Wickedly fast Explanation Generation for Multi-Objective Optimization

Authors suppressed for blind review

Institution suppressed for blind review

Abstract. WICKED is a near linear-time algorithm for summarizing trade-offs in multi-objective problems. Humans can read that summary to find recommendations for their systems. This paper evaluates those recommendations using data from (a) the POM3 model of agile selection of tasks; (b) the COCOMO-suite predictors for software development effort, months, defects and risk. WICKED's recommendations were found to be just as effective at improving objective scores as the actions of standard optimizers.
WICKED runs orders of magnitude faster than standard optimizers. For example, for one of our larger models, WICKET and NSGA-II terminated in 3 and 150 minutes, respectively. Hence, we recommend WICKED when a succinct summary has to be rapidly generated (e.g. in some interactive design meeting).

Keywords: Software engineering, explanation, optimization, multi-objective.

1 Introduction

"If you cannot- in the long run- tell everyone what you have been doing, your doing has been worthless." – Erwin Schrödinger

To better support decision making in SBSE, we must better explain SBSE results. Explaining the results of these multi-objective optimizers to a user can be problematic. A typical run of a multi-objective optimizer can process thousands to millions of examples. It is an overwhelming task for humans to certify the correctness of conclusions generated from so many results. Verrappa and Leiter warn that

> "..for industrial problems, these algorithms generate (many) solutions, which makes the tasks of understanding them and selecting one among them difficult and time consuming" [58].

Even if explanations are constrained to (say) just a few hundred examples taken from the Pareto frontier, this can still confuse the user. Valerdi notes that it can take days for panels of human experts to rigorously review even a few dozen examples [56]. For example, once we had a client demanding to audit our SBSE reasoner. When we delivered the of candidate solutions on the Pareto frontier, they were overwhelmed by the amount of information. Flustered, the client discounted our analysis.

Other researchers have recognized the importance of explanation. For example, in the field of machine learning, it is known to be a key factor in selecting algorithms. "each time one of our favorite approaches has been applied in industry, each time the comprehensibility of the results, though ill-defined, has been a decisive factor of choice over an approach by pure statistical means, or by neural networks." [4].

In spite of the importance of explanation, there are few papers on this topic in the context of SBSE. One notable exception comes from Veerappa and Lieter [58] who clustered examples from the Pareto frontier (examples generated from a goal graph of requirements for London ambulance services). In this approach, "instead of having to inspect a large number of individual solutions, (users) can look at a much smaller number of groups of related solutions, and focus their attention on the important characteristics of the group rather than the particularities of their individual solutions" [58].

The Veerappa and Lieter study is an interesting investigation of an important issue that is rarely explored in the SBSE literature. That said, is there a better approach? SBSE algorithms are slow enough without the additional cost of some $O(N^2)$ post-processing clustering. Also, given that both clustering and SBSE algorithms divide and explore examples, perhaps there is something we can do "under the hood" to unify SBSE and clustering, thus removing the need for extra architecture.

This paper introduces WICKED, an experiment with an "under the hood" unification of SBSE and a near-linear time clustering algorithm. WICKED has several parts:

- **W**HERE4 is a linear-time clusterer that divides the population;
- **I**NFOGAIN is a linear-time feature selector that prunes irrelevancies;
- **C**ART is a decision tree learning that find branches to different clusters;
- **K**ILL is a branch pruner that simplifies (shortens) the branches;
- **E**NVY is a contrast set learner that find close pairs of *worse,better* clusters;
- **D**ELTA generates recommendations from the difference of the branches between *worse, better* clusters.

The first steps (W.I.C.K.) generate a very small decision tree that divides up the multi-objective decision space in to many small regions. After that, the last steps (E.D.) can be automatic or manual tasks performed by users as they trace out effects over their multi-objective problems. The rest of this paper addresses four research questions:

RQ1: *Can WICKED generates succinct summaries of multi-objective problems.* We show that, at least for the models explored here, the trees generated by W.I.C.K. are very short (25 lines, or less).

RQ2: *Are WICKED's summaries explanations of multi-objective problems?* We will defend this claim using theory taken from cognitive psychology.

RQ3: *Are WICKED's explanations effective?* We will re-run our models use WICKED's recommendations as constraints. This will achieve optimizations comparable to standard MOEA algorithms (NSGA-II).

RQ4: *Does WICKED scale to large problems?* We show that WICKED's runtimes scale linearly with population size and terminates faster than standard multi-objective optimizers (e.g. for one model, one minute for WICKED and 150 for NSGA-II [14]).

Note that we do *not* claim WICKED is a *better* at improving objective scores than standard optimizers. To make that case that WICKED is useful, we need only show that WICKED's optimizations are no worse than standard methods, while at the same time show that WICKED offers important additional services (explanation) that scale to large problems.

2 A Motivating Example

As motivation, this section offers the real-world goal that sparked this work. The Software Engineering Institute (SEI: http://www.sei.cmu.edu/) at Carnegie Mellon University is a Federally-Funded Research and Development Center. As such, SEI works with many software and software-intensive systems, which are primarily undertaken for the

United States government (and the US Department of Defense in particular) but also in commercial industry as well. As a result of involvement in these programs, the SEI is a storehouse for multiple repositories of qualitative and quantitative data collected from software development. Developers and managers from across the country look to SEI for explanations of what factors effect their project (these explanations are used to manage their current projects and well as propose methods on how to better handle their future projects in a better manner). Also, SEI researchers may have a voice in large-scale governmental policy decisions about information technology.

Fact sheets are one tool used by the SEI for explaining its advice on best practices and lessons learned (with traceability back to the data on which they are based). These explanations are short reports (one to two pages) which are intended to give busy managers quick guidance for their projects. In the case of quantitative data, these explanations may contain some 2D plot showing how one objective (e.g. defects) changes in response to changes in one input variable (e.g. lines of code).

Since they are aimed at communicating with busy professionals on specific points, the fact sheets of necessity are tradeoffs between understandability and accuracy. Overly simplistic fact sheets, while approachable for a larger audience, can be misleading. For example, consider 2D plots: even for single goal reasoning such as defect reduction, these are poorly characterized via one input variable. Studies with SE data have compared models learned $N = 1$ and $N > 1$ input variables. The models using more than one input performed better [38]. Further, many recent SE research publications that propose multiple competing goals for SE models; e.g.

- Build software *faster* using *less* effort with *fewer* bugs [17];
- Return defect predictors that find *most* defects in the *smallest* parts of the [**?**].

As we move from single goal to multiple-goal reasoning, the value of examining important factors in isolation using relatively simple plots becomes even more questionable. The SBSE experience is that reasoning and trading off between multiple goals is much more complex than browsing effects related to a single isolated goal.

Additionally, given all the context variables that can be used to describe different software projects, we recognize it is unlikely that any *one* report can explain *all* the effects seen in all different kinds of software projects. Several reports in empirical SE offer the same *locality effect*; i.e. models built from *all* data perform differently, and often worse, than those learned from specific subsets [5, 36, 41, 51, 59].

How can we augment SEI's fact sheets to explain:

- The space of effects in multiple dimensions of inputs?
- The responses of multiple objectives to changes in those inputs?
- And do so across the space of multiple contexts?

We propose printing succinct decision trees (generated by WICKED) on the SEI fact sheets. Standard decision trees have leaves predicting for a single class variable. The leaves of WICKED's trees, on the other hand, comment on multiple objectives.

For example, Figure 1 shows a decision tree whose leaves divide 500 decisions into the eleven groups "abcdefghijk". These decisions lead to five objectives: delivered lines of code, efort, months, defects, risks (for details on how this model, see later in this paper). Some clusters dominate others[1] so we say that those clusters ENVY another (see last column of Figure 1).

[1] Cluster i dominates j if no objective of j scores worse in j and at least one objective is better.

When a cluster ENVY's many other clusters, we select the one with the minimum DELTA (for cluster j that envies i, compute the set difference $\Delta = b_i - b_j$ between the tree branches b_j, b_i that lead to j, i). For example,

As described below, these trees can be used by domain users to generate multiple explanations about their domain. Better yet, as shown by our experiments, these small trees actually produce recommendations that are effective at optimizing a domain across multiple objectives.

		Mean values				
	Id	Effort	Months	Defects	Risks	Evnies
if plex ≤ 1.1:						
.. if team ≤ 2.9:						
.. then if plex ≤ 1.0:						
.. .. .. then..	a	1467	21	19600	17	
.. .. .. else..	b	1539	22	23135	17	a,d,k
.. else if team ≤ 4.3:						
.. .. .. then if plex ≤ 1.0:						
.. .. then..	c	1570	20	14233	17	
.. .. else..	d	1468	21	15807	17	
.. .. .. else	e	1890	23	15132	17	c
else if team ≤ 3.8:						
.. .. then if team ≤ 2.3:						
.. .. then..	f	1344	20	16297	18	
.. .. else if team ≤ 3.1:						
.. .. then..	g	1351	21	23808	18	f
.. .. else..	h	1437	21	23988	17	
.. else if plex ≤ 1.1:						
.. .. .. then..	i	1735	22	14928	18	c
.. .. .. else if team ≤ 4.7:						
.. .. then..	j	1517	21	17211	18	d,f
.. .. else..	k	1467	21	19600	17	

Fig. 1: Sample output from W.I.C.K. (WHERE4, INFOGAIN, CART, KILL) for an SE process model with five objectives: deliver *more* lines of code with *fewer* defects in *least* effort and time (effort is total staff hours; time is calendar time start to end of the project).

When we show trees like the above, one question we are often is asked in how such a simple structure can summarize the complex trade offs in a complex four-objective SE problem?

3 Generating Explanations

In the SE literature, the general consensus in software engineering is that "good" explanations are succinct explanations [4, 15, 20]. On this score, MOEAs fare poorly since their output can be very verbose (hundreds or more examples from the Pareto frontier).

WICKED uses several techniques taken from cognitive science and data mining in order to generate very succinct explanations. For cognitive science, we use *constrast sets*. According to Kelly's personal construct theory (PCT) humans explain things via "constructs" that distinguish sets of examples [27]. So, for Kelly, human explanations are not about "things" in isolation but rather the *differences between groups of things*.

WICKED employees several techniques to operationalize Kelly's theory. Using *cluster and differentiate*, data is clustered and the subsequent reasoning is re-

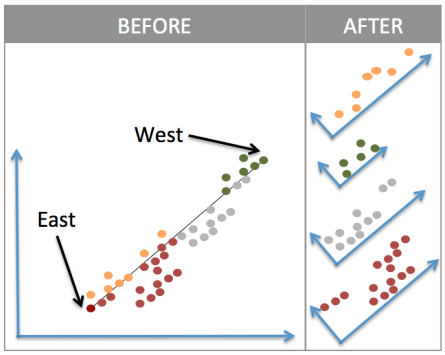

Fig. 2: WHERE4. Each dot is one set of de-

stricting to only the attributes that seperate those clusters. Next, using only the attributes found by cluster and differentiate, *contrast set learning* [47] (a) build rules to different clusters, (b) find pairs of clusters *worse, better*, then (c) report the delta between the branch to *worse* and the branches to *better*.

WHERE4: WICKED's clusterer [36] inputs a set of decisions $d_1, d_2...$, each of which specifies inputs to a multi-objective model. It separate the decisions along an approximate dimension of greatest variability: a line joining two very dissimilar decisions). This can be found in linear time after just $2N$ comparisons, using a technique proposed by Faloutsos and Lin [18]: pick any decision d_i at random; find *East*, the most distance decision d_j; find *West*, the most distant distance d_k to *East*[2]. Next, project all decisions d_i onto a two-dimensional (x, y) grid as follows: $a = dist(d_i, West)$; $b = dist(d_i, East)$; $c = dist(East, West)$; $x = (a^2 + c^2 - b^2)/(2c)$; $y = \sqrt{a^2 - x^2}$. As illustrated in Figure 2, WHERE4 then clusters the decisions on the median x, y values and recurses on each cluster (division of N examples terminates when a cluster contains less than $\sqrt{N}$).

INFOGAIN: To find attributes that differentiate between clusters, WICKED uses the INFOGAIN entropy measure adapted from Fayyad and Irani [19][3]. The values seen in N decisions are sorted, then recursively split at points that most simplifies the cluster distributions that use the values in each part of the split. To measure simplicity, we use the standard entropy measure. Let a split divides a decision's values into two lists v_1, v_2 of size n_1, n_2. Let the total counts of cluster c_k that use an element of v_i be c_{ik} and $p_{ik} = c_{ik}/n_i$. The entropy e_i of the list v_i can now be computed as $-\sum_k p_{ik} log_2 p_{ik}$ and the expected value of the simplicity of a list after splitting is $n_1/n * e_1 + n_2/n * e_2$ (where $n = n_1 + n_2$) and *lower* values are *better* since they contain *fewer* clusters. Using this method, WICKED's INFOGAIN operator ranks all attributes according to the simplicity of their associated cluster distribution. It then rejects all but the simplest 25% of the decisions and all subsequent reasoning in WICKED uses just this small set of decisions that most distinguish the cluster of decisions. Note that, like WHERE4, INFOGAIN is also very fast to execute: after an initial top-level $O(nlog(n))$ sort of the decision values, the rest is just a few linear passes through the data.

CART: Recall that in our operationalization of Kelly's theory, after *differentiate and contrast* (with WHERE4 and INFOGAIN), comes *constrast set learning* (with CART,KILL,ENVY and DELTA).

---

[2] For distance, we use Aha's Euclidean measure [2]; i.e. $\sqrt{\sum_x (d_{ix} - d_{jx})^2}$ where d_{ix}, d_{jx} are decisions values normalized 0..1 for the range min..max.

[3] For the reader who knows that work, we comment that standard FayyadIrani did not work for our models. We had to disable the MDL pruning of sub-ranges since, for an optimizer to explore small and interesting ranges of simulation data, it is necessary to let it "take a chance" on regions of the data that might not be strongly supported by many examples.

Apart from succinctness, cognitive science theory argues that there is more to "explaining" something that just presenting some ideas very succinctly. An important part of an explanation system is that it presents something that can be read in many different ways. Leake [32] lists a dozen different tasks that humans perform when "explaining" some phenomena (Leake does not claim that the following list is complete; just that it demonstrates a wide range of goal-based purposes for explanation). Leake's list includes:

1. Connect event to expected/believed conditions.
2. Connect event to previously unexpected conditions.
3. Find predictors for anomalous situations.
4. Find repair points for causes of an undesirable state.
5. Clarify current situation to predict effects or choose response.
6. Find controllable (blockable or achievable) causes.
7. Find actors contributions to outcome.
8. Find motivations for anomalous actions or decisions.
9. Find a within-theory derivation.

Generalizing from these examples, we say that a Leake-style explanation system weaves together a set of axioms into some story that overlaps, at least to some degree, with the known facts or outcomes. That story will have gaps containing things that are currently unknown but which might be assumed.

Working in the 1990s, Menzies formally characterized Leake's explanations as the *abductive* process where a theory T is augmented with assumptions A that (a) do not lead to contridictions and (b) also lead to desired goals G[4]. Classic implementions of Leake-style explanation suffer from computational problems.

That is, to Leake, explanation is akin to planning where the "explainer" is showing some audience how to find or connect together information. A system that supports such explanations makes it easier to "connect the dots". In practice that means an explanation system must:

- Input a large set of axioms: e.g. examples, pieces of background knowledge;
- Output a *reduced* set of axioms: e.g. rules, model fragments, or as done by Veerappa and Lieter [58], a small number of representative examples takes from centroids of clusters on the Pareto frontier;
- Such that, in the reduces space, it is simple and quick to generate goal-based explanations including the nine kinds listed above.

3.1 Decision Trees as "Explanation Tools"

Decision tree learning is a widely-used framework for data mining: given a single goal (called the "class"), find some attribute value that splits the data such that the distribution of classes in each split has been simplified (where the simplest distribution is one containing examples from only one class). Decision tree learners then grow sub-trees by recursing on the data in each split. Popular decision-tree learners include:

- CART [**?**, 8] which minimizes the variance of continuous classes in each split; or
- C4.5 [52] which minimizes the information content of the discrete classes in each split.

One reason to prefer decision trees is that they can very fast to execute. Each level of recursion processes progressively less data. Also, the computation at each level of the

[4] $T \cup A \vdash G$ and $T \cup A \vdash \not\perp$

recursion may be just a few linear passes through the data, followed by an sort of the attributes– so nothing more than $O(Nlog(N))$ at each level [**?**].

Another reason to prefer decision trees is that, as discussed below, they can operationalize much of Leake's and Kelly's cognitive models on explanation. Decision tree learners do have the disadvantage in that, as used in standard pratice, they only focus on one goal. The aim of this paper is to present a novel extension to standard decision tree learning that extends them to multi-objective optimization.

Given the above discussion, it is easy to see why that is so since decision tree learners can operationalize the above definitions of "explanation".

One reason for the popularity of decision tree learners Previously, work on constrast learning for single goal SE problems found that very succinct contrast sets could be generated as a post-processor to decision tree learning [42]:

- Building a decision tree to separate the different outcomes;
- Identifying leaves containing desired outcome X and undesired outcome Y;
- Querying that tree to find branches B_x and B_y that lead to X, Y.
- Computing $B_x - B_y$ which selects/rejects for desired/undesired outcomes.

In one spectacularly successful demonstration of this technique [39], it was found decision trees with 6,000 nodes had much superfluous information. Specifically, when some branch point high in the tree most separated the classes, then all contrast set learning had to do is report those branch decisions that selected for branches leading to the better classes. Using that approach, a contrast set learning could report contrasts with only one to four variables in each (and when applied to test data, those constrast sets were at pruning away all the undesired outcomes). Other studies with other data sets [40] confirmed the **the law of tiny constrasts**: *the minimal constrast set between things is usually much smaller than a complete description of those things.*

For simple goal classification, one way to operationalize Leake's framework is using decision trees. Given leaves of that tree $\{X, Y, Z, etc\}$, then exists some branch $\{B_x, B_y, B_z, etc\}$ that connects the root to the leaves as a conjunction of attribute/value pairs. Given some opinion about the value of the contents of each leaf $\{U_x, U_Y, U_Z, etc\}$, then the set difference $B_x - B_y$ is the contrast set of the differences that can drive examples on X over to Y.

"from here to there".

an explanation does not generate some single unique output. Rather, it inputs a set of axioms or examples and outputs a reduced set of axioms or examples within which it faster and simpler to generate explanations

Current MOEA algorithms are "instance-based methods" that return specific examples that perform "best" with respect to the multiple goals. The number of examples generated in this way can be overwhelming.

If a user wants to learn general principles from those examples, some secondary *explanation* process is required to group and generalize those examples. For example, Veerappa and Lieter [58] clustering examples from the Pareto frontier so users (at a minimum) need only browse the centroids of each clusters).

GAs flat vectors, not the trees explored by by ()say) Gouse et al.

Goals is performance just as good but explain better

One caveat before beginning: if the audience for the results of optimization are not human beings, then perhaps an explanation systems is not required. For example, Petke, Harman, Langdon, & Weimer [49] use evolutionary methods to rewrite code such that the new code executes faster. The audience for the rewritten code is a compiler. Such

compilers do not argue or and ask questions about the code they are given to process. Hence, that rewrite system does not necessary need an explanation system. That said, a succinct and useful description of the difference between passing and failing runs of the rewrite system could be useful when (e.g.) a human is trying to debug that code rewrite system.

Yet another model of "explanation" not explored here is the "surprise modeling" approach recommended by Freitas [21], Voinea&Tulea [1] and others including Horvitz [24]. In that approach, (a) some background knowledge (e.g. summaries of prior actions by users) is used to determine "normal" behavior; (b) users are only presented results that deviated from normal expectations. In analogous research, Koegh [28] argues that *time series discords* (infrequent sequential events in a times series) are a useful way to summarize reports from complex temporal streams. The premise of surprise modeling and reporting discords is that "rare events need to be explored". In non-temporal domains, time series discords becomes *anomaly detection* [9]. For example, in the SE domian, Voinea and Telea report tools that can quickly highlight regions of unusually active debugging (and such regions should be reviewed by management) [**?**] (see also the anomaly detection work of Gruska et al. [22]).

We do not dispute the importance of exploring anomalous outliers. On the other hand, when forming policies for software projects, we need treatments that are well supported by the data. Hence, our contrast sets report changes in the data that, in our data, were *frequently* seen to lead to change.

Also, time series discords and anomaly detection are reports on some variables. Hence, they have a different goal to WICKED that strives to report recommendations on how to change the system so to remove some problem.

Further, all the systems described above [1, 22, 24, 28] are either for unsupervised learning (where no objectives are known) or for single objective systems (where only one goal is known). WICKED, on the other hand, is more ambitious since it was designed for multi-objective systems.

Another potential issue with WICKED is correlation-vs-causation conflation. The issue here is that contrast sets will be useless if they report spurious correlations and not true causal effects. Proving that some effect is truly causal is a non-trivial task. The standard Hall criteria for causal effects [48] is so strict that, outside of highly controlled lab conditions, it rarely accepts that any effect is causal. Hence, in software engineering, when researchers talk of causality [6, 12, 25, 62] they use Granger's "predictive causality"; i.e. causality is the ability of predicting values seen in the future from values seen in the past. Elsewhere, Granger causality has been adapted to data mining by organizing cross-validations such that the test sets contain data collected at a later time than the training sets [34]. In this paper, we adapt Granger casualty to search-based methods by testing recommendations learned from M simulations on a subsequent round of N new simulations. Those recommendations satisfy Granger causality when the subsequent round of N simulations are changed in a manner predicted by the recommendations gleaned from the original M simulations.

"Data farming" is a technique used extensively by the U.S. Military [**?**]. Data farming builds a "landscape" of output that can be analyzed for trends, anomalies, and insights in multiple parameter dimensions. In a recent review of search-based and data mining methids in SE, we found numerous examples of data farming [**?**, **?**, **?**, **?**, 10, 11, 23, 46, 54, 57].

In theory. Once a project manager can view their project on the landscape, they can use this visualization to determine

We come to this work after attending a recent seminar at the US Department of Defence's Software Engineering Institute (SEI), Pittsburgh, USA. That seminar reflected on how to best broadcast the lessons learned by SEI to a very broad audience.

In the 21^{st} century, it is now impossible to manually browse very large quantities of software project data. For example, as of October 2012, Mozilla Firefox had 800K reports on software projects. While it is now possible to automatically analyze such data with data miners, at some stage a group of business users will have to convene to *interpret the results* (e.g., to decide if it is wise to deploy the results as a defect reduction method within an organization). These business users are now demanding that data mining tools be augmented with tools to support business-level interpretation of that data. For example,

> at a recent panel on software analytics at ICSE'12,
> industrial practitioners lamented the state of the art in data mining
> and software engineering [**?**]. Panelists commented that
> "prediction is all well and good, but what about decision
> making?". That is, these panelists are more interested in the interpretations
> that follow the mining, rather than just the mining.

4 Models

We have tested WICKED on numerous MOEA tasks including the standard laboratory problems (DTLZ, Schaffer, Fonseca, etc) and found it recommendations generated instances with objective scores competitive with those generated by NSGA-II or SPEA2.

Results from those standard lab problems are rarely convincing or interesting to software project managers. Hence, we show results from two business-level process models. POM3 [33, 50] implements the Boehm and Turner model [**?**, **?**, 50] of agile programming where teams select tasks as they appear in the scrum backlog. POM3 studies the implications of different ways to adjust task lists in the face of shifting priorities. XOMO [37, 43, 44] is four software process models from the University of Southern California. XOMO reports four-objective scores (which we will try to minimize): project *risk*; development *effort* and *defects*; and total *months* of development.

4.1 POM3

Turner and Boehm say that the agile management challenge is to strike a balance between the three objectives of *completion rates*, *idle rates*, and *overall cost* of a project. In the agile world, projects terminate after achieving a *completion rate* of $(X < 100)\%$

Short name	Decision	Description	Controllable
Cult	Culture	Number (%) of requirements that change.	yes
Crit	Criticality	Requirements cost effect for safety critical systems.	yes
Crit.Mod	Criticality Modifier	Number of (%) teams affected by criticality.	yes
Init. Kn	Initial Known	Number of (%) initially known requirements.	no
Inter-D	Inter-Dependency	Number of (%) requirements that have interdependencies. Note that dependencies are requirements within the *same* tree (of requirements), but interdependencies are requirements that live in *different* trees.	no
Dyna	Dynamism	Rate of how often new requirements are made.	yes
Size	Size	Number of base requirements in the project.	no
Plan	Plan	Prioritization Strategy (of requirements): 0= Cost Ascending; 1= Cost Descending; 2= Value Ascending; 3= Value Descending; 4 = $\frac{Cost}{Value}$ Ascending.	yes
T.Size	Team Size	Number of personnel in each team	yes

Fig. 3: List of Decisions used in POM3 (optimizers tune controllables, on right).

	POM3a A broad space of projects.	POM3b Highly critical small projects	POM3c Highly dynamic large projects
Culture	$0.10 \leq x \leq 0.90$	$0.10 \leq x \leq 0.90$	$0.50 \leq x \leq 0.90$
Criticality	$0.82 \leq x \leq 1.26$	$0.82 \leq x \leq 1.26$	$0.82 \leq x \leq 1.26$
Criticality Modifier	$0.02 \leq x \leq 0.10$	$0.80 \leq x \leq 0.95$	$0.02 \leq x \leq 0.08$
Initial Known	$0.40 \leq x \leq 0.70$	$0.40 \leq x \leq 0.70$	$0.20 \leq x \leq 0.50$
Inter-Dependency	$0.0 \leq x \leq 1.0$	$0.0 \leq x \leq 1.0$	$0.0 \leq x \leq 50.0$
Dynamism	$1.0 \leq x \leq 50.0$	$1.0 \leq x \leq 50.0$	$40.0 \leq x \leq 50.0$
Size	x ∈ [3,10,30,100,300]	x ∈ [3, 10, 30]	x ∈ [30, 100, 300]
Team Size	$1.0 \leq x \leq 44.0$	$1.0 \leq x \leq 44.0$	$20.0 \leq x \leq 44.0$
Plan	$0 \leq x \leq 4$	$0 \leq x \leq 4$	$0 \leq x \leq 4$

Fig. 4: Three classes of projects studied using POM3.

of its required tasks. Team members become *idle* if forced to wait for a yet-to-be-finished task from other teams. To lower *idle rate* and increase *completion rate*, management can hire staff- but this can increase *overall cost*.

When optimizing POM3, we seek changes to the controllables of Figure 3 that maximize *completion rate* while minimizing *cost* and *idle*ness. To make this task more realistic, we run POM for three different kinds of software projects, denoted POM3a, POM3b, POM3c shown in Figure 4. We make no claim that these three projects cover X% of all software projects– rather, our point here is that POM3 can handle different kinds of models.

scale factors (exponentially decrease effort)	prec: have we done this before? flex: development flexibility resl: any risk resolution activities? team: team cohesion pmat: process maturity
upper (linearly decrease effort)	acap: analyst capability pcap: programmer capability pcon: programmer continuity aexp: analyst experience pexp: programmer experience ltex: language and tool experience tool: tool use site: multiple site development sced: length of schedule
lower (linearly increase effort)	rely: required reliability data: secondary memory storage requirements cplx: program complexity ruse: software reuse docu: documentation requirements time: runtime pressure stor: main memory requirements pvol: platform volatility

Fig. 5: XOMO model decisions.

4.2 XOMO

The XOMO model enables an exploration of competing factors within software projects. Ideally, management decisions can minimize all of *months*, *effort*, *defects* and *risk*. However, there are many trade-offs to be considered. For example. increasing software reliability *reduces* the number of added defects while *increasing* the software development effort. For another example, better documentation can improve team communication and *decrease* the number of introduced defects. However, such increased documentation *increases* the development effort.

To explore those trade offs, XOMO uses the inputs of Figure 5 to drive four models. The *effort* model predicts for "development months" where one month is 152 work hours by one developer (and includes development and management hours):

$$effort = a \prod_i EM_i * KLOC^{b+0.01 \sum_j SF_j} \quad (1)$$

Here, *EM,SF* denote the effort multipliers and scale factors and a, b are the *local calibration* parameters which in COCOMO-II have default values of 2.94 and 0.91.

project		low	high
FL:	rely	3	5
JPL flight	data	2	3
software	cplx	3	6
	time	3	4
	stor	3	4
	acap	3	5
	apex	2	5
	pcap	3	5
	plex	1	4
	ltex	1	4
	pmat	2	3
	tool	2	2
	sced	3	3
	KSLOC	7	418

project		low	high
GR:	rely	1	42
JPL ground	data	2	3
software	cplx	1	4
	time	3	4
	stor	3	4
	acap	3	5
	apex	2	5
	pcap	3	5
	plex	1	4
	ltex	1	4
	pmat	2	3
	tool	2	2
	sced	3	3
	KSLOC	11	392

project		low	high
O2:	prec	3	5
Orbital Space	pmat	4	5
Place guidance	docu	3	4
navigation and	ltex	2	5
control (v2)	sced	2	4
	flex	3	3
	resl	4	4
	time	3	3
	stor	3	3
	data	4	4
	pvol	3	3
	reuse	4	4
	...		
	KSLOC	75	125

Fig. 6: Three case studies used in XOMO.

The XOMO *defect* model [7] assumes that certain variable settings *add* defects while others may *subtract* (and the final defect count is the number of additions, less the number of subtractions).

Also the *months* model predicts for total development time and can be used to determine staffing levels for a software project. For example, if *effort*=200 and *months*=10, then this project needs $\frac{200}{10} = 20$ developers.

Lastly, the *risk* model comments on how sets of management decisions decrease the odds of successfully completing a project. For example suppose a manager demands *more* reliability (*rely*) while *decreasing* analyst capability (*acap*). Such a project is "risky" since it means the manager is demanding more reliability from less skilled analysts. The XOMO *risk* model contains dozens of rules that trigger on each such "risky" combinations of decisions [35].

In the following, we run three different case studies of XOMO: one for each of the NASA projects described in Figure 6.

5 Code

WICKED's approach is to reply on *low dimensional approximations* and *multi-objective evolutionary algorithms* with decomposition Recent studies on of SE data show we do not need to reason about all the context variables (since many are redundant or noisy). When applied to SBSE, low dimensional approximations can guide mutation strategies to find better solutions one to two orders of magnitude faster than standard evolutionary optimizers [30, 31]. Secondly, work on MOEA/D [61] has shown the benefits of dividing problems into multiple cells, then optimizing each cell separately. Note that this approach is analogous (and actually pre-dates) work mentioned above on locality and context.

One way to describe the WICKED system is as a near-linear time variant on MOEA/D. [61]) WICKED began as an experiment with low-dimensional MOEA/D.r

```
if cplx ≤ 1.1:                              Effort  Months   Defect Risk
.. if resl ≤ 3.2:
.. .. if pvol ≤ 1.1:
.. .. .. if site ≤ 1.0:
.. .. .. ..  then: ['__25']  #                    ?       ?        ?       ?
.. .. .. ..  else: ['__20']  #              d     ?       ?        ?       ?
.. .. if rely ≤ 1.0:
.. .. .. if pcap ≤ 0.9:
.. .. .. .. if ltex ≤ 1.0:
.. .. .. .. .. if site ≤ 1.0:
.. .. .. .. .. ..  then: ['__3']  #         a     ?        ?       ?       ?
.. .. .. .. .. ..  else: ['__1']  #         c     ?        ?       ?       ?
```

```
.. .. .. .. .. else: ['__6']  #        b      ?        ?      ?       ?
.. .. .. .. if ltex ≤ 1.0:
.. .. .. .. .. if pmat ≤ 3.9:
.. .. .. .. .. ..  then: ['__15']  #          ?        ?      ?       ?
.. .. .. .. .. else: ['__15']  #       e     ?        ?      ?       ?
```

This section describes WHERE+CART+ENVY+DELTA.

5.1 WHERE

WHERE inputs a set of N examples, each of which is a set of decisions D mapped to a set of objectives O, so $N_i = (D, O)$ (and usually $D > 1$ and $O > 1$ and $O < D$). WHERE clusters the examples on the decisions and reports the average objective scores for each objective in each cluster.

WHERE uses a dimensionality reduction heuristic proposed by Faloutsos and Lin [18]. The method inputs N examples $N_1, N_2, ...$ Next, WHERE picks any point N_i at random. Thirdly, WHERE finds the point *West* $\in N$ that is furthest[5] from N_i. Finally, WHERE finds the point *East* $\in N$ that is furthest from *West* (and $c = dist(West, East)$).

To recursively cluster the data, WHERE iterates over $N_i \in N$ to find $a = dist(N_i, West)$, $b = dist(N_i, East)$, $x = (a^2 + c^2 - b^2)/(2c)$. This x value is the projection of N_i on the line running *East* to *West*. WHERE divides the examples on the median x value, then recurses on each half. Recursion on N initial examples stops when a sub-region contains less that M examples (e.g. $M = \sqrt{N}$).

Note that this four-step process requires only $2N$ distance comparisons per level of recursion and one call to a sorting routine to find the median value. The total time for WHERE is some linear multiple of the sorting time used to find the median at each level. Assuming sorting takes time $O(NlogN)$, then we can say that WHERE runs in near linear time (and not the $O(N^2)$ required for other clustering algorithms such as K-Means [**?**]).

5.2 CART

5.3 ENVY

5.4 DELTA

6 Methods

This study ranks methods using the Scott-Knott procedure recommended by Mittas & Angelis in their 2013 IEEE TSE paper [45]. This method sorts a list of l treatments with ls measurements by their median score. It then splits l into sub-lists m, n in order to maximize the expected value of differences in the observed performances before and after divisions. E.g. for lists l, m, n of size ls, ms, ns where $l = m \cup n$:

$$E(\Delta) = \frac{ms}{ls} abs(m.\mu - l.\mu)^2 + \frac{ns}{ls} abs(n.\mu - l.\mu)^2$$

Scott-Knott then applies some statistical hypothesis test H to check if m, n are significantly different. If so, Scott-Knott then recurses on each division. For example, consider the following data collected under different treatments *rx*:

[5] For this work, we use the standard Euclidean measure recommended for instance-based reasoning by Aha et al. [2]; i.e. $\sqrt{\sum_i (x_i - y_i)^2}$ where x_i, y_i are values normalized 0..1 for the range min..max.

```
rx1 = [0.34, 0.49, 0.51, 0.6]
rx2 = [0.6,  0.7,  0.8,  0.9]
rx3 = [0.15, 0.25, 0.4,  0.35]
rx4=  [0.6,  0.7,  0.8,  0.9]
rx5=  [0.1,  0.2,  0.3,  0.4]
```

After sorting and division, Scott-Knott declares:

- Ranked #1 is rx5 with median= 0.25
- Ranked #1 is rx3 with median= 0.3
- Ranked #2 is rx1 with median= 0.5
- Ranked #3 is rx2 with median= 0.75
- Ranked #3 is rx4 with median= 0.75

Note that Scott-Knott found little difference between rx5 and rx3. Hence, they have the same rank, even though their medians differ.

Scott-Knott is preferred to, say, hypothesis testing over all-pairs of methods[6]. To avoid an all-pairs comparison, Scott-Knott only calls on hypothesis tests *after* it has found splits that maximize the perfromance differences.

For this study, our hypothesis test H was a conjunction of the A12 effect size test of and non-parametric bootstrap sampling; i.e. our Scott-Knott divided the data if *both* bootstrapping and an effect size test agreed that the division was statistically significant (99% confidence) and not a "small" effect ($A12 \geq 0.6$).

For a justification of the use of non-parametric bootstrapping, see Efron & Tibshirani [16, p220-223]. For a justification of the use of effect size tests see Shepperd&MacDonell [55] and Kampenes [26]. These researchers warn that even if an hypothesis test declares two populations to be "significantly" different, then that result is misleading if the "effect size" is very small[7]. Hence, to assess the performance differences we first must rule out small effects. Vargha and Delaney's non-parametric A12 effect size test explores two lists M and N of size m and n. The counter $A12$ is incremented $\forall x \in M, y \in N$ as follows:

- If $x > y$ then add $1/(mn)$;
- If $x = y$ then add $0.5/(mn)$.

A12 reports the probability that numbers in one sample are bigger than in another. The A12 thresholds for "small,medium,large" effect are $\{0.56, 0.64, 0.71\}$ respectively where "small" is a euphemism for trivial or negligible effect. This test was recently endorsed by Arcuri and Briand at ICSE'11 [3].

7 Results

The following results come from a standard Python 2.7 interpreter (not PyPy) running on a 2.6 GHz Mac Os/X with 4 GB of ram. For NSGA-II, we used the out-of-the-box version from DEAP, https://github.com/DEAP/deap.

In the following, we compared WICKED's results with that of NSGA-II [**?**]. NSGA-II is a genetic algorithm (GA) with a highly optimized *select* operator:

- Each generation builds generation $G + 1$ by *selecting* better individuals, *combining* some of their parts, then *mutating* the results (a little).

[6] e.g. Six treatments can be compared $(6^2 - 6)/2 = 15$ ways. A 95% confidence test run 15 times total confidence $0.95^{15} = 46\%$.

[7] For example, Kocaguenli et al. [29] report on the misleading results of such hypothesis tests in software defect prediction (due to small size of the effect being explored).

- NSGA-II is a GA whose *select* operator uses a non-dominating sort procedure to divide the solutions into *bands* where $band_i$ dominates all of the solutions in $band_{j>i}$ (and NSGA-II favors the least-crowded solutions in the better bands).

One reason to favor NSGA-II as our comparison optimizer is *repeatability*. Many multi-objective optimizers as MOEA/D [60] and PSO [**?**] are really *frameworks* within which an engineer has free reign to make numerous decisions (evidence: review papers list dozens of variants on PSO and MOEA/D [**?**, 13]). Hence, in terms of *repeatability*, it can better to use precisely defined algorithms like NSGA-II.

Other reasons to use NSGA-II are that (a) it is very widely used and (b) there is no clear consensus that some other algorithm is better. When selecting a comparison algorithm, we reached out to our SBSE colleagues to find which algorithms are accepted as "best". However, no consensus was found. On the other hand, it can be shown that NSGA-II is widely used. In 2013, Sayyad and Ammar [53] surveyed 36 SBSE papers where $\frac{21}{36}$ used NSGA-II (of the others, 4 used some home-brew genetic algorithm and the remainder each used some MOEA not used by any other paper).

7.1 Runtimes

Standard MOEAs require at least N^2 comparisons between N candidates, for each generation G of the evolution. In theory, WICKED is much faster than that:

- The current implementation of WICKED uses $G = 1$ since its recommendations are the results of one analysis of the data (in future work, we plan to explore an iterative evolutionary version of the algorithm).
- All the sub-routines of WICKED take near linear-time (the slowest is CART that must sort all examples at each level of its trees).

On experimentation, WICKED's runtimes are consistent with its theoretical properties. Figure 7 show the effect on runtimes of increasing the initial population size for WICKED and NSGA-II. The solid lines denote WICKED's performance: note that they are always less that those seen with NSGA-II. The effect that WICKED runs faster than standard optimizers, is most pronounced in the more complex models. In the POM3 results, some of the POM3 variants take 100s of seconds to terminate. The same problems are handled by WICKER in under one minute.

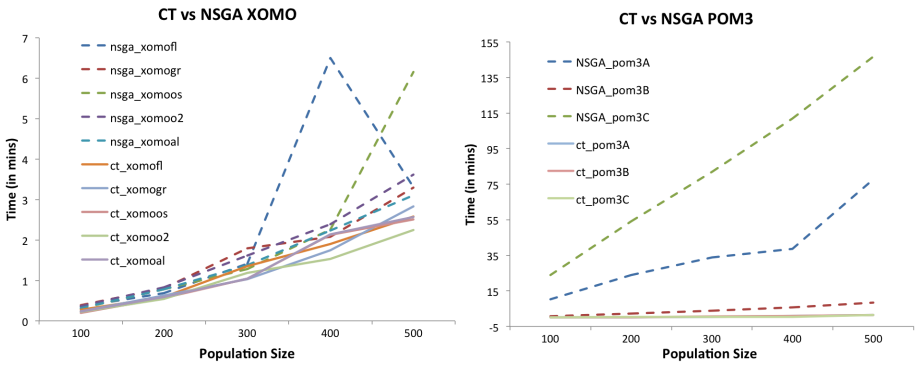

Fig. 7: Runtimes (minutes) for CT and NSGA II on POM Model (means over 20 repeats)

Note that these runtime results come from an optimized version of WICKED. Experiments are on-going with the Python profiler (to remove runtime bottlenecks). While those initial results are promising, we have nothing definitive to report at this time.

7.2 Optimization Improvements

To explore optimization improvements, we:

1. Collected *baseline* distributions seen in the objectives of the initial population (of 25 randomly generated individuals).
2. Using the baseline as generation $G = 1$, run NSGA-II until no improvement in any objective for three generations;
3. Using that baseline, run WICKED once. For each cluster:
   - Access the cluster items and the recommendation from the cluster;
   - Re-run the model that generated the data using constraints generated from that cluster;
4. Collected *treated* distributions from the output of steps two and three.

Figure **??** shows the *baseline* and *treated* distributions for:

- For all the objectives of:
  - The three variants of POM3 shown in Figure 4;
  - The three variants of XOMO shown in Figure 6;

That figure presents displays results from 20 repeated runs as horizontal quartile charts. In that figure, black dots denote median values and horizontal lines denote the 25 to 75th percentile range. To simplify readability, for each objective, all results and normalized 0..100 for the min to max values seen for that objective. Our three treatments are shown in the "Rx" column: "0" denotes the baseline; "W" denotes WICKED, and "N" denotes NSGA-II.

In all these results, *lower* values are *better* (exception: the *completion* goal in POM3 which we seek to *maximize*).

8 Related Work

Sayyad

GALE

WHERE4, fastmap, platt PDDF = where4

9 to do

Menzies has used combination of WHERE+ENVY has been used previously for finding context-specific rules for single-objective reasoning (reducing defects or software development effort [36]). What is new here is the addition of CART+CON as well as the application to multiple-objective reasoning.

References

1. Visual data mining and analysis of software repositories, 2007.
2. D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Mach. Learn.*, 6(1):37–66, January 1991.
3. A. Arcuri and L. Briand. A practical guide for using statistical tests to assess randomized algorithms in software engineering. In *ICSE'11*, pages 1–10, 2011.
4. I. Askira-Gelman. Knowledge discovery: Comprehensibility of the results. In *Hawaii International Conference on System Sciences*, 1998.
5. N. Bettenburg, M. Nagappan, and A. E. Hassan. Think locally, act globally: Improving defect and effort prediction models. In *MSR'12*, 2012.

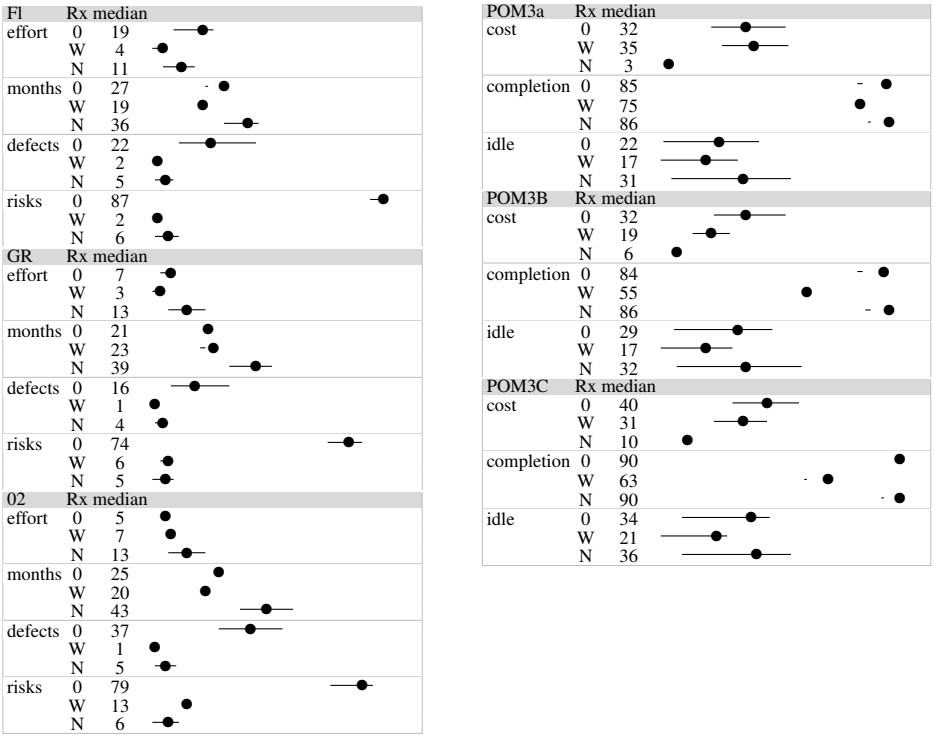

Fig. 8: XOMO results (left); POM3 results (right); all results from 20 runs with different random seeds. Big black dots show median values. Horizontal lines show 25th to 75th percentile. All results are normalized 0..100, min..max. Except for POM3's *completion* objective, *smaller* values are *better*. In the "Rx" column, "0,W,N" denotes results from baseline, WICKED, and NSGA-II (respectively).

6. P. Bhattacharya, M. Iliofotou, I. Neamtiu, and M. Faloutsos. Graph-based analysis and prediction for software evolution. In *Proceedings of the 34th International Conference on Software Engineering*, ICSE '12, pages 419–429, Piscataway, NJ, USA, 2012. IEEE Press.
7. B. Boehm, E. Horowitz, R. Madachy, D. Reifer, B. K. Clark, B. Steece, A. W. Brown, S. Chulani, and C. Abts. *Software Cost Estimation with Cocomo II*. Prentice Hall, 2000.
8. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. 1984.
9. V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41:15:1–15:58, July 2009.
10. E. Chiang and T. Menzies. Simulations for very early lifecycle quality evaluations. *Software Process: Improvement and Practice*, 7(3-4):141–159, 2003. Available from `http://menzies.us/pdf/03spip.pdf`.
11. L. Chung, B. Nixon, E. Yu, and J. Mylopoulos. *Non-Functional Requirements in Software Engineering*. Kluwer Academic Publishers, 2000.
12. C. Couto, M. Valente, P. Pires, A. Hora, N. Anquetil, and R. Bigonha. Bugmaps-granger: a tool for visualizing and predicting bugs using granger causality tests. *Journal of Software Engineering Research and Development*, 2, 1024.
13. S. Das and P. Suganthan. Differential evolution: A survey of the state-of-the-art. *Evolutionary Computation, IEEE Transactions on*, 15(1):4–31, Feb 2011.

14. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast elitist multi-objective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, 6:182–197, 2002.
15. K. Dejaeger, T. Verbraken, and B. Baesens. Toward comprehensible software fault prediction models using bayesian network classifiers. *Software Engineering, IEEE Transactions on*, 39(2):237–257, Feb 2013.
16. B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Mono. Stat. Appl. Probab. Chapman and Hall, London, 1993.
17. O. El-Rawas and T. Menzies. A second look at faster, better, cheaper. *Innovations in Systems and Software Engineering*, 6(4):319–335, 2010.
18. C. Faloutsos and K.-I. Lin. Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, SIGMOD '95, pages 163–174, 1995.
19. U. M. Fayyad and I. H. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.
20. N. E. Fenton and M. Neil. Software metrics: Success, failures and new directions. *J. Syst. Softw.*, 47(2-3):149–157, July 1999.
21. A. A. Freitas. On objective measures of rule surprisingness. In *Proceedings of the Second European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD'98*, pages 1–9. Springer-Verlag, 1998.
22. N. Gruska, A. Wasylkowski, and A. Zeller. Learning from 6,000 projects: lightweight cross-project anomaly detection. In *Proceedings of the 19th international symposium on Software testing and analysis*, ISSTA '10, pages 119–130. ACM, 2010.
23. W. Heaven and E. Letier. Simulating and optimising design decisions in quantitative goal models. In *Requirements Engineering Conference (RE), 2011 19th IEEE International*, pages 79–88, 2011.
24. E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *UAI'05*, pages 275–283, 2005.
25. B. Huberman, D. Romero, and F. Wu. Crowdsourcing, attention and productivity. *Journal of Information Science*, 35(6):758–765, December 2009.
26. V. B. Kampenes, T. Dybå, J. E. Hannay, and D. I. K. Sjøberg. A systematic review of effect size in software engineering experiments. *Information & Software Technology*, 49(11-12):1073–1086, 2007.
27. G. Kelly. *The Psychology of Persona] Constructs. Volume 1: A Theory of Personality. Volume 2: Clinical Diagnosis and Psychotherapy*. Norton, 1955.
28. E. Keogh, J. Lin, and A. Fu. Hot sax: Efficiently finding the most unusual time series subsequence. In *Proceedings of the Fifth IEEE International Conference on Data Mining*, ICDM '05, pages 226–233, Washington, DC, USA, 2005. IEEE Computer Society.
29. E. Kocaguneli, T. Zimmermann, C. Bird, N. Nagappan, and T. Menzies. Distributed development considered harmful? In *ICSE*, pages 882–890, 2013.
30. J. Krall and T. Menzies. Gale: Geometric active learning for search-based software engineering. *IEEE Transactions on Software Engineering (submitted)*, 2014.
31. J. Krall, T. Menzies, and M. Davies. Learning the task management space of an aircraft approach model. In *Modeling in Human Machine Systems: Challenges for Formal Verification, an AAAI 2014 Spring Symposium*, 2014.
32. D. Leake. Goal-based explanation evaluation. *Cognitive Science*, 15:509–545, 1991.
33. B. Lemon, A. Riesbeck, T. Menzies, J. Price, J. D'Alessandro, R. Carlsson, T. Prifiti, F. Peters, H. Lu, and D. Port. Applications of simulation and ai search: Assessing the relative merits of agile vs traditional software development. In *IEEE ASE'09*, 2009. Available from `http://menzies.us/pdf/09pom2.pdf`.

34. M. Lumpe, R. Vasa, T. Menzies, R. Rush, and R. Turhan. Learning better inspection optimization policies. *International Journal of Software Engineering and Knowledge Engineering*, 21(45):725–753, 2011.
35. R. Madachy. Heuristic risk assessment using cost factors. *IEEE Software*, 14(3):51–59, May 1997.
36. T. Menzies, A. Butcher, D. R. Cok, A. Marcus, L. Layman, F. Shull, B. Turhan, and T. Zimmermann. Local versus global lessons for defect prediction and effort estimation. *IEEE Trans. Software Eng.*, 39(6):822–834, 2013. Available from `http://menzies.us/pdf/12localb.pdf`.
37. T. Menzies, O. El-Rawas, J. Hihn, M. Feather, B. Boehm, and R. Madachy. The business case for automated software engineerng. In *ASE '07: Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*, pages 303–312, New York, NY, USA, 2007. ACM. Available from `http://menzies.us/pdf/07casease-v0.pdf`.
38. T. Menzies, J. Greenwald, and A. Frank. Data mining static code attributes to learn defect predictors. *IEEE Transactions on Software Engineering*, January 2007. Available from `http://menzies.us/pdf/06learnPredict.pdf`.
39. T. Menzies and Y. Hu. Data mining for very busy people. November 2003. Available from `http://menzies.us/pdf/03tar2.pdf`.
40. T. Menzies and Y. Hu. Just enough learning (of association rules): The TAR2 treatment learner. In *Artificial Intelligence Review*, 2007. Available from `http://menzies.us/pdf/07tar2.pdf`.
41. T. Menzies and M. Shepperd. Special issue on repeatable results in software engineering prediction. *Empirical Software Engineering*, 17:1–17, 2012.
42. T. Menzies and E. Sinsel. Practical large scale what-if queries: Case studies with software risk assessment. In *Proceedings ASE 2000*, 2000. Available from `http://menzies.us/pdf/00ase.pdf`.
43. T. Menzies, S. Williams, O. El-Rawas, D. Baker, B. Boehm, J. Hihn, K. Lum, and R. Madachy. Accurate estimates without local data? *Software Process Improvement and Practice*, 14:213–225, July 2009. Available from `http://menzies.us/pdf/09nodata.pdf`.
44. T. Menzies, S. Williams, O. El-Rawas, B. Boehm, and J. Hihn. How to avoid drastic software process change (using stochastic stability). In *ICSE'09*, 2009. Available from `http://menzies.us/pdf/08drastic.pdf`.
45. N. Mittas and L. Angelis. Ranking and clustering software cost estimation models through a multiple comparisons algorithm. *IEEE Trans. Software Eng.*, 39(4):537–551, 2013.
46. I. Myrtveit, E. Stensrud, and M. Shepperd. Reliability and validity in comparative studies of software prediction models. *IEEE Trans. Softw. Eng.*, 31(5):380–391, May 2005.
47. P. K. Novak, N. Lavrač, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, 10:377–403, June 2009.
48. L. Paul and N. Hall. *Causation : A Users Guide*. Oxford University Press, 2013.
49. J. Petke, M. Harman, W. Langdon, and W. Weimer. Using genetic improvement & code transplants to specialise a c++ program to a problem class. In *European Conference on Genetic Programming (EuroGP)*, 2014.
50. D. Port, A. Olkov, and T. Menzies. Using simulation to investigate requirements prioritization strategies. In *IEEE ASE'08*, 2008. Available from `http://menzies.us/pdf/08simrequire.pdf`.
51. D. Posnett, V. Filkov, and P. Devanbu. Ecological inference in empirical software engineering. In *Proceedings of ASE'11*, 2011.

52. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufman, 1992. ISBN: 1558602380.
53. A. Sayyad and H. Ammar. Pareto-optimal search-based software engineering (posbse): A literature survey. In *RAISE'13, San Fransisco*, May 2013.
54. M. Shepperd and G. F. Kadoda. Comparing software prediction techniques using simulation. *IEEE Trans. Software Eng*, 27(11):1014–1022, 2001.
55. M. J. Shepperd and S. G. MacDonell. Evaluating prediction systems in software project estimation. *Information & Software Technology*, 54(8):820–827, 2012.
56. R. Valerdi. Convergence of expert opinion via the wideband delphi method: An application in cost estimation models. In *Incose International Symposium, Denver, USA*, 2011. Available from http://goo.gl/Zo9HT.
57. A. van Lamsweerde and E. Letier. Integrating obstacles in goal-driven requirements engineering. In *Proceedings of the 20th International Conference on Software Engineering*, pages 53–62. IEEE Computer Society Press, 1998. Available from `http://citeseer.nj.nec.com/vanlamsweerde98integrating.html`.
58. V. Veerappa and E. Letier. Understanding clusters of optimal solutions in multi-objective decision problems. In *Proceedings of the 2011 IEEE 19th International Requirements Engineering Conference*, RE '11, pages 89–98, Washington, DC, USA, 2011. IEEE Computer Society.
59. Y. Yang, Z. He, K. Mao, Q. Li, V. Nguyen, B. W. Boehm, and R. Valerdi. Analyzing and handling local bias for calibrating parametric cost estimation models. *Information & Software Technology*, 55(8):1496–1511, 2013.
60. H. Zhang and X. Zhang. Comments on 'data mining static code attributes to learn defect predictors'. *IEEE Transactions on Software Engineering*, September 2007.
61. Q. Zhang and H. Li. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *Evolutionary Computation, IEEE Transactions on*, 11(6):712–731, Dec 2007.
62. P. Zheng, Y. Zhou, M. Lyu, and Y. Qi. Granger causality-aware prediction and diagnosis of software degradation. In *Services Computing (SCC), 2014 IEEE International Conference on*, pages 528–535, June 2014.