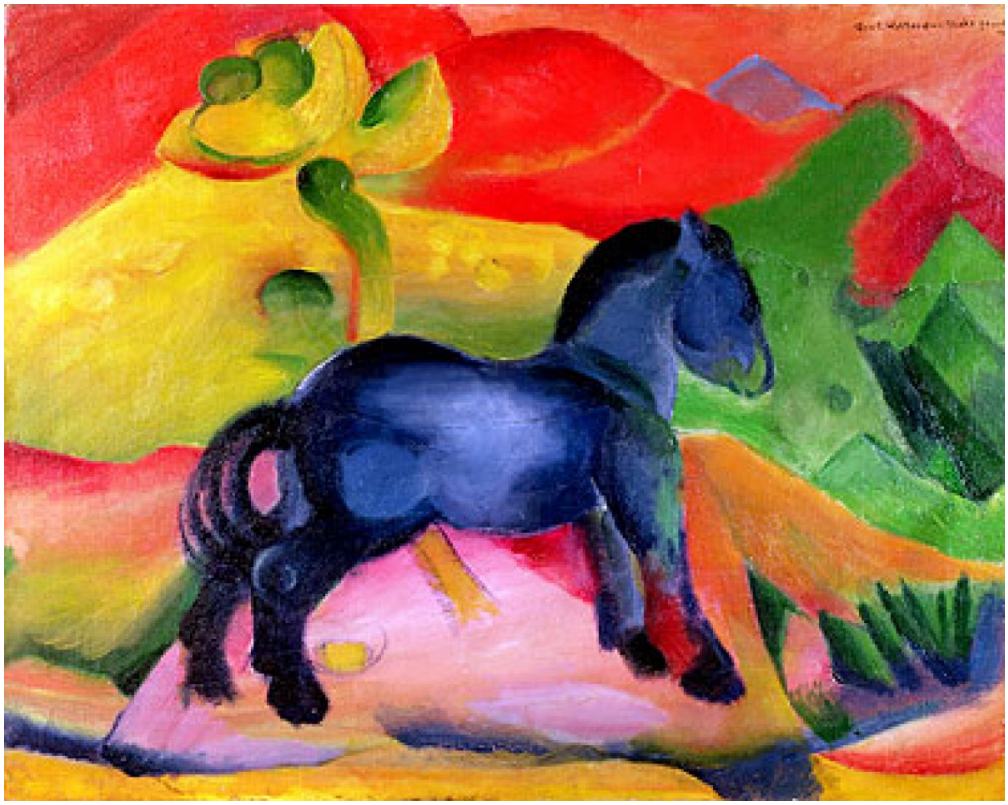
Course overview

Introduction to data science (DSC 105) Fall 2023

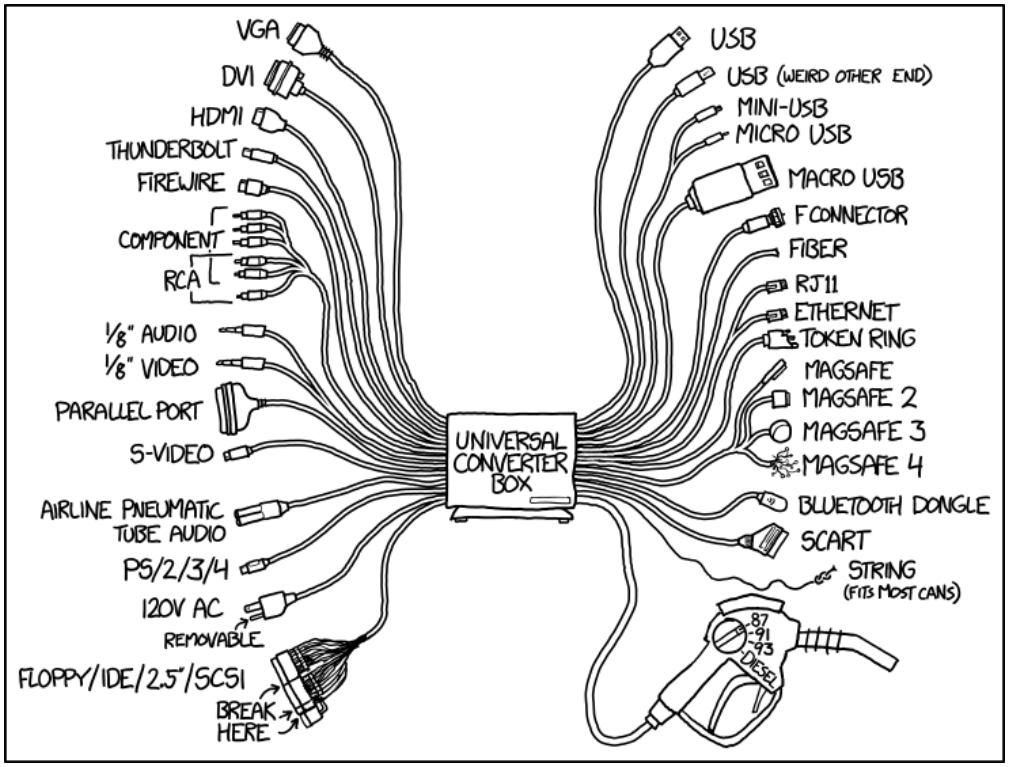
Table of Contents

About me



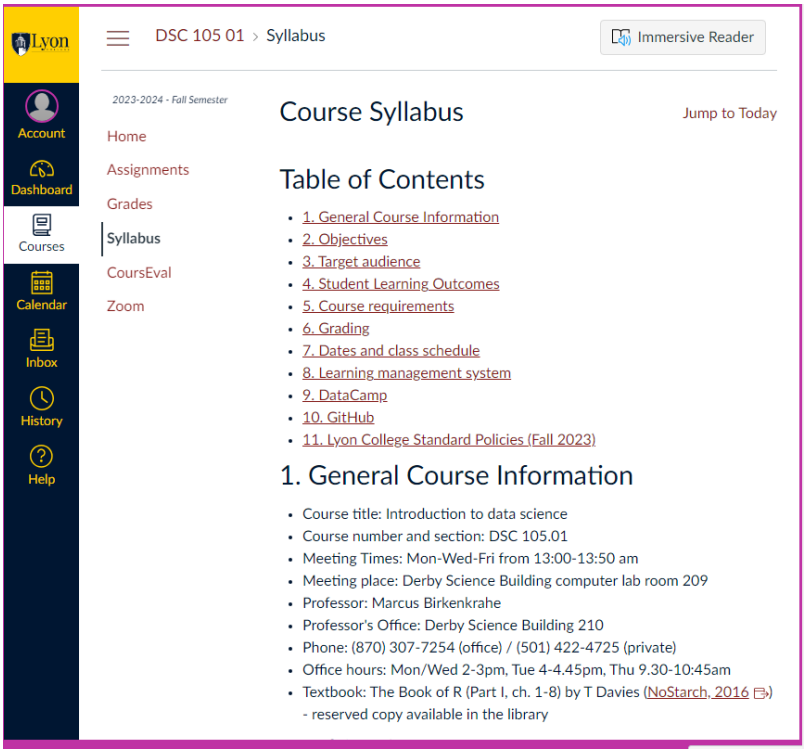
- PhD: theoretical particle physics
- Industry: Knowledge management
- Teaching data science / Coding w/AI

About you



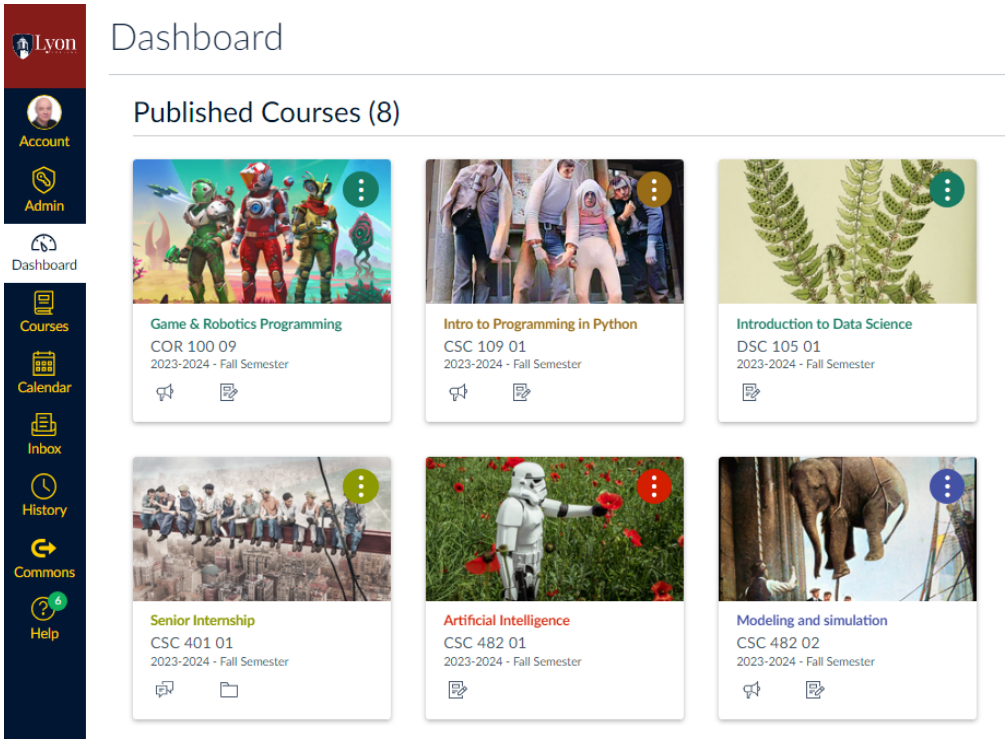
1. Why are you here?
2. What would delight you?
3. What would disappoint you?
4. Where are you headed?

Course syllabus



- Syllabus on Canvas
- General information & standard policies
- Course information (grading, schedule)

Canvas LMS

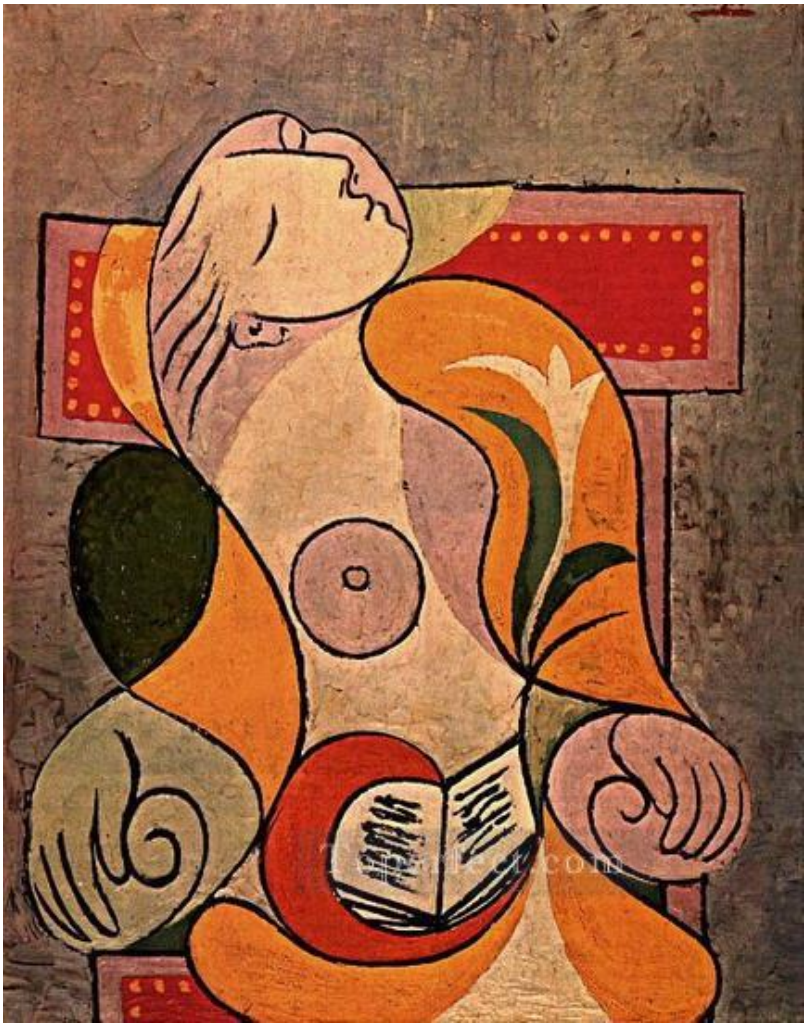


- All grades should be visible at all times
- Control your own notifications (email)
- Important course links on a page
- Add all your Canvas calendars to Google

Course topics

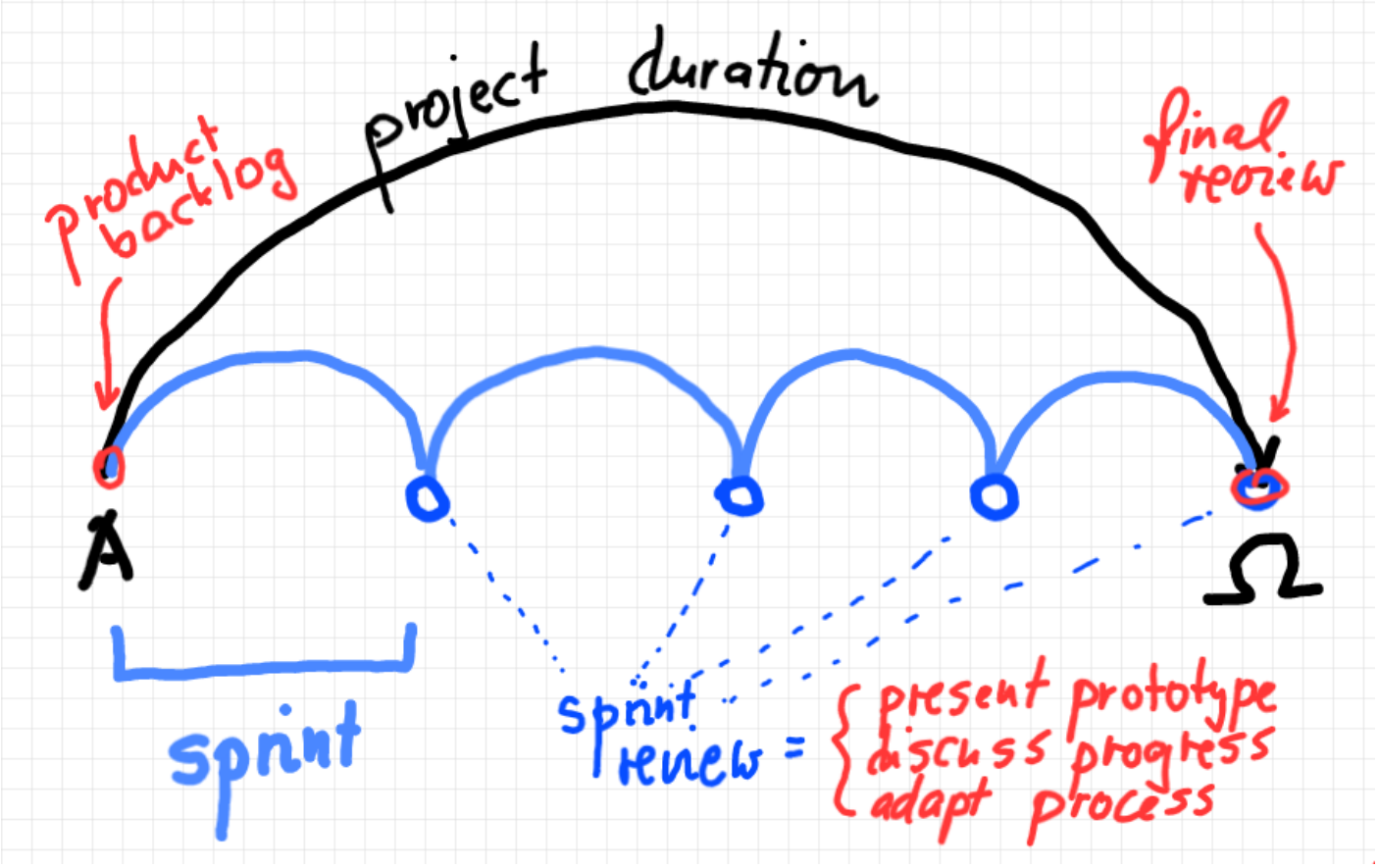
1. The R statistical programming language
2. Basics of data visualization with R
3. Software development methods
4. From R to Python … and back again

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

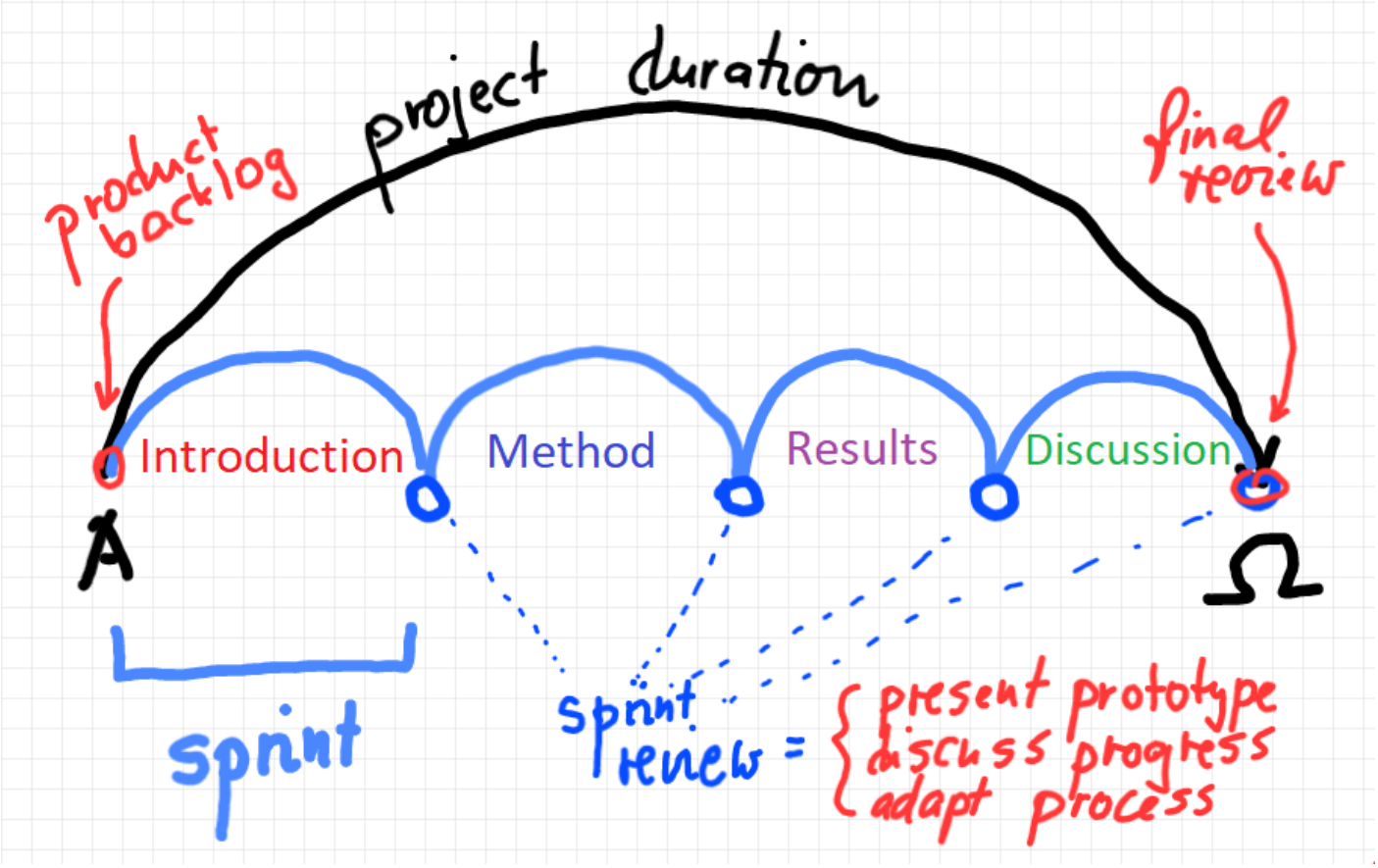
Agile project (with "Scrum")



- The project makes up 25% of your final grade for this course.
- What is a Scrum project? (GitHub 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 in the week of Sep-11. Use it to present your initial results (see FAQ on what to deliver, and 1st sprint review).

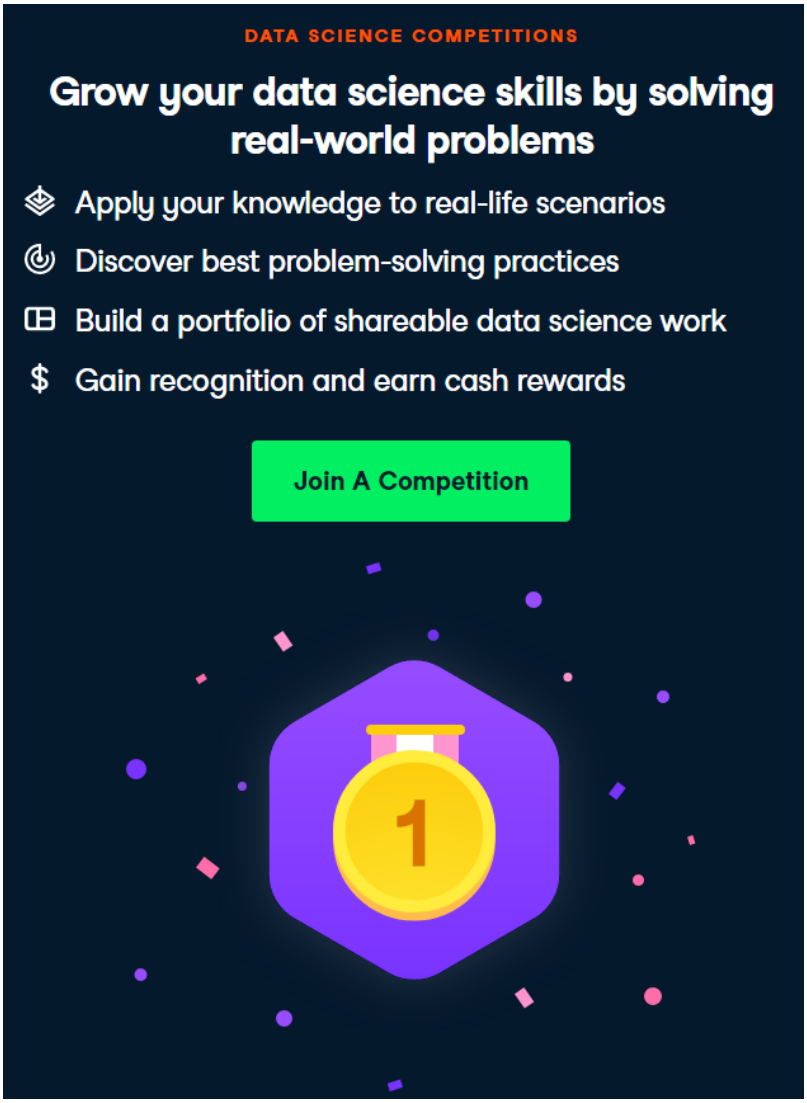
IMRaD and Scrum



- 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)

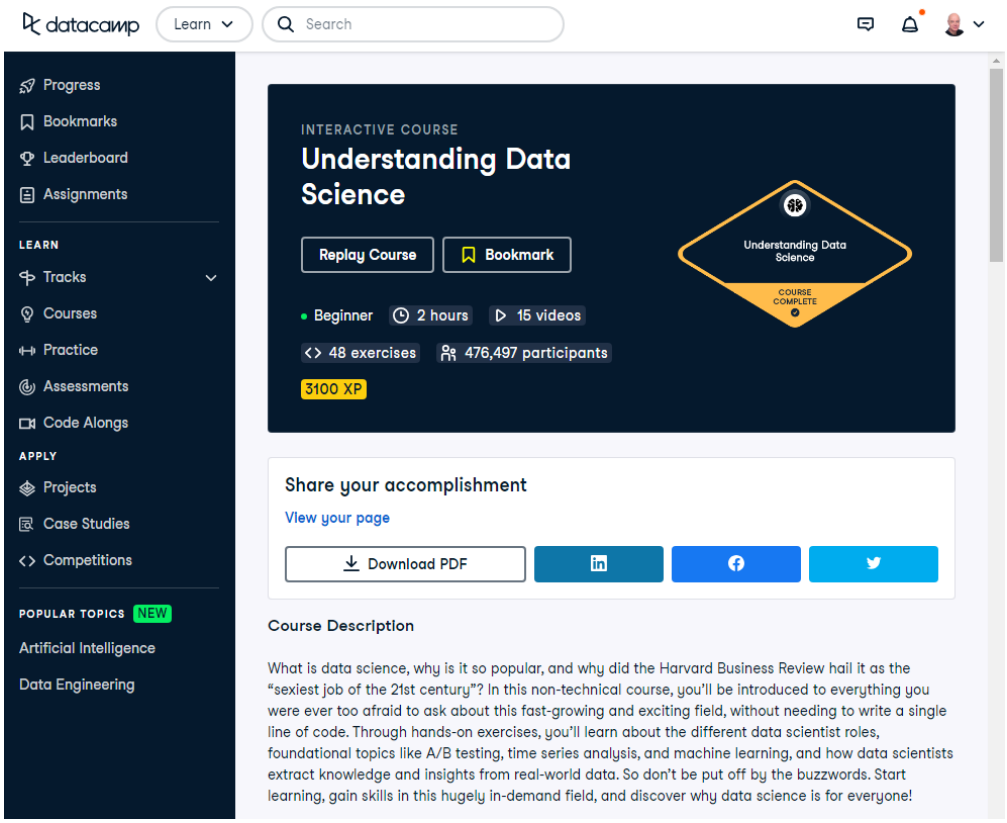
([Video: Research Writing with IMRaD])

Many project opportunities



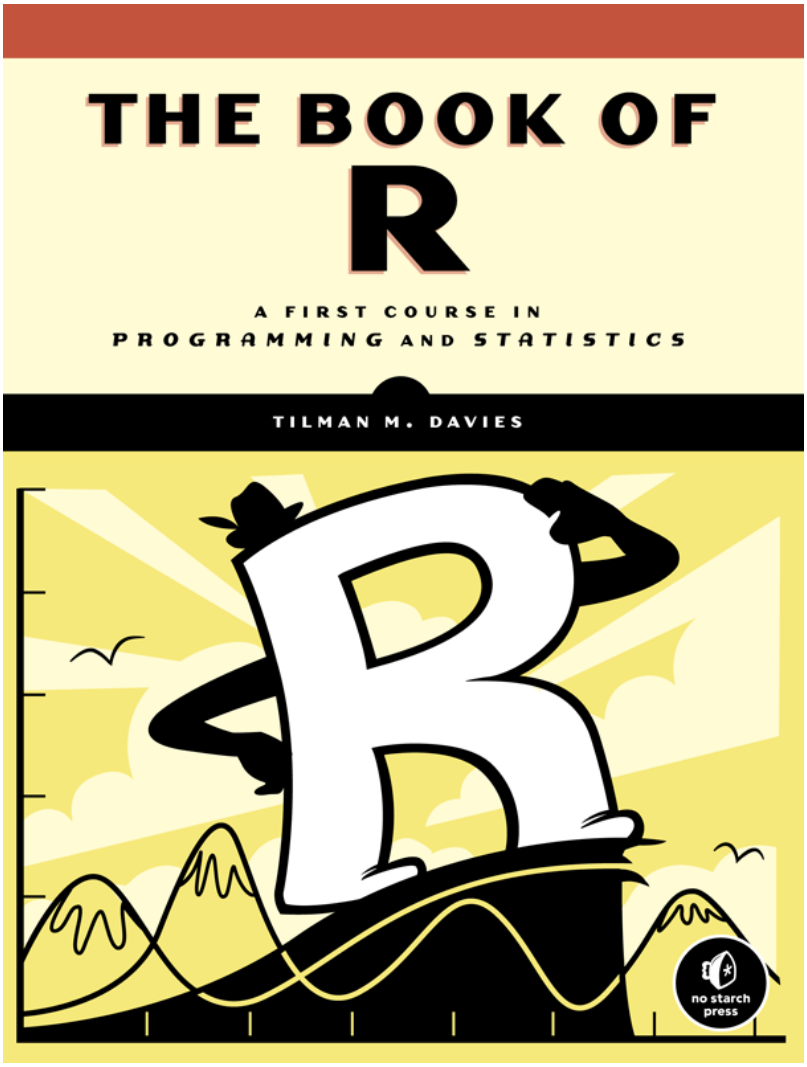
- Explore and document an R package
- Document an extended analysis example
- Explore a data set of your choice
- Complete a DataCamp competition
- See DataCamp projects for examples
- You can branch out: SQL, Python, Java etc.
- See GitHub issues for examples (e.g. whale song)
- Double/triple up if 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.

Introduction to DataCamp



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

NEXT Introduction to the textbook

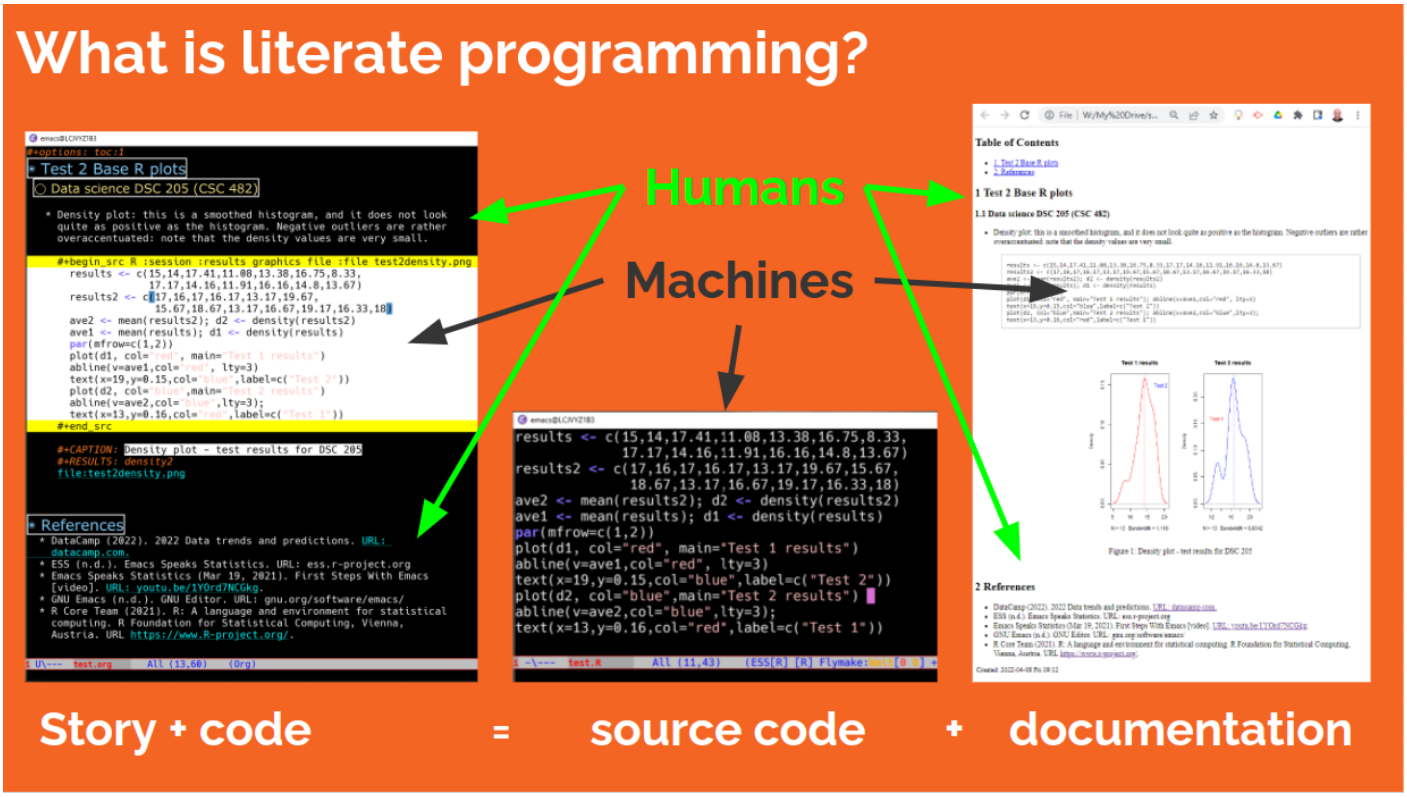


- R is *FOSS* with focus on stats and graphics
- Davies' "Book of R" is extensive (832p.)
- You don't have to read along but it might help
- Many other tutorials and textbooks available
- The best short online tutorial: Matloff's "fasteR"
- Beware of ideologies (cp. Matloff's "TidyverseSceptic")

Good-bye Emacs

- 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

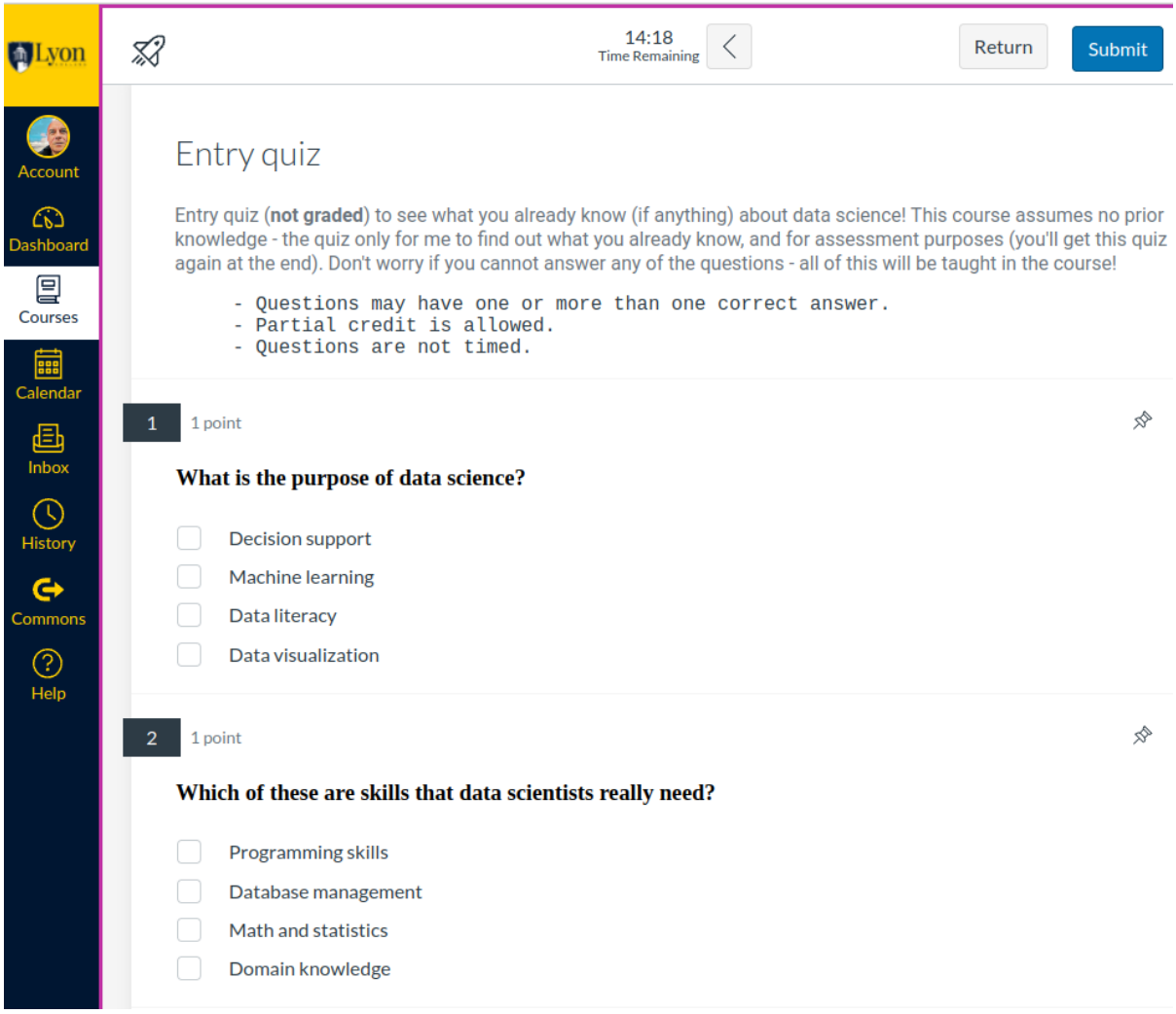
Literate programming



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

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

Tests (multiple-choice)



- 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 (25% of your final grade) - we will practice in the last week before the exam.

Home assignments

- Register with DataCamp now
- Read "Adventures with R" by Monday 28 August 1 PM (Canvas)

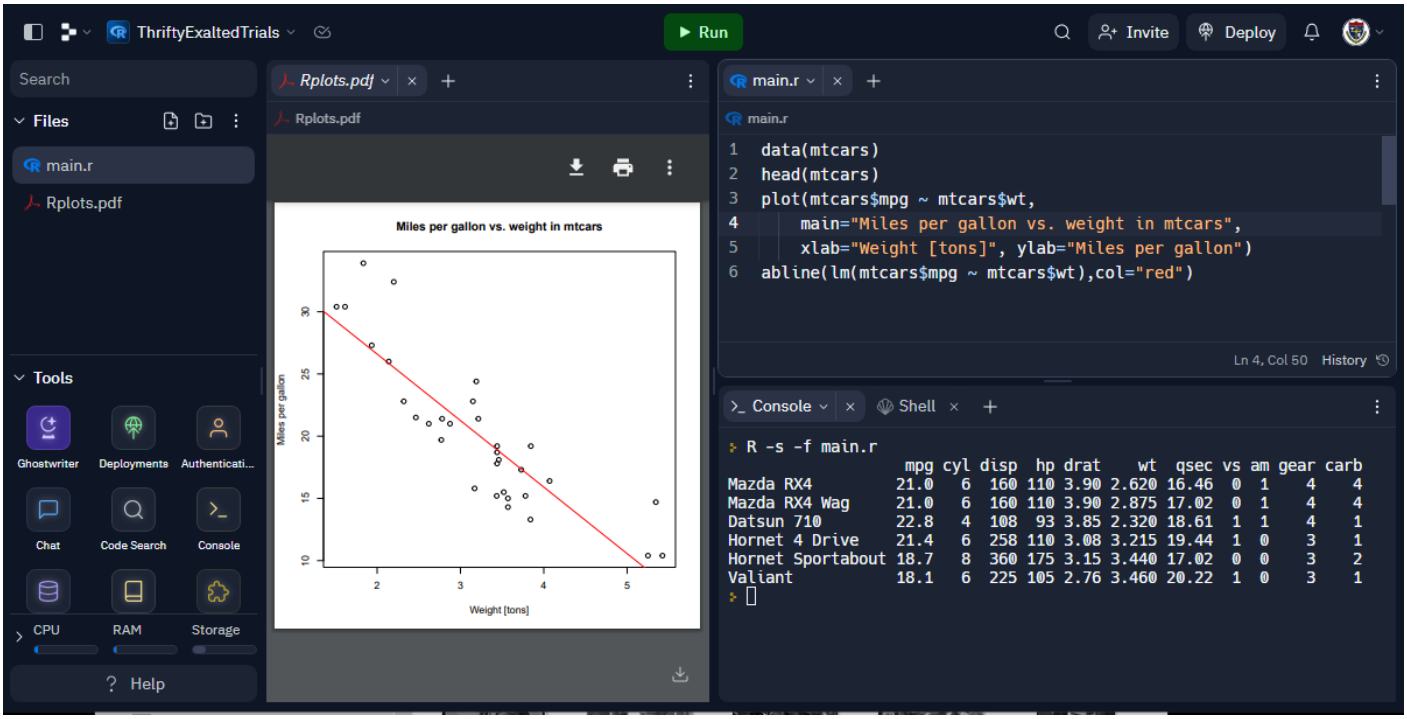
  We're going to look at DataCamp workspace tomorrow - here is a good short summary video of the AI coding capabilities: https://youtu.be/9XHop0xTTCM?si=bY4UsJmSrzxCYMUI - for this particular tool: what're your thoughts on this? How is it going to change coding and data science? For the better or for the worse?

- Complete 1st DataCamp assignment by Sunday 11.59 PM

Introduction to data science

- Data science definition
- Data science workflow
- Application to real-world problems
- Different professional data science roles

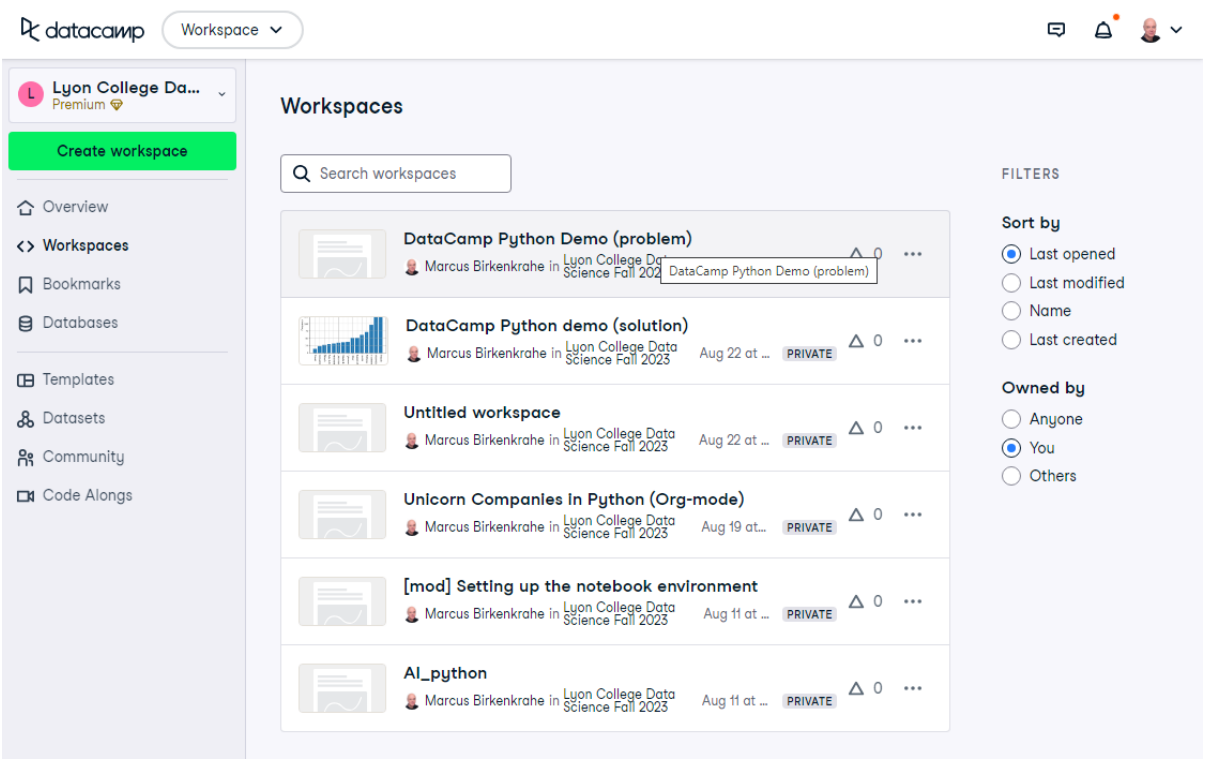
Practice - first R program



- In your browser, go to `replit.com`
- Register using your Lyon email
- Create a REPL using an R template

print("hello, world") data(mtcars) head(mtcars) plot(mtcars$mpg ~ mtcars$wt, main="Miles per gallon vs. weight in mtcars", xlab="Weight [tons]", ylab="Miles per gallon") abline(lm(mtcars$mpg ~ mtcars$wt),col="red")

Next: DataCamp Workspace



- Simple example: https://tinyurl.com/SimpleRexample
- Turn `mtcars` into code font and `view` the text cell.
- Load the `mtcars` data set and run the code cell.
- Print the first few lines of `mtcars`.
- Open the `help` for `mtcars`.
- Plot miles-per-gallon vs. weight for all cars.
- Compute and draw a red trendline through the points.

- Code:

```
data(mtcars)
head(mtcars)
#help(mtcars)
mpg = mtcars$mpg
wt = mtcars$wt
plot(x=wt, y=mpg)
abline(lm(mpg~wt),col="red")
```

Author: Marcus Birkenkrahe
Created: 2023-09-20 Wed 12:33