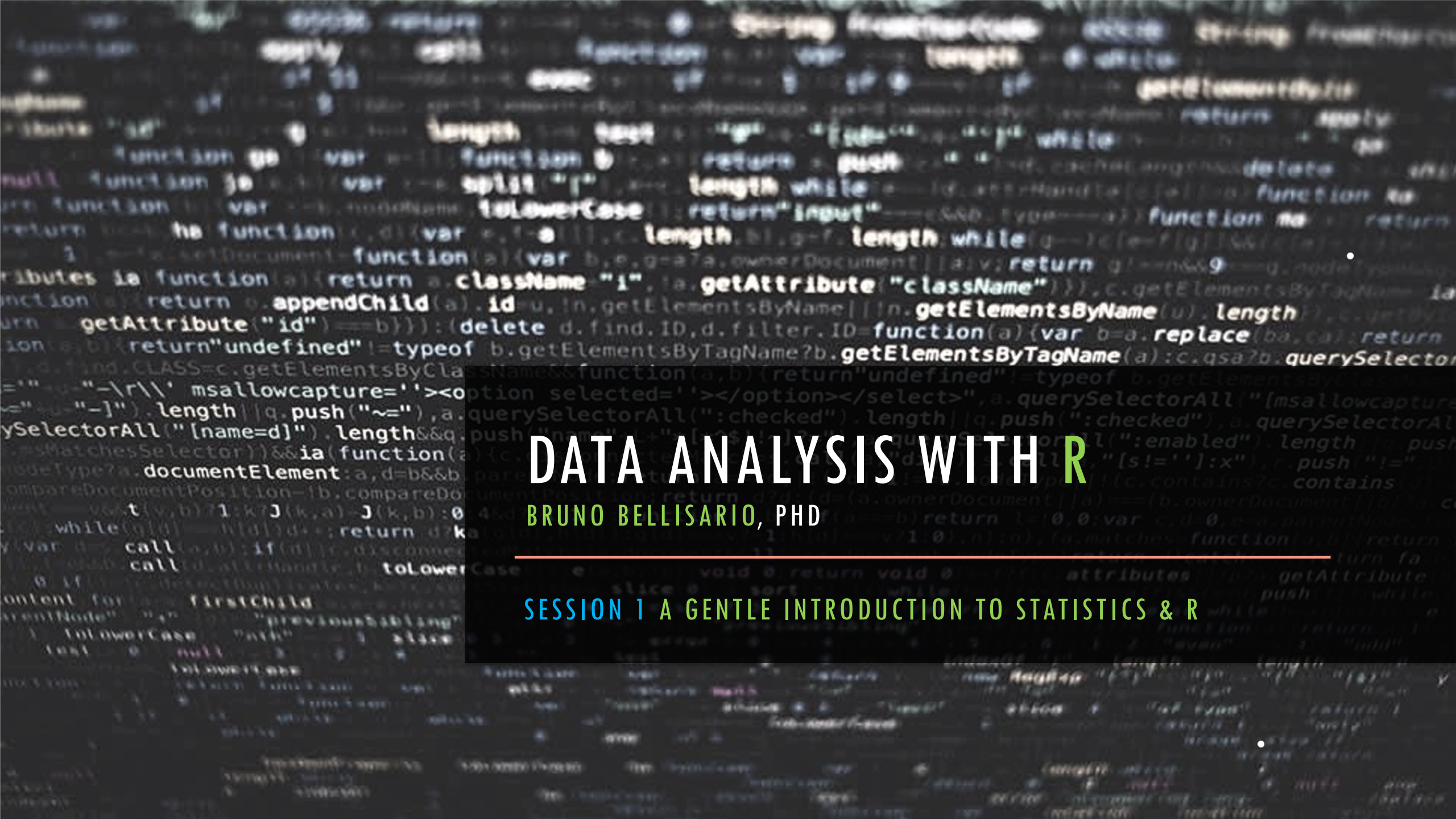


DATA ANALYSIS WITH R

BRUNO BELLISARIO, PHD

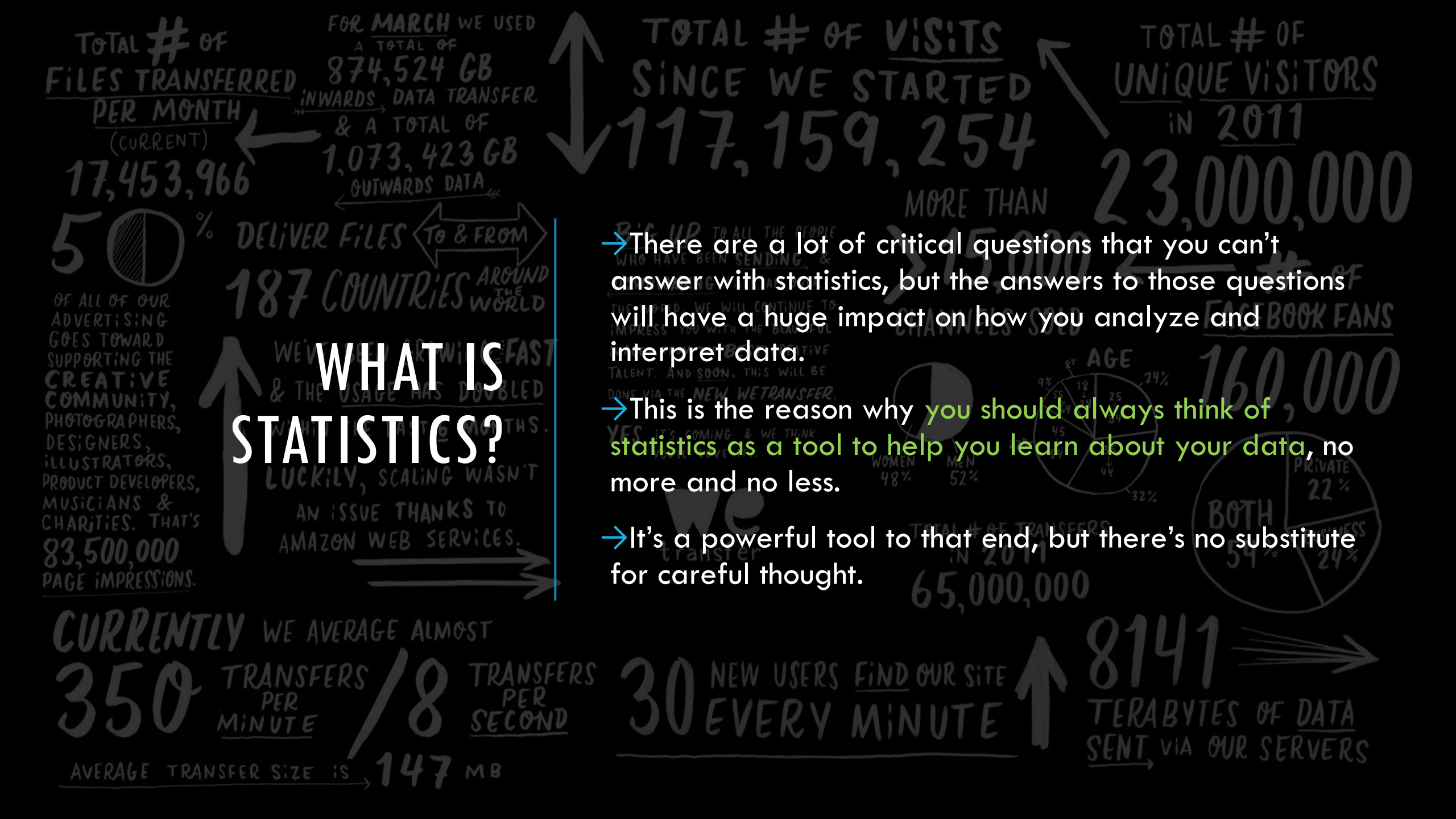
SESSION 1 A GENTLE INTRODUCTION TO STATISTICS & R

WHY DO WE LEARN STATISTICS?

- A big part of this issue at hand relates to the very idea of statistics. What is it? What's it there for? And why are scientists so bloody obsessed with it?
- As a group, scientists seem to be bizarrely fixated on running statistical tests on everything. In fact, we use statistics so often that we sometimes forget to explain to people why we do. It's a kind of article of faith among scientists that your findings can't be trusted until you've done some stats.

WHAT IS STATISTICS?

- The field of statistics is the science of learning from data.
- Statistical knowledge helps you use the proper methods to collect the data, employ the correct analyses, and effectively present the results.
- Statistics is a crucial process behind how we make discoveries in science, make decisions based on data, and make predictions.
- Statistics allows you to understand a subject much more deeply.



WHAT IS STATISTICS?

→There are a lot of critical questions that you can't answer with statistics, but the answers to those questions will have a huge impact on how you analyze and interpret data.

→This is the reason why you should always think of statistics as a tool to help you learn about your data, no more and no less.

→It's a powerful tool to that end, but there's no substitute for careful thought.

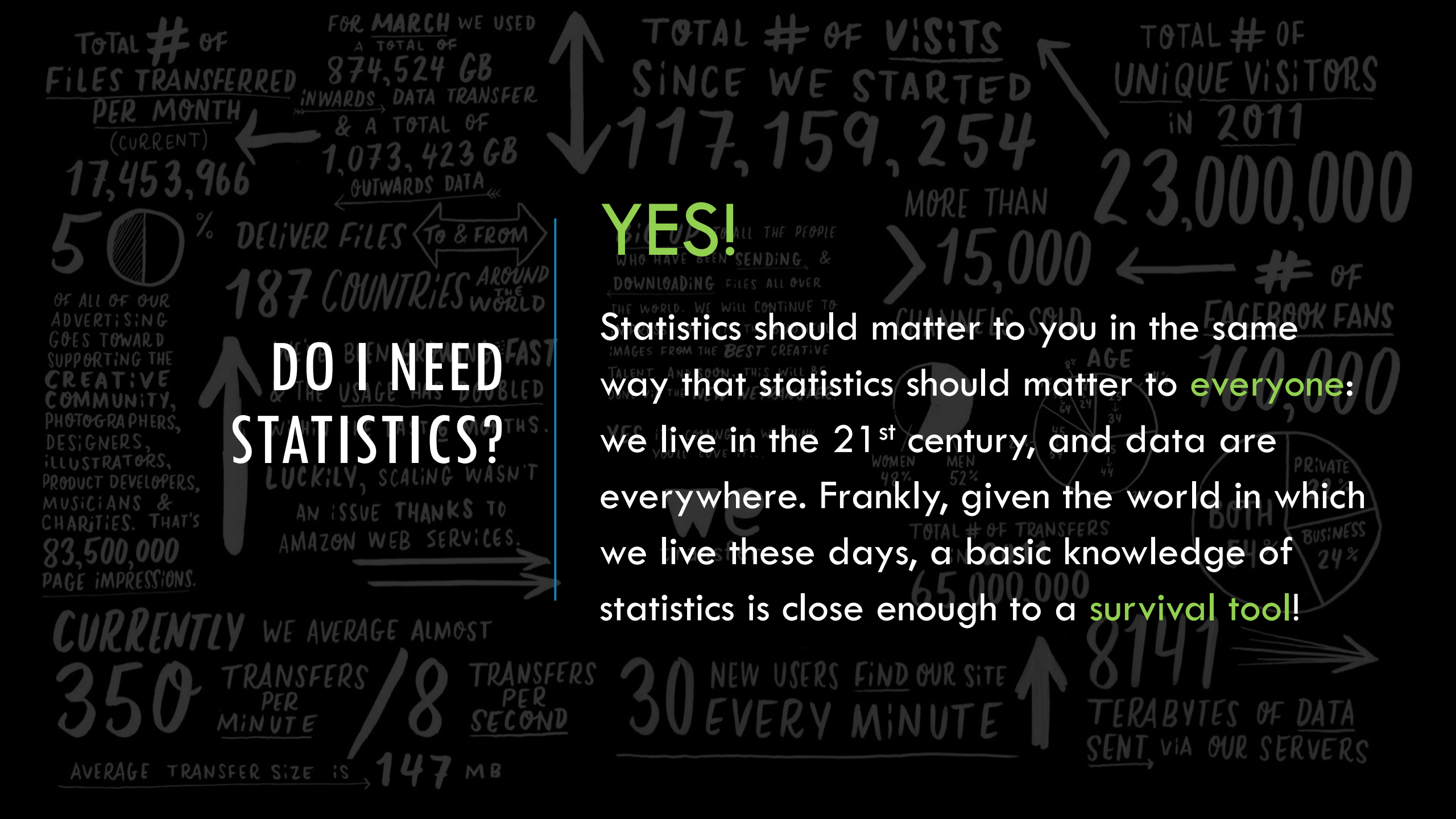
CAN'T SOMEONE ELSE DO THE STATISTICS?

You don't need to become a fully trained statistician. You do need to reach a certain level of statistical competence.

→Statistics is deeply intertwined with research design. If you want to be good at designing studies, you need to at least understand the basics of stats.

→If you want to be a good researcher, then you need to be able to understand the literature. But almost every paper reports the results of statistical analyses. So, if you really want to understand your research field, you need to be able to understand what other people did with their data. And that means understanding a certain amount of statistics.

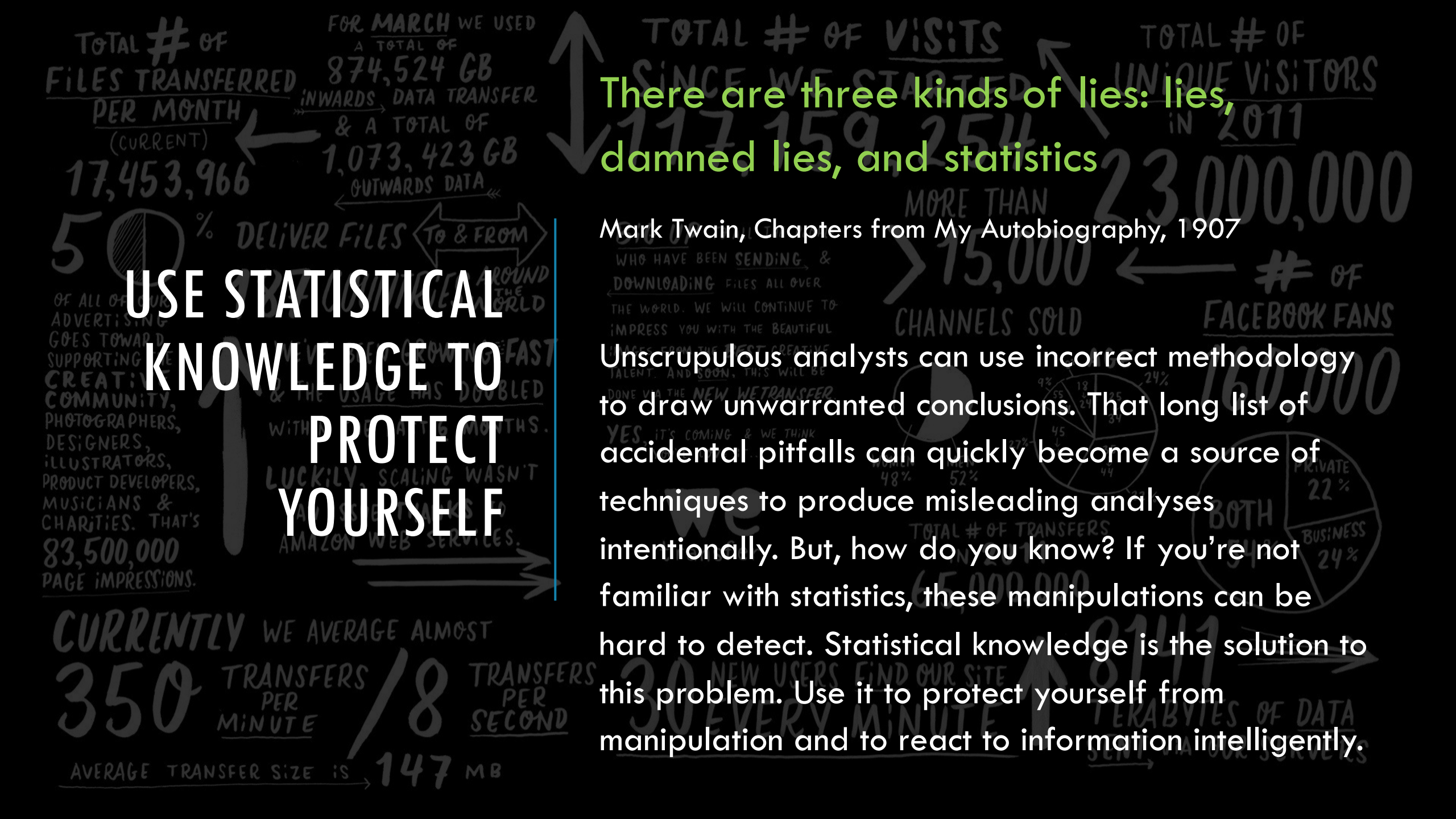
→There's a big practical problem with being dependent on other people to do all your statistics: statistical analysis is expensive.



DO I NEED STATISTICS?

YES!

Statistics should matter to you in the same way that statistics should matter to everyone: we live in the 21st century, and data are everywhere. Frankly, given the world in which we live these days, a basic knowledge of statistics is close enough to a survival tool!



USE STATISTICAL KNOWLEDGE TO PROTECT YOURSELF

There are three kinds of lies: lies, damned lies, and statistics

Mark Twain, Chapters from My Autobiography, 1907

Unscrupulous analysts can use incorrect methodology to draw unwarranted conclusions. That long list of accidental pitfalls can quickly become a source of techniques to produce misleading analyses intentionally. But, how do you know? If you're not familiar with statistics, these manipulations can be hard to detect. Statistical knowledge is the solution to this problem. Use it to protect yourself from manipulation and to react to information intelligently.

USE STATISTICAL KNOWLEDGE TO PROTECT OUR KNOWLEDGE

Open access, freely available online

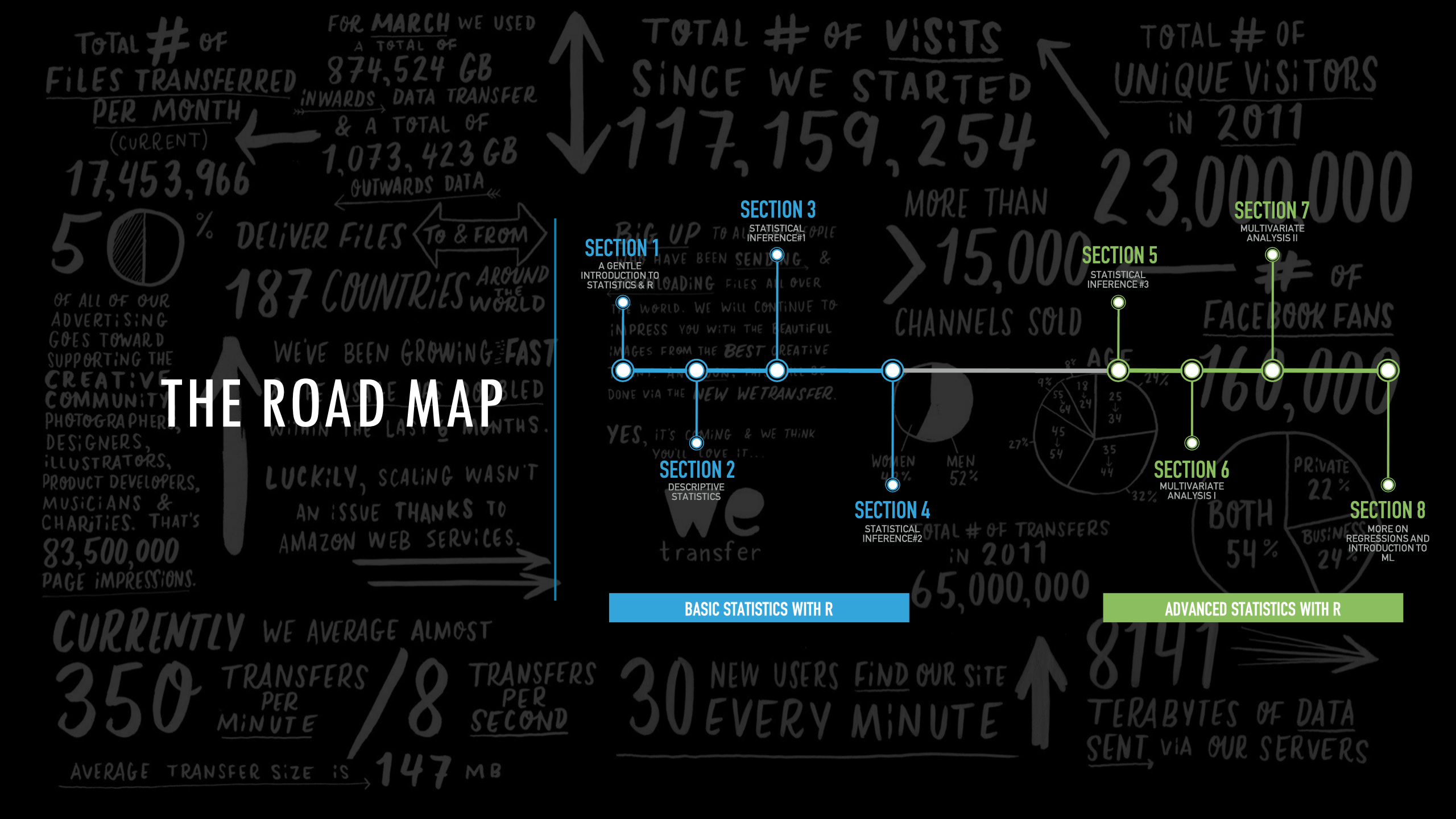
Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

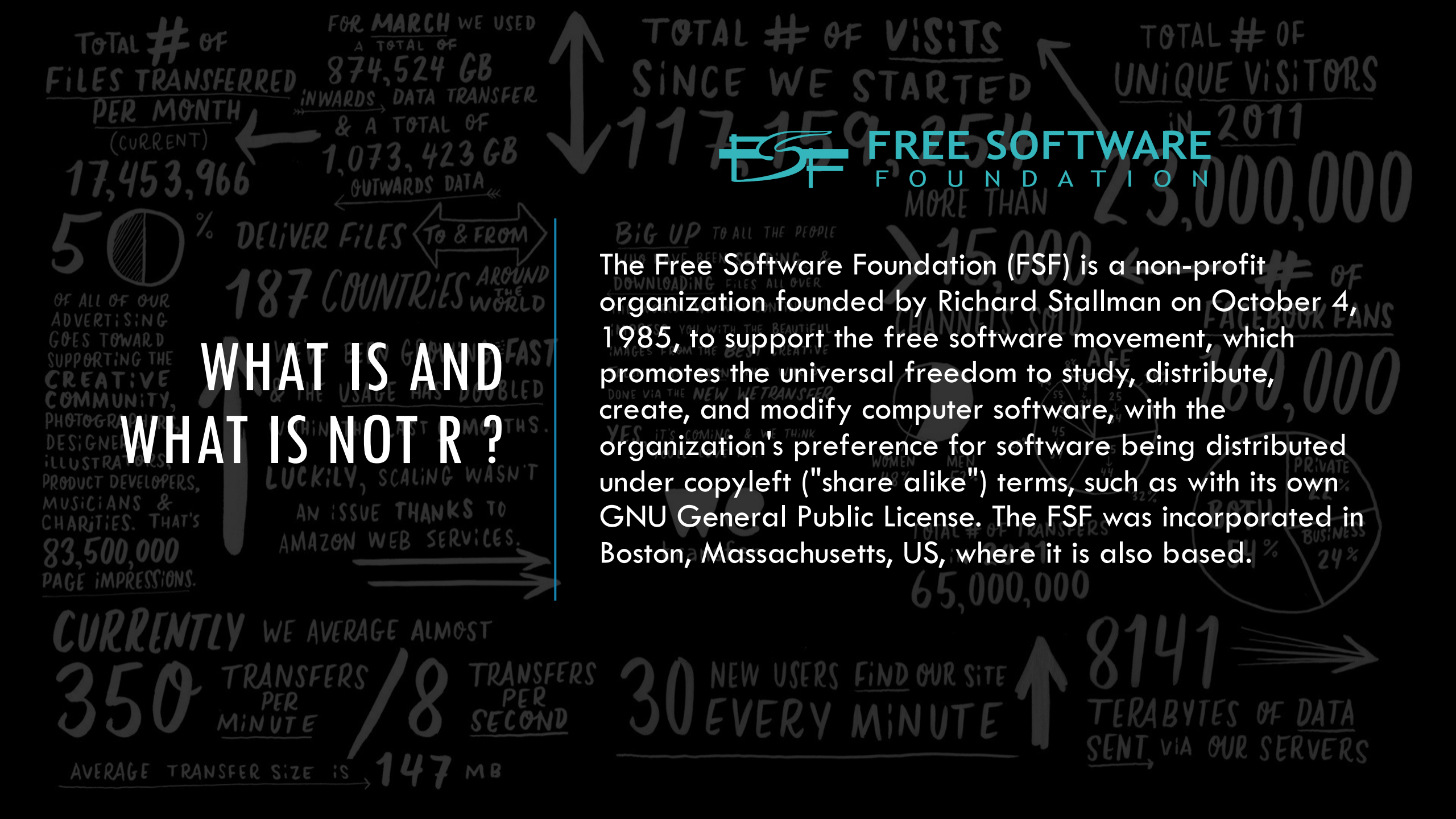
There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

THE ROAD MAP


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
TOTAL # OF VISITS SINCE WE STARTED 117,159,254
TOTAL # OF UNIQUE VISITORS IN 2011 23,000,000
50 % 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.
DELIVER FILES TO & FROM 187 COUNTRIES AROUND THE WORLD
WE'VE BEEN GROWING FAST
LUCKILY, SCALING WASN'T AN ISSUE THANKS TO AMAZON WEB SERVICES.
MORE THAN 15,000 CHANNELS SOLD
OF FACEBOOK FANS 160,000
SECTION 1 A GENTLE INTRODUCTION TO STATISTICS & R
SECTION 2 DESCRIPTIVE STATISTICS
SECTION 3 STATISTICAL INFERENCE#1
SECTION 4 STATISTICAL INFERENCE#2
SECTION 5 STATISTICAL INFERENCE #3
SECTION 6 MULTIVARIATE ANALYSIS I
SECTION 7 MULTIVARIATE ANALYSIS II
SECTION 8 MORE ON REGRESSIONS AND INTRODUCTION TO ML
WOMEN 48% MEN 52%
9% 18 ↓ 24 25 ↓ 34 24% 35 ↓ 44 27% 45 ↓ 54 55 ↓ 64 8%
PRIVATE 22% BUSINESS 24% BOTH 54%
BASIC STATISTICS WITH R
ADVANCED STATISTICS WITH R
TOTAL # OF TRANSFERS IN 2011 65,000,000
CURRENTLY WE AVERAGE ALMOST 350 TRANSFERS PER MINUTE / 8 TRANSFERS PER SECOND
AVERAGE TRANSFER SIZE IS 147 MB
30 NEW USERS FIND OUR SITE EVERY MINUTE
8141 TERABYTES OF DATA SENT VIA OUR SERVERS

WHAT IS AND WHAT IS NOT R ?

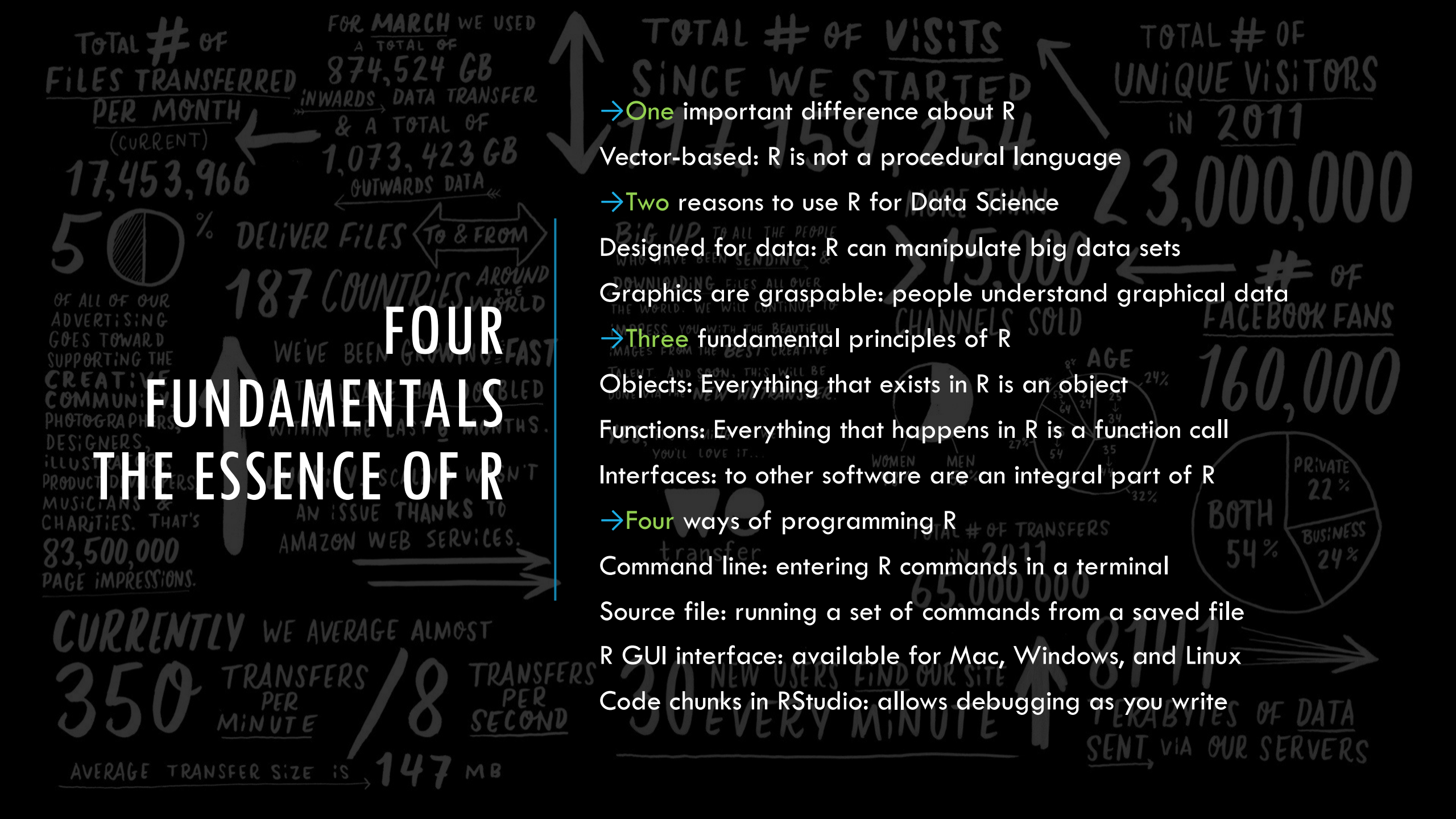
- R is a language and environment for statistical computing and graphics available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form.
- R is a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.
- It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.
- R is not a 'conventional' software nor a statistics system but an environment within which statistical techniques are implemented



WHAT IS AND WHAT IS NOT R ?

FREE SOFTWARE FOUNDATION

The Free Software Foundation (FSF) is a non-profit organization founded by Richard Stallman on October 4, 1985, to support the free software movement, which promotes the universal freedom to study, distribute, create, and modify computer software, with the organization's preference for software being distributed under copyleft ("share alike") terms, such as with its own GNU General Public License. The FSF was incorporated in Boston, Massachusetts, US, where it is also based.



FOUR FUNDAMENTALS THE ESSENCE OF R

→One important difference about R

Vector-based: R is not a procedural language

→Two reasons to use R for Data Science

Designed for data: R can manipulate big data sets

Graphics are graspable: people understand graphical data

→Three fundamental principles of R

Objects: Everything that exists in R is an object

Functions: Everything that happens in R is a function call

Interfaces: to other software are an integral part of R

→Four ways of programming R

Command line: entering R commands in a terminal

Source file: running a set of commands from a saved file

R GUI interface: available for Mac, Windows, and Linux

Code chunks in RStudio: allows debugging as you write

BASIC MATHS

R has all the basic mathematical functions and obeys the standard order of mathematical operations (PEMDAS):

→Parentheses ( )

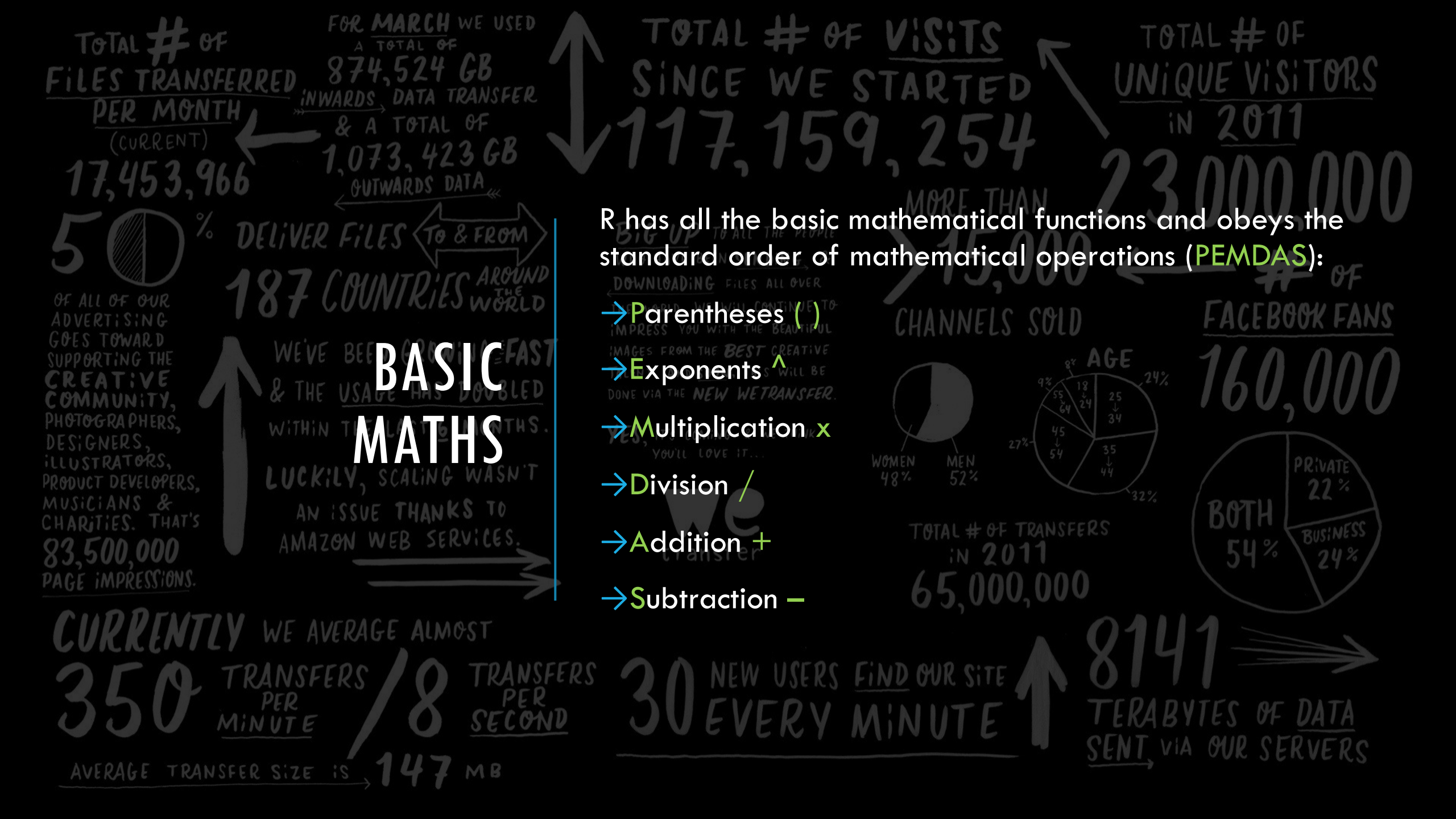
→Exponents ^

→Multiplication x

→Division /

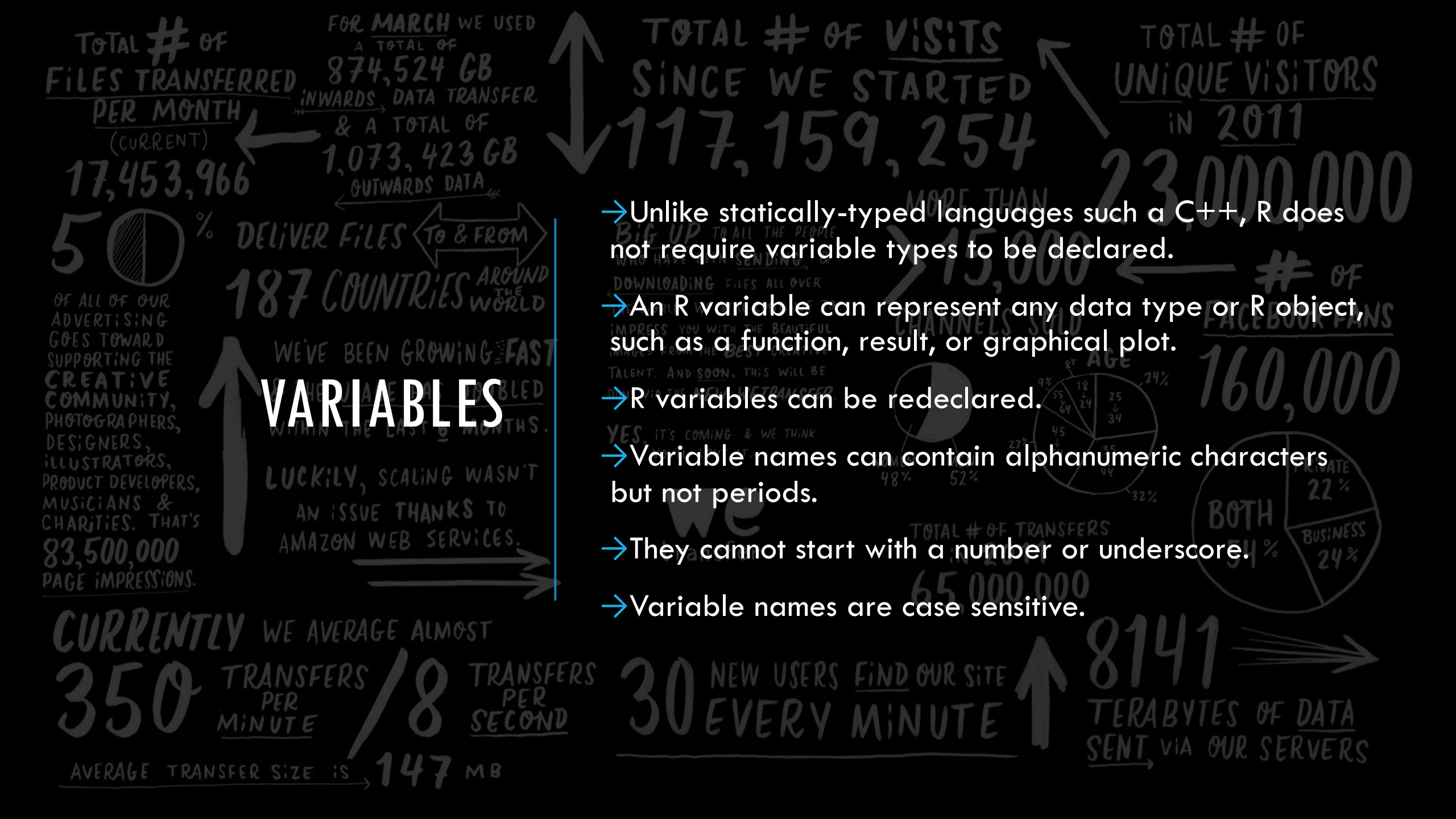
→Addition +

→Subtraction –

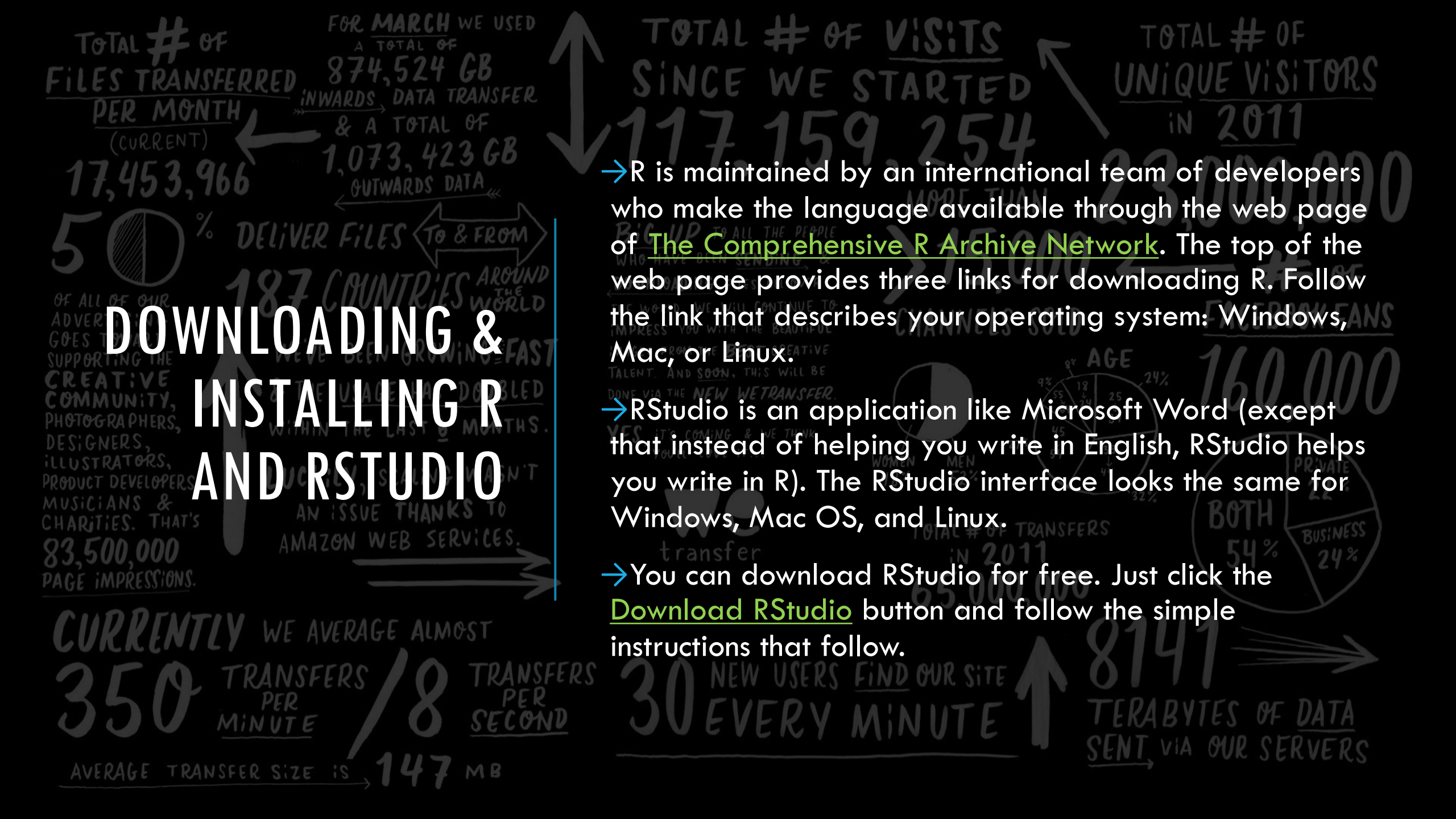


VARIABLES

- Unlike statically-typed languages such a C++, R does not require variable types to be declared.
- An R variable can represent any data type or R object, such as a function, result, or graphical plot.
- R variables can be redeclared.
- Variable names can contain alphanumeric characters but not periods.
- They cannot start with a number or underscore.
- Variable names are case sensitive.



DOWNLOADING & INSTALLING R AND RSTUDIO



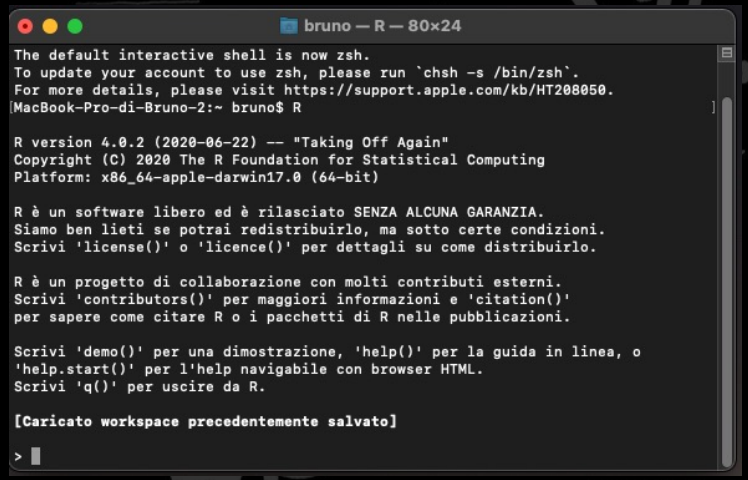
→R is maintained by an international team of developers who make the language available through the web page of The Comprehensive R Archive Network. The top of the web page provides three links for downloading R. Follow the link that describes your operating system: Windows, Mac, or Linux.

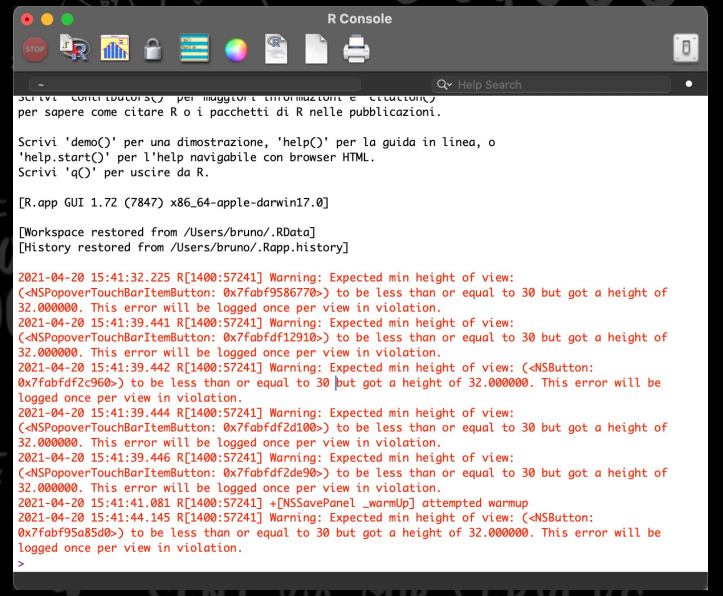
→RStudio is an application like Microsoft Word (except that instead of helping you write in English, RStudio helps you write in R). The RStudio interface looks the same for Windows, Mac OS, and Linux.

→You can download RStudio for free. Just click the Download RStudio button and follow the simple instructions that follow.

PANELS & TOOLBARS

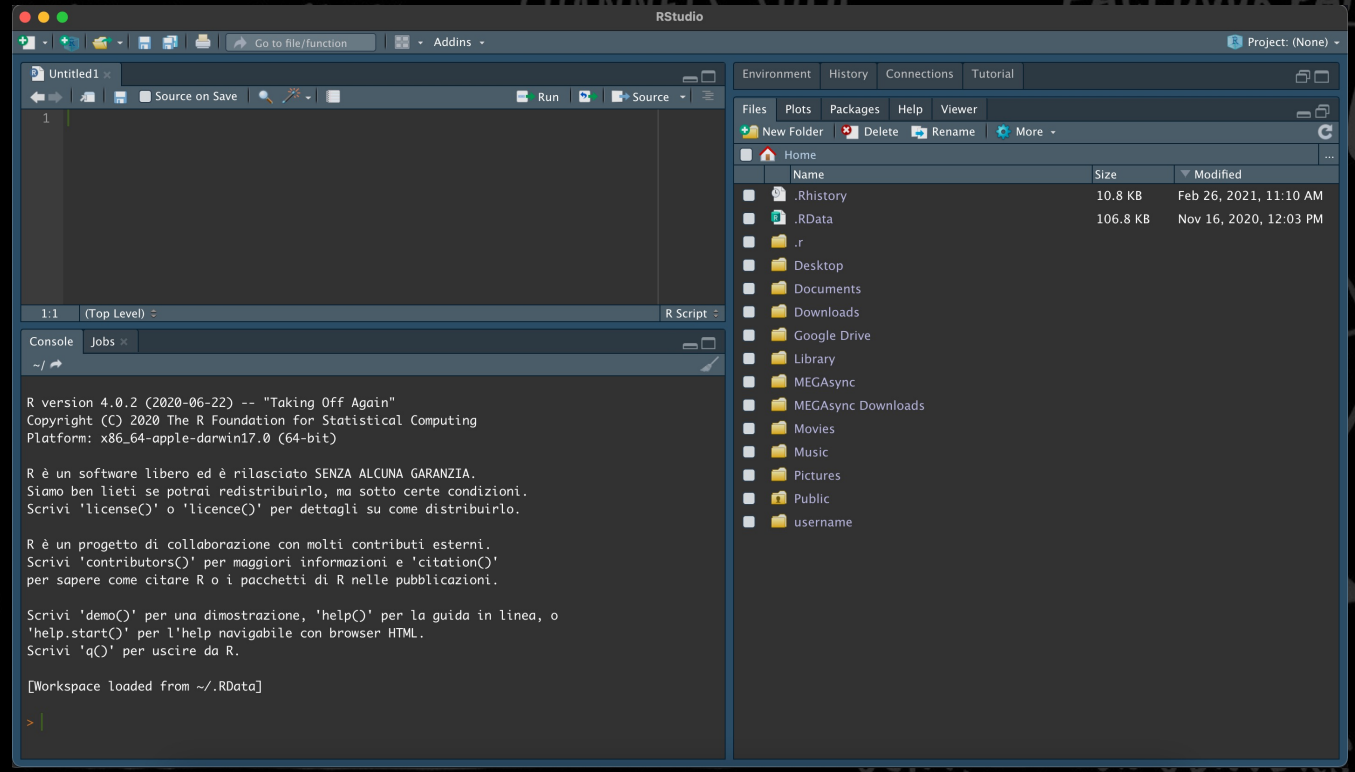
→R is a command line driven program. The user enters commands at the prompt (> by default) and each command is executed one at a time.





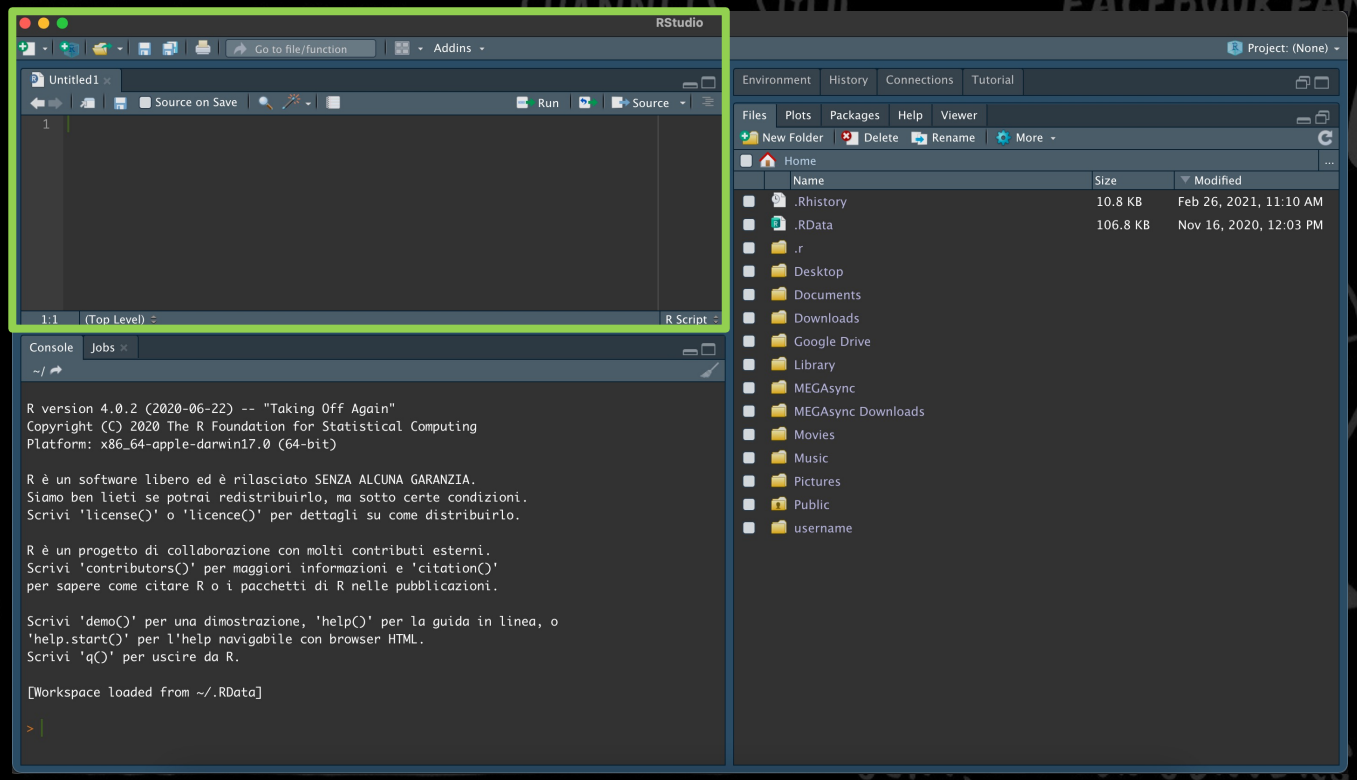
PANELS & TOOLBARS

RStudio is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing RStudio using a web browser.



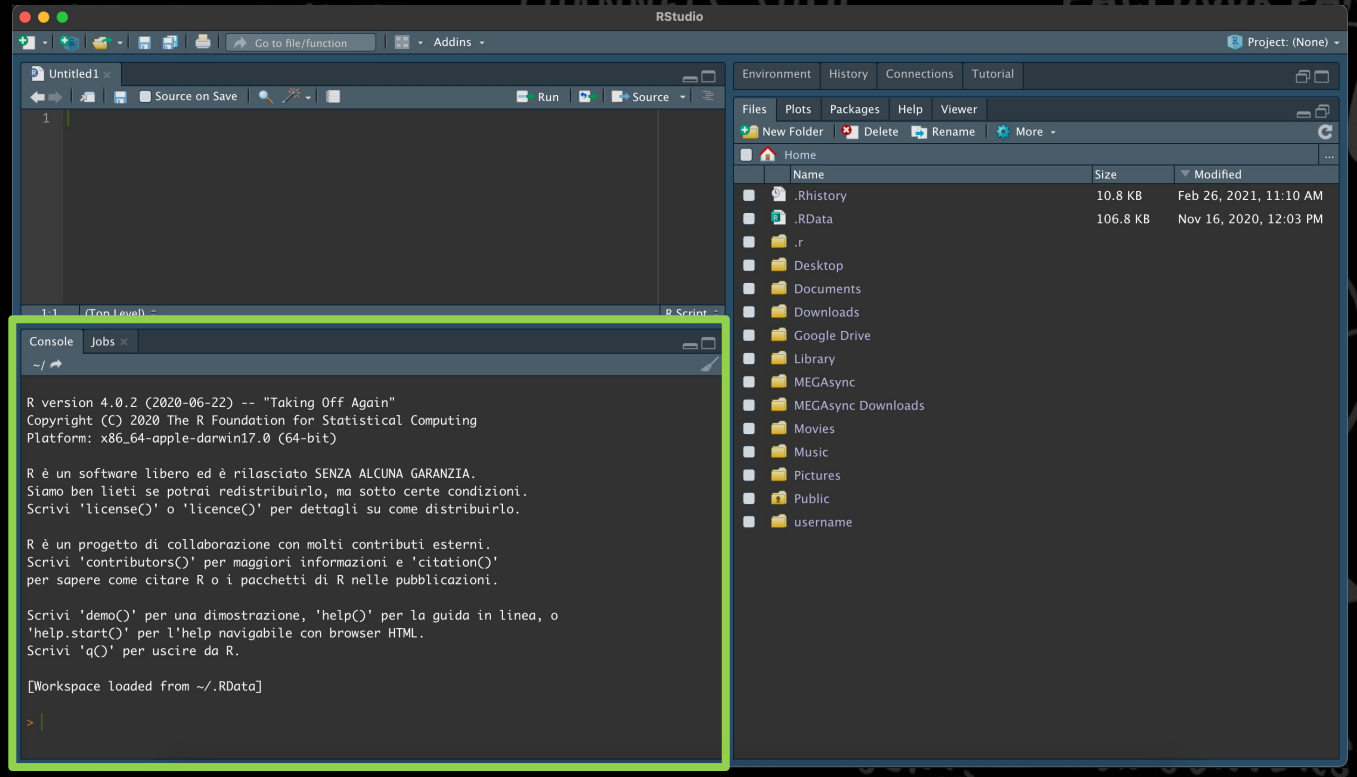
PANELS & TOOLBARS

The script editor is in the upper left of the RStudio environment. If you are starting a new R session, or have closed the last open R script, the script editor is not visible. You can open the script editor by either creating a new empty script or by opening an existing script.



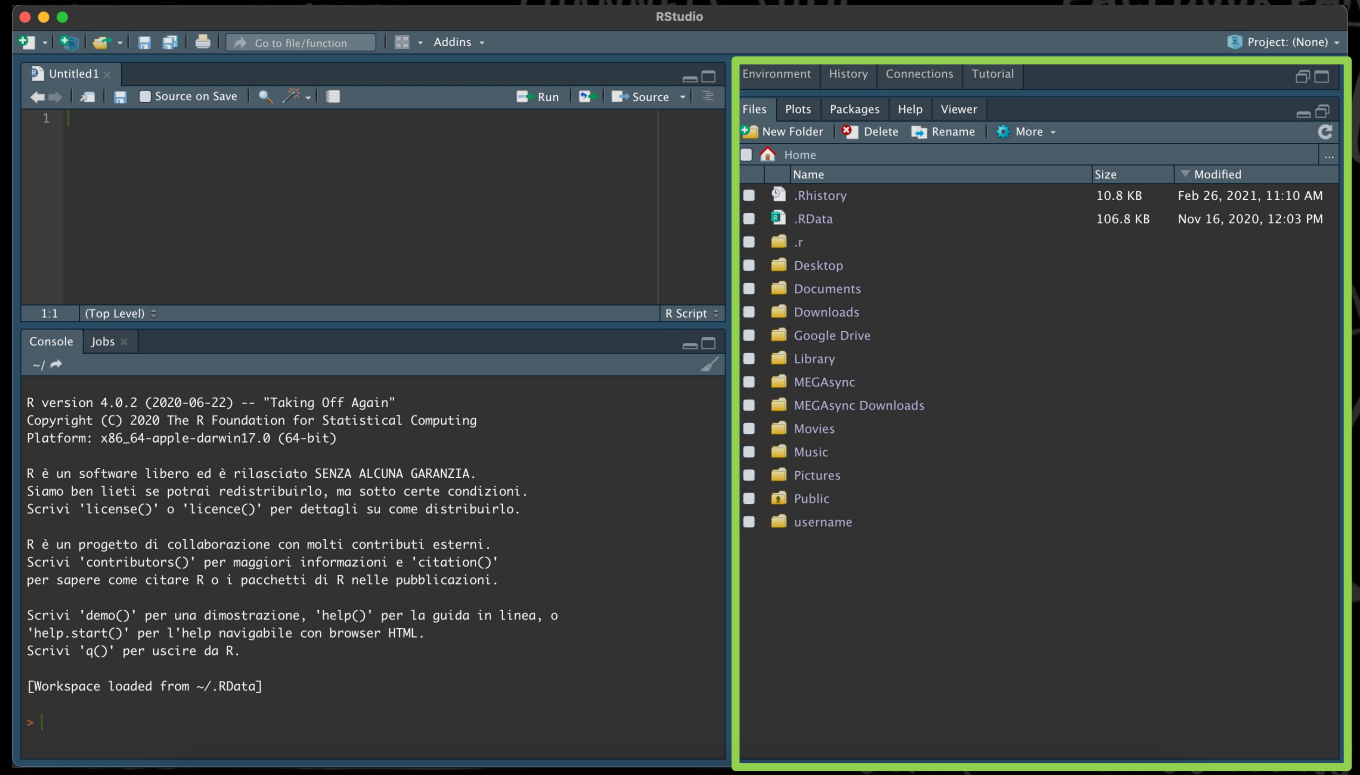
PANELS & TOOLBARS

The command console is where R executes commands and usually responds to those commands. The one exception is when R draws a plot. The command console is in the lower left of the environment.



PANELS & TOOLBARS

In the utility panel, you can find the plot area, the directory, the history and environment panel.

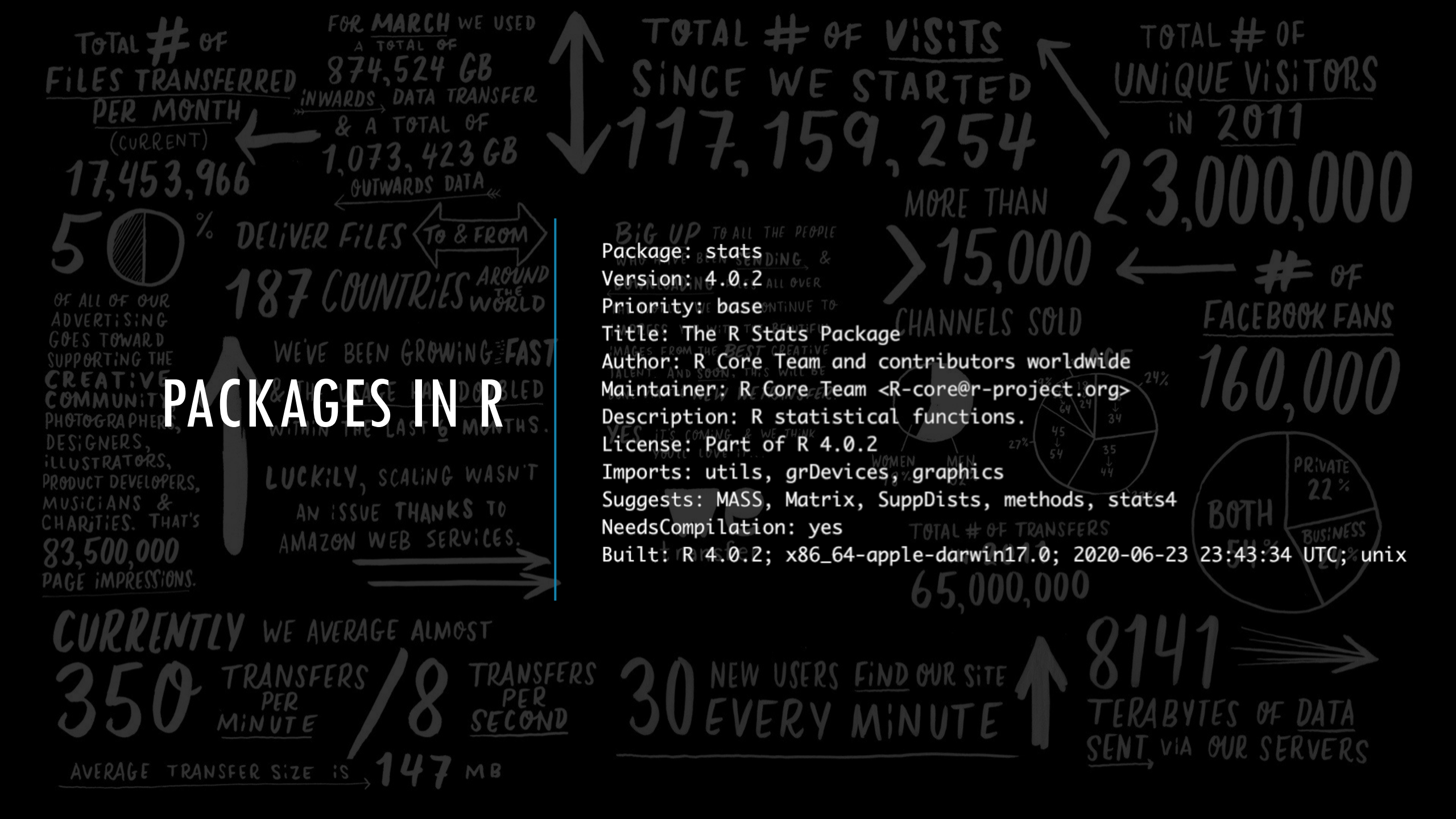


PACKAGES IN R

- Packages are collections of functions and data sets developed by the community. They increase the power of R by improving existing base R functionalities, or by adding new ones.
- A package includes code, documentation for the package and the functions inside, some tests to check everything works as it should, and data sets.
- The basic information about a package is provided in the DESCRIPTION file, where you can find out what the package does, who the author is, what version the documentation belongs to, the date, the type of license its use, and the package dependencies.

PACKAGES IN R

```
Package: stats
Version: 4.0.2
Priority: base
Title: The R Stats Package
Author: R Core Team and contributors worldwide
Maintainer: R Core Team <R-core@r-project.org>
Description: R statistical functions.
License: Part of R 4.0.2
Imports: utils, grDevices, graphics
Suggests: MASS, Matrix, SuppDists, methods, stats4
NeedsCompilation: yes
Built: R 4.0.2; x86_64-apple-darwin17.0; 2020-06-23 23:43:34 UTC; unix
```



PACKAGES IN R

Packages can be found in ad-hoc repositories

→CRAN: the official repository, it is a network of ftp and web servers maintained by the R community around the world.

→Bioconductor: this is a topic specific repository, intended for open-source software for bioinformatics

→Github : although this is not R specific, Github is probably the most popular repository for open-source projects

Packages can be installed by using the console panel and the syntax `install.packages("package.name")` or by using the toolbar “packages” in the utility panel

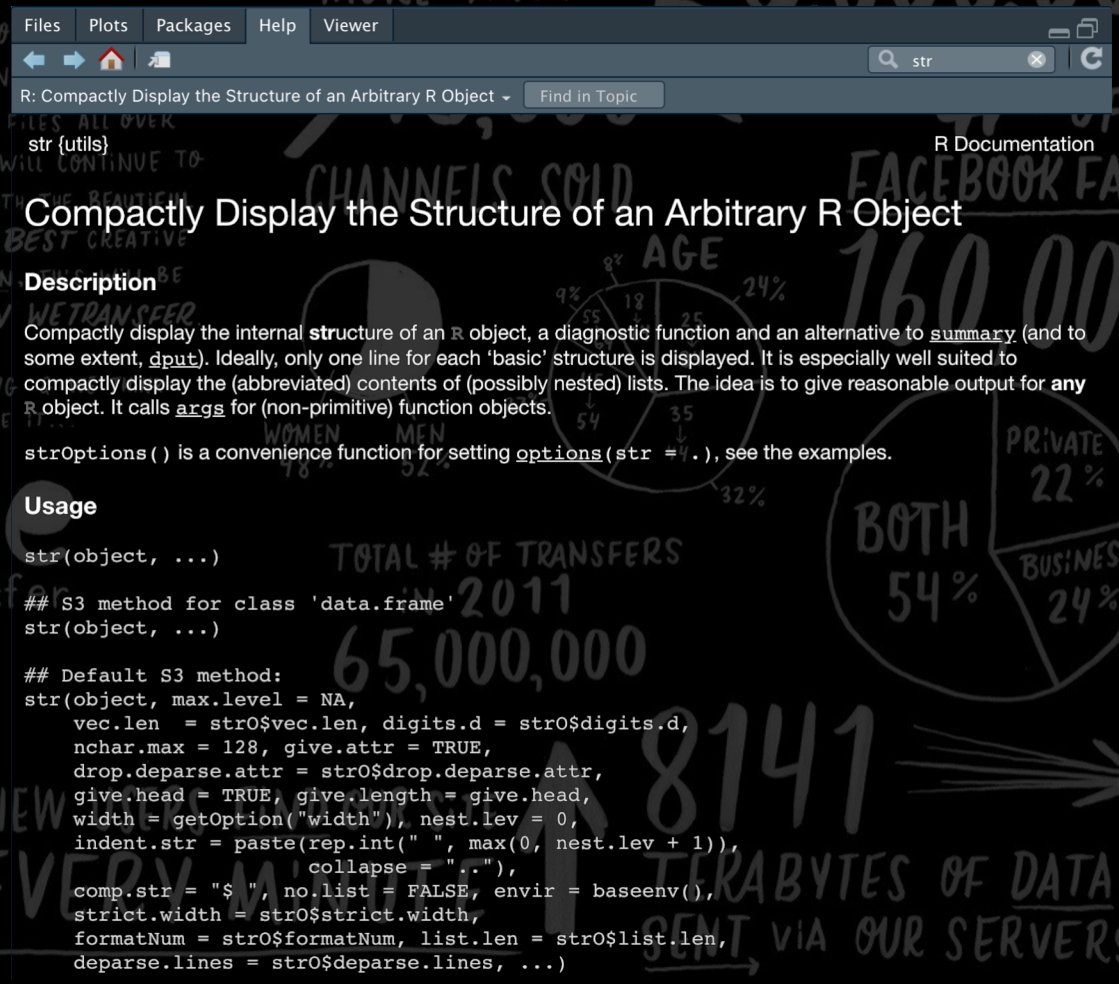
PACKAGES IN R

```
Installing package into
'/home/username/R/x86_64-pc-linux-gnu-
library/3.3' (as 'lib' is unspecified)
trying URL
'https://cran.rstudio.com/src/contrib/vioplot_0
.2.tar.gz' Content type 'application/x-gzip'
length 3801 bytes
==================================================
==== downloaded 3801 bytes

* installing *source* package 'vioplot' ... **
R ** preparing package for lazy loading ** help
*** installing help indices ** building package
indices ** testing if installed package can be
loaded * DONE (vioplot)
```

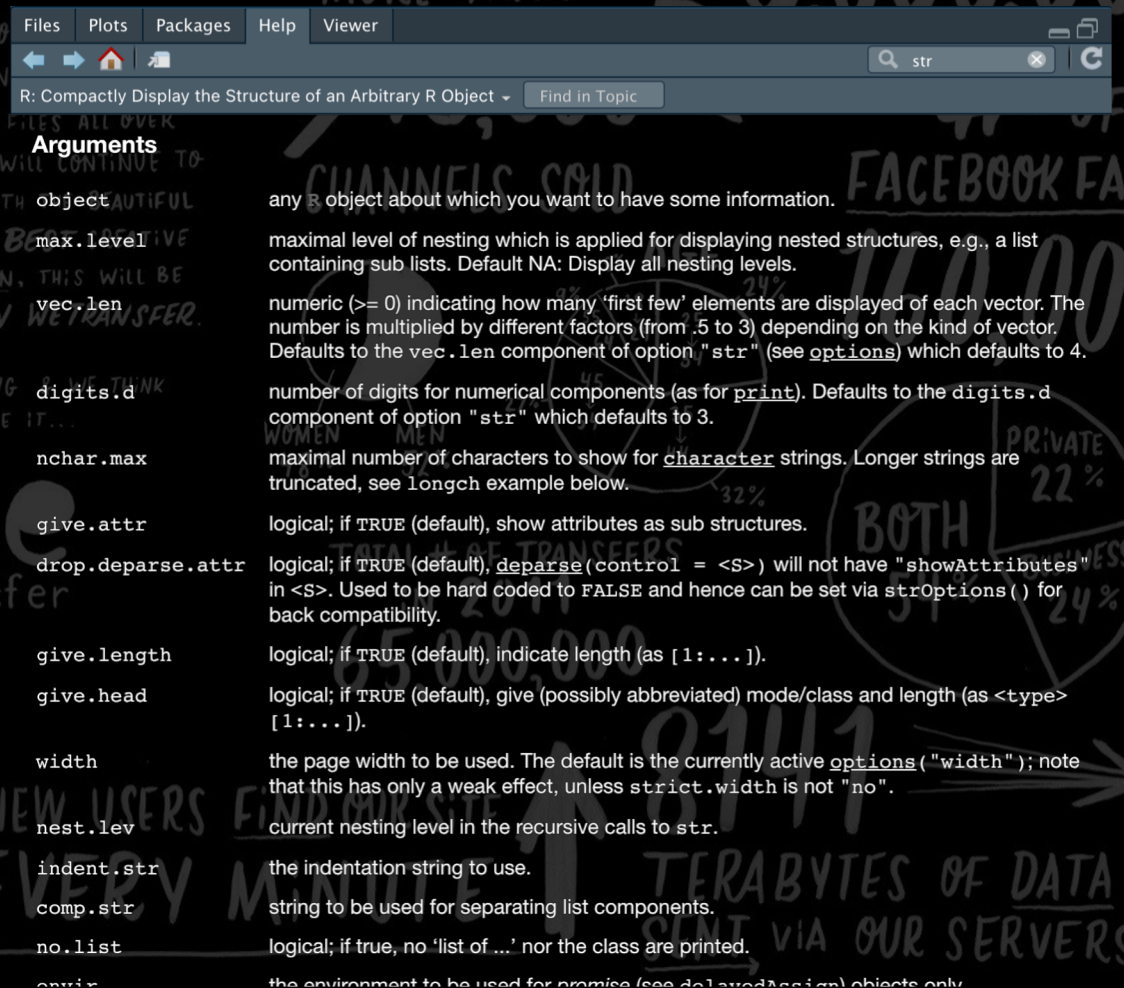
USING THE HELP TAB

The help panel has its own toolbar. This toolbar contains two tool buttons. They are home [this will return you to the help home page and clear you help search history] and search help.



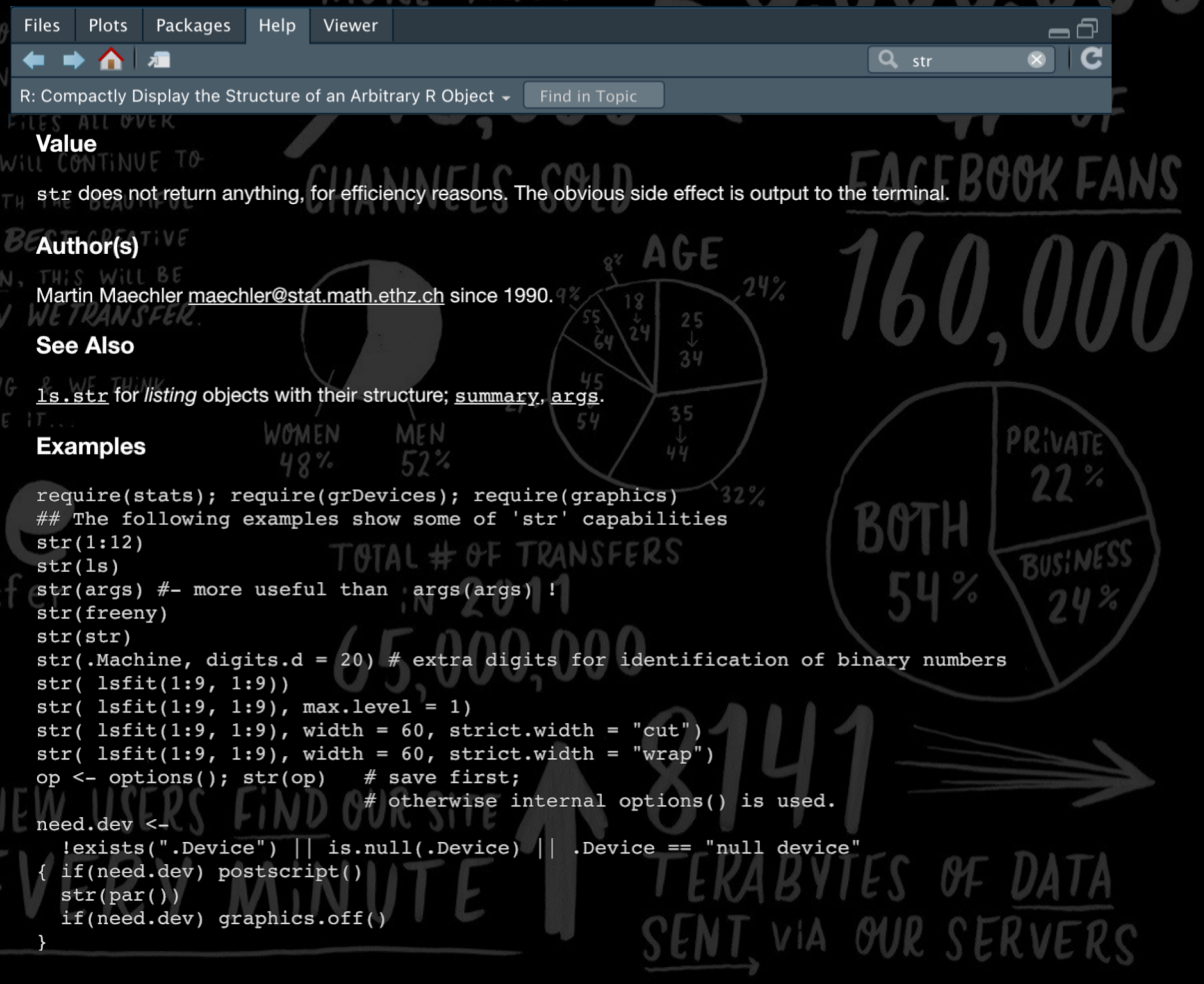
USING THE HELP TAB

The help panel has its own toolbar. This toolbar contains two tool buttons. They are home [this will return you to the help home page and clear you help search history] and search help.



USING THE HELP TAB

The help panel has its own toolbar. This toolbar contains two tool buttons. They are home [this will return you to the help home page and clear you help search history] and search help.



DATA STRUCTURE & ACQUISITION

Data types

→Plain text and delimited files (.txt, .csv, .tsv)

→Microsoft Word and Excel files (.doc, .docx, .xls, .xlsx)

→Databases (SQLite, MySQL, Microsoft Access)

→Data in statistical software (SAS, SPSS, NCSS, Octave)

Data source

→in R and R packages

→on your computer storage

→on a local or linked network

→anywhere on the global Internet

DATA STRUCTURE & ACQUISITION

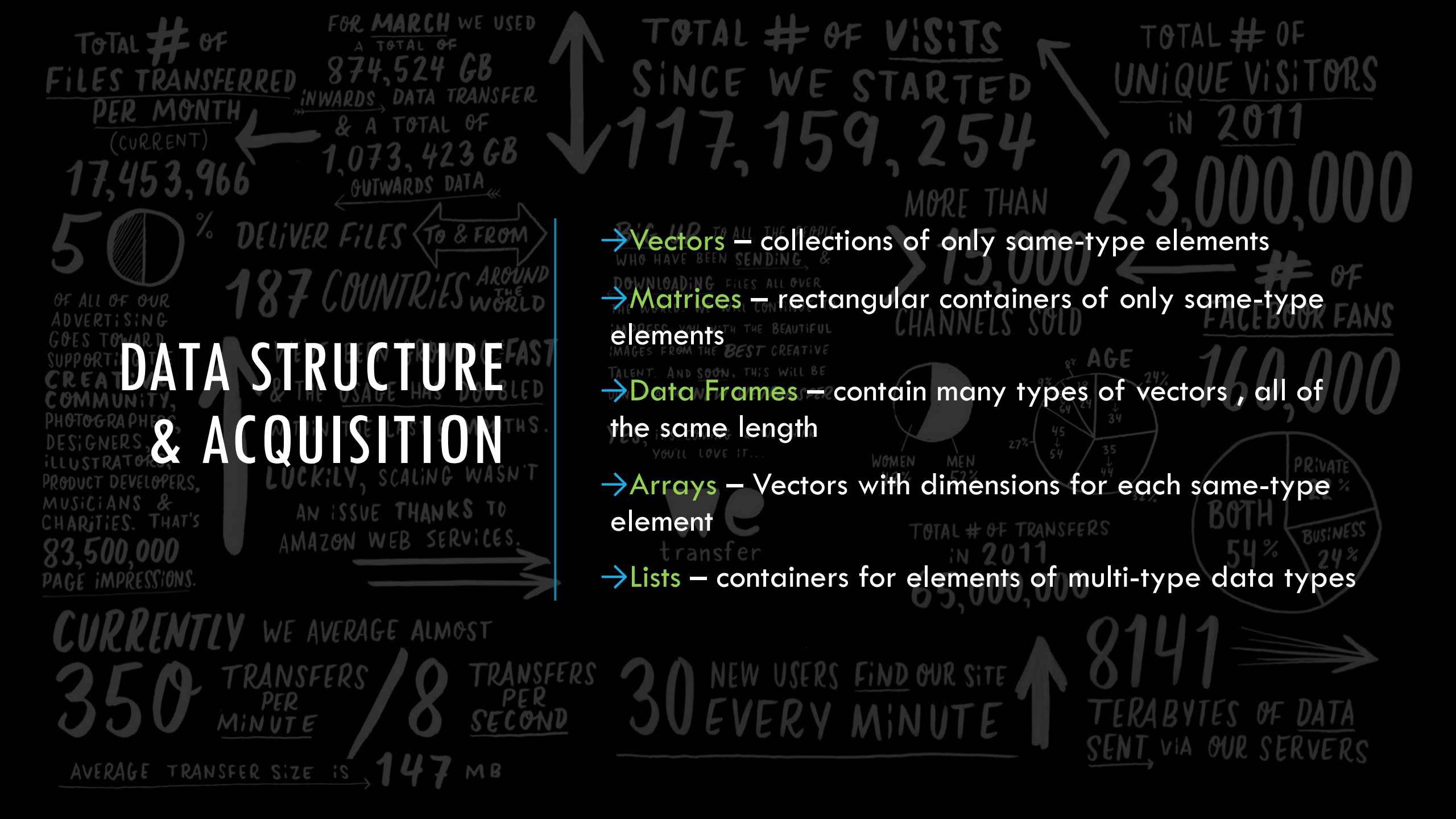
→Homogeneity: Tells us whether the data structure is 'homogeneous' i.e. it contains only similar types of data for instance all numeric, all string etc. , or a combination of multiple data types i.e. 'heterogeneous'.

→Dimension: Tells us in what fashion, order data will be stored, whether it is linear or 1D, tabular or 2D etc.

	1D	2D	Multi-D
Homogeneous	Vector	Matrix	Array
Heterogeneous	List	Dataframe	

DATA STRUCTURE & ACQUISITION

- →Vectors – collections of only same-type elements
- →Matrices – rectangular containers of only same-type elements
- →Data Frames – contain many types of vectors , all of the same length
- →Arrays – Vectors with dimensions for each same-type element
- →Lists – containers for elements of multi-type data types



DATA STRUCTURE & ACQUISITION

Importing data, using the `read()` function

It's the easiest way to import delimited data files, and results in a data.frame.

Three arguments:

→The path and filename or URL of the data file

→Whether the first line contains the column header labels

set to TRUE if and only if the first row contains one fewer fields than the number of columns

→The separator between data items

DATA AND SCRIPTS MUST ALWAYS BE CONTAINED IN THE SAME DIRECTORY!

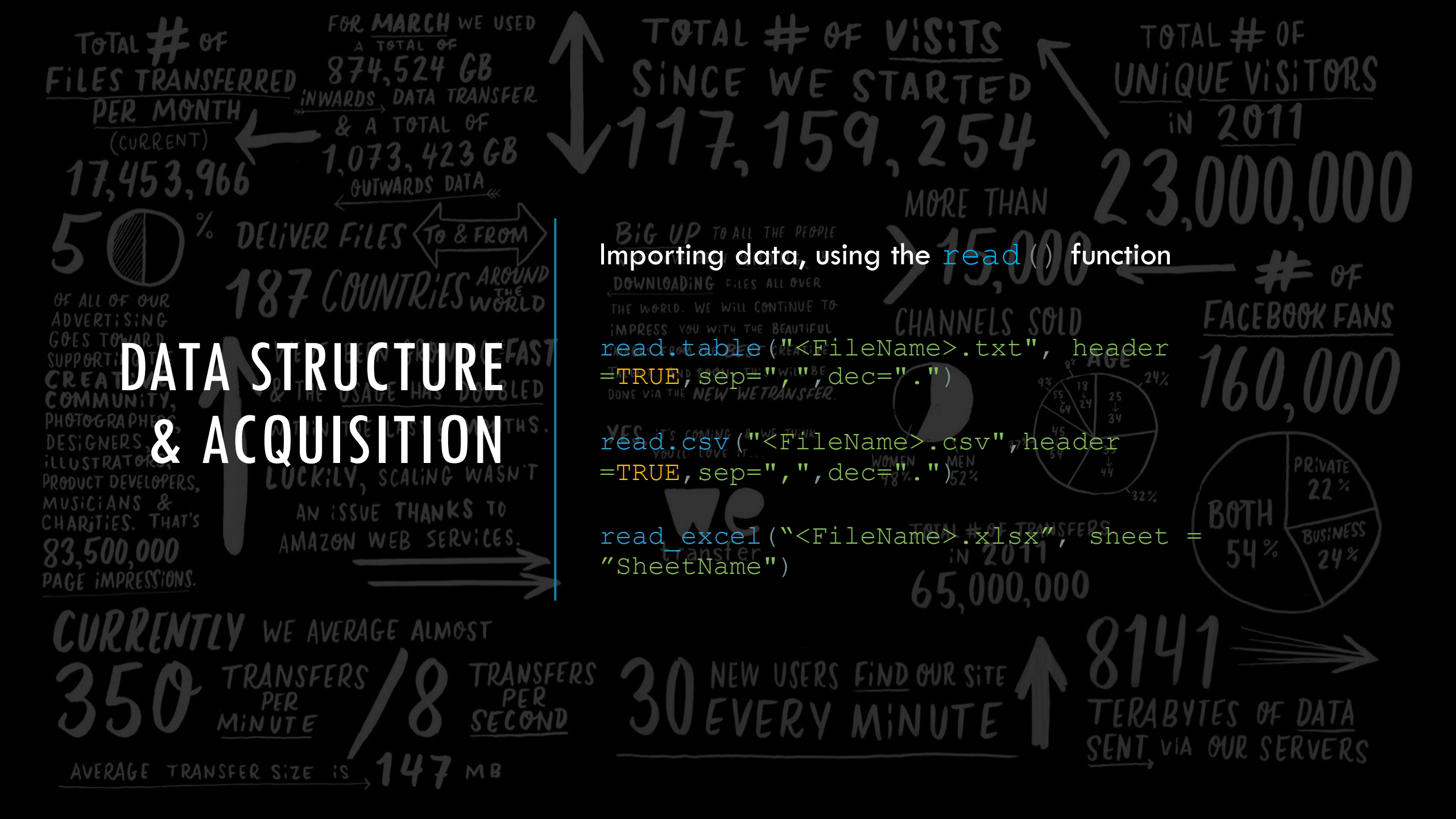
DATA STRUCTURE & ACQUISITION

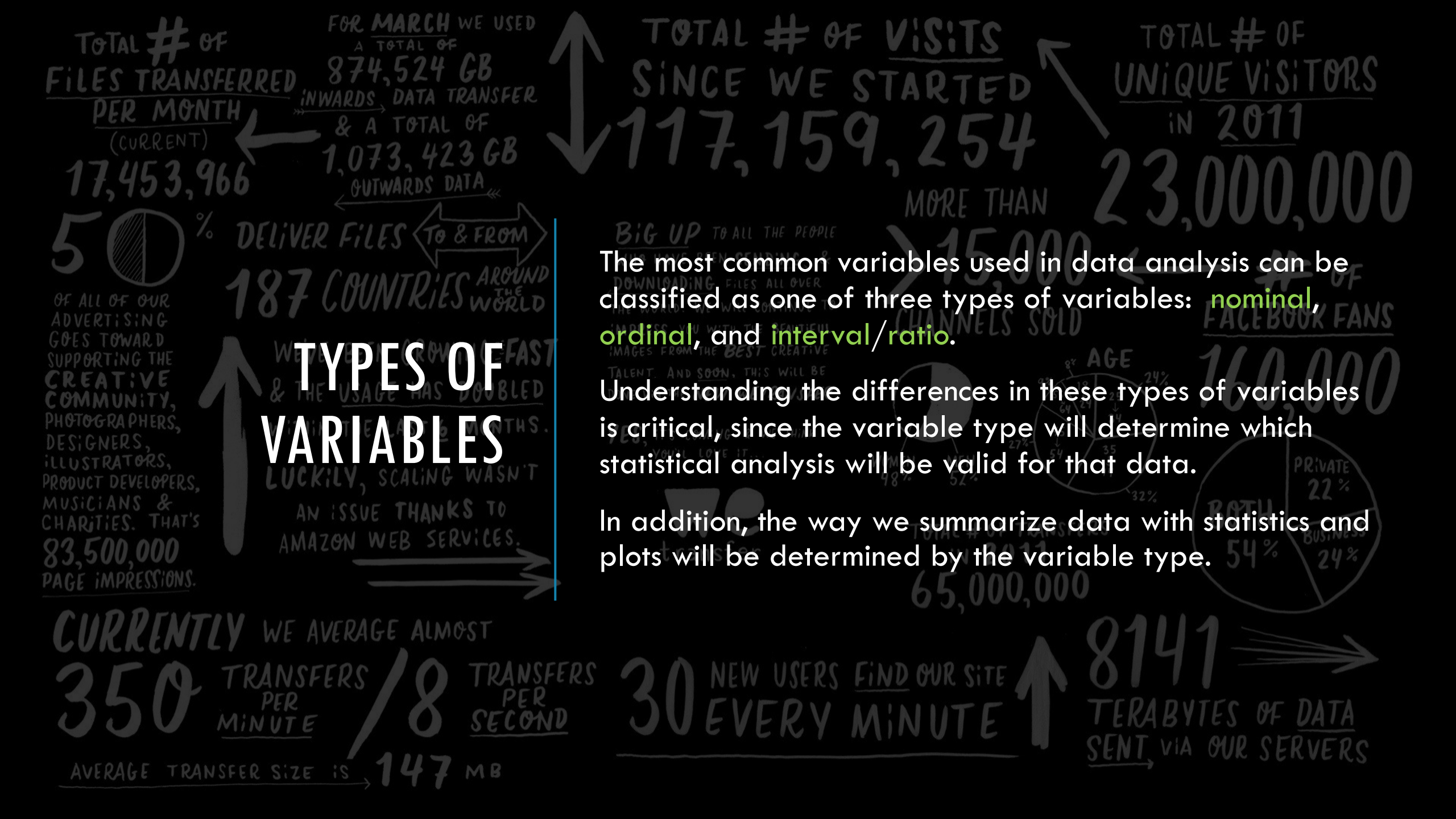
Importing data, using the `read()` function

```
read.table("<FileName>.txt", header
=TRUE,sep="","",dec=".")

read.csv("<FileName>.csv",header
=TRUE,sep="","",dec=".")

read_excel("<FileName>.xlsx", sheet =
"SheetName")
```





TYPES OF VARIABLES

The most common variables used in data analysis can be classified as one of three types of variables: nominal, ordinal, and interval/ratio.

Understanding the differences in these types of variables is critical, since the variable type will determine which statistical analysis will be valid for that data.

In addition, the way we summarize data with statistics and plots will be determined by the variable type.

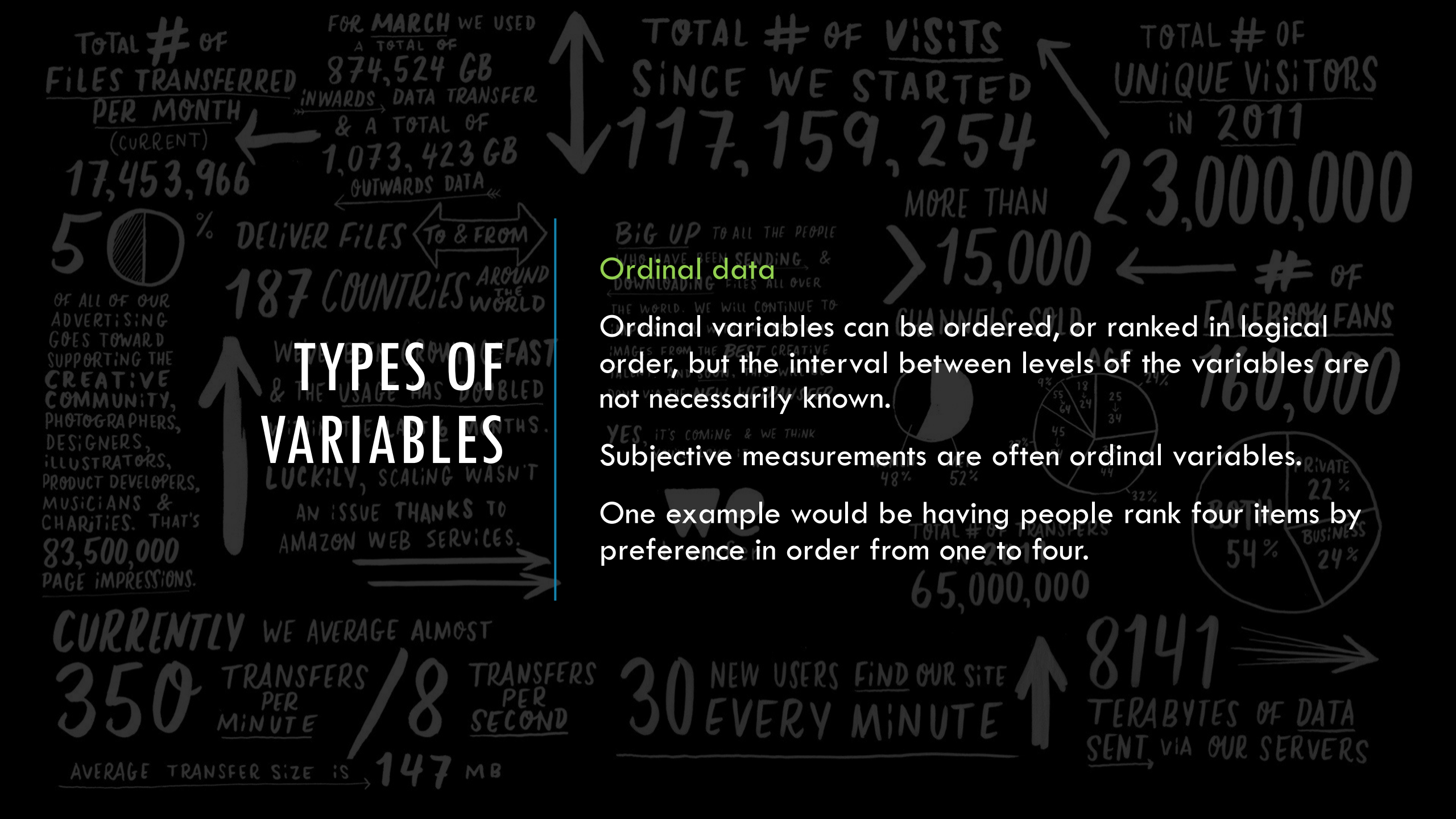
TYPES OF VARIABLES

Nominal data

Nominal variables are data whose levels are labels or descriptions, and which cannot be ordered. Examples of nominal variables are sex, school, and yes/no questions. They are also called “nominal categorical” or “qualitative” variables, and the levels of a variable are sometimes called “classes” or “groups”.

The levels of categorical variables cannot be ordered. For the variable sex, it makes no sense to try to put the levels “female”, “male”, and “other” in any numerical order.

If levels are numbered for convenience, the numbers are arbitrary, and the variable can’t be treated as a numeric variable.



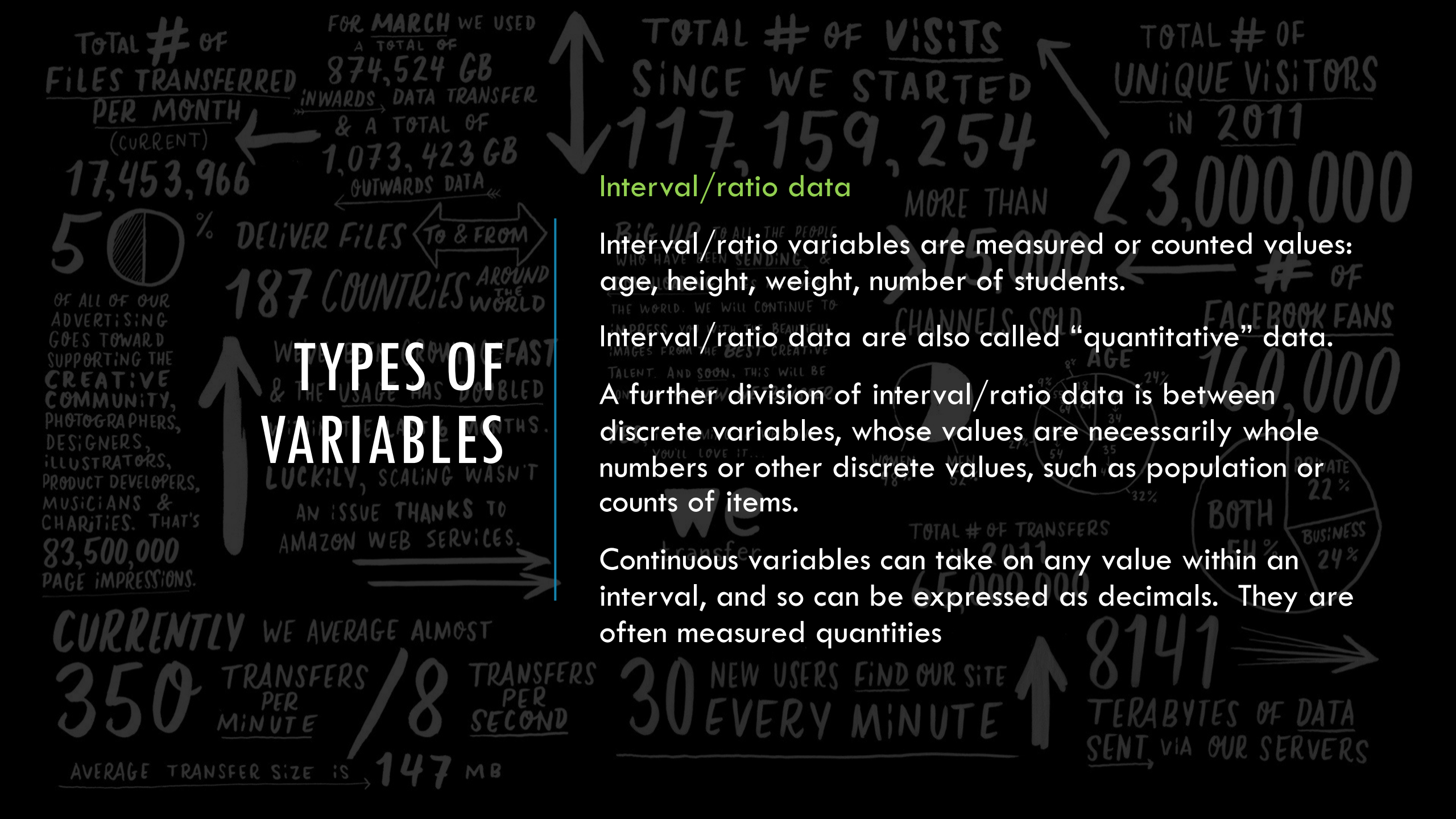
TYPES OF VARIABLES

Ordinal data

Ordinal variables can be ordered, or ranked in logical order, but the interval between levels of the variables are not necessarily known.

Subjective measurements are often ordinal variables.

One example would be having people rank four items by preference in order from one to four.



TYPES OF VARIABLES

Interval/ratio data

Interval/ratio variables are measured or counted values: age, height, weight, number of students.

Interval/ratio data are also called “quantitative” data.

A further division of interval/ratio data is between discrete variables, whose values are necessarily whole numbers or other discrete values, such as population or counts of items.

Continuous variables can take on any value within an interval, and so can be expressed as decimals. They are often measured quantities

TYPES OF VARIABLES IN R

R does not use the terms nominal, ordinal, and interval/ratio for types of variables.

In R, nominal variables can be coded as variables with factor or character classes.

Interval/ratio data can be coded as variables with numeric or integer classes.

An L used with values to tell R to store the data as an integer class.

```
BugCount.int = c(1L, 2L, 3L, 4L, 5L)

class(BugCount.int)

[1] "integer"
```