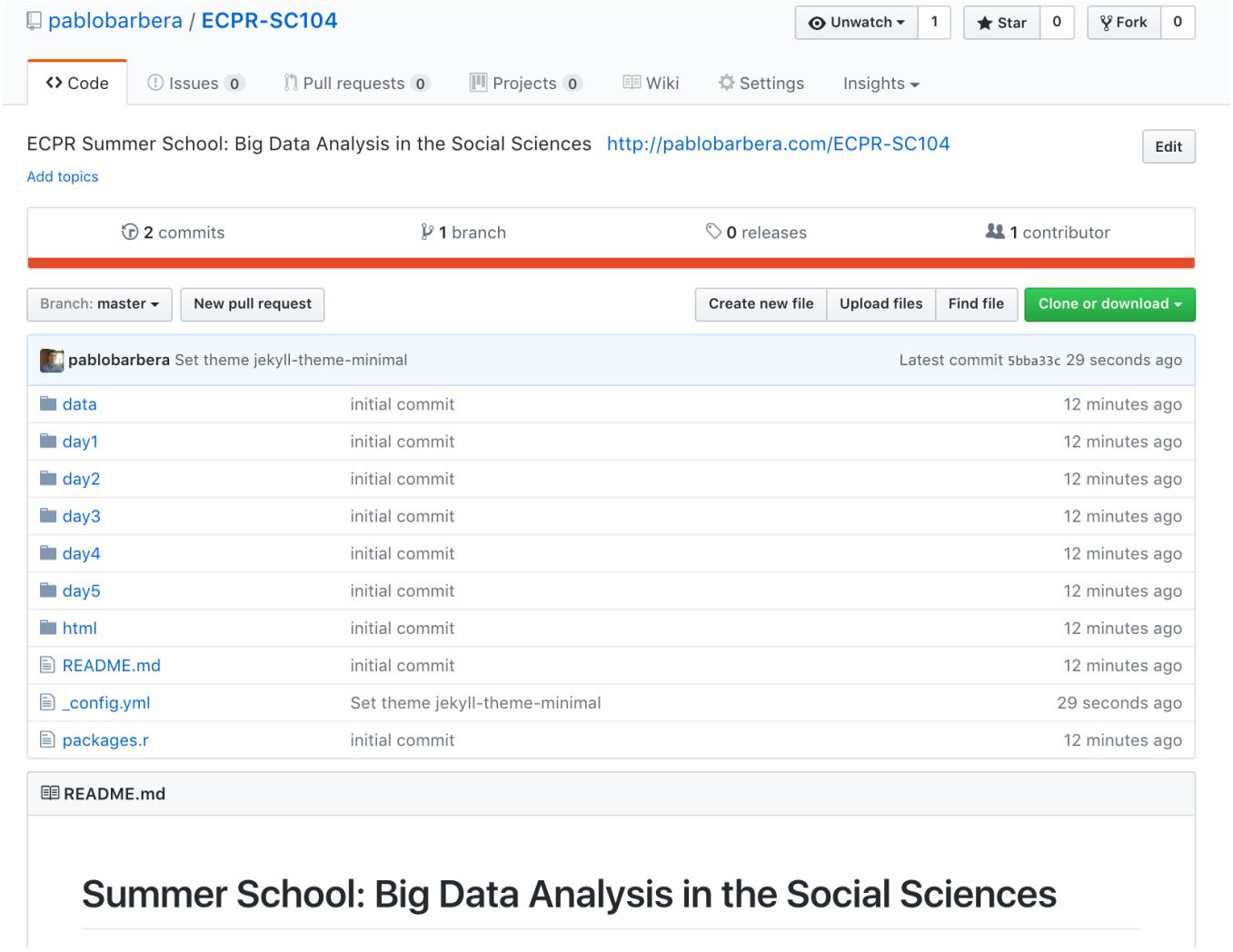
Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:
github.com/pablobarbera/ECPR-SC104

Course website



github.com/pablobarbera/ECPR-SC104

Introduction to cloud computing


There is no cloud
it's just someone else's computer

Cloud computing

Use of remote servers hosted online to collect, store, and manipulate data rather than a personal computer.

Why moving to the cloud?

- Scalability: ability to increase memory or computing power to that necessary for our application
- Cost: pay for what you use, no need to buy expensive hardware
- Speed: easy and fast to launch a service on demand
- Reliability: cloud services include backup, redundancy, disaster recovery...
- Accessability: multiple users can access and analyze data simultaneously, from anywhere with an internet connection

Cloud computing

What you need to know:

- UNIX commands to log in and interact with the server
- SQL to query large-scale databases

Companies offering cloud computing services:

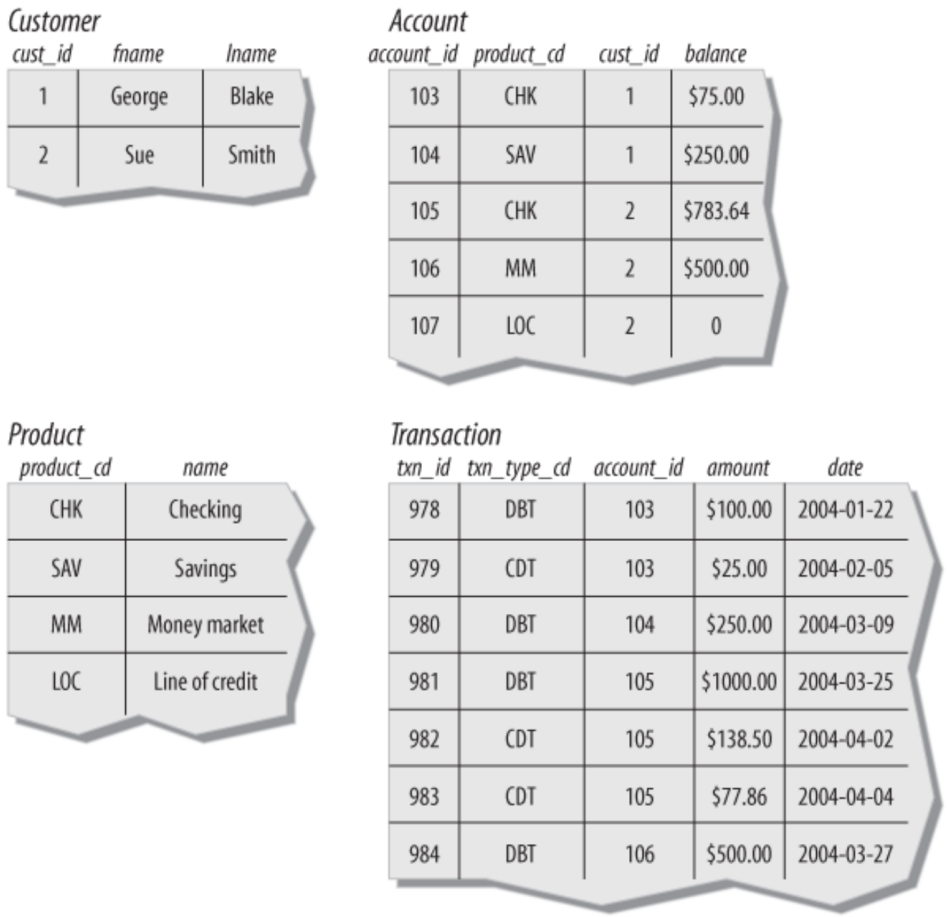
- Amazon Web Services (AWS)
- Google Cloud Platform
- Microsoft Azure
- Digital Ocean
- ...many others

Basic UNIX commands

`cd` *dirname*	Change directory
`mkdir` *dirname*	Create new directory
`cp` *oldfile newfile*	Copy a file
`mv` *oldfile newfile*	Move a file
`ls -lh`	List your files (with sizes)
`cat` *file*	Print file in console
`head` *file*	Print first lines of file
`tail` *file*	Print last lines of file
`wc -l` *file*	Count lines in file
`grep` *string file*	Regex on file text
`gzip` *file*	Compress file
`wget` *URL*	Download file from URL
`ps -u` *user*	See running processes by user
`kill` *process*	End running process

Databases

- Database systems: computerized mechanisms to store and retrieve data.
- Relational databases: data is represented as tables linked based on common keys (to avoid redundancy).



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 (pronounced S-Q-L or SEQUEL) is a language designed to query relational databases
- The result of an SQL query is always a table
- It's a nonprocedural language: define inputs and outputs; how the statement is executed is left to the *optimizer*
- How long SQL queries depends on optimization that is opaque to user (which is great!)
- SQL is a language that works with many commercial products:
  - Oracle Database, SQL Server (MS), MySQL, PostgreSQL, SQLite (all three open-source), Google BigQuery, Amazon Redshift...
  - Performance will vary, but generally faster than standard data frame manipulation in R (and much more scalable)