

COURSE OVERVIEW

(DSC10 Data Science Tools and Methods)

MARCUS BIRKENKRAHE

Created: 2021-06-16 Mi 21:08

TABLE OF CONTENTS

- What're you going to learn today?
- Who am I?
 - Science
 - Industry
 - Teaching
 - Pleasure
- What are your expectations?
- Which topics will we cover?
 - Introduction to data science
 - Introduction to R programming
 - Visualization using R
 - FasteR approach
 - Schedule
- How will we do it?
 - Classroom sessions
 - Lecture scripts with exercises
(GitHub)
 - Reading assignments
 - Video lectures (YouTube)
 - Online assignments (DataCamp)
 - Team EDA project
 - Agile project management
 - Tests and final exam
 - Podcasts and feeds
 - Summary of course activities
- What do you have to do to pass?

- DataCamp assignments (> 50%)
- Team project (> 50%)
 - What is a team project?
 - Do you have project examples?
 - Can I do a project as an absolute beginner?
- Final exam (> 50%)
- What's next?
 - Your challenges
- Any questions?

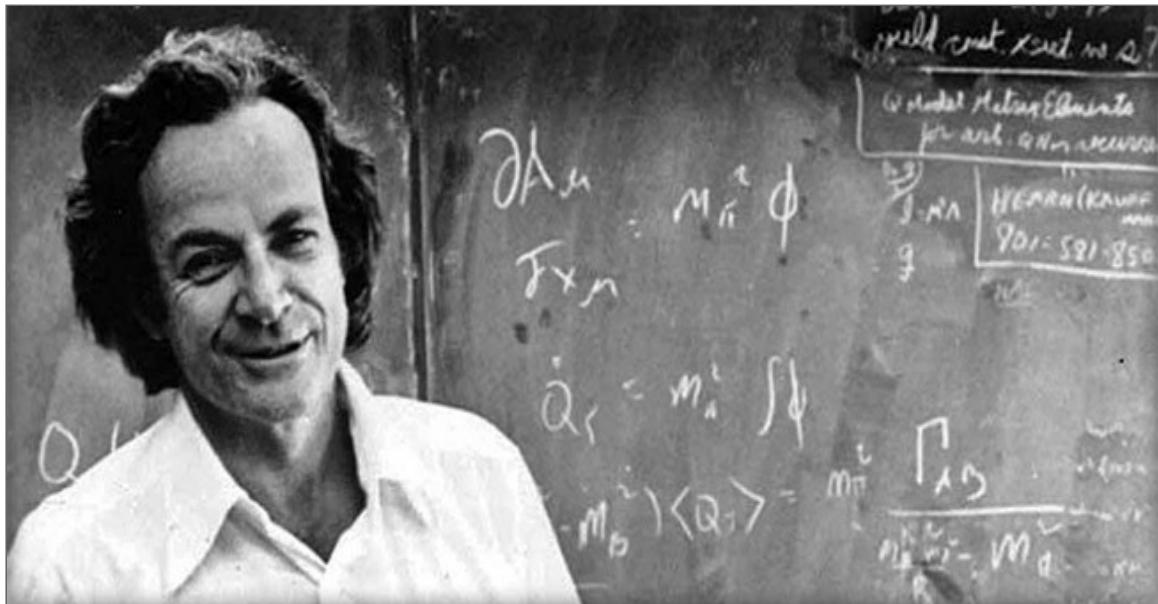
WHAT'RE YOU GOING TO LEARN TODAY?

- Who is your lecturer?
- Who are you and what do you want?
- Which topics will we cover?
- How will we do it?
- What do you have to do to pass?
- What's next?

WHO AM I?



SCIENCE



- Development of WWW
- PhD theoretical particle physics
- 60 research publications
- Assoc. Ed. Int. J. of Data Science
- Ed. Board Int. J. of Big Data Mgmt.
- Scientific member [d-cube@Berlin](#)

INDUSTRY



- Executive at Accenture & Shell
- Coach and consultant
- Certified psychotherapist
- Startup mentor

TEACHING



- Business informatics @HWR Berlin
- Visiting professor of data science @Lyon
- Adviser for CPU @LA
- Internship supervision

PLEASURE



- Playing: Assassin's Creed Valhalla
- Reading: Bernard Cornwell, The Burning Land
- Watching: Person of Interest

WHAT ARE YOUR EXPECTATIONS?

- What do you want to learn here?
- What would you like to avoid?
- What did you take away from another course?
- What did you really not like in another course?

WHICH TOPICS WILL WE COVER?



INTRODUCTION TO DATA SCIENCE

6

Data visualization will go mainstream

In 2020, data visualizations helped us make sense of an increasingly complex world. Creating, critically understanding, and evaluating data visualizations will become a foundational skill for every citizen.

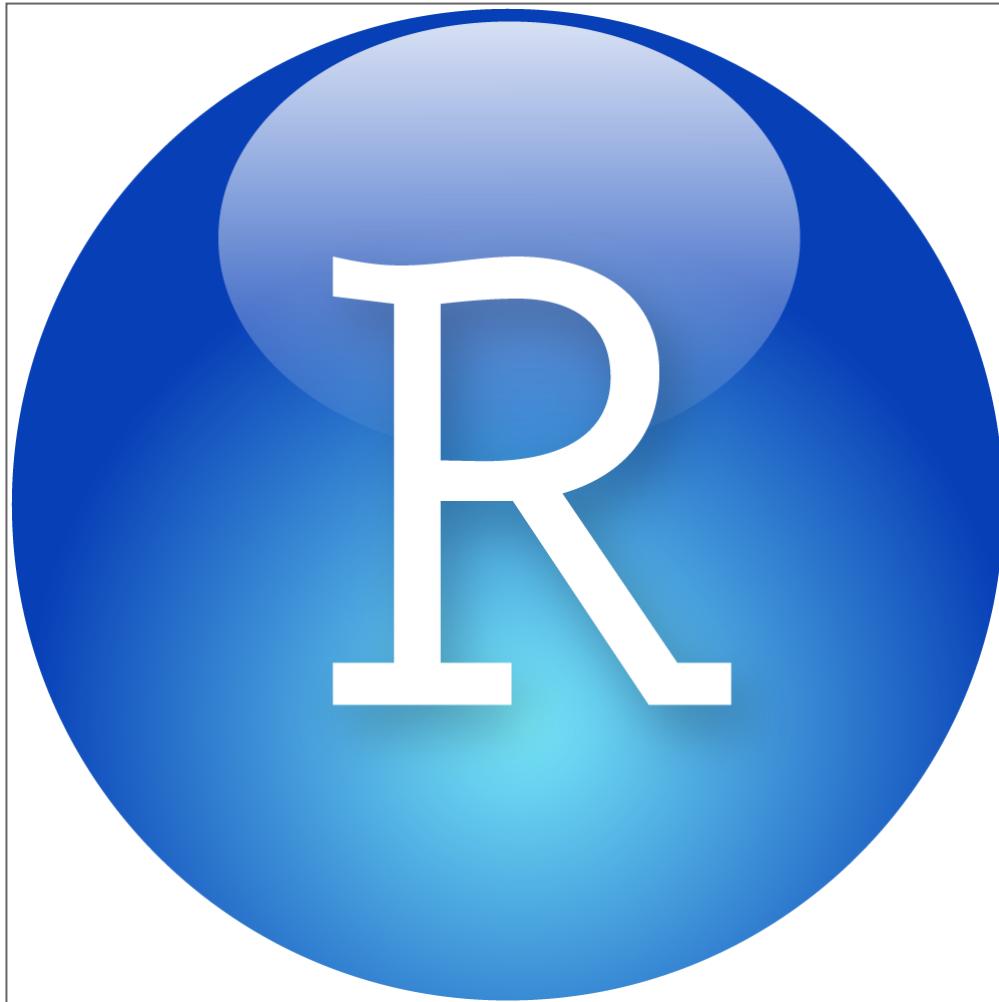
8

Data skills will cross over to every discipline

From primary to tertiary education, data literacy will become foundational for every discipline.

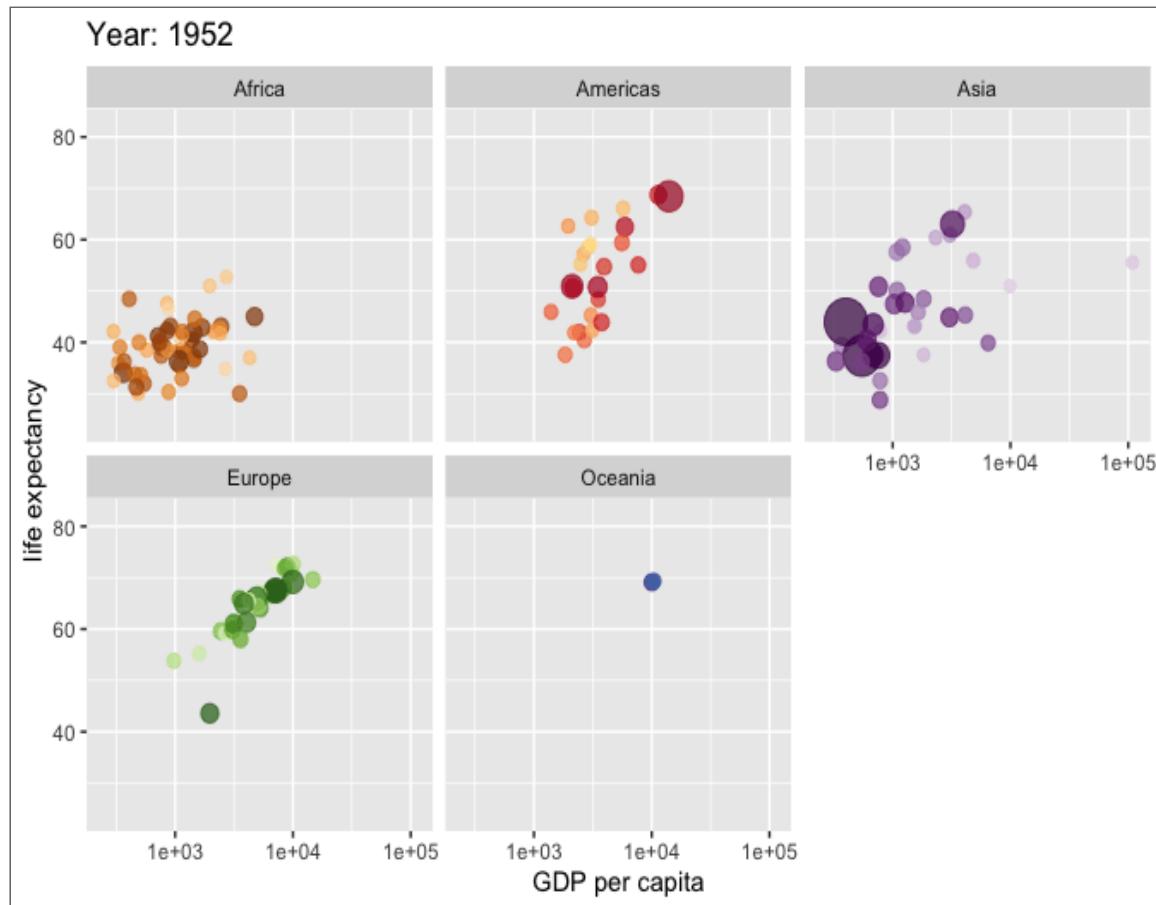
Source: [datacamp.com](https://www.datacamp.com)

INTRODUCTION TO R PROGRAMMING



Source: [RStudio](#)

VISUALIZATION USING R



Source: [Thomas Lin Pedersen](#)

FASTER APPROACH



- Focus on data exploration (EDA)
- Stay close to base R
- Use real data sets
- Compute interactively
- Prepare for DSC201 (ML)

SCHEDULE

No	Date	Lectures ¹	DataCamp ²	Tests/Quiz ³	fasterR ⁴
1	17-Aug	Overview			
2	19-Aug	On the R Shell			
3	24-Aug	Vectors in R			
4	26-Aug	Data frames in R			
5	31-Aug	Factors in R			
6	2-Sep	apply functions			
7	7-Sep	Cleaning data			
8	9-Sep	Lists in R			
9	14-Sep	Nile exploration			
10	16-Sep	Visualization			
11	21-Sep	Base R graphics			
12	23-Sep	Writing functions			
13	28-Sep	Iteration I			
14	30-Sep	Fibonacci series			
15	5-Oct	Literate Programming			
16	7-Oct	Conditions			
17	12-Oct	EDA example I			
18	14-Oct	Linear regression I			
19	19-Oct	Object-orientation			
20	21-Oct	EDA example II			
21	26-Oct	Packages			
22	28-Oct	Grammar of Graphics			
23	2-Nov	Functional Programming			
24	4-Nov	Text mining I			
25	9-Nov	Text mining II			
26	11-Nov	Linear regression II			
27	16-Nov	Dates and times			
28	18-Nov	Coding style			
29	23-Nov	Logistic regression			
30	25-Nov	Version control			
31	30-Nov	Iteration II			
32	2-Dec	Summary and outlook			
33	TBD				
VISUALIZATION					
APPLICATIONS					
				Final exam	

HOW WILL WE DO IT?



CLASSROOM SESSIONS

Data Science 101

"Ideation" / R
• Why are you here?
Excitement
 e^{-rt}
Fear

- add another axis
- variables?
- Qualitative vs. quantitative

ΔE = loss of excitement
("Confidence development")
loss of fear

14 Oct
 $\Delta F \rightarrow$ Time series

I was unsure as you could have maybe used a third-party program ;)

:Just out of interest:

Who commented / upvoted my

Recorded with BigBlueButton.

19:23

LECTURE SCRIPTS WITH EXERCISES ([GITHUB](#))

birkenkrahe / ds101

Code Issues Pull requests Discussions Actions Projects Wiki ...

master Go to file Add file Code

About

Course materials (except video lectures) for the introduction to data science with R, winter 2020. Some of the R material is inspired by, or based on sections in books by Davies (Book of R, 2016), Cotton (Learning R, 2013), and Irizarry (Intro to DS, 2019), and also on Matloff's tutorial "FasterR" (available on GitHub).

1000 data science bookmarks

File	Description	Last Commit
1_overView.zip	Add files via upload	4 months ago
2_introduction.zip	Add files via upload	5 months ago
3_arithmetic.zip	Add files via upload	3 months ago
4_vectors.zip	Add files via upload	3 months ago
6_lipRog.zip	Add files via upload	2 months ago
7_plotting.zip	Add files via upload	last month
9_ggplot.zip	Add files via upload	2 days ago
LICENSE	Initial commit	5 months ago
README.md	Add files via upload	5 months ago
description.pdf	Add files via upload	5 months ago
ds_bookmarks.md	Add files via upload	2 days ago

r datascience
introduction-to-r
Readme
GPL-3.0 License

READING ASSIGNMENTS



Norman Matloff (2020):
fasteR: Fast Lane to Learning R!

VIDEO LECTURES ([YOUTUBE](#))



Ease-of-use, Fun factor [PLAY ALL](#)

Vectors in R (part 1)

9 videos • 127 views • Last updated on Nov 23, 2020

Unlisted ▾

✖️ ➡️ ⋮

Everything is an object, and vectors are among the most important objects in R. In this video series, we cover creating, sorting, and measuring vectors. It may sound all a little technical, and it is, but it is an important building block towards our end game of data-driven storytelling. Once we understand objects and vectors, many other concepts will fall in our lap! - Note: try to complete the exercises at the end BEFORE looking at the solution video. The exercises are contained in the description of the solution video!

= SORT

**1 Everything is an object**
Marcus Birkenrahe
WATCHED 4:34

**2 Assigning Objects**
Marcus Birkenrahe
WATCHED 10:06

**3 Who needs vectors anyway?**
Marcus Birkenrahe
WATCHED 6:41

**4 Creating vectors**
Marcus Birkenrahe
WATCHED 5:12

**5 Down the river Nile**
Marcus Birkenrahe
WATCHED 4:34

**6 Plotting histograms**
Marcus Birkenrahe
WATCHED 4:11

ONLINE ASSIGNMENTS (DATACAMP)

The screenshot shows the 'Team Assignments' page for the 'BSEL Berlin Analytics 21' team. A prominent orange callout box with the text 'Don't miss the deadline' points to the 'ACTIVE' tab in the filter bar. The page lists four assignments:

Title	Assignees	Status	Due By	C	A	CR	Details
Data Science for Everyone Introduction to Data Science Chapter	Team	Active	Mar 10, 10:00 CET	0	0	0%	View
Data Science for Everyone Data Collection and Storage Chapter	Team	Active	Mar 15, 10:00 CET	0	0	0%	View
Data Science for Everyone Preparation, Exploration, and Visualization Chapter	Team	Active	Mar 22, 10:00 CET	0	0	0%	View
Data Science for Everyone Experimentation and Prediction Chapter	Team	Active	Mar 29, 10:00 CEST	0	0	0%	View

- Register at DataCamp

TEAM EDA PROJECT

Kaggle Notebook (Left): Pima Indians Diabetes Database Analysis

Sample R project on Kaggle

corrrplot(correlat)

RStudio Session (Right):

```

is.edu/FasterR/data/Pima.csv",
header=TRUE)

27
28
29 The data set is in a CSV ("comma-separated-values") file. Here we
read it using the *read.csv* function.
The
30 file header, if it exists, is the first
line in the file. If the file does not
have a header, you can set header to
31 *FALSE* and add one later using the names
function.
32
33 ##### Look at the *pima* dataset
34 Now, let's look at the data frame: in the
following code chunk, substitute the
correct function for *...* so that you
see the first few lines of the dataset
*pima*.
35 ````{r Print first few lines of pima}
36 head(pima)
37 ````
```

	pregnant	glucose	diastolic
1	6	148	72
2	1	85	66
3	8	183	64
4	1	89	66
5	0	137	40
6	5	116	74

6 rows | 4 of 9 columns

38 Execute the chunk to check that you did the right thing. The output will appear right below it. The data frame should have nine (numeric) variables and 768 observations.

34:91 Look at the "pima" dataset : R Markdown

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home > OneDrive > R > workspace

Name Size

- .Rhistory 1.9 KB
- Pima_solutions.docx 188.6 KB
- Pima_solutions.Rmd 6 KB
- Pima_solutions.html 897.9 KB
- workspace.Rproj 205 B
- Pima_solutions.aux 662 B
- Pima_solutions.log 20.4 KB
- Pima_solutions.tex 15.6 KB
- notebook_2.knit.md 476 B
- notebook_2.nb.html 892.3 KB
- Pima_problems.Rmd 5.4 KB
- notebook_2.Rmd 426 B

Console

AGILE PROJECT MANAGEMENT



TESTS AND FINAL EXAM

DS101 Entry Quiz 📝

Challenge 🏆 Ends in 5 days

Start date: Feb 22 2021, 5:06 pm

End date: Mar 3 2021, 10:00 am

Hosted by birkenkrahe

Challenges are available throughout the course

Summary Players (10) Questions (20)

All (20) Difficult questions (6)

Question ▾ Type

Question	Type	Progress
1 Which of these are good problems for ...	Quiz	0%
2 Which of these are skills that data scie...	Quiz	0%
3 Which of these things have to do with...	Quiz	60%
4 Which part of the data science proces...	Poll	?
5 What is "R" (in data science)?	Quiz	0%
6 According to the TIOBE ranking, R is t...	True or false	80%

Quiz questions will be recycled in the final exam

PODCASTS AND FEEDS

Google Podcasts

Search for podcasts

Build a Career in Data Science

Sep 10, 2020

Chapter 1: What is Data Science?

48 min left

Podcasts on:
**data science careers,
data-driven storytelling,
visualization,
applications**

People also listened to

- Data Futurology - Leadership... Felipe Flores
- Data Stories Enrico Bertini and Moritz Stefaner
- #144 Machine Learning: Getting th... We are joined by Alexey Grigorev for an episode that will be very useful for anyone wanting to identify the skills needed to g...
- 163 | svelte.js for web-based... 163 | svelte.js for web-based dataviz with Amelia Wattenberger

More episodes from Build a Career in Data Science

HWR Berlin

Feed Actions

Analytics Vidhya
<https://www.analyticsvidhya.com/feed/>
Learn everything about Analytics

BBC Technology News
<http://newsrss.bbc.co.uk/rss/bbcnewstechnology/rss.xml>
BBC News - Technology

R-Bloggers
<https://www.business-science.com/feed/>

CNET How-To
<http://feed.cnet.com/feed/how-to>
CNET editors and users share the top tech 'how to' tips and tricks with advice for getting the most out of all your gadgets.

CNET Tech News
<http://feed.cnet.com/feed/news>
CNET news editors and reporters provide top technology news, with investigative reporting and in-depth coverage of tech issues and events.

Data is beautiful
<https://www.reddit.com/r/dataisbeautiful/.rss>
A place to share and discuss visual representations of data: Graphs, charts, maps, etc.

SUMMARY OF COURSE ACTIVITIES

- Weekly classroom meetings
- Lecture scripts (GitHub)
- Reading assignments (Online)
- Video lectures (YouTube)
- Online assignments (DataCamp)
- Team EDA projects (Sprints)
- Tests and final exam
- Podcasts and feeds

WHAT DO YOU HAVE TO DO TO PASS?



DATA CAMP ASSIGNMENTS (> 50%)

1 Introduction to Data Science

100%

We'll start the course by defining what data science is. We'll cover the data science workflow and how data science is applied to real-world problems. We'll finish the chapter by learning about different roles within the data science field.

▶ What is data science?	Videos	✓ 50 xp
◀ Customer segmentation workflow		✓ 100 xp
◀ Building a customer service chatbot		✓ 100 xp
▶ Applications of data science		
◀ Assigning data science project		
☰ Investment research		
▶ Data science roles and tools		✓ 50 xp
☰ Editing a job post	Mixed exercises	✓ 50 xp
◀ Matching skills to jobs		✓ 100 xp
◀ Classifying data tasks		✓ 100 xp

Videos

Ca. 30 min per chapter

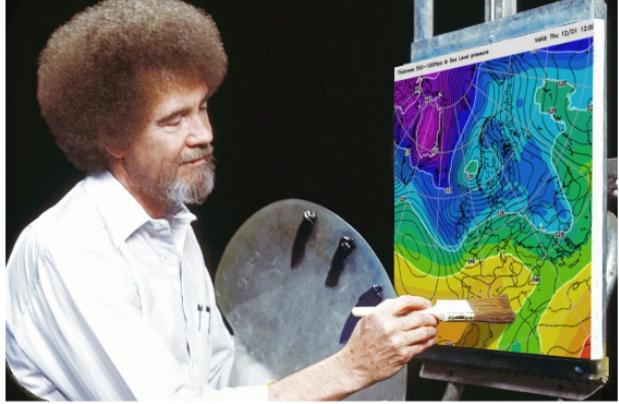
Mixed exercises

Complete 12 of 24 assignments

TEAM PROJECT (> 50%)

 **Election 2016 Trump-Clinton Spatial Visualization**
R notebook using data from [multiple data sources](#) · 1,275 views · 7mo ago
data visualization, exploratory data analysis, politics, +1 more

Kaggle sample R project



Version 10 of 10

[Notebook](#)

[Table Of Contents](#)

- [1. Packages](#)
- [2. How To Create A Map Using Ggplot2](#)
- [3. Election Data & First Map](#)
- [4. Trump Vs Clinton](#)
- [5. Statebins](#)
- [6. References](#)

[Input \(2\)](#)

[Output](#)

[Execution Info](#)

[Log](#)

[Comments \(12\)](#)

Table of Contents

- [1. Packages](#)
- [2. How to create a map using ggplot2](#)
- [3. Election Data & First Map](#)
- [4. Trump vs Clinton](#)
- [5. Statebins](#)
- [6. References](#)

Present on Nov 30 or Dec 2

WHAT IS A TEAM PROJECT?

- Description of the dataset
- Introduction of the problem statement
- Description of the methods used
- Visualization of the data (plots!)
- Analysis of the plots
- Limitations of own analysis
- References

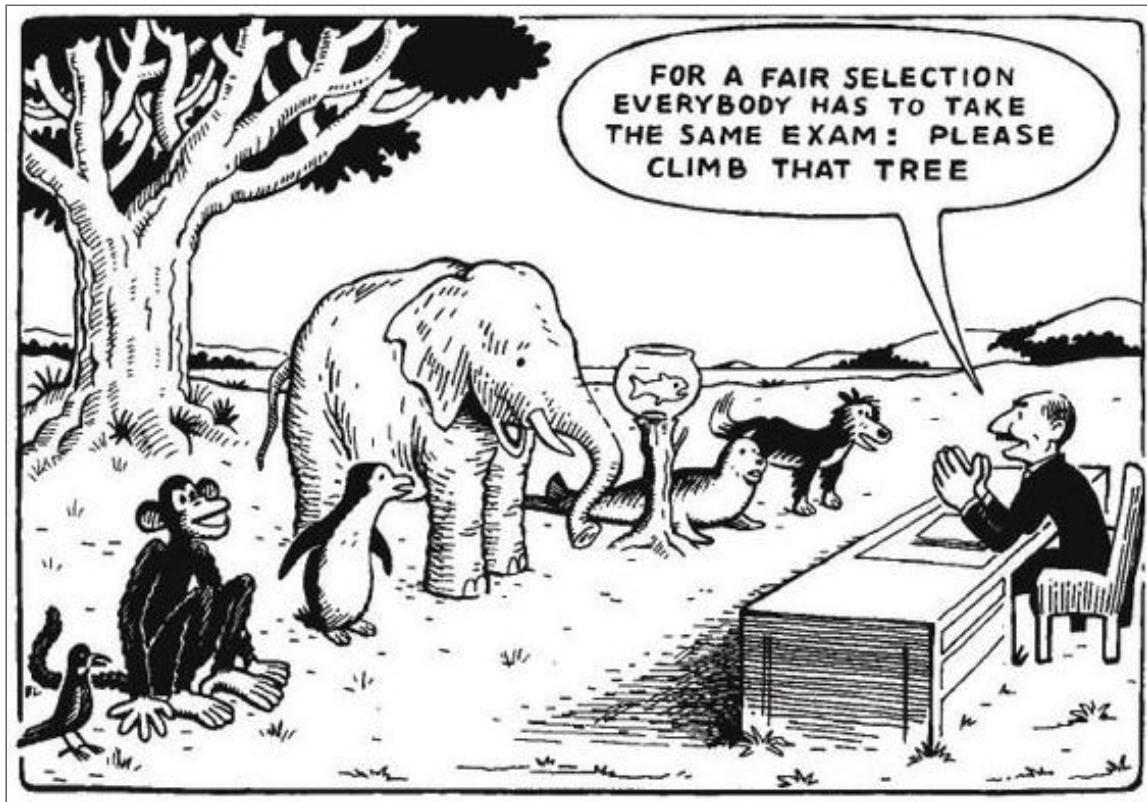
DO YOU HAVE PROJECT EXAMPLES?

- Examples on Kaggle ([example](#))
- Examples on data science blogs ([example](#))
- Translate from Python to R ([example](#))
- Extend someone else's EDA ([example](#))
- Document an R package ([example](#))
- Use your own data ([example](#))

CAN I DO A PROJECT AS AN ABSOLUTE BEGINNER?

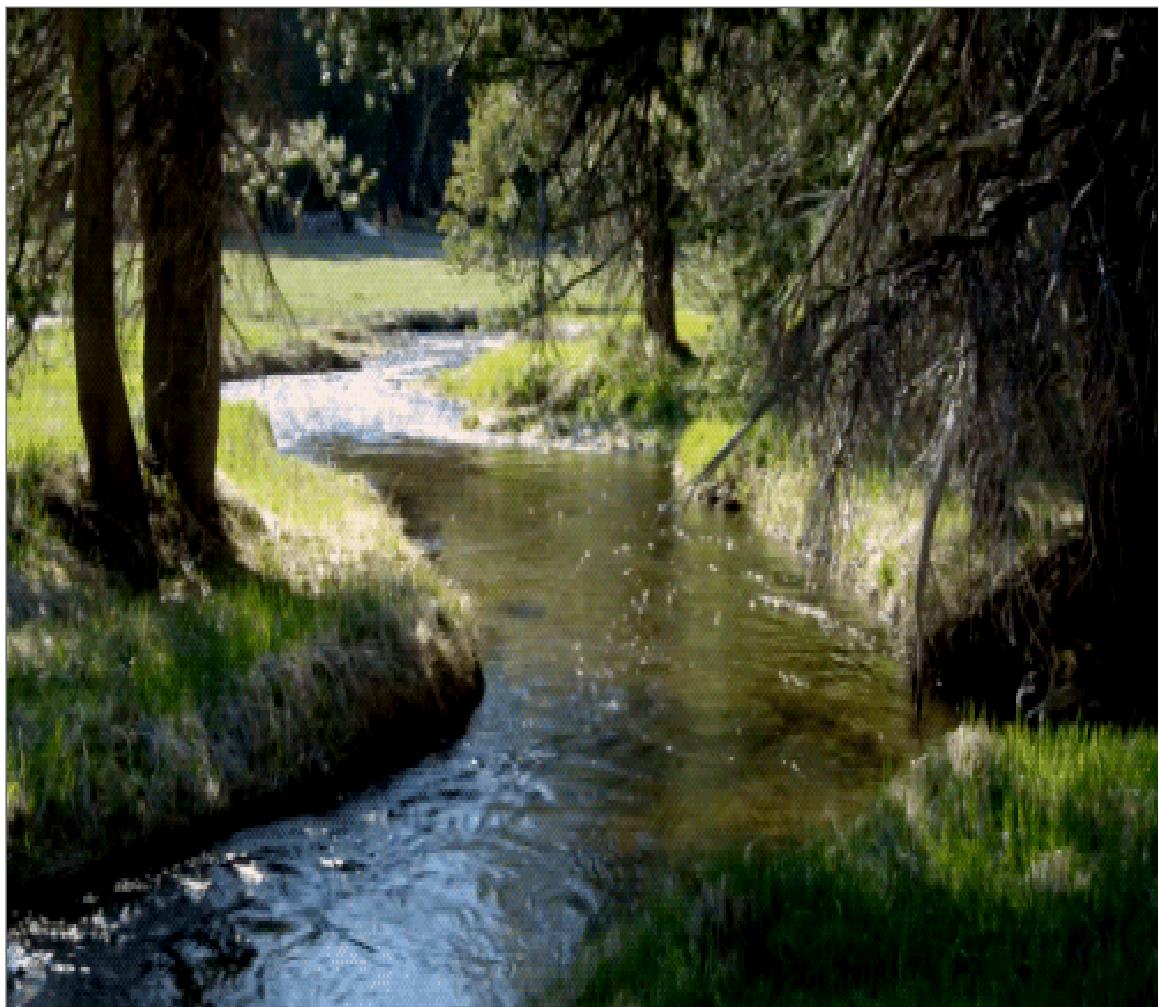
- Keep It Simply Scientific (IMRaD)
- Look at examples (e.g. **bookmarks**)
- Create data set (e.g. productivity)
- Researchers are beginners

FINAL EXAM (> 50%)



Final exam: date TBD

WHAT'S NEXT?



YOUR CHALLENGES

What?	When?
<u>Register with DataCamp</u>	Today
<u>Complete test challenge</u> *	Aug 19
<u>Complete DataCamp assignment</u> *	Aug 24
Set up <u>project</u> (2-3 ppl)*	Sep 2
Check FAQs	n.d.
Ask questions (class/ <u>forum</u>)	n.d.

**) do this every week until December*

ANY QUESTIONS?



A PDF of this presentation is available.