

How to write a great research paper

Simon Peyton Jones
Microsoft Research, Cambridge
Adapted for REU



Writing papers is a skill

- Many papers are badly written
- Good writing is a skill you can learn
- It's a skill that is worth learning:
 - You will get more brownie points (more papers accepted etc)
 - Your ideas will have more impact (people cite your papers)
 - You will have better ideas

Increasing importance





Do not be intimidated

Fallacy You need to have a fantastic idea before you can write a paper. (Everyone else seems to.)

Write a paper,
and give a talk, about
any idea,
no matter how weedy and insignificant it
may seem to you



Do not be intimidated

Write a paper, and give a talk, about any idea, no matter how insignificant it may seem to you

- **Writing the paper is how you develop the idea in the first place**
- It usually turns out to be more interesting and challenging than it seemed at first

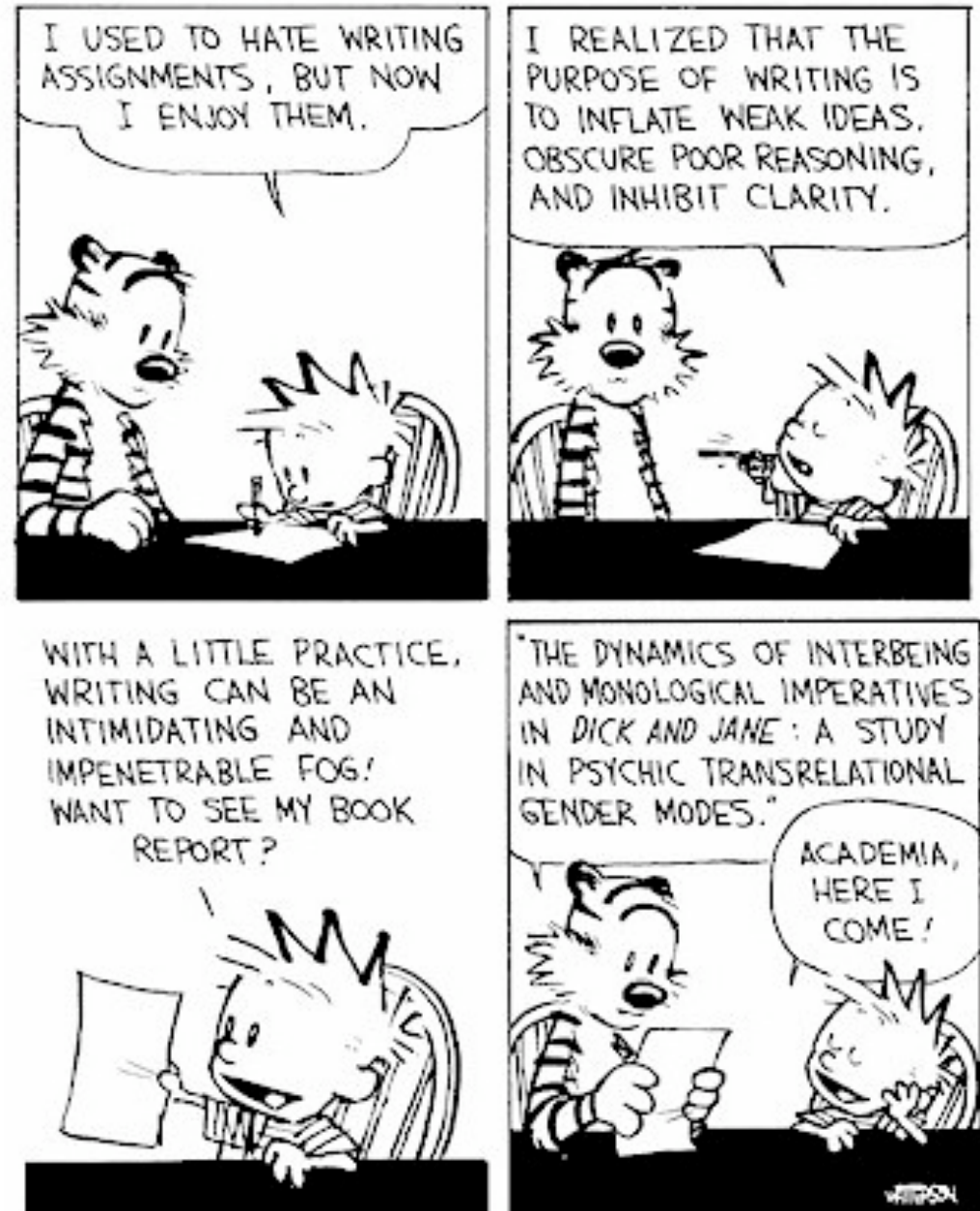


The purpose of your paper

Why bother?

Fallacy

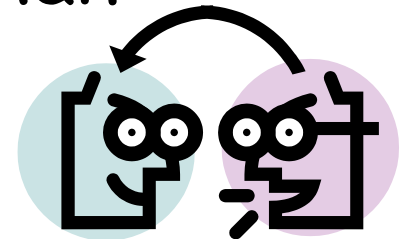
we write papers and give talks mainly to impress others, gain recognition, and get promoted





Papers communicate ideas

- Your goal: to infect the mind of your reader with **your idea**, like a virus
- Papers are far more durable than programs (think Mozart)



The greatest ideas are (literally)
worthless if you keep them to
yourself



The Idea

Idea

A re-usable insight,
useful to the reader

- Figure out what your idea is
- Make certain that the reader is in no doubt what the idea is. Be 100% explicit:
 - “The main idea of this paper is....”
 - “In this section we present the main contributions of the paper.”
- Many papers contain good ideas, but do not distil what they are.



One ping

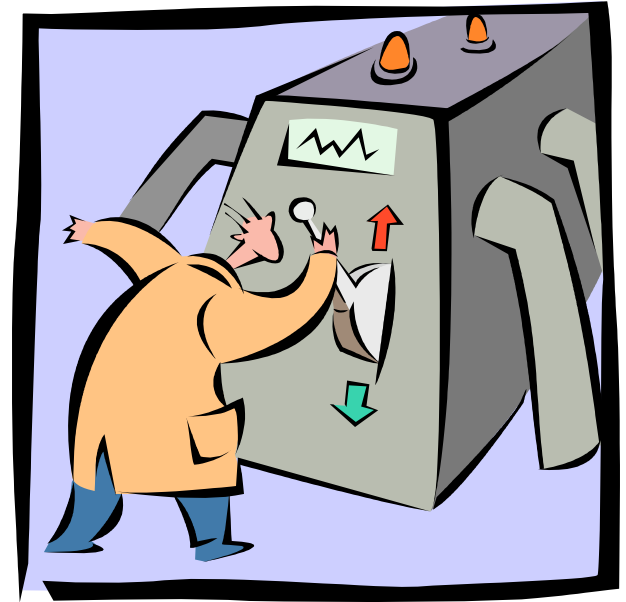
- Your paper should have just one “ping”: one clear, sharp idea
- Read your paper again: can you hear the “ping”?
- You may not know exactly what the ping is when you start writing; but you must know when you finish
- If you have lots of ideas, write lots of papers

Thanks to Joe Touch for “one ping”



The purpose of your paper is not...

To describe
the WizWoz
system

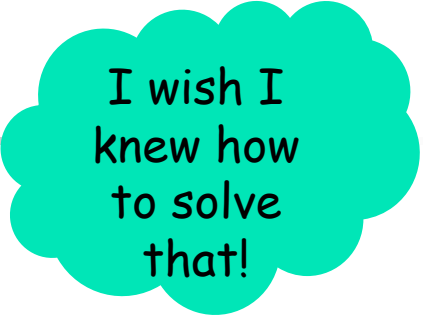


- Your reader does not have a WizWoz
- She is primarily interested in re-usable brain-stuff, not executable artefacts

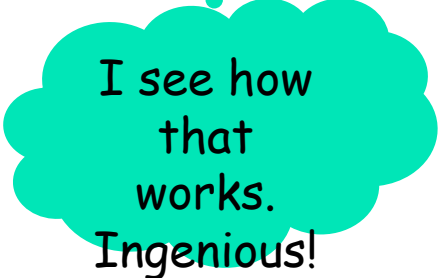


Your narrative flow

- Here is a problem
- It's an interesting problem
- It's an unsolved problem
- **Here is my idea**
- My idea works (details, data)
- Here's how my idea compares to other people's approaches



I wish I knew how to solve that!



I see how that works.
Ingenious!





Structure (conference paper)

- Title (1000 readers)
- Abstract (4 sentences, 100 readers)
- Introduction (1 page, 100 readers)
- The problem (1 page, 10 readers)
- My idea (2 pages, 10 readers)
- The details (5 pages, 3 readers)
- Related work (1-2 pages, 10 readers)
- Conclusions and further work (0.5 pages)



The abstract

- I usually write the abstract last
- Used by program committee members to decide which papers to read
- Is a summary of what the paper is about.
- Four sentences [Kent Beck]
 1. State the problem
 2. Say why it's an interesting problem
 3. Say what your solution achieves
 4. Say what follows from your solution



Example

1. Many papers are badly written and hard to understand
2. This is a pity, because their good ideas may go unappreciated
3. Following simple guidelines can dramatically improve the quality of your papers
4. Your work will be used more, and the feedback you get from others will in turn improve your research



Structure

- Abstract (4 sentences)
- **Introduction** (1 page)
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- Related work (1-2 pages)
- Conclusions and further work (0.5 pages)



The introduction (1 page)

1. Describe the problem
2. State your contributions

...and that is all

The introduction page gives people the first impression of your paper and 80% of the time that impression does not change. This is backed up by a study in psychology.



Framing your research problems

- X is good for Y (in the context of Z)
- X can be extended to achieve Y (in the context of Z)
- The application of X to Y produces an improved result
- The adoption of X facilitates Y (the data in Z format)
- An X approach to the problem of Y mitigates the need for Z



Important words/phrases

- **Optimal:** Does not mean "very good"
 - we picked the optimal value for X.. **NO!!** (unless you can prove it)
 - we picked a value for X that produced the best
- **Proved:** Does not mean "demonstrated"
 - with experiments we proved that our..**NO!** (experiments rarely prove things)
 - with experiments we offer evidence that our ..
- **Significant:** There is a danger of confusing the informal statement and the statistical claim
 - our idea is significantly better than Smiths
 - our idea is statistical significantly better than Smiths, at a confidence level of ..



Important words/phrases (Cont.)

Complexity: Has an overloaded meaning in computer science

- - *The X algorithms complexity means it is not a good solution (complex= intricate)*
- - *The X algorithms time complexity is $O(n^6)$ meaning it is not a good solution*
- • **It is easy to see: First, this is a cliché. Second, are you sure it is easy?**
 - - *It is easy to see that $P = NP$*
- • **Actual: Almost always has no meaning in a sentence**
 - - *It is an actual B-tree -> It is a B-tree*
 - - *There are actually 5 ways to hash a string -> There are 5 ways to hash a string*
- **Theoretically: Almost always has no meaning in a sentence**
 - - *Theoretically we could have jam or jelly on our toast.*
- • **etc : Only use it if the remaining items on the list are obvious.**
 - - *We named the buckets for the 7 colors of the rainbow, red, orange, yellow etc.*
 - - *We measure performance factors such as stability, scalability, etc. No!*



State your contributions

- Write the list of contributions first
- **The list of contributions drives the entire paper:** the paper substantiates the claims you have made
- Reader thinks "gosh, if they can really deliver this, that's be exciting; I'd better read on".



State your contributions

Which of the two is best in practice? The trouble is that the evaluation model has a pervasive effect on the implementation, so it is too much work to implement both and pick the best. Historically, compilers for strict languages (using call-by-value) have tended to use eval/apply, while those for lazy languages (using call-by-need) have often used push/enter, but this is 90% historical accident — either approach will work in both settings. In practice, implementors choose one of the two approaches based on a qualitative assessment of the trade-offs. In this paper we put the choice on a firmer basis:

- We explain precisely what the two models are, in a common notational framework (Section 4). Surprisingly, this has not been done before.
- The choice of evaluation model affects many other design choices in subtle but pervasive ways. We identify and discuss these effects in Sections 5 and 6, and contrast them in Section 7. There are lots of nitty-gritty details here, for which we make no apology — they were far from obvious to us, and articulating these details is one of our main contributions.

In terms of its impact on compiler and run-time system complexity, eval/apply seems decisively superior, principally because push/enter requires a stack like no other: stack-walking

Bulleted list
of
contributions

Do not leave the
reader to guess what
your contributions are!



Contributions should be refutable

NO!	YES!
We describe the WizWoz system. It is really cool.	We give the syntax and semantics of a language that supports concurrent processes (Section 3). Its innovative features are...
We study its properties	We prove that the type system is sound, and that type checking is decidable (Section 4)
We have used WizWoz in practice	We have built a GUI toolkit in WizWoz, and used it to implement a text editor (Section 5). The result is half the length of the Java version.



No "rest of this paper is..."

- **Not:** "The rest of this paper is structured as follows. Section 2 introduces the problem. Section 3 ... Finally, Section 8 concludes".
- **Instead, use forward references from the narrative in the introduction.**
The introduction (including the contributions) should survey the whole paper, and therefore forward reference every important part.



Structure

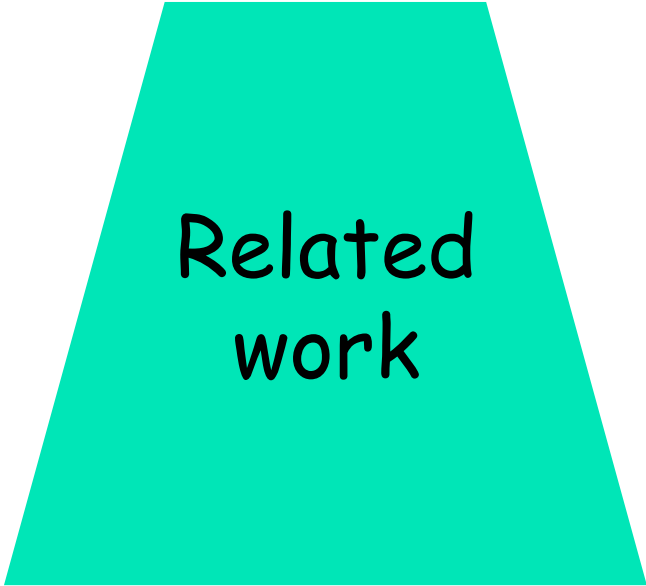
- Abstract (4 sentences)
- Introduction (1 page)
- ~~Related work~~
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- Related work (1-2 pages)
- Conclusions and further work (0.5 pages)



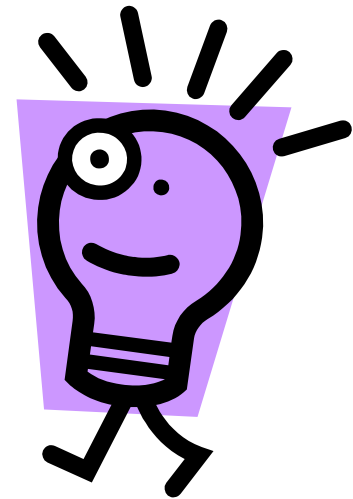
No related work yet!



Your reader



Related
work



Your idea

We adopt the notion of transaction from Brown [1], as modified for distributed systems by White [2], using the four-phase interpolation algorithm of Green [3]. Our work differs from White in our advanced revocation protocol, which deals with the case of priority inversion as described by Yellow [4].

No related work yet

- **Problem 1:** the reader knows nothing about the problem yet; so your (carefully trimmed) description of various technical tradeoffs is absolutely incomprehensible
- **Problem 2:** describing alternative approaches gets between the reader and your idea

I feel stupid



I feel tired



Structure

- Abstract (4 sentences) about 200 words
- Introduction (1 page) 1.5 pages
- The problem (1 page) 0.5 pages
- My idea (2 pages) overview of approach
1 page
- The details (5 pages) technical 2 pages,
experiment (2-3 pages)
- Related work (1-2 pages) 1.5 pages
- Conclusions and further work (0.5 pages)



Presenting the idea

3. The idea

Consider a bifurcated semi-lattice D , over a hyper-modulated signature S . Suppose p_i is an element of D . Then we know for every such p_i there is an epi-modulus j , such that $p_j < p_i$.

- Sounds impressive...but
- Sends readers to sleep
- In a paper you **MUST** provide the details, but **FIRST** convey the idea



Presenting the idea

- Explain it as if you were speaking to someone using a whiteboard
- **Conveying the intuition is primary**, not secondary
- Once your reader has the intuition, she can follow the details (but not vice versa)
- Even if she skips the details, she still takes away something valuable



Putting the reader first

- **Do not** recapitulate your personal journey of discovery. This route may be soaked with your blood, but that is not interesting to the reader.
- Instead, choose the most direct route to the idea.



The payload of your paper

Introduce the problem, and
your idea, using

EXAMPLES

and only then present the
general case



Using examples/figures

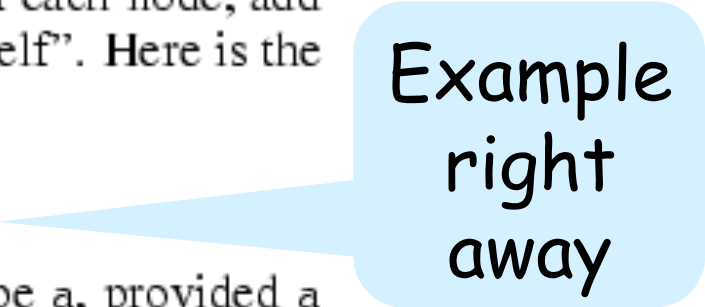
2 Background

To set the scene for this paper, we begin with a brief overview of the *Scrap your boilerplate* approach to generic programming. Suppose that we want to write a function that computes the size of an arbitrary data structure. The basic algorithm is “for each node, add the sizes of the children, and add 1 for the node itself”. Here is the entire code for `gsize`:

```
gsize :: Data a => a -> Int
gsize t = 1 + sum (gmapQ gsize t)
```

The type for `gsize` says that it works over any type `a`, provided `a` is a *data* type — that is, that it is an instance of the class `Data`¹. The definition of `gsize` refers to the operation `gmapQ`, which is a method of the `Data` class:

```
class Typeable a => Data a where
  ...other methods of class Data...
  gmapQ :: (forall b. Data b => b -> r) -> a -> [r]
```



Example
right
away

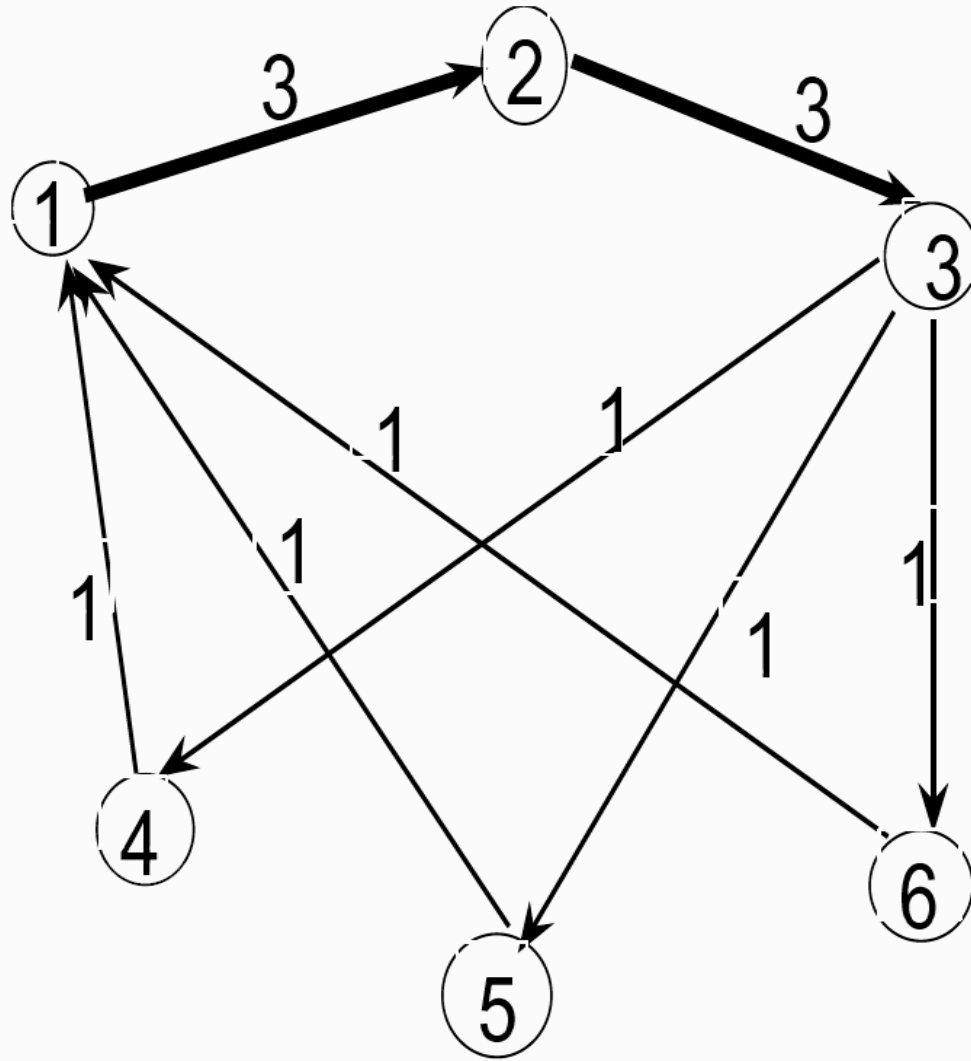
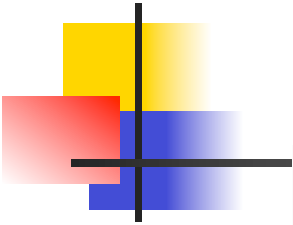


Fig 1. Sequence graph example



Reproducibility

- “Reproducibility is one of the main principles of the scientific method, and refers to the ability of a test or experiment to be accurately reproduced, or replicated, by someone else working independently.”
- Most papers have experiments that are not reproducible.
- However there is an increasing movement to insist upon reproducibility.
- For SIGKDD and SIGMOD, the reviewers can now reject a paper simply by saying it is not reproducible.



The details: evidence

- Your introduction makes claims
- The body of the paper provides **evidence to support each claim**
- Check each claim in the introduction, identify the evidence, and forward-reference it from the claim
- Evidence can be: analysis and comparison, theorems, measurements, case studies



Structure

- Abstract (4 sentences)
- Introduction (1 page)
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- **Related work** (1-2 pages)
- Conclusions and further work (0.5 pages)



Related work

Fallacy

To make my work look good, I have to make other people's work look bad



The truth: credit is not like money

Giving credit to others does not diminish the credit you get from your paper

- Warmly acknowledge people who have helped you
- Be generous to the competition. "In his inspiring paper [Foo98] Foogle shows.... We develop his foundation in the following ways..."
- Acknowledge weaknesses in your approach



Credit is not like money

Failing to give credit to others
can kill your paper

If you imply that an idea is yours, and the referee knows it is not, then either

- You don't know that it's an old idea (bad)
- You do know, but are pretending it's yours (very bad)



Structure

- Abstract (4 sentences)
- Introduction (1 page)
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- Related work (1-2 pages)
- **Conclusions and further work** (0.5 pages)



Conclusions and further work

- Be brief.



The process of writing



The process

- Start early. Very early.
 - Hastily-written papers get rejected.
 - Papers are like wine: they need time to mature
- Collaborate
- Use *CVS* to support collaboration



Getting help

Get your paper read by as many friendly guinea pigs as possible

- Experts are good
- Non-experts are also very good
- Each reader can only read your paper for the first time once! So use them carefully
- Explain carefully what you want ("I got lost here" is much more important than "Jarva is mis-spelt".)



Getting expert help

- A good plan: when you think you are done, send the draft to the competition saying "could you help me ensure that I describe your work fairly?".
- Often they will respond with helpful critique (they are interested in the area)
- They are likely to be your referees anyway, so getting their comments or criticism up front is Jolly Good.



Listening to your reviewers

Treat every review like gold dust
Be (truly) grateful for criticism as
well as praise

This is **really, really, really** hard

But it's
really, really, really, really, really, really,
really, really, really, really
important



Listening to your reviewers

- Read every criticism as a positive suggestion for something you could explain more clearly
- DO NOT respond "you stupid person, I meant X". Fix the paper so that X is apparent even to the stupidest reader.
- Thank them warmly. They have given up their time for you.



Language and style



Basic stuff

- Submit by the deadline
- Keep to the length restrictions
 - Do not narrow the margins
 - Do not use 6pt font
 - On occasion, supply supporting evidence (e.g. experimental data, or a written-out proof) in an appendix
- Always use a spell checker



Visual structure

- Give strong visual structure to your paper using
 - sections and sub-sections
 - bullets
 - italics
 - laid-out code
- Find out how to draw pictures, and use them

Visual structure

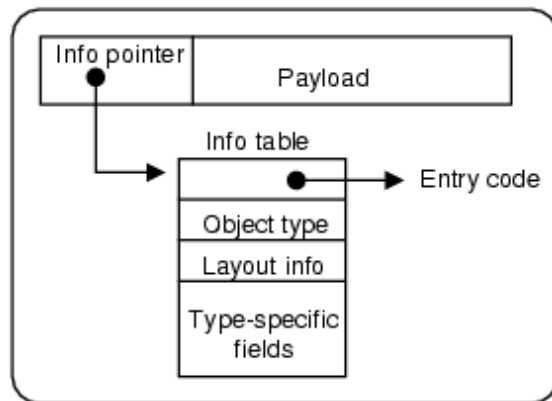


Figure 3. A heap object

The three cases above do not exhaust the possible forms of f . It might also be a *THUNK*, but we have already dealt with that case (rule *THUNK*). It might be a *CON*, in which case there cannot be any pending arguments on the stack, and rules *UPDATE* or *RET* apply.

4.3 The eval/apply model

The last block of Figure 2 shows how the eval/apply model deals with function application. The first three rules all deal with the case of a *FUN* applied to some arguments:

- If there are exactly the right number of arguments, we behave exactly like rule *KNOWNCALL*, by tail-calling the function. Rule *EXACT* is still necessary — and indeed has a direct counterpart in the implementation — because the function might not be statically known.
- If there are too many arguments, rule *CALLK* pushes a *call*

remainder of the object is called the *payload*, and may consist of a mixture of pointers and non-pointers. For example, the object $CON(C a_1 \dots a_n)$ would be represented by an object whose info pointer represented the constructor C and whose payload is the arguments $a_1 \dots a_n$.

The info table contains:

- Executable code for the object. For example, a *FUN* object has code for the function body.
- An object-type field, which distinguishes the various kinds of objects (*FUN*, *PAP*, *CON* etc) from each other.
- Layout information for garbage collection purposes, which describes the size and layout of the payload. By “layout” we mean which fields contain pointers and which contain non-pointers, information that is essential for accurate garbage collection.
- Type-specific information, which varies depending on the object type. For example, a *FUN* object contains its arity; a *CON* object contains its constructor tag, a small integer that distinguishes the different constructors of a data type; and so on.

In the case of a *PAP*, the size of the object is not fixed by its info table; instead, its size is stored in the object itself. The layout of its fields (e.g. which are pointers) is described by the (initial segment of) an argument-descriptor field in the info table of the *FUN* object which is always the first field of a *PAP*. The other kinds of heap object all have a size that is statically fixed by their info table.

A very common operation is to jump to the entry code for the object, so GHC uses a slightly-optimised version of the representation in Figure 3. GHC places the info table at the addresses *immediately*



Use simple, direct language

NO

The object under study was displaced horizontally

On an annual basis

Endeavour to ascertain

It could be considered that the speed of storage reclamation left something to be desired

YES

The ball moved sideways

Yearly

Find out

The garbage collector was really slow



Summary

If you remember nothing else:

- Identify your key idea
- Make your contributions explicit
- Use examples
- Make sure your experiments are reproducible

<http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/mleone/web/how-to.html>
A good starting point:
"Advice on Research and Writing"



Use the active voice

The passive voice is "respectable" but it DEADENS your paper. Avoid it at all costs.

NO

It can be seen that...

34 tests were run

These properties were
thought desirable

It might be thought that
this would be a type error

YES

We can see that...

We ran 34 tests

We wanted to retain these
properties

You might think this would
be a type error

"We" = you
and the
reader

"We" = the
authors

"You" = the
reader