Abstraction in Technical Computing

by

Jeffrey Werner Bezanson

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

Author .....
Department of Electrical Engineering and Computer Science
January 9, 2015

Certified by .....
Alan Edelman
Professor
Thesis Supervisor

Accepted by .....
Leslie Kolodziejski
Chairman, Department Committee on Graduate Students

Abstraction in Technical Computing

by

Jeffrey Werner Bezanson

Submitted to the Department of Electrical Engineering and Computer Science
on January 9, 2015, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Array-based programming environments are popular for scientific and technical computing. These systems consist of built-in function libraries paired with high-level languages for interaction. Although the libraries perform well, it is widely believed that scripting in these languages is necessarily slow, and that only heroic feats of engineering can at best partially ameliorate this problem.

In this thesis I argue that what is really needed is a more coherent structure for this functionality. To find one, we must ask what technical computing is really about. I suggest that this kind of programming is characterized by an emphasis on operator complexity and code specialization, and that a language can be designed to better fit these requirements.

The key idea is to integrate code *selection* with code *specialization*, using generic functions and data-flow type inference. Systems like these can suffer from inefficient compilation, or from uncertainty about what exactly to specialize on. I show that dispatch on structured type tags helps address these problems. The resulting language, Julia, achieves a Quine-style "explication by elimination" of many of the productive features technical computing users expect.

Thesis Supervisor: Alan Edelman
Title: Professor

Acknowledgments

thank Tim Holy for frequently and spontaneously writing explanations that I found myself wanting to copy and paste into this writeup.

Contents

Chapter 1

Introduction

Much of the history of programming languages has been about increasing abstraction and generalization. Today we take it for granted that a single suitably powerful language could be used for nearly any programming task, were it not for pragmatic factors beyond our control. Over time, many special-purpose languages have been subsumed by these more powerful and general languages. The field of technical computing — programming for applied math and the sciences — is something of a holdout from this trend, with its own purpose-built languages still dominating. There are probably many reasons for this, including historical, social and technical.

In this work we will focus on the technical reasons for this, and the problems that the current state of affairs causes.

What do we mean by a *language* for scientific computing? The prevailing answer is that it is about numeric arrays, perhaps matrices specifically, and numerical libraries that use them. Instead we intend to argue that, at a deeper level, it is about certain kinds of flexibility, particularly flexibility in the behavior of key operators and functions. This flexibility is needed to maximize composability and reusability in scientific code. Without it, certain programs may be easy to write today, but changing functional and performance requirements can become difficult to meet.

It is clear that scientific and technical users have their own particular needs. So far there

Mainstream PL	Technical computing
classes, single dispatch	complex operators
separate compilation	performance, inlining
parametric polymorphism	ad-hoc polymorphism, extensibility
static checking	experimental computing
modularity, encapsulation	large vocabularies
eliminating tags	self-describing data, acceptance of tags
data hiding	explicit memory layout

Figure 1-1: Priorities of mainstream object-oriented and functional programming language research and implementation compared to those of the technical computing domain.

have been two paths available for fulfilling these needs: using a mainstream general-purpose language like C++ or Java, or designing a new language like R or MATLAB. Designing your own language seems like a drastic thing to do, and yet it is unusually common in technical computing (see gang of 40).

1.1 The technical computing problem

Figure 1.1 compares the general design priorities of mainstream programming languages to those of technical computing languages. The priorities in each row are not necessarily opposites or even mutually exclusive, but rather are a matter of emphasis. For example, it is certainly possible to have both parametric and ad-hoc polymorphism within the same language, but syntax, recommended idioms, and the design of the standard library will tend to emphasize one or the other.

It is striking how different these priorities are. We believe these technical factors have contributed significantly to the persistence of specialized environments in this area.

1.2 Proposed solution

It is clear that any future scientific computing language will need to be able to match the performance of C, C++, and Fortran. To do that, it is almost certain that speculative optimizations such as tracing (TODO cite) will not be sufficient — the language will need to be able to *prove* facts about types, or at least let the user request specific types. It is also clear that such a language must strive for maximum convenience, or else the split between performance languages and productivity languages will persist.

It is fair to say that two approaches to this problem are being tried: one is to design better statically-typed languages, and the other is to apply program analysis techniques to dynamically-typed languages. Static type systems seem to be getting close to achieving the desired level of flexibility (as in Fortress [1], for instance), but it is still too early to call a winner between these two approaches (if, indeed, there even needs to be a winner).

Our contribution derives from observations of the second approach. Efforts to analyze and optimize dynamically-typed programs generally make two assumptions: (1) we should work on analyzing *existing* popular languages, and (2) users of these languages don't want to use types. The first assumption makes practical sense. Convenience is hard to quantify, so using existing languages that have already been deemed convenient by popular opinion puts us on solid footing. This thesis describes a new language, so we simply take the negation of the first assumption as a premise. Our work addresses the second assumption more directly. Specifically, we point out that types do not have to be used for static checking, and that using them for *code selection* and *code specialization* is particularly useful in technical computing. This perspective has not been explored thoroughly in the past.

what to dispatch on? dispatch power has been extended in many ways, but there is no real limit to what somebody might want to dispatch on. so what to do? some sets of values are more robust under computation than others (closure properties). identify those sets using dataflow concerns.

say we have a method defined for integers, and also for the special cases "2" and "odd integers". a realistic implementation will group all of these under "integer", and ideally

generate a couple branches to handle the other cases. we argue the concept of "integer" here is a more robust set, and so a more fundamental language concept.

somewhat counter-intuitively, dynamic dispatch can be good for performance since it permits invoking the most specialized possible method. static overloading can lead to calling a sub-optimal case when multiple overloads exist for the sake of performance.

(one way to describe it: you can start with the notation you want, or with the notation that people are using already, and come up with better ways to *explain* it. in other words, design a robust framework that assigns a satisfactory meaning to each thing we'd like to write. the amazing thing about programming languages it that a better explanation can directly lead to better performance!)

but what type system should we use?

it is quite possible that some static type system will work well for this

however we defer this question.

- mention dual traps of wasted power and divergence

- vs predicate dispatch: we extend dispatch power in a different direction, guided by semantic subtyping. goal is maximum power that still yields high likelihood of resolving many calls to a single implementation.

a multimethod system designed for type inference and specialization.

past (static) type systems for dispatch were designed to ensure the absence of no-method errors and ambiguities (completeness and uniqueness). our goal is instead to statically resolve methods. this is inherently heuristic and best-effort. since static types shouldn't affect program behavior, we conclude that the dispatch must be dynamic, which is happily the same conclusion you would reach if you simply wanted dynamic typing.

interesting variants: - require static single method matches - reject programs with no-method errors - reject programs that yield Unions

Solution: integrate code specialization and code selection

Why integrate code specialization and code selection?

- specialization requires selection anyway

- simpler system
- can sometimes collapse 2 layers of dispatch into 1
- can replace library code with a code generator without either changing client code *or* any extra overhead

Extensive code specialization is a key feature of technical computing. "what to specialize on" has been an open problem. our types are a possible solution to this for two reasons:

1. you can tune the amount of information they contain
2. everybody agrees to use them, which helps ensure that specializing on the types will actually do something useful.

The key ingredients:

1. Self-describing data model aware of memory layout
2. Type tags with nested structure
3. A fully-connected type tree
4. Dynamic multiple dispatch over all types
5. Dataflow type inference
6. Automatic code specialization

This list of features may appear somewhat ad-hoc. However, they turn out to be remarkably strongly coupled, and deeply constrained by our ultimate goal. Each of these features has appeared in some form before, but never in a way that fully solves the problems described here.

Challenges of this approach (why has this not been done before?)

1.3 Contributions

1 - an analysis of the nature of technical computing that suggests what sort of language would form a good base for it. it should emphasize complex operators and code generation/specialization.

2 - the idea of integrating specialization and selection, using multiple dispatch and semantic subtyping.

3 - an explication through elimination of technical computing language features

"One of the most fruitful techniques of language analysis is explication through elimination. The basic idea is that one explains a linguistic feature by showing how one could do without it." [7]

A main contribution of this thesis is the application of this approach to features of technical computing environments that have not been subject to such analysis before.

Chapter 2

Problems in technical computing

At a high level, the experiences and thoughts that give rise to our design decisions.

How should a computer scientist approach this space? We might try to maximize performance. Or it is also easy to make unfounded assumptions about "what users want". But instead we should study the real world, see what is happening and figure out how to steer it in a better direction.

Hypothesis: people don't know what they want. It's also hard to predict what people will want in the future. We need, in the words of Gerry Sussman, systems adaptable to uses not imagined by their designers.

2.1 What is a technical computing environment?

This question has not really been answered.

Views of this are strongly shaped by what systems happen to exist, and what people were exposed to as they learned to program.

Technical computing software has been designed haphazardly. Each system is a pile of features taken without argument.

2.2 The software stack is too complex

Collapsing abstraction layers

- for numerical debuggability in particular - debuggability in general - no time spent worrying about binding time - you will build a dynamic dispatch layer anyway, so build it in

How much of the past 30 years of handwritten Matlab internals can be autogenerated with a compiler? (A lot)

Can now get past traditional classifications of language boundaries - interpreted vs. compiled, high-level vs. low-level, procedural vs. imperative vs. functional, etc.

language performance psychology: if your language doesn't directly support efficient machine data types, users will rewrite their code in C in order to get them, and then be happy with the result (though not with the process).

so julia is an all-levels language

2.3 Code generation

- tensor contraction engine - Firedrake, pyop2 (vs. c++ libmesh) - JuMP vs. python puLP and pyomo - want to be able to pass functions, not C++ code as strings - FFTW - erfinv and horner

2.3.1 A compiler for every problem

2.4 Why dynamic typing?

Mathematical abstractions often frustrate our attempts to represent them on a computer. For example, mathematicians can move instinctively between isomorphic objects such as scalars and 1x1 matrices, but most programming languages would prefer to represent scalars and matrices quite differently. A system might be able to convert automatically from one to the other, but it cannot always know which one we *meant*.

2.4.1 Operational reasoning

people tend to think about programs operationally, i.e. what it *does* when it runs. for example writing if false code end the code does not "occur" and therefore does not need to be valid

there is less to learn. with static languages you have to learn what happens at both compile time and run time, when only run time really matters.

there is a desire to parameterize as much as possible. functions accept parameters, so function calling ought to be sufficient to express any desired parameterization.

2.4.2 I/O

inevitably there is a need to refer to a datatype at run time. the best example is file I/O, where you might want to say "read 500 double-precision numbers from file X". in static languages the syntax and identifiers used to specify such run-time types must be different from those used to specify static types. in C you see defined constants such as `DATATYPE_DOUBLE`.

2.4.3 Flat metadata hierarchy

static types are approximations of dynamic types, so languages with static types inevitably assign two types to a location (both a static type and a dynamic type) where one would do. in some languages, like C++, the desire for performance or ease of implementation leads the compiler to make some decisions based on static types. this is confusing. if type declarations can be omitted, as in a type-inferred language, the situation is even worse since the static type of a value might not be apparent.

2.4.4 Mismatches with mathematical abstractions

programs, in general, deal with values of widely varying disjoint types: functions, numbers, lists, network sockets, etc. type systems are good at sorting out values of these different

types. however, in mathematical code most values are numbers. numerical properties (such as positive, negative, even, odd, greater than 1, etc.) are what matter, and these are highly dynamic. the lattices involved are generally not of finite height.

different number representations exist only for efficiency, and are often incidental to the meaning of a piece of code. for example people want "y = sqrt(x)" to compute the square root of x whether it is positive or negative, and give a real or complex result accordingly. as another example, in many cases it is convenient not to need to worry about integer overflow.

multiple common features underlie mathematical objects of different types (e.g. numbers, sets, matrices). in some cases it makes sense to consider numbers and matrices as the same kind of thing, and in other cases it doesn't matter. A given type system is likely not to have anticipated the particular common features that matter to your program, making it more difficult to express an idea. A concrete example is the matlab fragment if condition idx = ':' else idx = 1 end where we want to select either an entire dimension or the first position alone. The ':' and 1 are both indexes in this context, though they would be of disjoint types in most programming languages.

2.5 Bad tradeoffs in current designs

- Ducking the issue of typing altogether - why this isn't possible IRL

Case in point: Python and NumPy.

NumPy tried to work within the confines of Python, but has more or less failed for technical and political reasons. (Technical: hard to work with types in language that deliberately tries to obscure them from the user. Political: unwillingness to fix broken package distribution system.) Consequences: Numba and Numba_lang, Anaconda. Essentially reinventing types in "Python".

This phenomenon is increasingly noticed in other domains, particularly JavaScript and web programming. Modern JavaScript implementations are quite fast, but Google's Dart language is based on the premise that we could have a web language that is even faster, and

offers more productivity as well. How so? Because Dart's designers observed that JavaScript programmers in practice often write code that could be defined using traditional OO classes, but the language does not support them.

dictionaries for everything (python, js) is the wrong default. almost every type somebody wants has a fixed number of fields with fixed types.

Python is often described as a good glue language. This means it is effectively used as an interface standard, a kind of extended C ABI that makes it easier for libraries to interoperate, and easier for users to access those libraries. Something as straightforward as providing a standard N-dimensional numeric array class (NumPy), which does not exist at the level of C, goes a long way.

However, we wish to point out that in this picture, Python is not doing as much work as it might first appear. Python does not make it easier to implement the functionality inside NumPy, or other "under the hood" scientific libraries. In many cases it creates more work, through the need to write wrappers and interfaces for native code.

Merely being "dynamic" (e.g. Python) should not be considered the gold standard of flexibility. Although these systems permit tricks that can solve otherwise difficult programming problems, this is not always the kind of flexibility that is needed. When faced with the need to describe many functions with elaborate behavior and many cases, one does not primarily need permissiveness, but rather powerful and descriptive organizing principles.

How does Julia break through the 4x-slower-than-C performance barrier?

bits types and parameters – can have abstract types whose size is known unconditionally

2.5.1 What needs to be built in?

On "built-in", why built-in is bad. Built-in-ness often conflates two aspects:

1. A feature being readily available and agreed-on by all language users

2. A feature tightly coupled to the rest of the system

(2) implies (1), but not the other way around. (2) is the only technically interesting item, since the other can be addressed e.g. just by including a library in the standard software distribution. Many technical computing languages have done a large amount of (2) while justifying it with point (1).

While a large part of our motivation is to move more decisions and functionality into libraries, it is equally important to identify what *must* be part of a language for the system to be successful. We believe that large amounts of functionality can be provided by add-ons, but that certain key features absolutely cannot be. Past failures to properly classify features this way have caused a lot of undue pain.

First, performance cannot be an add-on. If some users have a fast version of a language and others have a slow version (with the difference being an order of magnitude or more), library writers cannot be sure whether users will find their code fast enough to be useful. How are we to teach people to program in the language? Courses on MATLAB programming emphasize vectorized code, but if not all implementations had this performance characteristic the curriculum would become confused.

Psychologically, it may be difficult to accept a "non standard" extension that changes a language so fundamentally. There is a nagging, though perhaps totally unfounded, perception that something subtle may break. If indeed a bug arises due to the use of such an extension, a user is likely to conclude that the extension is dangerous or broken and stop using it. If, on the other hand, a bug arises due to a language's standard optimizing compiler, the user will simply file a bug report, then find a way to work around the problem.

Adding a JIT compiler to a language also requires acceptance of detriments like compilation pauses and pages with RWX permissions. In some cases this may lead to use of the extension being disallowed, perhaps for security reasons.

Type systems similarly fail when provided as optional extensions. Library writers face the same kinds of problems as with performance add-ons. Should I use type annotations in my library?

Dynamic dispatch mechanisms also make especially poor add-ons. Of course, every program makes decisions at run time, and so implements its own "dispatch" to some extent. But these behaviors are inextensible; if language users do not agree on a reusable dispatch framework their code will not be composable.

2.6 Social phenomena

Programming languages are observed to have strong network effects, and the difficulty of getting new languages adopted is well known. However based on [socio PLT paper] we believe this doesn't have to be the case. The formula of improving or redesigning general-purposes languages to be more appealing to domain experts might solve the problem. That way the new system has immediate appeal for at least some users, without the worry that a different tool will be needed as soon as requirements change slightly.

barrier to contributing

Software business is based on imposing restrictions

Debates about what abstractions mean – AbstractMatrix we thought we knew what it meant, but what about something like SymTridiagonal? It can implement most of what a dense matrix has, but it can't obey every invariant.

PackedQR

Chapter 3

Available technology

3.1 The language design space

It is helpful to begin with a rather coarse classification of programming languages, according to how expressive their type systems are, and whether their type systems are dynamic or static:

	More types	Fewer types
Dynamic	Dylan, Julia	Scheme, Python, MATLAB
Static	ML, Haskell, Scala	C

The lower right corner tend to contain older languages. The upper right corner contains many popular "dynamic" languages. The lower left corner contains many modern languages resulting from research on static type systems. We are most interested in the upper left corner, which is notable for being rather sparsely populated. It has been generally believed that dynamic languages do not "need" types, or that there is no point in talking about types if they are not going to be checked at compile time. These views have some merit, but as a result the top-left corner of this design space has been seriously under-explored.

3.1.1 Classes vs. functions

3.1.2 Separate compilation vs. performance

3.1.3 Parametric vs. ad-hoc polymorphism

The idea of specialization unites parametric and ad-hoc polymorphism. Beginning with a parametrically-polymorphic function, one can imagine a compiler specializing it for various cases, i.e. certain concrete argument values. These specializations would be stored in a lookup table, for use either at compile time or at run time.

Next we can imagine that the compiler might not optimally specialize certain definitions, and that the programmer would like to provide hints or hand-written implementations to speed up special cases. For example, imagine a function that traverses a generic array. A compiler-generated specialization might inline a specific array types's indexing operations, but a human might further realize that the loop order should be switched for certain arrays types based on their storage order.

However, if we are going to let a human specialize a function for performance, we might as well allow them to specialize it for some other reason, including entirely different behavior. But at this point separate ad-hoc polymorphism is no longer necessary; we are using a single overloading feature for everything.

Parametric polymorphism describes code that works for any object precisely because it does not do anything meaningful to the object, for example the identity function. In contrast, programming with tagged data (e.g. symbolic expression systems, XML) permits code to work for any object because every object has the same structure, allowing meaningful operations.

3.1.4 Static checking vs. flexibility

3.1.5 Modularity vs. large vocabularies

In the context of software engineering, modularity is a primary concern. To build large systems, separate components must be isolated to some degree. Reasons for this include simple concerns like avoiding name conflicts, and a desire to separate interface from implementation to allow a component to change without affecting the rest of the system.

Modularity is sometimes taken to an extreme, and one will see fully qualified names like `org.jboss.annotation.ejb.cache.simple.CacheConfig` (selected from the JBOSS Java APIs).

Technical computing languages have often avoided and discouraged such designs. For example MATLAB for most of its history supported only a single namespace, which comes pre-populated with thousands of functions.

3.1.6 Eliminating vs. embracing tagged data

3.1.7 Data hiding vs. explicit memory layout

Examples of CSC and CSR sparse representation. In one sense this is a perfect example of an interface with multiple implementations, and therefore a good use case for object-oriented programming. However in the technical computing world, *hiding* this difference in representation is not usually considered desirable. Clearly a sparse matrix class cannot contain all functions of matrices that users might want to compute. Yet when new functions are added, the programmer needs and wants to exploit representation details (CSC or CSR) for performance. The performance differences involved here are quite significant (TODO cite).

The loss of encapsulation due to multi-methods weighed in [2] is less of a problem for technical computing, and in some cases even advantageous.

3.1.8 Dynamic dispatch is key

It would be unpleasant if every piece of every program we wrote were forced to do only one specific task. Every time we wanted to do something slightly different, we'd have to write a different program. But if a language allows the same program element to do different things at different times, we can write whole classes of programs at once. This kind of capability is one of the main reasons *object-oriented* programming is popular: it provides a way to automatically select different behaviors according to some structured criteria.

In class-based OO there is essentially *no way* to create an operation that dispatches on existing types (the expression problem). This clearly does not match technical computing, where most programs deal with the same few types (e.g. number, array), and might sensibly want to write new operations that dispatch on them.

We use the non-standard term "criteria" deliberately, in order to clarify our point of view, which is independent of any particular object system.

The sophistication of the available "selection criteria" account for a large part of the perceived "power" or leverage provided by a language. In fact it is possible to illustrate a hierarchy of such mechanisms. As an example, consider a simple simulation, and how it can be written under a series of increasingly powerful paradigms. First, written-out imperative code:

```
while running
    for a in animals
        b = nearby_animal(a)
        if a isa Rabbit
            if b isa Wolf then run(a)
            if b isa Rabbit then mate(a,b)
        else if a isa Wolf
            if b isa Rabbit then eat(a,b)
            if b isa Wolf then follow(a,b)
        end
```

```
    end
end
```

We can see how this would get tedious as we add more kinds of animals and more behaviors. Another problem is that the animal behavior is implemented directly inside the control loop, so it is hard to see what parts are simulation control logic and what parts are animal behavior. Adding a simple object system leads to a nicer implementation [1]:

```
class Rabbit
    method encounter(b)
        if b isa Wolf then run()
        if b isa Rabbit then mate(b)
    end
end

class Wolf
    method encounter(b)
        if b isa Rabbit then eat(b)
        if b isa Wolf then follow(b)
    end
end

while running
    for a in animals
        b = nearby_animal(a)
        a.encounter(b)
    end
end
```

[1] A perennial problem with simple examples is that better implementations often make the code longer.

Here all of the simulation's animal behavior has been essentially compressed into a single program point: `a.encounter(b)` leads to all of the behavior by selecting an implementation based on the first argument, `a`. This kind of criterion is essentially indexed lookup; we can imagine that `a` is simply an integer index into a table of operations.

The next enhancement to "selection criteria" adds a hierarchy of behaviors, to provide further opportunities to avoid repetition:

```
abstract class Animal
    method nearby()
        # search within some radius
    end
end

class Rabbit <: Animal
    method encounter(b: Animal)
        if b isa Wolf then run()
        if b isa Rabbit then mate(b)
    end
end

class Wolf <: Animal
    method encounter(b: Animal)
        if b isa Rabbit then eat(b)
        if b isa Wolf then follow(b)
    end
end

while running
    for a in animals
```

```
        b = a.nearby()
        a.encounter(b)
    end
end
```

We are still essentially doing table lookup, but the tables have more structure: every `Animal` has the `nearby` method, and can inherit a general-purpose implementation.

This brings us roughly to the level of most popular object-oriented languages. But in this example still more can be done. Notice that in the first step to objects we replaced one level of `if` statements with method lookup. However, inside of these methods a structured set of `if` statements remains. We can replace these by adding another level of dispatch.

```
class Rabbit <: Animal
    method encounter(b: Wolf) = run()
    method encounter(b: Rabbit) = mate(b)
end

class Wolf <: Animal
    method encounter(b: Rabbit) = eat(b)
    method encounter(b: Wolf) = follow(b)
end
```

We now have a *double dispatch* system, where a method call uses two lookups, first on the first argument and then on the second argument.

This syntax might be considered a bit nicer, but the design clearly begs a question: why is $n = 2$ special? It isn't, and we could clearly consider even more method arguments as part of dispatch. But at that point, why is the first argument special? Why separate methods in a special way based on the first argument? It seems arbitrary, and indeed we can remove the special treatment:

```
abstract class Animal
end

class Rabbit <: Animal
end

class Wolf <: Animal
end

nearby(a: Animal) = # search
encounter(a: Rabbit, b: Wolf) = run(a)
encounter(a: Rabbit, b: Rabbit) = mate(a,b)
encounter(a: Wolf, b: Rabbit) = eat(a, b)
encounter(a: Wolf, b: Wolf) = follow(a, b)

while running
    for a in animals
        b = nearby(a)
        encounter(a, b)
    end
end
```

Here we made two major changes: the methods have been moved "outside" of any classes, and all arguments are listed explicitly. This change has fairly significant implications. Since methods no longer need to be "inside" classes, there is no syntactic limit to where definitions may appear. Now it is easier to add new methods after a class has been defined. Methods also now naturally operate on combinations of objects, not single objects.

The shift to thinking about combinations of objects is fairly revolutionary. Many interesting properties only apply to combinations of objects, and not individuals. We are also

now free to think of more exotic kinds of combinations.

We can define a method on *any number* of objects:

```
encounter(ws: Wolf...) = pack(ws)

encounter{T<:Animal}(a: T, b: T) = mate(a, b)
```

3.2 Multiple dispatch

CHART of different multiple dispatches. classify them

Why "just overloading" is not enough. You intercept every operation and rewrite it, kind of an escape hatch for a language you don't like. If somebody else also tries to do this, there won't necessarily be any coherence.

A helpful way to classify languages with some kind of generic programming support is to look at which language constructs are generic. For example, in C++ the syntax `object.method(x)` is generic: a programmer can get it to do different things by supplying different values for `object`. The C++ syntax `f(a,b)` or `a+b` is usually not generic, but can be overloaded by supplying definitions for different argument types. However, this overloading only uses compile-time types (no run-time information), and so is essentially a form of renaming — it can be implemented by renaming each definition with a unique name, and replacing overloaded calls with an appropriate name based on compile-time types. This is a much weaker form of generic programming, since different behaviors cannot be obtained by passing different values at run time.

For this reason, C++ programs are not fully generic: it is difficult to substitute new behaviors for every part of a program. Some languages take generic programming much further. For example, in MATLAB, *every* function is effectively generic, and can be overloaded by new classes. This enables a programmer to "intercept" every function call (and therefore, essentially everything a program does)

foo

3.2.1 Example: lattices

This example will illustrate possible benefits of multiple dispatch and dynamic typing for mathematical computing. The benefits are not absolute, but notational and semantic: they involve code size and clarity, and the extent to which the entities provided by the language match a mental model of the domain.

A *lattice* is an algebraic structure where some pairs of elements satisfy a reflexive, anti-symmetric, and transitive relation $\leq$. For purposes of this example, we will consider lattices that have a greatest, or *top*, element ($\top$), and a least, or *bottom* element ($\bot$). When working with lattices one often wants to compute a least upper bound, or *join* ($\sqcup$), or a greatest lower bound, or *meet* ($\sqcap$).

Several interesting concerns arise when modeling lattices in a programming language. First, the structure is very general, and so admits implementations for many different kinds of elements. We want to write code using the operators $\leq$, $\sqcup$, and $\sqcap$, and have it apply to any kind of lattice. Therefore some kind of overloading or object-oriented programming is desirable. Second, some properties apply to all lattices, and we would like to avoid implementing them repeatedly.

Using "duck typing", the problem of modeling an abstraction like lattices disappears almost entirely. One may simply define methods for $\leq$, $\sqcup$, and $\sqcap$ at any time, for any type, and that type will function as a lattice. That is certainly convenient, but it also fails to provide any reusable functionality for those defining lattices.

Figure 3.2.1 shows a small Julia library for lattices. It defines an abstract class `LatticeElement` that may be subclassed by objects that will be used primarily as elements of some lattice. The library also provides standard `LatticeElement` provides some useful default method definitions.

3.2.2 Symbolic programming and predicate dispatch

Systems based on symbolic rewrite rules arguably occupy a further tier of dispatch sophistication. In these systems, you can dispatch on essentially anything, including arbitrary values

```
abstract LatticeElement

<=(x::LatticeElement, y::LatticeElement) = x===y
==(x::LatticeElement, y::LatticeElement) = x<=y && y<=x
< (x::LatticeElement, y::LatticeElement) = x<=y && !(y<=x)

immutable TopElement <: LatticeElement; end
immutable BotElement <: LatticeElement; end

const ⊤ = TopElement()
const ⊥ = BotElement()

<=(::BotElement, ::TopElement) = true
<=(::BotElement, ::LatticeElement) = true
<=(::LatticeElement, ::TopElement) = true

⊔(x::LatticeElement, y::LatticeElement) =  # join
    (x <= y ? y : y <= x ? x : ⊤)

⊓(x::LatticeElement, y::LatticeElement) =  # meet
    (x <= y ? x : y <= x ? y : ⊥)
```

and structures. These systems are typically powerful enough to concisely define the kinds of behaviors we are interested in.

However, symbolic programming lacks data abstraction: the concrete representations of values are exposed to the dispatch system (e.g. there is no difference between being a list and being something represented as a list).

Patterns are very powerful, but the tradeoff is that there is not necessarily a useful relationship between what your program does and what a static analysis (based on a finite-height partial order over patterns) can discover. Maybe julia could be considered a sweet spot somewhere in between.

3.3 Domain theory

In the 1960s Dana Scott asked how to assign meanings to programs, which otherwise just appear to be lists of symbols. For example, given a program computing the factorial function, we want a process by which we can assign the meaning "factorial" to it.

This is about modeling the behavior of a program without running it.

The idea of analyzing computer programs began in earnest in the 1960s with Dana Scott's invention of domain theory. A "domain" in this theory is a partial order of sets of values that a program might manipulate. Domain theory models computation as follows: a program starts with no information, the lowest point in the partial order ("bottom"). Computation steps accumulate information, gradually moving higher through the order. The advantage of this model, in essence, is that it provides a way to think about the meaning of a program without running the program. Even the "result" of a non-terminating program has a representation — the bottom element. Other elements of the partial order might refer to intermediate results.

Domain theory gave rise to the study of denotational semantics and the design of type systems. However, the original theory is quite general and invites us to invent any domains we wish for any sort of language. For example, given a program that outputs an integer, we

might decide that we only care whether this integer is even or odd. Then our posets are the even and odd integers, and we will classify operations in the program according to whether they are evenness-preserving and so on. It should be clear that this sort of analysis, while clearly related to type systems, is fairly different from what most programmers think of as type checking.

3.4 Dynamic type inference

This is exciting because the type system can do useful computational work for you, instead of only checking the work done by the rest of the program.

Joins of quantified types

"Diagonal dispatch"

(T,T) (Int, Real)

\(Int,Int)/

Avoid hard coding logic into the compiler.

Exposing to the user what is part of the language anyway and not just hiding it in the compiler.

Data flow analysis has to actually work

Taking things out of the language that break dataflow analysis

Local reasoning

You are imposing a type system anyway, even implicitly perhaps, so might as well make that from the get-go.

3.4.1 Objections to dynamic typing

it may be that the "power" of a language is equal to the complexity of the criteria used by the language's run-time dispatch mechanisms.

pre-OO: pointer indirection OO: single dispatch, class hierarchies

Haskell: without typeclasses, a beautiful language, but one that nobody would use.

I claim that compile-time abstractions do not count. The problem is that at some point you have to *actually run the program*. Specifying the behavior of a program is hard; this is where we need help from the language.

Scripting languages are often defended using the observation that most code is not performance-critical. However, this fact does not mean that it's ok for a language design to ignore performance, nor does it mean that performance should be the default priority of every line of code. Rather it means that the default should be convenience and safety, with a path to performance easily in reach for when it's needed.

Chapter 4

The Julia approach

4.1 Language description

4.1.1 Type system

Our goal is to design a type system useful for describing method applicability, and (similarly) for describing classes of values for which to specialize code. Set-theoretic types are a natural basis for such a system. A set-theoretic type is a symbolic expression that denotes a set of values. In our case, these correspond to the sets of values methods are intended to apply to, or the sets of values supported by compiler-generated method specializations. Since set theory is widely understood, the use of such types tends to be intuitive.

These types are less coupled to the languages they are used with, since one may design a value domain and set relations within it without yet considering how types relate to program terms (TODO cite Castagna). Since our goals only include performance and expressiveness, we simply skip the later steps for now, and do not address how to type-check terms (or, indeed, the question of whether checking is even possible).

To avoid the dual traps of "wasted power" and divergence, the system we use must have a decidable subtype relation, and must be closed under data-flow operations (meet, join, and widen). It must also lend itself to a reasonable definition of specificity, so that methods can be

ordered automatically (a necessary property for extensibility). These requirements are fairly strict, but still admit many possible designs. The one we present here is aimed at providing the minimum level of sophistication needed to yield a language that feels "powerful" to most modern programmers. Beginning with the simplest possible system, we added features as needed either to satsify the aforementioned closure properties, or to allow us to write method definitions that seemed particularly useful (as it turns out, these two considerations lead to essentially the same features). The presentation that follows will partially reproduce the order of this design process.

We will define our types by formally describing their denotations as sets. We use the notation $[\![T]\!]$ for the set denotation of type expression T. Concrete language syntax and terminal symbols of the type expression grammar are written in typewriter font, and meta-symbols are written in mathematical italic. First there is a universal type `Any`, an empty type `Bottom`, and a partial order $\leq$:

$$[\![\texttt{Any}]\!] = \mathcal{D}$$
$$[\![\texttt{Bottom}]\!] = \emptyset$$
$$T \leq S \Leftrightarrow [\![T]\!] \subseteq [\![S]\!]$$

where $\mathcal{D}$ represents the domain of all values.

Next we add data objects with structured tags. The tag of a value is accessed with `typeof(x)`. Each tag consists of a declared type name and some number of sub-expressions, written as $\texttt{Name}\{E_1, \cdots, E_n\}$. The center dots ($\cdots$) are meta-syntactic and represent a sequence of expressions. Tag types may have declared supertypes (written as `super(T)`). Any type used as a supertype must be declared as abstract, meaning it cannot have direct instances.

$$[\![\texttt{Name}\{\cdots\}]\!] = \{x \mid \texttt{typeof}(x) = \texttt{Name}\{\cdots\}\}$$

$$[\![\texttt{Abstract}\{\cdots\}]\!] = \bigcup_{\texttt{super}(T)=\texttt{Abstract}\{\cdots\}} [\![T]\!]$$

These types closely resemble the classes of an object-oriented language with generic (parametric) types, invariant type parameters, and no concrete inheritance. We prefer parametric *invariance* for reasons that have been addressed in the literature [4]. Invariance preserves the property that the only subtypes of a concrete type are `Bottom` and itself. We also find that most uses of covariance are more flexibly handled by union type connectives, which will be introduced below.

Next we add conventional product (tuple) types, which are used to represent the arguments to methods. These are almost identical to the nominal types defined above, but are different in two ways: they are *covariant* in their parameters, and permit a special form ending in three dots (`...`) that denotes any number of trailing elements:

$$[\![\texttt{Tuple}\{P_1, \cdots, P_n\}]\!] = \prod_{1 \le i \le n} [\![P_i]\!]$$

$$[\![\texttt{Tuple}\{\cdots, P_n\ldots\}]\!], n \ge 1 = \bigcup_{i \ge n-1} [\![\texttt{Tuple}\{\cdots, P_n^i\}]\!]$$

P_n^i represents i repetitions of the final element P_n of the type expression.

The abstract tuple types ending in `...` correspond to variadic methods, which provide convenient interfaces for tasks like concatenating any number of arrays. Multiple dispatch has been formulated as dispatch on tuple types before [6]. This formulation has the advantage that *any* type that is a subtype of a tuple type can be used to express the signature of a method. It also makes the system simpler, since subtype queries can be used to ask questions about methods.

The types introduced so far would be perfectly sufficient for many programs, and are

roughly equal in power to several multiple dispatch systems that have been designed before. However, these types are not closed under data-flow operations. For example, when the two branches of a conditional expression yield different types, a program analysis must compute the union of those types to derive the type of the conditional. The above types are not closed under set union. We therefore add the following type connective:

$$[\![\texttt{Union}\{A, B\}]\!] = [\![A]\!] \cup [\![B]\!]$$

As if by coincidence, `Union` types are also tremendously useful for expressing method dispatch. For example, if a certain method applies to all 32-bit integers regardless of whether they are signed or unsigned, it can be specialized for `Union{Int32,UInt32}`.

`Union` types are easy to understand, but complicate the type system considerably. To see this, notice that they provide an unlimited number of ways to rewrite any type. For example a type `T` can always be rewritten as `Union{T,Bottom}`, or `Union{Bottom,Union{T,Bottom}}`, etc. Any code that processes types must "understand" these equivalences. `Union` types also commute with covariant type constructors (tuples in our case), providing even more ways to rewrite types:

$$\texttt{Tuple\{Union\{A,B\},C\}} = \texttt{Union\{Tuple\{A,C\},Tuple\{B,C\}\}}$$

This is one of a few reasons that union types are often considered undesirable. When used with type inference, such types can grow without bound, possibly leading to slow or even non-terminating compilation. Their occurrence also typically corresponds to cases that would fail most static type checkers. Yet from the perspectives of both data-flow analysis and method specialization, they are perfectly natural and even essential [5] [8] (TODO cite analyses that have used union types).

The next problem we need to solve arises from combining data-flow analysis and parametric invariance. When a type constructor `C` is applied to a type S that is known only approximately at compile time, the type `C`$\{S\}$ does not correctly represent the result if `C`

is invariant. The correct result would be the union of all types $\texttt{C}\{T\}$ where $T \le S$. Interestingly, there is again a corresponding need for such types in method dispatch. Often one has, for example, a method that applies to arrays of any kind of integer (`Array{Int32}`, `Array{Int64}`, etc.). These cases can be expressed using a `UnionAll` connective, which denotes an iterated union of a type expression for all values of a parameter in a specified range:

$$[\![\texttt{UnionAll } L\texttt{<:T<:}U\ A]\!] = \bigcup_{L \le T \le U} [\![A[T/\texttt{T}]]\!]$$

This is equivalent to an existential type [3]; for each concrete subtype of it there exists a corresponding T. Anecdotally, programmers have often found existential types confussing. We prefer the union interpretation because we are describing sets of values; the notion of "there exists" can be semantically misleading since it sounds like only a single T value might be involved.

Examples

`UnionAll` types are quite expressive. In combination with nominal types they can describe groups of containers such as `UnionAll T<:Number Array{Array{T}}` (all arrays of arrays of some kind of number) or `Array{UnionAll T<:Number Array{T}}` (an array of arrays of potentially different types of number).

In combination with tuple types, `UnionAll` types provide powerful method dispatch specifications. For example `UnionAll T Tuple{Array{T},Int,T}` matches three arguments: an array, an integer, and a value that is an instance of the array's element type. This is a natural signature for a method that assigns a value to a given index within an array.

Type constructors

It is important for any proposed high-level technical computing language to be simple and approachable, since otherwise the value over established powerful-but-complex languages like

C++ is less clear. In particular, type parameters raise usability concerns. Needing to write parameters along with every type is verbose, and requires users to know more about the type system and to know more details of particular types (how many parameters they have and what each one means). Furthermore, in many contexts type parameters are not directly relevant. For example, a large amount of code operates on `Array`s of any element type, and in these cases it should be possible to ignore type parameters.

Consider `Array{T}`, the type of arrays with element type `T`. In most languages with parametric types, the identifier `Array` would refer to a type constructor, i.e. a type of a different *kind* than ordinary types like `Int` or `Array{Int}`. Instead, we find it intuitive and appealing for `Array` to refer to any kind of array, so that a declaration such as `x::Array` simply asserts `x` to be some kind of array. In other words,

$$\texttt{Array} = \texttt{UnionAll T Array}'\texttt{\{T\}}$$

where $\texttt{Array}'$ refers to a hidden, internal type constructor. The `{ }` syntax can then be used to instantiate a `UnionAll` type at a particular parameter value.

Subtyping

Describe algorithm.

Very likely Π_2^{P}-hard. Checking a subtype relation with unions requires checking that for all choices on the left, there exists a choice on the right that makes the relation hold. This matches the quantifier structure of 2-TQBF problems of the form $\forall x_i.\exists y_i.F$ where F is a logical formula. If the formula is rewritten in conjunctive normal form, it corresponds to subtype checking between two tuple types, where the relation must hold for each pair of corresponding types. Now use a type `N{`x`}` to represent $\neg x$. The clause $(x_i \vee y_i)$ can be translated to x_i `<: Union{N{`y_i`}, True}` (where the x_i and y_i are type variables bound by `UnionAll` on the left and right, respectively).

Type system variants

features that are fairly straightforward to add:

- structurally-subtyped records
- mu-recursive types (regular trees)
- regular types (allowing ... in more places)

features that are difficult to add, or possibly break decidability:

- arrow types
- negations
- intersections, multiple inheritance
- universal quantifiers
- lower bounds in quantifiers
- arbitrary predicates, theory of natural numbers, etc.

4.1.2 Data model

4.1.3 Dispatch mechanism

4.2 Performance model

4.2.1 Type inference

4.2.2 Specialization

4.3 Explication through elimination

4.3.1 Conversion and other basic operations

This section will illustrate how we implement key features of technical computing systems using our methodology.

- The abstractions of equality and comparison. Different equivalence classes between is/===, isequal and ==

- Numeric vs lexicographic ordering?

cmp, lexcmp, vs isless, ¡

4.3.2 Numeric types and embeddings

We might prefer "number" to be a single, concrete concept, but the history of mathematics has seen the concept extended many times, from integers to rationals to reals, and then to complex, quaternion, and more. These constructions tend to follow a pattern: a new set of numbers is constructed around a subset isomorphic to an existing set of numbers. For example, the reals are isomorphic to the complex numbers with zero imaginary part.

Human beings happen to be good at equating and moving between isomorphic sets, so it is easy to imagine that the reals and complexes with zero imaginary part are one and the same. But a computer forces us to be specific, and admit that a real number is not complex, and a complex number is not real. And yet the close relationship between them is too compelling not to model in a computer somehow. Here we have a numerical analog to the famous "circle and ellipse" problem in object-oriented programming: the set of circles is isomorphic to the set of ellipses with equal axes, yet neither "is a" relationship in a class hierarchy seems fully correct. An ellipse is not a circle, and in general a circle cannot serve as an ellipse (for example, the set of circles is not closed under the same operations that the set of ellipses is, so a program written for ellipses might not work on circles). This problem implies that a single built-in type hierarchy is not sufficient: we want to model custom *kinds* of relationships between types (e.g. "can be embedded in" in addition to "is a").

Two further problems should also be kept in mind. First, the natural isomorphisms between sets of numbers might not be isomorphisms on a real computer. For example, due to the behavior of floating-point arithmetic, an operation on complex numbers with zero imaginary part might not give an answer equal to the same operation on real numbers.

Second, the contexts that demand use of one type of number or another are often not easily described by type systems. The classic example is square root (`sqrt`), whose result is complex for negative arguments. Including a number's sign in its type is a possibility, but this quickly gets out of hand — should a type system attempt to prove a matrix symmetric before we compute its eigenvalues? While we cannot offer a once-and-for-all solution to these problems, we will show how the flexibility of our proposed mechanism is useful for addressing them.

Implementing type embeddings

Most functions are naturally implemented in the value domain, but some are actually easier to implement in the type domain. One reason is that there is a bottom element, which most data types lack.

Diversity of number and number-like types in practice

Originally, our reasons for implementing all numeric types at the library level were not entirely practical. We had a principled opposition to including such definitions in a compiler, and guessed that being able to define numeric types would help ensure the language was "powerful enough". However, defining numeric and number-like types and their interactions turns out to be surprisingly useful. Once such types become easy to obtain, people find more and more uses for them.

Ordinal types: Pointer, Char

Integer types: Int8, Int16, Int32, Int64, Int128, UInt8, UInt16, UInt32, UInt64, UInt128, BigInt

Floating point types: Float16, Float32, Float64, BigFloat (Float128, double-double)

Fixed point: Fixed32b, Ufixed8, Ufixed16

Extensions: Complex, Quaternion, Interval, DualNumber

Number-like: Date, TimePeriod, DatePeriod, Color, (Units), DNA nucleotide type (bioseq.jl)

dates: different cardinal and ordinal behavior

musical notes

Applications

Ranges illustrate an interesting application of type promotion. A range data type, notated `a:s:b`, represents a sequence of values starting at `a` and ending at `b`, with a distance of `s` between elements (internally, this notation is translated to `colon(a, s, b)`). Ranges seem simple enough, but a reliable, efficient, and generic implementation is difficult to achieve. We propose the following requirements:

- The start and stop values can be passed as different types, but internally should be of the same type.
- Ranges should work with ordinal types, not just numbers (examples include characters, pointers, and calendar dates).
- If any of the arguments is a floating-point number, a special `FloatRange` type designed to cope well with roundoff is returned.

In the case of ordinal types, the step value is naturally of a different type than the elements of the range. For example, one may add 1 to a character to get the "next" encoded character, but it does not make sense to add two characters.

It turns out that the desired behavior can be achieved with six definitions:

First, given three floats of the same type we can construct a `FloatRange` right away:

```
colon{T<:FloatingPoint}(start::T, step::T, stop::T) = FloatRange{T}(start, step, stop)
```

Next, if `a` and `b` are of the same type and there are no floats, we can construct a general range:

```
colon{T}(start::T, step, stop::T) = StepRange(start, step, stop)
```

Now there is a problem to fix: if the first and last arguments are of some non-floating-point numeric type, but the step is floating point, we want to promote all arguments to a common floating point type. We must also do this if the first and last arguments are floats, but the step is some other kind of number:

```
colon{T<:Real}(a::T, s::FloatingPoint, b::T) = colon(promote(a,s,b)...)
```

```
colon{T<:FloatingPoint}(a::T, s::Real, b::T) = colon(promote(a,s,b)...)
```

These two definitions are correct, but ambiguous: if the step is a float of a different type than `a` and `b` both definitions are equally applicable. We can add the following disambiguating definition:

```
colon{T<:FloatingPoint}(a::T, s::FloatingPoint, b::T) = colon(promote(a,s,b)...)
```

All of these five definitions require `a` and `b` to be of the same type. If they are not, we must promote just those two arguments, and leave the step alone in case we are dealing with ordinal types:

```
colon{A,B}(a::A, s, b::B) = colon(convert(promote_type(A,B),a), s, convert(promote_type(
```

This example shows that it is not always sufficient to have a built-in set of "promoting operators". Library functions like this `colon` need more control.

Current approaches

Numbers tend to be among the most complex features of a language. Numeric types usually need to be a special case: in a typical language with built-in numeric types, describing their behavior is beyond the expressive power of the language itself. For example, in C arithmetic operators like `+` accept multiple types of arguments (ints and floats), but no user-defined C function can do this (this situation is of course improved in C++). In Python, a special arrangement is made for `+` to call either an `__add__` or `__radd__` method, effectively providing double-dispatch for arithmetic in a language that is idiomatically single-dispatch.

4.3.3 Multidimensional array indexing

One-dimensional arrays are a simple and essential data structure found in most programming languages. The multi-dimensional arrays required in scientific computing, however, are a

different beast entirely. Allowing any number of dimensions entails a significant increase in complexity. Why? The essential reason is that core properties of the data structure no longer fit in a constant amount of space. The space needed to store the sizes of the dimensions (the array shape) is proportional to the number of dimensions. This does not seem so bad, but becomes a large problem due to three additional facts.

First, code that operates on the dimension sizes needs to be highly efficient. Typically the overhead of a loop is unacceptable, and such code needs to be fully unrolled. Second, in some code the number of dimensions is a *dynamic* property — it is only known at run time. Third, programs may wish to treat arrays with different numbers of dimensions very differently. A vector (1d) might have rather different behaviors than a matrix (2d) (for example, to compute a norm). This kind of behavior makes the number of dimensions a crucial part of program semantics, preventing it from remaining a compiler implementation detail.

These facts pull in different directions. The first fact asks for static analysis. The second fact asks for run-time flexibility. The third fact asks for dimensionality to be part of the type system, but partly determined at run time (for example, via virtual method dispatch). Current approaches choose a compromise. In some systems, the number of dimensions has a strict limit (e.g. 3 or 4), so that separate classes for each case may be written out in full. Other systems choose flexibility, and just accept that most or all operations will be dynamically dispatched. Other systems might provide flexibility only at compile time, for example a template library where the number of dimensions must be statically known.

Whatever decision is made, rules must be defined for how various operators act on dimensions. For now we will focus on indexing, since selecting parts of arrays has particularly rich behavior with respect to dimensionality. For example, if a single row or column of a matrix is selected, does the result have one or two dimensions? Array implementations prefer to invoke general rules to answer such questions. Such a rule might say "dimensions indexed with scalars are dropped", or "trailing dimensions of size one are dropped", or "the rank of the result is the sum of the ranks of the indexes" (as in APL).

Our goal here is a bit unusual: we are not concerned with which rules might work best, but merely with how they can be specified, so that domain experts can experiment.

In fact different domains want different things. E.g. in images, each dimension might be quite different, e.g. time vs. space vs. color, so you don't want to drop or rearrange dimensions very often.

Here are our ground rules:

1. You can't manually implement the behavior inside the compiler
2. The compiler must be able to reasonably understand the program
3. The code must be reasonably easy to write

How are such rules implemented? For a language with built-in multidimensional arrays, the compiler will analyze indexing expressions and determine an answer using hard-coded logic. However, this approach is not satisfying: we would rather implement the behavior in libraries, so that different kinds of arrays may be defined, or so that rules of similar complexity may be defined for other kinds of objects. But these kinds of rules are unusually difficult to implement in libraries. If a library writes out its indexing logic using imperative code, the host language compiler is not likely to be able to analyze it. Using compile-time abstracton (templates) would provide better performance, but such libraries tend to be difficult to write (and read), and the full complement of indexing behavior expected by technical users strains the capabilities of such systems.

Our dispatch mechanism permits a novel solution. If a multiple dispatch system supports variadic functions and argument "splicing" (the ability to pass a structure of n values as n separate arguments to a function), then indexing behavior can be defined as method signatures.

This solution is still a compromise among the factors outlined above, but it is a new compromise that provides a net-better solution.

Below we define a function `index_shape` that computes the shape of a result array given

a series of index arguments. We show three versions, each implementing a different rule that users in different domains might want:

```
# drop dimensions indexed with scalars
index_shape() = ()
index_shape(i::Real, I...) = index_shape(I...)
index_shape(i, I...) = tuple(length(i), index_shape(I...)...)

# drop trailing dimensions indexed with scalars
index_shape(i::Real...) = ()
index_shape(i, I...) = tuple(length(i), index_shape(I...)...)

# rank summing (APL)
index_shape() = ()
index_shape(i, I...) = tuple(size(i)..., index_shape(I...)...)
```

Inferring the length of the result of `index_shape` is sufficient to infer the rank of the result array.

These definitions are concise, easy to write, and possible for a compiler to understand fully using straightforward techniques.

Here is a sample derivation for the call `index_shape(1:m,1,1:n)` (the argument type tuple is `(Range1,Int,Range1)`), using the first definition above (dropping scalar-indexed dimensions):

```
index_shape(1:n, 1, 1:m) => tuple(length(::Range1)::Int, index_shape((::Int, ::Range1)..

index_shape((::Int, ::Range1)...) => index_shape(::Int, ::Range1)

index_shape(::Int, ::Range1) => index_shape((::Range1,)...)

index_shape((::Range1,)...) => index_shape(::Range1)
```

```
index_shape(::Range1) => tuple(length(::Range1)::Int, index_shape(()...)...)

index_shape(()...) => index_shape()::()

back substitute => tuple(length(::Range1)::Int, index_shape(()...)::()...)::(Int,)

back substitute => tuple(length(::Range1)::Int, ::(Int,)...)

tuple(::Int, ::(Int,)...) => tuple(::Int, ::Int)

::(Int, Int)
```

The result type is determined using only dataflow type inference, plus a rule for splicing an immediate container (the type of `f((a,b)...)` is the type of `f(a,b)`). Argument list destructuring takes place inside the type intersection operator used to combine argument types with method signatures.

This approach does not depend on any heuristics. Each call to `index_shape` simply requires one recursive invocation of type inference. This process reaches the base case `()` for these definitions, since each recursive call handles a shorter argument list (for less-well-behaved definitions, we might end up invoking a widening operator instead).

```
diverge() = randbool() ? () : tuple(1, diverge()...)
```

4.3.4 Array views

4.3.5 Units

4.3.6 Even more elimination?

Some features of the language could be even further eliminated. For example data types could be implemented in terms of lambda abstractions. But certain patterns are so useful that they might as well be provided in a standard form. It also probably makes the compiler much more efficient not to need to pass around and repeatedly analyze full representations of the meanings of such ubiquitous constructs.

Chapter 5

Case studies

5.1 Numerical linear algebra

Multiple dispatch on special matrices

29 LAPACK types via composition of 9 types, issue 8240

5.2 Boundary-element method

There are lots of general packages for FEM problems, but it is much more difficult to create such a package for BEM problems. The method requires integrating functions with singularities, many times in the inner loop of code that builds the problem matrix. Integrating such functions numerically on each iteration is much too slow. As a result, many special-purpose implementations have been written by hand for different problems.

Some recent work (TODO cite Homer) managed a more general solution, using Mathematica to generate C++ code for different cases. This worked well, but was difficult to implement and the resulting system is difficult to use. We see the familiar pattern of using multiple languages and code-generation techniques, with coordination of the overall process done either manually or with ad-hoc scripts. To polish the implementation for use as a practical library, a likely next step would be to add a Python interface, adding yet another

layer of complexity.

Features of julia this demonstrates:

- functions that need to dispatch on more than one argument - staged methods providing a natural way to integrate code generation - doing specialization through the object system, leading to a "flat" system

The code can be structured as a simple function library.

5.3 Galerkin matrix assembly for singular kernels

A typical problem in computational science is to form a discrete approximation of some infinite-dimensional linear operator $\mathcal{L}$ with some finite set of basis functions $\{b_m\}$ via a Galerkin approach [refs], which leads to a matrix L with entries $L_{mn} = \langle b_m, \mathcal{L} b_n \rangle = \langle b_m, b_n \rangle_{\mathcal{L}}$ where $\langle \cdot, \cdot \rangle$ denotes some inner product (e.g. $\langle u, v \rangle = \int uv$ is typical) and $\langle \cdot, \cdot \rangle_{\mathcal{L}}$ is the *bilinear form* of the problem. Computing these matrix elements is known as the matrix *assembly* step, and its performance is a crucial concern for solving partial differential equations (PDEs) and integral equations (IEs).

5.3.1 The "easy" case: nonsingular assembly

For example, in the finite-element method (FEM) [refs], the basis functions b_m are typically low-order polynomials defined piecewise over geometric elements (typically triangles or tetrahedra), and $\mathcal{L}$ is typically a differential operator like $-\nabla \cdot c(x) \nabla$ for some coefficients $c(x)$, which leads to a bilinear form $\langle b_m, b_n \rangle_{\mathcal{L}} = \int \nabla b_m \cdot c(x) \nabla b_n$ (after integration by parts). Because the basis functions are localized and $\mathcal{L}$ consists of local operations, the matrix L is sparse and L_{mn} need only be computed for m and n corresponding to neighboring elements. Moreover, these integrals are straightforward to evaluate by standard cubature schemes because the integrands are *nonsingular*: they typically have no divergences or discontinuities. In particular, because the functions b_m and c are usually smooth within a single element, one can use a fixed low-order cubature rule: you evaluate the integrand at a handful of pre-

computed points within each element, multiply by precomputed weights, and sum to obtain the approximate integral.

Even so, the basis functions and the coefficient function $c(x)$ may need to be evaluated tens of millions of times for even a moderate-size mesh in three dimensions, so production FEM implementations in traditional high-level dynamic languages such as Matlab and Python are forced to offload matrix assembly to external C and C++ code. For example, the popular FEniCS [ref] and Firedrake [ref] FEM packages for Python both implement domain-specific compilers: a symbolic expression for the bilinear form is combined with fragments of user-specified C++ code to define functions like $c(x)$, compiled to C++ code, and then compiled to object code which is dynamically loaded. In Julia, we believe this could be simplified considerably because functions like $c(x)$ could be defined directly in Julia and code generation/compilation could be performed entirely within Julia without a C++ intermediary. Indeed, preliminary experiments with pure Julia FEM implementations [ref http://www.codeproject.com/Articles/579983/Finite-Element-programming-in-Julia] have demonstrated performance comparable to sophisticated solutions like FEniCS in Python and FreeFem++ in C++ [ref]

5.3.2 Singular assembly for integral operators

A much more challenging case of Galerkin matrix assembly arises for singular *integral* operators $\mathcal{L}$, which act by convolving their operand against a singular "kernel" function $K(x)$: $u = \mathcal{L}v$ means that $u(x) = \int K(x-x')v(x')dx'$. For example, in electrostatics and other Poisson problems, the kernel is $K(x) = 1/|x|$ in three dimensions and $\ln|x|$ in two dimensions, while in scalar Helmholtz (wave) problems it is $e^{ik|x|}/|x|$ in three dimensions and a Hankel function $H_0^{(1)}(k|x|)$ in two dimensions. Formally, Galerkin discretizations lead to matrix assembly problems similar to those above: $L_{mn} =: \langle b_m, \mathcal{L}b_n\rangle = \int b_m(x)K(x-x')b_n(x')dx\,dx'$. However, there are several important differences from FEM:

- The kernel $K(x)$ nearly always diverges for $|x| = 0$, which means that generic cubature schemes are either unacceptably inaccurate (for low-order schemes) or unacceptably

costly (for adaptive high-order schemes, which require huge numbers of cubature points around the singularity), or both.

- Integral operators typically arise for *surface* integral equations (SIEs) [ref], and involve unknowns on a surface. The analogue of the FEM discretization is then a boundary element method (BEM) [ref], which discretizes a surface into elements (e.g. triangles), with basis functions that are low-order polynomials defiend piecewise in the elements. However, there are also volume integral equations (VIEs) which have FEM-like volumetric meshes and basis functions.

- The matrix L is typically dense, since K is long-range. For large problems, L is often stored and applied implicitly via fast-multipole methods [refs] and similar schemes, but even in this case the diagonal L_{mm} and the entries L_{mn} for adjacent elements must typically be computed explicitly. (Moreover, these are the integrals in which the K singularity is present.)

These difficulties are part of the reason why there is currently *no* truly "generic" BEM software, analogous to FEniCS for FEM: essentially all practical BEM code is written for a specific integral kernel and a specific class of basis functions arising in a particular physical problem. Changing anything about the kernel or the basis—for example, going from two- to three-dimensional problems—is a major undertaking.

We believe that Julia should be an ideal platform on which to attack this problem:

- Multiple dispatch allows the cubature scheme to be selected at compile-time based on the dimensionality, the degree of the singularity, the degree of the polynomial basis, and so on, and allows specialized schemes to be added easily for particular problems with no runtime penalty.

- Staged functions allow computer-algebra systems to be invoked at compile-time to generate specialized cubature schemes for particular kernels. New developments in BEM integration schemes [ref Homer] have provided efficient cubature-generation algorithms

of this sort, but it has not yet been practical to integrate them with runtime code in a completely automated way.

A prototype implementation of this approach follows.

5.4 Dates

compare to python DateTime, compare code length

5.5 JUMP

Metaprogramming tools reused for symbolic algebra

5.6 Computational geometry

robust predicates (dispatch over points, lines) and VoronoiDelaunay.jl benchmarked against CGAL

5.7 Beating the incumbents

- erfinv and digamma using horner macro - randn beating matlab

If nothing else, demonstrates that removing glue code overhead is worthwhile.

grisu: 6kLOC to 1kLOC (PR 7291)

5.8 misc staged functions

tim holy in issue 8839:

"without staged functions in my initial post in 8235. The take-home message: generating all methods through dimension 8 resulted in more than 5000 separate methods, and required

over 4 minutes of parsing & lowering time (i.e., a 4-minute delay while compiling julia). By comparison, the stagedfunction implementation loads in a snap, and of course can go even beyond 8 dimensions."

Chapter 6

Conclusion

Here is a snapshot of Julia in April, 2014

Show all the stuff that will be fun to look at because it'll be so antiquated by 2019

by following this kind of approach computer scientists can have it both ways, and make new interesting things that are also immediately useful.

A generation of dynamic languages have been designed by trying variants of the class-based object oriented paradigm. This process has been aided by the development of standard techniques (e.g. bytecode VMs) and reusable infrastructure such as code generators, garbage collectors, and whole VMs like the JVM and CLR.

It is possible to envision a future generation of languages that generalize this design to set-theoretic subtyping instead of just classes. This next generation will require its own new tools, such as partial evaluators (already under development in PyPy and Truffle). One can also imagine these future language designers wanting reusable program analyses, and tools for developing lattices and their operators.

It is interesting to observe that the data model of a language like Julia consists of two key relations: the subtype relation, which is relatively well understood and enjoys useful properties like transitivity, but also the typeof relation, which relates individual values to their types (i.e. the 'typeof' function). The typeof relation appears not to be transitive, and also has a degree of arbitrariness: a value is of a type merely because it is labeled as such,

and because various bits of code conspire to ensure that this labeling makes sense according to various criteria.

We have speculated about whether future languages will be able to do away with this distinction. One approach is $\lambda_{\aleph}$. We have also speculated that this could be done using types based on non-well-founded set theory, combining the subset-of and element-of relations using self-containing sets. We are not yet sure what a practical language based on this idea might look like.

There are several key aspects of performance programming that our design does not directly address.

Talk about storage and in-place optimizations.

Appendix A

Subtyping algorithm

```
abstract Ty

type TypeName
  super::Ty
  TypeName() = new()
end

type TagT <: Ty
  name::TypeName
  params
  vararg::Bool
end

type UnionT <: Ty
  a
  b
end

type Var
  ub
end

type UnionAllT <: Ty
  var::Var
  T
end

## Any, Bottom, and Tuple
const AnyT = TagT(TypeName(), ());  AnyT.name.super = AnyT
type BottomTy <: Ty; end
const BottomT = BottomTy()
const TupleName = TypeName(); TupleName.super = AnyT

## type application
inst(t::TagT) = t
inst(t::UnionAllT, param) = subst(t.T, Dict{Any,Any}(t.var => param))
inst(t::UnionAllT, param, rest...) = inst(inst(t,param), rest...)
super(t::TagT) = inst(t.name.super, t.params...)

extend(d::Dict, k, v) = (x = copy(d); x[k]=v; x)
```

```
subst(t,           env) = t
subst(t::TagT,     env) =
  t===AnyT ? t : TagT(t.name, map(x->subst(x,env), t.params), t.vararg)
subst(t::UnionT,   env) = UnionT(subst(t.a,env), subst(t.b,env))
subst(t::Var,      env) = get(env, t, t)
subst(t::UnionAllT, env) = let newVar = Var(subst(t.var.ub, env))
  UnionAllT(newVar, subst(t.T, extend(env, t.var, newVar)))
end

rename(t::UnionAllT) = let v = Var(t.var.ub)
  UnionAllT(v, inst(t,v))
end

## subtype implementation
type Bounds
  lb           # current lower and upper bounds of a Var
  ub
  depth::Int   # invariant nesting depth of a Var's UnionAll
  right::Bool  # this Var is on the right-hand side of A <: B
end

type UnionSearchState
  i::Int       # (0 <= i < nbits(idxs))
  idxs::Int64  # bit vector representing combination being tested
  UnionSearchState() = new(0,0)
end

type Env
  vars::Dict{Var,Bounds}
  depth::Int
  Ldepth::Int  # # of union decision points we're inside
  Lnew::Int    # # unions found at next nesting depth
  Lunions::Vector{UnionSearchState}
  Rdepth::Int  # same on right-hand side
  Rnew::Int
  Runions::Vector{UnionSearchState}
  Env() = new(Dict{Var,Bounds}(), 1,
              1,0,UnionSearchState[], 1,0,UnionSearchState[])
end

issub(x, y) = forall_exists_issub(x, y, Env(), 0)
issub(x, y, env)         = (x === y)
issub(x::Ty, y::Ty, env) = (x === y) || x === BottomT

function forall_exists_issub(x, y, env, nL)
```

```
  for forall in 1:(1<<nL)
    if !isempty(env.Lunions)
      env.Lunions[end].idxs = forall
    end

    !exists_issub(x, y, env, 0) && return false

    if env.Lnew > 0
      push!(env.Lunions, UnionSearchState())
      sub = forall_exists_issub(x, y, env, env.Lnew)
      pop!(env.Lunions)
      !sub && return false
    end end
  return true
end

function exists_issub(x, y, env, nR)
  for exists in 1:(1<<nR)
    if !isempty(env.Runions)
      env.Runions[end].idxs = exists
    end
    for ru in env.Runions; ru.i = -1; end
    for lu in env.Lunions; lu.i = -1; end
    env.Ldepth = env.Rdepth = 1
    env.Lnew = env.Rnew = 0

    sub = issub(x, y, env)

    if env.Lnew > 0
      return true # return up to forall_exists_issub
    end
    if env.Rnew > 0
      push!(env.Runions, UnionSearchState())
      found = exists_issub(x, y, env, env.Rnew)
      pop!(env.Runions)
      if env.Lnew > 0
        return true # return up to forall_exists_issub
      end
    else
      found = sub
    end
    found && return true
  end
  return false
end
```

```
function union_issub(a::UnionT, b::Ty, env)
  if env.Ldepth > length(env.Lunions)
    env.Lnew += 1
    return true
  end
  L = env.Lunions[env.Ldepth]; L.i += 1
  env.Ldepth += 1
  ans = issub(a.((L.idxs&(1<<L.i)!=0) + 1), b, env)
  env.Ldepth -= 1
  return ans
end

function issub_union(a::Ty, b::UnionT, env)
  if env.Rdepth > length(env.Runions)
    env.Rnew += 1
    return true
  end
  R = env.Runions[env.Rdepth]; R.i += 1
  env.Rdepth += 1
  ans = issub(a, b.((R.idxs&(1<<R.i)!=0) + 1), env)
  env.Rdepth -= 1
  return ans
end

issub(a::UnionT, b::UnionT, env) = a === b || union_issub(a, b, env)
issub(a::UnionT, b::Ty, env) = union_issub(a, b, env)
issub(a::Ty, b::UnionT, env) = a === BottomT || issub_union(a, b, env)

function issub(a::TagT, b::TagT, env)
  a === b && return true
  b === AnyT && return true
  a === AnyT && return false
  if a.name !== b.name
    a.name === TupleName && return false
    return issub(super(a), b, env)
  end
  if a.name === TupleName
    va, vb = a.vararg, b.vararg
    la, lb = length(a.params), length(b.params)
    if va && (!vb || la < lb)
      return false
    end
    ai = bi = 1
    while true
```

```
      ai > la && return bi > lb || (bi==lb && vb)
      bi > lb && return false
      !issub(a.params[ai], b.params[bi], env) && return false
      ai==la && bi==lb && va && vb && return true
      if ai < la || !va
        ai += 1
      end
      if bi < lb || !vb
        bi += 1
      end
    end
  else
    env.depth += 1  # crossing invariant constructor, increment depth
    yes = true
    for i = 1:length(a.params)
      ai, bi = a.params[i], b.params[i]
      yes &= (issub(ai, bi, env) && issub(bi, ai, env))
    end
    env.depth -= 1
    return yes
  end
end

function issub(a::Var, b::Ty, env)
  aa = env.vars[a]
  # Vars are fully checked by the "forward" direction of A<:B in
  # invariant position. So just return true when checking the "flipped"
  # direction B<:A.
  aa.right && return true
  if env.depth != aa.depth
    # Var <: non-Var can only be true when there are no invariant
    # constructors between the UnionAll and this occurrence of Var.
    return false
  end
  return issub(aa.ub, b, env)
end

function issub(a::Var, b::Var, env)
  a === b && return true
  aa = env.vars[a]
  aa.right && return true
  bb = env.vars[b]
  if aa.depth != bb.depth
    return false
  end
```

```
  if env.depth > bb.depth
    # if there are invariant constructors between a UnionAll and
    # this occurrence of Var, then we have an equality constraint on Var.
    if isa(bb.ub,Var)
      # right-side Var cannot equal more than one left-side Var, e.g.
      # (L,L) <: (T,S) but not (T,S) <: (R,R)
      bb.lb = a
      return bb.ub === a
    end
    if isa(bb.lb,Var)
      bb.ub = a
      return bb.lb === a
    end
    if !(issub(bb.lb, aa.lb, env) && issub(aa.ub, bb.ub, env))
      # make sure equality constraint is within the current bounds of Var
      return false
    end
    bb.lb = bb.ub = a
  else
    !issub(aa.ub, bb.ub, env) && return false
    # track greatest lower bound from covariant position. e.g.
    # (Int, Real, Integer, Array{?}) <: (T, T, T, Array{T})
    # is only true if ? >: Real.
    if issub(bb.lb, aa.ub, env)
      bb.lb = a
    end
  end
  return true
end

function issub(a::Ty, b::Var, env)
  bb = env.vars[b]
  !bb.right && return true
  if env.depth > bb.depth
    if isa(bb.ub,Var) || !(issub(bb.lb, a, env) && issub(a, bb.ub, env))
      return false
    end
    bb.lb = bb.ub = a
  else
    !issub(a, bb.ub, env) && return false
    if issub(bb.lb, a, env)
      bb.lb = a
    end
  end
  return true
```

```
end

function issub_unionall(t::Ty, u::UnionAllT, env, R)
  (t === u || t === BottomT) && return true
  haskey(env.vars, u.var) && (u = rename(u))
  env.vars[u.var] = Bounds(BottomT, u.var.ub, env.depth, R)
  ans = R ? issub(t, u.T, env) : issub(u.T, t, env)
  delete!(env.vars, u.var)
  return ans
end

issub(a::UnionAllT, b::UnionAllT, env) = issub_unionall(a, b, env, true)
issub(a::UnionT, b::UnionAllT, env) = issub_unionall(a, b, env, true)
issub(a::UnionAllT, b::UnionT, env) = issub_unionall(b, a, env, false)
issub(a::Ty, b::UnionAllT, env) = issub_unionall(a, b, env, true)
issub(a::UnionAllT, b::Ty, env) = issub_unionall(b, a, env, false)
```

Bibliography

[1] Allen, E., Chase, D., Hallett, J., Luchangco, V., Maessen, J.-W., Ryu, S., Steele, Jr., G. L., and Tobin-Hochstadt, S. The fortress language specification version 1.0. Tech. rep., March 2008.

[2] Bruce, K., Cardelli, L., Castagna, G., Leavens, G. T., and Pierce, B. On binary methods. *Theor. Pract. Object Syst. 1*, 3 (Dec. 1995), 221–242.

[3] Cardelli, L., and Wegner, P. On understanding types, data abstraction, and polymorphism. *ACM Comput. Surv. 17*, 4 (Dec. 1985), 471–523.

[4] Day, M., Gruber, R., Liskov, B., and Myers, A. C. Subtypes vs. where clauses: Constraining parametric polymorphism. In *Proceedings of the Tenth Annual Conference on Object-oriented Programming Systems, Languages, and Applications* (New York, NY, USA, 1995), OOPSLA '95, ACM, pp. 156–168.

[5] Igarashi, A., and Nagira, H. Union types for object-oriented programming. In *Proceedings of the 2006 ACM Symposium on Applied Computing* (New York, NY, USA, 2006), SAC '06, ACM, pp. 1435–1441.

[6] Leavens, G. T., and Millstein, T. D. Multiple dispatch as dispatch on tuples. *ACM SIGPLAN Not. 33*, 10 (Oct. 1998), 374–387.

[7] Morris, J. H. J. *Lambda-Calculus Models of Programming Languages*. PhD thesis, Massachusetts Institute of Technology, 1968.

[8] Smith, D., and Cartwright, R. Java type inference is broken: Can we fix it? In *Proceedings of the 23rd ACM SIGPLAN Conference on Object-oriented Programming Systems Languages and Applications* (New York, NY, USA, 2008), OOPSLA '08, ACM, pp. 505–524.