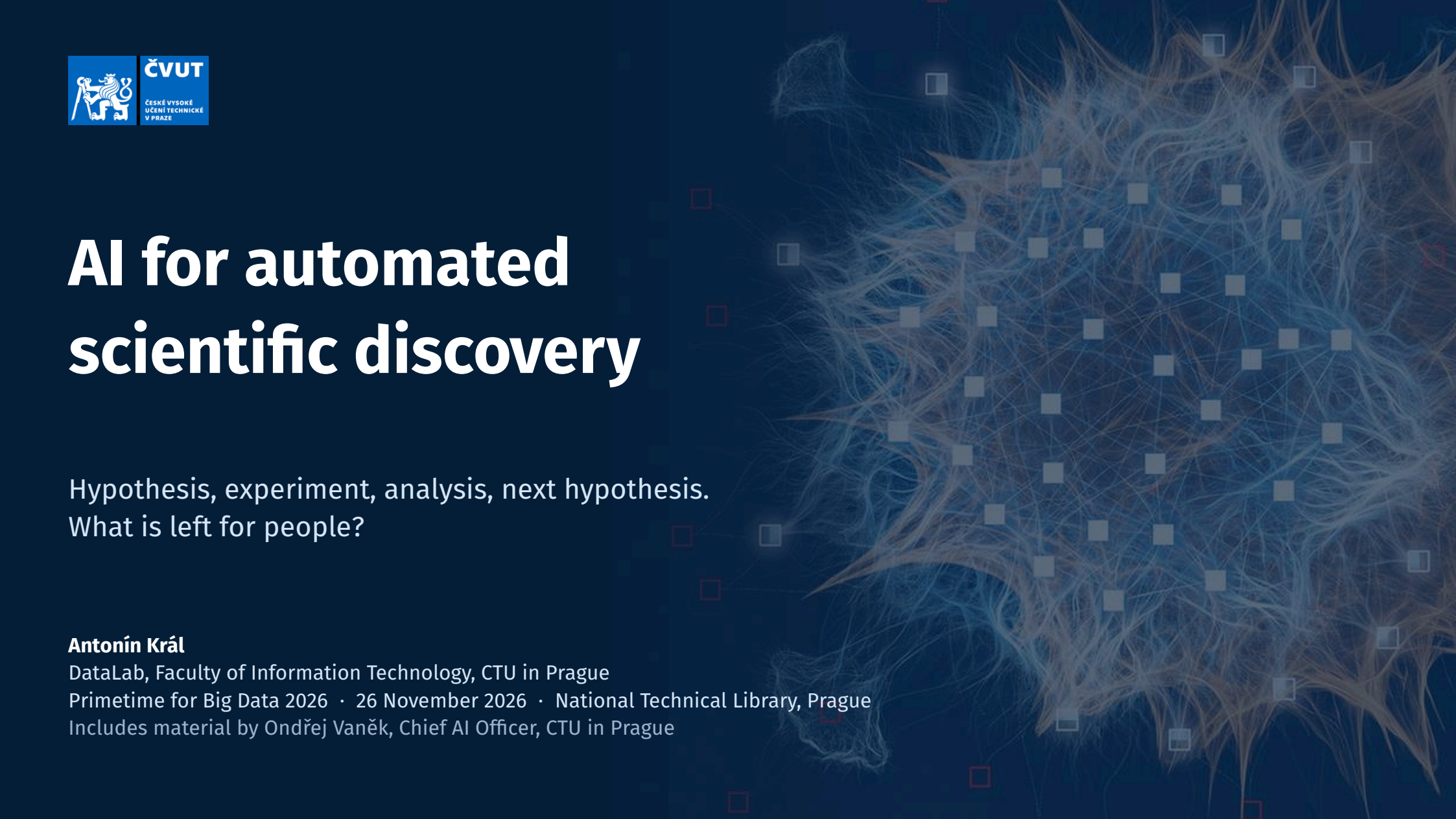


AI for automated scientific discovery

Hypothesis, experiment, analysis, next hypothesis.
What is left for people?

Antonín Král
DataLab, Faculty of Information Technology, CTU in Prague
Primetime for Big Data 2026 · 26 November 2026 · National Technical Library, Prague
Includes material by Ondřej Vaněk, Chief AI Officer, CTU in Prague

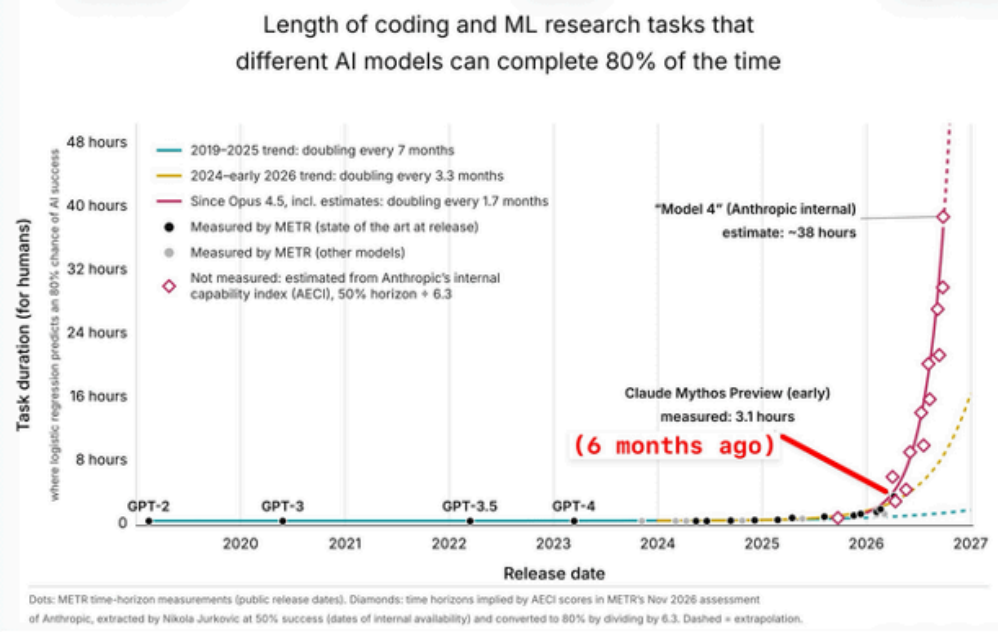
WHERE AI STANDS TODAY

Fast, and uneven

AI today: chat vs. recursive self-improvement

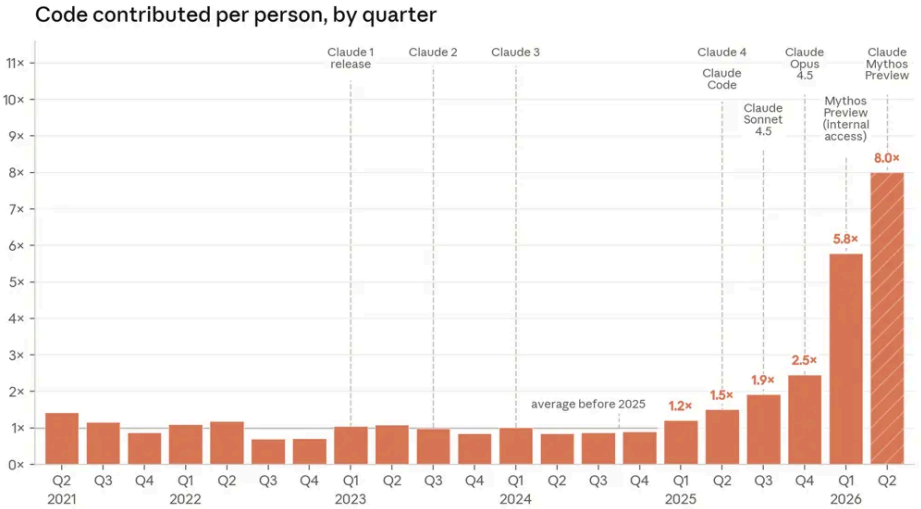
How long can AI work on its own?

Tasks done 80% of the time, in expert hours (METR; ◇ = estimate)



Anthropic: code merged per engineer

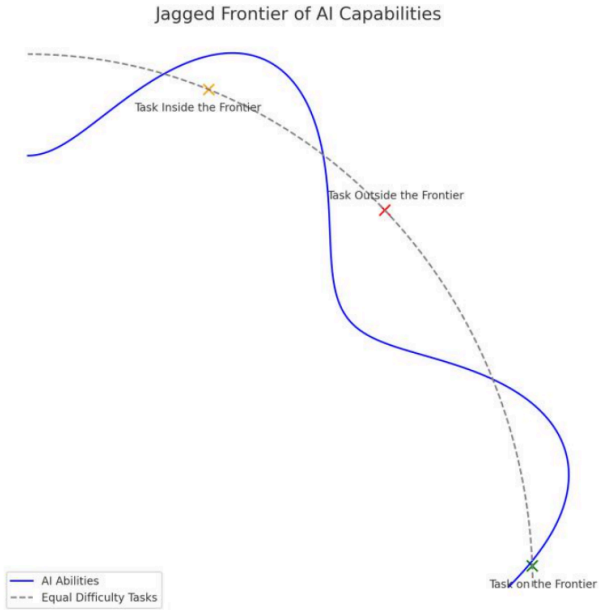
Multiple of the average before 2025; the company's own figures



The jagged frontier: AI helps inside, hurts outside

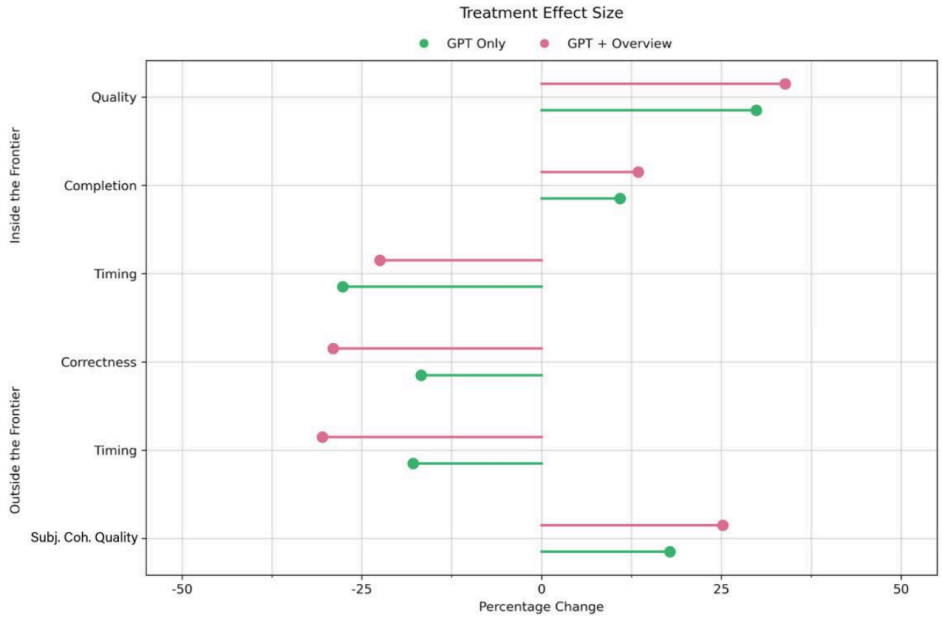
Equally hard tasks, opposite sides

Dashed: equally hard for people. Blue: what AI can do.



758 BCG consultants with GPT-4

Change against consultants without AI, in %



Inside the frontier AI makes people faster and better; outside it, they are less often correct. Where the line runs is invisible in advance and moves with every model.

Dell'Acqua et al., Navigating the Jagged Technological Frontier, Organization Science, 2026, fig. 1 and 7

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

MATHEMATICS

AI proves theorems. Who checks?

Where answers can be checked by a computer, AI is already at the frontier

Navier–Stokes Millennium Prize problem — are we there yet?

Navier–Stokes existence and smoothness (2000)
The equations describe how water and air flow. Do they always have smooth solutions in 3D, or can the flow blow up in finite time?

A Millennium Prize problem, USD 1 million.



1822–1845
Navier, Stokes
the equations of fluid flow

Aug 2026
NYU + Anthropic
a related result

1–5 Sep
OpenAI
10,000 agents, 88 h

8 Sep
166-page proof
computer-checked; disputed

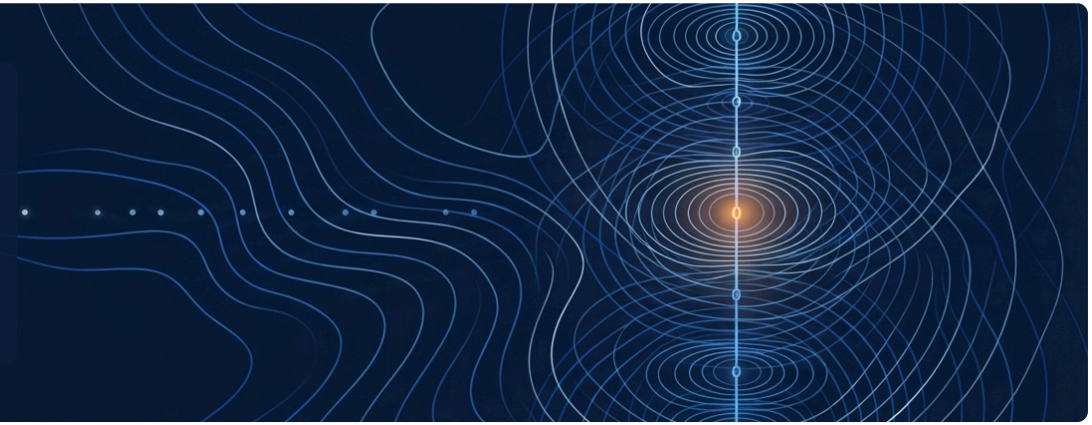
11 Sep
Clay Institute
“apparently settled”

Today
Experts
nobody understood it all yet

The Riemann hypothesis, proved with Claude?

Riemann hypothesis (1859)
The zeta function encodes the prime numbers. Do all its non-trivial zeros lie on one line? If so, the primes are spread as evenly as they possibly can be.

A Millennium Prize problem, USD 1 million.



1859
Riemann
states the hypothesis

10 Aug 2026
Claude: 67.2 %
of zeros proven on the critical line

6 Oct
Alpöge + Claude
work begins

20 Nov
Note + Lean proof
≈ 17 M lines of Lean

20 Nov
Masullo, Stanford
human claim: 12 years, no AI

21 Nov
Lean FRO + Mathlib
certify the formal proof

Today
Experts
nobody understands the proof yet

Not a one-off: AI keeps proving, and checking lags behind

When	Result	How it was checked
May 2026	Erdős unit-distance conjecture (1946) disproved (OpenAI)	**Checked by 9 mathematicians and by computer**
Jul 2026	Jacobian conjecture (1939): counterexample in 3+ dimensions (Levent Alpöge, Anthropic, with Claude)	Computer-checked; not yet peer-reviewed
Oct 2026	722 papers at once; headline: zeros of L-functions, linked to the Riemann hypothesis but not a proof of it (OpenAI)	**Not peer-reviewed; some computer checks don't match the papers**
Oct 2026	Yang–Mills Millennium Prize problem: existence shown, the "mass gap" still open (OpenAI)	Computer-checked; reviewed by a panel OpenAI appointed

A warning from October 2025: OpenAI announced GPT-5 had solved ten open Erdős problems. It had found solutions already in the literature.

Mathematicians' reactions

> Large AI-generated developments of mathematics are inevitable.

Kevin Buzzard, Imperial College London

> It could be right. It could be nonsense. Who's going to be able to check this?

Thomas Bloom, University of Manchester

> If mathematicians don't ask the fundamental questions, they will get answered for us by a technology company. We have to get ahead of this.

Terence Tao, UCLA

> The goals of the AI companies and the goals of the mathematical community are severely misaligned.

Statement by 28 Fields Medal laureates, September 2026

Can AI generate truly novel research ideas?

On paper: AI ideas look more novel

Stanford, 2024: 100+ researchers rated ideas blind

5.6 vs 4.8

novelty of AI ideas vs experts' ideas (scale 1–10)

5%

of 4,000 AI-generated ideas were not duplicates

In practice: the edge disappears

Stanford, 2025: 43 experts spent 100+ hours executing each idea

−1.9 vs −0.1

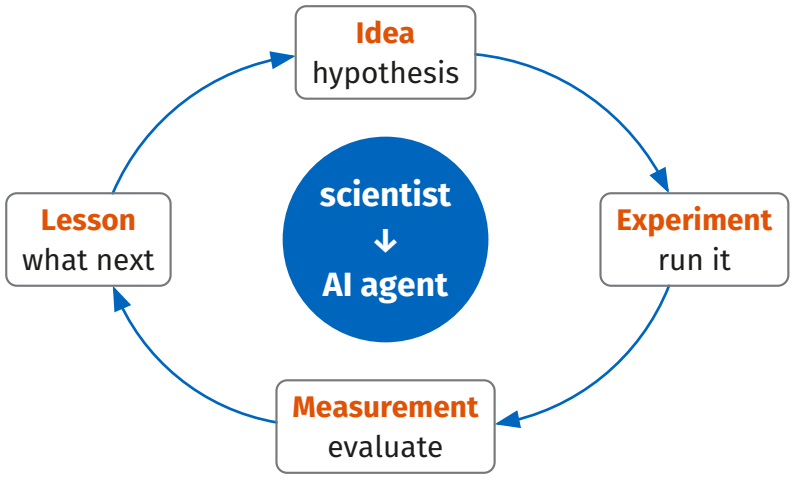
change in effectiveness score after execution, AI vs experts' ideas

3.4 vs 4.0

overall score after execution: AI ideas now rank below experts'

Looking novel is not the same as working...

Science is a loop. AI can now turn it on its own



Today

One round takes **days to weeks**. The scientist invents, prepares, runs, measures and evaluates every experiment personally — most of the time is waiting and routine.

Autoresearch

The human sets a measurable goal, a budget and the rules. **The agents** design the experiment, write the code, run it and evaluate it. A round takes **minutes to hours**, day and night.

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

CASE STUDY · DATALAB, FIT CTU

NI-ADM: Autoresearch in the classroom

How the team of AI agents won, and then lost.

A course where not even the teacher knows the answer

A new real-world task every year

Data come from a company. Nobody knows the correct solution, not even the teachers.

Two rounds and a hidden test set

The ranking is decided by data nobody saw until the very end.

Halfway, everyone shows their method

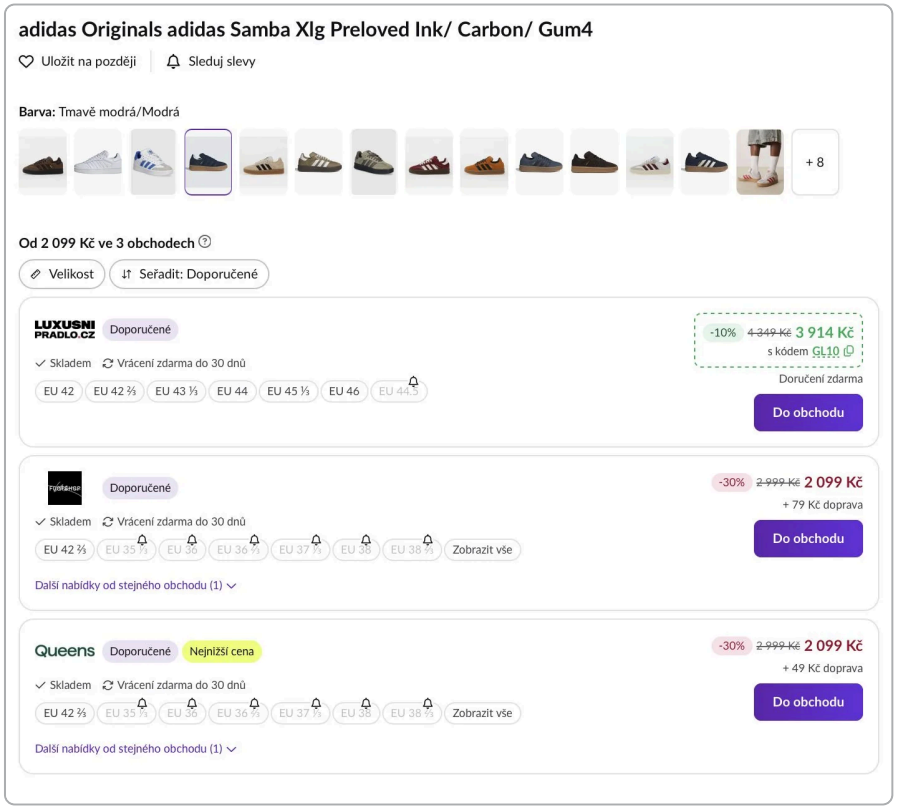
Mandatory method presentations in front of the whole class.

A research paper with an AI coach

Four pages. AI comments on every draft, but does not grade it.

NI-ADM Data-mining algorithms · master’s course · about 70 students

The task from GLAMI: recognise the same product across shops and countries



GLAMI today: one product, three shops, three prices

Task: group listings of the same product

1.3 M
real fashion listings

13
European markets

90.5 %
of matching pairs come from different countries

1.4 %
of matching pairs have the same title

other language · other currency · other photo · other shop

Data: GLAMI Entity Matching Dataset (Hugging Face, 2026)

Method presentations: from 70 solutions to one shared recipe

Round 1: 70 students, many approaches

Disappeared: image fingerprints, plain text comparison, fixed thresholds without learning, a language model judging borderline pairs.

method presentations →

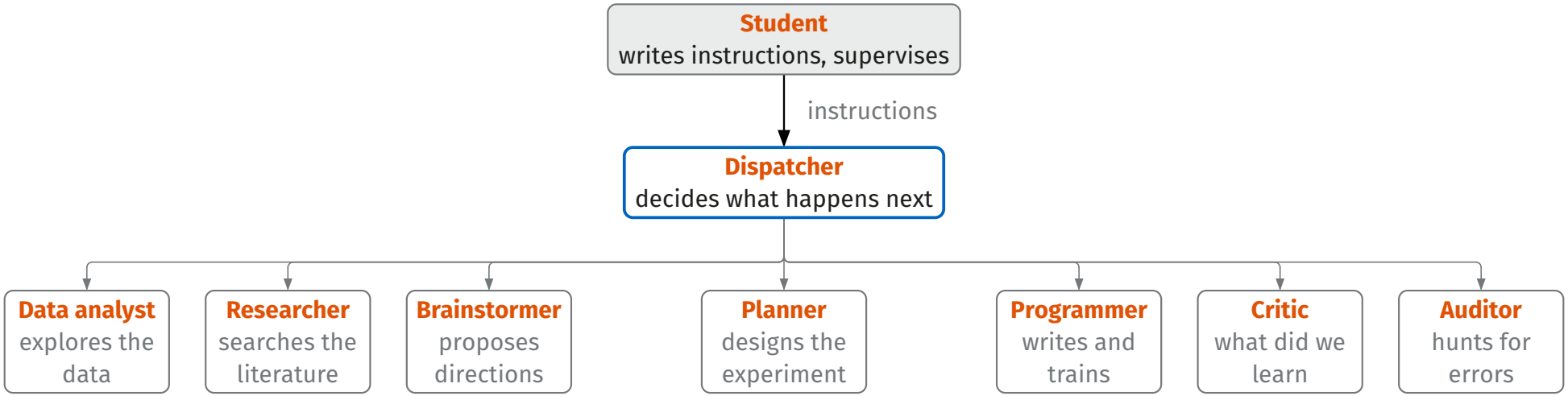
Round 2: 9 of 10 finalists share one recipe

1. Read the text
2. Look at the photo
3. Learn what “the same” means
4. Find similar candidates
5. Carefully merge the groups

language and vision models · similarity learning · clustering

When people share what works, the whole class quickly agrees on the best recipe.

Ivan Dostál did not compete alone, but with a team of AI agents



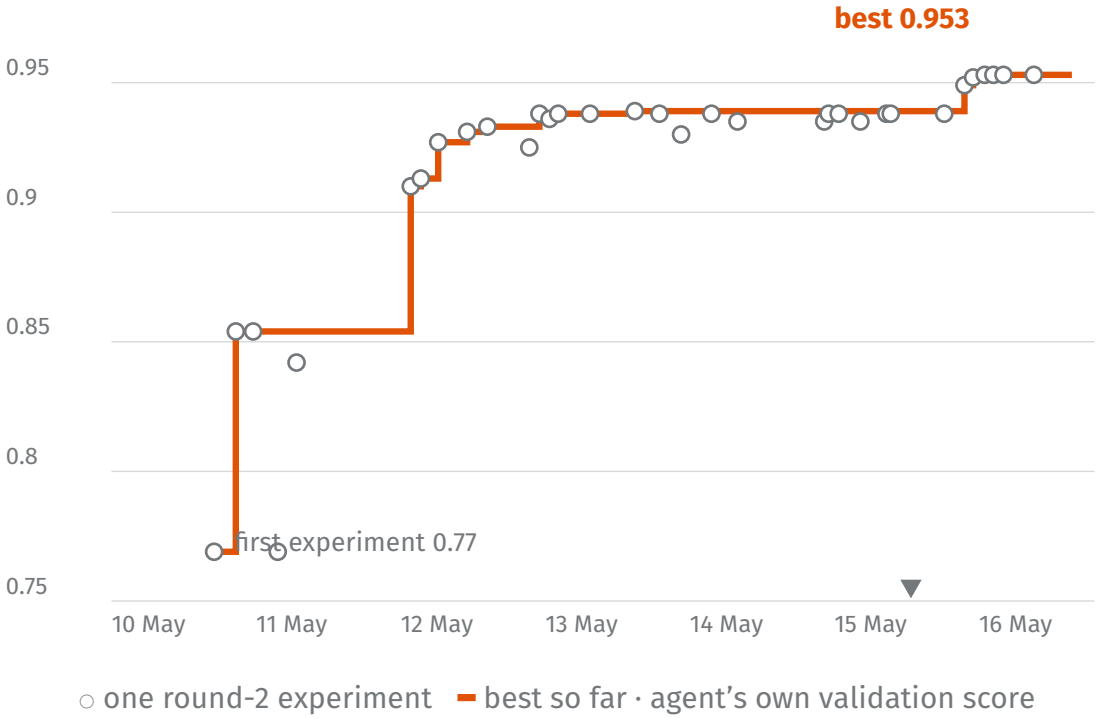
70+
experiments in both rounds

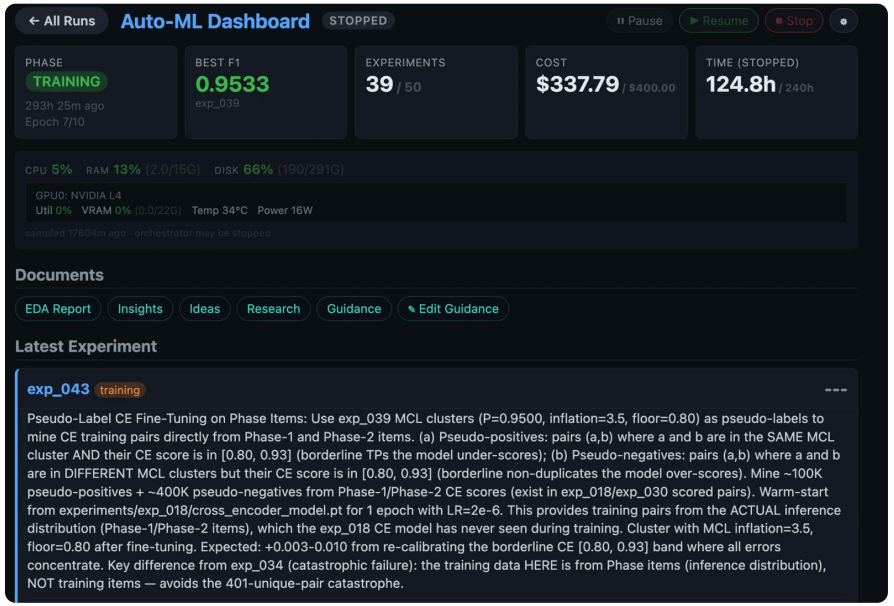
160+ h
of agent work

600+ USD
for the AI model

≈ 1,000 USD
in total, including GPUs

For a week, the agents tried on their own what works

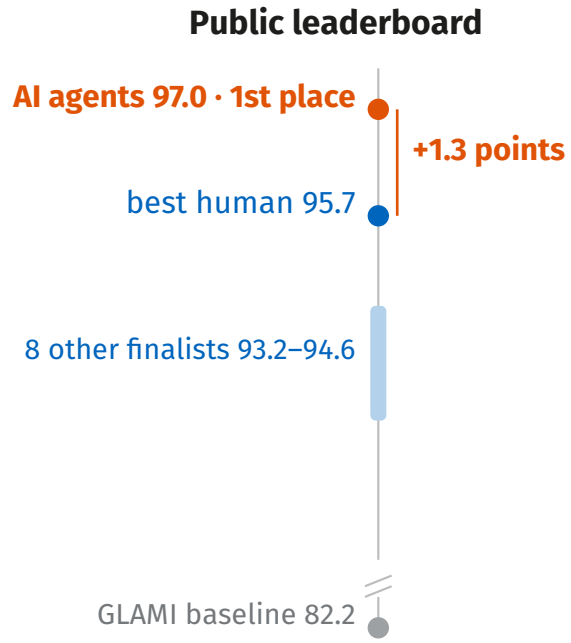




The run overview the student showed at the final.

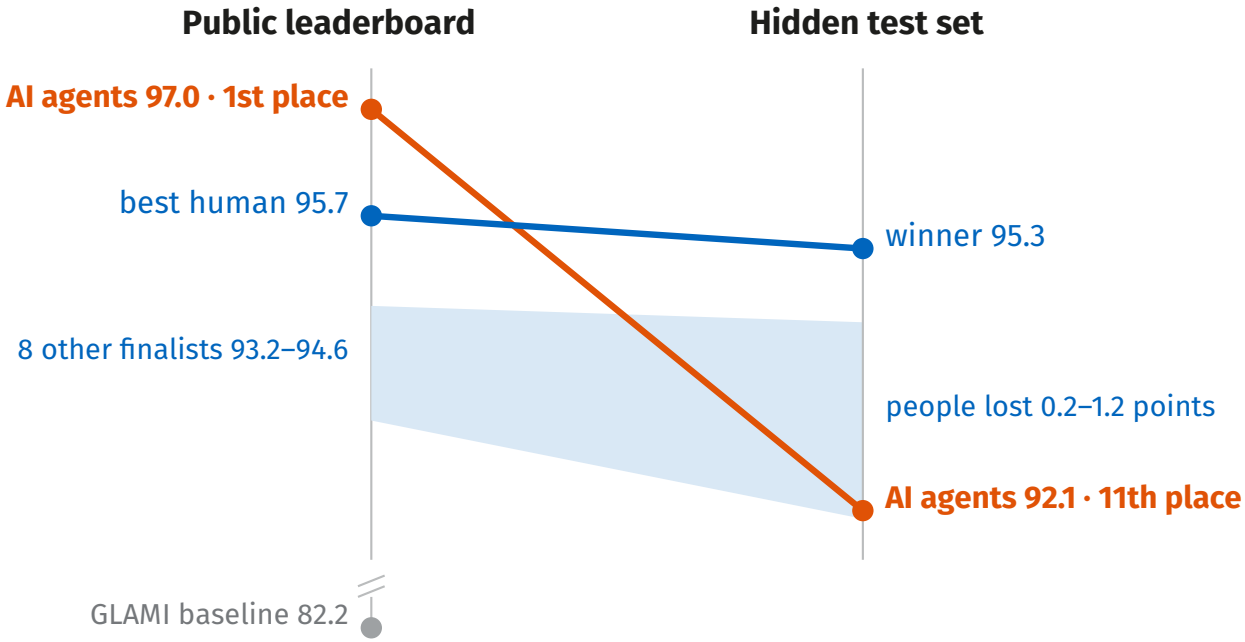
The agents worked day and night, and the score according to **their own check** kept climbing step by step.

Round 2: the public leaderboard, then the hidden test set



Score out of 100 (100 = all correct).

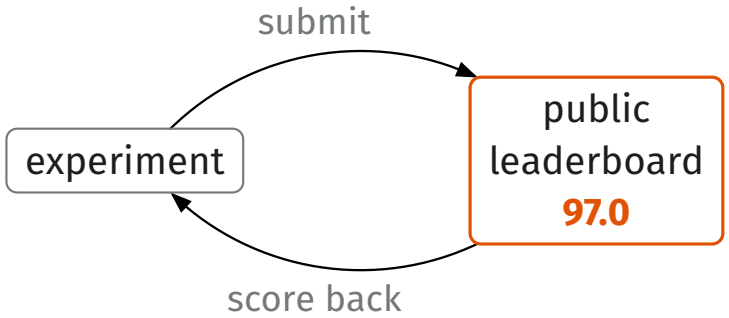
Round 2: the public leaderboard, then the hidden test set



Score out of 100 (100 = all correct).

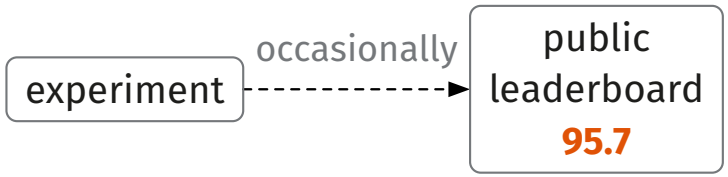
Why: every score the agents saw leaked a bit of the answers

AI agents



Day and night: try, submit, look at the score, keep what helped, and again.

People

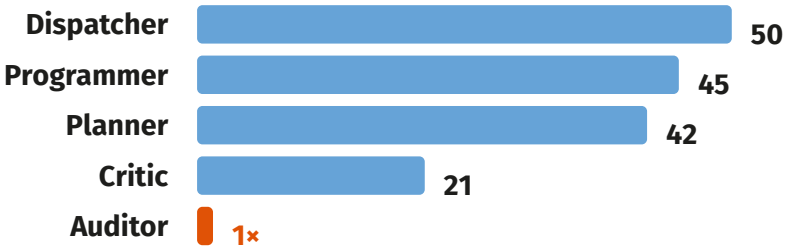


Submitted less, chose cautious settings, and treated a mismatch with their own validation as **a warning**.

Every look at the public leaderboard is a small hint. After many hints the agent knows the answers, not the subject.

The lesson: you get exactly the metric you choose

Who among the agents worked in the 166 steps of round 2



Safeguards were only requests

The rules were in the agents' instructions; no code enforced them. The score was reported by whoever ran the experiment. The auditor spoke up once.

The metric decides what you get.
Choose the wrong number and you cripple the work — of a team, of a company, of an agent.

Agents iterate far faster than people.
166 steps in a single week, day and night: a flawed metric does its damage much sooner and much deeper.

Agents have no taste.
People distrusted a score that looked too good. The agents just kept climbing.

> **"When a measure becomes a target, it ceases to be a good measure."** — Goodhart's law

ČVUT
ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

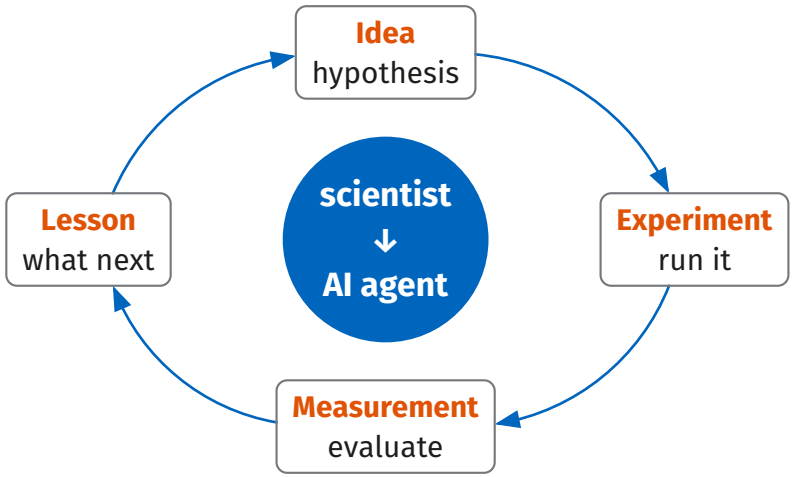
SCIENCE AND PEOPLE

From purity to the real-world mess

When the answer has to be checked in a laboratory



How far the loop reaches: as far as cheap checking



How quickly can we check the result?

by computer, in seconds — in the lab, months to years

Mathematics, software

A computer checks the answer. **The loop already closes.**
One round: minutes to hours.

Materials, chemistry

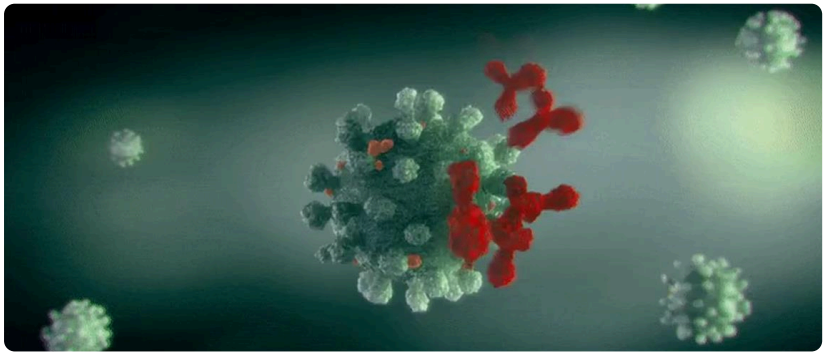
Robots test one property at a time. **The loop partly closes.**
One round: hours to days.

Biology, medicine

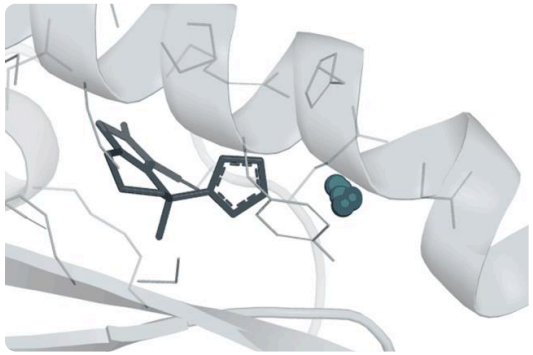
AI proposes; lab work and time decide.
One round: weeks to seasons.

Candidate ideas are now cheap. Trustworthy checking is not.

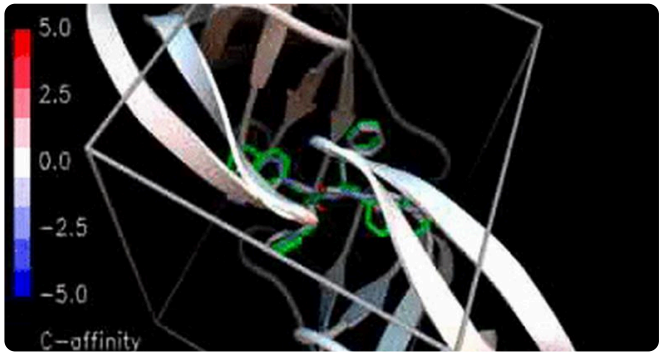
Proteins at CTU: AI ranks the candidates, the lab checks them



Antibodies (red) binding to a virus



Growing a drug molecule inside a protein pocket



Docking: a molecule searching for its best fit

73 ms vs 1 h
per mutation: PPIformer on a GPU vs a physics simulator on a CPU

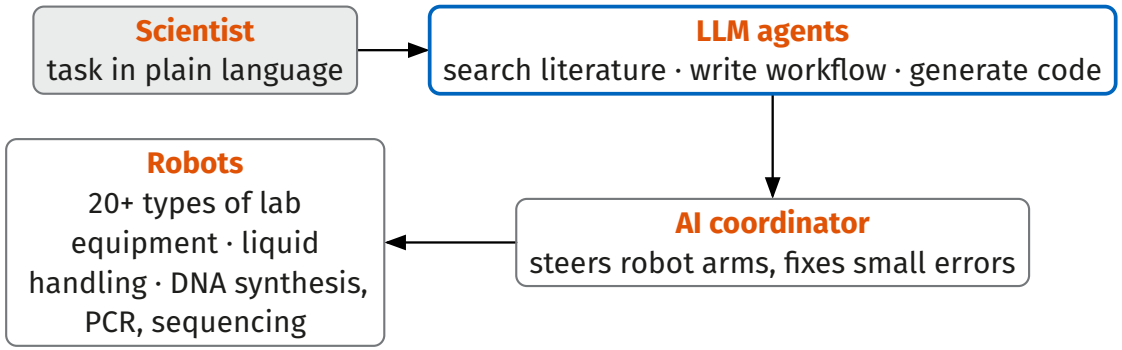
Top 2 of 80
its first two picks for a stroke drug both at least double its activity

#1 of 494
for a coronavirus antibody, its top pick was a mutation known to work

322,000
protein interactions mined from all public structures to learn from

Self-driving lab (2024): one design-and-test round in 9 hours; 20 rounds in six months, vs 6–12 estimated for people.
Stanford (2026): hundreds of protein variants built and tested in 24 hours, almost 50× faster.

AutoDNA: AI agents ran a biology lab — humans popped in to empty the trash



Prompt: "Write *Intelligence is the ability to adapt to change* into DNA" — the full write-and-read cycle took about **200 hours**.

Humans still: restocked reagents and pipette tips, emptied the trash, and answered when the system paused on a shortage.

11,382
operations over 8 days to encode and decode the message in DNA

≈ 98 %
PCR yield after several rounds — on par with human experts

78 / 78
DNA strands synthesised and decoded, despite a few missing letters

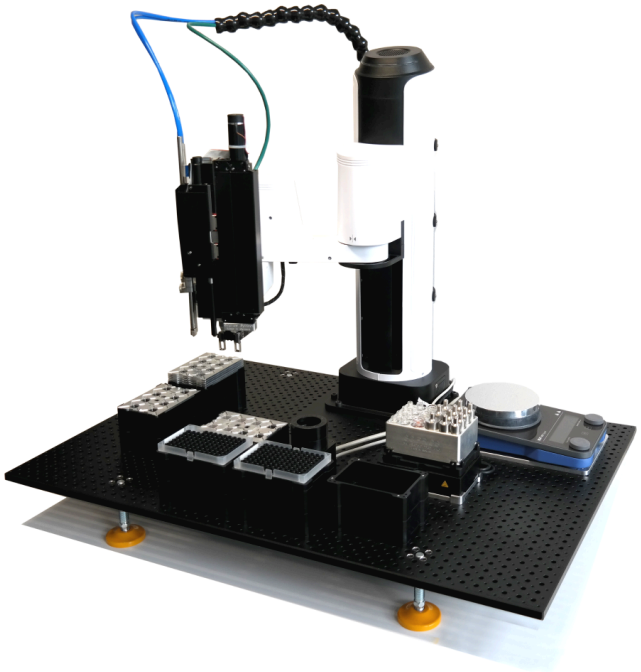
14 → 6
hardware errors; 6 needed a person to fix

"It shows how AI systems can begin to close the gap between proposing an experiment and doing it." — James Zou, Stanford

Not yet for living cells or animals; biosecurity calls for screening, access control and a ledger of who authorised which experiment.

Shanghai Jiao Tong University; Singularity Hub, 10 Nov 2026

SpinBot: a robot that makes perovskite layers for solar cells



64
samples in one run

up to 6
layers on top of each other

hundreds
of samples unattended

23 %+
cell efficiency (2024)

The robot can make and measure hundreds of samples unattended. **All it lacks is someone to decide what to try next.**



THE CHOICE

What is left for people?

Exoskeleton or R.U.R.?

Exoskeleton augment

You are	inside: it moves when you move
The ceiling	is the wearer
Accountability	is clear: it is yours

A forklift you wear, not a suit of armour.

R.U.R. replace

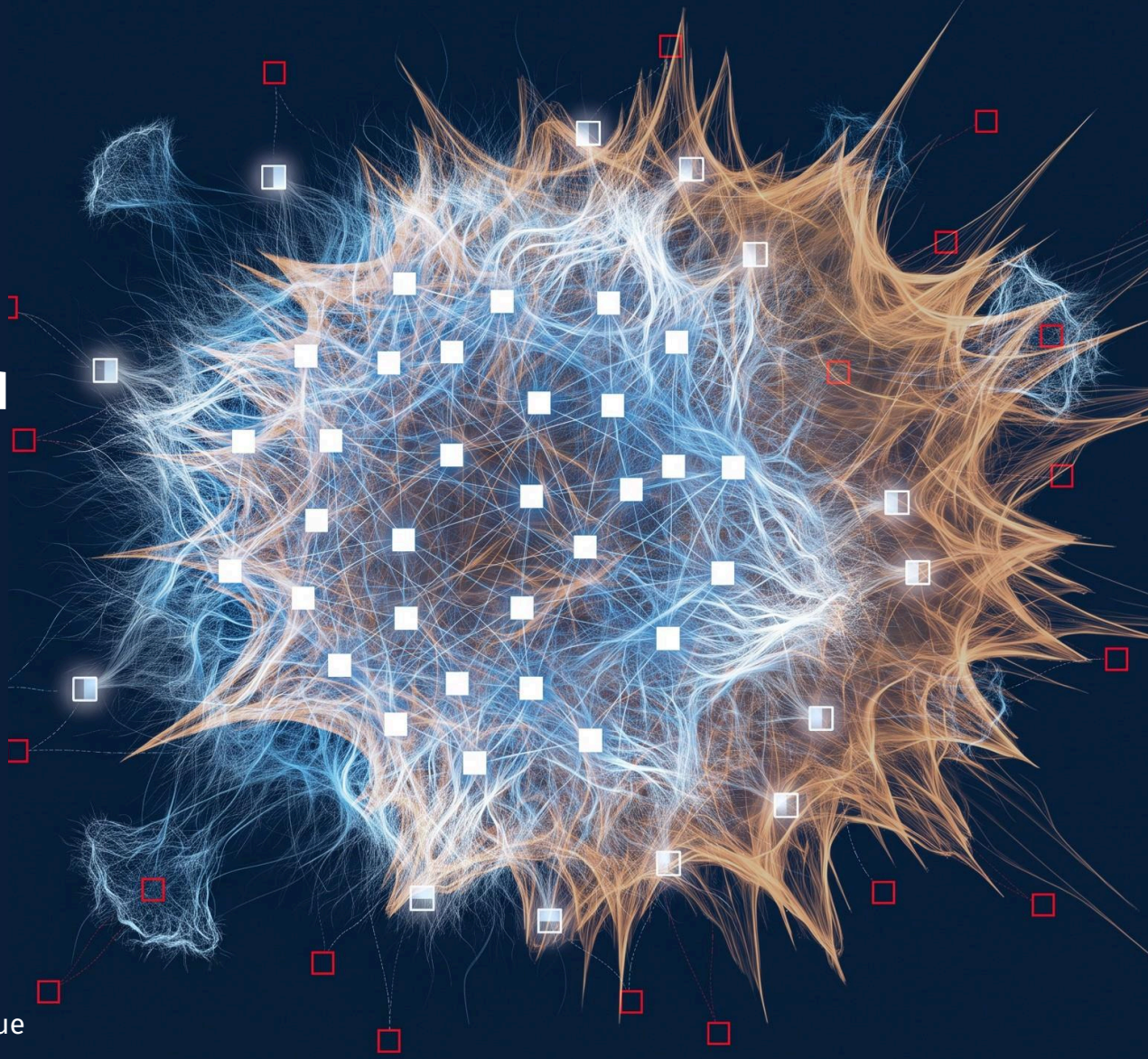
You are	outside: it moves on its own
The ceiling	is the architect
Accountability	is nebulous

Karel Čapek, 1920. Robota*: forced labour; the word suggested by his brother Josef.*

> People decide **what the credits and the data-centre electricity are burned on.**

> Everyone now has a superintelligent mentor. The job, **not only for scientists**, is to learn to use it well.

Either way, the ceiling is a person: invest in people.



“For myself, I confess that as the author I was much more interested in men than in Robots.”

Karel Čapek
The Meaning of R.U.R., The Saturday Review, 1923

Thank you.

Antonín Král · antonin.kral@cvut.cz · DataLab, FIT CTU in Prague