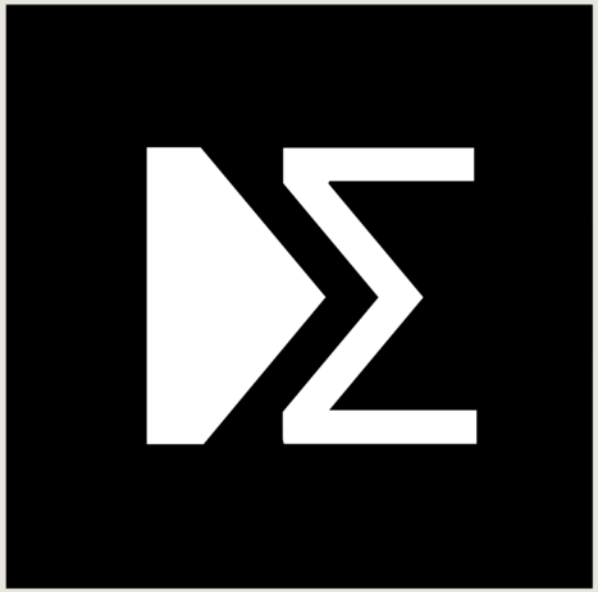
DSR ROADMAP

Lucas Aresin



DATA SCIENCE RETREAT®

SINCE 2014

AGENDA

INTRO ROUND — Let's get to know each other

Short break

THE DSR ROADMAP — What is the journey ahead like?

Short break

HANDS-ON — Set up a very simple data science project repo on GitHub

Short break

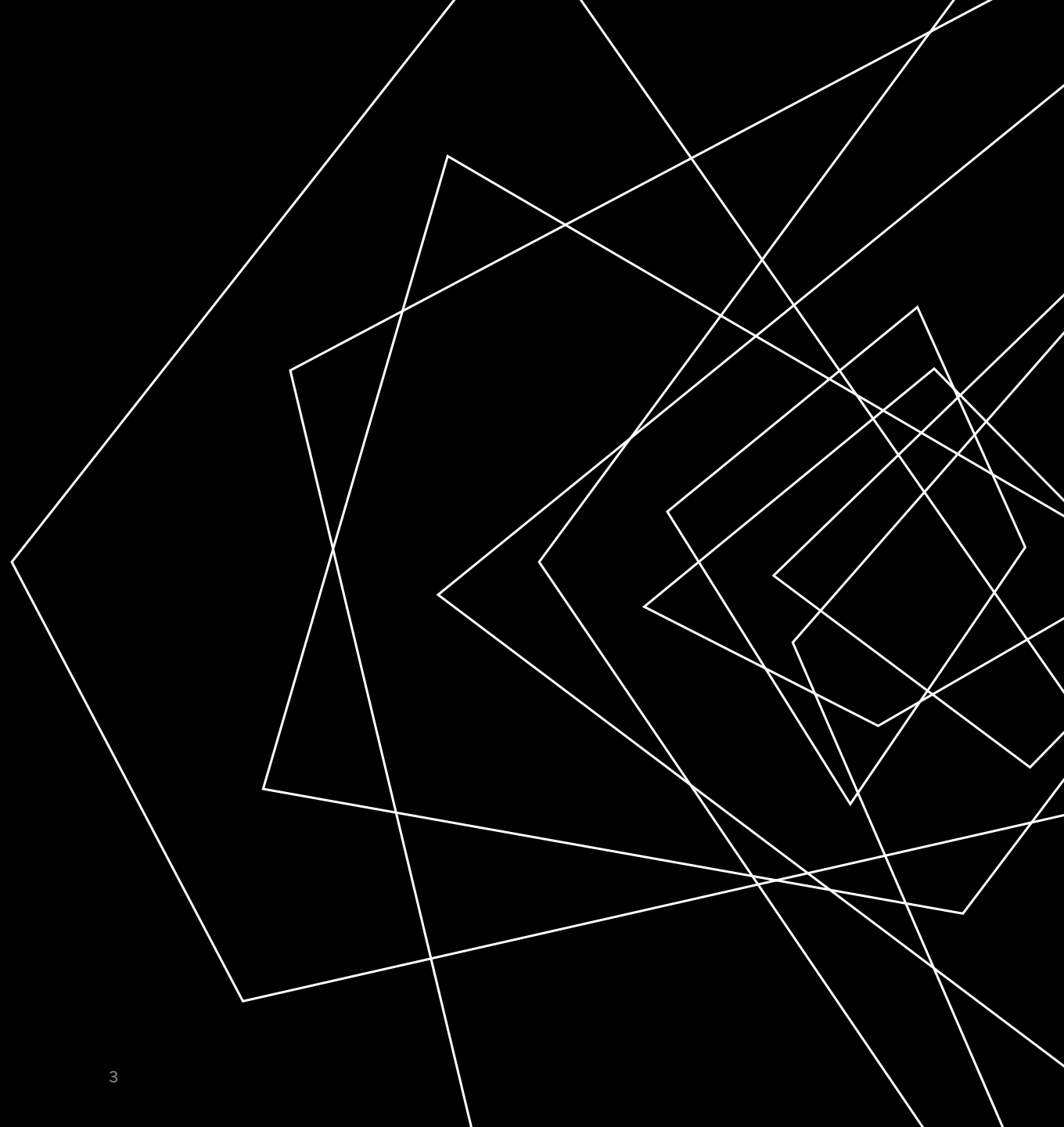
RESOURCES & QUESTIONS — Ask me anything + helpful links and tools

ABOUT ME

A self-taught data science and machine learning enthusiast. Speaks Java, JavaScript, Python.

I didn't understand how machines learn, so I learned how machines learn.

- Ask me anything
- Speak up if you don’t like something
- Contact me on LinkedIn



AND WHAT ABOUT YOU?

Your name

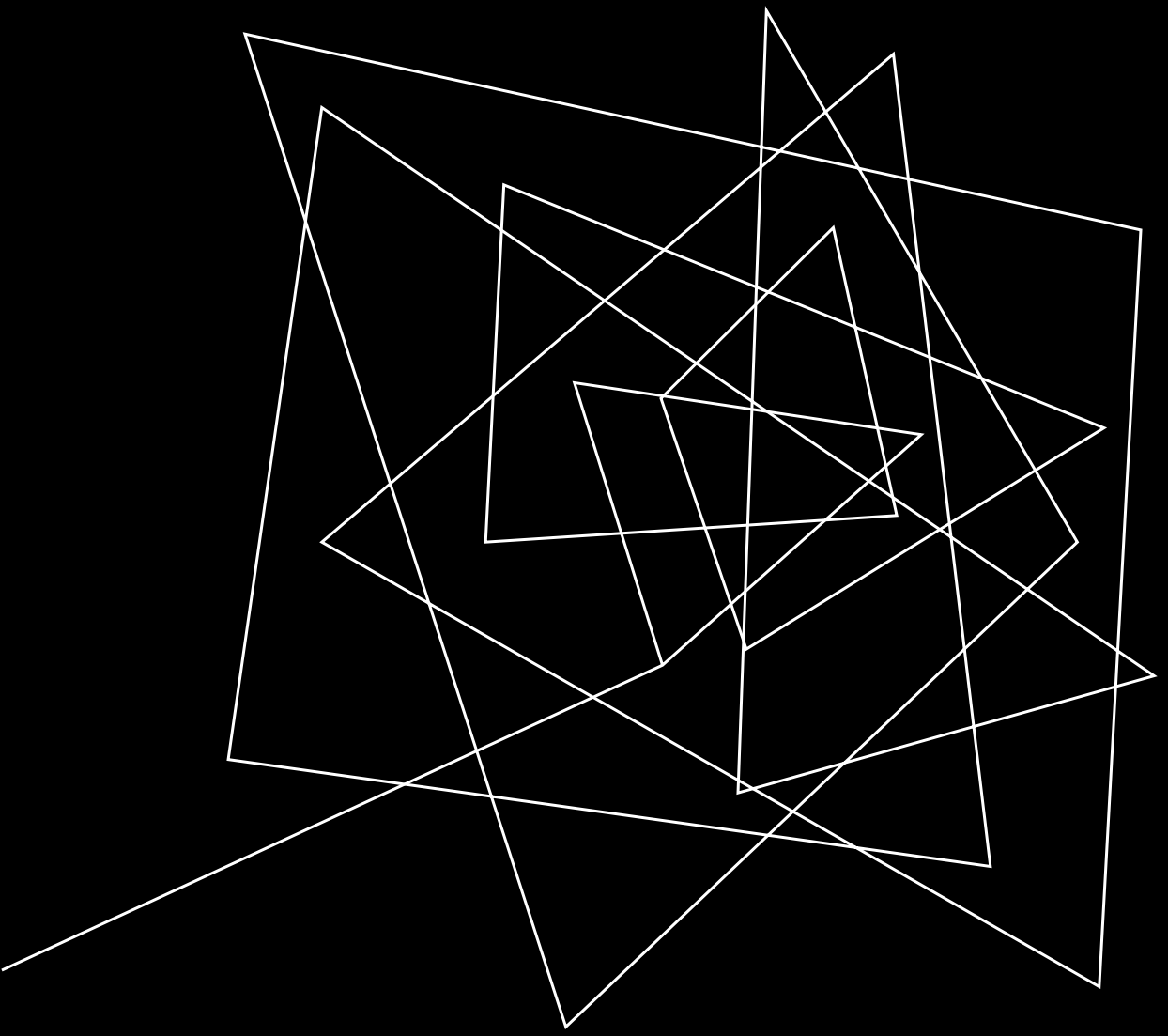
Your background

Why you came here

(and, if you'd like to share, your largest knowledge gap)

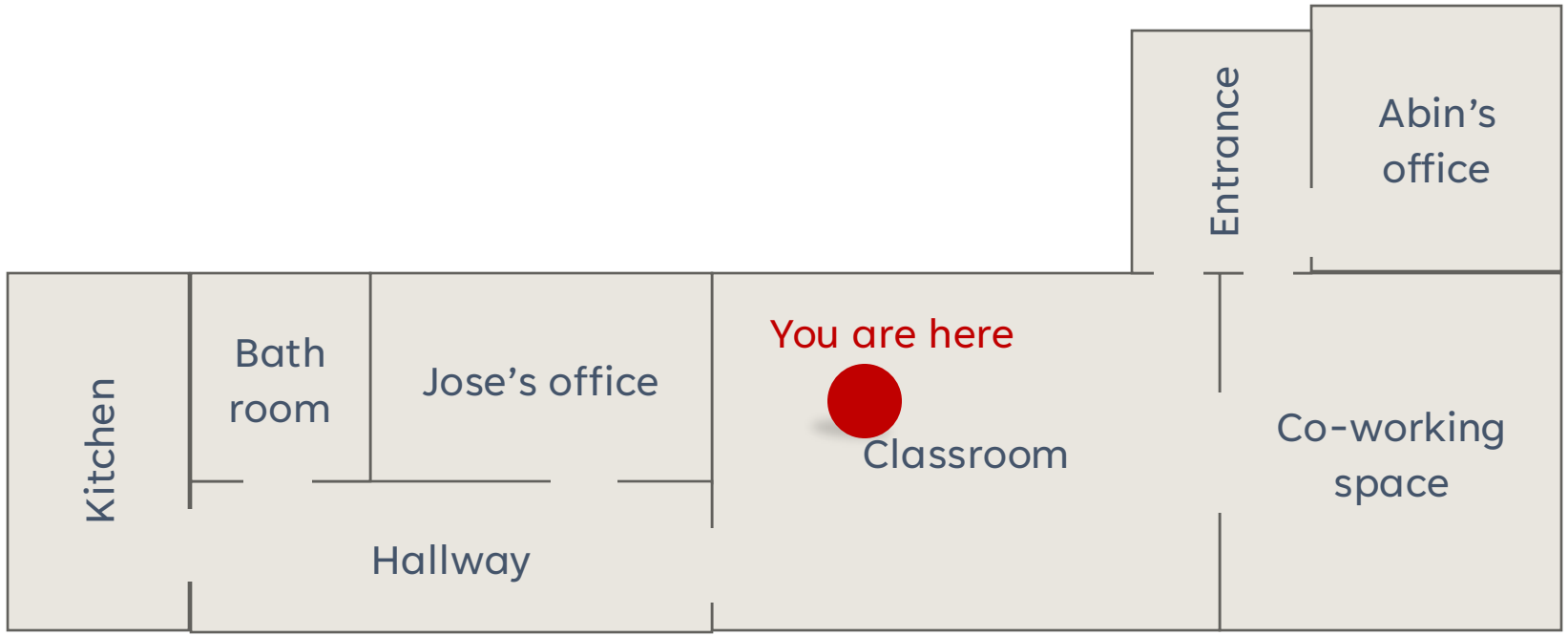
Hopes and dreams

(and, if you'd like to share, a book recommendation)



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

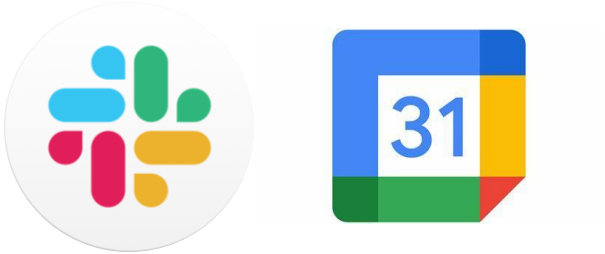
- They know what it’s like
- Ask them about their projects!

THEY ARE PEOPLE TOO ☺

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

LET’S GET SOME ADMIN OUT OF THE WAY

COMMUNICATION TOOLS



CLASS GUIDELINES

- Classes are from 09:30 to 17:30.
- Please be on time.
- Be prepared.

LECTURE ATTENDANCE

- If possible, be here in person
- Let people know when you are not coming

LET’S GET SOME ADMIN OUT OF THE WAY

“FREE” DAYS

- Project work
- 1:1s with Jose
- Prepare and reflect

FEEDBACK

- Communicate!
- Abin is always available

COURSE MATERIALS

- It can be a lot
- Don’t be overwhelmed
- Use it as your own knowledge base

DSR ROADMAP IN A NUTSHELL

1 TECHNICAL FUNDAMENTALS

2 DS AND ML FUNDAMENTALS

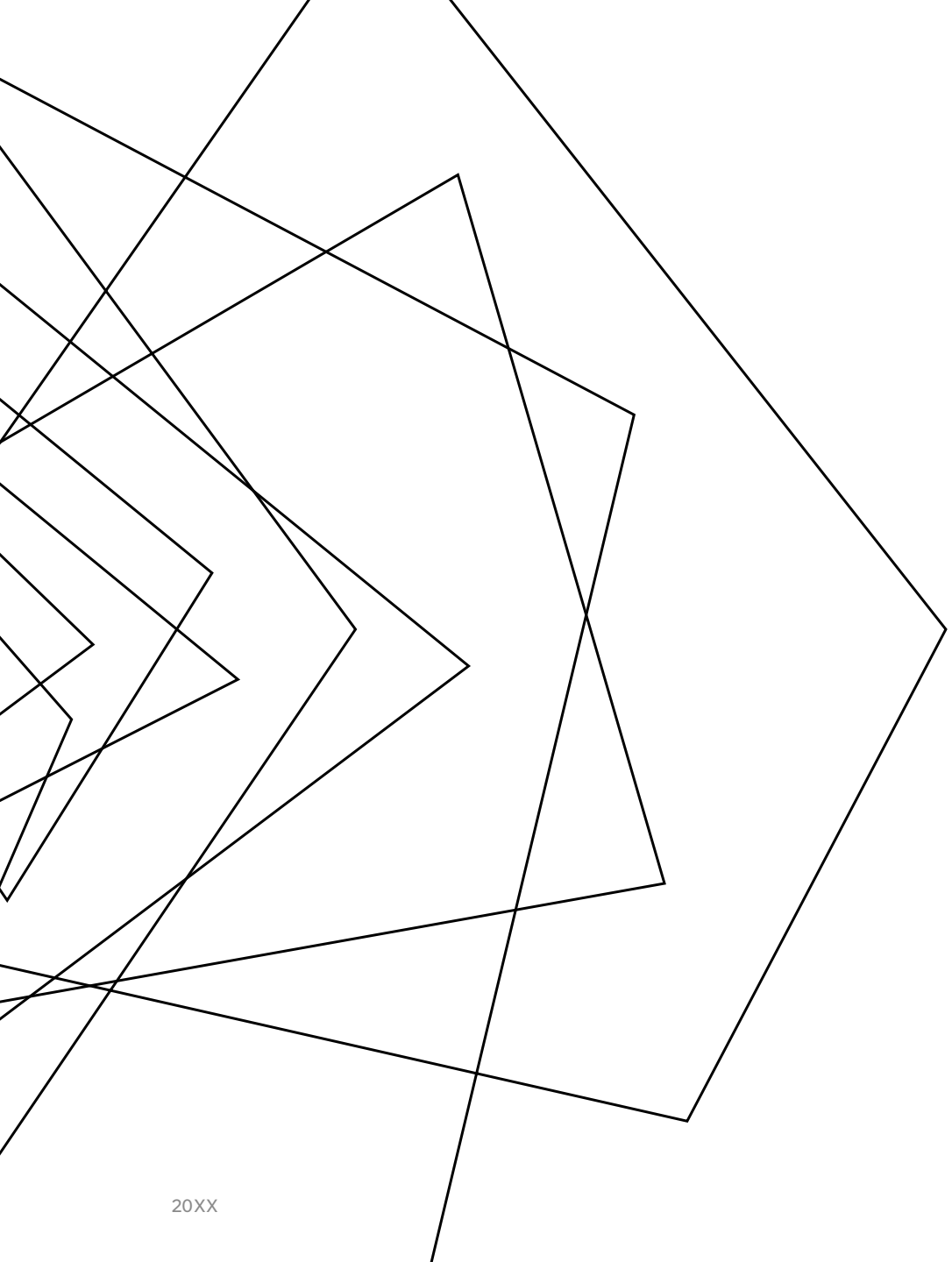
3 MINI COMPETITION

4 DEEP LEARNING / GENERATIVE AI

5 PRACTICAL DATA SCIENCE

6 SOFT SKILLS

7 THE PORTFOLIO PROJECT & DEMO DAY



1 TECHNICAL FUNDAMENTALS

DEV TOOLKITS AND ENVIRONMENTS

Git, Bash, Docker, Databases

PROGRAMMING AND DATA ANALYSIS

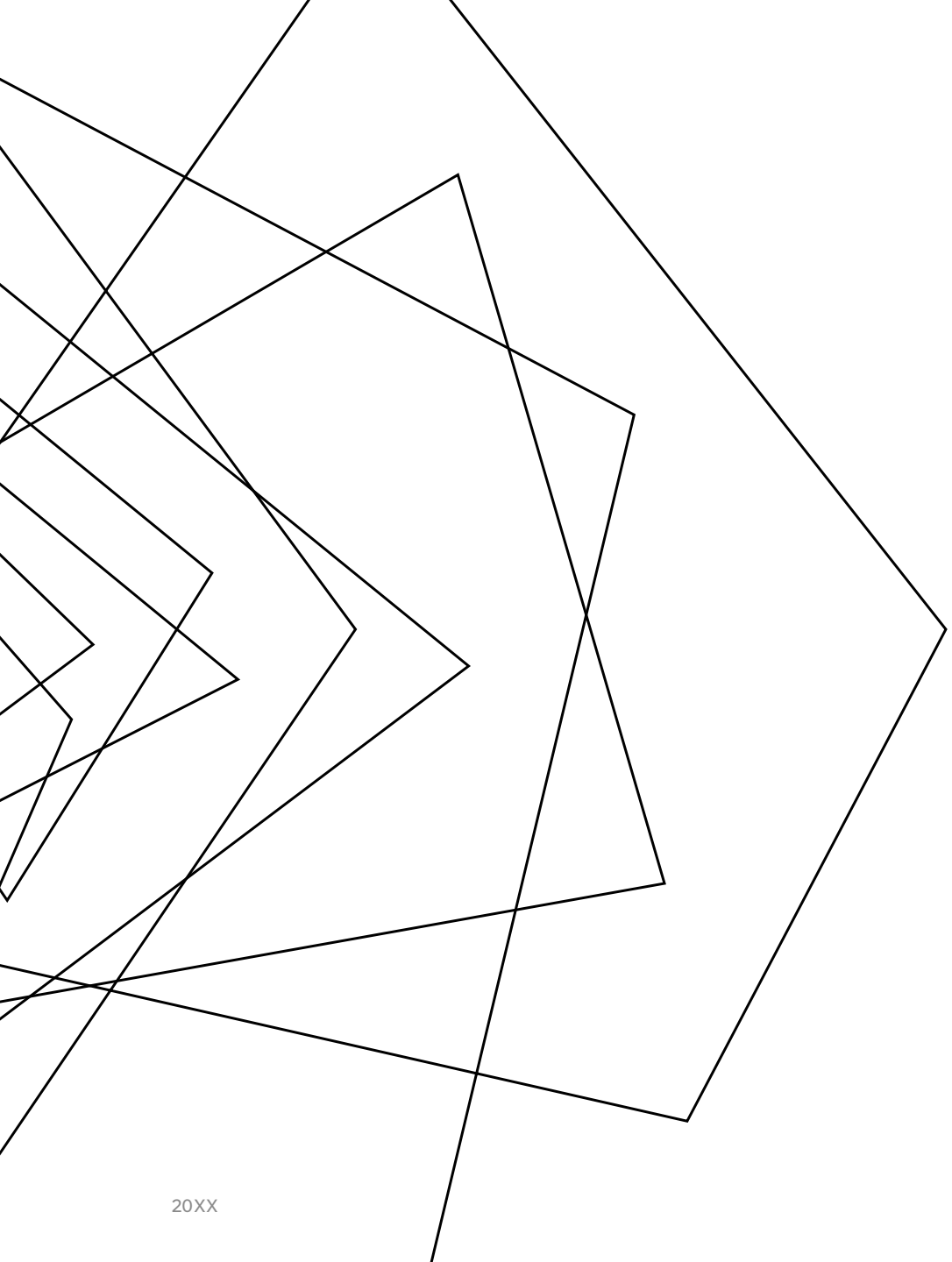
Python, NumPy, Pandas, SQL

SCARY MATH

Probability & Statistics

PRESENTATION

Data visualization (Example)



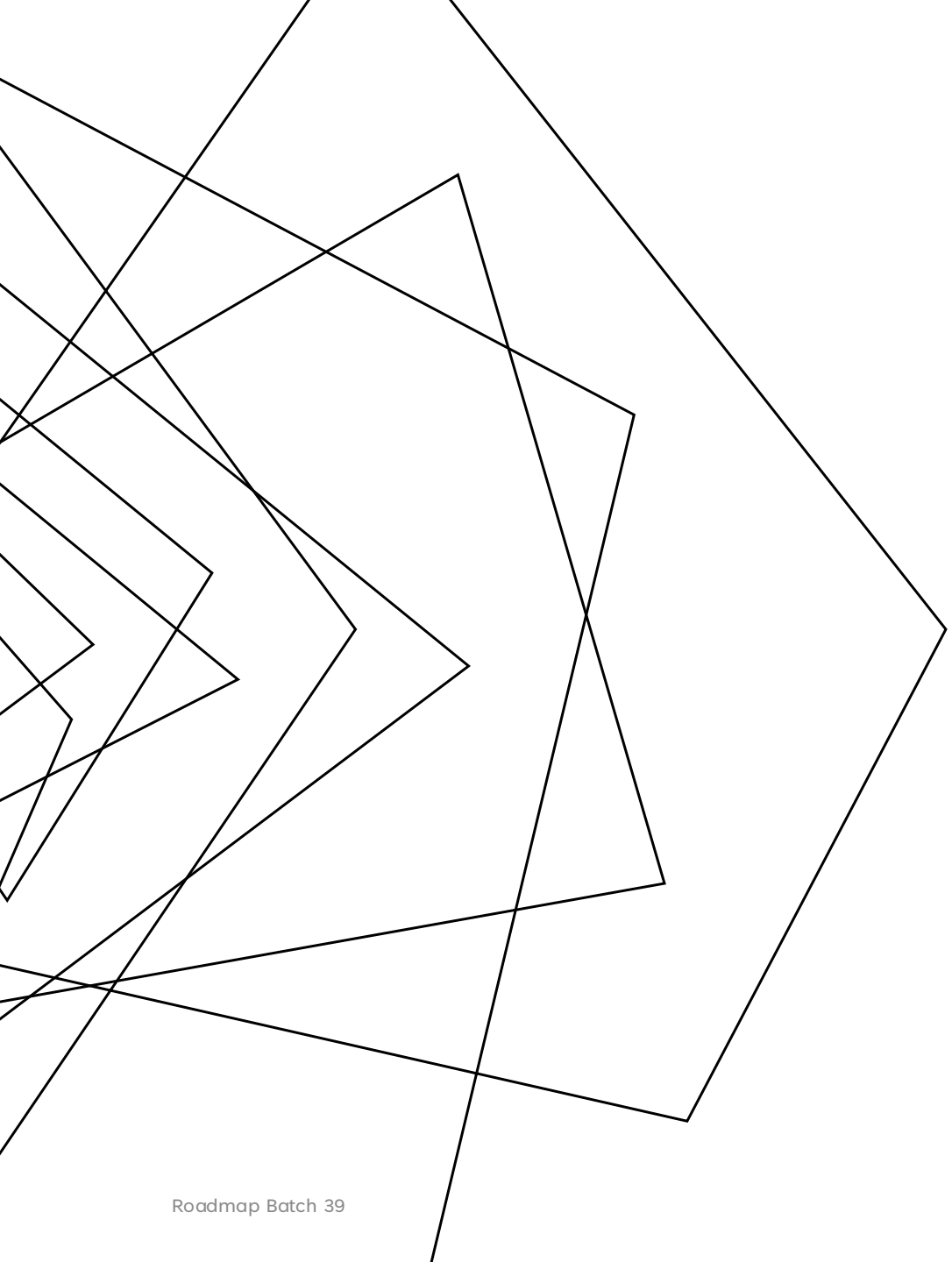
1 TECHNICAL FUNDAMENTALS

DO NOT UNDERESTIMATE THESE THINGS

- The foundation for everything else
- It will be hard to perform well without them

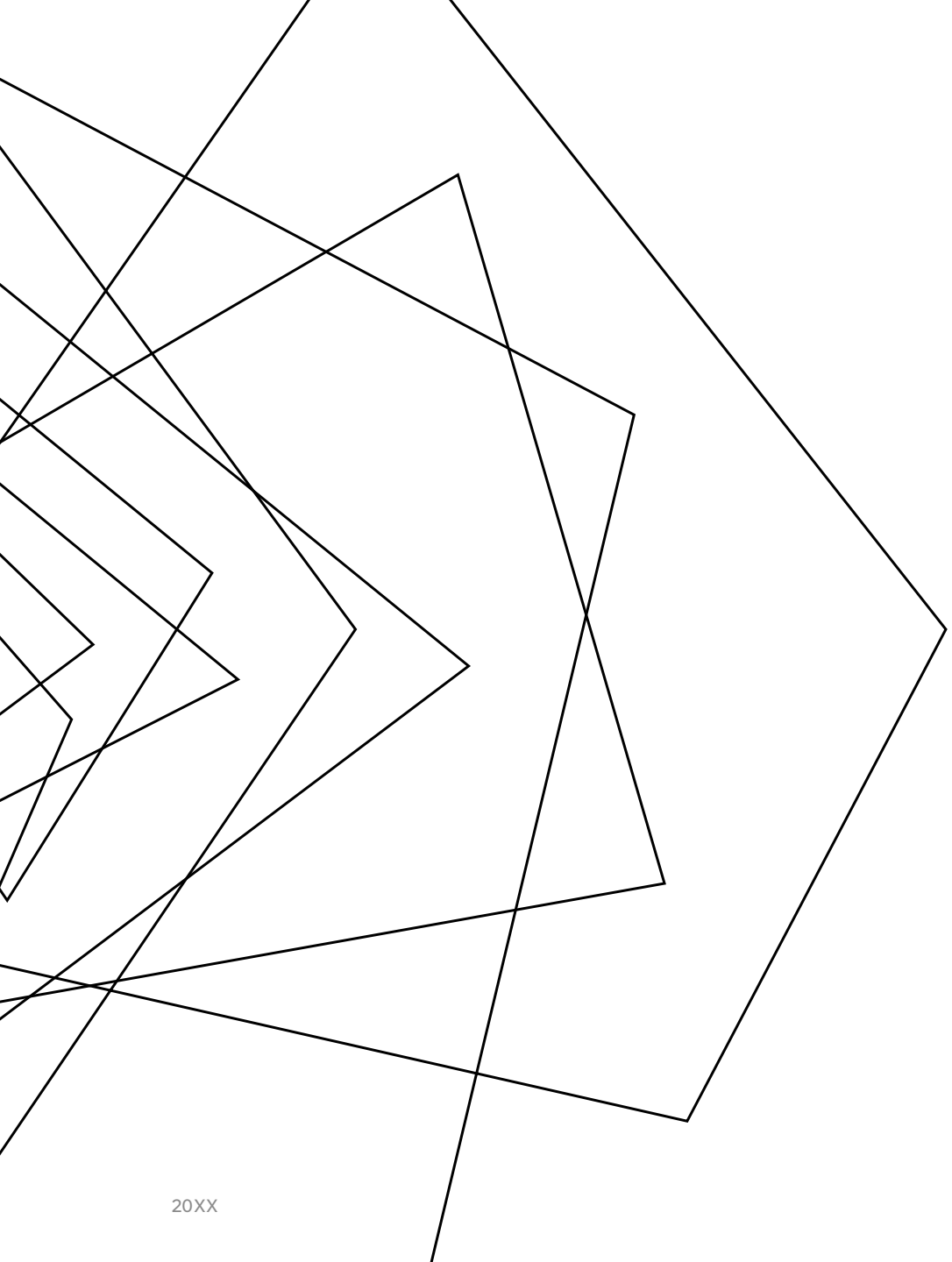
USEFUL LITERATURE

- Data Wrangling with Python (Jacqueline Kazil, Katharine Jarmul)
- Python for Data Analysis (Wes Mckinney)



THE FIRST TWO WEEKS

June / July	Mon 24	Tue 25	Wed 26	Thu 27	Fri 28	Sat 29	Sun 30	Mon 01	Tue 02	Wed 03	Thu 04	Fri 05
Topics	Intro	Numpy Pandas	Git Bash	SQL	-	Proba bility	-	Stats	Data Visualization		-	ML Fundamentals
		ONLINE				ONLINE						



2 DS AND ML FUNDAMENTALS

ML FUNDAMENTALS

How do machines learn? What are performance metrics? What is possible / not possible? Three types of learning.

OBJECT-ORIENTED PROGRAMMING

Objects & classes. Logging and error handling. Unit tests.

DS FUNDAMENTALS

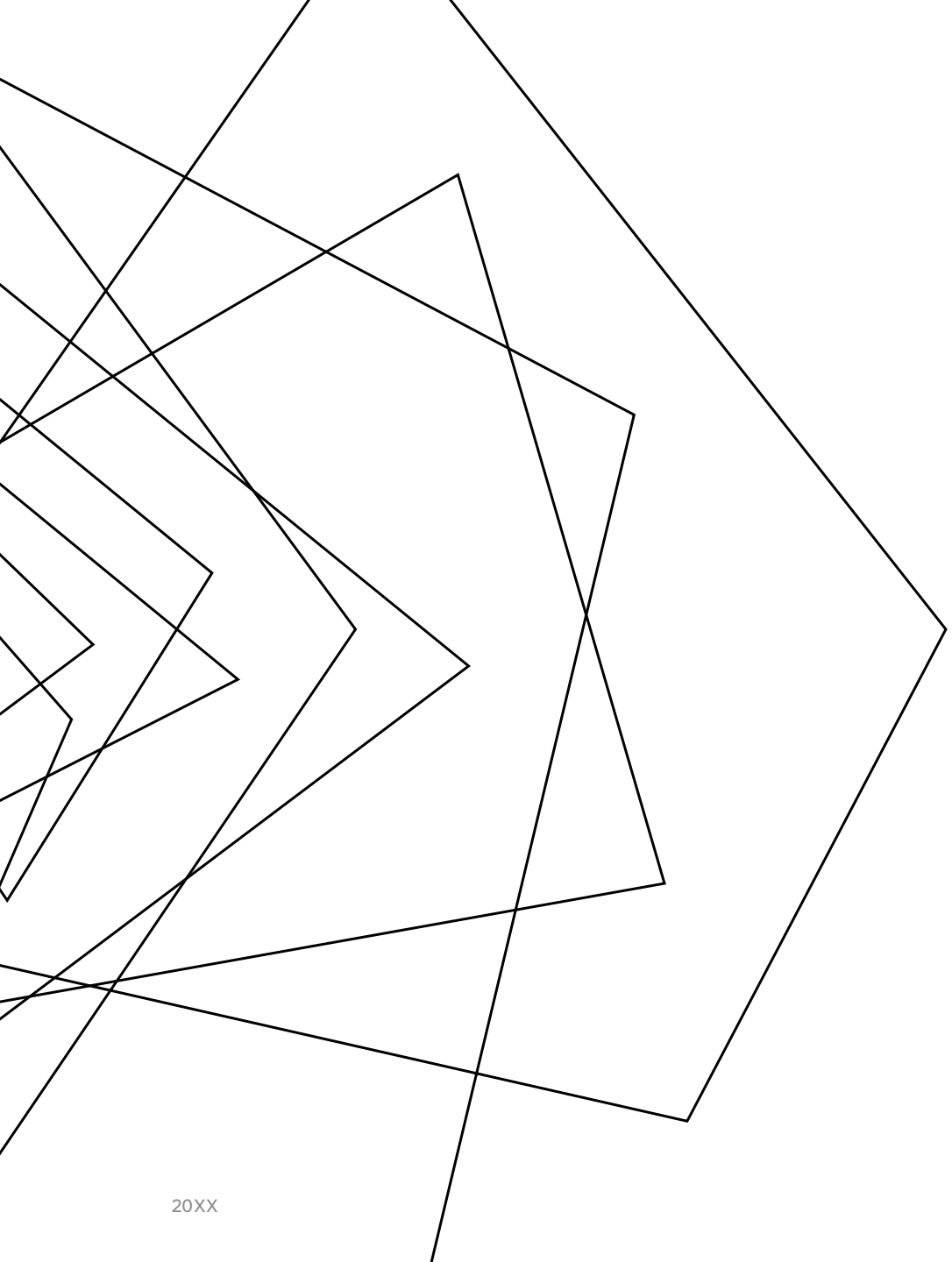
Develop a structured idea of the data science workflow. Implement a small example project. Data wrangling and model building.

TREES

Foundational tree models for classification / regression on tabular data. Bagging & Boosting.

TIME SERIES

Model changes over time. Forecasting. Decomposition, seasonality, trends and detrending. ARIMA, Holt-Winters method.



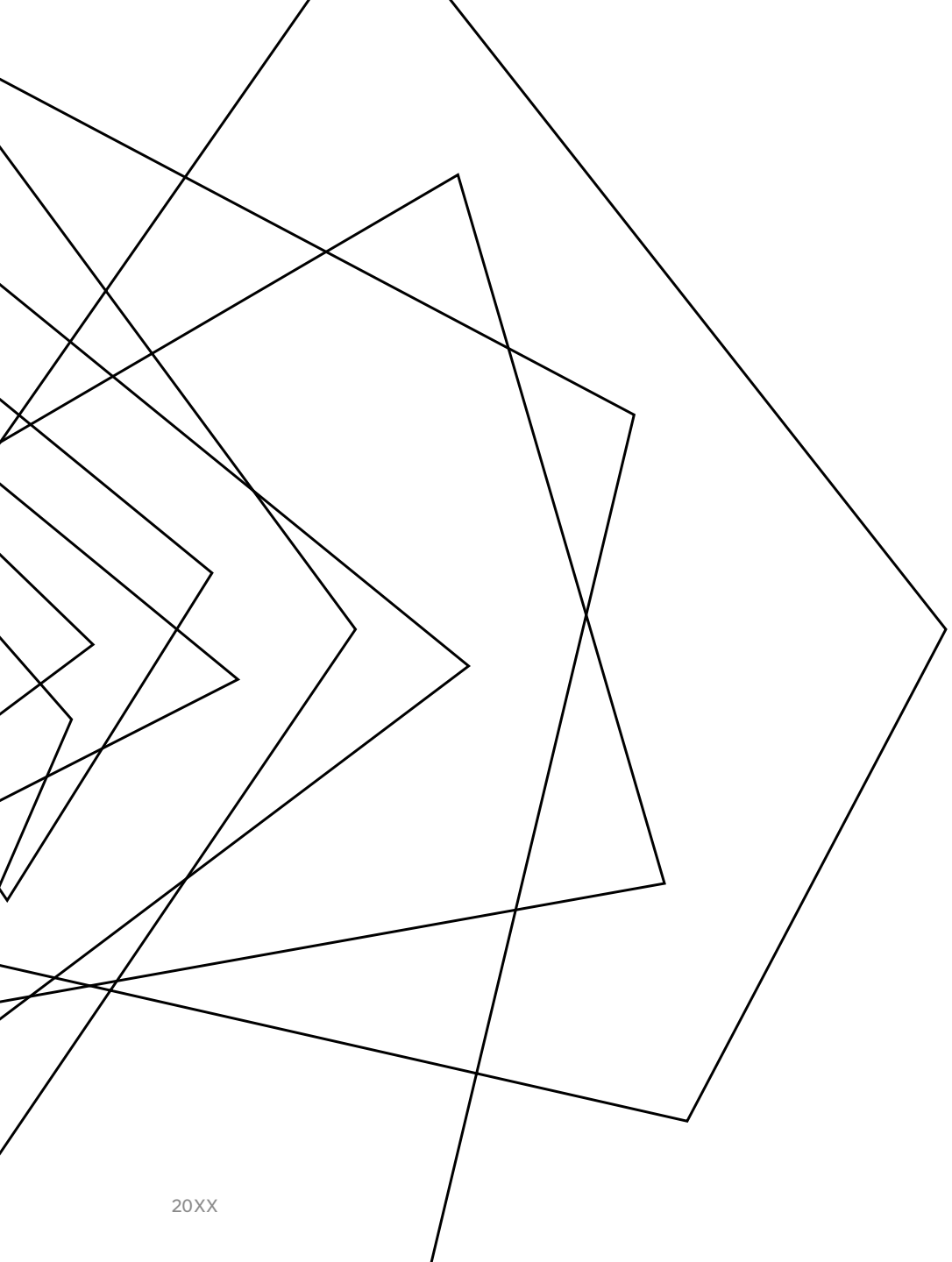
2 DS & ML FUNDAMENTALS

YOUR DATA SCIENCE TOOLKIT

- The key points required to pass an interview
- Day-to-day operations in data science
- Go through the notebooks if possible
- Practice these skills on Kaggle

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)
- Towardsdatascience.com articles



3 MINI COMPETITION

3 DAYS OF REAL DATA SCIENCE WORK

Put your skills to the test and aim for the high score.

WORK IN TEAMS

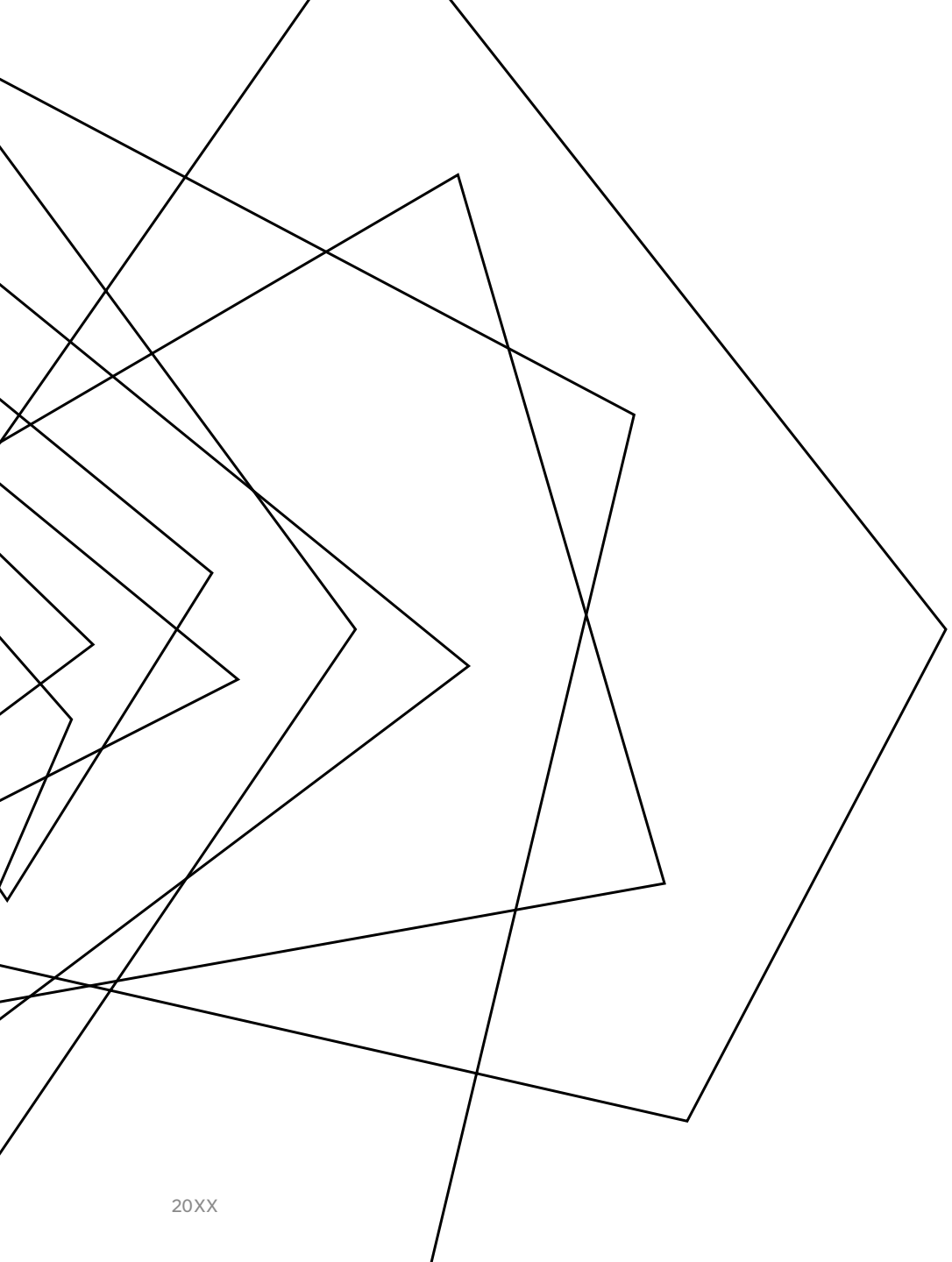
Collaboration is key. Remember your GitHub knowledge.

BUILD A CLEAN REPOSITORY

A high score is nice, but what matters is nice, clean code.

HAVE FUN!

Order in some pizza and stock the shared fridge.



4 DEEP LEARNING

FUNDAMENTAL DEEP LEARNING

BACKPROPAGATION! Neural Network architecture. Complexity theory.

POPULAR FRAMEWORKS

Tensorflow & PyTorch, mainly

TECHNIQUES AND FRAMEWORKS

NLP, transfer learning, representations, image processing, computer vision, geometric deep learning, reinforcement learning. And, of course, the mythical transformers.

ADVANCED TECHNIQUES

Debugging Deep Learning Models. Finetuning LLMs. RAG (Retrieval augmented generation).

4 DEEP LEARNING

WORD EMBEDDINGS

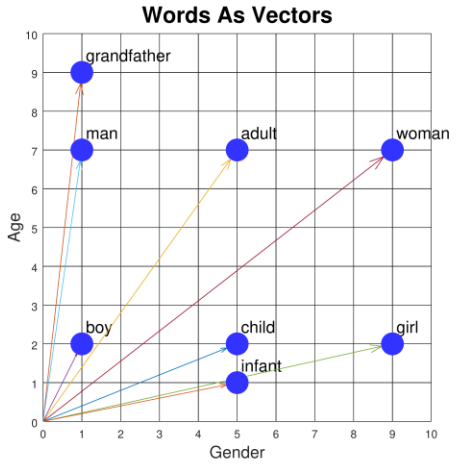
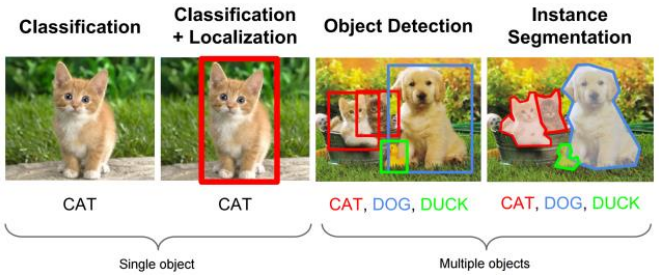


IMAGE CLASSIFICATION



SEMANTIC SEGMENTATION

4 DEEP LEARNING

HERE'S WHERE IT GETS REALLY FUN

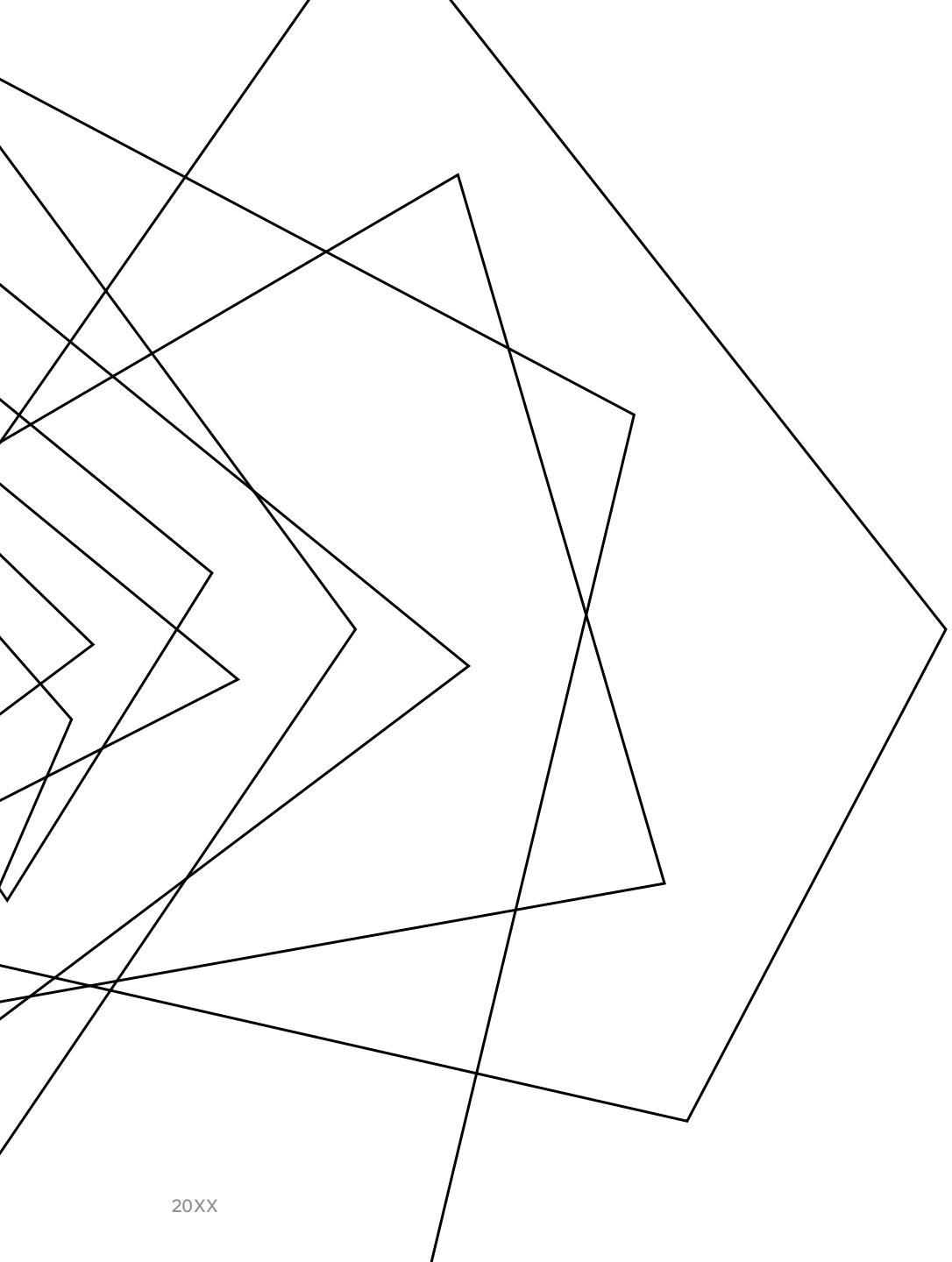
Building a GPT from scratch or designing your own RAG solution is very rewarding and builds deep understanding.

USEFUL VIDEOS

- Neural Network series by 3Brown1Blue

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



5 PRACTICAL DATA SCIENCE

ML OPS

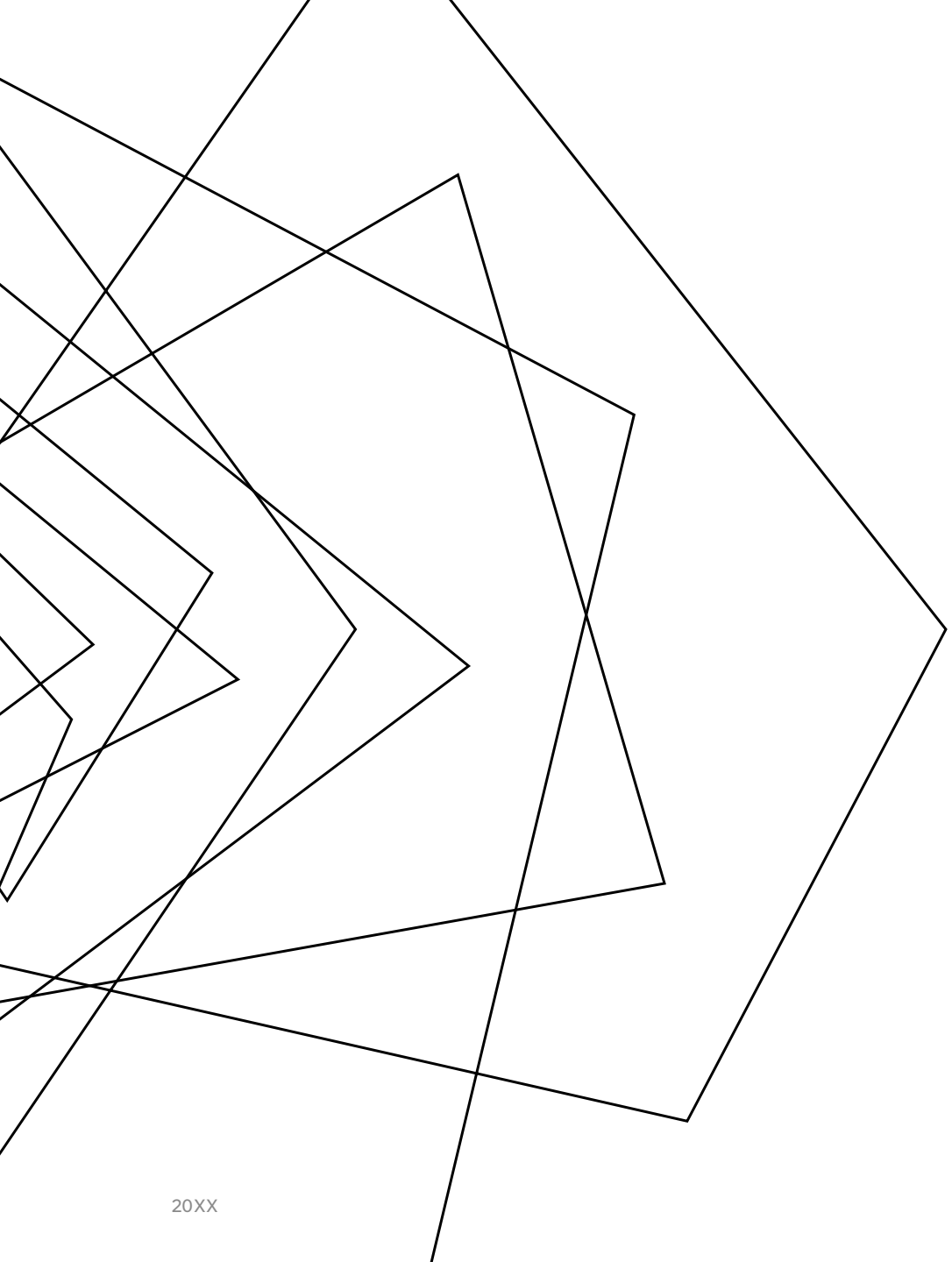
From training to deployment to API – bring your machine learning to life!

TEST-DRIVEN DEVELOPMENTS

Avoid catastrophic mistakes by writing good tests!

PRACTICAL DS WITH STREAMLIT

Build a useable application in minutes in Python.



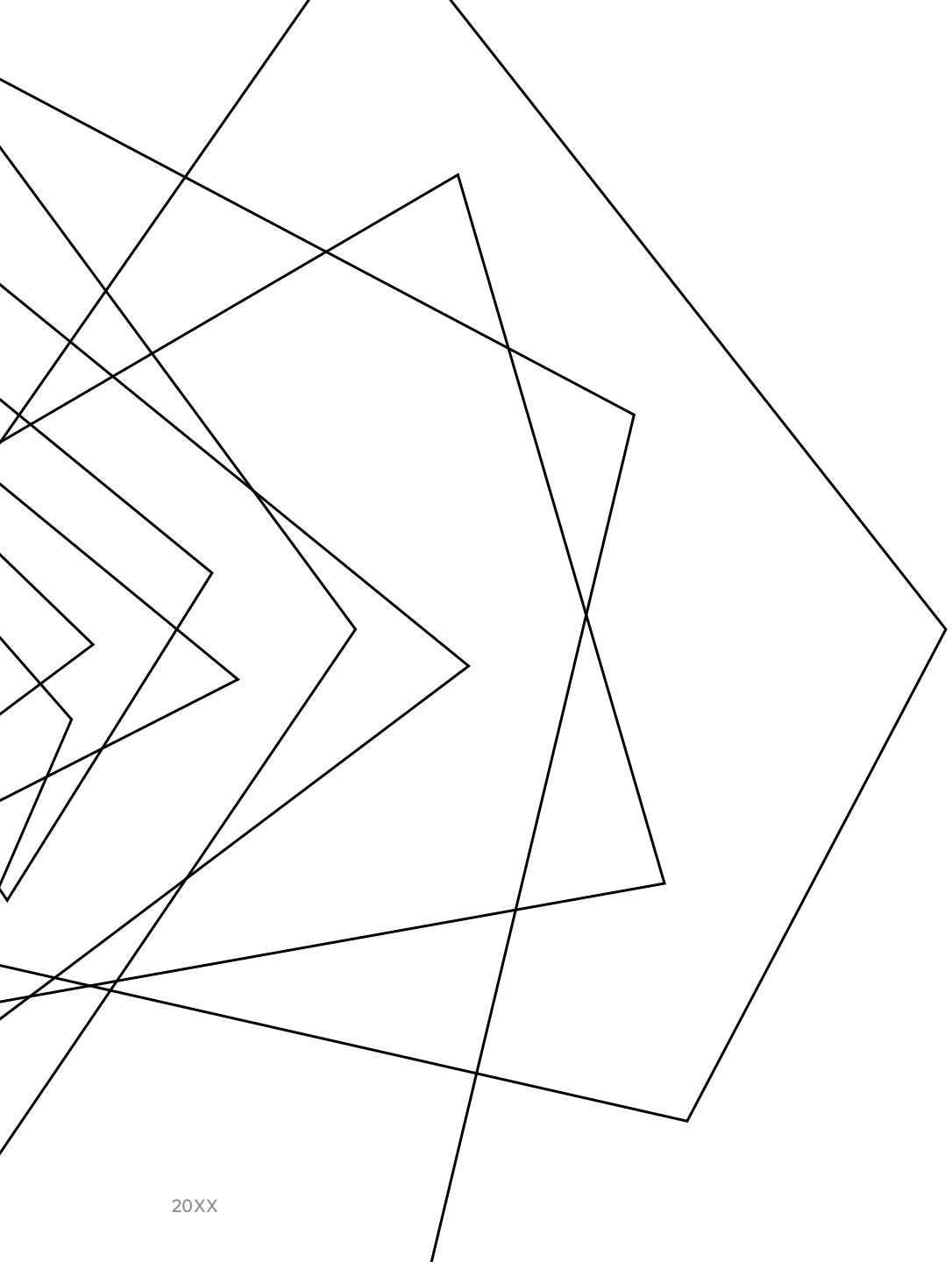
5 PRACTICAL DATA SCIENCE

CARRY YOUR MODELS INTO THE WORLD

Nice code in a Jupyter notebook is worth nothing if it’s not deployed and usable. Here you learn how to get things out there!

DO NOT MISS THESE CLASSES

- Highly applicable in real life
- Not as flashy as building a GPT, but very real and useful
- Immensely useful for portfolio project



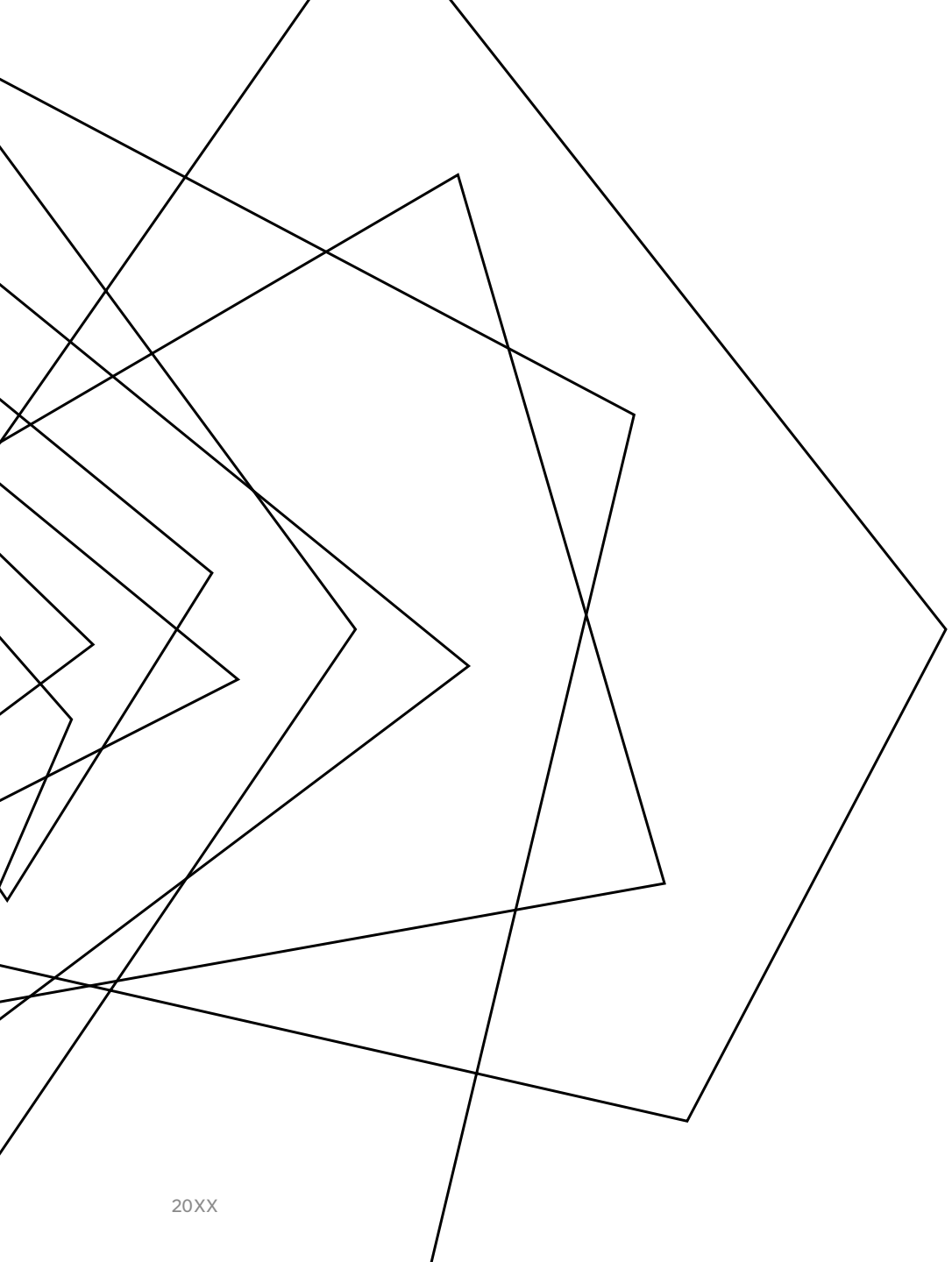
6 SOFT SKILLS

BUSINESS COMMUNICATION

Talk to stakeholders. Explain difficult concepts. Manage expectations. Nail that interview.

CAREER SUPPORT

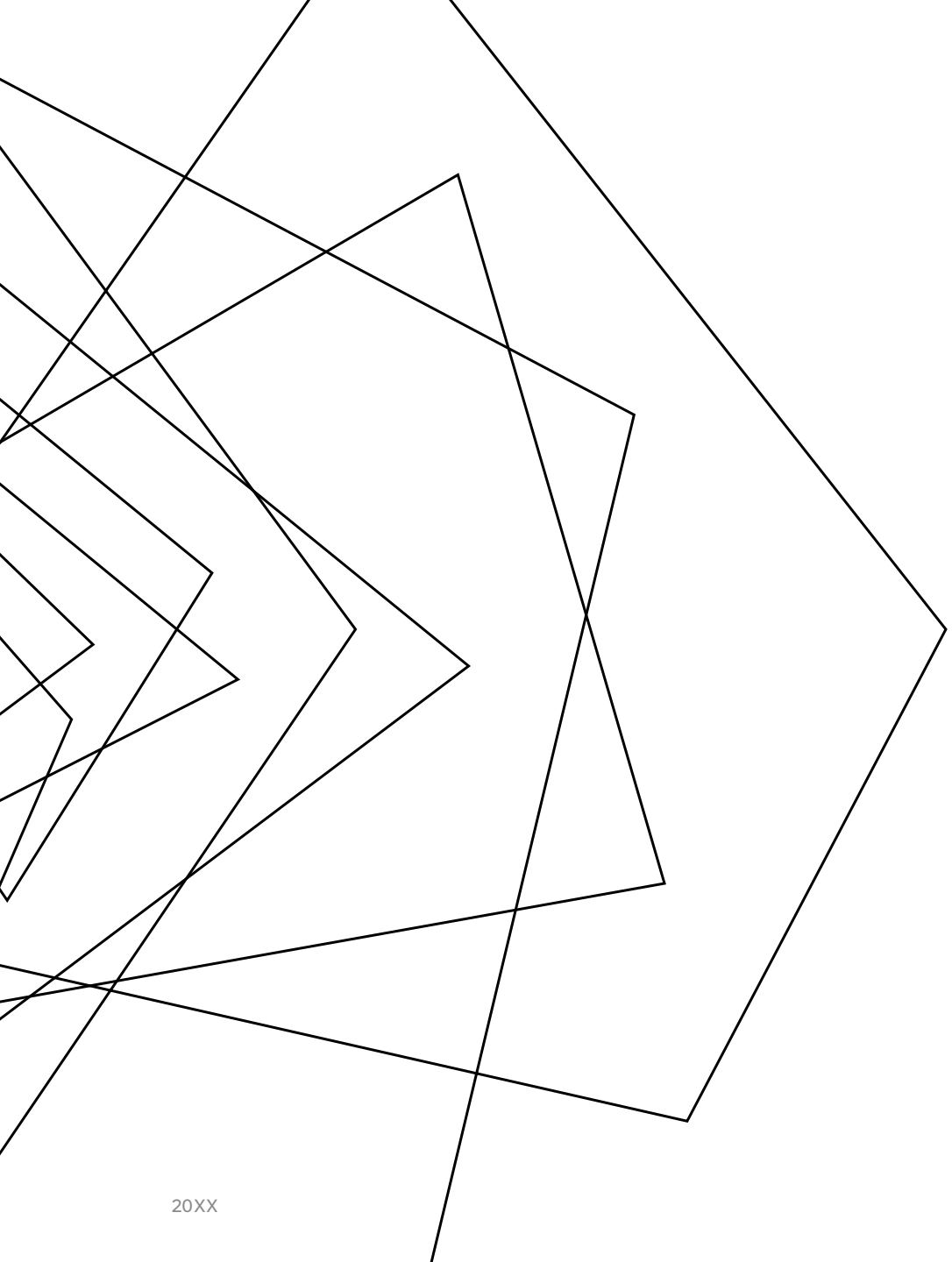
How to land a job in data science.



6 SOFT SKILLS

COME PREPARED

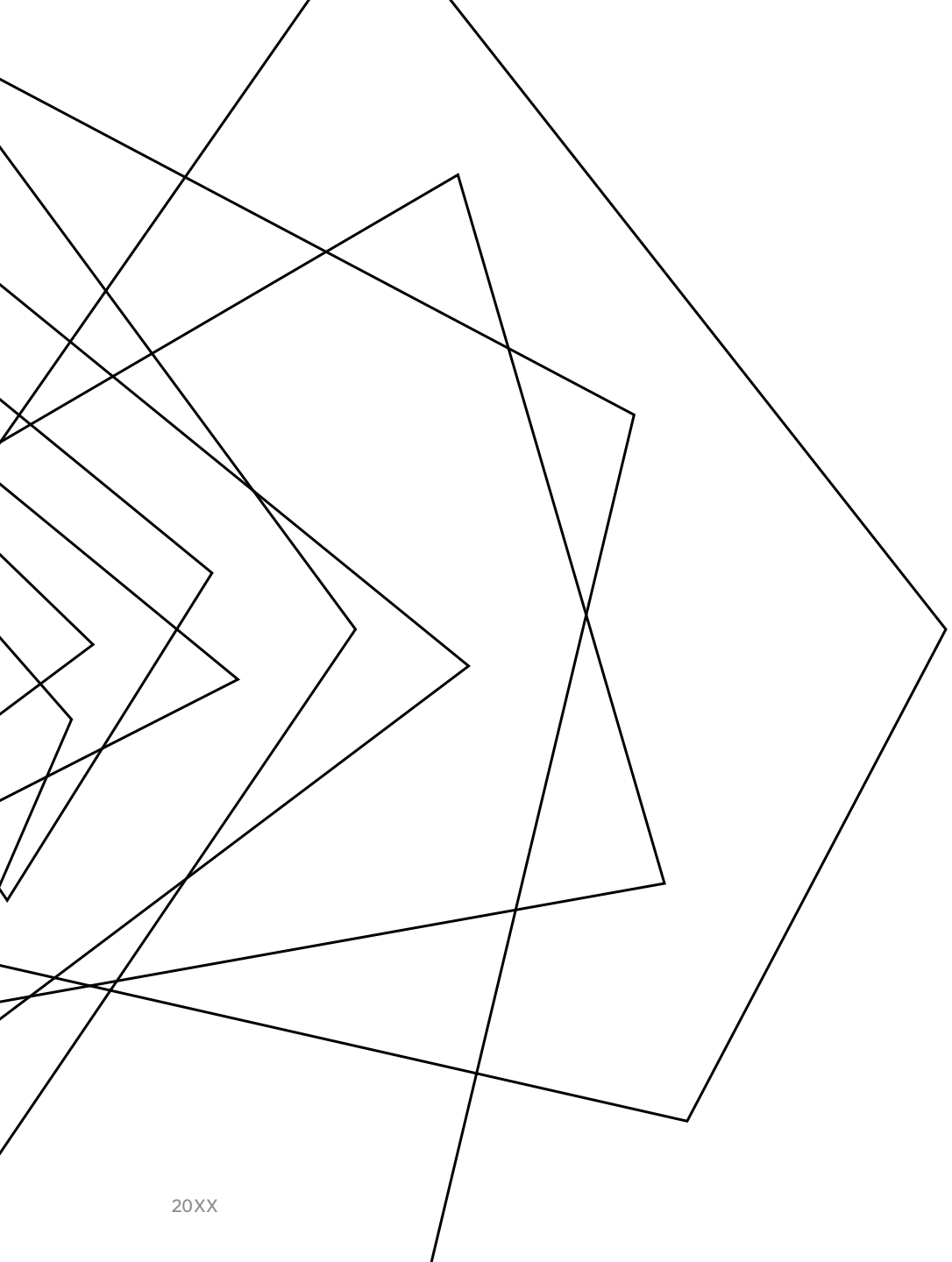
- For the communication class, pick a topic / situation that you found challenging in the past.
- For the career support class, prepare your CV



7 THE PORTFOLIO PROJECT

THIS IS WHERE IT GETS REAL!

- Find a problem
- Turn it into a data problem
- Find, create, retrieve and prepare data
- Deliver a result and present it



7 THE PORTFOLIO PROJECT

SHOWCASE WHAT YOU LEARNED

IDEALLY A TOPIC YOU FIND INTERESTING

WORK ALONE OR IN TEAMS

TIME MANAGEMENT

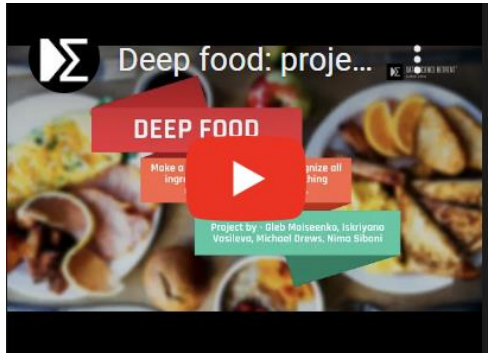
SOME COOL EXAMPLES

The sound of failure

Batch 25

Medical Advice Generator

Batch 34



Deep Food

Batch 22

Enhance doctors (not replace them)

Batch 34

STARTING POINTS

PAPERS WITH CODE

- Paperswithcode – it will provide you with trending machine learning research and the code to it

TECH & BUSINESS MAGAZINES

- MIT Technology Review
- The Economist
- Harvard Business Review

DISCUSSIONS

- With humans
- With ChatGPT or your local flavor

THERE IS NO SILVER BULLET

FINDING YOUR TOPIC IS HARD

BE AMBITIOUS BUT REALISTIC

DON’T ONLY LOOK BACK – ALSO LOOK AHEAD

START THINKING NOW

THE PERFECT PROJECT DOESN'T EX--

It makes you want to code late into the night

EXCITING

APPROVED

It got Jose's blessing

???

CHALLENGING

It lets you apply your skills

AMBITIOUS

It sets the bar high and impresses the audience

EVEN MORE ADVICE

COMMUNICATE EARLY

- Use the 1:1 sessions from the start!
- Provide a 1-paragraph description to Arun when you have a topic

MENTORS WILL SUPPORT YOU

- Teachers can mentor a project
- We try to match mentors to the topic
- Typical mentoring days are Mondays & Thursdays

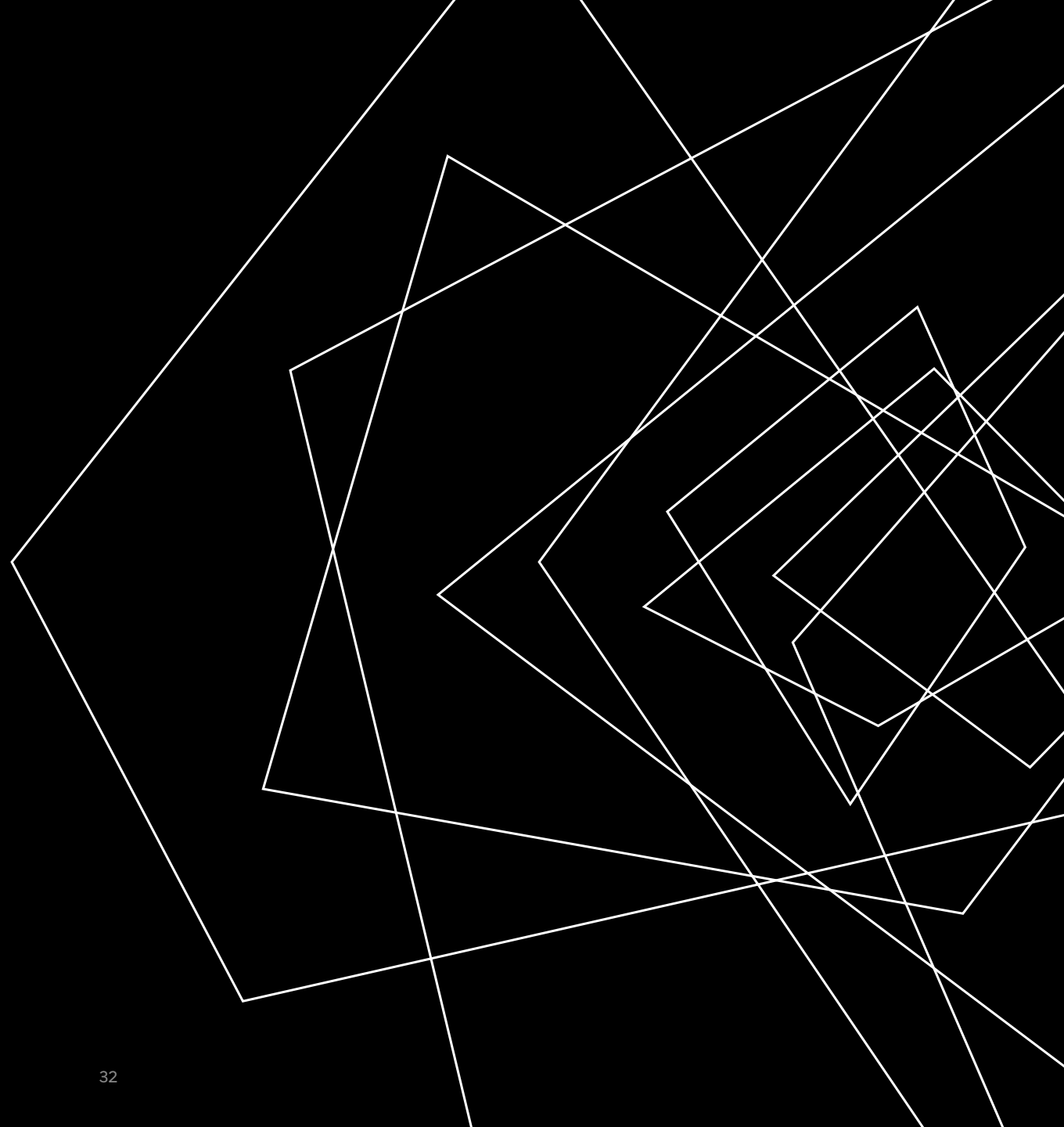
WHERE?

- You can be flexible – at home or at DSR

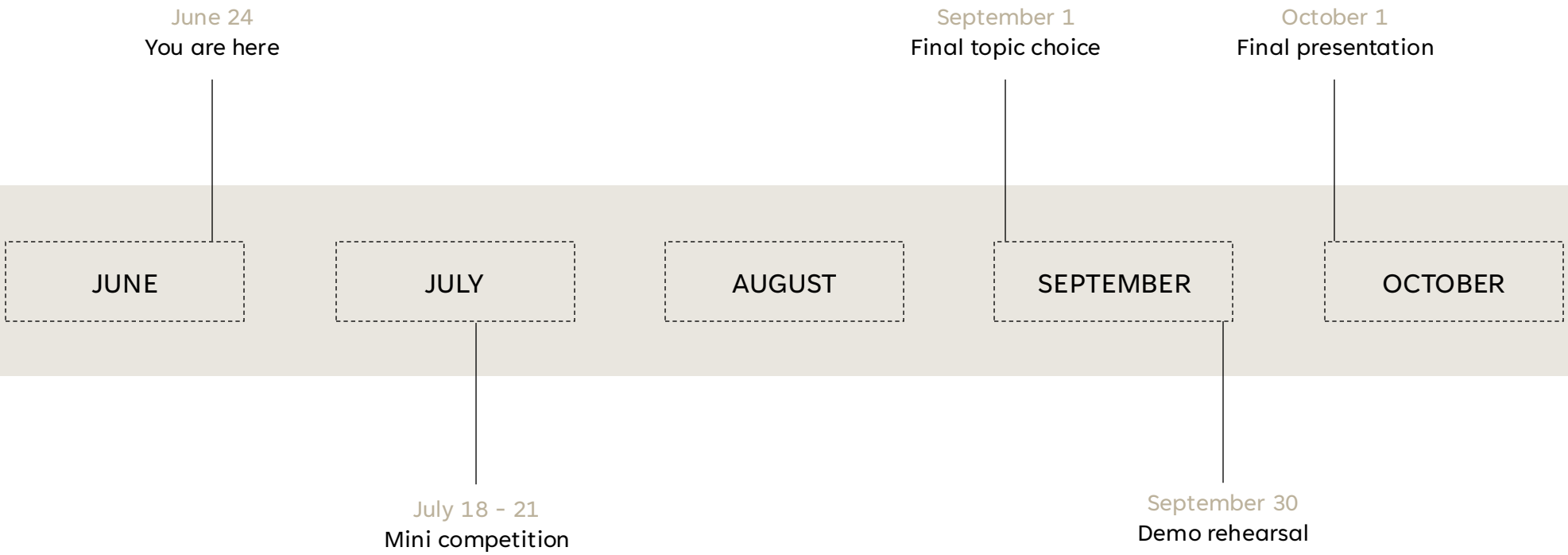
QUESTION: CAN I USE CHATGPT?

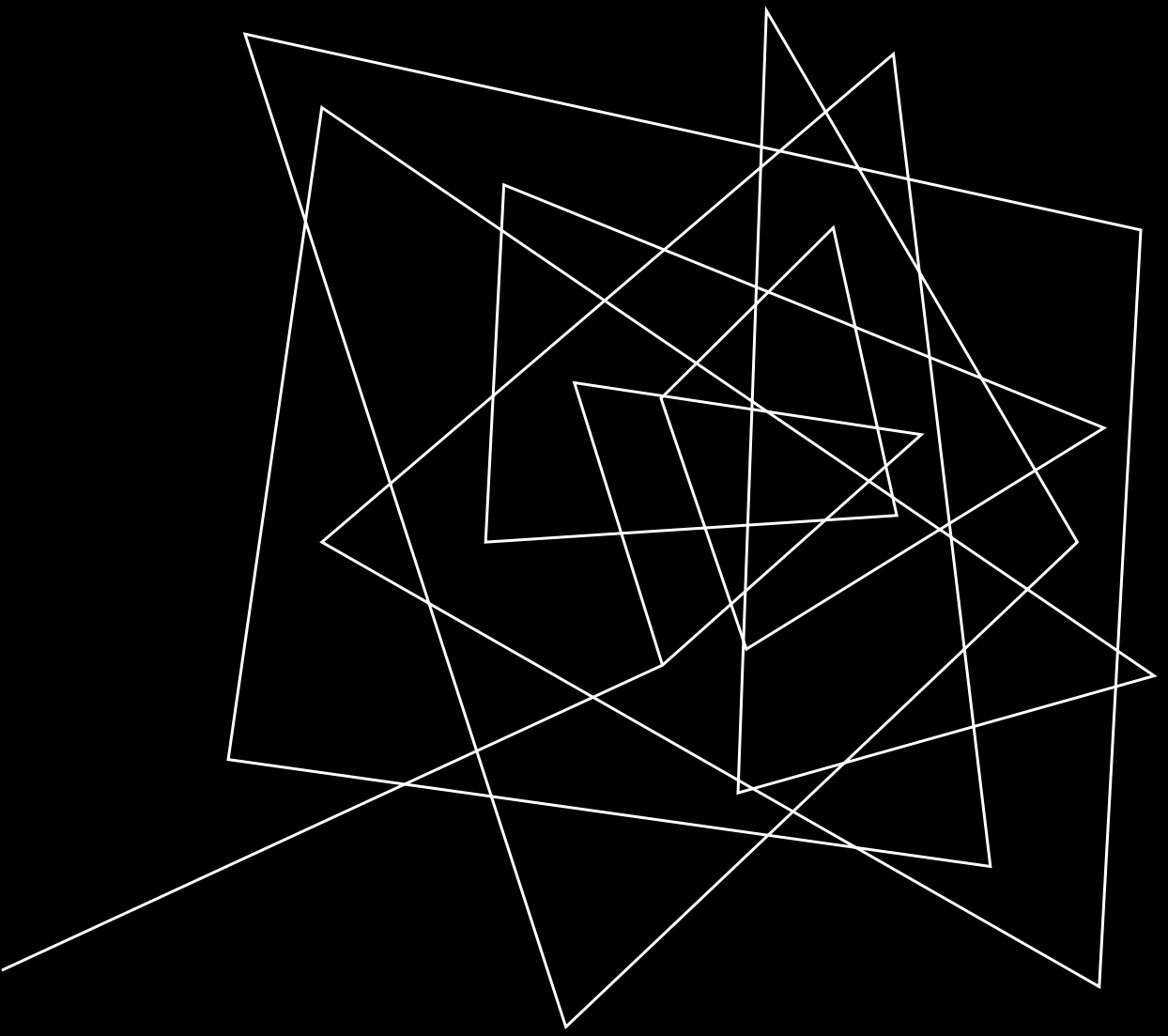
- Yes
- Resist the temptation (sometimes)
- Use it to learn, don't use it to skip

- Look into GitHub Copilot



MILESTONES


June 24
You are here
September 1
Final topic choice
October 1
Final presentation
JUNE
JULY
AUGUST
SEPTEMBER
OCTOBER
July 18 - 21
Mini competition
September 30
Demo rehearsal



SOFTWARE SETUP

THE SHELL

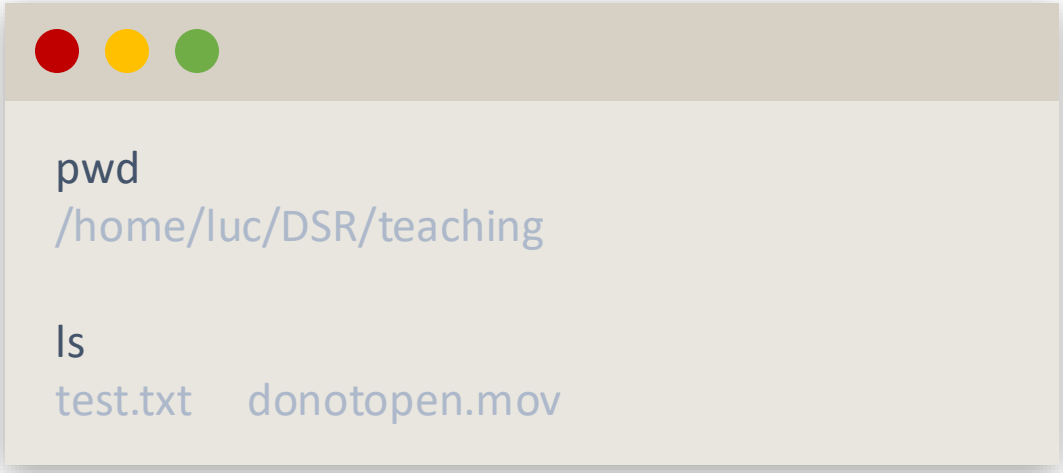
USER <> OPERATING SYSTEM INTERFACE

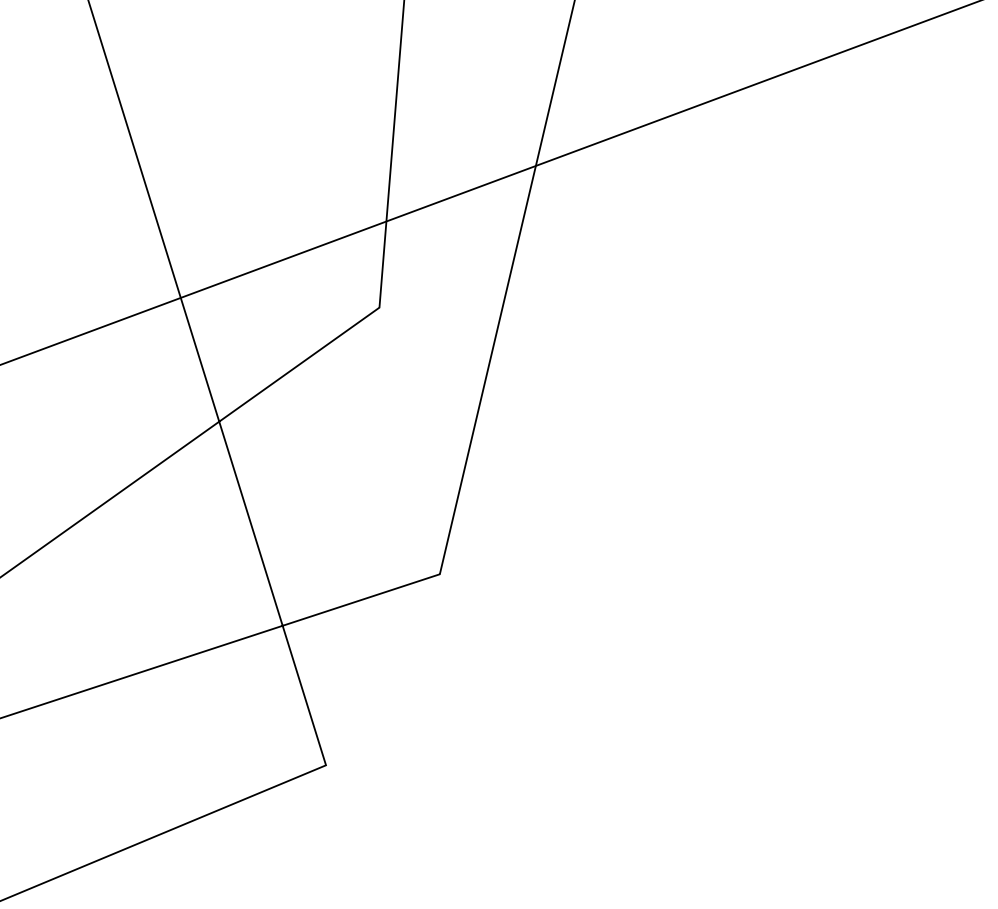
Navigate your file system. Setup virtual environments. Install packages. Talk to Git.

ACCESS IT VIA TERMINAL EMULATOR

For example: Terminal on Linux

USES BASH SCRIPTING LANGUAGE





ANACONDA

A Python distribution platform

Includes PyCharm, Jupyter Notebook, Jupyter Lab

Allows you to create virtual environments

GIT

Book of Git

Read everything about Git here.

A version control protocol

Records changes to a file or set of files over time so that you can recall any previous version at any point in time.

Branches for parallel development

People create **branches** to work on features in parallel and **merge** branches later.

GitHub is a popular platform for using Git

Highly recommended for your own projects

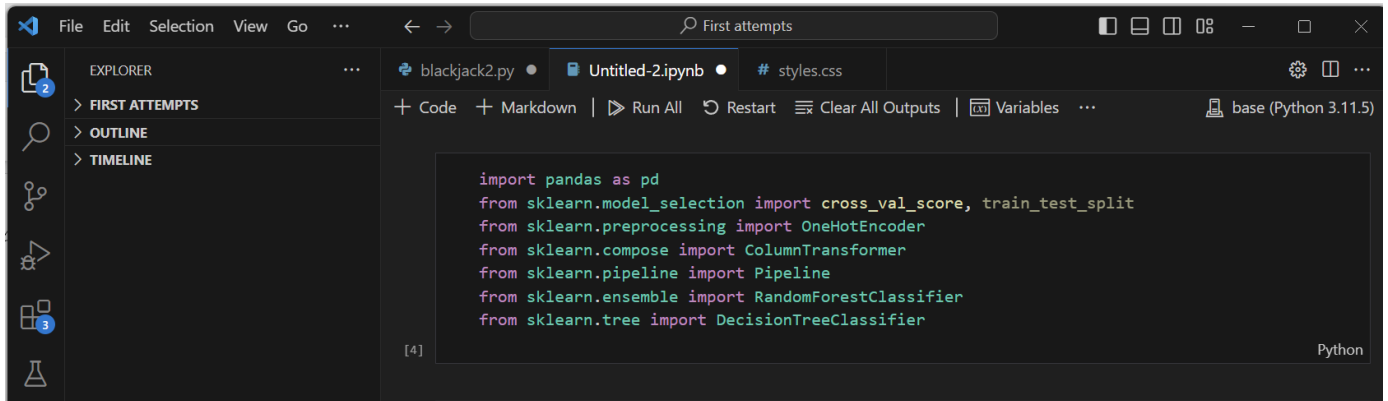
DEVELOPMENT ENVIRONMENTS

Integrated Development Environments (IDE)

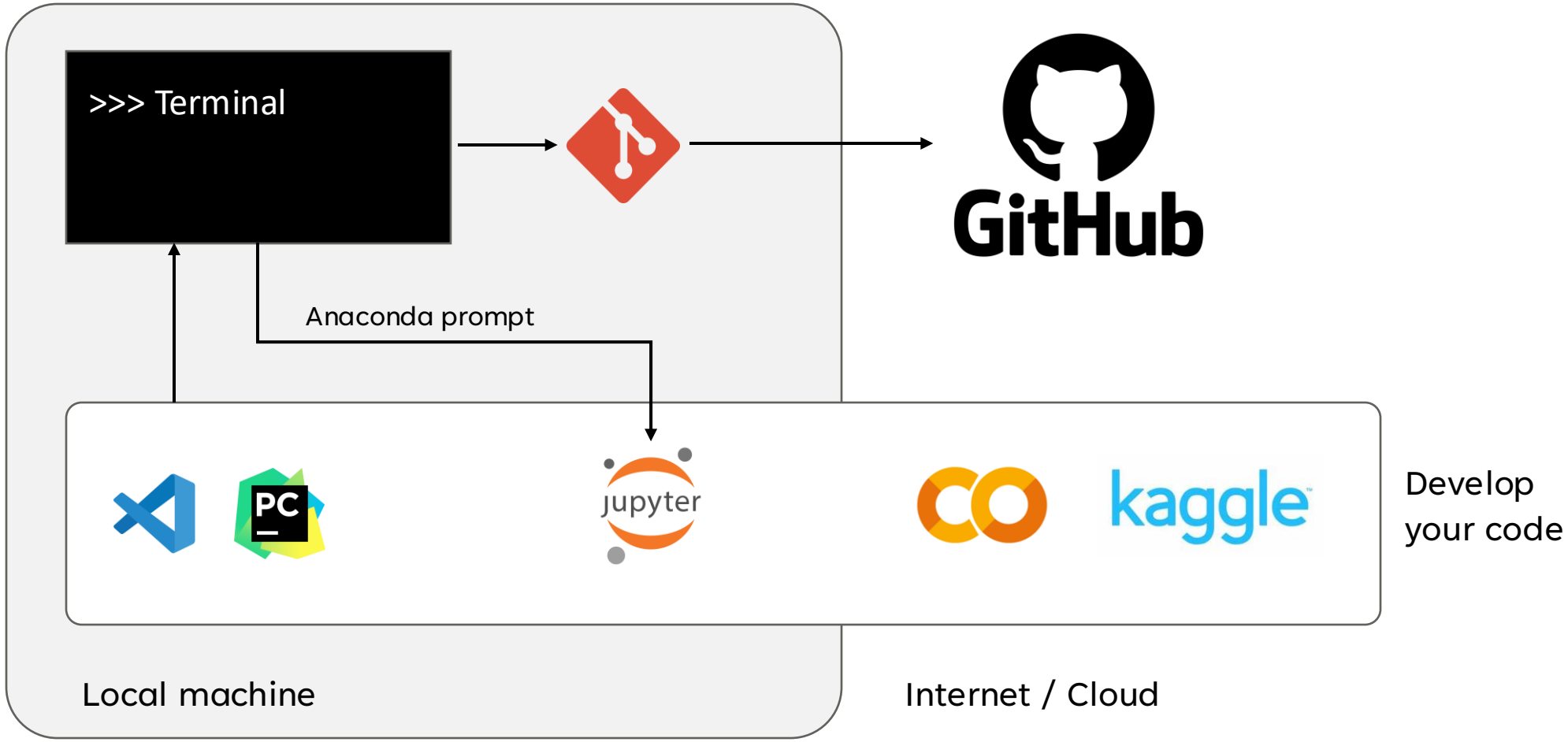
- Visual Studio Code
- PyCharm
- Spyder

Notebook Environments

- Jupyter Lab / Jupyter Notebooks
- Kaggle
- Google Colab

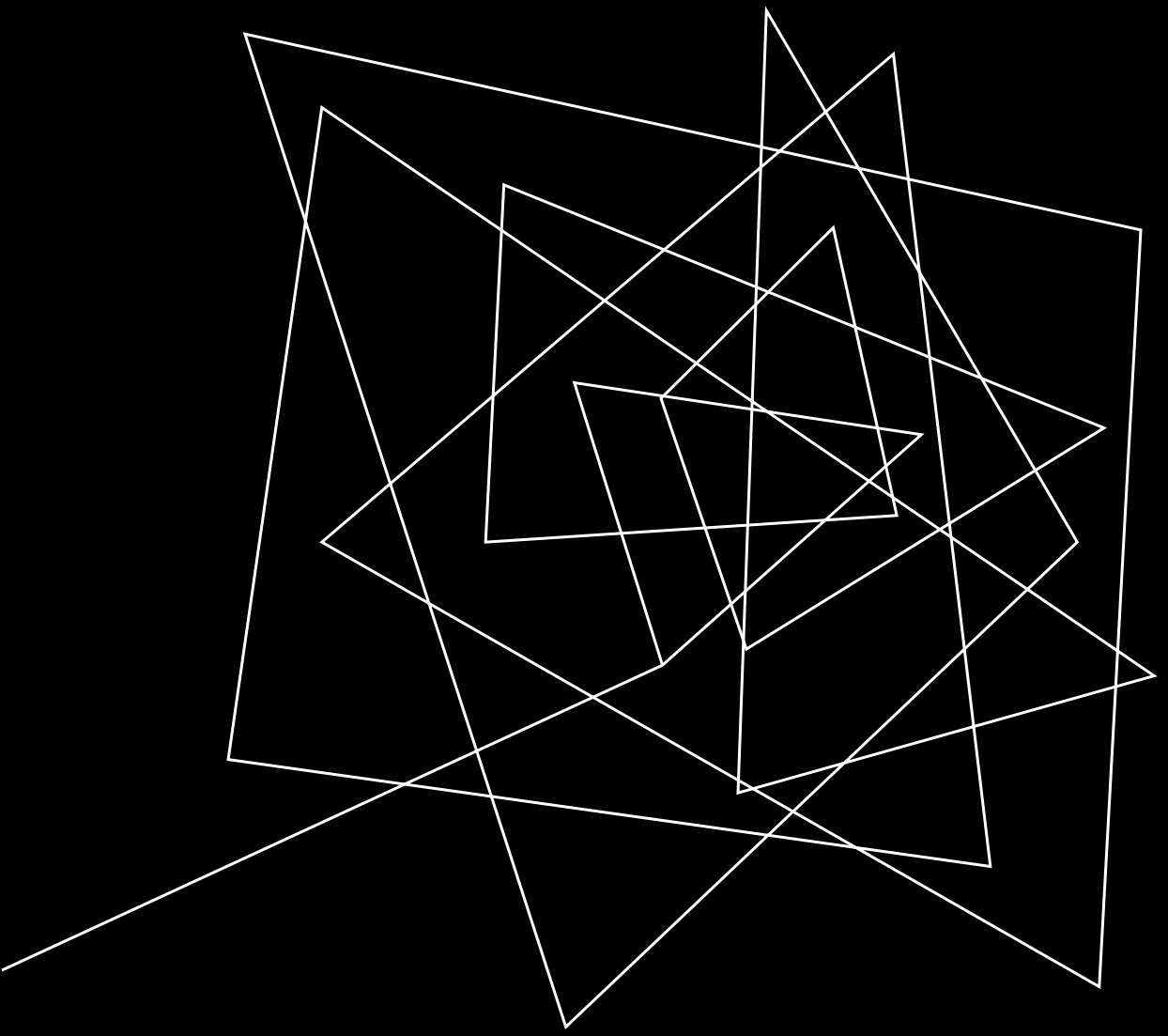


A (SIMPLIFIED) EXAMPLE SETUP



USEFUL

- 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



LET’S SETUP AN EXAMPLE PROJECT TOGETHER

MAKE SURE TO

Have your software installed

Terminal, Anaconda, VS Code or PyCharm

Have your accounts set up

Github

Follow these instructions

Lucamiras/dsr-teaching-setup: Introductory material for new students at Data Science Retreat Berlin. (github.com)

Stop me and ask

We're all learning here

GIT CLONE

HTTPS://GITHUB.COM/LUCAMIRAS/DSR-TEACHING-SETUP.GIT

MY APPROACH

ONE .VENV PER LECTURE

Ideally use the requirements.txt file provided by the teachers

Avoid package incompatibilities

LOCAL FOLDER / REPO PER LECTURE

Keep things organized

Make a separate folder for experiments

TAKE PLENTY OF NOTES

Nothing against physical notebooks – Thinking with your hands

Use your favorite notetaking app

MEETUPS

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

PyLadies Berlin

Women Techmakers Berlin

RESOURCES TO WATCH

StatQuest with Josh Starmer – YouTube

ritvikmath – YouTube

3Blue1Brown – YouTube

Kaggle – YouTube

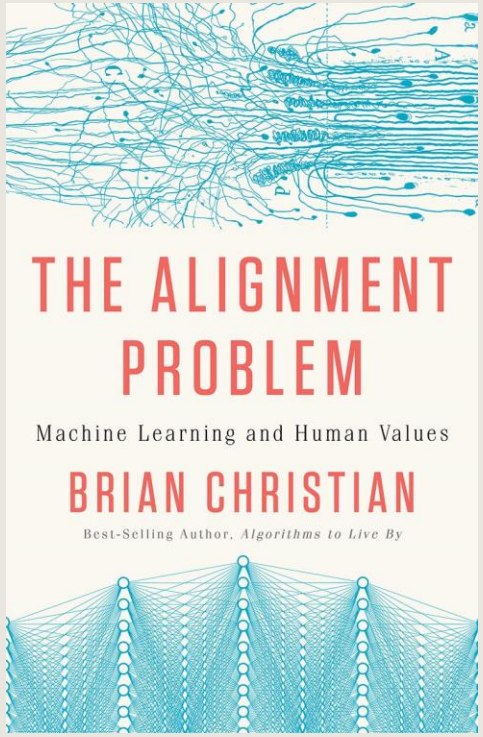
DataCamp -YouTube

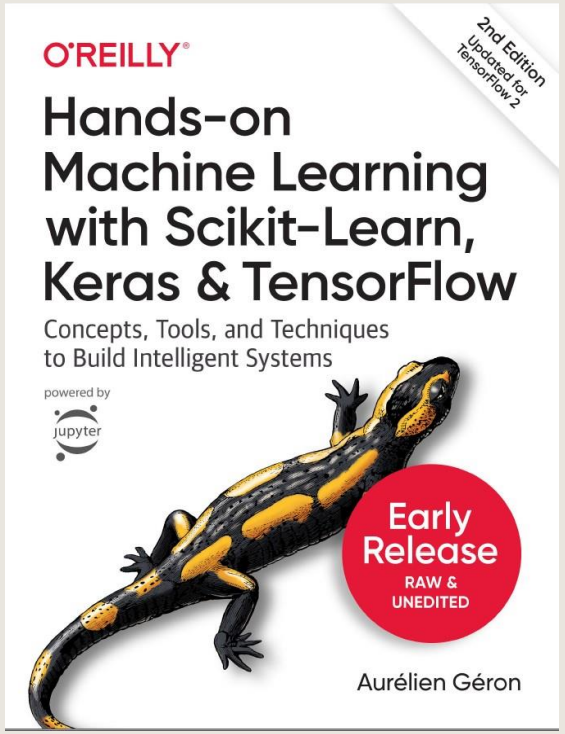
Two minute papers - YouTube

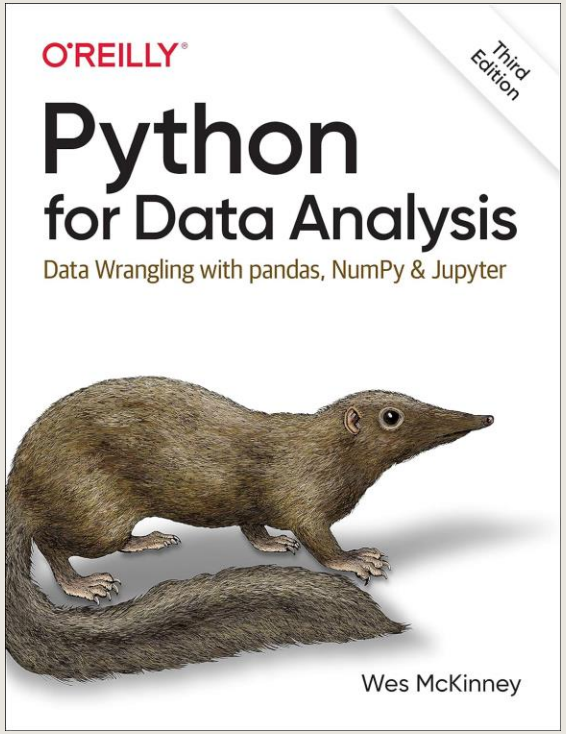
Machine Learning Mastery – YouTube

... and many more

PERSONAL BOOK RECOMMENDATIONS







UNSOLICITED ADVICE FROM A GRIZZLED ALUMNUS

- Learning is a privilege
- But time flies, and this will be over quick
- Pause for a moment every once in a while

OBLIGATORY EXPECTATION MANAGEMENT

The next three-and-a-half months will be challenging, demanding, exciting, eye-opening and, hopefully, fun.

Time is short, so work hard, but give yourself time to rest. Nobody gains anything if you burn out.

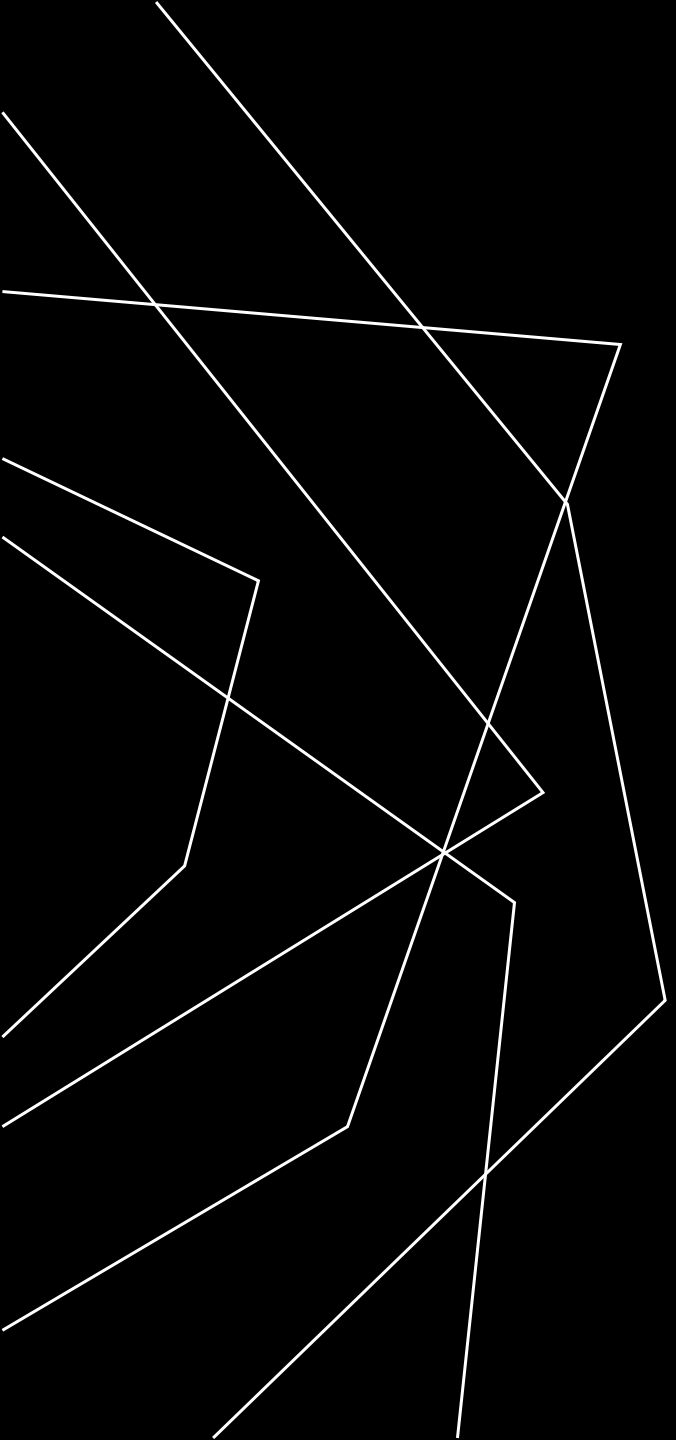
It’s a marathon, not a sprint – but even after a marathon you celebrate and don’t run the next one next morning.

Don’t be afraid to speak up. If something doesn’t work for you, it’s important for the team to know.

Believe in yourself.

Enjoy yourself.

You can do this!



THANK YOU

Any questions?