ENERGY DATA SCIENCE

Basic typography & Scientific data visualisation

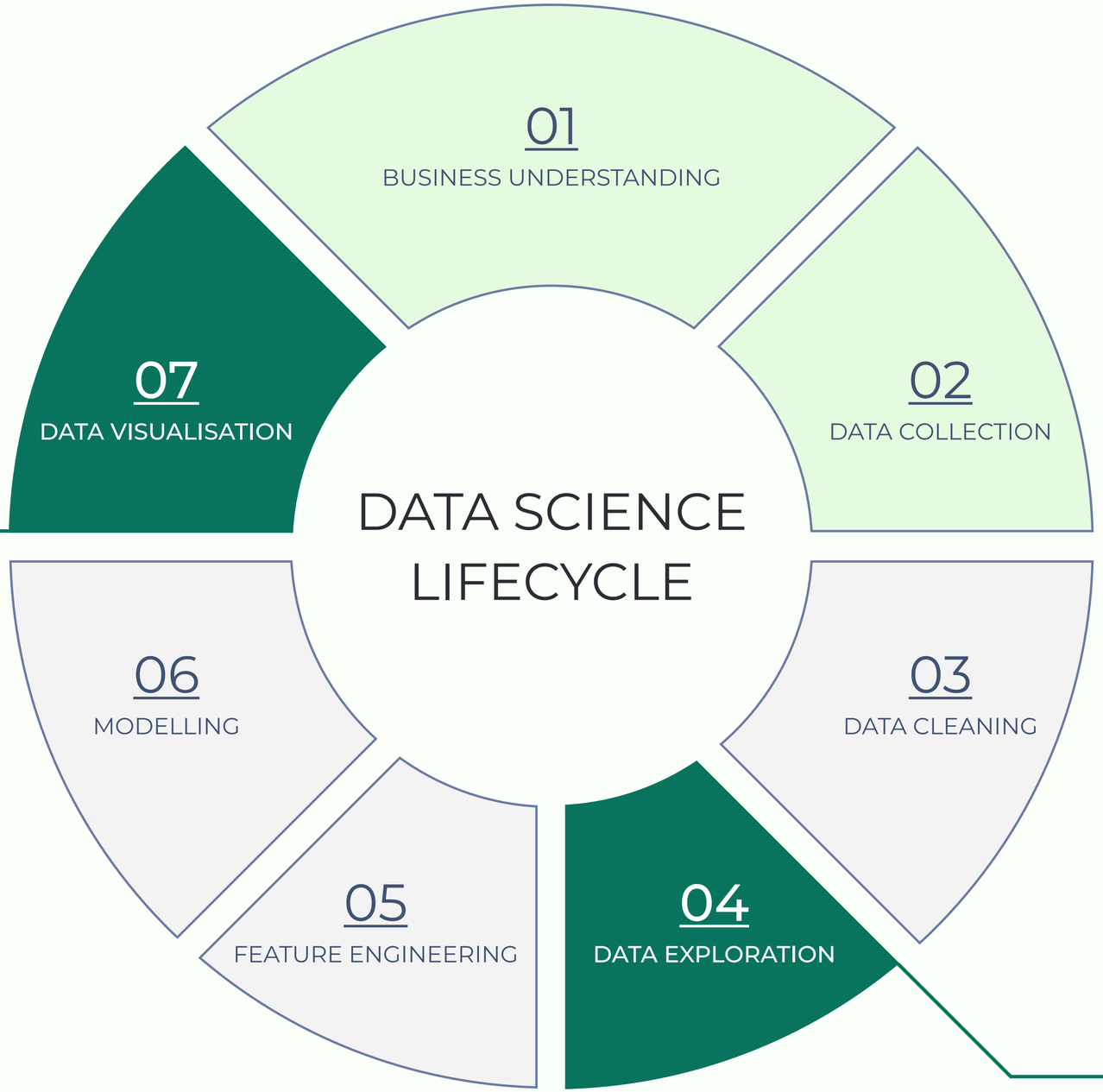
Prof. Juri Belikov

Department of Software Science
Tallinn University of Technology
juri.belikov@taltech.ee

PREVIOUSLY IN COURSE ...

Key takeaways:

- Data science life cycle
  - Business understanding
  - Data collection
  - Data cleaning
  - Data exploration
  - Feature engineering
  - Modelling
  - Data visualisation



Communicate the findings with key people using plots and interactive visualisations.

Convincing others it is true

Explore and visualise data to uncover insights from the data and form hypotheses.

Figuring out what is true

Typography

WHY TYPOGRAPHY MATTERS?

Typography **reinforces** the written message.

Although the world at large seems to be transitioning to other ways of communicating data, the written word is still the cornerstone of information exchange.

So at the very least, you should know the basics typography and why they matter.

A BRIEF ABOUT FONTS

Font (also *face*, *typeface*) is a set of characters including letter, punctuation marks, numbers and special symbols.

serif (Times New Roman)

Font Family is a group of typefaces sharing common features.

sans serif (Arial)

SERIF

Serif fonts have small lines or strokes attached to the larger strokes of the letters. Some examples:

Times New Roman
Charter
Palatino

serif (Times New Roman)

Serifs make these fonts easy to read. So for any large document use serif fonts for body text.

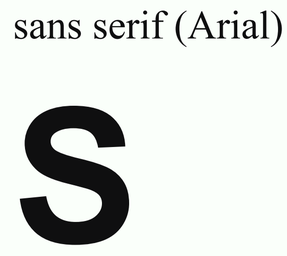
SANS SERIF

Sans serif fonts, on the other hand, do not have serifs which makes them legible even if typeset in very small point sizes. Some examples:

Calibri

Arial

Monserrat



Sans serif fonts are typically used on the web, in presentations, on billboards etc., where type must remain legible at any distance. But they do not work well for large manuscripts!

OTHER FONTS

Monospaced fonts originate from the use of typewriters. Nowadays they are mostly used for setting computer code. They should not be used for body text! Examples:

Courier New
Menlo

Other fonts-hand drawn, retro, novelty, comic and others which Matthew Butterick refers as “goofy fonts” should typically be avoided in professional documents.

[Accessed May 8, 2025], https://practicaltypography.com/goofy-fonts.html

NEVER!

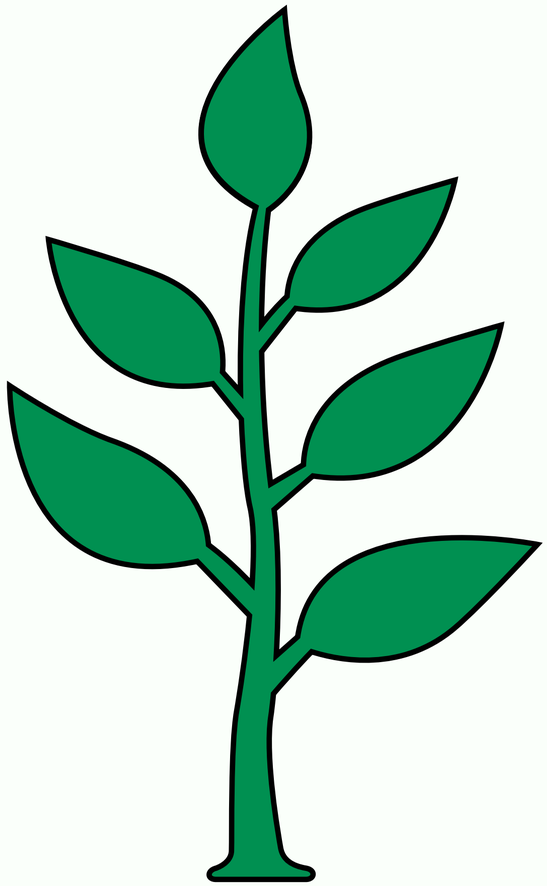
Graphics:

Vector vs Raster

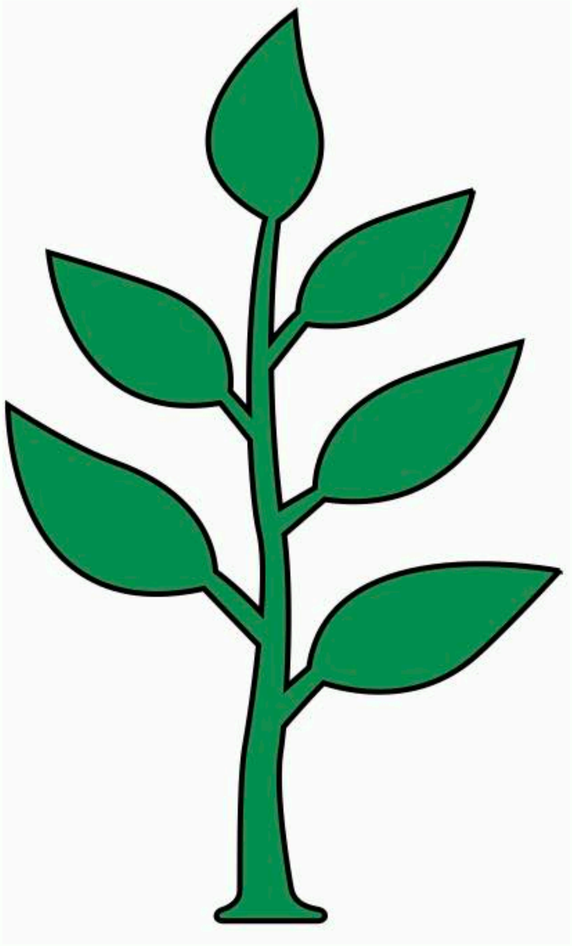
RASTER AND VECTOR: KEY DIFFERENCE

They may look similar from a distance ...

Vector



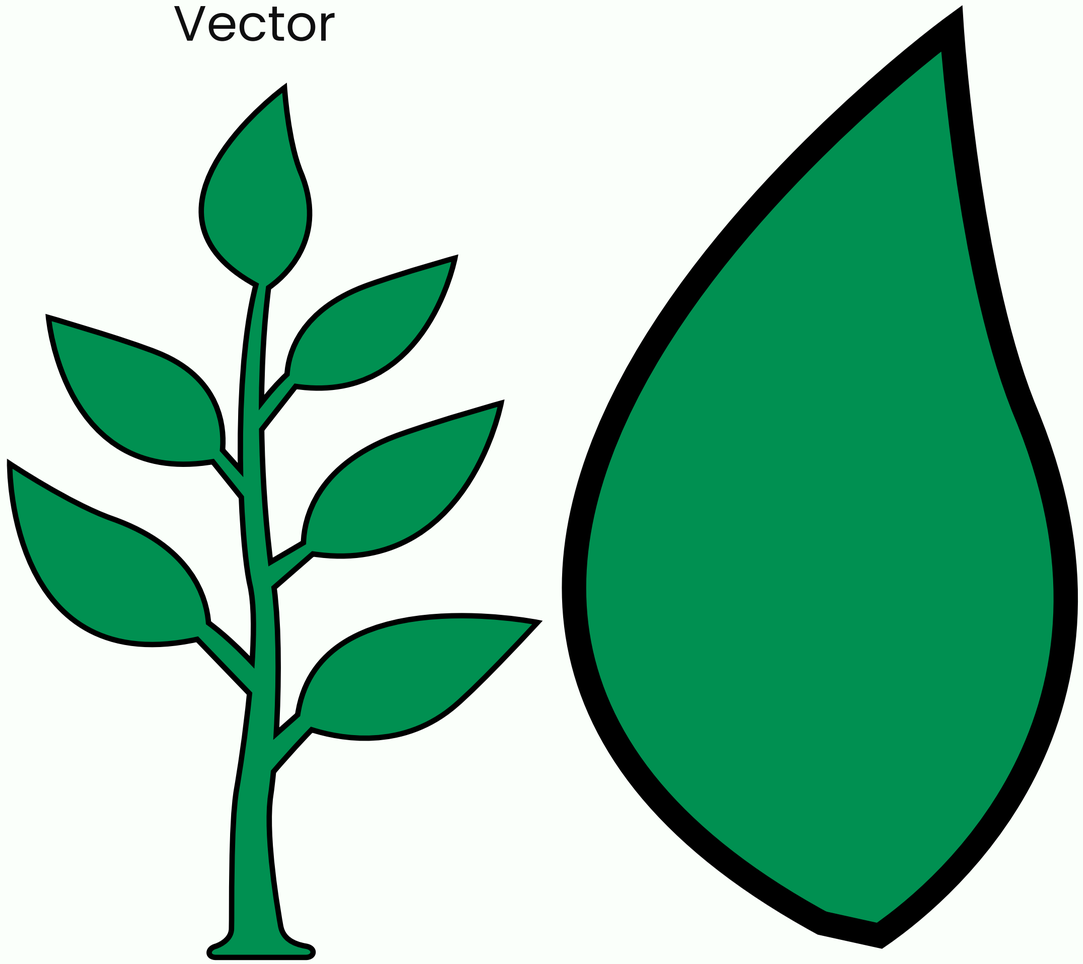
Raster



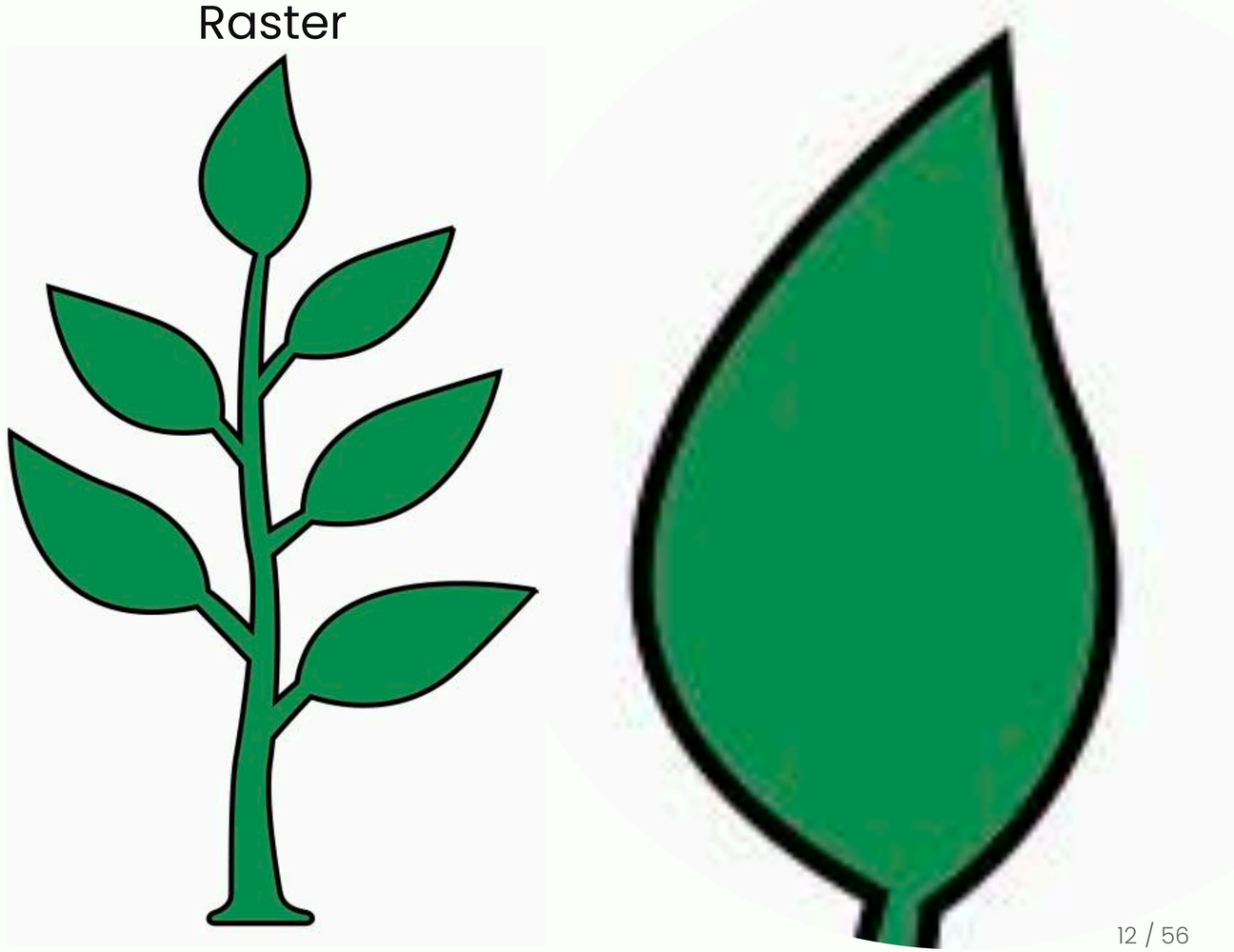
RASTER AND VECTOR: KEY DIFFERENCE

They may look similar from a distance ...

Vector



Raster



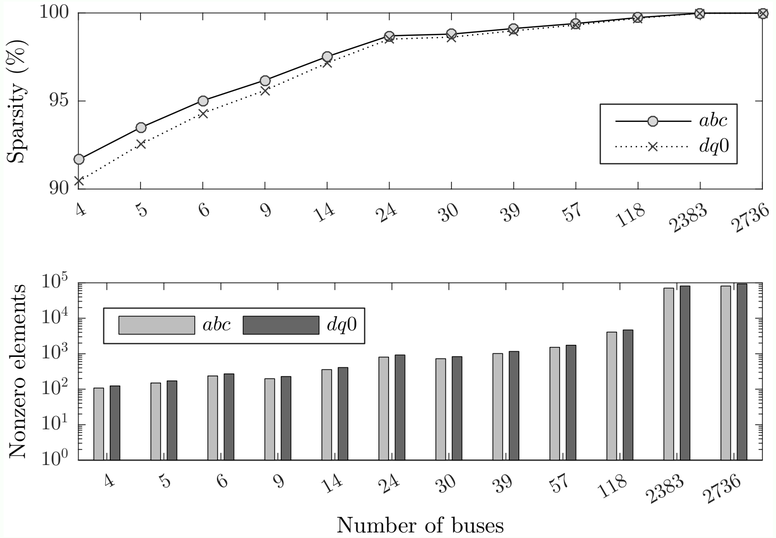
WHEN TO USE VECTOR IMAGES?

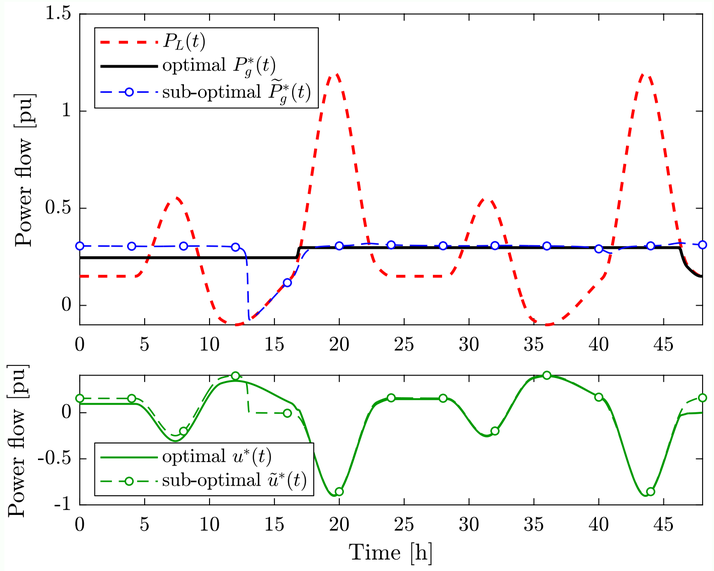
Vector images are based on information about shapes, their position, rotation, colour content, and relationships with other shapes.

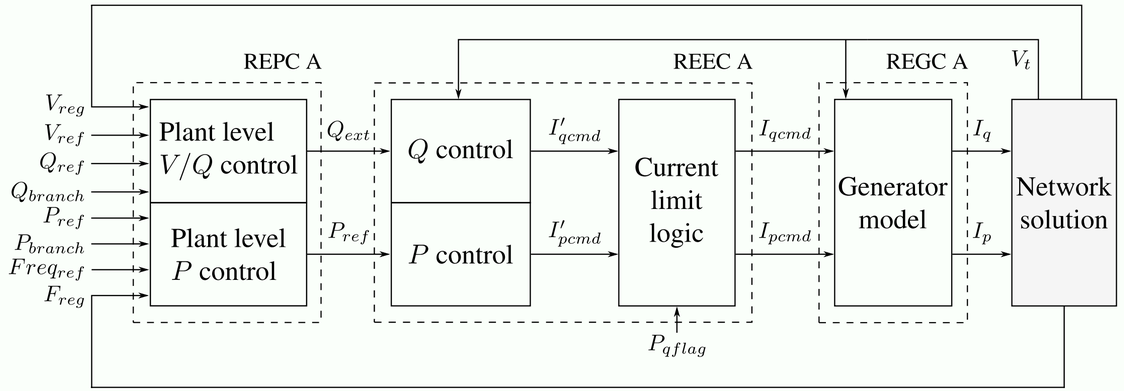
Vector images are often better because they can be made any size (which is good for screen reading) and take up much less storage space.

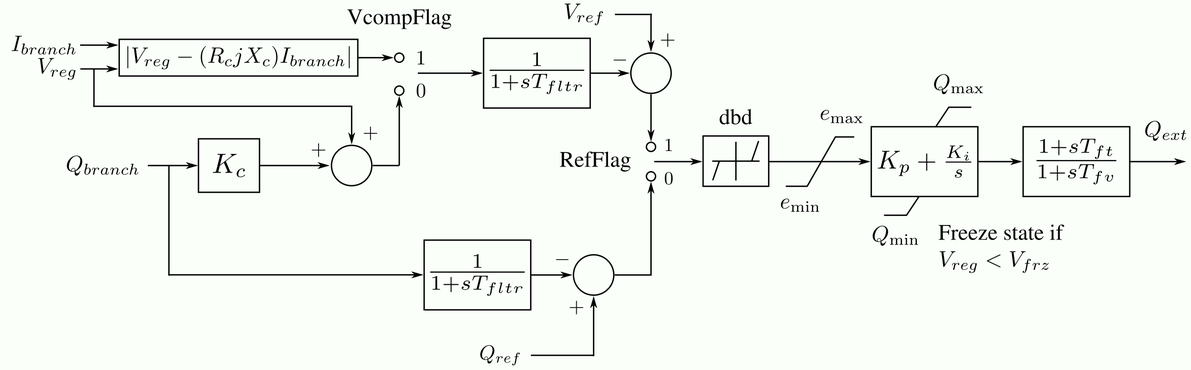
Examples: most scientific plots (like time series), diagrams, graphs, UML diagrams, various illustrations.

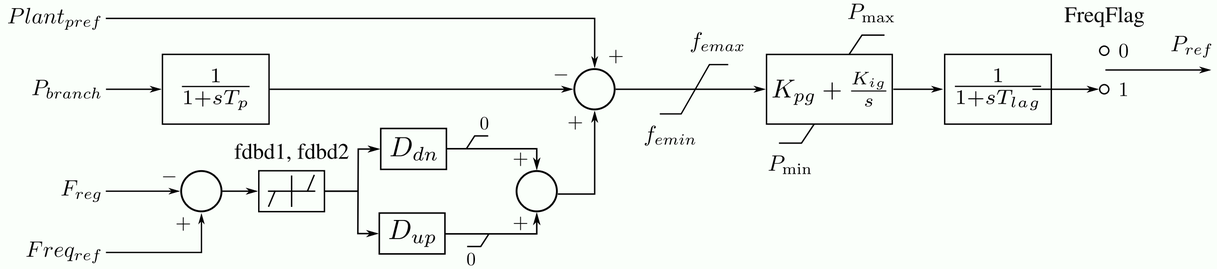
VECTOR IMAGES: EXAMPLES







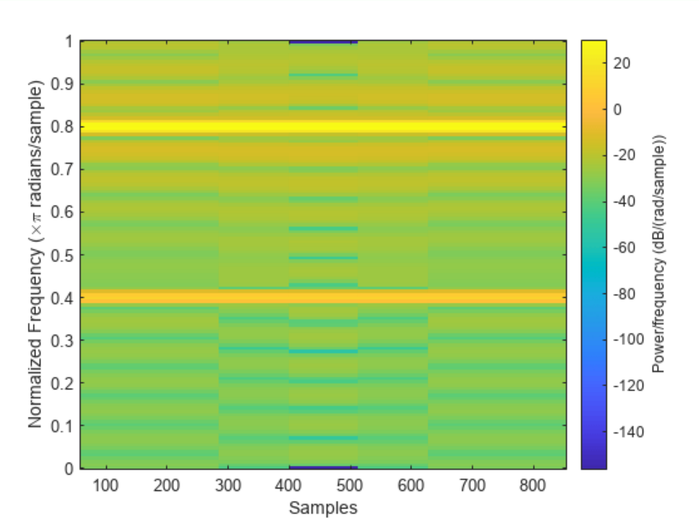


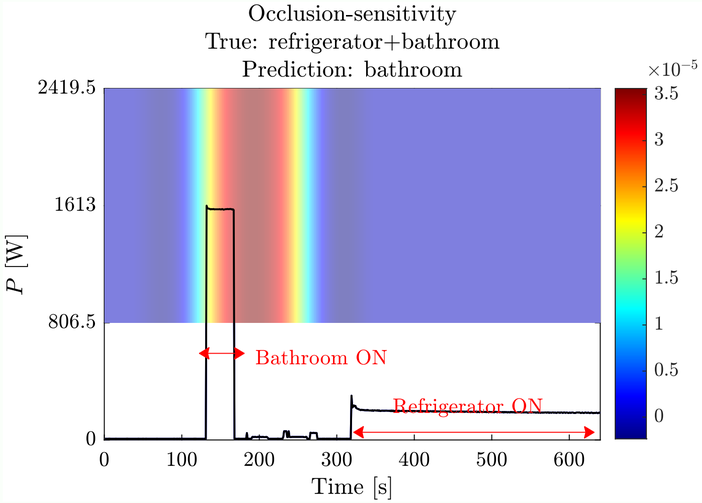


WHEN TO USE RASTER IMAGES?

Raster images are based on information about individual pixels. Thus, whenever images or illustrations cannot be efficiently represented otherwise, this is when you would use raster images.

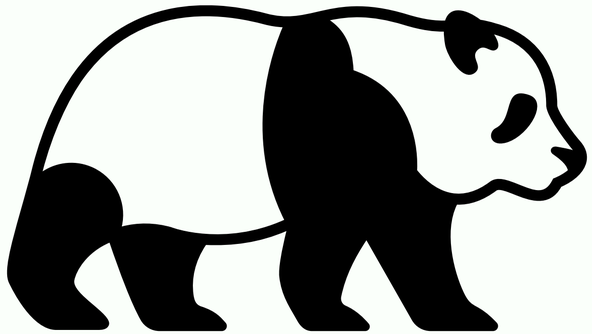
Examples: photographs, complicated plots (e.g., spectrograms, collages), logos (especially with gradients), and heat maps.



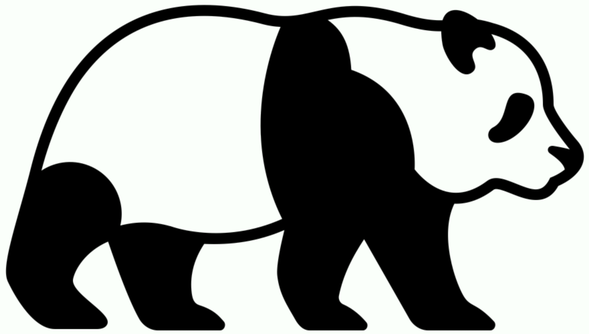


RASTER IMAGES: DPI MATTERS

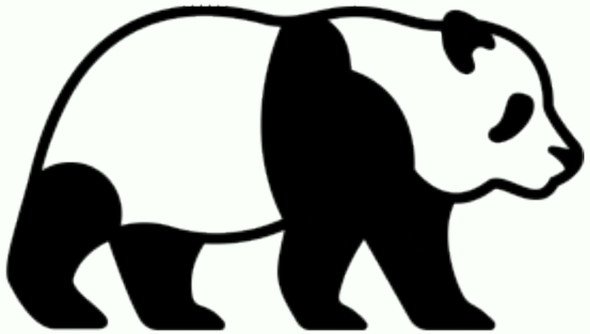
Vector



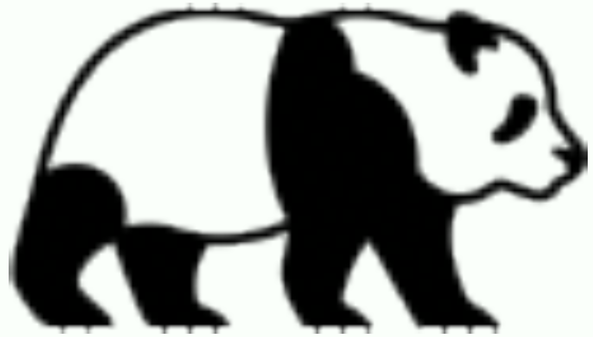
Raster 300 DPI



Raster 150 DPI



Raster 10 DPI



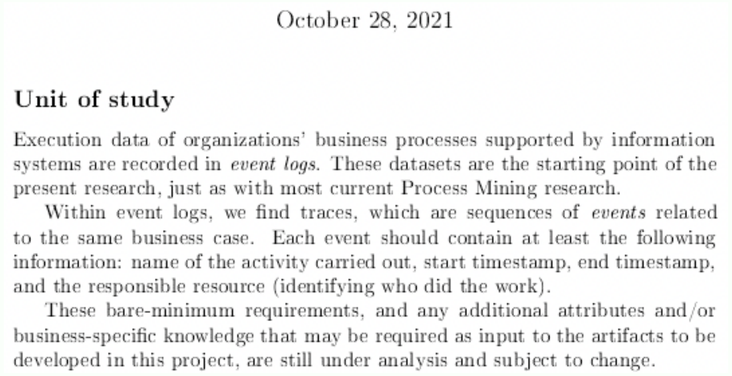
TOOLS TO USE

Professional typography: LyX/TeX/LaTeX. Can also use MS Word, but then you should know typography very well.

Vector drawing: pros use Adobe Illustrator, but there exist open source alternatives for different use cases: Inkscape for general vector drawing, Dia/yEd/draw.io for graphs, diagrams, UML, etc.

USE THEM RIGHT: 100% ZOOM

But it is not very visible on big screens. You want to zoom in



October 28, 2021

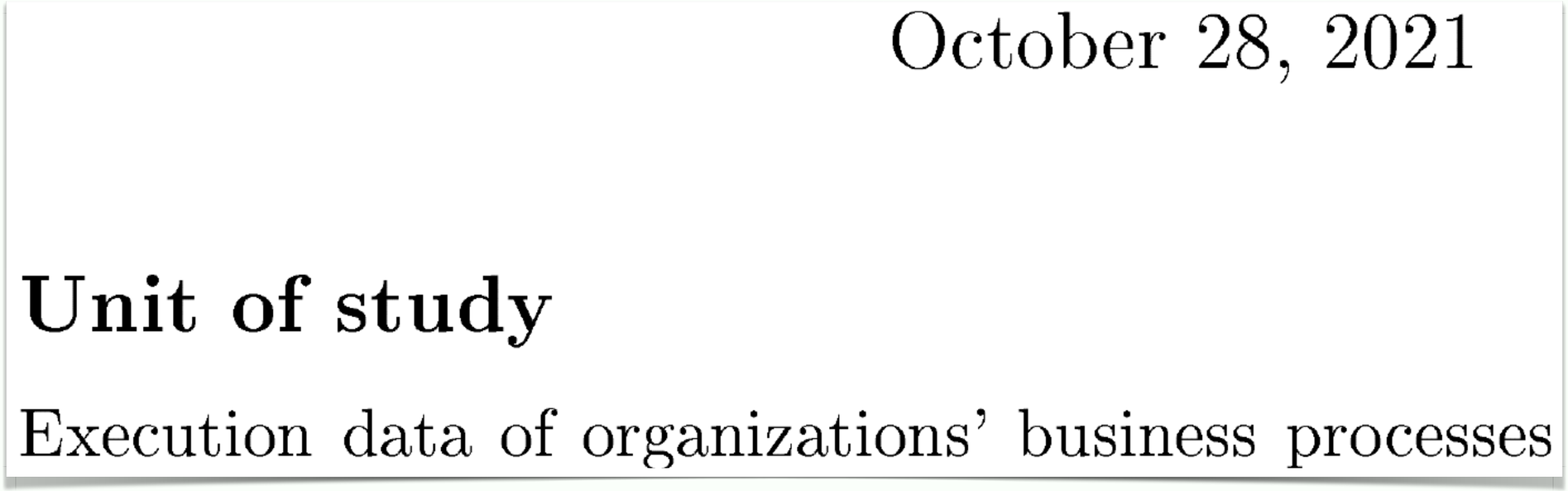
Unit of study

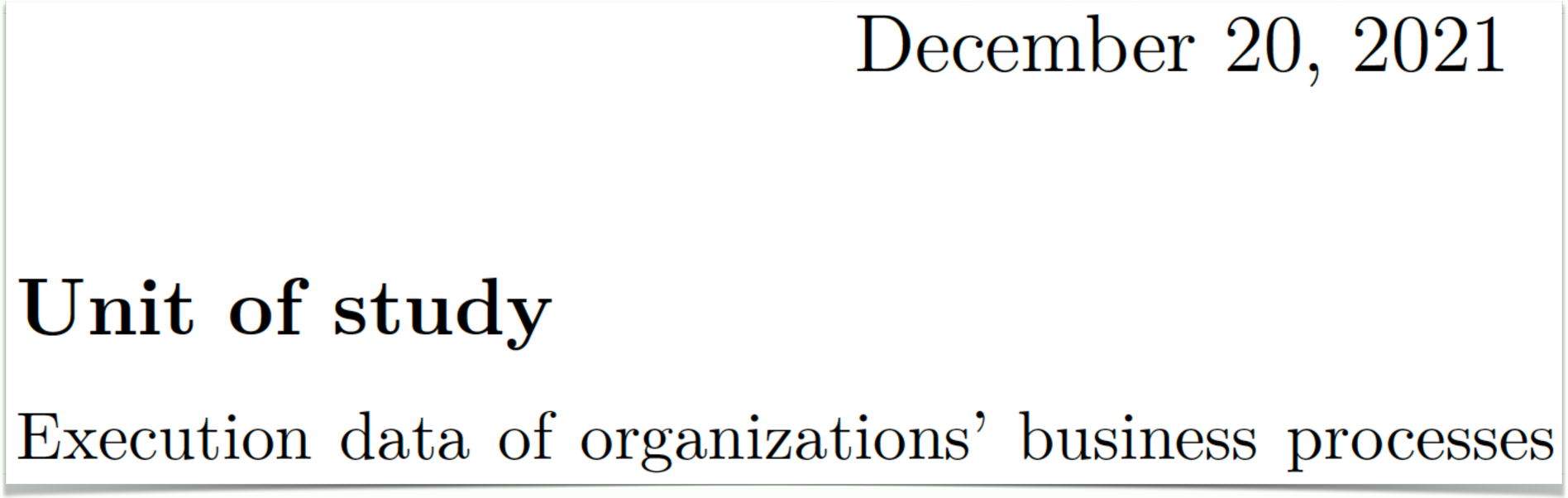
Execution data of organizations' business processes supported by information systems are recorded in *event logs*. These datasets are the starting point of the present research, just as with most current Process Mining research.

Within event logs, we find traces, which are sequences of *events* related to the same business case. Each event should contain at least the following information: name of the activity carried out, start timestamp, end timestamp, and the responsible resource (identifying who did the work).

These bare-minimum requirements, and any additional attributes and/or business-specific knowledge that may be required as input to the artifacts to be developed in this project, are still under analysis and subject to change.

USE THEM RIGHT: ZOOMED-IN





MORAL

When you begin working with LyX/LaTeX, please install the **cm-super** package. You will then be able to use a vector-based Computer Modern font instead of the default bitmap one.

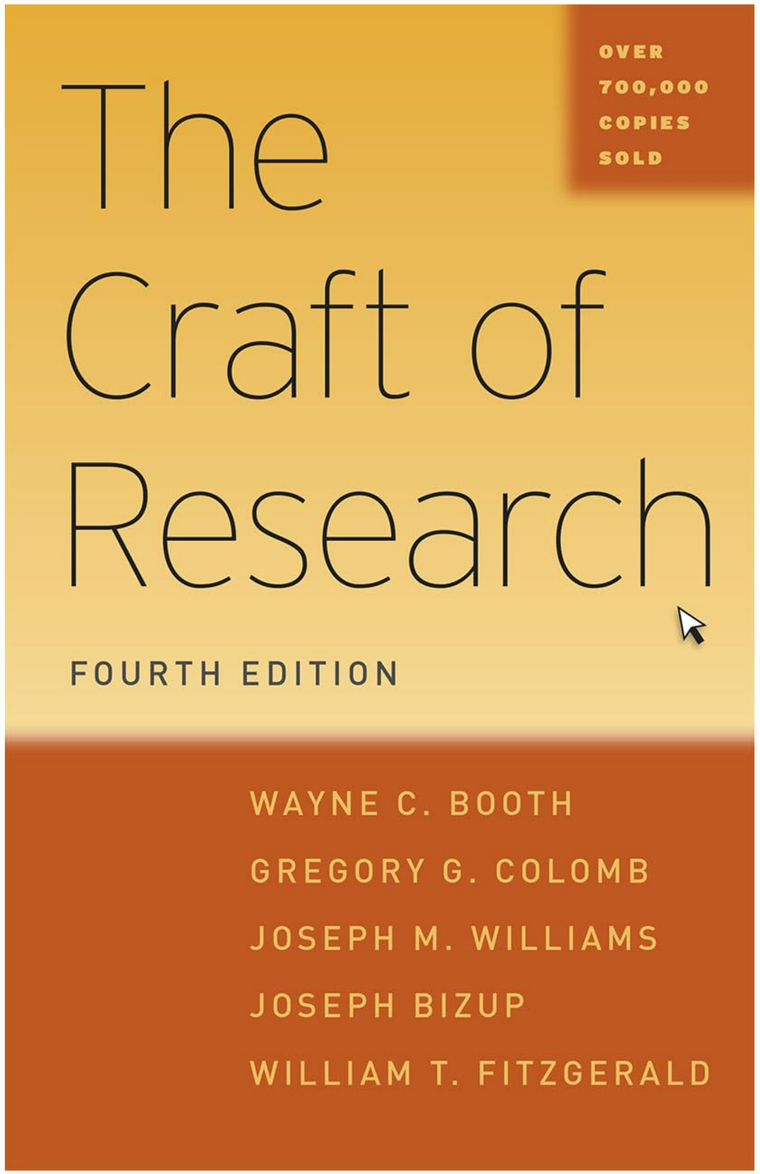
Note the reason for the bitmap font: TeX was originally intended for typesetting printed documents. However, nowadays we mostly read information on the screen ...

FURTHER READING

Please check out the excellent online book written and published by Matthew Butterick:

https://practicaltypography.com/

Data Visualisation



Further material is based on

“The Craft of Research”

by W. C. Booth, G. G. Colomb, J. M. Williams, J. Bizup, W. T. Fitzgerald, University of Chicago Press, Fourth edition, 2016

WHAT IS A DATA VISUALISATION?

Data visualisation is the representation of data information.

Traditionally graphics are divided into tables and figures.

- A table is a grid with columns and rows.
- Figures are all other graphic forms, including graphs, charts, photographs, drawings, and diagrams.

Figures that present quantitative data are divided into charts and graphs.

- Charts typically consist of bars, circles, points, or other shapes.
- Graphs consist of continuous lines.

WHY DO WE NEED DATA VISUALISATION?

Most readers grasp quantitative evidence more easily in tables, charts, and graphs than they do in words.

But some visual forms suit particular data and messages better than others.

VISUAL VS VERBAL?

Simple example

VERBAL:

In June 2023, Estonia imported 533 GWh of electricity and exported 384 GWh, resulting in a net import of 149 GWh.

VISUAL:

Import vs Export of electricity (GWh), June 2023

Import	533
Export	384
Net import	149

[Accessed May 8, 2025], https://andmed.stat.ee/en/stat/majandus__energeetika__energia-tarbimine-ja-tootmine__luhiajastatistika/KE21

VISUAL VS VERBAL: TEXT

Between 1995 and 2020, the final electricity consumption landscape in Estonia underwent significant changes across various sectors. Industrial consumption started at 38.43 percent in 1995 and gradually declined to 29.65 percent in 2020. Transport experienced a decrease until 2015, reaching a low of 0.69 percent, and then slightly increased to 0.95 percent in 2020. Residential consumption saw several fluctuations, with percentages ranging from 23.42 percent in 1995 to 27.83 percent in 2020. Commercial and Public services had a consistent increase until 2015, reaching 41.05 percent, followed by a decrease to 39.64 percent in 2020. Other consumption, including Agriculture, Forestry, Fishing, etc., followed a general trend of decrease, from 8.03 percent in 1995 to 1.94 percent in 2020.

Too long to be read ...

https://www.iea.org/reports/world-energy-balances-overview

VISUAL VS VERBAL: TABLES

To describe the data above the most common choices are tables, bar charts, and line graphs, each of which has a distinctive rhetorical effect.

A table seems precise and objective, but requires readers to infer relationships or trends on their own.

Share of different sectors in final electricity consumption in Estonia [%]

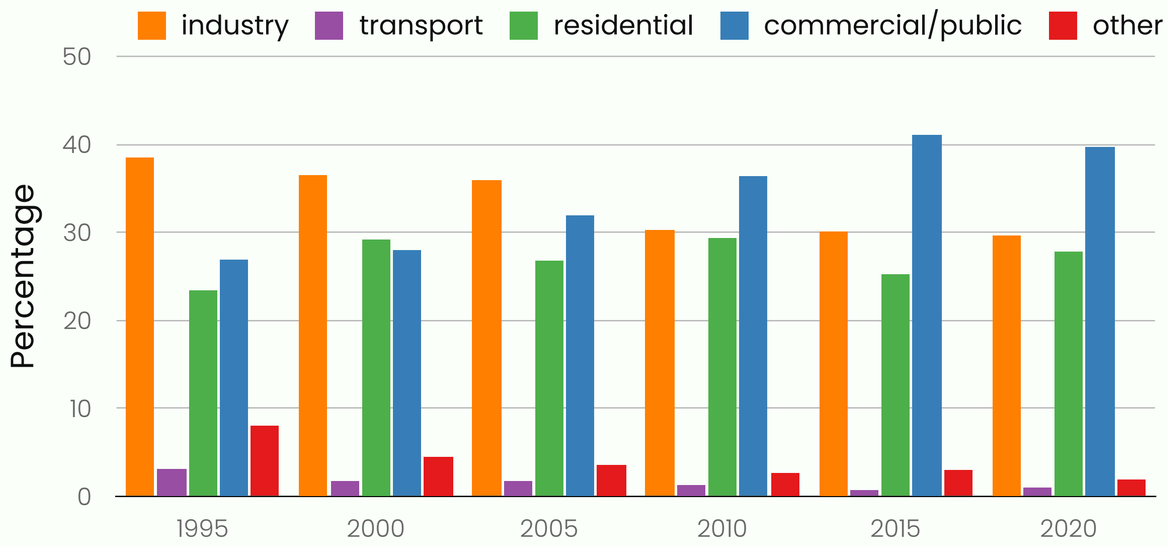
Sector	1995	2000	2005	2010	2015	2020
industry	38.43	36.55	35.89	30.33	30.05	29.65
transport	3.2	1.81	1.71	1.29	0.69	0.95
residential	23.42	29.23	26.82	29.28	25.22	27.83
commercial/public	26.91	27.94	31.94	36.35	41.05	39.64
other	8.03	4.47	3.64	2.75	2.99	1.94

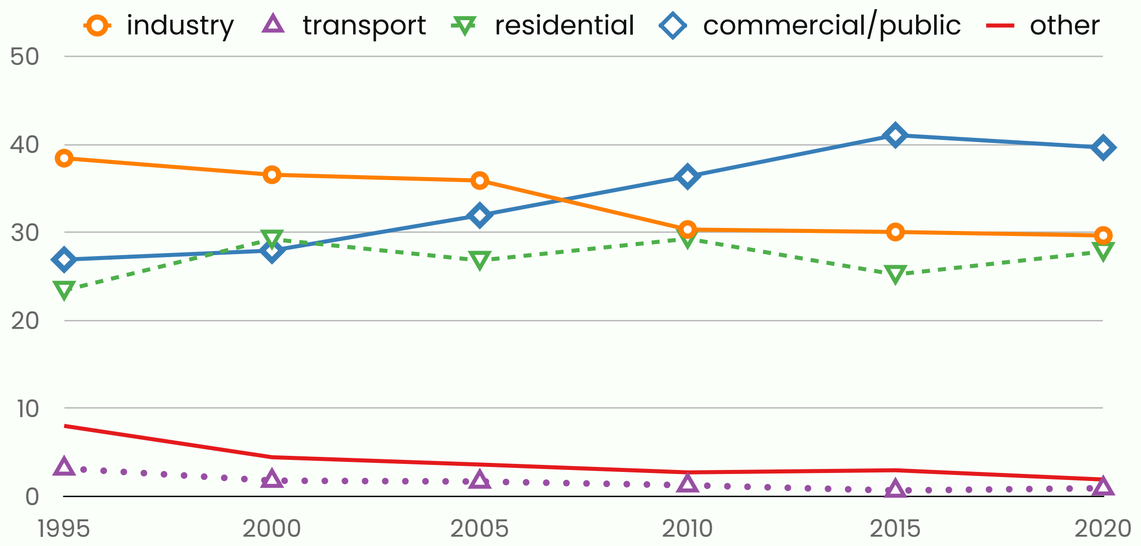
VISUAL VS VERBAL: CHARTS AND GRAPHS

Charts and line graphs present a visual image that communicates values less precisely than do the exact numbers of a table but with more impact.

But charts and graphs also differ.

A bar chart emphasises contrasts among discrete items, while a line graph suggests continuous change over time.





DESIGNING TABLES, CHARTS, AND GRAPHS

Readers do not care how fancy a graphic looks if it does not communicate your point **clearly**.

Frame each graphic to help your reader understand it:

- Label graphic in a way that describes its data
- Make the title or legend descriptive
- Do not include conclusions
- Insert into the table or figure information that helps readers see how the data support your point
- Introduce the table or figure with a sentence that explains how to interpret it
- Keep it simple
- Do not overreact with colours

SPECIFIC GUIDELINES: TABLES

Order the rows and columns by a principle that lets readers quickly and what you want them to see.

Round numbers to a relevant value.

Renewable energy consumption as % of total energy consumption in European economies, 2010-2020			
Country Name	2010	2020	Change
Czechia	10.95	16.97	6.02
Estonia	25.32	40	14.68
Hungary	13.46	14.76	1.3
Latvia	33.07	43.75	10.68
Lithuania	21.46	31.7	10.24
Poland	9.49	16.14	6.65
Slovak Republic	10.28	17.64	7.36
Slovenia	20.07	22.4	2.33

Change in renewable energy adoption as a percentage of total energy consumption in European economies, 2010-2020			
Baltic states vs. Central European states			
Country Name	2010	2020	Change
Estonia	25.32	40	14.68
Latvia	33.07	43.75	10.68
Lithuania	21.46	31.7	10.24
Slovak Republic	10.28	17.64	7.36
Poland	9.49	16.14	6.65
Czechia	10.95	16.97	6.02
Slovenia	20.07	22.4	2.33
Hungary	13.46	14.76	1.3

[Accessed May 8, 2025], https://databank.worldbank.org/source/world-development-indicators#
Series Code: EG.FEC.RNEW.ZS

SPECIFIC GUIDELINES: TABLES

Order the rows and columns by a principle that lets readers quickly and what you want them to see.

Round numbers to a relevant value.

Renewable energy consumption as % of total energy consumption in European economies, 2010-2020			
Country Name	2010	2020	Change
Czechia	10.95	16.97	6.02
Estonia	25.32	40	14.68
Hungary	13.46	14.76	1.3
Latvia	33.07	43.75	10.68
Lithuania	21.46	31.7	10.24
Poland	9.49	16.14	6.65
Slovak Republic	10.28	17.64	7.36
Slovenia	20.07	22.4	2.33

sorted

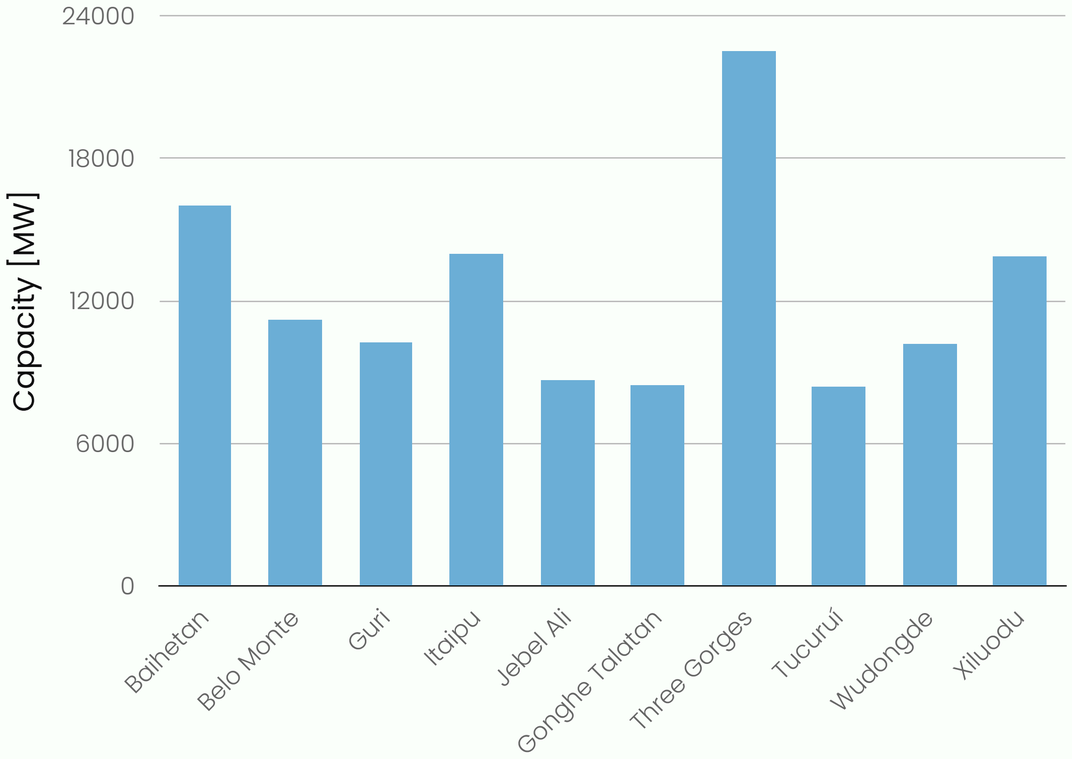
Change in renewable energy adoption as a percentage of total energy consumption in European economies, 2010-2020			
Baltic states vs. Central European states			
Country Name	2010	2020	Change
Estonia	25.32	40	14.68
Latvia	33.07	43.75	10.68
Lithuania	21.46	31.7	10.24
Slovak Republic	10.28	17.64	7.36
Poland	9.49	16.14	6.65
Czechia	10.95	16.97	6.02
Slovenia	20.07	22.4	2.33
Hungary	13.46	14.76	1.3

[Accessed May 8, 2025], https://databank.worldbank.org/source/world-development-indicators#
Series Code: EG.FEC.RNEW.ZS

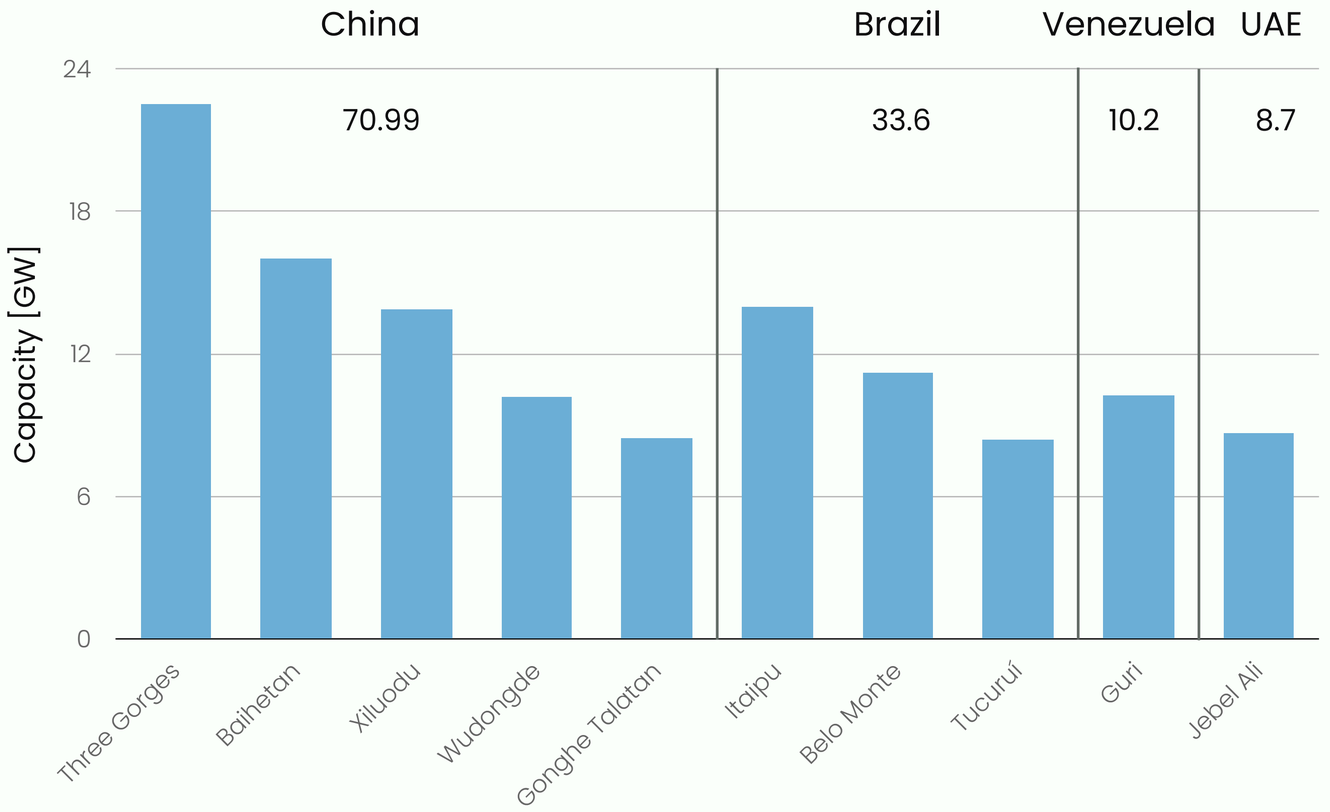
SPECIFIC GUIDELINES: BAR CHARTS

Bar charts communicate as much by visual impact as by specific numbers.

Example: Most of the largest power plants are located in China and Brazil.



Ten largest power plants by capacity



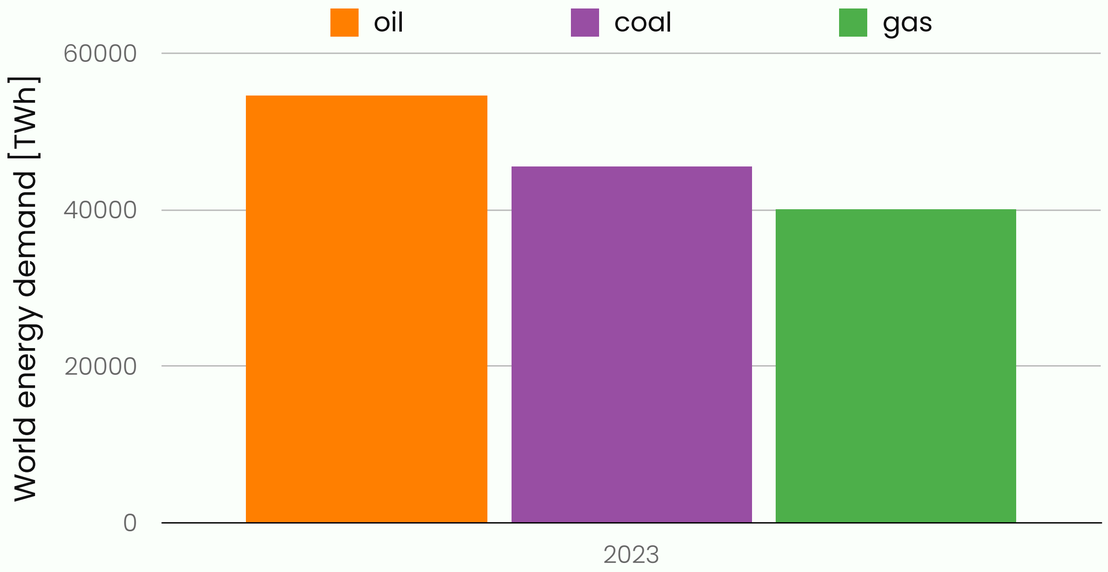
Distribution of the largest power plants by countries

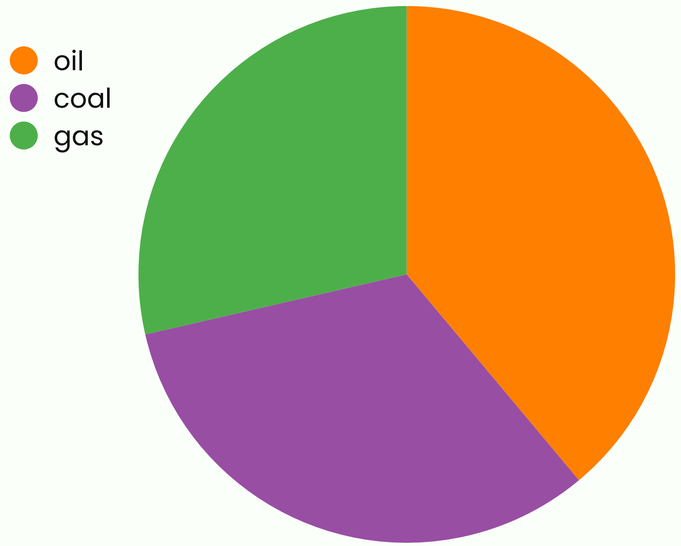
[Accessed May 8, 2025], https://en.wikipedia.org/wiki/List_of_largest_power_stations

BAR CHARTS VS PIE CHARTS

Most data that fit a bar chart fit in a pie chart. But while pie charts are popular in magazines, they are harder to read than bar charts and invite misinterpretation. Readers must mentally compare proportions of segments whose size is hard to judge in the first place.

Use bar charts instead.

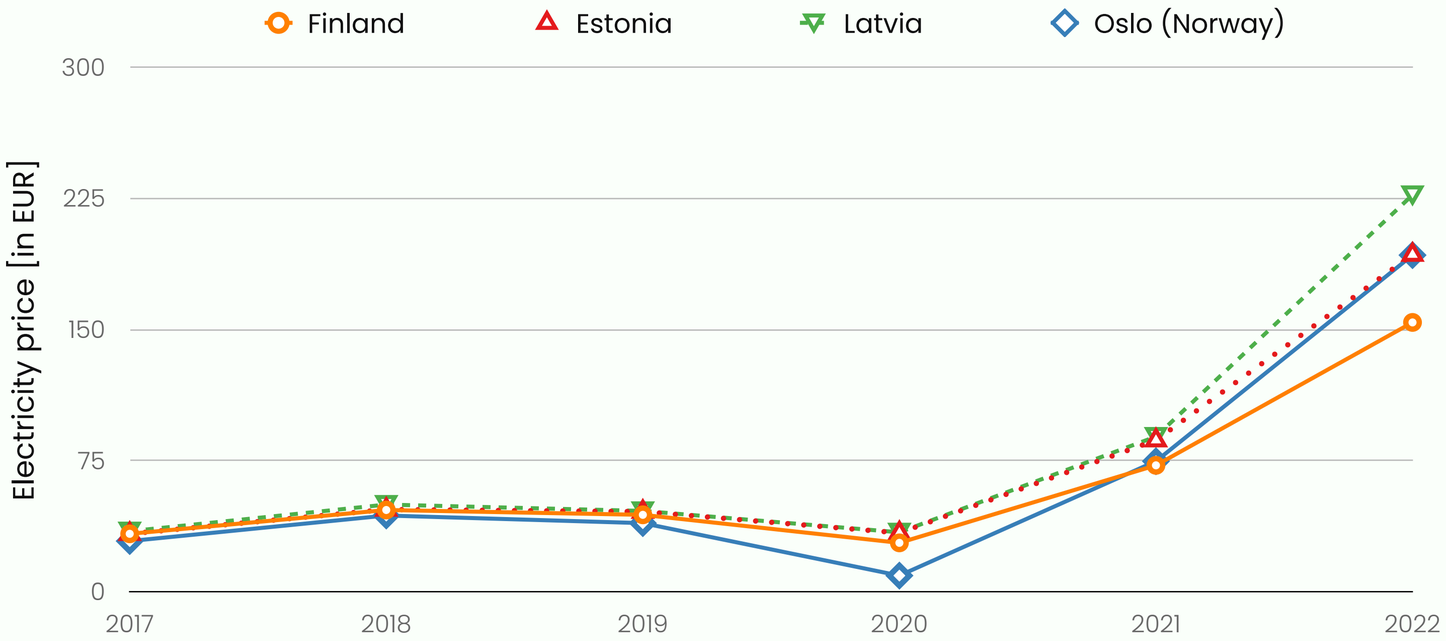




SPECIFIC GUIDELINES: LINE GRAPHS

Because a line graph emphasises trends, readers must see a clear image to interpret it correctly. Tips:

- ✓ Choose the variable that makes the line go in the direction, up or down, that supports your point.
- ✓ Plot more than six lines on one graph only if you cannot make your point in any other way.
- ✓ If you have fewer than ten or so data points, indicate them with dots. If only a few are relevant, insert numbers to show their exact value.



SPECIFIC GUIDELINES: ADVICE

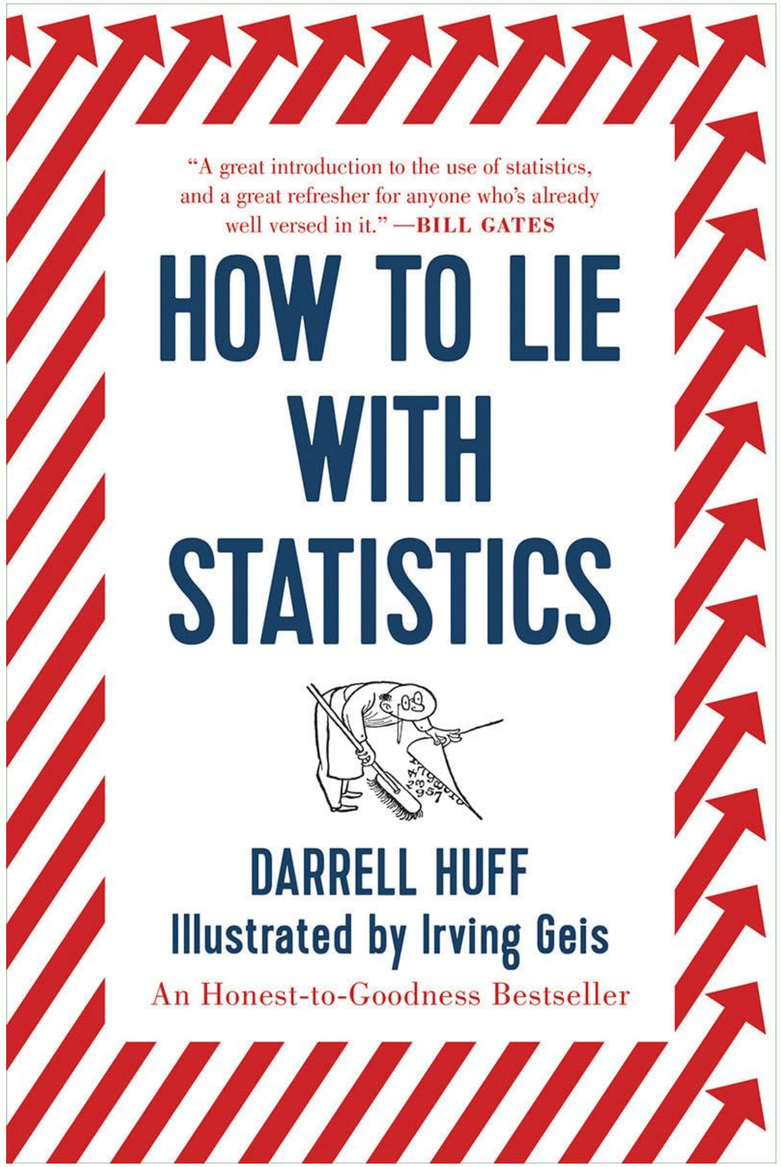
Different ways of showing the same data can be confusing.

To cut through that confusion, test different ways of representing the same data.

Construct alternative graphics; then ask someone unfamiliar with the data to judge them for impact and clarity.

Be sure to introduce the figures with a sentence that states the claim you want the figure to support.

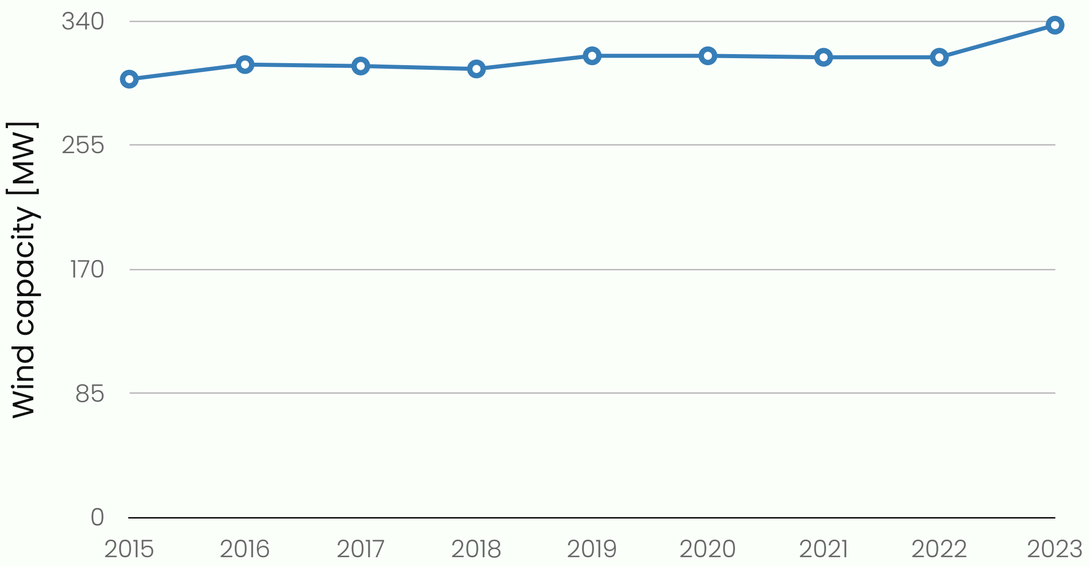
Communicating data ethically

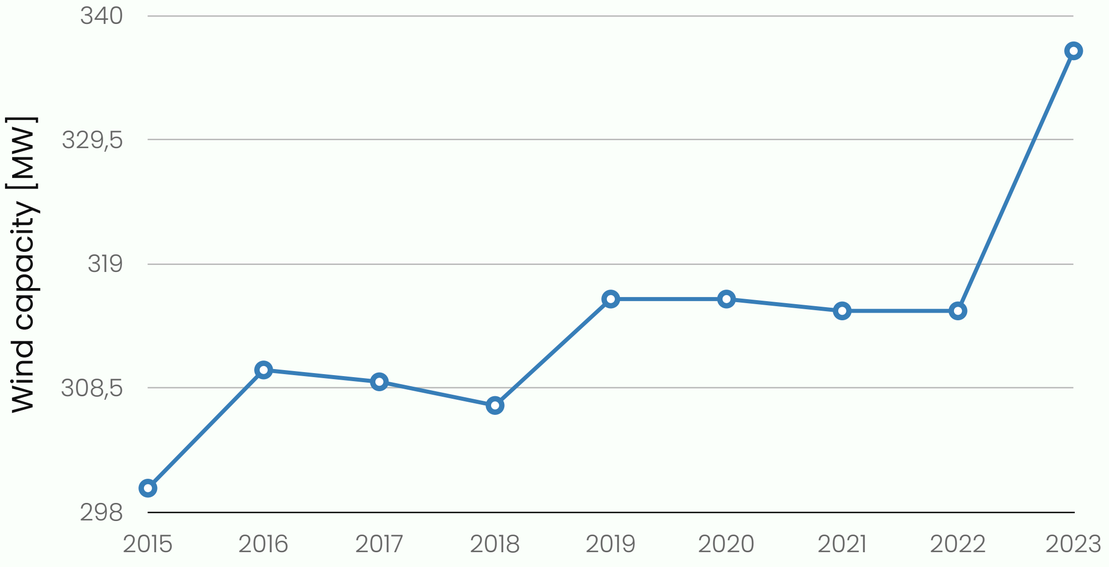


EXAMPLE

Graphics must be not only clear and accurate, but honest.

Total available capacity of wind plants in Estonia [MW], since 2015.



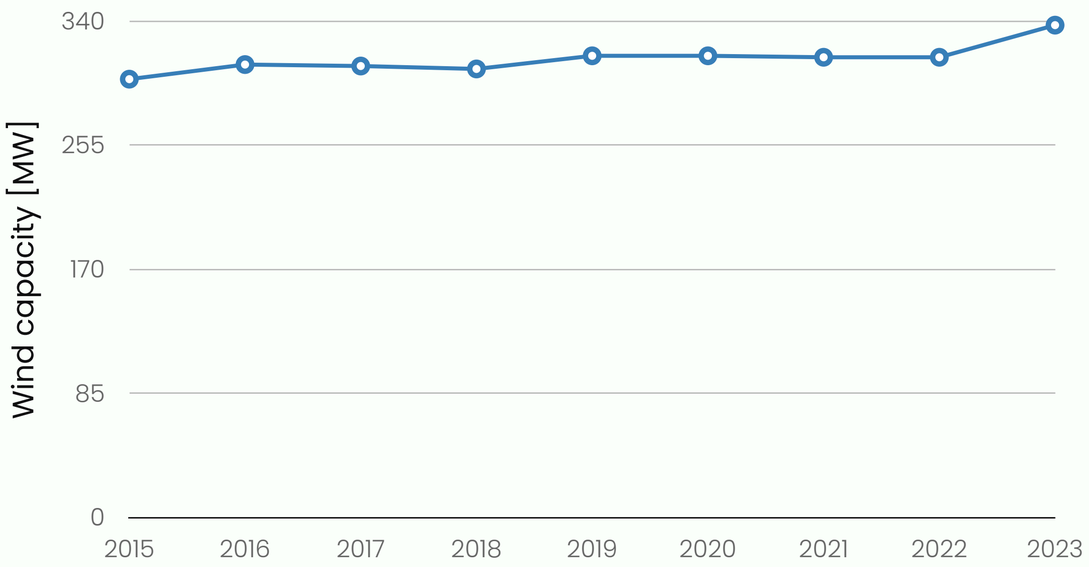


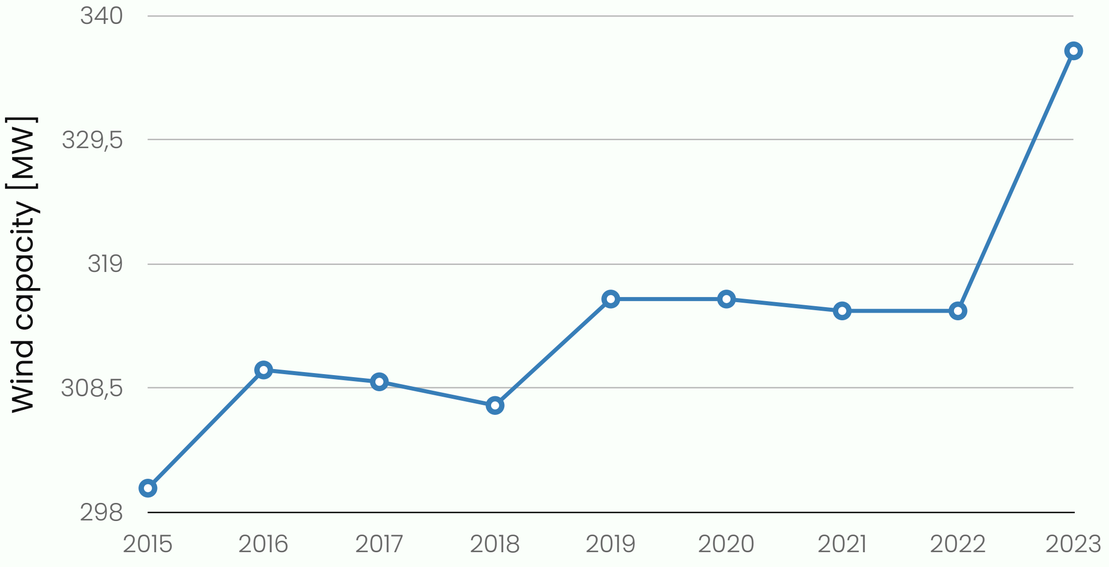
[Accessed September 8, 2025], https://andmed.stat.ee/en/stat/majandus__energeetika__energia-tarbimine-ja-tootmine__aastastatistika/KE032

EXAMPLE

Graphics must be not only clear and accurate, but honest.

Total available capacity of wind plants in Estonia [MW], since 2015.





NB! Left figure 0–320 scale: flat slope

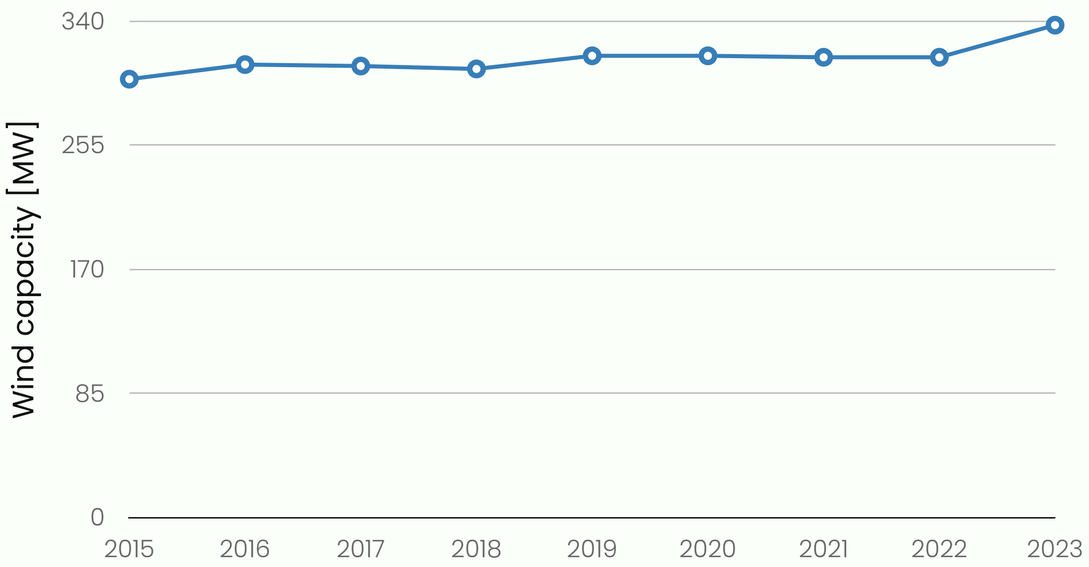
Right figure 290–320: more distinct slope

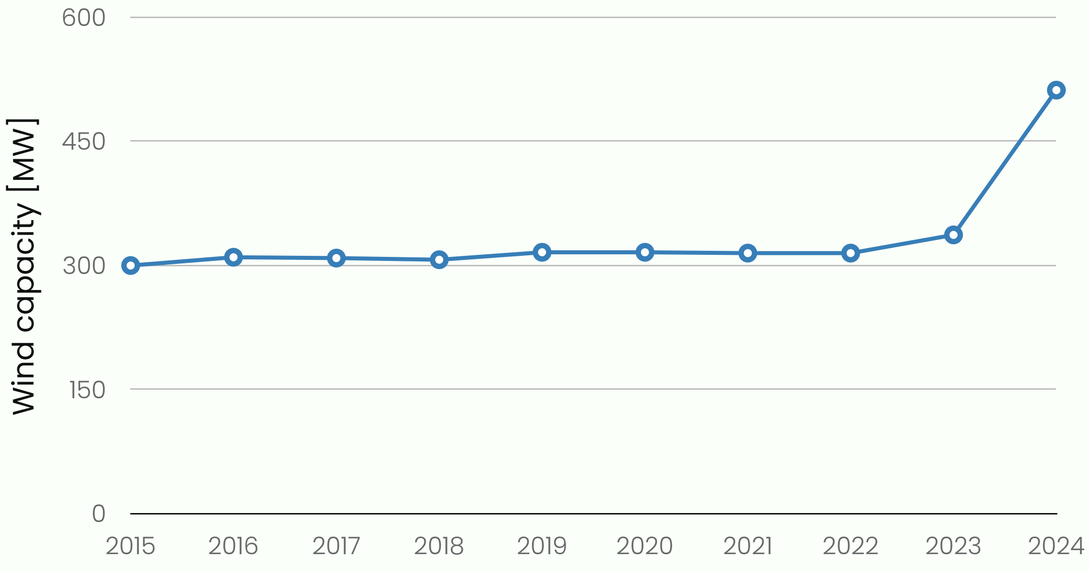
[Accessed September 8, 2025], https://andmed.stat.ee/en/stat/majandus__energeetika__energia-tarbimine-ja-tootmine__aastastatistika/KE032

EXAMPLE

Graphics must be not only clear and accurate, but honest.

Total available capacity of wind plants in Estonia [MW], since 2015.



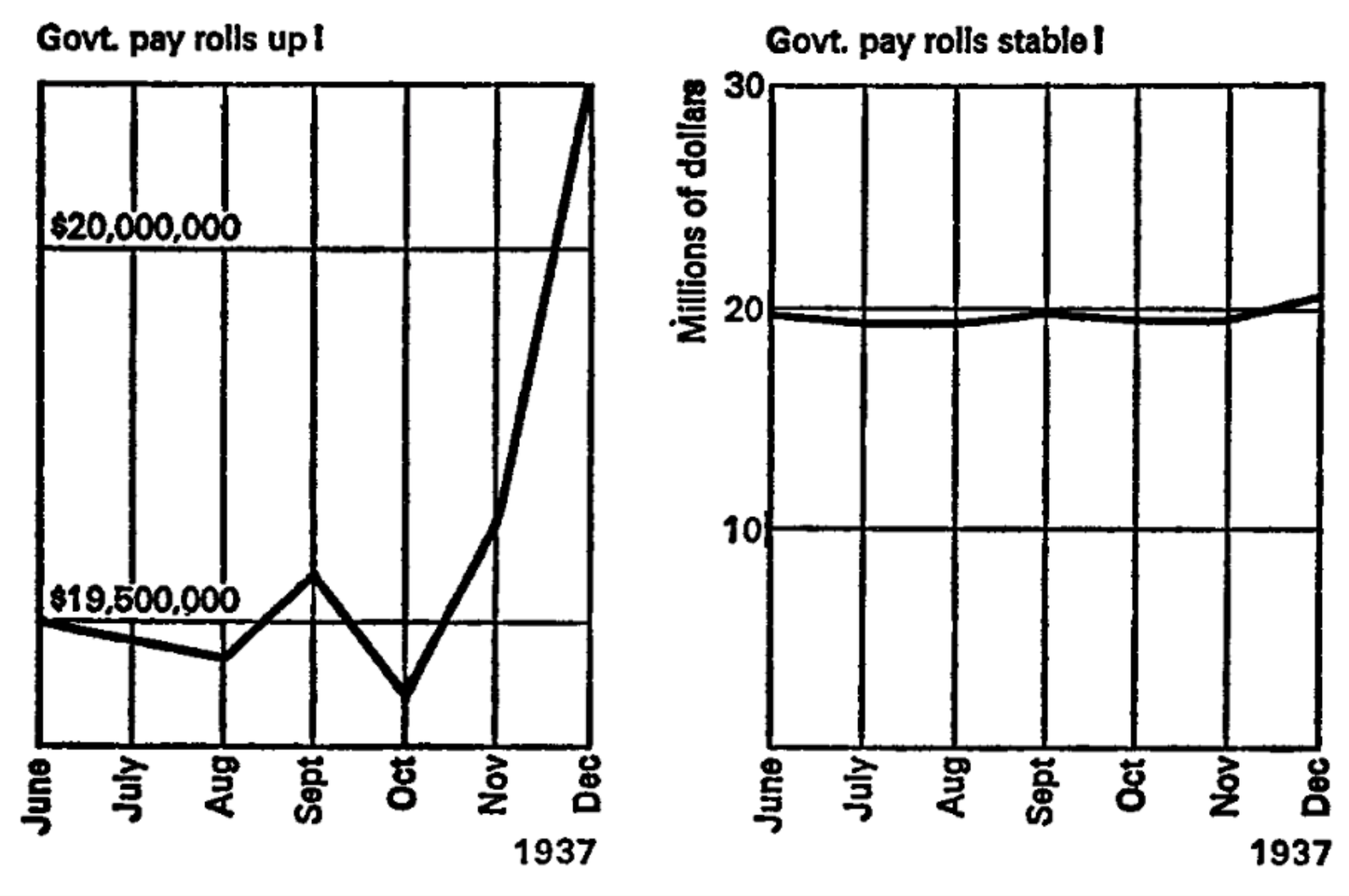


NB! Left figure 0–320 scale: flat slope

Right figure 290–320: more distinct slope

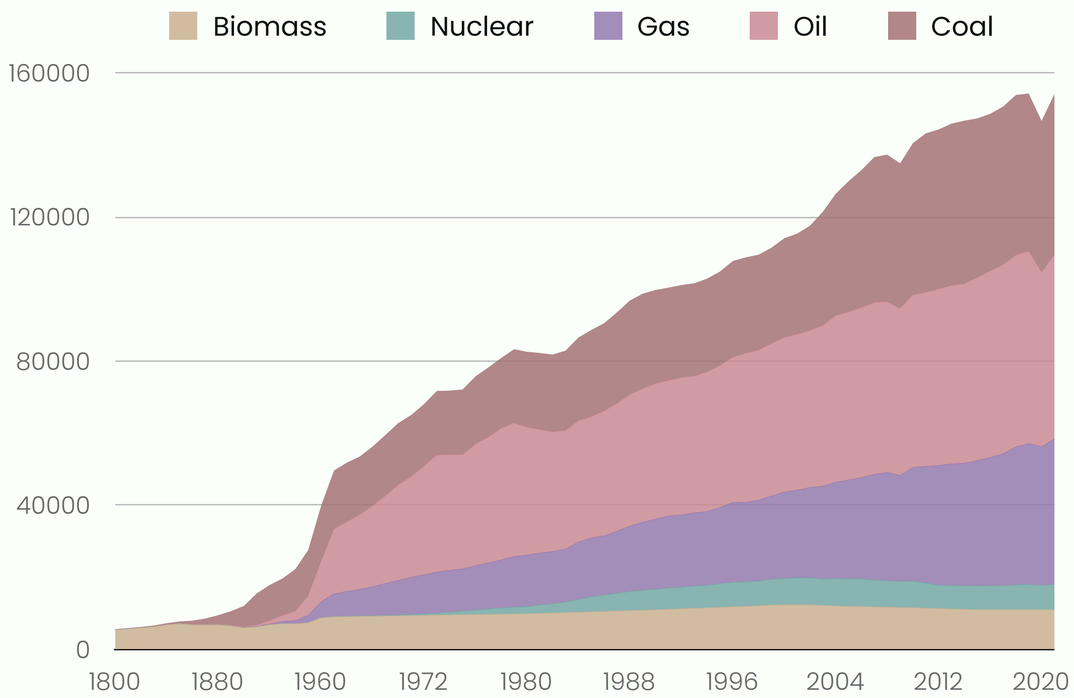
[Accessed September 8, 2025], https://andmed.stat.ee/en/stat/majandus__energeetika__energia-tarbimine-ja-tootmine__aastastatistika/KE032

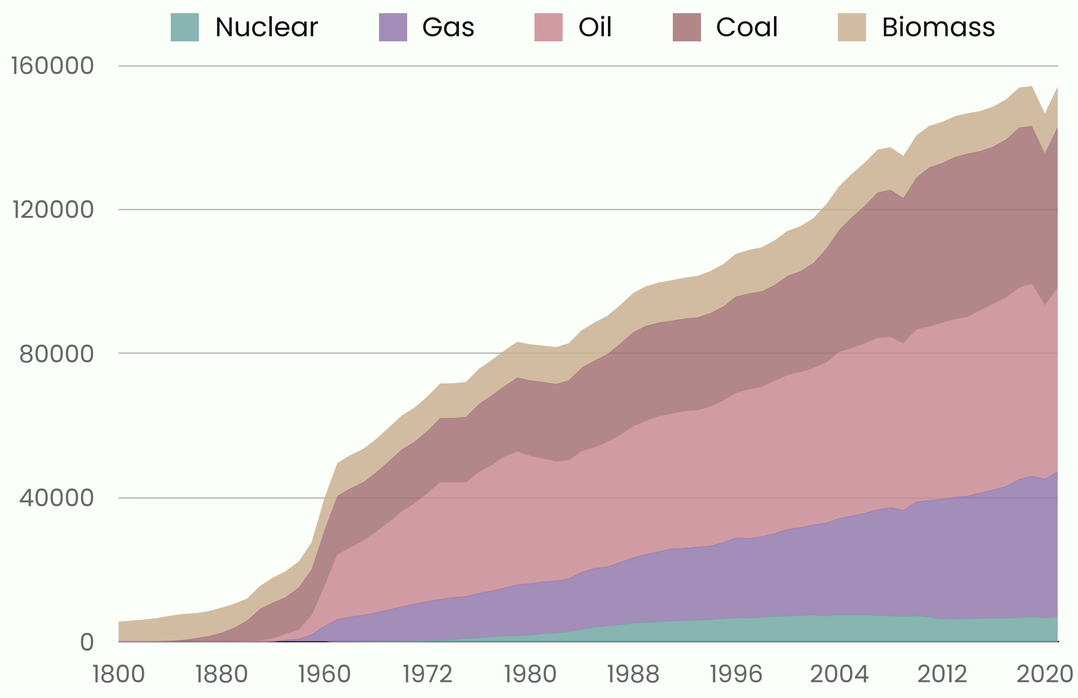
EXAMPLE (2)


Govt. pay rolls up!
$20,000,000
$19,500,000
June
July
Aug
Sept
Oct
Nov
Dec
1937
Govt. pay rolls stable!
Millions of dollars
30
20
10
June
July
Aug
Sept
Oct
Nov
Dec
1937

EXAMPLE (3)

Global primary energy demand by source [TWh].

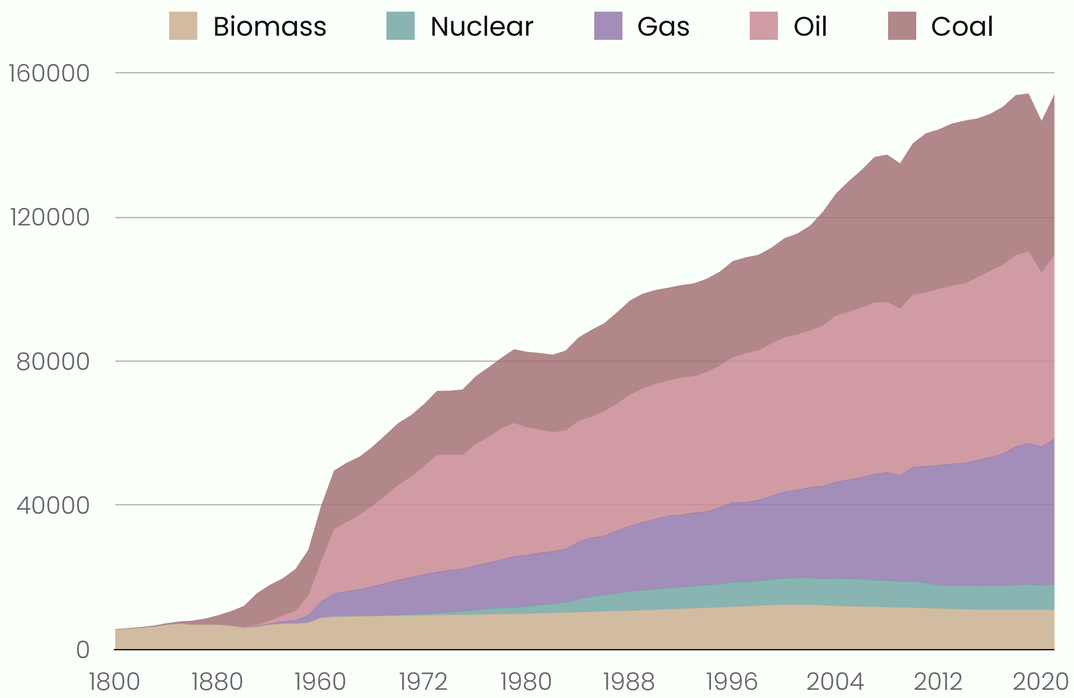


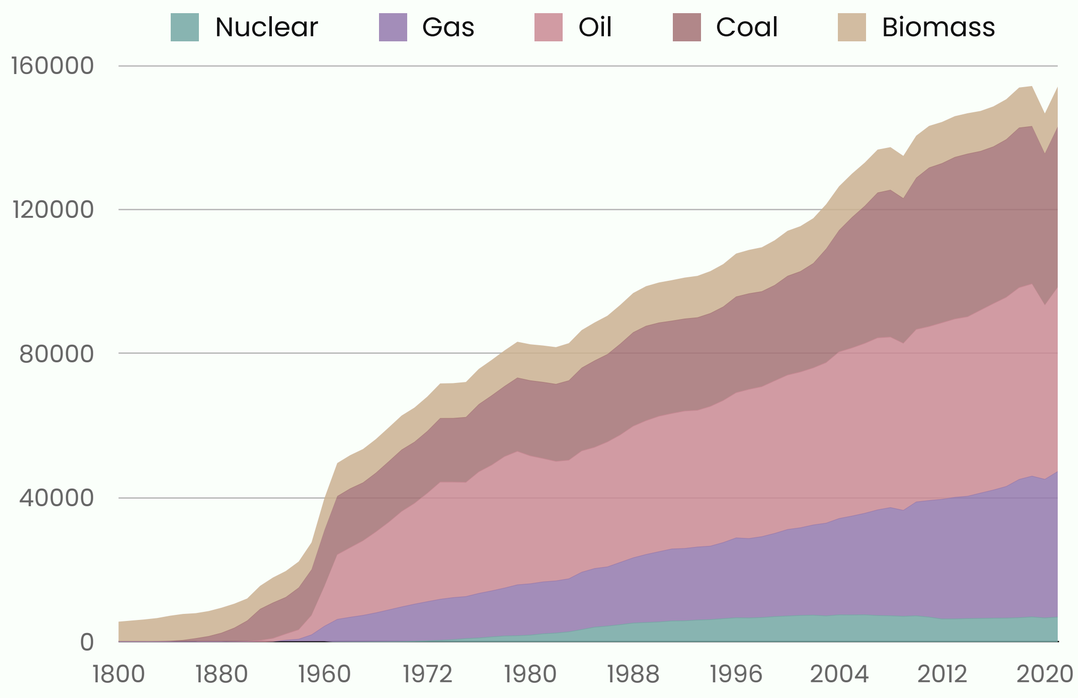


[Accessed May 6, 2025], https://ourworldindata.org/global-energy-200-years

EXAMPLE (3)

Global primary energy demand by source [TWh].





NB! Left-hand side plot: biomass and nuclear have very low trend.

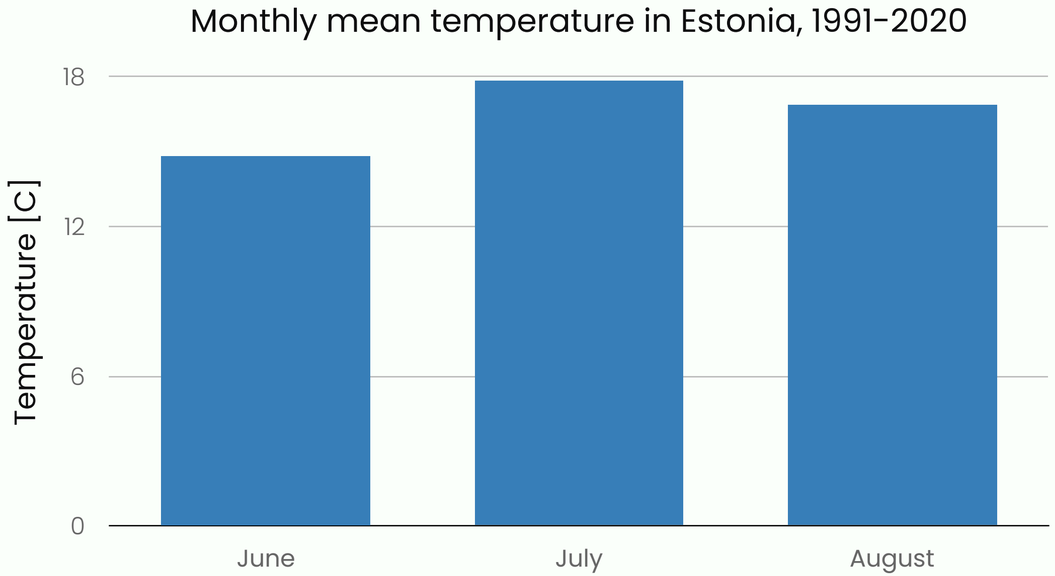
Right-hand side plot: it may seem that biomass has a more distinct trend.

[Accessed May 6, 2025], https://ourworldindata.org/global-energy-200-years

Common graphic forms and their uses

BAR CHART

Data: Compares the value of one variable across a series of items called cases (e.g., mean temperature$_{variable}$ in different months$_{cases}$).

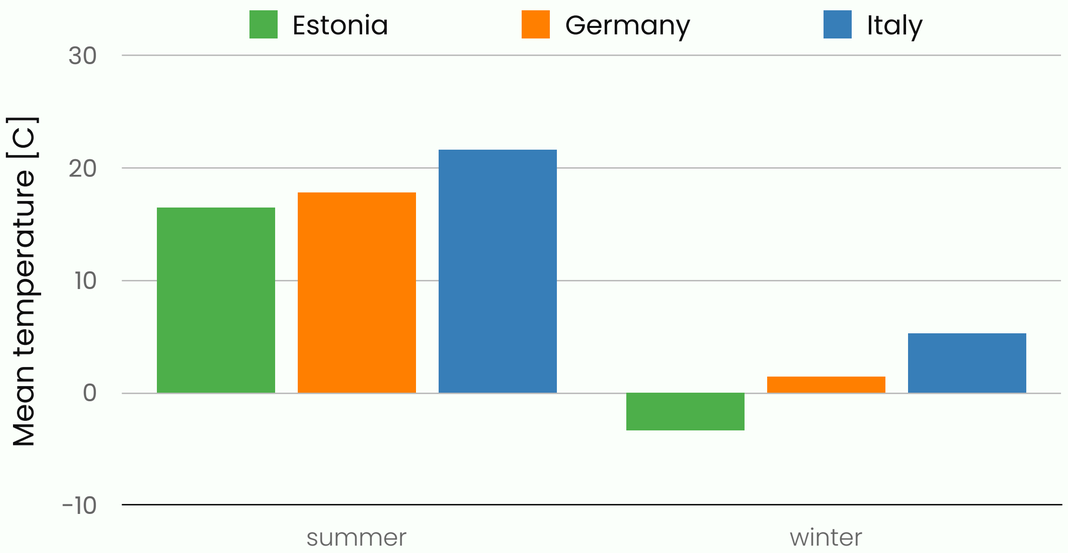


Usage:

- ✓ Creates strong visual contrasts among individual cases, emphasising individual comparisons.
- ✓ For specific add numbers to bars.
- ✓ Can show ranks or trends.
- ✓ Vertical bars (called columns) are most common, but can be horizontal if cases are numerous or have complex labels.

BAR CHART, GROUPED OR SPLIT

Data: Compares the value of one variable, divided into subsets, across a series of cases (e.g., mean temperature$_{variable}$ for summer and winter seasons$_{subsets}$ in three EU countries$_{cases}$).



Usage:

✓ Contrasts subsets within and across individual cases; not useful for comparing total values for cases.

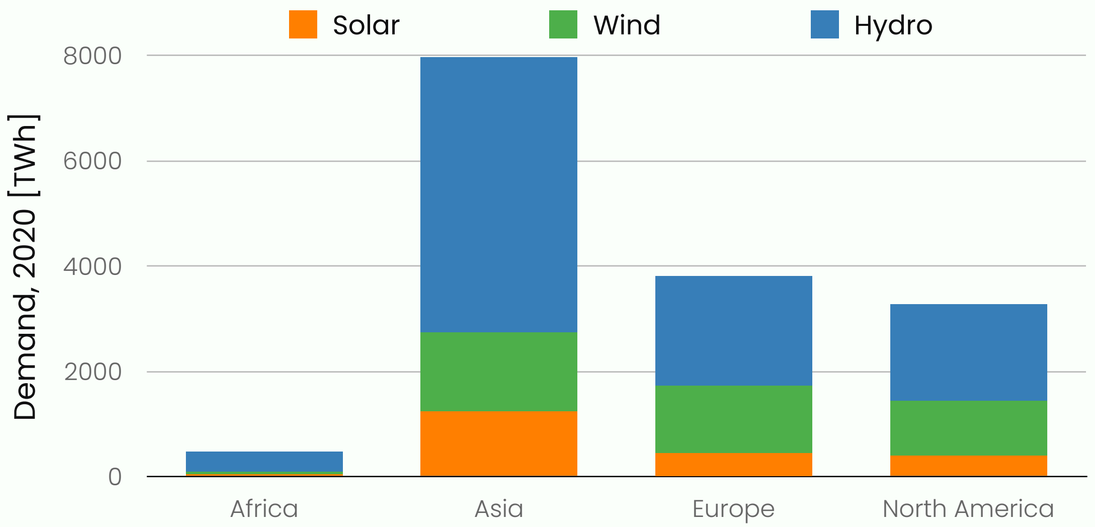
✓ For specific values, add numbers to bars.

✓ Grouped bars show ranking or trends poorly; useful for time series only if trends are unimportant.

https://climateknowledgeportal.worldbank.org/

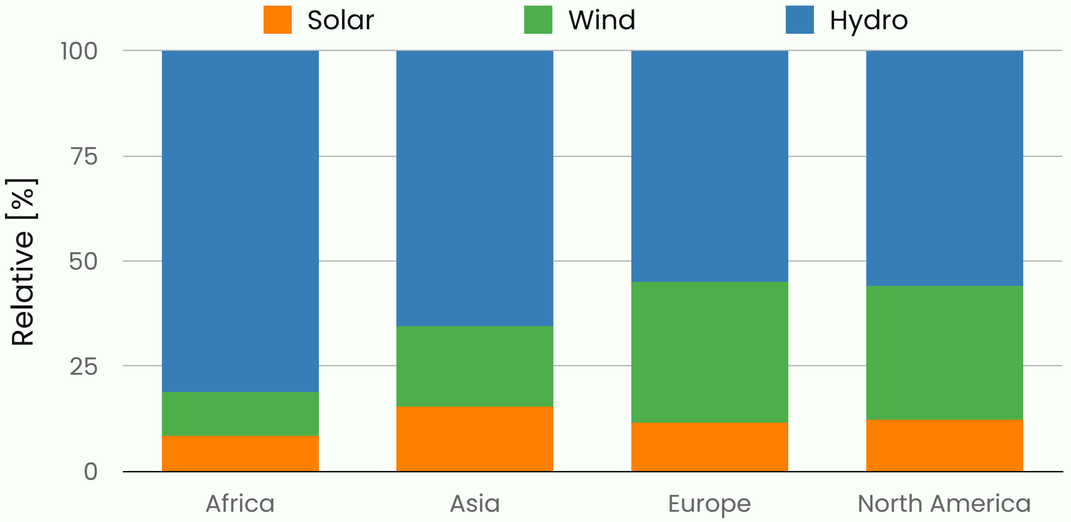
BAR CHART, STACKED

Data: Compares the value of one variable, divided into two or more subsets, across a series of cases (e.g., energy-mix$_{variable}$ segmented by source type$_{subsets}$ in four regions$_{cases}$).



Usage:

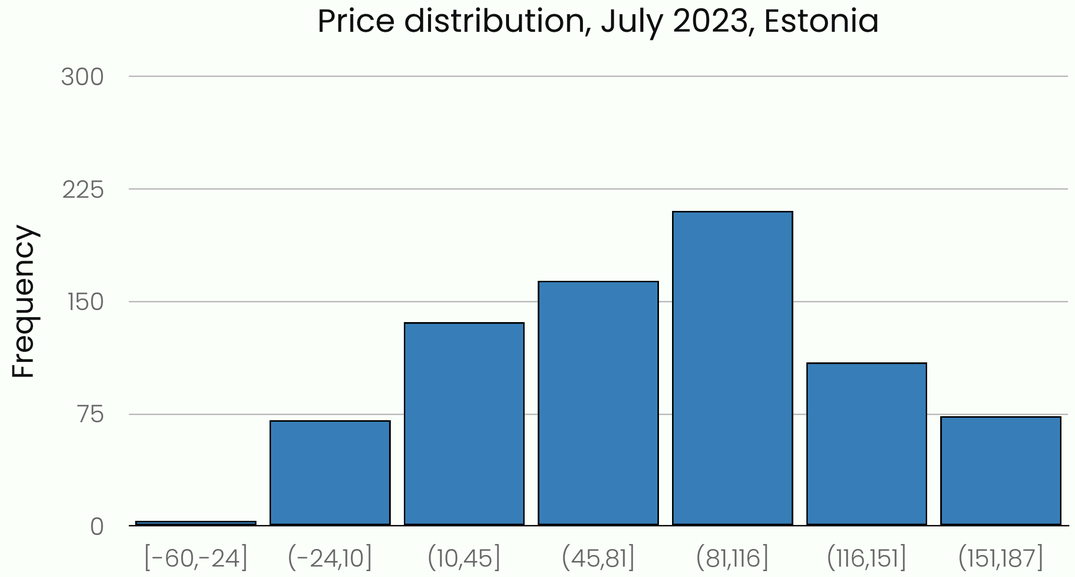
- ✓ Best for comparing totals across cases and subsets within cases; difficult to compare subsets across cases (use grouped bars).
- ✓ For specific values, add numbers to bars add segments.
- ✓ Useful for time series. Can show ranks or trends for total values only.



https://ourworldindata.org/energy-mix

HISTOGRAM

Data: Compares two variables, with one segmented into ranges that function like the cases in a bar graph (e.g., frequency$_{\textit{continuous variable}}$ with electricity prices$_{\textit{segmented variable}}$).



Usage:

✓ Best for comparing segments within continuous data sets.

✓ Shows trends, but emphasises segments.

✓ For specific values, add numbers to bars.

[Accessed May 8, 2025], https://dashboard.elering.ee/en/nps/price?interval=minute&period=search&start=2023-06-30T21:00:00.000Z&end=2023-08-01T20:59:59.000Z

IMAGE CHART

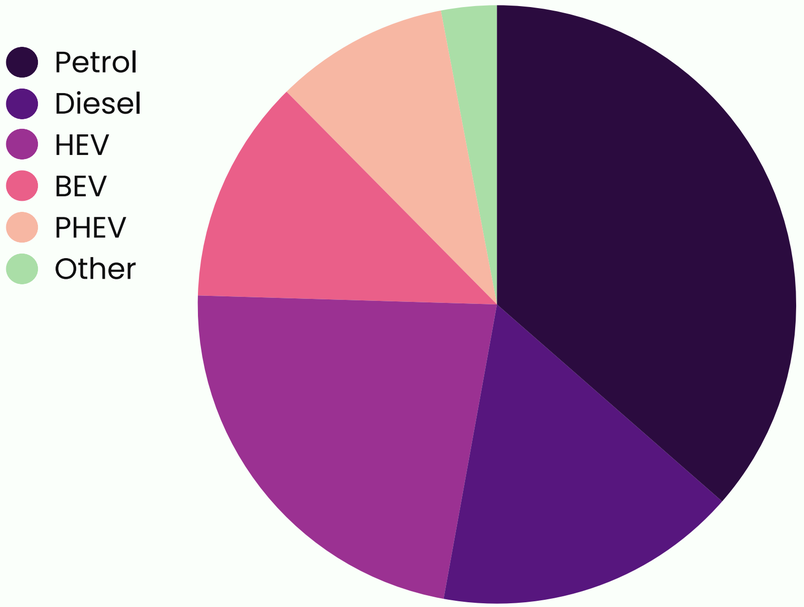
Data: Shows value of one or more variable for cases displayed on a map, diagram, or other image (e.g., bidding areas$_{cases}$ coloured in gradient to show electricity price$_{variable}$).

Usage:

- ✓ Shows the distribution of the data in relation to preexisting categories.
- ✓ De-emphasises specific values.
- ✓ Best when the image is familiar, as in a map or diagram of a process.

PIE CHART

Data: Shows the proportion of a single variable for a series of cases (e.g., the car market share$_{variable}$ by fuel type$_{cases}$).



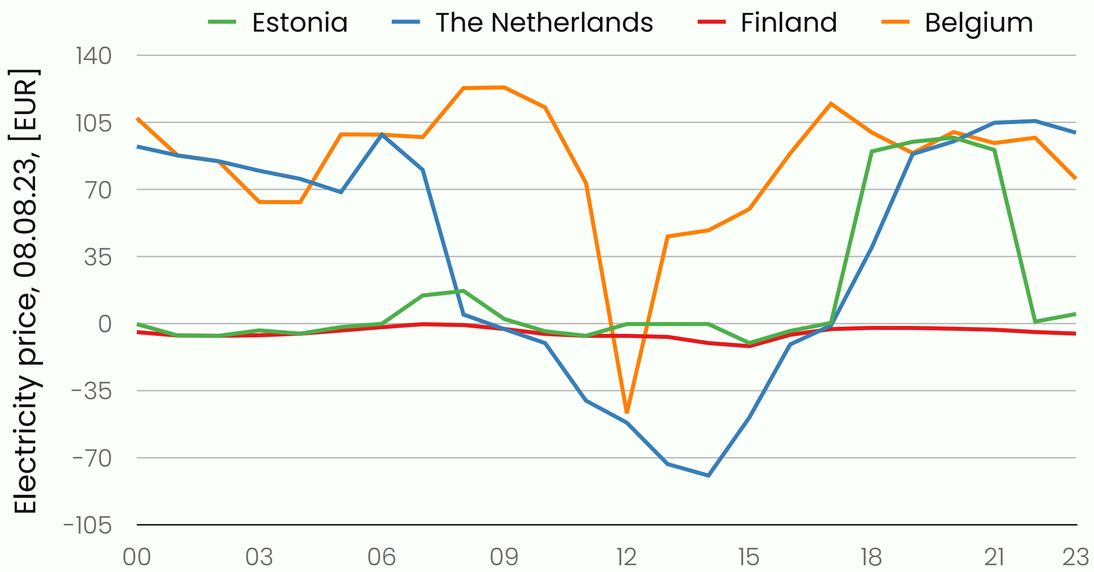
Usage:

- ✓ Best for comparing one segment to the whole.
- ✓ Useful only with few segments or segments that are very different in size; otherwise comparisons among segments are difficult.
- ✓ For specific values, add numbers to segments.
- ✓ Common in popular venues, frowned on by professionals.

[Accessed May 8, 2025], https://www.acea.auto/fuel-pc/fuel-types-of-new-cars-battery-electric-12-1-hybrid-22-6-and-petrol-36-4-market-share-full-year-2022/

LINE GRAPH

Data: Compares continuous variables for one or more cases (e.g., electricity price$_{variable}$ and time$_{variable}$ in four EU countries$_{case}$).

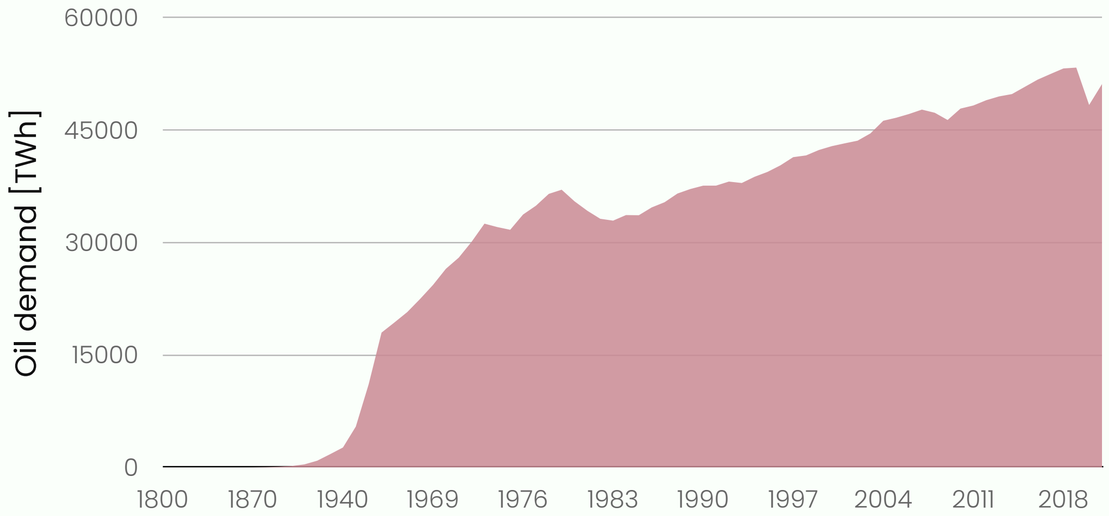


Usage:

- ✓ Best for showing trends; deemphasises specific values.
- ✓ Useful for time series.
- ✓ To show specific values, add numbers to data points.
- ✓ To show the significance of a trend, segment the grid (e.g., below or above average performance).

AREA CHART

Data: Compares two continuous variables for one or more cases (e.g., oil demand$_{variable}$ over time$_{variable}$ in the world$_{case}$).



Usage:

✓ Shows trends; deemphasises specific values.

✓ Can be used for time series.

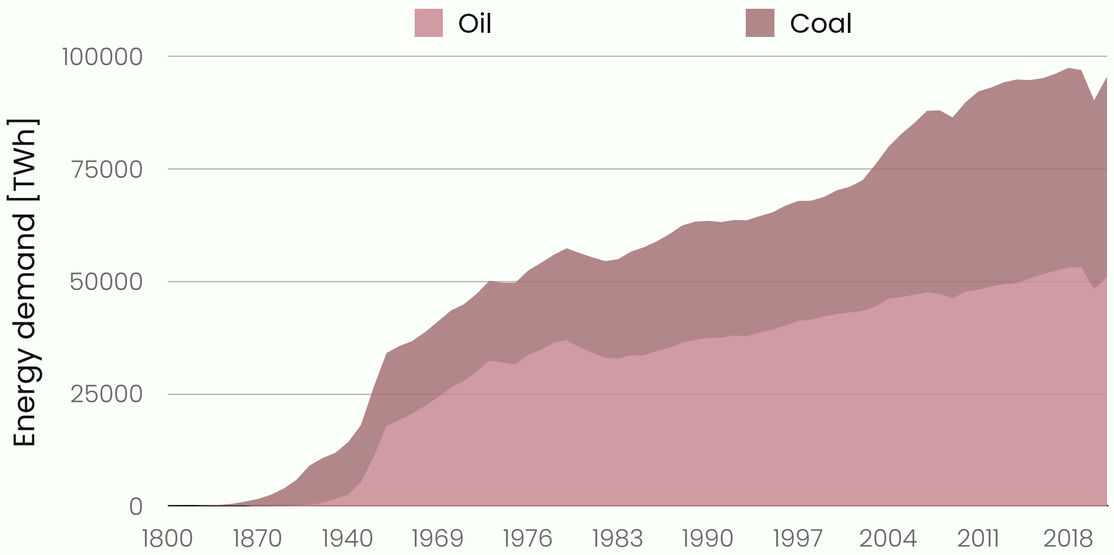
✓ To show specific values, add numbers to data points.

✓ Areas below the lines add no information, but will lead some readers to misjudge values.

✓ Confusing with multiple lines/areas.

AREA CHART, STACKED

Data: Compares two continuous variables for two or more cases (e.g., energy demand$_{variable}$ over time$_{variable}$ by energy source$_{case}$).



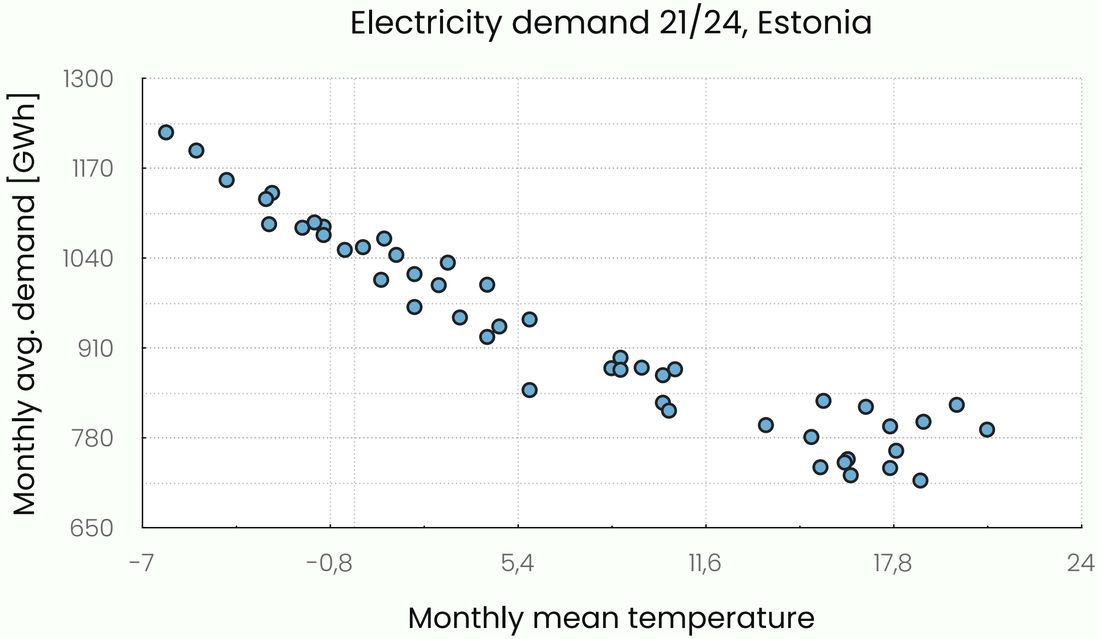
Usage:

✓ Shows the trend for the total of all cases, plus how much each case contributes to that total.

✓ Likely to mislead readers on the value or the trend for any individual case.

SCATTERPLOT

Data: Compares two variables at multiple data points for a single case (e.g., energy demand and temperature in one building).

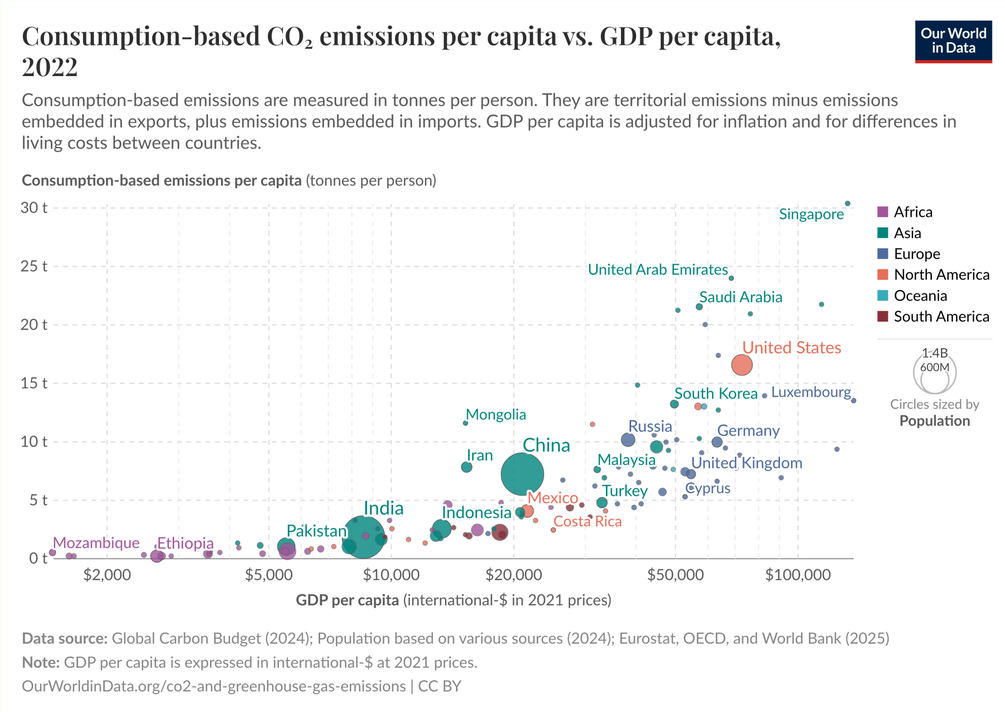


Usage:

- ✓ Best for showing the distribution of data, especially when there is no clear trend or when the focus is on outlying data points.
- ✓ If only a few data points are plotted, it allows a focus on individual values.

BUBBLE PLOT

Data: Compares three variables at multiple data points for a single case (e.g., GDP per capita, CO2 emissions per capita, and population size).

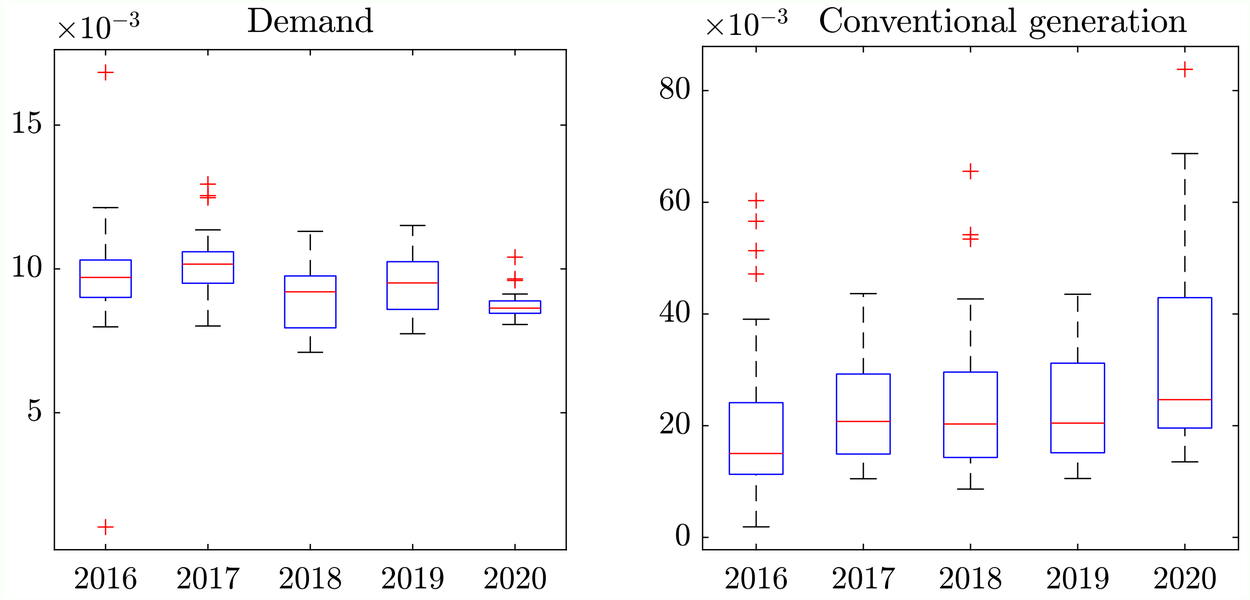


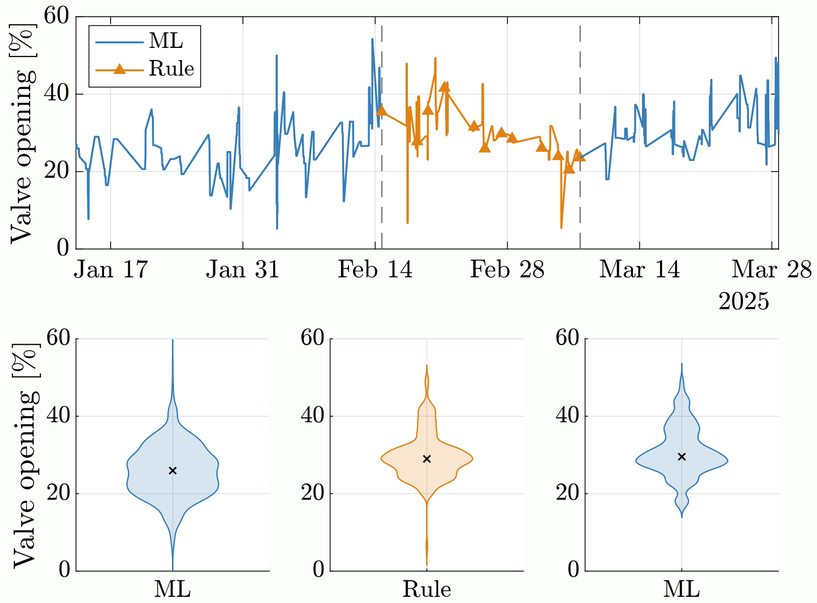
Usage:

- ✓ Emphasises the relationship between the third variable (bubbles) and the first two.
- ✓ Most useful when the question is whether the third variable is a product the others.
- ✓ Readers easily misjudge relative values shown by bubbles; adding numbers mitigates that problem.

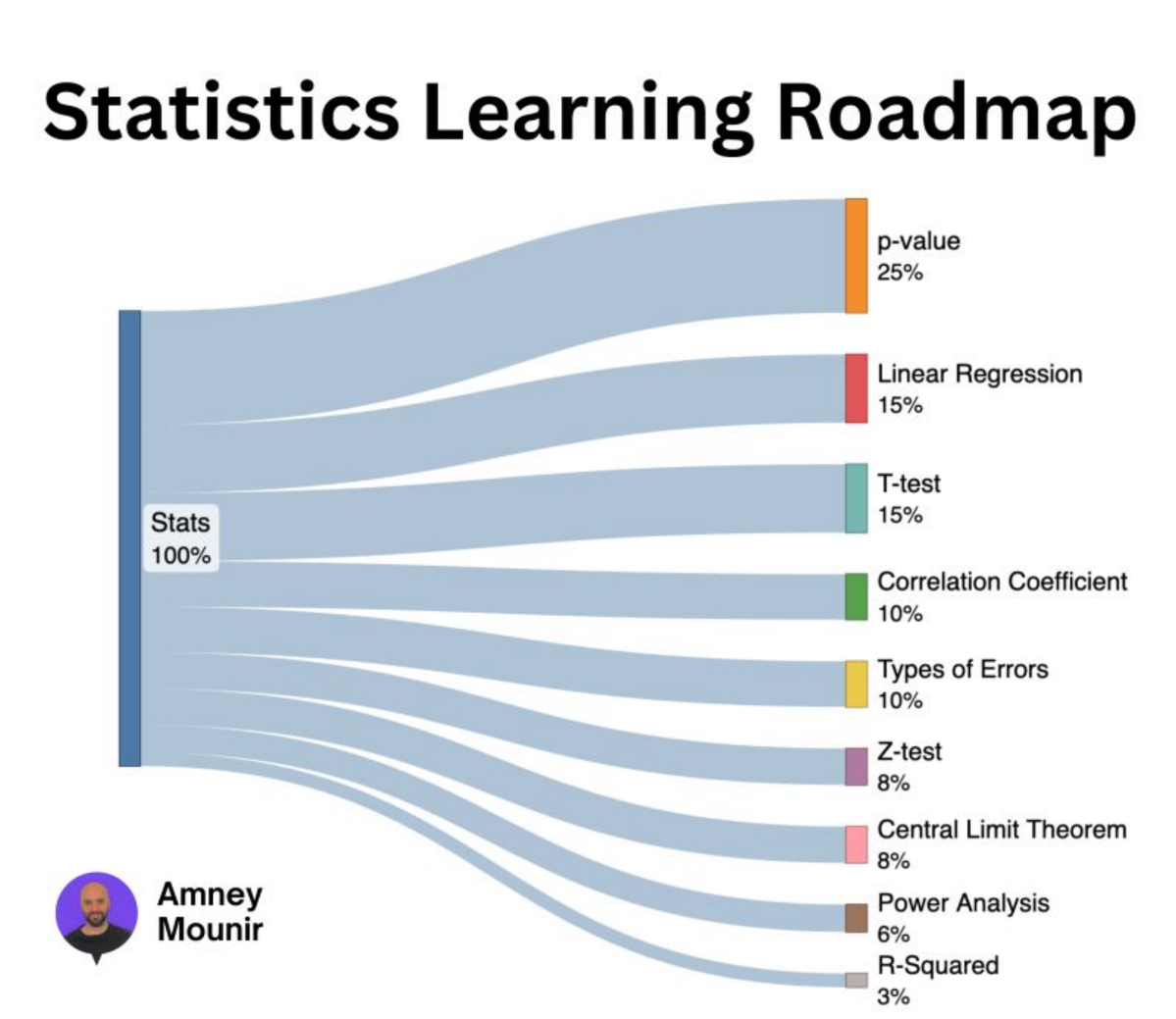
[Accessed May 6, 2025], https://ourworldindata.org/worlds-energy-problem

AND MORE (BOX, VIOLIN PLOT) ...



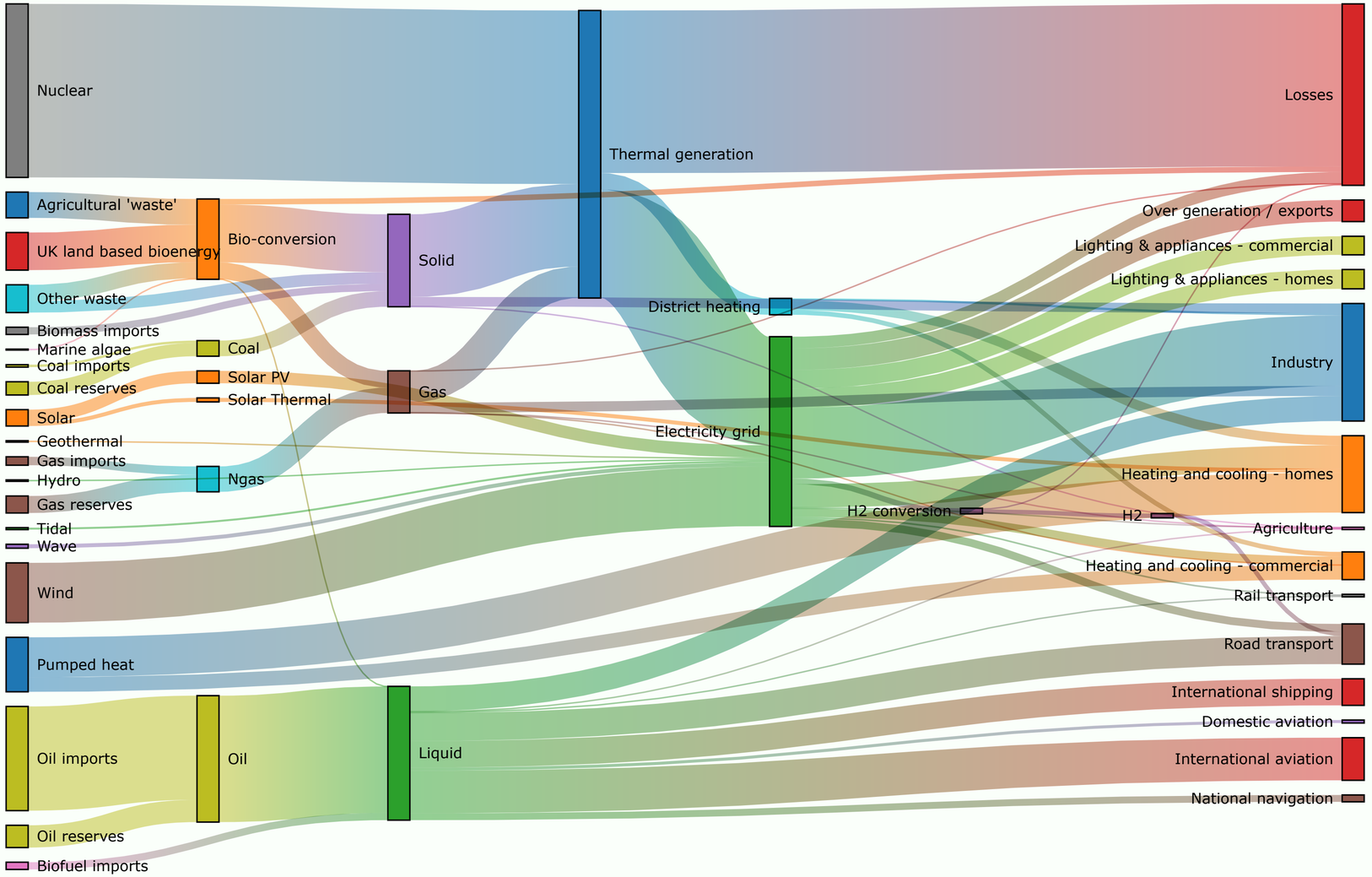


AND MORE (SANKEY DIAGRAM) … (2)



[Accessed May 19, 2025], https://www.linkedin.com/posts/amney-mounir_here-is-what-you-need-to-to-ace-your-statistics-activity-7316090162260631552-BkGQ/

AND MORE ... (3)



[Accessed May 6, 2025], https://observablehq.com/@d3/sankey/2

COMMON GRAPHIC FORMS AND THEIR USES

Further reading:

https://www.data-to-viz.com/

https://www.r-graph-gallery.com/index.html

http://pgfplots.sourceforge.net/gallery.html

https://colorbrewer2.org

https://inkscape.org/

https://www.overleaf.com/learn/latex/TikZ_package

Thank you!

Questions?