Haskell in the Newsroom

A little background

- Twitter: @erikhinton
- Been at the Times for 2.5 years
- Been using Haskell almost as long as I've been programming

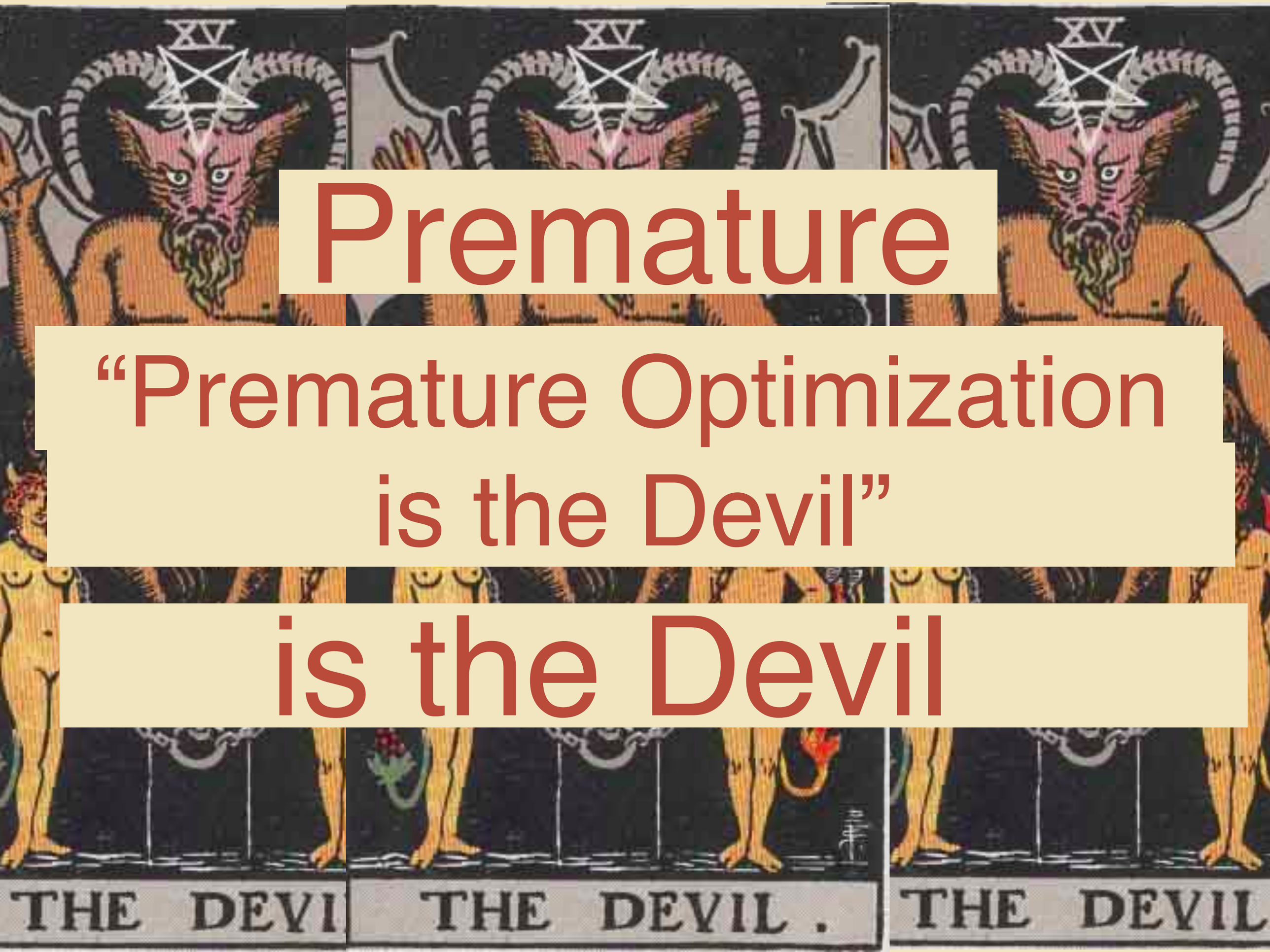
What to expect

- A perspective on integrating Haskell into a newsroom

- A paucity of actual code (not the point)

- The first google image I could find for each slide

“Premature Optimization is the Devil”



Premature

“Premature Optimization is the Devil”

is the Devil

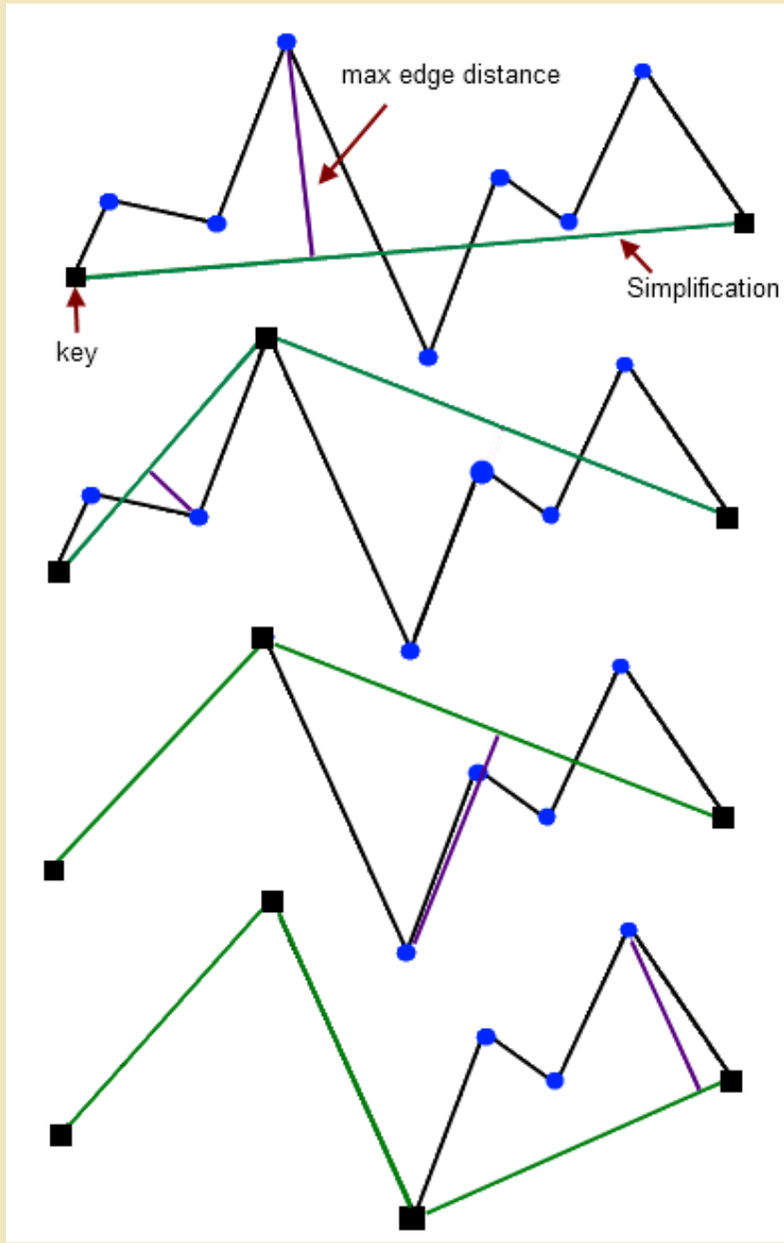
How it all began

P e u c k e r m y b r a i n s o u t

- “This algorithm seems simple, should be easy to whip up in Ruby.”

- ERRNO - How about I just crash on any moderately large shapes

- “I should learn Haskell.”



The agony of insulation

- You have dig into the Ruby source C (MRI) to even find the time complexity of array operations

- Insulated from the cost of our code, we lowball and undermine our ambition (in

What Haskell offers

Extra! Extra!

- Raw speed
- Data analysis parallelism
- Simple executables (No VM)
- Constant-space, resource-deterministic, modular stream processing
- Data abstraction

Speed

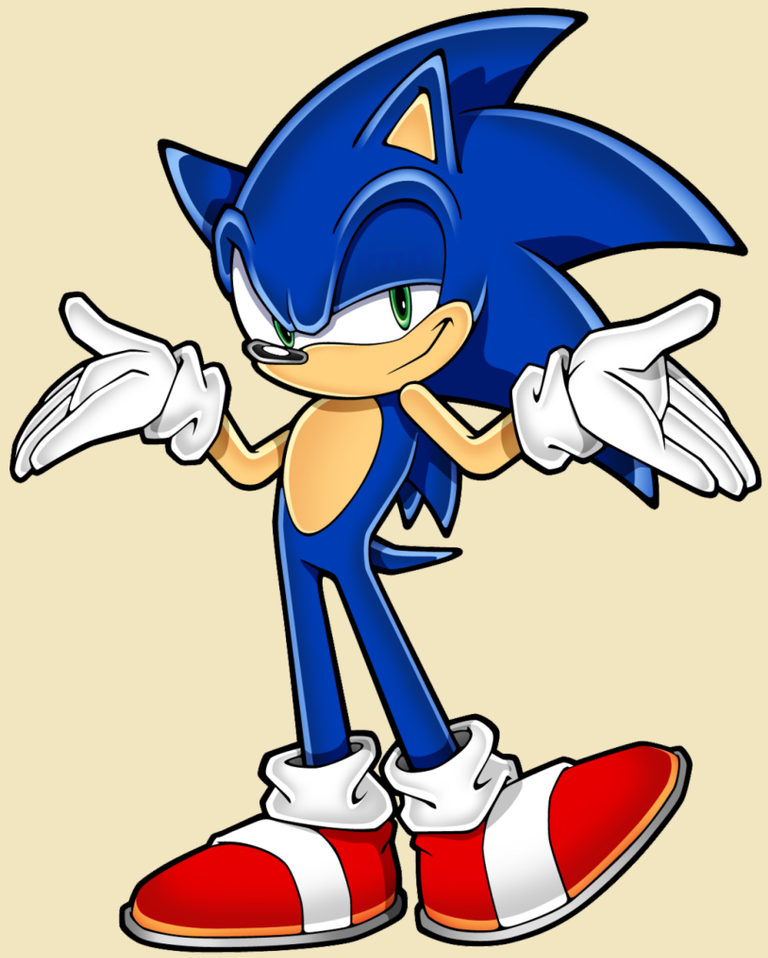
. . . a n d i t s d i s c o n t e n t s

- Not really a claim about benchmark speed (though true)

- Largely a product of hyper-focused libraries/structures

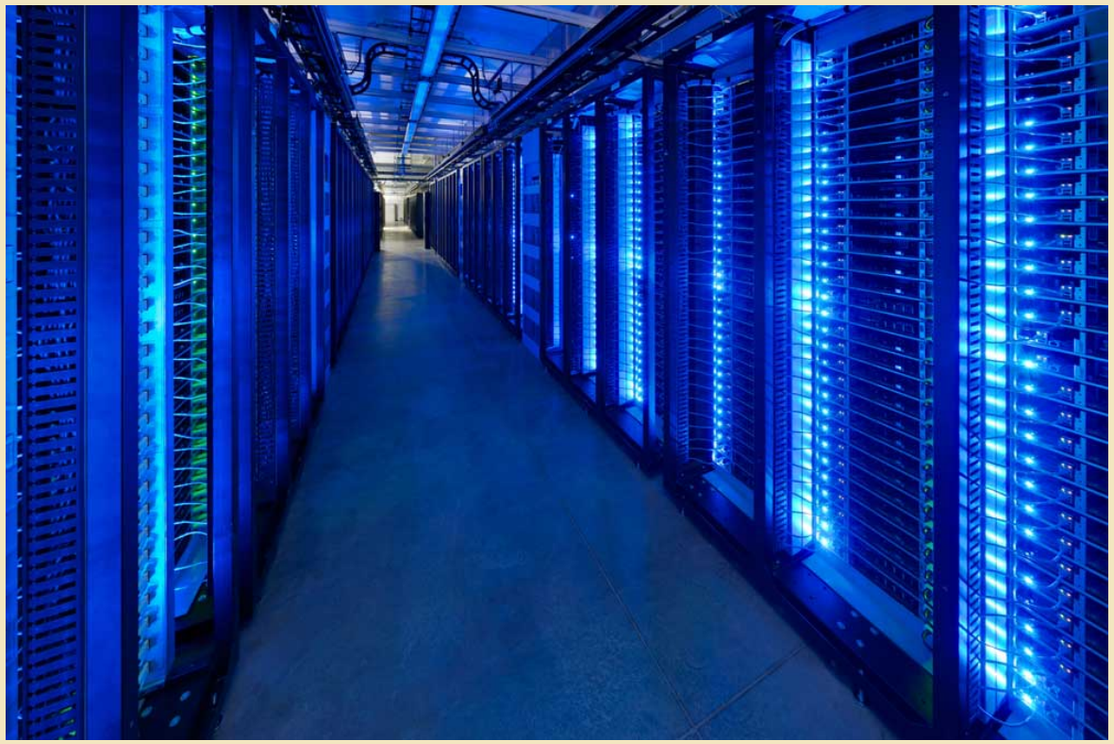
- Naive code on deadline - 60+ apps/yr

- Ruby’s insanely expensive ORMs



Throw EC2 at it?

- There's a misconception that since we can boot up XXL compute clusters, offline processing doesn't need to be optimized



- The cost of dev time

- Newspapers are poor(ish)

- Creative iteration

But ... queues!

- Now you have two problems
- Distributed systems are never as simple as they seem
- See Election Night

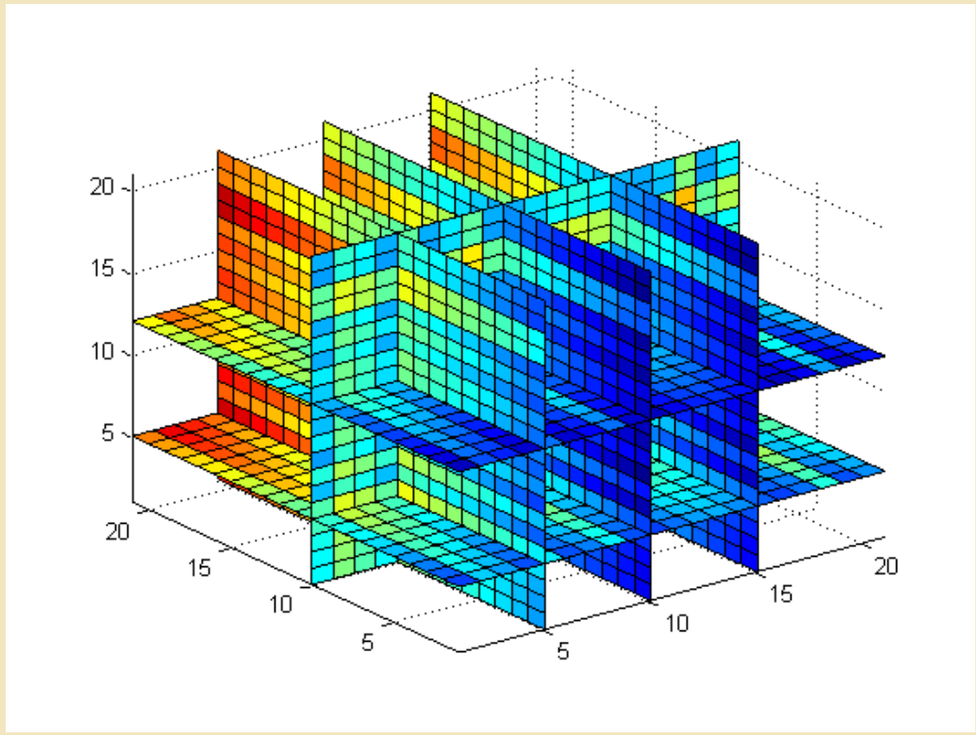
Parallel Dimensions

A Repa love story

- Single-machine

- Automatically parallel operations with regular, multi-dimensional arrays

- Moves data journalism out of get-a-coffee wait times



Flirting with a lisp

- Clojure, Java, and Scala are other languages we have used or tried to use for these data processing tasks.
  - Riemann
- The “complexities” of JVM-ed scripts on a cron/on a worker
- Daemons for short-lived apps
- Files over folders

Streaming data

- The mythical 400-gigabyte XML file
- Olympics, election
- Everyone loves a good firehose

Why conduits

- Both Ruby and Python offer constant-space file stream consumption with iterators that perform nicely
- Modularity
- Moving from fixture to real data, fast prototyping
- What about web streaming?
- “Use EventMachine/ Twisted”
- Handle low-level details with ‘socket’ or ‘httplib’
- In conduit: just replace source with http-conduit

Can-do-its

```
EventMachine.run do
  http = ...
  http.stream do |chunk|
        File.open("foo.txt","a"){|f|
                f.write(chunk)}
      puts chunk
  end
```

```
sourceList ["1","2","3"] $$
   conduitFile "foo.txt" =$
   sinkHandle stdout
responseBody res $$+-
   conduitFile "foo.txt" =$
   sinkHandle stdout
```

Can-do-its Pt. 2

```
conduit :: Conduit Int IO String
conduit = do
    -- Get all of the adjacent pairs from the stream
    mi1 <- await
    mi2 <- await
    case (mi1, mi2) of
      (Just i1, Just i2) -> do
        yield $ show (i1, i2)
        leftover i2
        conduit
```

https://www.fpcomplete.com/school/advanced-haskell/conduit-overview

Data, abstraction

- The purists' comma

- "We should stop focusing on writing functions to pull specific data out of objects, and start leaning on the structure of our data."

- The olympics data horrorshow

Olympic pains

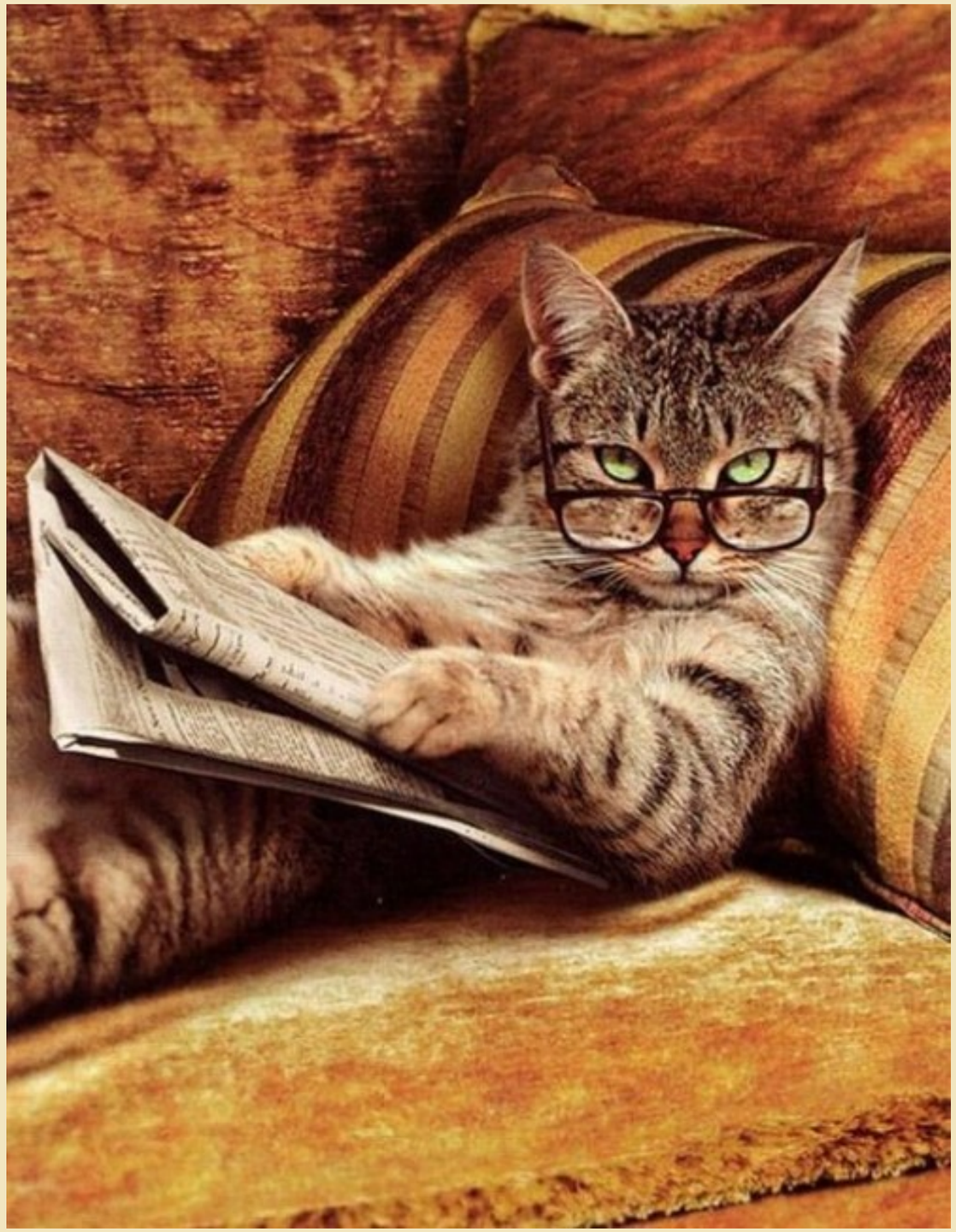
- A bag full of helper functions, ad hoc parsers

- Parsing/flattening data vs. querying a known structure

- Adventures in non-composability

Lenses

- Specify and spell out the abstract structure/ relationships in our data
- Free reference!
- Looping over objects and getting/setting sub-values vs. functional combinators



Lenses

A list of records with nulls

```
things.each do |t|
  if t
    t[:results].each do |r|
      r[:done] = true
    end
  end
end
things.map do |t|
  if t
    t[:results].map do |r|
      r[:done]
    end
  end
```

```
{- assume datatypes are above -}
done_t :: Traversal' [Maybe Record] Bool
done_t =
  traverse._Just.results.traverse.done


set (done_t) True things
toListOf (done_t) things
```

Our first foray

All the tweets

- Previously: a script running on a cron (the newsroom standard)

- But what of hurricanes!?

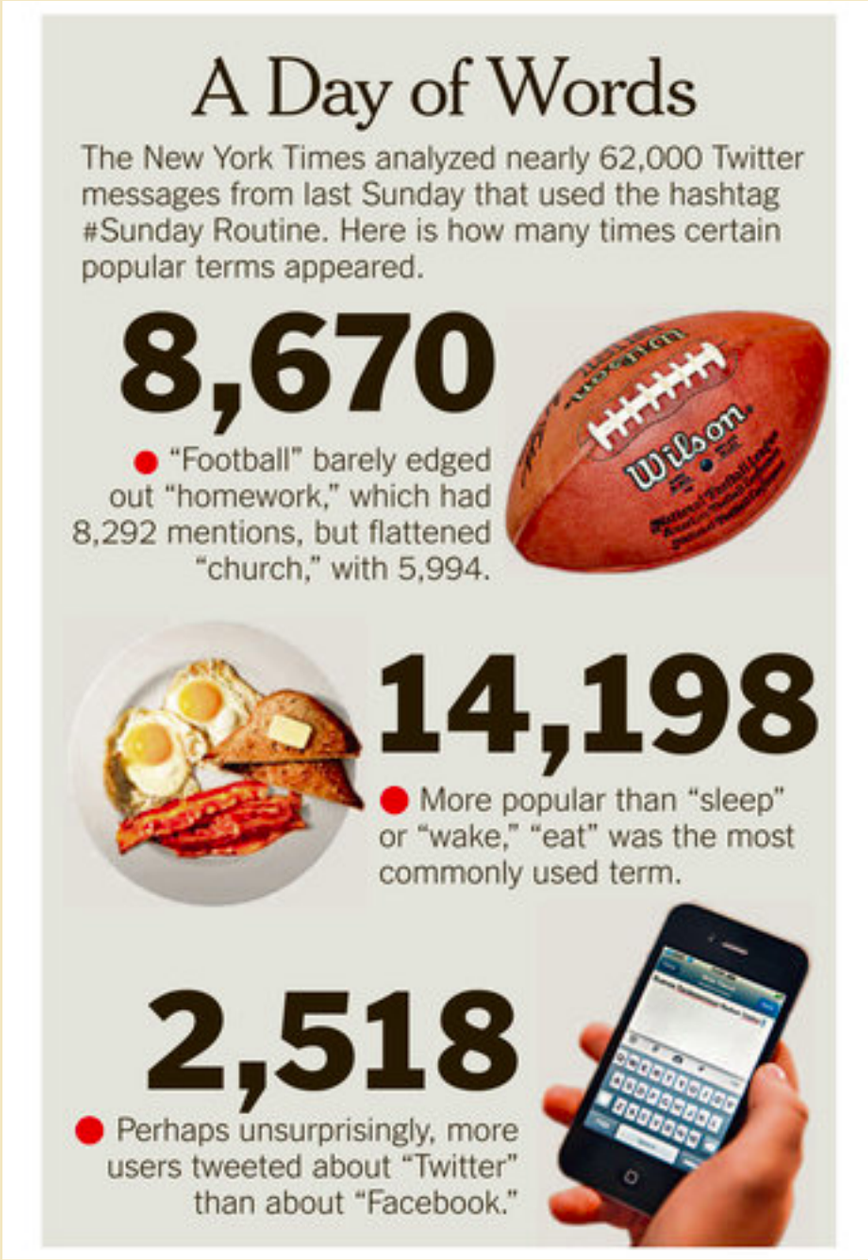
- Sunday routines

- Could n-gram at speed of firehose

A Day of Words

"I'll take it"

- Lick & Twobble
- Lick: enumerator, hashmap
- Twobble: enumerator, attoparsec, and aeson, never used
- Low bars and first runs: "We already use Haskell"



Fashion Fingerprints

The problem

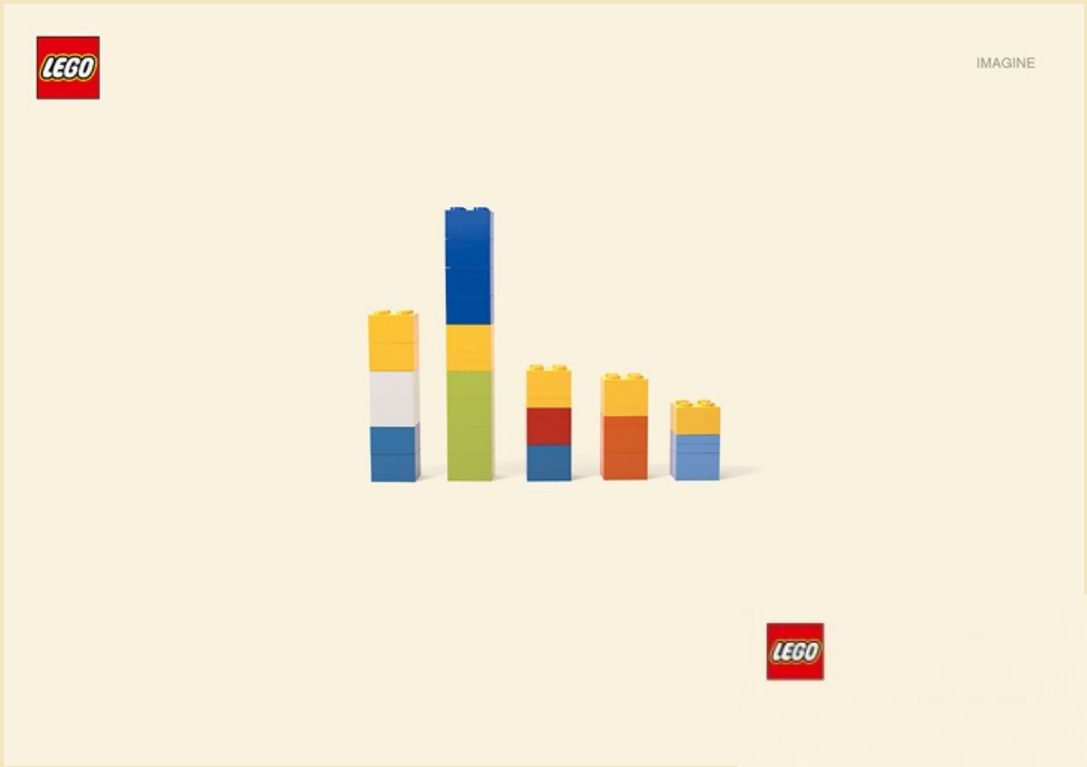
- Fashion week handwaving

- “I wish you’d do more Nate Silver stuff”

- No one understands fashion week

The Inspiration

Jung von Matt



Fingerprints

Marissa Webb

Costello Tagliapietra

Tanya Taylor

Creatures of the Wind

Jason Wu

Duckie Brown

Sally La Pointe

Cusnie et Ochs

A

B

C

D

E

F

G

The pieces

And months of soul searching

- Repa made the processing of 10,000 images at 640x480 easy, fast

- Repa solved resizing issue (do you crop first?)

- Running w/ DevIL

- Color distance and sliding scales of withiness

```
sqcie94 :: (Double,Double,Double) -> (Double,Double,Double) -> Double
sqcie94 (l,a,b) (l',a',b') = lterm ** 2 + cterm ** 2 + hterm ** 2
                           where lterm = deltal/(kl * sl)
                                 cterm = deltac/(kc * sc)
                                 hterm = deltah/(kh*sh)
                                 deltal = l - l'
                                 deltac = c - c'
                                 c = sqrt (a**2 + b**2)
                                 c' = sqrt (a'**2 + b'**2)
                                 deltah = sqrt (deltaa ** 2 + deltab ** 2 - deltac ** 2)
                                 deltaa = a - a'
                                 deltab = b - b'
                                 sl = 1
                                 sc = 1 + 0.048 * c
                                 sh = 1 + 0.014 * c
                                 kl = 2
                                 kc = 1
                                 kh = 1
```

$$\Delta E^*_{94} = \sqrt{\left(\frac{\Delta L^*}{k_L S_L}\right)^2 + \left(\frac{\Delta C^*_{ab}}{k_C S_C}\right)^2 + \left(\frac{\Delta H^*_{ab}}{k_H S_H}\right)^2}$$

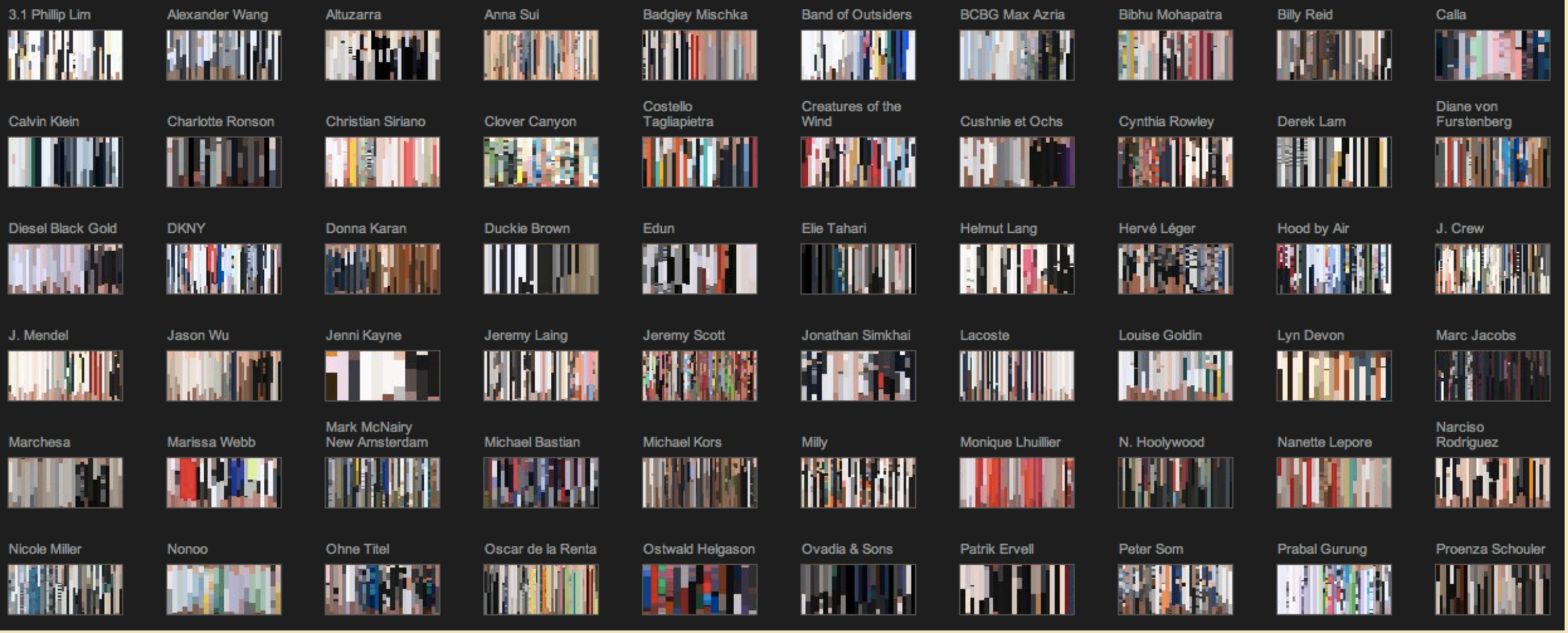
where:

$$\Delta L^* = L^*_1 - L^*_2$$
$$C^*_1 = \sqrt{{a^*_1}^2 + {b^*_1}^2}$$
$$C^*_2 = \sqrt{{a^*_2}^2 + {b^*_2}^2}$$
$$\Delta C^*_{ab} = C^*_1 - C^*_2$$
$$\Delta H^*_{ab} = \sqrt{{\Delta E^*_{ab}}^2 - {\Delta L^*}^2 - {\Delta C^*_{ab}}^2} = \sqrt{{\Delta a^*}^2 + {\Delta b^*}^2 - {\Delta C^*_{ab}}^2}$$
$$\Delta a^* = a^*_1 - a^*_2$$
$$\Delta b^* = b^*_1 - b^*_2$$
$$S_L = 1$$
$$S_C = 1 + K_1 C^*_1$$
$$S_H = 1 + K_2 C^*_1$$

Final Product

Fashion Fingerprints

Color is an important part of the ideas and trends that emerge during Fashion Week. Each look is broken into a minimal stack of color bars, then combined to abstract the collection's hues.



The Book Review

S h h h h

- Awful presentation of back catalog
- Common phrases - failure
- Quotation extraction - interesting
- 40k reviews in ~ 10 sec.
- Exploratory power

The New York Times
Book Review
FEBRUARY 3, 2013 $2

YUKO SHIMIZU

Folk Opera

By Ian Buruma

SANDALWOOD DEATH *By Mo Yan. Translated by Howard Goldblatt. 409 pp. University of Oklahoma Press. Paper, $24.95.*

POW! *By Mo Yan. Translated by Howard Goldblatt. 386 pp. Seagull Books. $27.50.*

Mo Yan, winner of the 2012 Nobel Prize in Literature, has a deft way with similes: salty, sometimes gross, usually unexpected. Comparing women's breasts to "ripe mangoes" is almost a cliché, but to describe the nipples as "rising gracefully, like the captivating mouths of hedgehogs" is arresting. Passengers disembarking from a train do so "like beetles rolling their precious dung." A rich meal of pork lies on a man's stomach, "churning and grinding like a litter of soon-to-be-born piglets."

What gives Mo Yan's novels their highly idiosyncratic tone is the combination of a great literary imagination and a peasant spirit. Howard Goldblatt's translations catch this atmosphere brilliantly. The prose reads well in English, without losing a distinctly Chinese feel, but it is very far from the classical Chinese tradition. There is nothing mandarin, or even urbane, about Mo Yan's work. He has retained the earthy character of rural Shandong, where he grew up in a farming family.

Continued on Page 10

Future projects

- Replacing our sad, overworked bakers
- Photostream and the agony of a heavy trigger finger (Usain Bolt)
- Our inter app messaging service (probably a pipe dream)

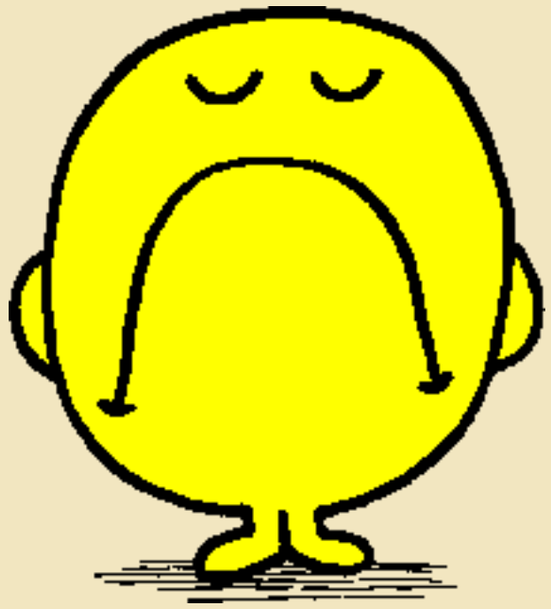
The Dark Side

Not all roses with Haskell

- The tarpits of laziness and profiling as a necessity

- News devs love generic, arbitrary JSON

- Having to hand someone Awodey before I can talk about the lens library type signatures



- Haskell needs that one headline app or framework to motivate the learning curve slog

Op-Eds

More Haskell temptations

- Hackage/ most Haskell documentation blows Ruby's out of the water: types and less folder hell
- The profile grapher
- Type holes and GHCi

Totem and Taboo

Haskell in a strange land

- “What a joke!”
- Hit-by-a-car anxiety
- The sedative of “offline”
- Monadic concerns
- Baby names

QUESTIONS?

- @erikhinton