GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

Classes teòriques:
Teoria 3, Teoria 6, Teoria 7, Teoria 9.

Alvaro Corral Cano
alvaro.corral@uab.cat

Judit Chamorro Servent
judit.chamorro@uab.cat

Departament de Matemàtiques

Data processing for visualization

- **Chapter 3 (today) - Data processing for visualization (I)**
  - Uncertainty and error
  - Transformations and data massage (seminars)
- **Chapter 6 (11/03) - Data processing for visualization (II)**
  - Dimensionality reduction
  - Computation and important metrics selection
- **Chapter 7 (21/03) - Advanced systems (I)**
  - Múltiples variables i múltiples dimensions
  - Xarxes
  - Camps de vectors
- **Chapter 9 (08/04) - Advanced systems (II)**
  - Dades 3D
  - Visualització científica
  - Mapes

GRAU EN ENGINYERIA DE DADES

104365 Visualització de Dades

Teoria 3. Tractament de dades I

3. Data processing for visualization (I). Contents:

1. Introduction of Visualizing errors & uncertainty

2. Error

3. Uncertainty

4. Transformation and data massage (mostly in seminar 4 & practices)

3. Visualizing errors & uncertainty. Contents:

1. Introduction of Visualizing errors & uncertainty

2. Error

1. Introduction
2. Residual error (absolute error, square error, percentage error)
3. Error (graded) bars and confidence bands
4. Systematic errors & random errors
5. Statistical concepts for visualization errors (descriptive analysis/distributions)
6. Visualizing random errors

3. Uncertainty

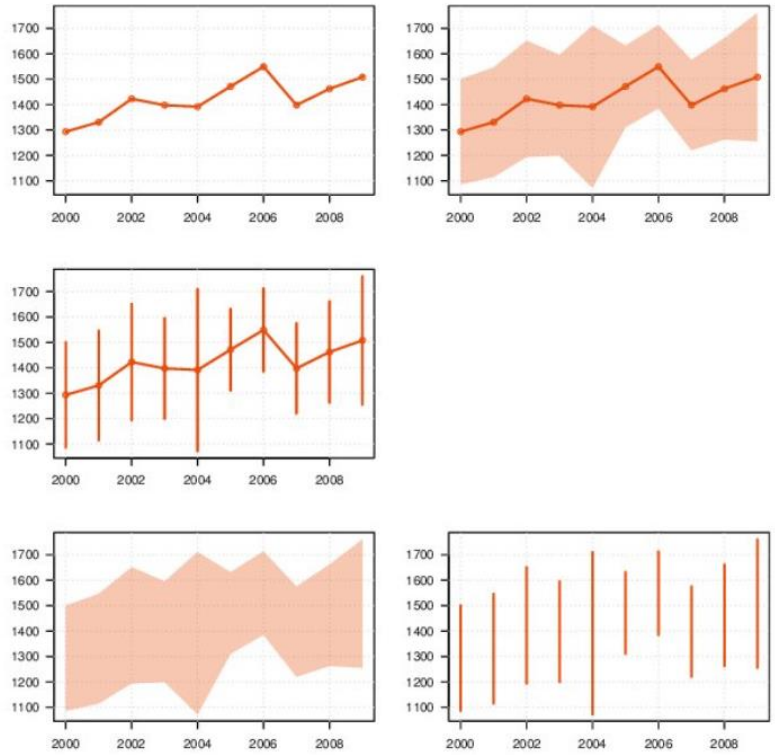
1. Introduction
2. Uncertainty visualization: Confidence bands. Frequency framing. Standard Error.
3. Dynamic uncertainty visualization: Curve fits and Hypothetical outcome plots
4. Bayesian tools to determine distributions (Monte Carlo simulation). And to normalize them (Central Limit Theorem)

3.1 Introduction: **Visualizing errors**

Effect of Displaying Uncertainty in Line and Bar Charts – Presentation and Interpretation

Article · January 2015

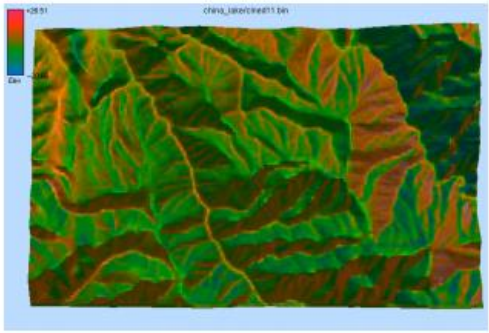
DOI: 10.5220/0005300702250232



Five types of line charts: *line, ribbon+line, error bar+line, ribbon, error bar*

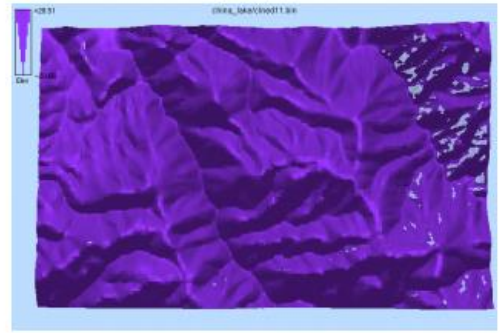
Edwin de Jonge

3.1 Introduction: Visualizing errors



(a)

(b)



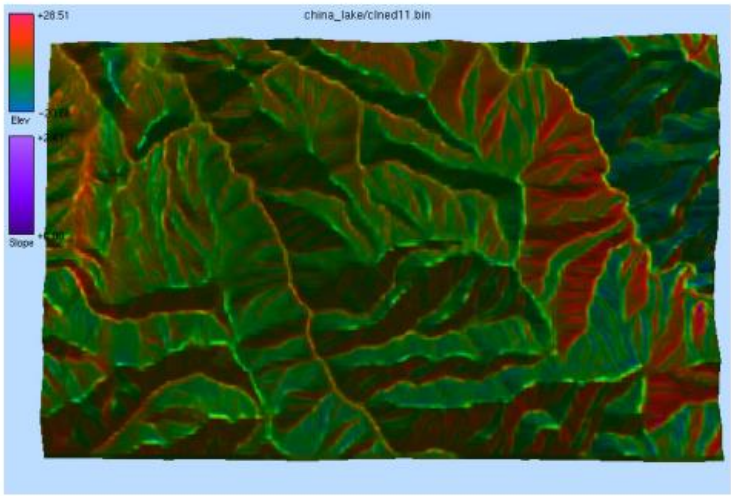
(c)

VisTRE: A Visualization Tool to Evaluate Errors in Terrain Representation

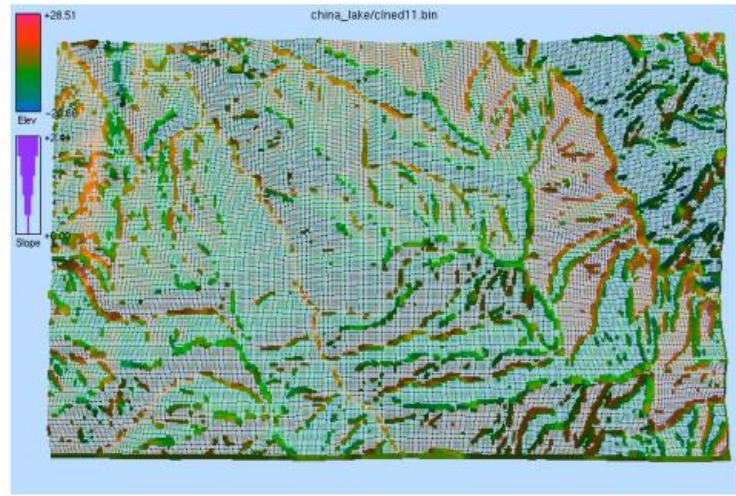
Christopher G. Healey
Computer Science Department
North Carolina State University
Raleigh, NC 27695-8206, USA

Jack Snoeyink
Computer Science Department
UNC Chapel Hill
Chapel Hill, NC 27599-3175, USA

Elevation error



(d)



(e)

Elevation error combined with Slope error

Figure 2: Examples of different visual features used to represent error values: (a) elevation error from the national elevation dataset (NED) terrain model at 1° resolution visualized using hue; (b) elevation error visualized using luminance; (c) elevation error visualized using size; (d) NED terrain model with elevation error visualized using hue and slope error visualized using luminance; (e) elevation error visualized using hue and slope error visualized using size

Healey & Snoeyink

3.1 Introduction: **Visualizing errors**

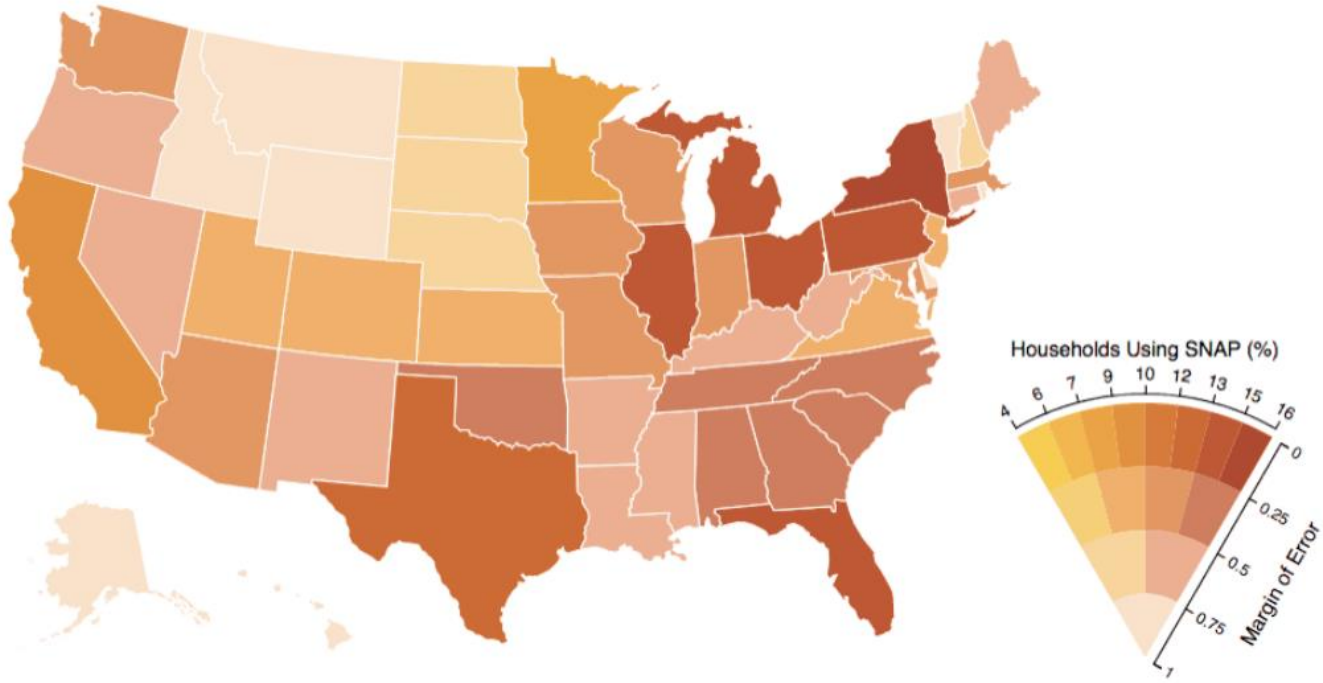
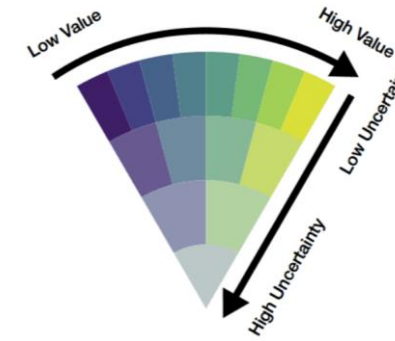


Fig. 9: A map combining measurements by region and showing the margin of error in the measurement using ranges of color saturation.

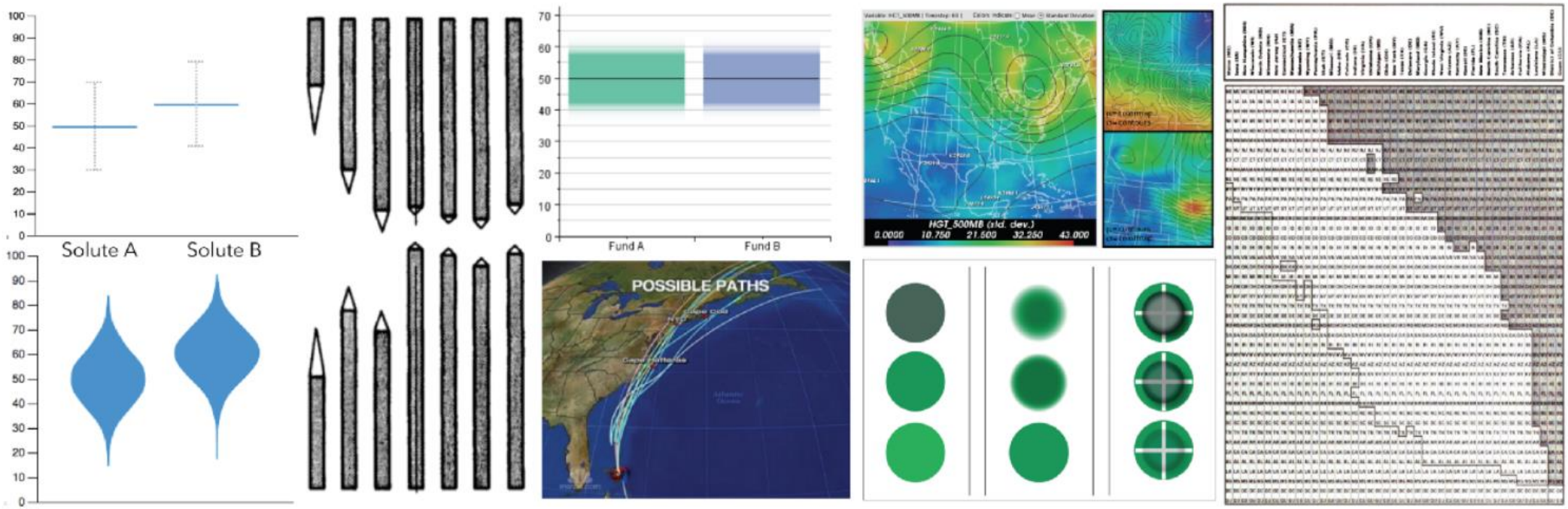


Also useful for uncertainty

Color saturation

Robert Falkowitz, 2020

3.1 Introduction: **Visualizing uncertainty**



Techniques for visualizing uncertainty (clockwise from top left): error bars, using negative space to convey confidence intervals on random variables, gradient plots, ensemble visualization, a matrix in which three different shades convey the reliability of precomputed comparisons, visual encodings like saturation, fuzziness, and transparency, possible hurricane paths on a news weather forecast, violin plots.

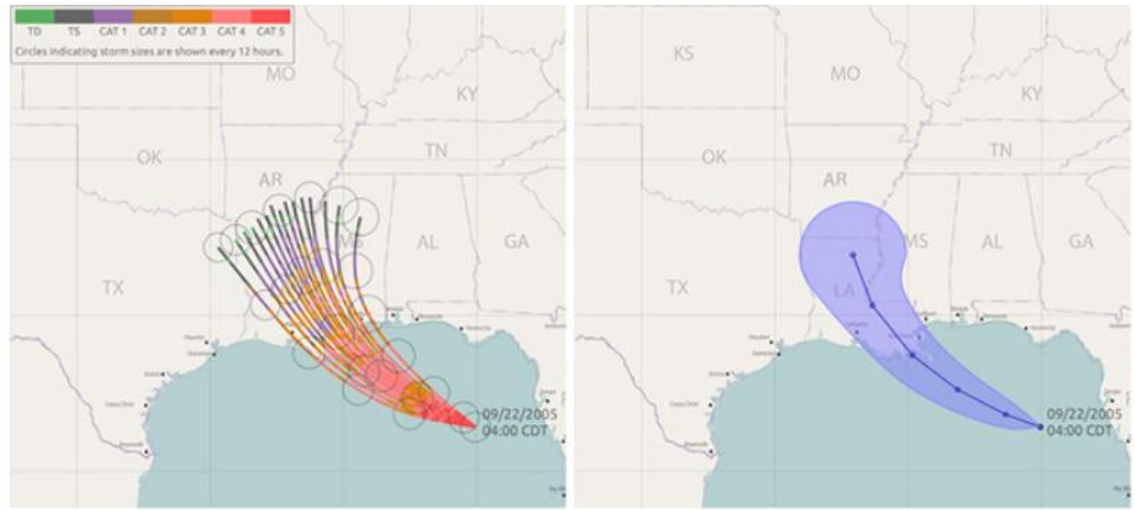
https://medium.com/hci-design-at-uw/hypothetical-outcomes-plots-experiencing-the-uncertain-b9ea60d7c740

3.1 Introduction: **Visualizing uncertainty**

How data visualizations can clarify and confound uncertainty

Jessica Hullman's Scientific American article weighs pros and cons of common data visualizations

OCT 15, 2019



Two approaches to visualizing uncertainty in hurricane paths. From Liu, Padilla, Creem-Regehr, and House. Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks (2019)

https://www.mccormick.northwestern.edu/computer-science/news-events/news/articles/2019/how-data-visualizations-can-clarify-and-confound-uncertainty.html

Spaghetti plot showing an ensemble of predictions

!!! Cone of uncertainty is NOT the area in which the hurricane may lie.

The cone is intended to indicate that the storm may take a multitude of paths, and that *it is harder to predict its location further into the future.*

3.1 Introduction: **Visualizing uncertainty**

A visual representation of temperature forecast uncertainty in the news. Visualizations for the general public (e.g., on TV or in newspapers) often provide **no explicit information on the uncertainty that is shown, and the viewer needs to adopt a model to interpret the underlying distribution.**

Alexander Toet, 2016

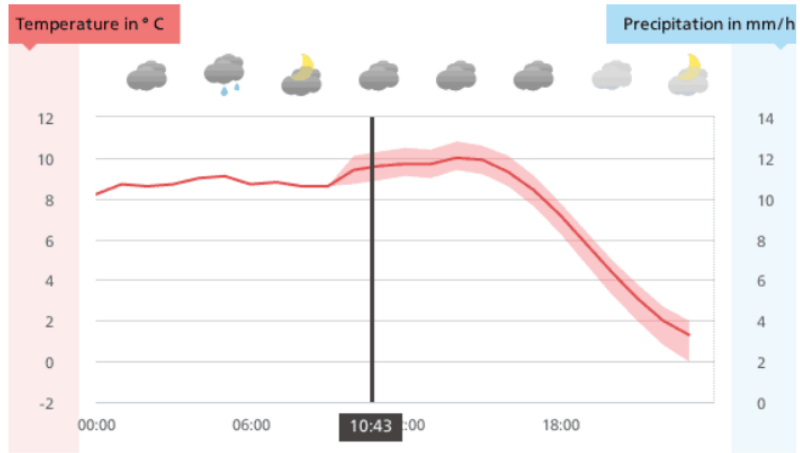


Fig. 13: Using shading to indicate a range of probable future measurements

Robert Falkowitz, 2020

3.1. Introduction: **Revealing Error & Uncertainty**

Revealing errors and uncertainty:

If you have information about the errors or uncertainty present in your data, whether it be

- **from a model**
- or **from distributional assumptions**,

it is a good idea display it.

3.2.1. Introduction: **Potential sources of ERROR**

In statistics, the entire set of raw data that you may have available for a test or experiment is known as the **population**.
Statistics allows us to take a **sample**, perform some computations on that set of data.

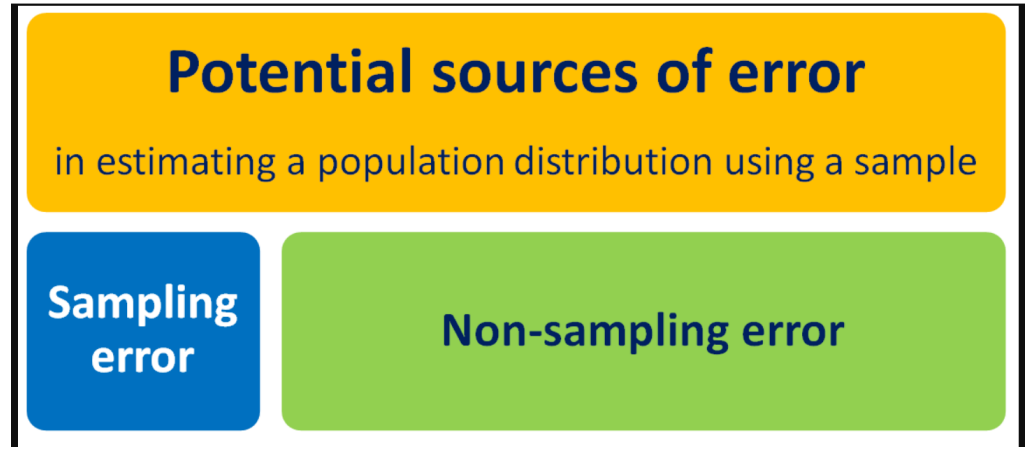


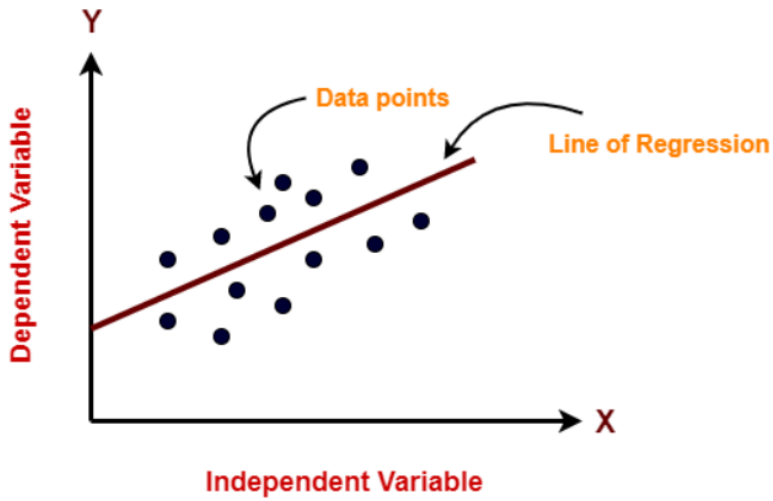
Image from Creative maths

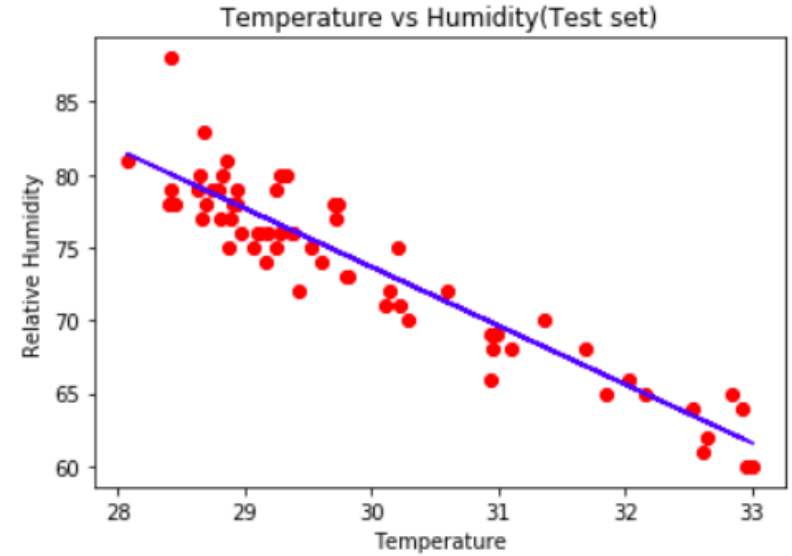
Using probability and some assumptions we can **with a certain degree of certainty** understand trends for the entire population or predict future events.

3.2.1. Introduction: **Example – quality of a model**

Example: linear regression – *used in supervised machine learning.*

As the name suggests, linear regression follows the **linear mathematical model for determining the value of one dependent variable from value of one given independent variable.**

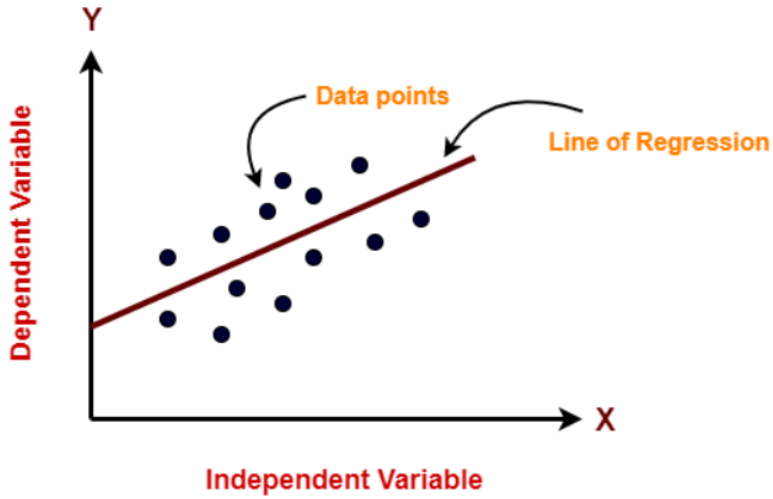




3.2.1. Introduction: **Example – quality of a model**

Example: linear regression – *used in supervised machine learning.*

As the name suggests, linear regression follows the linear mathematical model for determining the value of one dependent variable from value of one given independent variable. **Do you remember the linear equation from school?**



$$y = ax$$

where *y* is the dependent variable (usually quantitative), *a* is the slope, *x* is the independent variable.

If the line does not cross the origin (0,0) we can have *y=c+ax*.

3.2.1. Introduction: **Example – quality of a model**

Example: to judge the quality of a model and enable us **to compare regressions** against other regressions with different parameters, **error metrics can help.**

Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output — Coefficients — Input — Error

Linear regression

Linear Regression: Multiple Variables

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

3.2.1. Introduction: **Error metrics** Example - regression model

Error metrics will be able **to judge the differences between predictions and actual values.**

!!!But we cannot know how much the error has contributed to the discrepancy

The **fitted** (**or predicted**) values are the $\hat{y}$-values that you would expect for the given ***x***-values according to the built regression model (or visually, the best-fitting straight regression line).

The ***quality*** **of a regression model** is how well its predictions ($\hat{y}$) match up against actual values (x).

Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output — Coefficients — Input — Error

Linear Regression: Multiple Variables

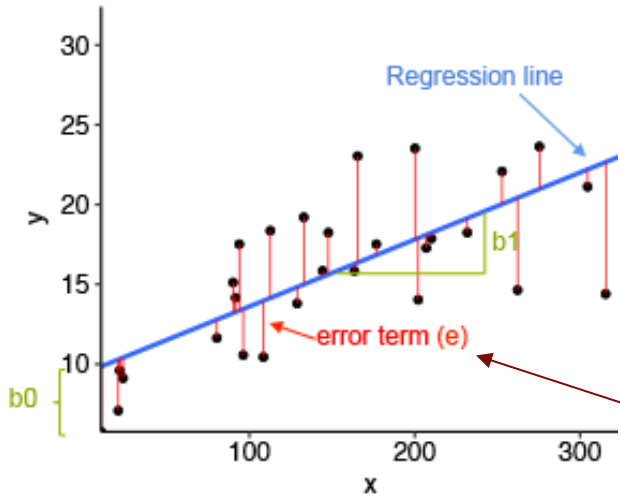
$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

3.2.2. Residual Error. Example - regression model

Error metrics will be able to judge the differences between predictions and actual values. *!!! But we cannot know how much the error has contributed to the discrepancy*

The *quality* of a regression model is how well its predictions match up against actual values.

Residual error: the difference between the actual value and the model's estimate.



Linear Regression: Single Variable

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

Predicted output · Coefficients · Input · Error

Linear Regression: Multiple Variables

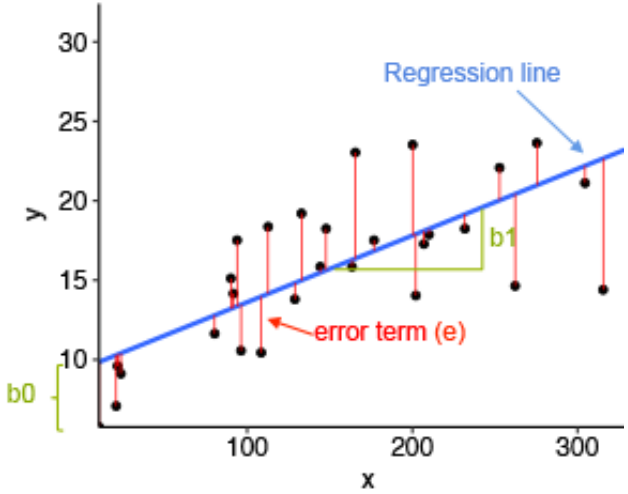
$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

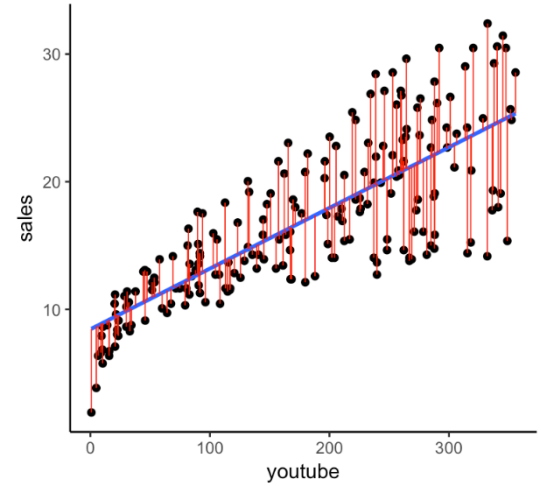
(e does not refer to the ∈)

3.2.2. Introduction: **Example – quality of a model**

Error metrics – Example: to judge the quality of a model and enable us **to compare regressions** against other regressions with different parameters

The *quality* of a regression model is how well its predictions match up against actual values.





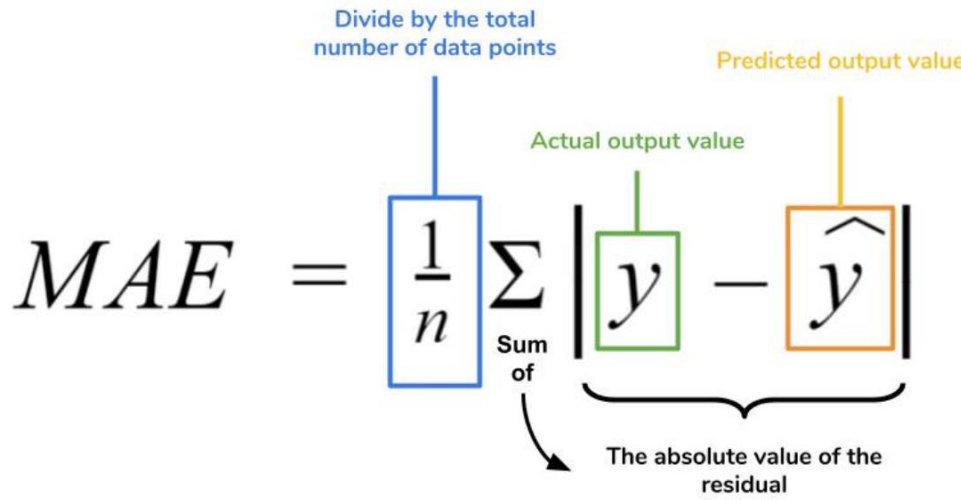
We build a model to predict sales on the basis of advertising budget spent in *youtube* medias

To check the regression assumptions, we'll examine the distribution residuals

3.2.2. Residual Error: Mean absolute error (MAE)

Mean absolute error (MAE): is the simplest regression error metric. It **describes** the typical **magnitude of the residuals**

1. We'll calculate the residual for every data point, taking only the absolute value of each (negative and positive residuals do not cancel out).
2. Take the average of all the residuals.

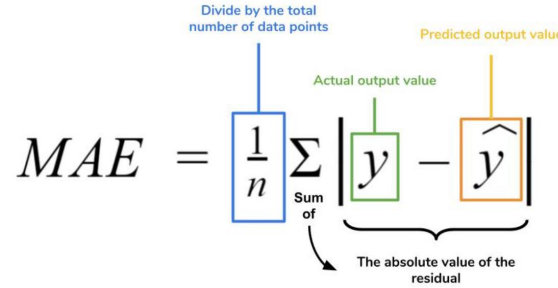


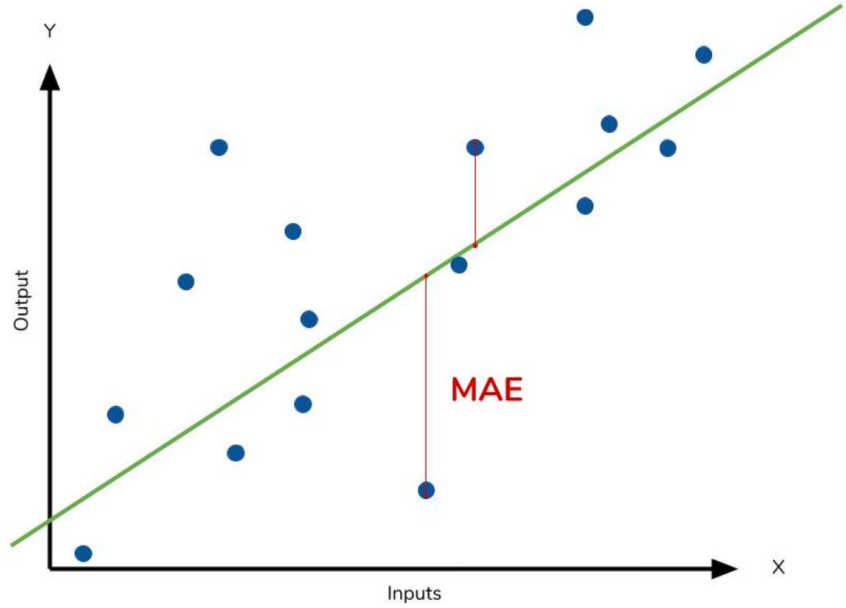
3.2.2. Residual Error: Mean absolute error (MAE)

Limitations of MAE:

- It does not indicate underperformance or overperformance of the model/experiment.

- It does not bring attention to the outliers (or extrema values)





Green line represents the model's predictions

Blue points represent the data

3.2.2. Residual Error: Mean square error (MSE)

Mean square error (MSE): is just like MAE, but squares the difference before summing them all instead of using the absolute value

$$MSE = \frac{1}{n}\sum\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

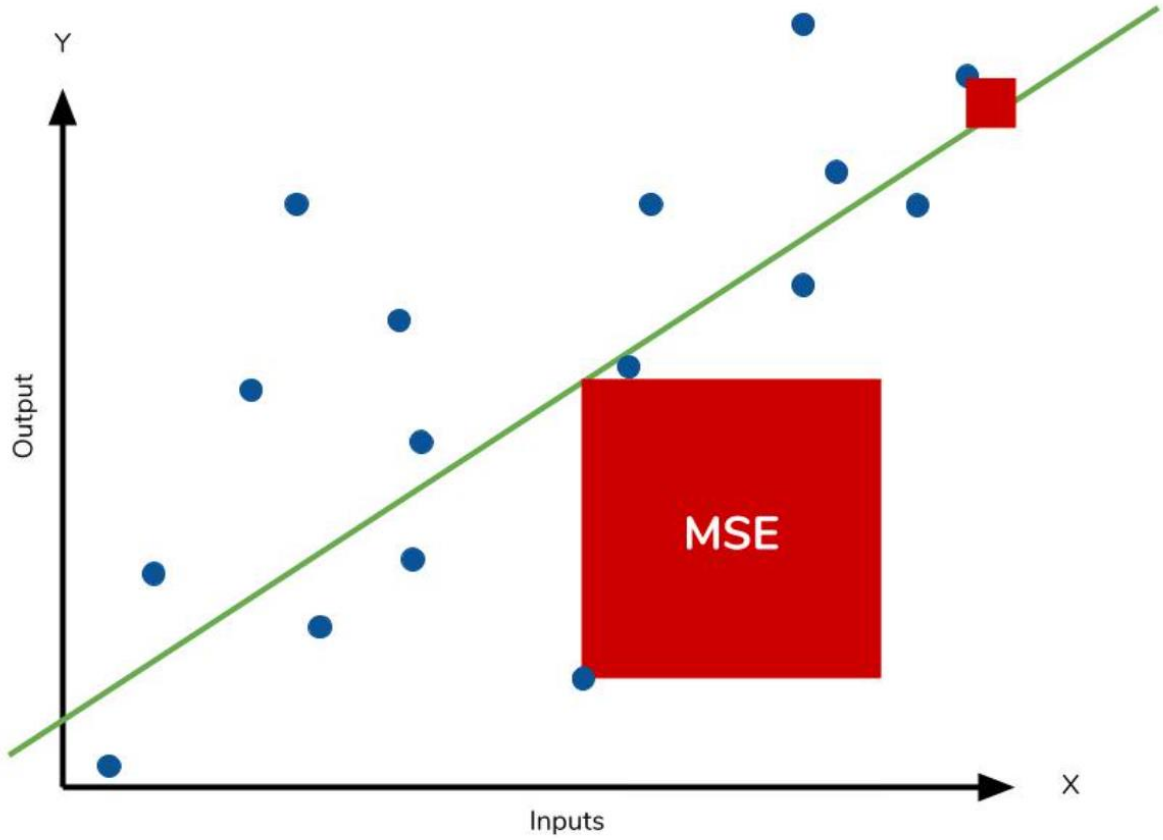
Because we are squaring the difference, the **MSE will almost always be bigger than the MAE** (we cannot compare MAE to the MSE)

The effect of the square term is most apparent with the presence of outliers in our data – **Outliers in our data will contribute to much higher total error in MSE than they would the MAE.**

3.2.2. Residual Error: Mean square error (MSE)

Mean square error (MSE):

!!! Outliers will produce these exponentially larger differences



Green line represents the model's predictions

Blue points represent the data

3.2.2. Residual Error: MAE vs MSE

MAE or MSE??

- To downplay **the outlier's significance**, we would use the MAE since the outliers' residuals won't contribute as much to the total error as MSE.
- **The choice** between MSE and MAE **is application-specific** and depends on how you want treat large errors.
- Both **MAE and MSE can range from 0 to positive infinity**, so as both measures get higher, it becomes harder to interpret how well your model is performing.

3.2.2. Residual Error: Root mean squared error (RMSE)

Root mean squared error (RMSE) is the square root of the MSE.

MSE vs RMSE:

- RMSE is **often used to convert the error metric back into similar units**, making interpretation easier.
- **The effect of the outliers** in MSE and RMSE is **similar**. They both square residual.
- The **RMSE is analogous to the standard deviation** (MSE variance) and measure of how large your residuals are spread out.

3.2.2. Residual Error: Mean percentatge error (MPE)

Mean percentage error (MPE): it lacks the absolute value operation

$$MPE = \frac{100\%}{n} \Sigma \left(\frac{y - \hat{y}}{y} \right)$$

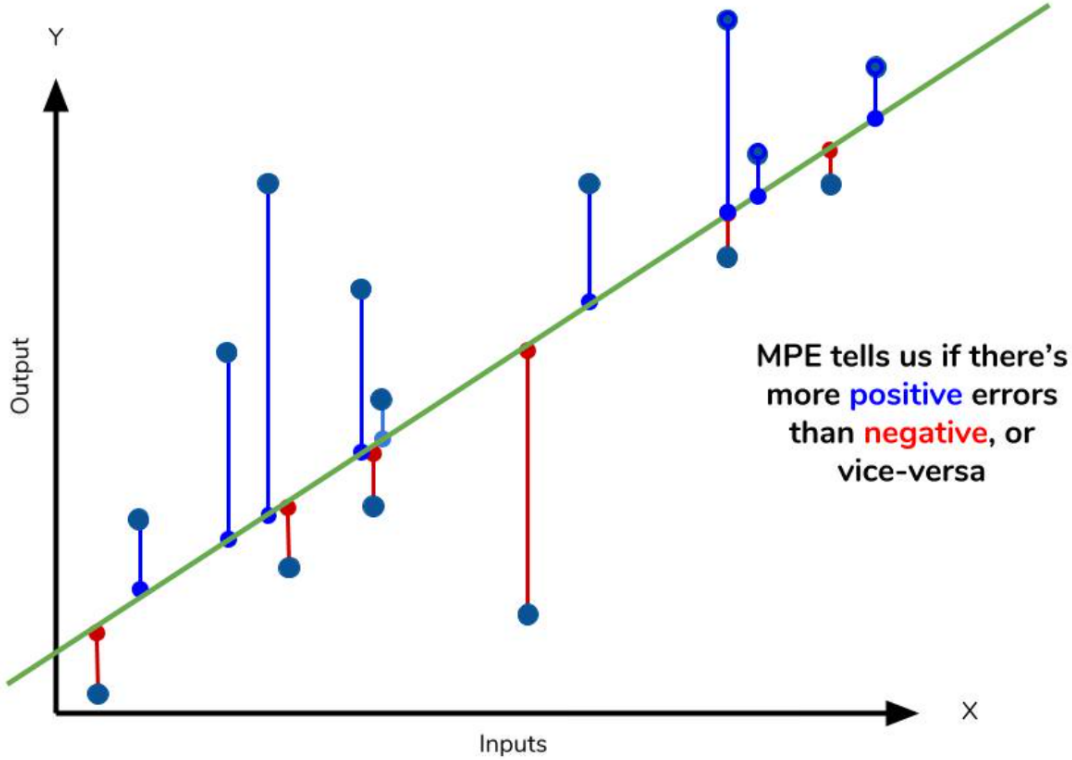
However, unlike MAE and MPE, MPE **allows us to see if our model systematically underestimates or overestimates** (bias).

Underestimates -> MPE more negative

Overestimates -> MPE positive

3.2.2. Residual Error: Mean percentatge error (MPE)

Mean percentage error (MPE):

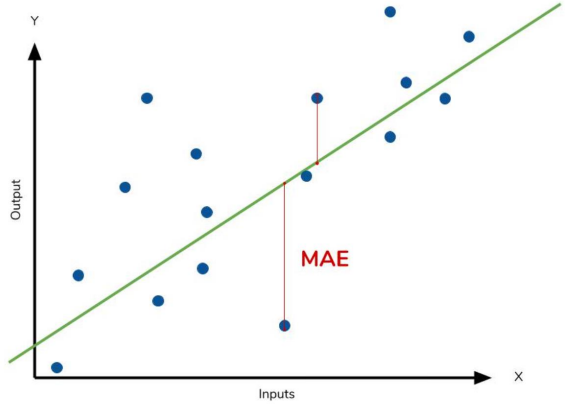


Green line represents the model's predictions

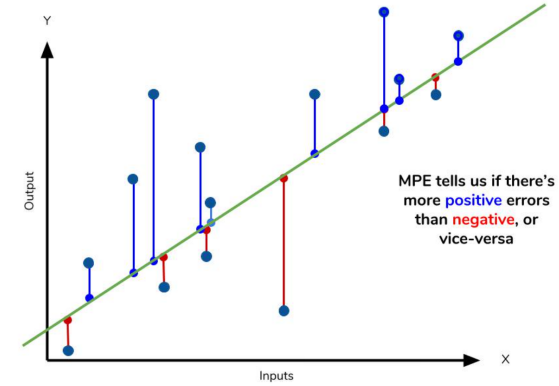
Blue points represent the data

3.2.2. Residual Error – Metric's summary

Acroynm	Full Name	Residual Operation?	Robust To Outliers?
MAE	Mean Absolute Error	Absolute Value	Yes
MSE	Mean Squared Error	Square	No
RMSE	Root Mean Squared Error	Square	No
MPE	Mean Percentage Error	N/A	Yes

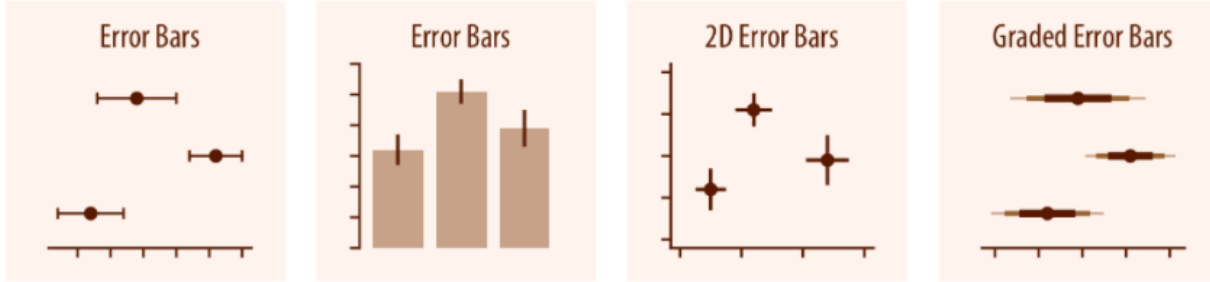


Be specific



To check the regression assumptions, we'll examine the distribution residuals.

5.3.2 Uncertainty visualization: **Error bars**

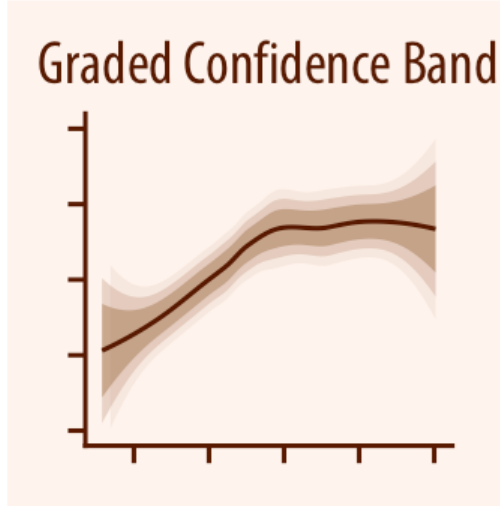


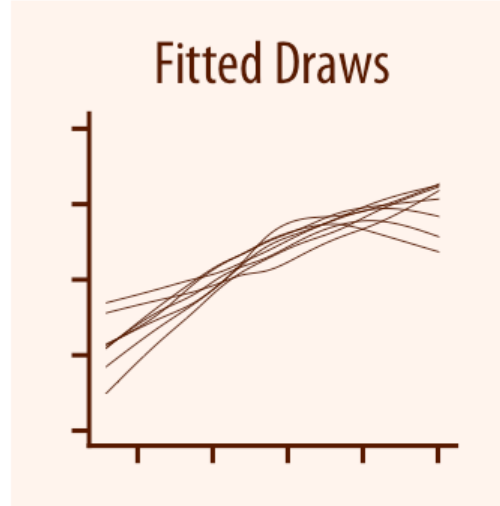
Claus Wilke

- **Error bars are meant to indicate the range of likely values for some estimate or measurement.** They extend horizontally and/or vertically from some reference point representing the estimate or measurement.
- **Graded error bars** show multiple ranges at the same time, where each range corresponds to a different degree of confidence. They are in effect multiple error bars with different line thicknesses plotted on top of each other.

3.2.3 Visualizing errors (confidence bands)

Confidence band: the equivalent of an error bar for smooth line graphs.



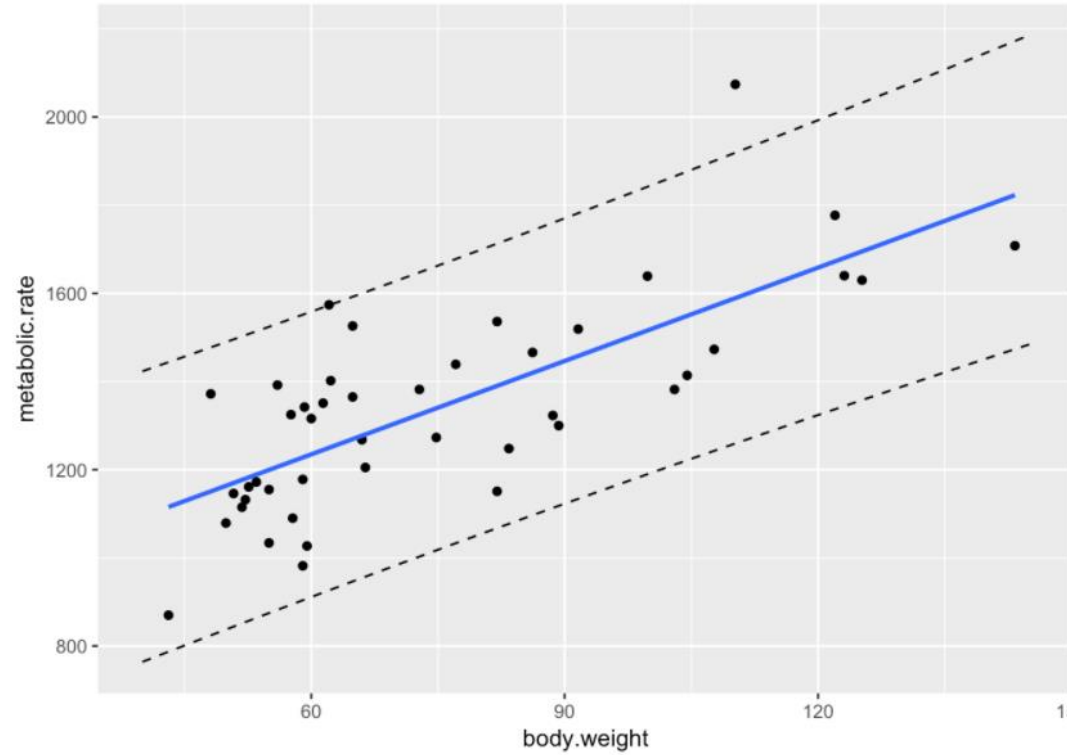


Claus O.Wilke

3.2.3. Confidence bands

Prediction Intervals - Suppose you want to predict the metabolic rate of a new patient, whose body weight is known. How large is your error?

Regression coefficients give you an estimate. However, **to know how large is the error, you need to use prediction intervals.**

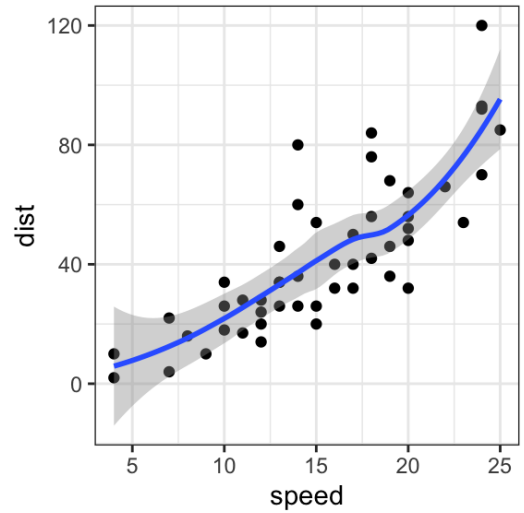


About 95% of the data points do fall within the bands

3.2.3. Confidence bands

- **One way to show uncertainty is to fit a smoothing model to the data and/or regression lines.**
- (seminars) GGPlot: Adding the canal **'geom_smooth' or 'stat_smooth()':** aids the eye in seeing patterns in the presence of **overplotting**. See ?geom_smooth or ?stat_smooth

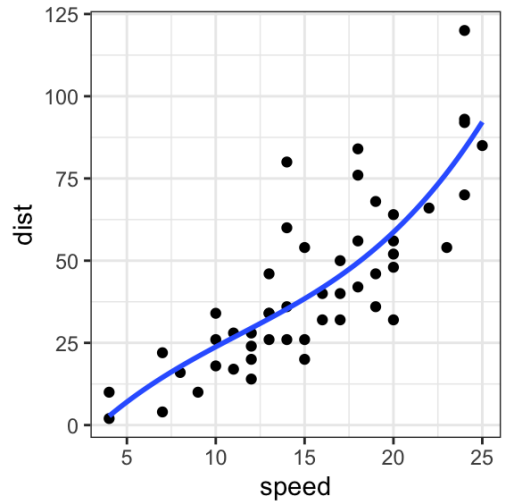
Example:



One can play with one of the arguments for stat-smooth which is level: level of confidence Interval to use (0.95 by default)

3.2.3. Confidence bands

- One way to show uncertainty is to fit a smoothing model to the data and/or regression lines.
- (seminars) GGPlot: Adding the canal **'geom_smooth' or 'stat_smooth()':** aids the eye in seeing patterns in the presence of **overplotting**. See ?geom_smooth or ?stat_smooth

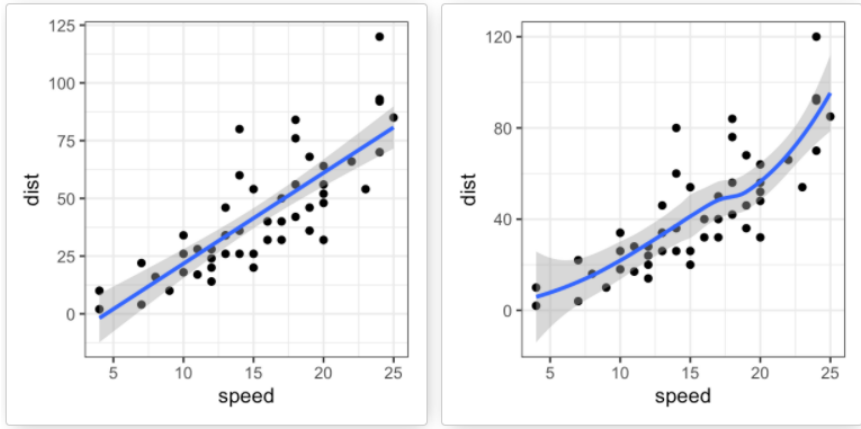


Polynomial interpolation

```
# Remove the confidence bande: se = FALSE
p + geom_smooth(method = "lm", formula = y ~ poly(x, 3), se = FALSE)
```

3.2.3. Confidence bands

- One way to show uncertainty is to fit a smoothing model to the data and/or regression lines.
- Adding the canal ‘geom_smooth’ or ‘stat_smooth()’: aids the eye in seeing patterns in the presence of overplotting. **By default, the method is: local regression fitting (“loess”)**




```
p <- ggplot(cars, aes(speed, dist)) +
  geom_point()
# Add regression line
p + geom_smooth(method = lm)

# loess method: local regression fitting
p + geom_smooth(method = "loess")
```

There is also the method “lm” (**linear regression**) & “glm” (**generalized linear model**)

3.2.4 Systematic vs. Random Errors

- **Systematic errors:**

– Errors that **affect all measurements in the same way**.

– Systematic errors **have a determinate origin** (**there is a cause**).

Example: Not calibrating the measuring instrument

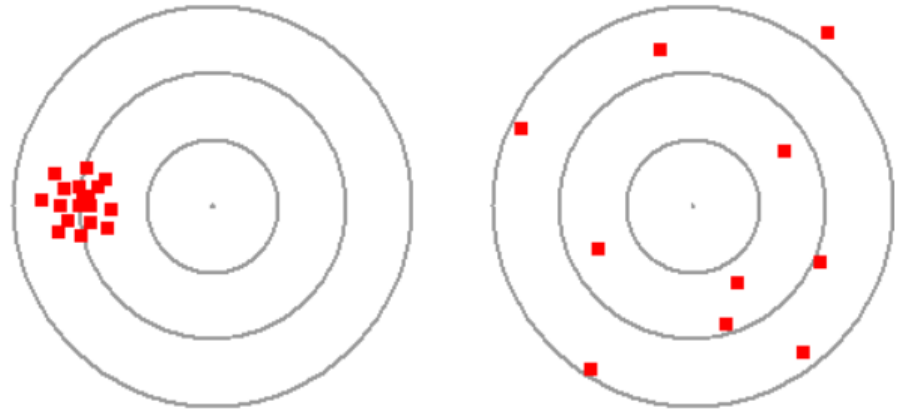
- **Random errors:**

– Errors that **occur randomly and affect measurements in an unpredictable manner**.

– **They are undetermined in origin and cause.** Random errors may occur due to carelessness or lack of concentration

3.2.4 Systematic vs. Random Errors

- **Systematic errors: Tend to be consistent in magnitude and/or direction** (the recorded values differ from the “true” values to be measured in a way that is both consistent and predictable).
- **Random errors: Vary in magnitude and direction**



3.2.4 Graphing: Systematic vs. Random Errors

– **Systematic errors:** Tend to be consistent in magnitude and/or direction. They have an origin/cause. Two types:

- ***Constant errors:*** the size of the error is often **independent of measurement magnitude (not correlated).**

It will be reflected in a *change in the 'y-axis'* intercept on the graph.

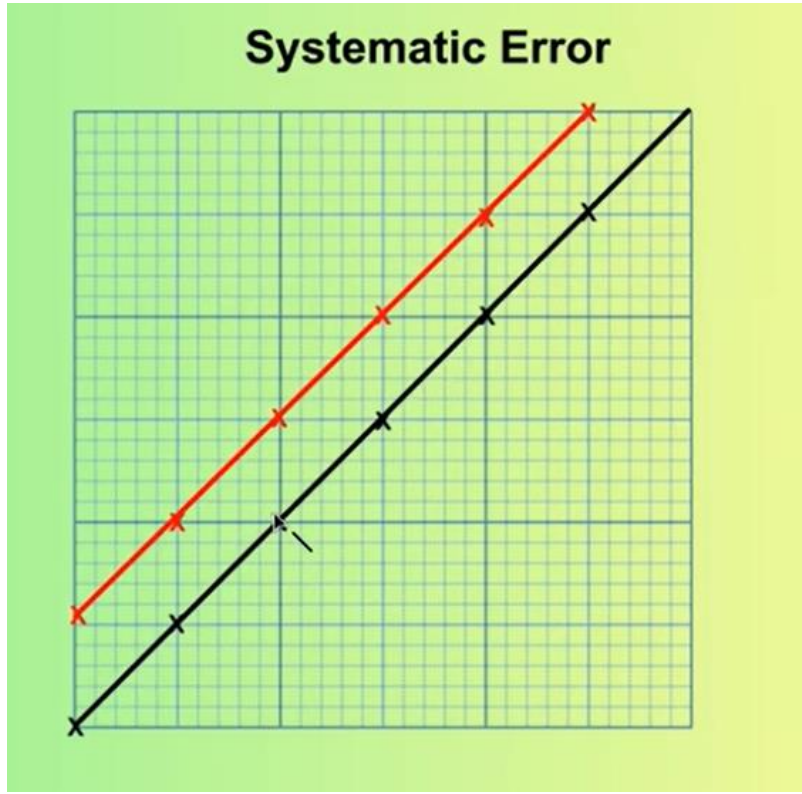
- ***Proportional errors:*** the error **may increase with the magnitude** of the measurement.

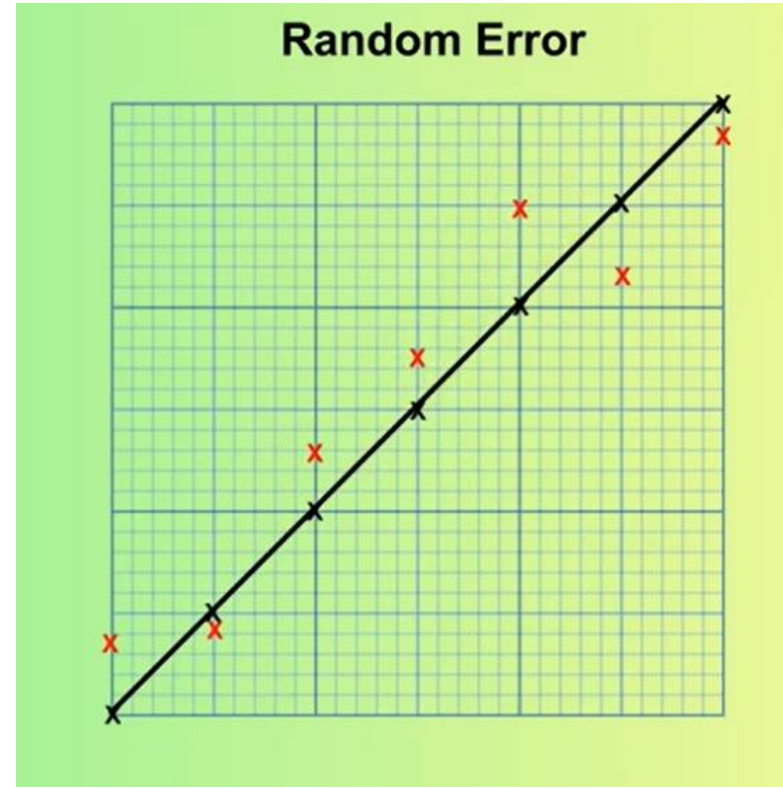
It will change *the slope* of the line of the graph.

– **Random errors: Vary in magnitude and direction**. Not origin/cause.

They will cause a *scatter plot effect* on the graph, *making the determination of the line of best fit impossible*.

3.2.4 Systematic vs. Random Errors





Scattered all around to the true value

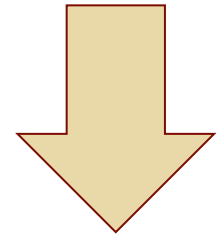
Black points represents the true data

Red points represent the experimental data

3.2.5 Visualization of Errors (statistical concepts)

– Before going depth, **let's do a summary about statistical concepts** that we will need:

1. **Descriptive statistics**
2. **Distributions**
3. **Variability**



Visualization of random errors

3.2.5 Visualization of Errors (statistical concepts)

– Before going depth, **let's do a summary about statistical concepts** that we will need**:**

1. **Descriptive statistics** (it simply provides a description of what the data sample we have looks like)

- **Mean**: the **central value**, commonly called the average
- **Median**: the **middle value** if we ordered the data from low to high and divide exactly in half
- **Mode**: the value which occurs more often

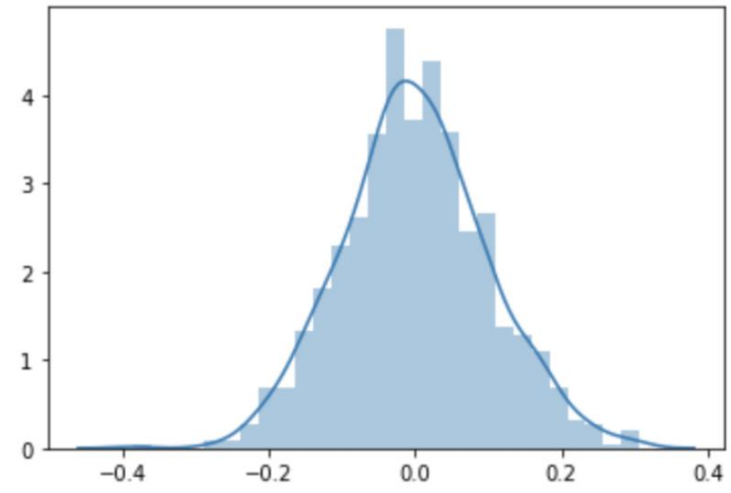
Descriptive statistics are useful, but they **can often hide important information about the data set** (*you will see it with the Anscombe dataset on Friday*)

3.2.5 Visualization of Errors (statistical concepts)

2. Distributions

- **A distribution** is a chart, for example a histogram, that displays **the frequency with which each value appears in a data set.** This type of chart **gives us information about the spread and skewness of the data.**

- One of the most important distributions is **the normal distribution. It is symmetrical in shape with most of the values clustering around the central peak and the further away values distributed equally on each side of the curve.** Many variables in nature will form a normal distribution.



3.2.5 Visualization of Errors (statistical concepts)

3. Variability

– **Variance** measures **how far each value in the data set is from the mean**

– **Standard deviation (SD)** is a common measure of variation for data that has a normal distribution. It gives a value to represent **how widely distributed the values are.**

- low SD: the values tend to lie quite close to the mean.

- high SD: the values are more spread out.

3.2.5 Visualization of Errors (statistical concepts)

3. Variability

- Variance: how far each value in the data set is from the mean
- Standard deviation: how widely distributed the values are.

If the data does not follow a normal distribution, then other measures of variance are used:

– The interquartile range:

- Derived by first ordering the values by rank and then dividing the data points into four equal parts, called quartiles.
- Each quartile describes where 25% of the data points lie according the median.
- The interquartile range is calculated by subtracting the median for the two central quarter (Q1 & Q3).

3.2.5 Visualization of Errors (statistical concepts)

3. Variability

- Variance: how far each value in the data set is from the mean
- Standard deviation: how widely distributed the values are.
- Interquartile (different distributions): Q1, median=Q2, Q3.

!!! Standard error versus standard deviation:

- **The standard deviation is a property of the population**. It tells us **how much spread** there is among individual observations we could make.
- **The standard error tells us how precisely we have determined a parameter estimate.**

*The standard error is approximately **given by the sample standard deviation divided by the square root of the sample size**, and confidence intervals are calculated by multiplying the standard error with small, constant values.*

3.2.5 Visualization of Random Errors

- **The form of the distribution of the random errors must be known**.
- Although the form of the probability distribution must be known, the parameters of **the distribution can be estimated from the data**.

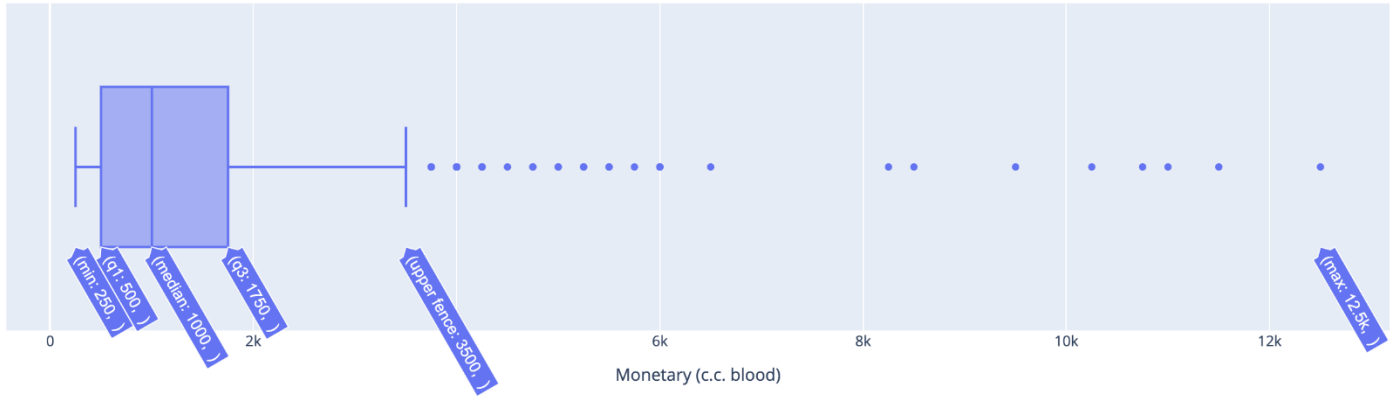
  The random errors from different types of processes could be **described by any one of a wide range of different probability distributions in general**, including: the uniform, triangular, double exponential, binomial and Poisson distributions.

- **The normal distribution often describes the actual distribution of the random errors in real-world processes reasonably well** (This is related to the CLT, that we will see later today). With most process modelling methods - inferences are based on the idea that the random errors are drawn from a normal distribution.

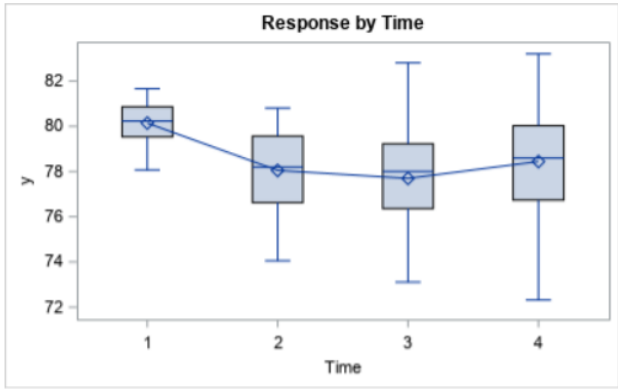
  The normal distribution is also used because **the mathematical theory behind it is well-developed.**

3.2.6 Visualization of Random Errors

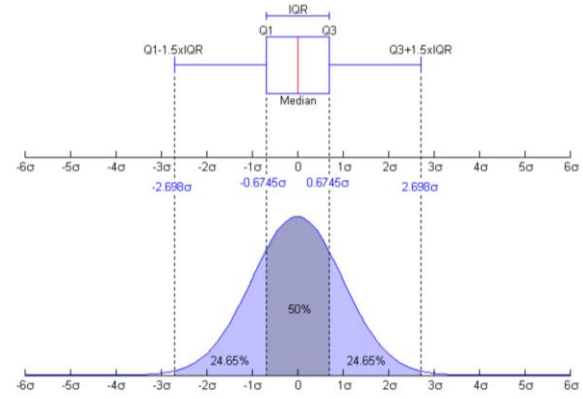
Statistical methods may be used to analyze the data & random errors



Example by Rebecca Vickery 2021: A boxplot provides a useful visualization of the interquartile range (IQR) .

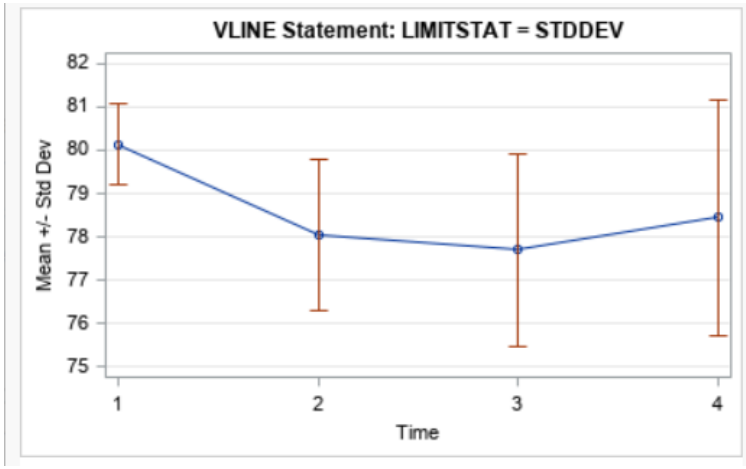


Example by Wicklin 2019: The boxplot shows the schematic distribution of the data at each point. The boxes use the **interquartile range and whiskers to indicate the spread of the data**. A line connects the means/medians of the response at each time point.

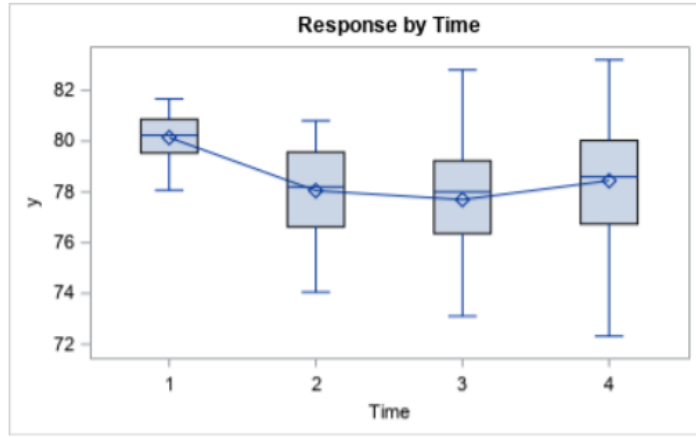


Example by Jhguch 2011: IQR of a normal distribution

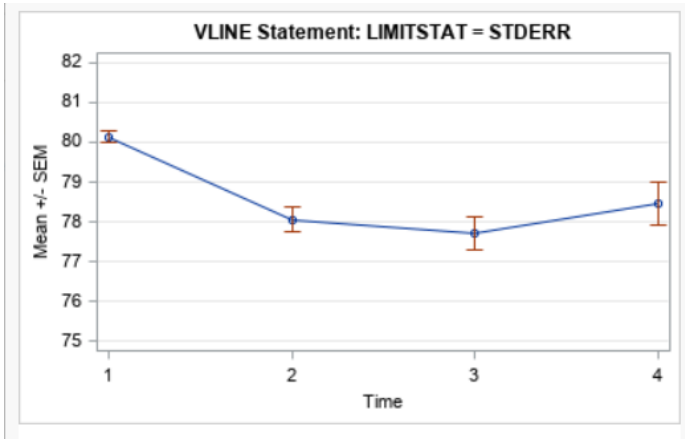
3.2.5 Visualization of Random Errors



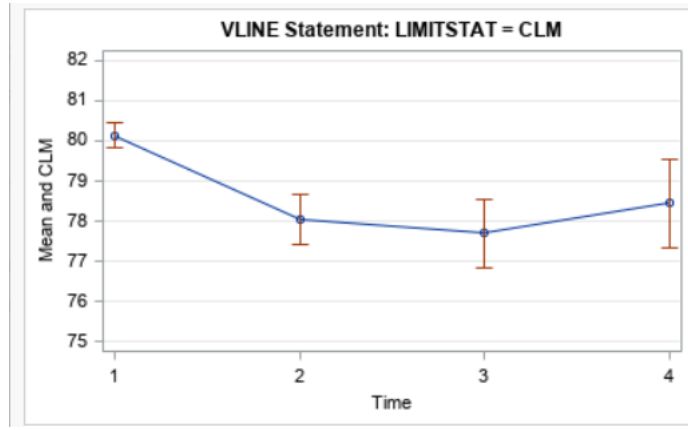
"**standard deviation**" is a term that is familiar to a lay audience



The boxes use the **interquartile range and whiskers to indicate the spread of the data**. A line connects the means/medians of the response at each time point.



The exact meaning of the "**standard error of the mean**" might be difficult to explain to a lay audience, but the qualitative explanation is often sufficient



The "**confidence interval of the mean (CLM)**" is hard to explain to a lay audience

3. Visualizing errors & uncertainty. Contents:

1. Introduction of Visualizing errors & uncertainty

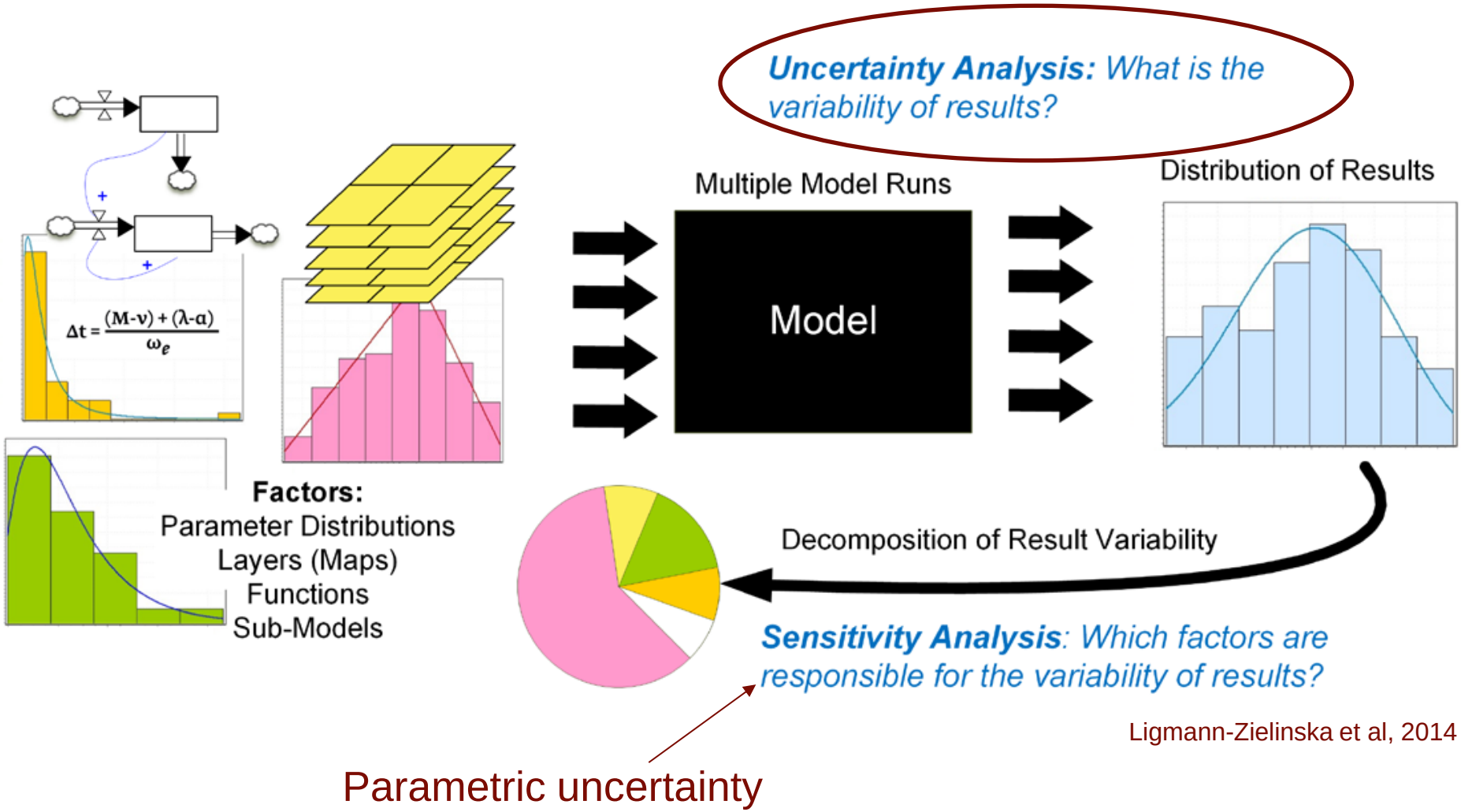
2. Error

1. Introduction
2. Residual error (absolute error, square error, percentage error)
3. Error (graded) bars, confidence strips, eyes, quantile dots and confidence bands
4. Systematic errors & random errors
5. Statistical concepts for visualization errors (descriptive analysis/distributions)
6. Visualizing random errors

3. Uncertainty

1. Introduction
2. Uncertainty visualization: Error bars. Confidence bands. Frequency framing. Standard Error.
3. Dynamic uncertainty visualization: Curve fits and Hypothetical outcome plots
4. Bayesian tools to determine distributions (Monte Carlo simulation). And to normalize them (Central Limit Theorem)

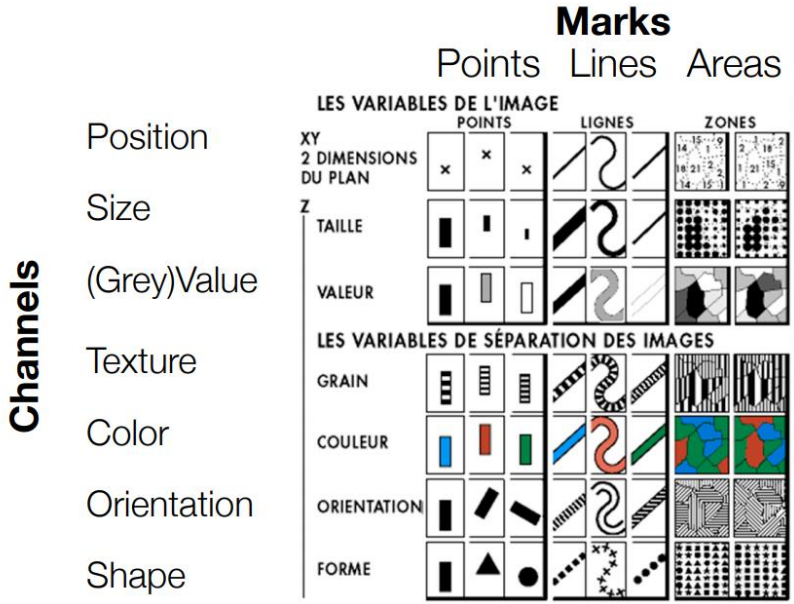
3.3.1 Introduction: **Uncertainty** vs sensitivity analysis



3.3.1 Introduction: **Uncertainty**

Limitations visualizing uncertainty?

- We are **limited by the number of visualization channels**.
- When **moving from quantified uncertainty to visualized uncertainty**, we often **simplify the uncertainty** to make it fit into the available visual representations.



Semiology Graphics (J.Bertin,67)

- We need a **balance between the degree of complexity and the audience that we want to reach.**

+Other channels such as opacity, 3D positions, motion…
! Still limited
! Channels overwhelmed when increasing the amount and dimensionality of the data

3.3.1 Introduction: **Uncertainty**

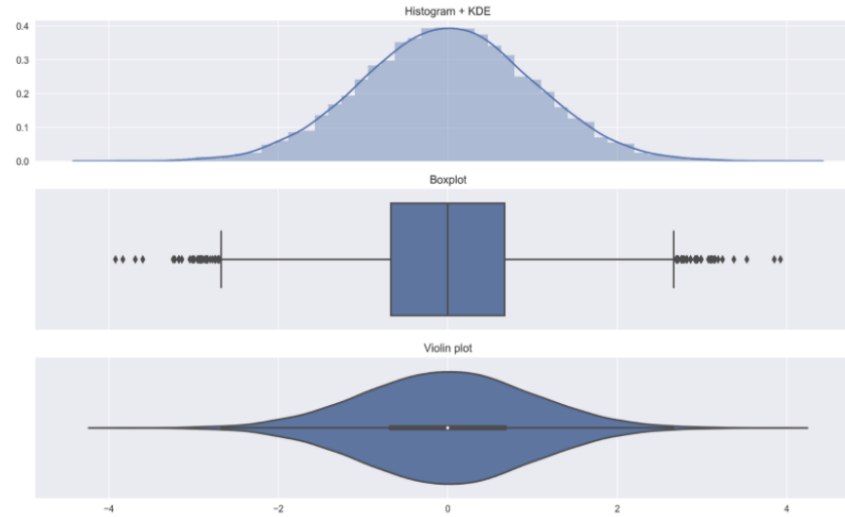
- **Epistemic uncertainties:** due to *lack of knowledge and limited data in practice* (deficient measurements, poor models, missing data)
- **Aleatoric uncertainty:** inherent random uncertainty (from running an experiment and getting slightly different results each time) *-> often represented as a probability density function (PDF)*

The most straightforward understanding of uncertainty is often the easiest to expose visually -> often thought that is entirely statistically defined

3.3.1 Introduction: **Uncertainty**

Non-spatial data uncertainty displays

- **Error bars**
- **Boxplots** to express variability by showing the quartiles.
- **Violin plots**, *additionally displaying the probability density* (kernel density estimation) of the data at each value.



3.3.1 Introduction: **Uncertainty**

Non-spatial data uncertainty displays

- **Error bars**
- **Boxplots** to express variability by showing the quartiles.
- **Violin plots**, additionally displaying the probability density (kernel density estimation) of the data at each value.

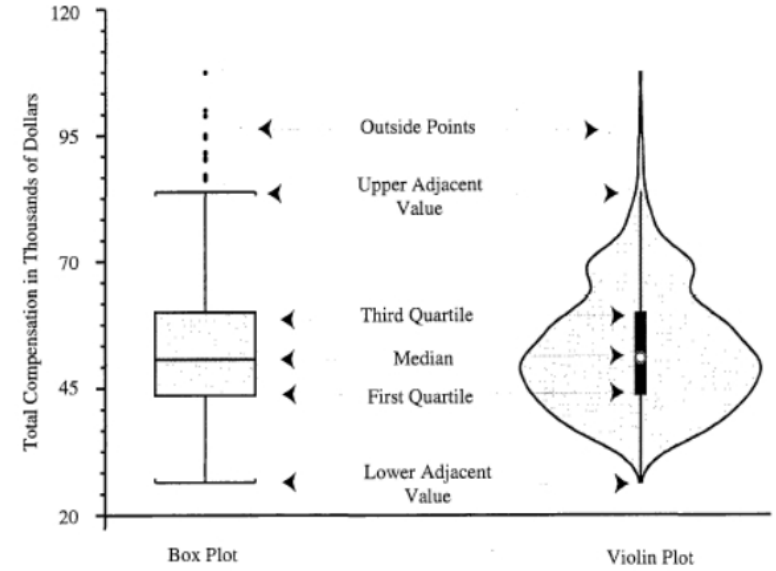


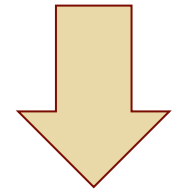
Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.

Hintze and Nelson 1998

3.3.1 Introduction: **Uncertainty**

Spatial data uncertainty metrics (like for errors)

- **A measure of central tendency**, such as the **mean.**
- **An indicator of dispersion**, for example the **variance or standard deviation.**
- **Extrema** (minimum and maximum, or extreme percentiles).



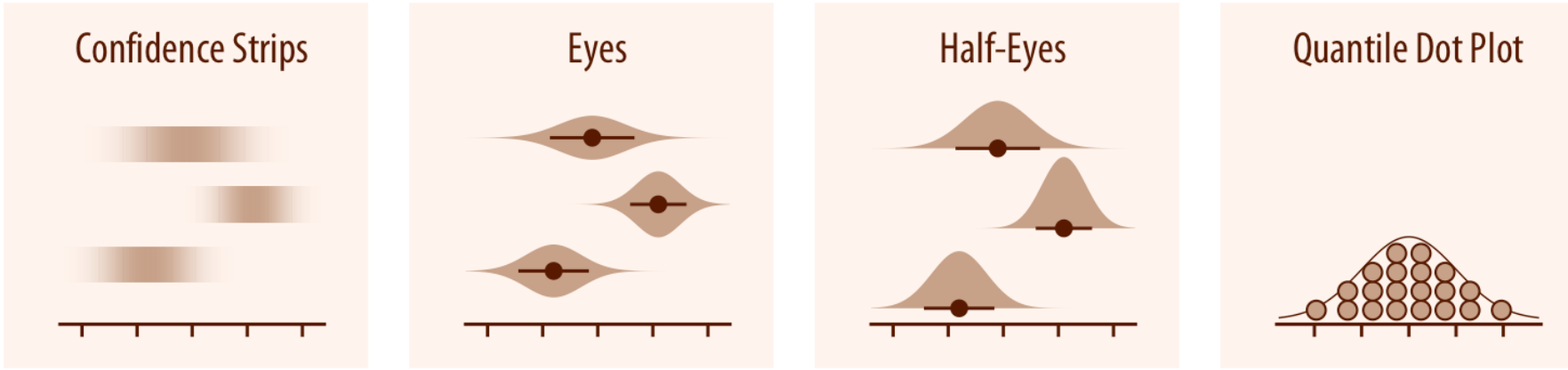
Two **commonly used approaches** to indicate uncertainty **are error bars and confidence bands.**

3.3.2 Visualizing uncertainty

Confidence strips: graduated.

Eyes and half eyes: combine error bars with methodologies to combine distribution (violins and ridgelines).

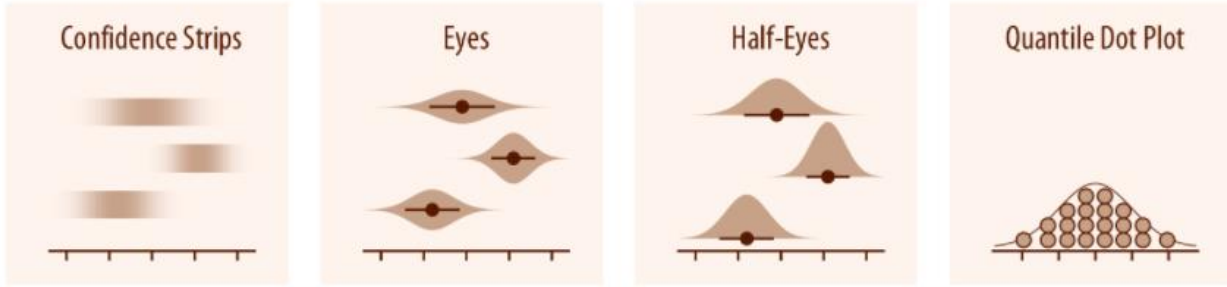
Quantile dots: the distribution in discrete units.



Claus O.Wilke

3.3.2 Visualizing uncertainty

To achieve a more detailed visualization than is possible with error bars or graded error bars, **we can visualize the actual confidence or posterior distributions**

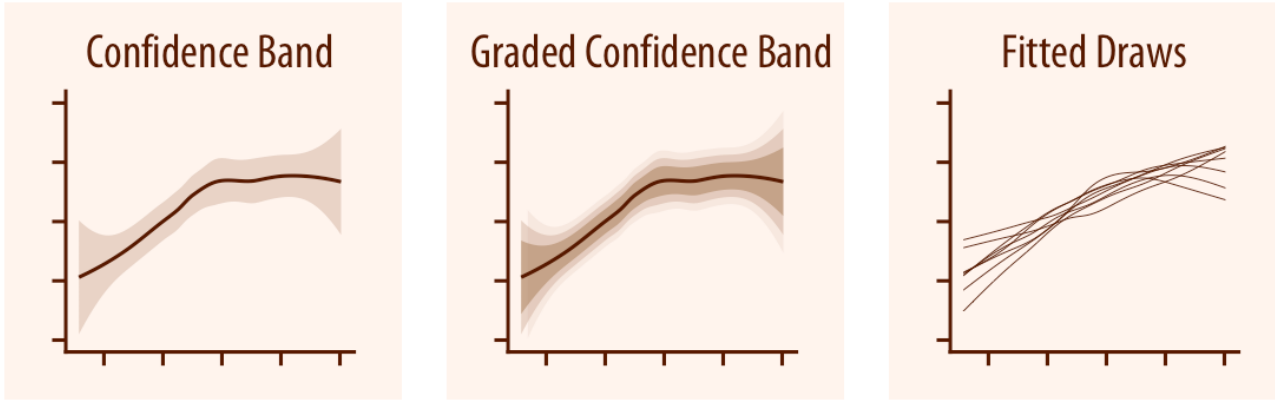


Claus Wilke

- **Confidence strips** provide a clear visual sense of uncertainty but are difficult to read accurately.
- **Eyes and half-eyes** combine error bars with approaches to visualize distributions (violins and ridgelines, respectively) -> show both precise ranges for some confidence levels and the overall uncertainty distribution.
- **A quantile dot plot** can serve as an alternative visualization of an uncertainty distribution (later)

3.3.2 Visualizing uncertainty

For smooth line graphs, the equivalent of an error bar is a confidence band (as we saw in the errors' subsection)



Claus Wilke

- It shows a range of values the line might pass through at a given confidence level.
- As in the case of error bars, we can draw graded confidence bands that show multiple confidence levels at once.
- We can also show individual fitted draws instead of / or in addition to the confidence bands.

3.3.2 Uncertainty visualization: **for a lay audience**

Then, two **commonly used approaches** to indicate uncertainty **are error bars and confidence bands.**

For a lay audience, however, **visualization strategies that create a strong intuitive impression of the uncertainty will be preferable**, even if they come at the cost of either reduced visualization accuracy or less data-dense displays.

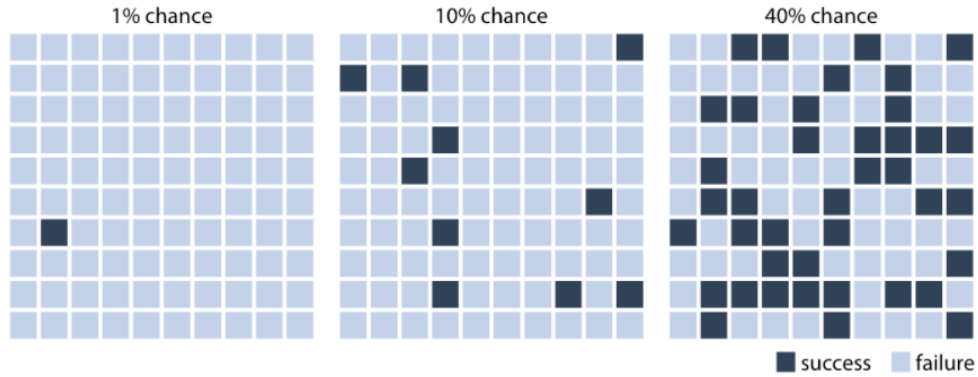
Options here include:

- **frequency framing**, where we explicitly draw different possible scenarios in approximate proportions
- **animations** that cycle through different possible scenarios.

3.3.2 Uncertainty visualization: **Frequency framing**

- Mathematically, we deal with uncertainty by employing the concept of probability. BUT **visualising a single probability is difficult**.
- For many problems of practical relevance, it is sufficient **to think about relative frequencies.**

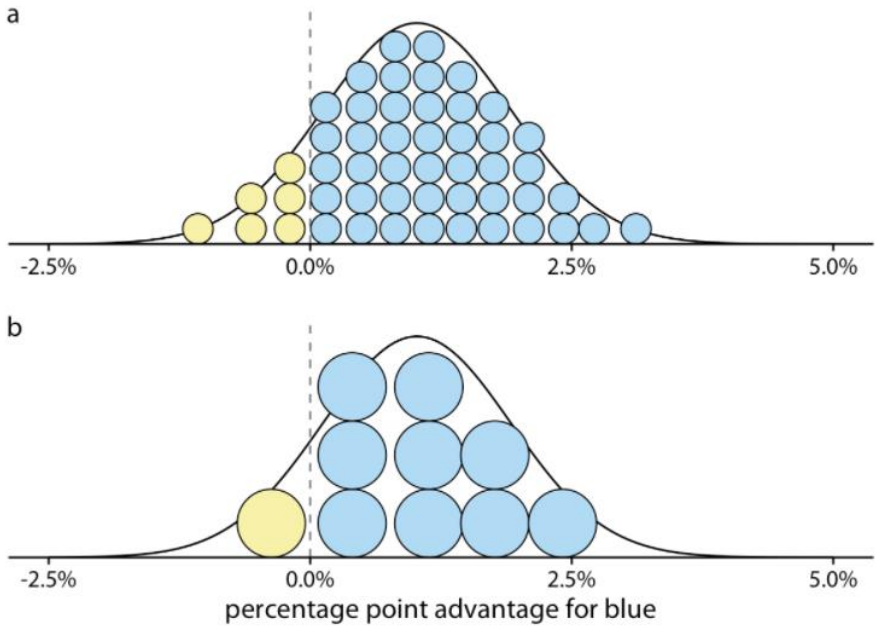
We can make the concept of probability tangible by creating a graph that emphasizes both the frequency aspect and the unpredictability of a random trial, for example by drawing squares of different colours in a random arrangement.



Claus Wilke

3.3.2 Uncertainty visualization: **Frequency framing**

- **What happens if we have more than two discrete outcomes (success or failure)?**



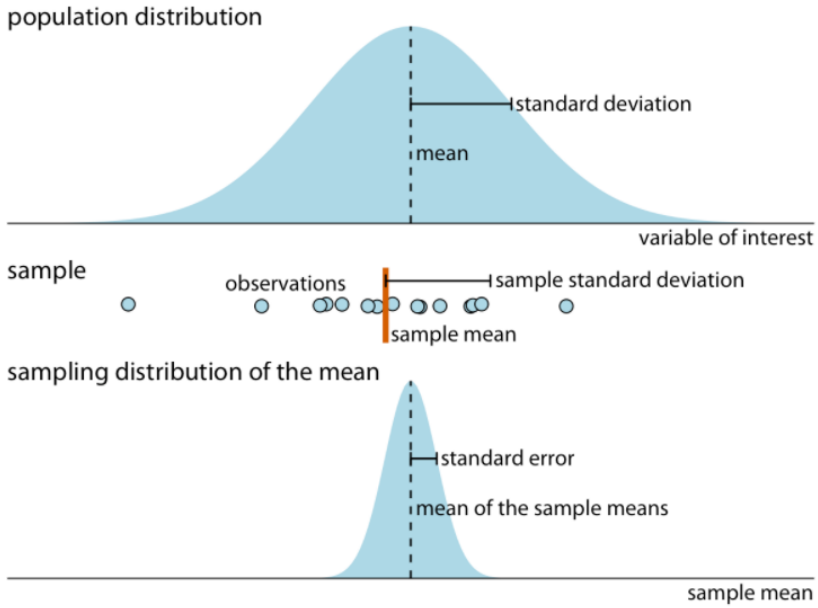
Quantile dotplot representations of an election outcome distribution.
The percentage chance in (b) is not accurate.
They serve to trade mathematical precision for more accurate human perception of the resulting visualization - communicating to a lay audience

Claus Wilke

As a general principle, **quantile dotplots** should use a small to moderate number of dots. **If there are too many dots, then we tend to perceive them as a continuum rather than as individual, discrete units.**

3.3.2 Uncertainty visualization: **Standard Error**

The standard error provides a measure of the uncertainty associated with our parameter estimate.



The variable of interest that we are studying has some true distribution in the population, with a true **population mean and standard deviation.**

Any finite **sample** of that variable will have a sample **mean and standard deviation that differ from the population parameters.**

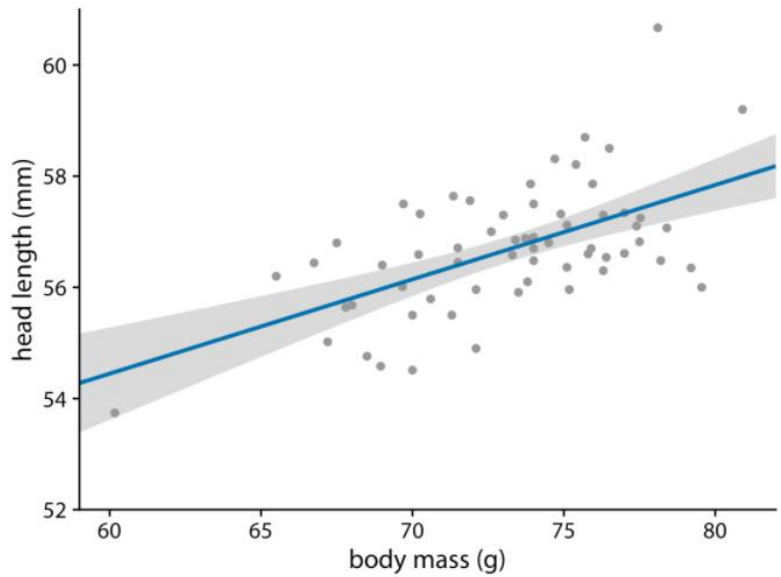
The standard error informs us about how precisely we are estimating the population mean (our parameter estimate)

Claus Wilke

The larger the sample size -> the smaller the standard error and thus the less uncertain the estimate

3.3.3 **Dynamic** uncertainty visualization: **Curve fits**

- **We can show a trend in a dataset by fitting a straight line or curve to the data**
- These trend estimates also have uncertainty, and it is customary **to show the uncertainty in a trend line with a confidence band**

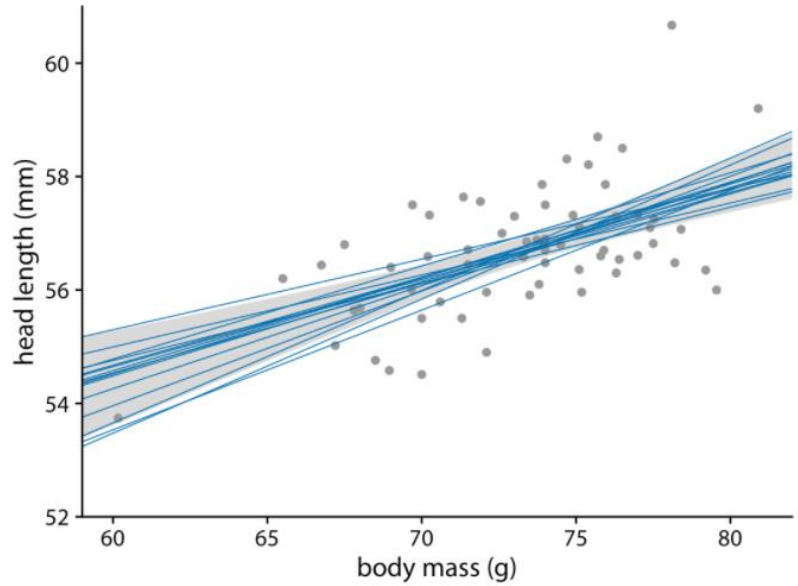


- The straight blue line represents **the best linear fit** to the data.
- The gray band around the line shows **the uncertainty in the linear fit**. The gray band represents a 95% confidence level.

Keith Tarvin, Oberlin College

3.3.3 **Dynamic** uncertainty visualization: **Curve fits**

- We can show a trend in a dataset by fitting a straight line or curve to the data
- These trend estimates also have uncertainty, and it is customary **to show the uncertainty in a trend line with a confidence band**

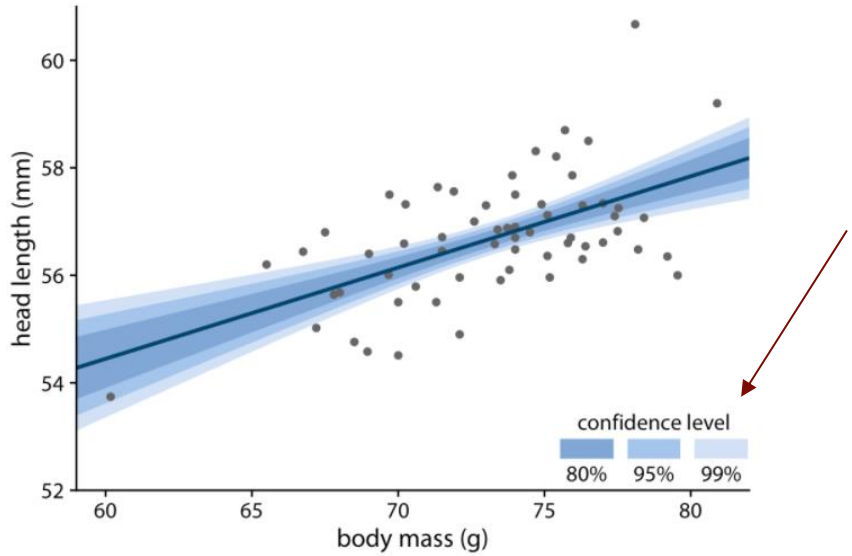


The straight blue lines now represent equally likely alternative fits randomly drawn from the posterior distribution.

Keith Tarvin, Oberlin College

3.3.3 **Dynamic** uncertainty visualization: **Curve fits**

- To draw a confidence band, we need to specify a confidence level, and **it can be useful to highlight different levels of confidence**.
- **A graded confidence band** enhances the sense of uncertainty in the reader, and it **forces the reader to confront the possibility that the data might support different alternative trend lines.**



We can draw **graded confidence bands to highlight the uncertainty in the estimate**.

Keith Tarvin, Oberlin College

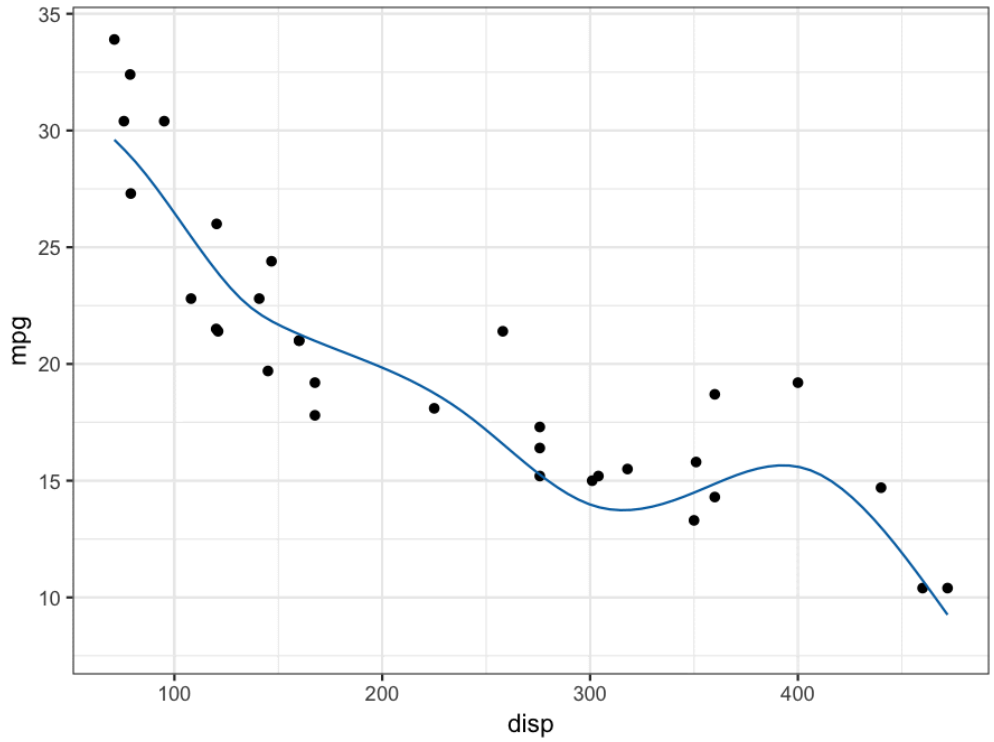
3.3.3 **Dynamic** uncertainty visualization: **HOPs**

Hypothetical outcomes plots (HOPs)

- HOPs visualize a set of draws from a distribution -> **each draw is shown as a new plot** in either a small multiples or animated form.
- Better than showing a continuous probability distribution.
- **HOPs require relatively little background knowledge to interpret**.

! Limitation: dynamically presenting draws introduces sampling error

3.3.3 **Dynamic** uncertainty visualization: **HOPs**



Hullman J. et al, 2015

Visualizing uncertainty with hypothetical outcomes plots (Claus Wilke) : https://www.youtube.com/watch?v=SjYwhku2si0

3.3.4 Bayesian - Monte Carlo Simulation

When to use it?

- When it is impossible (or impractical) to determine a distribution theoretically (o deterministically)

- In many areas: engineering, finance, management of risks of a project

What involves?

- One or more input variables X (some of which usually follow a probability distribution)
- One or more output variables Y (whose distribution is desired)
- A mathematical model coupling the X's and the Y's

3.3.4 Central limit theorem (CLT)

Idea : Given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.

Why is it important?

It implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions -> the Gaussian distribution is used for hypothesis testing using confidence intervals, or to look at the statistical significance of experiment results.

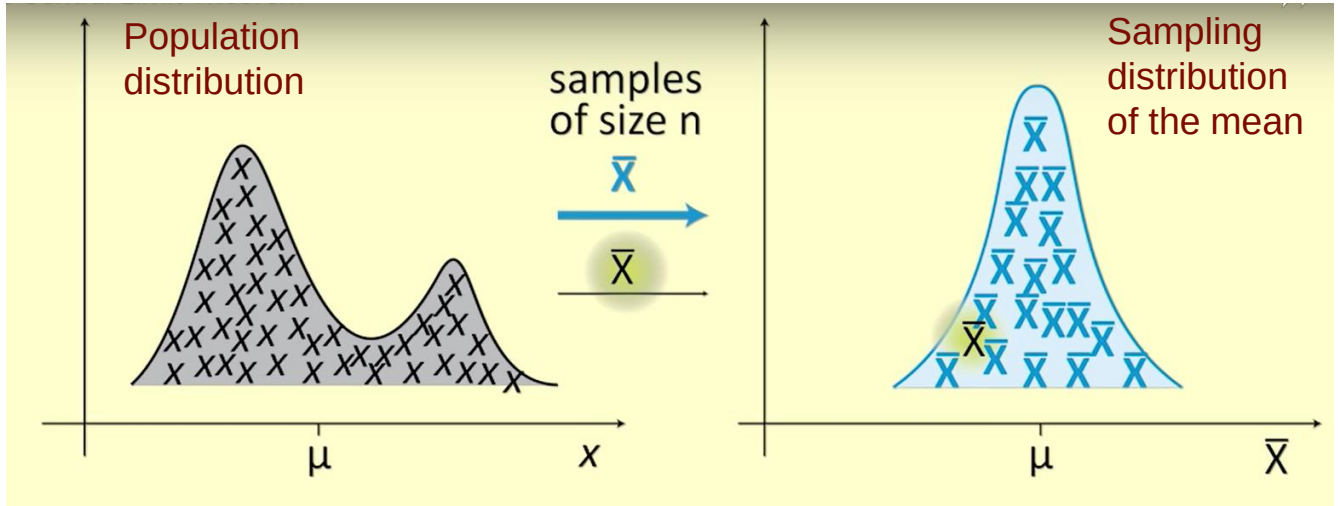


Image from Creative maths

3.3.4 Assumptions behind the CLT

- **Randomization Condition:**

  The data must be sampled randomly (or errors)

- **Independence Assumption:**

  The sample values must be independent of each other

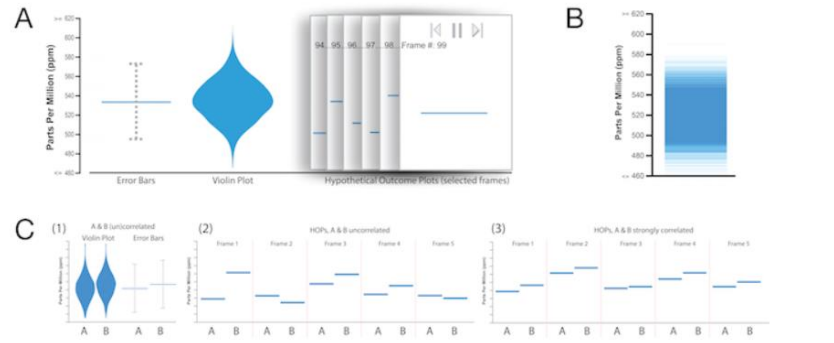
- **Sample Size:**

  When the sample is drawn without replacement (usually the case), the sample size, *n*, should be no more than 10% of the population.

  When the population is skewed or asymmetric, the sample size should be large. If the population is symmetric, then we can draw small samples as well.

3.3 More examples visualizing uncertainty

- Visualizing uncertainty by Robert Falkowitz
- Uncertainty Quantification and Visualization: Geo-Spatially Registered Terrains and Mobile Targets Suresh Lodha Computer Science, University of California, - ppt download (slideplayer.com)
- Visualizing Uncertainty About the Future: spiegelhalter_visualizing.pdf (berkeley.edu)
- Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering (manuscript)



(A) Representations of uncertainty compared in our study, (B) HOPs limiting case, (C) HOPs can express properties of a joint distribution.

3.4 Transformation & Data massage. Contents:

1. Introduction of Visualizing errors & uncertainty
2. Error
3. Uncertainty
4. **Transformation and data massage**
    1. Introduction
    2. Best practices
    3. Potential activity transforming your data
    4. Tidy and transform data in R (basics)

3.4.1 Introduction: Transformation & **Data massage.**

What if the database is not formatted in the way you expect? Or the data is completely unstructured?

It is rare that you get the data in exactly the right form you need

- Before data is loaded to visualize it, **it must be transformed** to meet any format and structural requirements
- ***Data massaging*, also known as data *cleansing* or *scrubbing*, is a process that eliminates unnecessary information from data or cleans a dataset to make it useable.**

3.4.1 Introduction: **Transformation** & Data massage**.**

What does it mean for your data to be “tidy”?

The word **“tidy”** in data science **using R means that your data follows a standardized format:**

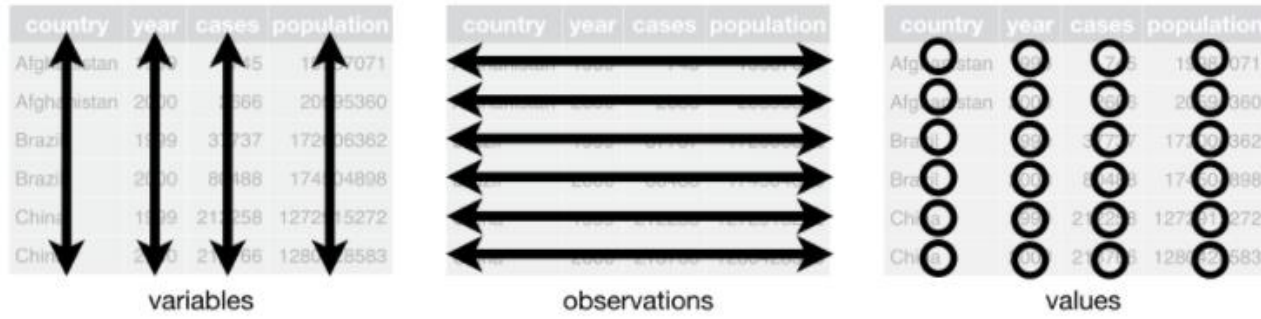
- A *dataset* is a collection of values, usually either numbers (if quantitative) or strings AKA text data (if qualitative/categorical). **Values are organised in two ways. Every value belongs to a variable and an observation**.
- **A variable contains all values that measure the same underlying attribute across units** (examples: weight, temperature, duration).
- **An observation contains all values measured on the same unit across attributes** (examples: person, day, village).
- **“Tidy” data is a standard way of mapping the meaning of a dataset to its structure.** A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

3.4.1 Introduction: **Transformation & Data massage.**

What does it mean for your data to be “tidy”?

In “tidy data”:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table



Tidy data graphic from R for Data Science

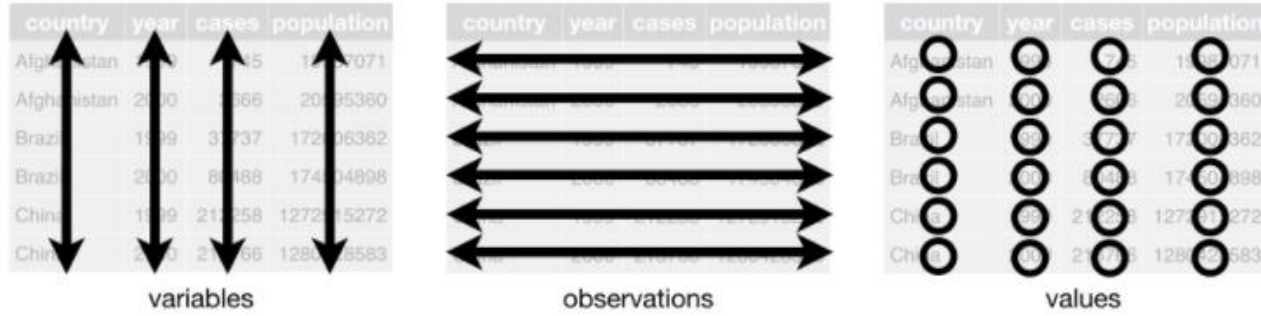
3.4.1 Introduction: **Transformation & Data massage.**

What does it mean for your data to be “tidy”?

Fecha	Nombre	Mate	Ingles
1-11-2015	Hernandez, Rodrigo	90	60

mes	año	primer	apellido	materia	puntos
11	2015	Rodrigo	Hernandez	mate	90
11	2015	Rodrigo	Hernandez	ingles	60

Edgar Ruiz
2018



Tidy data graphic from R for Data Science

3.4.1 Introduction: **Transformation & Data massage.**

- **Databases come in different shapes and sizes and each must be treated as unique.**

- A few data massaging techniques are required **to adapt the data** to the algorithms we are working with.

- Common tasks include stripping unwanted characters and whitespace, converting number and date values into desired formats, and organising data into a meaningful structure.

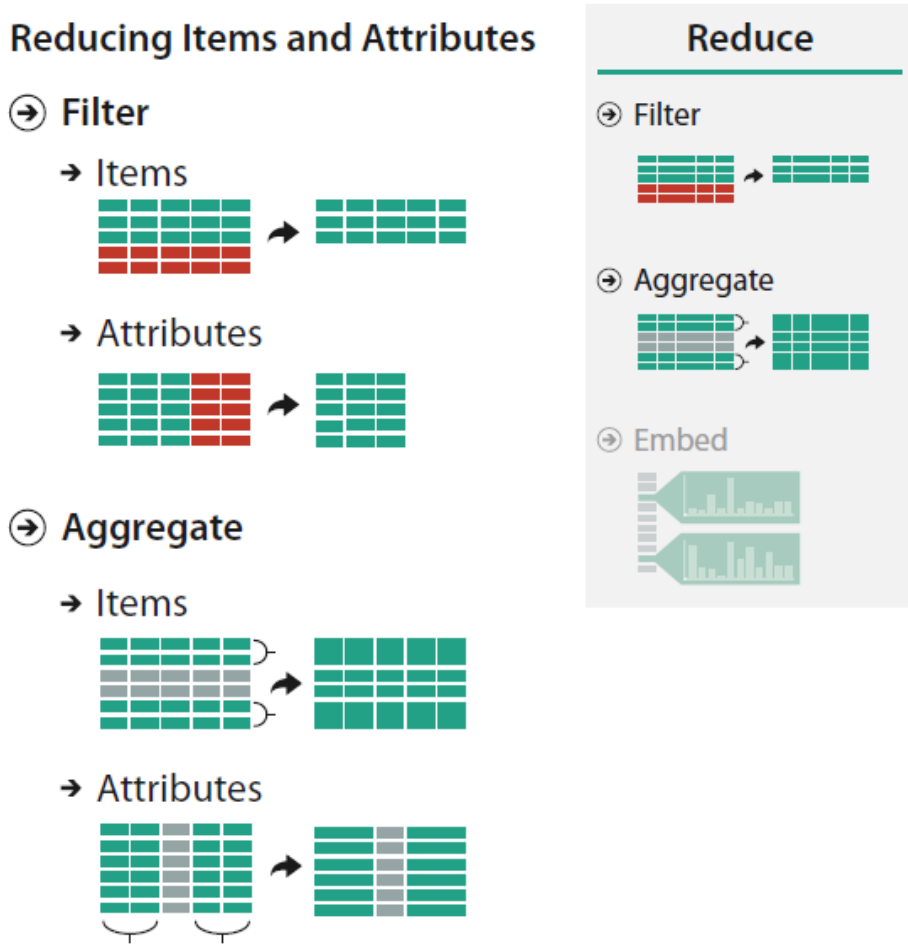
- ***Massaging the data is usually the "transform" step.*** In most cases**, one or more transformations are required.**

3.4.1 Introduction: **Transformation & Data massage.**

Things we do to massage the data include:

- **Change formats** from the standard source system emissions to the target system requirements, e.g. change date format from m/d/y to d/m/y, or sort the data.
- **Replace missing values** with defaults, e.g. "0" when a quantity is not given.
- **Filter out data** that is not desired in the destination system. Sub setting or removing observations based on some condition.
- **Check validity of data and fixing records:** ignore or report on rows that would cause an error, remove unwanted characters and duplicates.
- **Splitting and resampling**
- **Normalise/standardizing data** to remove variations that should be the same, e.g. replace upper case with lower case, replace "01" with "1".

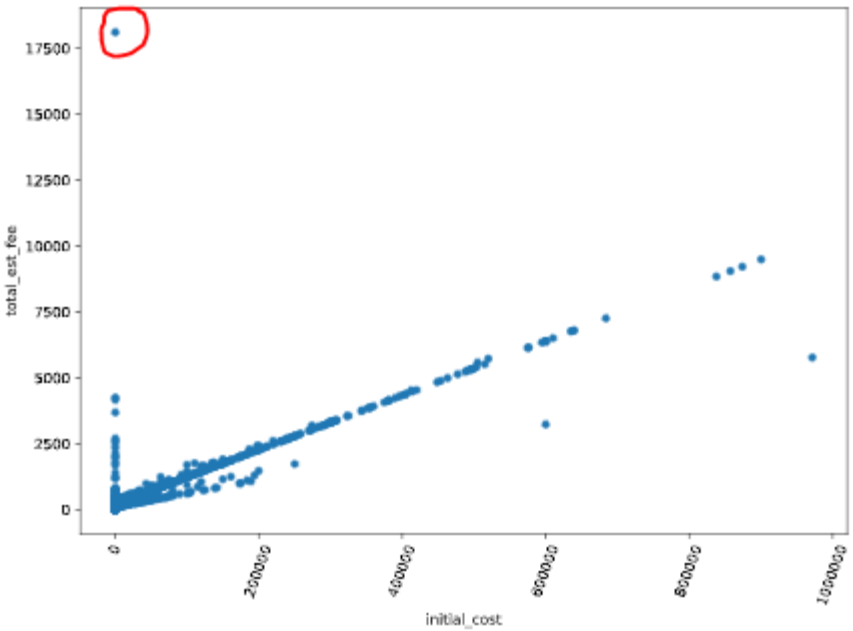
3.4.1 Introduction: **Transformation & Data massage.**



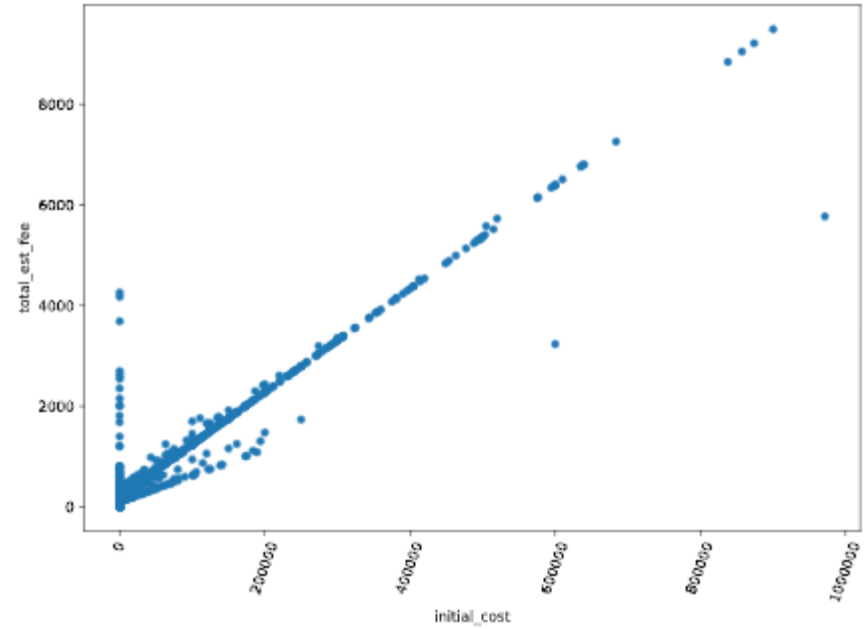
Design choices for reducing (or increasing) the amount of data items and attributes to show.

Tamara Munzner

3.4.1 Introduction: **Transformation & Data massage.**



Removing the outlier



	Name	Height	Roll
0	A	5.2	55
1	A	5.2	55
2	C	5.6	15
3	D	5.5	80
4	E	5.3	12
5	E	5.3	12
6	G	5.6	47
7	H	5.5	104

Removing duplicate rows

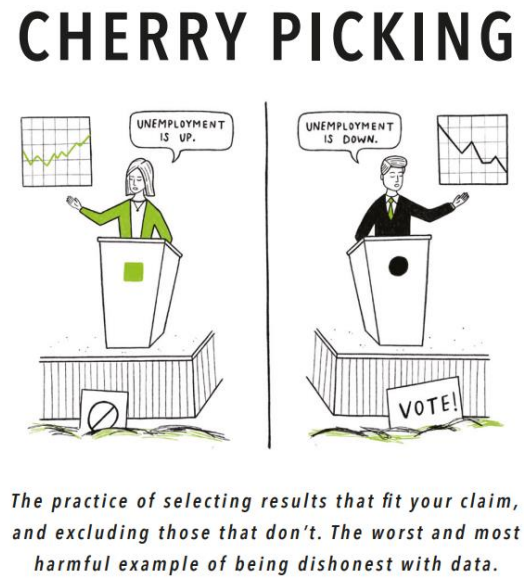
	Name	Height	Roll
0	A	5.2	55
2	C	5.6	15
3	D	5.5	80
4	E	5.3	12
6	G	5.6	47
7	H	5.5	104

3.4.2 Best practices

Unfortunately, there is such a thing as a bad massage. **The term "data massaging" is also associated with the practice of "cherry-picking",** selectively excluding or altering data based on what people want (or don't want) it to reflect.

Cherry-picking changes the message that the final visualisation communicates to the audience.

!!! Be careful



3.4.2 Best practices

The questions you need to ask of your data are:

- Does it represent genuine observations about a given phenomenon or is it influenced by the limitations of a collection method?
- Does your data reflect the entirety of a particular phenomenon, a recognised sample, or maybe even an obstructed view caused by hidden limitations in the availability of data about that phenomenon?

Once you complete your examination of your data you will have a good idea about what actions may be needed to transform your data.

In accordance with the desire for trustworthy design, any modifications or enhancements you apply to your data need to be noted and potentially explained to your audience.

3.4.3 Potential activity transforming your data

'Before you can plot or graph anything, you have to find the data, understand it, evaluate it, clean it, and perhaps restructure it.' (Marcia Gray, graphic designer)

Three different types of potential activity involved in transforming your data:

- **Cleaning:** resolve any data condition issues
- **Creating:** consider developing new calculations and value conversions
- **Consolidating:** think about introducing further data to expand or append to what you already have

3.4.3 Potential activity transforming your data

- **Cleaning:** There is no single approach for how best to conduct data cleaning- Issues may be resolved through manual intervention, sorting, filtering, isolating, modifying any problem values/characters.
- **Creating:** Expand your data to form new calculations and derive new groupings or any other mathematical treatments. This may include:

  - Creating percentage calculations based on existing quantities.

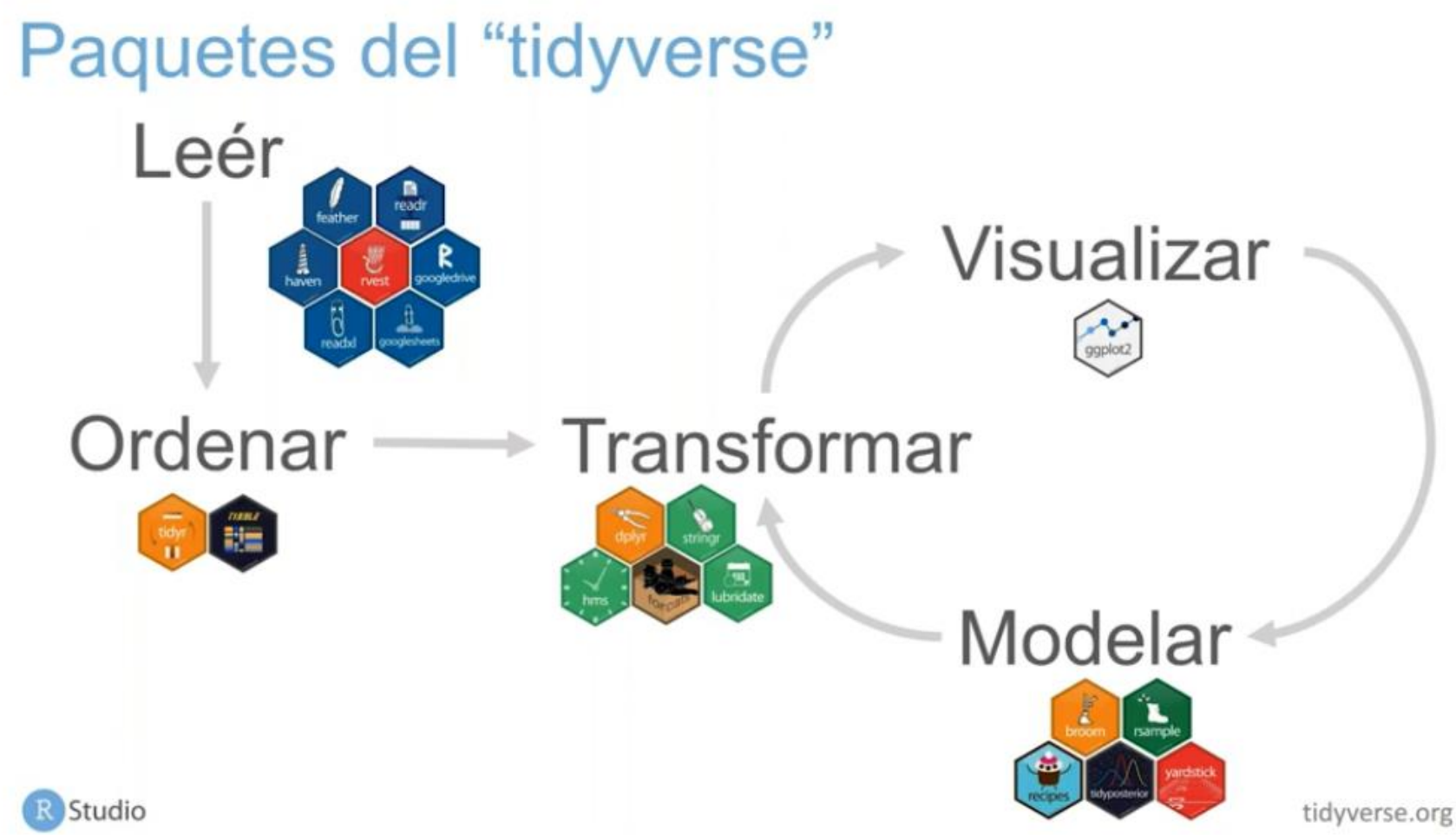
  - Using ‘start date’ and ‘end date’ values to calculate the duration in days.

  - Using logic-based formulae to create new categorical values out of quantities

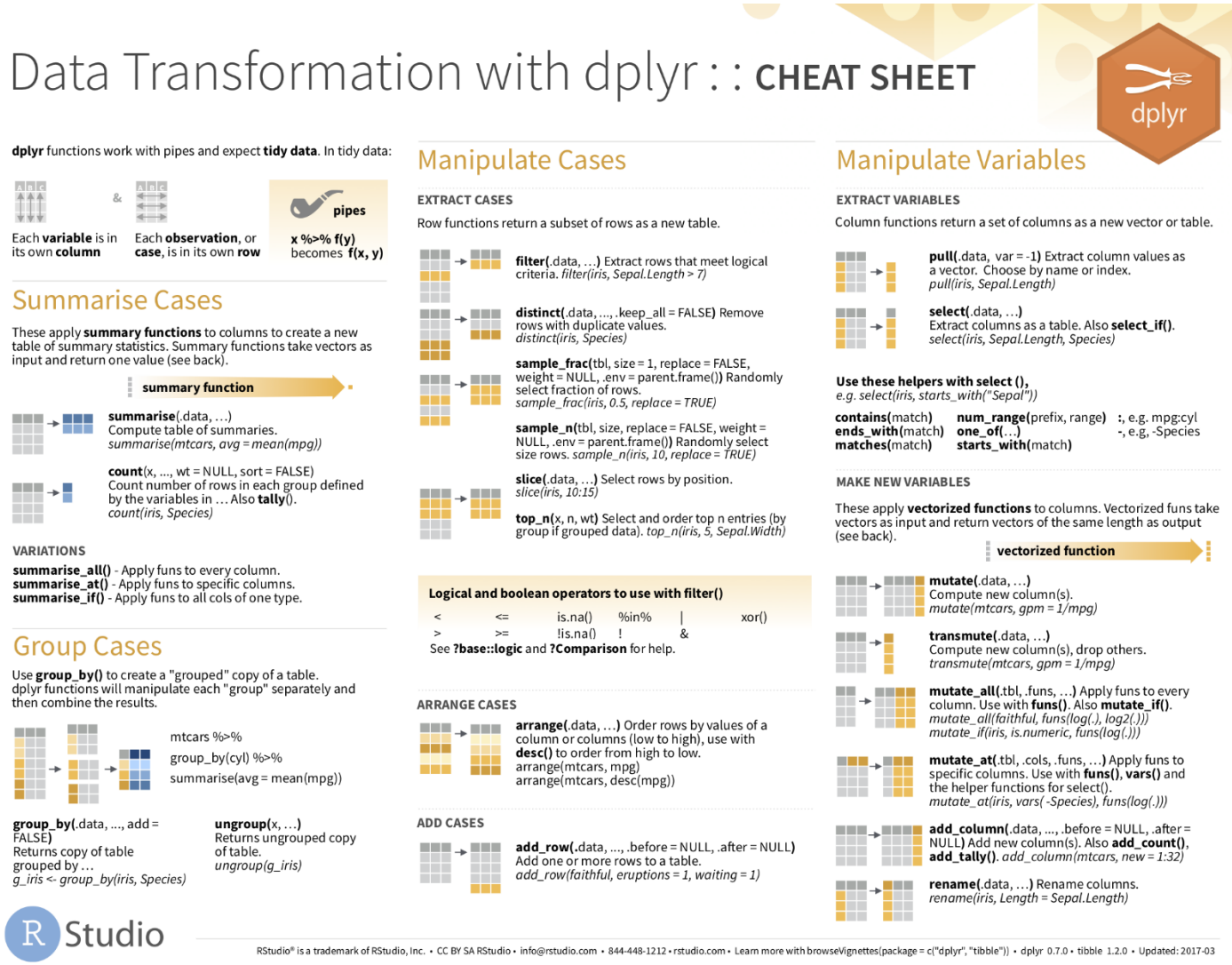
  - To derive reasonable categorical or quantitative values from the original form.

- **Consolidating:** you may seek to source and introduce additional data to **expand** (more variables) or **append** (more items) your data further in order to enhance its analytical potential

3.4.4 Tidy and transform data with R (basics)



3.4.4 Data Transforming with R



Data Transformation with dplyr : : **CHEAT SHEET**

dplyr

dplyr functions work with pipes and expect **tidy data**. In tidy data:

Each **variable** is in its own **column** & Each **observation**, or **case**, is in its own **row**

pipes
x %>% f(y) becomes **f(x, y)**

Summarise Cases

These apply **summary functions** to columns to create a new table of summary statistics. Summary functions take vectors as input and return one value (see back).

summary function

summarise(.data, …)
Compute table of summaries.
summarise(mtcars, avg = mean(mpg))

count(x, ..., wt = NULL, sort = FALSE)
Count number of rows in each group defined by the variables in … Also **tally**().
count(iris, Species)

VARIATIONS

summarise_all() - Apply funs to every column.
summarise_at() - Apply funs to specific columns.
summarise_if() - Apply funs to all cols of one type.

Group Cases

Use **group_by()** to create a "grouped" copy of a table. dplyr functions will manipulate each "group" separately and then combine the results.

mtcars %>%
group_by(cyl) %>%
summarise(avg = mean(mpg))

group_by(.data, ..., add = FALSE)
Returns copy of table grouped by …
g_iris <- group_by(iris, Species)

ungroup(x, …)
Returns ungrouped copy of table.
ungroup(g_iris)

Manipulate Cases

EXTRACT CASES

Row functions return a subset of rows as a new table.

filter(.data, …) Extract rows that meet logical criteria. *filter(iris, Sepal.Length > 7)*

distinct(.data, ..., .keep_all = FALSE) Remove rows with duplicate values.
distinct(iris, Species)

sample_frac(tbl, size = 1, replace = FALSE, weight = NULL, .env = parent.frame()) Randomly select fraction of rows.
sample_frac(iris, 0.5, replace = TRUE)

sample_n(tbl, size, replace = FALSE, weight = NULL, .env = parent.frame()) Randomly select size rows. *sample_n(iris, 10, replace = TRUE)*

slice(.data, …) Select rows by position.
slice(iris, 10:15)

top_n(x, n, wt) Select and order top n entries (by group if grouped data). *top_n(iris, 5, Sepal.Width)*

Logical and boolean operators to use with filter()

<	<=	is.na()	%in%	|	xor()
>	>=	!is.na()	!	&	

See **?base::logic** and **?Comparison** for help.

ARRANGE CASES

arrange(.data, …) Order rows by values of a column or columns (low to high), use with **desc()** to order from high to low.
arrange(mtcars, mpg)
arrange(mtcars, desc(mpg))

ADD CASES

add_row(.data, ..., .before = NULL, .after = NULL) Add one or more rows to a table.
add_row(faithful, eruptions = 1, waiting = 1)

Manipulate Variables

EXTRACT VARIABLES

Column functions return a set of columns as a new vector or table.

pull(.data, var = -1) Extract column values as a vector. Choose by name or index.
pull(iris, Sepal.Length)

select(.data, …)
Extract columns as a table. Also **select_if()**.
select(iris, Sepal.Length, Species)

Use these helpers with select (),
e.g. select(iris, starts_with("Sepal"))

contains(match)
ends_with(match)
matches(match)
num_range(prefix, range)
one_of(…)
starts_with(match)
:, e.g. mpg:cyl
-, e.g, -Species

MAKE NEW VARIABLES

These apply **vectorized functions** to columns. Vectorized funs take vectors as input and return vectors of the same length as output (see back).

vectorized function

mutate(.data, …)
Compute new column(s).
mutate(mtcars, gpm = 1/mpg)

transmute(.data, …)
Compute new column(s), drop others.
transmute(mtcars, gpm = 1/mpg)

mutate_all(.tbl, .funs, …) Apply funs to every column. Use with **funs()**. Also **mutate_if()**.
mutate_all(faithful, funs(log(.), log2(.)))
mutate_if(iris, is.numeric, funs(log(.)))

mutate_at(.tbl, .cols, .funs, …) Apply funs to specific columns. Use with **funs()**, **vars()** and the helper functions for select().
mutate_at(iris, vars(-Species), funs(log(.)))

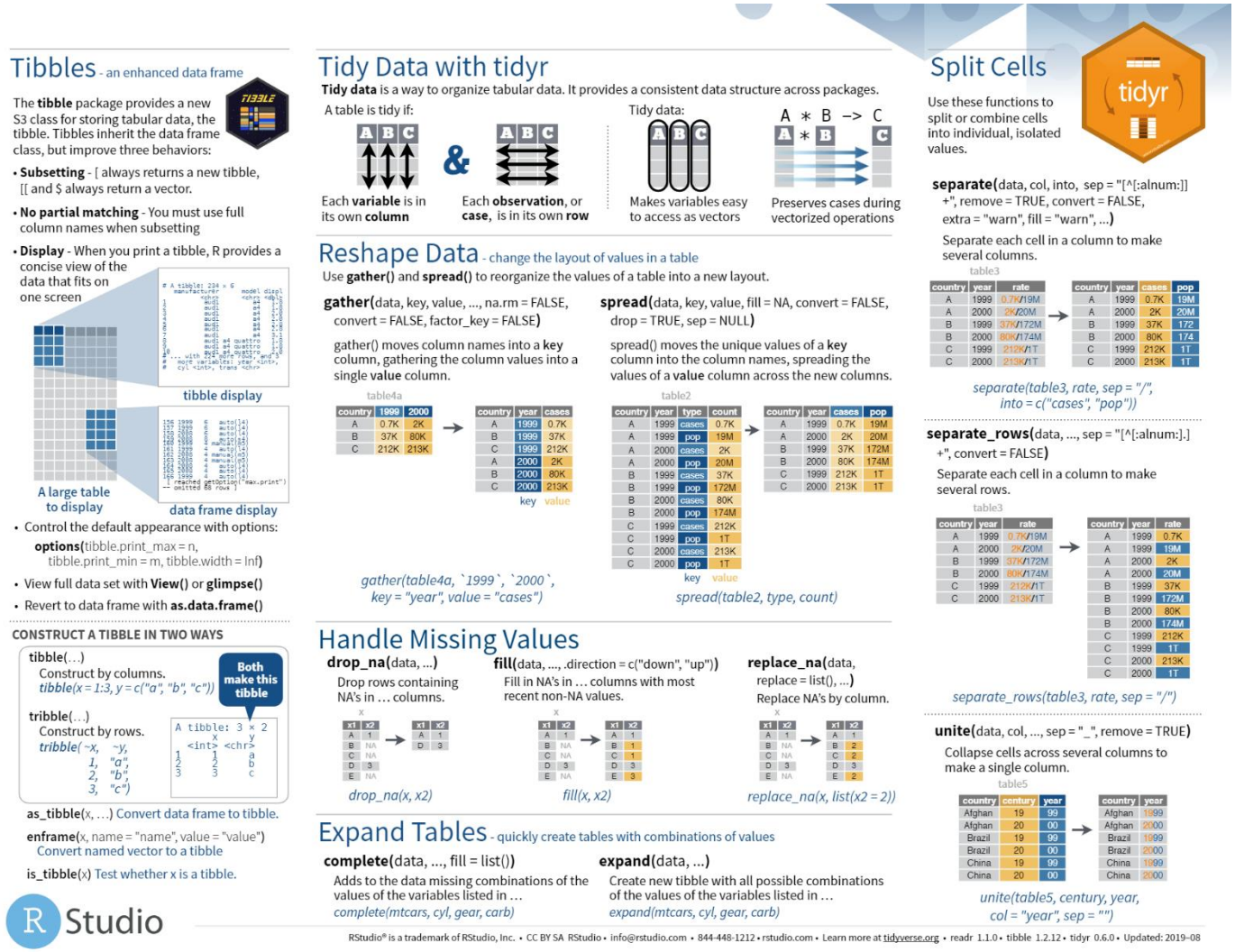
add_column(.data, ..., .before = NULL, .after = NULL) Add new column(s). Also **add_count()**, **add_tally()**. *add_column(mtcars, new = 1:32)*

rename(.data, …) Rename columns.
rename(iris, Length = Sepal.Length)

RStudio

We will see some examples during seminars

3.4.4 Tidy Data with R



We will see some functions during seminars too

Thanks for your attention!

Judit Chamorro Servent

Departament de Matemàtiques

judit.chamorro@uab.cat