VISUALIZATION

Data

Overview

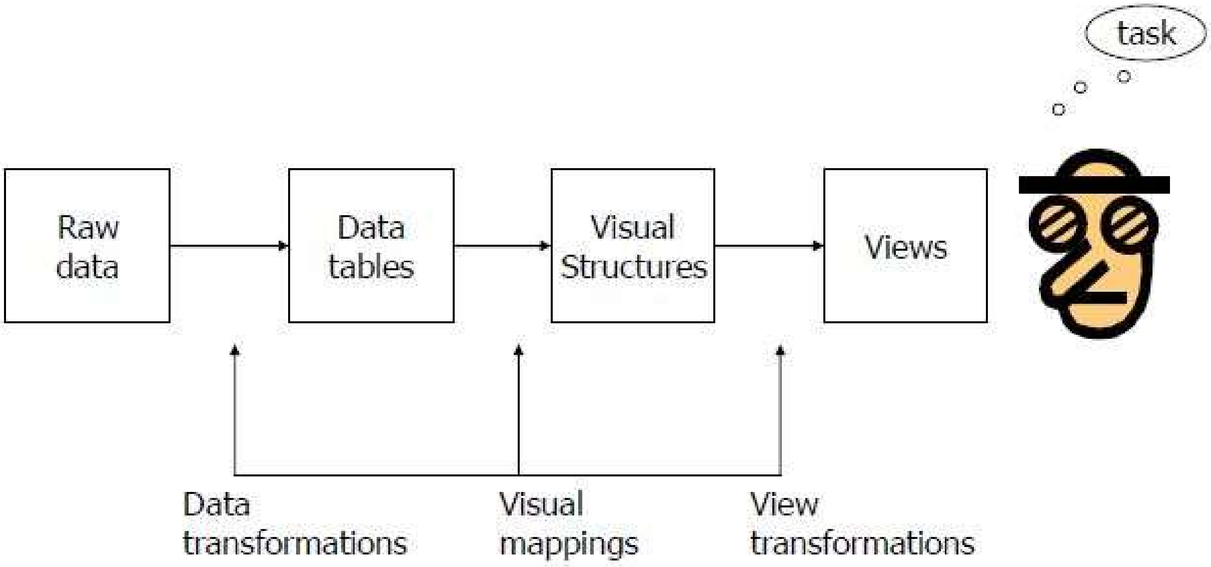
- Data
- Attribute Types
- Dataset Types

Overview

- Data
- Attribute Types
- Dataset Types

Visualization

- Visualization is the process that **transforms** (abstract) **data** into **interactive graphical representations** for the purpose of **exploration, confirmation, or presentation**
- **Goal:**
  - Record - Preserve information
  - Analyze - Reason about data
  - Present - Convey information



Data

- Data is taken from and/or representing some phenomena from the world
- Data Types are structural or mathematical interpretation of data
  - Fundamental unit
  - **Item, Link, Attribute, Position, Grid**
  - Different from data types in programming!

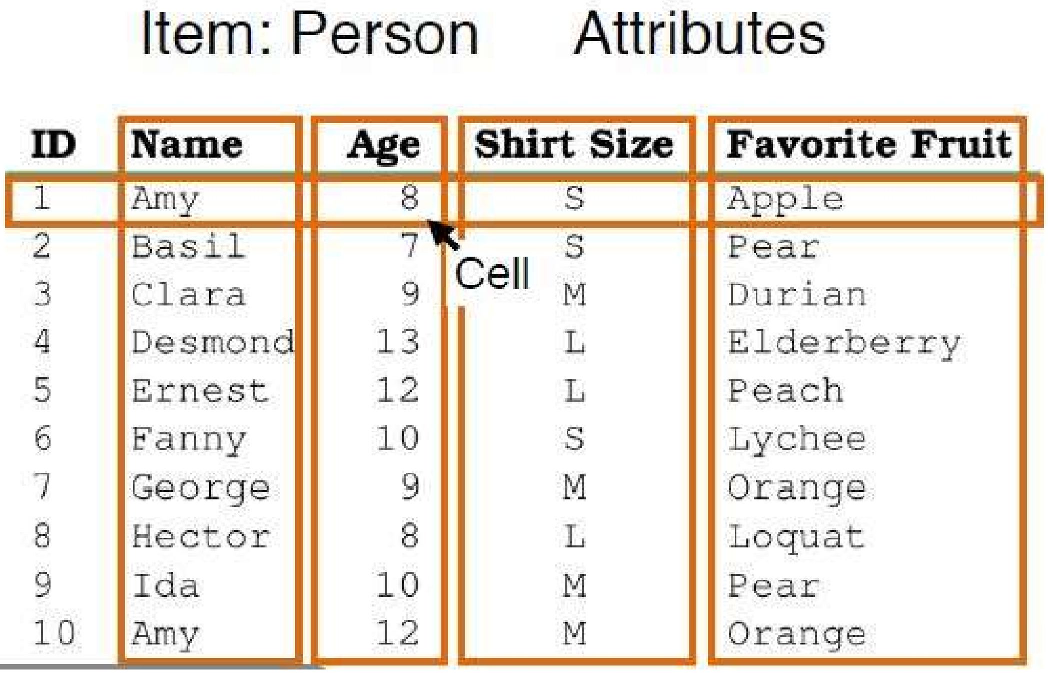
Items & Attributes

Item: individual entity, discrete

e.g., Patient, Car, Stock, City

Attribute: measured, observed, logged property

e.g., Patient: height, blood pressure; Car: horsepower, make

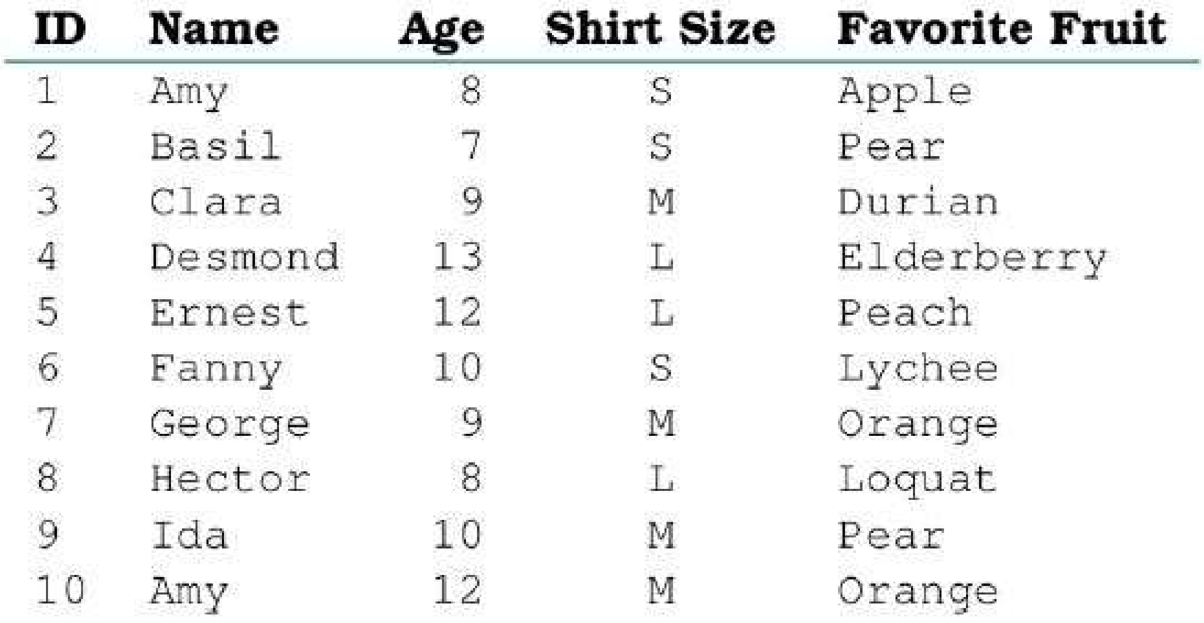

Item: Person
Attributes
Cell

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

Other Data Types

- Links/Relations
  - Express relationship between two items
  - Friendship on Facebook, Interaction between proteins
- Positions
  - Spatial data -> location in 2D or 3D
  - Pixels in photo, Voxels in MRI scan, latitude/longitude
- Grids
  - Sampling strategy for continuous data

Data Semantics



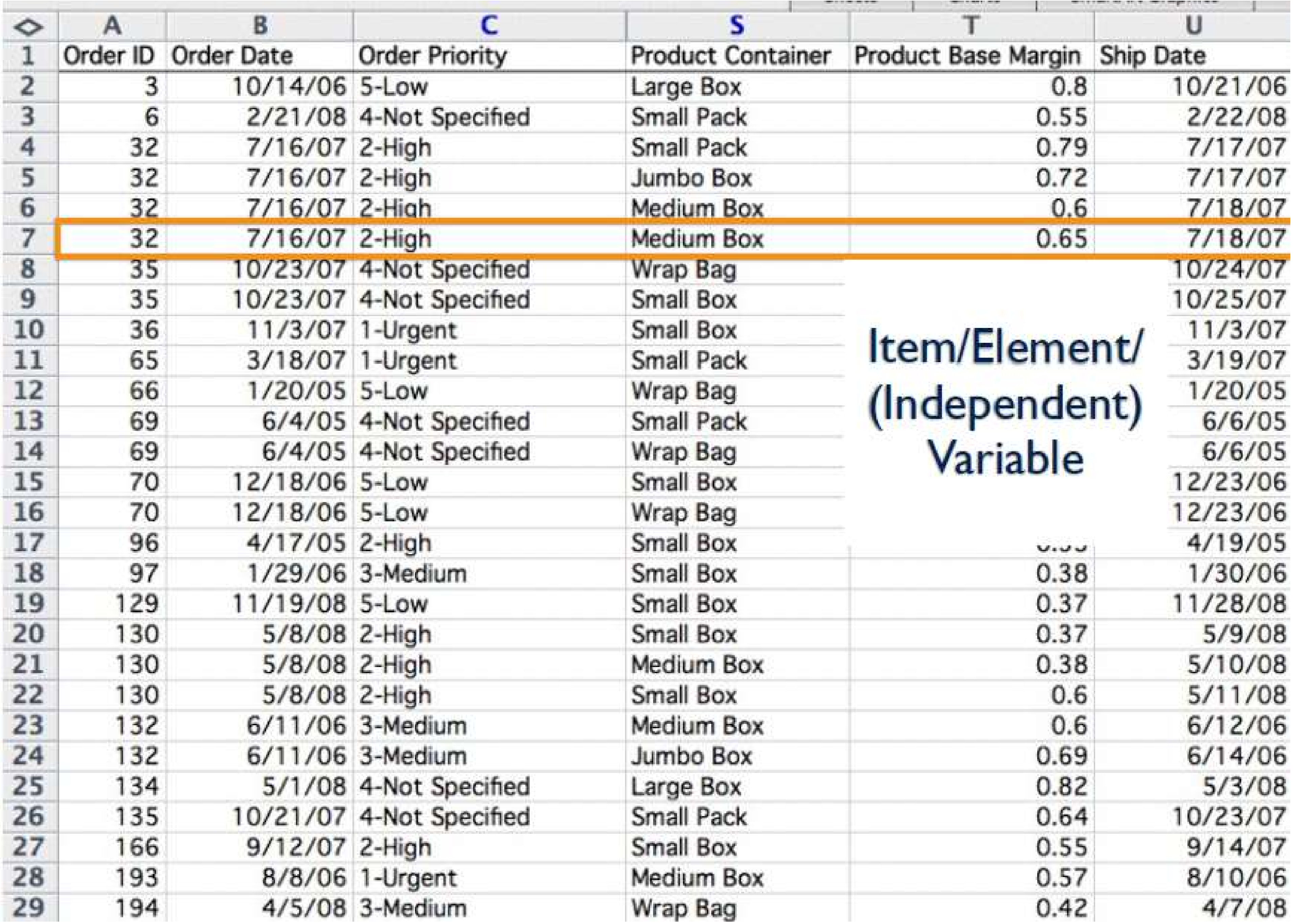
ID	Name	Age	Shirt Size	Favorite Fruit
1	Amy	8	S	Apple
2	Basil	7	S	Pear
3	Clara	9	M	Durian
4	Desmond	13	L	Elderberry
5	Ernest	12	L	Peach
6	Fanny	10	S	Lychee
7	George	9	M	Orange
8	Hector	8	L	Loquat
9	Ida	10	M	Pear
10	Amy	12	M	Orange

- Basil, 7, S, Pear
- What does it mean?
- **Semantics:** real world meaning
- Name? City? Fruit? Height? Age? Day of Month?
- Metadata: Descriptive information about the data
  - Is utilized to associate Semantics

Structured vs Unstructured

- Structured Data
  - Known data types, semantics
- Unstructured Data
  - No predefined data model
  - Text-heavy, interspersed with facts (dates, times, locations)
  - Video, images
  - Need to be converted to structured data for visualization
- Unstructured -> Structured
  - Natural Language Processing
  - Text mining (sentiment, keywords, concepts, categories)

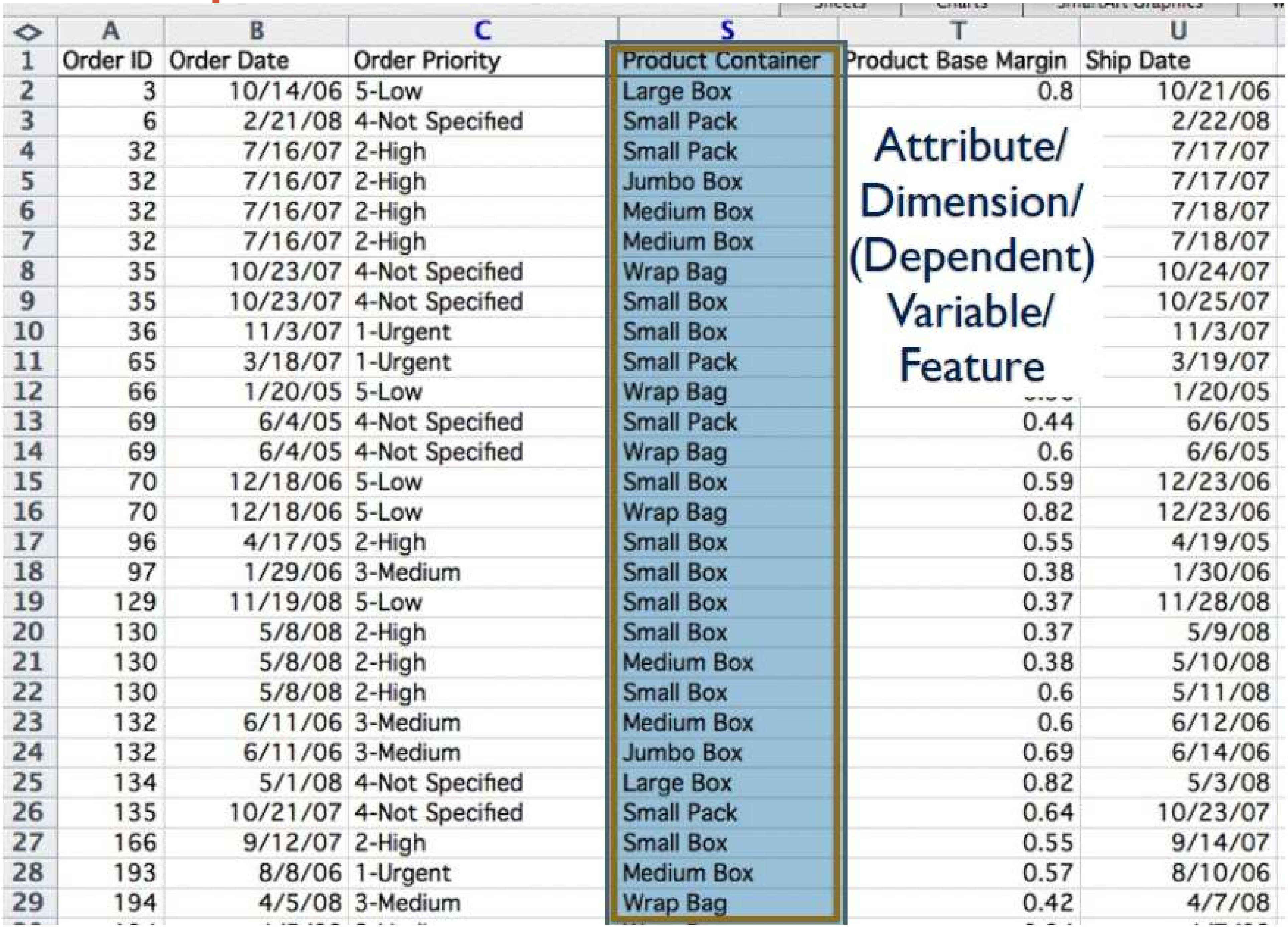
Example - Item



	A	B	C	S	T	U
1	Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
2	3	10/14/06	5-Low	Large Box	0.8	10/21/06
3	6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
4	32	7/16/07	2-High	Small Pack	0.79	7/17/07
5	32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
6	32	7/16/07	2-High	Medium Box	0.6	7/18/07
7	32	7/16/07	2-High	Medium Box	0.65	7/18/07
8	35	10/23/07	4-Not Specified	Wrap Bag		10/24/07
9	35	10/23/07	4-Not Specified	Small Box		10/25/07
10	36	11/3/07	1-Urgent	Small Box		11/3/07
11	65	3/18/07	1-Urgent	Small Pack		3/19/07
12	66	1/20/05	5-Low	Wrap Bag		1/20/05
13	69	6/4/05	4-Not Specified	Small Pack		6/6/05
14	69	6/4/05	4-Not Specified	Wrap Bag		6/6/05
15	70	12/18/06	5-Low	Small Box		12/23/06
16	70	12/18/06	5-Low	Wrap Bag		12/23/06
17	96	4/17/05	2-High	Small Box	[illegible]	4/19/05
18	97	1/29/06	3-Medium	Small Box	0.38	1/30/06
19	129	11/19/08	5-Low	Small Box	0.37	11/28/08
20	130	5/8/08	2-High	Small Box	0.37	5/9/08
21	130	5/8/08	2-High	Medium Box	0.38	5/10/08
22	130	5/8/08	2-High	Small Box	0.6	5/11/08
23	132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
24	132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
25	134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
26	135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
27	166	9/12/07	2-High	Small Box	0.55	9/14/07
28	193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
29	194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Item/Element/
(Independent)
Variable

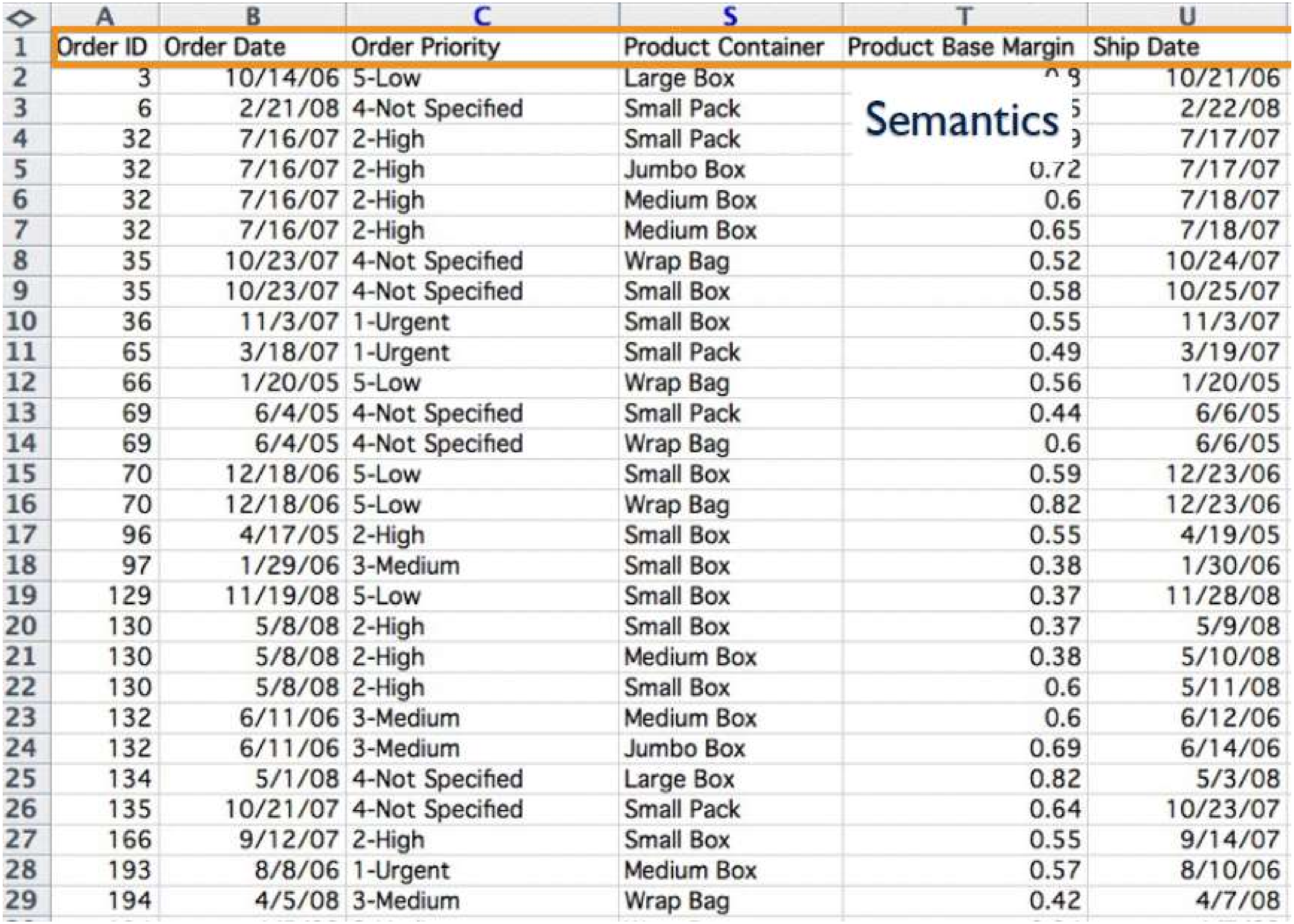
Example - Attribute



	A	B	C	S	T	U
1	Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
2	3	10/14/06	5-Low	Large Box	0.8	10/21/06
3	6	2/21/08	4-Not Specified	Small Pack		2/22/08
4	32	7/16/07	2-High	Small Pack		7/17/07
5	32	7/16/07	2-High	Jumbo Box		7/17/07
6	32	7/16/07	2-High	Medium Box		7/18/07
7	32	7/16/07	2-High	Medium Box		7/18/07
8	35	10/23/07	4-Not Specified	Wrap Bag		10/24/07
9	35	10/23/07	4-Not Specified	Small Box		10/25/07
10	36	11/3/07	1-Urgent	Small Box		11/3/07
11	65	3/18/07	1-Urgent	Small Pack		3/19/07
12	66	1/20/05	5-Low	Wrap Bag		1/20/05
13	69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
14	69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
15	70	12/18/06	5-Low	Small Box	0.59	12/23/06
16	70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
17	96	4/17/05	2-High	Small Box	0.55	4/19/05
18	97	1/29/06	3-Medium	Small Box	0.38	1/30/06
19	129	11/19/08	5-Low	Small Box	0.37	11/28/08
20	130	5/8/08	2-High	Small Box	0.37	5/9/08
21	130	5/8/08	2-High	Medium Box	0.38	5/10/08
22	130	5/8/08	2-High	Small Box	0.6	5/11/08
23	132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
24	132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
25	134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
26	135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
27	166	9/12/07	2-High	Small Box	0.55	9/14/07
28	193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
29	194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Attribute/
Dimension/
(Dependent)
Variable/
Feature

Example - Semantics



◇	A	B	C	S	T	U
1	Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
2	3	10/14/06	5-Low	Large Box	0.8	10/21/06
3	6	2/21/08	4-Not Specified	Small Pack	5	2/22/08
4	32	7/16/07	2-High	Small Pack	9	7/17/07
5	32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
6	32	7/16/07	2-High	Medium Box	0.6	7/18/07
7	32	7/16/07	2-High	Medium Box	0.65	7/18/07
8	35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
9	35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
10	36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
11	65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
12	66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
13	69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
14	69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
15	70	12/18/06	5-Low	Small Box	0.59	12/23/06
16	70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
17	96	4/17/05	2-High	Small Box	0.55	4/19/05
18	97	1/29/06	3-Medium	Small Box	0.38	1/30/06
19	129	11/19/08	5-Low	Small Box	0.37	11/28/08
20	130	5/8/08	2-High	Small Box	0.37	5/9/08
21	130	5/8/08	2-High	Medium Box	0.38	5/10/08
22	130	5/8/08	2-High	Small Box	0.6	5/11/08
23	132	6/11/06	3-Medium	Medium Box	0.6	6/12/06
24	132	6/11/06	3-Medium	Jumbo Box	0.69	6/14/06
25	134	5/1/08	4-Not Specified	Large Box	0.82	5/3/08
26	135	10/21/07	4-Not Specified	Small Pack	0.64	10/23/07
27	166	9/12/07	2-High	Small Box	0.55	9/14/07
28	193	8/8/06	1-Urgent	Medium Box	0.57	8/10/06
29	194	4/5/08	3-Medium	Wrap Bag	0.42	4/7/08

Semantics

Overview

- Data
- Attribute Types
- Dataset Types

Attribute Types

- Categorical (nominal)
  - Compare equality
  - *Fruit, Gender, Movie Genres, File Types*
- Ordered
  - Ordinal
    - Great/Less than defined
    - *Shirt size*
  - Quantitative
    - Arithmetic possible
    - *Length, Weight, Count*

Quantitative Types

- Interval (arbitrary zero)
  - Dates: Jan 19; Location: (Lat, Long)
  - Only differences (i.e., intervals) can be compared
- Ratio (true zero)
  - zero: there is nothing of the measured entity observed
  - Measurements: Length, Mass
  - Can measure ratios & proportions

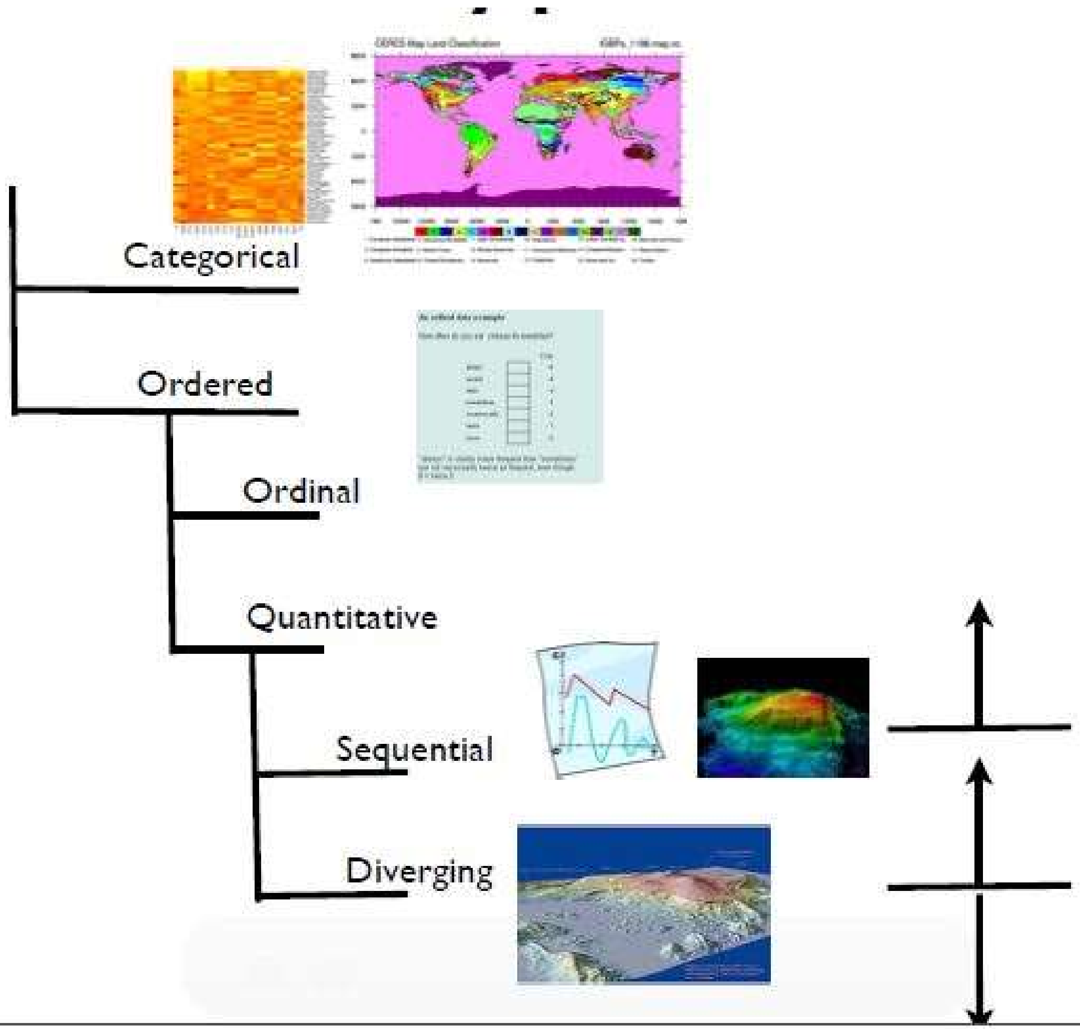
Operations

- Nominal (labels)
  - Operations: =, ≠
- Ordinal (ordered)
  - Operations: =, ≠, >, <
- Interval (location of zero arbitrary)
  - Operations: =, ≠, >, <, +, − (distance)
- Ratio (zero fixed)
  - Operations: =, ≠, >, <, +, −,×, ÷ (proportions)

Sequential vs Diverging data

- Sequential:
  - homogeneous from min to max
  - # people in countries
- Diverging:
  - two or multiple sequences that meet
  - Elevation dataset: above sea level & below sea level

Attribute Types (Summary)



Categorical

Ordered

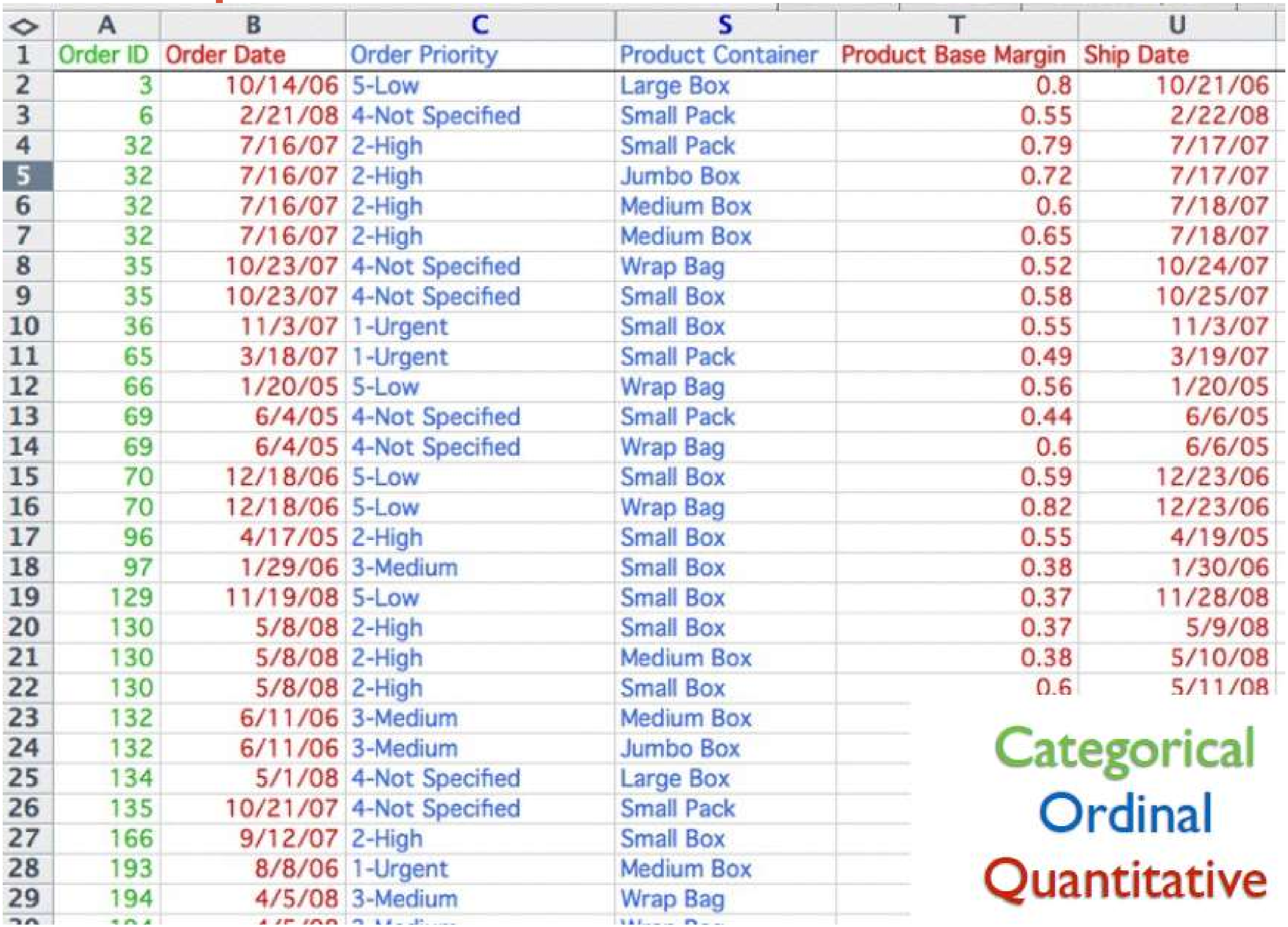
Ordinal

Quantitative

Sequential

Diverging

Example



◇	A	B	C	S	T	U
1	Order ID	Order Date	Order Priority	Product Container	Product Base Margin	Ship Date
2	3	10/14/06	5-Low	Large Box	0.8	10/21/06
3	6	2/21/08	4-Not Specified	Small Pack	0.55	2/22/08
4	32	7/16/07	2-High	Small Pack	0.79	7/17/07
5	32	7/16/07	2-High	Jumbo Box	0.72	7/17/07
6	32	7/16/07	2-High	Medium Box	0.6	7/18/07
7	32	7/16/07	2-High	Medium Box	0.65	7/18/07
8	35	10/23/07	4-Not Specified	Wrap Bag	0.52	10/24/07
9	35	10/23/07	4-Not Specified	Small Box	0.58	10/25/07
10	36	11/3/07	1-Urgent	Small Box	0.55	11/3/07
11	65	3/18/07	1-Urgent	Small Pack	0.49	3/19/07
12	66	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
13	69	6/4/05	4-Not Specified	Small Pack	0.44	6/6/05
14	69	6/4/05	4-Not Specified	Wrap Bag	0.6	6/6/05
15	70	12/18/06	5-Low	Small Box	0.59	12/23/06
16	70	12/18/06	5-Low	Wrap Bag	0.82	12/23/06
17	96	4/17/05	2-High	Small Box	0.55	4/19/05
18	97	1/29/06	3-Medium	Small Box	0.38	1/30/06
19	129	11/19/08	5-Low	Small Box	0.37	11/28/08
20	130	5/8/08	2-High	Small Box	0.37	5/9/08
21	130	5/8/08	2-High	Medium Box	0.38	5/10/08
22	130	5/8/08	2-High	Small Box	0.6	5/11/08
23	132	6/11/06	3-Medium	Medium Box		
24	132	6/11/06	3-Medium	Jumbo Box		
25	134	5/1/08	4-Not Specified	Large Box		
26	135	10/21/07	4-Not Specified	Small Pack		
27	166	9/12/07	2-High	Small Box		
28	193	8/8/06	1-Urgent	Medium Box		
29	194	4/5/08	3-Medium	Wrap Bag		

Categorical

Ordinal

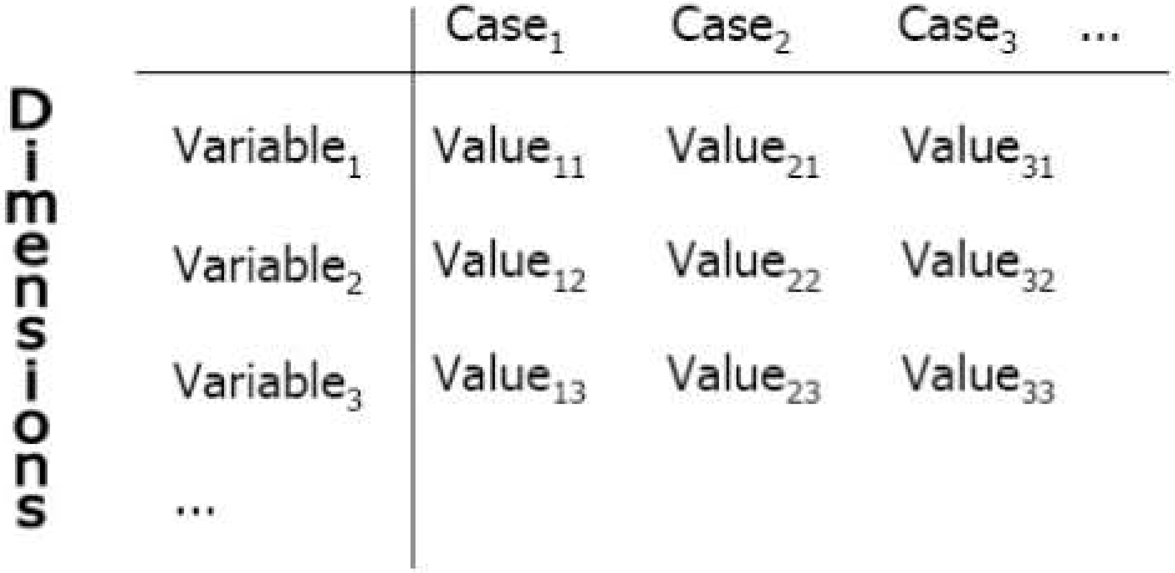
Quantitative

Dimensions

- Data sets of dimensions 1, 2, 3 are common
- Number of variables/attributes per class/item
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - >3 - Hypervariate data

Data Table

- Items (*cases*) have attributes (*variables*)
- Function $f(\mathrm{Case}_l) = \langle \mathrm{Value}_{l1}, \mathrm{Value}_{l2}, \ldots \rangle$



Dimensions

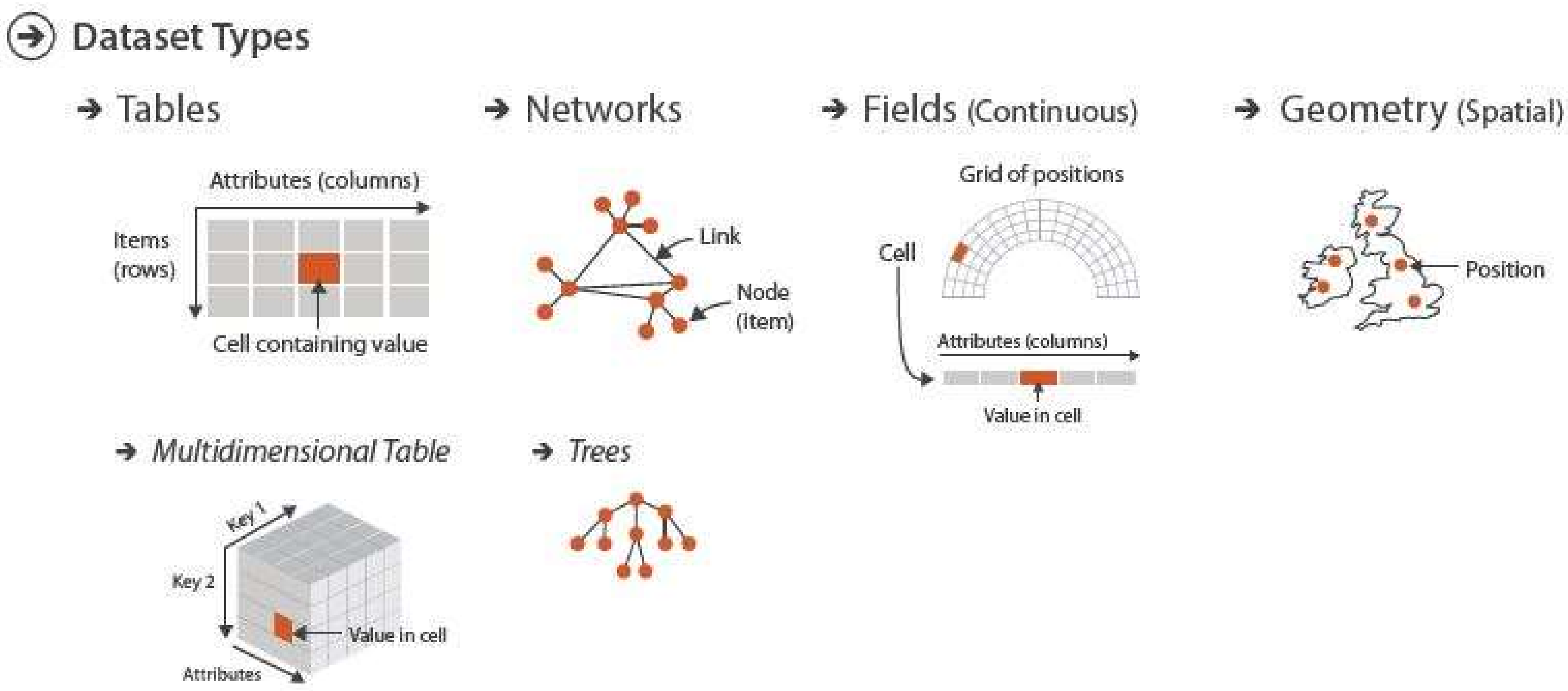
	Case_1	Case_2	Case_3	…
$\mathrm{Variable}_1$	Value_{11}	Value_{21}	Value_{31}	
$\mathrm{Variable}_2$	Value_{12}	Value_{22}	Value_{32}	
$\mathrm{Variable}_3$	Value_{13}	Value_{23}	Value_{33}	
…				

Overview

- Data
- Attribute Types
- Dataset Types

Dataset Types

- Data types combine to form Dataset Types



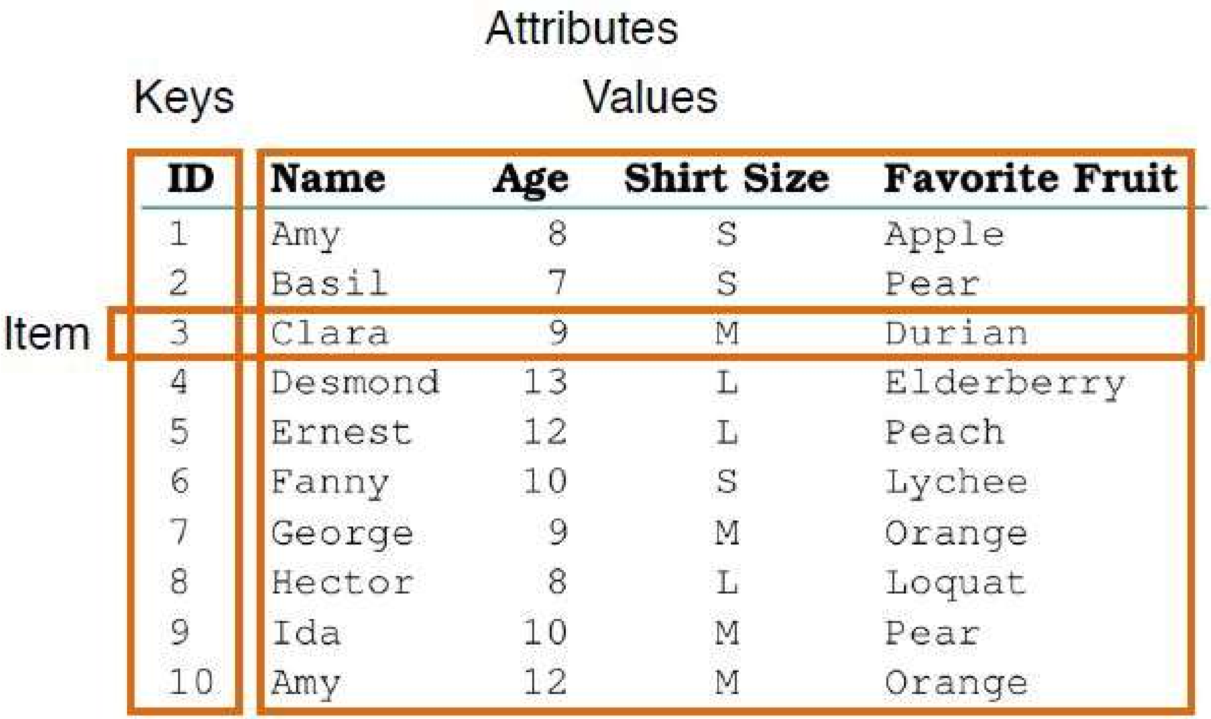
Tables

Flat Table

- one item per row
- each column is attribute
- unique (implicit) **key**
- no duplicates

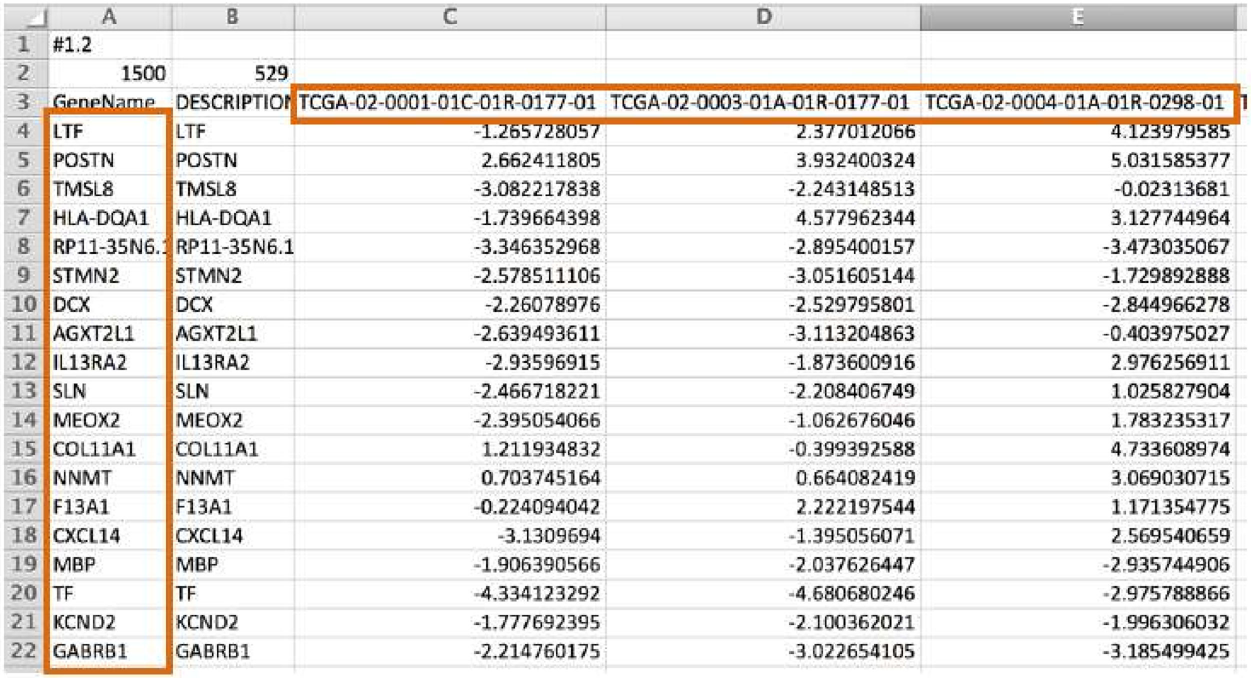
Multidimensional Table

- indexing based on multiple keys


Attributes
Keys
Values
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

Multidimensional Table



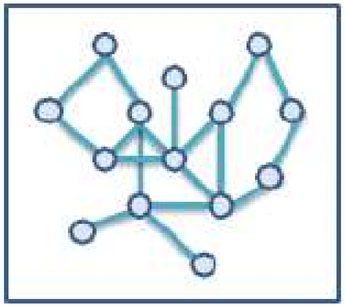
	A	B	C	D	E
1	#1.2				
2	1500	529			
3	GeneName	DESCRIPTION	TCGA-02-0001-01C-01R-0177-01	TCGA-02-0003-01A-01R-0177-01	TCGA-02-0004-01A-01R-0298-01
4	LTF	LTF	-1.265728057	2.377012066	4.123979585
5	POSTN	POSTN	2.662411805	3.932400324	5.031585377
6	TMSL8	TMSL8	-3.082217838	-2.243148513	-0.02313681
7	HLA-DQA1	HLA-DQA1	-1.739664398	4.577962344	3.127744964
8	RP11-35N6.1	RP11-35N6.1	-3.346352968	-2.895400157	-3.473035067
9	STMN2	STMN2	-2.578511106	-3.051605144	-1.729892888
10	DCX	DCX	-2.26078976	-2.529795801	-2.844966278
11	AGXT2L1	AGXT2L1	-2.639493611	-3.113204863	-0.403975027
12	IL13RA2	IL13RA2	-2.93596915	-1.873600916	2.976256911
13	SLN	SLN	-2.466718221	-2.208406749	1.025827904
14	MEOX2	MEOX2	-2.395054066	-1.062676046	1.783235317
15	COL11A1	COL11A1	1.211934832	-0.399392588	4.733608974
16	NNMT	NNMT	0.703745164	0.664082419	3.069030715
17	F13A1	F13A1	-0.224094042	2.222197544	1.171354775
18	CXCL14	CXCL14	-3.1309694	-1.395056071	2.569540659
19	MBP	MBP	-1.906390566	-2.037626447	-2.935744906
20	TF	TF	-4.334123292	-4.680680246	-2.975788866
21	KCND2	KCND2	-1.777692395	-2.100362021	-1.996306032
22	GABRB1	GABRB1	-2.214760175	-3.022654105	-3.185499425

Keys: Patients

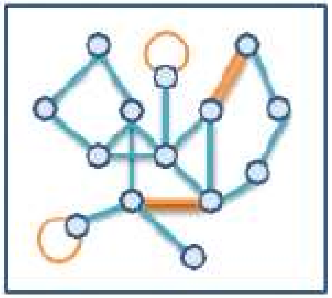
Keys: Genes

Graphs/Networks

- A graph G(V,E) consists of a set of **vertices (nodes)** V and a set of **edges (links)** E connecting these vertices.



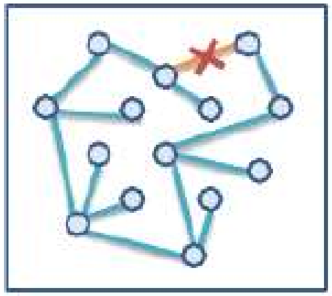
- A simple graph is a graph which contains:
  - No multi-edges
  - No loops



Not a simple graph!
→ A ***general graph***

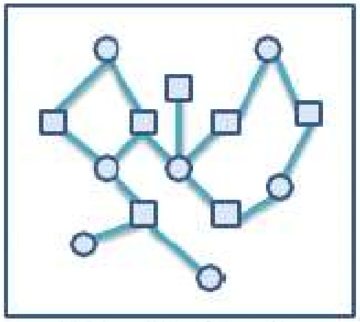
Special Graphs

- A ***tree/hierarchy*** is a graph with *no cycles*



Tree

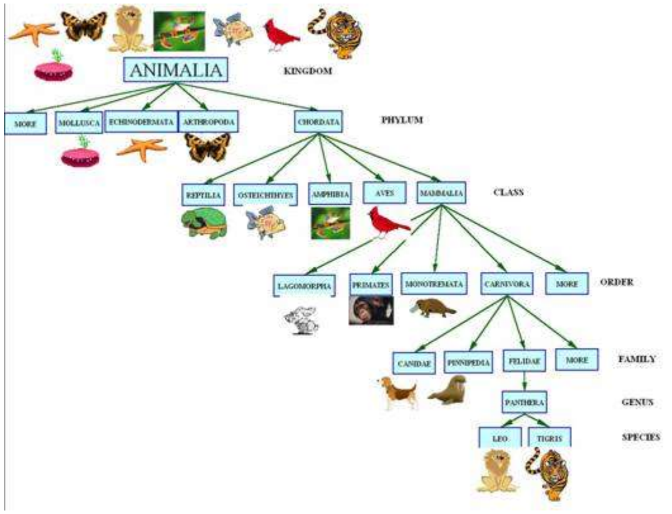
- A ***directed graph*** (digraph) is a graph that distinguishes between edges A-> B and A <- B
- A ***bipartite graph*** has vertices that can be partitioned into two independent sets



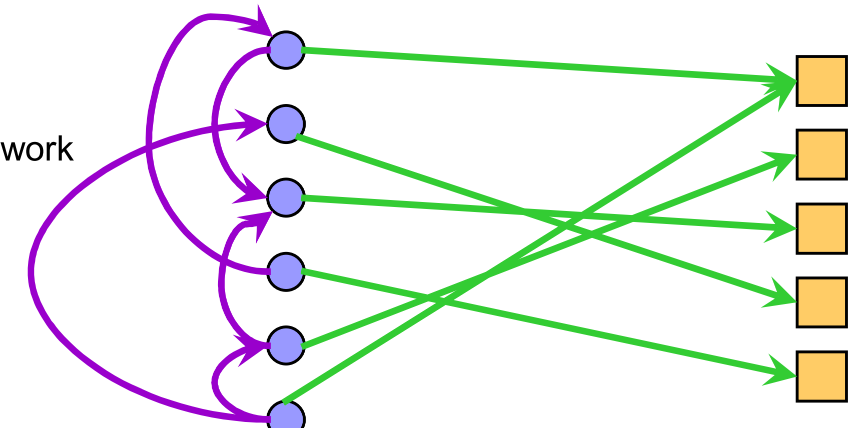
Bipartite Graph

Graphs in Real World

- Graph: WWW, Social Networks
- Tree: Org structures, Classifications
- Bipartite graph



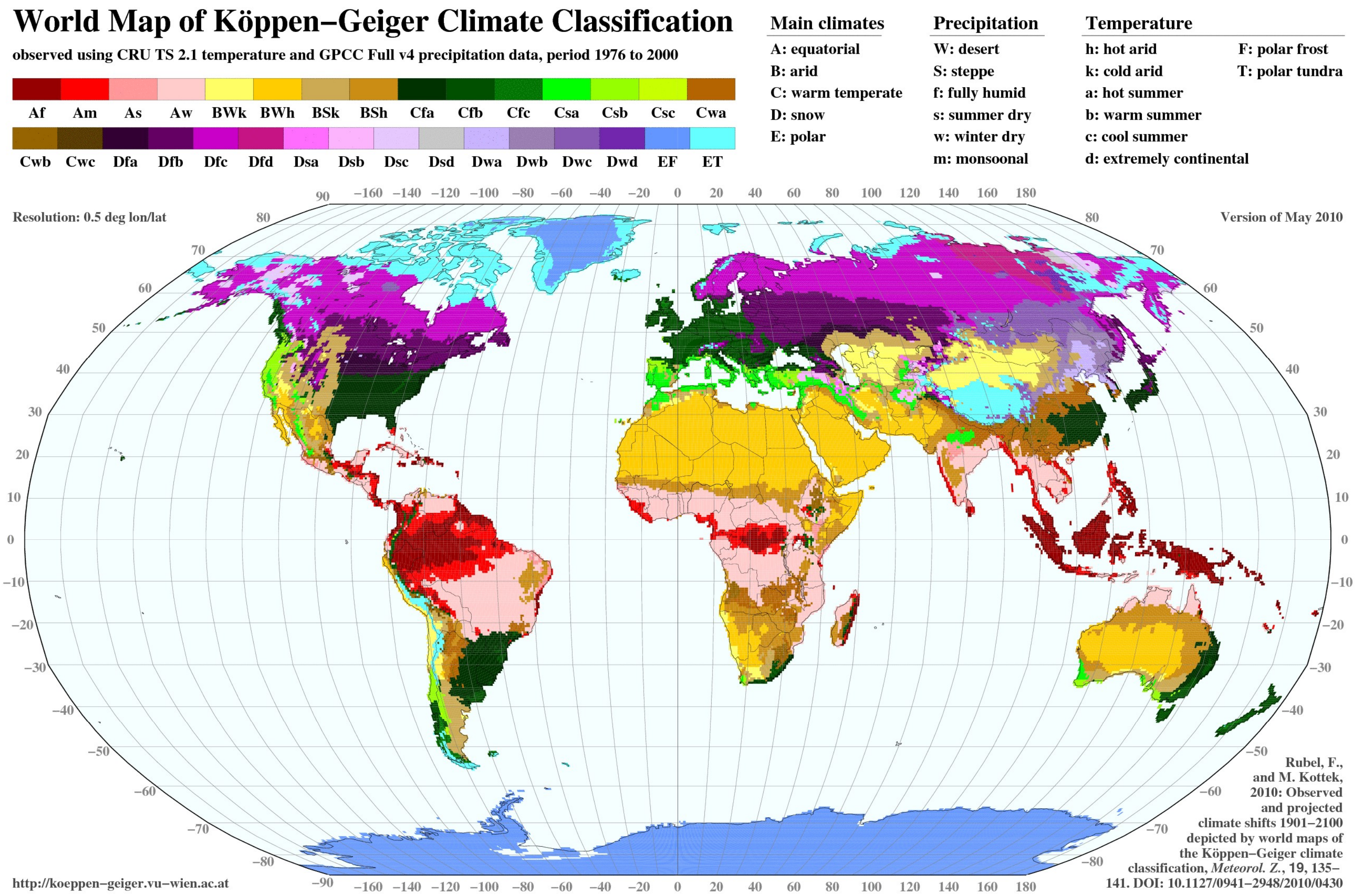
People
Social Network



Things (products, services,…)

Who’s buying what ?

Spatial Data



World Map of Köppen–Geiger Climate Classification

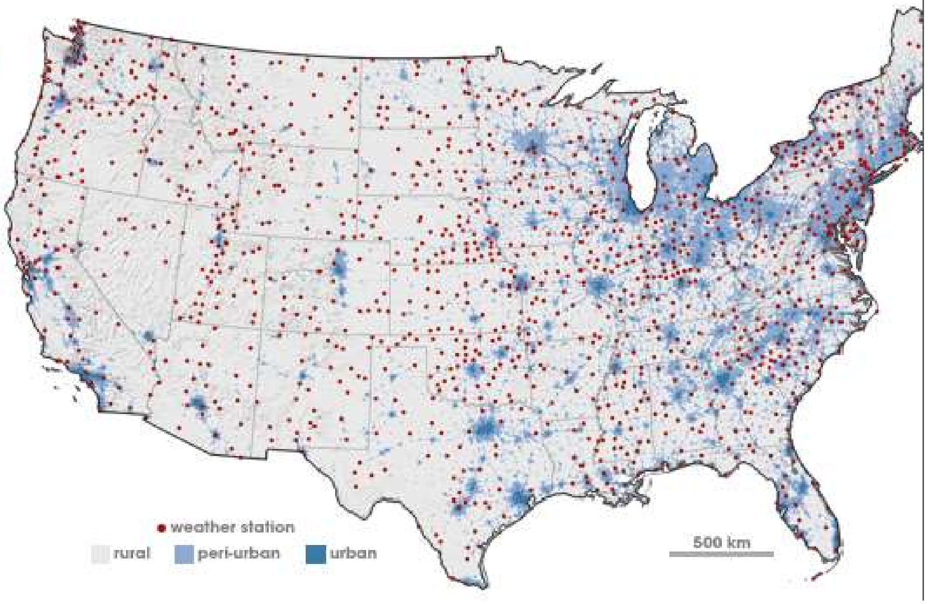
observed using CRU TS 2.1 temperature and GPCC Full v4 precipitation data, period 1976 to 2000

http://koeppen−geiger.vu−wien.ac.at

Rubel, F., and M. Kottek, 2010: Observed and projected climate shifts 1901−2100 depicted by world maps of the Köppen−Geiger climate classification, *Meteorol. Z.*, 19, 135−141. DOI: 10.1127/0941−2948/2010/0430

Fields

- Attribute values associated with cells
- Cell contains data from continuous domain
  - Temperature, pressure, wind velocity
- Measured or simulated
- Sampling & Interpolation
  - Signal processing & stats



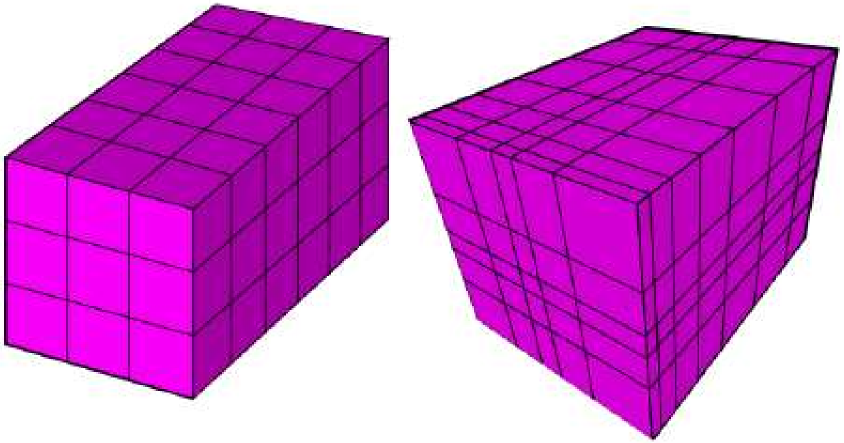
Fields: Grid Types

Uniform Grid

- Geometry & topology can be computed

Rectilinear Grid

- Nonuniform sampling

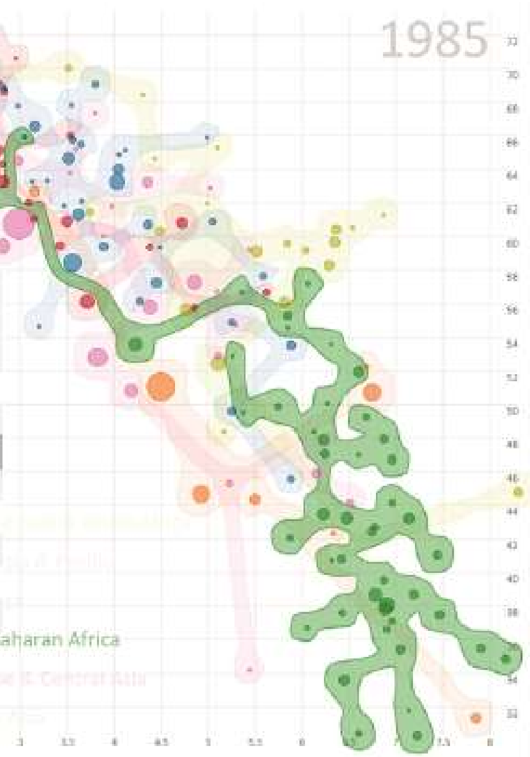


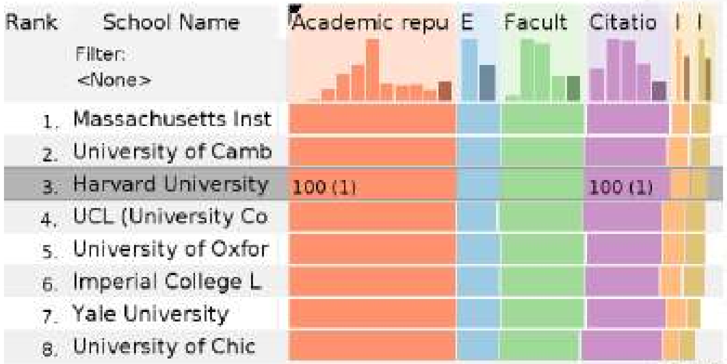
Information vs Scientific Visualization

- **Information Vis**
  - “Abstract Data”
  - Tables, Graphs
  - Free to choose spatial layout
- **Scientific Vis**
  - “Spatial Data” (Fields)
  - Not free to choose spatial layout
  - Find best way to depict reality

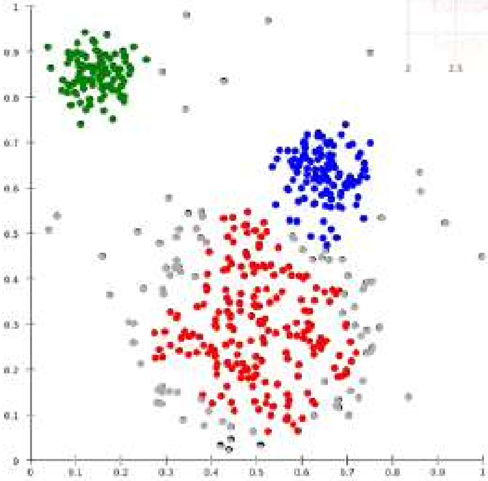
Other Collections

- Sets
  - Unique items, unordered
- Lists
  - Ordered, duplicates allowed
- Clusters
  - Groups of similar items





Rank	School Name	Academic repu	E	Facult	Citatio	I	I
	Filter: <None>						
1.	Massachusetts Inst						
2.	University of Camb						
3.	Harvard University	100 (1)			100 (1)		
4.	UCL (University Co						
5.	University of Oxfor						
6.	Imperial College L						
7.	Yale University						
8.	University of Chic						



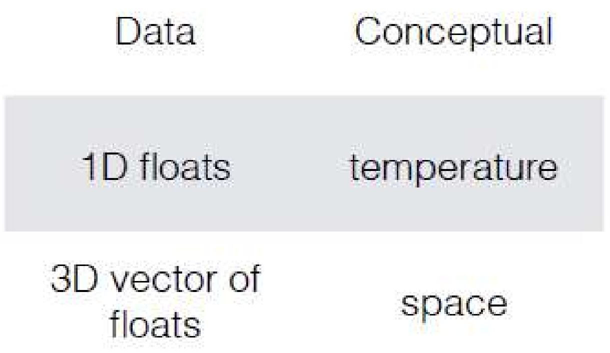
Data vs Conceptual Model

Data Model: Low-level description of the data

- Set with operations, e.g., floats with +, -, /, *

Conceptual Model: Mental construction

- Includes semantics, supports reasoning



Data	Conceptual
1D floats	temperature
3D vector of floats	space

Data Model -> Data Type

- From data model...
  - 32.5, 54.0, -17.3, … (floats)
- Using conceptual model...
  - Temperature
- To data type
  - Continuous to 4 significant digits (Q)
  - Hot, warm, cold (O)
  - Burned vs. Not burned (N)

Combinations, Derived Data

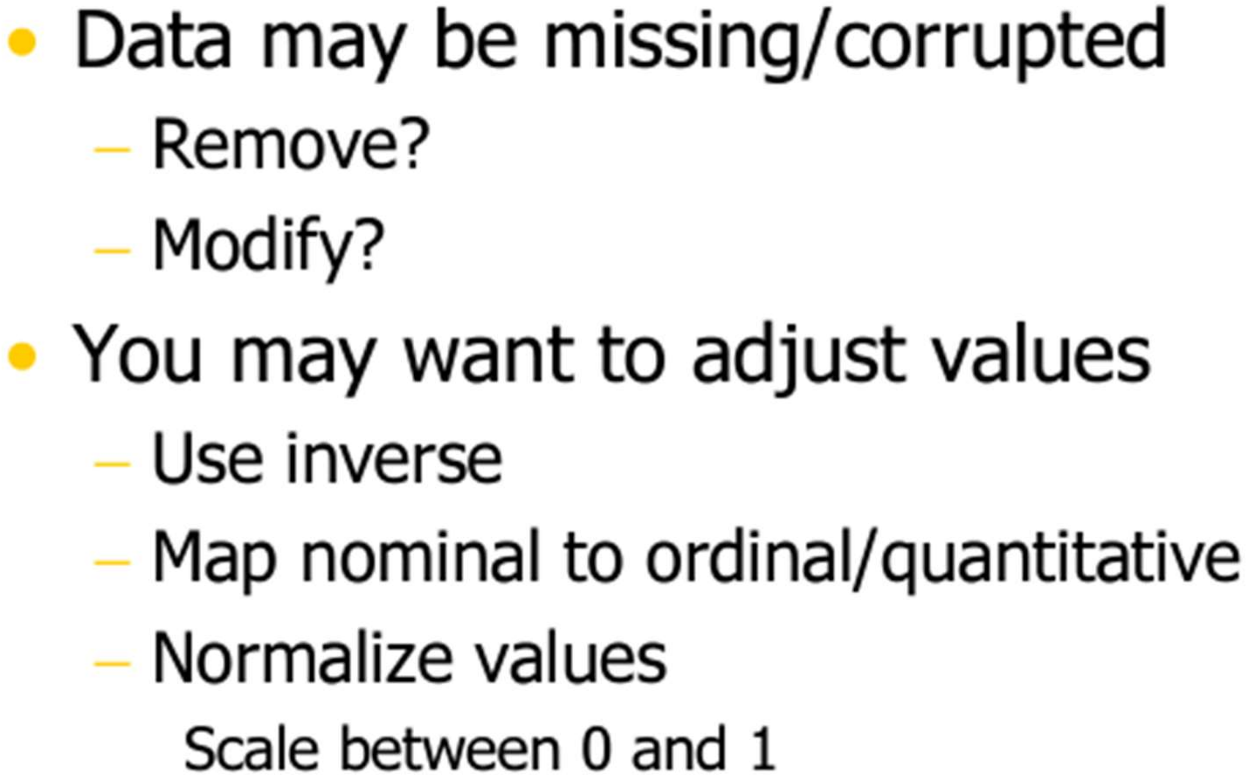
- Networks can have attributes
- Attributes have hierarchies
- Data types can be transformed

- Real life is complicated…

Metadata

- Descriptive information about the data
  - Might be something as simple as the type of a variable, or could be more complex
  - For times when the table itself just isn't enough
  - Example: if variable1 is "l", then variable3 can only be 3, 7 or 16

Data Cleansing



- Data may be missing/corrupted
  - Remove?
  - Modify?
- You may want to adjust values
  - Use inverse
  - Map nominal to ordinal/quantitative
  - Normalize values

    Scale between 0 and 1

Publicly Available Datasets

- http://www.kdnuggets.com/datasets/index.html
- http://www.google.com/publicdata/directory
- https://developers.facebook.com/docs/graph-api
- http://www.data.gov/
- *Many more!*