

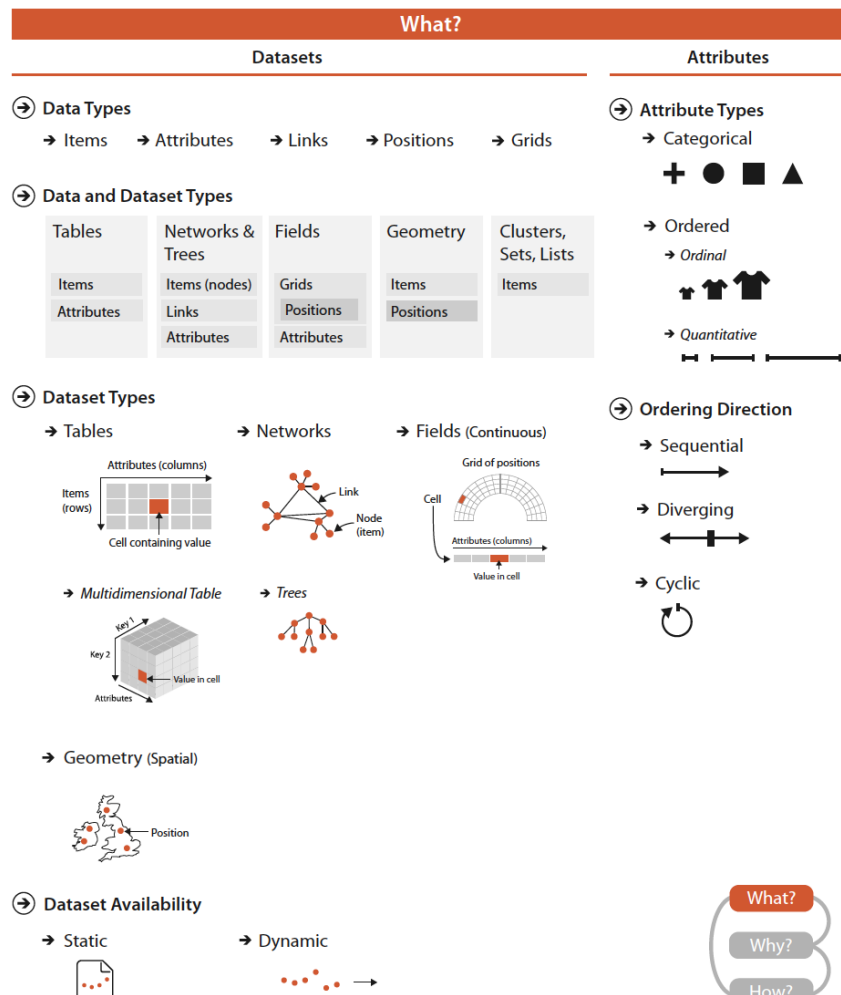
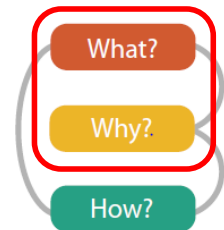
# Visualization Analysis and Design

## What: Data Abstraction

## Why: Task Abstraction

Jinwook Seo, Ph.D.

Department of Computer Science and Engineering  
Seoul National University



## Why Do Data Semantics and Types Matter?

Basil, 7, S, Pear

- What does each word mean?  
→ Semantics: real-world meaning

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

## Why Do Data Semantics and Types Matter?

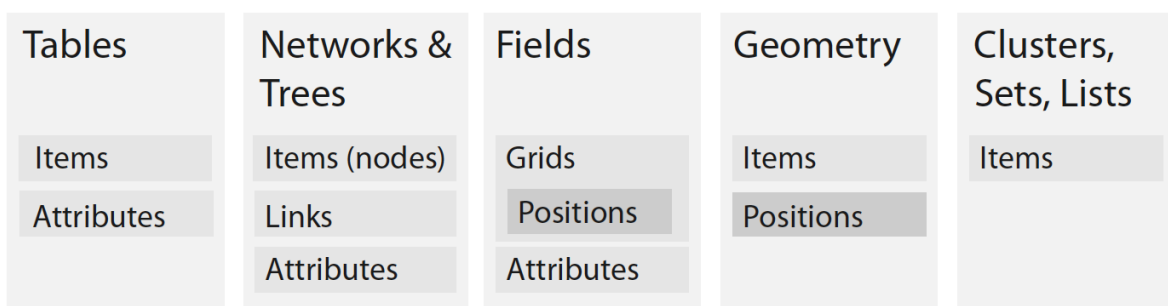
- Types of the data: structural or mathematical interpretation
  - Data level → **Data Types**
    - what kind of thing is it?
    - item, attribute, link, position, grid
  - Dataset level → **Dataset Types**
    - how are these data types combined into a larger structure
    - table, tree, field of sampled values
  - Attribute level → **Attribute Types**
    - what kinds of math operations are meaningful for it?
    - attribute: property that can be measured, observed, or logged
      - variable, dimension
        - Number of detergents: quantity – addition/subtraction
        - Postal code: code – category

## Data Types

- **Attribute:** specific property that can be measured, observed, or logged (a.k.a **variable** or **dimension**)
- **Item:** individual entity that is discrete
- **Link:** relationship between items
- **Grid:** specifies the *strategy for sampling continuous data* in terms of both geometric and topological relationships between its cells
- **Position:** spatial data, providing a location

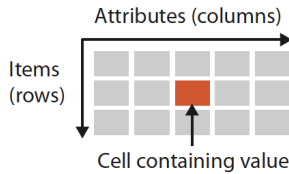
## Dataset Types and Data types

- Dataset
  - collection of info that is the target of analysis
  - arise from combinations of data types

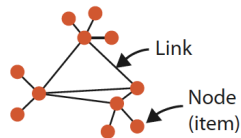


## Basic Dataset Types

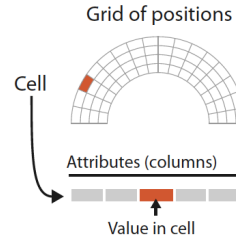
## → Tables



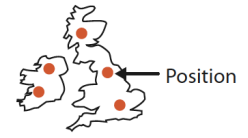
## → Networks



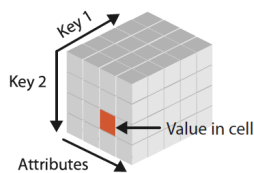
## → Fields (Continuous)



## → Geometry (Spatial)



## → Multidimensional Table



## → Trees



## Dataset Type: Tables

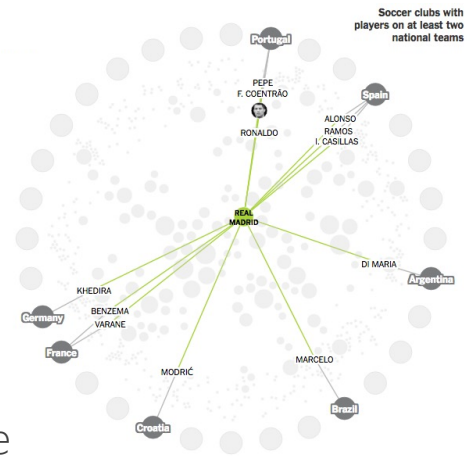
- item/record/tuple
- attribute/field/  
variable/dimension
- cell contains value
  - quantitative
  - ordinal
  - nominal

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	0.55	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	5	4-Not Specified	Small Pack	0.44	6/6/05
69	5	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: indexed with multiple keys

## Dataset Type: Networks and Trees

- well-suited when there is some kind of relationship b/w items
- **Node:** item
  - Can have associated attributes
- **Link:** relation between two items
- **Trees:** networks with hierarchical structure
  - Unlike network, there are no cycles

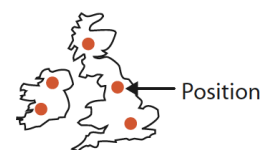


[http://www.nytimes.com/interactive/2014/06/20/sports/worldcup/how-world-cup-players-are-connected.html?\\_r=0](http://www.nytimes.com/interactive/2014/06/20/sports/worldcup/how-world-cup-players-are-connected.html?_r=0)

## Dataset Type: Geometry

- Specifies info about the **shape** of items with explicit spatial positions
- Often includes hierarchical structure at multiple scales
  - May be intrinsic or may be derived from the original data
- Not necessarily have attributes
- Simply showing a geometric dataset is not an interesting problem for a vis designer

## Geometry (Spatial)



## Dataset Type: Combinations

- **Set:** unordered group of items
- **List:** ordered group of items
- **Cluster:** grouping based on attribute similarity
- **Path:** ordered set + links connecting nodes
- **Compound network:** network with an associated tree

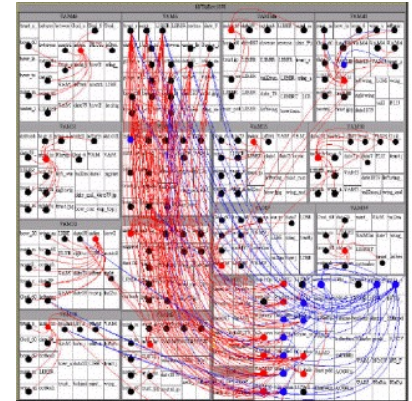
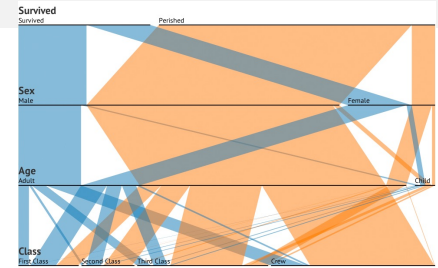


Figure 1: Directory structure of a Web site visualized as a Treemap with external links overlaid as curves. Blue curves are HTML links, red curves are image links.

## Abstraction and Availability

- Data Abstraction
  - domain-specific to GENERIC
  - translate domain-specific terms into words that are as **generic** as possible
- Data Availability
  - Static File: available all at once
  - Dynamic Streams

## Attribute Types

## Attributes

## ➔ Attribute Types

➔ Categorical

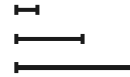


➔ Ordered

➔ Ordinal



➔ Quantitative



## ➔ Ordering Direction

➔ Sequential



➔ Diverging



➔ Cyclic



- Hierarchical Attributes: within an attribute or between multiple attributes

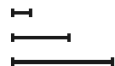
## Levels of Measurements

Ordered

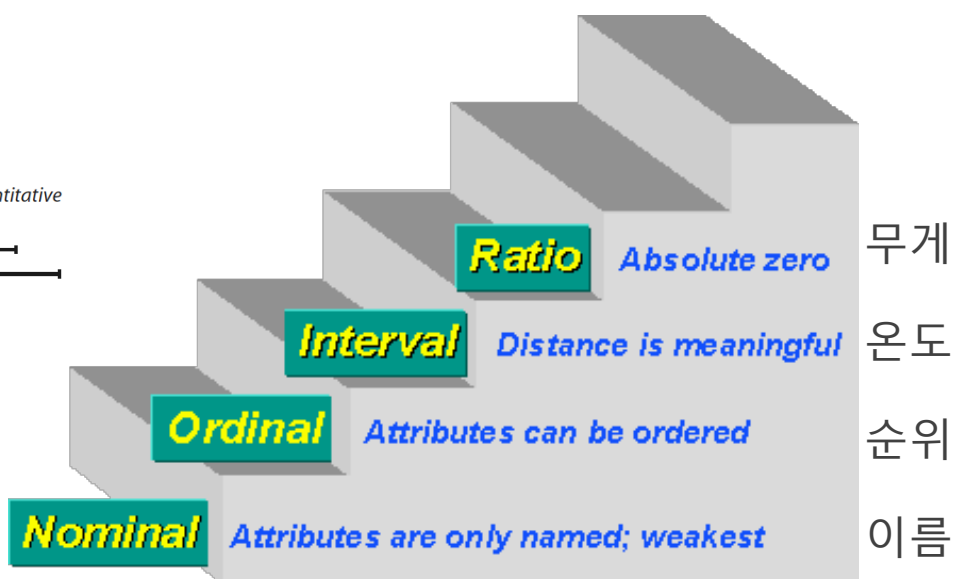
➔ Ordinal



➔ Quantitative



Categorical



## Semantics - Key vs. Value Semantics

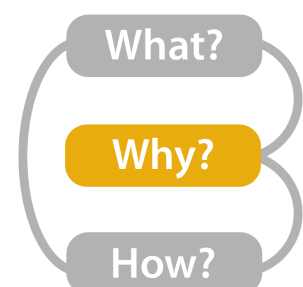
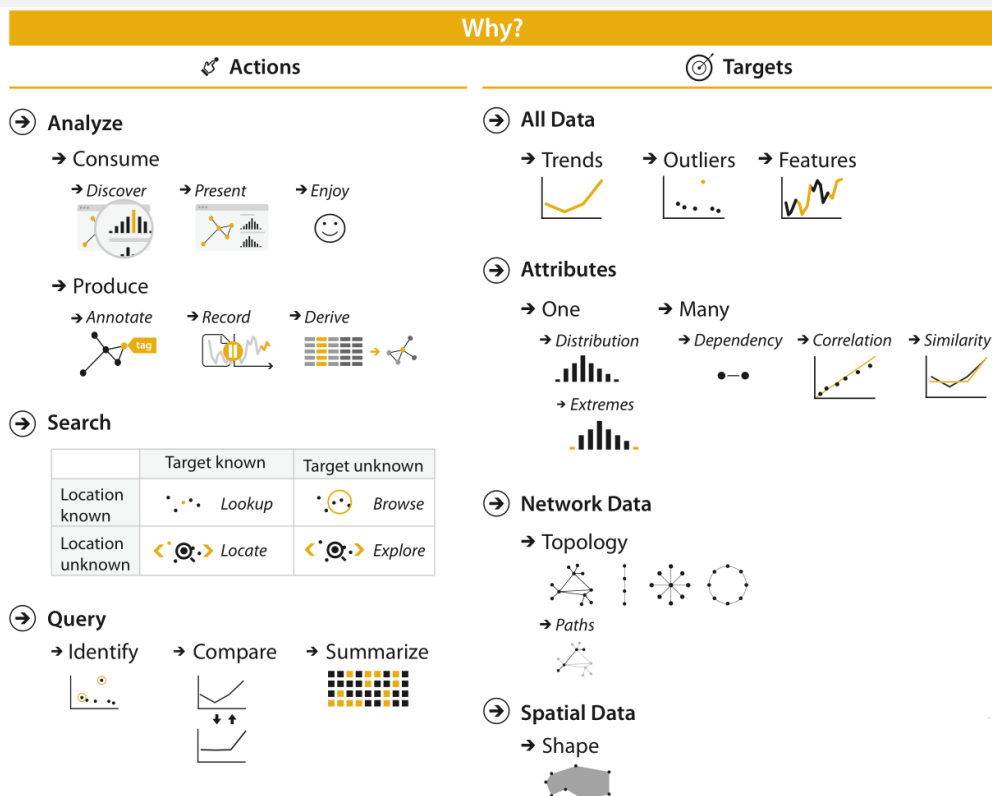
- **Key attribute** acts as an *index* used to look up **value attributes**
- Flat Tables: only one key
  - Implicit key: keys are simply the index of the row

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

- Explicit key: keys may be categorical or ordinal attributes (unique)
- Multidimensional Tables: multiple keys required
  - combination of all keys must be unique for each item, even though an individual key may contain duplicates
  - independent keys vs. dependent values

## The Big Picture

## Why People Are Using Vis? (for What Tasks?)





## Task Abstraction

- Consider tasks in **abstract form**, rather than **domain-specific way**
  - Otherwise, hard to make useful comparisons between domain situations
- domain-specific task descriptions



"Contrast the prognosis of patients who were intubated in the ICU more than one month after exposure to patients hospitalized within the first week"

"See if the results for the tissue samples treated with LL-37 match up with the ones without the peptide"



- Abstract form: *"compare values between two groups"*
- Task abstraction should guide data abstraction

## Actions

## Actions define User Goals

- High-level: **Analyze**
  - Consume, Produce
- Mid-level: **Search**
  - Lookup, Browse, Locate, Explore
- Low-level: **Query**
  - Identify, Compare, Summarize
- Choices at each level are independent
- Describe all of actions at all three levels

## → Analyze

## → Consume

## → Discover



## → Present



## → Enjoy



## → Produce

## → Annotate



## → Record



## → Derive



## → Search

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

## → Query

## → Identify



## → Compare



## → Summarize



## High-Level: Analyze – Consume (existing information)

- Discover (= explore)
  - Find new knowledge that was *not previously known*
    - **generate** a new hypothesis or **verify** an existing hypothesis
  - Often for scientific inquiry
- Present (= explain)
  - The communication of information that is *specific and already understood*
  - e.g. infographic (static information graphics)
  - output of a discover session -> input to a present session
- Enjoy
  - Motivated by users' enjoyment
  - Casual encounters with vis for enjoyment
  - Users are driven by Curiosity stimulated and satisfied by vis
  - e.g. Name Voyager

## → Consume

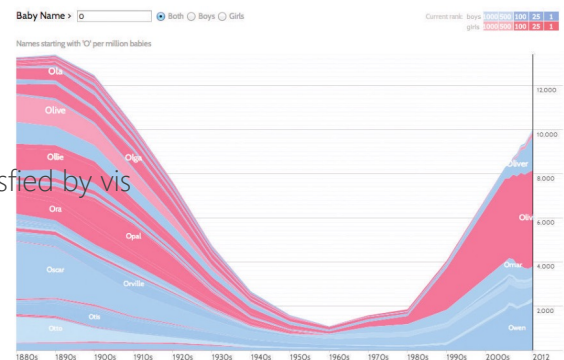
## → Discover



## → Present



## → Enjoy



## High-level: Analyze – Produce (new information as input to a next instance)

- Annotate (~ tag)
  - Attaches temporary info to existing elements
- Record
  - To save or capture vis elements as persistent artifacts
  - e.g. screen shots, interaction logs, etc.
- Derive (= transform)
  - To produce new data elements (= derived attributes) based on existing elements
    - Could expand the design space of possible vis idioms
  - Changing types of data
  - Transform with additional info
  - Using arithmetic/logical/statistical operations

## → Produce

## → Annotate



## → Record



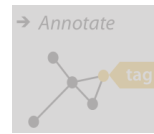
## → Derive



## High-level: Analyze – Produce - Record

- Record
  - graphical history of visual exploration

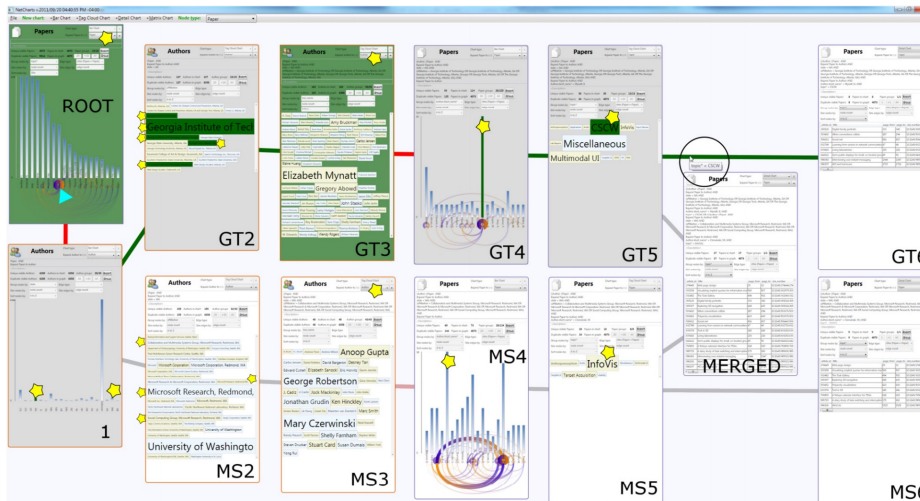
→ Produce



→ Record



→ Derive



A GraphTrail analysis showing two parallel exploration paths

[http://research.microsoft.com/en-us/um/people/nath/docs/graphtrail\\_chi2012.pdf](http://research.microsoft.com/en-us/um/people/nath/docs/graphtrail_chi2012.pdf)

## High-level: Analyze – Produce - Derive

- Do not just draw what you are given
  - Decide what the right thing to show is
  - Create it by transformations
  - and draw it!

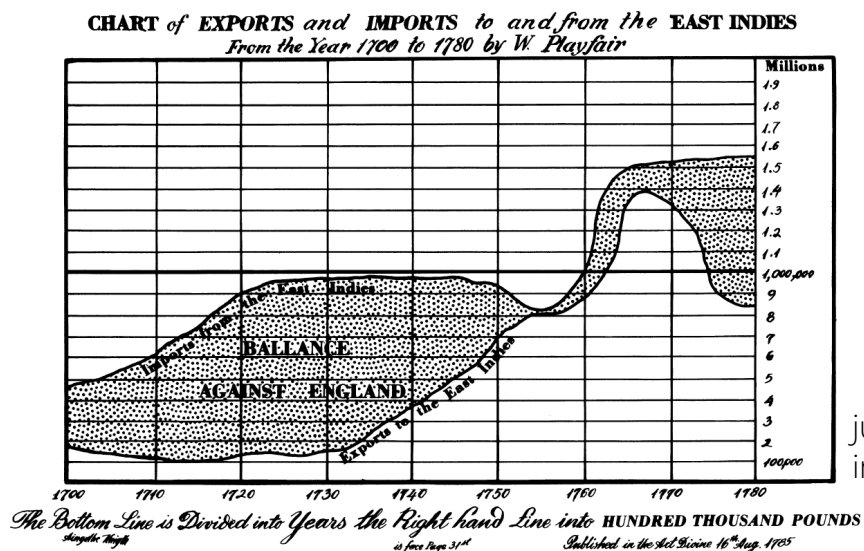
→ Produce



→ Record



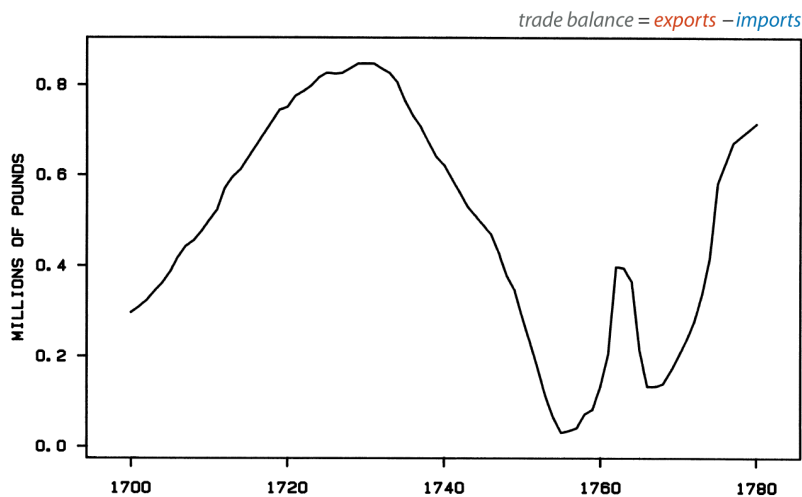
→ Derive

judging the difference  
in height

## High-level: Analyze – Produce - Drive

- Do not just draw what you are given
  - Decide what the right thing to show is
  - Create it by transformations
  - and draw it!
  - detail is aggregated away, but

## → Produce







## Mid-level: Search (Find with successful outcome)

- High-level analyze cases require the user to search for elements of interest

	Target known	Target unknown
Location known	<i>Lookup</i>	<i>Browse</i>
Location unknown	<i>Locate</i>	<i>Explore</i>

## Mid-level: Search (Find with successful outcome)

- e.g. a tree vis for Tree of Life
  - Lookup
    - Looking up* human (target O) knowing it belongs to mammals (location O)
  - Locate
    - Locating* rabbits (target O) not knowing where it belongs to (location X)
  - Browse
    - Browsing* all leafs (target X) of the mammal subtree (location O)
  - Explore
    - Exploring* for a family having the largest number of species (target X, location X)

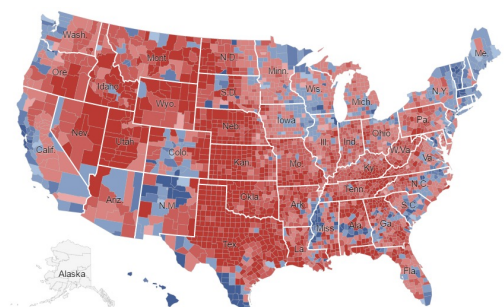
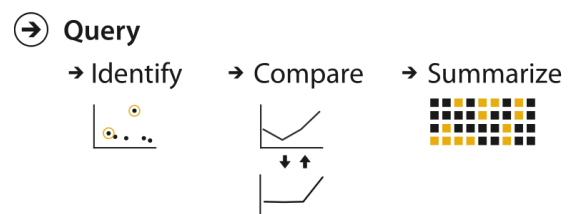
	Target known	Target unknown
Location known	 <i>Lookup</i>	 <i>Browse</i>
Location unknown	 <i>Locate</i>	 <i>Explore</i>



<http://tolweb.org/tree/home/pages/aboutoverview.html>

## Low-level: Query

- Once targets for a search has found,
  - QUERY the targets
- e.g. a choropleth map of US election results
  - Identify (a single target)
    - identify* the election results for one state
  - Compare (multiple targets)
    - compare* the election results of one state to another
  - Summarize / Overview (all possible targets)
    - summarize* the election results across all states to determine how many favored one candidate



## Targets that actions refer to

- Some aspect of the data that is interest of the user

→ All Data

→ Trends



→ Outliers



→ Features



→ Attributes

→ One

→ Distribution



→ Extremes



→ Many

→ Dependency



→ Correlation



→ Similarity



→ Network Data

→ Topology



→ Paths



→ Spatial Data

→ Shape



## Targets - All Data Level

- All Data level
  - Trends (= pattern)
    - a high-level characterization of a pattern in the data
    - e.g. increases, decreases, peaks, regression fit etc.
  - Outliers (= anomalies, novelties, deviants, surprises)
  - Features
    - Any particular structures of interest
    - Task-dependent definition
      - clusters in cluster analysis

→ Targets

→ All Data

→ Trends



→ Outliers



→ Features



→ Attributes

→ One

→ Distribution



→ Extremes



→ Many

→ Dependency



→ Correlation



→ Similarity



→ Network Data

→ Topology



→ Paths



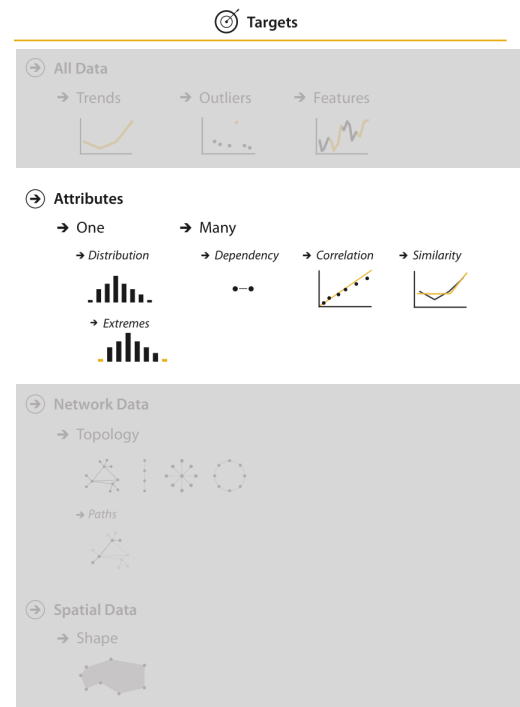
→ Spatial Data

→ Shape



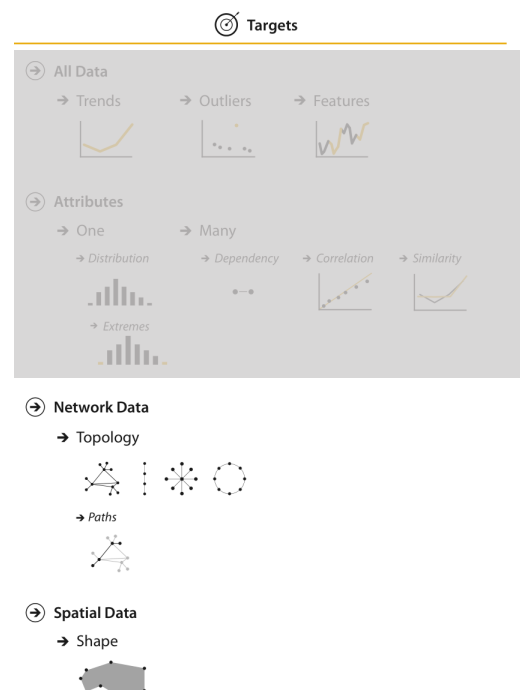
## Targets – Attributes Level

- Attributes level
  - One attribute
    - An individual value
    - extremes (maximum, minimum)
  - Multiple attributes
    - Dependency
      - An attribute depends on another attribute
    - Correlation
      - Tendency of values of two attributes are tied
    - Similarity
      - Quantitative measurement of how values of two attributes are similar



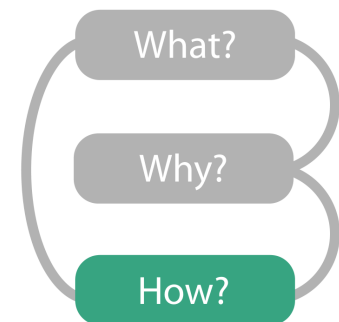
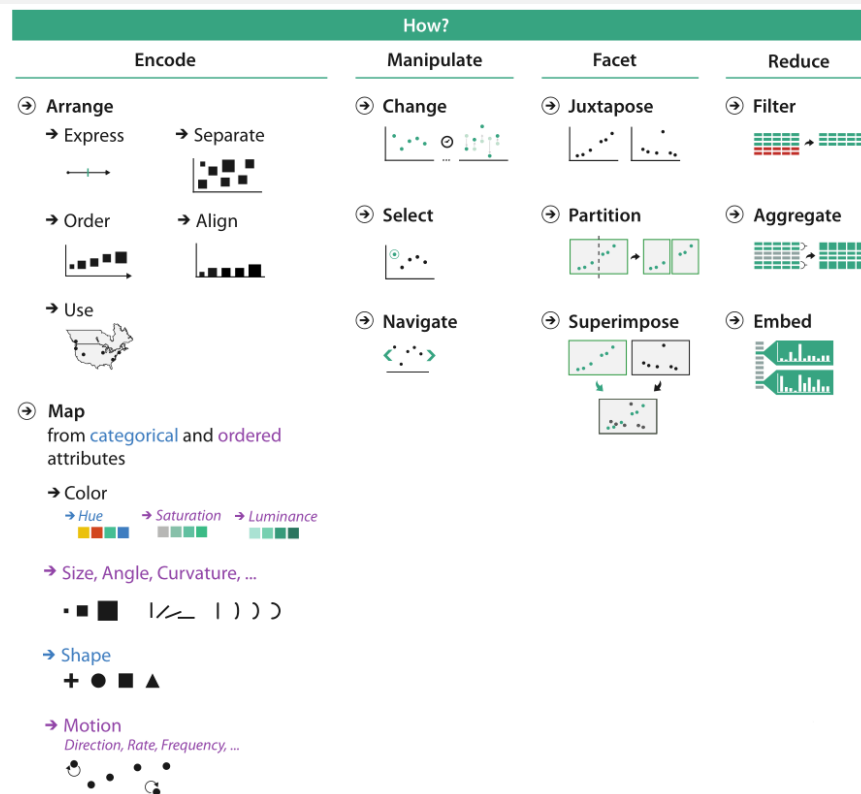
## Targets

- Targets for specific types of datasets
  - Network data
    - Topology
      - paths
  - Spatial data
    - Shape
  - And more..



- ✓ These targets can be expressed in domain-specific terms,  
But should always recognize these abstractions

## How a vis idiom is constructed? – a set of design choices



- Questions?