Data Visualisation
CSE613

Prof. Ramesh Ragala

February 1, 2016

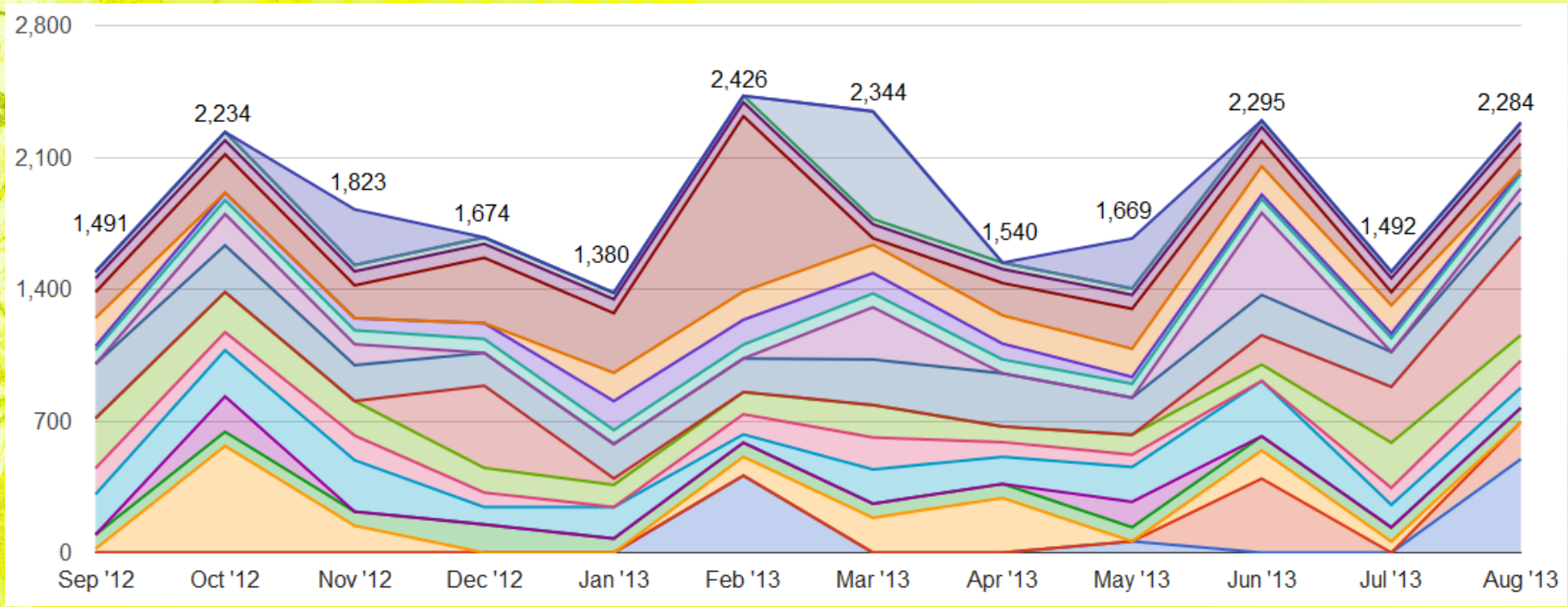
Introduction : Graphics

- We know the visual system parses the scenes of the physical world into objects objects perceived as having form and function.
- The mental representations of objects in the scene derive their meaning from experience,reflection, and reasoning.
- Graphics as external representations might also be considered scenes, but scenes of abstract worlds rather than the physical world.
- **Broad Definition of Graphics:**
  - **Graphics are visual representations of data that rely on graphical elements points, lines, areas, and volumes organized in a geometric space.**
  - the graphical elements represent objects, while relationships between them are represented by space.
  - Examples: Maps are oldest graphics

Introduction : Graphics

- From late 18^{th} century, graphics are used to represent abstract concepts like
  - Quantities
  - Time and Space
  - Preference
  - qualitative information such as various relationships between objects or sets of objects based on their properties.
- From late 20^{th} century, Interactive graphics and animation with the help of computer graphics and user interfaces $\rightarrow$ features
  - Ability to group
  - classify
  - Filter
  - Superimpose
  - Juxtapose
  - Permute the display elements

Introduction : Graphics



Introduction: Semiology of Graphics

- **Semiotics is a field that studies how signs and symbols become associated with meaning and the conventions or code by which signs are organized into systems to communicate and to model information about the external world**
- For many semiotic theorists $\rightarrow$ language was viewed as the central semiotic system and terms from linguistics have been borrowed and extended in its application to various disciplines outside linguistics.
- Example: Films and television programs are called "texts" which are "read."
- **http://www.digitalhumanities.org/dhq/vol/8/4/000190/000190.html**

Introduction: Semiology of Graph

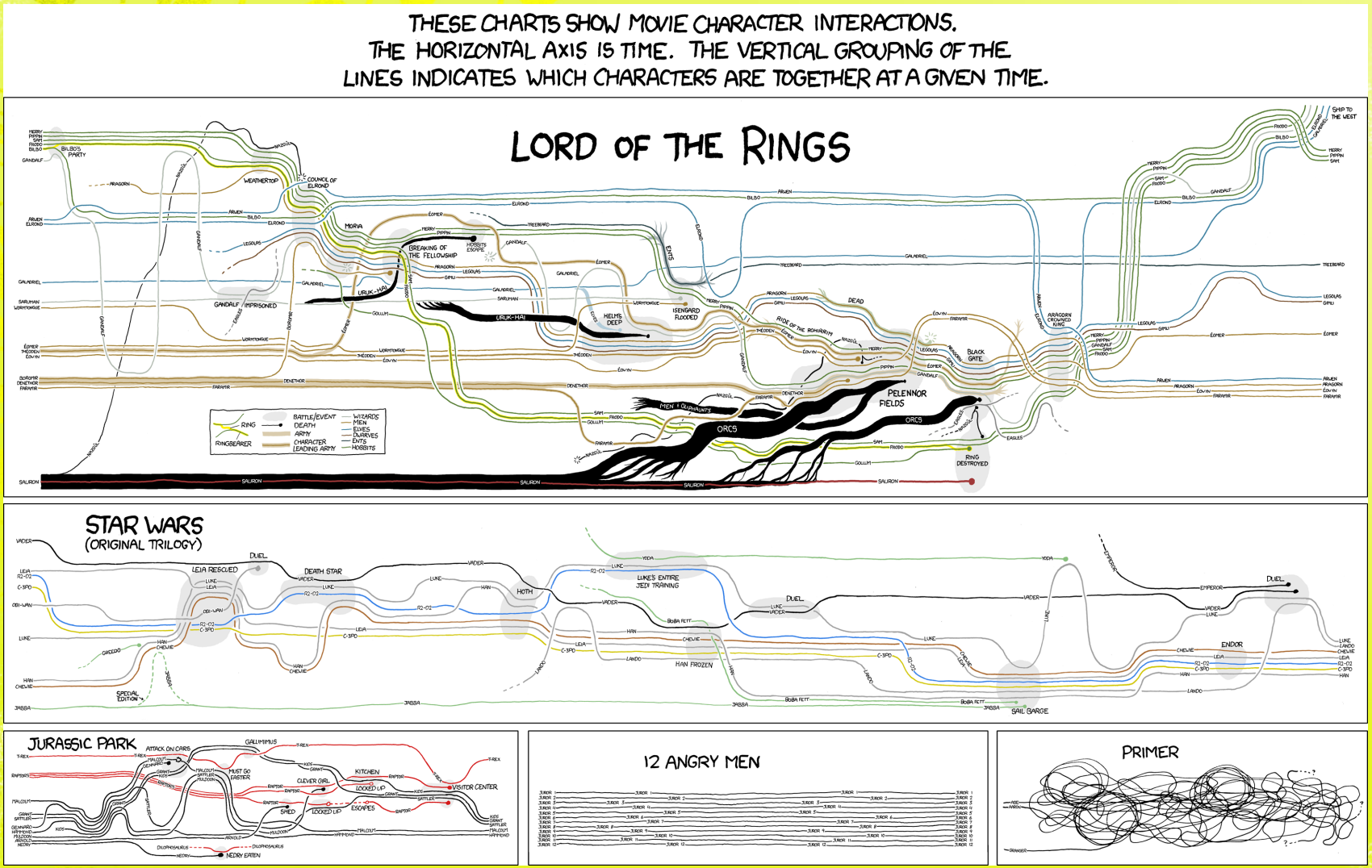


Figure: Charts of character interactions in movies

Introduction : Semiology of Graphics

- **Jacques Bertin(1918-2010)** used semiotics into graphics systems. French Cartographer
- **Jacques Bertin** developed a semiotic theory for graphic communication(1983),
  - This theory helps the designers to create diagrams, networks and maps with relevant information.
  - This theory helps the statisticians to understand the data and relations among the data more easily.
- He assumed graphic representation as a tool for discovery

Introduction : The Essence of Semiotics

- Humans communicate thoughts using systems of signs.
- Meaning is created in our minds when we "read" signs in "text."
- These carry meaning conveyed by learned conventions or codes that become transparent to us.
- A sign stands for something other than itself.
- It consists of two inseparable parts:
  - **Signifier:** Which is the form of the signifier
  - **Signified:** Which is the concept represent by the form
- We cannot make sense of the form unless we can relate it to a code.
- A code is a set of conventions: an arbitrary set of rules or standards in some domain that have been established to mean certain things.
- Example: Cartesian Coordinates system is a code → latitude on x-axis → longitude on y-axis.

Introduction : The essence of Semiotics

- The form is the connected set of shapes with black outlines and black and white shading.
- read the shapes (the signifier) as a person
- the Person with upraised arms (the signified)
- Assume this picture at airport ramp $\rightarrow$ signalling to the pilot
- signifier would be **upraised arms** and signified would be **this way**

Introduction : The essence of semiotics

- A signifier may have three possible fundamental modes of the relationship with it represents (the signified objects or concepts)
  - **Symbolic / Symbol:** The form of the symbol looks nothing like the thing that it represents, and the relationship is purely arbitrary and must be learned.
  - Example: word Tree
  - **Iconic / Icon:** The form resembles the thing it represents.
  - Example: folder in desktop
  - **Index / Indexed:** The form is connected directly or causally in a way that can be observed or inferred.
  - Example: smoke(signifier) means fire(signified)
- Sing's can be combined into texts $\rightarrow$ to avoid confusion

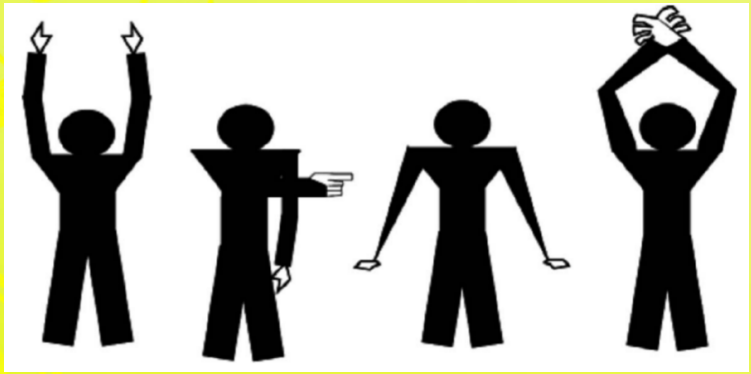
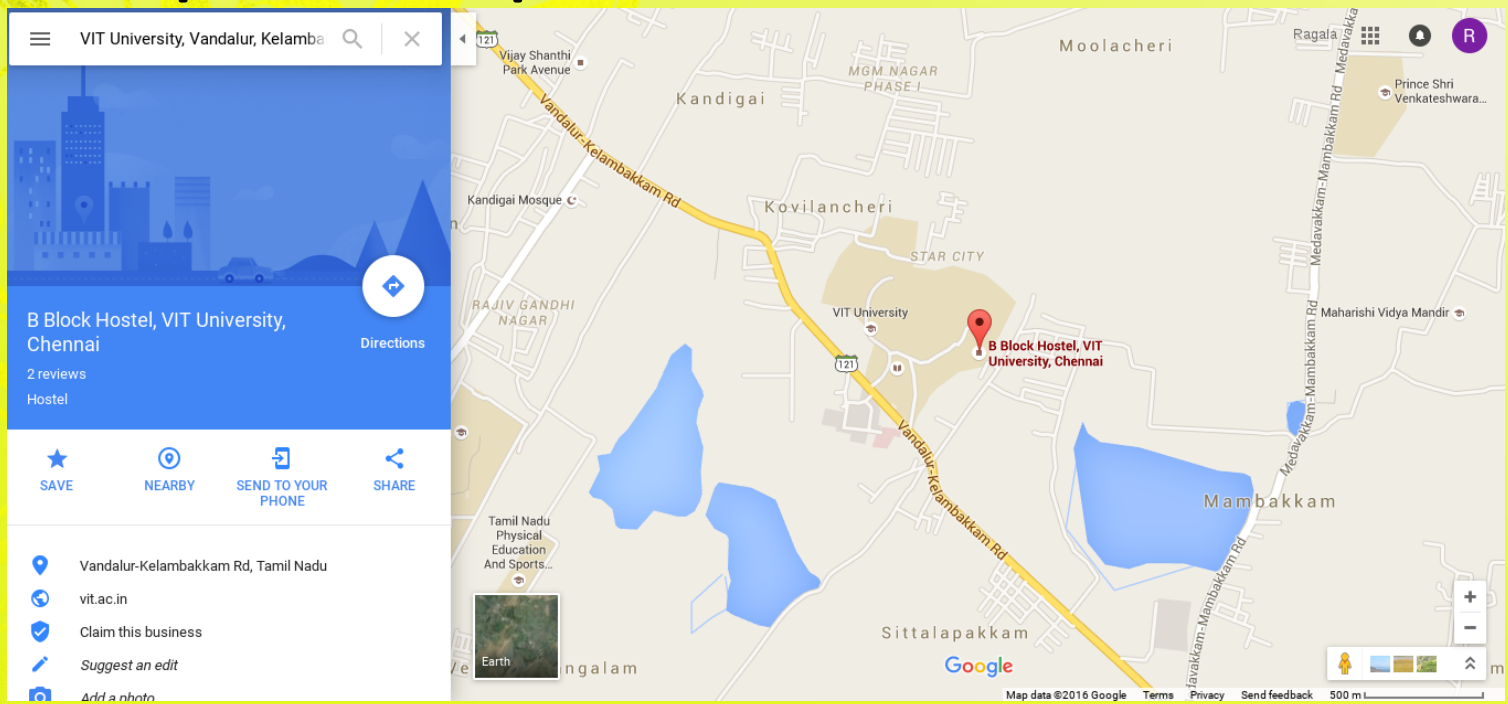


Figure: this way, turn to left, slowdown and stop

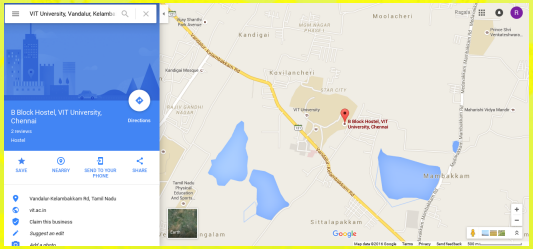
- The semiotics systems has following components or 3 branches of semiotics
  - **Syntax:** Concerned with how signs are recognized relative to each other.
  - **Semantics:** Concerned with how the meanings of signs are understood.
  - **Pragmatics:** Concerned with how signs convey the intended meaning within the context in which they are used.

- **Example on components of semiotics:**



- Features of the real world symbolized on maps as:
  - Points
  - Lines
  - Areas

- **Example on components of semiotics:**
  - **Syntax:** Relation between objects on a map and real world



  - **Semantic:** Relation between the objects on a map and the data they represent
    - Forest $\rightarrow$ blue color
    - Distance Guage
    - Identifying the labels
    - Population Areas
  - **Pragmatics:** How the map user responds to the map symbology
    - Context
    - Metal Model
    - Way finding

Introduction: the essence of semio

- Many Semiotic Systems to store, Communicate and understand different types of information such as music, language, Mathematics, abstract art and graphics.
- The signs associated with meaning in 3 different ways

- **Monosemic:** It is associated with one meaning. → mathematical problem, painting or a map
- **Polysemic:** It is associated with several meaning → figurative art
- **Pansemic:** It is associated any meaning → abstract art

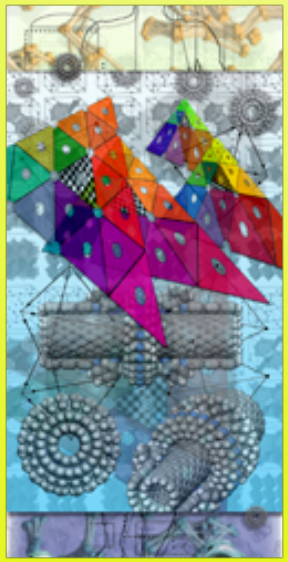


Figure: Ex: polysemic

Introduction: the essence of semio

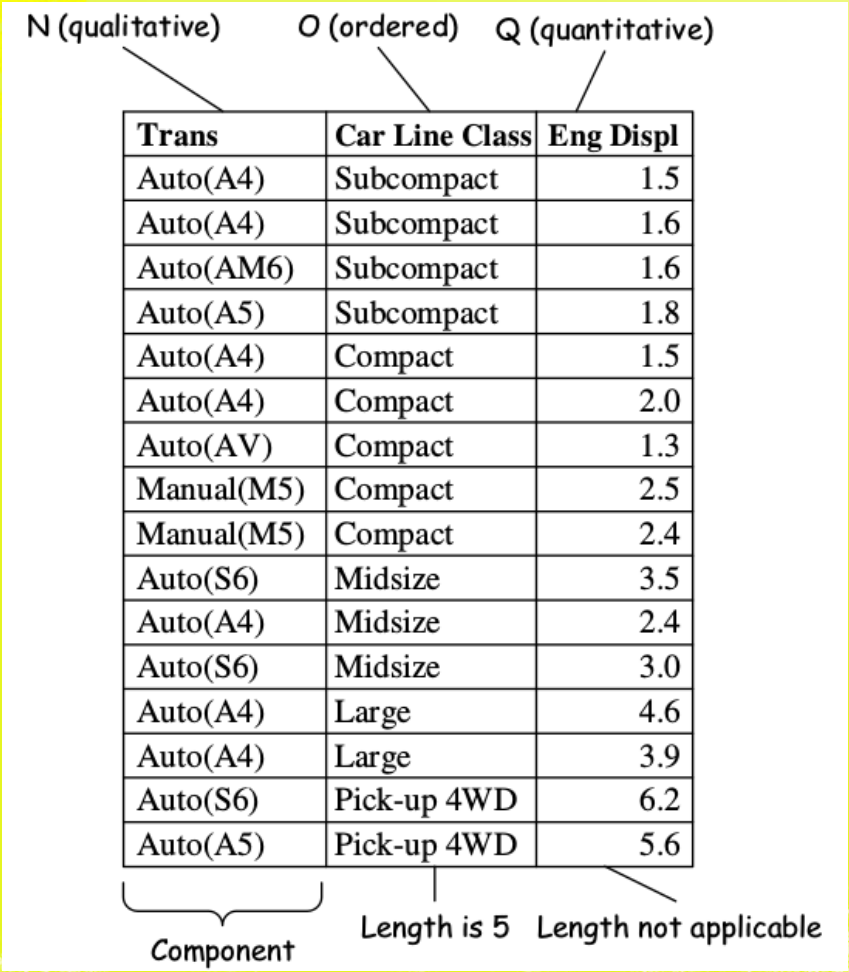
- Bertin applied the semiotic approach to diagrams, maps and networks
- 3 things must be understand to design efficient and effective graphics
  - **The properties and structure of information**
  - **The properties of graphical system**
  - **The rules for constructing an efficient graphic representation of the data**

- **The properties and structure of the information:**
  - The information used for diagrams are datasets or tables of data.
  - A diagram depicts representations of **components**, which vary across a fixed set of data called the **invariant**
  - Bertin outlined 3 factors that must be done to analyze the information before deciding how the information should be presented:
    - Determine the invariant and the number of components to include in the graphic.
    - Determine the level of organization of each included component.
    - Determine the length of each included component.

Introduction: The essence of Semiotics

- **The properties and structure of the information:**

- **Example**



Trans	Car Line Class	Eng Displ
Auto(A4)	Subcompact	1.5
Auto(A4)	Subcompact	1.6
Auto(AM6)	Subcompact	1.6
Auto(A5)	Subcompact	1.8
Auto(A4)	Compact	1.5
Auto(A4)	Compact	2.0
Auto(AV)	Compact	1.3
Manual(M5)	Compact	2.5
Manual(M5)	Compact	2.4
Auto(S6)	Midsize	3.5
Auto(A4)	Midsize	2.4
Auto(S6)	Midsize	3.0
Auto(A4)	Large	4.6
Auto(A4)	Large	3.9
Auto(S6)	Pick-up 4WD	6.2
Auto(A5)	Pick-up 4WD	5.6

- **Number of Components:**
  - A graphical image or series of images cannot be understood without knowing the **invariant** and the **information components** involved in the representations.

- Example-1:
- **Invariant:** all 2011 vehicles assessed by the EPA
- **Components:** average city fuel efficiency for each
  : car line class by
  : manufacturer

- Example - 2:
- **Invariant :** all 2011 autos assessed by the EPA
- **Components:** average city fuel efficiency by
  :car line class
  :number of gears
  :transmission types
  :engine displacements
  :number of engine cylinders

Introduction: The essence of Semiotics

- Each Data component can be perceived visually in graphic: a **mark** or more visual properties of mark.
- A mark's location in a diagram has two dimensions → Horizontal axis → Vertical axis
- simply a mark is made to represent some information other than itself. It is also referred to as a sign.
- it can be a points, lines, areas surfaces and volumes.
- These two dimensional are called **planner variables** of the graphic.
- The visual properties of the mark such as shape, color, or size are called **retinal variables**.
- The set of positional and retinal variables are called **Visual Variables**
- must be visual variable in graphic ≡ data components to be visualized
- large dataset has many data components → design efficient graphic is a problem

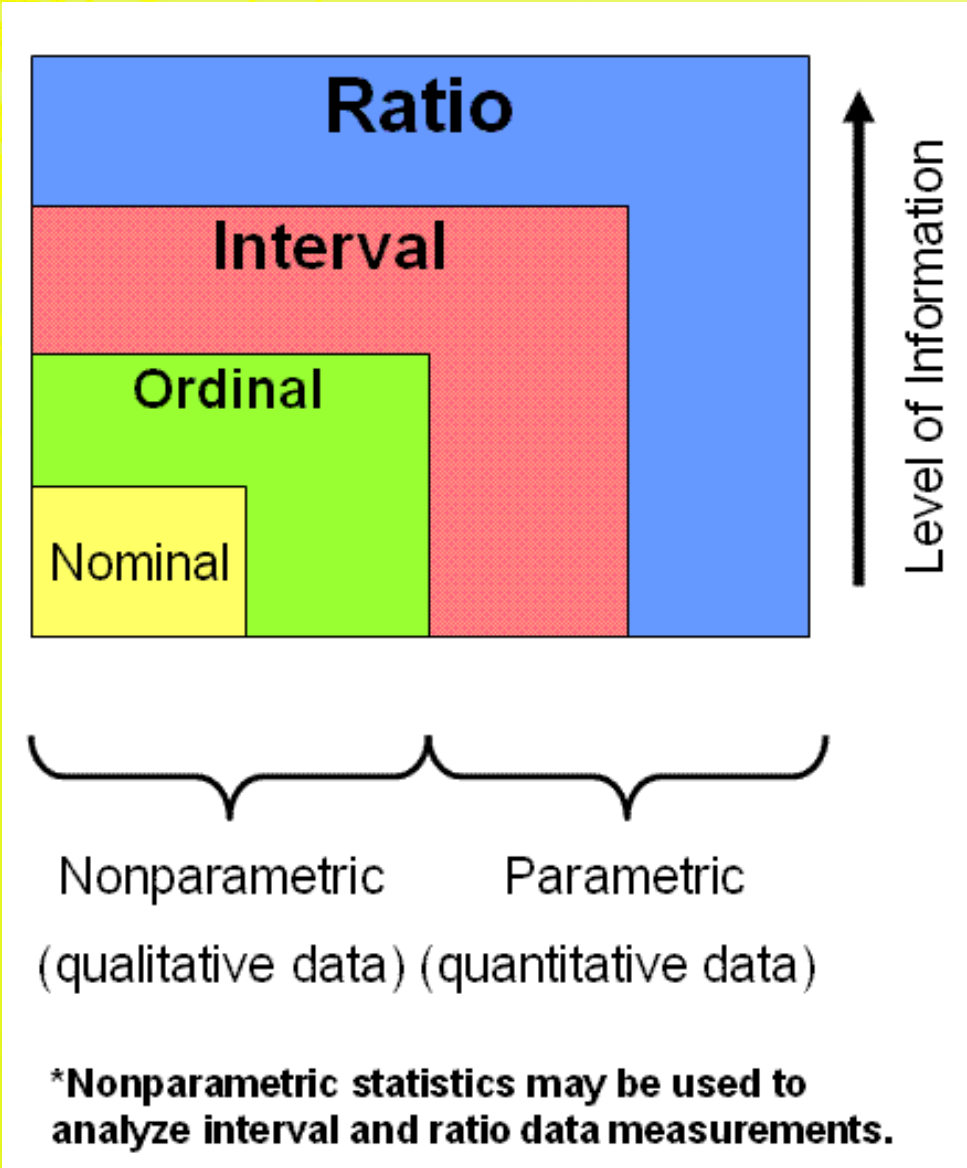
Introduction: The essence of semiotics

- The visual encoding is the way in which data is mapped into visual structures, upon which we build the images on a screen.
- Two types of Visual Encoding Variable:
  - **Planar Variables**
  - **Retinal Variables**
- Planar variables work for any data type.
- They work great to present any quantitative data.
- we can try to use Z-axis $\rightarrow$3-D Charts
- So what should we do then to present three or more variables?
- Humans are sensitive to the retinal variables.
- They easily differentiate between various colors, shapes, sizes and other properties.
- Retinal variables were introduced by Bertin, 40 years ago.

Different Types of Data in data visualization

- Quantitative Data
  - Continuous Data
  - Discrete Data
    - Interval Data
    - Ratio Data
- Qualitative Data
  - Ordinal Data
  - Nominal Data

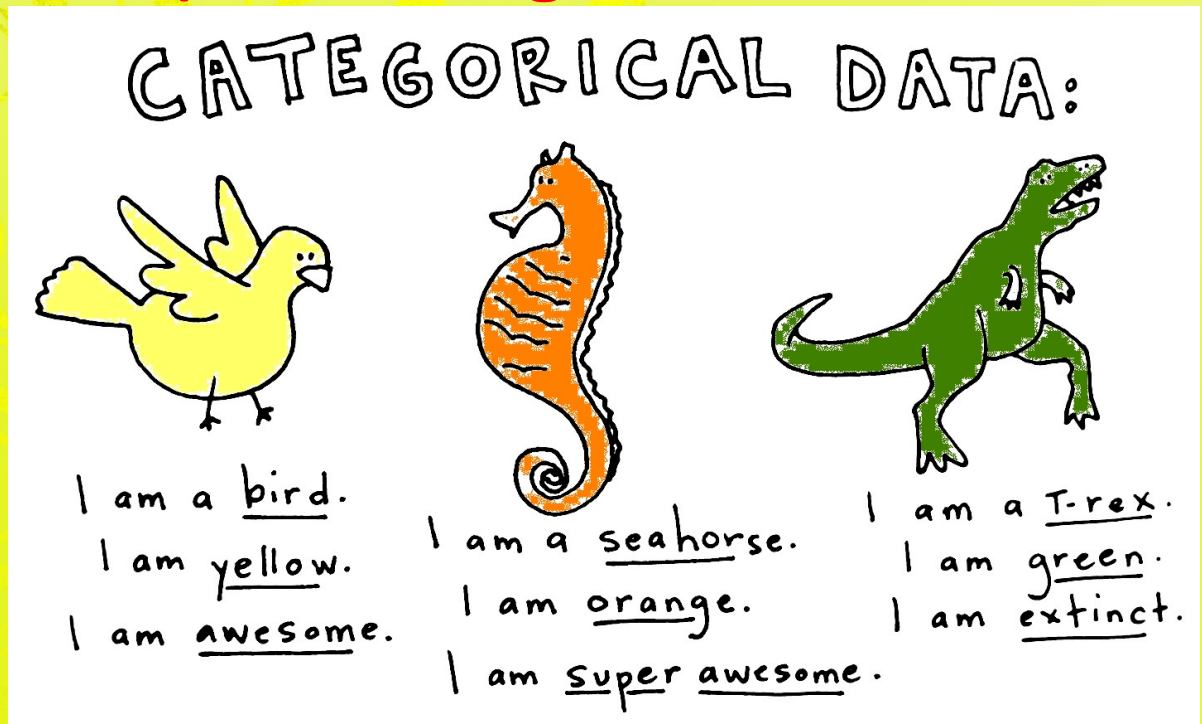
Introduction: Semiotics



Introduction: The essence of Semiotics

- **Level of Organisation:**
- A data variable is classified → 4 types → based on on the scale by which the values it contains are measured:
  - **Nominal/categorical data:**
    - The data values are categorical and not numeric.
    - A categorical variable is one that has two or more categories or labels or classes, but there is no intrinsic ordering to the categories.
    - simply Categorical variables represent types of data which may be divided into groups.
    - It is completely qualitative measurement.
    - Examples: age, gender, educational levels, countries, people names. operations: == and !=
    - Comparing two observations using the values for the variable, the observations will either be similar or different depending on whether the categorical value matches or not.

- **Example on Categorical Data:**



- if the categorical data has only two outcomes → binary or binomial data
- The Binomial data outcomes may pass/fail, live/dead or extinct/not extinct

Introduction: The essence of Semiotics

- **Examples on Categorical Variables**

	A	B	C	D	E	F	G	H	I
1	Name	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price	Country
2	Volkswagen Rabbit Dl	43,1	21,5	48	1985	4	78	2400	Germany
3	Ford Fiesta	36,1	14,4	66	1800	4	78	1900	Germany
4	Mazda GLC Deluxe	32,8	19,4	52	1985	4	78	2200	Japan
5	Datsun B210 GX	39,4	18,6	70	2070	4	78	2725	Japan
6	Honda Civic CVCC	36,1	16,4	60	1800	4	78	2250	Japan
7	Oldsmobile Cutlass	19,9	15,5	110	3365	8	78	3300	USA
8	Dodge Diplomat	19,4	13,2	140	3735	8	78	3125	USA
9	Mercury Monarch	20,2	12,8	139	3570	8	78	2850	USA

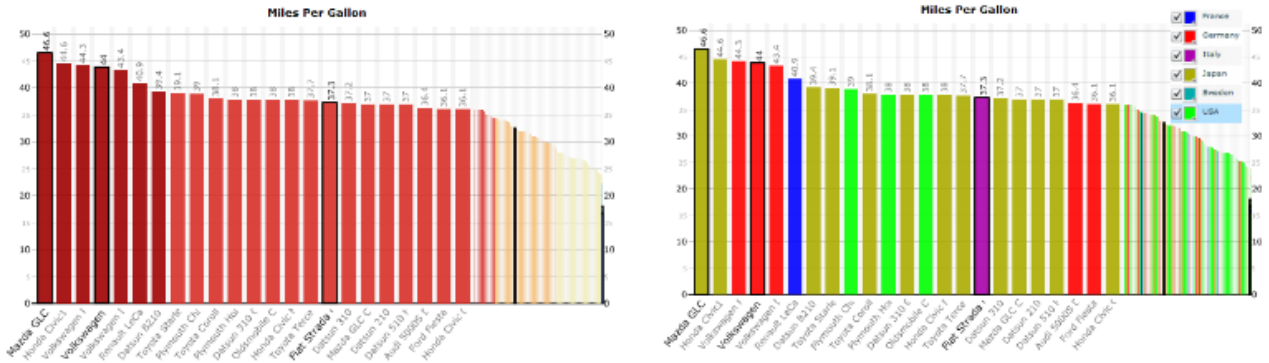


Figure: Classic car data set shown as bar chart for numerical variable "Miles per gallon" and coloured based on categorical variable Country.

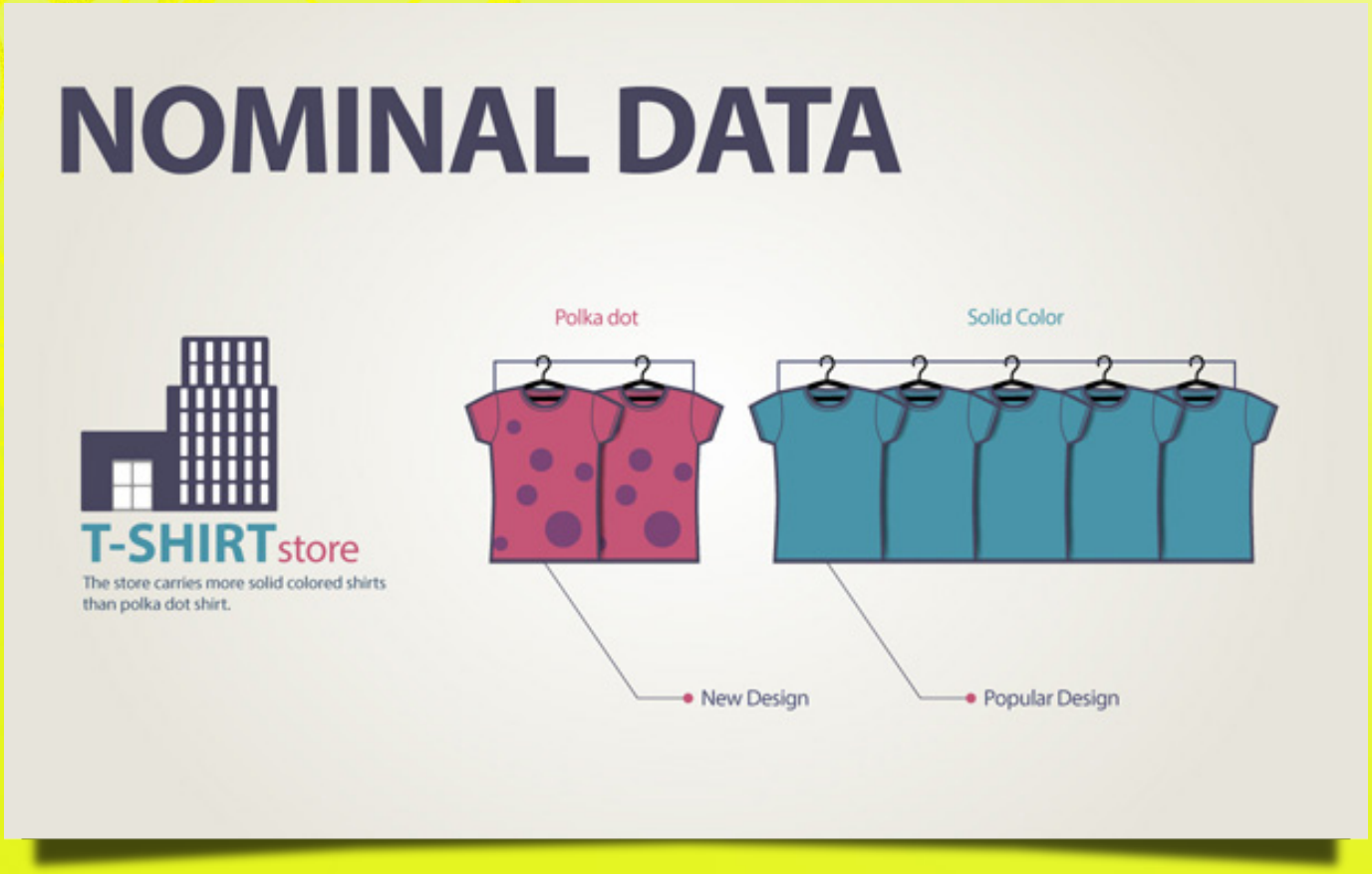


Figure: nominal level of measurement

Introduction: Semiotics

- A categorical variable (sometimes called a nominal variable) is one that has two or more categories, but there is no intrinsic ordering to the categories.
- A purely categorical variable is one that simply allows you to assign categories but you cannot clearly order the variables.
- If the variable has a clear ordering, then that variable would be an ordinal variable.
- The Nominal or categorical data has only meaning $\rightarrow$ how they are differing from one another.
- **Example:** Country names are Nominal data values $\rightarrow$ putting all country names in alphabetical order is not making any relationship to another.
- Assignment of numbers to categories has no mathematical meaning.
- Nominal categories should be mutually exclusive and exhaustive

Introduction: Semiotics

- **Where Can We Have Categorical Data:**
  - Social sciences : opinions on issues
  - Health sciences : response to treatments/drugs
  - Behavioral sciences : e.g. diagnose mental illness
  - Public health : AIDS awareness
  - Zoology : animals food preferences
  - Education : student's response to exams
  - Marketing : consumer preferences
  - Almost everywhere
- Distinction in categorical data are: Nominal Data and Ordinal Data

- **Ordinal data values:**
  - The data values are categorical but ordered.
  - Comparing two observations using the values for that variable.
  - Operations: $==, !=, \leq$ and $\geq$
  - it is mainly used for obey ordering relations among data values
  - Ordinal data is that which has inherent order, but no inherent degree of difference between what is being ordered.
  - **Example:** The I^{st}, II^{nd} and III^{rd} place winners in a race are on ordinal scale
  - But we do not know **how much faster** first place was than second place
  - But we know only that one was faster than other.

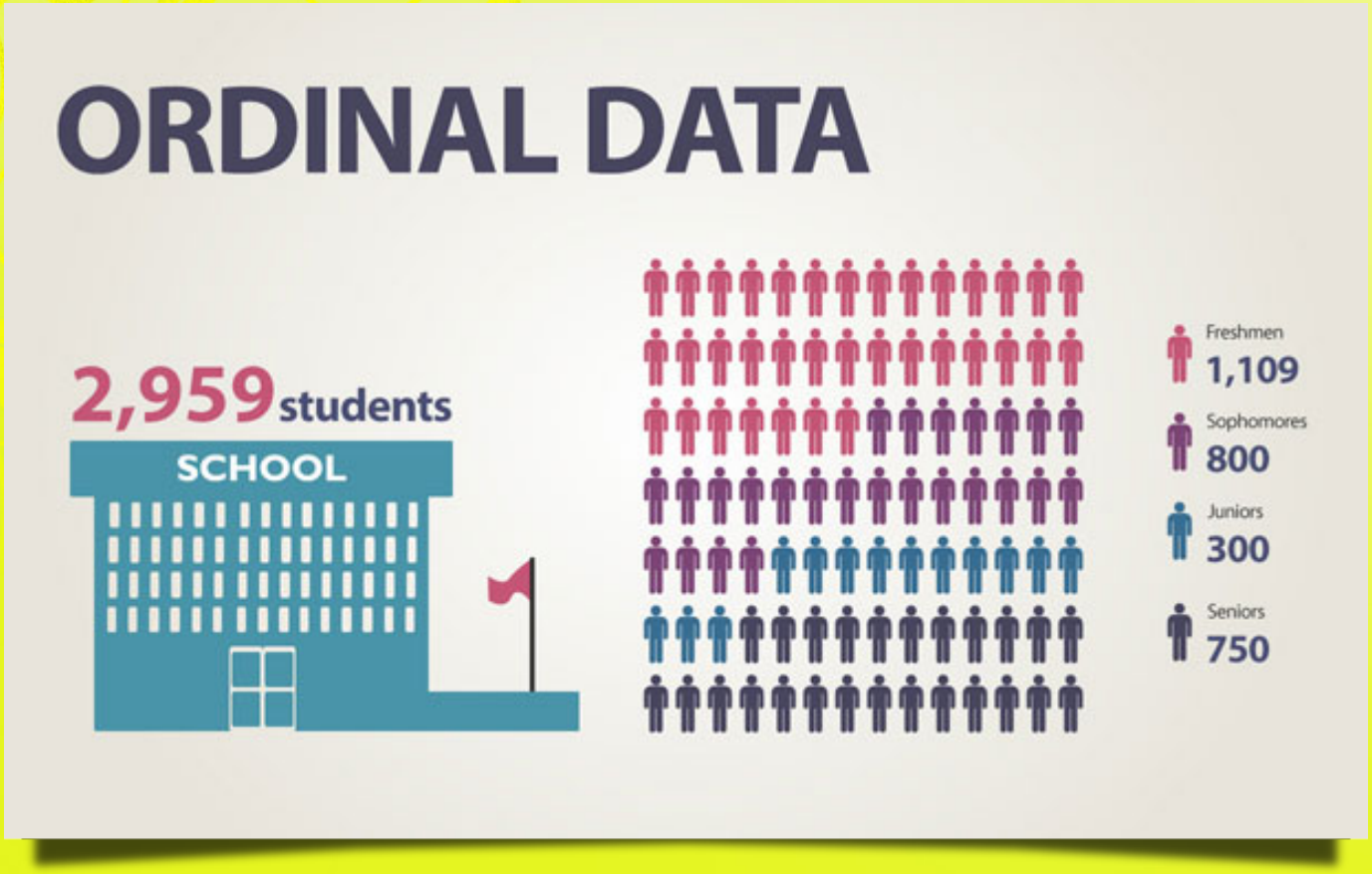
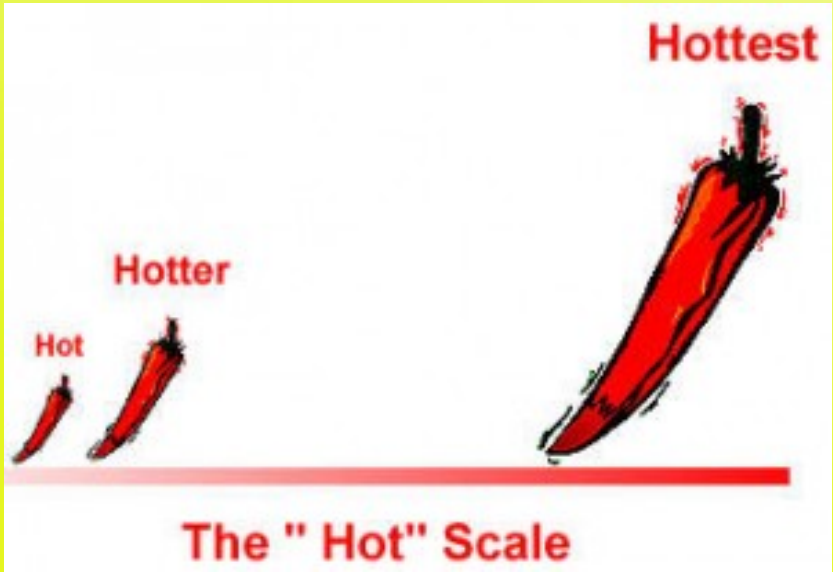


Figure: Ordinal level of measurement

Introduction: Semiotics



- **Interval Data:**
  - The data values are numeric.
  - It represents the more sensitive type of data or sophisticated form of measurement.
  - simply, Interval data is data which exists on a scale with meaningful quantitative magnitudes between values.
  -
  - Data values can be compared quantitatively using basic arithmetic operations +,-,* and / not the values themselves.
  - The values are ordered. it includes negative numbers and zero. But zero is not absolute reference point.
  - Scale data is usually aggregated or converted to averages.

- **Interval Data:**
  - **Example:1** The dataset does not contain an interval data variable, if there were a variable in a dataset that recorded the measurements of temperature. $\rightarrow$ it would be classified as a interval variable.
  - Temperature variable contains the values 40,60 and 80, we could say that compared with 40^{o}F, 80^{o}F is two times warmer than 60^{o}F (80-40)/(60-40), but not twice as hot because 0^{o}F is an arbitrarily chosen point on the scale.
  - **Example:2** if Sidda Reddy is rated as "6" on attractiveness and Durga Prasad a "3" $\rightarrow$ it does not mean Sidda Reddy is twice as attractive as Durga Prasad.

Introduction: Semiotics

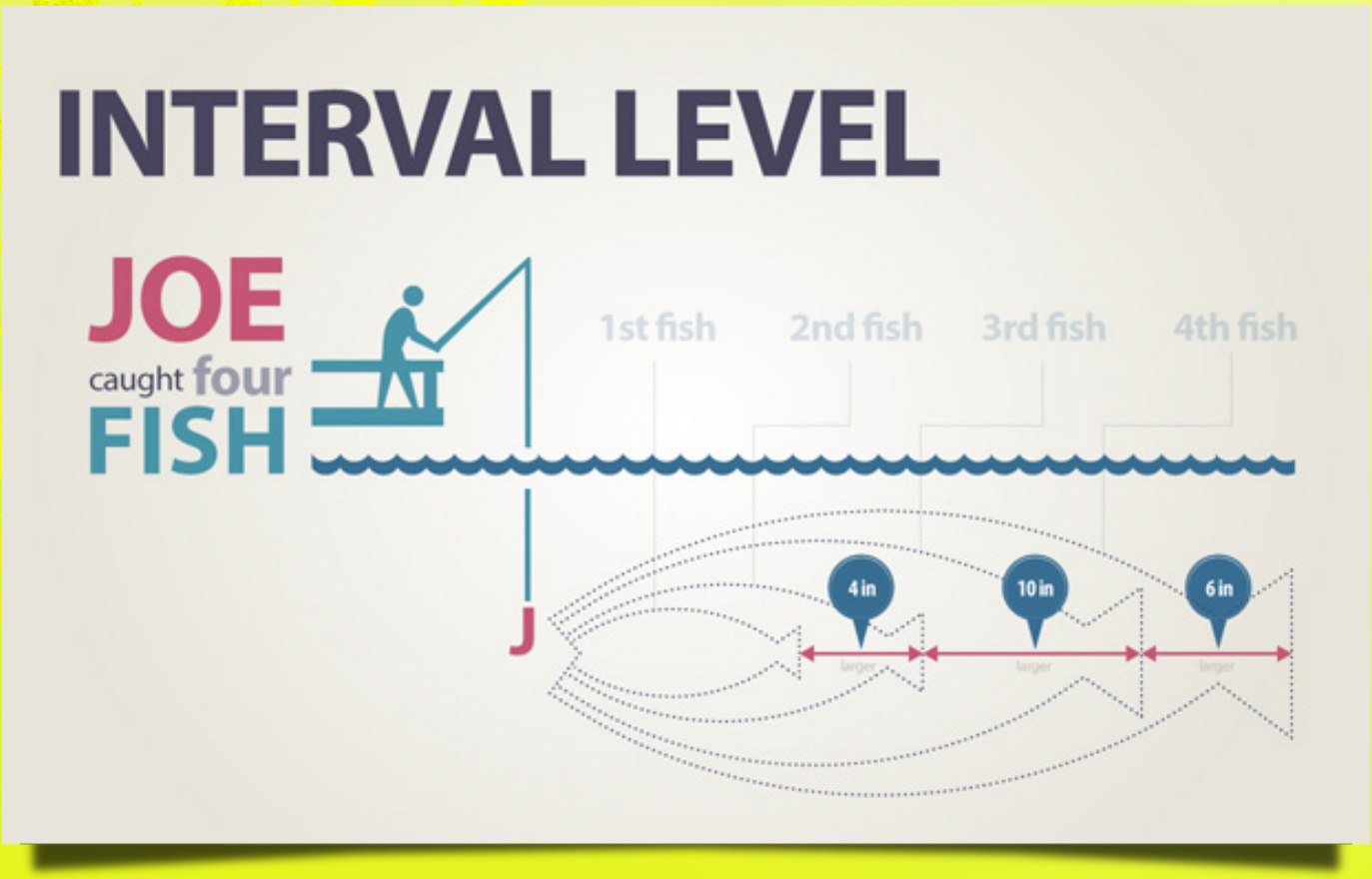


Figure: interval level of measurement

- The measurement between the sizes of the fish Joe caught in order of when he caught them.

- **Ratio Data:**
  - The Data Values are numeric and include an absolute zero.
  - This data values are allowed to compare quantitatively with other using basic arithmetic operations
  - Ratio data is data which, like interval data, has a meaningful order and a constant scale between ordered values, but additionally it has a meaningful zero value.
  - Supported Operations are ==, !=, $\leq$, $\geq$, -, / and *
  - The Ratio level of measurement applies to data that can be arranged in order.
  - In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.
  - **Example:** If one box weighs 50lbs and another 100lbs $\rightarrow$ the second box weighs twice as much as the first $\rightarrow$ this is not a case in interval data

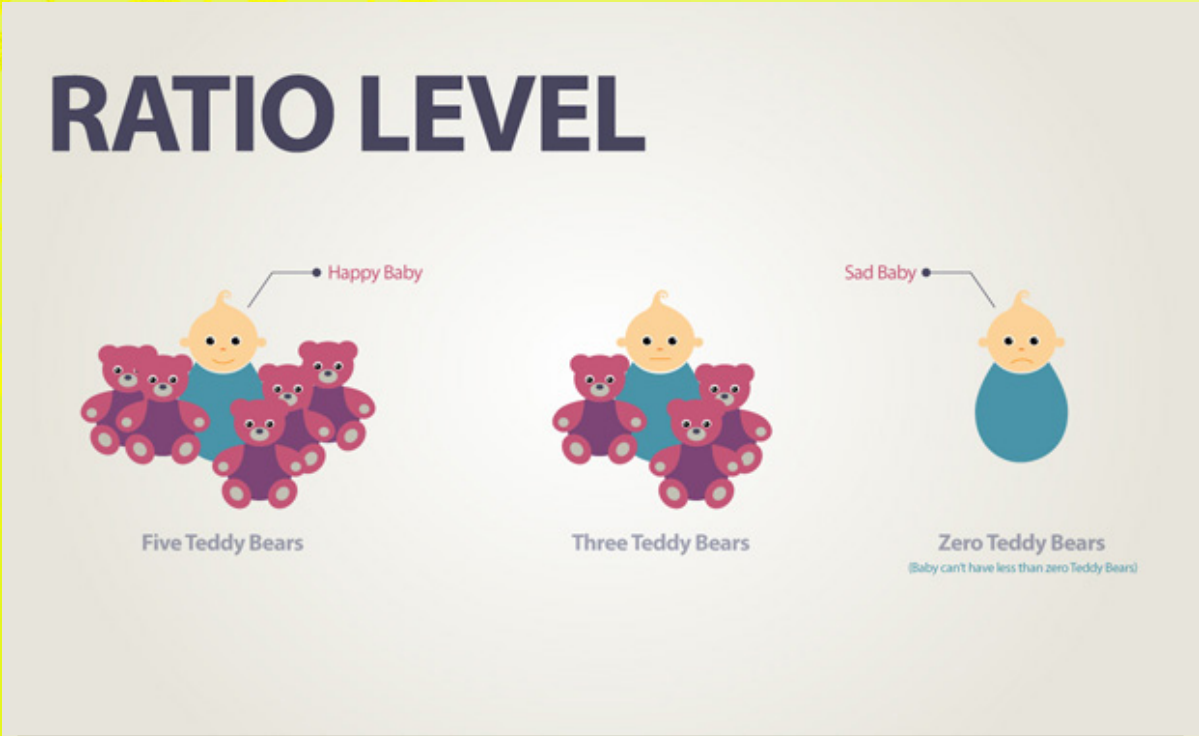


Figure: Ratio level of measurement

- The amounts of teddy bears a certain child has.
- Since we cant have less than zero teddy bears, then the ratio level has a true zero.

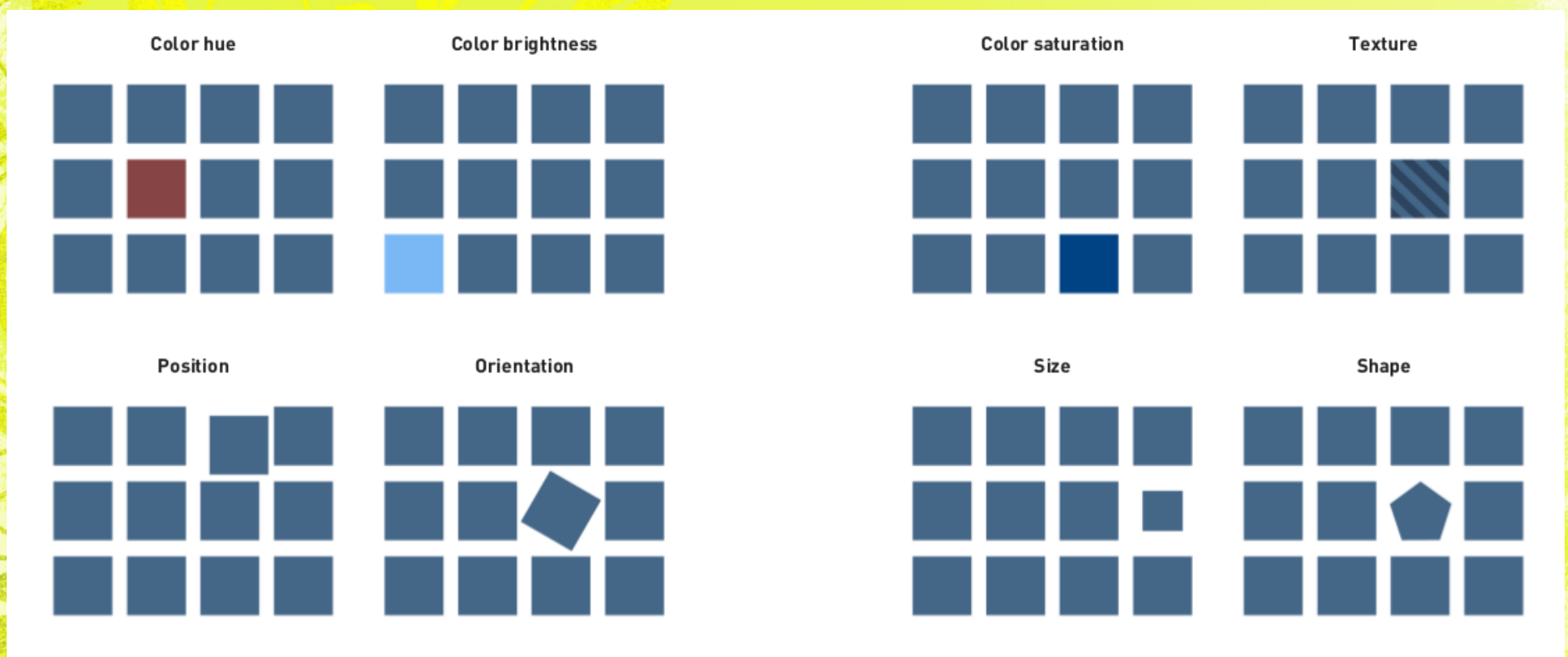
- **According to Bertin, The components are characterized by following way:**
  - Nominal Variables → Qualitative Components(N)
  - Ordinal Variables → Ordered Components(O)
  - Interval and Ration Variables → Quantitative Components(Q)

- **Length of the Components:**
  - Components may be divided
  - How they are divided depends upon the Component's level of Organization
  - **Example: Car Line Class** ordered component has length 5.
    - subcompact,compact, midsize, large and Pickup-Truck 4WD
  - Any Qualitative or Ordered component / Any Quantitative Component $\rightarrow$ not containing **Continuous** data $\rightarrow$ divided into elements or categories $\rightarrow$ The number of these is called **Component's Length**.
  - Component's Length is not applicable any Quantitative component with continuous data.
  - Component's Length is short if $<$ **4** and Long if $>$**15**

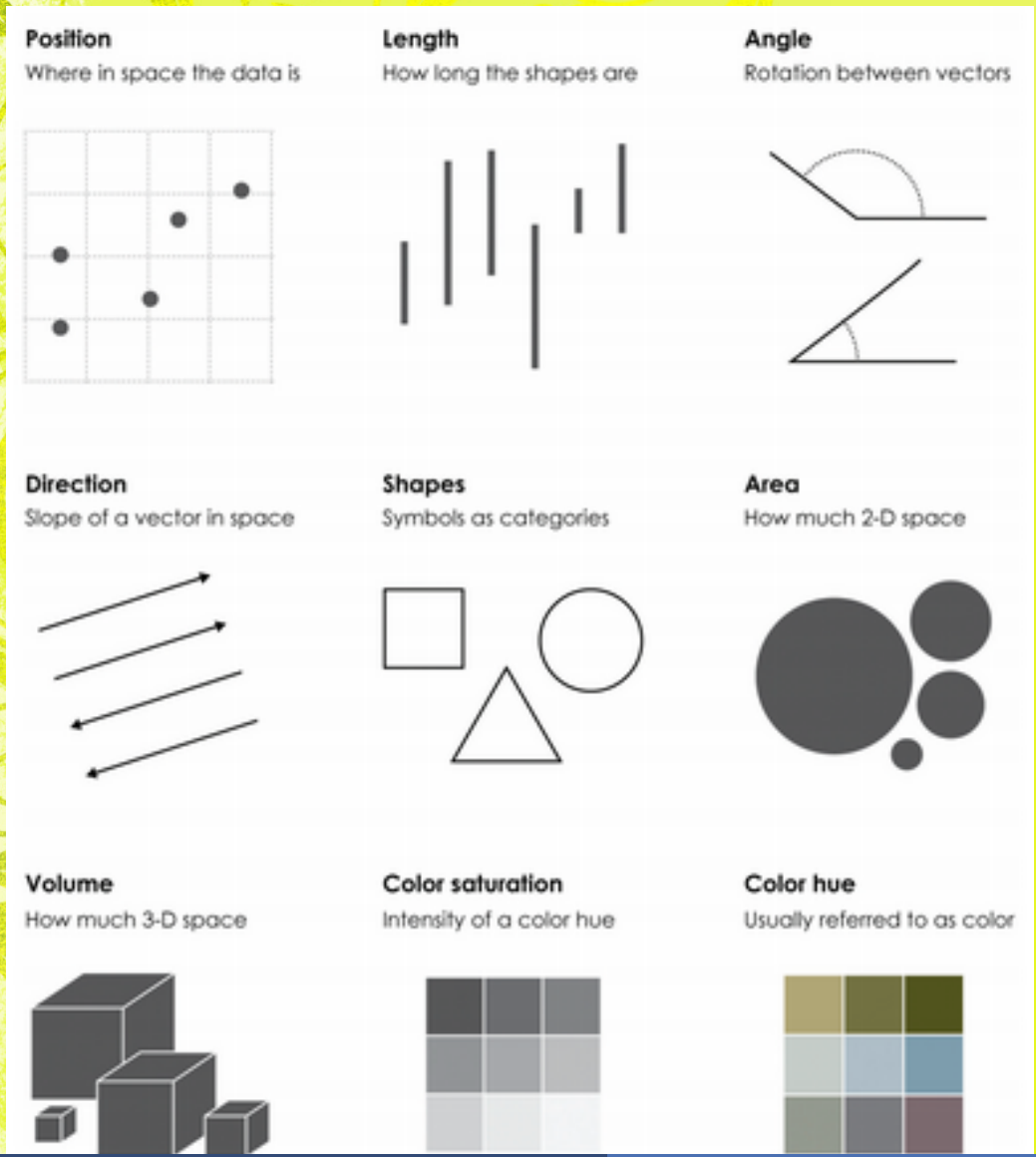
- **The Properties of Graphic System:**
  - Graphics are read in three stages:
    - Identify what is external to the Graphic
    - Identify the mapping between the visual variables and components
    - Perceive the relevant correspondences between the marks and the subset of the data that the graphic represent.
  - **External Identification:**
    - The first stage is to understand the invariant and components that are involved in the graphic.
    - Prior knowledge of domain or conventions used in data is required
    - It also need the knowledge on Graphical elements ;; color or shape
  - **Internal Identification:**
    - The next stage is to identify how the components are mapped to the visual variables.

- **Perceiving the marks:**
  - The reader perceives the meaning of each mark through its location and visual properties.
  - According to Bertin, the mark should one of three types of signs: Point, Line and Area.
  - According to Bertin, The mapping the mark to one these sign is called implementation.
  - The mark can also be styled with visual properties that vary based on values in the data components. → These properties are called → retinal properties → size, value(Brightness), Color(hue and saturation), Orientation, shape and texture → Total 8 visual variable → 2 planner variables (X and Y) and 6 visual variables.

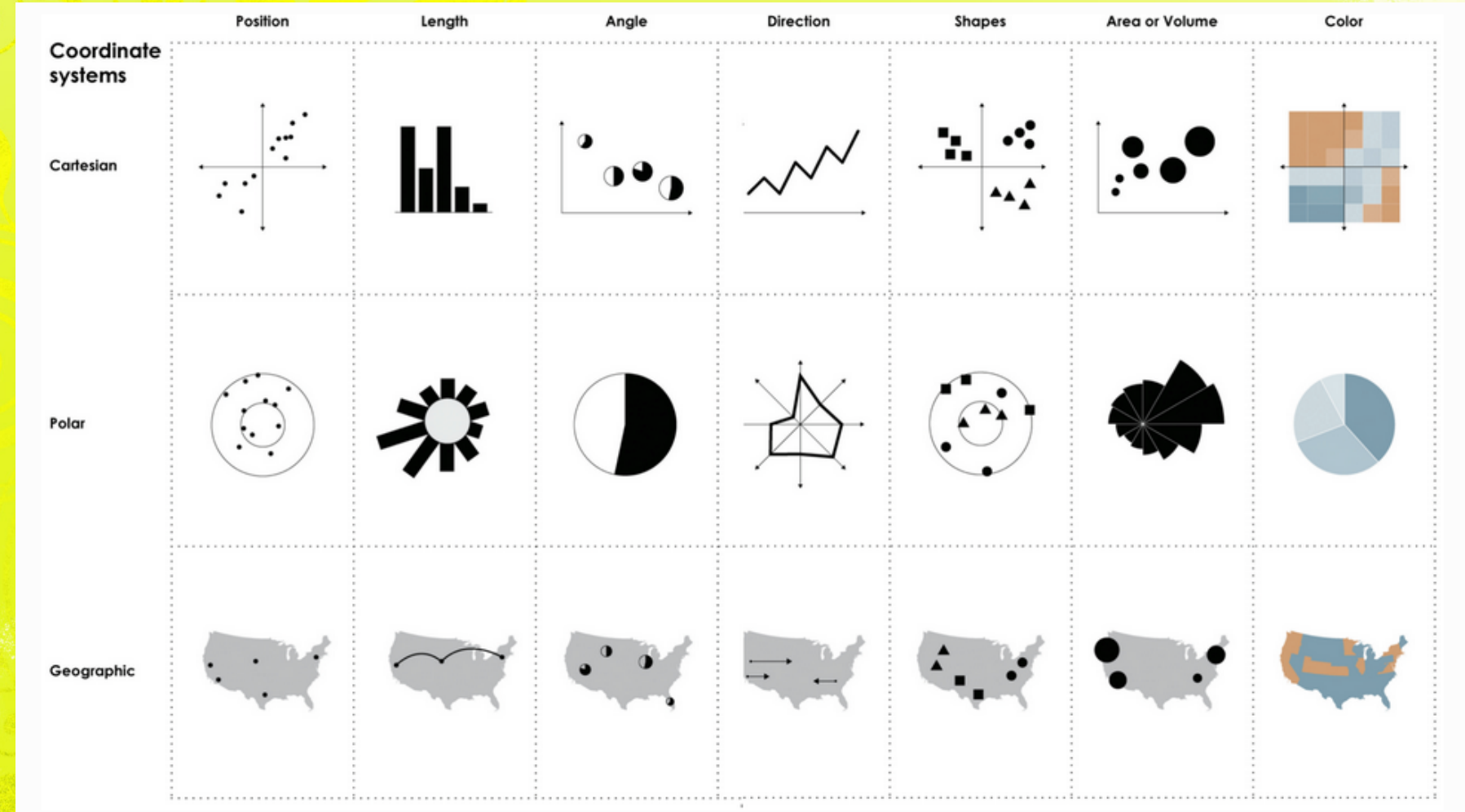
- **Visual Variable, (Bertin)**



- **Visual Variable, (Bertin)**



- **Visual Variable, (Bertin)**



- **Chronological presentation of Visual Variable, (Bertin)**

Visual Variable	Author	Example
Size	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Shape	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Lightness/ value	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Color *(hue+saturation)*	Bertin (1967/83).	
Orientation	Bertin (1967/83), Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Texture	Bertin (1967/83), Morrison (1974), MacEachre (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Tyner (2010).	
Location	Bertin (1967/83), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Hue	Morrison (1974), MacEachren (1995), Kraak & Ormeling (2003), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Saturation/ intensity	Morrison (1974), MacEachren (1995), Krygier & Wood (2005), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Arrangement	Morrison (1974), MacEachren (1995), Dent et al. (2009), Slocum et al. (2010), Tyner (2010).	
Focus/ crispness	MacEachren (1995).	
Resolution	MacEachren (1995).	
Transparency	MacEachren (1995).	
Spacing	Slocum et al (2010).	
Perspective Height	Slocum et al (2010).	

Visual Variable:

- **Location:**
  - It describes the position of the symbol relative to a coordinate frame.
- **Size:**
  - It describe the amount of space occupied by the symbol
  - Size is the primary visual variable, in propositional related visualization
  - Thickness of the Flow line $\rightarrow$ maps
- **Shape:**
  - External or Outline of the sign.
  - It is essential to the design of qualitative points in visualization
  - The shape of symbols can vary from highly abstract, such as circles, squares, or triangles to highly iconic.

Visual Variable:

- **Orientation:**
  - the direction or rotation of the symbol from normal.
  - Represents the Directionality of flow.
- **Color Hue:**
  - The dominant wavelength of the symbol on the visible portion of the electromagnetic spectrum.
  - A qualitative or spectral color scheme manipulates color hue while controlling the other components of color.
- **Color Value:**
  - The relative amount of energy emitted or reflected by the symbol.
  - Variation in color value results in the perception of shading, or areas of relative light and dark.

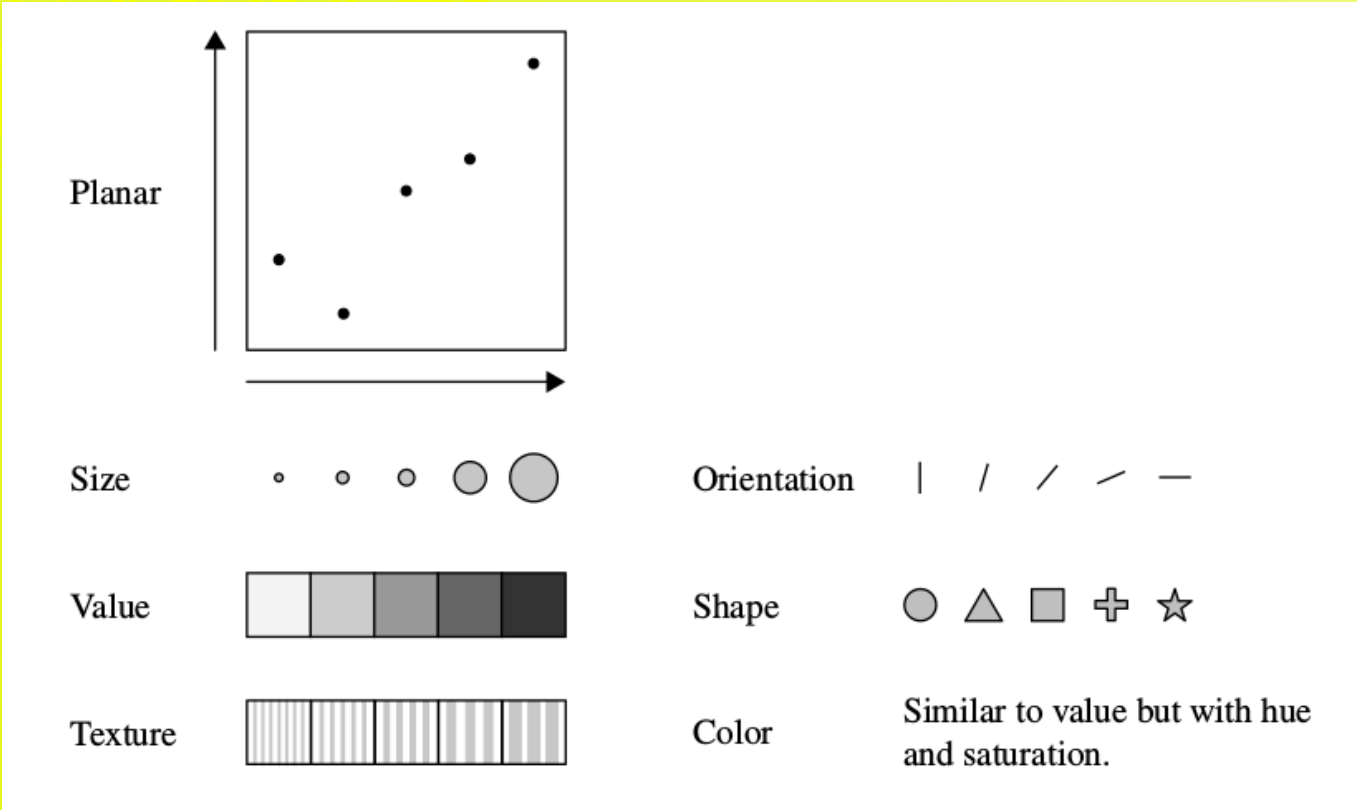
Visual Variable:

- **Texture:**
  - The coarseness of the fill pattern within the symbol
  - Represents the Directionality of flow.
- **Color saturation:**
  - The spectral peakedness of the symbol across the visible spectrum, and is the third of three visual variables associated with the perception of color.
  - Color saturation also is referred to as "chroma", "intensity", and "purity" in color theory.
- **Arrangement:**
  - The layout of graphic marks constituting a symbol
  - The visual variable arrangement varies from regular to irregular.
  - Arrangement differs from the visual variable texture.

Visual Variable:

- **Crispness:**
  - The sharpness of the boundary of the symbol.
  - Crispness also is referred to as "depth-of-field" and "fuzziness" in Information
  - crispness was the most effective visual variable for representing uncertainty in the context of point symbolization.
- **Resolution:**
  - The spatial precision at which the symbol is displayed.
  - Resolution as a visual variable leverages different levels of abstraction to encode information,
- **Transparency:**
  - The amount of graphic blending between a symbol and the background or underlying map symbols.
  - It is originally referred to transparency as "fog"

- **Length of the Visual Variable:**
  - The length of a variable is the number of perceptible divisions it supports.
  - Bertin Called these divisions as **steps**.



Visual Variables:

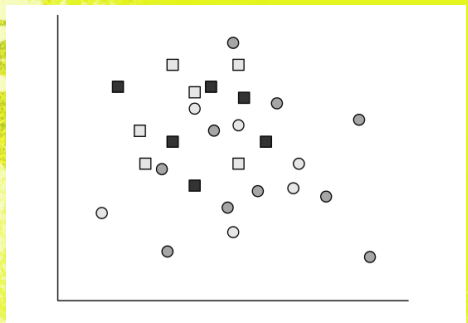
- **Level of Organization of the Visual Variable:**
  - The level of perceptual organization of a visual variable specifies its ability to convey the information of the component it represents.

TABLE 3.1 Level of Organization of Visual Variables

Variable	Associative (≡)	Selective (≠)	Ordered (O)	Quantitative
Planar	yes	yes	yes	yes
Size		yes	yes	yes
Value		yes	yes	
Texture	yes	yes	yes	
Color	yes	yes		
Orientation	yes	yes		
Shape	yes			

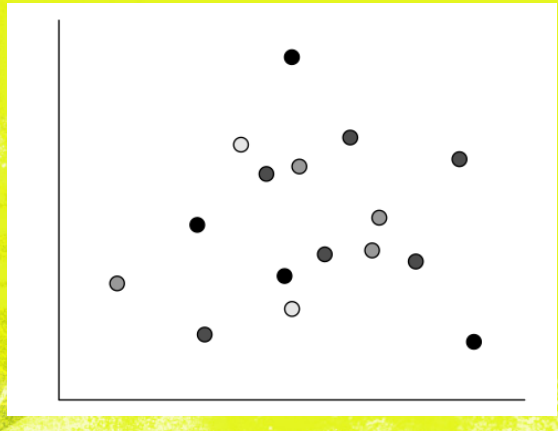
Visual Variables:

- **Level of Organization of the Visual Variable:**
  - Bertin divided the Qualitative Component into two ways based on the values it contains could be perceived
    - Reader want to see all the component's elements as a group → Associative
    - Reader want to see only the specific categories elements → Selective
  - **Associative Organization:**
    - This is the lowest level of organization
    - Variables that are associative allow all the elements of the qualitative component it represents to be instantly perceived as a group.
    - Disassociative Variables → Value



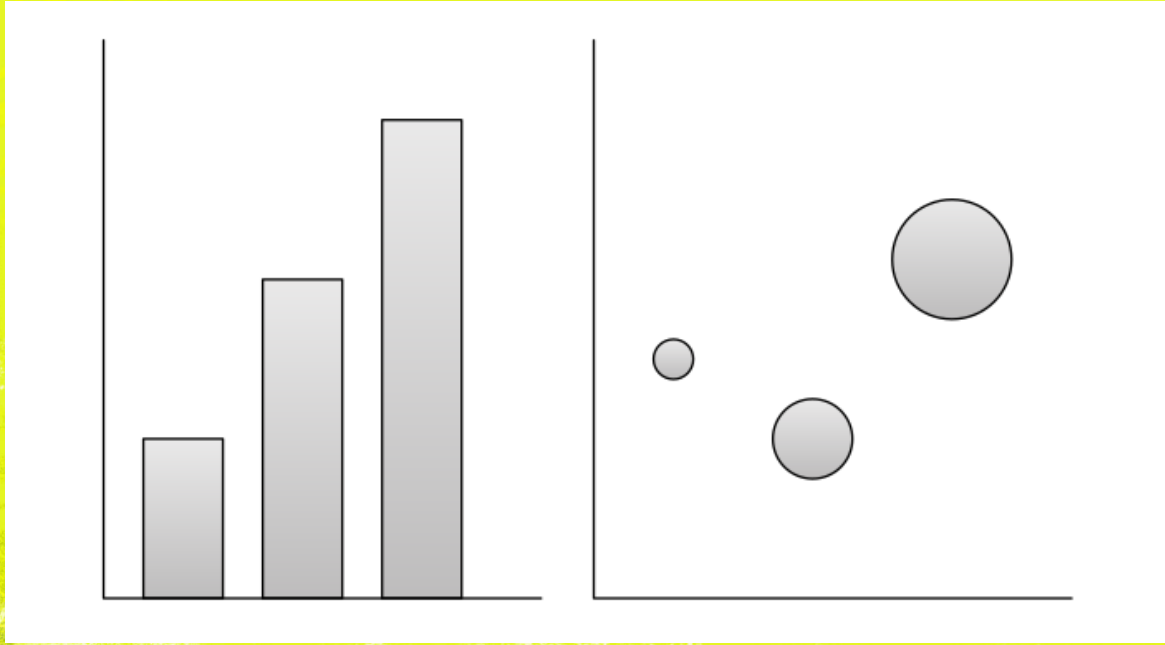
Visual Variables:

- **Level of Organization of the Visual Variable:**
  - **Selective Organization:**
    - This level allows the elements of only a specific category of the qualitative component it represents to be instantly perceived as a group.
    - It is the opposite of association.
    - Value is Selective but shape is not selective
  - **Ordered Organization:**
    - Variables that are ordered allow comparisons of relative magnitude to be made about the data values they signify because the steps can be perceived as increasing or decreasing.

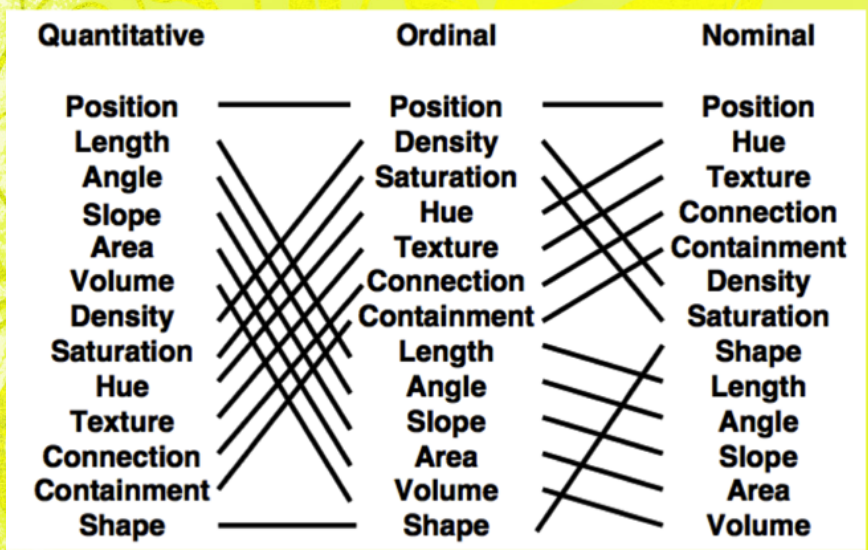


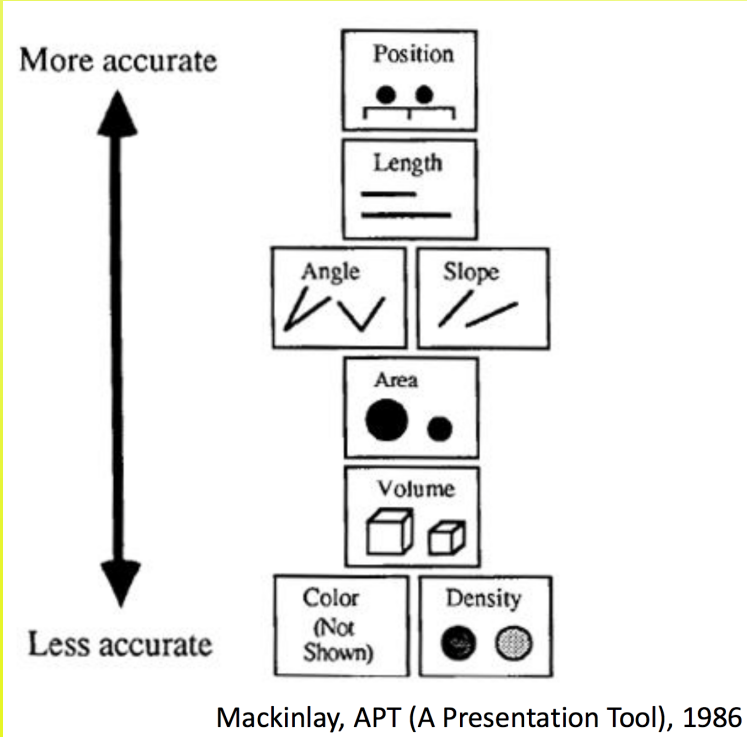
Visual Variables:

- **Level of Organization of the Visual Variable:**
  - **Quantitative Organization:**
    - This is the highest level.
    - Variables that are quantitative allow the ratios of the values of the component they represent to be compared because of the existence of an absolute zero.
    - These comparisons can be perceived directly without the need for a legend.



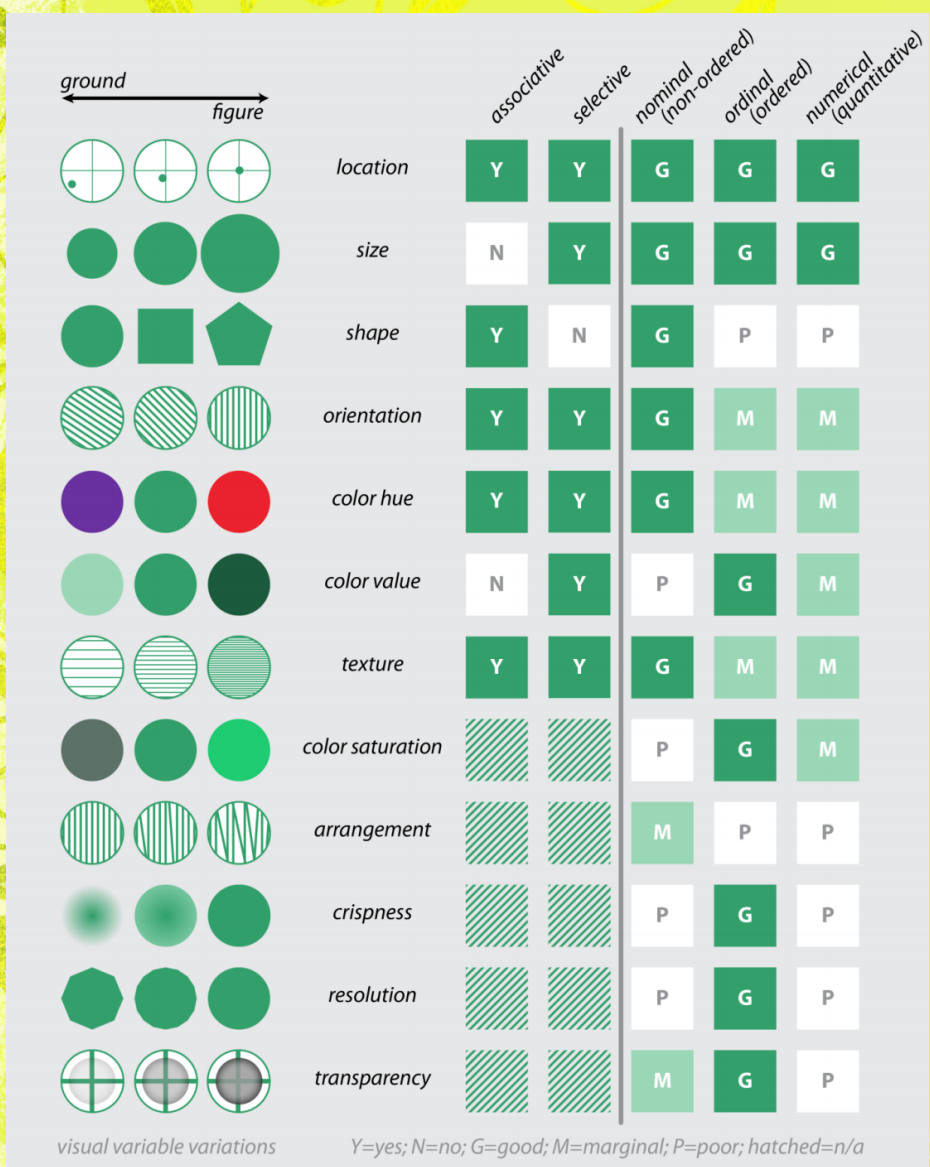
Mackinlay Effectiveness Criteria:





Mackinlay, APT (A Presentation Tool), 1986

- **Bertin's visual variables and their syntactics.**



Funny: Brain Work:

Look at the chart and say the COLOUR not the word

YELLOW BLUE ORANGE
BLACK RED GREEN
PURPLE YELLOW RED
ORANGE GREEN BLACK
BLUE RED PURPLE
GREEN BLUE ORANGE

Left – Right Conflict

Your right brain tries to say the colour but your left brain insists on reading the word.

How many 3's?

248721840123874092165901476098560
932472091256290650985265904827582
985680960986309584390564095878950
374509284750989475092984

How many 3's?

2487218401238740921659014760985609324720912562906509852659048275829856809609863095843905640958789503745092847509894750929 84

Introduction to Datasets

- A data set is defined as a collection of data items.
- The characteristics of a data item are described by a collection of variables.
- A variable can be defined as a property, or a characteristic, of a data item that may vary from one item to another or over time.
- Example: Data Item $\rightarrow$ country :: Variable $\rightarrow$ various characters of countries like population size, ageing groups etc..
- A multivariate data set is simply a data set including two or more variables.
- The items of a multivariate data set can be thought of as points in a multidimensional space where each dimension represents a variable.
- A standard format used for structuring multivariate data is to use an m-by-n matrix including m rows, usually representing data items, and n columns, usually representing variables.