

# Introduction to Data Visualization with R using ggplot2

Richard Johansen

Zhiyuan Yao

Jennifer Latessa

# Workshop Agenda

**Workshop Expectations**

**Understanding Data**

**Visualizations**

**The Scenario**

**Why R and ggplot2?**

# Workshop Agenda

## **Workshop Expectations**

Understanding Data

Visualizations

The Scenario

Why R and ggplot2?

# Workshop Expectations

- Assumptions:
  - You have some basic coding experience or familiarity with a coding environment
  - You have some general data visualization knowledge (i.e. What is a scatter plot)
- Prerequisites:
  - R and R studio Installed
  - Install the ggplot2 package
- Caveat
  - This is only a simple introduction to data visualization with R & ggplot2
  - We cannot cover everything but we will cover many of the common aspects you will use the majority of the time
- Goal
  - Conduct basic data exploration and data visualization
  - Allow you to (re)produce print-quality graphics in seconds

# Workshop Agenda

Workshop Expectations

**Understanding Data**

Visualizations

The Scenario

Why R and ggplot2?

# What is Data?

- Data is a collection of **objects** defined by **attributes**
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Synonyms: variables, fields, characteristics, features, columns, etc.
- A collection of attributes describe an object
  - Synonyms: records, points, cases, samples, instances, rows, etc.

# Attribute Values

- Each attribute has a potential set of values objects draw from.
- The same attribute can be mapped to different attribute values
  - Example: height can be measured in meters or feet
- Different attributes can be mapped to the same set of values
  - Example: Attribute values for ID and age are both integers

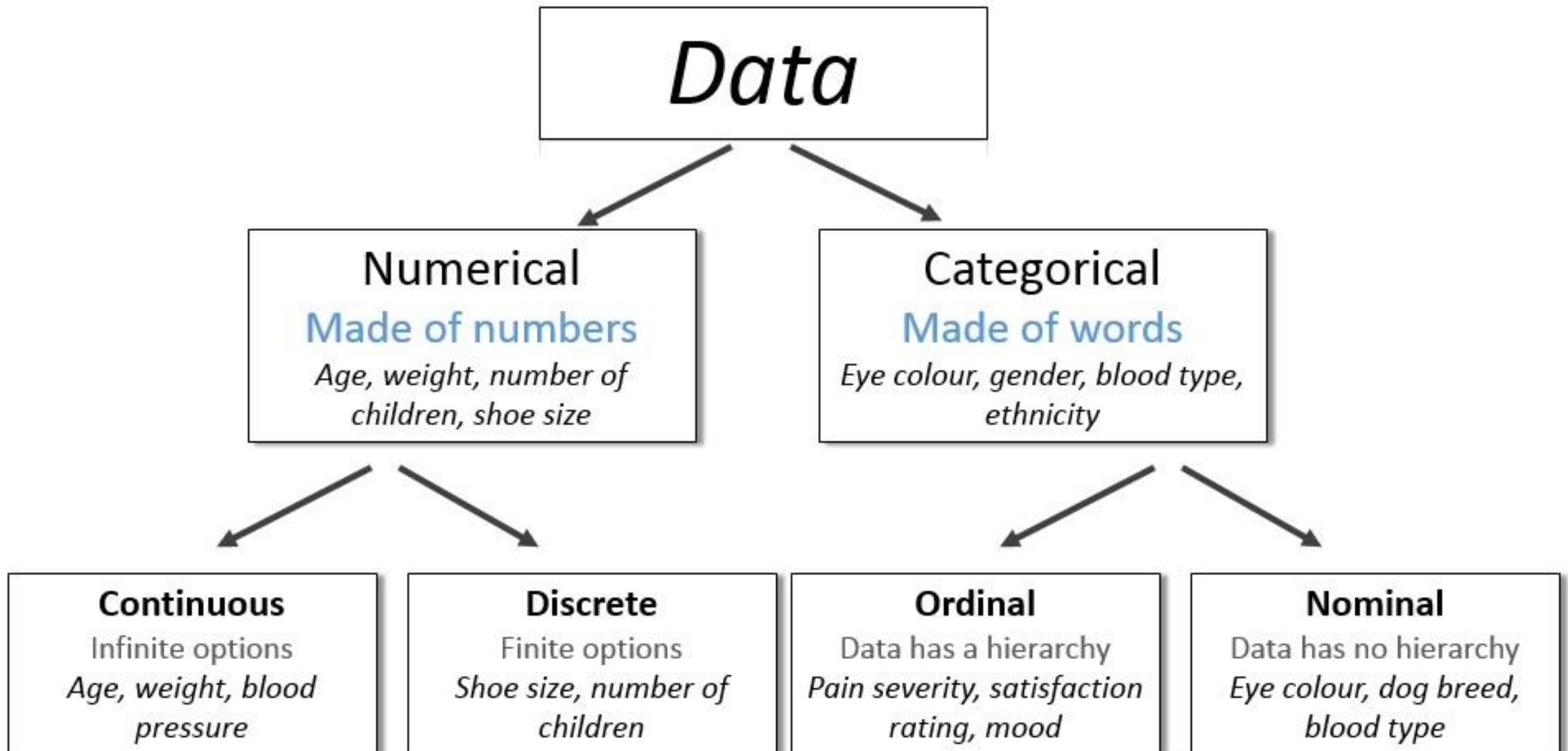
# Attribute Classification

- Discrete Attribute – has a infinite or countably infinite set of values
  - Examples: zip codes, number of words,
  - Typically represented as integers
- Continuous Attribute - has real numbers as attribute values
  - Example: temperature, height, weight
  - Typically represented as floating point (decimal)



# Important Attribute Classes

- Categorical
  - Nominal - Data that can be counted, but not aggregated or ordered
    - Examples: Eye Color, Zip Code, Music Genre
  - Ordinal - Data that can be counted and ordered, but not aggregated.
    - Examples: Grades, Clothing Size, Positions (in a race)
- Numerical
  - Interval (metrics) - The difference in values are constant and meaningful
    - Examples: The difference between a temperature of 100°F and 90°F is the same difference as between 90°F and 80°F.
  - Ratio - An interval scale with an absolute zero
    - Examples: Income, Height, Weight



# Data Quality

- Main Issues with Data Quality:
  - Noise and Outliers
  - Missing Values
  - Duplicate Data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

# Primitive Data Types

- Boolean:
  - True (T) or False (F)
- Char:
  - Characters and Strings – “A”, “Beta”, “There are different data types!”
- Factors:
  - Ordinal Data – 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup> Or High, Medium, Low
- Int:
  - Integers – (1, 2, 100)
- Float/Double:
  - Decimal – (0.1, 0.2, 0.1352)

# Workshop Agenda

Workshop Expectations

Understanding Data

**Visualizations**

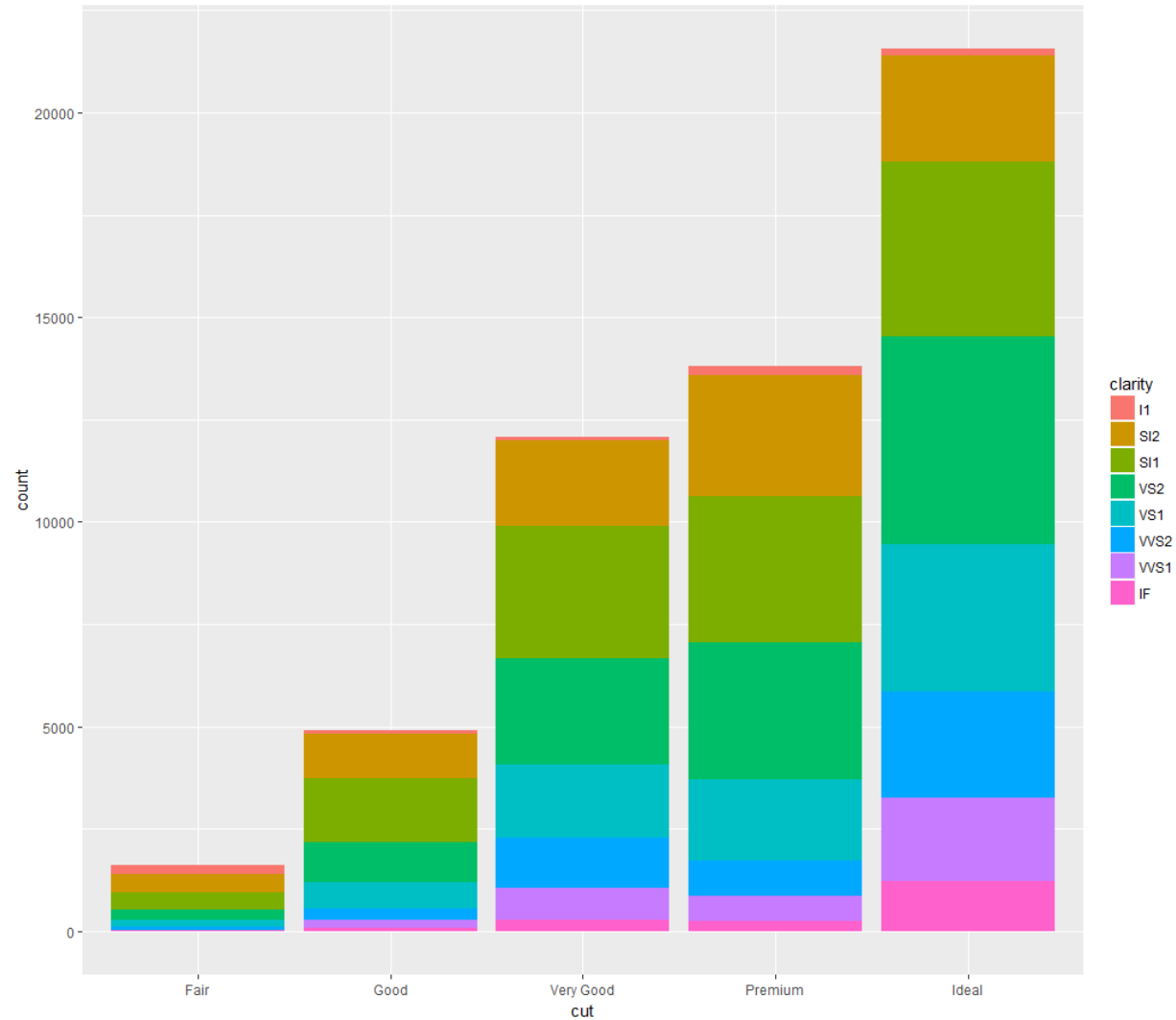
The Scenario

Why R and ggplot2?

# What is Data Visualization

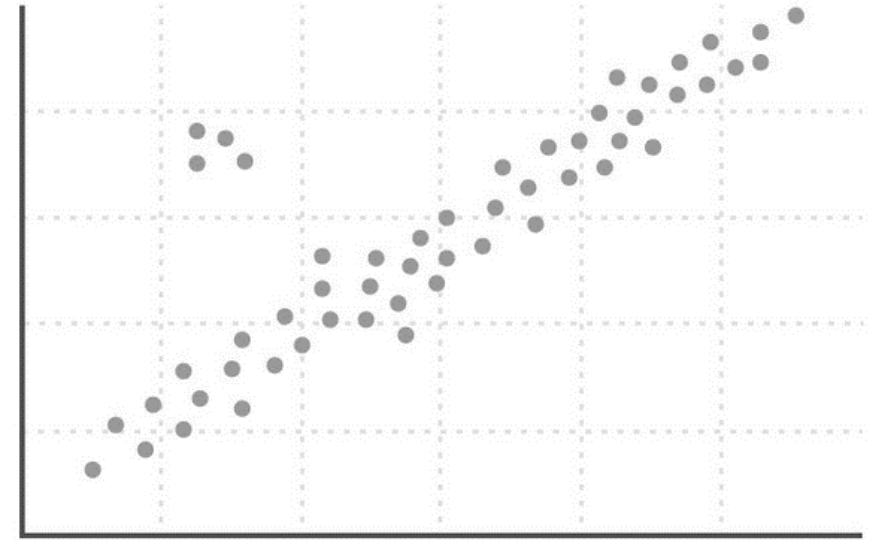
- “Visualization transforms data into images that effectively and accurately represent information about the data.”
  - Schroeder et al. The Visualization Toolkit, 2<sup>nd</sup> ed. 1998
- “Data graphics visually display measured quantities by means of the combined use of points, lines, a coordinate system, numbers, symbols, words, shading, and color.”
  - Edward Tufte The Visual Display of Quantitative Information (2<sup>nd</sup> Ed.)

# Why is Data Visualization Important?



# Why Do Visualization

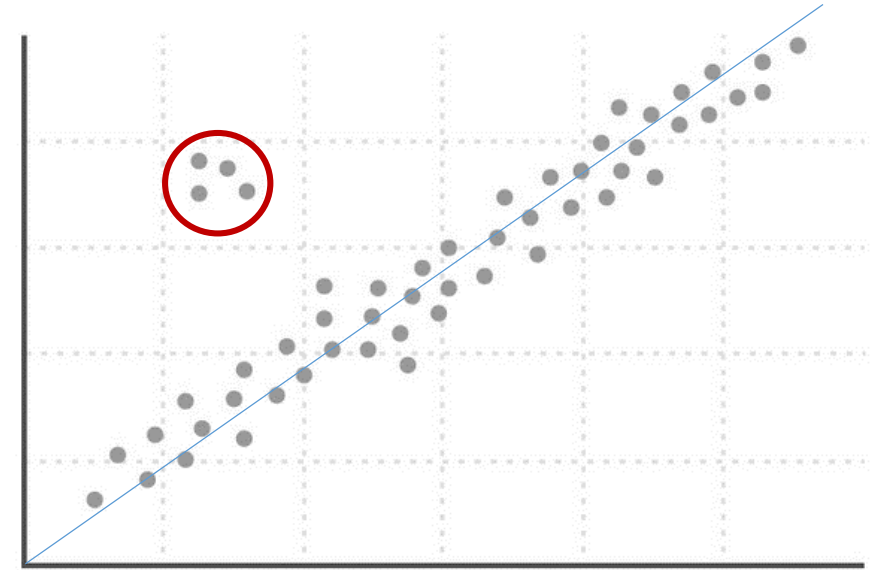
- Reasons for doing data visualization
- Exploration
  - Use visualizations as a means of data exploration
- Analysis
  - Verify (or Falsify) a hypothesis
- Presentation
  - Visualization is used to communicate results or findings





# Why Do Visualization

- Reasons for doing data visualization
- Exploration
  - Use visualizations as a means of data exploration
- Analysis
  - Verify (or Falsify) a hypothesis
- Presentation
  - Visualization is used to communicate results or findings



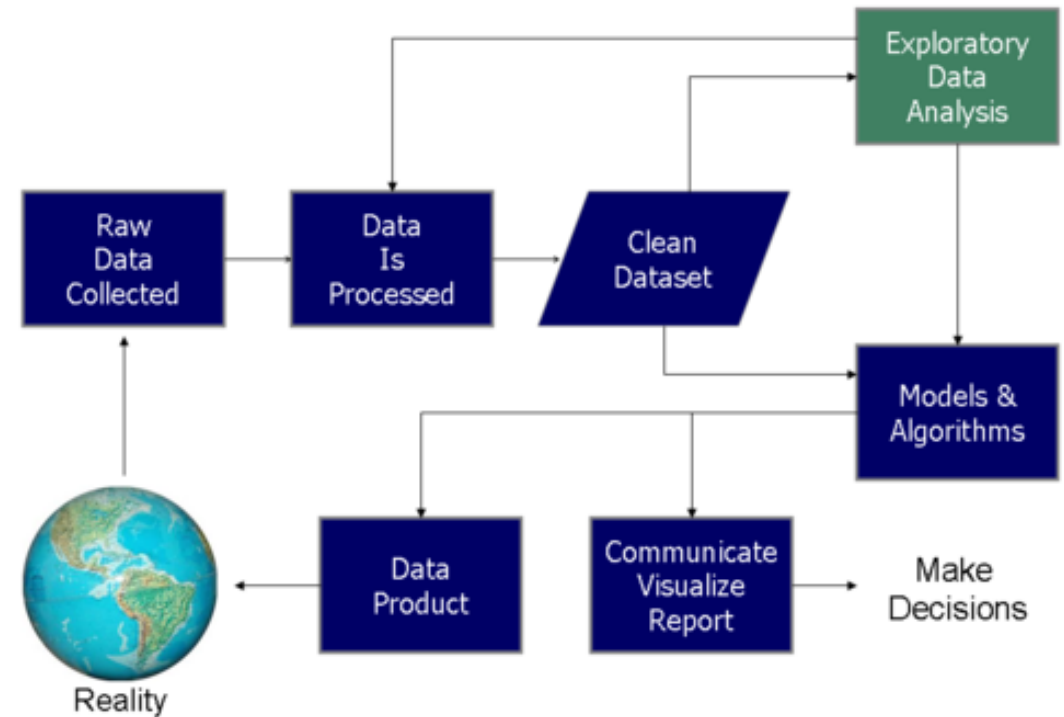
# Why Do Visualization

- Reasons for doing data visualization
- Exploration
  - Use visualizations as a means of data exploration
- Analysis
  - Verify (or Falsify) a hypothesis
- Presentation
  - Visualization is used to communicate results or findings



# Data Science Process

- Acquire Data
  - “Know your data”
- Clean and Pre-Process
- Visualize (explore)
- Model/Analyze
- Communicate Findings/  
Data Production



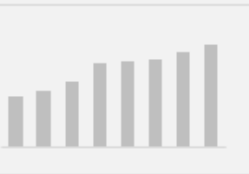

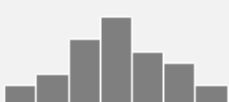


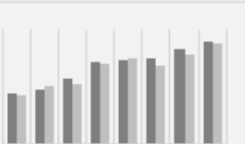







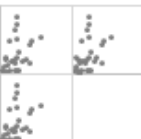











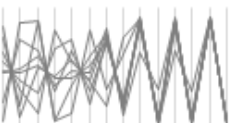
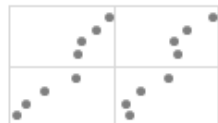
[https://commons.wikimedia.org/wiki/File:Data\\_visualization\\_process\\_v1.png](https://commons.wikimedia.org/wiki/File:Data_visualization_process_v1.png)

# Chart Types

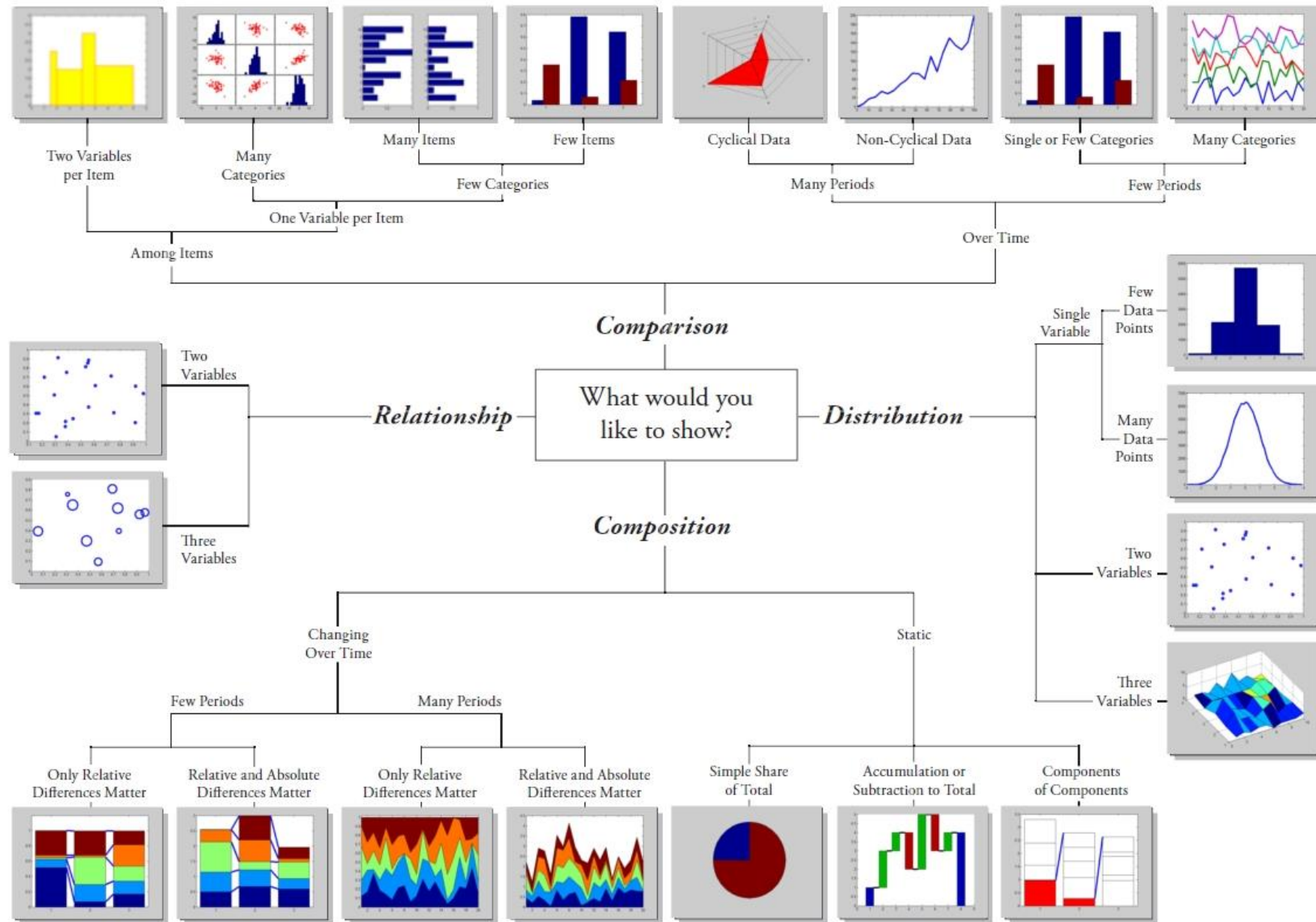
- Common Chart Types
  - Comparison - comparing and sorting data points;
  - Composition - part-to-whole comparisons;
  - Distribution - comparison of data points along an axis;
  - Relationship - relationship patterns between two or more variables;

## Data comparison charts

## Data reduction charts

Comparison		Composition	Distribution	Evolution	Relationship	Profiling	
<b>Bars</b> 		<b>Pie</b> 	<b>Histogram</b> 	<b>Line</b> 	<b>Scatterplot</b> 	<b>Grouped bars</b> 	
<b>Dot plot</b> 	<b>Bullet</b> 	<b>Pareto</b> 	<b>ID Scatterplot</b> 	<b>Horizon</b> 	<b>Connected Scatterplot</b> 	<b>Cycle plot</b> 	<b>Scatterplot matrix</b> 
<b>ID Scatterplot</b> 	<b>Heat map</b> 	<b>Multidimensional Pie</b> 	<b>Boxplot</b> 	<b>Step</b> 	<b>Bubble</b> 	<b>Reorderable matrix</b> 	<b>Horizon</b> 
<b>Slope</b> 	<b>Alert</b> 			<b>Connected Scatterplot</b> 		<b>Parallel Plot</b> 	<b>Trellis</b> 

# Chart Suggestions—A Thought-Starter



## Measure: ascertain the size, amount, or degree of (something)



A bar graph uses either horizontal or vertical bars to show comparisons among categories. They are valuable to identify broad differences between categories at a glance.



A treemap shows both the hierarchical data as a proportion of a whole and, the structure of data. The proportion of categories can easily be compared by their size.



Bubble charts represent numerical values of variables by area. With two variables (category and numeric), the circles placed so they are packed together.



A heat chart shows total frequency in a matrix. Values in each cell of the rectangular grid are symbolized into classes.

## Relationship: a connection or similarity between two or more things or, the state of being related to something else



A choropleth map allows quantitative values to be mapped by area. They should show normalized values not counts collected over unequal areas or populations.



A chord diagram visualizes the inter-relationships between categories and allows comparison of similarities within a dataset or, between different groups of data.



Scatterplots allow you to look at relationships between two numeric variables with both scales showing quantitative variables. The level of correlation can also be quantified.



Spider lines, also termed desire lines, show paths between origins and destinations. They show connections between places.

## Change: process through which something becomes different, often over time



A bar graph uses either horizontal or vertical bars to show comparisons among categories. They are valuable to identify broad differences between categories at a glance.



A heat chart shows total frequency in a matrix. Using a temporal axis values, each cell of the rectangular grid are symbolized into classes over time.



Bubble charts with three numeric variables are multivariate charts that show the relationship between two values while a third value is shown by the circle area.



Graduated symbol maps show a quantitative difference between mapped features by varying symbol size. Data are classified with a symbol assigned to each range.



A Density/heat map calculates spatial concentrations of events or values enabling the distribution to be visualized as a continuous surface.



A Data clock creates a circular chart of temporal data, commonly used to see the number of events at different periods of time.



Line graphs visualize a sequence of continuous numeric values and are used primarily for trends over time. They show overall trends and changes from one value to the next.



A combo chart combines two graphs where they share common information on the x-axis. They allow relationships between two datasets to be shown.

## Interaction: flow of information, products or goods between places



A chord diagram visualizes the inter-relationships between categories and allows comparison of similarities within a dataset or, between different groups of data.



Spider lines, also termed desire lines, show paths between origins and destinations. They show connections and flow between places.

## Distribution: the arrangement of phenomena, could be numerically or spatially



Histograms show the distribution of a numeric variable. The bar represents the range of the class bin with the height showing the number of data points in the class bin.



A box plot displays data distribution showing the median, upper and lower quartiles, min and max values and, outliers. Distributions between many groups can be compared.



A choropleth map allows quantitative values to be mapped by area. They should show normalized values not counts collected over unequal areas or populations.



Graduated symbol maps show a quantitative difference between mapped features by varying symbol size. Data are classified with a symbol assigned to each range.



A Density/heat map calculates spatial concentrations of events or values enabling the distribution to be visualized as a continuous surface.



A unique symbol map (areas or points) allows descriptive (qualitative) information to be shown by location. Areas have different fills and points can be geometric or pictorial.

## Part-to-whole: relative proportions or percentages of categories, showing the relationship between parts and whole



Donut charts are used to show the proportions of categorical data, with the size of each piece representing the proportion of each category.



A treemap shows both the hierarchical data as a proportion of a whole and, the structure of data. The proportion of categories can easily be compared by their size.

### Acknowledgement

Inspired by work by Jon Schwabish and Severino Ribeca, The Graphic Continuum, 2014 and, Alan Smith et al. Visual Vocabulary, The Financial Times, 2016



# Enhancing Visualizations

- 1 Dimensional Data
  - Length
- 2 Dimensional Data
  - Position
- 2+ Dimensional Data
  - Position
  - Color Hue/Saturation
  - Size
  - Shape



Length



Position



Color



Size



# Workshop Agenda

Workshop Expectations

Understanding Data

**The Scenario**

Visualizations

Why R and ggplot2?

# Titanic Data Set

- We will use the Kaggle Competition's Titanic Machine Learning from Disaster Dataset
  - Everyone is familiar with the Titanic
  - The data set is a good representation of real world data
- Following the teaching model from Dave Langer's presentation on Data Science Dojo

# Titanic Data Dictionary

## Variables:

- Survival – Survival (yes=1, no=0)
- Pclass – Ticket Class (1<sup>st</sup> class, 2<sup>nd</sup> class)
- Sex – Gender (Male or Female)
- Age – Passenger age
- Sibsp – # of Siblings/Spouse
- Parch – # of Parents/Children
- Ticket – Ticket Number
- Fare – Passenger Fare
- Cabin – Cabin Number
- Embarked – Port of Embarkation

# Your Job

- You are hired as a consultant and have been tasked with analyzing the titanic data set.
- Your goal is to explore patterns and trends to explain what influenced the survival rate of the passengers on the Titanic.

# Workshop Agenda

Workshop Expectations

Understanding Data

Visualizations

The Scenario

**Why R and ggplot2?**

# Why R?

- Pros:

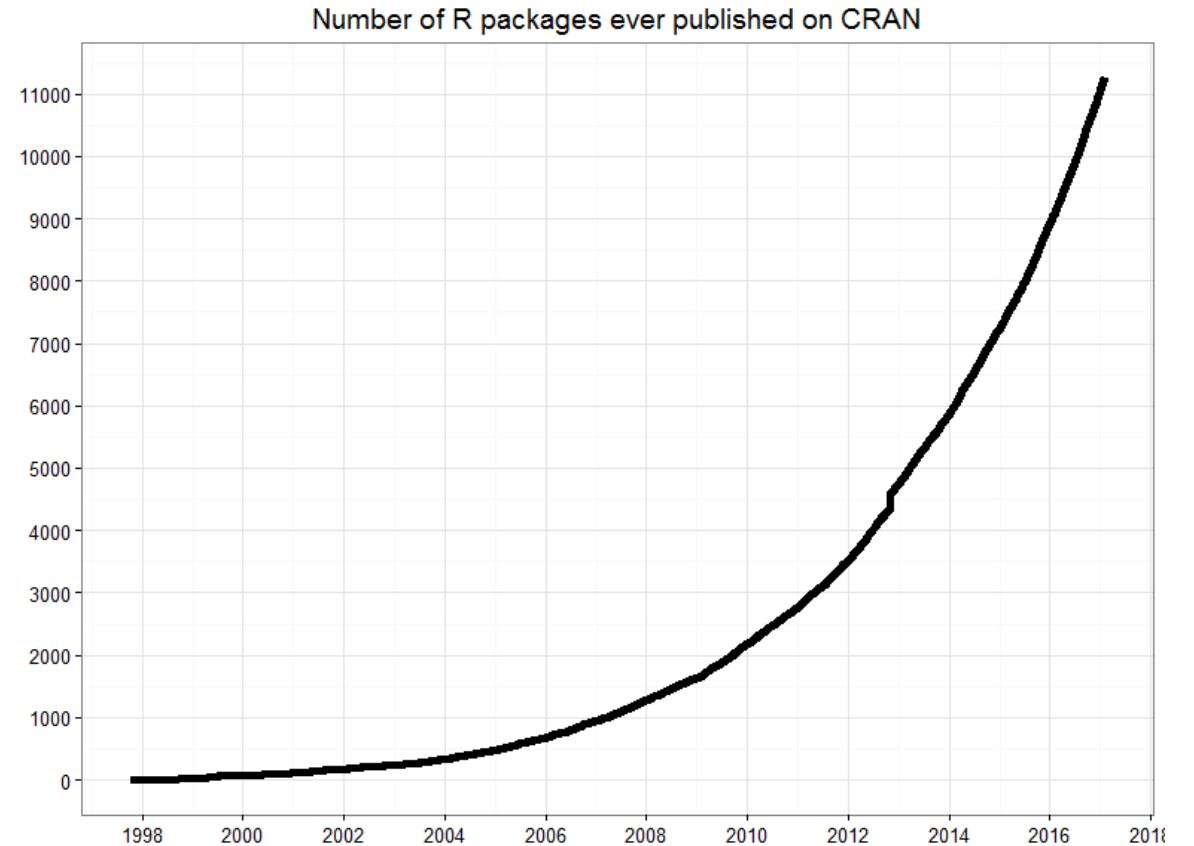
- Reproducibility
  - Explicit documentation of steps
- Versatile
  - Run on any operating system
  - Integrates with other software, languages, and data extensions
    - Python, Java, SAS, SPSS, Excel
- Free and Open-Source
- Large very active community
  - 10K packages CRAN, Twitter, GitHub, etc.
- Comprehensive
  - Eliminates the need for multiple software
    - GIS, Excel, ENVI, etc.
- Computationally Robust
  - Fast and allows for high level analysis

- Cons:

- Steep Learning Curve
  - Requires a significant time investment especially starting with little to no coding experience
- Limited “Point & Click”
  - R Commander or Radiant?
- Open-Source
  - Rely on creators to follow coding etiquette
  - Version control and instability

# Why R?

- Increasing in Popularity (especially in academia)
- User contributions significantly increased over the last decade



<http://blog.revolutionanalytics.com/2014/01/in-data-scientist-survey-r-is-the-most-used-tool-other-than-databases.html>

<http://blog.revolutionanalytics.com/2016/03/16-years-of-r-history.html>

# ggplot2 vs. base R Plots

- Ggplot2 is a package created by Hadley Wickham designed to follow the grammar of graphics.
- Although not included in Base R, it is very recognizable and widely used
- Both work well and have their respective advantages and disadvantages
  - This is beyond the scope of this workshop
  - <http://varianceexplained.org/r/why-i-use-ggplot2/>
  - <https://flowingdata.com/2016/03/22/comparing-ggplot2-and-r-base-graphics/>



# Grammar of Graphics

Originated by [Leland Wilkinson](#), simplified by [Hadley Wickham](#) and others.

Describe all the non-data ink

Plotting space for the data

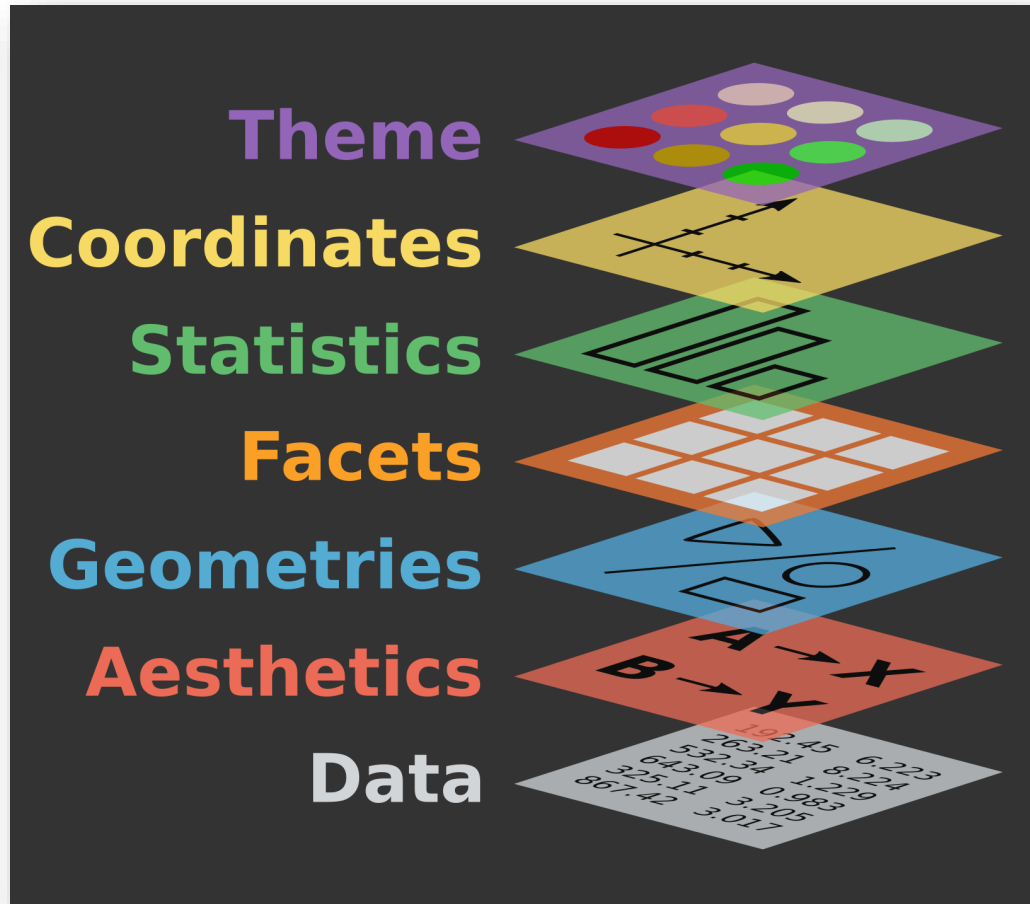
Statistical models & summaries

Rows and columns of sub-plots

Shapes used to represent the data

Scales onto which data is mapped

The actual variables to be plotted



# The Basics

- Data – The raw materials of your visualization
- Aesthetics – The mapping of your data to the visualization
  - X-axis is age
  - Y-axis is survival
- Layers – Any visualization requires at least one layer and in ggplot2 these are typically the geoms.
  - Example a barchart is `geom_bar()`

# Questions?



## Contact the Visualization Laboratory

**Email:** [AskData@uc.edu](mailto:AskData@uc.edu)

**Web:** <https://guides.libraries.uc.edu/VizLab>

**Visit:** 240 Braunstein Hall (Geology-Math-Physics Library)

### Consultation Hours

Tuesdays -10:00am-12:00pm

Thursdays: 2:00pm- 4:00pm