Data Visualization with Python

Introduction to Data Science

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

2024-II

Outline

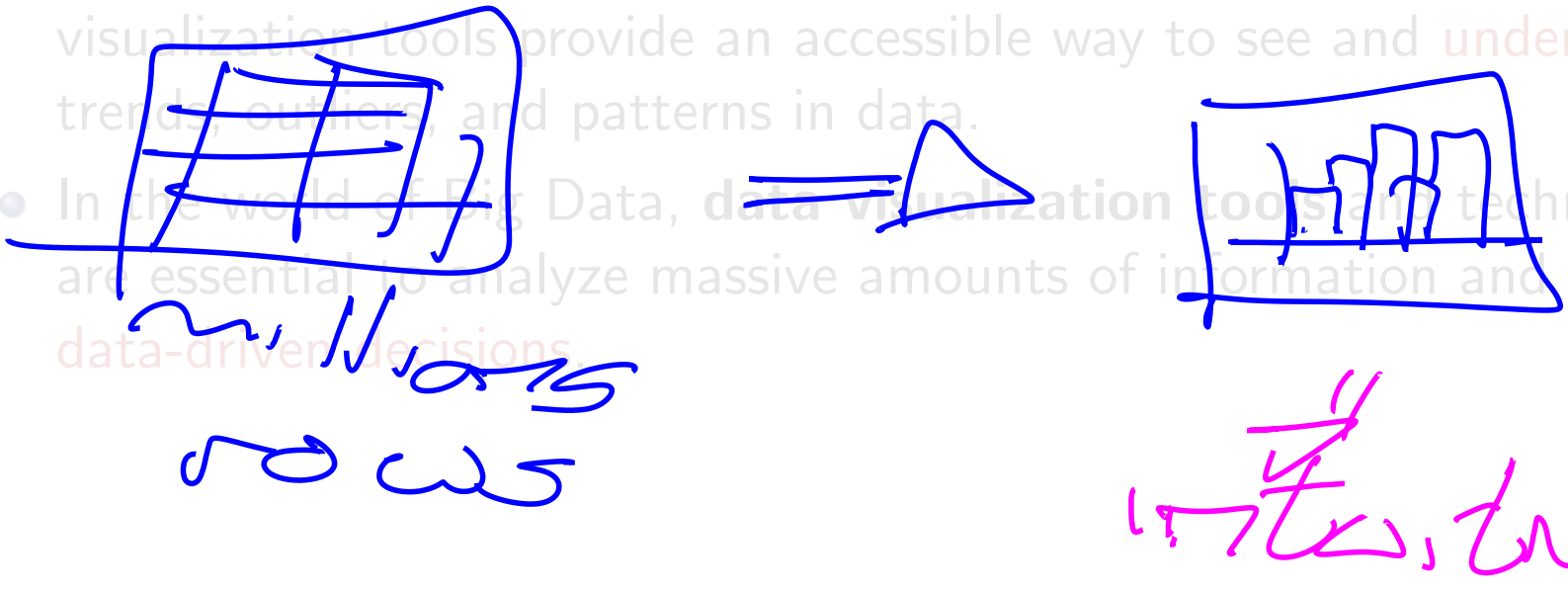
UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

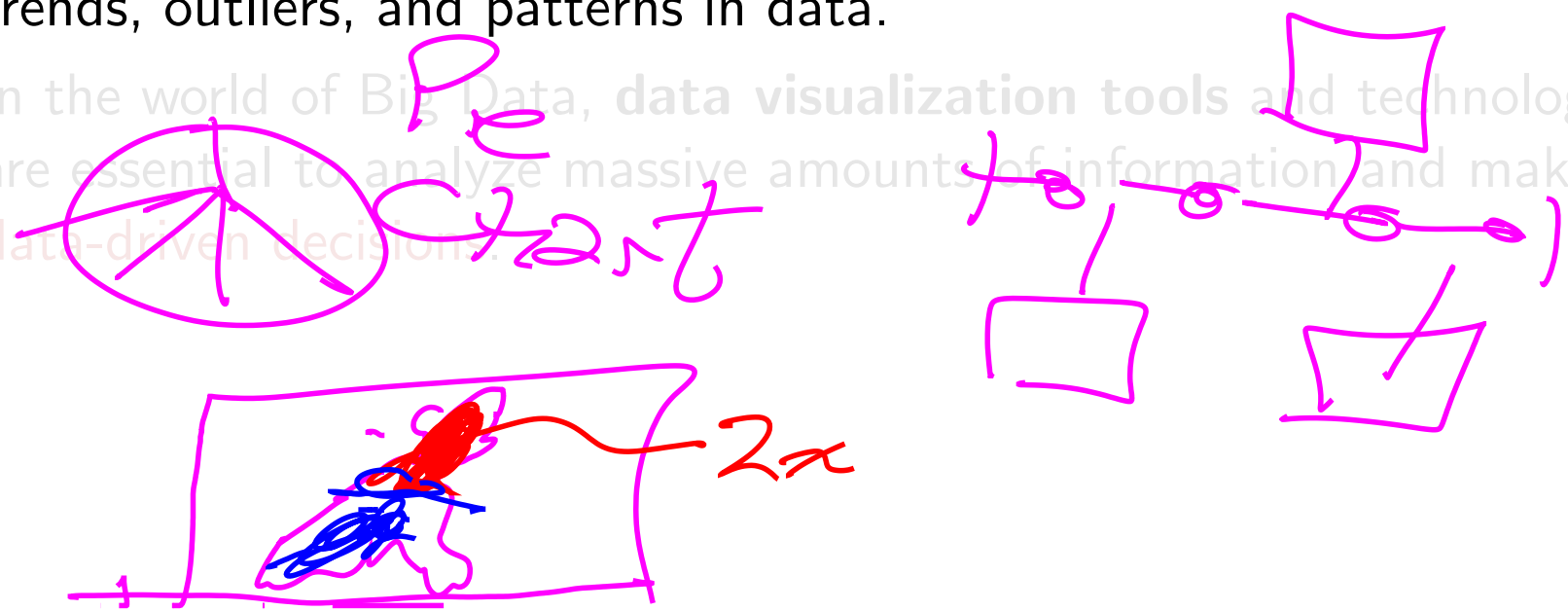
Introduction

- **Data visualization** is the graphical representation of information and data.
- By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, **data visualization tools** and technologies are essential to analyze massive amounts of information and make data-driven decisions.



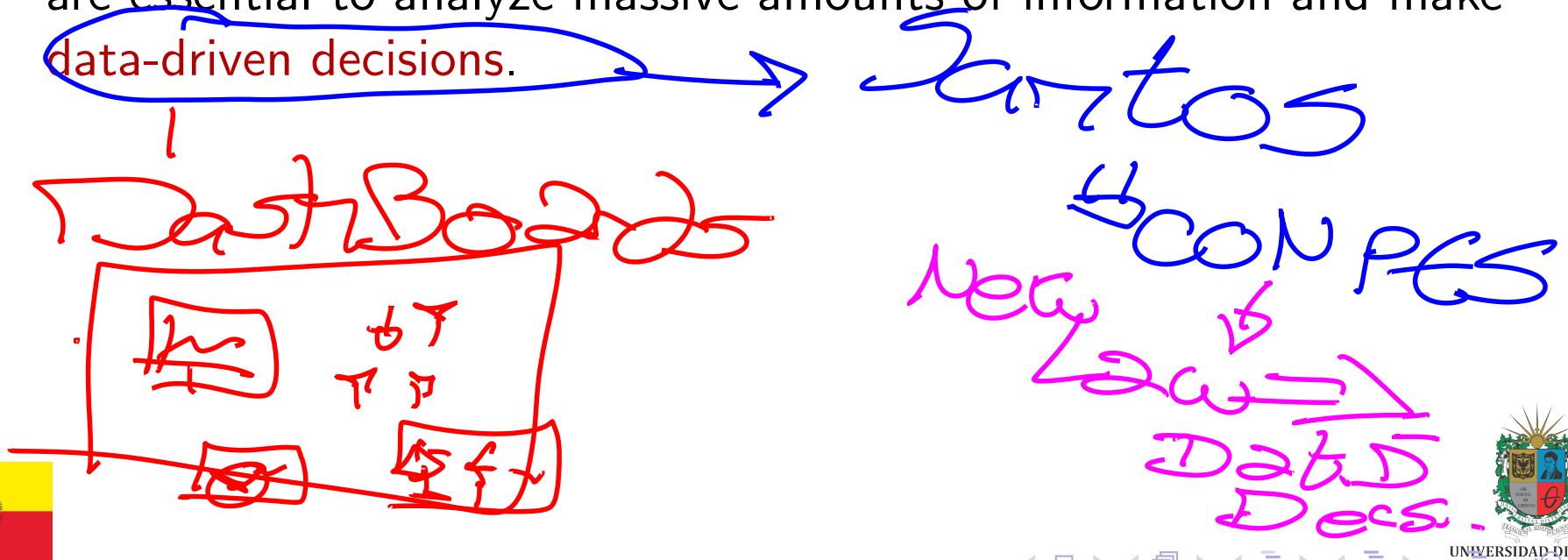
Introduction

- **Data visualization** is the graphical representation of information and data.
- By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, **data visualization tools** and technologies are essential to analyze massive amounts of information and make data-driven decisions.



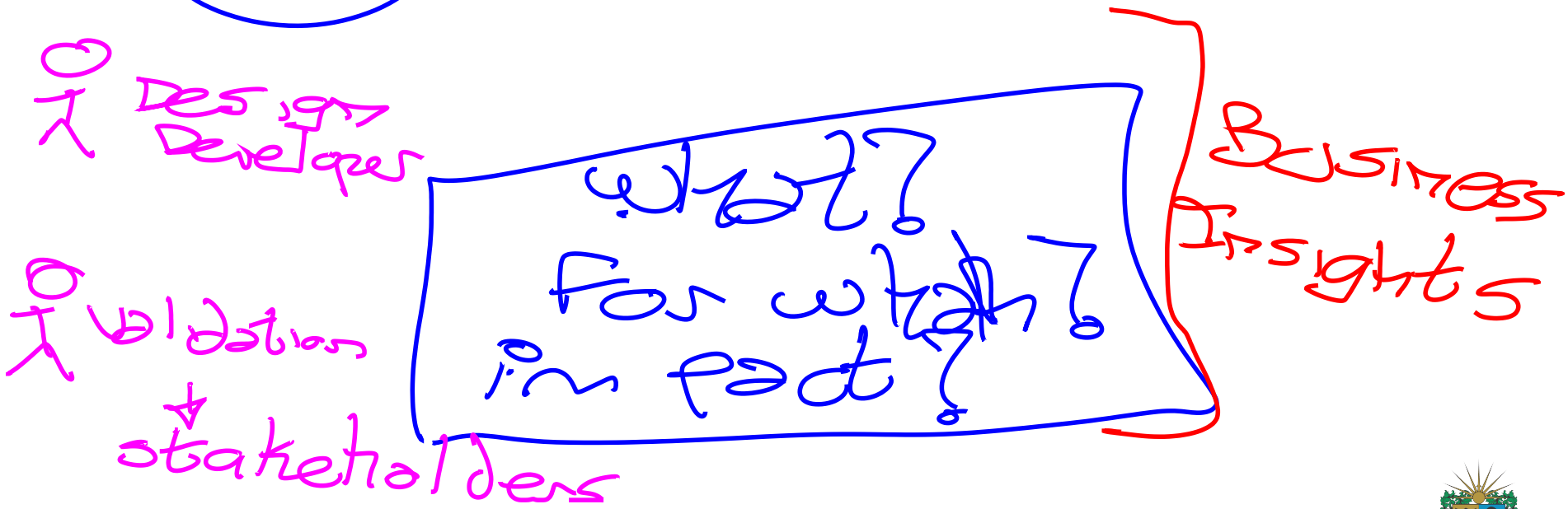
Introduction

- **Data visualization** is the graphical representation of information and data.
- By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- In the world of Big Data, **data visualization tools** and technologies are essential to analyze massive amounts of information and make data-driven decisions.



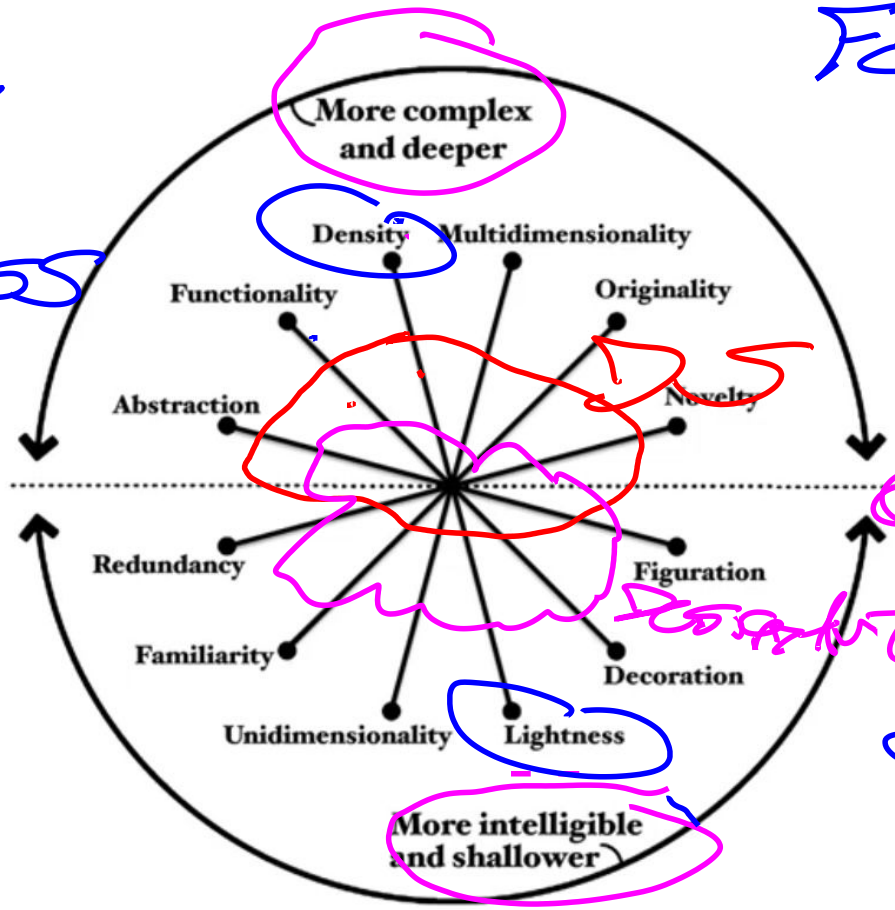
Tools for Thinking About Design

- Design is the process of creating a plan or convention for the construction of an object or a system.
- Design has different connotations in different fields.
- In some cases, the direct construction of an object is also considered to be design.



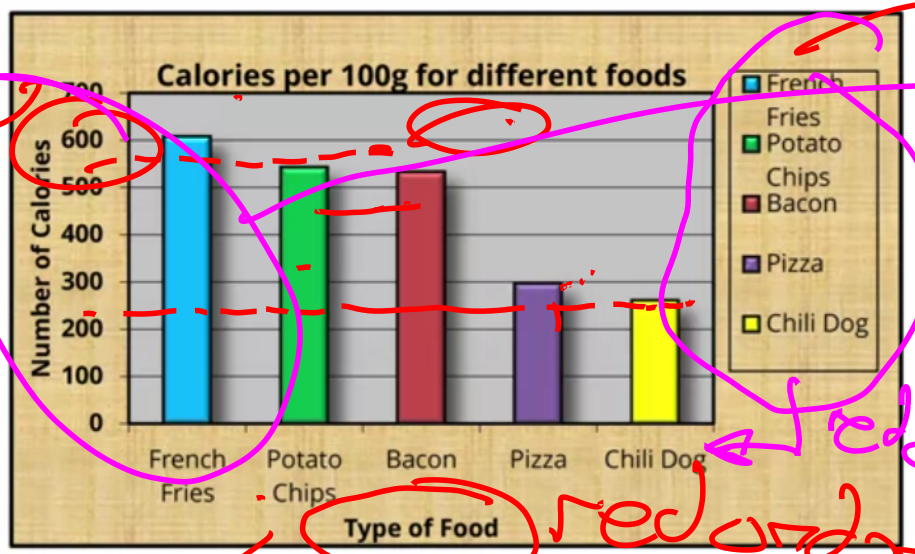
Visualization Wheel

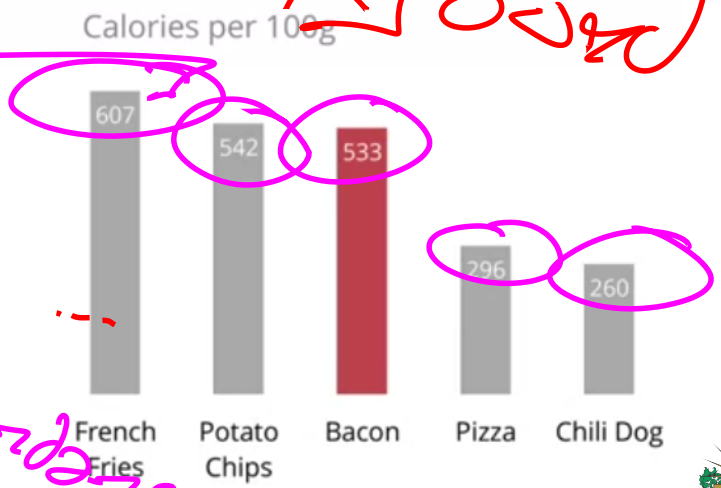
The Visualization Wheel is a tool that helps you to think about the design of your visualization.



Graphical Heuristics

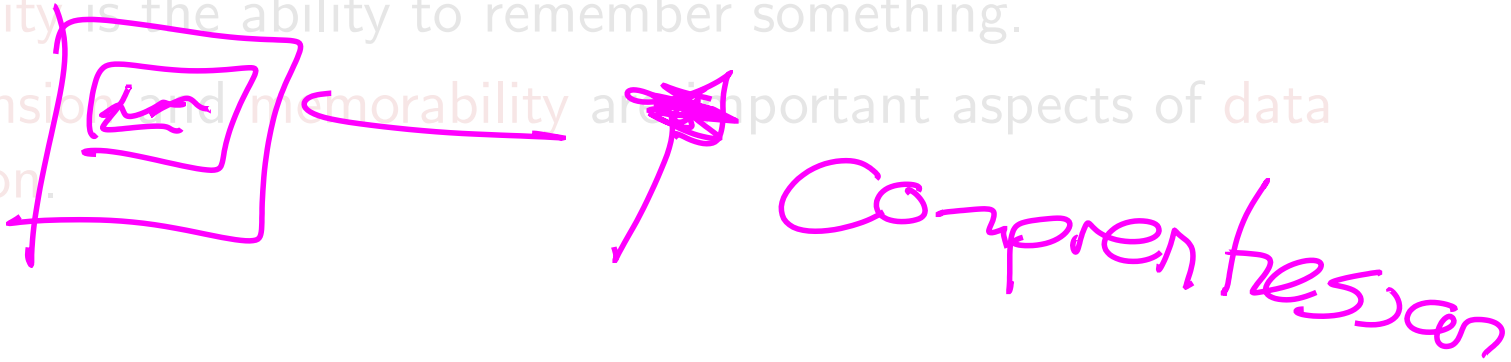
- Graphical heuristics are rules of thumb that help you to design effective visualizations.
- Graphical heuristics are based on the principles of perception and cognition.
- Graphical heuristics consist on pre-attentive processing, gestalt principles, and color theory.





Comprehension and Memorability of Charts

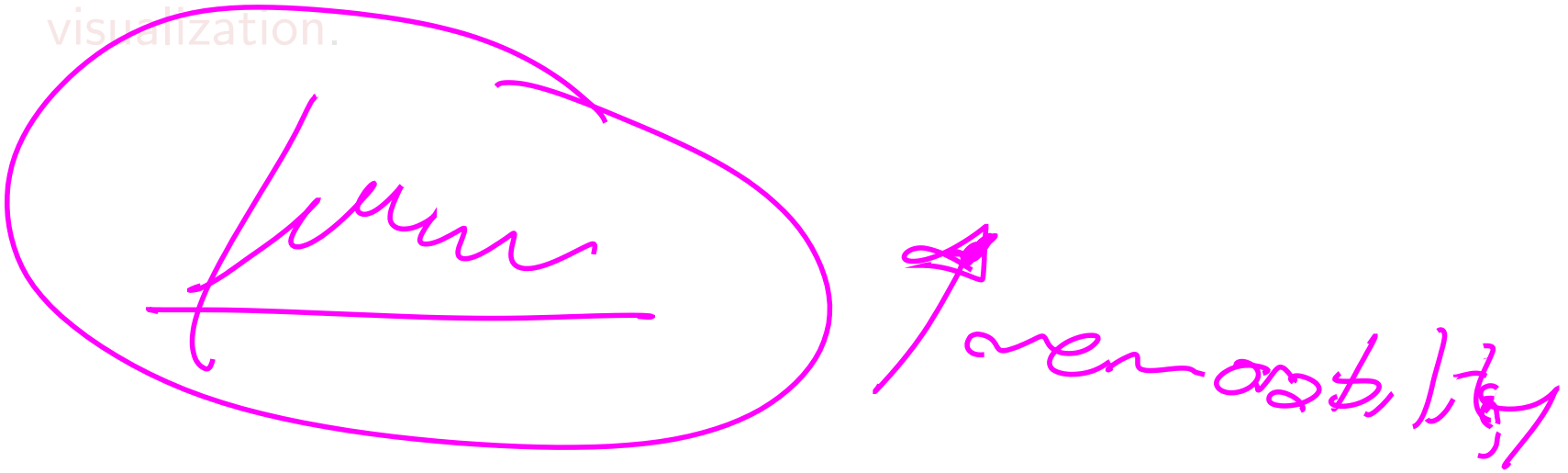
- Comprehension is the ability to understand something.
- Memorability is the ability to remember something.
- Comprehension and memorability are important aspects of data visualization.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Comprehension and Memorability of Charts

- Comprehension is the ability to understand something.
- Memorability is the ability to remember something.
- Comprehension and memorability are important aspects of data visualization.

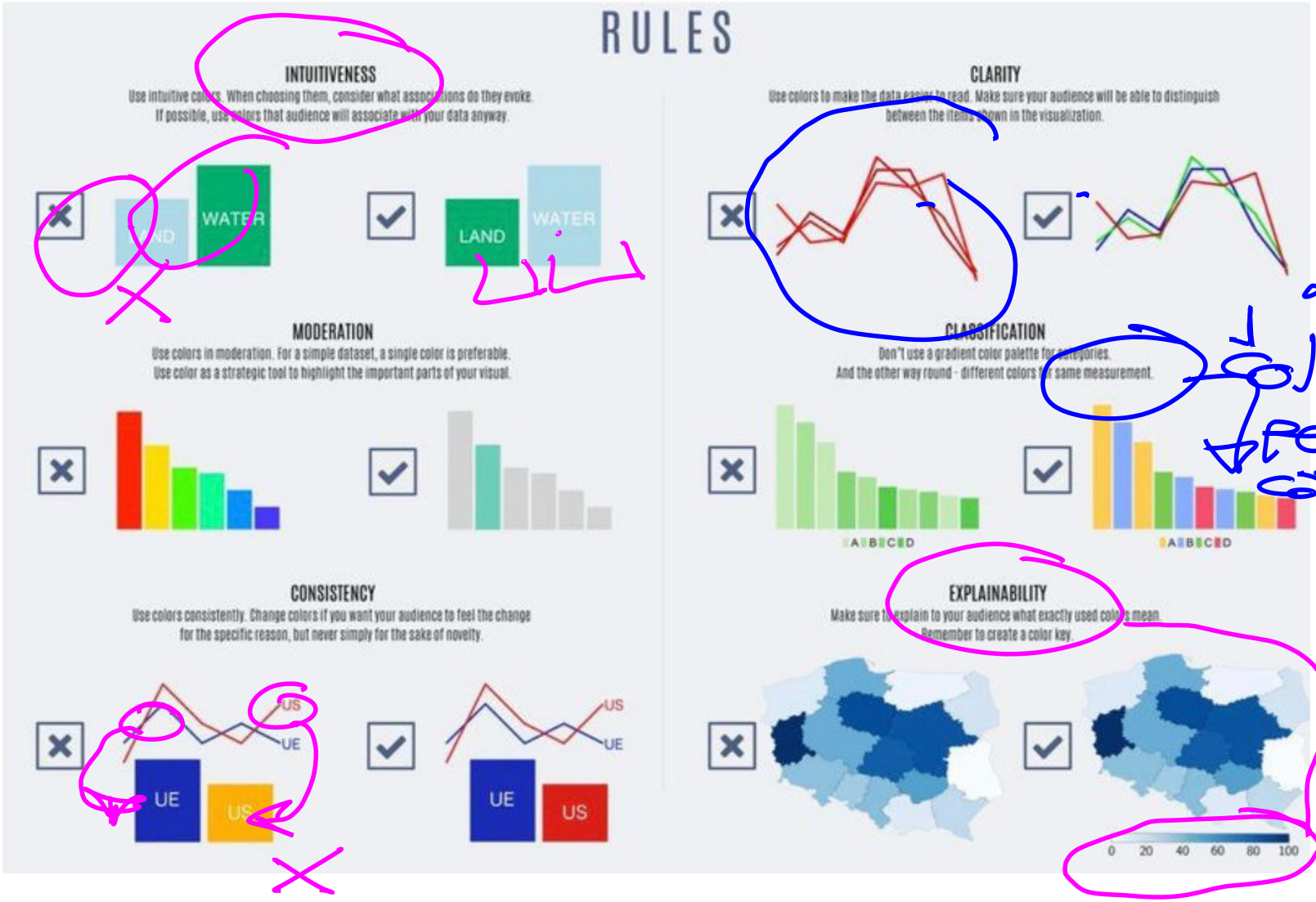


Comprehension and Memorability of Charts

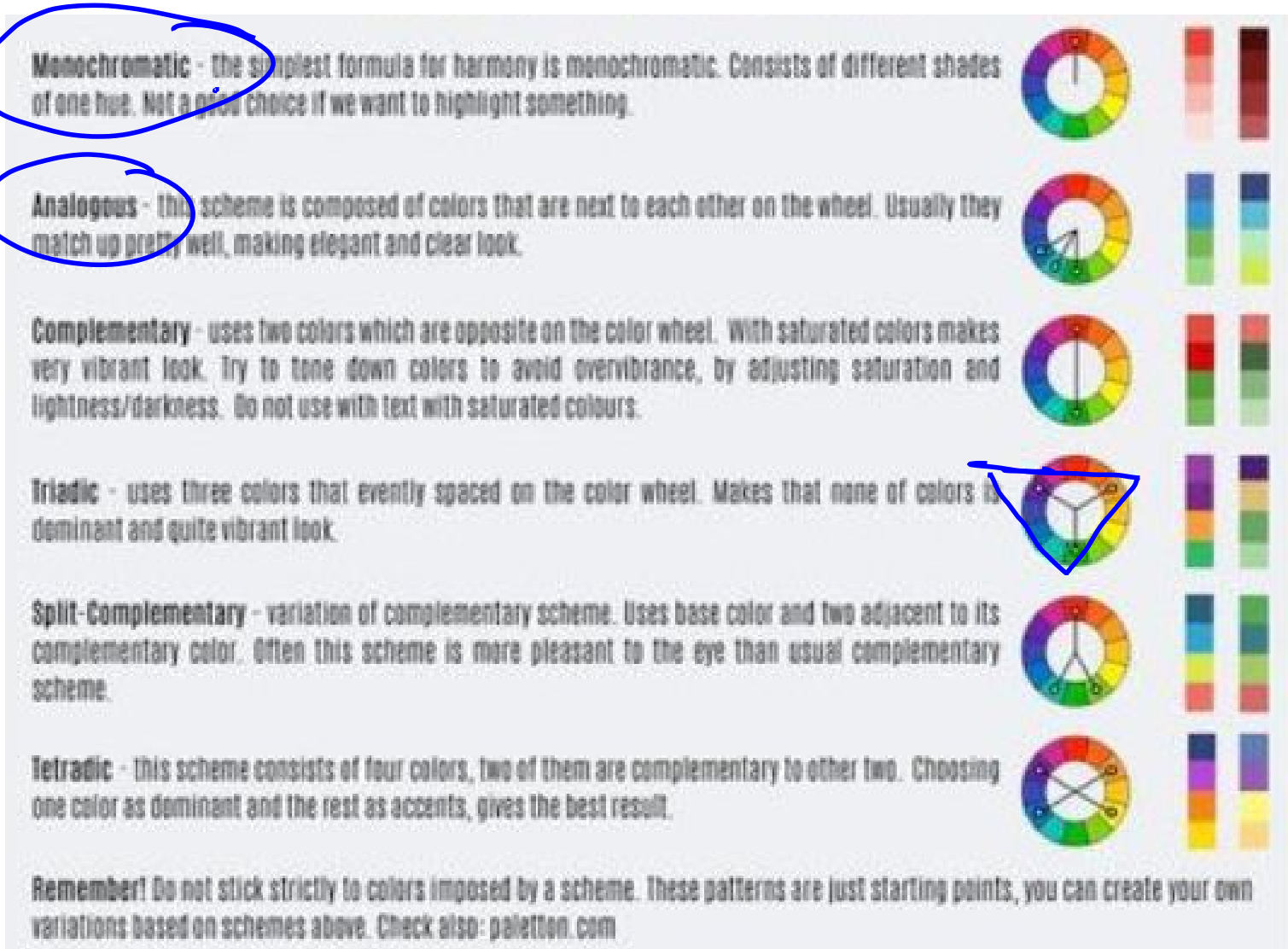
- Comprehension is the ability to understand something.
- Memorability is the ability to remember something.
- Comprehension and memorability are important aspects of data visualization.

1. Time - likelihood
2. Avoid confusions
3. Context
   - one interpretation

Colors in Data Visualization



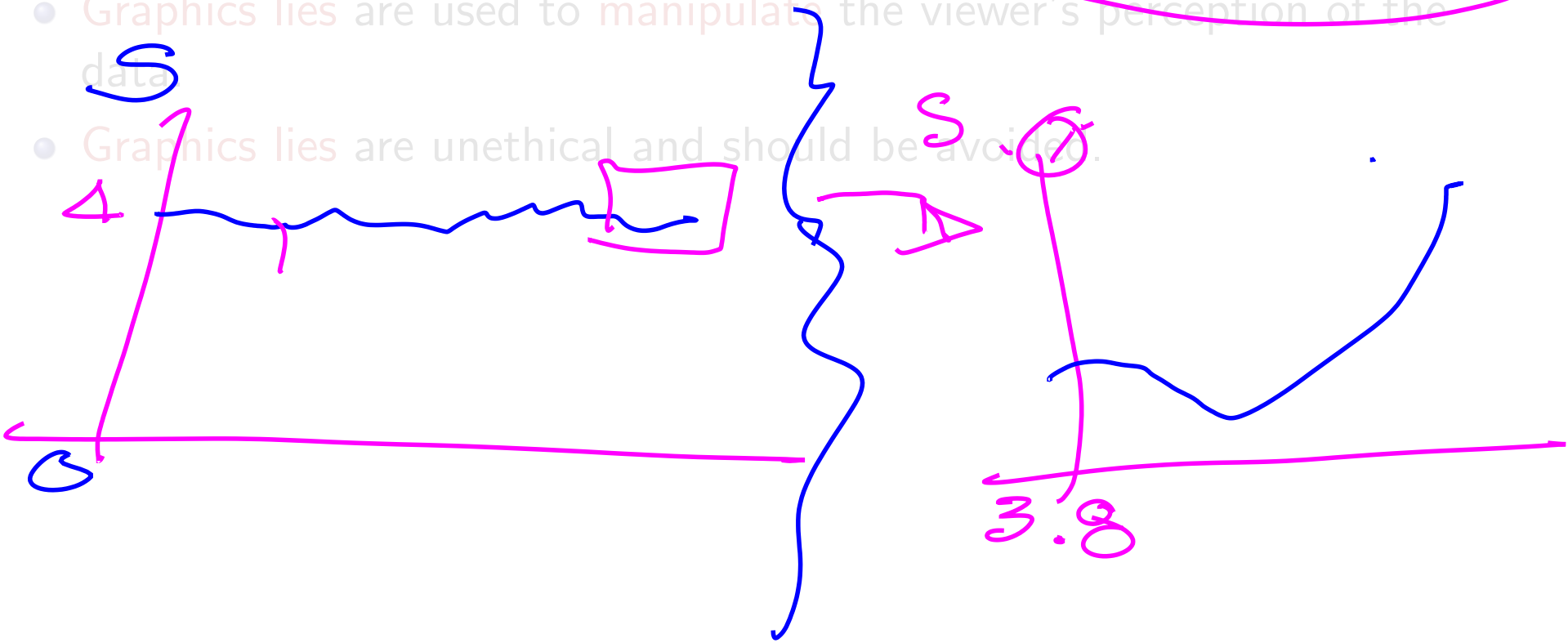
Color Schemas



Color Palletes

Graphics Lies, Misleading Visuals

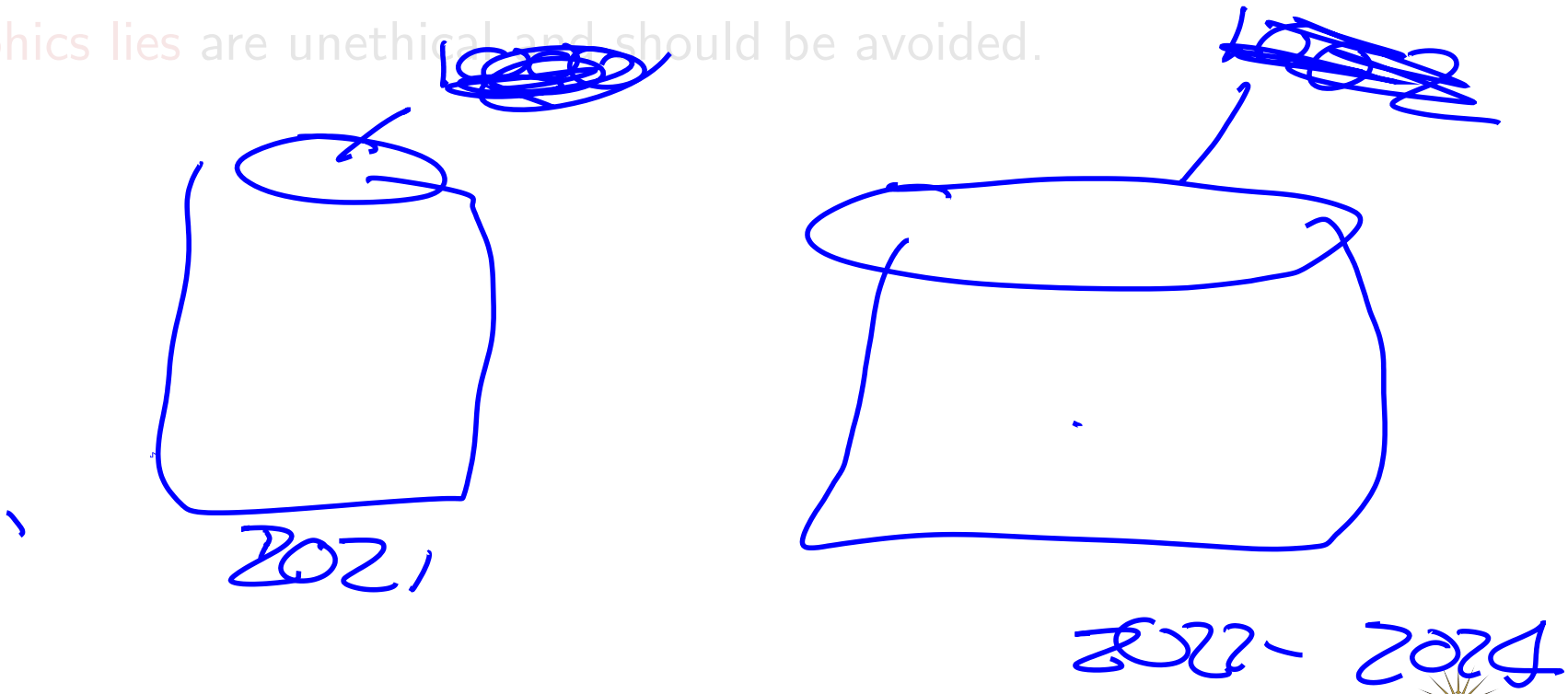
- Graphics lies are visualizations that are misleading or deceptive.
- Graphics lies are used to manipulate the viewer's perception of the data.
- Graphics lies are unethical and should be avoided.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Graphics Lies, Misleading Visuals

- Graphics lies are visualizations that are misleading or deceptive.
- Graphics lies are used to manipulate the viewer's perception of the data.
- Graphics lies are unethical and should be avoided.



Graphics Lies, Misleading Visuals

- Graphics lies are visualizations that are misleading or deceptive.
- Graphics lies are used to manipulate the viewer's perception of the data.
- Graphics lies are unethical and should be avoided.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

What is Matplotlib?

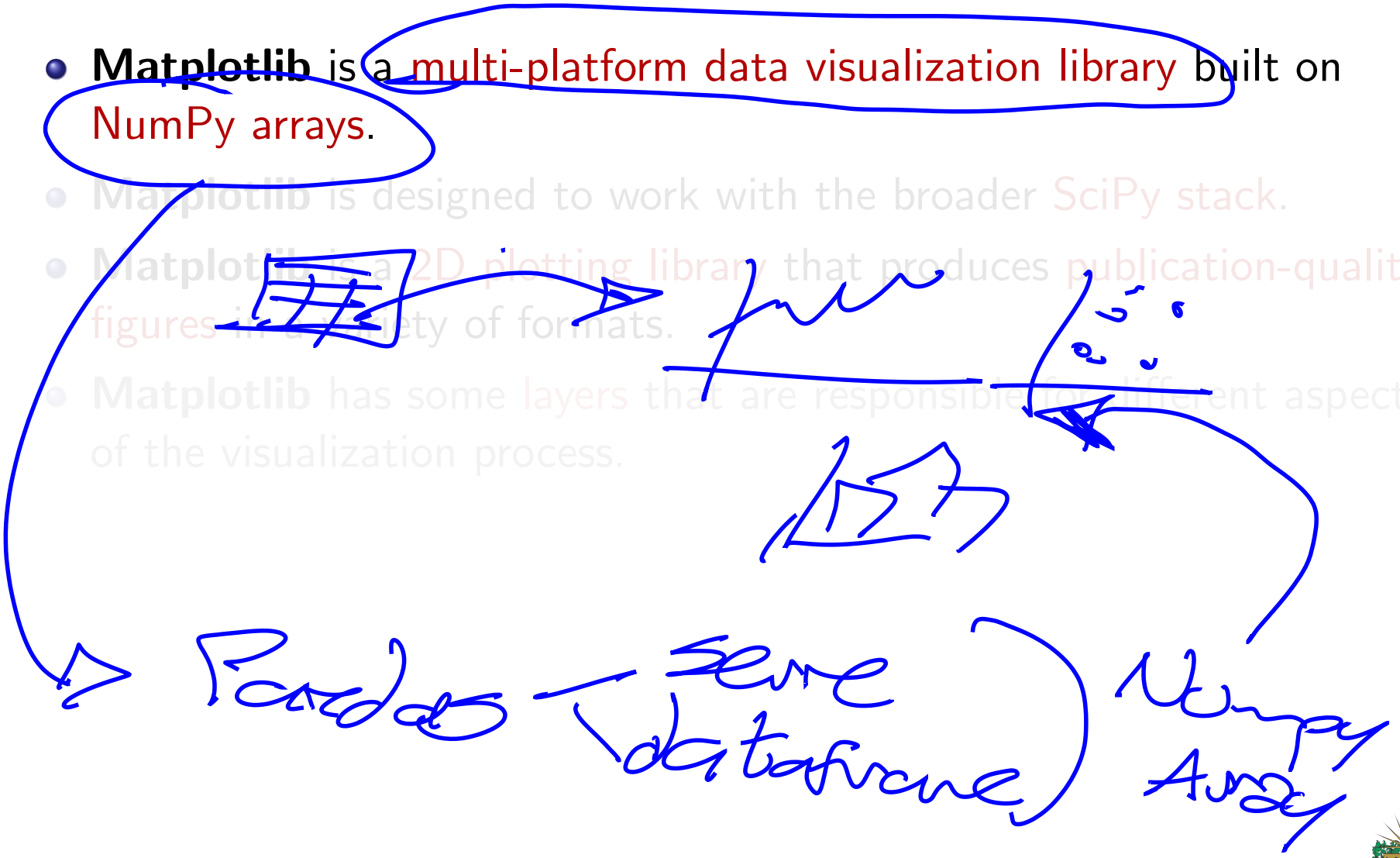
- **Matplotlib** is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- **Matplotlib** makes easy things easy and hard things possible.
- **Matplotlib** can be used in Python scripts, the Python and IPython shell, web application servers, and various graphical user interface toolkits.

MatLab → Canvas objects

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Matplotlib Architecture

- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- **Matplotlib** has some layers that are responsible for different aspects of the visualization process.



Matplotlib Architecture

- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- **Matplotlib** has some layers that are responsible for different aspects of the visualization process.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

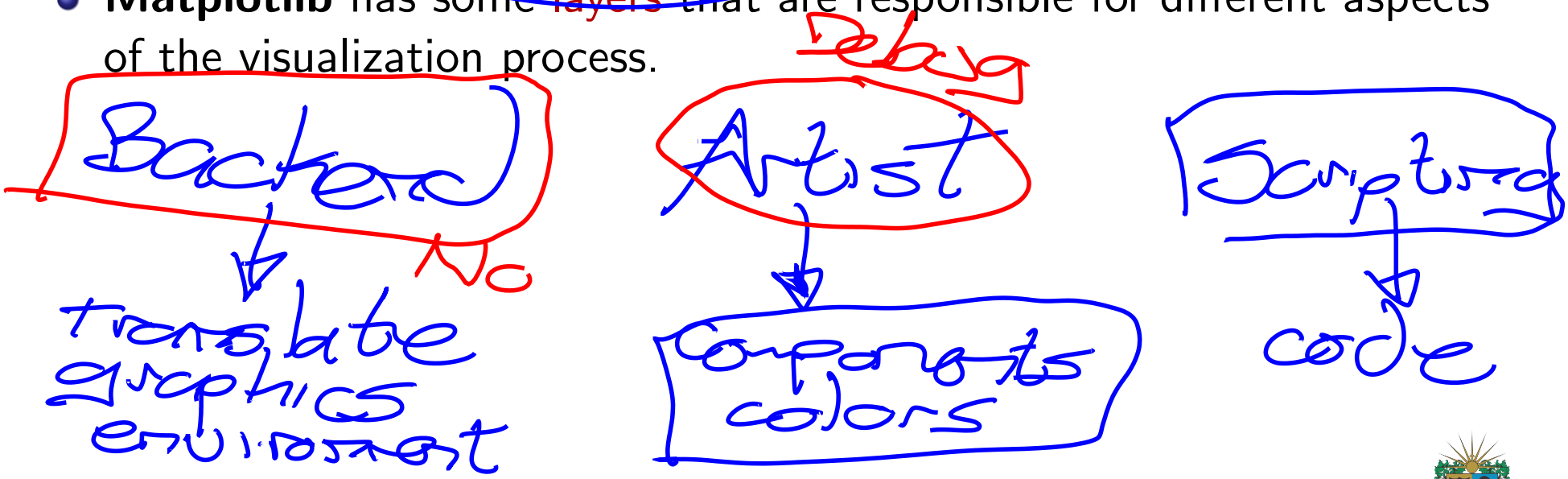
Matplotlib Architecture

- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- **Matplotlib** has some layers that are responsible for different aspects of the visualization process.

PNG → 720 x 1024

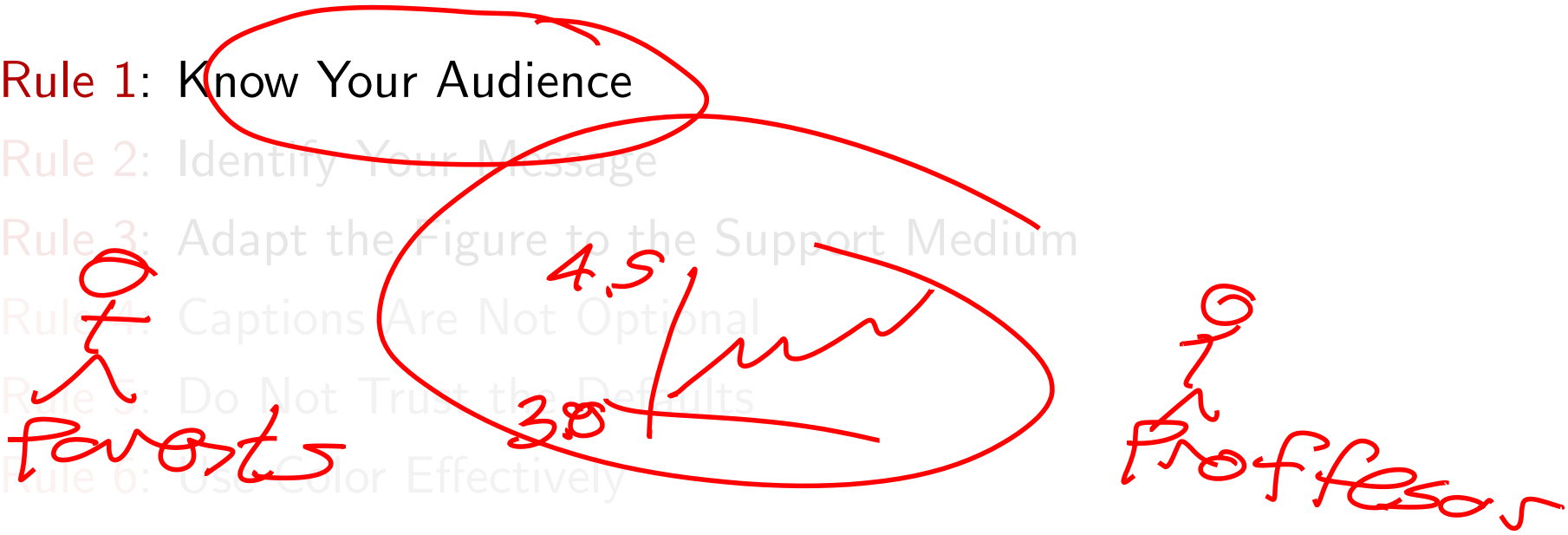
Matplotlib Architecture

- **Matplotlib** is a multi-platform data visualization library built on NumPy arrays.
- **Matplotlib** is designed to work with the broader SciPy stack.
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- **Matplotlib** has some layers that are responsible for different aspects of the visualization process.



Ten Simple Rules for Better Figures

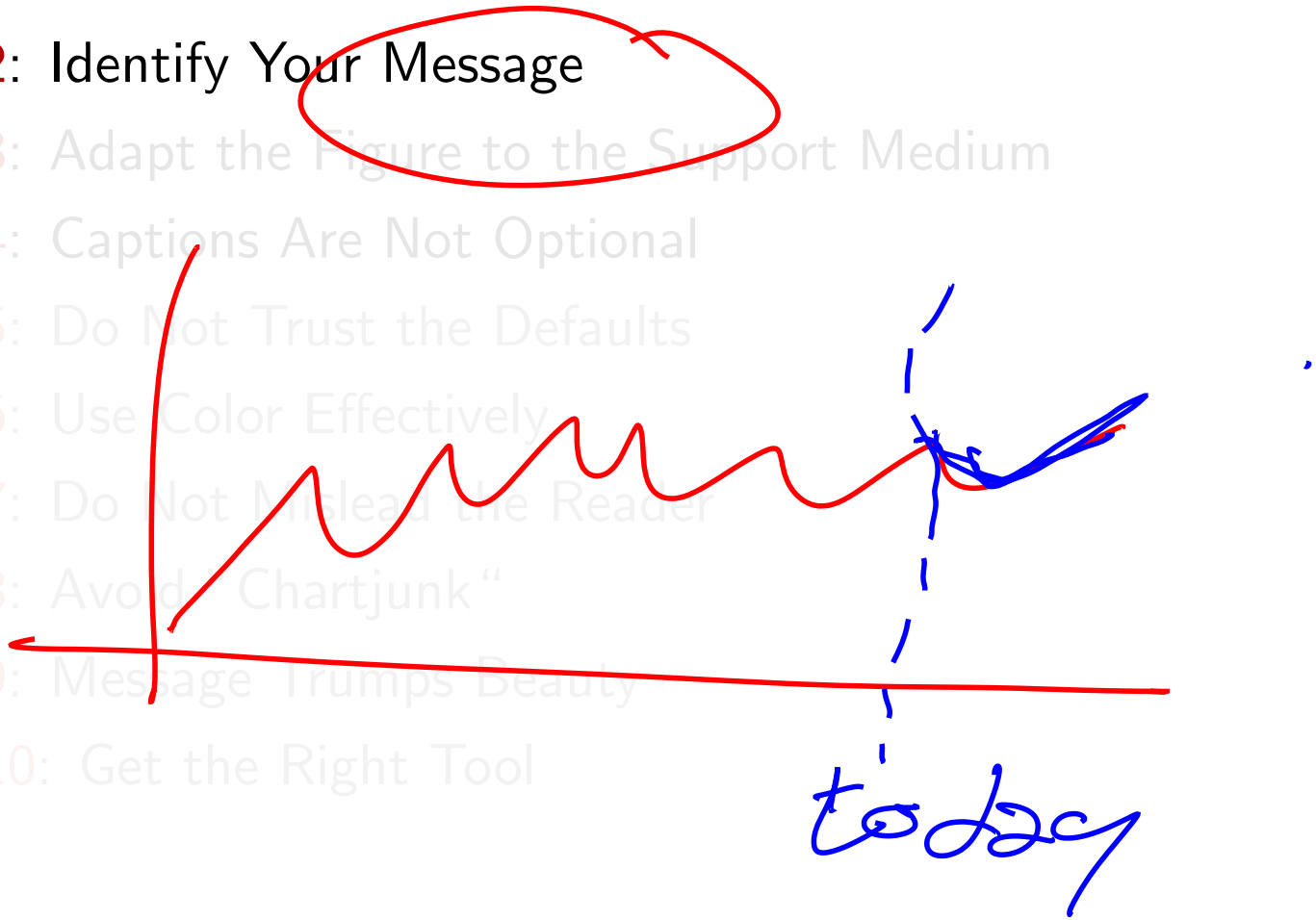
- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ten Simple Rules for Better Figures

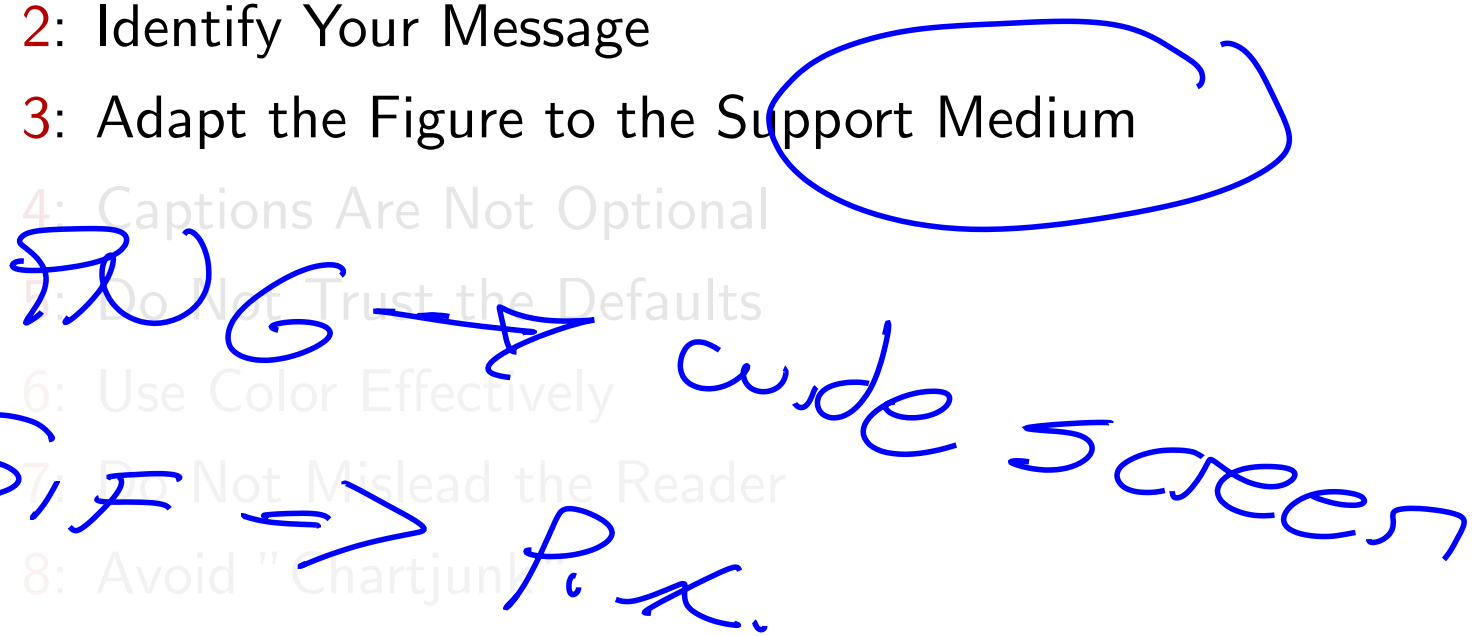
- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

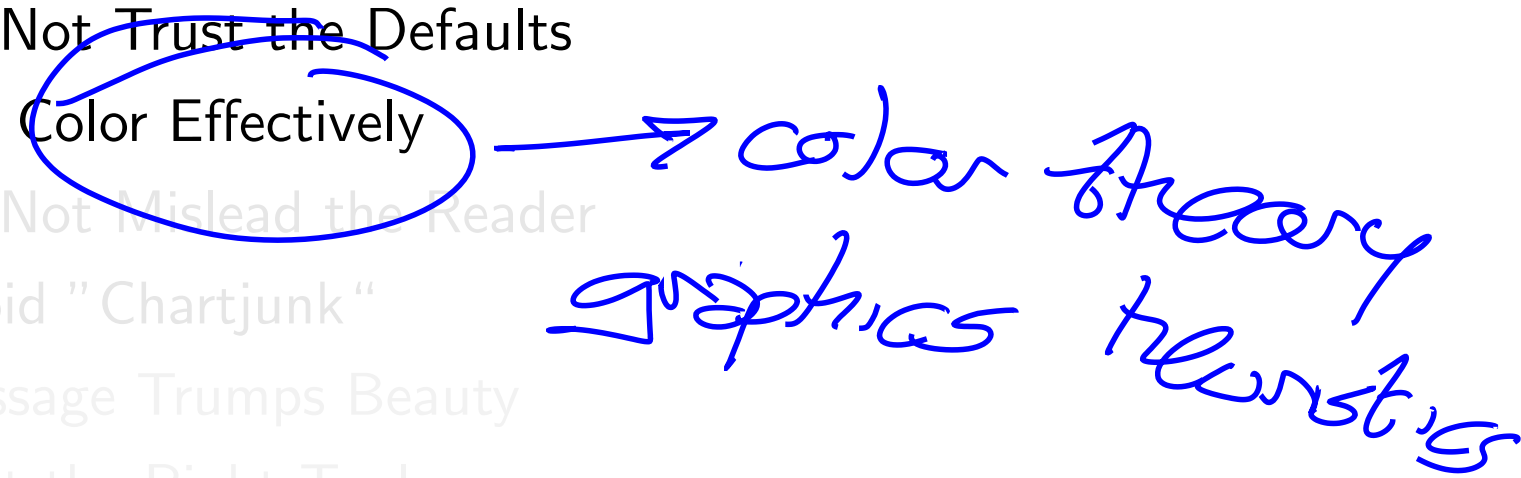
Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

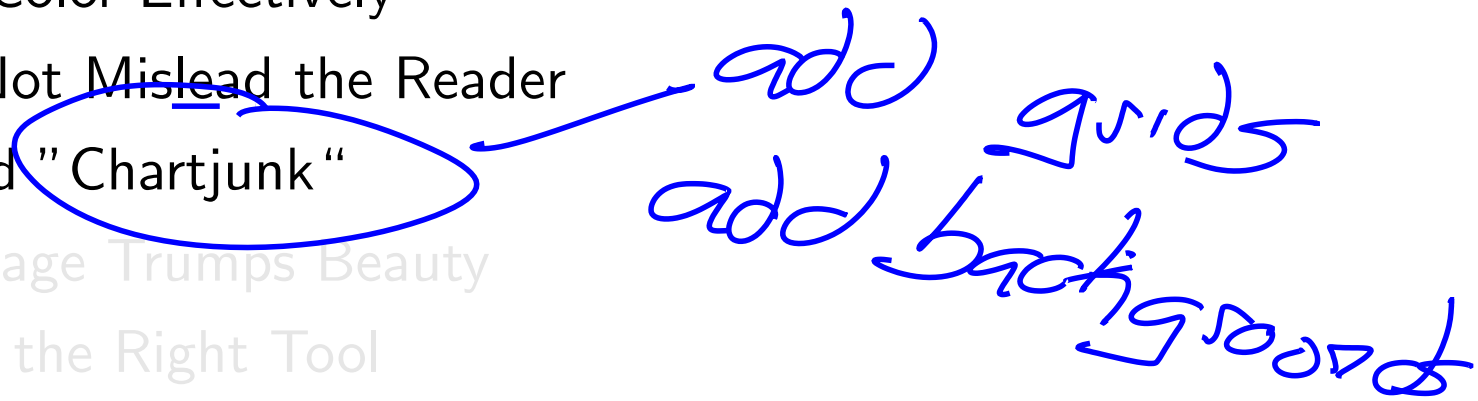
Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

No saturate → simple / clear

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

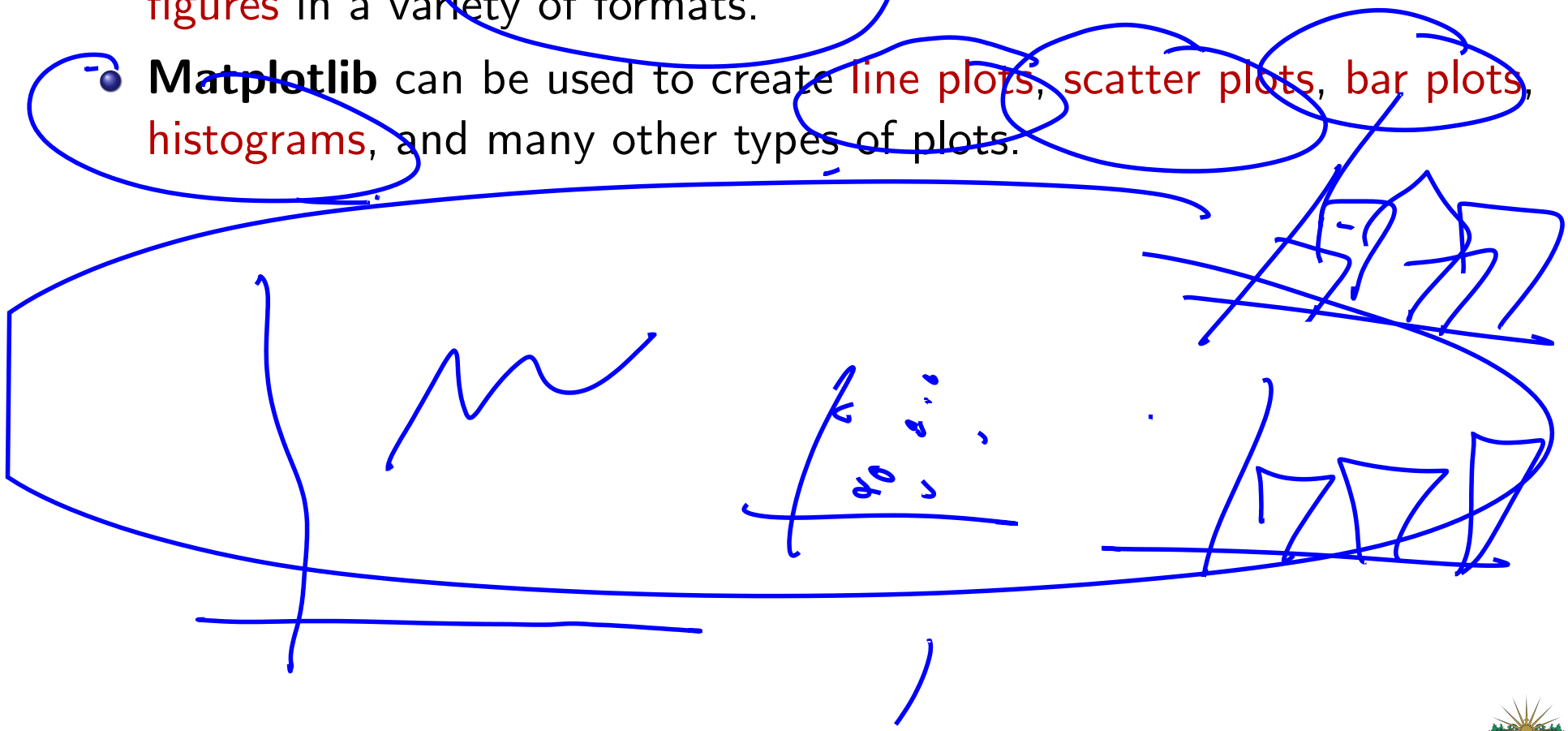
Ten Simple Rules for Better Figures

- Rule 1: Know Your Audience
- Rule 2: Identify Your Message
- Rule 3: Adapt the Figure to the Support Medium
- Rule 4: Captions Are Not Optional
- Rule 5: Do Not Trust the Defaults
- Rule 6: Use Color Effectively
- Rule 7: Do Not Mislead the Reader
- Rule 8: Avoid "Chartjunk"
- Rule 9: Message Trumps Beauty
- Rule 10: Get the Right Tool

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Basic Plotting with Matplotlib

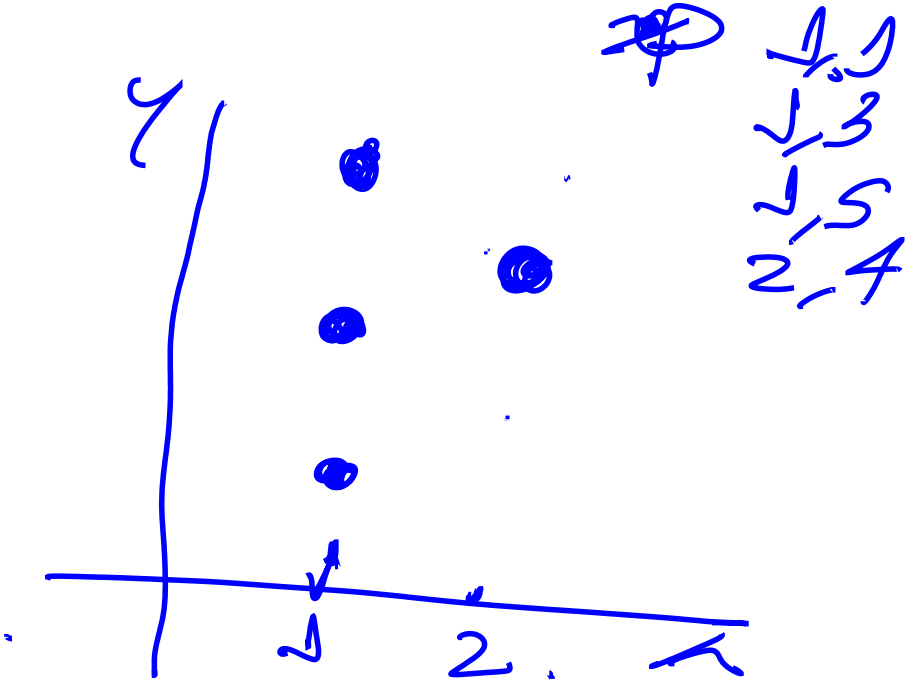
- **Matplotlib** is a 2D plotting library that produces publication-quality figures in a variety of formats.
- **Matplotlib** can be used to create line plots, scatter plots, bar plots, histograms, and many other types of plots.

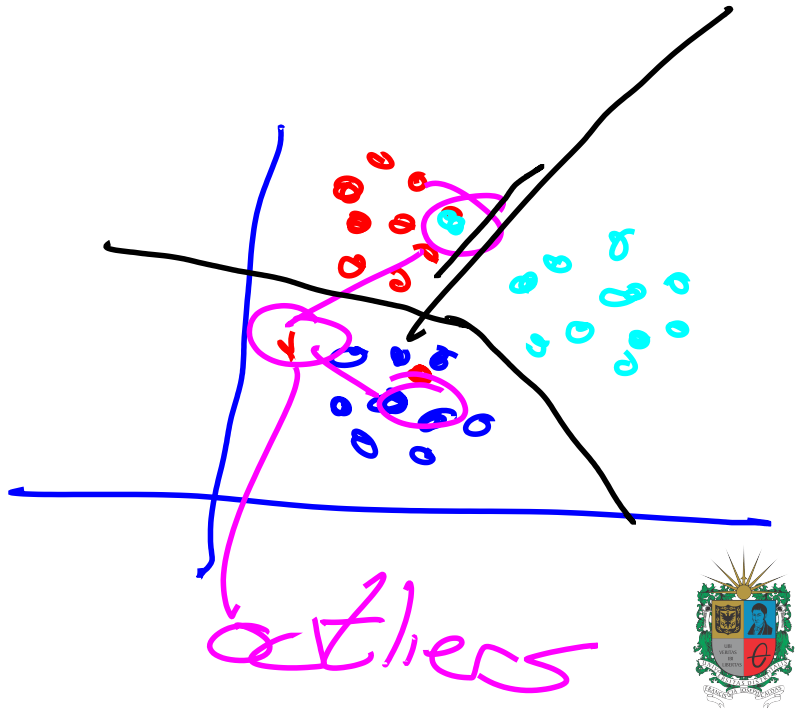


Scatter Plots

Definition

A scatter plot is a type of plot that shows individual data points along the x and y axis. Scatter plots are used to observe **relationships** between variables.

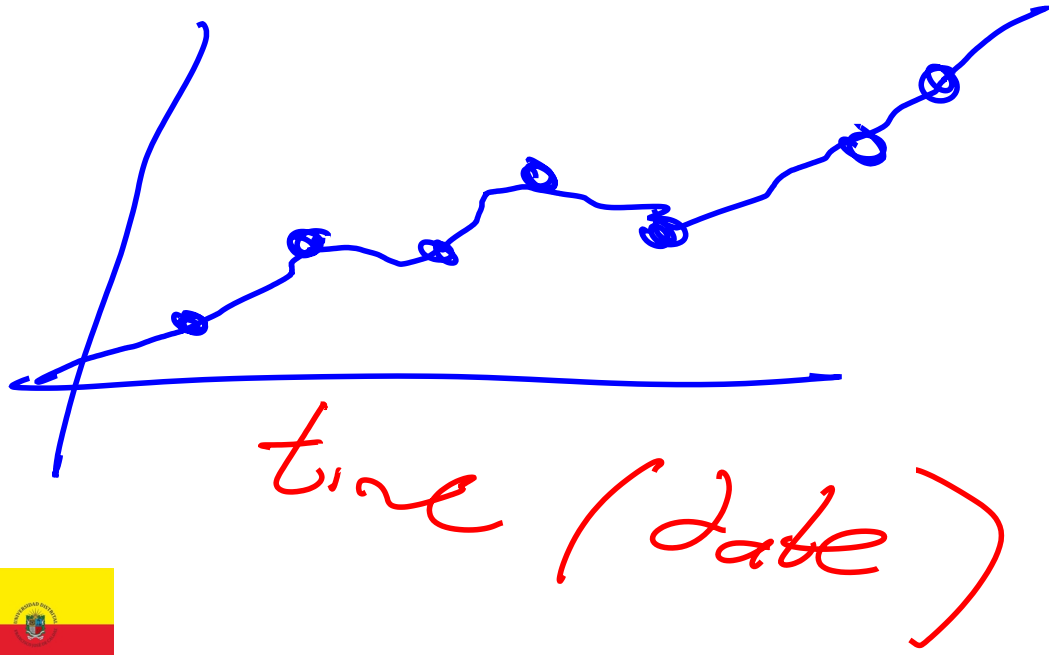




Line Plots

Definition

A line plot is a type of plot that displays information as a series of data points called markers connected by straight line segments. Line plots are used to observe the **trend** in data over intervals of time.



Bar Plots

Definition

A bar plot is a type of plot that presents categorical data with rectangular bars with lengths proportional to the values that they represent. Bar plots are used to compare **quantitative** data across different categories.

Subplots

- **Subplots** are groups of smaller axes that can exist together within a single figure.
- **Subplots** are useful when you want to show multiple plots in the same figure.
- **Subplots** are created using the `plt.subplots()` function.

Histograms

Definition

A histogram is a type of plot that displays the distribution of a dataset. Histograms are used to show the **frequency** of values in a dataset.

Box Plots

Definition

A box plot is a type of plot that displays the distribution of a dataset. Box plots are used to show the **spread** and **central tendency** of values in a dataset.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

HeatMaps

Definition

A heatmap is a type of plot that displays the intensity of data at the intersection of two variables. Heatmaps are used to show the **correlation** between variables in a dataset.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Animation

- Animation is the process of creating a sequence of images that change over time.
- Animation is created using the `FuncAnimation` class from the `matplotlib.animation` module.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Widgets

- Widgets can be used to create interactive visualizations in a Jupyter notebook.
- Widgets can be used to create sliders, buttons, and other interactive elements.
- Widgets are created using the `ipywidgets` library.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Widgets

- Widgets can be used to create interactive visualizations in a Jupyter notebook.
- Widgets can be used to create sliders, buttons, and other interactive elements.
- Widgets are created using the `ipywidgets` library.

Widgets

- Widgets can be used to create interactive visualizations in a Jupyter notebook.
- Widgets can be used to create sliders, buttons, and other interactive elements.
- Widgets are created using the `ipywidgets` library.

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

What is Seaborn?

- **Seaborn** is a Python data visualization library based on Matplotlib.
- **Seaborn** provides a high-level interface for creating attractive and informative statistical graphics.
- **Seaborn** is built on top of Matplotlib and closely integrated with the data structures from **pandas**.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Spurius Correlations

Definition

A spurious correlation is a statistical relationship between two variables that is not causally related. Spurious correlations are often the result of confounding variables.

Mapping and Geographic

Definition

Mapping is the process of creating a visual representation of an area.
Mapping is used to show the **geographic** distribution of data.

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Thanks!

Questions?

Repo: *https://github.com/EngAndres/ud-public/tree/main/courses/data-science-introduction*

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS