Exploratory Data Analysis & Visualization

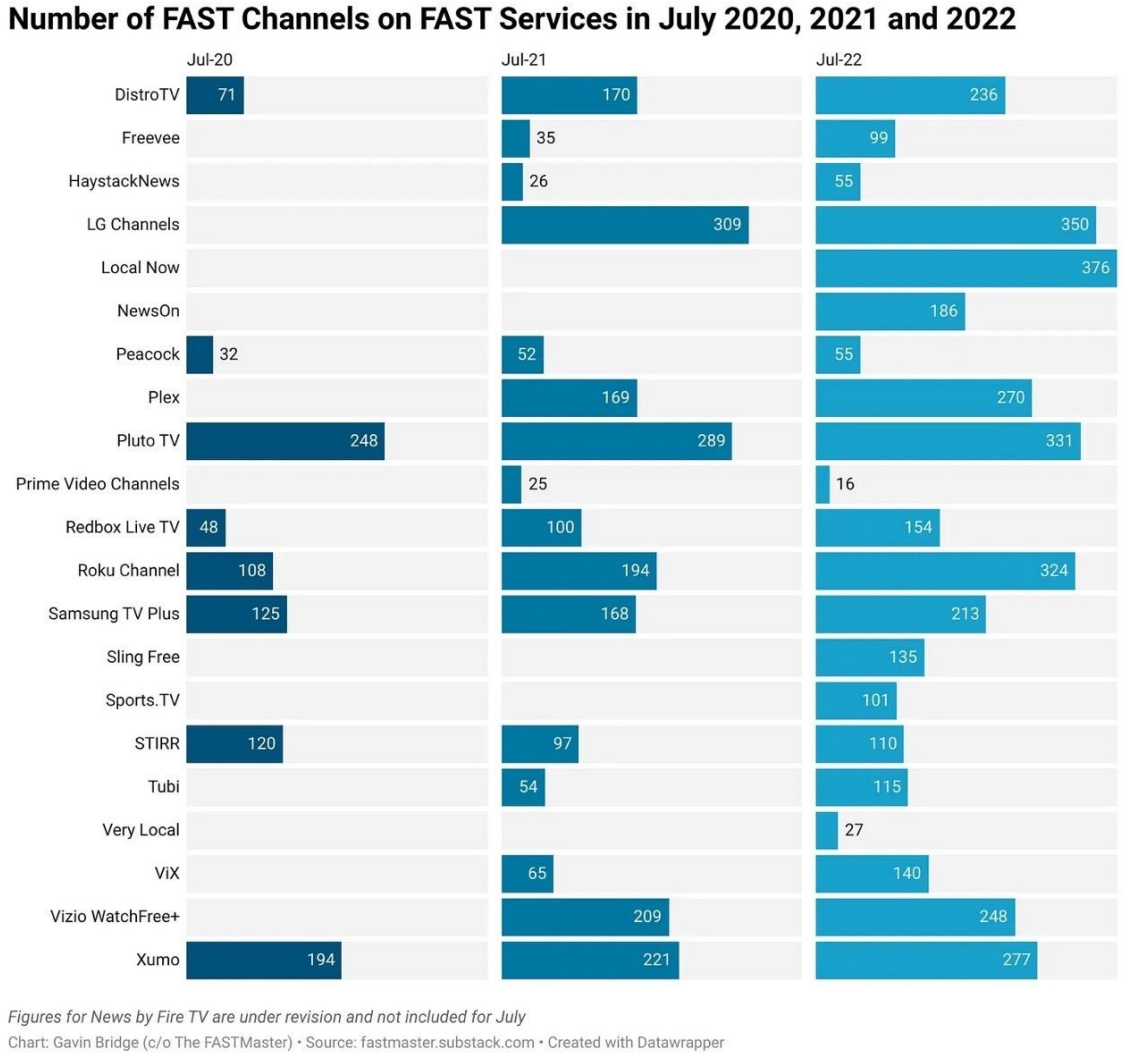
How good is your data?

How good is your visualization?

Ben Winjum

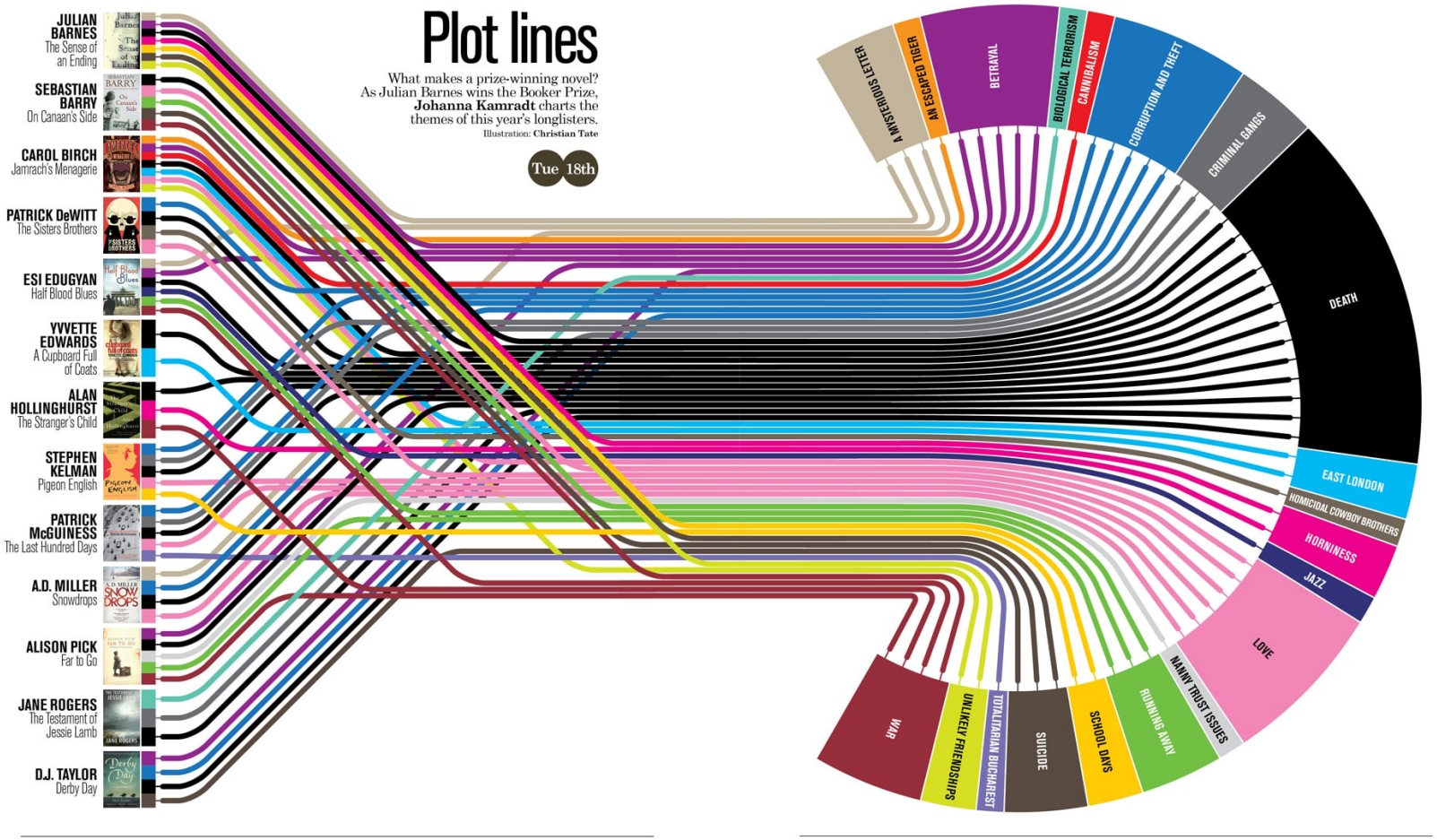
Questions about Setups? Materials? Logistics?

Example Visualizations from Discussions



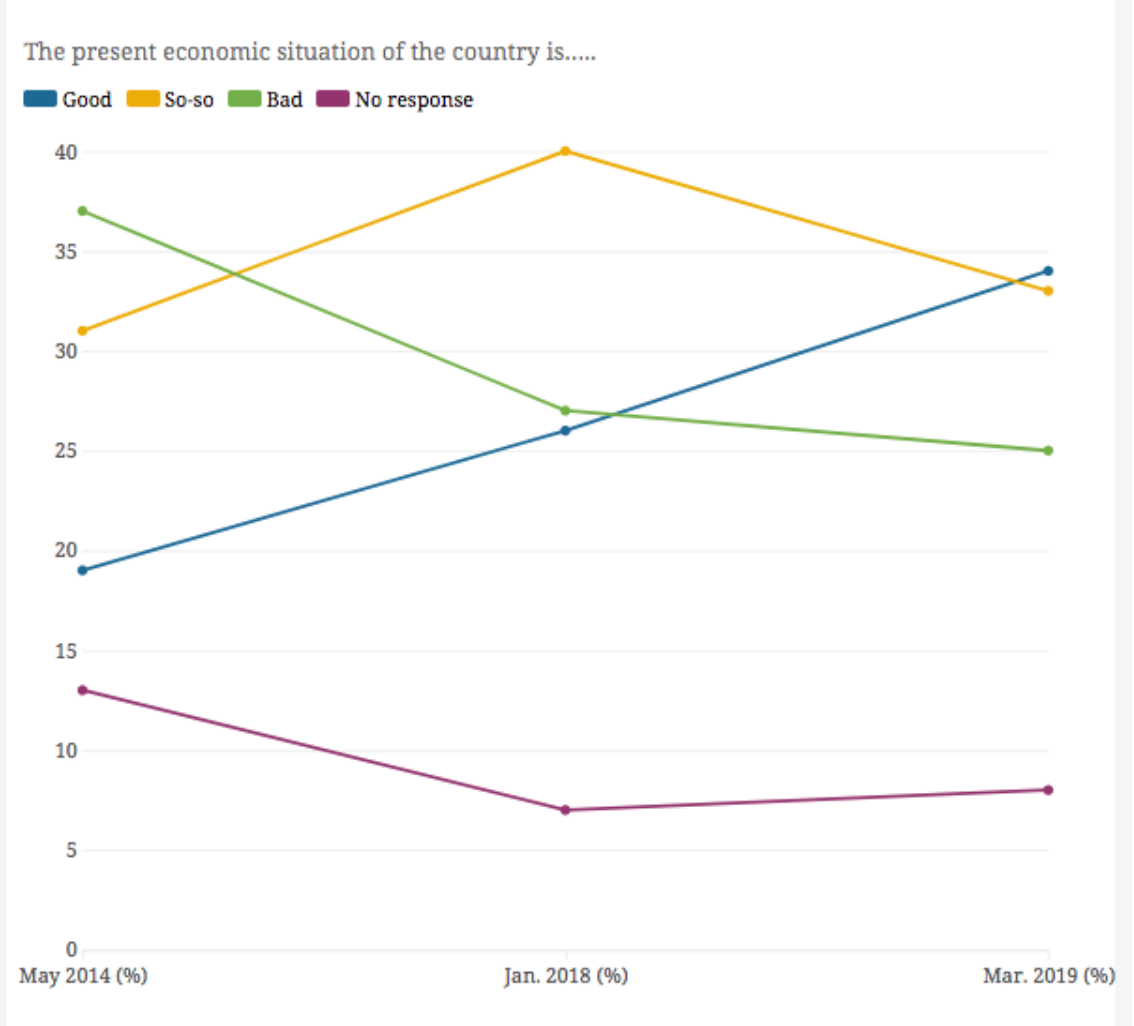
Gerardo

Example Visualizations from Discussions



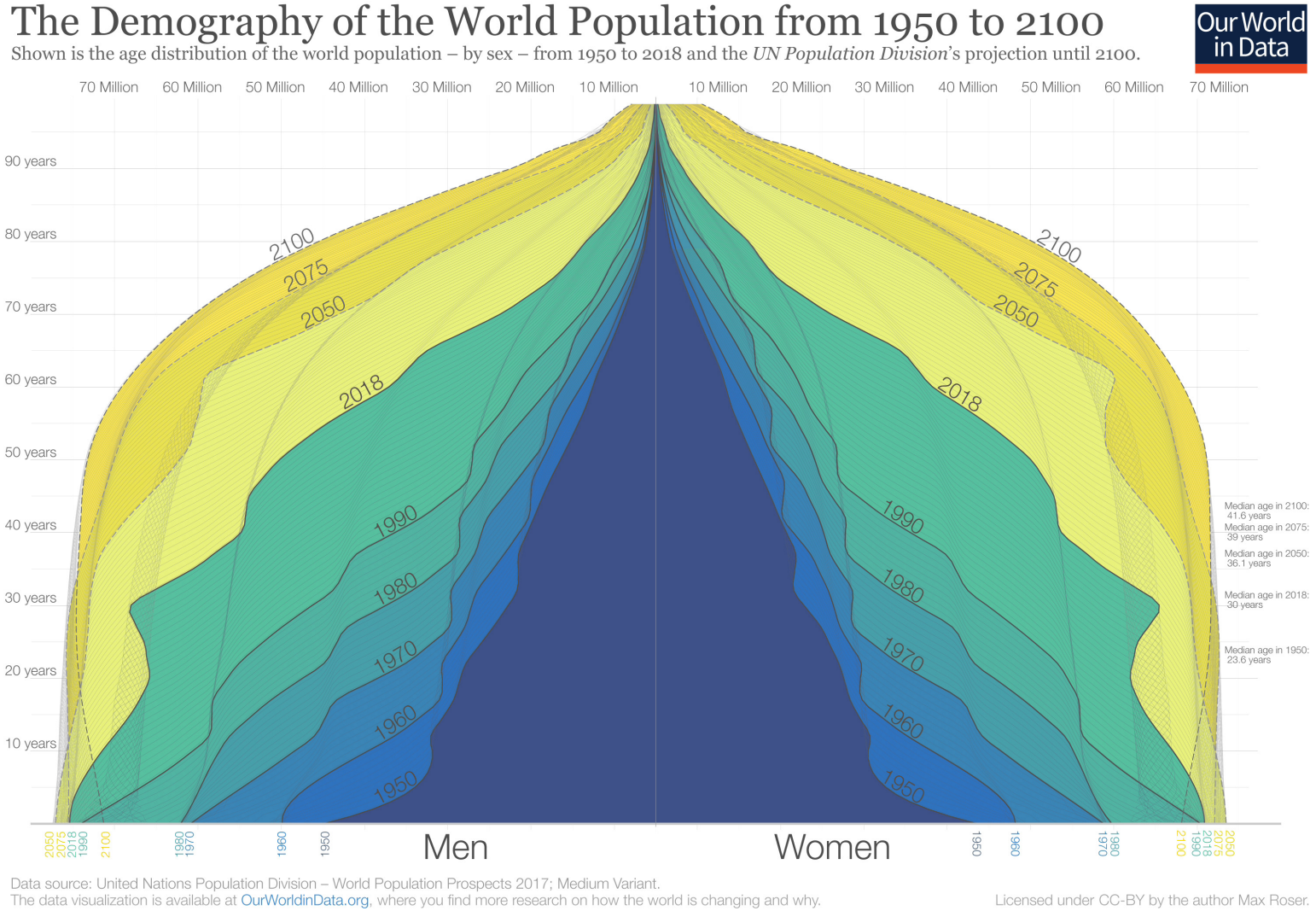
Miral

Example Visualizations from Discussions



Divya

Example Visualizations from Discussions



Tristan

https://ourworldindata.org/age-structure

Going Through the Assignments: Python

Going Through the Assignments: Python

- First let’s review some essential first steps in EDA

Look at your data to see if it passes the common-sense test

- Wrong values? Or metadata values (like number of rows)?
  - Can you detect them? And/or correct them?
  - Does it match with documentation?
  - Do the data have reasonable values?
- Unusable?
  - Does the data allow you to answer your question?
- Missing values?
  - How do you fill them in?
- Messy?
  - Can you use the data?
  - Does it need to be massaged?

Missing data can arise from various places

- A survey was conducted and values were just randomly missed when being entered in the computer.
- A respondent chooses not to respond to a question like `Have you ever recreationally used opioids?’
- You decide to start collecting a new variable (due to new actions: like a pandemic) partway through the data collection of a study.
- You want to measure the speed of meteors, and some observations are just 'too quick' to be measured properly.

Missing data type 1:
Missing Completely at Random

- Examples:
  - a coin is flipped to determine whether an entry is removed.
  - values were just randomly missed when being entered in the computer.
  - a cosmic particle hits your computer and destroys the bit that stores your data element

- Effect if you ignore:
  - there is no effect on inferences (or estimates)

- How to handle:
  - lots of options
  - best to impute (more on that soon)

Missing data type 2:
Missing at Random

- Examples:
  - men and women respond to the question "have you ever felt harassed at work?" at different rates (and may be harassed at different rates)
- Effect if you ignore:
  - inferences are biased and predictions are usually worsened
- How to handle:
  - use the information in the other predictors to build a model and impute a value for the missing entry
  - key: we can fix any biases by modeling and imputing the missing values based on what is observed!

Missing data type 3:
Missing Not at Random

- Example(s):
  - patients drop out of a study because they experience some really bad side effect that was not measured.
  - cheaters are less likely to respond when asked if you've ever cheated.

- Effect if you ignore:
  - major effects on inferences or predictions.

- How to handle:
  - you can 'improve' things by dealing with it like it is MAR, but you [likely] may never completely fix the bias
  - incorporating a missingness indicator variable may be the best approach (if it is in a predictor)

Correcting for missing data: we will consider basic approaches

- Delete it
  - Drop the feature
    - BUT: Is there so little data that this is reasonable? Or will you be dropping key info?
  - Drop the particular record
    - BUT: Is there a pattern to the data that's missing? And will this cause there to a resulting bias in your data if you drop the records that have missing data?
    - BUT: If you drop all the records, will what's left be too little to do modeling? And/or make your model brittle?

Missing data

- Delete it
- Fill it with another value (imputation)
  - Mean, median, mode
  - If you have other relevant values, you can use subset for calculations
    - Example, if a male height is missing, substitute with the median value only of other male heights
  - Use regression or classification
  - Fill with values of the “nearest” records
  - Interpolate from “nearest” records

Correcting for missing data

- We are not focused on modeling and will take the basic approach
  - Drop the observations that have any missing values.
    - Use pd.DataFrame.dropna(axis=0)
  - Discard the entire feature
    - Use pd.DataFrame.drop(column,axis=1)
  - Imputation: substitute a value like the mean or median (if quantitative) or most common class (if categorical) for all missing values
    - Use pd.DataFrame.fillna(value=x.mean())

Correcting for missing data

- We are not focused on modeling and will take the basic approach
  - Drop the observations that have any missing values.
    - Use pd.DataFrame.dropna(axis=0)
  - Discard the entire feature
    - Use pd.DataFrame.drop(column,axis=1)
  - Imputation: substitute a value like the mean or median (if quantitative) or most common class (if categorical) for all missing values
    - Use pd.DataFrame.fillna(value=x.mean())
- FYI: there are several alternative approaches to imputation
  - Create a new variable that is an indicator of missingness
  - Hot deck imputation: for each missing entry, randomly select an observed entry in the variable and plug it in.
  - Model the imputation: plug in predicted values from a model based on the other observed predictors.
  - Model the imputation with uncertainty: plug in predicted values plus randomness from a model based on the other observed predictors.

Outliers

- Method of finding
  - Boxplot
  - Scatter plot
  - Histogram – might notice that in the default plot, the scales are way off
- Methods for dealing with them
  - Similar to missing values (except for discarding an entire feature)

Going Through the Assignments: Python

Data

- What is it?
- How do we store it?
- How do we think about it?
- How do we use it?

What is data?

- Webster's Dictionary:
  - factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
  - information in digital form that can be transmitted or processed
  - information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

What is data?


KANYE WEST
@kanyewest
Follow
I hate when I'm on a flight and I wake up with a water bottle next to me like oh great now I gotta be responsible for this water bottle
RETWEETS
48,234
LIKES
52,776
3:09 PM - 27 Jan 2016

What is data?

Text

What is data?

Pictures

Text

What is data?

Pictures

Text

Numbers

What is data?

Pictures

Text

Numbers

Dates

What is data?

Pictures

Text

Numbers

States

Dates

What is data?

Tweet (aggregate)

Pictures

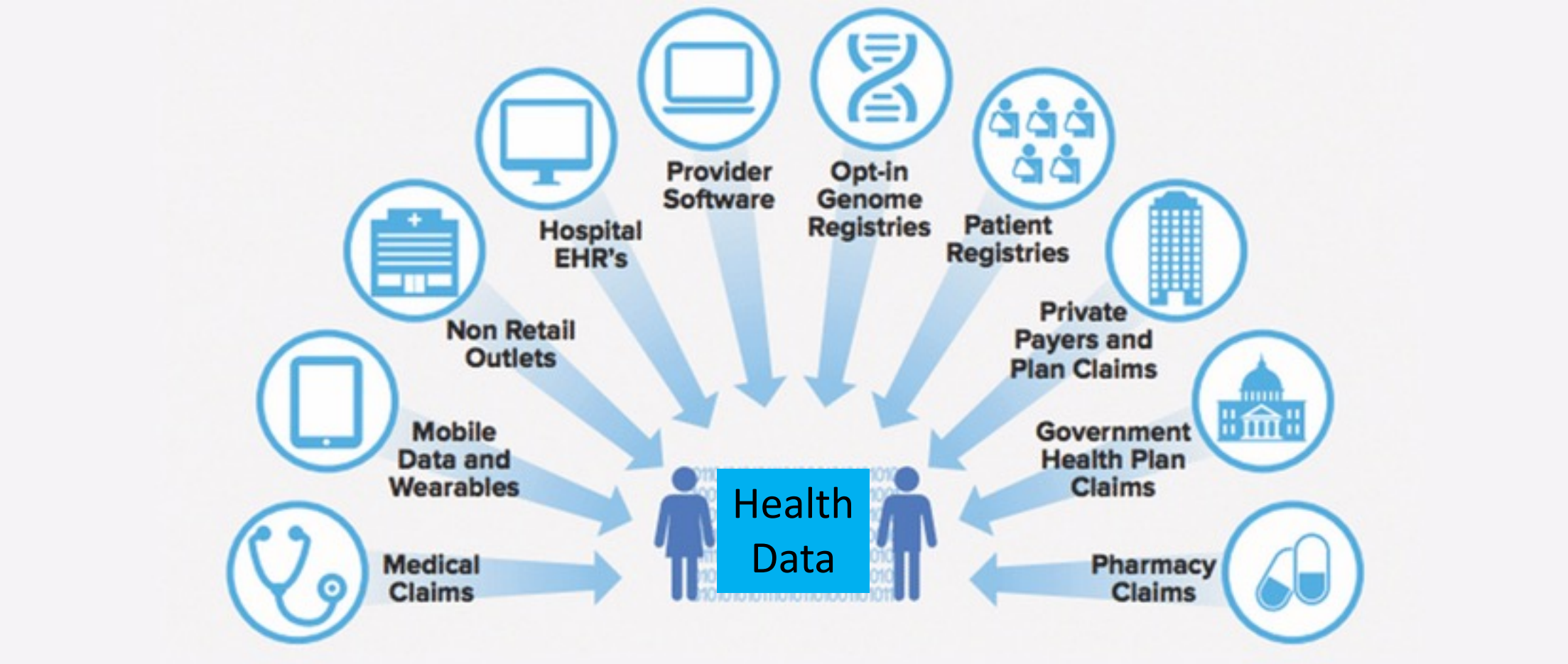
Text

Numbers

States

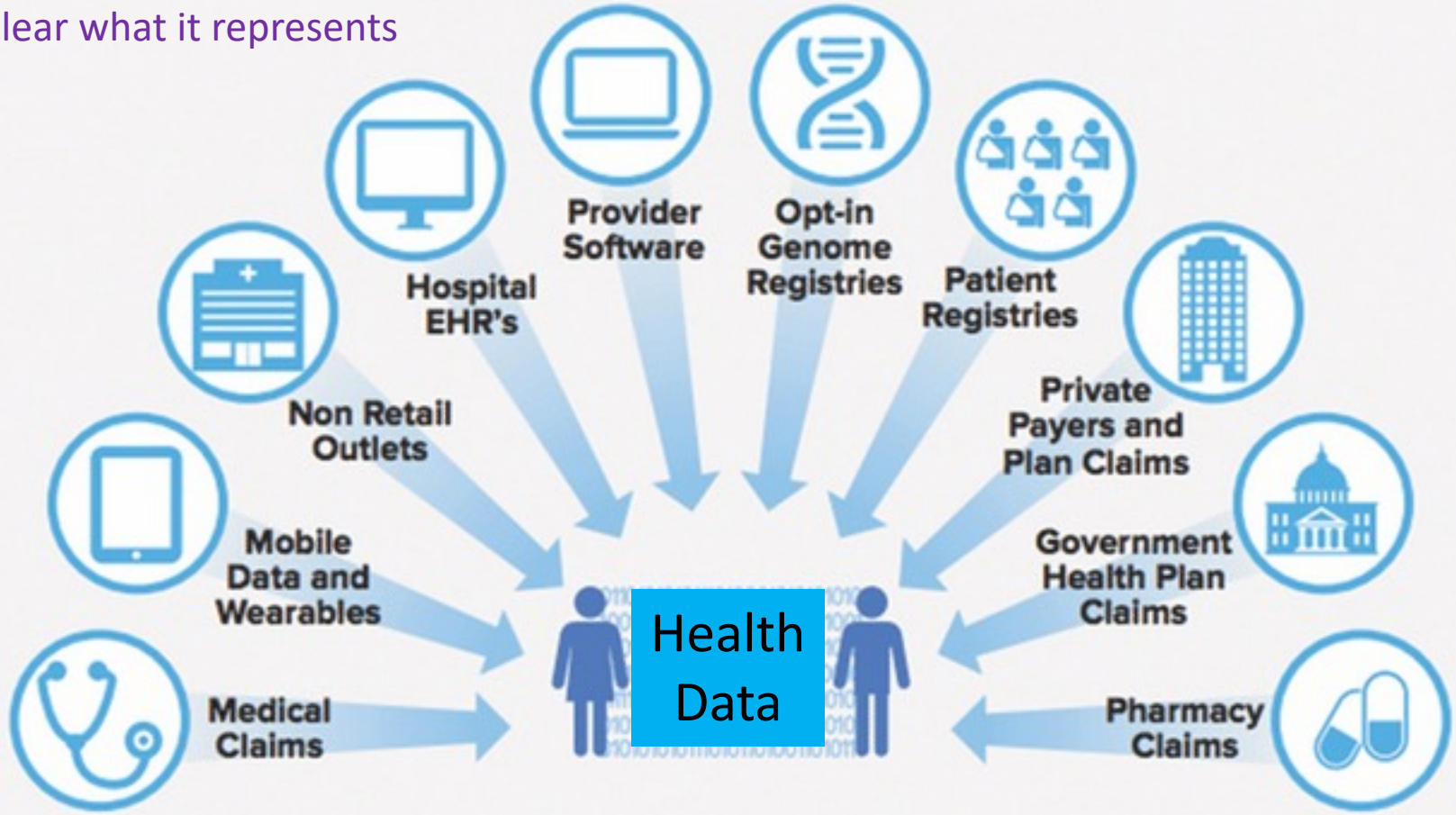
Dates

What is data?


Provider Software
Opt-in Genome Registries
Hospital EHR's
Patient Registries
Non Retail Outlets
Private Payers and Plan Claims
Mobile Data and Wearables
Government Health Plan Claims
Health Data
Medical Claims
Pharmacy Claims

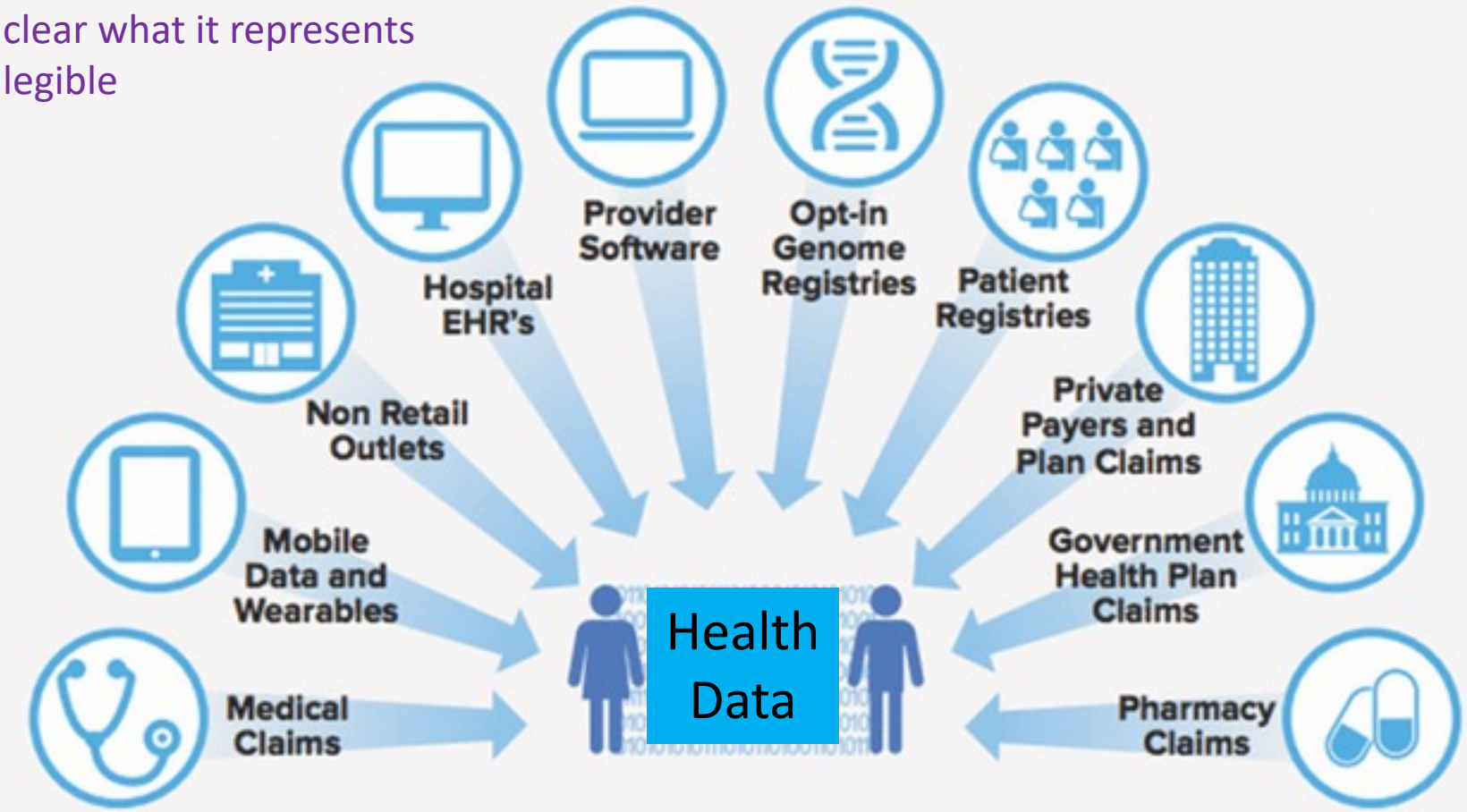
What is data?

It may not be clear what it represents



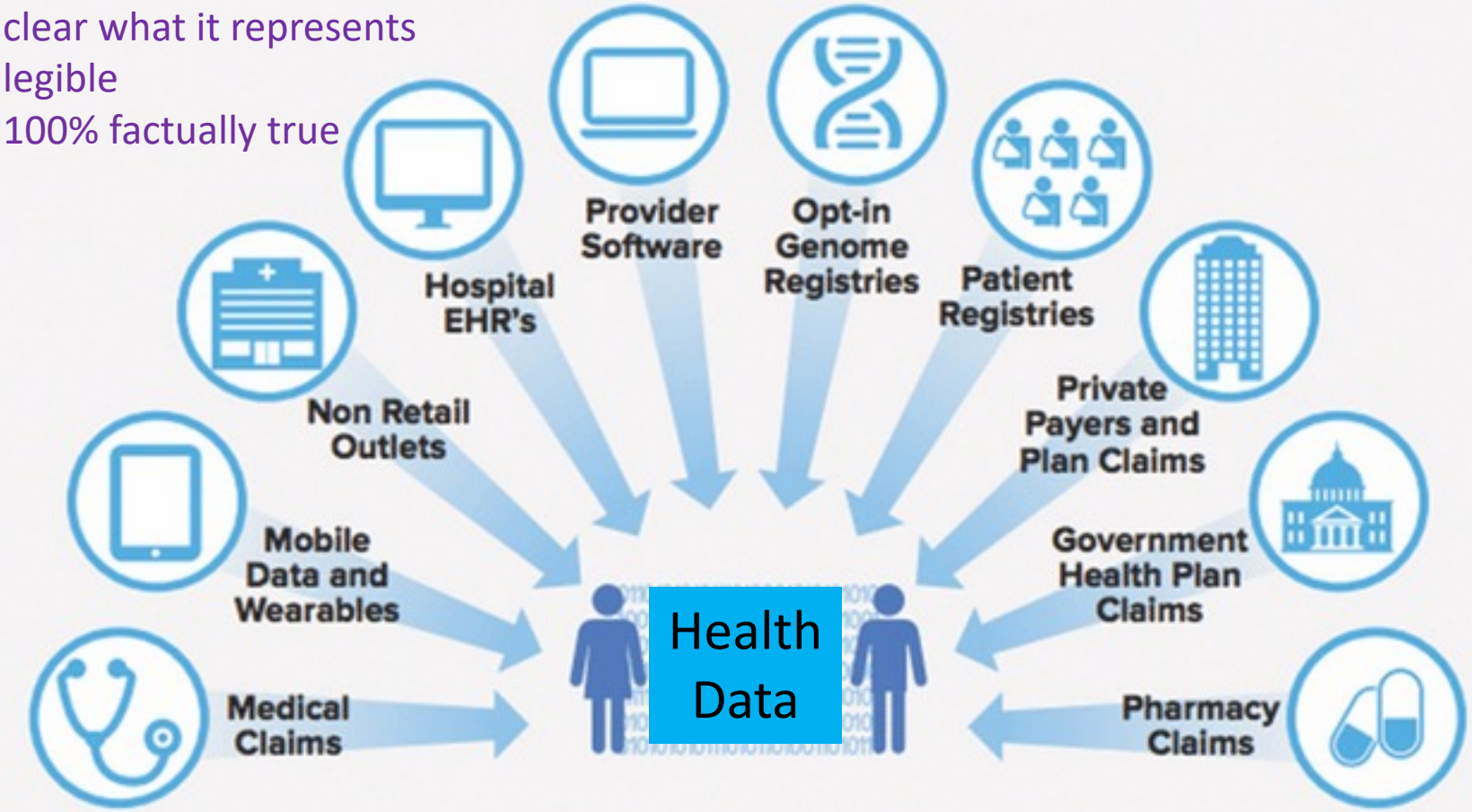
What is data?

It may not be clear what it represents
It may not be legible



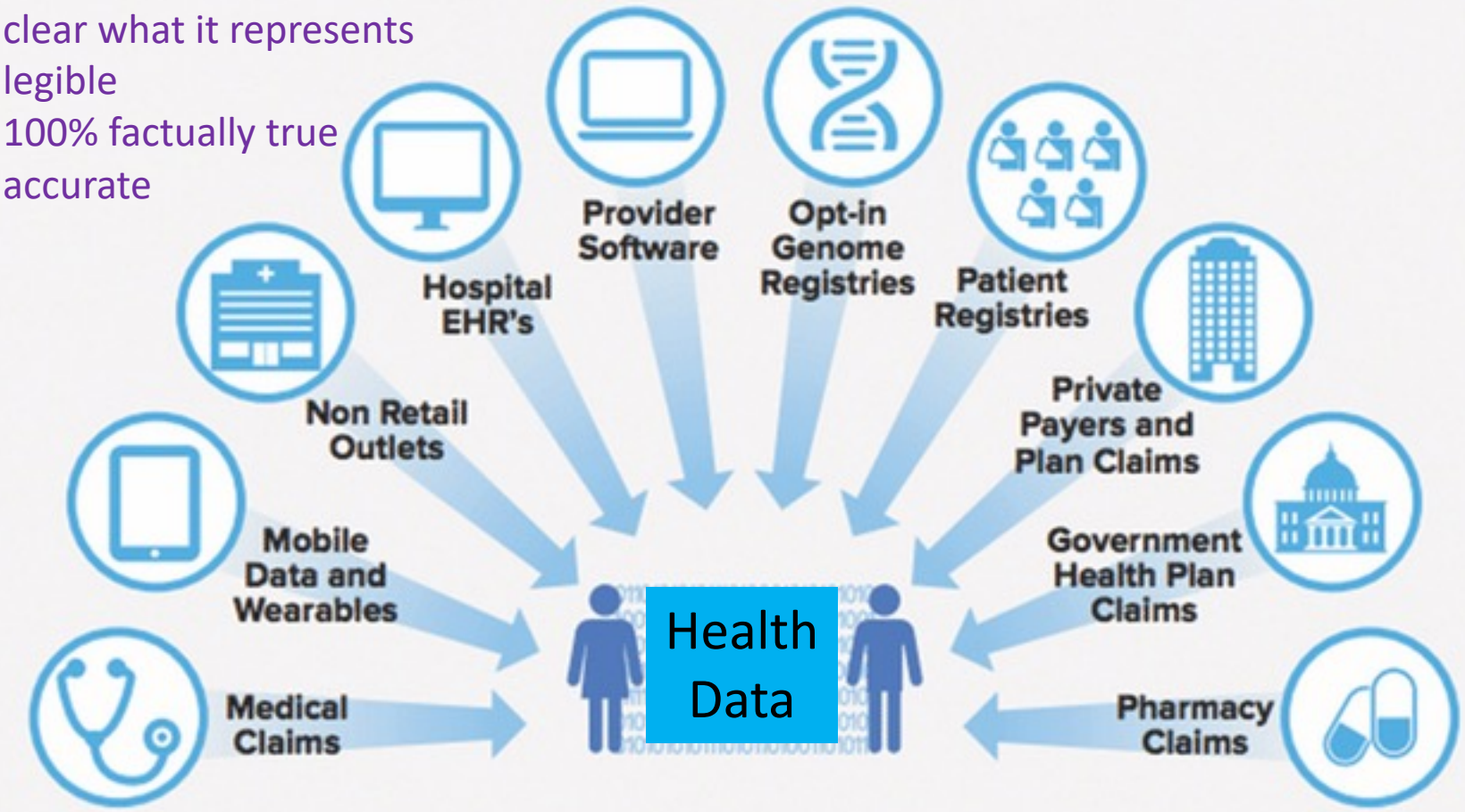
What is data?

It may not be clear what it represents
It may not be legible
It may not be 100% factually true



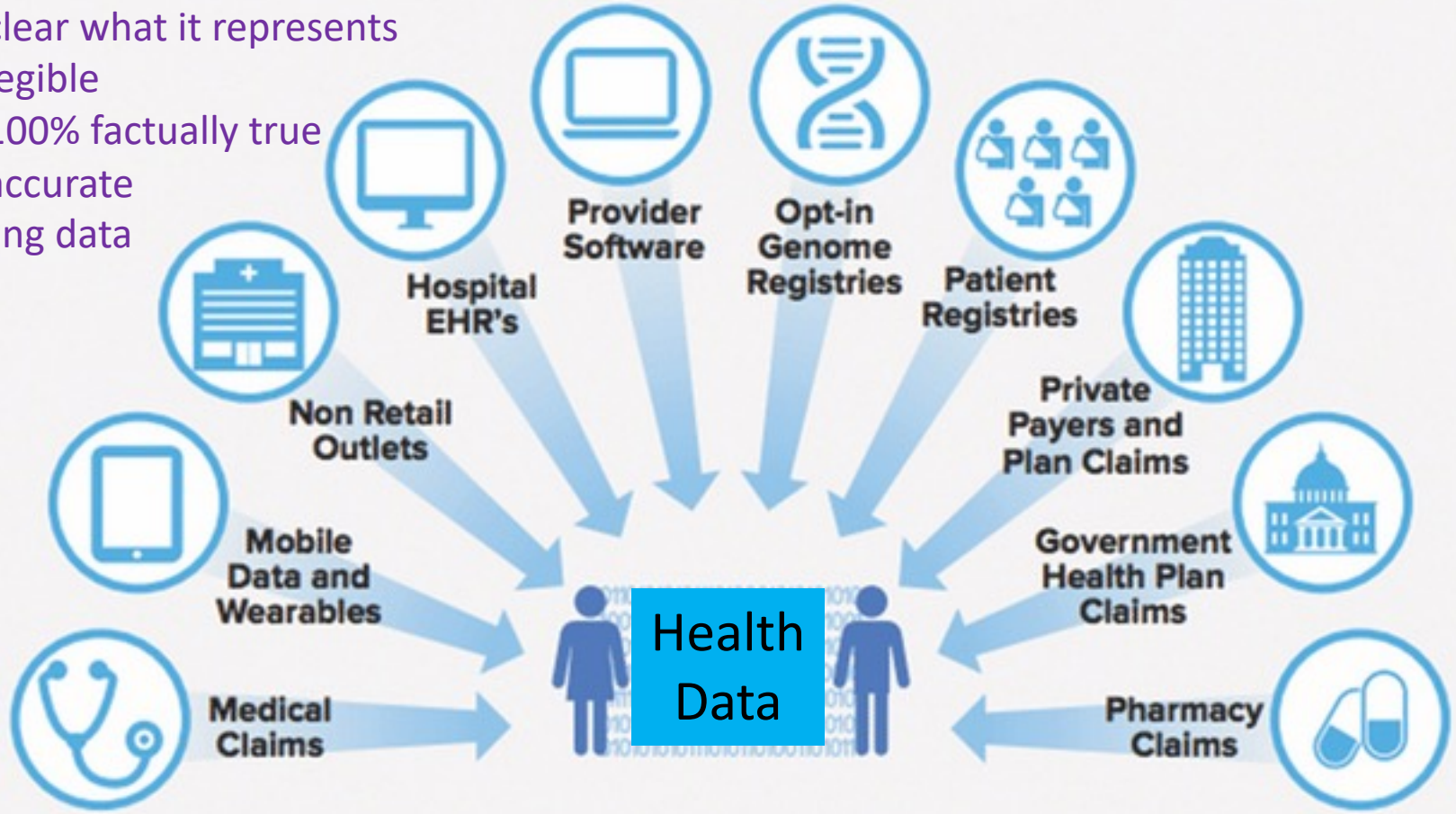
What is data?

It may not be clear what it represents
It may not be legible
It may not be 100% factually true
It may not be accurate

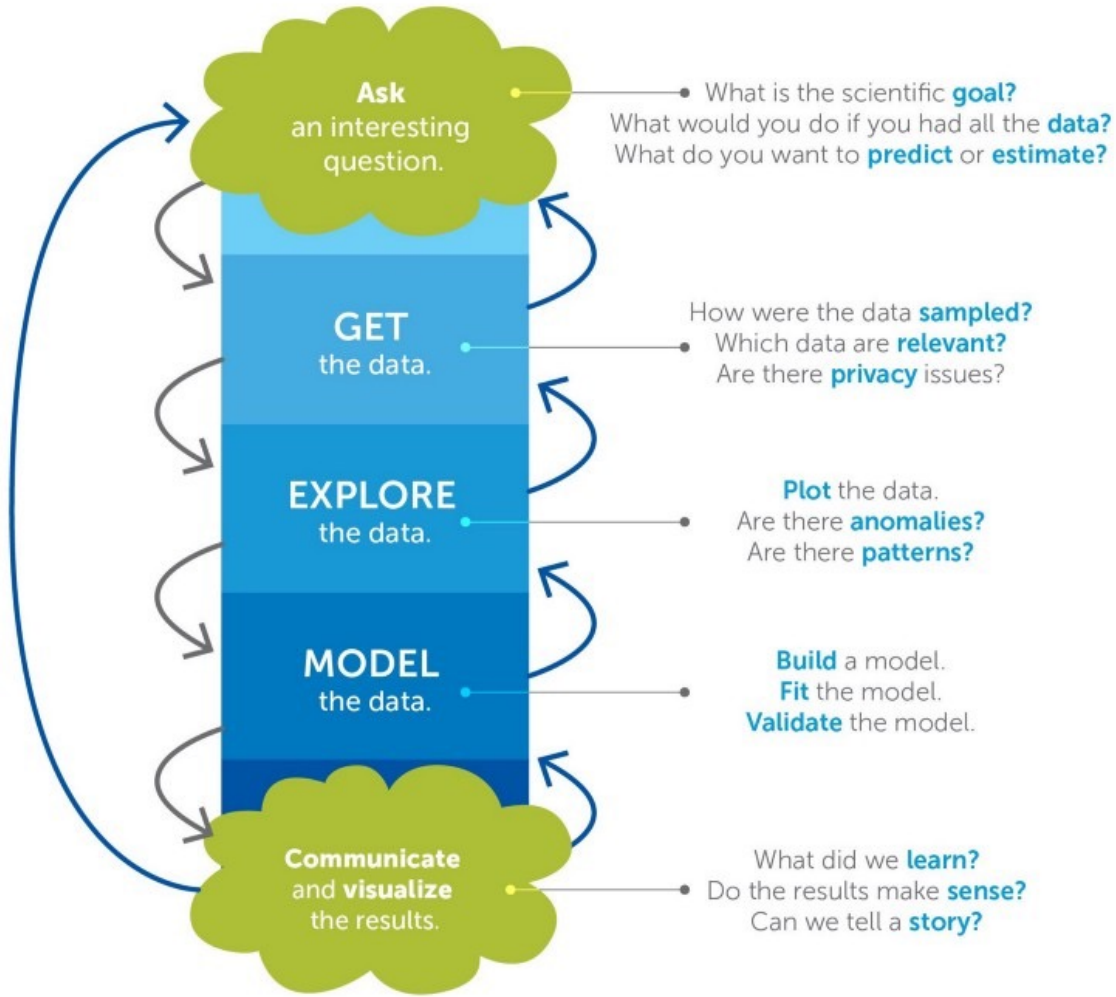


What is data?

It may not be clear what it represents
It may not be legible
It may not be 100% factually true
It may not be accurate
It may be missing data



The Data Science Process



Derived from the work of Joe Blitzstein and Hanspeter Pfister,
originally created for the Harvard data science course http://cs109.org/.

Getting data

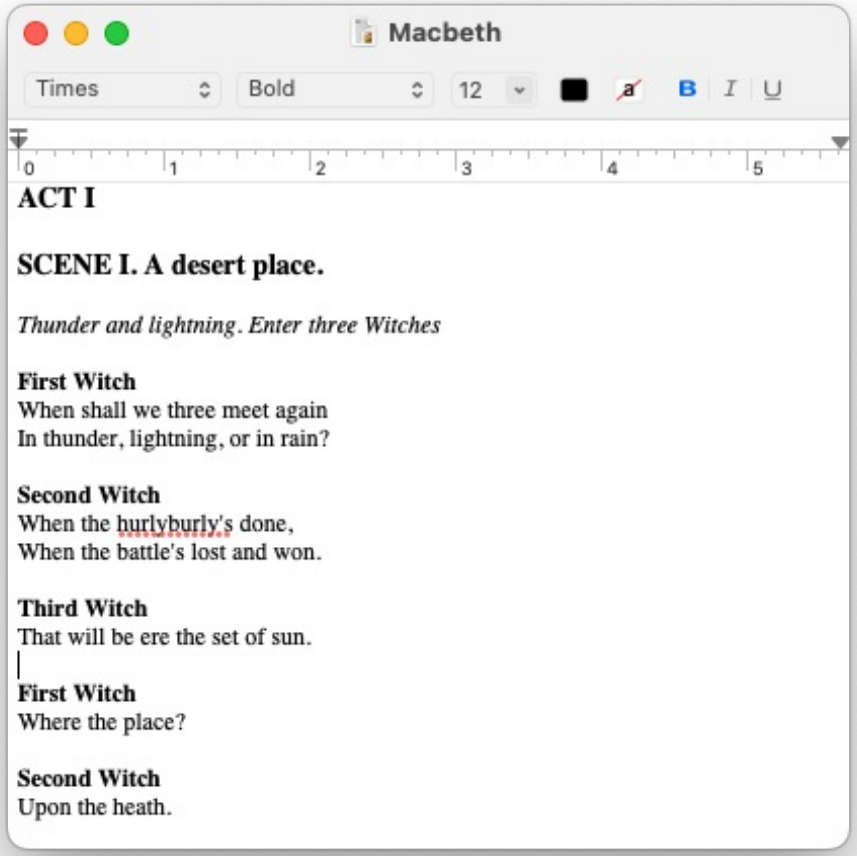
- What data is necessary to answer our question?
- Who collected it?
  - Is it from a reliable source? An authoritative source? (.com, .gov, .org)?
- When and where was it collected?
- How is it stored?
- How much data is necessary?
  - Comprehensive data set or sampled data set?
  - Are there any biases in the collection method?
- Are there restrictions or limitations for using it?
  - Is it licensed?
  - How difficult is it to analyze?
  - What is the allowed usage of data under its license?
- Are there humanistic concerns related to the data?
  - Did its collection need to be approved by an IRB?
  - Does its storage and use fall under HIPAA requirements? FERPA requirements?

Using data: formats

- Text
  - Web standards: html, xml, css, svg, json, ...
  - Source code: c, cpp, h, cs, js, py, java, rb, pl, php, sh, ...
  - Documents: txt, tex, markdown, asciidoc, rtf, ps, ...
  - Configuration: ini, cfg, rc, reg, ...
  - Tabular data: csv, tsv, ...
- Binary
  - Images: jpg, png, gif, bmp, tiff, psd, ...
  - Videos: mp4, mkv, avi, mov, mpg, vob, ...
  - Audio: mp3, aac, wav, flac, ogg, mka, wma, ...
  - Documents: pdf, doc, xls, ppt, docx, odt, ...
  - Archive: zip, rar, 7z, tar, iso, ...
  - Database: mdb, accde, frm, sqlite, ...
  - Executable: exe, dll, so, class, ...

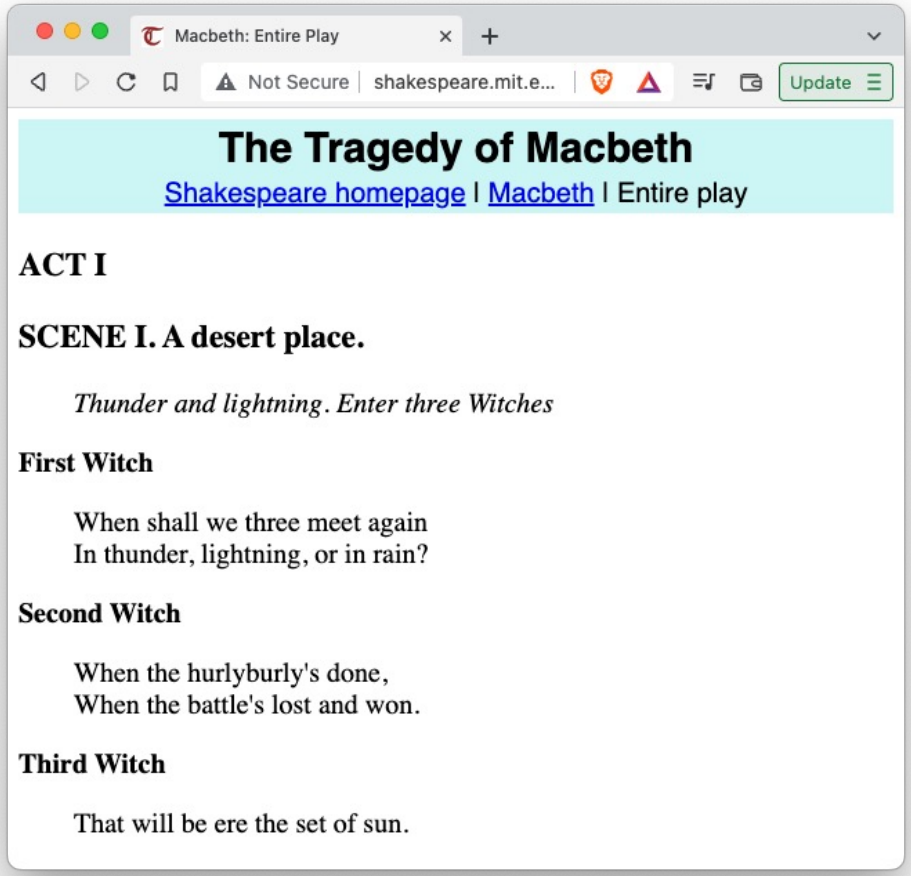
Using data: format examples

- Text
  - Plain text file (*.txt)



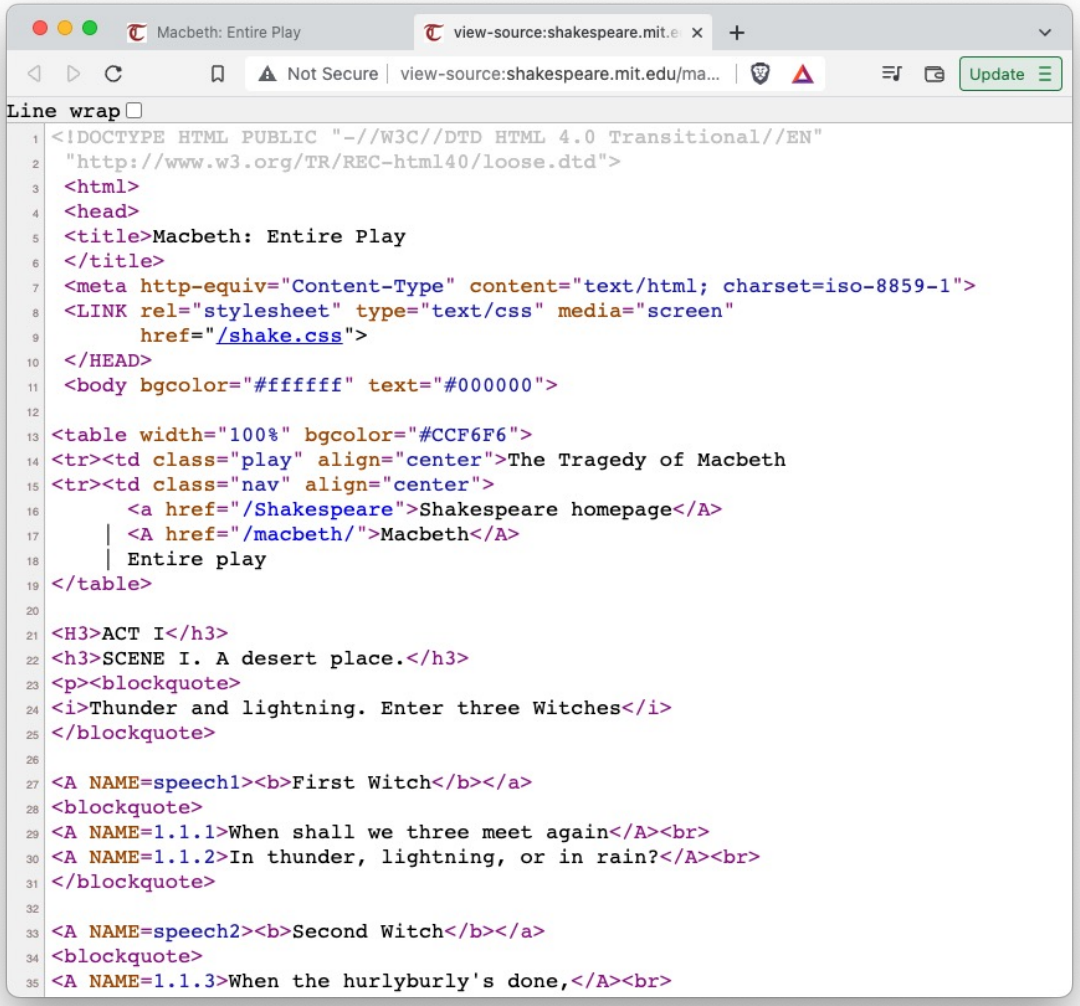
Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)



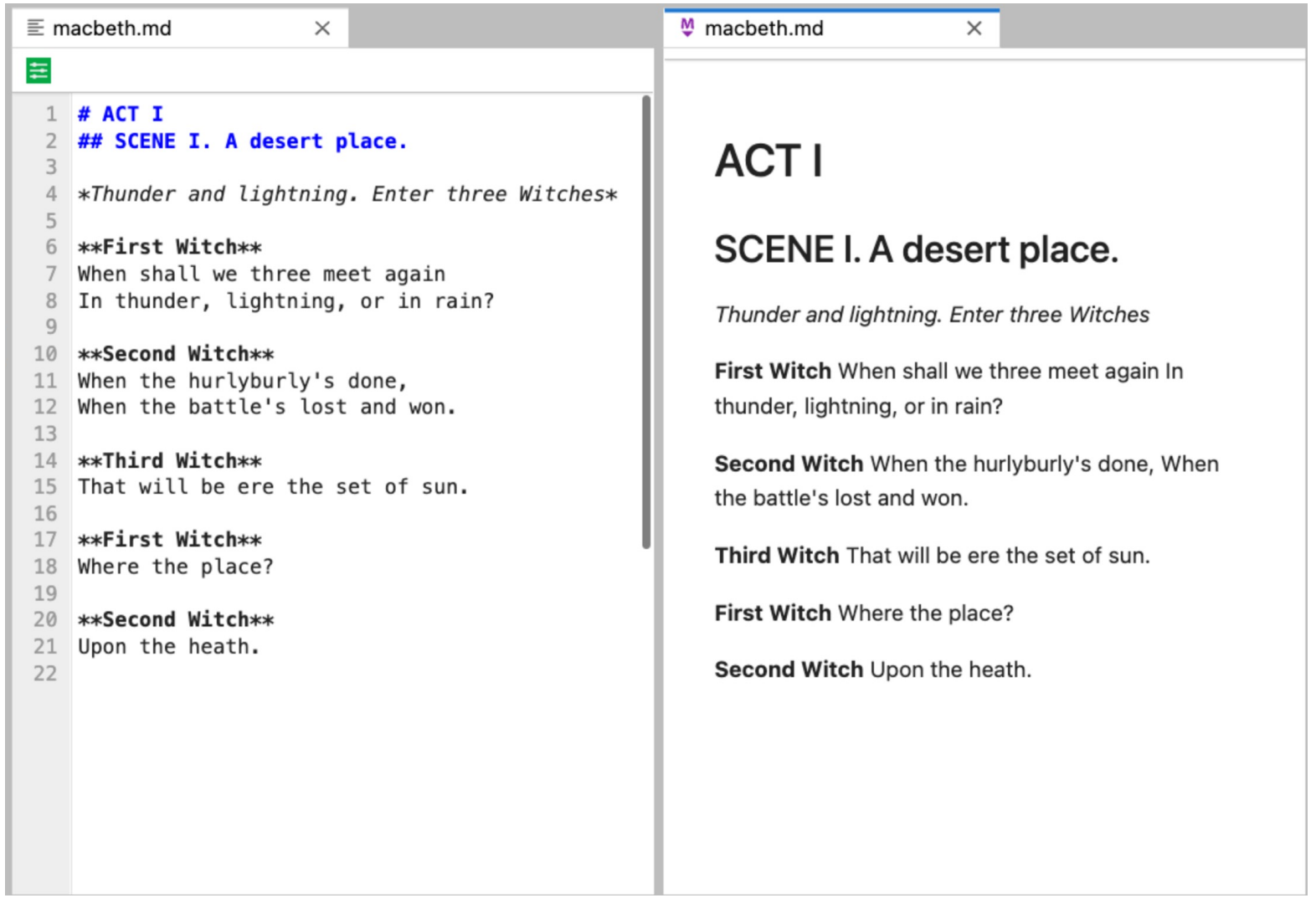
Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)



Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)
  - Markdown (*.md)



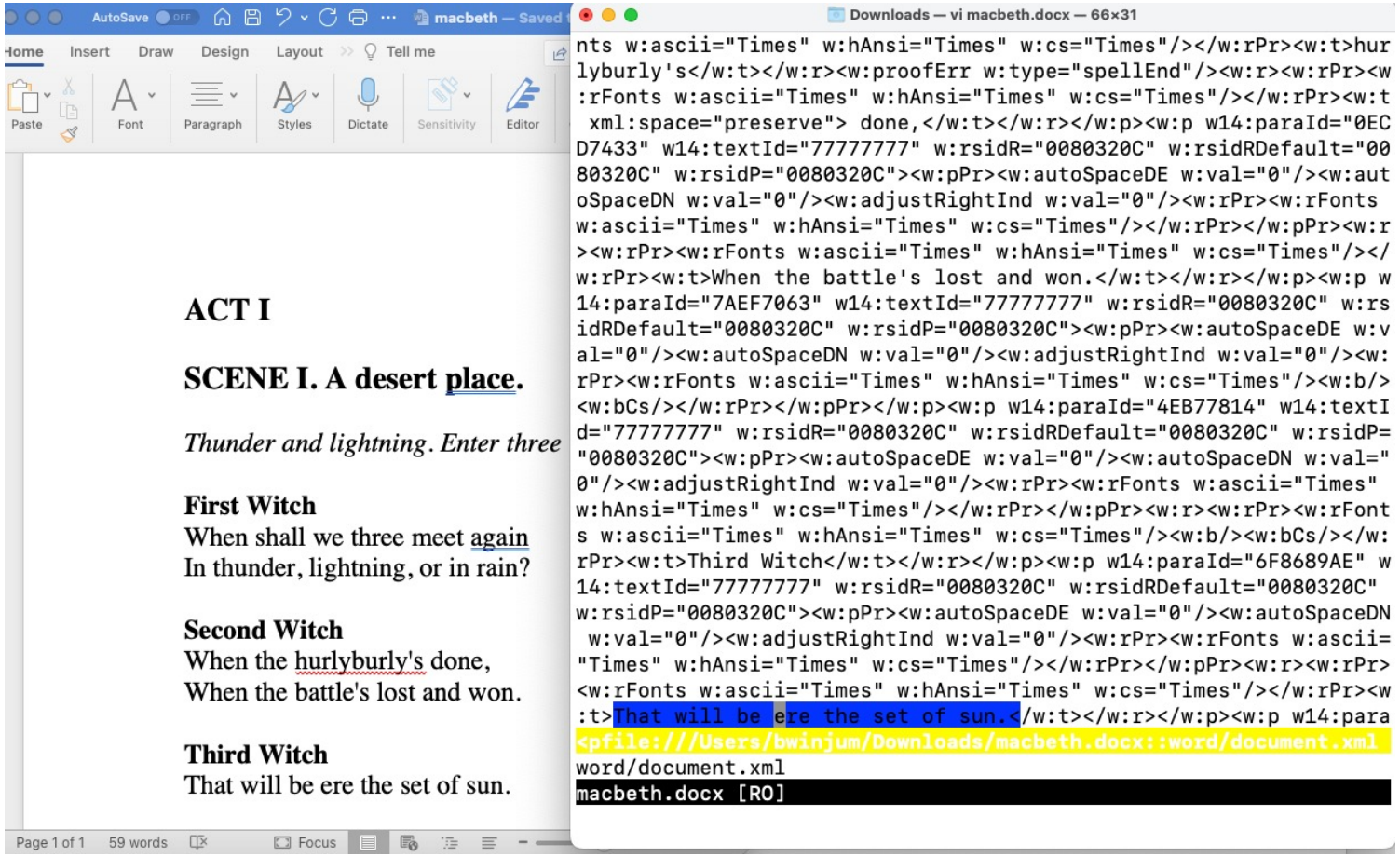
Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)
  - Markdown (*.md)
  - XML (*.xml)

```
<roll_call_vote>
 <congress>115</congress>
 <session>1</session>
 …
 <members>
  <member>
   <member_full>Alexander (R-TN)</member_full>
   <last_name>Alexander</last_name>
   <first_name>Lamar</first_name>
   <party>R</party>
   <state>TN</state>
   <vote_cast>Yea</vote_cast>
   …
  </member>
 </members>
</roll_call_vote>
```

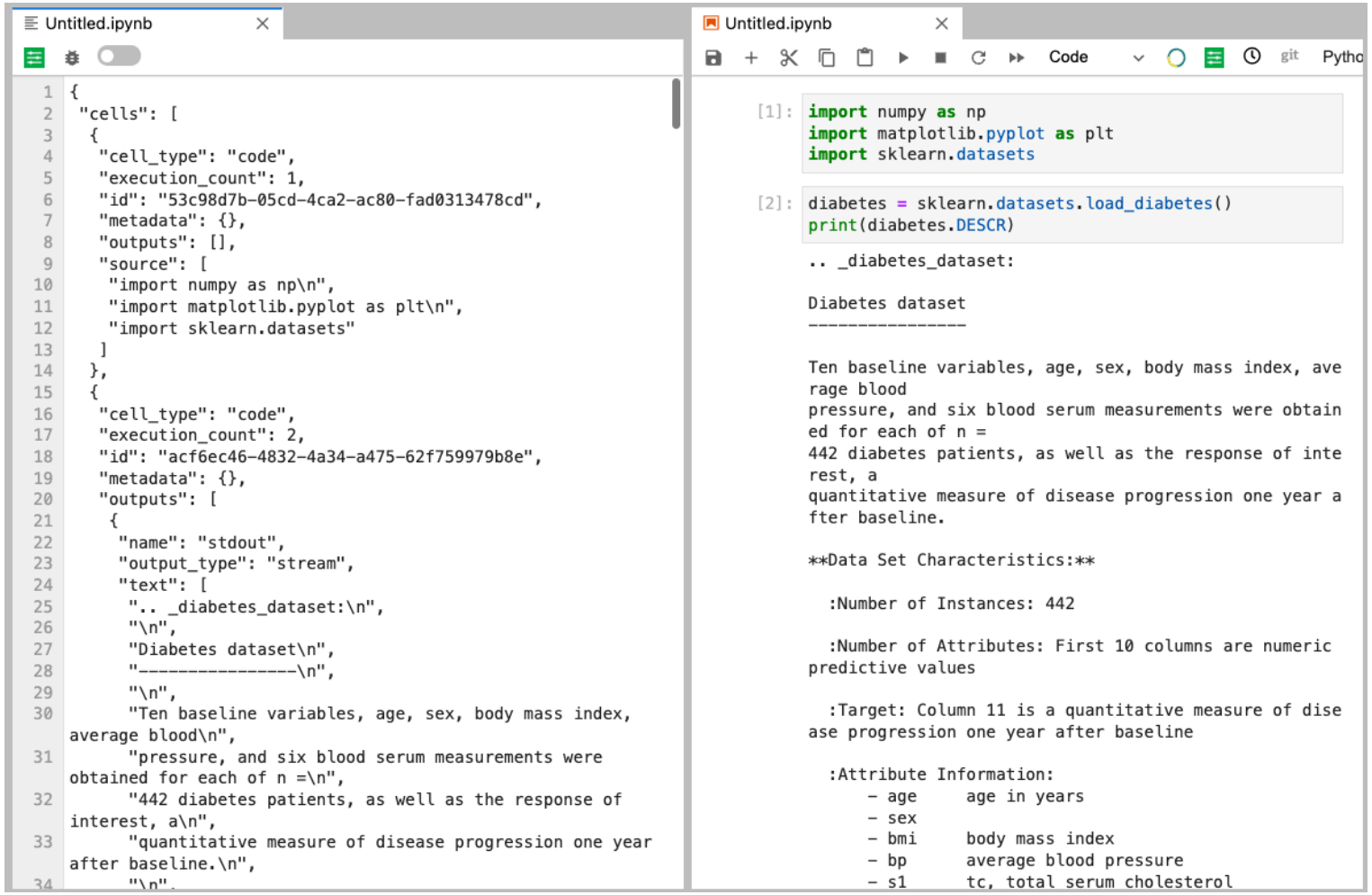
Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)
  - Markdown (*.md)
  - XML (*.xml)
    - MS Word docs (*.docx)



Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)
  - Markdown (*.md)
  - XML (*.xml)
    - MS Word docs (*.docx)
  - JSON (*.json)



Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)
  - Markdown (*.md)
  - XML (*.xml)
    - MS Word docs (*.docx)
  - JSON (*.json)
  - CSV (*.csv)

```
taken,person,quant,reading
619,dyer,rad,9.82
619,dyer,sal,0.13
622,dyer,rad,7.8
622,dyer,sal,0.09
734,pb,rad,8.41
734,lake,sal,0.05
734,pb,temp,-21.5
735,pb,rad,7.22
735,,sal,0.06
735,,temp,-26.0
751,pb,rad,4.35
751,pb,temp,-18.5
751,lake,sal,0.1
752,lake,rad,2.19
752,lake,sal,0.09
752,lake,temp,-16.0
752,roe,sal,41.6
837,lake,rad,1.46
837,lake,sal,0.21
837,roe,sal,22.5
844,roe,rad,11.25
```

Using data: format examples

- Text
  - Plain text file (*.txt)
  - HTML file (*.html)
  - Markdown (*.md)
  - XML (*.xml)
    - MS Word docs (*.docx)
  - JSON (*.json)
  - CSV (*.csv)
  - TSV (*.tsv)

taken	person	quant	reading
619	dyer	rad	9.82
619	dyer	sal	0.13
622	dyer	rad	7.8
622	dyer	sal	0.09
734	pb	rad	8.41
734	lake	sal	0.05
734	pb	temp	-21.5
735	pb	rad	7.22
735		sal	0.06
735		temp	-26.0
751	pb	rad	4.35
751	pb	temp	-18.5
751	lake	sal	0.1
752	lake	rad	2.19
752	lake	sal	0.09
752	lake	temp	-16.0
752	roe	sal	41.6
837	lake	rad	1.46
837	lake	sal	0.21
837	roe	sal	22.5
844	roe	rad	11.25

Essential Python skill: Importing data – we're going to start with basic files & Pandas

Format Type	Data Description	Reader	Writer
text	CSV	read_csv	to_csv
text	Fixed-Width Text File	read_fwf	
text	JSON	read_json	to_json
text	HTML	read_html	to_html
text	LaTeX		Styler.to_latex
text	XML	read_xml	to_xml
text	Local clipboard	read_clipboard	to_clipboard
binary	MS Excel	read_excel	to_excel
binary	OpenDocument	read_excel	
binary	HDF5 Format	read_hdf	to_hdf
binary	Feather Format	read_feather	to_feather
binary	Parquet Format	read_parquet	to_parquet
binary	ORC Format	read_orc	
binary	Stata	read_stata	to_stata
binary	SAS	read_sas	
binary	SPSS	read_spss	
binary	Python Pickle Format	read_pickle	to_pickle
SQL	SQL	read_sql	to_sql
SQL	Google BigQuery	read_gbq	to_gbq

Code time

Messy data?

	Lord of the Rings	Chronicles of Narnia	Dark Tower
Book #1	700	150	2300
Book #2	823	235	1600
Book #3	1432	176	666

This makes Python happy

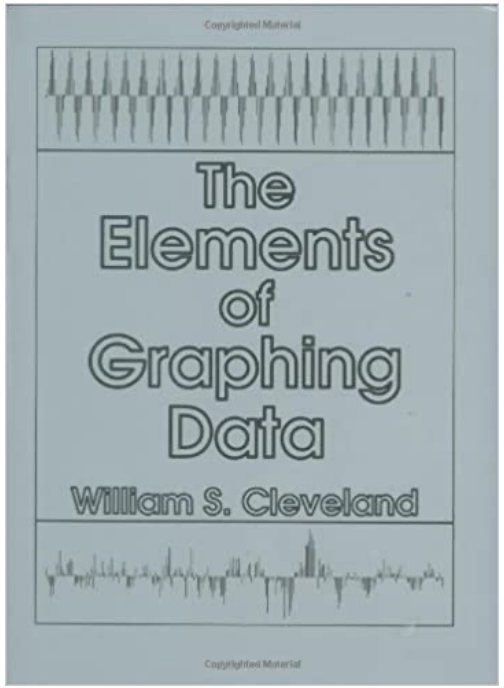
ID	Book Title	Book Number	Page Count
1	Lord of the Rings	Book #1	700
2	Lord of the Rings	Book #2	823
3	Lord of the Rings	Book #3	1432
4	Chronicles of Narnia	Book #1	150
5	Chronicles of Narnia	Book #2	235
6	Chronicles of Narnia	Book #3	176
7	Dark Tower	Book #1	2300
8	Dark Tower	Book #2	1600
9	Dark Tower	Book #3	666

Going Through the Assignments: Tableau

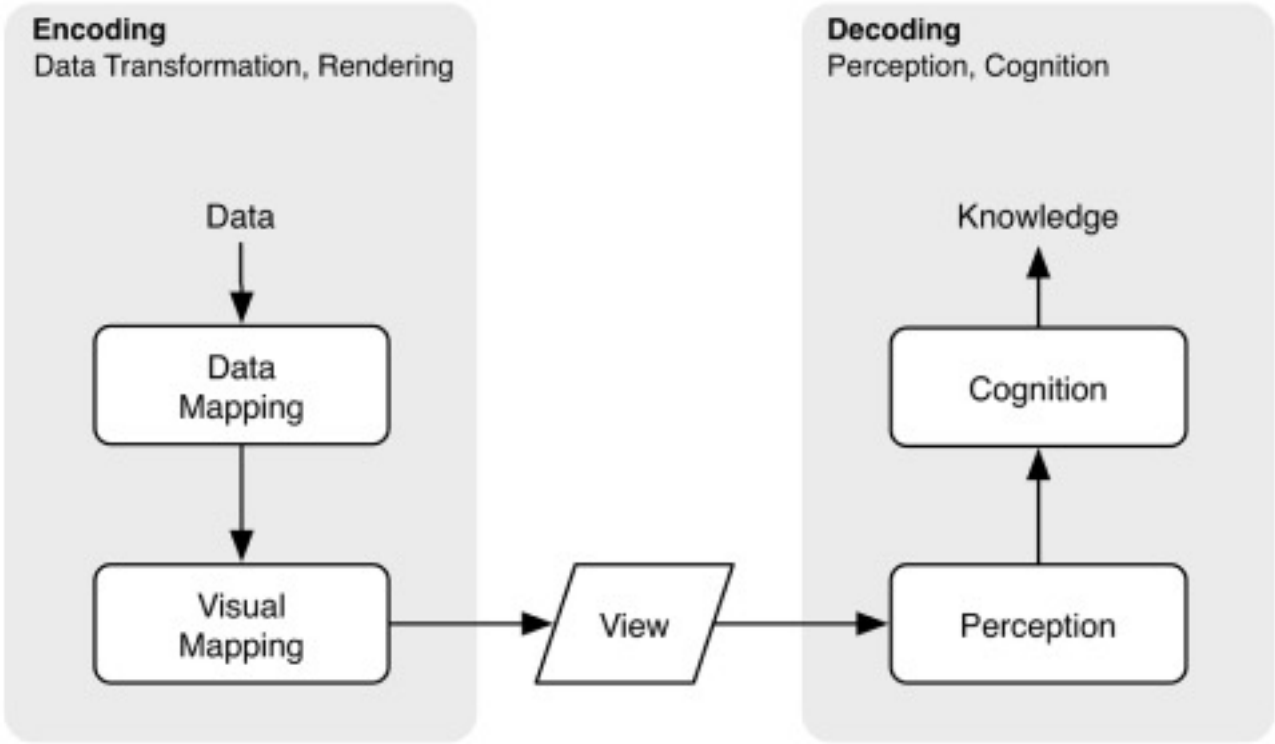
Elements of Graphing Data

Elements of Graphing Data

- The following slides all make use of points and graphics taken from Cleveland's "The Elements of Graphing Data"
- I highly recommend reading and thinking about this book



Good graphics must be visually clear and understandable



https://www.proprofsdiscuss.com/q/1693020/what-is-the-difference-between-encoding-and

Graphical Perception

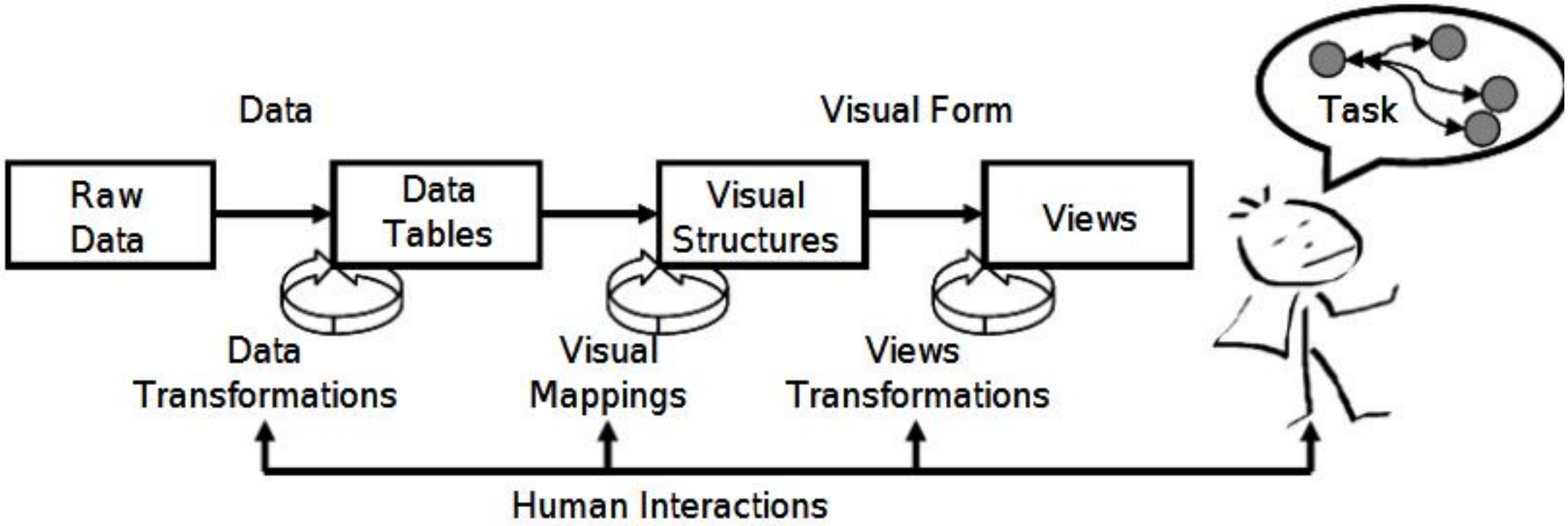
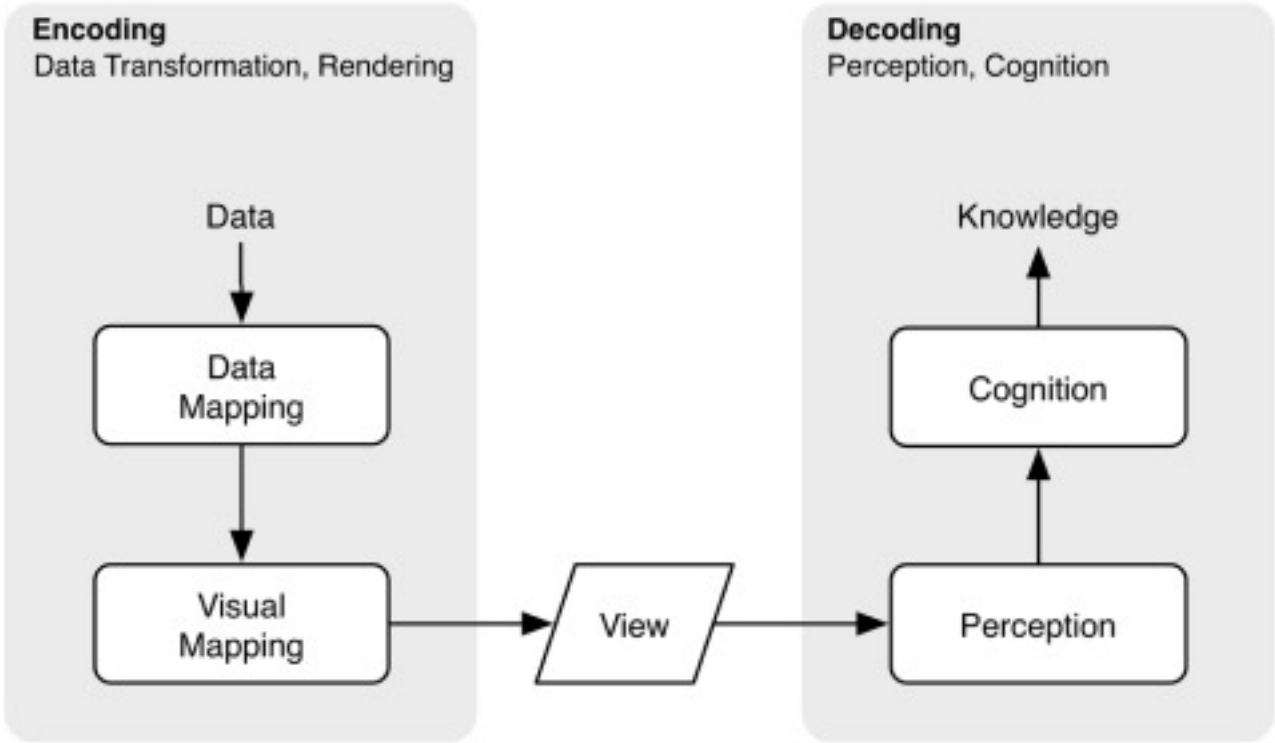


Figure 1 - Reference model for visualization (Card 1999).

https://rpubs.com/greddi/GR-DV-Perception-essay

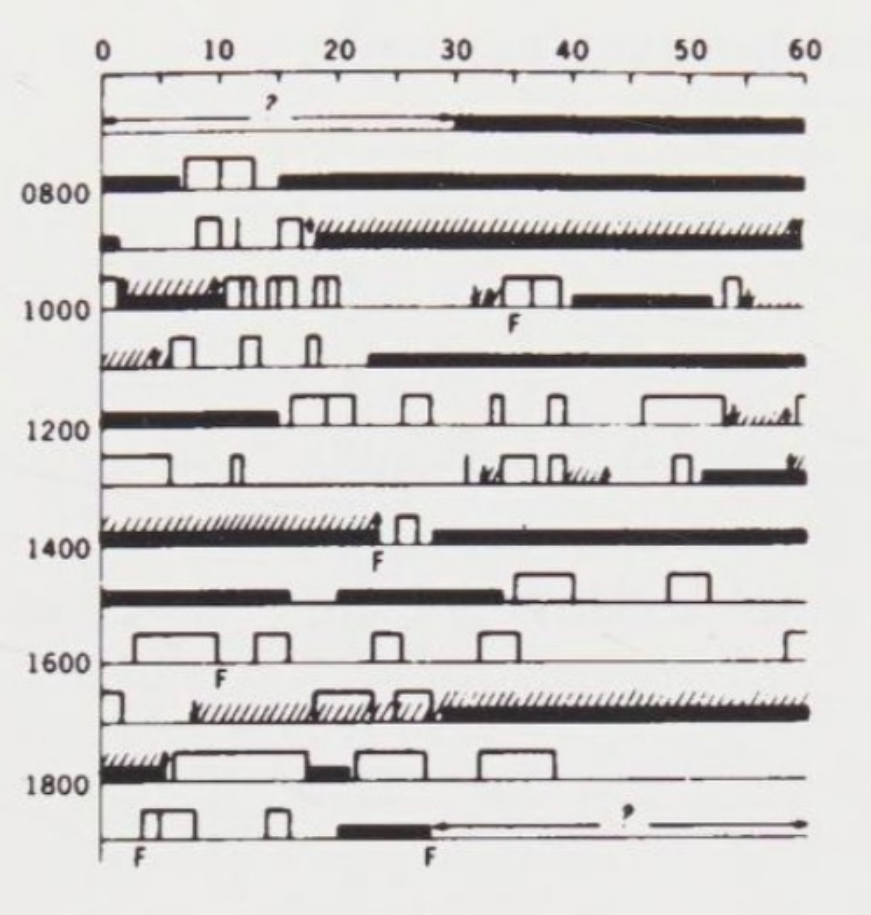
Graphical Perception

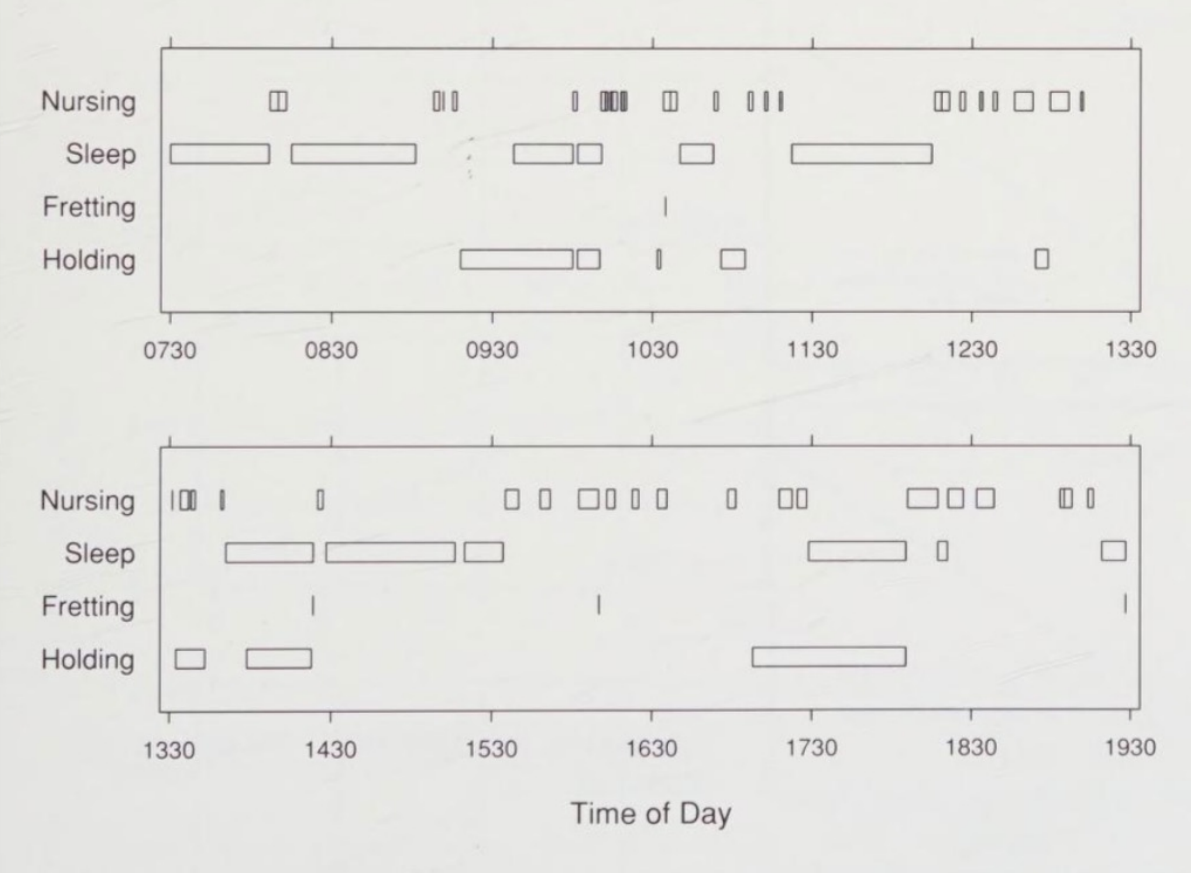


https://www.proprofsdiscuss.com/q/1693020/what-is-the-difference-between-encoding-and

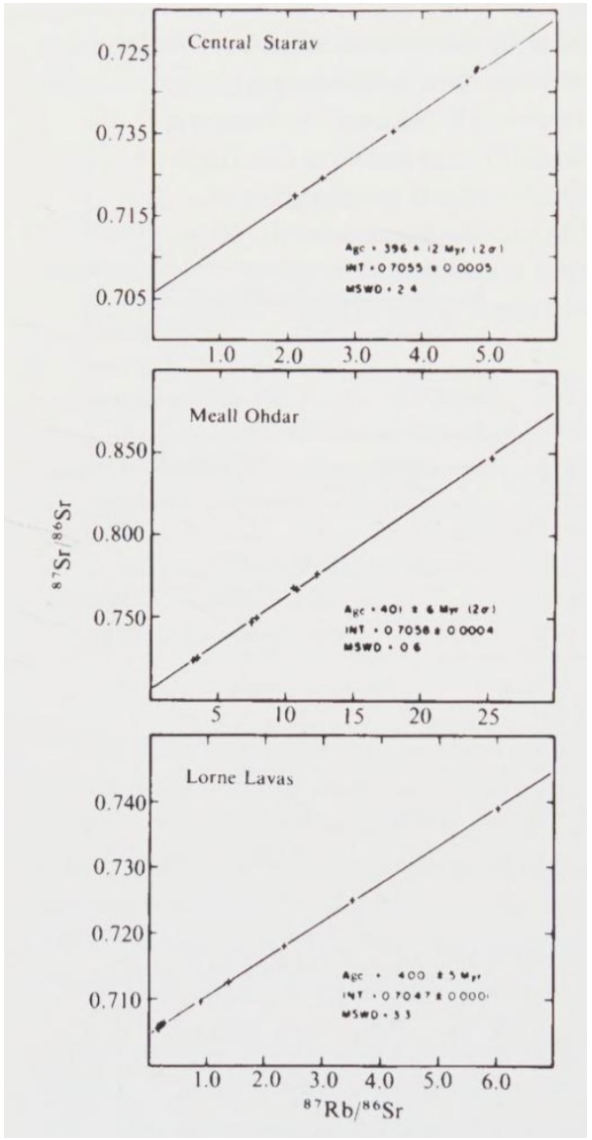
Clear Visual

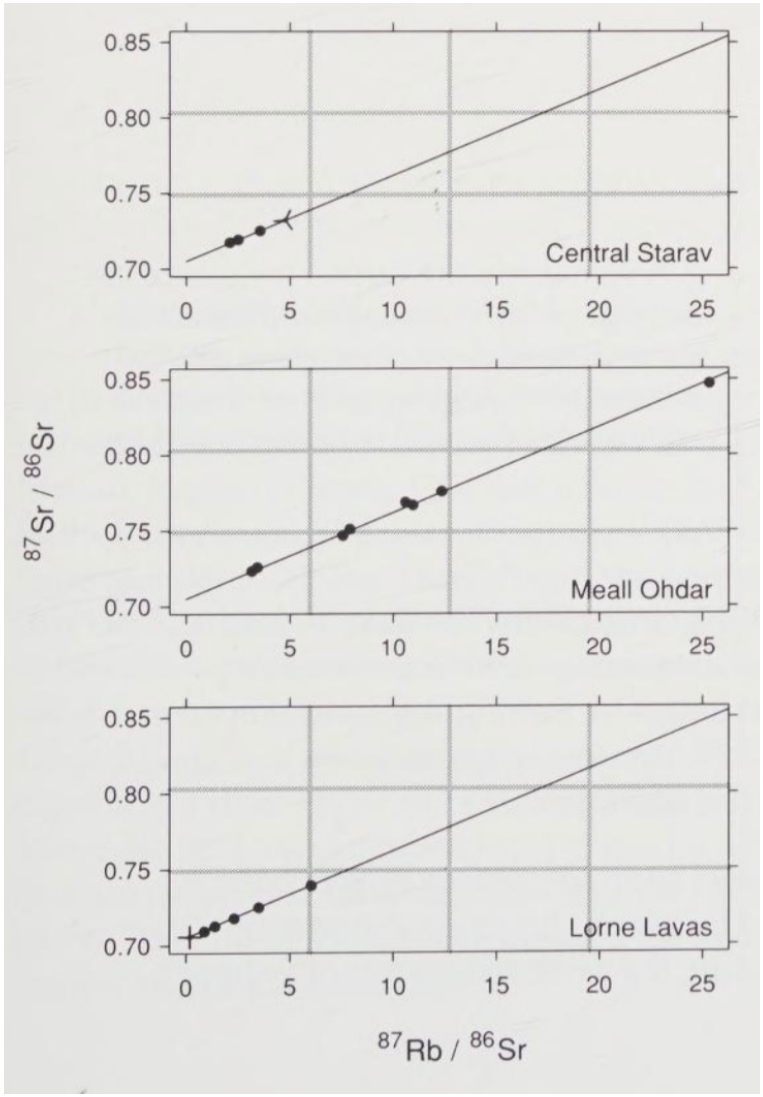
Make the data stand out; avoid superfluity



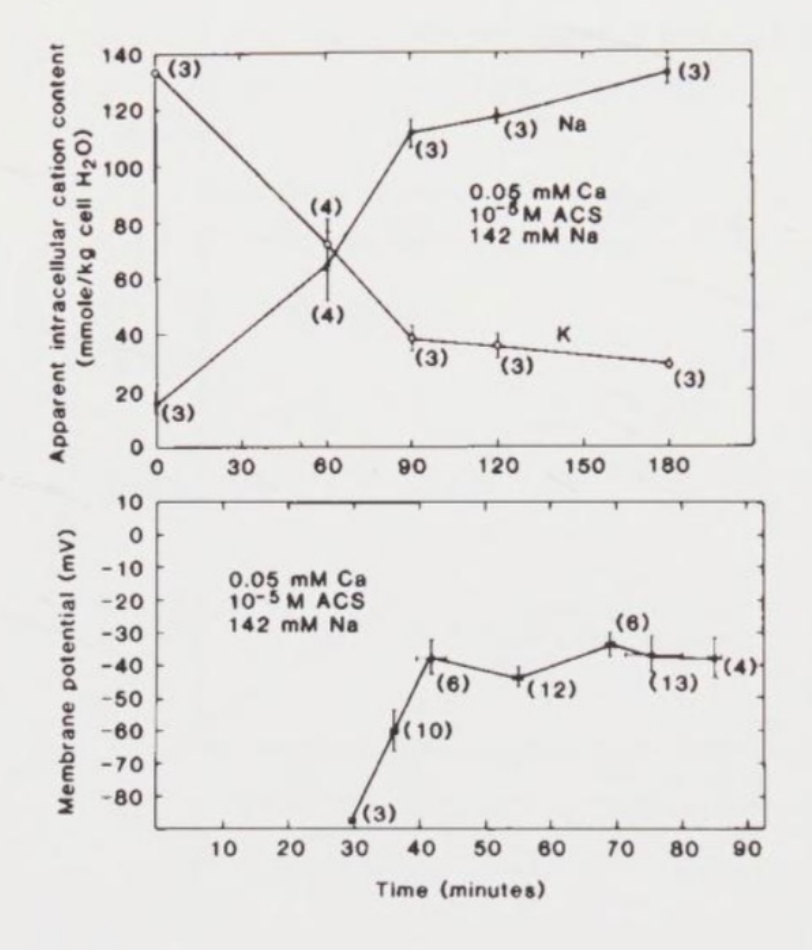


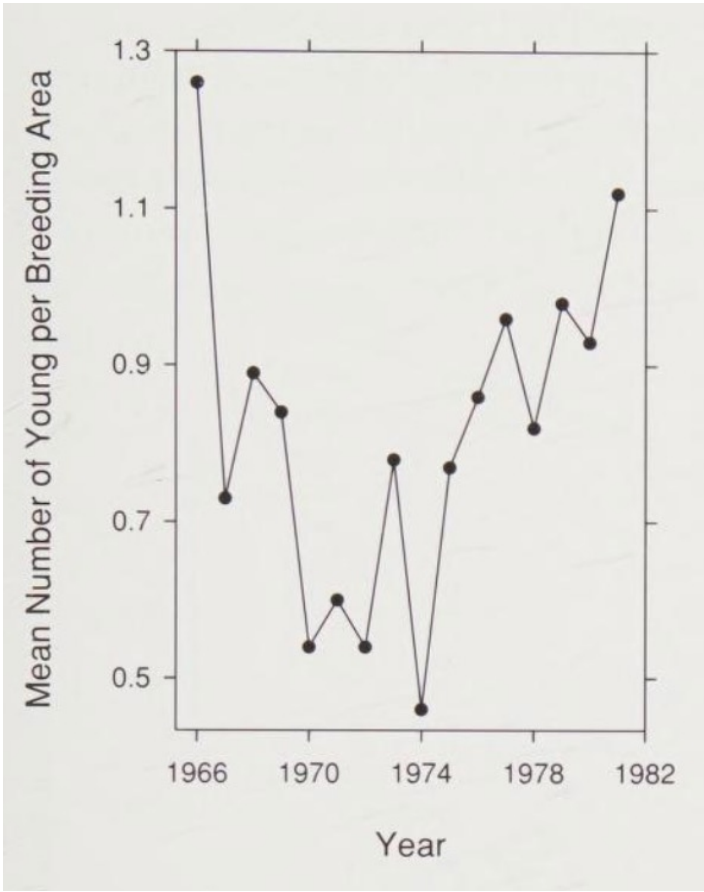
Use visually prominent graphical elements to show the data



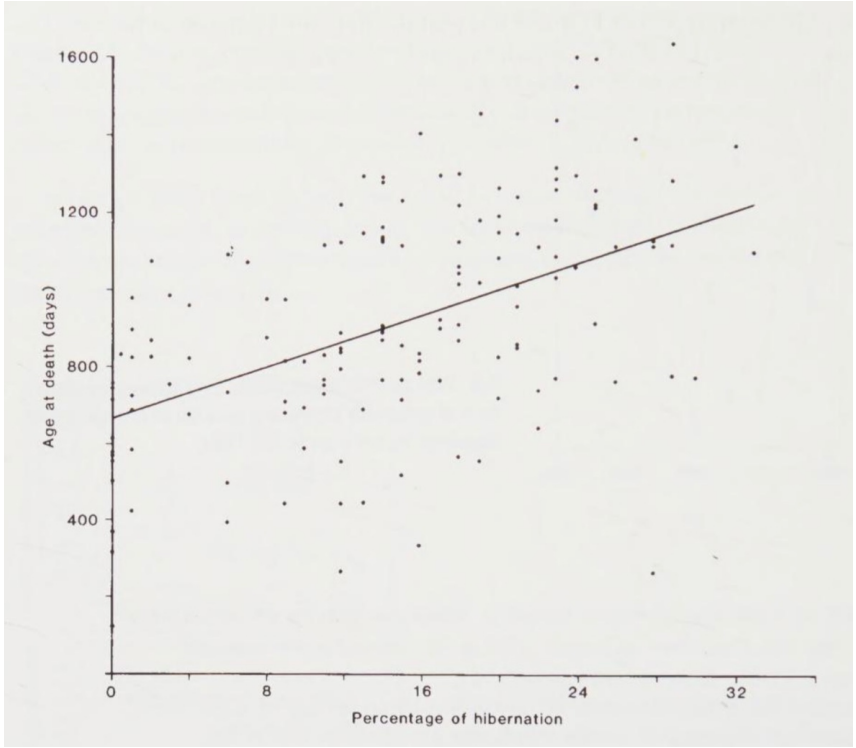


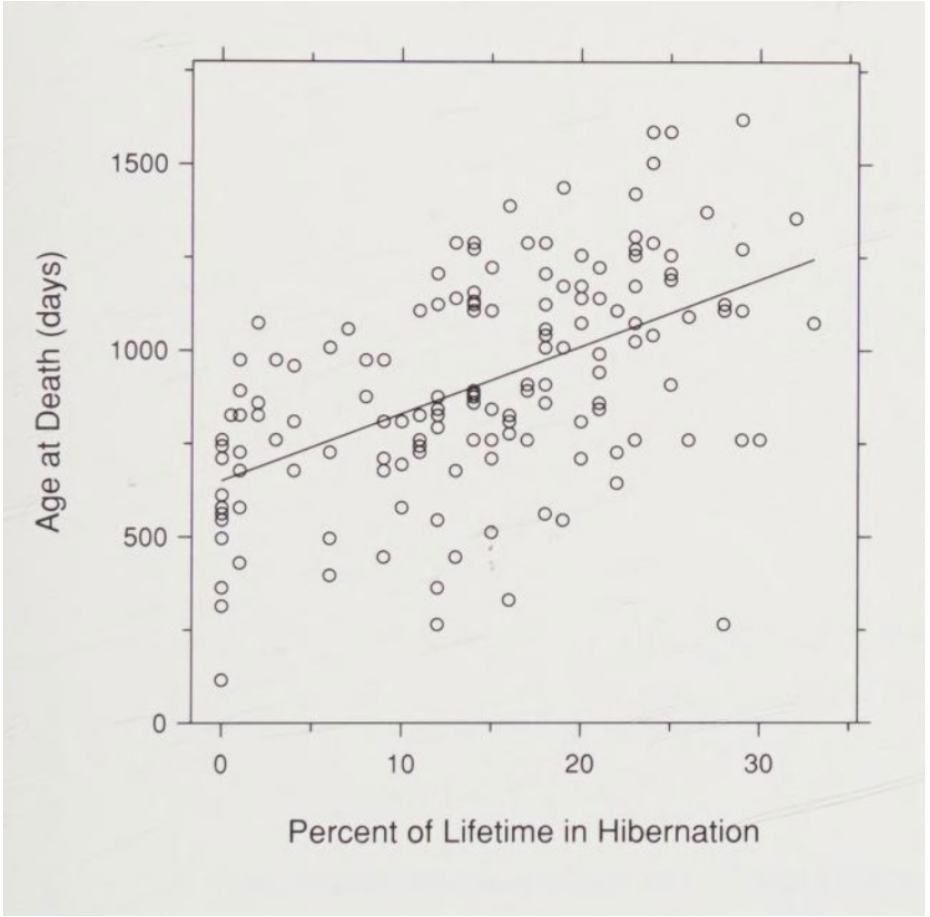
Use visually prominent graphical elements to show the data



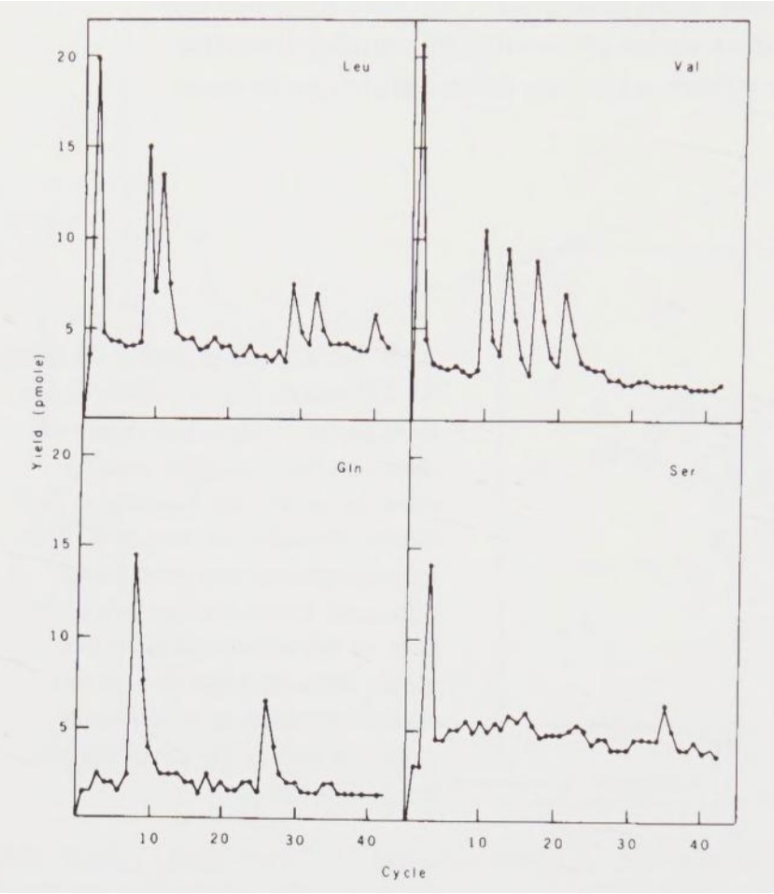


Use a pair of scale lines for each variable; make the data rectangle slightly smaller than the scale-line rectangle; tick marks should point outward

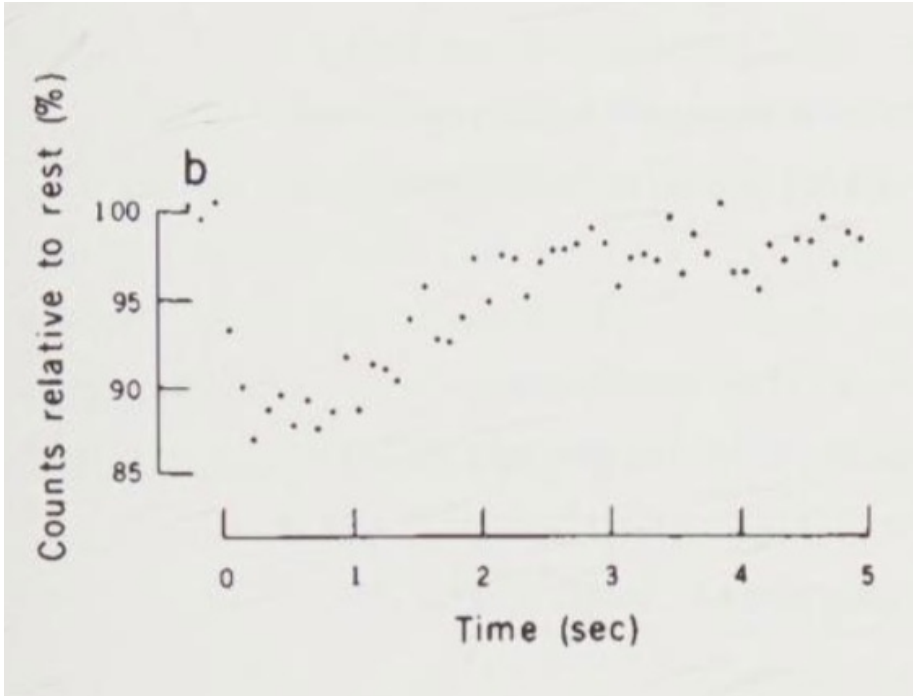


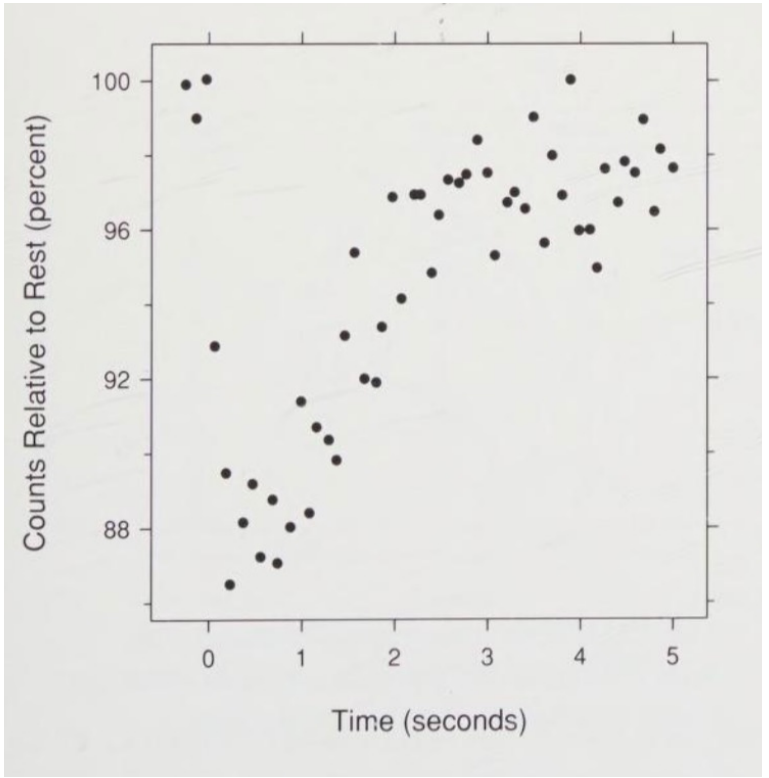


Use a pair of scale lines for each variable; make the data rectangle slightly smaller than the scale-line rectangle; tick marks should point outward

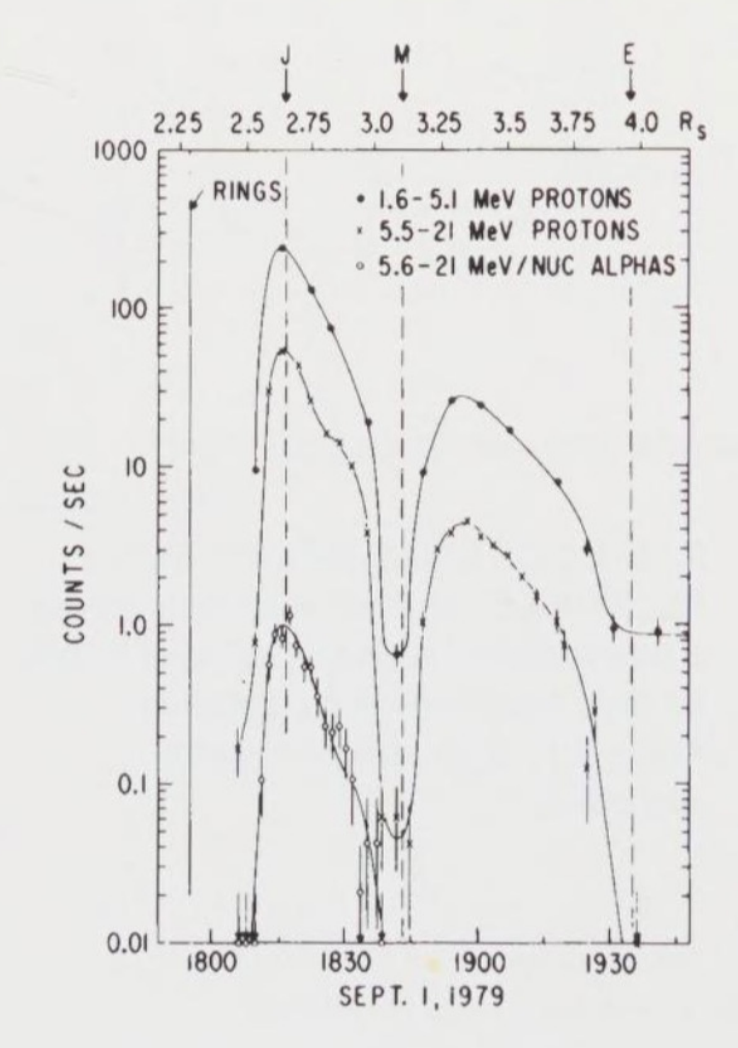


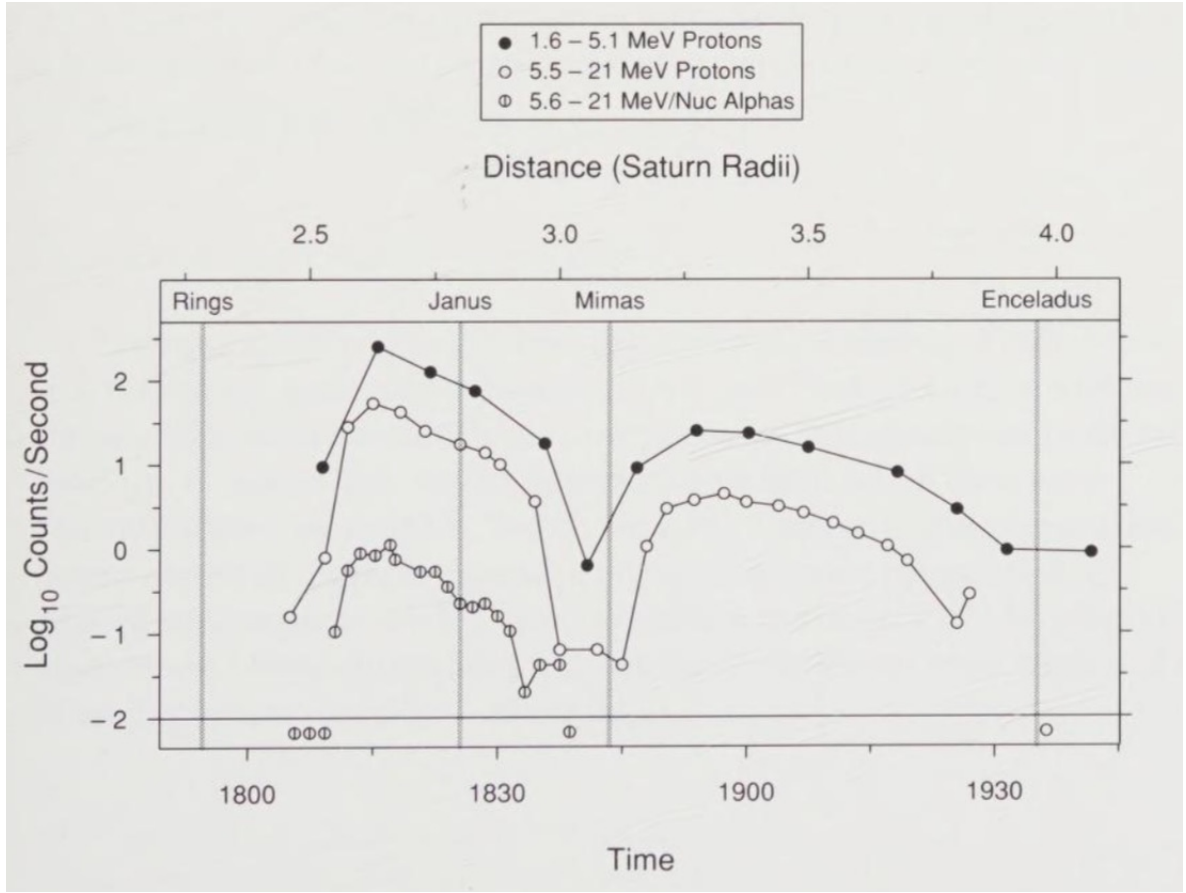
Use a pair of scale lines for each variable; make the data rectangle slightly smaller than the scale-line rectangle; tick marks should point outward



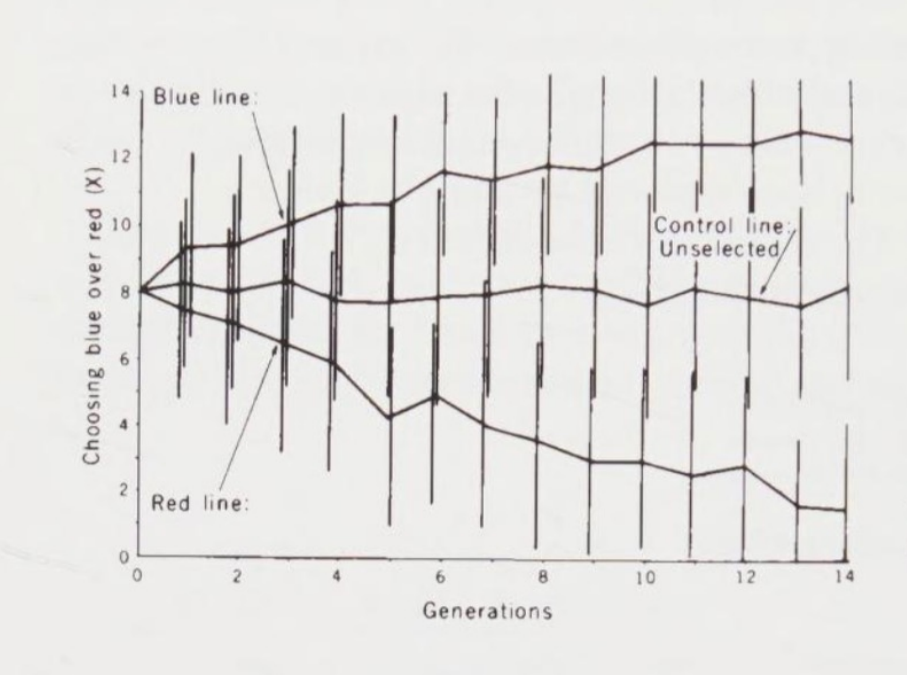


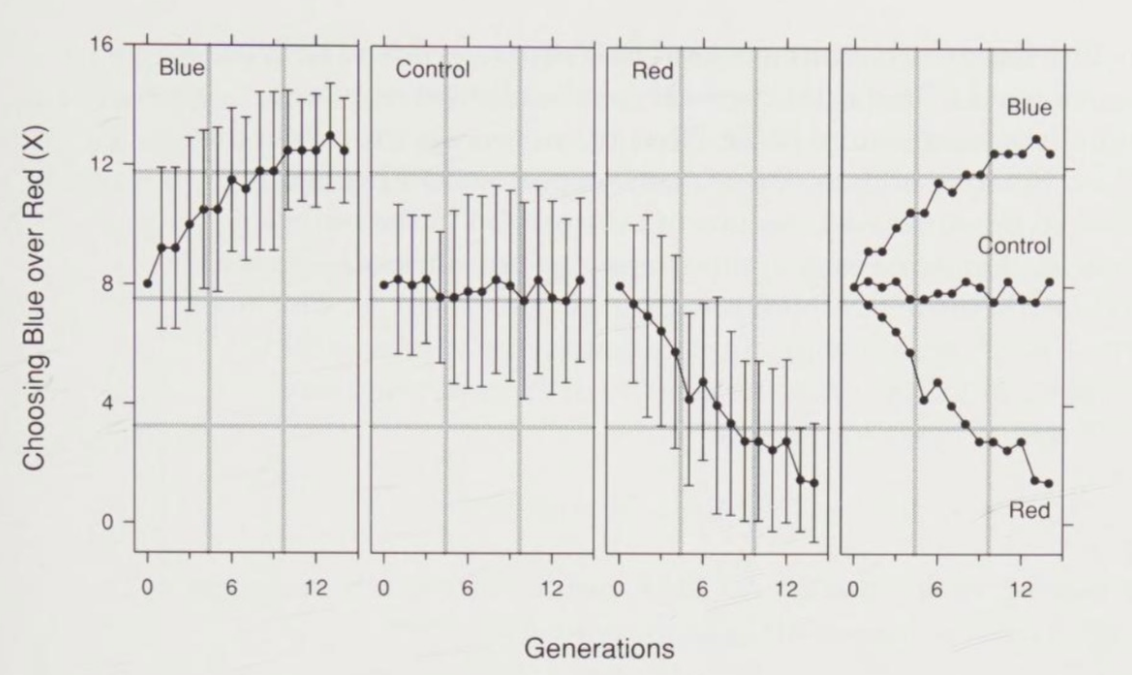
Do not clutter the interior of the scale-line rectangle



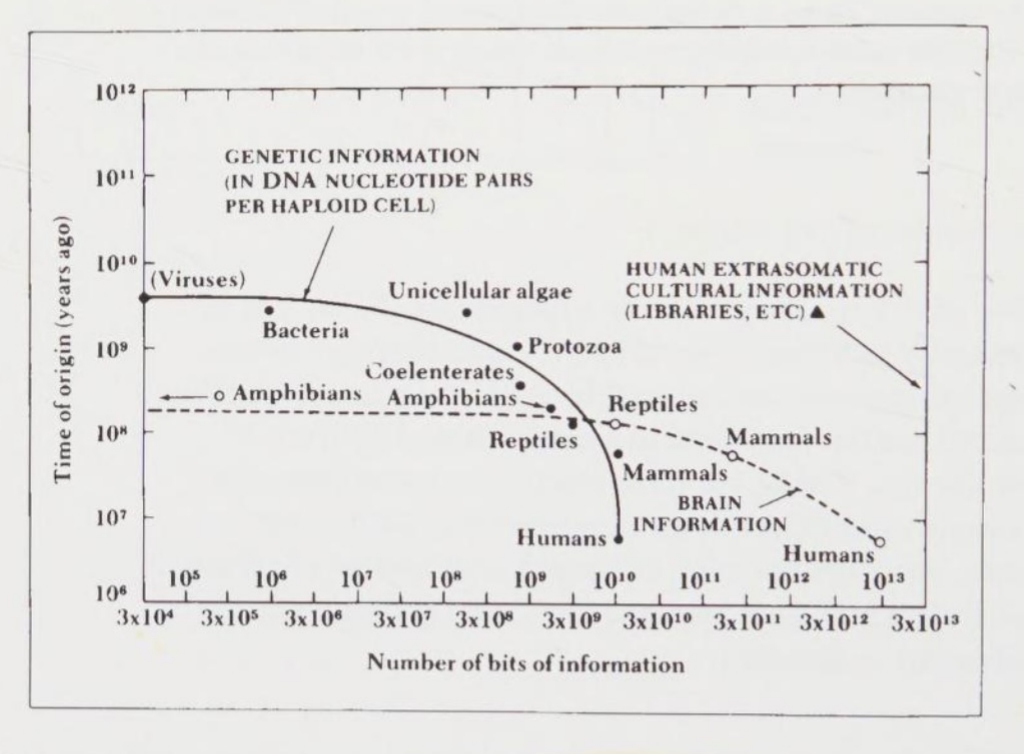


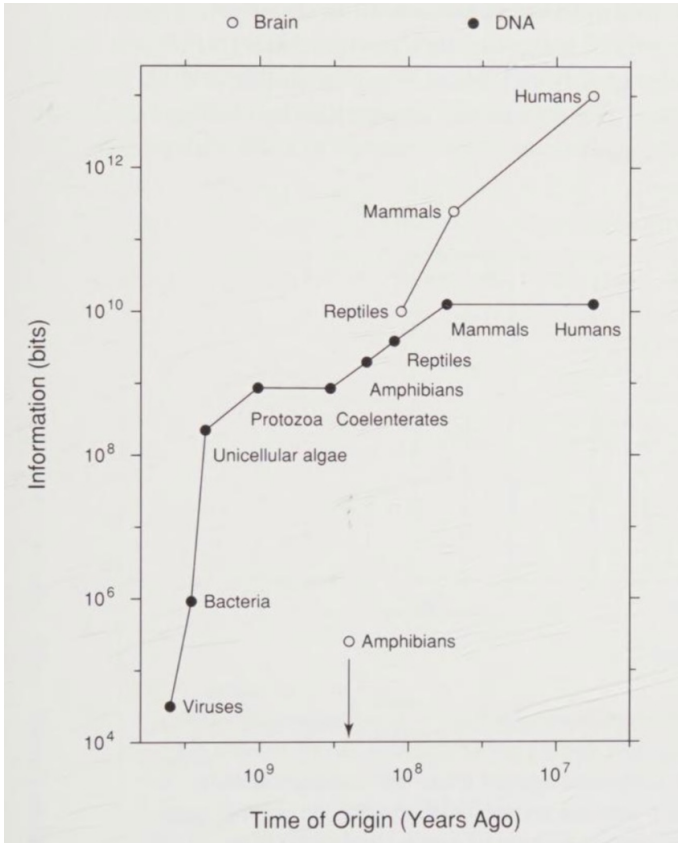
Do not clutter the interior of the scale-line rectangle



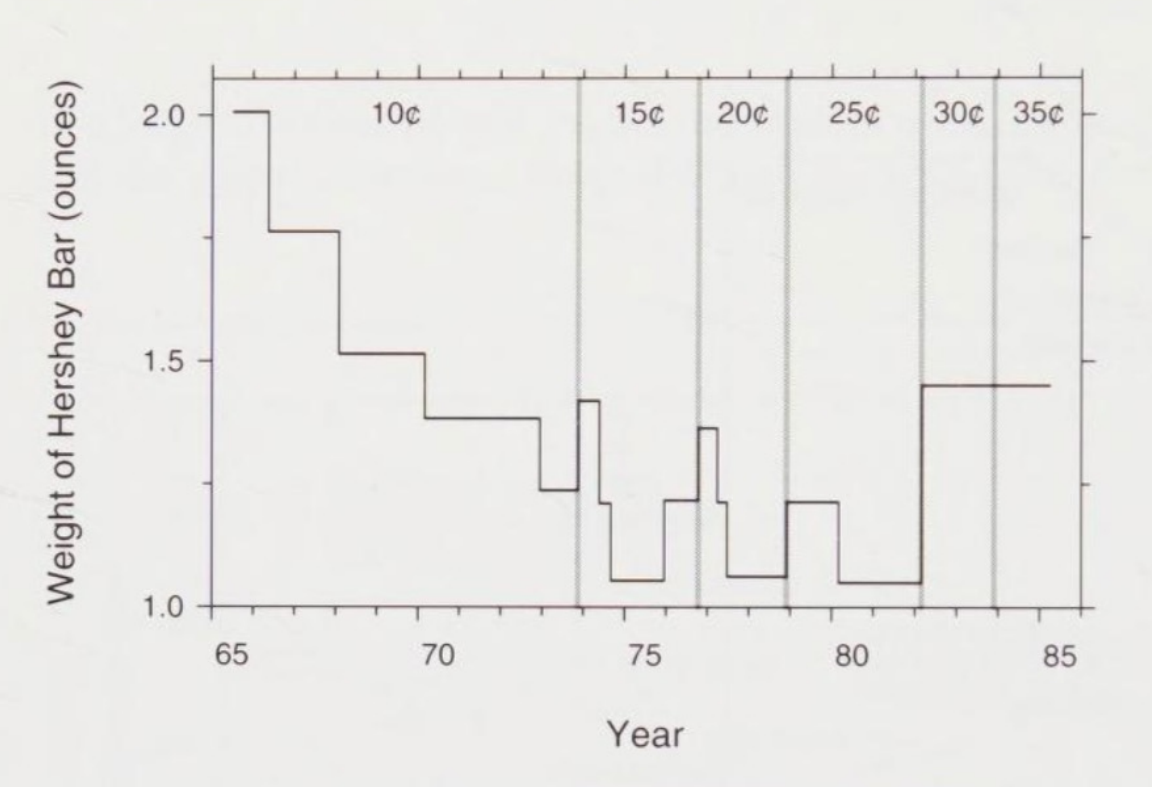


Do not overdo the number of tick marks

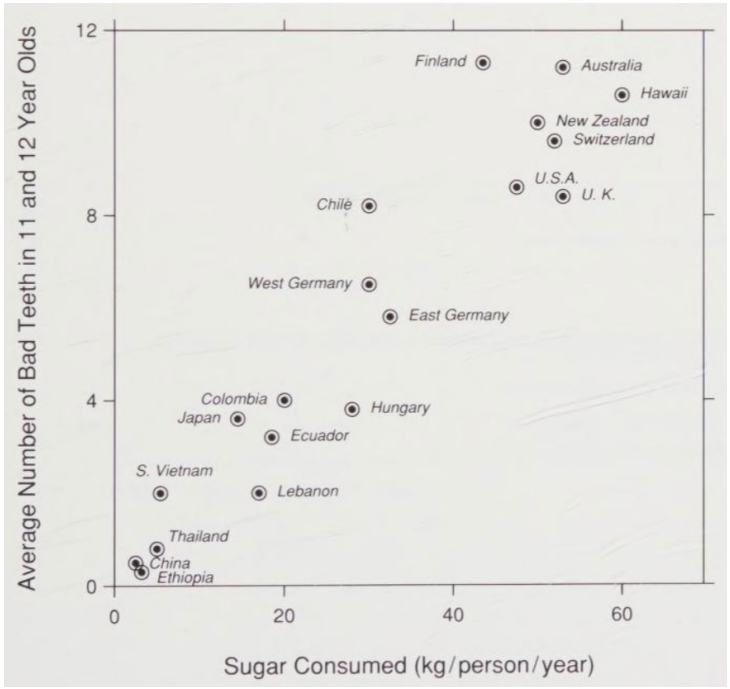


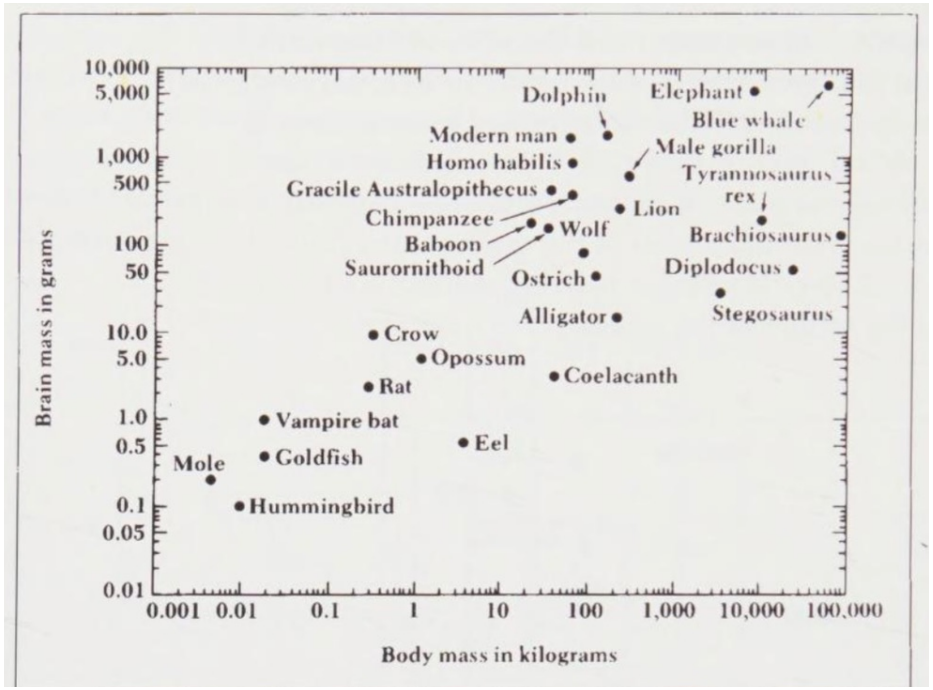


Use a reference line when there is an important value that must be seen across the entire graph, but do not let the line interfere with the graph



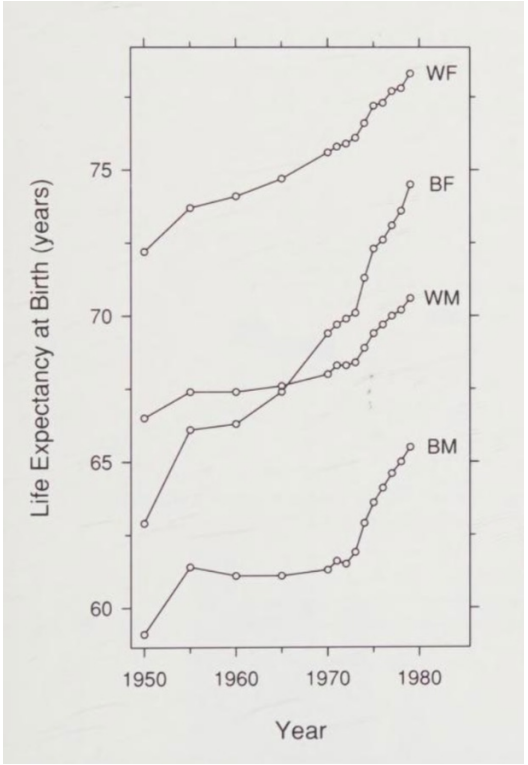
Do not allow labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph

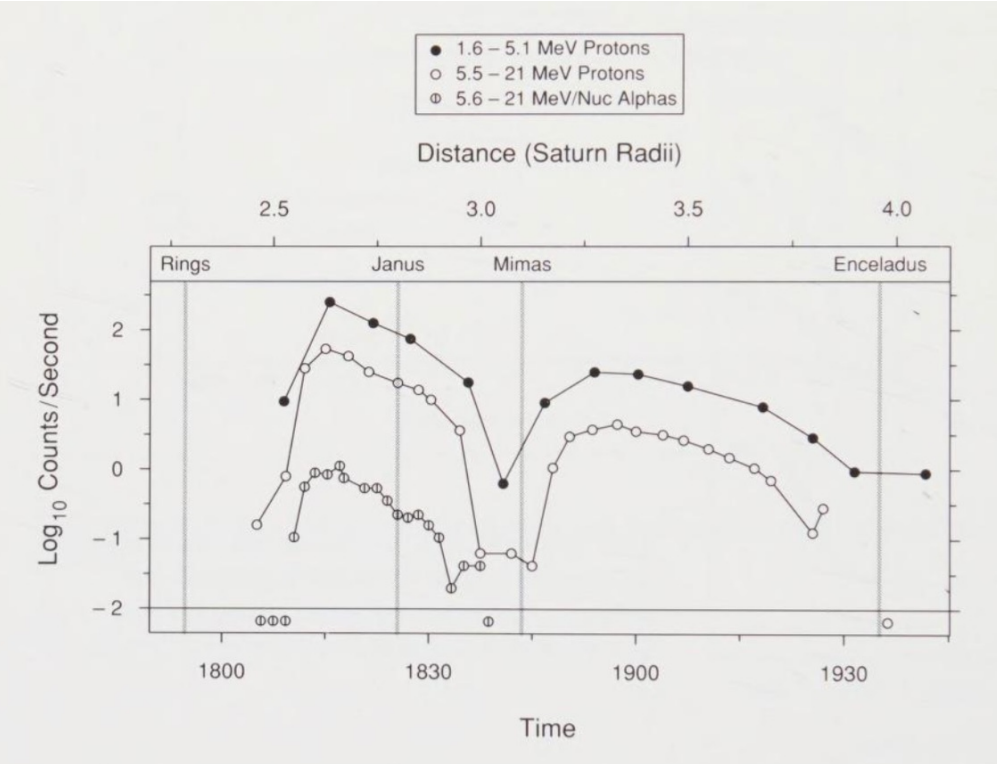




- External key makes identification harder but it reduces clutter

Do not allow labels in the interior of the scale-line rectangle to interfere with the quantitative data or to clutter the graph





- External key makes identification harder but it reduces clutter

Avoid putting notes and keys inside the scale-line rectangle; put a key outside, and put notes in the caption or in the text

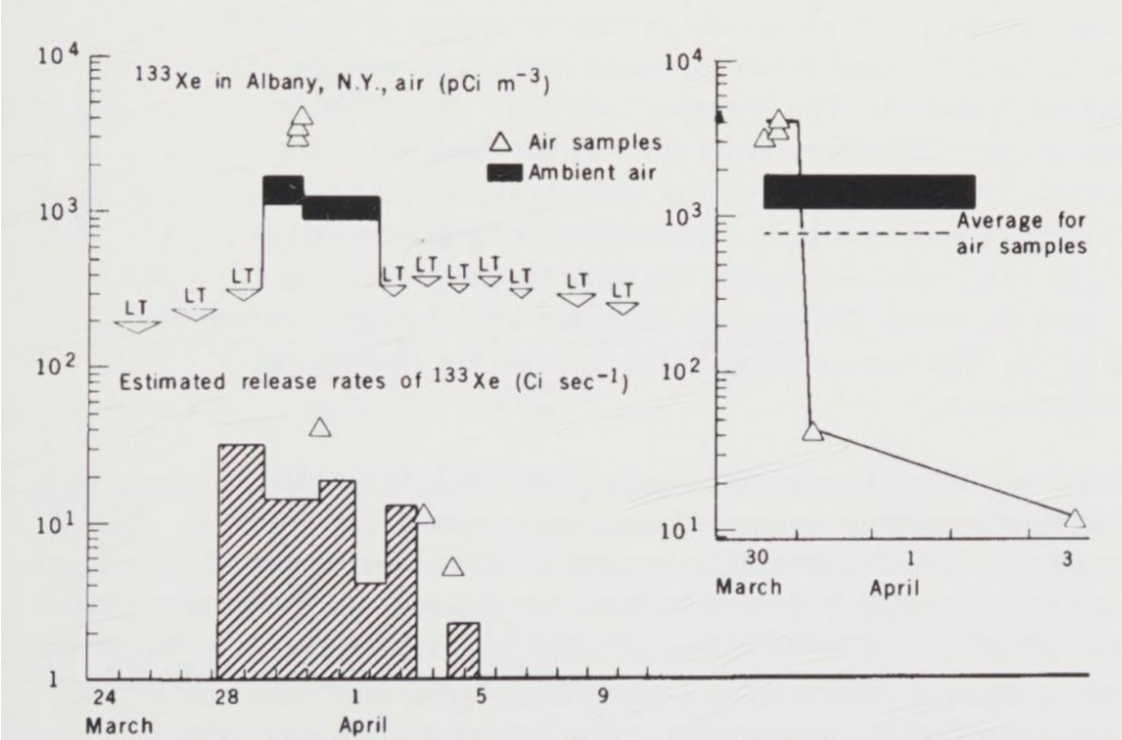
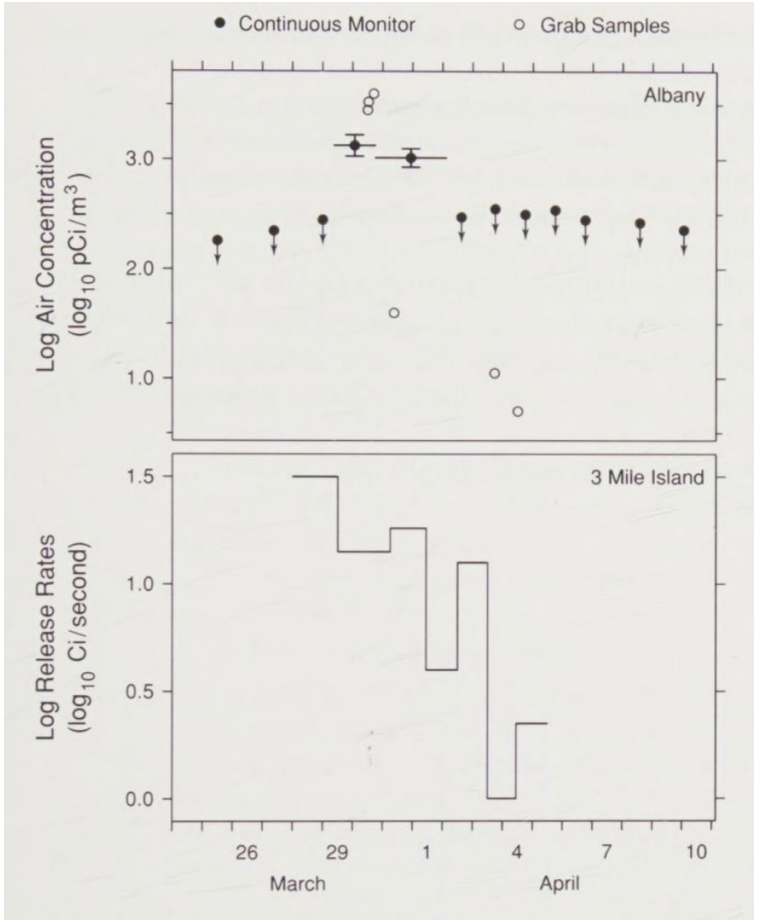
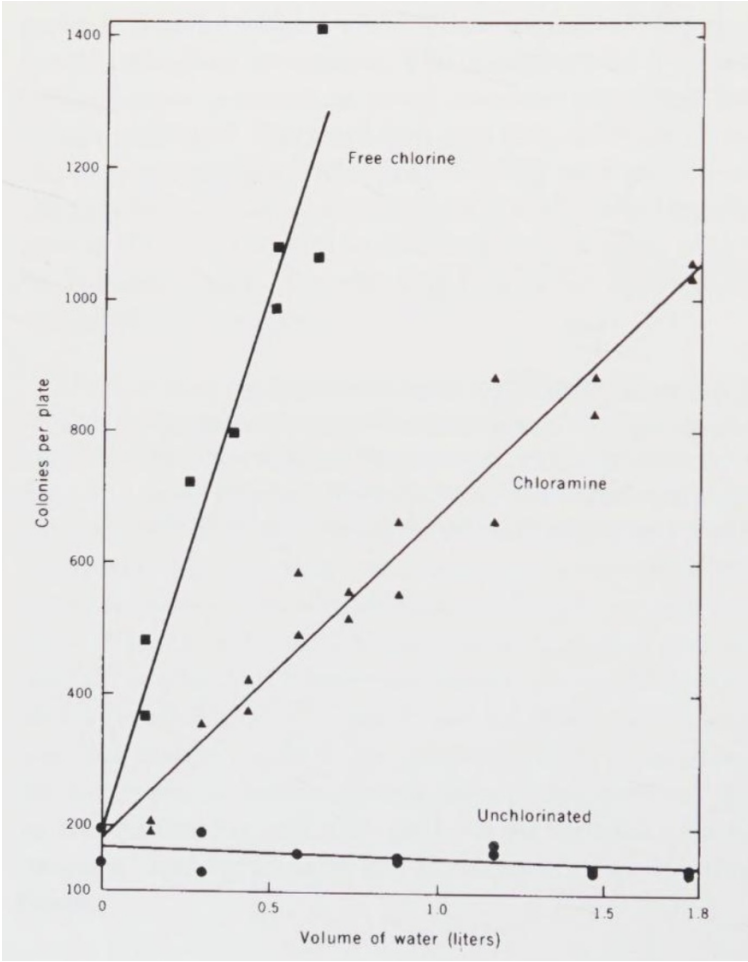


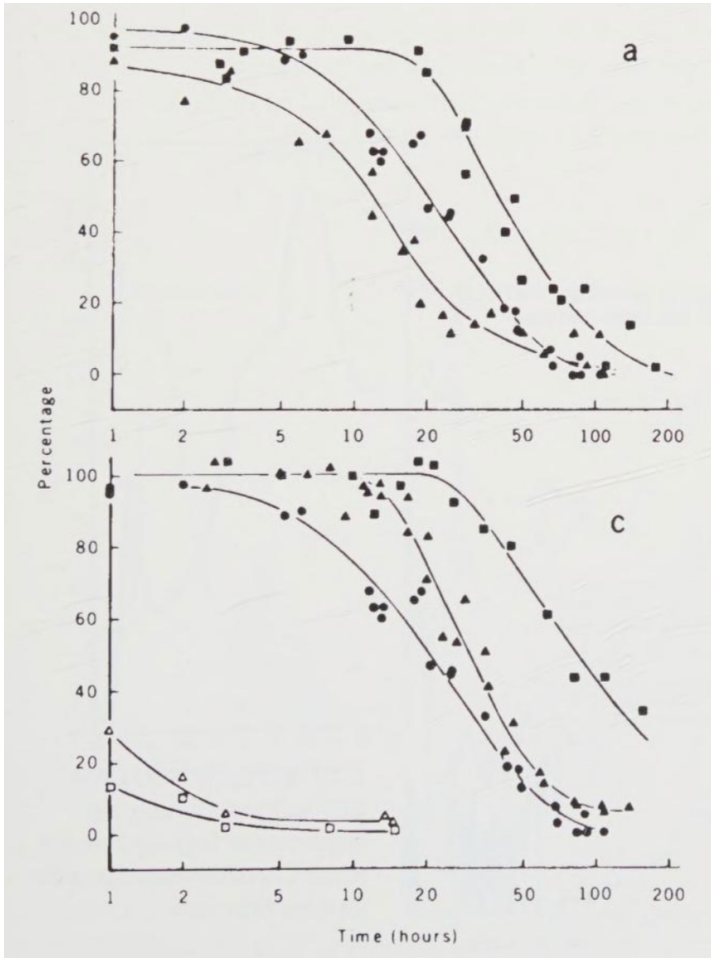
Fig. 1. Xenon-133 activity (picocuries per cubic meter of air) in Albany, New York, for the end of March and early April 1979. The lower trace shows the time-averaged estimates of releases (curies per second) from the Three Mile Island reactor (*2*). The inset shows detailed values for air samples (gas counting) and concurrent average values for ambient air (Ge diode). Abbreviation: *LT*, less than.

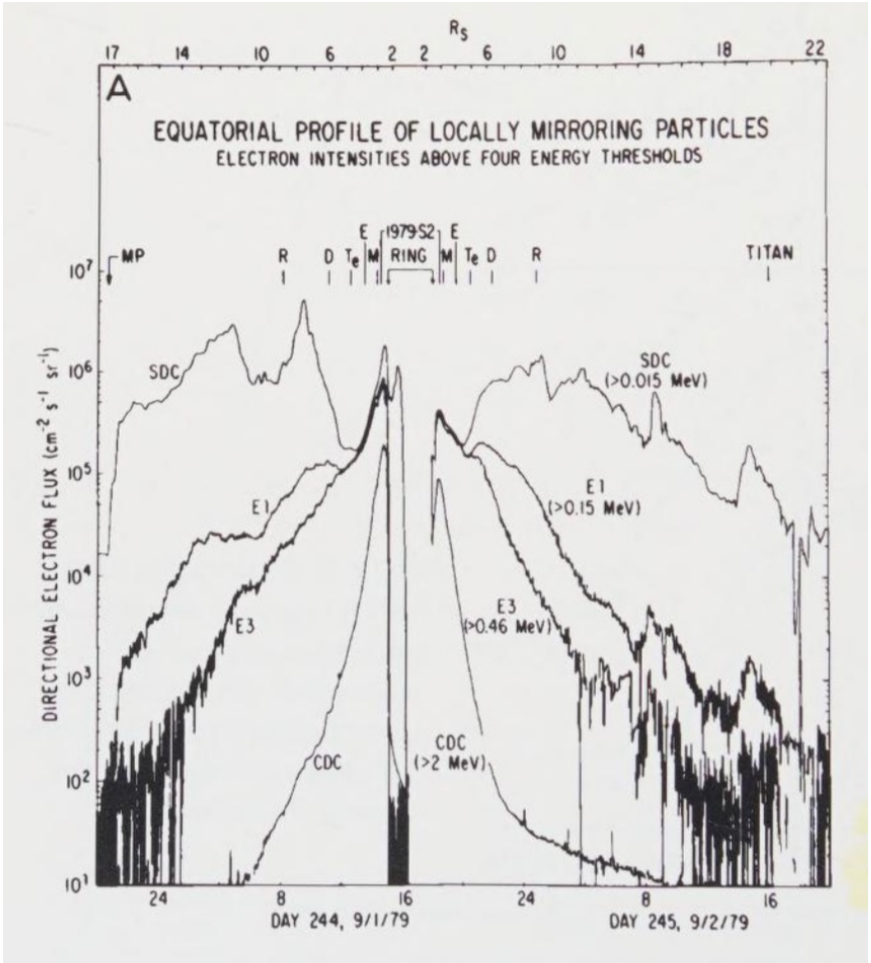


Overlapping plotting symbols must be visually distinguishable



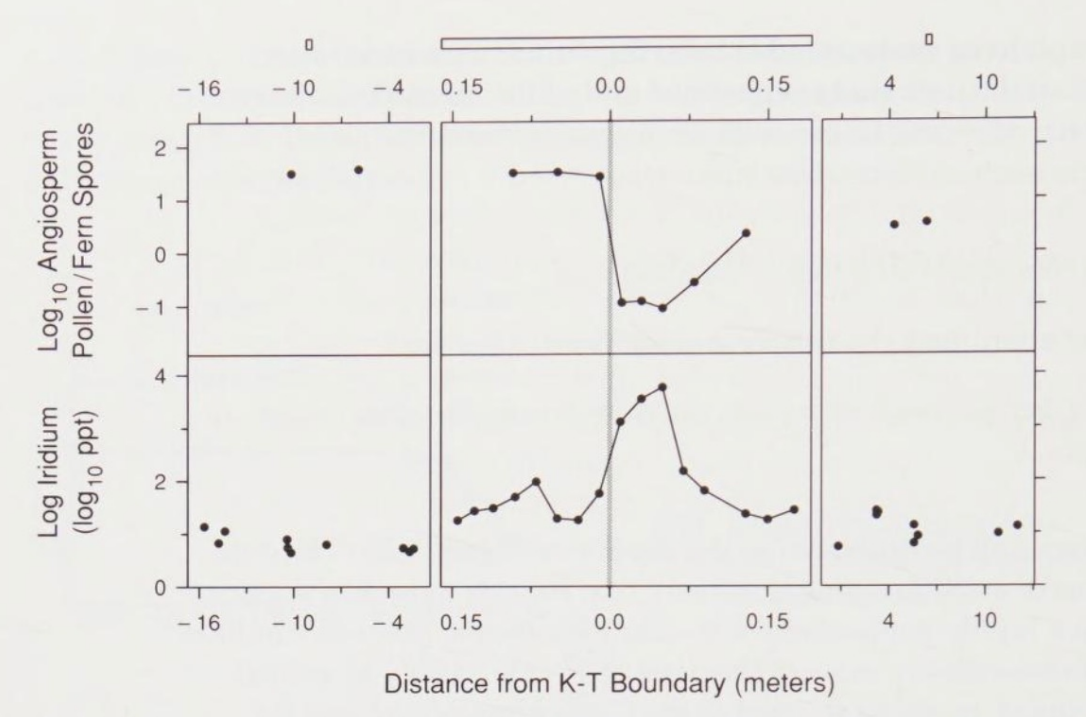
Superposed data sets must be readily visually assembled





Clear Understanding

Put major conclusions into graphical form; make captions comprehensive and informative



- Describe everything that is graphed
- Draw attention to the important features of the data
- Describe the conclusions that are drawn from the data on the graph

”**Angiosperm-Fern Ratio and Iridium Near the K-T Boundary.** The graph shows measurements of a core from northeastern New Mexico. The horizontal scale is in meters from the boundary between the Cretaceous and the Tertiary periods; negative values are below the K-T boundary so time goes from earlier to later in going from left to right. The widths of the three rectangles at the top of the graph show the same number of meters on the horizontal scales of the three panels. The top panel shows the ratio of angiosperm pollen to fern spores; the K-T boundary is taken to be the time point at which these values begin to decrease. The bottom panel shows concentrations of iridium; the concentrations begin a dramatic rise and fall at the boundary. Since the principal source of iridium is extraterrestrial, its rise and fall supports the hypothesis that an asteroid struck the earth causing a cloud of dust in the upper atmosphere; this is argued to have caused the large number of extinctions, including the dinosaurs, that occurred at the beginning of the Tertiary period.”

If you have error bars, make sure you explain

- What “error” are they?
  - Sample standard deviation
  - Standard error (estimate of the standard deviation)
  - Confidence interval

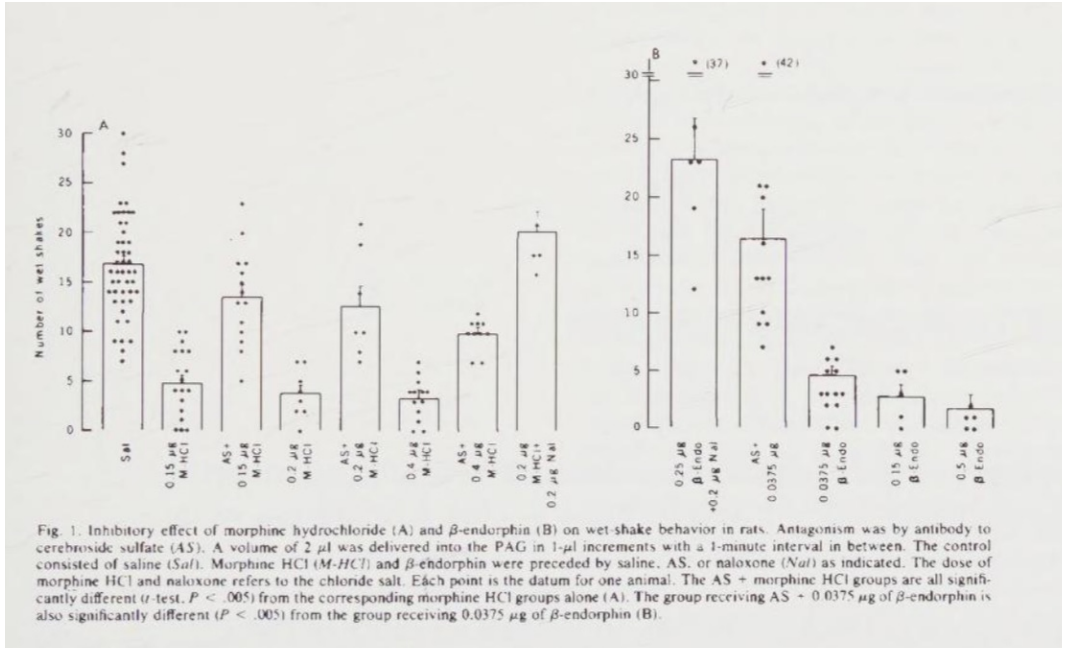
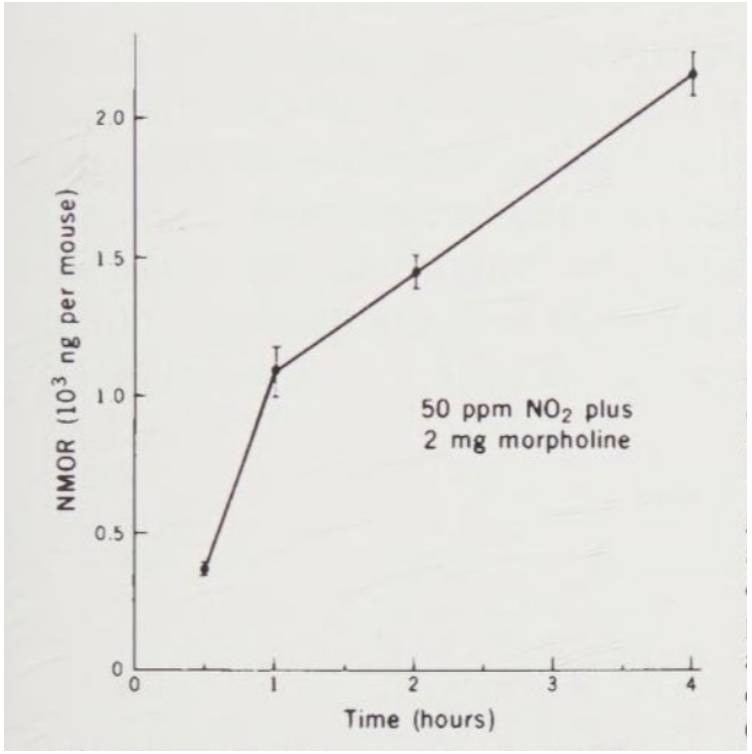
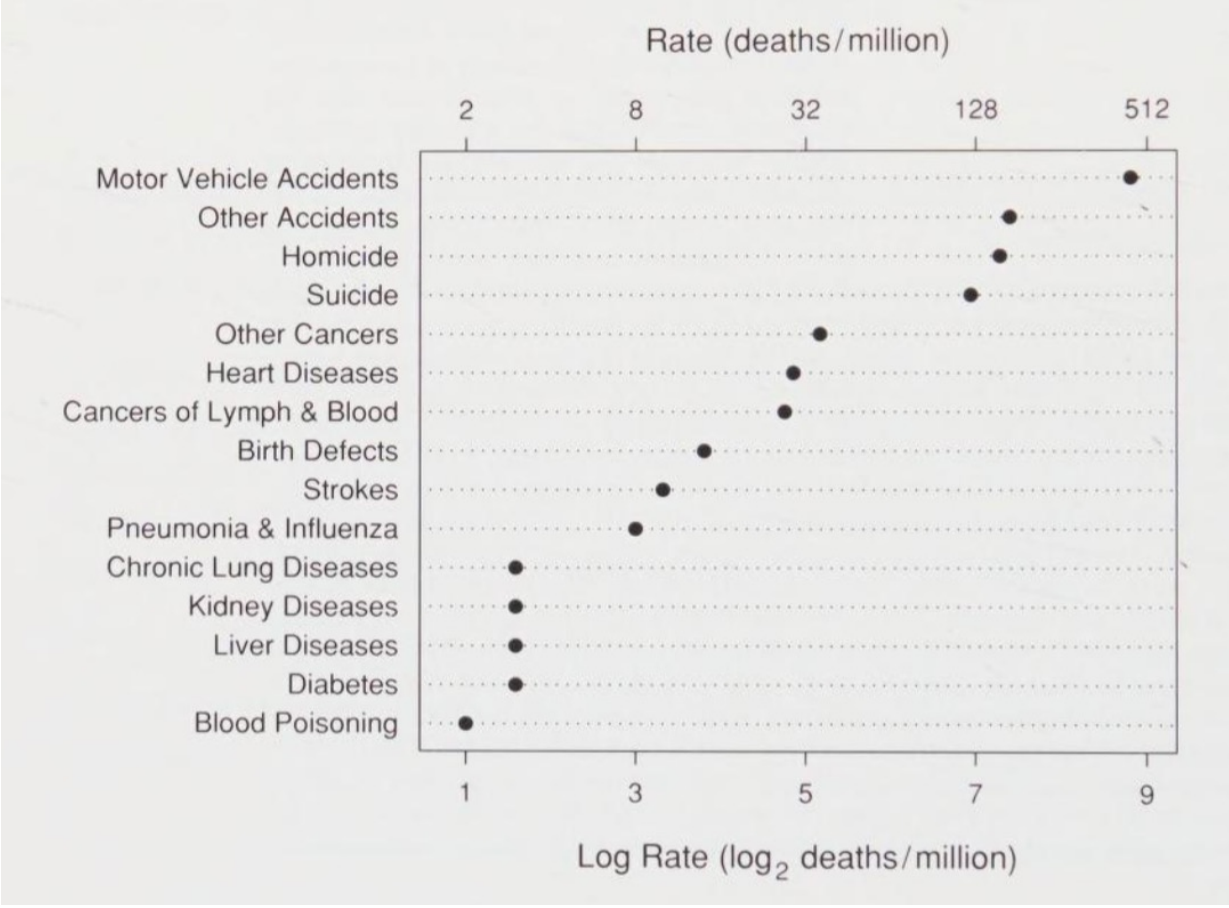


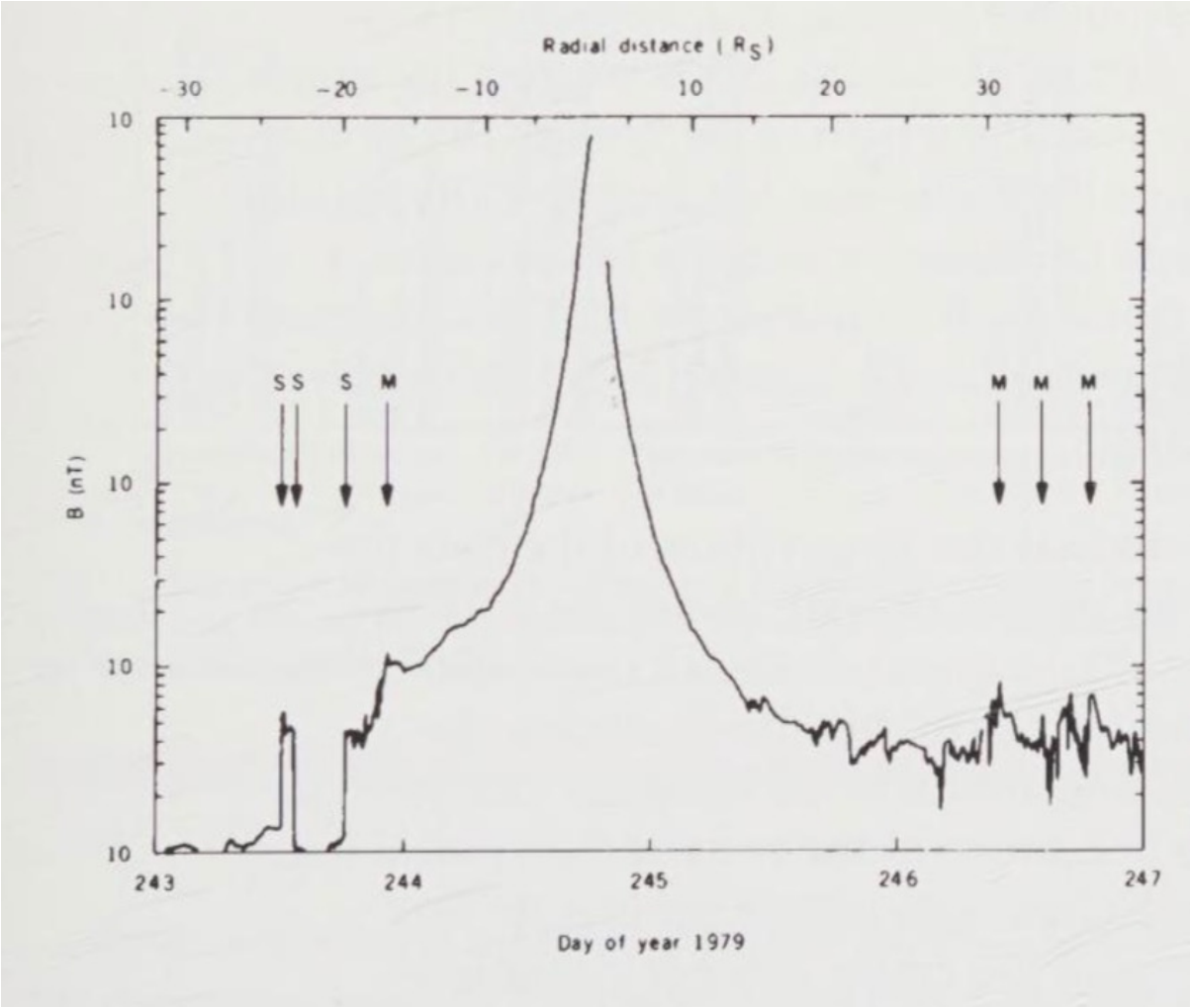
Fig. 1. Inhibitory effect of morphine hydrochloride (A) and β-endorphin (B) on wet-shake behavior in rats. Antagonism was by antibody to cerebroside sulfate (*AS*). A volume of 2 μl was delivered into the PAG in 1-μl increments with a 1-minute interval in between. The control consisted of saline (*Sal*). Morphine HCl (*M-HCl*) and β-endorphin were preceded by saline, AS, or naloxone (*Nal*) as indicated. The dose of morphine HCl and naloxone refers to the chloride salt. Each point is the datum for one animal. The AS + morphine HCl groups are all significantly different (*t*-test, $P < .005$) from the corresponding morphine HCl groups alone (A). The group receiving AS + 0.0375 μg of β-endorphin is also significantly different ($P < .005$) from the group receiving 0.0375 μg of β-endorphin (B).



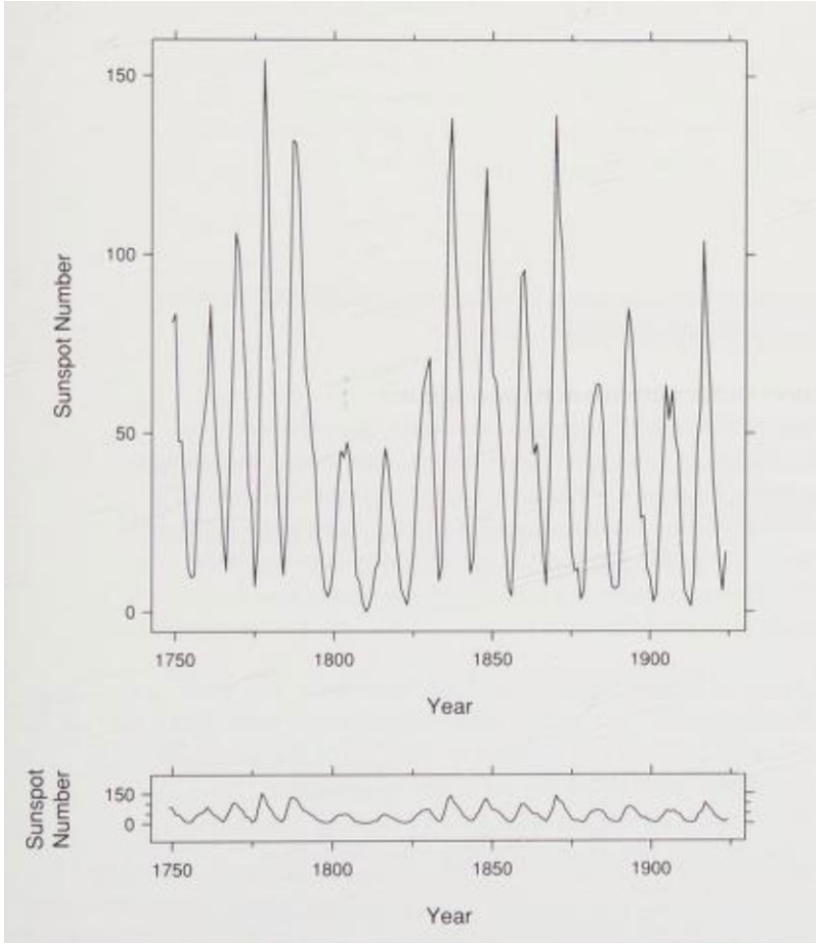
When logarithms of a variable are graphed, the scale label should correspond to the tick mark labels



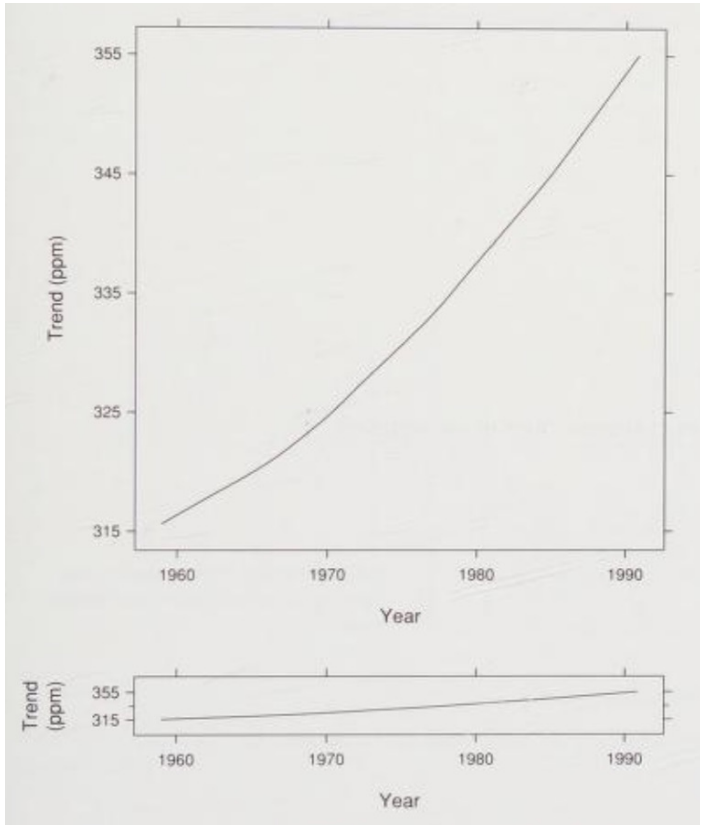
Proofread graphs

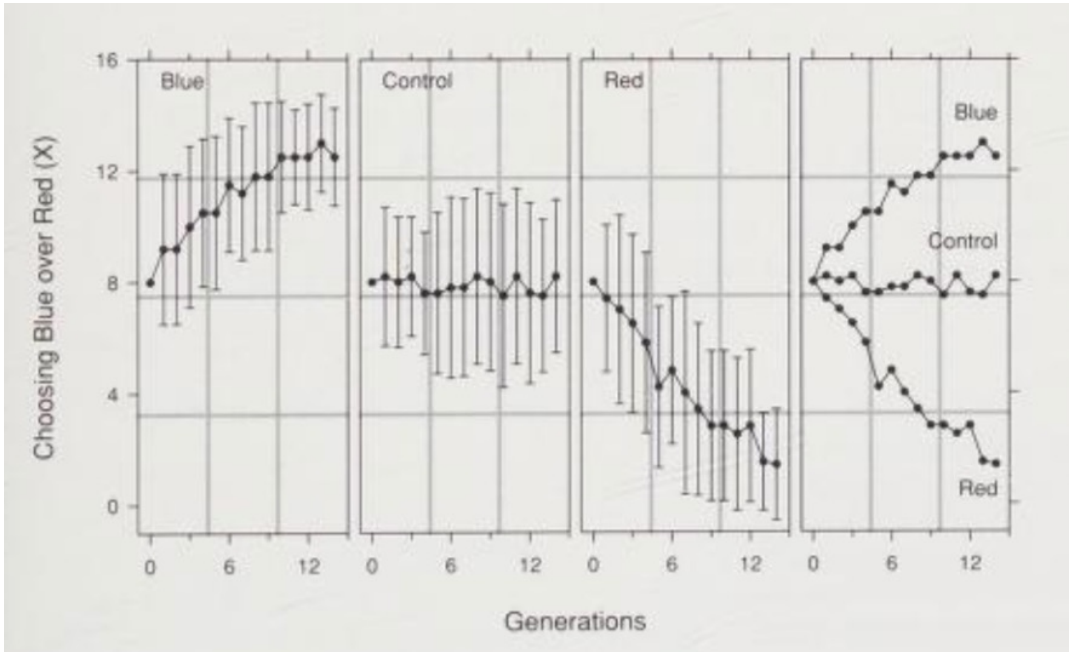


When the orientation of line segments are judged to decode information about rate of change, bank the segments to 45 degrees

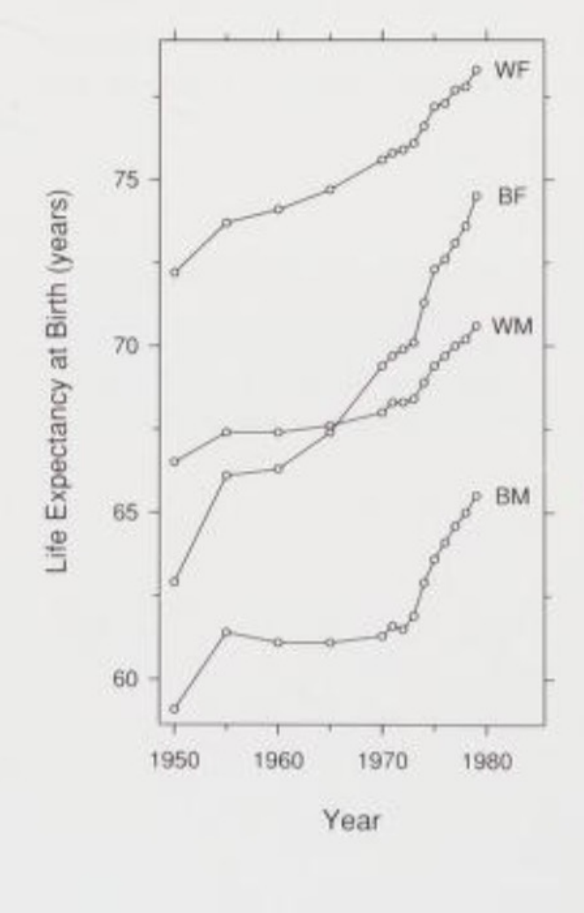


When the orientation of line segments are judged to decode information about rate of change, bank the segments to 45 degrees

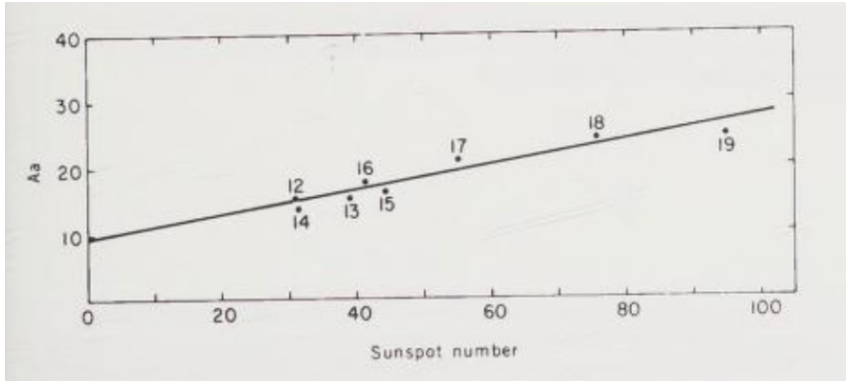


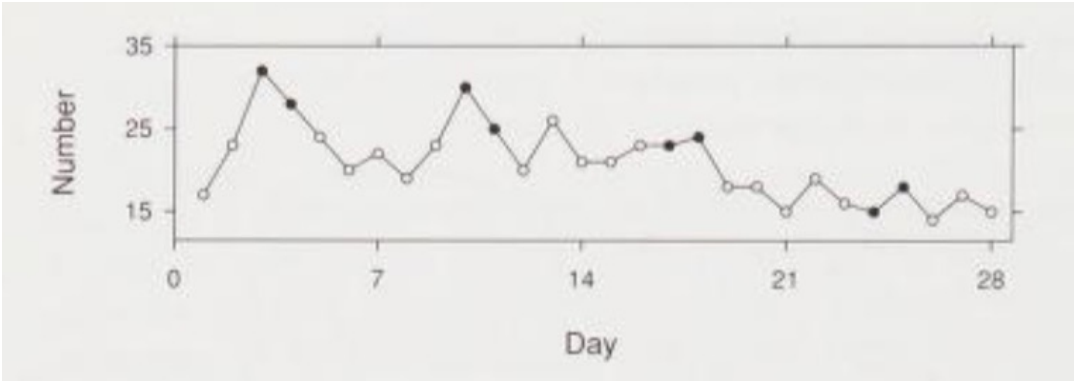


Choose the range of the tick marks to include or nearly include the range of the data

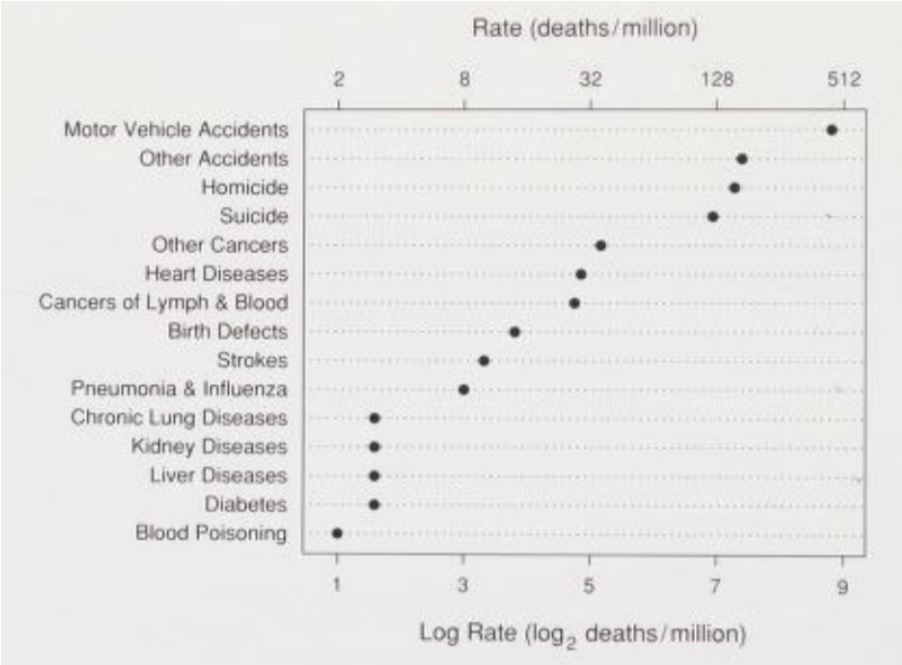


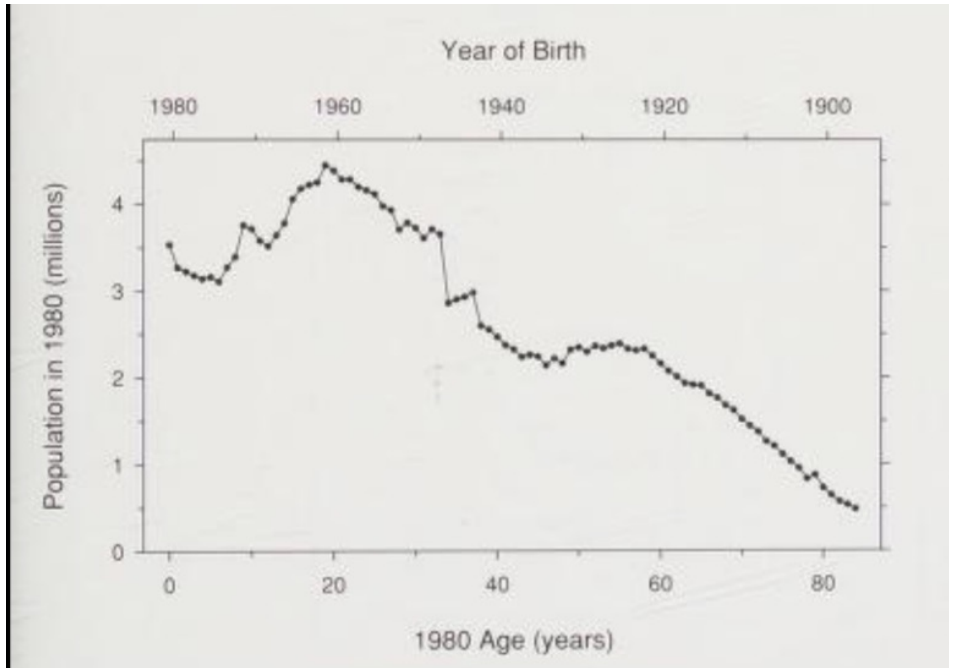
Subject to the constraints that scales have, choose the scales so that the data rectangle fills up as much of the scale-line rectangle as possible



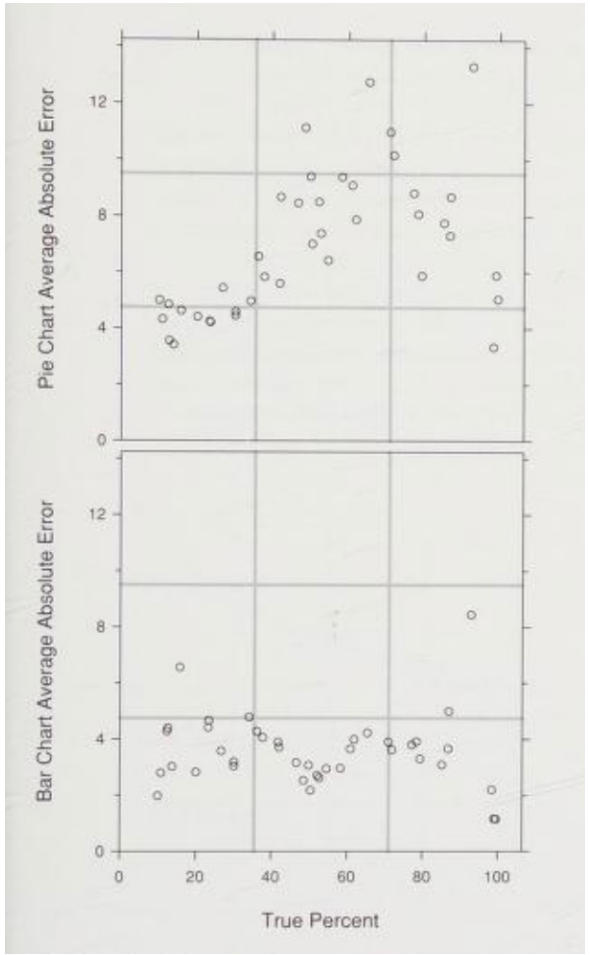


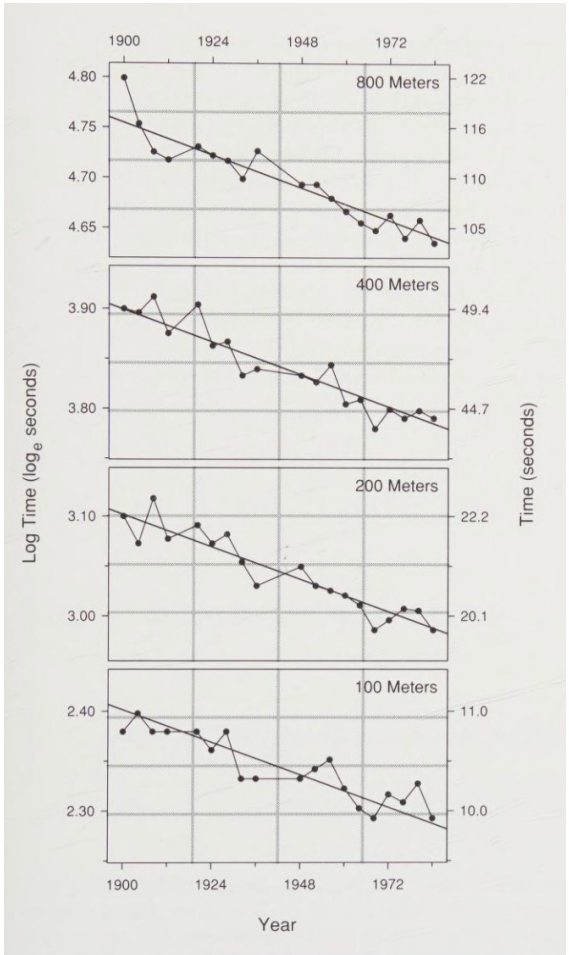
It is sometimes helpful to use the pair of scale lines for a variable to show two different scales

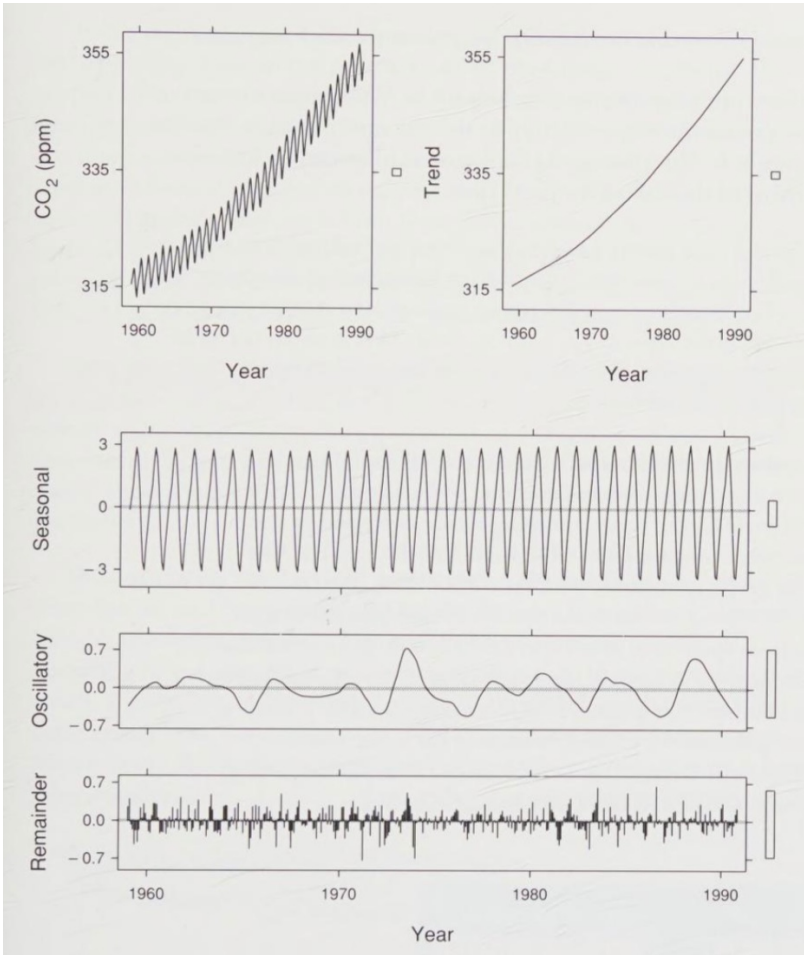




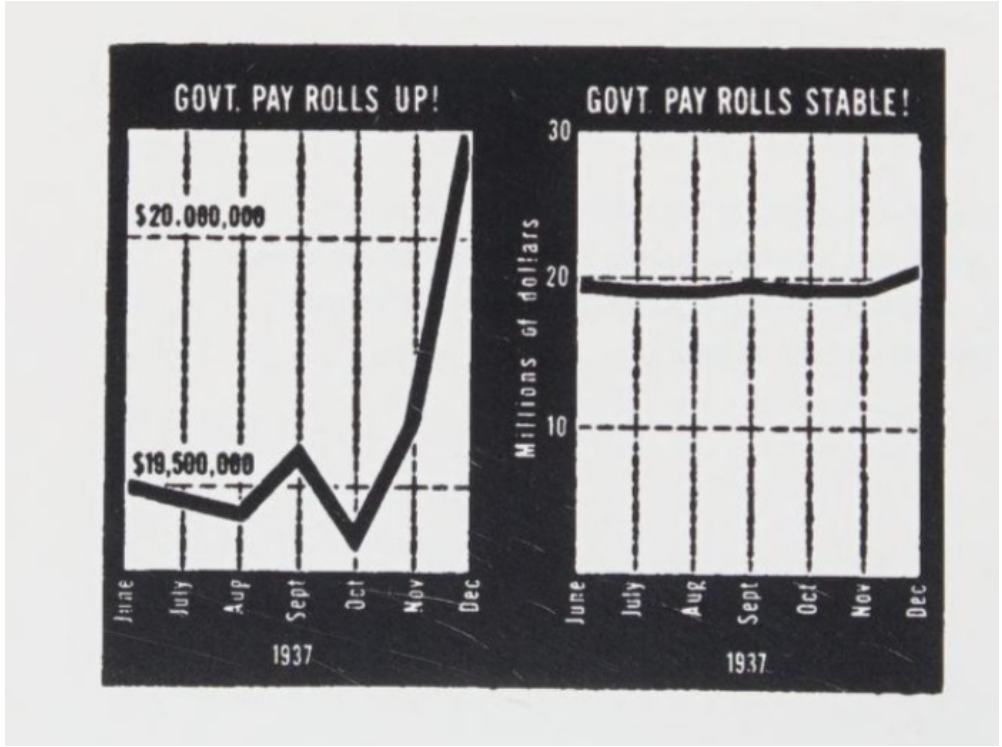
Choose appropriate scales when data on different panels are compared

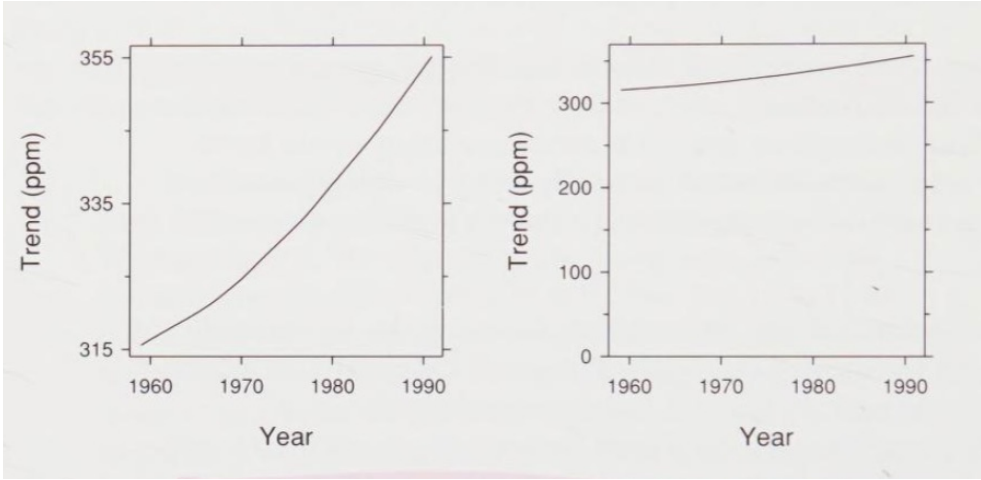




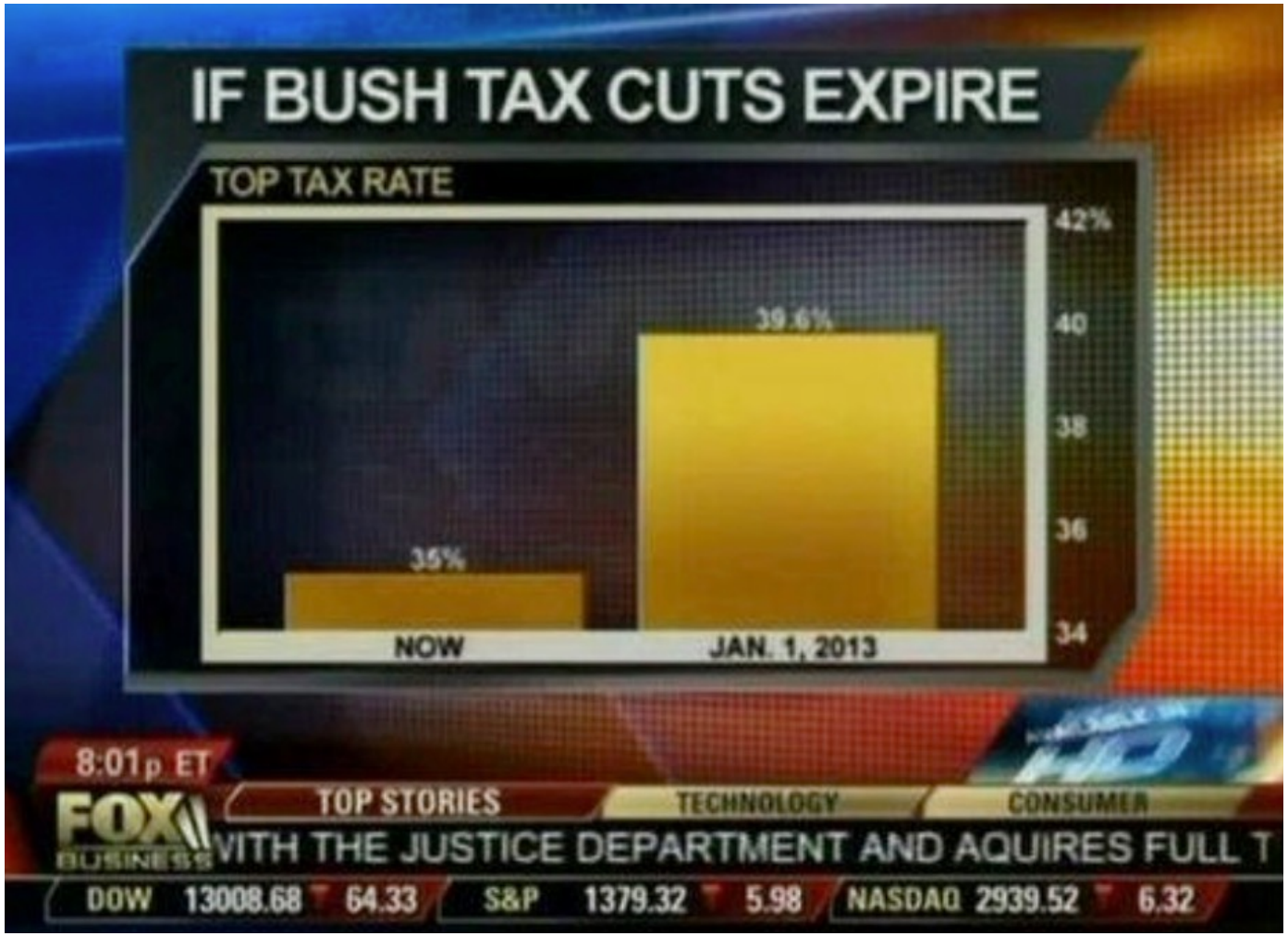


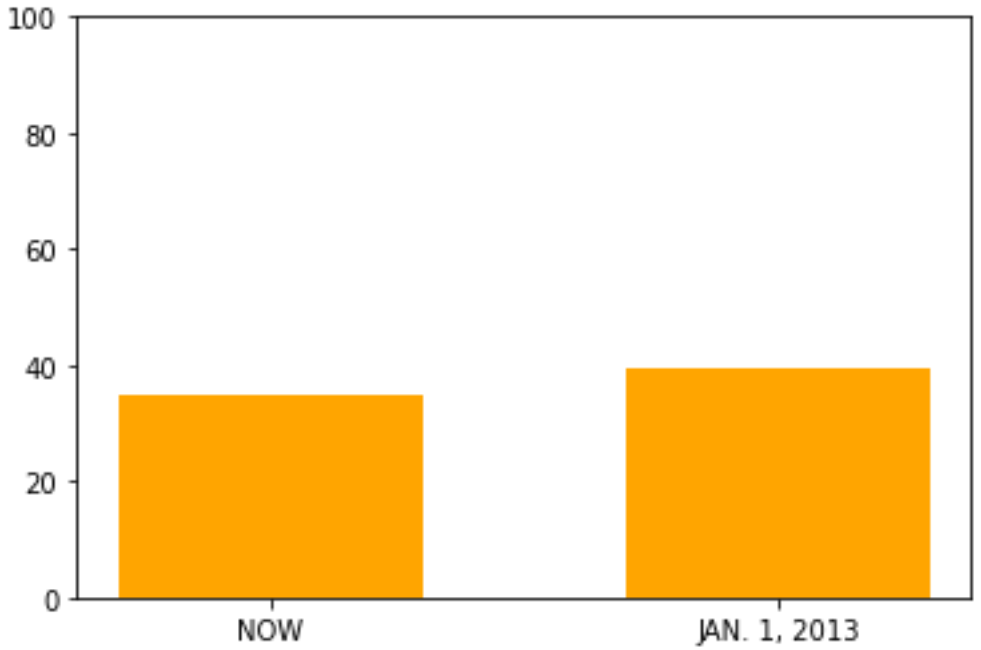
Do not insist that zero always be included on a scale showing magnitude



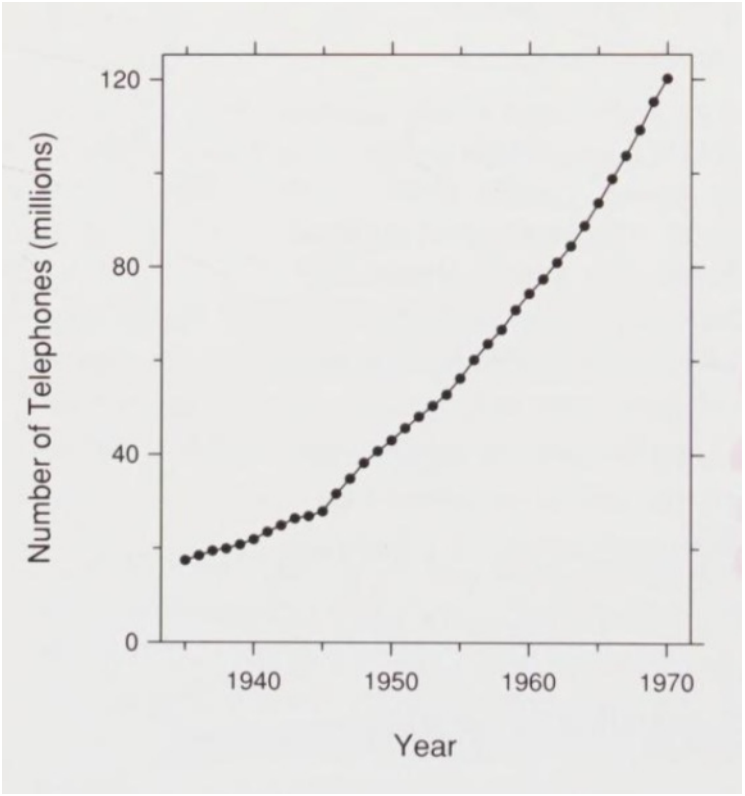


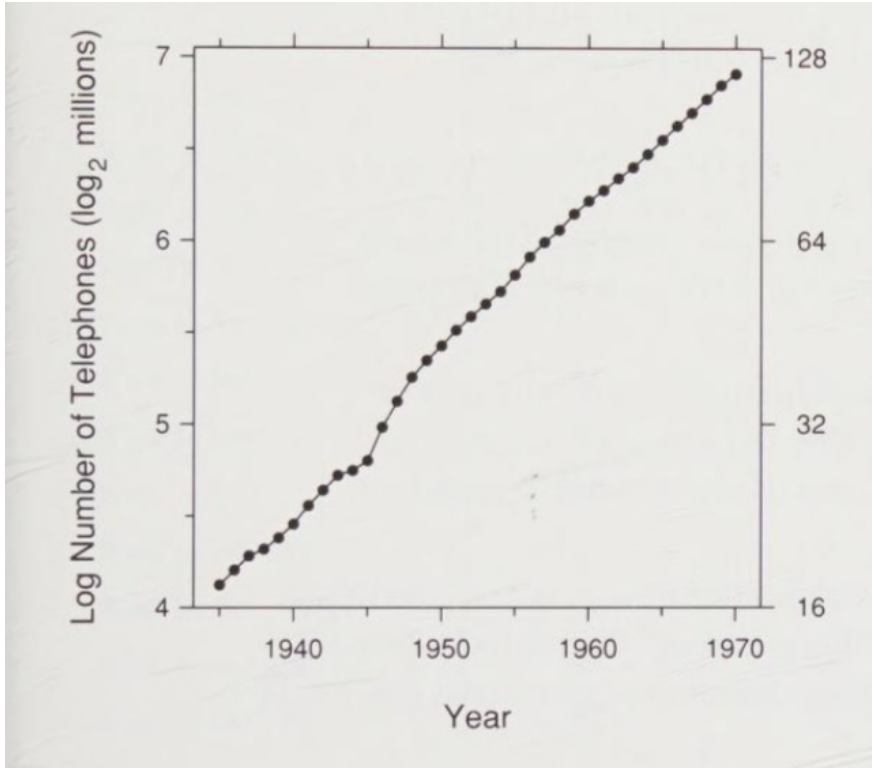
Is there a conflict with our rule here?



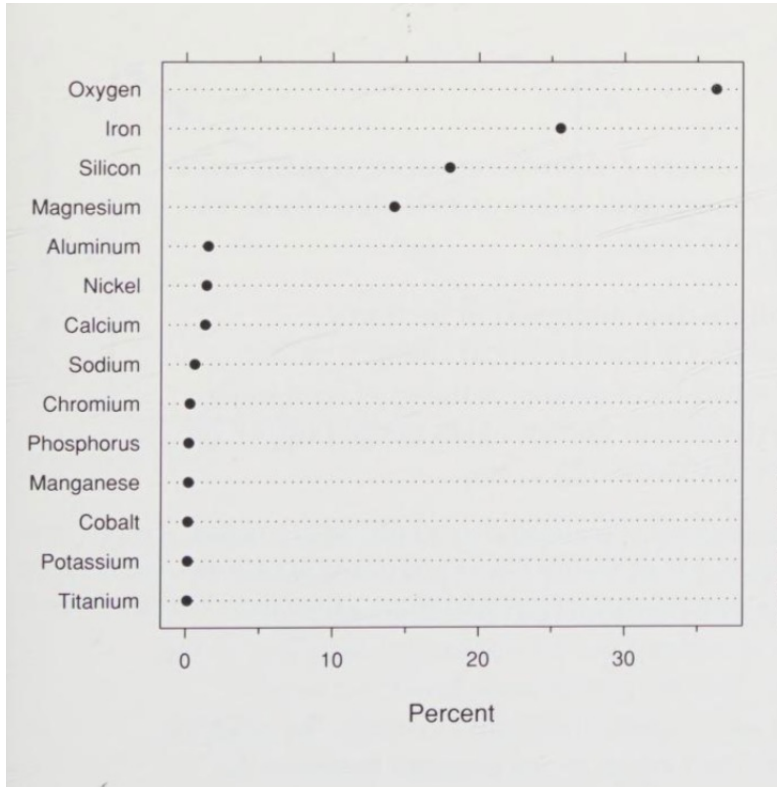


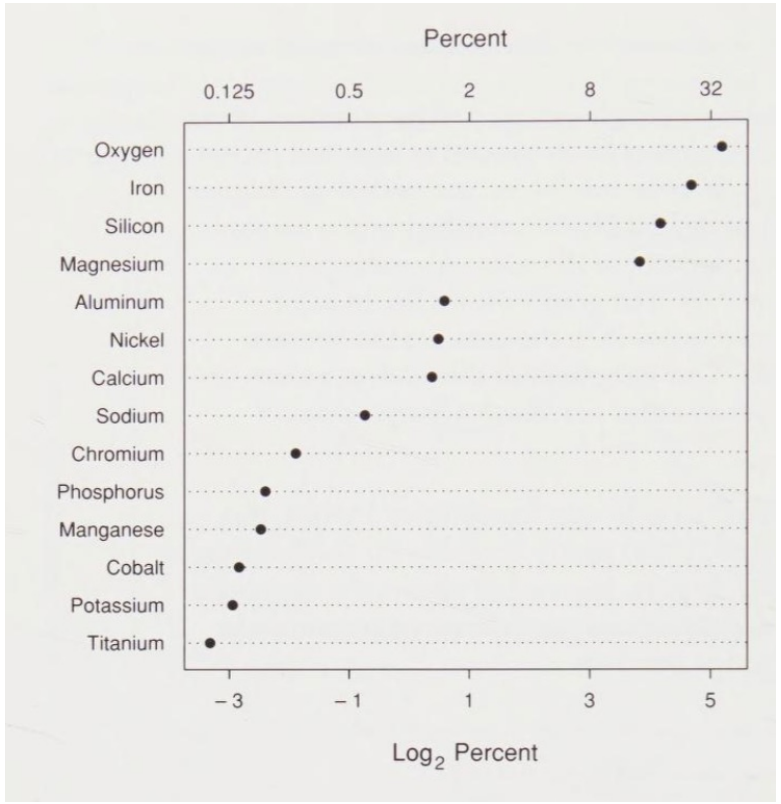
Use a logarithmic scale when it is important to understand percent change or multiplicative factors



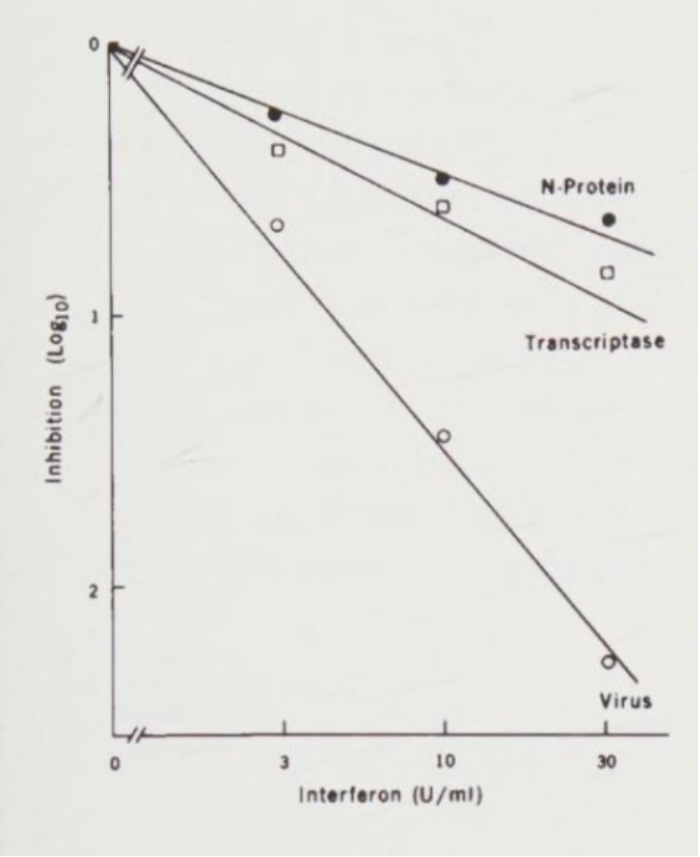


Showing data on a logarithmic scale can cure skewness toward large values

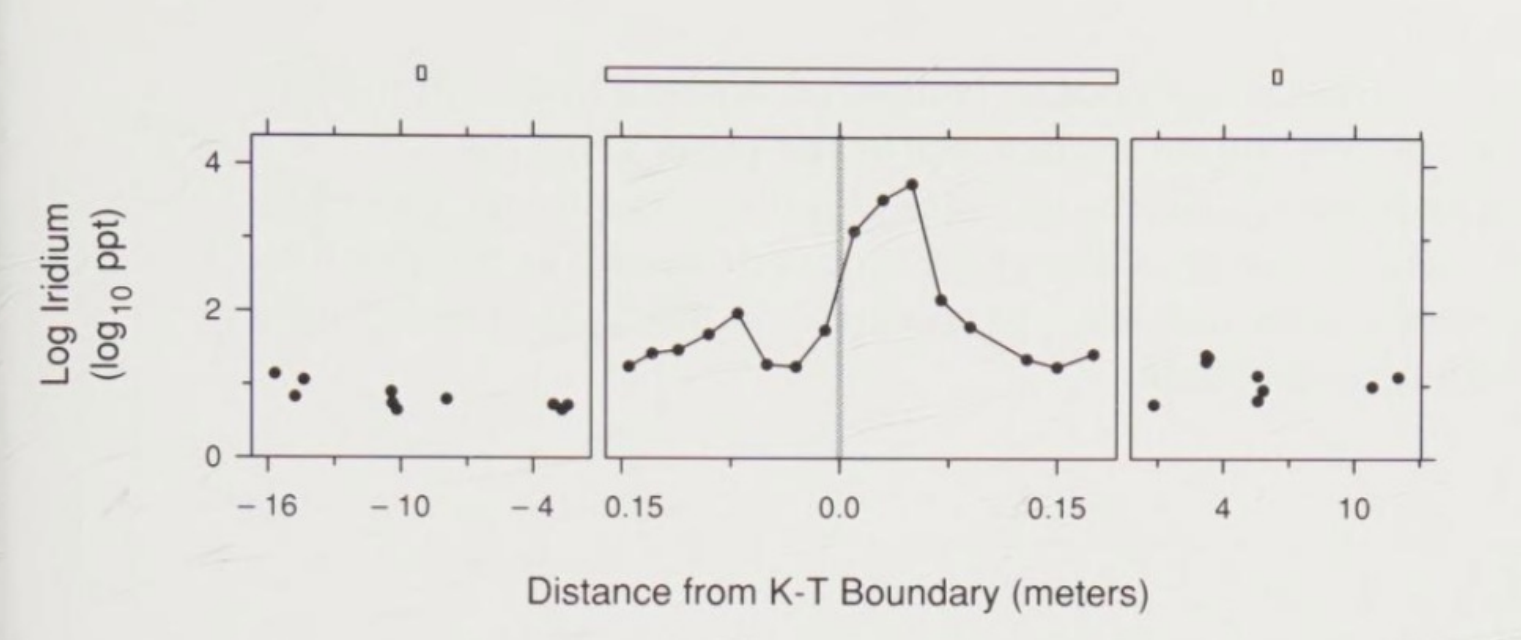




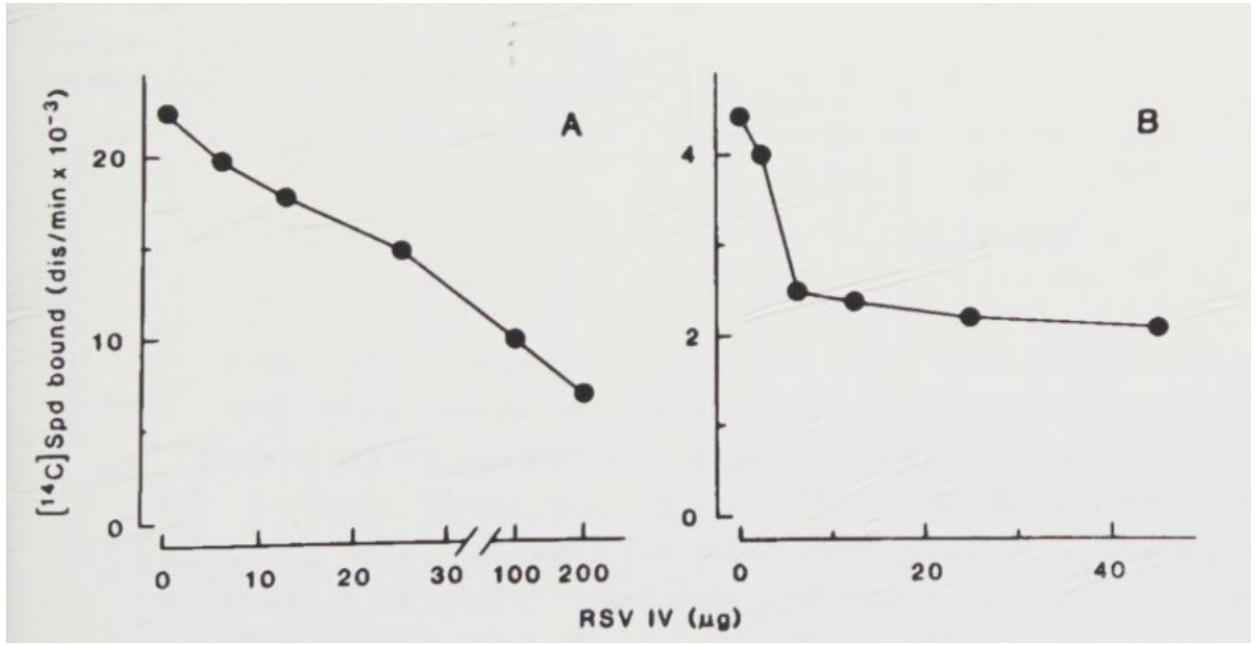
Use a scale break only when necessary; if a break cannot be avoided, use a full scale break; do not connect numerical values on two sides of a break; taking logs can cure the need for a break

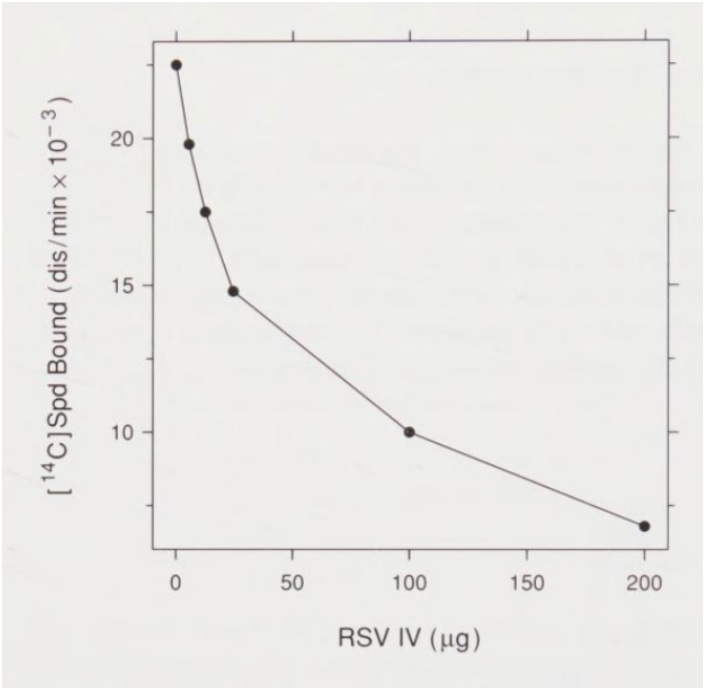


Use a scale break only when necessary; if a break cannot be avoided, use a full scale break; do not connect numerical values on two sides of a break; taking logs can cure the need for a break

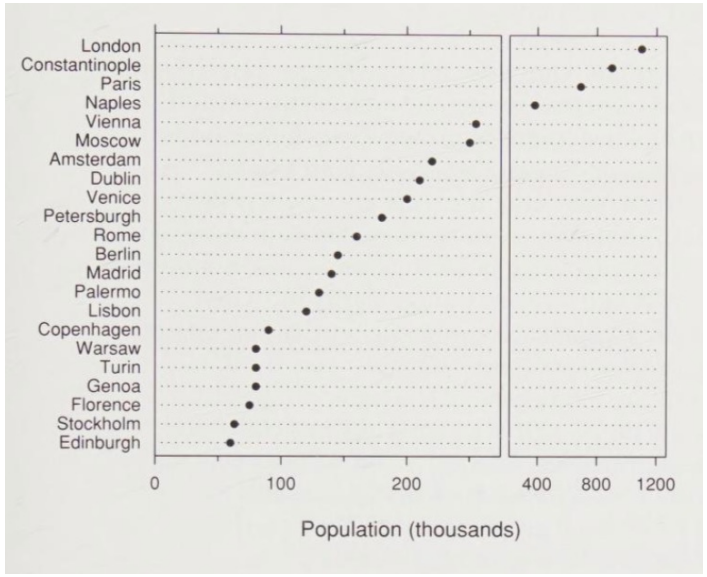


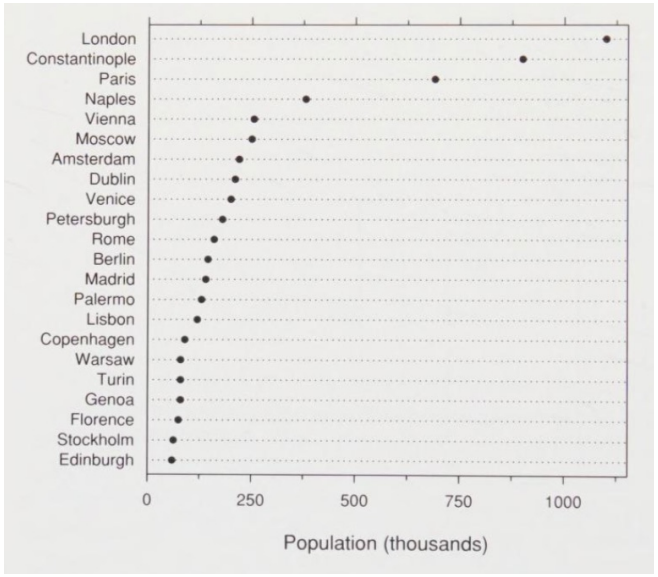
Use a scale break only when necessary; if a break cannot be avoided, use a full scale break; do not connect numerical values on two sides of a break; taking logs can cure the need for a break

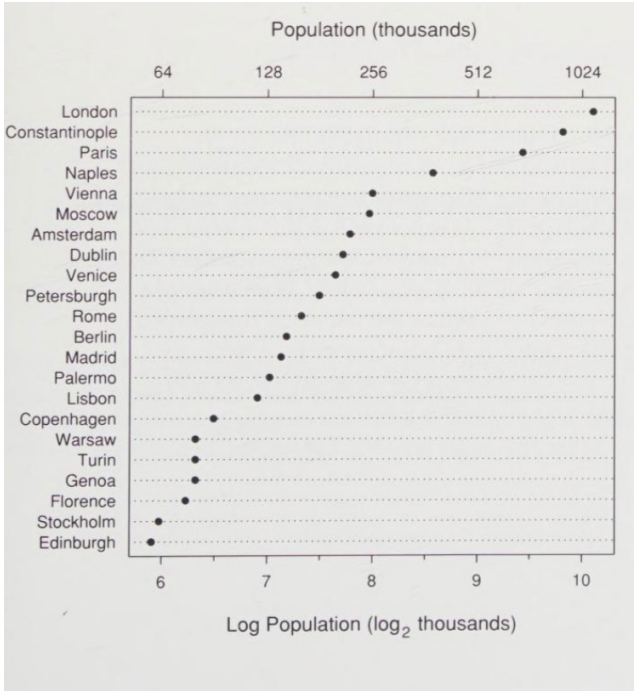




Use a scale break only when necessary; if a break cannot be avoided, use a full scale break; do not connect numerical values on two sides of a break; taking logs can cure the need for a break



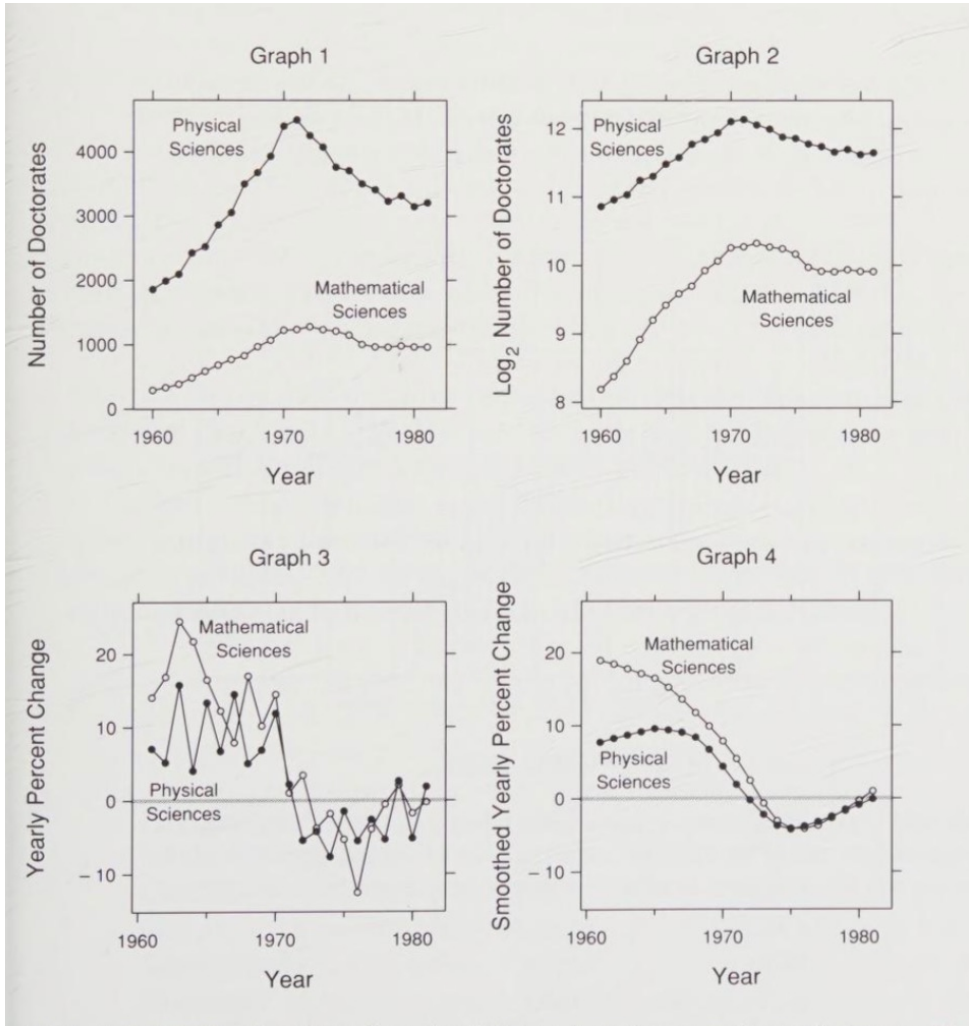




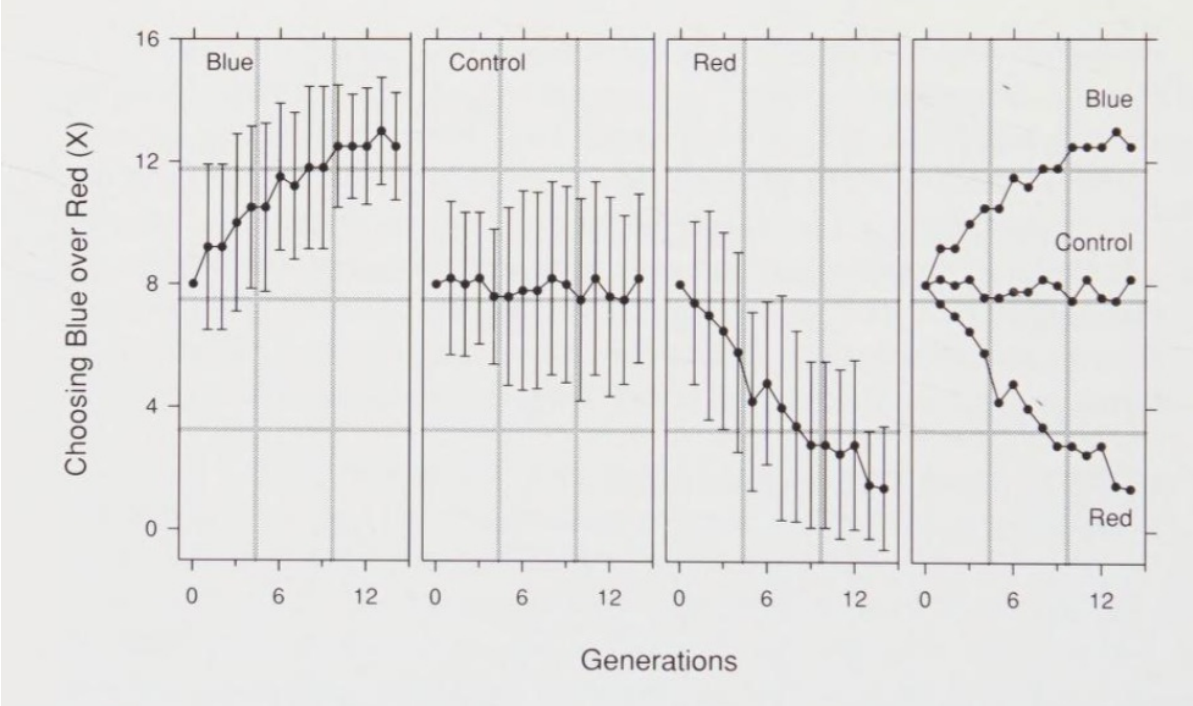
Principles of Graph Construction

- General strategy
  - A large amount of quantitative information can be packed into a small region
  - Graphing data should be an iterative, experimental process
  - Graph data two or more times when it is needed
  - Many useful graphs require careful, detailed study

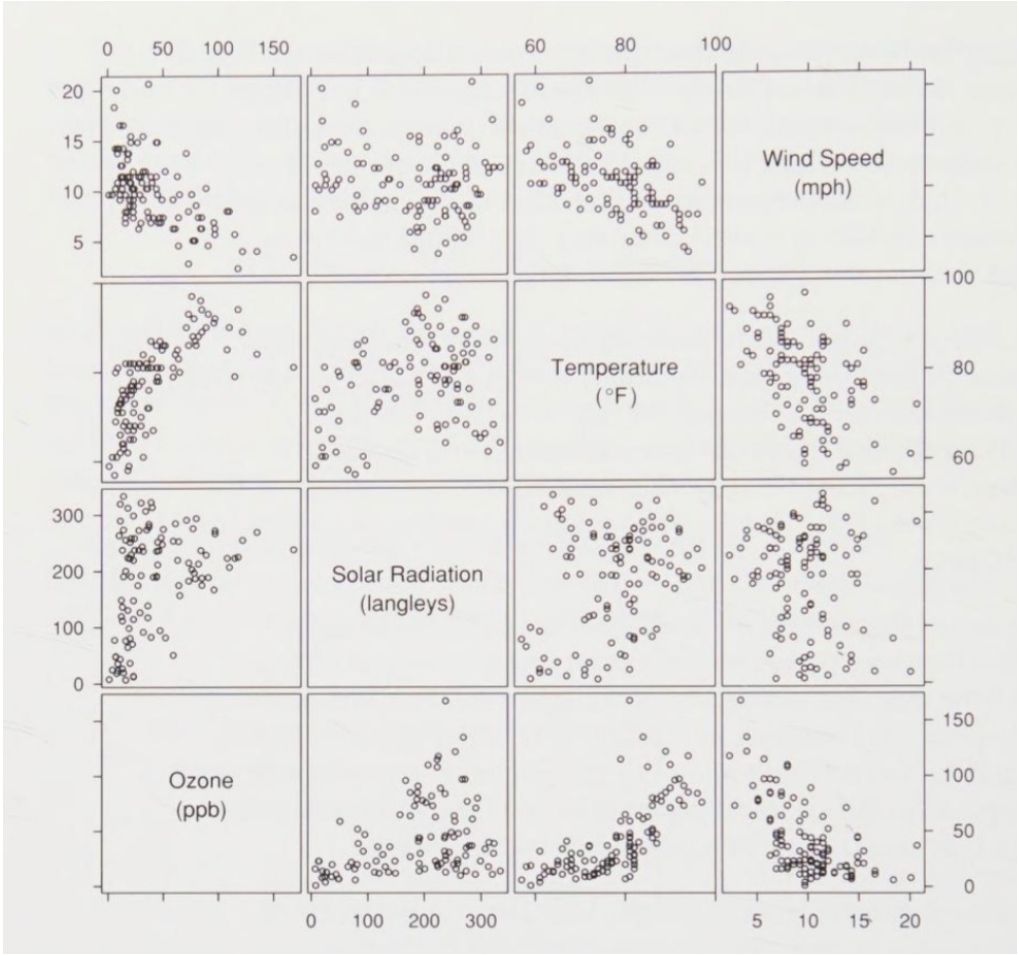
Graphing data should be an iterative, experimental process



Graph data two or more times when it is needed



Many useful graphs require careful, detailed study



Graphical Methods

- The usefulness of various types of plots –> this is left to be studied throughout our course
  - Logarithms
  - Residuals
  - Distributions
    - 1D scatterplots and histograms
    - Quantile plots
    - Box plots
    - Q-Q plots (quantile-quantile plots)
  - Dot plots and multi-dot plots
  - Plotting symbols
    - Overlap
    - Superposed plotting symbols
    - Superposed curves
  - Visual reference grids
  - Loess (**lo**cally **we**ight regre**ss**ion)
  - Time series
  - Scatter plots
  - Statistical variation

Graphical Perception

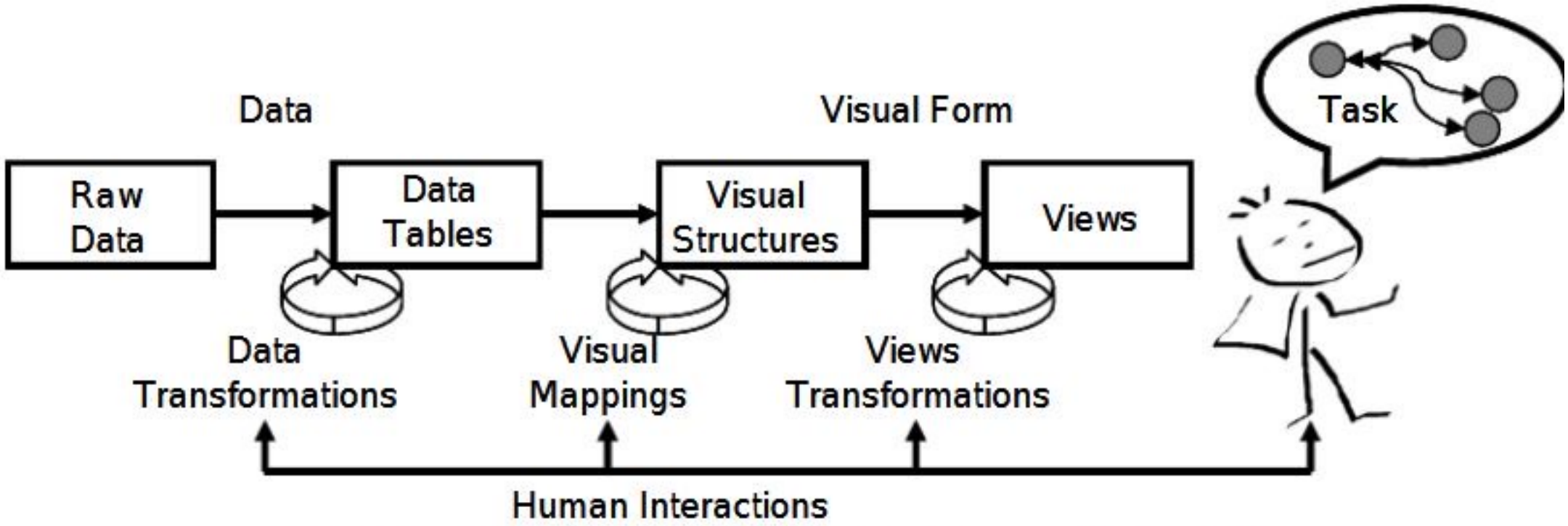
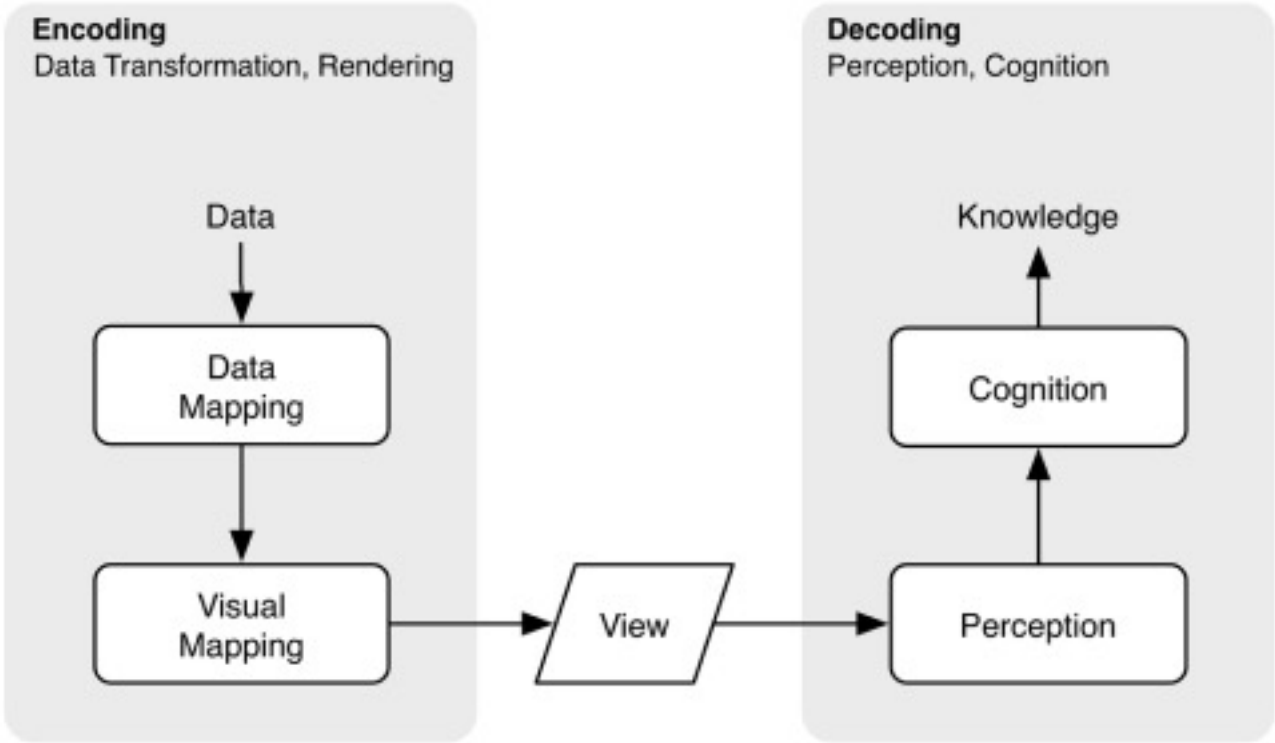


Figure 1 - Reference model for visualization (Card 1999).

https://rpubs.com/greddi/GR-DV-Perception-essay

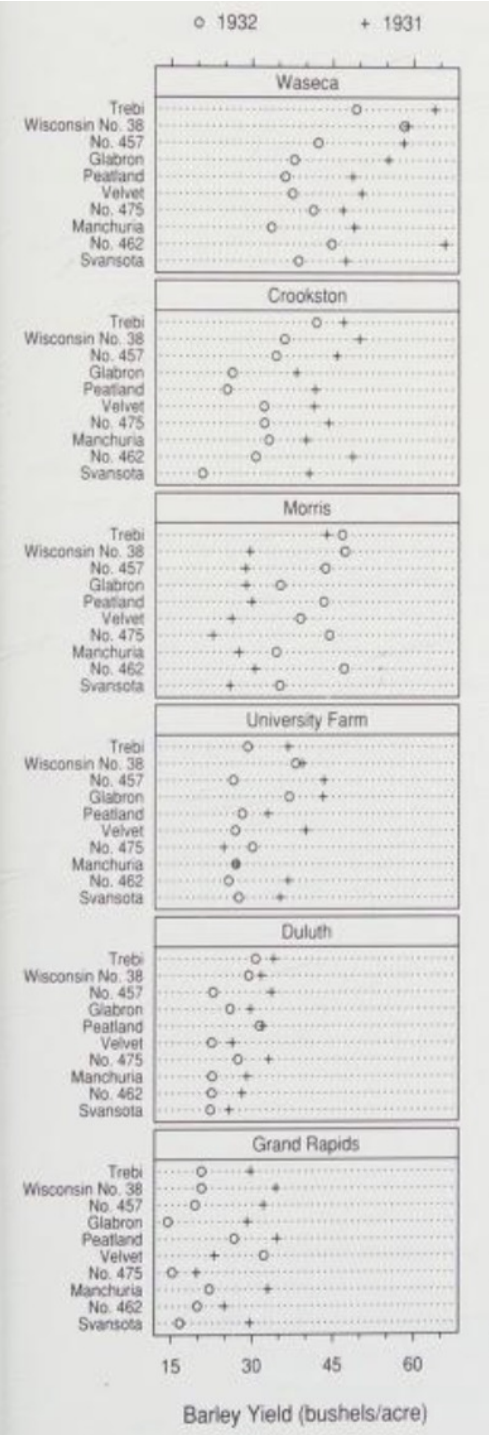
Graphical Perception



https://www.proprofsdiscuss.com/q/1693020/what-is-the-difference-between-encoding-and

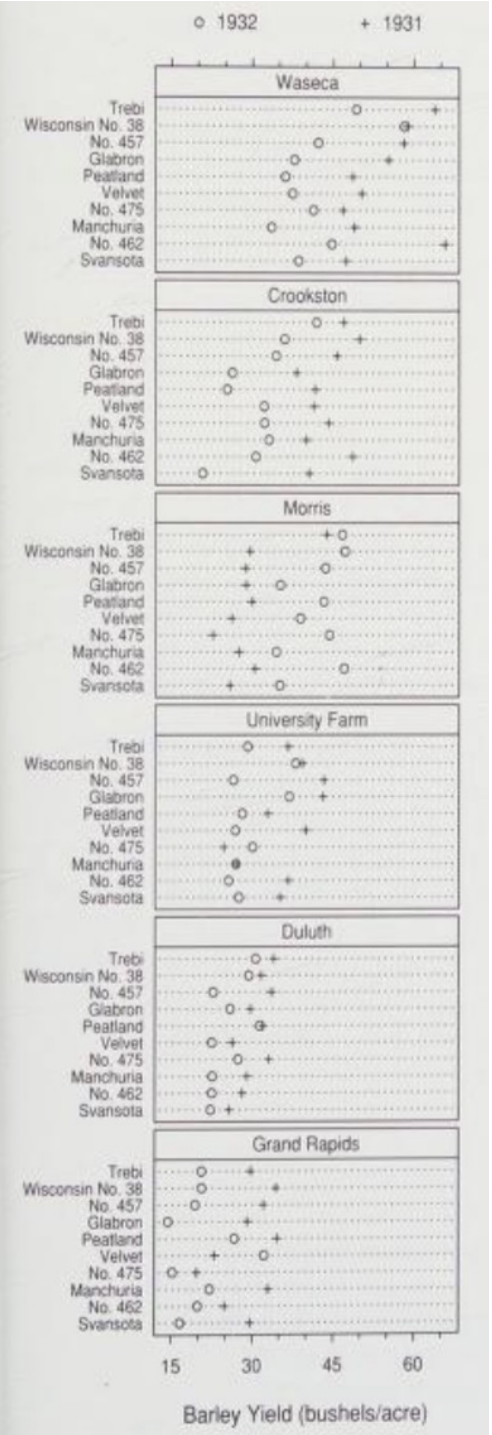
Graphical Perception

- Quantitative and categorical information
- Scale information and physical information
- Table look-up and pattern perception
  - Visual decoding of scale information
  - Visual decoding of physical information



Graphical Perception

- Three visual operations of table look-up
  - Scanning
    - Visual inspection of the physical elements
  - Interpolation
    - Visual estimation of values based on a set of discrete known data values
  - Matching
    - Visual identification of a geometric element with a labeled data value

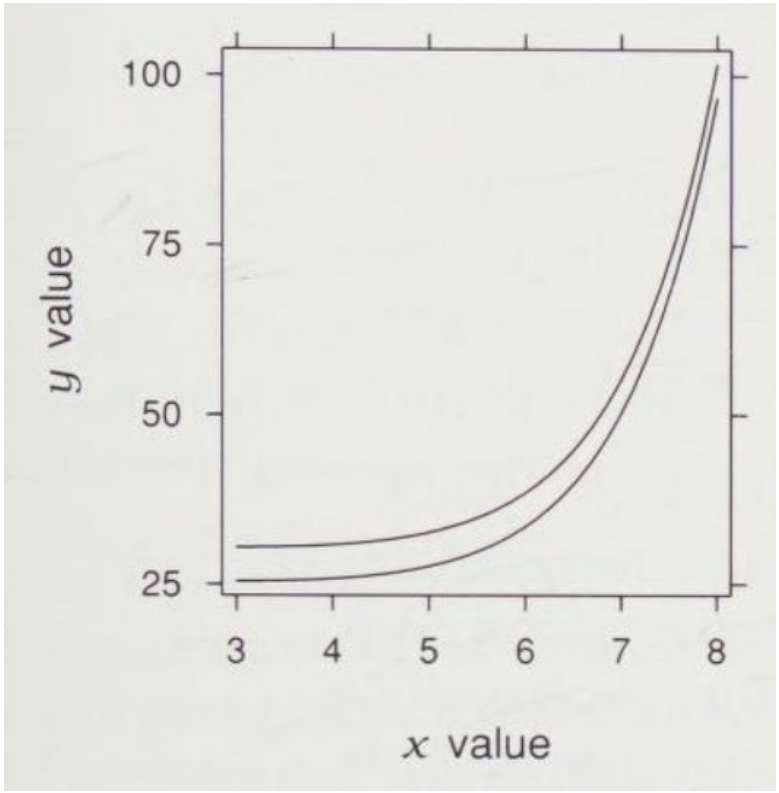


Graphical Perception

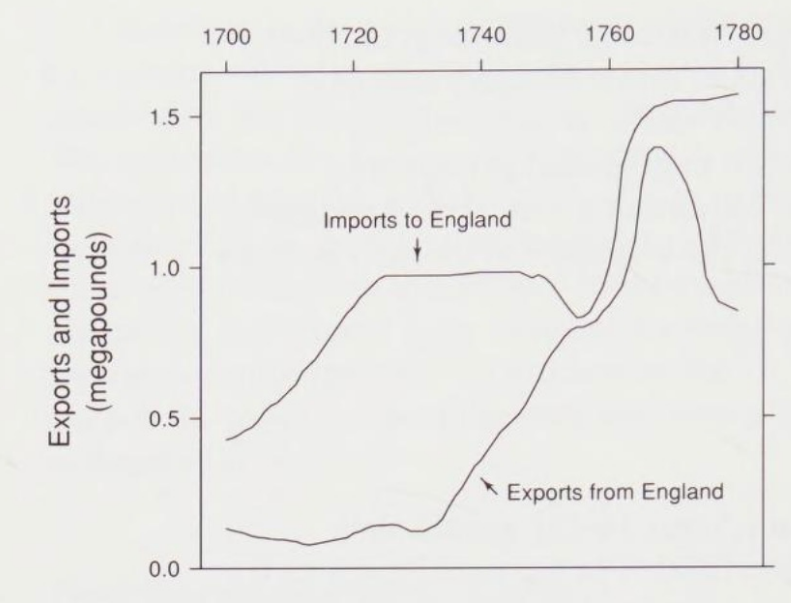
- Three visual operations of pattern perception
  - Detection
    - Visual recognition of a geometric aspect that encodes physical values
  - Assembly
    - Visual grouping of detected physical elements; discerning overall patterns in data
  - Estimation
    - Visual assessment of relative magnitudes of two or more physical values.
      - Discrimination (equal or not equal)
      - Ranking (less than or greater than)
      - Ratioing (a/b)

Pattern perception for curves can be inaccurate

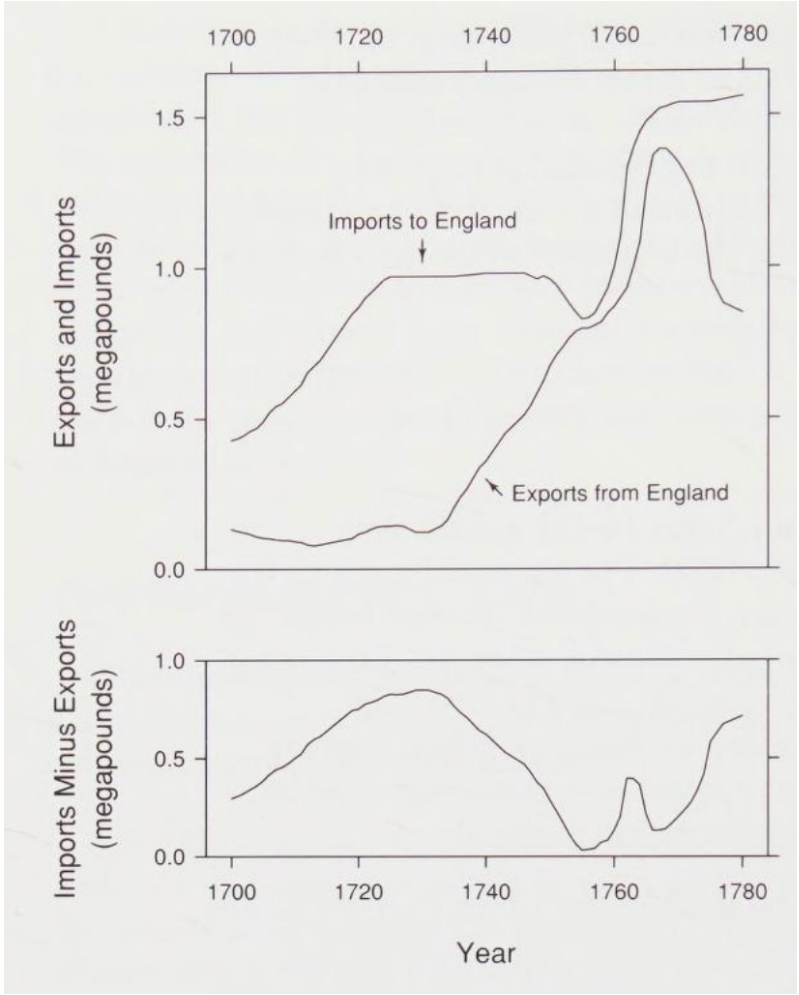
Not because we are inaccurate at estimating differences,
but because we're inaccurate at detecting differences



Pattern perception for curves can be inaccurate

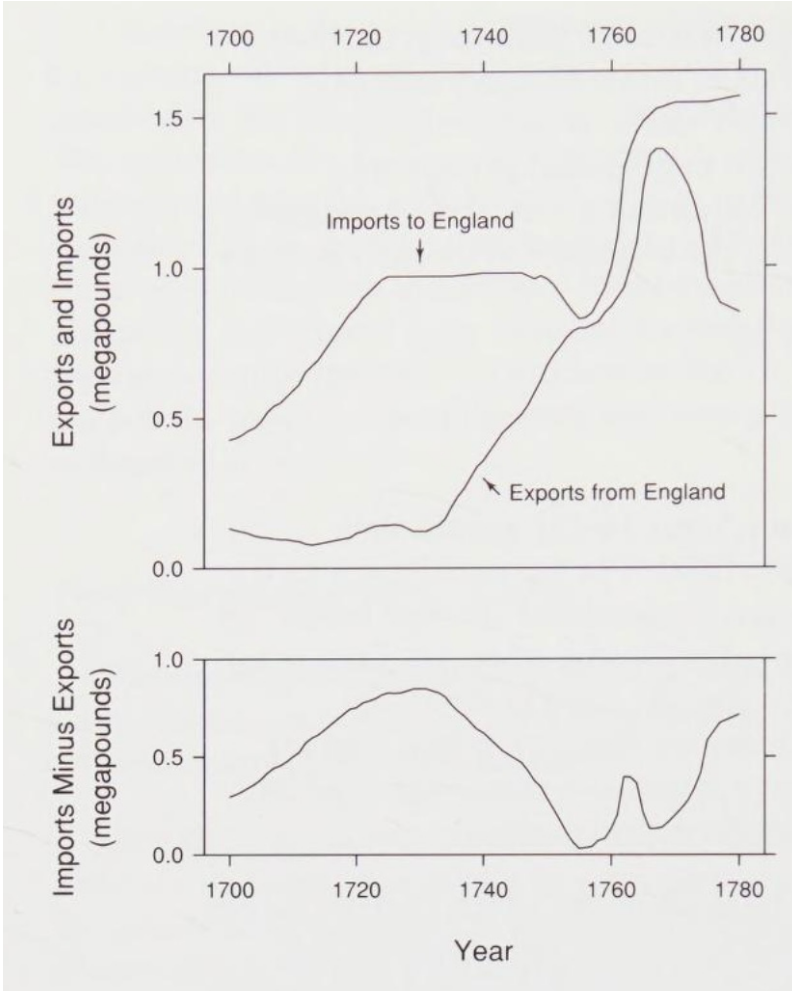


Pattern perception for curves can be inaccurate



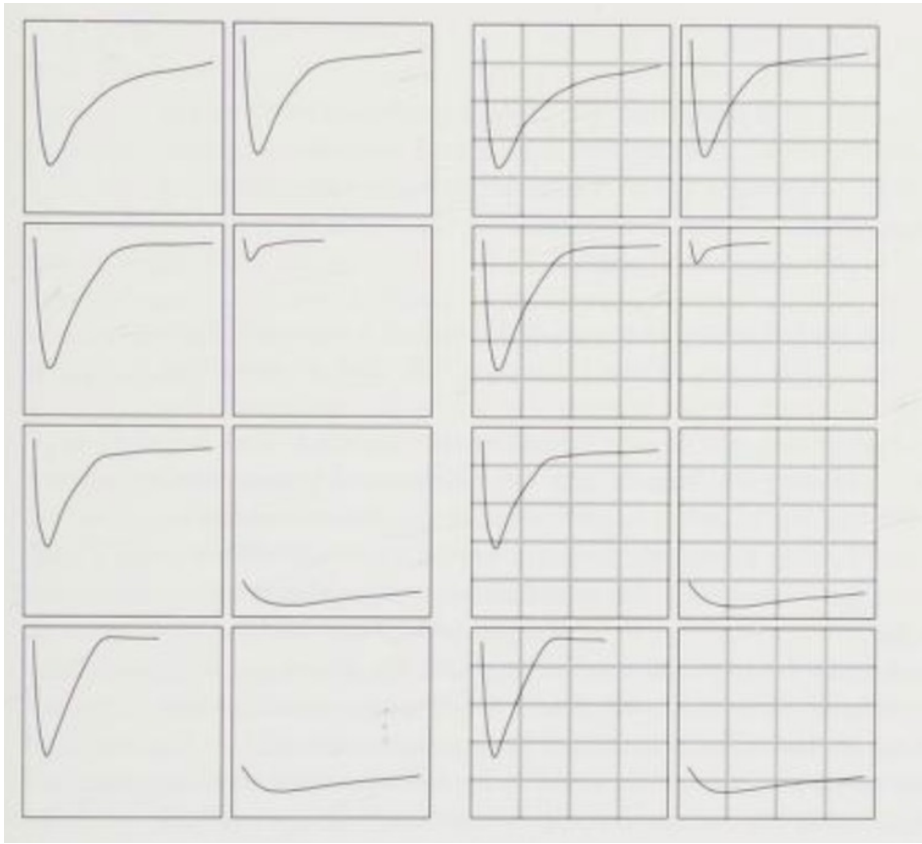
Pattern perception for curves can be inaccurate

- Table look-up is more efficient for the bottom panel than the top when calculating differences
- Remedy: superpose two curves on one panel and graph the differences on another



Pattern perception for curves can be inaccurate

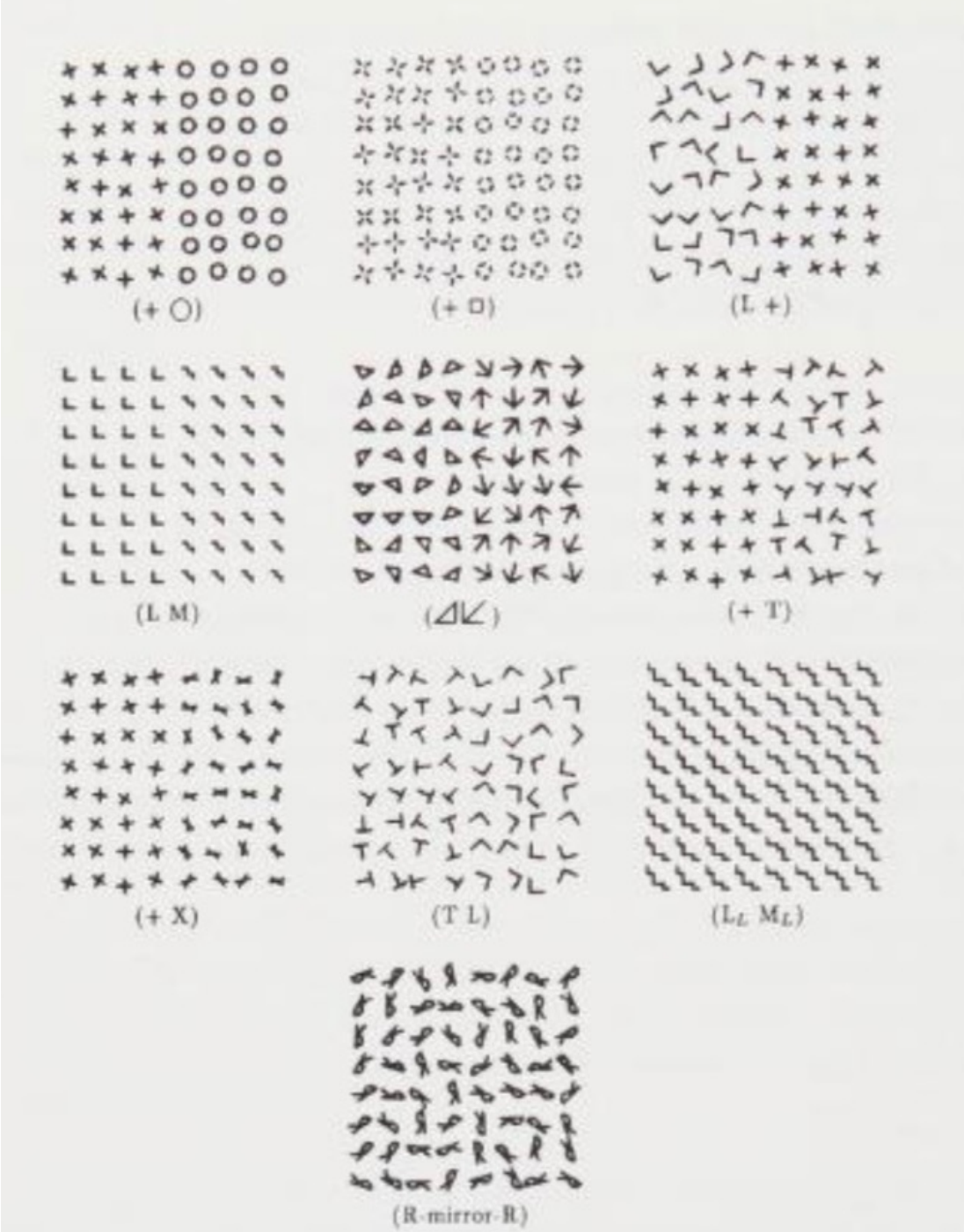
- Table look-up is more efficient for the bottom panel than the top when calculating differences
- Remedy: superpose two curves on one panel and graph the differences on another
- Another is to juxtapose curves on separate panels – eliminates the distortion caused by superposition



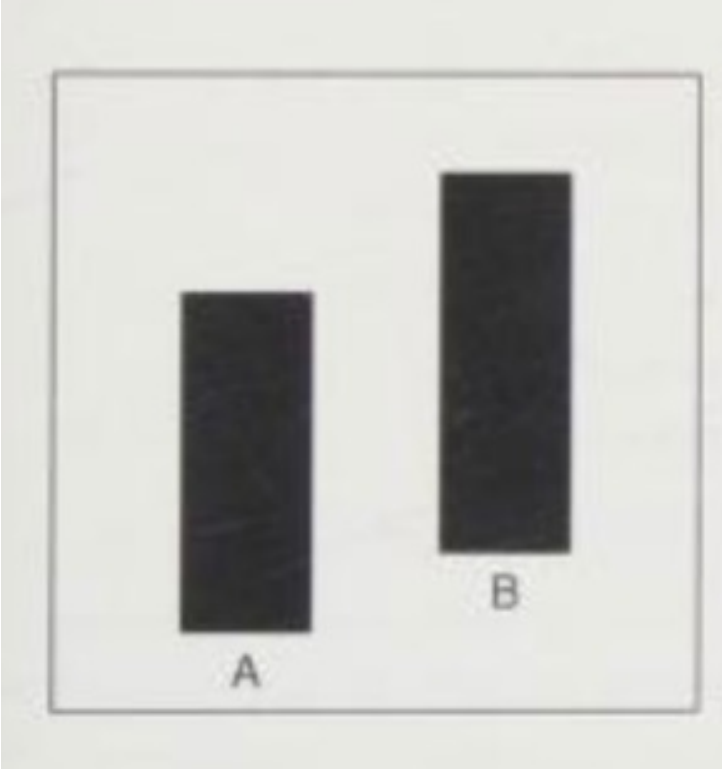
Color

- There are different approaches to color
  - HSL (Hue, Saturation, Lightness); CMYK (Cyan, Magenta, Yellow, Key (Black)), RGB (Red, Green, Blue) …..
- Changing hue can provide efficient discrimination of colors
  - For fixed hue, we can perceive ordering as either lightness changes or saturation changes
  - We cannot effortlessly perceive an ordering to changing hue.
  - Hue can be effective at perceiving boundaries between adjacent levels of color encoding
  - Two hues provide more color variation than just one
- We will show more on color in another session

- Marker identification
- Assembly (the grouping of elements) depends on effectively discriminating the texture of chosen markers

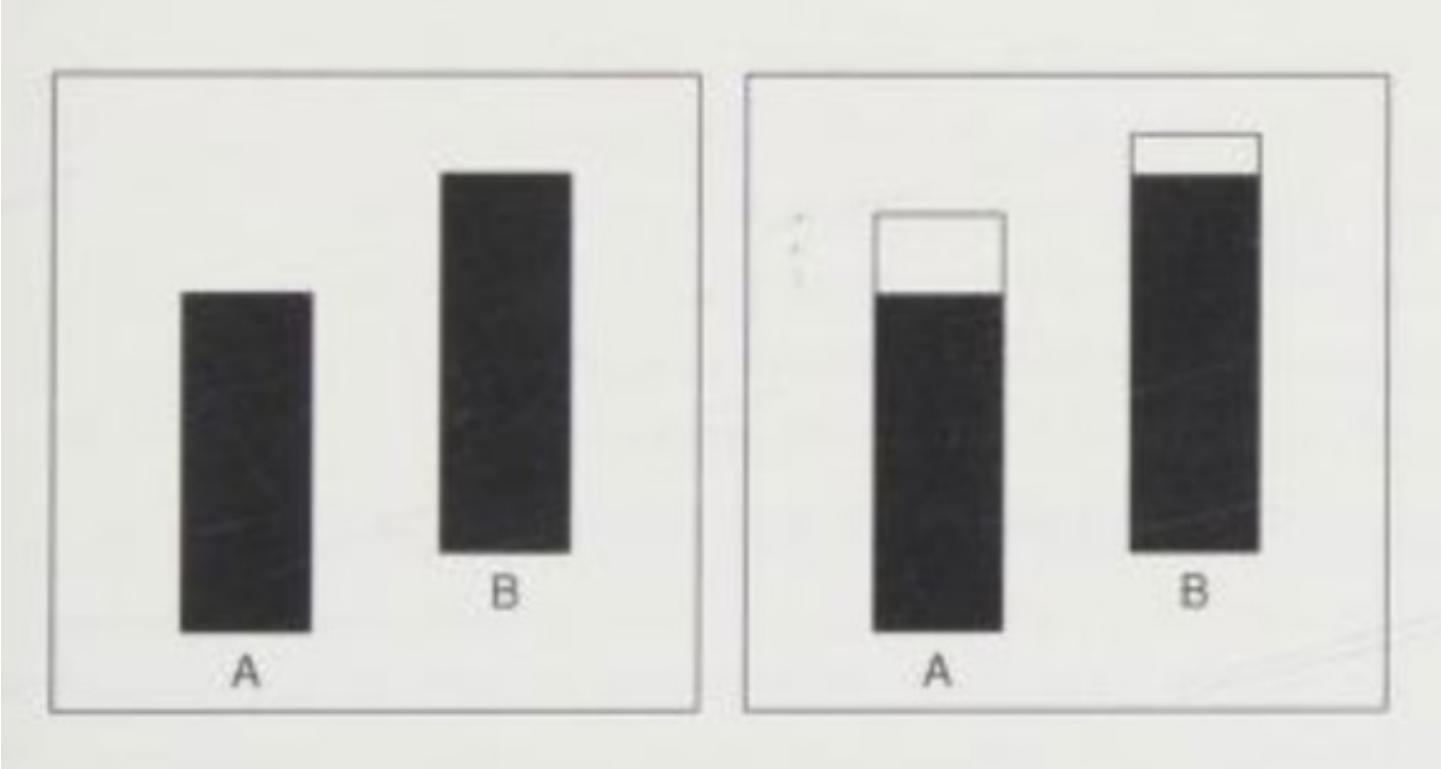


Detecting length differences is more accurate for large percent change

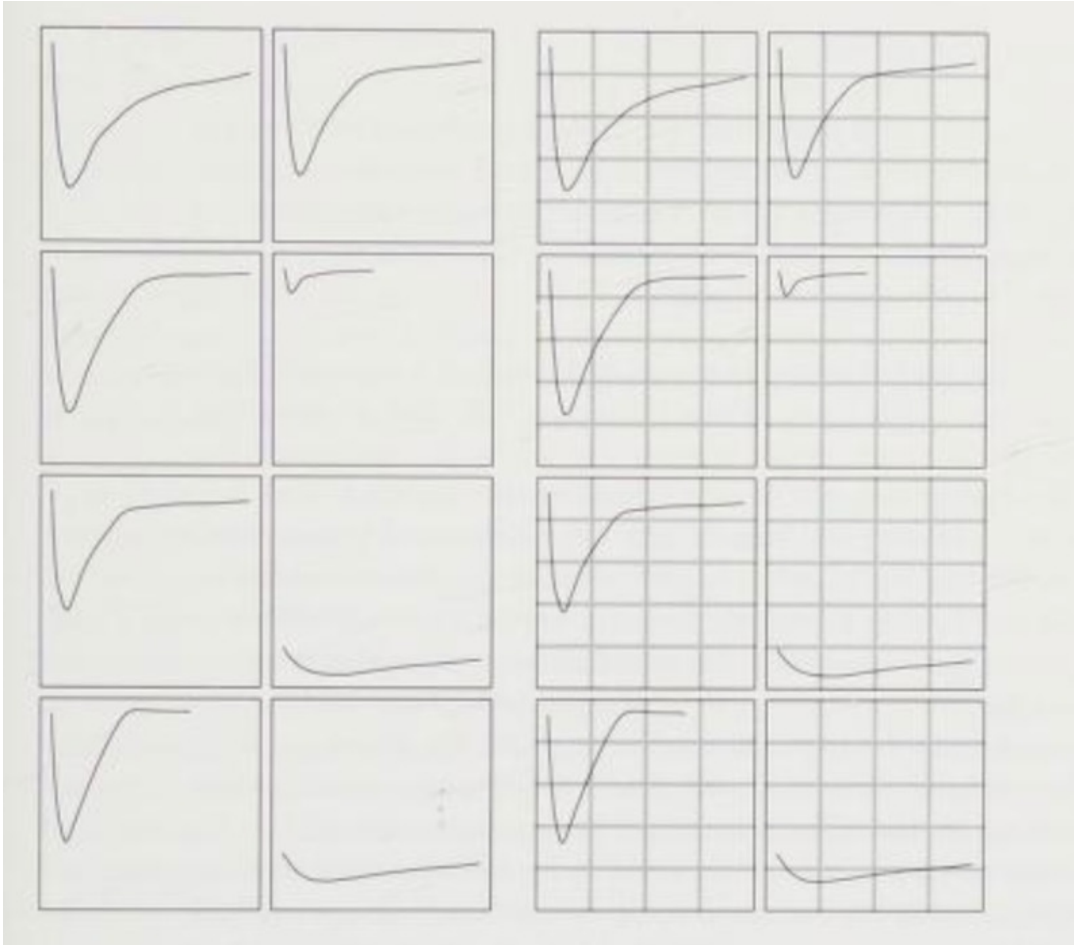


- Which bar is longer?

Detecting length differences is more accurate for large percent change

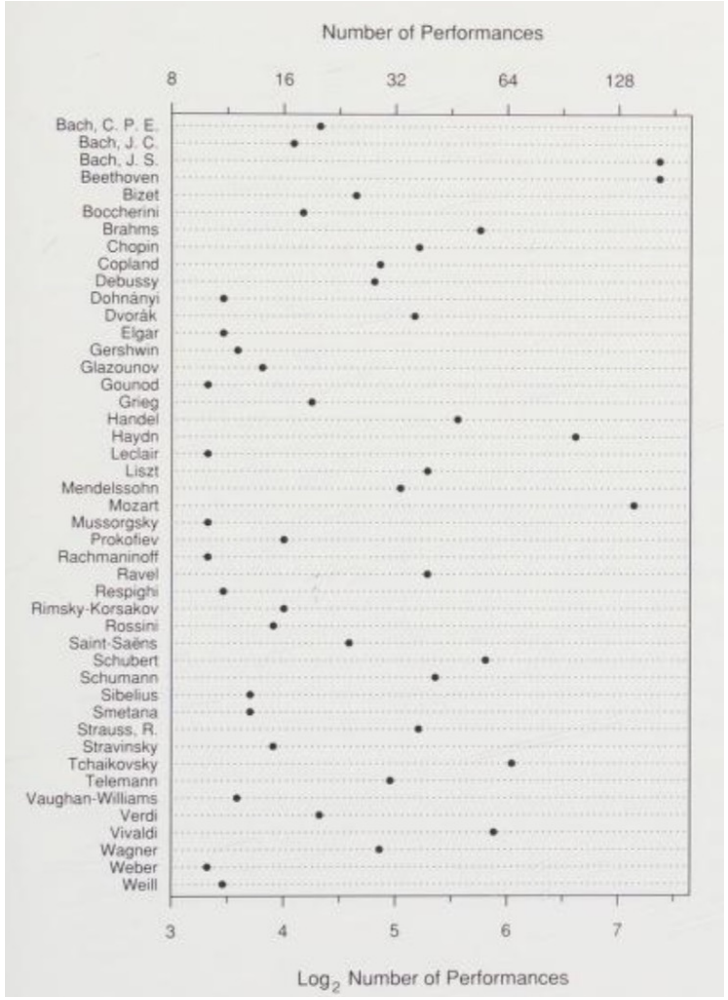


Grids can enhance pattern perception



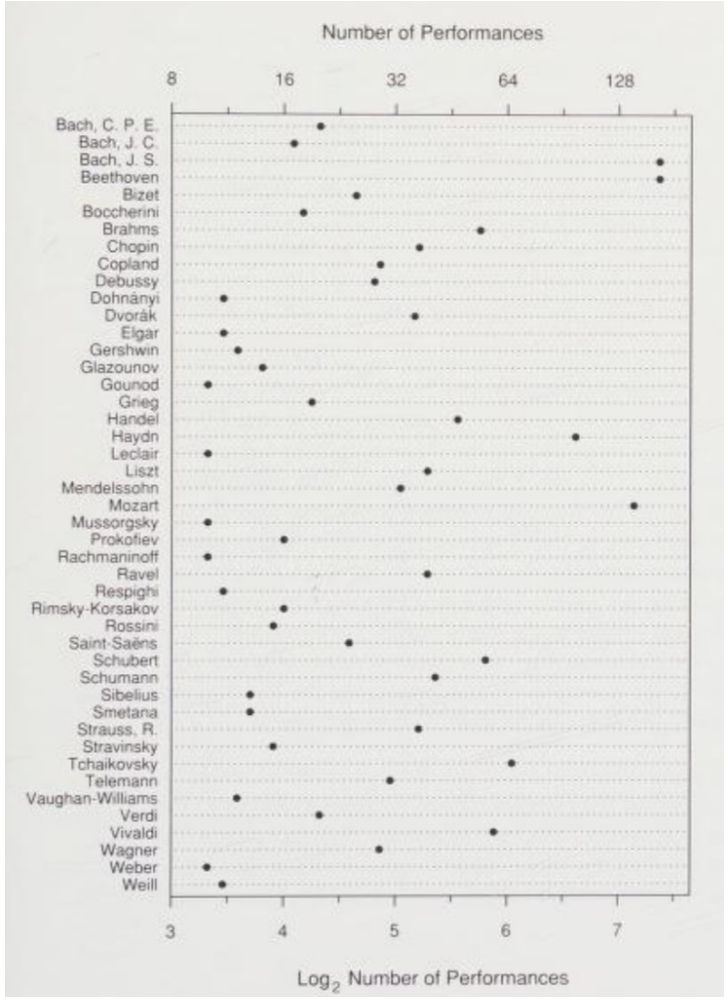
Ordering can make perception more effective

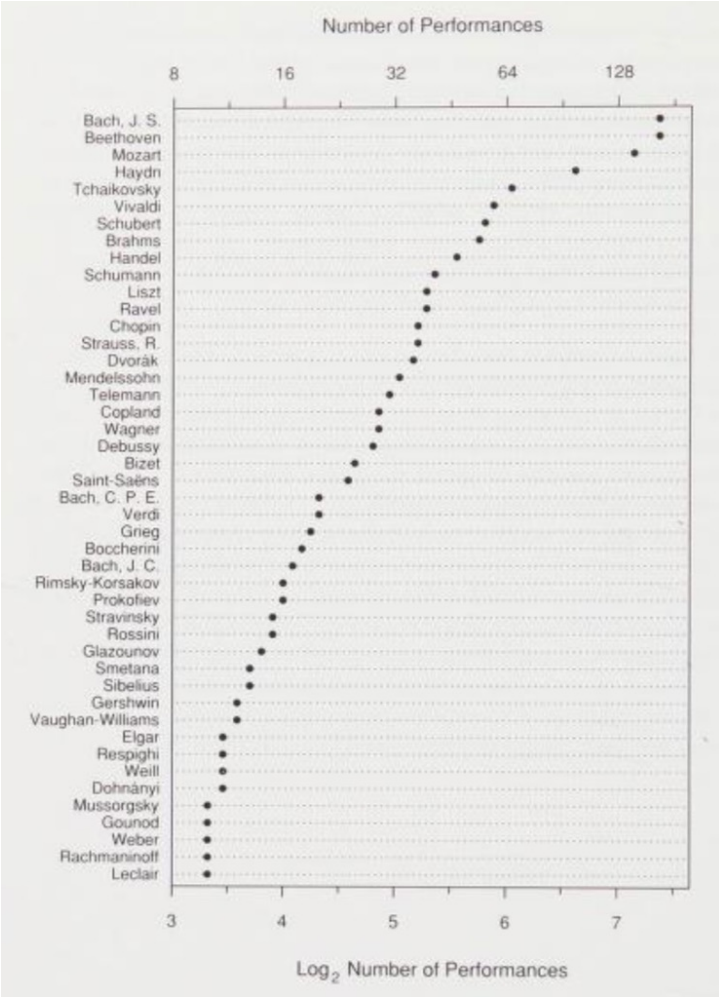
- Who are the top five composers?



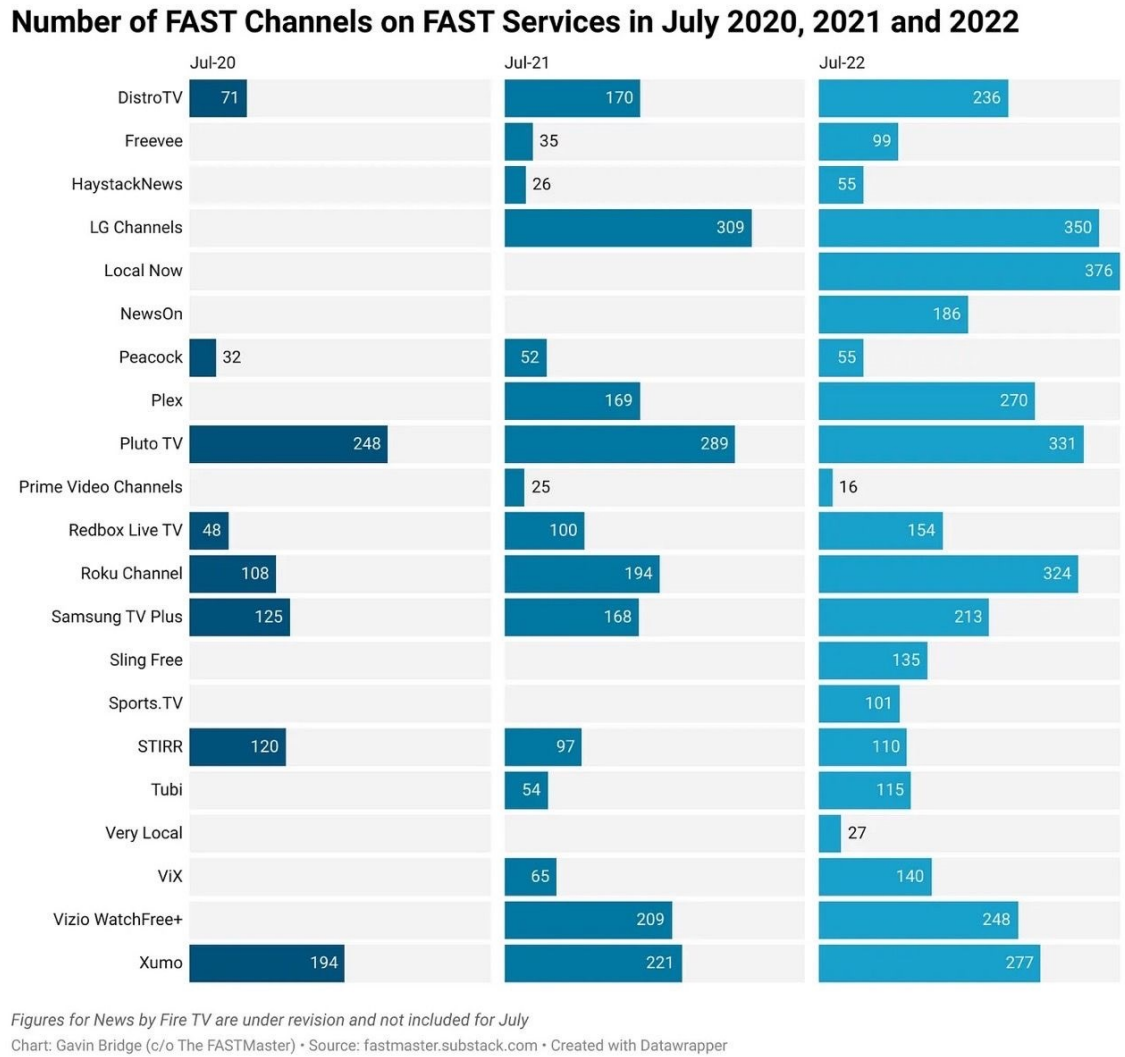
Ordering can make perception more effective

- Who are the top five composers?

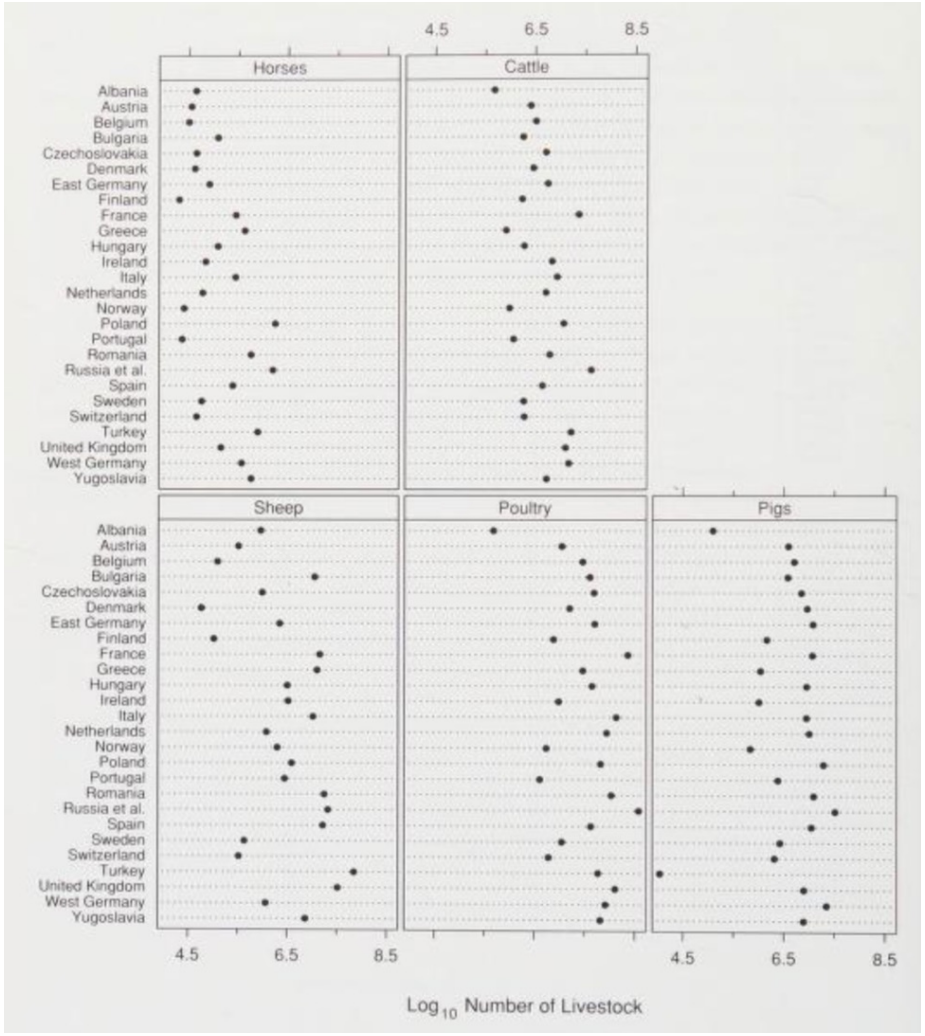


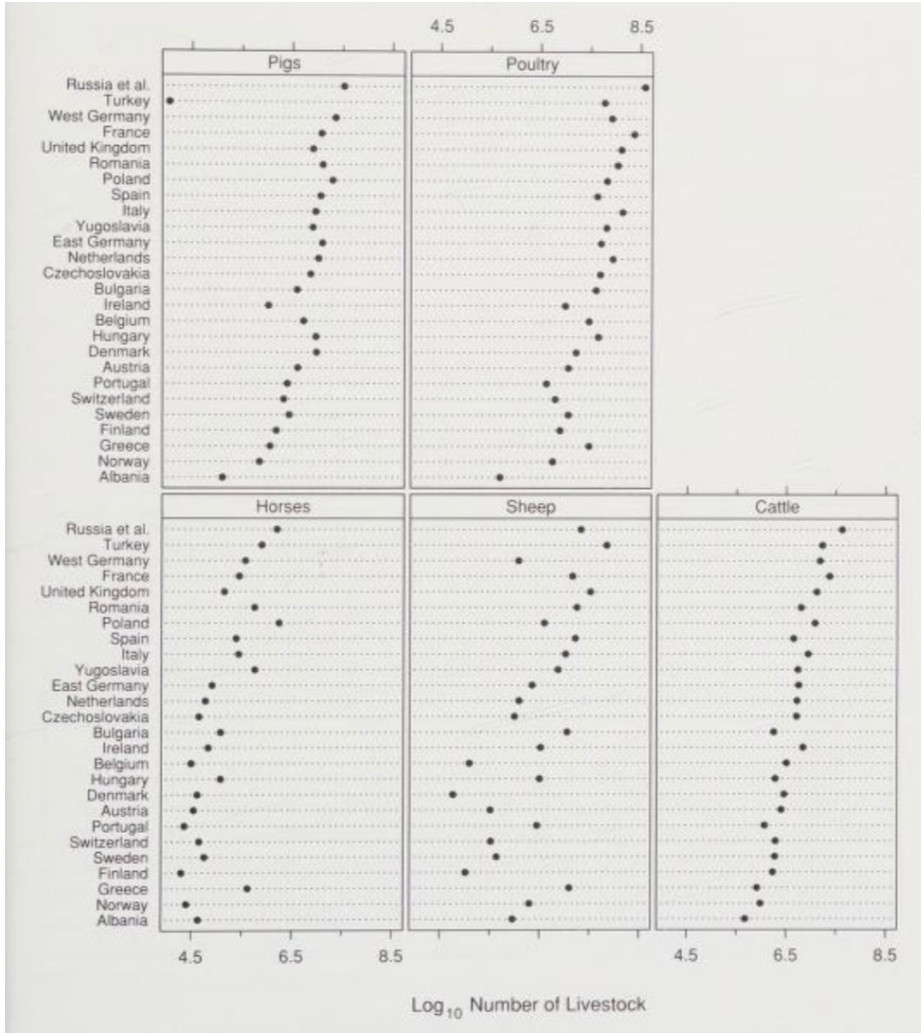


Ordering can make perception more effective



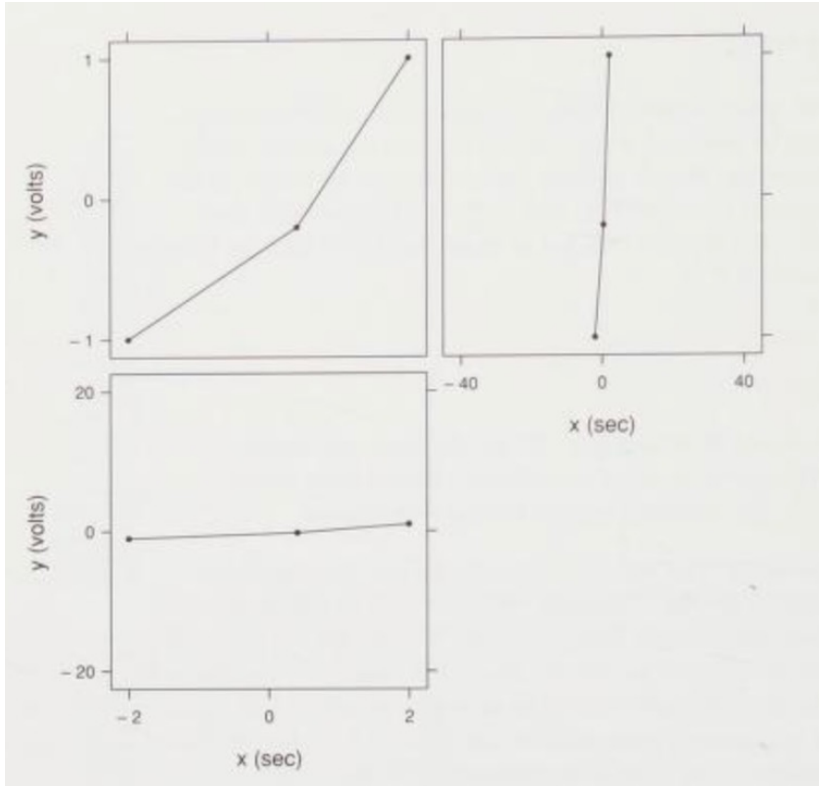
The same follows for multivariate data



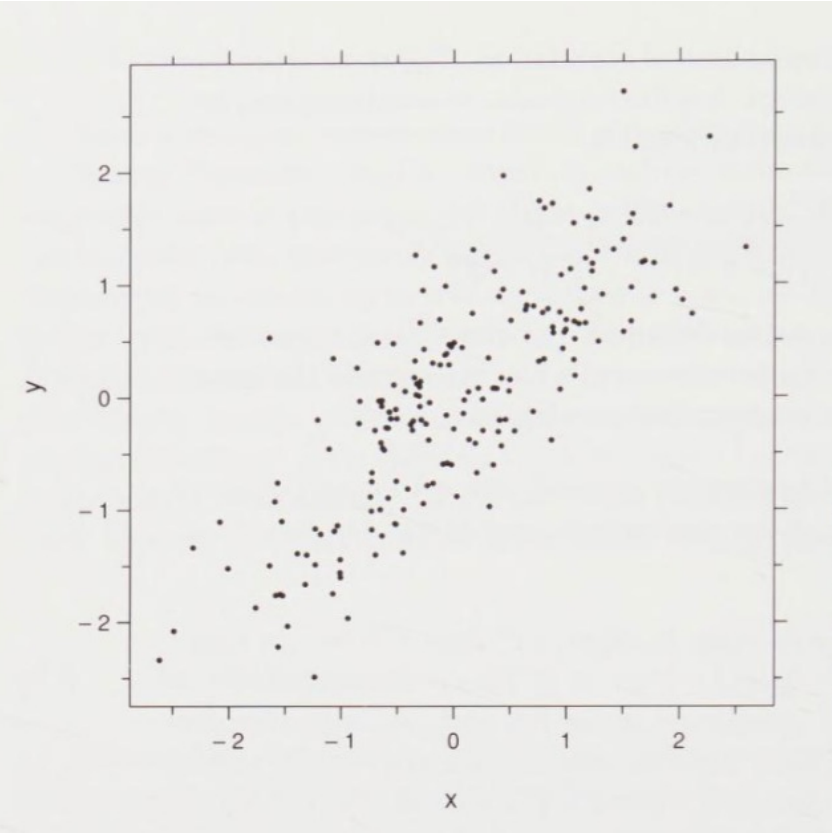


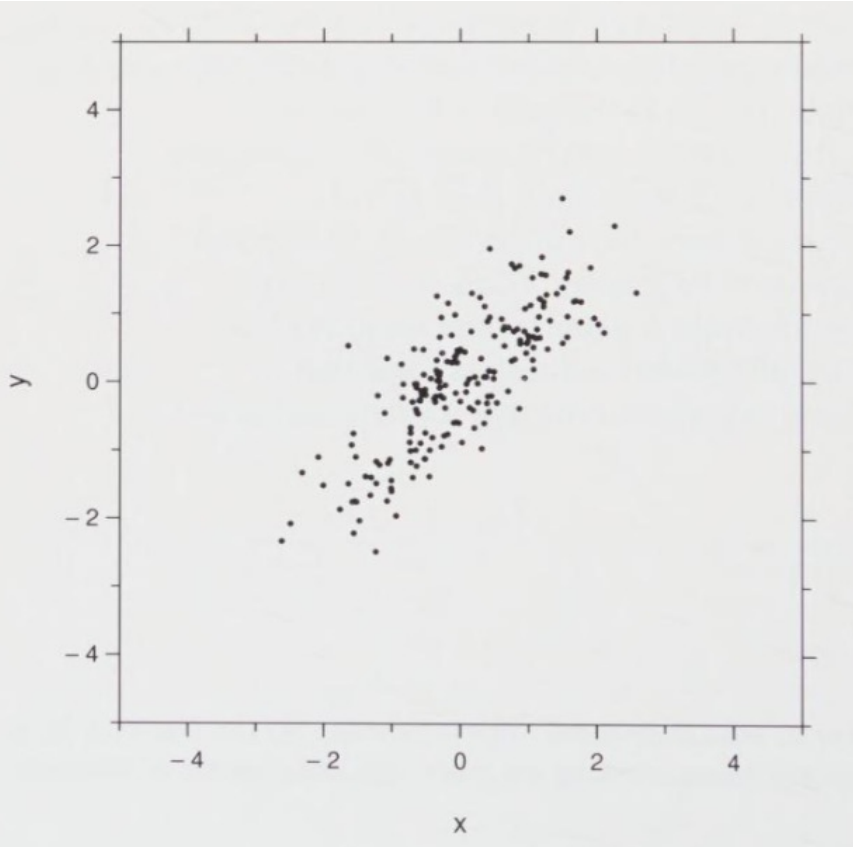
Aspect ratios should be set so that slopes average to be 45 degrees

- This allows us to optimally distinguish angular differences
- If you need to efficiently table look-up the slope, consider deriving and graphing the slope directly
  - Steps to find the slope:
    - scan horizontally for the two endpoints
    - interpolate to get two scale values
    - do mental arithmetic to get difference
    - do the same for the horizontal difference
    - mentally divide to get slope

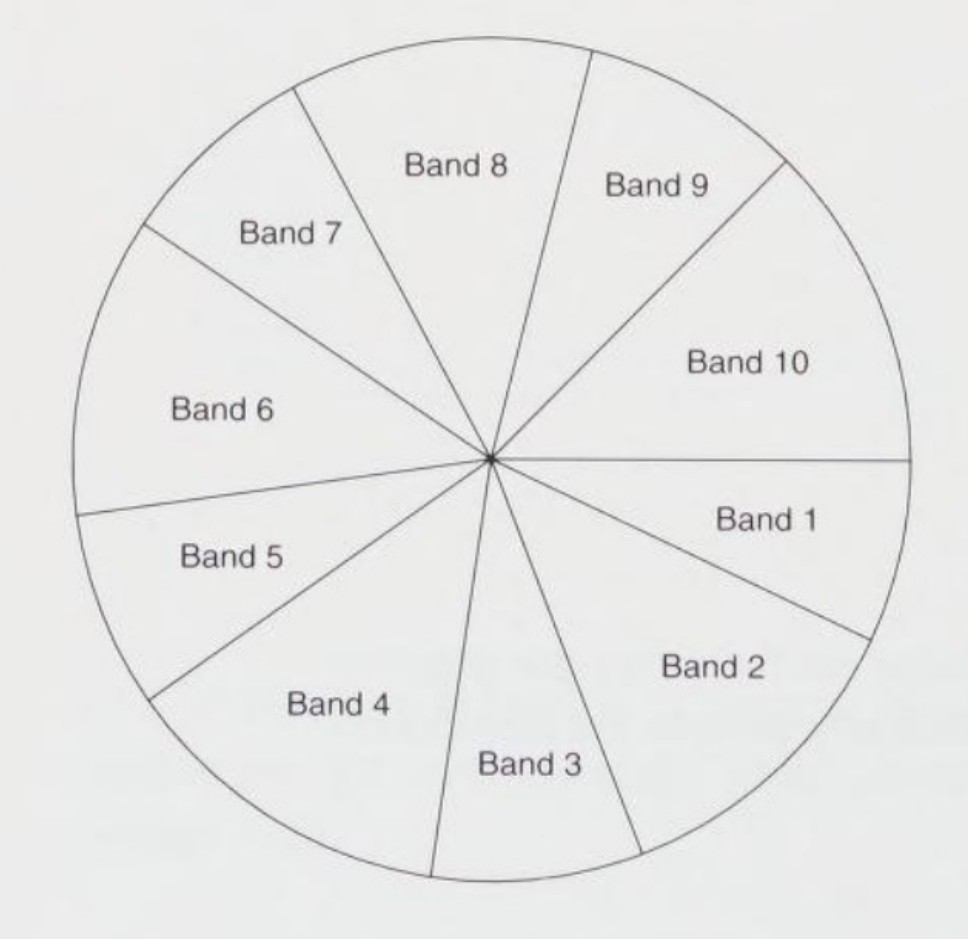


Correlation – when comparing, make sure that the ratio of the data rectangles to the scale-line rectangles are equal



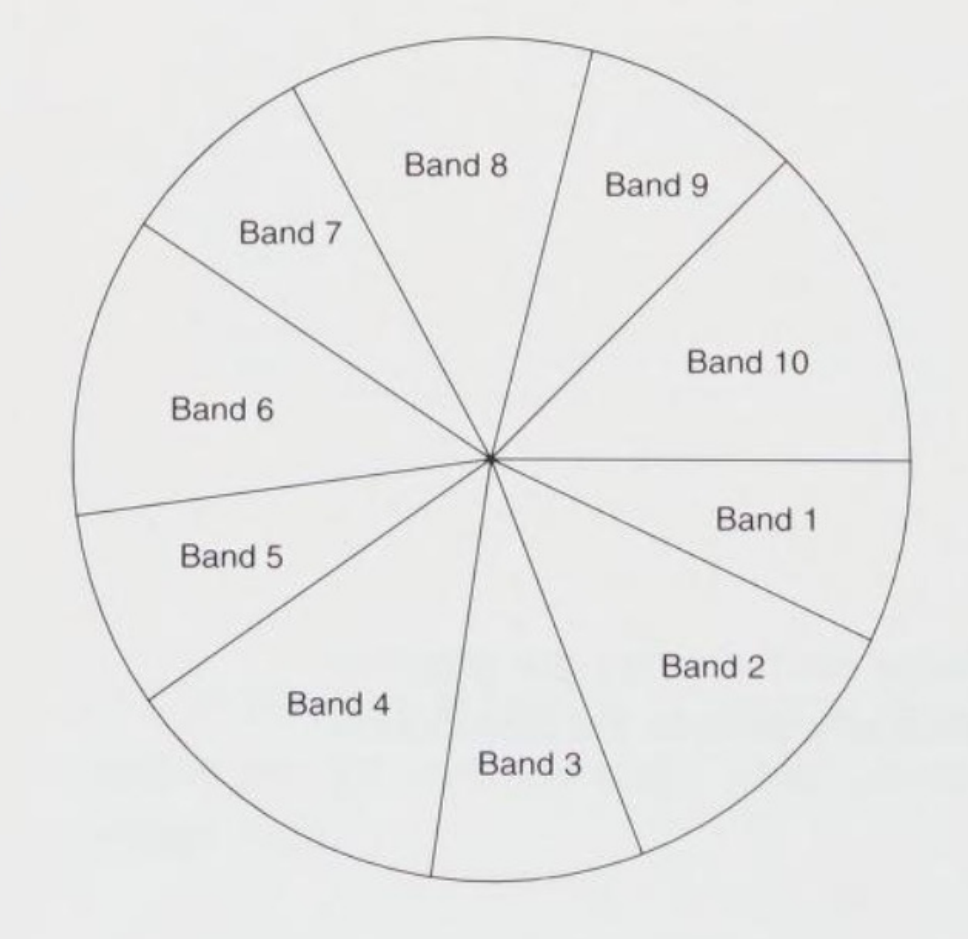


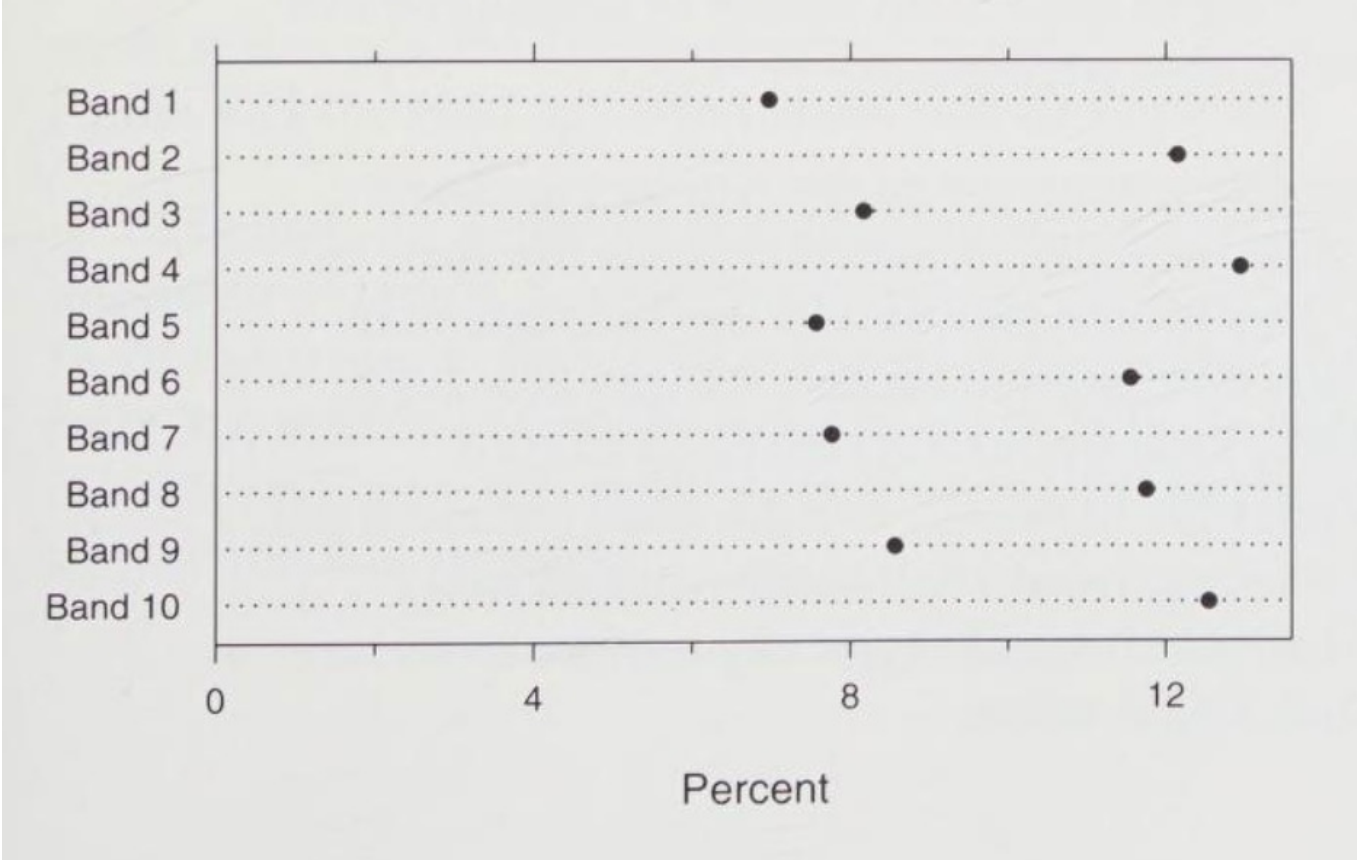
We just aren’t as good at estimating areas



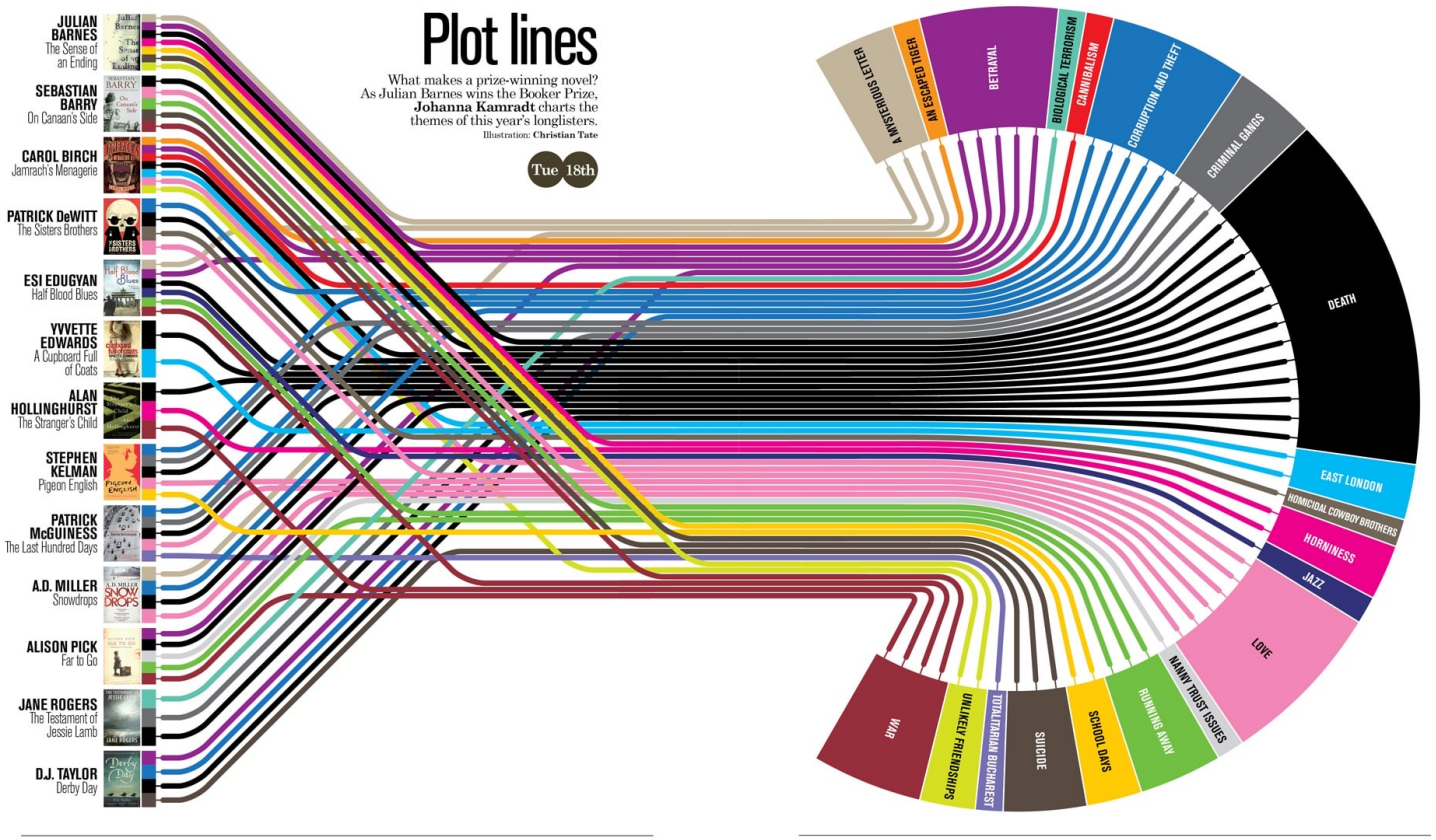
- Who are the top five bands?
- How much larger is #1 than #5?

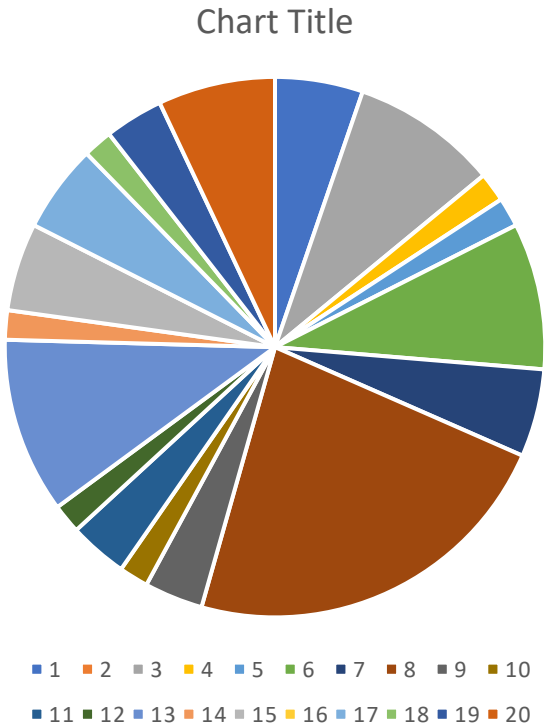
We just aren’t as good at estimating areas



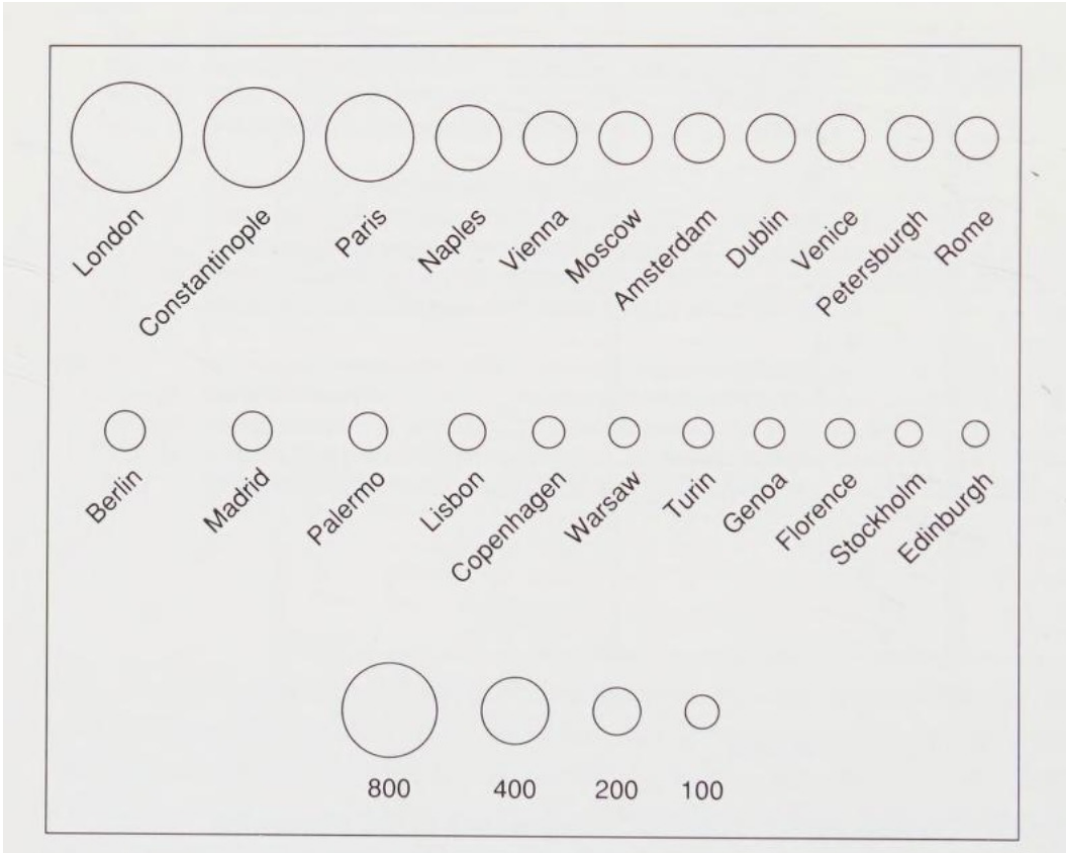


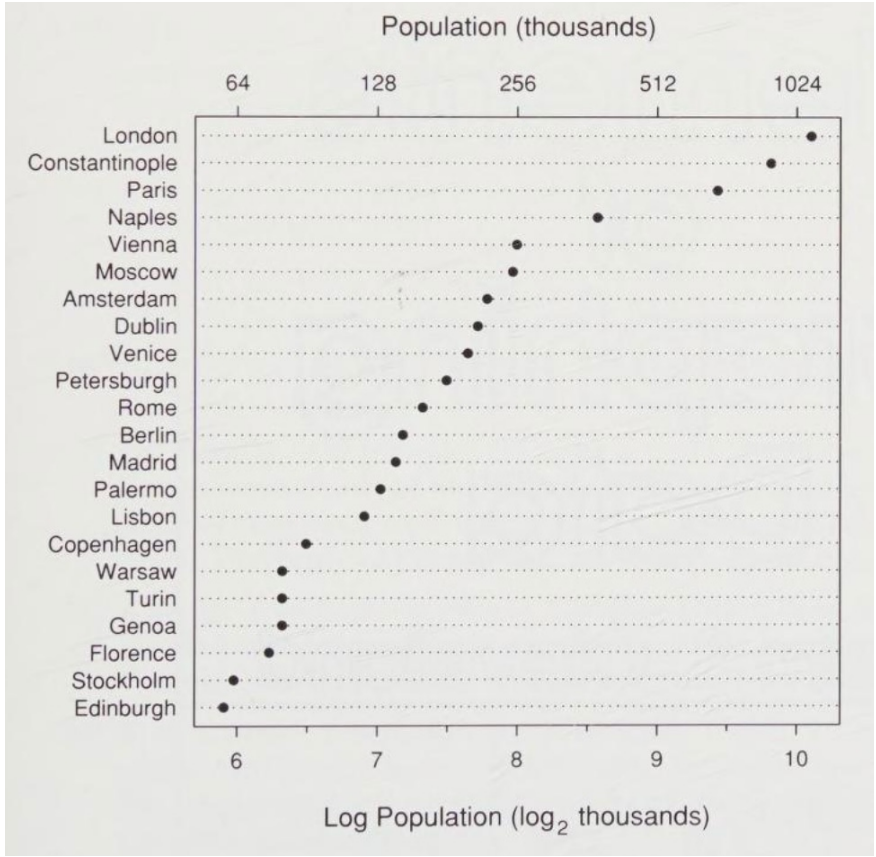
We just aren’t as good at estimating areas



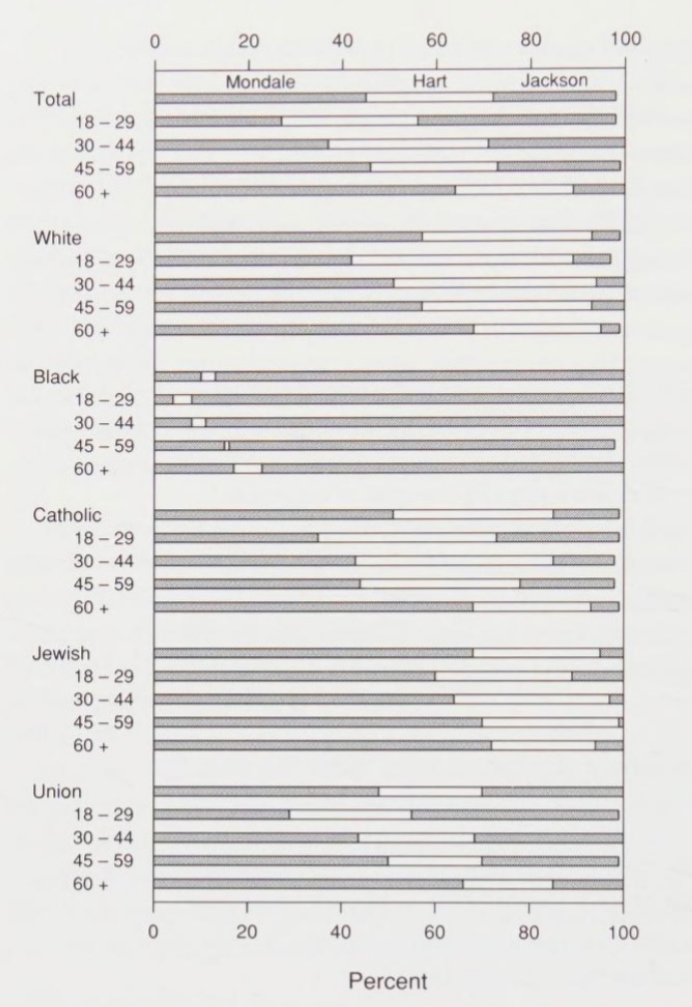


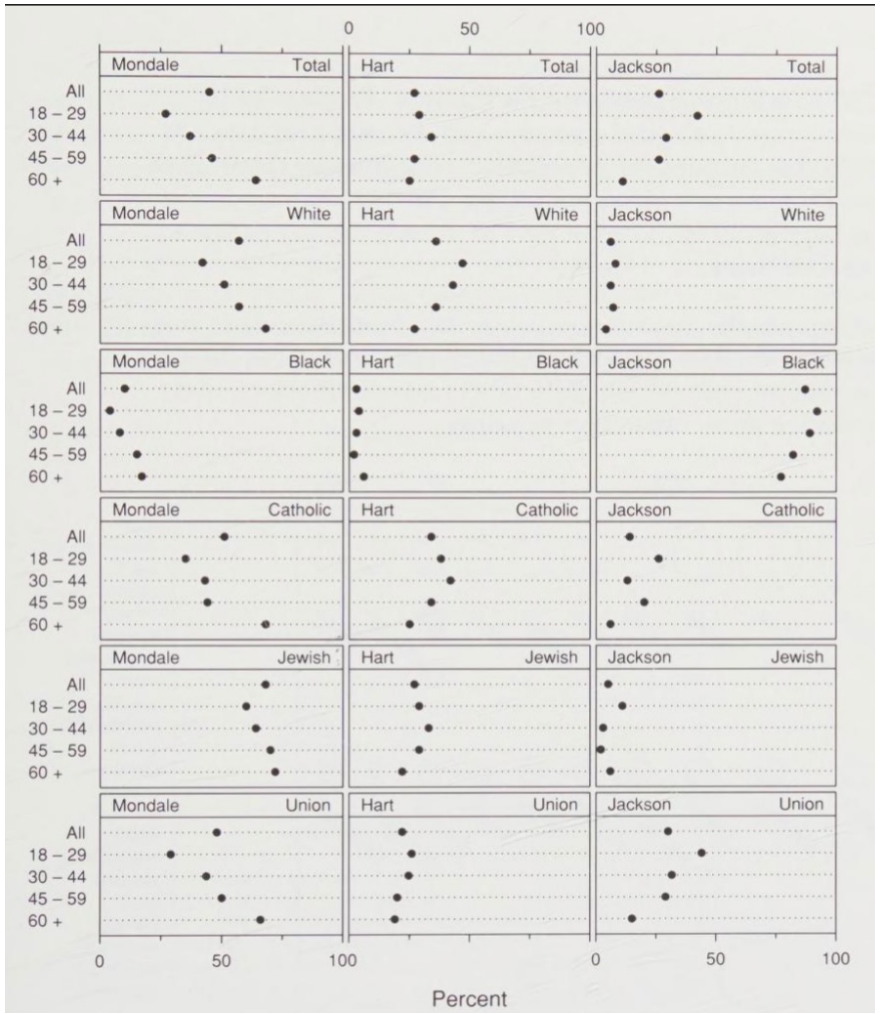
We just aren't as good at estimating areas



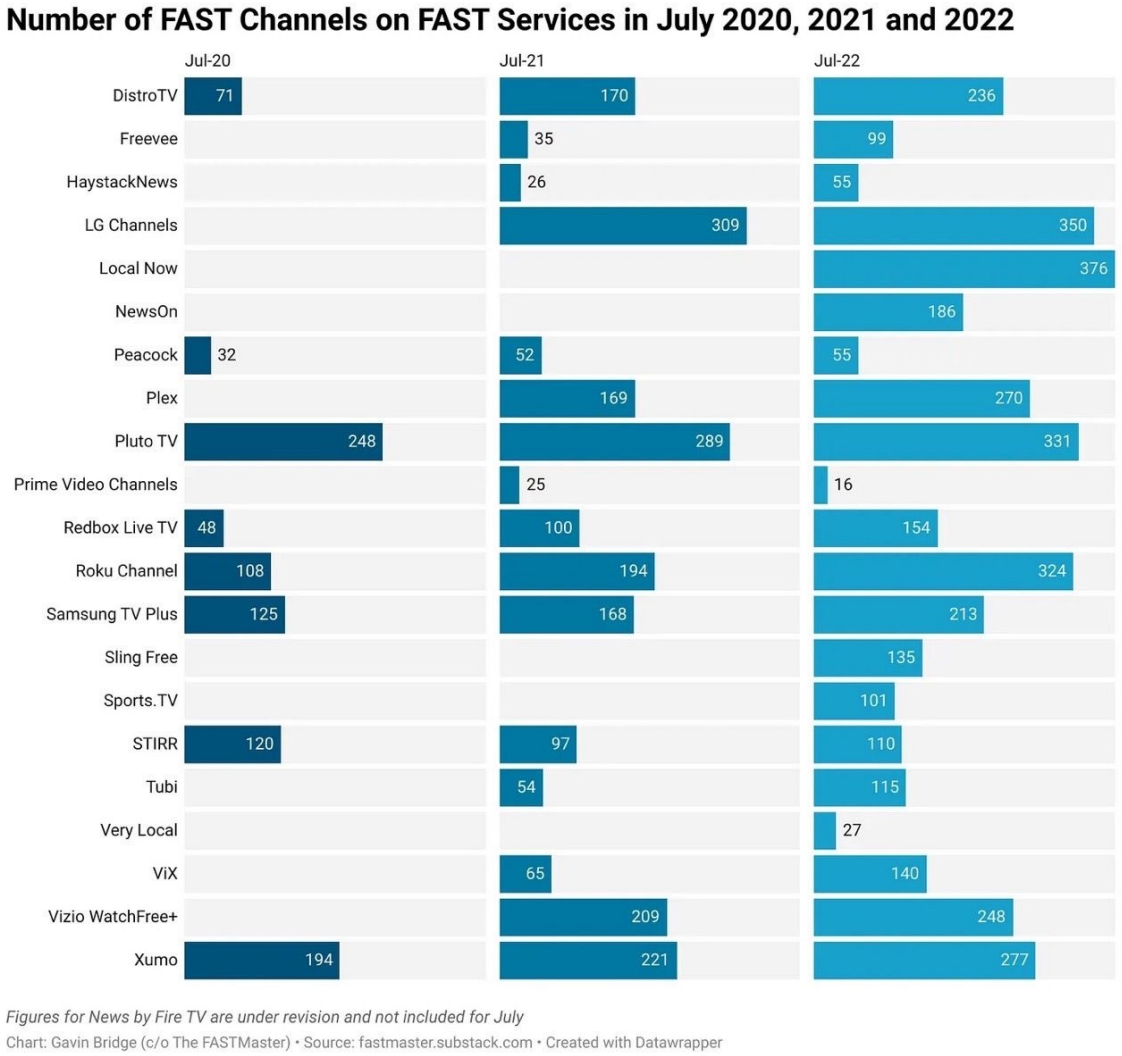


We aren’t as perceptive with stacked bar charts as we are with plots that have positions along common scales.





We aren’t as perceptive with stacked bar charts as we are with plots that have positions along common scales.



Next time:

We explore these analysis and visualization points
for
univariate data