CSI3005

# Advanced Data Visualization Techniques

# Module 1

**Introduction to Data Visualization**

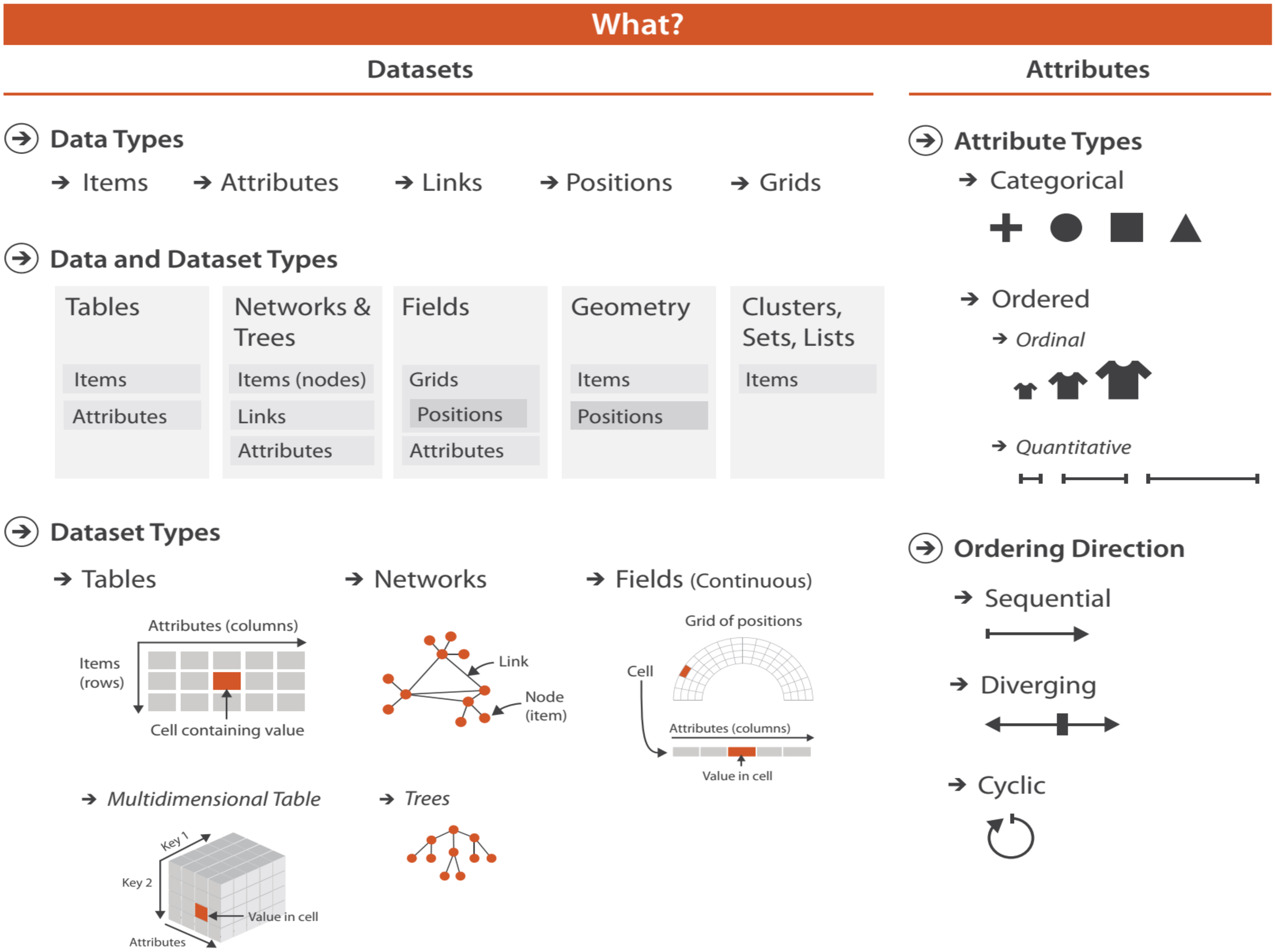
- ❑ Overview of data visualization
- ❑ **Data Abstraction**
- ❑ Task Abstraction
- ❑ Analysis: Four Levels for Validation

**Text Book**

Tamara Munzer**, Visualization Analysis and Design** -, CRC Press 2014 . **(Chapter 1, 2,3 and 4)**

# Data Abstraction

# Data abstraction



## 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*

### Attributes

**Attribute Types**

- Categorical
- Ordered
  - *Ordinal*
  - *Quantitative*

**Ordering Direction**

- Sequential
- Diverging
- Cyclic

# Data abstraction

- Geometry (Spatial)

## Dataset Availability

- Static

- Dynamic

What?

Why?

How?

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

# Data abstraction

- This figure shows the abstract types of *what* can be visualized.

- The four basic dataset types are **tables, networks, fields, and geometry**; other possible collections of items include clusters, sets, and lists.
- These 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 in the form of a static file, or it might be dynamic data processed gradually in the form of a stream.

- 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.

## Why Do Data Semantics and Types Matter?

- ❑ 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?

Suppose that you see the following data:

**14, 2.6, 30, 30, 15, 100001**

❖ What does this sequence of six numbers mean?

Similarly, suppose that you see the following data:

**Basil, 7, S, Pear**

❖ These numbers and words could have many possible meanings.

- To know about the data, two crosscutting pieces of information are required. Theses are:
    - Semantics of data
    - Types of data.

- The **semantics** of the data is its real-world meaning.
- For instance, does a word represent a human first name,
- or !!!!!!!!!!!
- is it the shortened version of a company name where the full name can be looked up in an external list,
- or !!!!!!!!!
- is it a city,
- or !!!!!!!!!!! is it a fruit?

- The **type** of the data is its structural or mathematical interpretation.
- Two levels:
    - At the data level, what kind of thing is it: an item, a link, an attribute?
    - At the attribute level: what kinds of mathematical operations are meaningful for it?

- For example: if a number represents a count of boxes of detergent, then its type is a quantity, and adding two such numbers together makes sense.

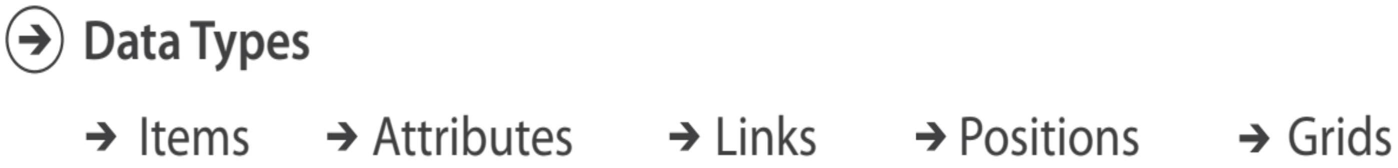
- If the number represents a postal code, then its type is a code rather than a quantity—it is simply the name for a category that happens to be a number rather than a textual name.

- Adding two of these numbers together does not make sense.

- Meta data:
  - Additional (textual information) information of the original dataset 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 | Desm | 13 | L | Elderberry |
| 5 | Ernest | 12 | L | Peach |
| 6 | Fanny | 10 | S | Lychee |
| 7 | Geore | 9 | M | Orange |
| 8 | Hect | 8 | L | Loquat |
| 9 | Ida | 10 | M | Pear |
| 10 | Amy | 12 | M | Orange |

# Data types



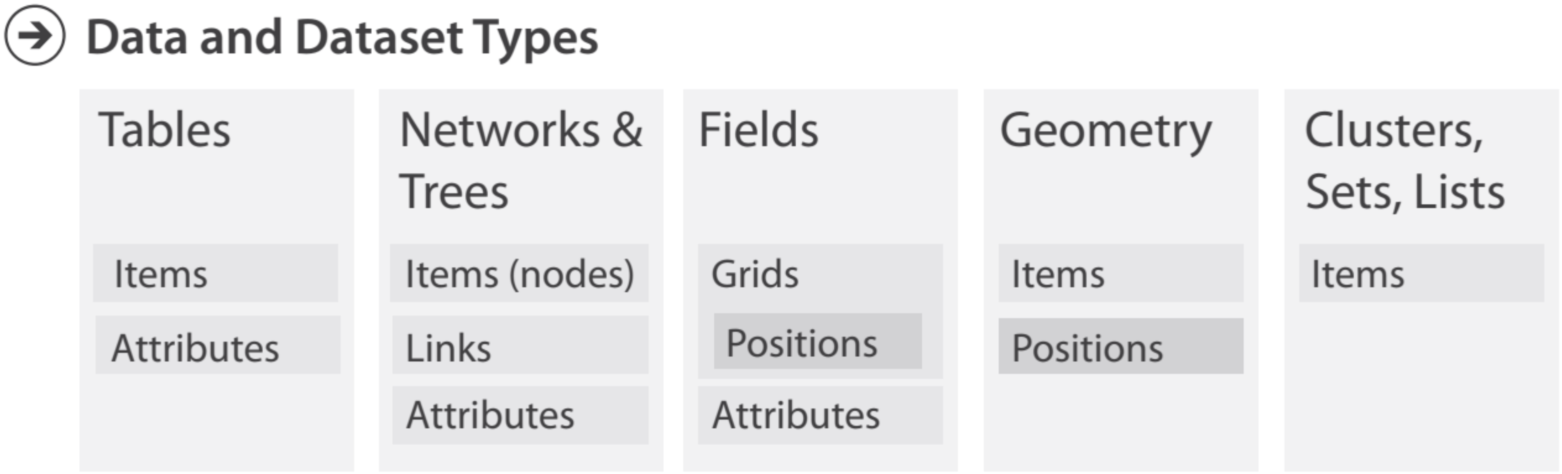
Data Types

→ Items → Attributes → Links → Positions → Grids

- An **attribute** is some specific property that can be measured, observed, or logged.
  - For Synonyms for *attribute* are **variable** and **data dimension**, or just **dimension** for short.
  - Example: attributes could be salary, price, number of sales, protein expression levels, or temperature, weather data.
- 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.

# 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.



Data and Dataset Types

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

**Figure 2.3.** The four basic dataset types are tables, networks, fields, and geometry; other possible collections of items are clusters, sets, and lists. These datasets are made up of five core data types: items, attributes, links, positions, and grids.

# Dataset types

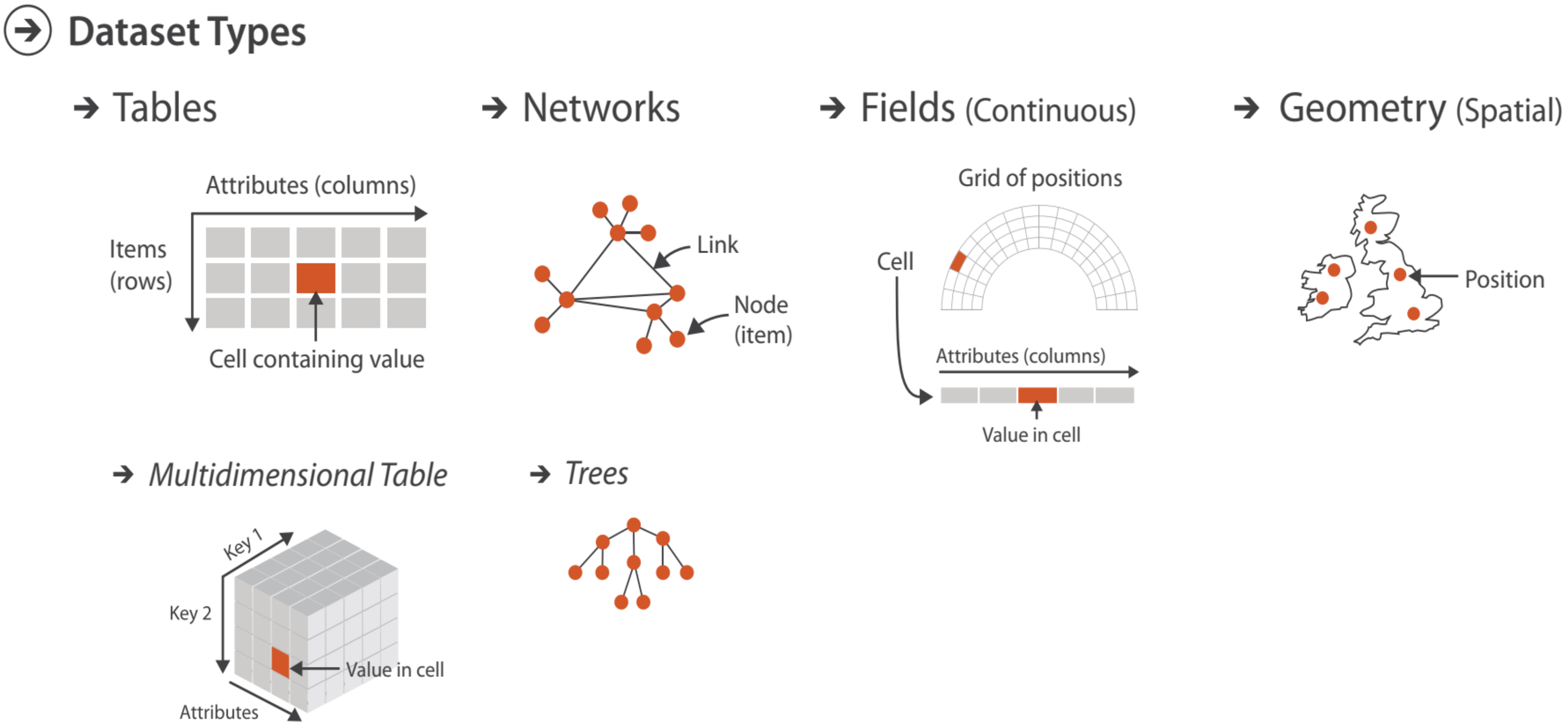


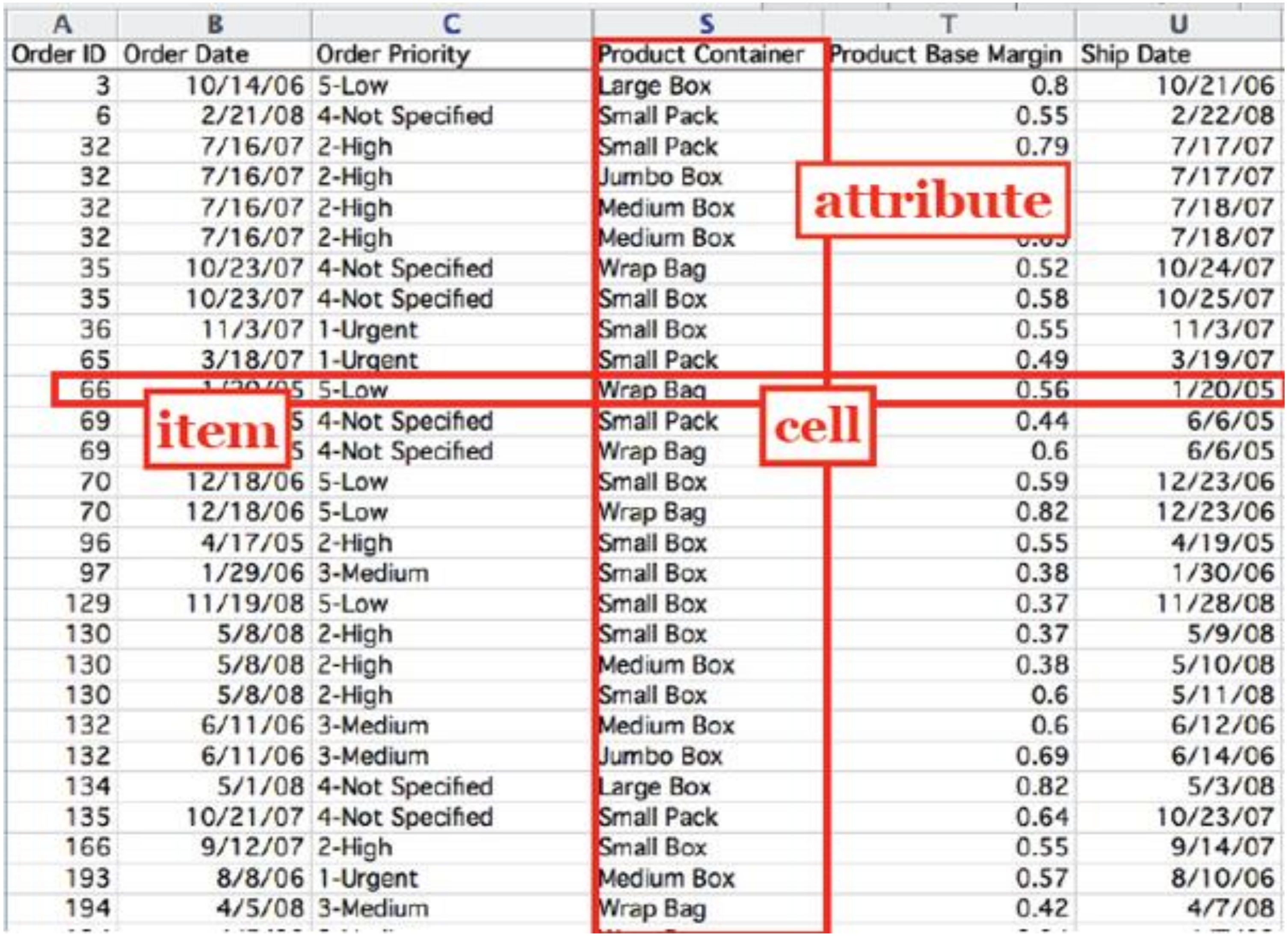
Figure 2.4. The detailed structure of the four basic dataset types.

**Tables:** made up of rows and columns: spreadsheet.

- **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.



| 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 | 1/20/05 | 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 |

**Networks: it** is well suited for specifying that there is some kind of relationship between two or more items.

- An item in a network is known as **node**
- A **link** is a relation between two items

- For example, in an articulated social network the nodes are people, and links mean friendship.
- In a gene interaction network, the nodes are genes, and links between them that these genes have been observed to interact with each other.

- Networks with hierarchical structure are called **trees**.

- Note: trees do not have cycles: each child node has only one parent node pointing to it.

**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**

  - **Sampling**: how frequently to take the measurements, and
  - **Interpolation:** how to show values in between the sampled points in a way that does not mislead.

- 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.
- Most datasets that contain inherently spatial data occur in the context of tasks that require understanding aspects of its spatial structure, especially shape
- Grids – Uniform Grid, Unstructured grid

- 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, and like spatial fields they typically occur in the context of tasks that require shape understanding

## Other Combinations

- Set
- Lists
- Cluster
- Path

# Data availability

The two kinds of dataset availability: *static* or *dynamic*.

- the entire dataset is available all at once, as a **static file**.

- Some datasets are available in **dynamic streams:** One kind of dynamic change is to add new items or delete previous items.

# Attribute types

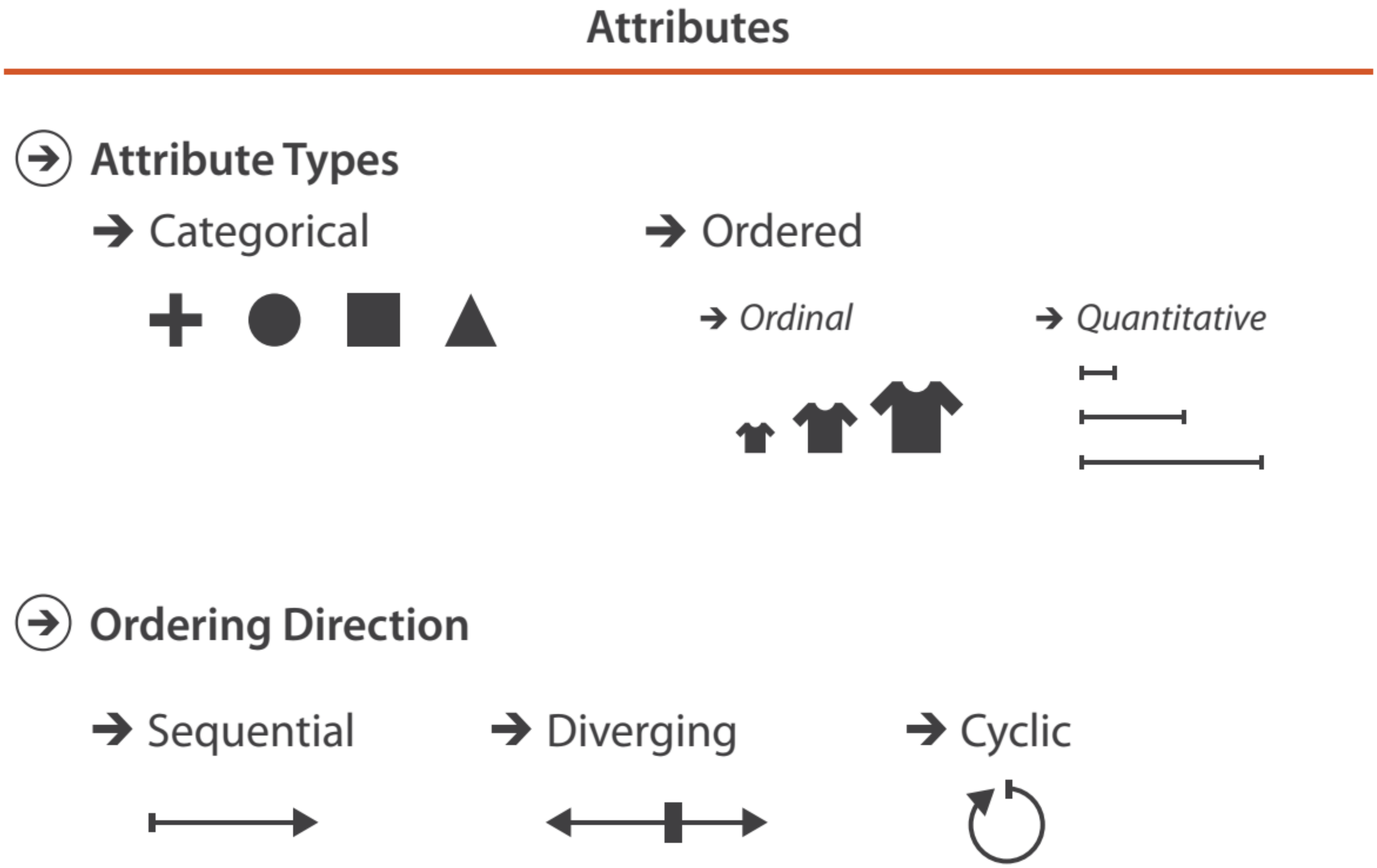


**Figure 2.7.** Attribute types are categorical, ordinal, or quantitative. The direction of attribute ordering can be sequential, diverging, or cyclic.

- The type of **categorical** data, such as favorite fruit or names, doesn’t have an implicit ordering, but it often has hierarchical structure.
- Examples of categorical attributes are fruits (apples, oranges, etc..), movie genres, file types, and city names.

- All **ordered** data does have an implicit ordering, as opposed to unordered *categorical* data.
- This type can be further subdivided such as ordinal and quantitative.
- With **ordinal** data, such as shirt size, we cannot do full-fledged arithmetic, but there is a well-defined ordering. For example, large minus medium is not a meaningful concept, but we know that medium falls between small and large.

- 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, and the answer of 26 inches can be calculated.
- Other examples of quantitative data are height, weight, temperature, stock price, number of calling functions in a program, and number of drinks sold at a coffee shop in a day.

# Attribute Semantics

- Key vs. value semantics
- The key attribute acts as an index to retrieve the data value
- Different data set types will have different ways to define the keys

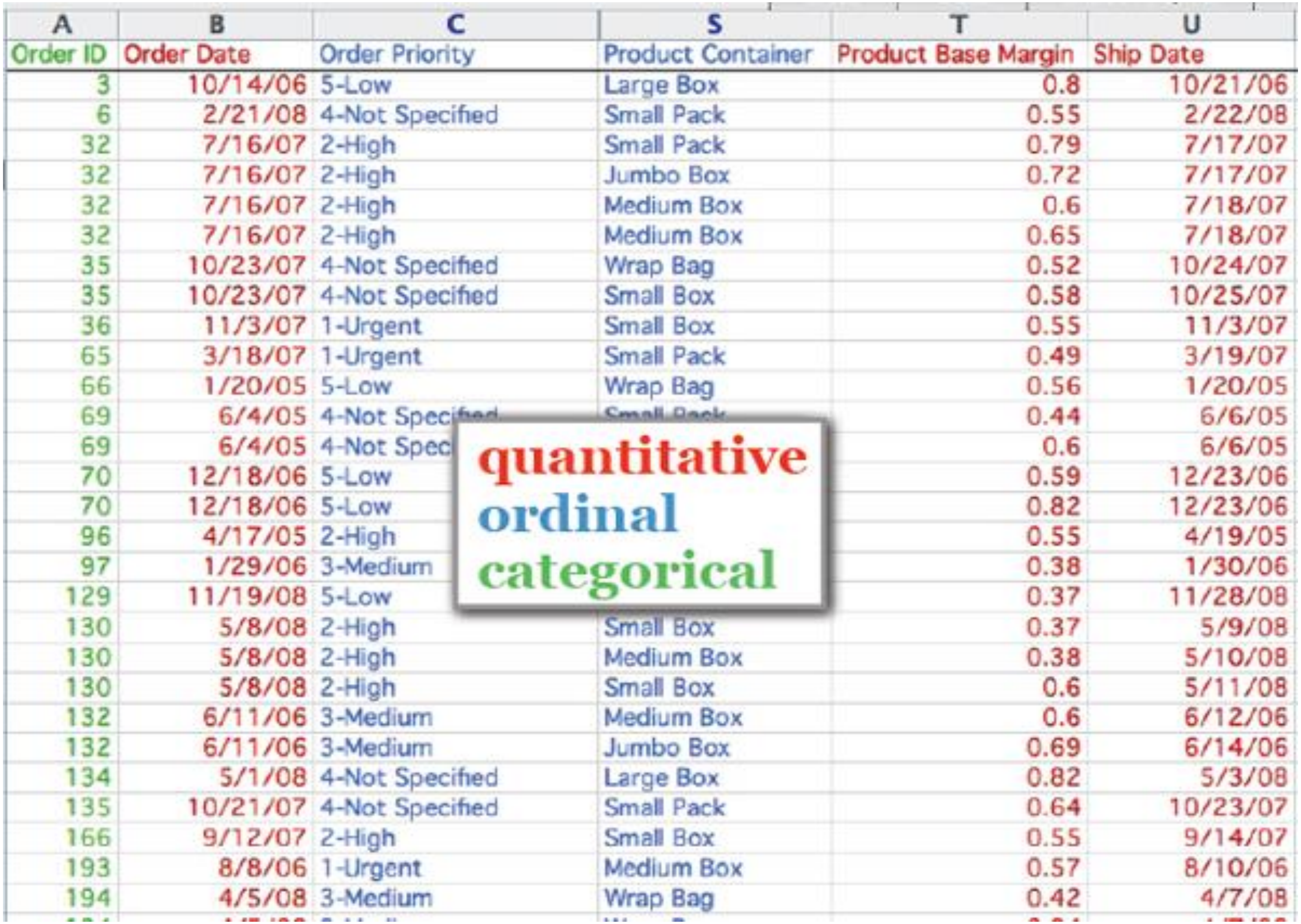
# Flat Table

An item

| 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 |

Can be used as a key

May not be a good choice of 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 |

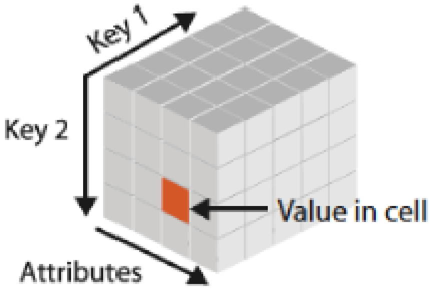
quantitative

ordinal

categorical

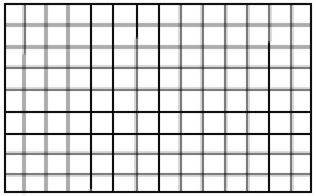
# Multi-dimensional Tables

- A key has multiple attributes and needs to be a unique combination of values
- It is not always clear what attributes are keys and what are values
  - Figuring out independent and dependent variables (cause-effect analysis)

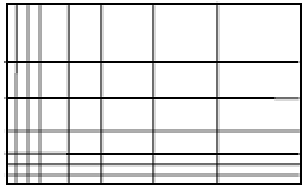


# Field Data

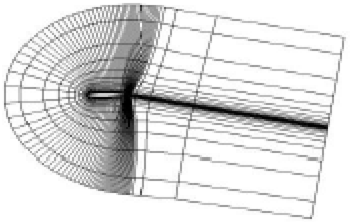
- Field data are mostly seen in scientific applications (temperatures, pressures, etc)
- Values are defined on grids, where the positions of the grid points are the key



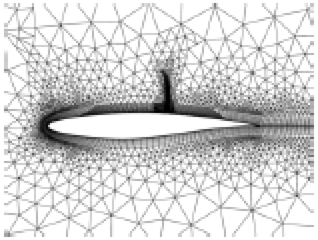
Cartesian Grid



Rectilinear Grid



Curvilinear Grid

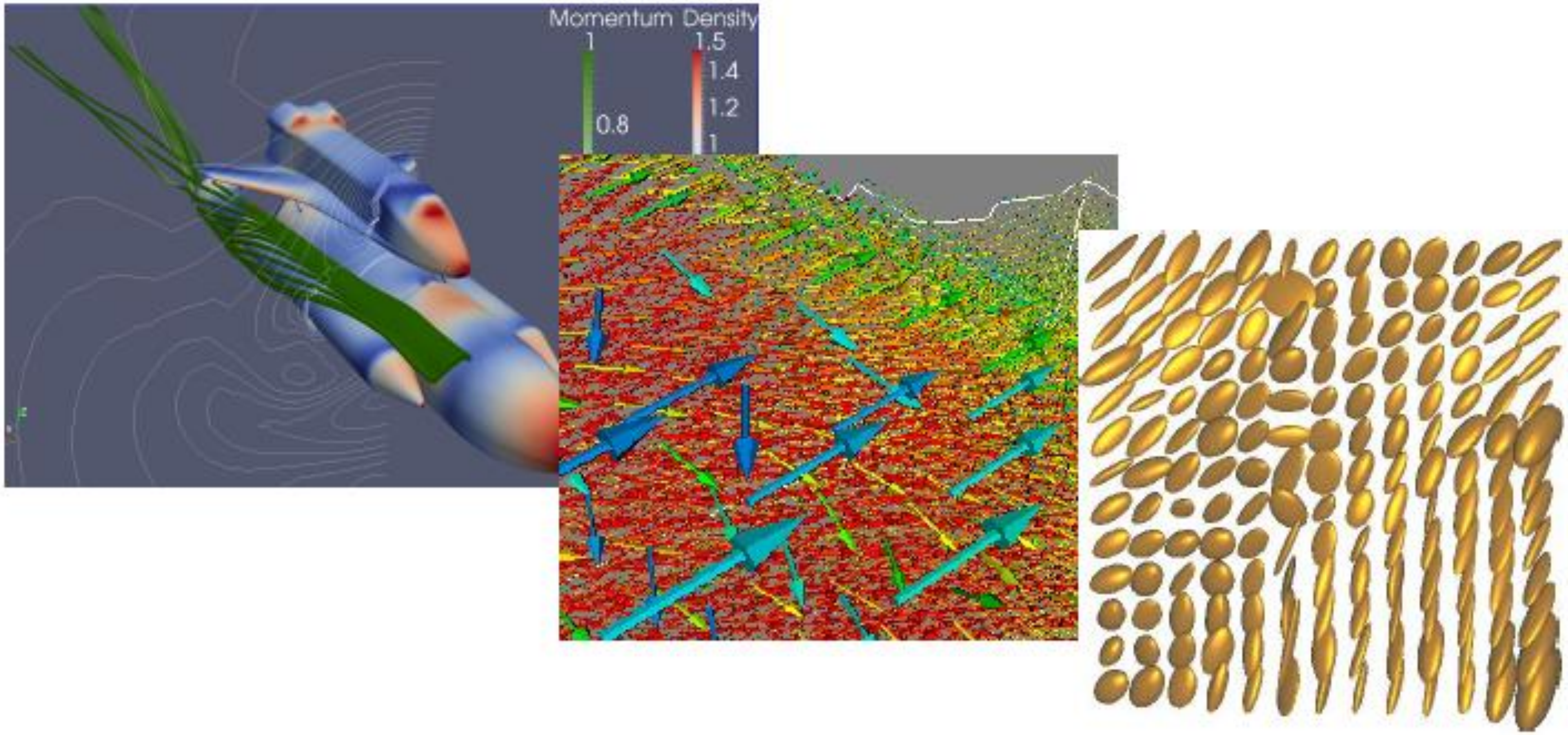


Irregular Grid

- Value attributes: scalar, vector, tensor

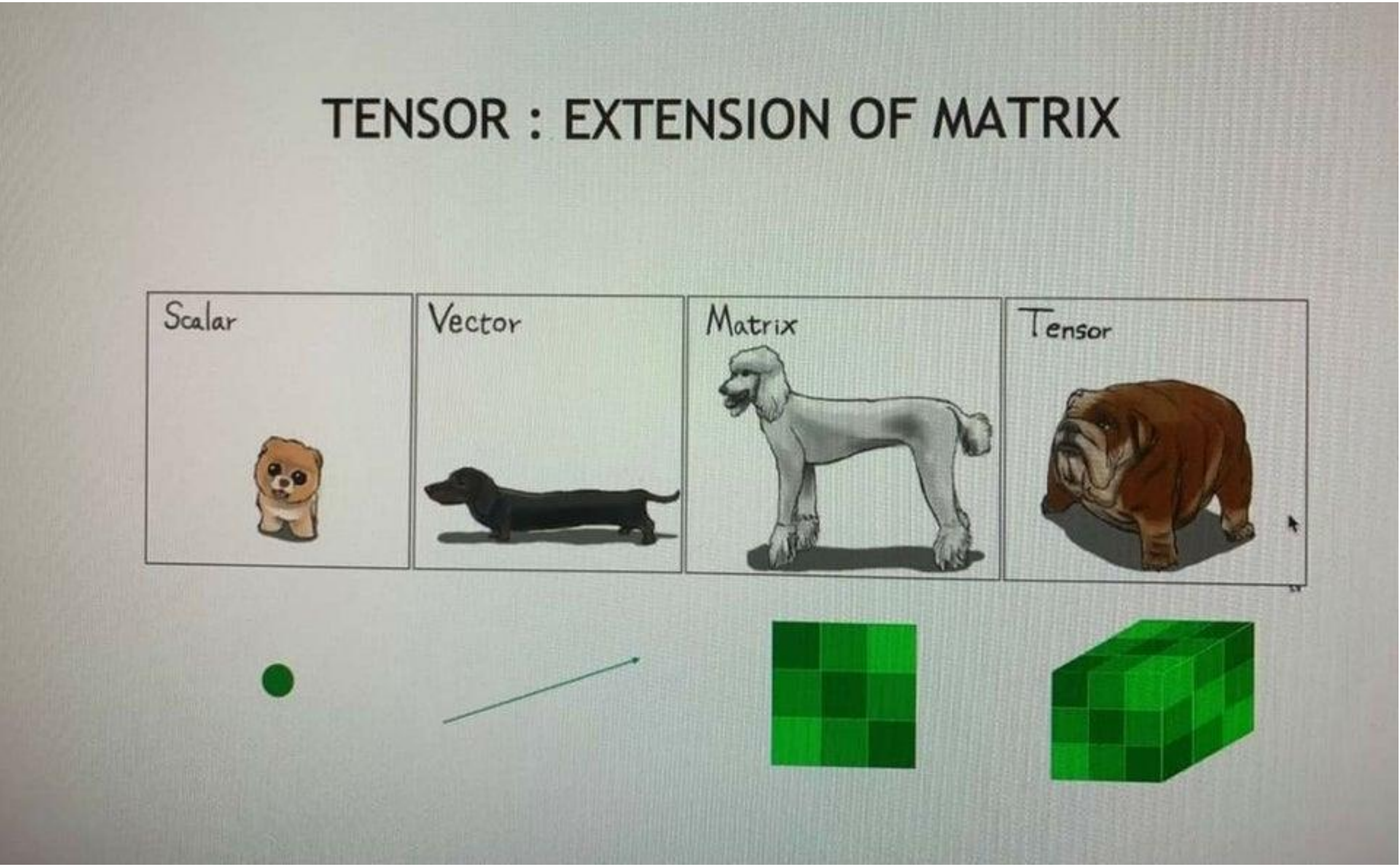
# Attributes

- Scalars (e.g. density), Vectors (e.g. momentum), , Tensors (e.g. stress tensor)



Scalar Vector Matrix Tensor

$$1 \qquad \begin{bmatrix} 1 \\ 2 \end{bmatrix} \qquad \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \qquad \begin{bmatrix} \begin{bmatrix} 1 & 2 \end{bmatrix} & \begin{bmatrix} 3 & 2 \end{bmatrix} \\ \begin{bmatrix} 1 & 7 \end{bmatrix} & \begin{bmatrix} 5 & 4 \end{bmatrix} \end{bmatrix}$$



# TENSOR : EXTENSION OF MATRIX

# Temporal Semantics

- Any kind of information that is related to time
- Temporal data are often more complex to deal with
- Temporal attributes can be either keys or values
- Time-varying data often means time is the key attribute
  - e.g  Time series data