On Presenting Benchmark Results

Mats Carlsson, SICS

December 6, 2016

The Issue

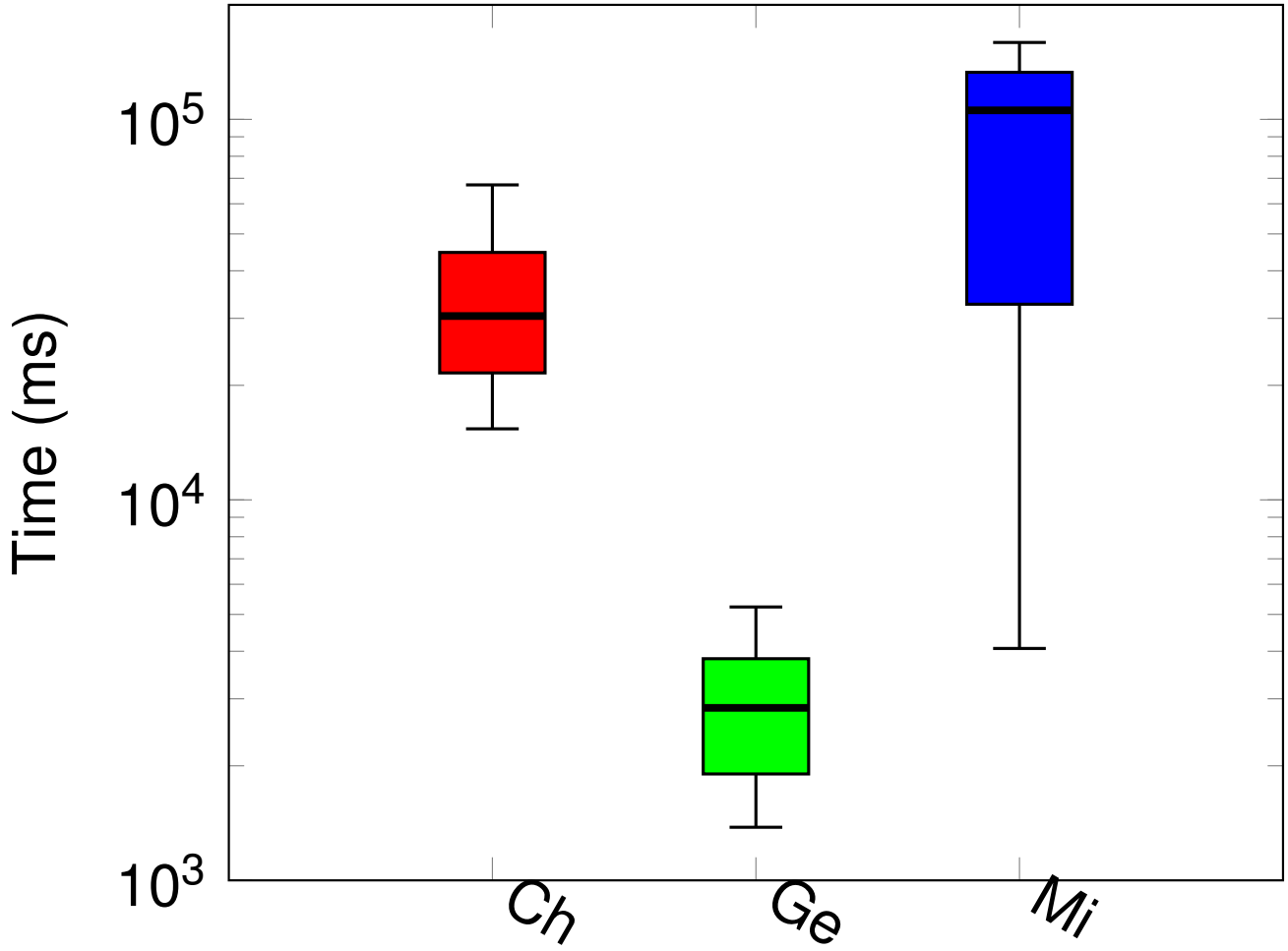
- One benchmark, multiple algorithms
- How to compare their performance
- Case study: 14 instances, 3 solvers, comparing runtime
- No instances time out

Table

Instance	Ch.	Ge.	Mi.
1	36550	**3207**	105624
2	30420	**2840**	32675
3	21560	**1904**	114081
4	39880	**3481**	106709
5	18830	**1847**	41669
6	44700	**3824**	137657
7	48550	**3825**	59086
8	15370	**1379**	159281
9	35670	**3127**	21932
10	22940	**2019**	147064
11	26110	**2316**	117292
12	20200	**1836**	9646
13	67270	**5230**	132960
14	54470	4989	**4069**
Ave	34466	**2987**	84982
StdDev	15223	**1194**	54575

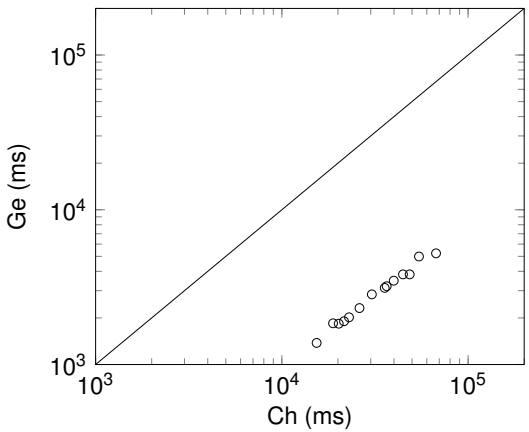
Box Plot

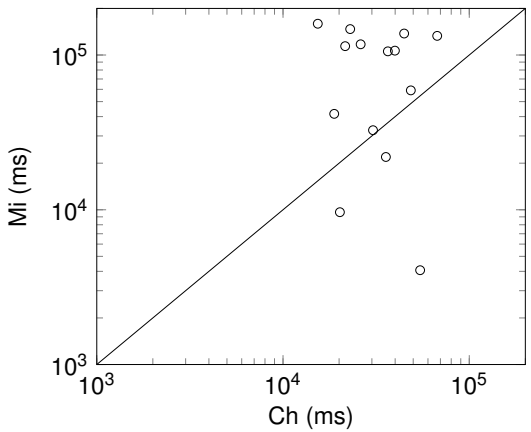
- Box shows first, second (median), third quartile
- Whiskers show min and max

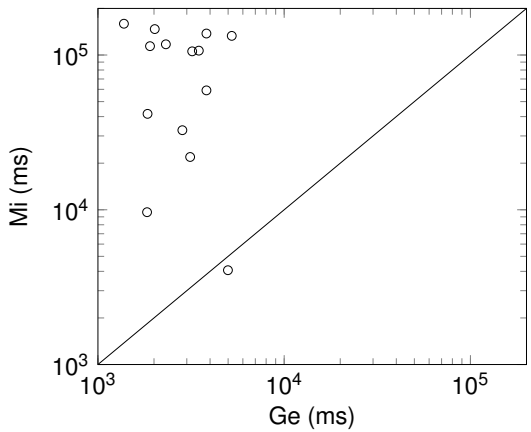


Scatter Plot

- Compare alg. A and B
- Point at (x, y): x ms for alg. A, y ms for alg. B
- Multiple data sets can be done in one plot

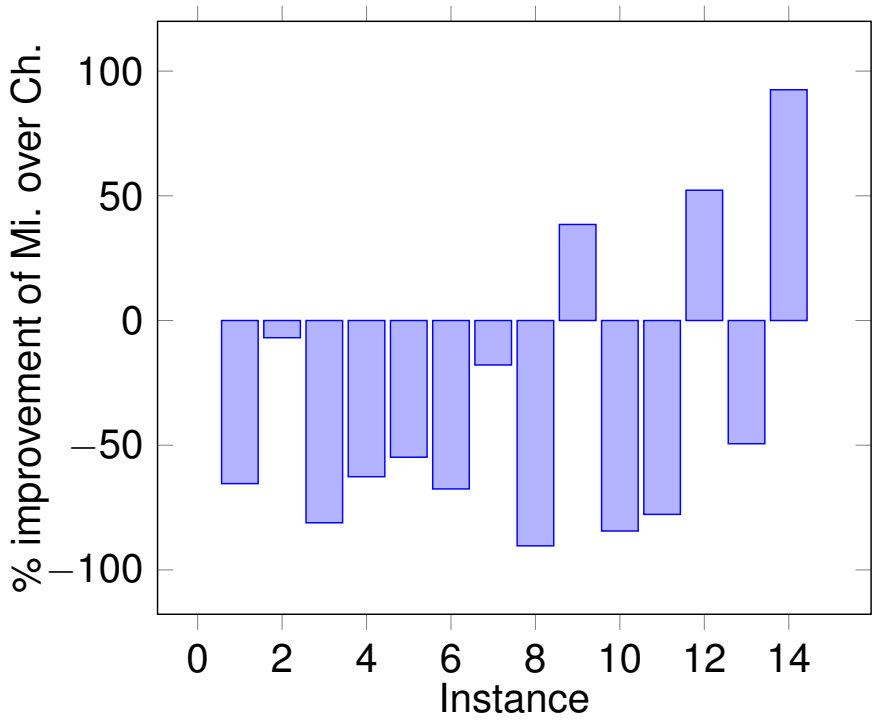






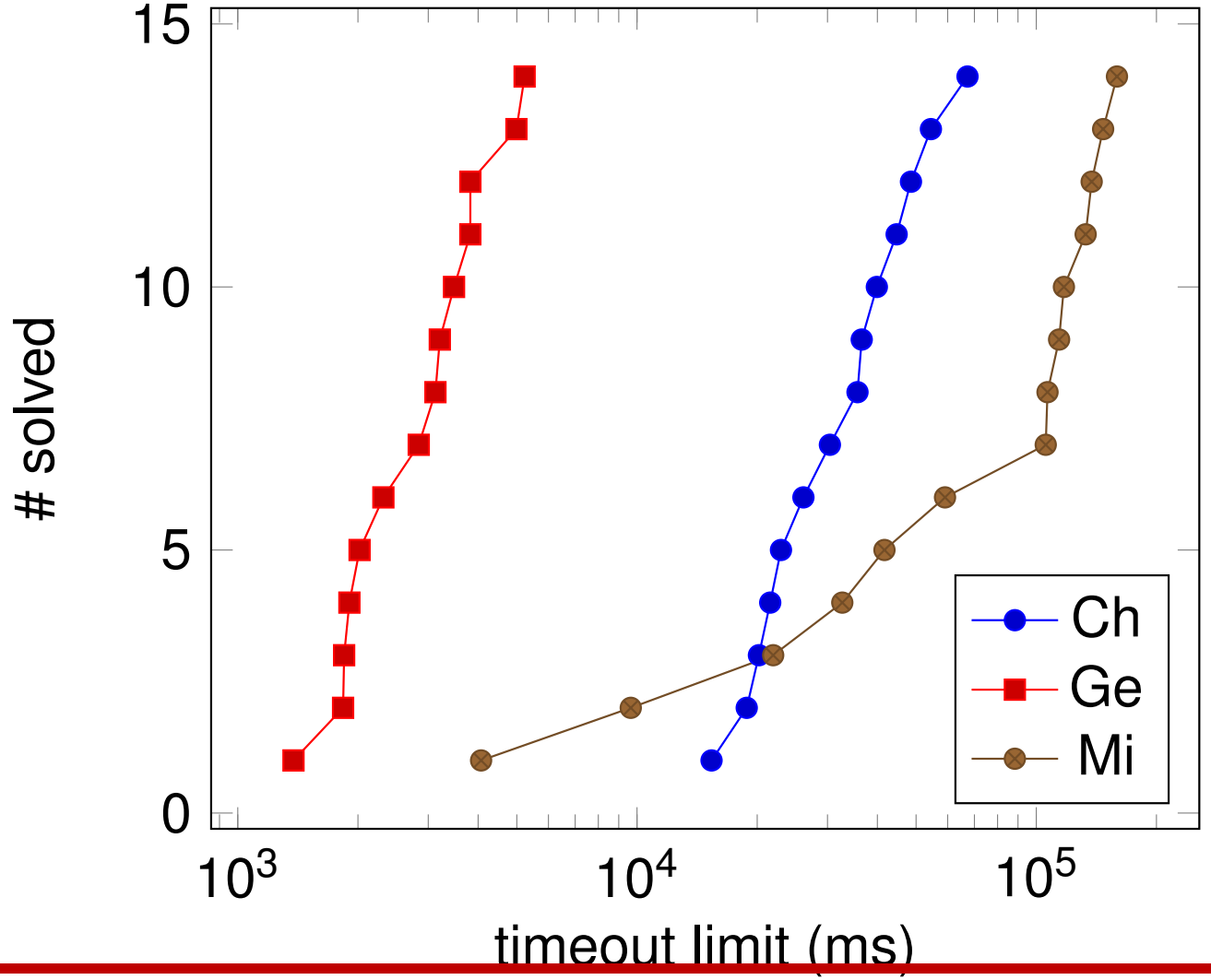
Baseline Histogram

- Algo. 1 (baseline) vs. algo. 2 (purportedly improved)
- Let t_{ij} denote runtime for instance i on algorithm j
- Let $h(i) = \begin{cases} 1 - t_{i2}/t_{i1} & \text{, if } t_{i1} > t_{i2} \quad \text{(improvement)} \\ t_{i1}/t_{i2} - 1 & \text{, if } t_{i1} < t_{i2} \quad \text{(deterioration)} \end{cases}$



Deadline Plot

- $f(x)$: #instances solved by deadline x



Performance Profile

- Dolan, E. D. and Moré, J. J. 2002. Benchmarking optimization software with performance profiles. *Mathematical programming* 91(2):201–213.
- $f(\tau)$: % instances that are $\leq \tau\times$ slower than the best algorithm