The Nyiragongo Crater in the Democratic Republic of Congo is the world's largest lava lake, one of the wonders of the African continent. The crater bubbles 1,300 feet (Olivier Grunewald) #

COMP 364 - Lecture #22
March 14, 2012
Mathieu Perreault

Data storage

Record-oriented data storage

- CSV (Comma-separated-value), so called “flat files”
- Database

CSV files

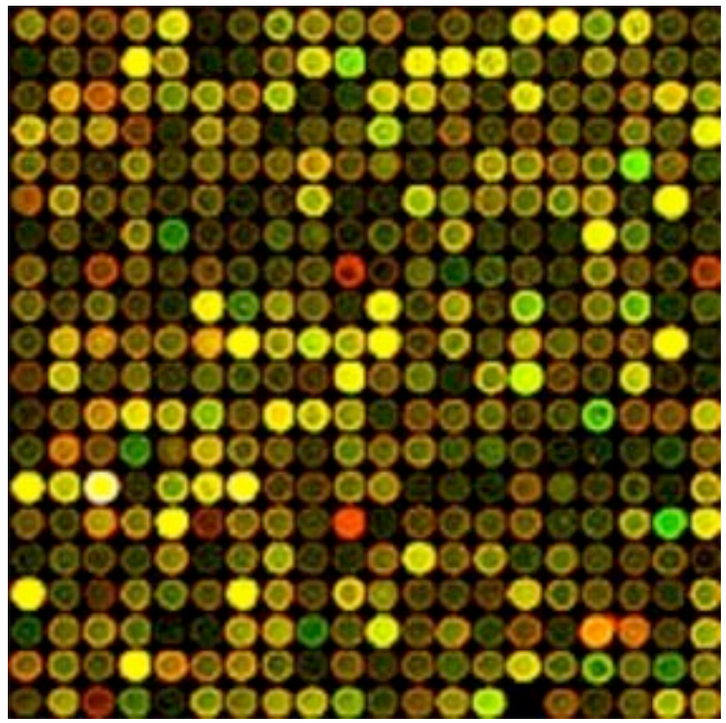
- Easy to read and write
- Easy to share
- What do you store?
- How do you write it to a file?

Goals:

1. Easy to read in (for analysis)
2. Easy to write out
3. *Human-readable*

Microarray example

1. Running a set of gene expression experiments
2. Multiple plates over several days
3. Experiments are run in triplicate



How do you store your data?

The format depends on what analysis you do

- Compare all gene profiles
    - Compare average profiles
    - Compare for different runs
        - Compare across runs
- Compare gene profiles by gene function
- Detect problems with plates and positions

The more complicated your analysis, the more complicated the storage requirements.

What is a database?

- Stores data in a structured format
- Data can have complicated structure
- Ability to query for and retrieve data using a (more) friendly language
- Ability to update data without completely overhauling database

Introducing SQLite

- SQLite is the most widely used database engine in the world
- The database is contained in a single file
- Used everywhere, even in your phone!
- Very light and simple to use (demo to come)
- Supports the same syntax as the big database engines (MySQL, PostgreSQL)

Difference with CSV files

- If SQLite is in a single file, how can it be *that* different from CSV.
- SQLite is a binary file, not character-based.
- SQLite supports datatypes! INTEGER, TEXT, BLOB, REAL and NULL.
- SQLite doesn’t need to be read in memory before accessing data, like CSV does.
- Many types of records can be in your database.

Example database

Microarray reads
gene_id
amplitude
plate
well_num
pos_x
pos_y

Genes
gene_id
gene_name
ncbi_url

Plate
plate_id
date_run
time_run
humidity
location

GO Terms
gene_id
go_term