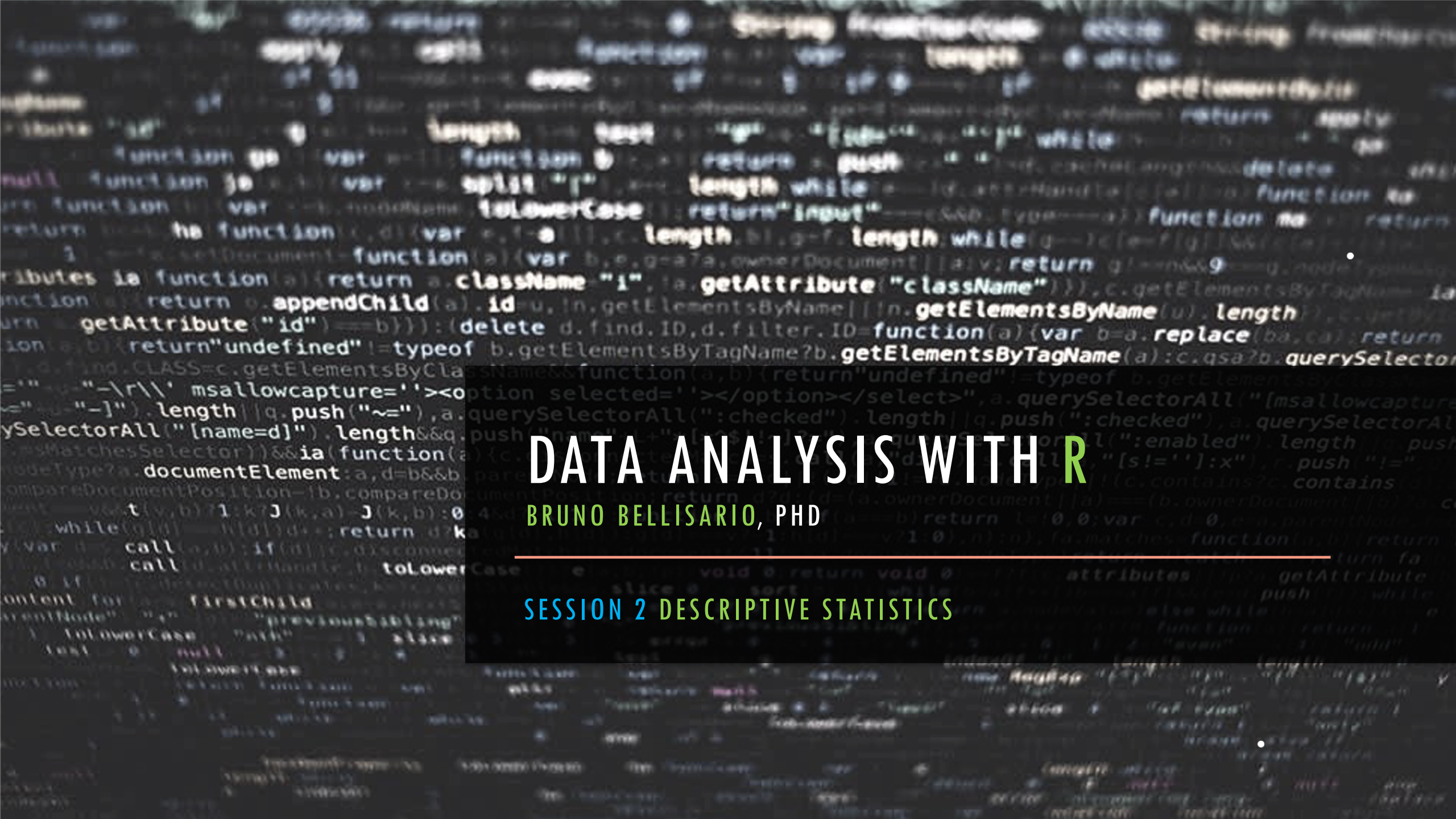


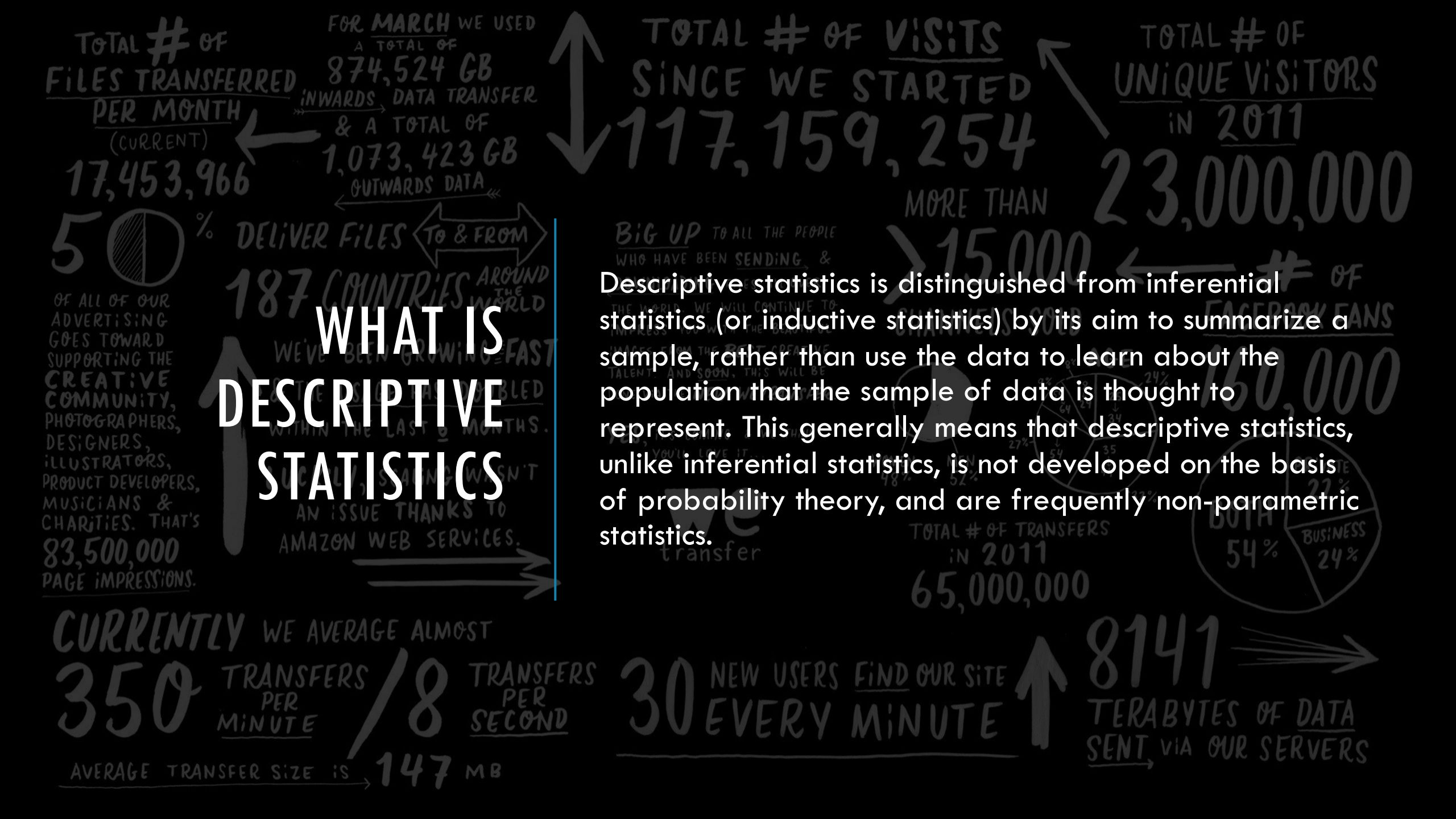
DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 2 DESCRIPTIVE STATISTICS

WHAT IS DESCRIPTIVE STATISTICS

Descriptive statistics is distinguished from inferential statistics (or inductive statistics) by its aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, is not developed on the basis of probability theory, and are frequently non-parametric statistics.



PRE-PROCESSING DATA

Preprocessing the data is one of the crucial steps of data analysis, one of the preliminary steps in that includes feature scaling. Often, programmers new to data science tend to neglect or bypass the step and directly go to analysing the data; this leads to bias and, in turn, influences the prediction accuracy.

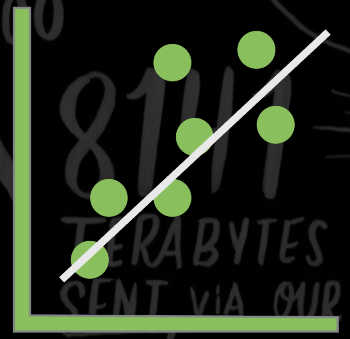
PRE-PROCESSING DATA

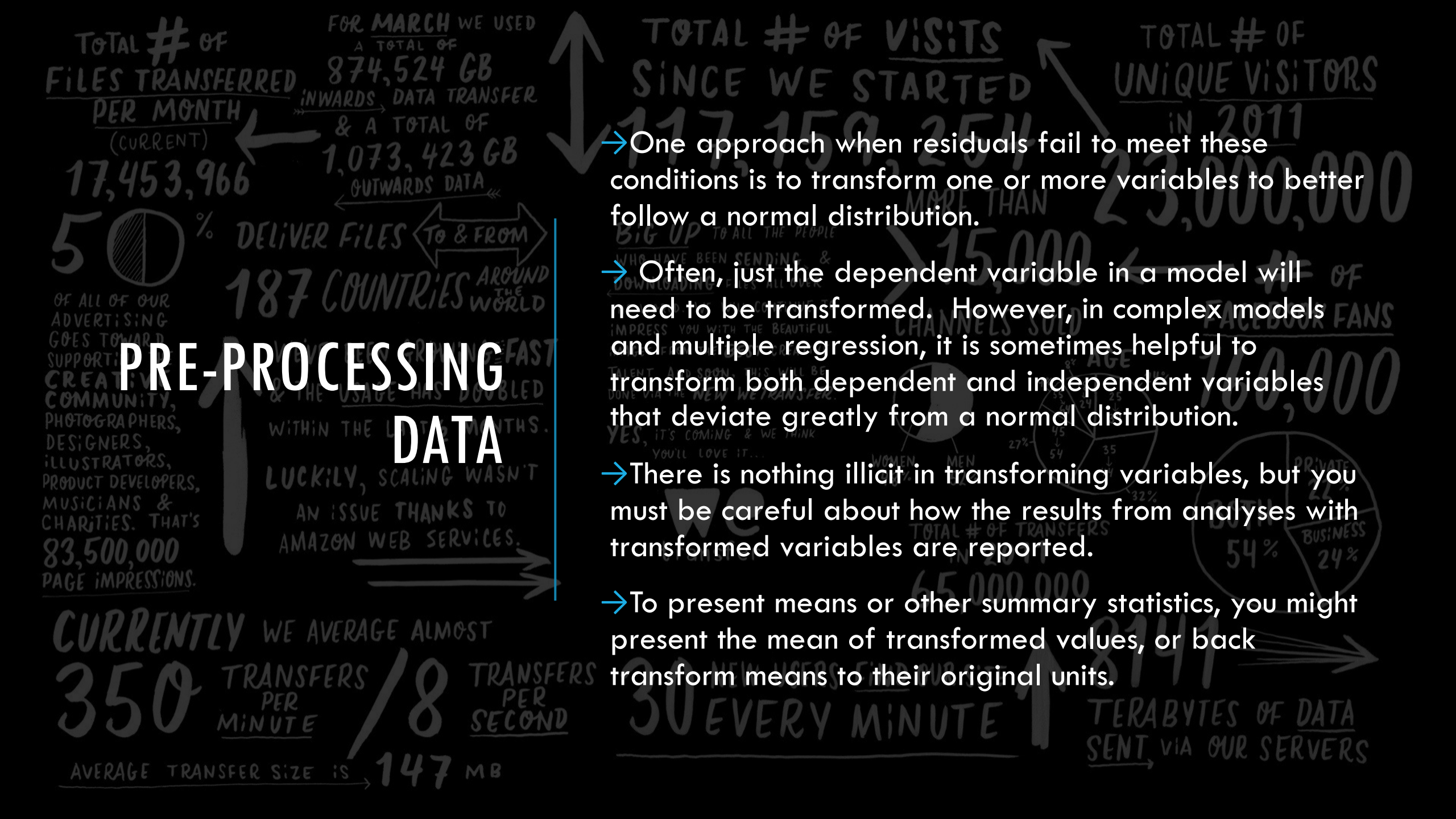
Transforming data

→Transforming data is one step in addressing data that do not fit model assumptions and is also used to coerce different variables to have similar distributions.

→Most parametric tests require that residuals be normally distributed and that the residuals be homoscedastic.

In statistics, a sequence (or a vector) of random variables is homoscedastic if all its random variables have the same finite variance. This is also known as homogeneity of variance





PRE-PROCESSING DATA

→One approach when residuals fail to meet these conditions is to transform one or more variables to better follow a normal distribution.

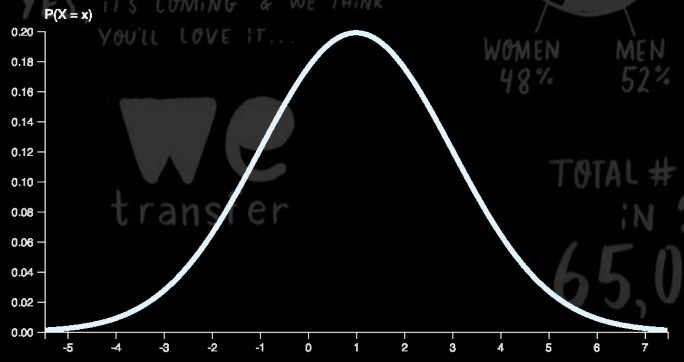
→ Often, just the dependent variable in a model will need to be transformed. However, in complex models and multiple regression, it is sometimes helpful to transform both dependent and independent variables that deviate greatly from a normal distribution.

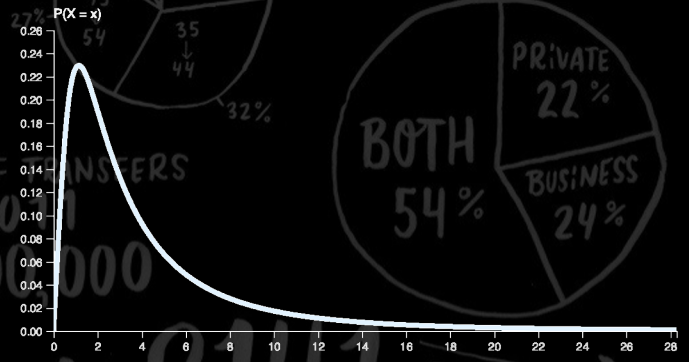
→There is nothing illicit in transforming variables, but you must be careful about how the results from analyses with transformed variables are reported.

→To present means or other summary statistics, you might present the mean of transformed values, or back transform means to their original units.

PRE-PROCESSING DATA

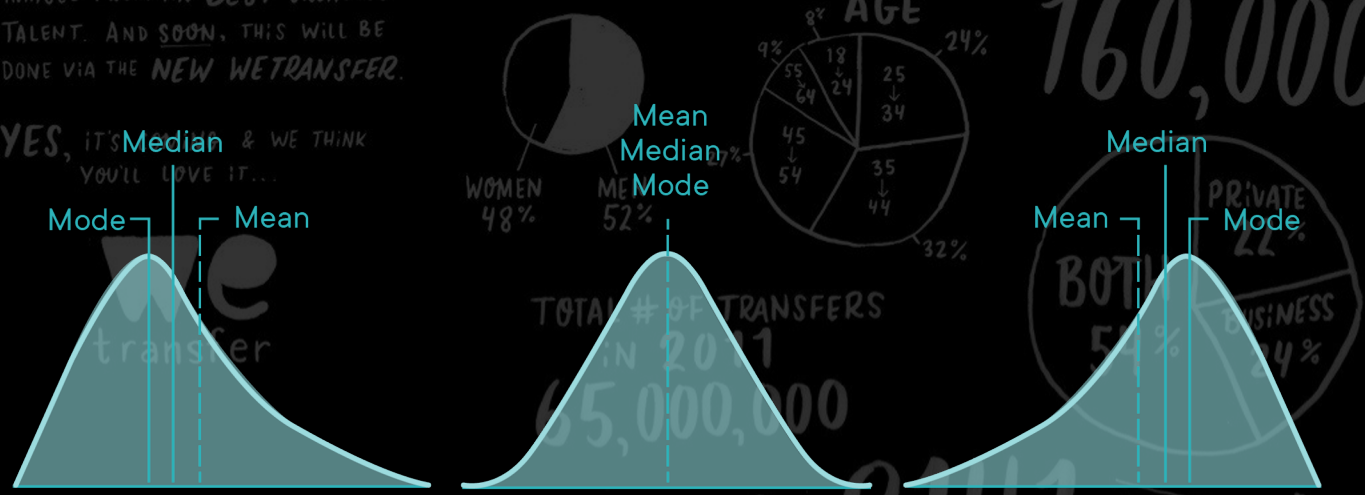
- Some measurements in nature are naturally normally distributed. Other measurements are naturally log-normally distributed.
- There may be many low values with fewer high values and even fewer very high values.





PRE-PROCESSING DATA

- Usually, data can deviate from normality, meaning that the distribution in skewed.
- Right-skewed data - tail is on the right, positive skew
- Left-skewed data - tail is on the left, negative skew



HOW TO MANAGE DATA DISTRIBUTIONS

- For right-skewed data common transformations include square root, cube root, and log.
- For left-skewed data common transformations include square root (constant – x), cube root (constant – x), and log (constant – x).
- Because log(0) is undefined, as is the log of any negative number, when using a log transformation a constant should be added to all values to make them all positive before transformation.
- Another approach is to use a general power transformation, such as Tukey's Ladder of Powers or a Box–Cox transformation.

DATA NORMALIZATION/ STANDARDIZATION

→Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

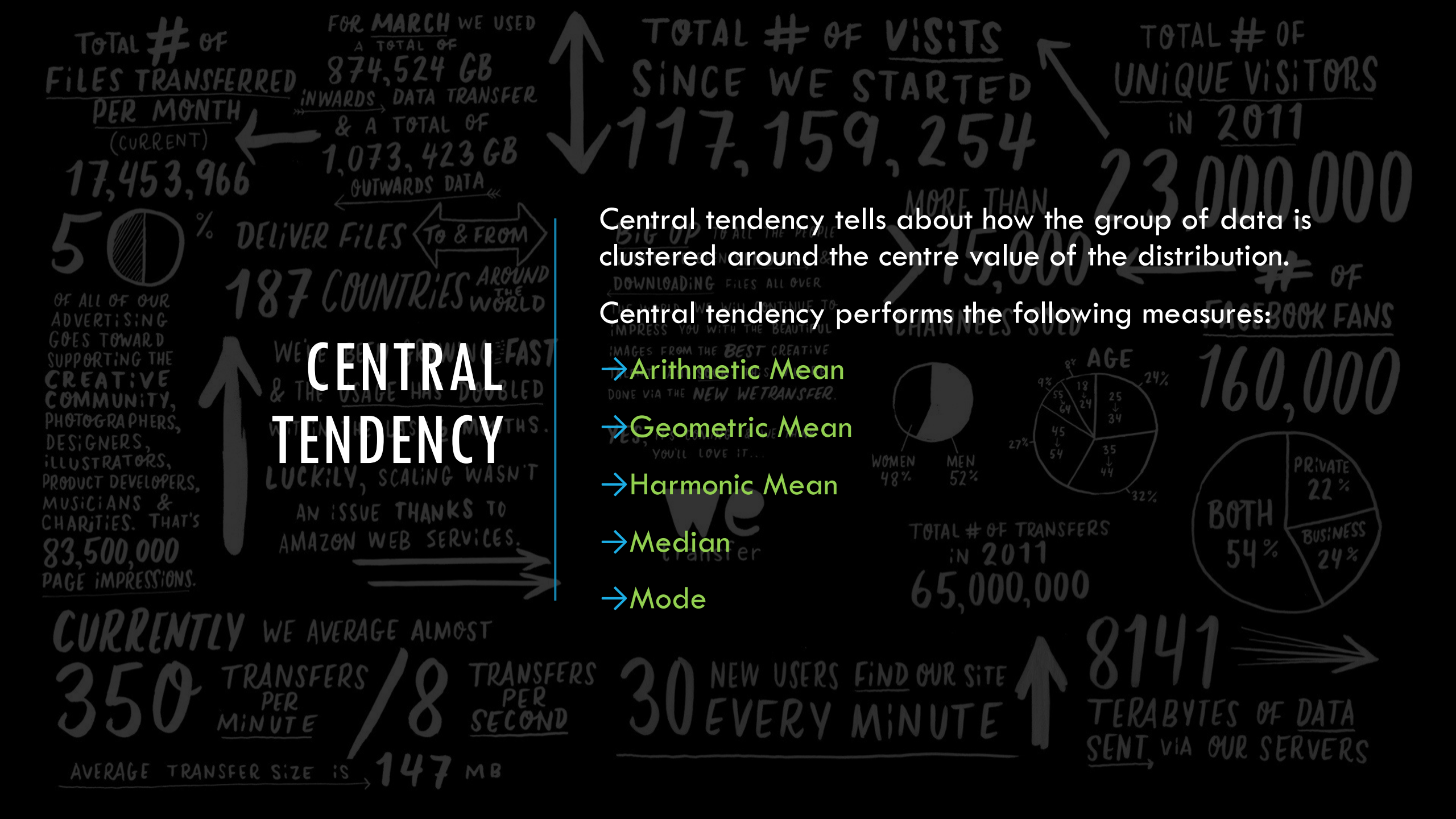
→Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation.

$$x_{std} = \frac{x - \mu}{\sigma}$$

THE BIG QUESTION: NORMALIZE OR STANDARDIZE?

→Normalization is good to use when you know that the distribution of your data does not follow a Gaussian distribution.

→Standardization, on the other hand, can be helpful in cases where the data follows a Gaussian distribution. However, this does not have to be necessarily true. Also, unlike normalization, standardization does not have a bounding range. So, even if you have outliers in your data, they will not be affected by standardization.



CENTRAL TENDENCY

Central tendency tells about how the group of data is clustered around the centre value of the distribution.

Central tendency performs the following measures:

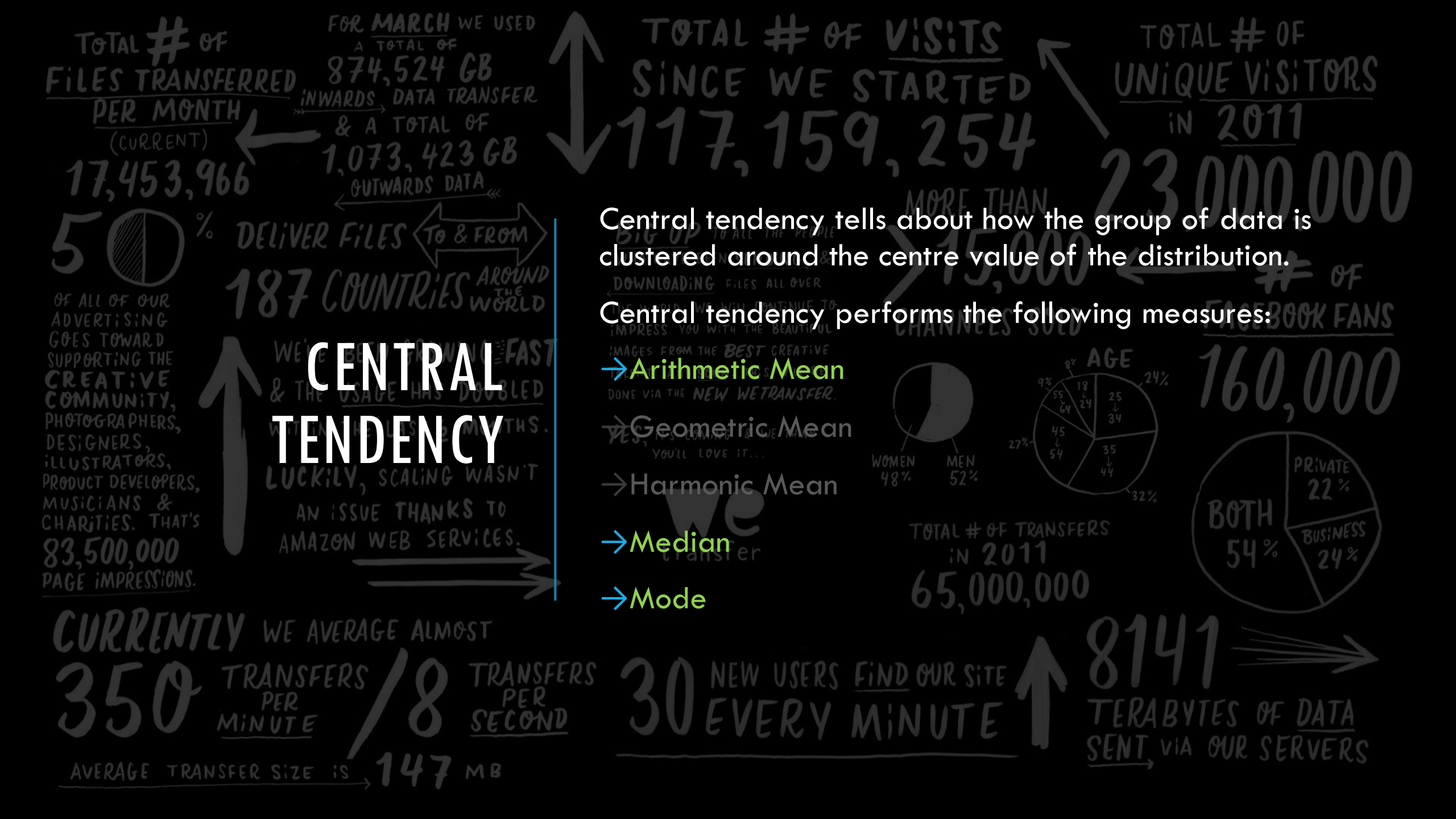
→Arithmetic Mean

→Geometric Mean

→Harmonic Mean

→Median

→Mode



CENTRAL TENDENCY

Central tendency tells about how the group of data is clustered around the centre value of the distribution.

Central tendency performs the following measures:

→Arithmetic Mean

→Geometric Mean

→Harmonic Mean

→Median

→Mode

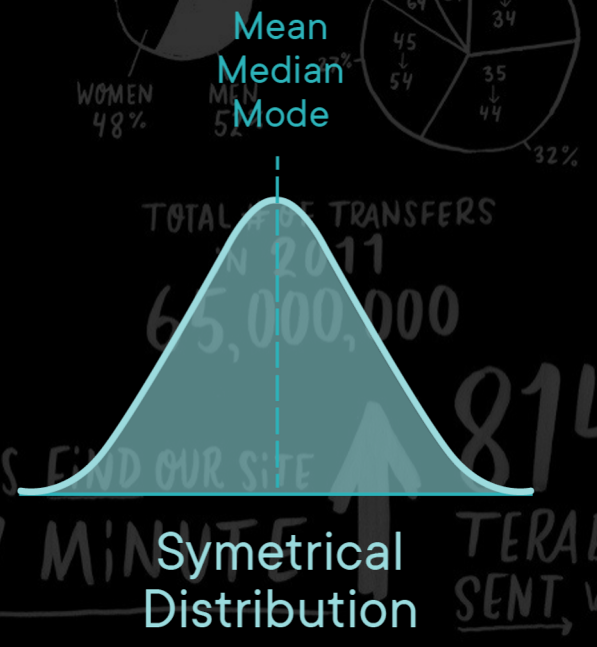
CENTRAL TENDENCY

- →Arithmetic Mean - is the sum of a collection of numbers divided by the count of numbers in the collection
- →Median - the value separating the higher half from the lower half of a data sample, a population, or a probability distribution.
- →Mode - the value that appears most often in a set of data values

CENTRAL TENDENCY

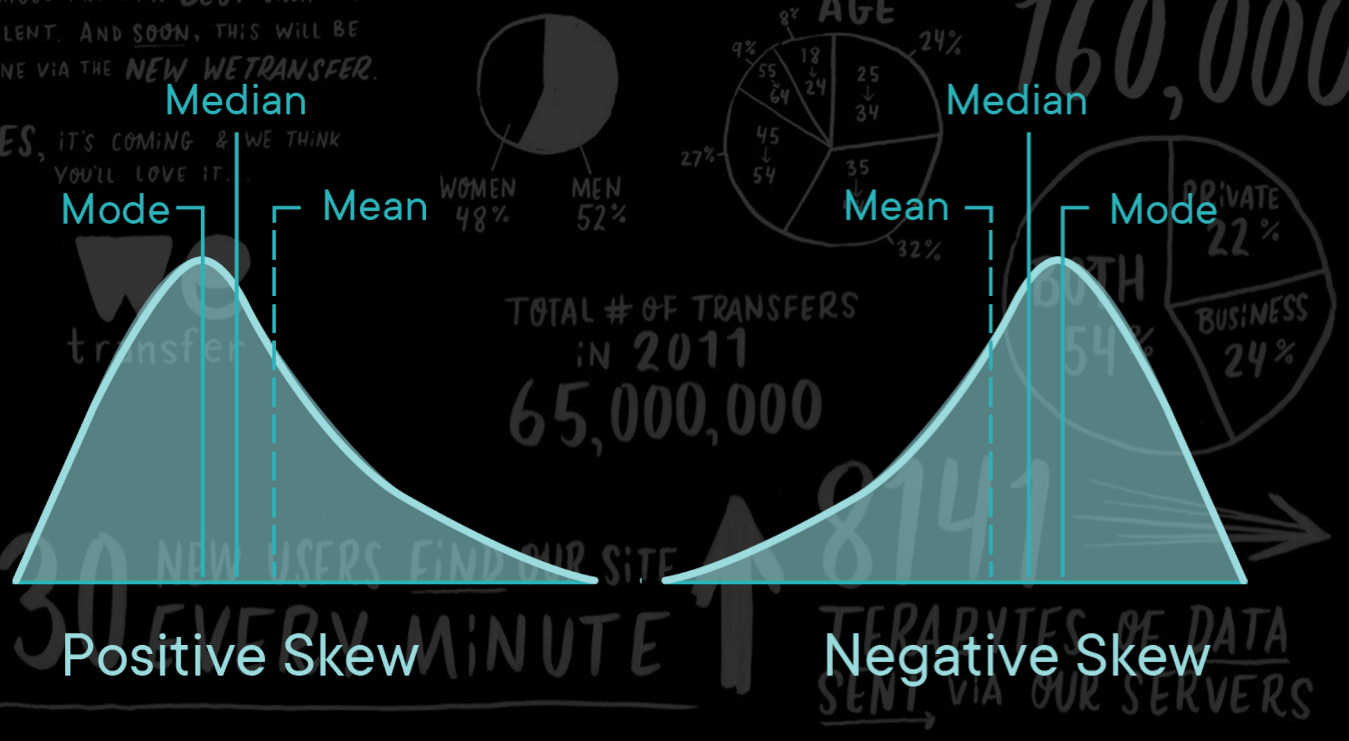
The preferred measure of central tendency often depends on the shape of the distribution. Of the three measures of tendency, the mean is most heavily influenced by any outliers or skewness.

In a symmetrical distribution, the mean, median, and mode are all equal. In these cases, the mean is often the preferred measure of central tendency.



CENTRAL TENDENCY

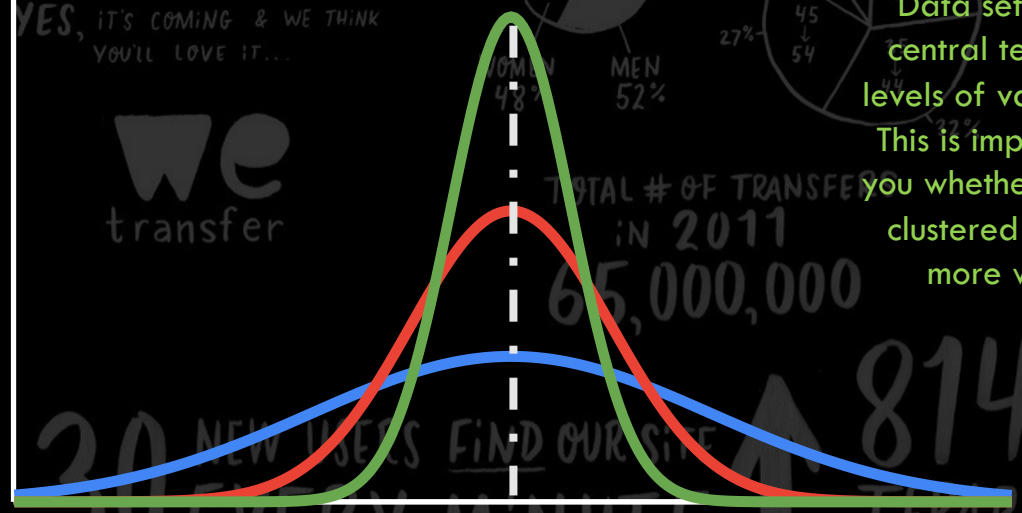
For distributions that have outliers or are skewed, the median is often the preferred measure of central tendency because the median is more resistant to outliers than the mean. Below you will see how the direction of skewness impacts the order of the mean, median, and mode. Note that the mean is pulled in the direction of the skewness (i.e., the direction of the tail).



VARIABILITY IN THE DISTRIBUTION OF DATA

In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution.

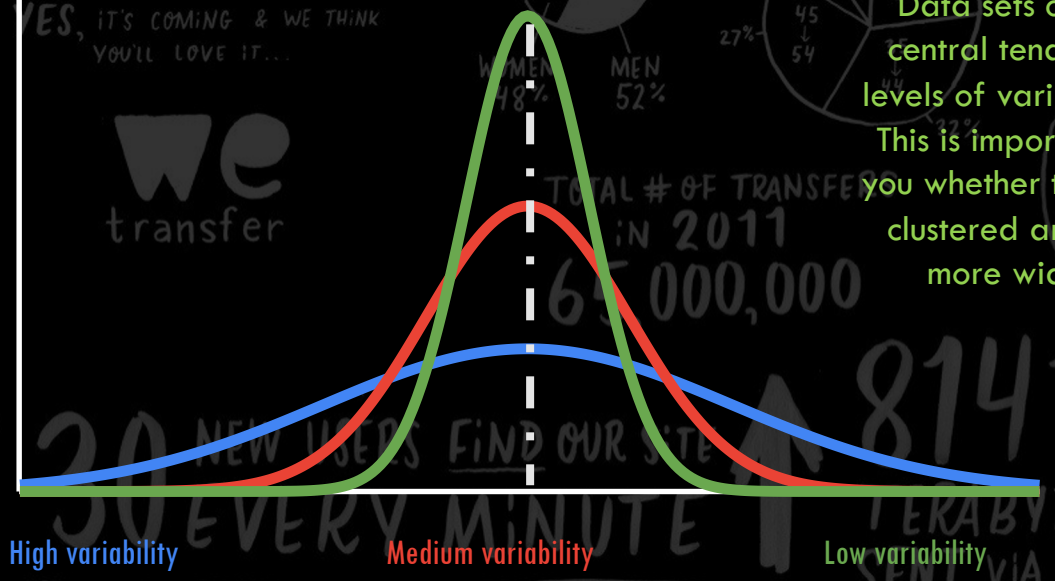
While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the centre.



Data sets can have the same central tendency but different levels of variability or vice versa. This is important because it tells you whether the points tend to be clustered around the center or more widely spread out.

VARIABILITY IN THE DISTRIBUTION OF DATA

Low variability is ideal because it means that you can better predict information about the population based on sample data. High variability means that the values are less consistent, so it's harder to make predictions.



Data sets can have the same central tendency but different levels of variability or vice versa. This is important because it tells you whether the points tend to be clustered around the center or more widely spread out.

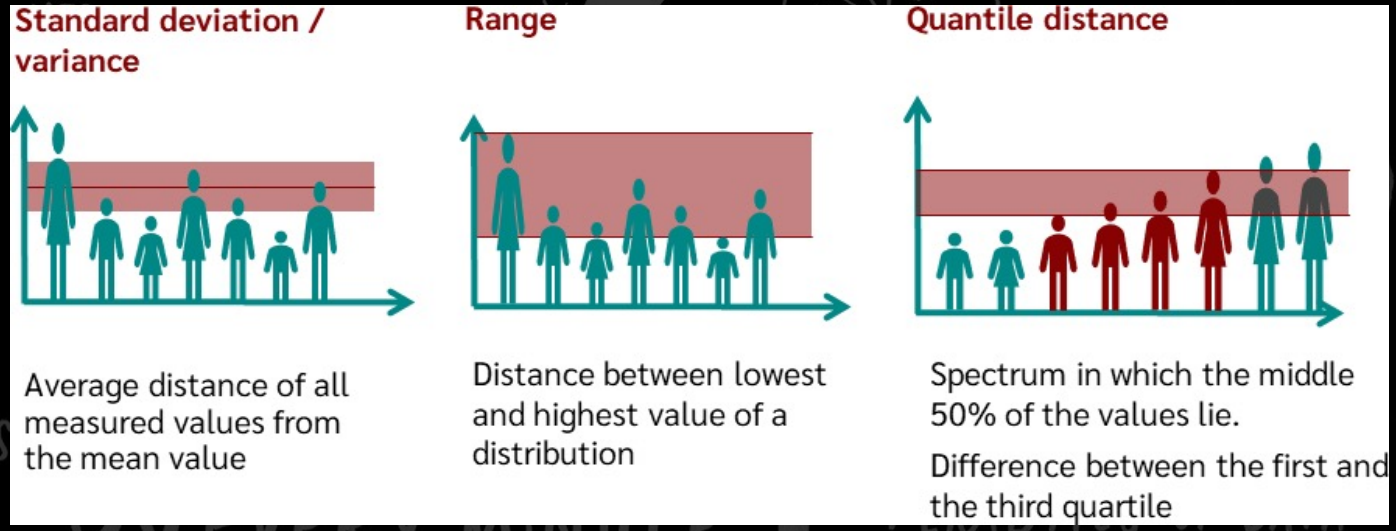
MEASURES OF VARIABILITY

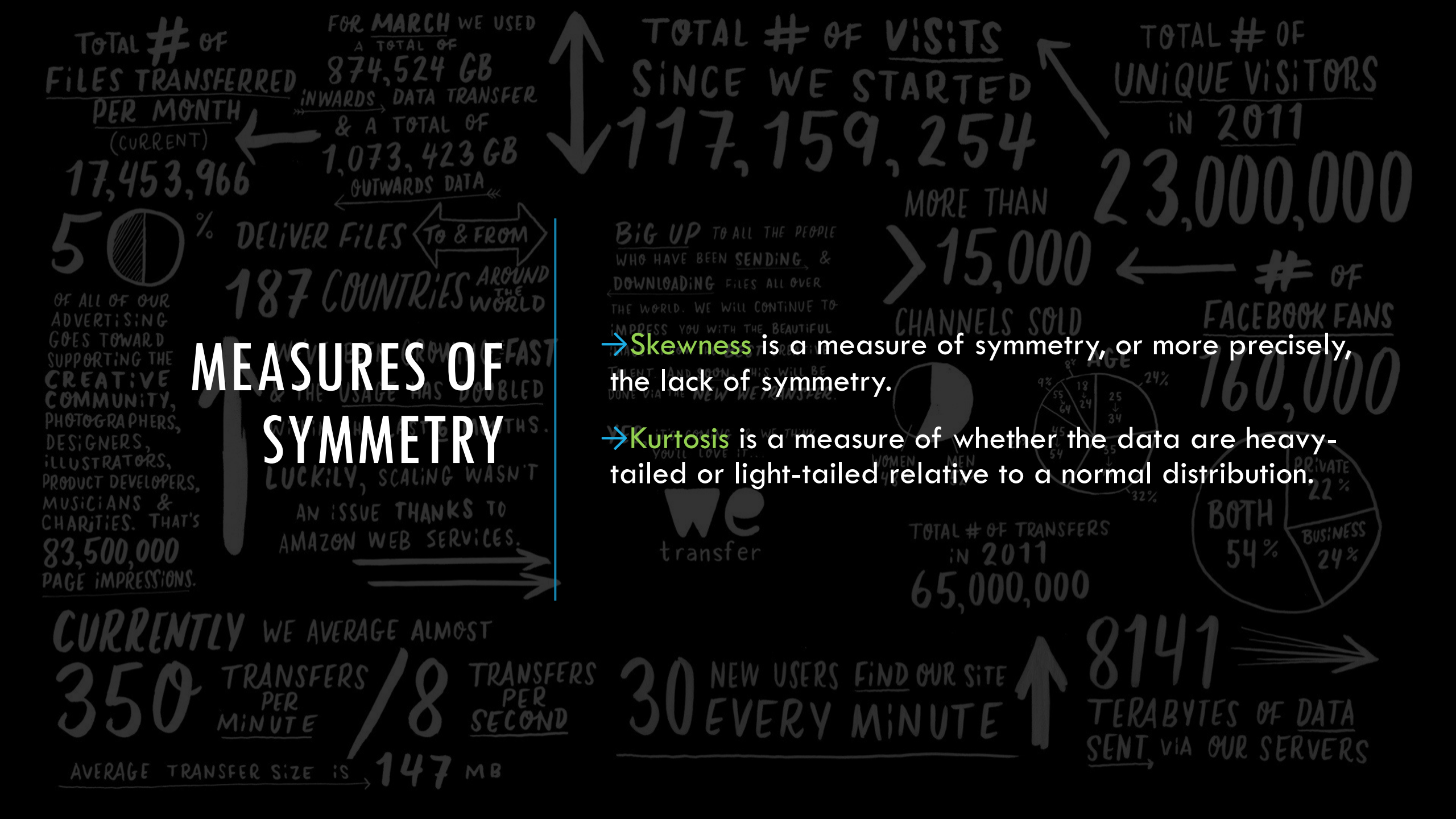
→Range - the difference between the highest and lowest values

→Interquartile range - the range of the middle half of a distribution

→Standard deviation - average distance from the mean

→Variance - average of squared distances from the mean





MEASURES OF SYMMETRY

→Skewness is a measure of symmetry, or more precisely, the lack of symmetry.

→Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution.

RANGE

The range tells you the spread of your data from the lowest to the highest value in the distribution. To find the range, simply subtract the lowest value from the highest value in the data set.

Example

You have 8 data points from Sample A.

Data 72-110-134-190-238-287-305-324

The highest value (H) is 324 and the lowest (L) is 72.

$$R = H - L$$

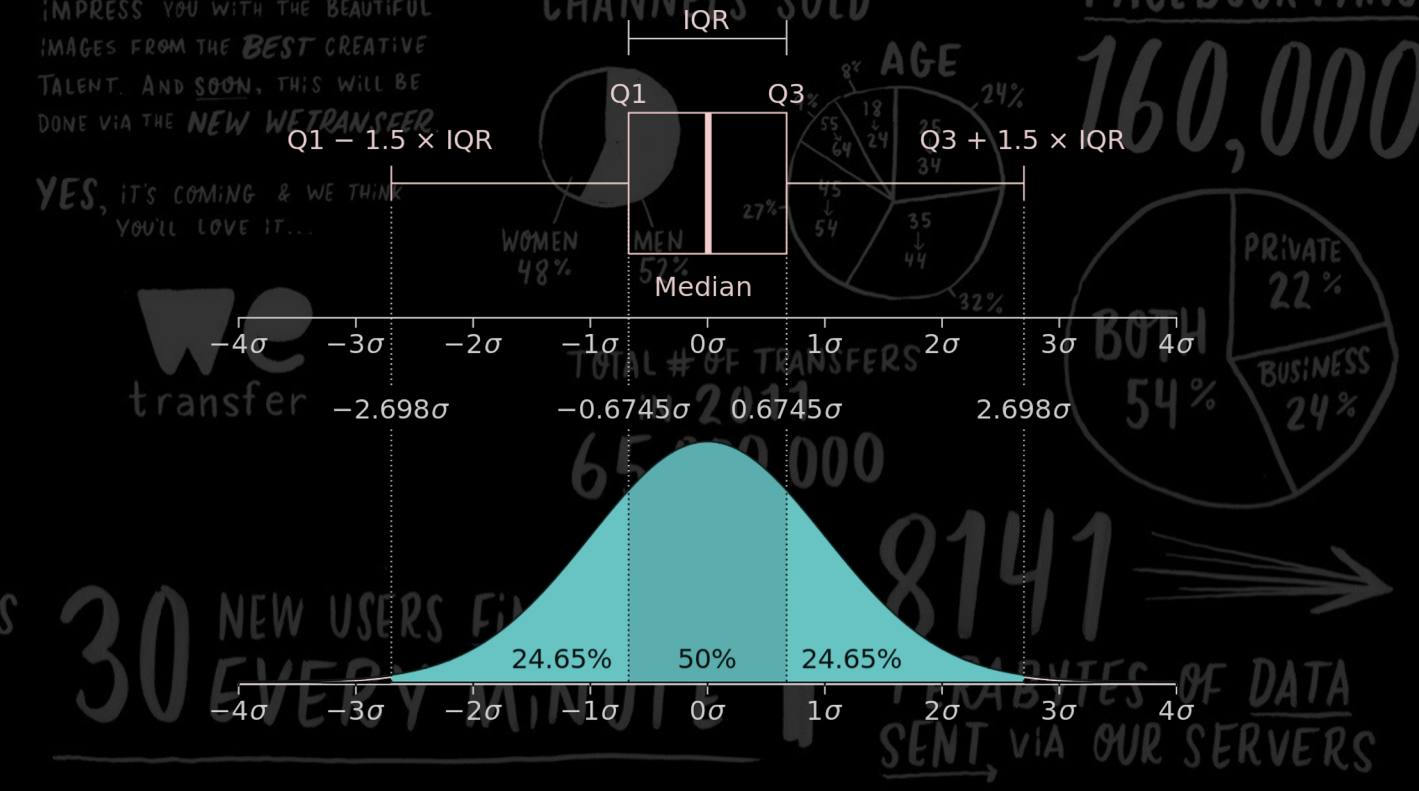
$$R = 324 - 72 = 252$$

The range of your data is 252

INTERQUARTILE RANGE

The interquartile range gives you the spread of the middle of your distribution.

The interquartile range contains half of the values. While the first quartile (Q1) contains the first 25% of values, the fourth quartile (Q4) contains the last 25% of values.



FIVE-NUMBER SUMMARY

Every distribution can be organized using a five-number summary:

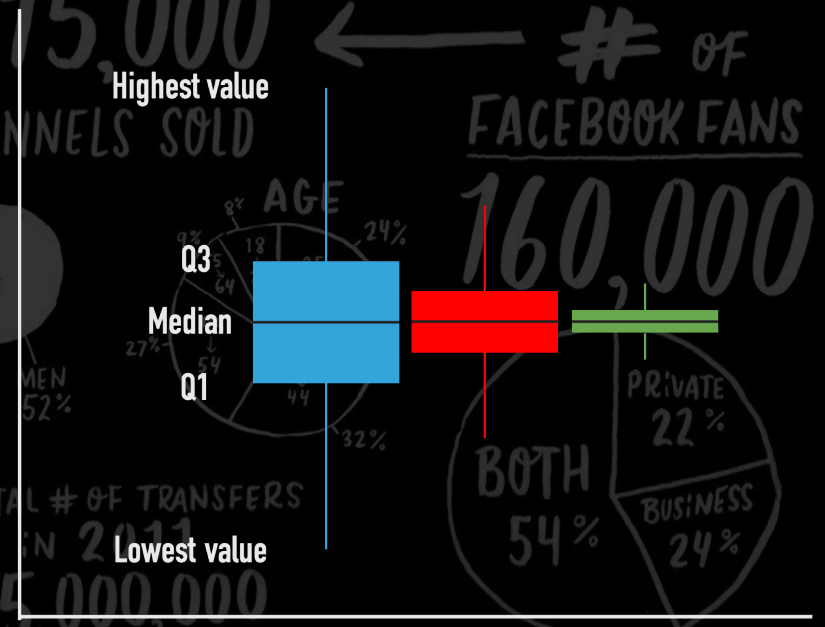
→Lowest value

→Q1: 25th percentile

→Q2: the median

→Q3: 75th percentile

→Highest value (Q4)



STANDARD DEVIATION

The standard deviation is the average amount of variability in your dataset.

It tells you, on average, how far each score lies from the mean. The larger the standard deviation, the more variable the data set is.

→List each score and find their mean.

→Subtract the mean from each score to get the deviation from the mean.

→Square each of these deviations.

→Add up all of the squared deviations.

→Divide the sum of the squared deviations by n – 1 (for a sample) or N (for a population).

→Find the square root of the number you found.

STANDARD DEVIATION

Standard deviation formula for populations

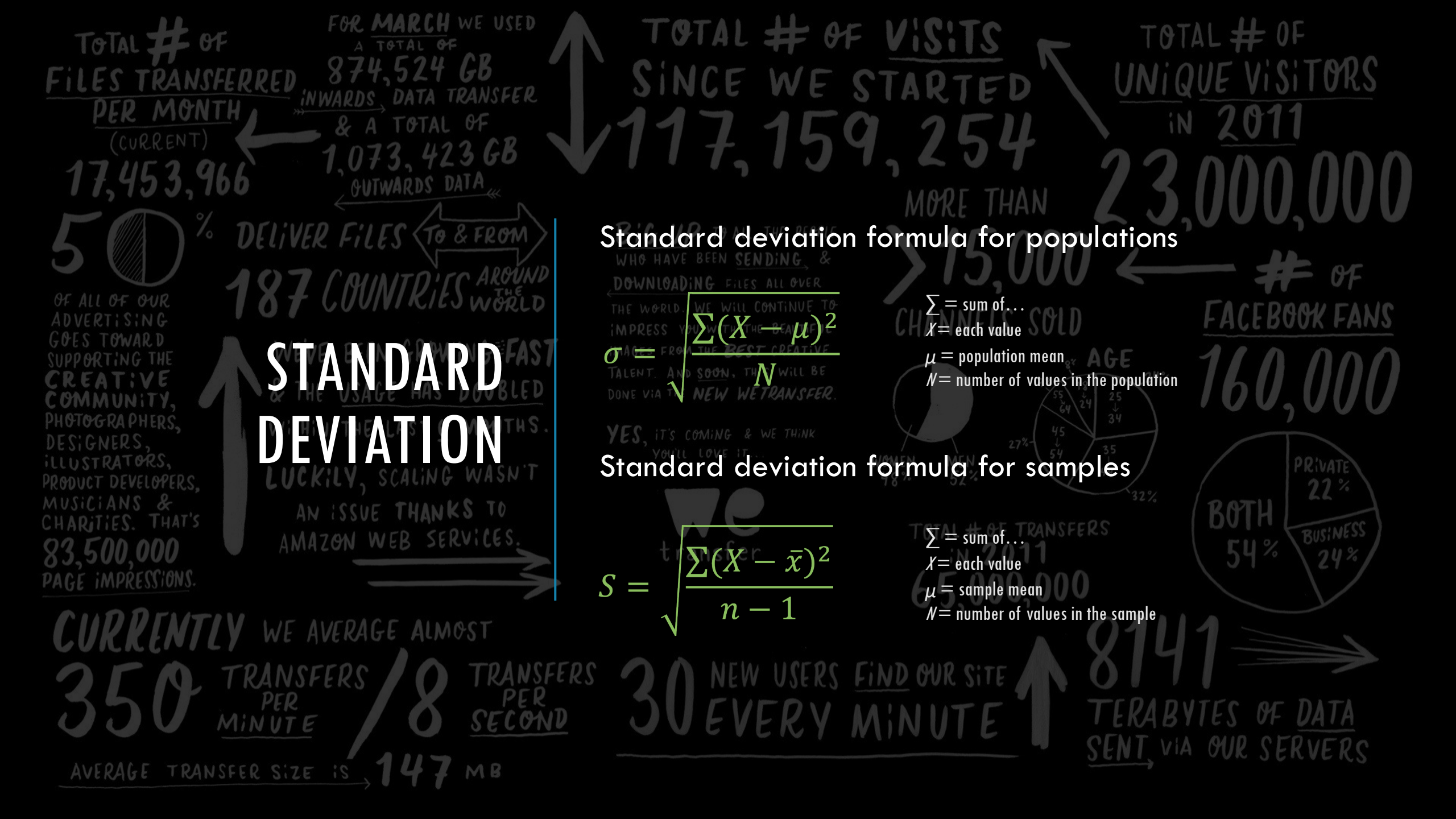
$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

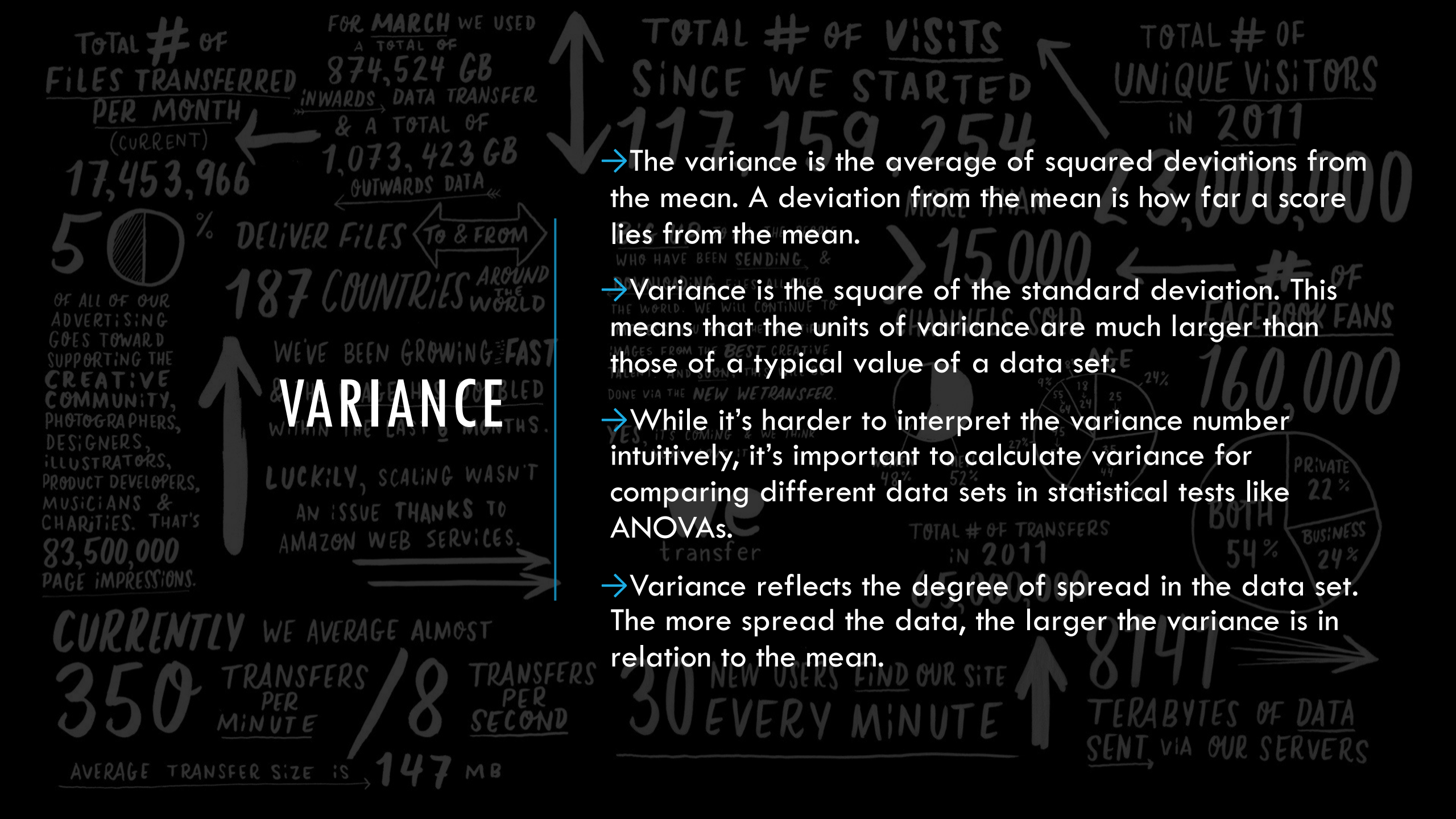
$\sum$ = sum of…
X = each value
μ = population mean
N = number of values in the population

Standard deviation formula for samples

$$S = \sqrt{\frac{\sum(X-\bar{x})^2}{n-1}}$$

$\sum$ = sum of…
X = each value
μ = sample mean
N = number of values in the sample





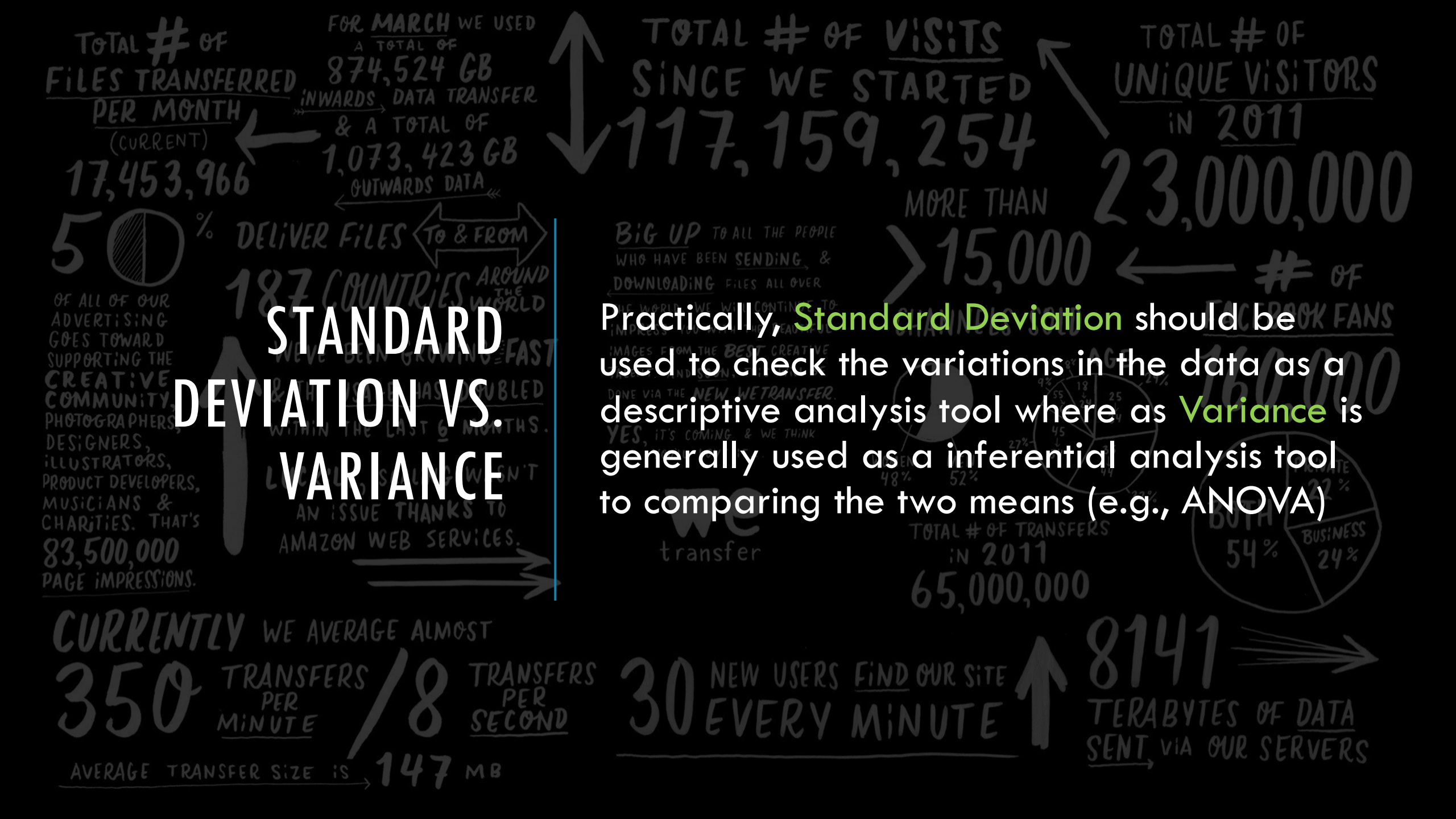
VARIANCE

→The variance is the average of squared deviations from the mean. A deviation from the mean is how far a score lies from the mean.

→Variance is the square of the standard deviation. This means that the units of variance are much larger than those of a typical value of a data set.

→While it's harder to interpret the variance number intuitively, it's important to calculate variance for comparing different data sets in statistical tests like ANOVAs.

→Variance reflects the degree of spread in the data set. The more spread the data, the larger the variance is in relation to the mean.



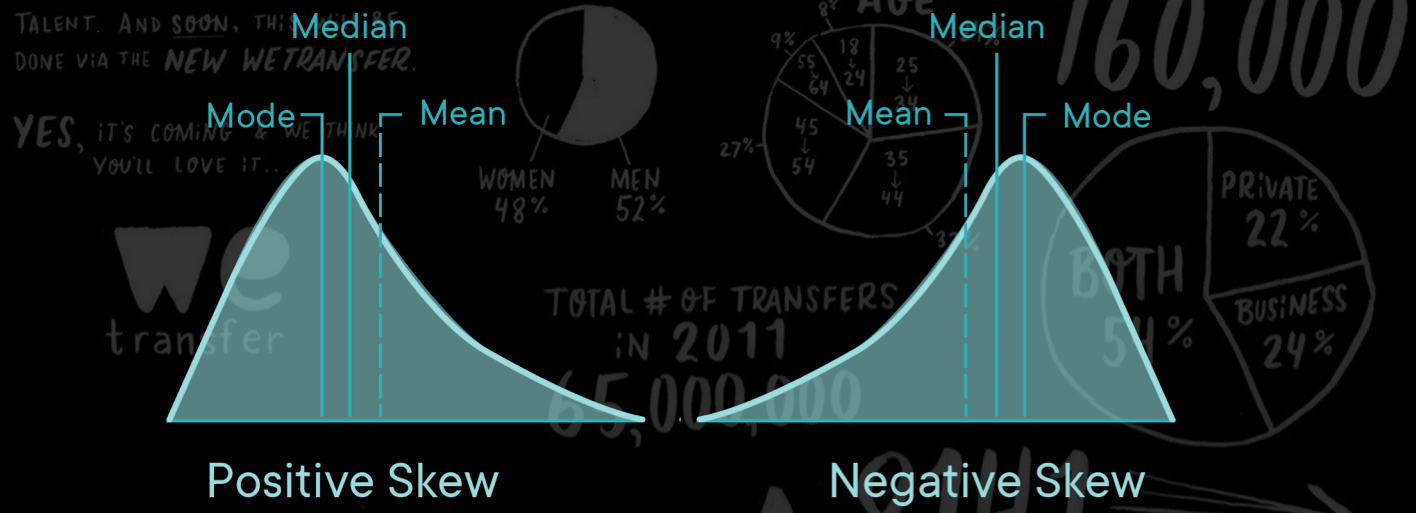
STANDARD DEVIATION VS. VARIANCE

Practically, Standard Deviation should be used to check the variations in the data as a descriptive analysis tool where as Variance is generally used as a inferential analysis tool to comparing the two means (e.g., ANOVA)

MEASURES OF VARIABILITY: SKEWNESS & KURTOSIS

Both measures are also known as statistical moments

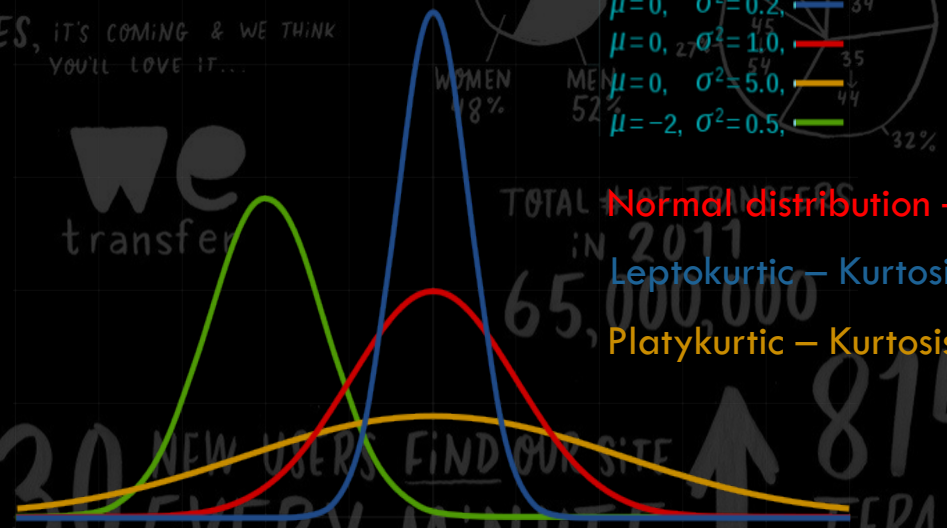
Skewness is a measure of the asymmetry of the probability distribution assuming a unimodal distribution and is given by the third standardized moment.



MEASURES OF VARIABILITY: SKEWNESS & KURTOSIS

Both measures are also known as statistical moments

Kurtosis describes the tailedness of the distribution as it describes the shape of it. It is also a measure of the peakedness of the distribution. A high kurtosis distribution has a sharper peak and longer fatter tails, while a low kurtosis distribution has a more rounded pean and shorter thinner tails.



Normal distribution – Kurtosis = 0

Leptokurtic – Kurtosis > 0

Platykurtic – Kurtosis < 0

UNDERSTANDING CONFIDENCE INTERVALS

Confidence intervals are used to indicate how accurate a calculated statistic is likely to be. Confidence intervals can be calculated for a variety of statistics, such as the mean, median, or slope of a linear regression.

Most of the statistics we use assume we are analysing a sample which we are using to represent a larger population.

If we collect a large sample and the values aren't too variable, then the sample mean should be close to the population mean. But if we have few observations, or the values are highly variable, we are less confident our sample mean is close to the population mean.

We will use confidence intervals to give a sense of this confidence.

UNDERSTANDING CONFIDENCE INTERVALS

Most of statistical tests calculate a probability (p-value) of the likelihood of data and draw a conclusion from this p-value.

The traditional method is the most commonly encountered, and is appropriate for normally distributed data or with large sample sizes. It produces an interval that is symmetric about the mean.

For skewed data, confidence intervals by bootstrapping may be more reliable.

UNDERSTANDING CONFIDENCE INTERVALS

Bootstrapping is a method that samples the data many times, each time calculating a statistic, and then determining a confidence interval or other statistic from these iterations.