Big Data Management and Analytics
Assignment 5

Assignment 5-1

(a) Explain each of the terms by providing a short definition

Aggregation: Matching of similar objects to groups and aggregation of the entire group

Compression: Compress received data

Data Reduction: reduce the size of received data

Histograms: Describes a method for approximating frequency distributions of elements in streams

Assignment 5-1

(a) Explain each of the terms by providing a short definition

Load Shedding: Given the case where the data in the stream arrives with such a high velocity that it could overburden the system, some part of the data will be discarded.

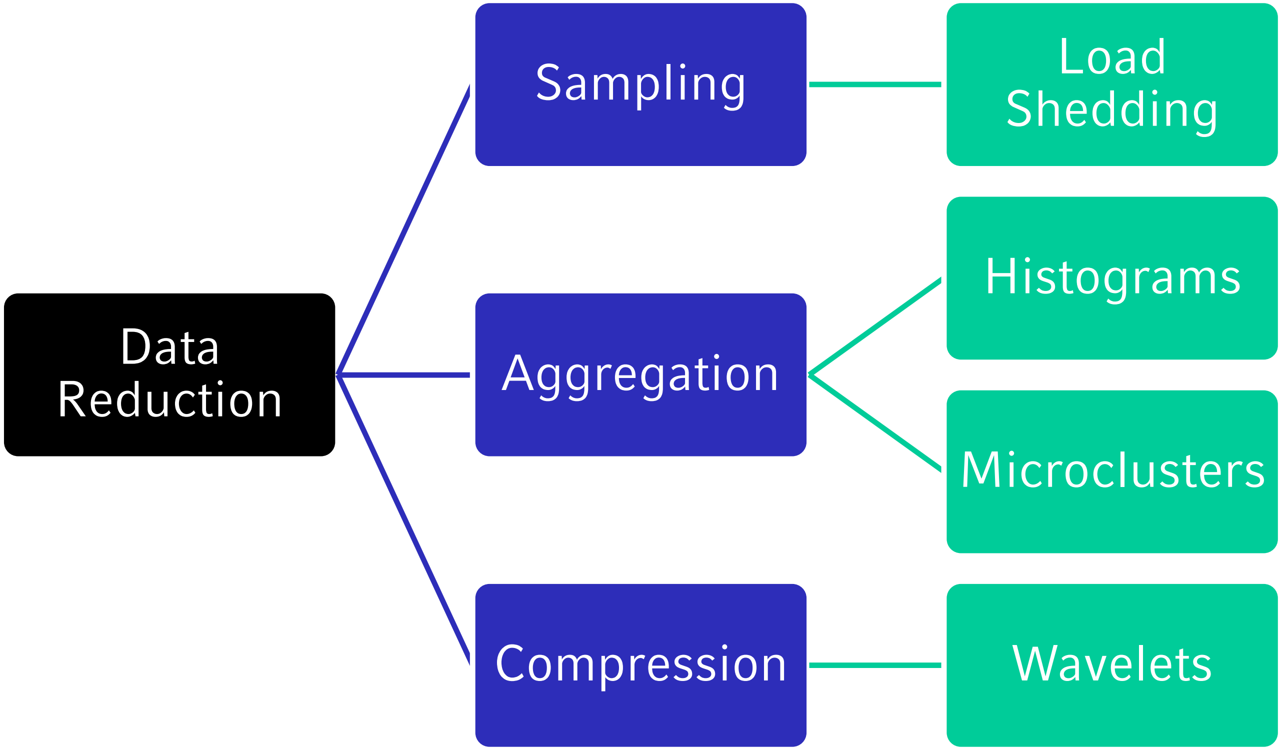
Microclusters: Describes a group of similar objects

Sampling: Selecting a subset of the data

Wavelets: Deconstruct a signal in several coefficients

Assignment 5-1

(b) Illustrate how the terms are related to each other.



Assignment 5-2

Given the following input sequence S = (4,1,2,3,6,1,7,6)

(a) Perform a Haar Wavelet Transformation on S and determine the Wavelet coefficients

Transform S into a sequence of two-component vectors $\left((s_1, d_1) \ldots (s_n, d_n)\right)$

where $\boxed{\forall i \; mod \; 2 = 0: \begin{pmatrix} s_i \\ d_i \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_i \\ x_{i+1} \end{pmatrix}}$

Haar matrix

Vector with current element of the sequence and its sucessor element

Assignment 5-2

Given the following input sequence S = (4,1,2,3,6,1,7,6)

(a) Perform a Haar Wavelet Transformation on S and determine the Wavelet coefficients

Step1:

$$s_1 = \left(\frac{4+1}{2}, \frac{2+3}{2}, \frac{6+1}{2}, \frac{7+6}{2}\right) = (2.5, 2.5, 3.5, 6.5)$$

$$d_1 = \left(\frac{4-1}{2}, \frac{2-3}{2}, \frac{6-1}{2}, \frac{7-6}{2}\right) = (1.5, -0.5, 2.5, 0.5)$$

Assignment 5-2

Given the following input sequence S = (4,1,2,3,6,1,7,6)

Step1:

$$s_1 = \left(\frac{4+1}{2}, \frac{2+3}{2}, \frac{6+1}{2}, \frac{7+6}{2}\right) = (2.5, 2.5, 3.5, 6.5)$$

$$d_1 = \left(\frac{4-1}{2}, \frac{2-3}{2}, \frac{6-1}{2}, \frac{7-6}{2}\right) = (1.5, -0.5, 2.5, 0.5)$$

Step2:

$$s_2 = \left(\frac{2.5+2.5}{2}, \frac{3.5+6.5}{2}\right) = (2.5, 5)$$

$$d_2 = \left(\frac{2.5-2.5}{2}, \frac{3.5-6.5}{2}\right) = (0, -1.5)$$

DATABASE SYSTEMS GROUP

Assignment 5-2

Given the following input sequence S = (4,1,2,3,6,1,7,6)

Step2:

$$s_2 = \left(\frac{2.5+2.5}{2}, \frac{3.5+6.5}{2}\right) = (2.5, 5)$$

$$d_2 = \left(\frac{2.5-2.5}{2}, \frac{3.5-6.5}{2}\right) = (0, -1.5)$$

Step3:

$$s_3 = \left(\frac{2.5+5}{2}\right) = (3.75)$$

$$d_3 = \left(\frac{2.5-5}{2}\right) = (-1.25)$$

Assignment 5-2

Given the following input sequence S = (4,1,2,3,6,1,7,6)

Mean	Coefficients
(4,1,2,3,6,1,7,6)	(-)
(2.5, 2.5, 3.5, 6.5)	(1.5, -0.5, 2.5, 0.5)
(2.5, 5)	(0, -1.5)
(3.75)	(-1.25)

Assignment 5-2

Given the following input sequence S = (4,1,2,3,6,1,7,6)

(b) Reconstruct the original sequence S using the Wavelet coefficients

For the reconstruction of a sequence S we use:

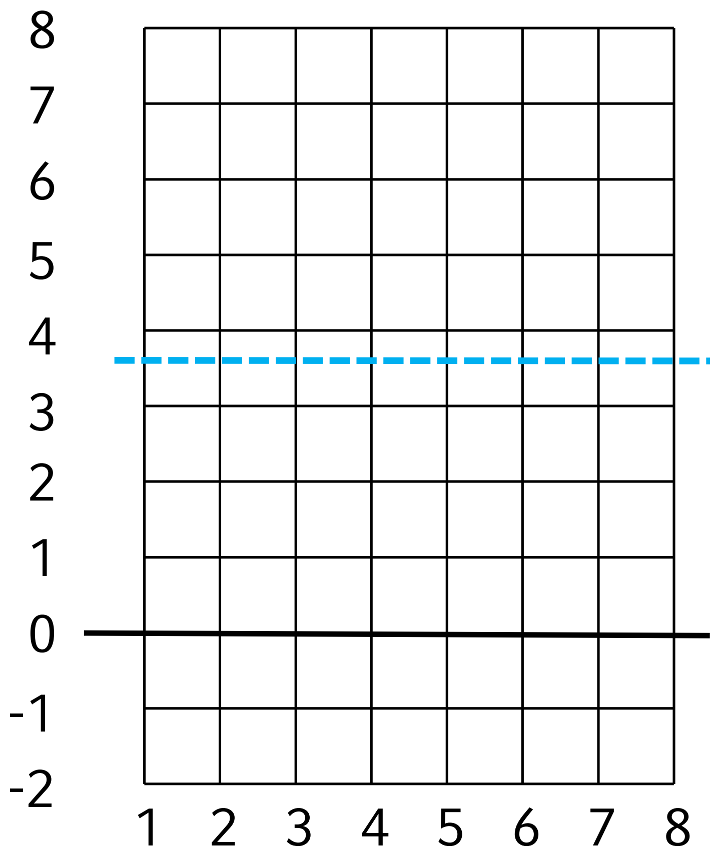
$$\boxed{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} s_i \\ d_i \end{pmatrix} = \begin{pmatrix} x'_i \\ x'_{i+1} \end{pmatrix}}$$

Assignment 5-2

The wavelet coefficients obtained from (a):

DWT(S) = (3.75, -1.25, 0, -1.5, 1.5, -0.5, 2.5, 0.5)

Loss-free reconstruction:

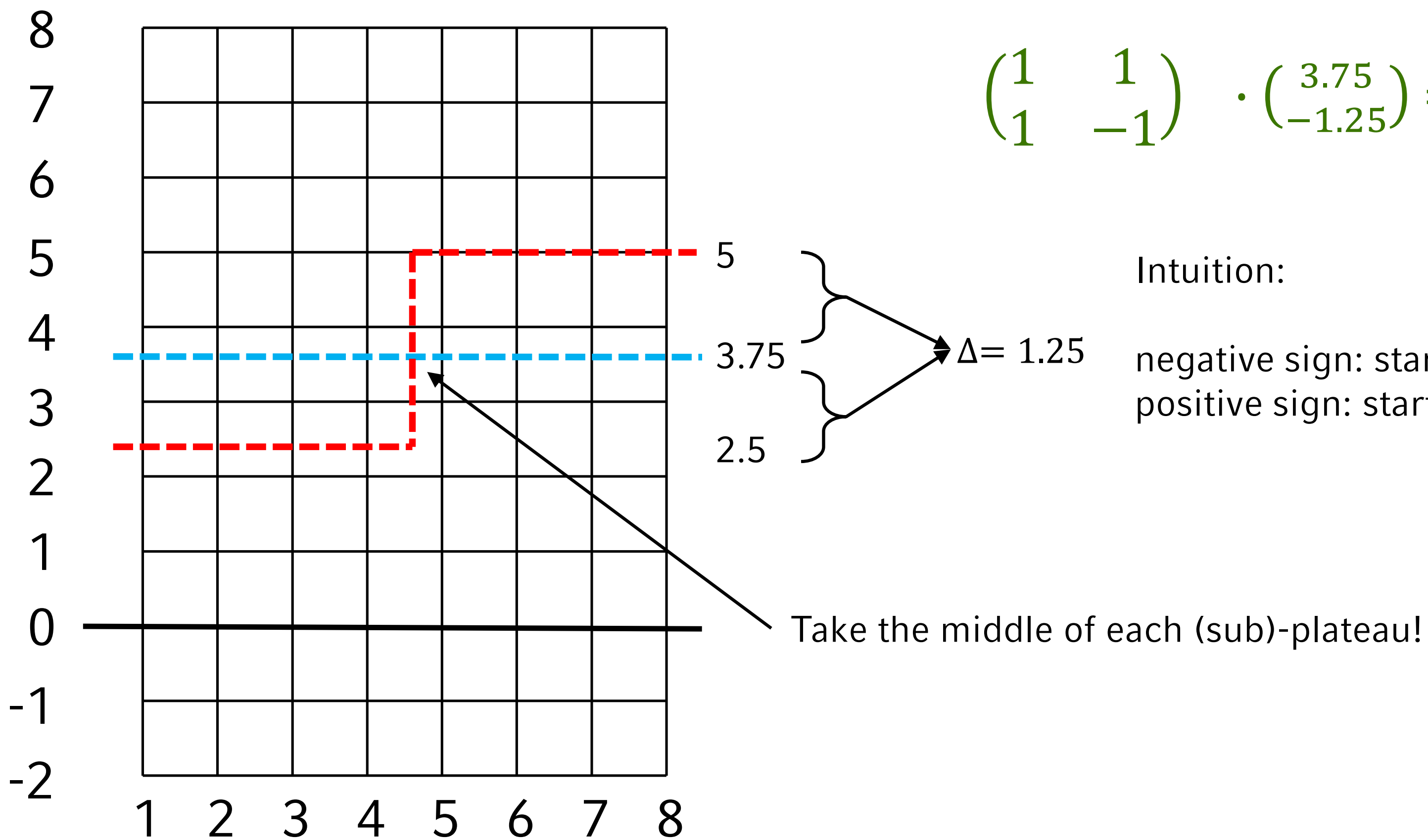


Assignment 5-2

The wavelet coefficients obtained from (a):

DWT(S) = (3.75, -1.25, 0, -1.5, 1.5, -0.5, 2.5, 0.5)

Loss-free reconstruction:



$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3.75 \\ -1.25 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 5 \end{pmatrix}$$

Intuition:

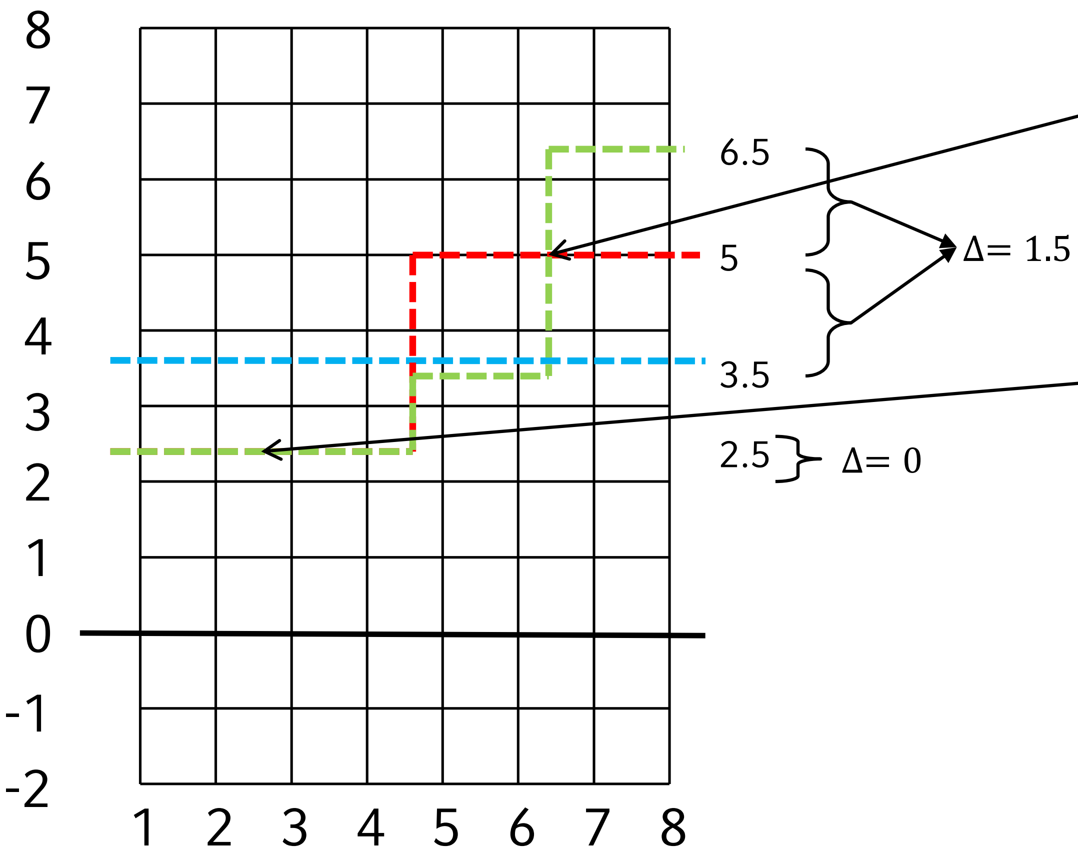
negative sign: start below the base line
positive sign: start above the base line

Assignment 5-2

The wavelet coefficients obtained from (a):

DWT(S) = (3.75, -1.25, 0, -1.5, 1.5, -0.5, 2.5, 0.5)

Loss-free reconstruction:



$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ -1.5 \end{pmatrix} = \begin{pmatrix} 3.5 \\ 6.5 \end{pmatrix}$$

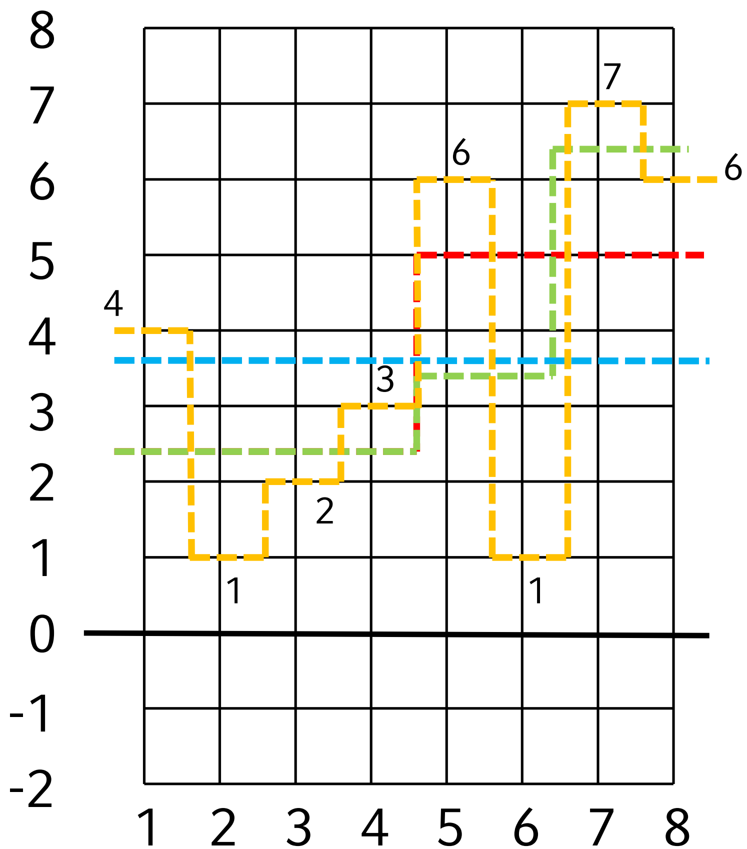
$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2.5 \\ 0 \end{pmatrix} = \begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}$$

Assignment 5-2

The wavelet coefficients obtained from (a):

DWT(S) = (3.75, -1.25, 0, -1.5, 1.5, -0.5, 2.5, 0.5)

Loss-free reconstruction:



$$\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix} \cdot \begin{pmatrix}2.5\\1.5\end{pmatrix} = \begin{pmatrix}4\\1\end{pmatrix}$$

$$\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix} \cdot \begin{pmatrix}2.5\\-0.5\end{pmatrix} = \begin{pmatrix}2\\3\end{pmatrix}$$

$$\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix} \cdot \begin{pmatrix}3.5\\2.5\end{pmatrix} = \begin{pmatrix}6\\1\end{pmatrix}$$

$$\begin{pmatrix}1 & 1\\1 & -1\end{pmatrix} \cdot \begin{pmatrix}6.5\\0.5\end{pmatrix} = \begin{pmatrix}7\\6\end{pmatrix}$$

The original sequence is

S = (4,1,2,3,6,1,7,6)

Assignment 5-2

(c) We assume that all coefficients of value [-0.5,0.5] are close to zero

DWT(S) = (3.75, -1.25, 0, -1.5, 1.5, -0.5, 2.5, 0.5)

Changes to:

DWT'(S) = (3.75, -1.25, 0, -1.5, 1.5, 0, 2.5, 0)

Reconstructing S with DWT'(S):

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} \boxed{3.75} \\ \boxed{-1.25} \end{pmatrix} = \begin{pmatrix} 2.5 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2.5 \\ \boxed{0} \end{pmatrix} = \begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 5 \\ \boxed{-1.5} \end{pmatrix} = \begin{pmatrix} 3.5 \\ 6.5 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2.5 \\ \boxed{1.5} \end{pmatrix} = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

$$\boxed{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 2.5 \\ \boxed{0} \end{pmatrix} = \begin{pmatrix} 2.5 \\ 2.5 \end{pmatrix}}$$

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 3.5 \\ \boxed{2.5} \end{pmatrix} = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$$

$$\boxed{\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \cdot \begin{pmatrix} 6.5 \\ \boxed{0} \end{pmatrix} = \begin{pmatrix} 6.5 \\ 6.5 \end{pmatrix}}$$

Assignment 5-2

(c) We assume that all coefficients of value [-0.5,0.5] are close to zero

DWT(S) = (3.75, -1.25, 0, -1.5, 1.5, -0.5, 2.5, 0.5)

Changes to:

DWT'(S) = (3.75, -1.25, 0, -1.5, 1.5, 0, 2.5, 0)

Using DWT'(S) leads to a loss afflicted reconstruction with:

S = (4,1,2,3,6,1,7,6)

S' = (4,1,2.5,2.5,6,1,6.5,6.5)

Now take each residue from S,S' and compute their difference and sum up the differences to the total approximation error:

$$\varepsilon_{err}(S, S') = \sum_{i=0}^{|S|-1} |S(i) - S'(i)|$$

Assignment 5-2

(c) We assume that all coefficients of value [-0.5,0.5] are close to zero

S = (4,1,2,3,6,1,7,6)

S‘ = (4,1,2.5,2.5,6,1,6.5,6.5)

Now take each residue from S,S‘ and compute their difference and sum up the differences to the total approximation error:

$$\varepsilon_{err}(S,S') = \sum_{i=0}^{|S|-1} |S(i) - S'(i)|$$

$\varepsilon_{err}(S,S') = |4-4| + |1-1| + |2-2.5| + |3-2.5| + |6-6| + |1-1| + |7-6.5| + |6-6.5| = 2$

Assignment 5-3

(a) Compute the reduced representation of S using PAA (box size M=4)

Hint: A PAA approximates a time series X of length N with a vector $\bar{X} = (\bar{x}_1, \ldots, \bar{x}_M)$ of arbitrary length $M \leq N$, where for each $\bar{x}_i$ holds:

$$\boxed{\bar{x}_i = \frac{M}{N} \sum_{j=\frac{N}{M}(i-1)+1}^{\frac{N}{M}i} x_j}$$

Assignment 5-3

(a) Compute the reduced representation of S using PAA (box size M=4)

position p = (1 ,2 ,3 ,4 ,5 ,6 ,7 ,8)

Initial sequence S = (4 ,1 ,2 ,3 ,6 ,1 ,7 ,6)

$$\bar{x}_1 = \frac{4}{8} \sum_{j=\frac{8}{4}(1-1)+1=1}^{\frac{8}{4}1=2} x_j = \frac{(4+1)}{2} = 2.5$$

$$\bar{x}_2 = \frac{4}{8} \sum_{j=\frac{8}{4}(2-1)+1=3}^{\frac{8}{4}2=4} x_j = \frac{(2+3)}{2} = 2.5$$

Assignment 5-3

(a) Compute the reduced representation of S using PAA (box size M=4)

position p = (1 ,2 ,3 ,4 ,5 ,6 ,7 ,8)

Initial sequence S = (4 ,1 ,2 ,3 ,6 ,1 ,7 ,6)

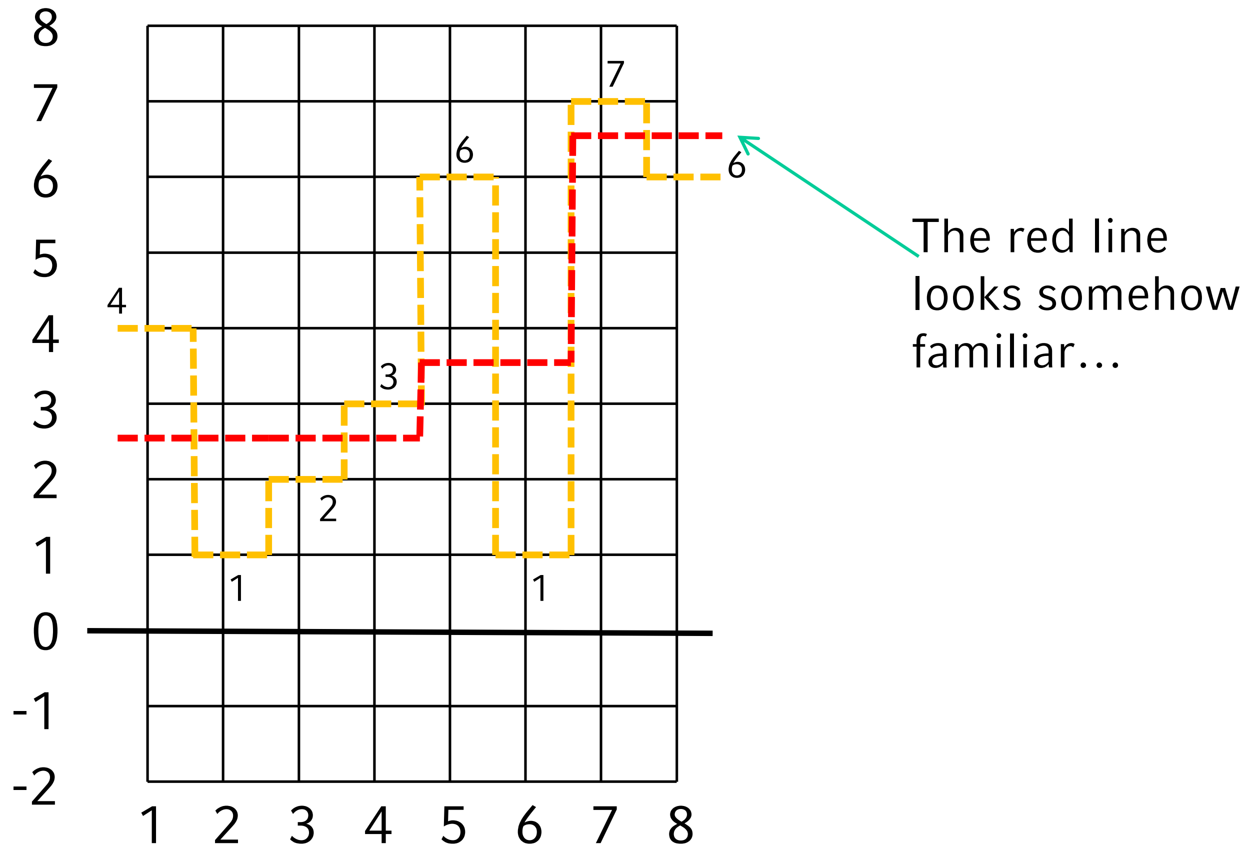
$$\bar{x}_3 = \frac{4}{8} \sum_{j=\frac{8}{4}(3-1)+1=5}^{\frac{8}{4}3=6} x_j = \frac{(6+1)}{2} = 3.5$$

$$\bar{x}_4 = \frac{4}{8} \sum_{j=\frac{8}{4}(4-1)+1=7}^{\frac{8}{4}2=8} x_j = \frac{(7+6)}{2} = 6.5$$

Assignment 5-3

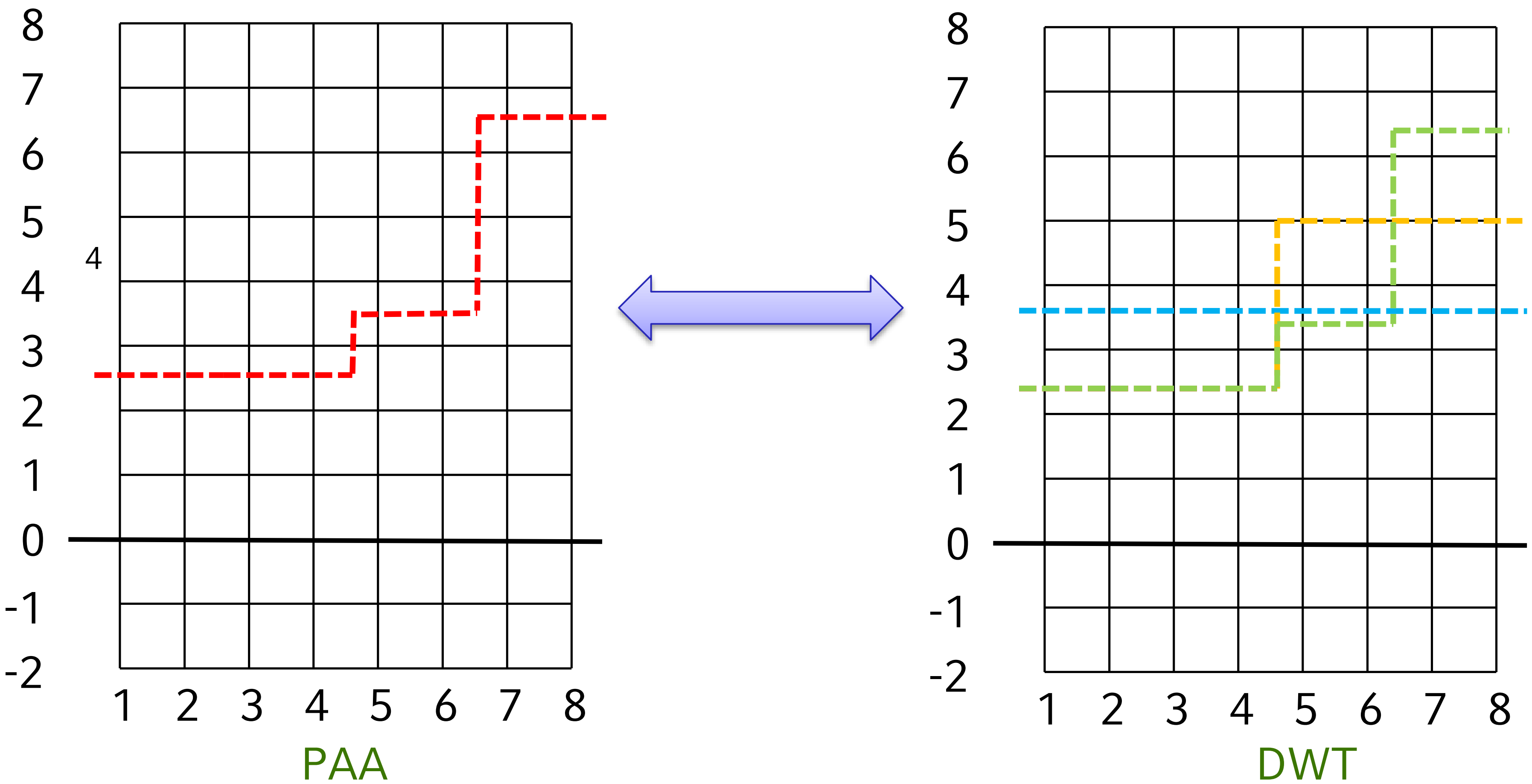
(a) Compute the reduced representation of S using PAA (box size M=4)

PAA(S) = (2.5, 2.5, 3.5, 6.5)



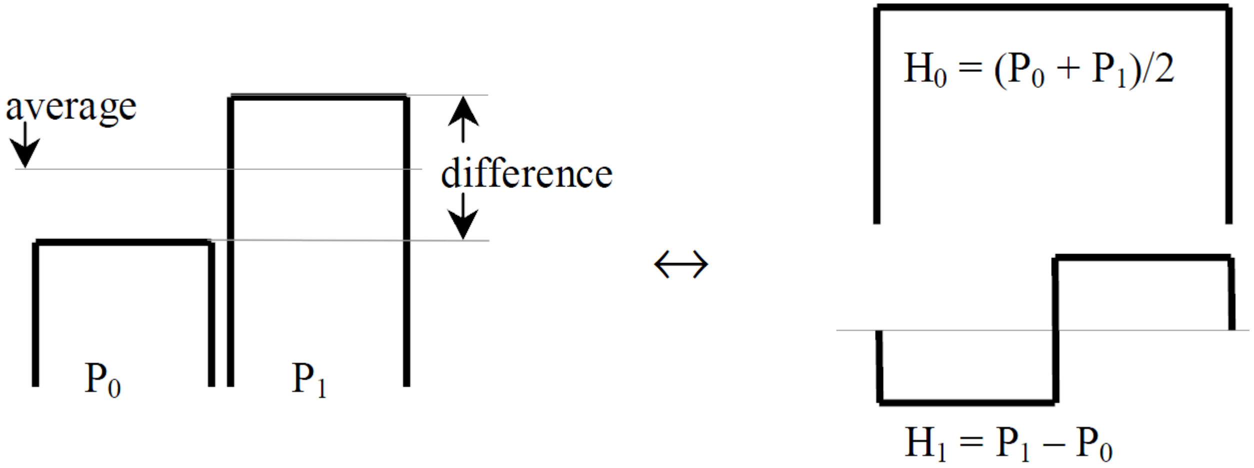
Assignment 5-3

(b) Convince yourself that PAA and DWT (using Haar Wavelets as basis function!) are equivalent!



Assignment 5-3

(b) Convince yourself that PAA and DWT (using Haar Wavelets as basis function!) are equivalent!



Given the case that the number of coefficients of the DWT is a power of two it is always possible to convert between the representations (PAA and Haar)

Source:
Keogh E. et. Al. - Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases; KAIS Long paper (2000)

Assignment 5-3

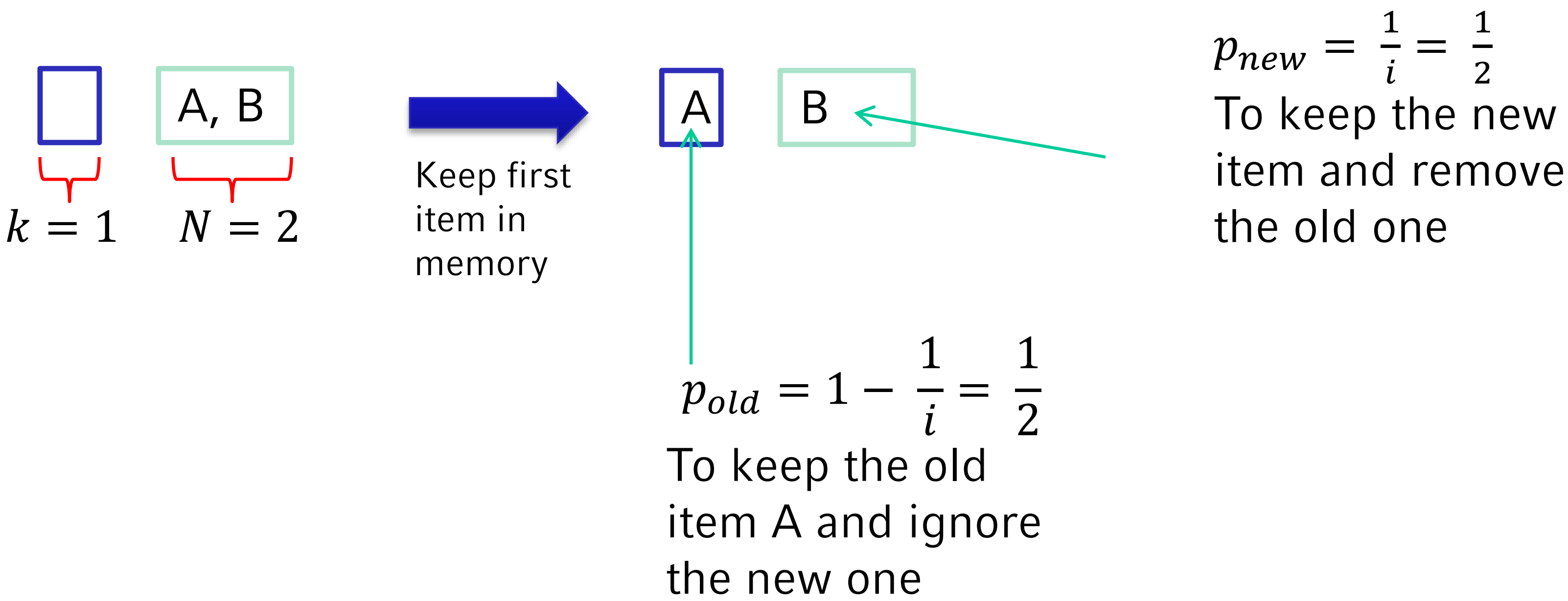
(b) Convince yourself that PAA and DWT (using Haar Wavelets as basis function!) are equivalent!

Mean	Coefficients
(4,1,2,3,6,1,7,6)	(-)
(2.5, 2.5, 3.5, 6.5)	(1.5, -0.5, 2.5, 0.5)
(2.5, 5)	(0, -1.5)
(3.75)	(-1.25)

Assignment 5-4

Given a data stream of size N. Randomly select $k \leq N$ elements from the stream. Here k represents the size of the reservoir.

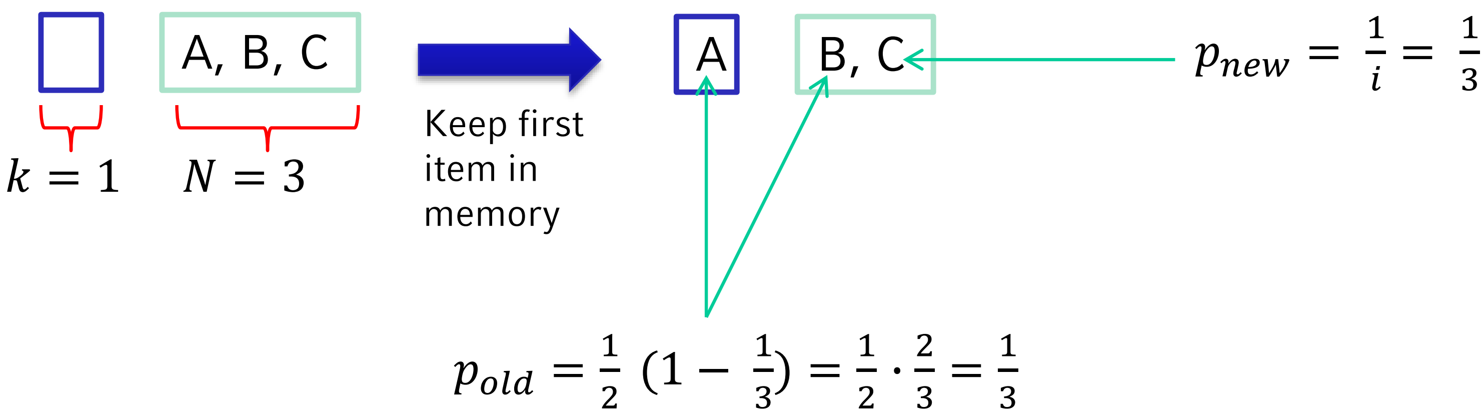
(a) Setting $k = 1, N = 2$. The first element is in the reservoir, the second is not. What is the probability of both elements to be in the reservoir?



Assignment 5-4

Given a data stream of size N. Randomly select $k \leq N$ elements from the stream. Here k represents the size of the reservoir.

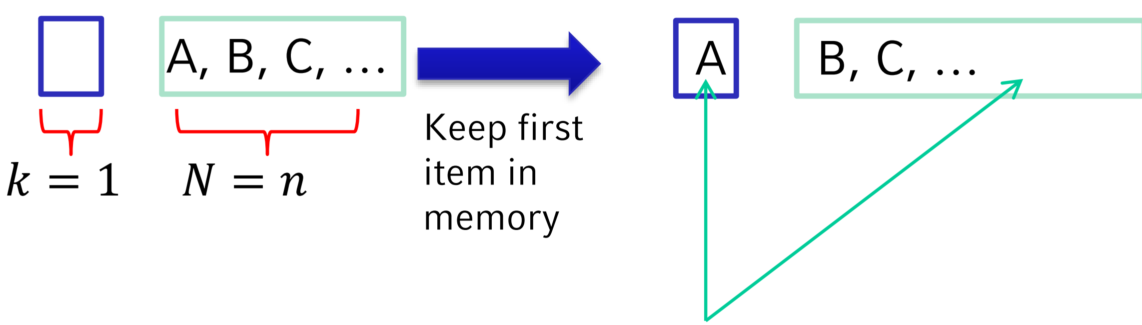
(b) Setting $k = 1, N = 3$. What is now the probability for each of the elements to be in the reservoir?



Assignment 5-4

Given a data stream of size N. Randomly select $k \leq N$ elements from the stream. Here k represents the size of the reservoir.

(c) Setting $k = 1$. What is the probability for any given N?

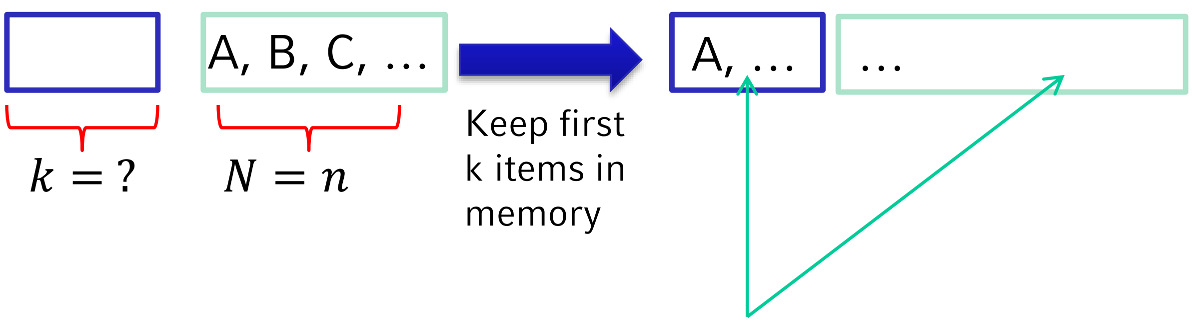


$$p_{old} = \frac{1}{2} \cdot \frac{2}{3} \cdot \frac{3}{4} \dots \cdot \frac{i-2}{i-1} \cdot \frac{i-1}{i} = \frac{1}{i}$$

Assignment 5-4

Given a data stream of size N. Randomly select $k \leq N$ elements from the stream. Here k represents the size of the reservoir.

(d) What is the probability for an arbitrary reservoir size k and an abitrary stream size N?



$$p_{new} = \frac{k}{i}$$

$$p_{old} = \frac{k}{k+1} \cdot \frac{k+1}{k+2} \cdot \frac{k+2}{k+3} \dots \cdot \frac{k+i-k-1}{i} = \frac{k}{i}$$