Information Visualization and Visual Analytics (M1522.000500)

What: Data Abstraction
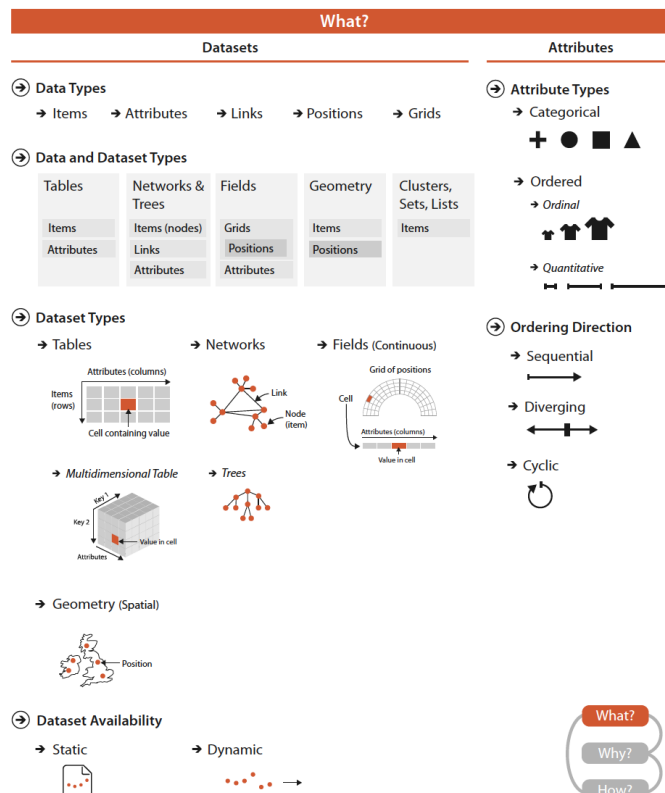
Jinwook Seo, Ph. D.

Professor, Dept. of Computer Science and Engineering

Seoul National University

What?

Why?

How?

Basil, 7, S, Pear

- What does each word mean?
  - → **Semantics**: real-world meaning

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

- Types of the data: structural or mathematical interpretation
  - Data level → Data Types
    - what kind of thing is it?
    - item, attribute, link, position, grid
  - Dataset level → Dataset Types
    - how are these data types combined into a larger structure
    - table, tree, field of sampled values

      Scientific visualization에서 많이 씀, information vis 에서 많이 안씀
  - Attribute level → Attribute Types
    - what kinds of math operations are meaningful for it?
    - attribute: property that can be measured, observed, or logged
      - → variable, dimension
      - Number of detergents: quantity – addition/subtraction
      - Postal code: code – category

Data Types

- **Attribute**: specific property that can be measured, observed, or logged (a.k.a **variable** or **dimension**)

- **Item**: individual entity that is discrete

- **Link**: relationship between items

- **Grid**: specifies the strategy for sampling continuous data in terms of both geometric and topological relationships between its cells

- **Position**: spatial data, providing a location

Dataset Types and Data types

- Dataset
  - collection of info that is the target of analysis
  - arise from combinations of data types

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

Basic Dataset Types

➔ **Tables**

Attributes (columns) →

Items
(rows)

Cell containing value

➔ **Networks**

Link

Node
(item)

➔ **Fields** (Continuous)

Grid of positions

Cell

Attributes (columns)

Value in cell

➔ **Geometry** (Spatial)

Position

➔ *Multidimensional Table*

Key 1

Key 2

Value in cell
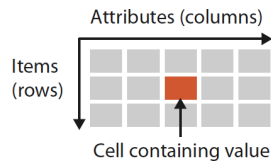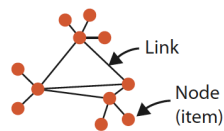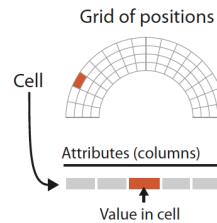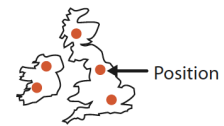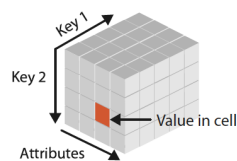
Attributes

➔ *Trees*

Dataset Type: Tables

- item / record / tuple

- attribute / field / variable / dimension

- cell contains value
  - quantitative
  - ordinal
  - nominal

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	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		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

attribute

item

cell

- multidimensional table: indexed with multiple keys

Dataset Type: Network and Trees

- Well suited when there is some kind of relationship b/w items

- **Node**: item
  - Can have associated attributes

- **Link**: relation between two items

- **Trees**: networks with a hierarchical structure
 - Unlike network, there are no cycles

Soccer clubs with players on at least two national teams
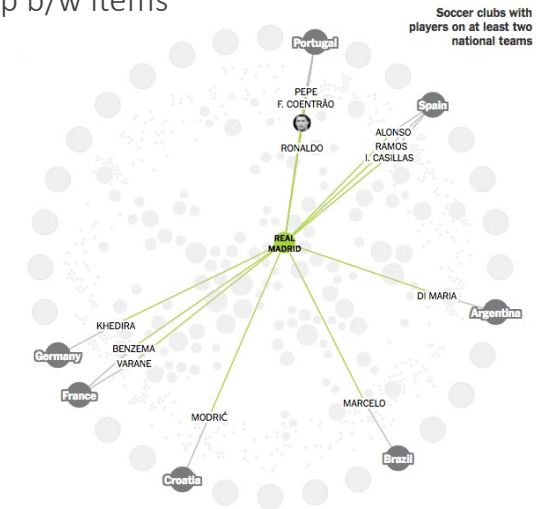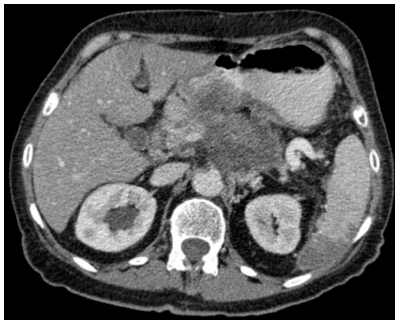
Dataset Type: Fields

- Contains data created by sampling on Grids

- Cell contains measurements from continuous domain
  - sampling, interpolation & reconstruction

- Spatial Fields: cell structure of the field is based on sampling at spatial positions
  - Nonspatial data: abstract data
  - Scivis: spatial position is given with the dataset
    - Handling continuous data appropriately within the math framework of signal processing
  - Infovis: the use of space in visual encoding is chosen by the designer
    - The chosen idiom is suitable for the combination of data and task

Fields (Continuous)

Grid of positions

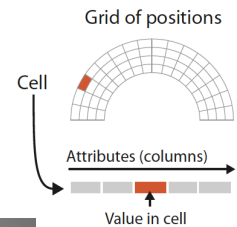Cell

Attributes (columns)

Value in cell
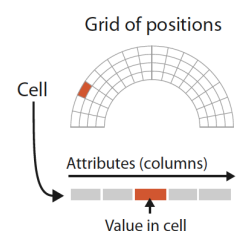
Dataset Type: Fields - Spatial Fields

- Spatial Fields: cell structure of the field is based on sampling at spatial positions
  - Scivis: spatial position is **given** with the dataset
   - Handling continuous data appropriately within the math framework of signal processing
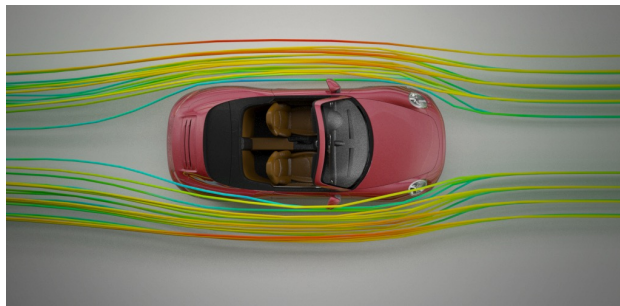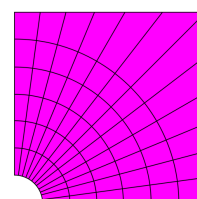
Fields (Continuous)

Dataset Type: Fields - Grid Types

- Grid needs
  - geometry: location in space
  - topology: how each cell connects with its neighboring cells

- Uniform grid: sampling at completely regular intervals
  - No need of grid geometry and topology
- Rectilinear grid: supports nonlinear sampling
  - some info about geometric location of each row
- Structured grid: allows curvilinear shapes
  - geometric location of each cell should be specified
- Unstructured grid: allows complete flexibility
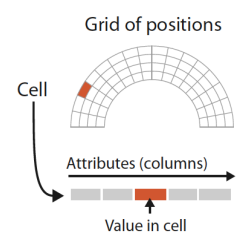 - topological & geometric information required
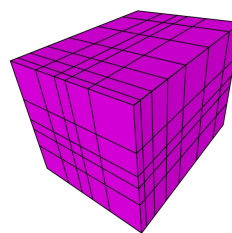
Fields (Continuous)

Rectilinear grid

Curvilinear grid
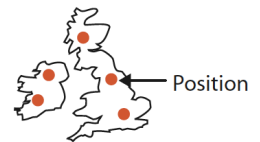
Dataset Type: Geometry

- Specifies info about the shape of items with explicit spatial positions

- Often includes hierarchical structure at multiple scales
    - May be intrinsic or may be derived from the original data

- Not necessarily have attributes

- Simply showing a geometric dataset is not an interesting problem for a vis designer

Geometry (Spatial)

Position

Dataset Type: Combinations

- Set: unordered group of items

- List: ordered group of items

- Cluster: grouping based on attribute similarity

- Path: ordered set + links connecting nodes

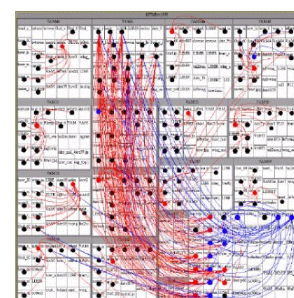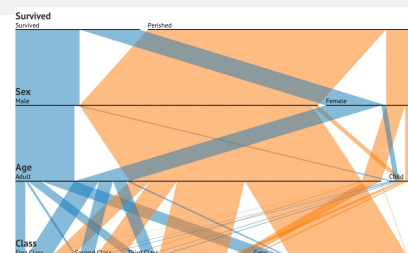- Compound network: network with an associated tree

Figure 1: Directory structure of a Web site visualized as a Treemap with external links overlaid as curves. Blue curves are HTML links, red curves are image links.
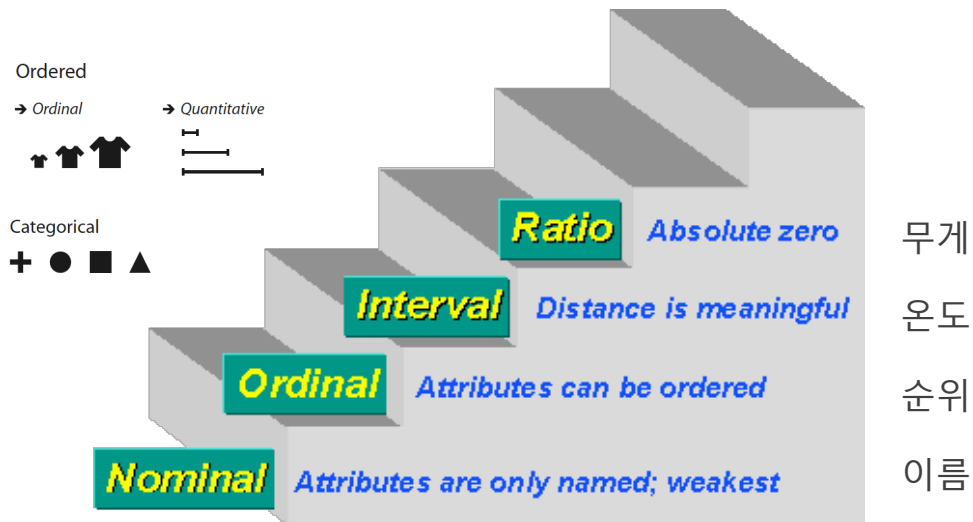
Abstraction and Availability

- Data Abstraction
  - domain-specific to GENERIC
  - translate domain-specific terms into words that are as generic as possible

- Data Availability
  - Static File: available all at once
  - Dynamic Streams

Attribute Types
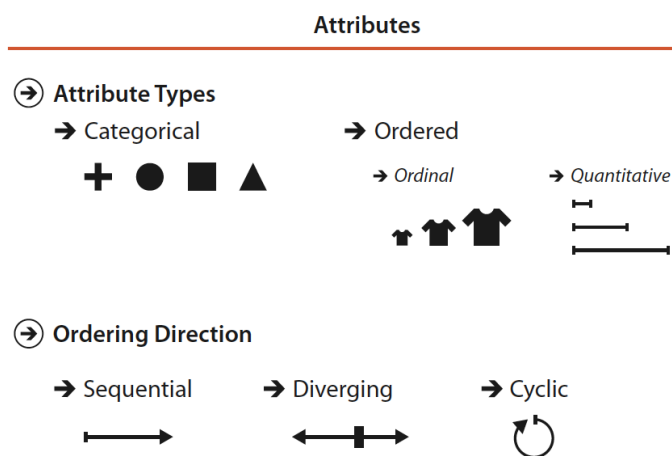
- Hierarchical Attributes: within an attribute or between multiple attributes

Levels of Measurements

Ordered

→ *Ordinal*   → *Quantitative*

Categorical

Ratio Absolute zero

Interval Distance is meaningful

Ordinal Attributes can be ordered

Nominal Attributes are only named; weakest

무게

온도

순위

이름

Semantics - Key vs. Value Semantics

- **Key attribute** acts as an index used to look up **value attributes**

- Flat Tables: only one key

  - Implicit key: keys are simply the index of the row

| 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

구분을 안함 multidimensional/flat

  - Explicit key: keys may be categorical or ordinal attributes (unique)

- Multidimensional Tables: multiple keys required

  - combination of all keys must be unique for each item, even though an individual key may contain duplicates

  - independent keys vs. dependent values

Semantics - Key vs. Value Semantics - Fields   의미가 조금 다름

- Fields differ from tables (Fields are characterized in terms of # of keys vs. values)
  - Each cell represents **continuous** data
  - Key, value → independent, dependent variable (spatial field data)
  - Multivariate structure depends on # of value attributes
  - Multidimensional structure depends on # of key attributes

- Scalar Field: univariate - one attribute per cell (e.g., CT scan)

- Vector Fields: multivariate with a list of multiple attribute values at each point

- Tensor Fields:
  - Representing more complex multivariate mathematical structure with an array of attributes at each point
  - e.g., stress in 3D field: defined by nine numbers that represent forces acting in three orthogonal directions → can be represented as a more complex shape, e.g., ellipsoid.
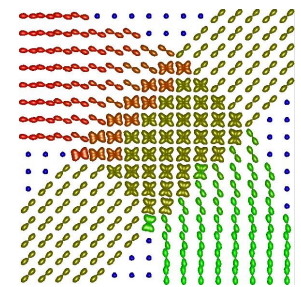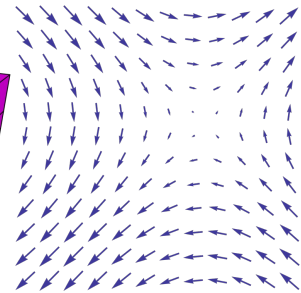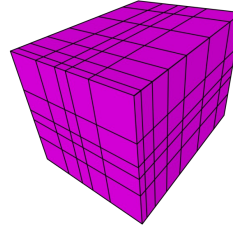
Semantics - Temporal Semantics

- Temporal attribute: any info relates to time

- Complicated to handle: rich hierarchical structure
  - Multiscale: nanoseconds ~ decades
  - Weeks do not cleanly fit into months

- Time-varying semantic: time is one of the key attributes
  - Time-series dataset: an ordered sequence of time-value pairs
    - Not always spaced at uniform intervals

- Dynamic
  - Time-varying semantics: time is a key attribute
  - Stream type dataset: items can change during a running session

Note

- Questions?