

Data Formats Continued

ME314: Introduction to Data Science and Machine Learning

Ken Benoit

12 July 2023

Plan today

- Alternative data formats
- Brief primer on relational databases
- Move on to regression

Alternative data formats

Database systems

Relational databases

- Mainly implementations and extensions of the SQL Standard ([ISO/IEC 9075:2016](#))
- Transactions are always **ACID** (atomic, consistent, isolated, durable)
- Data needs to be defined

Non-relational databases

- Key-value storage types (e.g. Amazon DynamoDB) or document storage types (e.g. CouchDB, MongoDB)
- Sometime labelled as providing **ACID** transactions but often only *eventually consistent*
- FYI for clicking on the SQL standard link: The standard is open, i.e. anyone can get it, but subject to a fee

—

- a lightweight data-interchange format that is (supposedly!) easy for humans to read and write, and easy for machines to generate or parse
- follows conventions from Javascript, but is language-independent
- Example: [Twitter data](#)
- built on two structures:
 - A collection of name/value pairs
 - An ordered list of values

object

- unordered set of name/value pairs. An object begins with { and ends with }
- each name is followed by : and the name/value pairs are separated by ,

array

- an ordered collection of values
- begins with [and ends with]
- array values are separated by ,

- can be a "string", a number, or true, false, or null, or an object or array
- can be nested

strings in JSON

- a sequence of zero or more Unicode characters, wrapped in double quotes
- uses backslash escapes, e.g.
- `"\u2708\u201c"` represents a plane
- `"this is \"quoted\""` represents "quoted"

```
print("It's a bird, it's a \u2708\u201c!!!")
```


Relational data structures

- invented by E. F. Codd at IBM in 1970
- A relational database is a collection of data organized as a set of formally defined tables
- These tables can be accessed or reassembled in many different ways without having to reorganize the underlying tables that organize the data
- RDBMS: a relational database management system. Examples include: MySQL, SQLite, PostgreSQL, Oracle. MS Access is a lite version of this too.
- The standard user and application programmer interface to a relational database is structured query language (SQL)

Example

**from Database of Parties, Elections, and Governments (DPEG)
relational database**

```
SELECT c.countryName, c.countryAbbrev, p.* FROM party AS p  
LEFT JOIN country AS c  
ON p.countryID = c.countryID
```