Sirius AHM October 2015

Mark Ainsworth

Division of Applied Mathematics, Brown University

&

Oak Ridge National Laboratory

Mark_Ainsworth@brown.edu

BROWN

OAK RIDGE National Laboratory

Outline

- Basic concepts from Information Theory
- Some thoughts on compression of scientific data
- Simple example
- Conclusions

BROWN

OAK RIDGE National Laboratory

Basic Quantities in Information Theory

- Data stream S and for $x \in S$ let

$$Pr(X = x) = p_x \in [0, 1]$$

- Shannon Information Content

$$h(x) = -\log_2 p_x$$

- Entropy

$$H(S) = -\sum_{x \in S} p_x \log_2 p_x$$

⇒ Noisy/random data has HIGH ENTROPY

BROWN

OAK RIDGE National Laboratory

Relevance to Data Storage

Source Coding Theorem

N items of data from S require at least $NH(S)$ bits of storage.

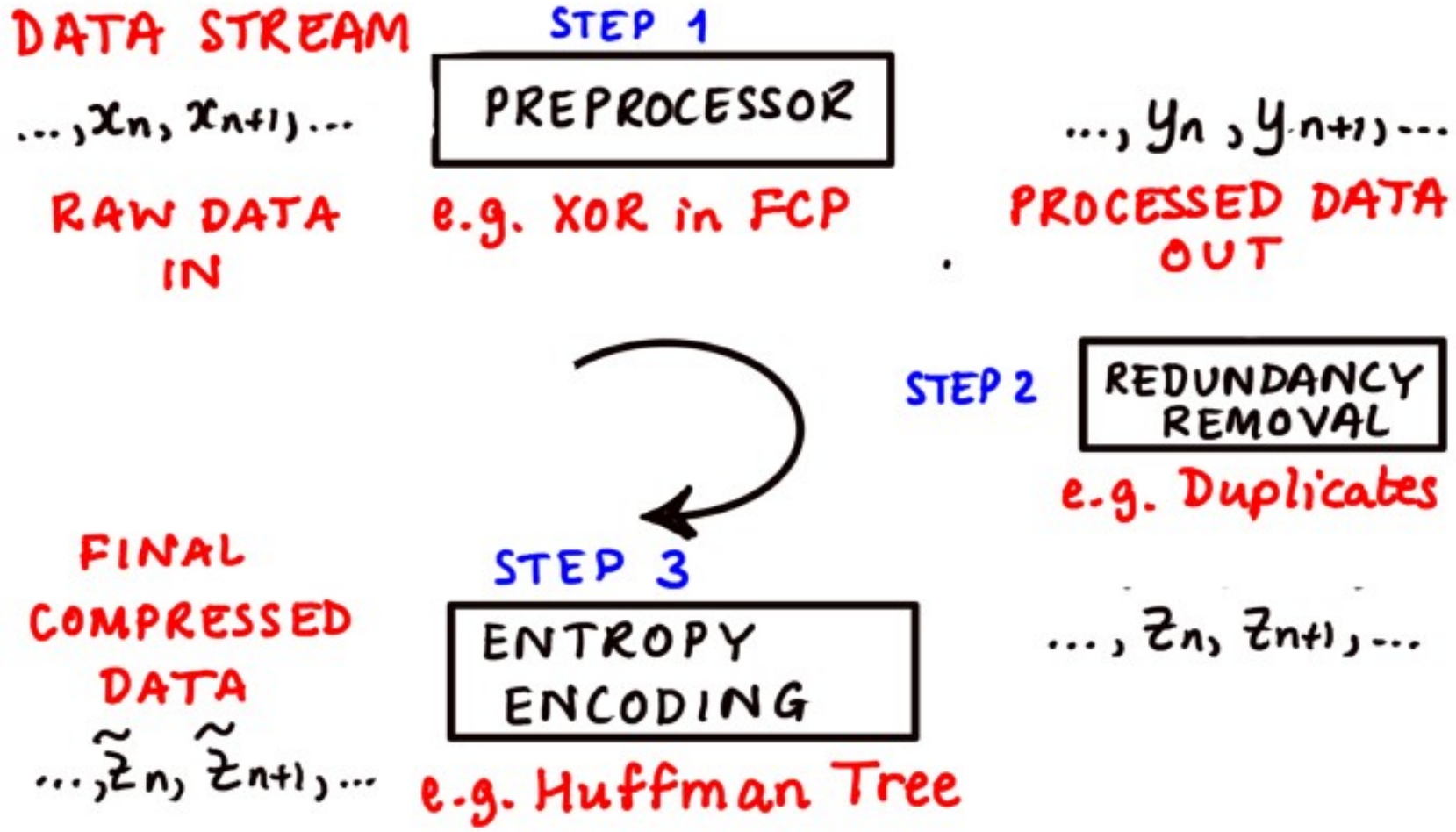
- Information content related to likelihood.
- Noisy/random data has LARGE $H(S)$ and hence ALMOST INCOMPRESSIBLE.

E.G. Gomez & Cappello 2013 show only 15% compression (lossless).

BROWN

OAK RIDGE National Laboratory

Data Compression



Data Compression

- Noisy / random data has HIGH $H(S)$

  $\Rightarrow$ limited scope for lossless compression

- Applying hierarchy of thresholds to data to reduce entropy $H(S)$

- Map onto storage hierarchy

HIGHER FIDELITY ; LESS COMPRESSION

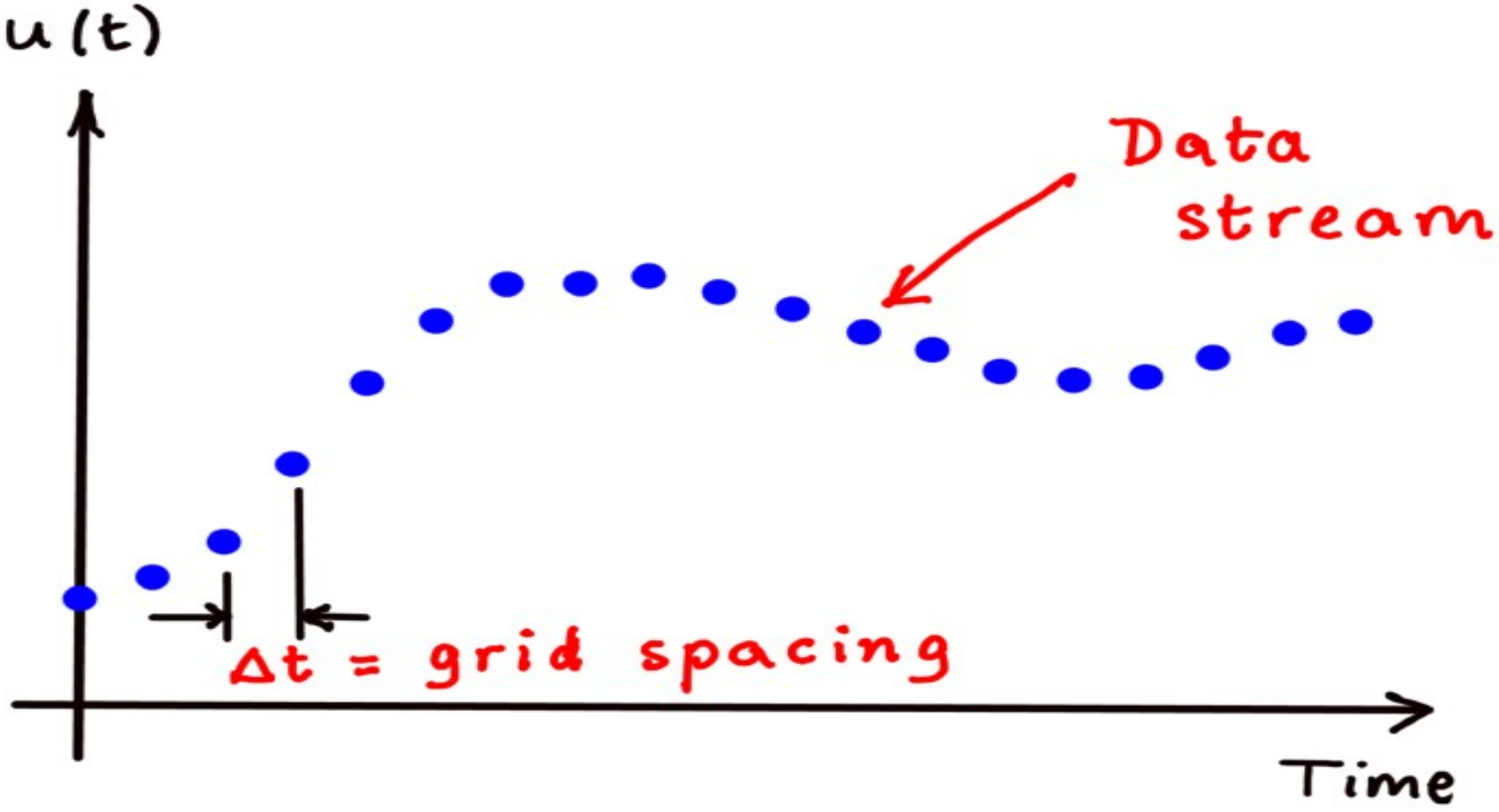
LOWER THRESHOLD
- LOW ENTROPY
- HIGH COMPRESSIBLE

VS

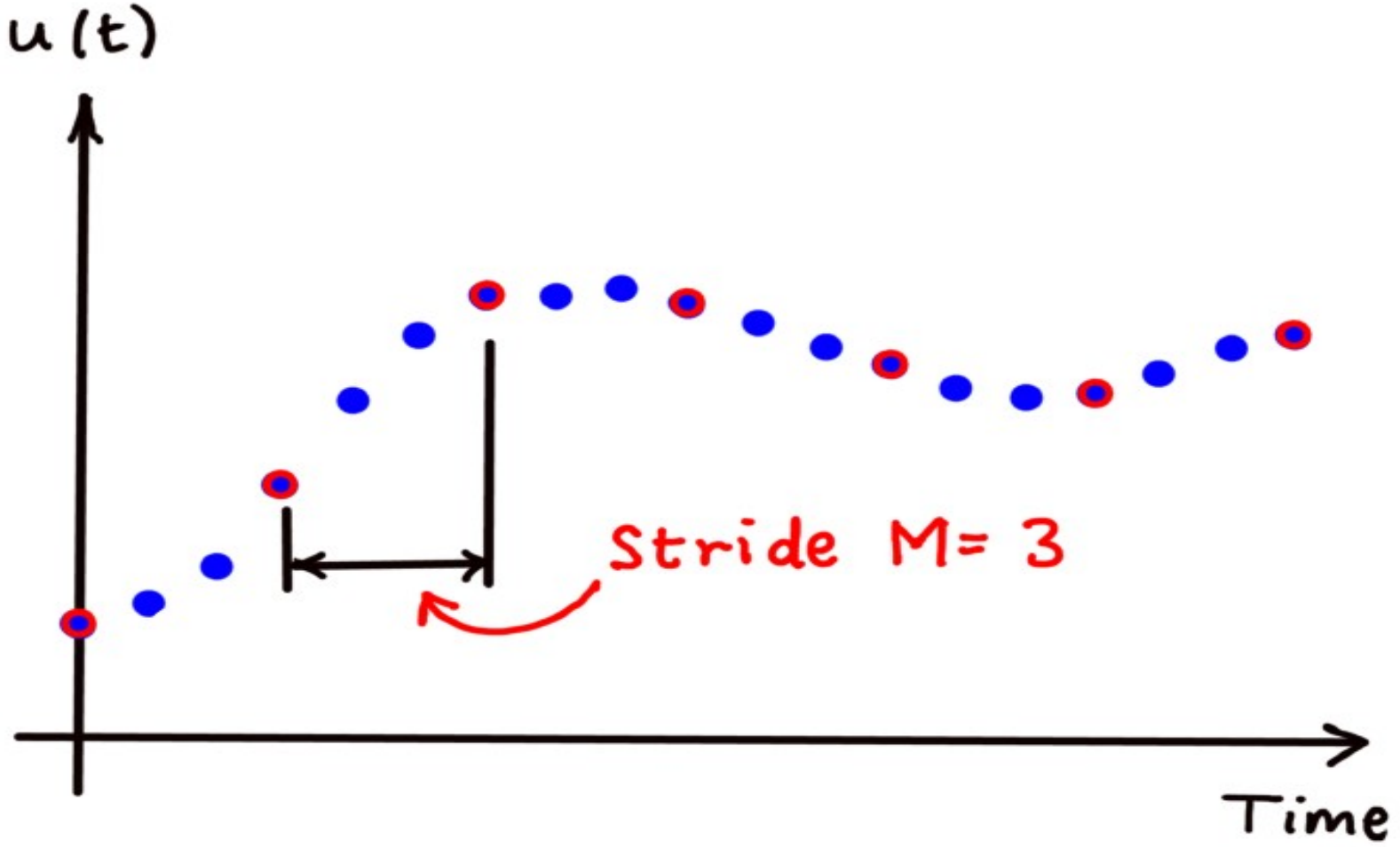
HIGHER THRESHOLD
- ALMOST LOSSLESS
- LOW COMPRESSIBLE

LOWER FIDELITY ; HIGH COMPRESSION.

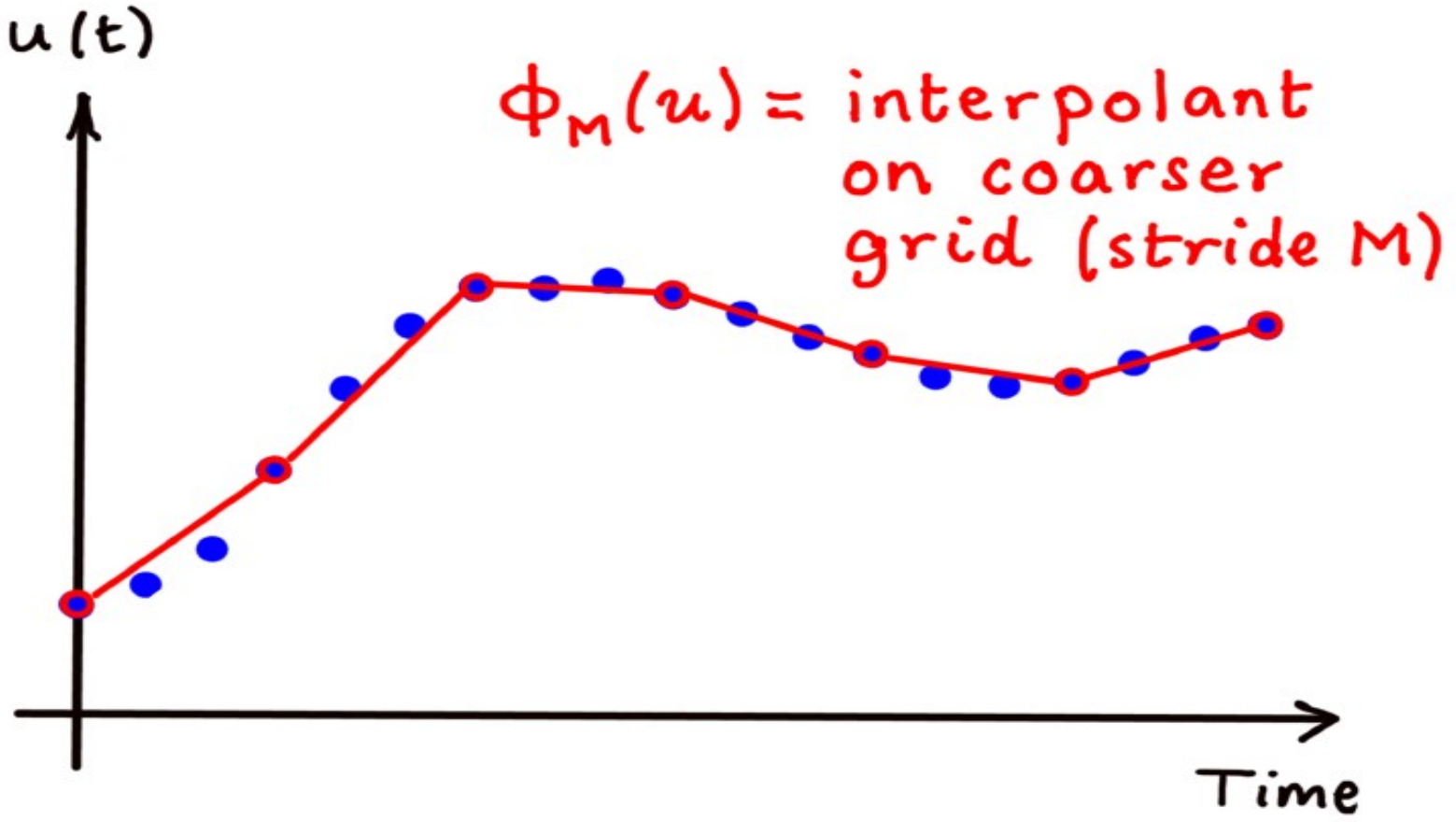
Simple Example



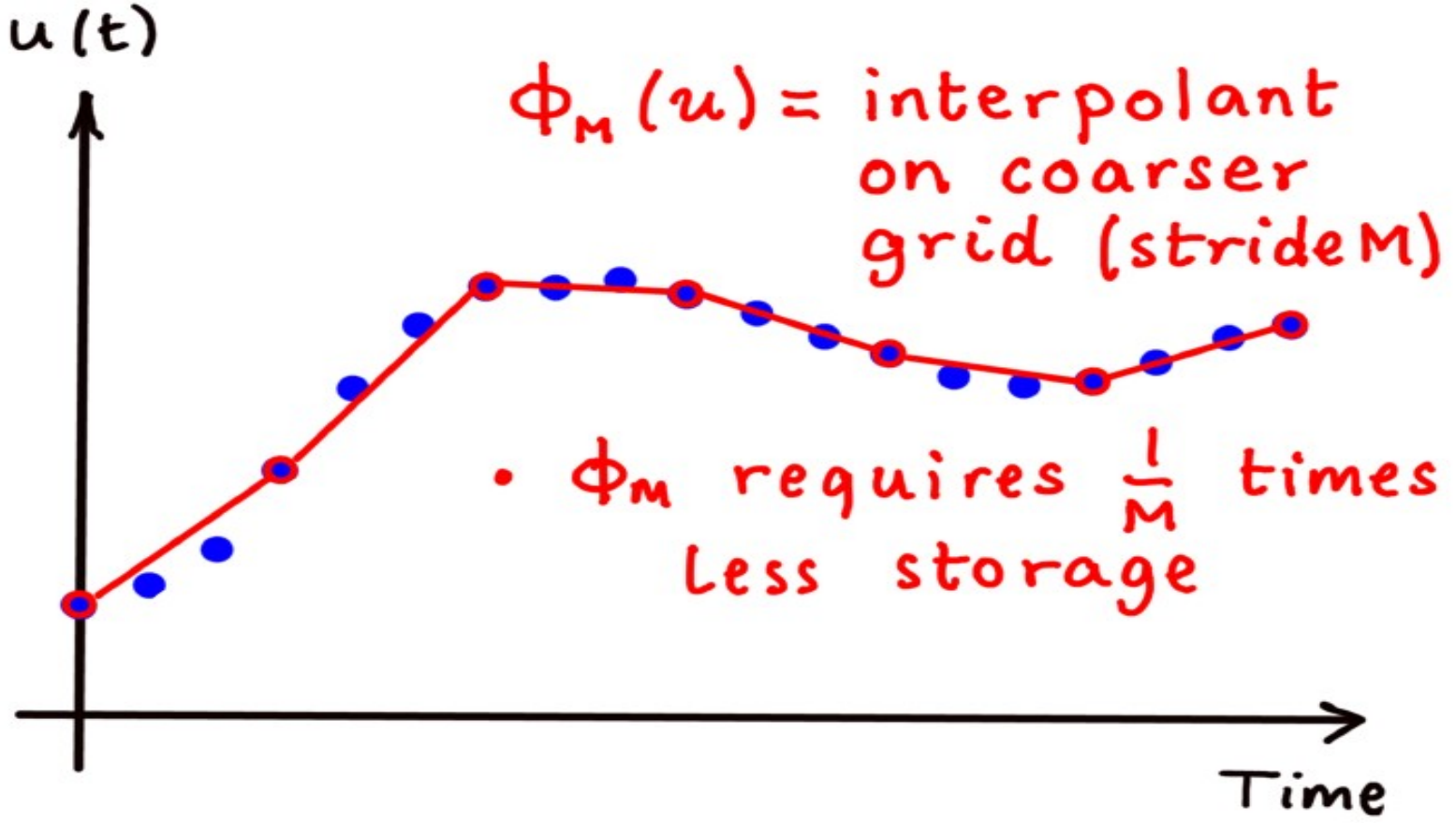
Simple Example



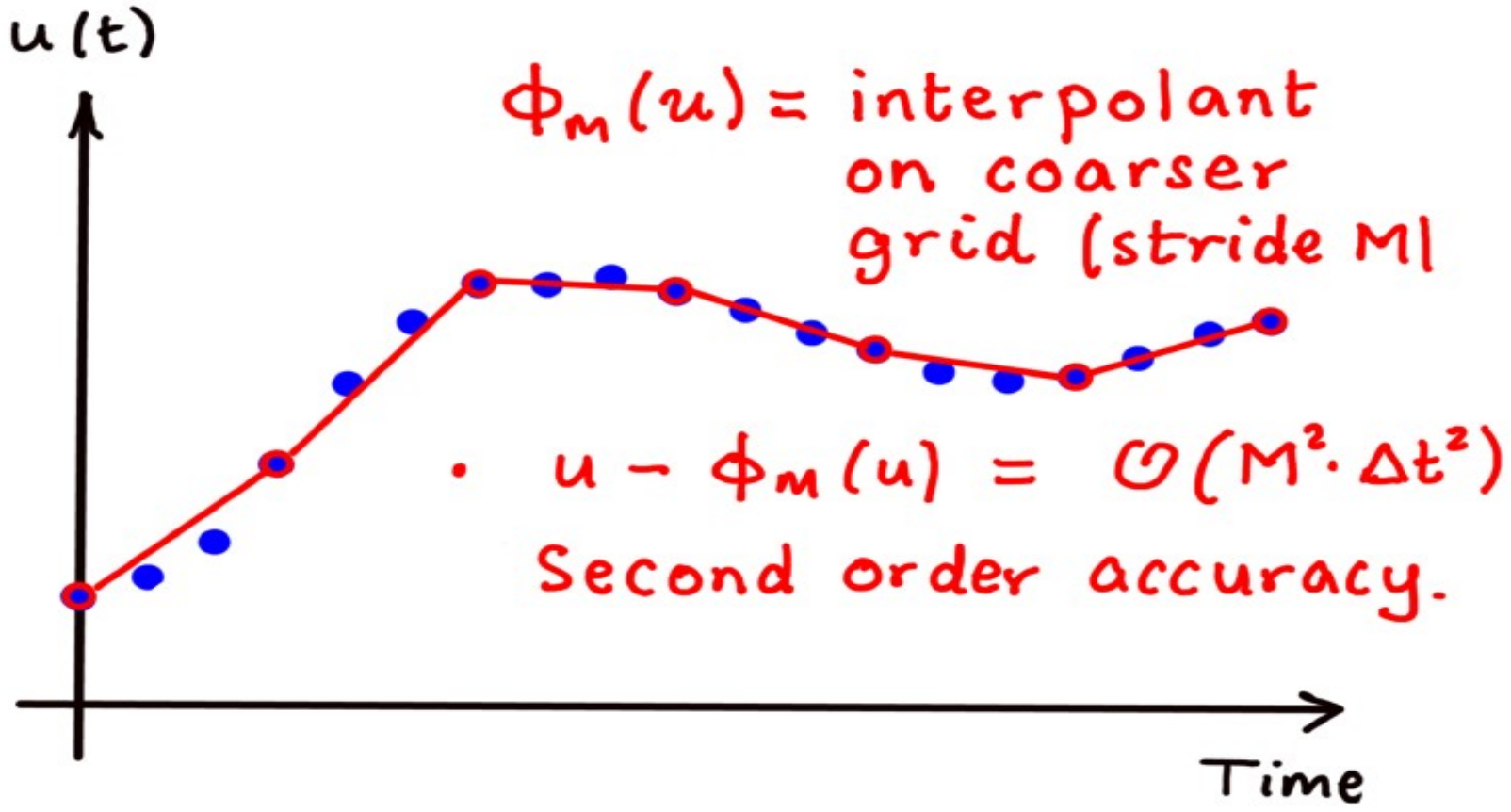
Simple Example



Simple Example

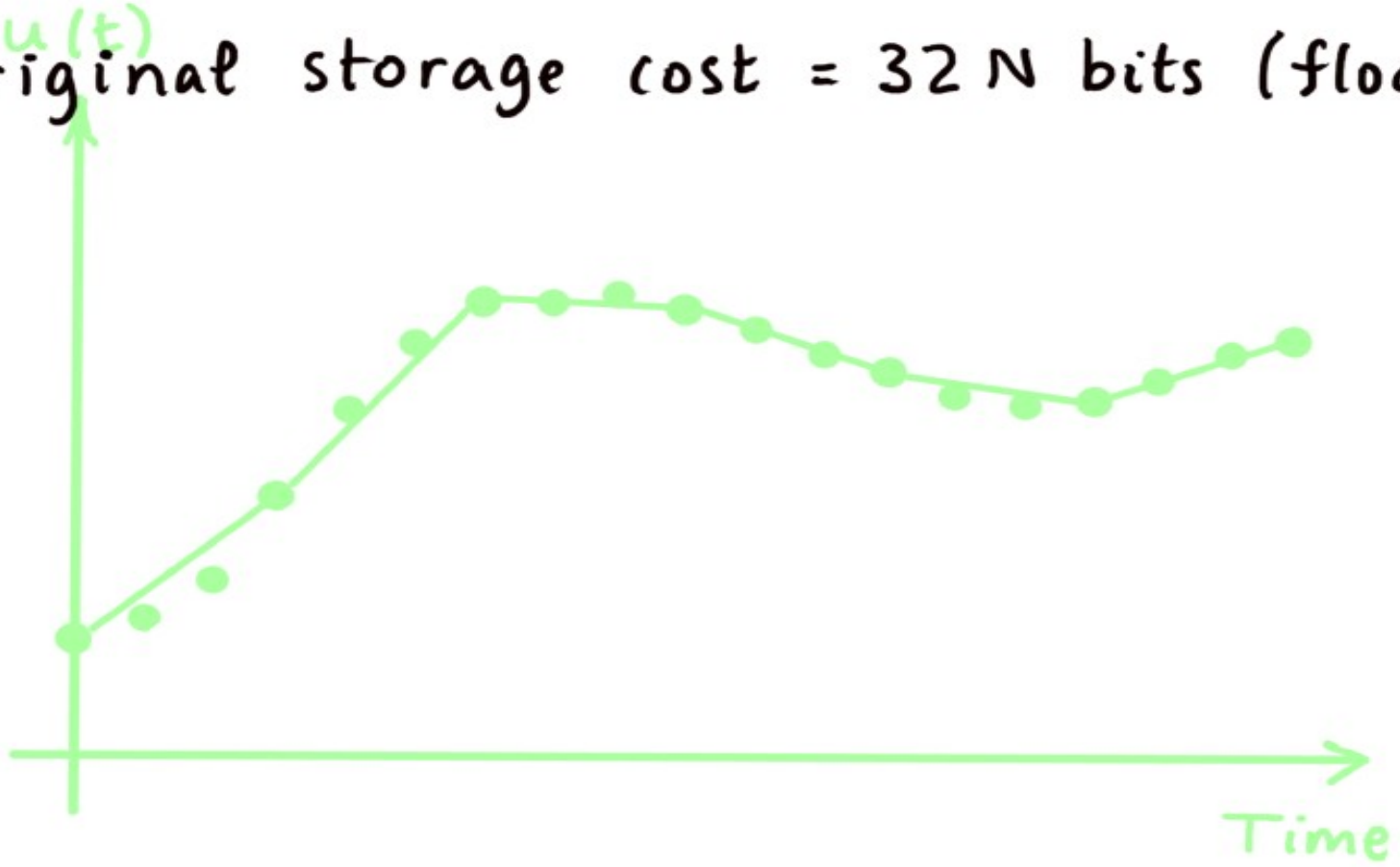


Simple Example



Simple Example

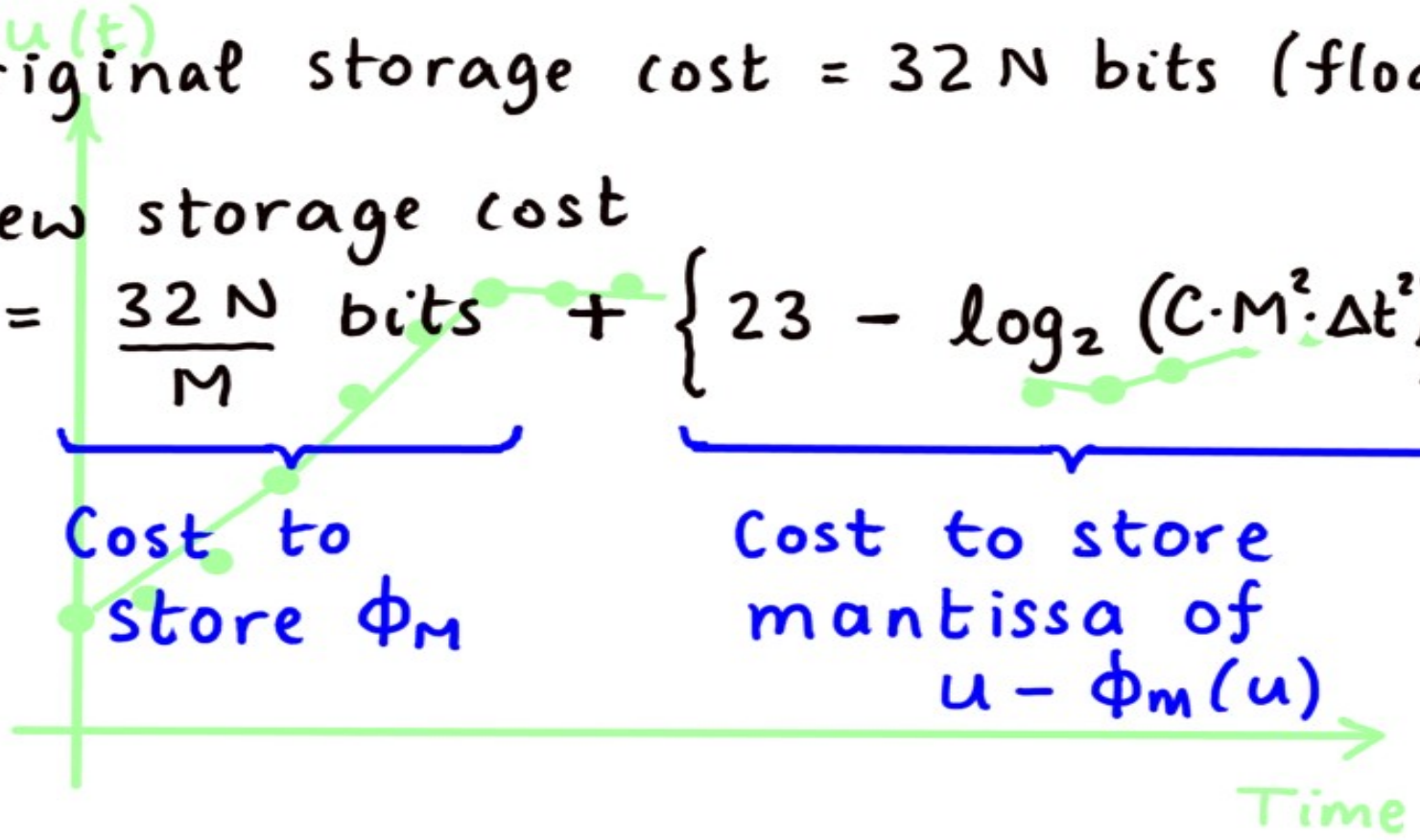
- Original storage cost = 32 N bits (floats)



Simple Example

- Original storage cost = 32 N bits (floats)
- New storage cost

$$= \underbrace{\frac{32\,N}{M}\ \text{bits}}_{\text{Cost to store } \phi_M} + \underbrace{\left\{23 - \log_2\left(C \cdot M^2 \cdot \Delta t^2\right)\right\} N}_{\text{Cost to store mantissa of } u - \phi_M(u)}$$



Simple Example

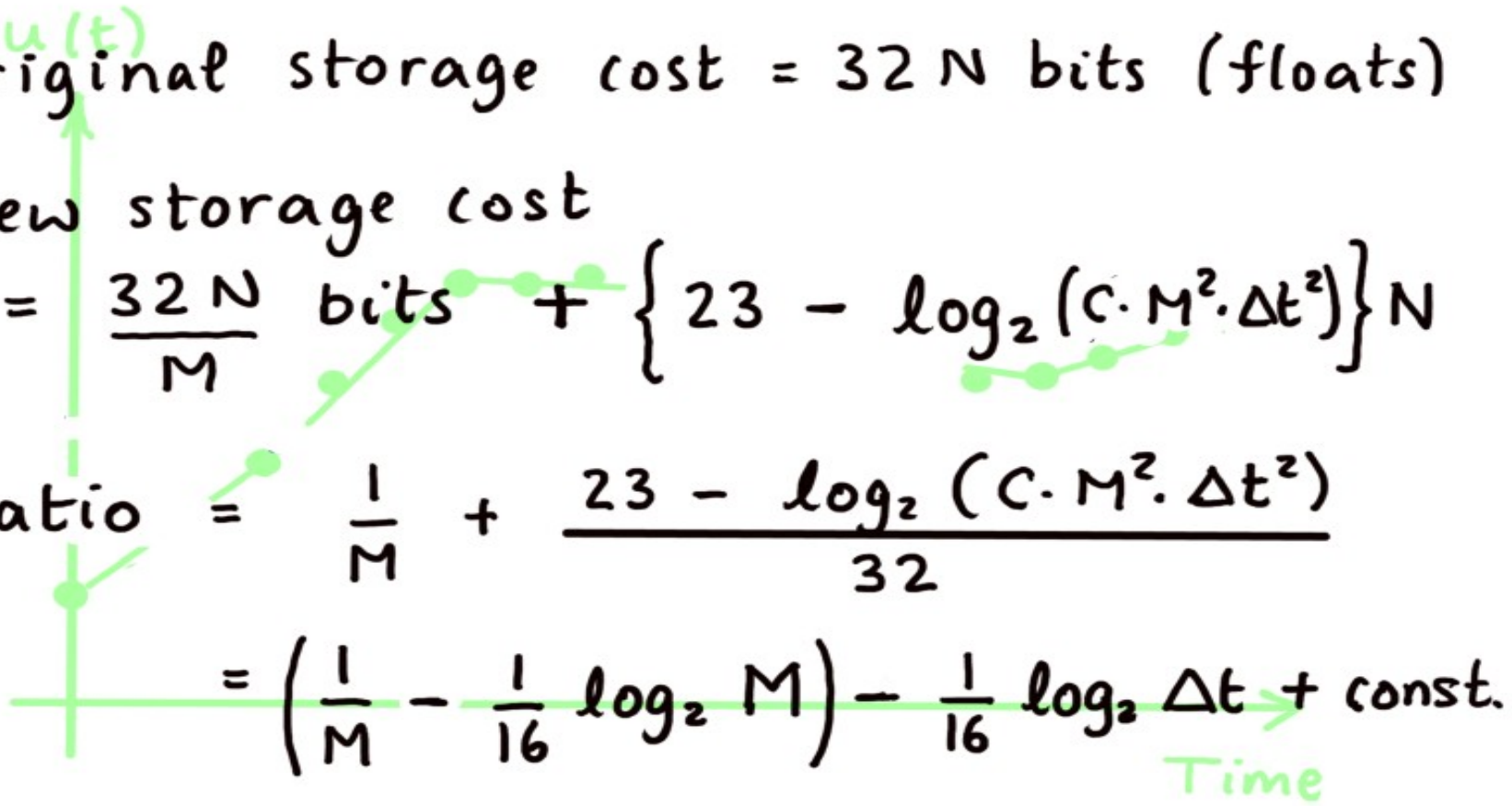
- Original storage cost = $32\,N$ bits (floats)
- New storage cost

$$= \frac{32\,N}{M} \text{ bits} + \left\{23 - \log_2\left(C \cdot M^2 \cdot \Delta t^2\right)\right\} N$$

- Ratio

$$= \frac{1}{M} + \frac{23 - \log_2\left(C \cdot M^2 \cdot \Delta t^2\right)}{32}$$

$$= \left(\frac{1}{M} - \frac{1}{16}\log_2 M\right) - \frac{1}{16}\log_2 \Delta t + \text{const.}$$

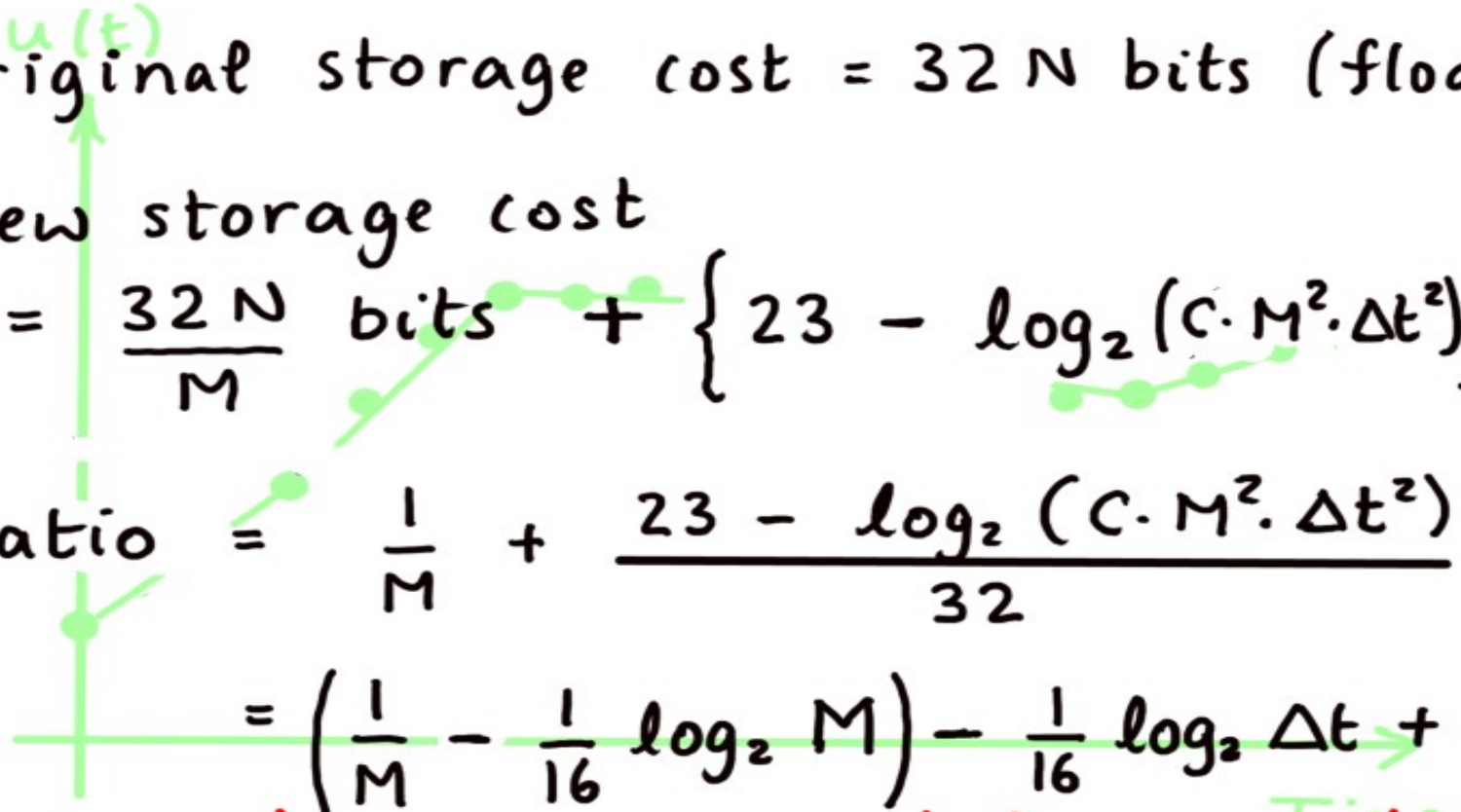


- Original storage cost = $32N$ bits (floats)
- New storage cost

$$= \frac{32N}{M} \text{ bits} + \left\{23 - \log_2\left(C \cdot M^2 \cdot \Delta t^2\right)\right\} N$$

- Ratio $= \frac{1}{M} + \frac{23 - \log_2\left(C \cdot M^2 \cdot \Delta t^2\right)}{32}$

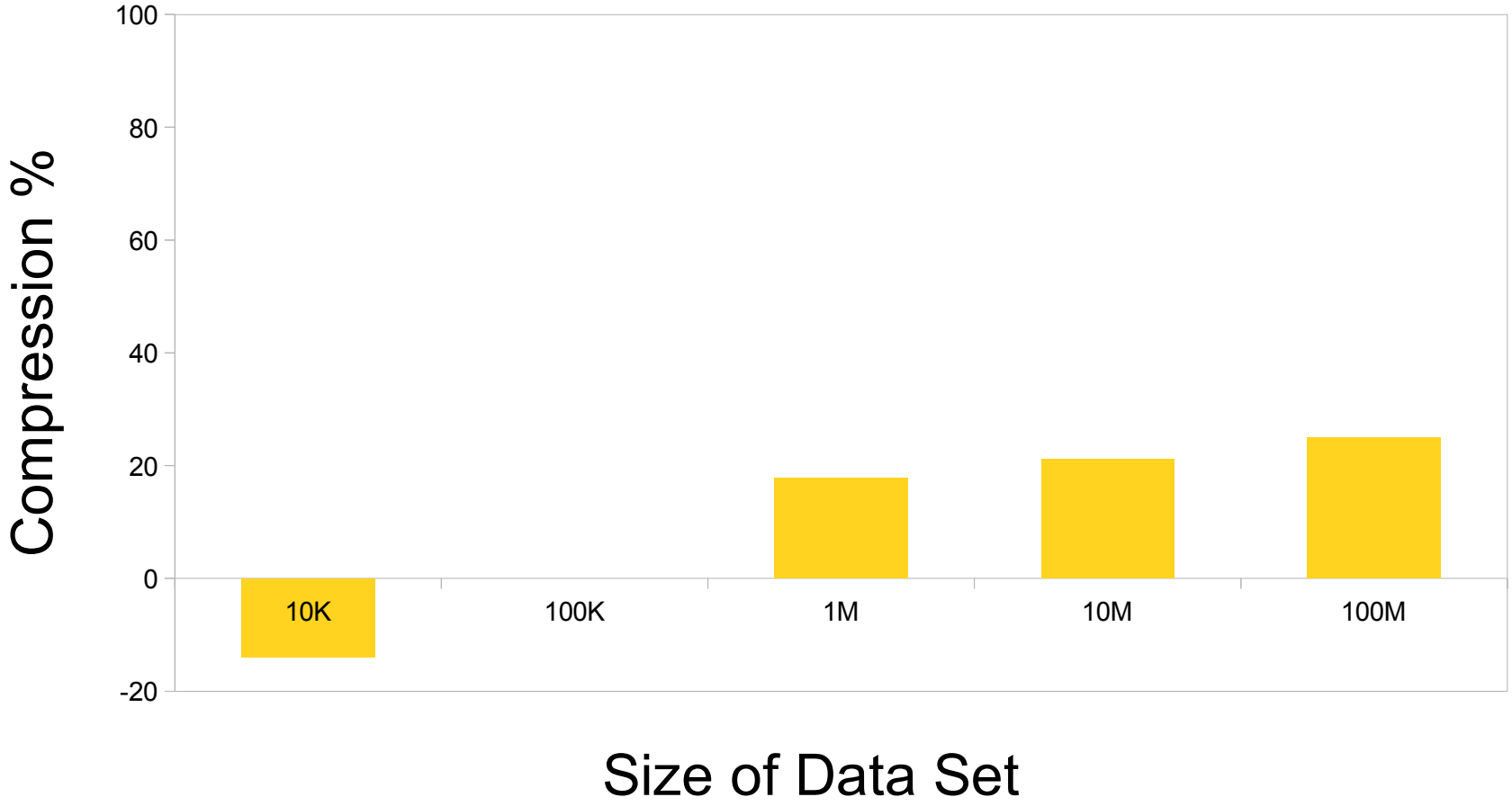
$$= \left(\frac{1}{M} - \frac{1}{16}\log_2 M\right) - \frac{1}{16}\log_2 \Delta t + \text{const.}$$



Min. independent of N, Δt

Benign

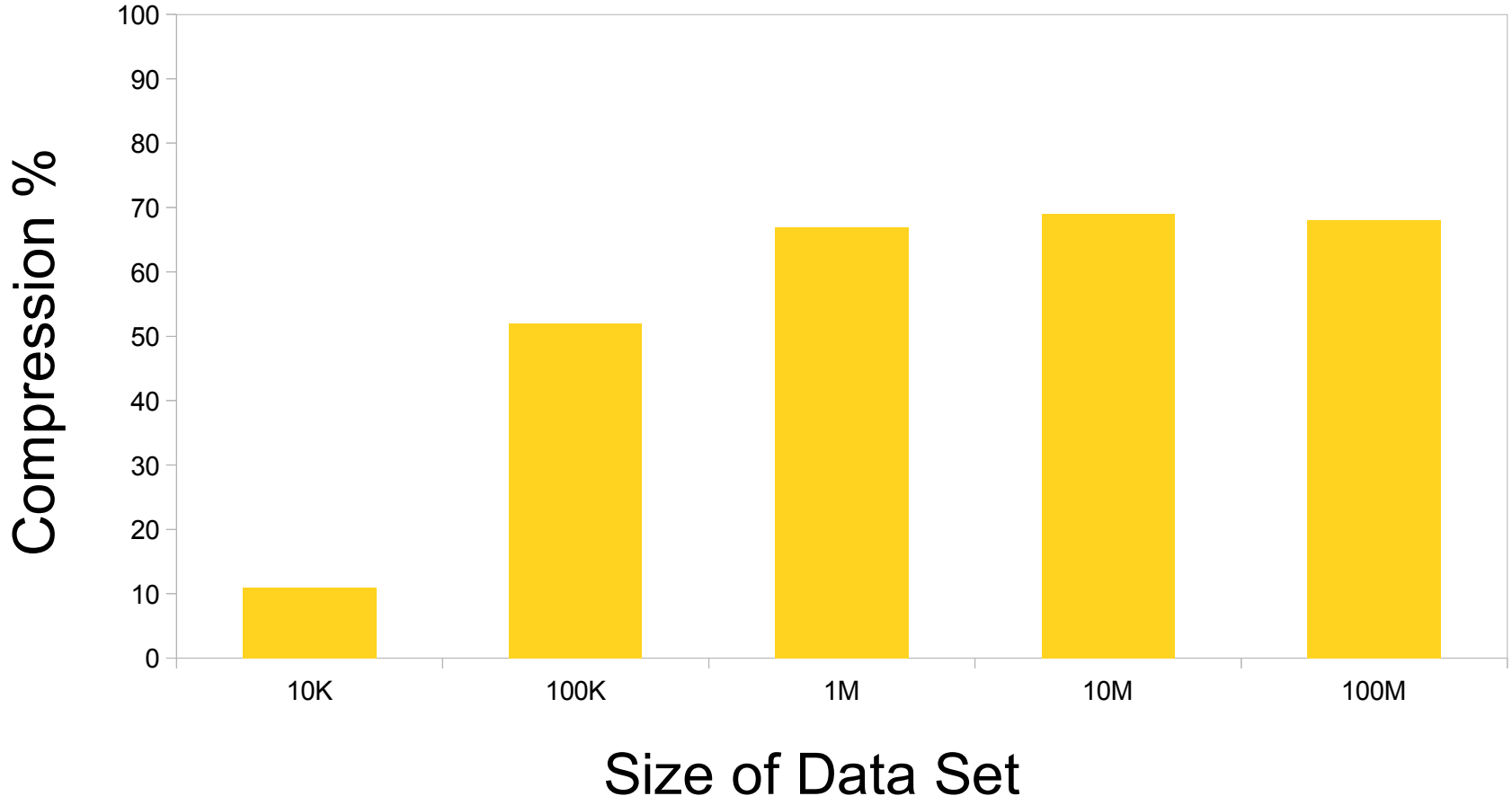
Compression (No auditor)



BROWN

OAK RIDGE National Laboratory

Compression Interpolation (Stride=5)

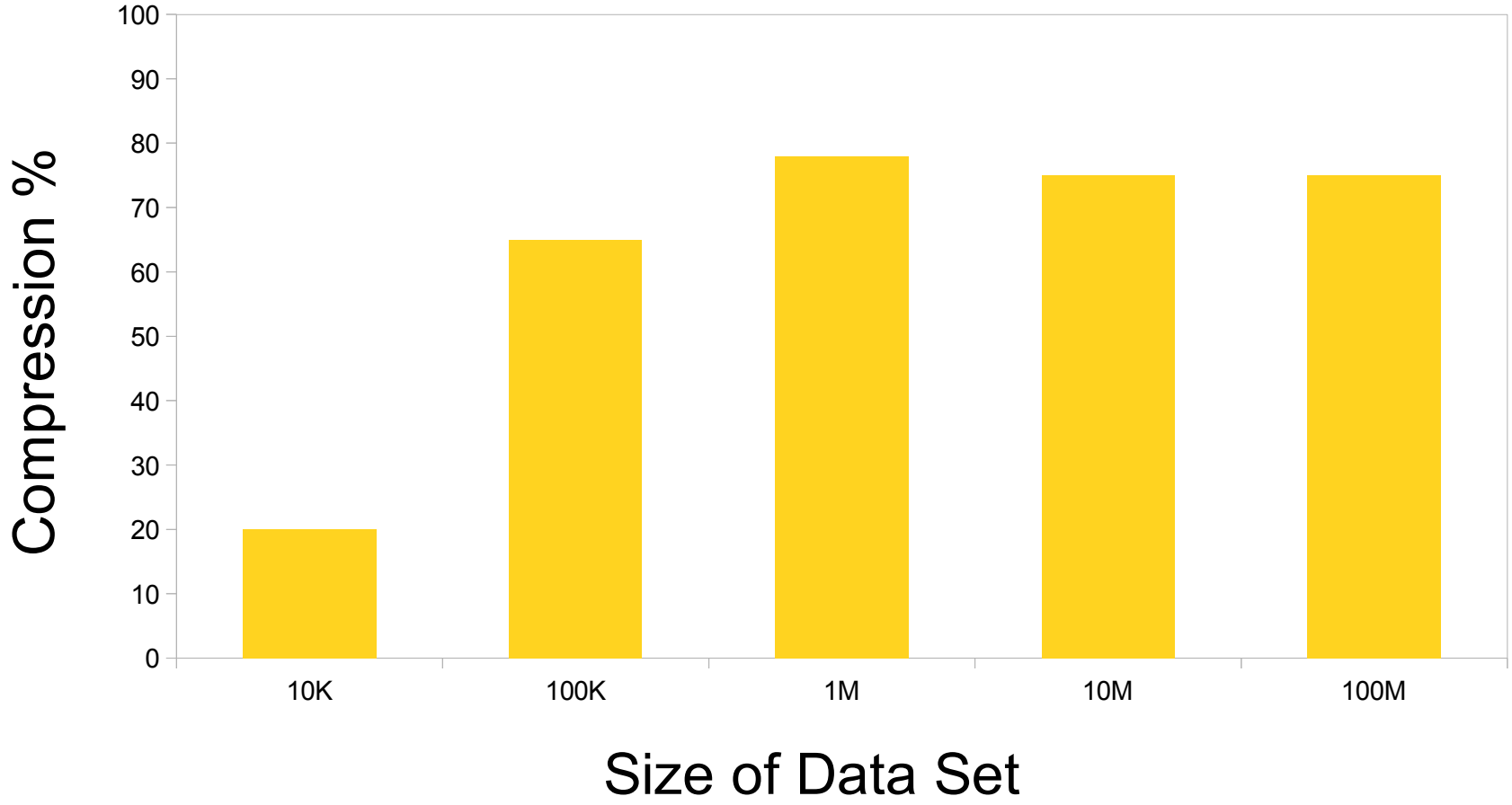


BROWN

OAK RIDGE National Laboratory

Compression

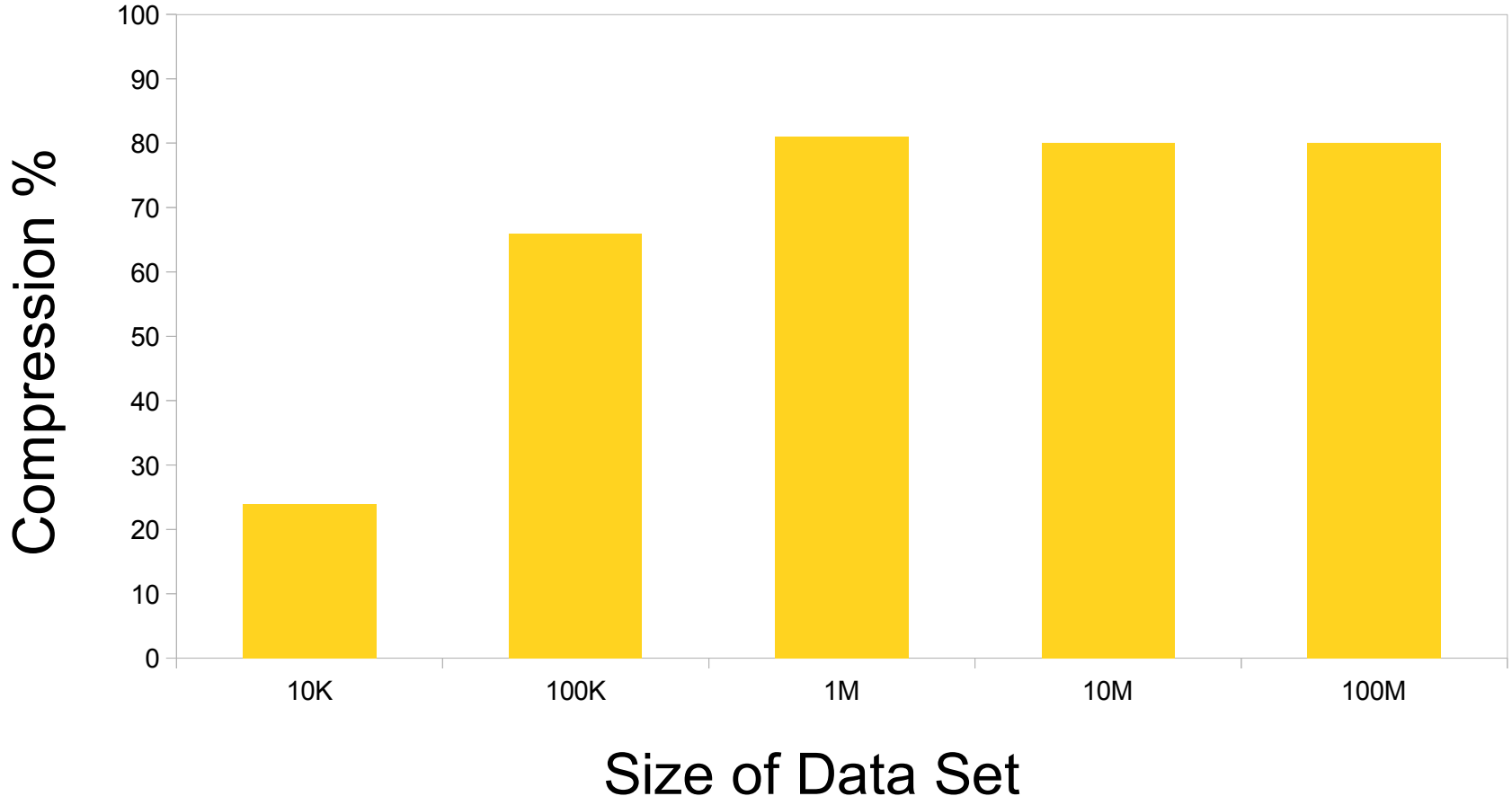
Interpolation (Stride=10)



BROWN

OAK RIDGE National Laboratory

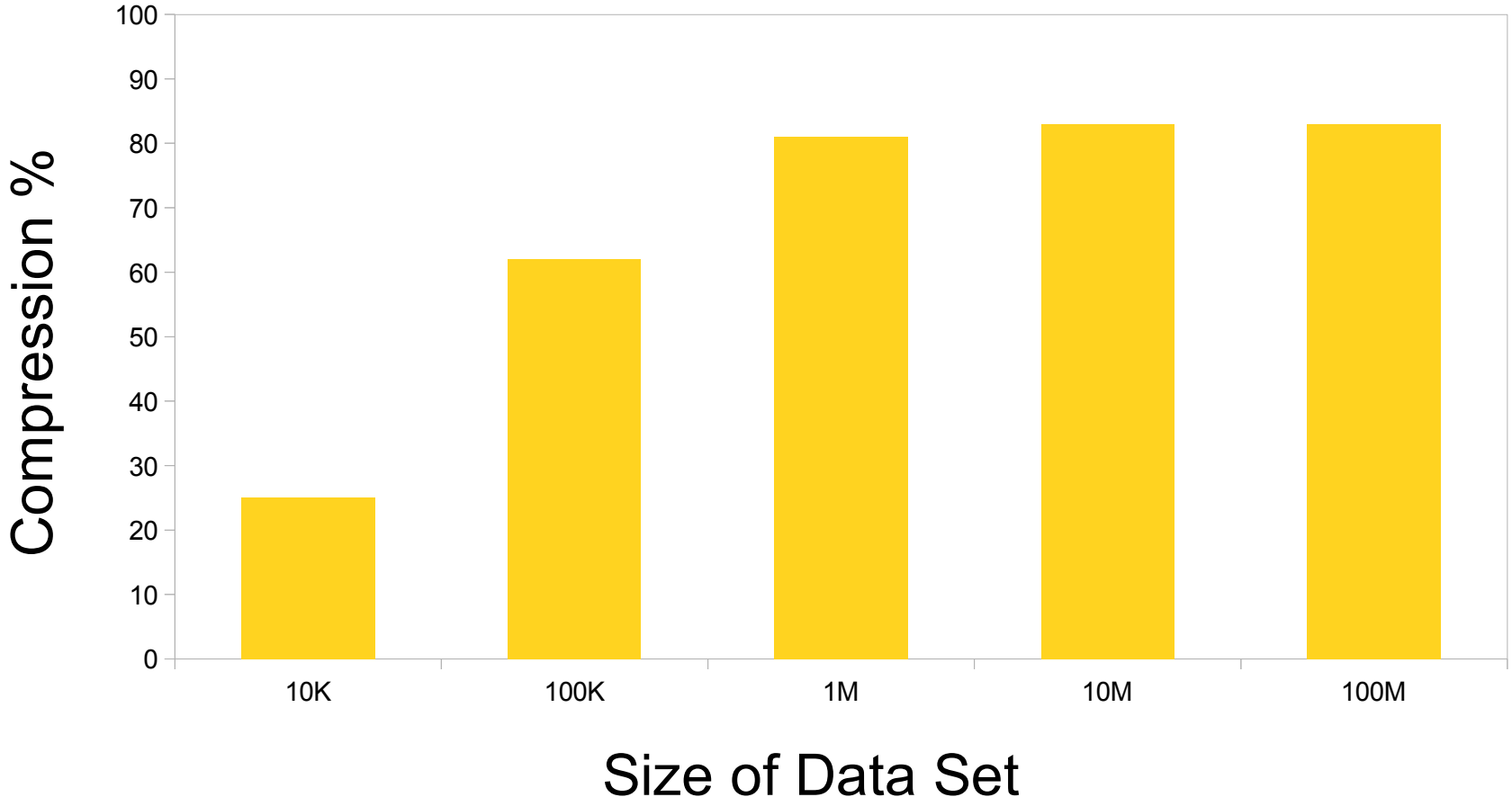
Compression Interpolation (Stride=20)



BROWN

OAK RIDGE National Laboratory

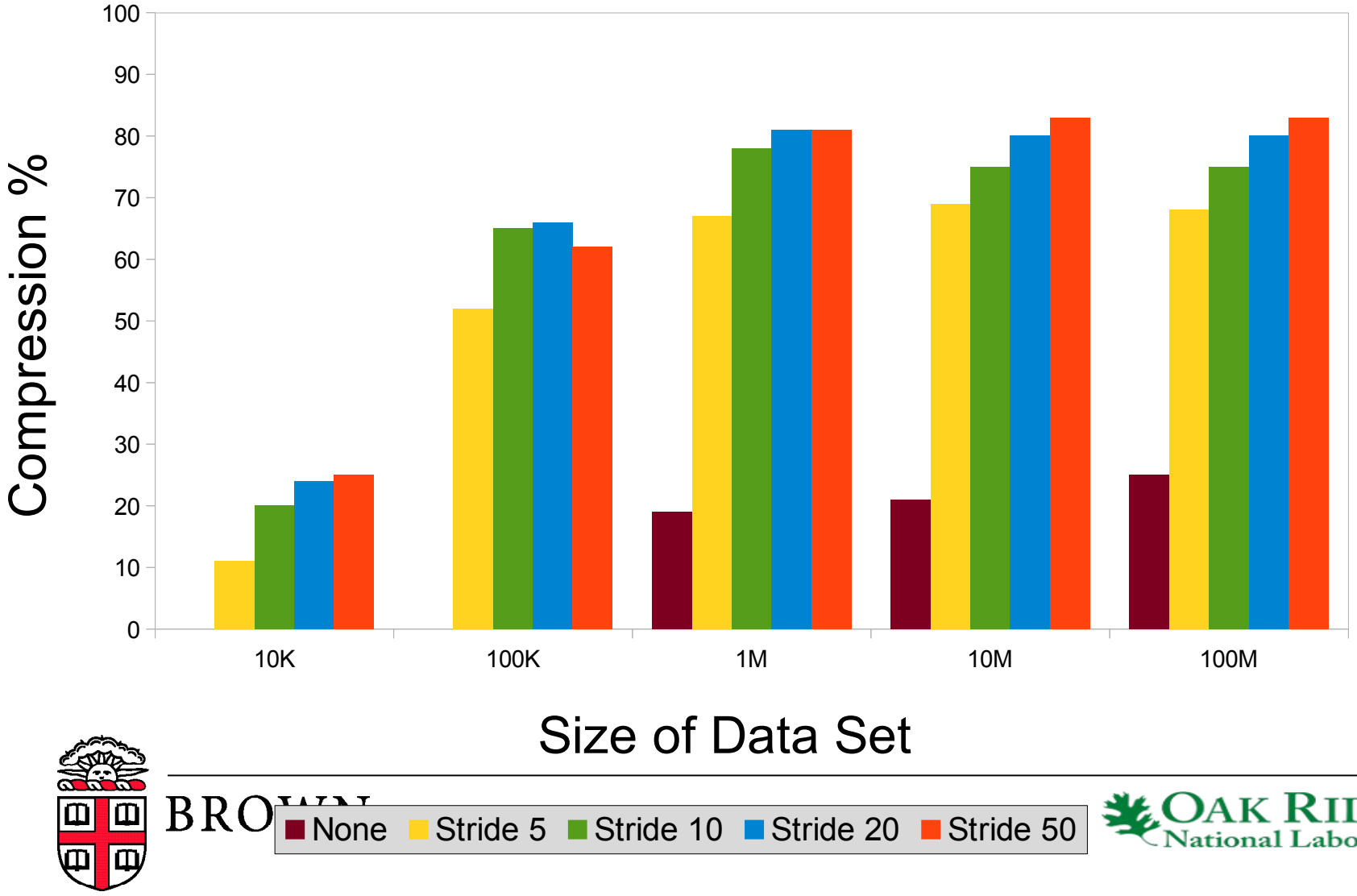
Compression Interpolation (Stride=50)



BROWN

OAK RIDGE National Laboratory

Compression Interpolation Auditor



BROWN

OAK RIDGE National Laboratory