A live coding environment for a data exploration language

Tomas Petricek[a]

a University of Kent, UK

Note to reviewers: This paper has been previously submitted to a different venue. The revisions made in response to past reviews are summarized in Appendix E (to be removed for publication).

Abstract A growing amount of code is written to explore and analyze data, often by data analysts who do not have a traditional background in programming, for example by journalists. The way such data anlysts write code is different from the way software engineers do so. They use few abstractions, work interactively and rely heavily on external libraries. We aim to capture this way of working and build a programming environment that makes data exploration easier by providing instant live feedback.

We combine theoretical and applied approach. We present the *data exploration calculus*, a formal language that captures the structure of code written by data analysts. We then implement a data exploration environment that evaluates code instantly during editing and shows previews of the results. We formally describe an algorithm for providing instant previews for the data exploration calculus that allows the user to modify code in an unrestricted way in a text editor. Supporting interactive editing is tricky as any edit can change the structure of code and fully recomputing the output would be too expensive. We prove that our algorithm is correct and that it reuses previous results when updating previews after a number of common code edit operations. We also illustrate the practicality of our approach with an empirical evaluation and a case study.

As data analysis becomes an ever more important use of programming, research on programming languages and tools needs to consider new kinds of programming workflows appropriate for those domains and conceive new kinds of tools that can support them. The present paper is one step in this important direction.

ACM CCS 2012

- *Human-centered computing → Interactive systems and tools;*
- *Information systems → Data mining;*
- *Software and its engineering → Compilers;*

Keywords Data exploration, live coding, data journalism, instant previews

The Art, Science, and Engineering of Programming

Perspective The Theoretical Science of Programming

Area of Submission Programming environments

1 Introduction

One of the aspects that make spreadsheets easier to use than other programming tools is their liveness. When you change a cell in Excel, the whole spreadsheet updates instantly and you immediately see new results, without having to explicitly trigger re-computation and without having to wait for an extensive period of time.

An increasing number of programming environments aim to provide a live development experience for standard programming languages, but doing this is not easy. Fully recomputing the whole program after every keystroke is inefficient and calculating how a change in the source code changes the result is extremely hard when the text editor allows arbitrary changes. Consider the following snippet that gets the release years of the 10 most expensive movies from a data set movies:

```
let top = movies.sortBy(λx → x.getBudget())
    .take(10).map(λx → x.getReleased().format("yyyy"))
```

A live coding environment computes and displays the list of years. Suppose that the programmer then makes 10 a variable and changes the date format:

```
let count = 10
let top = movies.sortBy(λx → x.getBudget())
    .take(count).map(λx → x.getReleased().format("dd-mm-yyyy"))
```

Ideally, the live coding environment should understand the change, reuse a cached result of the first two transformations (sorting and taking the first 10 elements) and only evaluate the map operation to differently format the release dates of top 10 movies. Our environment does this for a simple data exploration language in an unrestricted text editor. We discuss related work in Section 7, but we briefly review the most important directions here, in order to situate our contributions.

1.1 Related work

Simple data exploration performed, for example, by journalists [21] is done either programmatically or using spreadsheets. The latter is easy but error-prone while the former requires expert programming skills. We aim to bring liveness of spreadsheets to programmatic data exploration. This requires extending recomputation as done in spreadsheets [52] to code written in an ordinary text editor.

Many visual data exploration tools support interactivity [11, 25, 61] and some can even export the workflow as a script [28]. Data analysts who prefer working with code typically resort to notebook systems such as Jupyter or R Markdown [5, 29]. Those are text-based, but have a limited model of recomputation. Users structure code in cells and manually reevaluate cells. Many notebook systems do not track dependencies between cells which can lead to well-documented inconsistencies [30, 48, 50].

Existing text-based editors that provide instant feedback, such as Lighttable [20] and Chrome DevTools, often work only in certain situations or require full recomputation. A more principled approach can be used by systems based on structured editing [33, 42] where code is only modified via known operations with known effect on the computation graph (e.g. "extract variable" has no effect on the result; "change constant value" forces recomputation of subsequent code). We aim to cater for the many users who prefer to edit programs as free-form text. The technical aspects of our implementation are related to incremental computation [2, 23]. This focuses on cases where *data changes* but program stays the same, while our focus is on cases where *program changes* but data stays the same.

1.2 Contributions

We present the design and implementation of a live coding environment for a simple data exploration language that provides correct and efficient instant feedback, yet is integrated into an ordinary text editor. Our key contributions are:

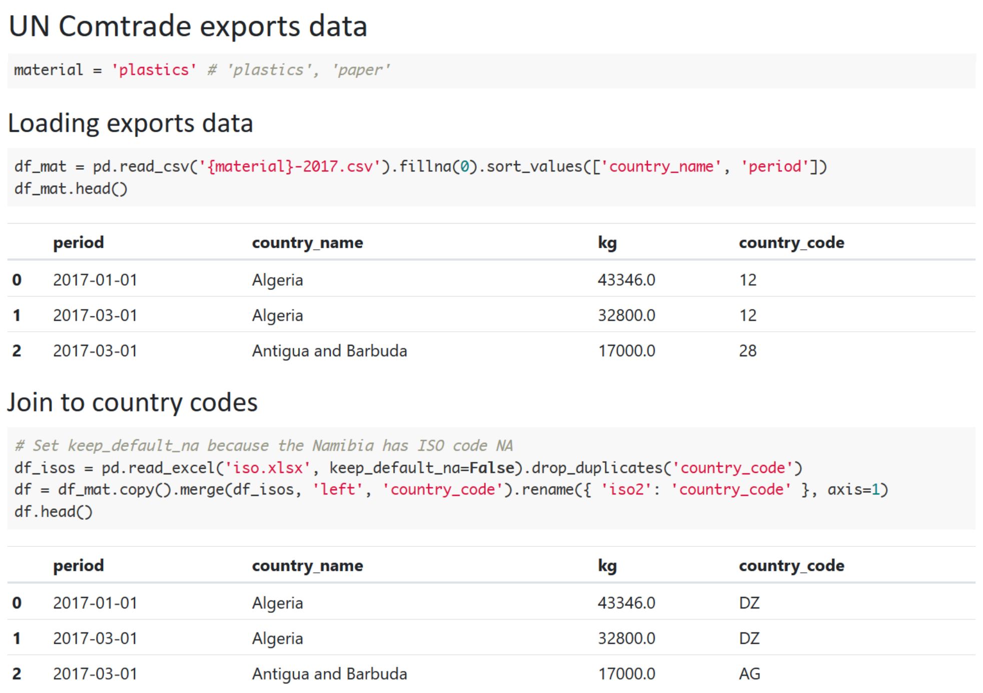
- We introduce the *data exploration calculus* (Section 3), a small formally tractable language for data exploration. The calculus is motivated by our review of how data analysts work (Section 2) and makes it possible to write simple, transparent and reproducible data analyses.
- A live coding environment does not operate in batch mode and so we cannot follow classic compiler literature. We capture the essence of such new perspective (Section 4) and use it to build our *live preview* mechanism (Section 5) that evaluates code instantly during editing.
- We prove that our live preview mechanism is correct (Section 6.1) and that it reuses previously evaluated values when possible (Section 6.2). We illustrate the practicality of the mechanism using an empirical evaluation (Section 6.3) and a case study (Section 6.4).

2 Understanding how data scientists work

Data scientists often use general-purpose programming languages such as Python, but the kind of code they write and the way they interact with the system is very different from how software engineers work [22]. This paper focuses on simple data wrangling and data exploration as done, for example, by journalists analysing government datasets. This new kind of non-expert programmers is worth our attention as they often work on informing the public. They need easy-to-use tools, but not necessarily a full programming langugae. In this section, we ilustrate how such data analyses look and we provide a justification for the design of our data exploration calculus.

2.1 Simple data exploration in Jupyter

Data analysts increasingly use notebook systems such as Jupyter, which make it possible to combine text, equations and code with results of running the code, such as tables or visualizations. Notebooks blur the conventional distinction between development and execution. Data analysts write small snippets of code, run them to see results immediately and then revise them.


UN Comtrade exports data

```
material = 'plastics' # 'plastics', 'paper'
```

Loading exports data

```
df_mat = pd.read_csv('{material}-2017.csv').fillna(0).sort_values(['country_name', 'period'])
df_mat.head()
```

	period	country_name	kg	country_code
0	2017-01-01	Algeria	43346.0	12
1	2017-03-01	Algeria	32800.0	12
2	2017-03-01	Antigua and Barbuda	17000.0	28

Join to country codes

```
# Set keep_default_na because the Namibia has ISO code NA
df_isos = pd.read_excel('iso.xlsx', keep_default_na=False).drop_duplicates('country_code')
df = df_mat.copy().merge(df_isos, 'left', 'country_code').rename({ 'iso2': 'country_code' }, axis=1)
df.head()
```

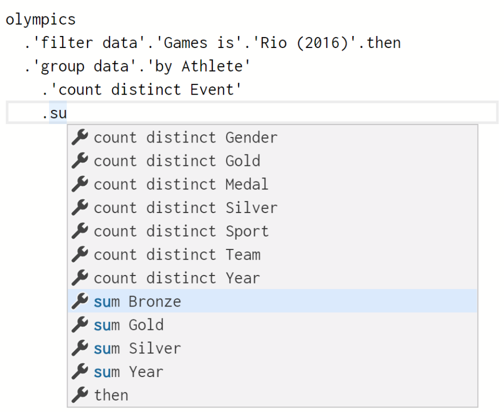
	period	country_name	kg	country_code
0	2017-01-01	Algeria	43346.0	DZ
1	2017-03-01	Algeria	32800.0	DZ
2	2017-03-01	Antigua and Barbuda	17000.0	AG

Figure 1 Financial Times analysis that joins UN trade database with ISO country codes.

Notebooks are used by users ranging from scientists who implement complex models of physical systems to journalists who perform simple data aggregations and create visualizations. Our focus is on the simplest use cases. Making programmatic data exploration more spreadsheet-like should encourage users to choose programming tools over spreadsheets, resulting in more reproducible and transparent data analyses.

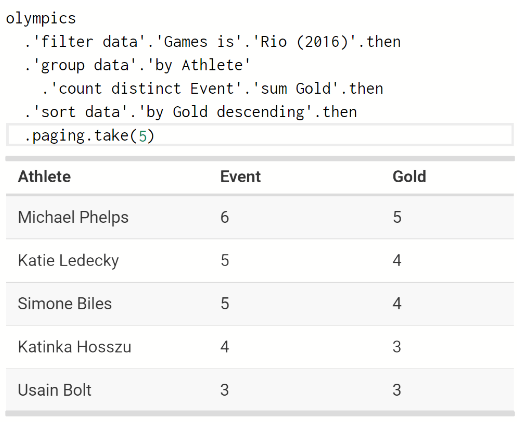
Consider the Financial Times analysis of plastic waste [7, 27]. It joins datasets from Eurostat, UN Comtrade and more, aggregates the data and builds a visualization comparing waste flows in 2017 and 2018. Figure 1 shows an excerpt from one notebook of the data analysis. The code has a number of important properties:

- There is no abstraction. The analysis uses lambda functions as arguments to library calls, but it does not define any custom functions. Code is parameterized by having a global variable material set to "plastics" and keeping other possible values in a comment. This lets the analyst run and check results of intermediate steps.
- The code relies on external libraries. Our example uses Pandas [36], which provides operations for data wrangling such as merge to join datasets or drop_duplicates to delete rows with duplicate column values. Such standard libraries are external to the data analysis and are often implemented in another language like C++.
- The code is structured as a sequence of commands. Some commands define a variable, either by loading data, or by transforming data loaded previously. Even in Python, data is often treated as immutable. Other commands produce an output that is displayed in the notebook.
- There are many corner cases, such as the fact that keep_default_na needs to be set to handle Namibia correctly. These are discovered interactively by running the code and examining the output, so providing an instant feedback is essential.



```
olympics
  .'filter data'.'Games is'.'Rio (2016)'.then
  .'group data'.'by Athlete'
    .'count distinct Event'
    .su
```

(a) The analysis counts the number of distinct events per athlete. After typing '.' the editor offers further aggregation operations.



```
olympics
  .'filter data'.'Games is'.'Rio (2016)'.then
  .'group data'.'by Athlete'
    .'count distinct Event'.'sum Gold'.then
  .'sort data'.'by Gold descending'.then
  .paging.take(5)
```

Athlete	Event	Gold
Michael Phelps	6	5
Katie Ledecky	5	4
Simone Biles	5	4
Katinka Hosszu	4	3
Usain Bolt	3	3

(b) Our live coding environment for The Gamma. The table is updated on-the-fly a and shows the result at the current cursor position.

Figure 2 Previous work on The Gamma (left) and our new editor with live preview (right).

Many Jupyter notebooks are more complex than the above example and might use helper functions or object-oriented code. However, simple data analyses such as the one discussed here are frequent enough and pose interesting problems for programming tools. This paper aims to bring such analyses to the attention of programming research community by capturing their essential properties as a formal calculus.

2.2 Dot-driven data exploration in The Gamma

Simple data exploration as discussed in the previous section has been the motivation for a scripting language The Gamma [47]. Scripts in The Gamma are sequences of commands that either define a variable or produce an output. It does not support top-level functions and lambda functions can be used only as method arguments.

Given the limited expressiveness of The Gamma, libraries are implemented in other languages, such as JavaScript. The Gamma uses type providers [55] for accessing and exploring external data sources. Type providers generate object types with members and The Gamma makes using those convenient by providing auto-complete when the user types the dot symbol ('.') to access a member. The combination of type providers and auto-complete makes it possible to solve a large number of data exploration tasks through the very simple interaction of selecting operations from a list.

The example in Figure 2a summarizes data on Olympic medals. Identifiers such as 'sum Bronze' are names of members generated by the type provider. The type provider used in this example generates an object with members for data transformations such as 'group data', which return further objects with members for specifying transformation parameters, such as selecting the grouping key using 'by Athlete'.

The Gamma language is richer, but the example in Figure 2a shows that non-trivial data exploration can be done using a very simple language. The assumptions about structure of code that are explicit in The Gamma are implicitly present in Python and R data analyses produced by journalists, economists and other users with other than programming background. When we refer to The Gamma in this paper, readers not familiar with it can consider a small subset of Python.

2.3 Live coding for data exploration

The implementation that accompanies this paper builds a live coding environment for The Gamma. The implementation is discussed in Section 6.4 and it replaces the original text editor with just auto-complete with a live coding environment that provides *live previews*. We choose to extend The Gamma over using a subset of Python as the simplicity of TheGamma lets us focus on the principles of live coding systems.

An example of a live preview is shown in Figure 2b. As noted earlier, The Gamma programs consist of lists of commands which are either expressions or let bindings. Our editor displays a live preview below the command that the user is currently editing. The preview shows the result of evaluating the expression or the value assigned to a bound variable. When the user changes the code, the preview is updated automatically, without any additional interaction with the user.

There are a number of guiding principles that inform our design. First, we allow the analyst to edit code in an unrestricted form in a text editor. Although structured editors provide an appealing alternative and make recomputation easier, we prefer the flexibility of plain text. Second, we focus on the scenario when code changes, but input does not. Rapid feedback allows the analyst to quickly adapt code to correctly handle corner cases that typical analysis involves. In contrast to work on incremental computation, we do not consider the case when data changes, although supporting interactive data exploration of streaming data is an interesting future direction.

3 Data exploration calculus

The *data exploration calculus* is a small formal language for data exploration. The calculus is not, in itself, Turing-complete and it can only be used together with external libraries that define what objects are available and what the behaviour of their members is. This is sufficient to capture the simple data analyses discussed in Section 2. We define the calculus in this section and then use it to formalise our live preview mechanism in Section 4. The live preview mechanism does not rely on types and so we postpone the discussion of static typing to Section C. Interestingly, it reuses the mechanism used for live previews.

3.1 Language syntax

The calculus combines object-oriented features such as member access with functional features including lambda functions. The syntax is defined in Figure 3. Object values o are defined by external libraries that are used in conjunction with the core calculus.

A program p in the data exploration calculus consists of a sequence of commands c. A command can be either a let binding or a term. Let bindings define variables x that can be used in subsequent commands. Lambda functions can only appear as arguments in method calls. A term t can be a value, variable or a member access, while an expression e, which can appear as an argument in member access, can be a lambda function or a term.

Programs, commands, terms, expressions and values

$$\begin{array}{lll} p ::= c_1;\ldots;c_n & t ::= o \mid x & e ::= t \mid \lambda x \to e \\ c ::= \mathsf{let}\ x = t \mid t & \quad\ \mid \ t.m(e,\ldots,e) & v ::= o \mid \lambda x \to e \end{array}$$

Evaluation contexts of expressions

$$\begin{array}{lll} C_e[-] & = & C_e[-].m(e_1,\ldots,e_n) \mid o.m(v_1,\ldots,v_m,C_e[-],e_1,\ldots,e_n) \mid - \\ C_c[-] & = & \mathsf{let}\ x = C_e[-] \mid C_e[-] \\ C_p[-] & = & o_1;\ \ldots;\ o_k;\ C_c[-];\ c_1;\ \ldots;\ c_n \end{array}$$

Let elimination and member reduction

$$\begin{array}{ll} o_1;\ \ldots;\ o_k;\ \mathsf{let}\ x = o;\ c_1;\ \ldots;c_n \rightsquigarrow & \\ \qquad o_1;\ \ldots;\ o_k;\ o; c_1[x \leftarrow o];\ \ldots;c_n[x \leftarrow o] & (\mathsf{let}) \\ o.m(v_1,\ldots,v_n) \rightsquigarrow_\epsilon o' \implies C_p[o.m(v_1,\ldots,v_n)] \rightsquigarrow C_p[o'] & (\mathsf{external}) \end{array}$$

■ **Figure 3** Syntax, contexts and reduction rules of the data exploration calculus

3.2 Operational semantics

The data exploration calculus is a call-by-value language. We model evaluation as a small-step reduction $\rightsquigarrow$. Fully evaluating a program results in an irreducible sequence of objects $o_1;\ \ldots;\ o_n$ (one object for each command, including let bindings) which can be displayed as intermediate results of the data analysis. The operational semantics is parameterized by a relation $\rightsquigarrow_\epsilon$ that models the functionality of the external libraries used with the calculus and defines the reduction behaviour for member accesses. The relation has the following form:

$$o_1.m(v_1,\ldots,v_n) \rightsquigarrow_\epsilon o_2$$

Here, the operation m is invoked on an object and takes values (objects or function values) as arguments. The reduction always results in an object. Figure 3 defines the reduction rules in terms of $\rightsquigarrow_\epsilon$ and evaluation contexts; C_e specifies left-to-right evaluation of arguments of a method call, C_c specifies evaluation of a command and C_p defines left-to-right evaluation of a program. The rule (external) calls a method provided by an external library in a call-by-value fashion and (let) substitutes a value of an evaluated variable in all subsequent commands and leaves the result in the list of commands. Note that our semantics does not define how λ functions are reduced. This is done by external libraries, which will typically supply functions with arguments using standard β-reduction. The behaviour is subject to constraints discussed next.

3.3 Properties

The data exploration calculus has a number of desirable properties. Some of those require that the relation $\rightsquigarrow_\epsilon$, which defines evaluation for external libraries, satisfies a number of conditions. We discuss *normalization* and *let elimination* in this section. Those two are particularly important as they will allow us to prove correctness of our method of evaluating live previews in Section 6.1.

Definition 1 (Further reductions)**.** We define two additional reduction relations:

- We write $\rightsquigarrow^*$ for the reflexive, transitive closure of $\rightsquigarrow$
- We write $\rightsquigarrow_{\text{let}}$ for a call-by-name let binding elimination $c_1;\ \ldots;\ c_{k-1};$ $\text{let}\ x = t;\ c_{k+1};\ \ldots; c_n \ \rightsquigarrow_{\text{let}}\ c_1;\ \ldots;\ c_{k-1};\ t; c_{k+1}[x \leftarrow t];\ \ldots; c_n[x \leftarrow t]$

We say that two expressions e and e' are *observationally equivalent* if, for any context C, the expressions $C[e]$ and $C[e']$ reduce to the same value. Lambda functions $\lambda x \rightarrow 2$ and $\lambda x \rightarrow 1+1$ are not equal, but they are observationally equivalent. We require that external libraries satisfy two conditions. First, when a method is called with observationally equivalent values as arguments, it should return the same value. Second, the evaluation of $o.m(v_1, \ldots, v_n)$ should be defined for all o, n and v_i. The external library will typically satisfy this by defining an error object and reducing all invalid calls to the error object [37].

Definition 2 (External library)**.** An external library consists of a set of objects O and a reduction relation $\rightsquigarrow_\epsilon$ that satisfies the following two properties:

- *Totality.* For all o, m, i and all $v_1, \ldots, v_i$, there exists o' such that $o.m(v_1, \ldots, v_i) \rightsquigarrow_\epsilon o'$.
- *Compositionality.* For observationally equivalent arguments, the reduction should always return the same object, i.e. given $e_0, e_1, \ldots, e_n$ and $e'_0, e'_1, \ldots, e'_n$ and m such that $e_0.m(e_1, \ldots, e_n) \rightsquigarrow^* o$ and $e'_0.m(e'_1, \ldots, e'_n) \rightsquigarrow^* o'$ then if for any contexts $C_0, C_1, \ldots, C_n$ it holds that if $C_i[e_i] \rightsquigarrow^* o_i$ and $C_i[e'_i] \rightsquigarrow^* o_i$ for some o_i then $o = o'$.

Compositionality is essential for proving the correctness of our live preview mechanism. Totality allows us to prove normalization, i.e. all programs reduce to a value – although the resulting value may be an error value provided by the external library.

Theorem 1 (Normalization)**.** *For all p, there exists $n, o_1, \ldots, o_n$ such that $p \rightsquigarrow^* o_1; \ldots; o_n$.*

Proof. A program that is not a sequence of values can be reduced and reduction decreases the size of the program. See Appendix A.1 for more detail. □

Although the reduction rules (let) and (external) of the data exploration calculus define an evaluation in a call-by-value order, eliminating let bindings in a call-by-name way using the $\rightsquigarrow_{\text{let}}$ reduction does not affect the result. This simplifies our later proof of live preview correctness in Section 6.1.

Lemma 2 (Let elimination for a program)**.** *Given any program p such that $p \rightsquigarrow^* o_1; \ldots; o_n$ for some n and $o_1, \ldots, o_n$ then if $p \rightsquigarrow_{\text{let}} p'$ for some p' then also $p' \rightsquigarrow^* o_1; \ldots; o_n$.*

Proof. By constructing $p' \rightsquigarrow^* o_1; \ldots; o_n$ from $p \rightsquigarrow^* o_1; \ldots; o_n$. See Appendix A.2. □

4 Formalising a live coding environment

A naive way of providing live previews during code editing would be to re-evaluate the code after each change. This would be wasteful – when writing code to explore data, most changes are additive. To update a preview, we only need to evaluate newly added code. We describe an efficient mechanism in this section.

$$\begin{tikzcd}
\mathsf{val}(10) & \mathsf{val}(15) \\
\mathsf{mem}(\mathsf{skip}, s_0) \arrow[u, "\mathsf{arg}(1)"] \arrow[d, "\mathsf{arg}(0)"'] & \mathsf{mem}(\mathsf{take}, s_1) \arrow[l, "\mathsf{arg}(0)"'] \arrow[u, "\mathsf{arg}(1)"'] \\
\mathsf{var}(\mathsf{data}) &
\end{tikzcd}$$

(a) Graph constructed from initial expression: let $x = 15$ in data.skip(10).take(x)

$$\begin{tikzcd}
\mathsf{val}(10) & \\
\mathsf{mem}(\mathsf{skip}, s_0) \arrow[u, "\mathsf{arg}(1)"] \arrow[d, "\mathsf{arg}(0)"'] & \mathsf{mem}(\mathsf{take}, s_2) \arrow[l, "\mathsf{arg}(0)"'] \arrow[ul, bend right, "\mathsf{arg}(1)"'] \\
\mathsf{var}(\mathsf{data}) &
\end{tikzcd}$$

(b) Updated graph after changing x to 10: let $x = 10$ in data.skip(10).take(x)

Figure 4 Dependency graphs formed by two steps of the live programming process.

4.1 Maintaining dependency graph

The key idea behind our method is to maintain a dependency graph [32] with nodes representing individual operations of the computation that can be evaluated to obtain a preview. Each time the program text is modified, we parse it afresh (using an error-recovering parser) and bind the abstract syntax tree to the dependency graph. When binding a new expression to the graph, we reuse previously created nodes as long as they have the same structure and the same dependencies. For expressions that have a new structure, we create new nodes.

The nodes of the graph serve as unique keys into a lookup table containing previously evaluated parts of the program. When a preview is requested for an expression, we use the graph node bound to the expression to find a preview. If a preview has not been evaluated, we force the evaluation of all dependencies in the graph and then evaluate the operation represented by the current node.

4.1.1 Elements of the graph

The nodes of the graph represent individual operations of the computation. In our design, the nodes are used as cache keys, so we attach a unique symbol s to some of the nodes. That way, we can create two unique nodes representing, for example, access to a member named take which differ in their dependencies.

The graph edges are labelled with labels indicating the kind of dependency. For a method call, the labels are "first argument", "second argument" and so on. Writing s for symbols and i for integers, nodes (vertices) v and edge labels l are defined as:

$$\begin{array}{lcll} v & = & \mathsf{val}(o) \mid \mathsf{var}(x) \mid \mathsf{mem}(m,s) \mid \mathsf{fun}(x,s) & (\textit{Vertices}) \\ l & = & \mathsf{body} \mid \mathsf{arg}(i) & (\textit{Edge labels}) \end{array}$$

The val node represents a primitive value and contains the object itself. Two occurrences of 10 in the source code will be represented by the same node. Member access mem contains the member name, together with a unique symbol s – two member access nodes with different dependencies will contain a different symbol. Dependencies of member access are labelled with arg indicating the index of the argument (0 for the instance and 1, . . . for the arguments). Finally, nodes fun and var represent function values and variables bound by λ abstraction.

4.1.2 Example graph

Figure 4 illustrates how we build the dependency graph. Node representing take(x) depends on the argument – the number 15 – and the instance, which is a node representing skip(10). This depends on the instance data and the number 10. Note that variables bound via let binding such as x do not appear as var nodes. The node using it depends directly on the node representing the expression assigned to x.

After changing the value of x, we create a new graph. The dependencies of the node mem(skip, s_0) are unchanged and so the symbol s_0 attached to the node remains the same and previously computed previews can be reused. This part of the program is not recomputed. The arg(1) dependency of the take call changed and so we create a new node mem(skip, s_2) with a fresh symbol s_2. The preview for this node is then computed as needed using the already known values of its dependencies.

4.1.3 Reusing graph nodes

The binding process takes an expression and constructs a dependency graph. It uses a lookup table to reuse previously created member access and function value nodes. The key in the lookup table is formed by a node kind together with a list of dependencies. A node kind includes the member or variable name; a lookup table Δ then maps a node kind with a list of dependencies to a cached node:

$$\begin{array}{ll} k \;::=\; \mathsf{fun}(x) \mid \mathsf{mem}(m) & (\textit{Node kinds}) \\ \Delta(k, [(v_1, l_1), \ldots, (v_n, l_n)]) & (\textit{Lookup for a node}) \end{array}$$

The example on the second line looks for a node of a kind k that has dependencies $v_1, \ldots, v_n$ labelled with labels $l_1, \ldots, l_n$. We write $\Delta(k, l)\downarrow$ when a value for a key is not defined. When creating the graph in Figure 4b, we perform the following lookups:

$$\Delta(\mathsf{mem}(\mathsf{skip}), [(\mathsf{var}(\mathsf{data}), \mathsf{arg}(0)), (\mathsf{val}(10), \mathsf{arg}(1))]) \qquad (1)$$
$$\Delta(\mathsf{mem}(\mathsf{take}), [(\mathsf{mem}(\mathsf{skip}, s_0), \mathsf{arg}(0)), (\mathsf{val}(10), \mathsf{arg}(1))]) \qquad (2)$$

First, we look for the skip member access. The result is the mem(skip, s_0) known from the previous step. We then look for the take member access. In the earlier step, the argument of take was 15 rather than 10 and so this lookup fails. We then construct a new node mem(take, s_2) and later add it to the cache.

4.2 Binding expressions to a graph

After parsing modified code, we update the dependency graph and link each node of the abstract syntax tree to a node of the dependency graph. This process is called binding and is defined by the bind-expr function (Figure 5) and bind-prog function (Figure 6). Both functions are annotated with a lookup table Δ and a variable context Γ. The variable context is a map from variable names to dependency graph nodes and is used for variables bound using let binding.

When invoked, bind-expr$_{\Gamma,\Delta}(e)$ returns a node v that corresponds to the expression e, paired with a dependency graph (V, E) formed by nodes V and labelled edges E. That edges are written as (v_1, v_2, l) and include a label l. The bind-prog$_{\Gamma,\Delta}$ function works similarly, but turns a sequence of commands into a sequence of nodes.

$$\text{bind-expr}_{\Gamma,\Delta}(e_0.m(e_1,\ldots,e_n)) = v, (\{v\} \cup V_0 \cup \ldots \cup V_n, E \cup E_0 \cup \ldots \cup E_n) \tag{1}$$
$$\quad \text{when } v_i, (V_i, E_i) = \text{bind-expr}_{\Gamma,\Delta}(e_i) \text{ and } v = \Delta(\text{mem}(m), [(v_0, \text{arg}(0)), \ldots, (v_n, \text{arg}(n))])$$
$$\quad \text{let } E = \{(v, v_0, \text{arg}(0)), \ldots, (v, v_n, \text{arg}(n))\}$$

$$\text{bind-expr}_{\Gamma,\Delta}(e_0.m(e_1,\ldots,e_n)) = v, (\{v\} \cup V_0 \cup \ldots \cup V_n, E \cup E_0 \cup \ldots \cup E_n) \tag{2}$$
$$\quad \text{when } v_i, (V_i, E_i) = \text{bind-expr}_{\Gamma,\Delta}(e_i) \text{ and } \Delta(\text{mem}(m), [(v_0, \text{arg}(0)), \ldots, (v_n, \text{arg}(n))])\downarrow$$
$$\quad \text{let } v = \text{mem}(m, s), s \text{ fresh and } E = \{(v, v_0, \text{arg}(0)), \ldots, (v, v_n, \text{arg}(n))\}$$

$$\text{bind-expr}_{\Gamma,\Delta}(\lambda x \to e) = v, (\{v\} \cup V, \{e\} \cup E) \tag{3}$$
$$\quad \text{when } \Gamma_1 = \Gamma \cup \{x, \text{var}(x)\} \text{ and } v_0, (V, E) = \text{bind-expr}_{\Gamma_1,\Delta}(e) \text{ and } v = \Delta(\text{fun}(x), [(v_0, \text{body})])$$
$$\quad \text{let } e = (v, v_0, \text{body})$$

$$\text{bind-expr}_{\Gamma,\Delta}(\lambda x \to e) = v, (\{v\} \cup V, \{e\} \cup E) \tag{4}$$
$$\quad \text{when } \Gamma_1 = \Gamma \cup \{x, \text{var}(x)\} \text{ and } v_0, (V, E) = \text{bind-expr}_{\Gamma_1,\Delta}(e) \text{ and } \Delta(\text{fun}(x), [(v_0, \text{body})])\downarrow$$
$$\quad \text{let } v = \text{fun}(x, s), s \text{ fresh and } e = (v, v_0, \text{body})$$

$$\text{bind-expr}_{\Gamma,\Delta}(o) = \text{val}(o), (\{\text{val}(o)\}, \emptyset) \tag{5}$$

$$\text{bind-expr}_{\Gamma,\Delta}(x) = v, (\{v\}, \emptyset) \text{ when } v = \Gamma(x) \tag{6}$$

Figure 5 Binding rules that define a construction of a dependency graph for an expression.

$$\text{bind-prog}_{\Gamma,\Delta}(\text{let } x = e; c_2; \ldots; c_n) = v_1; \ldots; v_n, (\{v_1\} \cup V \cup V_1, E \cup E_1) \tag{7}$$
$$\quad \text{let } v_1, (V_1, E_1) = \text{bind-expr}_{\Gamma,\Delta}(e_1) \text{ and } \Gamma_1 = \Gamma \cup \{(x, v_1)\}$$
$$\quad \text{and } v_2; \ldots; v_n, (V, E) = \text{bind-prog}_{\Gamma_1,\Delta}(c_2; \ldots; c_n)$$

$$\text{bind-prog}_{\Gamma,\Delta}(e; c_2; \ldots; c_n) = v_1; \ldots; v_n, (\{v_1\} \cup V \cup V_1, E \cup E_1) \tag{8}$$
$$\quad \text{let } v_1, (V_1, E_1) = \text{bind-expr}_{\Gamma,\Delta}(e) \text{ and } v_2; \ldots; v_n, (V, E) = \text{bind-prog}_{\Gamma_1,\Delta}(c_2; \ldots; c_n)$$

$$\text{bind-prog}_{\Gamma,\Delta}([\,]) = [\,], (\emptyset, \emptyset) \tag{9}$$

Figure 6 Binding rules that define a construction of a dependency graph for a program.

When binding a member access, we use bind-expr recursively to get a node and a dependency graph for each sub-expression. The nodes representing sub-expressions are then used for lookup into Δ, together with their labels. If a node already exists in Δ it is reused (1). Alternatively, we create a new node containing a fresh symbol (2). The graph node bound to a function depends on a synthetic node $\text{var}(x)$ that represents a variable of unknown value. When binding a function, we create a variable node and add it to the variable context Γ_1 before binding the body. As with member access, the node representing a function may (3) or may not (4) already exist in the lookup table.

When binding a program, we bind the first command and recursively process remaining commands (9). For let binding (7), we bind the expression e assigned to the variable to obtain a graph node v_1. We then store the node in the variable context Γ_1 and bind the remaining commands. The variable context is used when binding a variable in bind-expr (6) and so all variables declared using let will be bound to a graph node representing the value assigned to the variable. When the command is just an expression (8), we bind the expression using bind-expr.

$$\begin{aligned}
&\mathsf{update}_{V,E}(\Delta_{i-1}) = \Delta_i \text{ such that:}\\
&\quad \Delta_i(\mathsf{mem}(m), [(v_0, \mathsf{arg}(0)), \ldots, (v_n, \mathsf{arg}(n))]) = \mathsf{mem}(m, s)\\
&\qquad \text{when } \mathsf{mem}(m, s) \in V \text{ and } (\mathsf{mem}(m, s), v_i, \mathsf{arg}(i)) \in E \text{ for } i \in 0, .., n\\
&\quad \Delta_i(\mathsf{fun}(x), [(v_1, \mathsf{body})]) = \mathsf{fun}(x, s)\\
&\qquad \text{when } \mathsf{fun}(x, s) \in V \text{ and } (\mathsf{fun}(x, s), v_1, \mathsf{body}) \in E\\
&\quad \Delta_i(v) = \Delta_{i-1}(v) \quad \text{(otherwise)}
\end{aligned}$$

Figure 7 Updating the node cache after binding a new graph

4.3 Edit and rebind loop

During editing, the dependency graph is repeatedly updated according to the binding rules. We maintain a series of lookup table states $\Delta_0, \Delta_1, \Delta_2, \ldots$ The initial lookup table is empty, i.e. $\Delta_0 = \emptyset$. At a step i, we parse a program p_i and obtain a new dependency graph using the previous Δ. The result is a sequence of nodes corresponding to commands of the program and a graph (V, E):

$$v_1; \ldots; v_n, (V, E) = \mathsf{bind\text{-}prog}_{\emptyset, \Delta_{i-1}}(p_i)$$

The new state of the cache is computed using $\mathsf{update}_{V,E}(\Delta_{i-1})$ defined in Figure 7. The function adds newly created nodes from the graph (V, E) to the previous cache Δ_{i-1} and returns a new cache Δ_i.

5 Computing live previews

The binding process constructs a dependency graph after code changes. The nodes in the dependency graph correspond to individual operations that will be performed when running the program. When evaluating a preview, we attach partial results to nodes of the graph. Since the binding process reuses nodes, previews for sub-expressions attached to graph nodes will also be reused.

In this section, we describe how previews are evaluated. The evaluation is done over the dependency graph, rather than over the structure of program expressions as in the operational semantics given in Section 3.2. In Section 6, we prove that resulting previews are the same as the result we would get by directly evaluating code and we also show that no recomputation occurs when code is edited in certain ways.

5.1 Previews and delayed previews

Programs in the data exploration calculus consist of sequence of commands. Those are evaluated to a value with a preview that can be displayed to the user. However, we also support previews for sub-expressions. This can be problematic if the current sub-expression is inside the body of a function. For example:

```
let top = movies.take(10).map(λx → x.getReleased().format("dd-mm-yyyy"))
```

$$\text{(lift-expr)}\ \frac{v \Downarrow [\![e]\!]_\Gamma}{v \Downarrow_{\text{lift}} [\![e]\!]_\Gamma} \qquad \text{(fun-val)}\ \frac{(\text{fun}(x,s), v, \text{body}) \in E \qquad v \Downarrow p}{\text{fun}(x,s) \Downarrow \lambda x \to p}$$

$$\text{(lift-prev)}\ \frac{v \Downarrow p}{v \Downarrow_{\text{lift}} [\![p]\!]_\emptyset} \qquad \text{(fun-bind)}\ \frac{(\text{fun}(x,s), v, \text{body}) \in E \qquad v \Downarrow [\![e]\!]_x}{\text{fun}(x,s) \Downarrow \lambda x \to e}$$

$$\text{(val)}\ \frac{}{\text{val}(o) \Downarrow o} \qquad \text{(fun-expr)}\ \frac{(\text{fun}(x,s), v, \text{body}) \in E \qquad v \Downarrow [\![e]\!]_{x,\Gamma}}{\text{fun}(x,s) \Downarrow [\![\lambda x \to e]\!]_\Gamma}$$

$$\text{(var)}\ \frac{}{\text{var}(x) \Downarrow [\![x]\!]_x}$$

$$\text{(mem-val)}\ \frac{\forall i \in \{0 \ldots k\}.(\text{mem}(m,s), v_i, \text{arg}(i)) \in E \qquad v_i \Downarrow p_i \qquad p_0.m(p_1,\ldots,p_k) \rightsquigarrow_\epsilon p}{\text{mem}(m,s) \Downarrow p}$$

$$\text{(mem-expr)}\ \frac{\forall i \in \{0 \ldots k\}.(\text{mem}(m,s), v_i, \text{arg}(i)) \in E \qquad \exists j \in \{0 \ldots k\}.v_j \not\Downarrow p_j \qquad v_i \Downarrow_{\text{lift}} [\![e_i]\!]_{\Gamma_i}}{\text{mem}(m,s) \Downarrow [\![e_0.m(e_1,\ldots,e_k)]\!]_{\Gamma_0,\ldots,\Gamma k}}$$

Figure 8 Rules that define evaluation of previews over a dependency graph for a program

Here, we can directly evaluate sub-expressions movies and movies.take(10), but not x.getReleased() because it contains a free variable x. Our preview evaluation algorithm addresses this by producing two kinds of previews. A *fully evaluated preview* is just a value, while a *delayed preview* is a partially evaluated expression with free variables:

$$\begin{array}{lcll} p & = & o \mid \lambda x \to e & (\textit{Fully evaluated previews}) \\ d & = & p \mid [\![e]\!]_\Gamma & (\textit{Evaluated and delayed previews}) \end{array}$$

A fully evaluated preview p can be either a primitive object or a function value with no free variables. A possibly delayed preview d can be either an evaluated preview p or an expression e that requires variables Γ. We use an untyped language and so Γ is just a list of variables $x_1, \ldots, x_n$. As discussed in Appendix B, delayed previews have an interesting theoretical link with graded comonads. The body of a lambda function may have a fully evaluated preview if it uses only variables that are bound by earlier let bindings, but it will typically be delayed. We consider a speculative design for an abstraction mechanism that better supports live previews in Appendix D.

5.2 Evaluation of previews

The evaluation of previews is defined in Figure 8. Given a dependency graph (V, E), the relation $v \Downarrow d$ evaluates a sub-expression corresponding to the node v to a possibly delayed preview d. The nodes V and edges E of the graph are parameters of $\Downarrow$, but they do not change during the evaluation and so we do not explicitly write them.

The auxiliary relation $v \Downarrow_{\text{lift}} d$ always evaluates to a delayed preview. If the ordinary evaluation returns a delayed preview, so does the auxiliary relation (lift-expr). If the ordinary evaluation returns a value, the value is wrapped into a delayed preview requiring no variables (lift-prev). A node representing a value is evaluated to a value (val) and a node representing an unbound variable is reduced to a delayed preview that requires the variable and returns its value (var).

For member access, we distinguish two cases. If all arguments evaluate to values (member-val), then we use the evaluation relation defined by external libraries $\rightsquigarrow_\epsilon$ to immediately evaluate the member access and produce a value. If some of the arguments are delayed (member-expr), because the member access is inside the body of a lambda function, we produce a delayed member access expression that requires the union of the variables required by the individual arguments.

The evaluation of function values is similar, but requires three cases. If the body can be reduced to a value with no unbound variables (fun-val), we return a lambda function that returns the value. If the body requires only the bound variable (fun-bind), we return a lambda function with the delayed preview as the body. If the body requires further variables, the result is a delayed preview.

5.3 Caching of evaluated previews

For simplicity, the relation $\Downarrow$ in Figure 8 does not specify how previews are cached. In practice, this is done by maintaining a lookup table from graph nodes v to previews p. Whenever $\Downarrow$ is used to obtain a preview for a graph node, we first check the lookup table. If the preview has not been previously evaluated, we evaluate it and add it to the lookup. Cached previews can be reused in two ways. First, if the same sub-expression appears multiple times in the program, it will share a graph node and the preview will be resued. Second, when binding modified source code, the process reuses graph nodes and so previews are also reused during code editing.

6 Evaluating live coding environment

Computing previews using a dependency graph implements a correct and efficient optimization. In this section we show that this is the case, first theoretically in Section 6.2, and then empirically in Section 6.3. We also describe a case study where we developed an online service for data exploration based on the methods discussed in this paper (Section 6.4).

6.1 Correctness of previews

To show that the previews are correct, we prove two properties. Correctness (Theorem 6) guarantees that, the previews we calculate using a dependency graph are the same as the values we would obtain by evaluating the program directly. Determinacy (Theorem 7) guarantees that previews assigned to a graph node based on an earlier graph are the same as previews that we would obtain afres using an updated graph.

To simplify the proofs, we consider programs without let bindings. Eliminating let bindings does not change the result of evaluation, as shown in Lemma 2, and it also does not change the constructed dependency graph as shown below in Lemma 3.

Lemma 3 (Let elimintion for a dependency graph). *Given programs p_1, p_2 such that $p_1 \rightsquigarrow_{\mathsf{let}} p_2$ and a lookup table Δ_0 then if $v_1;\ldots;v_n,(V,E) = \mathsf{bind\text{-}prog}_{\emptyset,\Delta_0}(p_1)$ and $v'_1;\ldots;v'_n,(V',E') = \mathsf{bind\text{-}prog}_{\emptyset,\Delta_1}(p_2)$ such that $\Delta_1 = \mathsf{update}_{V,E}(\Delta_0)$ then for all i, $v_i = v'_i$ and also $(V,E) = (V',E')$.*

Proof. By analysis of the binding process. See Appendix A.3. □

The Lemma 3 provides a way of removing let bindings from a program, such that the resulting dependency graph remains the same. Here, we bind the original program first, which adds the node for e to Δ. In our implementation, this is not needed because Δ is updated while the graph is being constructed using bind-expr. To keep the formalisation simpler, we separate the process of building the dependency graph and updating Δ and thus Lemma 3 requires an extra binding step.

Now, we can show that, given a let-free expression, the preview obtained using a correctly constructed dependency graph is the same as the one we would obtain by directly evaluating the expression. This requires a simple auxiliary lemma.

Lemma 4 (Lookup inversion). *Given Δ obtained using* update *in Figure 7 then:*

– *If $v = \Delta(\mathsf{fun}(x),[(v_0,l_0)])$ then $v = \mathsf{fun}(x,s)$ for some s.*
– *If $v = \Delta(\mathsf{mem}(m),[(v_0,l_0),\ldots,(v_n,l_n)])$ then $v = \mathsf{mem}(m,s)$ for some s.*

Proof. By construction of Δ in Figure 7. □

Theorem 5 (Term preview correctness). *Given a term t that has no free variables, together with a lookup table Δ obtained from any sequence of programs using* bind-prog *(Figure 6) and* update *(Figure 7), then let $v,(V,E) = \mathsf{bind\text{-}expr}_{\emptyset,\Delta}(t)$.*

If $v \Downarrow p$ over a graph (V,E) then $p = o$ for some value o and $t \rightsquigarrow^* o$.

Proof. By induction over the binding process. See Appendix A.4. □

Theorem 6 (Program preview correctness). *Consider a program $p = c_1;\ldots;c_n$ that has no free variables, together with a lookup table Δ_0 obtained from any sequence of programs using* bind-prog *(Figure 6) and* update *(Figure 7). Assume a let-free program $p' = t_1;\ldots;t_n$ such that $p \rightsquigarrow^*_{\mathsf{let}} p'$.*

Let $v_1;\ldots;v_n,(V,E) = \mathsf{bind\text{-}prog}_{\emptyset,\Delta_0}(p)$ and define updated lookup table $\Delta_1 = \mathsf{update}_{V,E}(\Delta_0)$ and let $v'_1;\ldots;v'_n,(V',E') = \mathsf{bind\text{-}prog}_{\emptyset,\Delta_1}(p)$.

If $v'_i \Downarrow p_i$ over a graph (V',E') then $p_i = o_i$ for some value o_i and $t_i \rightsquigarrow o_i$.

Proof. Direct consequence of Lemma 3 and Theorem 5. □

Our implementation updates Δ during the recursive binding process and so a stronger version of the property holds: previews calculated over a graph obtained directly for the original program p are the same as the values of the fully evaluated program. Our formalisation omits this for simplicity.

The second important property is determinacy, which makes it possible to cache the previews evaluated via $\Downarrow$ using the corresponding graph node as a lookup key.

Edit contexts of expressions

$$\begin{array}{lcl} K_e[-] & = & K_e[-].m(e_1,\ldots,e_n) \;\mid\; e.m(e_1,\ldots,e_{l-1},K_e[-],e_{l+1},\ldots,e_n) \;\mid\; - \\ K_c[-] & = & \mathsf{let}\ x = K_e[-] \;\mid\; K_e[-] \end{array}$$

Code edit operations preserving preview for a sub-expression

(let-intro-var) $\overline{c_1};$ «e»$;\overline{c_2}$ changes to $\overline{c_1};\mathsf{let}\,x = e;$ «x»$;\overline{c_2}$ where x is fresh.

(let-intro-ins) $\overline{c_1};\overline{c_2};$ «$K_c[e]$»$;\overline{c_3}$ is changed to $\overline{c_1};\mathsf{let}\,x = e;\overline{c_2};$ «$K_c[x]$»$;\overline{c_3}$ via a semantically non-equivalent expression $\overline{c_1};\overline{c_2};K_c[x];\overline{c_3}$ where x is free.

(let-intro-del) $\overline{c_1};\overline{c_2};$ «$K_c[e]$»$;\overline{c_3}$ is changed to $\overline{c_1};\mathsf{let}\,x = e;\overline{c_2};$ «$K_c[x]$»$;\overline{c_3}$ via an expression $\overline{c_1};\mathsf{let}\,x = e;\overline{c_2};K_c[e];\overline{c_3}$ with unused variable x.

(let-elim-del) $\overline{c_1};\mathsf{let}\,x = e;\overline{c_2};$ «$K_c[x]$»$;\overline{c_3}$ is changed to $\overline{c_1};\overline{c_2};$ «$K_c[e]$»$;\overline{c_3}$ via a semantically non-equivalent expression $\overline{c_1};\overline{c_2};K_c[x];\overline{c_3}$ where x is free.

(let-elim-ins) $\overline{c_1};\mathsf{let}\,x = e;\overline{c_2};$ «$K_c[x]$»$;\overline{c_3}$ is changed to $\overline{c_1};\overline{c_2};$ «$K_c[e]$»$;\overline{c_3}$ via an expression $\overline{c_1};\mathsf{let}\,x = e;\overline{c_2};K_c[e];\overline{c_3}$ with unused variable x.

(edit-mem) $\overline{c_1};K_c[$ «e_0»$.m(\overline{e})];\overline{c_2}$ is changed to $\overline{c_1};K_c[$ «e_0»$.m'(\overline{e'})];\overline{c_2}$

(edit-let) $\overline{c_1};\mathsf{let}\,x = e_1;\overline{c_2};K_c[$ «e_2» $];\overline{c_3}$ is changed to $\overline{c_1};\mathsf{let}\,x = e_1';\overline{c_2};K_c[$ «e_2» $];\overline{c_3}$ when $x \notin FV(e_2)$.

Figure 9 Code edit operations that preserve previously evaluated preview

Theorem 7 (Preview determinacy). *For some Δ and for any programs p, p', assume that the first program is bound, i.e. $v_1;\ldots;v_n,(V,E) = \mathsf{bind\text{-}prog}_{\emptyset,\Delta}(p)$, the graph node cache is updated $\Delta' = \mathsf{update}_{V,E}(\Delta)$ and the second program is bound, i.e. $v_1';\ldots;v_m',(V',E') = \mathsf{bind\text{-}prog}_{\emptyset,\Delta'}(p')$. Now, for any v, if $v \Downarrow p$ over (V,E) then also $v \Downarrow p$ over (V',E').*

Proof. By induction over $\Downarrow$, show that the same evaluation rules also apply over (V',E'). This is the case, because graph nodes added to Δ' by $\mathsf{update}_{V,E}$ are added as new nodes in $\mathsf{bind\text{-}prog}_{\emptyset,\Delta'}$ and nodes and edges of (V,E) are unaffected. $\square$

The cache of previews (Section 5.3) associates a preview d with a node v as the key. Theorem 7 guarantees that this is valid. As we update dependency graph during code editing, previous nodes will continue representing the same sub-expressions.

6.2 Reuse of previews

In this section, we identify a number of code edit operations where the previously evaluated values for a sub-expression can be reused. This includes the motivating example from Section 1 where the data analyst extracted a constant into a let binding and modified a parameter of the last method call in a call chain.

The list of preview-preserving edits is shown in Figure 9. It includes several ways of introducing and eliminating let bindings and edits where the analyst modifies an unrelated part of the program. The list is not exhaustive. Rather, it illustrates typical edits that the data analyst might perform when writing code. To express the operations we define an editing context K which is similar to evaluation context C from Figure 3, but allows sub-expressions appearing anywhere in the program.

We use the notation «e» to mark parts of expressions that are not recomputed during the edit; we write $\overline{c}$ and $\overline{e}$ for a list of commands and expressions, respectively. In some of the edit operations, we also specify an intermediate program that may be semantically different and only has a partial preview. This illustrates a typical way of working with code in a text editor using cut and paste. For example, in (let-intro-ins), the analyst cuts a sub-expression e, replaces it with a variable x and then adds a let binding for a variable x and inserts the expression e from the clipboard. The (let-intro-del) operation captures the same edit, but done in a different order.

Theorem 8 proves that the operations given in Figure 9 preserve the preview for a marked sub-expression. It relies on a Lemma 13 given in Section A.5 that generally characterizes one common kind of edits. Given two versions of a program that both contain the same sub-expression e, if the let bindings that define the values of variables used in e do not change, then the graph node assigned to e will be the same when binding the original and the updated program.

Theorem 8 (Preview reuse). *Given the sequence of expressions as specified in Figure 9, if the expressions are bound in sequence and graph node cache updated as specified in Figure 7, then the graph nodes assigned to the specified sub-expressions are the same.*

Proof. Cases (edit-let) and (edit-mem) are direct consequences of Lemma 13; for (let-intro-var), the node assigned to x is the node assigned to e which is the same as before the edit from Lemma 13. Cases (let-intro-ins) and (let-intro-del) are similar to (let-intro-var), but also require using induction over the binding of $K_c[e]$. Finally, cases (let-elim-ins) and (let-elim-del) are similar and also use Lemma 13 together with induction over the binding of $K_c[x]$. □

6.3 Empirical evaluation of efficiency

The key performance claim about our method of providing instant feedback is that it is more efficient than recomputing values for the whole program (or the current command) after every keystroke. In the previous section, we formally proved that this is true and gave examples of code edit operations that do not cause recomputation. In this section, we further support this claim with an empirical evaluation. The purpose of this section is not to precisely evaluate overheads of our implementation, but to compare how much recomputation different evaluation strategies perform.

For the purpose of the evaluation, we use a simple image manipulation library that provides operations for loading, greyscaling, blurring and combining images. We compare delays in updating the preview for three different evaluation strategies, while performing the same sequence of code edit operations. Using image processing as an example gives us a way to visualize the reuse of previously computed values. As in a typical data exploration scenario, the individual operations are relatively expensive compared to the overheads of building the dependency graph.

(1) Enter the following code and then change parameter of blur from 4 to 8:

```
image.load("shadow.png").greyScale().blur(4)
```

(2) Assign the result to a variable and start writing code for further operations:

```
let shadow = image.load("shadow.png").greyScale().blur(8)
shadow.combine
```

(3) Finish code to combine two images and then change parameter of combine from 20 to 80:

```
let shadow = image.load("shadow.png").greyScale().blur(8)
shadow.combine(image.load("pope.png"), 20)
```

(4) Extract the parameter of combine to a let bound variable:

```
let ratio = 80
let shadow = image.load("shadow.png").greyScale().blur(8)
shadow.combine(image.load("pope.png"), ratio)
```

Figure 10 Code edit operations that are used in the experimental evaluation

To avoid clutter when visualizing the performance, we update the preview after complete tokens are added rather than after individual keystrokes. Figure 10 shows the sequence of edits that we use to measure the delays in updating a live preview. We first enter an expression to load, greyscale and blur an image (1) then introduce let binding (2) and add more operations (3). Finally, we extract one of the parameters into a variable (4). Most of the operations are simply adding code, but there are two cases where we modify existing code and change value of a parameter for blur and combine immediately after (1) and (3), respectively.

We implement the algorithm described in Sections 4 and 5 in a simple web-based environment that allows the user to modify code and explicitly trigger recomputation. It then measures time needed to recompute values for the whole program and displays the resulting image. If the parsing fails, we record only the time taken by parsing attempt. We compare the delays of three different evaluation strategies:

- *Call-by-value.* Following the semantics in Section 3.2, all sub-expressions are evaluated before an expression. This is often wasteful. For example, we parse the expression image.load("shadow.png").blur as a member access with no arguments. The evaluation loads the image, but then fails because blur requires one argument.
- *Lazy.* To address the wastefulness of call-by-value strategy, we simulate lazy evaluation by implementing a version of the image processing library where operations build a delayed computation and only evaluate it when rendering an image. Using this strategy, failing computations do not perform unnecessary work.
- *Live.* Finally, we use the algorithm described in Sections 4 and 5. The cache is empty at the beginning of the experiment and we update it after each token is added. This is the only strategy where evaluation does not start afresh after reparsing code.

The experimental environment is implemented in F# and compiled to JavaScript using the Fable compiler. We run the experiments in Firefox (version 64.0.2, 32bit) on Windows 10 (build 1809, 64bit) on Intel Core i7-7660U CPU with 16Gb RAM.

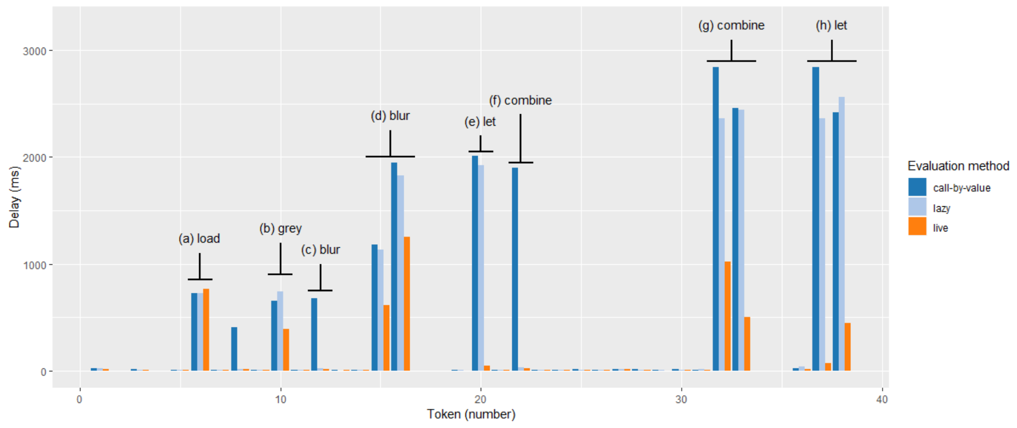


Figure 11 Time required to recompute the results of a sample program after individual tokens are added or modified for three different evaluations strategies.

Figure 11 shows times needed to recompute previews after individual tokens are added, deleted or modified, according to the script in Figure 10, resulting in 38 measurements. We mark a number of notable points in the chart:

a. Loading image for the first time incurs small extra overhead in the live strategy.
b. Greyscaling using the live strategy does not need to re-load the image.
c. Accessing the blur member without arguments causes delay for call-by-value.
d. When varying the parameter of blur, the live strategy reuses the greyscaled image.
e. Introducing let binding does not cause recomputation when using live strategy.
f. As in (c), accessing a member without an argument only affects call-by-value.
g. The live strategy is much faster when varying the parameter of combine.
h. Introducing let binding, again, causes full recomputation for lazy and call-by-value.

A summarized view of the delays is provided in Figure 12, which shows histograms illustrating the distribution of delays for each of the three evaluation methods. A large proportion of delays is very small (less than 15ms) because the parser used in our experimental environment often fails (e.g. for unclosed parentheses). The histogram on the right summarizes only delays for edit operations where the delay for at least one of the strategies was over 15ms. The histogram shows that the live strategy eliminates

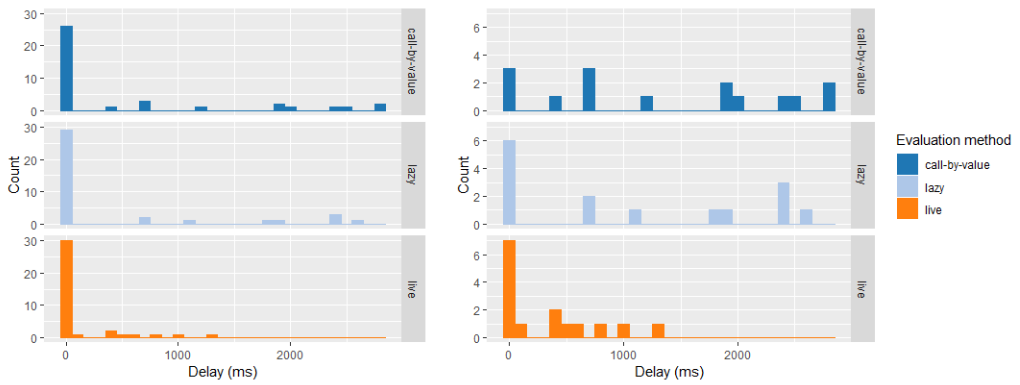
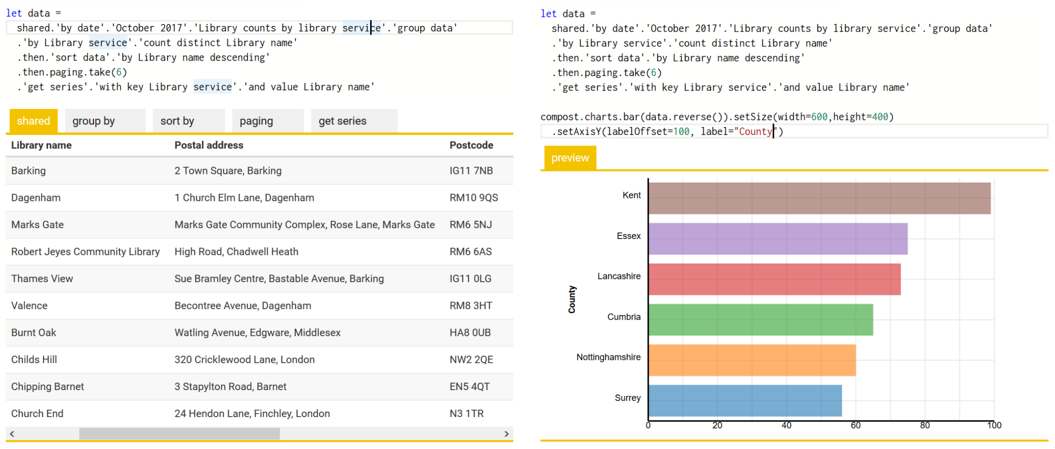


Figure 12 Distribution of delays incurred when updating previews. We show a histogram computed from all delays (left) and only from delays larger than 15ms (right).



```
let data =
  shared.'by date'.'October 2017'.'Library counts by library service'.'group data'
  .'by Library service'.'count distinct Library name'
  .then.'sort data'.'by Library name descending'
  .then.paging.take(6)
  .'get series'.'with key Library service'.'and value Library name'
```

Library name	Postal address	Postcode
Barking	2 Town Square, Barking	IG11 7NB
Dagenham	1 Church Elm Lane, Dagenham	RM10 9QS
Marks Gate	Marks Gate Community Complex, Rose Lane, Marks Gate	RM6 5NJ
Robert Jeyes Community Library	High Road, Chadwell Heath	RM6 6AS
Thames View	Sue Bramley Centre, Bastable Avenue, Barking	IG11 0LG
Valence	Becontree Avenue, Dagenham	RM8 3HT
Burnt Oak	Watling Avenue, Edgware, Middlesex	HA8 0UB
Childs Hill	320 Cricklewood Lane, London	NW2 2QE
Chipping Barnet	3 Stapylton Road, Barnet	EN5 4QT
Church End	24 Hendon Lane, Finchley, London	N3 1TR

```
let data =
  shared.'by date'.'October 2017'.'Library counts by library service'.'group data'
  .'by Library service'.'count distinct Library name'
  .then.'sort data'.'by Library name descending'
  .then.paging.take(6)
  .'get series'.'with key Library service'.'and value Library name'

compost.charts.bar(data.reverse()).setSize(width=600,height=400)
  .setAxisY(labelOffset=100, label="County")
```

Figure 13 Data analysis counting the number of libraries per county in the UK. We load and aggregate data (left) and then create a chart with a label (right). The example is adapted from http://gallery.thegamma.net/73/.

the longest delays (by caching partial results), with the exception of a few where the underlying operation takes a long time (such as blurring the image). The results would be even more significant with an error-recovering parser.

The purpose of our experiment is not to exactly assess the overhead of our implementation. Our goal is to illustrate how often can previously evaluated results be reused and the impact this has when writing code. The experiment presented in this section is small-scale, but it is sufficient for this purpose. When recomputing results after every edit using the *call-by-value* strategy, the time needed to update results grows continually. The *lazy* strategy removes the overhead for programs that fail, but keeps the same trend. Our *live* strategy reuses values computed previously. Consequently, expensive operations such as (d) and (g) in Figure 11 are significantly faster, because they do not need to recompute operations done previously when writing the code. As shown in Figure 12, there are almost no very expensive operations (taking over 1 second) in the *live* strategy in contrast to several in the other two strategies.

6.4 Transparent tools for data journalism

In Section 2.1, we motivated our work by considering how journalists explore open data. In addition to the theoretical and experimental work presented in this paper, we also implemented an online data exploration environment, equipped with live editor for The Gamma language that provides instant feedback during coding. The environment uses the principles presented in this paper to build a more comprehensive system that allows users, such as journalists, to analyse, summarize and visualize open data. In this section, we briefly report on our experience with the system. Two screenshots shown in Figure 13 illustrate a number of interesting features:

- The left screenshot uses a type provider for data aggregation [47]. Type provider are treated as an external library (with objects, members and reduction relation). Type providers also rely on type information to provide editor auto-complete, which we support by implementing type checking over a dependency graph (Section C).

- When displaying live preview for code written using the data aggregation type provider (left), our environment generates tabes that show individual steps of the data transformation. A tab is selected based on the cursor position and shows a preview for the current sub-expression (e.g. raw data before grouping).
- Another external library provides support for charting (right). Here, the environment displays the result of evaluating the whole command. The screenshot shows a case where the user modifies parameters of the chart. Thanks to our live evaluation strategy, this is done efficiently without reevaluating the data transformation.

The environment is available at http://gallery.thegamma.net. A user study to evaluate the usability of the system from a human-computer interaction perspective is left for future work. As an anecdotal evidence, the code in Figure 13 was developed by an attendee of a Mozilla Festival 2017 who had no prior programming experience.

7 Related and future work

We aim to make creating transparent data analyses, such as those done by journalists, easier and we present a tool that provides instant feedback when writing code in a text editor. There is a rich body of work exploring complementary designs.

Notebooks and data science tools. Basic feedback mechanism, the REPL (read-eval-print-loop), originated in Lisp [34] and has become standard [17]. Notebook systems such as Jupyter [5, 29, 45] structure code in cells, but still rely on executing individual cells using a stateful REPL. Adding provenance tracking [30, 48, 50] addresses the reproducibility issues of notebooks, but keeps cells as units of work. Ideas such as dependency tracking and efficient recomputation also exist in visual data exploration tools [11, 25, 61] and scientific workflow systems [6, 41]. The implementation techniques are related to our work, but we focus on text-based scripts. Tempe [13] focuses on streaming data, but provides a text-based scripting environment with automatic code update; its usability in contrast to REPLs has been empirically evaluated [12]. The challenges faced by Tempe are more akin to dynamic software update [26].

Live coding and editing. Live coding based on textual programs has been popularised by Victor [58, 59] and is actively developed in domains such as live coded music [1, 51]. Kubelka et al. [31] review the use of live coding in a Smalltalk derived environment. Lighttable [20] and Chrome DevTools provide limited live previews akin to those presented in this paper, but without well specified recomputation model. Finally, work on keeping state during code edits [8, 35] would be relevant for supporting streaming data. In the future, we also plan to explore how direct manipulation [53] could be integrated with our preview mechanism. Using structured editors [56] rather than text editors is a promising alternative direction as editors can recompute program based on understanding of edit operations [33, 42, 44]. This can be elegantly implemented using bi-directional lambda calculus [43], but it also makes us consider more human-centric abstractions [14, 15] further discussed in Section D.

Incremental computation and dependency analysis. Work on self-adjusting and incremental computation [2, 24] handles recomputation when the program stays the same, but input changes. It relies on dependency tracking similar to ours. We rely on static code analysis whereas most incremental systems, e.g. [3, 4, 23] evaluate the program and use programmer-supplied information to build a dependency graph. [24] implement a small adaptive interpreter that treats code as changing input data, suggesting an interesting implementation technique for live coding systems. Our use of dependency graphs [32] is static and first-order and can be seen as a form of program slicing [60], although our binding process is more directly inspired by Roslyn [40], which uses it for efficient background type-checking.

Semantics and partial evaluation. The evaluation of previews is a form of partial evaluation [10], done in a way that allows caching. This can be done implicitly or explicitly in the form of multi-stage programming [57]. Semantically, the evaluation of previews can be seen as a modality [16] and delayed previews are linked to contextual modal type theory [39], formally modelled using comonads [18]. This suggests a direction for rigorous analysis of the presented system.

8 Summary

One of the aspects that make spreadsheets easier to use than programming tools is that they provide instant feedback. We aim to make programming tools as instant as spreadsheets. We described a number of key aspects of simple data analyses such as those done by journalists and then used our observations to build both theory and simple practical data analytics tools.

Our *data exploration calculus* is a simple formally tractable language for data exploration. The calculus captures key observations about simple data analyses. They rely on logic defined by external libraries, implement few abstractions and are written as lists of commands.

Our main technical contribution is a *live preview* mechanism that efficiently evaluates code during editing and instantly provides a preview of the result. We allow users to edit code in unconstrained way in an text editor, which makes this particularly challenging. The key trick is to separate the process into a fast *binding phase*, which constructs a dependency graph and a slower *evaluation phase* that can cache results. This makes it possible to quickly parse updated code, reconstruct dependency graph and compute preview using previous, partially evaluated, results.

We evaluated our approach in three ways. First, we proved that our mechanism is correct and that it reuses evaluated values for many common code edit operations. Second, we conducted an experimental study that illustrates how often are previously evaluated results reused during typical programming scenario. Thirdly, we used our research as a basis for online data exploration environment, which shows the practical usability of our work.

References

[1] Samuel Aaron and Alan F Blackwell. “From sonic Pi to overtone: creative musical experiences with domain-specific and functional languages”. In: *Proceedings of the first ACM SIGPLAN workshop on Functional art, music, modeling & design*. ACM. 2013, pages 35–46.

[2] Umut A. Acar. “Self-adjusting Computation”. AAI3166271. PhD thesis. Pittsburgh, PA, USA, 2005. ISBN: 0-542-01547-1.

[3] Umut A. Acar, Guy E. Blelloch, and Robert Harper. “Adaptive Functional Programming”. In: *ACM Trans. Program. Lang. Syst.* 28.6 (Nov. 2006), pages 990–1034. ISSN: 0164-0925. DOI: 10.1145/1186632.1186634. URL: http://doi.acm.org/10.1145/1186632.1186634.

[4] Umut A. Acar and Ruy Ley-Wild. “Self-adjusting Computation with Delta ML”. In: *Advanced Functional Programming: 6th International School, AFP 2008, Heijen, The Netherlands, May 2008, Revised Lectures*. Edited by Pieter Koopman, Rinus Plasmeijer, and Doaitse Swierstra. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pages 1–38. ISBN: 978-3-642-04652-0. DOI: 10.1007/978-3-642-04652-0_1. URL: https://doi.org/10.1007/978-3-642-04652-0_1.

[5] J Allaire, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, Rob Hyndman, and R Arslan. “rmarkdown: Dynamic Documents for R”. In: *R package version* 1 (2016), page 9010.

[6] Ilkay Altintas, Chad Berkley, Efrat Jaeger, Matthew Jones, Bertram Ludascher, and Steve Mock. “Kepler: an extensible system for design and execution of scientific workflows”. In: *Scientific and Statistical Database Management*. IEEE. 2004, pages 423–424.

[7] David Blood. *Recycling is broken – notebooks*. Nov. 2018. URL: https://github.com/ft-interactive/recycling-is-broken-notebooks.

[8] Sebastian Burckhardt, Manuel Fahndrich, Peli de Halleux, Sean McDirmid, Michal Moskal, and Nikolai Tillmann. “It’s Alive! Continuous Feedback in UI Programming”. In: *PLDI*. ACM SIGPLAN, June 2013.

[9] David Raymond Christiansen. “Dependent type providers”. In: *Proceedings of the 9th ACM SIGPLAN workshop on Generic programming*. ACM. 2013, pages 25–34.

[10] Charles Consel and Olivier Danvy. “Tutorial notes on partial evaluation”. In: *Proceedings of the 20th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM. 1993, pages 493–501.

[11] Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. “Vizdom: Interactive Analytics Through Pen and Touch”. In: *Proc. VLDB Endow.* 8.12 (Aug. 2015), pages 2024–2027. ISSN: 2150-8097. DOI: 10.14778/2824032.2824127.

[12] R. DeLine and D. Fisher. “Supporting exploratory data analysis with live programming”. In: *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Oct. 2015, pages 111–119. DOI: 10.1109/VLHCC.2015.7357205.

[13] R. DeLine, D. Fisher, B. Chandramouli, J. Goldstein, M. Barnett, J. Terwilliger, and J. Wernsing. “Tempe: Live scripting for live data”. In: *2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. Oct. 2015, pages 137–141. DOI: 10.1109/VLHCC.2015.7357208.

[14] Jonathan Edwards. *Direct Programming*. June 2018. URL: https://vimeo.com/274771188.

[15] Jonathan Edwards. “Subtext: uncovering the simplicity of programming”. In: *ACM SIGPLAN Notices* 40.10 (2005), pages 505–518.

[16] Matt Fairtlough, Michael Mendler, and Eugenio Moggi. “Special issue: Modalities in type theory”. In: *Mathematical Structures in Computer Science* 11.4 (2001), pages 507–509.

[17] Robert Bruce Findler, John Clements, Cormac Flanagan, Matthew Flatt, Shriram Krishnamurthi, Paul Steckler, and Matthias Felleisen. “DrScheme: A programming environment for Scheme”. In: *Journal of functional programming* 12.2 (2002), pages 159–182.

[18] Murdoch J Gabbay and Aleksandar Nanevski. “Denotation of contextual modal type theory (CMTT): Syntax and meta-programming”. In: *Journal of Applied Logic* 11.1 (2013), pages 1–29.

[19] Marco Gaboardi, Shin-ya Katsumata, Dominic A Orchard, Flavien Breuvart, and Tarmo Uustalu. “Combining effects and coeffects via grading.” In: *ICFP*. 2016, pages 476–489.

[20] Chris Granger. “LightTable: A new IDE concept”. In: *http://www.chris-granger.com/2012/04/12/light-table-a-new-ide-concept/* (2012).

[21] Jonathan Gray, Lucy Chambers, and Liliana Bounegru. *The data journalism handbook: how journalists can use data to improve the news*. " O’Reilly Media, Inc.", 2012.

[22] P Guo. “Data science workflow: Overview and challenges”. In: *blog CACM, Communications of the ACM* (2013).

[23] Matthew A. Hammer, Joshua Dunfield, Kyle Headley, Nicholas Labich, Jeffrey S. Foster, Michael Hicks, and David Van Horn. “Incremental Computation with Names”. In: *Proceedings of the 2015 ACM SIGPLAN International Conference on Object-Oriented Programming, Systems, Languages, and Applications*. OOPSLA 2015. Pittsburgh, PA, USA: ACM, 2015, pages 748–766. ISBN: 978-1-4503-3689-5. DOI: 10.1145/2814270.2814305. URL: http://doi.acm.org/10.1145/2814270.2814305.

[24] Matthew A. Hammer, Khoo Yit Phang, Michael Hicks, and Jeffrey S. Foster. "Adapton: Composable, Demand-driven Incremental Computation". In: *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI '14. Edinburgh, United Kingdom: ACM, 2014, pages 156–166. ISBN: 978-1-4503-2784-8. DOI: 10.1145/2594291.2594324. URL: http://doi.acm.org/10.1145/2594291.2594324.

[25] J. M. Hellerstein, R. Avnur, A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth, and P. J. Haas. "Interactive data analysis: the Control project". In: *Computer* 32.8 (Aug. 1999), pages 51–59. ISSN: 0018-9162. DOI: 10.1109/2.781635.

[26] Michael Hicks, Jonathan T. Moore, and Scott Nettles. "Dynamic Software Updating". In: *Proceedings of the ACM SIGPLAN 2001 Conference on Programming Language Design and Implementation*. PLDI '01. Snowbird, Utah, USA: ACM, 2001, pages 13–23. ISBN: 1-58113-414-2. DOI: 10.1145/378795.378798. URL: http://doi.acm.org/10.1145/378795.378798.

[27] Leslie Hook and John Reed. *Why the world's recycling system stopped working*. Oct. 2018. URL: https://www.ft.com/content/360e2524-d71a-11e8-a854-33d6f82e62f8.

[28] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. "Wrangler: Interactive Visual Specification of Data Transformation Scripts". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: ACM, 2011, pages 3363–3372. ISBN: 978-1-4503-0228-9. DOI: 10.1145/1978942.1979444.

[29] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. "Jupyter Notebooks-a publishing format for reproducible computational workflows." In: *ELPUB*. 2016, pages 87–90.

[30] David Koop and Jay Patel. "Dataflow Notebooks: Encoding and Tracking Dependencies of Cells". In: *9th USENIX Workshop on the Theory and Practice of Provenance, TaPP 2017, Seattle, WA, USA, June 23, 2017*. 2017.

[31] Juraj Kubelka, Romain Robbes, and Alexandre Bergel. "The Road to Live Programming: Insights from the Practice". In: *Proceedings of the 40th International Conference on Software Engineering*. ICSE '18. Gothenburg, Sweden: ACM, 2018, pages 1090–1101. ISBN: 978-1-4503-5638-1. DOI: 10.1145/3180155.3180200. URL: http://doi.acm.org/10.1145/3180155.3180200.

[32] David J Kuck, Robert H Kuhn, David A Padua, Bruce Leasure, and Michael Wolfe. "Dependence graphs and compiler optimizations". In: *Proceedings of the 8th ACM SIGPLAN-SIGACT symposium on Principles of programming languages*. ACM. 1981, pages 207–218.

[33] Eyal Lotem and Yair Chuchem. "Lamdu Project". In: *https://github.com/lamdu/lamdu* (2018).

[34] John McCarthy. “History of LISP”. In: *SIGPLAN Not.* 13.8 (Aug. 1978), pages 217–223. ISSN: 0362-1340. DOI: 10.1145/960118.808387. URL: http://doi.acm.org/10.1145/960118.808387.

[35] Sean McDirmid. “Living it up with a live programming language”. In: *ACM SIGPLAN Notices*. Volume 42. 10. ACM. 2007, pages 623–638.

[36] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. O’Reilly Media, Inc., 2012.

[37] Robin Milner. “A theory of type polymorphism in programming”. In: *Journal of Computer and System Sciences* 17.3 (1978), pages 348–375. ISSN: 0022-0000. DOI: 10.1016/0022-0000(78)90014-4.

[38] Alan Mycroft, Dominic Orchard, and Tomas Petricek. “Effect systems revisited—control-flow algebra and semantics”. In: *Semantics, Logics, and Calculi*. Springer, 2016, pages 1–32.

[39] Aleksandar Nanevski, Frank Pfenning, and Brigitte Pientka. “Contextual modal type theory”. In: *ACM Transactions on Computational Logic (TOCL)* 9.3 (2008), page 23.

[40] Karen Ng, Matt Warren, Peter Golde, and Anders Hejlsberg. “The Roslyn Project, Exposing the C# and VB compiler’s code analysis”. In: *White paper, Microsoft* (2011).

[41] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R Pocock, Anil Wipat, et al. “Taverna: a tool for the composition and enactment of bioinformatics workflows”. In: *Bioinformatics* 20.17 (2004), pages 3045–3054.

[42] Cyrus Omar, Ian Voysey, Ravi Chugh, and Matthew A. Hammer. “Live Functional Programming with Typed Holes”. In: *PACMPL* 3.POPL (2019). DOI: 10.1145/3290327.

[43] Cyrus Omar, Ian Voysey, Michael Hilton, Jonathan Aldrich, and Matthew A Hammer. “Hazelnut: a bidirectionally typed structure editor calculus”. In: *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*. ACM. 2017, pages 86–99.

[44] Roland Perera. “Interactive functional programming”. PhD thesis. University of Birmingham, 2013.

[45] Fernando Pérez and Brian E Granger. “IPython: a system for interactive scientific computing”. In: *Computing in Science & Engineering* 9.3 (2007).

[46] Tomas Petricek. “Context-aware programming languages”. PhD thesis. University of Cambridge, Computer Laboratory, 2017.

[47] Tomas Petricek. “Data Exploration through Dot-driven Development”. In: *LIPIcs-Leibniz International Proceedings in Informatics*. Volume 74. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2017.

[48] Tomas Petricek, James Geddes, and Charles A. Sutton. “Wrattler: Reproducible, live and polyglot notebooks”. In: *10th USENIX Workshop on the Theory and Practice of Provenance, TaPP 2018, London, UK, July 11-12, 2018.* 2018.

[49] Tomas Petricek, Gustavo Guerra, and Don Syme. “Types from data: Making structured data first-class citizens in F”. In: *ACM SIGPLAN Notices* 51.6 (2016), pages 477–490.

[50] João Felipe Nicolaci Pimentel, Vanessa Braganholo, Leonardo Murta, and Juliana Freire. “Collecting and analyzing provenance on interactive notebooks: when IPython meets noWorkflow”. In: *Workshop on the Theory and Practice of Provenance (TaPP)*. 2015, pages 155–167.

[51] Charles Roberts, Matthew Wright, and JoAnn Kuchera-Morin. “Beyond editing: extended interaction with textual code fragment”. In: *NIME*. 2015, pages 126–131.

[52] Peter Sestoft. *Spreadsheet technology*. Technical report. Citeseer, 2012.

[53] Ben Shneiderman. “Direct manipulation: A step beyond programming languages”. In: *ACM SIGSOC Bulletin*. Volume 13. 2-3. ACM. 1981, page 143.

[54] Don Syme. “Leveraging. net meta-programming components from F#: integrated queries and interoperable heterogeneous execution”. In: *Proceedings of the 2006 workshop on ML*. ACM. 2006, pages 43–54.

[55] Donald Syme, Keith Battocchi, Kenji Takeda, Donna Malayeri, and Tomas Petricek. “Themes in information-rich functional programming for internet-scale data sources”. In: *Proceedings of the 2013 workshop on Data driven functional programming*. ACM. 2013, pages 1–4.

[56] Gerd Szwillus and Lisa Neal. *Structure-based editors and environments*. Academic Press, Inc., 1996.

[57] Walid Taha and Tim Sheard. “MetaML and multi-stage programming with explicit annotations”. In: *Theoretical computer science* 248.1 (2000), pages 211–242.

[58] Bret Victor. “Inventing on Principle”. In: *http://worrydream.com/InventingOnPrinciple* (2012).

[59] Bret Victor. “Learnable programming: Designing a programming system for understanding programs”. In: *http://worrydream.com/LearnableProgramming* (2012).

[60] Mark Weiser. “Program slicing”. In: *Proceedings of the 5th international conference on Software engineering*. IEEE Press. 1981, pages 439–449.

[61] Richard Wesley, Matthew Eldridge, and Pawel T. Terlecki. “An Analytic Data Engine for Visualization in Tableau”. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’11. Athens, Greece: ACM, 2011, pages 1185–1194. ISBN: 978-1-4503-0661-4. DOI: 10.1145/1989323.1989449.

A Details of proofs

A.1 Normalization for data exploration calculus

Theorem 9 (Normalization). *For all p, there exists n and $o_1, \ldots, o_n$ such that $p \rightsquigarrow^* o_1; \ldots; o_n$.*

Proof. We define size of a program in data exploration calculus as follows:

$$\begin{array}{rcl} \mathsf{size}(c_1; \ldots; c_n) &=& 1 + \Sigma_{i=1}^{n} \mathsf{size}(c_i) \\ \mathsf{size}(\mathsf{let}\ x = t) &=& 1 + \mathsf{size}(t) \\ \mathsf{size}(e_0.m(e_1, \ldots, e_n)) &=& 1 + \Sigma_{i=0}^{n} \mathsf{size}(e_i) \\ \mathsf{size}(\lambda x \rightarrow e) &=& 1 + \mathsf{size}(e) \\ \mathsf{size}(o) = \mathsf{size}(x) &=& 1 \end{array} \tag{1}$$

The property holds because, first, both (let) and (external) decrease the size of the program and, second, a program is either fully evaluated, i.e. $o_1; \ldots; o_n$ for some n or, it can be reduced using one of the reduction rules. □

A.2 Let elimmination for a program

Lemma 10 (Let elimination for a program). *Given any program p such that $p \rightsquigarrow^* o_1; \ldots; o_n$ for some n and $o_1, \ldots, o_n$ then if $p \rightsquigarrow_{\mathsf{let}} p'$ for some p' then also $p' \rightsquigarrow^* o_1; \ldots; o_n$.*

Proof. The elimination of let binding transforms a program $c_1; \ldots; c_{k-1}; \mathsf{let}\ x = t; c_{k+1}; \ldots; c_n$ to a program $c_1; \ldots; c_{k-1}; t; c_{k+1}[x \leftarrow t]; \ldots; c_n[x \leftarrow t]$. The reduction steps for the new program can be constructed using the steps of $p \rightsquigarrow^* o_1; \ldots; o_n$. The new command t reduces to an object o using the same steps as the original term t in $\mathsf{let}\ x = t$ but with context $C_c = -$ rather than $C_c = \mathsf{let}\ x = -$; the terms t introduced by substitution also reduce using the same steps as before, but using contexts in which the variable x originally appeared. □

A.3 Let elimination for a dependency graph

Lemma 11 (Let elimintion for a dependency graph). *Given programs p_1, p_2 such that $p_1 \rightsquigarrow_{\mathsf{let}} p_2$ and a lookup table Δ_0 then if $v_1; \ldots; v_n, (V, E) = \mathsf{bind\text{-}prog}_{\emptyset, \Delta_0}(p_1)$ and $v'_1; \ldots; v'_n, (V', E') = \mathsf{bind\text{-}prog}_{\emptyset, \Delta_1}(p_2)$ such that $\Delta_1 = \mathsf{update}_{V,E}(\Delta_0)$ then for all i, $v_i = v'_i$ and also $(V, E) = (V', E')$.*

Proof. Assume $p_1 = c_1; \ldots; c_{k-1}; \mathsf{let}\ x = e; c_{k+1}; \ldots; c_n$ and the let binding is eliminated resulting in $p_2 = c_1; \ldots; c_{k-1}; e; c_{k+1}[x \leftarrow e]; \ldots; c_n[x \leftarrow e]$. When binding p_1, the case $\mathsf{bind\text{-}prog}_{\Gamma, \Delta}(\mathsf{let}\ x = e)$ is handled using (7) and the node resulting from binding e is added to the graph V, E. It is then referenced each time x appears in subsequent commands $c_{k+1}; \ldots; c_n$. When binding p_2, the node resulting from binding e is a primitive value or a node already present in Δ_1 (added by $\mathsf{update}_{V,E}$) and is reused each time $\mathsf{bind\text{-}expr}_{\Gamma, \Delta_1}(e)$ is called. □

A.4 Term preview correctness

Theorem 12 (Term preview correctness). *Given a term t that has no free variables, together with a lookup table Δ obtained from any sequence of programs using* bind-prog *(Figure 6) and* update *(Figure 7), then let $v,(V,E) = \text{bind-expr}_{\emptyset,\Delta}(t)$. If $v \Downarrow p$ over a graph (V,E) then $p = o$ for some value o and $t \rightsquigarrow^* o$.*

Proof. When combining recursively constructed sub-graphs, the bind-expr function adds new nodes and edges leading from those new nodes. Therefore, an evaluation using $\Downarrow$ over a sub-graph will also be valid over the new graph – the newly added nodes and edges do not introduce non-determinism to the rules given in Figure 8.

We prove a more general property showing that for any e, its binding $v,(V,E) = \text{bind-expr}_{\emptyset,\Delta}(e)$ and any evaluation context C such that $C[e] \rightsquigarrow o$ for some o, one of the following holds:

a. If $FV(e) = \emptyset$ then $v \Downarrow p$ for some p and $C[p] \rightsquigarrow o$
b. If $FV(e) \neq \emptyset$ then $v \Downarrow [\![e_p]\!]_{FV(e)}$ for some e_p and $C[e_p] \rightsquigarrow o$

In the first case, p is a value, but it is not always the case that $e \rightsquigarrow^* p$, because p may be lambda function and preview evaluation may reduce sub-expression in the body of the function. Using a context C in which the value reduces to an object avoids this problem.

The proof of the theorem follows from the more general property. Using a context $C[-] = -$, the term t reduces $t \rightsquigarrow^* t' \rightsquigarrow_\epsilon o$ for some o and the preview p is a value o because $C[p] = p = o$. The proof is by induction over the binding process, which follows the structure of the expression:

(1) $\text{bind-expr}_{\Gamma,\Delta}(e_0.m(e_1,\ldots,e_n))$ – Here $e = e_0.m(e_1,\ldots,e_n)$, v_i are graph nodes obtained by induction for expressions e_i and $\{(v,v_0,\text{arg}(0)),\ldots,(v,v_n,\text{arg}(n))\} \subseteq E$. From lookup inversion Lemma 4, $v = \text{mem}(m,s)$ for some s.

If $FV(e) = \emptyset$, then $v_i \Downarrow p_i$ for $i \in 0\ldots n$ and $v \Downarrow p$ using (mem-val) such that $p_0.m(p_1,\ldots,p_n) \rightsquigarrow p$. From induction hypothesis and *compositionality* of external libraries (Definition 2), it holds that for any C such that $C[e_0.m(e_1,\ldots,e_n)] \rightsquigarrow o$ for some o then also $C[p_0.m(p_1,\ldots,p_n)] \rightsquigarrow C[p] \rightsquigarrow o$.

If $FV(e) \neq \emptyset$, then $v_i \Downarrow_{\text{lift}} [\![e'_i]\!]$ for $i \in 0\ldots n$ and $v \Downarrow [\![e'_0.m(e'_1,\ldots,e'_n)]\!]_{FV(e)}$ using (mem-expr). From induction hypothesis and *compositionality* of external libraries (Definition 2), it holds that for any C such that $C[e_0.m(e_1,\ldots,e_n)] \rightsquigarrow o$ for some o then also $C[e'_0.m(e'_1,\ldots,e'_n)] \rightsquigarrow o$.

(2) $\text{bind-expr}_{\Gamma,\Delta}(e_0.m(e_1,\ldots,e_n))$ – This case is similar to (1), except that the fact that $v = \text{mem}(m,s)$ holds by construction, rather than using Lemma 4.

(3) $\text{bind-expr}_{\Gamma,\Delta}(\lambda x \to e_b)$ – Here $e = \lambda x \to e_b$, v_b is the graph node obtained by induction for the expression e_b and $(v,v_b,\text{body}) \in E$. From lookup inversion Lemma 4, $v = \text{fun}(x,s)$ for some s. The evaluation can use one of three rules:

i. If $FV(e) = \emptyset$ then $v_b \Downarrow p_b$ for some p_b and $v \Downarrow \lambda x \to p_b$ using (fun-val). Let $e'_b = p_b$.
ii. If $FV(e_b) = \{x\}$ then $v_b \Downarrow [\![e'_b]\!]_x$ for some e'_b and $v \Downarrow \lambda x \to e'_b$ using (fun-bind).
iii. Otherwise, $v_b \Downarrow [\![e'_b]\!]_{x,\Gamma}$ for some e'_b and $v \Downarrow [\![\lambda x \to e'_b]\!]_\Gamma$ using (fun-expr).

For i.) and ii.) we show that a.) is the case; for iii.) we show that b.) is the case; that is for any C, if $C[\lambda x \to e_b] \rightsquigarrow o$ then also $C[\lambda x \to e'_b] \rightsquigarrow o$. For a given C, let $C'[-] = C[\lambda x \to -]$ and use the induction hypothesis, i.e. if $C'[e_b] \rightsquigarrow o$ for some o then also $C'[e'_b] \rightsquigarrow o$.

(4) $\mathsf{bind\text{-}expr}_{\Gamma,\Delta}(\lambda x \to e)$ – This case is similar to (3), except that the fact that $v = \mathsf{fun}(x,s)$ holds by construction, rather than using Lemma 4.

(5) $\mathsf{bind\text{-}expr}_{\Gamma,\Delta}(o)$ – In this case $e = o$ and $v = \mathsf{val}(o)$ and $\mathsf{val}(o) \Downarrow o$ using (val) and so the case a.) trivially holds.

(6) $\mathsf{bind\text{-}expr}_{\Gamma,\Delta}(x)$ – The initial Γ is empty, so x must have been added to Γ by case (3) or (4). Hence, $v = \mathsf{var}(x)$, $v \Downarrow [\![x]\!]_x$ using (var) and so $e_p = e = x$ and the case b.) trivially holds.

□

A.5 Binding sub-expressions

Lemma 13 (Binding sub-expressions)**.** *Assume we have programs $p_1 = c_1; \ldots; c_k; K_c[e]; c_{k+1}; \ldots; c_n$ and $p_2 = c'_1; \ldots; c'_k; K'_c[e]; c'_{k+1}; \ldots; c'_n$ and $I \subseteq \{1 \ldots k\}$ such that $\forall i \in I. c_i = c'_i$ and for each $x \in \bigcup_{i \in I} FV(c_i) \cup FV(e)$ there exists $j \in I$ such that $c_j = \mathsf{let}\, x = e$ for some e. Given any Δ, assume that the the first program is bound, i.e. $v_1; \ldots; v_n, (V,E) = \mathsf{bind\text{-}prog}_{\emptyset,\Delta}(p_1)$, the cache is updated $\Delta' = \mathsf{update}_{V,E}(\Delta)$ and the second program is bound, i.e. $v'_1; \ldots; v'_n, (V',E') = \mathsf{bind\text{-}prog}_{\emptyset,\Delta'}(p_2)$.*

Now, assume $v, G = \mathsf{bind\text{-}expr}_{\Gamma,\Delta}(e)$ and $v', G' = \mathsf{bind\text{-}expr}_{\Gamma',\Delta'}(e)$ are the recursive calls to bind e during the first and the second binding, respectively. Then, the graph nodes assigned to the sub-expression e are the same, i.e. $v = v'$.

Proof. First, assuming that $\forall x \in FV(e). \Gamma(x) = \Gamma'(x)$, we show by induction over the binding process of e for the first program that the result is the same. In cases (1) and (3), the updated Δ' contains the required key and so the second binding proceeds using the same case. In cases (2) and (4), the second binding reuses the node created by the first binding using case (1) and (3), respectively. Cases (5) and (6) are the same.

Second, when binding let bindings in $c_1; \ldots; c_k$, the initial $\Gamma = \emptyset$ during both bindings. Nodes added to Γ and Γ' for commands c_j such that $j \in I$ are the same and nodes added for remaining commands do not add any new nodes referenced from e and so $v = v'$ using the above. □

B Theories of delayed previews

The operational semantics presented in this paper serves two purposes. It gives a simple guide for implementing text-based live coding environments for data science and we use it to prove that our optimized way of producing live previews is correct. However, some aspects of our mechanism are related to important work in semantics of programming languages and deserve to be mentioned.

$$\begin{tikzcd}
\mathsf{mem}(m,s_0) \arrow[r, "\mathsf{arg}(0)"] \arrow[dr, bend right, "\mathsf{arg}(1)"'] & \mathsf{val}(o) & \mathsf{var}(x,s_1) \arrow[l, "\mathsf{callsite}(m,1)"'] \\
 & \mathsf{fun}(x,s_2) \arrow[u, "\mathsf{callsite}(m,1)"'] \arrow[ur, bend right, "\mathsf{body}"'] &
\end{tikzcd}$$

Figure 14 Dependency graph for $o.m(\lambda x \rightarrow x)$ with a newly added callsite edges.

The construction of delayed previews is related to meta-programming. Assuming we have delayed previews $[\![e_0]\!]_x$ and $[\![e_1]\!]_y$ and we invoke a member m on e_0 using e_1 as an argument. To do this, we construct a new delayed preview $[\![e_0.m(e_1)]\!]_{x,y}$. This operation is akin to expression splicing from meta-programming [54, 57].

The semantics of delayed previews can be more formally captured by Contextual Modal Type Theory (CMTT) [39] and comonads [18]. In CMTT, $[\Psi]A$ denotes that a proposition A is valid in context Ψ, which is similar to our delayed previews written as $[\![A]\!]_\Psi$. CMTT defines rules for composing context-dependent propositions that would allow us to express the splicing operation used in (mem-expr). In categorical terms, the context-dependent proposition can be modelled as a graded comonad [19, 38]. The evaluation of a preview with no context dependencies (built implicitly into our evaluation rules) corresponds to the counit operation of a comonad and would be explicitly written as $[\![A]\!]_\emptyset \rightarrow A$.

C Type checking

Live previews give analysts quick feedback when they write incorrect code, but having type information is still valuable. First, it can help give better error messages. Second, types can be used to provide auto-complete – when the user types '.' we can offer available members without having to wait until the value of the object is available.

Revised dependency graph. Type checking of small programs is typically fast enough that no caching is necessary. However, The Gamma supports *type providers* [9, 55], which can generate types based on an external file or a REST service call, e.g. [49]. For this reason, type checking can be relatively time consuming and can benefit from the same caching facilities as those available for live previews.

Adding type checking requires revising the way we construct the dependency graph introduced in Section 4. Previously, a variable bound by a lambda function had no dependencies. However, the type of the variable depends on the context in which it appears. Given an expression $o.m(\lambda x \rightarrow x)$, we infer the type of x from the type of the first argument of the member m. A variable node for x thus needs to depend on the call site of m. We capture that by adding an edge callsite(m,i) from x to o which indicates that x is the input variable of a function passes as the i^{th} argument to the m member of the expression represented by the target node. We also add callsite(m,i) as an edge from the node of the function. Figure 14 shows the revised dependency graph for $o.m(\lambda x \rightarrow x)$.

$$\text{(var)}\ \frac{(\mathsf{var}(x,s), v, \mathsf{callsite}(m,i)) \in E \qquad v \vdash \{..,m:(\tau_1,\ldots,\tau_k)\to\tau,..\} \qquad \tau_i = \tau' \to \tau''}{\mathsf{var}(x,s) \vdash \tau'}$$

$$\text{(mem)}\ \frac{\forall i \in \{0\ldots k\}.(\mathsf{mem}(m,s), v_i, \mathsf{arg}(i)) \in E \qquad v_0 \vdash \{..,m:(\tau_1,\ldots,\tau_k)\to\tau,..\} \qquad v_i \vdash \tau_i}{\mathsf{mem}(m,s) \vdash \tau}$$

$$\text{(fun)}\ \frac{\{(\mathsf{fun}(x,s), v_b, \mathsf{body}), (\mathsf{var}(x,s), v_c, \mathsf{callsite}(m,i))\} \subseteq E \qquad v_c \vdash \{..,m:(\tau_1,\ldots,\tau_k)\to\tau,..\} \qquad \tau_i = \tau' \to \tau'' \qquad v_b \vdash \tau''}{\mathsf{fun}(x,s) \vdash \tau' \to \tau''}$$

Figure 15 Rules that define type checking of terms and expressions over a dependency graph (V, E)

Type checking. The structure of typing rules is similar to the evaluation relation $v \Downarrow d$ defined earlier. Given a dependency graph (V, E), we define typing judgements in the form $v \vdash \tau$. The type τ can be a primitive type, a function $\tau \to \tau$ or an object type $\{m_1:\sigma_1,\ldots,m_n:\sigma_n\}$ with member types $\sigma = (\tau_1,\ldots,\tau_n) \to \tau$.

The typing rules for variables, functions and member access are shown in Figure 15. When type checking a member access (mem), we find its dependencies v_i and check that the instance is an object with the required member m. The types of arguments of the member then need to match the types of the remaining (non-instance) nodes. Type checking a function (fun) and a variable (var) is similar. In both cases, we follow the callsite edge to find the member that accepts the function as an argument and use the type of the argument to check the type of the function or infer the type of the variable.

The results of type checking can be cached and reused in the same way as live previews, although we leave out the details. A property akin to correctness (Theorem 6) requires defining standard type checking over the structure of expressions, which we also omit for space reasons.

D Feedback-friendly abstraction

The data analysis by Financial Times in Section 2.1 illustrates why notebook users often avoid abstraction. Wrapping code in a function makes it impossible to split code into cells and see results of intermediate steps. Instead, the analysis used a global variable with possible values in a comment.

Providing live previews inside ordinary functions is problematic, because we do not have readily available values for input parameters and our mechanism for lambda functions only provides delayed previews inside body of a function. We believe that extending the data exploration calculus with an abstraction mechanism that would support development with instant feedback is an interesting design problem and we briefly outline a possible solution here.

Data scientists often write code interactively using a sample data set and, when it works well, wrap it into a function that they then call on other data. Similarly, spreadsheet users often write equation in the first row of a table and then use the "drag down" operation to apply it to other rows. One way of adding similar functionality to the data exploration calculus is to label a sequence of commands such that the sequence can be reused later with different inputs:

$$\begin{array}{lcl} p & = & c_1; \ldots; c_n \\ c & = & \mathsf{let}\, x = t \;\mid\; t \;\mid\; lbl\colon p \;\mid\; lbl \end{array}$$

We introduce two new kinds of commands: a labelled sequence of commands and a reference to a label. When evaluating, the command *lbl* is replaced with the associated sequence of commands p before any other reductions. Consequently, variables used in the labelled block are dynamically scoped and we can use let binding to redefine a value of a variable before invoking the block repeatedly. The correct use of dynamic scoping can be checked using coeffects [46].

This minimalistic abstraction mechanism supports code reuse without affecting how live previews are computed. Commands in a labelled block require variables to be defined before the block. Those define sample data for development and can be redefined before reusing the block. We intend to implement this mechanism in a future version of our data exploration environment (Section 6.4).

E Manuscript version history (to be removed for publication)

This note briefly reviews the revisions made to this manuscript in response to reviews for an earlier version of the paper.

Unclear motivation. An earlier review pointed out that it was not clear who is the target user for the described system. The newly added Section 2.1 provides the answer. Our focus is on non-expert programmers conducting simple data exploration such as journalists working with government data. This is perhaps an unusual focus for a programming language research, but we believe it is increasingly important.

Unclear scope of the work. An earlier version of the paper did not make it clear what is prior work and what is new in this paper. The paper has been restructured to make this clear – we introduce the data exploration calculus (Section 3) and a theory for building a live coding environment for the calculus (Section 4) with live preview support (Section 5). This is accompanied by an implementation that extends an editor for The Gamma, described earlier [47] – we add live preview support and auto-complete using type checking based on the dependency graph (Section C).

Simplicity of the language. An earlier review argued that the language used in the paper is overly simplistic. The newly added Section 2.1 explains the motivation behind the data exploration calculus. The simplicity of the language reflects the use cases that motivated this paper. The newly added case study (Section 6.4) shows that the language is sufficient for building a simple, but realistic and practically useful data exploration environment. We also discuss extensions to the langauge in Section D.

Missing empirical evaluation. An earlier version of the paper lacked empirical evaluation. It is not our aim to measure the exact overhead of our implementation (which is typically negligible with respect to the cost of evaluating the data transformations), but the newly added empirical evaluation (Section 6.3) shows the effect of our evaluation strategy during a reprsentative coding task.

Missing related work. An earlier version of the paper lacked discussion of important related work, in particular that on incremental computation. This has now been addressed – incremental computation focuses on the case where data change, while our focus is on cases where code changes, but the implementation techniques are closely related.

About the author

Tomas Petricek Tomas is a lecturer at University of Kent. He is working on making programming and data science easier and more accessible. His research interests span a range of areas including theory of programming languages, history and philosophy of programming, as well as open data and data journalism. Previously, he contributed to functional programming and the development of the F# language and type providers at Microsoft Research and obtained PhD from University of Cambridge for his work on coeffects, a theory of context-aware programming languages. Along the way, he became interested in understanding programming through the perspective of history and philosophy of science and wrote papers about the evolution of programming concepts such as types and errors. Contact him at tomas@tomasp.net.