Why do we Eval?

A large-scale study of Eval usage in R

Anon
Under review

Abstract

Most dynamic languages allow users to turn text into code using various functions, often named `eval`, with language-dependent semantics. The widespread use of these reflective functions hinders static analysis and prevents compilers from performing optimizations. This paper aims to provide a better sense of why programmers use `eval`. Understanding why `eval` is used in practice is key to finding ways to mitigate its negative impact. We have reasons to believe that reflective feature usage is language and application domain specific; we focus on data science code written in R, and compare our results to previous work that analyzed web programming in JavaScript. Dynamic analysis of a corpus of 4.5M lines of libraries and end-user code confirms that `eval` is indeed in widespread use; R's `eval` is more dangerous in some ways, and safer in others, than what was previously reported for JavaScript.

2012 ACM Subject Classification Software and its engineering → Scripting languages; General and reference → Empirical studies; Software and its engineering → General programming languages

Keywords and phrases R, reflection, eval

Digital Object Identifier 10.4230/LIPIcs.ECOOP.2021.23

1 Introduction

Most dynamic languages provide their users with a facility to transform unstructured text into executable code and evaluate that code. We refer to this reflective facility as `eval` bowing to its origins in LISP, all the way back in 1956 [15]. `Eval` has been much maligned over the years. In computing lore, it is as close to a boogeyman as it gets. Yet, for McCarthy, `eval` was simply the way to write down the definition of LISP, he was surprised that someone coded it up and offered it to end users. Since then, reflective facilities have been used to parameterize programs over code patterns that can be provided after the program is written. The presence of such a feature in a language is a hallmark of dynamism; it is a form of delayed binding as the behavior of any particular call to `eval` will only be known when the program is run and that particular call site is evaluated.

Trouble in Paradise. Reflective facilities hinder most attempts to reason about, or apply meaning-preserving transformation to, the code using them. In practice, `eval` causes static analysis techniques to loose so much precision as to become pointless. For compilers, anything but the most trivial, local, optimizations are unsound after a use of `eval`. Furthermore, the addition of arbitrary code — code that could have been obtained from a network connection — as a program is running is a security vulnerability waiting to happen. To illustrate these challenges, consider the interaction of a static analysis tool with a dynamic language. An abstract interpretation-based program analyzer computes an over-approximation of the set of possible behaviors exhibited by the program under study [7]. A reflective may have *any* behavior that can be expressed in the target language; i.e. `eval` can be replaced by any legal sequence of instructions. As dynamic languages tend to be permissive, the analysis has to, for example, assume that many (or all) functions in scope may have been redefined, e.g. that `‘+‘` now opens a network connection or something equally surprising. A single occurrence of `eval` causes the static analyzer to loose all information about program state

European Conference on Object Oriented Programming (ECOOP 2021).
Editors: Eds.; Article No. 23; pp. 23:1–23:17
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

and meaning of identifiers. This loss of precision can sometimes be mitigated by analyzing the string argument [6] to bound its possible behavior but when the string comes from outside the program not much can be done. A frustrated group researchers argued giving up on soundness and, instead, under-approximating dynamic features (soundiness) [14]. In their words "a practical analysis, therefore, may pretend that `eval` does nothing, unless it can precisely resolve its string argument at compile time." Alas, assuming that `eval` does not have side-effects, or that side-effects will not affect the results of the analysis, may be unduly optimistic.

Is Past Prologue? Previous work investigated how `eval` is used in web programming, specifically in websites that use JavaScript [20]. In 2010, 17 of the largest website used the feature. In 2011, 82% of the 10,000 most accessed sites used `eval` [19]. Yet, the strings passed to `eval`, and their behaviors when executed, are far from random; it was shown that when one can observe several calls to `eval`, the "shape" of future calls can be predicted with 97% accuracy [16]. Overall, practical usage suggested that most reflective calls were relatively harmless. While this backs up the soundiness squad's approach, does it generalize to other application domains than web programming and to other languages?

The Here and Now. In this study, we investigate the usage of `eval` in programs written in the R programming language. R is language designed by statisticians for applications in data science [18, 11]. What makes looking at R after JavaScript interesting is that, while both languages are dynamic, they are quite different. While one can program in an object-oriented style in R, like in JavaScript, R is primarily a lazy, untyped, functional language. JavaScript was designed to run untrusted code in browser, while R is used for statistical computing on desktops. JavaScript is a general purpose language used by a wide community of programmers; while R is used for scientific computing by data scientists and domain experts with, often, limited programming experience. One can distinguish between library implementers, developers with some programming experience and a working knowledge of R, and end-users, who are typically not expert programmers and often have only a cursory knowledge of the language.[1] Our goal is thus to highlight the differences in usage between JavaScript and R, and try to explain those differences in terms of language features, application domain and programmer experience. Hopefully some of our observations will generalize to other languages.

The What and How. One significant benefit of choosing R is that every package in the CRAN repository is curated and comes with examples of typical usage. This gives us a large code base that we can analyze dynamically. To observe `eval` we built a two-level monitoring infrastructure:[2] we can monitor R programs by instrumentation — this gives us access to many user-visible properties of R programs — but we can also monitor the inner-workings of the R interpreter — this allows us to capture details not exposed at the source level. Dynamic analysis is limited, it can only observe behaviors triggered by the particular inputs passed to a program. Luckily, CRAN libraries come with many tests and use-cases. The choice of corpus is crucial. Our corpus has been constructed to reflect the levels of sophistication of the R community. We distinguish between *CRAN packages* (500

[1] Consider that R evaluates function argument lazily, just like Haskell. We informally surveyed end-users, including computer scientists, and did not find a single user aware of this fact. Library developers, on the other hand, know about laziness and program defensively around it.

[2] Our infrastructure is open source and publicly available, our anonymized code is at `http://url.com`, a more complete artifact will be submitted to the artifact evaluation committee.

curated packages that pass stringent quality checks and are equipped with tests and sample data) and *Kaggle scripts* (1,619 end-user written programs that performs a particular data analysis task). It is reasonable to expect that `eval` usage differ between these datasets: the libraries represent a lively ecosystem with new libraries added each day, while end user code is often thrown together, run once, and never revisited.

Why do we Eval? **A short summary of the results should go here.**

2 Background and Previous work

This section provides a short introduction to R as it has a few surprising features that impact the use of reflective features of the language. We then look at the semantics of `eval` in R and discuss some design choices. Lastly, we put this paper in context of previous work.

2.1 R in a nutshell

R is a lazy functional programming language [17] with dynamic features that allow to write object-oriented code. Most data types are vectorized. Values are constructed by the `c(...)` function, e.g. `c("a","bc")` creates a vector of two strings. To enable equational reasoning, R copies values accessible through multiple variables when they are written to. Values can be tagged by user-defined attributes. For instance, one can attach the attribute `dim` to the value `x<-c(1,2,3,4)` with `attr(x,"dim")<-c(2,2)`. This causes arithmetic functions to treat `x` as a 2×2 matrix. Another attribute is `class` which can be bound to a list of names, e.g., `class(x)<-"human"`. This sets the class of `x` to `human`; classes are used for object-oriented dispatch. Every R linguistic construct is desugared to a function call, even control flow statements, assignments, and bracketing. Furthermore, all functions can be redefined in user code. This makes R both flexible and challenging to compile [8]. Arguments to user-defined functions are bundled into thunks called *promises*. Logically, a promise combines an expression's code, its environment, and its value. To access the value of a promise, one must force it. Forcing a promise triggers evaluation and the computed value is captured for future reference.

2.2 Eval Semantics

The expressive power of `eval` depends on design decisions along two axes:

- **Scoping:** What environment does `eval` executes its argument in? JavaScript and R evaluate it in the current environment, thus exposing local variables and parameters and breaking the caller's abstraction boundary. Julia is more restrictive, `eval` runs at the "top level", in the global environment. JavaScript has a global `eval` that behaves like Julia, and a "strict mode" in which `eval` may access local variables, but fobidden from injecting new ones in the local environment. In Java, one can implement `eval` with Julia-semantics.[3]
- **Reflective API:** What other reflective operations are exposed to user code? JavaScript provides few reflective functions (other than enumeration of an object's properties and string-based indexing), the impact of `eval` is thus limited. R, on the other hand, has a

[3] While `eval` used to be the purview of interpreted languages, just-in-time compilation lifted this restriction. `Eval` takes its input, wraps it in a static method of a new anonymous class, generates bytecode for that class, invokes the class loader to install that code, and finally reflectively calls the method.

rich reflective API. For example, R allows user code to walk the call stack and arbitrarily add and delete local variables. This means that there are few limits to the side-effects of an `eval`.

These design choices impact our ability to reason about code when reflection is used. A more expressive `eval` translates directly into additional restrictions for program analysis and transformation tools. The reason Julia restricts `eval` to the top level is to preserve the compiler's ability to optimize code [2]. In particular, `eval` had to be prevented from changing the type of local variables as this would require recompilation of the method. Julia even adds a versoning mechanism, called world age, to ensure that code added during an `eval` does not invalidate inlining optimizations [1]. More permissive languages such as R are much harder to deal with from a compiler's perspective.

Eval in R. The reflective interface exposed by R is somewhat complex. The core library exports four functions with slightly different semantics, `eval`, `evalq`, `eval.parent`, and `local`. `Eval` is the more general function, and, unlike in JavaScript, it takes three arguments, an expression, its environment and the enclosing environment. The following discussion simplify some details which are not relevant to this paper. For the interested reader we recommend the excellent book by Hadley Wickham [23]. The definition of `eval` starts as follows:

```
eval
  <- function(expr,
              envir = parent.frame(),
              enclos = if(is.list(envir)) parent.frame() else baseenv())
        { ...body... }
```

The parameters are `expr`, the value to be evaluated, `envir`, the environment in which evaluation happens, and `enclos`, the environment to look up objects not found in `envir` (by default either the top level, `baseenv` or the environment of the caller). The `expr` parameter can take values of different type, for our purposes we focus on the most common, `expressions`. It is easiest to think of an `expression` as an abstract syntax tree as returned by R's parser. The simplest way to create an `expression` is to call `quote` passing some R expression:

```
> quote(a + b)
# a + b
```

Here the return value is an abstract syntax tree. From a string, use the `parse` function:

```
> parse(text="a+b")
# expression(a+b)
```

The most common way to create an expression is to extract it from an argument. Each argument comes packaged inside a promise which retains the source code of the argument. Consider the following function definition, `f` is a function with a single parameter `x`. When the call `f(a+b)` is evaluated, `a+b` is stored inside a newly created promise.

```
> f <- function(x) substitute(x), list(a=1)))
> f(a+b)
# expression(1+b)
> substitute(a+b,list(a=1,b=2))
# 1+2
```

The call to `substitute(x)` extracts, from the promise bound to `x`, the expression passed in the call to `f`, i.e., `a+b`. The `substitute(exp,env)` function takes two arguments, the second is an object that can be used as an environment. This can be either a list of named values, a data frame or an actual environment. `substitute` will look for occurrences of each symbol from `env` in `exp` and replace them with their value taken from `env`.

```
> environment()
# <environment: R_GlobalEnv>
```

The above shows how to access the current environment, this can belong to the current function or, as above, the top level. Environments are nested, with each environment having a parent. This nesting is used when looking up names. When calling `new.env`, the default parent is the current environment. The chain of environments can be traversed with `parent.env`, ultimately one arrives at `emptyenv`. `parent.frame` and `sys.frame` grant access to environments further up in the calling stack, up to the global environment `.GlobalEnv`, and packages. One can also directly read, modify or create new bindings, given any environment:

- `env$v` and `get("v",envir=env)` read variable `v` in environment `env`;
- `env$v<-2` and `assign("v",2,envir=env)` write 2 in `v`. If not found, `v` is created in `env`.

Environments are often used as hashtables by programmers as they have reference semantics (other are copy-on-write) and have a built-in string look up. across function calls, or to create package namespaces.

```
evalq <- function(exp,env,enclos) eval(quote(exp),env,enclos)
eval.parent <- function(exp,n) eval(exp,parent.frame(n))
local <- function(exp,env,enclos) evalq(exp,new.env())
```

The three `eval` variants can be expressed as calls to `eval`. The `evalq` form quotes the argument to prevent it from being evaluated in the current environment. The `eval.parent(e,n)` form evaluates `e` in the environment of the `n`-th caller of this function. Finally, `local` evaluates an `exp` in a new environment to avoid polluting the current one.

2.3 Eval Usage in R

The R language was intended to be extensible, the combination of lazy evaluation, `substitute` and `eval` are the tools given to developers to this end. This API is slightly more complex than just passing a string, it is conceivable that this may discourage some casual users. `Eval` is also being used to reduce boilerplate code and provide convenience features for programmers. We now give some representative examples of its usage.

Intercession. A common use case for `eval` is to be combined with `match.call`. `match.call` walks up the call stack, captures the code that invoked the currently executing function, and returns it as an unevaluated expression. The pattern is to transform a call to some function `f` into a call to `g` with some arguments retained and others modified. As an illustration, consider the `vcpart` package's function `tvcglm` that is translated to a call to `tvcm` with two modifications to the argument list: argument `control` can't be missing and `fit` is set `"glm"`. The function ends with a call to `eval.parent` to ensure that the rewritten call is evaluted in the same environment the original call was.

```
tvcglm <- function(formula, data, family, control=tvcglm_control(), ...) {
  Call <- match.call()
  Call[[1L]] <- as.name("tvcm")
  if (!"control" %in% names(Call)) Call$control<-formals(tvcglm)$control
  if ("fit" %in% names(list(...))) warning("’fit’␣is␣ignored.␣")
  Call$fit <- "glm"
  return(eval.parent(Call))
}
```

This pattern is recognizable by the fact that the expression is a call and the target environment is that of the parent.

Code Generation. The more traditional use of `eval` is to execute code that was assembled by the programmer into a string. Here we show the method `plot` for class `sback` in package `wsbackfit`, simplified for explanatory purposes. The function takes a long argument list, the names of which are captured in the list `opt`. The string `stub` is composed of a subset of the arguments passed to this function; the variable is used to construct a call to `base::plot` which will draw a plot. *Note: This example is a bit messed up. What is var? Does the use of ... in the eval grab the arguments of this call? If yes does it not end up with xlab twice? HELP* The `parse` and `eval` combination is used by the the `source` function in the base library to load R code from a file in the current workspace.

```
plot.sback <- function(x,...) {
  opt <- names(list(...))
  stub <- paste(
   ifelse("xlab" %in% opt,"",paste(",xlab=\"",var,"\"",sep="")),
   ifelse("main" %in% opt,"",main.aux),
   ifelse("type" %in% opt,"",",type=\"l\""),
   sep = "")
  plot <- paste("plot(x.data,",stub,",...)",sep="")
  eval(parse(text=plot))
  ...
```

This pattern is recognizable by its use of `parse` to turn a string into an expression.

Debloating. `Eval` is often used as a means to reduce boilerplate code; simple and repetitive code can easily be replaced with judcious use of `eval`. For example, the `data.table` package uses `eval` to calls the `options` function with named arguments taken from a vector of strings. While the benefits are limited in this example, it is an attractive tool for programmers.

```
 opts = c("datatable.verbose"="FALSE", # ... many others
 for (i in names(opts))
   eval(parse(text=paste0("options(",i,"=",opts[i],")")))
```

This pattern is a special case of code generation, recognizable by the fact that `eval` is executed in a loop.

Trivial. When values are passed to `eval`, they are returned unchanged. They are an example of trivial uses of `eval`. Another trivial use is the empty expression, often found in JavaScript, but rare in R.

2.4 Previous Work

Richards et al. [19] provided the first large-scale study of the runtime behavior of `eval` in JavaScript. They dynamically analyzed a corpus of the 10,000 most popular websites with an instrumented web browser to gather execution traces. They show that `eval` is pervasive with 82% of the most popular websites using it. The reasons for its use include the desire to load code on demand, deserialization of JSON dataq and lightweight meta-programming to customize web pages. While many uses were legitimate, just as many were unnecessary and could be replaced with equivalent and safer code. They categorized inputs to `eval` so as to cover the vast majority of input strings. Restricting themselves to `eval` in which all named variables refer to the global scope, many patterns could be replaced by more disciplined code [16, 12]. The work did not measure code coverage, so the numbers presented are a lower bound on the possible behaviors. Furthermore, JavaScript usage in 2011 is likely different from today, e.g. Node.js was not covered by Richards. More details about dynamic analysis of JavaScript can be found in [10].

Wang et al. [22] analyzed use of dynamic features in 18 Python programs to find if they affect file change-proneness. Files with dynamic features are significantly more likely to be the subject of changes, than other files. Chen et al. looked at the correlation between code changes and dynamic features, including `eval`, in 17 Python programs [5]. They did not observe many uses of `eval`. Callau et al. [4] performed an empirical study of the usage of dynamic features in 1,000 Smalltalk projects. While `eval` itself is not present, Smalltalk has a rich reflective interface. The authors found that reflective are used in less than 2% of methods. The most common reflective method is `perform:`; it send a message that is specified by a string. These features are mostly used in the core libraries.

Bodden et al. [3] looked at usage of reflection in the Java DaCapo benchmark suite. They found that dynamic loading was triggered by the benchmark harness. The harness then executes methods via reflection, this caused static analysis tools to generate an incorrect call graph for the programs in DaCapo.

Morandat et al. [17] had a short section on the usage of `eval` in R. They found the it widely used in R code with 8500 call sites in CRAN and 2.2 million dynamic calls. The 15 most frequent call sites account for 88% of those. The `match.arg` function is the highest used one with 54% of all calls. In the other call sites, they saw two uses cases. The most common is the evaluation of the source code of a promise retrieved by `substitute` in a new environment; e.g. as done in the `with` function. The other use case is the invocation of a function whose name or arguments are determined dynamically. For this purpose, R provides `do.call` and thus `eval` is overkill.

3 Methodology

This section explains our methodology for selecting the corpus that will be analyzed and how we implemented our dynamic analysis.

3.1 Corpus

Our corpus is assembled out of two set of programs, one consisting of popular libraries obtained from the Comprehensive R Archive Network[4] (we refer to it as **CRAN**) and the other from the Kaggle data science competition (written **Kaggle**). The intent here is to contrast code written by experienced developpers (CRAN) with code authored by typical end users of the language (Kaggle).

All the code in our is runnable, i.e. we have scripts that invoke the CRAN packages and the Kaggle programs have input data. Clearly code coverage will vary; as is the case for any dynamic analysis, the quality of results is predicated on how representative the executions

[4] CRAN is the largest repository of R code with over 16.2K packages. It receives about 6 new package submissions a day [13]. Unlike other open code repositories such as GitHub, CRAN is a curated repository. Each submitted package must abide to a number of well-formedness rules that are automatically checked asserting certain quality. Most relevant for this work is that all of the runnable code is tested and only a successfully running package is admitted in the archive.

we can observe are with respect to all possible runs of the corpus. To mitigate the threat to validity attached to limited coverage, we obtain clients that will generate multiple *runs* of CRAN packages. In this context, a run is the execution of a CRAN package or Kaggle program with a particular set of input values.

CRAN Packages. The packages that are included in this study are the top 500 packages based on their reverse dependencies. The set of reverse dependencies for some package P is transitive closure of packages that import P. The hypothesis underlying that choice is that packages with a large set of reverse depencies are likely to be higher in quality and their tests will provide better coverage. The packages are implemented in R with some C and Fortran code. Using `cloc`, we measure a total 2M lines of R code and 2.2M lines of native code. For each package, we use its runnable code (tests, etc.) to compute code coverage. On average, the cde coverage is 55.7% which is acceptable without being exhaustive. Figure 1 shows these packages, the size of the dots reflects the project's size in lines of code. The x-axis indicates code coverage in percents and the y-axis gives the number of call sites to `eval` that were traced, in log scale. Dotted lines indicate means. Packages with over 50 eval call sites are named.

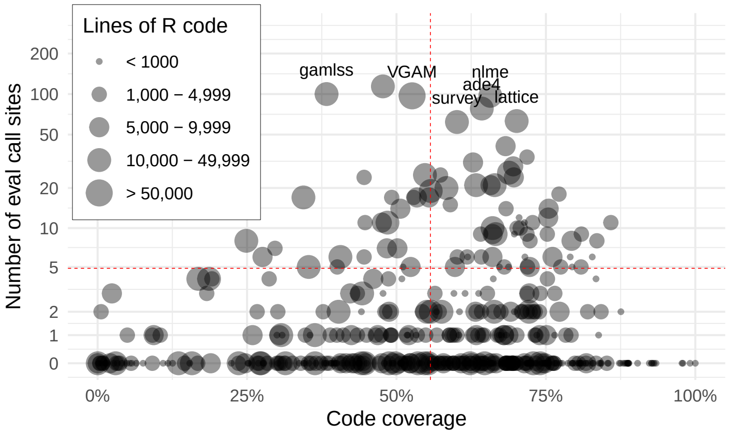

Figure 1 CRAN packages

To analyze a package, we need client code that invokes its methods. There are three sources of built-in runnable code that come with each CRAN package: *tests, examples* and *vignettes*. They are, respectively, traditional unit tests, code snippets from the documentation, and long-form use-cases written in Rmarkdown. Examples and vignettes are automatically extracted and turned into scripts, their input is bundled with the package. Most tests had to be discared due to a limitation of our pipeline: the `testthat` harness uses `eval` and thus causes the entire test to register as an `eval` call. The selected packages are bundled with 16,645 programs; 16.3K examples and 391 vignettes.

Kaggle Scripts. Kaggle is an online platform for data-science and machine-learning. The website allows people to submit data-analysis problems, users compete to find the best solution. The solutions are uploaded to the platform as either plain scripts or notebooks. We chose one of the most popular competition, predicting the survival of passengers the Titanic[5]. Unlike CRAN, Kaggle is not curated. After downloading the 2,890 solutions and extracting

[5] `https://www.kaggle.com/c/titanic`

the R code, we found that 1,042 were duplicates. From the remaining 1,848 solutions, 229 failed to execute. Next to various runtime exceptions, common problems were parse errors and misspelled package names. The final set contains 1,619 programs implemented in 119.1K lines of R code.

3.2 Analysis Pipeline

The results presented in this paper are the result of an automated analysis pipeline that acquires the code of packages, extract metadata, executes programs, traces their behavior and summarizes the observations. Figure 2 shows the main steps of the pipeline along with approximate time to execute each step, the data size, and the number of elements manipulated by the stage. Timings are for runs on an Intel Xeon 6140, 2.30GHz with 72 cores and 256GB of RAM.

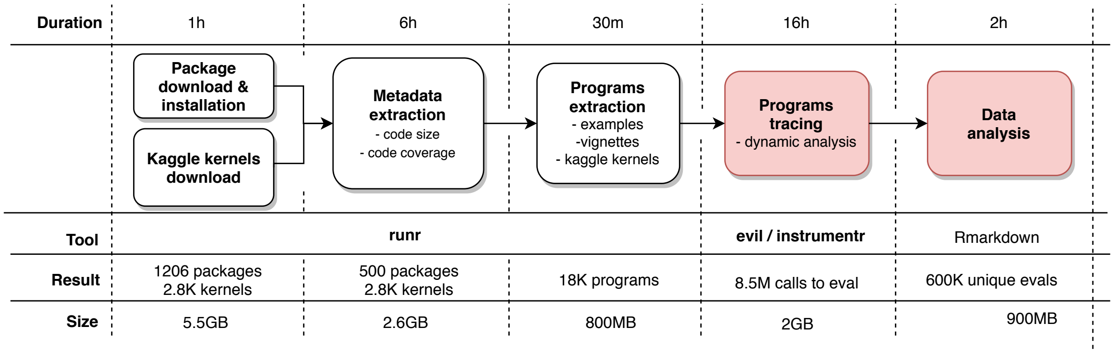

Figure 2 Analysis Pipeline

The first step in the pipeline consist of downloading CRAN packages along with their dependencies and acquiring Kaggle programs with the help of a web crawler. The second step, is to compute code size and coverage metrics for the CRAN packages (we use the *covr* package for coverage). The third step consist in extracting runnable programs from packages: i.e. the tests, examples and vignettes. Each extracted program is wrapped into a call to our dynamic analyzer — the tool is called *evil* for eval inspection library. This step is needed to ensure that we record `eval` usage only for the target package. Without this, the data would include eval calls from the unit testing frameworks as well as from bootstrapping R virtual machine itself. To avoid any interference, each program is run in its own R instance. The fourth step in the pipeline is to perform dynamic analysis for each run of a CRAN package or Kaggle program.

The dynamic analyzer builds upon the dynamic analysis framework, *instrumentr* that we have implemented to enable us to write dynamic analysis logic in R. *instrumentr* serves as an intermediary between *R-dyntrace* and *evil*, it intercepts the hooks exposed by *R-dyntrace* and attaches R functions exported by *evil* as callbacks. The *evil* callbacks execute on corresponding interpreter events.

The data extracted by *evil* from each program is concatenated, cleaned and summarized in the post-processing phase by custom R scripts. Finally, the summarized data is analyzed in RMarkdown notebooks to gather insights. Apart from the figures, the data points included in the paper are also generated by RMarkdown notebooks as latex macros.

The *evil* framework is implemented as a R package in 2K lines of R and 400 lines of C++ code. *instrumentr* is an R package implemented in 2.5K lines of R and 6K lines of C++

code. It internally uses a modified R interpreter, *R-dyntrace* [9], that exposes hooks from within the interpreter implementation for events of interest.

All steps of this pipeline are parallelized using GNU parallel [21] and orchestrated by GNU make. To schedule and parallelize extraction and analysis of programs, we use the *runr* package. Furthermore, *runr* gracefully handles and reports failures across large-scale program runs which greatly aids debugging of the analysis pipeline.

Limitations. Dynamic analysis can only observe calls that are triggered by the program's input. We believe that focusing on R packages with high code coverage does mitigate this to some extent. The results we report here were obtained with R's bytecode compiler turned off, this should not affect the results as the compiler does not optimize `eval`.

. We turn off the bytecode compiler for this study. The bytecode compiler can also call `eval`. We do not get source locations for 285K `eval` calls. In these cases `eval` is either passed as an argument to a higher-order functions or is defined in a function returned by a higher-order function and the R parser does not retain location information for `eval`. However, this is a meager 3.3% of all `eval` calls and is unlikely to affect our analysis. We ignore calls to the native `eval` function exposed by R. We also ignore the `rlang::tidy_eval` function which uses native `eval` internally because `rlang` is used to implement a DSL for data analysis in R. It introduces a new first-class promise object called `quosure` for which it implements special evaluation support in `tidy_eval`.

4 Usage Metrics

This section presents a high-level picture of the usage of `eval` in R.

(CHECK IF TRUE) As none of the Kaggle programs invoke `eval`, the remainder of the paper focuses solely on the CRAN data. We will return to Kaggle in our conclusion.

We recorded 20.4M calls to `eval` and its variants (20.3M for `eval`, 4 for `evalq`, 86.3K for `eval.parent` and 0 for `local`).

4,391 packages in our corpus call `eval` at least once.

Number of eval call sites. We observed that the number of `eval` call sites is small. Figure **??** shows a violin plot of the distribution of call sites to `eval` and its three variants. For packages that have call sites, the median is 3 call sites. 62 packages have a single call site, and one package, *gamlss*, has 107 call sites. 285 packages have no calls sites to `eval`. These numbers are under-approximations as `eval` may be aliased, called reflectively and given as argument to a higher-order function such as `map`.

Number of calls to eval. We observed an average 90.9 calls to `eval` per run. *Note: this is in packages that had any call to eval. How do we compute this? Say we are interested in package P. Do we only count runs that call eval from package P. Or do we also count runs that did not have an eval but executed some code from P? The latter may be easier. Which one is right?* Figure **??** shows a violin plot of the distribution of calls to eval per run. The largest number of calls per run is 2,736.8 and comes from the package minpack.lm.

Amount of code loaded by eval. We use `deparse` to turn the input to `eval` into strings and measure the size of that input. The largest input is 768 and the median input size is 8. Most inputs are small, less than 39 bytes. The average size is 14.5 bytes.

Amount of computation via eval. We compare the number of events intercepted during execution of `eval` across the whole corpus. For this, we use the number of interpreter steps as a proxy for events. We observe 39G *Note: Check that number.* events all told, and 3.6G events occur inside eval (5% of all events). Figure **??** show a violin plot of events.

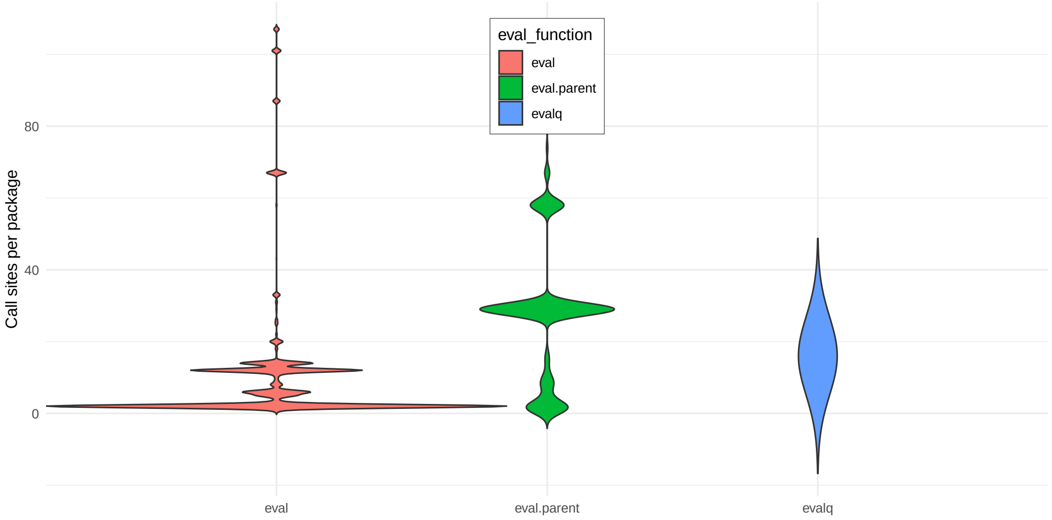

Figure 3 Violin plot of the call sites per packages per eval function.

We observe a very wide spread in the number of events generated by `eval` calls. The largest number of events generated by an `eval` call is 63.5M. However, a majority (96.4%) of `eval` calls performs up to 100 events suggesting that most expressions passed to `eval` are small. It is interesting to note that a higher proportion of Core R `eval` calls perform side effects in the 101–10,000 range. Only 2,997 Core R `eval` calls and 14,209 package `eval` call generate 10,001 to 20,000,000 events. These `eval` calls originate from statistical modeling packages such as `mlogit`, `mboost`, `metafor`, `lavaan`, `mclust` and `gamlss`.

Note: Talk about the percentage of code executed in eval

Eval Arguments. Figure 3 breaks down the expressions passed to `eval` by R data type. It separates the core packages from CRAN as they have different characteristics. In core, XXX036% of arguments in core are expressions. Whereas, in CRAN package XXX037% are expressions. When the argument is not an expression, it is a value. Values are trivial for `eval` as they are returned unchanged. 62% of CRAN arguments are environments. This is an anomaly as only four `eval` call sites dominate this result:

- `ggplot2::ggproto` contributes to 99.9% of package `eval` calls that receive environments as input.
- `R6::generator_funs` function contributes to only 936 `eval` calls and implements the same functionality as `ggplot2::ggproto` but for the *R6* OOP system.
- `future::backtrace` function applies `eval` to a future object which is implemented as an environment. This is called only once.
- `RModel::str.RMmodel` function overloads the core `str` method for its `RModel` objects and maps `eval` exactly like `str`. This is called only once.

The case of CORE libraries ... Note: Give numbers for the same things as above but for the base/core libraries Compare. *Note: Add XXX016 to XXX025.*

5 A Taxonomy of Eval

The previous section gave a quantitative view of the usage of `eval`; in this section we try to elucidate *what* it does.

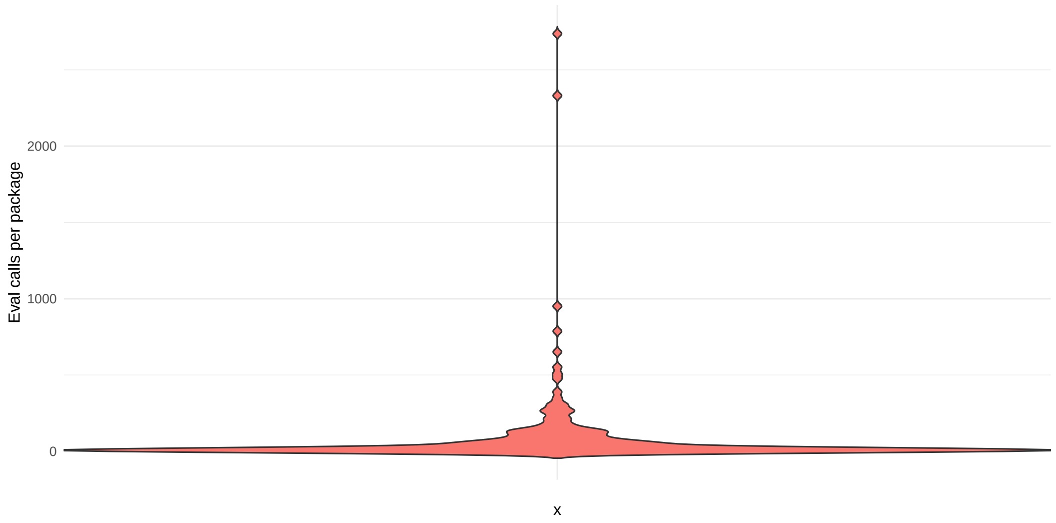

Figure 4 Violin plot of the normalized number of eval calls per package.

Side-effects. XXX026 total number of reads performed by `eval`. *Note: What is the definition of a read? An env lookup?* XXX027 average number of reads. XXX028 total number of writes performed by `eval`. Average number XXX029.

Note: Aviral wrote"measure number of reads and writes by eval outside of its bubble" — what is the bubble here?

XXX029 reads to caller environment – where caller is the caller of `eval`.

XXX030 reads to other caller environment – i.e. envs that are on the call stack.

XXX031 reads to caller environment – where caller is the caller of `eval`.

XXX032 reads to other caller environment – i.e. envs that are on the call stack.

Categorize side-effects between env passed to eval and other envs:

- ephemeral – read to envs that are created during `eval` XXX033
- local – reads to env of the function in which the call to `eval` lexically occurred; XXX034
- parent – reads to parent envs of the local one. XXX035
- call-stack – reads to envs that are on the call stack but not caller. XXX035.

In the corpus we observe 9.4G writes to variables of which 1.1G writes happen inside `eval`. However, all writes are not dangerous. Only writes to environments not local to the computation spawned by `eval` are side-effecting. These writes outlive the computation and hence are visible outside it. The remaining writes are local to the computation. We observe that 45M writes inside evals are side-effecting. This is only 4.3% of all variable writes inside `eval` and .5% of all variable writes in the corpus.

An `eval` is considered side-effecting if it performs a side-effecting write to a variable, directly or indirectly. Only 4.4% `eval` calls in Core R are side-effecting and 7.5% `eval` calls in CRAN packages are side-effecting.

New Bindings. Find number of times, eval introduces new bindings. There are many ways – library, load, attach, source, and explicitly introduce bindings using super-assign, assign and define.

C Calls. Find number of times, eval makes call to native code using - .Call, .External, etc. Find out if there are native functions that are called only from within eval.

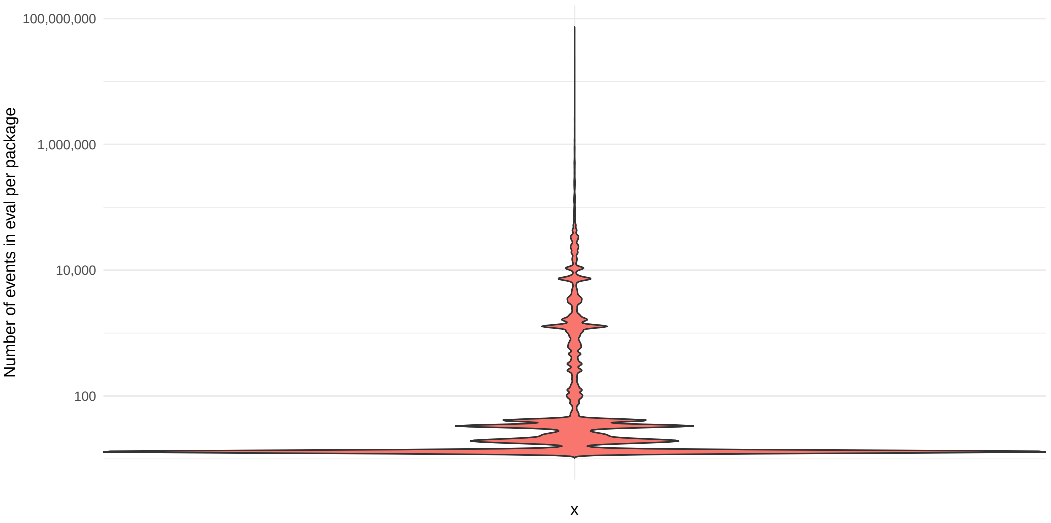

Figure 5 Violin plot of the number of events per packages (log scale).

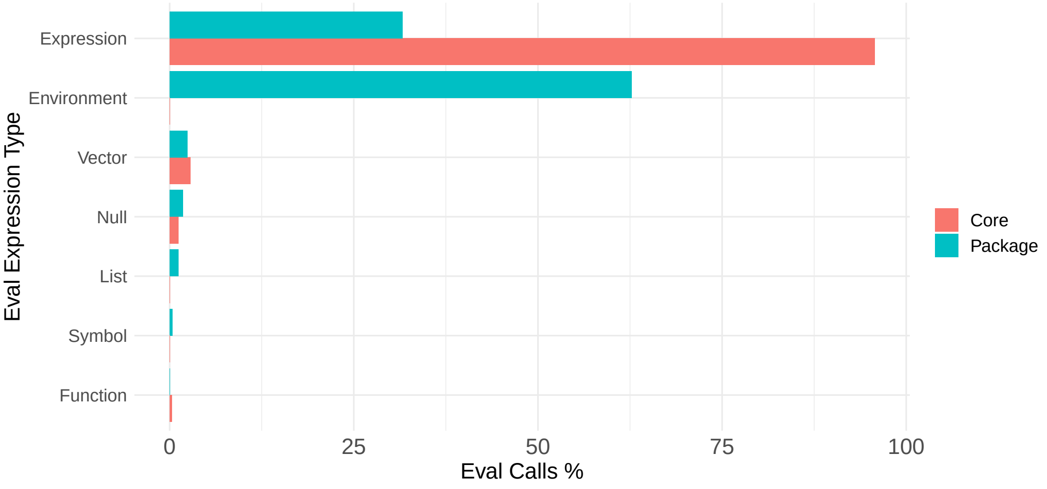

Figure 6 Type of Expression passed to `eval`

Non-local returns Eval can do non-local returns effective bypassing evaluation of the rest of the function. This can be useful for static analyzers.

Purity. Conclude with number of evals that are "pure", i.e. evals which could be ignored by a static analyzer without any problem.

Source. There are various ways to obtain expressions:

- `substitute` synthesizes ASTs from expressions by replacing symbols with their bindings in the specified environment.
- `expression` creates a vector of expression objects from text.
- `parse`, `str2expression` and `str2lang` turn strings into expressions.

We observe 5.2% cases where `eval` directly evaluates the output of `substitute`, 0.7%

cases where output of `parse` is read directly and only 704 cases for `expression`. Most expressions consumed by `eval` are generated by other functions.

`eval(parse(...))` can be used for dynamic code loading. This forms the core of `source` and `sys.source` functions in R that are commonly used for loading code in R files in interactive settings. We investigated the number of cases in which the output of `parse` and its variants is passed to `eval`, directly or transitively by tainting their output. This corresponds to 1% of the total `eval` call sites and .9% of the eval calls. We observed that very few of the eval call (10 in total) consume the result of calling `parse` on a file. Most of the eval calls consume the result of calling `parse` on a string. We also identified one function in core R, `invokeRestartInteractively` that prompts the user for input, parses it, and passes it to `eval`.

Scope. The environment argument can be `NULL`, a `List` or a data frame. This happens in 3.5% of the cases: `eval` copies the fields of the list or data frame and creates bindings for them in a new environment. This pattern is used to evaluate formulas which can directly refer to the fields of the data. The `envir` argument can also be a number n. It means that the environment in which the expressions is evaluated will be the result of `sys.call(n)` where n refers to the n-th stack frame.

The top-level environment in R is called the global environment. New environments can be created using `new.env`. They can be provided a parent environment which becomes the enclosing scope of the new environment.

We looked at the environments passed to all the `eval` calls in our corpus. Table **??** summarizes the results. A numeric environment class `n` denotes the environment of the n-th call stack frame from the current function. `global` denotes the top-level environment and `list` denotes a list passed for evaluation of formulas. Environment classes of the form n+ denote the n-th environment extended with a new environment. The new environment provides a limited form of sandboxing. All assignments using the `<-` function occur inside it and prevent the extended environment from mutation. However, it is still possible to mutate the extended environment using the `<<-` or `assign` functions; but, that happens rarely.

Core		Packages	
Environment Class	Eval Calls %	Environment Class	Eval Calls
2	83%	global+	74%
list	3.8%	3	5.6%
3	3.3%	2	5.3%
global+	3%	5	3.9%
1+	2.9%	list	3.5%

Table 1 Environments in terms of `eval` calls

We observe that a disproportionately high number of core R `eval` calls access the caller's caller's environment. This is because many core R functions call functions that pass the result of `parent.frame()` to an `eval`. A disproportionately high number of calls to `eval` happens in an extended top-level environment. This can be explained by the fact that many packages evaluate code passed from the user's workspace in the top-level environment to access the bindings.

The `imchange` function of package `imager` makes it possible to modify images using a dedicated formula syntax using ~.[6] Here, eval is evaluated in newenv, which creates a new environment that inherits from parent.frame() by default (classified as 1+).

```
newenv <- new.env()
...
fo <- parse(text=as.character(fo)[2])
im[where] <- eval(fo,envir=newenv,enclos=env)
```

adjCoef in package *actuar* find the root of an equation defined by a function h whose arguments must be named x and y. h is transformed into an auxiliary function h2 that can be optimized. Here, the list used for envir ensures the correspondance between the textual arguments of h and the arguments of h2.

```
sh <- substitute(h)
fcall <- paste(sh, "(x,␣y)")
...
h2 <- function(x, y)
    eval(parse(text = fcall),
    envir = list(x = x, y = y),
    enclos = parent.frame(2))
```

6 Case Studies

Note: 5(?) examples from real code?

6.1 LEGACY TEXT

We looked at the top ten expressions passed to core and package eval calls. The most frequent ten expressions to eval calls from core R contribute to 85% of all eval calls.

Expression	Eval Call	%
c("auto", "shell", "radix")	1,987,105	29%
c("auto", "shell", "quick", "radix")	1,593,169	23%
{info <-loadingNamespaceInfo(...	1,008,632	14%
c("onLoad", "attach", "detach", "onUnload")	470,566	6.9%
c("append", "prepend", "replace")	261,587	3.9%
c("left", "right", "centre", "none")	162,086	2.4%
c("no", "ifany", "always")	71,580	1.1%
c("pearson", "kendall", "spearman")	72,962	1.1%
NULL	75,330	1.1%
Symbol	66,279	1%

Table 2 Top ten eval calls in Core

The expression {info <-loadingNamespaceInfo(... is added by core R to a package directory during installation. To load the package, this code is executed. It creates a namespace for the package, injects the package bindings, and attaches the namespace to the program search path. The NULL comes from a call to substitute(subset) in

[6] Inspired by map in package *purr*.

stats::model.frame.default function which has a default value of subset as NULL. The Symbol arises from a call to as.name in base::str function that returns a symbol that is looked up by evaluating it in a specific environment. The remaining cases arise from calls to match.arg which is used to look up the default choices for a variable and match against the choice passed by the caller.

The most frequent ten expressions to eval calls from CRAN packages contribute to 77.1% of all eval calls.

Expression	Eval Call	%
Environment	989302	61%
column[rows] <<-what	55677	3.5%
function(value) freduce(value, ‘_function_list‘)	37251	2.3%
NULL	32005	2%
List	22293	1.4%
c("default", "default2012", "default2011"...	20610	1.3%
force(..1)	20461	1.3%
alist(‘_spec‘)	18532	1.2%
inner	18530	1.2%
String Vector	17487	1.1%

Table 3 Top ten eval calls in CRAN

The expression Environment occurs because of the four callsites explained above, ggplot2::ggproto, R6::generator_funs, future::backtrace and RModel::str.RMmodel. The next expression, column[rows] <<-what, is used inside the plyr::rbind.fill function to merge data frames by assigning concatenated vectors to rows. The <<- operator is interesting in that it skips the current scope and assigns in a parent scope in which the variable is already present. In our corpus, all these eval calls contribute to a single side-effect. The expression function(value) freduce(value, ‘_function_list‘) arises from the magrittr::%>% function which is a pipe operator that pipes the output of previous command to the next one. The expression is evaluated in a custom environment to create a function binding for evaluating the components of the pipe.String Vector and List also arise from the same function when a string or a list is piped using the %>% function into the next expression. The NULL arises from R6::generator_funs function when the eval is passed a NULL argument by the DataMask_generator package. The c("default", "default2012", "default2011"... pattern arises from copula::polyG where it reflectively access the default expression for its formal parameter and evaluates it. The force(..1) and alist(‘_spec‘) patterns occur in glue::glue_data function which concatenates and interpolates strings. The two patterns occur because the function captures unevaluated unnamed arguments and maps the evaluation of force(..1) on them. The force function forces promises and returns the result of evaluation. The inner pattern arises from glue::identity_transformer which enables the creation of custom transformation functions for affecting the interpolation and concatenation of input by the glue package.

7 Why do we Eval

7.1 Discussion

Yeah, why?

7.2 Can we do without?

We look at how *consistent* the `expr` argument of `eval` can be, *i.e.*, how many different types of the resolved `expr` there are per call sites. Most of the call sites, *i.e.*, 99%, are *consistent*, and this is similar to javascript. However, a few ones are highly *polymorphic* (10 different types). They are the pipe operators `%>%`, `%<>%` and `%$%` in package *magrittr*. It is effectively used to compose functions on their first argument, which can be of any type.

Similarly to JavaScript, there are also unnecessary uses of `eval`. For example, the `PerformanceAnalytics` package contains a function `chart.QQPlot` that uses `eval` to resolve a string into function and another to call it and assign its results into a variable:

```
function (R, d="norm", dp, ...) {
q.f <- eval(parse(text=paste("q",d,sep="")))
z <- NULL
eval(parse(text=paste("z<-q.f(",dp,",...)")))
}
```

In both cases, there is no need for `eval`:

```
function (R, d="norm", dp, ...) {
q.f <- get(paste0("q",d))
z <- q.f(dp, ...)
}
```

or even to a oneliner `do.call(paste0("q",d), as.list(dp, ...))`.

7.3 Comparison with Javascript

8 Conclusion

References

1 Julia Belyakova, Benjamin Chung, Jack Gelinas, Jameson Nash, Ross Tate, and Jan Vitek. World Age in Julia: Optimizing Method Dispatch in the Presence of Eval. *Proc. ACM Program. Lang.*, 4(OOPSLA), 2020. `doi:10.1145/3428275`.

2 Jeff Bezanson, Jiahao Chen, Ben Chung, Stefan Karpinski, Viral B. Shah, Jan Vitek, and Lionel Zoubritzky. Julia: Dynamism and Performance Reconciled by Design. *Proc. ACM Program. Lang.*, 2(OOPSLA), 2018. `doi:10.1145/3276490`.

3 Eric Bodden, Andreas Sewe, Jan Sinschek, Hela Oueslati, and Mira Mezini. Taming reflection: Aiding static analysis in the presence of reflection and custom class loaders. In *International Conference on Software Engineering (ICSE)*, 2011. `doi:10.1145/1985793.1985827`.

4 Oscar Callaú, Romain Robbes, Éric Tanter, and David Röthlisberger. How (and why) developers use the dynamic features of programming languages: the case of Smalltalk. *Empir. Softw. Eng.*, 18(6), 2013. `doi:10.1007/s10664-012-9203-2`.

5 Zhifei Chen, Wanwangying Ma, Wei Lin, Lin Chen, Yanhui Li, and Baowen Xu. A study on the changes of dynamic feature code when fixing bugs: towards the benefits and costs of Python dynamic features. *Sci. China Inf. Sci.*, 61(1), 2018. `doi:10.1007/s11432-017-9153-3`.

6 Aske Simon Christensen, Anders Møller, and Michael Schwartzbach. Precise analysis of string expressions. In *Static Analysis Symposium (SAS)*, 2003. `doi:10.1007/3-540-44898-5\_1`.

7 Patrick Cousot and Radhia Cousot. Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints. In *Principles of Programming Languages (POPL)*, 1977. `doi:10.1145/512950.512973`.

8 Olivier Flückiger, Guido Chari, Jan Jecmen, Ming-Ho Yee, Jakob Hain, and Jan Vitek. R Melts Brains: An IR for First-Class Environments and Lazy Effectful Arguments. In *International Symposium on Dynamic Languages (DLS)*, 2019. `doi:10.1145/3359619.3359744`.

9 Aviral Goel and Jan Vitek. On the Design, Implementation, and Use of Laziness in R. *Proc. ACM Program. Lang.*, 3(OOPSLA), 2019. doi:10.1145/3360579.

10 Liang Gong. *Dynamic Analysis for JavaScript Code.* PhD thesis, University of California, Berkeley, 2018. URL: http://www.escholarship.org/uc/item/7n30n4kd.

11 Ross Ihaka and Robert Gentleman. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996. URL: http://www.amstat.org/publications/jcgs/.

12 Simon Holm Jensen, Peter A. Jonsson, and Anders Møller. Remedying the Eval That Men Do. In *International Symposium on Software Testing and Analysis (ISSTA)*, 2012. doi:10.1145/2338965.2336758.

13 Uwe Ligges. 20 Years of CRAN (video on channel9). In *Keynote at UseR!*, 2017. URL: https://channel9.msdn.com/Events/useR-international-R-User-conferences/useR-International-R-User-2017-Conference/KEYNOTE-20-years-of-CRAN.

14 Benjamin Livshits, Manu Sridharan, Yannis Smaragdakis, Ondřej Lhoták, J. Nelson Amaral, Bor-Yuh Evan Chang, Samuel Z. Guyer, Uday P. Khedker, Anders Møller, and Dimitrios Vardoulakis. In Defense of Soundiness: A Manifesto. *Commun. ACM*, 58(2), January 2015. doi:10.1145/2644805.

15 John McCarthy. History of LISP. In *History of programming languages (HOPL)*, 1978. doi:10.1145/960118.808387.

16 Fadi Meawad, Gregor Richards, Floréal Morandat, and Jan Vitek. Eval Begone!: Semi-Automated Removal of Eval from Javascript Programs. In *Conference on Object-Oriented Programming, Systems, Languages, and Applications, (OOPSLA)*, 2012. doi:10.1145/2384616.2384660.

17 Floréal Morandat, Brandon Hill, Leo Osvald, and Jan Vitek. Evaluating the Design of the R Language: Objects and Functions for Data Analysis. In *European Conference on Object-Oriented Programming (ECOOP)*, 2012. doi:10.1007/978-3-642-31057-7_6.

18 R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2017. URL: https://www.R-project.org/.

19 Gregor Richards, Christian Hammer, Brian Burg, and Jan Vitek. The Eval that Men Do: A Large-scale Study of the Use of Eval in JavaScript Applications. In *European Conference on Object-Oriented Programming (ECOOP)*, 2011. doi:10.1007/978-3-642-22655-7_4.

20 Gregor Richards, Sylvain Lesbrene, Brian Burg, and Jan Vitek. An Analysis of the Dynamic Behavior of JavaScript Programs. In *Proceedings of the ACM Programming Language Design and Implementation Conference (PLDI)*, 2010.

21 Ole Tange. *GNU Parallel.* Ole Tange, 2018. doi:10.5281/zenodo.1146014.

22 Beibei Wang, Lin Chen, Wanwangying Ma, Zhifei Chen, and Baowen Xu. An empirical study on the impact of Python dynamic features on change-proneness. In *International Conference on Software Engineering and Knowledge Engineering*, 2015. doi:10.18293/SEKE2015-097.

23 Hadley Wickham. *Advanced R.* Chapman & Hall, 2014. doi:10.1201/b17487.