Visualization for Data Science
Task Abstraction

Administrivia – Instructor Absences

October 20th through October 27th

- Dr. K is in Gaborone for CompEd Conference
- October 20th Class8A. No in-person class. Two tutorial files this week.
  - Lecture 8A, which is the gallery of vizzes
  - Tutorial 7: is the last assessed programming content (Multiview)
  - Tutorial 8 - will be instructive for your EDA portion of your project.
- October 22nd Class8B. Matt will lead the lecture.
  - Conceptual foundations of Multi-views
- October 27th Class 9A. No in-person class. Zoom lecture. Flight lands at 1:30pm so we will zoom together at 3:30pm. The zoom lecture will also be recorded.

Administrivia – Quiz Update

Quiz 5: Error in AG will be manually fixed, stay tuned.

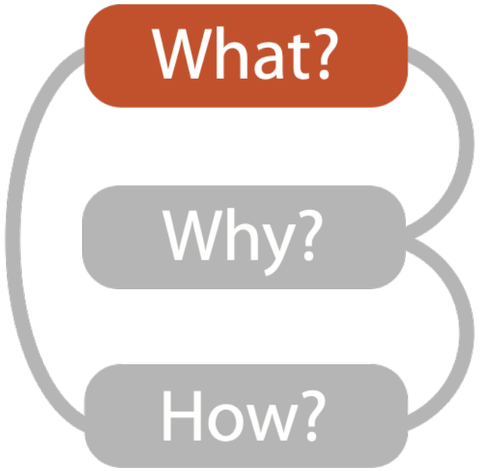
Quiz 6

- Tuesday and Wednesday this week
- All tests will be hidden

Quiz 7

- Covers Tutorial 7 and builds on viz you have seen up until this point
- Purely programming
- All tests will be hidden
- in 2 weeks (October 27 - 28th)

Quiz 1 Retake: November 3 – 4th)



What?

Datasets

Data Types

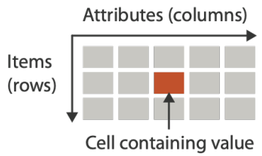
- Items
- Attributes
- Links
- Positions
- Grids

Data and Dataset Types

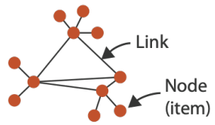
Tables	Networks & Trees	Fields	Geometry	Clusters, Sets, Lists
Items	Items (nodes)	Grids	Items	Items
Attributes	Links	Positions	Positions	
	Attributes	Attributes		

Dataset Types

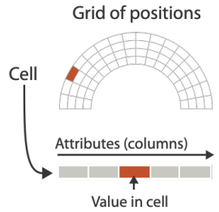
- Tables



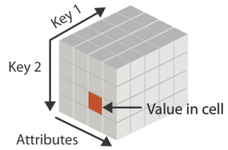
- Networks



- Fields (Continuous)



- *Multidimensional Table*



- *Trees*

- Geometry (Spatial)

Dataset Availability

- Static

- Dynamic

Attributes

Attribute Types

- Categorical

- Ordered
  - *Ordinal*

  - *Quantitative*

Ordering Direction

- Sequential

- Diverging

- Cyclic

How?

Encode

Arrange

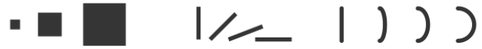
- Express
- Separate
- Order
- Align

- Use

Map
from categorical and ordered attributes

- Color
  - *Hue*
  - *Saturation*
  - *Luminance*

- Size, Angle, Curvature, ...



- Shape

- Motion
  Direction, Rate, Frequency, ...

Manipulate

- **Change**

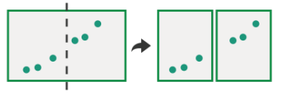
- **Select**

- **Navigate**

Facet

- **Juxtapose**

- **Partition**



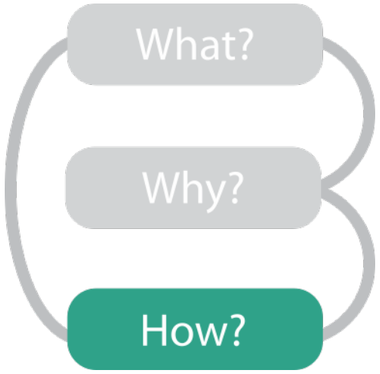
- **Superimpose**

Reduce

- **Filter**

- **Aggregate**

- **Embed**



COMPARING CATEGORIES

Compare values across categories

Bar

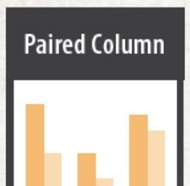

Paired Bar

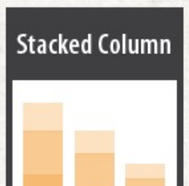

Stacked Bar


Span

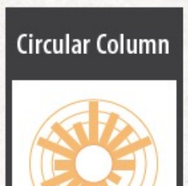

Unit

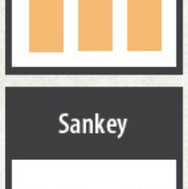

Column

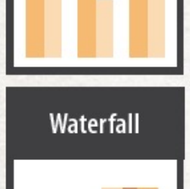

Paired Column


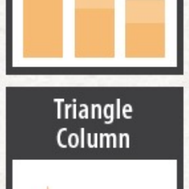
Stacked Column


Radial Bar


Circular Column


Sankey


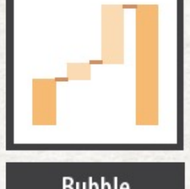
Waterfall


Triangle Column


Heatmap

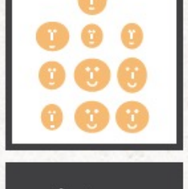

Chernoff

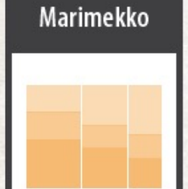

Marimekko

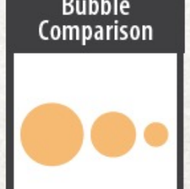

Bubble Comparison


Gauge

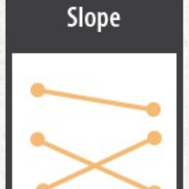

Slope

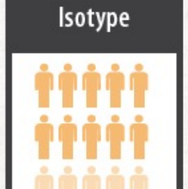

Isotype


Mosaic


Nested Bubbles


Bullet

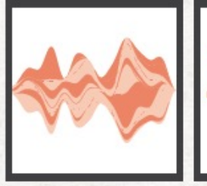

Dot Plot


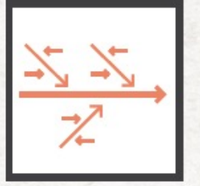
Word Cloud


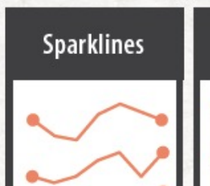
TIME

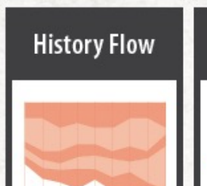
Track changes over time

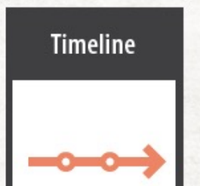
Line


Stream


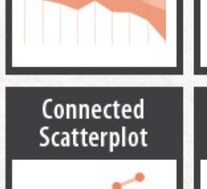
Cause-Effect


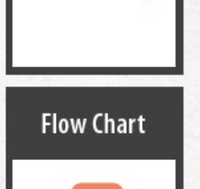
Sparklines


History Flow


Timeline


Horizon

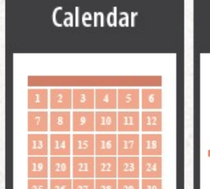

Connected Scatterplot


Flow Chart


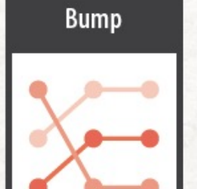
Calendar


Arc-Time


Bump


Stacked Area


Gantt


Cycle


DISTRIBUTION

Representation of the distribution of data

Histogram

Fan

Box & Whisker

Pyramid

Violin

Stem & Leaf

Stock

Stripe

Confidence Interval

Error Bars

GEOSPATIAL

Relates data to its geography

- Map
- Flow Map
- Icon Map
- Choropleth
- Map with Columns
- Isopleth
- Cartogram
- Map with Pie Charts
- Contour
- Non-Contiguous Cartogram
- Bubble Map
- Dorling Map
- Connection Map
- Point Map
- Subway Map

PART-TO-WHOLE

Relates the part of a variable to its total

- Pie
- Treemap
- Donut
- Square Cloud
- Arc
- Voronoi
- Nightingale
- Triangle Treemap
- Sunburst
- Waffle

RELATIONSHIP

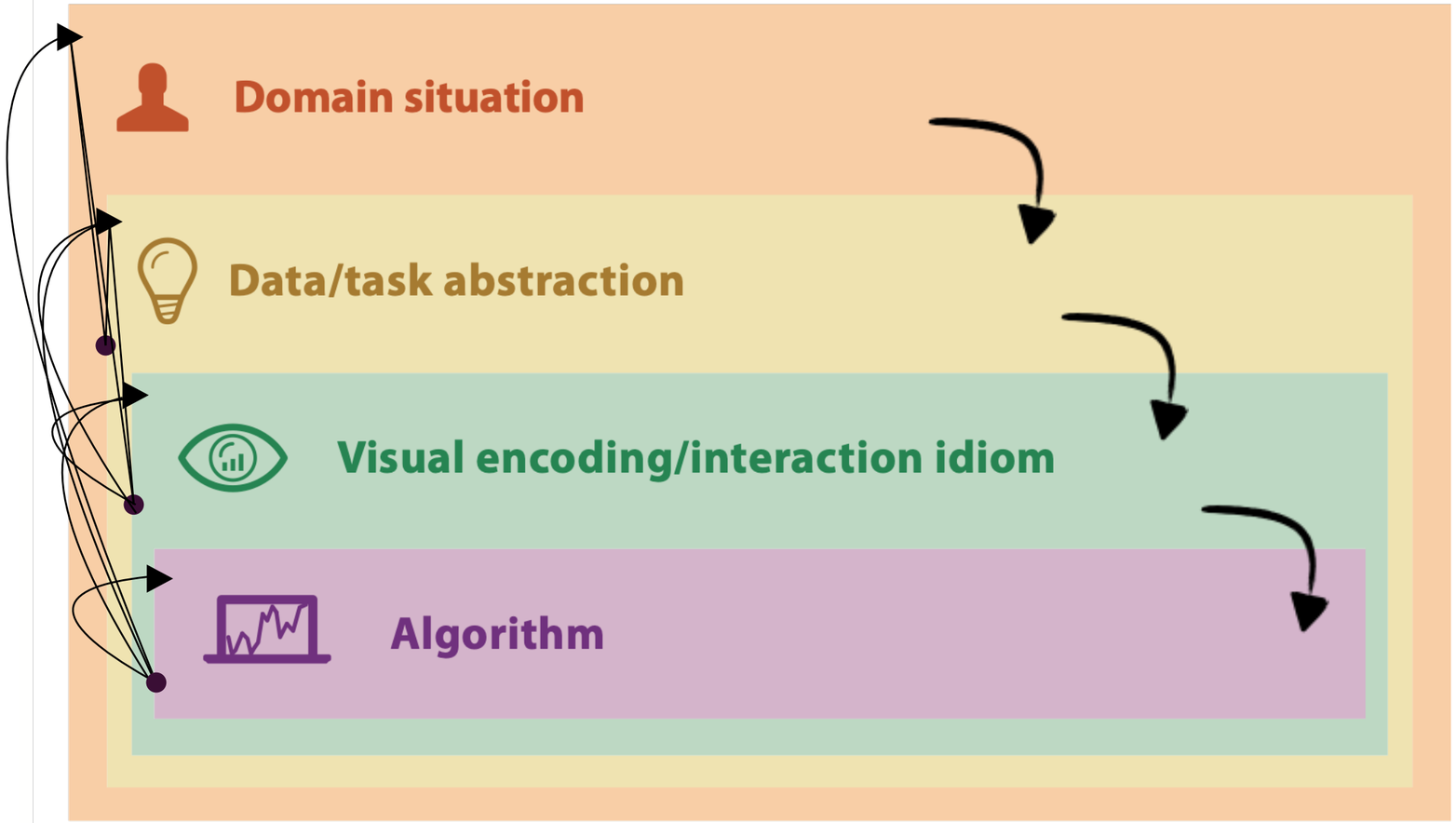
Illustrates correlations or relationships between variables

- Scatterplot
- Arc-Connection
- Dendrogram
- Word Tree
- Circle Packing
- Chord
- Tree
- Correlation Matrix
- Bubble
- Hive
- Double Tree
- Radar
- Parallel Coordinates
- Force-Directed
- Network
- Venn Diagram

@jschwabish @SR_Visual_Info

Nested model

- downstream: cascading effects
- upstream: iterative refinement

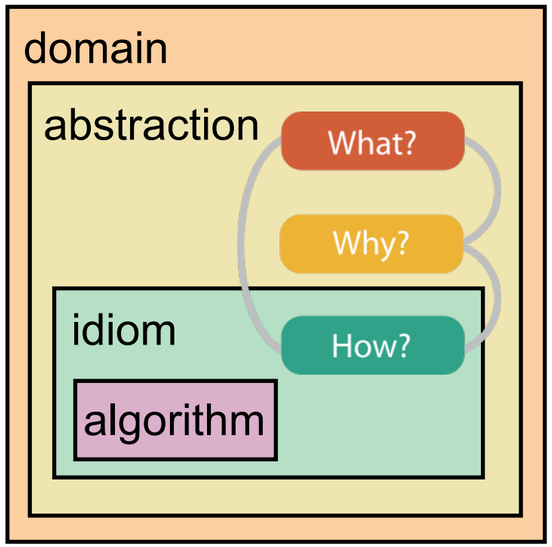


[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

Abstraction: Tasks (Why)

Analysis framework

- *domain situation*
  - who are the target users?
- *abstraction*
  - translate from specifics of domain to vocabulary of vis
    - what is shown? **data** abstraction
    - why is the user looking at it? **task** abstraction
- *idiom*
  - how is it shown?
    - **visual encoding** idiom: how to draw
    - **interaction** idiom: how to manipulate
- *algorithm*
  - efficient computation

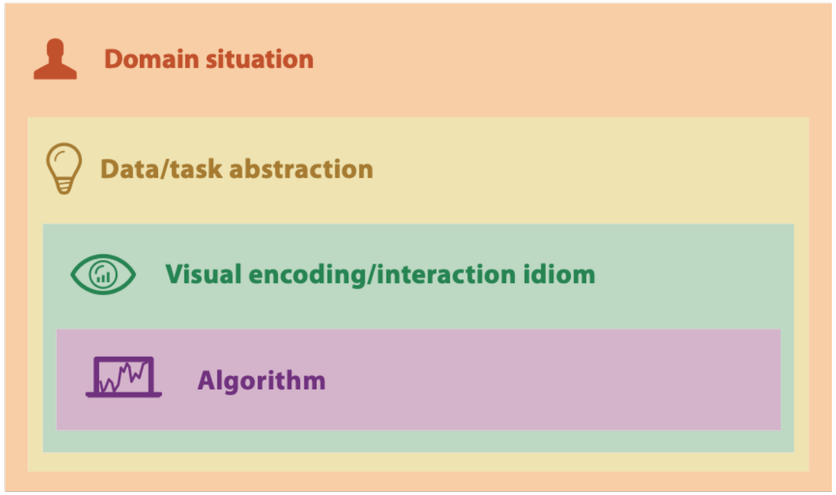


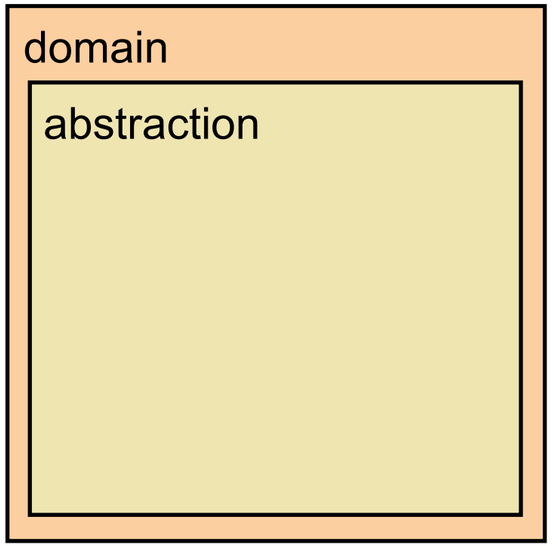
[A Multi-Level Typology of Abstract Visualization Tasks. Brehmer and Munzner. IEEE TVCG 19(12):2376-2385, 2013 (Proc. InfoVis 2013).]

[A Nested Model of Visualization Design and Validation. Munzner. IEEE TVCG 15(6):921-928, 2009 (Proc. InfoVis 2009).]

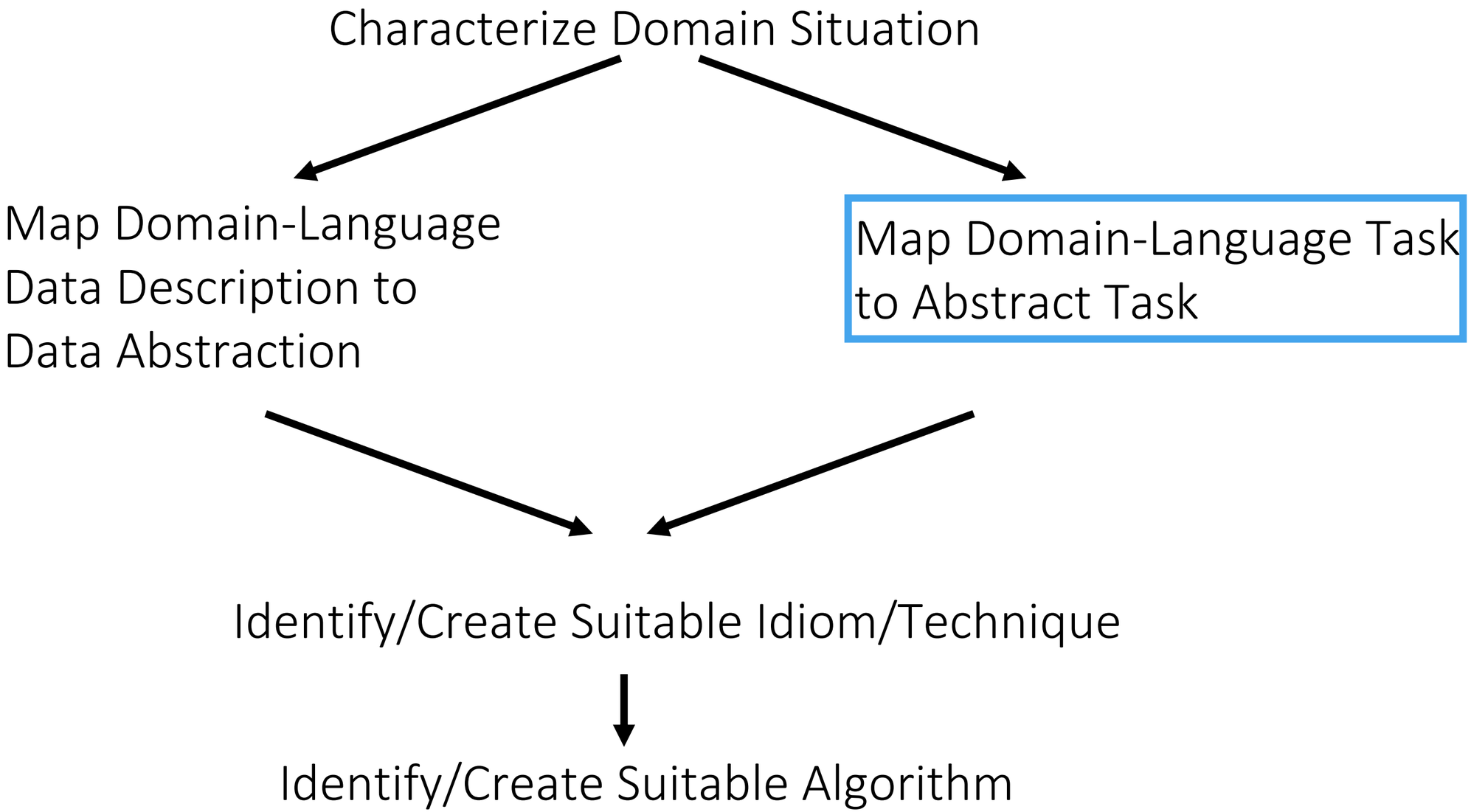
From domain to abstraction

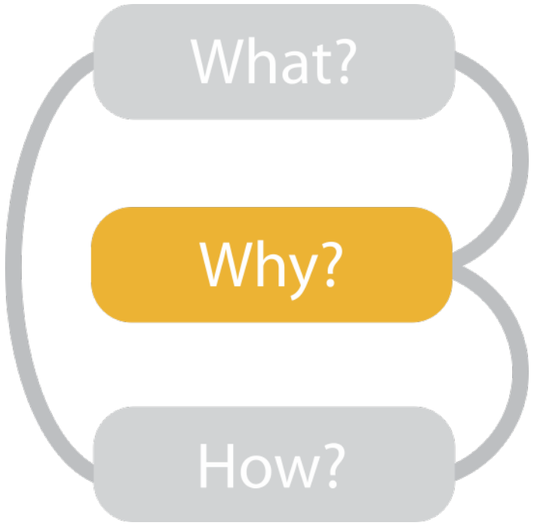
- domain characterization:
  details of application domain
  - group of users, target domain, their questions & data
    - varies wildly by domain
    - must be specific enough to get traction
  - domain questions/problems
    - break down into simpler abstract tasks
- abstraction: data & task
  - map ***what*** **and** ***why*** into generalized terms
    - identify tasks that users wish to perform, or already do
    - find data types that will support those tasks
      - possibly transform /derive if need be





Design process





- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Why?

Actions

Analyze

→ Consume

→ *Discover* → *Present* → *Enjoy*

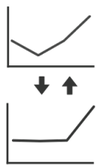


Search

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

Query

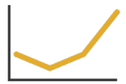
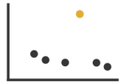
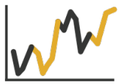
→ Identify → Compare → Summarize


Targets

All Data

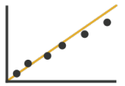
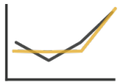
→ Trends → Outliers → Features

Attributes

→ One → Many

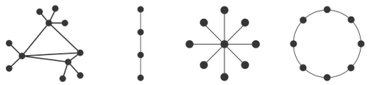
→ *Distribution* → *Dependency* → *Correlation* → *Similarity*

→ *Extremes*

Network Data

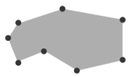
→ Topology



→ *Paths*

Spatial Data

→ Shape



Task Abstraction: Actions & Targets

High Level Conceptualization: Task = Action **on** Target

Actions – What is the user doing?

- The action the user is engaging in
- Examples: analyze, search, query, compare, identify, correlate

Targets – What is being acted upon?

- The target depends on how many attributes the user is focusing on
  - **Single Attribute:** Working with just one variable. Examples: distribution, extremes (min/max), outliers
  - **Multiple Attributes:** Working with several variables simultaneously Examples: correlation, dependency, relationships, patterns

Action – A low-level classification of tasks

- **Retrieve Value** - Find a specific value for a particular item or case in the dataset
- **Filter** - Select and display only the subset of data that meets certain criteria or conditions
- **Compute Derived Value** - Calculate a new value by aggregating, combining, or transforming existing data
- **Find Extremum** - Identify the maximum or minimum value within the data
- **Sort** - Order data items by the values of one or more attributes
- **Determine Range** - Find the span between the minimum and maximum values of an attribute
- **Characterize Distribution** - Understand the overall pattern and shape of how values are spread across an attribute
- **Find Anomalies** - Identify outliers, exceptions, or unusual cases that don't fit expected patterns
- **Cluster** - Discover groups of items that share similar characteristics or values
- **Correlate** - Determine whether a relationship or trend exists between two or more attributes

Movie Scenario

I'm curating a Martin Scorsese retrospective at our independent cinema. We have limited time slots - our evening screenings need to wrap up by 10:30pm with a 7pm start, so we're working with about a 3-hour window including intermission. I want to get a sense of his filmmaking style over the decades. Can you help me understand what his body of work looks like in terms of how he approaches film length? I've heard that older Hollywood films from the studio era tended to be more constrained in runtime - is that something I should expect with his earlier work, or does he break from those patterns?

Task Abstraction Exercise - PraireLearn

Read the scenario carefully and identify the underlying analytical tasks the curator needs to accomplish. For each task you identify:

Step 1: Write out the domain-specific question (viz-worthy?)

Step 2: Identify the Actions

- What analytical actions does the curator need to perform?
- Use the low-level task classification we discussed (Retrieve Value, Filter, Compute Derived Value, Find Extremum, Sort, Determine Range, Characterize Distribution, Find Anomalies, Cluster, Correlate)

Step 3: Identify the Targets

- What data attributes is the curator focused on?
- How many attributes are involved in each task? (Single attribute or multiple attributes?)
- For each target, do a data abstraction (what is the attribute type?, what is the cardinality?)

Step 4: Create Task Pairs

- Combine each action with its corresponding target(s)
- Format: **Action → Target(s)**

Example

Step 1: Domain Specific Question

- "What does the overall scope and scale of Scorsese's filmography look like, and how does it compare across different periods of his career?"

Step 2: Action

- **Compute Derived Value** (counts/aggregations across categories)

Step 3: Target

- **Attributes:** Films × Decade/Era
- **Number of attributes:** 2 (multiple attributes)
- **Attribute types:**
  - Film count: Quantitative (discrete)
  - Decade: Temporal (ordered)
- **Cardinality:**
  - Film count: ~25-30 total, varying by decade
  - Decade: ~6 distinct values (1970s-2020s)

Step 4: Task Pair

- **Compute Derived Value → Film count by decade (2 attributes)**

Diversity of Actions

- Retrieve Value - How long is the movie Gone with the Wind?
- Filter - What comedies have won awards?
- 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 won
- 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?

https://faculty.cc.gatech.edu/~stasko/papers/infovis05.pdf

Example: Find good movies

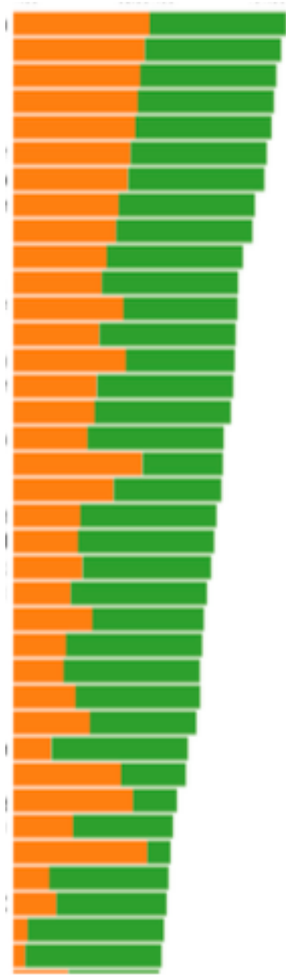
- identify good movies in genres I like
- domain:
  - general population, movie enthusiasts
- task: what is a good movie for me?
  - highly rated by critics?
  - highly rated by audiences?
  - successful at the box office?
  - similar to movies I liked?
  - matches specific genres?
- data: (is it available?)
  - yes! data sources IMDB, Rotten Tomatoes...

Example: Find good movies

- one possible choice for data and tasks, in domain language
  - data: combine audience ratings and critic ratings
  - task: find high-scoring movies for specific genre
- abstractions?
  - attribute: audience & critic ratings
    - ordinal
      - levels: 3 or 5 or 10…
  - attribute: genre
    - categorical
      - levels: < 20
  - items: movies
    - items: millions
  - task: find extreme (high) values

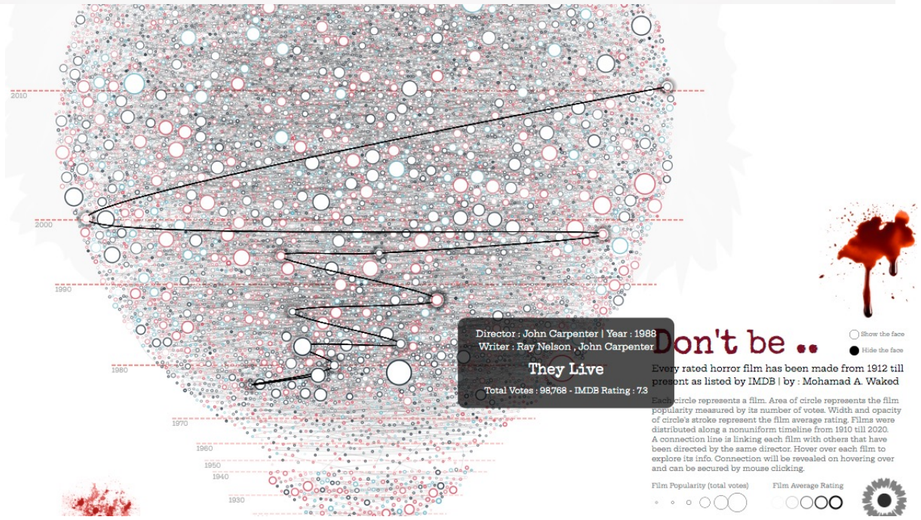
one possible idiom

- **stacked bar chart for ratings**



Example: Horrified

- same task: high-score movies
- slightly different data
  - 14K rated horror movies from IMDB
- very different visual encoding idiom
  - circle per item (movie)
  - circle area = popularity
  - stroke width/opacity = avg rating
  - year made = vertical position
- interaction idiom
  - lines connect movies w/ same director, on mouseover



https://www.alhadaqa.com/wp-content/uploads/2020/04/horrified.html

Example: Horrified vs stacked bars

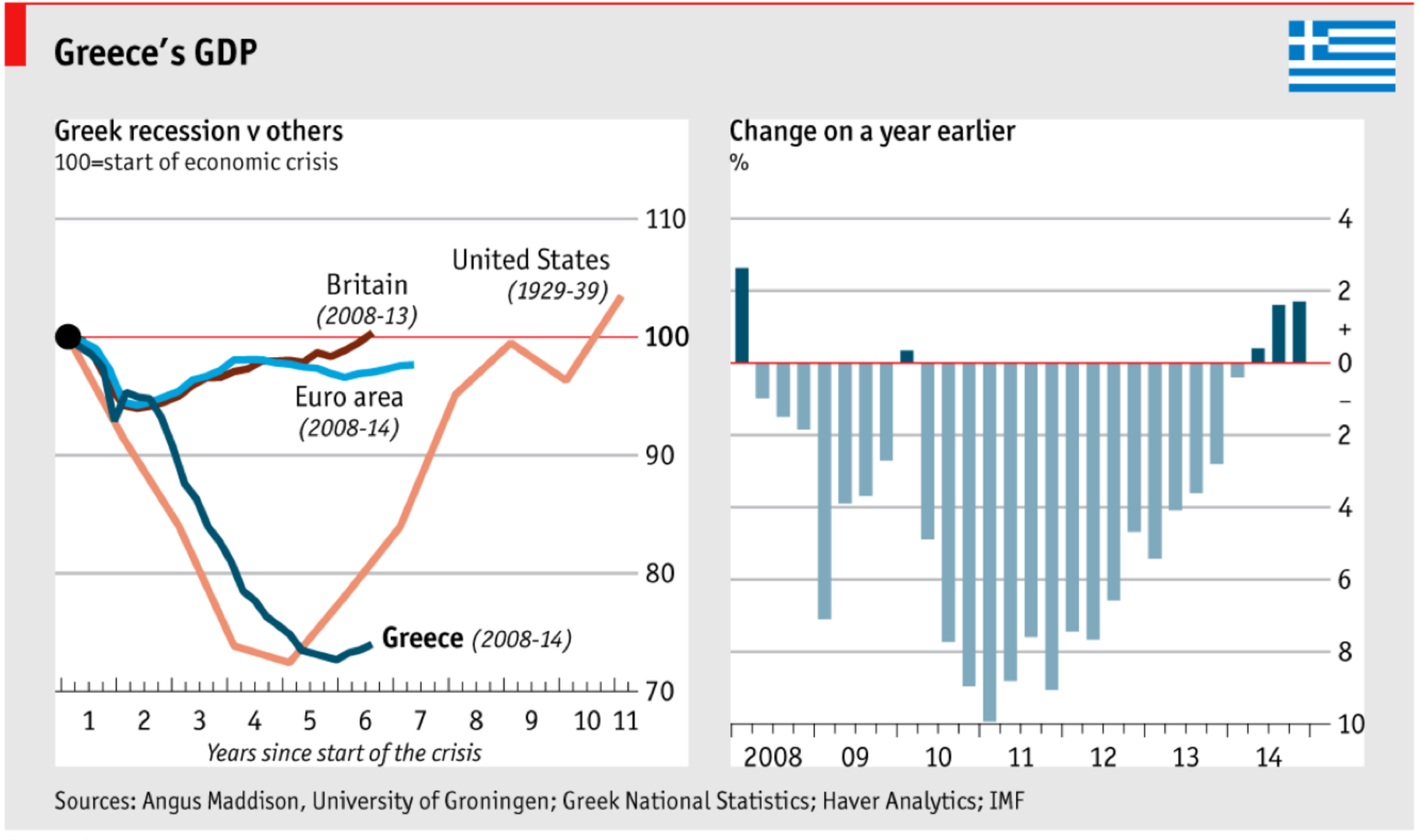
- horrified: browse/explore
- stacked bars: locate/lookup

- which is better?
  - depends on goals / task
    - enjoy, social context, lots of time
    - find 2nd-best rated movie of all time
      - Jeopardy call, < 10 seconds to respond!

http://alhadaqa.com/2019/10/horrified/

Example: Economics

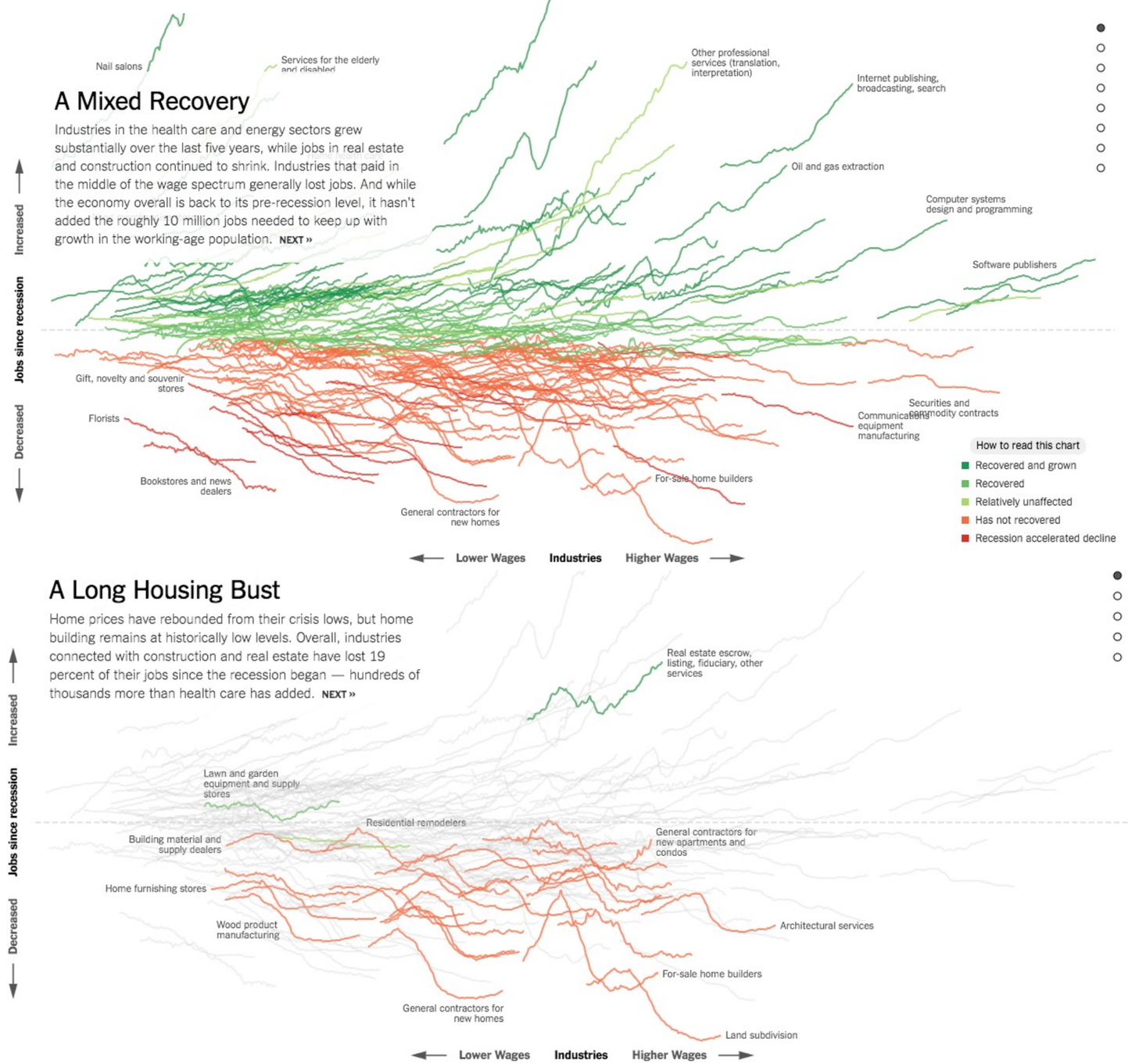
- task: compare and derive
- data: derive change



The Economist

Examples: Job market

- trends
  - how did job market develop since recession overall?
- outliers
  - real estate related jobs



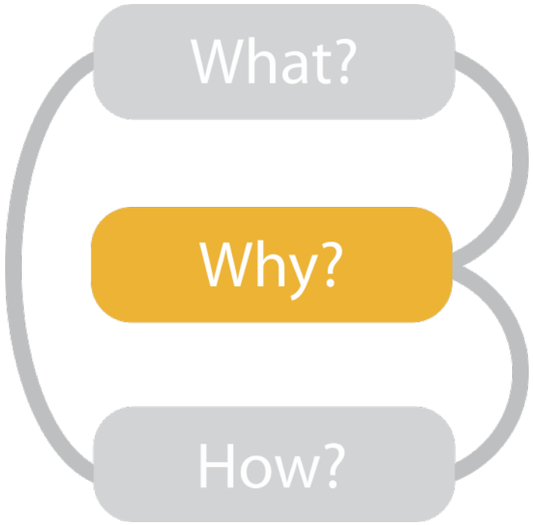
https://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html

Abstraction process

1. start by writing down the task, in the most natural ("domain") language
2. look for the actions (verbs) and targets (nouns)
3. add a data abstraction for each of the targets
   - then it's legitimate to use in the abstracted version of the task
   - you may need to recursively define other words along the way
   - you might realize you need to derive some data at this stage
4. consider how to translate actions into more generic ("abstract") words
   - either from the set of verbs from Tasks lecture, or come up with new ones
5. consider what other words/ideas could be abstracted
   - you might also derive some data at this stage

Task-Driven Visualization Design

- What is the overarching purpose of your visualization?
- What questions do you want to answer?
- What insights should the user be able to reach?
- What tasks should the user be able to accomplish?
  - More on this when we talk about Interaction



- {action, target} pairs
  - *discover distribution*
  - *compare trends*
  - *locate outliers*
  - *browse topology*

Why?

Actions

Analyze

→ Consume

→ *Discover* → *Present* → *Enjoy*

Search

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

Query

→ Identify → Compare → Summarize

Targets

All Data

→ Trends → Outliers → Features

Attributes

→ One → Many

→ *Distribution* → *Dependency* → *Correlation* → *Similarity*

→ *Extremes*

Network Data

→ Topology

→ *Paths*

Spatial Data

→ Shape

Task Abstraction: Actions & Targets

High Level Conceptualization: Task = Action **on** Target

Actions – What is the user doing?

- The action the user is engaging in
- Examples: analyze, search, query, compare, identify, correlate

Targets – What is being acted upon?

- The target depends on how many attributes the user is focusing on
  - **Single Attribute:** Working with just one variable. Examples: distribution, extremes (min/max), outliers
  - **Multiple Attributes:** Working with several variables simultaneously Examples: correlation, dependency, relationships, patterns