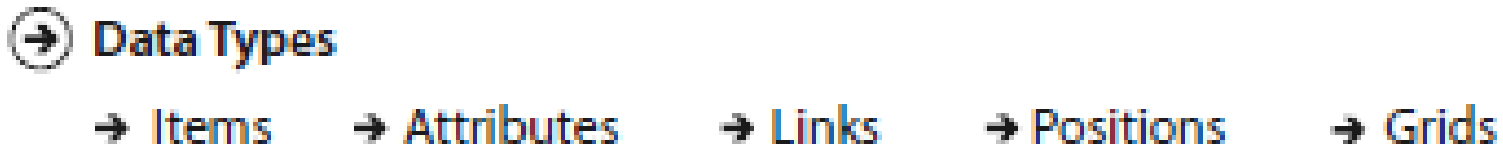
Lecture 2

9/1/2020

Data

- What is data?
- Semantics: real-world meaning of data
- Type: structural or mathematical
  - Structural – dataset type (table, network…)
  - Mathematical:
    - attributes - quantity (count) vs code (category)
    - Appropriate mathematical operations
- Both often require metadata
  - Sometimes we can infer some of this information
  - Line between data and metadata isn't always clear

Data Types



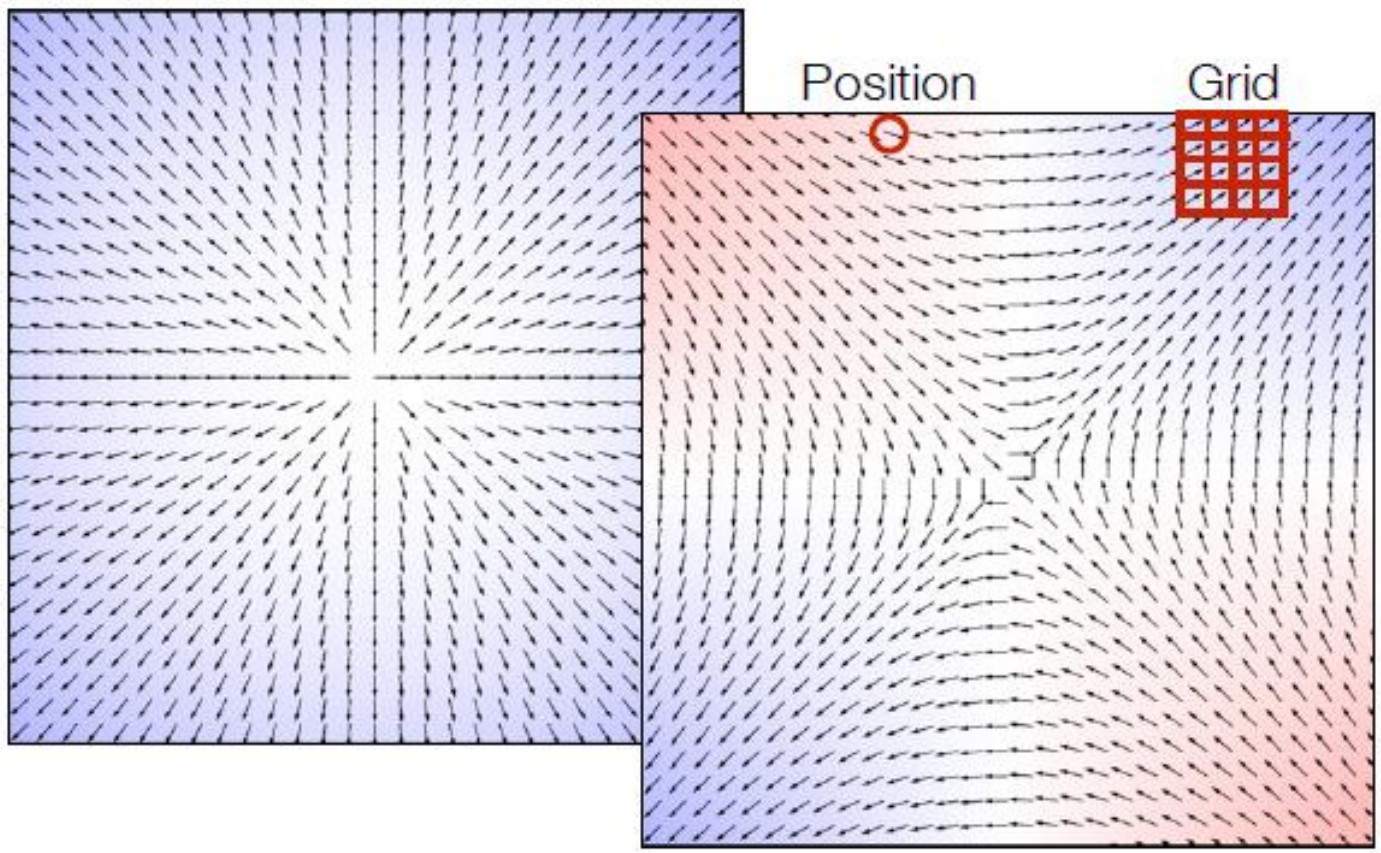
Data Types – Items and Attributes

- Items
  - An item is an individual discrete entity
  - For example, a row in a table, node in a network
- Attributes
  - An attribute is some specific property that can be measured, observed, or logged
  - Also known as: variable, (data) dimension
  - For example: salary, price, number of sales, temperature, protein expression levels

Data Types – Links (and Nodes)

- Nodes
  - Synonym for item but in context of networks (graphs)
- Links
  - Relation between two items
  - Ex: social network friends, computer network links

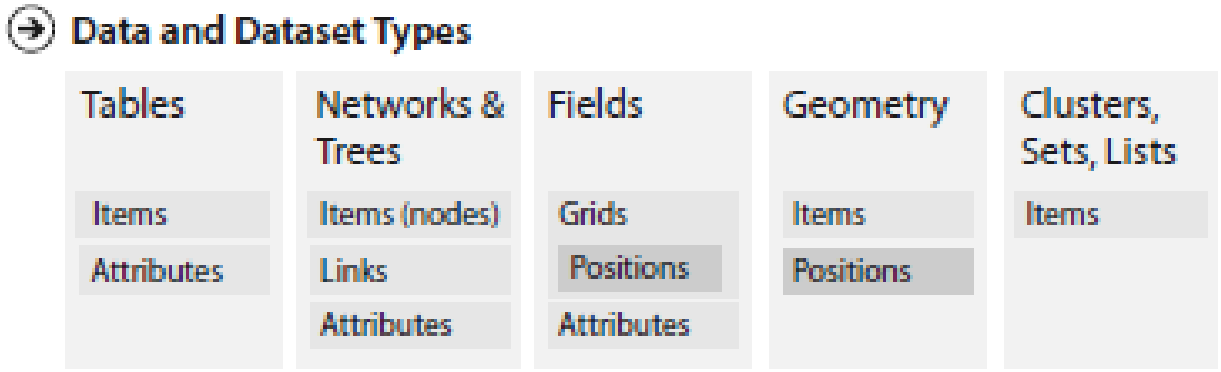
Data Types – Positions and Grids



Data Types – Positions and Grids

- Positions
  - A position is a location in space (usually 2D or 3D)
  - May be subject to projections
  - Ex: cities on a map, a sampled region in a CT scan
- Grids
  - A grid specifies how data is sampled both geometrically and topologically
  - Ex: how CT scan data is stored

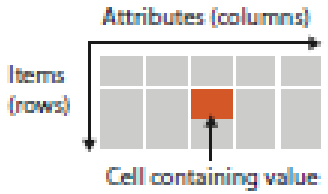
Dataset Types



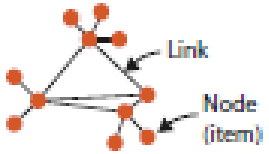
Dataset Types

Dataset Types

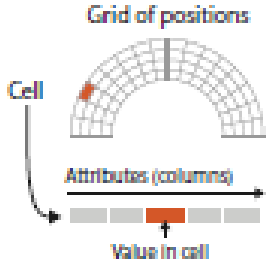
Tables



Networks

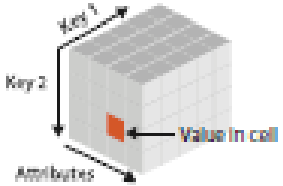


Fields (Continuous)

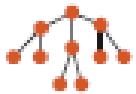


Geometry (Spatial)

Multidimensional Table



Trees



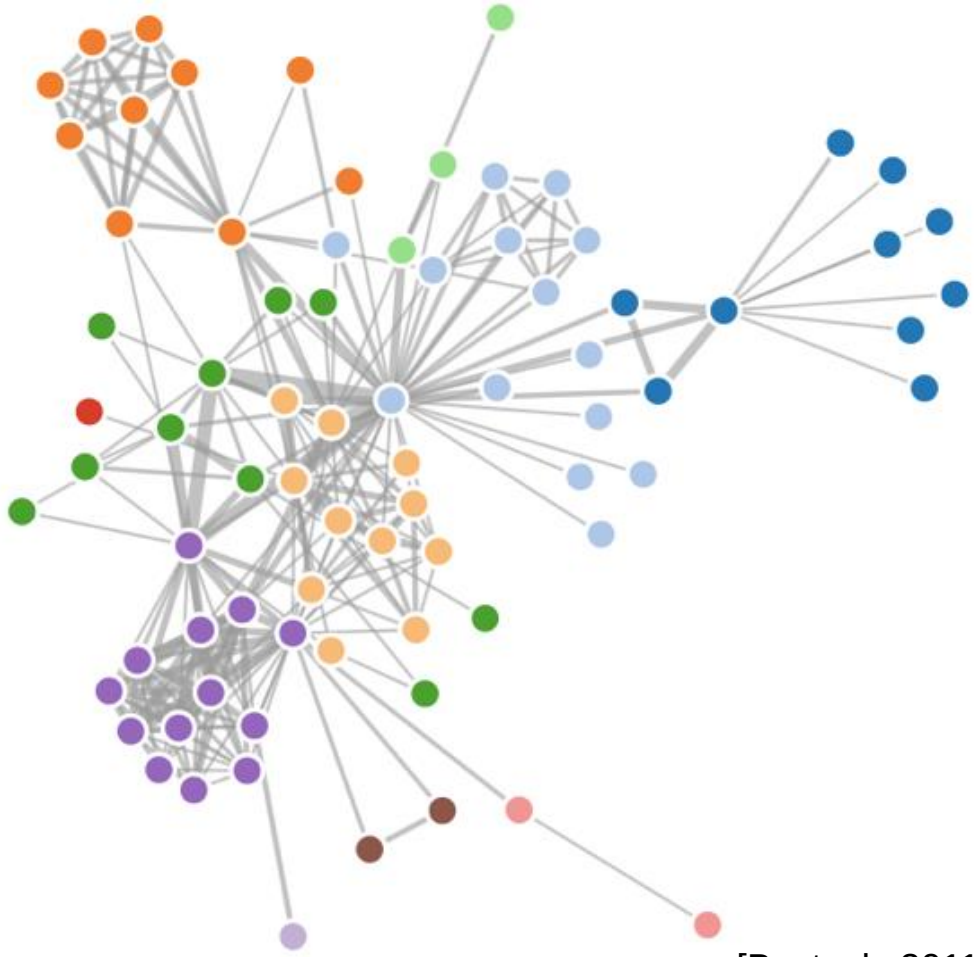
Dataset Types - Tables

- Data Types:
  - Items
  - Attributes
- Table Types:
  - Flat
  - Multidimensional

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	1/20/05	5-Low	Wrap Bag	0.56	1/20/05
69	6/6/05	4-Not Specified	Small Pack	0.44	6/6/05
69	6/6/05	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

Dataset Types - Networks

- Data Types
  - Nodes
    - Synonym for item but in the context of networks (graphs)
  - Links
    - A link is a relation between two items
    - For example, social network friends, computer network links



[Bostock, 2011]

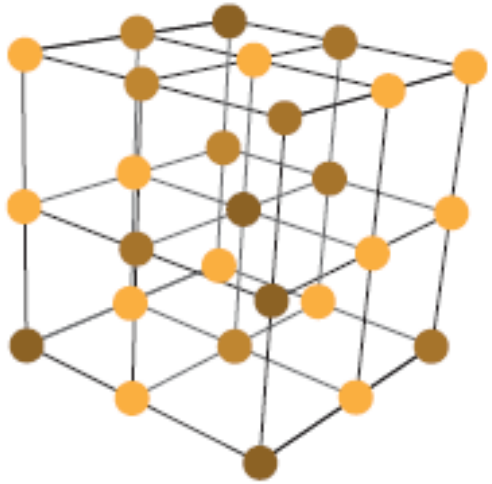
Dataset Types - Trees

- Networks with hierarchical structure
  - Do not have cycles: each child node has only one parent node pointing to it
  - Examples: organization chart of a company, evolutionary relationships between species in the biological tree of life

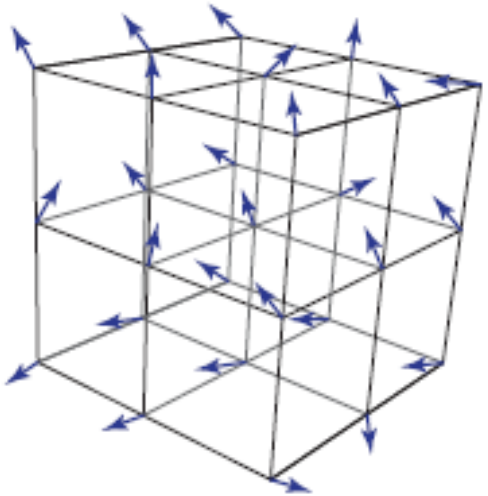
Dataset Types - Fields

- Contains attribute values associated with cells
- Each cell in a field contains measurements or calculations from a continuous domain
  - Conceptually infinitely many values that you might measure, so you could always take a new measurement between any two existing ones
- Examples: temperature, pressure, speed, force, and density
- Sampling – how frequently to take measurements
- Interpolation – how to show values in between the sampled points that is no misleading
  - appropriate interpolation between measurements allows you to reconstruct a new view of the data from an arbitrary viewpoint that is faithful to what you measured
  - In contrast, table and network datatypes are examples of discrete data
    - finite number of individual items exist
    - Interpolation is not a meaningful concept here

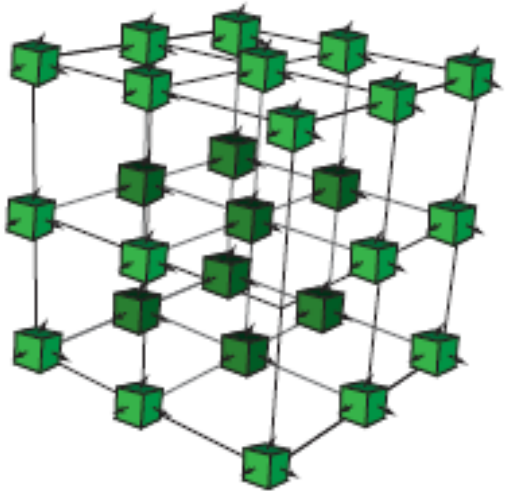
Dataset Types - Fields



Scalar Fields



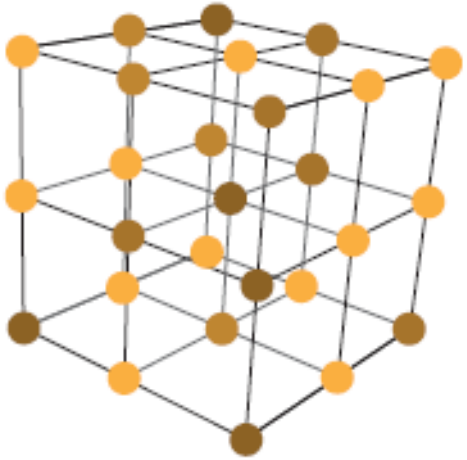
Vector Fields



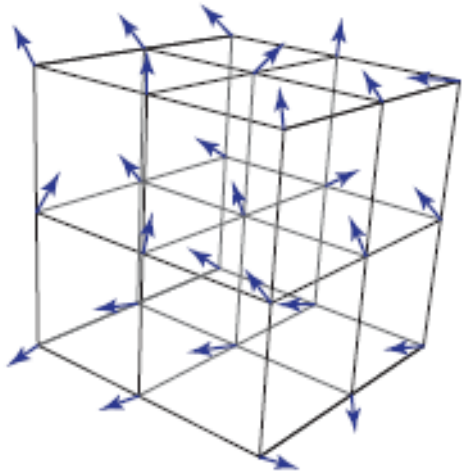
Tensor Fields

Each point in space has an associated...

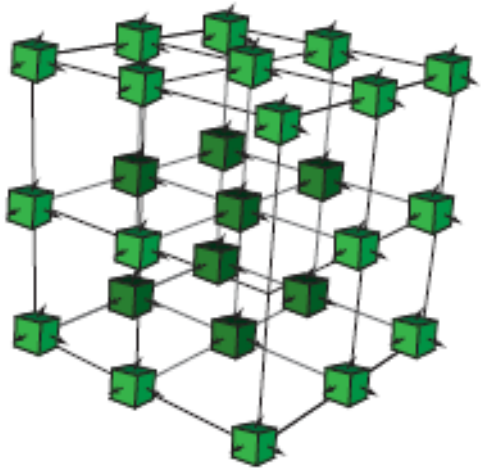
Dataset Types - Fields



Scalar Fields
(Order-0 Tensor Fields)



Vector Fields
(Order-1 Tensor Fields)



Tensor Fields
(Order-2+)

Each point in space has an associated...

s_0

Scalar

$$\begin{bmatrix} v_0 \\ v_1 \\ v_2 \end{bmatrix}$$

Vector

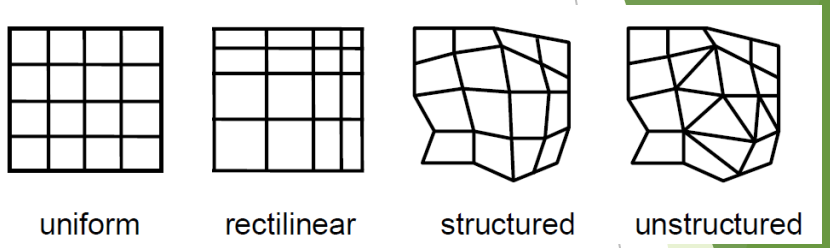
$$\begin{bmatrix} \sigma_{00} & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_{11} & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_{22} \end{bmatrix}$$

Tensor

Dataset Types – Fields – Spatial Fields

- Two subfields of visualization
  - Scivis (scientific visualization)
    - data where the spatial position is given with data; applied to data with an intrinsic spatial layout
    - Usually continuous data
    - Often displaying physical phenomena
    - Ex: flow simulation in 3D space
  - Infovis (information visualization)
    - Data has no set spatial representation
    - Usually discrete type data
    - Designer chooses how to visually represent data
    - Ex. Graphs of web links

Dataset Types – Fields - Grids



- Sampling at completely regular intervals
- Rectilinear
  - Supports non-uniform sampling
- Structured
  - allows curvilinear shapes, where the geometric location of each cell needs to be specified
- Unstructured
  - provide complete flexibility, but the topological information about how the cells connect to each other must be stored explicitly in addition to their spatial positions

[Weiskopf, Machiraju, Möller]

Dataset Types - Geometry

- Specifies information about the shape of items with explicit special positions
- Items can be:
  - Points
  - 1D lines or curves
  - 2D surfaces or regions
  - 3D volumes
- Intrinsically spatial, and like spatial fields they typically occur in the context of tasks that require shape understanding
- Sometimes shown alone (especially when shape understanding is the primary task
- Other times, it is a backdrop against which additional information is overlaid

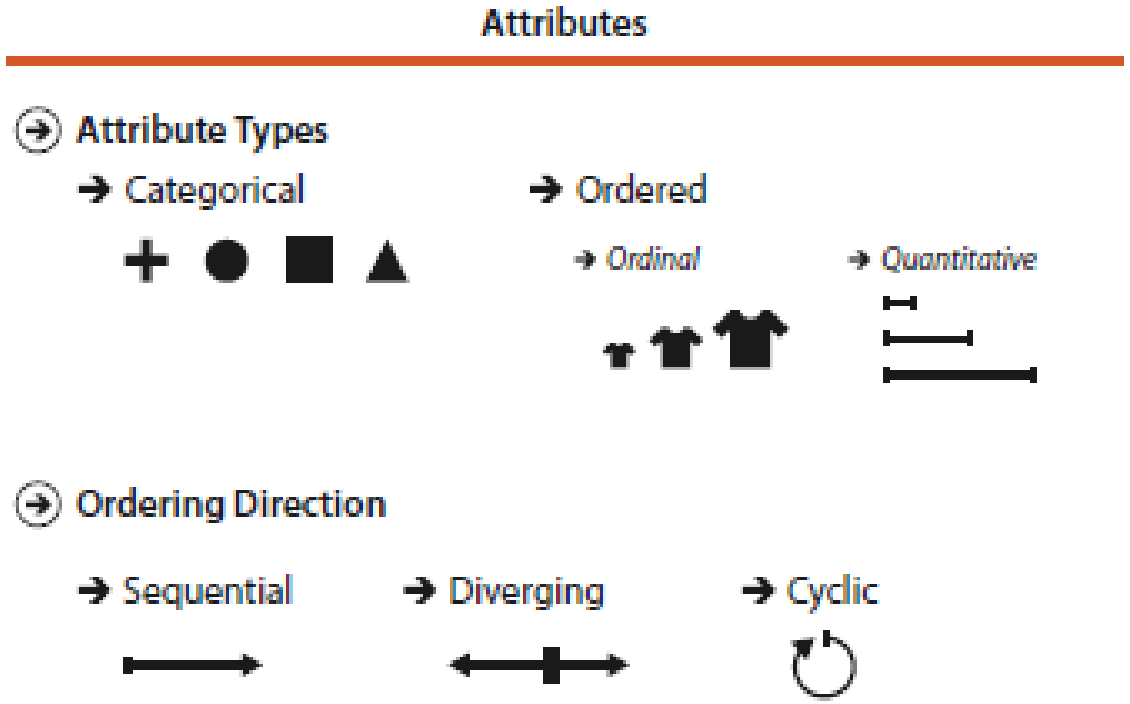
Other Combinations to Group Multiple Items

- Tables
  - Sets
    - Unordered group of items
  - Lists
    - Group of items with a specified ordering
    - Aka array in computer science
  - Cluster
    - Grouping based on attribute similarity, where items within a cluster are more similar to each other than to ones in another cluster
- Networks
  - Compound networks

Dataset Availability

- Static
  - Available all at once
  - Ex: Iris dataset
- Dynamic
  - Dataset info trickles in over the course of the vis session
    - Time varying
    - Stream type
  - Ex: real-time customer count at register

Attribute Types



Categorical, Ordinal, and Quantitative

Exercise: which attributes are categorical, ordinal, and quantitative?

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 Spec		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

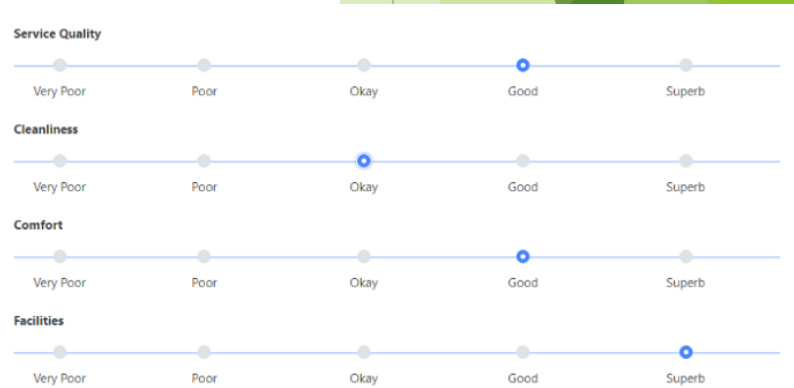
quantitative
ordinal
categorical

Categorical, Ordinal, and Quantitative

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	quantitative	0.6	6/6/05
70	12/18/06	5-Low	ordinal	0.59	12/23/06
70	12/18/06	5-Low	categorical	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

Attribute Types (from a stats perspective)

- Quantitative vs Qualitative
  - Quantitative
    - Numerical data, measurable, objective
    - Discrete vs Continuous
  - Qualitative
    - Non-numerical data, usually subjective
    - Categorical data
- Continuous: can be broken down into smaller units
  - Examples: height (72.5 inches), weight (120.4 lbs), …
- Discrete: whole number integer
  - Examples: number of dogs you have, number of trips to the grocery store in a month
- Categorical
    - Nominal: gender, eye color, state
    - Ordinal: Likert Scale



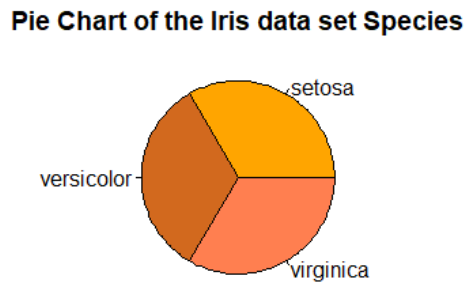
An example of a simple Likert scale

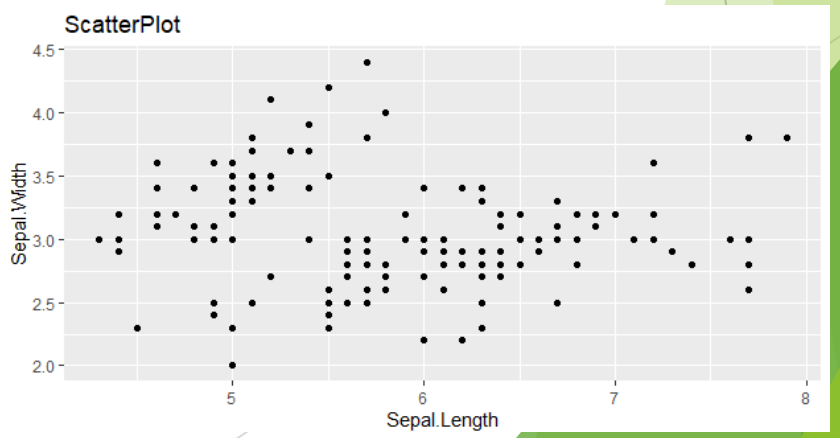
Semantics

- A key attribute acts as an index that is used to look up value attributes
- Key attribute
  - Independent (stats)
  - Dimension (data warehouses)
- Value attribute
  - Dependent (stats)
  - Measure (data warehouse)

Variety of Visualizations

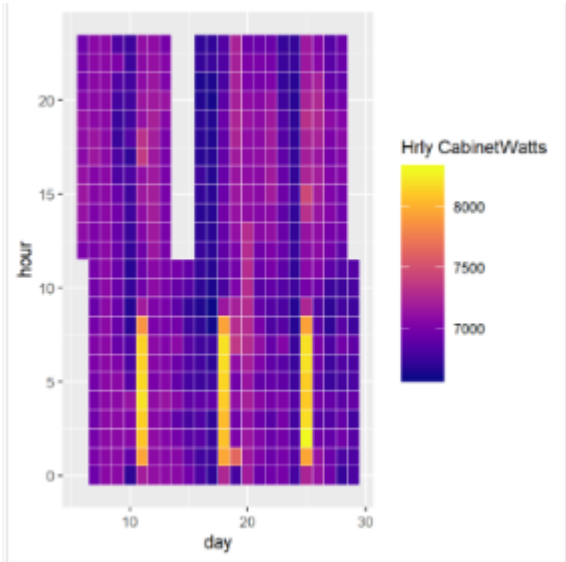
- Pie Charts
  - Percentages and proportions
- Bar Chart
  - Horizontal/vertical, discrete measurements of categorical variables
- Stacked Bar Graph
  - Larger category subdivided into smaller sections, impact of smaller units on total
- Line Graph
  - Measurements over a continuous interval or time period
- Scatterplot
  - Plot two variables, one on each axis, to view a relationship





Variety of Visualizations continued

- Box and Whisker Plots
  - Data distribution, quartiles, extremes, outliers
  - Comparison across groups/datasets easier to view
- Histograms
  - Data distribution over a continuous interval
- Heatmaps
  - Using color to show variance/relationship patterns across multiple variables or over time
  - Compare temp changes throughout the year across multiple cities (data centers)
- Tree Diagram
  - Hierarchy
  - Classification
  - Output of a regression tree model
- Tree Maps, Bubble Charts, and the list goes on…



Data

- Iris dataset found in R base package
- Excel
  - Pivot Table and Chart
- R
  - ggplot2 vignette
    - https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf
  - Basic plots
    - Using the base packages
  - ggplot2 plotting

Iris: setosa, versicolor, virginica

Iris setosa

Plants

Iris setosa, is a species in the genus Iris, it is also in the subgenus Limniris and in the series Tripetalae. It is a rhizomatous perennial from a wide range across the Arctic sea, including Alaska, Maine, Canada, Russia, northeastern Asia, China, Korea and southwards to Japan. Wikipedia

Scientific name: Iris setosa

Rank: Species

Higher classification: Iris subg. Limniris

Northern blue flag

Plants

Iris versicolor is also commonly known as the blue flag, harlequin blueflag, larger blue flag, northern blue flag, and poison flag, plus other variations of these names, and in Britain and Ireland as purple iris. It is a species of Iris native to North America, in the Eastern United States and Eastern Canada. Wikipedia

Scientific name: Iris versicolor

Higher classification: Irises

Rank: Species

Native Range: Eastern United States missouribotanicalgarden.org

Spread: 2.00 to 2.50 feet missouribotanicalgarden.org

Common Name: blue flag missouribotanicalgarden.org

Iris virginica

Plants

Iris virginica, with the common name Virginia iris, is a perennial species of flowering plant, native to eastern North America. It is common along the coastal plain from Florida to Georgia in the Southeastern United States. Wikipedia

Scientific name: Iris virginica

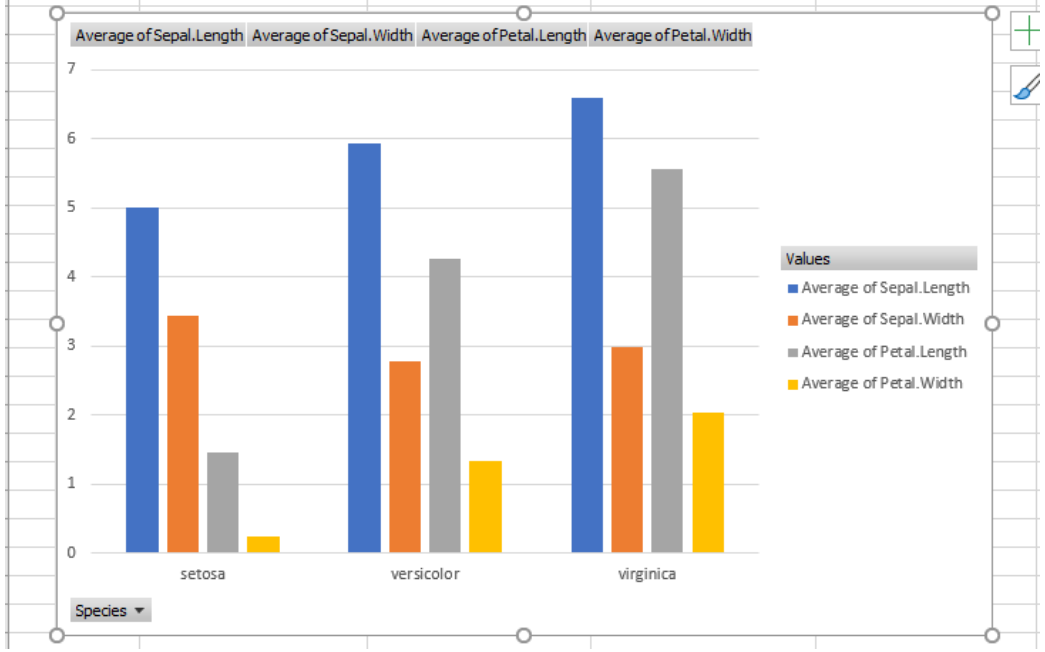
Higher classification: Irises

Rank: Species

Family: Iridaceae

Excel Pivot Table and Chart Demo

Row Labels	Average of Sepal.Length	Average of Sepal.Width	Average of Petal.Length	Average of Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.77	4.26	1.326
virginica	6.588	2.974	5.552	2.026
Grand Total	**5.843333333**	**3.057333333**	**3.758**	**1.199333333**



Sources/Credits

- Tamara Munzner, Visualization Analysis & Design, A K Peters Visualization Series, CRC Press, 2014.
- Utah, Miriah Meyer, Visualization (2014).
- UMass Dartmouth, David Koop, Data Visualization (2015).