BBOB Black-Box Optimization Benchmarking with CoCO (Comparing Continuous Optimizers)

The Turbo-Intro

Black-Box Optimization (Search)

Minimize (or maximize) a continuous domain objective (cost, loss, error, fitness) function

$$f : \mathbb{R}^d \to \mathbb{R}$$

in a black-box scenario (direct search)

$$x \longrightarrow \blacksquare \longrightarrow f(x)$$

where

- gradients are not available or useful
- problem specific knowledge is used only *within* the black box, e.g. with an appropriate encoding

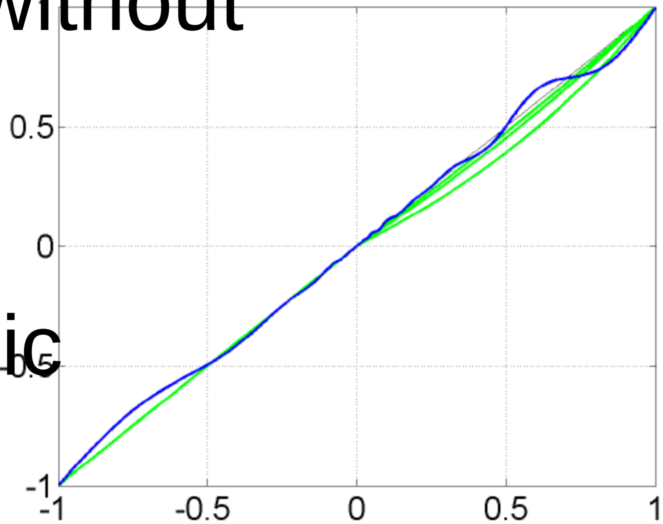
The search costs are the number of function evaluations

CoCO: the noiseless functions

24 functions within five sub-groups

- Separable functions
- Essential unimodal functions
- Ill-conditioned unimodal functions
- Multimodal structured functions
- Multimodal functions with weak or without structure

functions are not perfectly symmetric
and are locally deformed



CoCO: the noisy functions

three noise-”models”, so-called:

- Gauss, Uniform (severe), Cauchy (outliers)
- Utility-free noise

$$E(f(x)) \leq E(f(y)) \Rightarrow U(f(x)) \leq U(f(y)) \; \forall x, y, U$$

30 functions with three sub-groups

- 2x3 functions with weak noise
- 5x3 unimodal functions
- 3x3 multimodal functions

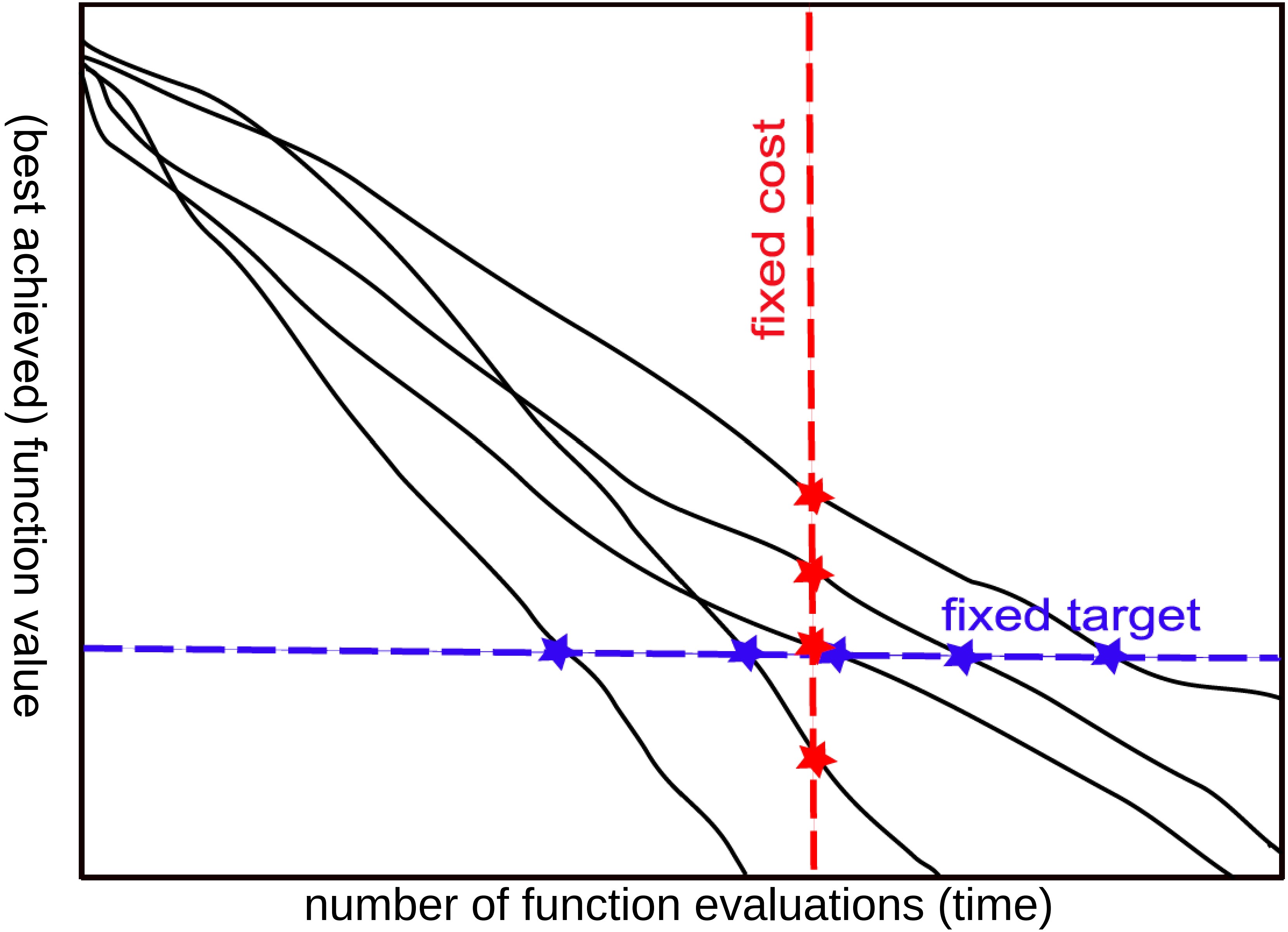
Measuring Performance

convergence graphs is

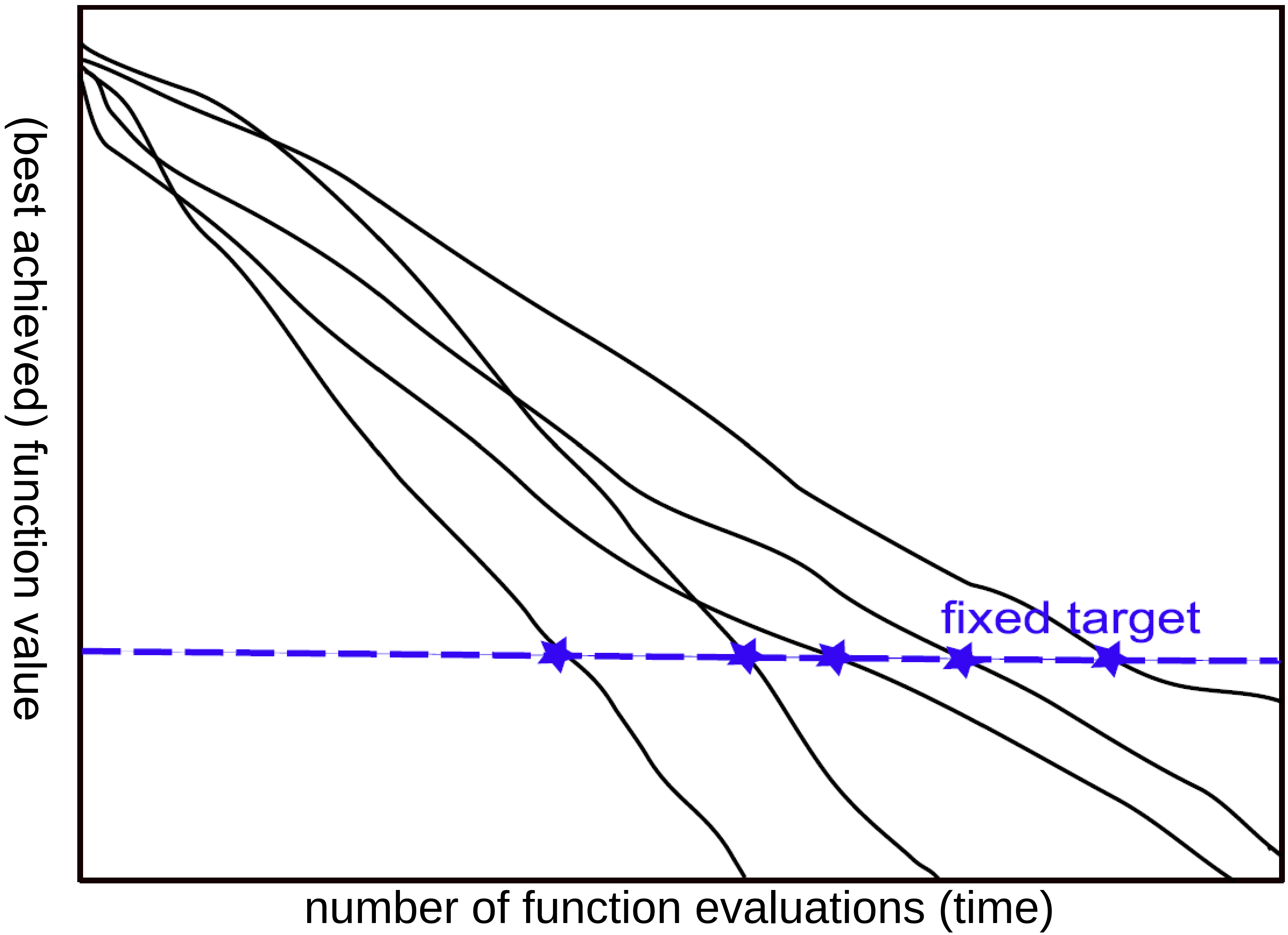
all we have to start with

Measuring Performance from Convergence Graphs

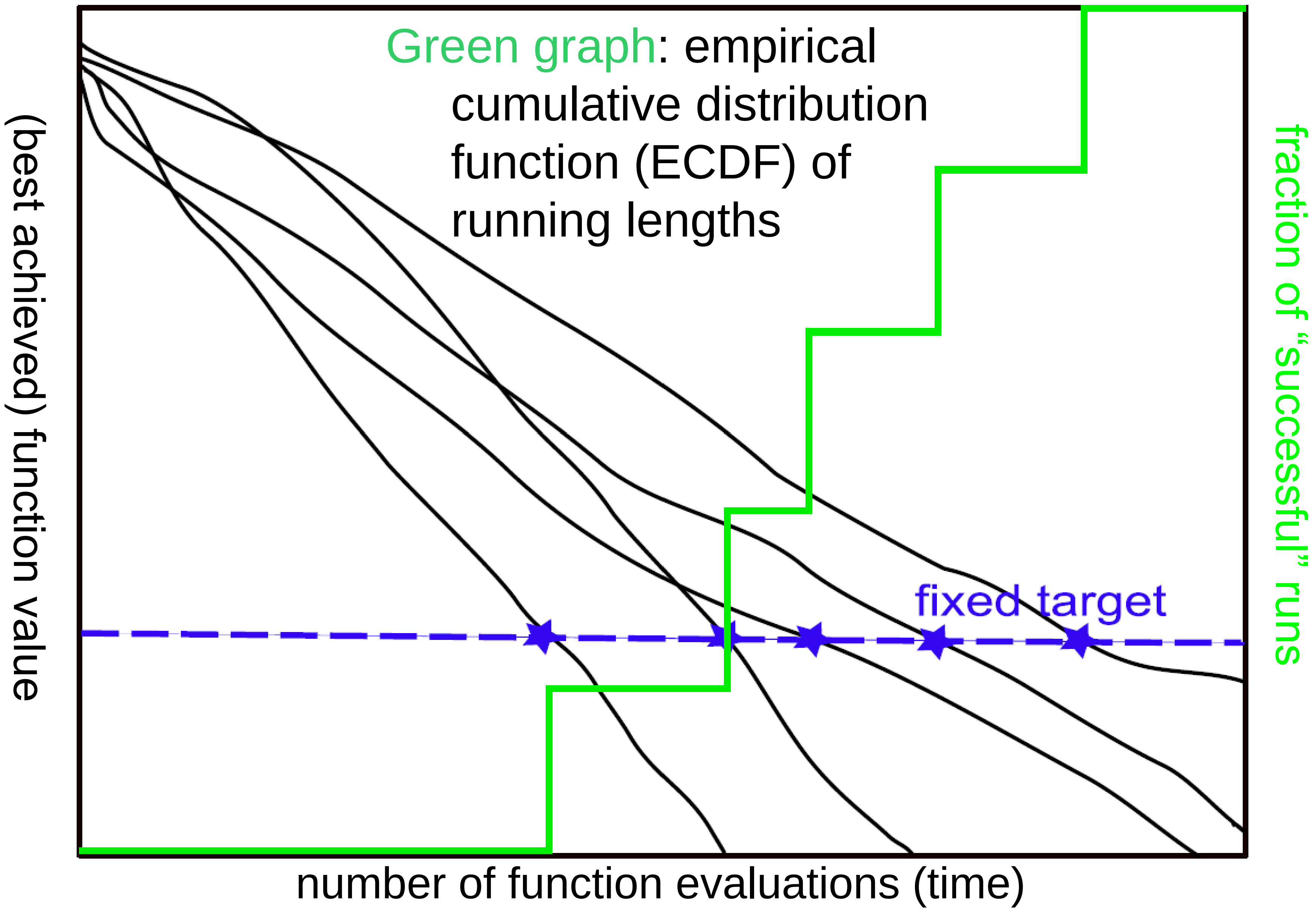
fixed-cost versus fixed-target



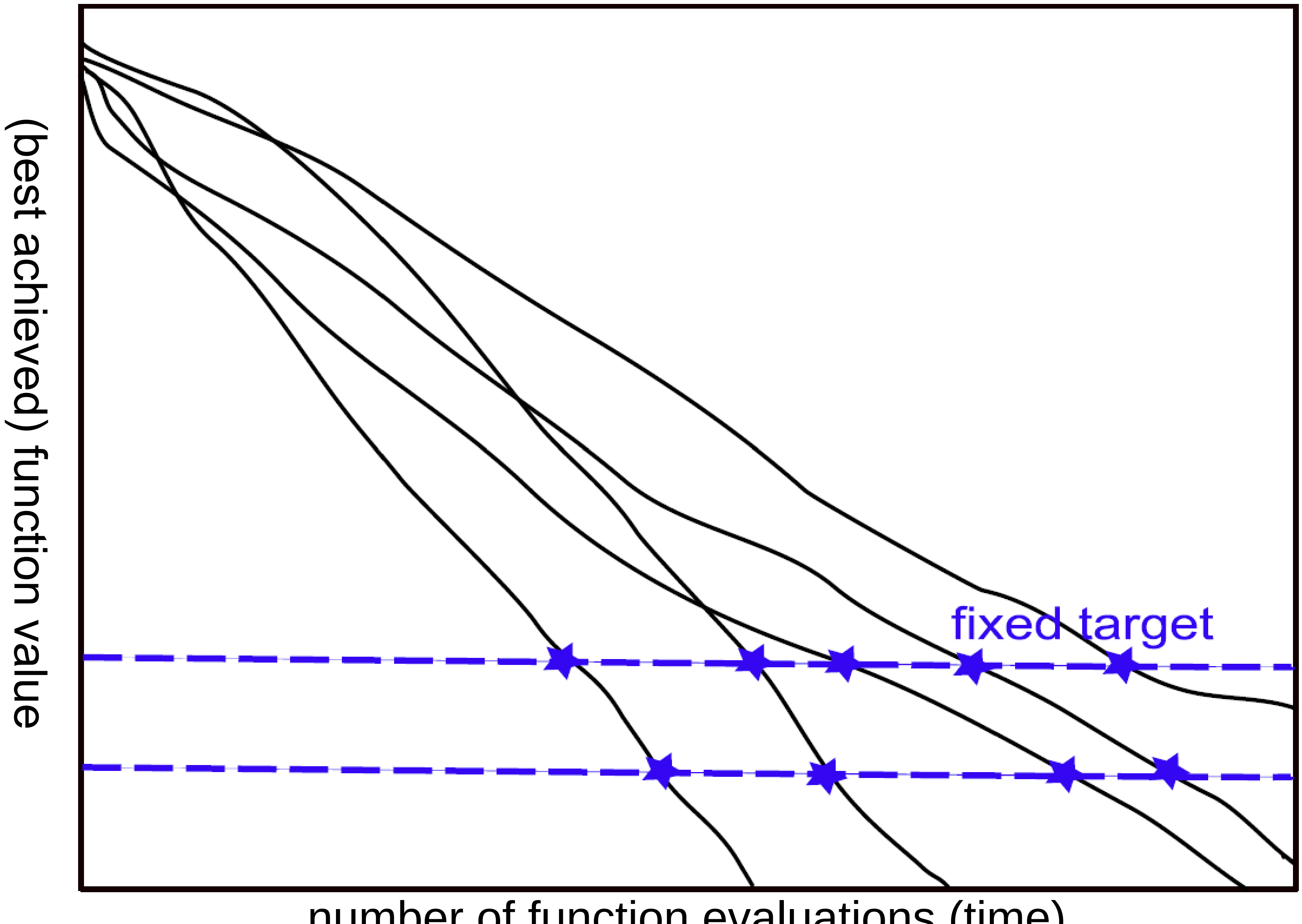
Empirical Cumulative Distribution with a given target value



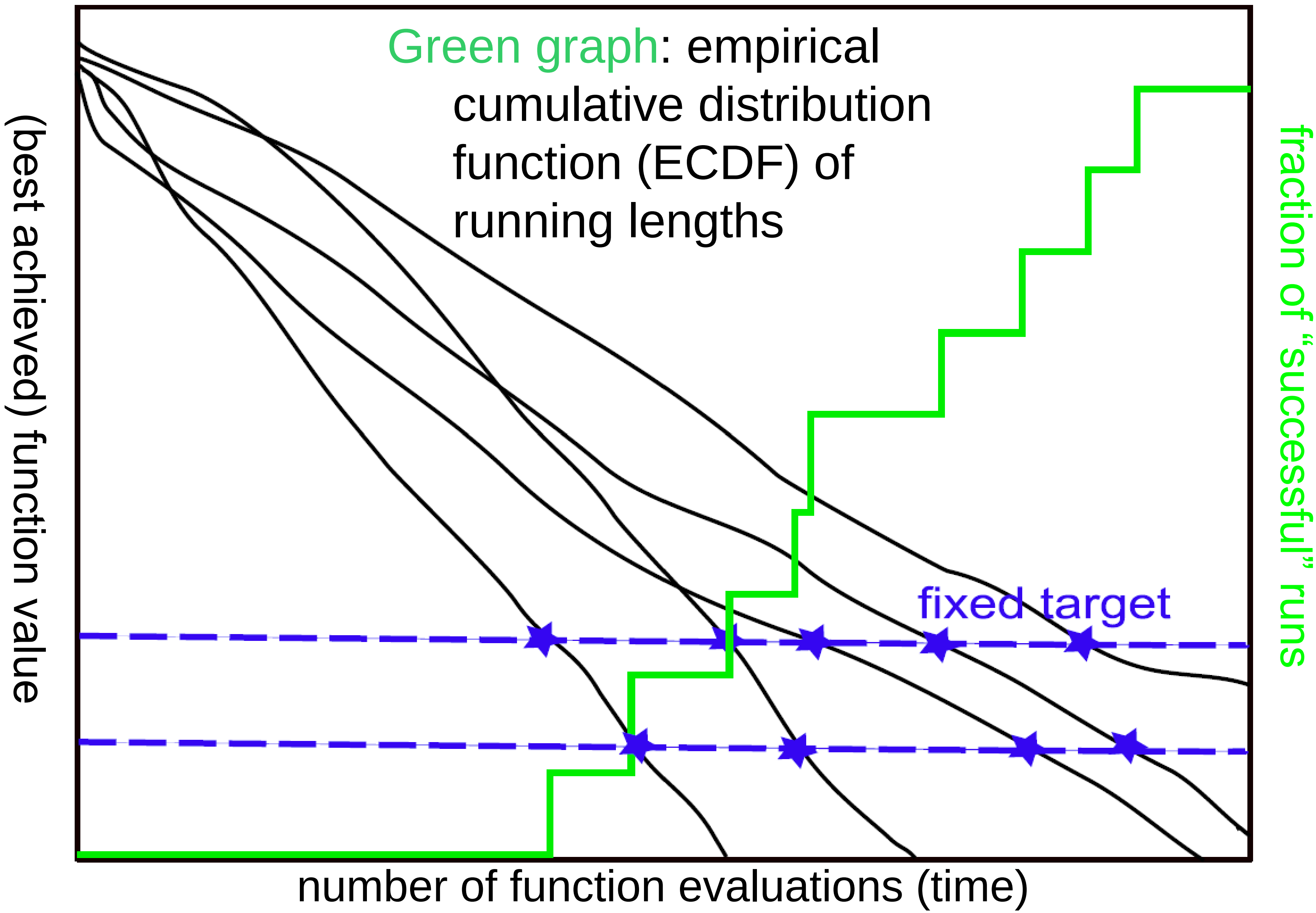
Empirical Cumulative Distribution with a given target value



Empirical Cumulative Distribution with two given target values



Empirical Cumulative Distribution with two given target values



Cumulative Distribution of Runtimes

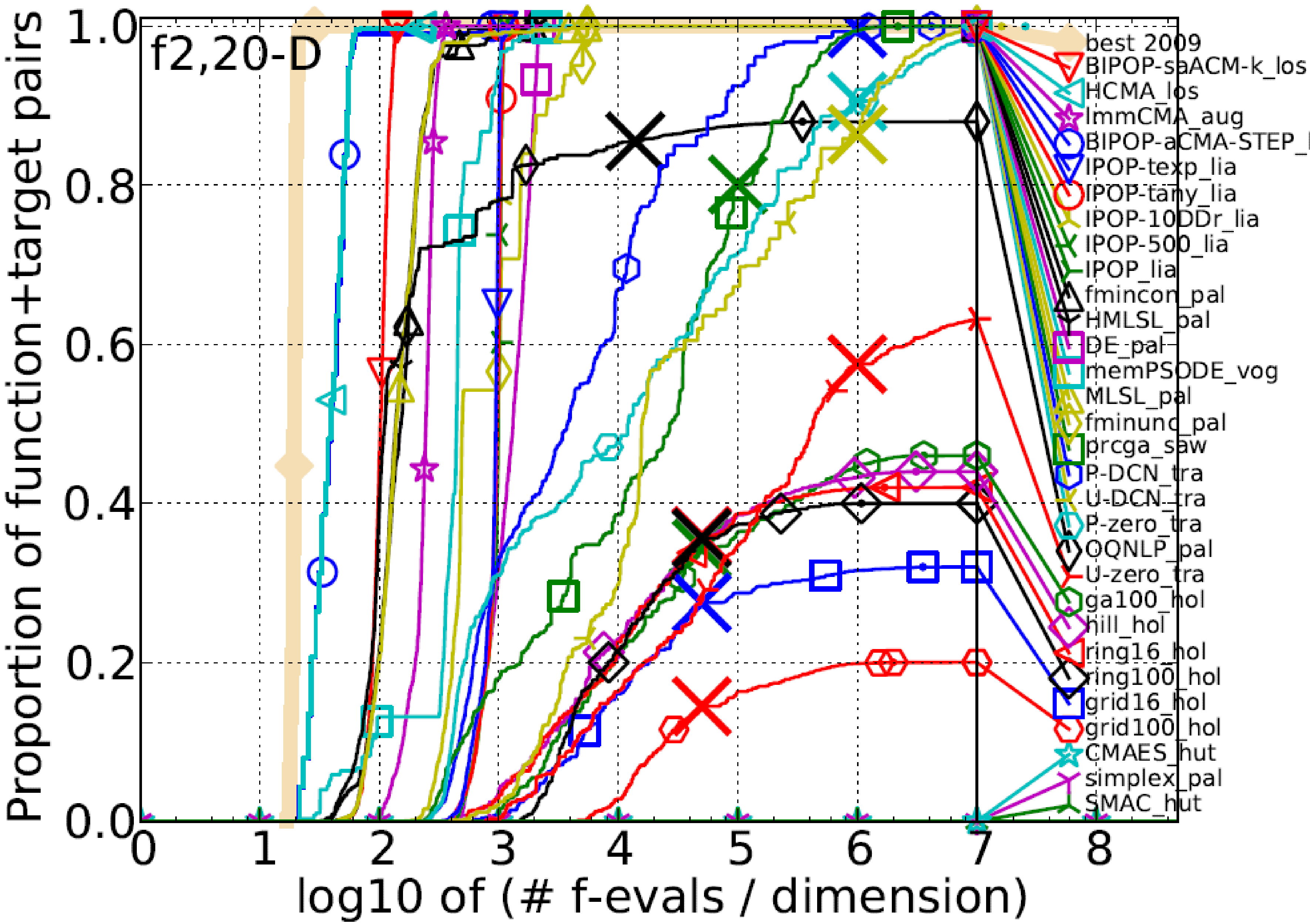
Runtime ECDFs (empirical cumulative distribution function) display a set of runlengths

- they can aggregate over any set of functions and target values

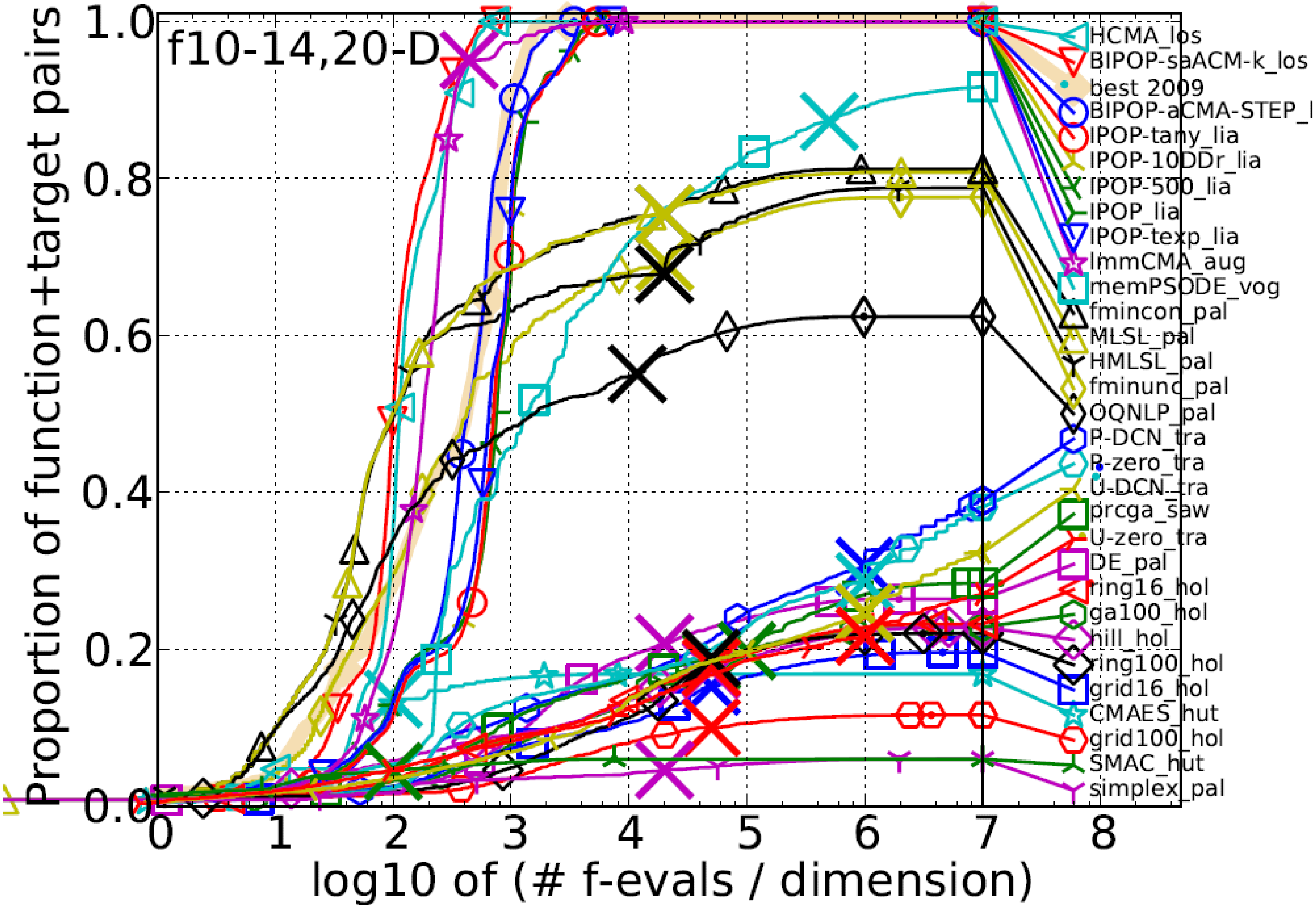
with the least amount of information loss into a single curve

- in BBOB:
  - 50 target values (log-uniform in [1e-8,100]) and 15 trials per function = 750 runlength values per function
  - aggregate of one to 30 functions
  - for unsuccessful runs: simulated restart within 15 instances

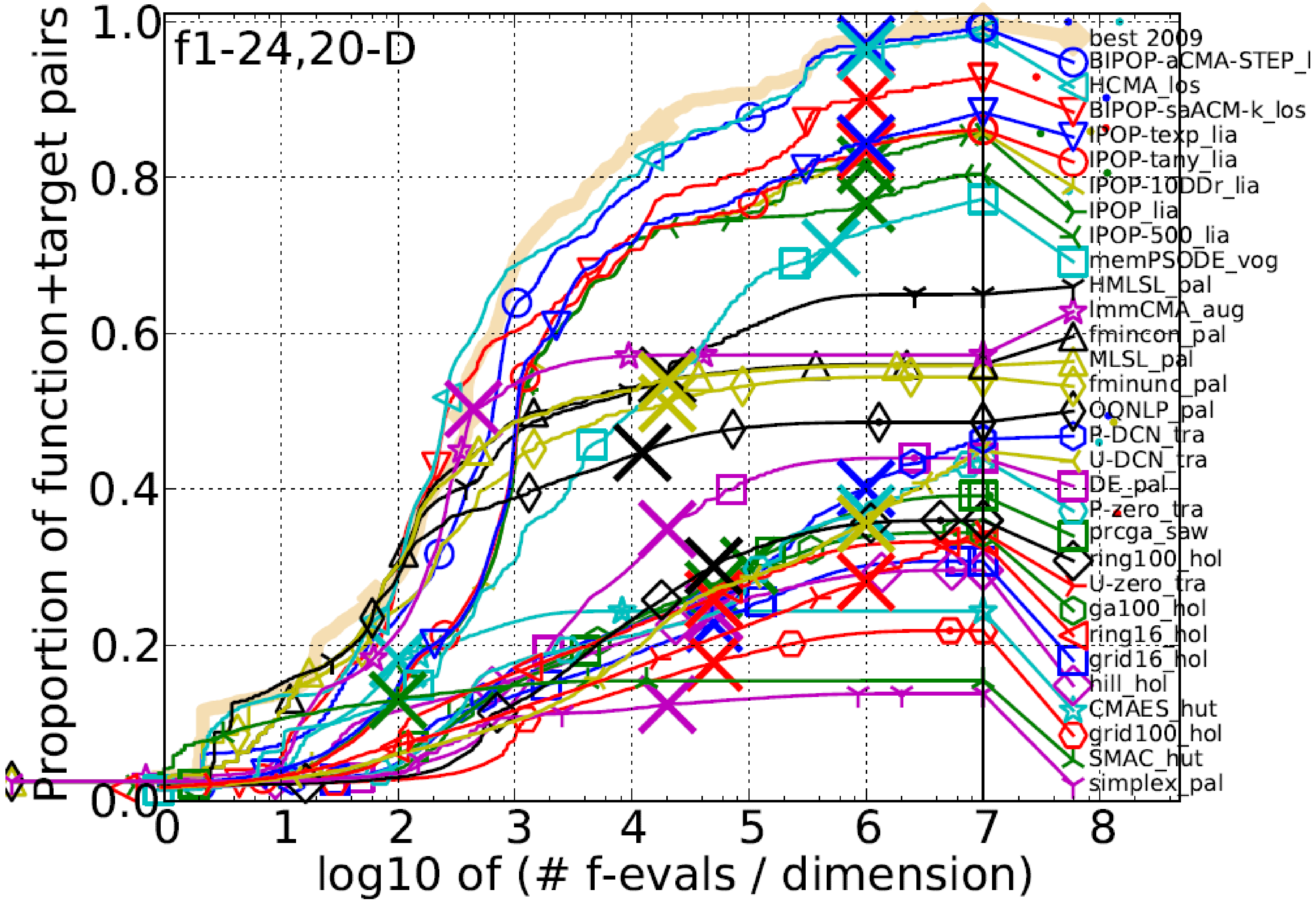
Examples of ECDFs



Examples of ECDFs



Examples of ECDFs



Evaluation of Search Algorithms

a performance should be

- quantitative on the ratio scale (highest possible)

  + "algorithm A is two *times* better than algorithm B" is a meaningful statement

  + can assume a wide range of values

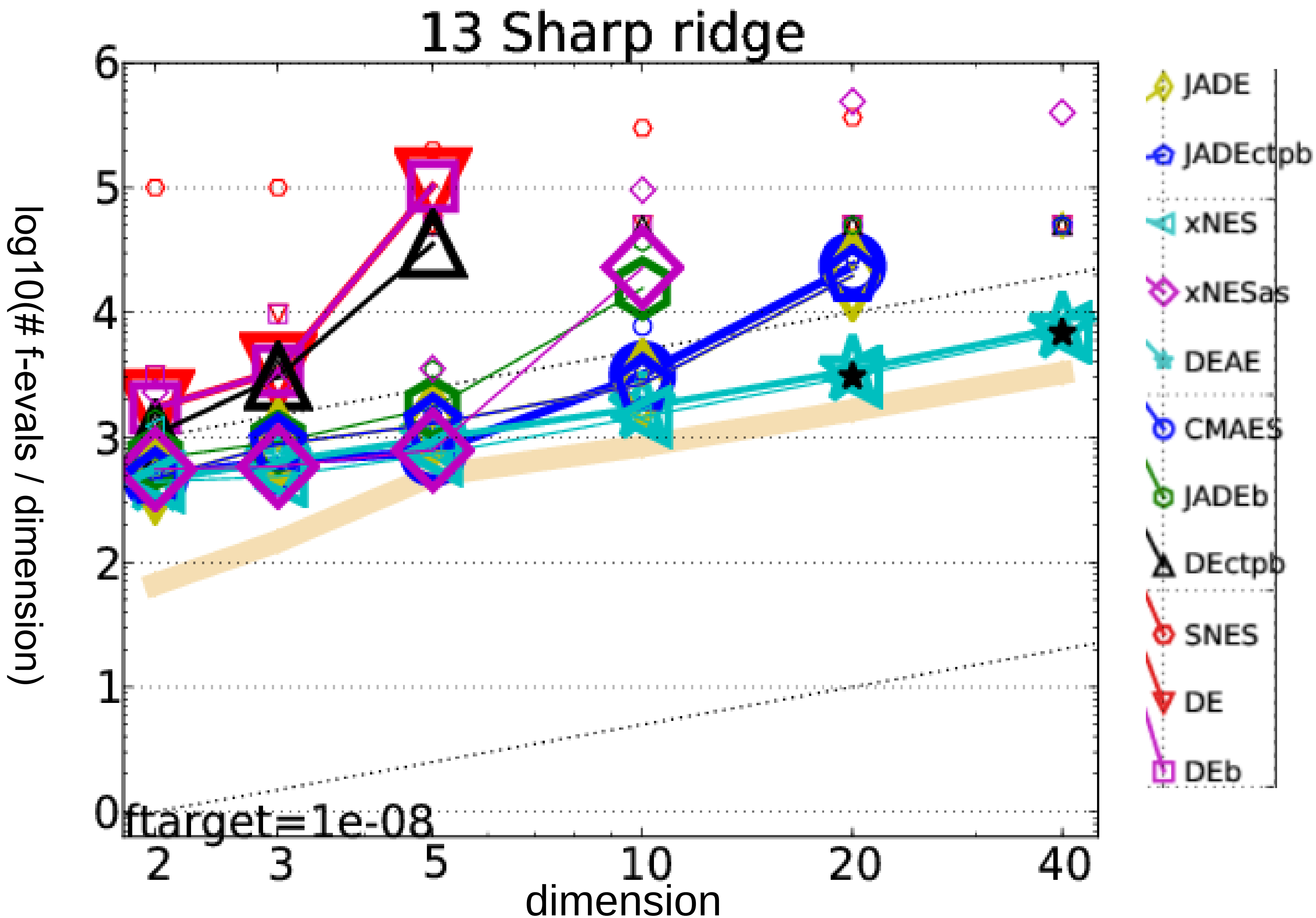
- meaningful (interpretable) with regard to the real world

  possible to transfer from benchmarking to real world

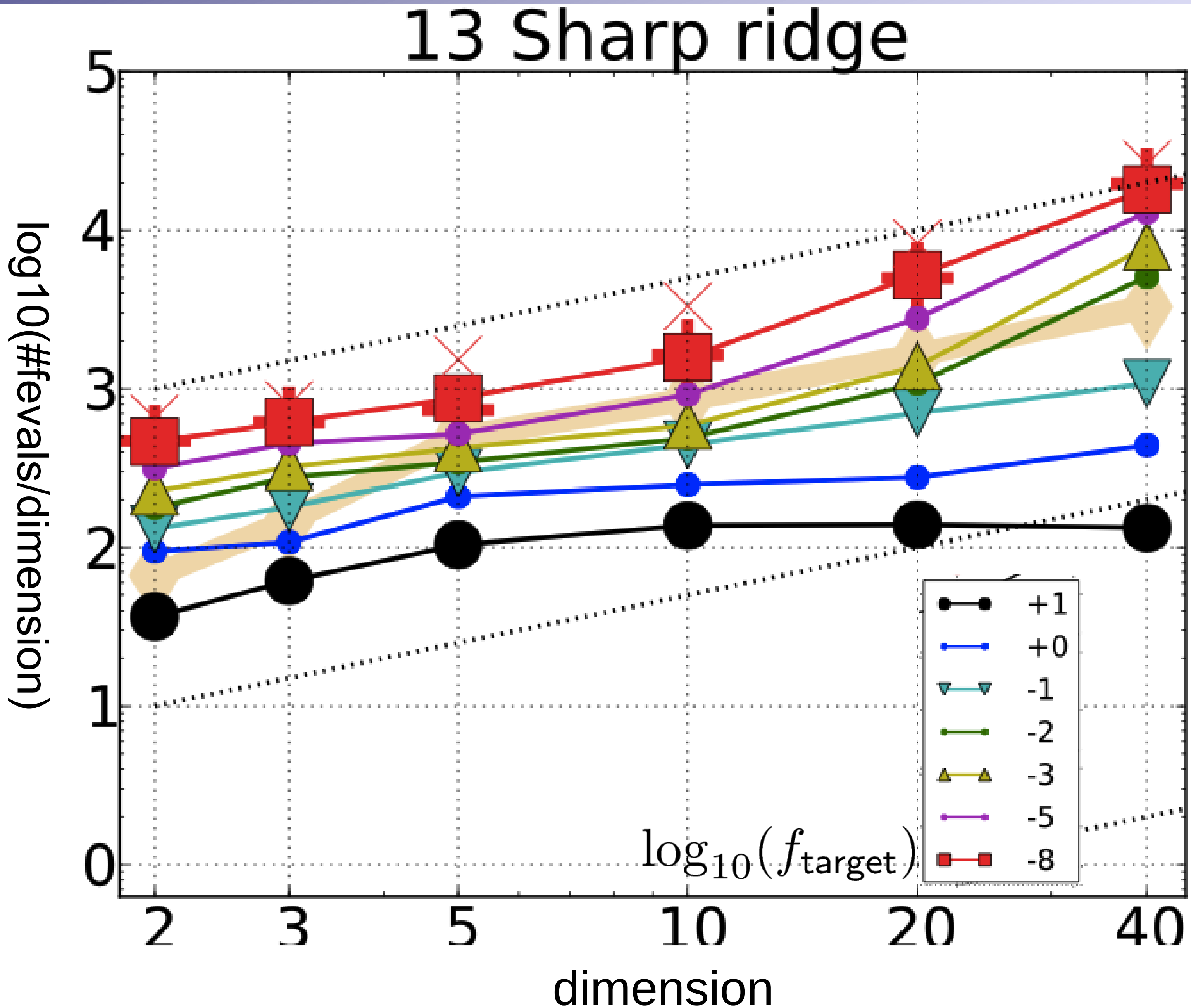
runtime is the prime candidate (we don't have many choices anyway)

other plots for single functions

Scaling Behaviour with Dimension

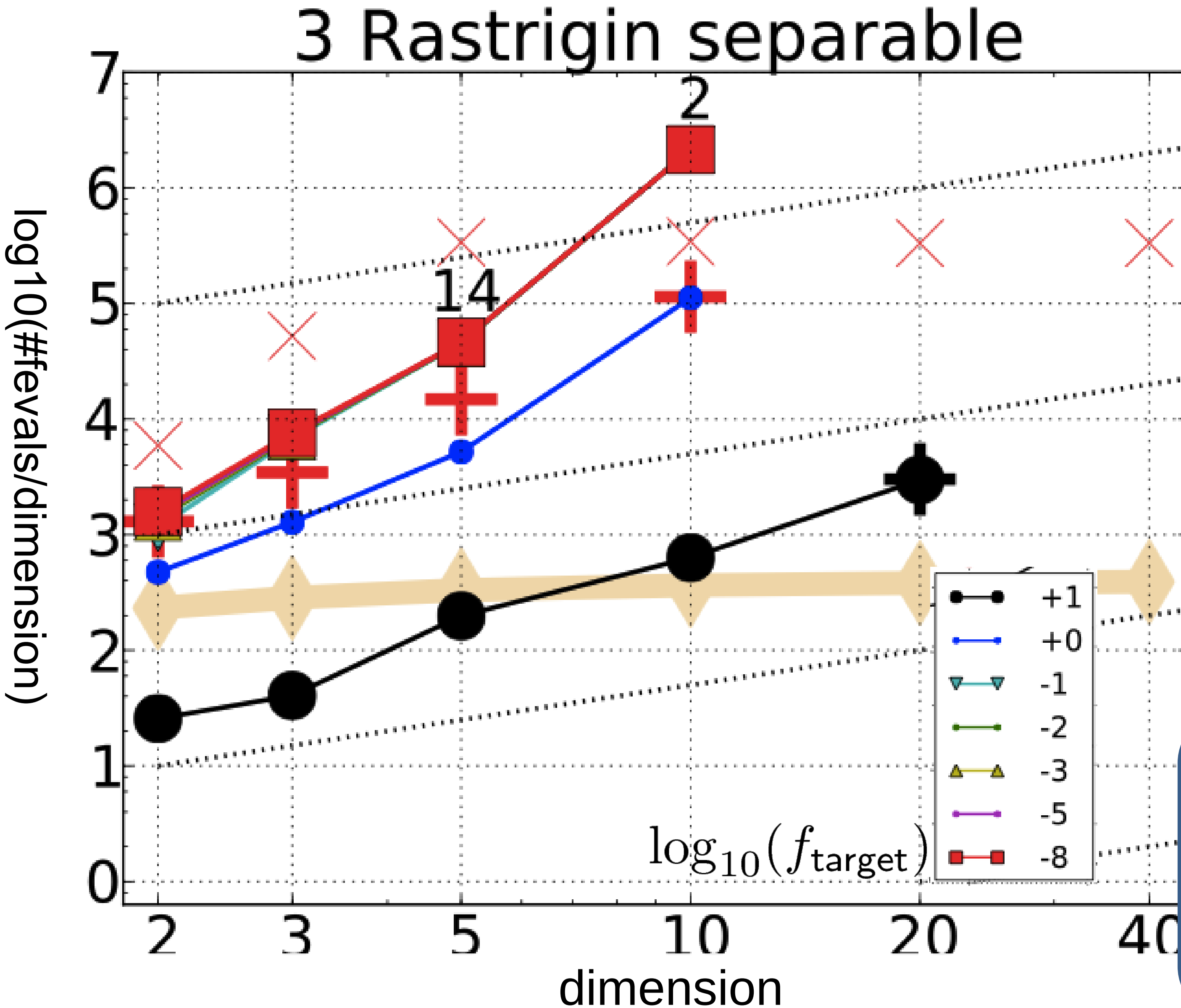


Scaling Behaviour with Dimension



- slanted grid lines: quadratic scaling
- horizontal lines: linear scaling
- light brown: artificial best 2009

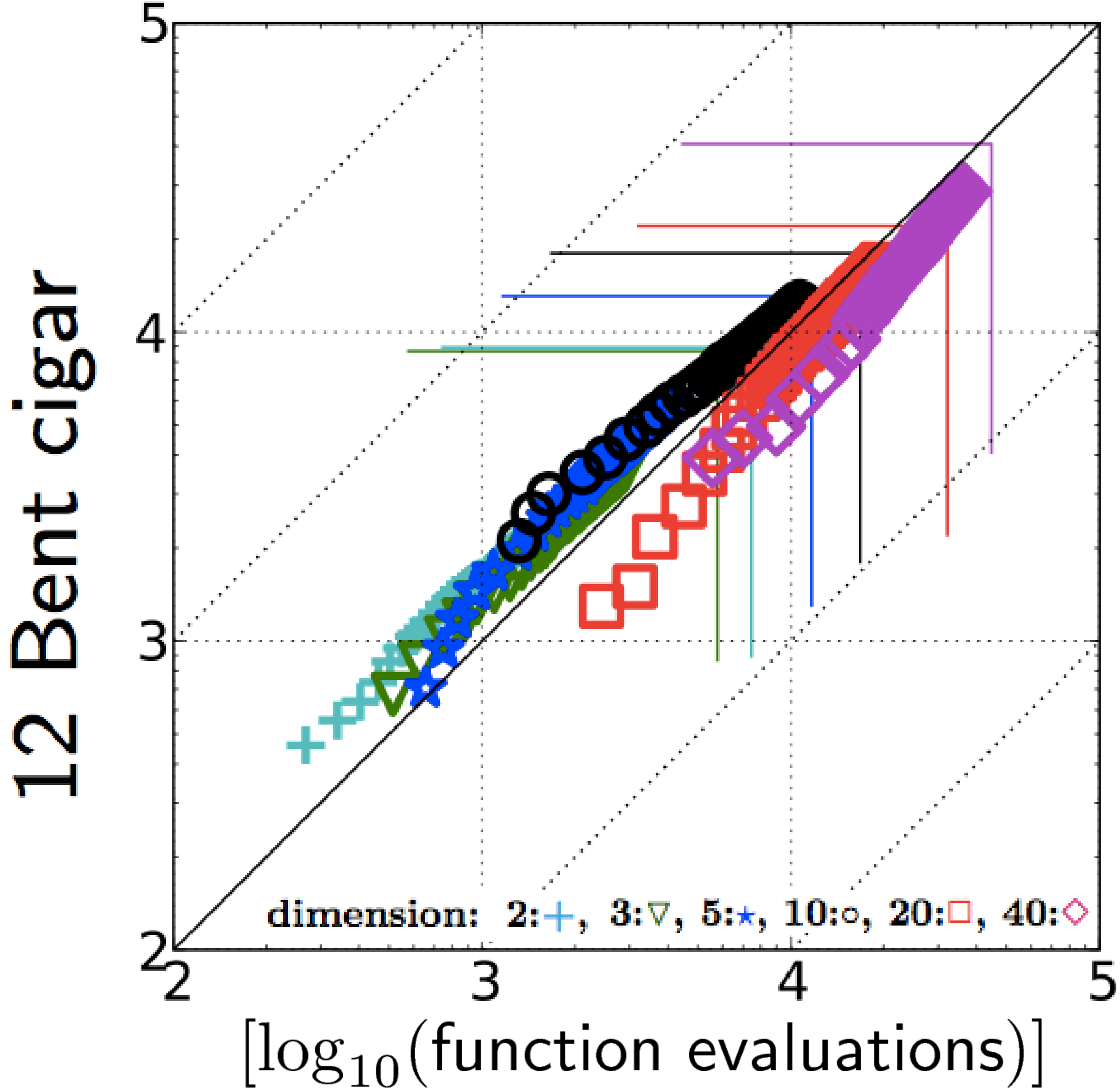
Example: Scaling Behaviour



- slanted grid lines: quadratic scaling
- horizontal lines: linear scaling
- light brown: artificial best 2009

$\Longrightarrow$ Experiments in >40-D are more often than not virtually superfluous

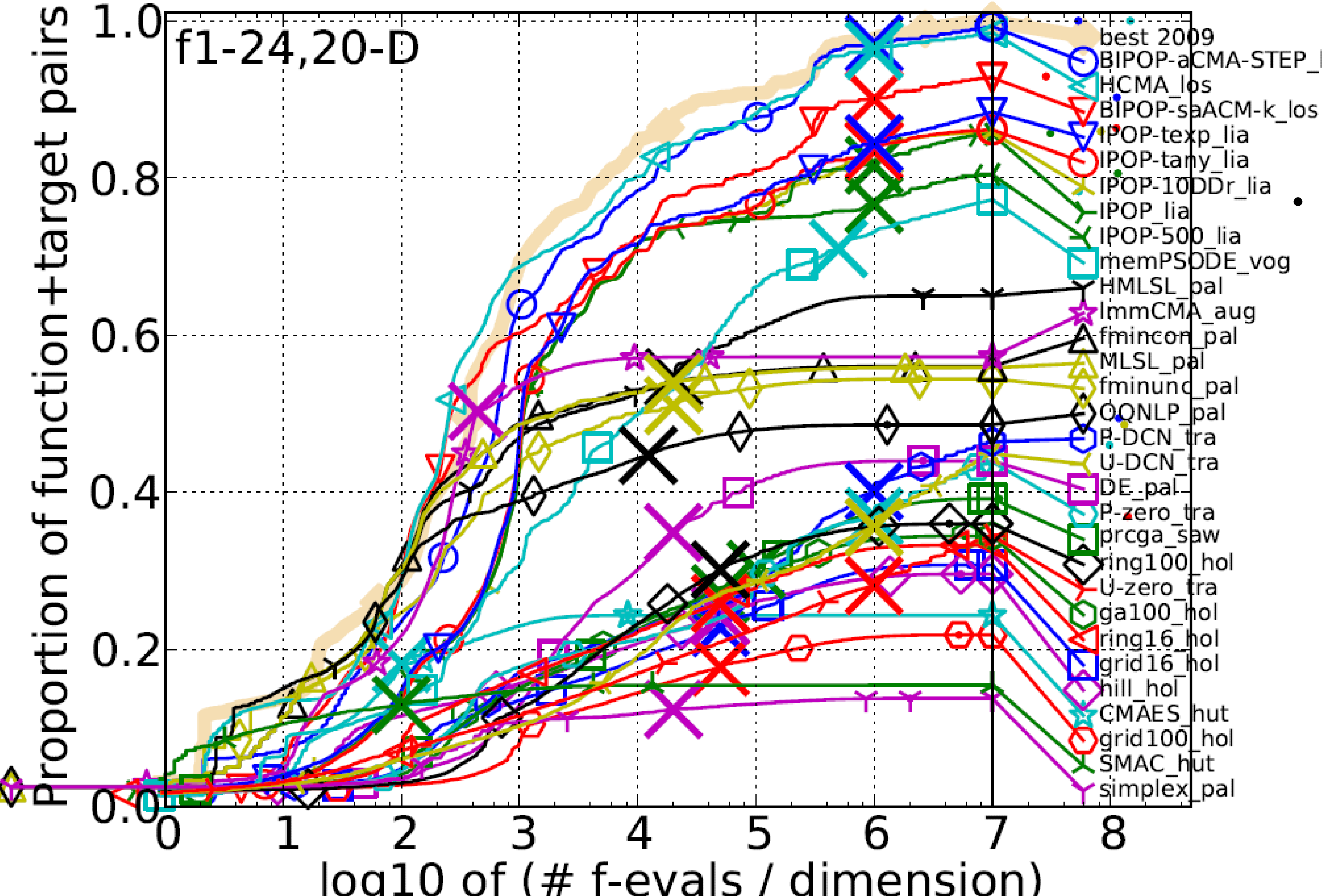
ERT scatter plot, all dimensions & targets



- estimated Expected Run Time (ERT), two algorithms
- 2-10 D: first algorithm "dominates"
- 20 & 40 D: second algorithm "dominates"

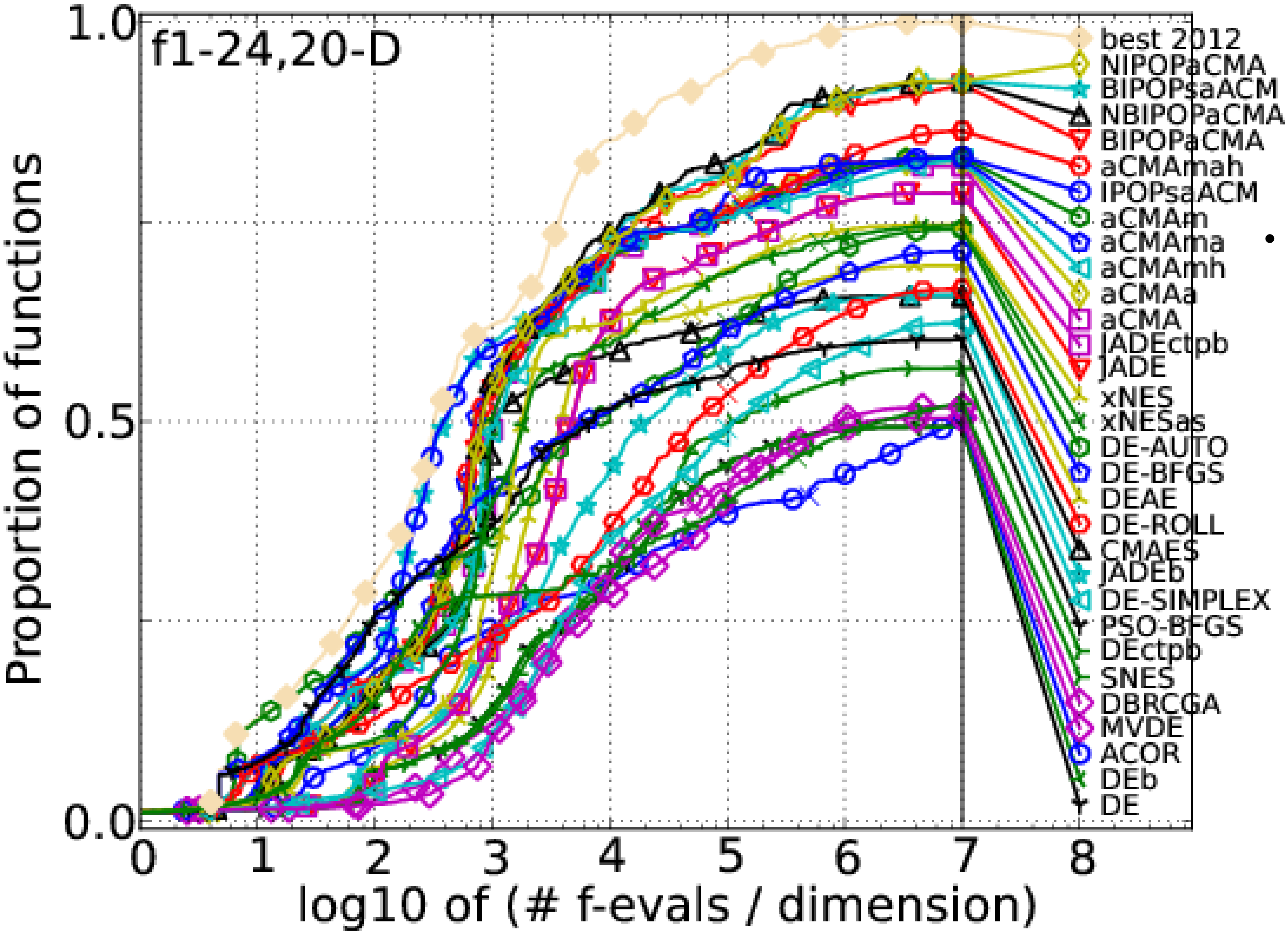
Questions?

All data 2013



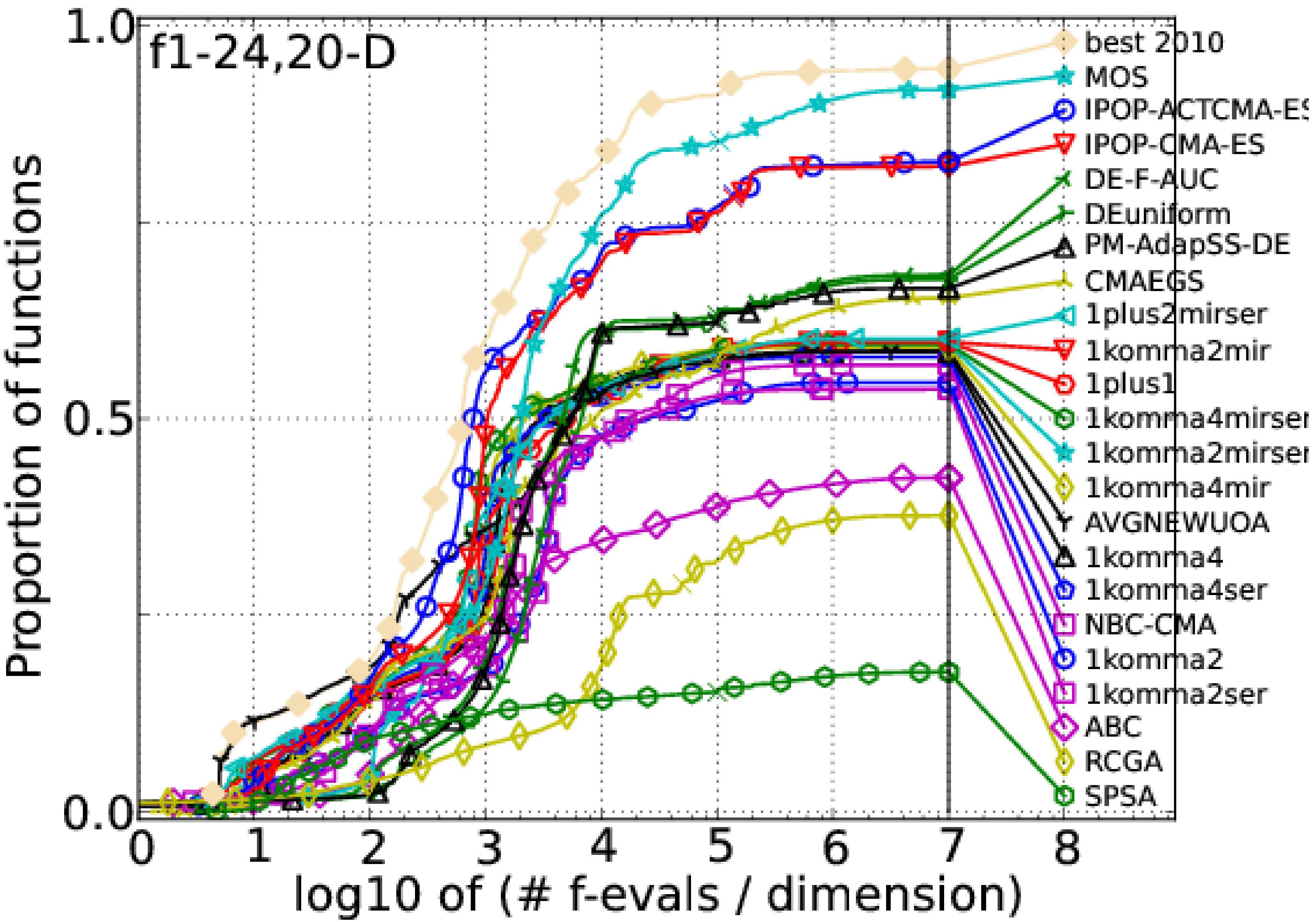
- “two objectives”:
  - fast
  - successful
- overfitting?

All data 2012

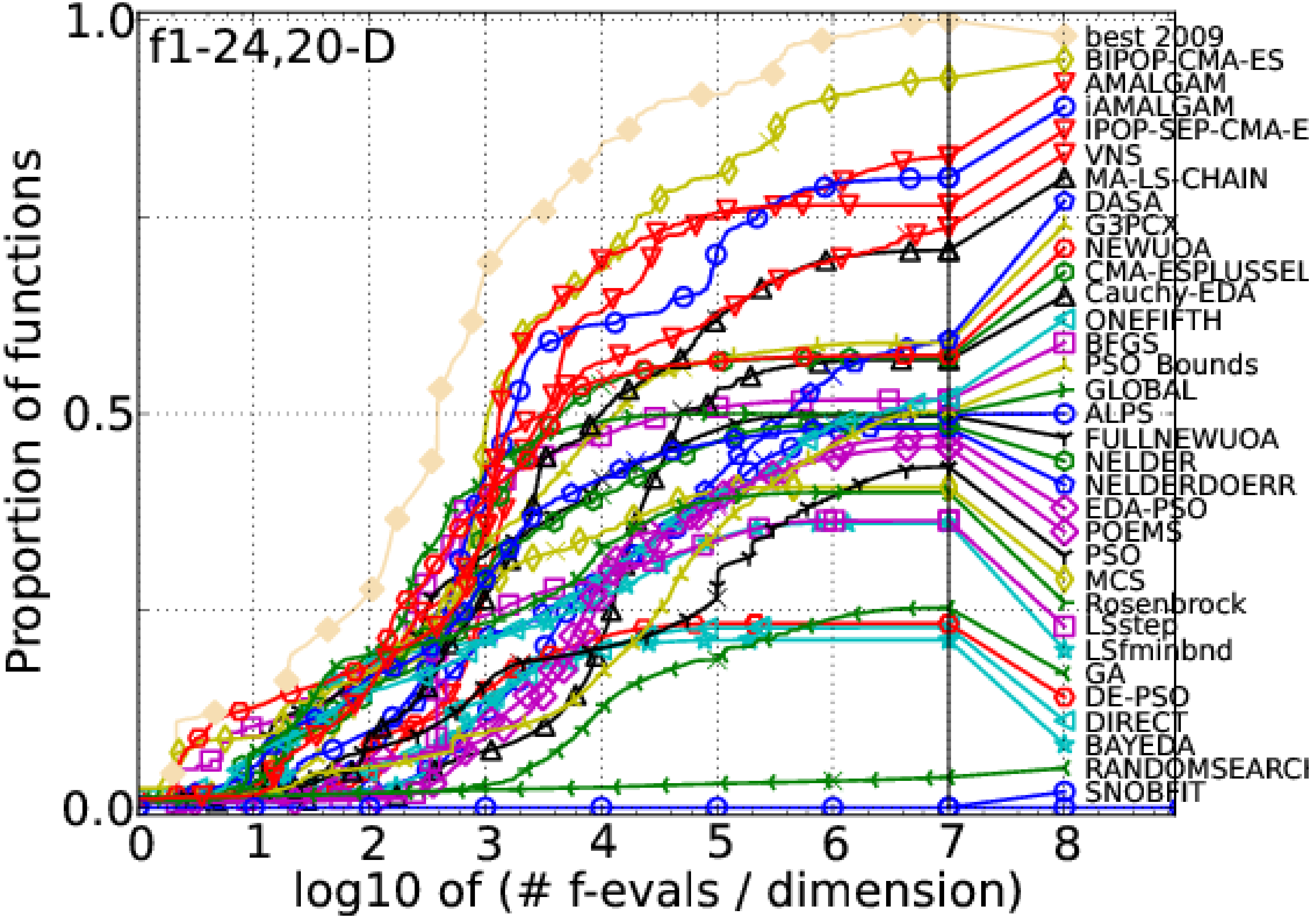


- "two objectives":
  - fast
  - successful
- overfitting?

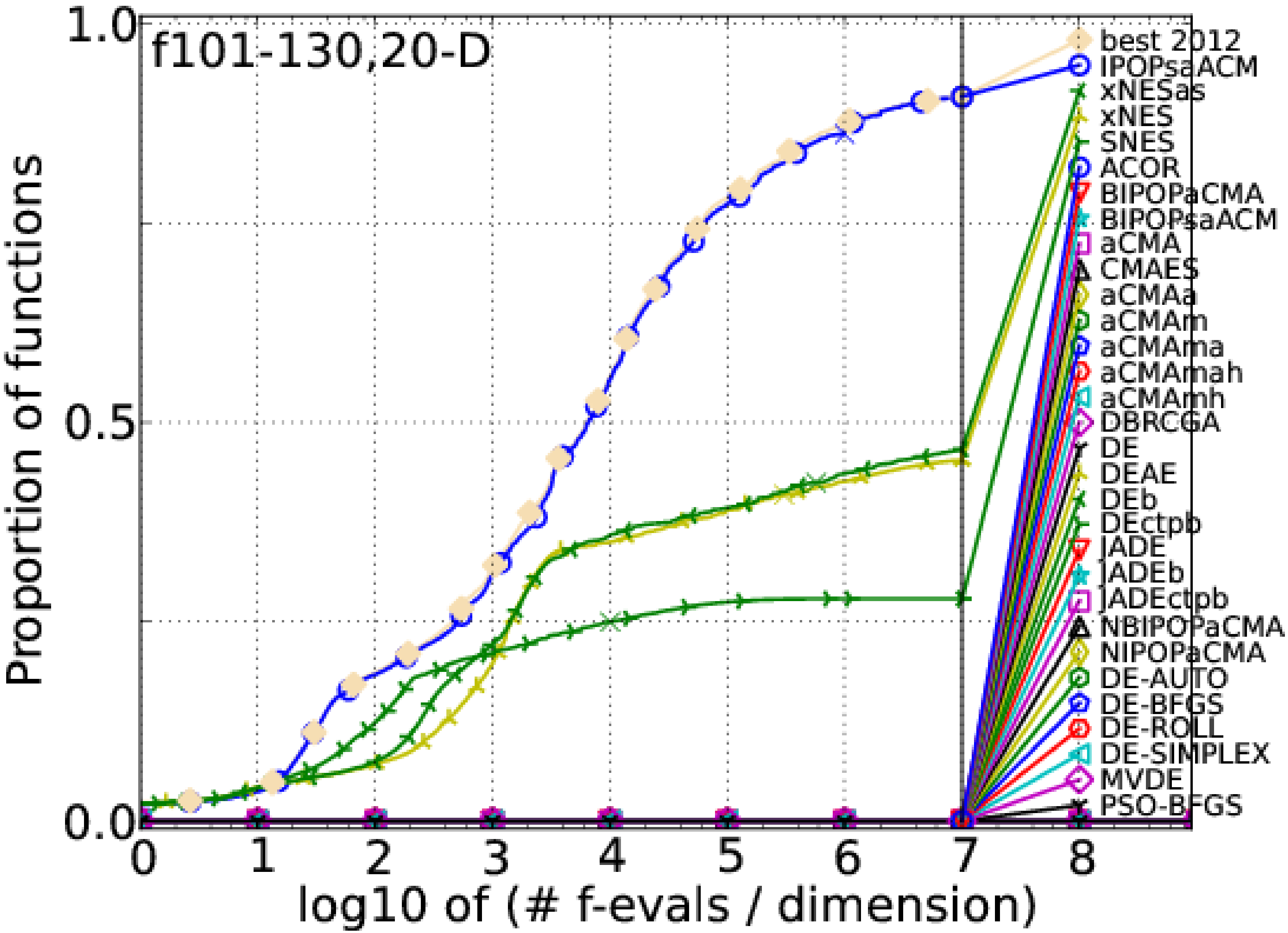
All data 2010



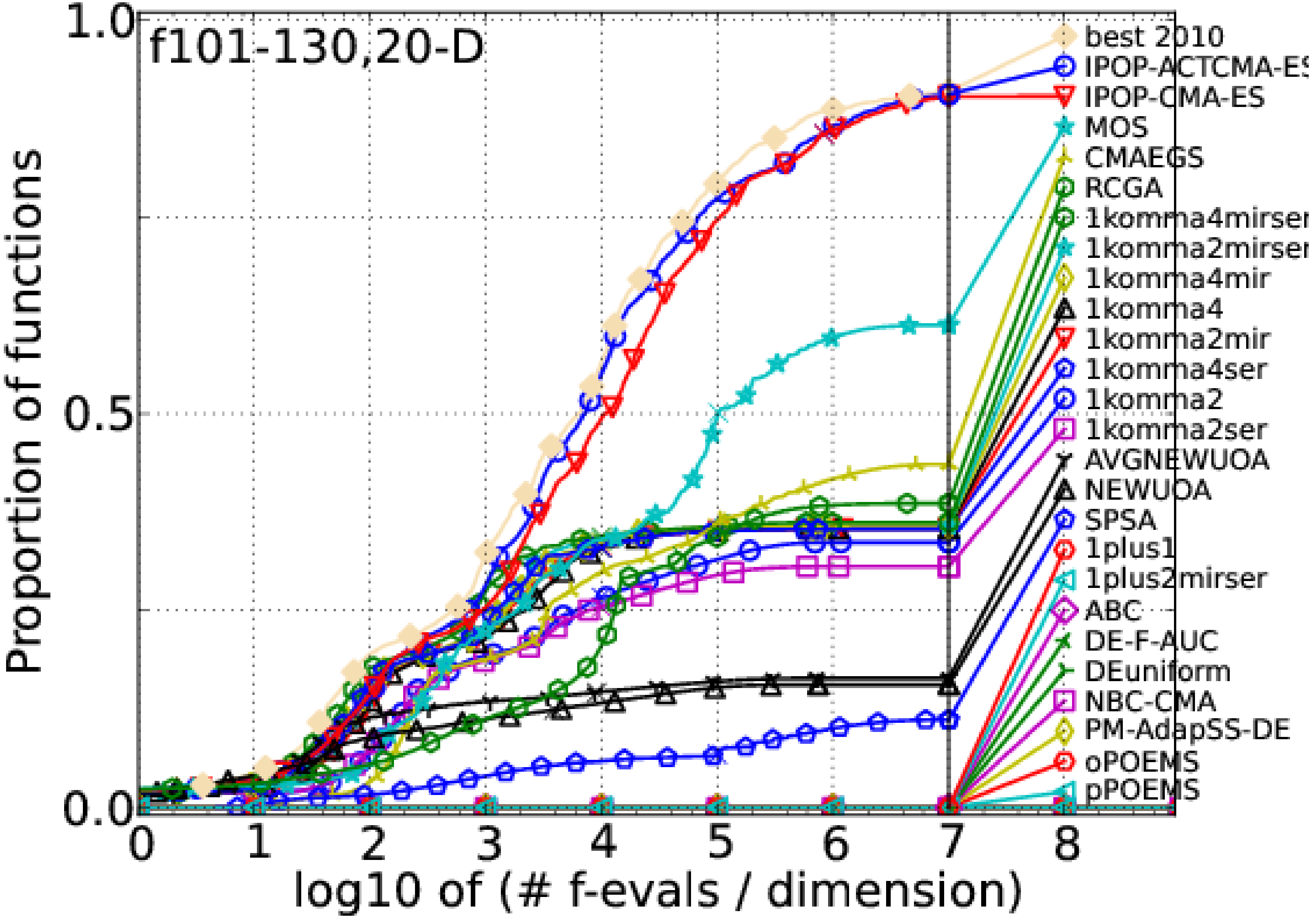
All data 2009



All data 2012 (noisy)



All data 2010 (noisy)



All data 2009 (noisy)