DS4200: Information Presentation and Visualization

Task abstraction and Visual encoding

Xiaoyi Yang
Khoury College of Computer Sciences
Northeastern University

Goals for Today

- What "Task abstraction" is and why it is so important.
- The differences between high, mid, and low level task classifications.
- How to classify tasks.
- How to pick appropriate visual representations based on attribute type and perceptual properties

Review on Data abstraction

What?

Datasets

Data Types

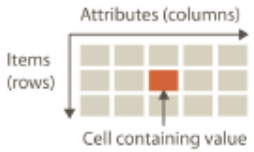
→ Items → Attributes → Links → Positions → Grids

Data and Dataset Types

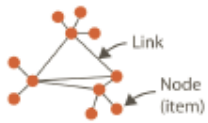
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

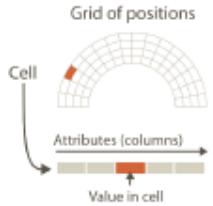
→ Tables



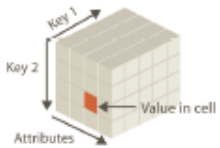
→ Networks



→ Fields (Continuous)



→ *Multidimensional Table*



→ *Trees*

Attributes

Attribute Types

→ Categorical

→ Ordered

→ *Ordinal*

→ *Quantitative*

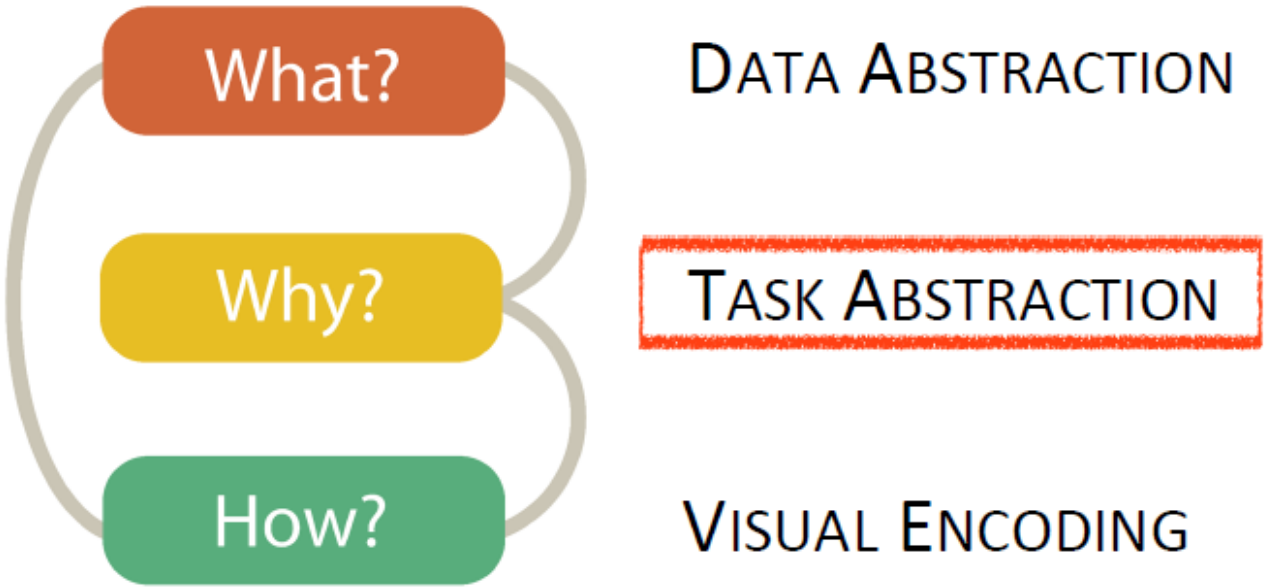
Ordering Direction

→ Sequential

→ Diverging

→ Cyclic

Data analysis



Task abstraction

Task: a specific goal or objective that a user or analyst wants to achieve by examining and interpreting data through visual representation

Task abstraction: the process of defining and simplifying the specific tasks or goals a visualization should achieve

Task abstraction

Task: a specific goal or objective that a user or analyst wants to achieve by examining and interpreting data through visual representation

Task abstraction: the process of defining and simplifying the specific tasks or goals a visualization should achieve

Why abstract?

Task abstraction example

Consider following two cases:

- Sales performance: A manager is supervising the sale performance for multiple stores. Each month, each store need to report the total sales, revenue, and customer retention for the manager to compare.
- Scientific research: A researcher is working on a project to contrast the effect of a medicine with different levels of dose. He will record the blood pressure, blood glucose levels and cholesterol for each group of patients.

What is the similarity and difference here?

Task abstraction example

Both examples try to compare some metrics across different group.

Task abstraction example

Both examples try to **compare** *some metrics across different group*.

- **Verb.**: Actions
- *Noun.*: Target

Task abstraction example

Both examples try to **compare** *some metrics across different group*.

- **Verb.**: Actions
- *Noun.*: Target

To describe complex activities, you can specify a chained sequence of tasks, where the output of one becomes the input to the next.

Why abstract?

- Better understand user's task
- Avoids domain specific terms thus easier to apply to other cases (broadly applicable results).
- understand whether and how to transform the user's original data into different forms by deriving new data.

Action: define user goals.

There are three levels of the action:

- High-level: Analyze, either to consume existing data or to also produce additional data or identity the overall trend in the data
- Mid-level: Search or filter, looking for target or location
- low-level: Query, like compare or summarize, or look for a specific pattern

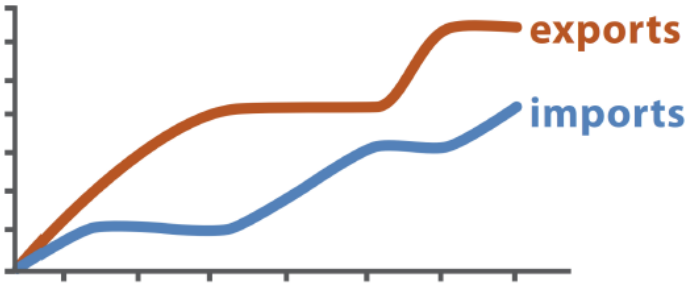
The choices at each of these three levels are independent from each other, and it's usually useful to describe actions at all three of them.

High level: Analysis

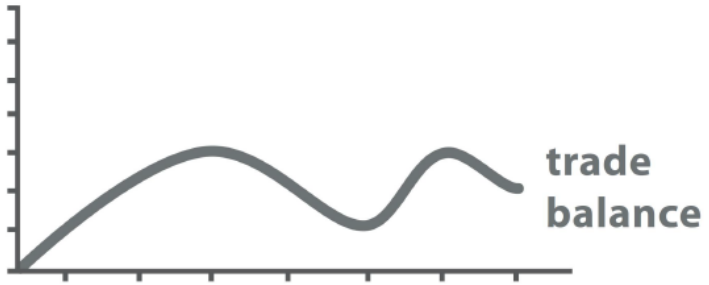
- Analyze
  - Consume
    - *Discover*
    - *Present*
    - *Enjoy*

  - Produce
    - *Annotate*
    - *Record*
    - *Derive*

High level: Analysis



Original Data



trade balance = *exports* – *imports*

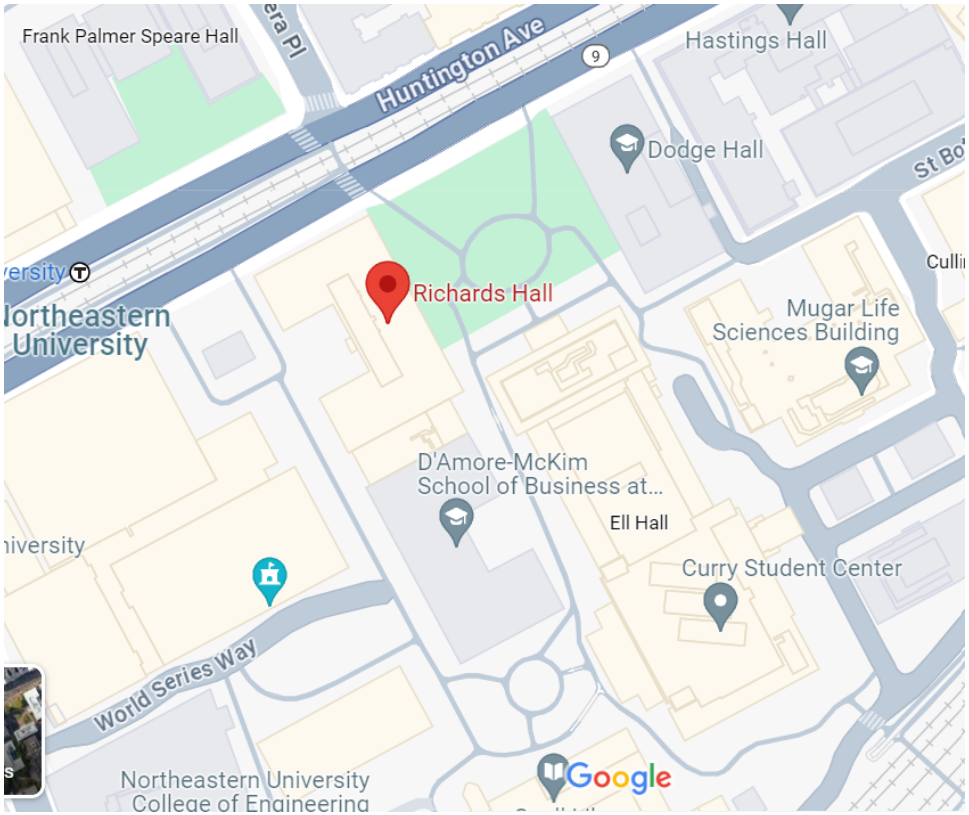
Derived Data

Mid level: Search

Search

	Target known	Target unknown
Location known	 *Lookup*	 *Browse*
Location unknown	 *Locate*	 *Explore*

Mid level: Search



Search

	Target known	Target unknown
Location known	*Lookup*	*Browse*
Location unknown	*Locate*	*Explore*

- What is the address?
- Where is Richards Hall?
- What buildings are near Richards Hall?
- What is in the center of the campus?

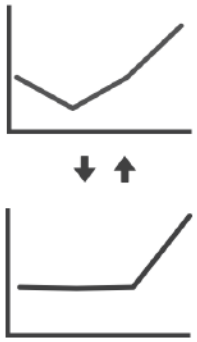
Low level: Query

Query

- Identify

single target

- Compare



multiple targets

- Summarize

all targets

Why high/mid/low?

- The word "high" and "low" do not mean "good" and "bad"!
- High-level because they involve broader data exploration and interpretation
- Low-level because they are more specific and involve precise data retrieval or manipulation
- Low level has lots of other task taxonomies...
- Both types of tasks play crucial roles in data visualization!

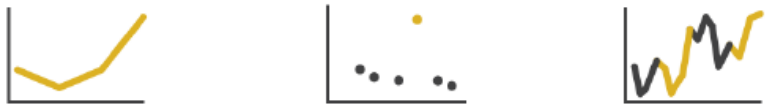
Low-level Task Taxonomy

- Retrieve Value: How long is the movie Gone with the Wind?
- Compute Derived Value: How many awards have MGM studio won in total?
- Find Extremum: What director/film has won the most awards?
- Sort: Rank movies by most number of awards.
- Determine Range: What is the range of film lengths?
- Characterize Distribution: What is the age distribution of actors?
- Find Anomalies: Are there exceptions to the relationship between number of awards won and total movies made by an actor?
- Cluster: Is there a cluster of typical film lengths?
- Correlate: Is there a trend of increasing film length over the years?

Targets

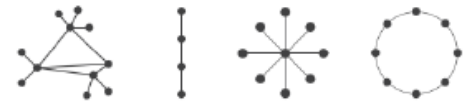
Targets: Aspects of the data that are interest to the user.

- **All Data**
  - Trends
  - Outliers
  - Features



- **Attributes**
  - One
    - *Distribution*
    - *Extremes*
  - Many
    - *Dependency*
    - *Correlation*
    - *Similarity*

- **Network Data**
  - Topology



  - *Paths*

- **Spatial Data**
  - Shape

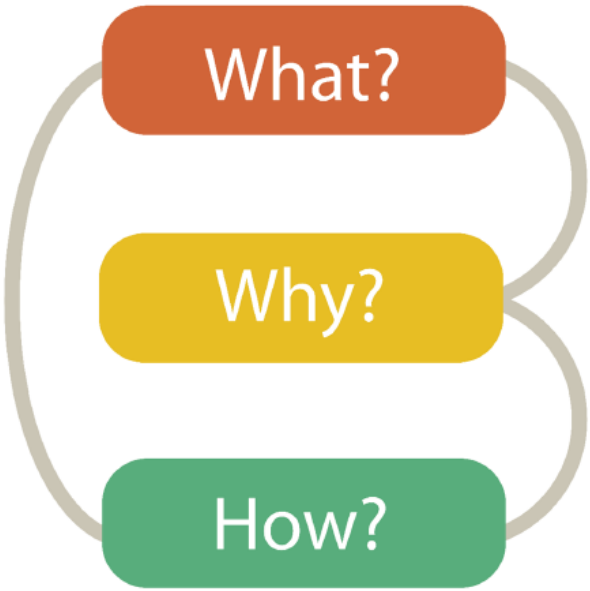
In-class activities

- Scenario: You are provided with a dataset containing information about various movies, including their budget, box office revenue, genre, and rating. The task is to explore relationships between the budget and box office revenue to see if higher-budget movies tend to have higher revenues. Additionally, you need to check if the relationship differs between genres and ratings.
- Scenario: You have a dataset about global temperature change over time, containing information on year, temperature anomaly, CO2 levels, and geographic region. Your task is to explore how the global temperature has changed over time and whether certain regions are experiencing temperature changes differently.
- Scenario: You have a dataset of athletes, including their height, weight, and sport. Your goal is to search for athletes who fall within a specific range of heights and weights, and who participate in a particular sport (e.g., "Basketball").

Link Task and figure choices

Task	Figure Choice	Why
Trend Identification	Line chart, Area chart	Best for showing continuous data over time.
Find Patterns	Scatter plot, Heatmap	Great for finding relationships, clusters, or trends.
Search	Scatter plot with filters Bar chart with filters	Allows searching for specific ranges or conditions.
Compare Values	Bar chart Box plot	Compares categories or groups across variables.
Find Extremes	Histogram, Bar chart Any chart with highlights	Highlights maximum or minimum values easily.
Explore Distribution	Histogram, Box plot	Shows how a variable is distributed across its range.

Data analysis



DATA ABSTRACTION

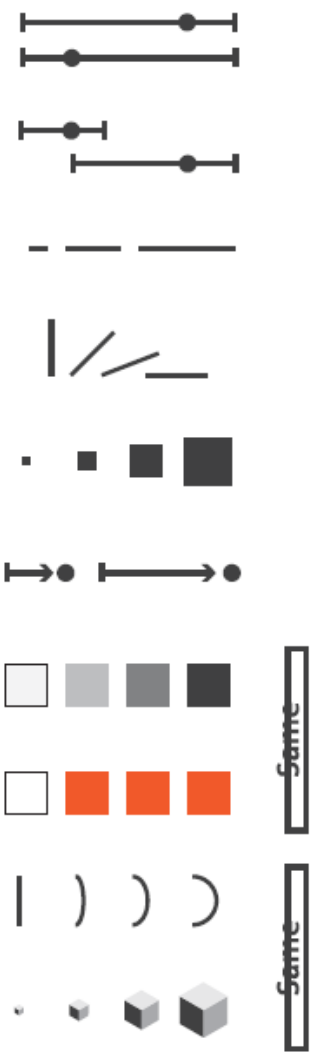
TASK ABSTRACTION

VISUAL ENCODING

Recall: Visualization Building Blocks

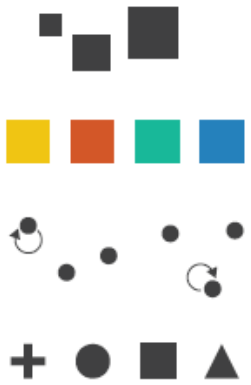
Magnitude Channels: Ordered Attributes

- Position on common scale
- Position on unaligned scale
- Length (1D size)
- Tilt/angle
- Area (2D size)
- Depth (3D position)
- Color luminance
- Color saturation
- Curvature
- Volume (3D size)



Identity Channels: Categorical Attributes

- Spatial region
- Color hue
- Motion
- Shape



Recall: Visualization Building Blocks

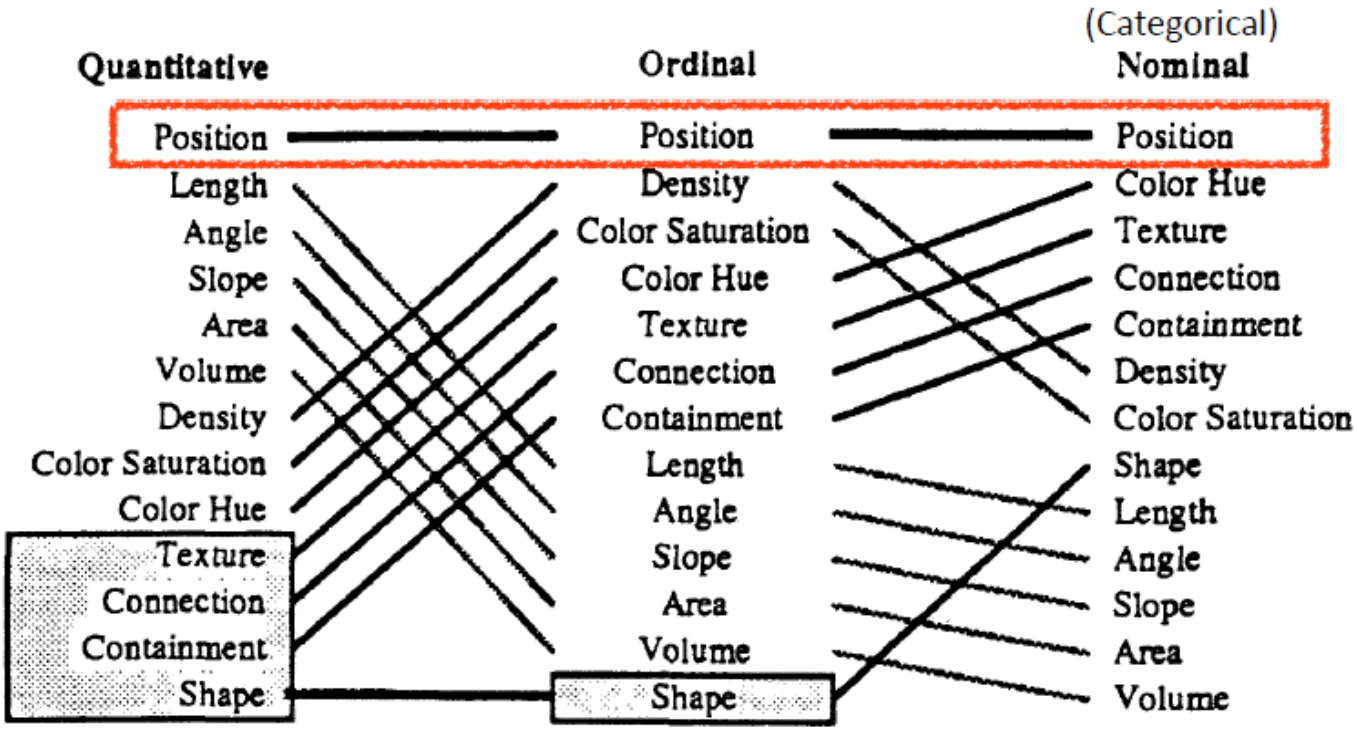


Figure 15: Ranking of Perceptual Tasks. *The tasks shown in the gray boxes are not relevant to that type of data.*

Mackinlay (1986)

Visual encoding

Visual encoding: the use of visual properties such as color, size, shape, position, and texture to represent data values or attributes in a data visualization.

Visual encoding

Visual encoding: the use of visual properties such as color, size, shape, position, and texture to represent data values or attributes in a data visualization.

Other things you can do when designing a visualization:

- Manipulate
- Facet
- Reduce

How?

Encode

- **Arrange**
  - → Express
  - → Separate
  - → Order
  - → Align
  - → Use

- **Map**
  from categorical and ordered attributes
  - → Color
    - → *Hue*
    - → *Saturation*
    - → *Luminance*
  - → Size, Angle, Curvature, ...
  - → Shape
  - → Motion
    Direction, Rate, Frequency, ...

Manipulate

- **Change**

- **Select**

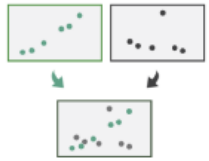
- **Navigate**

Facet

- **Juxtapose**

- **Partition**

- **Superimpose**

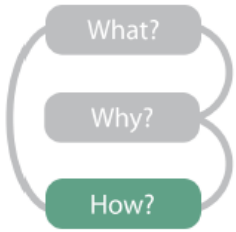


Reduce

- **Filter**

- **Aggregate**

- **Embed**



Arrange Tables

Separate, Order, Align Regions

- Separate

- Order

- Align

- *1 Key*
  List

- *2 Keys*
  Matrix

- *3 Keys*
  Volume

- *Many Keys*
  Recursive Subdivision

- Key: an independent attribute that can be used as a unique index
- Value: a dependent attribute

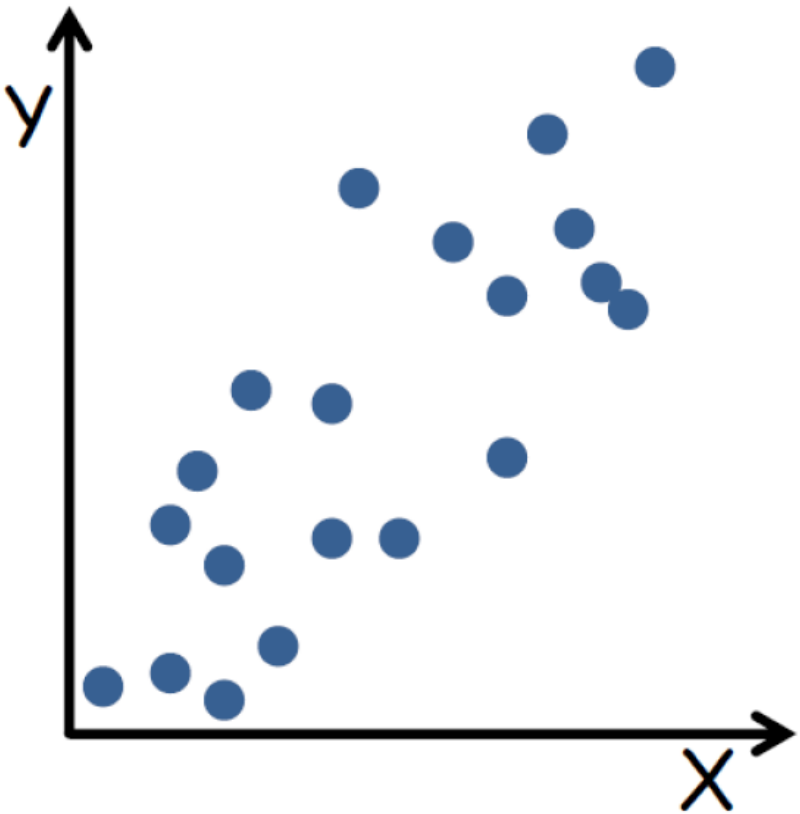
Example keys

Date	Precipitation	High Temperature
May 1, 2016	0”	60
May 2, 2016	0.3”	62
May 3, 2016	1”	55
May 4, 2016	0”	67

Student	College	HW1 grade (out of 10)
John	COS	9
Jane	Khoury	10
June	Khoury	8
Joe	Khoury	8

First question: How many keys do you have?

No key



Scatterplots

- What: two quantitative value attributes.
- How: Express values with horizontal and vertical spatial position and point marks.
- Why: Find trends, outliers, distribution, correlation; locate clusters.

Arrange table

- Spatial position is an ordered magnitude visual channel
- Categorical attributes have unordered identity semantics
- Solution: define spatial region

Separate, Order, Align Regions

- Separate

- Order

- Align

- *1 Key*
  List

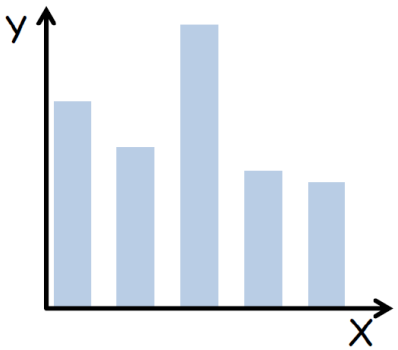
- *2 Keys*
  Matrix

- *3 Keys*
  Volume

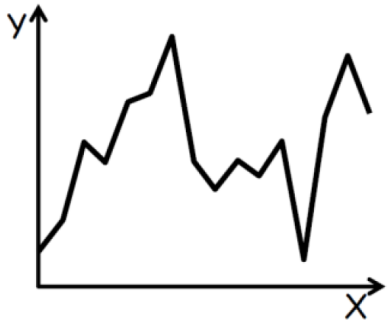
- *Many Keys*
  Recursive Subdivision

One key

One-dimensional list alignment, either horizontal or vertical.



BAR CHART



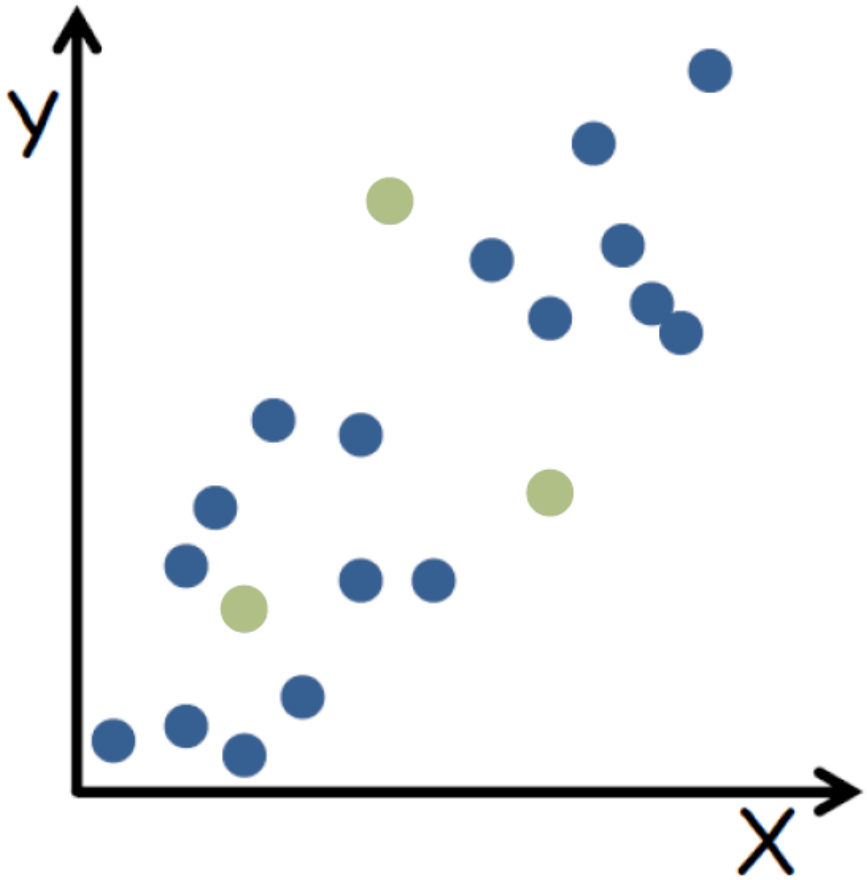
LINE GRAPH

Bar chart

- What: one quantitative value attribute, one categorical key attribute.
- How: Line marks, express value attribute with aligned vertical position, separate key attribute with horizontal position.
- Why: Lookup and compare values.

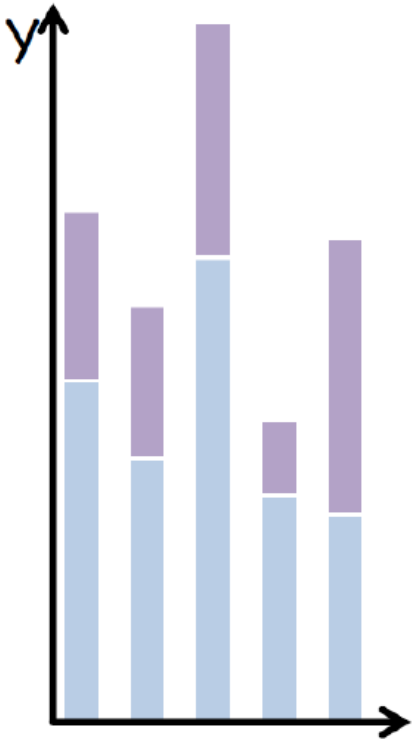
One key

What is "what" / "how" / "why"?

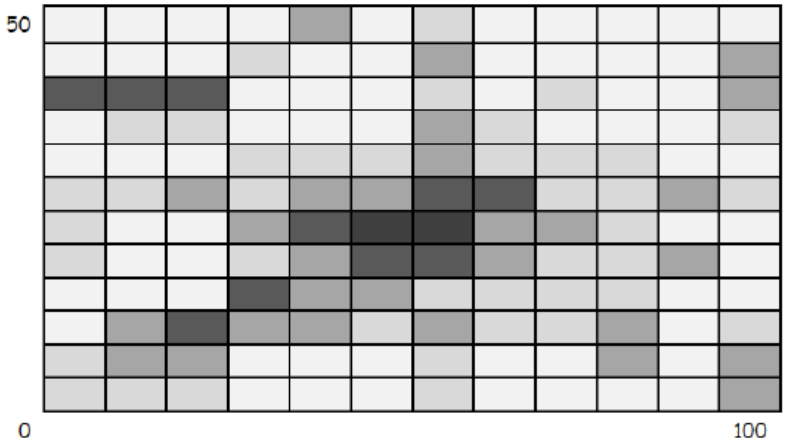


Two keys

What is "what" / "how" / "why"?



STACKED BAR CHART



HEATMAP

Two keys

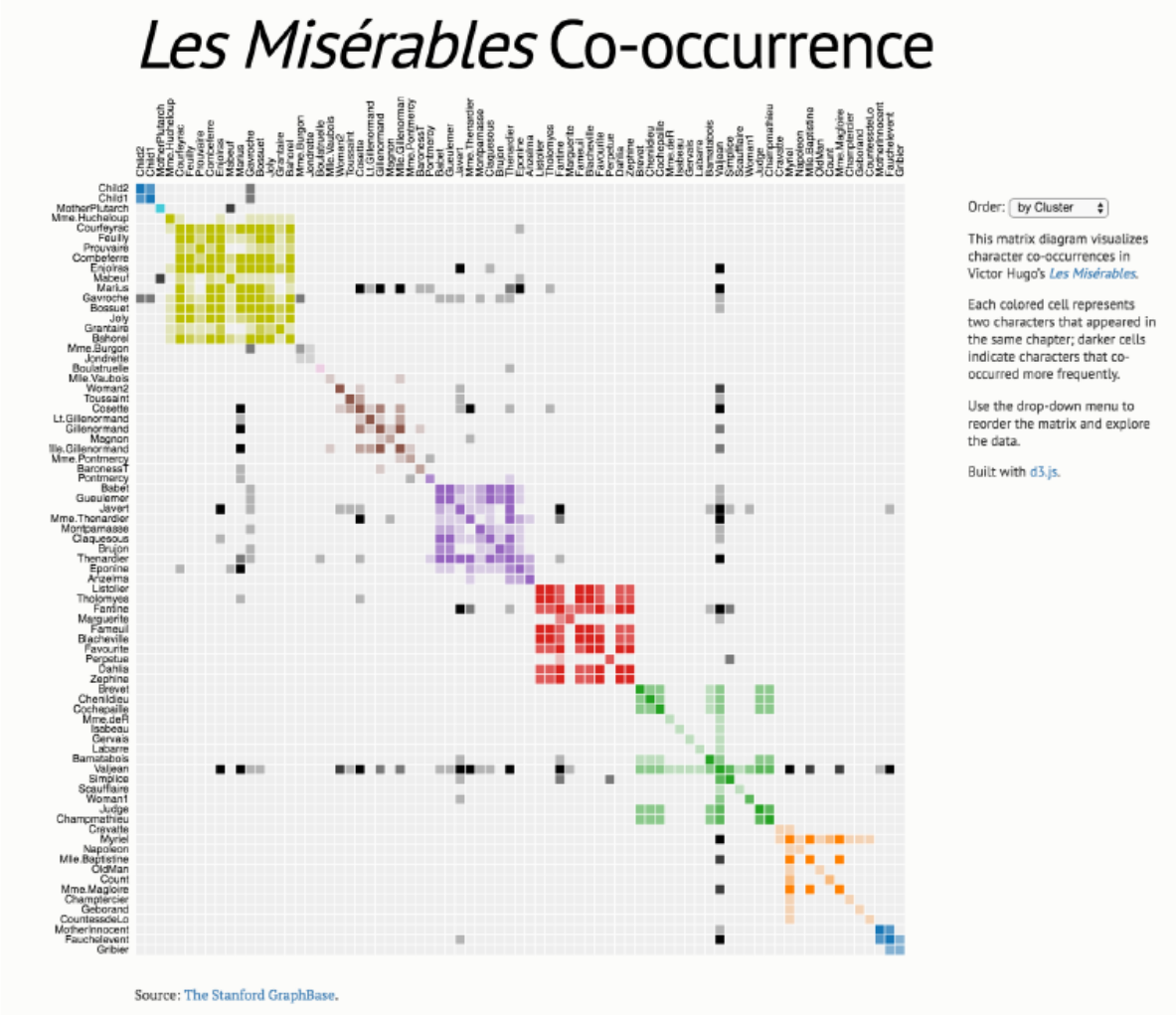
Stacked bar chart

- What: One quantitative value attribute, two categorical key attributes.
- How: Bar glyph with length-coded subcomponents of value attribute for each category of secondary key attribute. Separate bars by category of primary key attribute
- Why: Part-to-whole relationship, lookup values, find trends.

Heatmap

- What: One quantitative value attribute, two categorical key attributes.
- How: 2D matrix alignment of area marks, diverging colormap.
- Why: Find clusters, outliers; summarize.

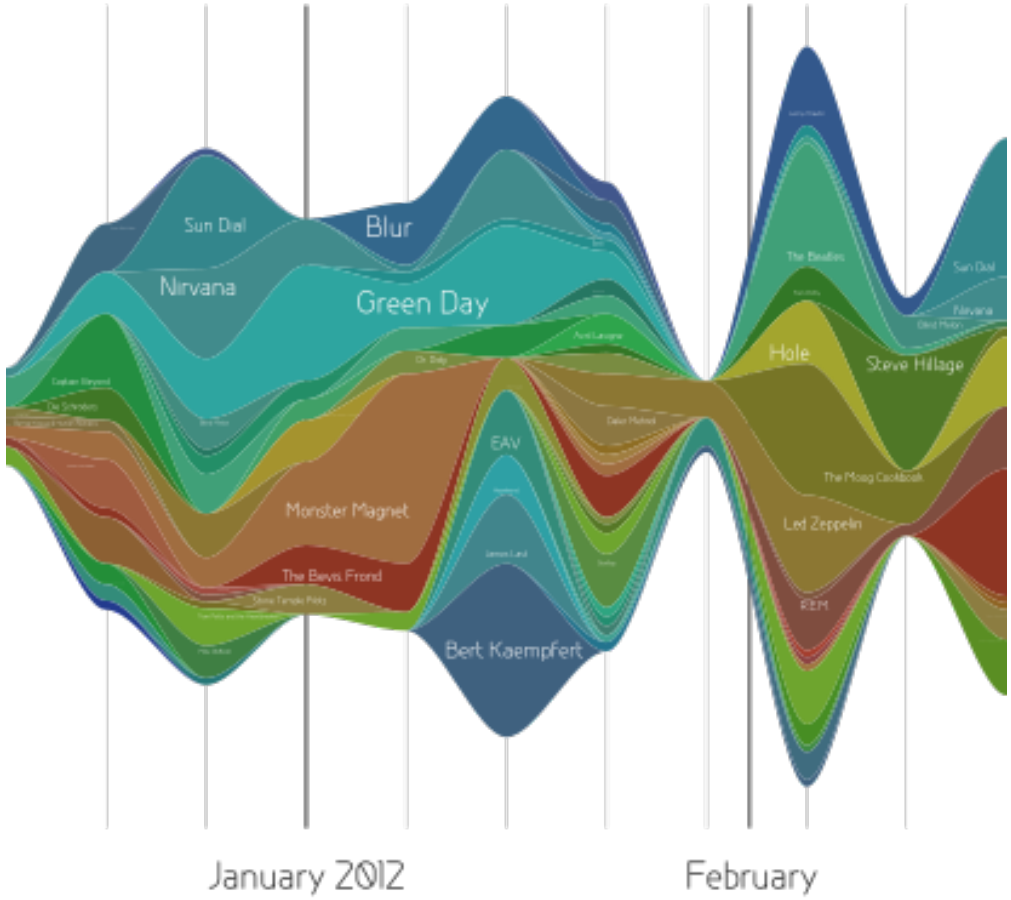
Two keys (Network)



https://bost.ocks.org/mike/miserables/

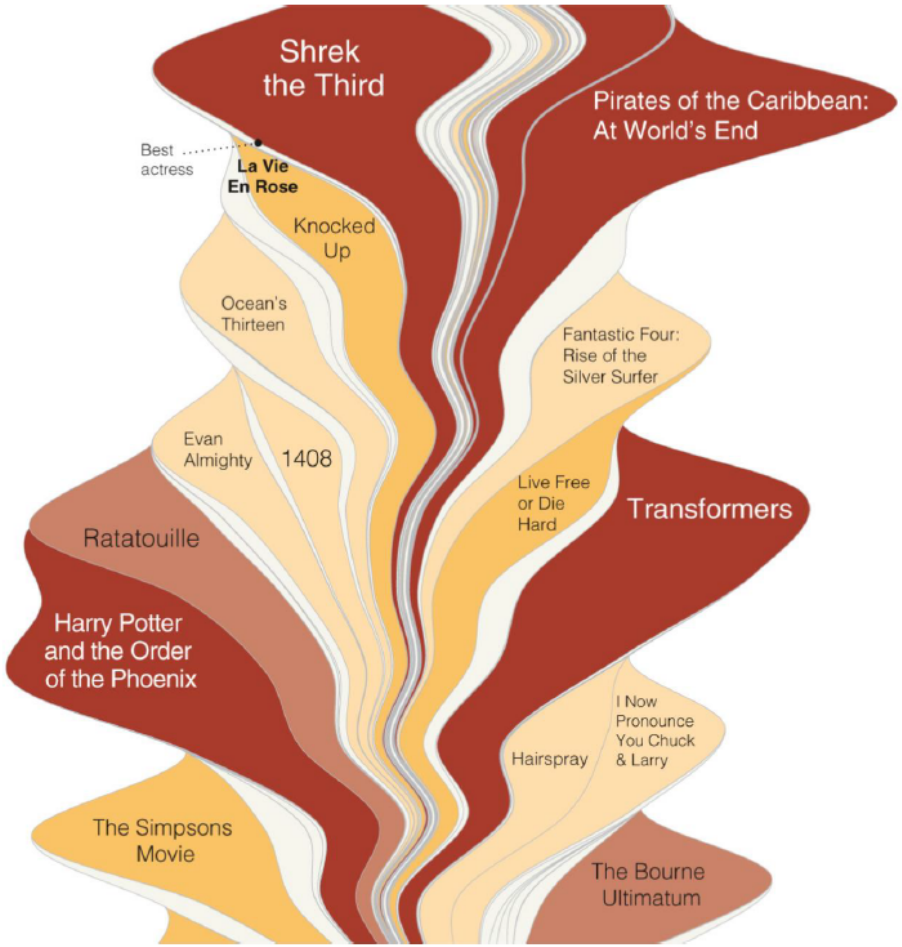
Two keys (Streamgraph)

What is "what" / "how" / "why"?



https://gist.github.com/mbostock/4060954

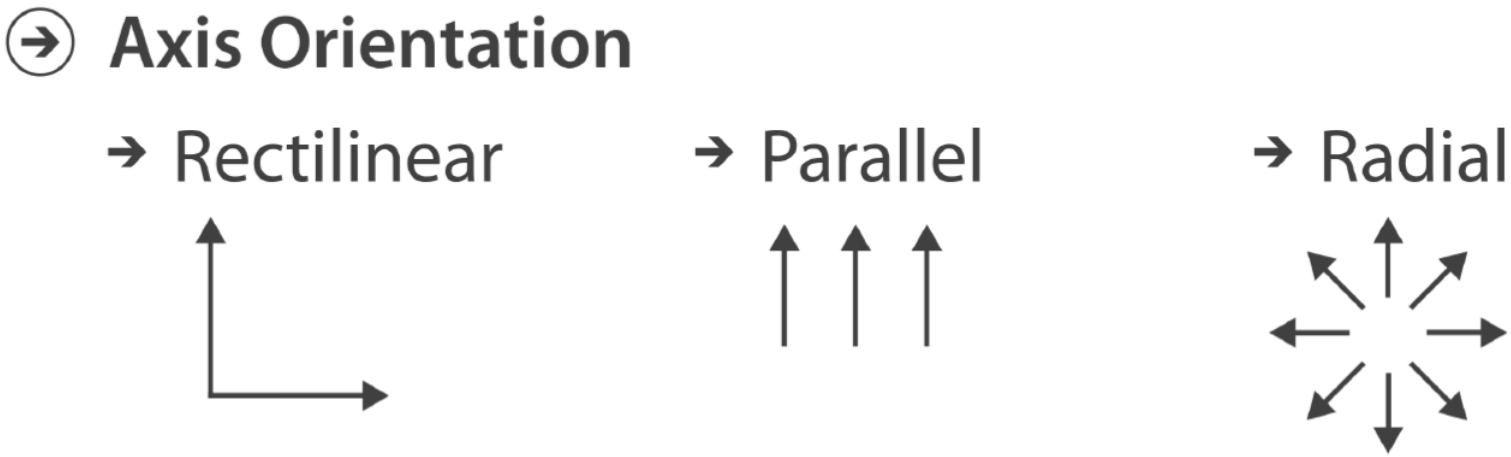
Two keys (Streamgraph)



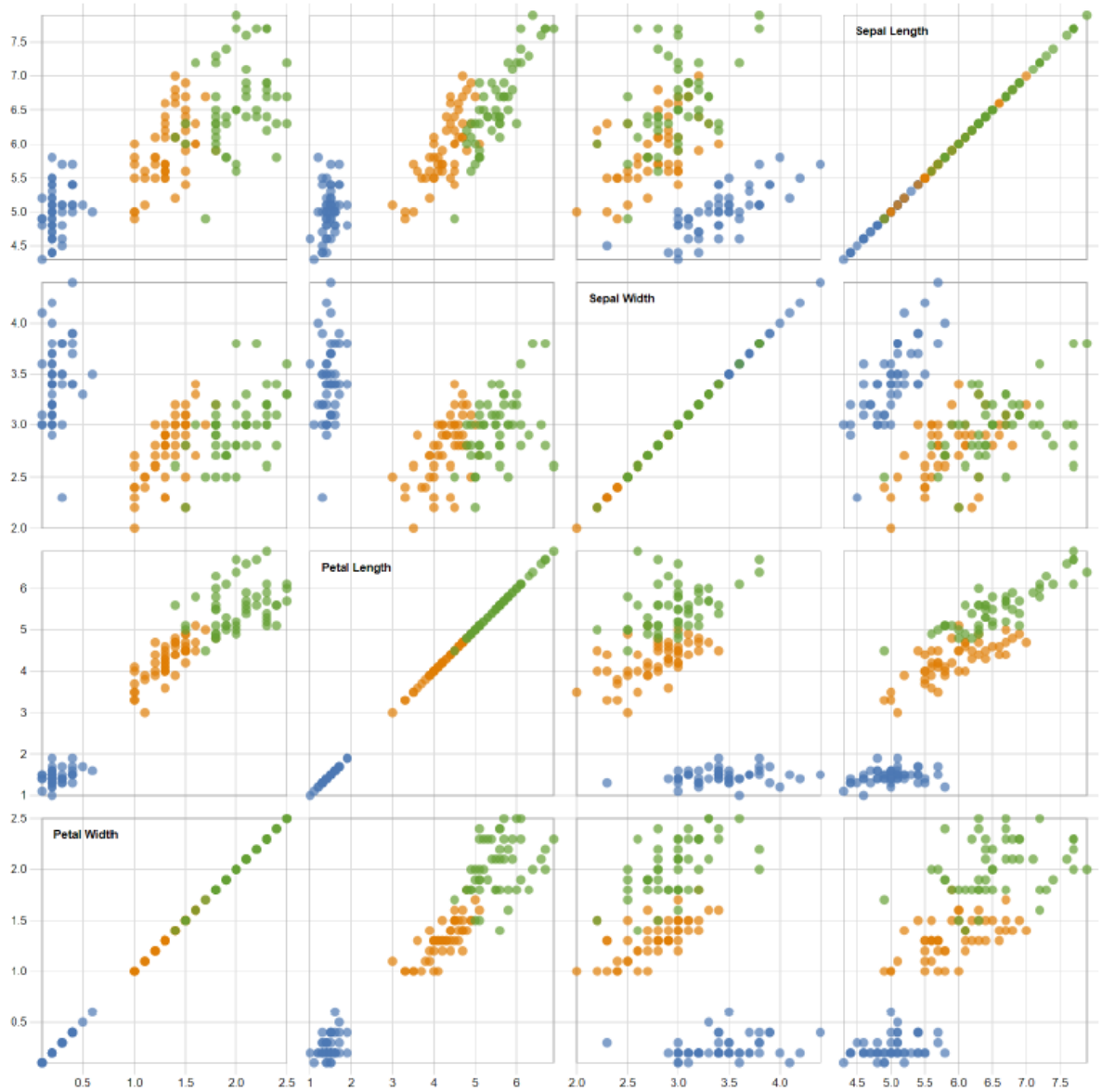
https://leebyron.com/streamgraph/

Extend your design

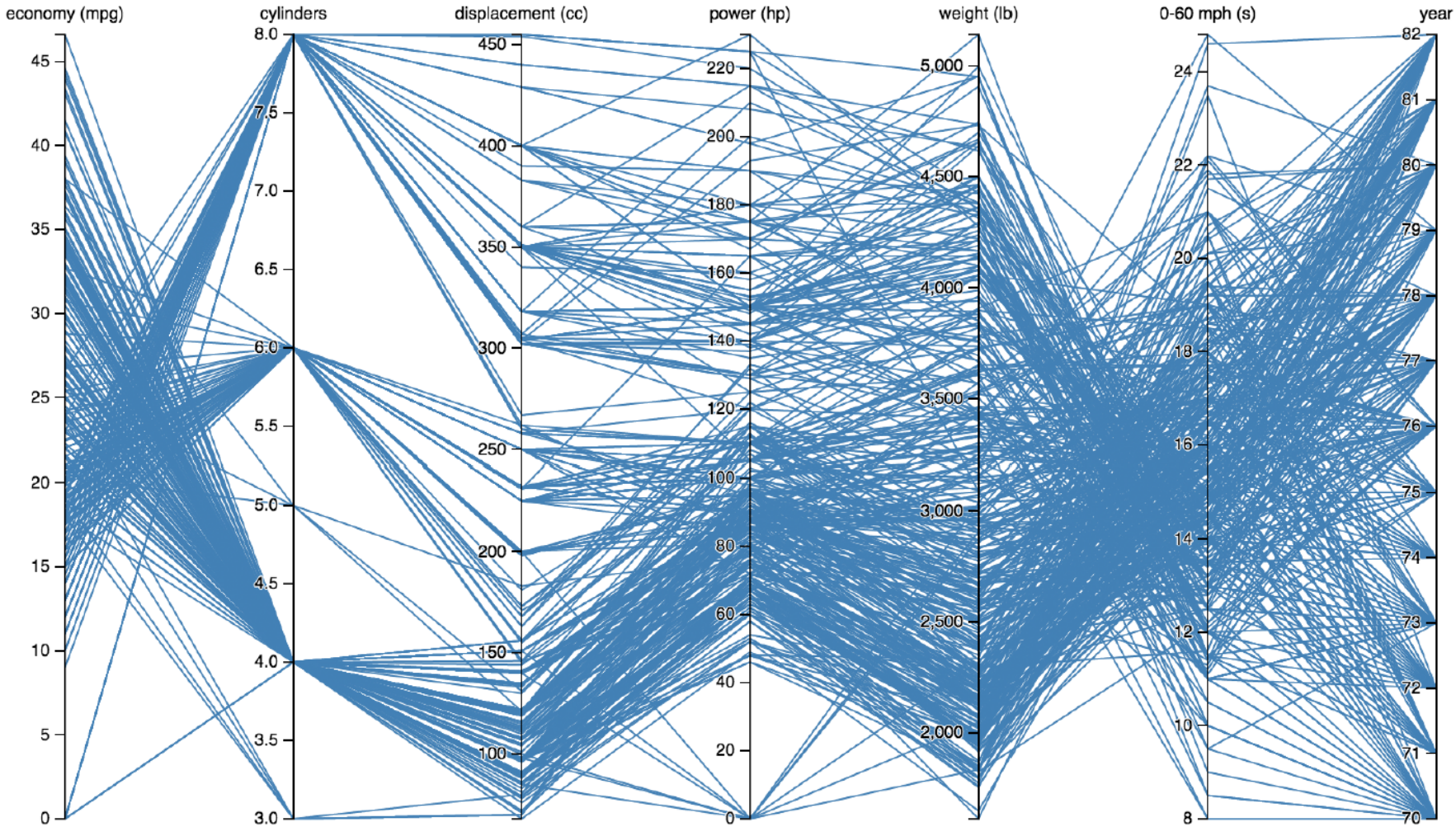
Another choice is to arrange the axis.



Rectilinear design



Parallel design



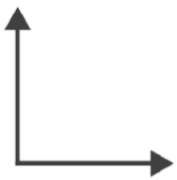
Radial design

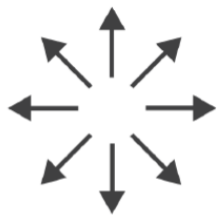
Spatial Axis Orientation

Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

- **Axis Orientation**
  - Rectilinear
  - Parallel
  - Radial



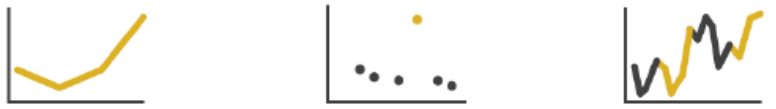


Axis summary

- Rectilinear layouts are more effective at showing the differences between linear and nonlinear trends.
- Parallel layouts are more effective at showing the changes among the keys.
- Radial layouts are more effective at showing cyclic patterns.
- Only Rectilinear layouts are effective when the data size is large

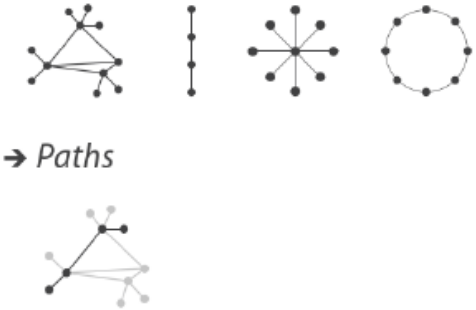
One special case: Distributions & Correlations

- **All Data**
  - Trends
  - Outliers
  - Features



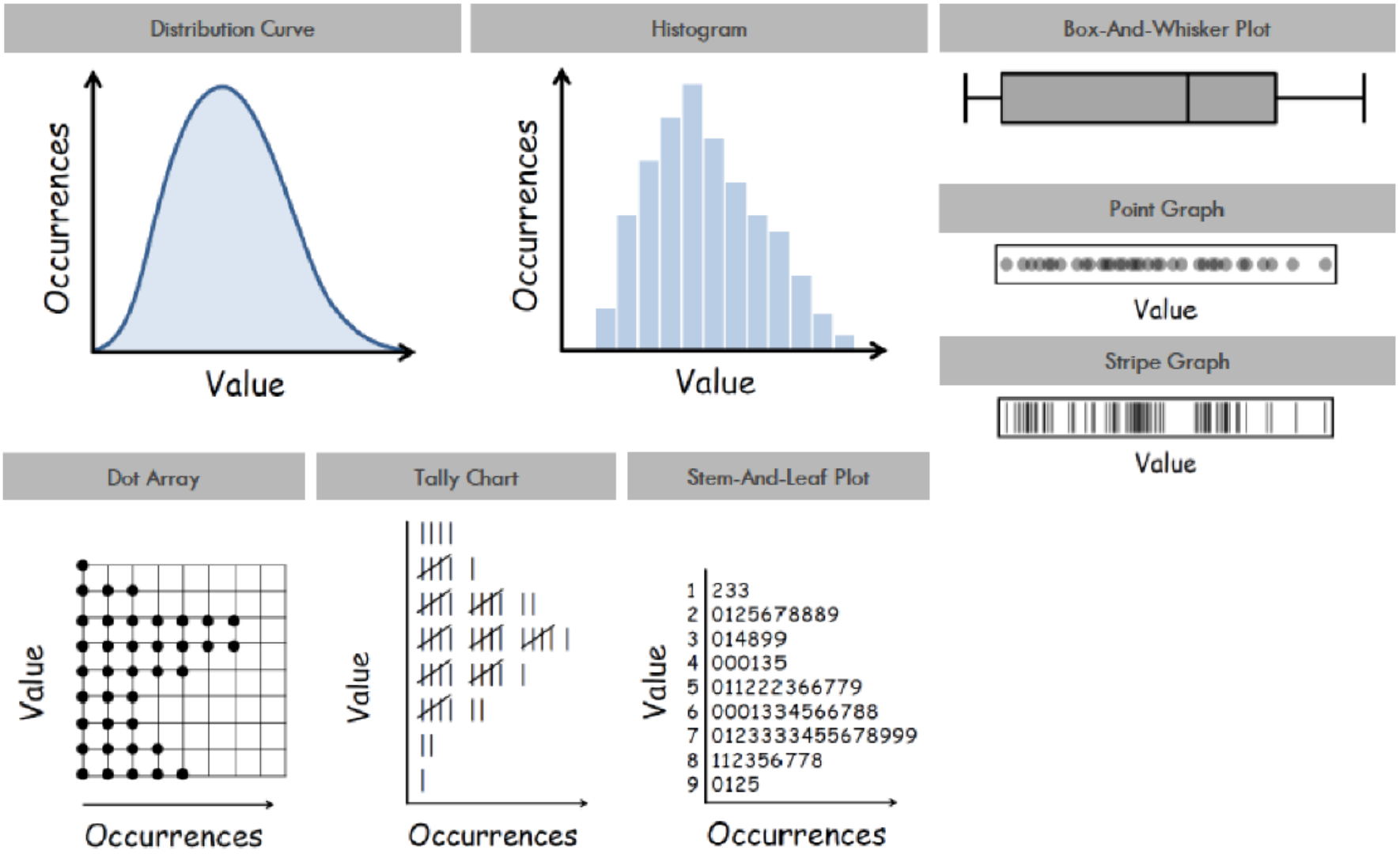
- **Attributes**
  - One
    - *Distribution*
    - *Extremes*
  - Many
    - *Dependency*
    - *Correlation*
    - *Similarity*

- **Network Data**
  - Topology
  - *Paths*

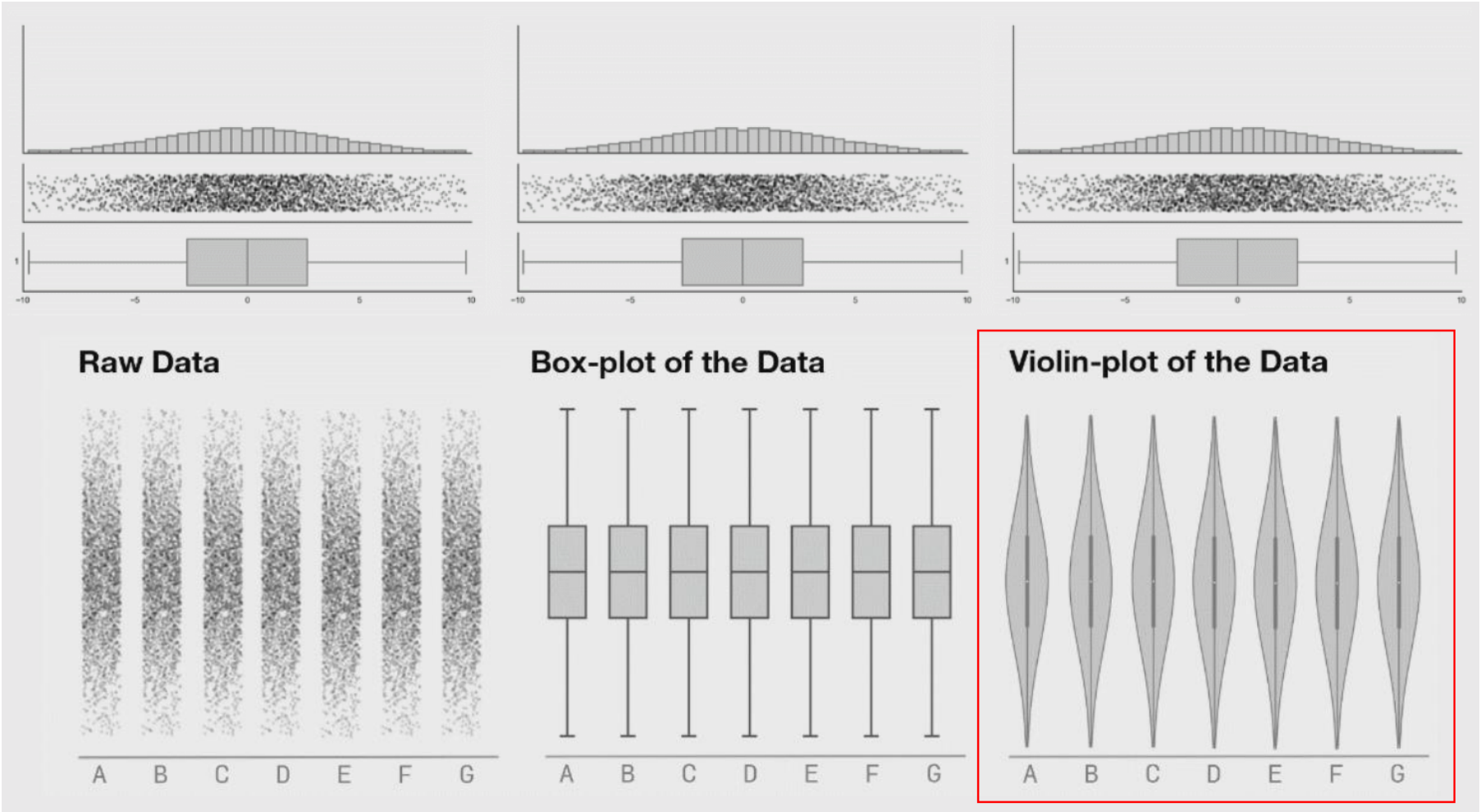


- **Spatial Data**
  - Shape

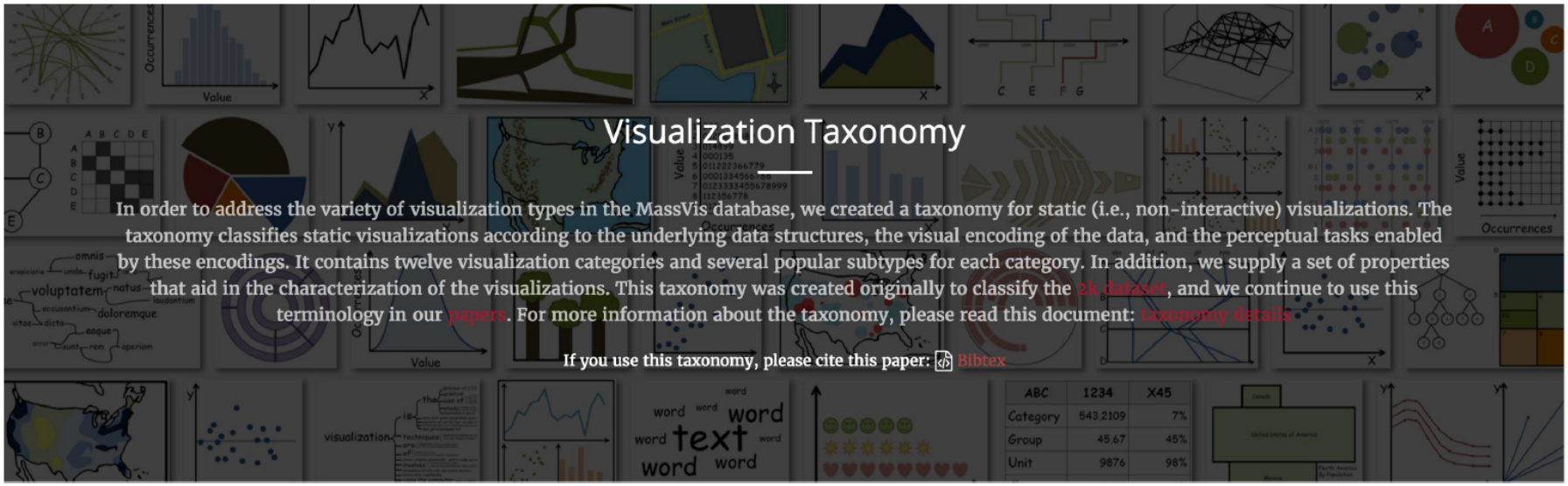
One special case: Distributions & Correlations



Violin plot


Raw Data
Box-plot of the Data
Violin-plot of the Data
A B C D E F G
A B C D E F G
A B C D E F G

Taxonomy



http://massvis.mit.edu/[1]

Great resource for categorizing visualizations, and brainstorming!

[1]Borkin, M., Vo, A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H., 2013, "What Makes a Visualization Memorable?", IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis 2013), 19, 12, 2306 2315.

Other resources

- DataVizProject (https://datavizproject.com/)
- The Data Visualization Catalogue (https://datavizcatalogue.com/)
- Matplotlib (https://matplotlib.org/gallery.html)
- Plotly (https://plot.ly/python/)
- D3 (https://github.com/d3/d3/wiki/Gallery)

In-class activities

For the following three cases, please discuss what tasks may be involved and classify the tasks:

- Stock Market Analysis
- Healthcare Data Visualization
- Supply Chain Management

In-class activities

For the previous three cases:

- Stock Market Analysis
- Healthcare Data Visualization
- Supply Chain Management

Choose one of the cases and then:

- Choose one of the tasks we have just described for the task. (everyone in the group should choose a different one)
- List what is the target and action for the task.
- Sketch the visualization for the task
- List what marks and channels you have chosen and discuss why.
- If you can add interaction to the visualization, describe what that should look like