

COURSE OVERVIEW

DSC 305 - MACHINE LEARNING - SPRING 2023

Marcus Birkenkrahe

January 9, 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

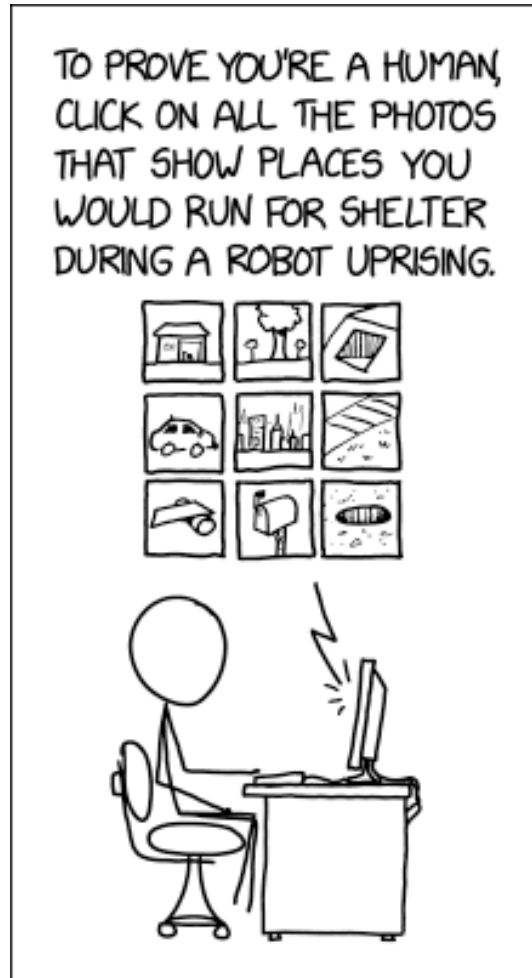


Figure 2: xkcd, <https://xkcd.com/2228/>, Machine Learning Captcha

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 and assignments are largely aligned with a textbook by Lantz (2019) and the DataCamp lessons in the "Machine Learning with R" track.

How will you be evaluated?

REQUIREMENT	UNITS	PPU	TOTAL	% of TOTAL
Final exam	1	100	100	20.
Home assignments	10	10	100	20.
Class assignments	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
- The project will be presented 4 times (sprint reviews)

What are "sprint reviews"?

- Scrum is an important software engineering technique

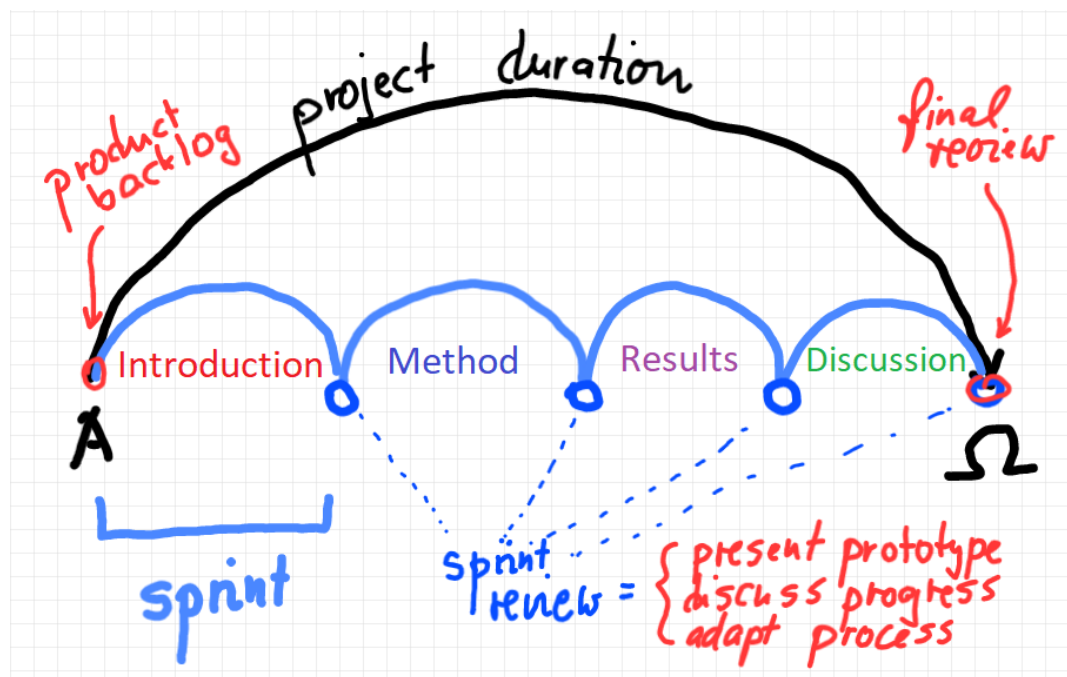


Figure 5: Scrum sprint review and IMRaD publishing framework

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

What kind of projects do you want?

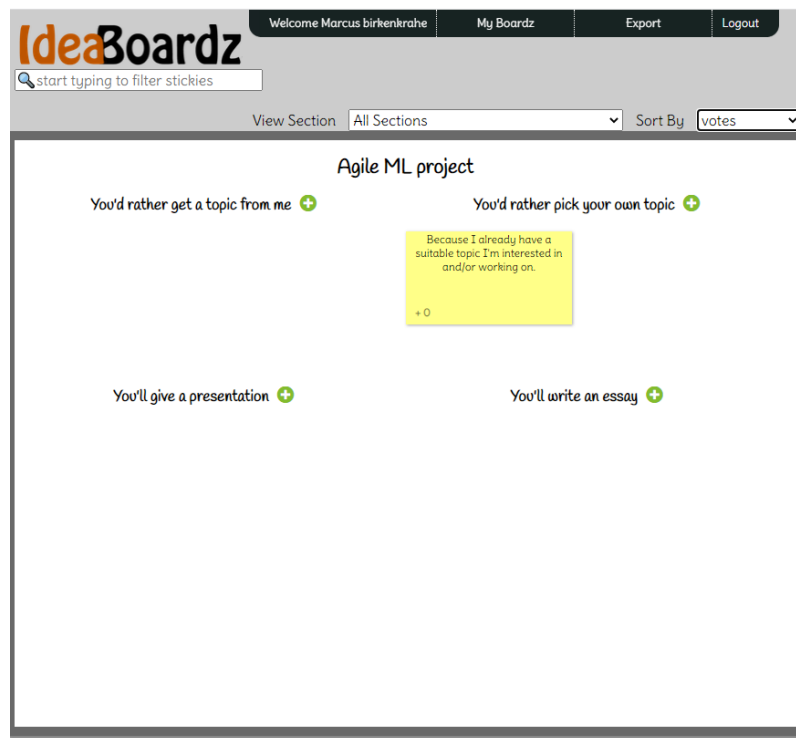


Figure 6: ideaboardz.com Kanban board

- 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?
 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?

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?

- 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







	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

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.

What else could you do for a good start?

1. 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
2. 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 9: Unsplash, test

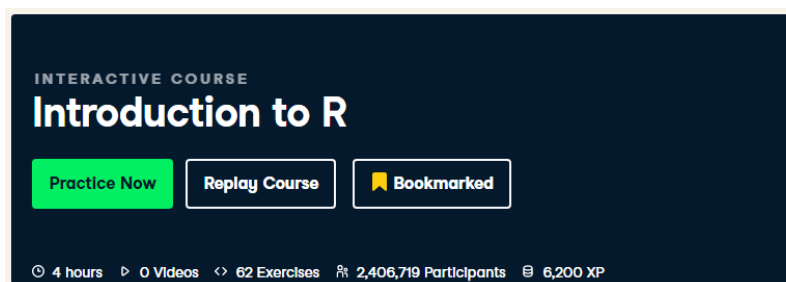


Figure 10: DataCamp course dashboard banner

- Learn basic cursor control (moving around)
 - Learn basic file management (open/close/find/save files)
 - Learn basic windows (buffer) management
3. Get the 2019 textbook by Lantz, Machine Learning with R (3e) and read the first chapter (it's free even without buying it).
 - There are many other books (most of them not all that good)
 - Stay away from "cookbooks" (ML with Keras or TensorFlow)

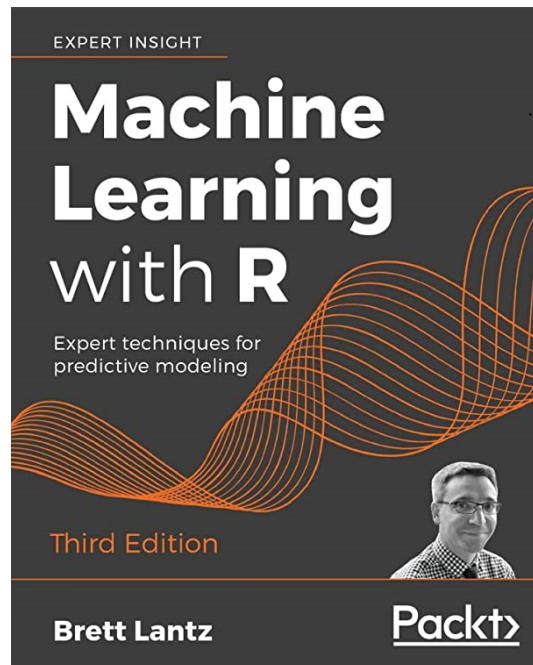


Figure 11: Book cover, ML with R 3rd ed. by Brett Lantz (Packt, 2019)

4. Install WSL (Windows Subsystem for Linux) on your PC, then learn the command line with Shotts' book (5e, 2023). ChatGPT: *Is Linux relevant for machine learning?*

What are you looking forward to?

Next

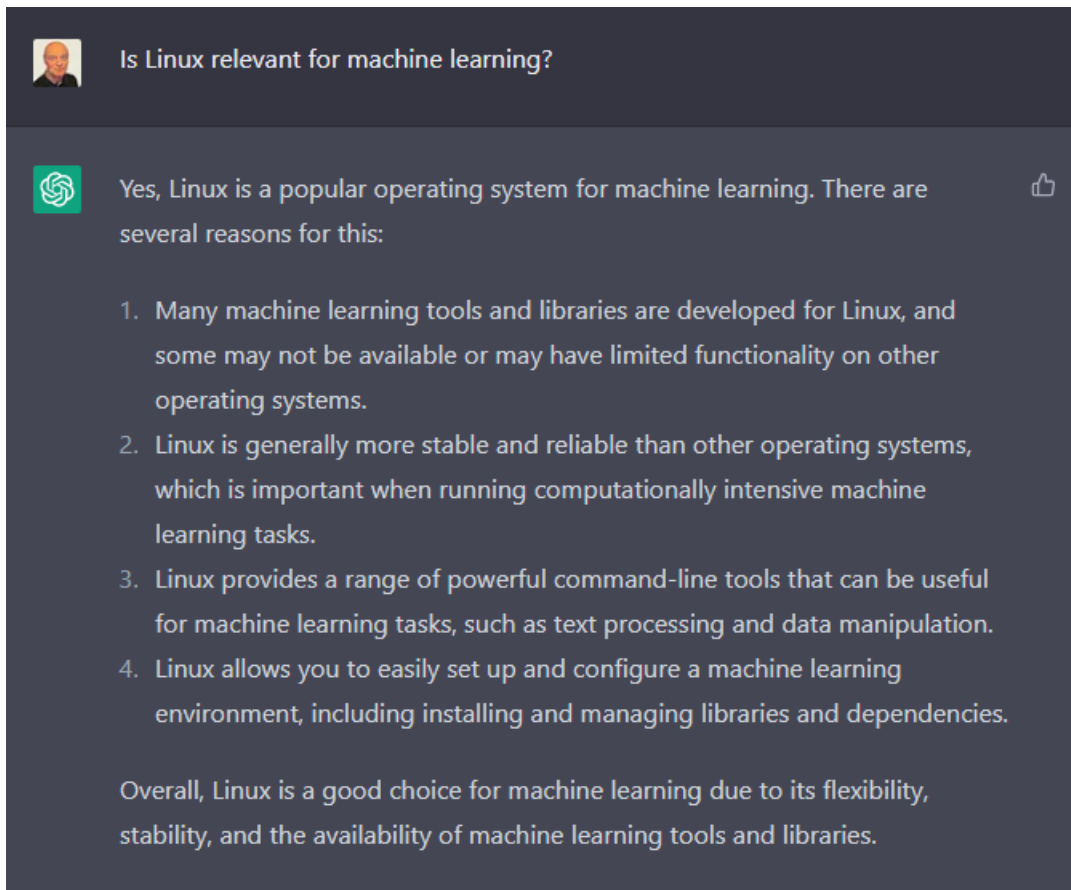


Figure 12: Conversation with ChatGPT by OpenAI

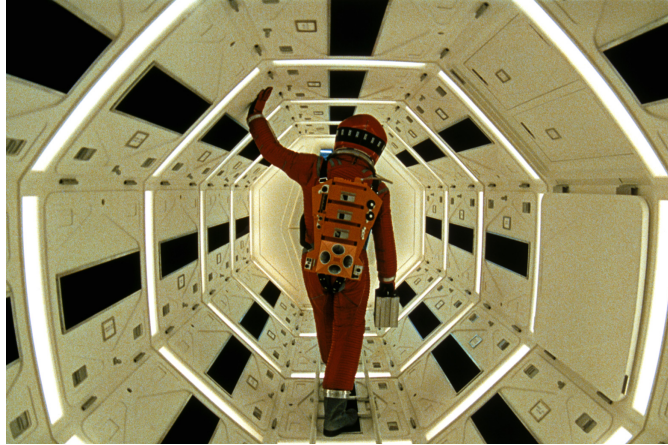


Figure 13: "2001: A Space Odyssey" (Kubrick and Clarke, 1968)

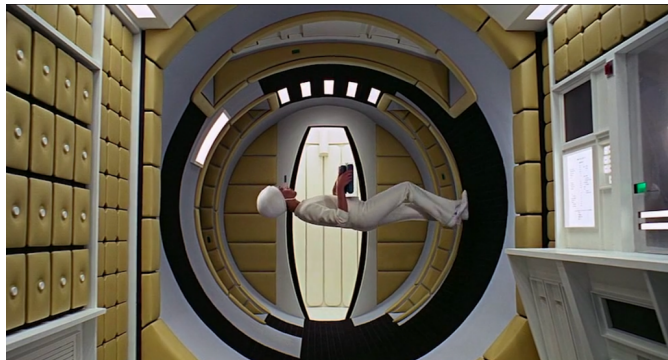


Figure 14: "2001: A Space Odyssey" (Kubrick and Clarke, 1968)



Figure 15: R logo, by the R Project, r-project.org