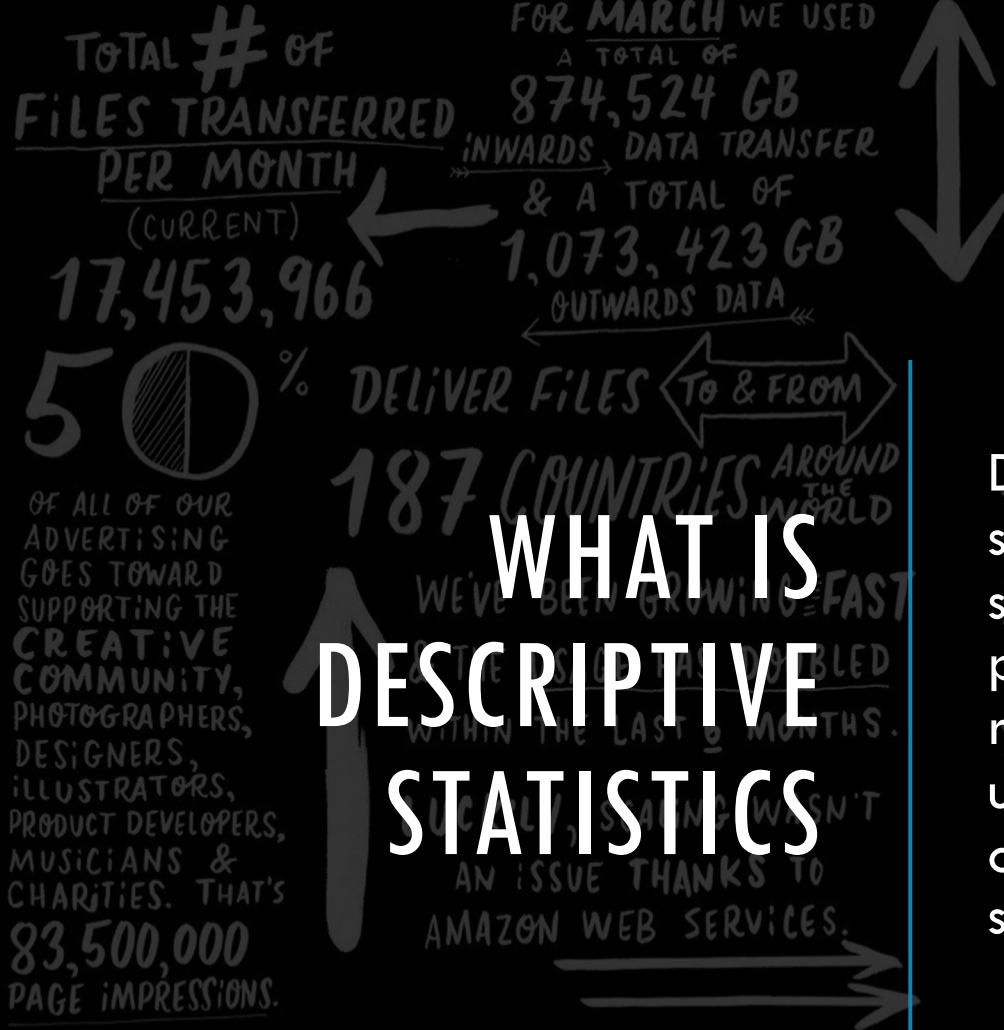


DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

SESSION 2 DESCRIPTIVE STATISTICS



TOTAL # OF VISITS SINCE WE STARTED
117,159,254

MORE THAN 15000

BIG UP TO ALL THE PEOPLE WHO HAVE BEEN SENDING & RECEIVING. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVES IN THE WORLD. AND SOON, THIS WILL BE THE BIGGEST CHANNEL IN 2013.

TOTAL # OF TRANSFERS IN 2011
65,000,000

TOTAL # OF UNIQUE VISITORS IN 2011
23,000,000

OF FACEBOOK FANS
160,000

WE transfer

30 NEW USERS FIND OUR SITE EVERY MINUTE

8141 TERABYTES OF DATA SENT VIA OUR SERVERS

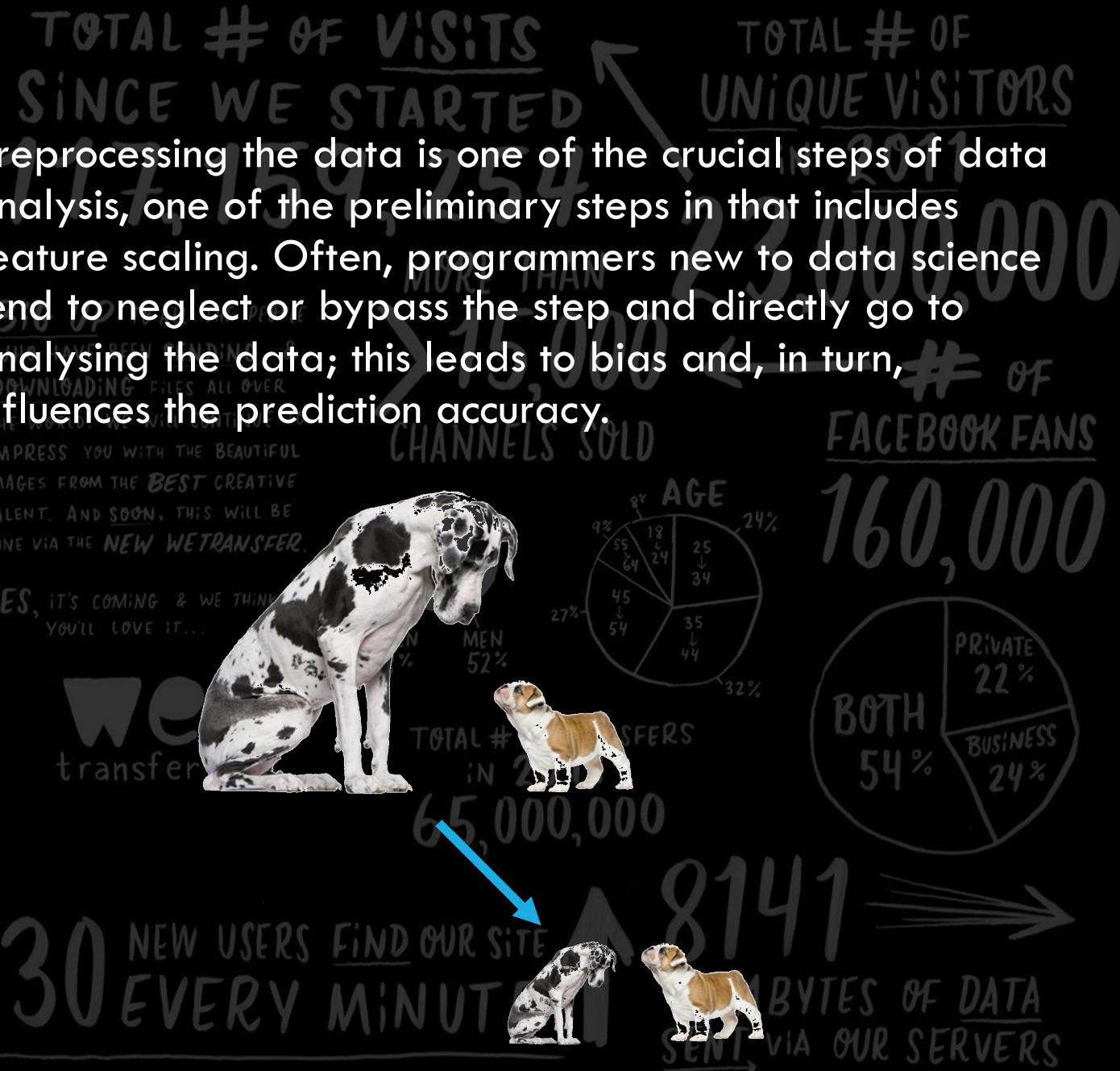
Descriptive statistics is distinguished from inferential statistics (or inductive statistics) by its aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential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



CURRENTLY WE AVERAGE ALMOST 350 TRANSFERS PER MINUTE

18 TRANSFERS PER SECOND

AVERAGE TRANSFER SIZE IS 147 MB



PRE-PROCESSING DATA

TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT) **17,453,966**

FOR MARCH WE US
A TOTAL OF
874,524 GB

INWARDS DATA TRANSF
& A TOTAL OF
1,073,423 GB

OUTWARDS DATA

50% DELIVER FILES 
TO & FRO
OF ALL OF OUR
ADVERTISING
187 COUNTRIES  ARE
THROUGHOUT THE WORLD

PRE-PROCESSING DATA

CURRENTLY WE AVERAGE ALMOST
350 TRANSFERS PER MINUTE / 18 TRANSFERS PER SECOND
AVERAGE TRANSFER SIZE IS 147 MB

TOTAL # OF VISITS SINCE WE STARTED

117,159,254

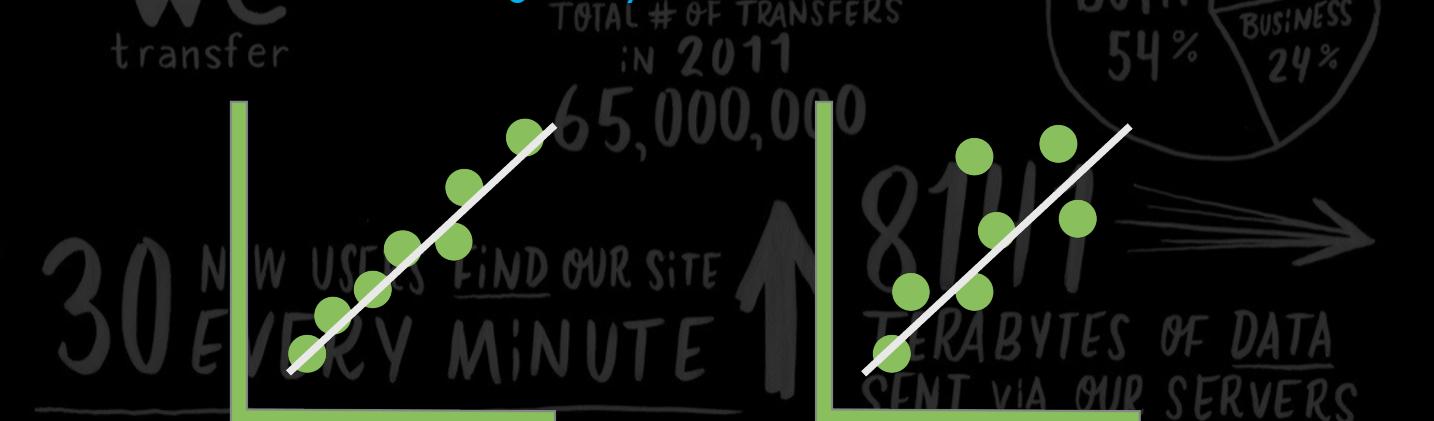
Transforming data

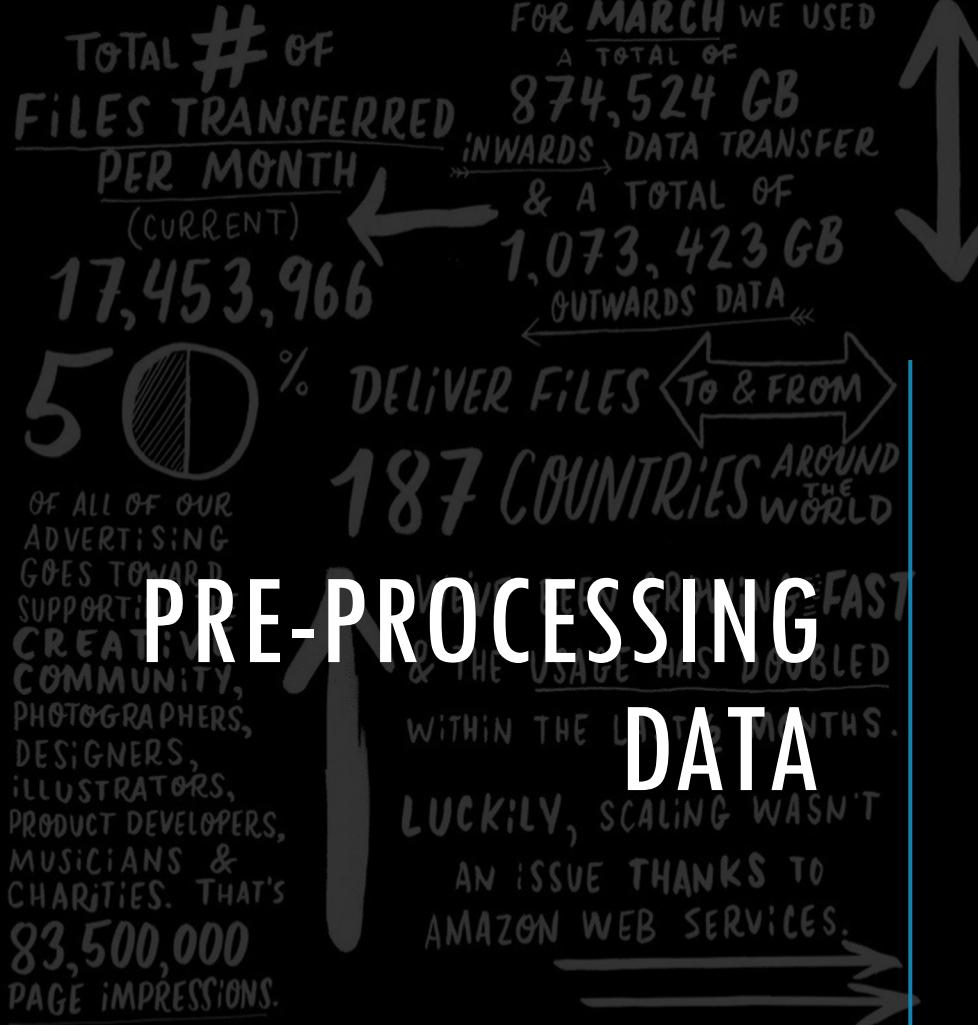
TOTAL # OF UNIQUE VISITORS IN 2011

23,000,000

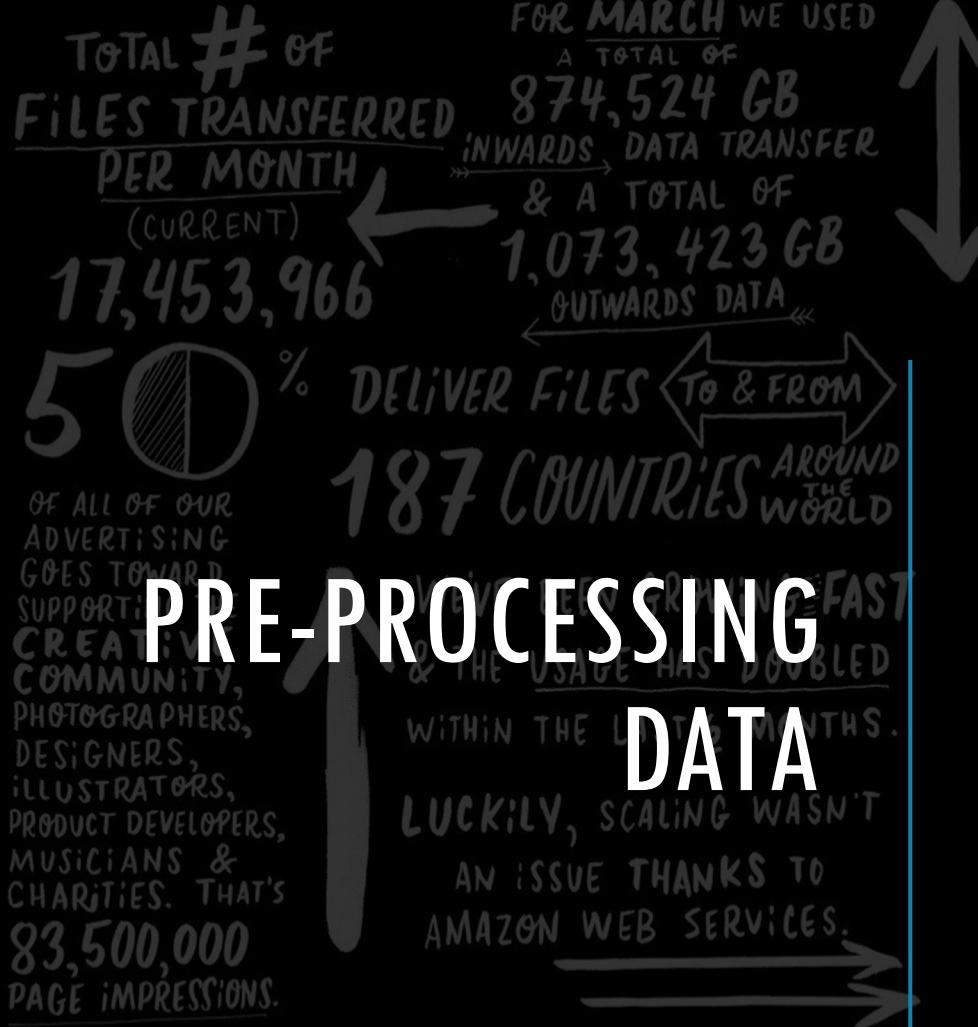
→ Most parametric tests require that residuals be normally distributed and that the residuals be homoscedastic.

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.





- TOTAL # OF VISITS SINCE WE STARTED**
- TOTAL # OF UNIQUE VISITORS IN 2011**
- 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.



CURRENTLY WE AVERAGE ALMOST 350 TRANSFERS PER MINUTE / 18 TRANSFERS PER SECOND

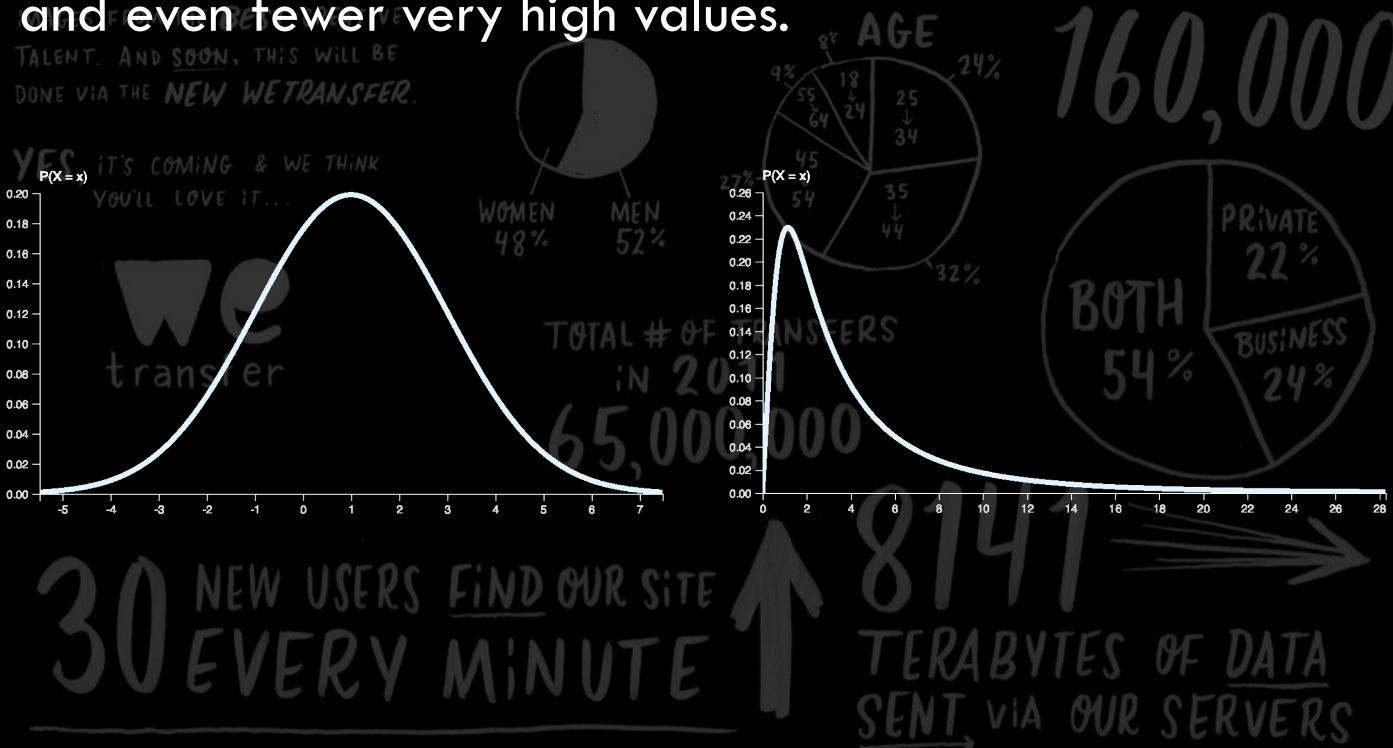
AVERAGE TRANSFER SIZE IS → 147 MB

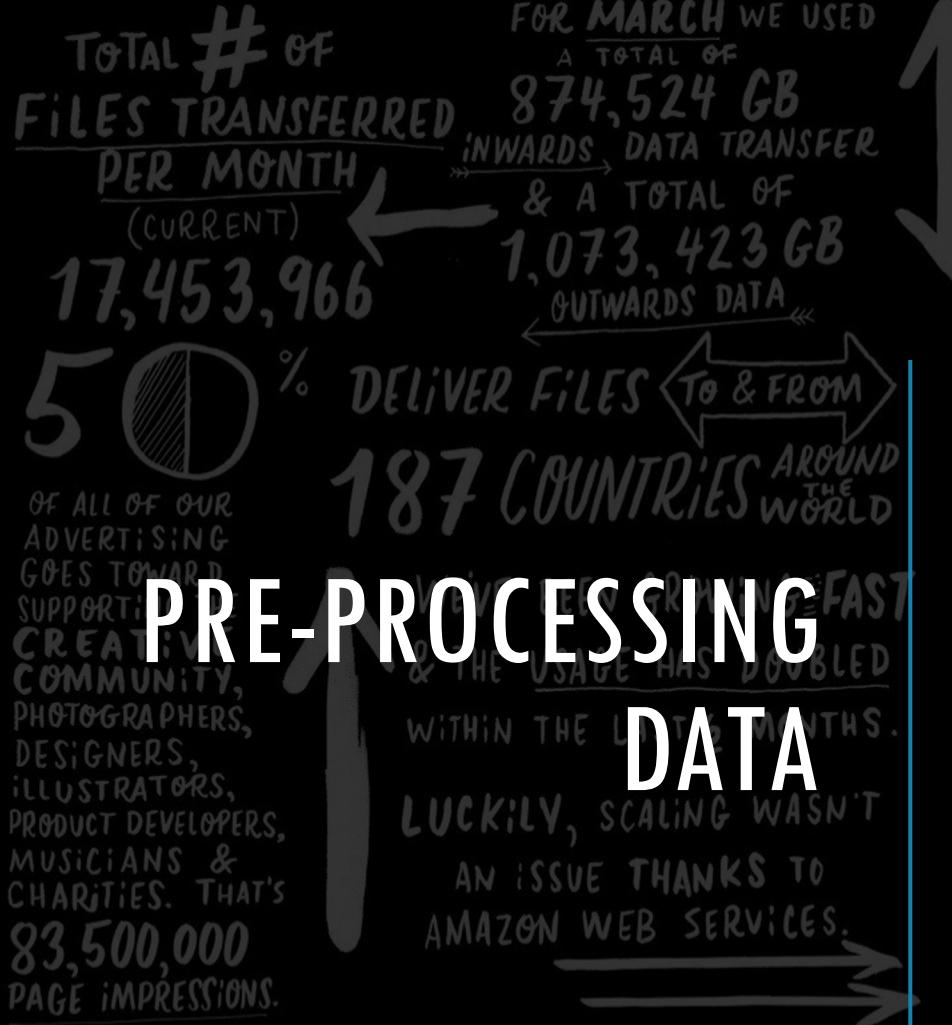
TOTAL # OF VISITS SINCE WE STARTED

117,159,254

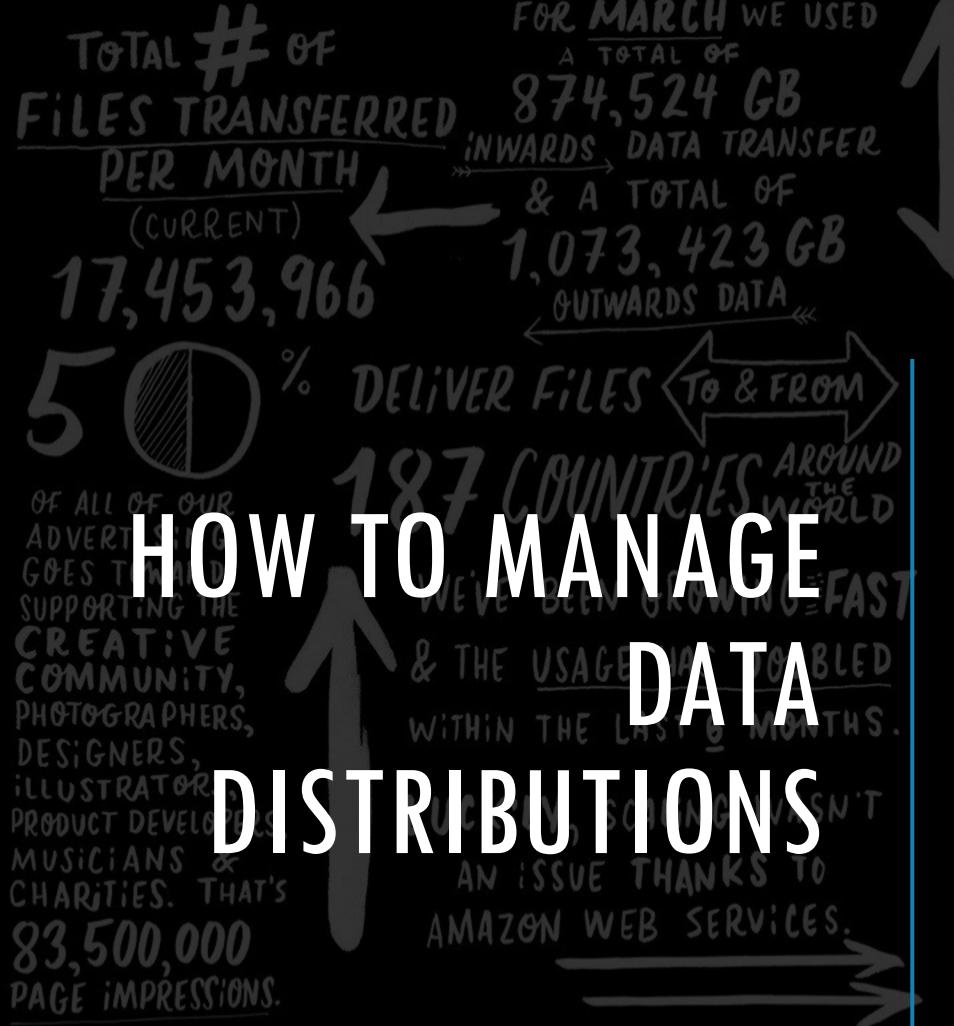
→ Some measurements in nature are naturally normally distributed. Other measurements are naturally log-normally distributed.

→ There may be many low values with fewer high values and even fewer very high values.





HOW TO MANAGE DATA DISTRIBUTIONS



- TOTAL # OF VISITS SINCE WE STARTED**: 117,159,254
- TOTAL # OF UNIQUE VISITORS IN 2011**: 23,000,000
- 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.
- NEW USERS FIND OUR SITE EVERY MINUTE**: 30
- TERABYTES OF DATA SENT VIA OUR SERVERS**: 8141

TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT)
 FOR MARCH WE USED A TOTAL OF 874,524 GB INWARDS DATA TRANSFER & A TOTAL OF 1,073,423 GB OUTWARDS DATA



DATA NORMALIZATION/ STANDARDIZATION

TOTAL # OF VISITS SINCE WE STARTED
 117,159,254

→ 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.

BIG UP TO THE PEOPLE WHO HAVE BEEN SENDING & DOWNLOADING FILES ALL OVER THE WORLD. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVE TALENT. AND SOON, THIS WILL BE DONE VIA THE NEW WETRANSFER.

YES, IT'S COMING & WE THINK YOU'LL LOVE IT.

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

TOTAL # OF TRANSFERS IN 2011
 65,000,000

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



TOTAL # OF TRANSFERS IN 2011
 65,000,000

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

30 NEW USERS FIND OUR SITE EVERY MINUTE

8141 TERABYTES OF DATA SENT VIA OUR SERVERS



TOTAL # OF UNIQUE VISITORS IN 2011
 23,000,000

OF FACEBOOK FANS
 160,000



THE BIG QUESTION: NORMALIZE OR STANDARDIZE?

TOTAL # OF VISITS SINCE WE STARTED

117,159,254

MORE THAN 15,000 CHANNELS SOLD

→ 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.

TOTAL # OF UNIQUE VISITORS IN 2011

23,000,000

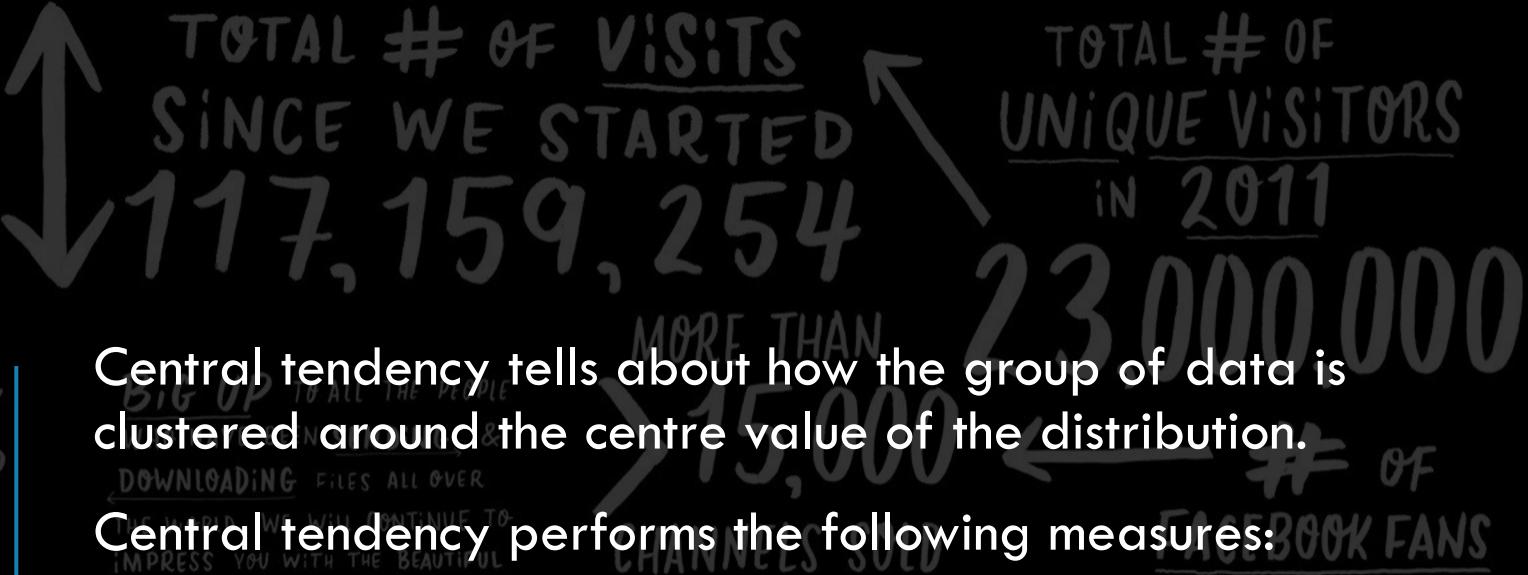
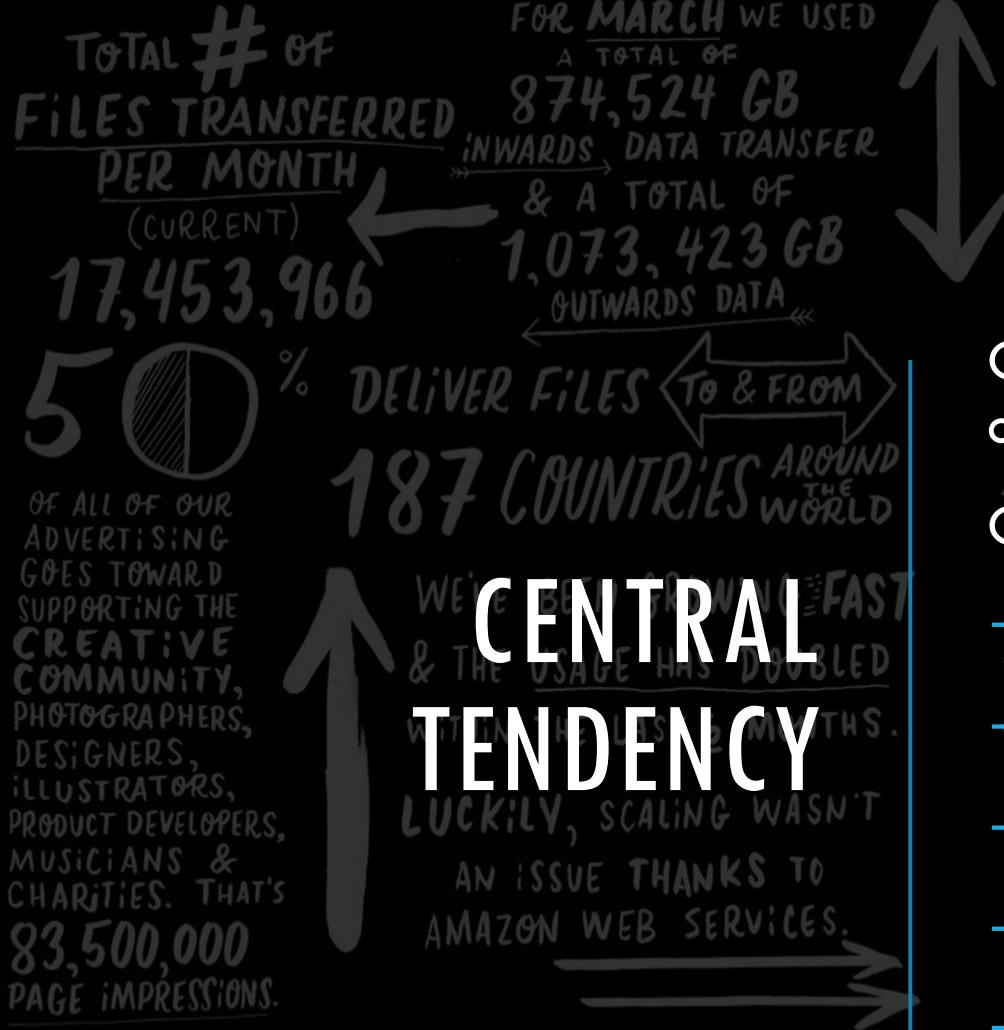
OF FACEBOOK FANS

110,000

This section of the infographic displays visitor and fan statistics. It shows the total number of unique visitors since the start (117,159,254), more than 15,000 channels sold, and the total number of Facebook fans (23,000,000). It also includes a pie chart showing the distribution of users by age group (18-24, 25-34, 35-44, 45-54, 55-64, 65+), gender (52% male, 48% female), and business type (24% business, 51% personal, 25% both).

30 NEW USERS FIND OUR SITE EVERY MINUTE

8141 TERABYTES OF DATA SENT VIA OUR SERVERS



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

Central tendency performs the following measures:

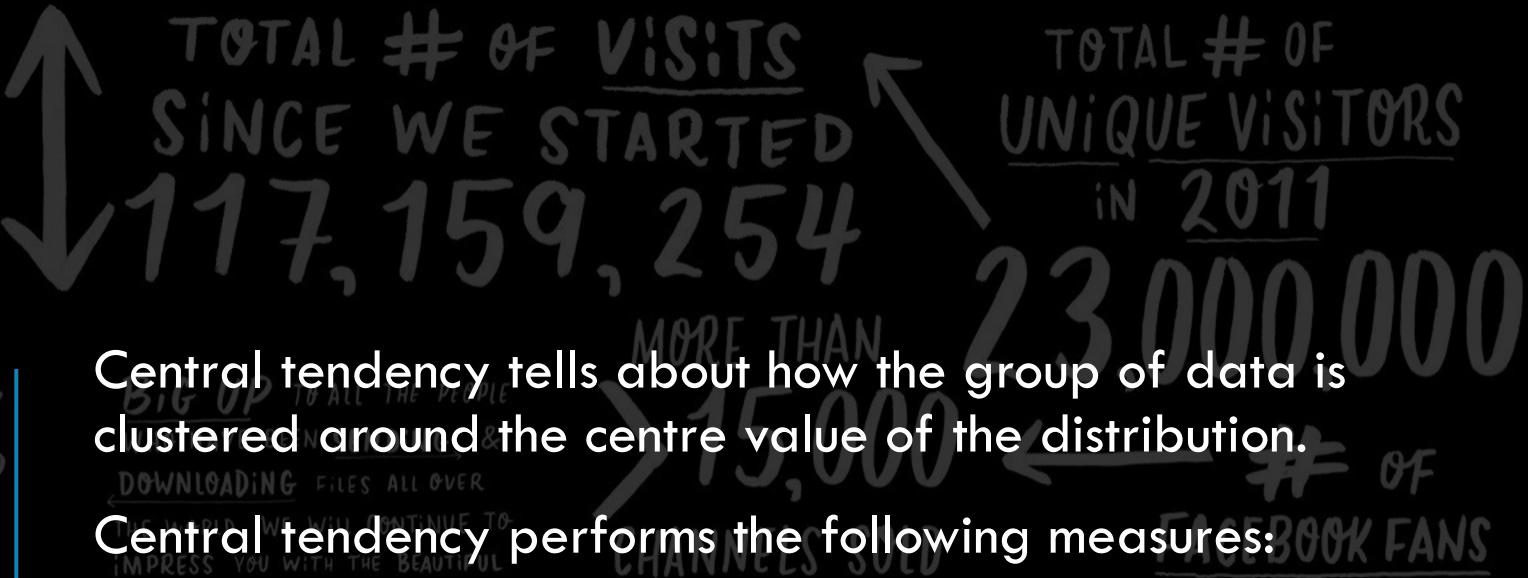
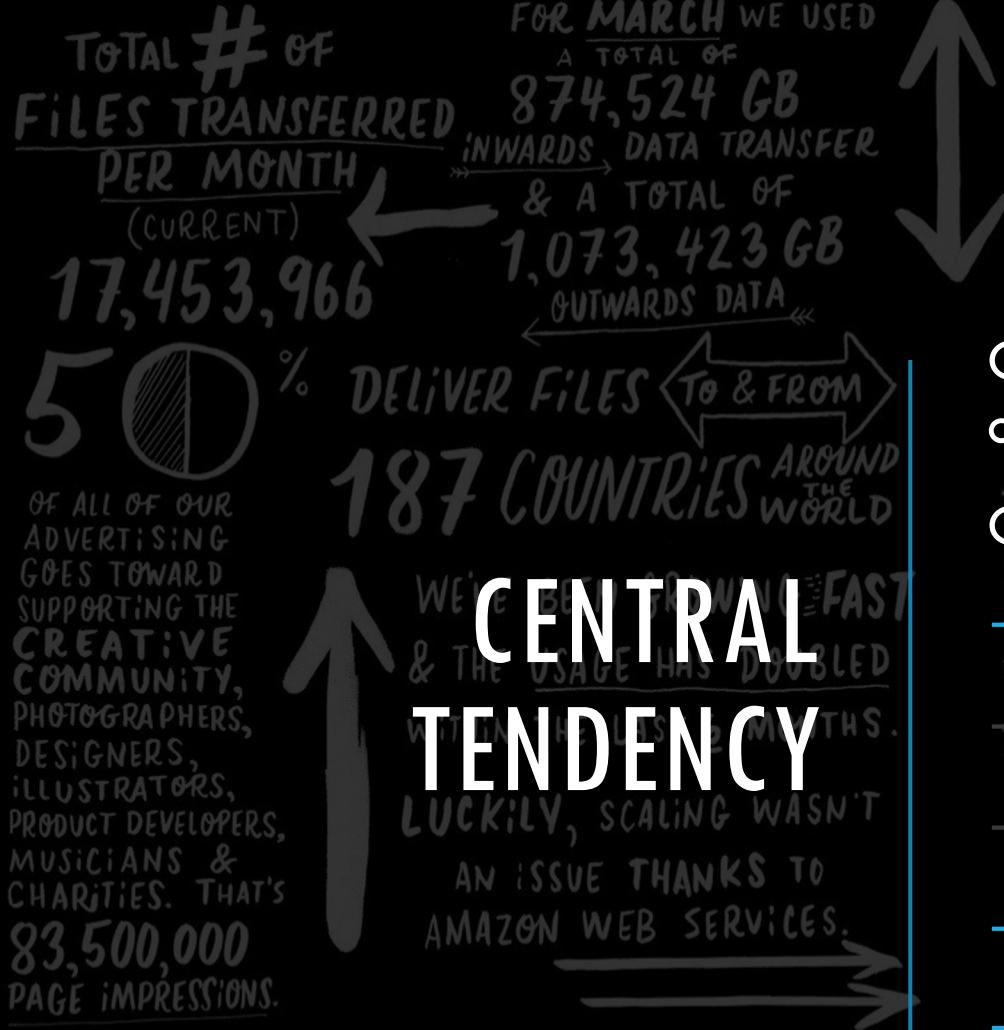
→ Arithmetic Mean

→ Geometric Mean

→ Harmonic Mean

→ Median

→ Mode



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

Central tendency performs the following measures:

→ Arithmetic Mean

DONE VIA THE NEW WETRANSFER.

→ Geometric Mean

YES, IT'S LOOKING & FEELING YOU'LL LOVE IT...

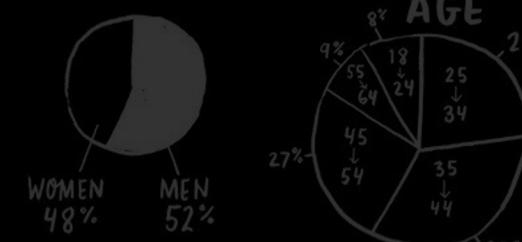
→ Harmonic Mean

We transfer

→ Median

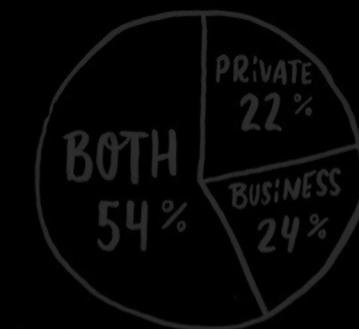
We transfer

→ Mode



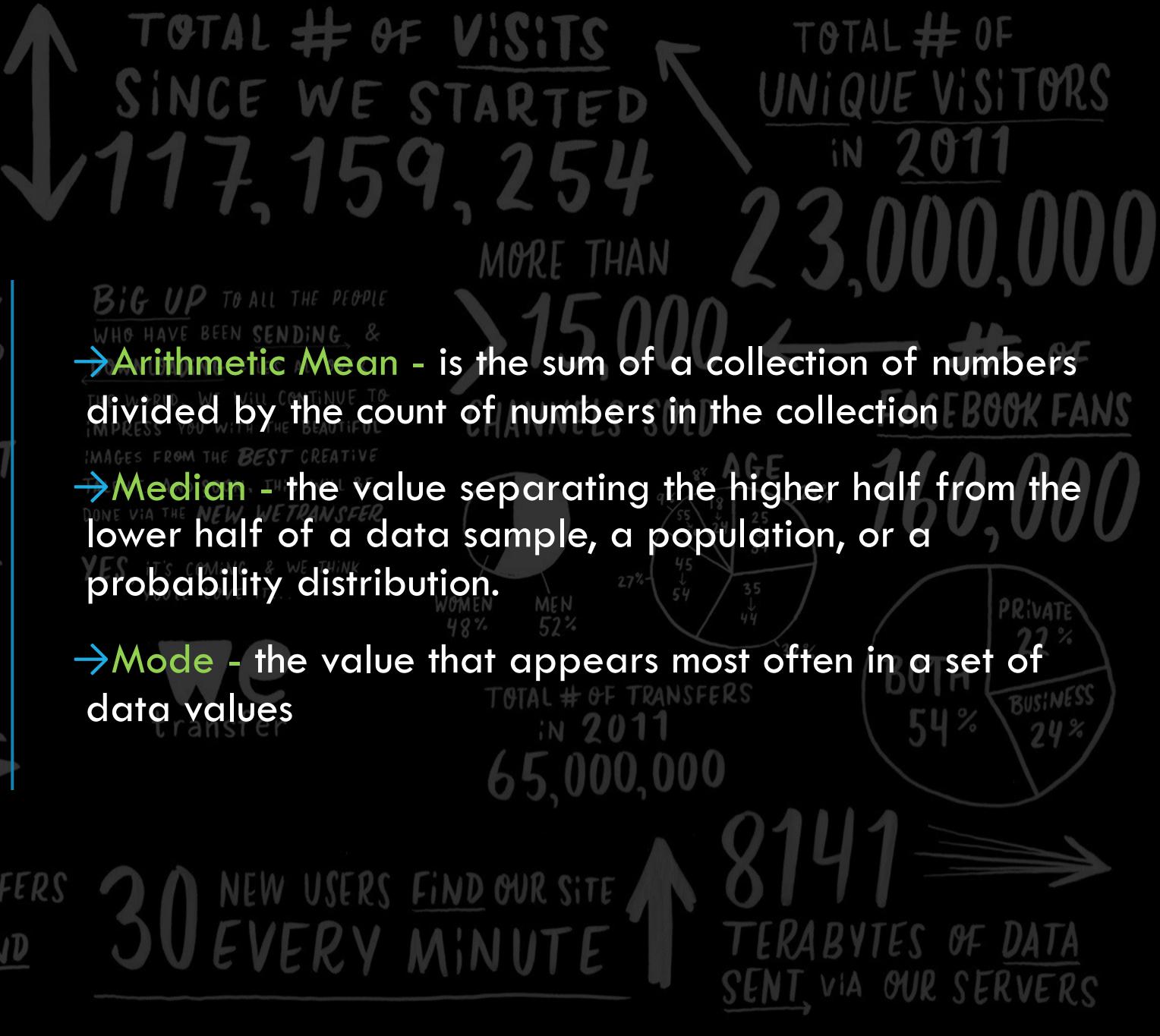
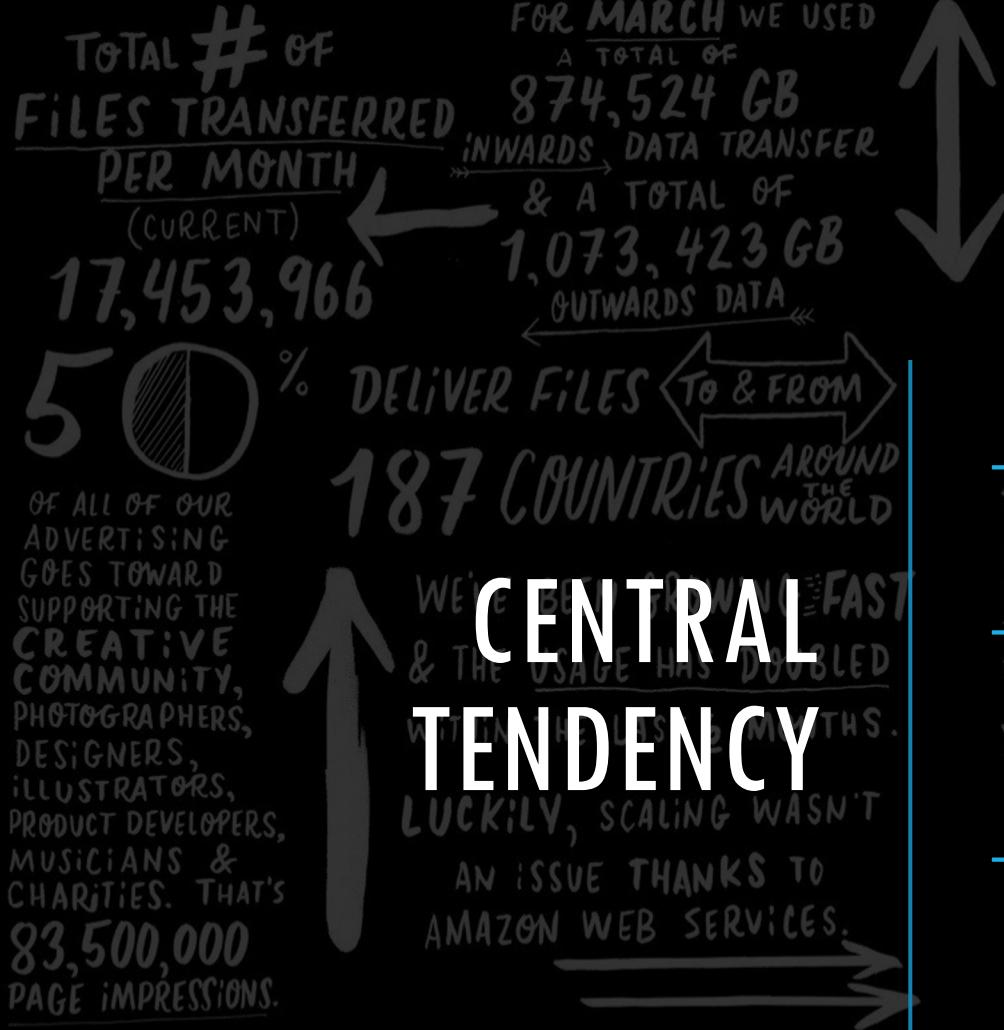
TOTAL # OF TRANSFERS IN 2011

65,000,000



30 NEW USERS FIND OUR SITE EVERY MINUTE

8141 TERABYTES OF DATA SENT VIA OUR SERVERS



TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT)

17,453,966

50% DELIVER FILES **TO & FROM 187 COUNTRIES AROUND THE WORLD**

FOR MARCH WE USED A TOTAL OF 874,524 GB INWARDS DATA TRANSFER & A TOTAL OF 1,073,423 GB OUTWARDS DATA

WE'RE BEING DOWNRIGHT FAST & THE USAGE HAS DOUBLED

WITHIN THE LAST 2 MONTHS.

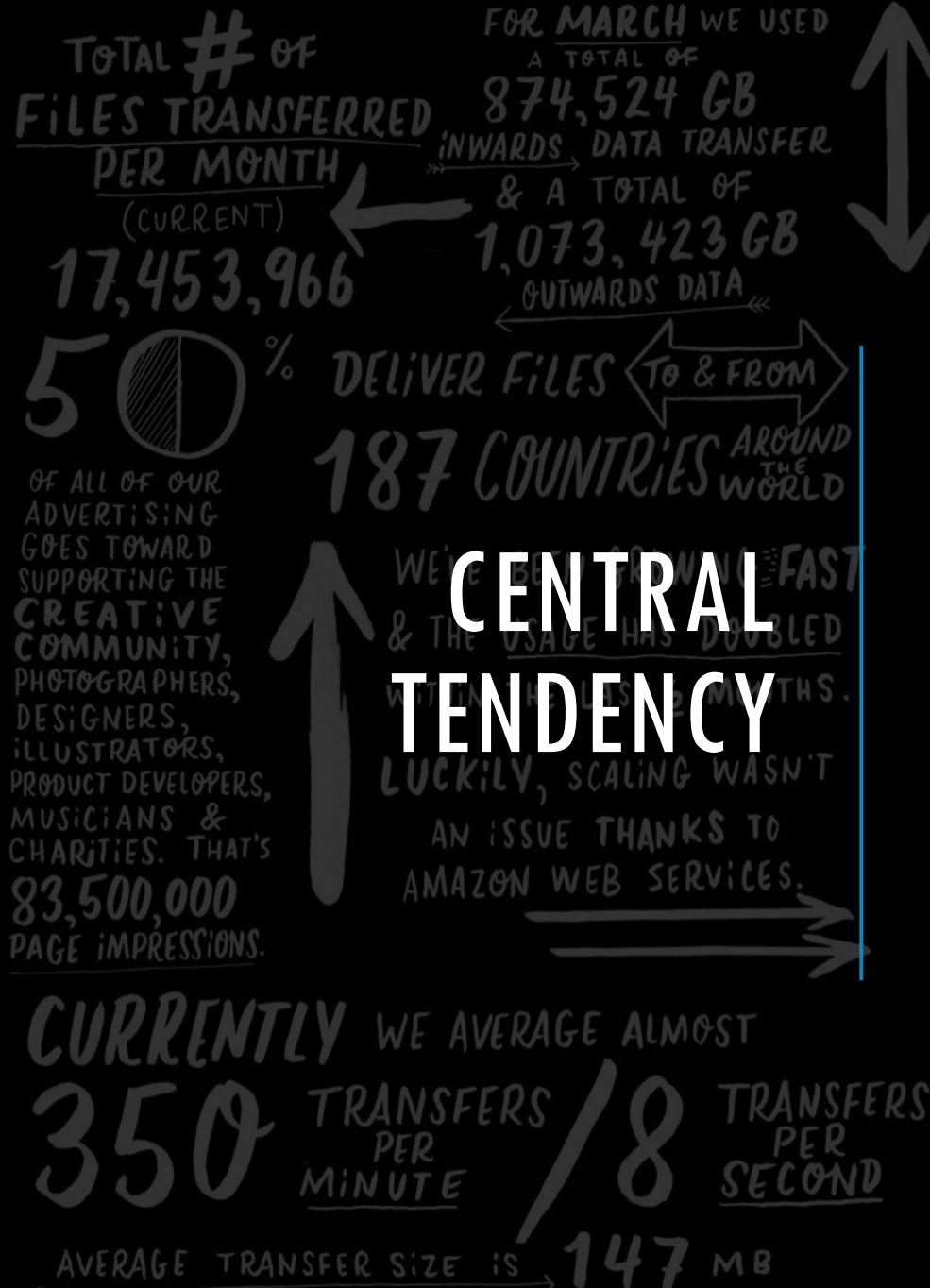
LUCKILY, SCALING WASN'T AN ISSUE THANKS TO AMAZON WEB SERVICES.

CENTRAL TENDENCY

↑

OF ALL OF OUR ADVERTISING GOES TOWARD SUPPORTING THE CREATIVE COMMUNITY, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S **83,500,000 PAGE IMPRESSIONS.**

CURRENTLY WE AVERAGE ALMOST
350 TRANSFERS PER MINUTE / 18 TRANSFERS PER SECOND
AVERAGE TRANSFER SIZE IS → 147 MB



CENTRAL TENDENCY

The infographic illustrates two types of skewed distributions: Positive Skew and Negative Skew.

Positive Skew: This section shows a bell-shaped curve shifted to the right, indicating a long tail of outliers. It features a pie chart showing gender distribution (Men 52%, Women 48%) and a bar chart showing age groups (18-24, 25-34, 35-44, 45-54, 55-64, 65+). The mean is the far-right point of the distribution, while the median is closer to the center, and the mode is the peak on the left.

Negative Skew: This section shows a bell-shaped curve shifted to the left, indicating a long tail of outliers. It features a pie chart showing transfer purposes (Business 24%, Private 22%, Both 54%) and a bar chart showing total transfers by month (January 10M, February 12M, March 15M, April 18M, May 20M, June 22M, July 25M, August 28M, September 30M, October 35M, November 40M, December 45M). The mean is the far-left point of the distribution, while the median is closer to the center, and the mode is the peak on the right.

TOTAL # OF

FILES TRANSFERRED
PER MONTH
(CURRENT)

17,453,966



%

DELIVER FILES  

187 COUNTRIES AROUND THE WORLD

OF ALL OF OUR ADVERTISING GOES TOWARD SUPPORTING THE CREATIVE COMMUNITY, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S 83,500,000 PAGE IMPRESSIONS.

VARIABILITY IN THE DISTRIBUTION OF DATA

CURRENTLY WE AVERAGE ALMOST

350 TRANSFERS PER MINUTE

18 TRANSFERS PER SECOND

AVERAGE TRANSFER SIZE IS

147 MB

FOR MARCH WE USED

A TOTAL OF

874,524 GB

INWARDS DATA TRANSFER

& A TOTAL OF

1,073,423 GB

OUTWARDS DATA

TOTAL # OF VISITS

SINCE WE STARTED

1,171,159,1254

TOTAL # OF

UNIQUE VISITORS

1,3811

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.

YES, IT'S COMING & WE THINK YOU'LL LOVE IT...

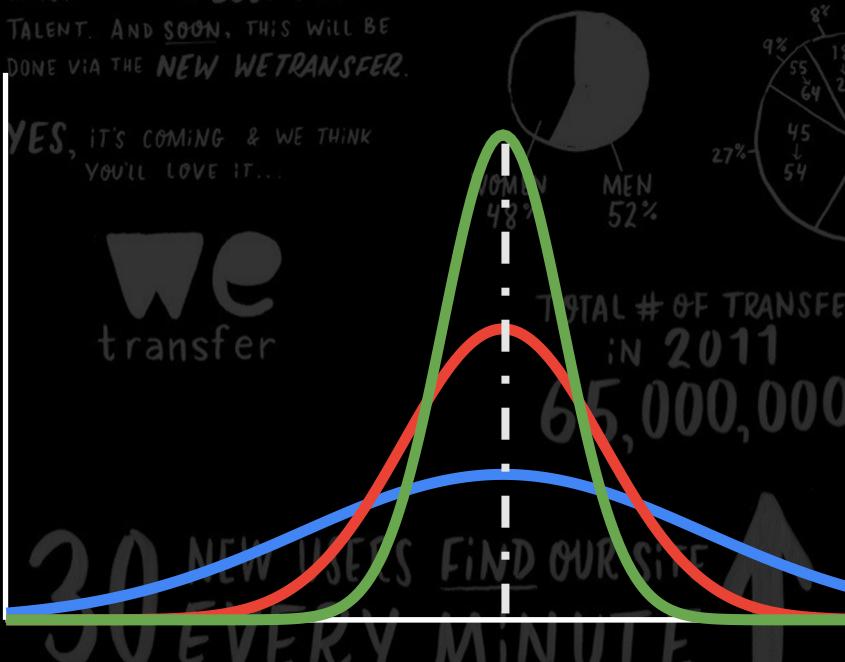


CHANNELS SOLD



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.



TOTAL # OF

FILES TRANSFERRED
PER MONTH
(CURRENT)

17,453,966



OF ALL OF OUR ADVERTISING GOES TOWARD SUPPORTING THE CREATIVE COMMUNITY, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S 83,500,000 PAGE IMPRESSIONS.

VARIABILITY IN THE DISTRIBUTION OF DATA

CURRENTLY WE AVERAGE ALMOST

350 TRANSFERS PER MINUTE

18 TRANSFERS PER SECOND

AVERAGE TRANSFER SIZE IS

147 MB

FOR MARCH WE USED

A TOTAL OF

874,524 GB

INWARDS DATA TRANSFER

& A TOTAL OF

1,073,423 GB

OUTWARDS DATA

TOTAL # OF VISITS

SINCE WE STARTED

117,159,254

TOTAL # OF

UNIQUE VISITORS
IN 2011

22,000,000

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.



High variability

Medium variability

Low variability

CHANNELS SOLD



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.



TOTAL # OF TRANSFERS IN 2011

65,000,000

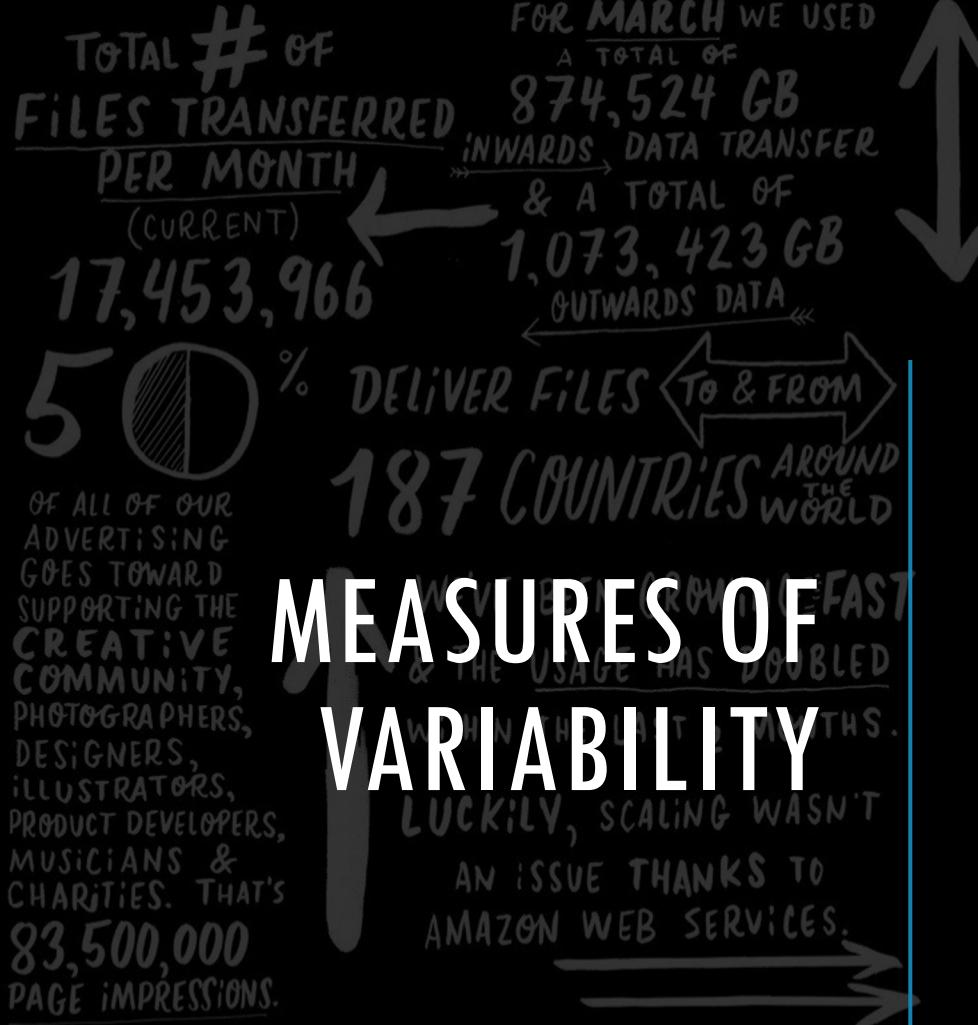
20 NEW USERS FIND OUR SITE

EVERY MINUTE

8141

TERABYTES OF DATA

SENT VIA OUR SERVERS



CURRENTLY WE AVERAGE ALMOST 350 TRANSFERS PER MINUTE / 18 TRANSFERS PER SECOND

AVERAGE TRANSFER SIZE IS 147 MB

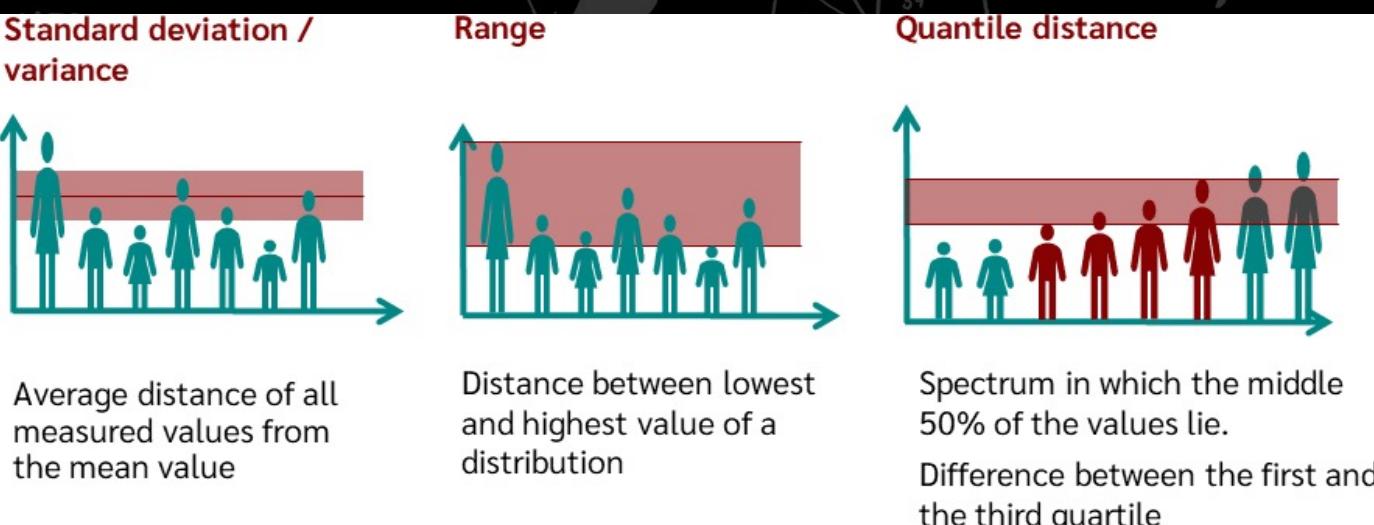
TOTAL # OF VISITS SINCE WE STARTED

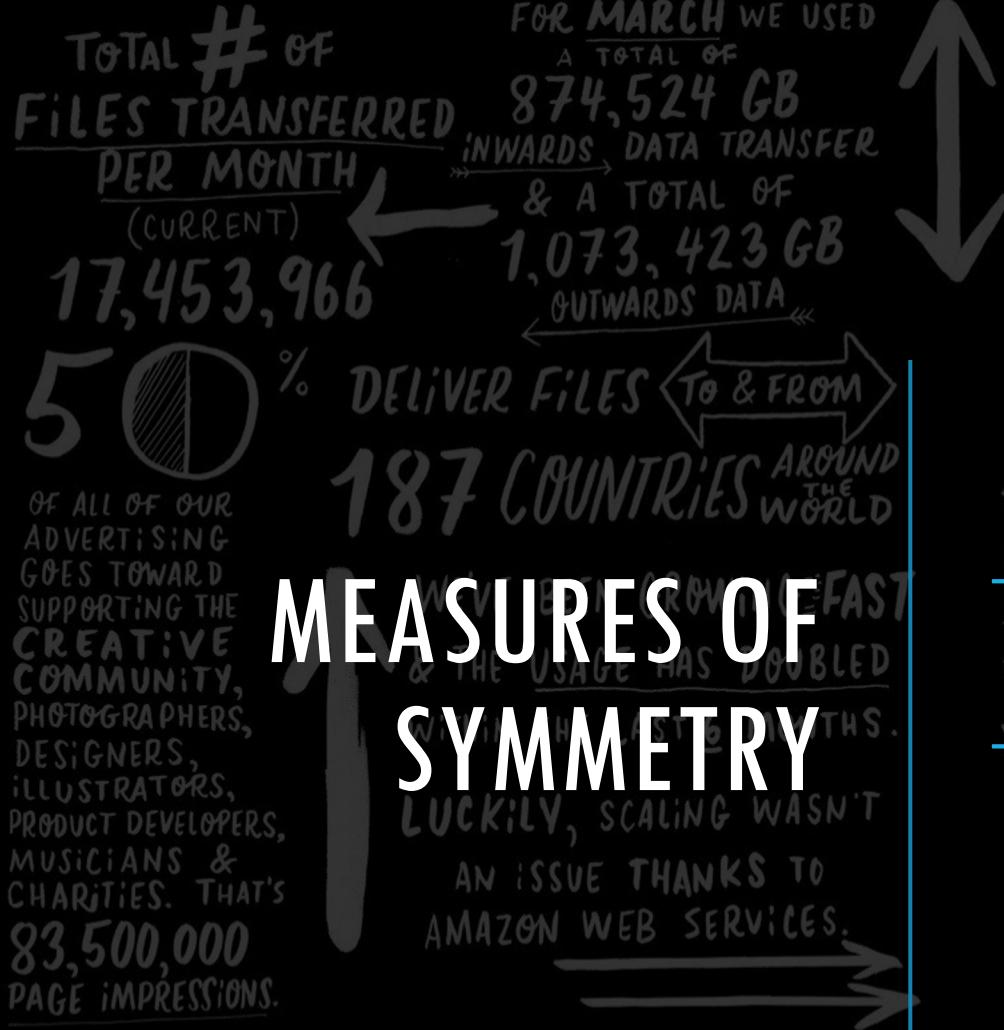
→ Range - the difference between the highest and lowest values

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

→ Standard deviation - average distance from the mean

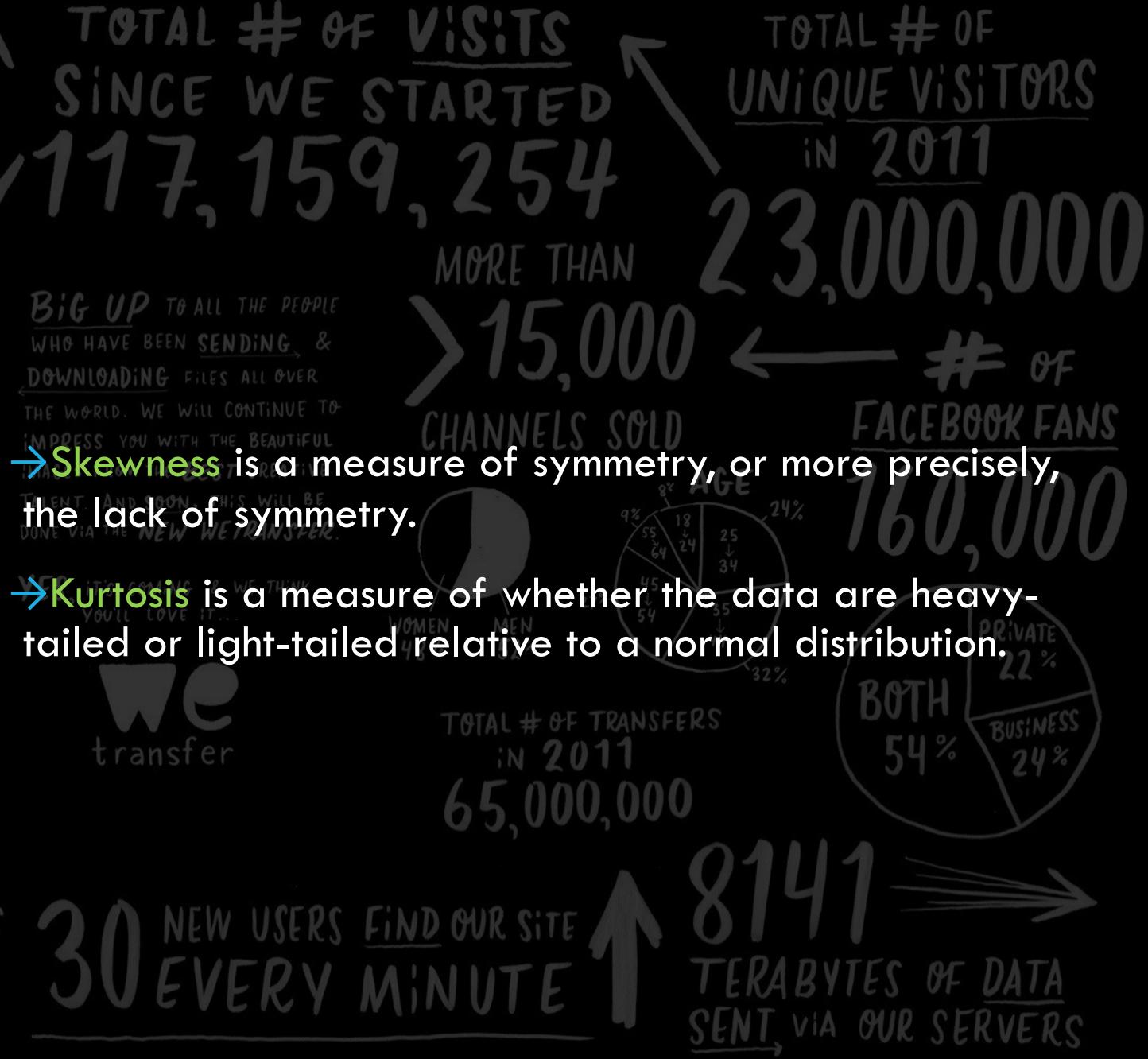
→ Variance - average of squared distances from the mean

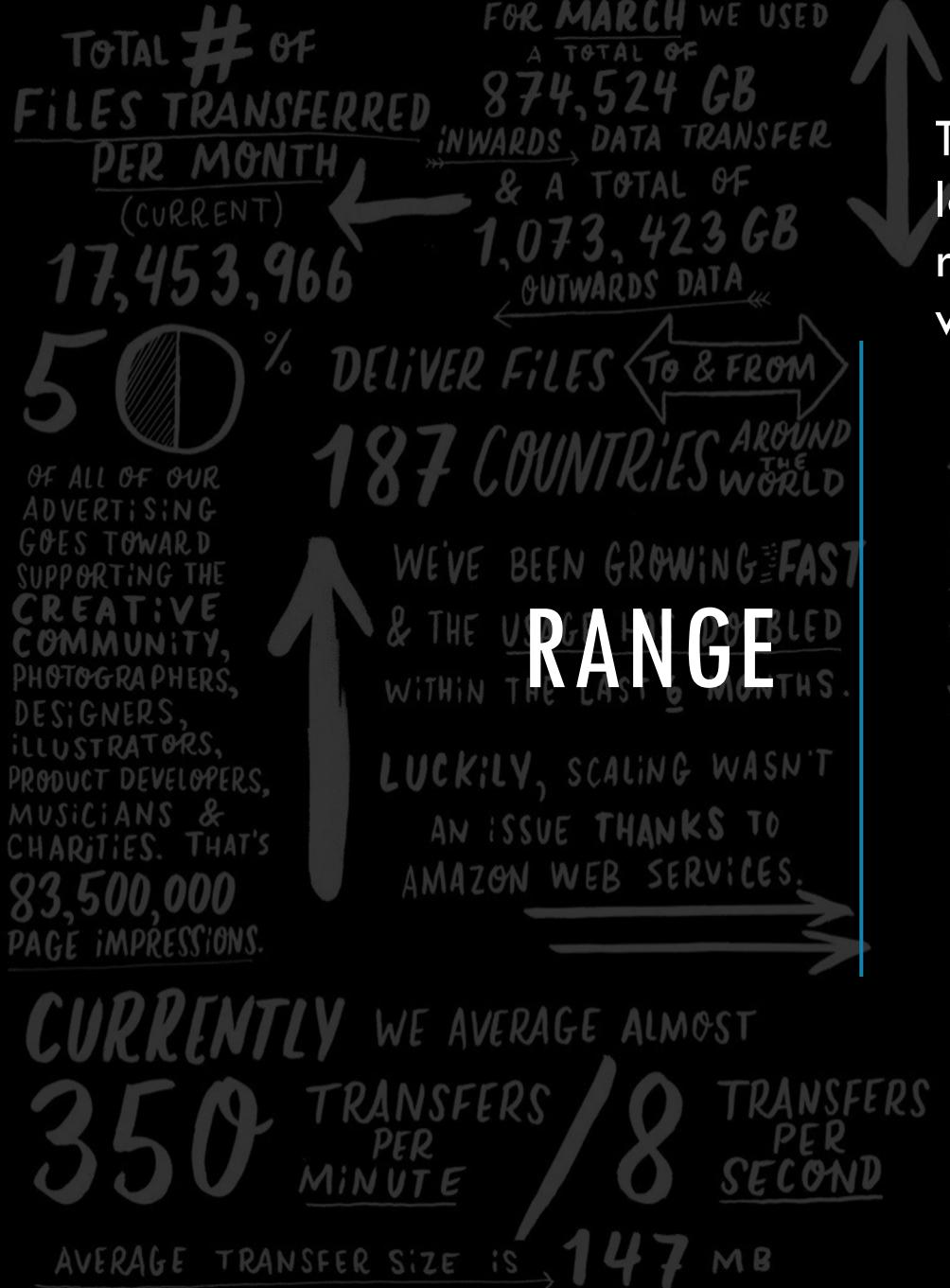




CURRENTLY WE AVERAGE ALMOST **350 TRANSFERS PER MINUTE** / **18 TRANSFERS PER SECOND**

AVERAGE TRANSFER SIZE IS → **147 MB**





TOTAL # OF VISITS SINCE WE STARTED

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.

BIG UP TO ALL THE PEOPLE WHO HAVE BEEN SENDING & DOWNLOADING FILES ALL OVER THE WORLD. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVE TALENT. AND SOON, THIS WILL BE DONE VIA THE NEW WETRANSFER.

YES, IT'S COMING & WE THINK YOU'LL LOVE IT...

Example

You have 8 data points from Sample A.

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

Women: 48% Men: 52%

R = H - L

R = 324 - 72 = 252

The range of your data is 252

INTERQUARTILE RANGE

TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT)

17,453,966

50% DELIVER FILES **TO & FROM 187 COUNTRIES AROUND THE WORLD**

OF ALL OF OUR ADVERTISING GOES TOWARD SUPPORTING THE **CREATIVE COMMUNITY**, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S **83,500,000 PAGE IMPRESSIONS.**

FOR MARCH WE USED A TOTAL OF **874,524 GB INWARDS DATA TRANSFER** & A TOTAL OF **1,073,423 GB OUTWARDS DATA**

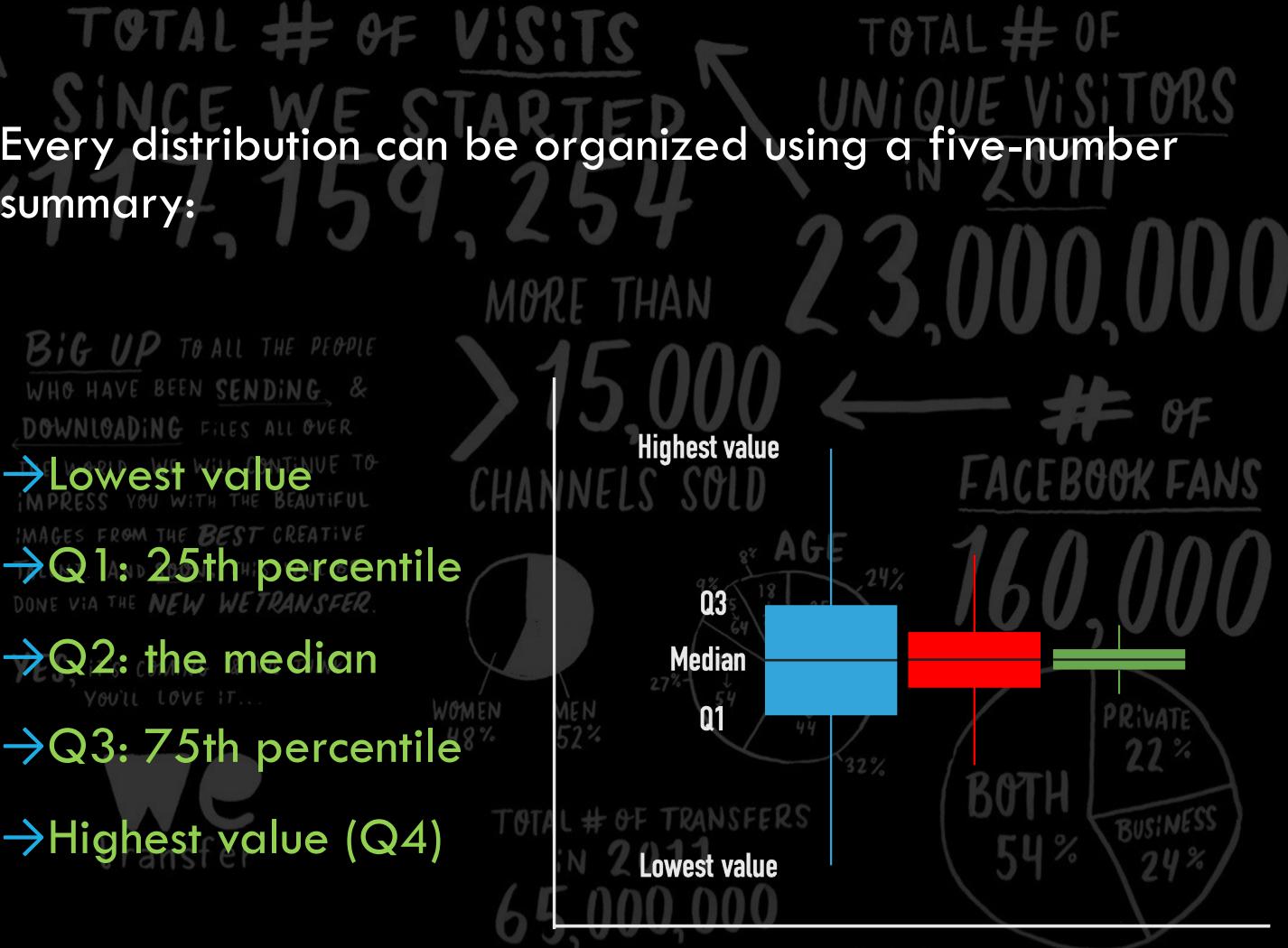
INTERQUARTILE RANGE

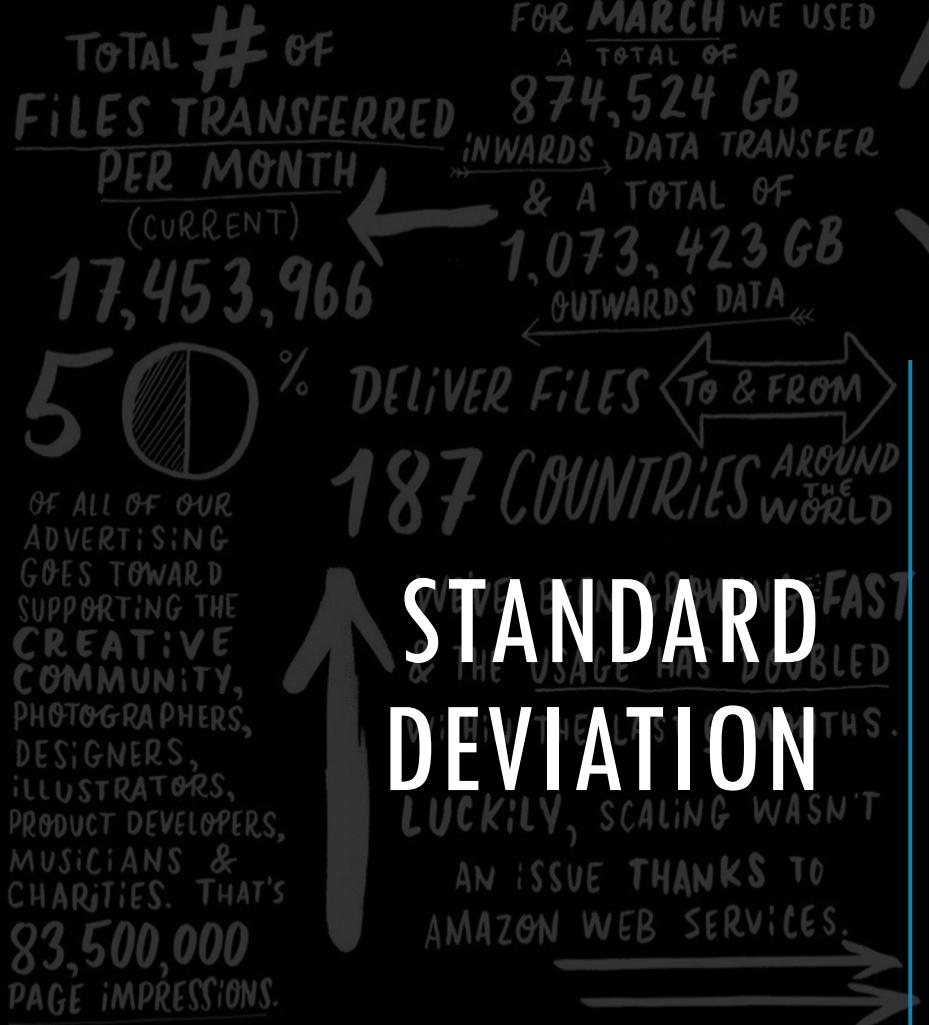
WITHIN THE PAST 2 MONTHS. & THE USAGE HAS DOUBLED LUCKILY, SCALING WASN'T AN ISSUE THANKS TO AMAZON WEB SERVICES.

**CURRENTLY WE AVERAGE ALMOST
350 TRANSFERS PER MINUTE / 18 TRANSFERS PER SECOND
AVERAGE TRANSFER SIZE IS 147 MB**

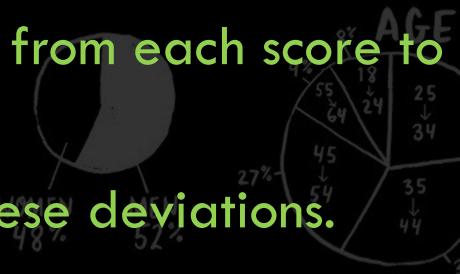
The figure consists of several data visualizations from WeTransfer's website, overlaid on a central normal distribution curve. The x-axis of the distribution is labeled with standard deviations from -4σ to 4σ , with the mean at 0σ . The area under the curve is divided into three main sections: 24.65% on the left, 50% in the center, and 24.65% on the right.

- Top Left:** A box plot showing the distribution of transfers. The y-axis ranges from Q1 - 1.5 × IQR to Q3 + 1.5 × IQR. The box spans from Q1 to Q3, with a median line at the center. Whiskers extend to the minimum (27%) and maximum (52%) values. Labels include "WOMEN 48%" and "MEN 52%".
- Top Right:** A pie chart titled "AGE" showing the distribution of users by age group: 18-24 (24%), 25-34 (25%), 35-44 (32%), 45-54 (18%), and 55-64 (8%).
- Bottom Right:** A pie chart titled "BOTH" showing the distribution between "PRIVATE" (22%) and "BUSINESS" (24%) users.
- Bottom Left:** A box plot for "CHANNELS SOLD" with a y-axis from Q1 - 1.5 × IQR to Q3 + 1.5 × IQR. The box spans from Q1 to Q3, with a median line at the center. Whiskers extend to the minimum (18%) and maximum (55%) values. Labels include "18%", "24%", "25%", "34%", "45%", "54%", and "55-64".
- Background Text:** Various statistics from WeTransfer's website are visible as text overlays:
 - "TOTAL # OF VISITS SINCE WE STARTED" (with a large number 254)
 - "TOTAL # OF UNIQUE VISITORS IN 2011" (with a large number 23,000,000)
 - "# OF DOWNLOADING FILES ALL OVER THE WORLD. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVE TALENT. AND SOON, THIS WILL BE DONE VIA THE NEW WETRANSFER"
 - "YES, IT'S COMING & WE THINK YOU'LL LOVE IT..."
 - "WE transfer"
 - "TOTAL # OF TRANSFERS 2011" (with a large number 65,000,000)
 - "30 NEW USERS FIND EVERY MINUTE" (with a large number 8141)
 - "8141 BYTES OF DATA SENT VIA OUR SERVERS"





STANDARD DEVIATION

- TOTAL # OF VISITS SINCE WE STARTED**
- The standard deviation is the average amount of variability in your dataset.
- TOTAL # OF UNIQUE VISITORS IN 2011**
- 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.
- # OF FACEBOOK FANS**
- 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.
- 
- | Age Group | Percentage |
|-----------|------------|
| 18 | 18% |
| 24 | 24% |
| 34 | 25% |
| 44 | 27% |
| 54 | 52% |
- 
- | Purpose | Percentage |
|----------|------------|
| PRIVATE | 22% |
| BOTH | 54% |
| BUSINESS | 24% |

TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT) **17,453,966**

FOR MARCH WE USED A TOTAL OF **874,524 GB** INWARDS DATA TRANSFER & A TOTAL OF **1,073,423 GB** OUTWARDS DATA

50% DELIVER FILES 

187 COUNTRIES 

OF ALL OF OUR ADVERTISING GOES TOWARD SUPPORTING THE CREATIVE COMMUNITY, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S **83,500,000** PAGE IMPRESSIONS.

STANDARD DEVIATION

WEVE BEEN GROWING FAST & THE USAGE HAS DOUBLED WITHIN THE LAST 6 MONTHS. LUCKILY, SCALING WASN'T AN ISSUE THANKS TO AMAZON WEB SERVICES.

CURRENTLY WE AVERAGE ALMOST
350 TRANSFERS PER MINUTE / 8 TRANSFERS PER SECOND
AVERAGE TRANSFER SIZE IS 147 MB

**TOTAL # OF VISITS
SINCE WE STARTED
117,159,254**

Standard deviation formula for populations

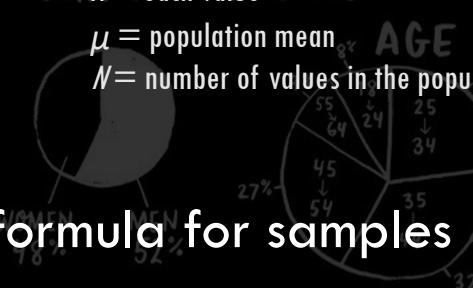
$$\sigma = \sqrt{\frac{1}{N} \sum (X - \mu)^2}$$

\sum = sum of...
 χ = each value

μ = population mean

N= number of v

55 25



Standard deviation formula for sample

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

Σ = sum of...

x = each value

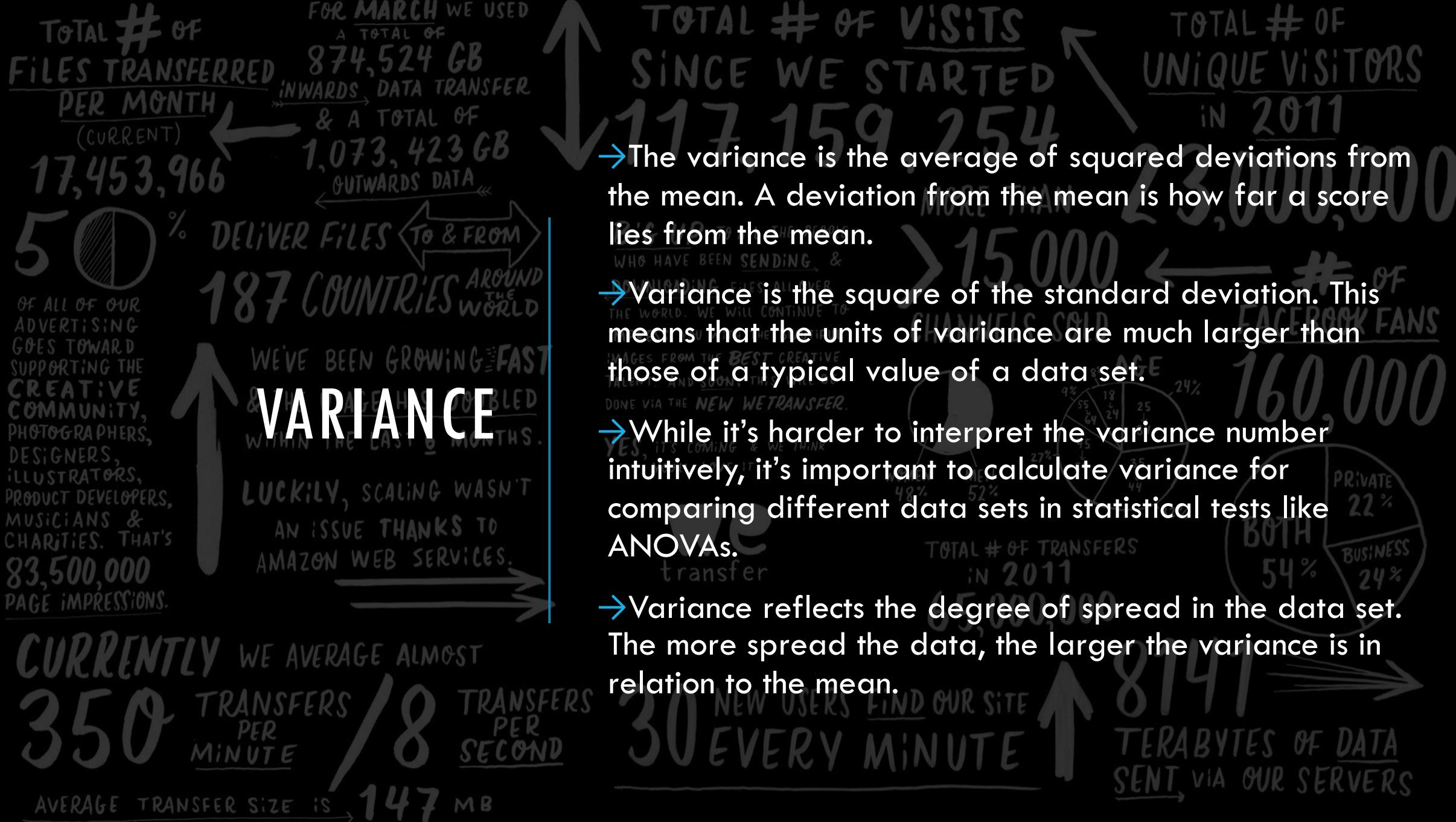
X cash value

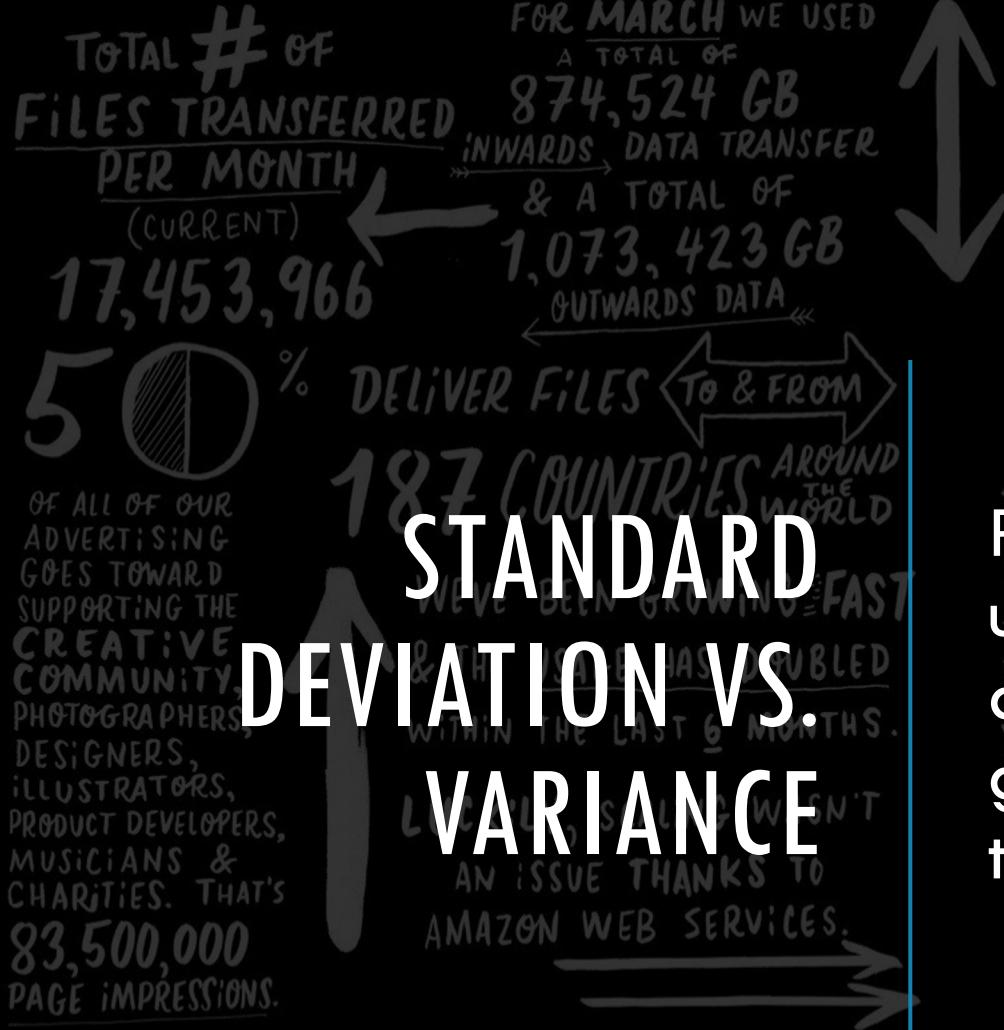
\bar{x} = sample mean
 N = number of values in the sample

**30 NEW USERS FIND OUR SITE
EVERY MINUTE**

in the sample
8141 →
TERABYTES OF DATA
SENT VIA OUR SERVERS

**TOTAL # OF
UNIQUE VISITORS
IN 2011**
23,000,000
lations ← **# OF
FACEBOOK FANS**
160,000





STANDARD DEVIATION VS. VARIANCE

TOTAL # OF VISITS SINCE WE STARTED

117,159,254

MORE THAN >15,000

TOTAL # OF UNIQUE VISITORS IN 2011

23,000,000

OF FACEBOOK FANS

160,000

TOTAL # OF TRANSFERS IN 2011

65,000,000

30 NEW USERS FIND OUR SITE EVERY MINUTE

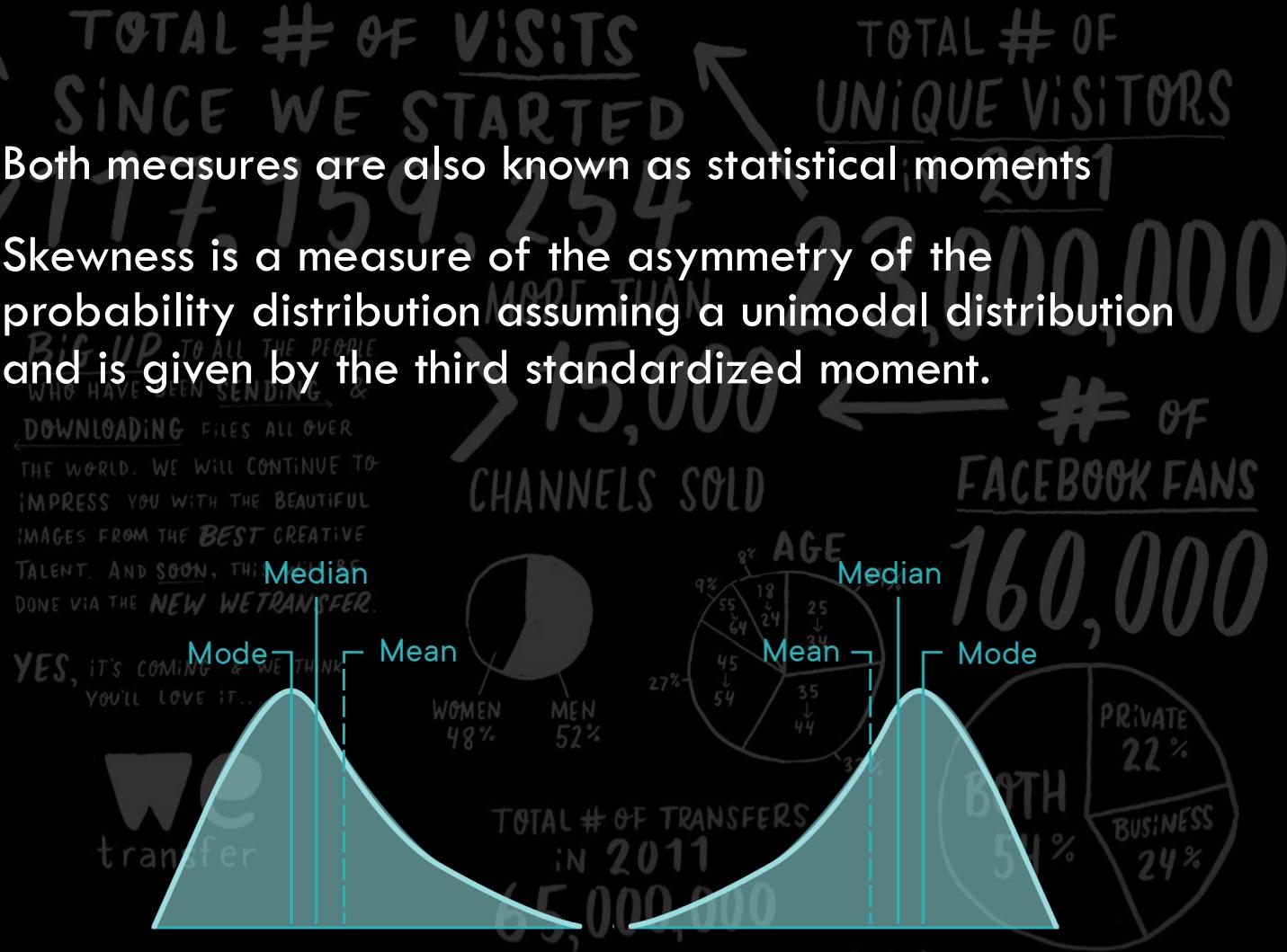
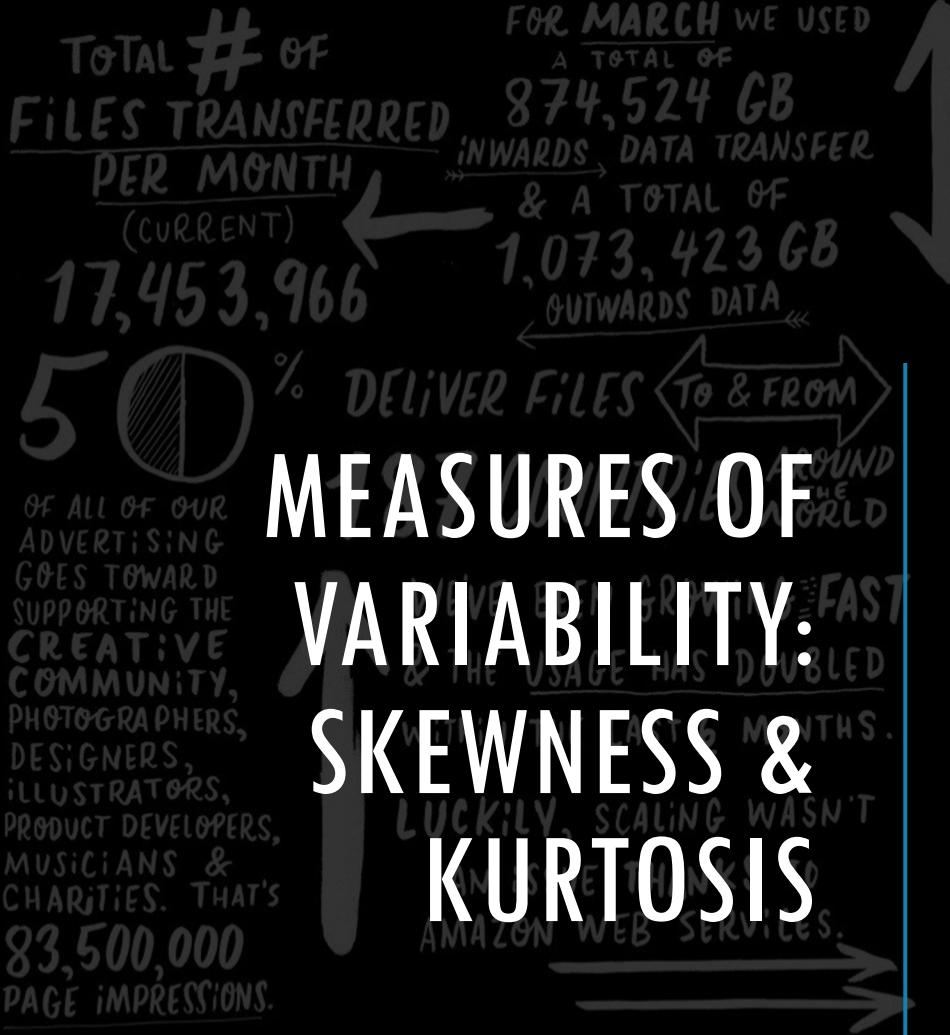
8141 TERABYTES OF DATA SENT VIA OUR SERVERS

BIG UP TO ALL THE PEOPLE WHO HAVE BEEN SENDING & DOWNLOADING FILES ALL OVER THE WORLD. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVE INDUSTRY. YES, IT'S COMING & WE THINK DONE VIA THE NEW WETRANSFER.

WE TRANSFER

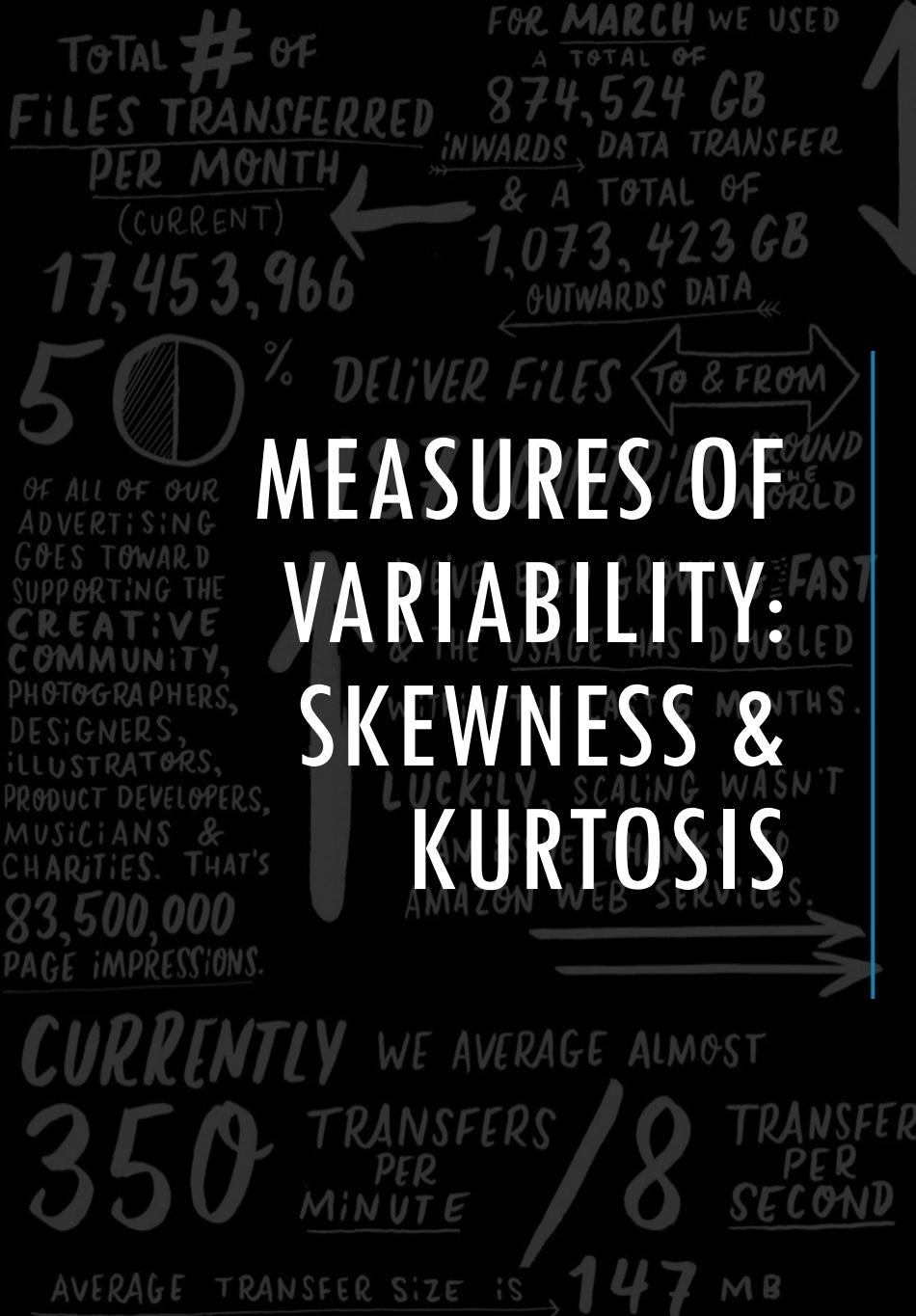
Practically, Standard Deviation should be used to check the variations in the data as a descriptive analysis tool whereas Variance is generally used as an inferential analysis tool to compare the two means (e.g., ANOVA)

MEASURES OF VARIABILITY: SKEWNESS & KURTOSIS



30 NEW USERS FIND OUR SITE EVERY MINUTE

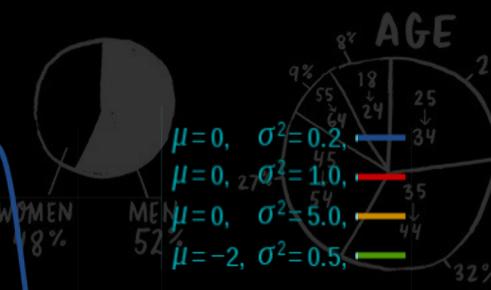
MEASURES OF VARIABILITY: SKEWNESS & KURTOSIS



TOTAL # OF VISITS SINCE WE STARTED

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 peak and shorter thinner tails.



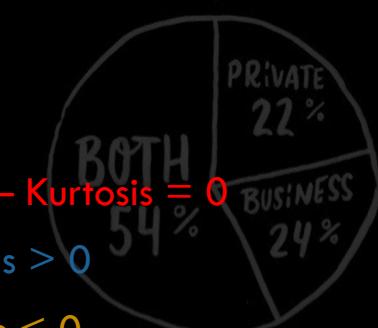
Normal distribution – Kurtosis = 0

Leptokurtic – Kurtosis > 0

Platykurtic – Kurtosis < 0

TOTAL # OF UNIQUE VISITORS IN 2011

160,000



30 NEW USERS FIND OUR SITE
EVERY MINUTE

8141

TERABYTES OF DATA SENT VIA OUR SERVERS

UNDERSTANDING CONFIDENCE INTERVALS

TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT)

17,453,966

50 % DELIVER FILES TO & FROM

187 COUNTRIES AROUND THE WORLD

FOR MARCH WE USED A TOTAL OF 874,524 GB INWARDS DATA TRANSFER & A TOTAL OF 1,073,423 GB OUTWARDS DATA

OF ALL OF OUR ADVERTISING GOES TOOUR SUPPORTING THE CREATIVE COMMUNITY, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S 83,500,000 PAGE IMPRESSIONS.

WE'VE BEEN GROWING FAST WITHIN THE LAST 9 MONTHS. DOUBLE, DOUBLE, DOUBLE WASN'T AN ISSUE THANKS TO AMAZON WEB SERVICES.

UNDERSTANDING CONFIDENCE INTERVALS

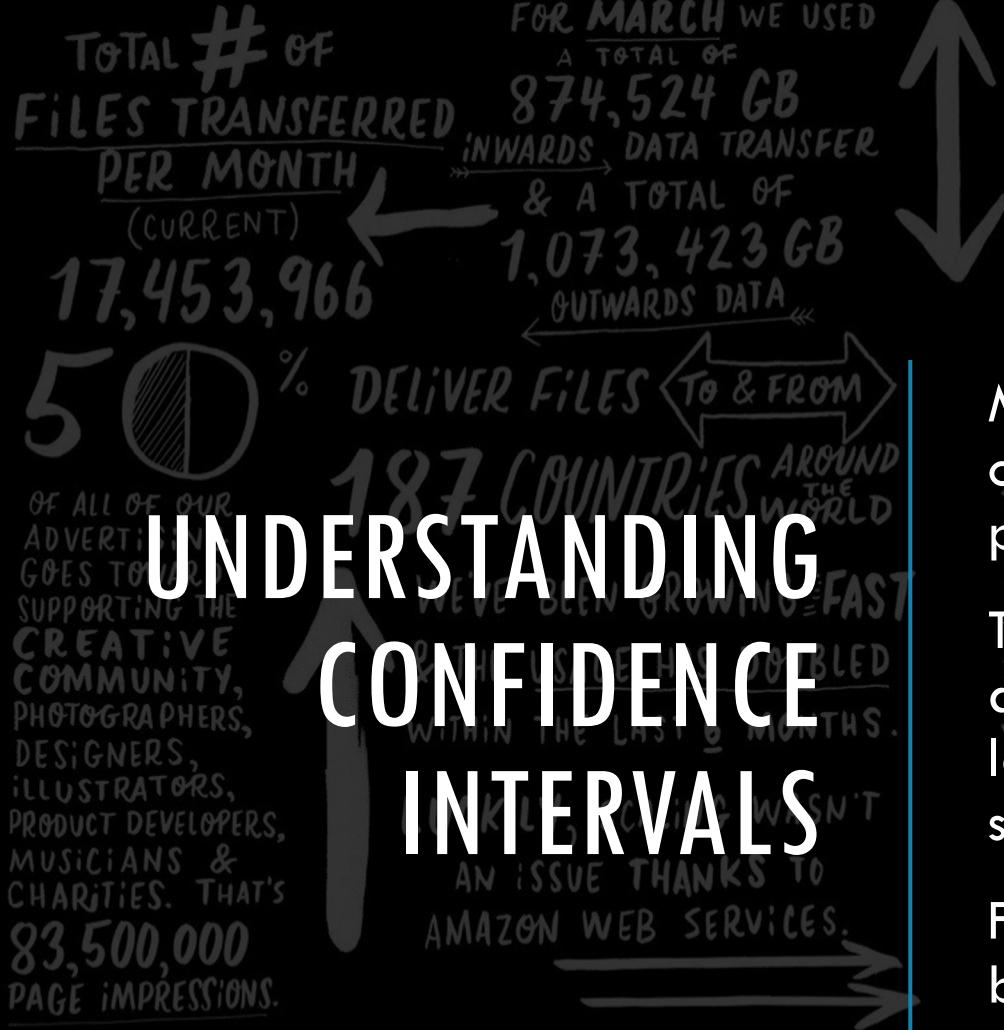
CURRENTLY WE AVERAGE ALMOST
350 TRANSFERS / 8 TRANSFER
PER MINUTE PER SECOND
AVERAGE TRANSFER SIZE IS 147 MB

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

TOTAL # OF VISITS SINCE WE STARTED

117,159,254

MORE THAN 1500 CHANNELS SOLD

BIG UP TO ALL THE PEOPLE WHO HAVE BEEN SENDING & RECEIVING. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVE STUDIO IN THE BUSINESS. WE ARE PROUDLY DONE VIA THE NEW WETRANSFER. YES IT'S COMING & WE THINK IT'S GOING TO BE GREAT.

AGE

WOMEN 48% MEN 52%

65,000,000 TOTAL # OF TRANSFERS IN 2011

160,000 FACEBOOK FANS

PRIVATE 22% BOTH 54% BUSINESS 24%

8141 TERABYTES OF DATA SENT VIA OUR SERVERS

30 NEW USERS FIND OUR SITE EVERY MINUTE

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.

This section features a large central text area containing several statistics: 'TOTAL # OF VISITS SINCE WE STARTED 117,159,254', 'MORE THAN 1500 CHANNELS SOLD', 'BIG UP TO ALL THE PEOPLE WHO HAVE BEEN SENDING & RECEIVING. WE WILL CONTINUE TO IMPRESS YOU WITH THE BEAUTIFUL IMAGES FROM THE BEST CREATIVE STUDIO IN THE BUSINESS. WE ARE PROUDLY DONE VIA THE NEW WETRANSFER. YES IT'S COMING & WE THINK IT'S GOING TO BE GREAT.', 'AGE WOMEN 48% MEN 52%', '65,000,000 TOTAL # OF TRANSFERS IN 2011', 'FACEBOOK FANS 160,000', 'PRIVATE 22% BOTH 54% BUSINESS 24%', '8141 TERABYTES OF DATA SENT VIA OUR SERVERS', and '30 NEW USERS FIND OUR SITE EVERY MINUTE'. To the left of the text, there's a large 'UNDERSTANDING CONFIDENCE INTERVALS' heading. Arrows point from the text to various parts of the infographic.

UNDERSTANDING CONFIDENCE INTERVALS

TOTAL # OF FILES TRANSFERRED PER MONTH (CURRENT)

17,453,966

50% DELIVER FILES **TO & FROM**

187 COUNTRIES AROUND THE WORLD

WE'VE BEEN GROWING FAST

RATHER THAN DEATH, DOUBLED

WITHIN THE LAST 6 MONTHS.

LIKELY, CLOUD WASN'T AN ISSUE THANKS TO AMAZON WEB SERVICES.

FOR MARCH WE USED A TOTAL OF 874,524 GB INWARDS DATA TRANSFER & A TOTAL OF 1,073,423 GB OUTWARDS DATA

OF ALL OF OUR ADVERTISING GOES TO SUPPORTING THE CREATIVE COMMUNITY, PHOTOGRAPHERS, DESIGNERS, ILLUSTRATORS, PRODUCT DEVELOPERS, MUSICIANS & CHARITIES. THAT'S 83,500,000 PAGE IMPRESSIONS.

UNDERSTANDING CONFIDENCE INTERVALS

**CURRENTLY WE AVERAGE ALMOST
350 TRANSFERS PER MINUTE / 18 TRANSFERS PER SECOND
AVERAGE TRANSFER SIZE IS 147 MB**

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.

Original Population With Sample Size N

B set Bootstrap Sample Size of N

B set Bootstrap Estimate Population Parameters

Further Inference