Course overview

Introduction to data science (DSC 105) Fall 2022

Marcus Birkenkrahe

July 27, 2022

Contents

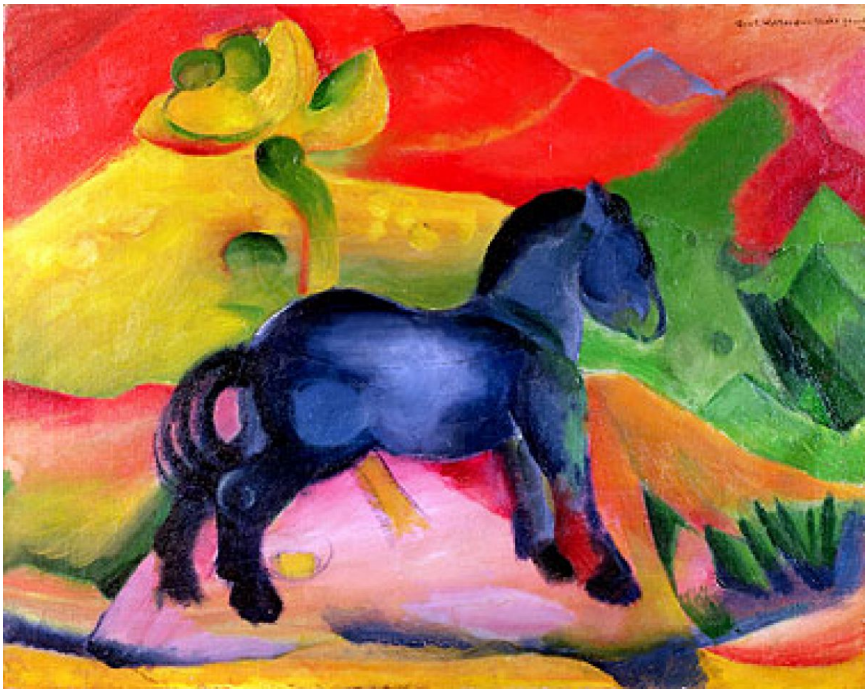

Figure 1: Blaues Pferd I (Franz Marc, 1911)

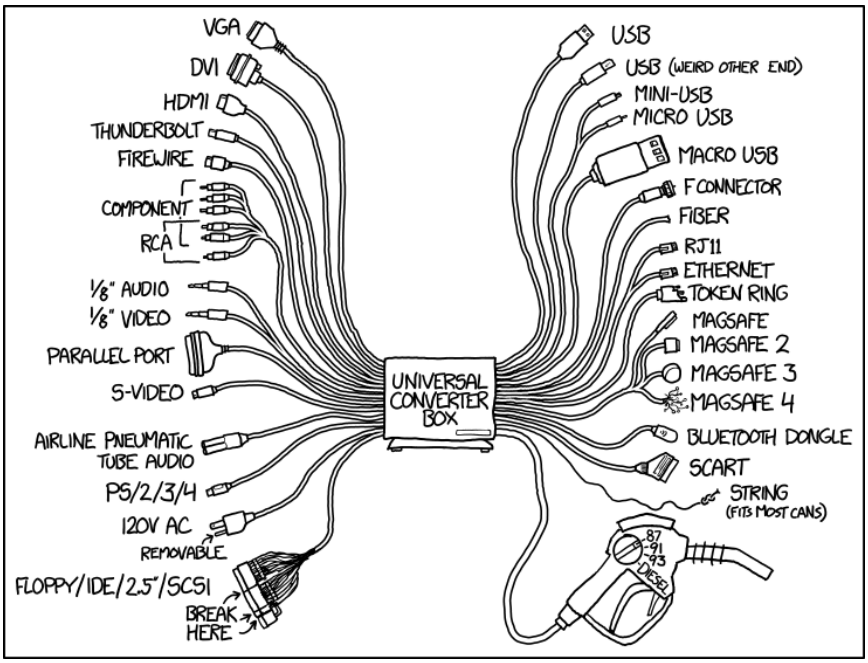

Figure 2: xkcd: Universal Converter Box

1 MUTUAL INTRODUCTIONS

1. Why are you here?
2. What do you want to get out of this class?
3. What would disappoint you?
4. Where are you headed?

2 COURSE SYLLABUS (on GitHub and on Canvas)

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

3 COURSE TOPICS

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

4 VIDEO LECTURES

- 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

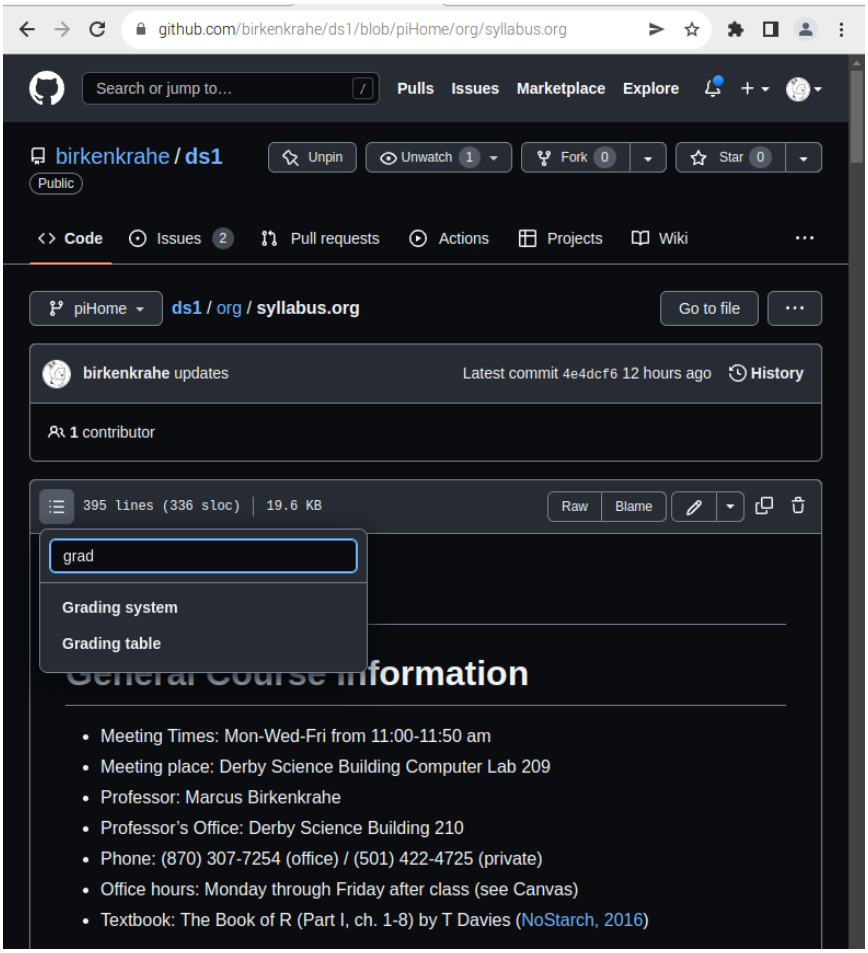

Figure 3: Syllabus on GitHub

Figure 4: Course topics

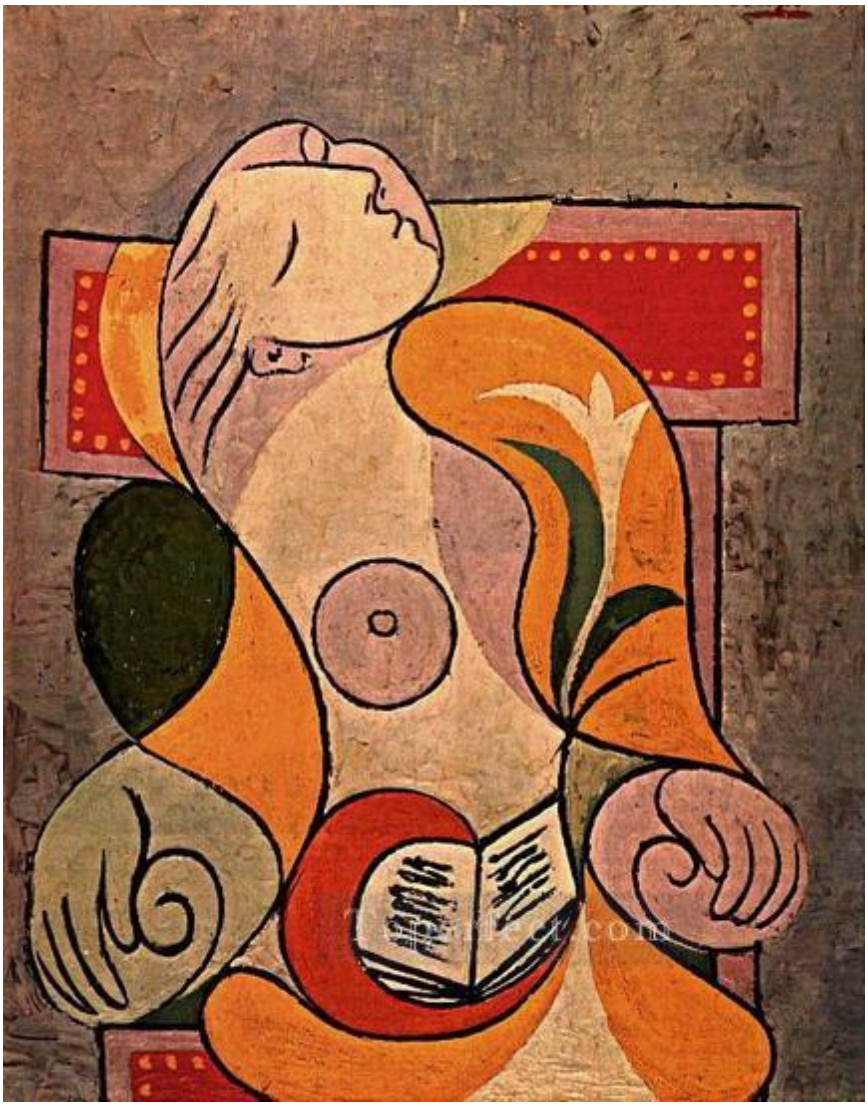

Figure 5: La lecture Marie Therese (Picasso, 1932)

- Data import with R
- RStudio R Notebooks and literate programming

5 AGILE TEAM PROJECT

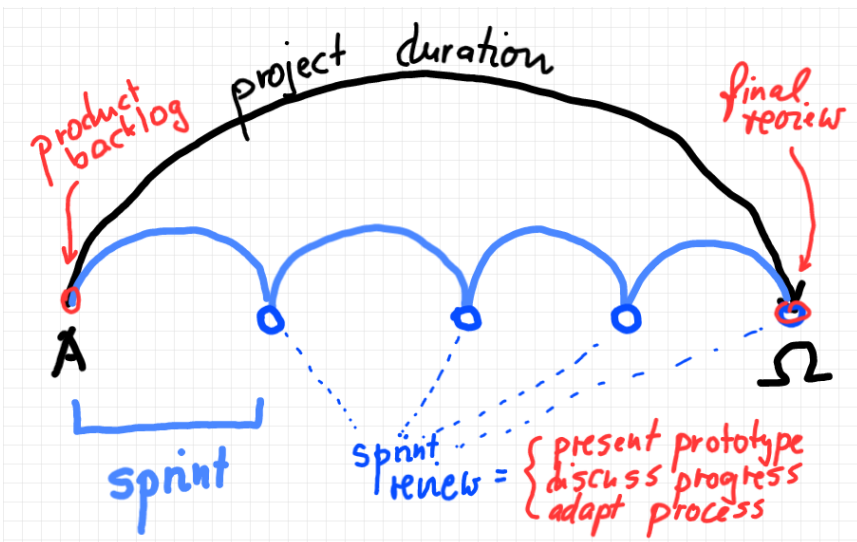

Figure 6: Agile (Scrum) project

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).

6 INTRODUCTION TO DataCamp (assignments)

- DataCamp is a data science learning platform
- Access for you is free (classroom license)

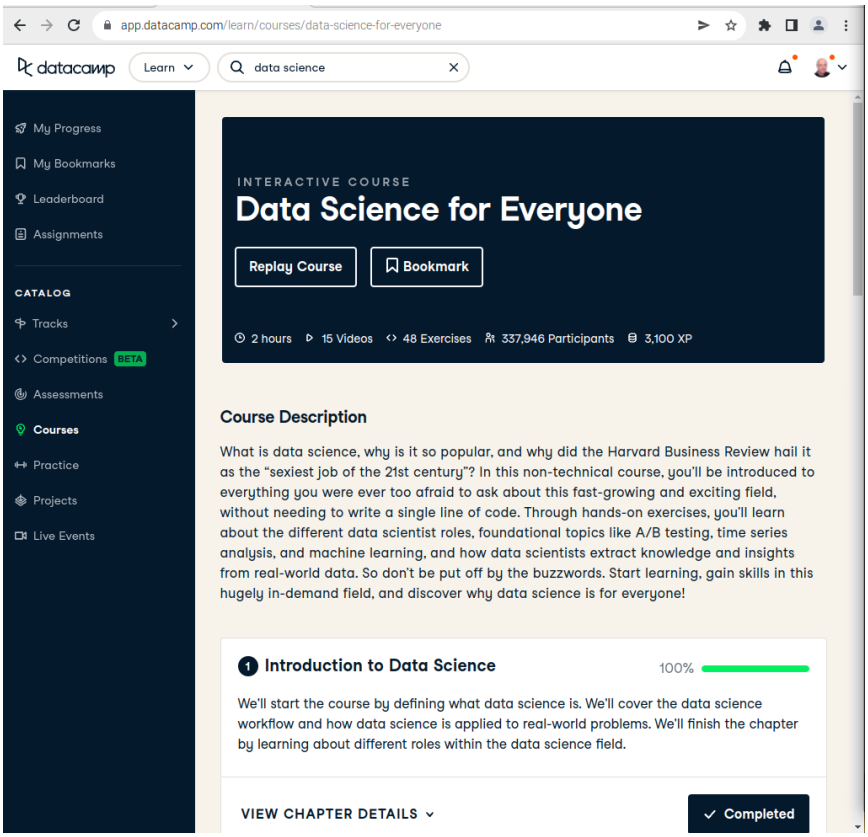

Figure 7: DataCamp course start page

- 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

7 INTRODUCTION TO THE TEXTBOOK

- 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")

8 INTRODUCTION to GNU Emacs + ESS + Org-mode

- 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

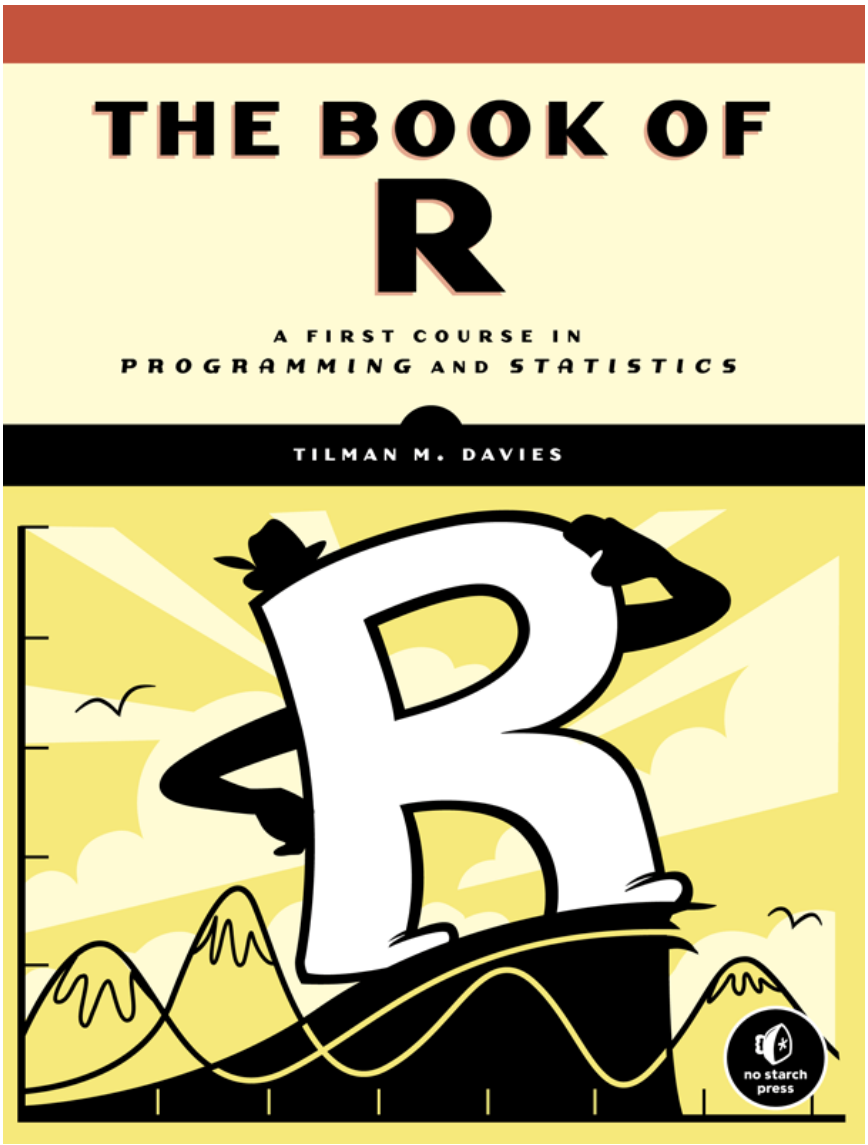

Figure 8: Cover of Book of R (Davies, 2016)

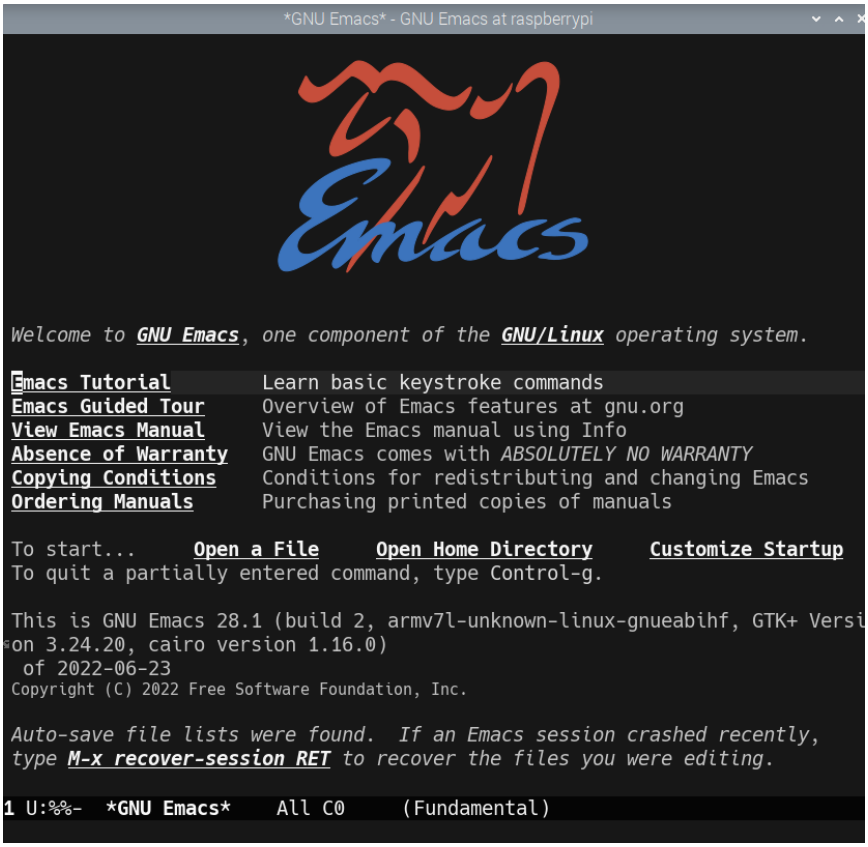

Figure 9: GNU Emacs start page

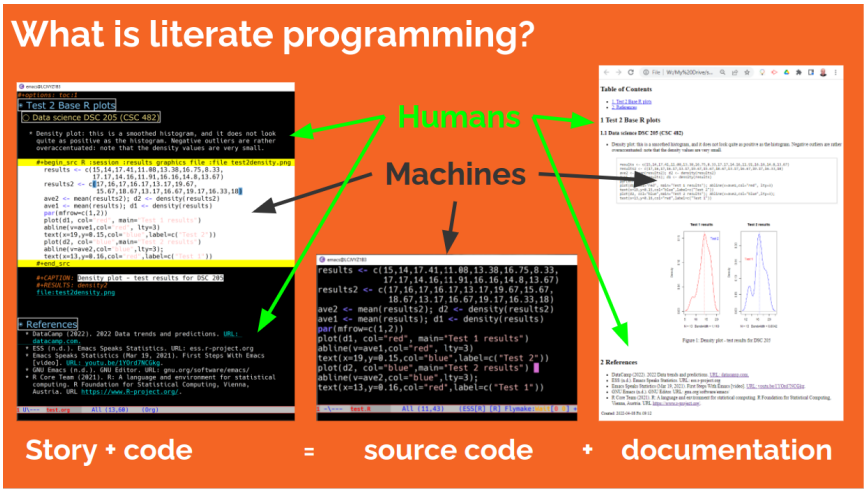

Figure 10: What is literate programming?

9 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

10 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.

11 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

12 TESTS (NOT GRADED)

- 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)

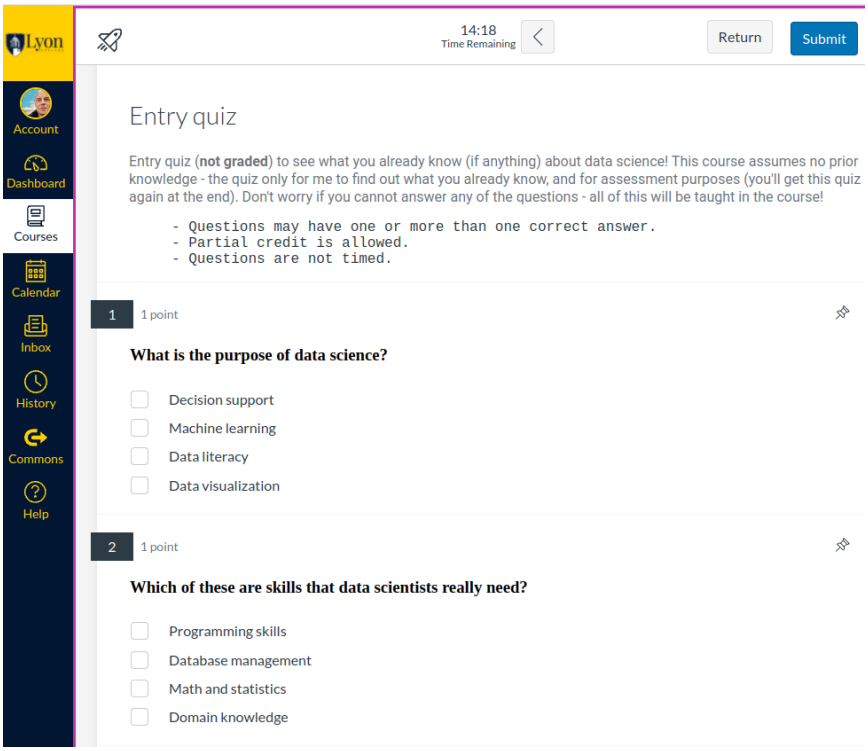

Figure 11: Start page of the entry quiz on Canvas

13 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