Automatic Instance Generation for Classical Planning

Álvaro Torralba,[1] Jendrik Seipp,[2,3] Silvan Sievers[3]

[1]Aalborg University, Denmark
[2]Linköping University, Sweden
[3]University of Basel, Switzerland

Empirical Evaluation – The ICAPS/IPC Way

The ICAPS/IPC Way

- Measure coverage
- Time limit 5 or 30 minutes
- Memory limit 2-8 GB
- Use the benchmarks from the International Planning Competition

		Blind Search										A* with $h^{\text{LM-cut}}$									
		Dominance Pruning						Duplicate Checking				Dominance Pruning						Duplicate Checking			
Domain	#	$\prec_B$	$\prec_B^{IT}$	$\prec_B^{gIT}$	$\prec_G^{IT}$	$\prec_G^{g}$	$\prec_G^{gIT}$	$\prec_D$	$\prec_D^{I}$	$\prec_D^{g}$	$\prec_D^{gI}$	$\prec_B$	$\prec_B^{IT}$	$\prec_B^{gIT}$	$\prec_G^{IT}$	$\prec_G^{g}$	$\prec_G^{gIT}$	$\prec_D$	$\prec_D^{I}$	$\prec_D^{g}$	$\prec_D^{gI}$
DataNet	20	**9**	**9**	5	**9**	**9**	**9**	5	5	5	5	**14**	**14**	12	**14**	**14**	**14**	12	12	12	12
Depots	22	3	3	**4**	**4**	**4**	**4**	2	2	**4**	**4**	**7**	**7**	**7**	**7**	**7**	**7**	5	5	**7**	**7**
Driverlog	20	**11**	**11**	**11**	**11**	**11**	**11**	9	9	10	10	13	13	13	13	13	13	13	13	13	13
Elevators	30	6	6	9	12	**16**	**16**	0	0	10	10	10	11	22	13	**23**	**23**	0	0	22	22
Floortile	40	**2**	**2**	**2**	**2**	**2**	**2**	0	0	0	0	**10**	**10**	**10**	**10**	**10**	**10**	5	5	5	5
Freecell	42	0	0	0	0	0	0	0	0	**2**	**2**	1	1	**2**	**2**	**2**	**2**	1	1	**2**	**2**
GED	20	13	13	**15**	**15**	**15**	**15**	7	7	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	13	13	**15**	**15**
Grid	5	1	1	1	1	1	1	1	1	1	1	**2**	**2**	**2**	**2**	**2**	**2**	1	1	**2**	**2**
Logistics	63	24	25	25	**26**	25	**26**	22	22	24	24	34	34	35	34	**36**	**36**	30	30	34	34
Miconic	145	46	**47**	**47**	**47**	45	**47**	42	42	42	42	135	135	135	135	135	135	135	135	135	135
NoMystery	20	**20**	**20**	**20**	**20**	19	**20**	16	16	16	16	**20**	**20**	**20**	**20**	**20**	**20**	19	19	19	19
OpenSt14	20	1	**2**	1	**2**	**2**	1	**2**	**2**	**2**	**2**	**2**	**2**	1	**2**	**2**	1	**2**	**2**	**2**	**2**
PSR	48	**48**	**48**	46	**48**	**48**	**48**	42	42	46	46	**48**	**48**	47	**48**	**48**	**48**	45	45	47	47
Rovers	40	**7**	**7**	**7**	**7**	**7**	**7**	6	6	**7**	**7**	8	8	8	8	8	8	8	8	8	8
Satellite	36	5	5	5	5	5	5	5	5	5	5	7	7	**9**	7	**9**	**9**	7	7	**9**	**9**
Tidybot14	10	3	3	3	3	3	3	3	3	3	3	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	5	**6**
Transport	28	10	11	13	14	**15**	**15**	0	0	13	13	12	12	**14**	13	**14**	**14**	6	6	**14**	**14**
Woodwork	26	7	7	7	7	7	7	7	7	7	7	16	16	16	**17**	**17**	**17**	16	16	16	16
Zenotravel	20	8	**9**	8	**9**	**9**	**9**	6	6	6	7	12	12	12	**13**	**13**	**13**	8	8	11	11
Others	239	67	67	67	67	67	67	67	67	67	67	87	87	87	87	87	87	87	87	87	87
$\sum$	894	291	296	296	309	310	**313**	242	242	285	286	459	460	473	466	**481**	480	419	419	465	466

Empirical Evaluation – The ICAPS/IPC Way

The ICAPS/IPC Way

- Measure coverage
- Time limit 5 or 30 minutes
- Memory limit 2-8 GB
- Use the benchmarks from the International Planning Competition

		Blind Search										A* with $h^{\text{LM-cut}}$									
		Dominance Pruning						Duplicate Checking				Dominance Pruning						Duplicate Checking			
Domain	#	$\prec_B$	$\prec_B^{IT}$	$\prec_B^{gIT}$	$\prec_G^{IT}$	$\prec_G^{g}$	$\prec_G^{gIT}$	$\prec_D$	$\prec_D^{I}$	$\prec_D^{g}$	$\prec_D^{gI}$	$\prec_B$	$\prec_B^{IT}$	$\prec_B^{gIT}$	$\prec_G^{IT}$	$\prec_G^{g}$	$\prec_G^{gIT}$	$\prec_D$	$\prec_D^{I}$	$\prec_D^{g}$	$\prec_D^{gI}$
DataNet	20	**9**	**9**	5	**9**	**9**	**9**	5	5	5	5	**14**	**14**	12	**14**	**14**	**14**	12	12	12	12
Depots	22	3	3	**4**	**4**	**4**	**4**	2	2	**4**	**4**	**7**	**7**	**7**	**7**	**7**	**7**	5	5	**7**	**7**
Driverlog	20	**11**	**11**	**11**	**11**	**11**	**11**	9	9	10	10	13	13	13	13	13	13	13	13	13	13
Elevators	30	6	6	9	12	**16**	**16**	0	0	10	10	10	11	22	13	**23**	**23**	0	0	22	22
Floortile	40	**2**	**2**	**2**	**2**	**2**	**2**	0	0	0	0	**10**	**10**	**10**	**10**	**10**	**10**	5	5	5	5
Freecell	42	0	0	0	0	0	0	0	0	**2**	**2**	1	1	**2**	**2**	**2**	**2**	1	1	**2**	**2**
GED	20	13	13	**15**	**15**	**15**	**15**	7	7	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	13	13	**15**	**15**
Grid	5	1	1	1	1	1	1	1	1	1	1	**2**	**2**	**2**	**2**	**2**	**2**	1	1	**2**	**2**
Logistics	63	24	25	25	**26**	25	**26**	22	22	24	24	34	34	35	34	**36**	**36**	30	30	34	34
Miconic	145	46	**47**	**47**	**47**	45	**47**	42	42	42	42	135	135	135	135	135	135	135	135	135	135
NoMystery	20	**20**	**20**	**20**	**20**	19	**20**	16	16	16	16	**20**	**20**	**20**	**20**	**20**	**20**	19	19	19	19
OpenSt14	20	1	**2**	1	**2**	**2**	1	**2**	**2**	**2**	**2**	**2**	**2**	1	**2**	**2**	1	**2**	**2**	**2**	**2**
PSR	48	**48**	**48**	46	**48**	**48**	**48**	42	42	46	46	**48**	**48**	47	**48**	**48**	**48**	45	45	47	47
Rovers	40	**7**	**7**	**7**	**7**	**7**	**7**	6	6	**7**	**7**	8	8	8	8	8	8	8	8	8	8
Satellite	36	5	5	5	5	5	5	5	5	5	5	7	7	**9**	7	**9**	**9**	7	7	**9**	**9**
Tidybot14	10	3	3	3	3	3	3	3	3	3	3	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	5	**6**
Transport	28	10	11	13	14	**15**	**15**	0	0	13	13	12	12	**14**	13	**14**	**14**	6	6	**14**	**14**
Woodwork	26	7	7	7	7	7	7	7	7	7	7	16	16	16	**17**	**17**	**17**	16	16	16	16
Zenotravel	20	8	**9**	8	**9**	**9**	**9**	6	6	6	7	12	12	12	**13**	**13**	**13**	8	8	11	11
Others	239	67	67	67	67	67	67	67	67	67	67	87	87	87	87	87	87	87	87	87	87
$\sum$	894	291	296	296	309	310	**313**	242	242	285	286	459	460	473	466	**481**	480	419	419	465	466

Having a standard evaluation setting is generally beneficial:

- Reproducibility
- Interpretability
- Avoids hand picking results

Empirical Evaluation – The ICAPS/IPC Way

The ICAPS/IPC Way

- Measure coverage
- Time limit 5 or 30 minutes
- Memory limit 2-8 GB
- ~~Use the benchmarks from the International Planning Competition~~
- Use the Autoscale benchmark set

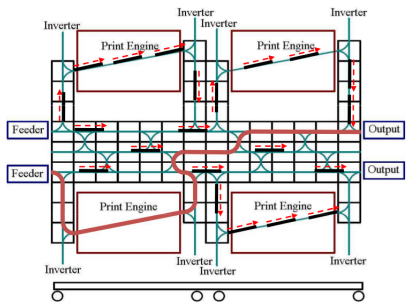
		Blind Search										A* with $h^{\text{LM-cut}}$									
		Dominance Pruning						Duplicate Checking				Dominance Pruning						Duplicate Checking			
Domain	#	$\prec_B$	$\prec_B^{IT}$	$\prec_B^{gIT}$	$\prec_G^{IT}$	$\prec_G^{g}$	$\prec_G^{gIT}$	$\prec_D$	$\prec_D^{I}$	$\prec_D^{g}$	$\prec_D^{gI}$	$\prec_B$	$\prec_B^{IT}$	$\prec_B^{gIT}$	$\prec_G^{IT}$	$\prec_G^{g}$	$\prec_G^{gIT}$	$\prec_D$	$\prec_D^{I}$	$\prec_D^{g}$	$\prec_D^{gI}$
DataNet	20	**9**	**9**	5	**9**	**9**	**9**	5	5	5	5	**14**	**14**	12	**14**	**14**	**14**	12	12	12	12
Depots	22	3	3	**4**	**4**	**4**	**4**	2	2	**4**	**4**	**7**	**7**	**7**	**7**	**7**	**7**	5	5	**7**	**7**
Driverlog	20	**11**	**11**	**11**	**11**	**11**	**11**	9	9	10	10	13	13	13	13	13	13	13	13	13	13
Elevators	30	6	6	9	12	**16**	**16**	0	0	10	10	10	11	22	13	**23**	**23**	0	0	22	22
Floortile	40	**2**	**2**	**2**	**2**	**2**	**2**	0	0	0	0	**10**	**10**	**10**	**10**	**10**	**10**	5	5	5	5
Freecell	42	0	0	0	0	0	0	0	0	**2**	**2**	1	1	**2**	**2**	**2**	**2**	1	1	**2**	**2**
GED	20	13	13	**15**	**15**	**15**	**15**	7	7	**15**	**15**	**15**	**15**	**15**	**15**	**15**	**15**	13	13	**15**	**15**
Grid	5	1	1	1	1	1	1	1	1	1	1	**2**	**2**	**2**	**2**	**2**	**2**	1	1	**2**	**2**
Logistics	63	24	25	25	**26**	25	**26**	22	22	24	24	34	34	35	34	**36**	**36**	30	30	34	34
Miconic	145	46	**47**	**47**	**47**	45	**47**	42	42	42	42	135	135	135	135	135	135	135	135	135	135
NoMystery	20	**20**	**20**	**20**	**20**	19	**20**	16	16	16	16	**20**	**20**	**20**	**20**	**20**	**20**	19	19	19	19
OpenSt14	20	1	**2**	1	**2**	**2**	1	**2**	**2**	**2**	**2**	**2**	**2**	1	**2**	**2**	1	**2**	**2**	**2**	**2**
PSR	48	**48**	**48**	46	**48**	**48**	**48**	42	42	46	46	**48**	**48**	47	**48**	**48**	**48**	45	45	47	47
Rovers	40	**7**	**7**	**7**	**7**	**7**	**7**	6	6	**7**	**7**	8	8	8	8	8	8	8	8	8	8
Satellite	36	5	5	5	5	5	5	5	5	5	5	7	7	**9**	7	**9**	**9**	7	7	**9**	**9**
Tidybot14	10	3	3	3	3	3	3	3	3	3	3	**6**	**6**	**6**	**6**	**6**	**6**	**6**	**6**	5	**6**
Transport	28	10	11	13	14	**15**	**15**	0	0	13	13	12	12	**14**	13	**14**	**14**	6	6	**14**	**14**
Woodwork	26	7	7	7	7	7	7	7	7	7	7	16	16	16	**17**	**17**	**17**	16	16	16	16
Zenotravel	20	8	**9**	8	**9**	**9**	**9**	6	6	6	7	12	12	12	**13**	**13**	**13**	8	8	11	11
Others	239	67	67	67	67	67	67	67	67	67	67	87	87	87	87	87	87	87	87	87	87
$\sum$	894	291	296	296	309	310	**313**	242	242	285	286	459	460	473	466	**481**	480	419	419	465	466

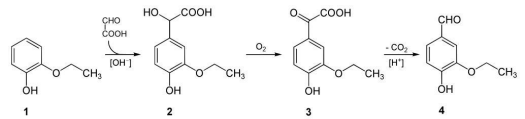
Having a standard evaluation setting is generally beneficial:

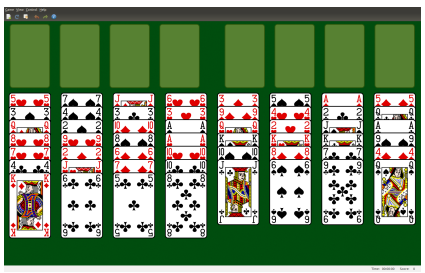
- Reproducibility
- Interpretability
- Avoids hand picking results

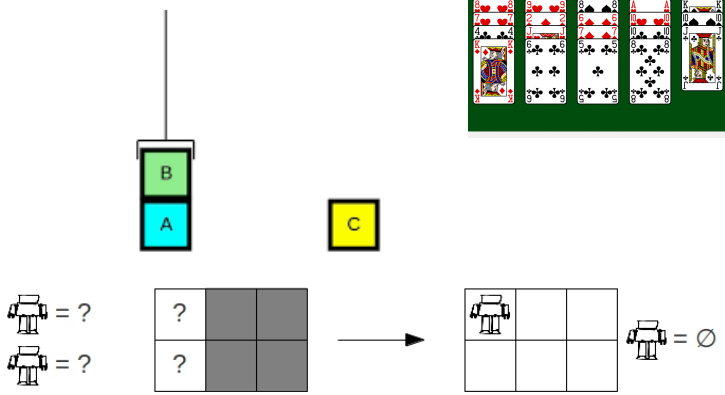
The IPC Benchmark Set

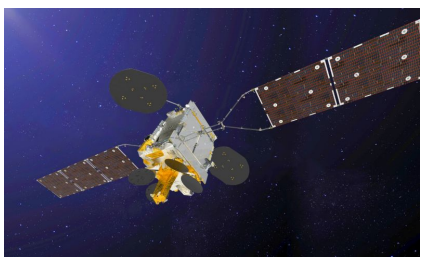
A collection made in 9 editions of the IPC: from IPC'1998 until IPC'2018
(Since 2008: separated instances for Optimal and Satisficing planning)











Thank you to all IPC organizers and everyone who contributed domains!

So, What's Wrong with the IPC Benchmark Set?

		IPC		
	#	L	D	O
Grid	5	**5**	**5**	**5**
Driverlog	20	**20**	**20**	**20**
Rovers	40	**40**	**40**	**40**
Snake	20	5	**15**	12
Total	85	70	**80**	77

Table: Coverage of three planners: L, D, and O.

So, What's Wrong with the IPC Benchmark Set?

		IPC		
	#	L	D	O
Grid	5	**5**	**5**	**5**
Driverlog	20	**20**	**20**	**20**
Rovers	40	**40**	**40**	**40**
Snake	20	5	**15**	12
Total	85	70	**80**	77

Table: Coverage of three planners: L, D, and O.

- Different number of instances per domain

So, What's Wrong with the IPC Benchmark Set?

		IPC		
	#	L	D	O
Grid	5	**5**	**5**	**5**
Driverlog	20	**20**	**20**	**20**
Rovers	40	**40**	**40**	**40**
Snake	20	5	**15**	12
Total	85	70	**80**	77

Table: Coverage of three planners: L, D, and O.

- Different number of instances per domain
- Instance scaling: Experiments on some domains of the IPC benchmark set may not observe any difference between planners even if it exists!

So, What's Wrong with the IPC Benchmark Set?

	IPC				Autoscale			
	#	L	D	O	#	L	D	O
Grid	5	**5**	**5**	**5**	30	**17**	14	16
Driverlog	20	**20**	**20**	**20**	30	15	10	**25**
Rovers	40	**40**	**40**	**40**	30	**30**	23	28
Snake	20	5	**15**	12	30	6	**19**	16
Total	85	70	**80**	77	120	68	66	**85**

Table: Coverage of three planners: L, D, and O.

- Different number of instances per domain
- Instance scaling: Experiments on some domains of the IPC benchmark set may not observe any difference between planners even if it exists!

Contribution

1. Autoscale: An automated tool to select instances from a given domain

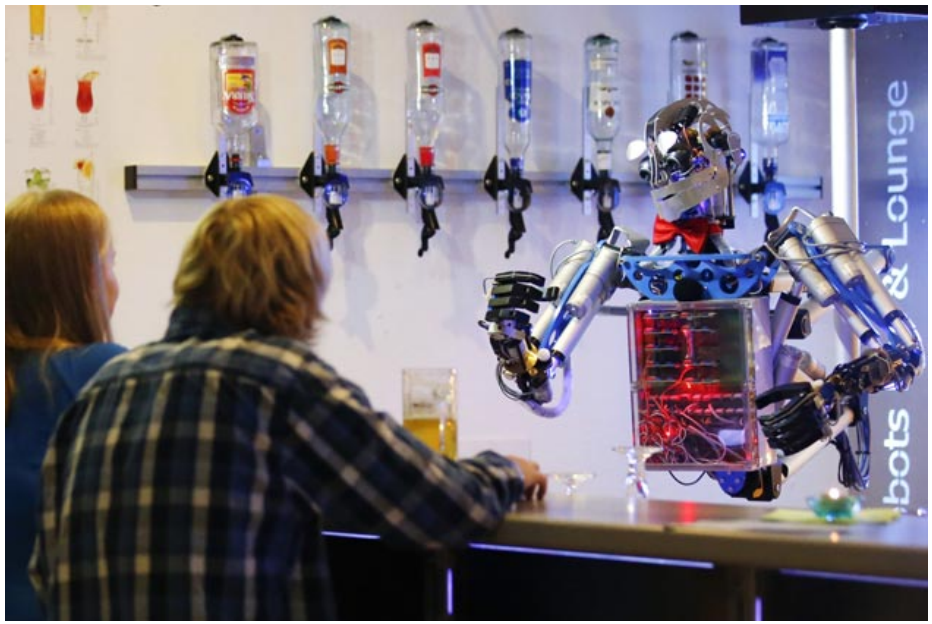
Contribution

1. Autoscale: An automated tool to select instances from a given domain

2. Two new benchmark sets:
   - Autoscale'21 Optimal
   - Autoscale'21 Agile/Satisficing planning

$\rightarrow$Better than the IPC set to evaluate current and future planners!

Example Domain: Barman



Instance Generator

```
./barman-generator.py <num_cocktails> <num_ingredients>
                      <num_shots> [<random_seed>]
  num_cocktails (min 1)
  num_ingredients (min 2)
  num_shots (min num_cocktails+1)
  random_seed (min 1, optional)
```

Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners

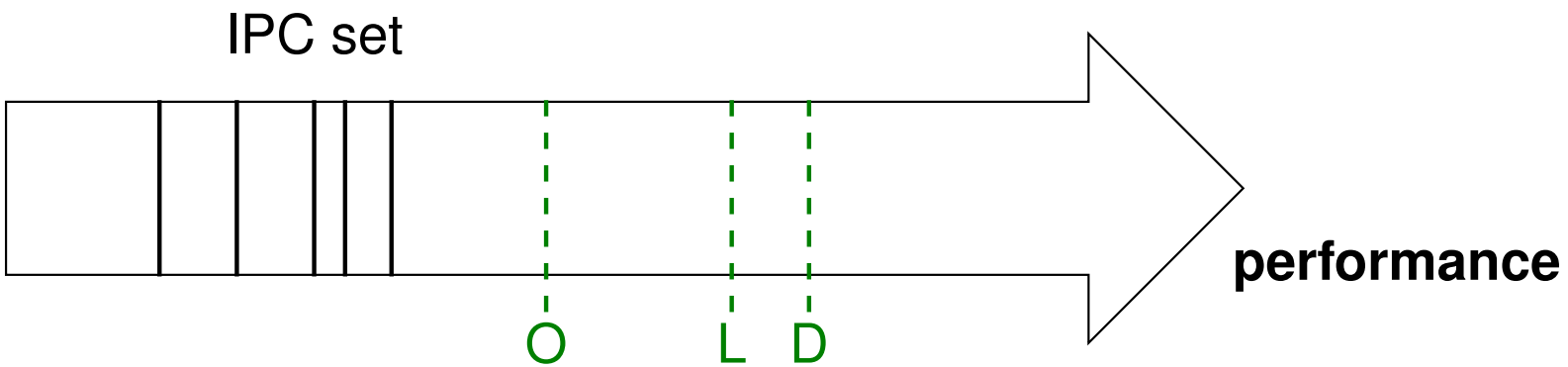
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners



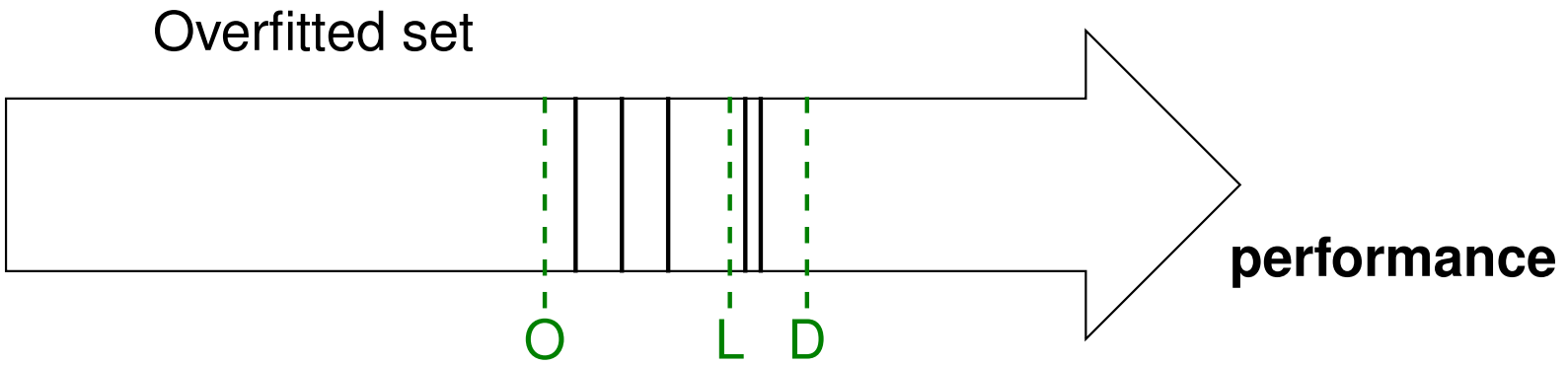
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners
- Avoids bias



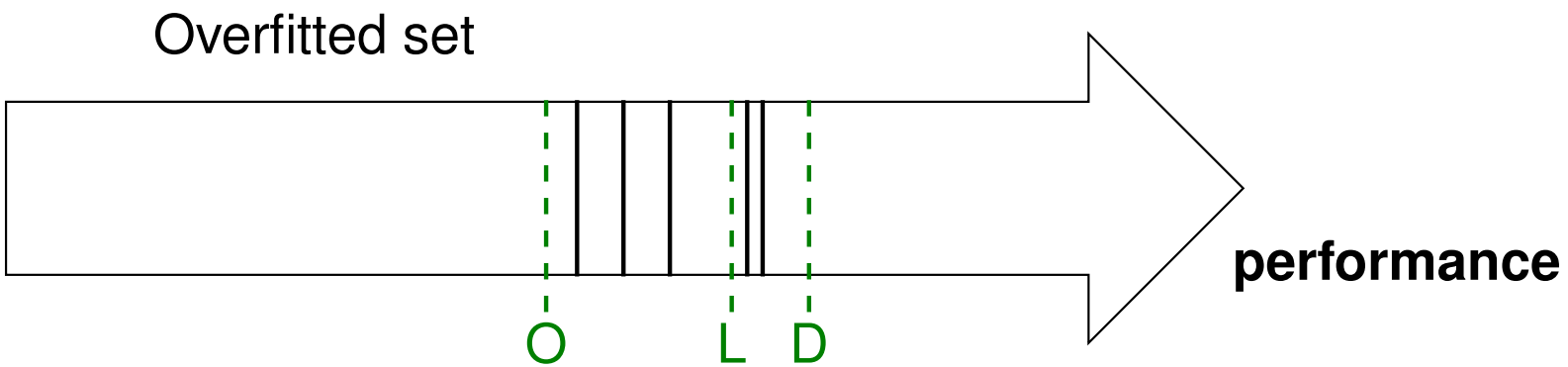
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners
- Avoids bias



Rule 1: Agnostic to Individual Planner Performance

Don't consider the individual results of all planners

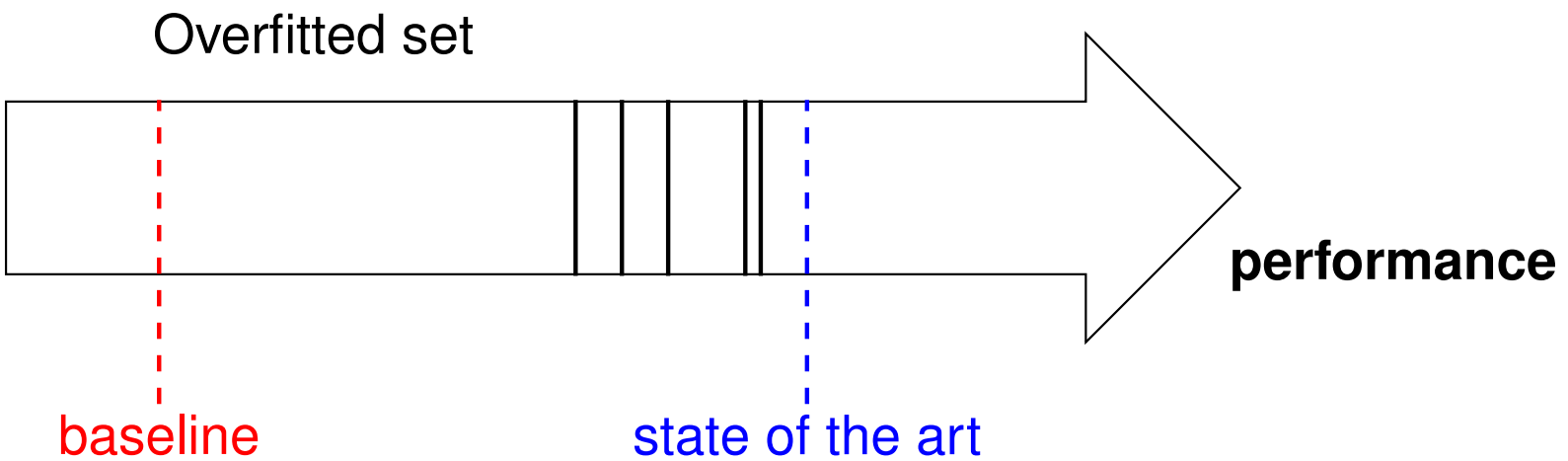
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners
- Avoids bias



Rule 1: Agnostic to Individual Planner Performance

Don't consider the individual results of all planners

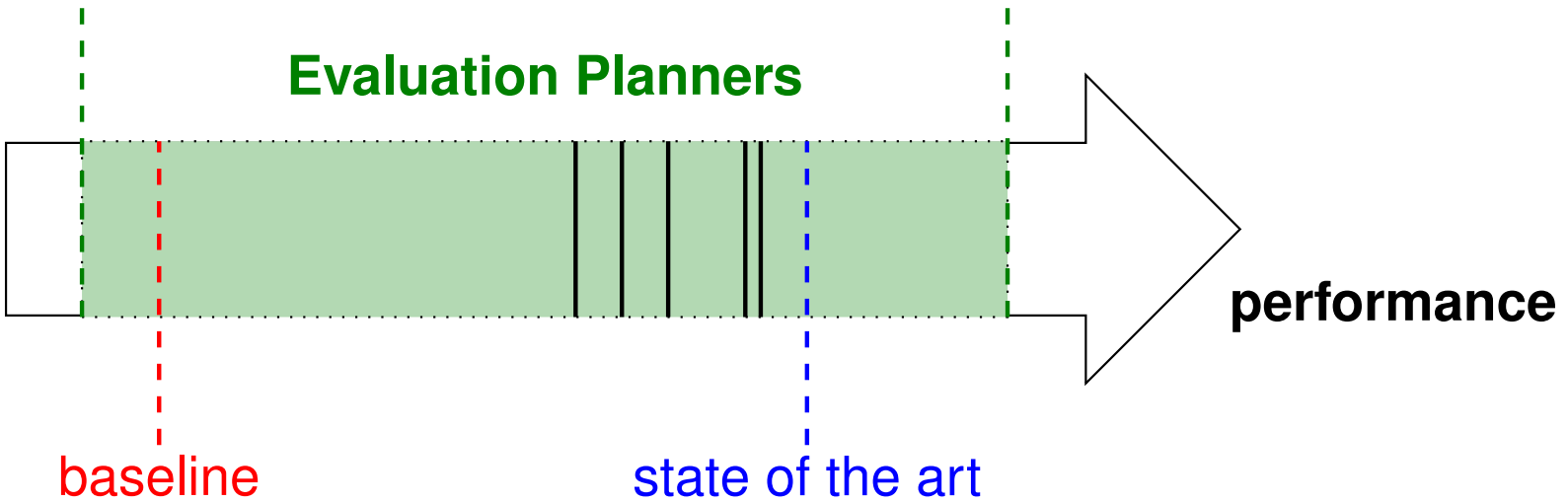
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners
- Avoids bias



Rule 1: Agnostic to Individual Planner Performance

Don't consider the individual results of all planners

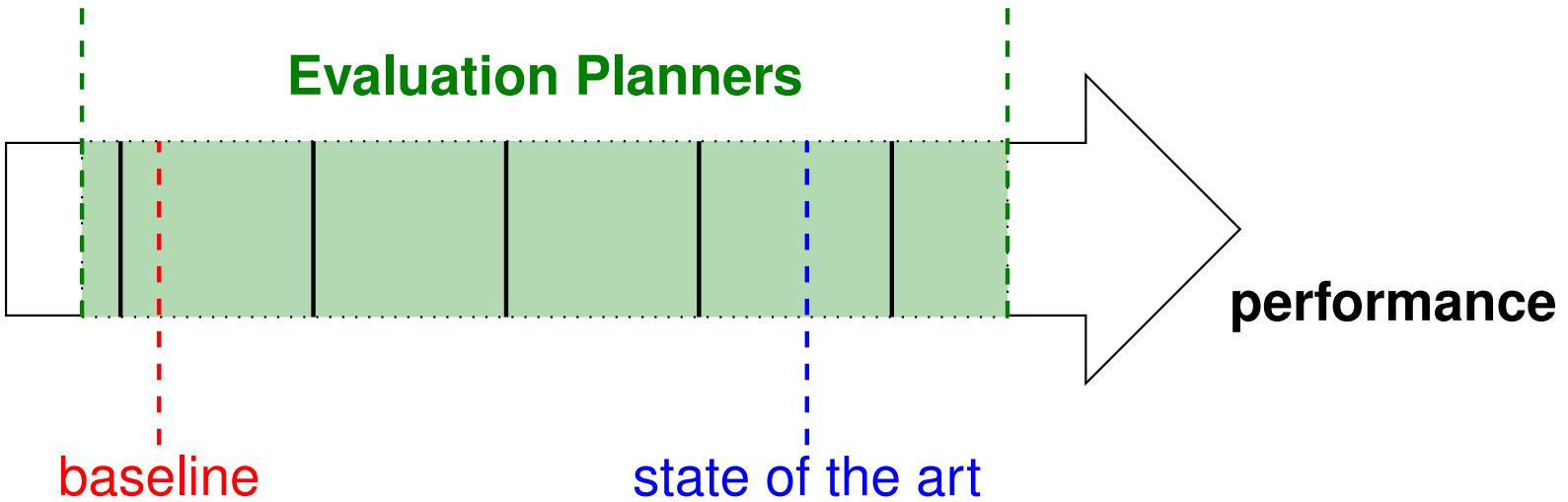
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners
- Avoids bias



Rule 1: Agnostic to Individual Planner Performance

Don't consider the individual results of all planners

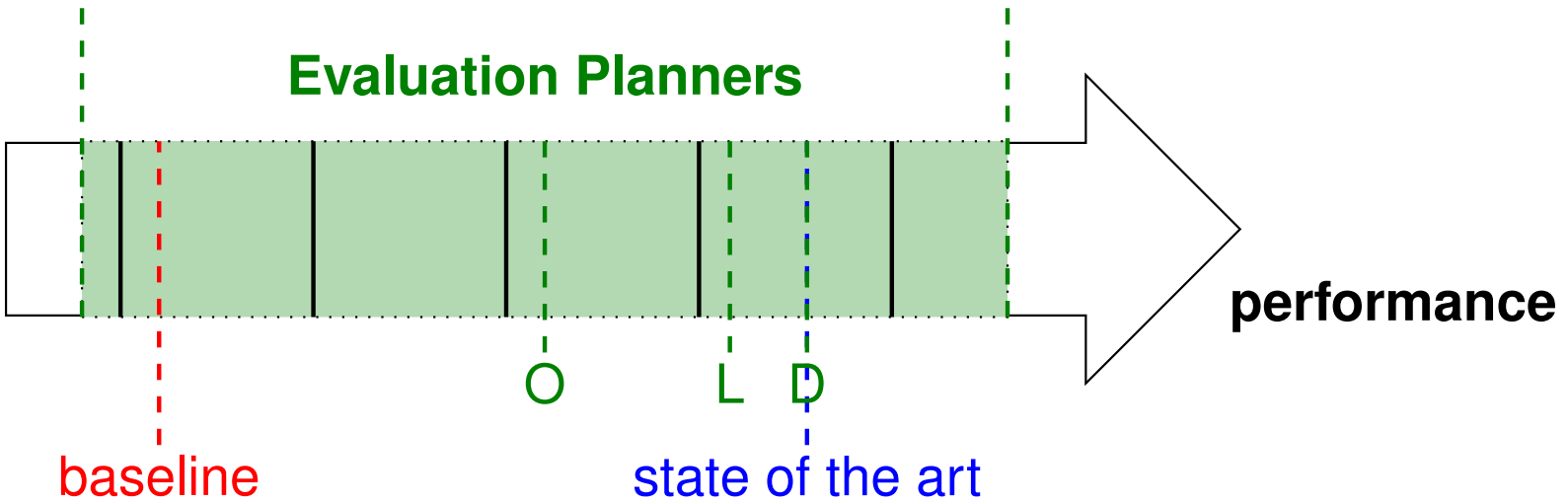
Instance Generation Problem

Input:

- Domain
- Instance generator
- A set of baseline planners
- A set of state-of-the-art planners

Output: Set of instances that:

- Is useful to evaluate current/future planners
- Avoids bias



Rule 1: Agnostic to Individual Planner Performance

Don't consider the individual results of all planners

Useful to Evaluate Planners

Rule 2: Smooth Scaling

The instance set should:

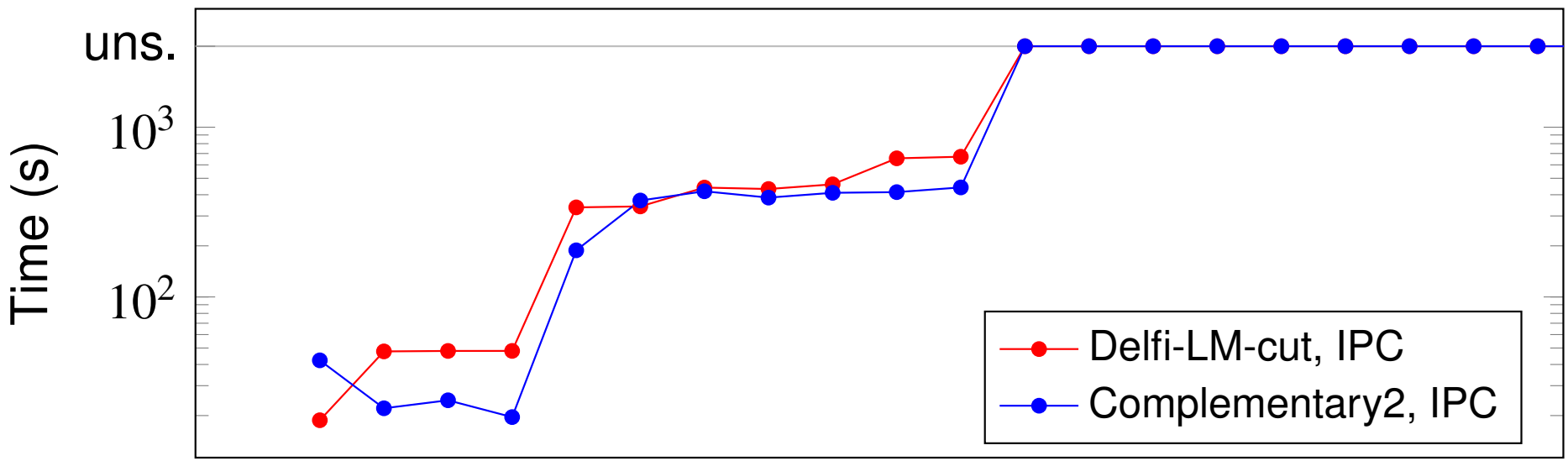
- Have easy instances (solvable by the baseline planners)
- Have hard instances (out of reach for the state of the art)
- Scale smoothly

Useful to Evaluate Planners

Rule 2: Smooth Scaling

The instance set should:

- Have easy instances (solvable by the baseline planners)
- Have hard instances (out of reach for the state of the art)
- Scale smoothly

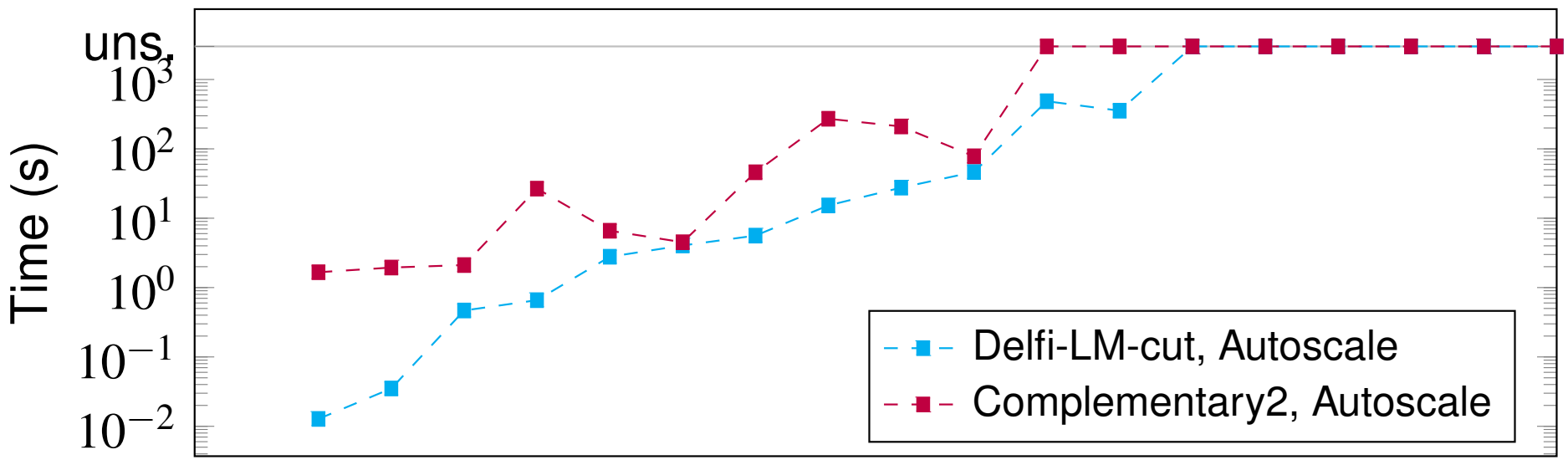


Useful to Evaluate Planners

Rule 2: Smooth Scaling

The instance set should:

- Have easy instances (solvable by the baseline planners)
- Have hard instances (out of reach for the state of the art)
- Scale smoothly



Parameter- and Sequence-Based Instance Selection

Rule 3: Parameter-based Selection

Avoid selecting the random seed

→Select # cocktails, # shots and # ingredients, but not the concrete instance

Parameter- and Sequence-Based Instance Selection

Rule 3: Parameter-based Selection

Avoid selecting the random seed

→Select # cocktails, # shots and # ingredients, but not the concrete instance

Rule 4: Sequence-based Selection

The parameter configurations can be organized in one or more sequences

cocktails	shots	ingredients
$(b = 5,$ $m = 1.34)$	$(b = 1, m = 0,$ +cocktails)	$(v = 3)$
5	6	3
6	7	3
7	8	3
9	10	3
10	11	3
11	12	3
13	14	3

Keep the Spirit of the Domain

Rule 5: User Constraints

The domain designer specifies guidelines on which parameters to scale

cocktails	shots	ingredients
$b \in [1,6]$ $m \in [0.1,5]$	$b \in [1,6]$ $m \in [1,5]$ + cocktails	$v \in \{2,3,4,5\}$

Optimization Process

For each domain we obtain a set of sequences by:

1. Generate candidate sequences that scale smoothly

C	S	I
5	6	3
6	7	3
7	8	3
9	10	3
10	11	3
11	12	3
13	14	3

Optimization Process

For each domain we obtain a set of sequences by:

1. Generate candidate sequences that scale smoothly

C	S	I	time(s)
5	6	3	10.1
6	7	3	25.6
7	8	3	101.7
9	10	3	300
10	11	3	900
11	12	3	2700
13	14	3	8100

Optimization Process

For each domain we obtain a set of sequences by:

1. Generate candidate sequences that scale smoothly

C	S	I	time(s)
5	6	3	10.1
6	7	3	25.6
7	8	3	101.7
9	10	3	300
10	11	3	900
11	12	3	2700
13	14	3	8100

C	S	I	time(s)
1	3	2	1.8
1	4	2	2.2
1	5	2	2.9
1	6	2	4.5
1	7	2	8.3
1	8	2	26
1	9	2	120

C	S	I	time(s)
1	5	4	4.2
1	6	4	21
1	7	4	62
1	8	4	250
2	10	4	990
2	11	4	4000
2	12	4	16000

C	S	I	time(s)
1	3	5	2.8
1	4	5	3.7
1	5	5	6.1
1	6	5	16
1	7	5	62
2	9	5	200
2	10	5	660

$\rightarrow$We use SMAC, a model-based optimization procedure

Optimization Process

For each domain we obtain a set of sequences by:

1. Generate candidate sequences that scale smoothly

C	S	I	time(s)
5	6	3	10.1
6	7	3	25.6
7	8	3	101.7
9	10	3	300
10	11	3	900
11	12	3	2700
13	14	3	8100

C	S	I	time(s)
1	3	2	1.8
1	4	2	2.2
1	5	2	2.9
1	6	2	4.5
1	7	2	8.3
1	8	2	26
1	9	2	120

C	S	I	time(s)
1	5	4	4.2
1	6	4	21
1	7	4	62
1	8	4	250
2	10	4	990
2	11	4	4000
2	12	4	16000

C	S	I	time(s)
1	3	5	2.8
1	4	5	3.7
1	5	5	6.1
1	6	5	16
1	7	5	62
2	9	5	200
2	10	5	660

→We use SMAC, a model-based optimization procedure

2. Choose selected (sub-)sequences to include easy and hard instances

→We use CPLEX to solve a MIP problem

Experiments

Compare our new benchmark sets against the IPC

- 26 domains
- Agile/Satisficing and Optimal track
- Autoscale'14: using 6 planners up to IPC'14
- Evaluation based on 8 planners from IPC'18

Experiments

Compare our new benchmark sets against the IPC

- 26 domains
- Agile/Satisficing and Optimal track
- Autoscale'14: using 6 planners up to IPC'14
- Evaluation based on 8 planners from IPC'18

How to evaluate the quality of a benchmark set?

$\rightarrow$Comparisons: number of pairs (X, Y) of planners, such that coverage(X) $\neq$ coverage(Y)

Results

Out of 28 pairs of planners, in how many Autoscale observed a difference in coverage that is not observed with the IPC set?

Domain	#IPC	OPT	AGL	Domain	#IPC	OPT	AGL
Barman	34/40	+12	+19	Nomystery	20	+10	+4
Blocksworld	35	+6	+26	Openstacks	70	-17	+25
Childsnack	20	+8	+1	Parking	40	-2	+5
Data-Network	20	-2	+2	Rovers	40	-4	+20
Depots	22	0	+25	Satellite	36	+5	+2
Driverlog	20	+5	+25	Scanalyzer	50	0	+8
Elevators	50	-3	+11	Snake	20	-1	0
Floortile	40	-3	+7	Storage	30	+6	+1
Grid	5	+7	+21	TPP	30	+2	+11
Gripper	20	0	+7	Transport	70	-8	+14
Hiking	20	+4	+3	Visitall	40	0	+17
Logistics	63	-3	+4	Woodworking	50	+5	+14
Miconic	150	0	0	Zenotravel	20	+4	+22

Results

Out of 28 pairs of planners, in how many Autoscale observed a difference in coverage that is not observed with the IPC set?

Domain	#IPC	OPT	AGL	Domain	#IPC	OPT	AGL
Barman	34/40	+12	+19	Nomystery	20	+10	+4
Blocksworld	35	+6	+26	Openstacks	70	-17	+25
Childsnack	20	+8	+1	Parking	40	-2	+5
Data-Network	20	-2	+2	Rovers	40	-4	+20
Depots	22	0	+25	Satellite	36	+5	+2
Driverlog	20	+5	+25	Scanalyzer	50	0	+8
Elevators	50	-3	+11	Snake	20	-1	0
Floortile	40	-3	+7	Storage	30	+6	+1
Grid	5	+7	+21	TPP	30	+2	+11
Gripper	20	0	+7	Transport	70	-8	+14
Hiking	20	+4	+3	Visitall	40	0	+17
Logistics	63	-3	+4	Woodworking	50	+5	+14
Miconic	150	0	0	Zenotravel	20	+4	+22

Conclusion

1. Autoscale: New tool to automatically select instances
   - Useful to Evaluate Current Planners
   - Avoid Bias
   - Keep the Spirit of the Domain

2. New benchmark set: Autoscale'21
   - Used IPC'18 planners as state-of-the-art planners

     →Useful for the next years!

     →Afterwards we can use Autoscale to update the benchmark set
   - Includes almost all IPC STRIPS domains!
     →Also domains without an instance generator

Autoscale tool and benchmarks are ready to be used, try them out!

https://github.com/AI-Planning/autoscale

https://github.com/AI-Planning/autoscale-benchmarks