PROGRAM ROADMAP

Iskriyana Vasileva

AGENDA

- Intro Round
- The DSR Roadmap
- Get You Ready for Data Science
- Set Up a Simple Data Science Project
- Some Meetup groups (in Berlin)

INTRO ROUND

INTRO

- About me:
  - A data enthusiast - 11 years in the field
  - Did DSR to transition from BI Analyst to Data Scientist
  - https://www.linkedin.com/in/iskriyanavasileva/

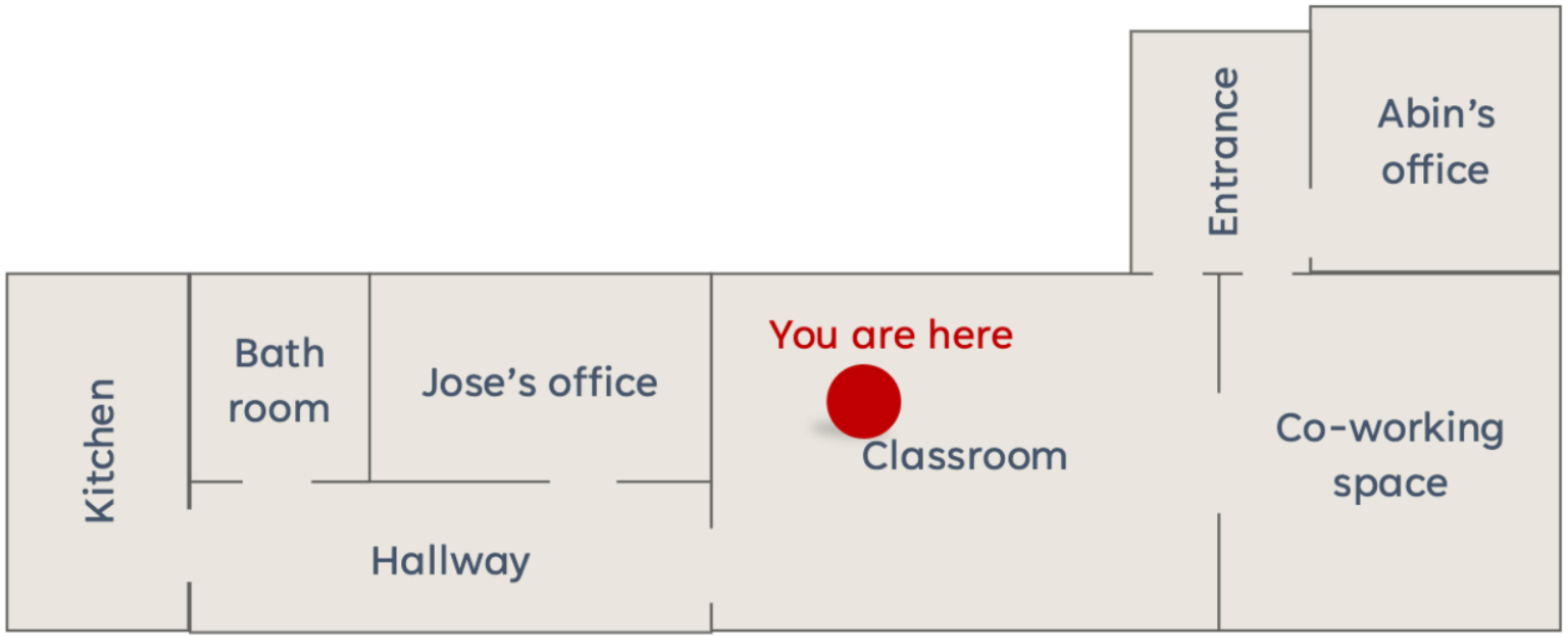
- About this class:
  - You can ask questions at any time
  - If you'd like me to change something during the lecture – let me know
  - If you think of something after the lecture – write to me on LinkedIn or Slack

INTRO

- What about you?
  - Your name (feel free to remind me about it ;))
  - Your background
  - What brought you here?
  - What fascinates you about Data Science?

THE DSR ROADMAP

THE RETREAT


Entrance
Abin's office
Kitchen
Bath room
Jose's office
You are here
Classroom
Co-working space
Hallway

THE TEACHERS

INDUSTRY PROFESSIONALS

- Practical, up-to-date knowledge
- Less theory, more practice

SOME ARE DSR GRADUATES

- They know what it's like
- Ask them about their projects

THEY ARE PEOPLE TOO

- Working on weekends
- Teaching while having a full-time jobs

ADMIN

Tools Access and Communication

- → test with 👍 in Slack #general

Class Guidelines

- The office is usually open from 8:30 AM to 6:30 PM; lectures are between 9:30 AM and 5:30 PM.
- Please, be punctual! Notify the teacher and peers on Slack if you'll be late.
- Teachers will aim to share materials in advance when possible.

Course Materials

- Don't feel overwhelmed by the volume of materials; use them as additional resources post-bootcamp.
- The flow of the class is dynamic and depends on the entire group's interaction.

ADMIN

Utilize free days for

- Project preparation & 1:1 sessions with Jose
- Reflecting on material covered
- Preparing for upcoming classes
- Relaxing and personal time

Feedback

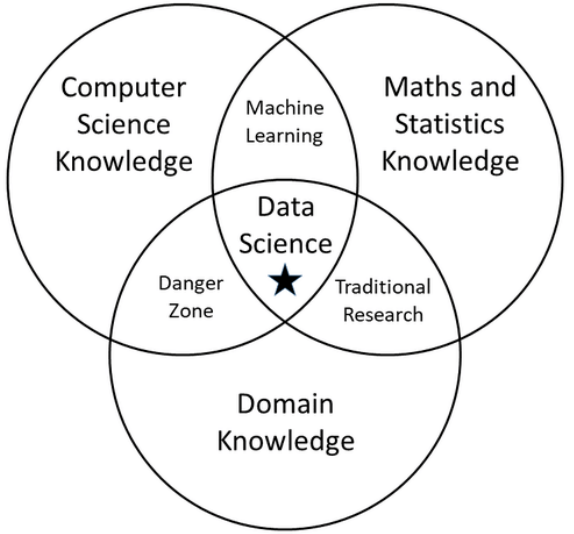
- Proactively communicate to the teacher, Arun, Abin, or Jose any feedback you have
- You can address concerns about the class directly during sessions – the teachers and staff will try to accommodate your needs
- Your early feedback allows for timely enhancements to the program – for e.g. QA sessions, if you are interested / need help on certain topics

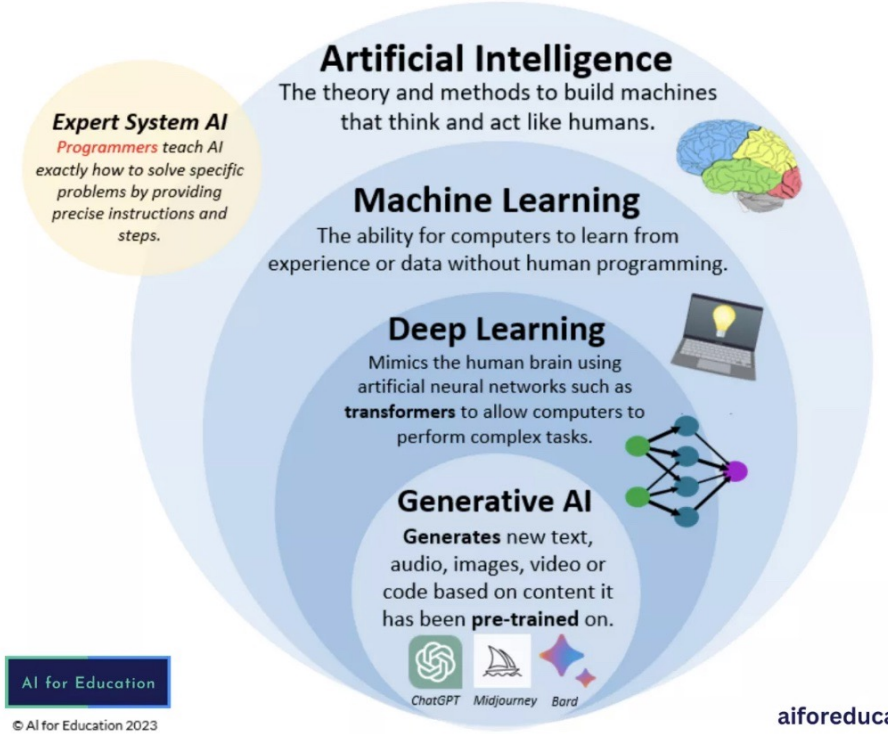
Attendance

- Strive to attend in person
- In case of illness or childcare needs, remote participation is an option
- Inform Abin or Arun as early as possible to facilitate remote access

THE DSR ROADMAP

- Theoretical & Technical Fundamentals
- Machine Learning Fundamentals
- Mini Competition
- Deep Learning - Non-generative & GenAI
- Practical Data Science
- Soft Skills
- The Final Project





DSR ROADMAP: THEORETICAL & TECHNICAL FUNDAMENTALS

Development Tools and Environments

- Git & Bash
- Docker & Databases

Programming and Data Analysis

- Python, NumPy & Pandas
- SQL

Theoretical Foundations

- Probability & Statistics

Data Presentation

- Visualisation

DSR ROADMAP:
THEORETICAL & TECHNICAL FUNDAMENTALS

Purpose

Git & Bash - command-line proficiency and version control for efficient software development

Docker & Databases - deploy containerized applications and manage databases for scalable and reliable software solutions

Python, NumPy & Pandas – essential for proficient data handling & advanced data science applications

SQL - leverage structured querying for data retrieval

Probability & Statistics – understanding the role of probability in statistical reasoning & using statistics for in-depth data analysis

Visualization - create compelling visual narratives from data to inform and persuade effectively

Tips - do not underestimate them!

Even if you know, use these lecture to refresh your skills & strengthen them.

Prepare questions you encountered during your preparation for the bootcamp but could not find the answer to them. It can be useful for everyone including your teachers.

Useful literature

Data Wrangling with Python (Jacqueline Kazil, Katharine Jarmul)

Python for Data Analysis (Wes Mckinney)

GIT & BASH

Purpose

Reproducibility and organization of code

Bash

- **What:** Bash is a shell for accessing and controlling various tools and services. Bash allows you to navigate your file system, install and manage packages, run scripts, and interact with other tools like Anaconda, Python, and Git
- **Why:** It provides fast access to important tools for Data Science such as Anaconda and Python
- **Focus:** Basic Bash command line usage and file system navigation.
- **Activities:** Hands-on exercises including file operations, text manipulation, and how the shell interacts with Python scripts.

Git

- **What:** Git is a version control tool
- **Why:** It allows for easy collaboration with other developers
- **Focus:** Introduction to Git fundamentals and version control concepts.
- **Activities:** Theory mixed with practical exercises on basic Git commands and collaborative features.

Note: Adjustments to content pacing based on class response to ensure comprehension.

DOCKER & DATABASES

Purpose

Get to know docker as tool enabling consistent and scalable environments for data science projects, simplifying collaboration and deployment.

Docker

- Learn to navigate and utilize Docker for deploying containerized applications.
- Work with existing Docker images to understand container functionality.
- Build custom Docker images, integrating data science tools and environments.

Database Concepts

- File Formats
- NoSQL Databases

PYTHON, NUMPY & PANDAS

Purpose

- Essential for efficiently managing, analysing, and manipulating data
- The foundational elements necessary for advanced data science applications

Python

- Data types: Numbers, Strings, Lists, Dictionaries, Tuples.
- Control with if/for/while, iterations, functions, and lambdas.
- File operations and data type nuances.

Numpy

- Array essentials: creation, attributes, and operations.
- Key functions: reshaping, Ufuncs (np.add), broadcasting, advanced indexing.

Pandas in Brief

- Data structures: Series and DataFrames
- Data handling: indexing, missing data, Ufuncs
- Data manipulation: combining, grouping, pivoting

SQL

Purpose

Manage relational databases effectively

- **Syntax Mastery:** Practice with SQLBolt, HackerRank, and quizzes.
- **Advanced Application:** Explore Google BigQuery
- **Clarify** questions

PROBABILITY, STATISTICS

Purpose

- Grasping probability's role in statistical reasoning
- Employing statistics for insightful data analysis

Probability Essentials

- Core Concepts: Probability, Random Variables, Distributions.
- Calculations: Marginal, Conditional Probability, Chain Rule.
- Relationships: Independence, Expectation, Variance, Covariance.
- Advanced: Common Distributions, Bayes' Rule, Information Theory, Monte Carlo, Markov Chains.

Statistics Fundamentals

- Basics: Descriptive Statistics, Combinatorics.
- Deep Dive: Distributions, Sampling, Hypothesis Testing, Model Estimation.

VISUALISATION

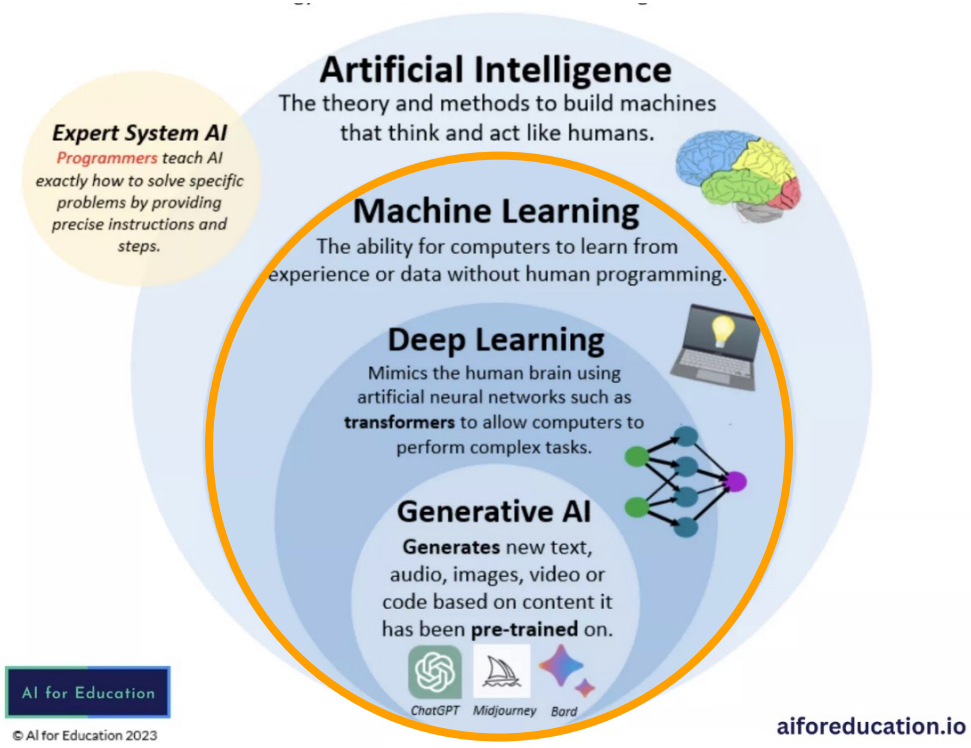
Purpose

Learn to craft web pages, render interactive charts, and build web apps efficiently.

- **D3 Module:** Master interactive visualizations using JavaScript, handling the DOM (Document Object Model), SVG (Scalable Vector Graphics) /CANVAS (HTML Canvas element for bitmap rendering), and data dynamics.
- **Plotly Module:** Create advanced visualizations with Plotly Express and develop web apps with Plotly DASH.

DSR ROADMAP: MACHINE LEARNING FUNDAMENTALS

- DS Fundamentals
- ML fundamentals
- Object-oriented Python and Software Architecture
- Trees
- Time Series (online)



DSR ROADMAP:
MACHINE LEARNING FUNDAMENTALS

Purpose

ML Fundamentals – foundational tools to construct, evaluate, and refine ML models for real-world applications

Object-Oriented Python & Software Architecture - to ensure the development of maintainable, scalable, and reliable data science systems with high code quality and efficiency

DS Fundamentals – introduce you to a data science workflow - exploratory data analysis (EDA), data cleaning & preparation (fill in NULL values, oversampling etc.), feature engineering, model selection & training, model evaluation, results delivery

Trees – understand how trees capture complex relationships & provide robust results (bagging vs. boosting)

Time Series – understand, model, and forecast data that changes over time

Tips – absolute must-haves

Some HRs may have questions with ready answers to filter out candidates already during the screening interview

If you get the chance, go through the notebooks in advance & identify questions or potential points you have to optimise

The best way to internalise these concepts is by coding – take a look at projects on Kaggle, towardsdatascience.com and re-do them. Even if you copy code, type it! Speaking from experience – it really does make a difference, as it makes you aware of things you would not notice if only copy-paste-ing!

Useful literature

The Elements of Statistical Learning (Jerome H. Friedman, Robert Tibshirani, & Trevor Hastie)

Data Analysis and Data Mining: An Introduction (Adelchi Azzalini & Bruno Scarpa)

MACHINE LEARNING FUNDAMENTALS

Purpose

Foundational tools to construct, evaluate, and refine ML models for real-world application

Machine Learning Overview

- What ML is (not) able to achieve
- The ML Project lifecycle
- Machine learning post-ChatGPT

Data Understanding

- Labeled vs unlabeled data (Supervised vs Unsupervised Learning)
- Structured vs unstructured data (Tables vs Images/Text/Sound)

Core ML Concepts and Techniques

- Loss functions
- Distance Metrics
- Embeddings aka representation learning
- Performance metrics vs loss functions
- The bias vs variance trade-off
- Evaluating performance with train, test, and validation sets
- Hyperparameters vs parameters
- Gradient descent

MACHINE LEARNING FUNDAMENTALS

Supervised Learning

- Tasks of supervised learning (regression and classification)
- Baseline models
- Algorithms for Supervised learning:
    - Linear and logistic regression
    - Decision trees, random forests, and gradient boosting
    - Neural networks

Unsupervised Learning

- Tasks of unsupervised learning (clustering and dimensionality reduction)
- Algorithms for Unsupervised learning:
    - K-means
    - Principal component analysis

OBJECT-ORIENTED PYTHON & SOFTWARE ARCHITECTURE

Purpose
To ensure the development of maintainable, scalable, and reliable data science systems with high code quality and efficiency.

- **OOP Fundamentals**
  - Utilize classes and objects for modelling data science problems.
  - Implement robust logging to track the execution and issues.
  - Develop error handling strategies to manage exceptions gracefully.

- **Testing Approaches**
  - Write acceptance tests to validate software meets business requirements.
  - Conduct unit tests to ensure individual components function correctly.

- **Software Architecture Concepts**
  - Adopt software architecture patterns that suit data science and AI projects, facilitating future growth and integration.

DS FUNDAMENTALS

Purpose

To develop a structured idea of the data science workflow.

Visualization Techniques

- Mastering Matplotlib for data representation.
- Dimensionality reduction with t-SNE and PCA for insightful visuals.

Data-Cleaning Essentials

- In-place column cleaning methodologies.
- Necessary steps for dataset preparation.

Feature Engineering Strategies

- Techniques for crafting new, informative features.
- Enhancements to boost model performance.

DS FUNDAMENTALS

Linear Models and Beyond

- Introduction to linear modelling techniques.

Model Selection & Evaluation

- Guidelines for experimenting with different models.
- Criteria for model selection and hyperparameter optimization.

Interpretation and Feature Selection

- Understanding linear model coefficients and applying LIME for interpretability.
- Univariate and stability selection methods for feature importance.

TREES

Foundational tree models for classification and regression tasks.

Ensemble Methods - Strength in Numbers:

- **Bagging:** Combining multiple trees to reduce variance and improve stability.
- **Random Forest:** A bagging technique with a forest of decision trees for robust predictions.
- **Boosting:** Sequentially building trees to correct previous errors and enhance performance.
- **Adaboost & Gradient Boosted Trees:** Specialized boosting methods focusing on different aspects of error reduction.

Advanced Techniques

- **Encoding Categorical Data:** Techniques like mean, label, and target encoding to transform categorical variables for improved tree and ensemble model performance.

TIME SERIES (ONLINE)

Purpose
To enable participants to understand, model, and forecast data that change over time, providing insights that are crucial for decision-making in various domains.

Understanding Time Series Data
- **Fundamentals:** Tasks in Time Series Analysis and unique challenges.
- **Key Properties:** Explore trend, seasonality, and noise within time series data.

Analytical Techniques
- **Autocorrelation Studies:** Delve into Autocorrelation and Partial Autocorrelation to understand data dependencies.
- **Decomposition:** Break down time series into its components to analyse separately.

Smoothing Methods
- **Moving Averages and Exponential Smoothing:** Techniques to smooth out short-term fluctuations and highlight longer-term trends or cycles.

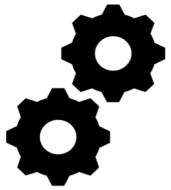
Predictive Techniques
- **Baseline Approaches:** Next day prediction, Moving Averages, Exponential Moving Averages.
- **Classical Models:** ARIMA, Holt-Winters method.
- **Modern Approaches:** Leveraging Machine Learning for advanced time series forecasting.

Toolkits for Implementation
- **SkTime & Darts:** Introduction to specialized libraries for time series analysis and modelling.

DSR ROADMAP: MINI COMPETITION

- 3 days of real data science work 😍👨‍💻👩‍💻

DSR ROADMAP: MINI COMPETITION



Purpose – during the competition you get to experience:

How to work in a team – both technically (Git & Bash) & conceptually (task distribution, tackling dependencies, data strategy)

The difference stages of a data science project – data exploration, data cleaning & preparation, feature engineering, model choice, model training, model testing & results delivery

Working on a realistic project

Tips

Enjoy it!

If you have the time after that, develop a clean solution on your own. It can be a good template for future reference

DSR ROADMAP: DEEP LEARNING

Fundamental Deep Learning Concepts

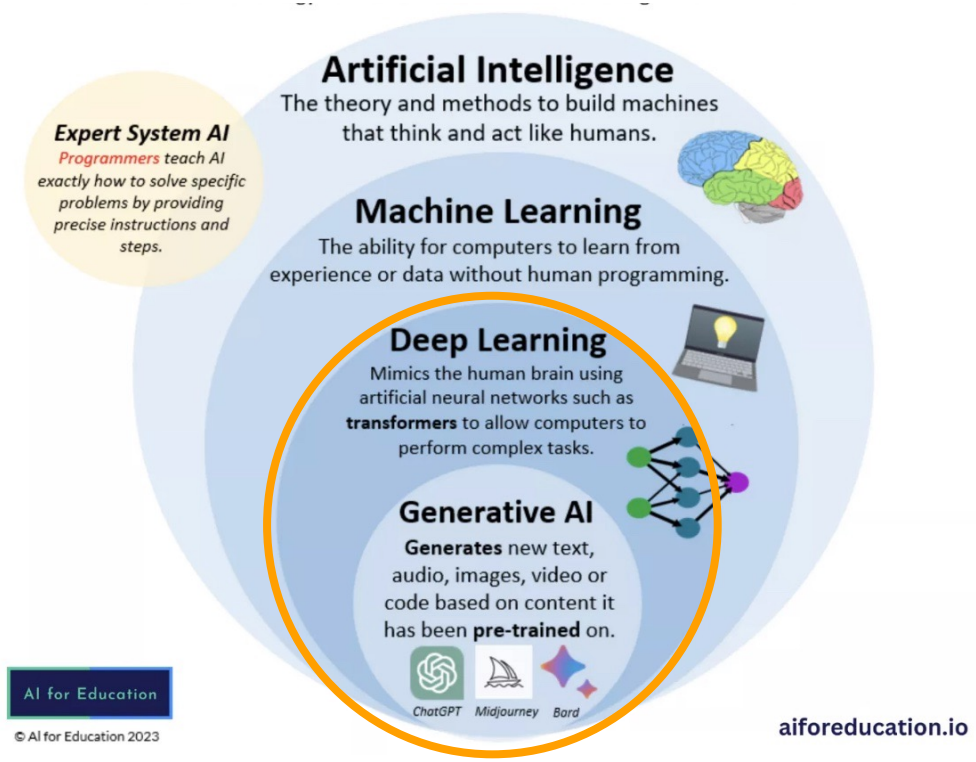
- Backpropagation
- Practical Reasoning
- Complexity Theory Fundamentals

Deep Learning Techniques & Frameworks

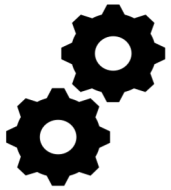
- Deep Learning with TensorFlow & PyTorch
    - Basics (PyTorch)
    - NLP, Transfer Learning & Representation (TensorFlow)
    - Image Processing (TensorFlow)
    - Computer Vision (PyTorch)
- Deep Reinforcement Learning
- Geometric Deep Learning

Advanced Deep Learning Applications

- Debugging Deep Learning Models
- LLM fine-tuning & Quantisation (online)
- RAG (Retrieval Augmented Generation)



DSR ROADMAP:
DEEP LEARNING FUNDAMENTALS & FRAMEWORKS



Purpose – this is where the real fun starts!

Backpropagation - to understand the core mechanism by which neural networks learn and adjust to improve accuracy (very efficient derivative calculation).

Deep Learning Basics – shows you the "deep" in deep learning - layers of increasingly meaningful representations.

Natural Language Processing (NLP) – to understand how text is represented in deep learning & how the attention mechanism work

Computer Vision – understand how images are represented in deep learning, processed & learn about computer vision tasks such as image classification, object detection, image segmentation

Deep Reinforcement Learning - uncover how agents learn from interactions to make decisions, applying reinforcement learning strategies

Geometric Deep Learning - To understand how deep learning can be applied to data structured as graphs and manifolds.

Tips

If you can, prepare before the lectures. Do an online course in deep learning. For ex. Coursera's Deep Learning

Definitely pay attention in these lectures – for some of you, it can be a lot to take in. Take some time after the lectures to revisit what was done during the day

PyTorch vs. Tensorflow before & PyTorch vs. Tesorflow now

Useful literature

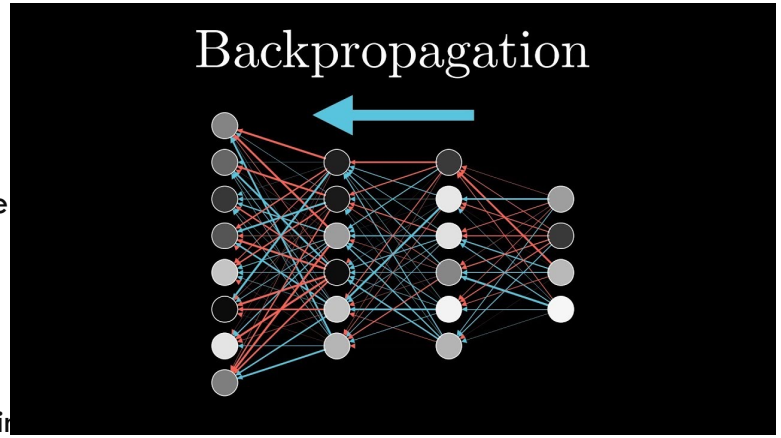
Deep Learning (Ian Goodfellow, Yoshua Bengio, Aaron Courville) – THE BIBLE of deep learning

Deep Learning with Python (François Chollet)

Set a GPU on AWS

Set a GPU on Google Cloud

DSR ROADMAP:
DEEP LEARNING FUNDAMENTALS



(Amazing!) Source:
https://www.youtube.com/watch?v=Ilg3gGewQ5U

Backpropagation in Neural Networks

- Master the backpropagation algorithm from the basics of perceptrons to advanced multilayer networks.
- Implement Forward (calculate outputs using activation functions) and Backward Pass (calculate gradients using the chain rule)
- Implement neural network training and optimization using Python.

Practical Reasoning for AI

- Sharpen skills in argument analysis and applied logic to understand and construct strong logical connections within AI systems
- Use cases: prompting, construction of RAG solutions / Agents

Complexity Theory in Machine Learning

- Computational complexity
- Big O Notation (upper bound of an algorithm's running time or space requirements in terms of the size of the input data.)
- Trade-offs between memory and computation in algorithms & data structures
- Learn to select & optimise AI models for efficiency, sclability, and real-world deployment

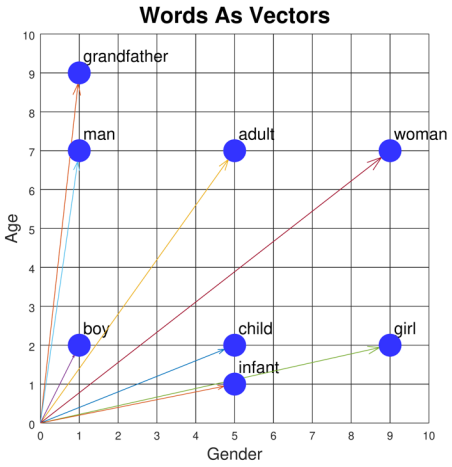
DSR ROADMAP:
DEEP LEARNING TECHNIQUES & FRAMEWORKS

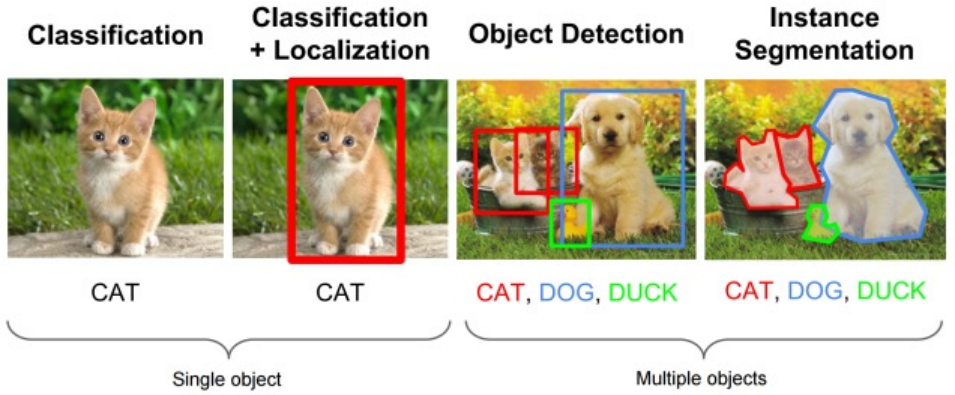
Basics (PyTorch)
- Fundamental PyTorch operations
- Structure multilayer perceptrons
- Understand convolutional networks, and transformers to grasp the layered representations.

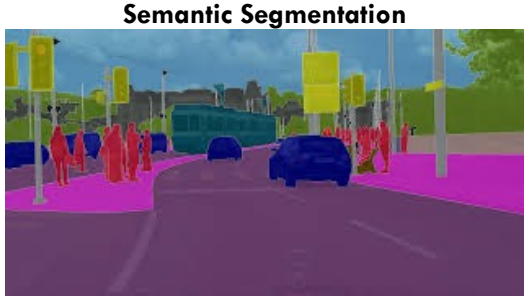
NLP, Transfer Learning & Representation (TensorFlow) (everything from scratch)
- Deep dive in word embeddings
- Attention-based models
- Build your own GPT

Image Processing (TensorFlow)
- Basic Image classification using convolutional neural networks (CNNs)
- Transfer Learning with CNNs
- Region-based CNNs for object detection tasks
- Region of Interest Identification and Intersection over Union (IoU) as a metric for evaluating object detection models
- U-Net architecture for image segmentation tasks
- GoogLeNet (Inception architecture) for image classification

Computer Vision (PyTorch)
- Image similarity
- Image data pipelining - processing, augmentation, shuffling, batching
- Training classifiers with FastAI
- Object detection & image segmentation using Meta's Detectron2







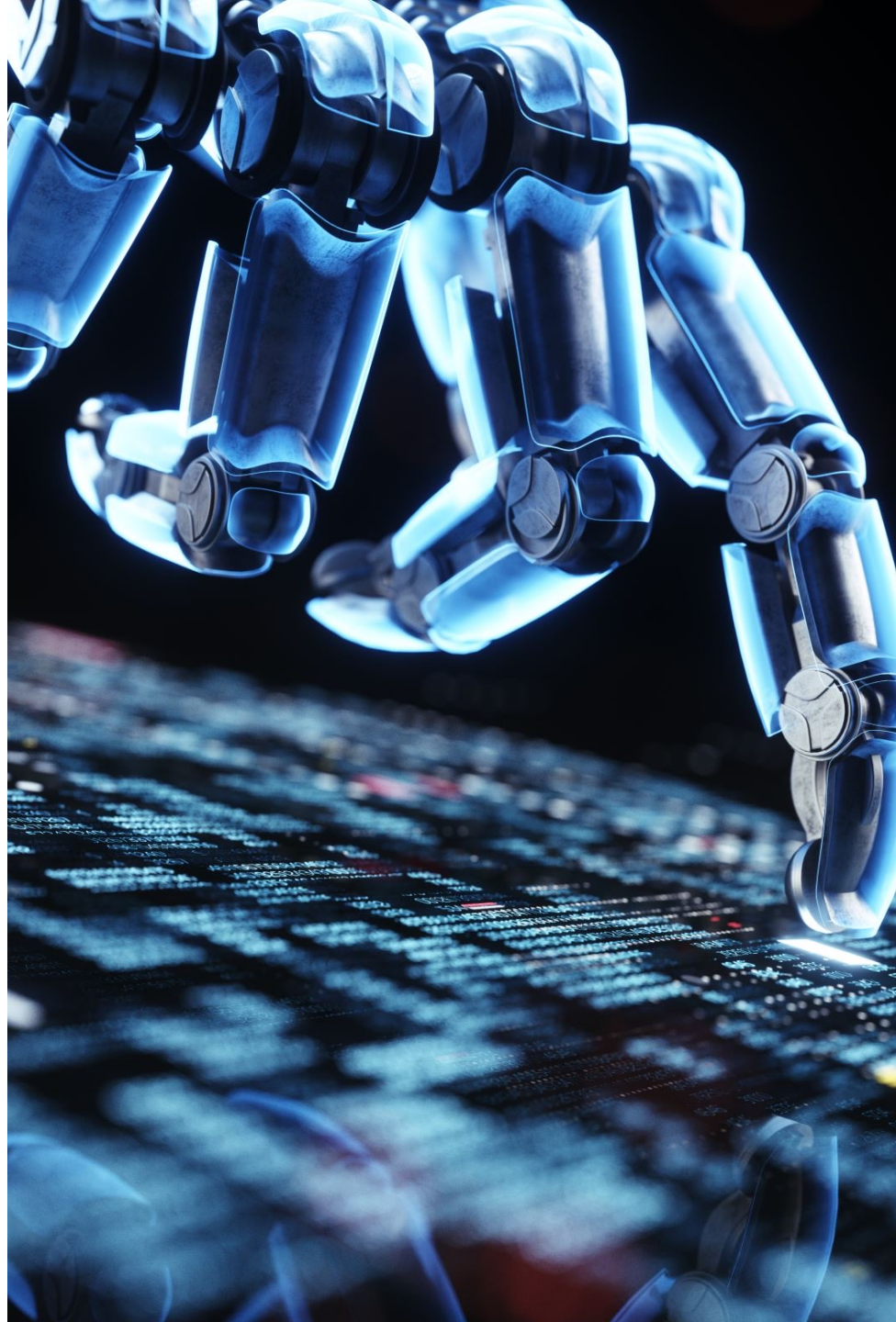
DSR ROADMAP: DEEP LEARNING TECHNIQUES & FRAMEWORKS

Deep Reinforcement Learning

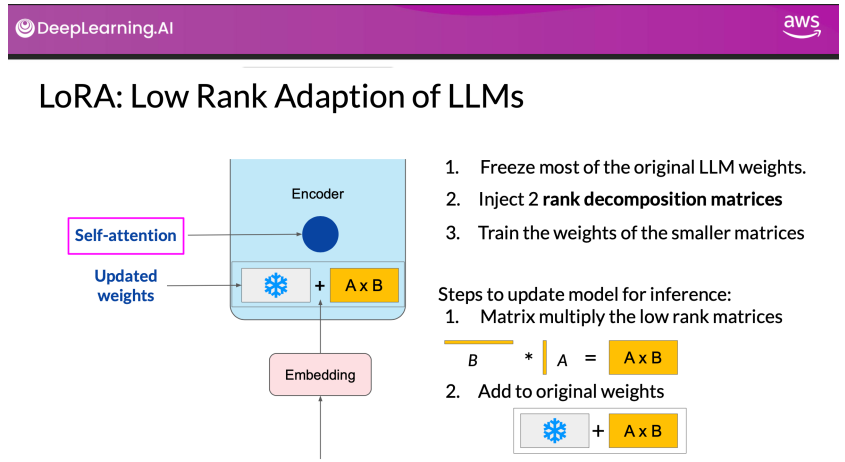
- Concentrates on teaching agents to make decisions by trial and error. Typically applied in areas like robotics, autonomous systems, game playing, and real-time decision-making in finance
- Core Principles of RL - understand RL's fundamentals, including agents, environments, actions, states, rewards, and policies, and how they guide decision-making through trial and error.
- Practical Frameworks - explore RL's implementation in various sectors using frameworks like OpenAI Gym, focusing on applications in robotics, gaming, and finance.
- Advanced Techniques and Trends - delve into the mechanisms and recent developments of RL, covering both theoretical methods like value and policy iteration and cutting-edge advancements like deep RL and multi-agent systems.

Geometric Deep Learning

- Focus on GDL's application in Graph Neural Networks (GNNs). Commonly applied in social network analysis, 3D shape processing, and molecule structure prediction.
- Architecture & functioning of various GNNs including
    - Graph Convolutional Networks (GCNs)
    - Attention-based GNNs
    - Message Passing Neural Networks (MPNNs).
- Equivariant GNNs, understanding how they maintain symmetry and consistency in data representations.



DSR ROADMAP: ADVANCED DL APPLICATIONS



Debugging Deep Learning Models

- Most common challenges & bugs with Deep Learning Models
- Intuition on how to identify & treat arising issues with tunning code in deep learning
- LLM-base debugging approach will be shown in detail
- A small hand-on project to apply the newly acquired skills

LLM Quantisation and Fine-tuning (online)

- Typical Architecture & Steps
- Model Quantization: Quantization reduces a model's precision by converting floating-point numbers to lower-bit representations, enhancing computational efficiency and reducing memory footprint.
    - Example: Deploying a quantized neural network on a mobile device to enable real-time image recognition with limited hardware resources.
- Flash Attention: Flash Attention is a technique to optimize the attention calculation in transformers for improved speed and efficiency in processing sequences.
    - Example: Using Flash Attention to speed up language translation tasks in an AI-powered chatbot, resulting in quicker response times.
- RoPE (Rotary Positional Embeddings): RoPE integrates positional information into attention mechanisms, allowing the model to better capture the order and relationship between elements in a sequence.
    - Example: Implementing RoPE in a text generation model to more accurately reflect syntactic structures based on word positions.
- Parameter-Efficient Fine-Tuning (PEFT) and Low-Rank Adaptation (LoRA): LoRA is a method for adapting large pre-trained models to specific tasks without extensively altering the original parameters, focusing on efficiency and adaptability.
    - Example: Fine-tuning a pre-trained language model for a legal document analysis task using LoRA to introduce task-specific adjustments while retaining the model's extensive general knowledge.

DSR ROADMAP: ADVANCED DL APPLICATIONS

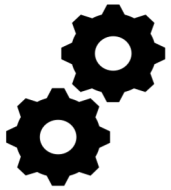
Retrieval Augmented Generation (RAG)

- A technique combining retrieval of information with generative models for advanced NLP tasks.
- RAG Components and Functionality: Explore RAG's system design including text embedding, vector storage, retrieval, and optimization strategies.
- Practical Application: Engage with various tools and frameworks for RAG, addressing bias mitigation and ethical considerations in NLP.
- If you'd like, check https://www.langchain.com/ & https://www.llamaindex.ai/

DSR ROADMAP: PRACTICAL DATA SCIENCE

- Practical Aspects of Data Science
- Practical Aspects of Machine Learning / MLOps
- Test-driven development

DSR ROADMAP: PRACTICAL DATA SCIENCE



Purpose – practical advice, real-life project approach, deployment

Practical DS – learn to transform data science projects into interactive web applications using Streamlit.

ML Ops – learn to expose your data science projects by deploying endpoints & developing ML pipelines

Test-driven development - to equip participants with comprehensive skills in Test-Driven Development (TDD) using Python, fostering clean and maintainable code practices.

Tips

Try not to skip these lectures. They showcase what you have to deal with once you start working as a data scientist..

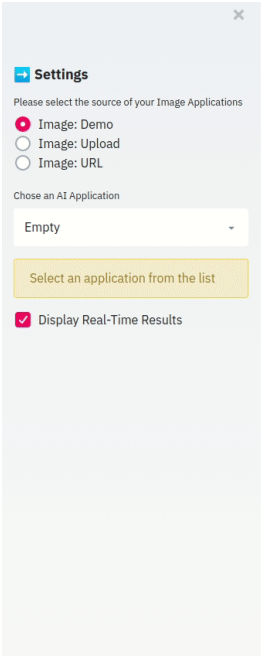
DSR ROADMAP: PRACTICAL DATA SCIENCE

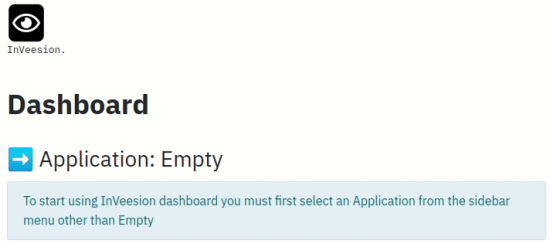
Practical DS

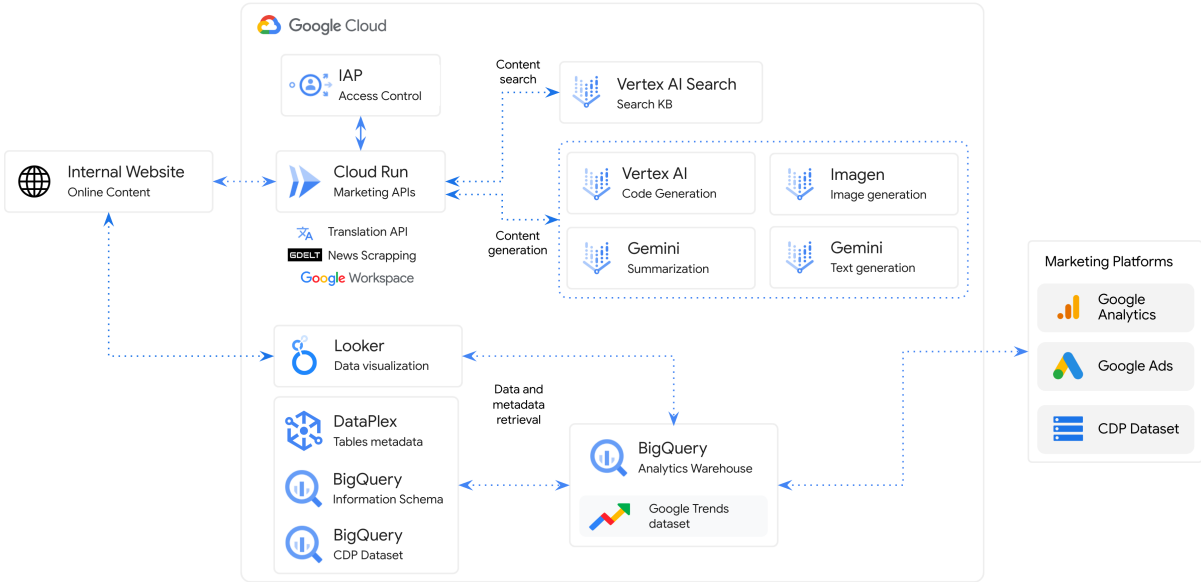
- Introduction to Streamlit: Gain a foundational understanding of Streamlit for web application deployment in data science.
- Project Deployment: Learn to transform data science projects into interactive web applications.
- Hands-On Experience: Apply your knowledge by building and deploying a small project with Streamlit by the end of the session.

ML Ops

- FastAPI and VertexAI: Learn to deploy machine learning models using FastAPI and manage ML pipelines with VertexAI.
- Preparation: Ensure a Google Cloud Platform account is set up prior to the session for practical deployment exercises.







DSR ROADMAP: PRACTICAL DATA SCIENCE

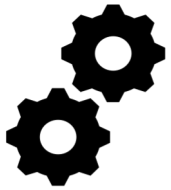
Test-driven development

- TDD Fundamentals: Master the TDD cycle (Red-Green-Refactor), understand its benefits in software development, and learn the distinctions between TDD and Behavior-Driven Development (BDD).
- Testing Techniques: Gain proficiency in writing, organizing, and managing test cases including unit and integration testing, and applying testing best practices.
- Practical TDD Applications: Engage in hands-on exercises to implement TDD on Python projects, including writing tests that initially fail and refactoring code to pass tests.
- Advanced TDD with GitLab: Utilize GitLab for issue tracking, milestone management, and setting up continuous integration and deployment pipelines.
- Best Coding and Testing Practices: Explore best practices in coding and test-driven development, including the use of code formatters and the Python Enhancement Protocol (PEP).

DSR ROADMAP: SOFT SKILLS

- Business Communication
- Career Support

DSR ROADMAP: SOFT SKILLS



Purpose – what to expect at interviews, strategy & effective communication.

Business Communication – improve interview skills and strategic communication techniques for effective leadership and stakeholder management

Career Support – ask questions related to how to get a job in data science

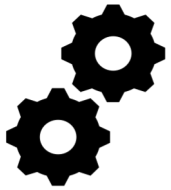
Tips

Prepare questions for the Career Support Class

For the communication class, pick a topic / situation that was challenging for you. You will get very useful feedback & actionable advise.

DSR ROADMAP:
THE FINAL PROJECT

DSR ROADMAP:
THE FINAL PROJECT – LAST 3 WEEKS



Purpose

”Probieren geht über Studieren” (“The proof of the pudding is in the eating.”) - practice all you have learned so far.

Data science is about doing and showcasing your skills with meaningful projects:

- You found a topic, for which data science offers a solution / optimisation / invention
- You dealt with the issue of finding or creating data
- You delivered some kind of result

DSR ROADMAP:
THE FINAL PROJECT - HOW TO START RESEARCH ON FINDING ONE

Disclaimer – no golden formula for this

One way is to pick a topic, that addresses a challenge in an industry, in which you later want to work
- Batch 25 "Sound of failure" reduce industrial downtime by diagnosing the failure of machines using their acoustic footprint; (Non-GenAI)
- Batch 34 MAG (Medical Advice Generator) – a LLM bot, that can help medical professionals studying for the Kenntnisprüfung

Address an issue that can solve an every-day problem of many
- Batch 22 "Deep Food" – an app that recognises ingredients from your fridge and gives you recipe suggestions (Non-GenAI)
- Batch 34 "AI to enhance doctors (not replace them)" – involves recording patient-doctor consultations, transcribing them with audio-to-text models, and then using Large Language Models (LLMs) & vector database to summarize and format into professional clinical documents; and facilitate efficient searching and interactive querying of clinical guidelines. (GenAI)

Take a look at existing research and see if you could improve it or it could inspire you to build upon it
- Paperswithcode – it will provide you with trending machine learning research and the code to it

Take a look at business magazines (for e.g. MIT Technology Review, The Economist, Harvard Business Review) and see what are popular data topics

Discuss with ChatGPT or / and BARD

DSR ROADMAP: THE FINAL PROJECT

Admin

Project formation process:

- Seek communication as soon as possible
- Brain-storm with each other
- The idea should be yours - you are expected to be creative

Provide an abstract (1 page / 1 paragraph) to Arun. Several advantages:

- You will have more out of your project and will save time and nerves ;) → if your project idea is similar to a past project, Arun will put you in touch with the respective team
- Gives you exposure to potential employers → there is the possibility to reach out to companies, if you'd like to do it in cooperation with one. **!Once you have decided to do the project this way, you are committed!**
- Better mentor match → Arun has enough time to reach out to mentors

Mentors - common mentoring (usually Mondays & Thursdays) and separate mentors, if a project topic was picked on time

DSR ROADMAP: THE FINAL PROJECT

Admin

Where – flexible – in the office or remote

Topic Choice – the latest by **26th – 28th of November**

Mandatory Presentation Rehearsal, i.e. Portfolio Project Review – **16.12.2024**

- Test the tech
- You will receive good feedback and may be some last-minute ideas

Your demo day: **17.12.2024**

DSR ROADMAP:
THE FINAL PROJECT

Tips

Mentors – they are there to guide you, to discuss your ideas, but not to micro-manage or check your code.

Idea and teams – stick to them. This will help both you and your mentors.

Data availability is crucial – you can either find a good data set, generate it and / or use transfer learning. Either way – make sure you start early enough with gathering the data

The most complex model is not always the best one. Start with simple & quick solutions to test your approach

Your results may not always be what you'd expect them to be. However, the way to them is also interesting. Share this!

Idea for a presentation structure:

- Why did you choose this project – motivation, use cases, background info
- Data used and interesting aspects, challenges and how did you tackle them
- Deep dive into the models used – what models did you try, which one is the final one and why, final model performance
- Next steps
- Demo / Recording

If you need a GPU, try to set up one before that

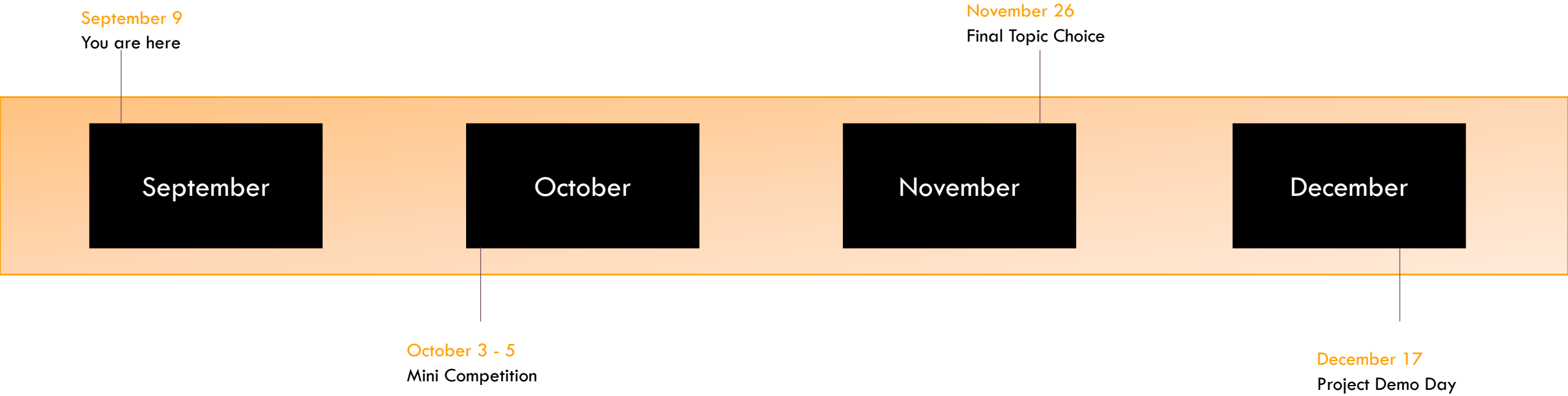
TO GPT OR NOT TO GPT

TO GPT OR NOT TO GPT

- Sparing partner – definitely
- Still learn – if you are under time pressure, use the code. However, make sure you understand it and you come back later to practice it
- Modern GPTs
  - ChatGPT (by OpenAI)
  - Gemini (by Google)
  - GitHub Copilot (specifically for code)

MILESTONES

MILESTONES

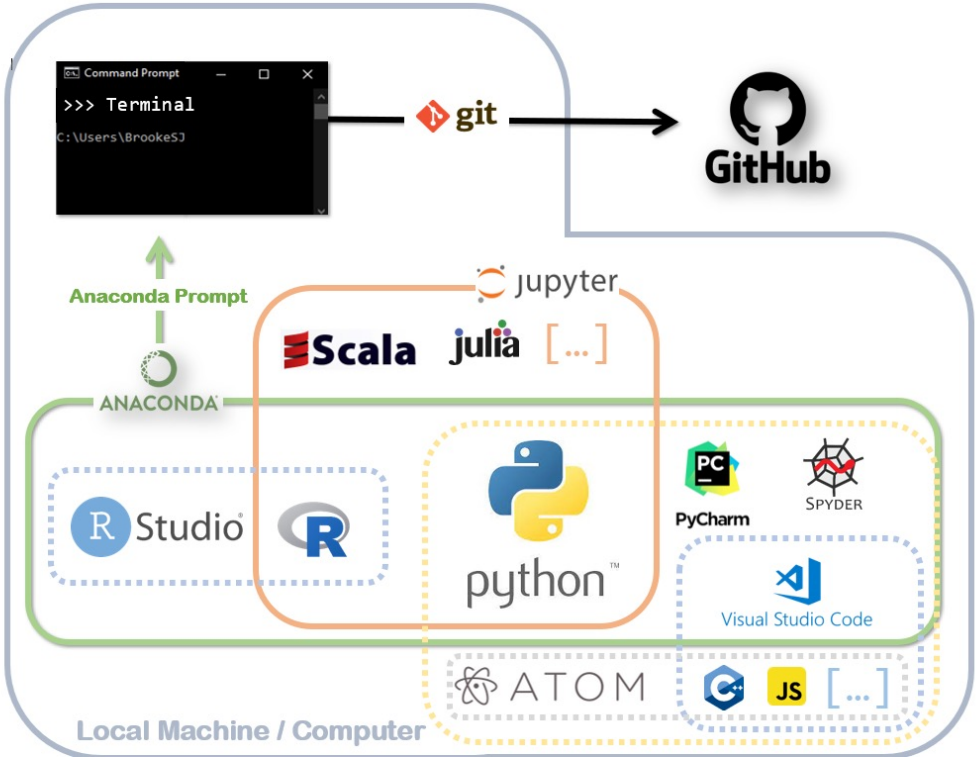

September 9
You are here
November 26
Final Topic Choice
September
October
November
December
October 3 - 5
Mini Competition
December 17
Project Demo Day

GET YOU READY FOR DS

GET YOU READY FOR DS

- Shell
  - A computer program that serves as an interface between the user and the operating system.
    It allows you to access and control various tools and services.
    With the Shell, you can navigate your file system, install and manage packages, run scripts, and interact with other tools like Anaconda, Python, and Git.
  - You access it via a terminal emulator – ex. for macOS - Terminal, iTerms. List of Terminals
  - Command-line interface shells use specific scripting language - bash / zsh
- Anaconda – a Python 🙃 distribution platform including some useful applications such as PyCharm, Jupyter Notebook & Lab
- Git – a version control tool. "Version control is a system that records changes to a file or set of files over time so that you can recall specific versions later."
- Integrated Development Environment (IDE) – Visual Studio Code, PyCharm, Spyder
- Jupyter Lab / Jupyter Notebook – Interactive data science environment – more visual than IDE, therefore it is used for educational & presentation purposes

GET YOU READY FOR DS



Source: https://www.sianbrooke.co.uk/dr-brookes-blog/coding-in-python-managing-packages-with-conda-and-pip

GET YOU READY FOR DS … ALSO

- Google Colab(oratory)

GET YOU READY FOR DS

- Useful:
  - How to Set Up a Data Science Project
  - Terminal for Mac users – iTerms (features)
  - Terminal modification, if you'd like it to be more colourful
  - Bash vs. zsh
  - Difference between conda and pip
  - Git in a nutshell
  - Fundamentals of computing & programming

SET UP A SIMPLE DATA SCIENCE PROJECT

HOW TO SET UP A DATA SCIENCE PROJECT

Make sure you have Git, Anaconda, Visual Studio Code & PyCharm installed and ideally have a GitHub account.

Let's set a Data Science Project from scratch by following the instructions here: https://github.com/Iskriyana/dsr-teaching-setup

- Maintain a learning structure - all of the notebooks can be used for future reference while you are preparing for an interview or when you are working on a future task
- My approach
  - Environment per lecture – most of the teachers already have it in their prep instructions
  - Folder / Repo per lecture
  - Notes per lecture
  - Bookmarks per lecture

SET UP A SIMPLE DATA SCIENCE PROJECT

Let's now do git clone https://github.com/lskriyana/dsr-teaching-setup.git

MEETUPS

MEETUPS IN BERLIN

- Data Science Retreat
- Generative AI on AWS (San Francisco, Global)
- GenAI Gurus - Generative Artificial Intelligence
- Berlin DataTalks Club & their slack
- Berlin Machine Learning Group
- meetup.ai
- Deep Learning Würzburg
- PyData
- Google Developer Group
- Berlin Computer Vision Group
- Advanced Machine Studying Group
- Women Who Code Berlin
- PyLadies Berlin
- Women Techmakers Berlin

THANK YOU & HAVE FUN!