Course overview

Introduction to data science (DSC 105) Fall 2022

Table of Contents

- 1. About me
- 2. Mutual introductions
- 3. Course syllabus (on GitHub and on Canvas)
- 4. Course "learning" platform: Canvas
- 5. Course topics
- 6. Video lectures
- 7. Agile [team] project (with "Scrum")
- 8. IMRaD and Scrum
- 9. Many project opportunities
- 10. Introduction TO DataCamp
- 11. Introduction to the textbook
- 12. Introduction to GNU Emacs + ESS + Org-mode
- 13. Literate programming
- 14. Home assignments
- 15. Tests (not graded except final exam)
- 16. Practice - course infrastructure
- 17. Glossary

Figure 1: Blaues Pferd I (Franz Marc, 1911)

1 About me

Figure 2: Teddy Roosevelt at Harvard (ca. 1877)

- PhD theoretical particle physics (mostly worked alone)
- Data science interests: languages, infrastructure, culture
- Research options: quantum computing, medical imaging, cybersecurity

2 Mutual introductions

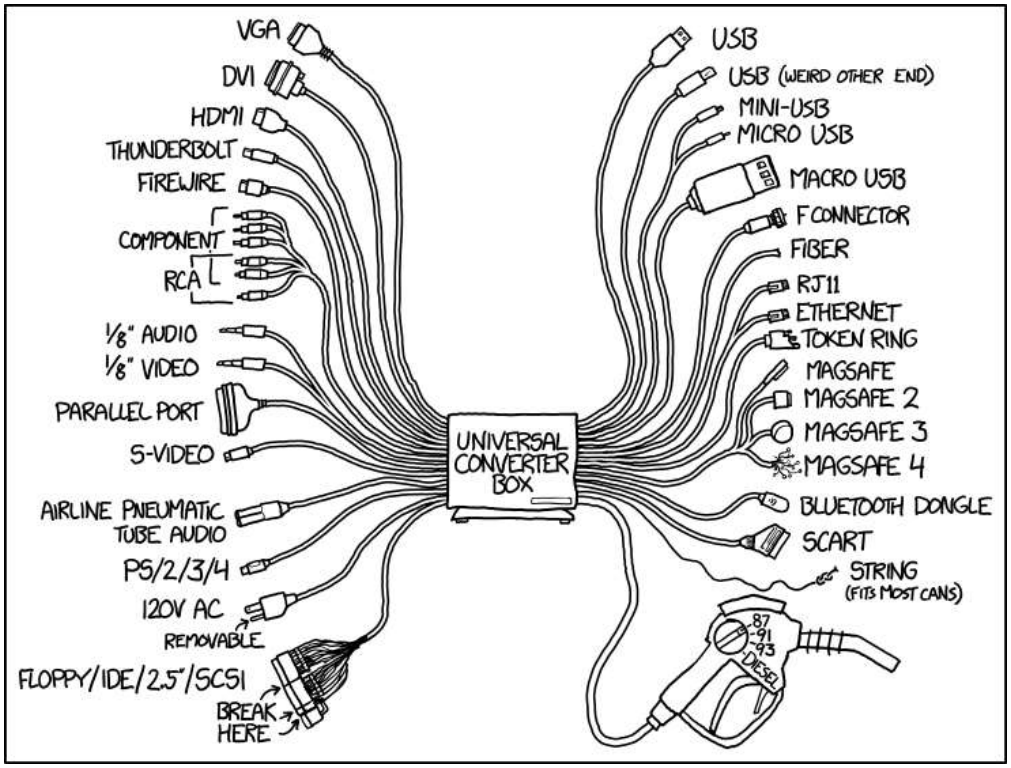

Figure 3: xkcd: Universal Converter Box

1. Why are you here?
2. What would delight you?
3. What would disappoint you?
4. Where are you headed?

3 Course syllabus (on GitHub and on Canvas)

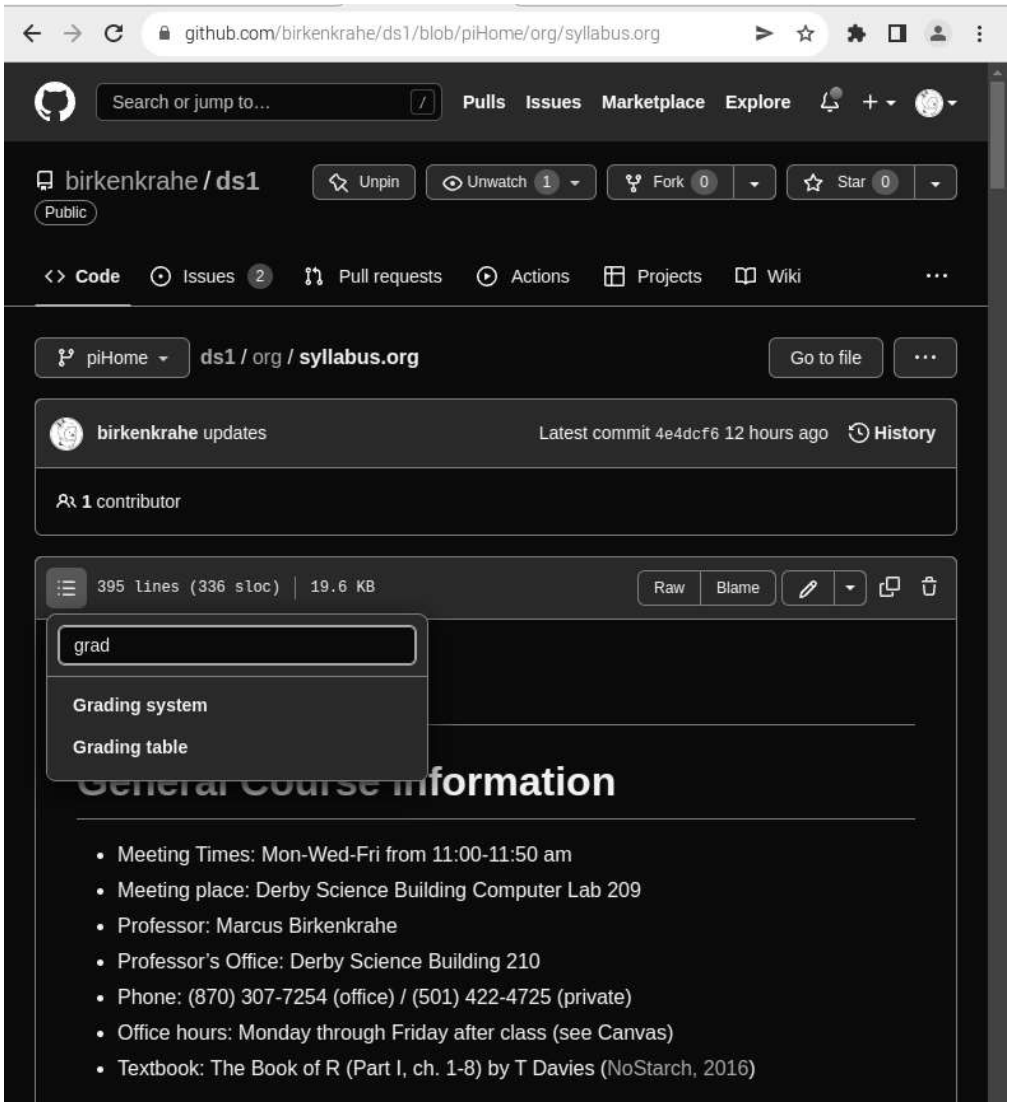

Figure 4: Syllabus on GitHub

- General information & standard policies
- Course information (grading, attendance)
- Schedule with dates of tests and assignments
- The GitHub repo contains course material

4 Course "learning" platform: Canvas

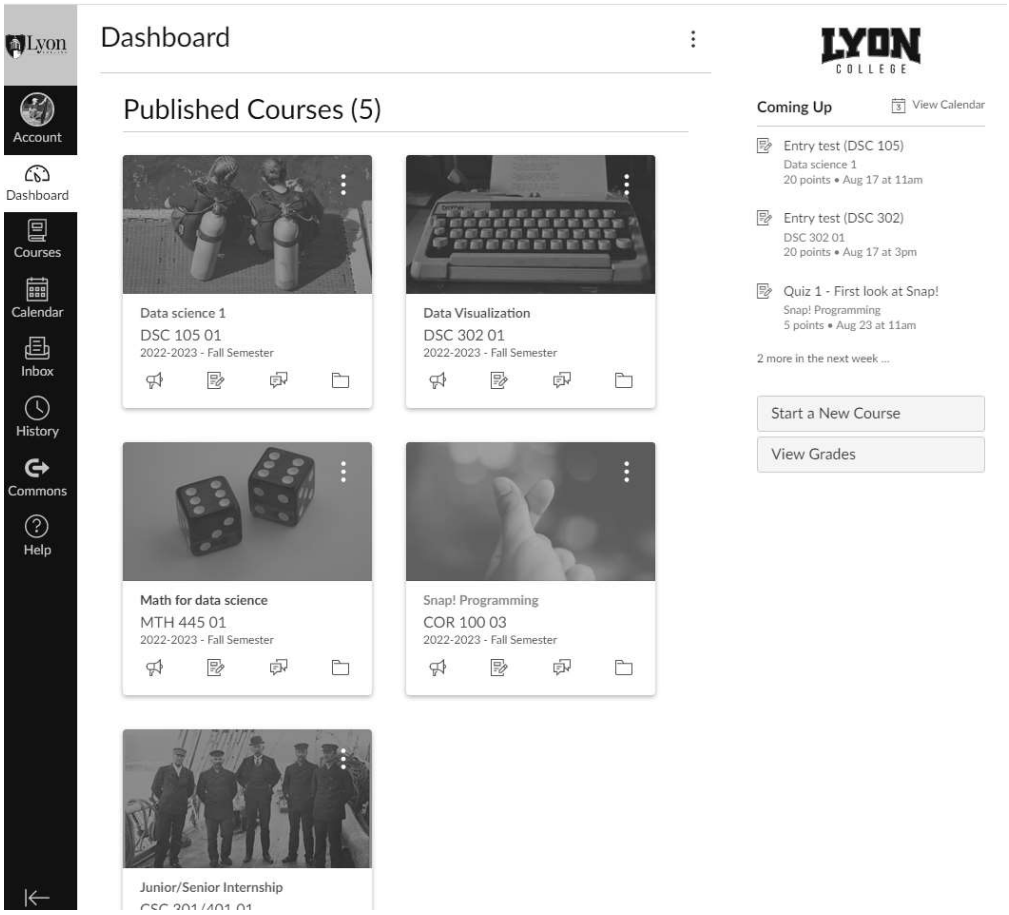

Figure 5: Course topics

- All grades should be visible at all times
- Control your own notifications (email)
- Important course links on a page
- New CMS for me & for Lyon: bear with us[1]
- Lecture *notes* (from Emacs Org-mode) will not show in GitHub

5 Course topics

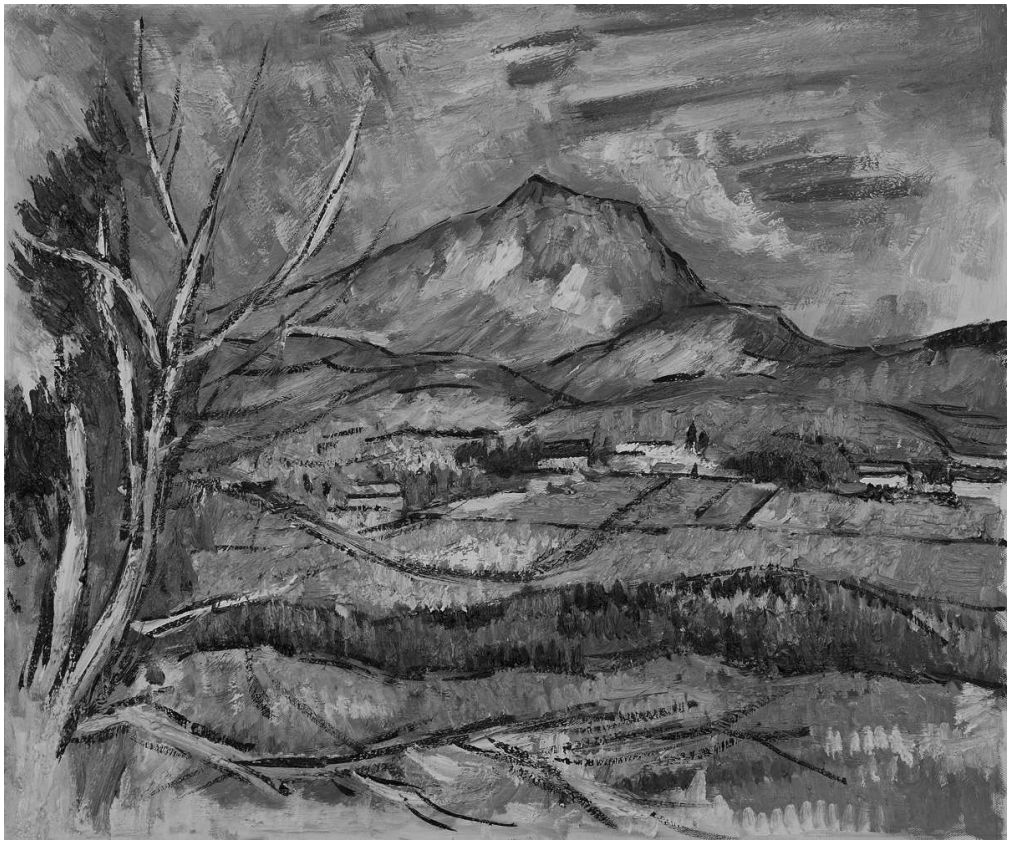

Figure 6: Course topics

1. The R statistical programming language
2. Basics of data visualization with R
3. Professional software development methods

6 Video lectures

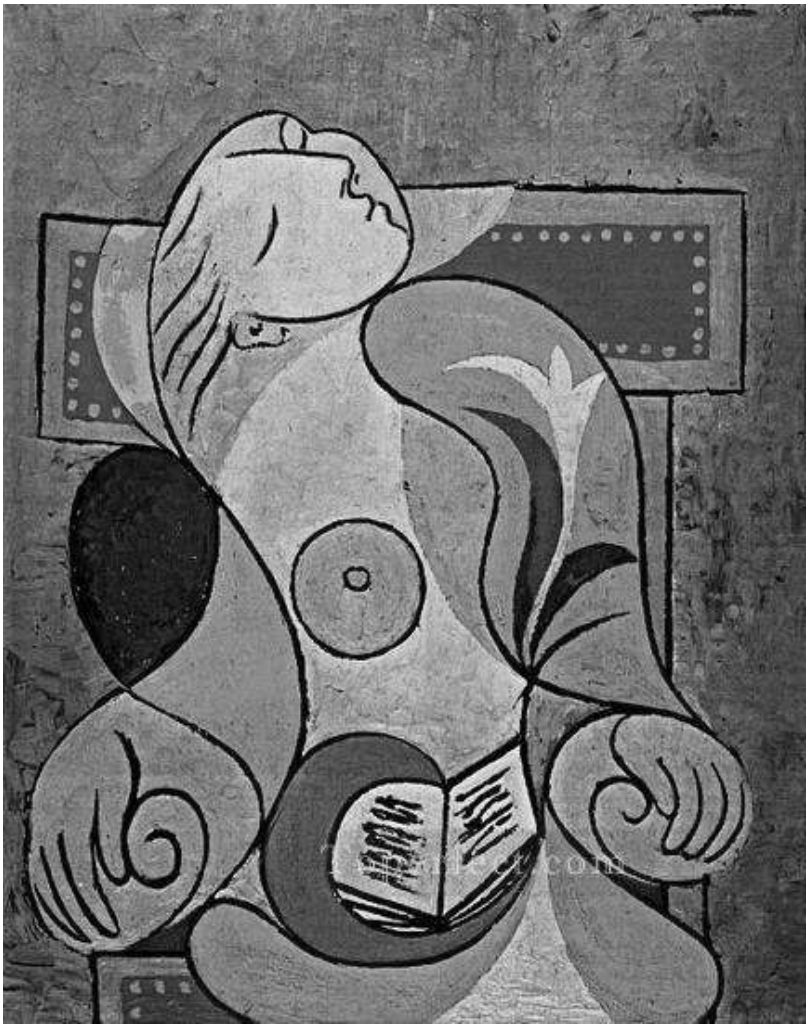

Figure 7: La lecture Marie Therese (Picasso, 1932)

- Emacs + Org-mode + R (Tutorial videos Spring '22)
- Introduction to R: installation and shell
- Vectors in R (part 1, part 2, part 3)
- Data frames, matrices, lists, factors in R
- Data frames in R
- Base R plotting
- Plotting with ggplot2
- Data import with R
- RStudio R Notebooks and literate programming

7 Agile [team] project (with "Scrum")

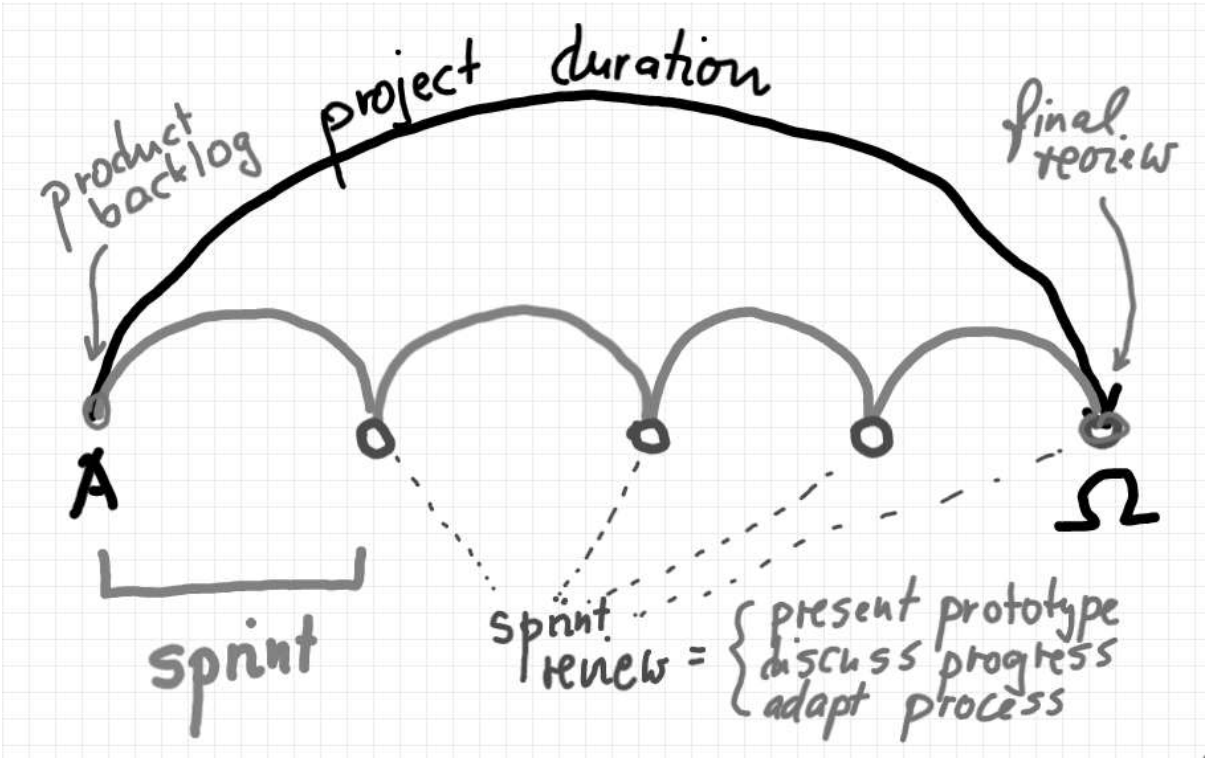

Figure 8: Agile (Scrum) project

The team project makes up 20% of your final grade for this course.

- What is a team project? (FAQ)
- Do you have examples for data science projects? (FAQ)
- Can you do a project as an absolute beginner? (FAQ)

Note: the first *sprint review* is on August 31. Use it to present your initial results (see FAQ on what to deliver, and 1st sprint review).

8 IMRaD and Scrum

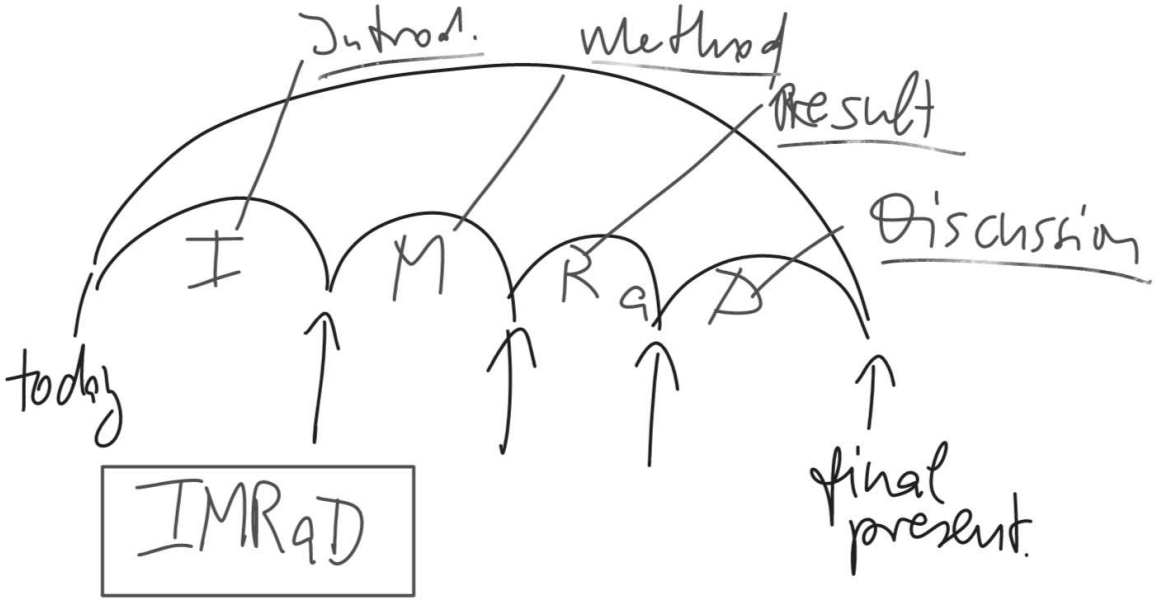

Figure 9: Agile (Scrum) project

- Introduction (research question - what you want to find out)
- Method (how you want to do it)
- Results (what you found out)
- Discussion (what it means)

(Video: Research Writing with IMRaD)

9 Many project opportunities

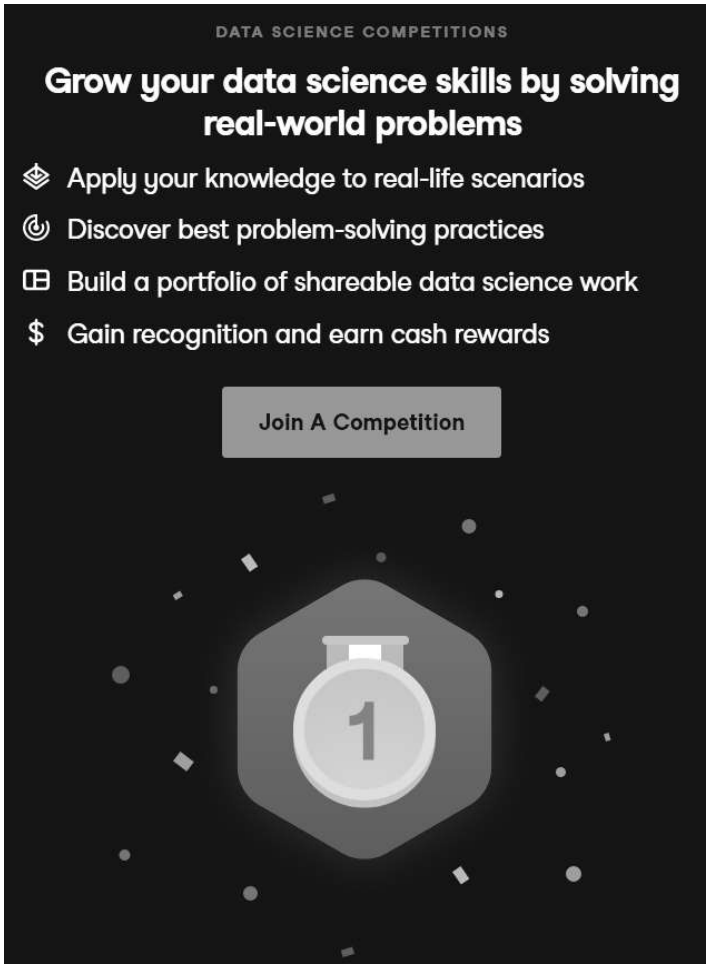

Figure 10: DataCamp competition announcement

- Explore and document an R package
- Document an extended analysis example
- Explore a data set of your choice
- Complete a DataCamp competition
- See DataCamp projects for examples
- You can branch out: SQL, Python, Java etc.
- See GitHub issues for examples (e.g. whale song)
- Double/triple up if you're in > 1 of my courses
- Use problems from other courses for your project, e.g. data collected by yourself, or data in economics, business, art etc.

10 Introduction TO DataCamp

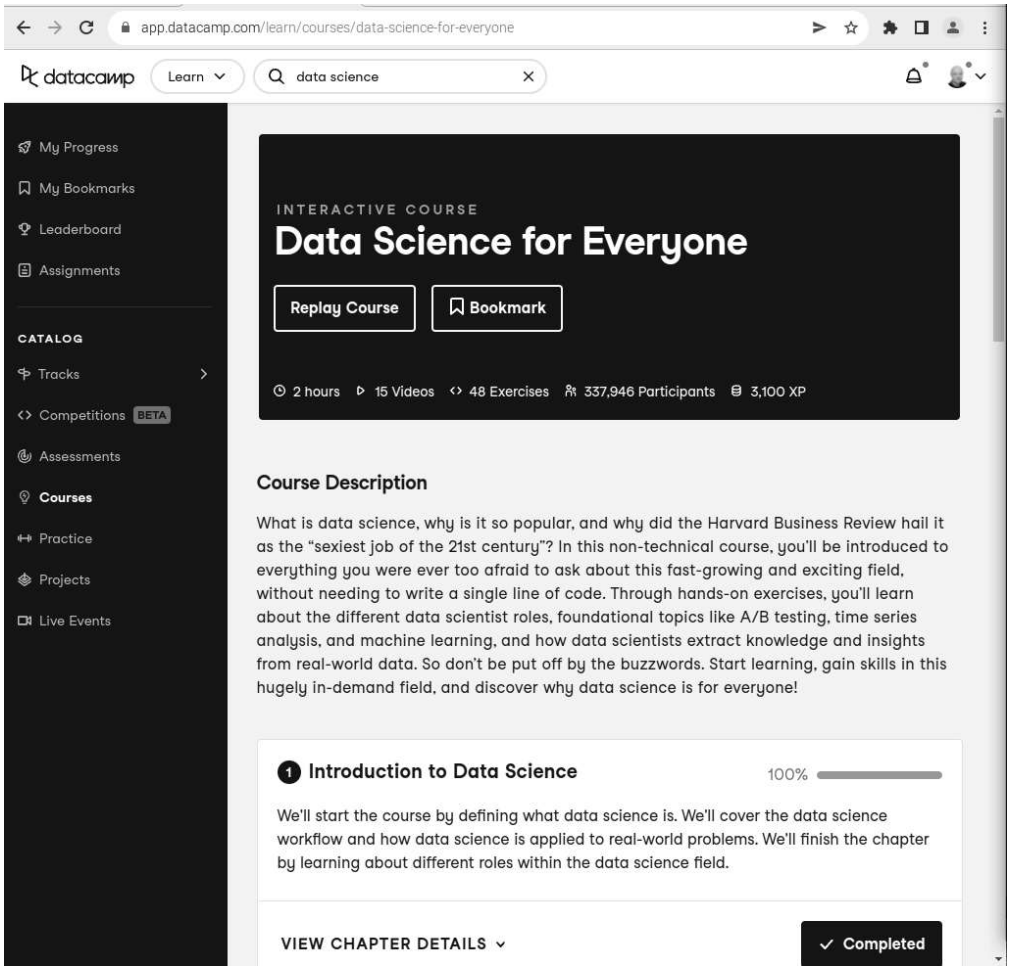

Figure 11: DataCamp course start page

- DataCamp is a data science learning platform
- Access for you is free (academic alliance)
- 9/15 assignments are DataCamp assignments
- Assignments are drawn from 4 courses
    1. Data science for everyone
    2. Introduction to R
    3. Data visualization with R
    4. Introduction to data visualization with `ggplot2`
- Complete them on time to get full points
- Completed DataCamp courses can support your resume

11 Introduction to the textbook

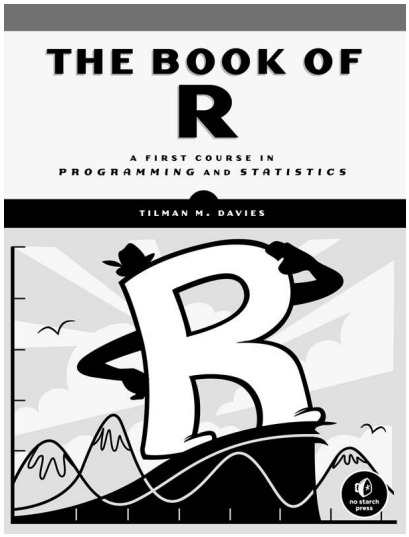

Figure 12: Cover of Book of R (Davies, 2016)

- R is *FOSS* with focus on stats and graphics
- Davies' "Book of R" is extensive (832p.)
- You don't have to read along but it might help
- Many other tutorials and textbooks available
- The best short online tutorial: Matloff's "fasteR"
- Beware of ideologies (cp. Matloff's "TidyverseSceptic")

12 Introduction to GNU Emacs + ESS + Org-mode

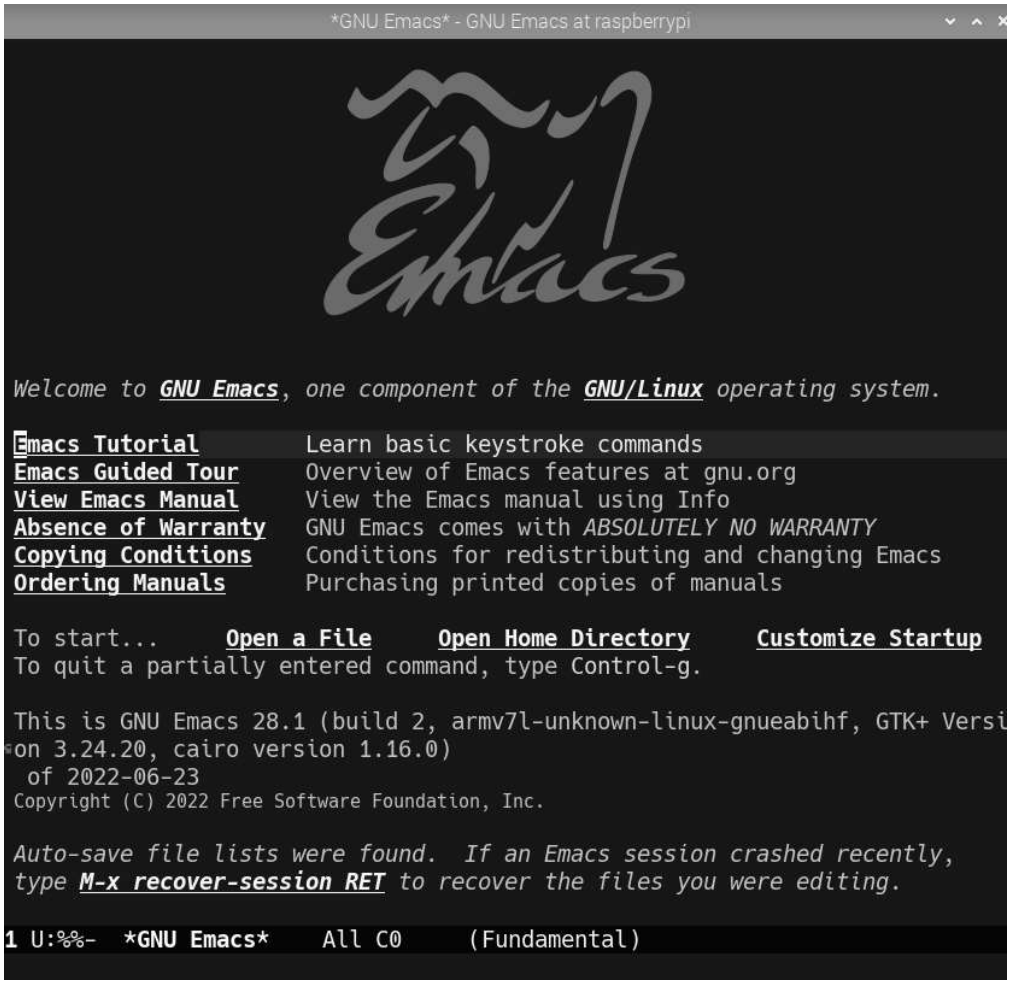

Figure 13: GNU Emacs start page

- Emacs: self-documenting, extensible *FOSS* text editor
- Process, file and package management (like an OS)
- *Literate programming* environment for 43 languages
- *IDE* for R programming and *REPL* for interactive coding

13 Literate programming

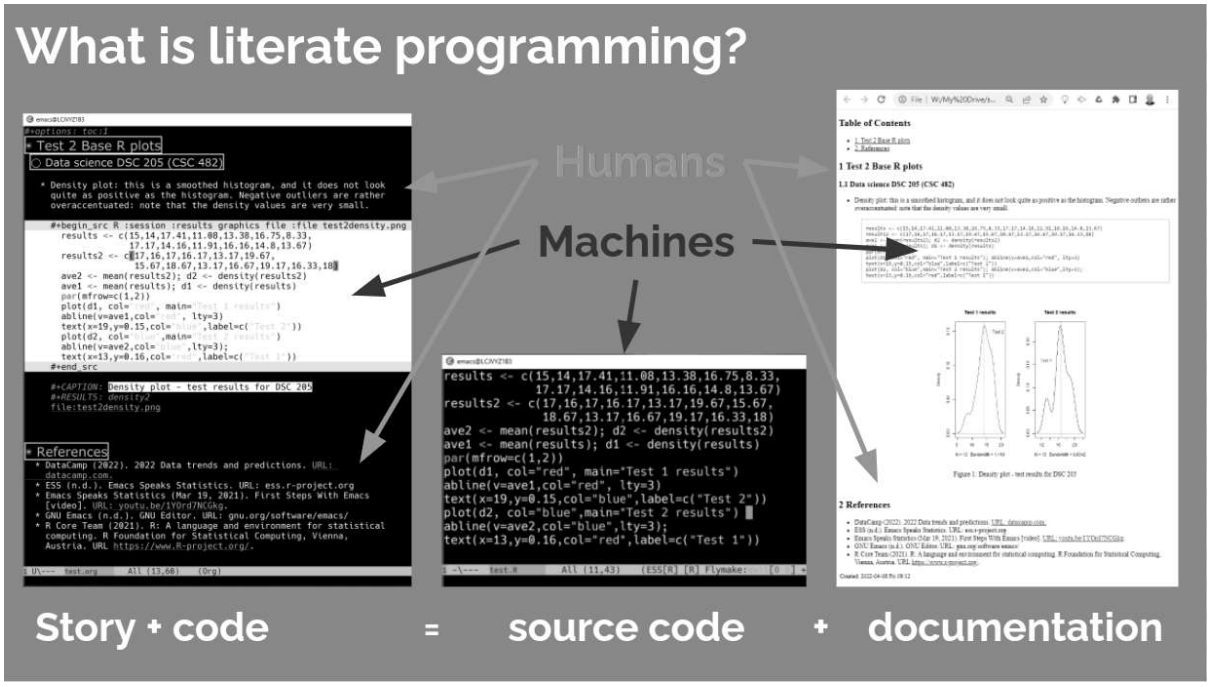

Figure 14: What is literate programming?

Source: "Teaching data science with hacker tools" (2022)

- Common practice among data scientists
- *Paradigm* behind interactive computing notebooks
- Useful when learning any programming language

14 Home assignments

- There are 15 programming assignments altogether = 10 points each, or 30% of your final grade.
- Register with DataCamp and complete the DataCamp chapter Introduction to data science by Monday, 22 August at 11 am (ca. 20 min).
  - Data science definition
  - Data science workflow
  - Application to real-world problems
  - Different professional data science roles
- Complete the Emacs on-board tutorial and upload an edited copy to Canvas by Friday, 26 August at 11 am (ca. 60 min).
  - Get comfortable with Emacs keyboard bindings
  - Learn how to create, view, edit, save files
  - Learn how to insert a time stamp automatically

15 Tests (not graded except final exam)

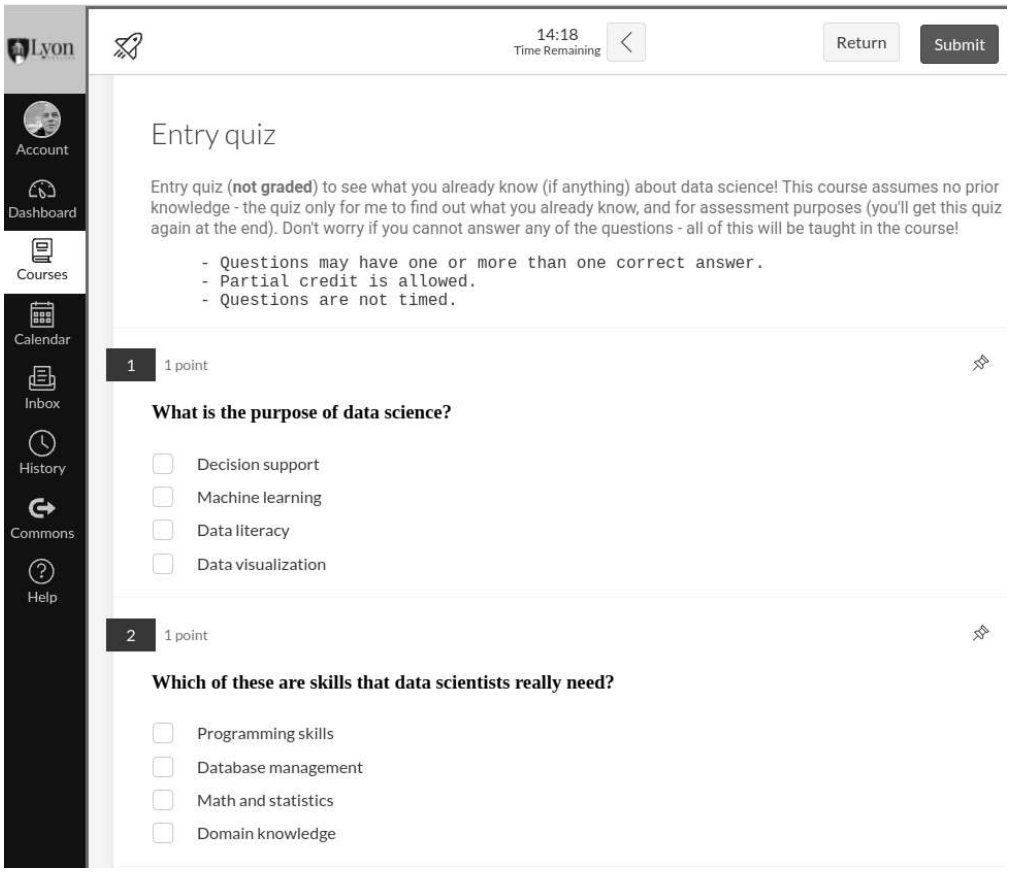

Figure 15: Start page of the entry quiz on Canvas

- Tests have to be completed online, are timed, and have a deadline; after the deadline, you can play them an unlimited number of times
- There will be a revision quiz on Canvas every week, consisting of 5-10 multiple choice, matching and true/false questions.
- A subset of the test questions will form the final exam (20% of your final grade) - we will practice in the last week before the exam.

16 Practice - course infrastructure

Useful: take notes! Practice leads to mastery and the practice exercises will often come back to haunt you in the tests.

1. Open a browser
2. Find the GitHub repos (birkenkrahe/ds1 and /org)
3. Open the command line terminal
4. Open/close R
5. Open Emacs
6. Find the Emacs tutorial
7. Open/close R inside Emacs
8. Run R in an Org-mode file
9. Close Emacs
10. Close the command line terminal

Note: Class room practice completion = 10 points each for active participation.

17 Glossary

TERM	MEANING
Command line	aka terminal/shell to talk to the OS
Emacs	GNU self-extensible text editor
FOSS	Free and Open Source Software
GitHub	Software development platform
Git	Version control software
GNU	GNU's not Unix
IDE	Integrated Development Environment
"Literate Programming"	Story + code => source code + doc
Paradigm	A standard way of looking at things
R	FOSS statistical programming language
REPL	Read-Eval-Print-Loop
Repo	Code repository
"Tidyverse"	Popular R package bundle
Scrum	Agile project management method
Sprint review	Period to complete a prototype
Prototype	Intermediate (not perfect) solution

Footnotes:

[1] CMS = Content Management System; these are the most common systems in business applications - present whenever people create 'content' of any sort (documents e.g.) and need to store it for later. CMS systems rely on database technology. In the case of Canvas, that's MySQL.

Author: Marcus Birkenkrahe
Created: 2022-08-27 Sat 13:38