CS 625: Data Visualization

Data and Data Cleaning

Bhanuka Mahanama

Department of Computer Science
Old Dominion University
bhanuka@cs.odu.edu

Based on lecture materials by Dr. Michele Weigle, and Dr. Tamara Munzner

Meaning of Data

- What does the sequence mean?
  - 70, 75, 80

Data Has No Meaning by Itself

Example: 70, 75, 80

- Without context it could be anything!
- With context, it could be
  - Highway speed limits between A and B
  - River temperature reading on X date
  - Financial transactions between parties P and Q
  - Network speeds on three links
- Data only becomes useful when we understand the meaning

Semantics and Data Types

- Semantics: real-world meaning
  - For visualization
- Data types: structural or mathematical interpretation
  - Item, link, attribute, position, grid
  - Different from data types in programming

Student	Homework (%)	Midterm (%)	Final Exam (%)
Alex	70	75	80
Maya	82	88	91
Jordan	65	72	78
Sam	90	85	87
Priya	76	81	84

Data Types: Items and Attributes

- Item: an individual entity that is discrete
  - Individual things we are discussing
  - People, stocks, transactions
- Attribute: a specific property that can be measured, observed or logged
  - Property of an item
  - Salary of a person
  - Price of a stock
  - Status of a transaction

Name	Salary	Education Level	Favorite Drink
Alex	48,000	Undergraduate	Coffee
Maya	72,000	Graduate	Tea
Jordan	115,000	PhD	Coffee
Sam	65,000	Undergraduate	Juice
Priya	98,000	Graduate	Tea
Leo	120,000	High School	Coffee
Nina	42,000	Undergraduate	Soda
Omar	88,000	Graduate	Water

Data Types: Links, Grids, and Positions

- Link: a relationship between items
  - Friendship on facebook, stock purchase by a person
- Grid: sampling strategy for continuous data
  - Weather temperature/precipitation map
- Position: spatial data, location in 2D or 3D
  - Pixels in a photo, latitude and longitude

Dataset Types

- Any collection of information for analysis
- Four basic of dataset types
  1. Tables: rows and columns
  2. Networks: connected relationships
  3. Fields: continuous spaces
  4. Geometry: shapes and sizes

Alternative ways to group items: sets, lists, and clusters

Dataset Types: Tables

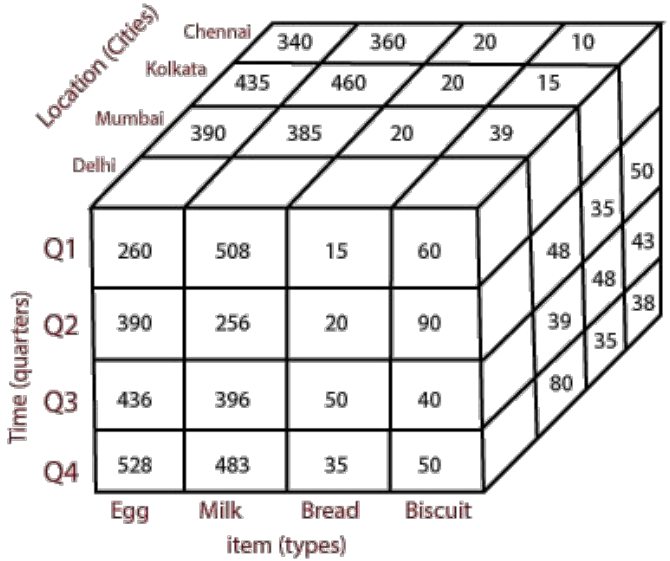
Simplest and most common form of a table is a flat table

- Row is an item of the dataset
- Colum is an attribute of the dataset
- Each cell contains a value specified by the combination of
  - Row and column
  - Item and attributes

Name	Salary	Education Level	Favorite Drink
Alex	48,000	Undergraduate	Coffee
Maya	72,000	Graduate	Tea
Jordan	115,000	PhD	Coffee
Sam	65,000	Undergraduate	Juice
Priya	98,000	Graduate	Tea
Leo	120,000	High School	Coffee
Nina	42,000	Undergraduate	Soda
Omar	88,000	Graduate	Water

Dataset Types: Multidimensional Tables

- Has complex structure for indexing a cell with multiple keys
- Each cell contains a value specified by the combination of *item, attributes, and additional keys*



https://www.javatpoint.com/data-warehouse-what-is-multi-dimensional-data-model

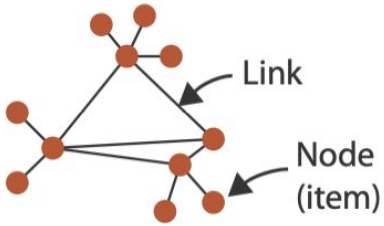
Dataset Types: Networks

Suited for specifying relationships between two or more items

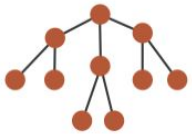
- Node (vertex): item in a network
- Link (edge): relation between two items
- Nodes and Links can have associated attributes

- Trees: networks with hierarchical structure
  - Special case of networks
  - No cycles

➔ Networks



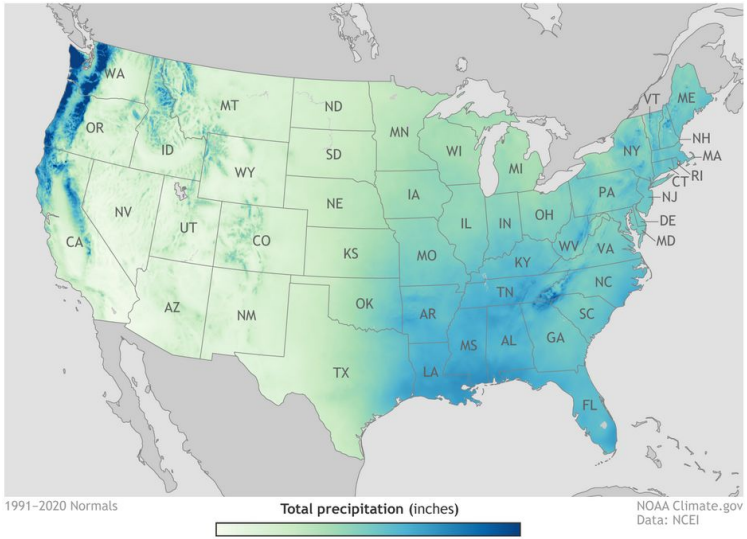
➔ *Trees*



Munzner, Figure 2.1

Dataset Types: Fields and Spatial Fields

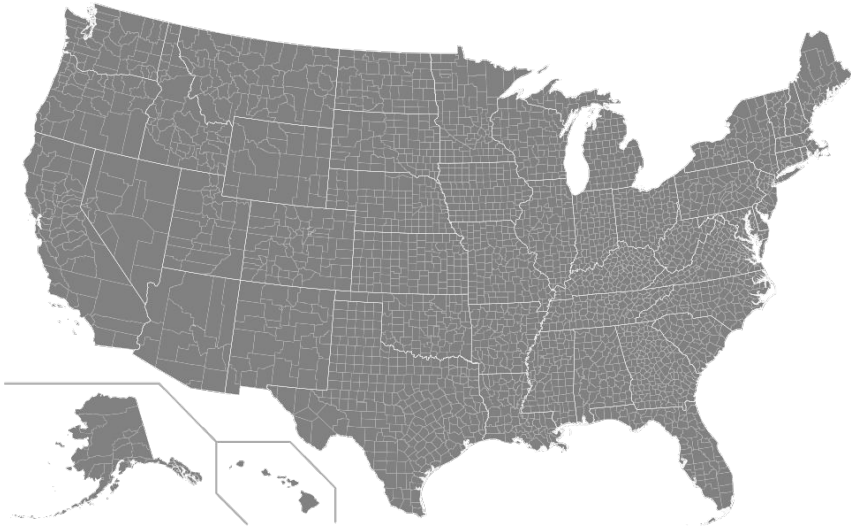
- Contains attribute values associated with cells
  - Spatial fields: cell structure is based on spatial positions
- Each cell contains value from a continuous domain
  - Measured, simulated
  - Example: temperature, precipitation
- Concerns
  - Sampling: where attributes are measured
  - Interpolation: how to model attributes elsewhere
  - Type of grid
- Attributes can be scalar, vector, or tensor



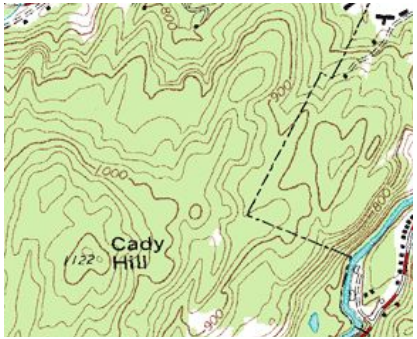
https://www.noaa.gov/climate

Dataset Types: Geometry

- Specifies information about the shape of items with explicit spatial positions/regions
  - Items can be points, lines, curves, surfaces, volumes
- Do not necessarily have attributes
- Examples:
  - Boundaries of a county, a state, or a country
  - Contours derived from a spatial field



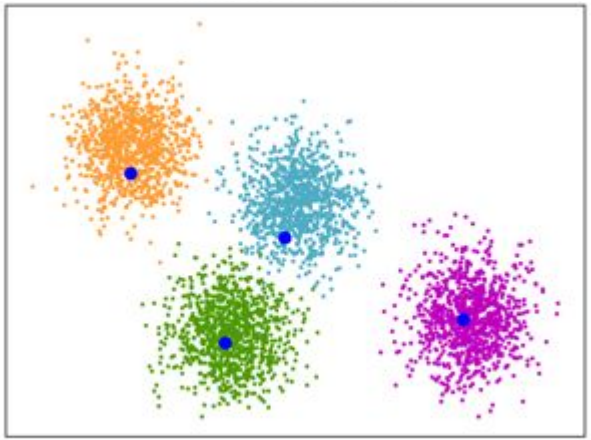
https://en.wikipedia.org/wiki/List_of_United_States_counties_and_county_equivalents



https://en.wikipedia.org/wiki/Contour_line

Other Combinations: Collections

- Form of grouping of items
- Sets: unordered groups of unique items
- Lists: groups of items with specified ordering
  - Possible duplicates
- Cluster: groups of items with attribute similarity



https://scikit-learn.org/

Data Types in Dataset Types

Different combinations of data types give rise to different dataset types

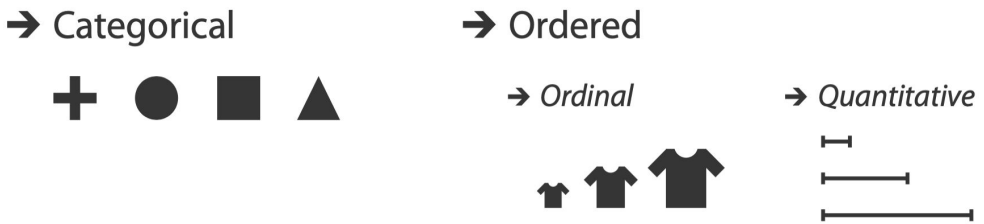
- Tables: items and attributes
- Networks and trees: items, links, and attributes
- Fields: grids, positions, and attributes
- Geometry: items, and positions
- Collections: items

Dataset Availability

- Static/offline dataset: entire dataset is available at once
  - Average daily traffic volume or accidents per intersection in January 2024
- Dynamic/online dataset: dataset information arrives or updates over the course of visualization session
  - Add new items
  - Change values of existing items
  - Live vehicle counts, live traffic incidents
- Any of the basic dataset types can be either static or dynamic

Attribute Types

- Classes of values and measurements
- Categorical attributes
  - No implicit ordering
  - Compare equality (same or different)
    - Apples versus oranges
- Ordered attributes
  - Have an implicit ordering
    - Shirt sizes (XS, S, M, L, XL)
    - Temperature (32F, 60F, 90F)
  - Ordinal data
  - Quantitative data



Munzner, Figure 2.7

Ordered Attributes

- Ordinal data
  - Well defined ordering
    - Less/greater than defined
    - XL > L > M > S > XS
  - Arithmetic operations not possible
    - L - M ?, XL - XS?
- Quantitative data
  - Well defined ordering and arithmetic operations are possible
  - 90 F > 80 F
  - 90F - 80F = 10F

Munzner, Figure 2.7

Attribute Types: Example

- What is/are categorical?
- What is/are ordinal?
- What is/are quantitative?
- What are the reasons?

Name	Salary	Education Level	Favorite Drink
Alex	48,000	Undergraduate	Coffee
Maya	72,000	Graduate	Tea
Jordan	115,000	PhD	Coffee
Sam	65,000	Undergraduate	Juice
Priya	98,000	Graduate	Tea
Leo	120,000	High School	Coffee
Nina	42,000	Undergraduate	Soda
Omar	88,000	Graduate	Water

Attribute Types: Example

- What is/are categorical?
  - Name, Favourite Drink
  - No implicit ordering
- What is/are ordinal?
  - Education level
  - There's ordering, but no arithmetic operations
- What is/are quantitative?
  - Salary
  - There's ordering and meaningful arithmetic operations

Name	Salary	Education Level	Favorite Drink
Alex	48,000	Undergraduate	Coffee
Maya	72,000	Graduate	Tea
Jordan	115,000	PhD	Coffee
Sam	65,000	Undergraduate	Juice
Priya	98,000	Graduate	Tea
Leo	120,000	High School	Coffee
Nina	42,000	Undergraduate	Soda
Omar	88,000	Graduate	Water

Attribute Types: Example 2

What are the types of attributes?

Order ID	Customer	Order Date	Order Priority	Items Purchased	Total
1001	Alex	2025-01-05	Low	2	48.75
1002	Maya	2025-01-06	Medium	5	129.40
1003	Jordan	2025-01-06	High	1	19.99
1004	Sam	2025-01-08	Medium	3	76.20
1005	Priya	2025-01-10	High	4	210.00
1006	Leo	2025-01-12	Low	6	95.50

Attribute Types: Example 2

Categorical: customer, orderID

Ordinal: priority

Quantitative: date, total , items purchases

Order ID	Customer	Order Date	Order Priority	Items Purchased	Total
1001	Alex	2025-01-05	Low	2	48.75
1002	Maya	2025-01-06	Medium	5	129.40
1003	Jordan	2025-01-06	High	1	19.99
1004	Sam	2025-01-08	Medium	3	76.20
1005	Priya	2025-01-10	High	4	210.00
1006	Leo	2025-01-12	Low	6	95.50

Ordered Data: Direction

- Sequential: values range from minimum to maximum
  - Mountain height data (minimum = 0 for MSL, maximum = height of Everest)
  - Bathymetry data (minimum = depth of Challenger Deep, maximum = 0 for MSL)
- Diverging: values are from two sequences pointing in opposite directions that meet at a common point
  - Full elevation dataset (- values for ocean valleys, + for mountains, 0 for MSL)
- Cyclic: values wrap around back to a starting point
  - Days of the week, hour of the day
  - Direction of wind

Data Abstraction

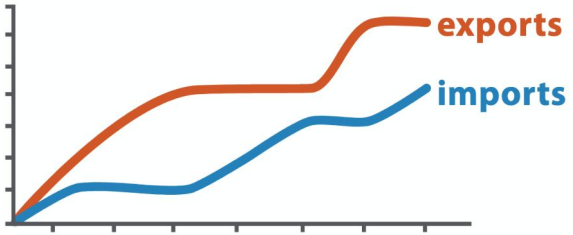
- Translate from domain-specific language to generic visualization language
- Identify dataset type(s), attribute types
- Identify cardinality
  - How many items in the dataset?
  - What is cardinality of each attribute?
    - Number of levels for categorical data
    - Range for quantitative data
- Consider whether to transform data
  - Guided by understanding of task

Data versus Conceptual Model: Example

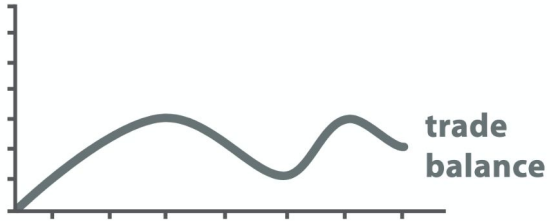
- Data model: floats
  - 32.52, 54.06, -14.35, ...
- Conceptual model
  - Temperature
- Possible data abstractions
  - Forecasting weather
    - Continuous to 2 significant figures: quantitative
  - Deciding if bath water is ready
    - Hot, warm, cold: ordinal
  - Decide if I should leave the house
    - Above freezing, below freezing: categorical

Derived Attributes

- Compute attributes from given data
  - Simple change of type
    - Date to Day
    - 01/26/2025 to Monday
  - Acquire additional data
    - 36.85° N, 76.29° W is Norfolk Downtown
  - Complex transformation
- Can reduce cognitive burden



Original Data



trade balance = *exports* – *imports*

Derived Data

Data Cleaning

Data Formats

Data comes in many formats!!

- Delimited Text
  - Tab separated (TSV)
  - Comma separated (CSV)
- Extensible Markup Language (XML)
  - Looks a bit like HTML
  - User-defined tags to identify data
- JavaScript Object Notation (JSON)
  - Collection of name/value pairs
  - Smaller than XML
  - Easier to parse
- Hierarchical Binary Formats
  - Not human readable
  - Used in scientific applications
  - HDF5
- Columnar Storage Formats
  - Big data analytics
  - Parquet, ORC (Optimized Row Columnar)
- …

Data Formats: Example

```
order_id,customer,date,total
1001,Alex,2025-03-01,48.75
1002,Maya,2025-03-02,72.00
1003,Jordan,2025-03-03,115.50
```

CSV

```
<orders>
  <order>
    <order_id>1001</order_id>
    <customer>Alex</customer>
    <date>2025-03-01</date>
    <total>48.75</total>
  </order>
</orders>
```

XML

```
[
  {
    "order_id": 1001,
    "customer": "Alex",
    "date": "2025-03-01",
    "total": 48.75
  },
  {
    "order_id": 1002,
    "customer": "Maya",
    "date": "2025-03-02",
    "total": 72.00
  }
]
```

JSON

Converting Data Between Different Formats

- Write your own program (Python, Perl, …)
  - Built-in packages (JSON, CSV, PyYAML…)
  - External libraries (Pandas, NumPy, …)
- Data tools
  - Spreadsheets (Excel, Google Sheets, …)
  - OpenRefine
  - Tableau
- Online tools
  - Cloudconvert: https://cloudconvert.com/
  - Mr. Data Converter: https://shancarter.github.io/mr-data-converter/
  - Search for “csv to json”, ….

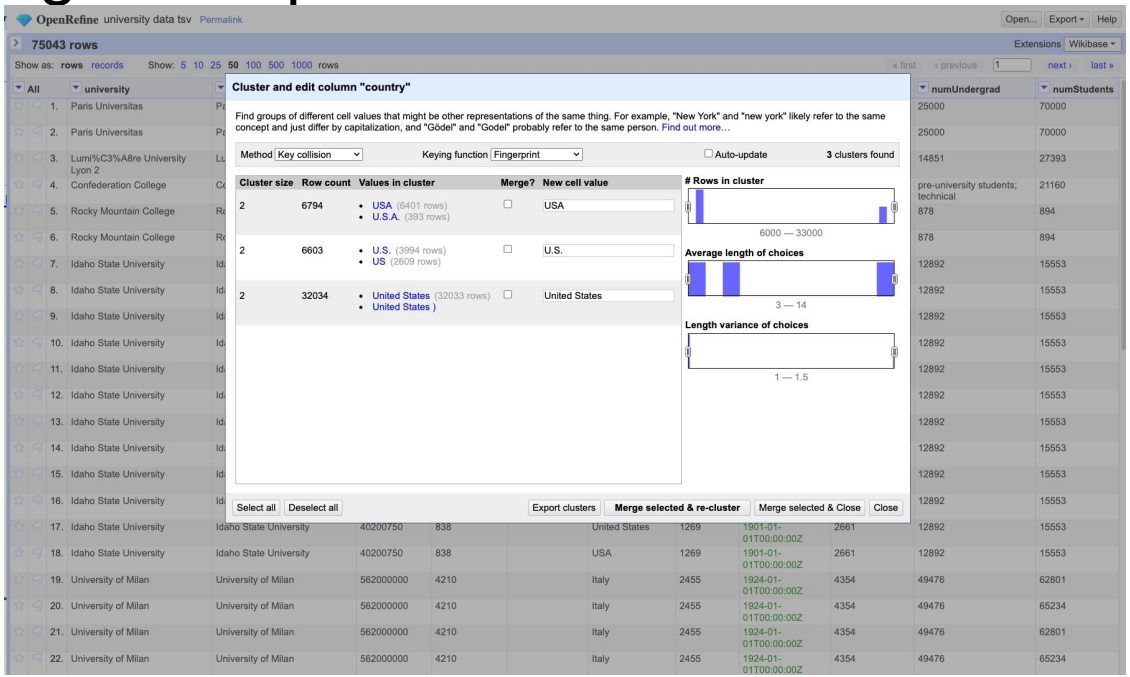
Real World Data is Messy!

In practice, datasets often contain:

- Missing data
- Invalid values
- Misfielded values
- Spelling errors
- Formatting inconsistencies
- …

order_id	customer	order_date	order_total
1001	Alex	03/01/2025	48.75
1002		2025-03-02	72
1003	Jordon	2025/03/03	one hundred
10O4	Maya	2025-13-05	65.00

Data Cleaning with OpenRefine



About expressions: https://openrefine.org/docs/manual/expressions
GREL: https://openrefine.org/docs/manual/grel
Regex101: https://regex101.com/