

SMT-LIB Benchmarks

Hans-Jörg Schurr

Dagstuhl Seminar 25441
October 30, 2025



Usage

- Competition
- Starting point for research
- Testing, Validation

Usage

- Competition
- Starting point for research
- Testing, Validation

Submission

- Open submission repository with CI
 - Dolmen for compliance
 - Personal interaction
- Yearly release
 - on Zenodo
 - Github for non-compliance
 - Internal staging repositories

Usage

- Competition
- Starting point for research
- Testing, Validation

Submission

- Open submission repository with CI
 - Dolmen for compliance
 - Personal interaction
- Yearly release
 - on Zenodo
 - Github for non-compliance
 - Internal staging repositories

```
non-incremental/UFNIA/2019-Preiner/  
partial/t3_rw617.smt2
```

```
(set-info :smt-lib-version 2.6)  
(set-logic UFNIA)  
(set-info :source |  
  Generated by: Mathias Preiner  
  Generated on: 2019-03-22  
  Application:  
    Verifying bit-vector rewrite rule candidates.  
  Target solver: CVC4, Z3, Vampire |)  
(set-info :license  
  "https://creativecommons.org/licenses/by/4.0/")  
(set-info :category "crafted")  
(assert [...]) [...]  
(set-info :status unknown)  
(check-sat) (exit)
```

The Big Pile of Benchmarks

- 495 177 SMT-LIB benchmarks
- 34 614 311 queries
- 90 logics
- 287 families
 - From 1 to 43 342 benchmarks
- Old benchmarks lack metadata



First Step: Collect Data

One nifty **database** with all you need to answer these questions!

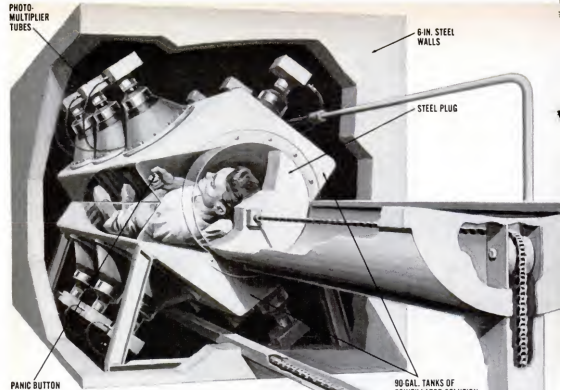
<http://explore.smt-lib.org>

- Download a single SQLite file and
 - use your favorite language to perform queries,
 - or just use raw SQL.
- Around 5.5 GiB of juicy data
- Benchmark and query metadata!
- Results from 20 years of SMT-COMP!

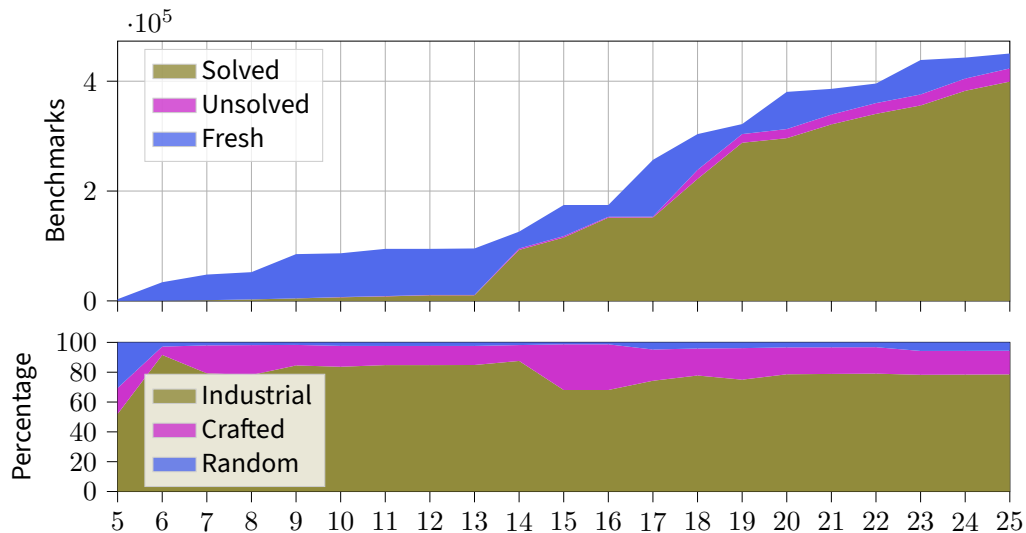


Part I

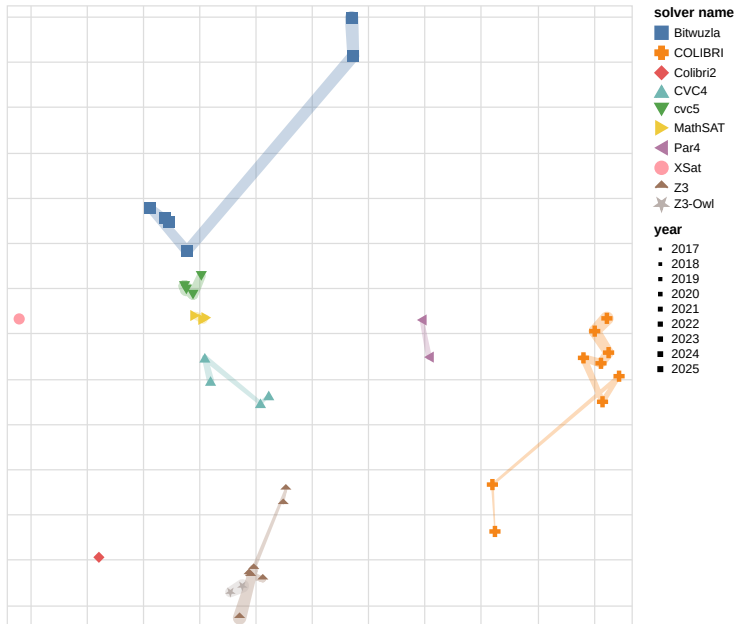
Examples



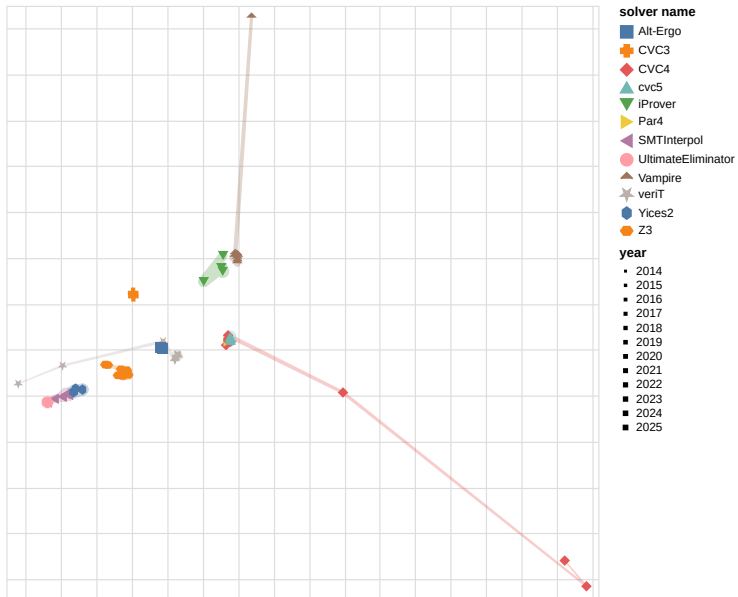
New Non-Incremental Benchmarks



Isomap: QF_FP



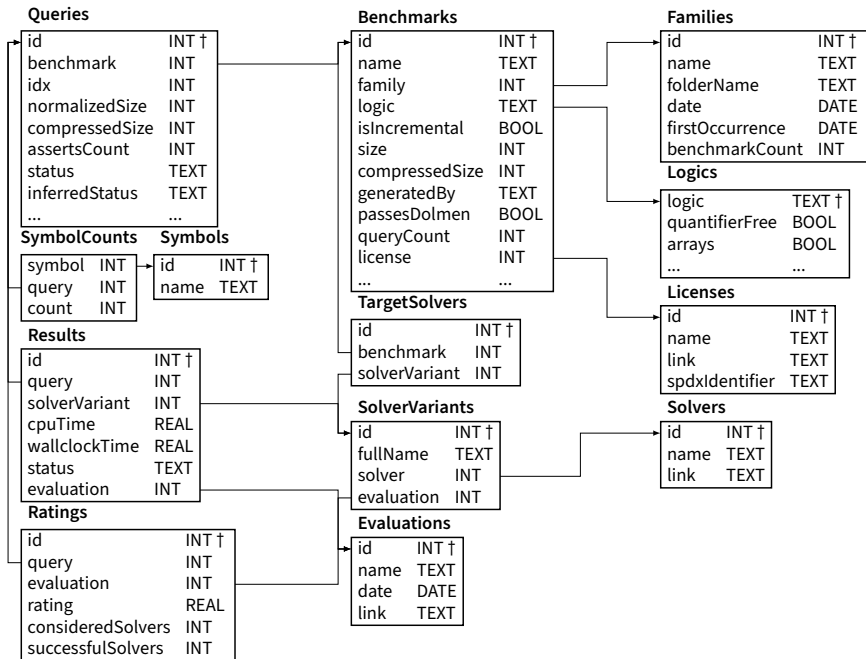
Isomap: UF



Part II

Database Structure





Queries

id	INT ↑
benchmark	INT
idx	INT
normalizedSize	INT
compressedSize	INT
assertsCount	INT
status	TEXT
inferredStatus	TEXT
...	...

SymbolCounts

symbol	INT
query	INT
count	INT

Symbols

id	INT ↑
name	TEXT

Results

id	INT ↑
query	INT
solverVariant	INT
cpuTime	REAL
wallclockTime	REAL
status	TEXT
evaluation	INT

Ratings

id	INT ↑
query	INT
evaluation	INT
rating	REAL
consideredSolvers	INT
successfulSolvers	INT

Benchmarks

id	INT ↑
name	TEXT
family	INT
logic	TEXT
isIncremental	BOOL
size	INT
compressedSize	INT
generatedBy	TEXT
passesDolmen	BOOL
queryCount	INT
license	INT
...	...

TargetSolvers

id	INT ↑
benchmark	INT
solverVariant	INT

SolverVariants

id	INT ↑
fullName	TEXT
solver	INT
evaluation	INT

Evaluations

id	INT ↑
name	TEXT
date	DATE
link	TEXT

Families

id	INT ↑
name	TEXT
folderName	TEXT
date	DATE
firstOccurrence	DATE
benchmarkCount	INT

Logics

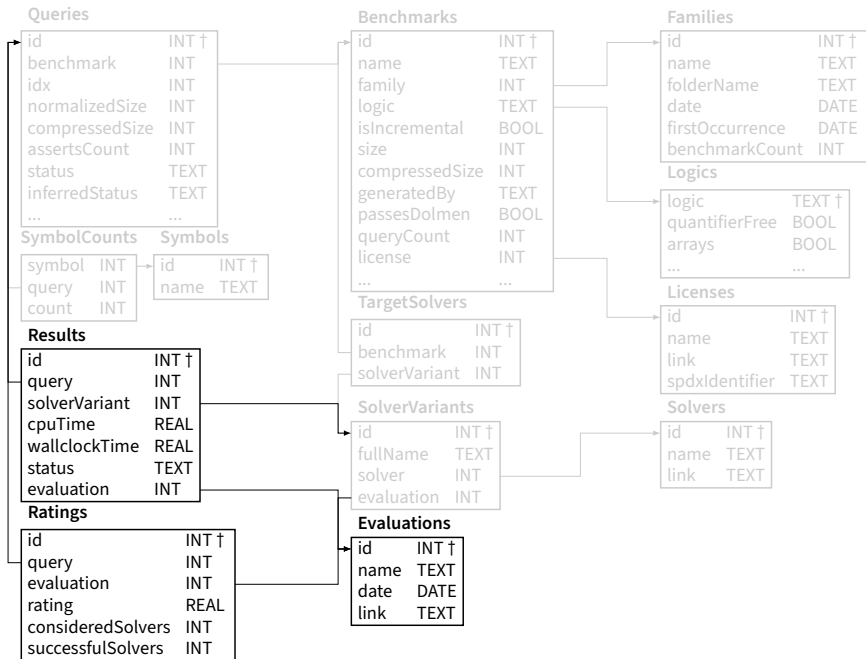
logic	TEXT ↑
quantifierFree	BOOL
arrays	BOOL
...	...

Licenses

id	INT ↑
name	TEXT
link	TEXT
spdxIdentifier	TEXT

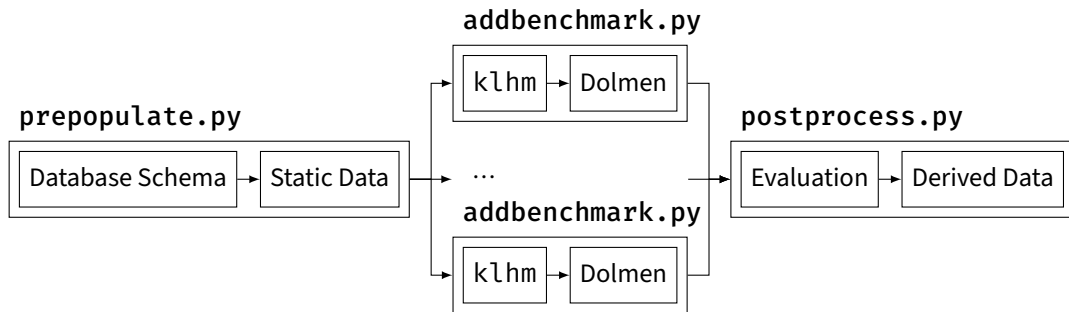
Solvers

id	INT ↑
name	TEXT
link	TEXT



How We Build The Catalog

Pipeline implemented as a set of Python scripts. Hand-curated data is pure Python.



Klammerhammer

- Standalone tool
- Produces JSON
 - one object per query
- Handles push/pop

Implementation

- Written in Zig
- Hardcoded symbol lists
- One linear scan
- Compression using **zstd** C API (quadratic!)

We collect results from **all** SMT-COMPs for the **current** benchmarks.

2005, 2006

primeval competitions – raw HTML webpages

2007–2012

SMTexec. Results were lost, we found the MySQL database.

2013

SMT-EVAL. Excellent documentation. CSV.

2014–2023

StarExec. CSV until 2016, then JSON. Some larger changes to the benchmark set.

2024–

SoSy Benchcloud. same JSON file. Incremental tracks!

$$1 - \frac{\# \text{ solvers with a successful variant}}{\# \text{ solvers with a variant that attempted the query}}$$

Caveats

- Differences between competitions
 - Hardware
 - Time limit
 - Memory limit
- Small number of solvers

Differences to TPTP

TPTP omits superseded solvers

- Intention: ignore experimental variants
- We track variants explicitly!
- Every solver represents serious work.

Conclusion

Releases

- Yearly release with benchmarks
 - Schema might change
 - Will include new SMT-COMP results
- We want you to play around with it
- and welcome feature requests!

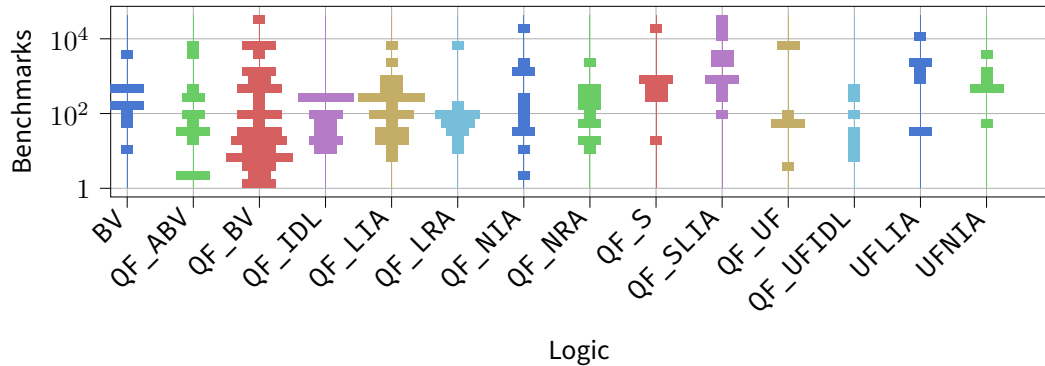
<https://zenodo.org/records/16290040>

Fun Facts

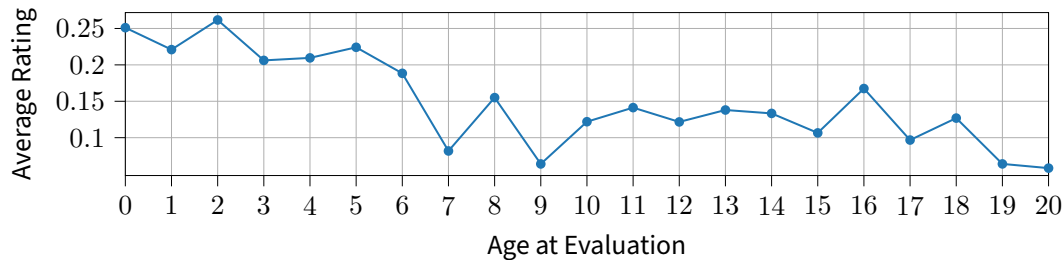
- Most queries: 2 630 828
- Largest query: 1.9 GB
(48.20 MB compressed)
- Deepest term: 3 515 188
- Best solver name:
Tiffany de Wintermonte



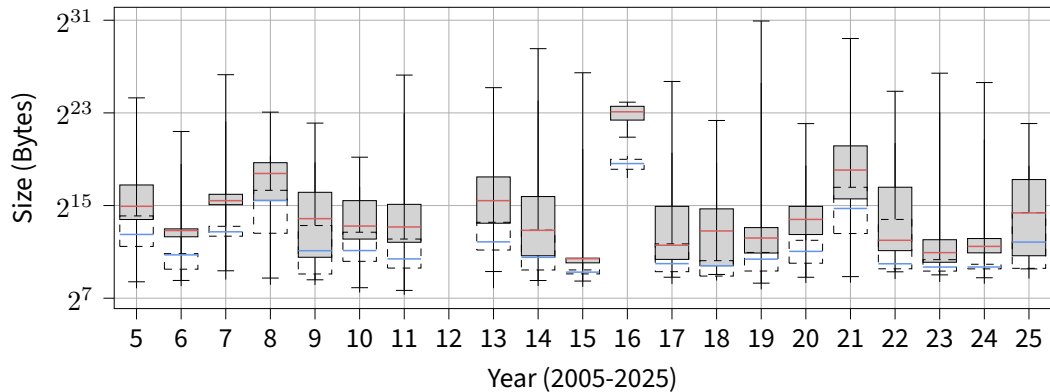
Histograms



Ratings By Age



New Benchmark Sizes



Solvers

