Chapter 3

What: Data Abstraction

The Big Picture

- **Four basic dataset type**
  - **Tables, networks, fields, and geometry**
  - **Other possible collections of items**
    - **Clusters, sets, lists**
- **Datasets are made up of different combinations of the five data types**
  - **Items, attributes, links, positions, grids**
  - **Attribute type**
    - **Categorical**
    - **Ordered**
      - **Ordinal and quantitative**
      - **ordering direction: sequential, diverging, cyclic**
- **Datasets can be static or dynamic (stream)**

What?

Datasets

Data Types

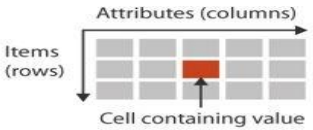
➔ Items ➔ Attributes ➔ Links ➔ Positions ➔ Grids

Data and Dataset Types

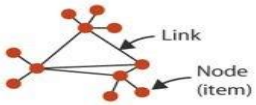
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

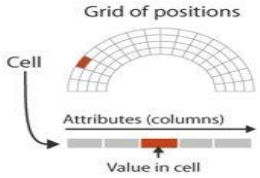
➔ Tables



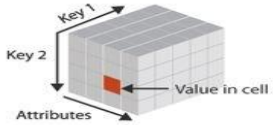
➔ Networks



➔ Fields (Continuous)

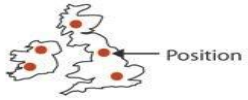


➔ *Multidimensional Table*



➔ *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 vis design are driven by the kind of data that you have at your proposal**
  - **What kind of data are you given?**
  - **What information can you figure out from the data?**
  - **What high-level concepts will allows you to split datasets apart into general and useful pieces?**
- **To move beyond guesses, we need to know their semantics and types**
  - **Semantics of the data: its real-world meaning**
  - **Types of the data: its structural or mathematical interpretation**

Why Do Data Semantics and Types Matter?

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

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

Five Basic Data Types

- **Attribute**
  - **Some specific property that can be measured, observed, or logged.**
- **Item**
  - **An individual entity that is discrete, such as a row in a simple table or a node in a network.**
- **Link**
  - **A relationship between items, typically within a network.**
- **Grid**
  - **Specifying the strategy for sampling continuous data in terms of both geometric and topological relationships between its cells.**
- **Position**
  - **It is spatial data, providing a location in two-dimensional or three-dimensional space.**

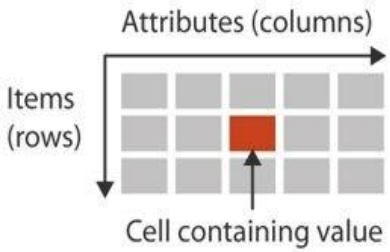
Dataset Types

- A dataset is any collection of information that is the target of analysis
- The four basic types
  - tables, networks, fields, and geometry
- Other ways to group items together
  - clusters, sets, and lists
- In real world, complex combinations of these basic types are common

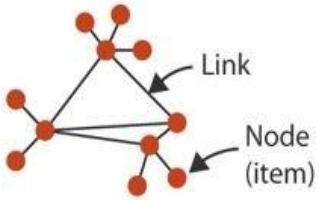
Dataset Types

Dataset Types

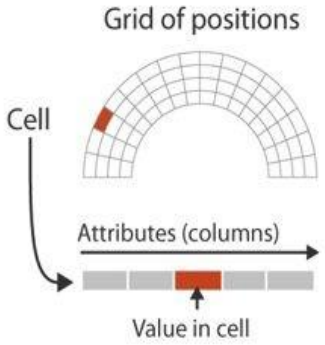
Tables



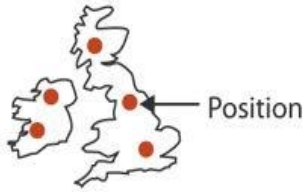
Networks



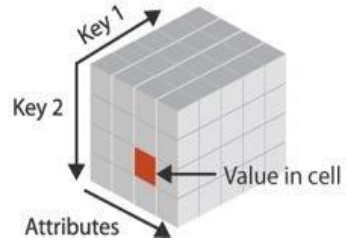
Fields (Continuous)



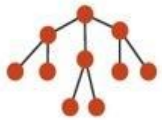
Geometry (Spatial)



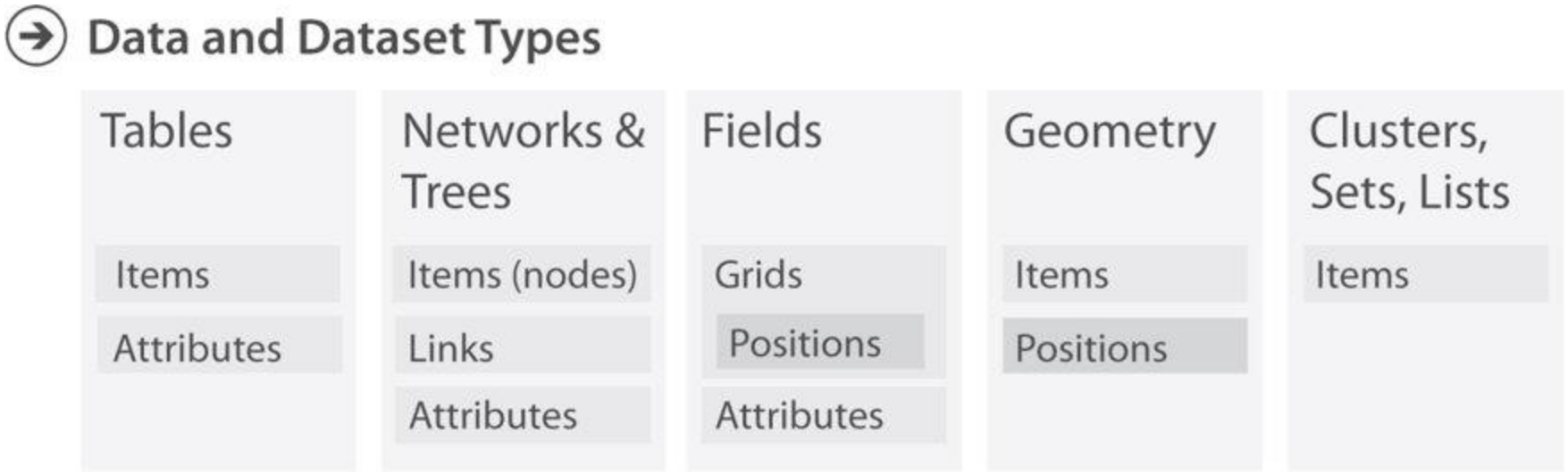
Multidimensional Table



Trees



Dataset Types



Datasets are made up of five basic data types

Dataset Types
Tables

- Tables are made up of rows and columns
- Flat table
  - Each row represents an item of data
  - Each column is an attribute of the dataset
  - Cell
    - specified by a row and a column, contains a value
- Multidimensional table
  - Has a more complex structure for indexing into a cell, with multiple keys

Dataset Types
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]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

Dataset Types
Networks and Trees

- **Networks are some kind of relationship between two or more items.**
  - **An item in a network is often called a node**
    - **Nodes can have associated attributes, just like item in a table**
  - **A link is a relation between two nodes**
    - **Links could have associated attributes; may be partly or wholly disjoint from the node attributes**
- **Trees: Networks with hierarchical structure**
  - **Trees do not have cycles**

Dataset Types
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**
    - **Ex. Temperature, pressure, speed, force, density**
- **Continuous data requires careful treatment that consider questions**
  - **Sampling**
    - **How frequent to take the measurements**
  - **Interpolation**
    - **How to show values in between the samples points in a way that does not mislead**

Dataset Types
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**
  - **Most tasks of the datasets aim to understand its spatial structure, especially shape**
  - **Ex. Medical imaging, CT or MRI**
    - **Locate suspected tumors that can be recognized through distinctive shapes or densities**

Dataset Types
Fields: Grid Types

- **Uniform grid: axis-aligned domain**
- **Rectilinear grid**
  - **Axis-aligned domain, same topology, Non-uniform sampling**
  - **Store location of each row**
- **Structured grid**
  - **Can be seen as a deformation of uniform grids**
    - **Topology stays the same**
    - **Need to store spatial positions which vary freely**
- **Unstructured grid, must store**
  - **Topological information, Spatial positions**

Dataset Types
Geometry

- **Specify information about the shape of items with explicit spatial positions**
  - **The items could be points, line, curves, or 2D surfaces, or 3D volumes**
- **Geometry datasets do not necessarily have attributes**
- **Is interesting in vis only when it is derived or transformed in a way that requires design choice consideration**
  - **Ex. When contours are derived from a spatial field**

Dataset Types
Other Combinations

Other ways to group multiple items together

- **Set**
  - **Simply an unordered group of items**
- **List**
  - **Grouping items with a specified ordering**
- **Cluster**
  - **A grouping based on attribute similarity**

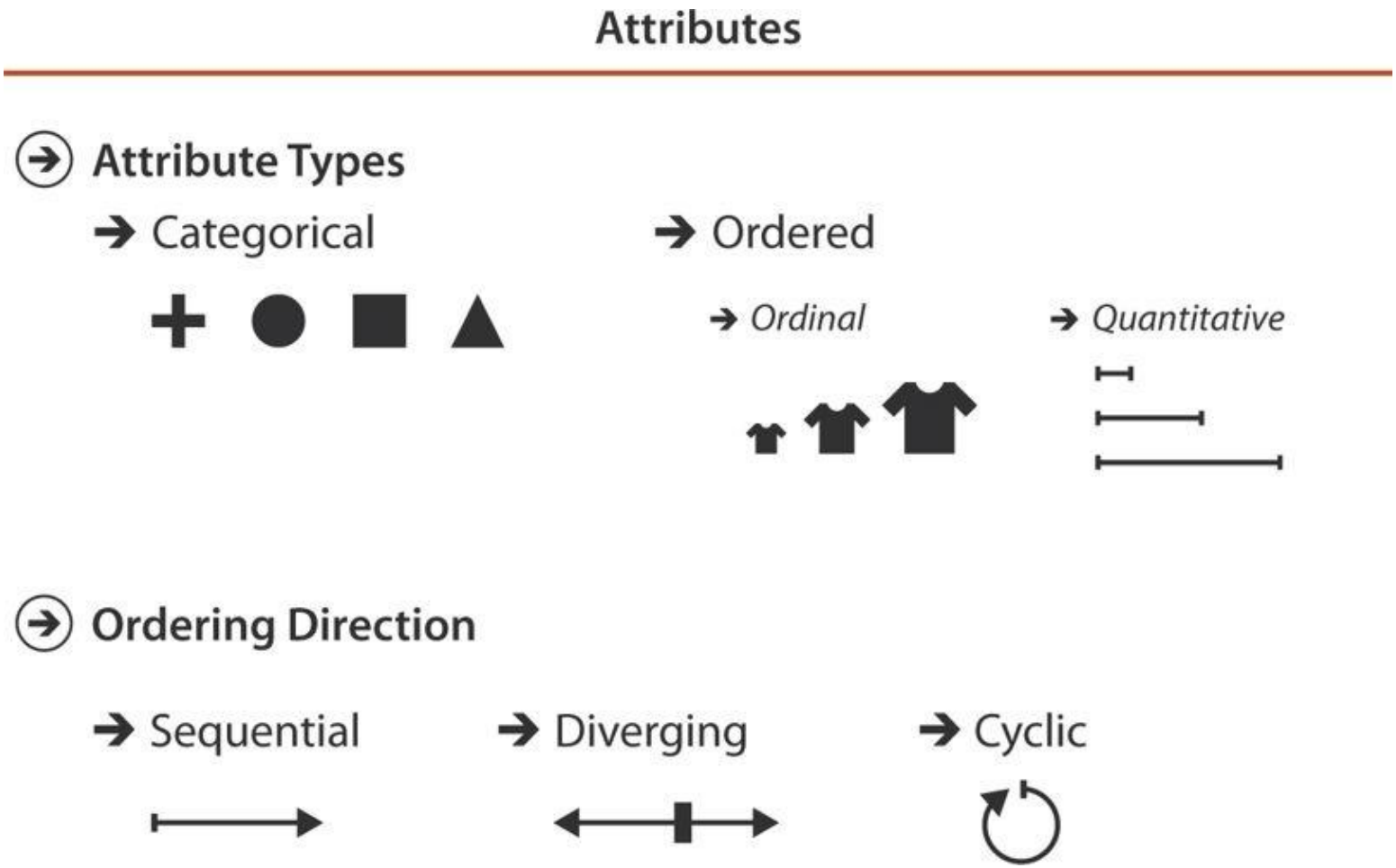
Dataset Types
Other Combinations

- **Path**
  - **It is an ordered set of segments formed by links connecting nodes.**
- **Compound network**
  - **It is a network with an associated tree.**
  - **All of the nodes in the network are the leaves of the tree.**
  - **Interior nodes in the tree provide a hierarchical structure for the nodes that is different from network links between them.**

Dataset Types
Dataset Availability

- **Static**
  - **Static file or Offline**
  - **Entire datasets are available all at once**
- **Dynamic**
  - **Dynamic stream or Online**
  - **Dataset information trickles in over the course of vis session**
    - **Add new items or delete previous items**
    - **Change the values of existing items**

Attribute Types



Attribute Types

- **Categorical**
  - **Does not have an implicit ordering**
  - **Ex. Fruit name, city names**
- **Ordered**
  - **Has an implicit ordering**
  - **Ordinal**
    - **There is a well-defined ordering, cannot do arithmetic**
    - **EX. Skirt size: small, medium, large, X**
  - **Quantitative**
    - **Can do arithmetic**
    - **Ex. Height, weight, temperature…**

Attribute Types

- **Ordering direction**
  - **Sequential**
    - **From a minimum to a maximum**
    - **Ex. Mountain height: from sea level**
  - **Diverging**
    - **Be deconstructed into two sequences pointing in opposite directions**
    - **Ex. A full elevation: values go up for mountains on land and down for undersea valleys**
  - **Cyclic**
    - **Values wrap around back to a starting point**
    - **Ex. Hour of the day, day of the week, month of the year**

Attribute Types

- **Hierarchical attributes**
  - **There may be hierarchical structure within an attribute or between multiple attributes**
- **Examples**
  - **Time-series dataset**
    - **Time can be aggregated hierarchically**
      - **From days up to weeks, up to months, up to years**
  - **Geographic data**
    - **Postal code can be aggregated up to the city level or state or entire country**

Semantics

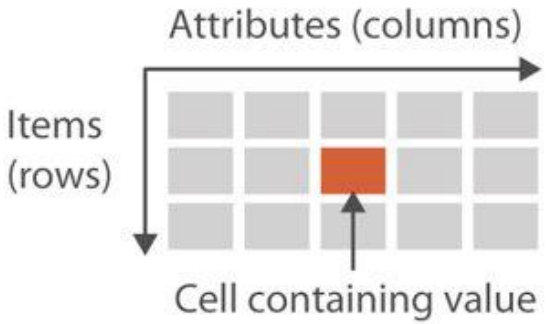
- **Knowing the type of an attribute does not tell us about its semantics**
- **Classification**
  - **Key vs. value semantics**
  - **Spatial/continuous data vs. non-spatial/discrete data**
  - **Temporal**

Semantics
Key vs. Value semantics

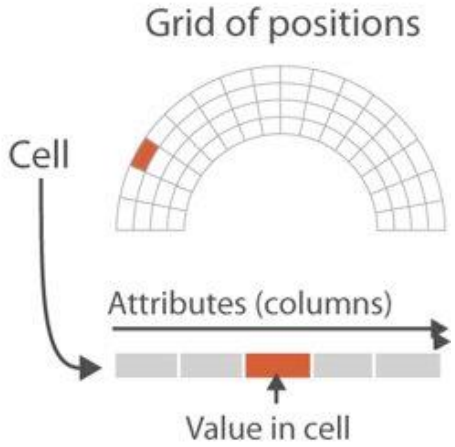
- **A key attribute acts as an index used to look up value attributes**
  - **Important for dataset types of tables and fields**
  - **Key might be**
    - **Completely implicit, simply the index of row**
    - **Explicit, contained within the table as an attribute**
      - **There must be no duplicate values**
  - **In table, keys may be categorical or ordinal attributes**

Semantics
Key vs. Value semantics

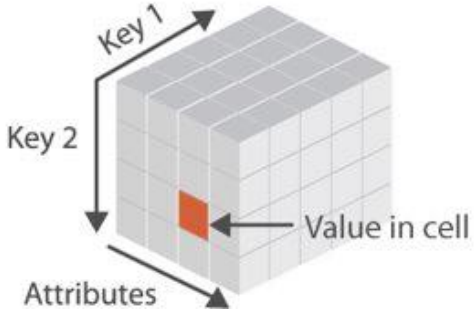
➔ Tables



➔ Fields (Continuous)



➔ *Multidimensional Table*



Semantics
Flat Tables

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

ID: key

Semantics
Flat 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	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

No explicit key: even Order ID has duplicates.
An implicit key: row number

Semantics
Multidimensional Tables

- **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.**

Semantics
Fields

- **In spatial fields, spatial location acts as a quantitative key**
- **Fields are typically characterized in terms of the number of keys versus values.**
  - **Multivariate structure depends on the number of value attributes**
  - **Multidimensional structure depends on the number of keys**
    - **2D, 3D**
    - **2D, 3D + 1: time-varying**

Semantics
Fields

- **Scalar Fields**
  - **A scalar field is univariate, with a single value attribute at each point in space.**
- **Vector Fields**
  - **A vector field is multivariate, with a list of multiple attribute values at each point.**
- **Tensor Fields**
  - **A tensor field has an array of attributes at each point, representing a more complex multivariate mathematical structure.**

Semantics
Temporal Semantics

- **A temporal attribute is simply any kind of information that relates to time.**
- **Temporal data is complicated to handle**
  - **Rich hierarchical structure that we use to reason about the time**
    - **Time hierarchy is deeply multiscale**
    - **Temporal scales of interest do not all fit into a strict hierarchy; ex., weeks do not fit cleanly into months**
  - **The potential for periodic structure**
    - **Finding or verifying periodicity either at predetermined scale or at some scale not known in advance**

Semantics
Time-Varying Data

- **A dataset has time-varying semantics when time is one of the key attributes**

- **A common case of temporal data occurs in a time-series dataset**
  - **An ordered sequence of time-value pairs**
  - **A special case of tables, where time is the key**
  - **Typical tasks**
    - **Finding trends, correlations, and variations at multiple time scales such as hourly, daily, weekly, and seasonal**