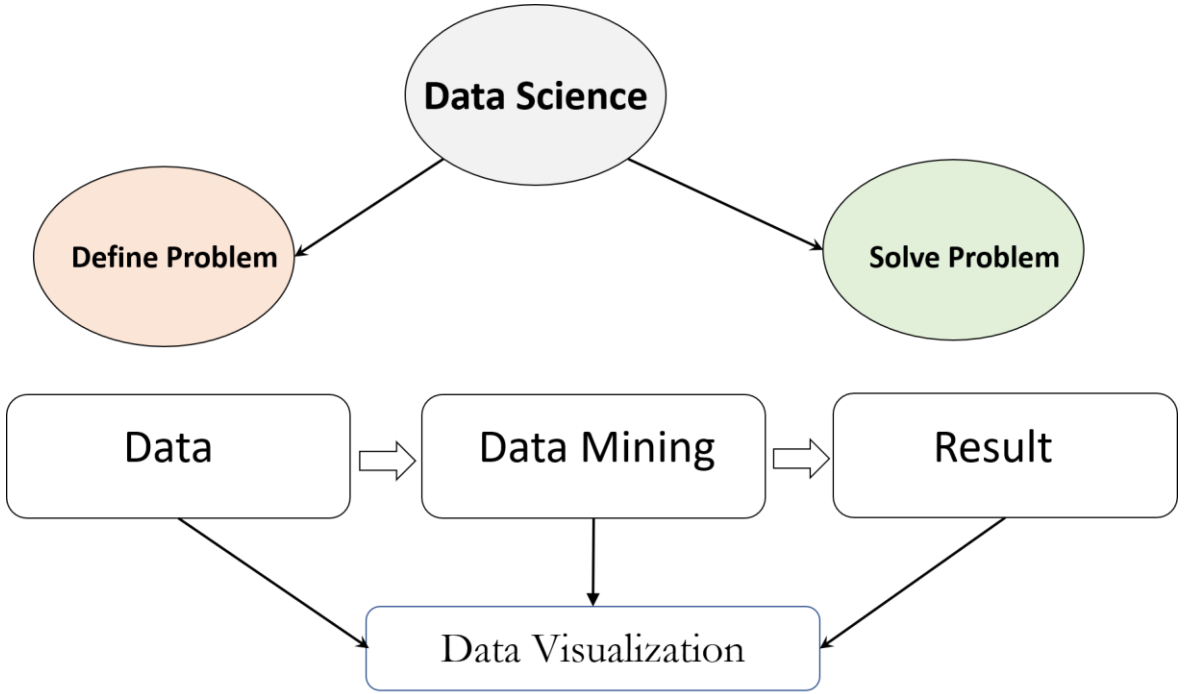
Data Visualization- What and why

Dr. Aihua Li

Any data science or data analytics project can be generally described with the following steps:

- Acquiring the business problems
- Getting data
- Processing data **(may also need data visualization for the details inside data)**
- Analyzing and modeling data
- **Visualizing data**
- Deploying the model
- Scoring its performance



What is data visualization?

- Data visualization is a graphical representation of any data or information.
- Tamara Munzner (2000, Visualization Analysis & Design): Computer-based visualization systems provide visual representations of datasets designed to help people carry out tasks more effectively.

Why data visualization?

Computer-based visualization systems provide visual **representations** of **datasets** designed to help **people** carry out tasks more **effectively**.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

- **Why have a human in the decision-making loop?**
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

- ✓ Visualization is suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods.
- ✓ Visualization is not needed when fully automatic solution exists and it is trusted.
- ✓ Many analysis problems are ill-specified: don't know exactly what questions to ask in advance.

- Why have a human in the decision-making loop?
- **Why have a computer in the loop?**
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ You can build tools that allow people to explore or present large datasets

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- **Why use an external representation?**
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ to surpass the limitations of our own internal cognition and memory

✓ e.g., two dimensional display

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- **Why depend on vision?**
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ Visualization, is based on exploiting the human visual system as a means of communication.

✓ The visual system provides a very high-bandwidth channel to our brain.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- **Why show the data in detail?**
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ Vis tools help people in situations where seeing the dataset structure in detail is better than seeing only a brief summary of it.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- **Why use interactivity?**
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ Interactivity is crucial for building vis tools that handle complexity.

✓ When datasets are large enough, the limitations of both people and displays preclude just showing everything at once.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- **Why is the vis idiom design space huge?**
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ There are many ways to create a visual encoding of data as a single picture. The design space of possibilities gets even bigger when you consider how to manipulate one or more of these pictures with interaction.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- **Why focus on tasks?**
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ A tool that serves well for one task can be poorly suited for another, for exactly the same dataset.

✓ Reframing the users' task from domain-specific form into abstract form allows you to consider the similarities and differences between what people need across many real-world usage contexts

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?

- **Why are most designs ineffective?**
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ A poor match with the properties of the human perceptual and cognitive systems.

✓ Be comprehensible by a human in some other setting, but it's a bad match with the intended task.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- **Why care about effectiveness?**
- Why is validation difficult?
- Why are there resource limitations?
- Why analyze visualization?

✓ Goal: support user tasks.

✓ This goal leads to concerns about correctness, accuracy, and truth playing a very central role in visualization.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- **Why is validation difficult?**
- Why are there resource limitations?
- Why analyze visualization?

✓ The problem of validation for a visualization design is difficult because there are so many questions that you could ask when considering whether a vis tool has met your design goals.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- **Why are there resource limitations?**
- Why analyze visualization?

✓ When designing or analyzing a vis system, you must consider at least three different kinds of limitations: computational capacity, human perceptual and cognitive capacity, and display capacity.

- Why have a human in the decision-making loop?
- Why have a computer in the loop?
- Why use an external representation?
- Why depend on vision?
- Why show the data in detail?
- Why use interactivity?
- Why is the vis idiom design space huge?
- Why focus on tasks?
- Why are most designs ineffective?
- Why care about effectiveness?
- Why is validation difficult?
- Why are there resource limitations?
- **Why analyze visualization?**

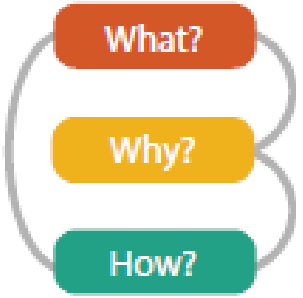


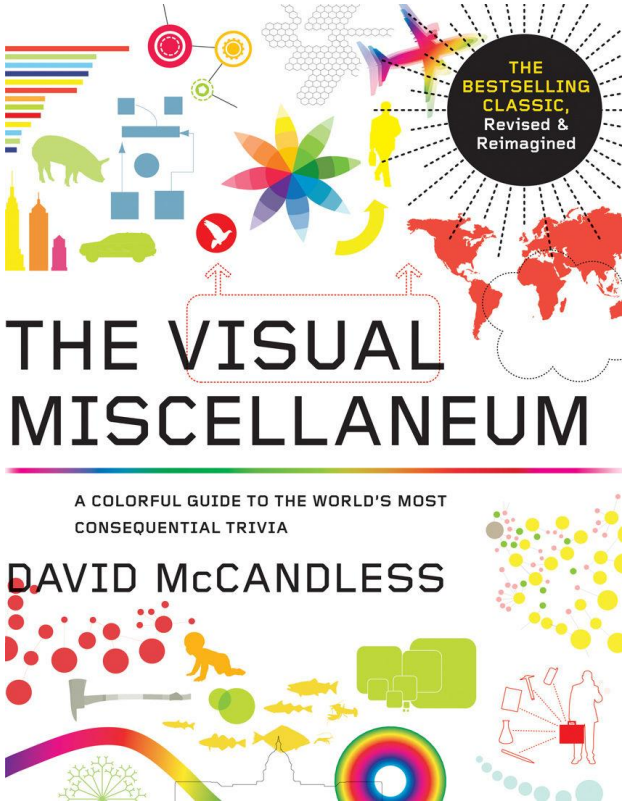
Figure 1.7. Three-part analysis framework for a vis instance: *why* is the task being performed, *what* data is shown in the views, and *how* is the vis idiom constructed in terms of design choices.

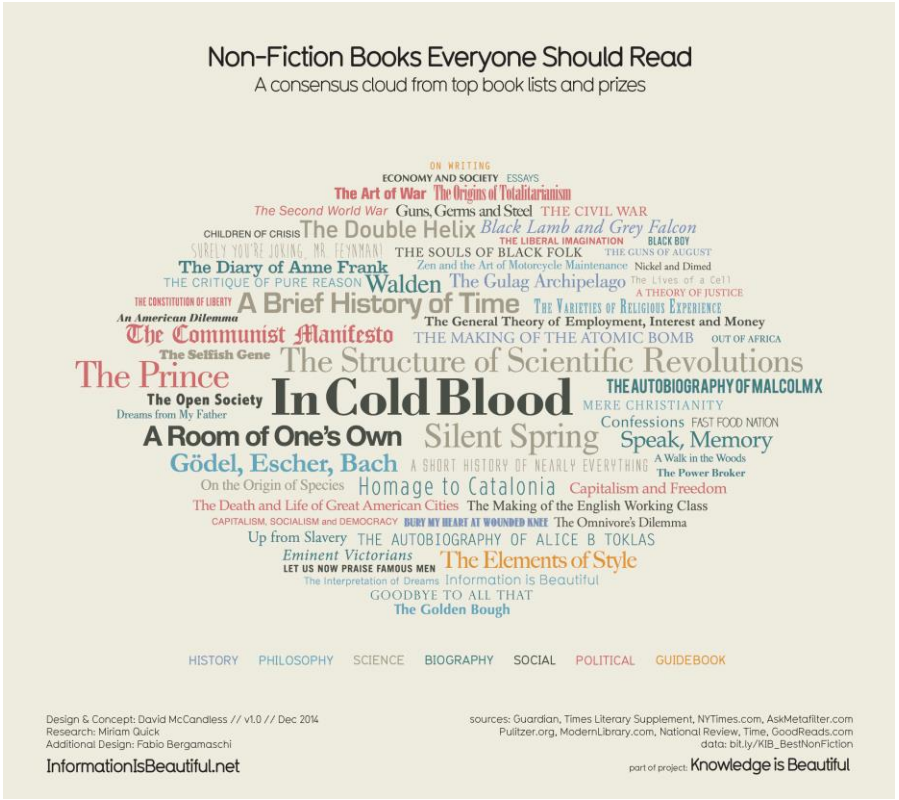
Possibility of data visualization:

- Long-term use of end users (e.g. exploratory analysis of scientific data)
- Presentation of known results
- Stepping stone to better understanding of requirements before developing models
- Help developers of automatic solution refine/debug, determine parameters
- Help end users of automatic solution verify, build trust

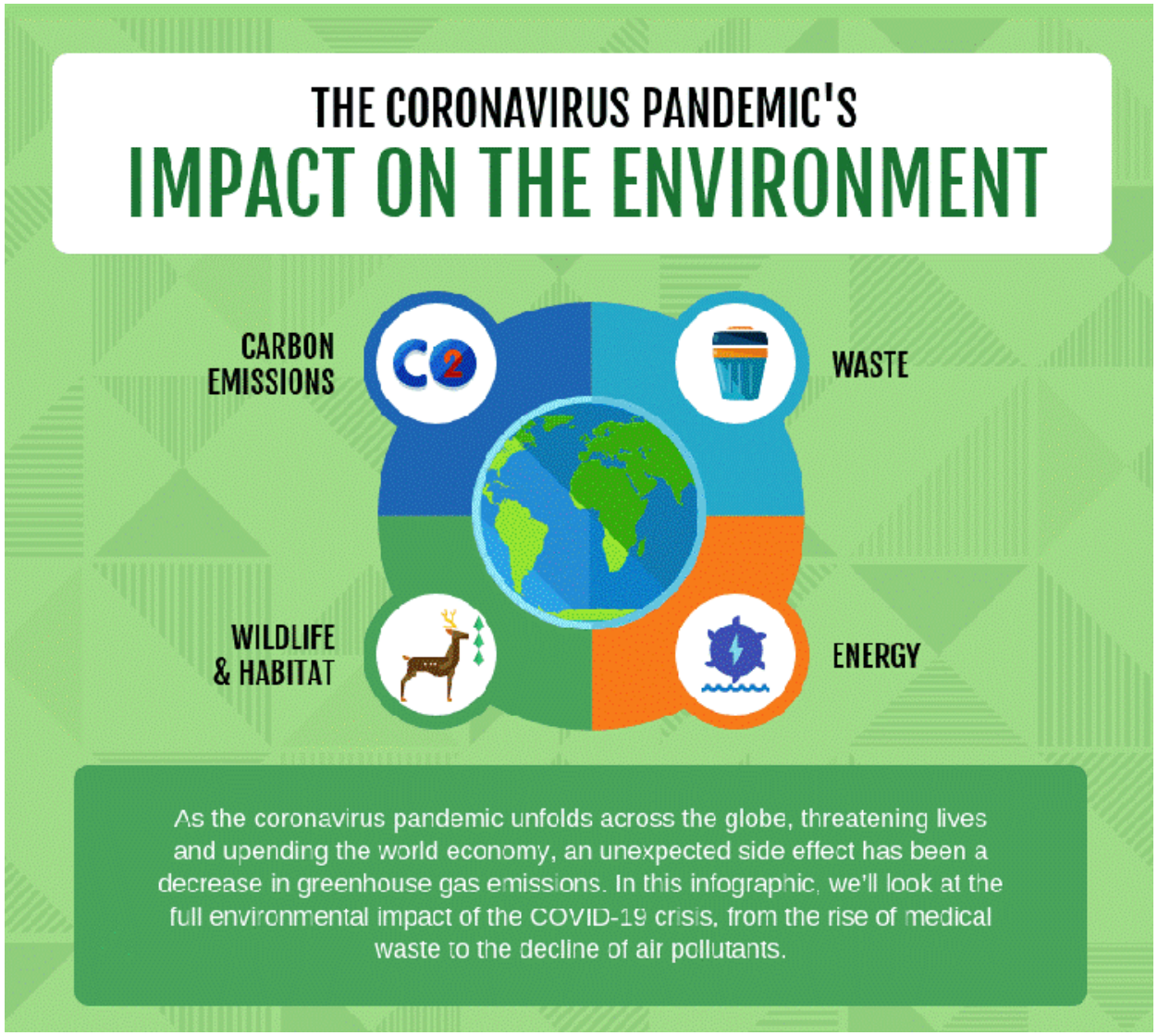
The beauty of data visualization

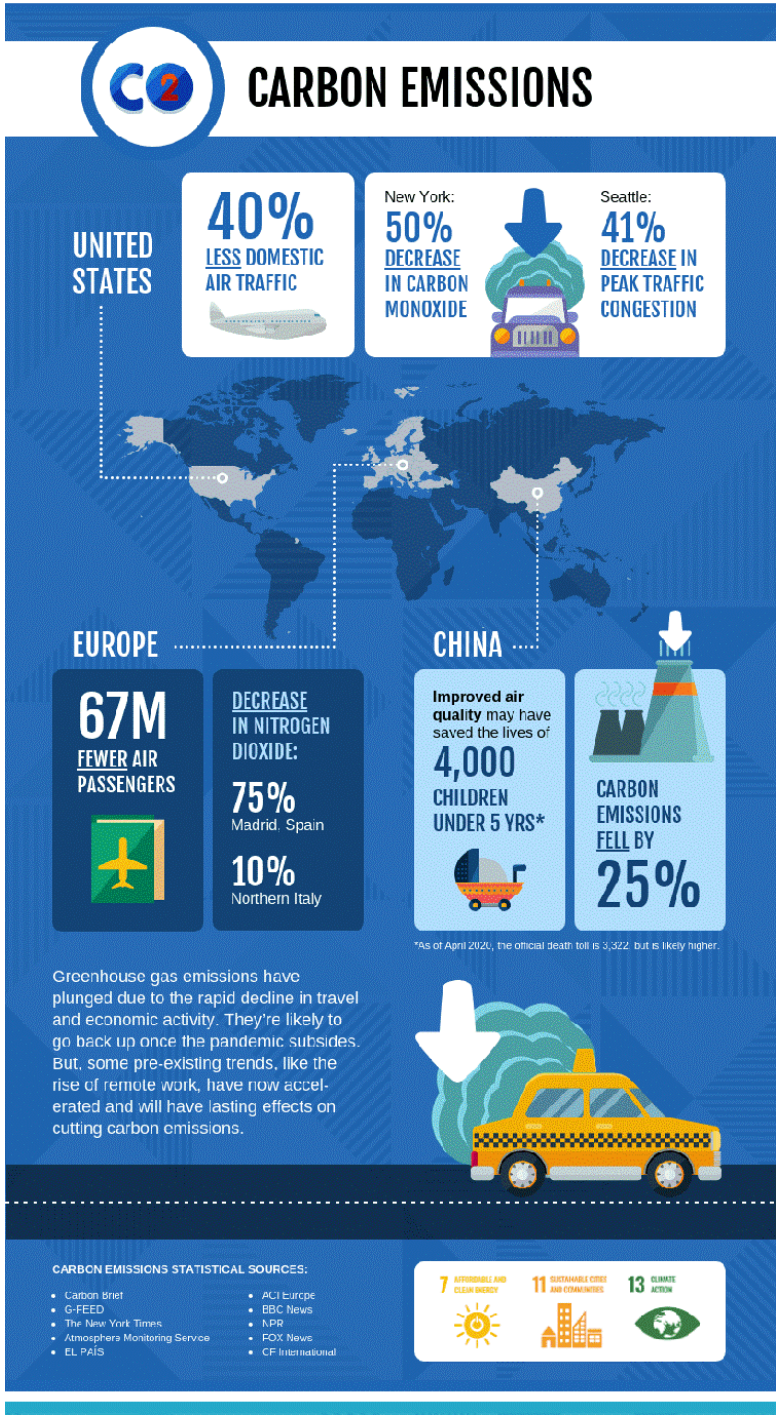
David McCandless: Good design is the best way to navigate information glut -- and it may just change the way we see the world.





Some examples





science-2013-global-forest

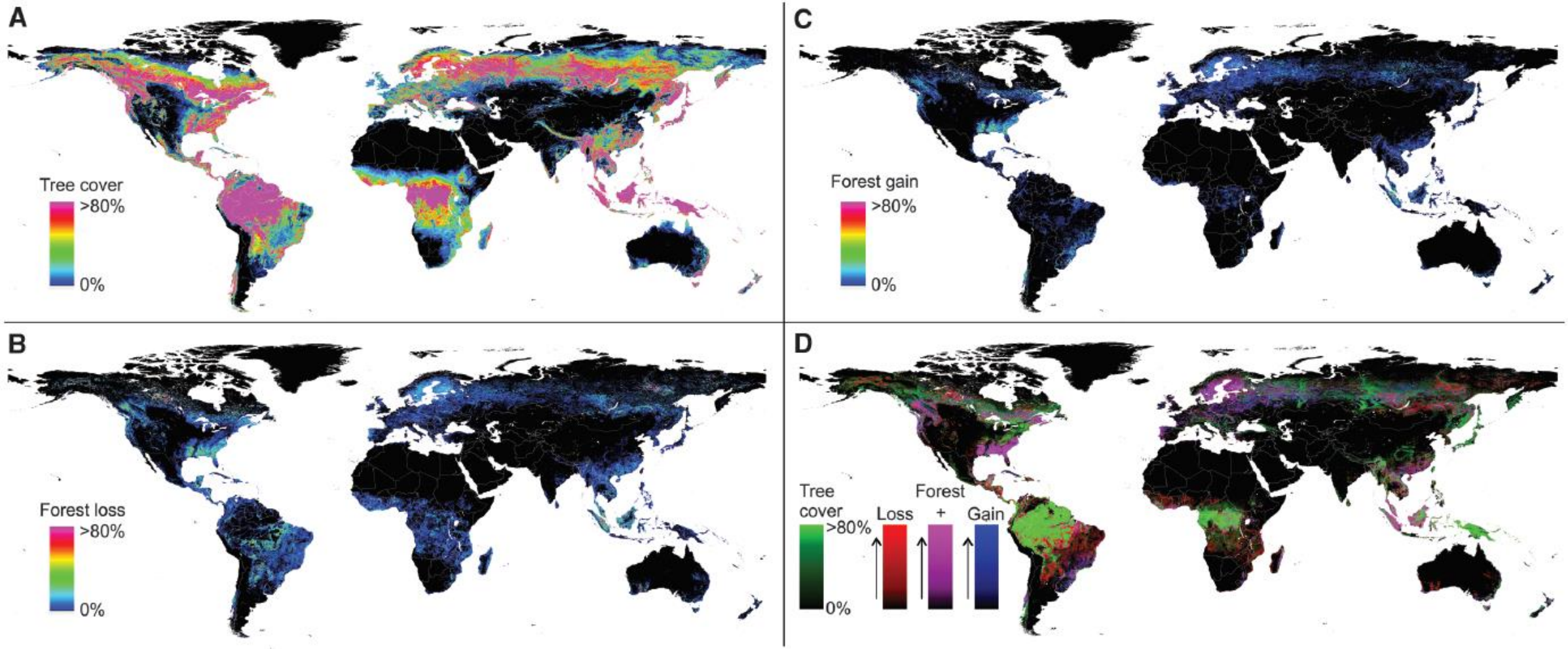


Fig. 1. (A) Tree cover, (B) forest loss, and (C) forest gain. A color composite of tree cover in green, forest loss in red, forest gain in blue, and forest loss and gain in magenta is shown in (**D**), with loss and gain enhanced for improved visualization. All map layers have been resampled for display purposes from the 30-m observation scale to a 0.05° geographic grid.

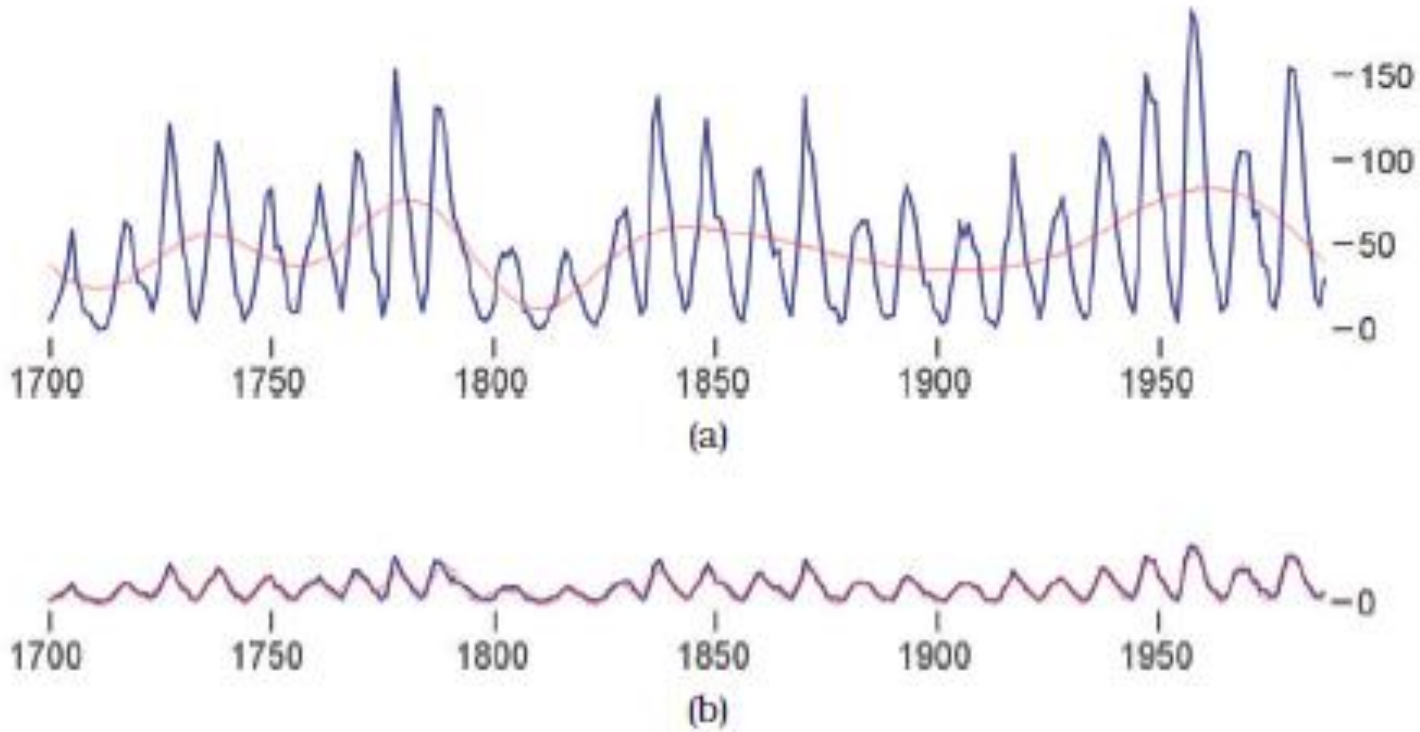


Figure 7.10. Sunspot cycles. The multiscale banking to 45° idiom exploits our orientation resolution accuracy at the diagonal. (a) An aspect ratio close to 4 emphasizes low-frequency structure. (b) An aspect ratio close to 22 shows higher-frequency structure: cycle onset is mostly steeper than the decay. From [Heer and Agrawala 06, Figure 5].

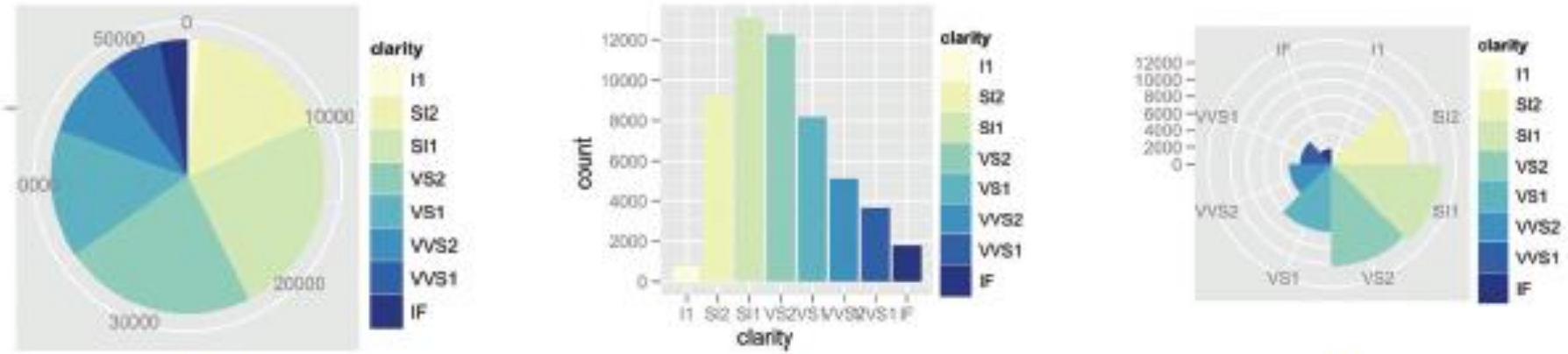


Figure 7.17. Pie chart versus bar chart accuracy. (a) Pie charts require angle and area judgements. (b) Bar charts require only high-accuracy length judgements for individual items. (c) Polar area charts are a more direct equivalent of bar charts, where the length of each wedge varies like the length of each bar. From [Wickham 10, Figures 15 and 16].

Anscombe's quartet

Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

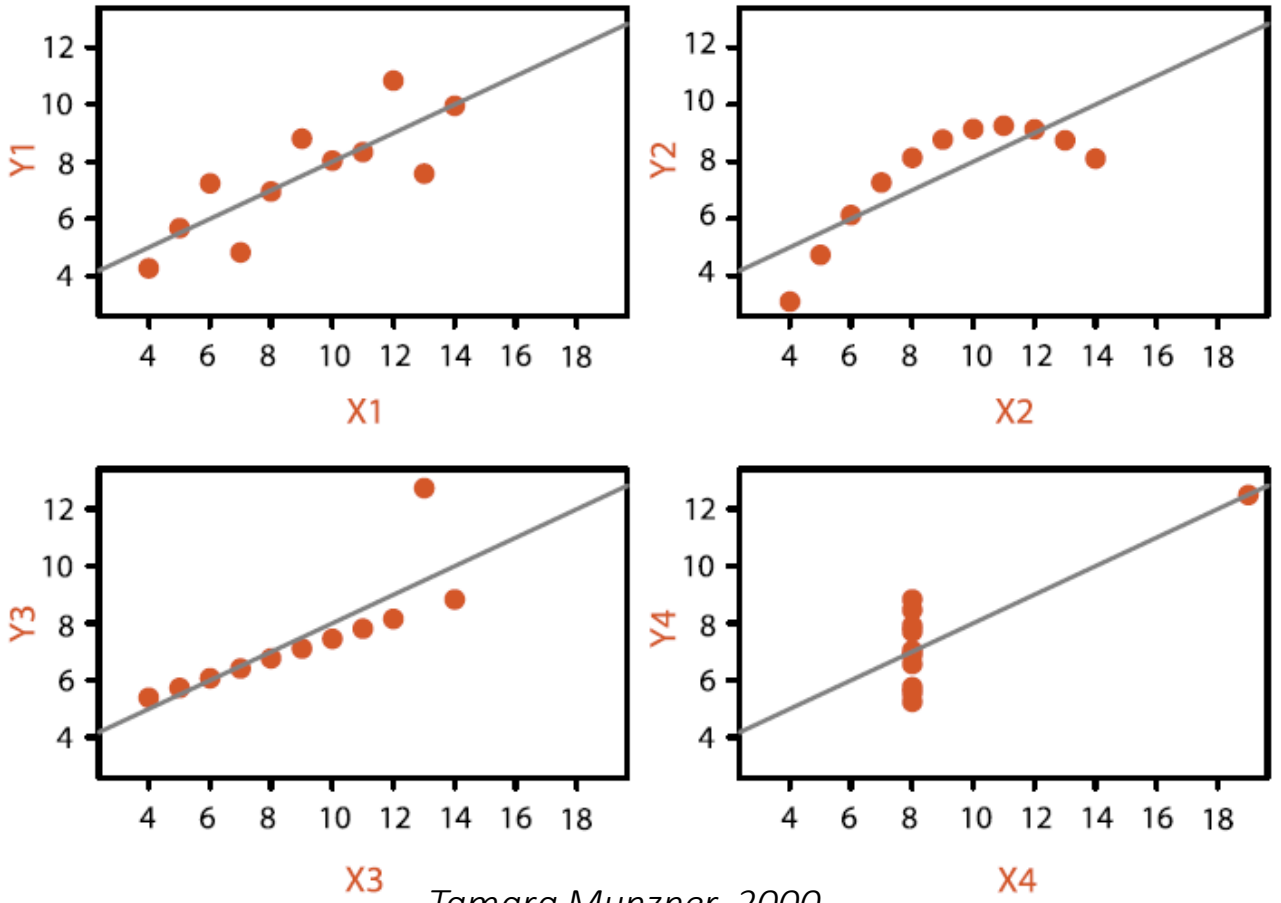
All of these four data sets have different distributions and consists of 11 points marked on x and y-axis.

Tamara Munzner, 2000

Anscombe's quartet

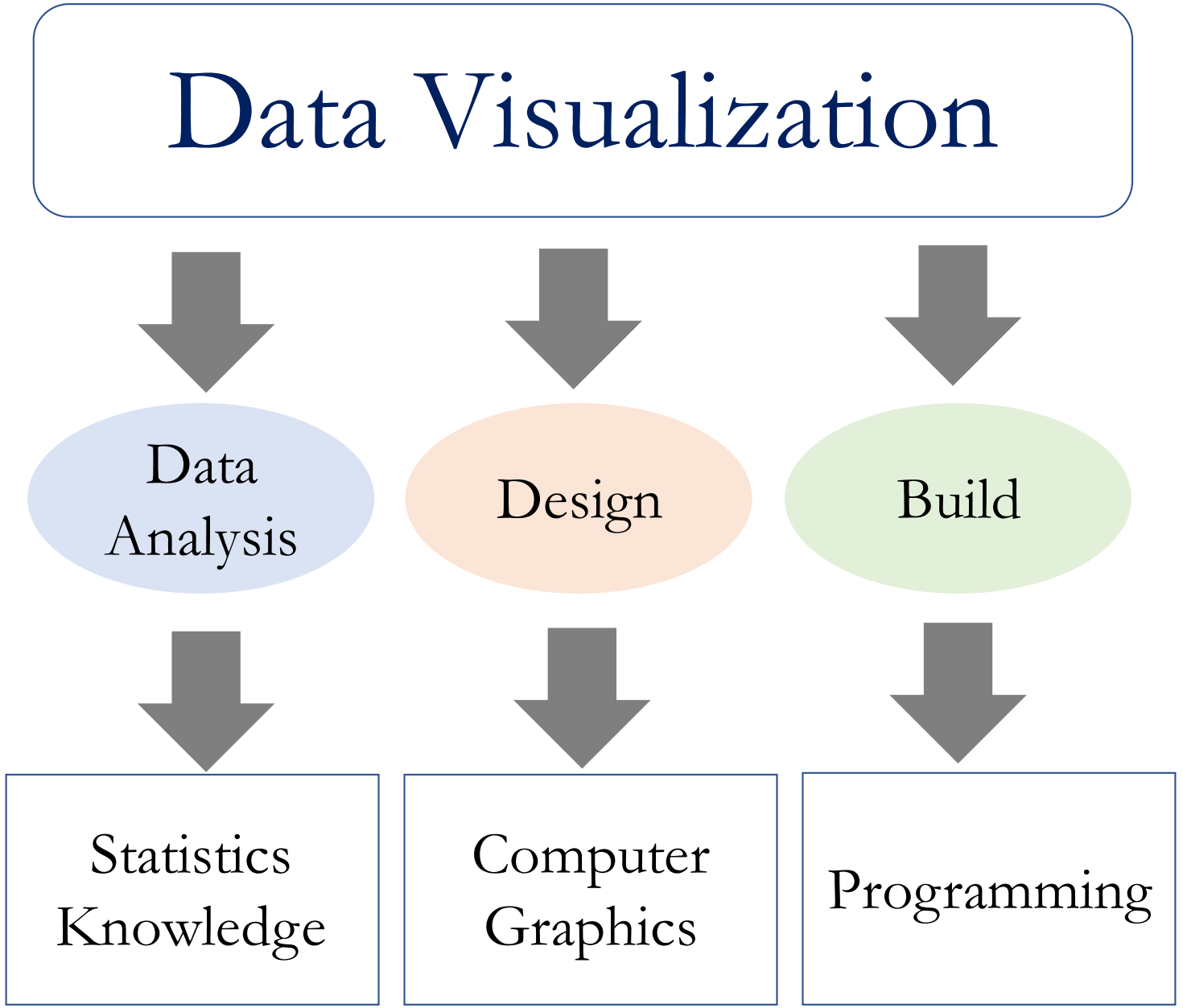
Anscombe's Quartet: Raw Data

	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.			
	14.0	9.96	14.0	8.10	14.			
	6.0	7.24	6.0	6.13	6.			
	4.0	4.26	4.0	3.10	4.			
	12.0	10.84	12.0	9.13	12.			
	7.0	4.82	7.0	7.26	7.			
	5.0	5.68	5.0	4.74	5.			
Mean	9.0	7.5	9.0	7.5	9.			
Variance	10.0	3.75	10.0	3.75	10.			
Correlation	0.816		0.816		(			



All of these four data sets have different distributions and consists of 11 points marked on x and y-axis.

Tamara Munzner, 2000


Data Visualization
Data Analysis
Design
Build
Statistics Knowledge
Computer Graphics
Programming

- **Data Analysis versus Data Visualization**
- Data analysis is an exploratory process that often starts with specific questions. It requires curiosity, the desire to find answers and a good level of tenacity, because those answers aren’t always easy to come by.
- Data visualization involves the visual representation of data, ranging from single charts to comprehensive dashboards. Effective visualizations significantly reduce the amount of time it takes for your audience to process information and access valuable insights.

Data Visualization Tools

Data visualization tool helps in, well, visualizing data. Using these tools, data and information can be generated and read easily and quickly.

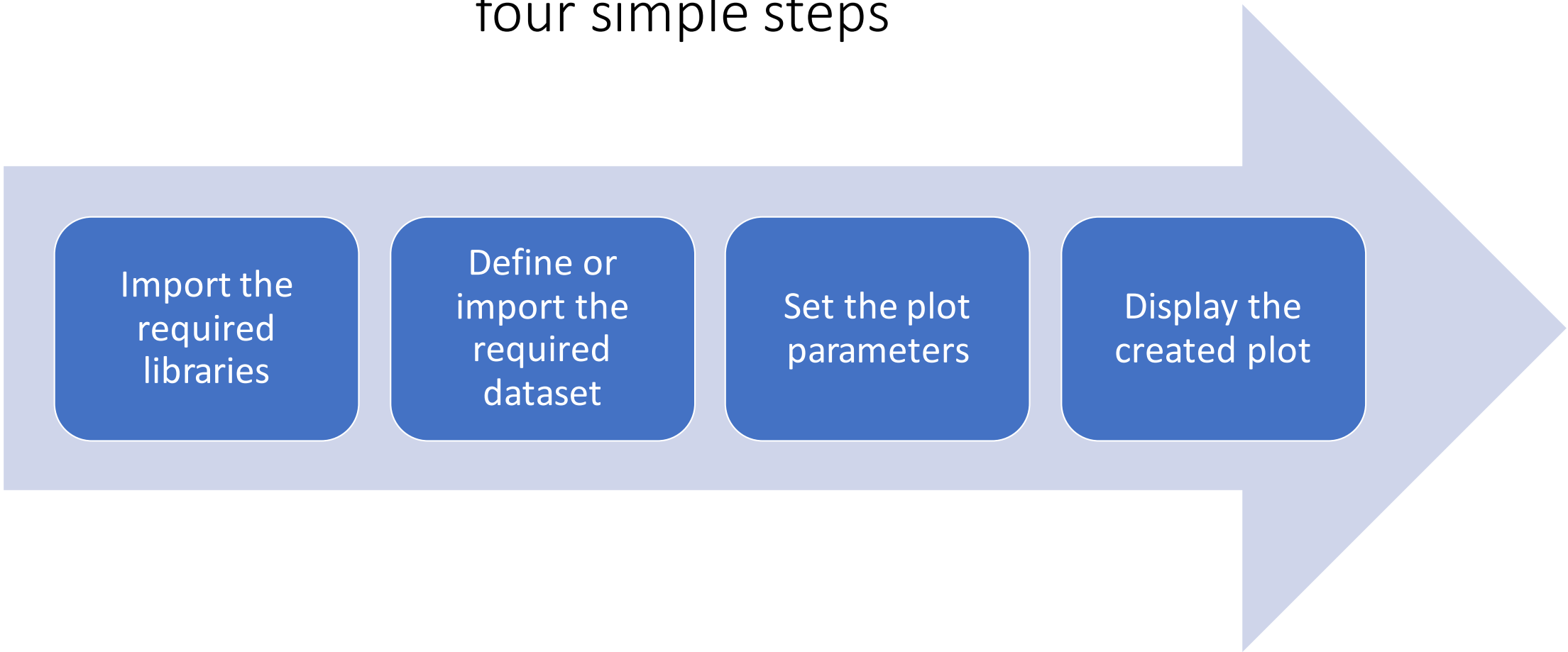
- Tableau Desktop – A business intelligence tool which helps you in visualizing and understanding your data.
- Microsoft Power BI – Developed by Microsoft, this is a suite of business analytics tools that allows you to transform information into visuals.
- MATLAB – A detailed data analysis tool that has an easy-to-use tool interface and graphical design options for visuals.
- R, Python & JavaScript

Data Visualization Techniques

- Know the target audience
- Create a goal
- Choose the chart
- Context
- Use tools

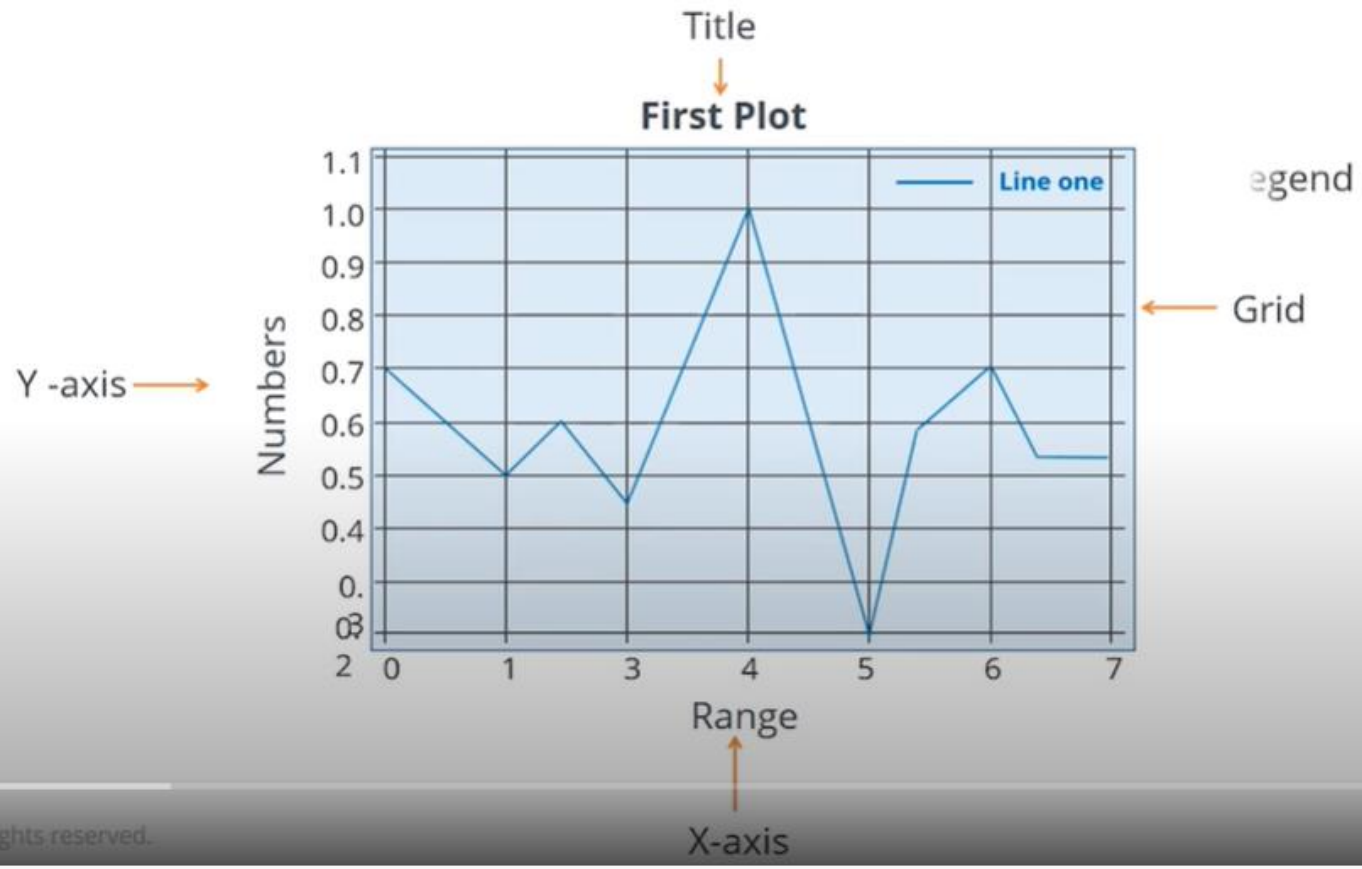
Steps to Create A Plot with R

four simple steps



Understanding the Plot

A plot is a graphical representation of data which shows relationship between two variables or the distribution of data.



- You cannot become a Data scientist without coding knowledge. Data Science is not all about coding but coding is an essential part of Data Science.