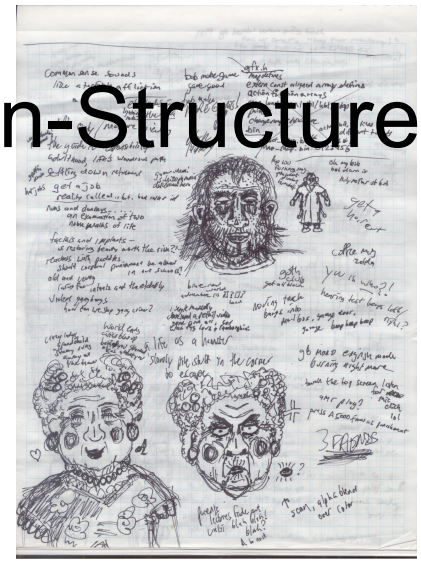
GENERAL ASSEMBLY

DATA SCIENCE
CLASS 11: TYPES AND BASES

I. ORGANIZING DATA
II. FORMATS
III. APIS [LAB]
IV. DATABASES
V. SQL [LAB] [LAB]

I. ORGANIZING DATA

Un-Structured Data



age	sex	brushing	cavities	cavity	timely	elective
9	female	monthly	1	TRUE	0.21	0.07
8	female	daily	0	FALSE	0.2	0.06
34	female	daily	1	TRUE	0.46	0.11
13	male	monthly	2	FALSE	0.13	0.06
20	female	daily	0	FALSE	0.32	0.03
24	female	daily	0	FALSE	0.36	0.06
17	male	daily	0	FALSE	0.17	0.05
42	female	daily	4	TRUE	0.54	0.36
19	male	weekly	7	TRUE	0.19	0.36

Structured Data

- Tabular data
- Nested data
- Graph data

Tabular data

- Rows and columns
- Implicit or explicit “schema”, tidy or not
- Very common and well-handled by software

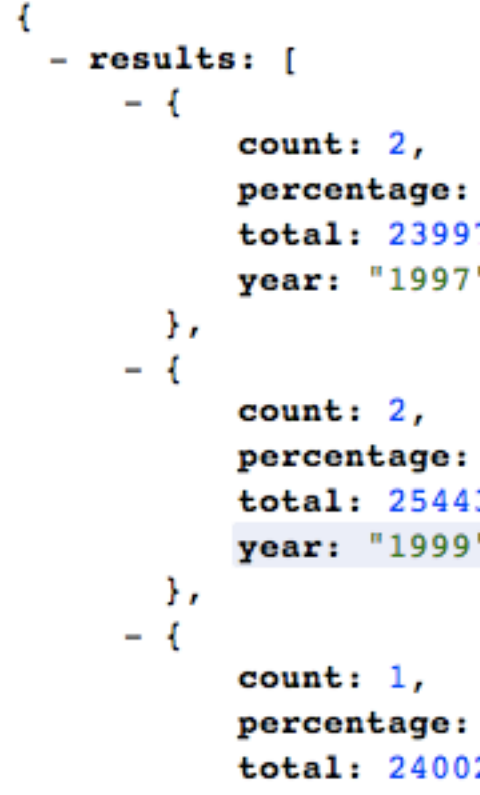
“Tidy Data” is a conceptual framework for organizing data

- Each variable is a *column*
- Each observation is a *row*
- Each type of observational unit is a *table*

Usually tabular; related to relational “normal forms”

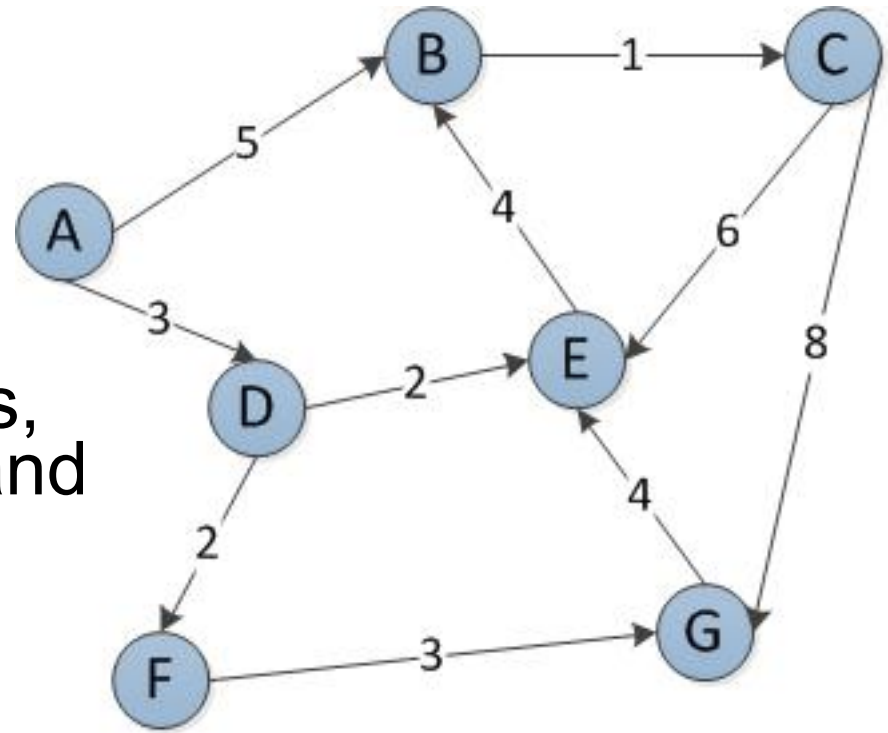
Nested data

- “Hierarchical”, “Tree”, “Object”
- Often schema-less
- Often multiple data types (lists, etc.)
- Often natural and fairly common



Graph data

- Nodes and Edges
- Schema?
- Fits particular applications, such as social networks and semantic web



II. FORMATS

- Tabular data
- Nested data
- Graph data

Tabular data

- Fixed-width text files: “ `field007 blue`” etc.
- Delimited text files (Comma Separated Values and the much-ignored RFC 4180, plus tabs, pipes, and more)
- Spreadsheets such as Excel
- Database formats

Tabular data

- Fixed-width text files: “ `field007 blue` ” etc.
- Delimited text files (Comma Separated Values and the much-ignored RFC 4180, plus tabs, pipes, and more)
- Spreadsheets such as Excel
- Database formats

ONE MORE THING!

We might include matrix formats, images, and perhaps even more here. (They could also be considered separately.)

Nested data

- Extensible Markup Notation (XML)
- JavaScript Object Notation (JSON)
- Other serialization systems such as Apache Thrift and Avro, and Google Protocol Buffers – even just pickled Python objects

Graph data

- Various formats, built on lower-level formats.

III. APIS [LAB]

IV. DATABASES

- Tabular data
- Nested data
- Graph data

Tabular data

- Relational DataBase Management Systems (RDBMS) – broadly implementing SQL
- Serverless/embedded: SQLite, H2, etc.
- Client-Server: MySQL/MariaDB, PostgreSQL, Oracle, MSSQL, etc.

Nested data

- Document stores: MongoDB, CouchDB, etc.
- Serialization schemes, often on HDFS
- Key-value stores: Redis, Voldemort, etc.
- Blurred Lines: Object-Relational Mapping, Cassandra, column orientation, etc.

Graph data

- Client-server (general): Neo4j, perhaps more
- Client-server (triple store): AllegroGraph, etc.
- Query languages: SPARQL, RDF++, Prolog…

V. SQL [LAB] [LAB]

INTRO TO DATA SCIENCE