Course overview

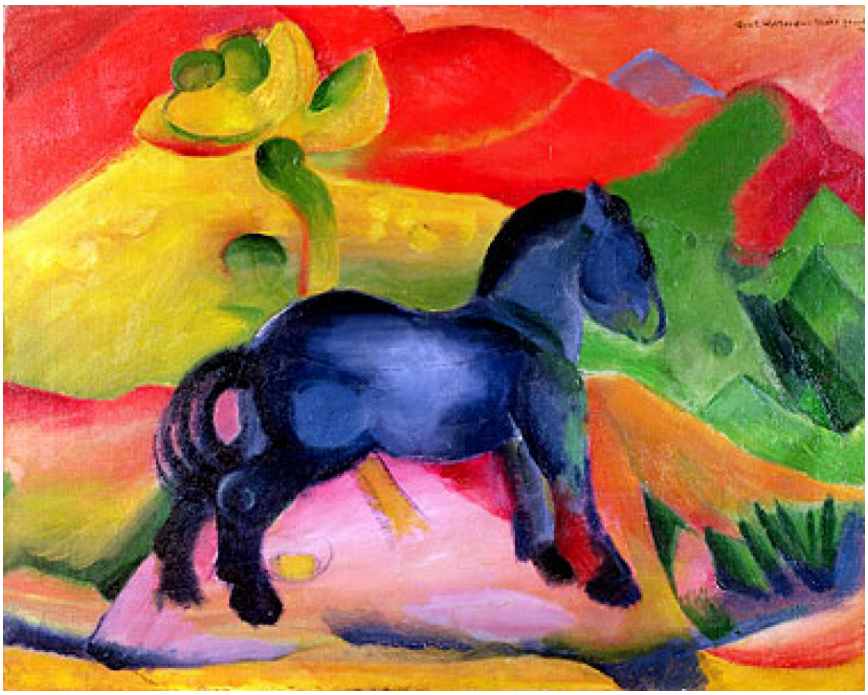
Introduction to data science (DSC 105) Fall 2024

Marcus Birkenkrahe

September 7, 2024

Contents

About me



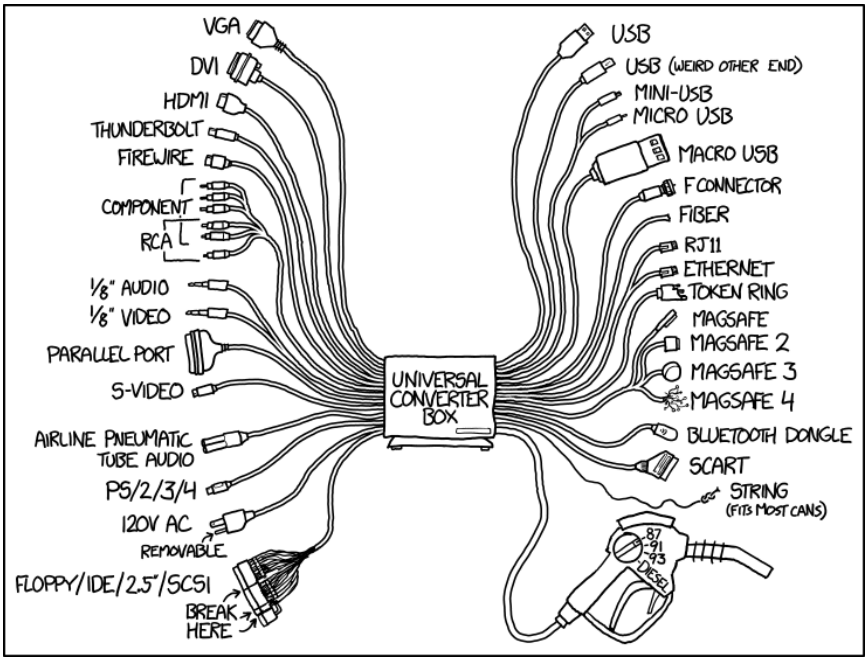
- PhD: theoretical particle physics (mathematical data science)
- Industry: Knowledge management & Cybersecurity
- Research: see researchgate.net, Google Scholar
- Profile: see ChatGPT's Aug 2024 memory of me:

You are a professor of computer and data science with a strong background in theoretical particle physics, holding a PhD with a focus on multigrid toolkits for lattice gauge theory simulations. You are passionate about teaching and strive to be the best instructor possible. Your expertise spans several programming languages, including C, C++, FORTRAN, R, and Python, and you are proficient with Linux Mint as your operating system.

In your teaching, you emphasize the importance of literate programming and use tools like Emacs, Org-mode, Jupyter notebooks, and GitHub to enhance learning experiences. You design courses that blend practical programming skills with theoretical knowledge, often focusing on data science and algorithms. You also engage with students through workshops and interactive sessions, encouraging them to explore programming as a fun and intellectually stimulating activity.

In addition to your academic pursuits, you are considering writing a book on teaching data science with literate programming and have plans to interview Donald Knuth on the subject. You prefer your content in Org-mode format for easy integration into your workflow.

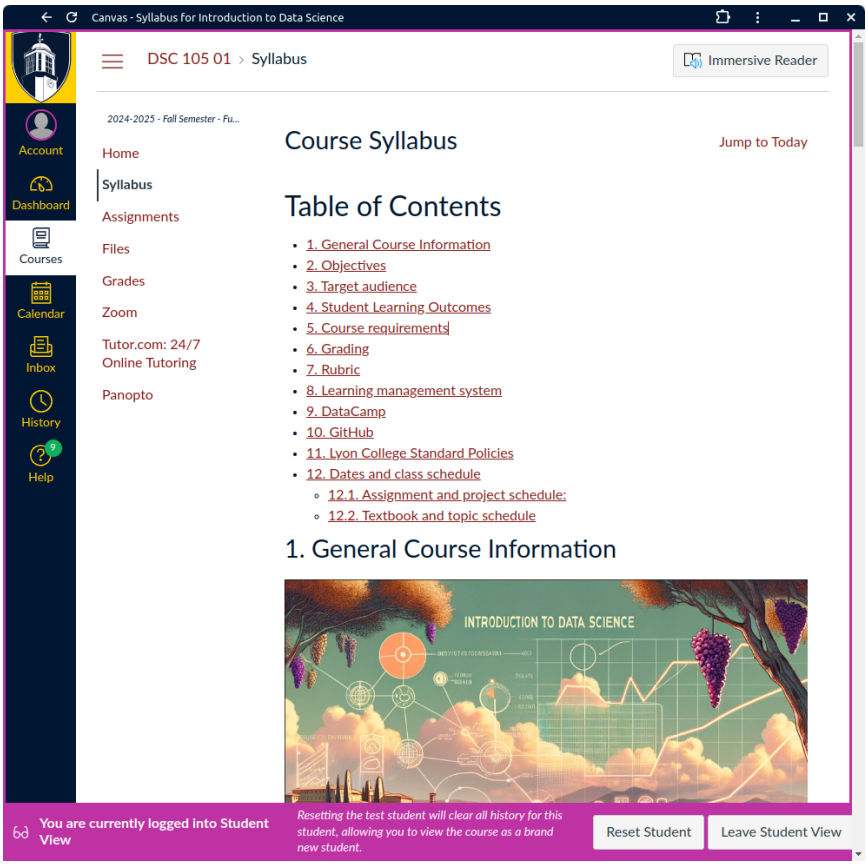
About you



1. Why are you here (*in this course*)?
2. What would delight you (*in this course*)?
3. What would disappoint you (*in this course*)?

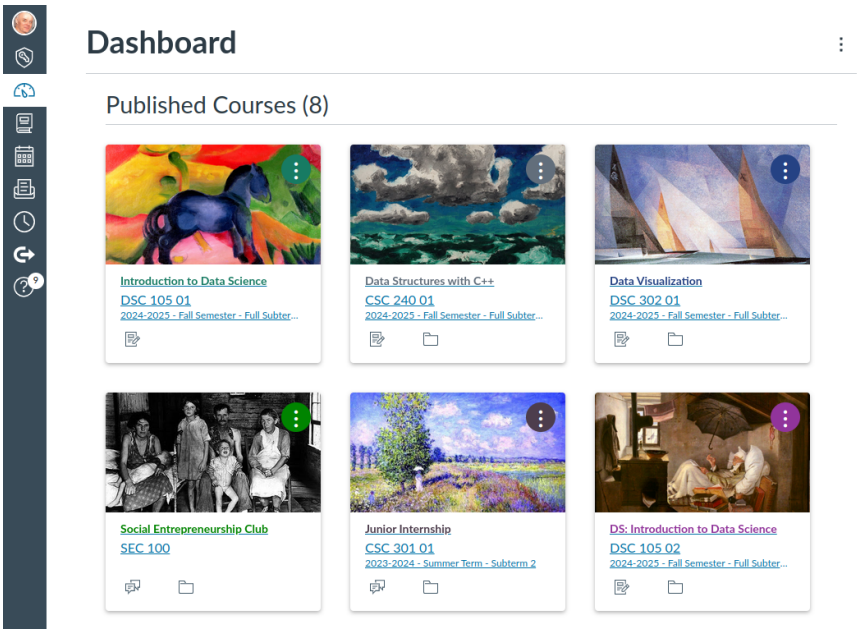
4. Where are you headed (*after this course*)?

Course syllabus



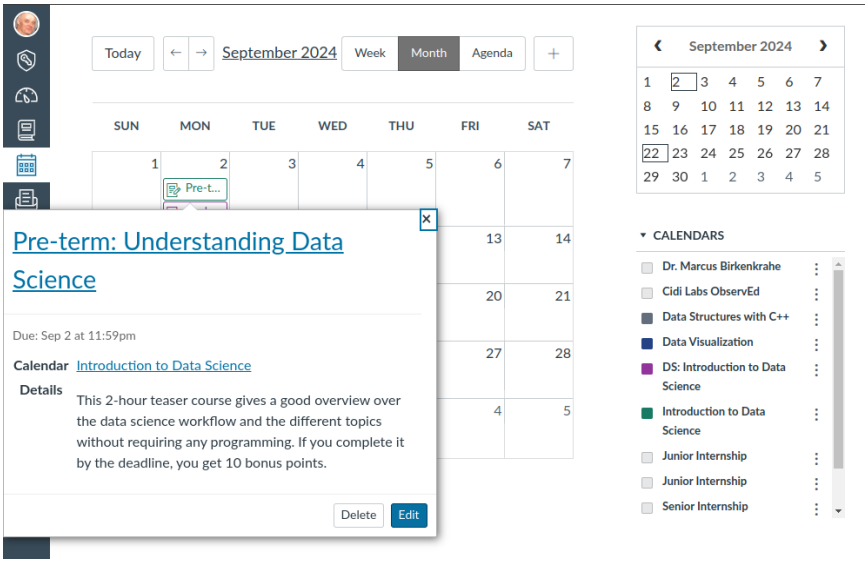
- Syllabus on Canvas
- General information & standard policies
- Course information (grading, schedule)
- **New:** About [not] Using AI to Write Code For You*

Canvas LMS



- All grades should be visible in the gradebook (with delays).
- Control your own notifications (especially email).
- Important course links on a page (see sidebar).

Canvas calendar

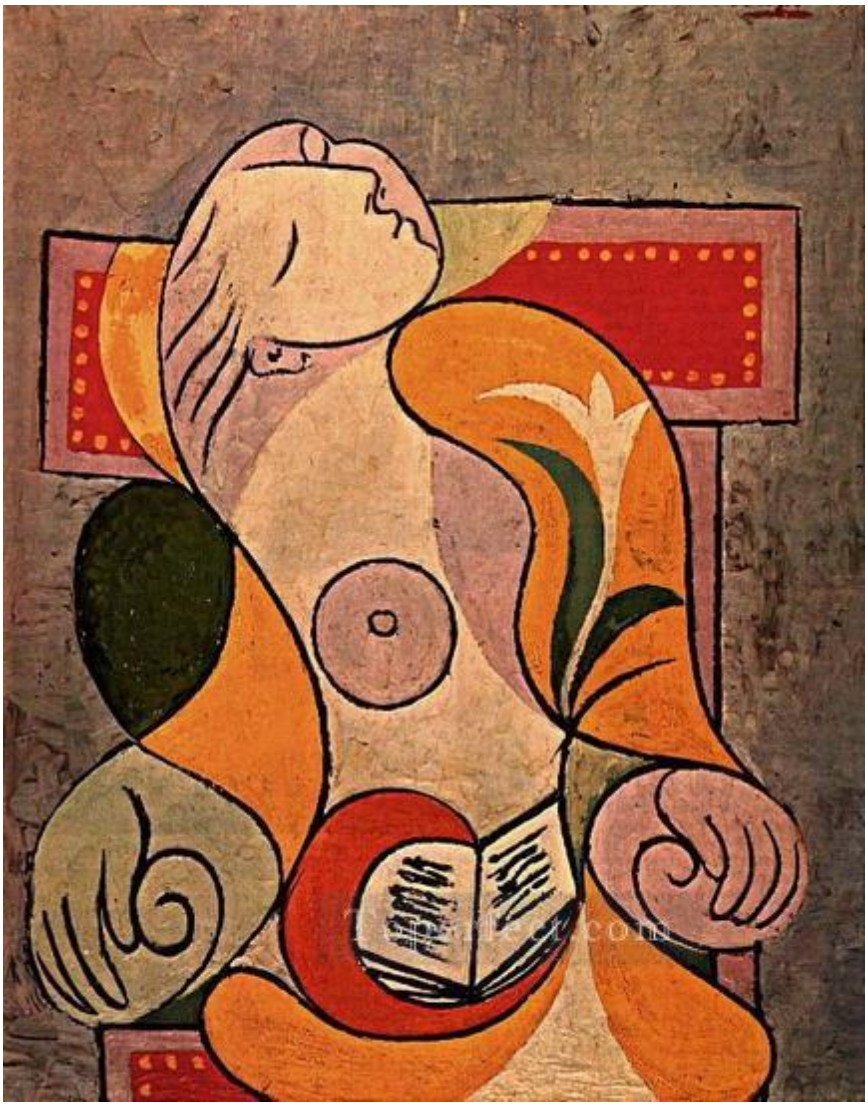


- Add all your Canvas calendars to Google from the Calendar page
- Activate all your courses on the Calendar page.
- Regularly check for upcoming deadlines.

Course topics

1. Getting started with literate programming
2. Introductory R and Python programming
3. Exploratory Data Analysis (EDA) using R (and Python)
4. Plotting data with base R, ggplot2 and matplotlib

Video lectures (old and new)



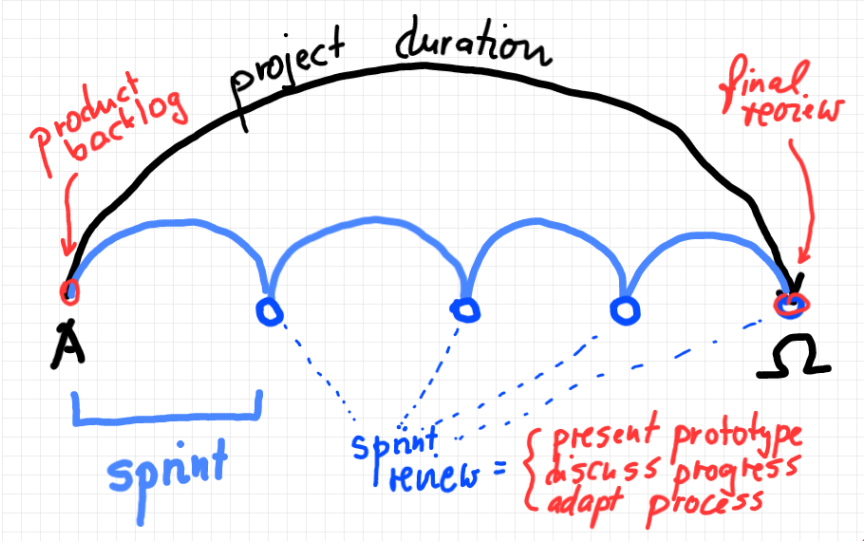
Taste of R: Lectures using the R programming language (2021-2024) - take a look at these, if you're bored with or put off by Python[1].

- Emacs + Org-mode + R (Tutorial videos Spring '22)

[1] There are good reasons to choose R over Python as a first language for data science students. Unfortunately, the dominance of the "Tidyverse" ideology negates this slight advantage. Still, for visualization and statistical analysis, R is still superior, IMHO.

- 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

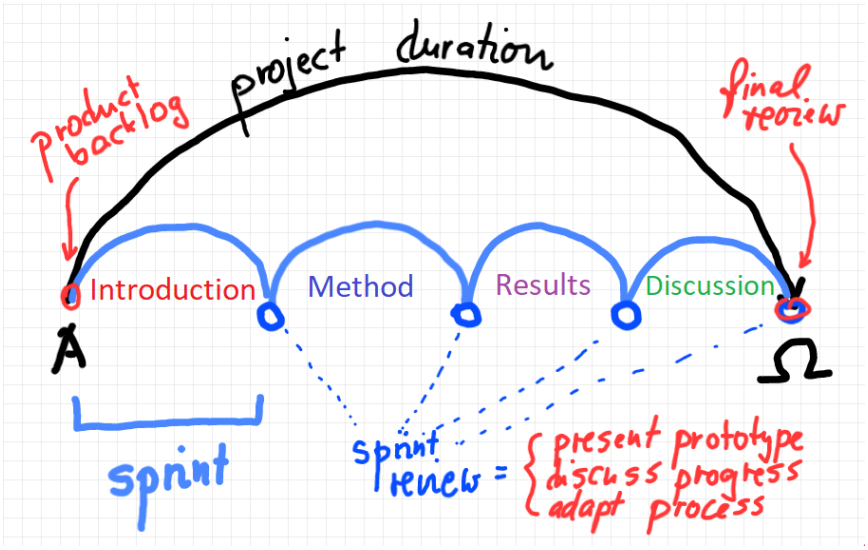
Agile project (with "Scrum")



- 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 four weeks already. Use it to present your initial results (see FAQ on what to deliver, and 1st sprint review).

IMRaD and Scrum

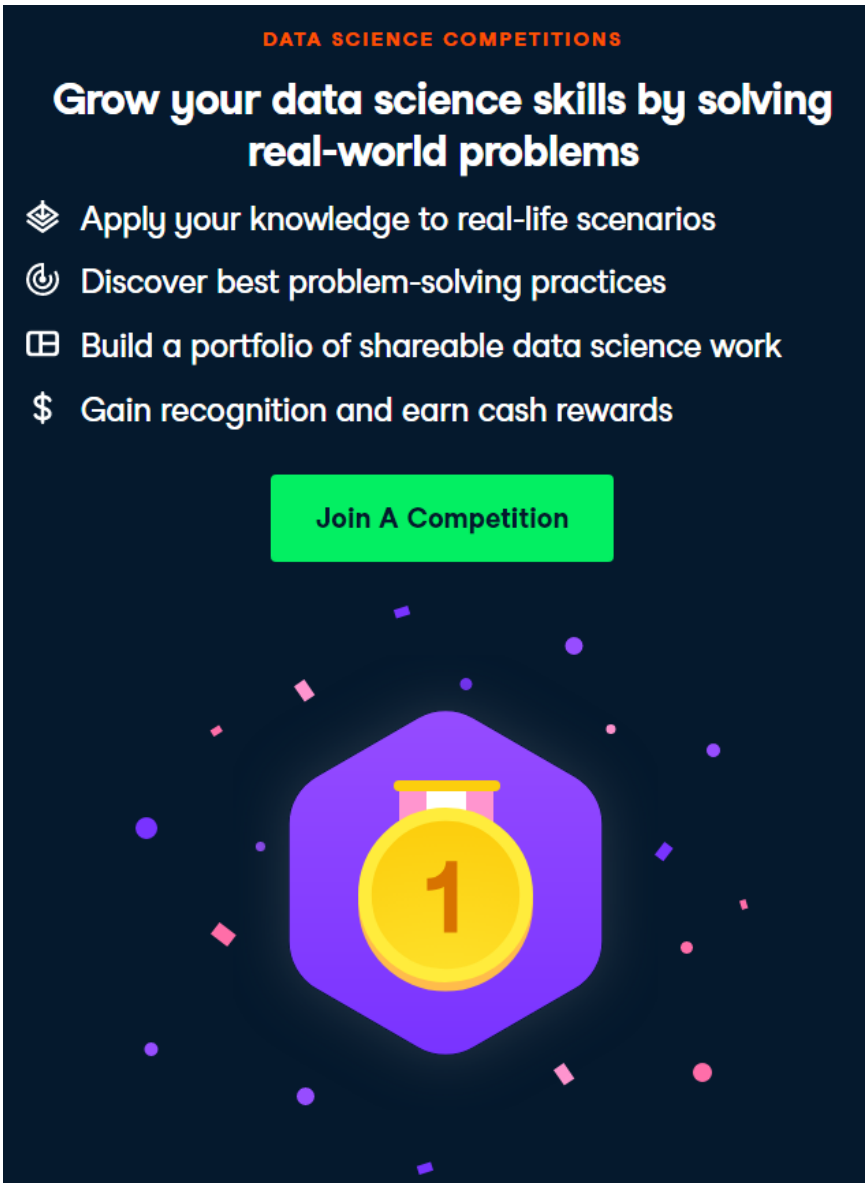


IMRaD is the framework for all scientific publications:

- 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

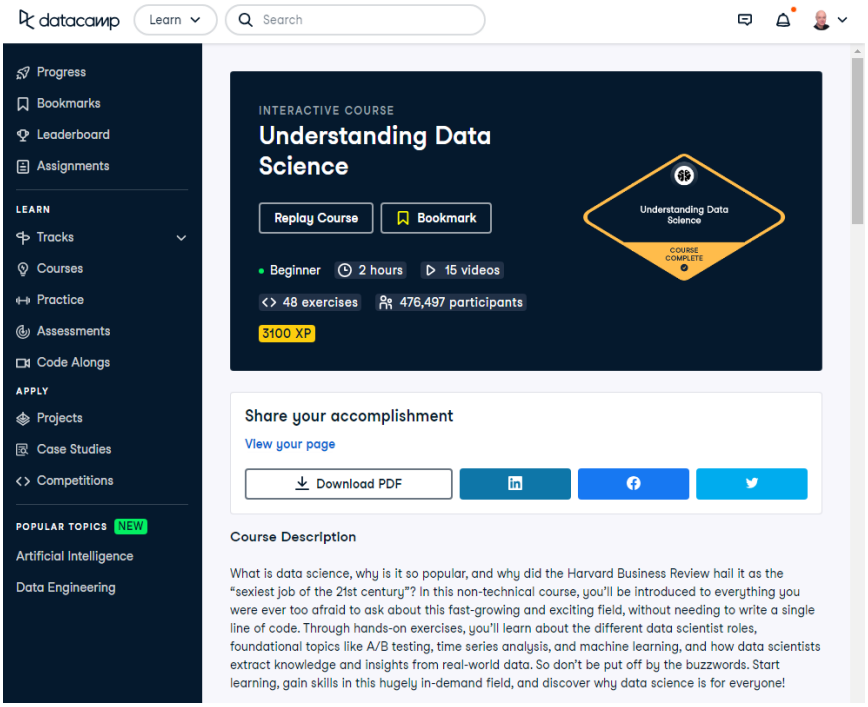


- Explore and document an R or Python package of your choice
- Document an extended analysis example (in R, Python, or

SQL[2])

- Explore a data set of your choice (e.g. soccer, finance, sales data)
- Complete a DataCamp competition and report on it!
- See DataCamp projects for more examplesx
- You can double/triple up on projects if you're in >1 of my courses[3]
- 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

[2]Or in another language: Julia, bash, or data analysis apps like PowerBi or Tableau come to mind - you can find examples at DataCamp.

[3]If you do use the same project topic in more than one of my courses, talk to me to make sure that the projects differ sufficiently.

- Access for you is free (academic alliance until end of the term[4])
- Most if not all term assignments are DataCamp assignments
- Assignments are drawn from several courses:
  1. Understanding data science (bonus, pre-term)
  2. Introduction to R
  3. Intermediate R
  4. Explorative Data Analysis with R
  5. Python for R Users
- Complete them on time to get full points (late submission: 50%)
- DataCamp certificates can support your resume (LinkedIn example)

[4] If you wish to use DataCamp beyond the end of the term, contact me and I can add you to next term's workspace.

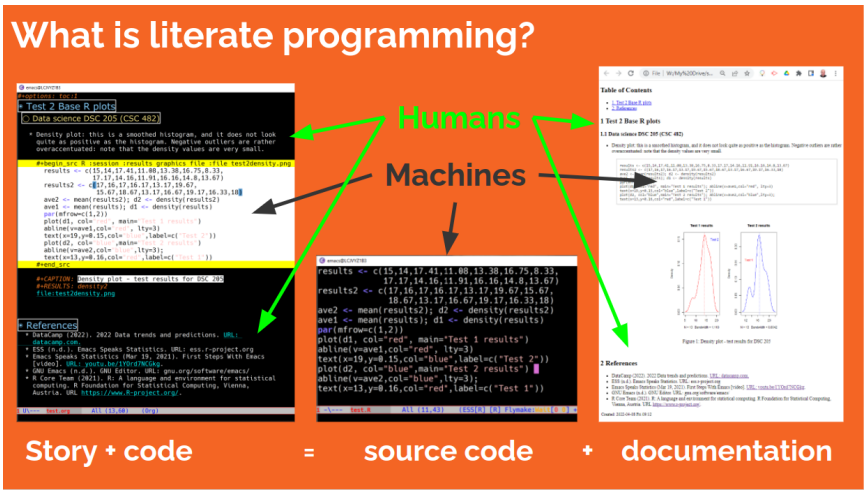
Good-bye Jupyter, Hello (again) Emacs + Org-mode!

You will learn to use Emacs and the Org-mode extension to master literate programming.

Attributes:

- Emacs: self-documenting, extensible *FOSS* text editor
- Process, file and package management (like an OS)
- *Literate programming* environment for $>$ 45 languages
- *IDE* for R programming and *REPL* for interactive coding
- Must use to mix different languages in one notebook
- Helps to understand and master data science infrastructure

What is literate programming?



Paper: "Teaching Data Science with Literate Programming Tools" (2023)

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

Using GitHub

- GitHub is the top software development platform now owned (but not ruined yet) by Microsoft.
- The term 'Git' refers to the version control software of the same name. You can also use it as a central document repository.
- All course materials (data/src/org/pdf/img) are on GitHub at github.com/birkenkrahe/ds1. You can watch/fork this repository.
- You can use GitHub yourself to synchronize content between different computers, e.g. the lab and your PC.
- Recommended: the Hello World exercise at GitHub. This teaches you the typical software project development workflow.

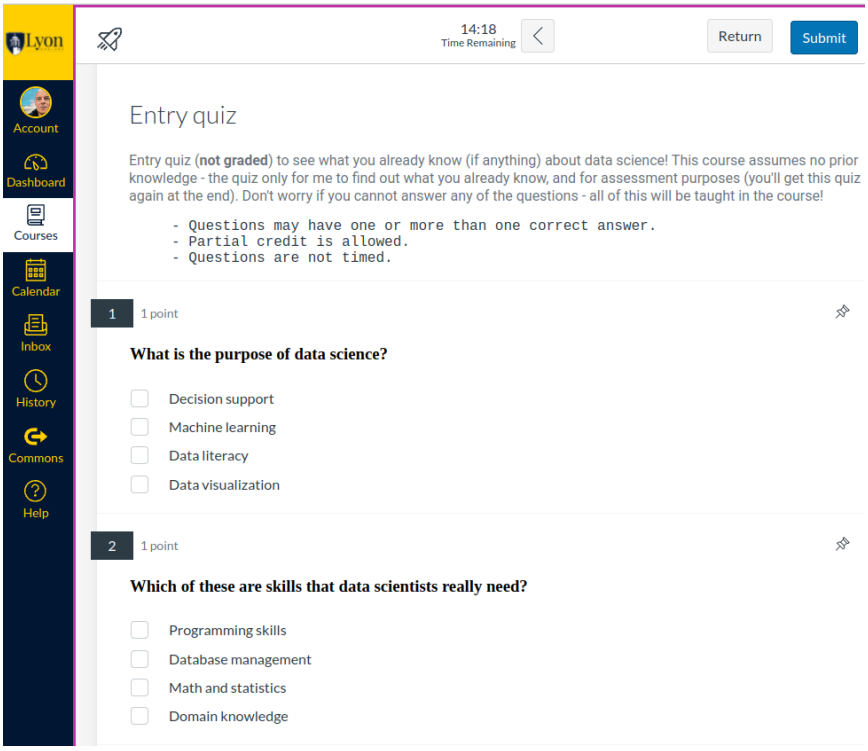
Using Linux

Linux is the world's most used operating system (the software between you and your computer, or your phone) based on 1970s Unix developed by Linus Torvalds (who also wrote the Git version control software).

With some luck, we're going to get lots of Linux love this term: IT is working on a server with 50 virtual Linux boxes just for you and me!

While they sort us out, we're going to take a look at Google Cloud Shell, which is free for the time being and will allow us to spin up Linux + Emacs + R.

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 several multiple choice, matching and true/false questions.
- A subset of the test questions will form the final exam (25% of your final grade) - the exam is optional for you to improve your grade.

First home assignments

- Register with DataCamp now if you haven't done it yet (links).
- Complete chapter 1 of the course "Understanding Data Science", Introduction to data science on the DataCamp platform.

Next: Infrastructure Exercise