COURSE OVERVIEW

Applied math for data science (DSC 482/MTH 445) Fall 2022

Table of Contents

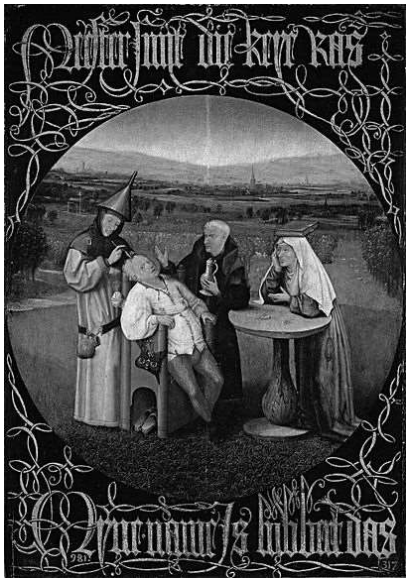

Figure 1: Cutting the stone/cure of folly, Hieronymous Bosch (1494)

Cutting the Stone, also called The Extraction of the Stone of Madness, or The Cure of Folly, is a painting by Hieronymous Bosch from 1494. The painting depicts a surgeon, wearing a funnel hat, removing the stone of madness from a patient's head by trepanation. The stone (Dutch: *kei*) appears as a flower bulb. The Gothic inscription reads: *Master, cut the stone out, fast. My name is Lubbert Das*. (Source: Wikipedia)

1 MUTUAL INTRODUCTIONS

Figure 2: Detail from the Garden of Earthly Delights (Hell) by Bosch (1503)

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

2 COURSE SYLLABUS (on GitHub and on Canvas)

Figure 3: DSC 302 Syllabus on GitHub

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

3 COURSE TOPICS (ILLUSTRATED)

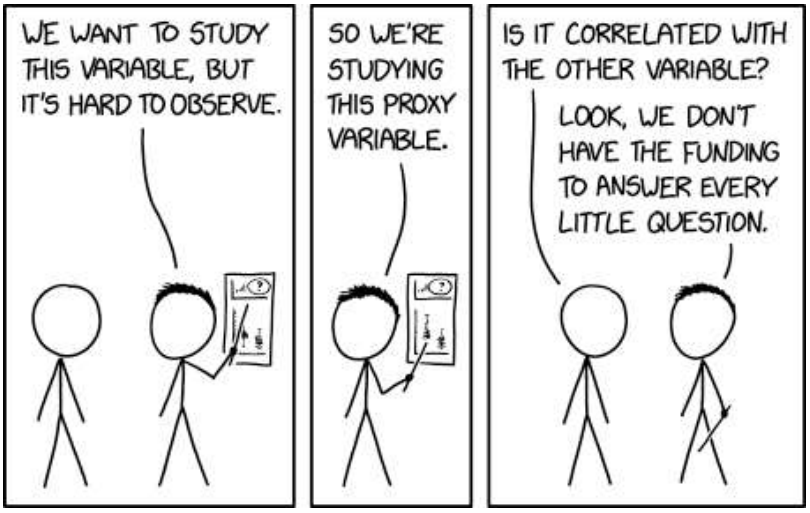

Figure 4: xkcd 2652: Proxy Variable (explainxkcd.com, July 29, 2022)

Title: 'Our work has produced great answers. Now someone just needs to figure out which questions they go with." In statistics, a proxy variable is used as a stand-in for one or more other variables that are difficult to measure. In order to be useful as such, proxy variables must be correlated with what they are intended to represent. For example, a drug might aim to reduce deaths from a slow-acting disease. But testing if it reduces deaths might take many years, so researchers might test for a proxy outcome instead, like whether it results in loss of bone density or damage to cells. Physicians use blood pressure as one of many proxies for cardiovascular health. Source: explanation (July 29, 2022)

4 COURSE TOPICS (SPELLED OUT)

Figure 5: Pablo Picasso, Guernica (1937), grayscale painting

1. Elementary statistics
2. Basic data visualization
3. Probability
4. Distributions

5. Linear regression

Guernica by Picasso (1937) depicts the bombing on the Spanish city of Guernica during the Spanish Civil War. It is also a typical cubist-surrealist painting where reality is dissolved in geometric patterns and symbols, distorted and disfigured to achieve a heightened effect. Our upcoming discussion on the origins of statistics and probability will make it clearer why I chose this image to illustrate the list of topics.

5 WHY "MATH FOR DATA SCIENCE"?

Figure 6: Joseph Wright of Derby: An Experiment on a Bird in an Air Pump (1768)

- The central purpose of data science is *pattern identification*
- Patterns cannot usually be discerned directly from the data
- In experiments, data must be sampled and analyzed
- Data analysis is primarily a mathematical discipline
- Core: statistical summaries and probability distributions
- Alternative name: causal inferential statistics

- Distinct from: machine learning, big data (massive datasets)

  "The painting departed from convention of the time by depicting a scientific subject in the reverential manner formerly reserved for scenes of historical or religious significance. Wright was intimately involved in depicting the Industrial Revolution and the scientific advances of the Enlightenment." (Source: Wikipedia)

6 THE WHOLE DATA SCIENCE PIPELINE

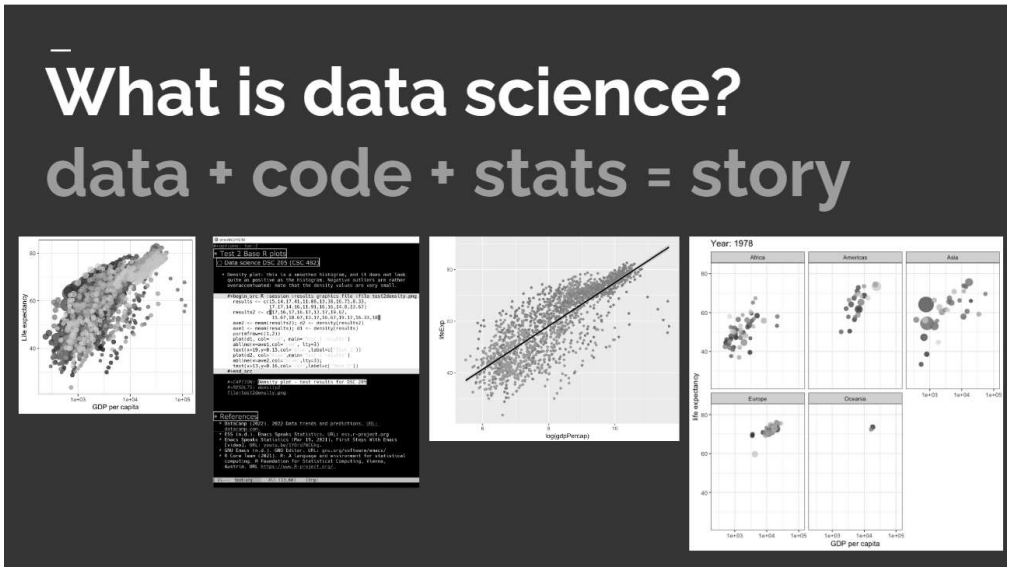

Figure 7: Data science pipeline

- Math important for: data cleaning and data modeling
- Need to know: coding (R programming) and story (visualization)
- This course focuses on descriptive rather than predictive stats
- The followup course ("Machine learning") is about prediction
- Missing (among other things): measure theory (project?)[1]

7 AGILE TEAM PROJECT

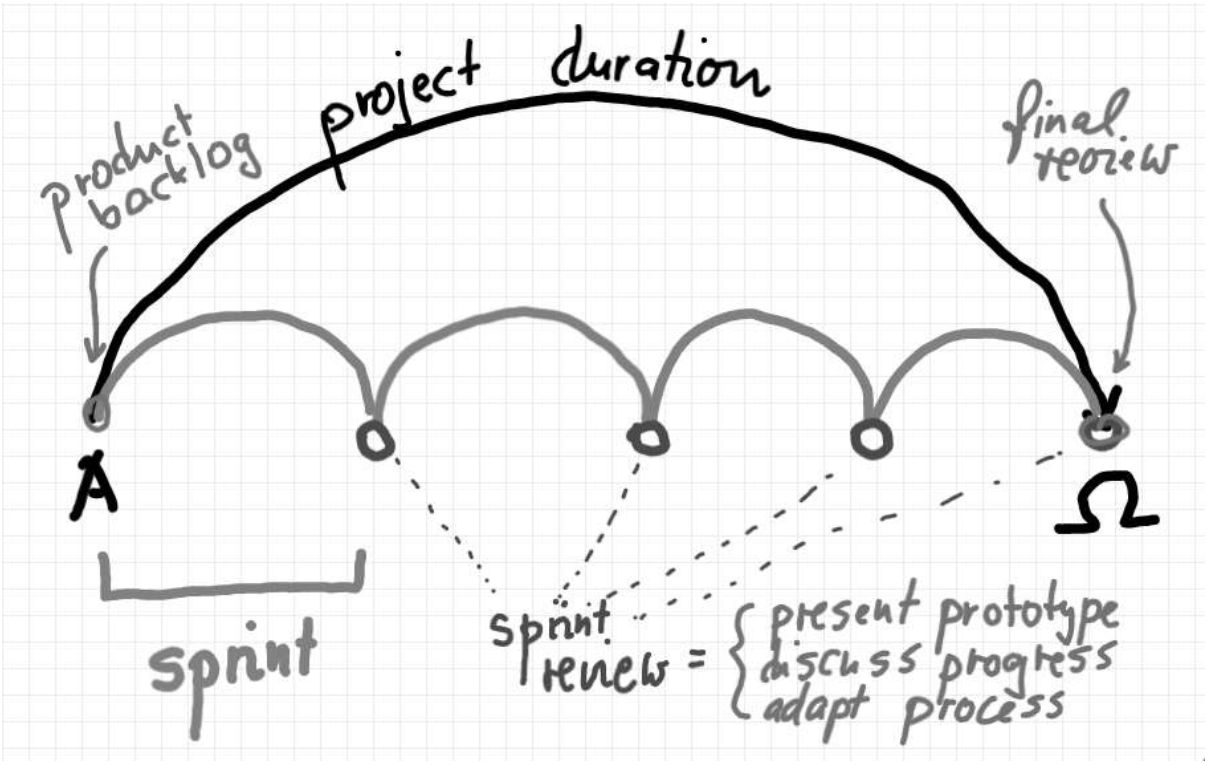

Figure 8: Agile (Scrum) project

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

8 MANY PROJECT OPPORTUNITIES

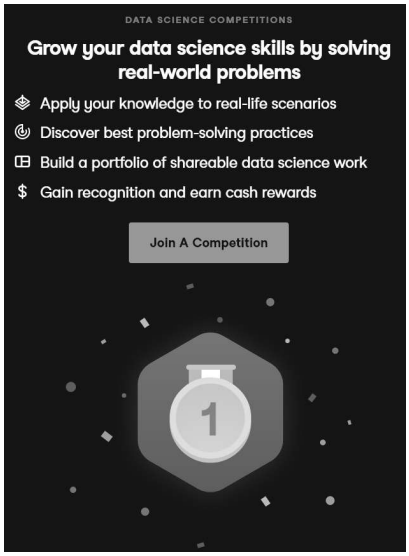

Figure 9: DataCamp competition announcement

- Analyze an interesting data visualization (explore math content)
- Explore a statistical package or platform (e.g. SPSS, MATLAB[2])
- Explore an R package (e.g. `data.table`[3], MASS)
- Solve a real-world problem (you can decide how much math you need)
- See DataCamp projects for examples (the math is often missing)
- Example: visualize whale song / double up between 2 or 3 courses
- Deepen any of our topics with current or classic scientific papers
- Deepen any of the course topics with a detailed applied example[4]
- If you can use, topics, experiments etc. from other courses!

9 INTRODUCTION TO DataCamp

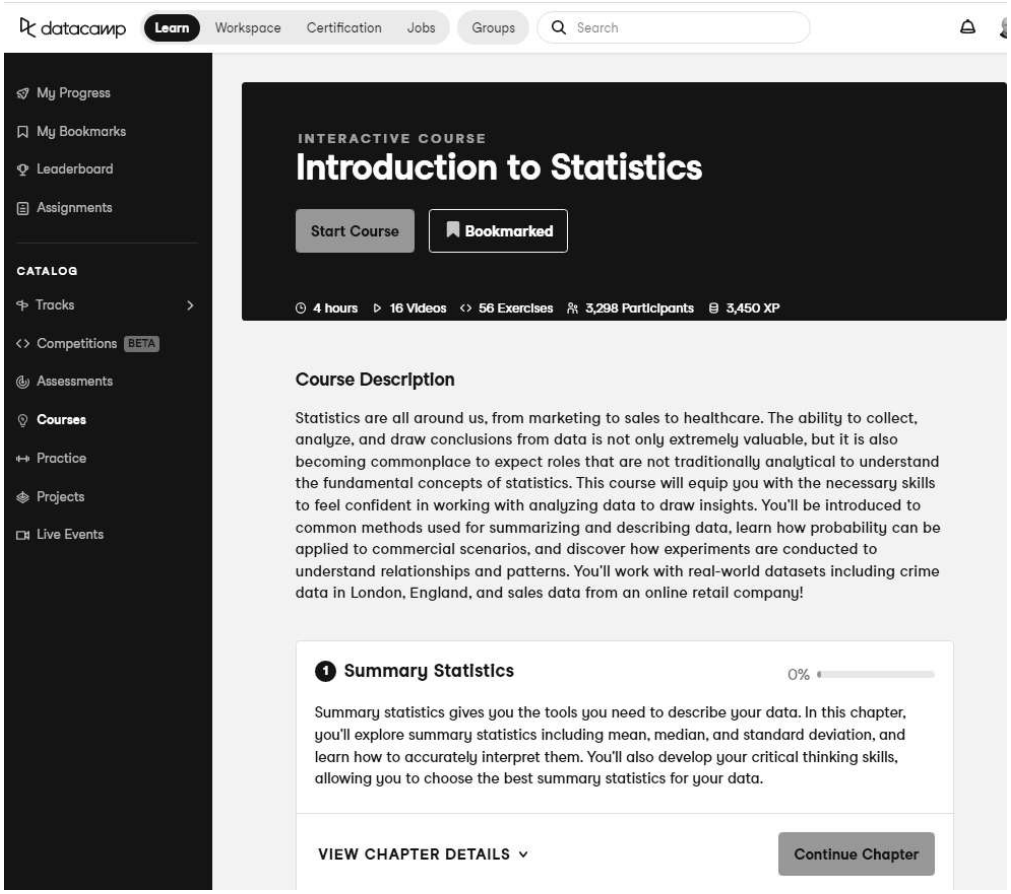

Figure 10: DataCamp course "Introduction to statistics" start page

- **DataCamp** is a data science learning platform
- Access for you is **free** (classroom license)
- 14/15 assignments are DataCamp assignments
- Assignments are drawn from 3 courses
    1. Introduction to statistics (4/4)
    2. Introduction to statistics in R (4/4)
    3. Foundations of probability in R (4/4)
    4. Introduction to regression in R (2/4)
- Complete them on time to get full points
- Completed DataCamp courses can support your resume

10 INTRODUCTION TO THE TEXTBOOK

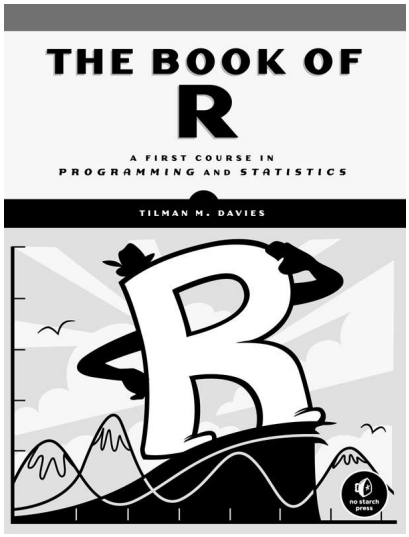

Figure 11: Cover of Book of R (Davies, 2016)

- R is *FOSS* with focus on stats and graphics
- Davies' "Book of R" is extensive (832p.) => library
- We will (hopefully) cover most of Part III (ca. 120 p.)
- You don't have to read along but it might help

11 OTHER SOURCES

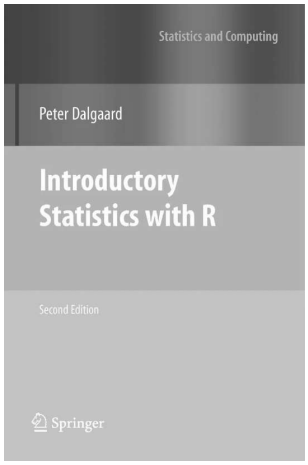

Figure 12: Peter Dalgaard, Introductory Statistics with R (2008)

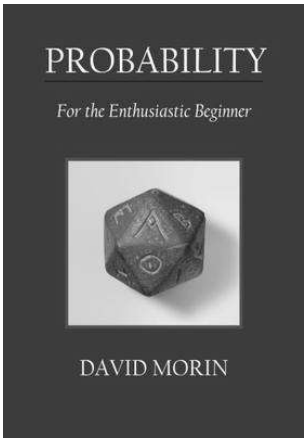

Figure 13: David Morin, Probability for the enthusiastic beginner (2016)

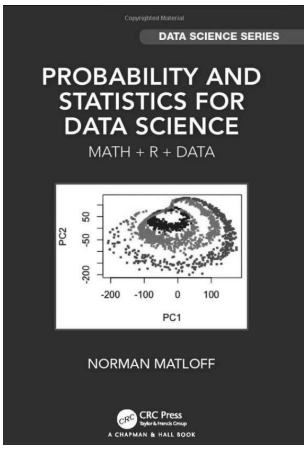

Figure 14: Norman Matloff, Probability and Statistics for Data Science (2020)x

- Matloff, Probability & statistics for data science (2020) => library
- Good (free) short online tutorial for R: Matloff's "fasteR"
- Beware of ideologies in science(cp. Matloff's "TidyverseSceptic")

12 INTRODUCTION to GNU Emacs + ESS + Org-mode

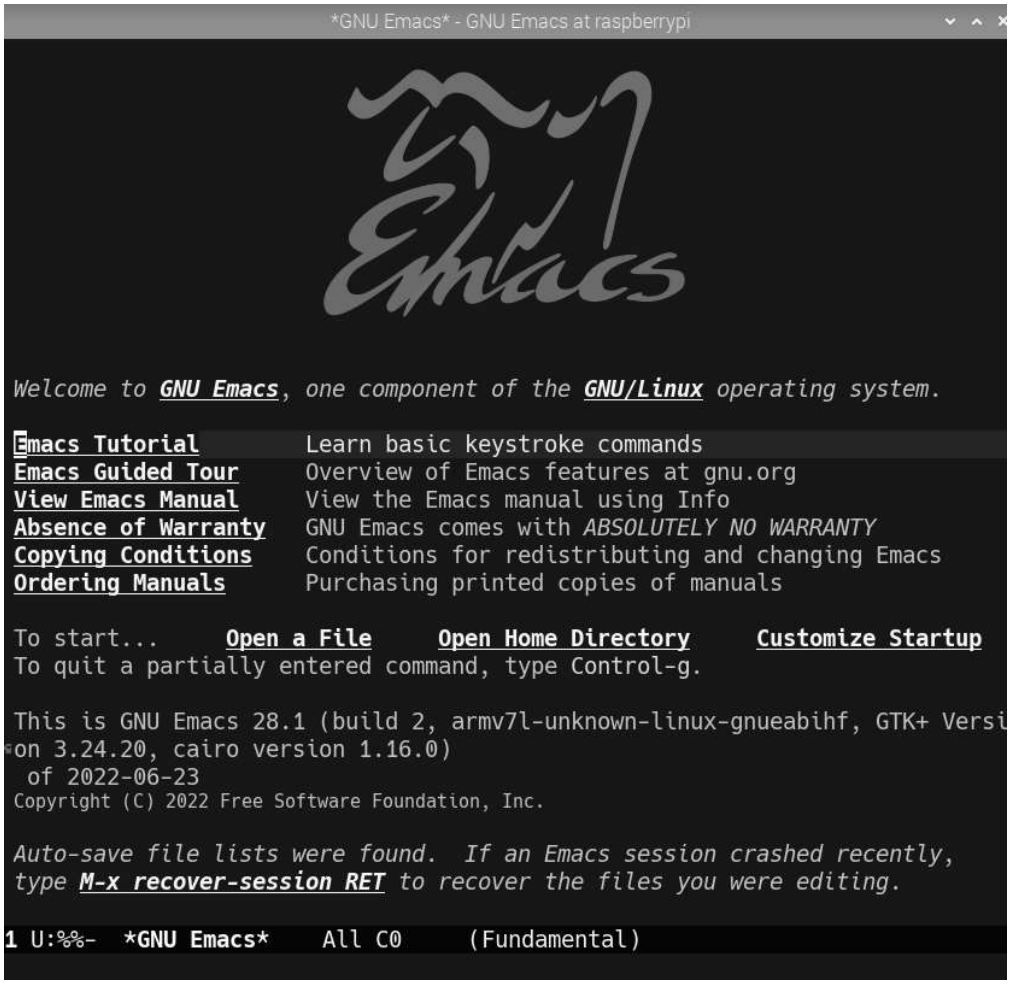

Figure 15: GNU Emacs start page

- 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

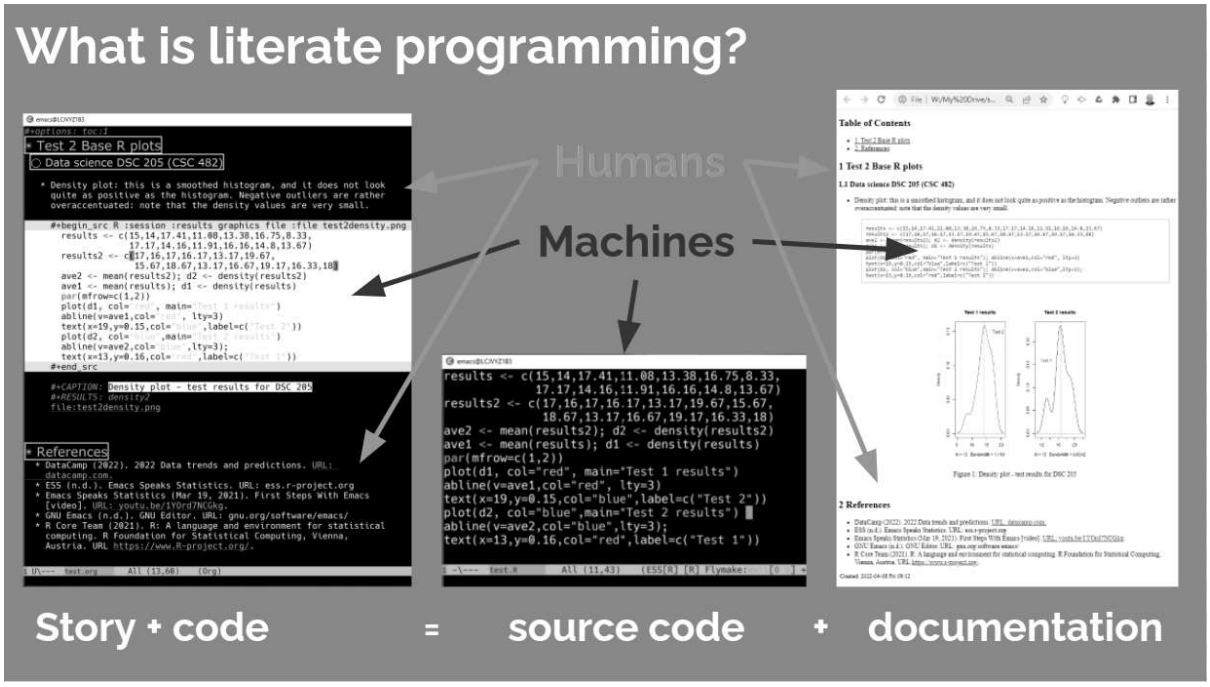

Figure 16: What is literate programming?

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.

1. Complete the Emacs on-board tutorial and upload an edited copy to Canvas by Thursday, 25 August at 8 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
2. Register with DataCamp and complete the DataCamp chapter "Summary statistics" from the course "Introduction to statistics" by Tuesday, 30 August at 8 am.
   - Motivating summary statistics
   - Mean, median, standard deviation
   - Interpretation of statistical summaries

15 TESTS (NOT GRADED)

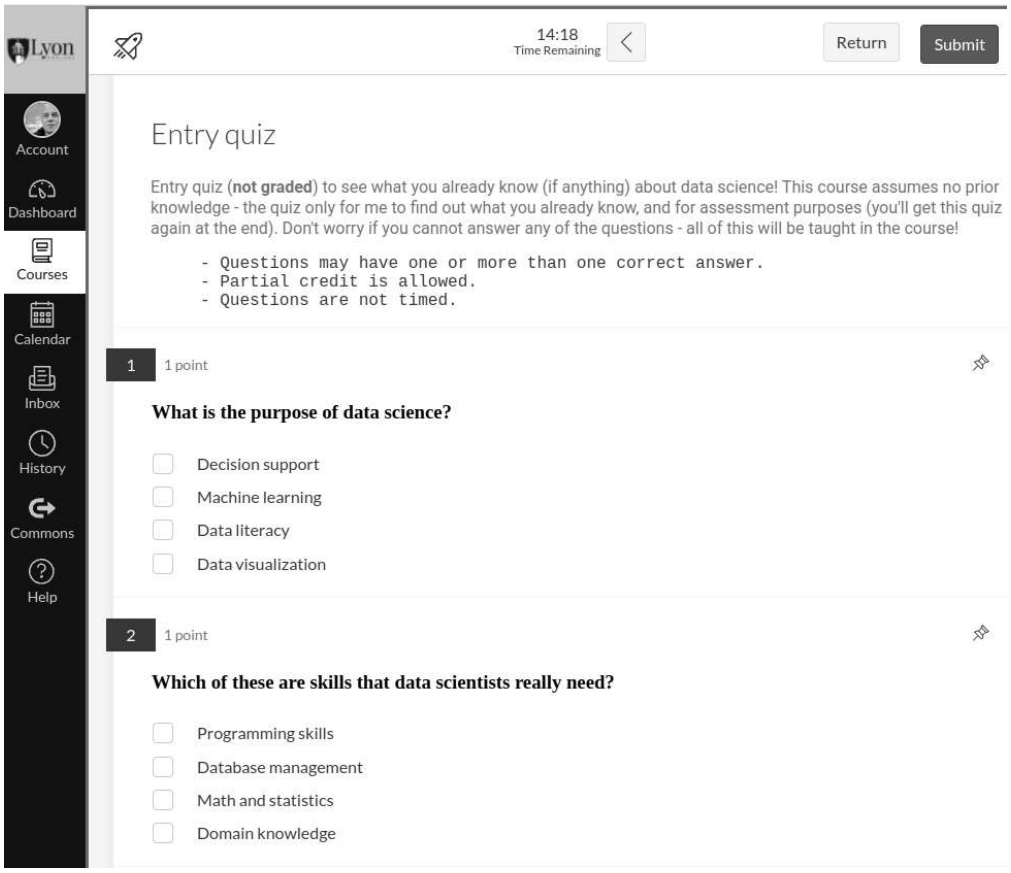

Figure 17: Start page of the entry quiz on Canvas

- 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 GLOSSARY

TERM	MEANING
Proxy variable	Observable stand-in for the real thing
Enlightenment	17th/18th century cultural movement
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

TERM	MEANING
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

17 REFERENCES

- Davies T D (2016). The Book of R. NoStarch Press.
- Dalgaard P (2008). Introductory Statistics with R. Springer.
- Matloff N (2020). Probability and stats for data science. CRC Press.
- Matloff N (2022). fasteR: fast Lane to Learning R! Github.
- Morin D (2016). Probability For the Enthusiastic Beginner. Harvard.

Footnotes:

[1] A measure assigns a probability to sets of events where each individual event has zero probability so that expectations for continuous random variables can be defined. This mini lecture (Lawrence, 2012) from an advanced probability seminar addresses answers the question why measure theory is needed here.

[2] Both of these are commercial, but there are other languages and platforms, e.g. Tableau (also featured on DataCamp), or GNU Octave.

[3] This is a great package whose abilities will remind those of you with SQL knowledge of the database course. To learn more about it, the DataCamp course is a good starting point.

[4] Time series is the example featured here, with important applications in environmental science, finance, portfolio analysis

Author: MARCUS BIRKENKRAHE
Created: 2022-08-09 Tue 13:13