Hertie School
Data Science Lab

Data Science Summer School 2025

# Python for Data Science

ds3.ai - Dang Ngoc Huy

Hertie School
Data Science Lab

# This course will be quite different

ds3.ai

# What is Python?

Python is an open-source, general-purpose, scripting language

# What is Python?

Python is an open-source, general-purpose, scripting language

**Open-Source**

- Developed and maintained by a global community
- Free to use for anyone, anywhere
- Constantly improved and updated
- Transparent and collaborative development

# What is Python?

Python is an open-source, general-purpose, scripting language

**General-Purpose**

- You can automate a lot of repetitive tasks with Python.
- Python isn't limited to data science, but it's very popular with data scientists!

# What is Python?

Python is an open-source, general-purpose, scripting language

**Scripting**

- No strict definition, but:
- Think of it as a sequence of instructions or commands
- Automates a task or process
- Like a pipeline: Inputs → Processing → Outputs

# What is Python?

Python is an open-source, general-purpose, scripting language

**Language**

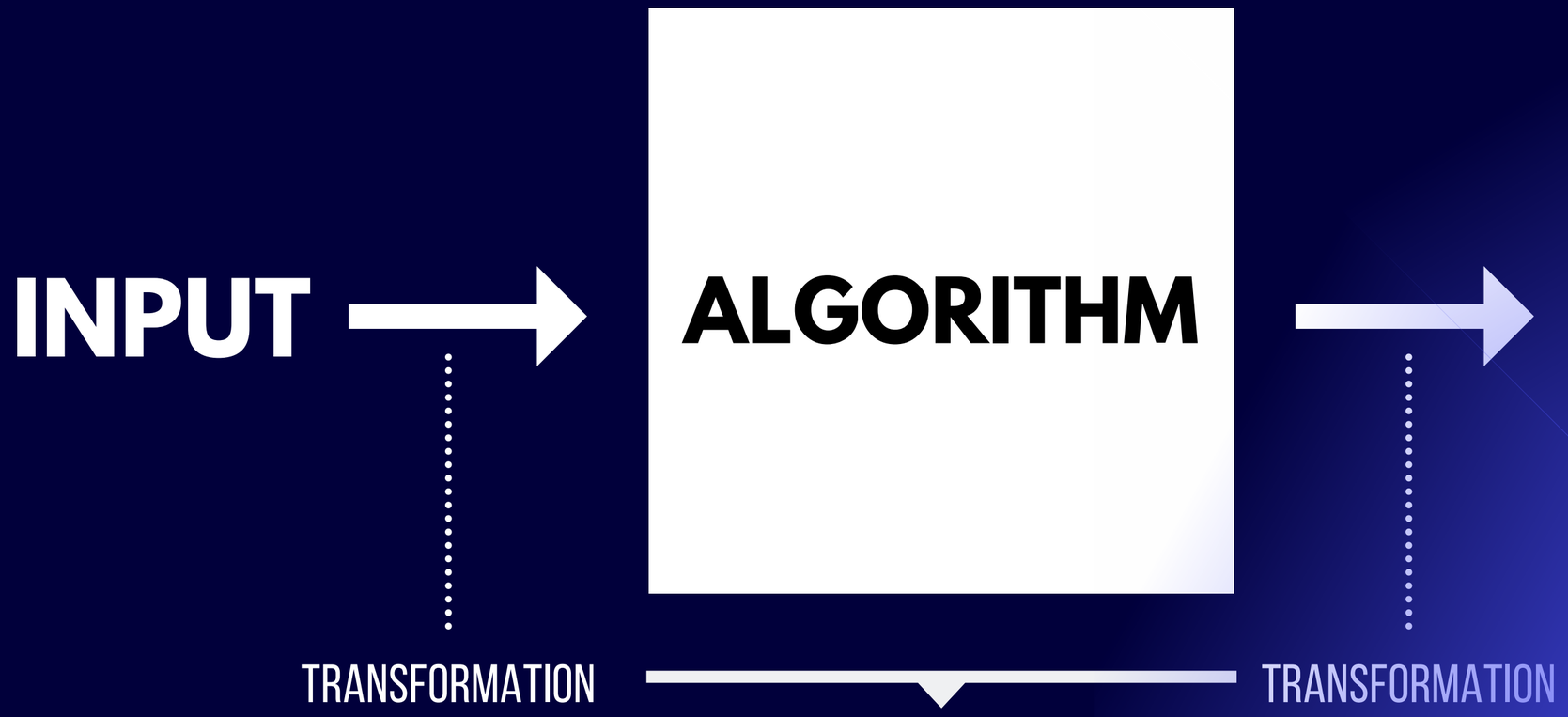
- Python is a programming language, not a ready-made app
- It follows grammar and syntax rules — like any language
- You write commands that the computer interprets and runs
- Because of its flexibility, you can build almost anything with Python!

Hertie School
Data Science Lab

# How does scripting work?

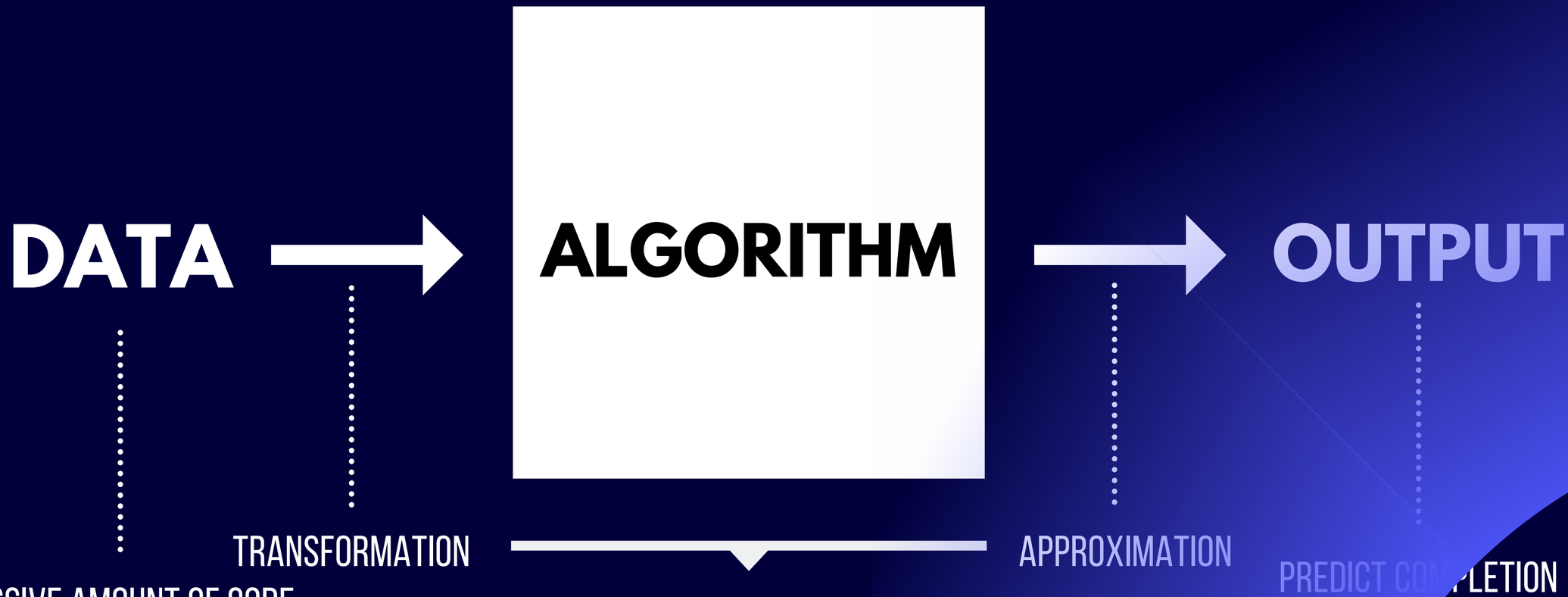
Hertie School
Data Science Lab

# How does scripting work?



Hertie School
Data Science Lab

# How does AI work?



Hertie School
Data Science Lab

# For Researchers: What Can We Use Python For?

Python is great for automating any repetitive computer-based task.
Here are some common research applications:

- Data Collection: e.g., Web scraping, API access
- Data Cleaning & Analysis: Handle messy datasets, transform data for insights
- Data Visualization: Create plots, dashboards, and visual reports
- Machine Learning: Train models to detect patterns or make predictions
- Deep Learning: Work with neural networks for advanced AI tasks

If it’s repetitive and done on a computer — Python can probably automate it.

Hertie School
Data Science Lab

# For Engineers: How Is Research Different from Development?

| Aspect | **Research** | **Development** |
|---|---|---|
| **Usage** | Scripting, prototyping, interactive work | Building full applications |
| **Goal** | Quick experimentation, testing ideas | Production-ready systems |
| **Priorities** | Ease of use, speed of iteration | Portability, resource efficiency, deployment readiness |

In research, we care about flexibility and speed.
In development, we optimize for stability and performance.

Hertie School
Data Science Lab

# Python Research Toolkit

## Code Editors & IDEs

Jupyter Notebook / JupyterLab
→ Interactive, visual, great for exploration and documentation

VS Code
→ Lightweight, supports extensions (Python, Jupyter, Git, etc.)

Cursor (AI-powered)
→ Built on VS Code, integrates AI to help write, edit, and explain code

Pro Tip: Choose your tools based on your project stage:
→ Use Colab/Jupyter to explore, VS Code/Cursor to refine, and GitHub to collaborate

Hertie School
Data Science Lab

# Python Research Toolkit

## Cloud-Based Tools

Google Colab
→ Free cloud-hosted Jupyter notebooks
→ No setup needed, GPU/TPU support, great for sharing

Deepnote
→ Real-time collaboration on notebooks (like Google Docs + Jupyter)
→ Clean UI, great for teaching or teamwork

Kaggle Notebooks
→ Online environment with direct access to datasets and competitions

Hertie School
Data Science Lab

# Python Research Toolkit

**Package & Environment Management**

Conda
→ Popular for managing Python/R environments and dependencies

Mamba
→ Fast alternative to Conda

pip + venv
→ Native Python tools, lightweight and simple for small projects

Hertie School
Data Science Lab

# Python Research Toolkit

**Version Control & Reproducibility**

Git + GitHub
→ Track changes, collaborate, and share code

DVC (Data Version Control)
→ Version control for datasets and ML models

ReproZip / Binder
→ Tools to reproduce and share research environments

Hertie School
Data Science Lab

# Python Research Toolkit

**Version Control & Reproducibility**

Git + GitHub
→ Track changes, collaborate, and share code

DVC (Data Version Control)
→ Version control for datasets and ML models

ReproZip / Binder
→ Tools to reproduce and share research environments

Hertie School
Data Science Lab

# R VS. PYTHON – THE ETERNAL DEBATE

A question that comes up often is why the MDS programme focuses on 2 programming languages, when Python is clearly leading the pack as the default language in machine learning, deep learning and many other data and devops workflow. There are a few reasons:

R is still the superior language for building statistical model with a robust ecosystem of packages maintained by scientists and statisticians. This is an area that is still quite lacking in Python;

Python is the desired language for industry roles, but if you want to work for government agencies, research foundations, think tank, international organisations or remain in academia, R (and Excel) is still the go-to tool in many of these institutions. Being proficient in both gives you the best exposure to the job market and the ability to work in & manage cross-language teams;

Being able to code effectively in multiple languages is a sign of good technical talent to potential employers.

Hertie School
Data Science Lab

# Is knowing how to code still useful?

In the age of AI, what's the point of learning how to code when AI can do a much better and faster job than you?

**KNOWING HOW TO USE PYTHON EMPOWERS YOU TO UNDERSTAND, BE IN CONTROL, AND BE ABLE BUILD BETTER WITH AI**

ds3.ai

Hertie School
Data Science Lab

# Is knowing how to code still useful?

**Coding alone is no longer the star**

What's More Important Now:

Problem Solving:
→ Understanding the why behind the code
→ Framing meaningful questions

Critical Thinking:
→ Choosing the right method, not just using the latest tool
→ Interpreting results in real-world context

Communication:
→ Explaining insights clearly to non-technical audiences
→ Storytelling with data

Hertie School
Data Science Lab

# Is knowing how to code still useful?

**From “Coder” to “Scientist”**

Being a scientist means:

- Asking the right questions
- Designing experiments
- Making sense of complexity
- Using code as your instrument — not your identity

It’s the science in data science that sets you apart.

Hertie School
Data Science Lab

# How to learn coding fast and not forget it

**The Wrong Approach**

- Binge-watch tutorials
- Read tons of docs
- Take notes...
- Then forget everything a week later

ds3.ai

Hertie School
Data Science Lab

# How to learn coding fast and not forget it

ds3.ai

**The Right Way: Learn by Doing — With Purpose**

**Step 1: Start with a specific goal**

Learning without a goal = wandering without a map.
Examples:

- “Build a sales forecast system using Python and Machine Learning”
- “Get a job as a Python developer in 6 months”
- “Automate my reports at work using Python scripts”

Ask yourself: Why are you learning to code?

Hertie School
Data Science Lab

# How to learn coding fast and not forget it

**The Right Way: Learn by Doing — With Purpose**

**Step 2: Create a Smart Roadmap**

- Don’t learn randomly
- Do research on how to achieve your goal
- Use tools like NotebookLM to:
  - Collect relevant materials
  - Ask focused questions
  - Build a personal learning path
  - Stay organized
- Use a Python roadmap to keep track of what other things you can learn about the language

ds3.ai

Hertie School
Data Science Lab

# How to learn coding fast and not forget it

## The Right Way: Learn by Doing — With Purpose

### Step 3: Execute and track Your Progress

- Mark what you've done each day
- Log resources used
- Know what's next
- See your growth visually (charts, streaks, checklists)

ds3.ai

Hertie School
Data Science Lab

# How to learn coding fast and not forget it

**The Right Way: Learn by Doing — With Purpose**

**Step 4: Challenge Yourself, Every Day**

- Use AI (like ChatGPT) to generate coding challenges
- Apply what you learn immediately
- Write code everywhere, from your coding notebooks to your terminal
- Fail fast, make mistakes and learn deeply: Every error = a new brain connection and new lesson

ds3.ai

Hertie School
Data Science Lab

# How to learn coding fast and not forget it

**The Right Way: Learn by Doing — With Purpose**

**Step 5: Reinforce the cycle of Learn → Practice → Apply → Review**

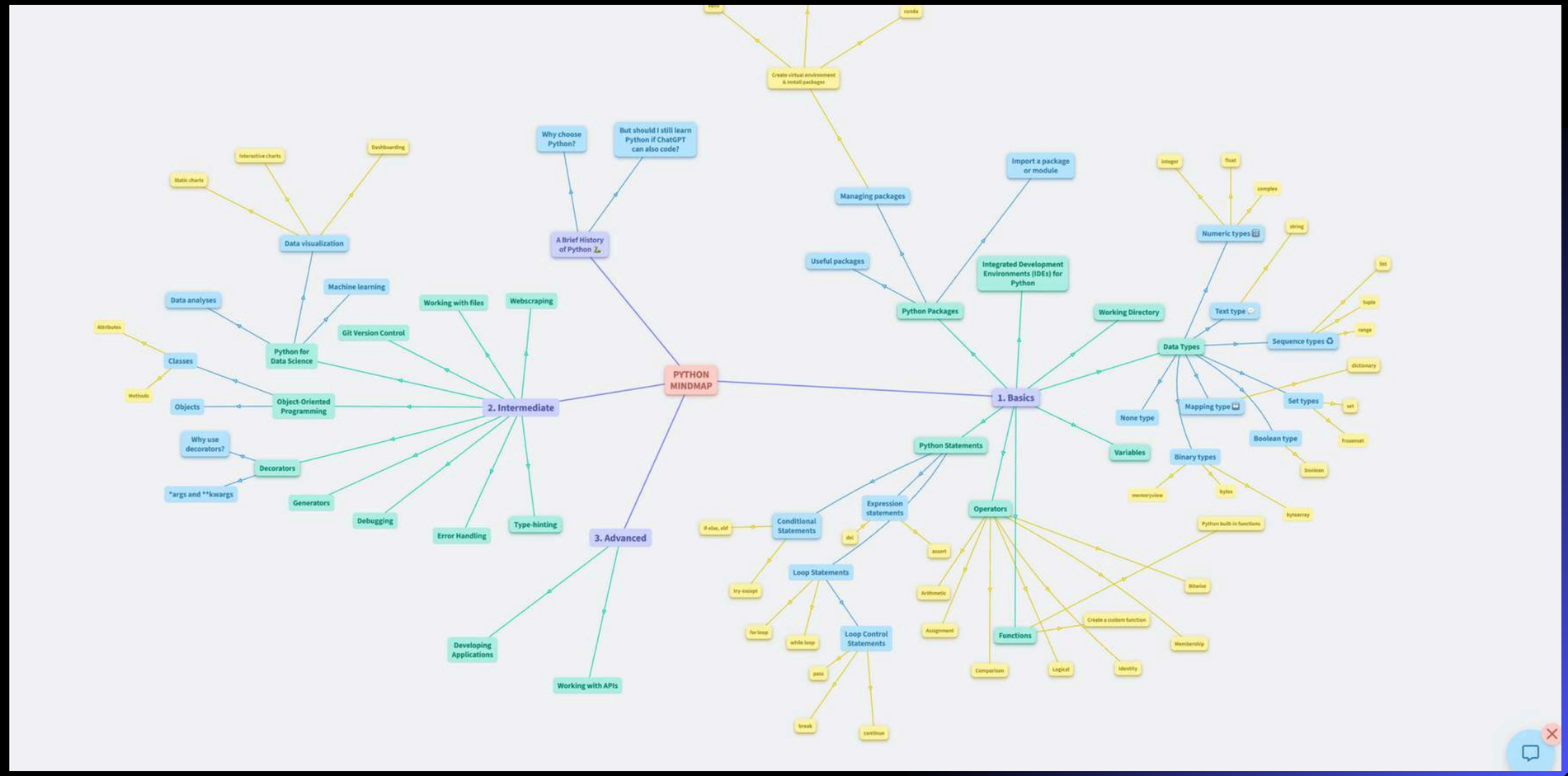
- Don’t just learn facts — build mental connections
- Reflect: how does this concept relate to what you already know?
- Build a personal "knowledge map"
- Teach others or write your own summary

ds3.ai

Hertie School
Data Science Lab

Code is a tool, not the goal.
Use it to build, solve, explore. That’s how you’ll remember it for life

Hertie School
Data Science Lab




# Python Roadmap

ds3.ai

Hertie School
Data Science Lab

# Thank **You**

ds3.ai