zip code:

unpacking
data compression

by Scott Vokes
@silentbicycle

Why Compression Matters

It’s essential infrastructure.

Content-Encoding: gzip

Send to compressed (zipped) folder

Multimedia encoding

Design for Compression

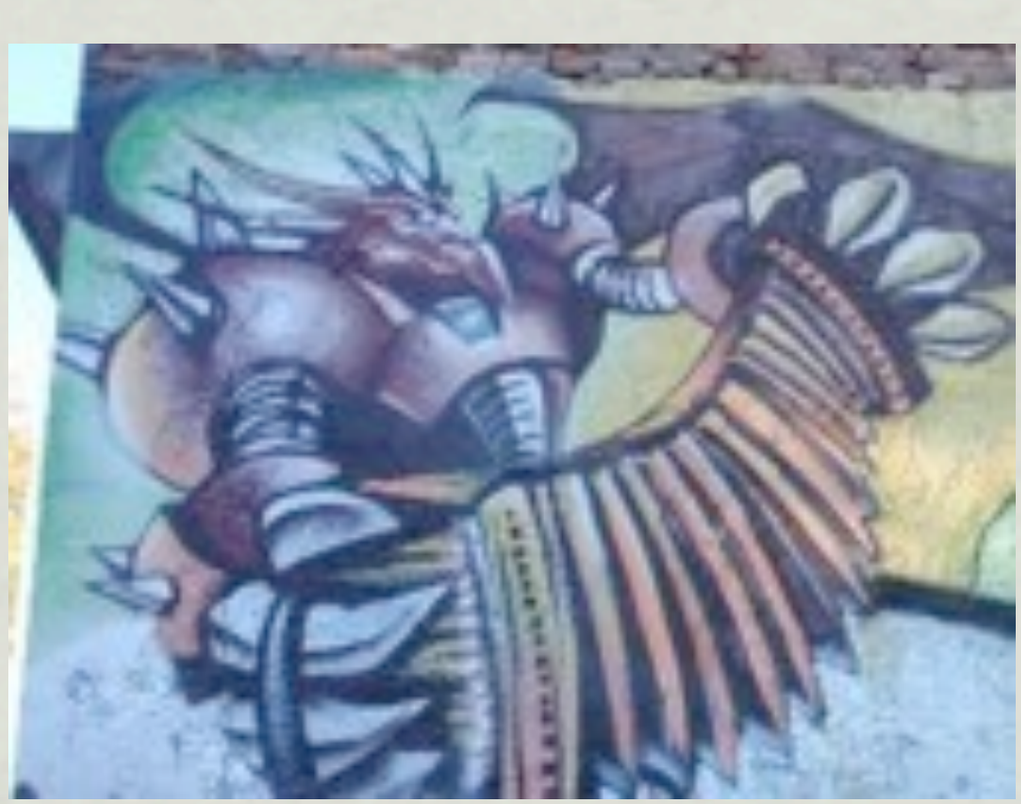
curses

XML

What is Data Compression?

heuristically detecting patterns

reducing duplication



FLICKR: WALLSTALKING.ORG

What is Data Compression?

"Friendly reminder:
Compression is machine learning."

- Paul Snively

Patterns & Repetition

NO PATTERN

hhluabsolsgtcoor

OBVIOUS PATTERN

abababababababab

Patterns & Repetition

NO PATTERN

hhluabsolsgtcoor

OBVIOUS PATTERN

abababababababab

Kolmogorov Complexity:

the smallest way to describe something
with 100% accuracy

Flavors of Compression

- lossless
  - gzip
  - zlib
  - GIF
- lossy
  - JPEG
  - MP3

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

Run-Length Coding

a b b b b b c d d d d d d d d d d e

Run-Length Coding

a b b b b b c d d d d d d d d d d e

a

Run-Length Coding

a b b b b b c d d d d d d d d d d e

a, 5 x b

Run-Length Coding

a b b b b b c d d d d d d d d d d e

a, 5 x b, c

Run-Length Coding

a b b b b b c d d d d d d d d d d e

a, 5 x b, c, 10 x d

Run-Length Coding

a b b b b b c d d d d d d d d d d e

a, 5 x b, c, 10 x d, e

Run-Length Coding

a b b b b b c d d d d d d d d d d e

a, 5 x b, c, 10 x d, e

or 1a 5b 1c 10d 1e

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

Delta Coding

32491
32492
32495
32500
32507
32516
32527

Delta Coding

32491	@32491
32492	+1
32495	+3
32500	+5
32507	+7
32516	+9
32527	+11

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

Huffman Coding (1952)

Variable-length bit patterns,
most common are shortest

Morse Code

COMMON	E .	T -
RARE	X -..-	J .---

Huffman Coding

sort tokens by frequency

merge nodes w/ lowest frequencies

build an unbalanced binary tree

Huffman Coding

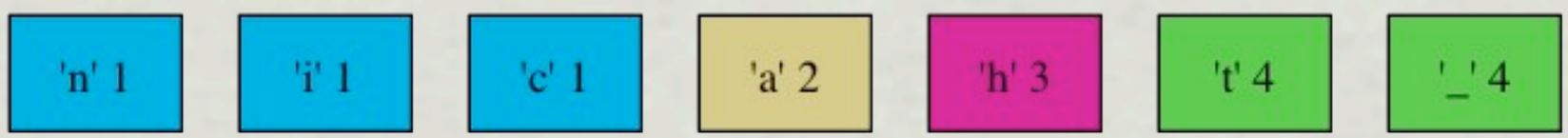
adaptive to frequencies in the data

COMMON the cat in the hat

UNUSUAL syzygy of zephyrs

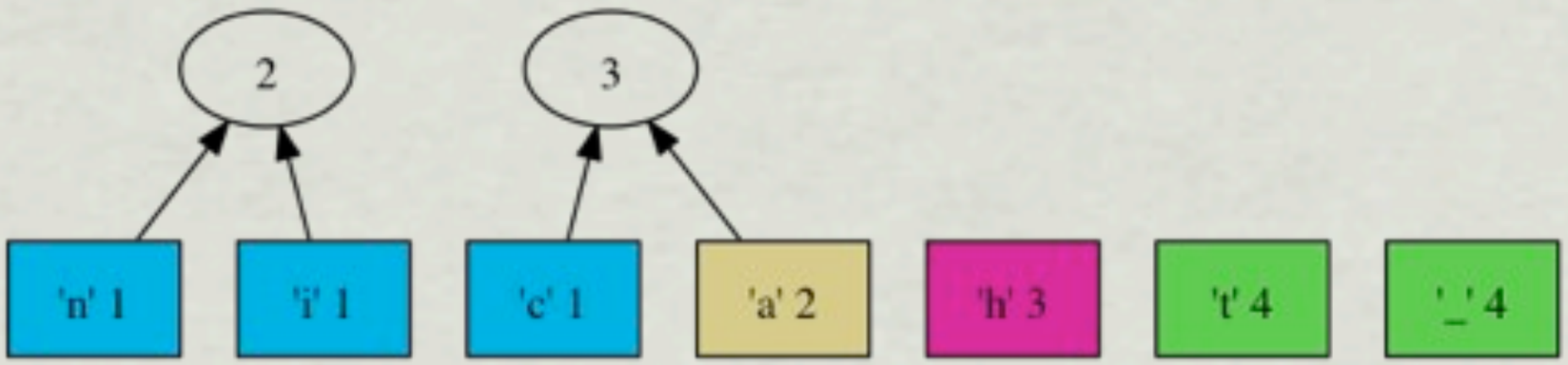
NARROW humulus lupulus

Huffman Coding

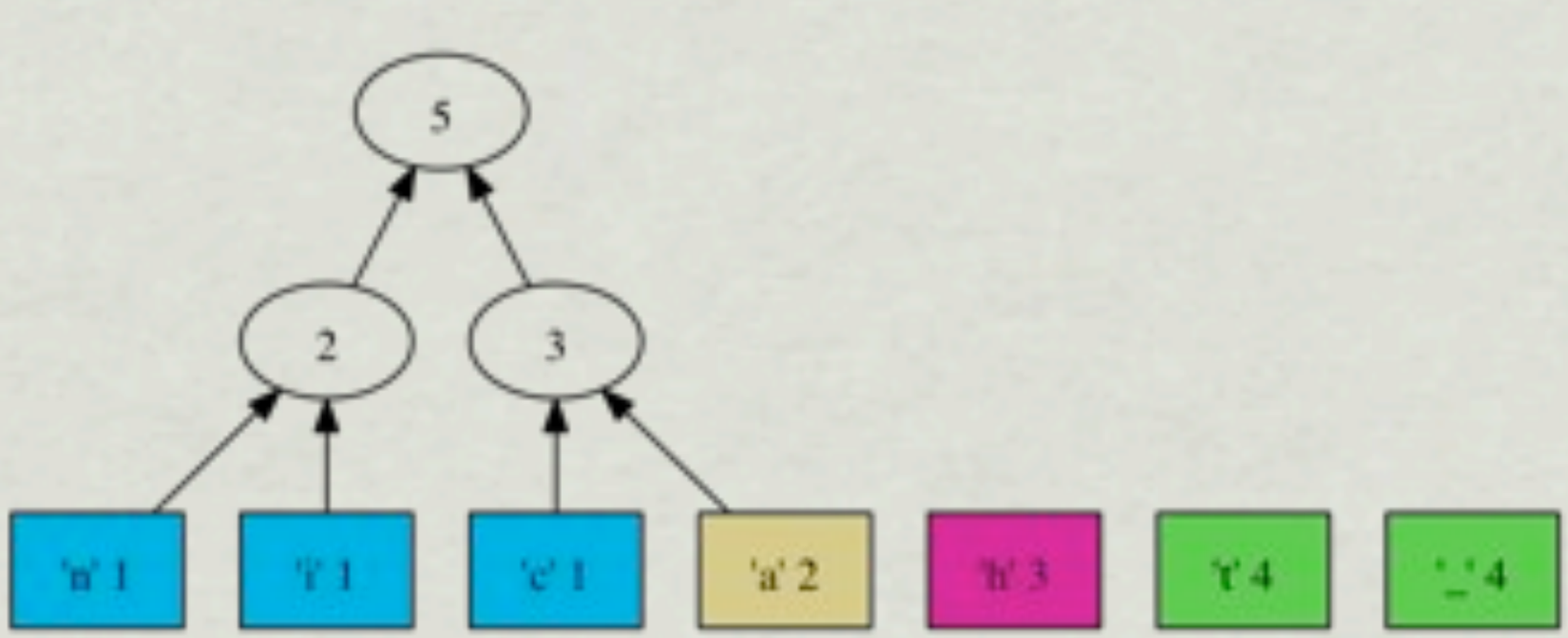


Huffman Coding

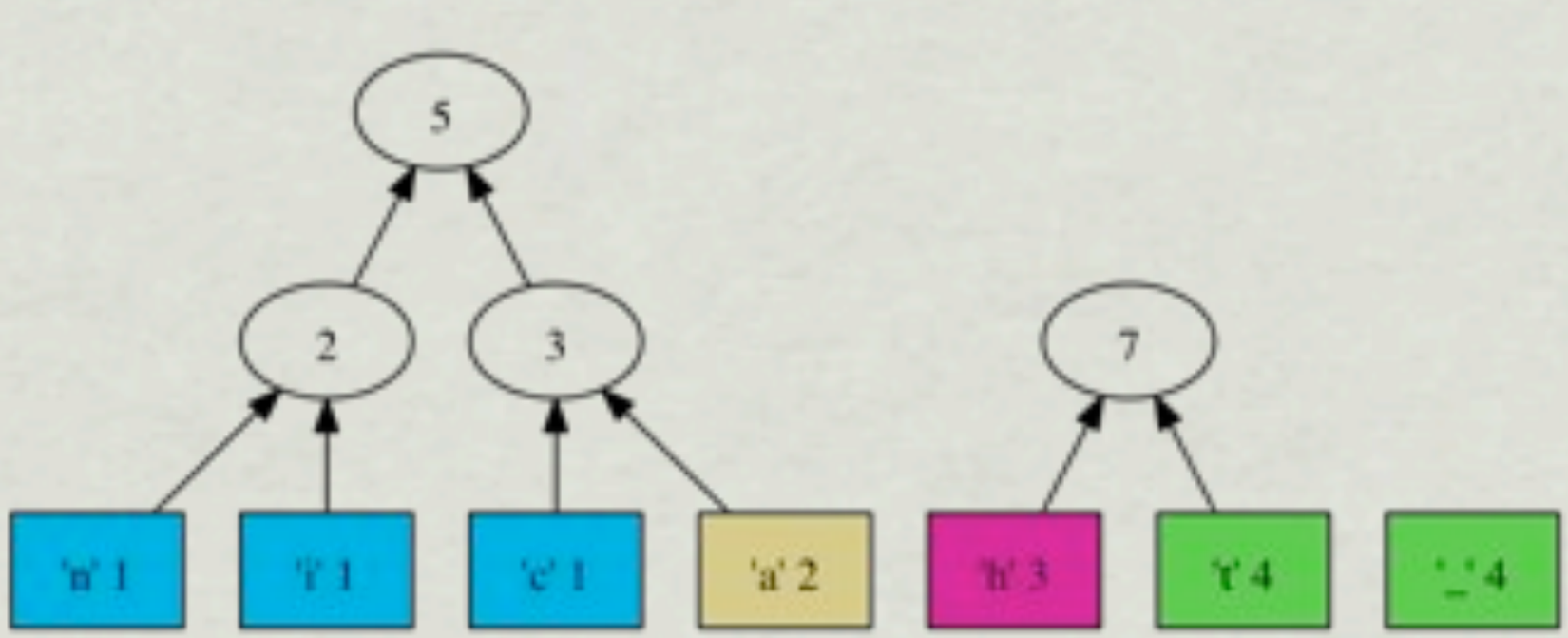
Huffman Coding



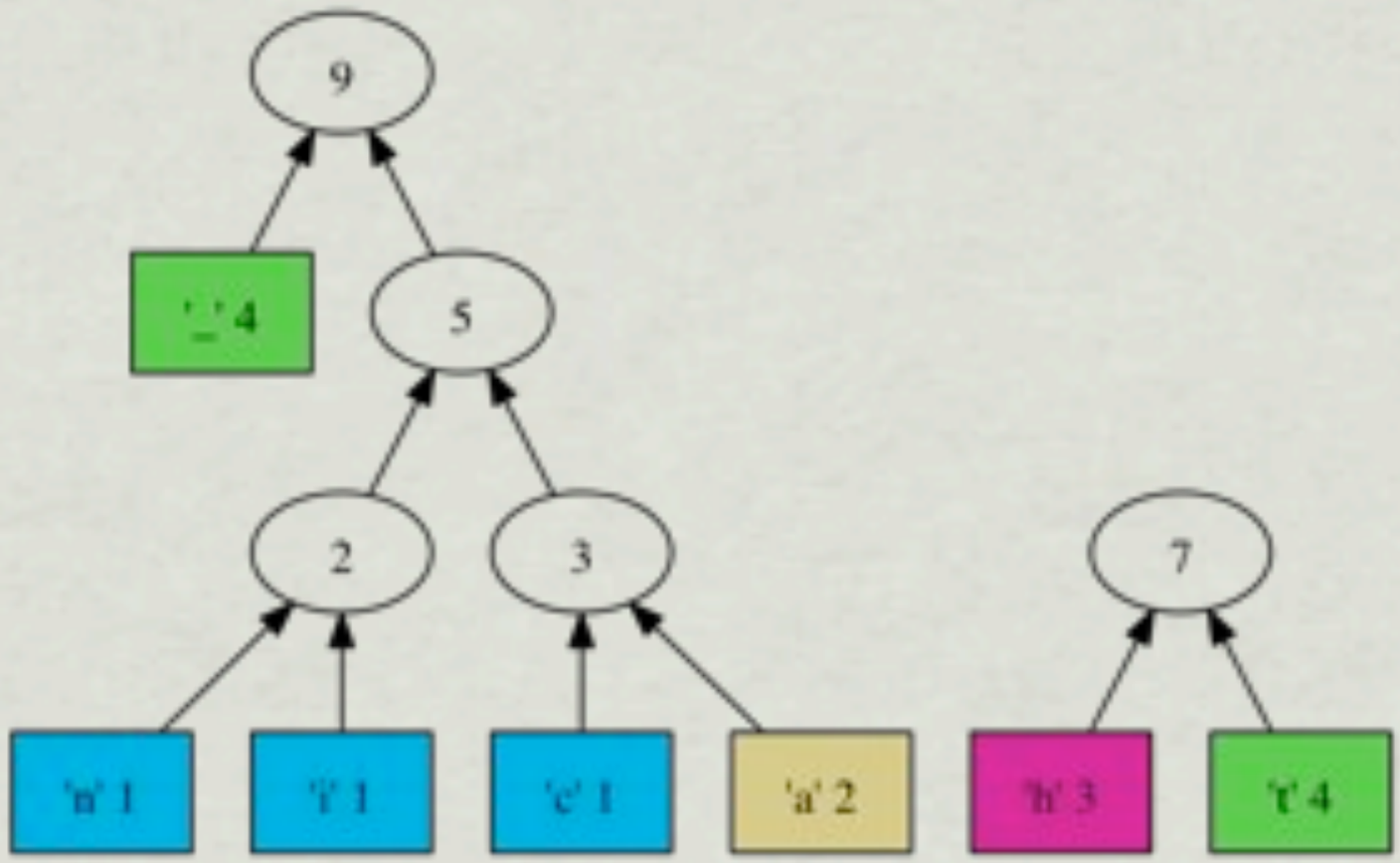
Huffman Coding



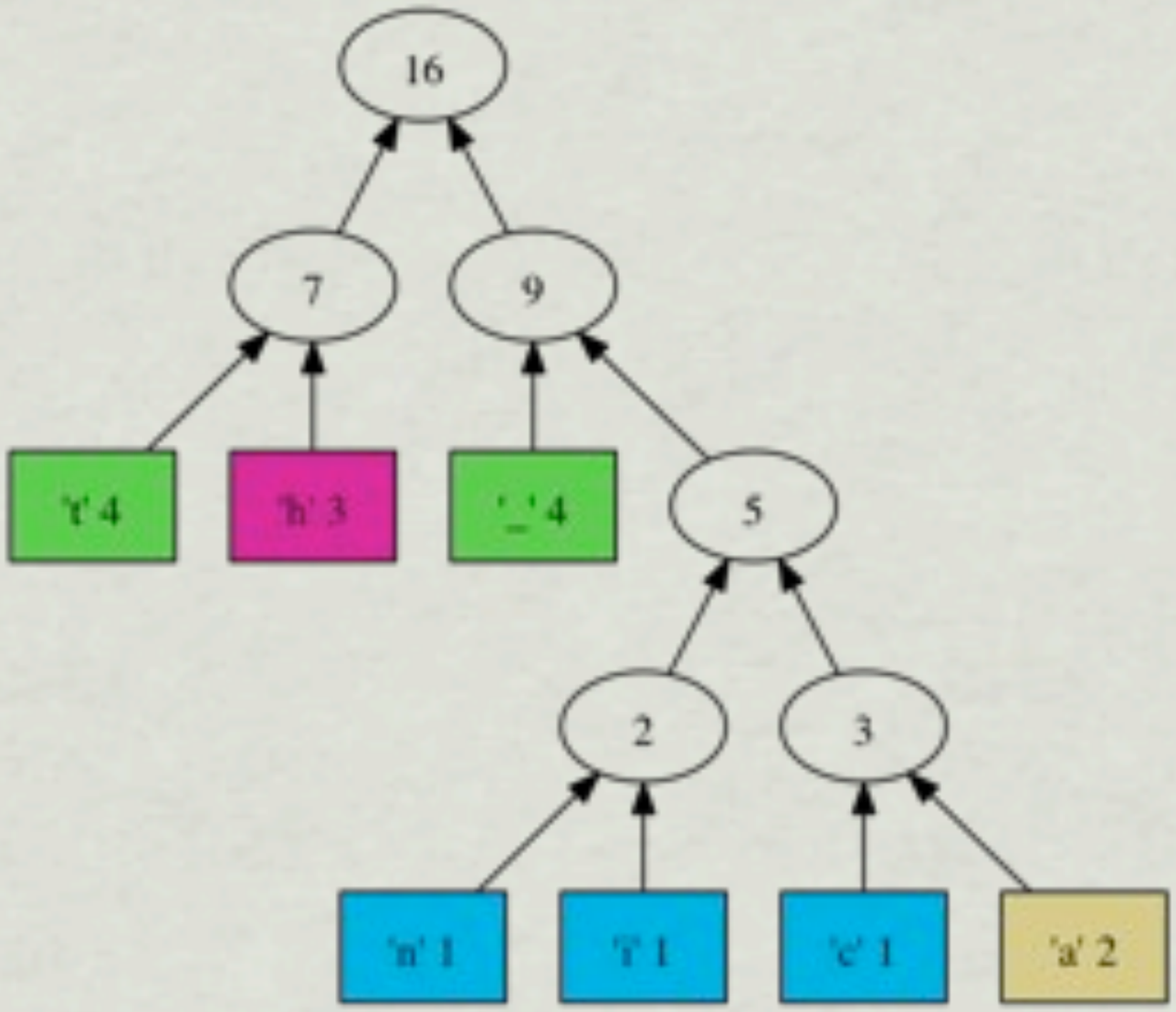
Huffman Coding



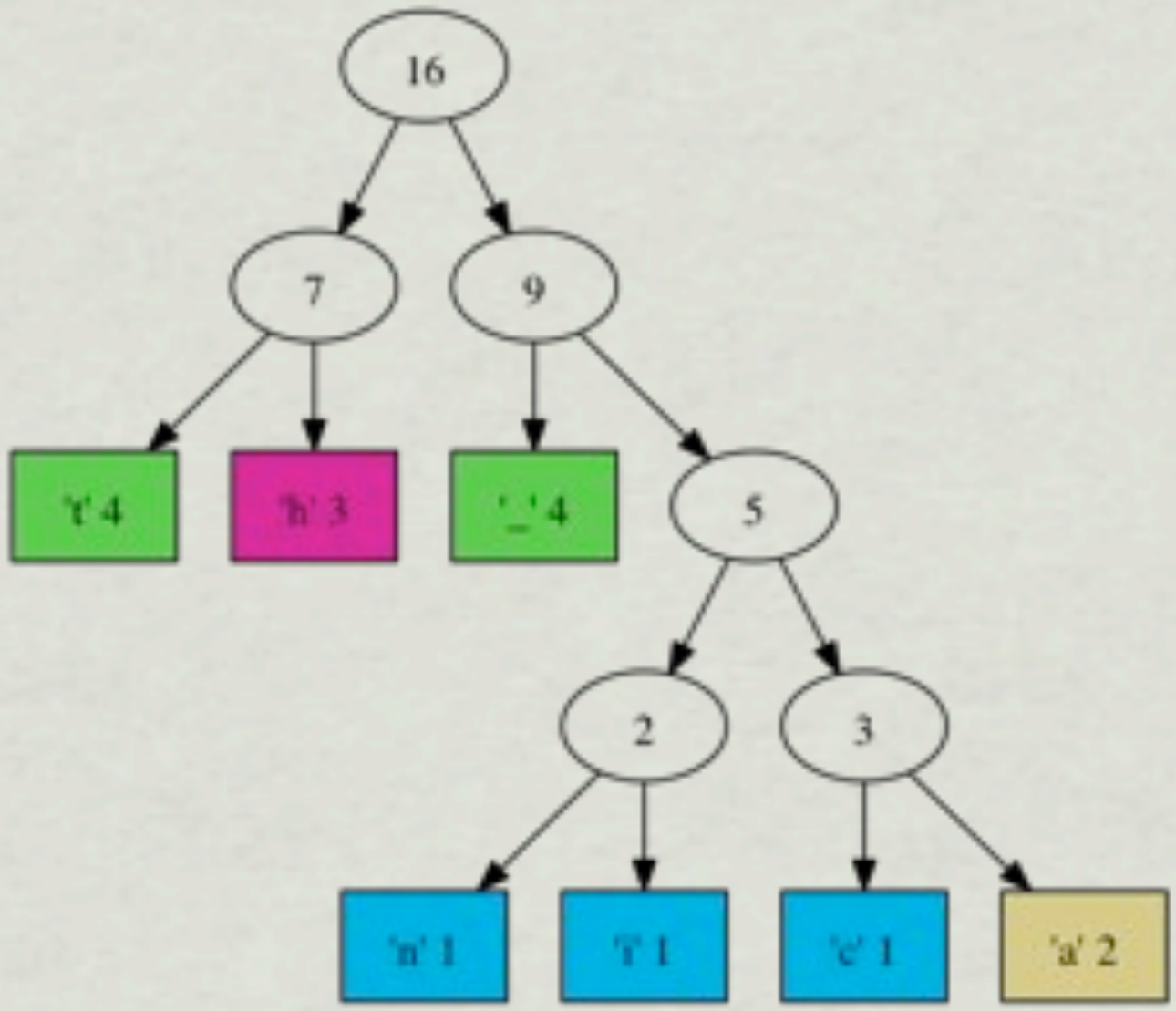
Huffman Coding



Huffman Coding



Huffman Coding




```
T 00
H 01
_ 10
N 1100
I 1101
C 1110
A 1111
```

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

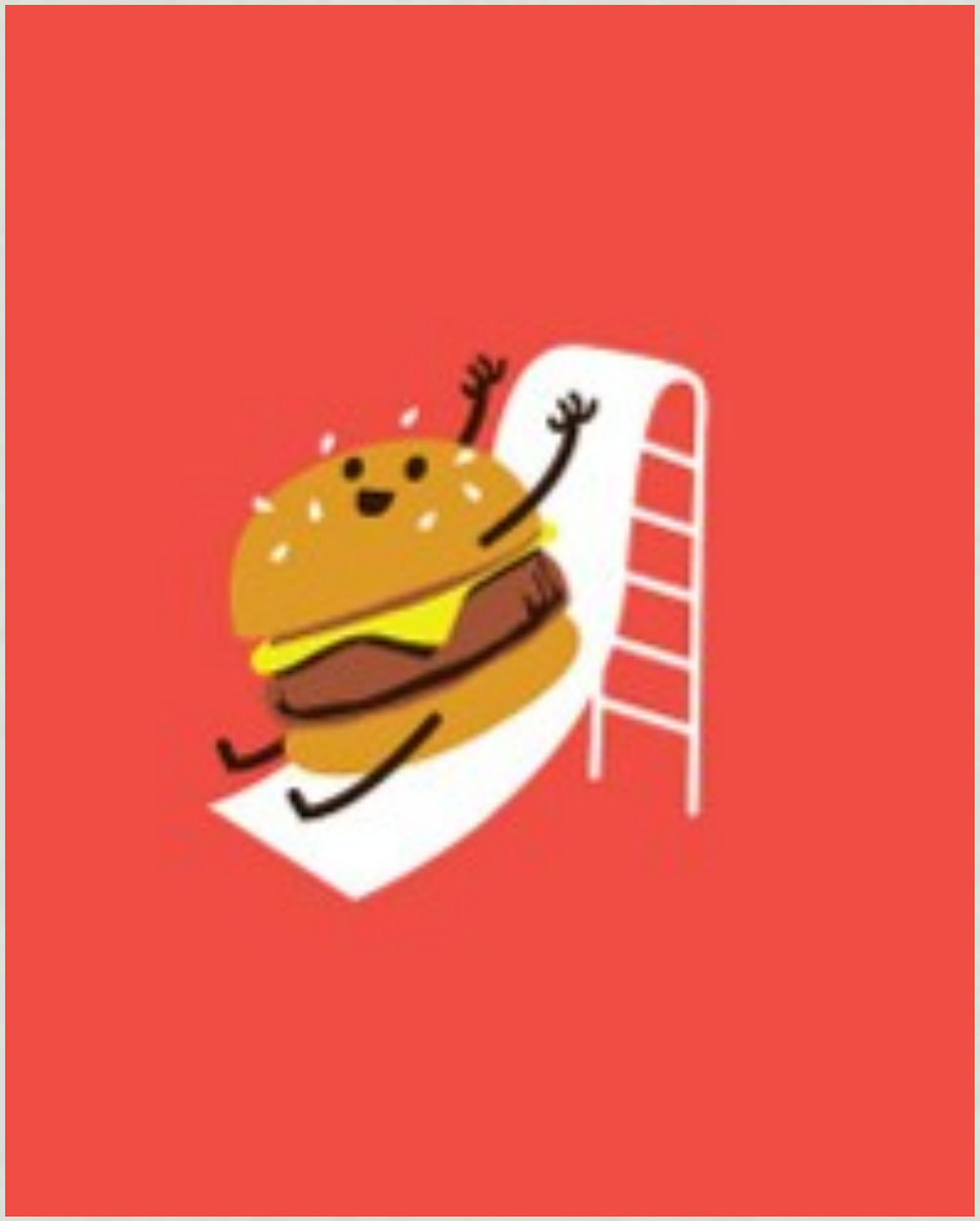
LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

LZ77 (1977)

sliding window compression

invented by Jacob Ziv and Abraham Lempel



SLIDER BURGER BY MARCO ANGELES WWW.MARCOANGELES.COM

LZ77

abcabcdabcdefghabcabchij

LZ77

abcabcdabcdefghabcabchij

abc

LZ77

abcabcdabcdefghabcabchij

abc# (back-reference)

(-3,+3)

LZ77

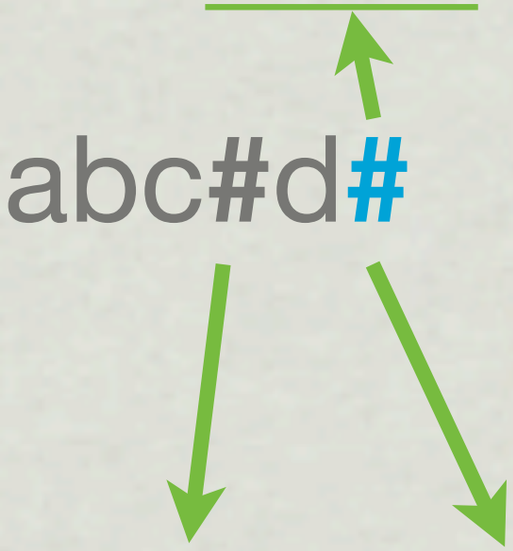
abcabcdabcdefghabcabchij

abc#d

(-3,+3)

LZ77

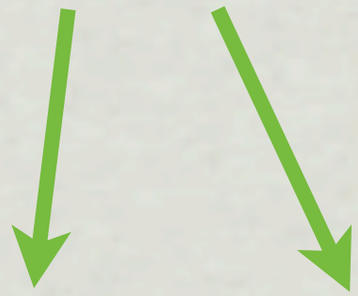
abcabcdabcdefghabcabchij



LZ77

abcabcdabcdefghabcabchij

abc#d#efgh



(-3,+3) (-4,+4)

LZ77

abcabcdabcdefghabcabchij

LZ77

abcabcdabcdefghabcabchij

abc#d#efgh#hij

(-3,+3) (-4,+4) (-15,+6)

LZ77

arrrrrrrrrr

LZ77

arrrrrrrrrr

ar

LZ77

arrrrrrrrrr

ar

LZ77

arrrrrrrrrr

ar# repeating past into the future

(-1,+9)

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

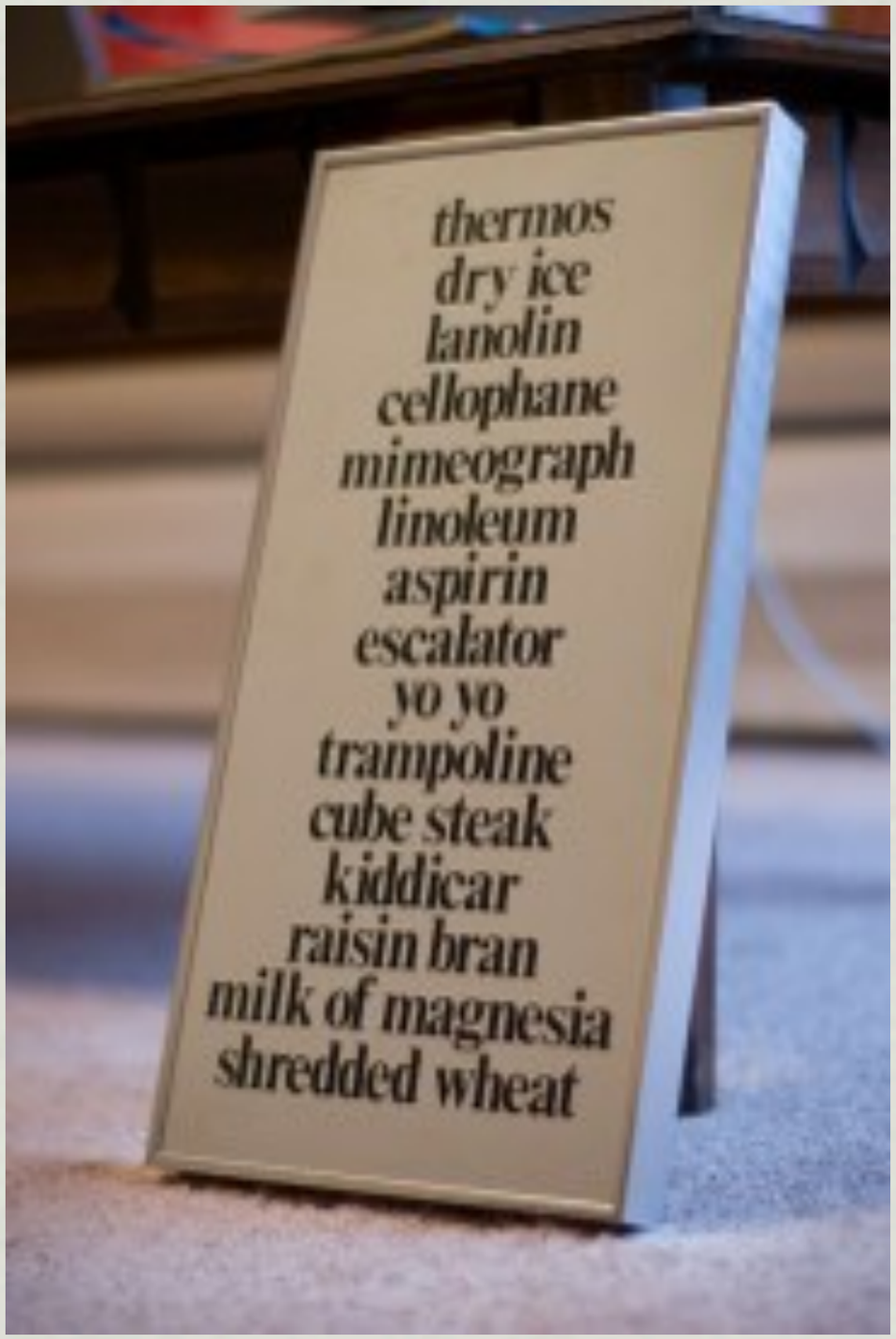
LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

LZ78

dictionary compression

find longest match; useful patterns grow in dictionary



FLICKR: @S4XTON

LZ78

when dictionary is “too full”,
throw it out and start over

can run in constant space

Variants

many things in common use are variations on LZ77 or LZ78

often combined with Huffman Coding, or with other simple adaptations

LZSS (1982)

sliding window based

Lempel-Ziv-Storer-Szymanski

only substitutions that break even

single bit markers

LZW (1984)

Lempel-Ziv-Welch

a better LZ78 (dictionary-based)

starts w/ a smart default dictionary

LZW (1984)

nasty patent situation

(expired, as of June 2003)

DEFLATE:
LZSS + Huffman

PKZIP. Also, gzip.

Lossless Compression

Run-Length Coding

Delta Coding

Huffman Coding

LZ77 family (e.g., LZSS, DEFLATE)

LZ78 family (e.g., LZW)

The Burrows-Wheeler Transform

Transformations

sorted data compresses much better

NORMAL the_cat_in_the_hat

SORTED _____aaceehhhintttt

Transformations

sorted data compresses much better

NORMAL the_cat_in_the_hat

SORTED _____aaceehhhintttt

unfortunately, sorting is a one-way process...

Burrows-Wheeler Transform (1994)

a reversible, partial sort

collates together common substrings

transformed data compresses better

used in bzip

Burrows-Wheeler Transform

repetition

Burrows-Wheeler Transform

```
^repetition
repetition^
epetition^r
petition^re
etition^rep
tition^repe
ition^repet
tion^repeti
ion^repetit
on^repetiti
n^repetitio
```

Burrows-Wheeler Transform

```
^repetition
epetition^r
etition^rep
tion^repet
ion^repetit
ition^repet
n^repetitio
on^repetiti
petition^re
repetition^
tion^repeti
tition^repe
```

Burrows-Wheeler Transform

```
..........n          nrpttoie^ie
..........r
..........p
..........t
..........t
..........o
..........i
..........e
..........^
..........i
..........e
```

Burrows-Wheeler Transform

```
..........^
.........e
.........e
.........i
.........i
.........n
.........o
.........p
.........r
.........t
.........t
```

Burrows-Wheeler Transform

```
.........n^
.........re
.........pe
.........ti
.........ti
.........on
.........io
.........ep
.........^r
.........it
.........et
```

Burrows-Wheeler Transform

```
.........^r
.........ep
.........et
.........io
.........it
.........n^
.........on
.........pe
.........re
.........ti
.........ti
```

Burrows-Wheeler Transform

```
........n^r
........rep
........pet
........tio
........tit
........on^
........ion
........epe
........^re
........iti
........eti
```

Burrows-Wheeler Transform

```
........^re
.......epe
.......eti
.......ion
.......iti
.......n^r
.......on^
.......pet
.......rep
.......tio
.......tit
```

Burrows-Wheeler Transform

```
.......n^re
.......repe
.......peti
.......tion
.......titi
.......on^r
.......ion^
.......epet
.......^rep
.......itio
.......etit
```

Burrows-Wheeler Transform

```
.......^rep
.......epet
.......etit
.......ion^
.......itio
.......n^re
.......on^r
.......peti
.......repe
.......tion
.......titi
```

Burrows-Wheeler Transform

```
......n^rep
......repet
......petit
......tion^
......titio
......on^re
......ion^r
......epeti
......^repe
......ition
......etiti
```

Burrows-Wheeler Transform

```
......^repe
......epeti
......etiti
......ion^r
......ition
......n^rep
......on^re
......petit
......repet
......tion^
......titio
```

Burrows-Wheeler Transform

```
.....n^repe
.....repeti
.....petiti
.....tion^r
.....tition
.....on^rep
.....ion^re
.....epetit
.....^repet
.....ition^
.....etitio
```

Burrows-Wheeler Transform

```
.....^repet
.....epetit
.....etitio
.....ion^re
.....ition^
.....n^repe
.....on^rep
.....petiti
.....repeti
.....tion^r
.....tition
```

Burrows-Wheeler Transform

```
....n^repet
....repetit
....petitio
....tion^re
....tition^
....on^repe
....ion^rep
....epetiti
....^repeti
....ition^r
....etition
```

Burrows-Wheeler Transform

```
....^repeti
....epetiti
....etition
....ion^rep
....ition^r
....n^repet
....on^repe
....petitio
....repetit
....tion^re
....tition^
```

Burrows-Wheeler Transform

```
...n^repeti
...repetiti
...petition
...tion^rep
...tition^r
...on^repet
...ion^repe
...epetitio
...^repetit
...ition^re
...etition^
```

Burrows-Wheeler Transform

```
...^repetit
...epetitio
...etition^
...ion^repe
...ition^re
...n^repeti
...on^repet
...petition
...repetiti
...tion^rep
...tition^r
```

Burrows-Wheeler Transform

```
..n^repetit
..repetitio
..petition^
..tion^repe
..tition^re
..on^repeti
..ion^repet
..epetition
..^repetiti
..ition^rep
..etition^r
```

Burrows-Wheeler Transform

```
..^repetiti
..epetition
..etition^r
..ion^repet
..ition^rep
..n^repetit
..on^repeti
..petition^
..repetitio
..tion^repe
..tition^re
```

Burrows-Wheeler Transform

```
.n^repetiti
.repetition
.petition^r
.tion^repet
.tition^rep
.on^repetit
.ion^repeti
.epetition^
.^repetitio
.ition^repe
.etition^re
```

Burrows-Wheeler Transform

```
.^repetitio
.epetition^
.etition^re
.ion^repeti
.ition^repe
.n^repetiti
.on^repetit
.petition^r
.repetition
.tion^repet
.tition^rep
```

Burrows-Wheeler Transform

```
n^repetitio
repetition^
petition^re
tion^repeti
tition^repe
on^repetiti
ion^repetit
epetition^r
^repetition
ition^repet
etition^rep
```

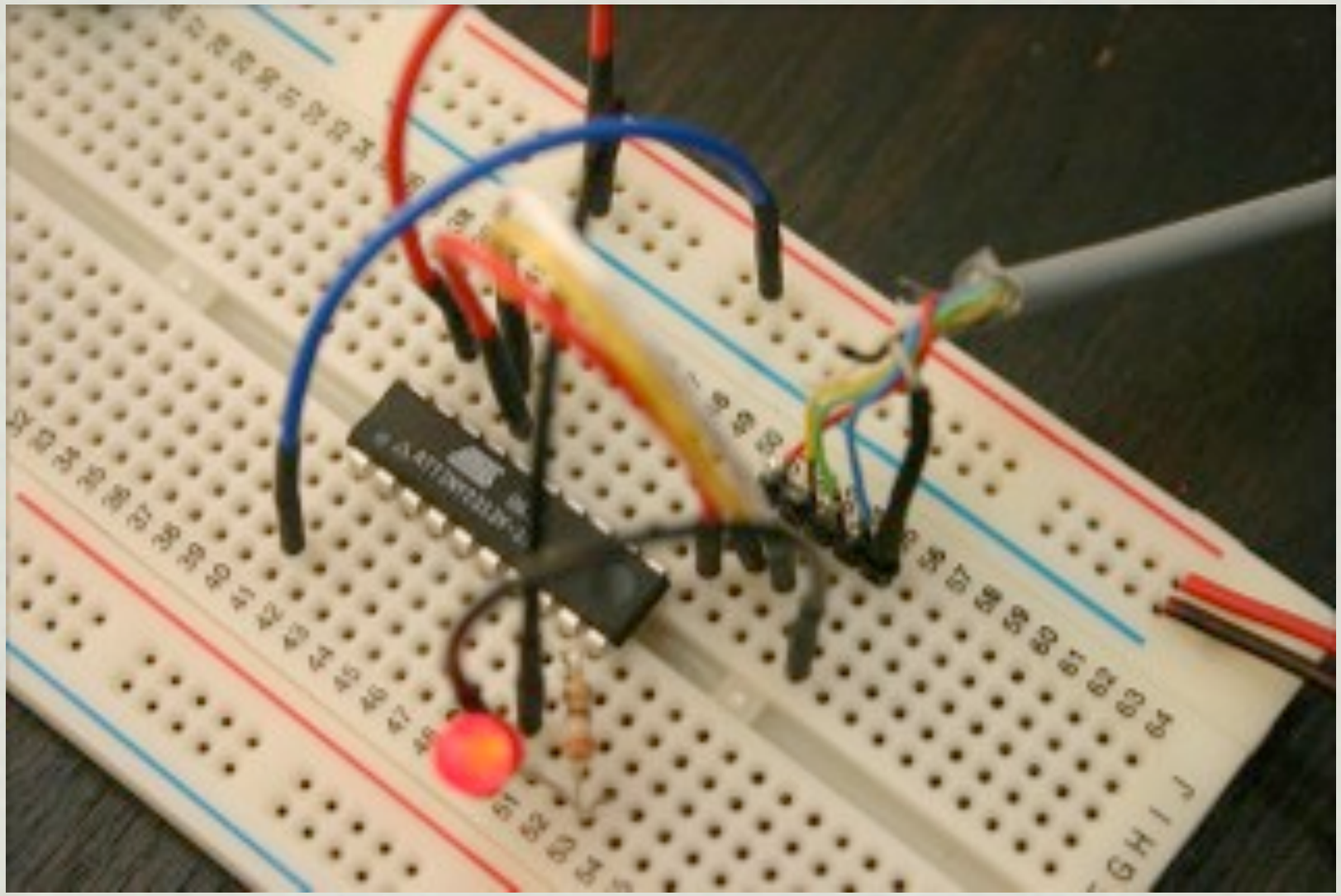
Burrows-Wheeler Transform

```
^repetition
epetition^r
etition^rep
tion^repet
ion^repetit
ition^repet
n^repetitio
on^repetiti
petition^re
repetition^
tion^repeti
tition^repe
```

Burrows-Wheeler Transform

```
^repetition
```

Compression for embedded?



FLICKR: @BARNOID

heatshrink

FLICKR: @MIGHTYOHM

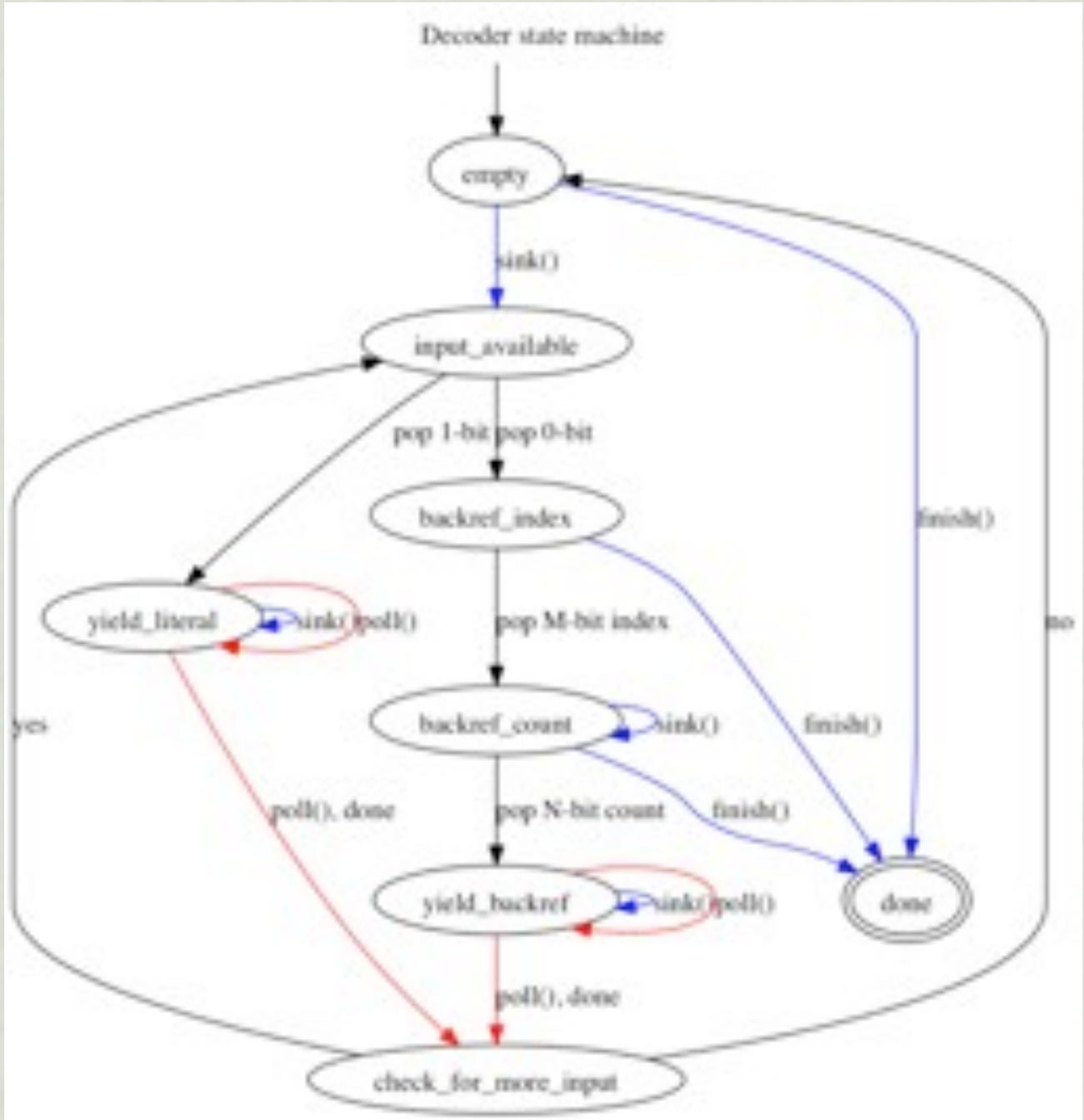
LZSS (sliding window)

hard real-time:
suspend/resume at any bit of I/O

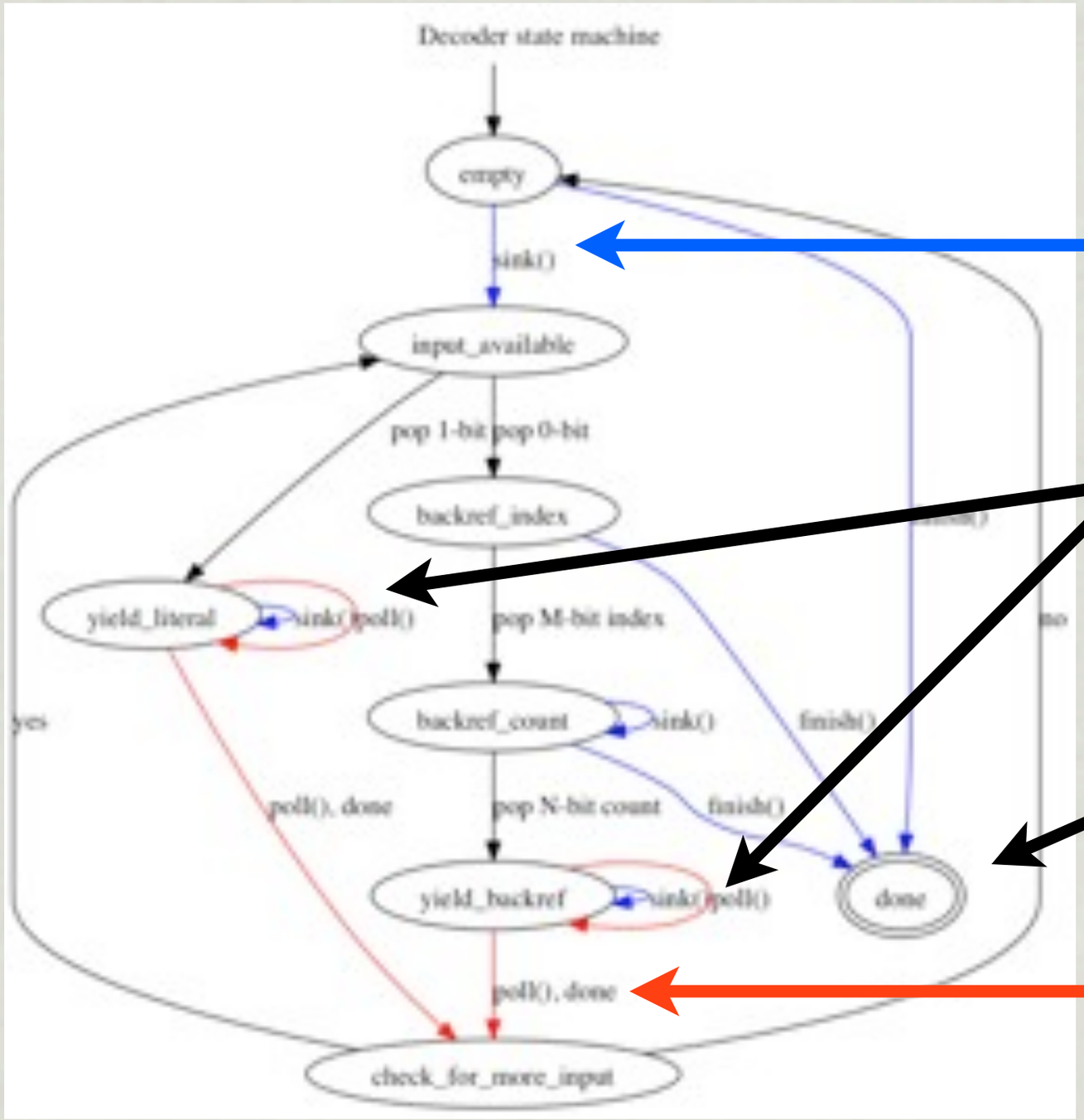
decompress in < 50 bytes RAM
compress in < 100 bytes RAM

BSD-style license

LZSS, for embedded.



LZSS, for embedded.



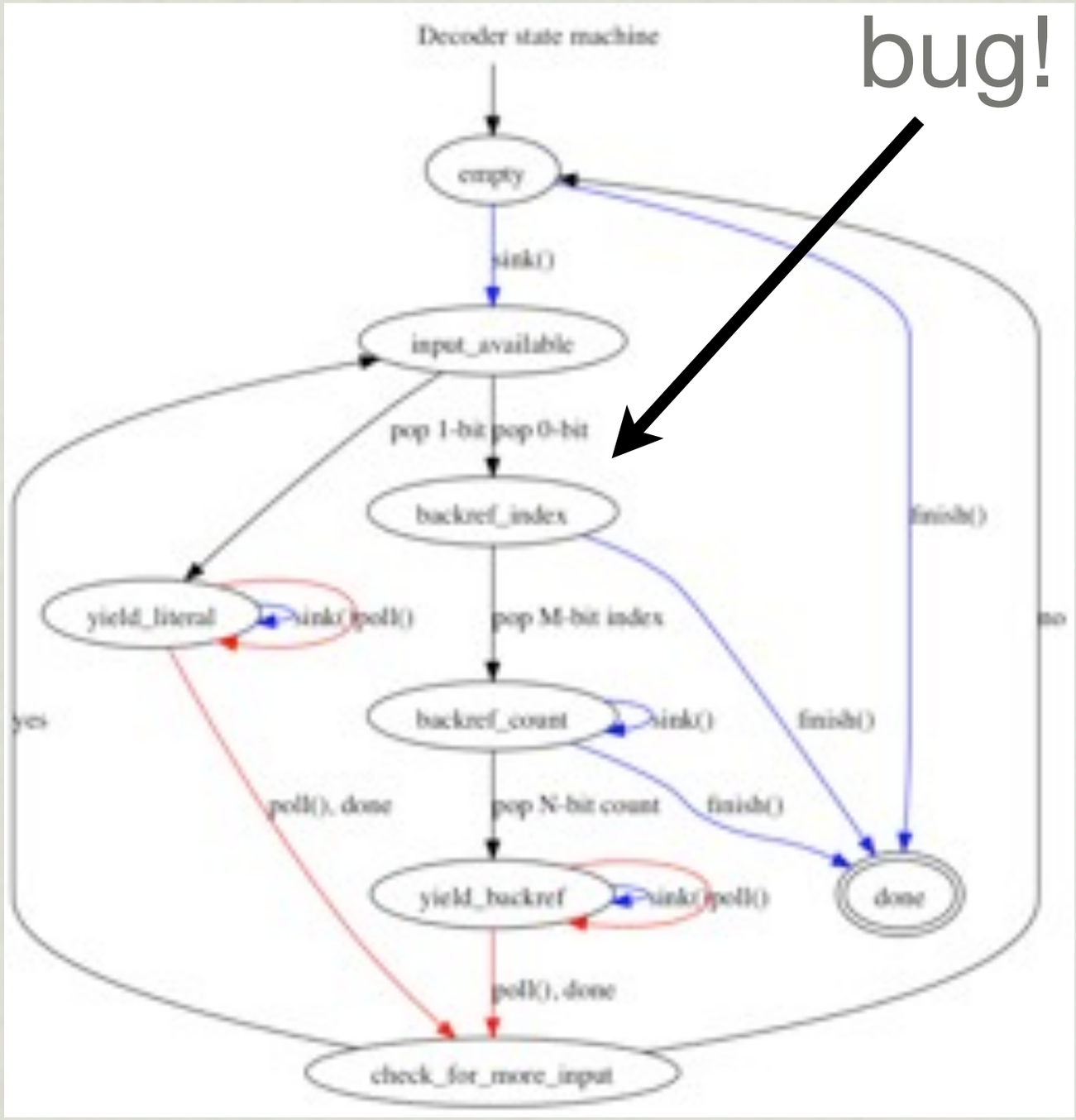
input

suspend / resume loops

done

output

LZSS, for embedded.



heatshrink (LZSS) demo

g. You may not use or otherwise export or re-export the Licensed Application except as authorized by United States law and the laws of the jurisdiction in which the Licensed Application was obtained. In particular, but without limitation, the Licensed Application may not be exported or re-exported (a) into any U.S.-embargoed countries or (b) to anyone on the U.S. Treasury Department's Specially Designated Nationals List or the U.S. Department of Commerce Denied Persons List or Entity List. By using the Licensed Application, you represent and warrant that you are not located in any such country or on any such list. You also agree that you will not use these products for any purposes prohibited by United States law, including, without limitation, the development, design, manufacture, or production of **nuclear, missile, or chemical or biological weapons.**

heatshrink (LZSS) demo

g#######-##########U##########isdi####w####,###########(a)##U.S.-embargo##(b)####20T#s####'####Design#N#s######Commer#D####E##################l############h#######d#lop###,#uf#u###nuc###l#hem###og#wea#

heatshrink (LZSS) demo

g#######-#########U#########isdi####w####,##########(a)##U.S.-embargo##(b)####20T#s####'####Design#N#s######Commer#D####E##################l##########h######d#lop###,#uf#u###nuc###l#hem###og#wea#

Some substitutions:

"the Licensed Application"
"may not be "
"United States law"
"ithout limitation, the "

heatshrink (LZSS) demo

g#######-#########U#########isdi####w####,##########(a)##U.S.-embargo##(b)####20T#s####'####Design#N#s######Commer#D####E##################l##########h######d#lop###,#uf#u###nuc###l#hem###og#wea#

Some substitutions:

"ction "
"ational"
"ed in any "

Lossy Compression

inherently data-specific

smart degradation

usually a quality/size continuum

common in multimedia

Lossy Compression

example: JPEG

51 KB

27 KB

14 KB

Closing

why compression matters

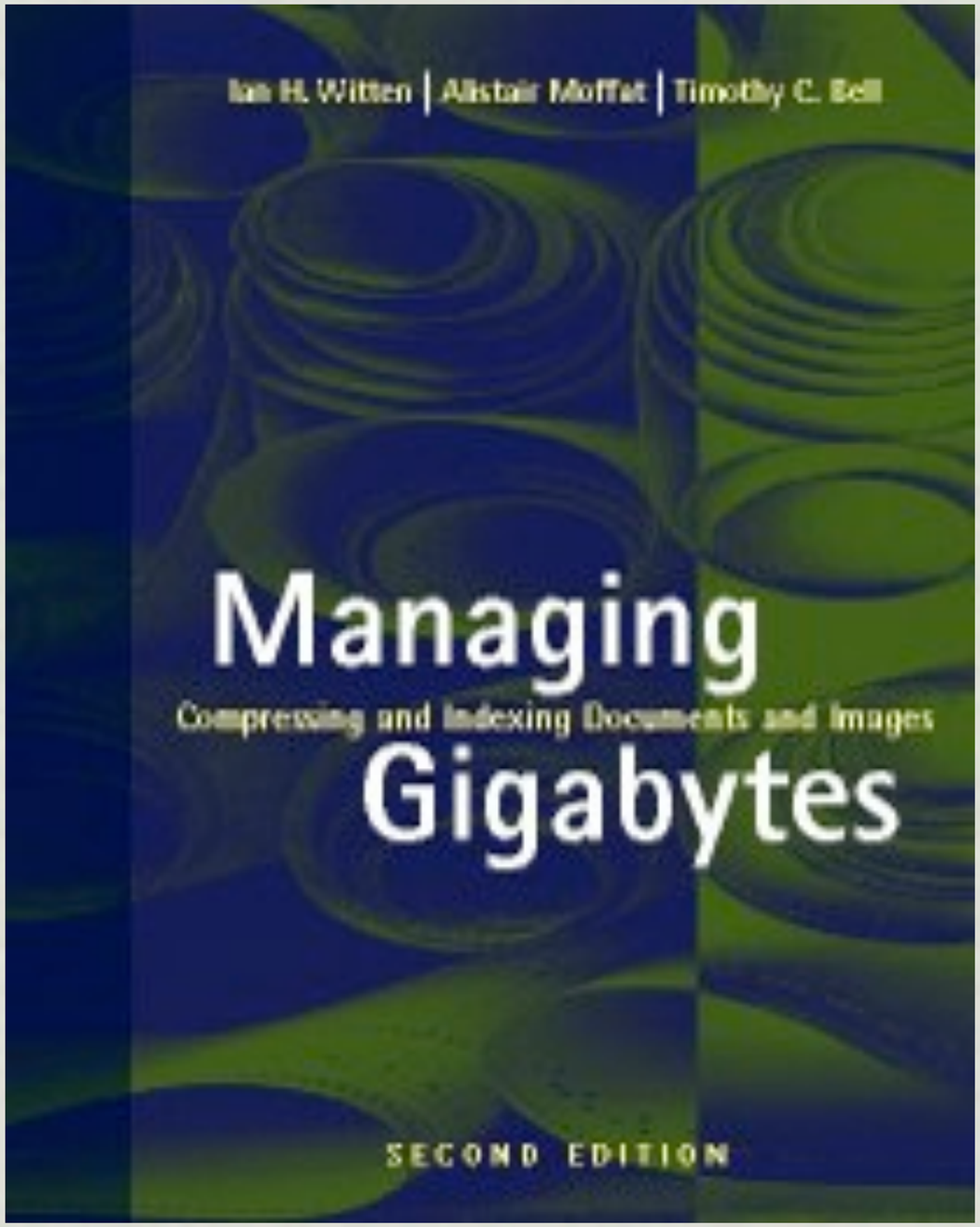
designing for compression

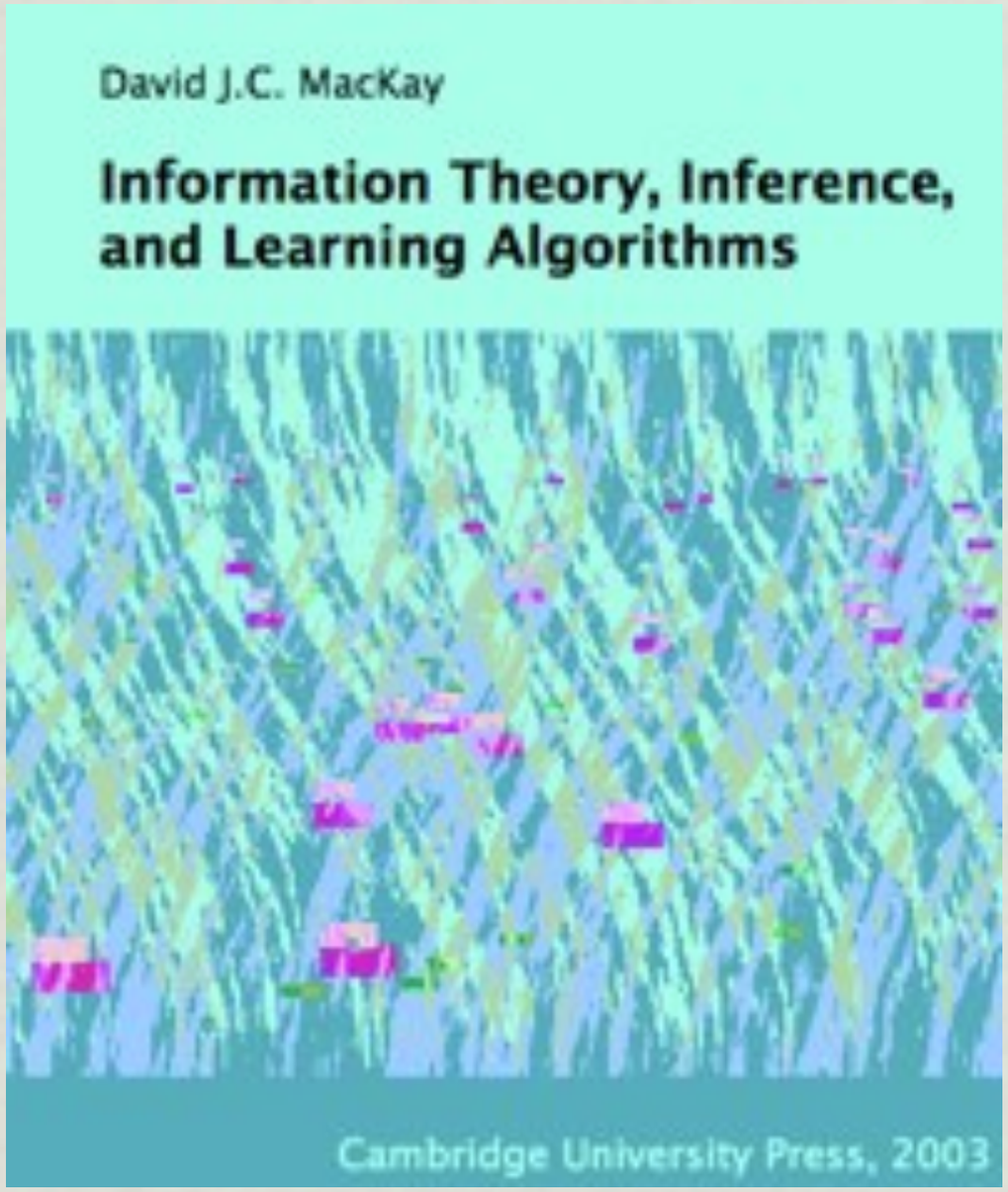
examples of lossless compression

case study: heatshrink

examples of lossy compression

To learn more





We're Hiring!
Detroit & Ann Arbor, MI

Questions?

@silentbicycle

github.com/silentbicycle