WHAT: DATA ABSTRACTION

Bùi Tiến Lên

01/01/2020

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

Contents

The Big Picture

Why Do Data Semantics and Types Matter?

- The **semantics** of the data is its real-world meaning
- The **type** of the data is its structural or mathematical interpretation

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

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**

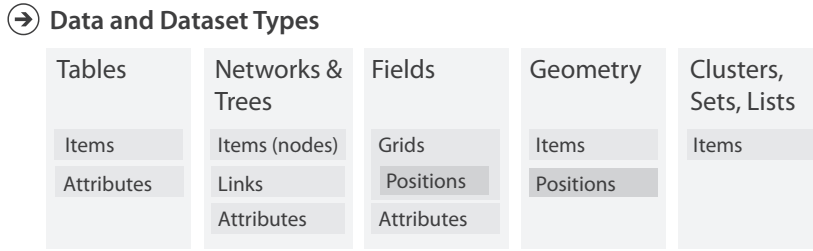
➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

Dataset Types

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

Dataset Types

- A **dataset** is any collection of information that is the target of analysis
- These basic **dataset types** arise from combinations of the data types of items, attributes, links, positions, and grids.

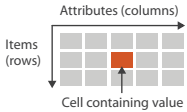


Dataset Types (cont.)

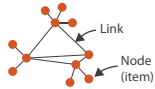
- The detailed structure of the four basic dataset types

Dataset Types

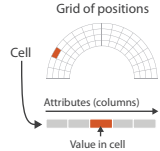
→ Tables



→ Networks

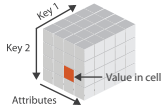


→ Fields (Continuous)



→ 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		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]5	4-Not Specified	Small Pack	0.44	6/6/05
69	[illegible]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

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

- Categorical
- Ordered
- Hierarchical Attributes

Attribute types

- The major disinction is between categorical versus ordered
  - **Attribute Types**
    - Categorical
    - Ordered
      - *Ordinal*
      - *Quantitative*

  - **Ordering Direction**
    - Sequential
    - Diverging
    - Cyclic

Categorical

- **Categories** can only distinguish whether two things are the same (apples) or different (apples versus oranges)
- The type of categorical data, such as favorite fruit or names, does not have an implicit ordering, but it often has hierarchical structure

Ordered

- All **ordered** data does have an implicit ordering, as opposed to unordered categorical data
- A subset of ordered data is **quantitative** data, namely, a measurement of magnitude that supports arithmetic comparison

Sequential versus Diverging

- 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

Cyclic

- Ordered data may be **cyclic**, where the values wrap around back to a starting point rather than continuing to increase indefinitely.
- Many kinds of time measurements are cyclic, including the hour of the day, the day of the week, and the month of the year.

Hierarchical Attributes

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

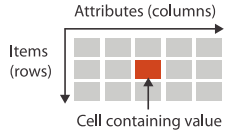
Semantics

- Key versus Value Semantics
- Temporal Semantics

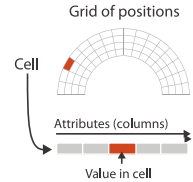
Key versus Value Semantics

- Knowing the type of an attribute does not tell us about its semantics
- A **key** attribute acts as an index that is used to look up value attributes
- The distinction between key and value attributes is important for the dataset types of tables and fields

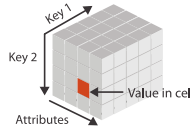
➔ Tables



➔ Fields (Continuous)



➔ *Multidimensional Table*



Flat Tables

- A simple **flat table** has only one key, where each item corresponds to a row in the table, and any number of value attributes.
- In this case, the key might be completely implicit, where it's simply the index of the row.
- In tables, keys may be categorical or ordinal attributes, but quantititive attributes are typically unsuitable as keys because there is nothing to prevent them from having the same values for multiple items.

Flat Tables (cont.)

- The order table with the attribute columns colored by their type; none of them is a key.

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

Multidimensional Tables

- The more complex case is a **multidimensional table**, where multiple keys are required to look up an item.
- The combination of all keys must be unique for each item, even though an individual key attribute may contain duplicates.

Fields

- **Fields** are typically characterized in terms of the number of keys versus values.
- Their **multivariate** structure depends on the number of value attributes, and their **multidimensional** structure depends on the number of keys.

Scalar Fields

- A **scalar field** is univariate, with a single value attribute at each point in space.
- One example of a 3D scalar field is the time-varying medical scan

Vector Fields

- A **vector field** is multivariate, with a list of multiple attribute values at each point
- The geometric intuition is that each point in a vector field has a direction and magnitude, like an arrow that can point in any direction and that can be any length

Tensor Fields

- A **tensor field** has an array of attributes at each point, representing a more complex multivariate mathematical structure than the list of numbers in a vector
- A physical example is stress, which in the case of a 3D field can be defined by nine numbers that represent forces acting in three orthogonal directions

Field Semantics

- This categorization of spatial fields requires knowledge of the attribute semantics and cannot be determined from type information alone.

Time-Varying Data

- A **temporal** attribute is simply any kind of information that relates to time.
- Data about time is complicated to handle because of the rich hierarchical structure that we use to reason about time, and the potential for periodic structure
- A dataset has **time-varying** semantics when time is one of the key attributes, as opposed to when the temporal attribute is a value rather than a key
- A common case of temporal data occurs in a **time-series** dataset, namely, an ordered sequence of time–value pairs

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.