3.Data Abstraction

Prof. Tulasi Prasad Sariki
SCSE, VIT, Chennai
www.learnersdesk.weebly.com

Outline

- What can be visualized?
- Why Do Data Semantics and Types Matter?
- Data Types
  - Items, Attributes, Links, Positions, and Grids
- Dataset Types
  - Tables, Networks, Fields, and Geometry
- Attribute Types
  -

What can be visualized?

- The four basic dataset types are
  - tables, networks, fields, and geometry;
  - other items : clusters, sets, and lists.
- The datasets are made up of different combinations of the five data types: items, attributes, links, positions, and grids.
- For any of these dataset types, the full dataset could be
  - Available immediately(static file)
  - Stream data processed gradually(dynamic file)
-

What can be visualized?

- The type of an attribute can be categorical or ordered, with a further split into ordinal and quantitative. The ordering direction of attributes can be sequential, diverging, or cyclic.

What?

Datasets

Data Types

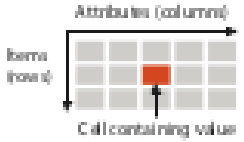
- Items
- Attributes
- Links
- Positions
- Grids

Data and Dataset Types

Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

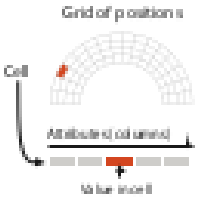
Dataset Types

- Tables

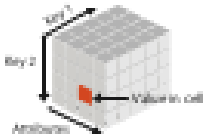


- Networks

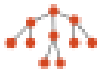
- Fields (Continuous)



- *Multidimensional Table*



- *Trees*



- Geometry (Spatial)

Dataset Availability

- Static

- Dynamic

Attributes

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

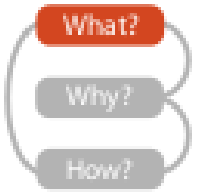
  - *Quantitative*

Ordering Direction

- Sequential

- Diverging

- Cyclic



Why Do Data Semantics and Types Matter?

- Many aspects of vis design are driven by
  - What kind of data are you given?
  - What information can you figure out from the data, versus the meanings that you must be told explicitly?
  - What high-level concepts will allow you to split datasets apart into general and useful pieces?

Why Do Data Semantics and Types Matter?

- What does these sequences mean?
  - 14, 2.6, 30, 30, 15, 100001
  - VIT, 7, S, Chennai

- To move beyond guesses, you need to know their semantics and types.
  - The semantics of the data is its real-world meaning.
  - The type of the data is its structural or mathematical interpretation.

Why Do Data Semantics and Types Matter?

- Sometimes types and semantics can be correctly inferred simply by observing the ***syntax of a data file or the names of variables*** within it, but often they must be provided along with the dataset in order for it to be interpreted correctly. Sometimes this kind of additional information is called ***metadata***;

ID	Name	Age	Shirt Size	Favorite Fruit
1	Amy	8	S	Apple
2	Basil	7	S	Pear
3	Clara	9	M	Durian
4	Desmond	13	L	Elderberry
5	Ernest	12	L	Peach
6	Fanny	10	S	Lychee
7	George	9	M	Orange
8	Hector	8	L	Loquat
9	Ida	10	M	Pear
10	Amy	12	M	Orange

Data Types

Five basic data types :

- Items, Attributes, Links, Positions, and Grids.
- An attribute(variable /dimension) is some specific property that can be measured, observed, or logged.
- For example, attributes could be salary, price, number of sales, protein expression levels, or temperature.
- An item is an individual entity that is discrete, such as a row in a simple table or a node in a network.

Data Types

- For example, items may be people, stocks, coffee shops, genes, or cities.
- A link is a relationship between items, typically within a network.
- A grid specifies the strategy for sampling continuous data in terms of both geometric and topological relationships between its cells.
- A position is spatial data, providing a location in two-dimensional (2D) or three-dimensional (3D) space.
- For example, a position might be a latitude–longitude pair describing a location on the Earth's surface.

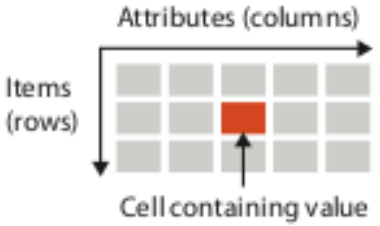
Dataset Types

- A dataset is any collection of information that is the target of analysis.
- The four basic dataset types are tables, networks, fields, and geometry.
- Other ways to group items together include clusters, sets, and lists.
- In real-world situations, complex combinations of these basic types are common.

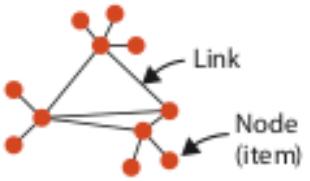
Dataset Types

Dataset Types

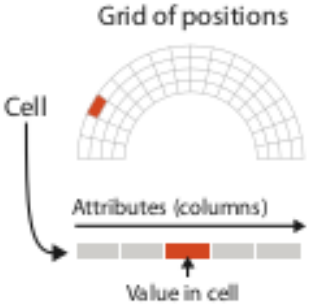
→ Tables



→ Networks

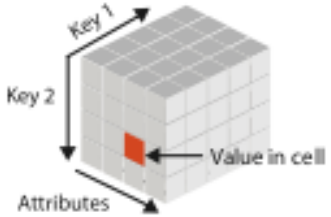


→ Fields (Continuous)



→ Geometry (Spatial)

→ *Multidimensional Table*



→ *Trees*

Tables

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box		7/17/07
32	7/16/07	2-High	Medium Box		7/18/07
32	7/16/07	2-High	Medium Box		7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	5	4-Not Specified	Small Pack	0.44	6/6/05
69	5	4-Not Specified	Wrap Bag	0.6	6/6/05
70	12/18/06	5-Low	Small Box	0.59	12/23/06
70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
96	4/17/05	2-High	Small Box	0.55	4/19/05
97	1/29/06	3-Medium	Small Box	0.38	1/30/06
129	11/19/08	5-Low	Small Box	0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

attribute

item

cell

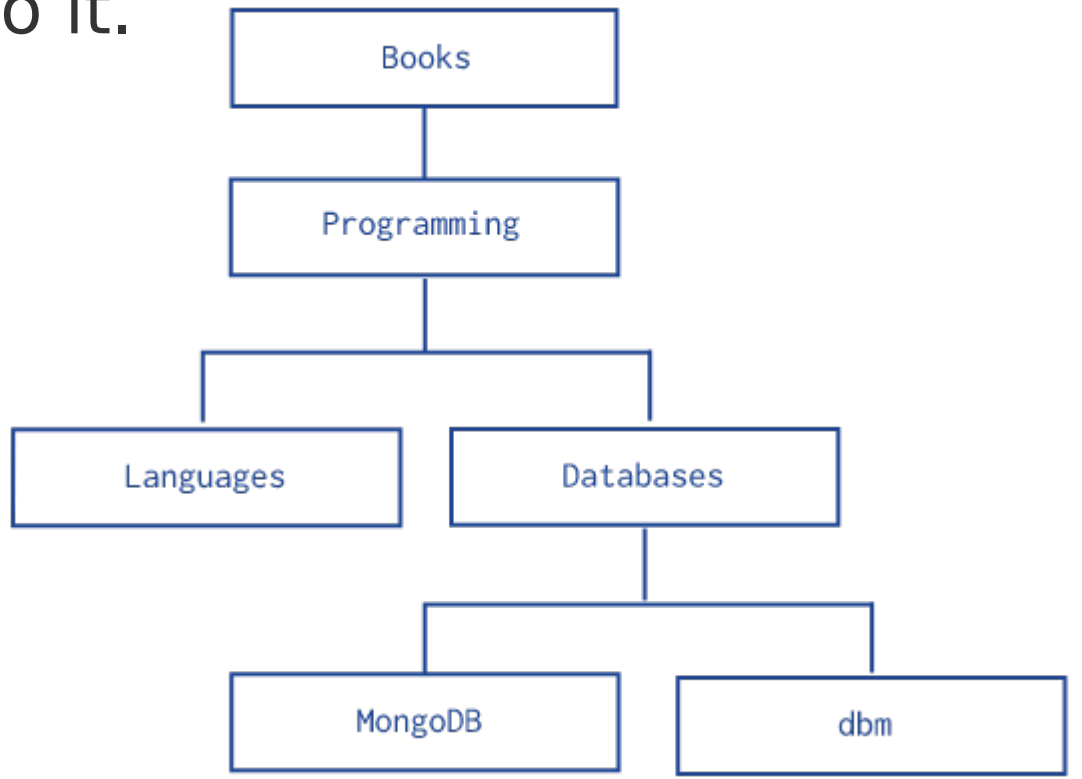
A multidimensional table has a more complex structure for indexing into a cell, with multiple keys.

Networks

- The dataset type of networks is well suited for specifying that there is some kind of relationship between two or more items.
- An item in a network is often called a node.
- A link is a relation between two items.
- For example, in an articulated social network the nodes are people, and links mean friendship.
- Nodes & Links can have associated attributes, just like items in a table.

Trees

- Networks with hierarchical structure are more specifically called trees.
- In contrast to a general network, trees do not have cycles: each child node has only one parent node pointing to it.
- Example



Fields

- The field dataset type also contains attribute values associated with cells. Each cell in a field contains measurements or calculations from a ***continuous*** domain.
- Continuous data requires careful treatment
  - Sampling : How frequently to take measurements
  - Interpolation : How to show values in between the sampled points in a way that does not mislead.

Geometry

- The geometry dataset type specifies information about the shape of items with explicit spatial positions. The items could be points, or one-dimensional lines or curves, or 2D surfaces or regions, or 3D volumes.
- Spatial data often includes hierarchical structure at multiple scales.

Other Combinations

- There are many ways to group multiple items together, including sets, lists, and clusters.
- A ***set*** is simply an un-ordered group of items.
- A group of items with a specified ordering could be called a ***list***.
- A ***cluster*** is a grouping based on attribute similarity(items within a cluster are more similar to each other than to ones in another cluster).

Dataset Availability

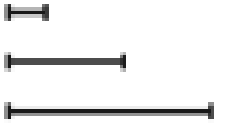
- The default approach to vis assumes that the entire dataset is available all at once, as a ***static*** file(offline). However, some datasets are instead ***dynamic*** streams(online).
- Designing for streaming data adds complexity to many aspects of the vis process that are straightforward when there is complete dataset availability up front.

Attribute Types

- **Attribute Types**
  - Categorical

  - Ordered
    - *Ordinal*

    - *Quantitative*



- **Ordering Direction**
  - Sequential

  - Diverging

  - Cyclic

Attribute Types

- ***Categorical data*** - such as favorite fruit or names, does not have an implicit ordering, but it often has hierarchical structure.
- Other examples of categorical attributes are movie genres, file types, and city names.
- ***Ordered: Ordinal and Quantitative*** - does have an implicit ordering, as opposed to un-ordered categorical data.

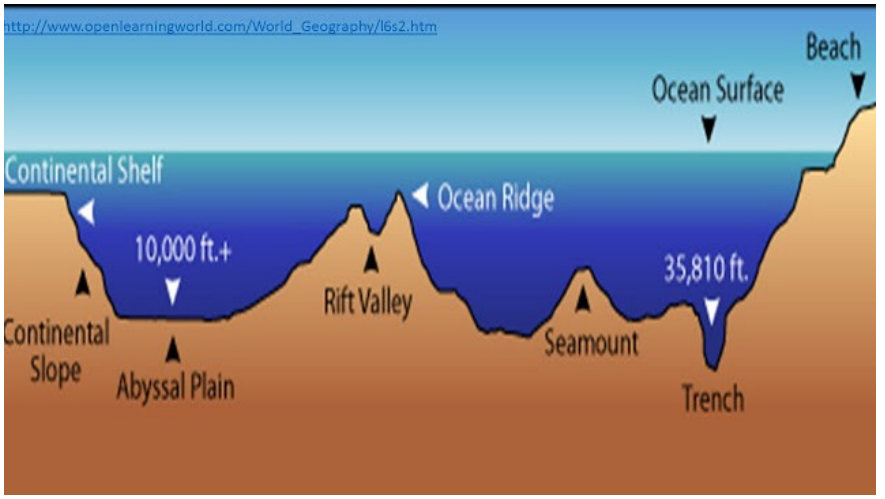
Attribute Types

- This type can be further subdivided. With ***ordinal*** data, such as shirt size, we cannot do full-fledged arithmetic, but there is a well-defined ordering.
- A subset of ordered data is ***quantitative*** data, namely, a measurement of magnitude that supports arithmetic comparison.
- For example, the quantity of 68 inches minus 42 inches is a meaningful concept

Attribute Types

- Ordered data can be either ***sequential***, where there is a homogeneous range from a minimum to a maximum value, or ***diverging***, which can be deconstructed into two sequences pointing in opposite directions that meet at a common zero point.
- For example mountain height dataset is sequential, when measured from a minimum point of sea level to a maximum point of Mount Everest.
- The full elevation dataset would be diverging, where the values go up for mountains on land and down for undersea valleys, with the zero value of sea level being the common point joining the two sequential datasets.

Attribute Types



- Ordered data may be ***cyclic***, where the values wrap around back to a starting point rather than continuing to increase indefinitely.
- Examples like hour of the day, the day of the week, and the month of the year.

Attribute Types

A	B	C	S	T	U
Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
3	10/14/06	5-Low	Large Box	0.8	10/21/06
6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
32	7/16/07	2-High	Small Pack	0.79	7/17/07
32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
32	7/16/07	2-High	Medium Box	0.6	7/18/07
32	7/16/07	2-High	Medium Box	0.65	7/18/07
35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/4/05	4-Not Spec	quantitative	0.6	6/6/05
70	12/18/06	5-Low	ordinal	0.59	12/23/06
70	12/18/06	5-Low	categorical	0.82	12/23/06
96	4/17/05	2-High		0.55	4/19/05
97	1/29/06	3-Medium		0.38	1/30/06
129	11/19/08	5-Low		0.37	11/28/08
130	5/8/08	2-High	Small Box	0.37	5/9/08
130	5/8/08	2-High	Medium Box	0.38	5/10/08
130	5/8/08	2-High	Small Box	0.6	5/11/08
132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
166	9/12/07	2-High	Small Box	0.55	9/14/07
193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Thank You