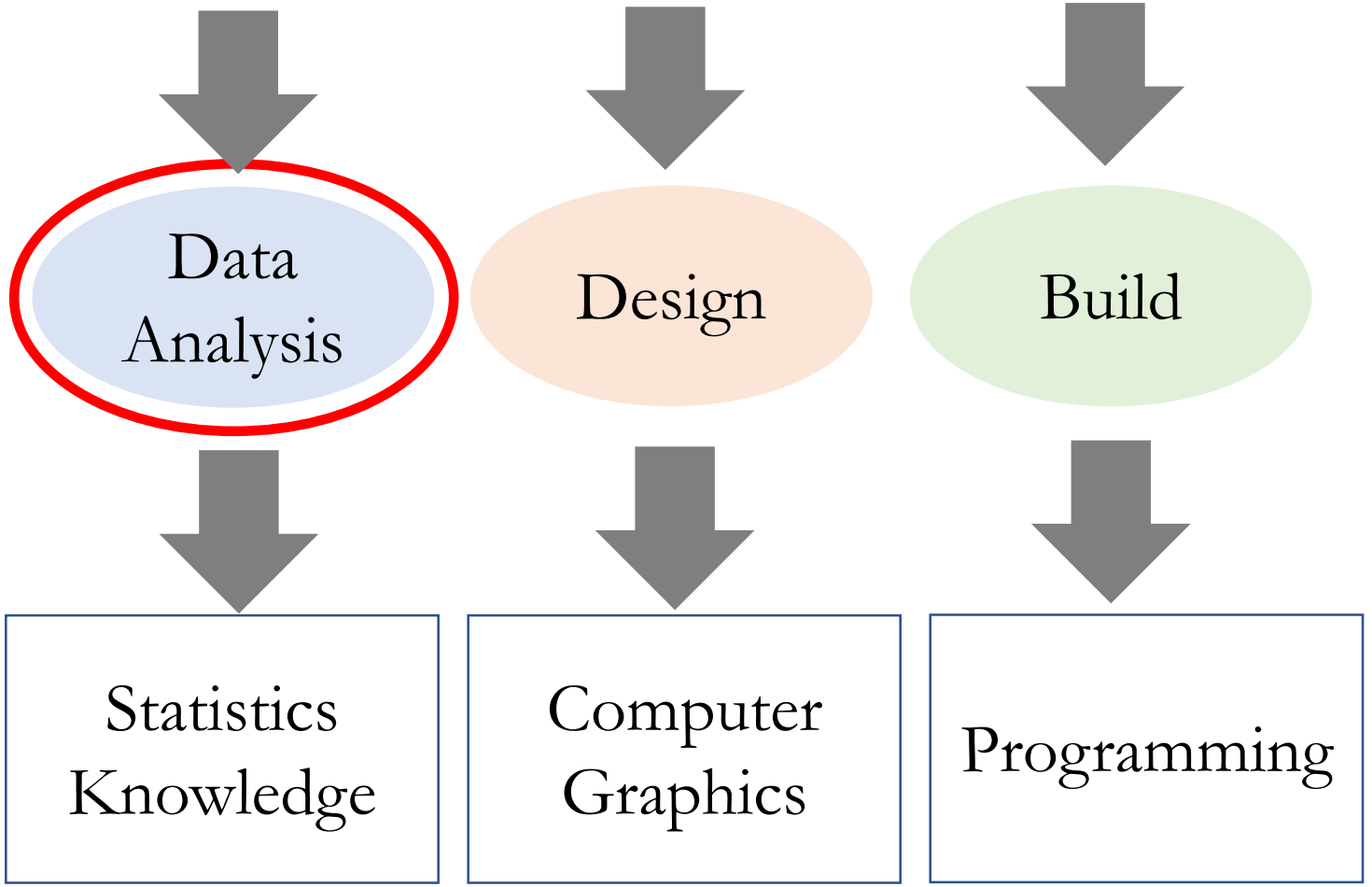
Data Visualization



Data types and data analysis

Aihua Li

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

Table 2.1. A full table with column titles that provide the intended semantics of the attributes.

Data types and data sources

(Tamara Munzner, 2000)

- Dataset Types: What can be visualized?
- Data Type: Fundamental unit combinations make up datasets type

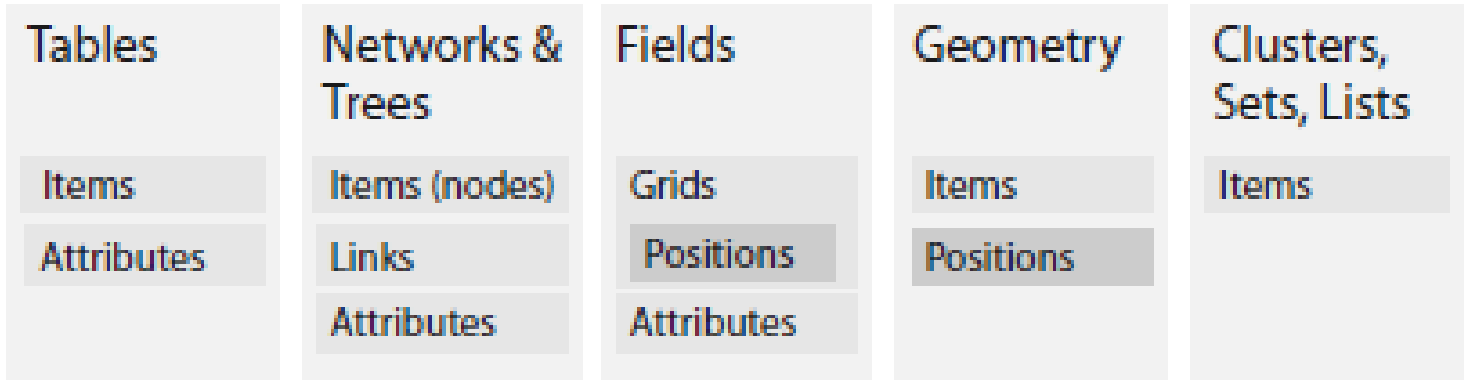
Figure 2.1. *What* can be visualized: data, datasets, and attributes.

Data Types

→ Items → Attributes → Links → Positions → Grids

- Five basic **data types**: items, attributes, links, positions, and grids.
- An **attribute** 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. 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 or three numbers specifying a location within the region of space measured by a medical scanner.

Data and 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.

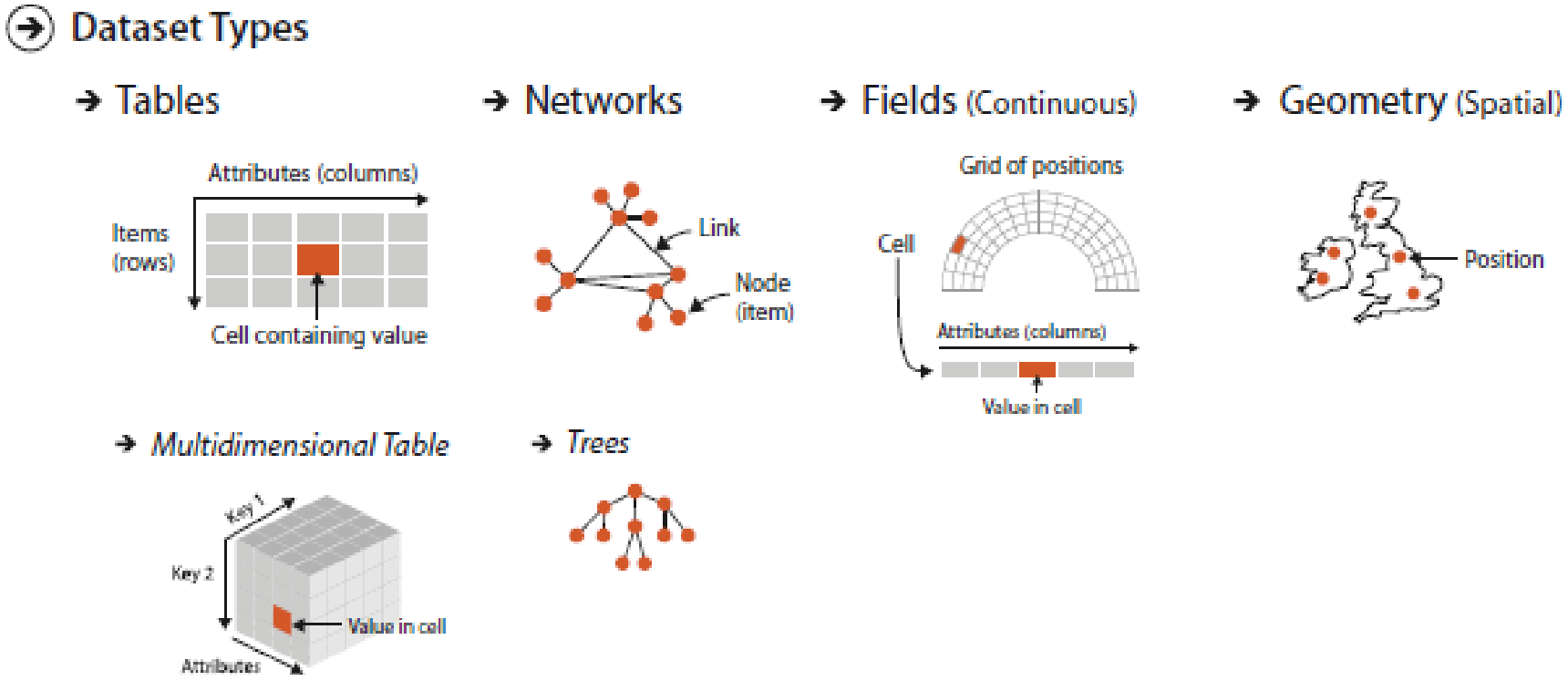


Figure 2.4. The detailed structure of the four basic dataset types.

- 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.

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		5-Low	Wrap Bag	0.56	1/20/05
69		4-Not Specified	Small Pack	0.44	6/6/05
69		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

Figure 2.5. In a simple table of orders, a row represents an *item*, a column represents an *attribute*, and their intersection is the *cell* containing the value for that pairwise combination.

Data Analysis (DCSI607, DCSI602)

- Basic Statistical Analysis: Mean, Max, Median, Min, Variance, and Standard Deviation

- Time Series Analysis:

- Spatial pattern Analysis:

- Basic Statistical Analysis: Sum, Mean, Max, Median, Min, Variance, and Standard Deviation

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

Table 2.1. A full table with column titles that provide the intended semantics of the attributes.

X=c(7,8,8,9,9,10,10,12,12,13)

Median: middle point value, (9+10)/2=9.5

Mean: the average value of all these numbers, 9.8

$$\mu = \frac{\sum X}{N} = \frac{98}{10} = 9.8$$

Min: 7; Max=13,

In R, using functions like:

median()
mean()
min()
max()
sum()

- The sample variance is the variability measured relative to the arithmetic average of n data values x_i comprising a sample, $\bar{X}$ is the mean of the sample

$$\text{var}(X) = s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

- This is the average of the square of the deviations from the sample mean. Alternatively,

$$\text{var}(X) = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

where $n - 1$ is used to account for the fact that the sample mean was already estimated from the n values.

7,8,8,9,9,10,10,12,12,13

- The sample variance is s_X^2= 3.96
- The sample standard deviation is s_X =1.99

var(X)

sd(X)