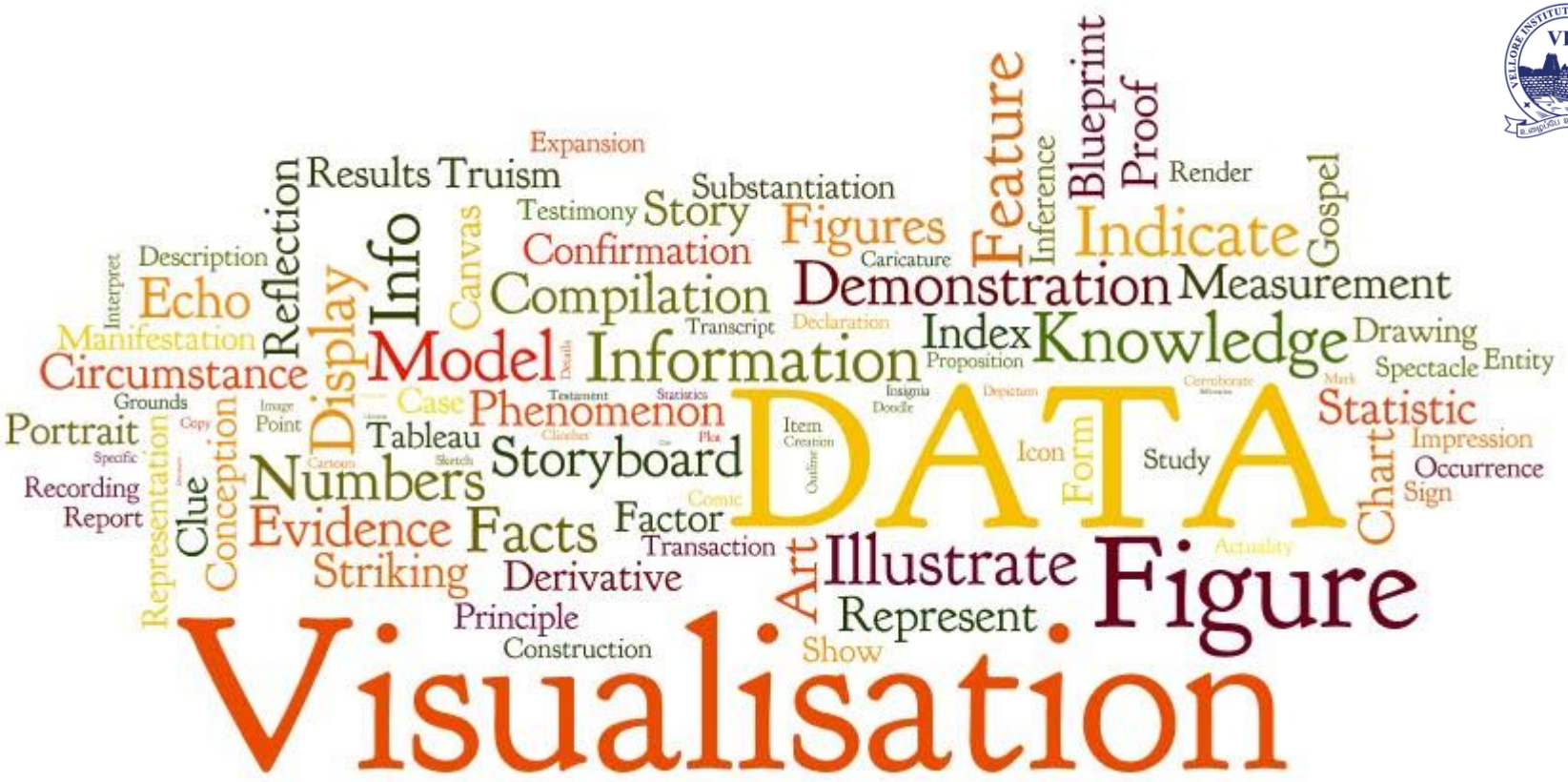


3.Data Abstraction

Prof. Ramesh Ragala

SCSE, VIT, Chennai

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

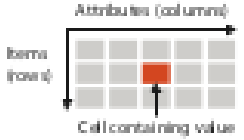
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

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

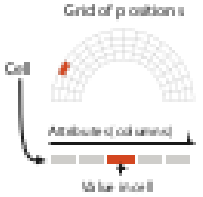
Dataset Types

→ Tables

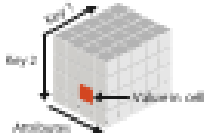


→ Networks

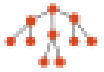
→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*



→ Geometry (Spatial)

Dataset Availability

→ Static

→ Dynamic

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

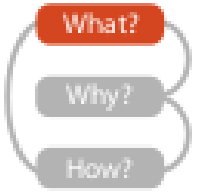
→ *Quantitative*

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 → many guesses
  - VIT, 20000, S, Chennai → many guesses
- You should know the interpretation to understand
- To move beyond guesses, you need to know their semantics and types.
  - The semantics of the data is its real-world meaning.
    - Word represent the first name of a human or city
  - The type of the data is its structural or mathematical interpretation.
    - @ data level → what kind of things is it.. An item, a link, an attribute?
    - @ dataset level →
      - how these data types are combined into a larger structure. Table, tree
      - Attribute level: what type of mathematical operations can be

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.
  - Example: 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

  - Example: 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.
  - If points are **arbitrarily distributed** and there is no connectivity between them, the data is called **scattered**
  - Otherwise, data is composed of cells bounded by grid lines.
- A **position** is spatial data, providing a location in two-dimensional (2D) or three-dimensional (3D) space.
  - 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, which is the target of analysis.
- The **four basic dataset types** are
  - tables,
  - Networks
  - fields
  - geometry.
- Other ways to **group items** together include clusters, sets, and lists.
- In real-world situations, complex combinations of these basic types are common.

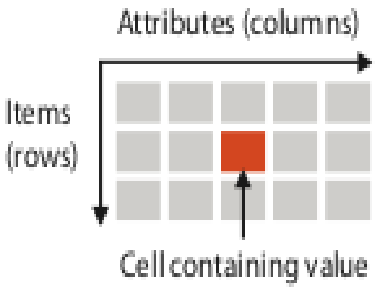
Data and Dataset Types

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

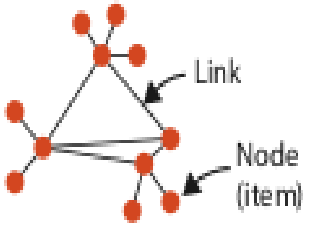
Dataset Types

Dataset Types

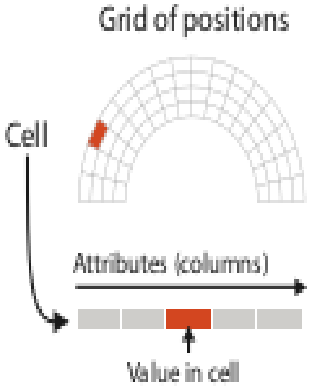
→ Tables



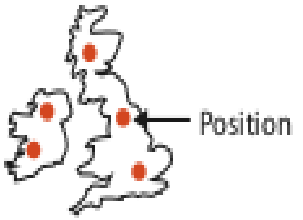
→ Networks



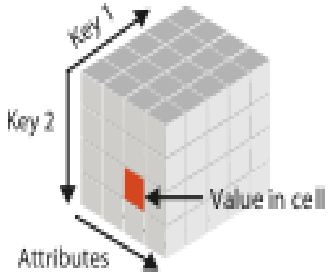
→ Fields (Continuous)



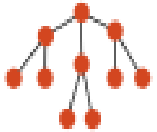
→ Geometry (Spatial)



→ *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	[illegible]	5-Low	Wrap Bag	0.56	1/20/05
69	[illegible]	4-Not Specified	Small Pack	0.44	6/6/05
69	[illegible]	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

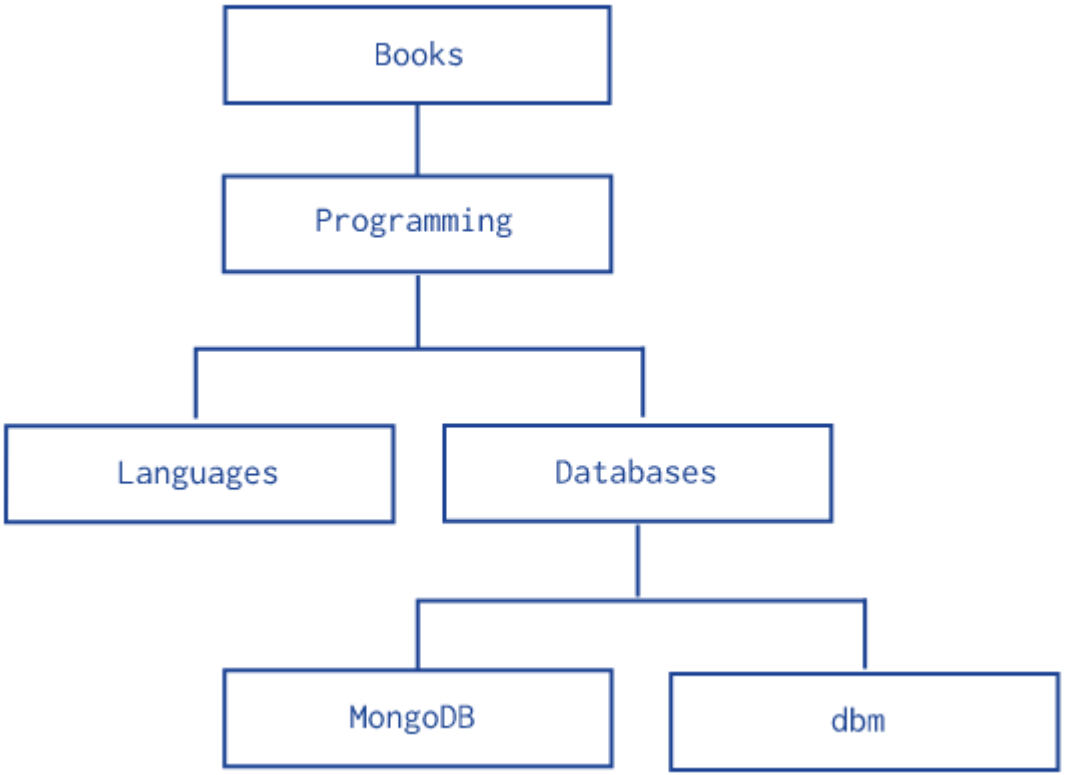
A multidimensional table has a more complex structure for indexing into a cell, with multiple keys. → cube cells

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.
- Example: 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 → temperature data, medical scan of human body indicating density of tissues at many places.
- 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.

Fields

- Spatial fields:
  - Continuous data is often found in the form of a **spatial field,** where the cell structure of the field is based on sampling at spatial positions.
  - Example: an X-ray image of the human body and to use color coding to highlight suspected tumors (with shape).
- Grid Types:
  - When a field contains data created by sampling at completely regular intervals, the cells form a **uniform grid**.

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.
- Geometry datasets **do not** necessarily have **attributes**, in contrast to the other three basic dataset types.

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).

Other Combinations

- A **compound network**
  - **It** is a network with an associated tree: all of the nodes in the network are the leaves of the tree, and interior nodes in the tree provide a hierarchical structure for the nodes that is different from network links between them.
- A **path** through a network is an ordered set of segments formed by links connecting nodes.
- *what* part of an analysis instance that pertains to data; that is, the **data abstraction**

Other Combinations

- In simple cases, it may be possible to describe your data abstraction using only that set of terms.

- In complex cases, you may need additional description as well.
- If so, your goal should be to translate domain-specific terms into words that are as generic as possible.

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).
  - One kind of dynamic change is to add new items or delete previous items.
  - Another is to change the values of existing items.
- 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

- Different Types of Data in data visualization:
  - Quantitative Data
    - Continuous Data
    - Discrete Data
      - Interval Data
      - Ratio Data
  - Qualitative Data
    - Ordinal Data
  - Nominal Data

Attribute Types

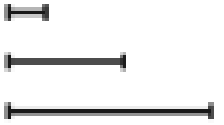
	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	Units (example)	Ordinal (example)	Nominal (example)	Binary (example)
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date	Month, date, year	Jan., Feb., Mar., etc.	N/A	Before / After
Cycle time	Hours, minutes, seconds, month, date, year	10, 20, 30, etc.	N/A	Before / After
Speed	Miles per hour/centimeters per second	10, 20, 30, etc.	N/A	Fast / Slow
Brightness	Lumens	Light, medium, dark	N/A	On / Off
Temperature	Degrees C or F	10, 20, 30, etc.	N/A	Hot / Cold
<Count data>	Number of things	10, 20, 30, etc.	N/A	Large / Small
Test scores	Percent, number correct	F, D, C, B, A	N/A	Pass / Fail
Defects	N/A	Number of cracks	N/A	Good / Bad
Defects	N/A	N/A	Oversized, missing	Good / Bad
Color	N/A	N/A	Red, blue, green	N/A
Location	N/A	N/A	East, West, South	Domestic / International
Groups	N/A	N/A	HR, Legal, IT	Exempt / Non-exempt
Anything	Percent	10, 20, 30, etc.	N/A	Above / Below

Attribute Types

- **Attribute Types**
  - Categorical

  - Ordered
    - *Ordinal*

    - *Quantitative*



- **Ordering Direction**
  - Sequential

  - Diverging

  - Cyclic

Attribute Types

- ***Continuous data***:
  - Theoretically has an infinite number of measurements depending on the resolution of the measurement system. There are no limits to the gaps between the measurements.
  - It is data that can be expressed on an infinitely divisible scale.
  - Even if the measurements range from 0-1 there may be an infinite number of measurements within (0.000000000000... to 0.999999999999...)

Attribute Types

- ***Discrete data***:
  - Data types that have a finite number of measurements and are based on counts.
  - Data that can be sorted into distinct, countable, and in completely separate categories.
  - The count value can not be divided further on an infinite scale with meaning.
  - Example: How many people can comfortably sit in class?

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.

Attribute Types

- Nominal/categorical data:
  - The data values are categorical and not numeric.
  - A categorical variable is one that has two or more categories or labels or classes, but there is no intrinsic ordering to the categories.
  - simply Categorical variables represent types of data which may be divided into groups.
  - It is completely qualitative measurement.
  - Examples: age, gender, educational levels, countries, people names. operations: == and !=
  - Comparing two observations using the values for the variable, the observations will either be similar or different depending on whether the categorical value matches or not.

Attribute Types

- Nominal/categorical data:

- From another slides

Attribute Types

- ***Ordered: Ordinal and Quantitative***
  - does have an implicit ordering, as opposed to un-ordered categorical data.

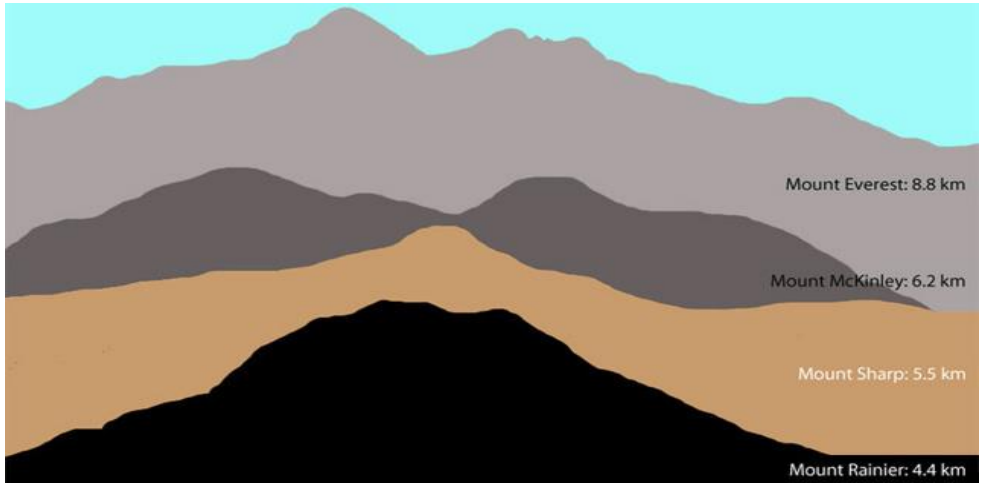
Attribute Types

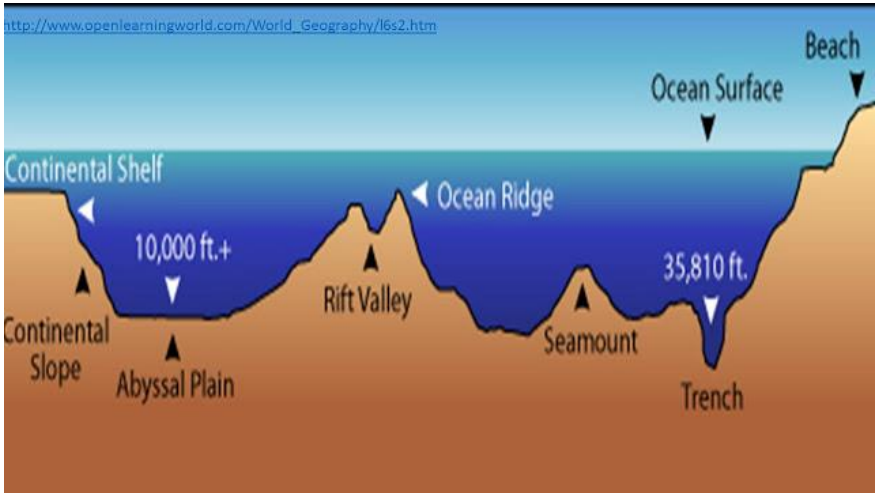
- This type can be further subdivided. With ***ordinal*** data, such as shirt size, we cannot do full-fledged arithmetic, but there is a well-defined ordering.
- A subset of ordered data is ***quantitative*** data, namely, a measurement of magnitude that supports arithmetic comparison.
- For example, 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		0.6	6/6/05
70	12/18/06	5-Low		0.59	12/23/06
70	12/18/06	5-Low		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

quantitative
ordinal
categorical

Thank You