

COURSE OVERVIEW

DSC 305 - MACHINE LEARNING - SPRING 2023

Marcus Birkenkrahe

January 10, 2023



What is "machine learning" about?

What do you think "machine learning" is about?

Why is machine learning important?

What do you think - is machine learning important? Why or why not?



Figure 1: xkcd, <https://xkcd.com/1838/>, Machine Learning

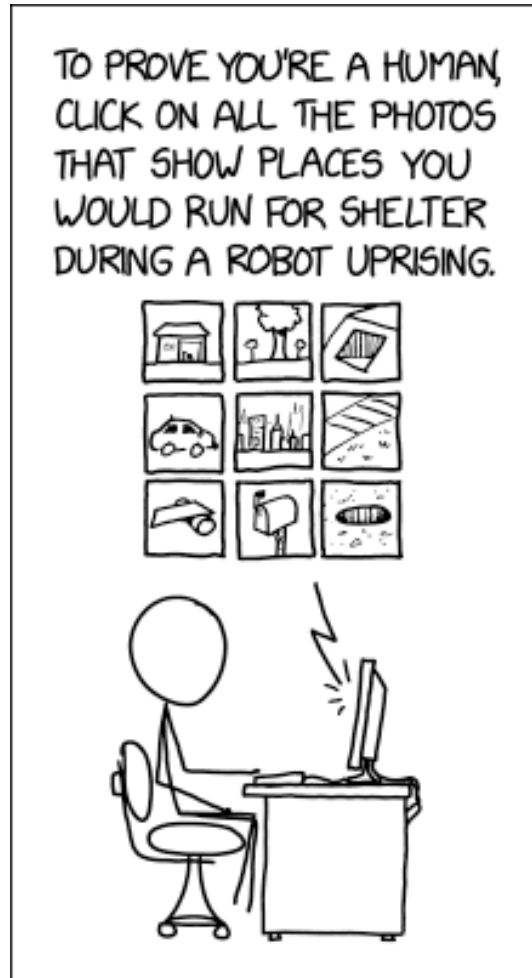


Figure 2: xkcd, <https://xkcd.com/2228/>, Machine Learning Captcha

WEEK	DATE	TOPICS and ASSIGNMENTS
1	Jan 10,12	R Review
2	Jan 17,19	What is Machine Learning?
3	Jan 24,26	Machine Learning Models
4	Jan 31, Feb 2	k-Nearest Neighbors (kNN)
5	Feb 7,9	Naive Bayes
6	Feb 14,16	Logistic Regression
7	Feb 21,23	Classification Trees
8	Mar 2	k-means clustering
9	Mar 7,9	Hierarchical clustering
10	Mar 14,16	Dimensionality reduction
11	Mar 28,30	Cancer data case study
12	Apr 4,6	Artificial Neural Networks
13	Apr 11,13	Modeling with ANNs
14	Apr 18,20	Support Vector Machines
15	Apr 25,27	Performing OCR with SVMs
16	May 2	

Figure 3: Source: syllabus, Canvas (lyon.instructure.com) or GitHub (github.com/birkenkrahe/ml)

What will we do in this course?

- Topics: supervised learning, unsupervised learning, deep learning
- Assignments aligned with Lantz (2019) and some DataCamp lessons
- You will work with R and its ML packages

How will you be evaluated?

REQUIREMENT	UNITS	PPU	TOTAL	% of TOTAL
Final exam	1	100	100	20.
DataCamp home assignments	10	10	100	20.
Class practice	10	10	100	20.
Project sprint reviews	5	20	100	20.
Multiple-choice tests	10	10	100	20.
TOTAL			500	100.

Figure 4: Source: syllabus, Canvas (lyon.instructure.com) or GitHub (github.com/birkenkrahe/ml)

- All course requirements have deadlines
- Late submissions will be penalized (loss of points)
- Final exam will be sourced by term test questions (graded)
- Home assignments: 10 DataCamp lessons (ungraded)
- Class practice: in-class interactive notebooks (ungraded)
- The project will be presented 4 times (sprint reviews)
- Tests: multiple choice questions (graded)

What are "sprint reviews"?

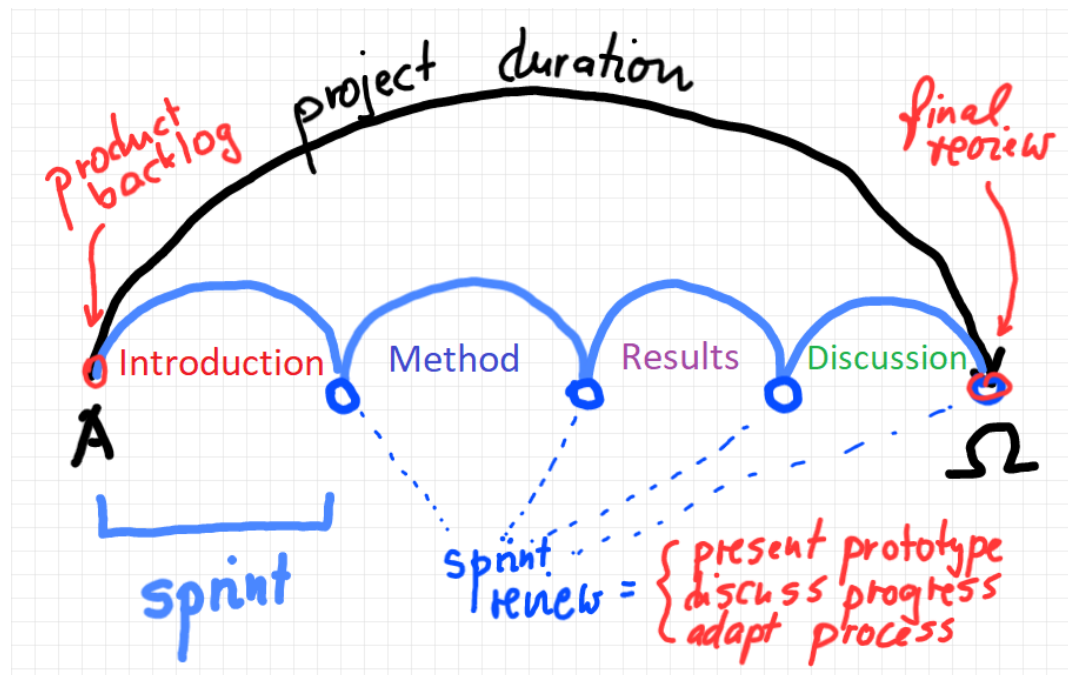


Figure 5: Scrum sprint review and IMRaD publishing framework

- Scrum is an important software engineering technique
- IMRaD is an important framework to publish scientific papers
- MLOps requires improved project management and reading papers

What kind of projects do you want?

- All projects are 3-people group projects - 7 projects in total
- Let me know what kind of project you'd like to work on this term!
- Turn to your neighbor(s) and discuss with them
- Fill in a post-it note and/or vote on existing notes: go to tinyurl.com/2s38bdtk:
 1. I can give you a project topic to work on - IF SO, WHY?

TYPE	PROS	CONS
Independent	More freedom	Time, structure, choice
Chosen	Expectations	Communication, motivation
Presentation	Shorter, visuals	Greater risk, visuals
Essay	Complete, concise	Writing, experience

Figure 6: Type of projects on offer

2. Or you can pick your own project topic - IF SO, WHY?
3. You can do the work and present the results in class - IF SO, WHY?
4. Or you can write an essay instead - IF SO, WHY?

Project examples

Examples for projects chosen by me:

- Work through, check and present a research article given to you
- Create and/or present a case study on one of my topics
- Explain my choice of concept or method with applications

Examples for projects chosen by yourself:

- Use an ML method on a dataset of your choice
- Pick your own research article to study and present
- Pick a concept or method with applications and explain it

Essays: you can use ChatGPT as long as you're open about it (quote it to avoid plagiarism) and can survive a cross-examination on the material.

Concrete examples: Analyse, test and present

- OpenAI Online Hackathons

- Teachable Machine
- Machine Learning for Kids
- ML with Snap!
- Magenta TensorFlow
- OpenAI ChatGPT Playground
- ML with Tensorflow (quickstart)
- DeepBlue defeats Gary Kasparov (Fridman interview)
- AlphaGo wins Go against human (documentary, 2017)

Which tools are you going to use?



Figure 7: Unsplash, workshop

- DataCamp courses (10 weekly home assignments)
- GitHub repository (all course materials except tests)
- GNU Emacs + ESS + R (literate programming environment)
- Canvas (learning management system)

How can you register at DataCamp?







	Understanding Machine Learning What is Machine Learning? Chapter	Team	Active	Jan 19, 13:00 CST			0%
	Understanding Machine Learning Machine Learning Models Chapter	Team	Active	Jan 26, 13:00 CST			0%

Figure 8: DataCamp assignments for January

- You find the invitation link for Spring 23 in Canvas.
- You will automatically be subscribed to the ML team
- If you are in more than one course, I will add you later manually
- These accounts will be valid until July 8, 2023 only

When is the first assignment due?

- The first DataCamp home assignment is due on January 19. For late submissions, you lose 1 point per day (out of 10 possible points)
- The first in-class assignment is due on January 19. For late submissions, you lose 1 point per day (out of 10 possible points)
- We'll write the first weekly multiple-choice test on January 19.



Figure 9: Unsplash, test

What else could you do for a good start?

R proficiency

Complete/review introductory R or statistics courses:

- Introduction to R" in DataCamp (data structures)
- Intermediate R (conditionals, functions, loops, utilities)
- Introduction to statistics
- fasteR by Norman Matloff (GitHub) - fast lane to R
- fastStat by Norman Matloff (GitHub) - fast lane to statistics (**new!**)

(I might give an introduction to R in a biostatistics course.)

Literate Programming

If you do not have any experience with Emacs, work through the **online tutorial** (open it in Emacs with `CTRL + h t`) - ca. 1 hour.

- Learn to open/close the editor

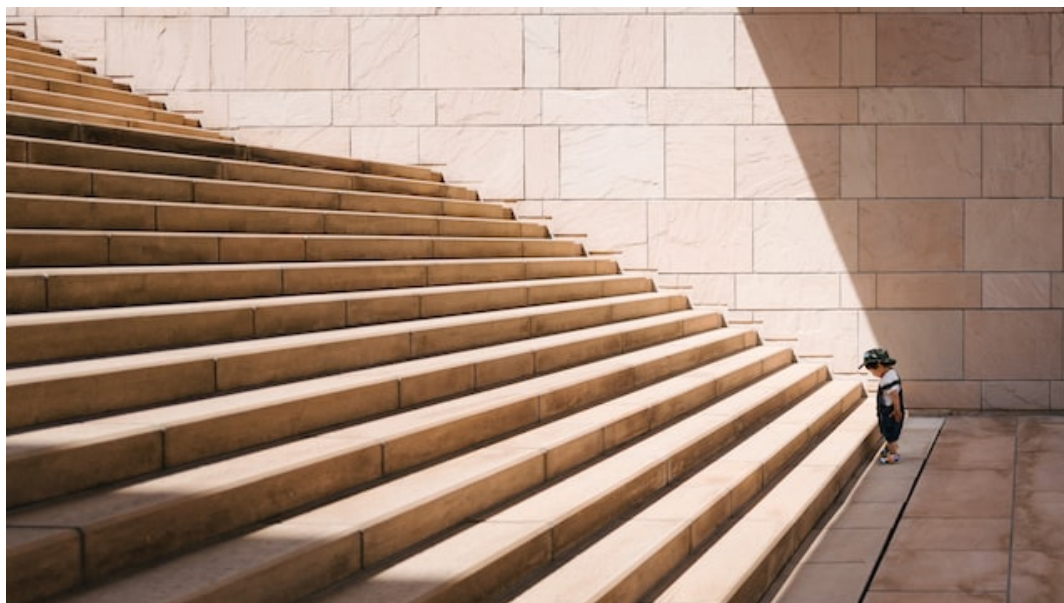


Figure 10: Off to a good start

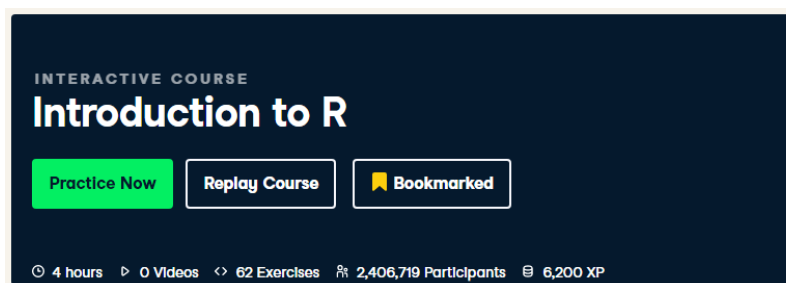
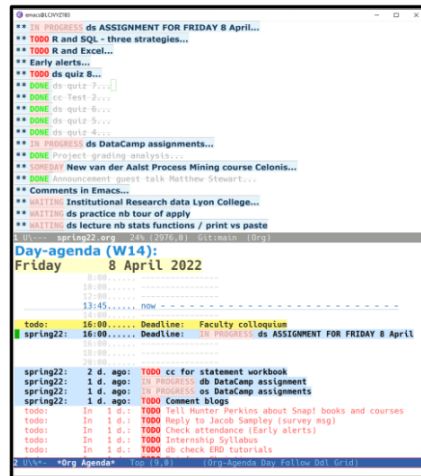


Figure 11: DataCamp course dashboard banner

GNU Emacs



LITERATE PROGRAMMING

- Programmable platform
- Self-documenting
- Fully extensible & transparent
- Text editor
- Keyboard-heavy
- Lisp machine
- Free software
- UNIX / Linux methodology
- Created 1975, launched 1985
- Used by me since 1991
- Hard to learn, easy to use



Figure 12: Literate Programming with GNU Emacs (illustration)

- Learn basic cursor control (moving around)
- Learn basic file management (open/close/find/save files)
- Learn basic windows (buffer) management

Visit me during office hours to get a personal introduction to Emacs.

Course textbook

- Get the 2019 textbook by Lantz, Machine Learning with R (3e) and read the first chapter (it's free even without buying it).
- I'm working through it myself. One assigned DataCamp course is by the same author, who is now senior data scientist at SONY Playstation.
- "Cookbook" machine learning books present only the recipes but offer no serious explanation. Same for many Kaggle projects - code only.

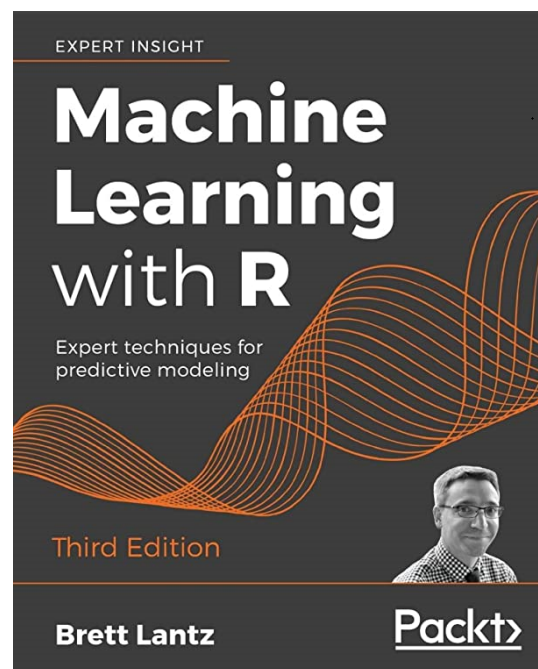


Figure 13: Book cover, ML with R 3rd ed. by Brett Lantz (Packt, 2019)

- I usually work with 3-10 different books but I keep coming back to the best ones that I have really worked through. Another good idea: pick a mathematical text on machine learning

Linux

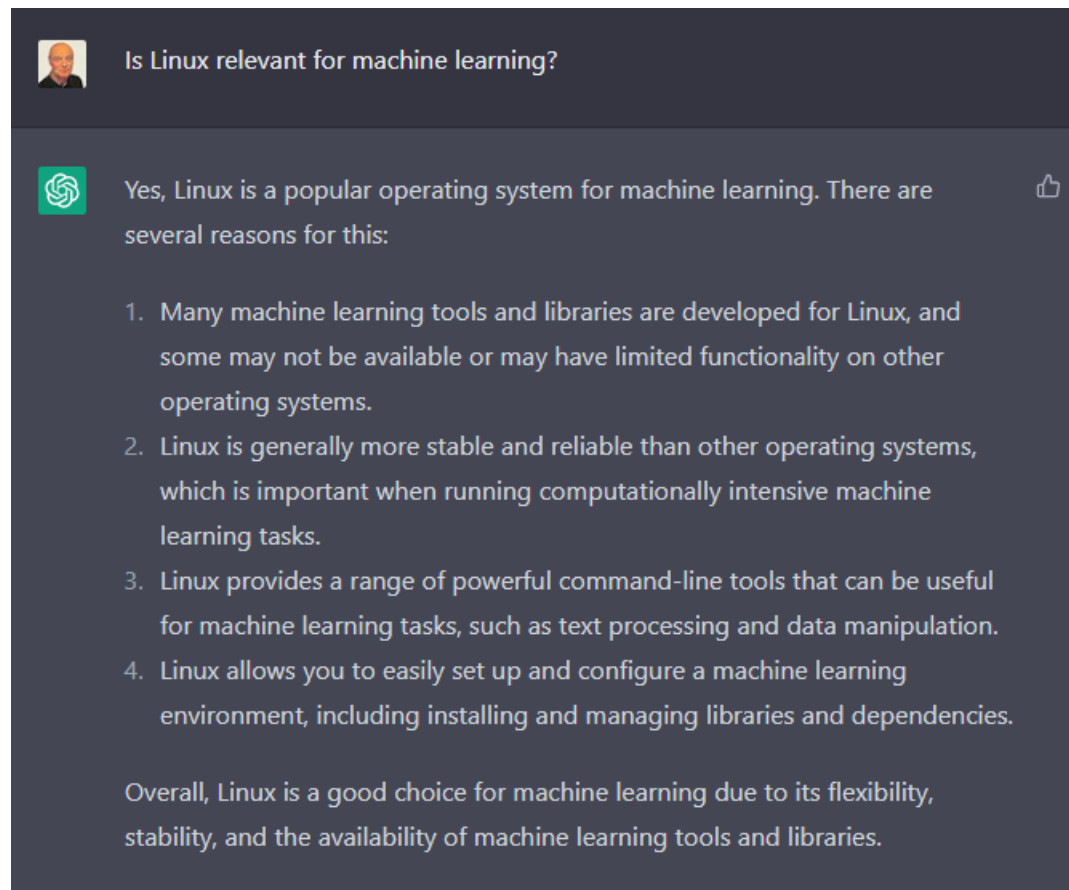


Figure 14: Conversation with ChatGPT by OpenAI

Install WSL (Windows Subsystem for Linux) on your PC, then learn the command line with Shotts' book (5e, 2023).

Caveat: several ML packages did not install under WSL Ubuntu 22 - however, they do install in Google Colaboratory for R (<https://colab.to/r>).

What are you looking forward to?

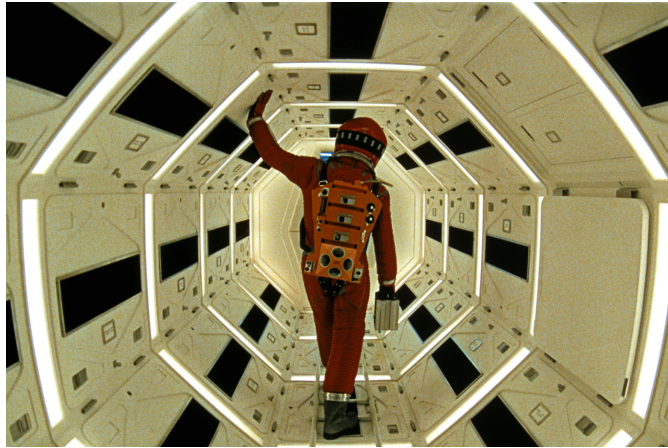


Figure 15: "2001: A Space Odyssey" (Kubrick and Clarke, 1968)

- Reacquainting myself with Neural Nets (1992)
- Starting my own machine learning research project (Medical imaging)
- Training you for opportunities (Stone Ward)

Next

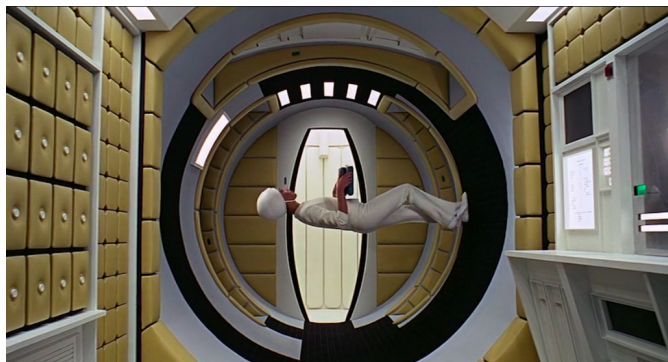


Figure 16: "2001: A Space Odyssey" (Kubrick and Clarke, 1968)



Figure 17: R logo, by the R Project, r-project.org