Lecture 4. Data Abstraction

(Munzner Ch.2)

Sim, Min Kyu, Ph.D.

mksim@seoultech.ac.kr

SEOULTECH 서울과학기술대학교 데이터사이언스학과

Why do data semantics and types matter?

- Let's look at the following data
  - 14, 2, 6, 30, 30, 15, 10001
  - Basil, 7, S, Pear
  - Can't be properly understood without knowing its semantic and type.
- Semantic
  - Its real world meaning
  - e.g. name, city, fruit, month
- Type
  - Its structural and mathematical interpretation
  - e.g. item, link, table, tree
- Metadata
  - Data that helps understanding the main data.

Data types

- Items
  - Individual and discrete entity
  - e.g.) table의 하나의 행, graph의 node
- Attribute
  - 측정, 관찰, 기록될 수 있는 특정한 성질
  - e.g.) salary, price, number of sales, protein levels, temperature
- Link
  - 아이템 들간의 관계 (대부분의 경우 graph에 존재)
  - e.g.) 두 도시 사이의 거리
- Grid
  - Continuous data를 sampling하기 위한 구체적 전략
- Position
  - 2D, 3D 공간에서에 위치정보 데이터
  - e.g.) latitude-longitude pair

Name	Age	Shirt Size	Favorite Fruit
Amy	8	S	Apple
Basil	7	S	Pear
Clara	9	M	Durian
Desmond	13	L	Elderberry
Ernest	12	L	Peach
Fanny	10	S	Lychee
George	9	M	Orange
Hector	8	L	Loquat
Ida	10	M	Pear
Amy	12	M	Orange

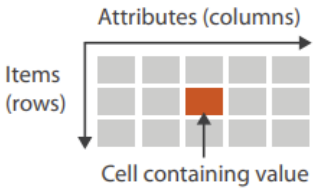
Dataset types

Data and Dataset Types

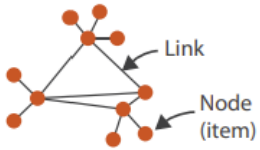
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

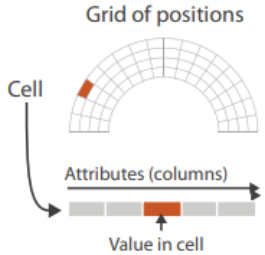
→ Tables



→ Networks

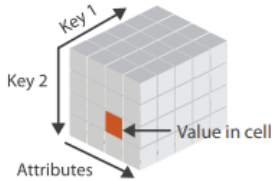


→ Fields (Continuous)



→ Geometry (Spatial)

→ *Multidimensional Table*



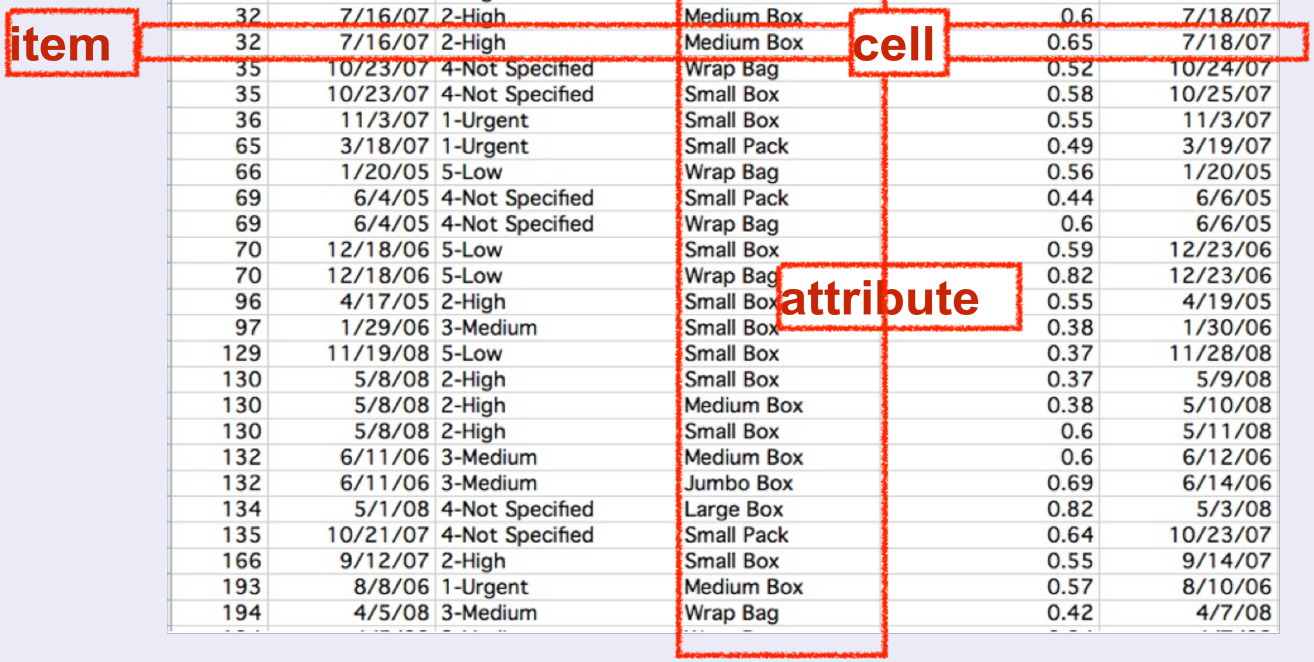
→ *Trees*

Tables

Flat table (*tidy* dataset in R)

- row(item of data)
- columns(attribute of the dataset)
- cell(=value)

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 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



Multidimensional table[1]

- More complex indexing
- Multiple keys

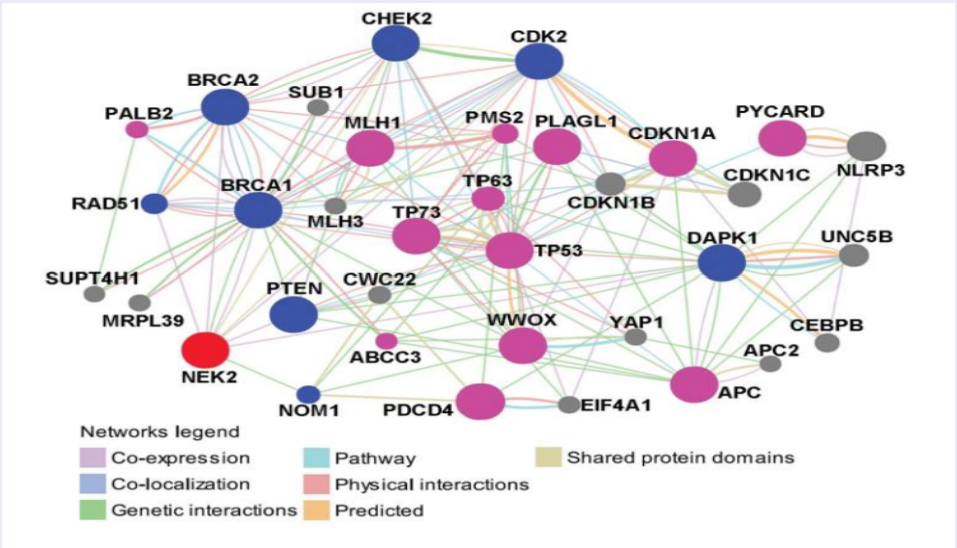
		USA			Europe		
		During last 12 months	Earlier than 12 months	Never	During last 12 months	Earlier than 12 months	Never
Men	1997	9.5%	8.3%	82.2%	9.2%	11.3%	79.5%
	2002	16.6%	7.8%	75.6%	14.4%	8.2%	77.4%
	2012	12.2%	22.9%	64.9%	13.7%	18.5%	67.8%
Women	1997	12.1%	7.0%	80.9%	12.6%	8.8%	78.6%
	2002	15.5%	8.4%	76.1%	17.6%	8.2%	74.2%
	2012	13.4%	23.6%	63.0%	14.9%	16.9%	68.2%
Total	1997	10.8%	7.7%	81.5%	11.0%	10.0%	79.0%
	2002	16.0%	8.1%	75.9%	16.0%	8.2%	75.8%
	2012	12.8%	23.3%	63.9%	14.3%	17.7%	68.0%

[a] https://stackoverflow.com/questions/50734739/what-is-the-best-structure-json-structure-for-a-multi-dimensional-array/50735012

Network and Tree

Network[1]

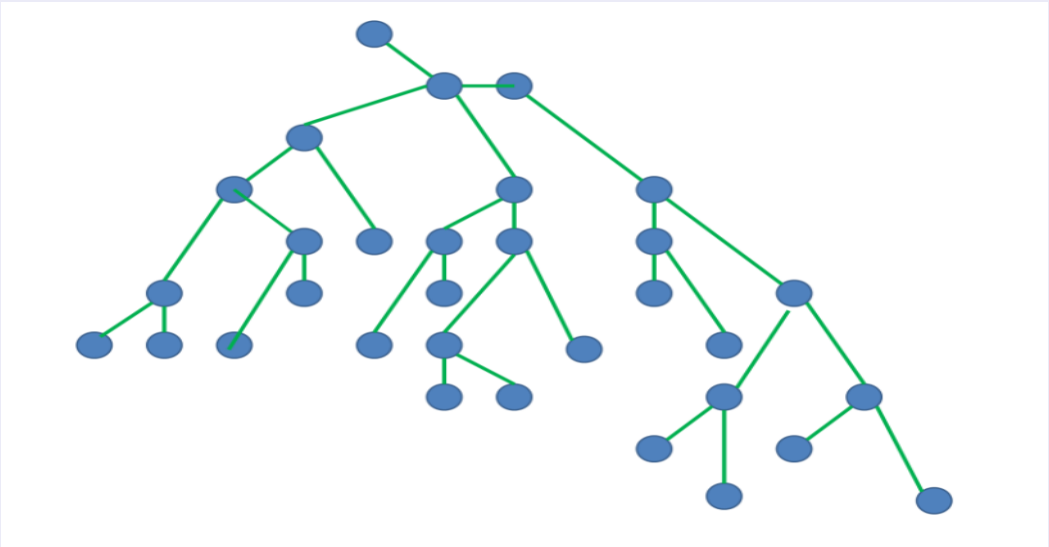
- Node (item), Link (relation between two items)
- e.g.) social network, gene interaction network
- Another name
  - network = graph
  - node = vertex
  - link = edge



[a] https://www.spandidos-publications.com/10.3892/or.2013.2910

Tree[1]

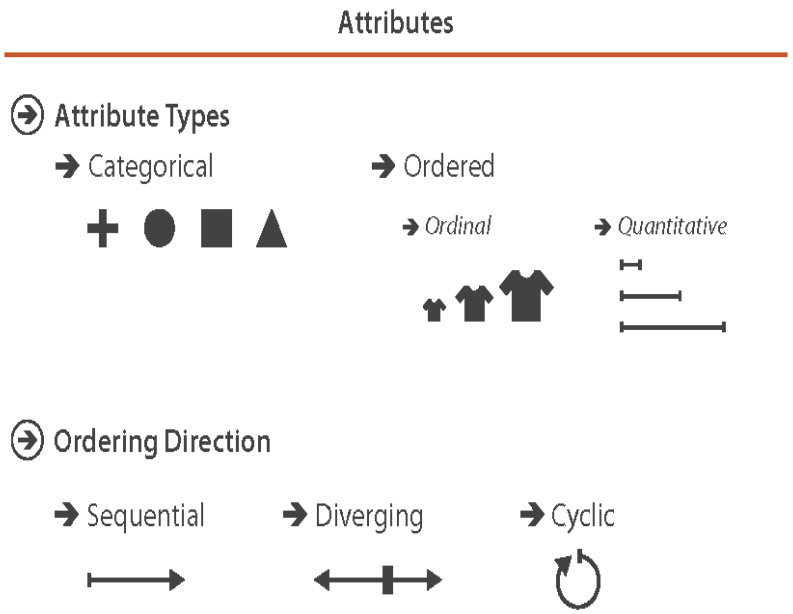
- Network with hierarchical structure
- No cycle
- 1 child has only 1 parent
- e.g.) organizational chart, biological tree



[b] https://forum.unity.com/threads/dynamic-way-of-displaying-node-graph-talent-tree.411263/

Taxonomy of attribute types

- Attribute types
  1. Categorical
  2. Ordered
     - Ordinal
     - Quantitative
- Ordering direction
  - Sequential
  - Diverging
  - Cyclic



1. Categorical

- No implicit ordering
- 같은지 다른지만을 구분
- e.g.) fruits, movie genres, file types
- c.f.) arbitrary ordering is possible like by alphabetical order or by price

2. Ordered

- Has implicit ordering
- **Ordinal** vs **Quantitative**
  - **Ordinal**
    - 연산은 불가능함
    - e.g.) shirt size, ranking
    - e.g.) 셔츠사이즈 L은 S보다 크지만 L에서 S를 빼는 것은 불가능
  - **Quantitative**
    - 연산이 가능함
    - e.g.) height, weight, temperature, stock price
- Direction - **Sequential** vs **Diverging** vs **Cyclic**
  - **Sequential**
    - Minimum과 Maximum하의 범위에서 제한됨
    - e.g.) mountain height
  - **Diverging**
    - 방향이 다르지만 0을 공통으로 가지는 수치로 분해가 가능함
    - e.g.) elevation data
  - **Cyclic**
    - 원점으로 돌아와서 다시 진행하는 데이터
    - e.g.) hour of the day, the day of the week

Hierarchical attributes

- 많은 attribute가 계층적 구조를 가짐
- 대표적으로 시공간에 관련된 attribute가 계층적 구조를 가짐
  - 시간: 초 -> 분 -> 시 -> 일 -> 주 -> 월 -> 연 -> decade
  - 공간: 국가 -> 도 -> 시 -> 구 -> 동

```
"Data Visualization"
```

```
## [1] "Data Visualization"
```