

Demo Day & the Road to HPC

Assemble the Studio, present your work, and race matrix multiplication

Day 10 · Week 2

🎲 Demo Day · Full Board

1 Mission briefing

Ten days ago the Studio was empty. Today it is a full board of seven AI capabilities you built with your own hands. Demo Day runs in three acts: **assemble** the whole Studio, **present** it like a scientist, then run the **Grand Benchmark** that reveals the single operation hiding beneath all of it — and points straight at the world of high-performance computing. *Previously:* you unlocked 🎮 Game Master — agents that learn from reward alone.

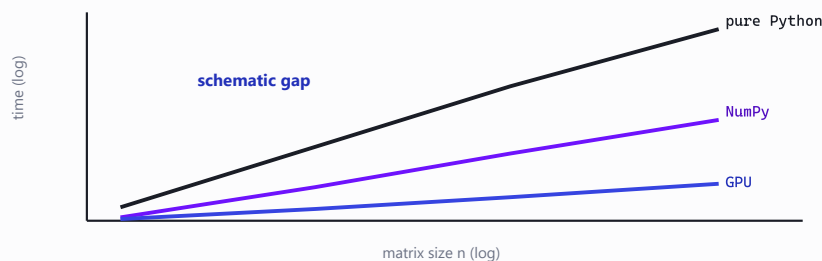
- ☐ Assemble all **7 modules** into one running Studio
- ☐ Present a tight **5-minute demo** with honest model cards
- ☐ Benchmark matmul: **pure Python vs NumPy vs GPU**
- ☐ Compare **$O(n^3)$** work and **measured** speedups
- ☐ Connect every module to `C = A @ B`
- ☐ Keep the plot — it's your ticket into **HPC**

2 Field guide the three acts

Act I — Assemble. Every notebook exported its model files into `ai_studio/models/`. Launch the Studio and each satisfied module unlocks its tab. Ten files, seven modules, one app.

Act II — Present. A demo is a science talk: **hook** → two favorite modules **live** → your **hardest bug** → what **surprised** you. A good **model card** states honestly what a model does, how accurate it is, and where it fails. Honesty is the science.

Act III — the Grand Benchmark. Multiply two big matrices three ways — a hand-written triple loop, NumPy, and the GPU — then time them and plot the result on log-log axes. The cost grows like **$O(n^3)$** , and the measured speedup depends on the hardware, size and numeric precision.



The money plot: the same math, three speeds. Log-log axes help reveal scaling trends; measured slopes and speedups vary with size and hardware.

📊 **Matrix Watch — the payoff.** Every conv, every attention head, every DQN prediction you built runs on `C = A @ B`. Matrix multiplication was the whole game all along — and high-performance computing makes it scream. **Keep this plot.**

A precise reading

The classical dense matmul counts about $2n^3$ floating-point operations, but measured timing curves need not have slope 3 at every size. Compare the same size and precision, warm up, synchronize the GPU, and state whether transfer time is included. Speedups are measurements, not guarantees. Matrix products power many neural layers; tabular Q-learning is a scalar update, and other operations still matter.

3 Lab missions your Demo Day checklist

ACT I · ASSEMBLE THE STUDIO

MODULE	FROM	MODEL FILE(S)	✓
 Prediction Machine	Day 2	prediction_machine.joblib	☐
 Pattern Finder	Day 3	pattern_finder.json	☐
 Digit Reader	Day 5	digit_reader.pt	☐
 Photo Detective	Day 6	photo_detective.pt · photo_detective_classes.json	☐
 RPS Arena	Day 7	rps_arena.pt · rps_labels.json	☐
 Language Lab	Day 8	language_lab.json	☐
 Game Master	Day 9	game_master_qtable.npy · game_master_dqn.zip	☐

1 Take the roll

Tick every file in the table above.

Expect: all 10 present in `ai_studio/models/` — nothing missing.

2 How big are your models?

Measure each saved file's size on disk.

Expect: a few KB for the JSON tokens up to tens of MB for the `.pt` models.

3 Launch the Studio

Run `python ai_studio/app.py` from the repo root.

Expect: the banner read **7 / 7 modules unlocked**.

ACT II · PRESENT LIKE A SCIENTIST

1 Write an honest model card

For one module, note what it does, its accuracy, and where it fails.

Expect: real numbers and honest failure cases — not hype.

2 Sketch your 5-minute demo

Hook → two favorite modules live → your hardest bug → what surprised you.

Expect: a run-of-show you can deliver without notes.

ACT III · THE GRAND BENCHMARK

1 Rebuild matmul from memory

Write the triple-loop `matmul` with no peeking.

Expect: it match NumPy within floating-point tolerance on a small test case.

2 Time the hand-written version

Clock your loop, NumPy, and the GPU (best of 3 each).

Expect: milliseconds for NumPy vs whole seconds for the loop.

3 The money plot

Plot time vs n on log-log axes for all three methods.

Expect: timing curves whose gaps and slopes depend on size.

4 Put a number on the gap

Divide the slow time by the fast one.

Expect: a measured speedup at matching matrix sizes.

5

Graduate 🎓

Run the final cell and save your money plot — a souvenir worth keeping.

Expect: a saved figure and a send-off to the HPC weeks.



4 Cheat sheet benchmarking matmul

<code>for i,j,k: C[i,j] += A[i,k]*B[k,j]</code>	Matmul by definition — the $O(n^3)$ triple loop.
<code>t = time.perf_counter()</code>	Read a high-resolution clock (take best of 3).
<code>torch.cuda.synchronize()</code>	Wait for the async GPU before you stop the clock.
<code>gflops = 2*n**3 / t / 1e9</code>	Turn seconds into GFLOPs (a matmul is $2n^3$ ops).
<code>C = A @ B</code>	The one line many neural-network layers rely on.
<code>sys.path.insert(0, str(REPO))</code>	Let the notebook import the Studio package.
<code>from ai_studio.studio import build_demo</code>	Assemble all seven modules in code.
<code>build_demo().launch()</code>	Bring the whole Studio to life.

5 🎲 Light up the full board

🎲 The full board — 7 / 7 modules

No new module today. Instead, every tab you built lights up at once. Launch it and confirm all seven are live:

```
$ python ai_studio/app.py
```

✓ 🤖 Prediction Machine · Day 2

✓ 📄 Digit Reader · Day 5

✓ 🎮 RPS Arena · Day 7

✓ 🎮 Game Master · Day 9

✓ 🎨 Pattern Finder · Day 3

✓ 🕵️ Photo Detective · Day 6

✓ 💬 Language Lab · Day 8

6 Mini-glossary

benchmark

A controlled speed test comparing methods fairly.

Big-O / $O(n^3)$

How cost grows with size; matmul scales as n cubed.

HPC

High-performance computing — big math, made fast at scale.

FLOP / GFLOPs

A floating-point operation · billions of them per second.

vectorization

Replacing Python loops with fast bulk array operations.

Julia

A fast scientific language built for exactly this kind of number-crunching.

7 Show & tell

1. Which two modules will you demo live, and what's your 10-second hook for each?
2. What was your hardest bug across ten days, and how did you finally kill it?
3. Your triple loop vs the GPU: how many times faster? Why does that gap matter for real AI?

THE LAB

Certificate of Completion

This certifies that

has completed AI Studio — Intro to AI and is hereby recognized
as a

Certified AI Explorer



Prediction Machine



Pattern Finder



Digit Reader



Photo Detective



RPS Arena



Language Lab



Game Master

Instructor

Date

The Lab