Lazy Evaluation and Profiling

Advanced Haskell

Andres Löh

10–11 October 2013 — Copyright © 2013 Well-Typed LLP

Well-Typed
The Haskell Consultants

The plan

We want to get a better idea about what gets evaluated when:

- This is relevant once performance or space use are unexpectedly bad.
- It is also quite important for computations that should run in parallel.
- We will first look at different evaluation strategies.
- Then, we discuss how we can influence the default evaluation in Haskell when we are not happy with it.

Reduction

Reduction

A subexpression that can be reduced is called a redex.

Reduction

A subexpression that can be reduced is called a redex.

Most typical form of reduction in Haskell: replacing the left hand side of a function definition by a corresponding right hand side (this is essentially beta reduction from lambda calculus).

Reduction

A subexpression that can be reduced is called a redex.

Most typical form of reduction in Haskell: replacing the left hand side of a function definition by a corresponding right hand side (this is essentially beta reduction from lambda calculus).

Question

What if there are multiple redexes in one term?

Multiple redexes

Many terms have multiple redexes.

How many redexes are in the following term?

$\text{id}\ (\text{id}\ (\lambda z \to \text{id}\ z))$

Multiple redexes

Many terms have multiple redexes.

How many redexes are in the following term?

$$\text{id}\ (\text{id}\ (\lambda z \rightarrow \text{id}\ z))$$

Three redexes:

$$\begin{array}{l} \text{id}\ (\text{id}\ (\lambda z \rightarrow \text{id}\ z)) \\ \phantom{\text{id}\ }(\text{id}\ (\lambda z \rightarrow \text{id}\ z)) \\ \phantom{\text{id}\ (\text{id}\ (\lambda z \rightarrow }\text{id}\ z \end{array}$$

Multiple redexes

Many terms have multiple redexes.

How many redexes are in the following term?

$\mathrm{id}\ (\mathrm{id}\ (\lambda z \rightarrow \mathrm{id}\ z))$

$(\lambda x \rightarrow \lambda y \rightarrow x * x)\ (1 + 2)\ (3 + 4)$

Three redexes:

$\mathrm{id}\ (\mathrm{id}\ (\lambda z \rightarrow \mathrm{id}\ z))$

$(\mathrm{id}\ (\lambda z \rightarrow \mathrm{id}\ z))$

$\mathrm{id}\ z$

Multiple redexes

Many terms have multiple redexes.

How many redexes are in the following term?

$$\mathsf{id}\ (\mathsf{id}\ (\lambda z \to \mathsf{id}\ z))$$

$$(\lambda x \to \lambda y \to x * x)\ (1 + 2)\ (3 + 4)$$

Three redexes:

$$\mathsf{id}\ (\mathsf{id}\ (\lambda z \to \mathsf{id}\ z))$$
$$(\mathsf{id}\ (\lambda z \to \mathsf{id}\ z))$$
$$\mathsf{id}\ z$$

Three as well:

$$(\lambda x \to \lambda y \to x * x)\ (1 + 2)$$
$$(1 + 2)$$
$$(3 + 4)$$

Example

Let us play through the possible reductions for the following terms:

```
head (repeat 1)
```

Example

Let us play through the possible reductions for the following terms:

```
head (repeat 1)
```

```
let minimum xs = head (sort xs)
in minimum [4, 1, 3]
```

Evaluation strategies

Haskell's lazy evaluation

In Haskell,

- expressions are only evaluated if actually required,
- the leftmost outermost redex is chosen to achieve this,
- sharing is introduced (whenever an identifier is bound to an expression) in order to prevent evaluating expressions multiple times.

Haskell's lazy evaluation

In Haskell,

- expressions are only evaluated if actually required,
- the leftmost outermost redex is chosen to achieve this,
- sharing is introduced (whenever an identifier is bound to an expression) in order to prevent evaluating expressions multiple times.

If no redexes are left, an expression is in normal form. If the top-level of an expression is a constructor or lambda, then the expression is in (weak) head normal form.

Common evaluation strategies

Call by value / eager (strict) evaluation

Most common. Arguments are reduced as far as possible before reducing a function application, usually left-to-right.

Common evaluation strategies

Call by value / eager (strict) evaluation

Most common. Arguments are reduced as far as possible before reducing a function application, usually left-to-right.

Call by name

Functions are reduced before their arguments. Used by some macro languages (TeX, for instance).

Common evaluation strategies (contd.)

Call by need / lazy evaluation

Optimized version of “Call by name”: function arguments are only reduced when needed, but shared if used multiple times.

λf g x $\rightarrow$ combine (f x) (g x)

Church-Rosser

Theorem (Church-Rosser)

If a term e can be reduced to e_1 and e_2, there is a term e_3 such that both e_1 and e_2 can be reduced to e_3.

Church-Rosser

Theorem (Church-Rosser)

If a term e can be reduced to e_1 and e_2, there is a term e_3 such that both e_1 and e_2 can be reduced to e_3.

Corollary

Each term has at most one normal form.

Church-Rosser

Theorem (Church-Rosser)

If a term e can be reduced to e_1 and e_2, there is a term e_3 such that both e_1 and e_2 can be reduced to e_3.

Corollary

Each term has at most one normal form.

Theorem

If a term has a normal form, then lazy evaluation arrives at this normal form.

Example: the first 100 odd square numbers

$$
\begin{aligned}
&\text{example} :: [\text{Int}] \\
&\text{example} = \qquad\qquad\qquad\qquad [1\,..]
\end{aligned}
$$

We start by generating all numbers (lazy evaluation in action).

Example: the first 100 odd square numbers

$$
\begin{aligned}
&\mathsf{example} :: [\mathsf{Int}] \\
&\mathsf{example} = \quad \mathsf{map}\ (\lambda \mathsf{x} \rightarrow \mathsf{x} * \mathsf{x})\ \ [1\,..]
\end{aligned}
$$

We use map to compute the square numbers.

Example: the first 100 odd square numbers

$$\begin{aligned}&\mathsf{example} :: [\mathsf{Int}]\\&\mathsf{example} = (\qquad\qquad\quad \mathsf{filter\ odd} \circ \mathsf{map}\ (\lambda \mathsf{x} \rightarrow \mathsf{x} * \mathsf{x}))\ [1\,..]\end{aligned}$$

We use function composition composition (and partial application) to subsequently filter the odd square numbers.

Example: the first 100 odd square numbers

$$\text{example} :: [\text{Int}]$$
$$\text{example} = (\text{take}\ 100 \circ \text{filter odd} \circ \text{map}\ (\lambda x \rightarrow x * x))\ [1\,..]$$

Finally, we use composition again to take the first 100 elements of this list.

What drives the evaluation?

If we type an expression in at the GHCi prompt:

- GHCi wants to print its result,
- and for printing, we need that expression in normal form,
- that then demands other expressions to be evaluated.

Similarly for a complete program.

What drives the evaluation?

If we type an expression in at the GHCi prompt:

- GHCi wants to print its result,
- and for printing, we need that expression in normal form,
- that then demands other expressions to be evaluated.

Similarly for a complete program.

Within a function, it is most often pattern matching that drives the evaluation:

- in order to produce part of the output, we have to select a case;
- in order to be able to choose a case, we have to evaluate some of the arguments just far enough.

What drives the evaluation?

If we type an expression in at the GHCi prompt:

- GHCi wants to print its result,
- and for printing, we need that expression in normal form,
- that then demands other expressions to be evaluated.

Similarly for a complete program.

Within a function, it is most often pattern matching that drives the evaluation:

- in order to produce part of the output, we have to select a case;
- in order to be able to choose a case, we have to evaluate some of the arguments just far enough.

Evaluating a term to weak head normal form (WHNF) reveals its outermost constructor and allows us to potentially make a choice in a pattern match.

Tracking demand

If we want to understand what gets evaluated when, we have to track the demand, and this usually happens backwards.

Typical question

If we need the result of a function to be evaluated to WHNF, what effect (if any) does this have on the argument(s) of the function?

Let us look at a few examples.

Tracking demand (contd.)

$\mathsf{id}\ \mathsf{x} = \mathsf{x}$

Demanding the result causes demanding the argument.

Tracking demand (contd.)

$\text{id}\ x = x$

Demanding the result causes demanding the argument.

$\text{const}\ x\ y = x$

Demanding the result causes demanding the first argument (but not the second).

Tracking demand (contd.)

```
id x = x
```

Demanding the result causes demanding the argument.

```
const x y = x
```

Demanding the result causes demanding the first argument (but not the second).

```
True  || _ = True
False || y = y
```

Demanding the result causes demanding the first argument, and depending on that might cause demanding the second.

Tracking demand (contd.)

$$\begin{aligned} &\mathsf{map}\ f\ [\,] &&= [\,] \\ &\mathsf{map}\ f\ (x : xs) &&= f\ x : \mathsf{map}\ f\ xs \end{aligned}$$

Demanding the result causes demanding the second argument (but not the first).

Devising a test for demand propagation

Do we have to look at the code in order to track demand?

Devising a test for demand propagation

Do we have to look at the code in order to track demand?

Idea:

- Feed a non-terminating computation into a function and demand its result.

Devising a test for demand propagation

Do we have to look at the code in order to track demand?

Idea:

- Feed a non-terminating computation into a function and demand its result.
- If the function produces a result, then it cannot have demanded its argument.

Devising a test for demand propagation

Do we have to look at the code in order to track demand?

Idea:

- Feed a non-terminating computation into a function and demand its result.
- If the function produces a result, then it cannot have demanded its argument.
- If the function loops, then we do not know whether we demanded the argument, or the loop arises from elsewhere – but (assuming we identify all loops and run-time exceptions), it is safe to assume that it did, as it would not have changed the result.

Strict functions

Definition

A (one-argument) function f is called strict if and only if

$$f\ \bot = \bot$$

Here, $\bot$ denotes any crashing or looping computation.

Strict functions

Definition

A (one-argument) function `f` is called `strict` if and only if

`f` $\perp = \perp$

Here, $\perp$ denotes any crashing or looping computation.

Note

In a strict language, all functions are strict.

In a non-strict language, such as Haskell, we have both strict and non-strict functions.

Lazy evaluation quiz

$(\lambda x \rightarrow x)\ \mathsf{True} \quad \rightsquigarrow^*$

$(\lambda x \rightarrow x)\ \bot \quad \rightsquigarrow^*$

$(\lambda x \rightarrow ())\ \bot \quad \rightsquigarrow^*$

$(\lambda x \rightarrow \bot)\ () \quad \rightsquigarrow^*$

$(\lambda x\ f \rightarrow f\ x)\ \bot \quad \rightsquigarrow^*$

$\mathsf{length}\ (\mathsf{map}\ \bot\ [1, 2]) \quad \rightsquigarrow^*$

Lazy evaluation quiz

$$
\begin{array}{lll}
(\lambda x \to x)\ \mathsf{True} & \rightsquigarrow^{*} & \mathsf{True} \\
(\lambda x \to x)\ \bot & \rightsquigarrow^{*} & \\
(\lambda x \to ())\ \bot & \rightsquigarrow^{*} & \\
(\lambda x \to \bot)\ () & \rightsquigarrow^{*} & \\
(\lambda x\ f \to f\ x)\ \bot & \rightsquigarrow^{*} & \\
\mathsf{length}\ (\mathsf{map}\ \bot\ [1, 2]) & \rightsquigarrow^{*} &
\end{array}
$$

Lazy evaluation quiz

$$
\begin{array}{lll}
(\lambda x \rightarrow x)\ \mathsf{True} & \leadsto^{*} & \mathsf{True} \\
(\lambda x \rightarrow x)\ \bot & \leadsto^{*} & \bot \\
(\lambda x \rightarrow ())\ \bot & \leadsto^{*} & \\
(\lambda x \rightarrow \bot)\ () & \leadsto^{*} & \\
(\lambda x\ f \rightarrow f\ x)\ \bot & \leadsto^{*} & \\
\mathsf{length}\ (\mathsf{map}\ \bot\ [1, 2]) & \leadsto^{*} &
\end{array}
$$

Lazy evaluation quiz

$$
\begin{array}{lll}
(\lambda x \rightarrow x)\ \mathsf{True} & \leadsto^* & \mathsf{True} \\
(\lambda x \rightarrow x)\ \bot & \leadsto^* & \bot \\
(\lambda x \rightarrow ())\ \bot & \leadsto^* & () \\
(\lambda x \rightarrow \bot)\ () & \leadsto^* & \\
(\lambda x\ f \rightarrow f\ x)\ \bot & \leadsto^* & \\
\mathsf{length}\ (\mathsf{map}\ \bot\ [1,2]) & \leadsto^* &
\end{array}
$$

Lazy evaluation quiz

$$
\begin{array}{lll}
(\lambda x \rightarrow x)\ \mathsf{True} & \leadsto^* & \mathsf{True} \\
(\lambda x \rightarrow x)\ \bot & \leadsto^* & \bot \\
(\lambda x \rightarrow ())\ \bot & \leadsto^* & () \\
(\lambda x \rightarrow \bot)\ () & \leadsto^* & \bot \\
(\lambda x\ f \rightarrow f\ x)\ \bot & \leadsto^* & \\
\mathsf{length}\ (\mathsf{map}\ \bot\ [1, 2]) & \leadsto^* &
\end{array}
$$

Lazy evaluation quiz

$$
\begin{array}{lll}
(\lambda x \to x)\ \mathsf{True} & \leadsto^* & \mathsf{True} \\
(\lambda x \to x)\ \bot & \leadsto^* & \bot \\
(\lambda x \to ())\ \bot & \leadsto^* & () \\
(\lambda x \to \bot)\ () & \leadsto^* & \bot \\
(\lambda x\ f \to f\ x)\ \bot & \leadsto^* & \lambda f \to f\ \bot \\
\mathsf{length}\ (\mathsf{map}\ \bot\ [1, 2]) & \leadsto^* &
\end{array}
$$

Lazy evaluation quiz

$(\lambda x \to x)\ \mathsf{True}$	$\leadsto^*$	True
$(\lambda x \to x)\ \bot$	$\leadsto^*$	$\bot$
$(\lambda x \to ())\ \bot$	$\leadsto^*$	$()$
$(\lambda x \to \bot)\ ()$	$\leadsto^*$	$\bot$
$(\lambda x\ f \to f\ x)\ \bot$	$\leadsto^*$	$\lambda f \to f\ \bot$
$\mathsf{length}\ (\mathsf{map}\ \bot\ [1, 2])$	$\leadsto^*$	2

Observing evaluation

For debugging purposes, you can observe when something gets evaluated using `trace` from `Debug.Trace`:

```
trace     :: String → a → a
traceShow :: Show a ⇒ a → b → b
```

The `trace` functions print their first argument as soon as the second is being evaluated.

Observing evaluation

For debugging purposes, you can observe when something gets evaluated using `trace` from `Debug.Trace`:

```
trace     :: String → a → a
traceShow :: Show a ⇒ a → b → b
```

The `trace` functions print their first argument as soon as the second is being evaluated.

Question: How many symbols are printed, in what order?

```
x .: xs = trace ":" (x : xs)
nil     = trace "." []
num x   = trace "0" x
list    = foldr (.:) nil (map num [1 .. 10])
main    = print $ sum (take 2 (drop 3 list))
```

Space leaks and profiling

Haskell data in memory

As we've discussed earlier:

- nearly all Haskell data lives on the heap,
- nearly all Haskell data is immutable,
- operations do not change data but rather create new data on the heap,
- a lot of data is shared.

Haskell data in memory

As we've discussed earlier:

- nearly all Haskell data lives on the heap,
- nearly all Haskell data is immutable,
- operations do not change data but rather create new data on the heap,
- a lot of data is shared.

Sharing is easy because everything is immutable.

Laziness on the heap

Bindings are not evaluated immediately:

- Instead, suspended computations (called thunks) are created on the heap.
- Thunks can be shared just as other subterms.
- If a thunk is required, it is evaluated and destructively updated on the heap.
- However, this is a safe and even desirable update – we don't change the value stored, we just change its representation.
- Other computations sharing the updated thunk won't have to recompute the expression.

Garbage collection

GHC uses a generational garbage collector:

- Optimized for lots of short-lived data, as is common in a purely functional language.
- New data is allocated in the “young” generation.
- The young generation is rather small and collected often.
- After a while, data that is still alive is moved to the “old” generation.
- The old generation is larger and collected rarely.
- The heap of a Haskell program can grow dynamically if more memory is needed.

The lifetime of data

Data is alive as long as there are references to it.

The lifetime of data

Data is alive as long as there are references to it.

In a lazy setting, it is sometimes hard to predict how long we retain references to data.

The lifetime of data

Data is alive as long as there are references to it.

In a lazy setting, it is sometimes hard to predict how long we retain references to data.

Space leak

A data structure which grows bigger, or lives longer, than we expect.

As space is a limited resource, we might run (nearly) out of it. Consequences:

- more garbage collections cost extra time,
- swapping,
- program might get killed.

Computing a large sum

$$\begin{aligned} \mathrm{sum}_1\ [\,] &= 0 \\ \mathrm{sum}_1\ (x : xs) &= x + \mathrm{sum}_1\ xs \end{aligned}$$

- A straight-forward definition, following the standard pattern of defining functions on lists.
- What is the problem?

Computing a large sum

$$
\begin{aligned}
\mathrm{sum}_1\ [\,] &= 0 \\
\mathrm{sum}_1\ (x : xs) &= x + \mathrm{sum}_1\ xs
\end{aligned}
$$

- A straight-forward definition, following the standard pattern of defining functions on lists.
- What is the problem?
- If we try to evaluate this function for larger and larger input lists, we note that it takes more and more memory, and significant amounts of time, or we get an error indicating it runs out of stack space.
- But certainly we should be able to sum a list in (nearly) constant (stack) space? What is going on?

Obtaining more information

Haskell's run-time system (RTS) can be instructed to spit out additional information:

- RTS options can be passed to Haskell binaries on the command line by placing them after `+RTS` or enclosing them between `+RTS` and `-RTS`.
- Many RTS flags require the binary to be compiled (or rather linked) using the `-rtsopts` GHC flag.
- You can obtain info about available RTS flags by invoking a compiled binary with `+RTS --help`.
- Very interesting are GC statistics (available in various amounts of detail via `-t`, `-s` or `-S`).
- You can increase the stack space by saying something like `-K50M` or `-K500M`.

GC statistics

```
$ ./Sum1 10000000 +RTS -s -K500M
50000005000000
   1,532,401,936 bytes allocated in the heap
     788,992,048 bytes copied during GC
     457,301,152 bytes maximum residency (10 sample(s))
         740,216 bytes maximum slop
             633 MB total memory in use (0 MB lost due to fragmentation)

                                    Tot time (elapsed)  Avg pause  Max pause
  Gen  0      2299 colls,     0 par    0.83s    0.83s     0.0004s    0.0008s
  Gen  1        10 colls,     0 par    0.60s    0.60s     0.0602s    0.2877s

  INIT    time    0.00s  (  0.00s elapsed)
  MUT     time    0.46s  (  0.46s elapsed)
  GC      time    1.43s  (  1.43s elapsed)
  EXIT    time    0.00s  (  0.00s elapsed)
  Total   time    1.89s  (  1.88s elapsed)

  %GC     time      75.8%  (75.8% elapsed)

  Alloc rate    3,352,283,510 bytes per MUT second

  Productivity  24.2% of total user, 24.3% of total elapsed
```

MUT (mutator) time is good, GC time is bad.

Maximum residency and percentage of GC time are revealing.

Heap profiling

More detailed information can be obtained using heap profiling.

- Requires recompilation of the program (makes program larger and overall slower).
- All used libraries must have profiling versions, too.
- In your cabal-install `config` file, put

```
library-profiling: True
```

  for the future.
- Compile a program with profiling enabled:

```
$ ghc --make -prof -auto-all -rtsopts Sum1
```

  The `-auto-all` is optional. It is more important for larger programs where you not only want to know how much space is being used, but also where it is being used.

Heap profiling – contd.

- Run with profiling enabled:

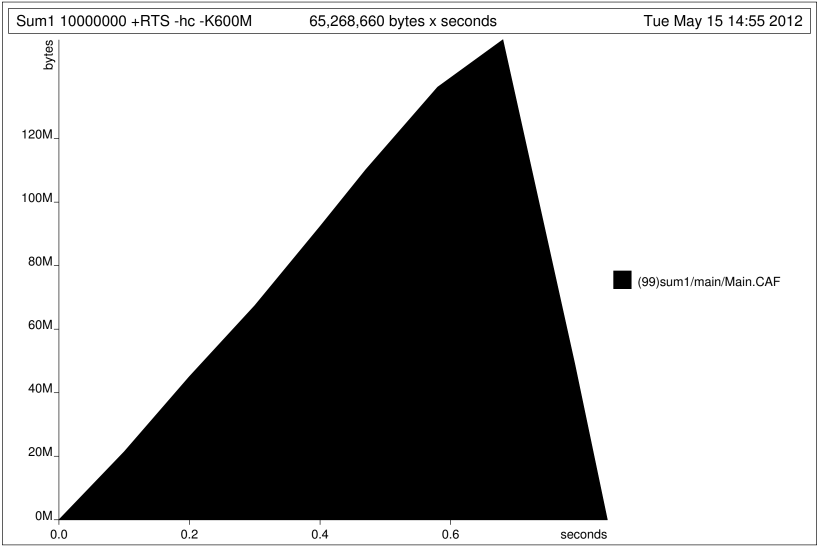
```
$ ./Sum1 10000000 +RTS -K800M -hc
```

  Again, there are many different `-h` flags.

- The `-hc` is for cost-center profiling.
- A very simplistic form of heap profiling via just `-h` is available even without compiling the program for profiling. It would be sufficient here!
- Files `Sum1.prof` and `Sum1.hp` are produced.
- The `.hp` file can be transformed into PostScript format using the `hp2ps` tool.

```
$ hp2ps Sum1.hp
```

Heap profile for sum$_1$



The problem

$$
\begin{aligned}
& \mathrm{sum}_1\ [1,2,3,4,\ldots] \\
\equiv\ & \quad \{\text{ Definition of } \mathrm{sum}_1 \ \} \\
& 1 + \mathrm{sum}_1\ [2,3,4,\ldots] \\
\equiv\ & \quad \{\text{ Definition of } \mathrm{sum}_1 \ \} \\
& 1 + (2 + \mathrm{sum}_1\ [3,4,\ldots]) \\
\equiv\ & \quad \{\text{ Definition of } \mathrm{sum}_1 \ \} \\
& 1 + (2 + (3 + \mathrm{sum}_1\ [4,\ldots])) \\
\equiv\ & \\
& \ldots
\end{aligned}
$$

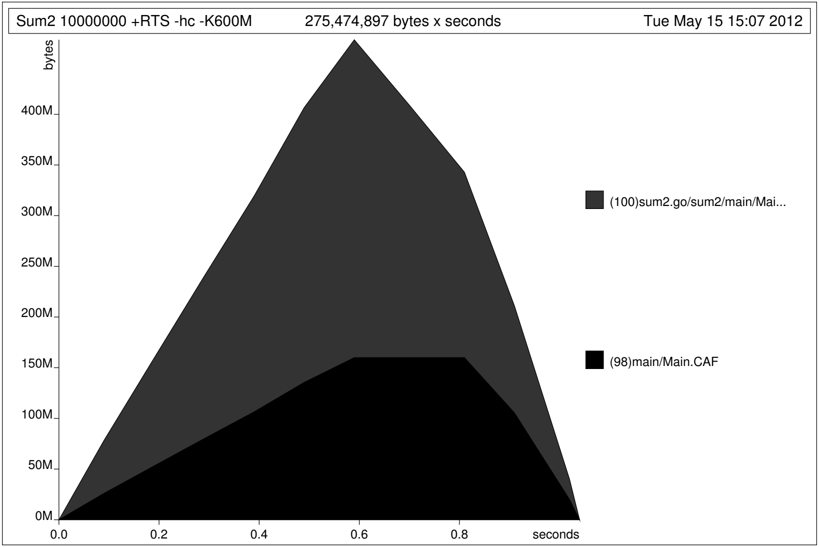
The whole recursion has to be unfolded before the first addition can be reduced!

Attempting a tail-recursive version

$$
\begin{aligned}
&\mathrm{sum}_2\ xs = go\ 0\ xs \\
&\quad \textbf{where} \\
&\qquad go\ acc\ [\,] \qquad\quad = acc \\
&\qquad go\ acc\ (x : xs) = go\ (acc + x)\ xs
\end{aligned}
$$

We hope that tail-recursion improves stack usage, and might thereby improve space behaviour as well, but …

Heap profile for sum_2



The new problem

$$
\begin{array}{ll}
 & \mathit{sum}_2\ [1, 2, 3, 4, \ldots] \\
\equiv & \quad \{\ \text{Definition of } \mathit{sum}_2\ \} \\
 & \mathit{sum}'_2\ 0\ [1, 2, 3, 4, \ldots] \\
\equiv & \quad \{\ \text{Definition of } \mathit{sum}_2\ \} \\
 & \mathit{sum}'_2\ (0 + 1)\ [2, 3, 4, \ldots] \\
\equiv & \quad \{\ \text{Definition of } \mathit{sum}_2\ \} \\
 & \mathit{sum}'_2\ ((0 + 1) + 2)\ [3, 4, \ldots] \\
 & \ldots \\
\equiv & \\
 & \mathit{sum}'_2\ (\ldots((0 + 1) + 2)\ldots)\ [\,] \\
\equiv & \quad \{\ \text{Definition of } \mathit{sum}_2\ \} \\
 & (\ldots(0 + 1) + 2)\ldots)
\end{array}
$$

We still build up the whole addition, but now in an accumulating argument! Evaluating that still takes stack!

Controlling evaluation

We need more control

Sometimes, we want to make things stricter than they are by default. Here:

- we have a computation that will be evaluated anyway,
- storing it in delayed form costs much more space than storing its result.

Forcing evaluation

Haskell has the following primitive function

```
seq :: a → b → b    -- primitive
```

The call `seq x y` is strict in `x` and returns `y`.

Forcing evaluation

Haskell has the following primitive function

```
seq :: a → b → b    -- primitive
```

The call `seq x y` is strict in `x` and returns `y`.

The function `seq` can be used to define strict function application:

```
($!) :: (a → b) → a → b
f $! x = x `seq` f x
```

Recall sharing!

Forcing quiz

The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

Forcing quiz

The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$(\lambda x \to ())\ \$!\ \bot \qquad \rightsquigarrow^*$

$\mathrm{seq}\ (\bot, \bot)\ () \qquad \rightsquigarrow^*$

$\mathrm{snd}\ \$!\ (\bot, 1) \qquad \rightsquigarrow^*$

$(\lambda x \to ())\ \$!\ (\lambda x \to \bot) \qquad \rightsquigarrow^*$

$\mathrm{length}\ \$!\ \mathrm{map}\ \bot\ [1, 2] \qquad \rightsquigarrow^*$

$\mathrm{seq}\ (\bot + 1)\ () \qquad \rightsquigarrow^*$

$\mathrm{seq}\ (1 : \bot)\ () \qquad \rightsquigarrow^*$

Forcing quiz

The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$$(\lambda x \rightarrow ())\ \$!\ \bot \quad \rightsquigarrow^* \quad \bot$$

$$\mathsf{seq}\ (\bot, \bot)\ () \quad \rightsquigarrow^*$$

$$\mathsf{snd}\ \$!\ (\bot, 1) \quad \rightsquigarrow^*$$

$$(\lambda x \rightarrow ())\ \$!\ (\lambda x \rightarrow \bot) \quad \rightsquigarrow^*$$

$$\mathsf{length}\ \$!\ \mathsf{map}\ \bot\ [1, 2] \quad \rightsquigarrow^*$$

$$\mathsf{seq}\ (\bot + 1)\ () \quad \rightsquigarrow^*$$

$$\mathsf{seq}\ (1 : \bot)\ () \quad \rightsquigarrow^*$$

Forcing quiz

The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$(\lambda x \to ())\ \$!\ \bot$	$\leadsto^*$	$\bot$
$\mathsf{seq}\ (\bot, \bot)\ ()$	$\leadsto^*$	$()$
$\mathsf{snd}\ \$!\ (\bot, 1)$	$\leadsto^*$	
$(\lambda x \to ())\ \$!\ (\lambda x \to \bot)$	$\leadsto^*$	
$\mathsf{length}\ \$!\ \mathsf{map}\ \bot\ [1, 2]$	$\leadsto^*$	
$\mathsf{seq}\ (\bot + 1)\ ()$	$\leadsto^*$	
$\mathsf{seq}\ (1 : \bot)\ ()$	$\leadsto^*$	

Forcing quiz

The function seq only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$(\lambda x \rightarrow ())\ \$!\ \bot$	$\leadsto^*$	$\bot$
$\mathsf{seq}\ (\bot, \bot)\ ()$	$\leadsto^*$	$()$
$\mathsf{snd}\ \$!\ (\bot, 1)$	$\leadsto^*$	1
$(\lambda x \rightarrow ())\ \$!\ (\lambda x \rightarrow \bot)$	$\leadsto^*$	
$\mathsf{length}\ \$!\ \mathsf{map}\ \bot\ [1, 2]$	$\leadsto^*$	
$\mathsf{seq}\ (\bot + 1)\ ()$	$\leadsto^*$	
$\mathsf{seq}\ (1 : \bot)\ ()$	$\leadsto^*$	

Forcing quiz

The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$(\lambda x \to ())\ \$!\ \bot$	$\leadsto^*\ \bot$
$\mathrm{seq}\ (\bot, \bot)\ ()$	$\leadsto^*\ ()$
$\mathrm{snd}\ \$!\ (\bot, 1)$	$\leadsto^*\ 1$
$(\lambda x \to ())\ \$!\ (\lambda x \to \bot)$	$\leadsto^*\ ()$
$\mathrm{length}\ \$!\ \mathrm{map}\ \bot\ [1, 2]$	$\leadsto^*$
$\mathrm{seq}\ (\bot + 1)\ ()$	$\leadsto^*$
$\mathrm{seq}\ (1 : \bot)\ ()$	$\leadsto^*$

Forcing quiz

The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$(\lambda x \rightarrow ())\ \$!\ \bot$	$\leadsto^*$	$\bot$
$\mathsf{seq}\ (\bot, \bot)\ ()$	$\leadsto^*$	$()$
$\mathsf{snd}\ \$!\ (\bot, 1)$	$\leadsto^*$	1
$(\lambda x \rightarrow ())\ \$!\ (\lambda x \rightarrow \bot)$	$\leadsto^*$	$()$
$\mathsf{length}\ \$!\ \mathsf{map}\ \bot\ [1, 2]$	$\leadsto^*$	2
$\mathsf{seq}\ (\bot + 1)\ ()$	$\leadsto^*$	
$\mathsf{seq}\ (1 : \bot)\ ()$	$\leadsto^*$	

Forcing quiz

The function seq only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$$(\lambda x \rightarrow ())\ \$!\ \bot \quad \leadsto^* \quad \bot$$

$$\mathrm{seq}\ (\bot, \bot)\ () \quad \leadsto^* \quad ()$$

$$\mathrm{snd}\ \$!\ (\bot, 1) \quad \leadsto^* \quad 1$$

$$(\lambda x \rightarrow ())\ \$!\ (\lambda x \rightarrow \bot) \quad \leadsto^* \quad ()$$

$$\mathrm{length}\ \$!\ \mathrm{map}\ \bot\ [1, 2] \quad \leadsto^* \quad 2$$

$$\mathrm{seq}\ (\bot + 1)\ () \quad \leadsto^* \quad \bot$$

$$\mathrm{seq}\ (1 : \bot)\ () \quad \leadsto^*$$

Forcing quiz

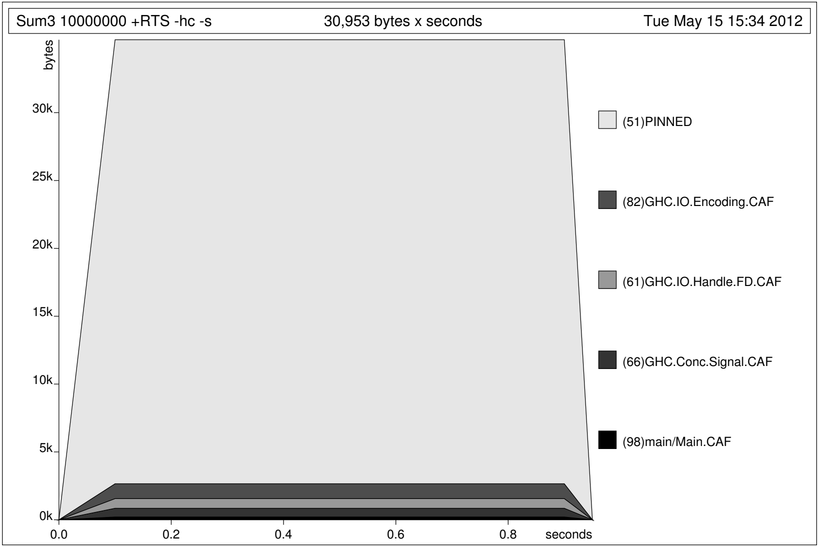
The function `seq` only evaluates to WHNF (i.e., a lambda abstraction, literal or constructor application).

$$
\begin{array}{ll}
(\lambda x \to ())\ \$!\ \bot & \leadsto^{*}\ \bot \\
\mathsf{seq}\ (\bot, \bot)\ () & \leadsto^{*}\ () \\
\mathsf{snd}\ \$!\ (\bot, 1) & \leadsto^{*}\ 1 \\
(\lambda x \to ())\ \$!\ (\lambda x \to \bot) & \leadsto^{*}\ () \\
\mathsf{length}\ \$!\ \mathsf{map}\ \bot\ [1, 2] & \leadsto^{*}\ 2 \\
\mathsf{seq}\ (\bot + 1)\ () & \leadsto^{*}\ \bot \\
\mathsf{seq}\ (1 : \bot)\ () & \leadsto^{*}\ ()
\end{array}
$$

Using seq to force the addition

$$
\begin{aligned}
&\mathrm{sum}_3\ \mathrm{xs} = \mathrm{go}\ 0\ \mathrm{xs}\\
&\quad \textbf{where}\\
&\qquad \mathrm{go\ acc}\ [\,] \qquad\quad = \mathrm{acc}\\
&\qquad \mathrm{go\ acc}\ (\mathrm{x} : \mathrm{xs}) = (\mathrm{go}\ \$!\ \mathrm{acc} + \mathrm{x})\ \mathrm{xs}
\end{aligned}
$$

Heap profile for sum$_3$



GC statistics

```
$ ./Sum3 10000000 +RTS -hc -s
50000005000000
   2,560,118,208 bytes allocated in the heap
         714,144 bytes copied during GC
          62,104 bytes maximum residency (10 sample(s))
          26,344 bytes maximum slop
               1 MB total memory in use (0 MB lost due to fragmentation)

                                    Tot time (elapsed)  Avg pause  Max pause
  Gen  0      4873 colls,     0 par    0.02s    0.02s     0.0000s    0.0000s
  Gen  1        10 colls,     0 par    0.00s    0.00s     0.0001s    0.0001s

  INIT    time    0.00s  (  0.00s elapsed)
  MUT     time    0.95s  (  0.95s elapsed)
  GC      time    0.02s  (  0.02s elapsed)
  RP      time    0.00s  (  0.00s elapsed)
  PROF    time    0.00s  (  0.00s elapsed)
  EXIT    time    0.00s  (  0.00s elapsed)
  Total   time    0.98s  (  0.98s elapsed)

  %GC     time       2.3%  (2.2% elapsed)

  Alloc rate    2,684,947,785 bytes per MUT second

  Productivity  97.6% of total user, 97.6% of total elapsed
```

Look at the maximum residency and GC time / productivity now.

Standard recursion patterns

The three versions of `sum` we have seen correspond to using `foldr`, `foldl` and `foldl'`, respectively:

$$\mathit{sum}_1 = \mathit{foldr}\ (+)\ 0$$
$$\mathit{sum}_2 = \mathit{foldl}\ (+)\ 0$$
$$\mathit{sum}_3 = \mathit{foldl}'\ (+)\ 0$$

Question

Is using foldl′/strictness always preferable?

Question

Is using `foldl'`/strictness always preferable?

For example, what about defining `map` ...

Rules of thumb

- If you expect partial results or want to use infinite lists, use `foldr`.

  Examples: `map`, `filter`.
- If the operator is strict, use `foldl'`.

  Examples: `sum`, `product`.
- Otherwise, use `foldl`.

  Examples: `reverse`.
- Use the GHC optimizer by passing `-O`. GHC performs strictness analysis to optimize your code – but don't rely on it to always figure out everything!

Strictness analysis

Let us look at our original tail-recursive version of `sum` again:

```
sum₂ :: [Int] → Int
sum₂ xs = go 0 xs
  where
    go acc []       = acc
    go acc (x : xs) = go (acc + x) xs
```

Let's perform strictness analysis!

Looking at the Core code

You can spit out GHC Core after optimization by saying `-ddump-simpl` – reading the result requires some experience, but can show what optimizations GHC performs.

The result of sum_2 without optimization:

```
SumStrict.sum :: [GHC.Types.Int] -> GHC.Types.Int
[GblId, Arity=1]
SumStrict.sum =
  \ (xs_aan :: [GHC.Types.Int]) ->
    letrec {
      go_aao [Occ=LoopBreaker]
        :: GHC.Types.Int -> [GHC.Types.Int] -> GHC.Types.Int
      [LclId, Arity=2]
      go_aao =
        \ (acc_aap :: GHC.Types.Int) (ds_daU :: [GHC.Types.Int]) ->
          case ds_daU of _ {
            [] -> acc_aap;
            : x_aar xs1_aas ->
              go_aao
                (GHC.Num.+ @ GHC.Types.Int GHC.Num.$fNumInt acc_aap x_aar) xs1_aas
          }; } in
    go_aao (GHC.Types.I# 0) xs_aan
```

Looking at the Core code (contd.)

Some observations about GHC Core:

- all names are fully qualified;
- symbolic names are escaped;
- nested pattern matches are compiled to simple pattern matches;
- everything is type-annotated;
- literals are unboxed, boxing is explicit;
- polymorphic functions have explicit type arguments (`@GHC.Types.Int`);
- overloaded functions are passed dictionaries (`GHC.Num.$fNumInt`).

Looking at the Core code (contd.)

Excerpts from the optimized version:

```
SumStrict.sum :: [GHC.Types.Int] -> GHC.Types.Int
SumStrict.sum =
  \ (xs_aan :: [GHC.Types.Int]) ->
    case SumStrict.$wgo 0 xs_aan of ww_sbN { __DEFAULT ->
    GHC.Types.I# ww_sbN
    }
```

Looking at the Core code (contd.)

Excerpts from the optimized version:

```
SumStrict.sum :: [GHC.Types.Int] -> GHC.Types.Int
SumStrict.sum =
  \ (xs_aan :: [GHC.Types.Int]) ->
    case SumStrict.$wgo 0 xs_aan of ww_sbN { __DEFAULT ->
    GHC.Types.I# ww_sbN
    }
```

```
SumStrict.$wgo [Occ=LoopBreaker]
  :: GHC.Prim.Int# -> [GHC.Types.Int] -> GHC.Prim.Int#
SumStrict.$wgo =
  \ (ww_sbI :: GHC.Prim.Int#) (w_sbK :: [GHC.Types.Int]) ->
    case w_sbK of _ {
      [] -> ww_sbI;
      : x_aar xs_aas ->
        case x_aar of _ { GHC.Types.I# y_abo ->
        SumStrict.$wgo (GHC.Prim.+# ww_sbI y_abo) xs_aas
        }
    }
```

Looking at the Core code (contd.)

Excerpts from the optimized version:

```
SumStrict.sum :: [GHC.Types.Int] -> GHC.Types.Int
SumStrict.sum =
  \ (xs_aan :: [GHC.Types.Int]) ->
    case SumStrict.$wgo 0 xs_aan of ww_sbN { __DEFAULT ->
    GHC.Types.I# ww_sbN
    }
```

```
SumStrict.$wgo [Occ=LoopBreaker]
  :: GHC.Prim.Int# -> [GHC.Types.Int] -> GHC.Prim.Int#
SumStrict.$wgo =
  \ (ww_sbI :: GHC.Prim.Int#) (w_sbK :: [GHC.Types.Int]) ->
    case w_sbK of _ {
      [] -> ww_sbI;
      : x_aar xs_aas ->
        case x_aar of _ { GHC.Types.I# y_abo ->
        SumStrict.$wgo (GHC.Prim.+# ww_sbI y_abo) xs_aas
        }
    }
```

Due to strictness analysis, GHC can actually unbox the `Int` accumulator.

More on controlling strictness

The type of seq

Question: Why is the type of seq

```
seq :: a → b → b
```

and not, for example,

```
seq' :: a → a
seq' x = seq x x    -- example definition
```

?

The type of `seq`

Question: Why is the type of `seq`

$$\mathsf{seq} :: a \to b \to b$$

and not, for example,

```
seq′ :: a → a
seq′ x = seq x x    -- example definition
```

?

Remember: Everything is demand driven!

In `seq x y`, the evaluation of `x` is tied to the demand for `y`.

In `seq x x`, the evaluation of `x` is tied to the demand for `x` – which it is anyway.

Bang patterns

```
sum₃ xs = go 0 xs
  where
    go acc  []      = acc
    go acc  (x : xs) = (go $! acc + x) xs
```

Instead of using seq or ($!), we can also make the pattern match evaluate more than it normally would . . .

Bang patterns

```
sum₄ xs = go 0 xs
  where
    go !acc []       = acc
    go !acc (x : xs) = go (acc + x) xs
```

…we can also use bang patterns.

A `!` in front of a variable pattern makes GHC evaluate the matched term to WHNF before continuing, even if this evaluation is not needed to make the decision for a particular case.

This requires the `BangPatterns` extension.

Irrefutable (lazy) patterns

A less known feature of Haskell is that it supports (without an extension) irrefutable patterns.

Consider this definition:

$$\mathsf{unzip} :: [(\mathsf{a},\mathsf{b})] \rightarrow ([\mathsf{a}],[\mathsf{b}])$$
$$\mathsf{unzip} = \mathsf{foldr}\ (\lambda(\mathsf{x},\mathsf{y})\ (\mathsf{xs},\mathsf{ys}) \rightarrow (\mathsf{x} : \mathsf{xs},\mathsf{y} : \mathsf{ys}))\ ([\,],[\,])$$

Irrefutable (lazy) patterns

A less known feature of Haskell is that it supports (without an extension) irrefutable patterns.

Consider this definition:

$$\mathsf{unzip} :: [(\mathsf{a}, \mathsf{b})] \to ([\mathsf{a}], [\mathsf{b}])$$
$$\mathsf{unzip} = \mathsf{foldr}\ (\lambda(\mathsf{x}, \mathsf{y})\ (\mathsf{xs}, \mathsf{ys}) \to (\mathsf{x} : \mathsf{xs}, \mathsf{y} : \mathsf{ys}))\ ([\,], [\,])$$

Now both of

$$\textbf{let}\ \mathsf{zs} = [1\,..]\quad \textbf{in}\ \mathsf{unzip}\ (\mathsf{zip}\ \mathsf{zs}\ \mathsf{zs})$$
$$\textbf{let}\ \mathsf{zs} = [1\,..\,10000000]\ \textbf{in}\ \mathsf{unzip}\ (\mathsf{zip}\ \mathsf{zs}\ \mathsf{zs})$$

result in a stack overflow. Why?

Unfolding the too strict unzip

$$
\begin{array}{ll}
 & \text{unzip}\ [(1,1),(2,2)] \\
\equiv & \quad \{\ \text{Definition of}\ \ \text{unzip}\ \ \} \\
 & \text{foldr}\ (\lambda(x,y)\ (xs,ys) \rightarrow (x:xs, y:ys))\ ([\,],[\,])\ [(1,1),(2,2)] \\
\equiv & \quad \{\ \text{Definition of}\ \ \text{foldr}\ \ \} \\
 & (\lambda(x,y)\ (xs,ys) \rightarrow (x:xs, y:ys))\ (1,1)\ (\text{foldr} \ldots [(2,2)]) \\
\equiv & \quad \{\ \text{Matching the first pair}\ \} \\
 & (\lambda(xs,ys) \rightarrow (1:xs, 1:ys))\ (\text{foldr} \ldots [(2,2)])
\end{array}
$$

At this point, we cannot reduce the function without first evaluation foldr further.

Introducing lazy patterns

A better definition is

$$\mathsf{unzip} :: [(a, b)] \rightarrow ([a], [b])$$
$$\mathsf{unzip} = \mathsf{foldr}\ (\lambda(x, y)\ {\sim}(xs, ys) \rightarrow (x : xs, y : ys))\ ([\,], [\,])$$

which in this case is equivalent to

$$\mathsf{unzip} :: [(a, b)] \rightarrow ([a], [b])$$
$$\mathsf{unzip} = \mathsf{foldr}\ (\lambda(x, y)\ xys \rightarrow (x : \mathsf{fst}\ xys, y : \mathsf{snd}\ xys))\ ([\,], [\,])$$

Introducing lazy patterns

A better definition is

$$\mathsf{unzip} :: [(a,b)] \rightarrow ([a],[b])$$
$$\mathsf{unzip} = \mathsf{foldr}\ (\lambda(x,y)\ {\sim}(xs,ys) \rightarrow (x : xs, y : ys))\ ([\,],[\,])$$

which in this case is equivalent to

$$\mathsf{unzip} :: [(a,b)] \rightarrow ([a],[b])$$
$$\mathsf{unzip} = \mathsf{foldr}\ (\lambda(x,y)\ xys \rightarrow (x : \mathsf{fst}\ xys, y : \mathsf{snd}\ xys))\ ([\,],[\,])$$

A $\sim$ in front of a pattern will make it match always (hence irrefutable). Only if the components of the pattern are demanded, they will be extracted.

Lazy **let** and **where**, strict **case**

The outermost pattern in a **let** or **where** is lazy by default:

```
unzip :: [(a, b)] → ([a], [b])
unzip []                 = ([], [])
unzip ((x, y) : xys) = (x : xs, y : ys)
  where
    (xs, ys) = unzip xys    -- works, bad with  !
```

Lazy **let** and **where**, strict **case**

The outermost pattern in a **let** or **where** is lazy by default:

```
unzip :: [(a, b)] → ([a], [b])
unzip []             = ([], [])
unzip ((x, y) : xys) = (x : xs, y : ys)
  where
    (xs, ys) = unzip xys   -- works, bad with  !
```

On the other hand, patterns in a **case**, lambda or left hand side are strict:

```
unzip :: [(a, b)] → ([a], [b])
unzip []             = ([], [])
unzip ((x, y) : xys) = case unzip xys of
  ∼(xs, ys) → (x : xs, y : ys)   -- requires ∼ , bad without
```

Strict data, lazy functions

Some advice:

- Don't worry about strictness too much or too early.
- Do not just assume that making things stricter is always positive – understanding evaluation is the key, laziness in some places is as desirable as strictness is in others.
- Do not spread strictness annotations (bang patterns, `seq`) over a large amount of functions.
- Try to make certain pieces of data strict, either by using strict fields or by establishing strictness invariants in a small interface.
- Other functions using the interface will then automatically maintain the strictness invariants.
- Never forget that `seq`, bang patterns and strict fields only force WHNF.

Normal form

When WHNF is not enough . . .

- For structured types, WHNF and NF are not the same.
- Sometimes – in particular in the context of parallel programming – we want data to be evaluated completely.
- Unlike `seq`, this functionality is not built into the language, but rather provided by a library `Control.DeepSeq` in the `deepseq` package.

Control.DeepSeq library

```
class NFData a where
   rnf :: a → ()
   rnf x = x `seq` ()    -- suitable default for flat types

deepseq :: NFData a ⇒ a → b → b
deepseq x y = rnf x `seq` y

($!!) :: NFData a ⇒ (a → b) → a → b
f $!! x = x `deepseq` f x

force :: NFData a ⇒ a → a
force x = x `deepseq` x
```

Control.DeepSeq library

```
class NFData a where
  rnf :: a → ()
  rnf x = x `seq` ()    -- suitable default for flat types

deepseq :: NFData a ⇒ a → b → b
deepseq x y = rnf x `seq` y

($!!) :: NFData a ⇒ (a → b) → a → b
f $!! x = x `deepseq` f x

force :: NFData a ⇒ a → a
force x = x `deepseq` x
```

Note that what doesn't make sense for WHNF (a function like `force`) does make some sense for NF.

Defining NFData instances

Example instance:

```
instance NFData a ⇒ NFData [a] where
  rnf []       = ()
  rnf (x : xs) = rnf x `seq` rnf xs
```

Note:

- The definition traverses the entire list before returning ().
- The call rnf x requires that the element type is an instance of NFData, too.
- We use seq in the second case to ensure that rnf x actually completes and produces the ().
- Calling deepseq has a cost. Don't force large structures unnecessarily often.
- Control.DeepSeq exports many basic instances.

Thought experiments

- Including partially defined values, how many different elements of type $(\mathrm{Bool}, \mathrm{Bool})$ are there?
- And how many lists of type $[()]$ and length at most 2?
- Can you write programs to distinguish all of these?

Lessons

- Both laziness and strictness can be desirable in certain situations.
- In Haskell, you are lazy by default, and have different options to make things more strict, selectively.
- Flat data often "wants" to be strict. For structured data, laziness is often desirable.
- Do not worry about strictness too early.
- Try to establish simple invariants on top of your datatypes – do not use `seq` or bang patterns in "random" places throughout your code.
- Use GHC's runtime statistics or profiling in order to pinpoint how much space is used and where time is spent.