

# Course overview

## Data Science Tools and Methods

Marcus Birkenkrahe

August 17, 2021

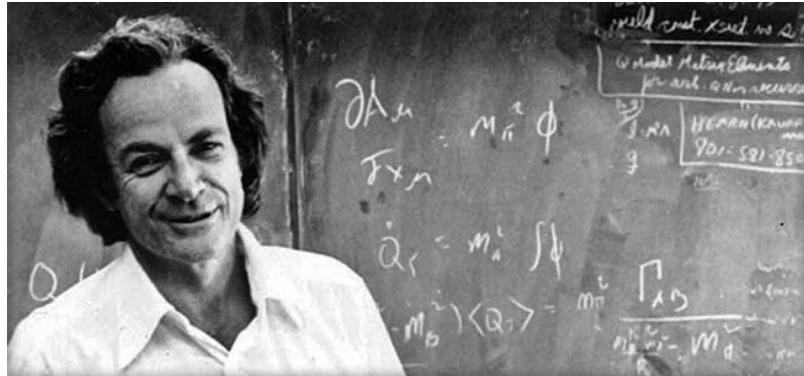
### What're you going to learn today?

- 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 (2020)
- Reading: Waugh, Sword of Honour (1952-1961)
- Watching: The Middle (2009-2018)

## **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?**

`./img/lavaflow.gif`

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.

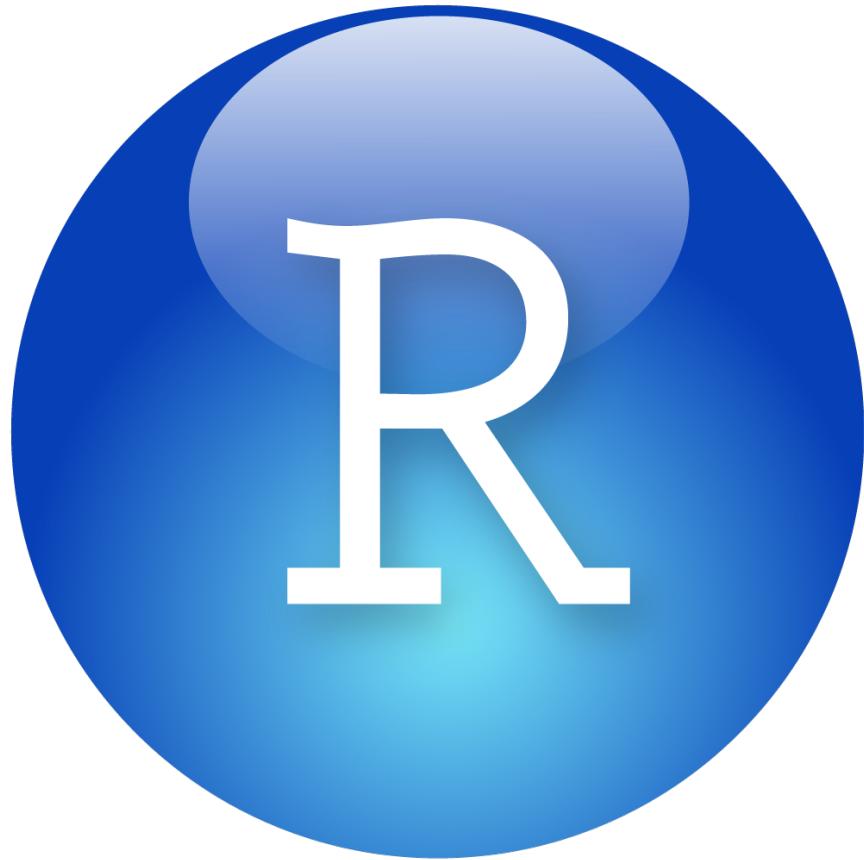
8

## **Data skills will cross over to every discipline**

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

Source: datacamp.com

## Introduction to R programming



Source: RStudio

## Visualization using R

`./img/gapminder.gif`

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)

Image source: unsplash

## Schedule (see Syllabus)

No	Date	Lectures <sup>1</sup>	DataCamp <sup>2</sup>	Tests/Quiz <sup>3</sup>	fasterR <sup>4</sup>
1	17-Aug	Overview			Overview and Getting Started
2	19-Aug	On the R Shell		Test 1	Installing R and first steps
3	24-Aug	Vectors in R		Test 2	More on Vectors
4	26-Aug	Data frames in R		Test 3	On to Data Frames!
5	31-Aug	Factors in R		Test 4	R Factor Class
6	2-Sep	apply functions			The tapply Function
7	7-Sep	Cleaning data			Data Cleaning
8	9-Sep	Lists in R			R List Class
9	14-Sep	Nile exploration			Another Look at the Nile Data
10	16-Sep	Visualization	<b>BASICS</b>	Test 7	Introduction to Base R Graphics
11	21-Sep	Base R graphics		Test 8	More on Base Graphics
12	23-Sep	Writing functions		Test 9	Writing Your Own Functions
13	28-Sep	Iteration I		Test 10	for Loops
14	30-Sep	Fibonacci series		Test 11	Functions with Blocks
15	5-Oct	Literate Programming		Test 12	
16	7-Oct	Conditions		Test 13	IDEs: Text Editing, Saving, Executing
17	12-Oct	EDA example I		Test 14	If, Else, Ifelse: Conditions
18	14-Oct	Linear regression		Test 15	Do Pro Athletes Keep Fit?
19	19-Oct	Object-orientat		Test 16	Regression Analysis, I
20	21-Oct	EDA example II			issues
21	26-Oct	Packages		Test 19	Baseball Player Analysis (cont'd)
22	28-Oct	Grammar of Graphics		Test 20	R Packages, CRAN, Etc.
23	2-Nov	Functional Programming	<b>VISUALIZATION</b>	Test 21	A First Look at ggplot2
24	4-Nov	Text mining I		Test 22	Should You Use Functional Programmin
25	9-Nov	Text mining II		Test 23	? Simple Text Processing I
26	11-Nov	Linear regression II		Test 24	Simple Text Processing II
27	16-Nov	Dates and times			Regression Analysis, II
28	18-Nov	Coding style			ng with the R Date Class
29	23-Nov	Logistic regression			
30	25-Nov	Version control	<b>APPLICATIONS</b>		Tips on R Coding Style and Strategy
31	30-Nov	Iteration II			The Logistic Model
32	2-Dec	Summary and outlook			Files and Directories
33	TBD				R while loops
					Summary and Outlook
					Final exam

## How will we do it?

./img/deer.gif

## Classroom sessions



## Lecture scripts with exercises (GitHub)

The screenshot shows a GitHub repository page for 'dsc101'. The URL in the browser bar is highlighted with a red box and labeled 'URL'. The repository owner is 'birkenkrahe'. Key interface elements highlighted with colored boxes are: 'Raise issues' (orange), 'Watch' (pink), 'Issues' (orange), 'Discussions' (green), and 'Discuss' (green). The repository description is 'Repository for DSC 101 - Data science methods and tools'. The file list includes '1\_overview' (highlighted with a blue box), 'LICENSE', and 'README.md'. The '1\_overview' file has a large blue box over it with the word 'Lectures' written in white. The README.md content includes a section titled 'dsc101' with a bulleted list:

- Repository for DSC 101 - Data science methods and tools.
- Contains Emacs Org-mode files (rendered as Markdown), Wiki, Discussion forum.
- First offered: fall 2021.

## Reading suggestions



- Matloff: fasteR: Fast Lane to Learning R! (2021)
- Matloff: The Art of R Programming (2011)
- Matloff TARP available for free from the Internet Archive
- Davies, The Book of R, NoStarch Press (2016)
- Irizarry, Introduction to Data Science (2020)

## Video lectures (YouTube)

The screenshot shows a YouTube channel page for 'Vectors in R (part 1)'. The channel has 9 videos and 127 views, last updated on Nov 23, 2020. It is unlisted. The channel description states: '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!' The channel is sorted by 'WATCHED' count.

Rank	Title	Length	Uploader
1	1 Everything is an object	4:34	Marcus Birkenrahe
2	2 Assigning Objects	10:06	Marcus Birkenrahe
3	3 Who needs vectors anyway?	6:41	Marcus Birkenrahe
4	4 Creating vectors	5:12	Marcus Birkenrahe
5	5 Down the river Nile	4:34	Marcus Birkenrahe
6	6 Plotting histograms	4:11	Marcus Birkenrahe

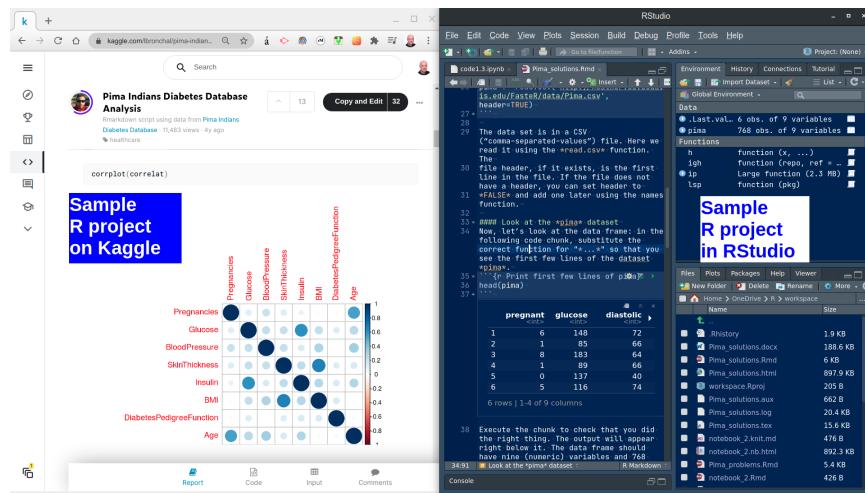
## Online assignments (DataCamp)

The screenshot shows the DataCamp interface for the team 'BSEL Berlin Analytics 21'. The 'Team Assignments' section is displayed, with the 'ACTIVE' tab selected. An orange callout box highlights the text 'Don't miss the deadline'.

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%	<a href="#">View</a>
Data Science for Everyone Data Collection and Storage Chapter	Team	Active	Mar 15, 10:00 CET	0	0	0%	<a href="#">View</a>
Data Science for Everyone Preparation, Exploration, and Visualization Chapter	Team	Active	Mar 22, 10:00 CET	0	0	0%	<a href="#">View</a>
Data Science for Everyone Experimentation and Prediction	Team	Active	Mar 29, 10:00 CEST	0	0	0%	<a href="#">View</a>

- Register at DataCamp today!

## Team EDA project



## 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**

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
- #144 Machine Learning: Getting th... We are joined by Anney Grigoriev for an episode that will be very useful for anyone wanting to identify the skills needed to g...
- Data Stories Enrico Bertini and Moritz Stefaner
- 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 BBC News - Technology rss.xml
- R-Bloggers https://www.business-scientist.com/ Blog feeds on technology, applications, current events
- CNET How-To http://feed.cnet.com/feed/howto 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/datavisualization/rss A place to share and discuss visual representations of data: Graphs, charts, maps, etc.

Shared via GitHub / Schoology

### **Summary of course activities**

- Twice 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?**

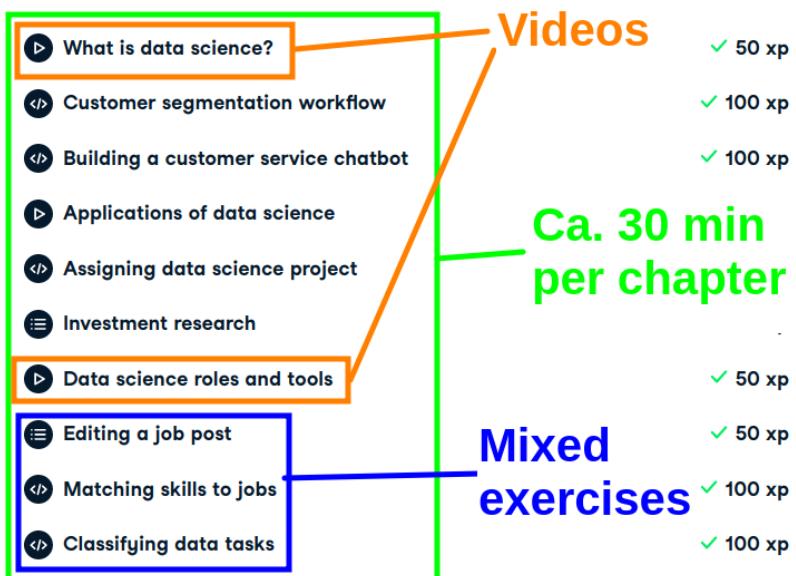
`./img/oceanrock.gif`

## DataCamp 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.



Complete at least 8 of 15 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

40 Copy and Edit 7 ...

# 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

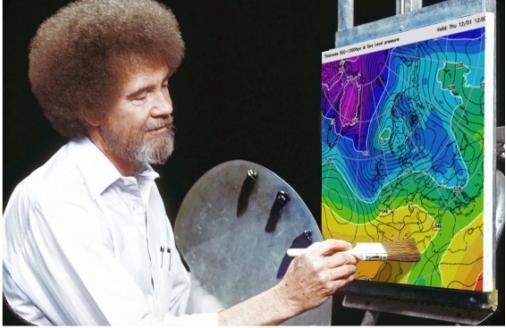
Execution Info

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

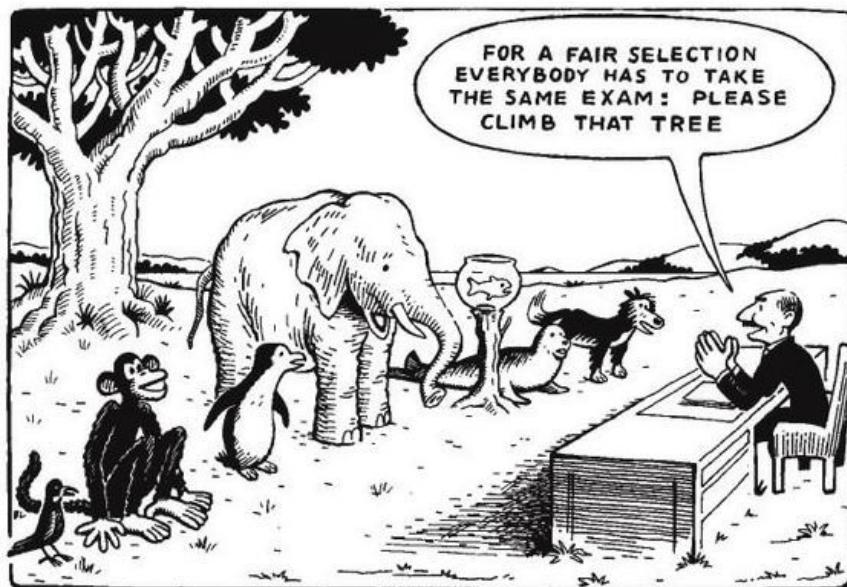
### Do you have project examples?

- 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. in my bookmarks)
- Create data set (e.g. your productivity)
- Researchers are beginners

### Final exam (> 50%)



Final exam: date TBD

## What's next?

./img/river.gif

### In the course

- Intro to Data science (Lecture)
- Intro to DataCamp (Practice)
- Intro to GitHub (Productivity)
- Intro to R (Language)

### Your challenges

What?	When?
<b>Register at DataCamp</b>	Today
Register at GitHub	Today
<b>Complete test challenge</b>	Aug 24
<b>Complete DataCamp assignment</b>	Aug 24
<b>Set up team project (2-3 ppl)</b>	Sep 2
Check FAQs x 2 in GitHub	n.d.
Ask questions (class/GitHub)	n.d.

*\*) do this every week until December*

## Any questions?

./img/stonehenge.gif

A copy of this presentation is available.