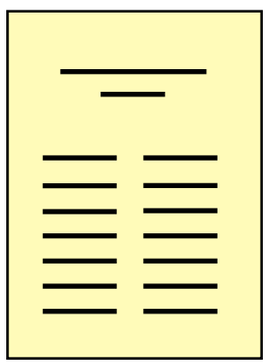
Make your empirical research more persuasive

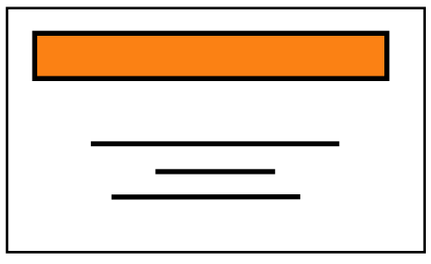
Simon Poulding, University of York
SBST @ ICST, April 2012

Context

SBSE / SBST Empirical Research Tutorials

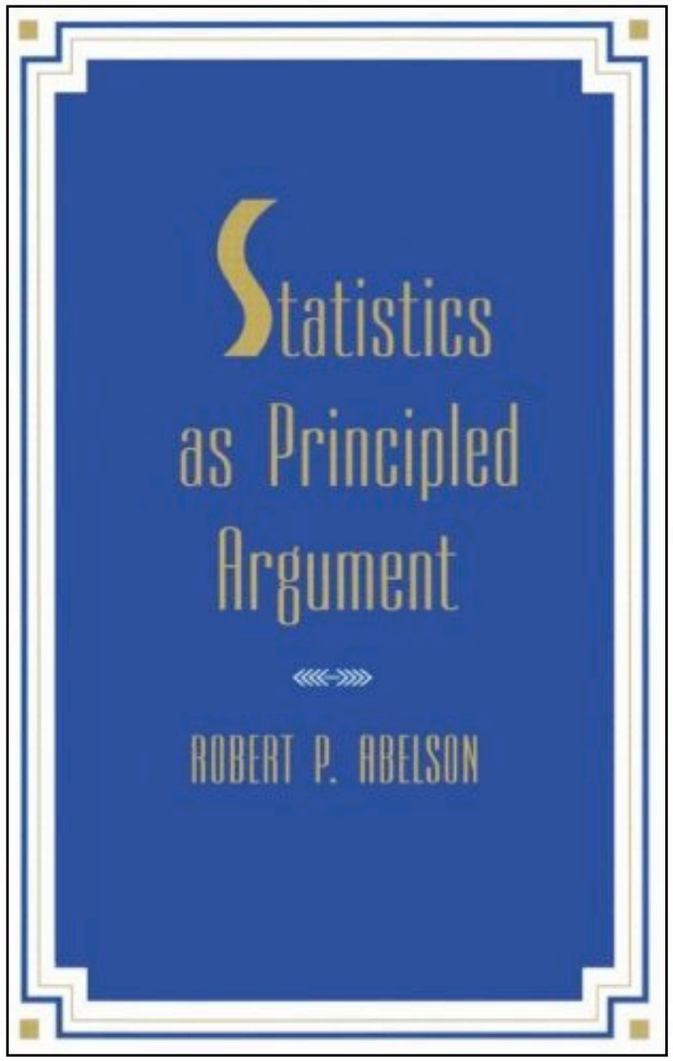


Andrea Acuri and Lionel Briand
“A Practical Guide for Using Statistical Tests to Assess Randomized Algorithms in Software Engineering”
ICSE 2011



Lionel Briand
“Conducting and Analyzing Empirical Studies in Search-based Software Engineering”
SSBSE 2011

Statistics as Principled Argument



statistics as evidence supporting an argument

... and this argument should be part of an engaging narrative about the research

Robert P. Abelson, “Statistics as Principled Argument”, 1995

Abelson's MAGIC Criteria

Magnitude	the size of the quantitative support for the claim
Articulation	amount of comprehensible detail in the conclusions
Generality	are the conclusions broadly applicable?
Interestingness	how important is the result; does it change belief?
Credibility	methodological soundness and coherence with theory

Example - Effect of Reward

participant performs a very boring task

↓

researcher asks her to tell next subject that it was actually very interesting - will pay $1 / $20

↓

later participant rates task from -5 (very boring) to +5 (very interesting)

amount	mean response
$1	1.35
$20	-0.05

Festinger (1957) via Abelson

Example - Effect of Reward

Magnitude	difference in mean score of 1.40 t-test gave a p-value < 0.03
Articulation	“smaller reward causes greater tendency to change belief to conform to behaviour”
Generality	does it only occur in this specific situation? only under lab conditions?
Interestingness	counter to expectations
Credibility	coherency with cognitive dissonance theory; some minor questions as to statistical analysis; open to other interpretations

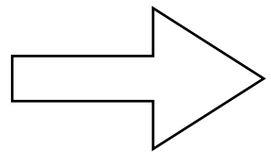
Abelson's 'Rule' of Two Criticisms

Magnitude

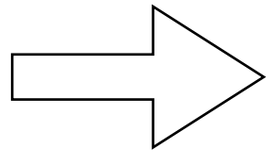
Articulation

Generality

Interestingness

2 or more criticisms  rejection

Credibility

lacks credibility  debate

Abelson: Rhetorical Style

My contention ...

statistical analysis more widely used in the field
(still room for improvement)

argument and narrative could be better

reviews often too conservative

Characteristics of SBST empirical research

SE research

stochastic algorithms

algorithms with many parameters

incredible variety of software

can often automate experiments

Magnitude

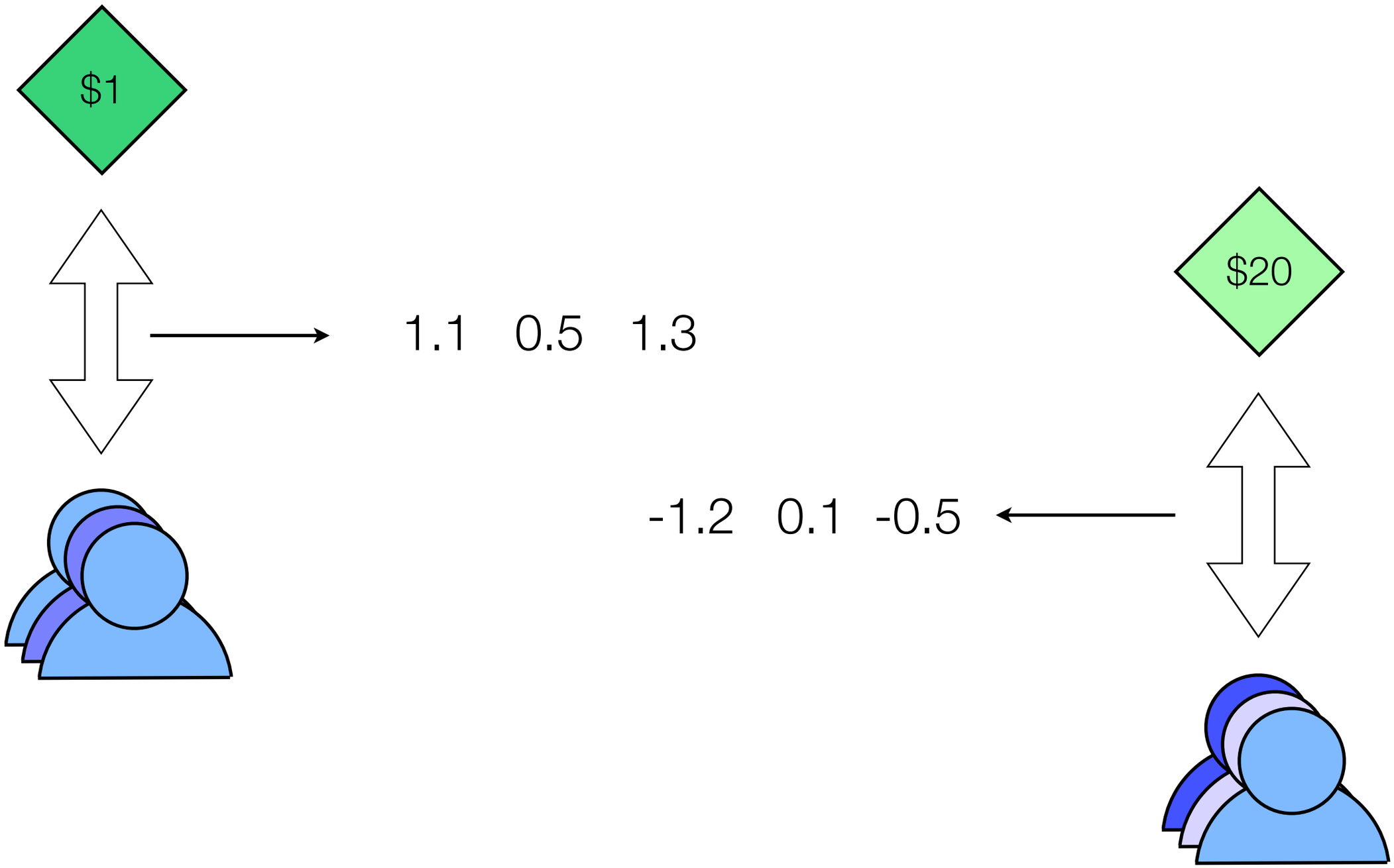
Articulation

Generality

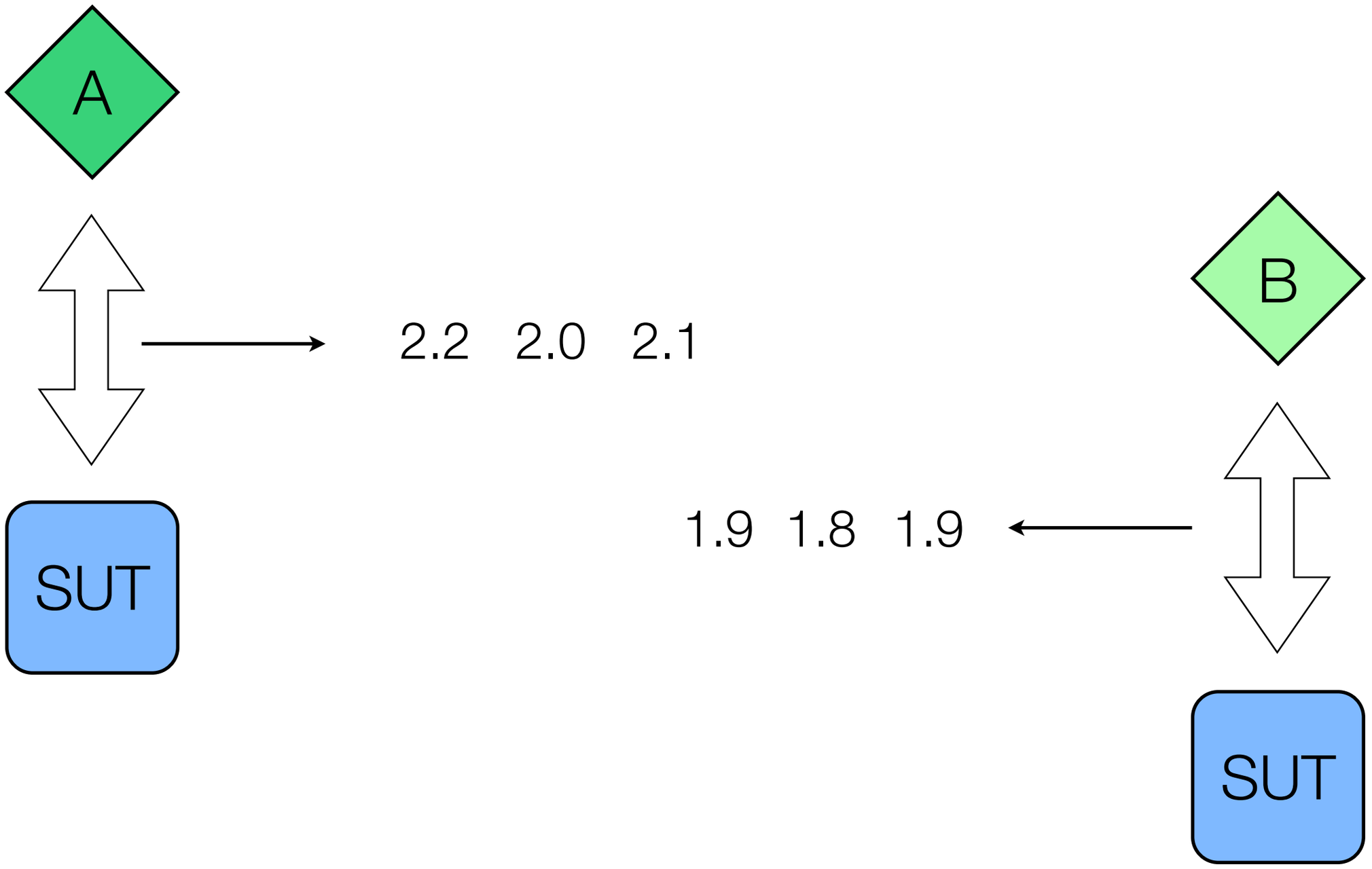
Interestingness

Credibility

Effect of Reward



SBST Comparison Experiment



Hypothesis Testing

Three possible explanations for differences:

(1) only a systematic factor

(2) only a chance factor

(3) a combination of both systematic and chance factors

Hypothesis Testing

Assume that difference in responses caused by:

(2) only a chance factor

Calculate how unlikely observed data is under this assumption.

If observed data is unlikely, then assumption is wrong, and so this is evidence that difference in responses caused by:

(3) a combination of both systematic and chance factors

Aside: Parametric or Non-Parametric?

	Parametric Tests	Non-Parametric Tests
	assume data has a specific (Normal) distribution 	less restrictive assumptions about distribution 
		can be less ‘efficient’? 
		not all parametric tests have a non-parametric equivalent 
Hypothesis Testing	Student’s t-test	Mann-Whitney-Wilcoxon ranksum test

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

$p = 2.3 \times 10^{-7}$

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

$p = 2.3 \times 10^{-7}$

$p = 0.0076$

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

$p = 2.3 \times 10^{-7}$

$p = 0.0076$

$p = 0.041$

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

$p = 2.3 \times 10^{-7}$

$p = 0.0076$

$p = 0.041$

$p = 0.052$

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

$p = 2.3 \times 10^{-7}$

$p = 0.0076$

$p = 0.041$

$p = 0.052$

$p = 0.081$

p-value

probability of obtaining observed data, or more extreme data, under assumption of only a chance factor

$p = 2.3 \times 10^{-7}$

$p = 0.0076$

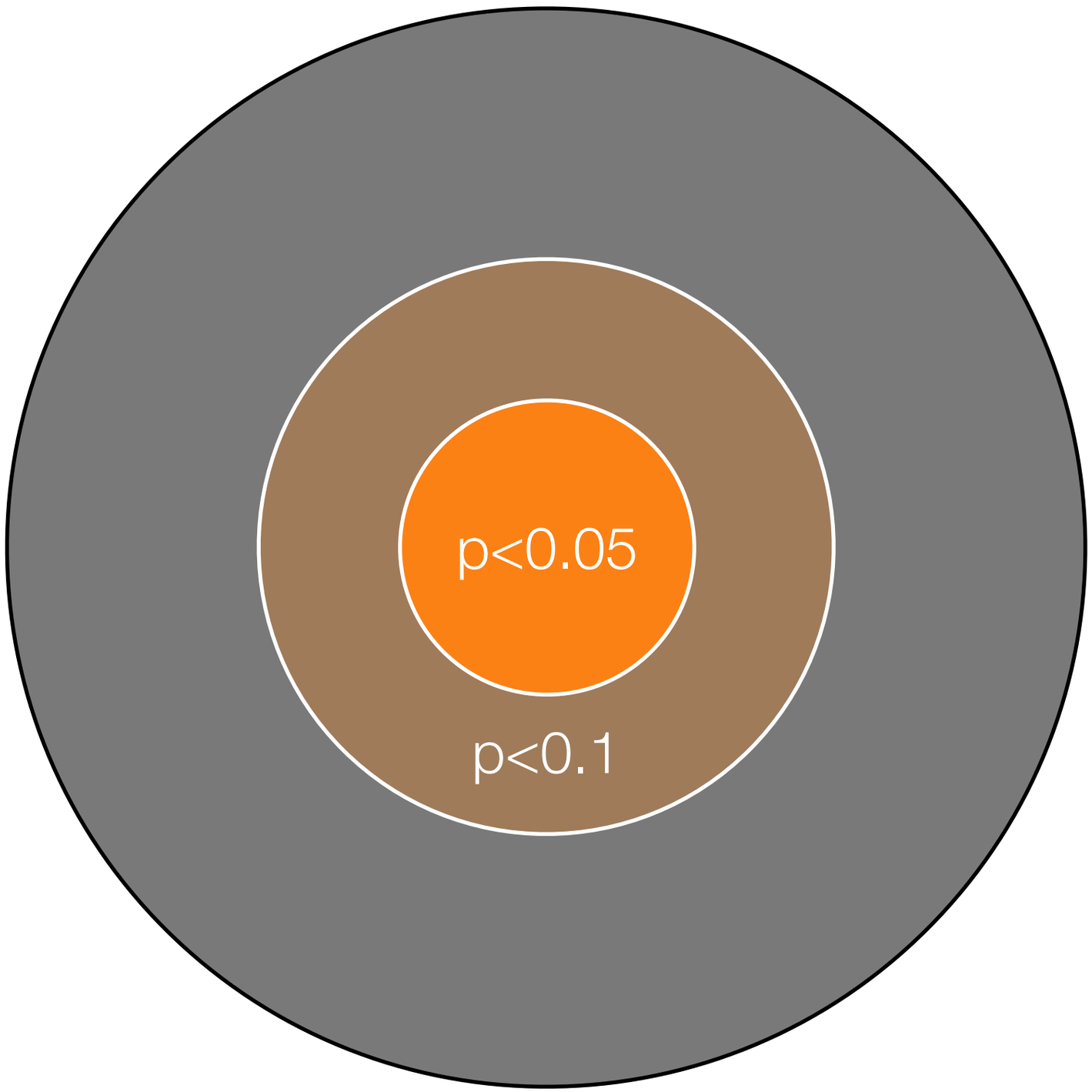
$p = 0.041$

$p = 0.052$

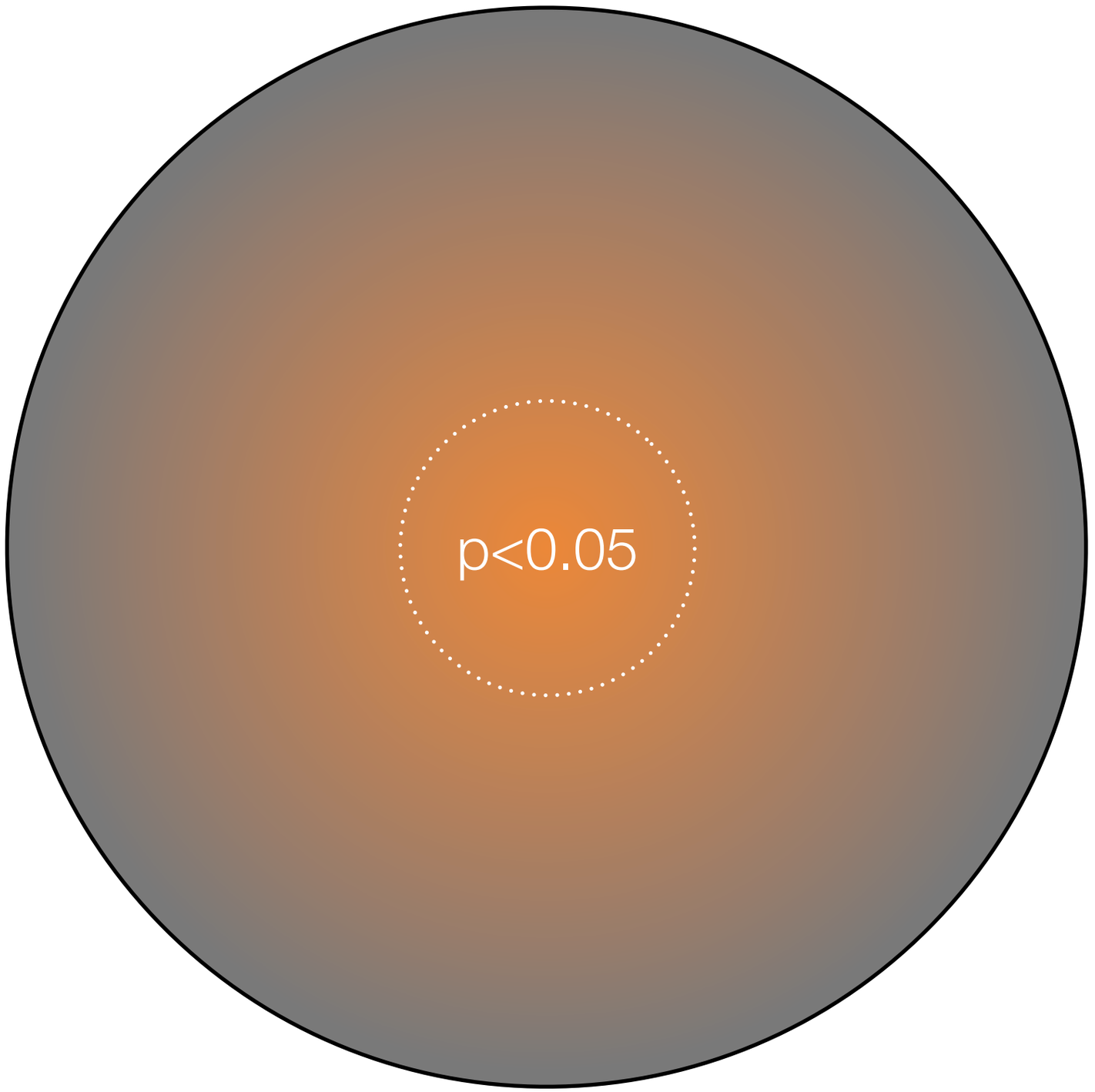
$p = 0.081$

$p = 0.37$

Over-categorisation



Over-categorisation



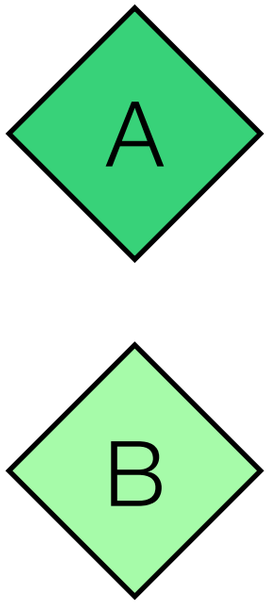
In Context of Argument

Experiment 1 - Coverage — p = 0.03

Experiment 2 - Fault-Detection — p = 0.02

Experiment 3 - Speed — p = 0.06

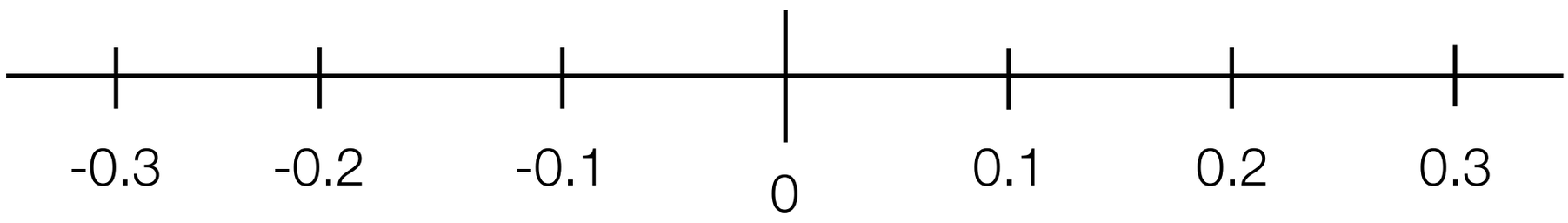
Alternative: Confidence Limits



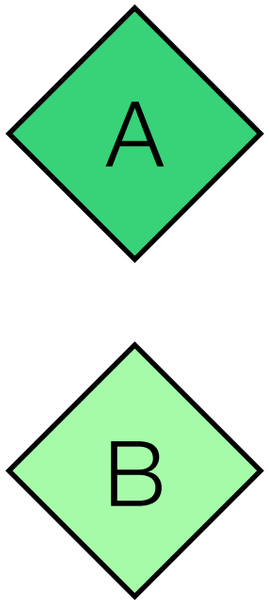
2.2 2.0 2.1 1.9 2.2 2.1 1.9 1.8

1.9 1.8 1.9 2.0 1.9 1.7 1.6 2.0

95% confidence limit for mean difference in response:



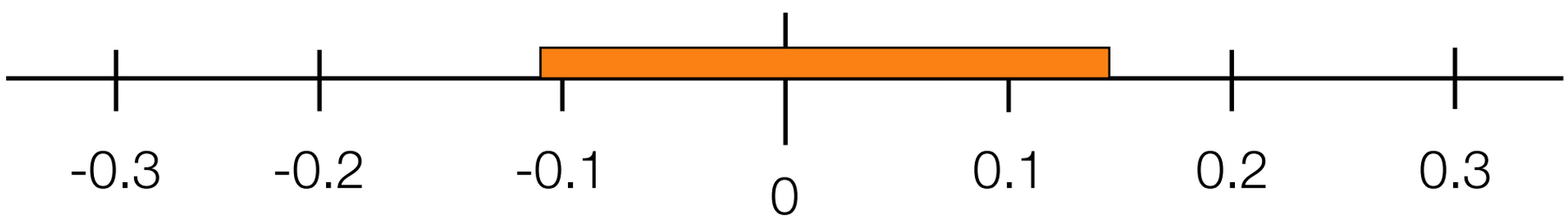
Alternative: Confidence Limits



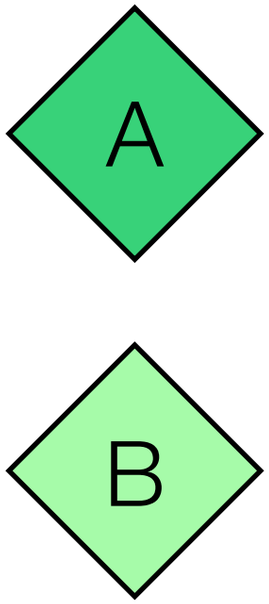
2.2 2.0 2.1 1.9 2.2 2.1 1.9 1.8

1.9 1.8 1.9 2.0 1.9 1.7 1.6 2.0

95% confidence limit for mean difference in response:



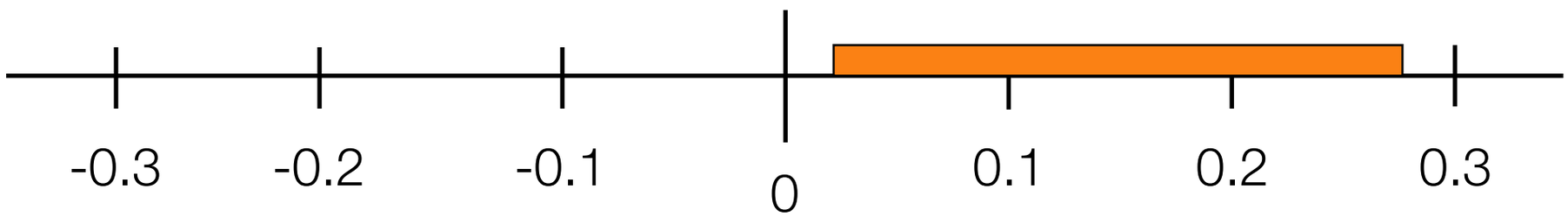
Alternative: Confidence Limits



2.2 2.0 2.1 1.9 2.2 2.1 1.9 1.8

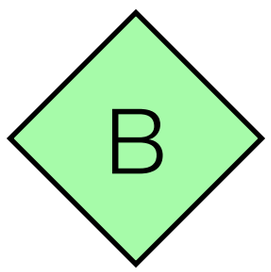
1.9 1.8 1.9 2.0 1.9 1.7 1.6 2.0

95% confidence limit for mean difference in response:



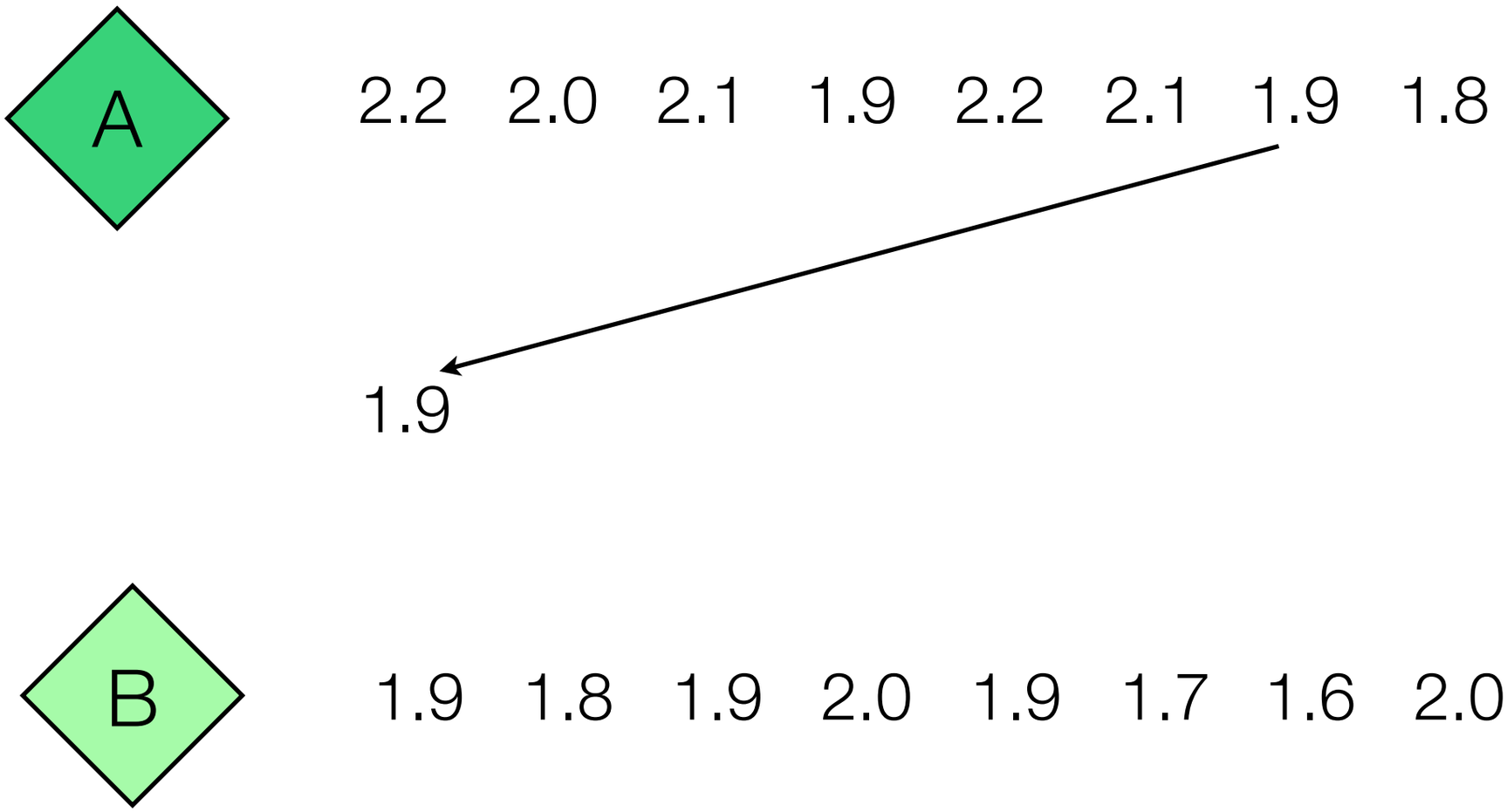
Bootstrapping

2.2 2.0 2.1 1.9 2.2 2.1 1.9 1.8

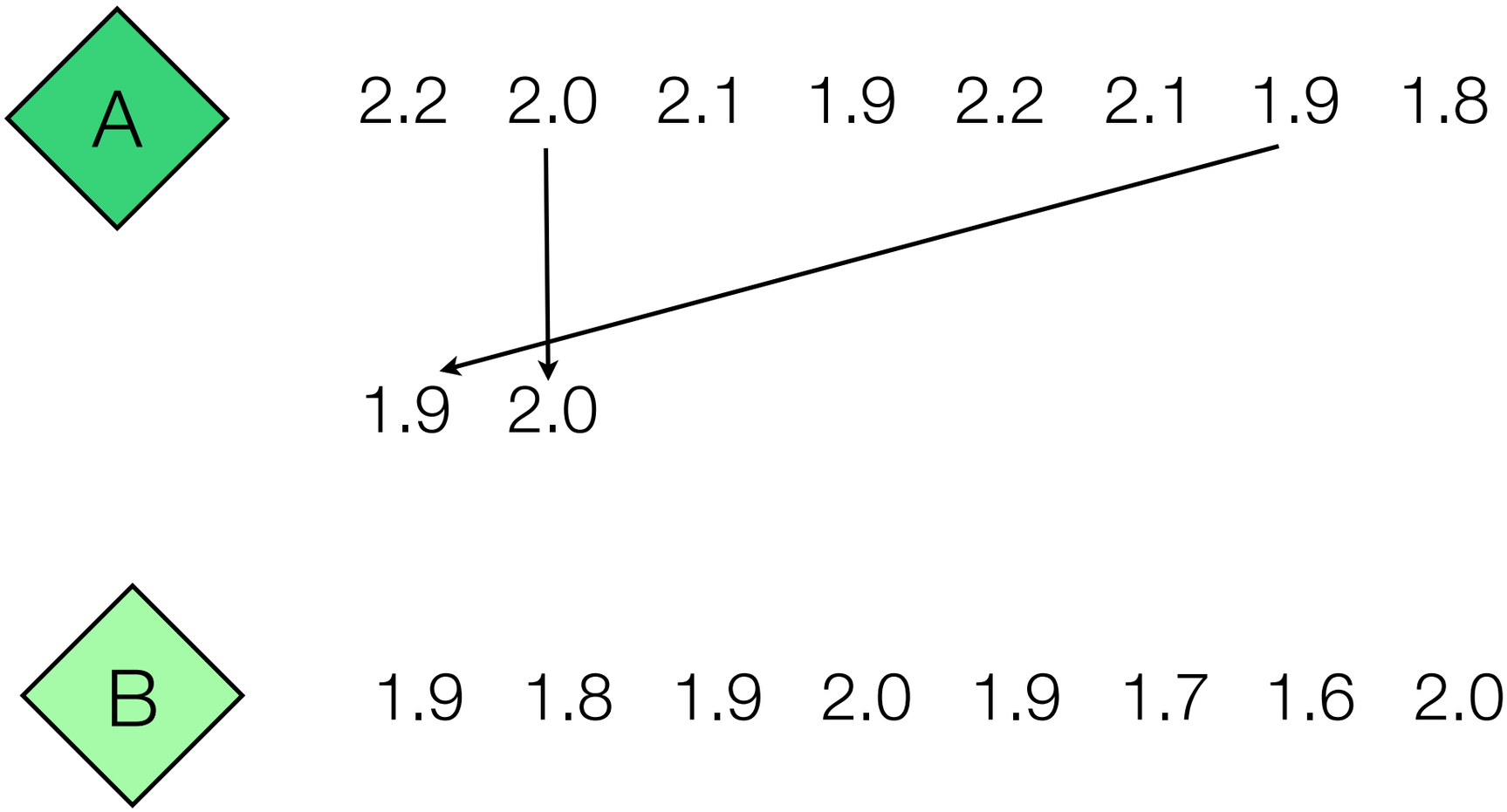


1.9 1.8 1.9 2.0 1.9 1.7 1.6 2.0

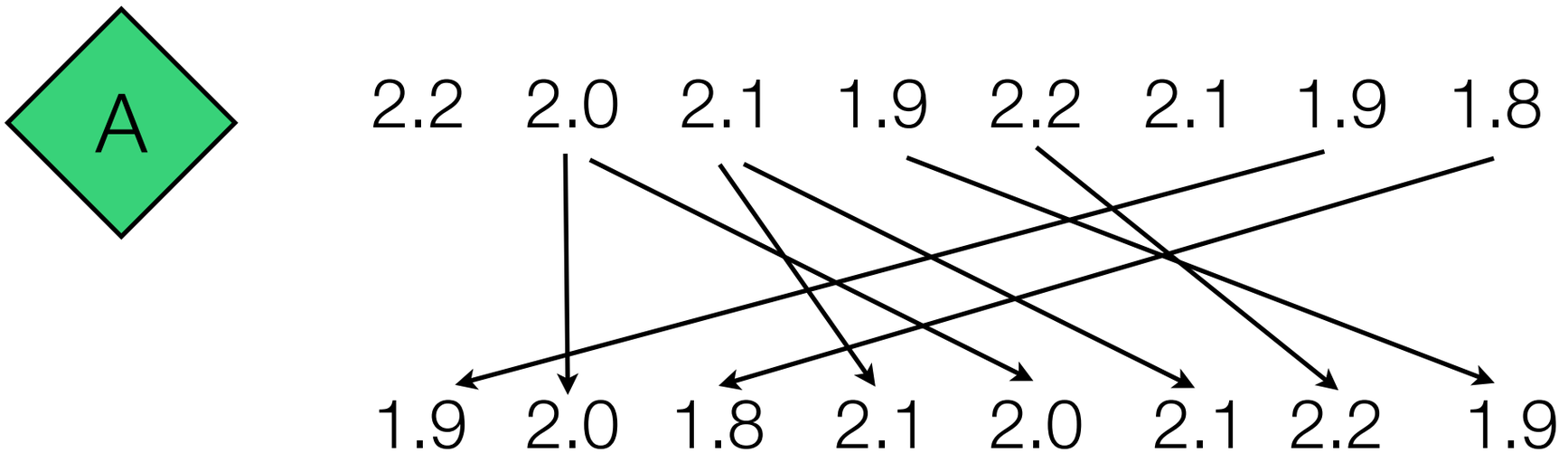
Bootstrapping

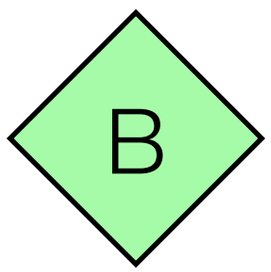


Bootstrapping



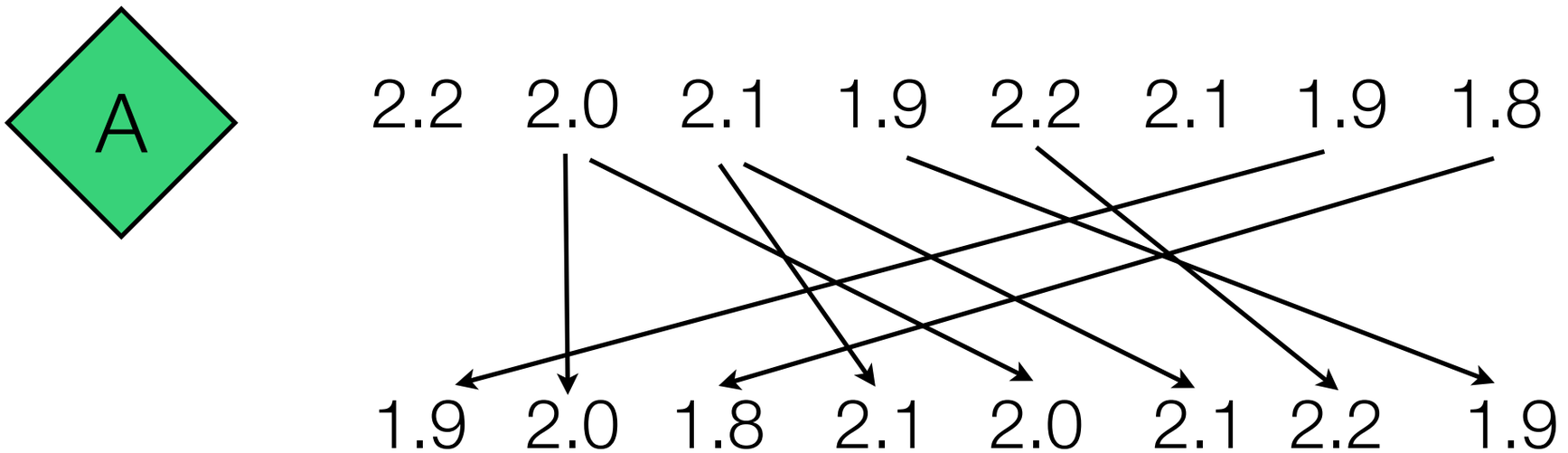
Bootstrapping

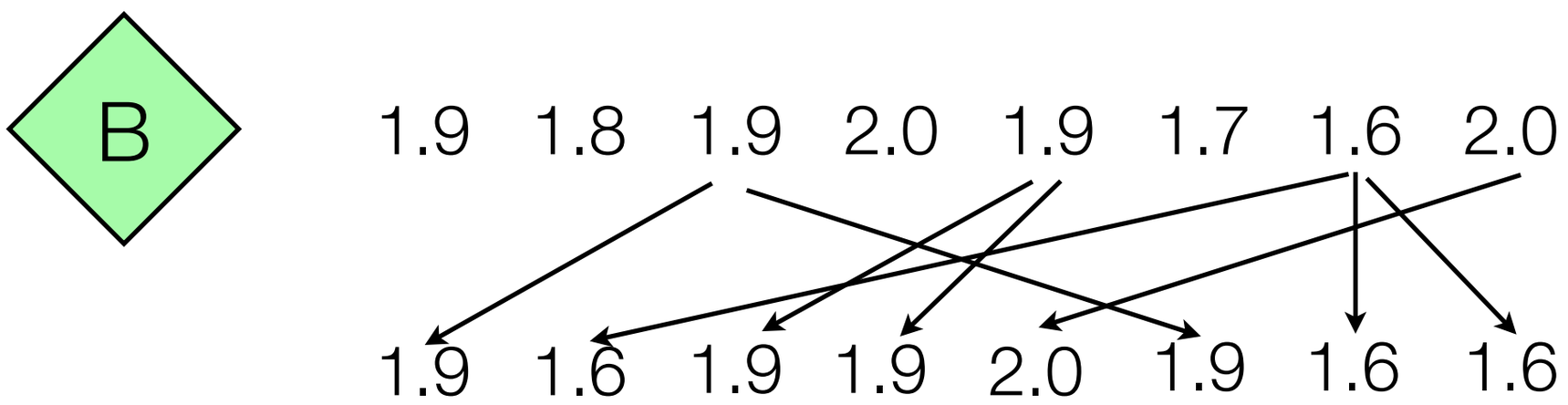




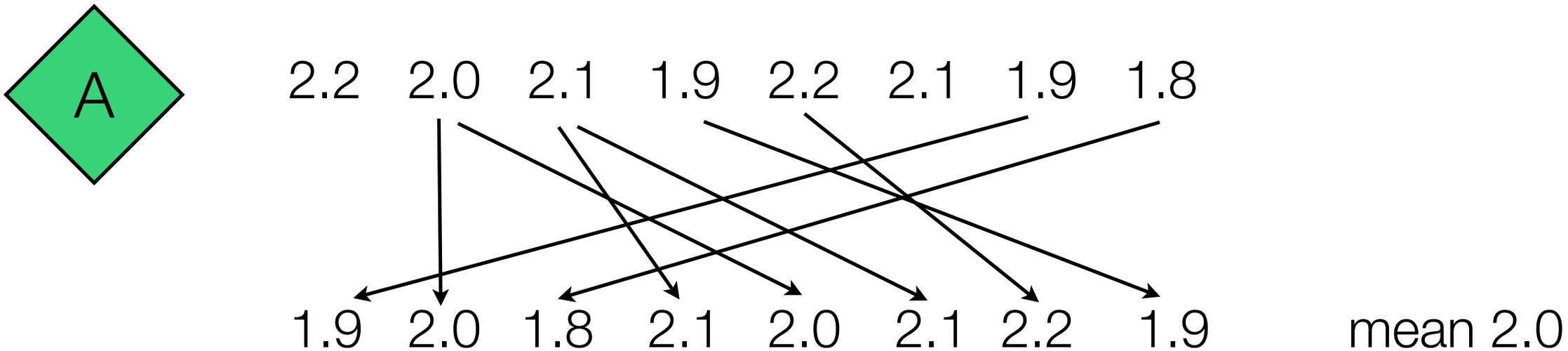
1.9 1.8 1.9 2.0 1.9 1.7 1.6 2.0

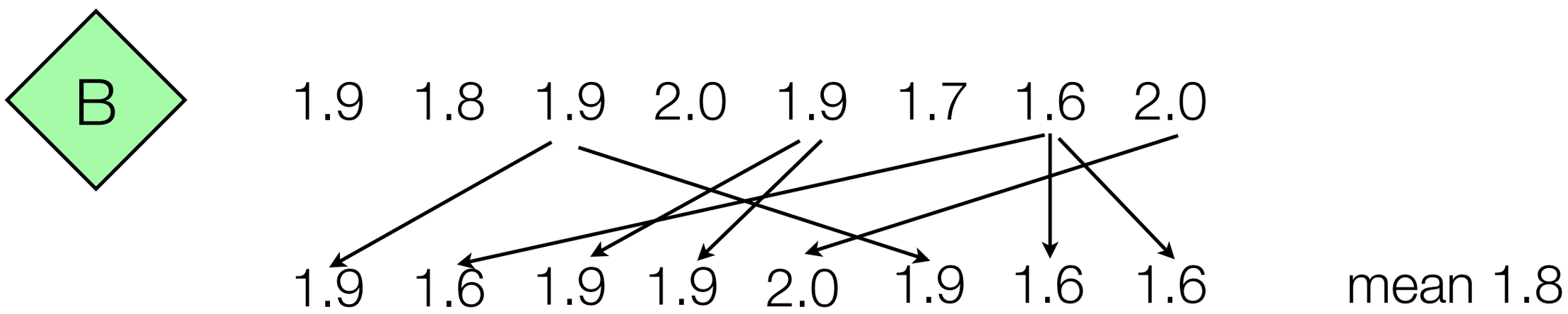
Bootstrapping





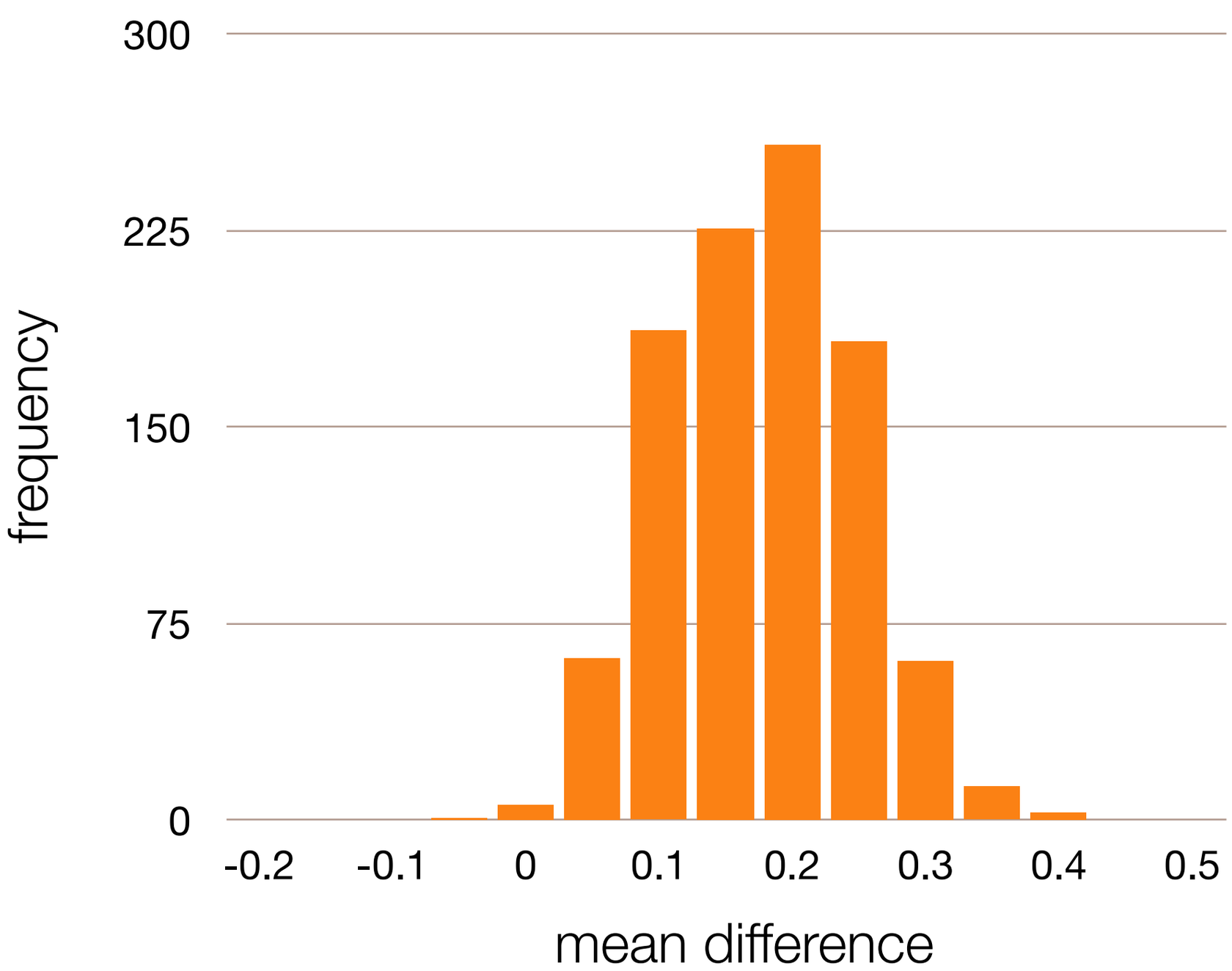
Bootstrapping





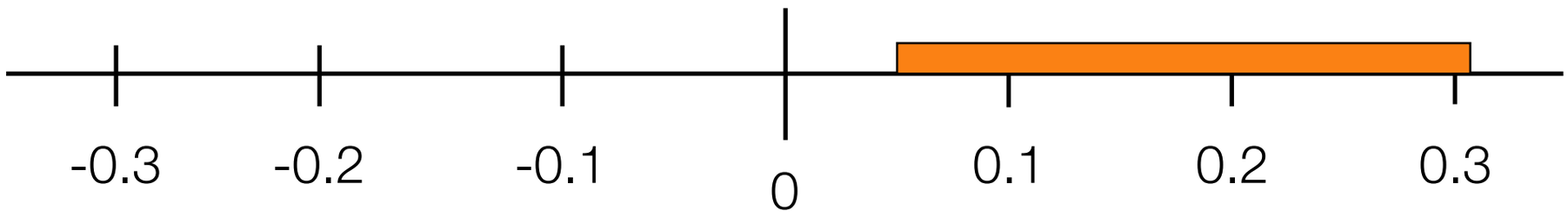
difference 0.2

Bootstrapping



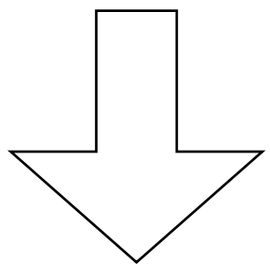
Bootstrapping

95% confidence limit: 0.05 to 0.3125

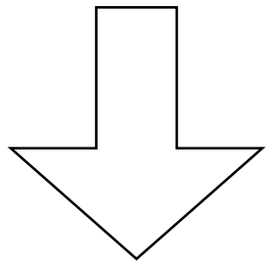


Very Large Samples

automated experimental trials



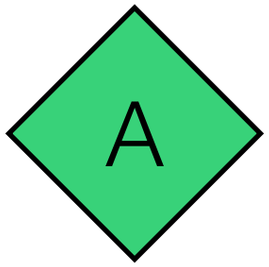
very large sample sizes



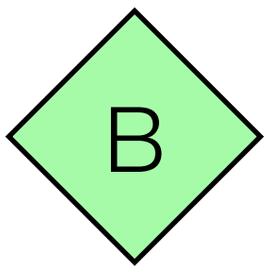
always demonstrate significance

Effect Size

raw effect size — e.g. mean difference in algorithm response



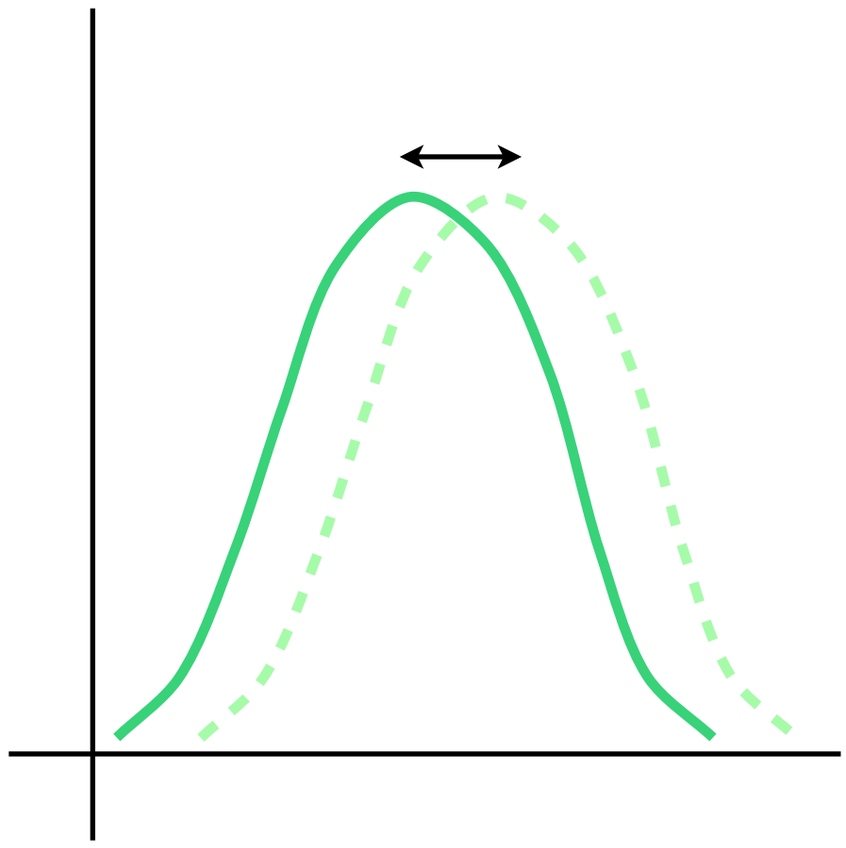
2.2 2.0 2.1 1.9 2.2 2.1 1.9 1.8

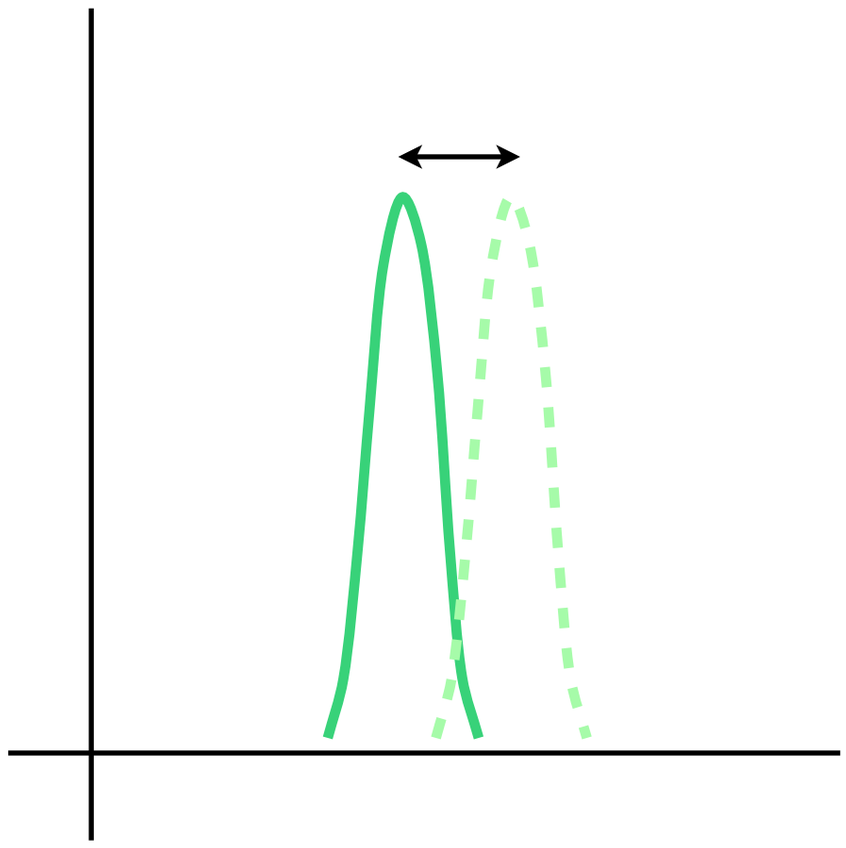


1.9 1.8 1.9 2.0 1.9 1.7 1.6 2.0

mean a - mean b = 0.175

Standardised Effect Size

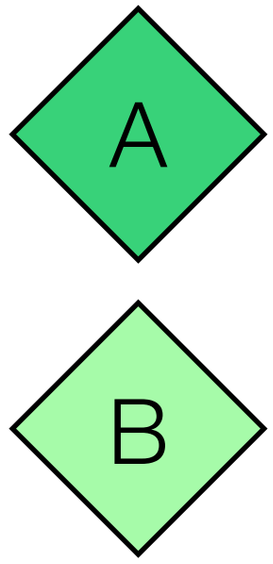




Standardised Effect Size - Parametric

Cohen's d

$$\frac{\text{raw effect size}}{\text{standard deviation}}$$



$$d = \frac{0.175}{0.1358} = 1.29$$

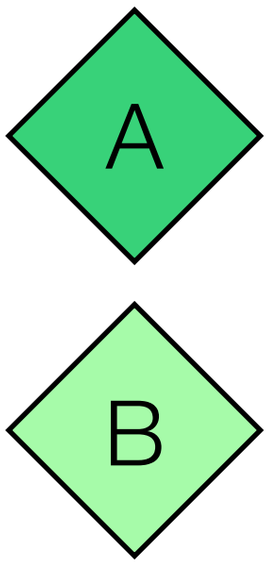
Standardised Effect Size - Non-Parametric

Vargha-Delaney A

calculated easily from ranksum statistic and sample size

values between 0 and 1; 0.5 when no effect

very nice real-world interpretation in terms of single samples



A = 0.79

Vargha and Delaney, “A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong”, J. Educational and Behavioral Statistics, 2001

Guidelines

	Cohen's d	Vargha-Delaney A
'small'	0.2	0.56
'medium'	0.5	0.64
'large'	0.8	0.71

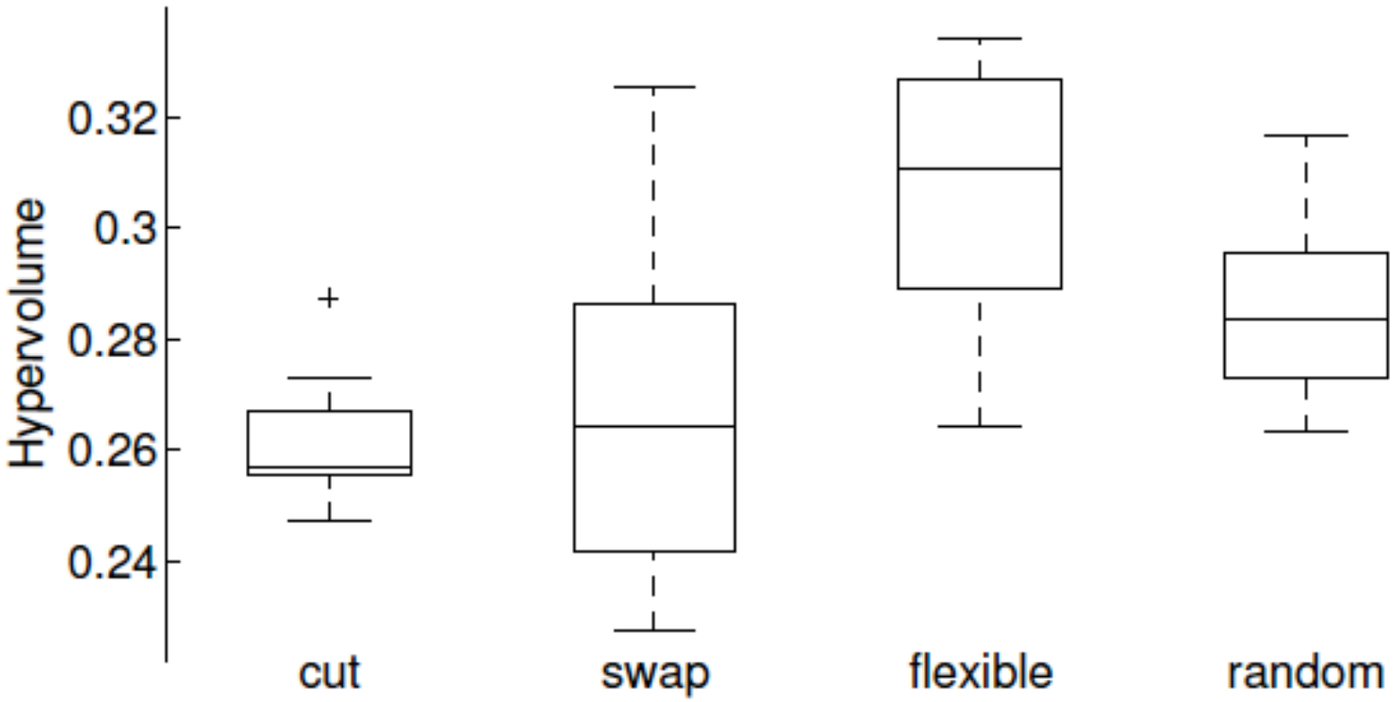
SBST/ICST 2012

Table II
FOR EACH PROJECT, AVERAGE COVERAGE ON ALL OF ITS CLASSES WHEN NO BYTECODE CONSTANT IS USED ($P_{\text{BYTECODE}} = 0$) AND WHEN THEY ARE USED WITH PROBABILITY $P_{\text{BYTECODE}} = 0.2$. THE $\hat{A}_{12}$ OF THESE COMPARISONS ARE CALCULATED BY AGGREGATING ALL RUNS OF ALL CLASSES PER PROJECT (IN BOLD IF STATISTICALLY SIGNIFICANT). ON HIGHER GRANULARITY, IT IS REPORTED THE PERCENTAGE % OF CLASSES FOR WHICH WE HAVE A SIGNIFICANT $\hat{A}_{12} > 0.5$ AND $\hat{A}_{12} < 0.5$.

Project	$P = 0$	$P = 0.2$	$\hat{A}_{12}$	$\% > 0.5$	$\% < 0.5$
COL	0.74	0.73	**0.48**	0.04	0.27
CCL	0.87	0.90	**0.56**	0.21	0.07
CCD	0.87	0.88	0.52	0.29	0.00
CCO	0.91	0.91	0.50	0.02	0.01
CMA	0.75	0.75	0.51	0.12	0.02
CPR	0.93	0.94	**0.52**	0.15	0.005
GCO	0.74	0.74	0.50	0.04	0.01
ICS	0.85	0.86	0.50	0.05	0.00
JCO	0.82	0.82	0.50	0.07	0.00
JDO	0.73	0.73	0.50	0.12	0.00
JGR	0.75	0.75	0.50	0.03	0.01
JTI	0.84	0.85	**0.52**	0.18	0.00
NXM	0.59	0.59	0.51	0.00	0.00
NCS	0.97	0.97	0.51	0.09	0.00
REG	0.75	0.75	0.50	0.00	0.00
SCS	0.63	0.85	**0.77**	0.75	0.00
TRO	0.88	0.87	**0.46**	0.005	0.32
XEN	0.65	0.72	**0.57**	0.29	0.00
XOM	0.76	0.77	0.51	0.17	0.00
ZIP	0.80	0.83	**0.69**	1.00	0.00
Average	0.79	0.81	0.53	0.18	0.04

Fraser and Arcuri, “The Seed is Strong: Seeding Strategies in Search-Based Software Testing”, ICST 2012

SBST/ICST 2012



(a) Comparison of the crossover operators.

Woehrle, “Search-based Stress Testing of Wireless Network Protocol Stacks”, SBST 2012

Magnitude

Articulation

Generality

Interestingness

Credibility

Abelson's 'Ticks', 'Buts', and 'Blobs'

Tick "claim of specific comparative difference"

But "qualify or constrain ticks"

Blob "cluster undifferentiated research results"

Examples

"Algorithm A is better than B"

"Algorithm A is better than B, and the larger cyclometric complexity of the code, the greater the improvement"

"Algorithm A is better than B, except for object-oriented software"

"For SUTs 1 and 4, A is better than B. For SUT 2, there is no significant difference. For SUTs 3, 5, and 6, B is better than A, although the difference isn't significant for SUT 5."

SBST/ICST 2012

"Our experiments show with strong statistical confidence that, even for a testing tool that is already able to achieve high coverage, the use of appropriate seeding strategies can further improve performance."

Fraser and Arcuri, "The Seed is Strong: Seeding Strategies in Search-Based Software Testing", ICST 2012

Magnitude

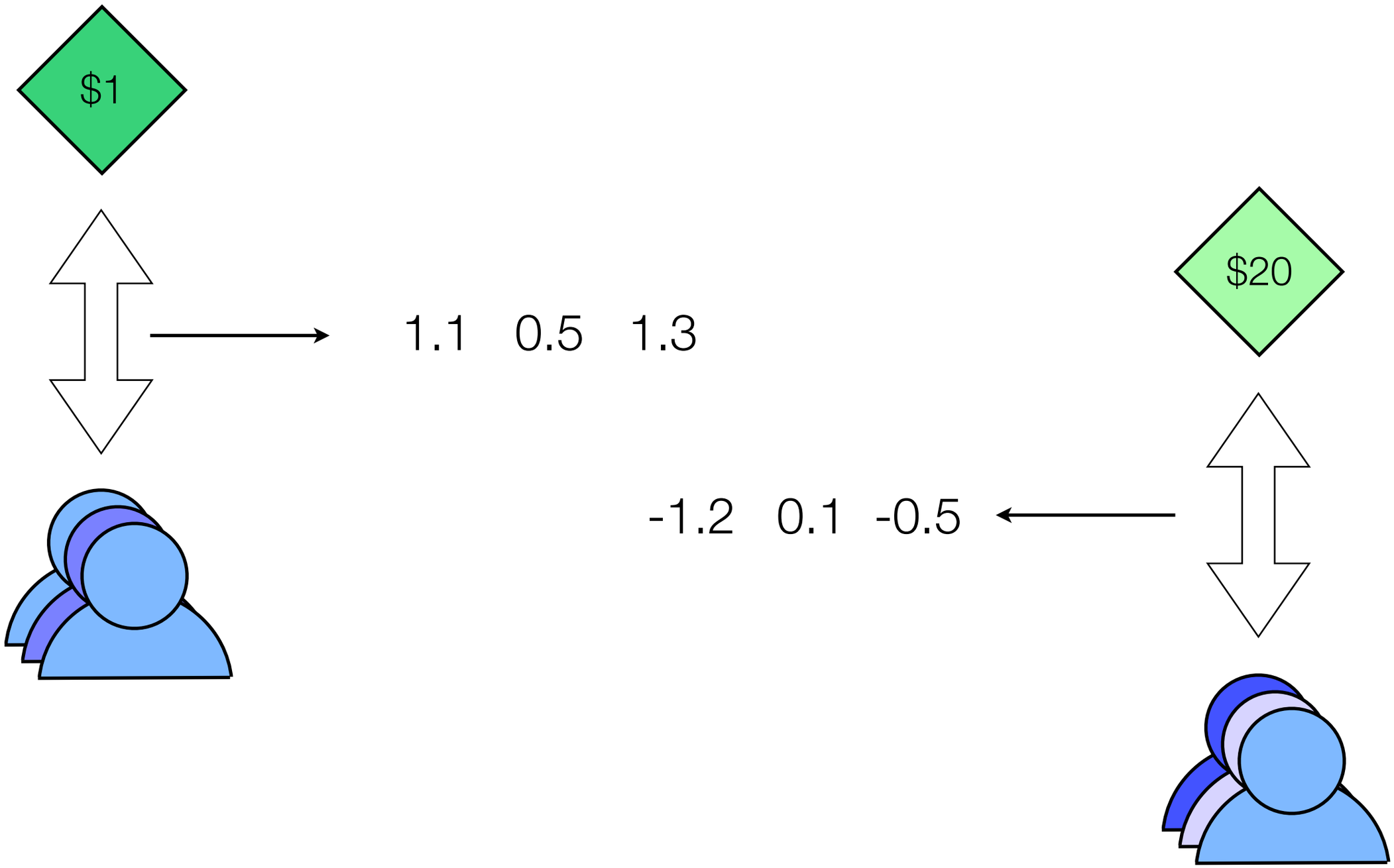
Articulation

Generality

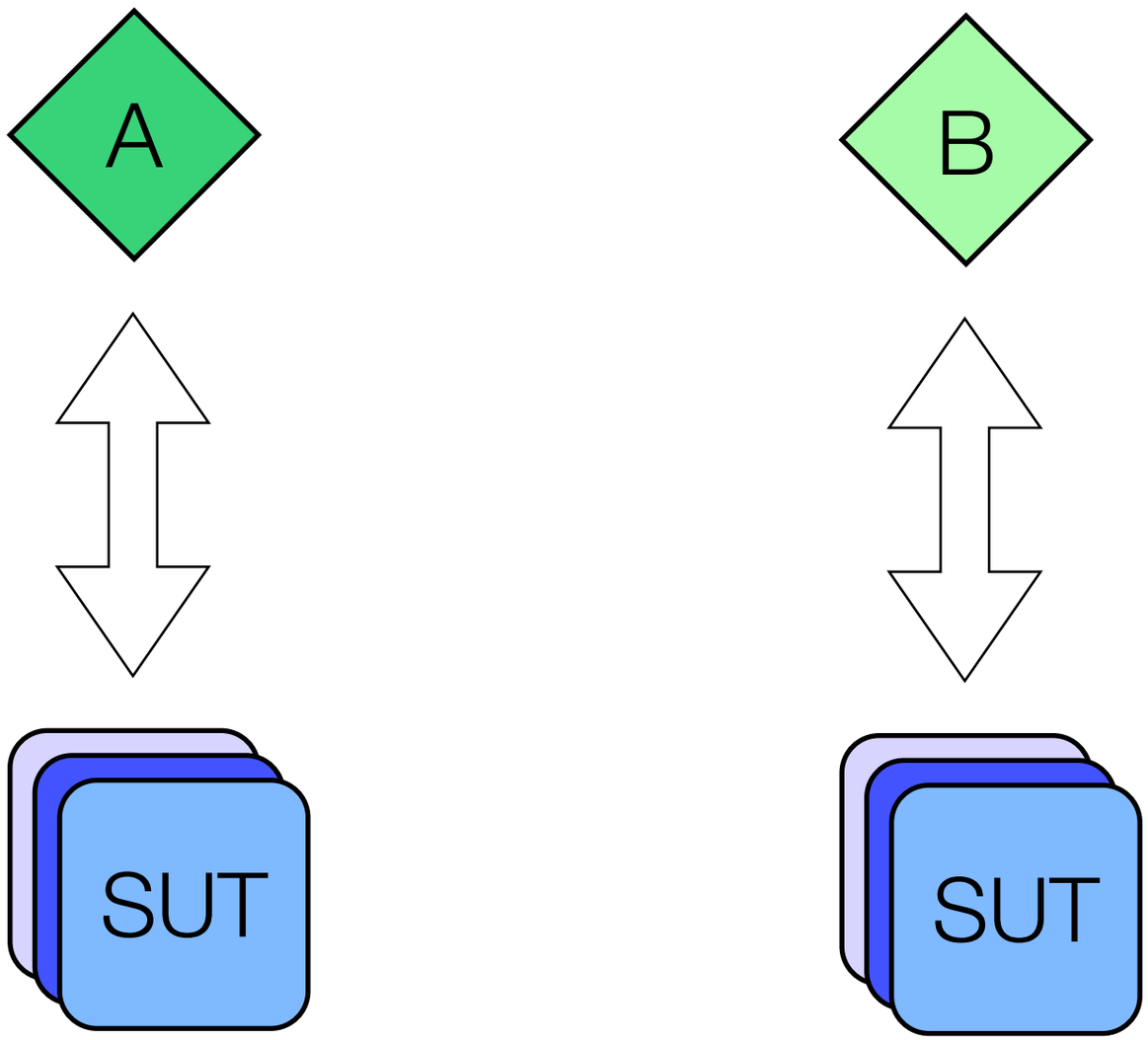
Interestingness

Credibility

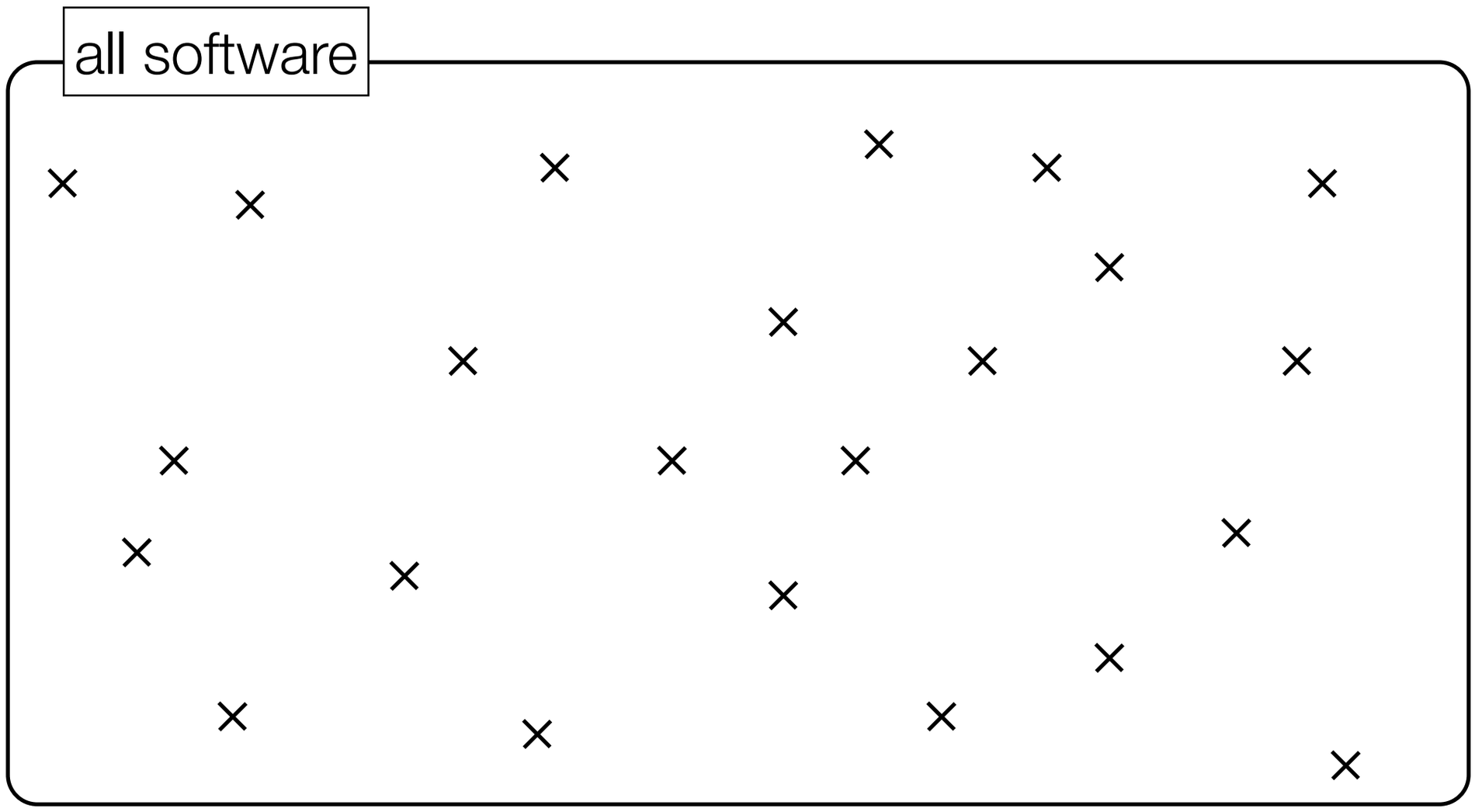
Effect of Reward



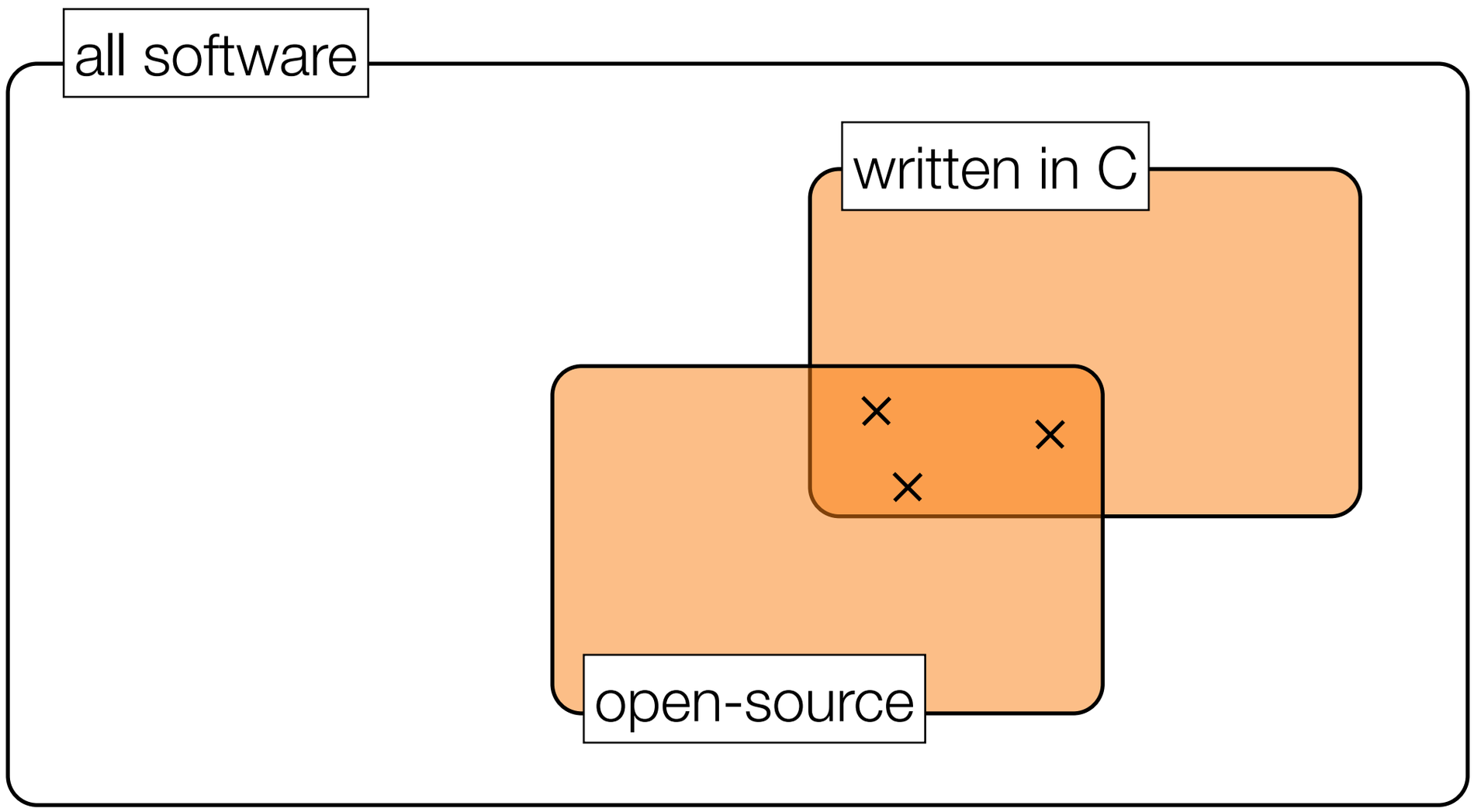
Representative Software



Representative Software & Selection Bias



Representative Software & Selection Bias



ICST/SBST 2012

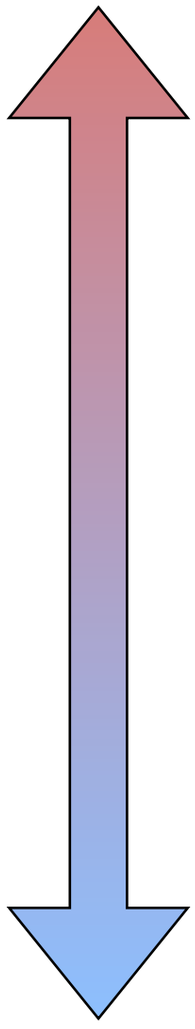
“ To avoid a bias in our case study class selection, we therefore randomly chose 20 classes out of the SF100 corpus of Java projects randomly selected from Sourceforge ...”

Pavlov and Fraser, “Semi-Automatic Search-based Test Generation”, SBST 2102

Style

"Algorithm A significantly better than B for ..."

liberal



conservative

"... software"

"... software written in C"

"... open-source software written in C"

"... a range of open-source software written in C"

" ... on the pieces of software we tried"

Benchmark SUTs

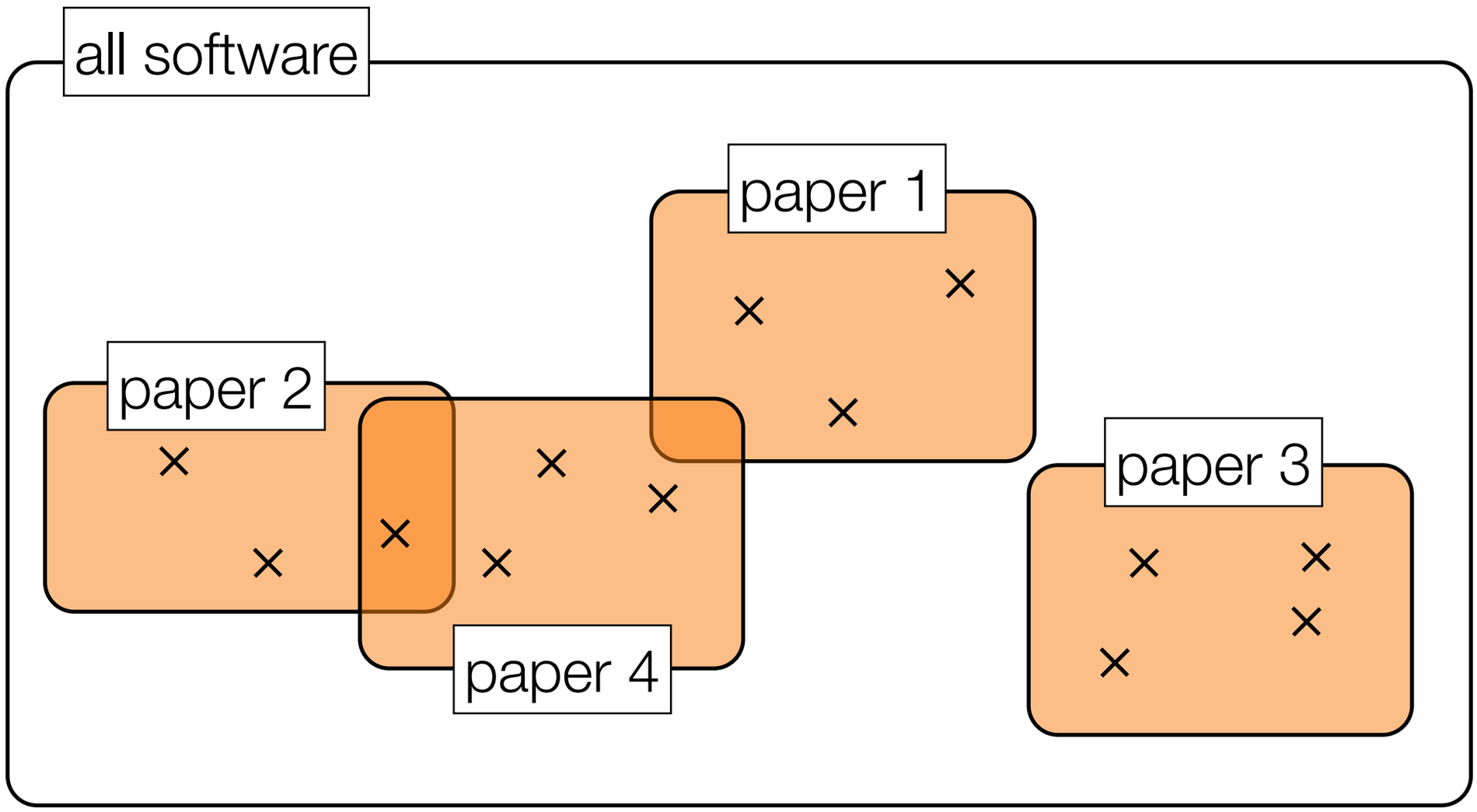
useful for comparison with previous results

research field becomes focussed on small set of software

benchmarks tend to be SUTs for which existing techniques do well

... emergent benchmarks possible only if SUTs are made available

Meta-Analysis



... depends on reproducibility of technique and access to experimental data

Magnitude

Articulation

Generality

Interestingness

Credibility

Interestingness

potential belief change:

strengthening / weakening existing beliefs

creating new beliefs

importance:

propositions (e.g. theories) requiring modification

SBST/ICST 2012

"This paper presents an approach in which examples of inputs are sought from the Internet by reformulating program identifiers into web queries."

McMinn, Shahbaz, and Stevenson, "Search-Based Test Input Generation for String Data Types Using the Results of Web Queries", ICST 2012

Abelson's Surprisingness Coefficient

m_{obs} observed magnitude

m_{exp} expected magnitude

$$S = \frac{(m_{obs} - m_{exp})^2}{|m_{obs}| + |m_{exp}|}$$

m_{exp}	m_{obs}	S
0	1.29	1.29
0.5	1.29	0.35
0.5	0.0	0.5
0.7	0.6	0.0077

Magnitude

Articulation

Generality

Interestingness

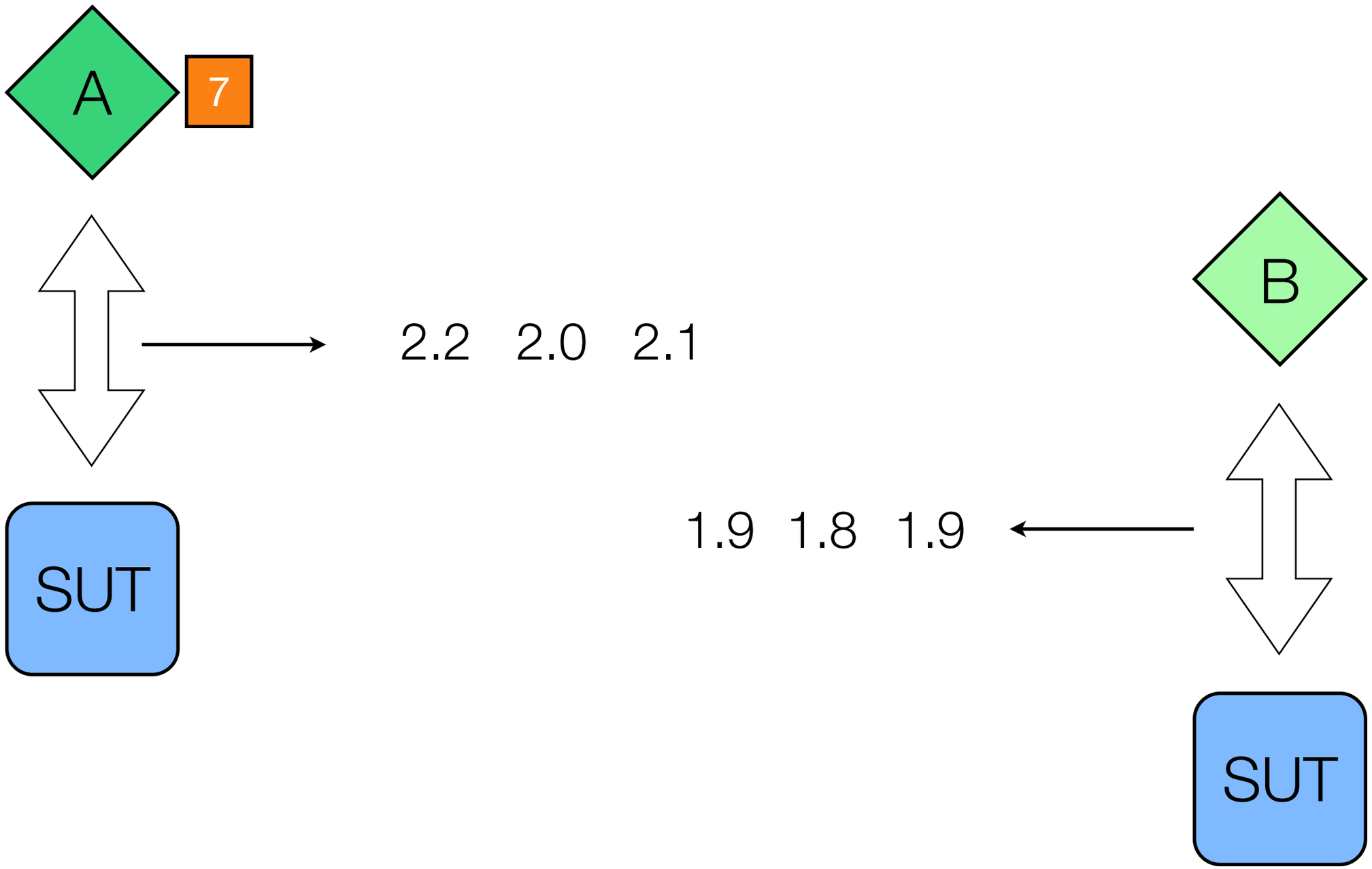
Credibility

Credibility

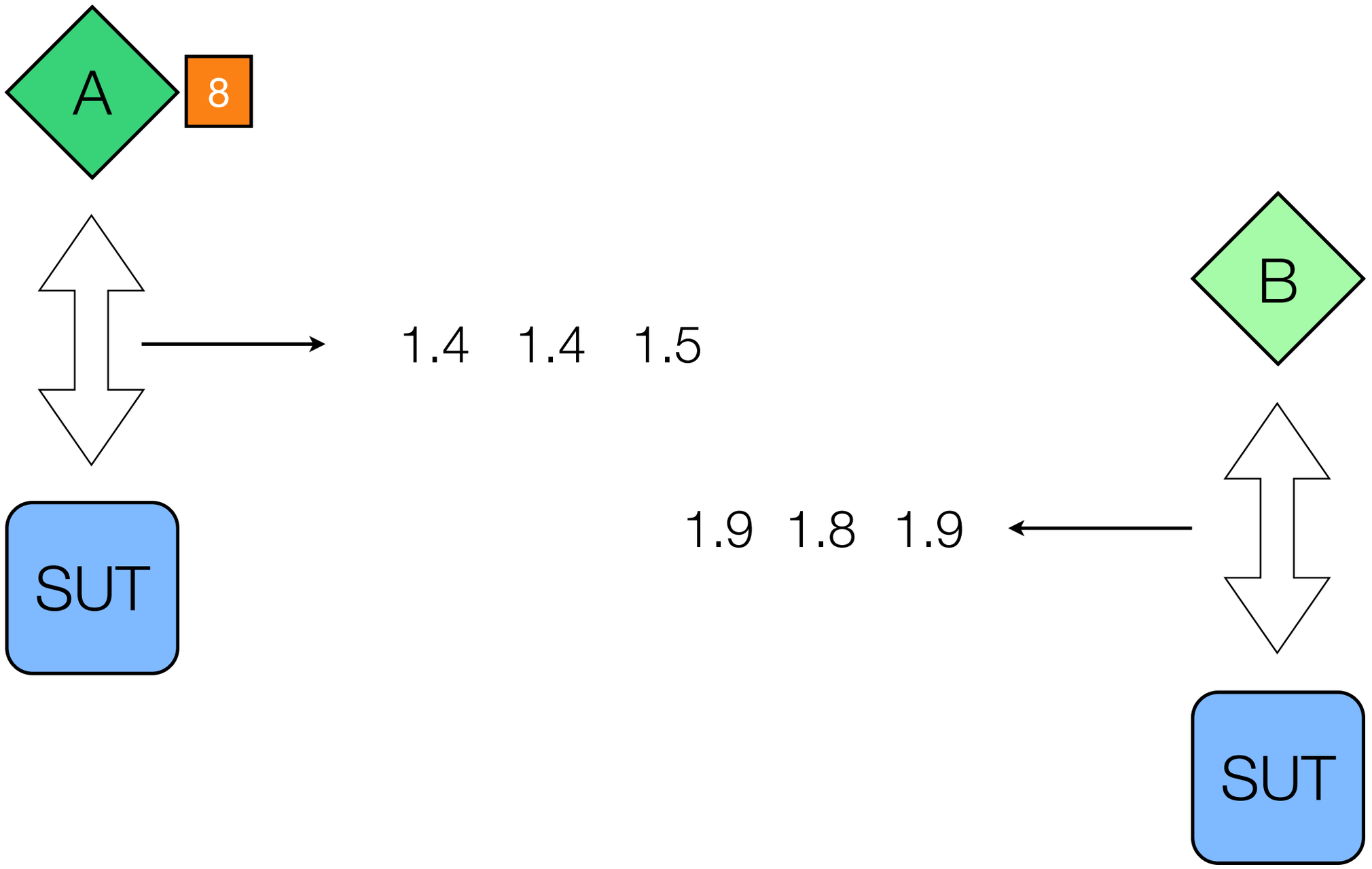
methodological soundness

coherence with theory

Algorithm Parameters

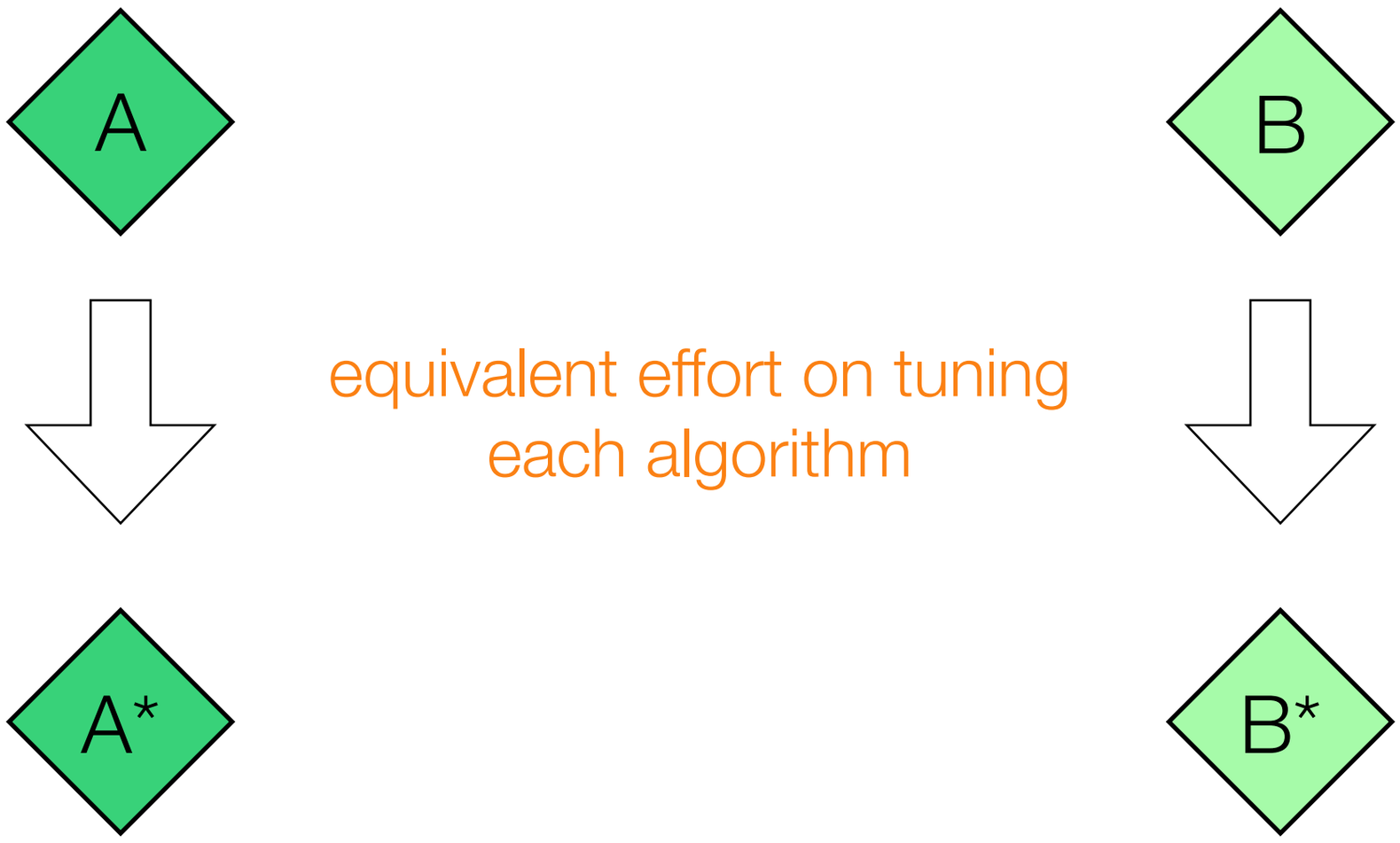


Algorithm Parameters

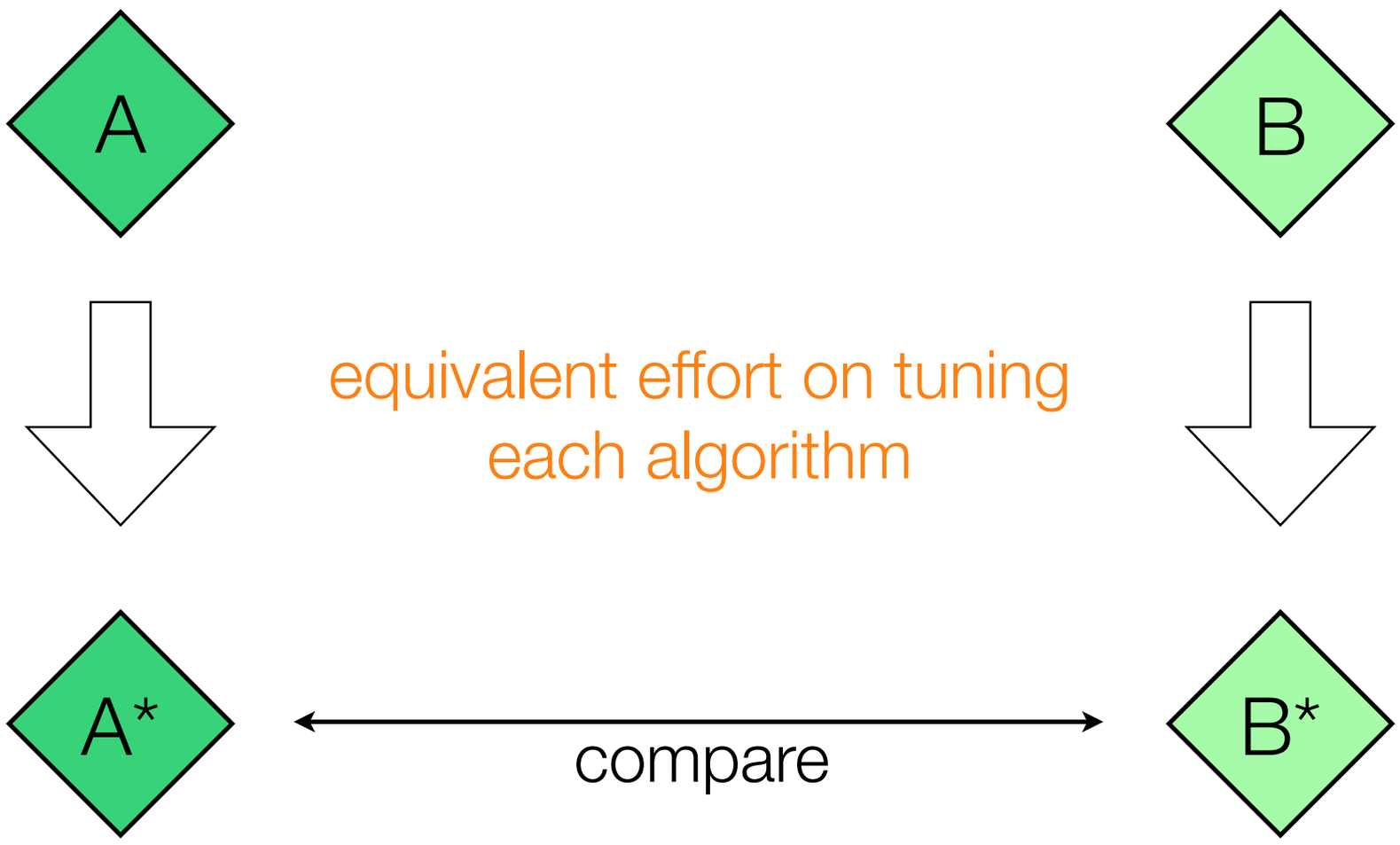


‘Best-Effort’ Tuning

‘Best-Effort’ Tuning



‘Best-Effort’ Tuning

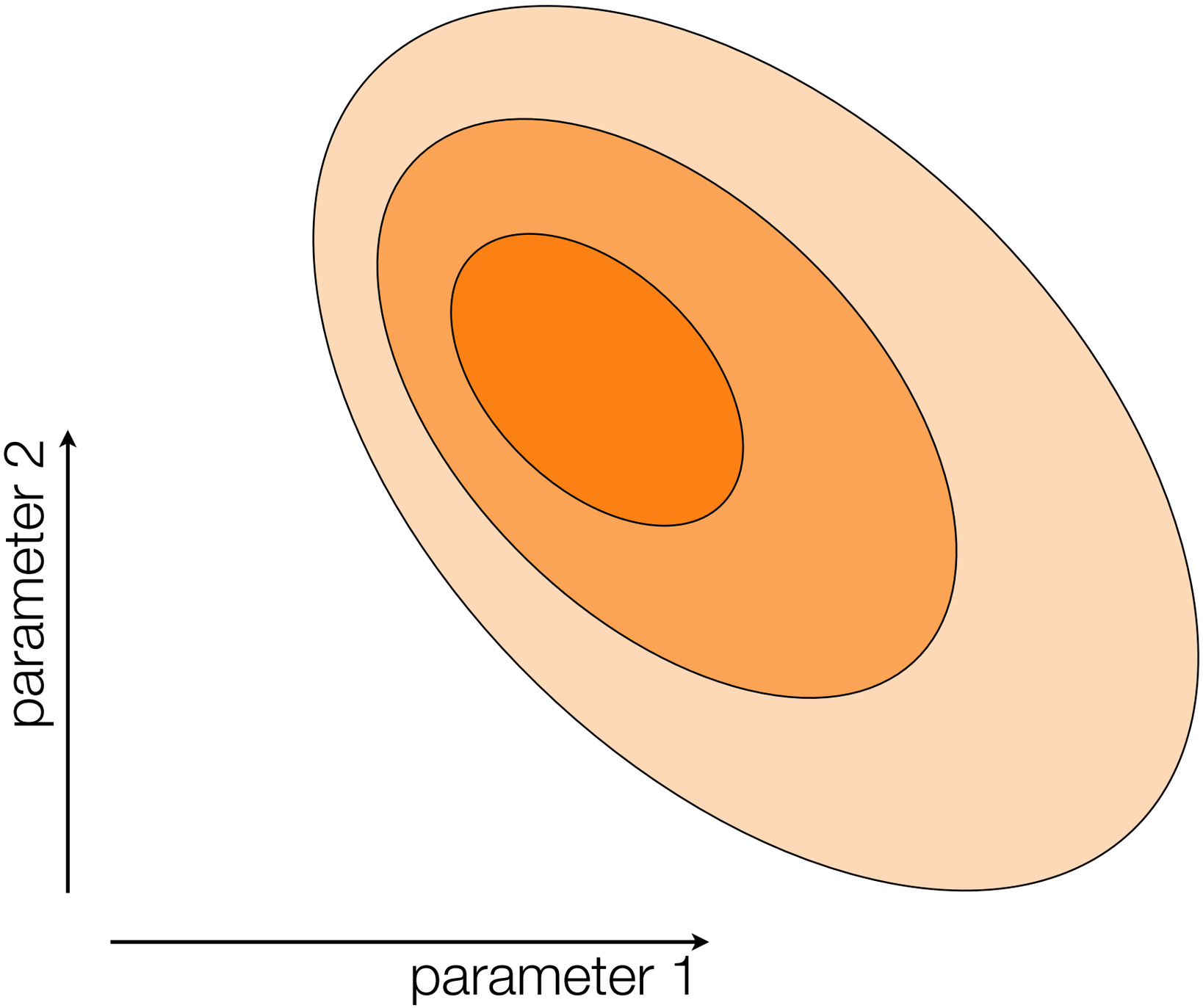


SBST/ICST 2012

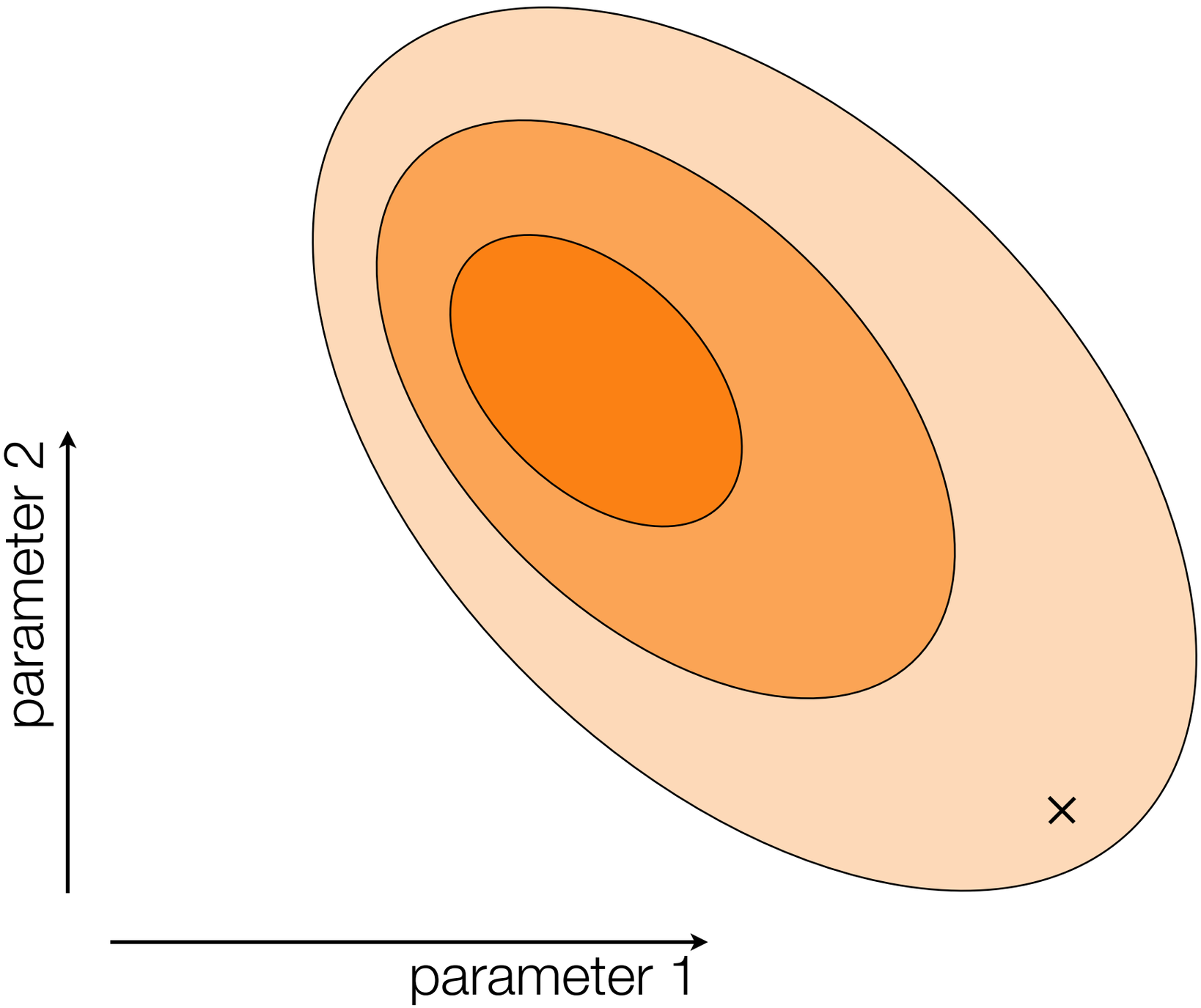
"One of the critical aspects of the application of search based techniques is finding the right configuration of parameters. We sampled the set of all parameter values combinations ... by selecting all pair-wise interactions between values ..."

Gómez, Baudry, Sahraoui, "Searching the boundaries of a modeling space to test metamodels", ICST 2012

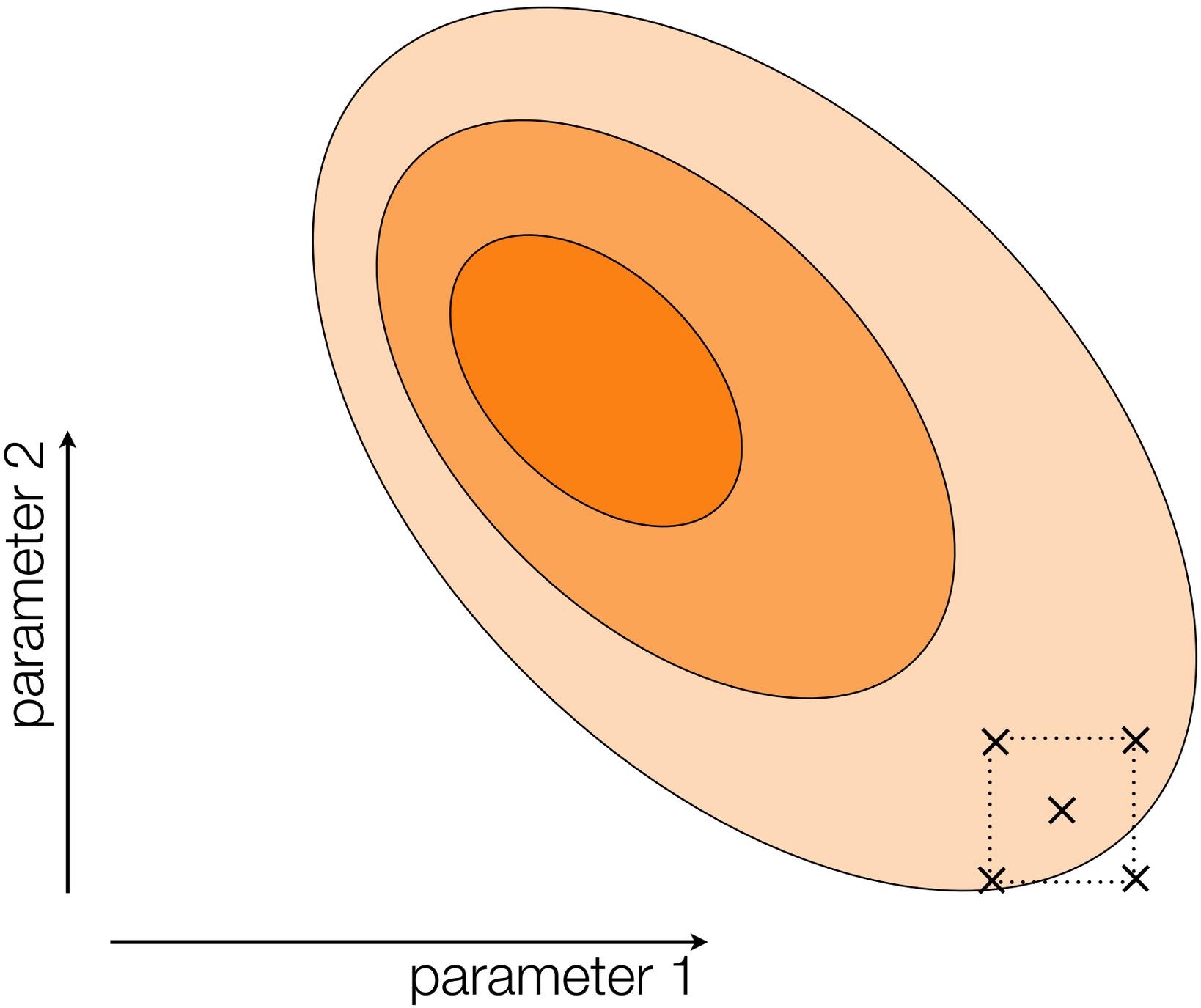
Response Surface Methodology



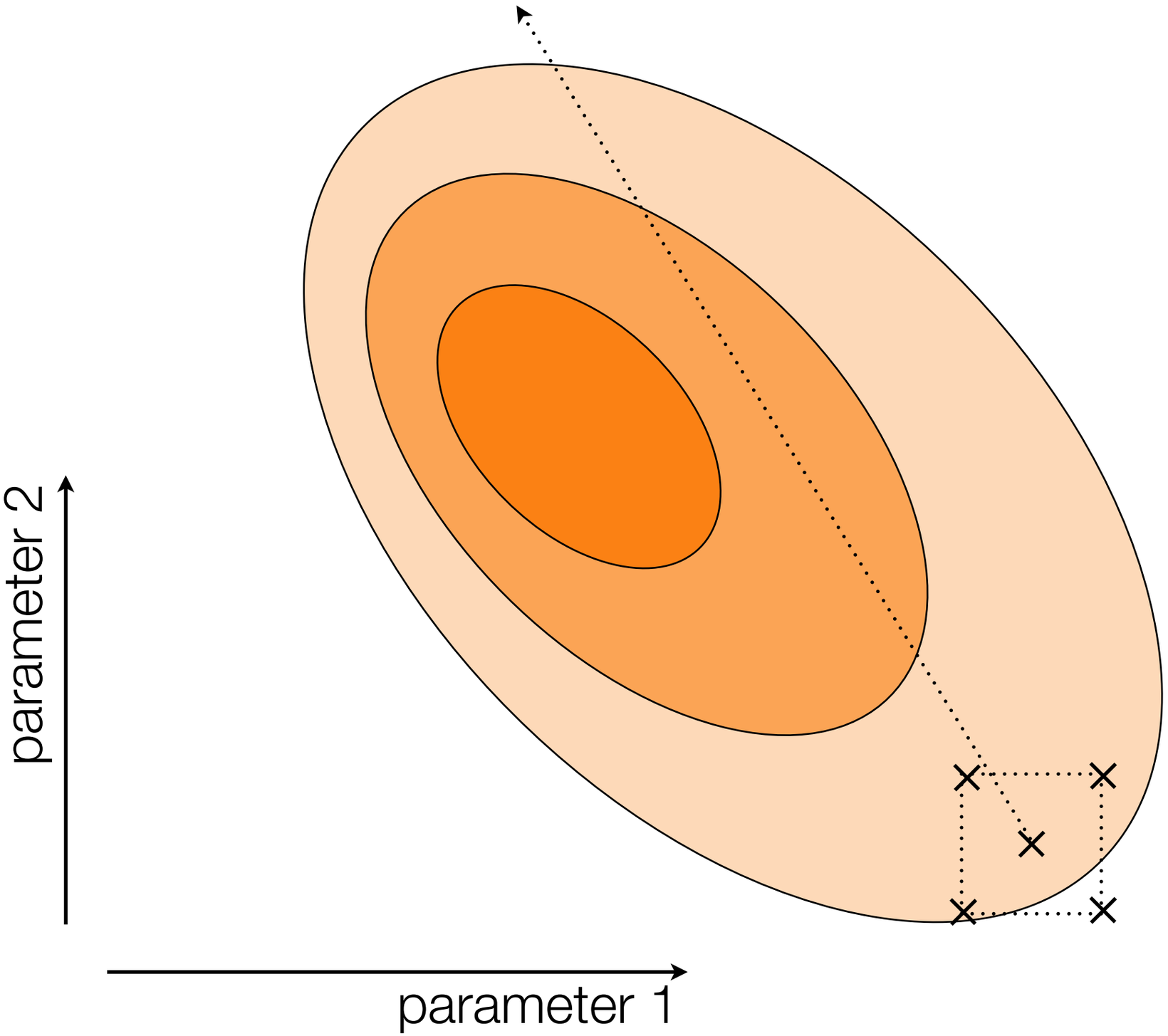
Response Surface Methodology



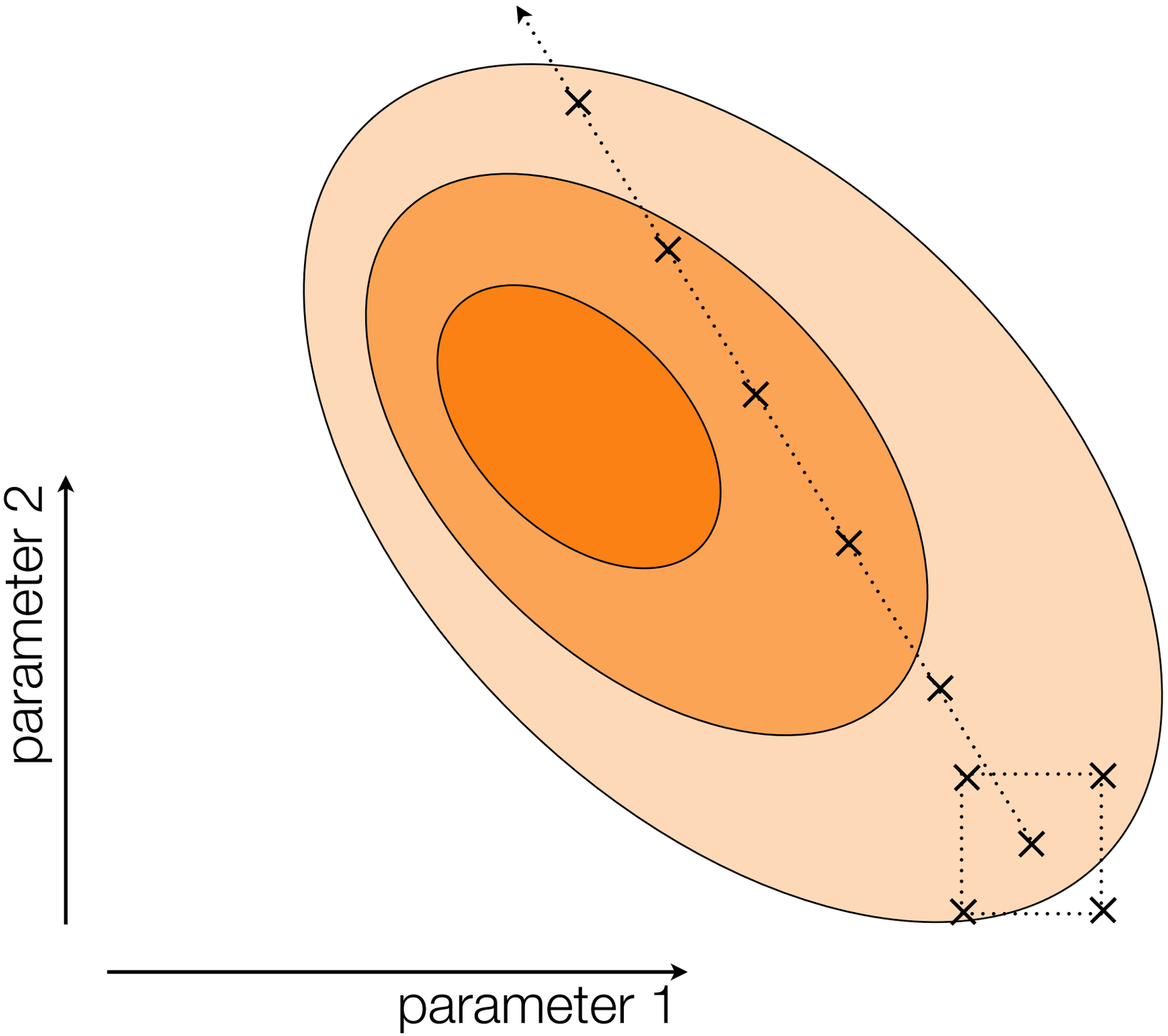
Response Surface Methodology



Response Surface Methodology



Response Surface Methodology



Reproducibility - Make Resources Available

benchmarks

wider experimentation on techniques

meta-analysis

establish credibility

Summary

Summary

statistics as evidence supporting an argument

... and this argument should be part of an
engaging narrative about the research

Magnitude

Articulation

Generality

Interestingness

Credibility