Data Analysis and Visualization

Topics to cover…

- Creating interesting stories with data
- Perception and presentation methods
- Best practices for visualization
- Visualization tools in Python
- Numerical computing and Interactive visualization

Creating interesting stories with data

Visualization requires planning

- Acquire or gather data from an external source, a website, or from a file on a disk
- Parse and filter data using programming methods to parse, clean, and reduce the data
- Analyze and refine to remove noise and unnecessary dimensions and find patterns
- Represent and interact to present the data in ways that are more accessible and understandable

Creating interesting stories with data

- Data visualization regularly promotes its ability to reveal stories with data
- Understand that data would be understood and remembered better if presented in the right way
- Reader-driven narratives and author-driven narratives

Author-driven narratives

- Has data and visualization that are chosen by the author and presented to the public reader
- Has a strict linear path through the visualization, relies on messaging, and has no interactivity
- Example: How the recession reshaped the economy in 255 charts (NY Times) https://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html?_r=0

Reader-driven narratives

- Provides tools and methods for the reader to play with the data, which gives the reader more flexibility and choices to analyze and understand the visualization
- Has no prescribed ordering of images, no messaging, and has a high degree of interactivity
- Example: Gapminder (a collection of over 600 data indicators in international economy, environment, health, technology, etc) https://www.gapminder.org/tools/?from=world

Example: Climate Change Visualization — Author-driven narratives

- Start with a compelling opening statement about the urgency of climate change.
- Provide visualization of data:
  - Line Chart: Show historical global temperature increases over the last century.
  - Bar Chart: Present data on CO2 emissions by country, highlighting the largest contributors.
- Highlight key findings
- End with a strong call to action

Example: Climate Change Visualization — Reader-driven narratives

- Pose an open-ended question to engage the audience.
- Provide interactive maps and charts that allow readers to explore the data:
  - Interactive Map: Users can hover over regions to see data on temperature changes, sea-level rise, and extreme weather events.
  - Dynamic Line Graph: Let users select different countries to compare their CO2 emissions and temperature changes.
- Invite readers to share their insights or take action based on their findings

Perception and presentation methods

Visualization systems

- Most visualization systems are designed so that human and computer can cooperate, each performing the following tasks:
  - Visually representing data to enhance data analysis
  - Visually displaying models, interpretations of data, ideas, hypotheses, and insight
  - Helping users to improve their models by finding either supporting or contradictory evidence for their hypotheses
  - Helping users to organize and share their ideas

The Gestalt principles of perception

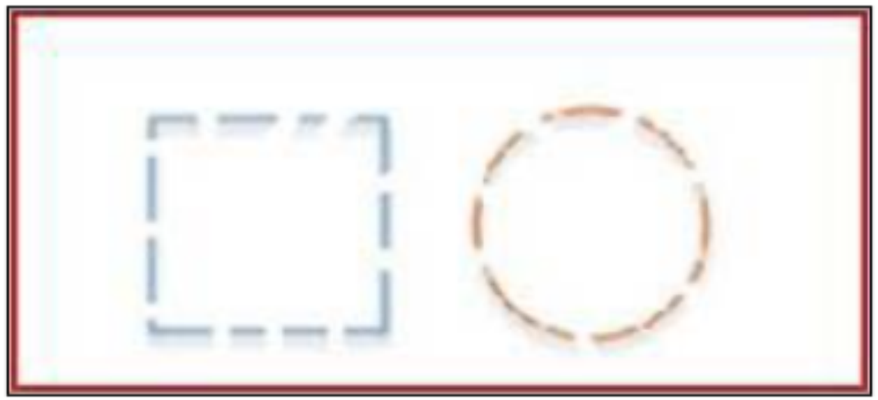
- “Gestalt” is German for “unified whole”. German psychologists Max Wertheimer, Kurt Koffka, and Wolfgang Kohler created the Gestalt Principles in the 1920s.
  - They wanted to understand how people make sense of the confusing things they see and hear.
  - They identified a set of laws that address the natural compulsion to find order in disorder.
  - According to this, the mind “informs” what the eye sees by perceiving a series of individual elements as a whole.

The Gestalt principles of perception

- Closure
- Continuity
- Figure / ground
- Proximity
- Similarity
- Enclosure
- Connection
- Common fate
- Emergence
- Multistability
- Invariance
- Pragnanz

Closure

- Also known as Reification.
- Even if a part of the border of a shape is missing, we still tend to see the shape as completely enclosed by the border and ignore the gaps



Continuity

- We group elements that seem to follow a continuous path in a particular direction.
- The human eye follows the paths, lines, and curves of a design and prefers to see a continuous flow of visual elements rather than separated objects. The human eye continues to follow the path even if an obstacle hides it or its flow is “broken” by interlinking or bisecting visual elements.

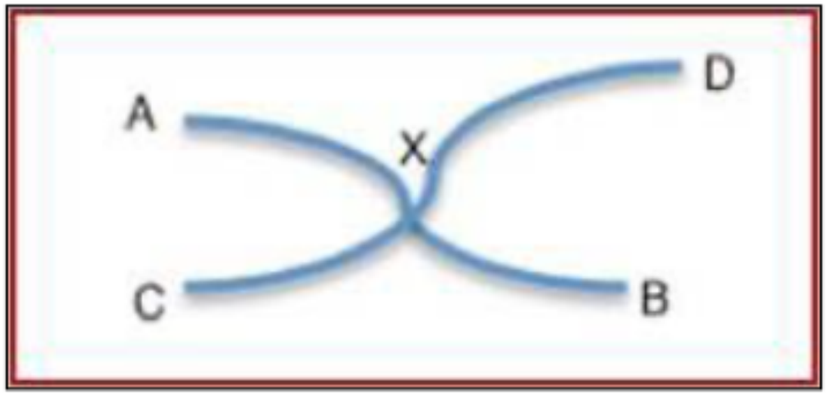


Figure / Ground

- People instinctively perceive objects as either being in the foreground or the background.
- Either stand out prominently in the front (the figure) or recede into the back (the ground).

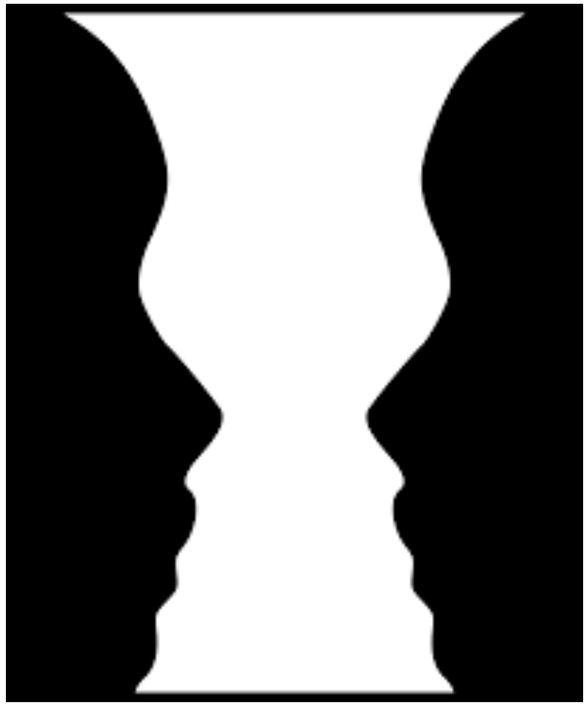
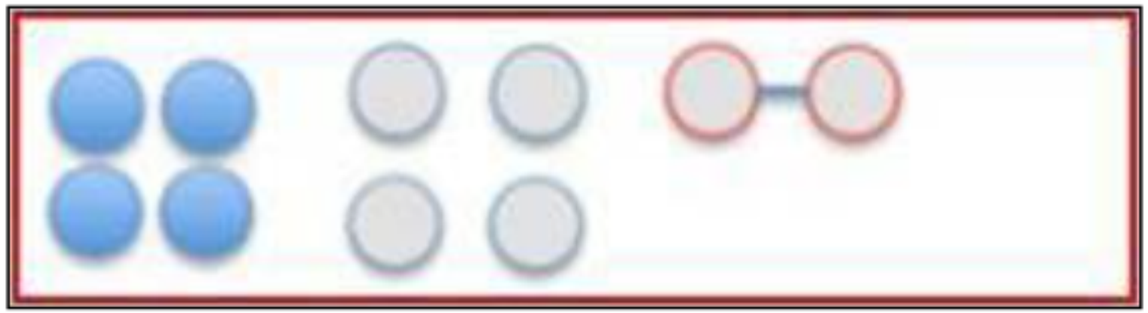


image source: https://en.wikipedia.org/wiki/Figure–ground_(perception)

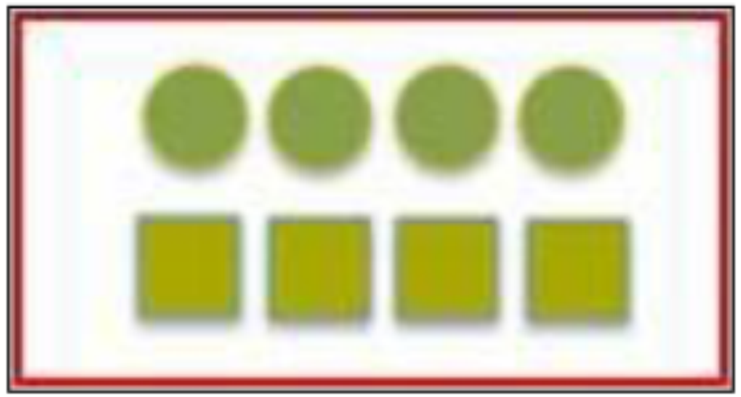
Proximity

- Objects that are close together or connected to each other are perceived as a group, reducing the need to process smaller objects separately



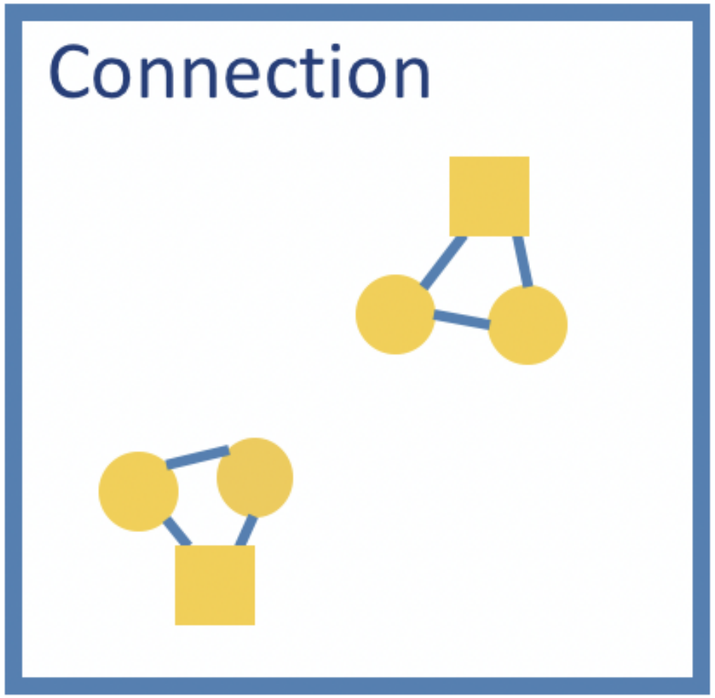
Similarity

- Objects that share similar attributes, color, or shape are perceived as a group



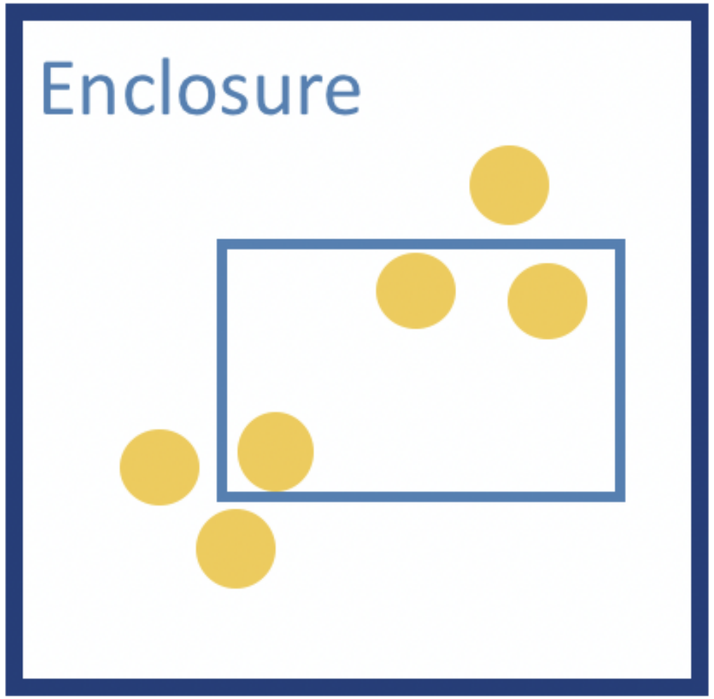
Connection

- Our tendency to perceive objects that are physically connected are part of the same group, whether or not that is true or relevant to the data at hand.
- For example, our eyes connect the group of three shapes in the lower left together and the three shapes in the upper right together, before connecting the dots and squares together.
- The connection principle often is stronger than proximity or similarity principles, but not as strong as enclosure.



Enclosure

- Reflects our tendency to perceive objects physically enclosed together in a ways that seems to create a boundary around that which is related or belonging to part of the same group, whether or not that is true or relevant to the data at hand.
- For example, our eyes see the yellow dots within the blue rectangle as having something in common with one another as they are all “in the box” together.
- We attribute meaning to the “box” whether that is accurate or not given what we are looking at.



Common fate

- When both the principles of proximity and similarity are in place, a movement takes place. Then they appear to change grouping

Emergence

- The principle of emergence is key to Gestalt thinking.
- We naturally perceive our surroundings without needing to analyze every detail, which is crucial for survival. — If we took too long to understand our environment, we might have been in danger from wild animals!

Gestalt Rule: **Emergence**

Interaction Design Foundation
interaction-design.org

https://www.interaction-design.org/literature/topics/gestalt-principles?srsltid=AfmBOooWFp_gqZGiGyOR1nSzPoSCvp0AuGlgX7Pz2xpTKf1D8d0jXxES

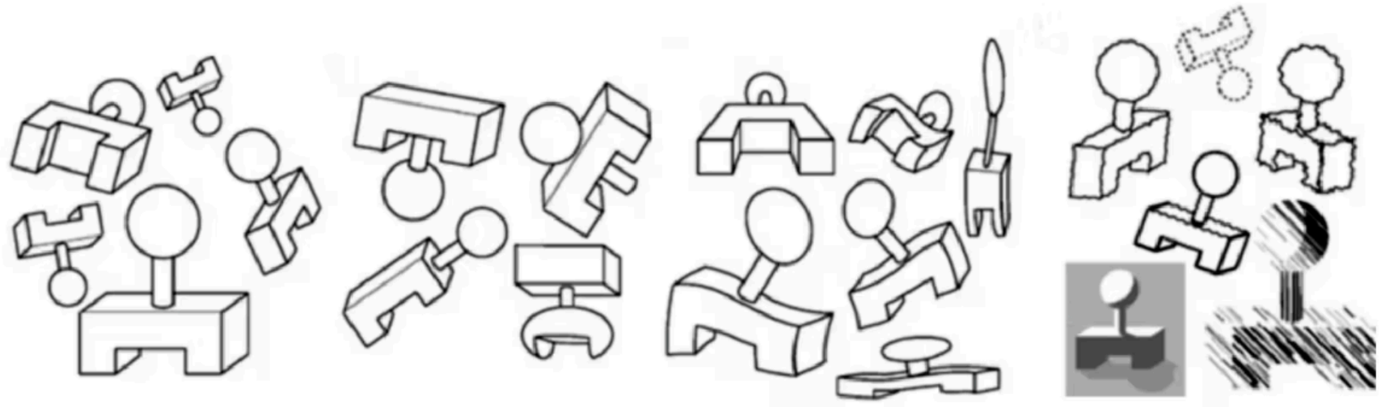
Multistability

- When images are ambiguous and present two or more meaningful interpretations, we experience the sensation of switching between them.
- We cannot see the multiple versions simultaneously. This switching sensation is called multistability.

https://www.interaction-design.org/literature/topics/gestalt-principles?srsltid=AfmBOooWFp_gqZGiGyOR1nSzPoSCvp0AuGlgX7Pz2xpTKf1D8d0jXxES

Invariance

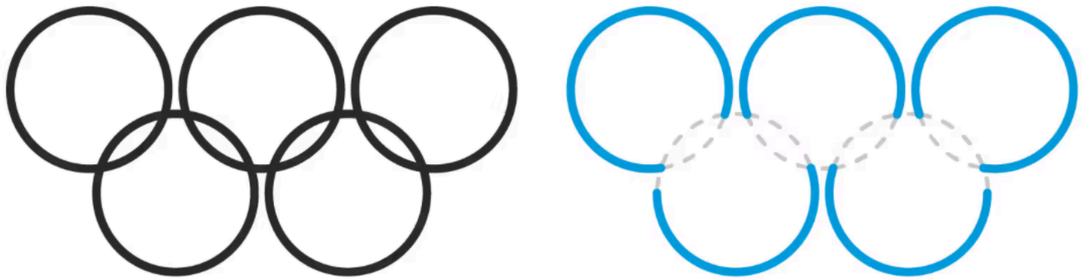
- Explains how we perceive basic shapes as identical despite various transformations. These transformations include rotation, movement, size alteration, stretching, different lighting conditions, and variations in parts.



https://www.interaction-design.org/literature/topics/gestalt-principles?srsltid=AfmBOooWFp_gqZGiGyOR1nSzPoSCvp0AuGlgX7Pz2xpTKf1D8d0jXxES

Pragnanz

- Pragnanz describes the human tendency to simplify complexity.
- Our environment constantly bombards our senses with stimuli, while we have limited attention and processing capacity to handle all the complexity. Pragnanz helps us see order and regularity in a world of visual competition.



https://www.interaction-design.org/literature/topics/gestalt-principles?srsltid=AfmBOooWFp_gqZGiGyOR1nSzPoSCvp0AuGlgX7Pz2xpTKf1D8d0jXxES

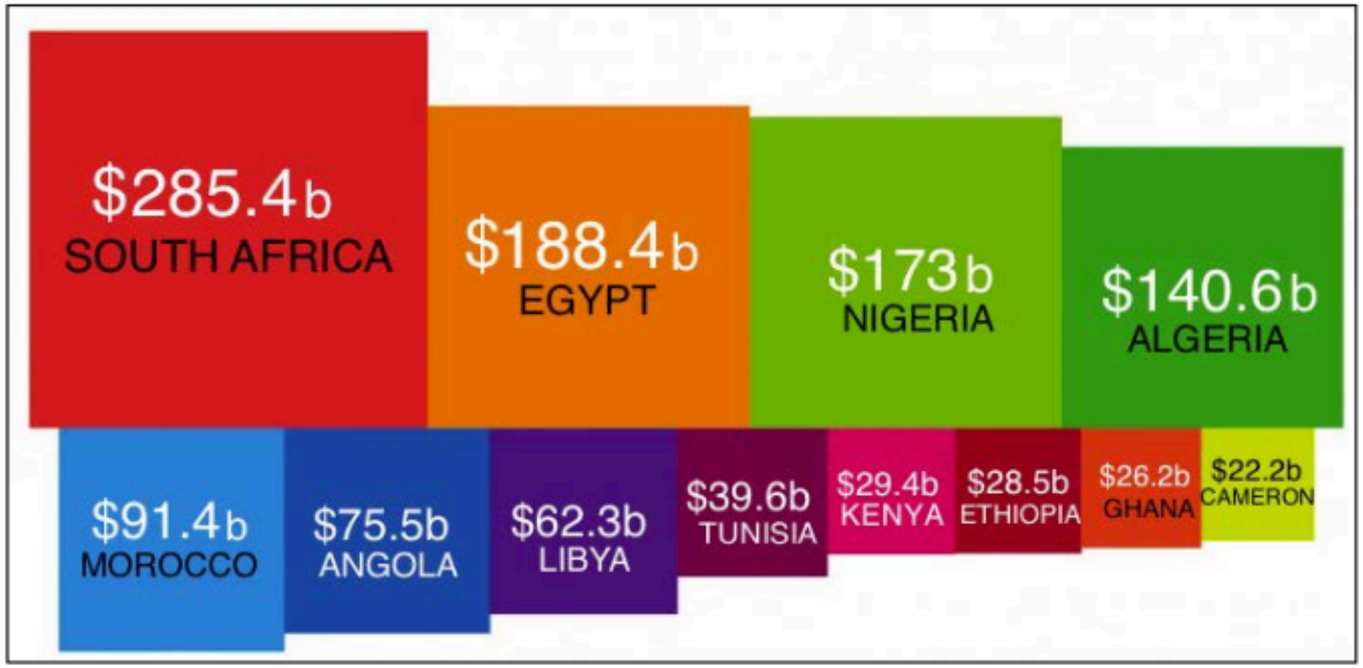
Best practices for visualization

Best practices for visualization

- Comparison and ranking
- Correlation
- Distribution
- Location-specific or geodata
- Part-to-whole relationships
- Trends over time

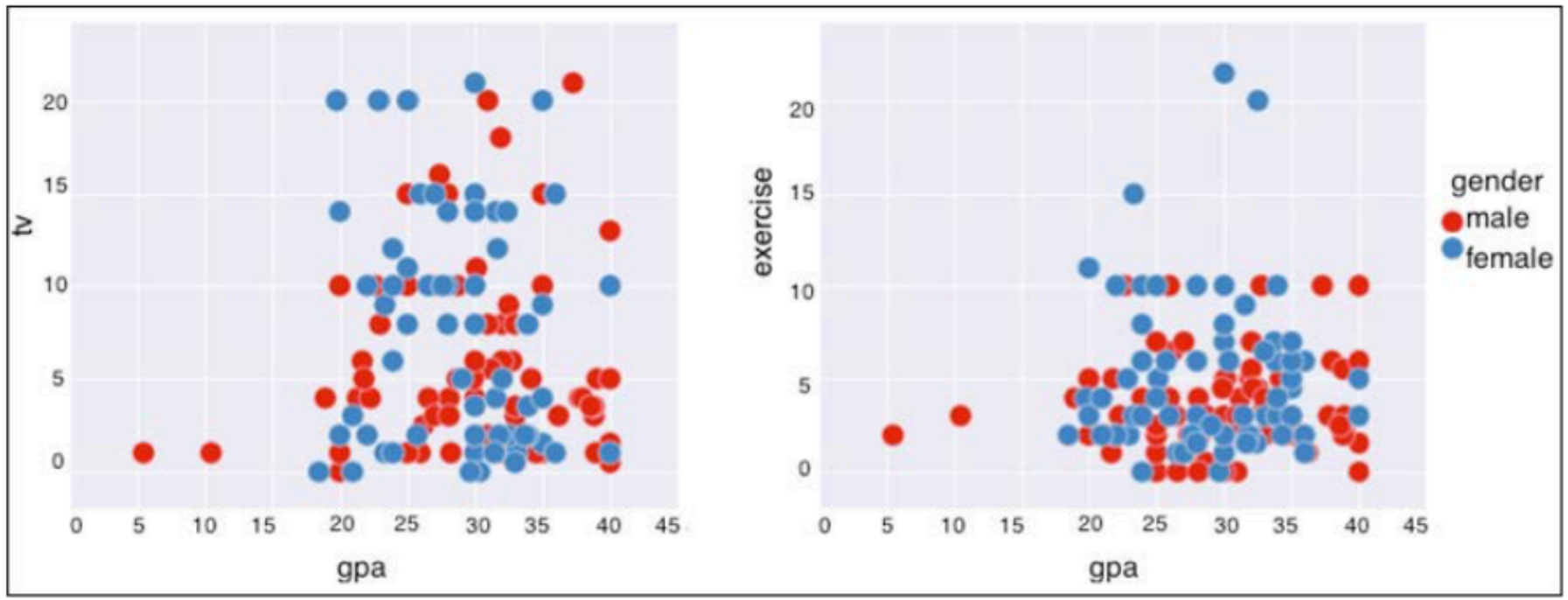
Comparison and ranking

- Typically using bar charts



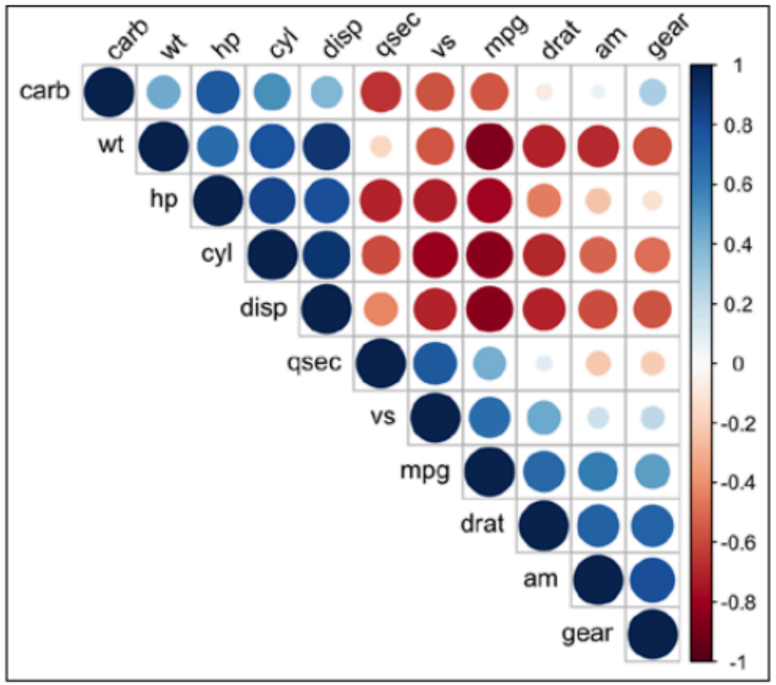
Correlation

- Typically using scatter plot to detect the correlations between two factors



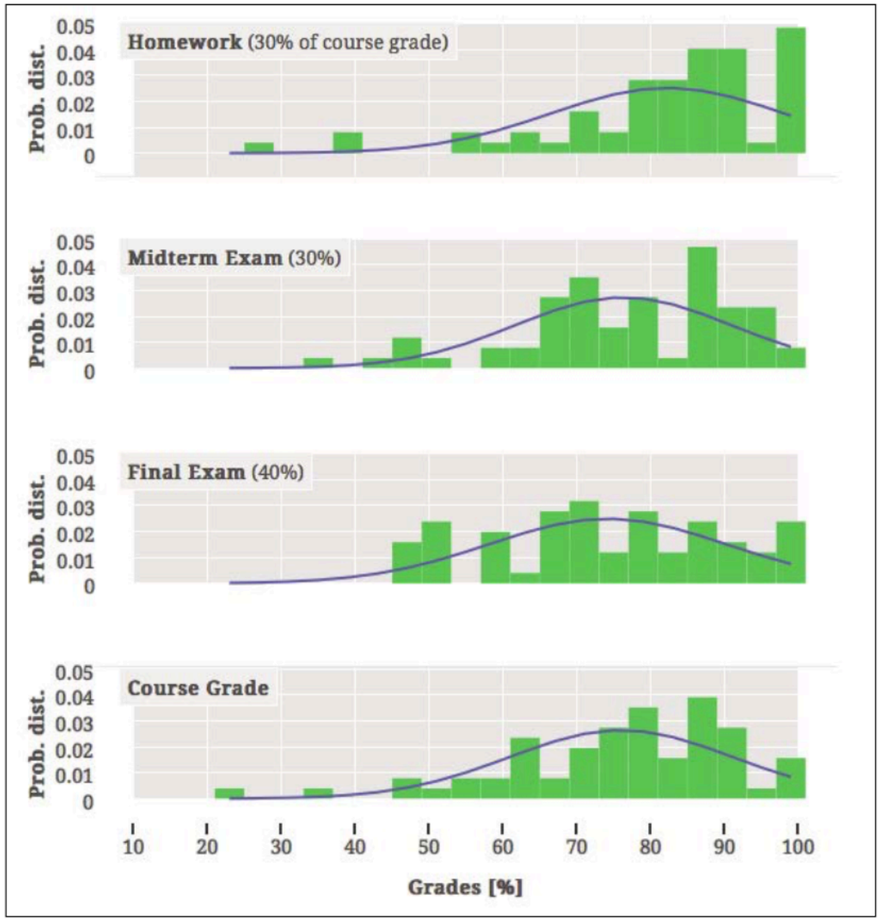
Correlation

- Use of correlation matrix to investigate the dependence between multiple variables at the same time



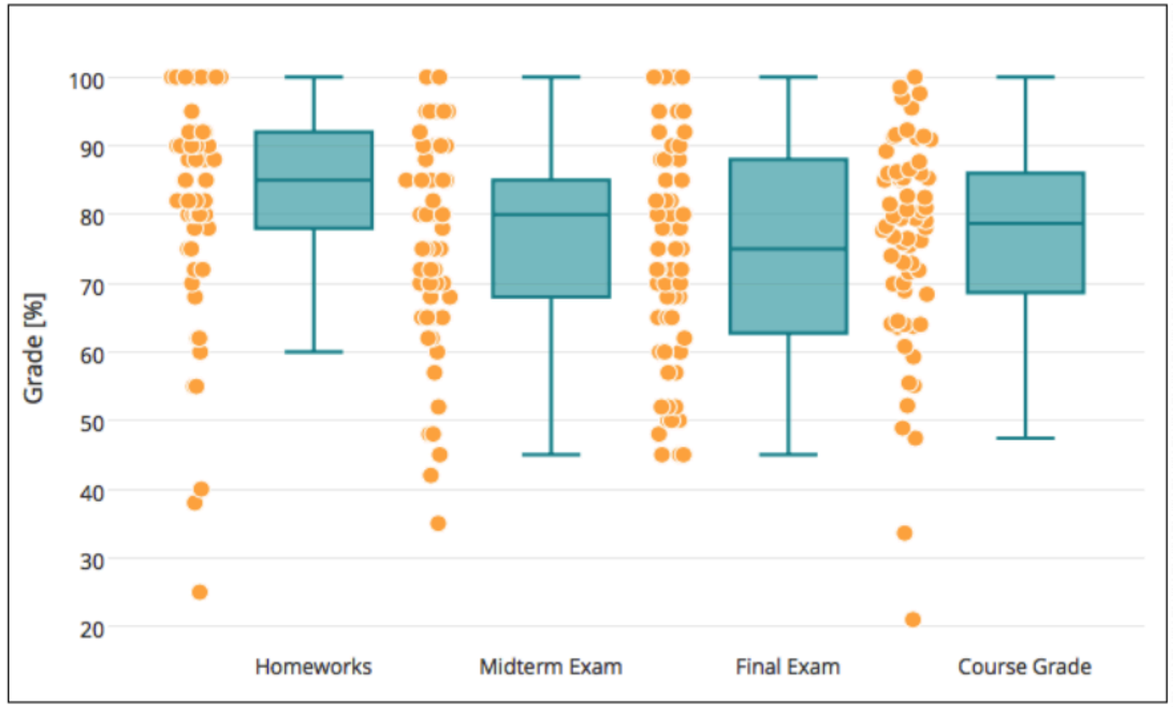
Distribution

- A distribution analysis shows how the quantitative values are distributed across their range
- Histogram, box plot, or box-and-whisker plot



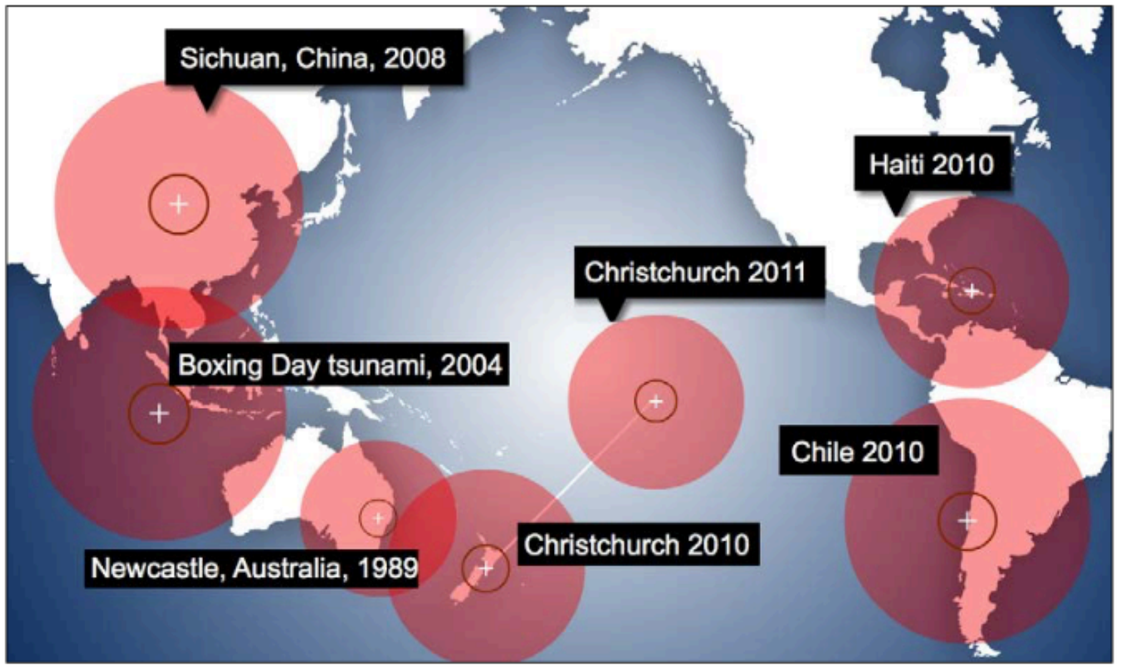
Distribution

- The box-and-whisker plots are excellent for displaying multiple distributions
- Can easily identify the low values, the 25th-percentile values, the medians, the 75th-percentiles, and the maximum values across all categories—all at the same time



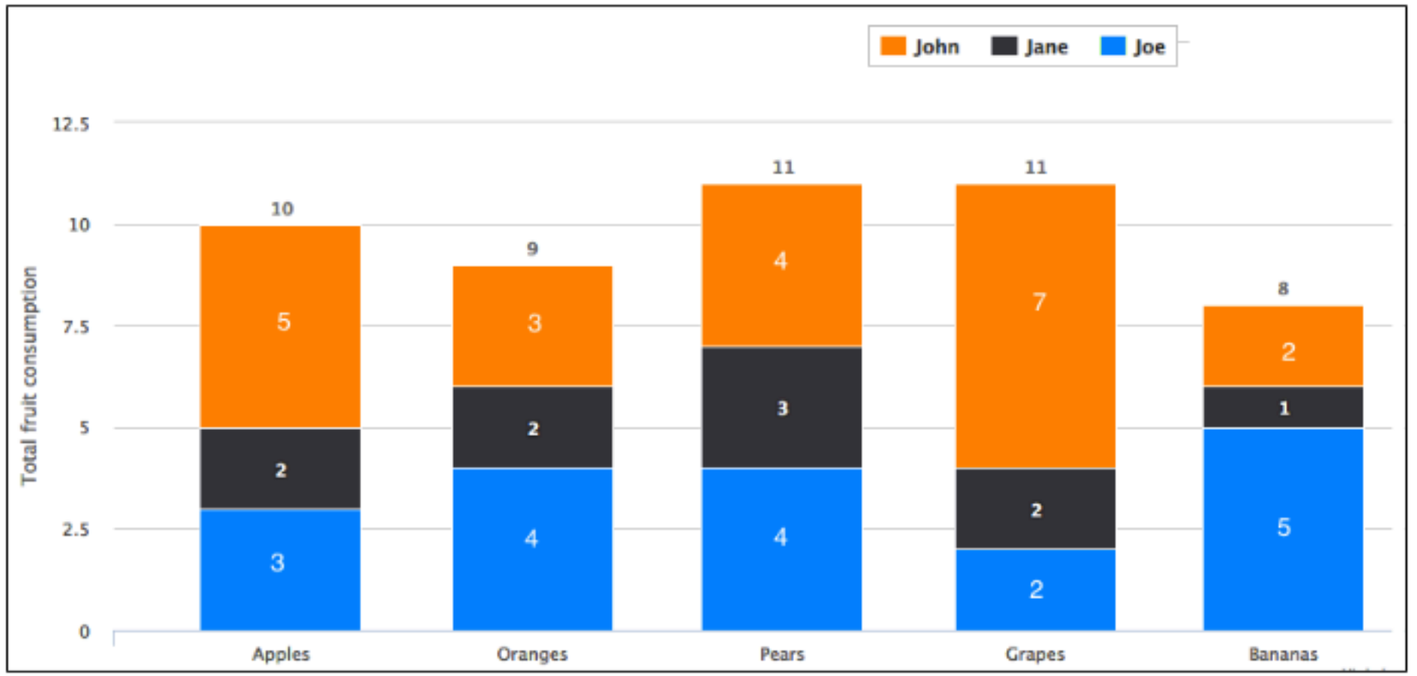
Location-specific or geodata

- Paired with another chart that details what the map is displaying

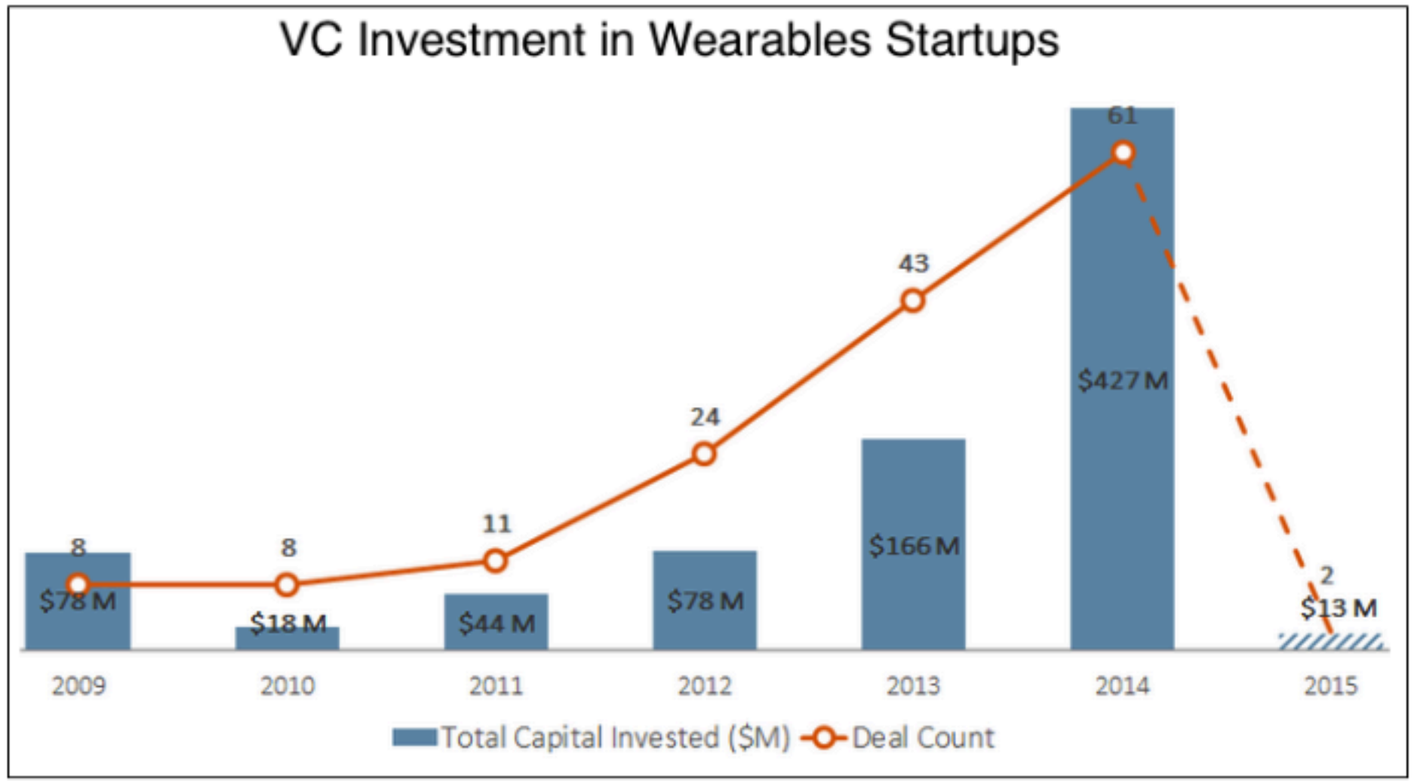


Part-to-whole relationships

- Pie chart, grouped bar charts or stacked column charts



Trends over time



Numerical computing and interactive visualization

Numerical computing and interactive visualization

- May need to visualise different kinds of data
  - Examples: Statistical, numerical, scientific
- Visualization can be static or interactive

Numerical computing

- Popular python packages include:
  - Numerical Python Package (NumPy)
  - Scientific Python Package (SciPy)
  - MKL functions

NumPy

- NumPy not only uses array objects, but also linear algebraic functions that can be conveniently used for computations
- Provides a fast implementation of arrays and associated array functionalities
- Using an array object, one can perform operations that include matrix multiplication, transposition of vectors and matrices, solve systems of equations, perform vector multiplication and normalization, and so on

SciPy

- SciPy is an extension of NumPy for mathematics, science, and engineering that has many packages available for linear algebra, integration, interpolation, fast Fourier transforms, large matrix manipulation, statistical computation, and so on

MKL functions

- Intel® Math Kernel Library (Intel® MKL)
- Provides high-performance routines on vectors and matrices
- Includes FFT functions and vector statistical functions that have been enhanced and optimized to work efficiently on Intel processors

Interactive visualization

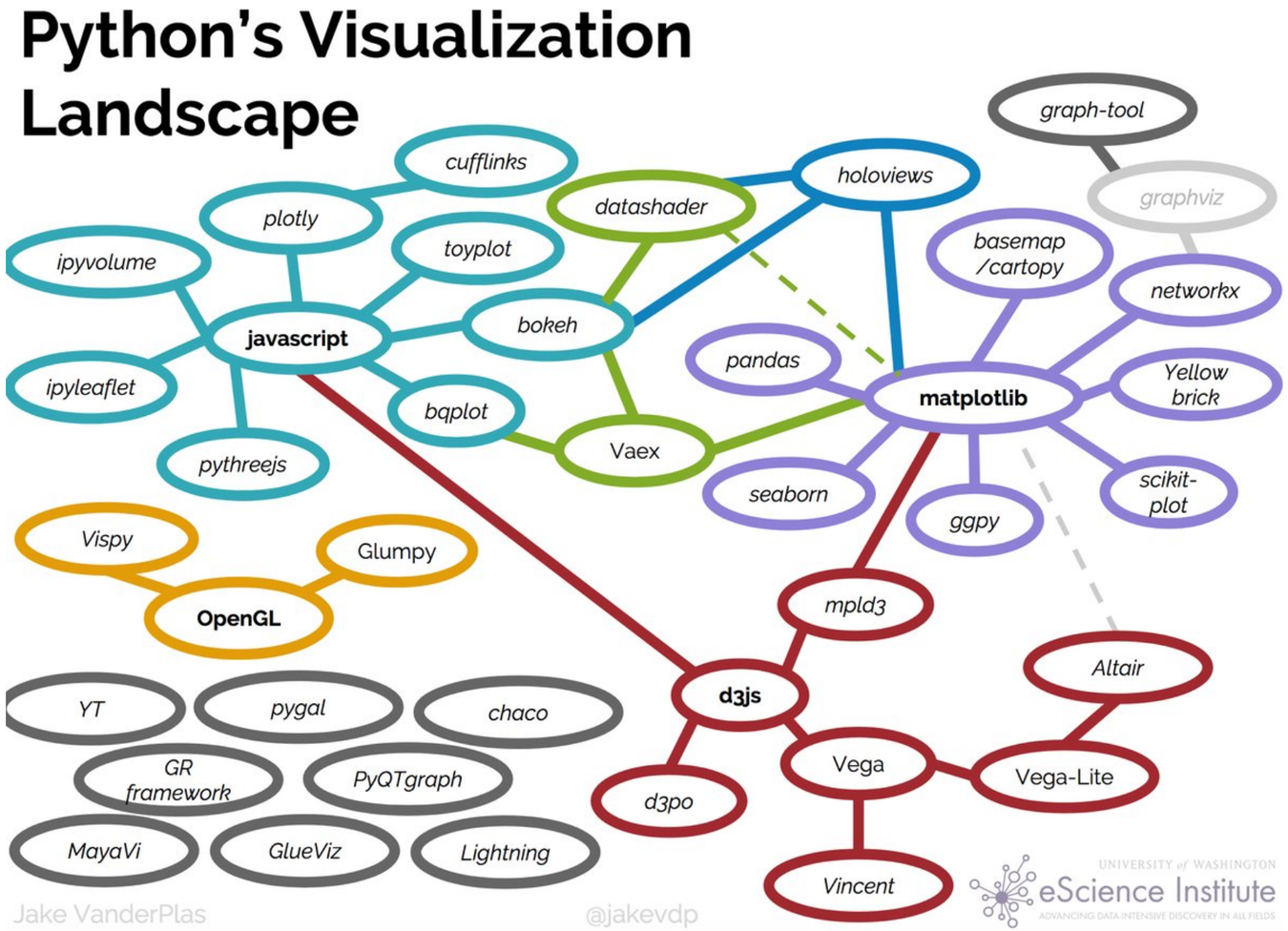
- For a visualization to be considered interactive, it must satisfy two criteria:
  - Human input: The control of some aspect of the visual representation of information must be available to humans
  - Response time: The changes made by humans must be incorporated into the visualization in a timely manner

Advantages and Disadvantages

- Advantage: People can explore a larger information space in a shorter time, which can be understood through one platform
- Disadvantage: It requires a lot of time to exhaustively check every possibility to test the visualization system

Visualization tools in Python

Python’s visualization landscape



https://www.anaconda.com/blog/python-data-visualization-2018-why-so-many-libraries

- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python
- https://matplotlib.org

- Seaborn is a Python data visualization library based on matplotlib
- It provides a high-level interface for drawing attractive and informative statistical graphics
- https://seaborn.pydata.org

- Bokeh is an interactive visualization library for modern web browsers
- It provides elegant, concise construction of versatile graphics, and affords high-performance interactivity over large or streaming datasets.
- Can make interactive plots, dashboards, and data applications
- https://bokeh.org

Vega-Altair

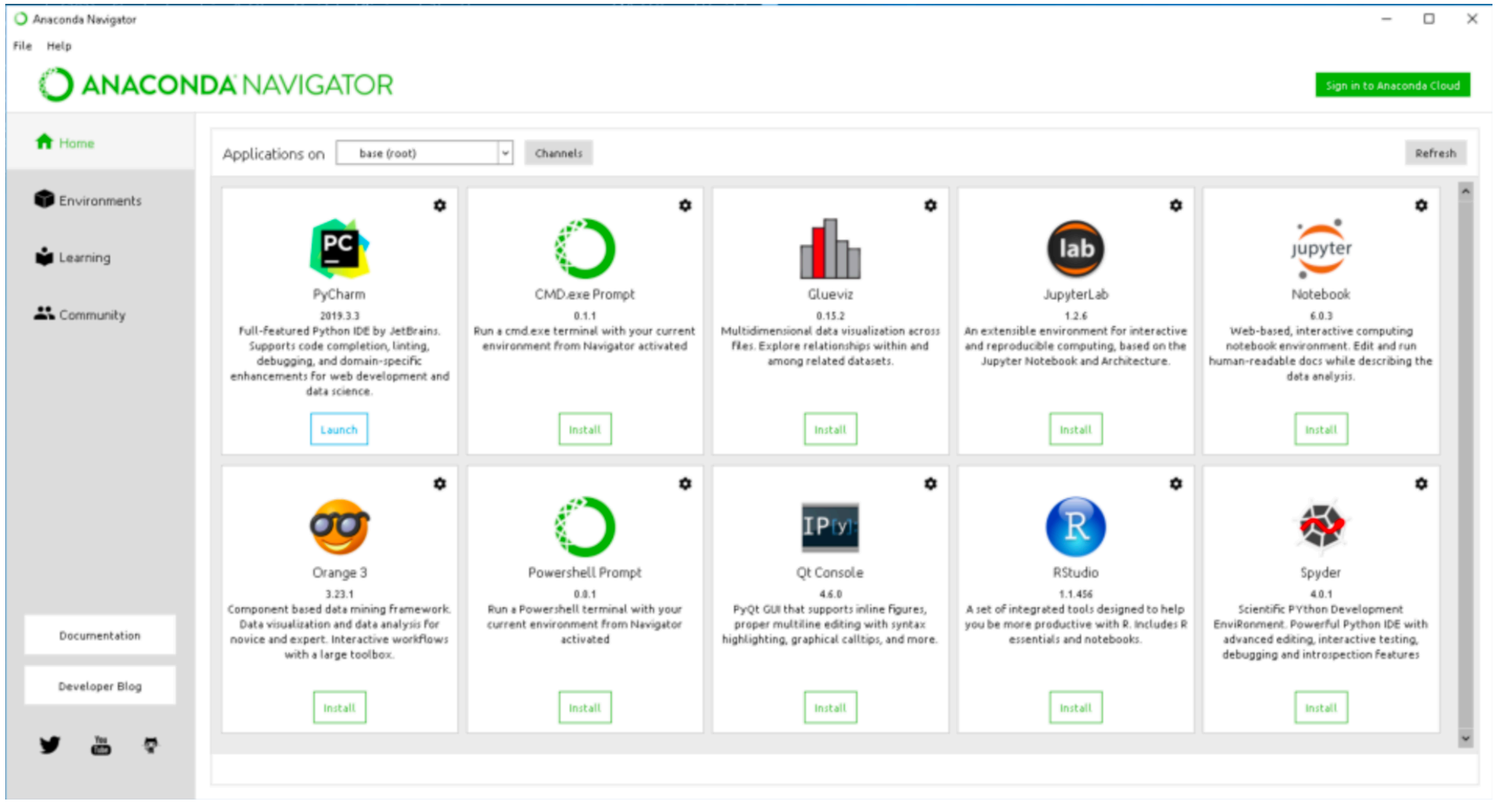
- A declarative statistical visualization library based on Vega and Vega-Lite.
- It has simple syntax for creating complex visualizations with minimal code.
- Supports interactivity and animations.
- Built-in support for data transformations and aggregations.
- https://altair-viz.github.io/

Which tool should be used?

- Plot types
- Data size
- User interface and publishing
- API types

Anaconda IDE

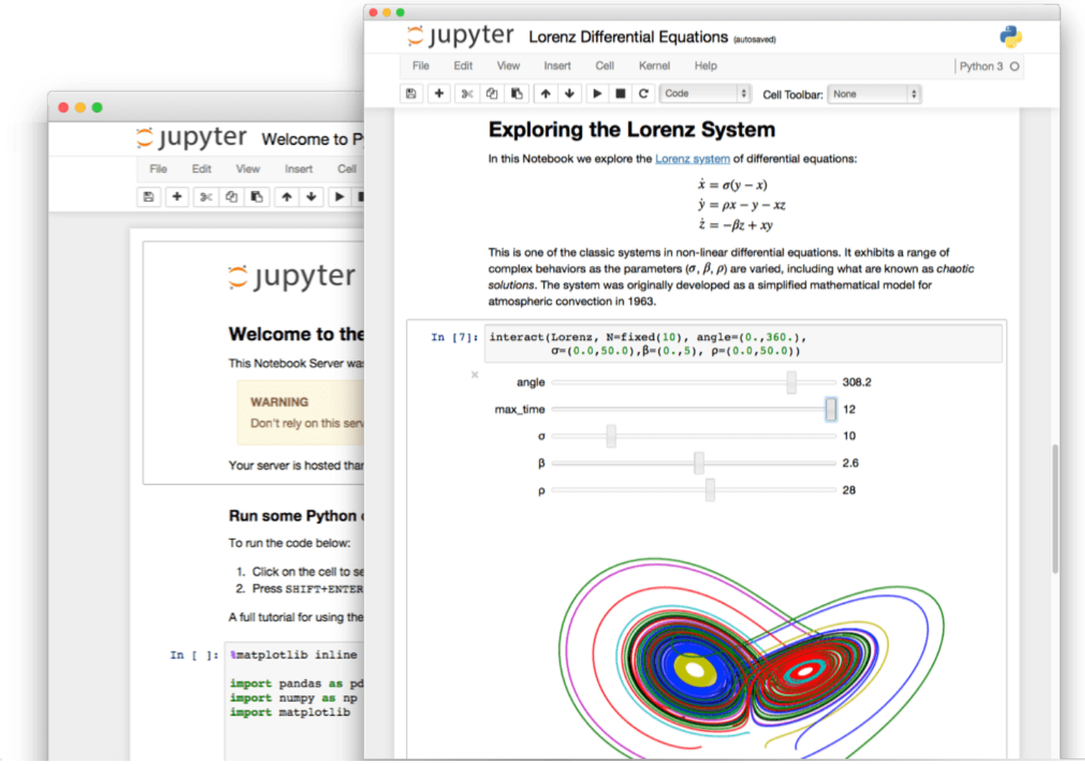
- https://docs.anaconda.com/anaconda/navigator/



Jupyter Notebook

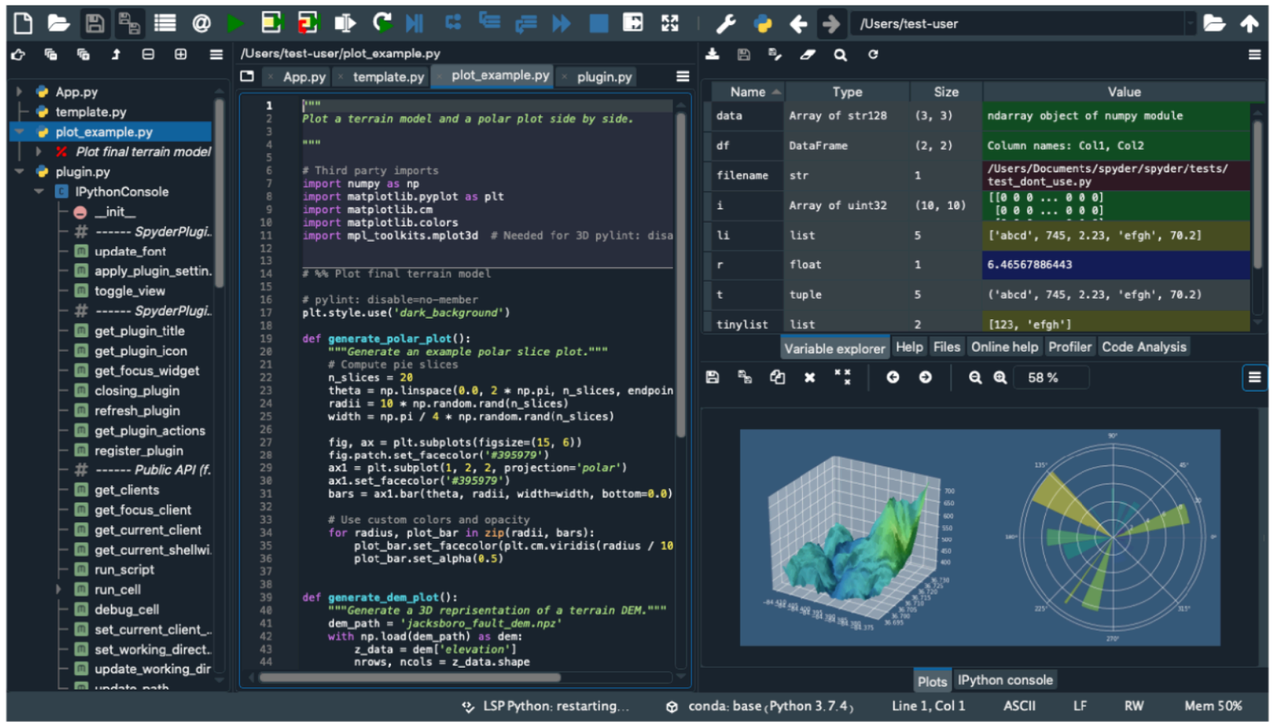
- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text

- https://jupyter.org



Spyder

- Spyder is a free and open source scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts.
- It features a unique combination of the advanced editing, analysis, debugging, and profiling functionality of a comprehensive development tool with the data exploration, interactive execution, deep inspection, and beautiful visualization capabilities of a scientific package.
- https://www.spyder-ide.org



Basic visualization with pandas

Pandas and matplolib

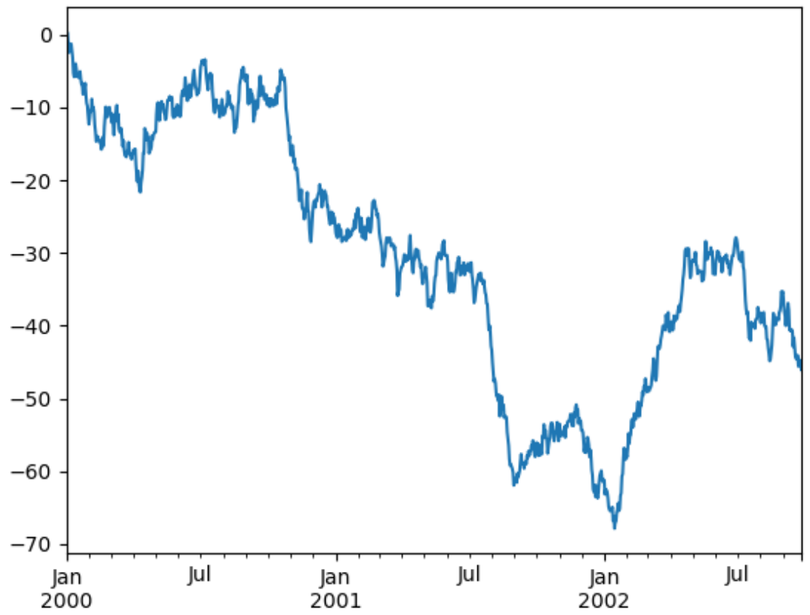
- **pandas** has build-in wrapper around the matplotlib.pyplot library that allows the visualization of data directly from Series or DataFrame
- Load your dataset into a DataFrame.
- Use the `.plot()` method for different types of visualizations:

Series

```
In [3]: ts = pd.Series(np.random.randn(1000), index=pd.date_range("1/1/2000", periods=1000))

In [4]: ts = ts.cumsum()

In [5]: ts.plot();
```



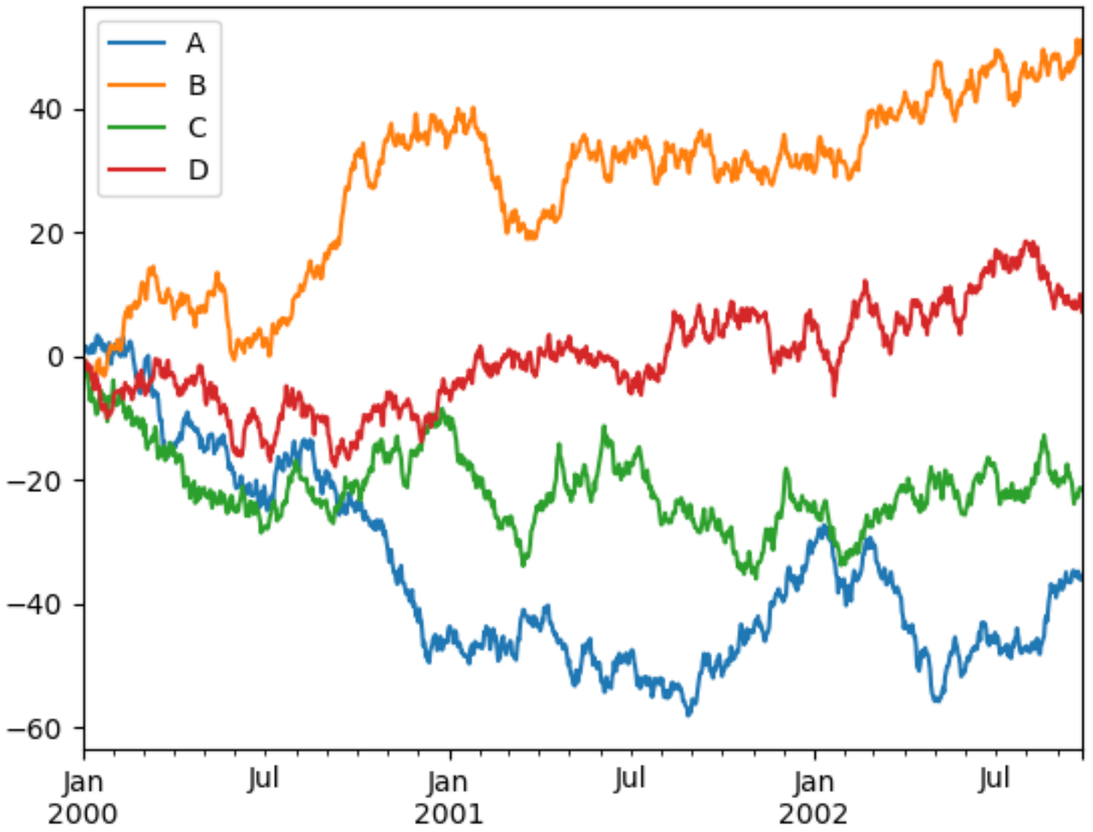
DataFrame

```
In [6]: df = pd.DataFrame(np.random.randn(1000, 4), index=ts.index, columns=list("ABCD"))

In [7]: df = df.cumsum()

In [8]: plt.figure();

In [9]: df.plot();
```



References

- Part of this slide set is prepared or/and extracted from the following book / websites:
  - MKirthi Raman (2015) , “Mastering Python Data Visualization”, Packt Publishing
  - pandas, https://pandas.pydata.org/
  - Dataspire. (n.d.). *Leveraging perception science to our advantage*. Retrieved January 21, 2025, from https://dataspire.org/blog/leveraging-perception-science-to-our-advantage
  - Interaction Design Foundation. (n.d.). *What are the Gestalt Principles?* Retrieved January 21, 2025, from https://www.interaction-design.org/literature/topics/gestalt-principles?srsltid=AfmBOooWFp_gqZGiGyOR1nSzPoSCvp0AuGlgX7Pz2xpTKf1D8d0jXxE
- This set of slides is for teaching purpose only