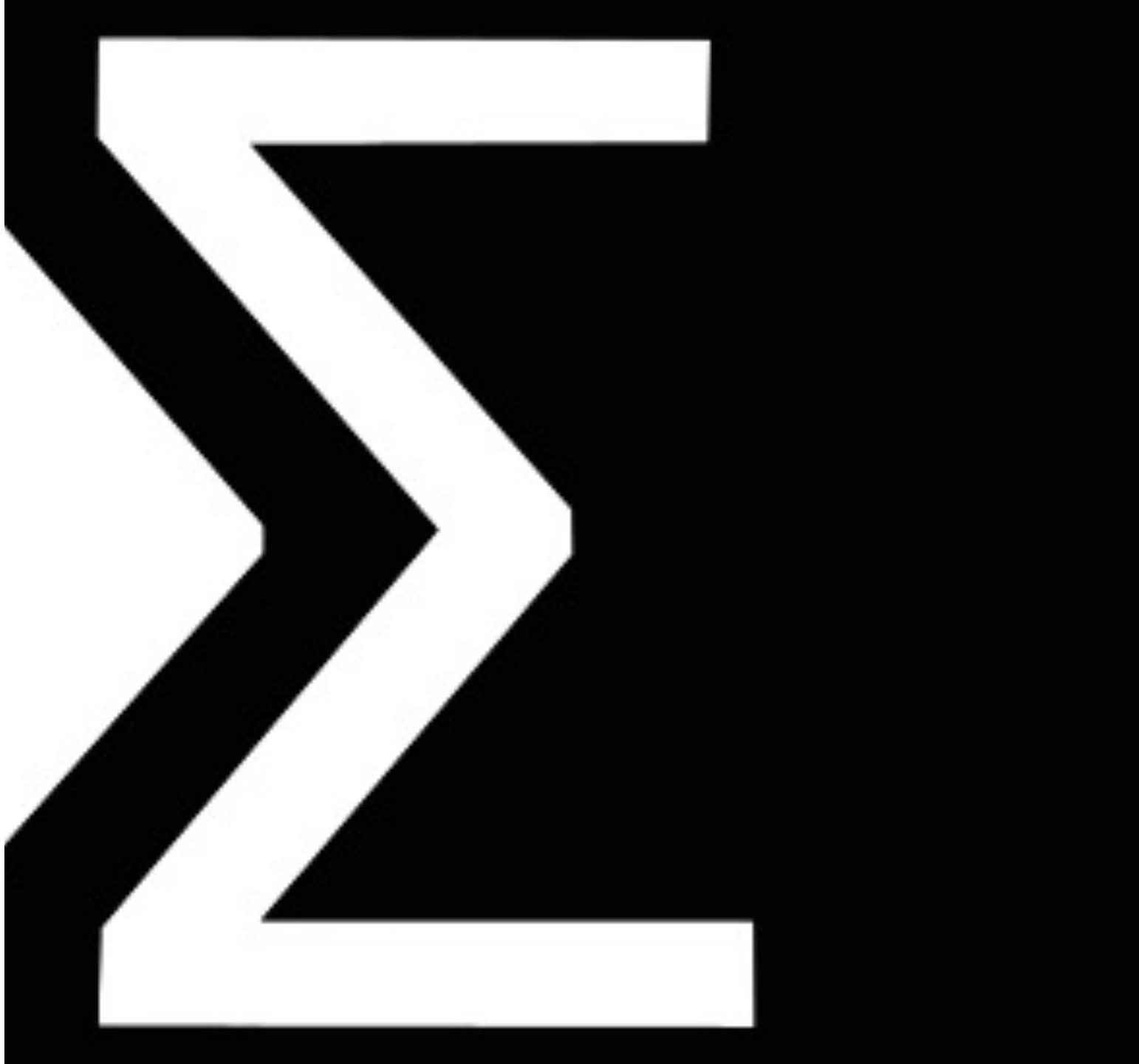
PROGRAM ROADMAP

Iskriyana Vasileva

Berlin, 26.04.2021



AGENDA

- Intro Round
- The DSR Roadmap
- Get You Ready for Data Science
- How to Set Up a Data Science Project
- An Example of a Data Science Project
- Cheat Sheets
- Some Meetup groups (in Berlin)

INTRO ROUND

INTRO

- About me:
  - A data enthusiast - 7 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

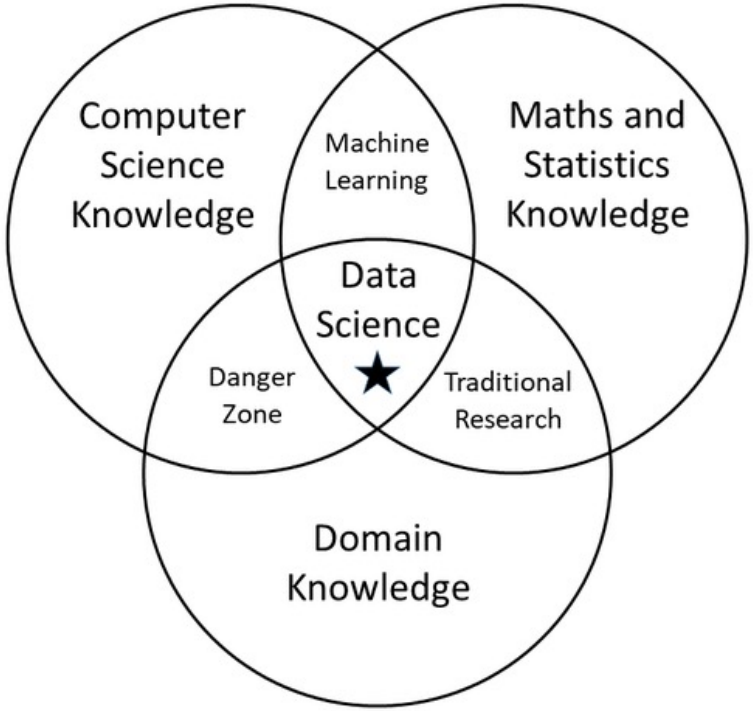
INTRO

- What about you?

THE DSR ROADMAP

THE DSR ROADMAP

- Theoretical & Technical Fundamentals
- Machine Learning Fundamentals
- Mini Competition
- Deep Learning (or the really “cool stuff”)
- Practical Data Science
- Soft Skills
- The Final Project



DSR ROADMAP: THEORETICAL & TECHNICAL FUNDAMENTALS

- Python
- Git & Bash
- NumPy
- Pandas
- SQL
- Visualisation
- Statistics, Probability
- Docker & Databases

DSR ROADMAP: THEORETICAL & TECHNICAL FUNDAMENTALS

Purpose – the tools enabling you to work as a data scientist

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

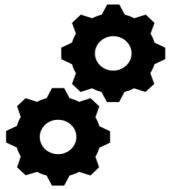
Data Wrangling with Python (Jacqueline Kazil, Katharine Jarmul)

Python for Data Analysis (Wes Mckinney)

DSR ROADMAP: MACHINE LEARNING FUNDAMENTALS

- DS Fundamentals
- ML fundamentals
- Trees
- Time Series

DSR ROADMAP:
MACHINE LEARNING FUNDAMENTALS



Purpose

DS Fundamentals – helps you structure a data science project from a data perspective. You get to do steps such as exploratory data analysis (EDA), data cleaning & preparation (fill in NULL values, oversampling etc.), feature engineering, model development & training, model testing, results delivery

ML Fundamentals – here you go through the "classics" of ML such as linear regression, logistic regression, evaluation metrics of classification etc.

Trees & Time Series – dedicated lectures only for these topics

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

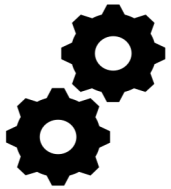
The Elements of Statistical Learning (Jerome H. Friedman, Robert Tibshirani, & Trevor Hastie)

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

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 (repository) & 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

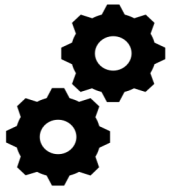
Enjoy it!

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

- Backpropagation
- Sequences
- Deep Learning Basics
- Natural Language Processing (NLP)
- Computer Vision
- Transfer Learning
- Reinforcement Learning

DSR ROADMAP: DEEP LEARNING



Purpose – this is where the real fun starts! Most probably this is why you chose this bootcamp

Backpropagation – central in deep learning – a mathematical technique for quickly calculating derivatives (the gradients).

Deep Learning Basics – shows you the "deep" in deep learning - layers of increasingly meaningful representations. The number of layers define the depth of the model.

Computer Vision – image classification using convolutional layers, data pipelining (processing, augmentation, shuffling, batching), regularisation, transfer learning or how to use pre-trained models, sound processing as spectrograms and as time series, variational autoencoders, GANs, briefly about image semantics

Natural Language Processing (NLP) – how to do a sentiment analysis with bag-of-words, non-DL approach to it (for e.g. using Gaussian NB), simple DL approach (for e.g. building Dense-Dropout layers using Sequential), what are word embeddings, LSTM, briefly about text generation, transfer learning and attention-based models

Reinforcement Learning - an agent receives information about its environment and learns to choose actions that will maximize some reward. For instance, a neural network that "looks" at a video- game screen and outputs game actions in order to maximize its score can be trained via reinforcement learning.

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

Try to set a GPU on your own (see literature section). If you decide to do a project with deep learning, you will most probably need a GPU.

Useful literature

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

Deep Learning with Python (François Chollet)

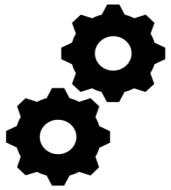
Set a GPU on AWS

Set a GPU on Google Cloud

DSR ROADMAP: PRACTICAL DATA SCIENCE

- Debugging
- Serverless
- Practical Aspects of Data Science
- Practical Aspects of Machine Learning

DSR ROADMAP: PRACTICAL DATA SCIENCE



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

Tips

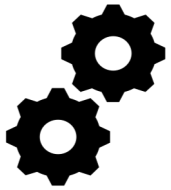
Try not to skip these lectures. They showcase what you have to deal with once you start working as a data scientist.

Debugging is extremely useful. In this class you will also get to know PyCharm.

DSR ROADMAP: SOFT SKILLS

- Communication
- Career Support

DSR ROADMAP: SOFT SKILLS



Purpose – what to expect at interviews, strategy & effective communication. Both teachers are extremely capable & know their stuff.

Tips

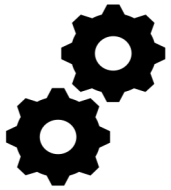
Prepare questions for the Career Support Class

Please, do respond on time for the preparation of the communication class. Kevin prepares accordingly

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



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;
- Batch 24 "CAN DEEP LEARNING RELIABLY PREDICT THE CONSUMPTION AND GENERATION OF RENEWABLE ENERGY?"

Address an issue that can solve an every-day problem of many
- Batch 25 "Check your pose" an app that detects your pose and provides you with feedback on the proper posture
- Batch 22 "Deep Food" – an app that recognises ingredients from your fridge and gives you recipe suggestions

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. The Economist, Harvard Business Review) and see what are popular data topics

DSR ROADMAP:
THE FINAL PROJECT

Tips

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

GET YOU READY FOR DS

GET YOU READY FOR DS

- Shell
  - A computer program which exposes an operating system's services to a human user or other program.
    It is named a shell because it is the outermost layer around the operating system.
    You can steer Anaconda, Python, Git etc. with it.
  - 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 – provides you with Python 🙃 and with some useful application around them 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) – 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

- 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

HOW TO SET UP A DATA SCIENCE PROJECT

HOW TO SET UP A DATA SCIENCE PROJECT

Make sure you have Git, Anaconda & 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

A DS PROJECT

A DS PROJECT

https://github.com/Iskriyana/nlp-product-sentiment-classification

CHEAT SHEETS

CHEAT SHEETS

- Anaconda:
- ohmyzsh:
- More:
  - https://github.com/Iskriyana/dsr-teaching-setup/tree/main/cheatsheets
  - https://github.com/ADGEfficiency/programming-resources/tree/master/cheat-sheets

MEETUPS

MEETUPS IN BERLIN

- Data Science Retreat
- 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!