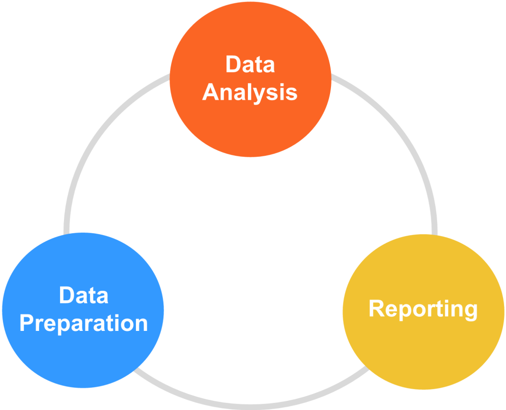
# Graphing Process

## Intro to Data Visualization

Gaston Sanchez

# Cycle of Data Analysis Projects

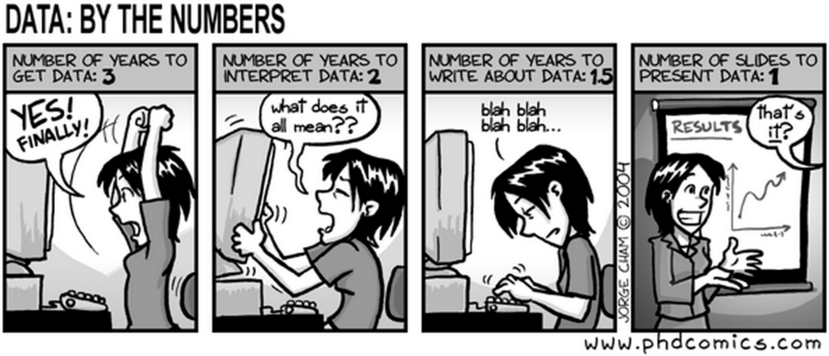
# Cycle of Data Anlaysis Projects



# Cycle of DAP and data visualizations

- Data visualization may be present at any stage of a data analysis project (DAP).
- The data preparation stage is usually supported with quick check-up displays.
- Exploratory displays constantly appear in the actual data analysis part.
- Communication displays are central to the reporting stage.

# Understanding the Data Analysis Process


DATA: BY THE NUMBERS
NUMBER OF YEARS TO GET DATA: 3
YES! FINALLY!
NUMBER OF YEARS TO INTERPRET DATA: 2
what does it all mean??
NUMBER OF YEARS TO WRITE ABOUT DATA: 1.5
blah blah blah blah...
JORGE CHAM © 2004
NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?
www.phdcomics.com

http://www.phdcomics.com/comics/archive.php/archive/tellafriend.php?comicid=462

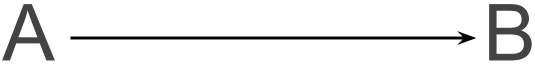
# Data Preparation

# Core Data Analysis

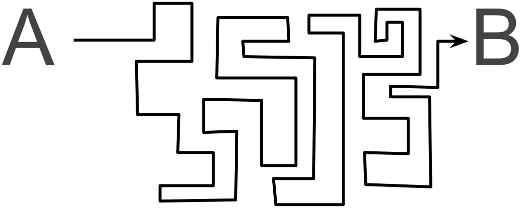
# Reporting

# Communication

NUMBER OF SLIDES TO PRESENT DATA: 1
RESULTS
that's it?



ideal linear data analysis process



data analysis process in practice

# Data Visualization Process

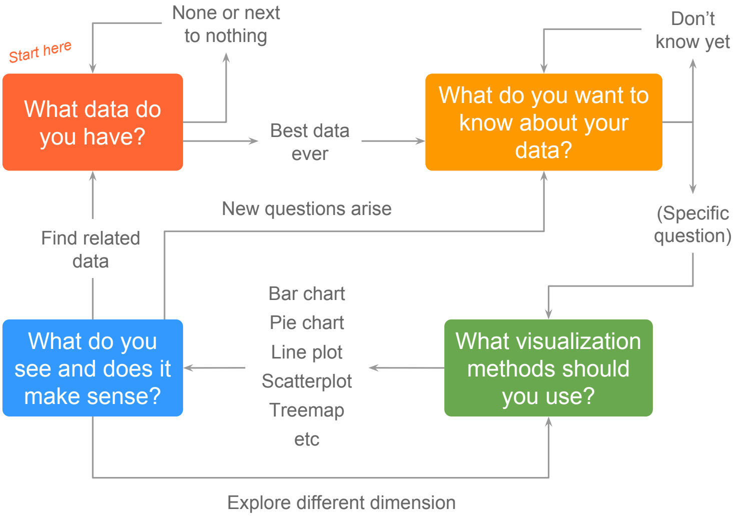
# Datavis Process

## Main Considerations

The plotting steps vary by dataset and project. But you should consider:

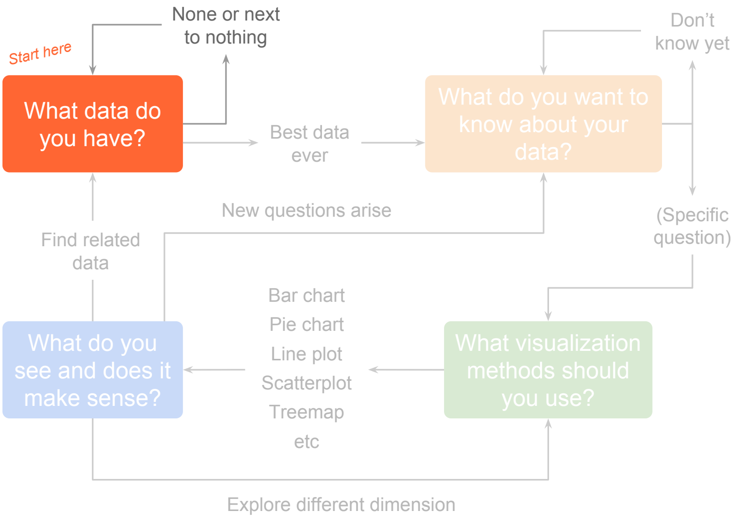
1. What data do you have?
2. What do you want to know about the data?
3. What visualization methods should you use?
4. What do you see and does it make sense?

# Exploration Process



(based on Nathan Yau)

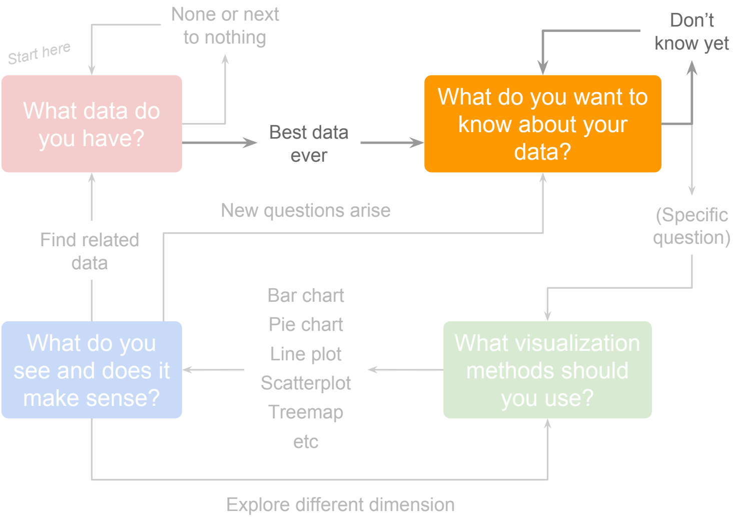
# Exploration Process



# What data you have?

- Garbage in ⇒ Garbage out.
- Good visualizations depend on good data.
- Data: can be small, medium, large, big.
- The important thing is its quality.

# Exploration Process



Start here

What data do you have?

None or next to nothing

Best data ever

What do you want to know about your data?

Don't know yet

(Specific question)

What visualization methods should you use?

Bar chart

Pie chart

Line plot

Scatterplot

Treemap

etc

What do you see and does it make sense?

Find related data

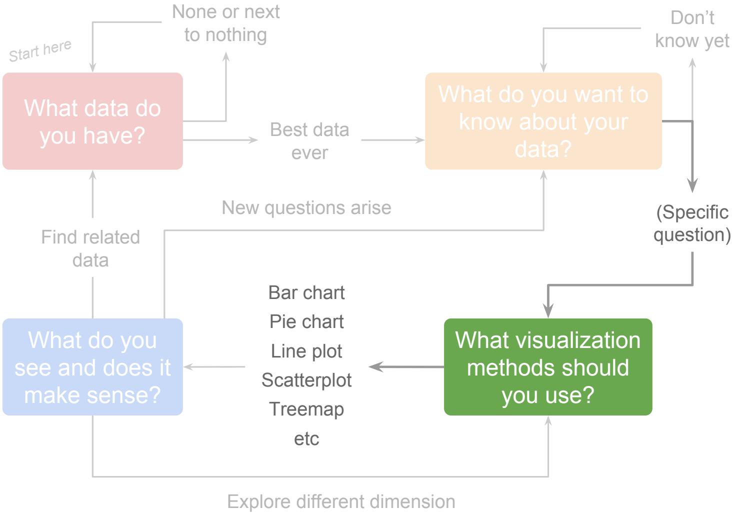
New questions arise

Explore different dimension

# What do you want to know?

- What is the research question?
- Focus on one or two questions that you want to answer.
- Talk with the experts in the field of application.
- At the beginning of a project, there will be more exploratory tasks.
- As a project evolves, the more refined and targeted the visualizations.

# Exploration Process



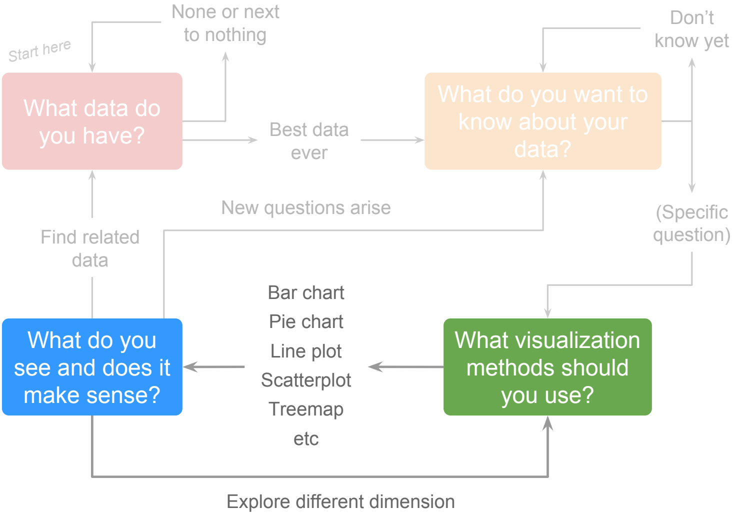
# What visualization to use?

- How many variables?
  - One variable
  - Two variables
  - Three or more
- What type of variables?
- Quantitative, qualitative, time

# What visualization to use?

- Distributions
- Parts of a whole
- Trends in time
- Maps and geographic displays
- Bivariate Relationships
- Networks and other associations
- Facetting

# Exploration Process

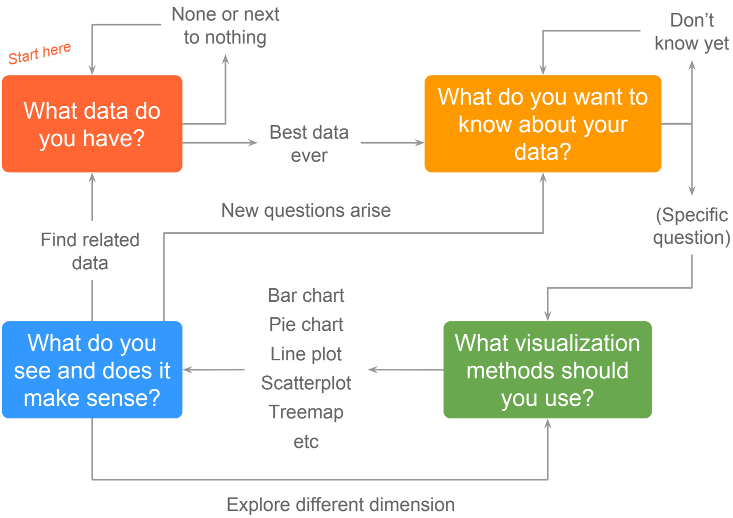


# What do you see?

Things to pay attention to:

- Systematic variation
- Increasing patterns
- Decreasing patterns
- Atypical values (outliers)
- Similarities and differences

# Exploration Process



# Visualization Process

- Iterative Process
- Back and Forth
- Trial and Error
- Target questions

# Things to keep in mind

# It's (almost) all about the questions

- Great visualization never start from the standpoint of the data set.
- Great visualization starts with questions.
- Why was the data collected?
- What is interesting about the data set?
- What stories can it tell?

# It's (almost) all about the questions

- An fundamental skill in understanding data is asking good questions.
- Identify the question that you want to answer.
- Think about how it will be used and work backward to what was collected.
- The more specific the question, the better visualizations tend to be.

# Data Visualization Purposes

# Stats Graphics

Graphics for

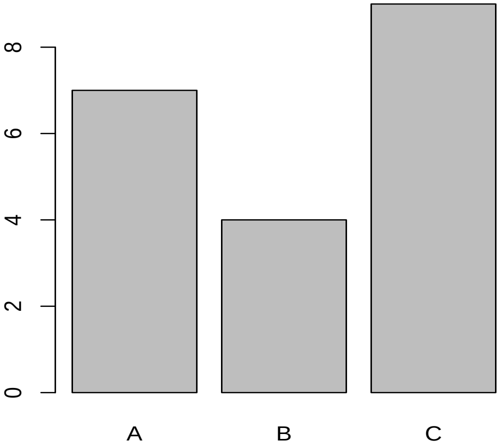
Exploration & Communication

# Analysis and Exploration

## Graphics for Exploration

- Graphics for understanding data.
- The analyst is the main (and usually only) consumer.
- Typically quick & dirty.
- Not much care about visual appearance and design principles.
- Lifespan of a few seconds.
- Basic plots (deault parameters) in R are designed for this purpose.

# Graphics for Exploration Example

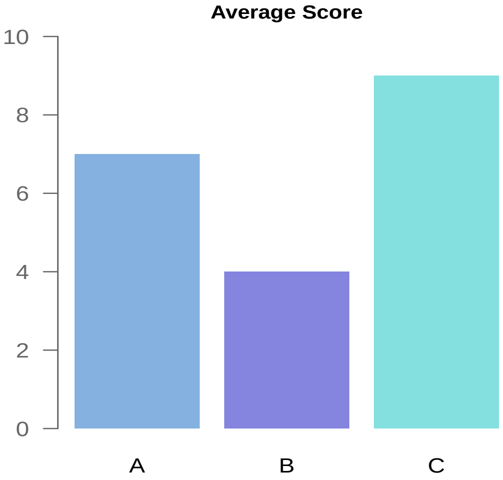


# Communication and Presentation

## Graphics for Communication

- Graphics for presenting data
- To be consumed by others
- Must care about visual appearance and design
- Require a lot of iterations in order to get the final version
- What's the message?
- Who's the audience?
- On what type of media / format?
- Very time consuming!

# Graphics for Communication



# Graphics for Communication

Use visualization to communicate ideas, influence, explain persuade

Visuals can serve as evidence or support

# Graphical output in different formats

# Plotting options

When creating a plot in R ...

Screen display OR Save in File

# File Acronyms

**File Acronyms**

| Acronym | Description |
|---|---|
| PDF | Portable Document Format |
| SVG | Scalable Vector Graphics |
| PNG | Portable Network Graphics |
| JPEG | Joint Photographic Experts Group |
| BMP | Bitmap |
| TIFF | Tagged Image File Format |

# Output Formats

## Graphics devices from the output format

Vector -vs- Raster

# Output Formats

## Vector Formats

An image is described by a set of mathematical shapes (e.g. PDF, PostScript, SVG)

## Raster Formats

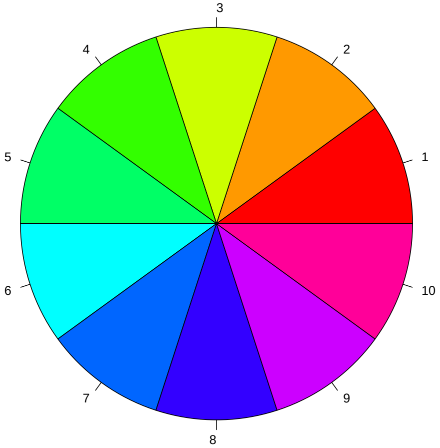
An image consists of an array of pixels, with information such as color recorded for each pixel (e.g. PNG, JPEG, TIFF, all screen devices)

# Vector or Raster?

## Vector Formats

Vector formats are superior for images that need to be viewed at a variety of scales (i.e. zoom in and out).

# Example: vector image (pdf)

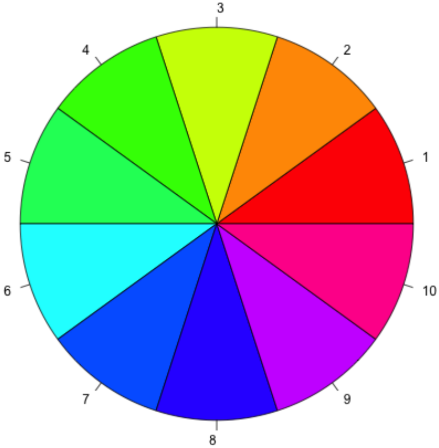


# Vector or Raster?

## Raster Formats

Raster formats tend to be preferred when a plot is visually complex (e.g. many data points), and it will produce smaller files if the image is very complex.

# Example: raster image (png)



# Vector or Raster?

If further modifications to an R plot will be made using third-party software:

- removing a particular form are only possibe with vector format
- modifying pixels of a particular color are only possible with raster formats

Keep in mind: It is easy to convert a vector format to a raster format, while the reverse is almost impossible

# Vector Formats

## PDF

- Good choice of static format
- Resizes well, usually portable
- Less efficient if a plot has many objects/points

# Vector Formats

## SVG

- XML-based format
- Good choice for web pages
- `svg()` available in Linux and Mac
- SVG output in Windows requires package `"Cairo"`
- Some advanced SVG features are limitted in R

# Raster (Bitmap) Formats

## PNG

- Desirable format for simple images (most statistical graphics)
- Good for line drawings or images with solid colors
- Good for many, many objects, points=
- PNG uses **lossless** compression: compresses the image without losing information
- PNG does not resize well
- Consequently, PNG files can be edited without reducing quality
- Most web browsers can read this format natively

# Raster Formats

## JPEG

- Good for photographs or natural scenes
- JPEG uses **lossy** compression: compresses the image with some information loss
- Consequently, repeatedly editing a JPEG filewill result in quality reduction
- JPEG does not resize well
- Better suited for complex images with lots of different regions (like photographs)

# Raster Formats

## TIFF

- Sophisticated format that allows multiple pages of raster output within a single file
- Supports lossless compression
- Less supported by web browsers
- Preferred format for publishers of books or journal articles

# Raster Formats

## Image Size

- Size of Raster images is specified in number of pixels (rahter than physical size in inches)
- The physical size of a raster image is determined by the **resolution** at which it is viewed
- e.g. PNG image 72 pixels wide will be 1 inch wide when viewed on a screen with a resolution of 72 dpi (dots per inch)
- e.g. PNG image 72 pixels wide will be 0.75 inches wide on a screen with a resolution of 96 dpi

# Considerations

Plots on Screen

-vs-

Plots on Print

# David Smith's Recommendations

- Use pdf for printing
- Use png for web displays
- For documents or for detail, go hi-resolution
- Choose your dimensions carefully
- Think about aspect ratio
- Vector formats are good for line drawings and plots with solid colors
- Remove the outer margins, if you're not using them
- Make sure anti-aliasing is enabled
- Avoid using JPEG
- Be creative

http:

//blog.revolutionanalytics.com/2009/01/10-tips-for-making-your-r-graphics-look-their-best.html

# PDF

## Use pdf for printing

- Use pdf if you plan to print your graphic
- The graphic is scale-independent
- PDF viewers are ubiquitous these days
- Easy to create a high-quality printout of a PDF file on almost any printer
- Best choice whenever you want to send the graph as a file via email, and the recipient needs the best quality possible

# PNG

## For Web display, use PNG

- These days, the best choice is the PNG format
- Most browsers can display PNG graphics without trouble
- The main choice you need to make when using `png()` is the dimensions of the graphic in pixels
- Slides 4x3 png plots: `width=1024` and `height=768` pixels
- Slides 16x9 png plots: `width=1920` and `height=1080` pixels

# Dimensions

## Choosing dimensions

- For PDF graphs this is easiest to deal with, where you specify width and height in inches anyway
- For raster images is a bit trickier:
- R assumes 72 pixels to the inch
- When you increase the pixel dimensions you're also increasing the implicit size of the graph area

# Summary

- Plots are created on a graphics device
- There are screen devices and file devices
- Default graphics on screen are good for exploratory analysis
- File devices are useful for presentation-consumption of graphics
- File devices are divided in *Vector* and *Raster* formats
- Vector formats are good for line drawings and plots with solid colors
- Bitmap formats are good for plots with a large number of points