GOOD PLOTS FOR PUBLICATION

Bioinformatics Awareness Days @ TIGEM
July 10th, 2023

Eugenio Del Prete, M. Eng., Ph.D.
BIOINFORMATICS CORE
e.delprete@tigem.it

Bioinformatics Core: Tasks

- **STATISTICAL DATA ANALYSIS**
  Experimental Design, Hypothesis Testing, Power Analysis Differential Expression Analysis, Cluster Analysis, Time Series Data Analysis, Survival Analysis, Correlation Analysis
- **OMICS**
  Microarray Analysis, Gene Networks, Pathway Analysis, TFBS Identification, Gene Annotation, Integration, Protein Analysis, Drug Networks
- **NEXT GENERATION SEQUENCING**
  Whole Exome, Targeted Gene, RNA, miRNA, ChIP, Visualization, Interpretation
- **DATABASE AND SOFTWARE**
  DB Creation, DB Maintenance, Web Sites Creation, Web Service Support
- **BIOINFORMATICS AND (BIO)STATISTICS TRAINING**

Bioinformatics Core: People

DIEGO DI BERNARDO

https://www.tigem.it/research/facilities/core-facilities/bioinformatics

https://bioinformatics.tigem.it/

DIEGO CARRELLA

ROSSELLA DE CEGLI

XAVIER BUJANDA CUNDIN

EUGENIO DEL PRETE

Bioinformatics Core: Something about Me

- **TLC ENGINEER @ UNIVERSITY OF ROME 'SAPIENZA'**
  MAIN TOPICS: Signal Processing, Remote Sensing, Bioinformatics
  THESIS: miRNA Analysis, Genomic Data Mining, Consensus Analysis, PSSM Creation
- **BIOINFORMATICS RESEARCH FELLOW @ INSTITUTE OF FOOD SCIENCES (CNR)**
  Protein Prediction and Classification, Protein Analysis, Proteomic Mass Spectra Analysis, Sequence Alignment and Phylogenetic Tree, Docking
- **PHD IN APPLIED BIOLOGY @ UNIVERSITY OF BASILICATA**
  Celiac Disease and Comorbities, Microarray Data Analysis, Ontologies, Gene Set Enrichment Analysis, Semantic Similarity, Proteomic Mass Spectra Analysis
- **BIOINFORMATICS RESEARCH FELLOW @ INSTITUTE OF APPLIED MATHEMATICS (CNR)**
  Proteomic Mass Spectra Analysis, Metabolomic (Lipidomic) Data Analysis, Web Tools Developer, Hypothesis Tests, Omics Data Integration
- **BIOSTATISTICIAN AND DATA SCIENTIST @ TIGEM**

Outline

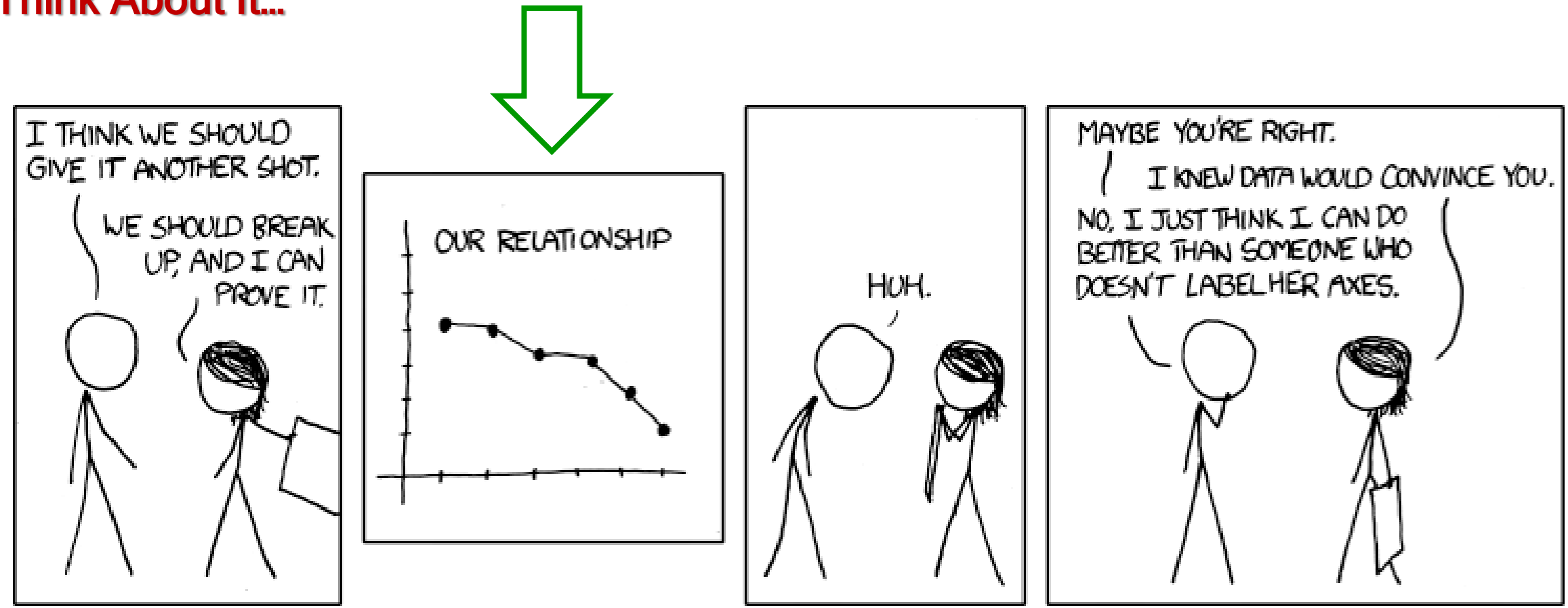
BAD PLOTS
- The Worst Error
- Bad Habits: Examples

GOOD PLOTS
- Good Habits: Rules
- Top Science Visualization Trends in 2022

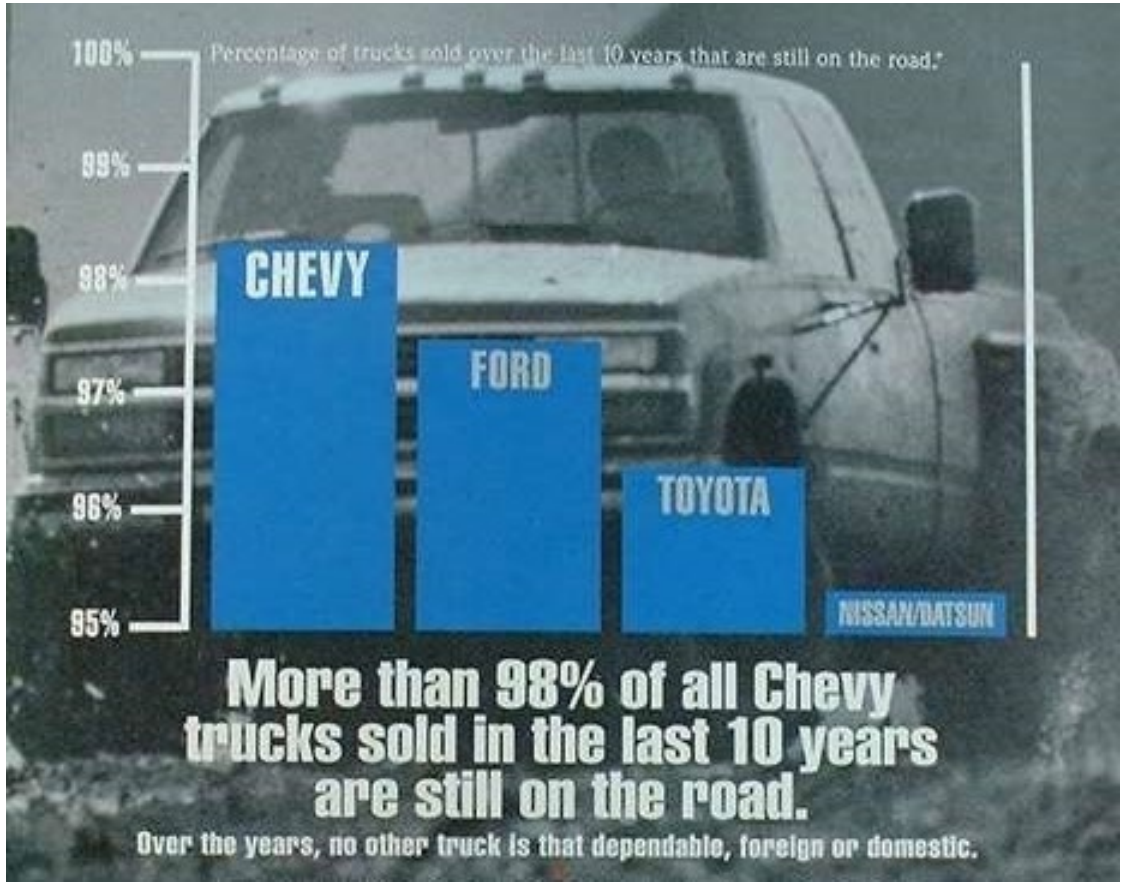
EXAMPLES
- Example 1: SuperPlot with Prism
- Example 1: SuperPlot with R
- Example 2: PCA with Prism

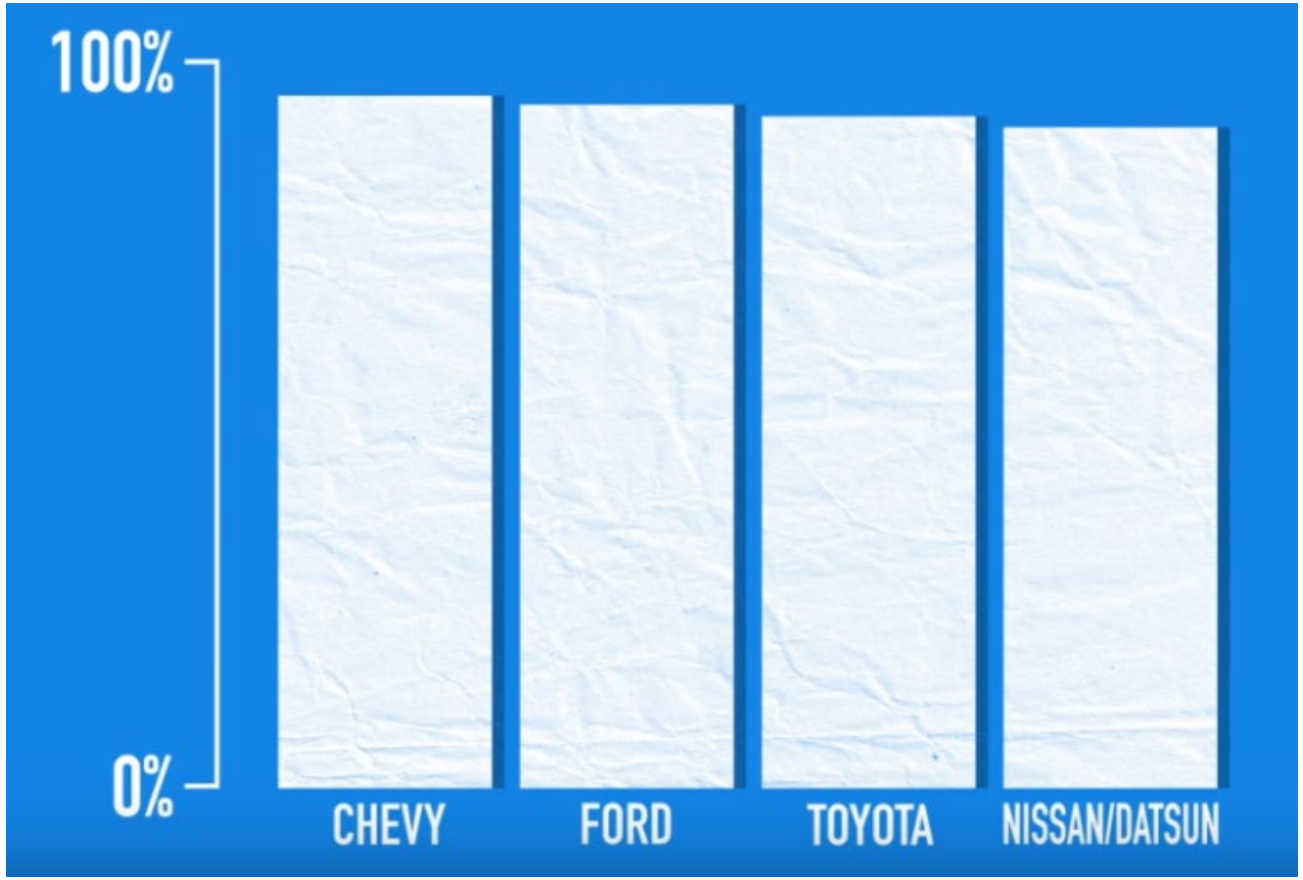
CONCLUSION
- Take Home Message
- Final Remarks

Think About It...



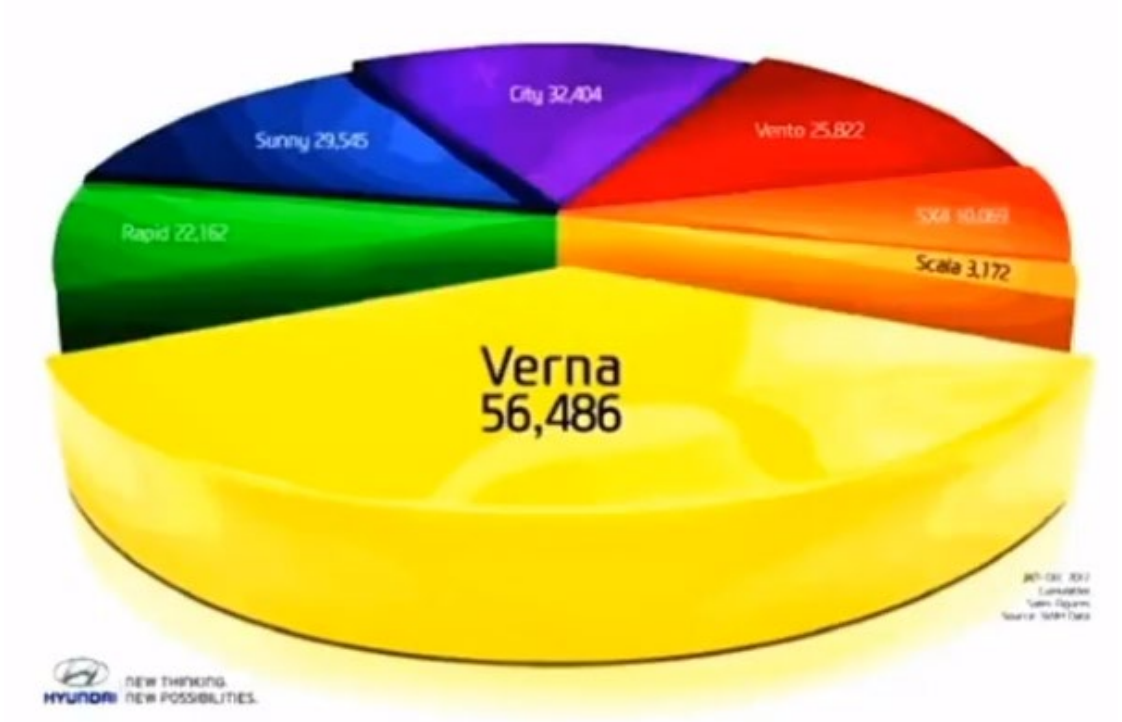
Bad Habits

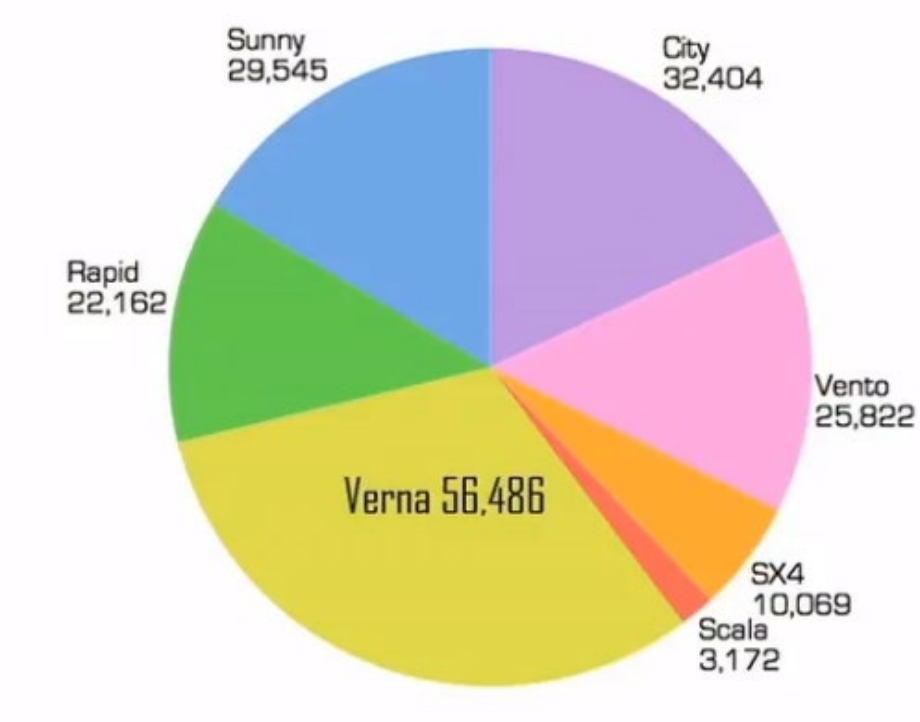




MYSTIFICATION

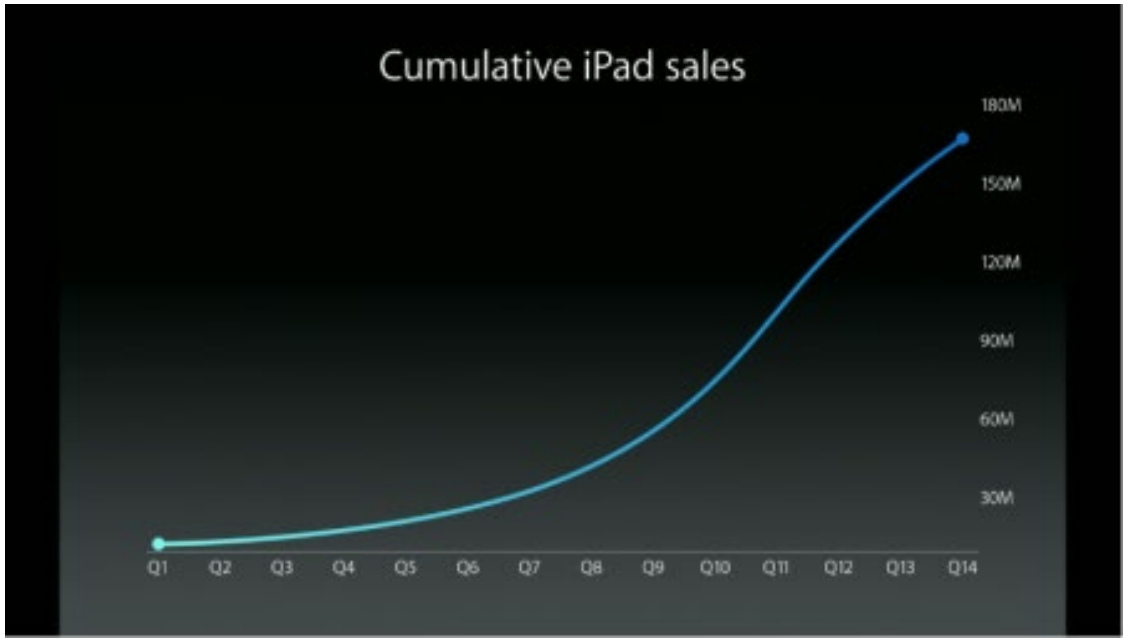
Bad Habits

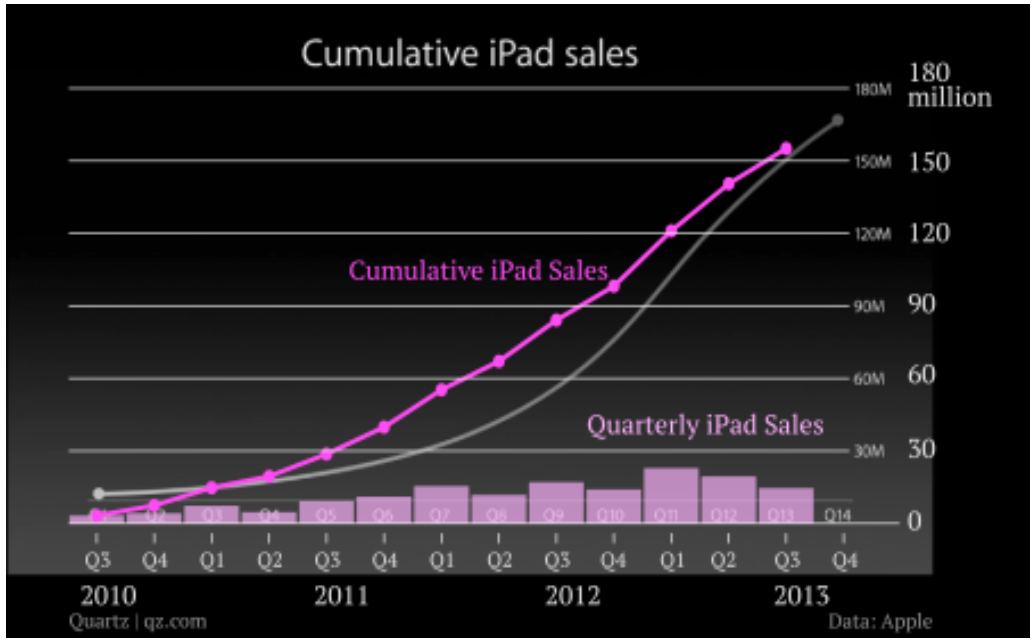




CONFOUNDING

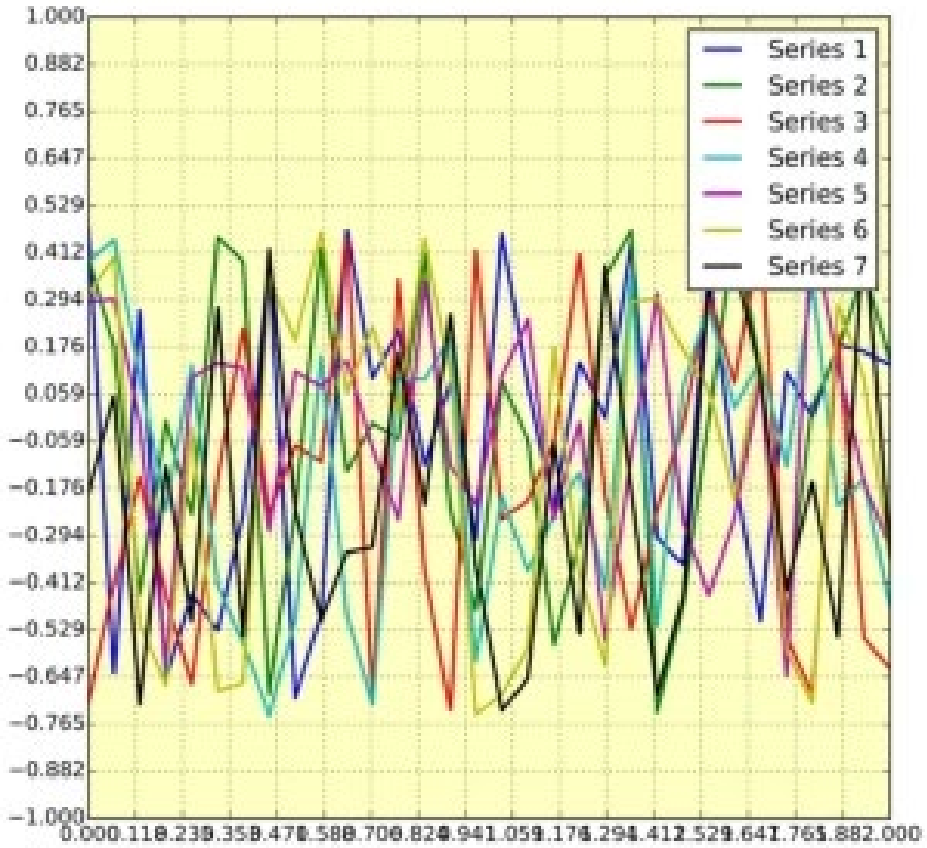
Bad Habits



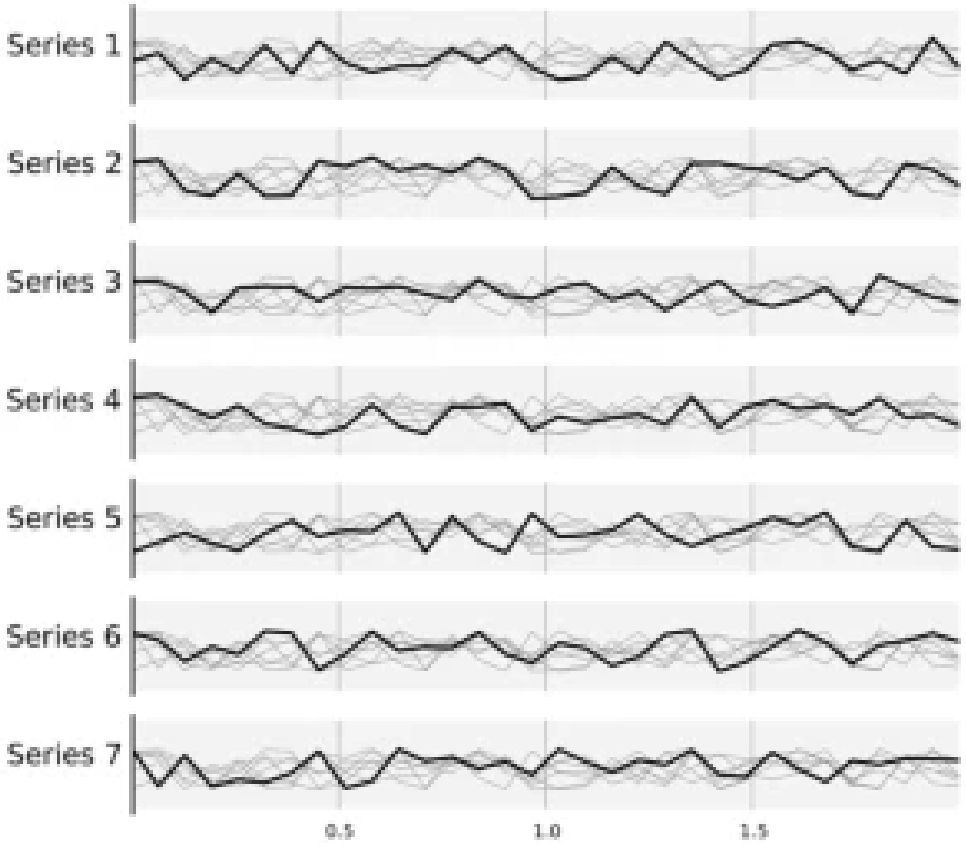


MISLEADING

Bad Habits



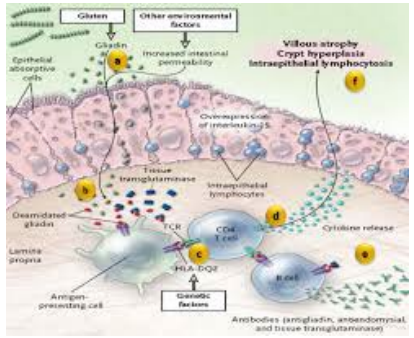
UNREADABLE

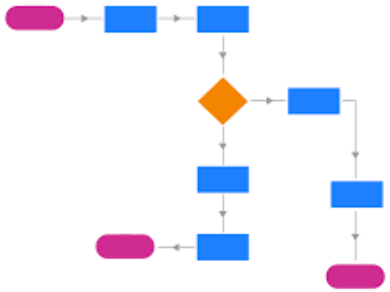


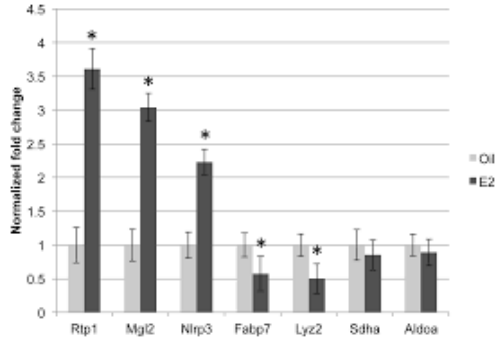
Good Habits

Audience
- Generic audience
- Specific audience
- Scientific journal

Message
- Express an idea
- Define a problem
- Report a result







Adaptation
- Support media

Good Habits

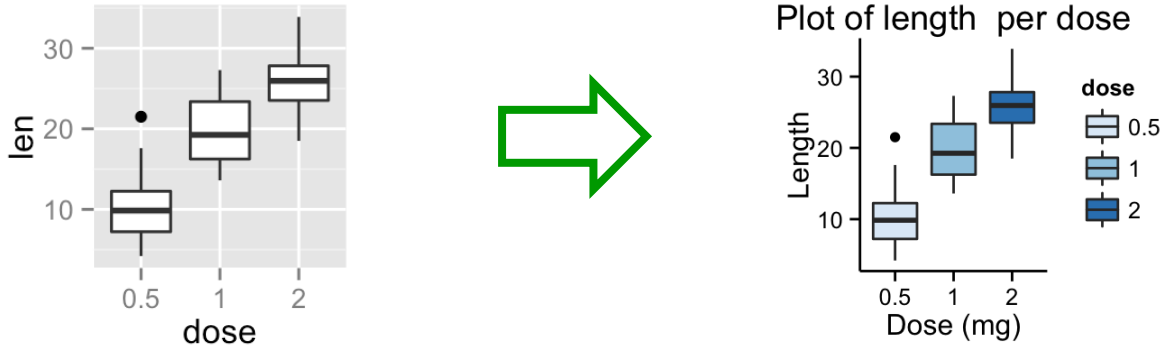
Caption

- Define the figure
- Describe the figure
- Report important values

Figure 1. Normalized fold change among conditions. The x-axis reports the genes for both the conditions (treatment A, treatment B), the y-axis reports the normalized expression in term of fold change (…) Legend: * p-value < 0.05, ** p-value < 0.01.

Default & Colors

- Title
- Background
- Suitable colors



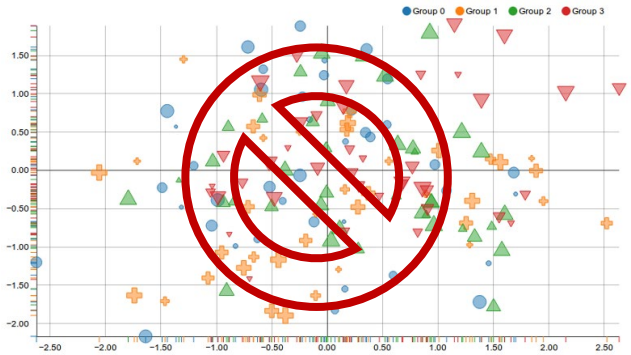
Axes Modification

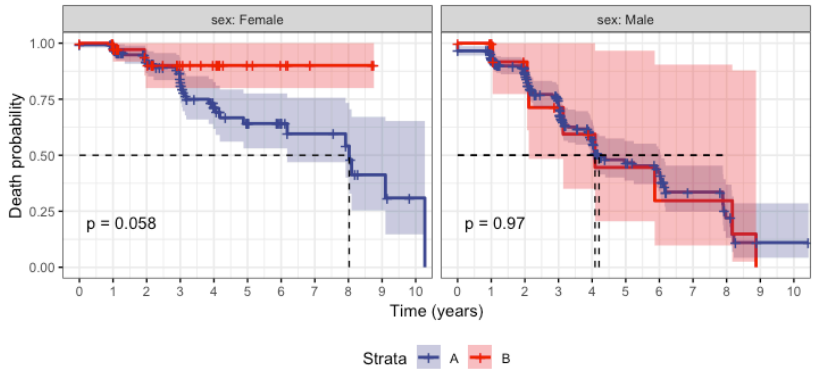
- Cut-off
- Scaling
- Normalization

Good Habits

"Chartjunk"

- No quibbling
- No redundancy
- Use facets





Hints

- State-of-the-art
- Scientific journal rules
- Straight to the goal

Tools

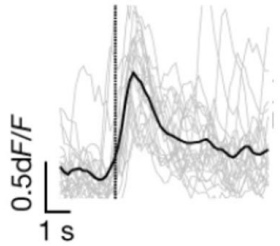
- “Only for the brave”
- Suitable programs

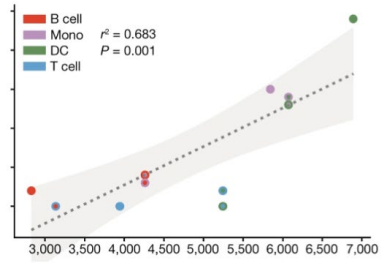
Prism (GraphPad)
Matplotlib (Python)
Ggplot (R)
GIMP (Linux)
Paint (Windows)
Adobe Photoshop
Cytoscape

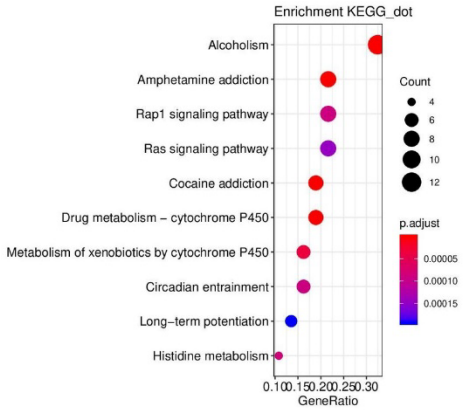
Top Science Visualization Trends in 2022

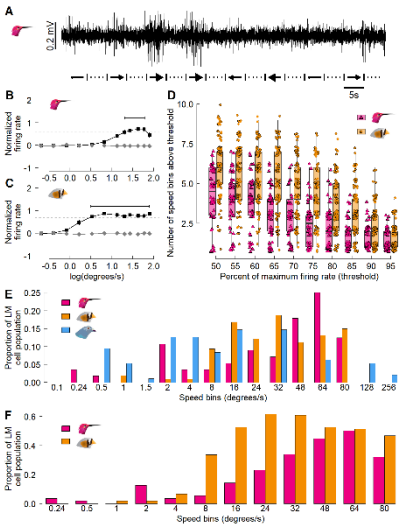
Top Trends in 2022

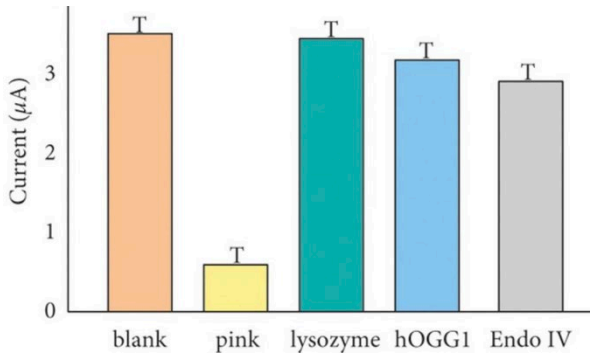
- Color blindness palette
- Grey shades (E. Tufte)
- Hybrid chart - tables
- Multi panel plots
- Ts error bars
- Axis breaks
- Icon libraries (Biorender)
- Climate stripes
- Space for figures

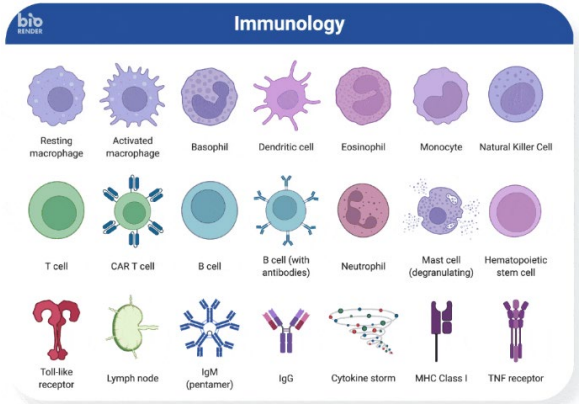






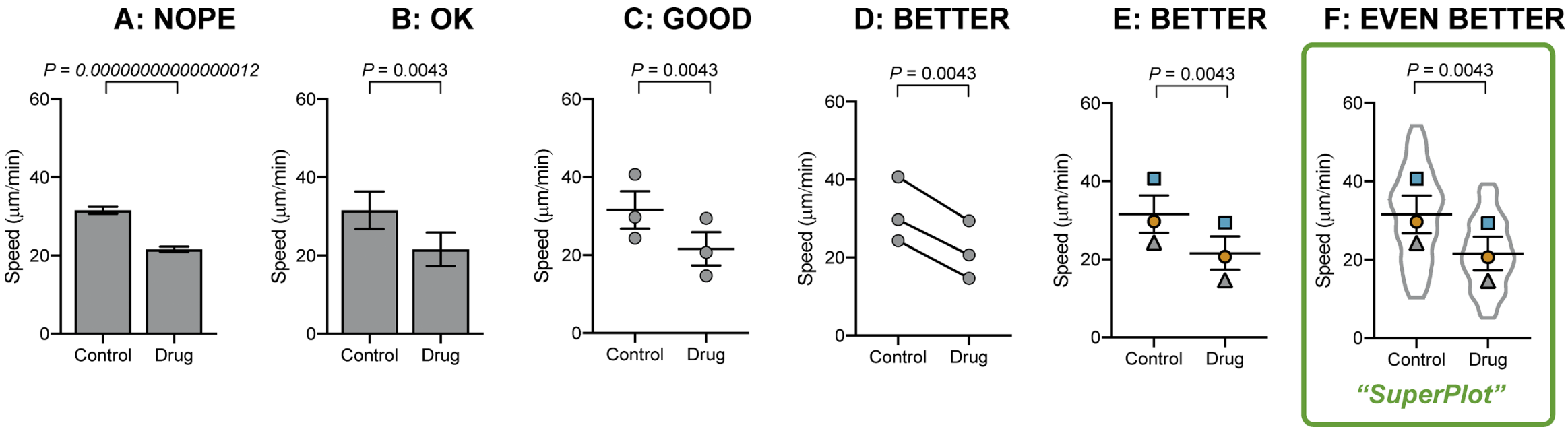






Example 1: Cell-level Variability and Reproducibility

- Suppose to test a **treatment** that could change the **speed** of crawling cells



Cells info | Biological replicates info | Paired info | Overall info

- **SuperPlot** convey more information: replicates, samples, pairment, statistics

Example 1: SuperPlot with Prism 9.4.0 (GraphPad)

- **Upload dataset and mean values**
  - control-placebo and drug (2 conditions)
  - 3 biological replicates per condition (6 samples)
  - 50 measurements per sample (300 values)

- **Depict the jitter plot of the data**
  - Show the cell-level variability
  - Select colors per biological replicates

- **Depict the error plot of the mean values**
  - Show the sample-level variability
  - Select colors per biological replicates

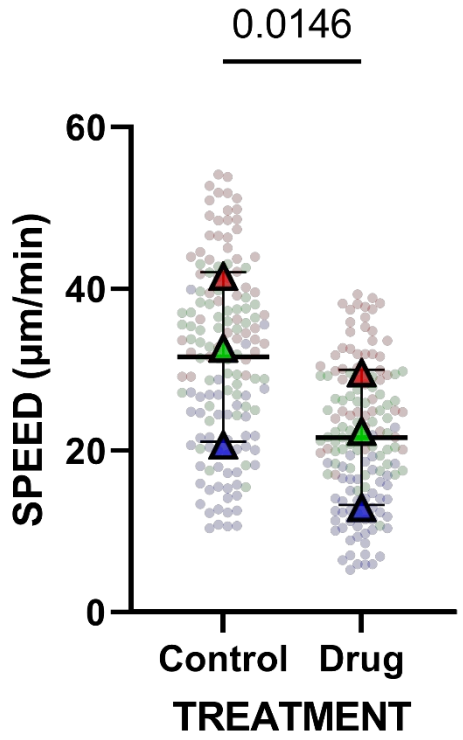
1. File → New → New Project File → Column
2. Paste data (one blank cell per replicates)
3. Paste mean values
4. Rename Data Tables

5. Graphs → New Graphs → Table: Data → Individual values: Scatter plot → Plot: No line or error bar
6. Define title and labels
7. Select first replicate → Change → Format Points → Symbol Color

8. Graphs → New Graphs → Table: Mean → Individual values: Scatter plot → Mean with SD
9. Define title and labels
10. Select all → Change → Format Points → Symbol Shape (Symbol Size, Symbol Color)

Example 1: SuperPlot with Prism 9.4.0 (GraphPad)

- **Superimpose the plots**
  - Adjust the y-axis (same range)
  - Add the statistical significance
  - Combine the two levels of variability

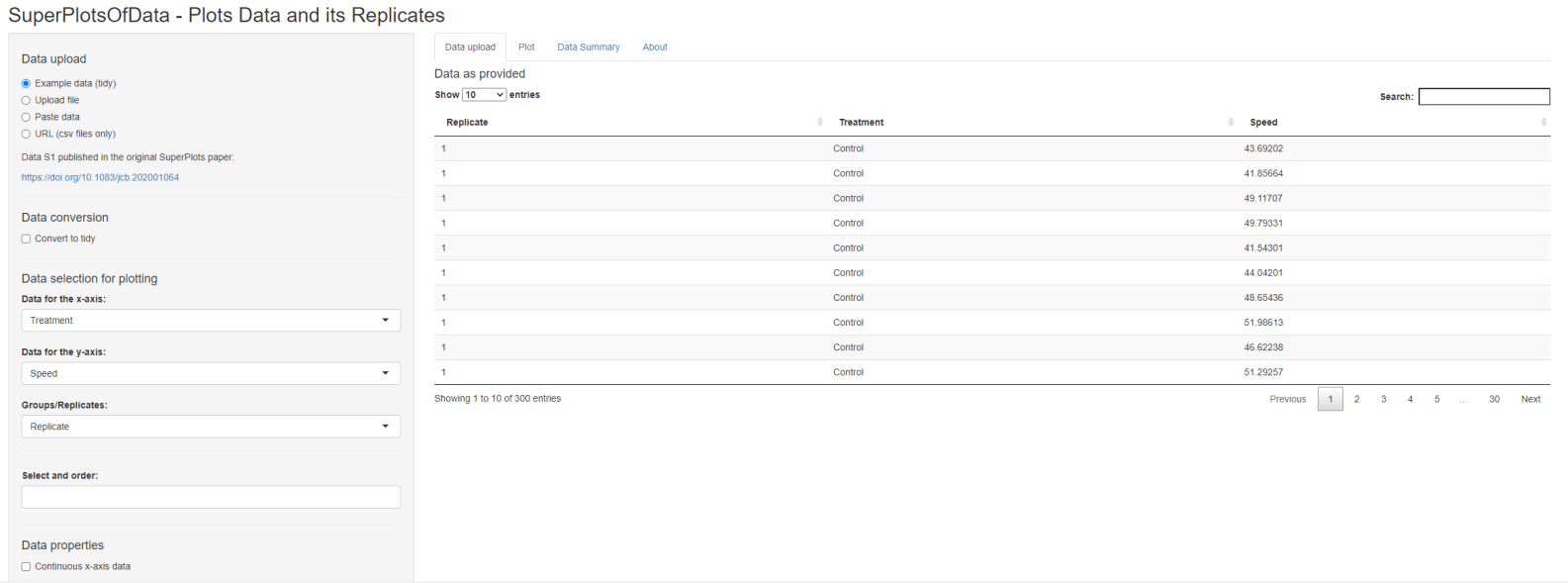


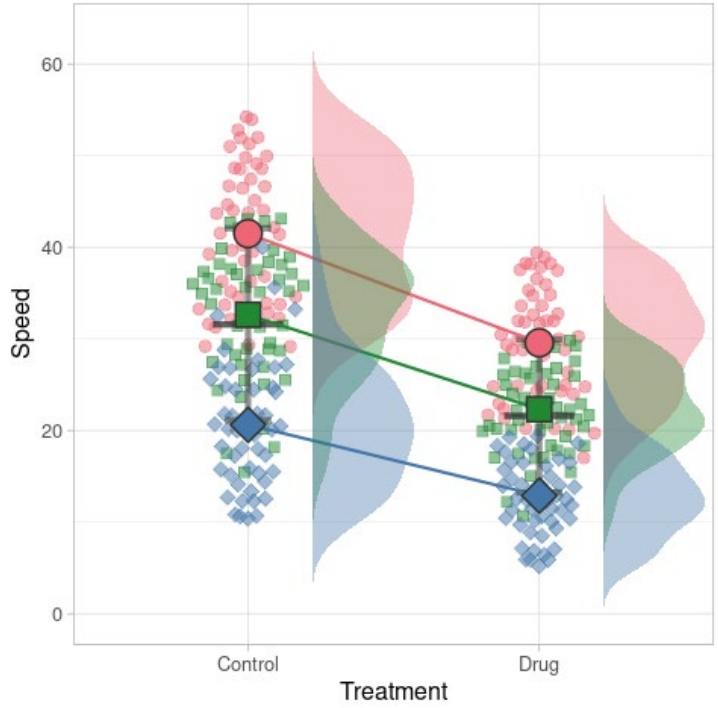
11. Double click on y-axis (Mean) → Left Y axis → Deselect 'Automatically…' → Range, Maximum: 60
12. Double click on y-axis (Data) → Left Y axis → All ticks, Ticks direction: None & Location: None → Same for X axis
13. Results → New Analysis → Column Analyses: t-tests → Paired
14. Draw (Mean) → Format pairwise comparisons → Appearance, Display options: P value (numbers) → Lift up the p-value
15. Layouts → New Layout → (Standard) → Drag the plots → Select the error plot → Change → Equalize scaling factor → Change…: Increase… → Superimpose the plots
16. File → Export

Example 1: SuperPlot with R (Shiny app)

- Several tools for depicting similar plots

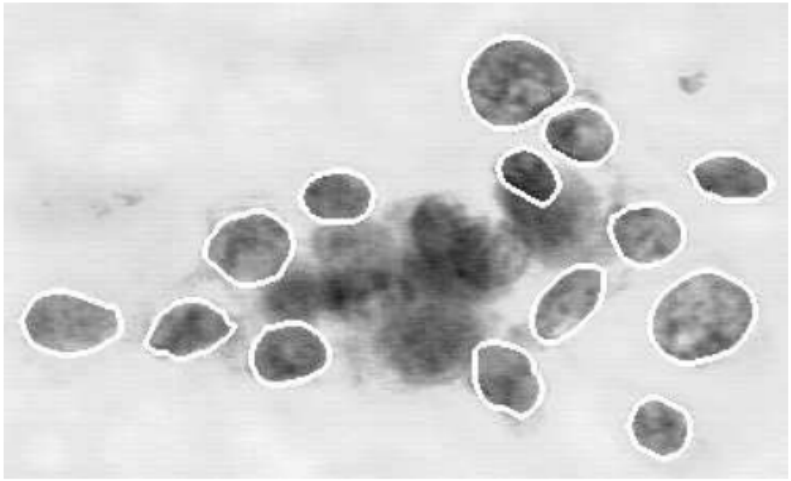
R, RStudio, Shiny (package)





Example 2: PCA plots with Prism 9.4.0 (GraphPad)

- Suppose to discriminate the **malignancy** of tumor by **images of cells** from breast cancer **tissue biopses**



- **Principal Component Analysis (PCA)** provide a dimensionality reduction (linear) method to cope with the presence of multiple features and ‘resume’ their information

Example 2: PCA plots with Prism 9.4.0 (GraphPad)

- **Upload dataset**
  - malignancy of cells (1 categorical variable)
  - 10 features of cells (10 continuous variable)
  - 569 cells

- **Perform PCA**
  - Select the method for selecting PCs
  - Select colors per diagnosis

- **Adjust the colors for the different plots**
  - Select colors per Loadings
  - Select colors per diagnosis in PC scores
  - Select colors and labels in Proportion of variance

1. File → New → New Project File → Multip. variables
2. Paste data (first row as column names)
3. Check the nature of the variables

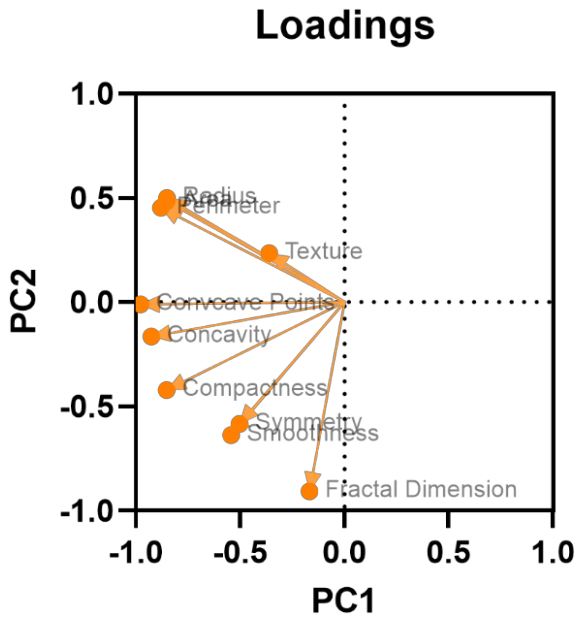
4. Analysis → Analyze → Multiple variable analysis: PCA
5. Options → Method for selecting PCs: % of total explained variance (80)
6. Output → Additional variables for graphing → Labels: ID Number & Symbol fill color: Diagnosis
7. Graphs → Select all

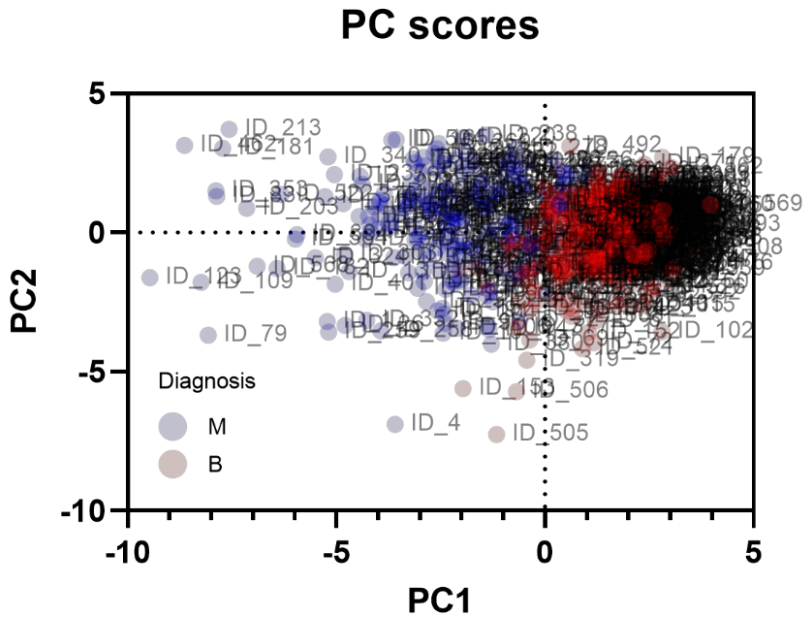
8. Loadings → Double click on one point → Symbols & Connecting Lines
9. PC Scores → Change → Change colors → Colors (Double click on one point)
10. Proportion of variance → Double click on legend → Bars and boxes & Symbols & Lines

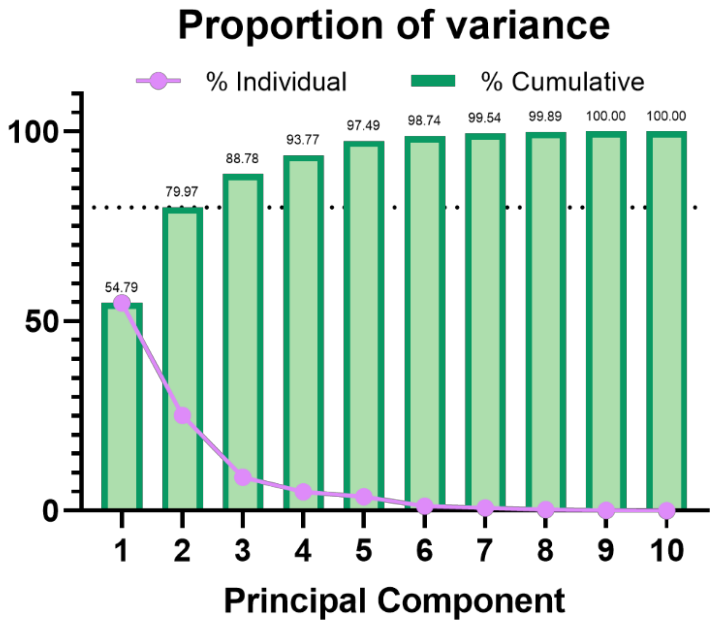
Example 2: PCA plots with Prism 9.4.0 (GraphPad)

- **Multiple plot**
  - Adjust colors and labels
  - Assemble plots in one multiple plot

11. Layouts → New Layout → (Standard) → Drag the plots → Select PC scores plot → Change → Equalize scaling factor → Change…: Increase…
12. Change labels position (if necessary)
13. File → Export







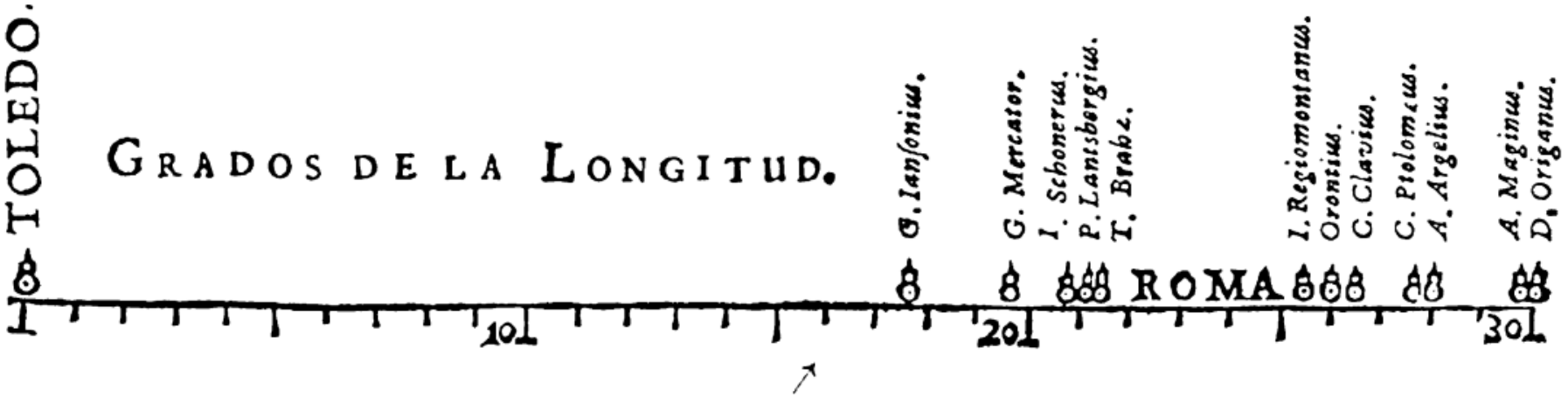
Take Home Message

- Investigating in good figures at the start of your research **will save you time and frustration**
- Save images as **TIFFs** (not JPGs or other low-quality formats) and **keep all your original files**
- After you have considered the **purpose of your figure**, choose the right graph to represent it
- Check **journal guidelines** before you build your figure to save yourself time
- Evaluate the necessity of every aspect of your figures and **eliminate any unnecessary clutter**

Final Remarks

FIGURE QUALITY IS A PAPER'S "SUIT AND TIE".

American Journal Expert (AJE)



Eugenio Del Prete, M.Eng., Ph.D.
Biostatistician and Data Analyst
Telethon Institute of Genetics and Medicine (TIGEM)
Pozzuoli (NA), Italy
e-mail: e.delprete@tigem.it

References

[1] Woolston, C. **How to dodge the pitfalls of bad illustrations**. Nature (2014).
[2] Rougier, N. **Ten Simple Rules for Better Figures**. PLoS Comput Biol (2014).
[3] Lord, S.J. **SuperPlots: Communicating reproducibility and variability in cell biology.** J Cell Biol (2020).
[4] Goedhart, J. **SuperPlotsOfData - a web app for the transparent display and quantitative comparison of continuous data from different conditions.** Mol Biol Cell (2021).
[5] Friendly, M. **The First (Known) Statistical Graph: Michael Florent van Langren and the "Secret" of Longitude.** The American Statistician (2010).

[h1] *http://www.sthda.com/english/wiki/ggplot2-box-plot-quick-start-guide-r-software-and-data-visualization*
[h2] *https://helenajambor.wordpress.com/2023/01/04/science-visualization-trends-of-2022/*
[h3] *https://huygens.science.uva.nl/SuperPlotsOfData/*
[h4] *https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)*