

o#+TITLE:COURSE OVERVIEW



What is an "Advanced intro to data science"?

Data science: Journey to the "Lonely Mountain"

Glóin. Thorin and Company:

THE HOBBIT	THE COURSE
Treasure hunters	Big Data
Skilled experts	Data science training
Magic ring owners	R + Emacs + ESS + Org-mode
Fighters against evil	Statistics for good or evil
Band of brothers	Scrum projects

What will you do in this course?

- Last DSC 205 course in spring'22 was too "tidy"¹
- Mixture of DataCamp lessons and current topics

¹My view towards the "Tidyverse" is well summarized in Matloff's essay "TidyverseS-ceptic" (Matloff, 2022).



Figure 1: Map of the Lonely Mountain (Tolkien, The Hobbit)



Figure 2: Map of the Lonely Mountain (Tolkien, The Hobbit)

- Learn R programming (functions, conditions, loops, utilities)
- Improve performance (e.g. `data.table`, fast reading/writing)
- Special data science topics (e.g. NLP, ML, projects)
- Transcend R programming (e.g. command line data science)

How will you be evaluated?

- All course requirements have deadlines
- Late submissions will be penalized (loss of points)
- Final exam will be sourced by term test questions
- The project topic can come from any of the course sub-topics
- The project deliverable is a working `literate` program

WEEK	DATE	TOPICS & ASSIGNMENTS	TESTS
1	Jan 11,13	Calling functions	
2	Jan 18,20	Intermed R: Conditionals	Test 1
3	Jan 23,25,27	Intermed R: Loops	Test 2
4	Jan 30, Feb 1,3	Intermed R: Functions	Test 3
5	Feb 6,8,10	Writing functions	
6	Feb 13,15,17	Intermediate R: apply	Test 4
7	Feb 20,22,24	Intermed R: Utilities	Test 5
8	Mar 1,3	Introduction to Bag-of-Words	Test 6
9	Mar 6,8,10	Natural language processing	
10	Mar 13,15,17	Introduction to data.table	Test 7
11	Mar 27,19,31	Importing and exporting data	Test 8
12	Apr 3,5	Introduction to shell	Test 9
13	Apr 10,12,14	Downloading data on shell	
14	Apr 17,19,21	Data cleaning and munging	Test 10
15	Apr 24,26,28	Machine learning	
16	May 1, 3	Project presentations	

Figure 3: Syllabus, Canvas (lyon.instructure.com) or GitHub (github.com/birkenkrahe/ds2)

REQUIREMENT	UNITS	PPU	TOTAL	% of TOTAL
Final exam	1	100	100	20.
Home assignments	10	10	100	20.
Class assignments	10	10	100	20.
Project sprint reviews	5	20	100	20.
Multiple-choice tests	10	10	100	20.
TOTAL			500	100.

Figure 4: Source: syllabus, Canvas (lyon.instructure.com) or GitHub (github.com/birkenkrahe/ml)

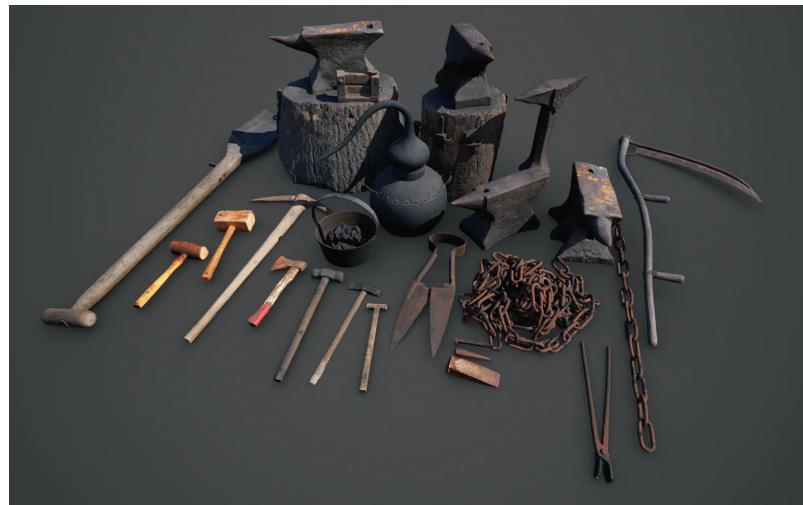


Figure 5: Unsplash, workshop

Which tools are you going to use?

- DataCamp courses (10 weekly home assignments)
- GitHub repository (all course materials except tests)
- GNU Emacs + ESS + R (literate programming environment)
- Canvas (learning management system)

How can you register at DataCamp?

 Intermediate R Conditionals and Control Flow Chapter	Team	Active	Jan 20, 13:00 CST
<hr/>			
 Intermediate R Loops Chapter	Team	Active	Jan 27, 13:00 CST

Figure 6: DataCamp assignments for January 2023

- You find the invitation link to the group for Spring 23 in Canvas.
- You will automatically be subscribed to the DSC 205 team
- If you are in more than one course, I will add you later manually
- These accounts will be valid until July 8, 2023 only

When is the first assignment due?



- The first DataCamp home assignment is due on Friday, January 20. For late submissions, you lose 1 point per day (out of 10).
- The first in-class assignment is due on Friday, January 20. For late submissions, you lose 1 point per day (out of 10 possible points)
- We'll write the first weekly multiple-choice test on January 20, covering all of the material that was covered until then.

What should we do as a project?

- For your term project, **pick an R package** to explore and explain
- The exploration should include **examples and tests**
- Your final deliverable should be a **literate program** (Org-mode)
- The final presentation should include **exercises** for the audience
- We can only accommodate **four (4) projects** altogether (groups of 3)
- Deliverables: proposal, literature review, abstract, Org-mode file



Which R packages are suitable for student projects?



There are many R packages that could be suitable for student projects, depending on the subject area and the goals of the project. Some packages that might be particularly useful for student projects include:

- `ggplot2`: a popular package for creating data visualizations
- `dplyr`: a package for data manipulation and preparation
- `tidyverse`: a package for cleaning and reshaping data
- `lme4`: a package for fitting linear mixed-effects models
- `caret`: a package for training and evaluating machine learning models
- `shiny`: a package for building interactive web applications with R

These are just a few examples, and there are many other packages that could be useful for student projects as well. It's always a good idea to explore the available packages and see what might be useful for your specific project.

Figure 7: R packages suitable for student projects by ChatGPT

Don't be fooled by ChatGPT

The screenshot shows a ChatGPT interface. The user asks, "create a vector of 10 integer numbers 1 to 10 in R". ChatGPT responds with two methods: using `seq()` and `c()`, or using the `1:10` operator. It also suggests checking the class and content of the vector with `class(x)` and `x` respectively.

You can create a vector of the integers from 1 to 10 in R using the `seq()` function and the concatenation operator `c()`:

```
x <- c(seq(1,10))
```

You can also use `1:10` operator which will create the same vector:

```
x <- 1:10
```

Also, to check the class and content of the vector you can use `class(x)` and `x` respectively.

```
x1 <- 1:10
x2 <- seq(from=1,to=10)
x3 <- c(1,2,3,4,5,6,7,8,9,10)
x1; x2; x3
class(x1)
```

What are you looking forward to?



- Developing lectures and exercises beyond the "Tidyverse"
- Learning more about R packages through your projects
- Having fun with R programming and real applications
- Graphics compilation in a short time
- Another tool under (better: on) your belt
- I want the data science minor and the skills
- Learn more about the functional aspects of R
- Excited to use Emacs!
- The projects!
- I learn better when I get to work hands-on
- Close to work environment
- I want to see how far I can take R

Next



Figure 8: Our next topic is "Argument matching"