DATA ABSTRACTION

Bùi Tiến Lên

2023

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

Contents

The Big Picture

Data Types

Data Types

The five basic **data types**:

1. An **item** is an individual entity that is discrete, such as a row in a simple table or a node in a network
2. An **attribute** is some specific property that can be measured, observed, or logged
3. A **link** is a relationship between items, typically within a network
4. A **position** is spatial data, providing a location in two-dimensional (2D) or three-dimensional (3D) space
5. A **grid** specifies the strategy for sampling continuous data in terms of both geometric and topological relationships between its cells

➔ **Data Types**

➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

Dataset Types

- Tables
- Networks and Trees
- Fields
- Geometry
- Other Combinations
- Dataset Availability

Dataset Types

Concept 1

A **dataset** is any collection of information that is the target of analysis. Data sets are made up of data objects.

- These basic **dataset types** arise from combinations of the data types of items, attributes, links, positions, and grids.

Data and Dataset Types

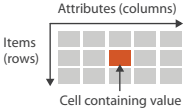
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types (cont.)

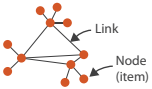
- The detailed structure of the four basic dataset types

Dataset Types

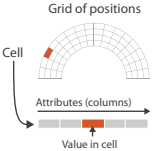
Tables



Networks

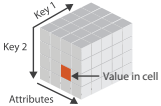


Fields (Continuous)



Geometry (Spatial)

Multidimensional Table



Trees

Tables

- Many datasets come in the form of **tables** that are made up of rows and columns, a familiar form to anybody who has used a spreadsheet
- For a simple **flat table**
  - Each row represents an **item** of data, and each column is an **attribute** of the dataset
  - Each **cell** in the table is fully specified by the combination of a row and a column—an item and an attribute—and contains a value for that pair
- A **multidimensional table** has a more complex structure for indexing into a cell, with multiple keys

Tables (cont.)

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	[illegible]	7/17/07
32	7/16/07	2-High	Medium Box	[illegible]	7/18/07
32	7/16/07	2-High	Medium Box	[illegible]	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

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.

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

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 that takes into account the mathematical questions of **sampling** and **interpolation**
- In contrast, the table and network datatypes discussed above are an example of **discrete** data where a finite number of individual items exist, and interpolation between them is not a meaningful concept.

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

Grid Types

- When a field contains data created by sampling at completely regular intervals, the cells form a **uniform grid**
- There is no need to explicitly store the **grid geometry** in terms of its location in space, or the **grid topology** in terms of how each cell connects with its neighboring cells

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.
- Geometry datasets are intrinsically spatial. 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

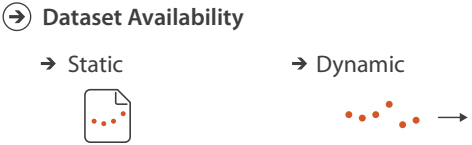
- A **set** is simply an unordered group of items
- A group of items with a specified ordering could be called a **list**
- A **cluster** is a grouping based on attribute similarity

There are also more complex structures built on top of the basic network type

- A **path** through a network is an ordered set of segments formed by links connecting nodes
- A **compound network** is a network with an associated tree

Dataset Availability

- The default approach to vis assumes that the entire dataset is available all at once, as a **static file**
- Some datasets are **dynamic streams**, where the dataset information trickles in over the course of the vis session



Attribute Types

Attribute types

Concept 2

An attribute (also called dimension, feature, variable) is a data field, representing a characteristic or feature of a data object.

- At the top level,
  - we can differentiate **qualitative** (or **categorical**) and **quantitative** (or **numerical**) attribute.
- At a second level,
  - we can categorize qualitative data into **nominal** and **ordinal** attribute,
  - and quantitative data into **discrete** and **continuous** attribute.

Level of measurement

- Describe the nature of information within the values assigned to variables

Type	Measure property	Mathematical operators	Advanced operations	Central tendency	Variability
Nominal	Classification, membership	$=, \neq$	Grouping	Mode	Qualitative variation
Ordinal	Comparison, level	$>, <$	Sorting	Median	Range, interquartile range
Interval	Difference, affinity	$+, -$	Comparison to a standard	Arithmetic mean	Deviation
Ratio	Magnitude, amount	$*, /$	Ratio	Geometric mean, harmonic mean	Coefficient of variation, studentized range

Nominal Attribute

Concept 3

Nominal attribute represents *things*

- His *name* is Brent Spiner.
- By *profession* he is an actor.
- He played the *character* Data in the TV show *Star Trek: The Next Generation*.

Ordinal Attribute

Concept 4

Ordinal attribute is similar to categorical data, except it has a clear order

- Brent Spiner's *date of birth* is Wednesday, February 2, 1949.
- He appeared in all seven *seasons* of *Star Trek: The Next Generation*.
- Data's *rank* was lieutenant commander.

Discrete Attribute

Concept 5

Discrete data are numeric data whose domain can be equated to the set of whole numbers $\mathbb{Z}$

An example of discrete data would be *the number of people* visiting a doctor.

Continuous Attribute

Concept 6

Continuous data are numeric data whose domain can be equated to the set of real numbers $\mathbb{R}$.

An example of continuous data would be *temperature values* as measured hourly by a weather station.

Sequential, Diverging and Cyclic

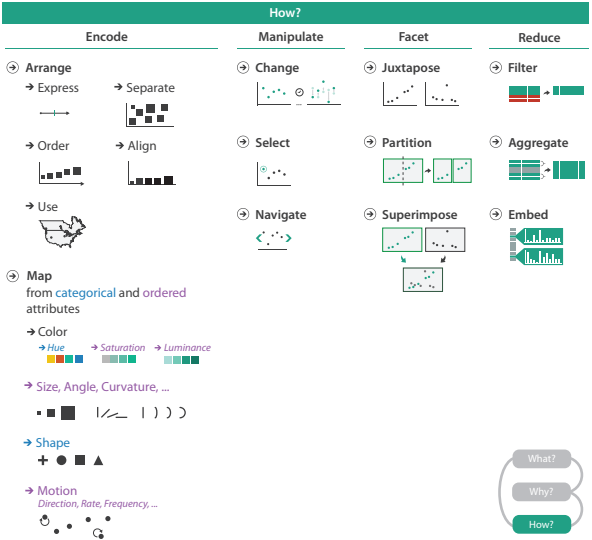
Ordered data can be

- **sequential**, where there is a homogeneous range from a minimum to a maximum value,
- **diverging**, which can be deconstructed into two sequences pointing in opposite directions that meet at a common zero point
- **cyclic**, where the values wrap around back to a starting point rather than continuing to increase indefinitely.

Hierarchical Attributes

- There may be hierarchical structure within an attribute or between multiple attributes.

The Big Picture



Data Processing

Dataset

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

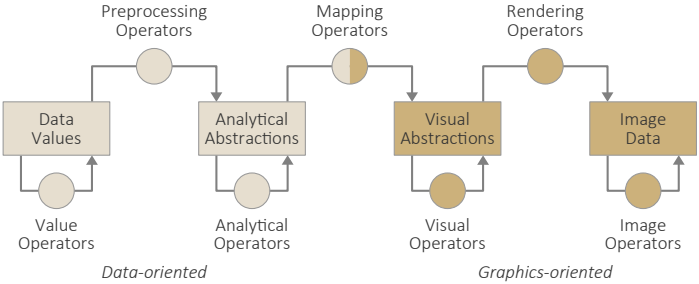
Functional dependency

$$f : (D_1 \times D_2 \times \cdots \times D_n) \rightarrow (A_1 \times A_2 \times \cdots \times A_m) \tag{1}$$

where D_i denote the *dimensions* (*independent variables*) and A_i the *attributes* (*dependent variables*)

D_1	D_2	$\cdots$	D_n	A_1	A_2	$\cdots$	A_m
$d_{1,1}$	$d_{1,2}$	$\cdots$	$d_{1,n}$	$a_{1,1}$	$a_{1,2}$	$\cdots$	$a_{1,m}$
$\vdots$							$\vdots$
$d_{k,1}$	$d_{k,2}$	$\cdots$	$d_{k,n}$	$a_{k,1}$	$a_{k,2}$	$\cdots$	$a_{k,m}$

Data Transformation



References

Goodfellow, I., Bengio, Y., and Courville, A. (2016).
Deep learning.
MIT press.

Munzner, T. (2014).
Visualization analysis and design.
CRC press.

Russell, S. and Norvig, P. (2016).
Artificial intelligence: a modern approach.
Pearson Education Limited.

Ward, M. O., Grinstein, G., and Keim, D. (2015).
Interactive data visualization: foundations, techniques, and applications.
CRC Press.