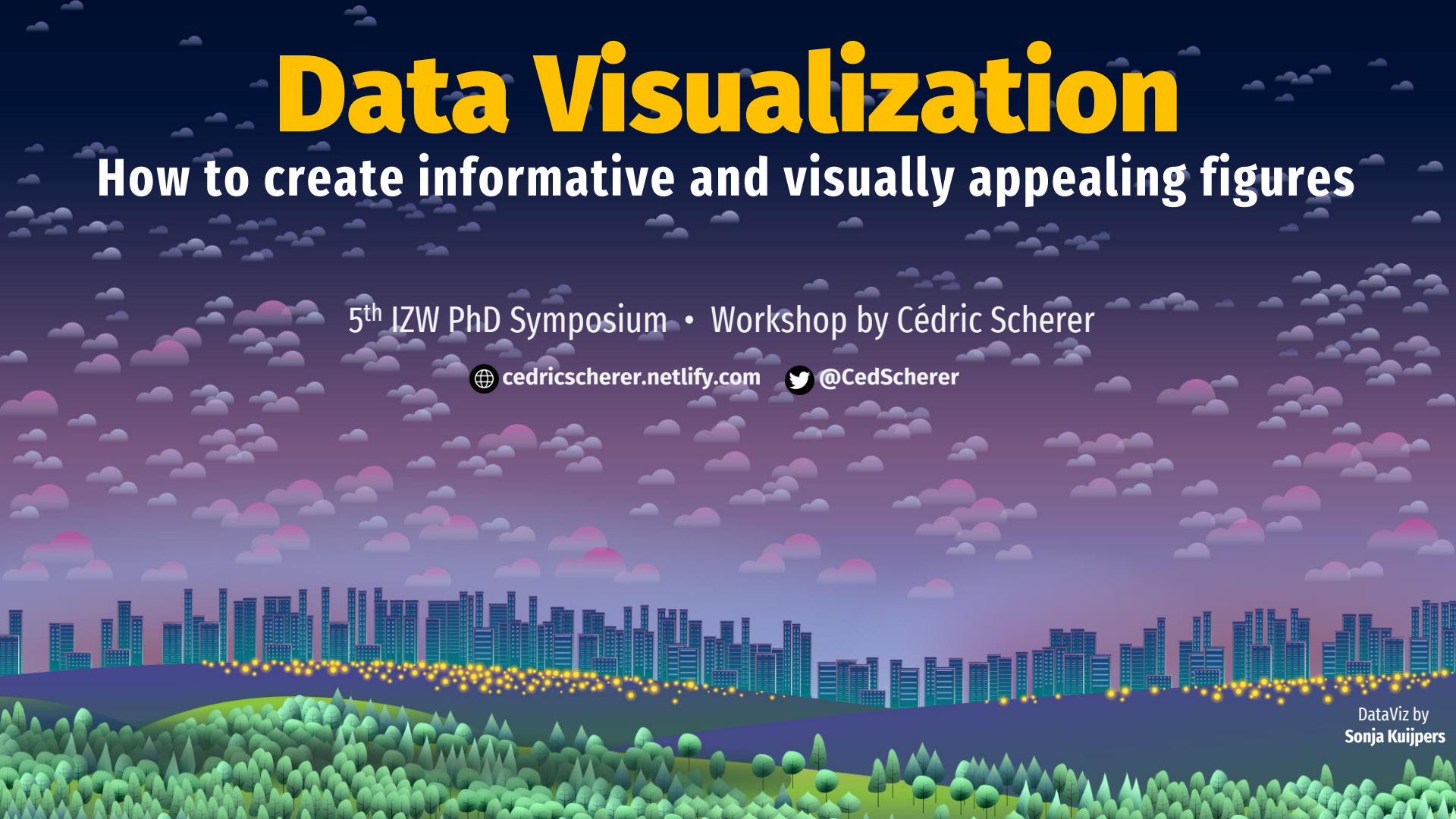
Data Visualization

How to create informative and visually appealing figures

5th IZW PhD Symposium • Workshop by Cédric Scherer

cedricscherer.netlify.com @CedScherer

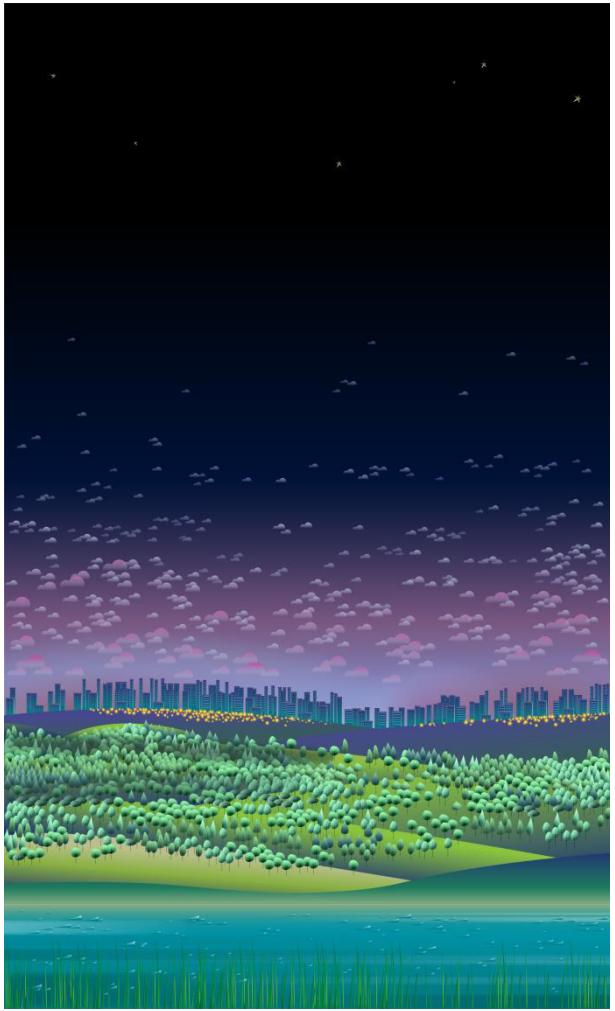


DataViz by
Sonja Kuijpers

You might be wondering what you are viewing here.

You might be wondering what you are viewing here.

Each element represents a person who committed suicide in the Netherlands in the year 2017.



Each category/method of suicide is represented by a certain element:

hanging (strangulation)

taking drugs/alcohol/medicines

in front of train or metro

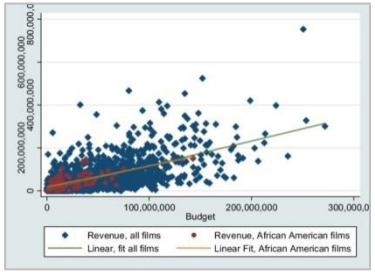
drowning

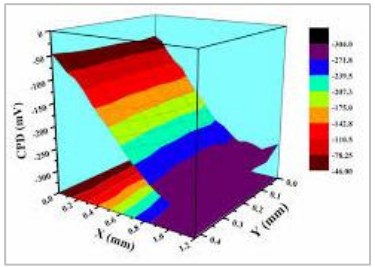
jumping from height

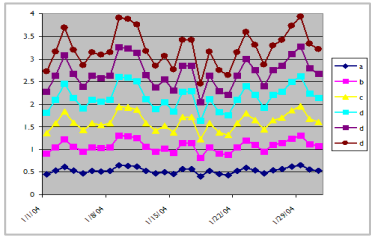
other method*

unknown method

Data visualization is any graphical representation of information and data

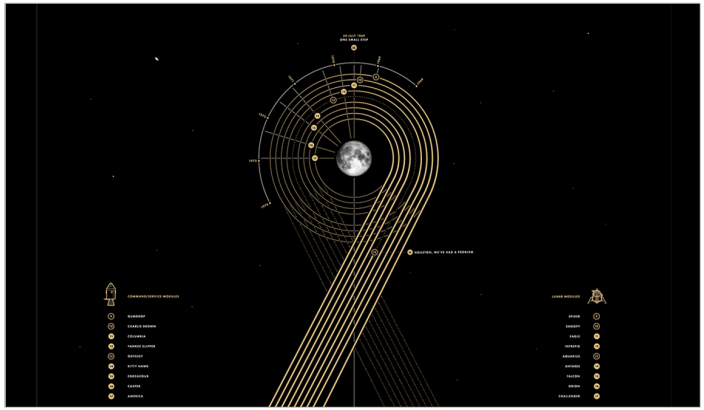


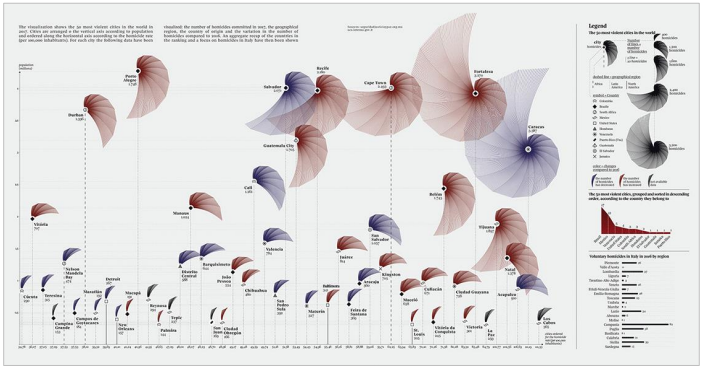




Anonymous

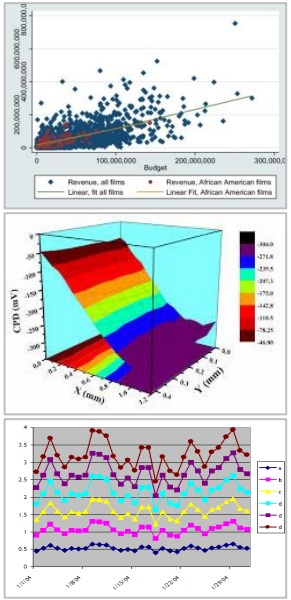
Sonia Kuijpers





Upper: Paul Button
Lower: Frederica Fragapane

Data visualization is any graphical representation of information and data



Anonymous

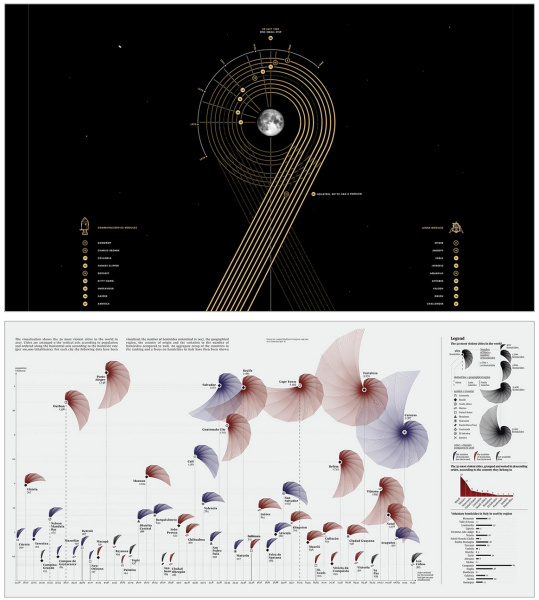
We aim for DataViz that:

- is informative
- reduces complexity
- is easy to grasp
- is visually appealing
- draws attention

but:

- is not too abstract
- is not too “unusual”

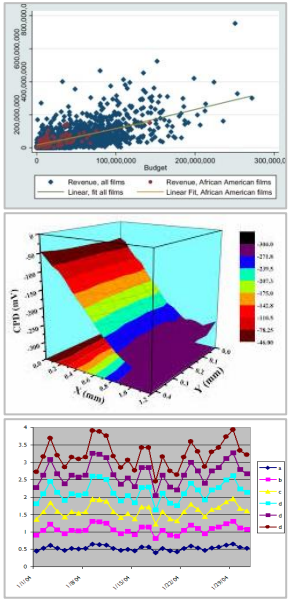
Sonia Kuijpers



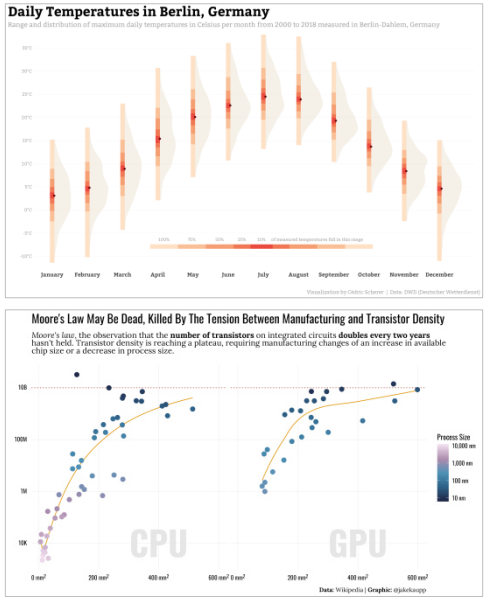
Upper: Paul Button
Lower: Frederica Fragapane

Gradient from poorly designed & uninformative data visualization to data art

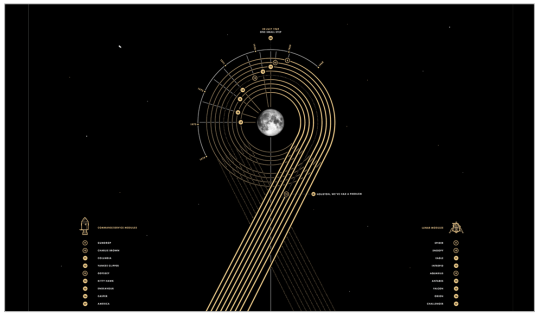
Data visualization is any graphical representation of information and data

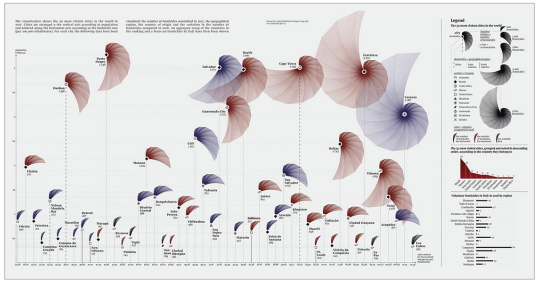


Anonymous



Upper: Cédric Scherer
Lower: Jake Kaupp





Sonia Kuijpers

Upper: Paul Button
Lower: Frederica Fragapane

Gradient from poorly designed & uninformative data visualization to data art

Know Your
Data Types

Data Types

- **Quantitative (numerical)** versus **qualitative (categorical)** data
- **Continuous** versus **discrete** data
- **Ordered** versus **unordered** data

Data Types

- **Quantitative (numerical)** versus **qualitative (categorical)** data
- **Continuous** versus **discrete** data
- **Ordered** versus **unordered** data

- "female" ⟶ qualitative + discrete + unordered
- 2019/09/26 "17:01:35" ⟶ quantitative + continuous + ordered
- 1 ⟶ quantitative + continuous + ordered

... or: quantitative + discrete + ordered
... or: qualitative + discrete + ordered
... or: qualitative + discrete + unordered

Data Types

- Quantitative (numerical):

  continuous or discrete, ordered variables

  e.g. temperatures or concentrations (continuous),
  age groups or species number (discrete)

- Qualitative (categorical):

  discrete factors with different ordered or unordered levels (categories)

  e.g. qualities or seasons (ordered),
  species or sex (unordered)

Data Types

- Quantitative (numerical):

  continuous or discrete, ordered variables

  e.g. temperatures or concentrations (continuous),
  age groups or species number (discrete)

- Qualitative (categorical):

  discrete factors with different ordered or unordered levels (categories)

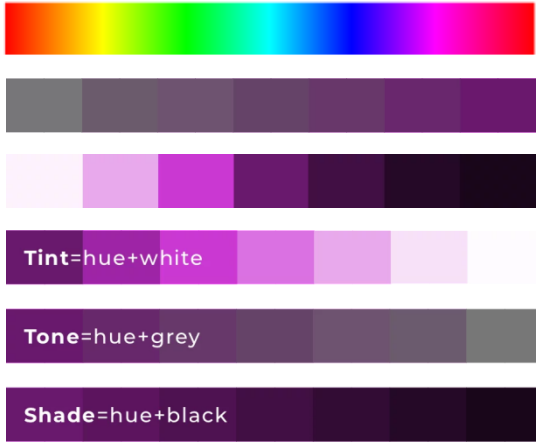
  e.g. qualities or seasons (ordered),
  species or sex (unordered)

- Plus special cases:
  dates (discrete ordered) and **text** (discrete unordered)

Choice of the Color Palette

Color Terminology

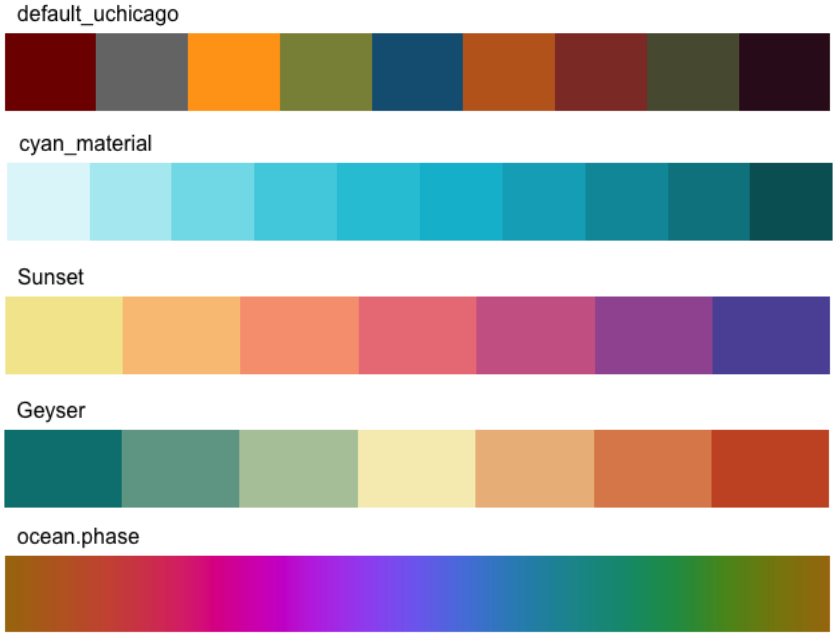
- **Hue**: color, like blue or red
- **Chroma**: how pure a color is (saturation)
- **Value**: how light or dark a color is
- **Tint**: created by adding white to a hue
- **Tone**: created by adding grey to a hue
- **Shade**: created by adding black to a hue



visualcomposer.com/blog/ultimate-guide-to-colors-and-color-palettes

Color Palette Types

- Four main types of color palettes:
  - **Categorical**
  - **Sequential:** Single-Hue
  - **Sequential:** Multi-Hue
  - **Diverging**
  - **Cyclical**



Choice of the Color Palette

Journals & Magazines > IEEE Computer Graphics and Ap... > Volume: 27 Issue: 2

Rainbow Color Map (Still) Considered Harmful

Publisher: IEEE

2 Author(s) David Borland ; Russell M. Taylor Ii **View All Authors**

172	3	9091
Paper Citations	Patent Citations	Full Text Views

Medical Physics

Current Issue | Authors | Submissions | Advertise | Search

Med Phys. 2015 Jun; 42(6): 2942–2954. PMCID: PMC5148121
Published online 2015 May 20. doi: 10.1118/1.4921125 PMID: 26127048

Effect of color visualization and display hardware on the visual assessment of pseudocolor medical images

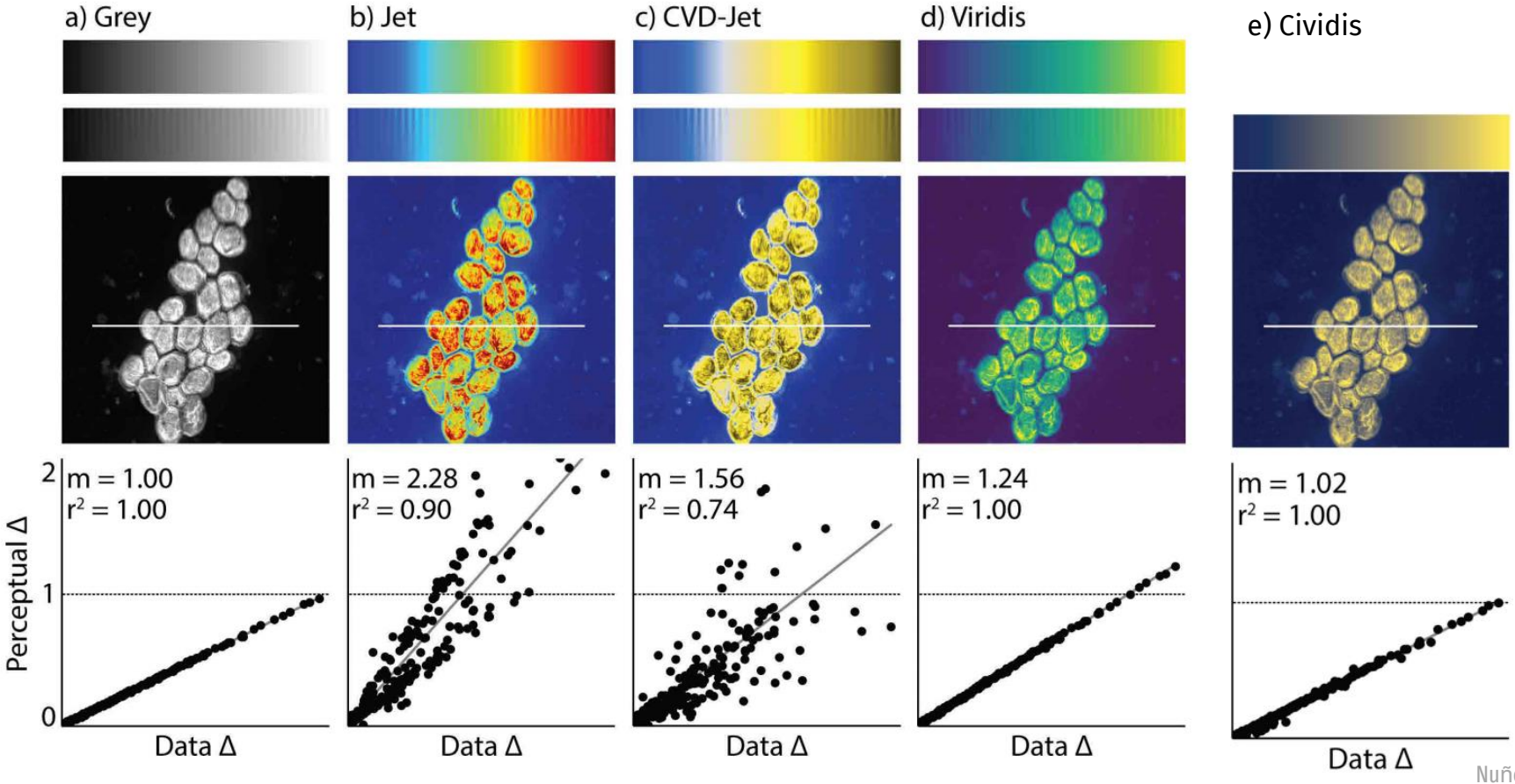
Silvina Zabala-Travers, Mina Choi, Wei-Chung Cheng, and Aldo Badano[a)]

10 March 2017

Interpretation of the rainbow color scale for quantitative medical imaging: perceptually linear color calibration (CSDF) versus DICOM GSDF

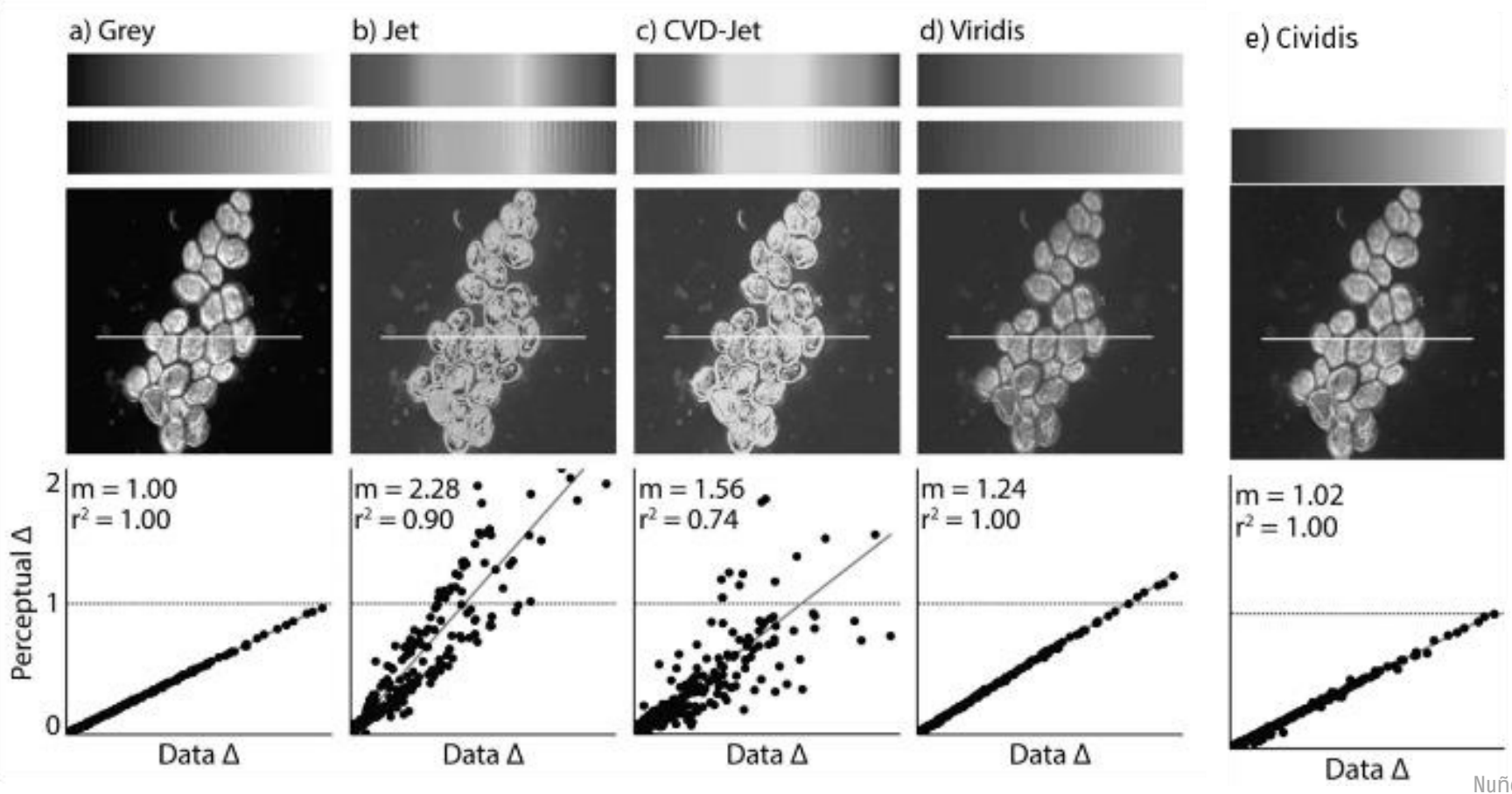
Frédérique Chesterman; Hannah Manssens; Céline Morel; Guillaume Serrell; Bastian Piepers; Tom Kimpe

Choice of the Color Palette



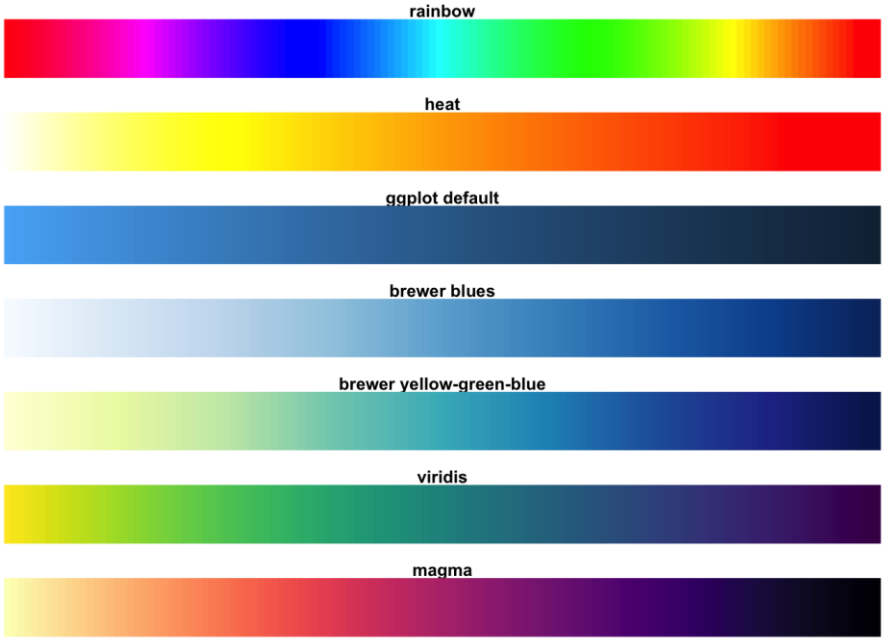
Nuñez, Anderton & Renslow (2018) PLOSone

Choice of the Color Palette



Nuñez, Anderton & Renslow (2018) PLOSone

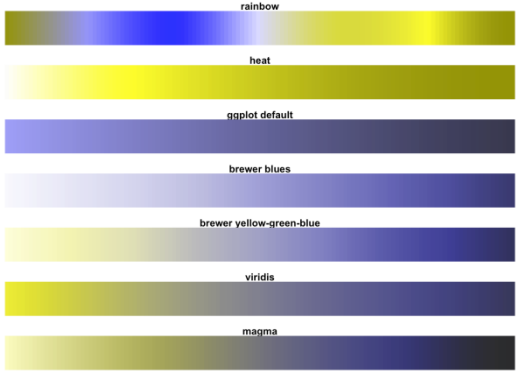
Choice of the Color Palette



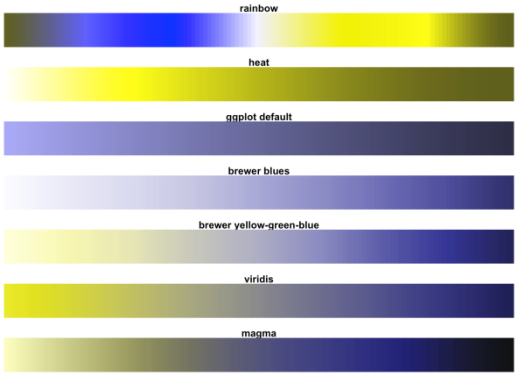
https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

Color Blind Awareness

Deutranopia: present in 6% of males



Protanopia: present in 1% of males

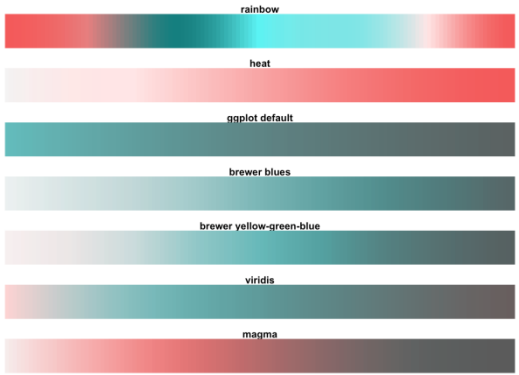


Make sure your plots are readable for color blind people and when printed in grey scale!

e.g. color-blindness.com/coblis-color-blindness-simulator

projects.susielu.com/viz-palette

Tritanopia: present in 0.008% of humans

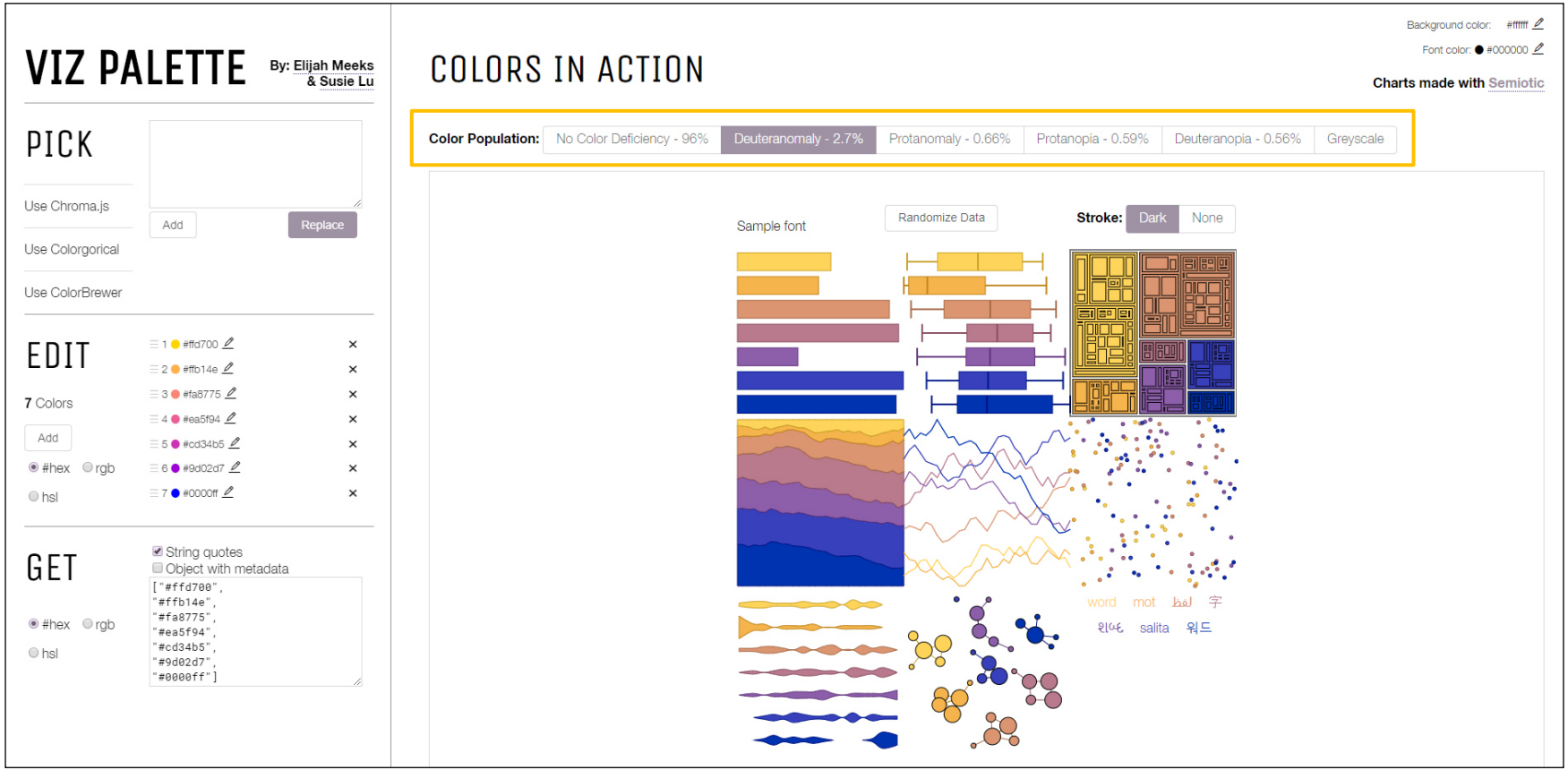


Monochromacy: present in 0.001% of humans

https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

... and present in ~75% of university printers! ☺

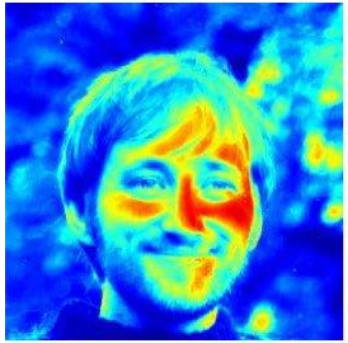
Color Blind Awareness



projects.susielu.com/viz-palette

Choice of the Color Palette

true-colour Phil



rainbow Phil

is distorted

batlow Phil

is flawless

www.fabiocrameri.ch/batlow

fabiocrameri.ch/batlow

Choice of the Color Palette

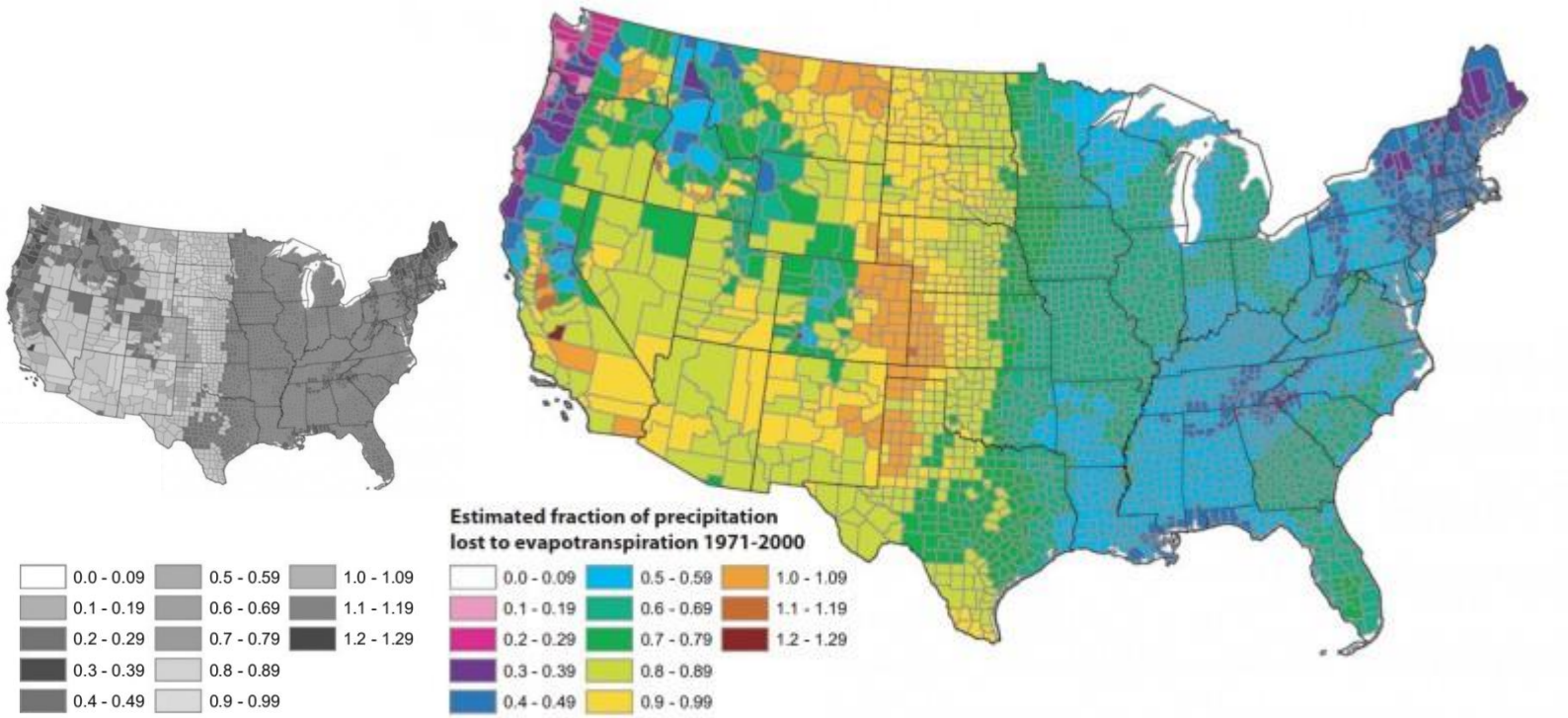


FIGURE 13. Estimated Mean Annual Ratio of Actual Evapotranspiration (ET) to Precipitation (P) for the Conterminous U.S. for the Period 1971-2000. Estimates are based on the regression equation in Table 1 that includes land cover. Calculations of ET/P were made first at the 800-m resolution of the PRISM climate data. The mean values for the counties (shown) were then calculated by averaging the 800-m values within each county. Areas with fractions >1 are agricultural counties that either import surface water or mine deep groundwater.

eagereyes.org/basics/rainbow-color-map

Choice of the Chart Type

Chart Types

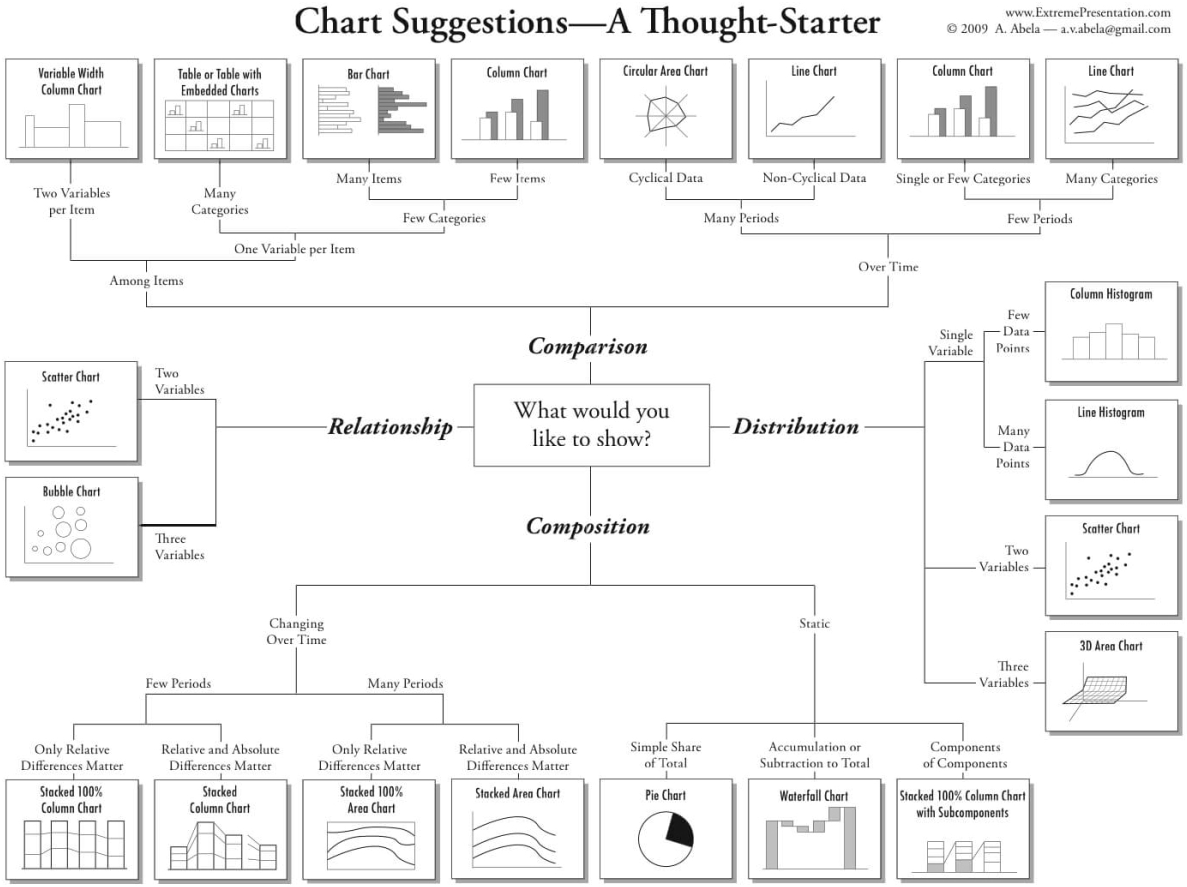


Chart Suggestions—A Thought-Starter
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Column Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Few Categories
Cyclical Data
Non-Cyclical Data
Many Periods
Single or Few Categories
Many Categories
Few Periods
One Variable per Item
Among Items
Over Time
Comparison
Scatter Chart
Two Variables
Relationship
What would you like to show?
Distribution
Bubble Chart
Three Variables
Composition
Single Variable
Few Data Points
Column Histogram
Many Data Points
Line Histogram
Two Variables
Scatter Chart
Three Variables
3D Area Chart
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart
Stacked 100% Column Chart with Subcomponents

Chart Types

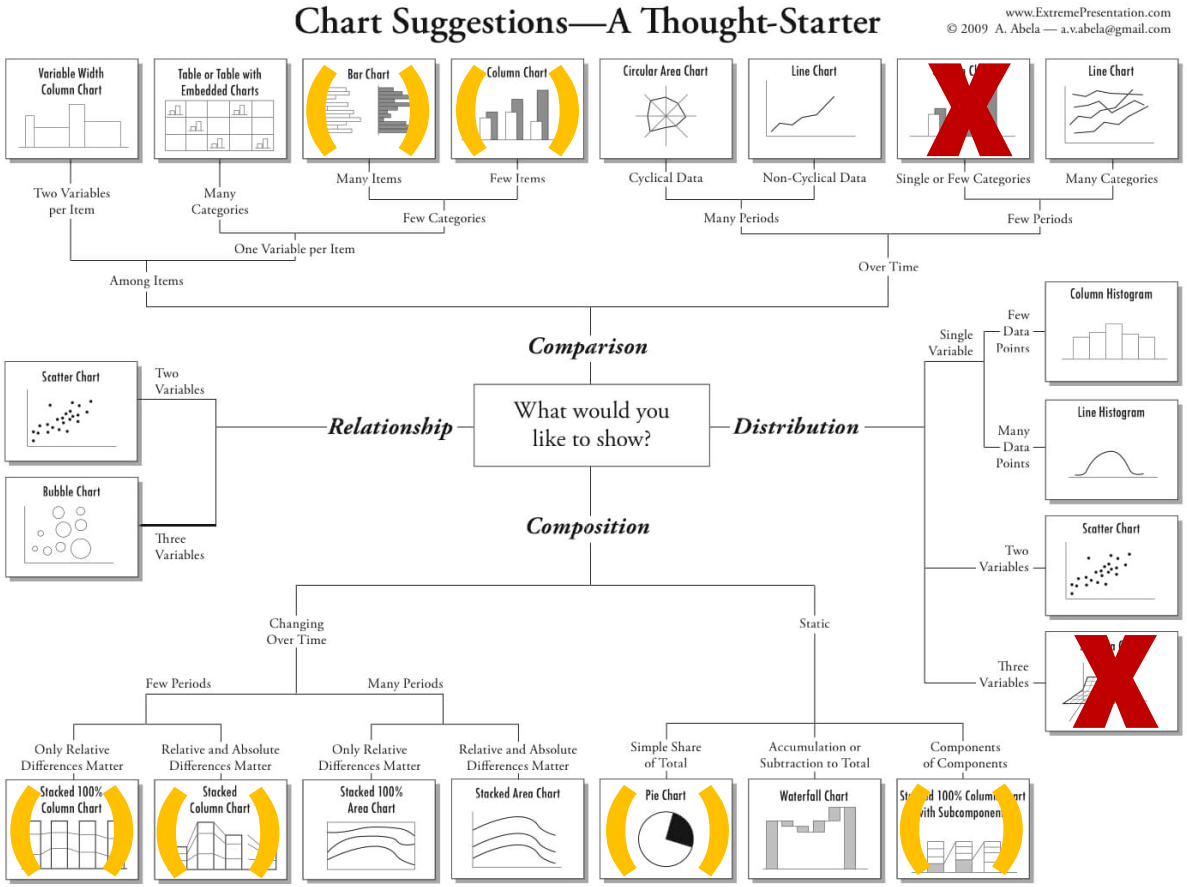
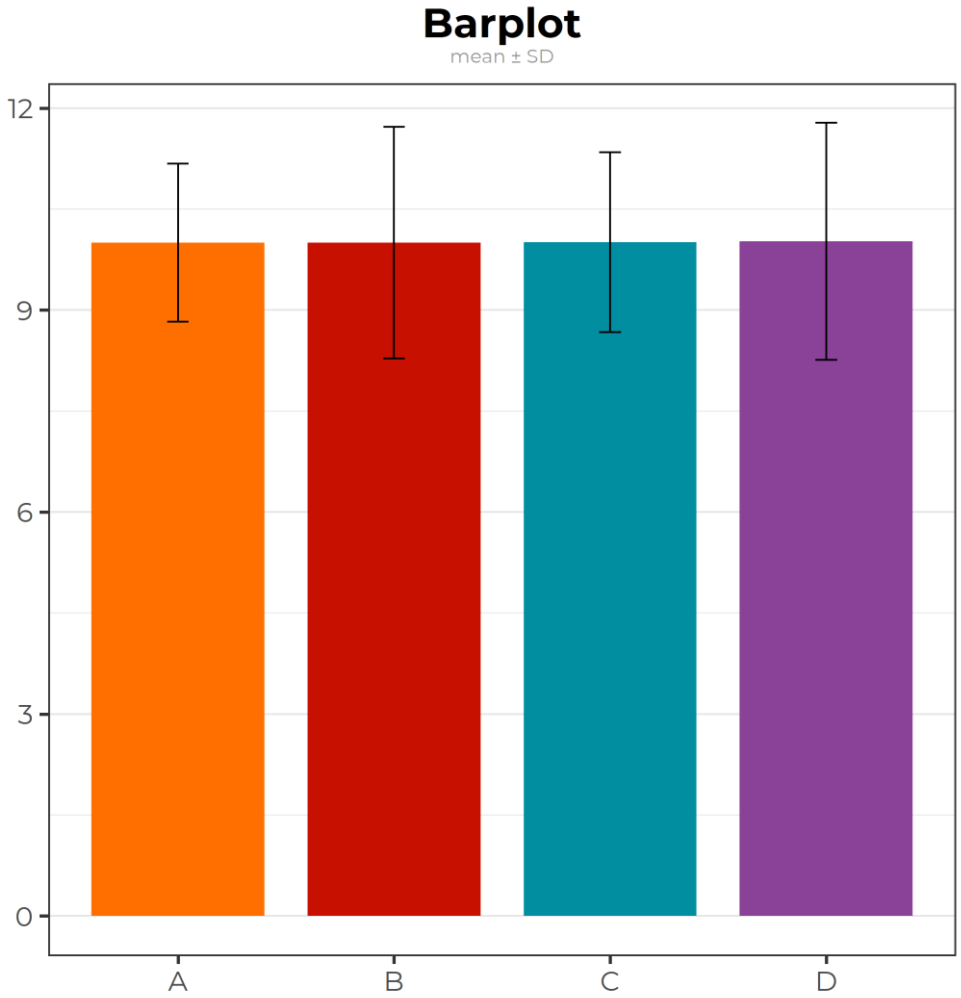
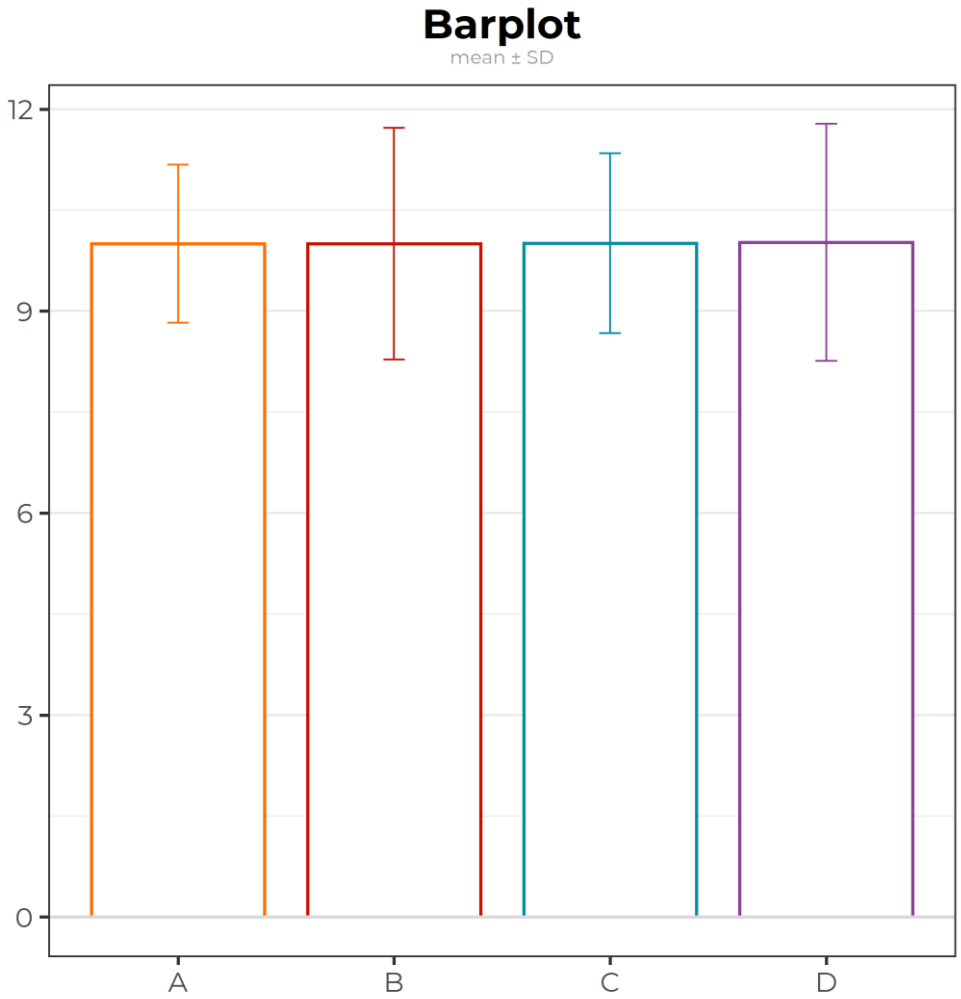


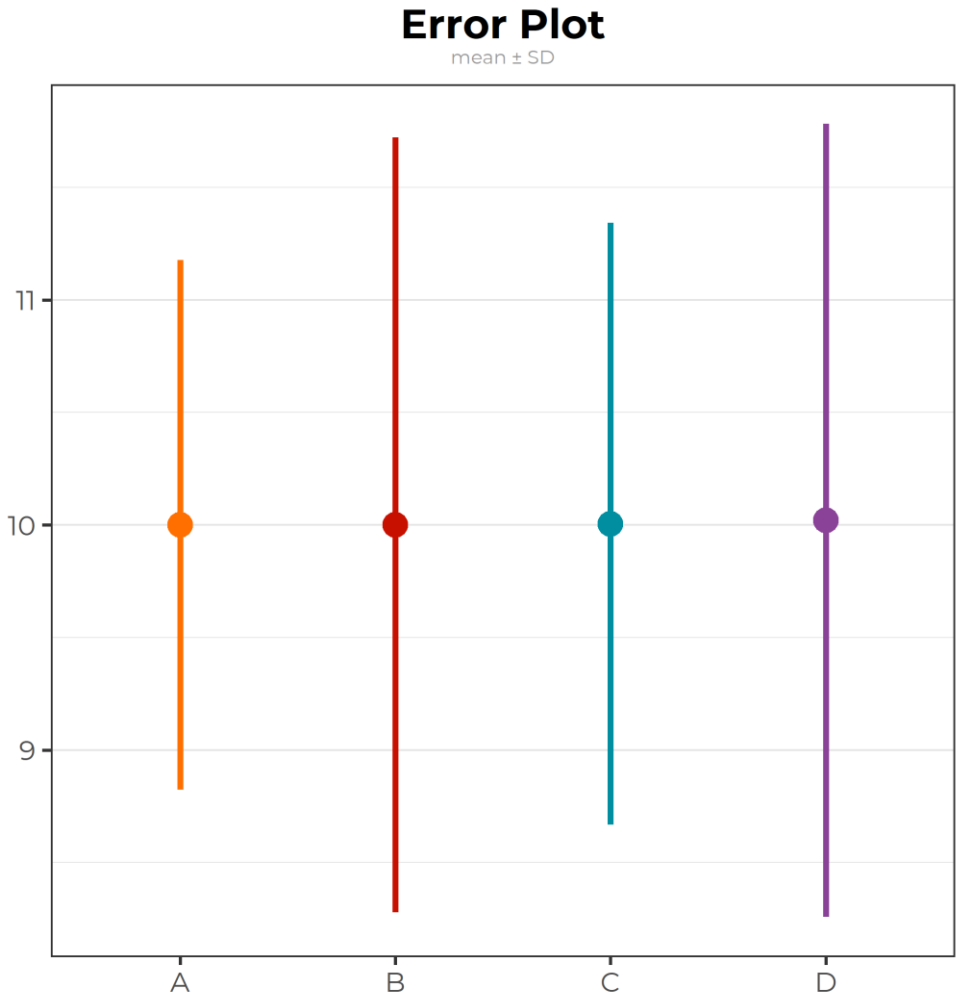
Chart Suggestions—A Thought-Starter
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com
Variable Width Column Chart
Table or Table with Embedded Charts
Bar Chart
Column Chart
Circular Area Chart
Line Chart
Line Chart
Two Variables per Item
Many Categories
Many Items
Few Items
Few Categories
Cyclical Data
Non-Cyclical Data
Many Periods
Single or Few Categories
Many Categories
Few Periods
One Variable per Item
Among Items
Over Time
Comparison
Scatter Chart
Two Variables
Bubble Chart
Three Variables
Relationship
What would you like to show?
Distribution
Composition
Single Variable
Few Data Points
Column Histogram
Many Data Points
Line Histogram
Two Variables
Scatter Chart
Three Variables
Changing Over Time
Static
Few Periods
Many Periods
Only Relative Differences Matter
Relative and Absolute Differences Matter
Only Relative Differences Matter
Relative and Absolute Differences Matter
Simple Share of Total
Accumulation or Subtraction to Total
Components of Components
Stacked 100% Column Chart
Stacked Column Chart
Stacked 100% Area Chart
Stacked Area Chart
Pie Chart
Waterfall Chart



- Boring
- Not accurate
  (only in case of amounts or proportion)
- Lots of useless space
  (but never omit zero!)



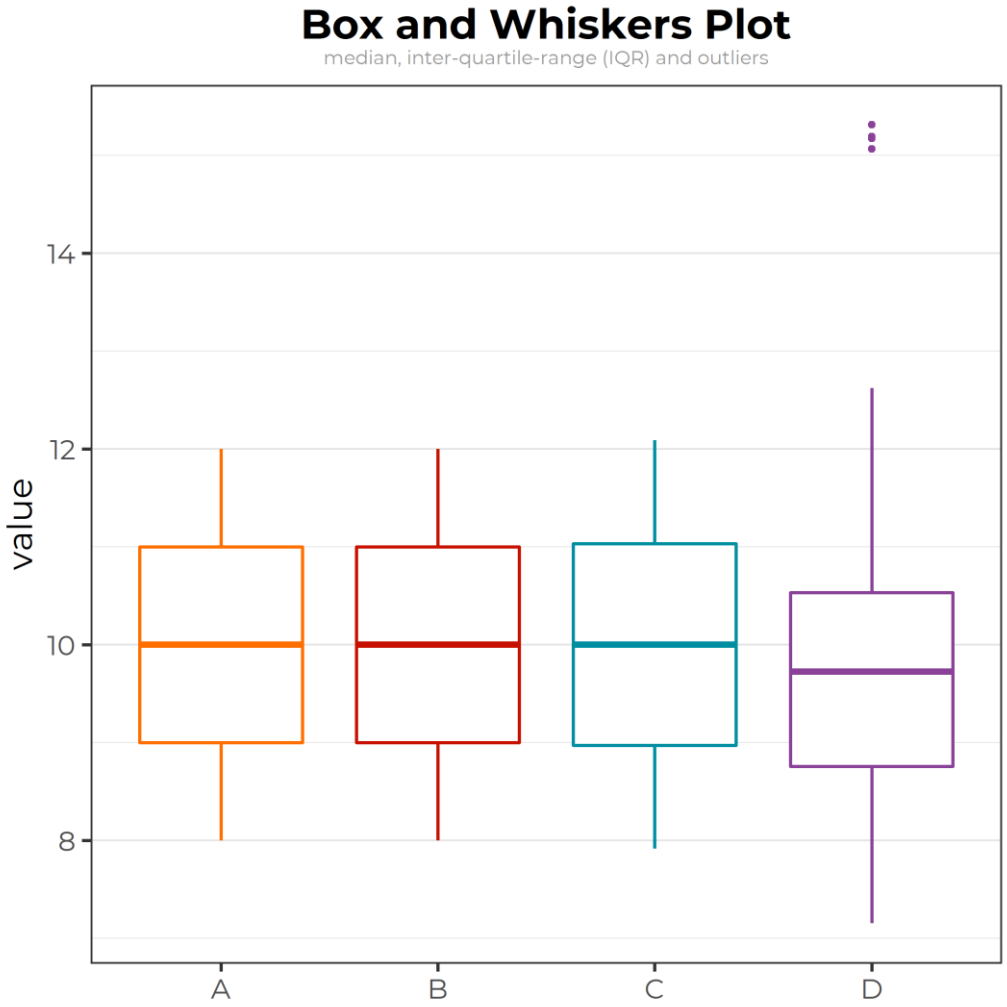
- Boring
- Not accurate
  (only in case of amounts or proportion)
- Lots of useless space
  (but never omit zero!)



- More accurate than bar plot
- Better data-ink ratio

But:

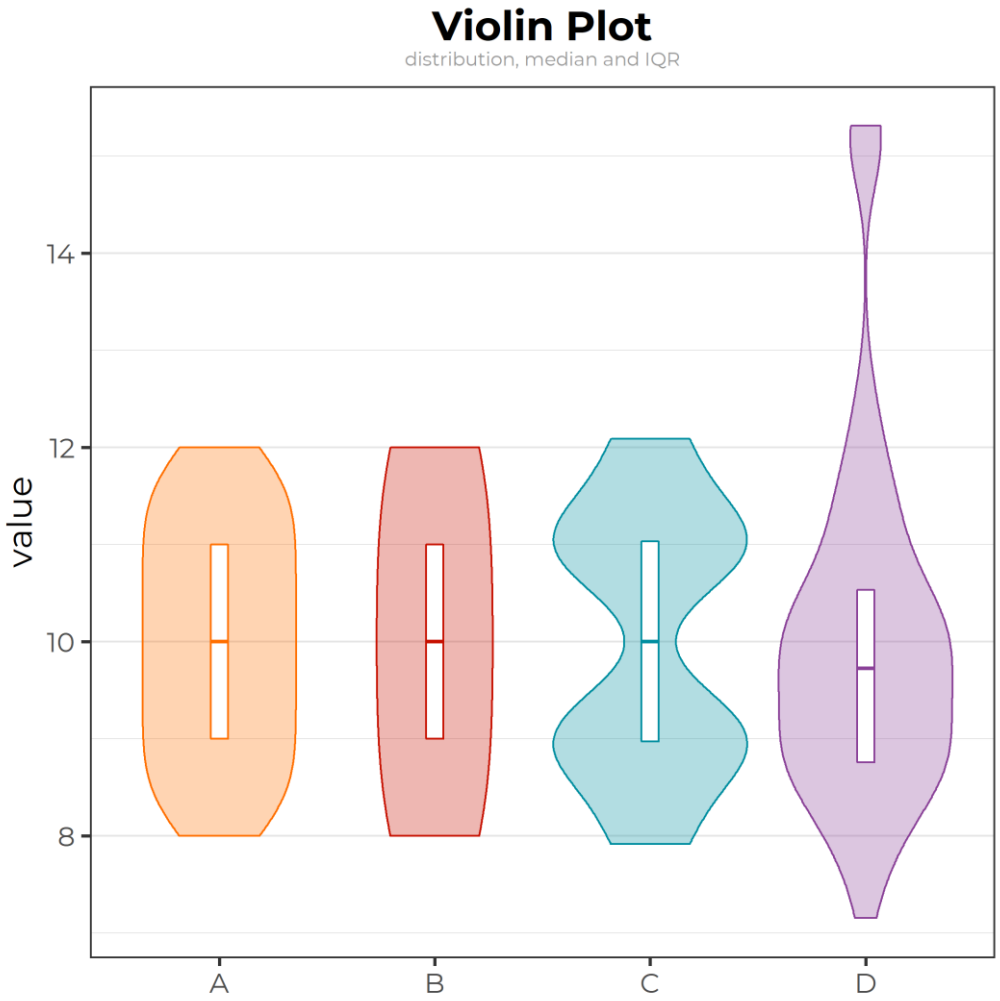
- unclear what's shown



- Shows important summary stats
- Looks “scientifically”

But:

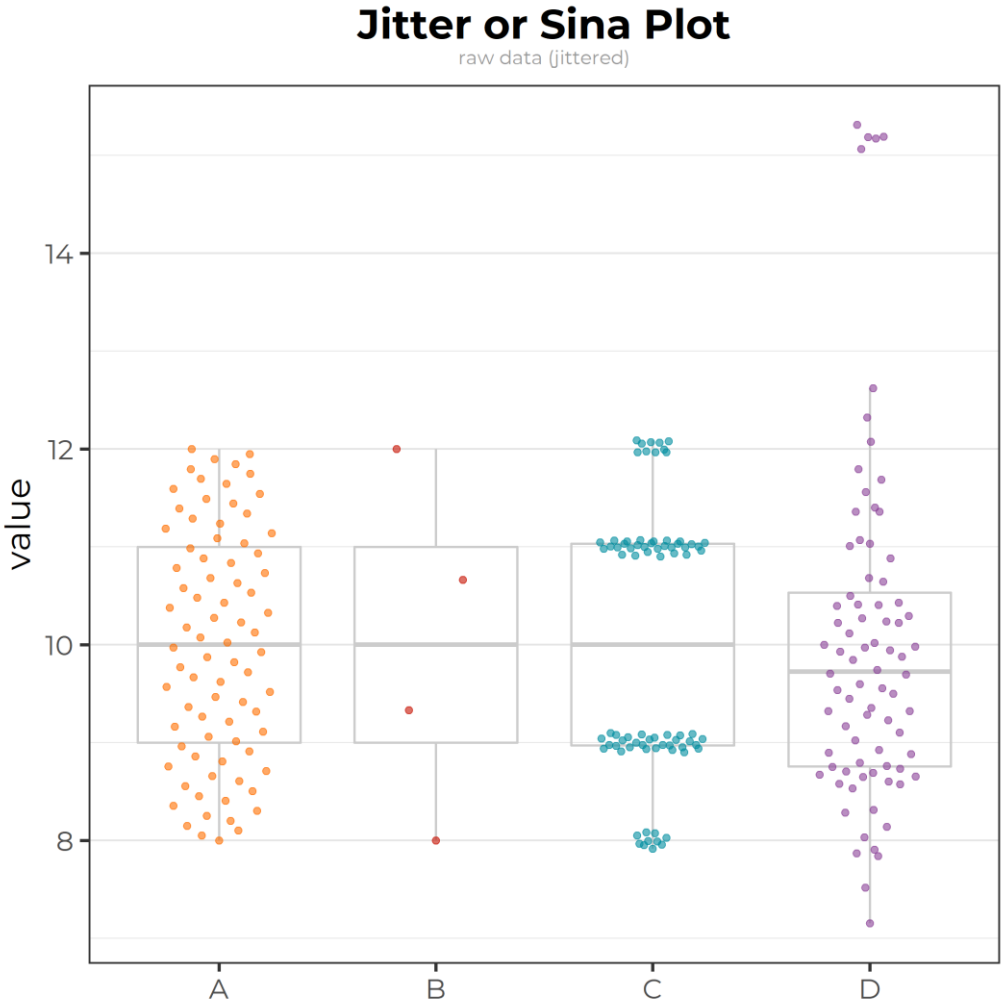
- No info about sample size
- Hard to grasp for broad audience



- Shows distribution and sample size
- Can be combined with summary stats

But:

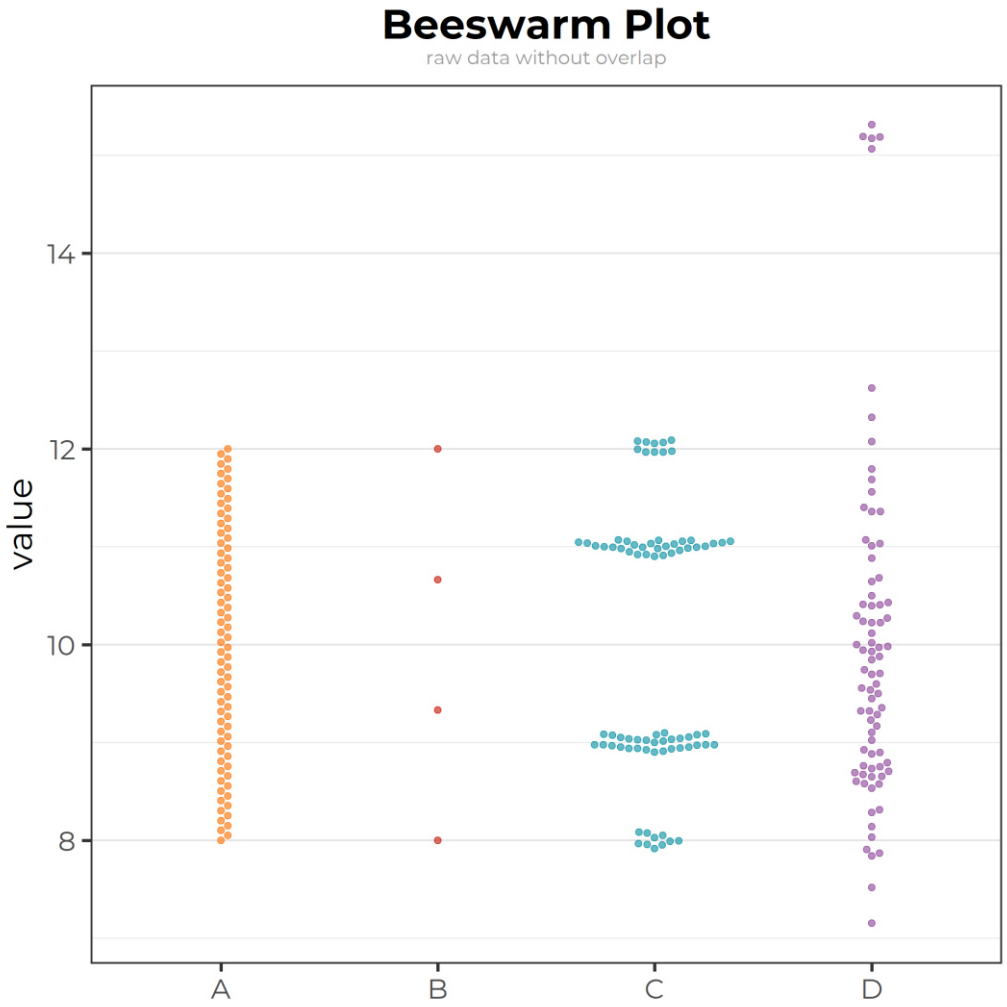
- Sample size hard to estimate



- Allows to estimate sample size and distribution
- Avoid overplotting by using transparency and/or jitter

But:

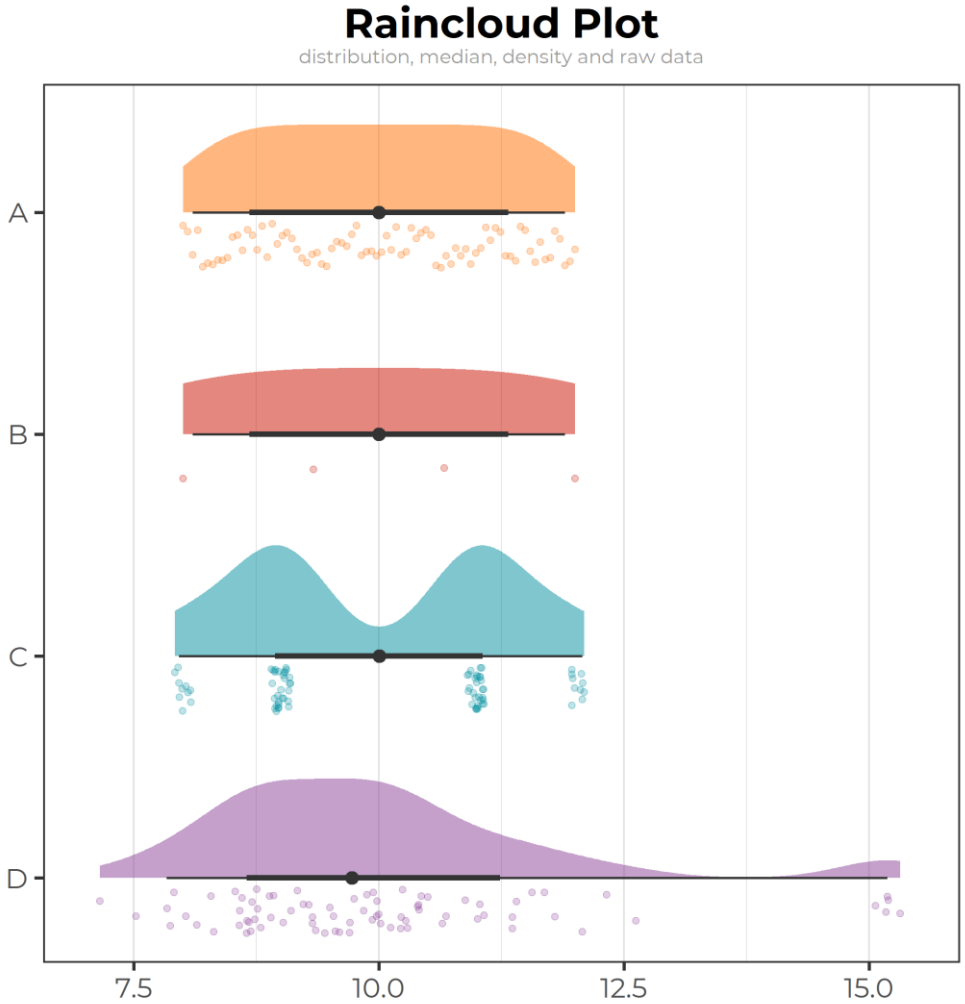
- Difficult in case of large sample size



- Allows to estimate sample size and distribution
- Prevents overplotting

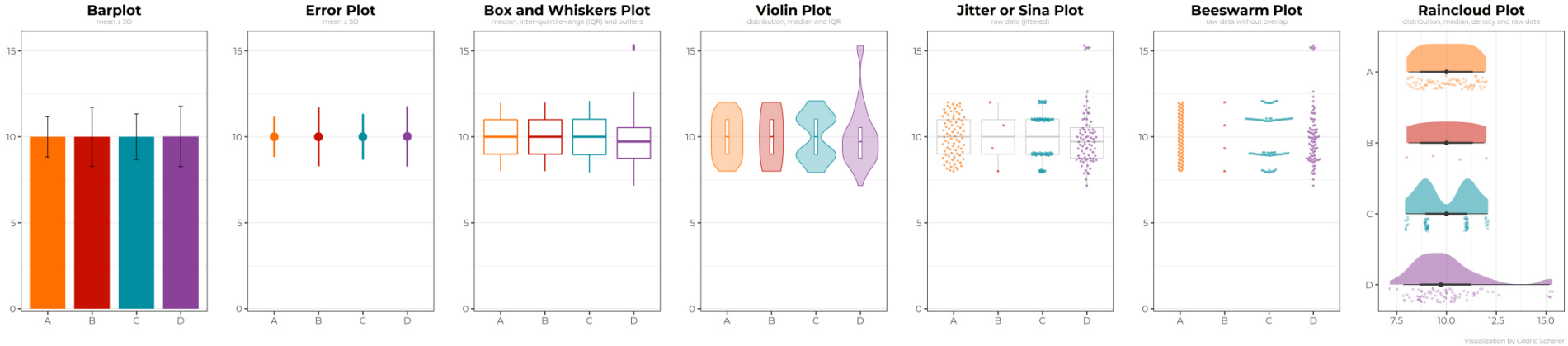
But:

- Difficult in case of large sample size



- Allows to estimate distribution, sample size and summary stats

Chart Type Choice

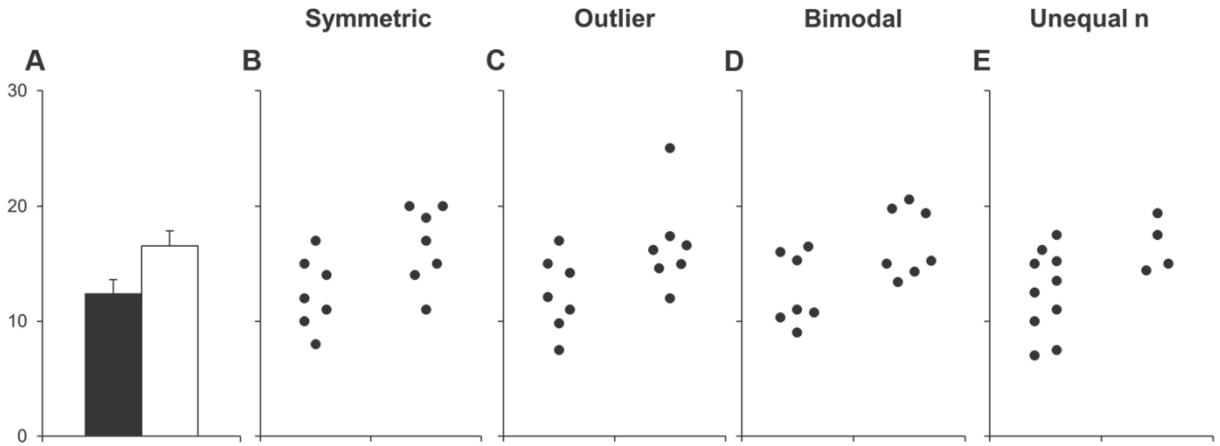


- Try several plot types and combinations
- Always check raw data and sample size
- Be as precise as possible
- Adjust chart type to audience

PLOS BIOLOGY

Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm

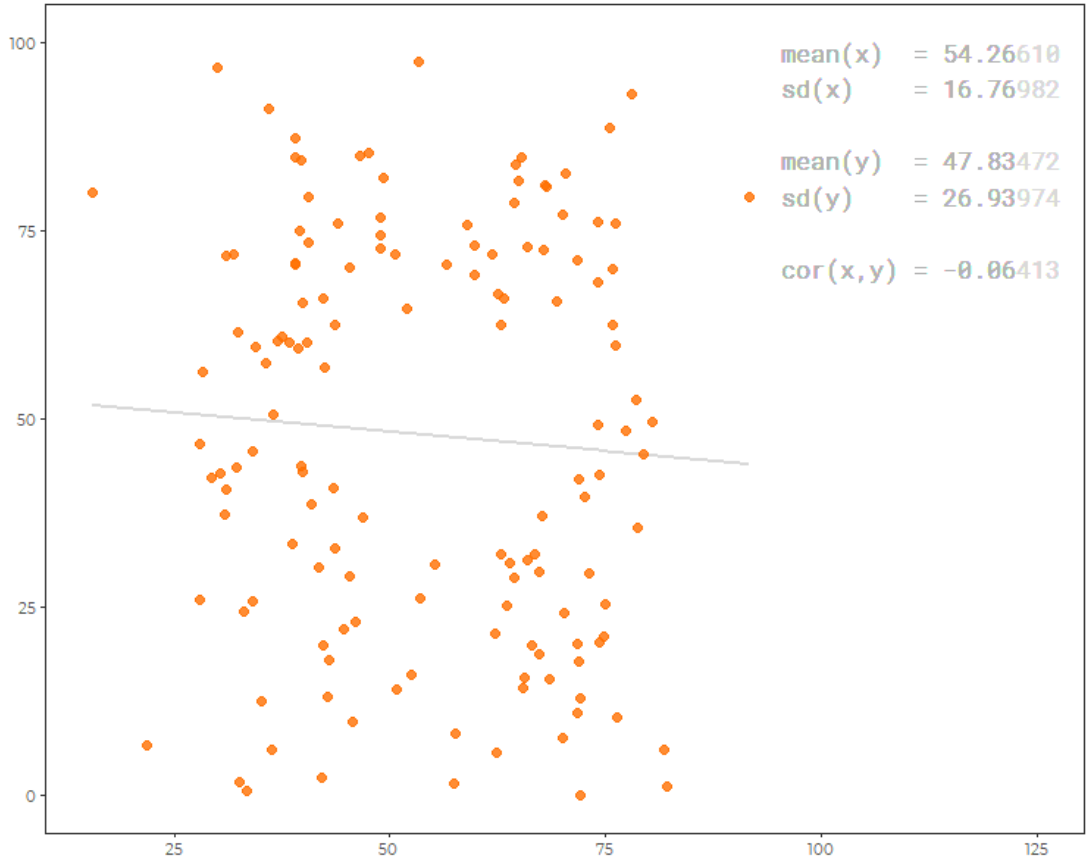
Tracey L. Weissgerber[1]*, Natasa M. Milic[1,2], Stacey J. Winham[3], Vesna D. Garovic[1]



Test			p value	
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

The Datasaurus Dozen

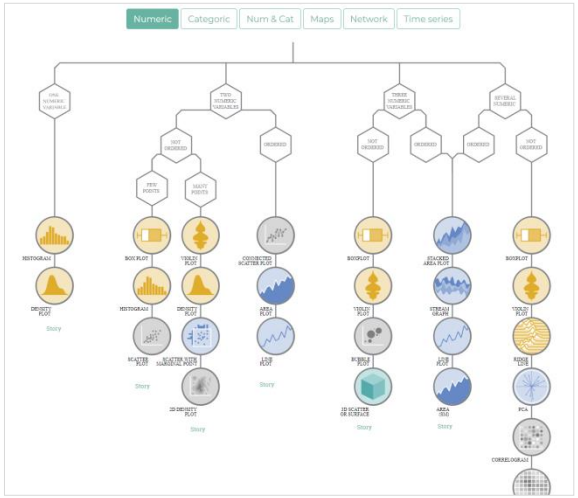
Different datasets – nigh-identical summary statistics



"The Datasaurus Dozen"
based on Anscombe's Quartet and Alberto

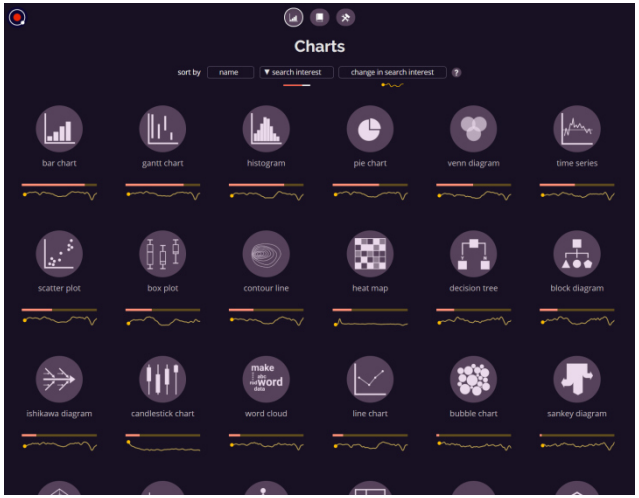
Idea by Alberto Cairo, Justin Matejka & George Fitzmaurice
Visualization by Tom Westlake & Cédric Scherer

Chart Type Choice



data-to-viz.com

datavizproject.com



visualizationuniverse.com/charts

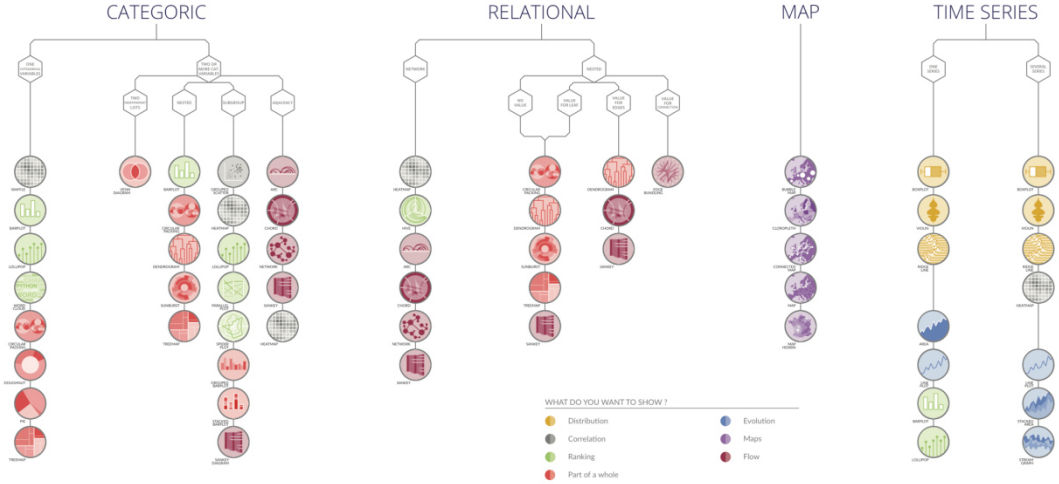
from Data to Viz

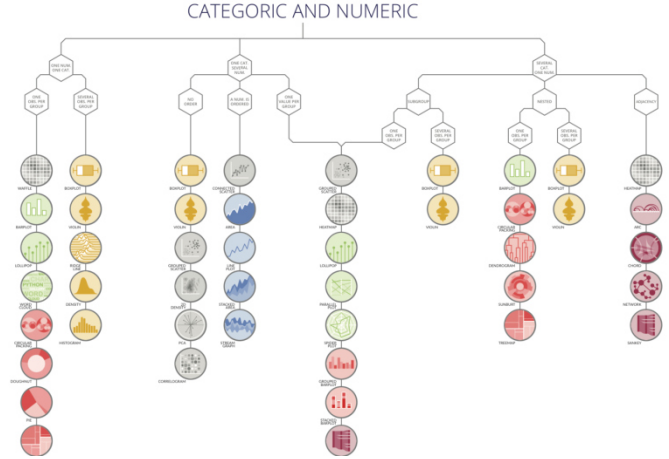
'From Data to Viz' is a classification of chart types based on input data format. It will help you find the perfect chart in three simple steps :

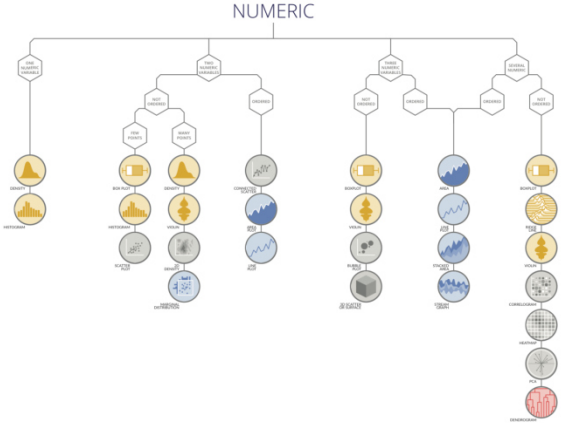
1. Identify what type of data you have.
2. Go to the corresponding decision tree and follow it down to a set of possible charts.
3. Choose the chart from the set that will suit your data and your needs best.

Dataviz is a world with endless possibilities and this project does not claim to be exhaustive. However it should provide you with a good starting point. For an interactive version and much more, visit:

data-to-viz.com







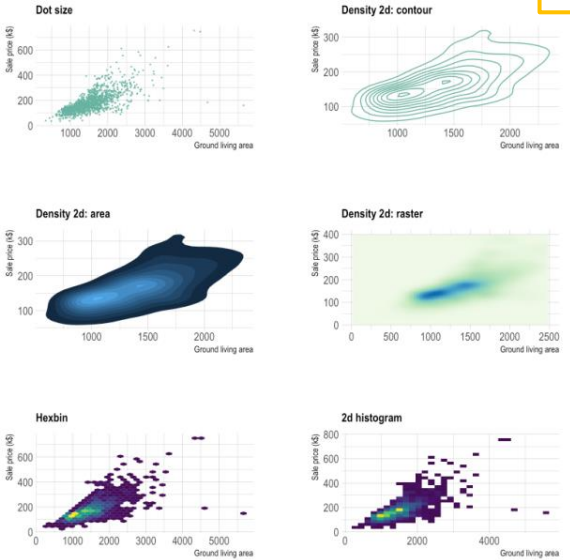
https://github.com/holtzy/data_to_viz/blob/master/img/poster/poster_screenshot.png

data-to-viz.com

Overplotting

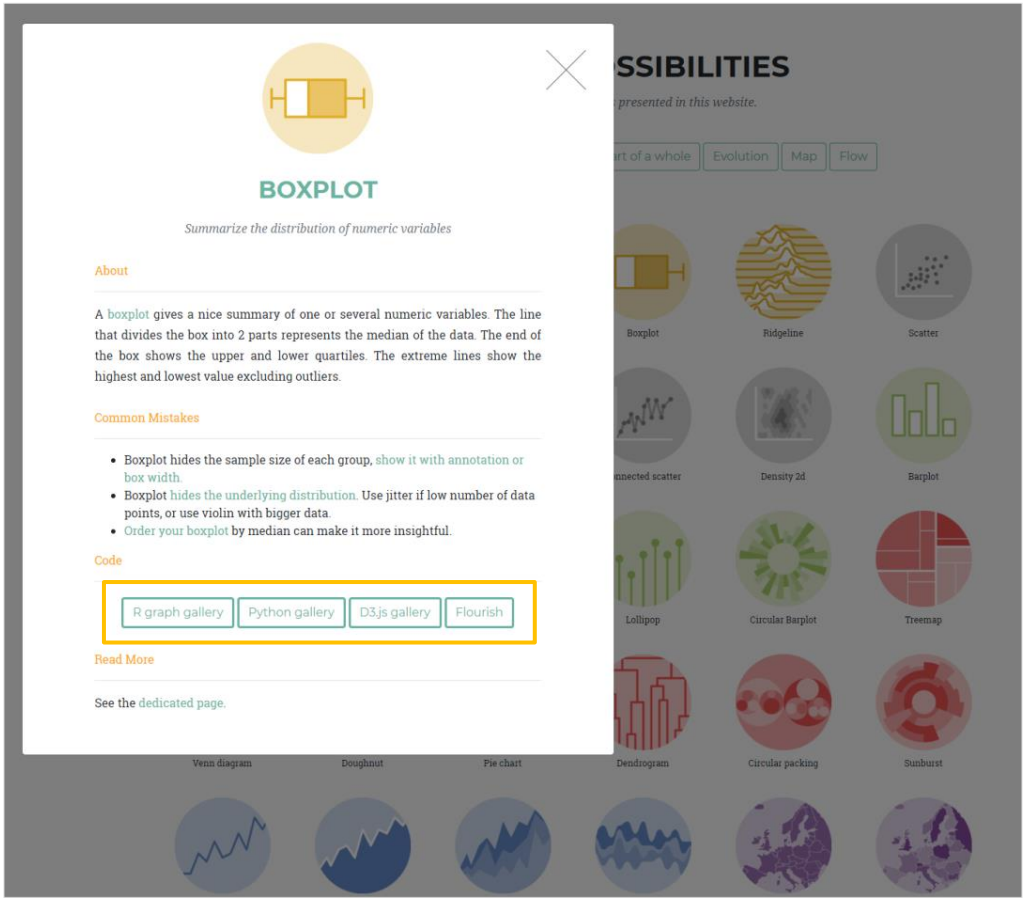
The most common pitfall with scatterplot is overplotting: when the sample size gets big, dots are plotted on top of each other what makes the chart unreadable. There are several work around to avoid this issue as describe in this specific post. Here is a summary of the different offered techniques:

CODE



Going further

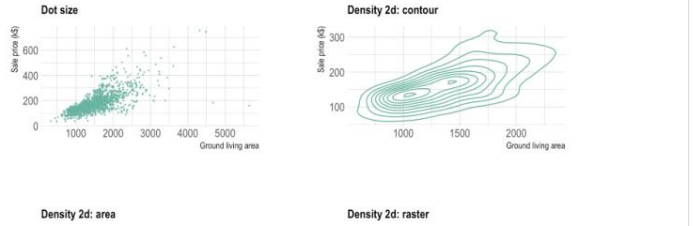
You can learn more about each type of graphic presented in this story in the dedicated



data-to-viz.com

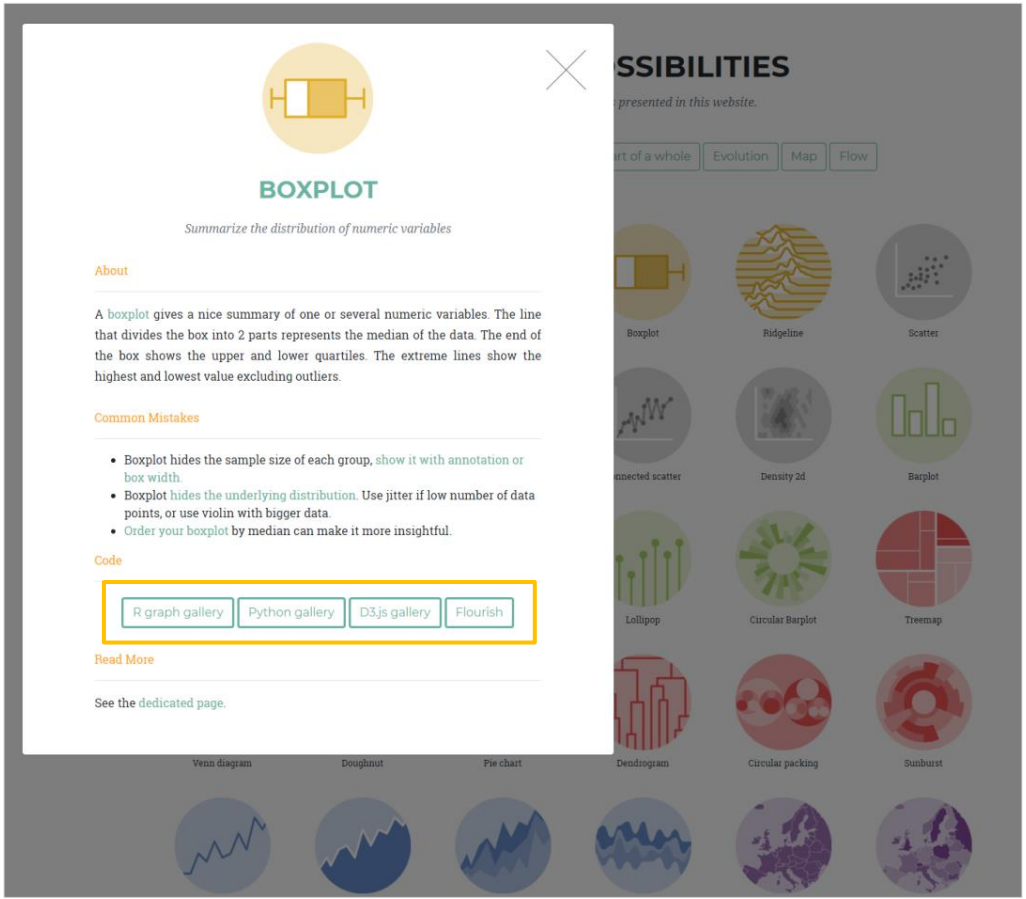
HIDE

```
# code for all graphics:
p <- data %>%
  ggplot( aes(x=GrLivArea, y=SalePrice/1000)) +
    theme_ipsum() +
    theme(
      plot.title = element_text(size=12)
    ) +
    ylab('Sale price (k$)') +
    xlab('Ground living area')

# Reduce dot size
p1 <- p + geom_point(color="#69b3a2", alpha=0.8, size=0.2) + ggtitle("Dot size")

# Use density estimate
p2 <- p + geom_density2d(color="#69b3a2") + ggtitle("Density 2d: contour")

# Use density estimate (area)
p3 <- p + stat_density_2d(aes(fill = ..level..), geom = "polygon") + ggtitle("Density 2d: area") + theme(legend.positio
n="none")

# With raster
p4 <- p +
  stat_density_2d(aes(fill = ..density..), geom = "raster", contour = FALSE) +
  scale_fill_distiller(palette=4, direction=1) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  theme(
    legend.position='none'
  ) +
  ggtitle("Density 2d: raster") +
  xlim(0,2500) +
  ylim(0,400)

# Hexbin
p5 <- p + geom_hex() +
    scale_fill_viridis() +
    theme(legend.position="none") +
    ggtitle("Hexbin")

# 2d histogram
p6 <- p + geom_bin2d( ) +
    scale_fill_viridis( ) +
    theme(legend.position="none")  +
    ggtitle("2d histogram")

p1 + p2 + p3 + p4 + p5 + p6 + plot_layout(ncol = 2)
```



Density 2d: area

Density 2d: raster



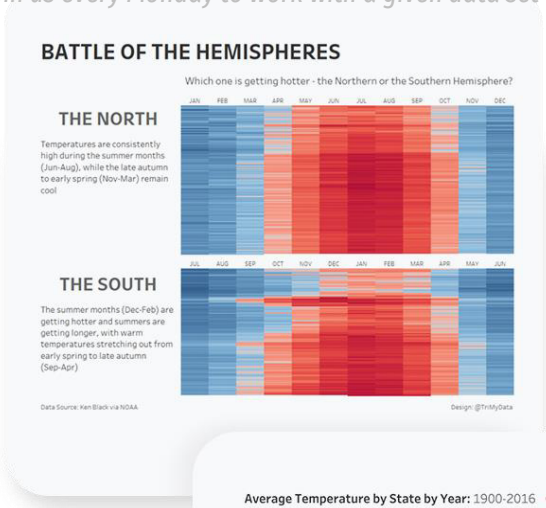
Train Your Skills and Get Inspired!

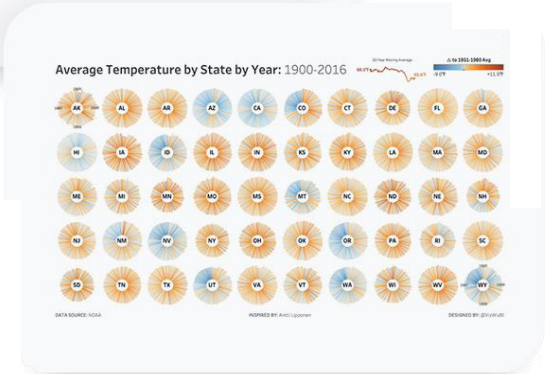
Social DataViz Projects

Social DataViz Projects

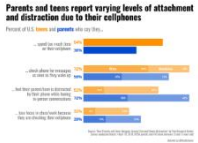
- **#MakeoverMonday** – weekly; any tool (but mainly contributions with Tableau)

"Join us every Monday to work with a given data set and create better, more effective visualizations and help us make information more accessible."

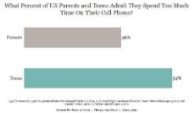




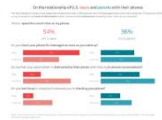
FAVORITES



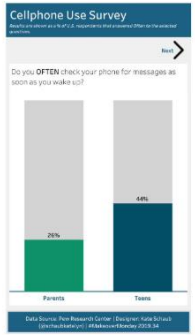
Author: Cédric Scherer
GGPlot



Author: Anna Foard
Link: Tableau Public



Author: Alex Avallone
Link: Tableau Public



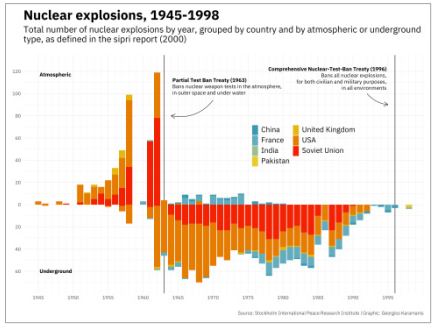
Author: Kate Schaub
Link: Tableau Public

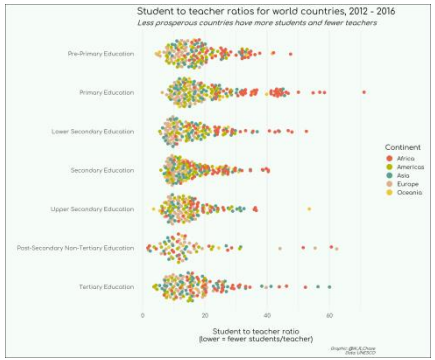
Author: Erik Rettmann
Link: Tableau Public

Social DataViz Projects

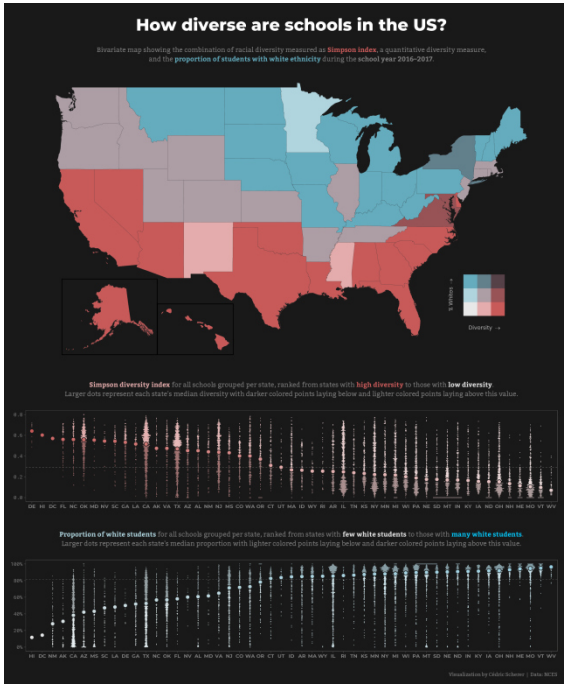
- **#TidyTuesday** – weekly; aimed at `R` and `tidyverse/ggplot2` users

  "A weekly data project aimed at the R ecosystem [with] an emphasis on understanding how to summarize and arrange data to make meaningful charts with ggplot2, tidyr, dplyr, and other tools in the tidyverse ecosystem."

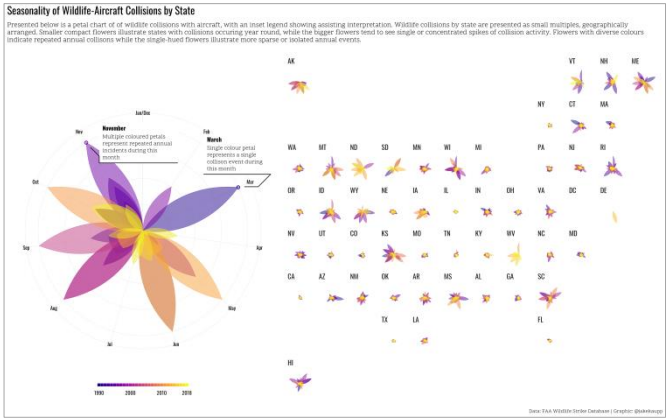


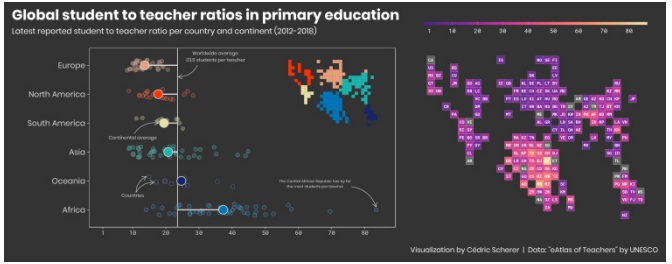


Upper: Georgios Karamanis
Lower: Will Chase



Cédric Scherer



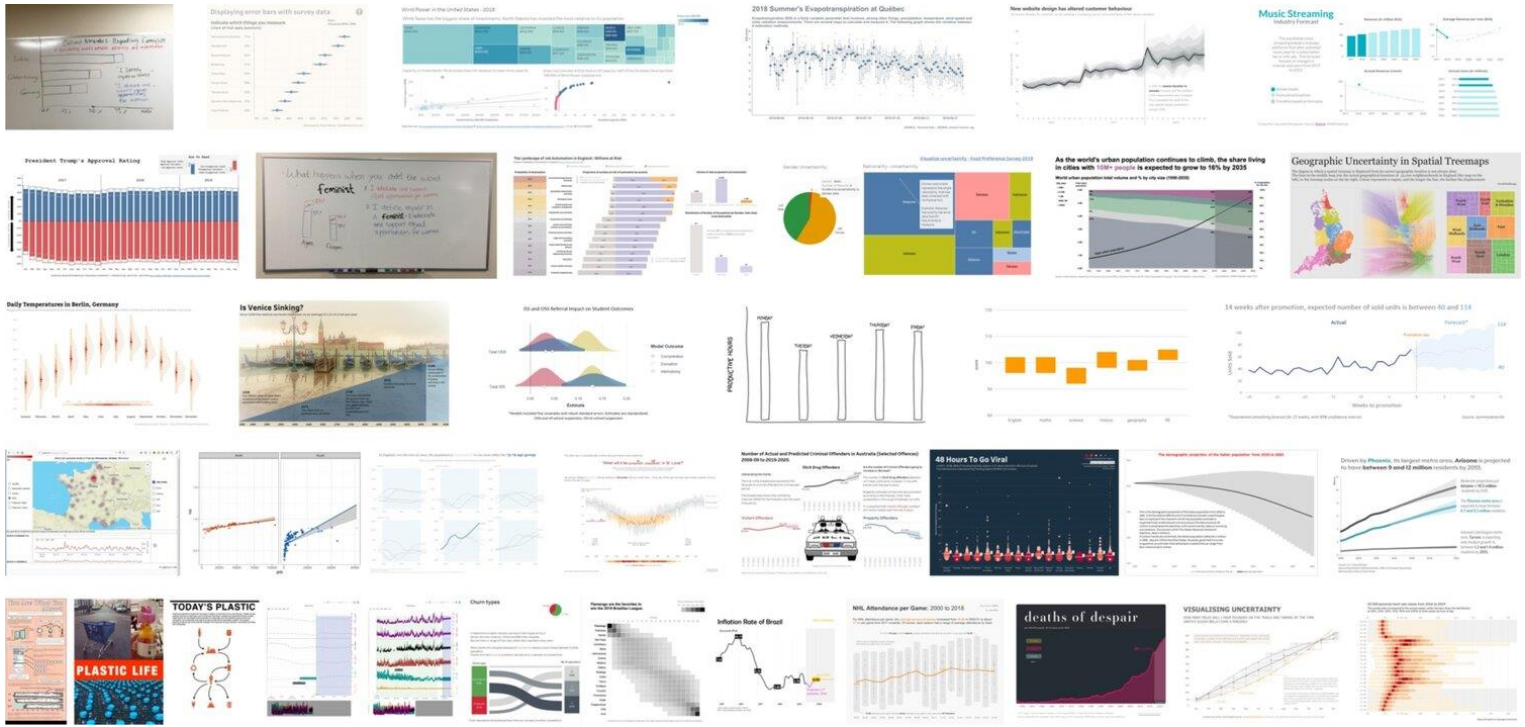


Upper: Jake Kaupp
Lower: Cédric Scherer

Social DataViz Projects

- **#SWDchallenge** – monthly; diverse set of tools and techniques

  "The #SWDchallenge is a monthly challenge where you can practice and apply data visualization and storytelling skills. Think of this as a safe space to try something new: test out a new tool, technique, or approach."



Contributions to #SWDchallenge September 2019