The Planner Museum

Evaluating Classical Planners Over Time

Arnaud Lequen Oliver Joergensen Windy Phung
Elliot Gestrin Damien Van Meerbeeck Markus Fritzsche Dominik Drexler
Jendrik Seipp

Machine Reasoning Lab
Linköping University, Sweden

Outline

Some results from IPC 2023

The International Planning Competition (IPC)

Running since 1998, it lets planners compete on a common set of benchmarks.

In **2023**, **LAMA**, whose first version dates back from **2008**:

- ranked **4th out of 22** in the Satisficing Track,
- **outperformed every planner** in the Agile Track.

Some results from IPC 2023

The International Planning Competition (IPC)

Running since 1998, it lets planners compete on a common set of benchmarks.

In **2023**, **LAMA**, whose first version dates back from **2008**:

- ranked **4th out of 22** in the Satisficing Track,
- **outperformed every planner** in the Agile Track.

Have we actually made progress in (non-optimal) classical planning over the past decades?

Our answer: the Planner Museum

Inspired by the **SAT Museum**, we revisit the history of the IPC.

Contribution 1: A digital museum of planners

Public repository for the source code for **29 IPC planners**, patched when needed, each shipped as an **Apptainer container** that runs on modern hardware.

Contribution 2: A cross-era evaluation

We run them all on the **same hardware** over adapted benchmarks spanning the entire history of the IPC, and quantify the rate of improvement.

Classical planning

We focus on **non-optimal**, **domain-independent classical planning**

- **1998:** first IPC, with the introduction of **PDDL** for expressing planning problems.
- Held every 2 to 3 years since
- Scope broadened to temporal, probabilistic, learning planners and to multiple tracks (Satisficing, Agile, Optimal).

Automated planning and the IPC

Classical planning

We focus on **non-optimal**, **domain-independent classical planning**

- **1998:** first IPC, with the introduction of **PDDL** for expressing planning problems.
- Held every 2 to 3 years since
- Scope broadened to temporal, probabilistic, learning planners and to multiple tracks (Satisficing, Agile, Optimal).

How to compare planners from different eras?

Coverage: The number of instances solved within time and memory limits.

Which planners make it into the museum?

- Recent IPCs feature 20+ planners per track
- We only keep some distinguished ones

Selection criteria

- **Top 3** per track (Agile / Satisficing), per competition.
- Exclude planners already distinguished earlier, to highlight **algorithmic** progress, not just code optimization.

Making 25-year-old planners run today

Planner adaptation

- Same OS for all (Ubuntu 24.04 LTS) & same versions of compilers
- Modern GCC sometimes rejects old code
- PDDL features that old planners can not parse (e.g. `:equality`) are compiled away

Reproducibility

One **Apptainer recipe per planner** $\Rightarrow$ standardized images with a unified interface.

- All artifacts online (code, containers, benchmarks)
- Experiments run with Downward Lab

A benchmark set that challenges everyone

The difficulty problem

Original IPC instances are often too easy for modern planners; naively harder ones may favor recent approaches.

Autoscale (Torralba et al., ICAPS 2021)

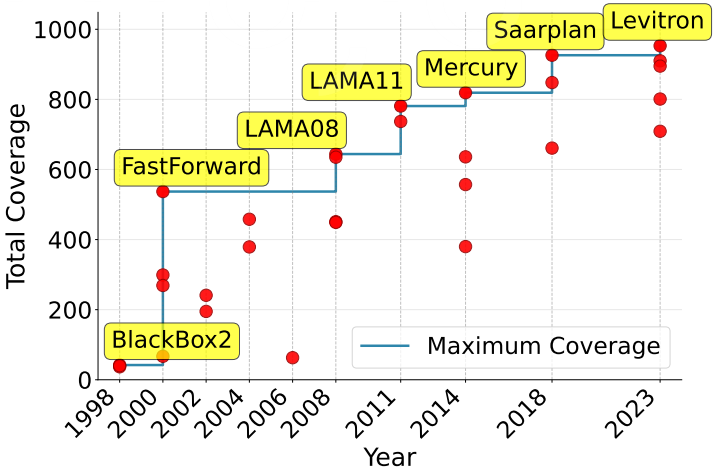
Many IPC domains ship with **parameterized instance generators**.
Autoscale automatically tunes their parameters so that:

- the instances are in in a **calibrated difficulty range**: challenging but not out of reach for state-of-the-art planners.
- every domain contributes the **same number of instances**

29 planners across 25 years of the IPC

IPC	Selected planners (paradigm)
1998	BlackBox2 (SAT), HSP (heuristic search), IPP (Graphplan)
2000	**Fast Forward**, HSP2, MIPS (model checking), System R (Prolog)
2002	LPG (local search), SimPlanner
2004	Fast Downward, Fast Diagonally Downward
2006	MIPS-XXL (external memory)
2008	C3, FFSA, **LAMA08** (landmarks)
2011	FDSS11 (portfolio), LAMA11
2014	Probe, Mercury (red-black), Madagascar (SAT), YAHSP3
2018	FD-Remix, LAPKT-BFWS (width search), Saarplan
2023	ANS, DecStar (decoupled), FDSS23, Maidu, Levitron

The big picture: recent planners win



But some old planners still shine

Coverage per IPC year (full table in the paper)

Highest coverage achieved by most recent planners. Yet there are some exceptions on specific domains:

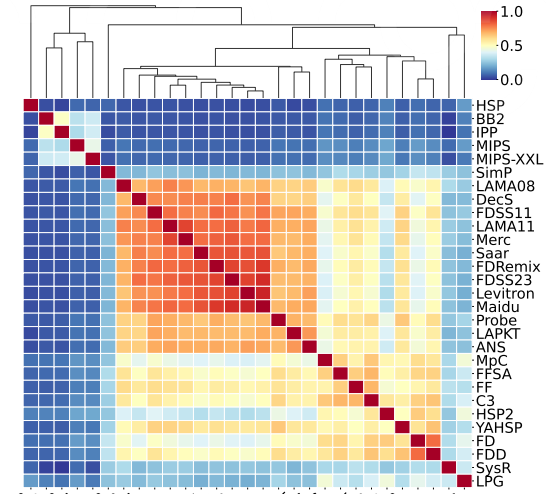
- **Fast Forward** (2000) outperforms some 2023 planners on 2008 domains.
- **System R** (2000) solves **all 30** Scanalyzer instances (a 2008 domain), while almost no other planner exceeds 20.
- **SimPlanner** (2002) attains the best coverage on Satellite, solving $\sim$**2**$\times$ as many instances as recent planners.

First-place rankings, year by year

Year	2000			2002		2004		2008			2011		2014				2018			2023				
	FF	HSP2	SysR	LPG	SimP	FD	FDD	C3	FFSA	LAMA08	FDSS11	LAMA11	Probe	Merc	MpC	YAHSP	FDRemix	LAPKT	Saar	ANS	DecS	FDSS23	Levitron	Maidu
2000	29	10	10																					
2002	23	8	9	2	8																			
2004	18	4	9	2	6	7	15																	
2006	18	4	9	2	6	7	15																	
2008	10	4	8	1	4	5	11	9	4	17														
2011	8	3	8	1	3	4	6	6	4	8	16	18												
2014	6	3	8		3	4	5	6	4	5	13	13	9	18	5	7								
2018	6	3	8		2	4	5	6	4	4	7	9	7	14	5	7	13	10	19					
2023	5	3	8		2	4	5	6	4	4	6	9	6	13	5	6	12	8	17	11	11	14	14	13

#domains where each planner takes first place, counting only planners available up to that year.
(darker = more wins)

How similar are the planners?



Similarity = **Jaccard index** on the set of instances each planner solves:

- Based on *behavior* only, not on the algorithms.
- Clusters emerge: early planners apart, the modern heuristic-search family together.
- Low-similarity pairs solve different instances.

Take-aways

The Planner Museum

- **29 historical IPC planners**, patched, containerized to run on modern hardware, and made public.
- A **fair, cross-era evaluation** using coverage on rescaled benchmarks spanning the IPC's full history.

What we learned

- **Yes, we made progress**: recent planners are best overall, and improvement has been steady since 2008.
- But some **old planners remain state-of-the-art** on specific domains.

- Some planners could not be found anywhere! (Especially for those from before IPC 2008).
- Some older planners couldn't be made to run at all.

The museum is open, and accepting donations!

Visit the museum!

- Free entrance for ICAPS attendees

- Additional visualizations

mrlab.ai/planner-museum.html

Did your planner compete at the IPC?

Help us preserve it: **submit it as a pull request!**

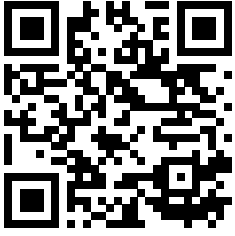
- Add the source code and an Apptainer recipe.
- We provide a uniform interface and template.

github.com/mrlab-ai/ planner-museum

Thank you!

Questions?

The museum is open — code, containers and benchmarks are online.



`mrlab.ai/planner-museum.html`

Backup — Coverage per IPC year

Year	1998			2000				2002		2004		2006	2008			2011		2014				2018			2023				
	BB2	HSP	IPP	FF	HSP2	MIPS	SysR	LPG	SimP	FD	FDD	MIPSXXL	C3	FFSA	LAMA08	FDSS11	LAMA11	Probe	Merc	MpC	YAHSP	FDRemix	LAPKT	Saar	ANS	DecS	FDSS23	Levitron	Maidu
1998	3	5	2	68	41	14	30	38	5	57	58	0	48	61	61	73	63	53	66	48	55	74	52	66	57	84	**89**	82	84
2000	4	2	4	62	57	2	61	40	21	49	62	7	63	50	70	76	79	75	79	45	79	80	64	78	69	72	84	**85**	**85**
2002	2	9	4	95	33	13	62	47	100	79	83	1	61	71	105	89	102	80	109	64	79	101	89	115	87	103	101	**116**	102
2004	1	0	1	37	37	0	0	7	46	61	77	1	46	43	26	77	66	71	65	36	76	75	84	82	83	67	77	**88**	83
2006	0	0	0	40	10	2	60	2	4	16	27	4	29	23	52	39	62	35	64	23	33	59	51	**78**	41	57	67	69	65
2008	23	21	26	119	90	33	42	47	38	90	100	46	121	99	118	125	130	110	130	91	105	124	93	137	113	116	133	**145**	134
2011	3	0	2	48	19	2	11	6	13	23	28	1	41	57	101	89	122	107	143	50	83	140	94	**149**	93	103	141	144	134
2014	6	0	0	15	12	1	3	8	13	4	23	2	19	17	32	86	67	48	64	5	38	99	74	113	69	107	**114**	108	107
2018	0	0	1	53	0	0	0	0	1	0	0	1	23	28	79	83	90	57	99	18	9	96	60	108	97	92	104	**116**	101
Total	**42**	**37**	**40**	**537**	**299**	**67**	**269**	**195**	**241**	**379**	**458**	**63**	**451**	**449**	**644**	**737**	**781**	**636**	**819**	**380**	**557**	**848**	**661**	**926**	**709**	**801**	**910**	**953**	**895**