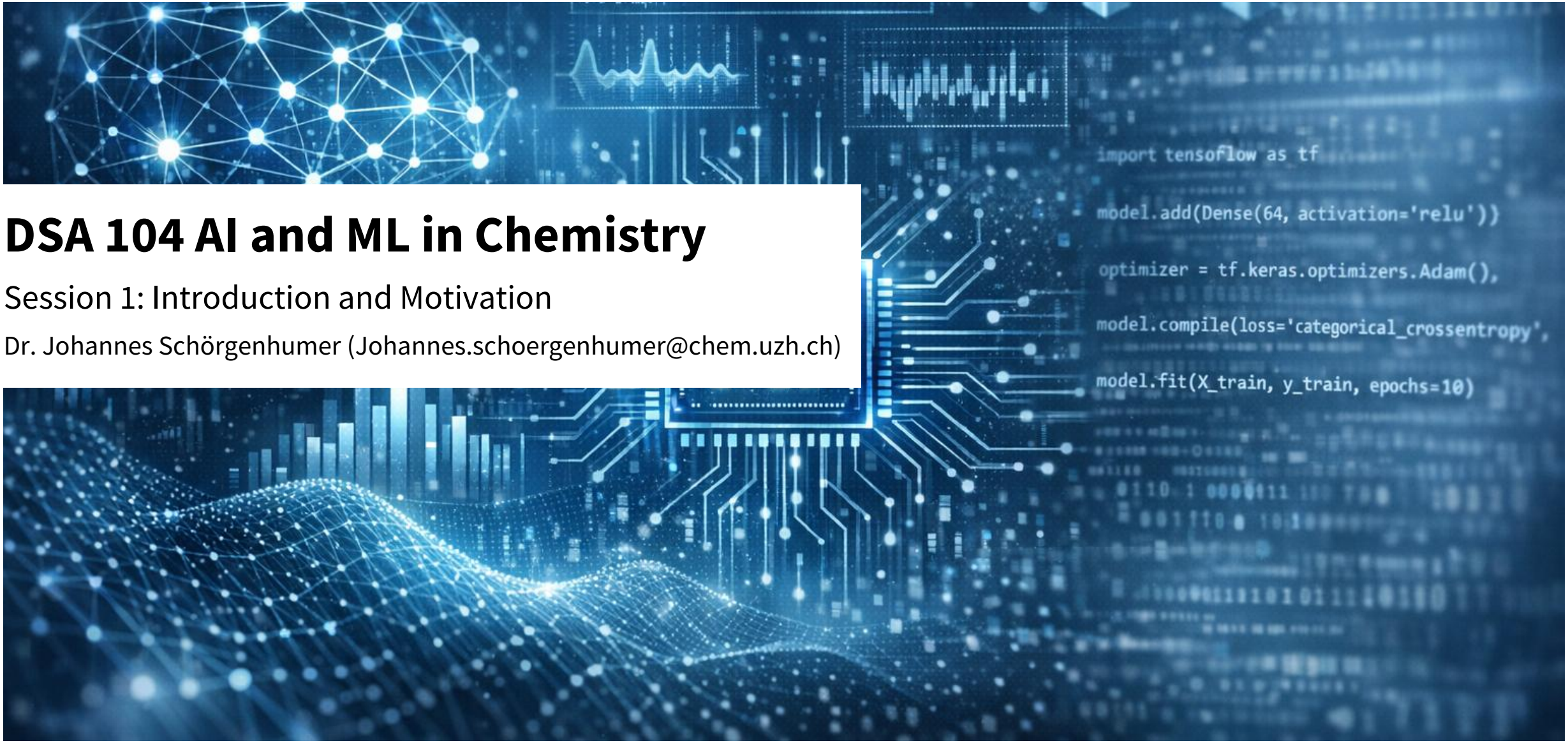
Universität Zürich[UZH]

Department of Chemistry



DSA 104 AI and ML in Chemistry

Session 1: Introduction and Motivation

Dr. Johannes Schörgenhumer (Johannes.schoergenhumer@chem.uzh.ch)

Who are we?

Dr. Jasmin Hafner

Bioinformatics Specialist
Postdoc Environmental Chemistry

Department of Chemistry and EAWAG

Dr. Johannes Schörgenhumer

Head of UZH High Throughput Experimentation Lab
Organizer of DSA Minor

Department of Chemistry

Who are you?

Your background?

Why are you taking the course?

What do you expect?

About this course

This course is …

… a very **basic introduction** to ML and AI in the LifeSciences.

… meant to equip you with basic tools in the **field to overcome the initial barrier**.

… aiming to **create some awareness** of benefits, possibilities, limitations, pitfalls and risks

This course is not…

… providing an **in-depth account** for the topic.

… aiming to make AI model architects out of you.

Connects to **DSA102** (lab automation), **DSA103** (data science), **DSA105** (applying contents)

For a more comprehensive insight (electives!), consider other modules, e.g.:

— 06SM521-533 Advanced Machine Learning

— 03SM22MI0027 Deep Learning

Course organisation

2 Session each week (integrated exercises, weekly assignments):

— Mo, 8:00 – 9:45
We, 13:00 – 14:45

— Lectures are streamed and recorded (MS Teams join code on OLAT)

— **Personal attendance encouraged** (exercises, discussions in small groups, …**Goodies :D**)

Special dates:

— Symposium: May 27 – further details and agenda to be announced

— Exam: We, June 3

— Mock Exam: Mo, May 18

Note: Schedule might be adapted slightly!

Session	Date	Weekday	Topic	Assignments	Lecturer
1	16.02.	Mo	Introduction & Motivation		Johannes
2	18.02.	We	Recap - Tabular Data		Johannes
3	23.02.	Mo	Definitions, Supervised ML		Johannes
4	25.02.	We	More ML models	Assignment 1	Johannes
5	02.03.	Mo	ML Metrics and Evaluation		Johannes
6	04.03.	We	Unsupervised ML	Assignment 2	Johannes
7	09.03.	Mo	Data types - Feature Selection		Johannes
8	11.03.	We	Q&A, Recap	Assignment 3	Johannes
9	16.03.	Mo	Regression models		Johannes
10	18.03.	We	Deep Learning	Assignment 4	Johannes
11	23.03.	Mo	Neural networks		Johannes
12	25.03.	We	Q&A, Recap, **DSA 105 organisation**	Assignment 5	Johannes
13	30.03.	Mo	Neural networks		Johannes
14	01.04.	We	GNN	Assignment 6	Johannes
15	13.04.	Mo	Transfer learning & foundation models		Johannes
16	15.04.	We	Q&A, Recap	Assignment 7	Johannes
17	20.04.	Mo	Generative models		Johannes
18	22.04.	We	AI and Automation	Assignment 8	Johannes
19	27.04.	Mo	Applications: Case studies		Johannes
20	29.04.	We	Applicability domain	Assignment 9	Jasmin
21	04.05.	Mo	Q&A, Recap		Johannes
22	06.05.	We	Uncertainty modelling	Assignment 10	Jasmin
23	11.05.	Mo	Modern Applications in LifeSciences		Johannes
24	13.05.	We	Ethics		Both
25	18.05.	Mo	**Mock Exam**		Johannes
26	20.05.	We	Q&A: Solutions Mock Exam; Feedback		Both
27	27.05.	We	**Symposium**		Both
28	03.06.	We	**Exam**		Both

Course organisation: Grading

No mandatory attendance for grades

Main grade determined by **Exam**:

— graded in half steps

— On paper only, no PC tasks

Assignments for bonus (+ up to half a grade):

— 10 assignments in total, mainly handed out on Wednesday (see schedule)

— For each complete assignment 0.05 bonus grade improvement

— Complete: All questions answered, all tasks completed

— Hand in by mail to johannes.schoergenhumer@chem.uzh.ch

— No individual feedback will be given, just a confirmation if complete or not

— Solutions will be discussed in the Wednesday session afterwards (i.e. 1 week to work)

Course organisation: “Disclaimer”

Course offered for the first time, hence:

- We are trying a few new tools and formats (e.g. stream via Teams)
- Schedule may be adjusted over the course of the semester
- All parts of the course may need to be adapted / improved for the next iteration:
  - Content (focus, balance life sciences, etc.)
  - Structure (content distribution, Q&A sessions etc.)
  - Format (exercises, live discussions, etc.)
  - Connectivity to other courses (DSA105!)
  - …

Your feedback is needed and welcome at any time!!!

Tools and materials

Prerequisites:

— Python skills (pandas, rdkit, seaborn, …)

— Basics in data science (and some statistics)

— Use of a suitable environment: IDE, Git, Jupyter

Platforms:

— **DSA 104 Repo: https://github.com/schoergj/DSA104/**

— OLAT (Course communication, lecture slides for reference)

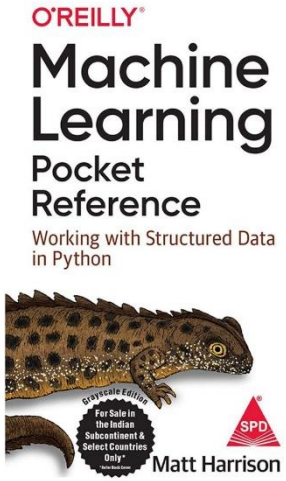
— Streams and Recordings on Teams

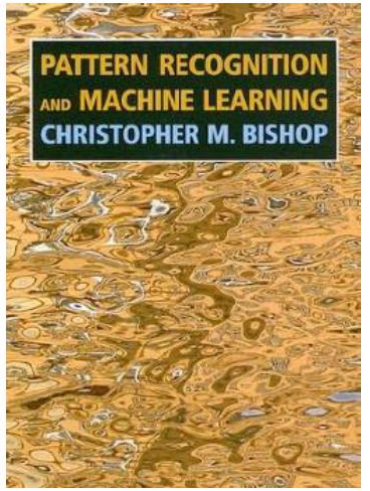
Materials:

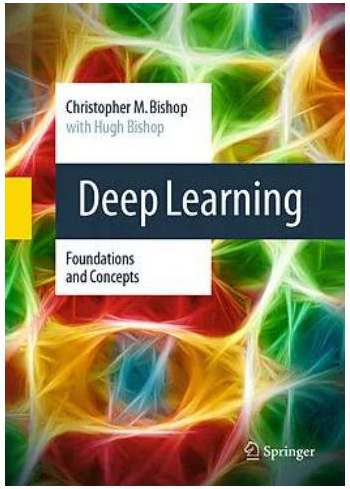
— Books (some recommendations)

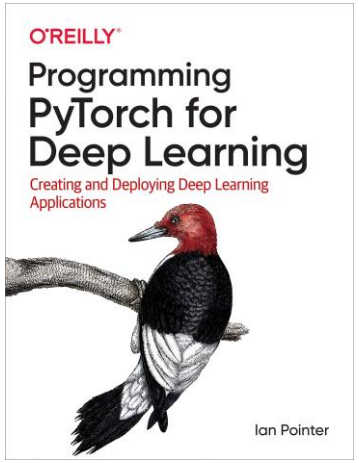
— Good resources available online

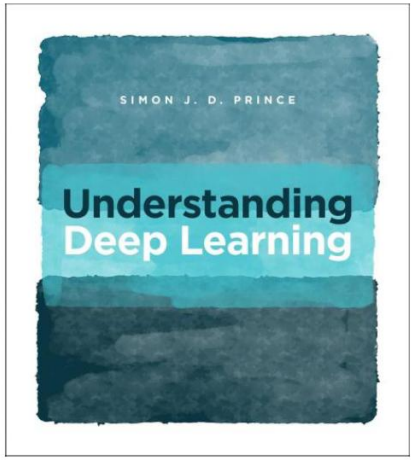
— Domain specific literature











https://www.geeksforgeeks.org/machine-learning/machine-learning/
https://developers.google.com/machine-learning/crash-course
https://www.w3schools.com/ai/

https://schwallergroup.github.io/ai4chem_course/

Why? A personal anecdote from a couple of years ago

2020(?): My (completely ignorant) brainstorming for grant proposal: ML model to predict reaction outcome based on reaction data and some preset rules

Me, PhD with highest honours and honorary ring of the Republic of Austria, apparently completely clueless

"How difficult is that? Or how long would that take to build and train some ML model and all?"

"Depends. A couple of minutes maybe?"

(…under certain presets and conditions. Then explains to me some basics in ML.)

Dr. A. Schörgenhumer

Bears the same honorary ring, but actually knows about stuff

(Computer Scientist, AI Coach, Senior Software Developer at *Transformas Consulting Solutions GmbH*)

Lessons learned:

… novel tool with **perceived high entry barrier**

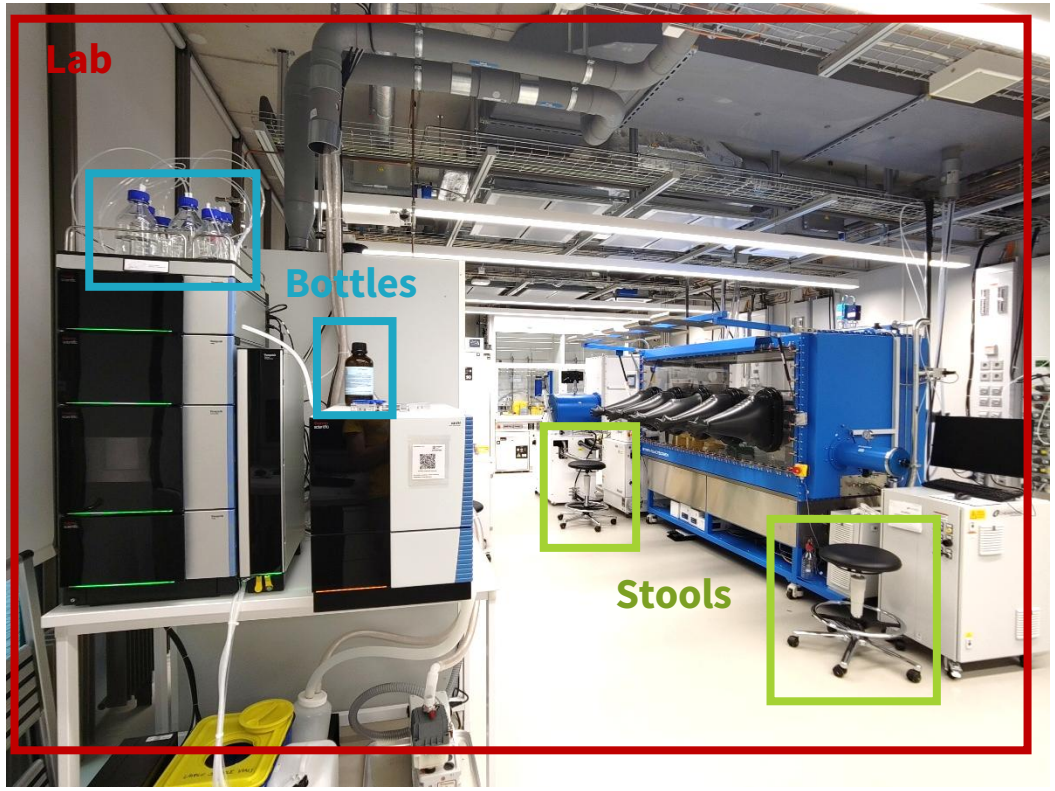
… but not true, as with other modern tools (e.g. lab automation – DSA 102, Prototyping and hardware programming – CHE 725, …)

Motivation – the human problem with data

We are good at…

…Detecting and distinguishing familiar classes like “human”, “laboratory”, “bottle”, “water”

…Discriminating solvents, bacteria strands, reaction types, *…if trained!*



PBS: Phosphate solution, odourless, colourless, harmless

Ethanol: organic solvent, structure & properties, certain scent, flammable, potentially intoxicating

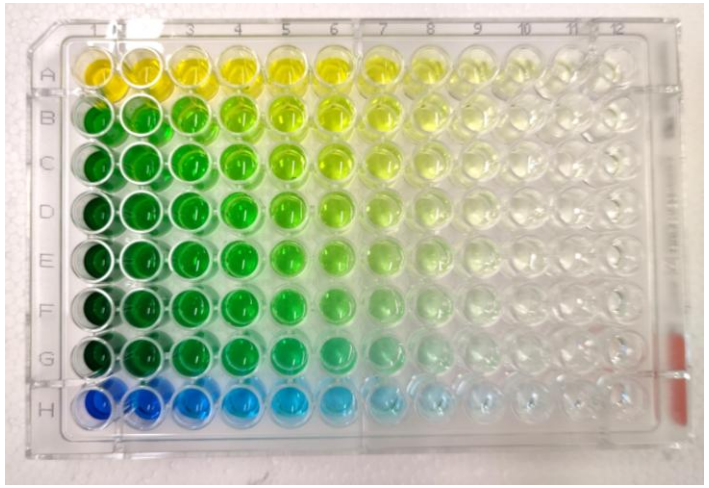
Motivation – the human problem with data

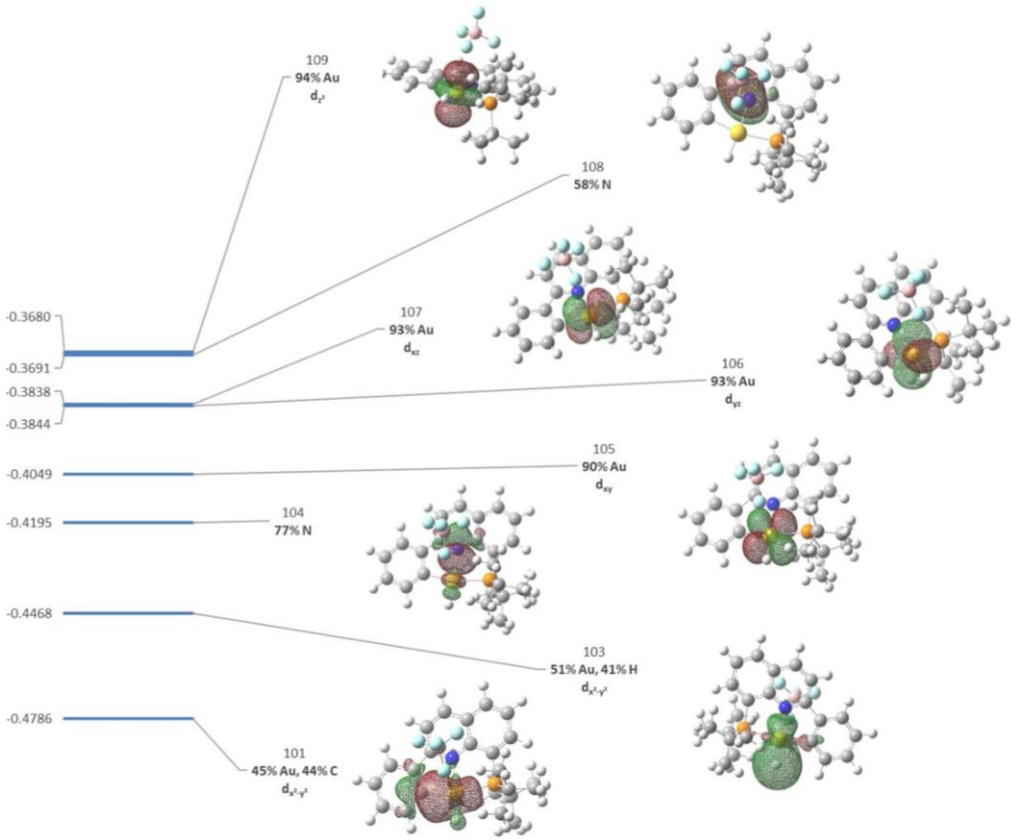
We are not good at…

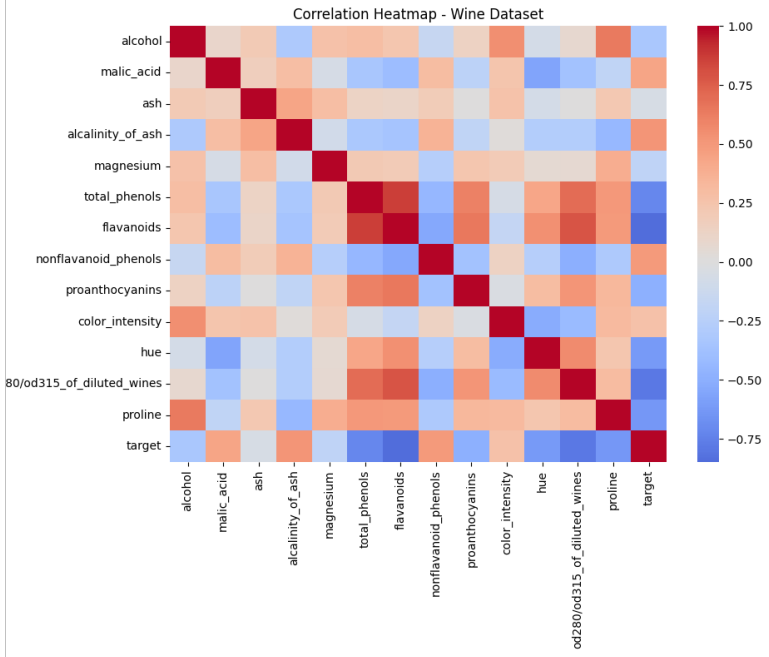
…Measuring continuous parameters: concentration, turbidity, cell viability, turnover, polarization, …

…Working with unfamiliar categories: quantum yield, circular dichroism, long-range electronic interactions, stereogenic surfaces …

…Recognizing complex patterns in multivariate data: Dependence of a fragrance on molecular properties







1) Find out more in DSA102
2) *JACS Au* **2025**, *5*, 1439–1447
3) More in DSA103

Motivation: Automated data analysis - the solution?

Input: observations, meta-data (ground truth, acquisition time, location, etc.)
Output: prediction of target variable (label, maps, extrapolation, etc.)

Pros:

— Fast, scales to massive amounts of data (only limited by available computational budget)

— Implements validated, proven theory

— Discovers patterns that would remain hidden to humans

— Result is objective and (mostly) repeatable

Cons:

— “absurd” errors that humans would never do

— Usually uses only partial available expertise (i.e., not everything can be encoded in software)

— *Much expertise needed for success in both,* ***machine learning and the problem****!*

…not a universal solution!!

Why this course?

Data driven algorithms ubiquitous!

All fields of applications massively impacted:

— Information technology

— Manufacturing and supply chains

— Medicine and healthcare

— Education

— Finance and trading

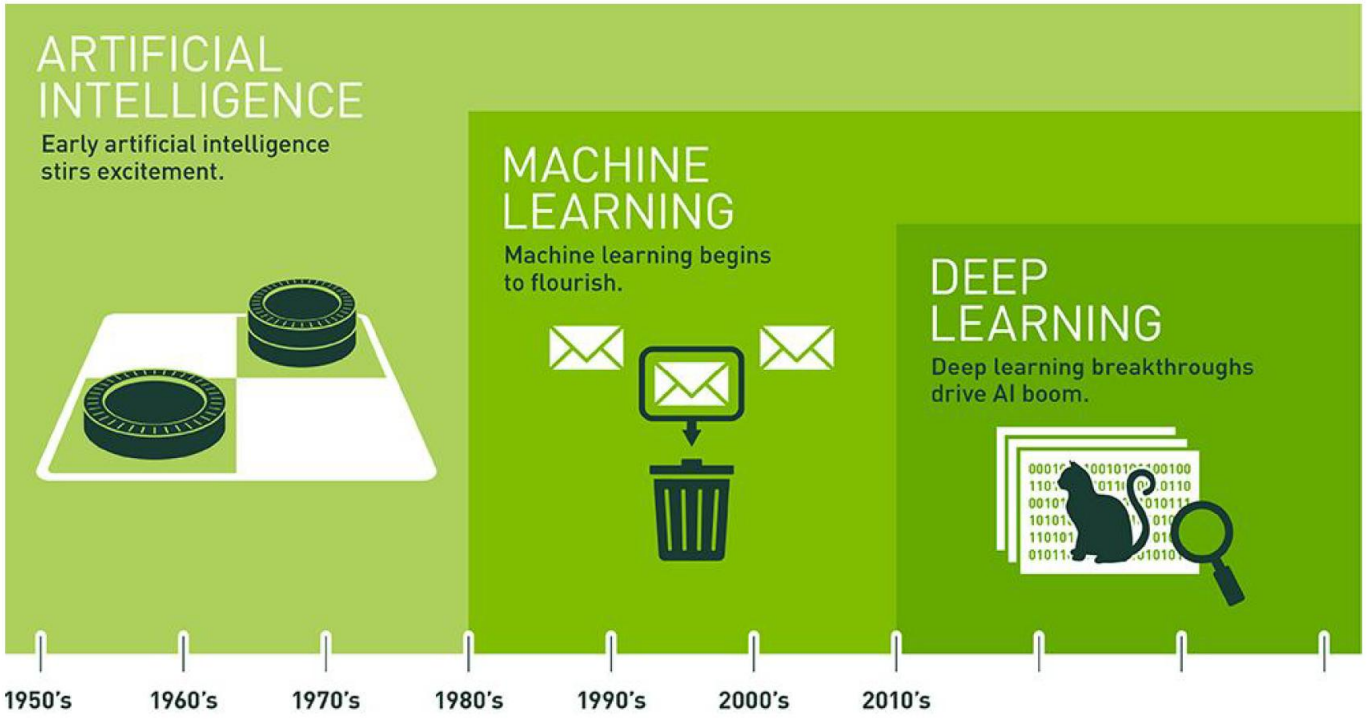
— News and publishing

— Advertisement

— Transportation

— ...

— Science



Adapted from A. Schörgenhumer, *Hands-on AI I*, Lecture materials, **2023**, JKU Linz.

Discussion

Where have you already encountered / or worked with ML or AI?

How and what for have you used generative AI?

Discuss in small groups and share then your results!

ChatGPT - generated image: AI communicating with humans, discussing about life sciences

Promises and applications in the Life Sciences

… cut development time for drugs

… facilitate complicated and time-intensive processes

… "brute force" old problems

… provide new research avenues

… avoid material waste

… deplace human expertise

Welcome to OpentronsAI

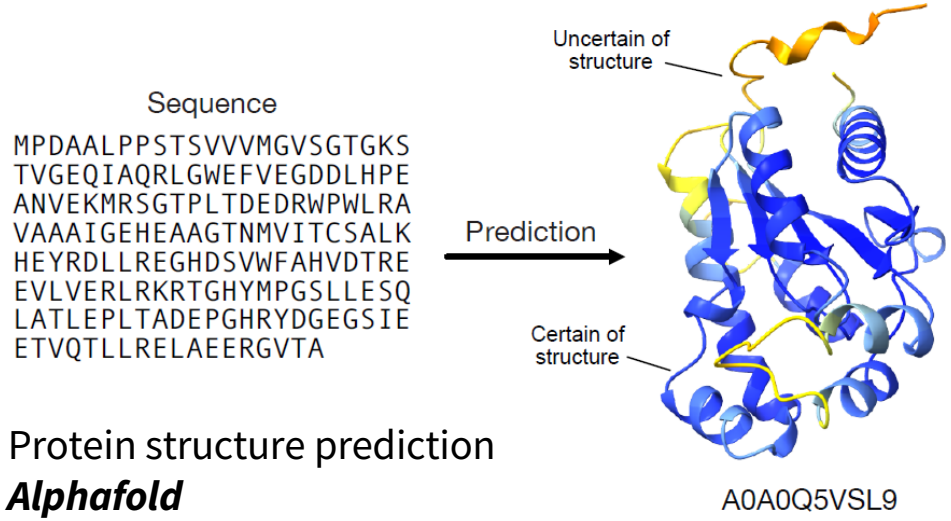
Get started creating and optimizing protocols for your Opentrons robot.

Update an existing protocol
Upload your existing protocol and explain what you'd like to change
Update a protocol ▸

Help with a new protocol
Go through our wizard to create a new protocol from scratch
Create a new protocol ▸

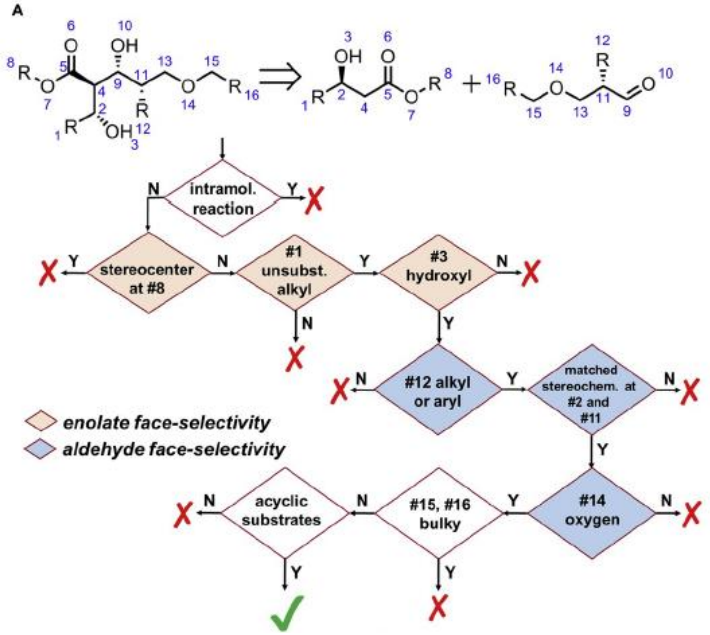
Go to chat
Head directly to OpentronsAI chat to ask a question or paste an existing prompt
Chat now ▸

NLM based protocol designer for Opentrons robots:
OpentronsAI *https://ai.opentrons.com/*



Protein structure prediction
Alphafold

Nature **2021**, *596*, 583-589



Rule-based & DL Retrosynthesis: ***Synthia***

Chem **2018**, *4*, 522-532

The holy grail? No, but…

…definitely here to stay!

— Large investments by both industry and academia

— Thrilling research groups (Switzerland), e.g.:
  — Philippe Schwaller (EPFL)
  — Kjell Jorner (ETHZ)

https://schwallergroup.github.io/

https://dcl.ethz.ch/

— New infrastructures - automated labs, e.g.
  — SWISSCAT+ at EPFL and ETHZ
  — UZH HTEL

— Finally, also some education: Many “Digital Chemistry” courses and some dedicated programmes; unique hands-on DSA Minor

ML, AI, the modern world … all about data

Data is this century's oil!

Data = Foundation of any ML / AI workflow, and thus: **1st step**!

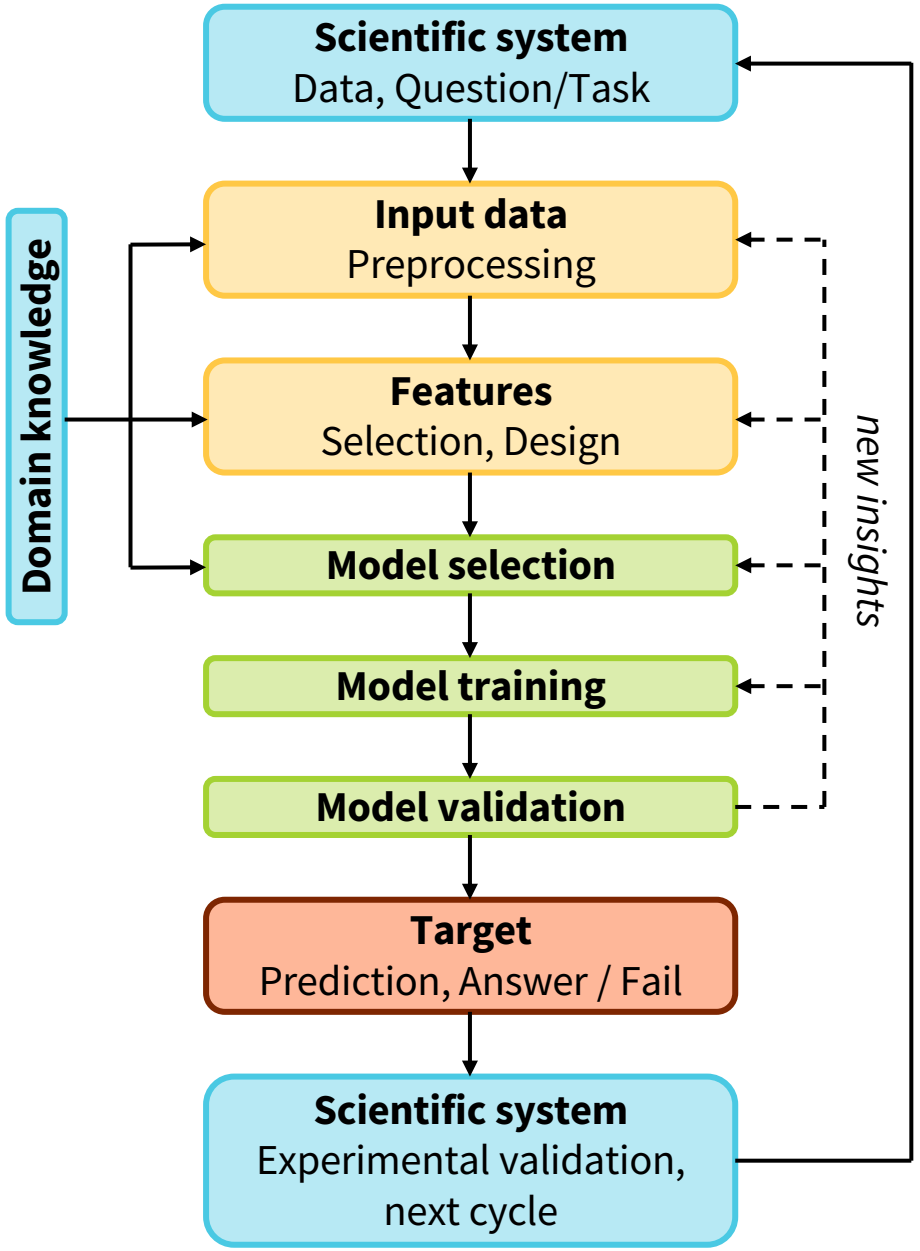
Heavy lifting in the workflow mainly associated with the data:

1. **Convert science problem to machine-readable form (Input data)**
2. **Preprocessing and Feature selection**
3. Define Target
4. Select model, train, validate

Requirements:

— Set up the working environment for this course (run a small first ML example)

— How to handle Data using Python (DSA 103) – small recap next time!



Exercise: Laying the foundation and how to work with the repo

1) Tasks for coffee break (If you haven't done so already anyway):

- Get on Git
- Install IDE (e.g. VS Code (VSC) or Pycharm)

2) Quick walkthrough: how to use the repo and the environment

- Fork and clone the DSA 104 repo: **https://github.com/schoergj/DSA104/**
- Create a virtual environment and sync dependencies

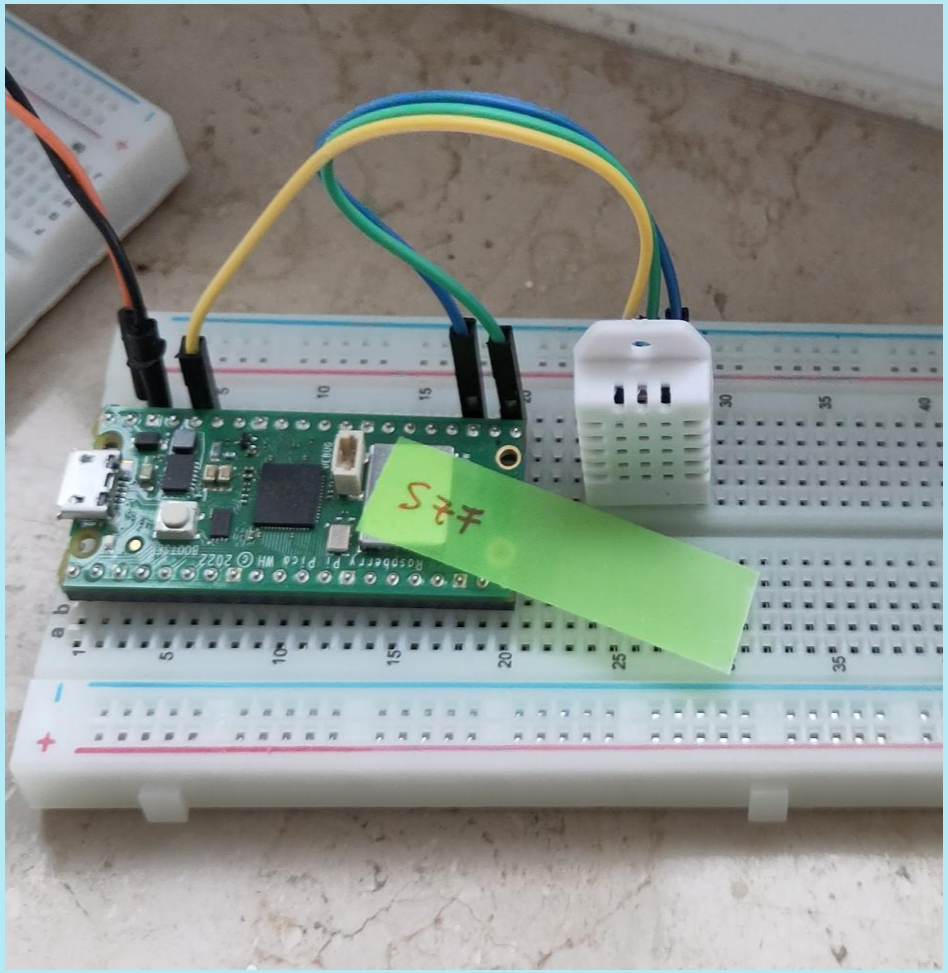
Exercise: Predicting sensor data

Scenario: Heating in a room in our apartment not sufficient, high temperature gradient towards outer wall and floor.

Several sensors (DHT22) have been placed around the apartment and on the balcony to record temperature and relative humidity. On a Raspberry Pi Pico, the data is timestamped and written into a .csv file.

Tasks:

1) Functionality check: Run the provided notebook “dht_prediction”
2) Predict missing temperature and humidity values by regression models
3) Find out which model works best for the prediction and play around with some parameters



Learn how you can do this (and more fun stuff):
CHE 725 Hacking for Chemists (2 ECTS elective, HS)