Imprs Summer school 2023

Graphs Do’s & Don’ts

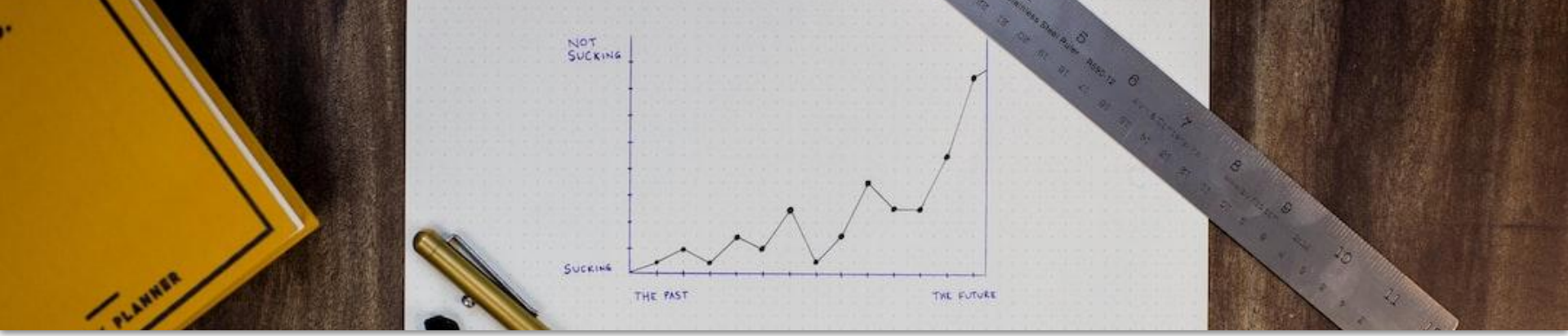
MORGAN FOUESNEAU & IVA MOMCHEVA

ds@mpia.de @mpi-astronomy

Max-Planck-Institut für Astronomie Heidelberg
Max Planck Institute for Astronomy Heidelberg

Content

- Making the right plot – a tool for efficient communication of findings
  - Going beyond defaults
  - Avoid common traps
- Effective analysis of large, multi-dimensional collections of data
  - Accessing online data: a brief introduction to ADQL
  - Exploring on disk data: dask, vaex, and TopCAT



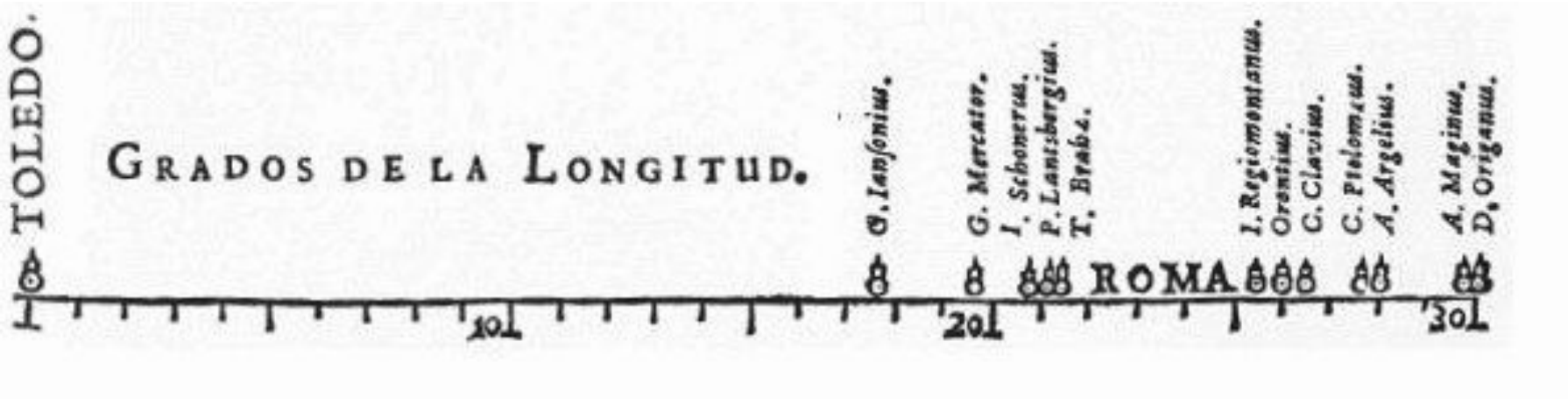
Data Visualization

a fundamental element of ~~astronomy~~ science

Efficient communication of findings

1643 - First statistical graphic

Stem plot

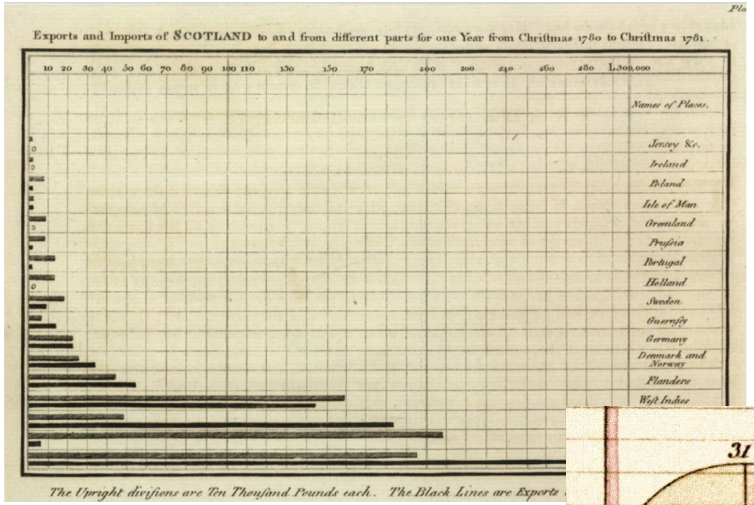


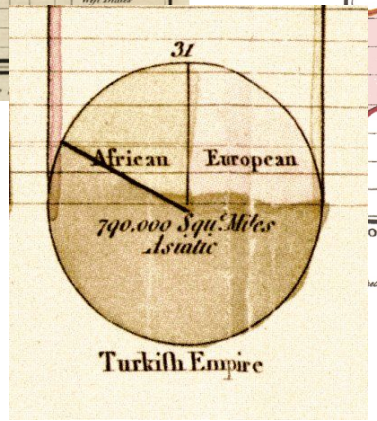
Graph of statistical data, showing the wide range of estimates of the distance in longitude between Toledo and Rome.

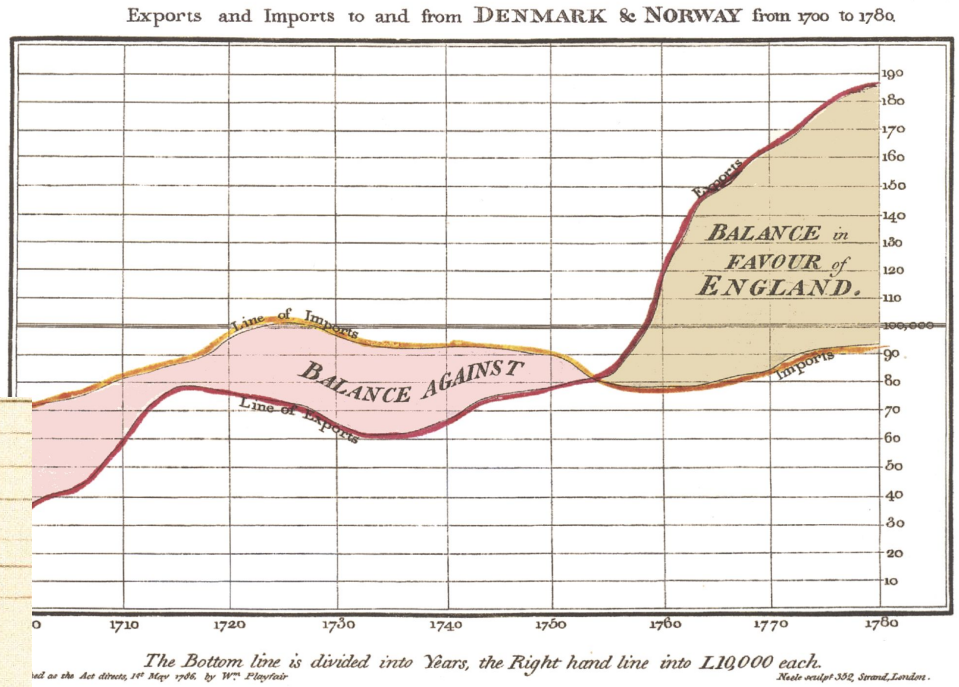
By Flemish cartographer [Michaël Florent van Langren (Langrenus)]

Late 18th Century - William Playfair invents

Bar plot
Pie chart
scatter plot







The ancestor of

Robert FitzRoy daily weather forecast chart

vice-admiral
Robert FitzRoy

1861

THE WEATHER.

METEOROLOGICAL REPORTS.

Wednesday, July 31, 8 to 9 a.m.	B.	E.	M.	D.	F.	C.	I.	S.
Nairn..	29·54	57	56	W.S.W.	6	9	o.	3
Aberdeen	29·60	59	54	S.S.W.	5	1	b.	3
Leith	29·70	61	55	W.	3	5	c.	2
Berwick	29·69	59	55	W.S.W.	4	4	c.	2
Ardrossan . ..	29·73	57	55	W.	5	4	c.	5
Portrush	29·72	57	54	S.W.	2	2	b.	2
Shields	29·80	59	54	W.S.W.	4	5	o.	3
Galway	29·83	65	62	W.	5	4	c.	4
Scarborough ..	29·86	59	56	W.	3	6	c.	2
Liverpool.. ..	29·91	61	56	S.W.	2	8	c.	2
Valentia	29·87	62	60	S.W.	2	5	o.	3
Queenstown ..	29·88	61	59	W.	3	5	c.	2
Yarmouth.. ..	30·05	61	59	W.	5	2	c.	3
London	30·02	62	56	S.W.	3	2	b.	—
Dover..	30·04	70	64	S.W.	3	7	o.	2
Portsmouth ..	30·01	61	59	W.	3	6	o.	2
Portland	30·03	63	59	S.W.	3	2	c.	3
Plymouth.. ..	30·09	62	59	W.	5	1	b.	4
Penzance	30·04	61	60	S.W.	2	6	c.	3
Copenhagen ..	29·94	64	—	W.S.W.	2	6	c.	3
Helder	29·99	63	—	W.S.W.	6	5	c.	3
Brest	30·09	60	—	S.W.	2	6	c.	5
Bayonne	30·13	68	—	—	—	9	m.	5
Lisbon	30·18	70	—	N.N.W.	4	3	b.	2

General weather probable during next two days in the—
North—Moderate westerly wind ; fine.
West—Moderate south-westerly ; fine.
South—Fresh westerly ; fine.

Explanation.

B. Barometer, corrected and reduced to 32° at mean sea level ; each 10 feet of vertical rise causing about one-hundredth of an inch diminution, and each 10° above 32° causing nearly three-hundredths increase. E. Exposed thermometer in shade. M. Moistened bulb (for evaporation and dew-point). D. Direction of wind (true—two points *left* of magnetic). F. Force (1 to 12—estimated). C. Cloud (1 to 9). I. Initials :—b., blue sky ; c., clouds (detached); f., fog ; h., hail ; l., lightning ; m., misty (hazy) ; o., overcast (dull) ; r., rain ; s., snow ; t., thunder. S. Sea disturbance (1 to 9).

Robert FitzRoy daily weather forecast chart

vice-admiral
Robert FitzRoy

1861

THE WEATHER.

METEOROLOGICAL REPORTS.

Wednesday, July 31, 8 to 9 a.m.	B.	E.	M.	D.	F.	C.	I.	S.
Nairn..	29·54	57	56	W.S.W.	6	9	o.	3
Aberdeen	29·60	59	54	S.S.W.	5	1	b.	3
Leith	29·70	61	55	W.	3	5	c.	2
Berwick	29·69	59	55	W.S.W.	4	4	c.	2
Ardrossan . ..	29·73	57	55	W.	5	4	c.	5
Portrush	29·72	57	54	S.W.	2	2	b.	2
Shields	29·80	59	54	W.S.W.	4	5	o.	3
Galway	29·83	65	62	W.	5	4	c.	4
Scarborough ..	29·86	59	56	W.	3	6	c.	2
Liverpool.. ..	29·91	61	56	S.W.	2	8	c.	2
Valentia	29·87	62	60	S.W.	2	5	o.	3
Queenstown ..	29·88	61	59	W.	3	5	c.	2
Yarmouth.. ..	30·05	61	59	W.	5	2	c.	3
London	30·02	62	56	S.W.	3	2	b.	—
Dover..	30·04	70	64	S.W.	3	7	o.	2
Portsmouth ..	30·01	61	59	W.	3	6	o.	2
Portland	30·03	63	59	S.W.	3	2	c.	3
Plymouth.. ..	30·00	62	59	W.	5	1	b.	4
Penzance	30·04	61	60	S.W.	2	6	c.	3
Copenhagen ..	29·94	64	—	W.S.W.	2	6	c.	3
Helder	29·99	63	—	W.S.W.	6	5	c.	3
Brest	30·09	60	—	S.W.	2	6	c.	5
Bayonne	30·13	68	—	—	—	9	m.	5
Lisbon	30·18	70	—	N.N.W.	4	3	b.	2

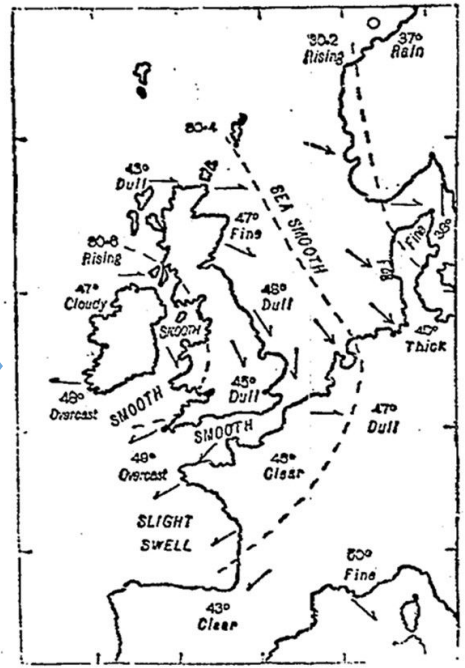
General weather probable during next two days in the—
North—Moderate westerly wind ; fine.
West—Moderate south-westerly ; fine.
South—Fresh westerly ; fine.

Explanation.

B. Barometer, corrected and reduced to 32° at mean sea level ; each 10 feet of vertical rise causing about one-hundredth of an inch diminution, and each 10° above 32° causing nearly three-hundredths increase. E. Exposed thermometer in shade. M. Moistened bulb (for evaporation and dew-point). D. Direction of wind (true—two points *left* of magnetic). F. Force (1 to 12—estimated). C. Cloud (1 to 9). I. Initials :—b., blue sky ; c., clouds (detached) ; f., fog ; h., hail ; l., lightning ; m., misty (hazy) ; o., overcast (dull) ; r., rain ; s., snow ; t., thunder. S. Sea disturbance (1 to 9).

1875

WEATHER CHART, MARCH 31, 1875.



The dotted lines indicate the gradations of barometric pressure The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

weather forecast today's evolution

1861

THE WEATHER.

METEOROLOGICAL REPORTS.

Wednesday, July 31, 8 to 9 a.m.	B.	E.	M.	D.	F.	C.	I.	S.
Nairn..	29·54	57	56	W.S.W.	6	9	o.	3
Aberdeen	29·60	59	54	S.S.W.	5	1	b.	3
Leith	29·70	61	55	W.	3	5	c.	2
Berwick	29·69	59	55	W.S.W.	4	4	c.	2
Ardrossan . ..	29·73	57	55	W.	5	4	c.	5
Portrush	29·72	57	54	S.W.	2	2	b.	2
Shields	29·80	59	54	W.S.W.	4	5	o.	3
Galway	29·83	65	62	W.	5	4	c.	4
Scarborough ..	29·86	59	56	W.	3	6	c.	2
Liverpool	29·91	61	56	S.W.	2	8	c.	2
Valentia	29·87	62	60	S.W.	2	5	o.	3
Queenstown ..	29·88	61	59	W.	3	5	c.	2
Yarmouth.. ..	30·05	61	59	W.	5	2	c.	3
London	30·02	62	56	S.W.	3	2	b.	—
Dover..	30·04	70	64	S.W.	3	7	o.	2
Portsmouth ..	30·01	61	59	W.	3	6	o.	2
Portland	30·03	63	59	S.W.	3	2	c.	3
Plymouth.. ..	30·00	62	59	W.	5	1	b.	4
Penzance	30·04	61	60	S.W.	2	6	c.	3
Copenhagen ..	29·94	64	—	W.S.W.	2	6	c.	3
Helder	29·99	63	—	W.S.W.	6	5	c.	3
Brest	30·09	60	—	S.W.	2	6	c.	5
Bayonne	30·13	68	—	—	—	9	m.	5
Lisbon	30·18	70	—	N.N.W.	4	3	b.	2

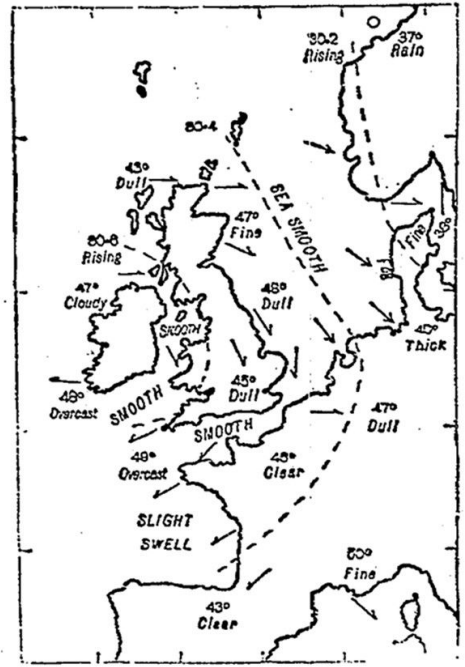
General weather probable during next two days in the—
North—Moderate westerly wind; fine.
West—Moderate south-westerly; fine.
South—Fresh westerly; fine.

Explanation.

B. Barometer, corrected and reduced to 32° at mean sea level; each 10 feet of vertical rise causing about one-hundredth of an inch diminution, and each 10° above 32° causing nearly three-hundredths increase. E. Exposed thermometer in shade. M. Moistened bulb (for evaporation and dew-point). D. Direction of wind (true—two points *left* of magnetic). F. Force (1 to 12—estimated). C. Cloud (1 to 9). I. Initials:—b., blue sky; c., clouds (detached); f., fog; h., hail; l., lightning; m., misty (hazy); o., overcast (dull); r., rain; s., snow; t., thunder. S. Sea disturbance (1 to 9).

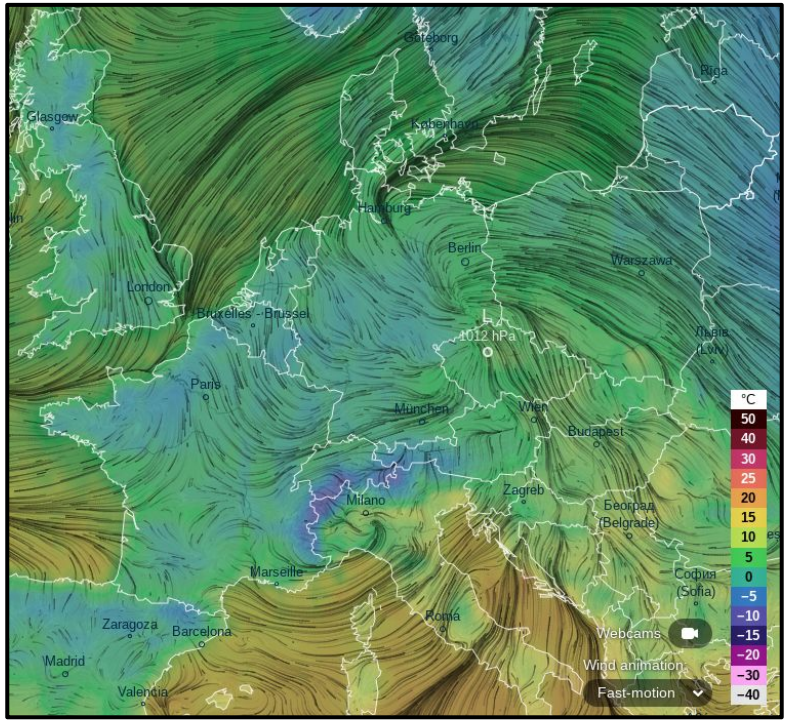
1875

WEATHER CHART, MARCH 31, 1875.



The dotted lines indicate the gradations of barometric pressure. The variations of the temperature are marked by figures, the state of the sea and sky by descriptive words, and the direction of the wind by arrows—barbed and feathered according to its force. ⊙ denotes calm.

2022



Forecast for Dec 5, 2022
One example of many maps (ventusky.com)

Any graphical display should

- **Show the data**
  - Avoid distorting what the data have to say
- Induce the reader to **think about the message**
  - Not about methodology, graphics design, …
- Present **concisely** many numbers in a small space
- serve a reasonably **clear purpose**
  - Description, exploration, tabulation, or decoration (not in science)
- **closely integrated** with the descriptions of a data set.

What makes **good** graphics?

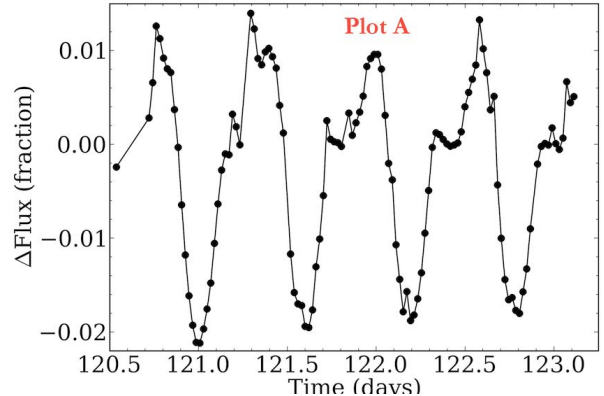
slido

What makes good graphics/plots?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

What **tool** to use?

The tool is **not an excuse** for bad graphics

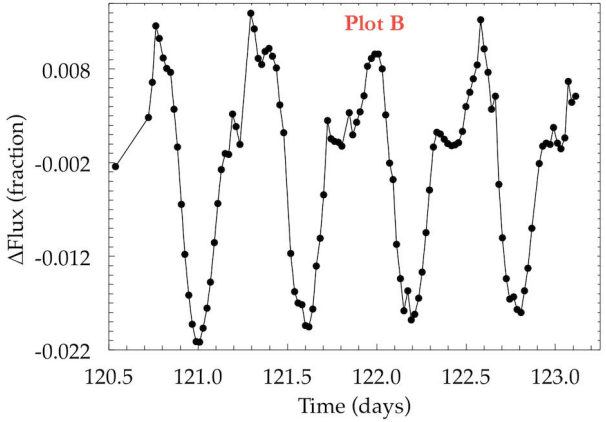


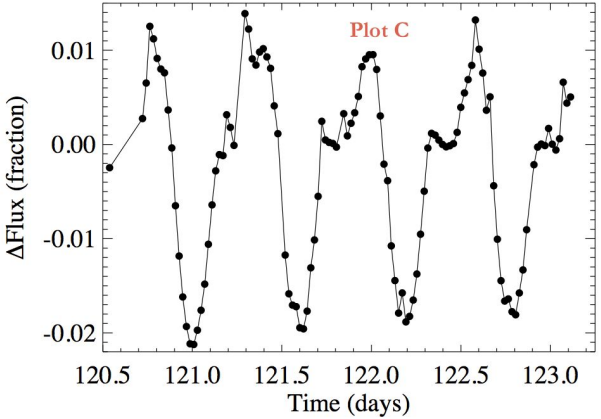
Excel vs Python vs IDL

Any tool can create the same figure

(with more or less work)

Question the defaults!





blog post
by Jim Davenport (2013)

A=Python, B=Excel, C=IDL

Important aspects to consider

- Match the graph
- Color choices
- Fonts size, notations, axis labels
- Wise use of the figure caption to complement the figure
- Think of reusability to increase your impact

Question the defaults!

Key **aspects** to **design** a figure

Important aspects to consider

- Match the graph
- Color choices
- Fonts size, notations, axis labels
- Wise use of the figure caption to complement the figure
- Think of reusability to increase your impact

Question the defaults!

Match the graph type to what you are communicating

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith, Chris Campbell, Ian Bott, Liz Faunce, Graham Parrish, Billy Ehrenberg-Shannon, Paul McCallum, Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar
A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar
Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine
Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line
The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot
The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline
A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot
Usually used to show how the relationship between 2 variables has changed over time.

Bubble
Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap
A good way of showing the patterns between 2 categories of data. Less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar
Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column
See above.

Ordered proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot
Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope
Perfect for showing how ranks have changed over time or vary between categories.

Lollipop
Lollipops draw more attention to the data value than standard barcharts and can also show rank and value effectively.

Bump
Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram
The standard way to show a statistical distribution – keeps the gaps between columns small to highlight the 'shape' of the data.

Dot plot
A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot
Good for showing individual values in a distribution; can be a problem when too many dots have the same value.

Barcode plot
Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot
Summarise multiple distributions by showing the median (centre) and range of the data.

Violin plot
Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid
A standard way for showing the age and sex breakdown of a population distribution, effectively, back to back histograms.

Cumulative curve
A good way of showing how unequal a distribution is: y axis is always cumulative frequency; x axis is always a measure.

Frequency polygons
For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Beeswarm
Use to emphasise individual points in a distribution. Points can be linked to an additional variable. Best with medium-sized datasets.

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line
The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column
Columns work well for showing change over time – but usually best with only one series of data or a time.

Column + line timeline
A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope
Good for showing changes data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart
Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick
Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)
Use to show the uncertainty in future projections – usually this grows the further forward the projection.

Connected scatterplot
A good way of showing changes for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap
A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline
Great when date and duration are key elements of the story in the data.

Circle timeline
Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline
Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Sonogram
Another alternative to the circle timeline for showing series where there are big variations in the data.

Streamgraph
A type of area chart; use when seeing changes in proportions over time is more important than individual values.

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per-cent.

Example FT uses
Fiscal budgets, company structures, national election results

Column
The standard way to compare the size of things. Must always start at 0 on the axis.

Bar
See above. Good when the data are not time series and labels have long category names.

Paired column
As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar
See above.

Marimekko
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol
Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictograph)
Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop
Lollipop charts draw more attention to the data value than standard barcharts – does not have to start at zero (but preferable).

Radar
A space-efficient way of showing values of multiple variables - but make sure they are organised in a way that makes sense to reader.

Parallel coordinates
An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet
Good for showing a measurement against the context of a target or performance range.

Grouped symbol
An alternative to barcharts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar
A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko
A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie
A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut
Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap
Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi
A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc
A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot
Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn
Generally only used for schematic representation.

Waterfall
Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)
The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)
Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map
For showing unambiguous movement across a map.

Contour map
For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values.

Equalised cartogram
Converting each unit on a map to a regular and equally sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)
Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density
Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map
Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey
Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall
Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord
A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network
Used for showing the strength and inter-connectedness of relationships of varying types.

© Financial Times 2016-2019
This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time - but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

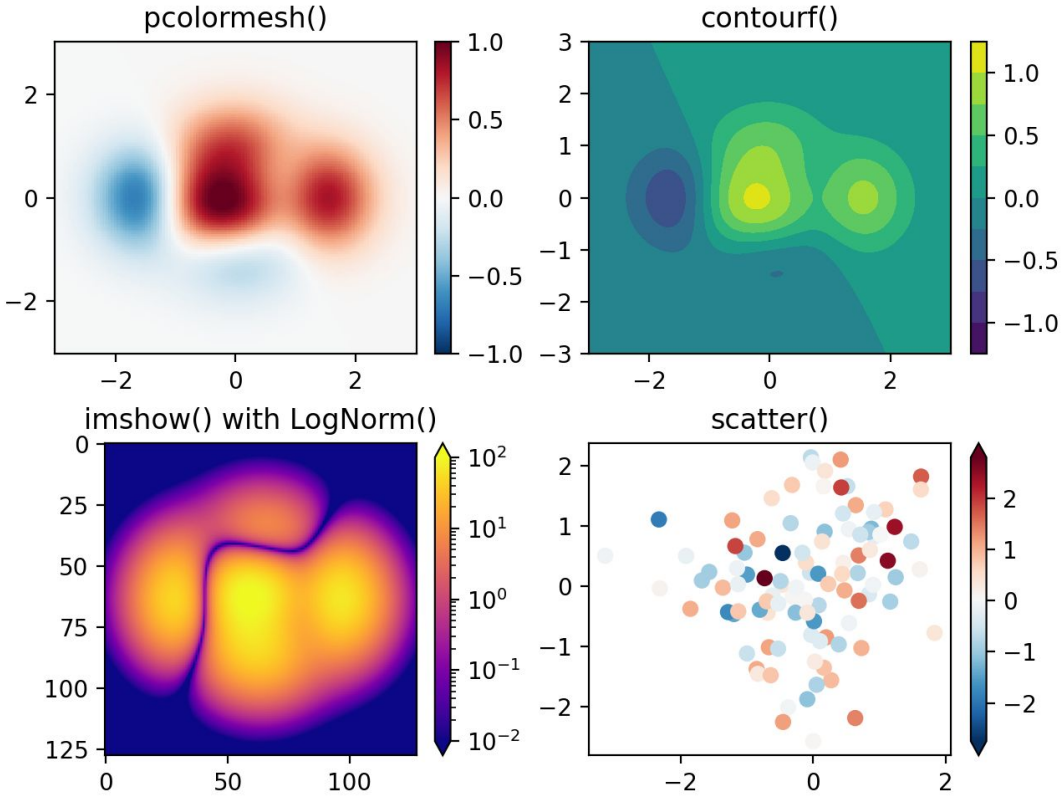
Isotype (pictogram)

“Efficient” communication of findings

choices affect **messages**

- type of chart
- colormaps
- normalization
- …

Do you **take-away** the same *“things”* on all the panels?

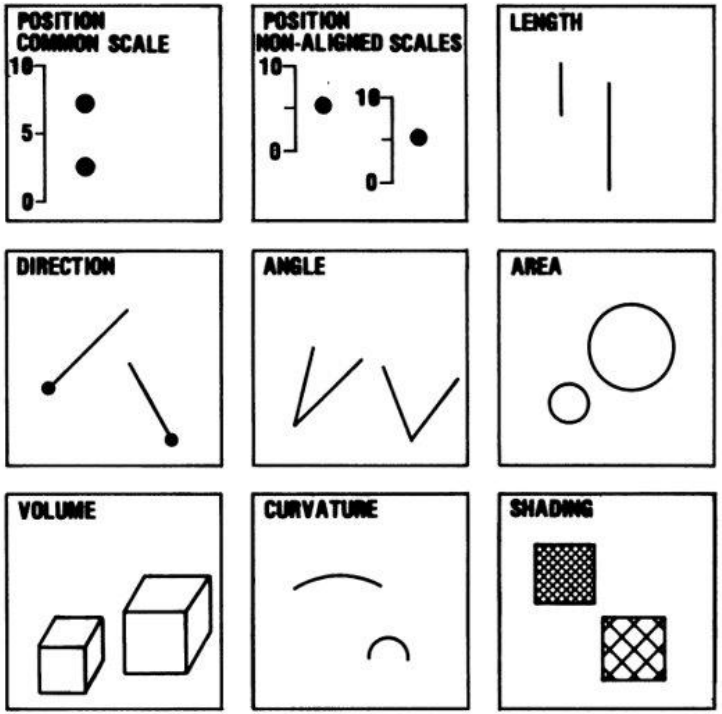


Example from matplotlib/tutorials

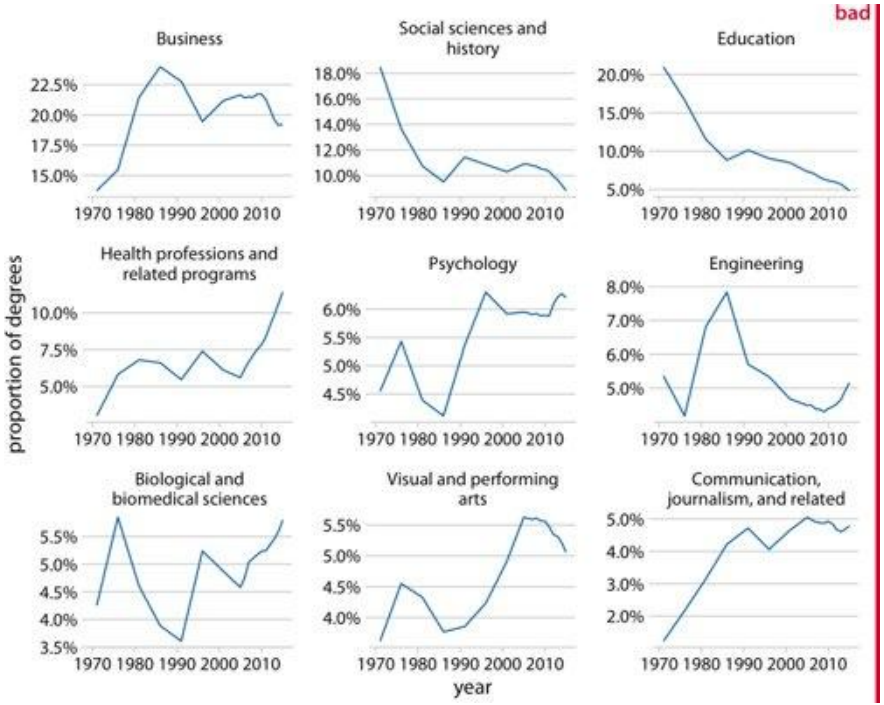
Hierarchy of **perception**

Comparison difficulty levels

- **Easiest**: comparing lengths and positions along a single axis
- **Moderate**: comparing angles, directions and areas
- **Hard**: comparing volumes, curves or colors



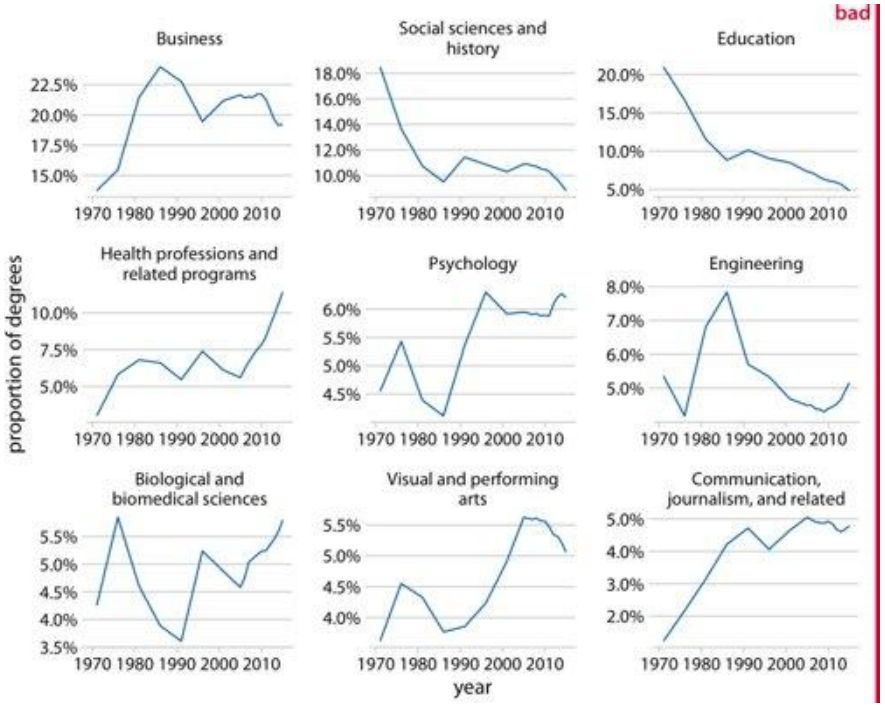
Multi-panel figures

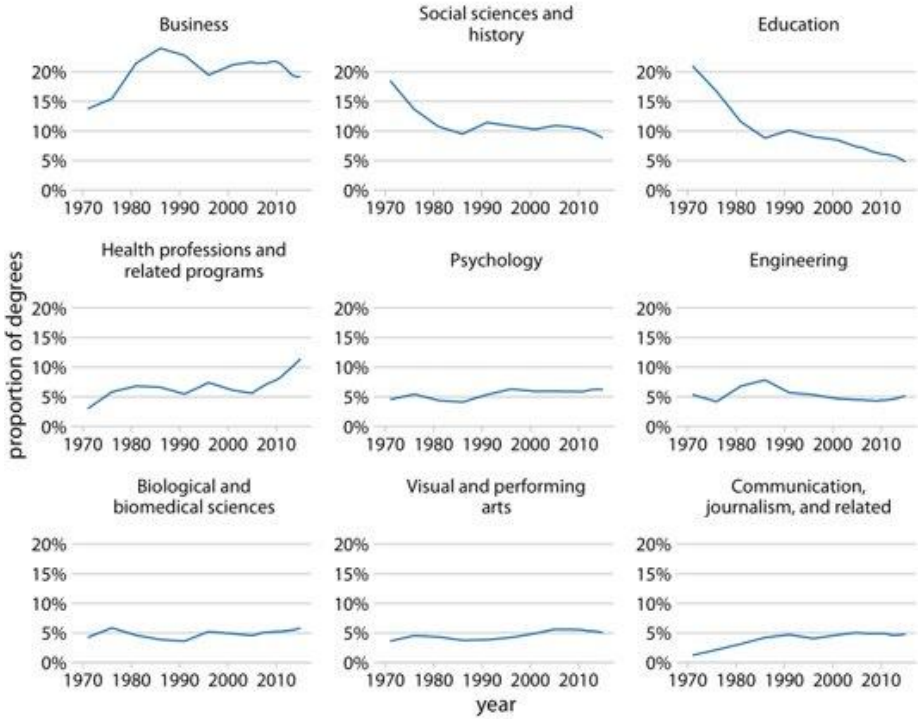


https://clauswilke.com/dataviz/

Multi-panel figures

Same axis scale to help the comparison

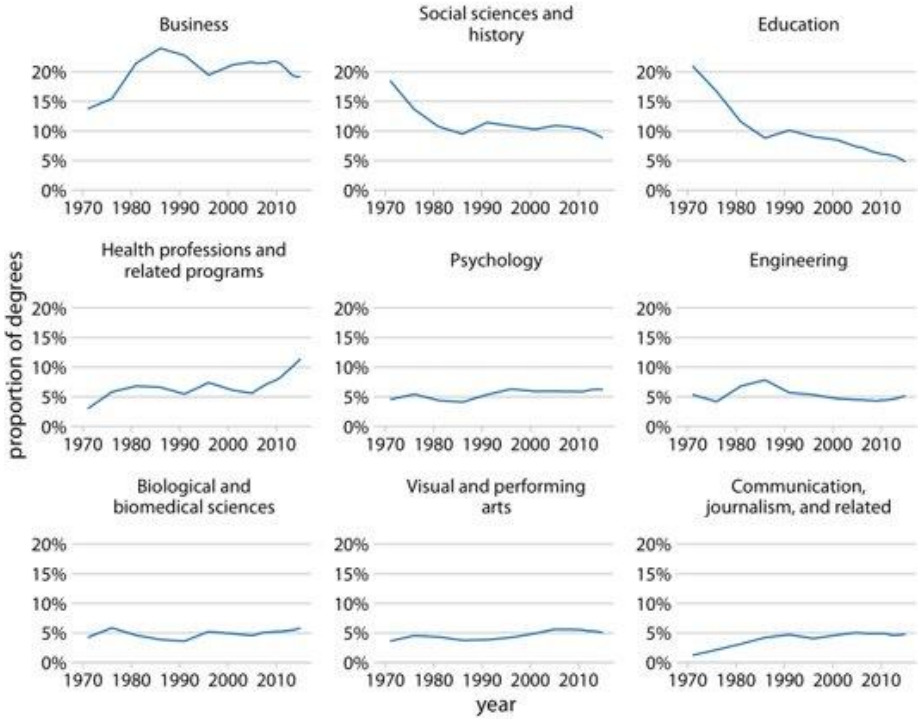




https://clauswilke.com/dataviz/

Multi-panel figures

- Use the **same (xy)-axis range**
- If you absolutely cannot, then draw attention to the differences in the caption
- Consider **ordering the panels** (here, most to least popular major)

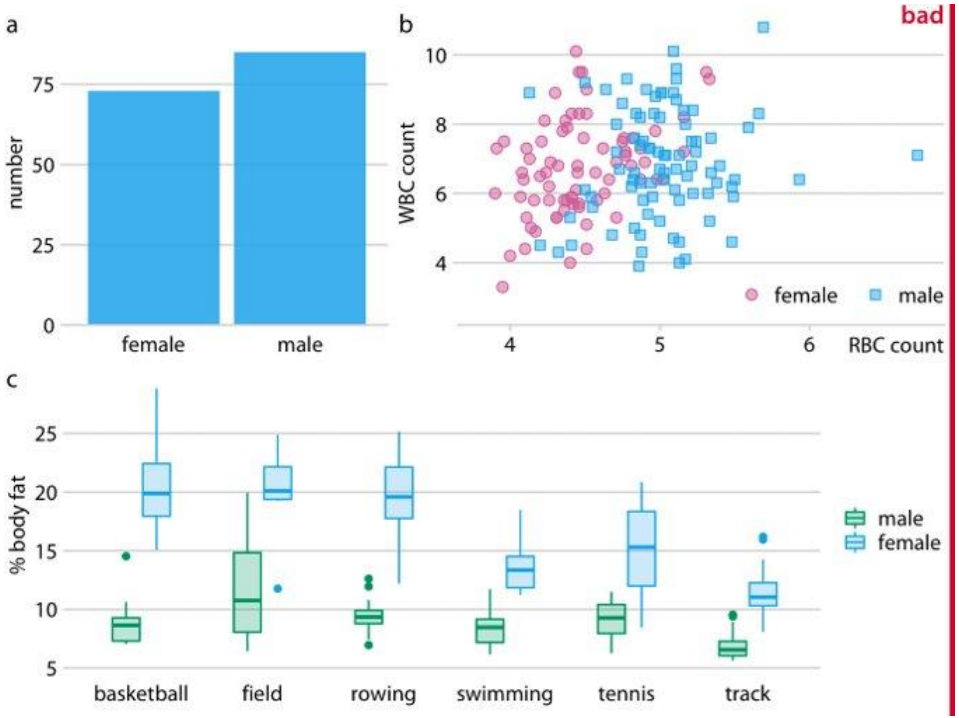


https://clauswilke.com/dataviz/

Compound figures

- The elements need to be connected logically

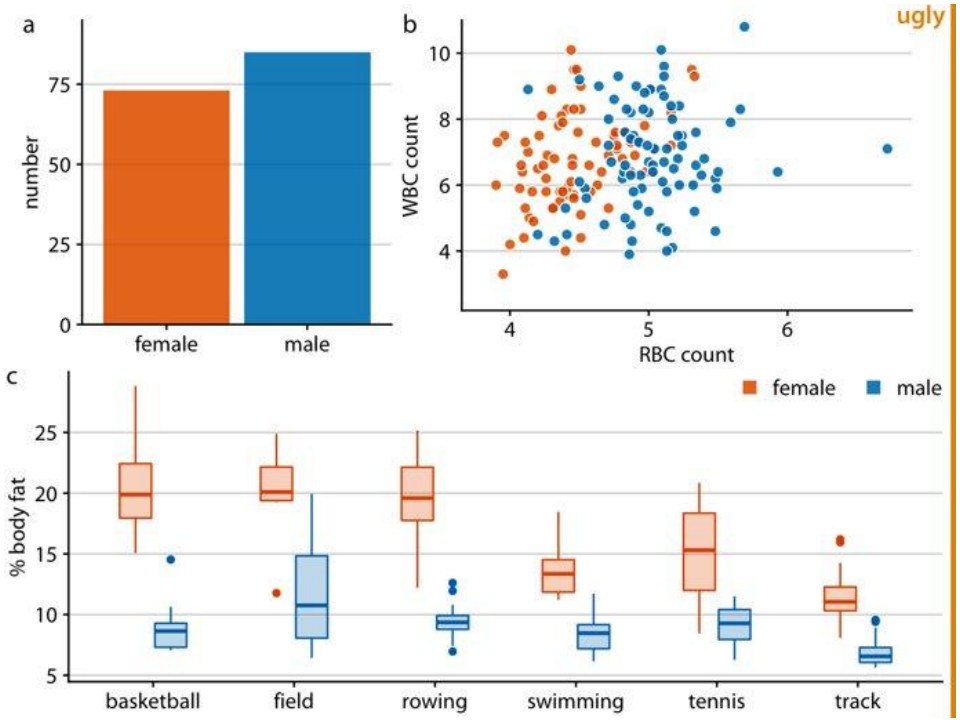
- The visual language needs to be consistent between the different parts



https://clauswilke.com/dataviz/

Compound figures

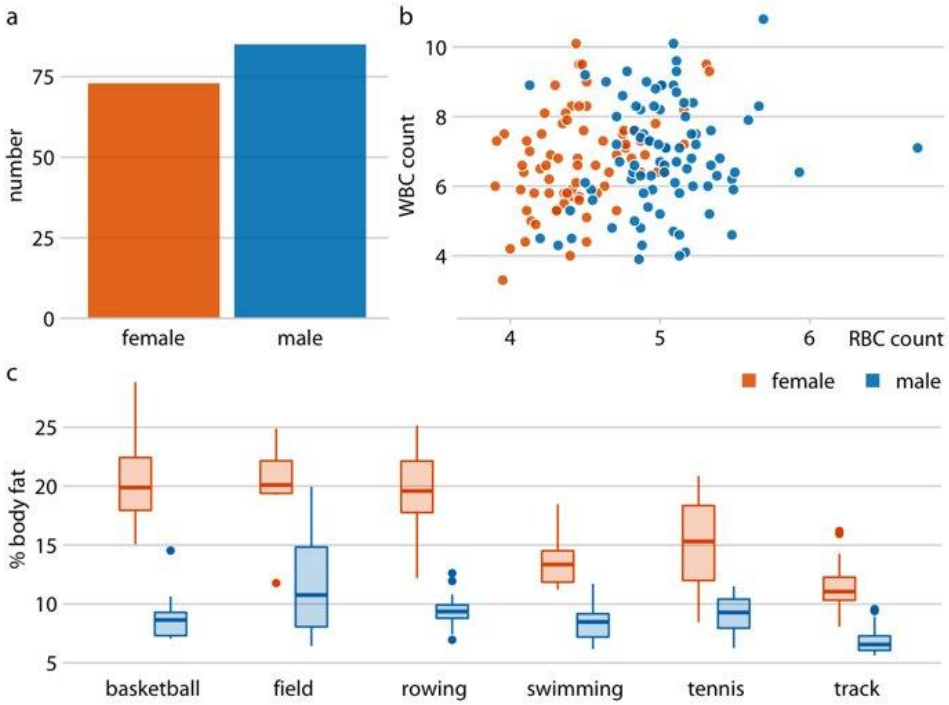
- The elements need to be connected logically
- The visual language needs to be consistent between the different parts
- Alignment of the visual elements (axis, grids) makes it less visually taxing



https://clauswilke.com/dataviz/

Compound figures

- The elements need to be connected logically
- The visual language needs to be consistent between the different parts
- Alignment of the visual elements (axis, grids) makes it less visually taxing



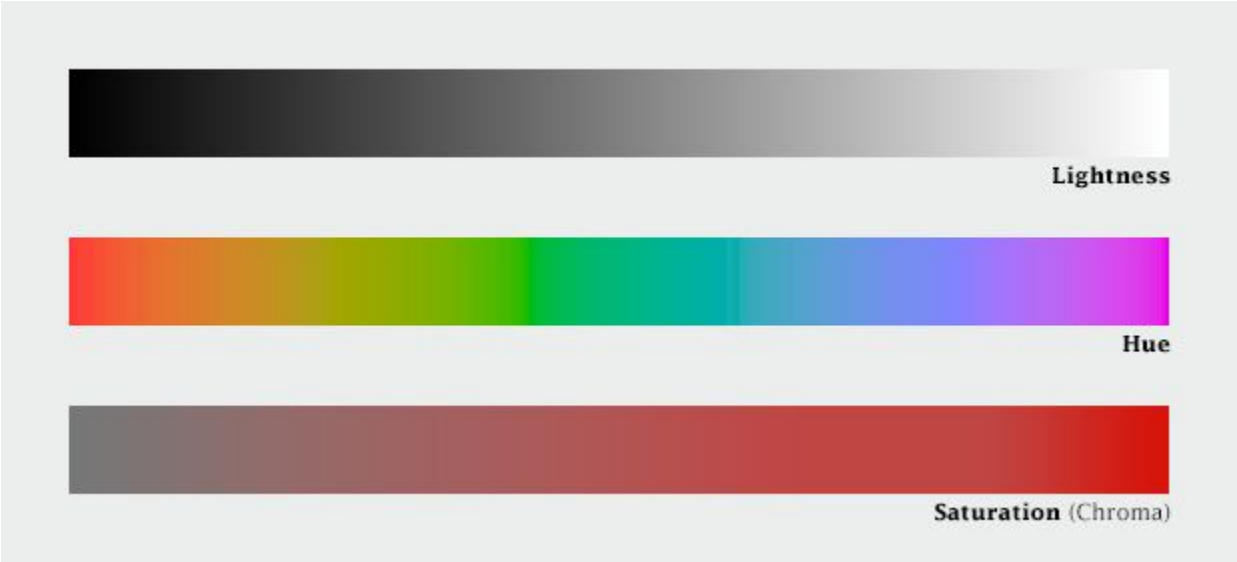
https://clauswilke.com/dataviz/

Important aspects to consider

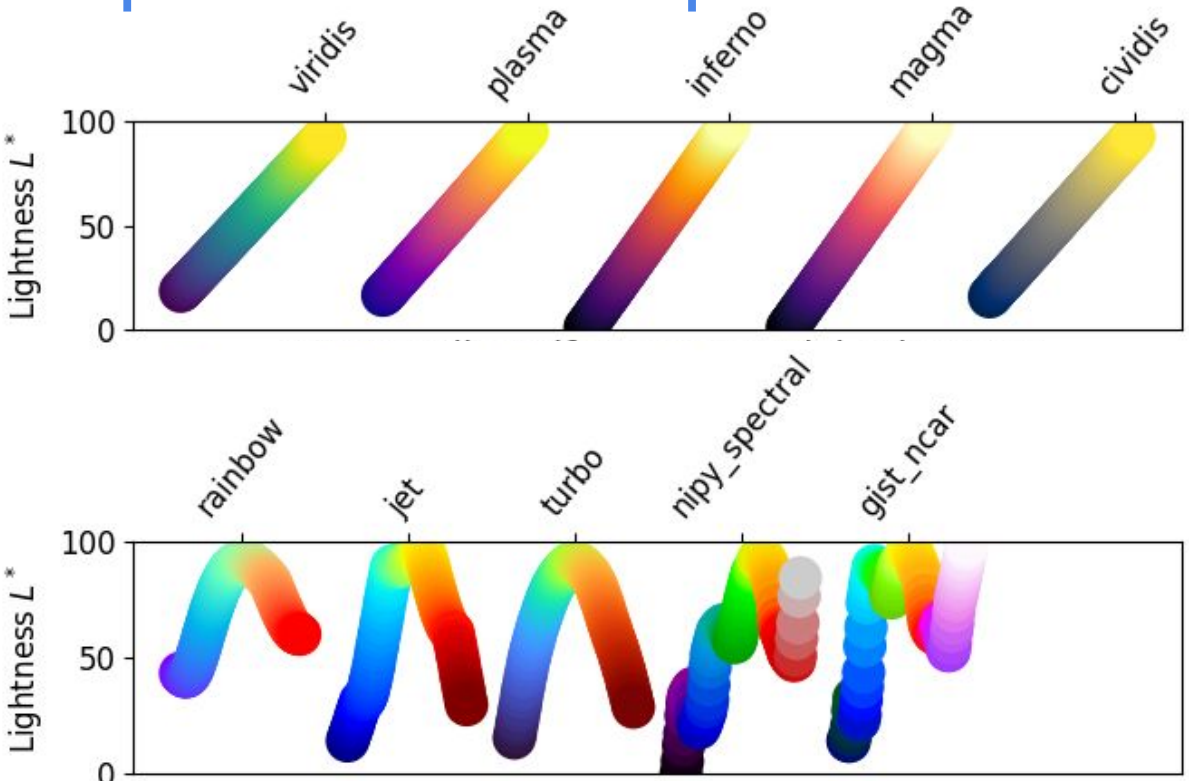
- Match the graph
- Color choices
- Fonts size, notations, axis labels
- Wise use of the figure caption to complement the figure
- Think of reusability to increase your impact

Question the defaults!

Color perception - Basically 3 dimensions



Color perception - Colormaps



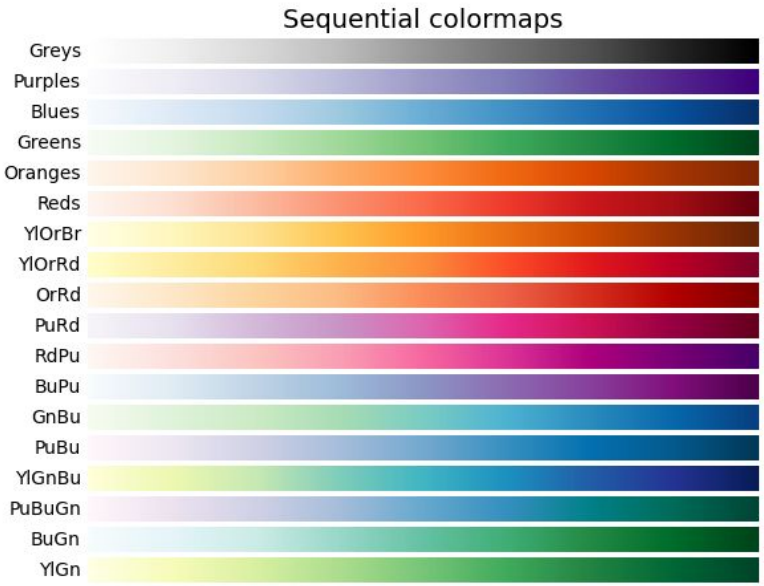
Choosing Colormaps in Matplotlib

4 types of color sequences

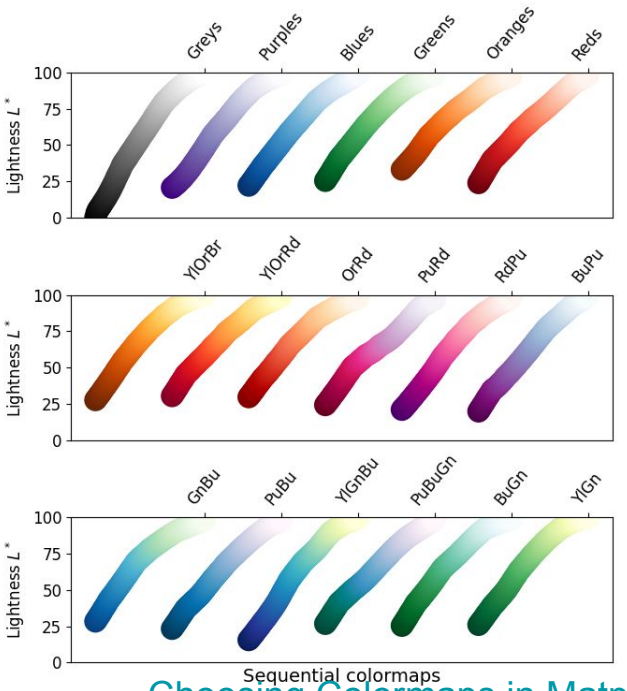
Sequential, cyclic, diverging, qualitative

4 types of color sequences

Sequential, *cyclic, diverging, qualitative*



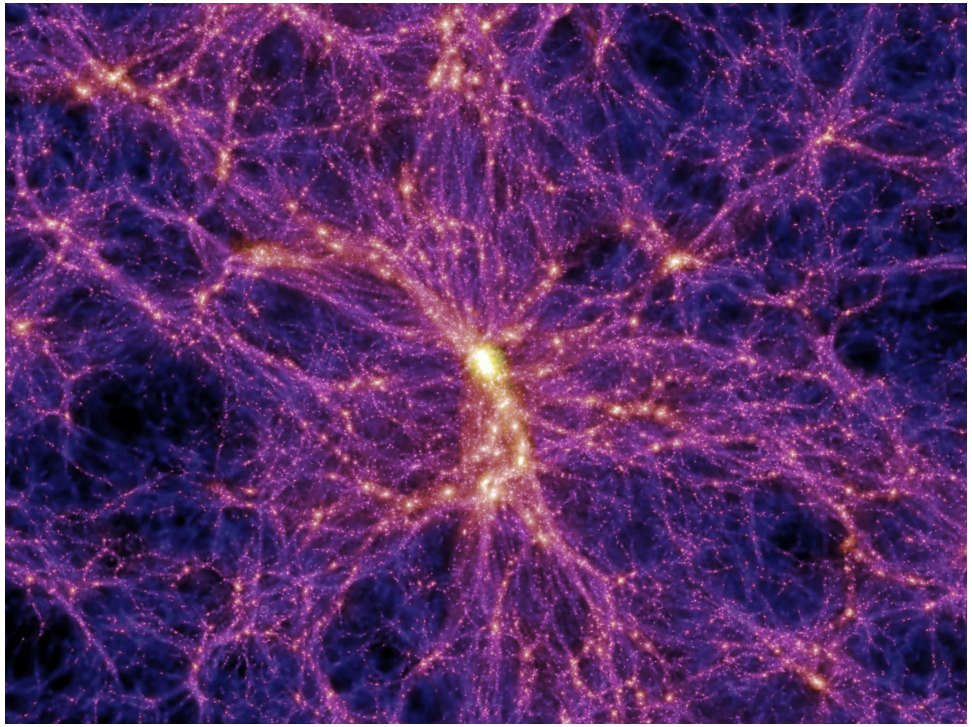
Note: *these are NOT perceptually uniform*



Choosing Colormaps in Matplotlib

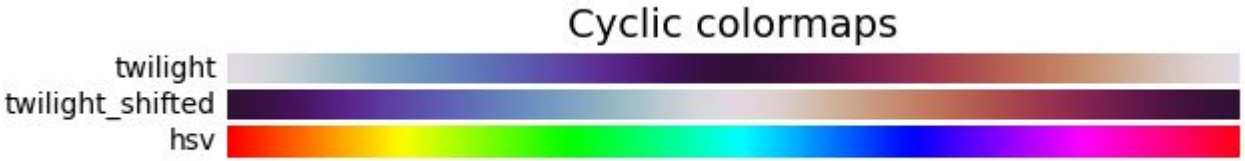
4 types of color sequences

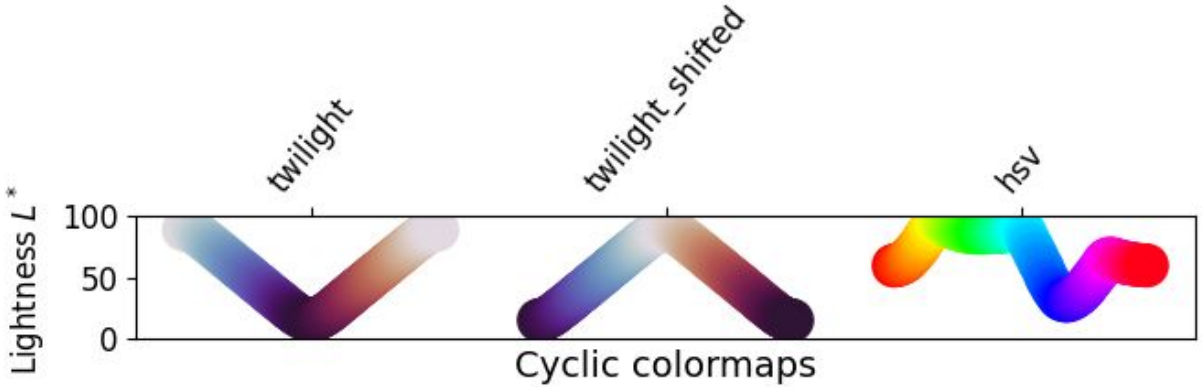
***Sequential**, cyclic, diverging, qualitative*



4 types of color sequences

*Sequential, **cyclic**, diverging, qualitative*

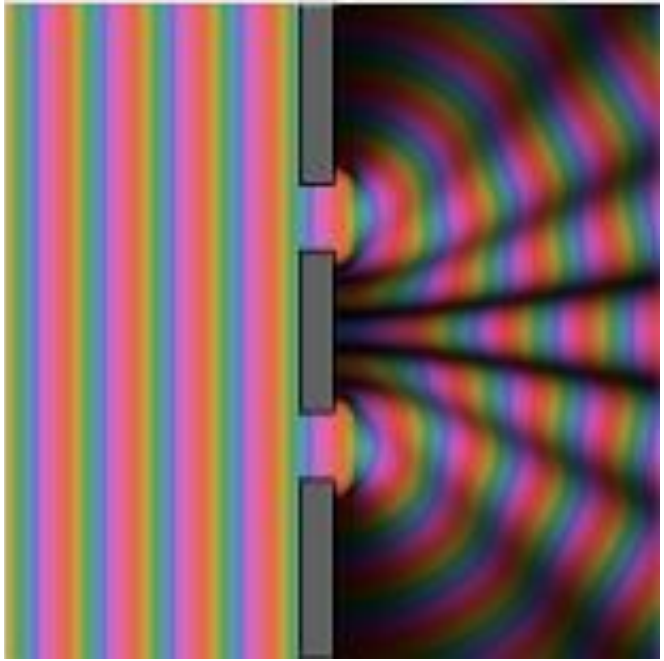




Choosing Colormaps in Matplotlib

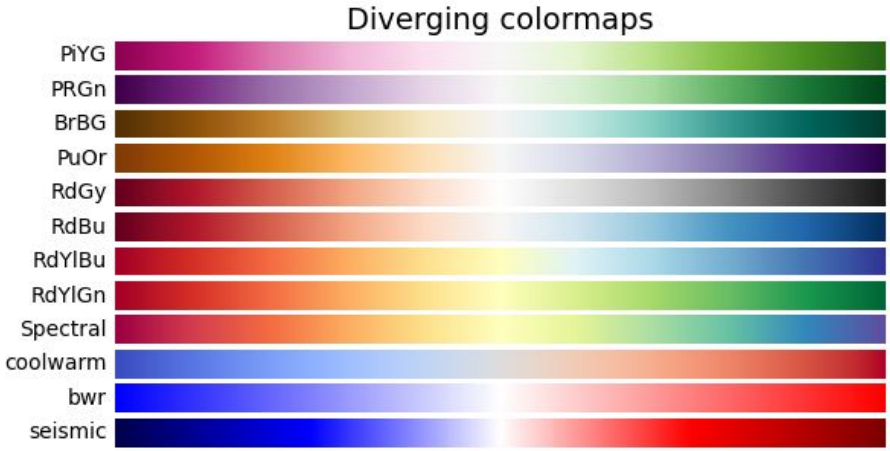
4 types of color sequences

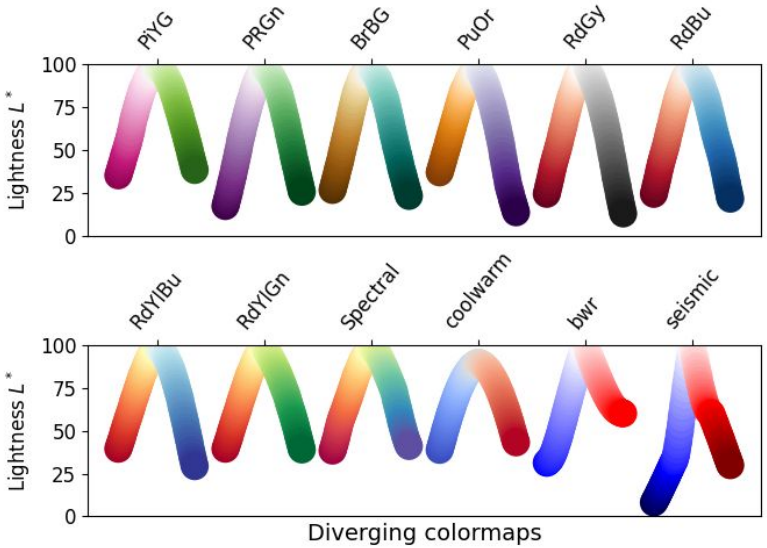
*Sequential, **cyclic**, diverging, qualitative*



4 types of color sequences

*Sequential, cyclic, **diverging**, qualitative*

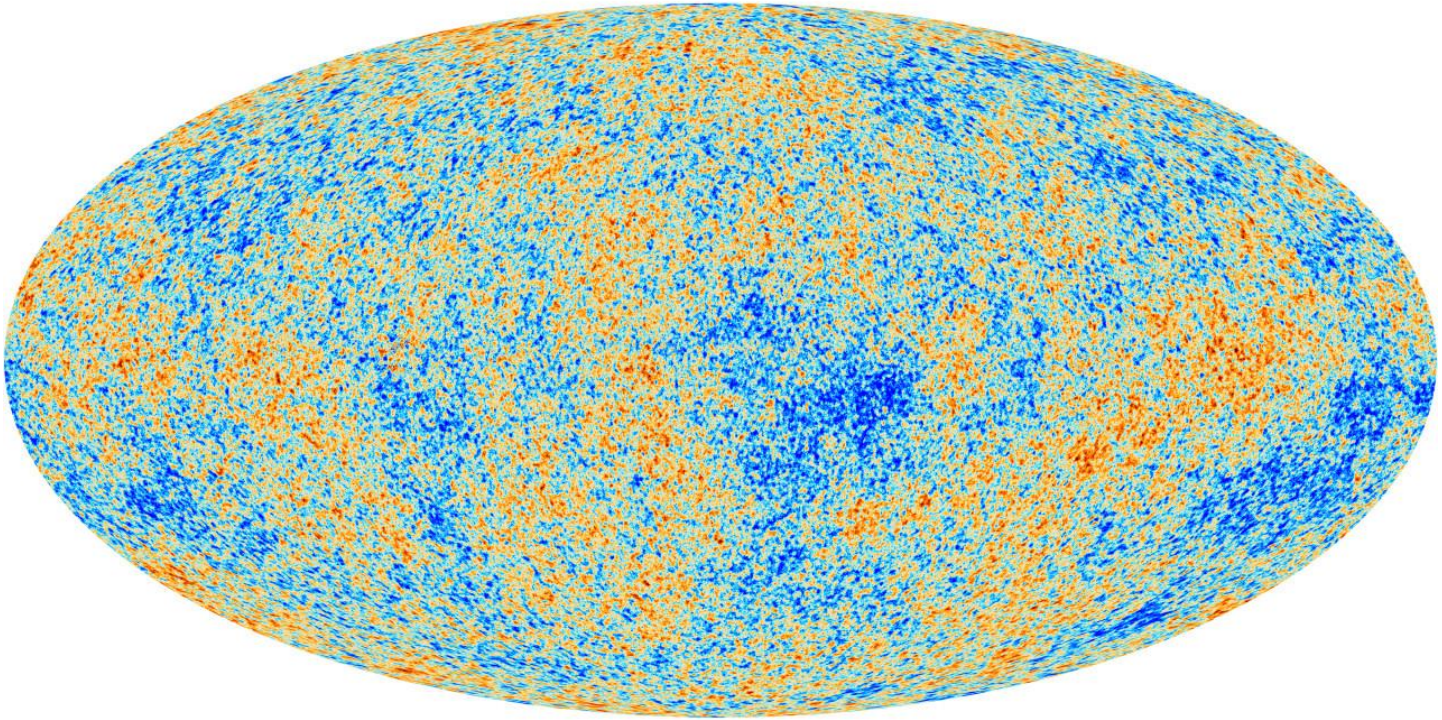




Choosing Colormaps in Matplotlib

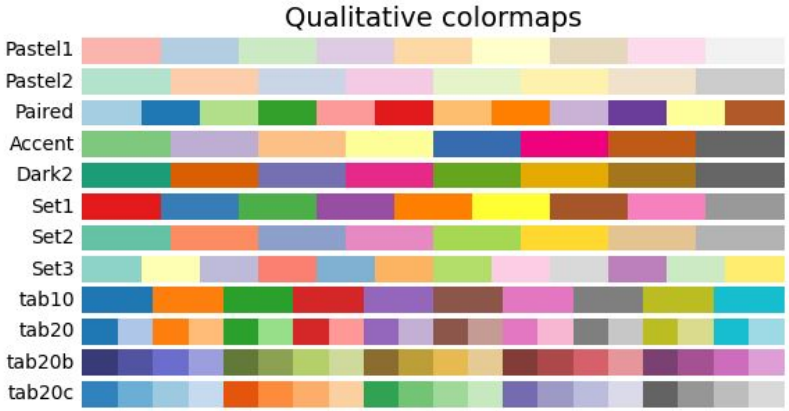
4 types of color sequences

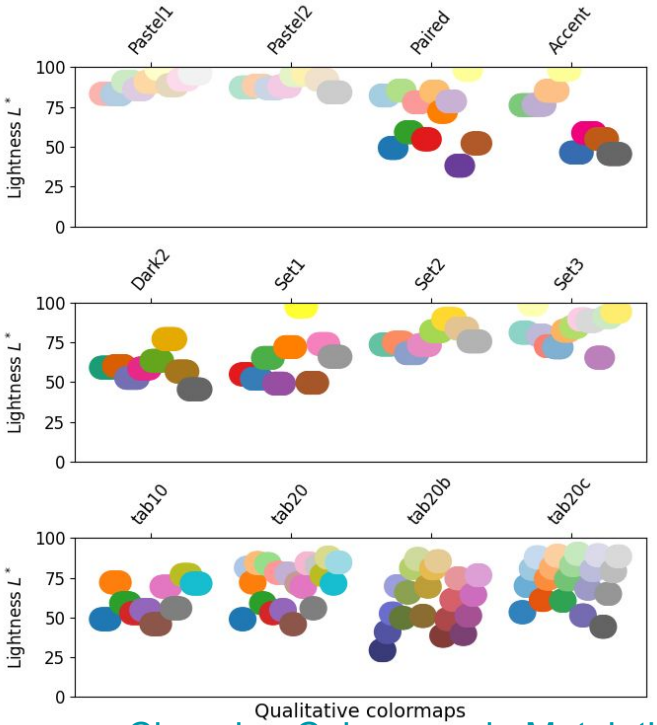
*Sequential, cyclic, **diverging**, qualitative*



4 types of color sequences

*Sequential, cyclic, diverging, **qualitative***



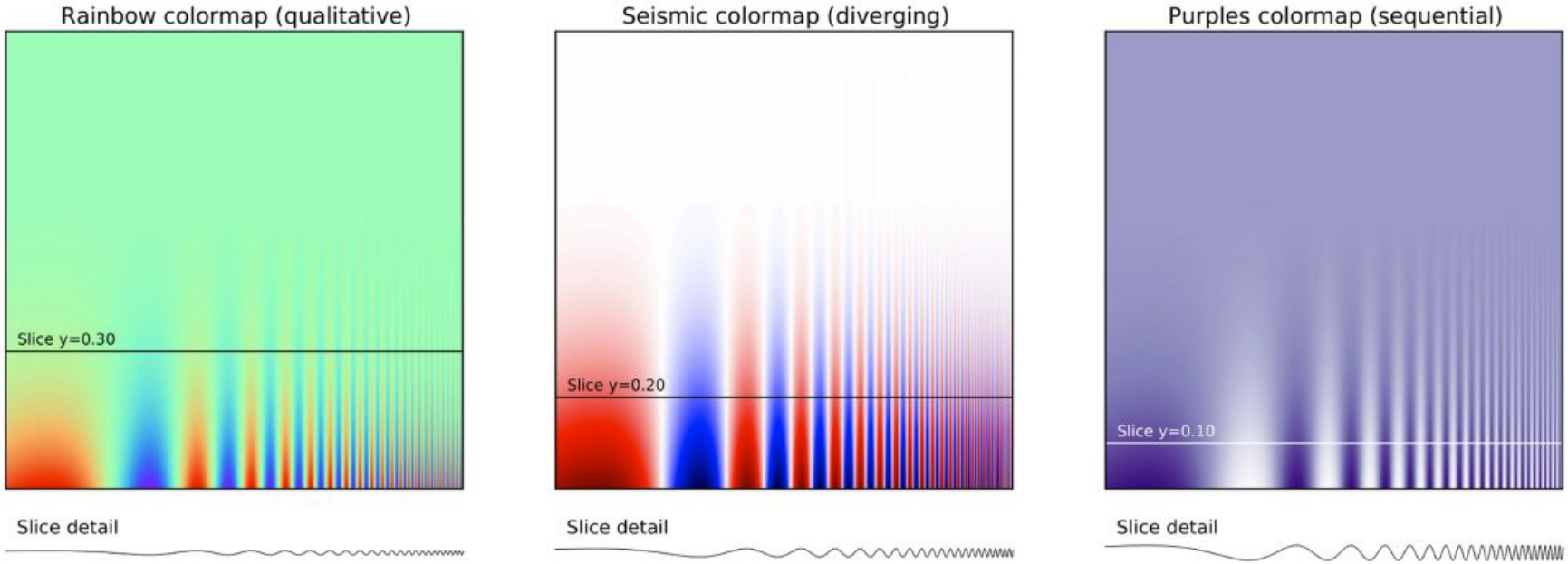


Choosing Colormaps in Matplotlib

Contrasts and nuances:

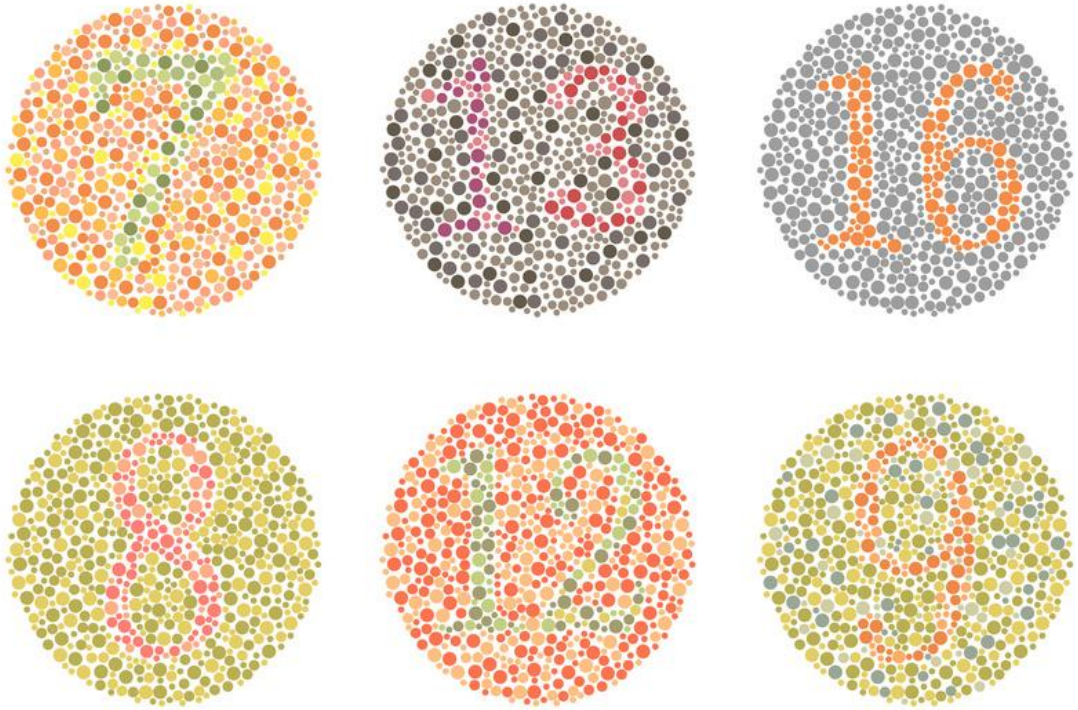
Use the complementary, analogous, triad

Color schemes may change perceptions



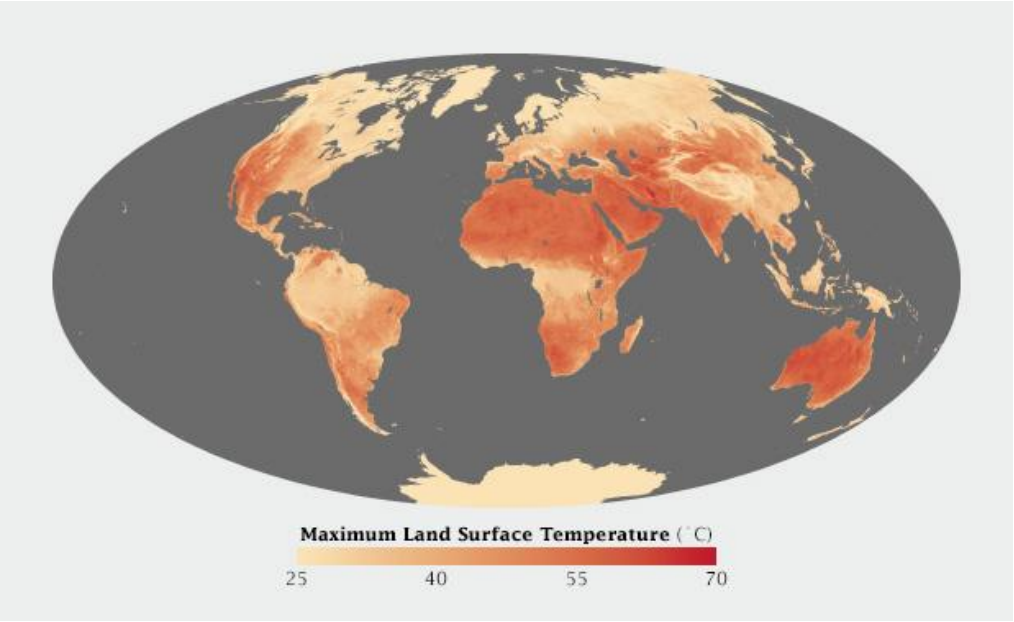
[Ten Simple Rules for Better Figures | PLOS Computational Biology]

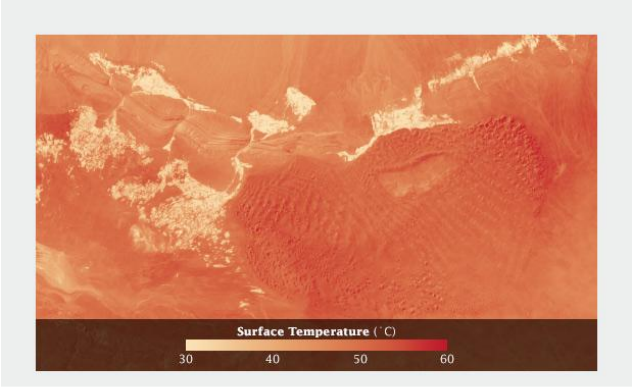
Color Blindness



- Red-green: 8% males, 0.5% females (Northern European descent)
  commonly inherited from mutations on the X chromosome,
  Source: Facts About Color Blindness
- Tom, Eduardo, Christian, Rene, ...

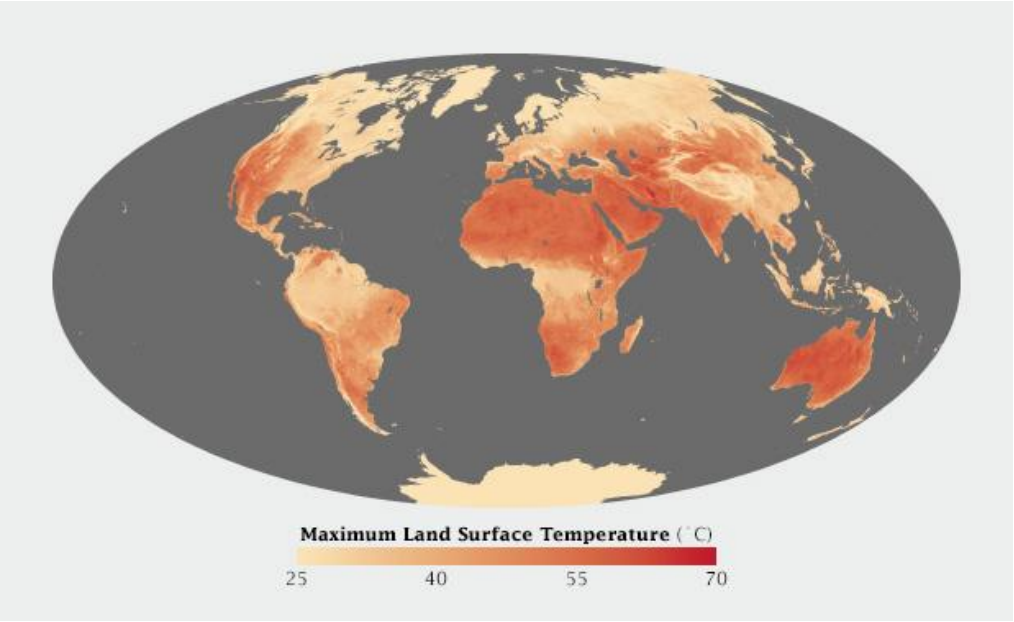
There is no general/perfect solution !

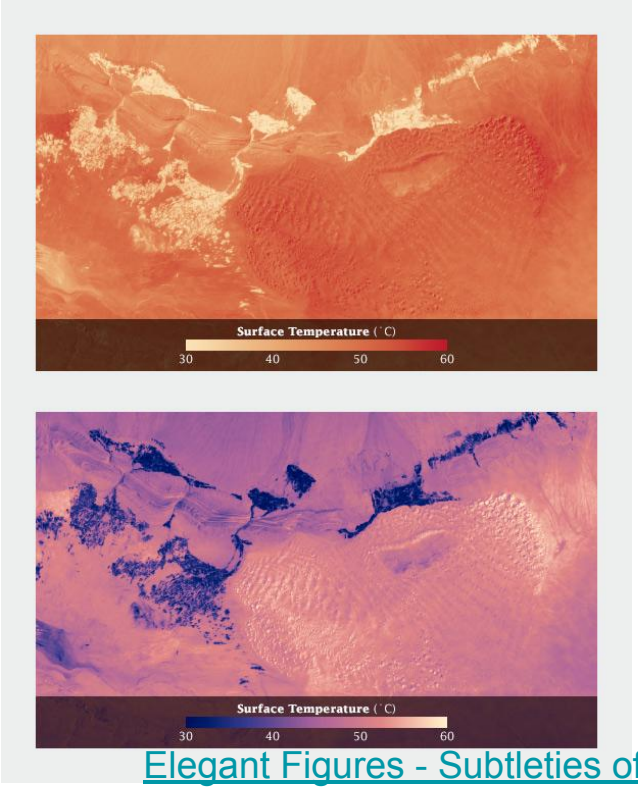




Elegant Figures - Subtleties of Color

There is no general/perfect solution !





Elegant Figures - Subtleties of Color

Some advice

- Use color **consciously** - less is more!
- Apply color **appropriately**
  - for qualitative, sequential, diverging, cyclic data
- Design color **inclusively**
  - color palettes, thickness of lines & symbols
- Remember:
  - There is no general/perfect solution.
  - Do not go against intuition.

Further reading

- Elegant Figures - Subtleties of Color (Part 1 of 6)
- COLOR IN SCIENTIFIC FIGURES
- Ten Simple Rules for Better Figures | PLOS Computational Biology
- Points of view: Color blindness | Nature Methods
- The Right Colors Make Data Easier To Read
- Rainbow Color Map (Still) Considered Harmful | IEEE Journals & Magazine
- How Bad Is Your Colormap? | Pythonic Perambulations
- How to Graph Badly or What NOT to Do

Important aspects to consider

- Match the graph
- Color choices
- Fonts size, notations, axis labels
- Wise use of the figure caption to complement the figure
- Think of reusability to increase your impact

Question the defaults!

Use **common sense** for figure size, fonts, notations, axis labels

“Efficient” communication of findings

choices affect **perception**

	Fonts name
You will always be mine	tahoma
You will always be mine	kiss me or not
You will always be mine	plasma drip

Do you **feel** the same with all fonts?

1001fonts.com

Example inspired from Don McMillan

Figure size - example

arXiv:1810.04449v2

Faster Hamiltonian Monte Carlo by Learning Leapfrog Scale

Changye Wu* · Julien Stoehr* · Christian P. Robert

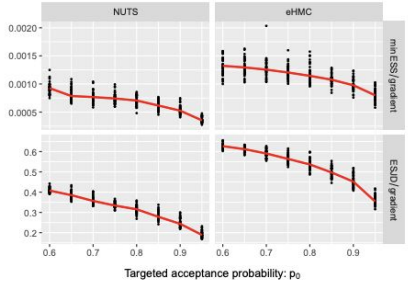


Fig. 1 Comparison between NUTS and eHMC samplers for a multivariate Normal distribution target: the x-axis is indexed by the targeted acceptance probability, p_0, associated with the primal-dual averaging step, and the y-axis in both plots corresponds to the minimum ESS across all parameter coordinates (first row) and to the ESJD (second row) per evaluation of ∇U, respectively. A black dot indicates the performance for a single experiment and the red curve is the resulting median of the 40 black-dot values for a particular p_0.

Figure size - example

arXiv:1810.04449v2

Faster Hamiltonian Monte Carlo by Learning Leapfrog Scale

Changye Wu* · Julien Stoehr* · Christian P. Robert

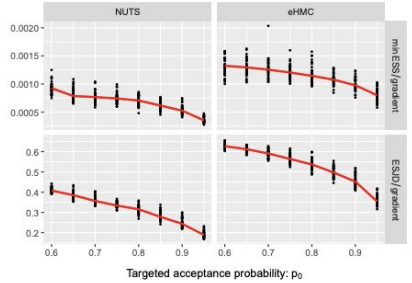


Fig. 1 Comparison between NUTS and eHMC samplers for a multivariate Normal distribution target: the x-axis is indexed by the targeted acceptance probability, p_0, associated with the primal-dual averaging step, and the y-axis in both plots corresponds to the minimum ESS across all parameter coordinates (first row) and to the ESJD (second row) per evaluation of ∇U, respectively. A black dot indicates the performance for a single experiment and the red curve is the resulting median of the 40 black-dot values for a particular p_0.

Key plot demonstrating the superiority of the eHMC sampling over NUTS

Selling well don't you think?

Figure size, use all the available space

make sure you have attractive key figures

arXiv:1810.04449v2

Faster Hamiltonian Monte Carlo by Learning Leapfrog Scale

Changye Wu* · Julien Stoehr* · Christian P. Robert

10

columnwidth

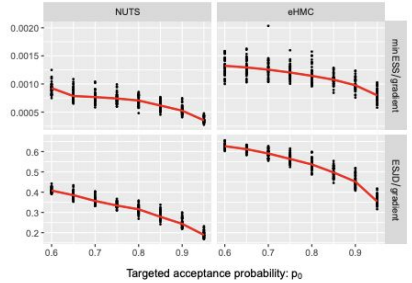
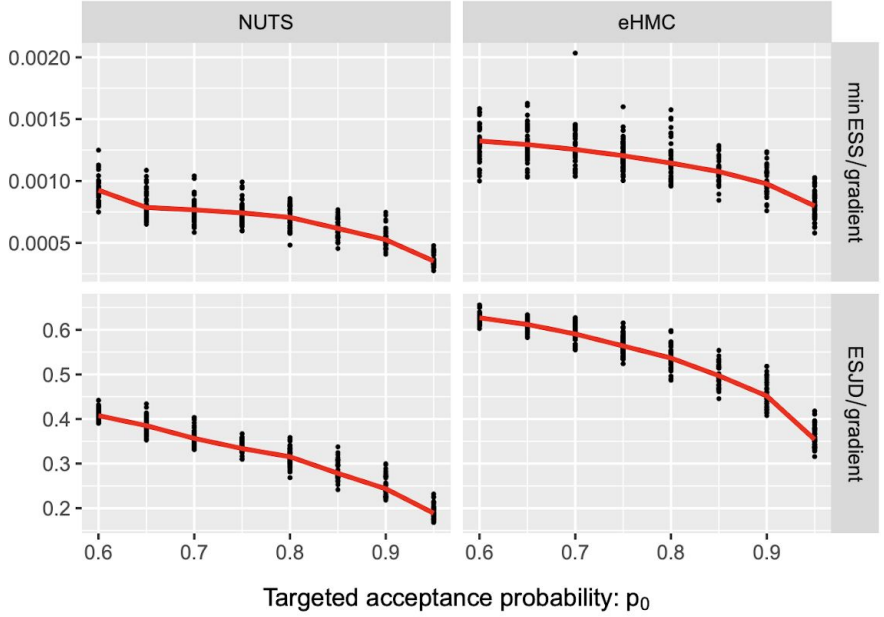


Fig. 1 Comparison between NUTS and eHMC samplers for a multivariate Normal distribution target: the x-axis is indexed by the targeted acceptance probability, p_0, associated with the primal-dual averaging step, and the y-axis in both plots corresponds to the minimum ESS across all parameter coordinates (first row) and to the ESJD (second row) per evaluation of ∇U, respectively. A black dot indicates the performance for a single experiment and the red curve is the resulting median of the 40 black-dot values for a particular p_0.

pagewidth

Fonts size - same example

Axis-labels & legends should never be smaller than the article's text



ggplot defaults

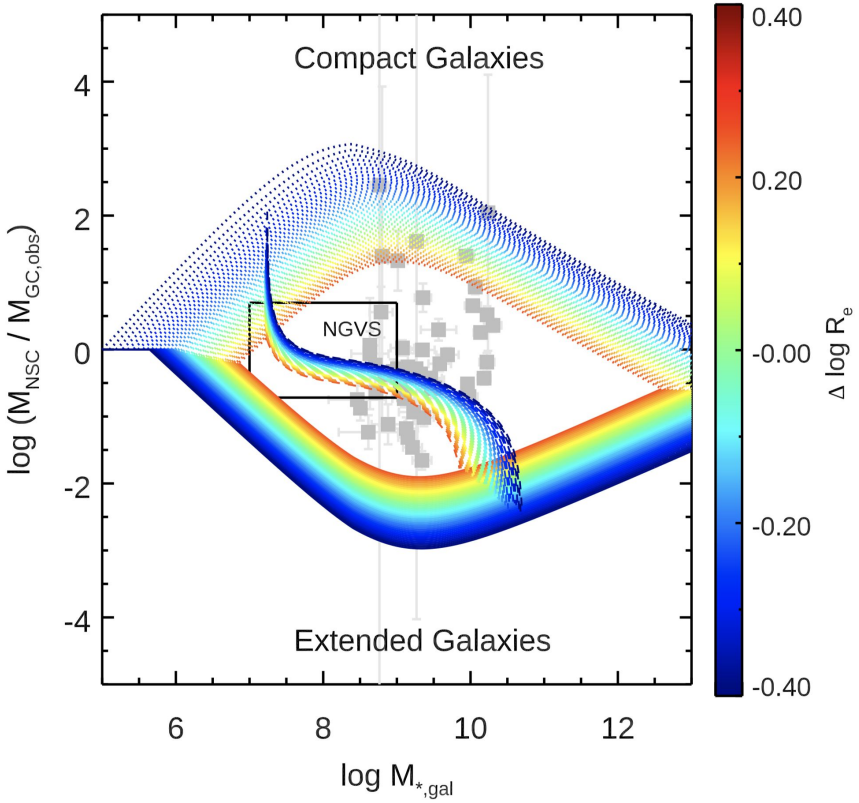
We'll come back to that

arXiv:1810.04449v2

Fig. 1 Comparison between NUTS and eHMC samplers for a multivariate Normal distribution

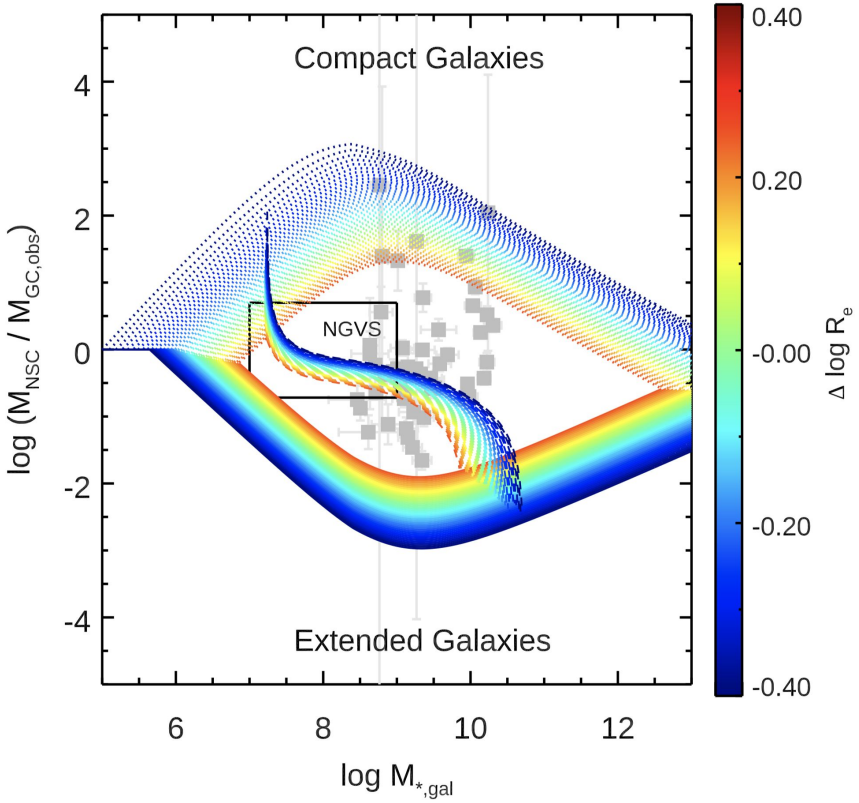
Avoid Overcrowding

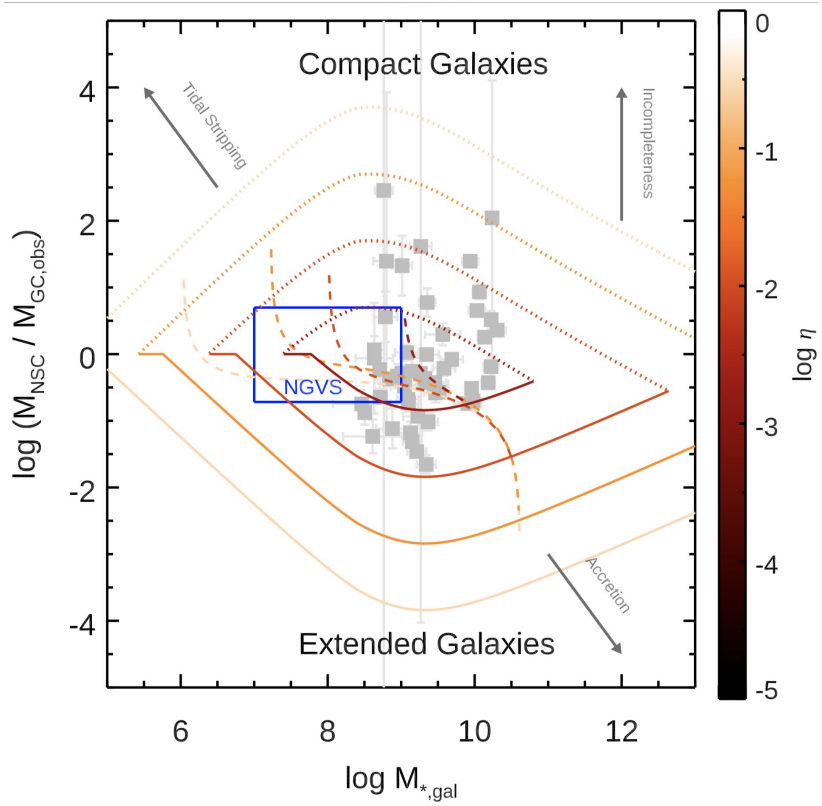
Leaman et al. 2020



Use **annotations** to help

Leaman et al. 2020





Rule of thumb: “< 30 seconds to parse”

- Pay attention to the **readability without zooming in**
  - figure size, fonts, notations, axis labels
  - Remember units

- **Use the space but avoid overcrowding**
  - Ink to space ratio > 70%
  - Simplify to the intended message

- **Less is more**
  - Stick to at most 3 colors/shape/symbols/linestyles
  - Avoid more than 4 axis (x, y, shape/color, symbol)

Warning: <u>Never</u> use “*this work*” in legends

Figures of context, reused by others in talks etc.

Important aspects to consider

- Match the graph
- Color choices
- Fonts size, notations, axis labels
- Wise use of the figure caption to complement the figure
- Think of reusability to increase your impact

Question the defaults!

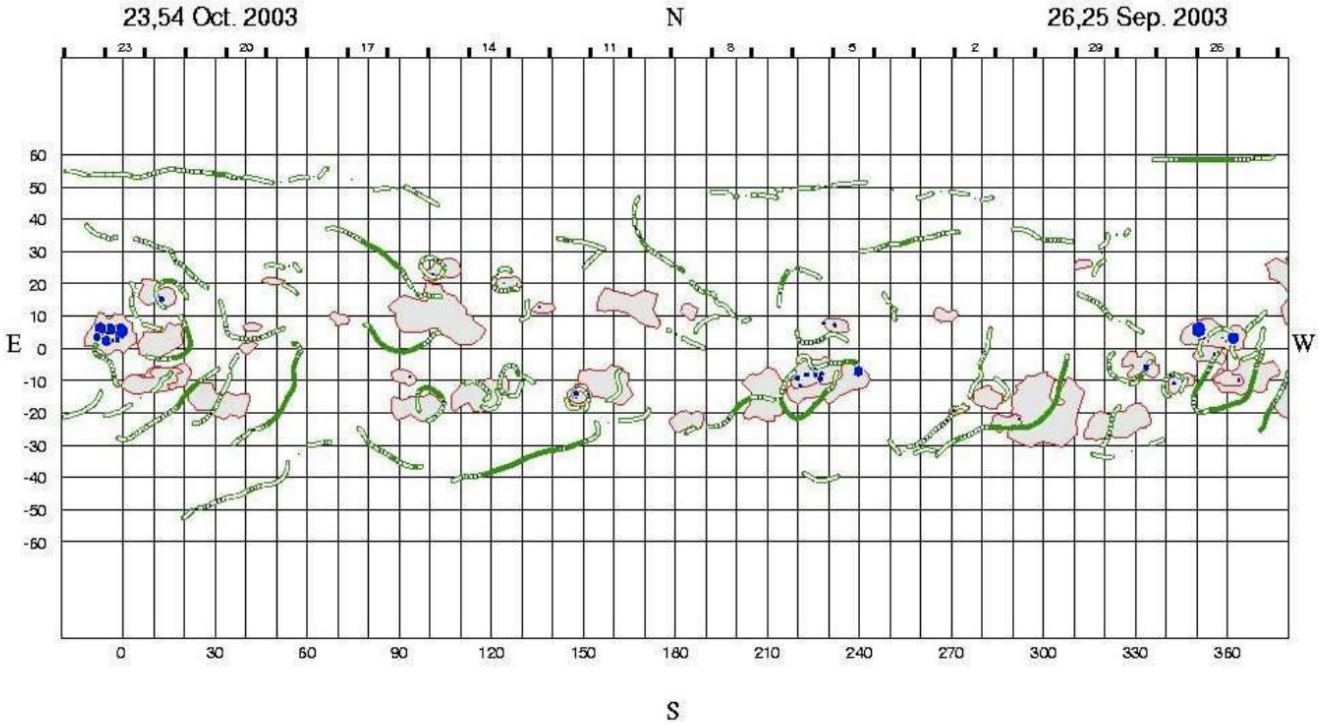
Use **wisely** figure captions to complement the figures

slido

What makes a good caption?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

Example of **bad use of caption**



Very useful! So what are the green worms living on the Sun?

Fig. 2. Example of a Meudon Synoptic Map of Solar Activity.

From Aboudarham, J., et al., 2008 (incl. Fouesneau)

Example of **bad use of caption**

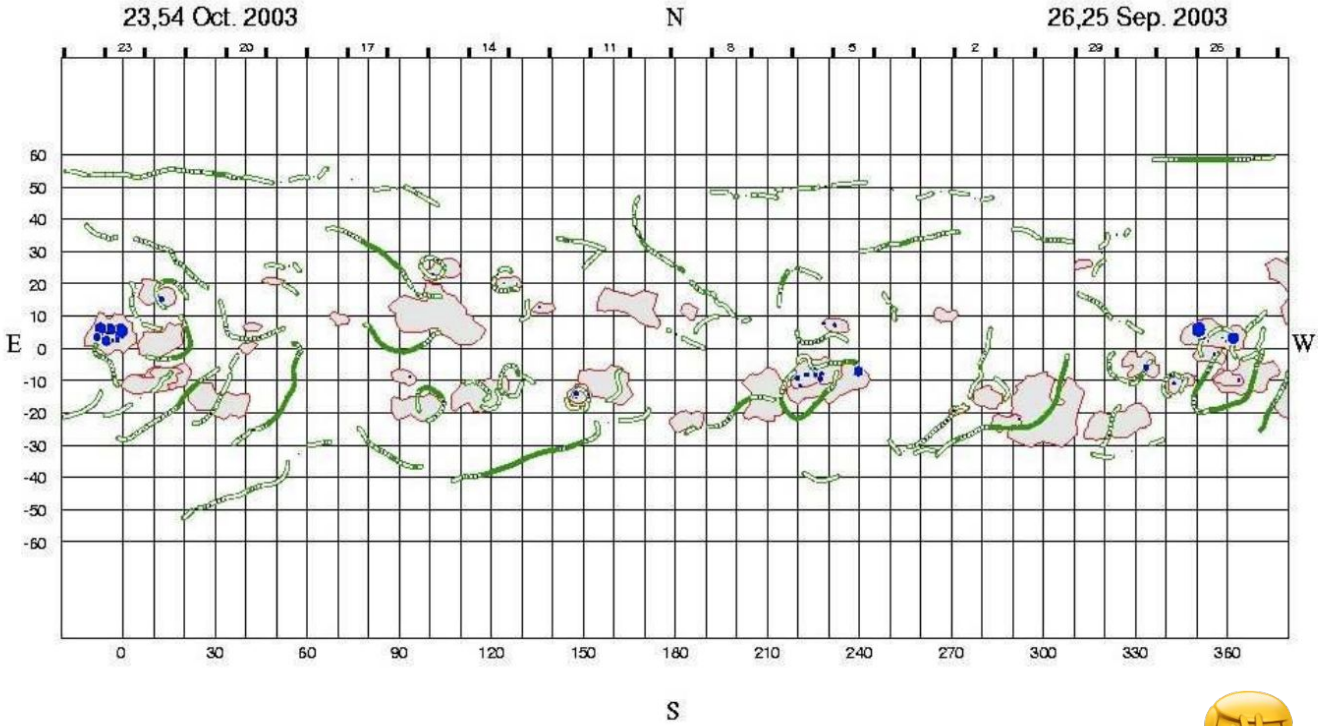


Fig. 2. Example of a Meudon Synoptic Map of Solar Activity.

Very useful!
So what are the green worms living on the Sun?

I was so young and blindly walking in the steps of the senior scientists...

From Aboudarham, J., et al., 2008 **(incl. Fouesneau)**

Example of **corrected** caption

Example of an automated Meudon Synoptic Map of Solar activity. Automated construction detailed in sect. 2. This map summarizes the time tracking of the filaments (green), the faculae (gray), and sunspots (blue) in the Carrington reference frame (see sect. 1). Solar features are to scale. Filament's higher hashing density encodes higher magnetic intensity (sect. 2).

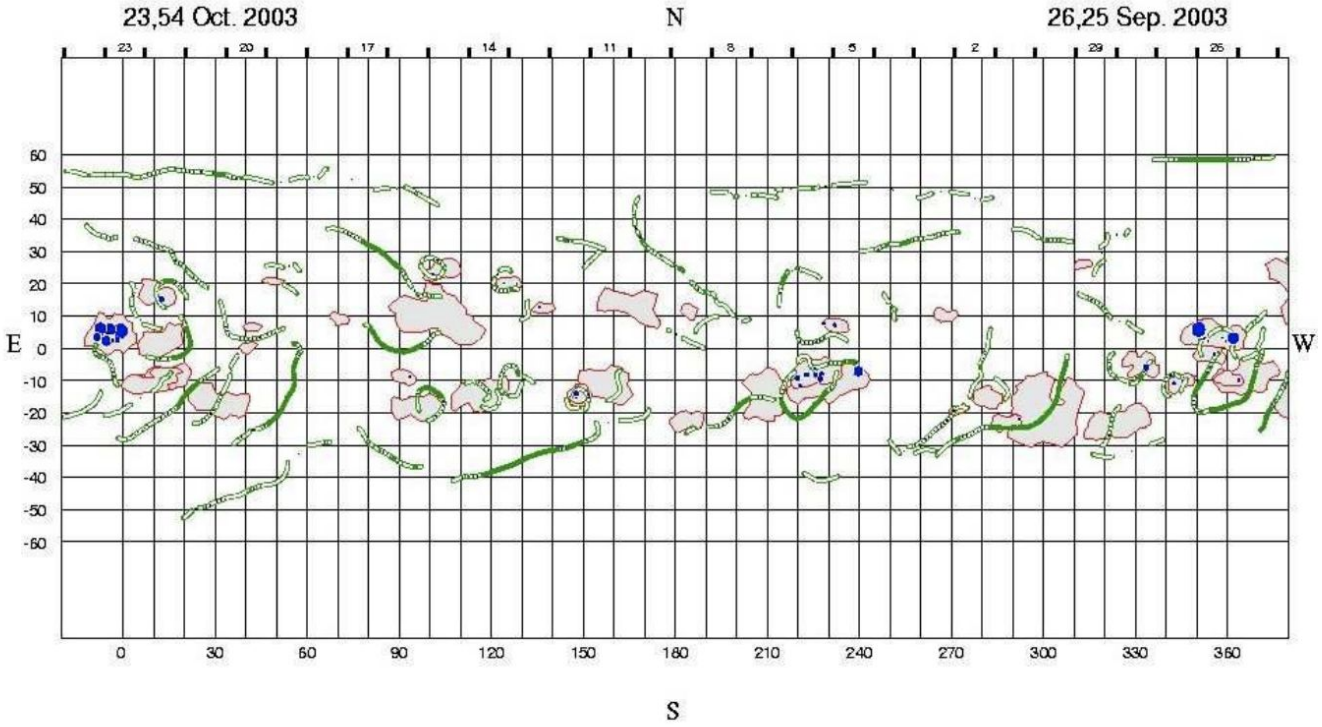


Fig. 2. Example of a Meudon Synoptic Map of Solar Activity.

Elements of **corrected** caption

Example of an automated Meudon Synoptic Map of Solar activity. Automated construction detailed in sect. 2. This map summarizes the time tracking of the filaments (green), the faculae (gray), and sunspots (blue) in the Carrington reference frame (see sect. 1). Solar features are to scale. Filament's higher hashing density encodes higher magnetic intensity (sect. 2).

Take away-message

Note the text links to find rapidly details if interested.

Describes the various symbols and colors

Warning: you do not always write a proposal!

Figure simplified as much as possible.

Annotated

and very detailed caption.

Clear take-away message

Never a good practice for a paper!

Example from a
granted 300 HST orbit proposal

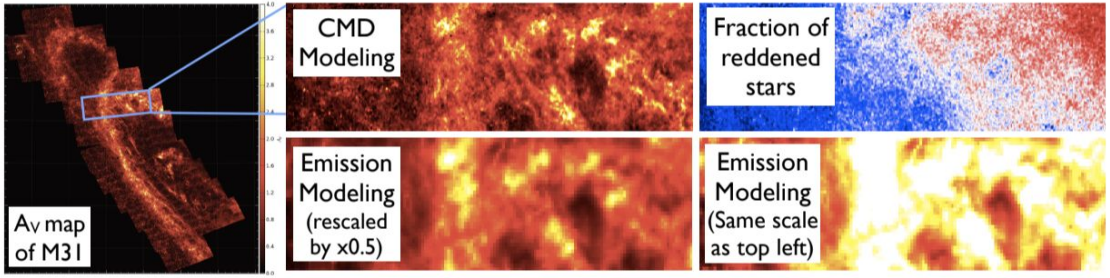


FIGURE 5: *Dust Extinction from NIR CMDs vs from MIR/FIR Emission:* We infer the A_V column in 25 pc pixels by modeling the distribution of RGB stars in a CMD (lower left) as an unreddened foreground and a reddened background that samples a log-normal distribution of A_V. **Left:** The resulting map of the mean A_V of the reddened stars for M31. **Top Middle:** Detail of the map derived from the NIR CMD. **Bottom Middle & Right:** Comparison with Draine et al. (2014) maps derived from dust emission, divided by 2 to better show the agreement in the morphology of structures (left). However, the true scale (right), shows the dramatic overprediction of the extinction by the emission-based model. **Top Right:** The fraction of reddened stars (where redder = higher reddening fraction) derived from the NIR CMD. The spatial variation is due to the geometry of a thick inclined disk, and thus constrains the 3-dimensional structure of the stellar disk. This proposal will allow identical mapping for M33, for a less-inclined galaxy with lower metallicity, higher SFR density, and spiral arms.

Good captions should

- **Describe the figure** and **<u>all</u> its relevant part**/components
  - left panel shows x in farenheit vs. y in candella color coded by z in stones. Solid lines represent …
- **State the takeaway message** of the figure!
  - A good figure will make it redundant but “teaching is the art of repetition”
- **Standalone** without the rest of the paper
  - Refer to the text sections/equations for specific details. (easier to skim your papers)
- Clearly **mention units** either on the figure or in the caption
  - What magnitude system is on your color-magnitude diagram? (Vega, AB, ST…)

Warning: do not use with excess

complement the figure but not a paragraph of the main text (unless writing a proposal)

Important aspects to consider

- Match the graph
- Color choices
- Fonts size, notations, axis labels
- Wise use of the figure caption to complement the figure
- Think of reusability to increase your impact

Question the defaults!

Discussion: Designing for Reusability

Pros:

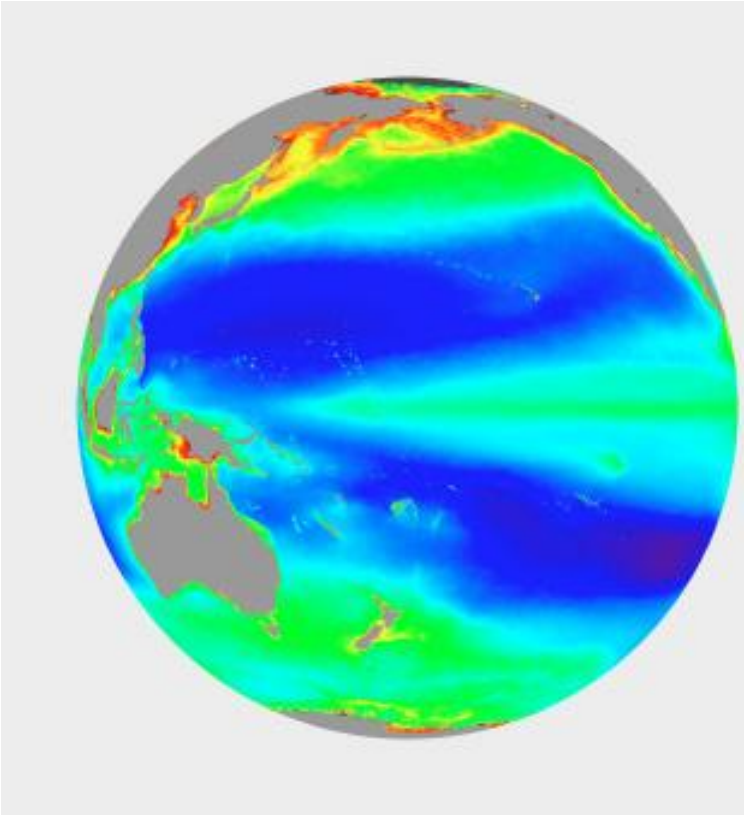
- One can design figures for both papers and talks.
- Ease to make a presentation out of a paper.

Cons:

- designing the same figure for both talk and paper may not use optimally both media.

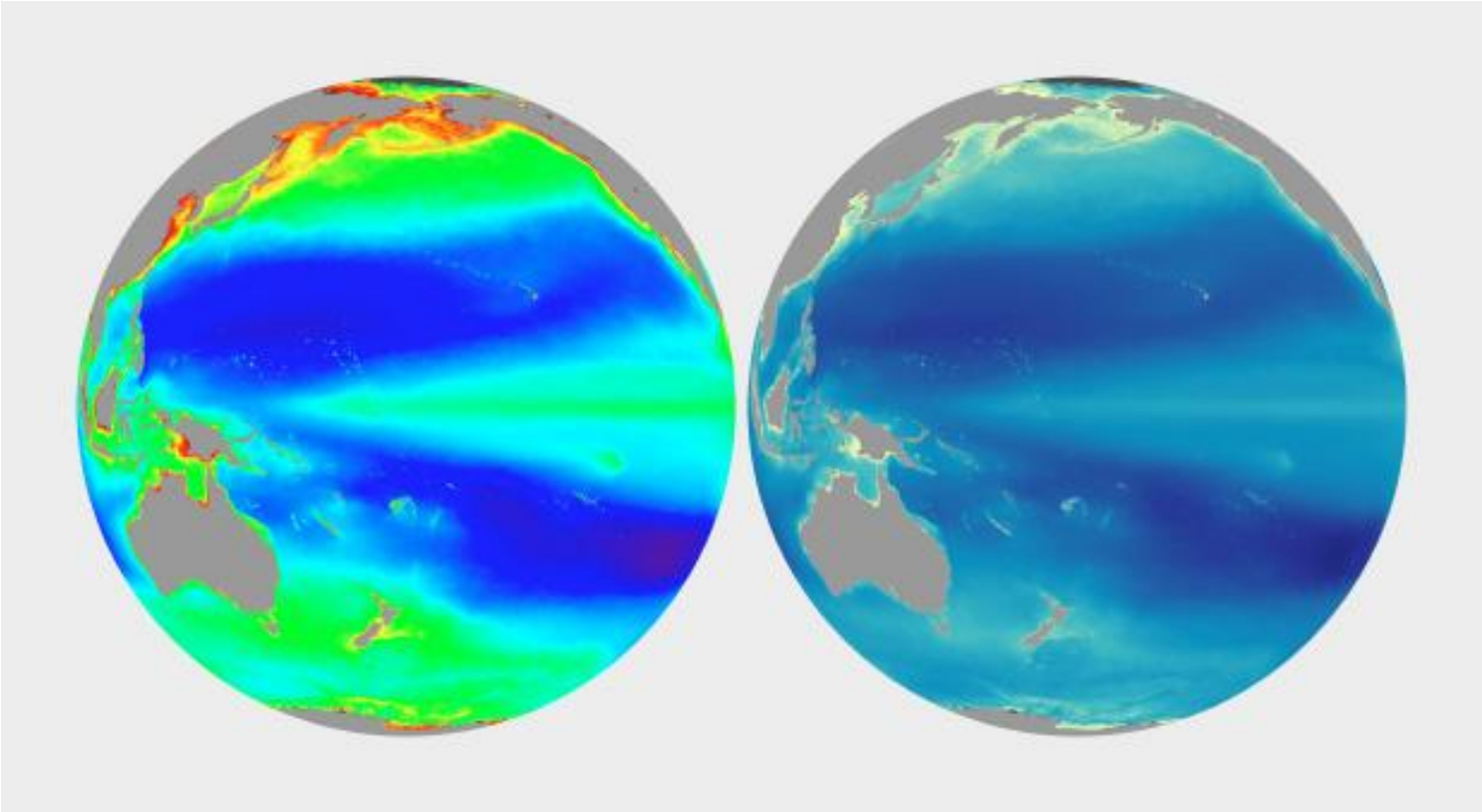
Do not go against **intuition** or **conventions**

Avoid going against Intuition !



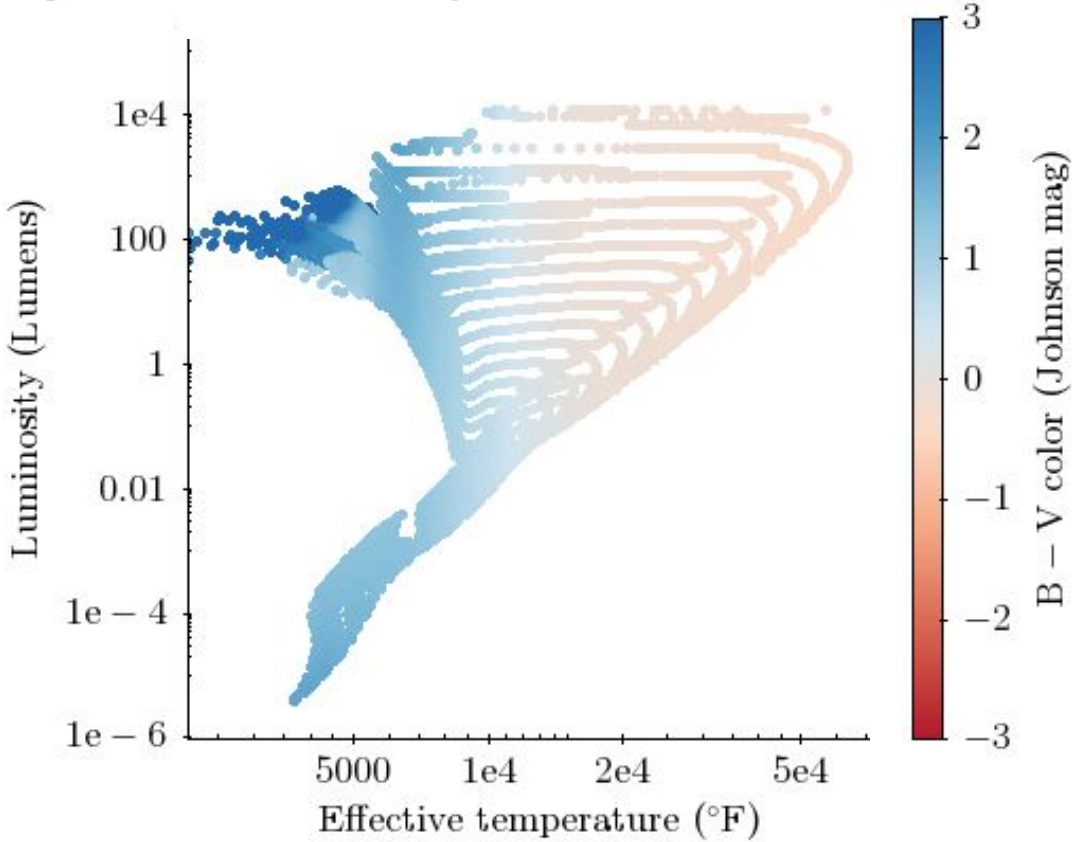
How long did you take to see colors associate with water?

Avoid going against Intuition !



[2]

Astronomy has strange conventions but still!



“Efficient” communication of findings

Keep in mind that **visualization** is always a **biased / restricted view**

Try various angles,
esp. when complex data sets.

slido

What are the most important rules for you?

Click **Present with Slido** or install our Chrome extension to activate this poll while presenting.

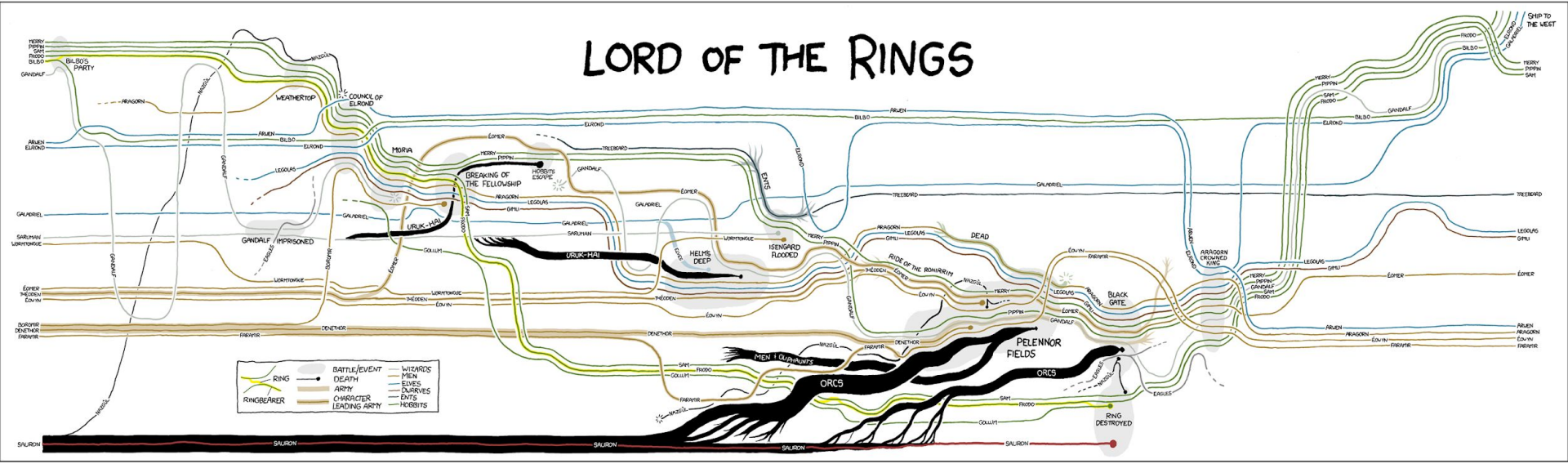
Most important rules

- Need **less than 1 minute to present** a figure
  - Decompose into panels if needed
  - Annotations: balance between cartoonish and pure data
  - Don't go against intuition or culture/conventions

- **No more than 2 messages!** (only one is best but not always feasible)
  - Pick appropriate colors for your message (sequential, divergence,...)
  - About 3 color max to be colorblind friendly, gradient if need more.

- **Use the caption wisely,** figure-to-caption-text ratio > 1
  - e.g., add the message of the figure, refer to text section(s) etc.

- Think about **reusability**:
  - figure for paper used in your talk **and other's talks**.
  - Make it **easy to dissect** = more citations!

Know the rules to break them all!

THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.