EDIS

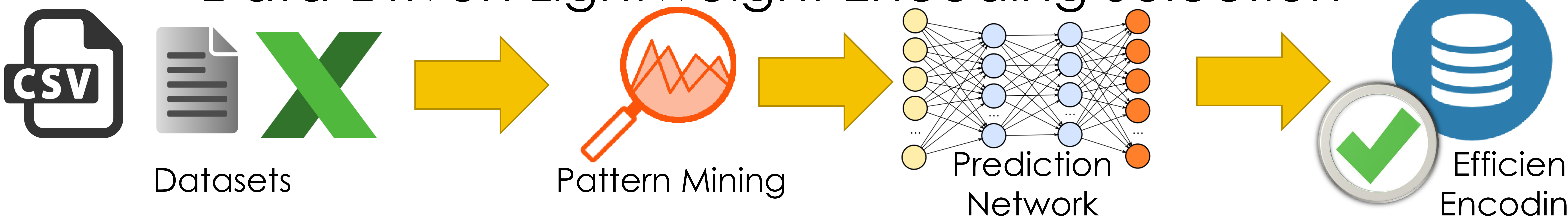
THE UNIVERSITY OF CHICAGO

EDIS: Efficient Data Management for IoT and Data Streams

Hao Jiang, Aaron J. Elmore

CERES Center for Unstoppable Computing

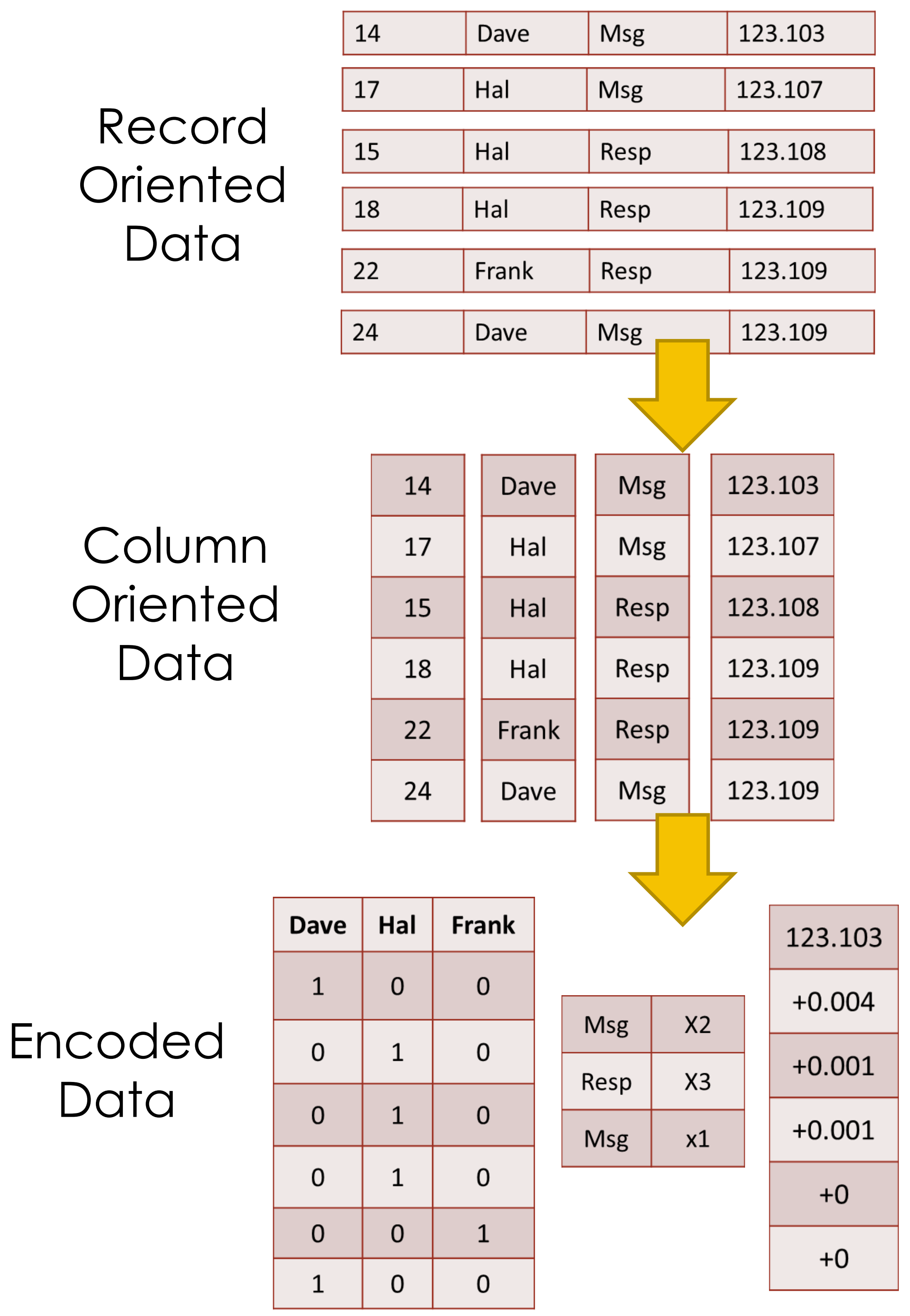
Data-Driven Lightweight Encoding Selection



Datasets → Pattern Mining → Prediction Network → Efficient Encoding

What is Encoding?

Space saving and query efficient data representation, which is used in Dremel, Carbon, Parquet, Vertica, etc.



Record Oriented Data

14	Dave	Msg	123.103
17	Hal	Msg	123.107
15	Hal	Resp	123.108
18	Hal	Resp	123.109
22	Frank	Resp	123.109
24	Dave	Msg	123.109

Column Oriented Data

14	Dave	Msg	123.103
17	Hal	Msg	123.107
15	Hal	Resp	123.108
18	Hal	Resp	123.109
22	Frank	Resp	123.109
24	Dave	Msg	123.109

Encoded Data

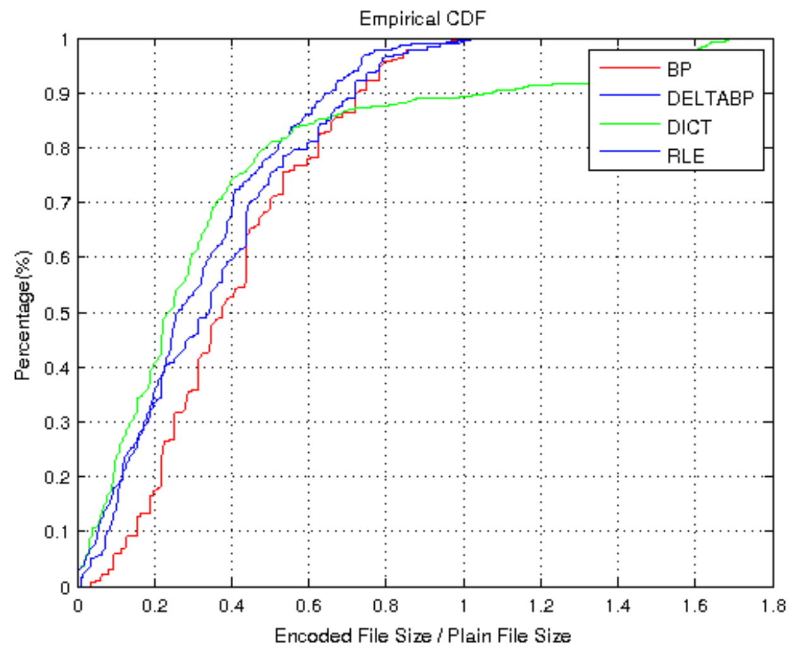
Dave	Hal	Frank
1	0	0
0	1	0
0	1	0
0	1	0
0	0	1
1	0	0

Msg	X2
Resp	X3
Msg	x1

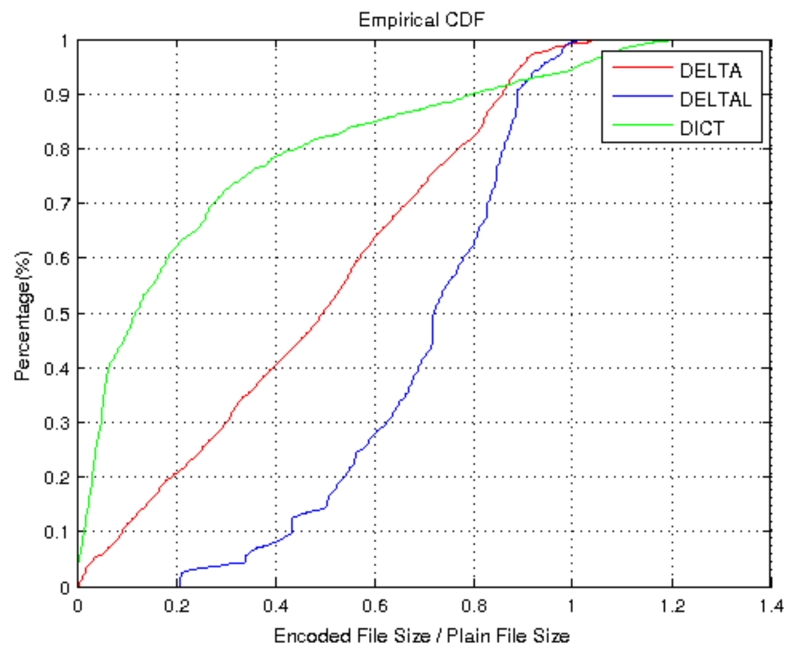
123.103
+0.004
+0.001
+0.001
+0
+0

Dataset Analysis

To study the efficacy of dataset encodings we have collected over 7,000 columns/attributes from 1,200 public datasets, and measure the time and space savings from different encodings and file formats.

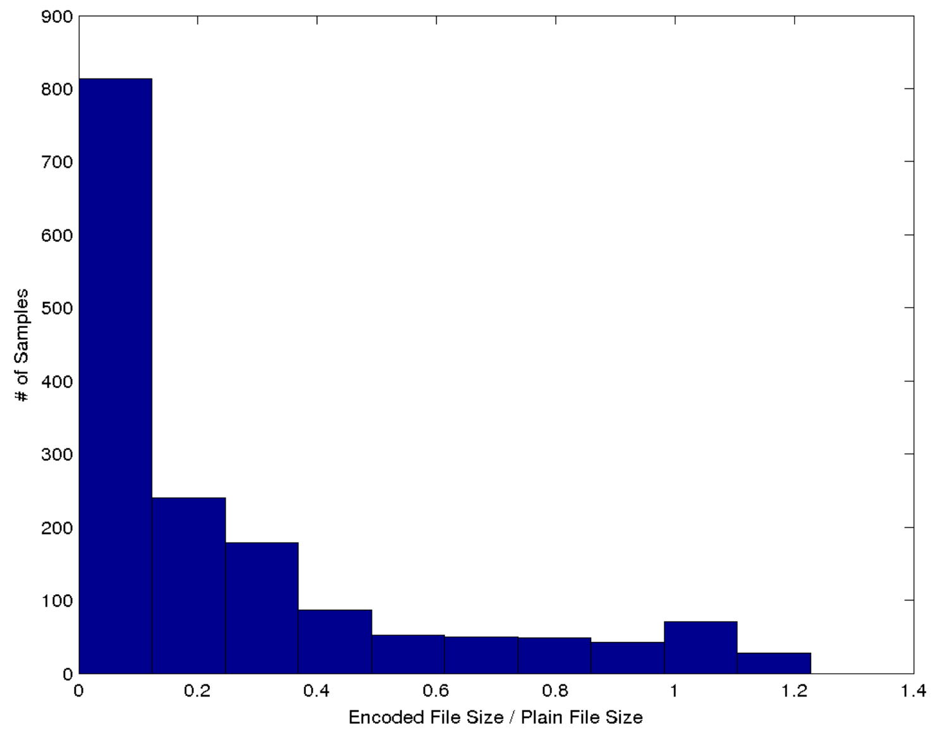


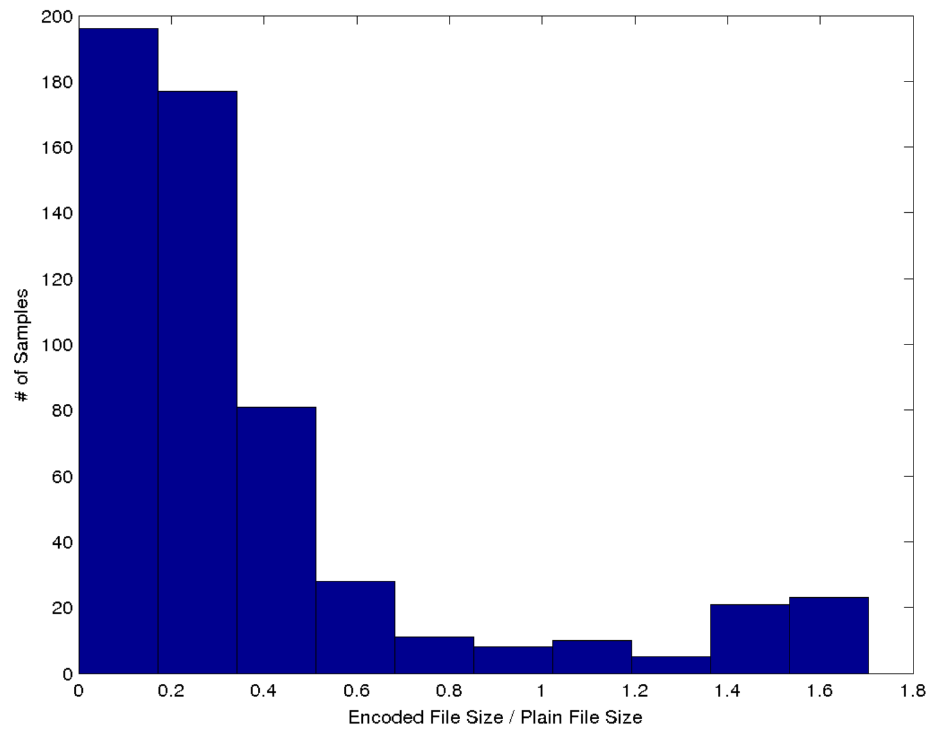
Compression Ratio for Integer Columns



Compression Ratio for String Columns

There is no single best encoding scheme.



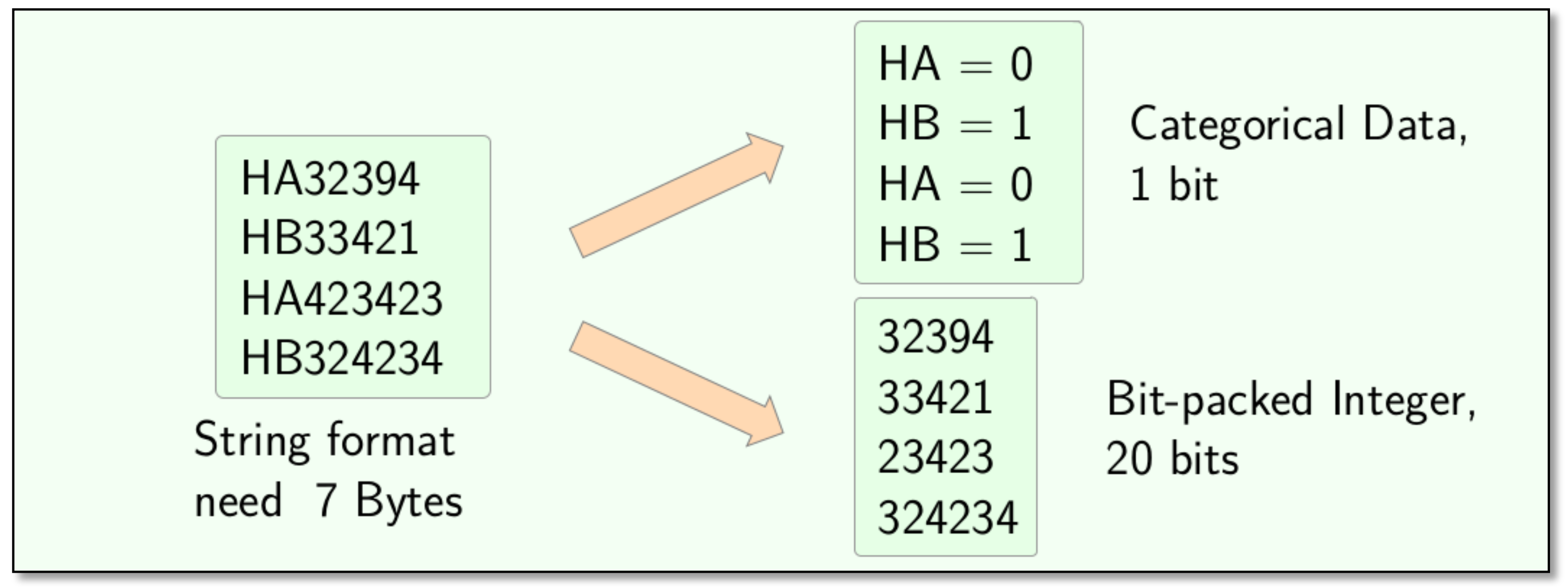


Dictionary Encoding is often adopted as default in many systems, however it performs sub-optimal in many cases

Automated one-pass encoding selector in progress

New Opportunities

Sub-attributes and embedded data does not encode well, can we identify this data and decompose attributes?



HA32394
HB33421
HA423423
HB324234

String format need 7 Bytes

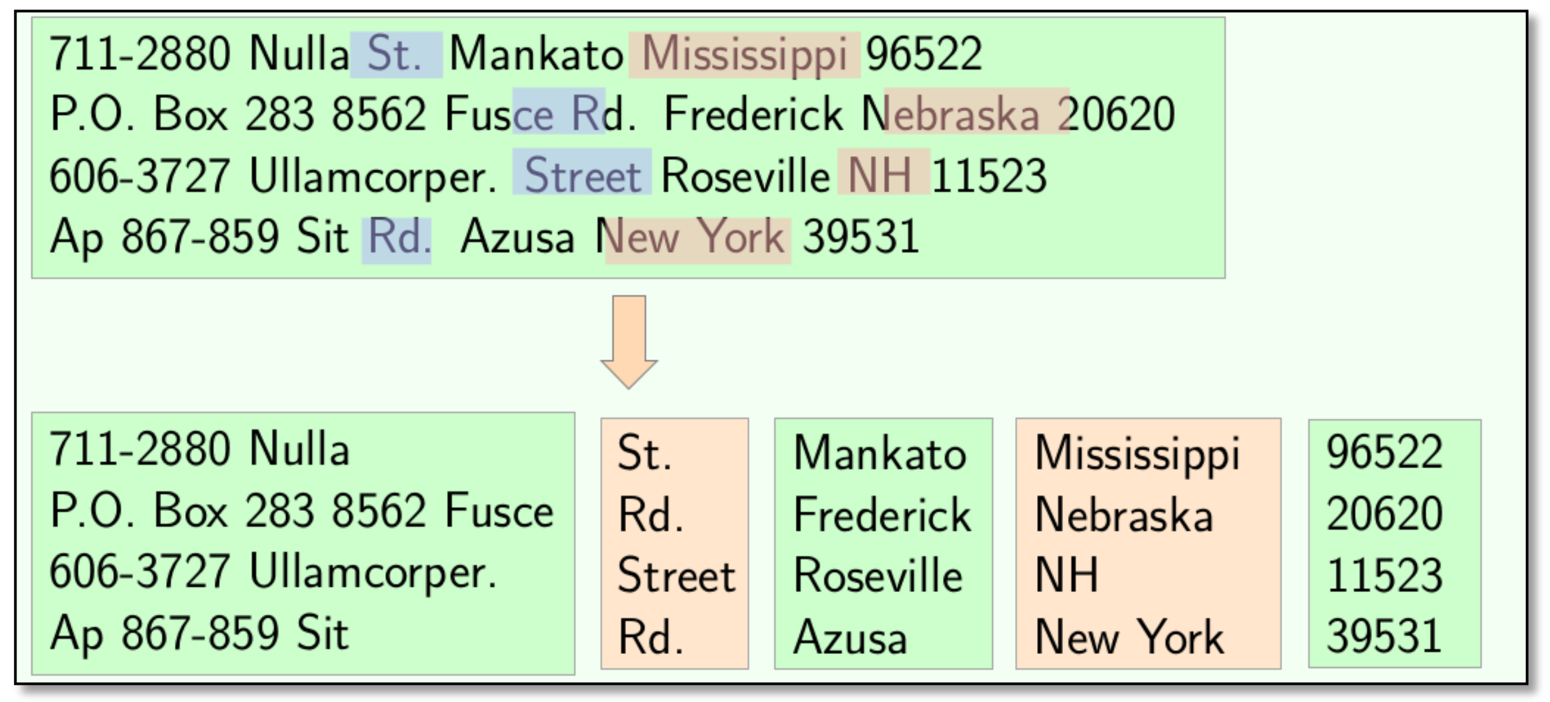
HA = 0
HB = 1
HA = 0
HB = 1

Categorical Data, 1 bit

32394
33421
23423
324234

Bit-packed Integer, 20 bits

Observing the pattern <STR><NUM> from dataset allowing us to separate them and apply more efficient encoding for each part



711-2880 Nulla St. Mankato Mississippi 96522
P.O. Box 283 8562 Fusce Rd. Frederick Nebraska 20620
606-3727 Ullamcorper. Street Roseville NH 11523
Ap 867-859 Sit Rd. Azusa New York 39531

711-2880 Nulla	St.	Mankato	Mississippi	96522
P.O. Box 283 8562 Fusce	Rd.	Frederick	Nebraska	20620
606-3727 Ullamcorper.	Street	Roseville	NH	11523
Ap 867-859 Sit	Rd.	Azusa	New York	39531

Use NLP techniques to look for similar words in text and extract patterns