Foundation of Data Science
Lecture 2, Module 2
Fall 2022

Rumi Chunara

<u>Fine Print</u>: these slides are, and always will be a work in progress. The material presented herein is original, inspired, or borrowed from others' work. Where possible, attribution and acknowledgement will be made to content's original source. Do not distribute without the instructor's permission.

Rules of Thumb for *All* Data Types

1. Know where your data comes from (believability)

2. Know what your data looks like
   - Data distributions (exploratory data analysis)
   - Data types **focus today!**

3. Know the limits of your data
   - What questions cannot be answered?
   - Can I gather more data? https://auctus.vida-nyu.org/
   - Is the data a good sample? **focus today!**

Structured and Unstructured Data

- Is the data in a standard format or encoding? (structured)
  - **Tabular data**: CSV, TSV, Excel, SQL
  - **Nested data**: JSON or XML
- Is the data organized in "records"? (structured)
  - No? But can we define records by parsing the data?
- Is the data nested, i.e., records contained within records? (structured)
  - Yes? Can we reasonably un-nest the data?
    - This would make the data *flat* (more efficient and easier to iterate over, easier to understand).
- Can we join/merge the data with other data? (structured)
  - Yes: can we join/merge the data
- What are the fields in each record?
  - How are they encoded? (strings, numbers, binary, dates)
  - What are the **types** present in the data?

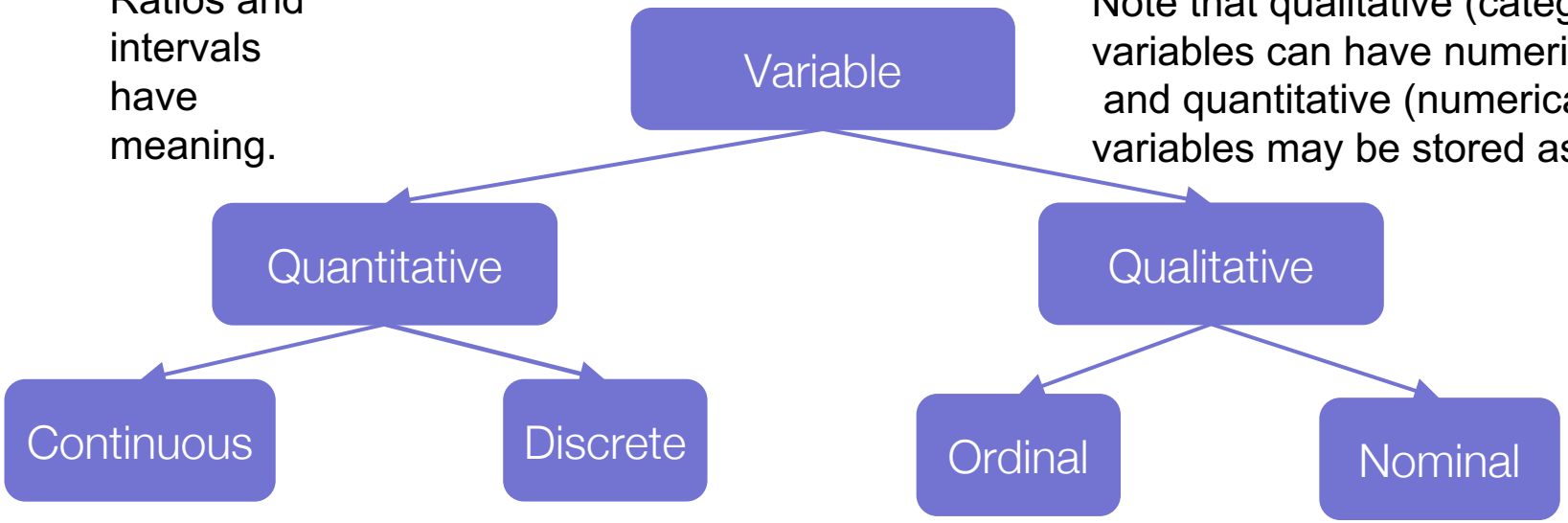
Structured and Unstructured Data

- If the answers to those questions are *"No"*, the data is unstructured or semi-structured.
  - Unstructured data: does not fit into a neat box.
    - Photos, videos, PDF files, webpages, emails, text in general.
  - Semi-structured data: a cross between structured and unstructured.
    - Emails or photos with tags or hashtags, where these last two bring some structure. Text documents with dates, location and keywords.

Types of Variables -- Structured Scope

Ratios and intervals have meaning.

Note that qualitative (categorical) variables can have numeric levels and quantitative (numerical) variables may be stored as strings.



Could be measured to arbitrary precision.

Examples:
- Price
- Temperature

Finite possible values

Examples:
- Number of siblings
- Yrs of education

Categories w/ levels but no consistent meaning to difference

Examples:
- Preferences
- Level of education

Categories w/ no specific ordering

Examples:
- Political Affiliation
- CallD number

What is the type of variable?

	Quantitative Continuous	Quantitative Discrete	Qualitative Ordinal	Qualitative Nominal
CO_2 level (PPM)				
Number of siblings				
GPA				
Income bracket (low, med, high)				
Gender				
Number of years of education				
Yelp Rating				

What is the type of variable?

	Quantitative Continuous	Quantitative Discrete	Qualitative Ordinal	Qualitative Nominal
CO_2 level (PPM)				
Number of siblings				
GPA				
Income bracket (low, med, high)				
Gender				
Number of years of education				
Yelp Rating				

Example: Flat File Data

n rows

2.3	-1.5	…	-1.3
1.1	0.1	…	-0.1
…	…	…	…

p columns

- Rows = objects/observations/instances
- Columns = measurements on objects (variables/features/attributes)
- Both *n* and *p* can be very large in data mining (often $p \ll n$)
- Matrix can be quite sparse
- ML rule of thumb: *n* = ~1000 (unless *p* very large)

Example: Transactional Data

Date stamped events (logs, phone calls):

128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.101, -, 3/22/00, 16:18:50, W3SVC, SRVR1, 128.200.39.181, 60, 425, 72, 304, 0, GET, /top.html, -,
128.195.36.101, -, 3/22/00, 16:18:58, W3SVC, SRVR1, 128.200.39.181, 8322, 527, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.101, -, 3/22/00, 16:18:59, W3SVC, SRVR1, 128.200.39.181, 0, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:54:37, W3SVC, SRVR1, 128.200.39.181, 140, 199, 875, 200, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 17766, 365, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:54:55, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:39, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:03, W3SVC, SRVR1, 128.200.39.181, 1081, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:56:04, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:56:33, W3SVC, SRVR1, 128.200.39.181, 0, 262, 72, 304, 0, GET, /top.html, -,
128.200.39.17, -, 3/22/00, 20:56:52, W3SVC, SRVR1, 128.200.39.181, 19598, 382, 414, 200, 0, POST, /spt/main.html, -,

Can be represented as a time series:

User 1	2	3	2	2	3	3	3	1	1	1	3	1	3	3	3	3
User 2	3	3	3	1	1	1										
User 3	7	7	7	7	7	7	7	7								
User 4	1	5	1	1	1	5	1	5	1	1	1	1	1	1		
User 5	5	1	1	5												
…		…														

Example: Relational Data

128.200.39.17, -, 3/22/00, 20:55:07, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 1061, 382, 414, 200, 0, POST, /spt/main.html, -,
128.200.39.17, -, 3/22/00, 20:55:36, W3SVC, SRVR1, 128.200.39.181, 0, 258, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
128.195.36.195, -, 3/22/00, 10:35:11, W3SVC, SRVR1, 128.200.39.181, 781, 363, 875, 200, 0, GET, /top.html, -,
128.195.36.195, -, 3/22/00, 10:35:16, W3SVC, SRVR1, 128.200.39.181, 5288, 524, 414, 200, 0, POST, /spt/main.html, -,
128.195.36.195, -, 3/22/00, 10:35:17, W3SVC, SRVR1, 128.200.39.181, 30, 280, 111, 404, 3, GET, /spt/images/bk1.jpg, -,
…,

128.195.36.195, Doe, John, 12 Main St, 973-462-3421, Madison, NJ, **07932**
114.12.12.25,Trank, Jill, 11 Elm St, 998-555-5675, Chester, NJ, 07911
…

07911, Chester, NJ, 07954, 34000, , 40.65, -74.12
07932, Madison, NJ, 56000, 40.642, -74.132
…

- Most large data sets are stored in relational data sets
- Data query via SQL

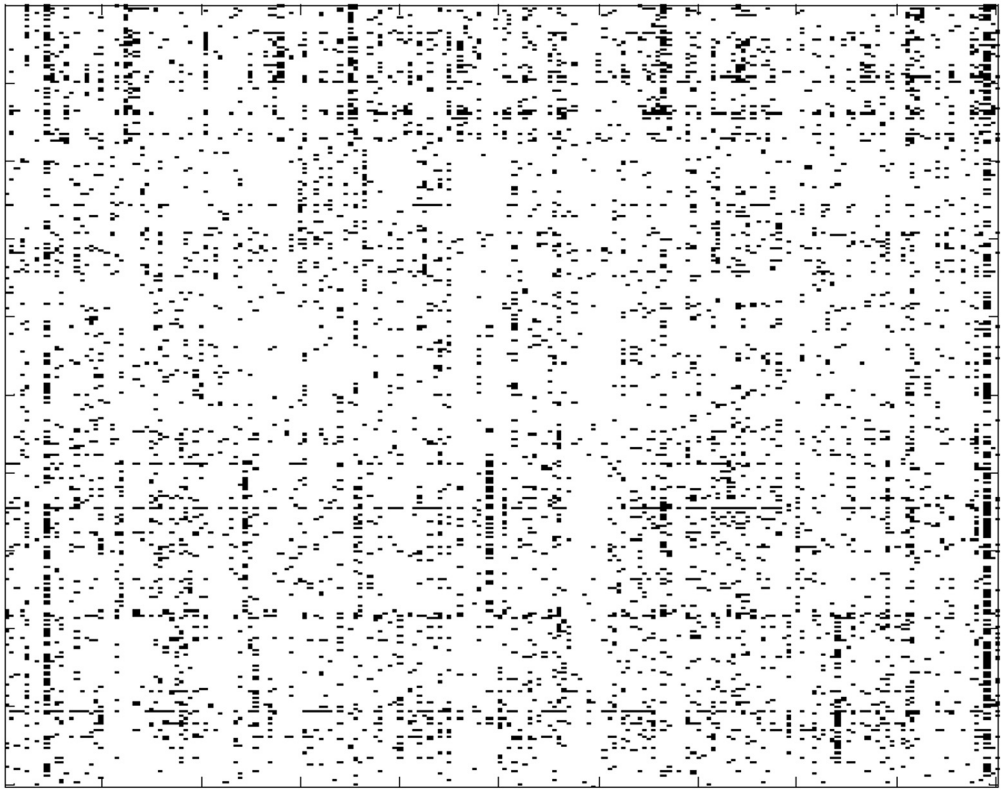
Example: Time Series Data

Is it structured, unstructured or semi-structured? Why?

Example: Text Data

Can be represented as a sparse matrix

Text Documents



Word ID

Example: Image Data

Is it structured, unstructured or semi-structured? Why?

Example: Spatio-Temporal Data

@b_mc817
Glendaaaaa

Omg earthquake!!!