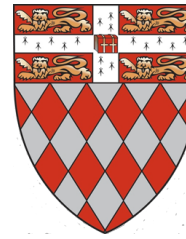


MRC

Laboratory of  
Molecular Biology



# Introduction to R

Alexey Morgunov

Fitzwilliam College & MRC-LMB, Cambridge

Previous experience with R?

Start here:

[github.com/alexeymorgunov/Rcourse](https://github.com/alexeymorgunov/Rcourse)

# What is R?

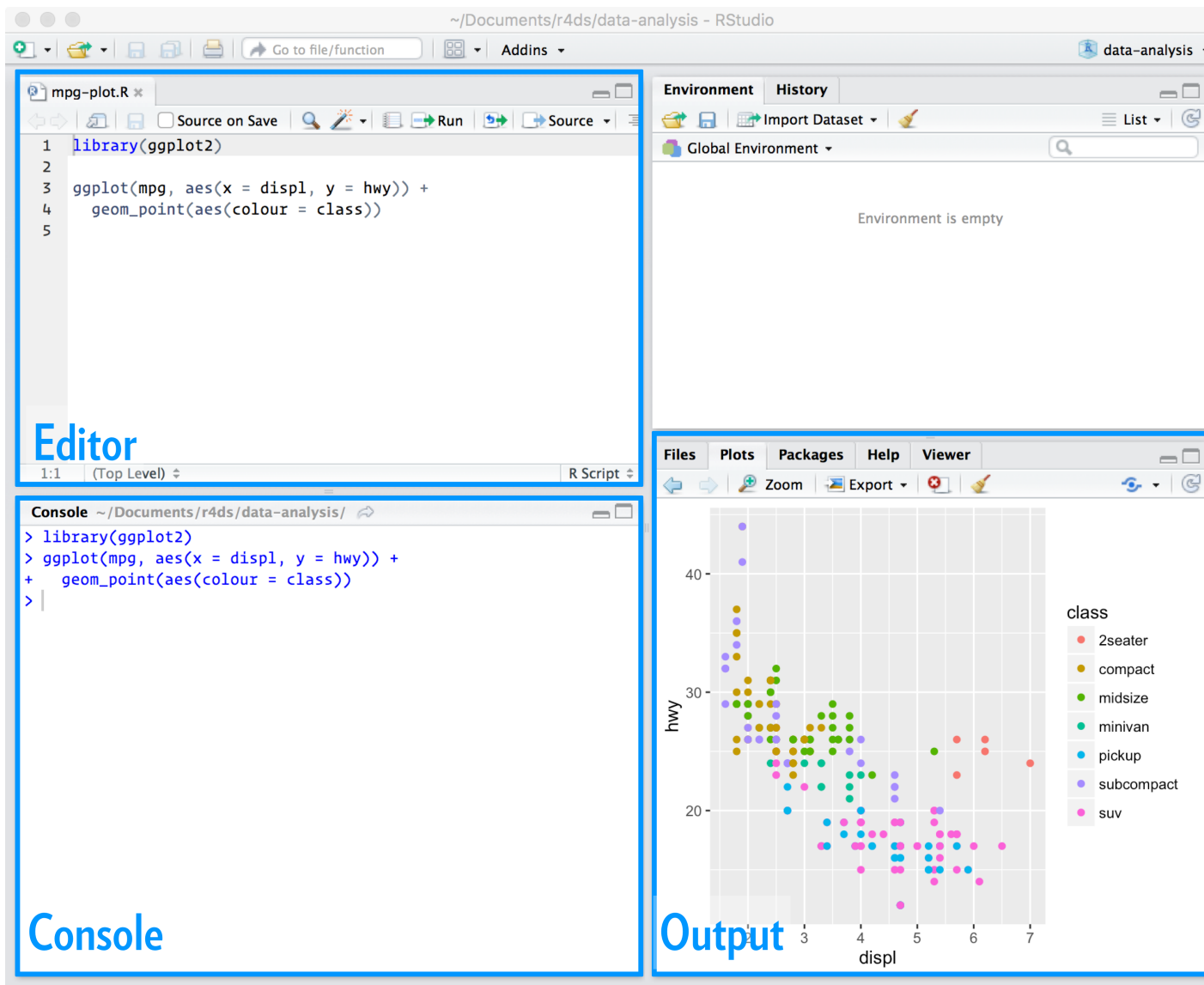
- 12<sup>th</sup> most popular programming language (TIOBE index, Jan 2019)
- Statistical computing, data analysis and graphics
- Interpreted language with a command line interface
- Several IDEs, e.g. RStudio
- Supports matrix arithmetic and data frames (c.f. tables in a relational database)
- Many packages available (CRAN, Bioconductor)
- tidyverse!



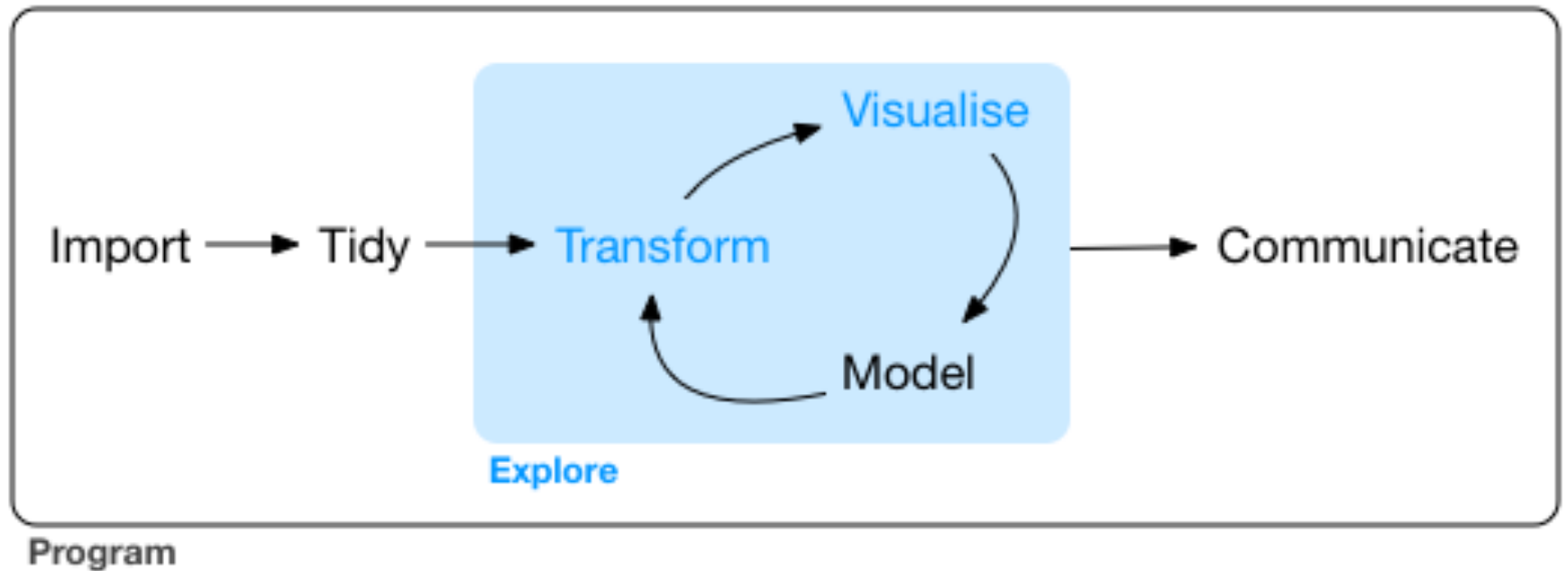
# Useful R links

- [r-project.org](https://r-project.org) (project homepage)
- [cran.r-project.org](https://cran.r-project.org) (download R)
- [rstudio.com](https://rstudio.com) (download RStudio)
- [bioconductor.org](https://bioconductor.org) (bioinformatics packages)
- [tidyverse.org](https://tidyverse.org) (more about tidyverse packages)





# Data Science



# Un-tidy data

```
> table2
```

```
# A tibble: 12 x 4
```

	country	year	type	count
	<chr>	<int>	<chr>	<int>
1	Afghanistan	1999	cases	745
2	Afghanistan	1999	population	19987071
3	Afghanistan	2000	cases	2666
4	Afghanistan	2000	population	20595360
5	Brazil	1999	cases	37737
6	Brazil	1999	population	172006362
7	Brazil	2000	cases	80488
8	Brazil	2000	population	174504898
9	China	1999	cases	212258
10	China	1999	population	1272915272
11	China	2000	cases	213766
12	China	2000	population	1280428583

```
> |
```

```
> table4a
```

```
# A tibble: 3 x 3
```

	country	`1999`	`2000`
*	<chr>	<int>	<int>
1	Afghanistan	745	2666
2	Brazil	37737	80488
3	China	212258	213766

```
> table4b
```

```
# A tibble: 3 x 3
```

	country	`1999`	`2000`
*	<chr>	<int>	<int>
1	Afghanistan	19987071	20595360
2	Brazil	172006362	174504898
3	China	1272915272	1280428583

```
> |
```

# Tidy data

```
> table1
# A tibble: 6 x 4
  country    year cases population
  <chr>    <int> <int>    <int>
1 Afghanistan 1999     745 19987071
2 Afghanistan 2000    2666 20595360
3 Brazil      1999   37737 172006362
4 Brazil      2000   80488 174504898
5 China       1999  212258 1272915272
6 China       2000  213766 1280428583
> |
```

- Each variable must have its own column
- Each observation must have its own row
- Each value must have its own cell

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observations

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

values



# Start here

[github.com/alexeymorgunov/Rcourse](https://github.com/alexeymorgunov/Rcourse)

# Thank you!

Have any questions, comments?

Email me: [asm63@cam.ac.uk](mailto:asm63@cam.ac.uk)