

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:



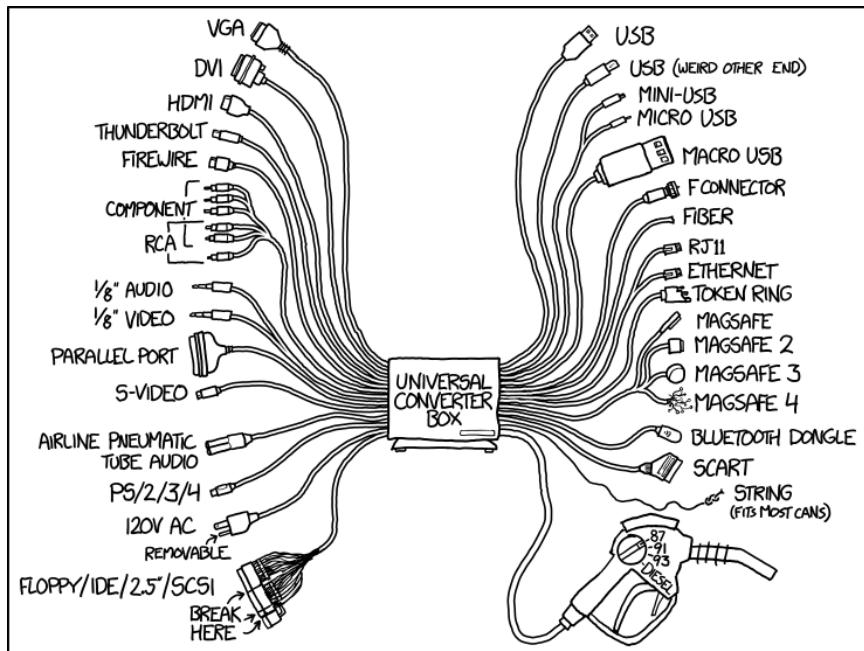
You are a professor of computer and data science with a strong background in theoretical particle physics, holding a PhD with a focus on multigrid toolkits for lattice gauge theory simulations. You are passionate about teaching and strive to be the best instructor possible. Your expertise spans several 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

The screenshot shows the Canvas Syllabus page for the course DSC 105 01. The left sidebar has a dark theme with icons for Account, Dashboard, Courses, Calendar, Inbox, History, and Help. The main content area has a light theme. The title is "Course Syllabus" and the subtitle is "Table of Contents". The table of contents lists 12 sections: 1. General Course Information, 2. Objectives, 3. Target audience, 4. Student Learning Outcomes, 5. Course requirements, 6. Grading, 7. Rubric, 8. Learning management system, 9. DataCamp, 10. GitHub, 11. Lyon College Standard Policies, 12. Dates and class schedule, 12.1 Assignment and project schedule, and 12.2 Textbook and topic schedule. Below the table of contents is a section titled "1. General Course Information" with a decorative background image of a tree and data charts. At the bottom, there is a note about resetting the student view and two buttons: "Reset Student" and "Leave Student View".

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

Canvas LMS

The screenshot shows the Canvas LMS dashboard with a sidebar on the left containing icons for user profile, notifications, calendar, grades, assignments, and help. The main area is titled "Published Courses (8)" and displays eight course cards in a grid:

- Introduction to Data Science**
DSC 105.01
2024-2025 - Fall Semester - Full Subterm...
View Details | View Syllabus
- Data Structures with C++**
CSC 240.01
2024-2025 - Fall Semester - Full Subterm...
View Details | View Syllabus
- Data Visualization**
DSC 302.01
2024-2025 - Fall Semester - Full Subterm...
View Details | View Syllabus
- Social Entrepreneurship Club**
SEC 100
View Details | View Syllabus
- Junior Internship**
CSC 301.01
2023-2024 - Summer Term - Subterm 2
View Details | View Syllabus
- DS: Introduction to Data Science**
DSC 105.02
2024-2025 - Fall Semester - Full Subterm...
View Details | View Syllabus

- 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

The screenshot shows the Canvas calendar interface. On the left, there's a sidebar with icons for Home, Profile, Notifications, Courses, and Help. The main area has tabs for Today, September 2024 (which is selected), Week, Month (selected), Agenda, and a plus sign. Below these are buttons for SUN through SAT. A specific event is highlighted: "Pre-term: Understanding Data Science" on Sep 2 at 11:59pm, part of the "Introduction to Data Science" calendar. The event details mention a 2-hour teaser course overview. To the right is a monthly calendar for September 2024, showing days from 1 to 30. At the bottom right of the calendar is a "CALENDARS" section with a dropdown menu containing several course names.

SUN	MON	TUE	WED	THU	FRI	SAT
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

CALENDARS

- Dr. Marcus Birkenkrahe
- Cidi Labs ObservEd
- Data Structures with C++
- Data Visualization
- DS: Introduction to Data Science
- Introduction to Data Science
- Junior Internship
- Junior Internship
- Senior Internship

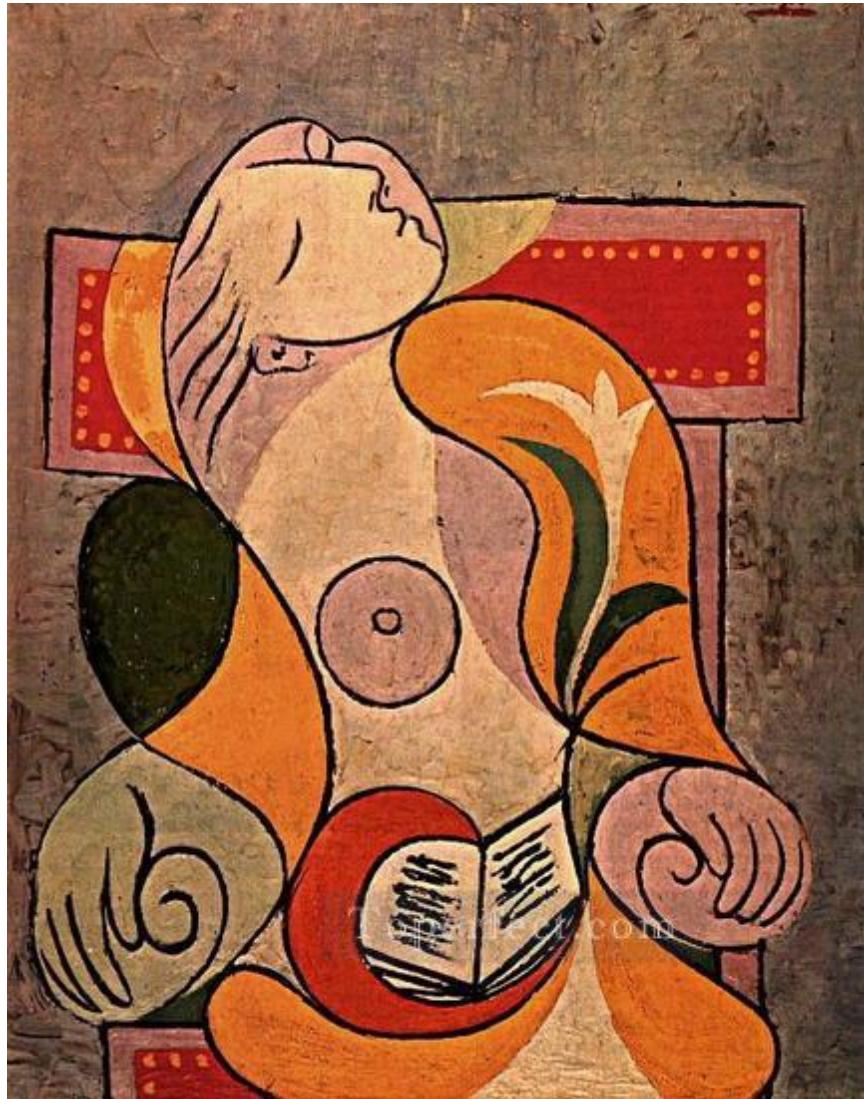
- Add all your Canvas calendars to Google from the Calendar page
- Activate all your courses on the Calendar page.
- Regularly check for upcoming deadlines.

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)



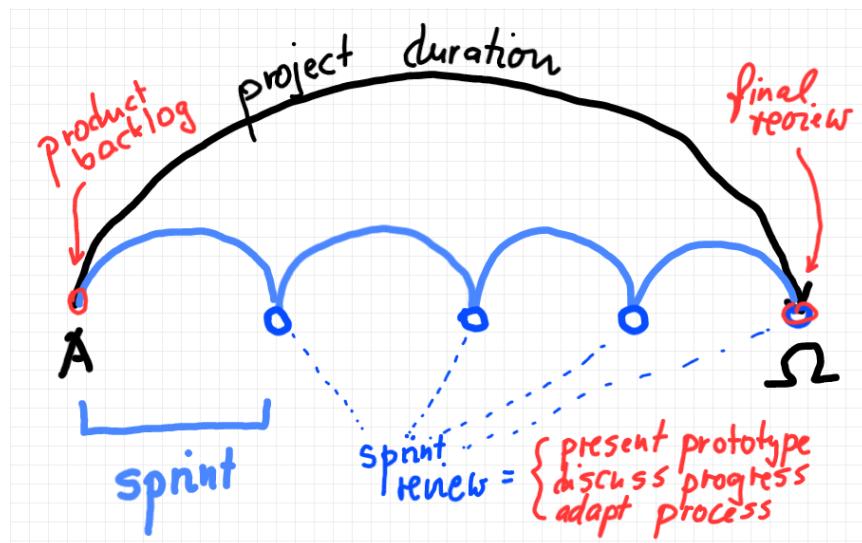
Taste of R: Lectures using the R programming language (2021-2024) - take a look at these, if you're bored with or put off by Python¹.

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

¹There 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

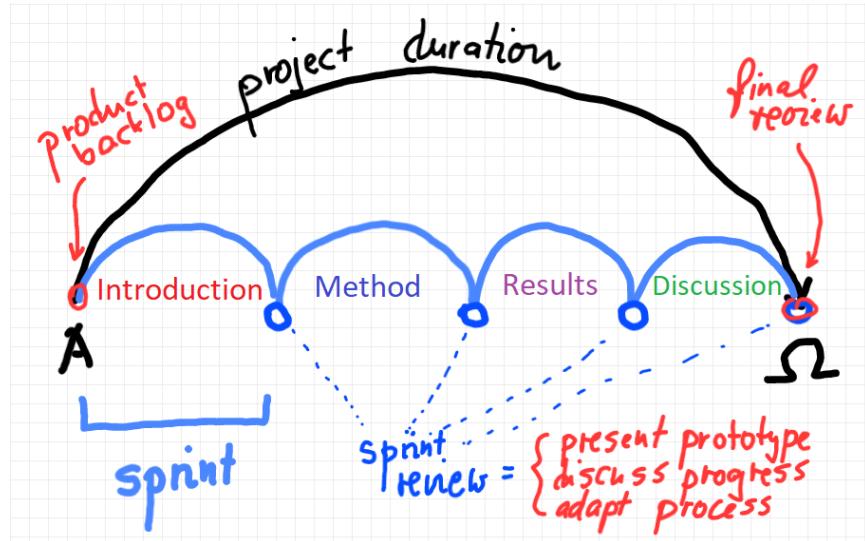
Agile project (with "Scrum")



- 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

DATA SCIENCE COMPETITIONS

Grow your data science skills by solving real-world problems

- ❖ Apply your knowledge to real-life scenarios
- ⌚ Discover best problem-solving practices
- ▣ Build a portfolio of shareable data science work
- \$ Gain recognition and earn cash rewards

[Join A Competition](#)



A purple hexagonal badge containing a gold medal with the number '1' on it, surrounded by a ribbon. The badge is set against a dark background with scattered small, colorful dots (purple, pink, yellow) resembling stars or confetti.

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

SQL²)

- 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 courses³
- Use problems from other courses for your project, e.g. data collected by yourself, or data in economics, business, art etc.

Introduction to DataCamp

The screenshot shows the DataCamp website interface. On the left, there's a sidebar with navigation links: Progress, Bookmarks, Leaderboard, Assignments, Learn (Tracks, Courses, Practice, Assessments, Code Alongs), Apply (Projects, Case Studies, Competitions), and Popular Topics (Artificial Intelligence, Data Engineering). The main content area displays an 'INTERACTIVE COURSE' titled 'Understanding Data Science'. It features a 'Replay Course' button, a 'Bookmark' button, and course statistics: Beginner, 2 hours, 15 videos, 48 exercises, 476,497 participants, and 3100 XP. To the right of the course title is a yellow diamond-shaped badge with the text 'Understanding Data Science' and 'COURSE COMPLETE'. Below the course details is a section titled 'Share your accomplishment' with a 'View your page' link and social sharing buttons for LinkedIn, Facebook, and Twitter. At the bottom is a 'Course Description' section with a paragraph of text.

- DataCamp is a data science learning platform

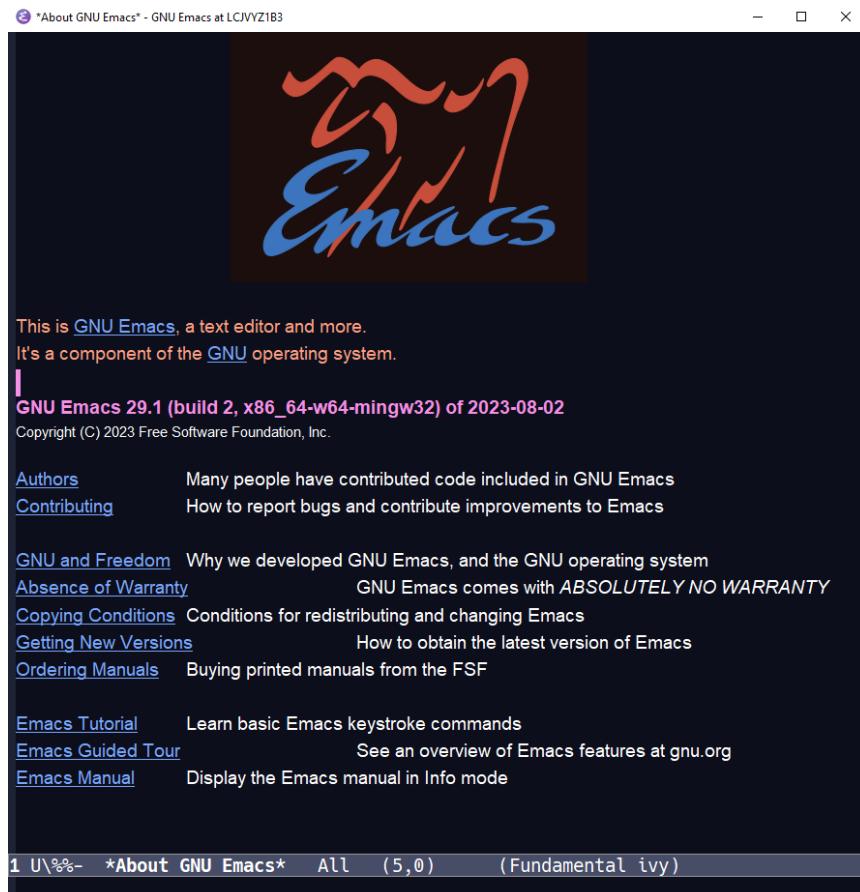
²Or in another language: Julia, bash, or data analysis apps like PowerBi or Tableau come to mind - you can find examples at DataCamp.

³If 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 term⁴)
- 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)

⁴If 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!

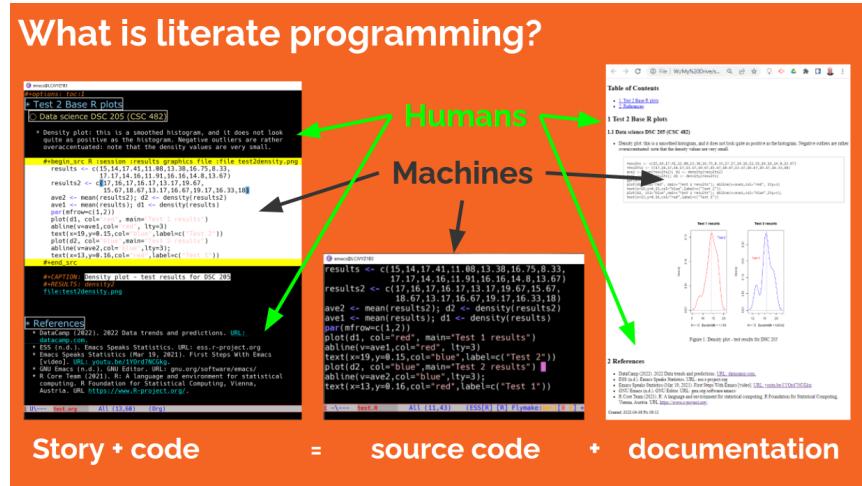


You will learn to use Emacs and the Org-mode extension to master literate programming.

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)

The screenshot shows a quiz interface titled "Entry quiz". The left sidebar contains a yellow header with "Lyon" and a user icon, followed by a vertical menu with icons for Account, Dashboard, Courses, Calendar, Inbox, History, Commons, and Help. The main area displays two questions:

1 1 point

What is the purpose of data science?

Decision support
 Machine learning
 Data literacy
 Data visualization

2 1 point

Which of these are skills that data scientists really need?

Programming skills
 Database management
 Math and statistics
 Domain knowledge

At the top right, there are buttons for "Return" and "Submit". The top center shows the time as "14:18" and "Time Remaining".

- 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