Data

Definition

- (1) Factual information used as a basis for reasoning, discussion, or calculation
- (2) Information output [acquired] by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful
- (3) Information in numerical form that can be digitally transmitted or processed

  – Merriam Webster

Analog Data

- Data represented in continuous form

Analog Data

- Data represented in continuous form

The Emir of Bukhara (1911) and *Supervisor of Chernigov Floodgate (1909).* Prokudin-Gorskii, photographer to the tsar.

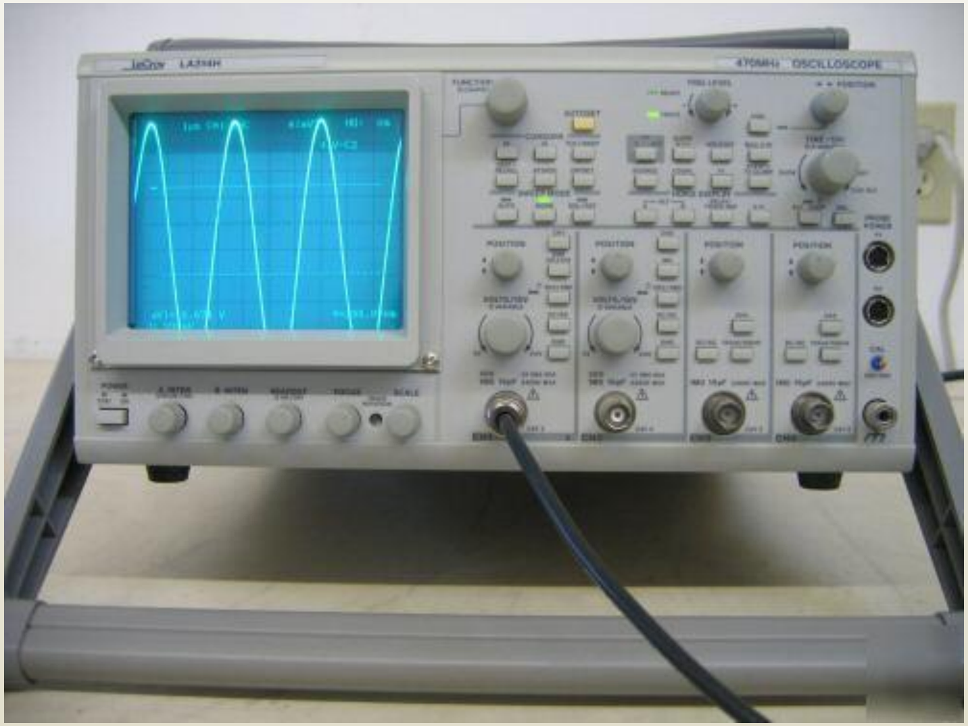
Analog Data

- Data represented in continuous form

Gramophone and records

Analog Data

- Data represented in continuous form



Analog oscilloscope

Analog Data

- Data represented in continuous form
- Challenges: difficult to
  - Store
  - Modify level of detail
  - Transmit
  - Replicate

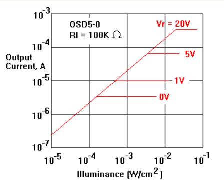
Digital data

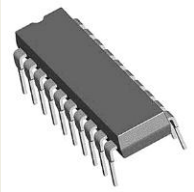
- Data represented in discrete form, using numbers
- World is not discrete
  - digital data is created through analog to digital conversion (i.e. digitization)

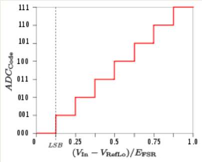
Digitization example

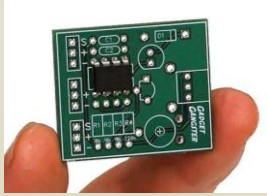
- Goal: acquire digital data to record brightness variation at given outdoor location

Photodiode → Analog to digital converter → Timer / recorder →









000000
134577
7777…

Digitization example

- Goal: acquire digital data to record brightness variation at given outdoor location

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0	0	0	0	1	3	4	5	5	6	7	7	7	7	7	6	6	6	5	4	2	1	0
0	0	0	0	0	0	0	0	2	4	4	5	5	5	5	4	3	1	0	0	0	0	0	0
3	3	3	3	3	4	5	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	6	5
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Digitization example

- Goal: acquire digital data to record brightness variation at given outdoor location

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0	0	0	0	1	3	4	5	5	6	7	7	7	7	7	6	6	6	5	4	2	1	0
0	0	0	0	0	0	0	0	2	4	4	5	5	5	5	4	3	1	0	0	0	0	0	0
3	3	3	3	3	4	5	6	7	7	7	7	7	7	7	7	7	7	7	7	7	7	6	5
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Row 1: summer day in IN; Row 2: winter day in IN;
Row 3: summer day in AK; Row 4: winter day in AK

Digitization examples

- Music encoded digitally
  - Microphone transforms sound into current (signal)
  - Analog to Digital Converter transforms continuous signal into discrete signal
  - Discrete signal is recorded as sequence of numbers
- Digital (video) camera
- Scanner

Advantages of digital data

- Easy to replicate without loss
  - No need for “master copy”
  - Any copy is as good as original
  - (Napster)

Advantages of digital data

- Good control of level of detail
  - If brightness is desired only every 4 hours

<table>
<tr><th>0</th><th>1</th><th>2</th><th>3</th><th>4</th><th>5</th><th>6</th><th>7</th><th>8</th><th>9</th><th>10</th><th>11</th><th>12</th><th>13</th><th>14</th><th>15</th><th>16</th><th>17</th><th>18</th><th>19</th><th>20</th><th>21</th><th>22</th><th>23</th></tr>
<tr><td>0</td><td>0</td><td>0</td><td>0</td><td>0</td><td>1</td><td>3</td><td>4</td><td>5</td><td>5</td><td>6</td><td>7</td><td>7</td><td>7</td><td>7</td><td>7</td><td>6</td><td>6</td><td>6</td><td>5</td><td>4</td><td>2</td><td>1</td><td>0</td></tr>
<tr><td colspan="4">0</td><td colspan="4">2</td><td colspan="4">6</td><td colspan="4">7</td><td colspan="4">6</td><td colspan="4">2</td></tr>
</table>

Advantages of digital data

- Good control of level of detail
  - If only three levels of brightness are needed

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0	0	0	0	1	3	4	5	5	6	7	7	7	7	7	6	6	6	5	4	2	1	0
0	0	0	0	0	0	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1	0	0

Challenges of digital data

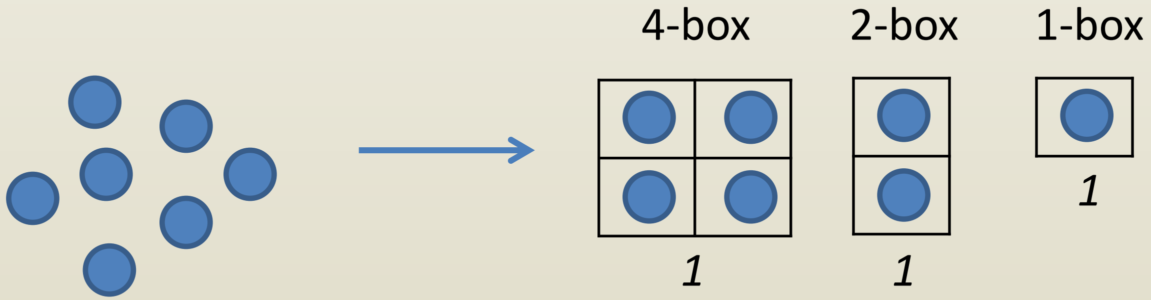
- Limited precision
  – Digital data provides an approximation
- Multiple discrete levels are difficult to implement in computing hardware
  – Base 10 requires implementing 10 digits in hardware: 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9
  – Solution: base 2, “binary”

Base 2—binary

- Only 2 digits: 0 and 1
- Any number can be represented in base 2
  - More binary digits are needed
- Not human friendly
  - We prefer base 10, and higher bases in general
- Hardware friendly
  - It is easier to distinguish quickly and robustly between two digits (e.g. 0 Volts and 5 Volts)
  - One binary digit is stored in one bit of memory
- Advantages overweigh disadvantages
  - All computers use base 2

Base 2

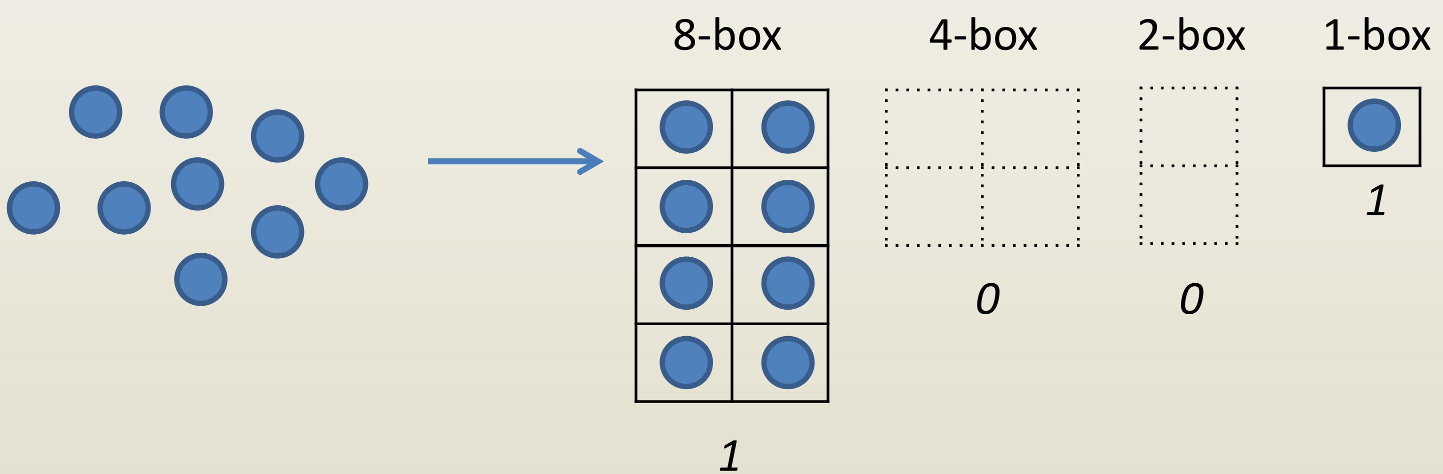
- Boxes of size that are powers of 2
  - 1, 2, 4, 8, 16, 32, etc.
  - In base 10 boxes are of size 1, 10, 100, 1000, etc.
- Always use biggest box to pack the elements you want to count



7_{10} → 111_2

Base 2

- Boxes of size that are powers of 2
  - 1, 2, 4, 8, 16, 32, etc.



9_{10} → 1001_2


THERE ARE 10 TYPES
OF PEOPLE IN THIS WORLD:
THOSE WHO UNDERSTAND BINARY
AND THOSE WHO DON'T

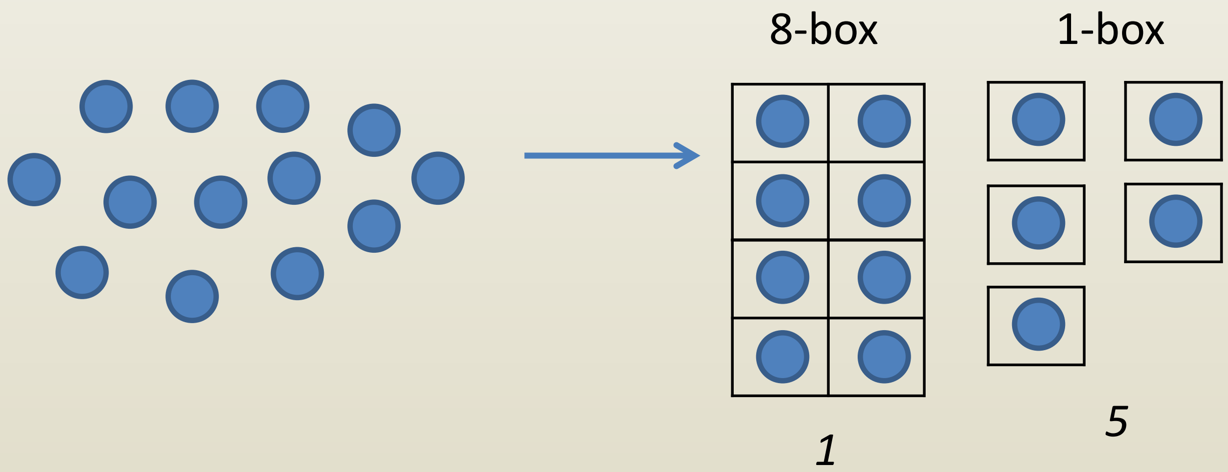
iClicker question

- Convert 1010_2 from binary to base 10

A. 6_{10}
B. 12_{10}
C. 101_{10}
D. 10_{10}
E. 1010_{10}

Base 8

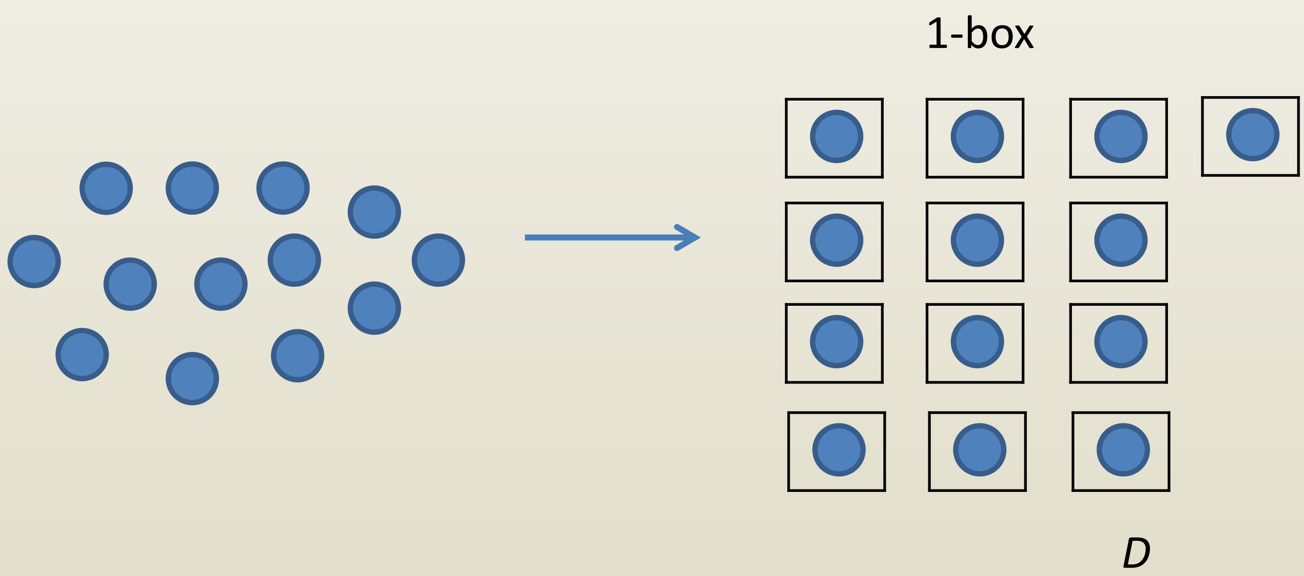
- Boxes of size that are powers of 8
  - 1, 8, 64, 512, etc.
  - 8 digits: 0, 1, 2, 3, 4, 5, 6, 7



$13_{10} \rightarrow 15_{8}$

Base 16

- Boxes of size that are powers of 16
  - 1, 16, 256, 4096, etc.
  - 16 digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F



13_{10} → D_{16}

Base 2 to base 16 conversions

- Base 16 is used to make base 2 manageable by humans
- 1 base 16 (i.e. hexadecimal) digit corresponds to 4 base 2 digits

Base 16	0	1	2	3	4	5	6	7
Base 2	0000	0001	0010	0011	0100	0101	0110	0111

Base 16	8	9	A	B	C	D	E	F
Base 2	1000	1001	1010	1011	1100	1101	1110	1111

Base 16	14	1D	AA	FF	AB89
Base 2	0001 0100	0001 1101	1010 1010	1111 1111	1010 1011 1000 1001

iClicker question

- Convert $DEED_{16}$ from base 16 to base 2

A. $1010\ 1011\ 1011\ 1010_2$

B. $1110\ 1101\ 1101\ 1110_2$

C. $1101\ 1110\ 1110\ 1101_2$

D. $1110\ 1111\ 1111\ 1110_2$

E. $1101\ 1111\ 1111\ 1101_2$

Data types

- Characters, to encode textual data
  - Lower case: a, b, c, …
  - Upper case: A, B, C, …
  - Digits: 0, 1, 2, …
  - Special characters: *space ( ), column (:), question mark (?), …*
  - There are fewer than 256 characters, so 8 bits are enough to encode a character
    - 8 bits are called a byte

Bits and bytes

- 1 kilobit (1kb) is 1,024 bits
  - And not 1,000 bits
- 1 megabit (1Mb) is 1,024 kilobits
- 1 kilobyte (1kB) is 1,024 bytes
  - or 8 kilobits
  - or 8x1,024 bits
- b stands for bit, B stands for byte
  - bits are typically used for networking bandwidths or memory address sizes
    - 100kbps (kilobits per second), 32 bit addresses
  - Bytes are typically used for memory capacity
    - 1GB (1,024 MB; 1,024x1,024KB; 1,024x1,024x1,024B)

iClicker question

- A 3-minute song is stored in a 1MB file. Can the song be streamed over a 256kbps network?

A yes

B no

C wrong answer

D wrong answer

E wrong answer

Memory addresses

- Smallest addressable memory location 1B
  - You cannot read or write less than 1 byte
- Sufficient binary digits needed to uniquely name all bytes
  - 1KB total memory size requires 10 bit memory addresses (2^{10} = 1,024)
- For a long time, computers used 32bit (4byte) addresses
  - Maximum memory size that can be addressed: 2^{32} = 4GB
- Switch to 64bit to allow for larger memories
  - Memories larger than 2^{64}—*never*
  - Number of particles in the universe: 10^{87}

Data types

- Characters, to encode textual data
- Integer numbers
  - Minimum and maximum representable number depends on number of bits used and on whether you allow for negative numbers or not
  - Unsigned byte: from 0 to 255
  - Signed byte: from -127 to 127
  - Unsigned 4 bytes: from 0 to over 4 billion

Data types

- Characters, to encode textual data
- Integer numbers
- Real numbers
  - Fixed point
    - Example: 8 bits for the integer part, 8 bits for the fractional part
    - Cannot encode very small or very large numbers

Data types

- Characters, to encode textual data
- Integer numbers
- Real numbers
  - Fixed point
  - Floating point
    - Example: 1 bit for the sign, 8 bits for the exponent, 23 bits for the mantissa
    - The decimal point is “floating”

Data types

- Characters, to encode textual data
- Integer numbers
- Real numbers
  - Fixed point
  - Floating point
  - Precision is limited
    - Numbers are approximate to begin with
    - After arithmetic operations, approximation error increases
    - Understanding and controlling numerical error is a fundamental problem in computer science

Data types

- Characters, to encode textual data
- Integer numbers
- Real numbers
- Compound data types
  - Strings: an array of characters
  - Vectors: an array of floating point numbers
  - Medical records: a combination of strings, vectors, etc.

CAD data of a car

- Car

CAD data of a car

- Car
  - Chassis
  - Power train
  - Body

CAD data of a car

- Car
  - Chassis
    - Wheels
    - Undercarriage
  - Power train
    - Engine
    - Gear box
    - Exhaust
    - Breaks
  - Body
    - Doors
    - Windows
    - Hood
    - Trunk lid

CAD data of a car

- Car
  - Chassis
    - Wheels
    - Undercarriage
  - Power train
    - Engine
      - Cylinders
      - Pistons
      - Spark plugs
        - » Body
        - » Ceramic insulator
        - » Electrodes
      - Valves
    - Gear box
    - Exhaust
    - Breaks
  - Body
    - Doors
    - Windows
    - Hood
    - Trunk lid

Modeling and abstraction

- Compound data types allow modeling complex entities hierarchically, through abstraction
  - Hide details irrelevant in given context
- Hierarchical modeling and abstraction supports
  - Creativity: avoids unnecessary cognitive burden, improves focus
  - Repair: enables systematic approach to tracking down problem
  - Interoperability: enables developing part that works with system without knowledge of system details

Examples of data processing

- Blurring
- Sorting
- Down-sampling
- Feature extraction
- Encryption/decryption
- Compression/decompression
- Statistical analysis

Blurring

- Filtering out high frequencies or abrupt changes
- Data sample replaced with average of neighboring samples

Original image

Blurred image

Sorting

- Permute data according to a total order relation
  - Example: sorting credit card transactions based on amount (decreasing) and then on transaction date (from recent to old)

Date	Amount
02.07.11	$4.60
01.12.11	$100.00
02.05.11	$34.35
02.02.11	$100.00

Original data

Date	Amount
02.02.11	$100.00
01.12.11	$100.00
02.05.11	$34.35
02.07.11	$4.60

Sorted data

Down sampling

- Reducing data
  - Fewer measurements in unit of time (i.e. reducing temporal resolution)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0	0	0	0	1	3	4	5	5	6	7	7	7	7	7	6	6	6	5	4	2	1	0
0				2				6				7				6				2			

Middle row: original data.
Bottom row: data down sampled in time

Down sampling

- Reducing data
  - Fewer measured levels

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	0	0	0	0	1	3	4	5	5	6	7	7	7	7	7	6	6	6	5	4	2	1	0
0	0	0	0	0	0	1	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1	0	0

Middle row: original data.
Bottom row: data down sampled by reducing number of levels

Down sampling

- Reducing data
  - Fewer measurements in unit of length, area, or volume (i.e. reducing spatial resolution)

Original image

Image down sampled 4x4

Feature extraction

- Edge extraction

Original image

Edge image

Encryption/decryption

- Encryption
  - Transform original data to hide its content
- Decryption
  - Revert data to original form
- Example
  - Original data: CS17700
  - Encryption scheme: replace letter with following letter in alphabet and digit with following digit
    - Encrypted data: DT28811
  - Decryption scheme: replace letter with preceding letter in alphabet and digit with preceding digit
    - Decrypted data: CS17700

Encryption/decryption

- Encryption
  - Transform original data to hide its content
- Decryption
  - Revert data to original form
- Example CS17700 -> DT28811
- A good encryption scheme
  - Cannot be decrypted by anyone other than intended recipient
  - Does not increase data size
  - Is fast

Enigma

"The enigma is a machine that is used to cipher and decipher messages. The result was a polyalphabetic substitution cipher that is nearly impossible to break"

"However, the machine did have some weaknesses which were found through the efforts at Bletchley Park. The use and breaking of the enigma machine had great impacts on WWII."

Compression/decompression

- Data compression
  - Exploiting data redundancy to derive a more compact data representation
- Data decompression
  - Reverting compressed data to a form similar to the original data
- Non-lossy compression
  - Decompressed data identical to original data
- Lossy compression
  - Decompressed data similar to original data

Compression / decompress. example

- Original data
  - 0000 0000 0011 1100 1111 1111 0000 0000 0000
- Data compressed by run length encoding
  - 1010 0 0100 1 0010 0 1000 1 1100 0
  - 10 0's 4 1's 2 0's 8 1's 12 0's
  - Non-lossy

Compression / decompress. example

- Original data
  - 0000 0000 0011 1100 1111 1111 0000 0000 0000
- Lossy compression: ignore sequences shorter than 3
  - 1010 0 1110 1 1100 0
  - 10 0's 14 1's 12 0's
- Decompressed data, not identical to original
  - 0000 0000 0011 1111 1111 1111 0000 0000 0000

iClicker question

- A book has 2^{20} words out of which only 2^8 are unique.
- The average length of a unique word is 4 characters. A character is stored in one byte.
- You compress the book by storing the unique words once and then storing indices of the unique words as they appear in the text.
- What is the size in bytes of the compressed book?

A. $2^8*4 + 2^{20}*1$

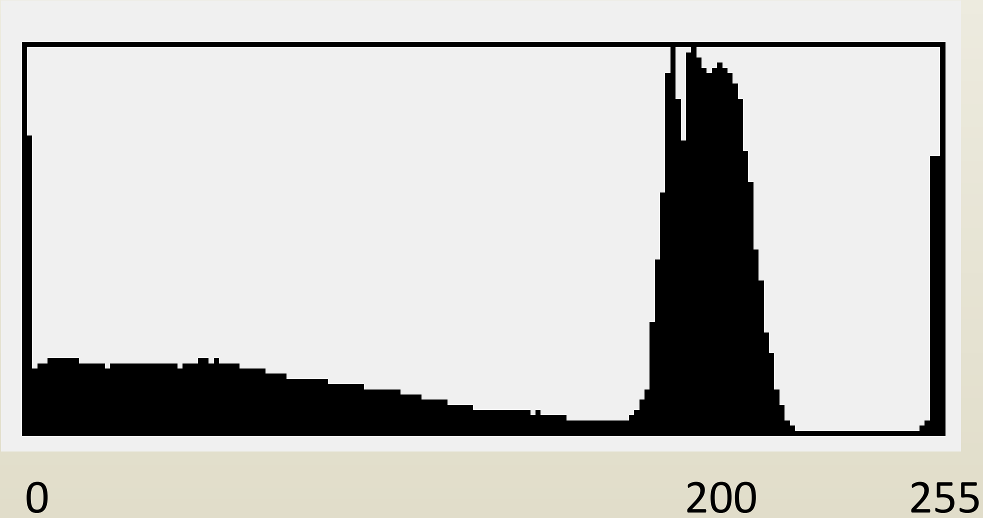
B. $2^8*4*8 + 2^{20}*8$

C. $2^8*4 + 2^{20}*8$

D. $2^{20}*4$

Statistical analysis

- Examples
  - Min, max, average, standard deviation, regression, ANOVA, ANCOVA etc.
  - Histogram



Blue channel histogram