McGill initiative in Computational Medicine

HEALTHY BRAINS FOR HEALTHY LIVES

A picture is worth 1000 statistics
Effective data visualization for scientific communication

Eisha Ahmed, PhD Candidate

Department of Experimental Medicine, McGill University

HBHL Science Communication Day – January 24, 2020

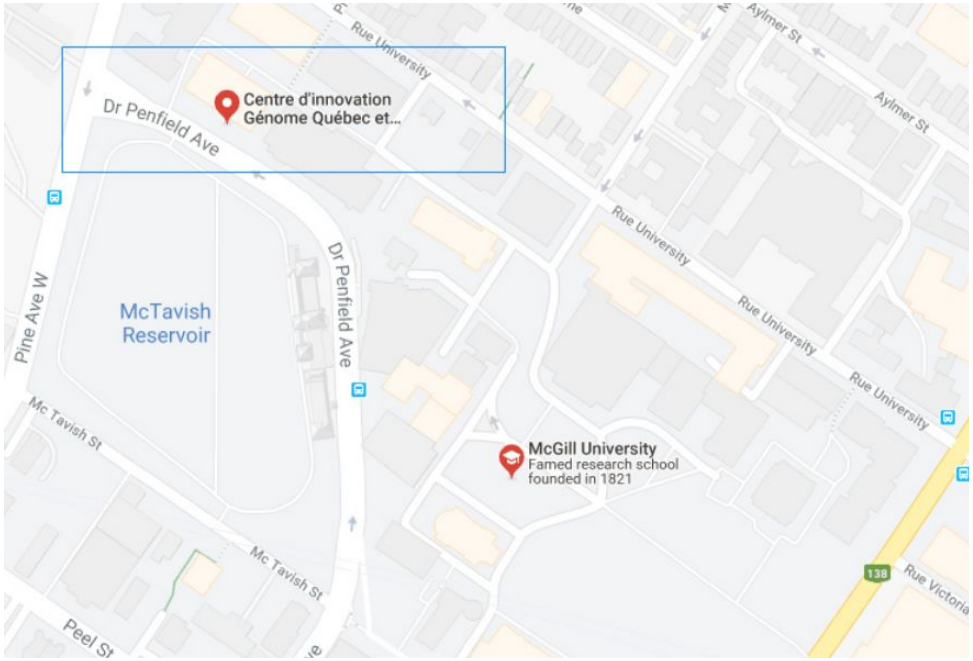
Mission: To deliver inter-disciplinary research programs and empower the use of data in health research and health care delivery.

McGill initiative in Computational Medicine

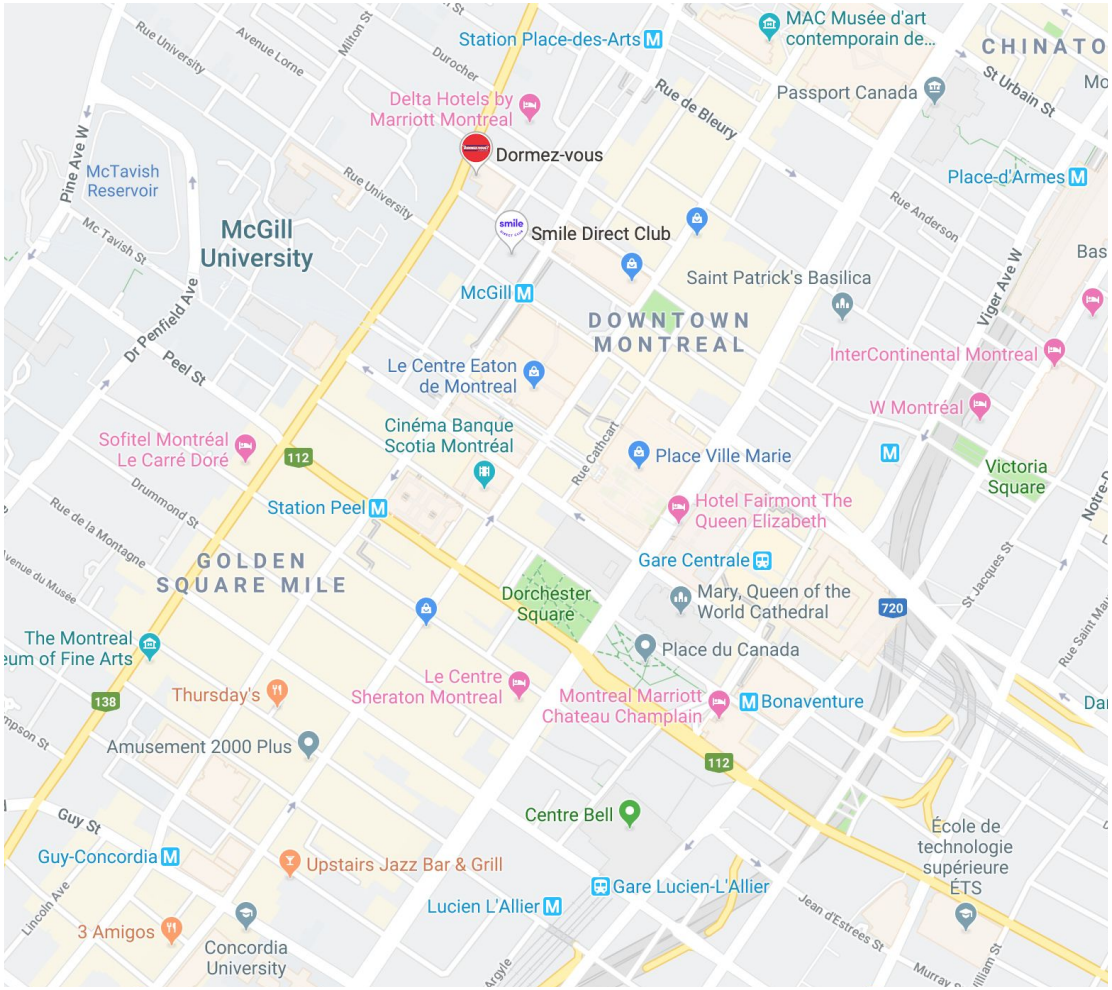
740, Dr. Penfield Avenue
Montreal, QC
Canada, H3A OG1

info-micm@mcgill.ca

https://www.mcgill.ca/micm



What is **data visualization (DV)**?



Google Maps

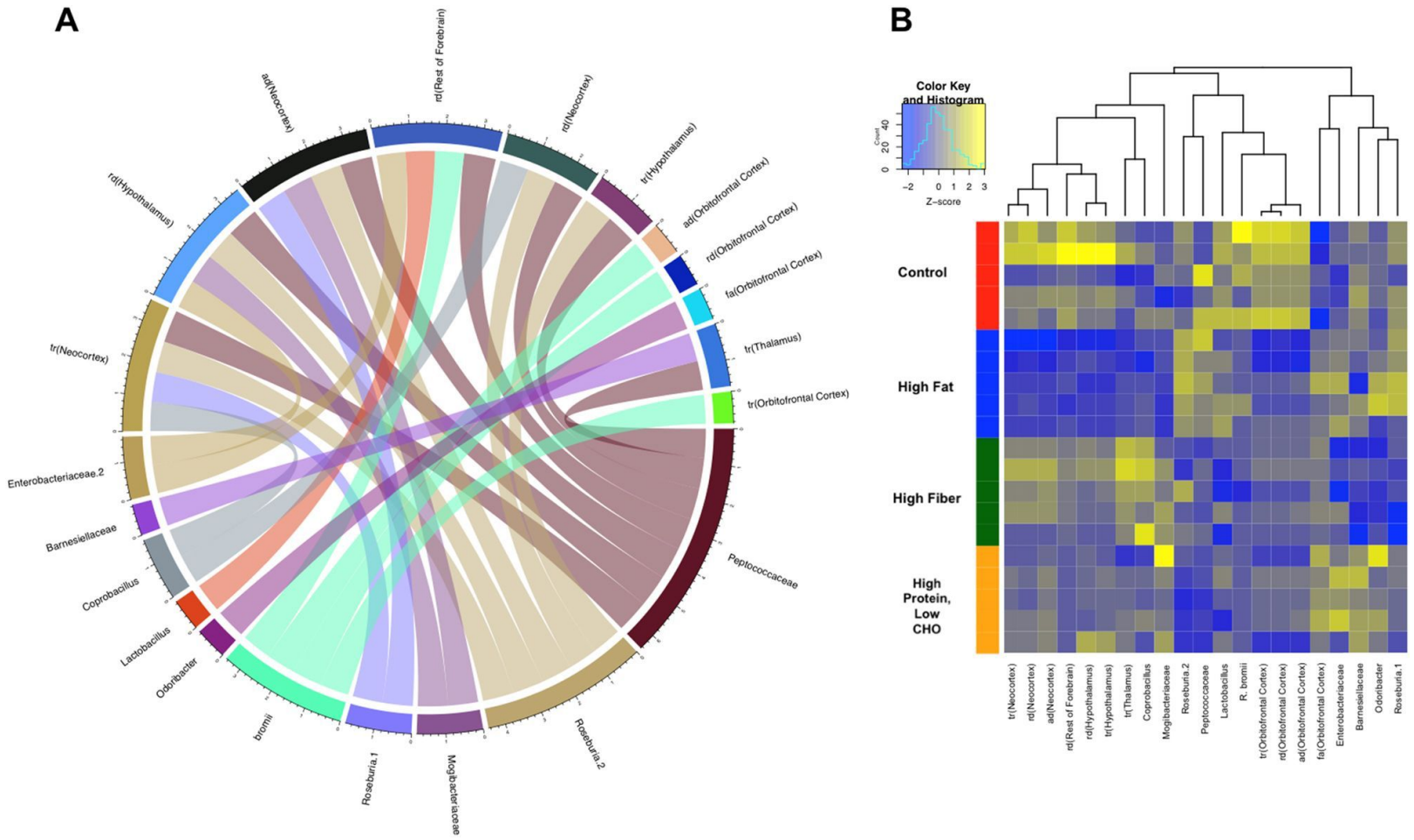
Is this data visualization?

Region, subregion, country or area *	Notes	Country code	Type	Parent code	Total population, both sexes combined, as of 1 July (thousands)					
					1950	1951	1952	1953	1954	1955
WORLD		900	World	0	2 536 431	2 584 034	2 630 862	2 677 609	2 724 847	2 773 020
UN development groups	a	1803	Label/Separator	900	...	...	...	...	...	...
More developed regions	b	901	Development Grou	1803	814 819	824 003	833 720	843 788	854 060	864 430
Less developed regions	c	902	Development Grou	1803	1 721 612	1 760 031	1 797 141	1 833 821	1 870 786	1 908 590
Least developed countries	d	941	Development Grou	902	195 428	199 180	203 015	206 986	211 133	215 486
Less developed regions, excluding least developed	e	934	Development Grou	902	1 526 184	1 560 850	1 594 126	1 626 836	1 659 653	1 693 104
Less developed regions, excluding China		948	Development Grou	1803	1 157 420	1 179 933	1 203 963	1 229 440	1 256 303	1 284 497
Land-locked Developing Countries (LLDC)	f	1636	Special other	1803	103 803	105 870	108 079	110 423	112 894	115 488
Small Island Developing States (SIDS)	g	1637	Special other	1803	23 771	24 209	24 684	25 187	25 710	26 249
World Bank income groups		1802	Label/Separator	900	...	...	...	...	...	...
High-income countries	h	1503	Income Group	1802	694 989	703 004	711 534	720 436	729 596	738 929
Middle-income countries	h	1517	Income Group	1802	1 703 597	1 741 086	1 777 129	1 812 536	1 847 973	1 883 962
Upper-middle-income countries	h	1502	Income Group	1517	938 931	962 816	984 350	1 004 408	1 023 703	1 042 796
Lower-middle-income countries	h	1501	Income Group	1517	764 666	778 270	792 779	808 129	824 269	841 166
Low-income countries	h	1500	Income Group	1802	137 042	139 119	141 352	143 769	146 387	149 214
No income group available		1518	Income Group	1802	804	826	847	868	891	915
Geographic regions	i	1840	Label/Separator	900	...	...	...	...	...	...
Africa	j	903	Region	1840	227 794	232 328	237 097	242 092	247 311	252 749
Asia	k	935	Region	1840	1 404 909	1 435 819	1 464 834	1 492 895	1 520 768	1 549 042
Europe	l	908	Region	1840	549 329	554 324	559 694	565 282	570 970	576 679
Latin America and the Caribbean	m	904	Region	1840	168 821	173 281	177 916	182 709	187 648	192 727
Northern America	n	905	Region	1840	172 603	175 017	177 779	180 813	184 052	187 430
Oceania	o	909	Region	1840	12 976	13 266	13 543	13 818	14 099	14 393
Sustainable Development Goal (SDG) regions	p	1828	Label/Separator	900	...	...	...	...	...	...
SUB-SAHARAN AFRICA		947	SDG region	1828	179 007	182 373	185 912	189 614	193 474	197 489
Eastern Africa		910	Subregion	947	66 145	67 603	69 109	70 669	72 289	73 975
Burundi		108	Country	910	2 309	2 360	2 406	2 449	2 492	2 537
Comoros		174	Country	910	159	163	167	170	173	176
Djibouti		262	Country	910	62	63	65	66	68	70
Eritrea		232	Country	910	822	835	849	865	882	900
Ethiopia		231	Country	910	18 128	18 467	18 820	19 184	19 560	19 947
Kenya		404	Country	910	6 077	6 242	6 416	6 598	6 789	6 988
Madagascar		450	Country	910	4 084	4 168	4 257	4 349	4 444	4 544
Malawi		454	Country	910	2 954	3 012	3 072	3 136	3 202	3 271
Mauritius	1	480	Country	910	493	506	521	537	554	571
Mayotte		175	Country	910	15	16	16	17	18	19
Mozambique		508	Country	910	5 959	6 059	6 165	6 275	6 390	6 508
Réunion		638	Country	910	248	259	268	277	284	292
Rwanda		646	Country	910	2 186	2 251	2 314	2 380	2 450	2 527
Seychelles		690	Country	910	36	37	37	38	39	39
Somalia		706	Country	910	2 264	2 308	2 352	2 397	2 444	2 492
South Sudan		728	Country	910	2 482	2 502	2 525	2 553	2 585	2 620
Uganda		800	Country	910	5 158	5 308	5 453	5 596	5 741	5 889
United Republic of Tanzania	2	834	Country	910	7 650	7 845	8 051	8 267	8 494	8 730
Zambia		894	Country	910	2 310	2 369	2 431	2 498	2 570	2 645
Zimbabwe		716	Country	910	2 747	2 832	2 922	3 016	3 113	3 213
Middle Africa		911	Subregion	947	26 454	26 924	27 431	27 966	28 524	29 098
Angola		24	Country	911	4 548	4 617	4 713	4 823	4 936	5 043
Cameroon		120	Country	911	4 307	4 384	4 462	4 542	4 623	4 707

UN World Population Prospects 2019

Is this data visualization?

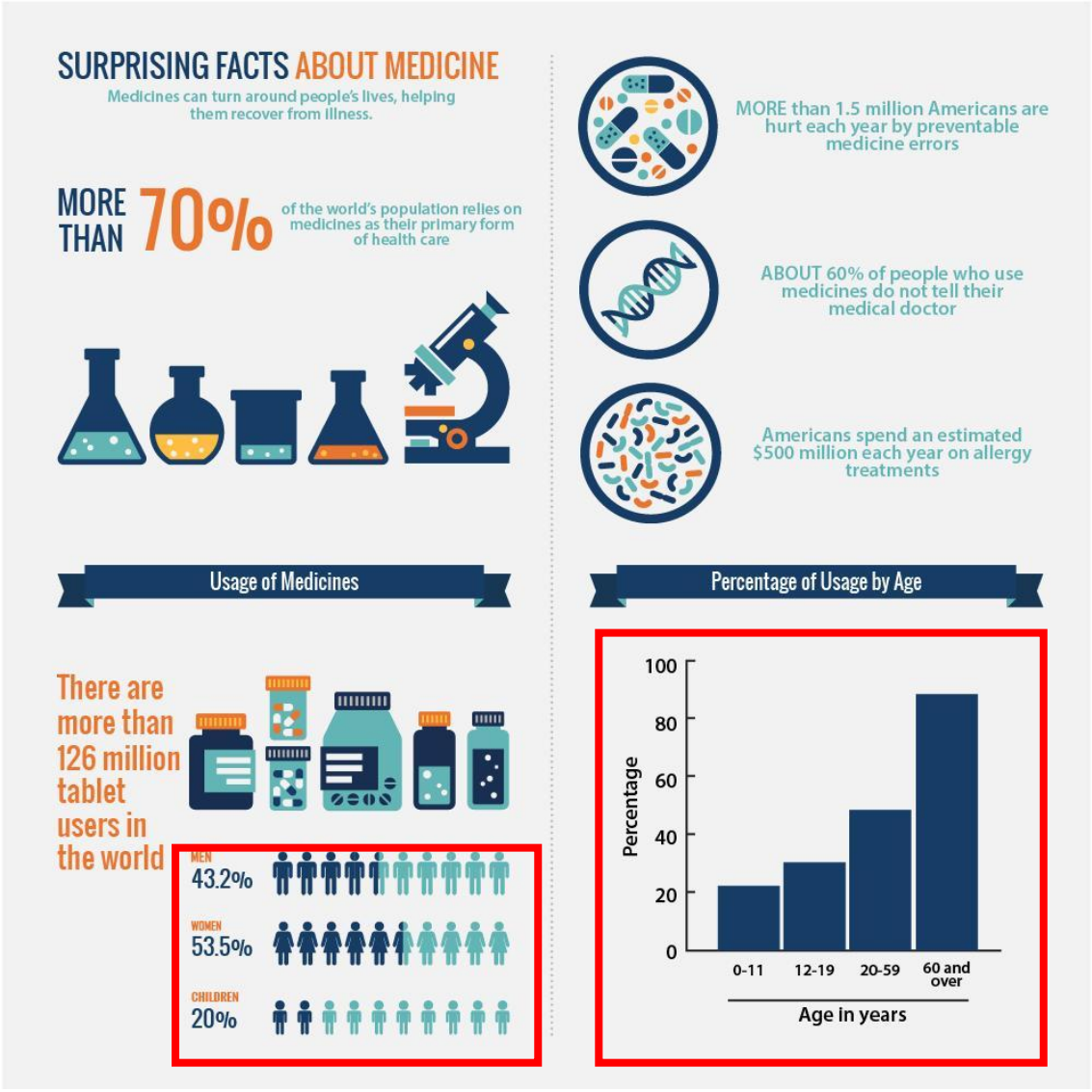
Is this data visualization?



Ong et al. 2018, Translational Psychiatry

Is this data visualization?

Tweet Action List, Twitter



Tarresa Muffet, https://visual.ly/

Is this data visualization?

Data visualization

is the graphic representation of data through pictorial design.

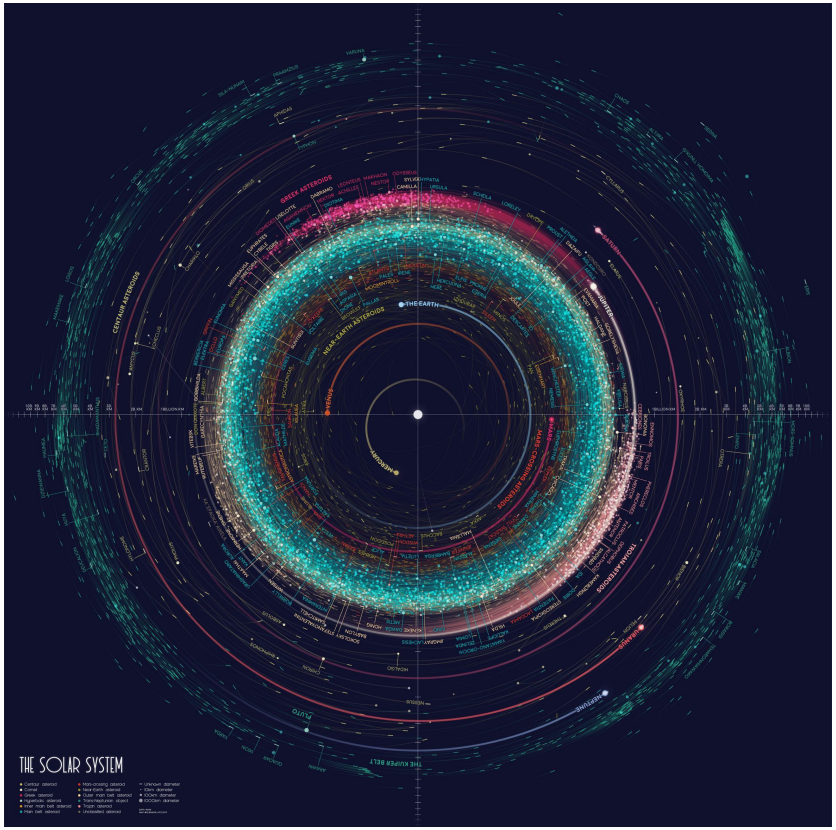
PART 1

Principles of good graphics

Why do we use **data visualization (DV)**?

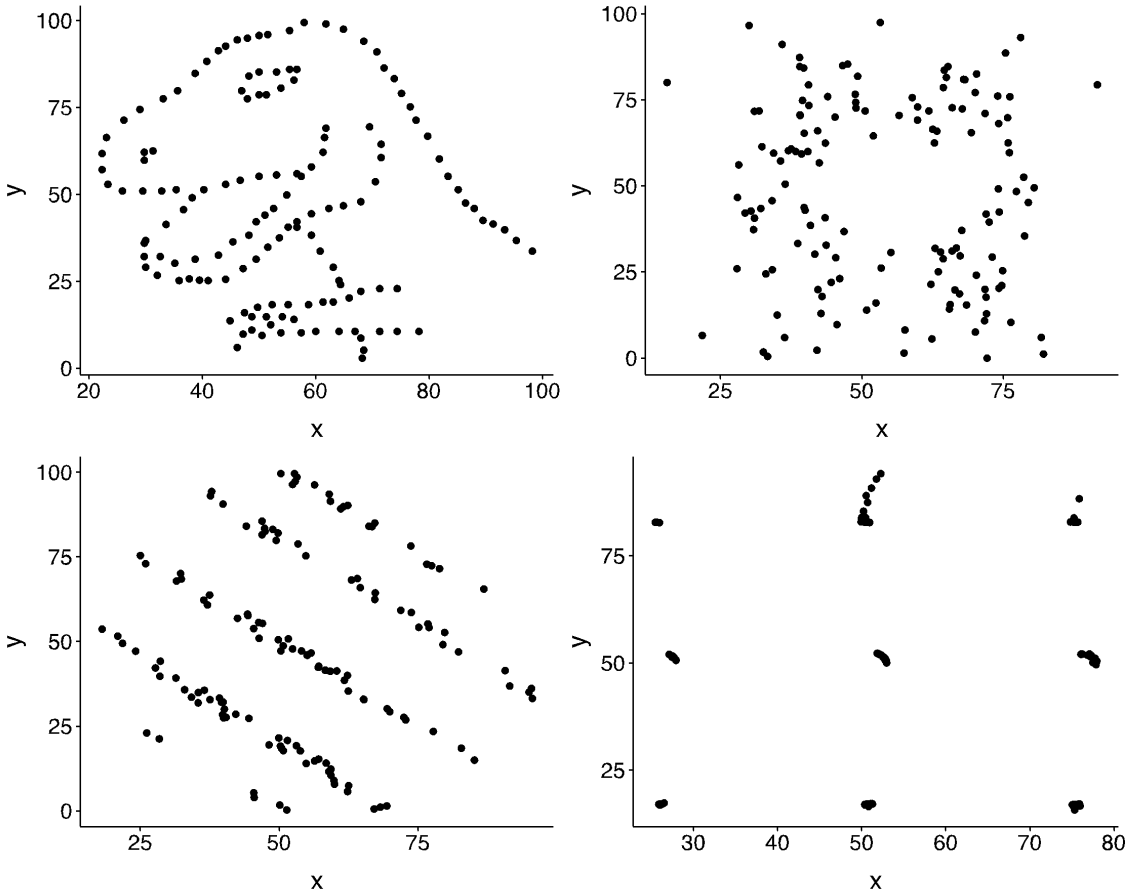
Why visualize data?

- Comprehend information quickly
- Identify relationships and patterns
- Communicate information to others
- Visual appeal



Map of the Solar System, Eleanor Lutz

Same statistics, different plots



X mean: 54.26

Y mean: 47.83

X SD: 16.76

Y SD: 26.93

Corr.: -0.06

Data source: *Matejka et al 2017, InProceedings of the 2017 CHI Conference on Human Factors in Computing Systems*

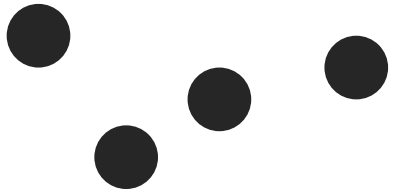
Hue

Size

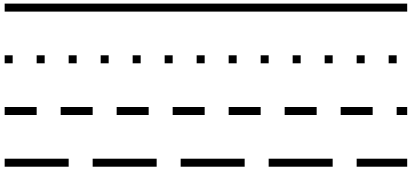
Value

Shape

Position



Pattern



What makes for effective
data visualization (DV)?

10 Commandments of Data Visualization

1. Thou shalt identify thy message and tell a clear story.
2. Know thy audience.
3. Thou shalt consider thy medium.
4. Thou shalt keep it simple.
5. Thou shalt reduce chart junk.
6. Thou shalt not mislead the reader.
7. Thou shalt not abuse colours.
8. Thou shalt use chart types that best fit thy message.
9. Thou shalt avoid 3D plots in static media.*
10. Thou shalt not trust default formatting.

**Exception made for brain images*

Let your message guide the figure design.

- What is the underlying message? How can a figure best express this message?
- Figure is meant to express an idea or introduce some facts that would be too long to explain with words
- Every figure should have a clear role
- Tighten the bond between the message and the visual

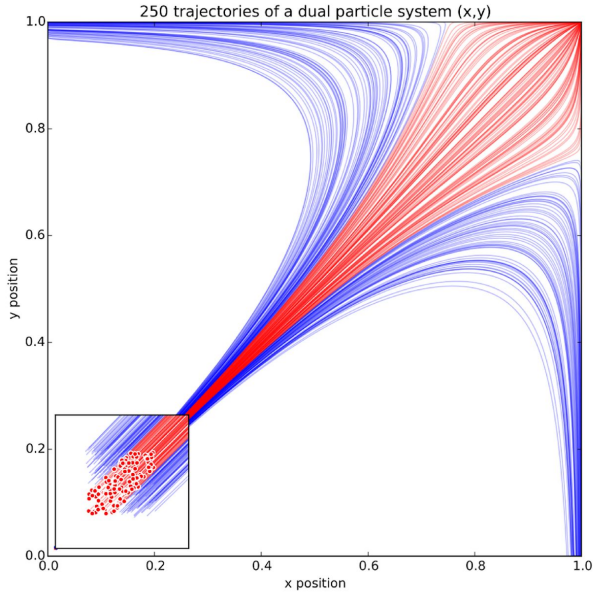
Design with your audience in mind.

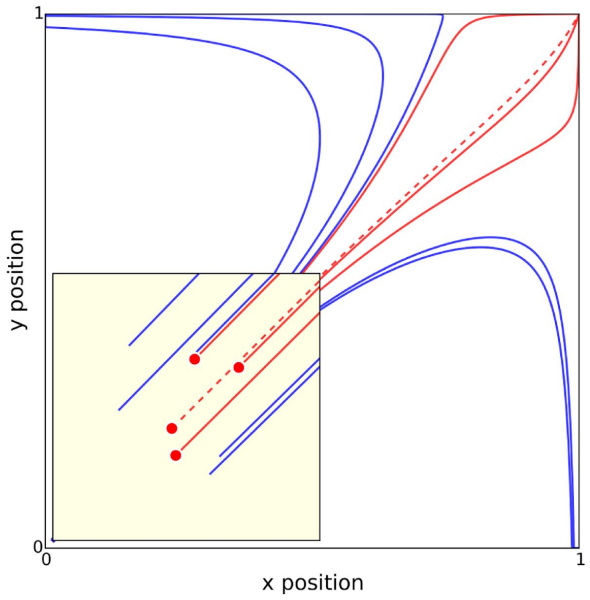
- Graphical visual should be designed with the audience in mind
- Is the figure for:
  - You alone?
  - Your collaborators?
  - A scientific audience or publication?
  - The general public?

Design with your medium in mind.

- Where will your visualization be used?
- How large will it be?
- How long will the audience interact with it?
  - E.g. Poster, computer monitor, projector, online, printed article

Which would be more appropriate for a scientific poster?

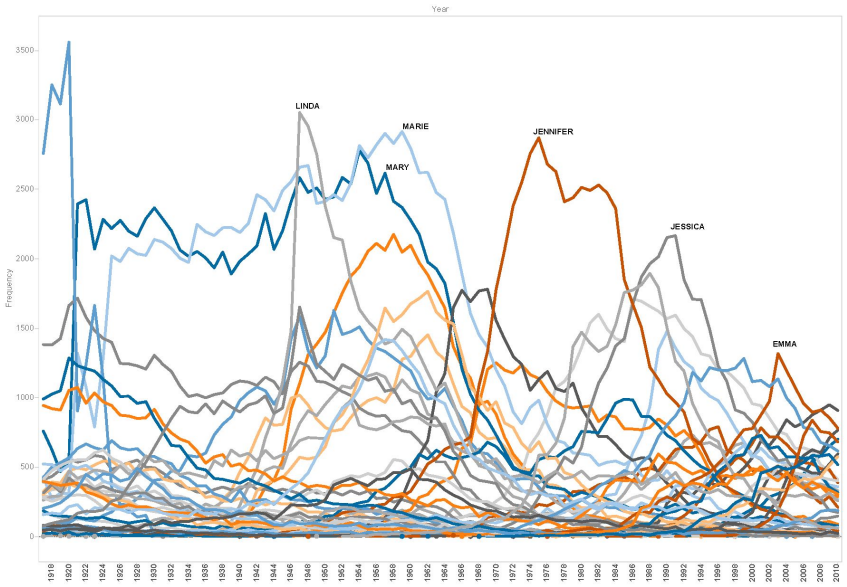




Rougier et al 2014, PLOS Computational Biology

Focus on patterns in your data.

- Don’t plot everything you have – focus on what is interesting!
- Reduce data complexity
  - Don’t add too much information to a single chart
  - If necessary, split data into multiple charts, simplify colours, or change the chart type

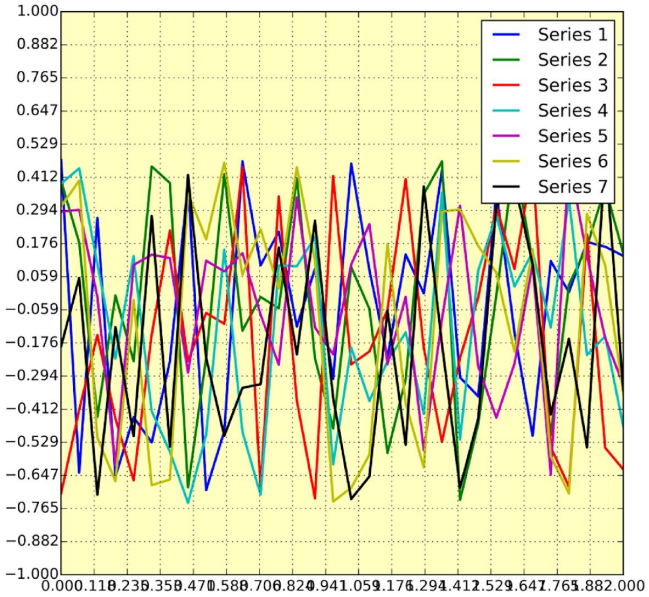


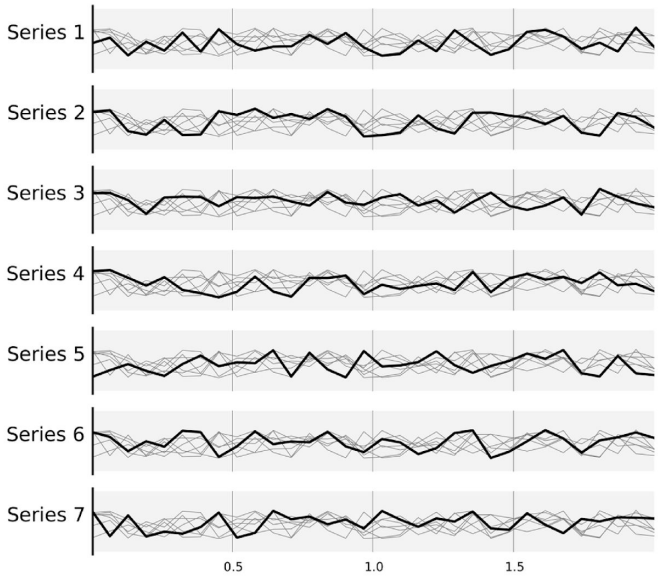
Haiku Analytics

All graphic elements should have meaning.

- Chart junk: all unnecessary or confusing visual elements in a figure that do not improve the message, or adds confusion
- Improve the data-ink ratio: remove non-data ink everywhere possible
- Message > Beauty (especially in science)

Which is easier to read?





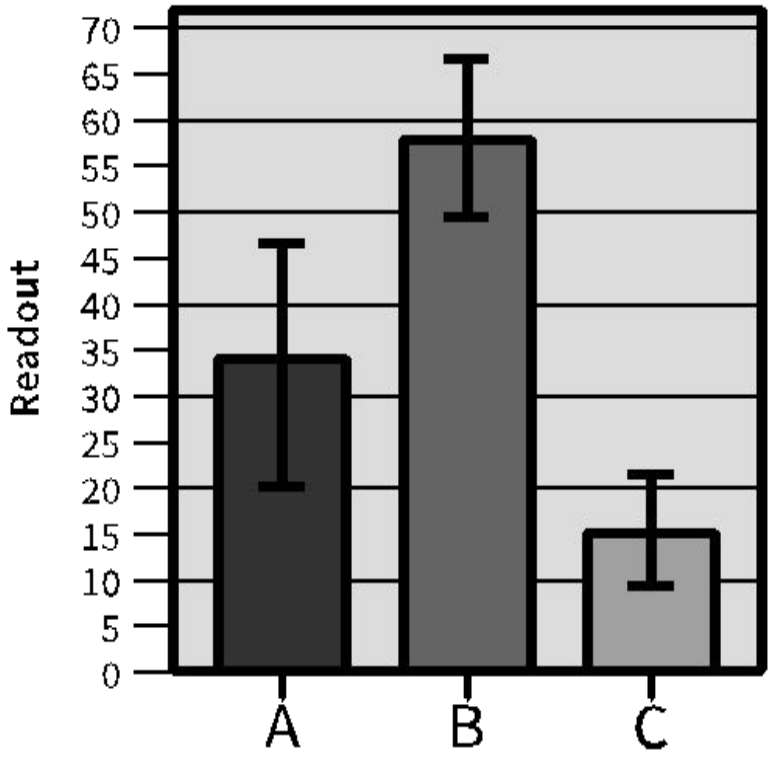
Rougier et al 2014, PLOS Computational Biology

Mini-Challenge

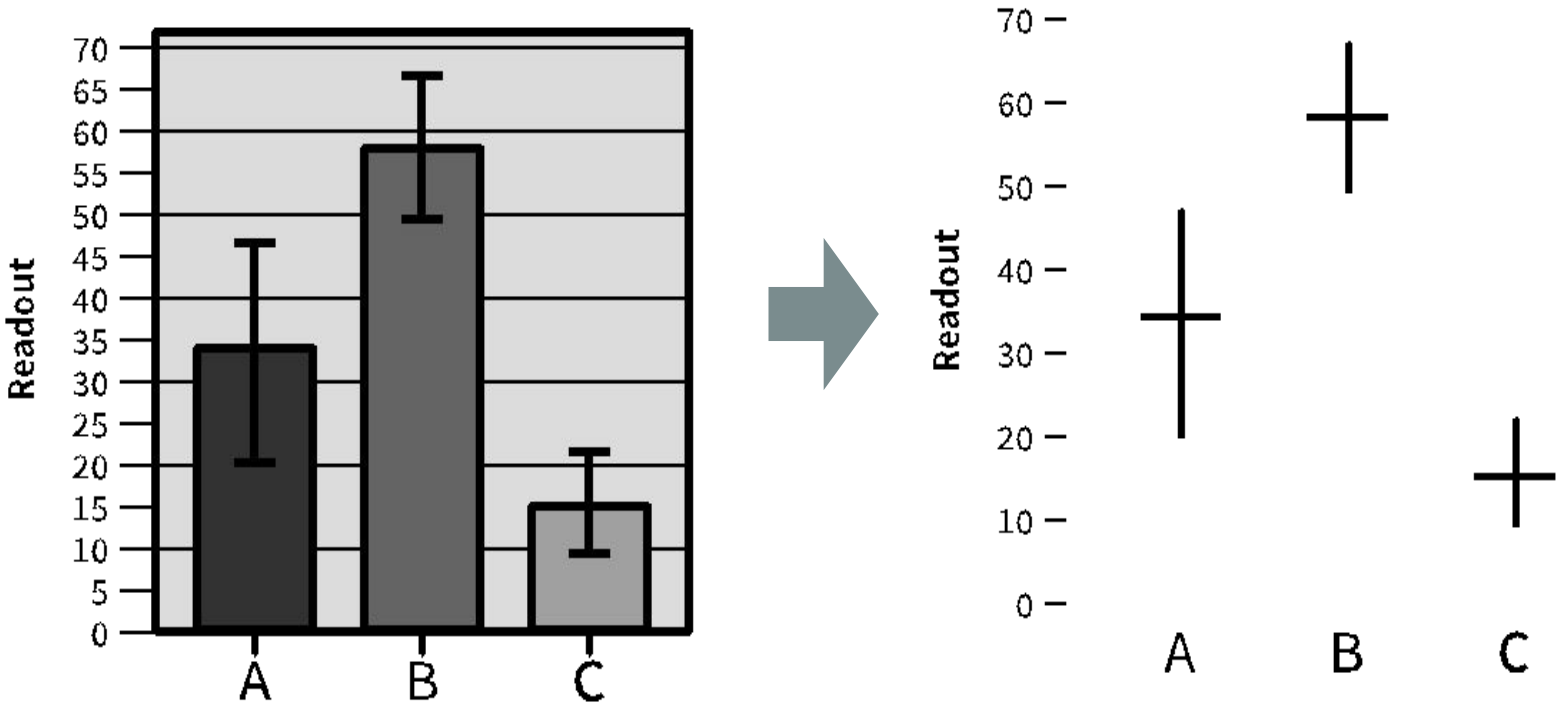
Individuals, 2 minutes

How much can you remove from the following column chart without losing information?

Sketch your solution using as little ink as possible!



Can you still interpret this plot?

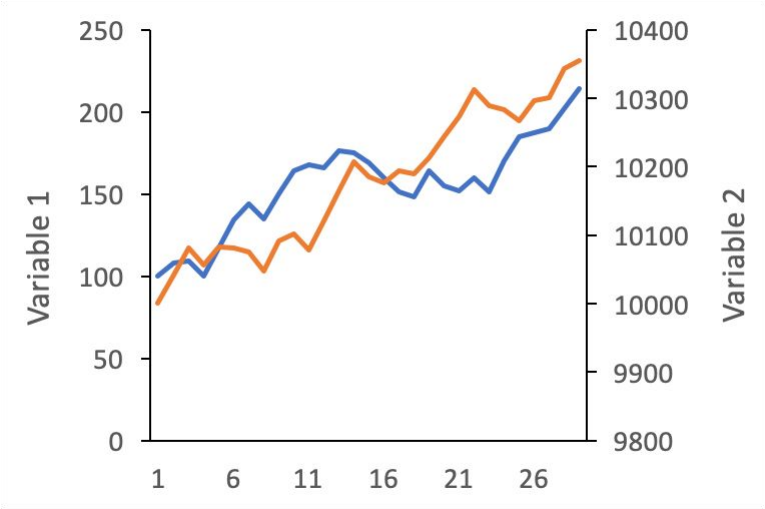


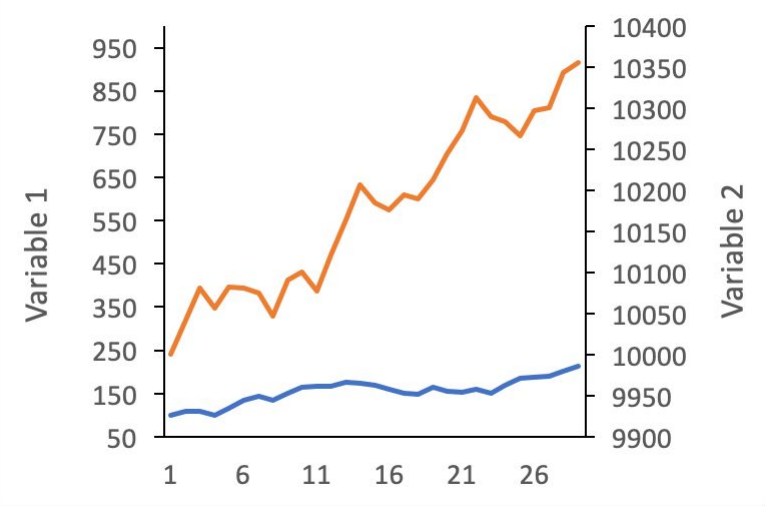
Though super minimalistic is not always the way to go, carefully consider whether each addition facilitates communication!

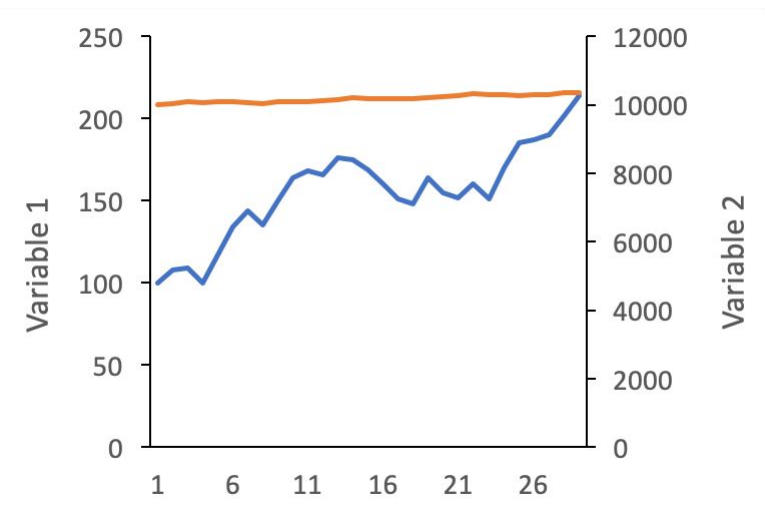
Represent your data accurately.

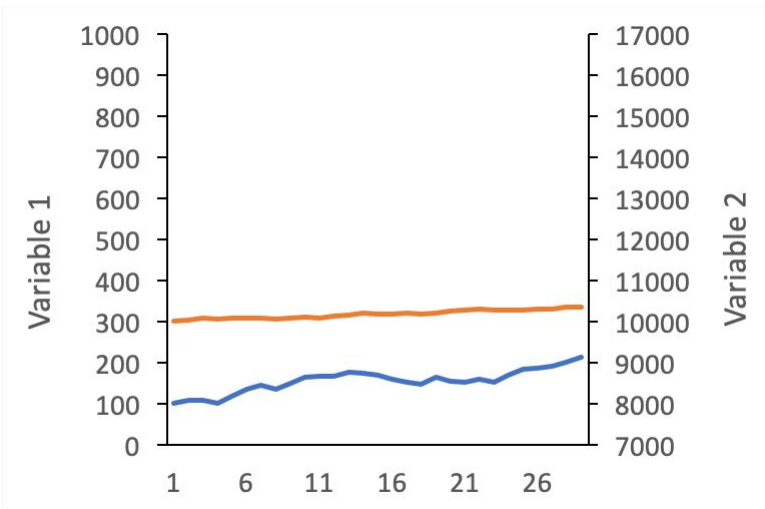
- **Proportional Values:** Numbers in a chart (bar, area, bubble etc.) should be directly proportional to the numerical quantities
- Do not oversimplify data– provide enough context
- Don’t deliberately truncate y-axis
- Pie charts must add up to 100%
- Avoid dual-axes line charts

Dual axis chart - Same data, different scales



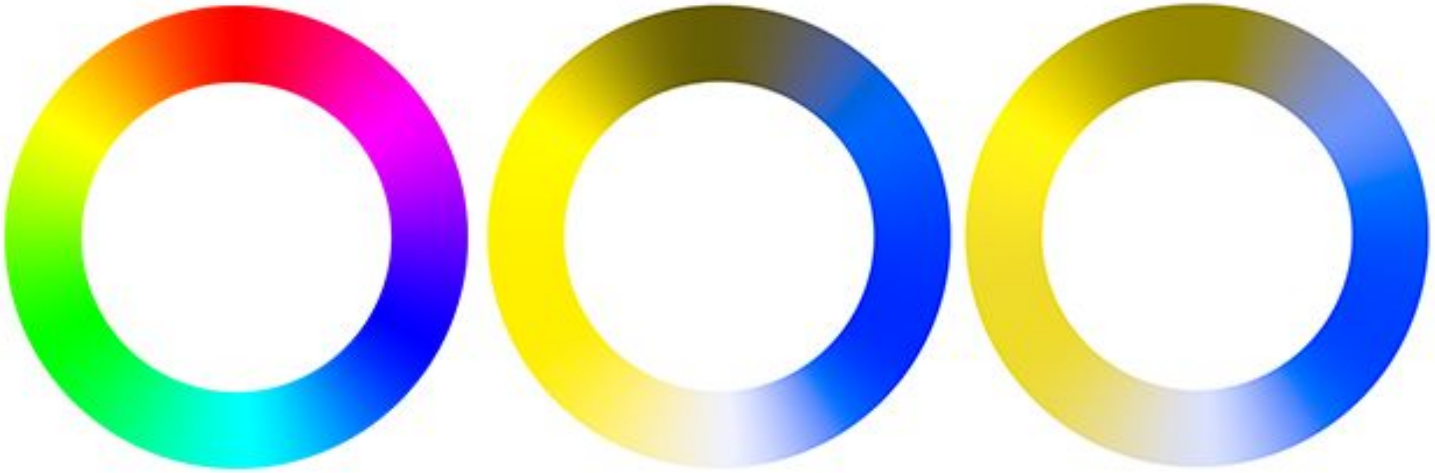






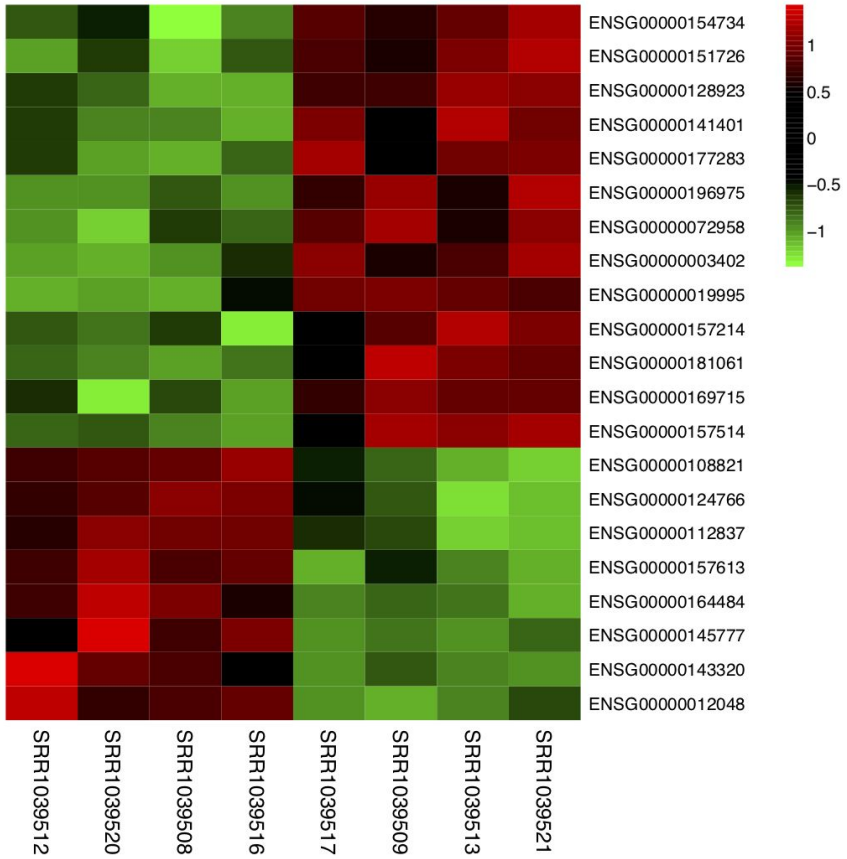
Colours should have purpose.

- **Limit your colour palette:** In ANY chart, do not use more than 6 colours (exceptions for gradients)
- Keep the consistent colour palette or style for all charts in the series
- 5-10% of men are colour-blind
- Colour:
  - Darker/brighter colours will read as “more”
  - Use gray to denote neutral or unimportant material
  - Use white space
  - Every color used must have a reason, and should be distinct
  - Choose appropriate colour palette and keep them consistent

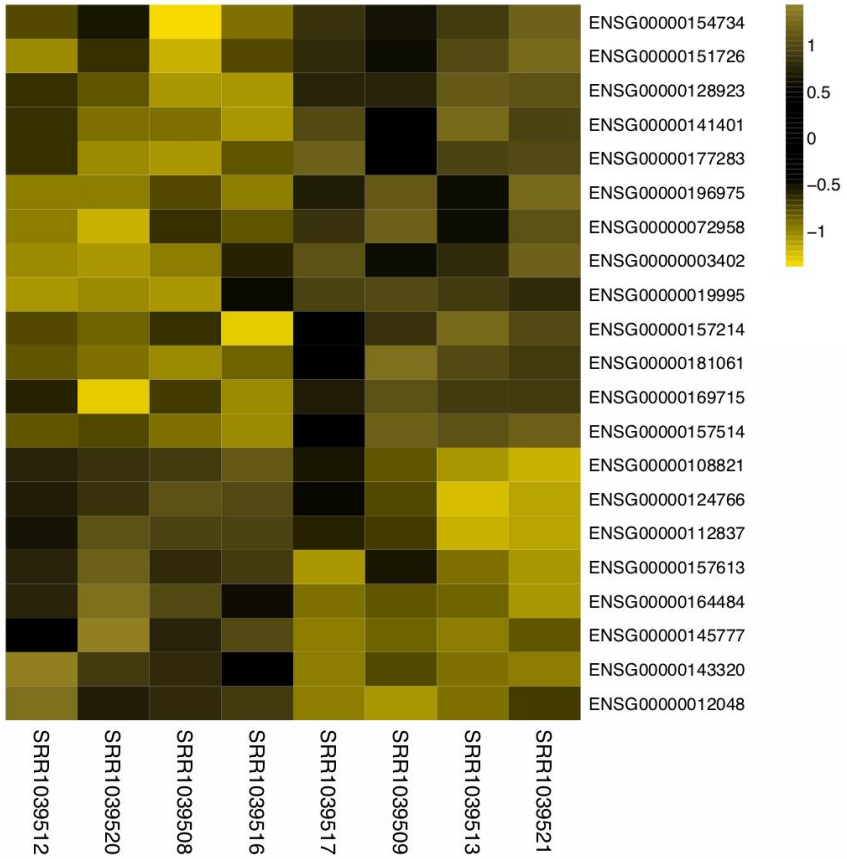

Color Vision
Portanopia
Deuteranopia

Is this a good colour palette selection?

Original

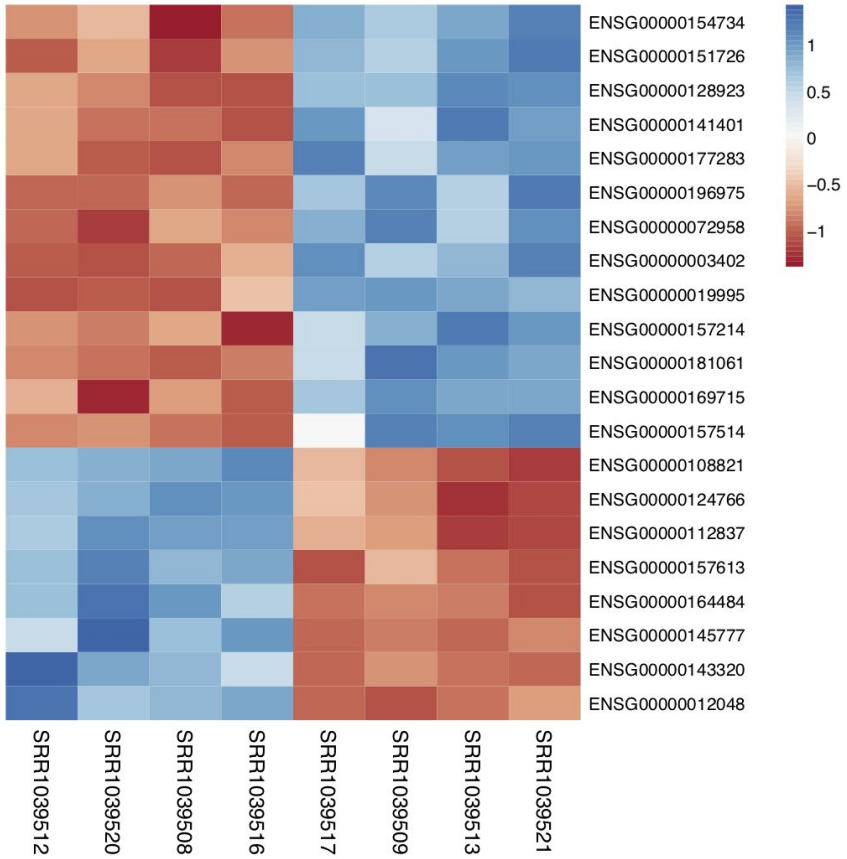


Protanopia

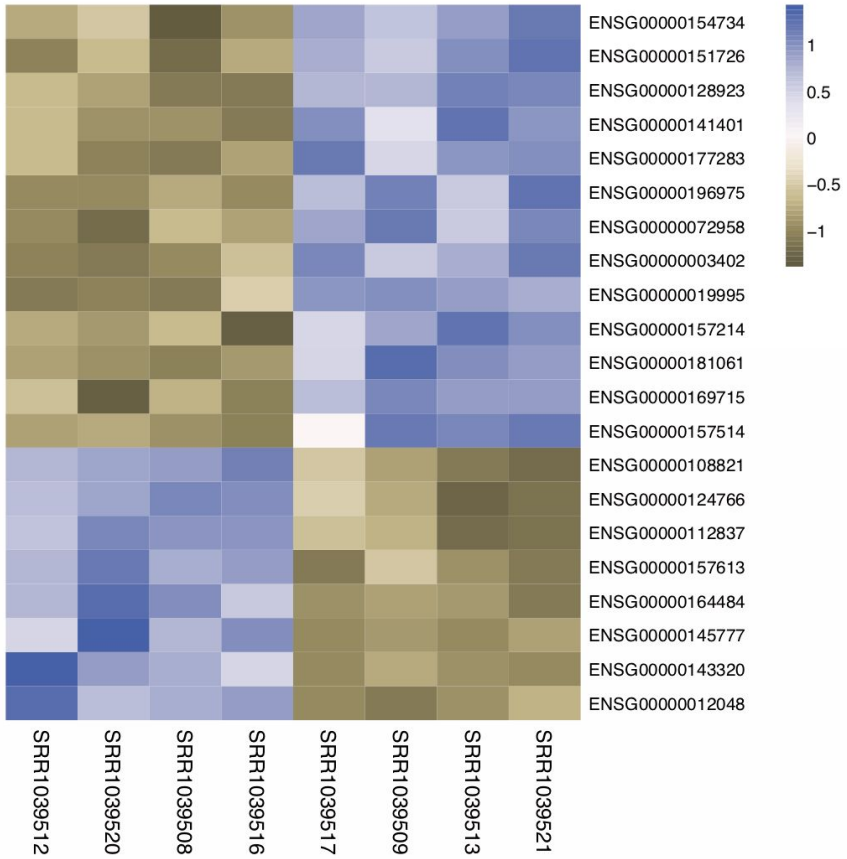


Is this a good colour palette selection?

Original



Protanopia

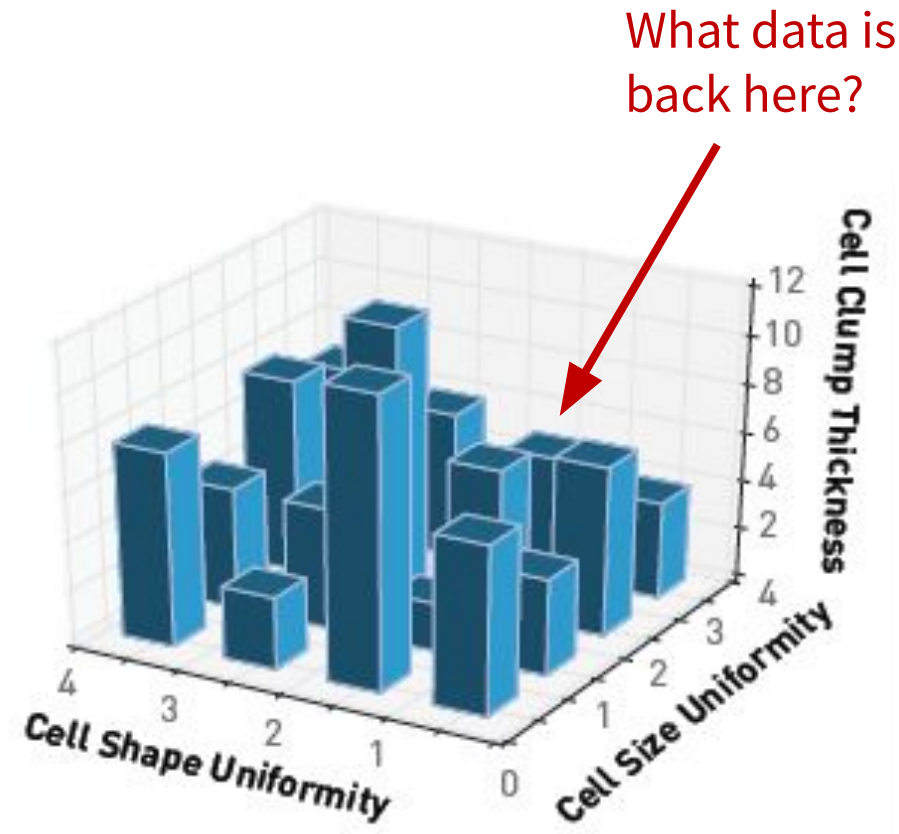


Graphic type influences your message.

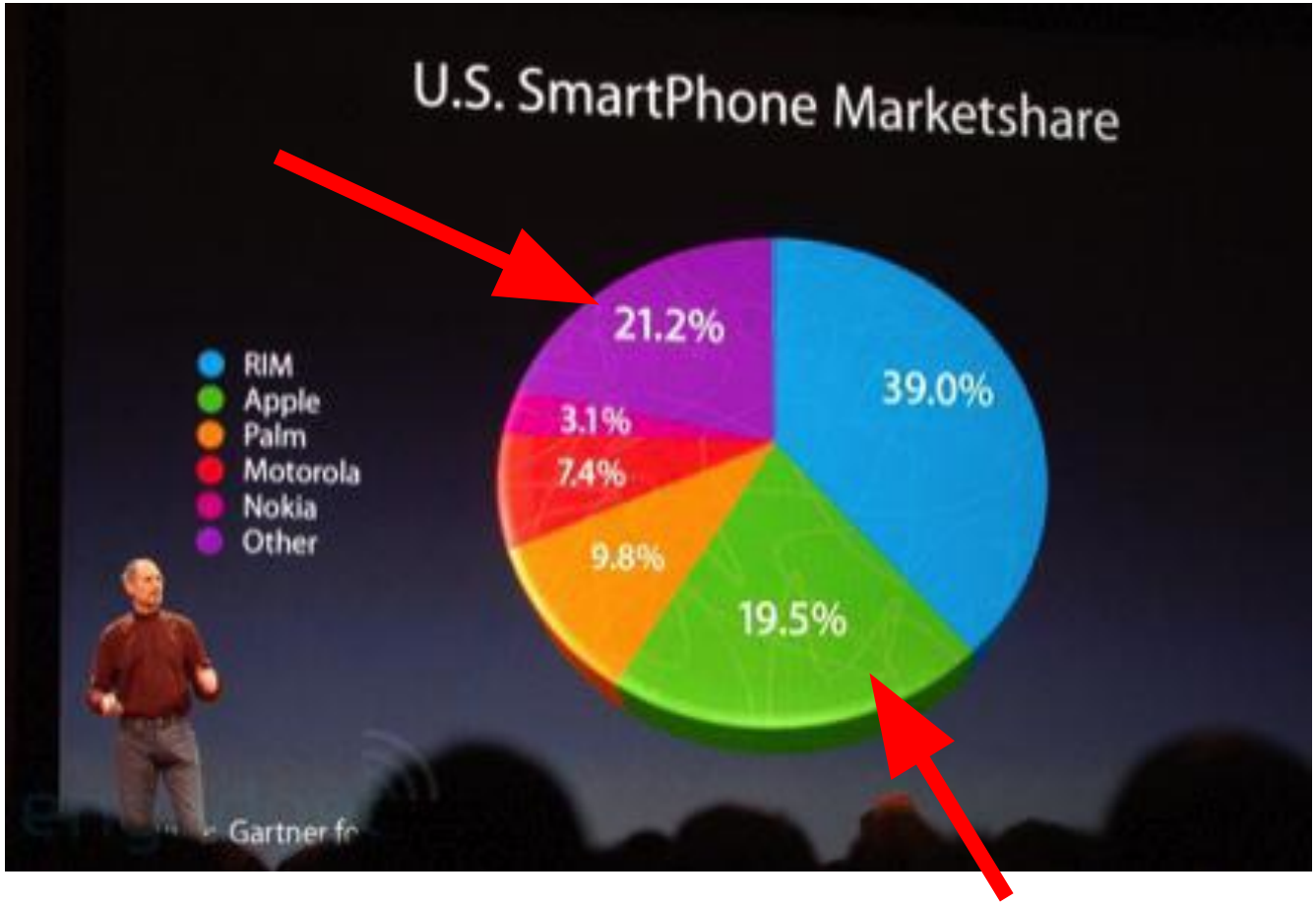
- Each type of plot communicates or emphasizes an aspect of your data
- Also consider how transforming your data effects your message.
- Example: Consider world population data, divided by content.
  - Pie Chart for 2018: “Majority of the people live in Asia”
  - Bar Chart for 2018: “Asia has the most people.”
  - Line Chart for 2002-2018: “The population is increasing in all continents”
  - Line Chart for 2002-2018 (growth rate): “The fastest growth rate is in Africa”

3D plots are for interactive media only.**

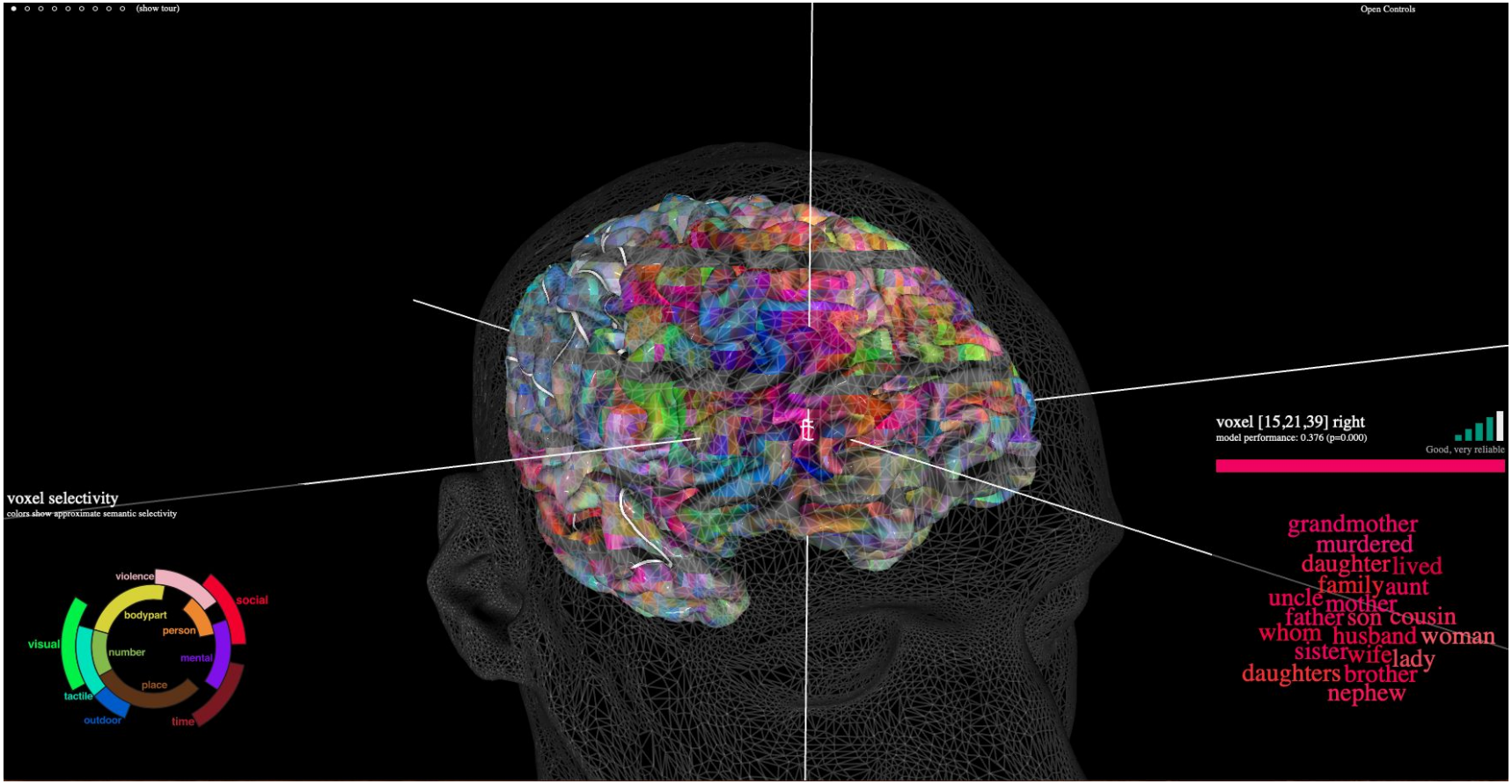
- 3D plots are unjustified for data with two axes!
- Obstruction of data in static plots with 3 axes



What's wrong with this chart?



A good use of 3D data visualization



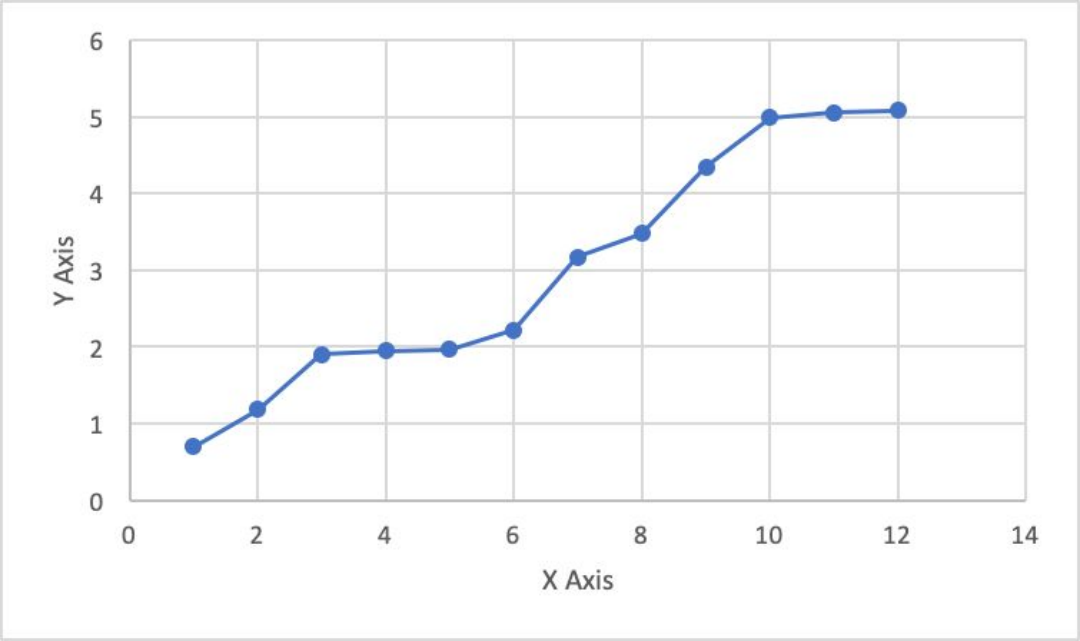
Browser-based, interactive semantic brain visualization tool

Gallant Lab @ UC Berkeley. Website: https://gallantlab.org/huth2016/

Visualizations require customization.

- Any plotting library/software has default settings
  - *Size, font, colours, styles, ticks, markers, axis-scales*
- Settings are good for any plot, best for none
- Almost all plots require at least some manual tuning
  - *E.g. Colour map selection for data, adding annotations*

In what software was this plot made?



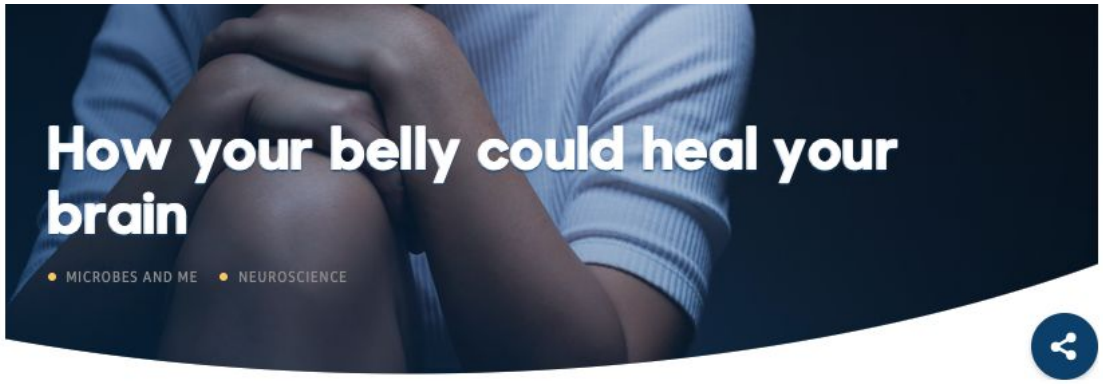
Original Article | Published: 12 April 2016

Gut microbiome remodeling induces depressive-like behaviors through a pathway mediated by the host's metabolism

P Zheng, B Zeng, C Zhou, M Liu, Z Fang, X Xu, L Zeng, J Chen, S Fan, X Du, X Zhang, D Yang, Y Yang, H Meng, W Li, N D Melgiri, J Licinio ✉, H Wei ✉ & P Xie ✉

Molecular Psychiatry **21**, 786–796(2016) | Cite this article

Scientific publication featured in a recent BBC article



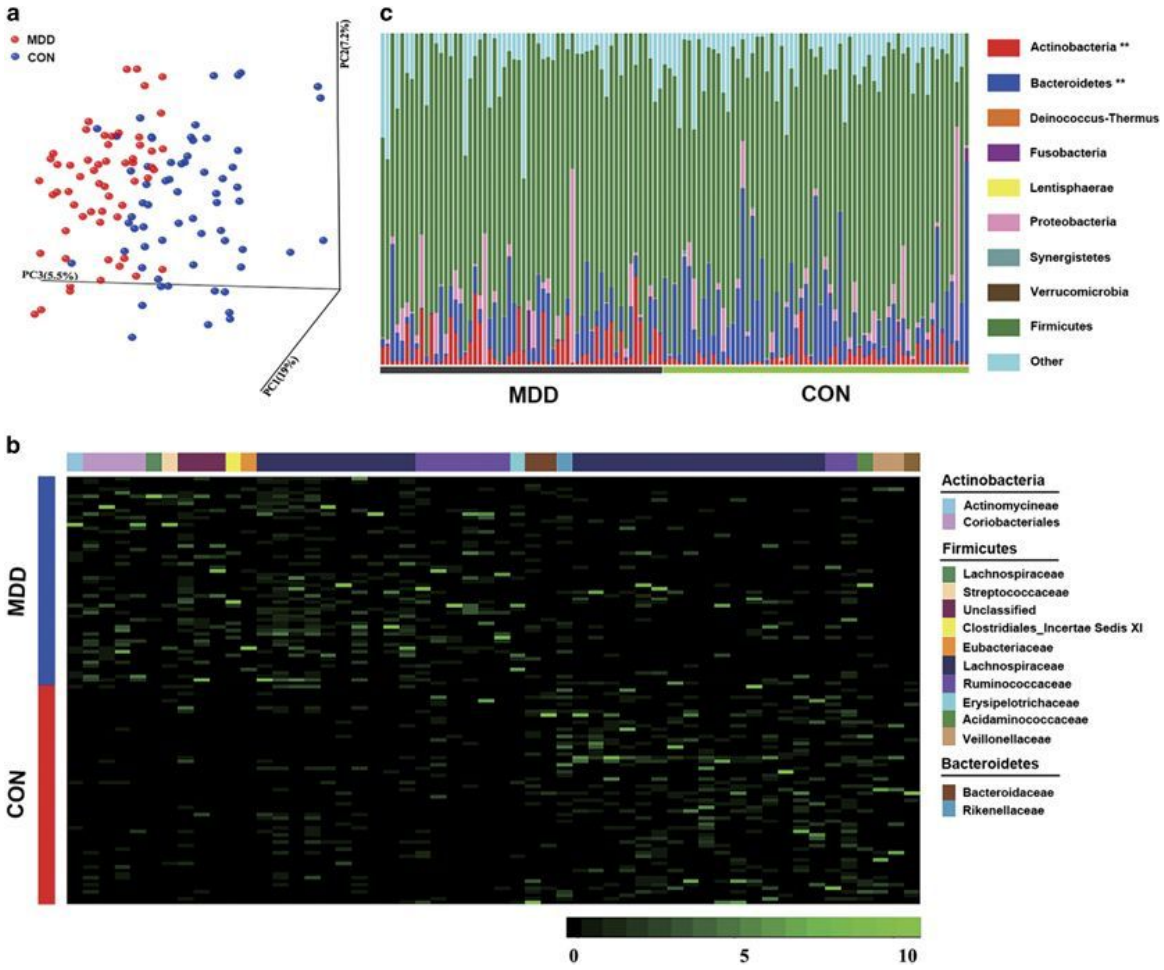
BBC FUTURE

By David Robson
21st February 2019

A century ago, a few isolated studies found a link between diet and mental health. Now, it's emerging that the bacteria inside us could be a crucial link between the food we eat and how we feel.

From the original manuscript

Figure 2. 16S rRNA gene sequencing reveals changes to microbial diversity in major depressive disorder (MDD).



16S sequencing of fecal microbiome from controls and patients with depression.

a) PCoA plot showing (dis)similarities in gut composition

b) Heatmap of 54 discriminative OTU abundances

c) Relative abundances of phyla between groups

Exercise

Teams of ~10 people, 15 minutes

As a group, your challenge is to improve this figure.

Collect your ideas on sticky notes, then sketch your new figure on a new sheet of chart paper.

- What is the **key message**, in plain language?
- **What is wrong** what these figures?
- Change(s) and improvement(s) to figure? (sketch)
- Challenge: What figure could you use to communicate the overall message to a general (non-scientific) audience? (sketch)

Sample chart paper layout

Key Message: Your simple sentence here.

a.

YOUR SKETCH

b.

YOUR SKETCH

c.

YOUR SKETCH

Problems and improvements

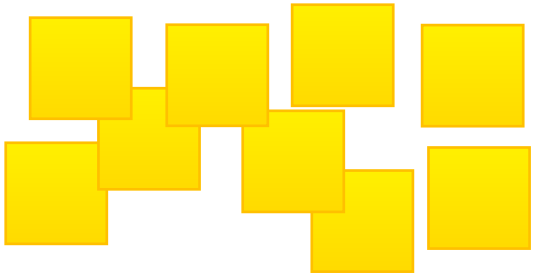
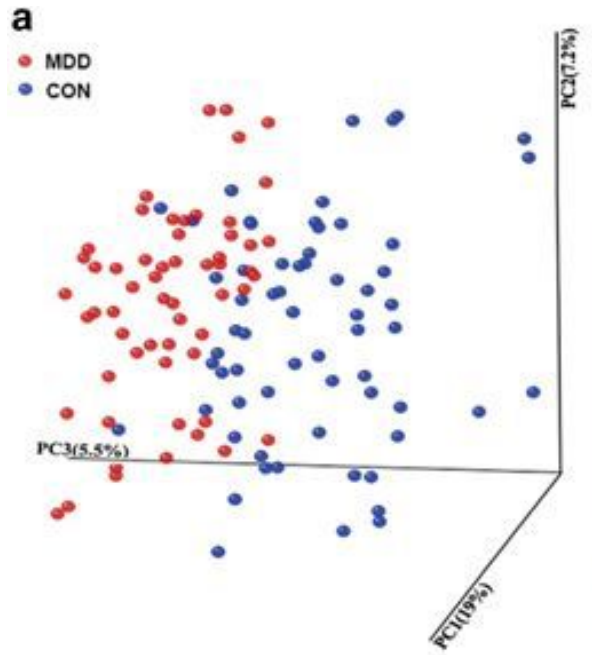


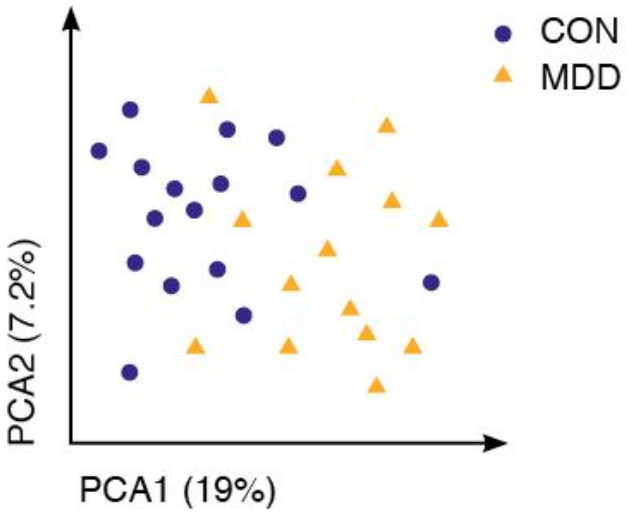
Figure for general audience:

YOUR SKETCH

Example figure improvements

Key Message: People with depression have differences in gut microbiome composition.





Problems

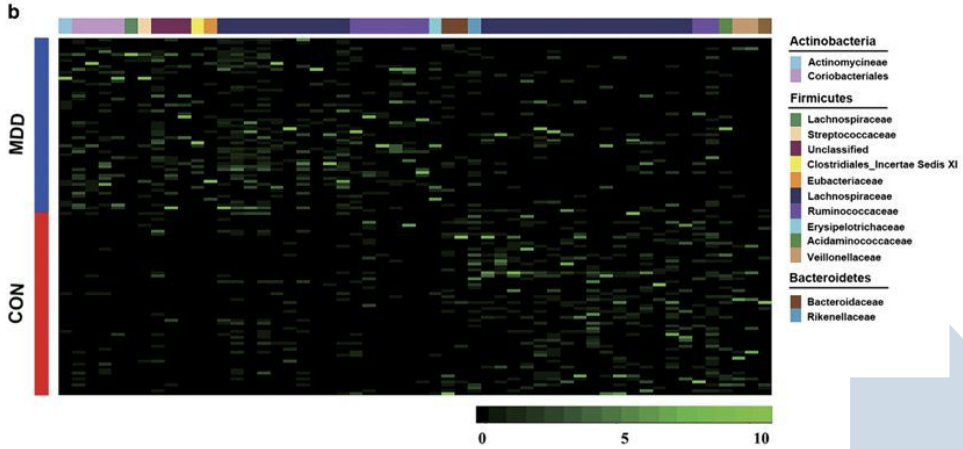
- Data on 3D axes difficult to read

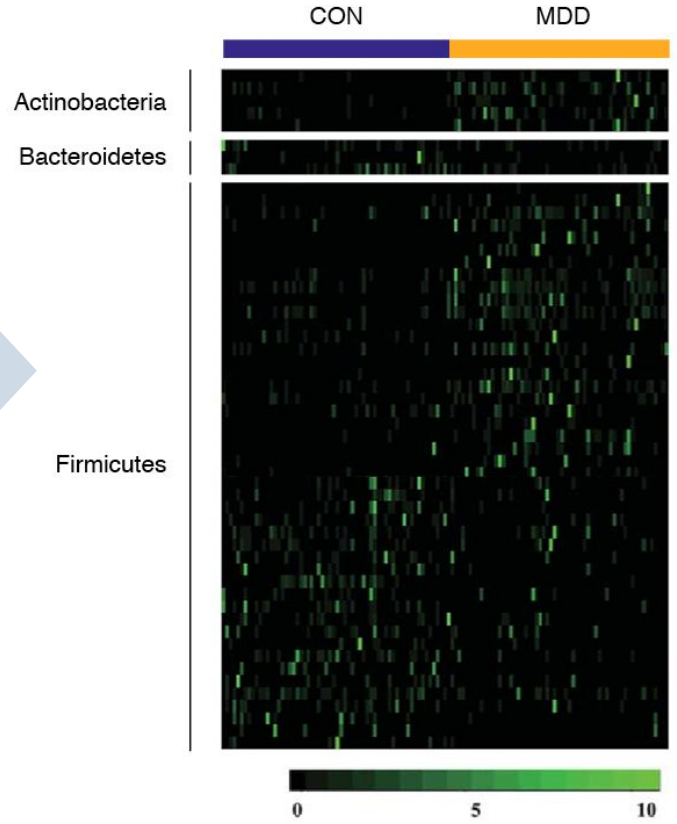
Suggested Improvements

- Restrict to 2 axes
- Use shape, colour, and shade to distinguish points (works in greyscale)

Example figure improvements

Key Message: People with depression have differences in gut microbiome composition.





Problems

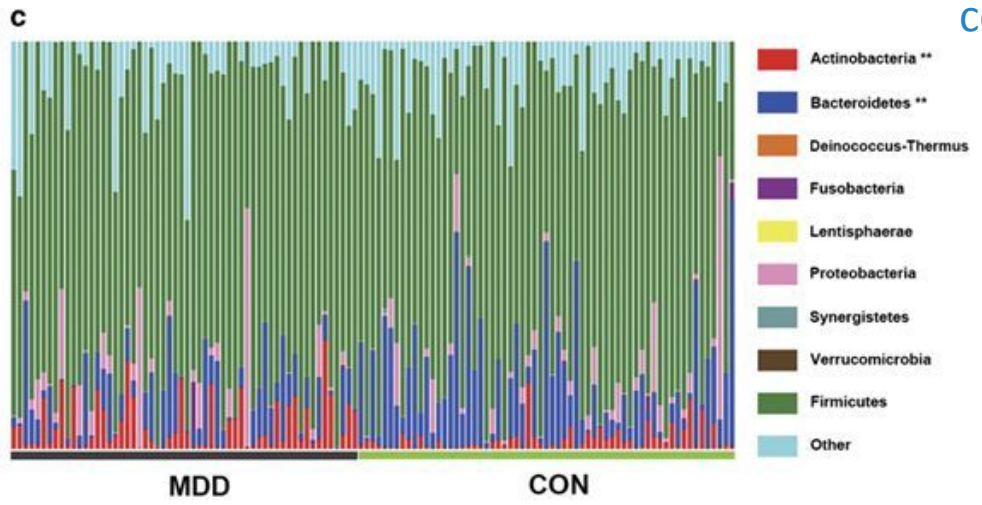
- Can't easily discriminate between taxonomic assignments
- Key message is obscured - what is the difference in message to Fig. C?

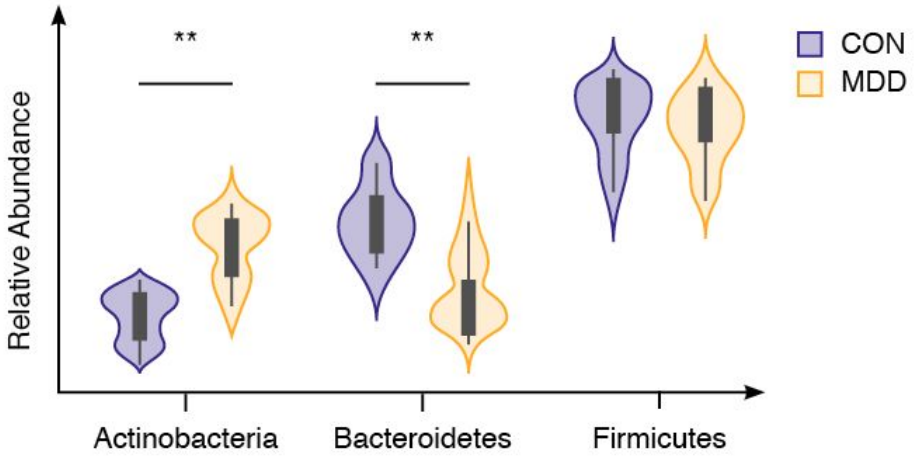
Suggested Improvements

- Remove detailed labeling of OTUs
- Rearrange heatmap to emphasize figure message (change in Firmicutes is hallmark of MDD)

Example figure improvements

Key Message: People with depression have differences in gut microbiome composition.





Problems

- Too many stacked bars - cannot easily distinguish phylla
- Missing y-axis label
- Inconsistent colours for CON/MDD labels

Suggested Improvements

- Change plot type to focus on key message
- Consistent colouring for CON/MDD labels

PART 2

Types of visualization graphics

Visualization type should suit the data and message

- What are you representing?
  - Relationships, distributions, proportions, changes over time
- What are your variable types?
  - Discrete? Continuous? Categorical? Spatial? Structural?

Great reference catalogues of data visualization types:

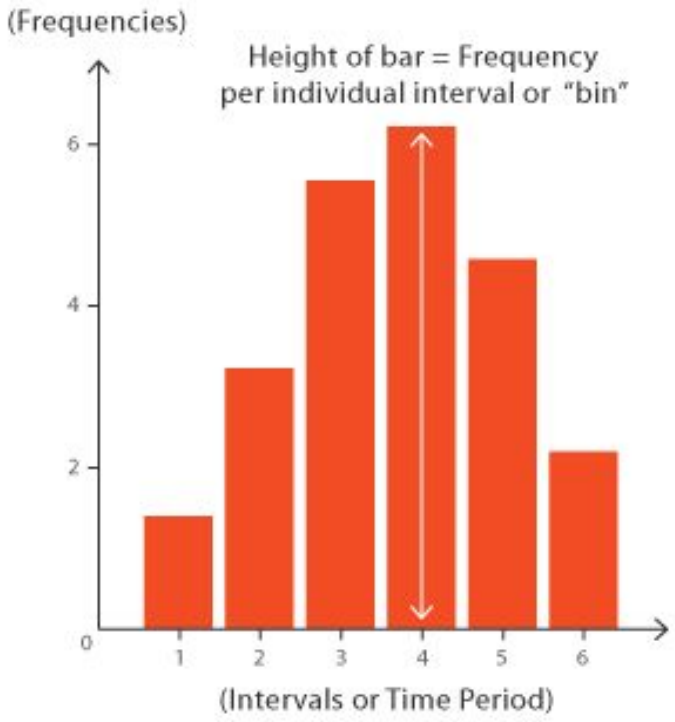
https://datavizcatalogue.com

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

Distributions

Histogram

One continuous variable



Visualizes data over a continuous interval or time period

Pros:

- Good at giving rough view of probability distribution

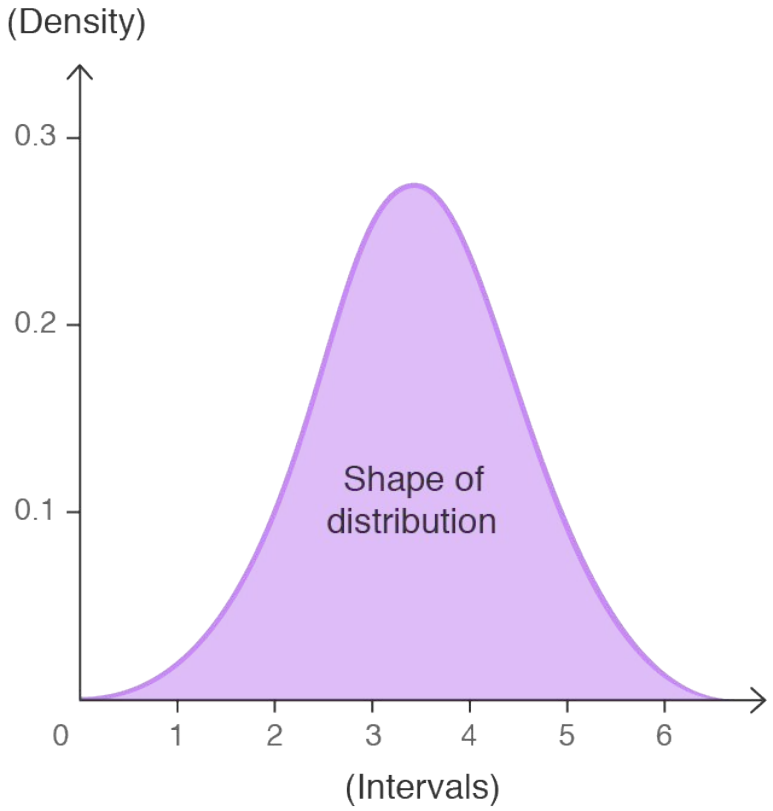
Cons:

- Appearance sensitive to selected bin width

Distributions

Density Plot

One continuous variable



Visualizes data over a continuous interval or time period

Pros:

- Better at determining distribution shape (not affected by bins)

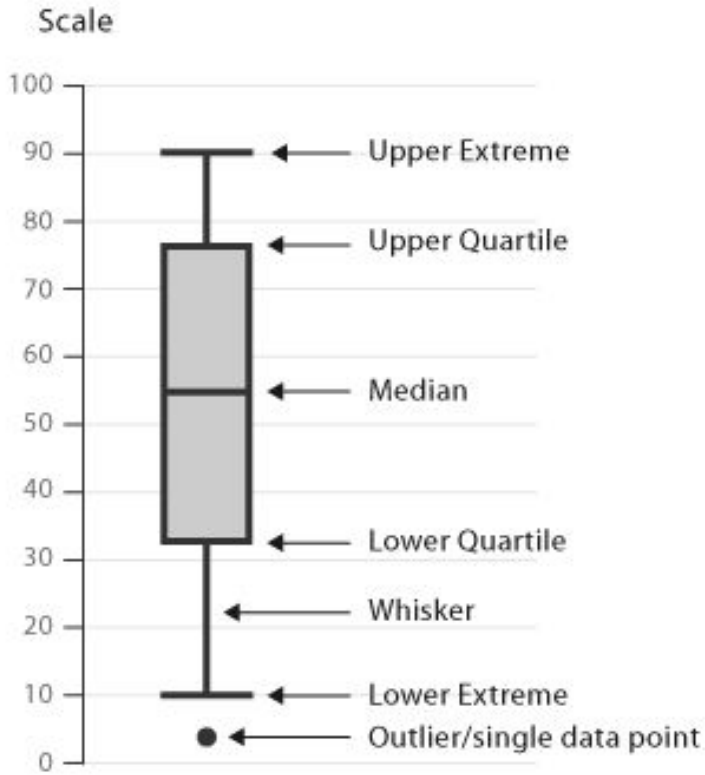
Cons:

- Can only show a very limited number of distributions one one plot

Distributions

Box & Whisker Plot

One continuous variable
0-2 categorical variables



Visualizes distribution of data through their quartiles.

Pros:

- Take up less space than histograms or density plots (can compare many groups)
- Shows skew in data

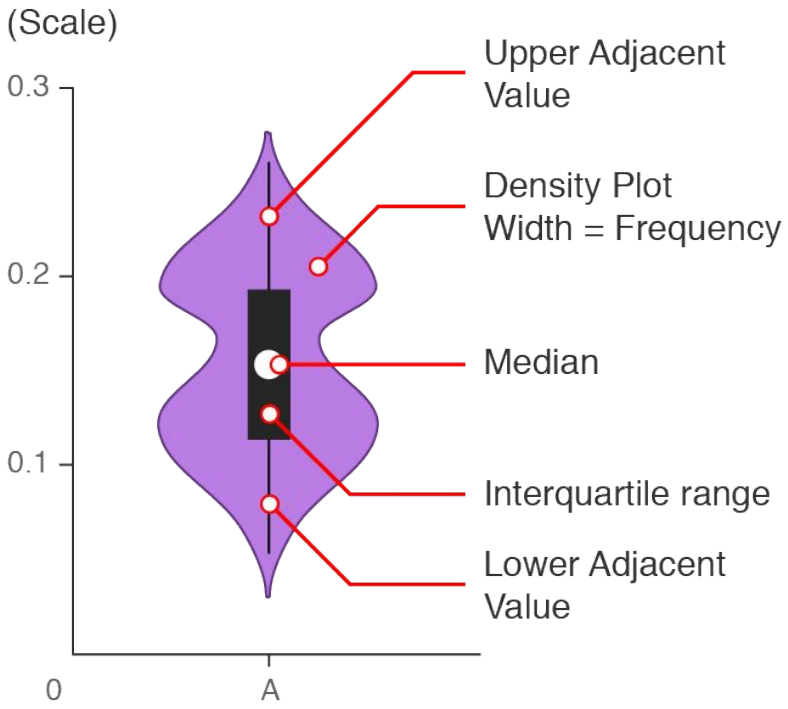
Cons:

- Simplicity can hide details in distribution of data

Distributions

Violin Plot

One continuous variable
0-2 categorical variables



Visualizes distribution of data through probability density; combines box-plot and density plot

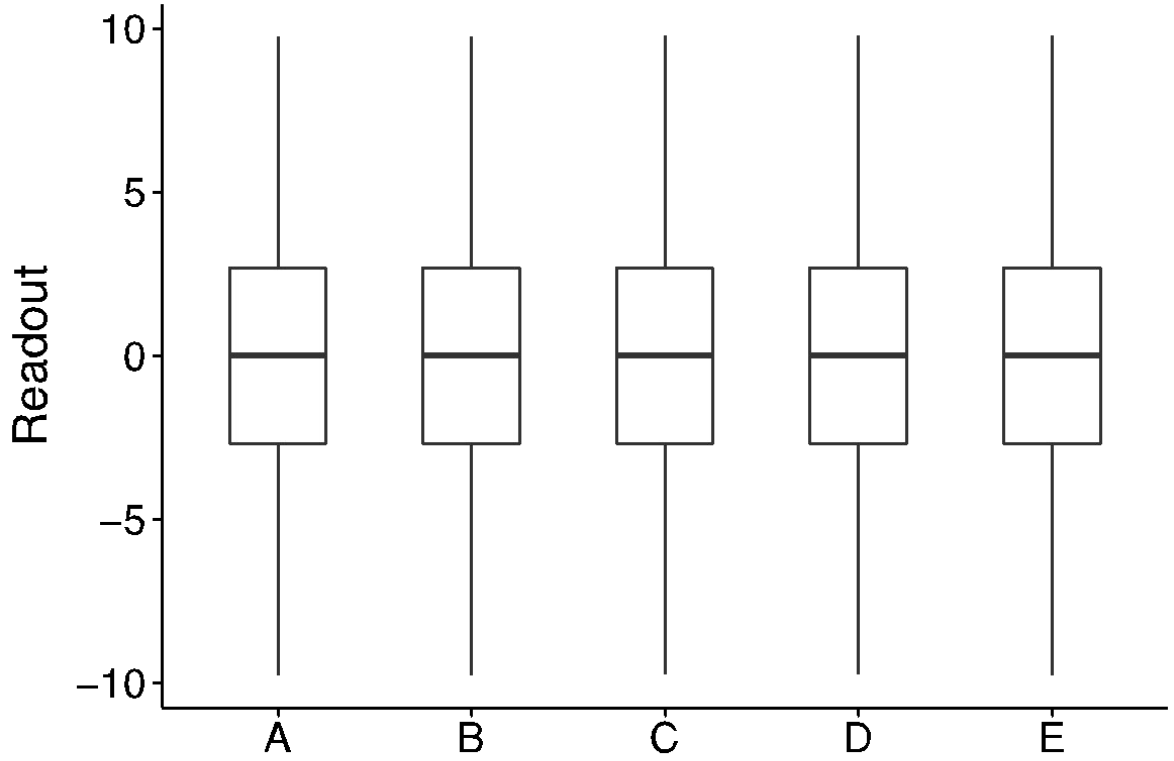
Pros:

- Can show much more information about distribution than box plots

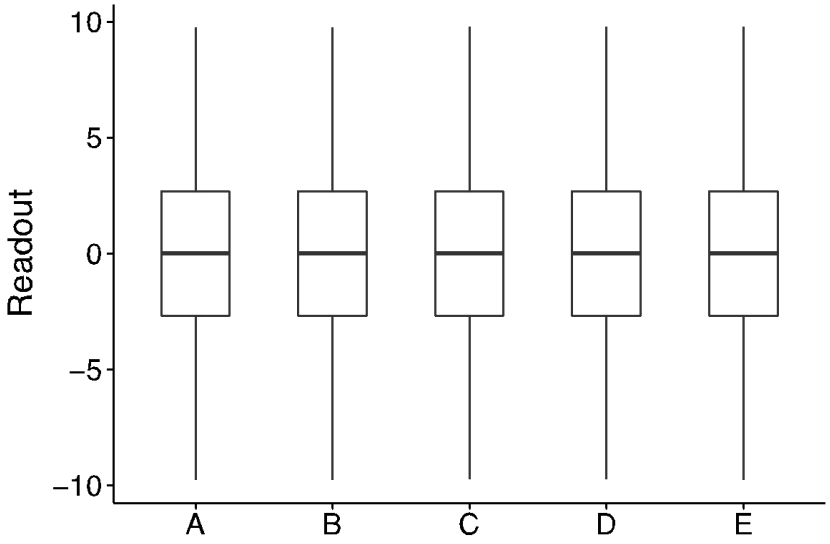
Cons:

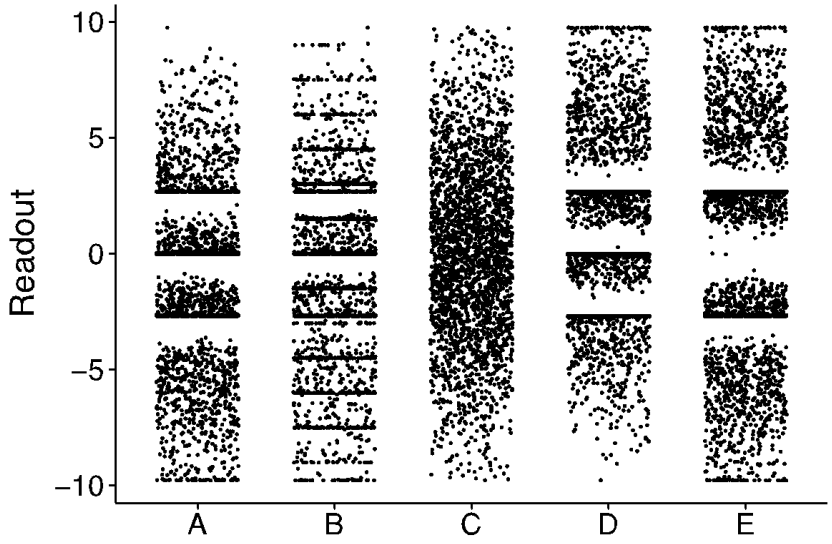
- More information → higher risk of noise

Are groups A-E the same?

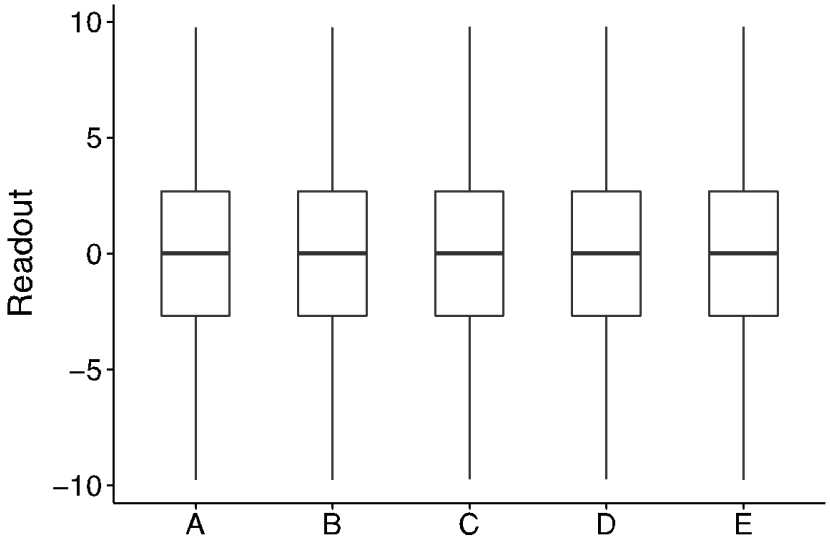


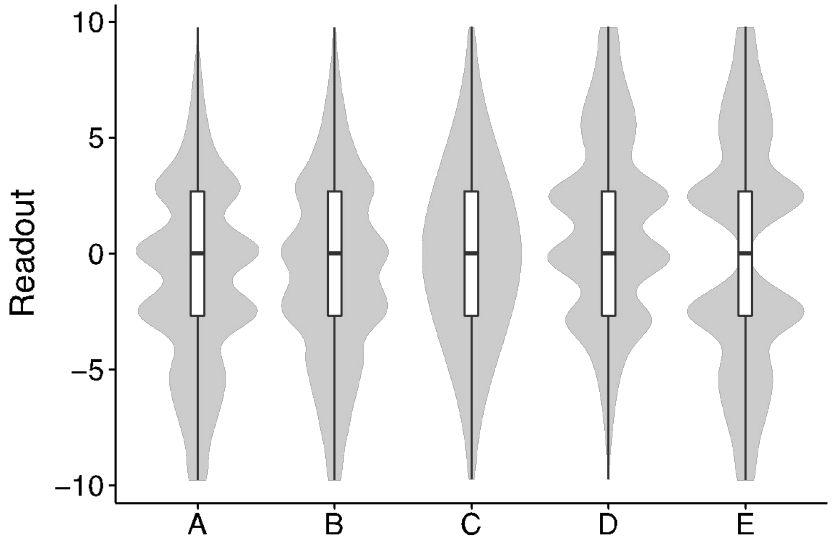
Are groups A-E the same?





Are groups A-E the same?

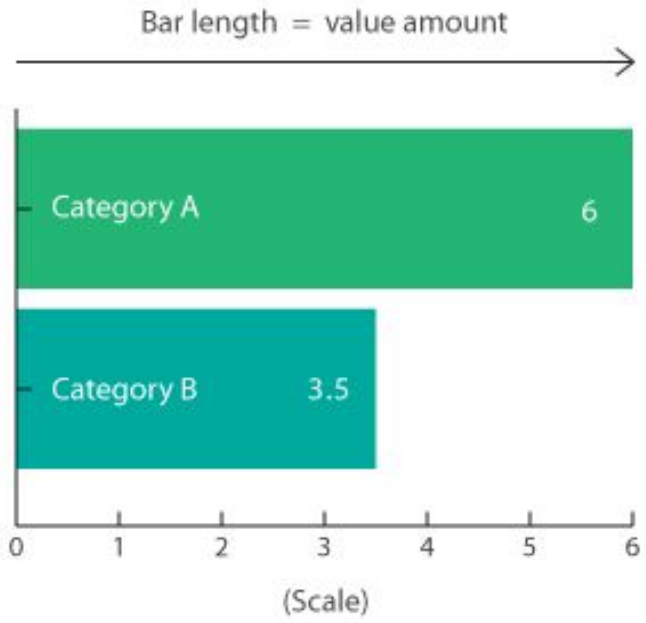




Categories

Bar (Column) Chart

One continuous variable
0-2 categorical variables



Uses horizontal/vertical bars to compare numerical values across categories.

Pros:

- Simple, easily to understand

Cons:

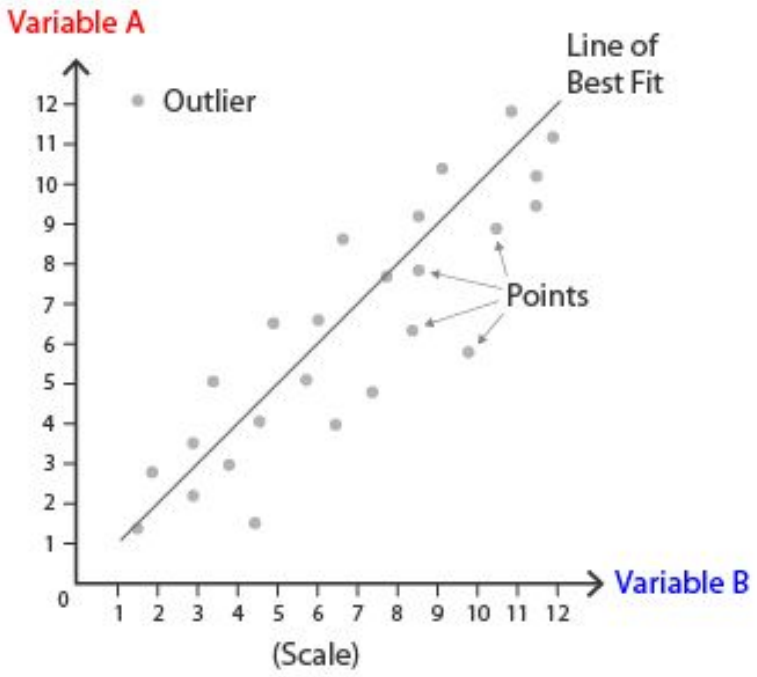
- Hides distribution of data
- Labeling and clutter becomes problematic as number of bars increases

Warning: Axis must start from zero!

Relationships

Scatter Plot

Two continuous variables



Collection of points placed on Cartesian coordinates (2 continuous variables)

Pros:

- Great to quickly identify correlations between variables
- Can include best-fit function to model correlations
- Useful for viewing clusters in data
- Include a 3rd categorical variable by changing point colour or shape

Cons:

- Difficult to read if too many data points close together

Relationships

Bubble Chart

3-4 continuous variable

Cross between scatter plot and proportional area chart

Pros:

- Can view distribution of 3-4 variables at once
- Great for interactive graphs

Cons:

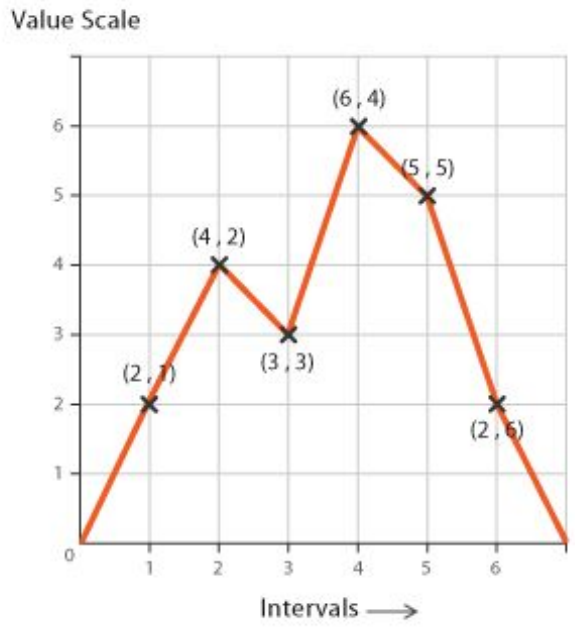
- Too many bubbles makes it difficult to read (limited capacity)

Warning: Value must be proportional to area, NOT radius!

Time Series

Line Graph

Two continuous variables



Displays values over a continuous interval or time period

Pros:

- Easy interpretation and identification of patterns in a dataset

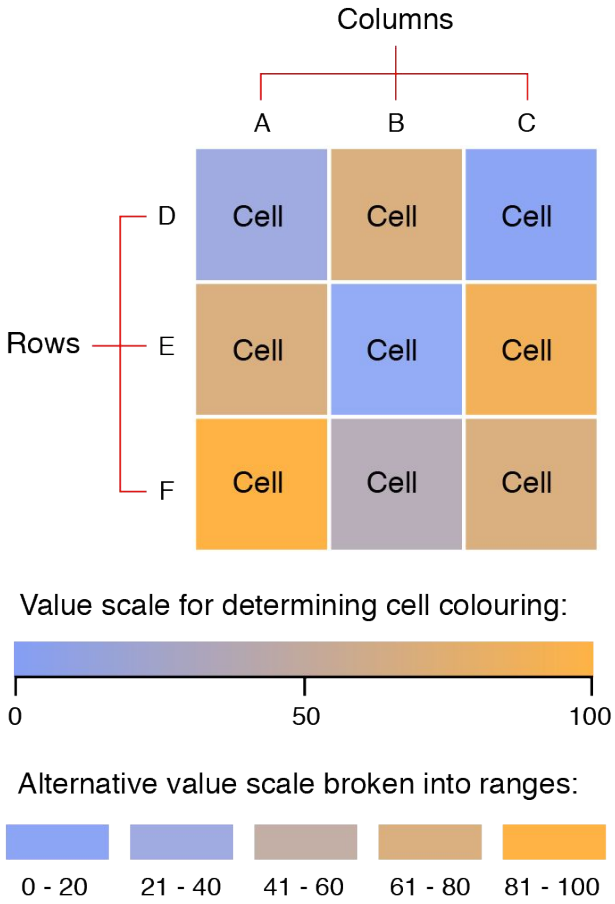
Cons:

- More than 3-4 lines are difficult to read (clutter)

Time Series

Heatmap

1-2 categorical variables
1 continuous/discrete variable



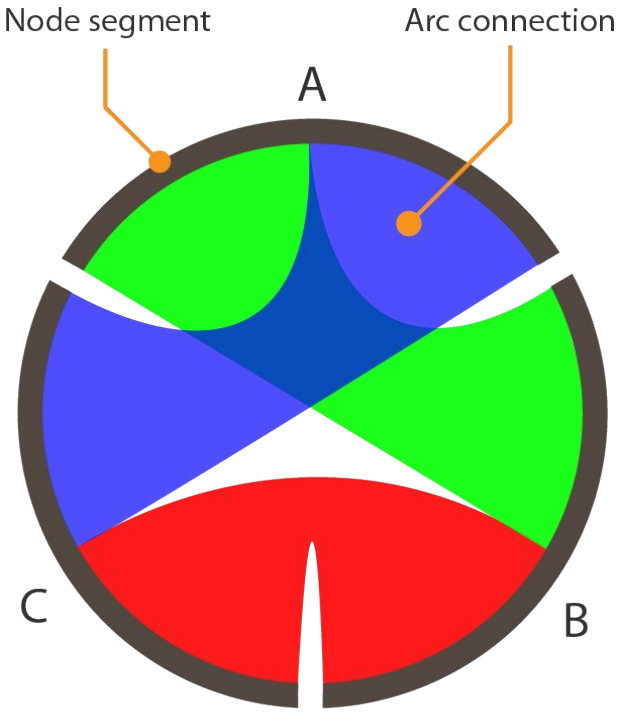
Visualize data through variation in colouring; useful for examining multivariate data

Pros:

- Qualitative view of large datasets make it easy to pick out patterns

Cons:

- Axis of large heat-maps difficult to read
- Less quantitative (precise) visualization



	A	B	C
A		10	10
B	10		10
C	10	10	

A $\xrightarrow{10}$ B

B $\xrightarrow{10}$ C

C $\xrightarrow{10}$ A

Relationships

Chord Diagrams

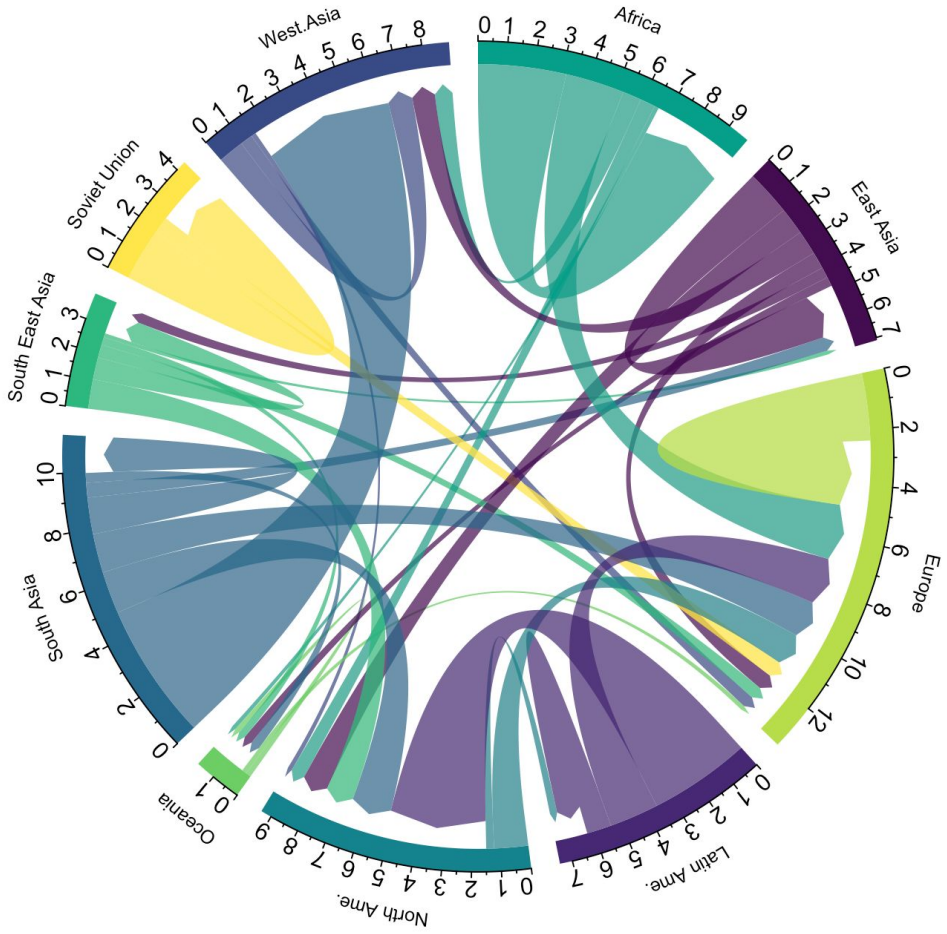
Visualizes relationships between entities (used to show what two groups have in common)

Pros:

- Visually appealing
- Useful for comparing similarities between different data subsets

Cons:

- Over-cluttering major issue with too many connections (hierarchical edge bundling used to reduce visual complexity)



EXAMPLE

Global migration rates

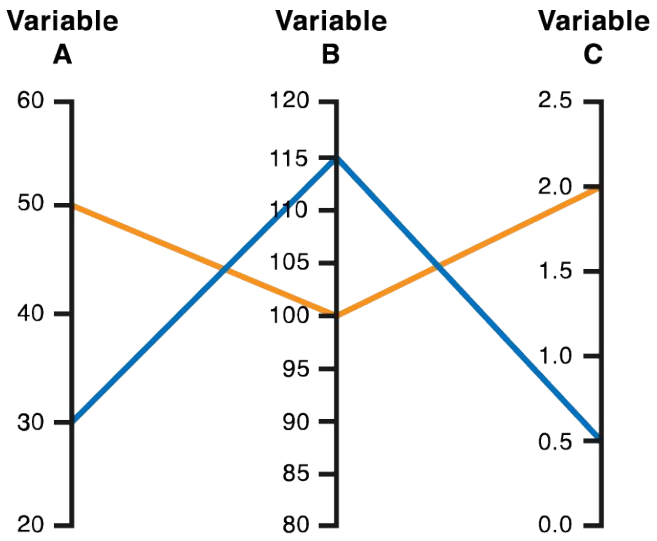
Data source: Abel GJ (2017). Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015.

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

3+ continuous variables

Relationships

Parallel Coordinates Plot



Data			
	Variable A	Variable B	Variable C
Item 1	50	100	2.0
Item 2	30	115	0.5

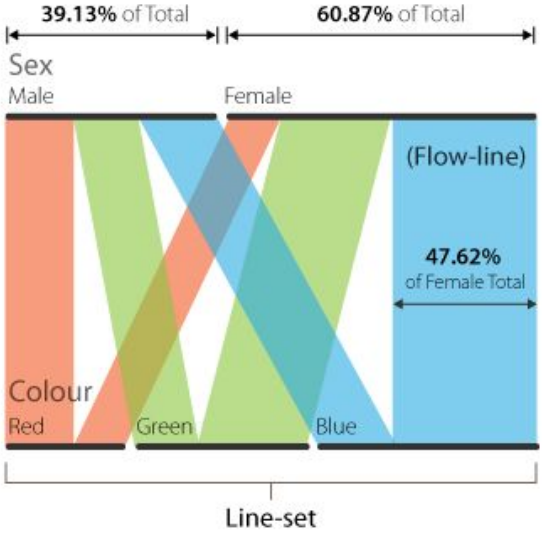
Used to plot multivariate, numerical data.

Pros:

- Helps identify relationships between adjacent variables

Cons:

- Easily become over cluttered (combat by using “brushing”)
- Data interpretation sensitive to variable ordering and axis-scales



Sex	Colour	Count	% of Sex TOTAL
Male	Red	35	32.41
	Green	33	30.56
	Blue	40	37.04
Female	Red	28	16.67
	Green	60	35.71
	Blue	80	47.62

Relationships

Allivual Diagrams

Used to show flow and proportions in data; similar to parallel coordinate plots (but show proportions)

Pros:

- Great for exploratory analysis on how groups relate to each other

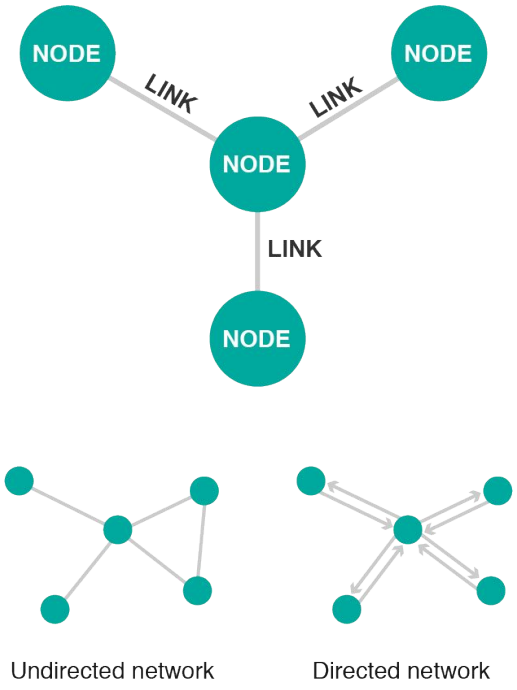
Cons:

- Difficult to read with more variables
- Interpretation sensitive to variable ordering

Relationships

Network Diagram

(Weighted) relationships



Uses nodes and links (edges) to represent relationships between different entities.

Pros:

- Intuitive visualization for relationships between entities
- Can display both directed and undirected relationships

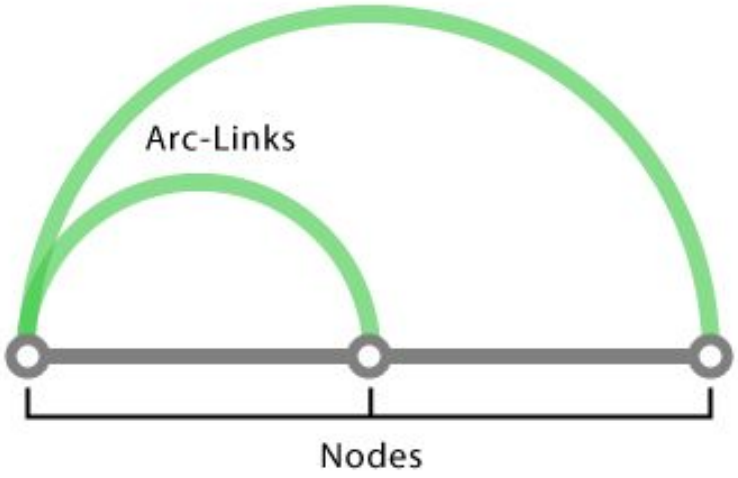
Cons:

- Limited data capacity; become hard to read “hairballs” with too many nodes
- Appearance sensitive to layout algorithm

Relationships

Arc Diagram

(Weighted) relationships



Method to represent 2D network diagrams

Pros:

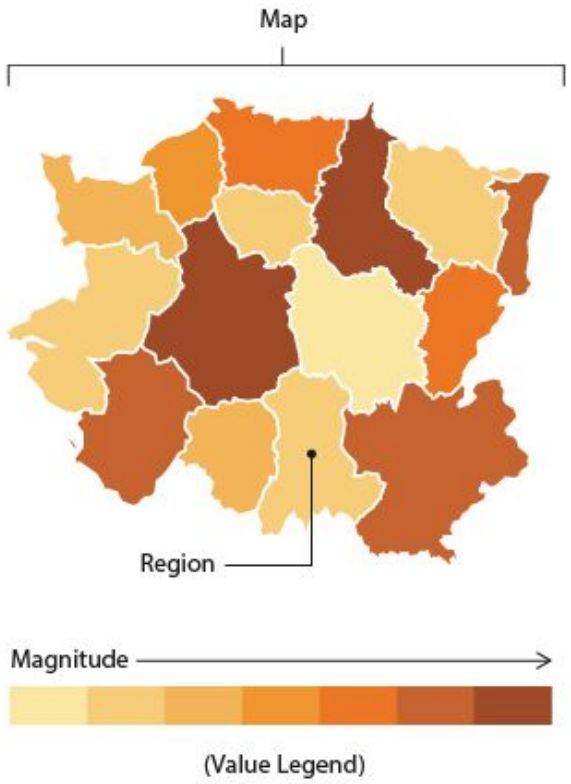
- Good at highlighting clusters and bridges

Cons:

- Too many links make diagram hard do read (clutter)
- Order of nodes is critical to quality of visualization
- Less effective at displaying overall network structure

Location

Choropleth Map



Visualizes values over a geographical area; geographical “Heat-map”

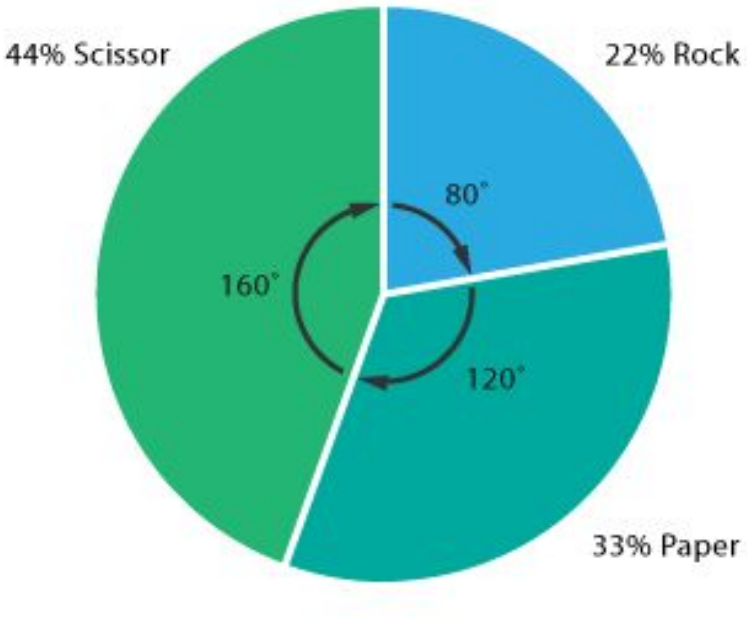
Pros:

- Common method to visualize data over geographic region (spatial distribution)
- Can identify patterns in data related to geography & location

Cons:

- Can’t accurately read or compare colours on map
- Larger regions more emphasized than smaller ones

For completion… Pie Charts



Show proportions or percentages between categories by dividing a circle into segments.

Pros:

- Gives a quick idea of the proportional distribution of the data

Cons:

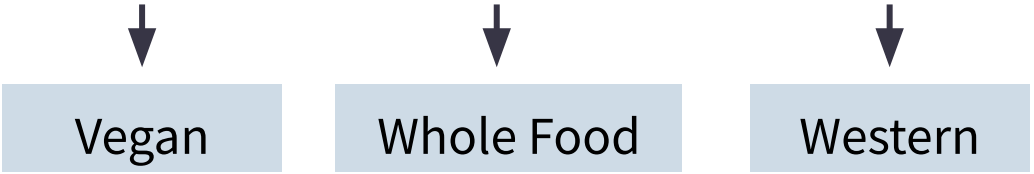
- Poor for making accurate comparisons across groups
- Cannot be used to show more then a few values
- Hard to distinguish proportions that are close in size

Exercise

Setting the scene

A team of researchers are interested in understanding **how various diets affects brain development during adolescence**. They conduct a longitudinal study to follow children from ages 12 to 18.

450 subjects (age 12)



Recorded at start: Sex, date of birth

Measured every year (until age 18)

- **Physical:** Weight (kg), height (cm)
- **Brain volume:** Gray matter volume (cm3), White matter volume (cm3)
- **Cognitive performance:** Reaction time (ms), Recall Score (Groton Maze Learning Test)

Exercise

Teams of ~10 people, 15 minutes

Your task is to design scientific figures that could help you explore and understand this dataset.

Develop your plan as a team, then sketch your ideas for scientific figure(s) on chart paper.

Consider:

- **The Story:** What specific questions might be asked? What chart(s) would best communicate the answer?
- **Clarity:** Tip - don't try to put all the data in one figure!
- **Who is the audience:** Your collaborators? (e.g. exploratory data visualization) Scientific publication?
- **Aesthetic:** What colours (if any) would you use? How would the plot look overall?

> You can achieve simplicity in the design of effective charts, graphs and tables by remembering three fundamental principles: **restrain**, **reduce**, **emphasize**.
>
> - Garr Reynolds

A small selection of data visualization libraries

Python
- Matplotlib, Seaborn, Plotly, ggPlot

R
- Base R, ggPlot2, Lattice, Highcharter

Javascript (for web)
- D3.js, Chart.js, Chartist.js, Highcharts

For questions or conversations on data visualization:

eisha.ahmed@mail.mcgill.ca

HEALTHY BRAINS FOR HEALTHY LIVES