Why is there a need to consider purely functional data structures?

- The standard implementations of many data structures assume imperative update. To what extent truly necessary?
- Purely functional data structures are **persistent**, while imperative ones are **ephemeral**:
  - Persistence is a useful property in its own right.
  - Can’t expect added benefits for free.
Purely Functional Data structures (4)

This lecture draws from:

We will look at some examples of how numerical representations can be used to derive purely functional data structures.

Numerical Representations (1)

Strong analogy between lists and the usual representation of natural numbers:

```haskell
data List a = Nil | Cons a (List a)
tail (Cons _ xs) = xs
append Nil ys = ys
append (Cons x xs) ys = Cons x (append xs ys)
```

```haskell
data Nat = Zero | Succ Nat
tail (Succ n) = n
append Nil ys = ys
append (Succ m n) = Succ (plus m n)
```

Numerical Representations (2)

This analogy can be taken further for designing container structures because:

- inserting an element resembles incrementing a number
- combining two containers resembles adding two numbers

Thus, representations of natural numbers with certain properties induce container types with similar properties. Called Numerical Representations.

Random Access Lists

We will consider Random Access Lists in the following. Signature:

```haskell
data RList a

empty :: RList a
isEmpty :: RList a -> Bool
cons :: a -> RList a -> RList a
head :: RList a -> a
tail :: RList a -> RList a
lookup :: Int -> RList a -> a
update :: Int -> a -> RList a -> RList a
```
Positional Number Systems (1)

- A number is written as a sequence of digits $b_0 b_1 \ldots b_{m-1}$, where $b_i \in D_i$ for a fixed family of digit sets given by the positional system.
- $b_0$ is the least significant digit, $b_{m-1}$ the most significant digit (note the ordering).
- Each digit $b_i$ has a weight $w_i$. Thus:

$$\text{value}(b_0 b_1 \ldots b_{m-1}) = \sum_{0}^{m-1} b_i w_i$$

where the fixed sequence of weights $w_i$ is given by the positional system.

Positional Number Systems (2)

- A number is written in base $B$ if $w_i = B^i$ and $D_i = \{0, \ldots, B - 1\}$.
- The sequence $w_i$ is usually, but not necessarily, increasing.
- A number system is redundant if there is more than one way to represent some numbers (disallowing trailing zeroes).
- A representation of a positional number system can be dense, meaning including zeroes, or sparse, eliding zeroes.

Exercise 1: Positional Number Systems

Suppose $w_i = 2^i$ and $D_i = \{0, 1, 2\}$. Give three different ways to represent 17.

Exercise 1: Solution

- 10001, since $\text{value}(10001) = 1 \cdot 2^0 + 1 \cdot 2^4$
- 10002, since $\text{value}(10002) = 1 \cdot 2^0 + 2 \cdot 2^3$
- 1021, since $\text{value}(1021) = 1 \cdot 2^0 + 2 \cdot 2^2 + 1 \cdot 2^3$
- 1211, since $\text{value}(1211) = 1 \cdot 2^0 + 2 \cdot 2^1 + 1 \cdot 2^2 + 1 \cdot 2^3$
From Positional System to Container

Given a positional system, a numerical representation may be derived as follows:

- for a container of size \( n \), consider a representation \( b_0b_1 \ldots b_{m-1} \) of \( n \),
- represent the collection of \( n \) elements by a sequence of trees of size \( w_i \) such that there are \( b_i \) trees of that size.

For example, given the positional system of exercise 1, a container of size 17 might be represented by 1 tree of size 1, 2 trees of size 2, 1 tree of size 4, and 1 tree of size 8.

What Kind of Trees?

The kind of tree should be chosen depending on needed sizes and properties. Two possibilities:

- **Complete Binary Leaf Trees**
  ```haskell
data Tree a = Leaf a | Node (Tree a) (Tree a)
```
  Sizes: \( 2^n, n \geq 0 \)

- **Complete Binary Trees**
  ```haskell
data Tree a = Leaf a | Node (Tree a) a (Tree a)
```
  Sizes: \( 2^{n+1} - 1, n \geq 0 \)
  (Balance has to be ensured separately.)

Example: Complete Binary Leaf Tree

Size \( 2^3 = 8 \):

Example: Complete Binary Tree

Size \( 2^4 - 1 = 15 \):
**Binary Random Access Lists (1)**

**Binary Random Access Lists** are induced by
- the usual binary representation, i.e. \( w_i = 2^i \), \( D_i = \{0, 1\} \)
- complete binary leaf trees

Thus:

```haskell
data Tree a = Leaf a
               | Node Int (Tree a) (Tree a)
data Digit a = Zero | One (Tree a)
type RList a = [Digit a]
```

The `Int` field keeps track of tree size for speed.

---

**Binary Random Access Lists (2)**

Example: Binary Random Access List of size 5:

```
[ One , Zero, One ]
```

---

**Binary Random Access Lists (3)**

The increment function on dense binary numbers:

- \( \text{inc} [ ] = [\text{One}] \)
- \( \text{inc} (\text{Zero} : \text{ds}) = \text{One} : \text{ds} \)
- \( \text{inc} (\text{One} : \text{ds}) = \text{Zero} : \text{inc} \text{ds} \) -- Carry

---

**Binary Random Access Lists (4)**

Inserting an element first in a binary random access list is analogous to `inc`:

```haskell
cons :: a -> RList a -> RList a
cons x ts = consTree (Leaf x) ts
consTree :: Tree a -> RList a -> RList a
consTree t [] = [One t]
consTree t (Zero : ts) = (One t : ts)
consTree t (One t' : ts) =
  Zero : consTree (link t t') ts
```
Binary Random Access Lists (5)

The utility function `link` joins two equally sized trees:

\[
\text{link } t_1 \ t_2 = \text{Node} (2 \times \text{size } t_1) \ t_1 \ t_2
\]

Exercise 2: `unconsTree`

The decrement function on dense binary numbers:

\[
\begin{align*}
\text{dec } [\text{One}] &= [] \\
\text{dec } (\text{One} : ds) &= \text{Zero} : ds \\
\text{dec } (\text{Zero} : ds) &= \text{One} : \text{dec } ds \quad \text{-- Borrow}
\end{align*}
\]

Define `unconsTree` following the above pattern:

\[
\text{unconsTree} : \text{RList } a \rightarrow (\text{Tree } a, \text{RList } a)
\]

And then `head` and `tail`:

\[
\begin{align*}
\text{head} : \text{RList } a &\rightarrow a \\
\text{tail} : \text{RList } a &\rightarrow \text{RList } a
\end{align*}
\]

Exercise 2: Solution (1)

\[
\begin{align*}
\text{unconsTree} : \text{RList } a &\rightarrow (\text{Tree } a, \text{RList } a) \\
\text{unconsTree } [\text{One } t] &= (t, []) \\
\text{unconsTree } (\text{One } t : ts) &= (t, \text{Zero} : ts) \\
\text{unconsTree } (\text{Zero} : ts) &= (\text{t1, One } t_2 : ts') \\
\text{where} \\
&(\text{Node } _{t1} \ t_2, ts') = \text{unconsTree } ts
\end{align*}
\]

Note: partial operation.
Exercise 2: Solution (2)

head :: RList a -> a
head ts = x
  where
    (Leaf x, _) = unconsTree ts

tail :: RList a -> RList a
tail ts = ts'
  where
    (_, ts’) = unconsTree ts

Binary Random Access Lists (7)

Lookup is done in two stages: first find the right tree, then lookup in that tree:

lookup :: Int -> RList a -> a
lookup i (Zero : ts) = lookup i ts
lookup i (One t : ts)
  | i < s = lookupTree i t
  | otherwise = lookup (i - s) ts
  where
    s = size t

Note: partial operation.

Binary Random Access Lists (8)

lookupTree :: Int -> Tree a -> a
lookupTree _ (Leaf x) = x
lookupTree i (Node w t1 t2)
  | i < w `div` 2 = lookupTree i t1
  | otherwise = lookupTree (i - w `div` 2) t2

The operation update has exactly the same structure.

Binary Random Access Lists (9)

Time complexity:
- cons, head, tail, perform $O(1)$ work per digit, thus $O(\log n)$ worst case.
- lookup and update take $O(\log n)$ to find the right tree, and then $O(\log n)$ to find the right element in that tree, so $O(\log n)$ worst case overall.

Time complexity for cons, head, tail disappointing: can we do better?
Skew Binary Numbers (1)

Skew Binary Numbers:
• \( w_i = 2^{i+1} - 1 \) (rather than \( 2^i \))
• \( D_i = \{0, 1, 2\} \)

Representation is redundant. But we obtain a canonical form if we insist that only the least significant non-zero digit may be 2.

Note: The weights correspond to the sizes of complete binary trees.

Skew Binary Numbers (2)

Theorem: Every natural number \( n \) has a unique skew binary canonical form. Proof sketch. By induction on \( n \).
• Base case: the case for 0 is direct.

Skew Binary Numbers (3)

• Inductive case. Assume \( n \) has a unique skew binary representation \( b_0b_1\ldots b_{m-1} \)
  - If the least significant non-zero digit is smaller than 2, then \( n + 1 \) has a unique skew binary representation obtained by adding 1 to the least significant digit \( b_0 \).
  - If the least significant non-zero digit \( b_i \) is 2, then note that \( 1 + 2(2^{i+1} - 1) = 2^{i+2} - 1 \). Thus \( n + 1 \) has a unique skew binary representation obtained by setting \( b_i \) to 0 and adding 1 to \( b_{i+1} \).

Exercise 3a: Skew Binary Numbers

• Give the canonical skew binary representation for 31, 30, 29, and 28.
• Solution: 00001, 0002, 0021, 0211
Exercise 3b: Skew Binary Numbers

Assume a **sparse** skew binary representation of the natural numbers

```
type Nat = [Int]
```

where the integers represent the **weight** of each **non-zero** digit. Assume further that the integers are stored in increasing order, except that the first two may be equal indicating that the smallest non-zero digit is 2. E.g. $28 = [3,3,7,15]$.

Implement a function `inc` to increment a natural number. (E.g. `inc [3,3,7,15] = [7,7,15]`)

Exercise 3b: Solution

```
inc :: Nat -> Nat
inc (w1 : w2 : ws)
  | w1 == w2 = w1 * 2 + 1 : ws
inc ws = 1 : ws
```

**Note! No carry propagation!**

E.g.:

```
inc [1,3,7,15] = [1,1,3,7,15]
inc [1,1,3,7,5] = [3,3,7,15]
inc [3,3,7,15] = [7,7,15]
```

Skew Binary Random Access Lists (1)

```
data Tree a = Leaf a | Node (Tree a) a (Tree a)
type RList a = [(Int, Tree a)]
```

```
empty :: RList a
empty = []
```

```
cons :: a -> RList a -> RList a
cons x ((w1, t1) : (w2, t2) : wts)
  | w1 == w2 = (w1 * 2 + 1, Node t1 x t2) : wts
cons x wts = ((1, Leaf x) : wts)
```

Example: Consing onto list of size 5:

```
cons [ ] = [ ]
```

Skew Binary Random Access Lists (2)

Example: Consing onto list of size 5:

```
cons [ ] = [ ]
```

```
```
Skew Binary Random Access Lists (3)

Example: Consing onto list of size 6:

\[
\text{cons } [\_, \_, ] = [\_, ]
\]

Skew Binary Random Access Lists (4)

Example: Consing onto list of size 7:

\[
\text{cons } [\_, ] = [\_, ]
\]

Skew Binary Random Access Lists (5)

\[
\begin{align*}
\text{head} & : \text{RList a} \to a \\
\text{head } ((\_, \text{Leaf } x) : _) & = x \\
\text{head } ((\_, \text{Node } x _) : _) & = x \\
\text{tail} & : \text{RList a} \to \text{RList a} \\
\text{tail } ((\_, \text{Leaf } _): \text{wts}) & = \text{wts} \\
\text{tail } ((\text{w}, \text{Node } t1 _ t2) : \text{wts}) & = (\text{w'}, t1) : (\text{w'}, t2) : \text{wts} \\
& \quad \text{where} \\
& \quad \quad \quad \text{w'} = \text{w } \text{'div' } 2 \\
\text{Note: again, partial operations.}
\end{align*}
\]

Skew Binary Random Access Lists (6)

\[
\begin{align*}
\text{lookup} & : \text{Int} \to \text{RList a} \to a \\
\text{lookup } i ((\text{w}, t) : \text{wts}) & \quad | \quad i < \text{w} \quad = \text{lookupTree } i \text{ w t} \\
& \quad | \quad \text{otherwise} \quad = \text{lookup } (i - \text{w}) \text{ wts} \\
\text{lookupTree} & : \text{Int} \to \text{Int} \to \text{Tree a} \to a \\
\text{lookupTree } _ _ (\text{Leaf } x) & = x \\
\text{lookupTree } i \text{ w } (\text{Node } t1 \text{ t2}) & \quad | \quad i = 0 \\
& \quad | \quad i < \text{w'} \\
& \quad | \quad \text{otherwise} \\
& \quad \quad \quad = \text{lookupTree } (i - 1) \text{ w'} \text{ t1} \\
& \quad \quad \quad | \quad \text{w'} = \text{w } \text{'div' } 2
\end{align*}
\]
Skew Binary Random Access Lists (7)

Time complexity:

- `cons`, `head`, `tail`: $O(1)$.
- `lookup` and `update` take $O(\log n)$ to find the right tree, and then $O(\log n)$ to find the right element in that tree, so $O(\log n)$ worst case overall.

Okasaki:

“Although there are better implementations of lists, and better implementations of (persistent) arrays, none are better at both.”