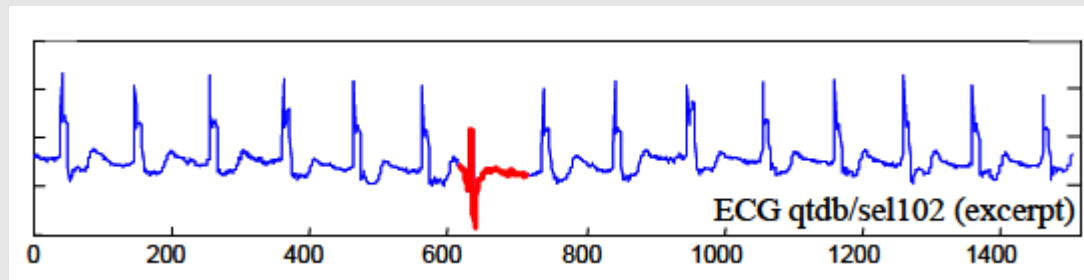


Idősor bányászat – abnormális jelenségek keresése

- hibadetektálás
- érdekesség
- anomáliadetektálás
- meglepetésszerű viselkedés ...



Feladatok:

indexelés (adott egy jel és egy hasonlósági mérce, melyik jel a leginkább hasonló rá?)

klaszterezés (adott egy hasonlósági mérce, mely jelek önmagukban leginkább egymásra hasonlítanak?)

osztályozás (adott egy jel és egy hasonlósági mérce, melyik jelosztályba sorolható?)

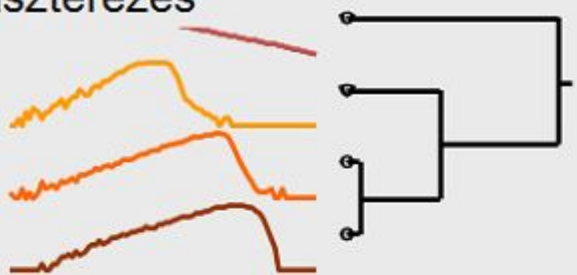
kivonatoltatás (adott egy extrémálisan hosszú jel, mi a lényegesen rövidített, de alakilag rá hasonlító közelítése?)

anomáliadetektálás (adott egy jel és a “normális” viselkedés valamilyen modellje, hol van a jelben a “meglepő”, az „érdekes”, a „váratlan”, az „újszerű”, ...?)

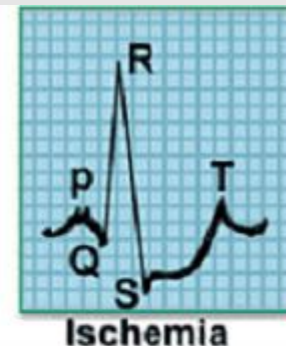
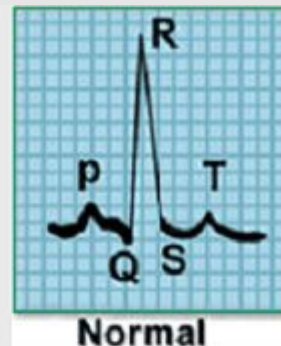
motivúm-felfedezés (többszörös, igen hasonló szakaszok észrevétele)

Mindenhol: hasonlóság detektálása!

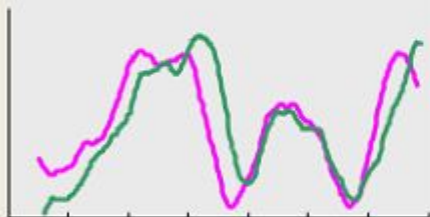
klaszterezés



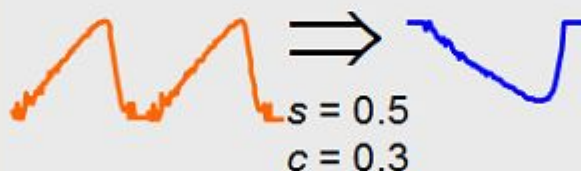
osztályozás



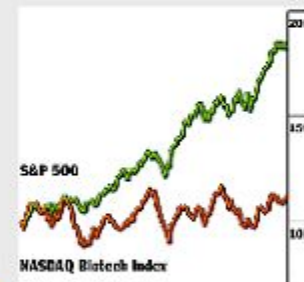
motivúm



szabály felfedezés

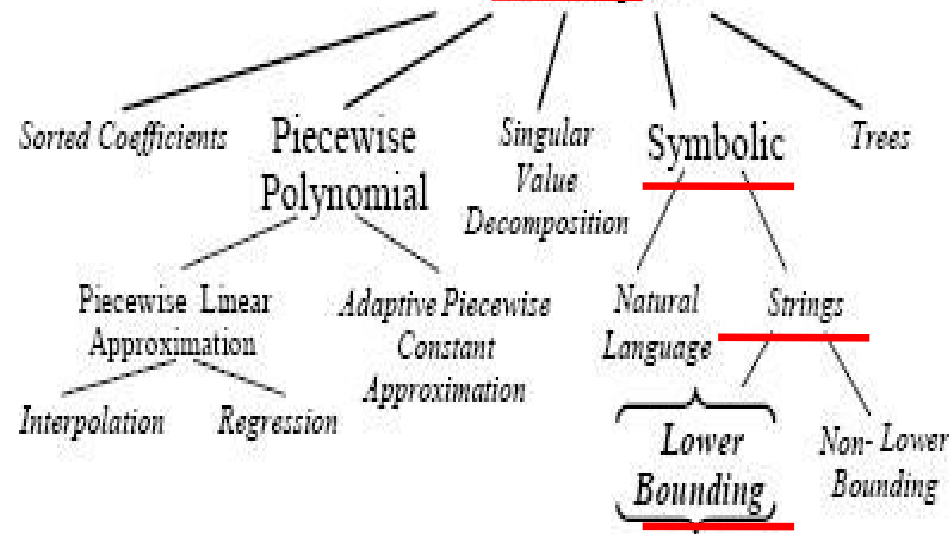


lekérdezés

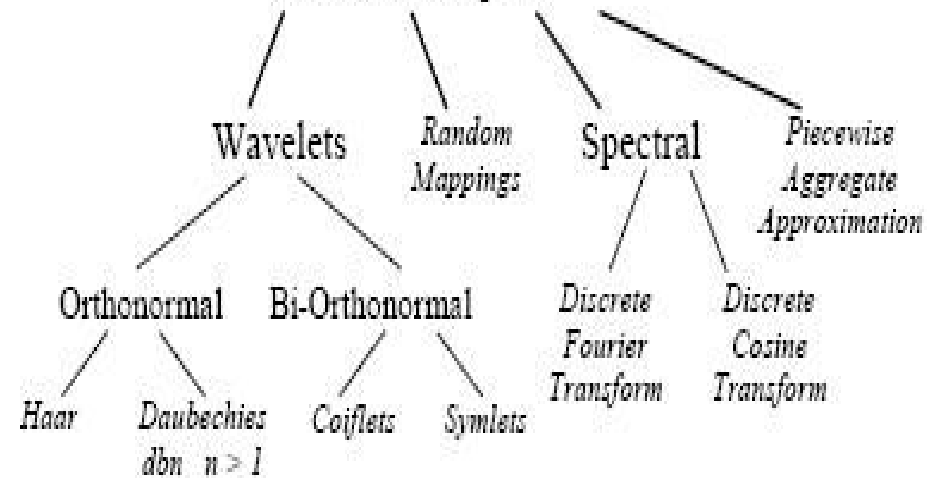


Time Series Representations

Data Adaptive



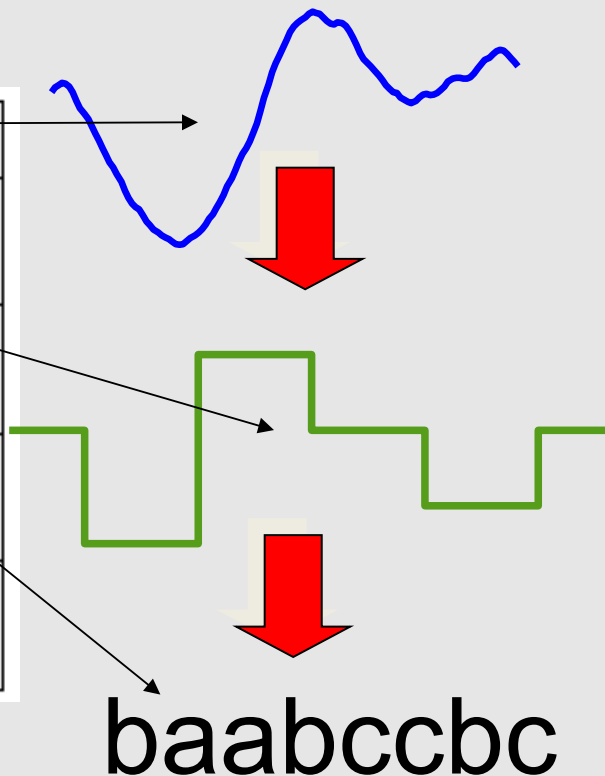
Non Data Adaptive



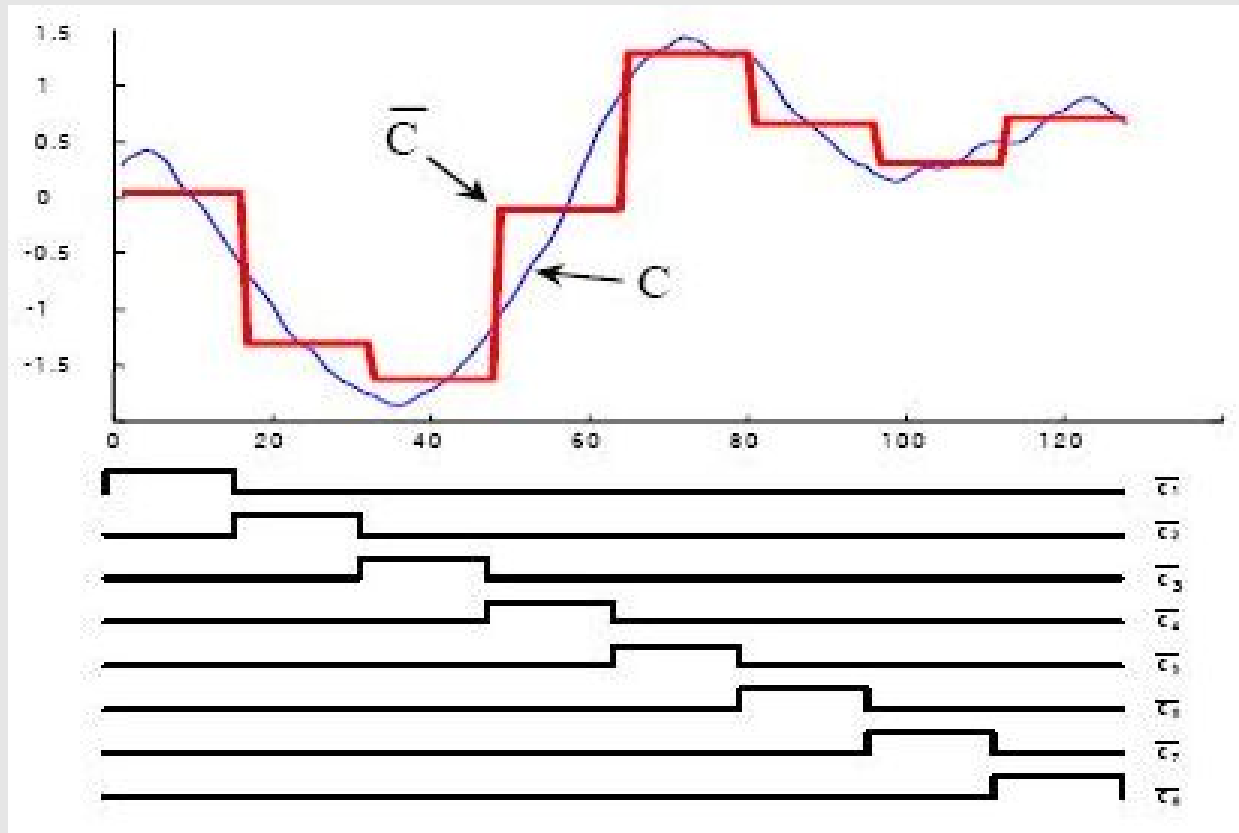
Idősor szimbolikus reprezentációja

SAX - **S**ymbolic **A**ggregate **A**ppro**X**imation

C	A time series $C = c_1, \dots, c_n$
$\overline{C}$	A Piecewise Aggregate Approximation of a time series $\overline{C} = \overline{c}_1, \dots, \overline{c}_w$
$\hat{C}$	A symbol representation of a time series $\hat{C} = \hat{c}_1, \dots, \hat{c}_w$
w	The number of PAA segments representing time series C
a	Alphabet size (e.g., for the alphabet = $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, $a = 3$)

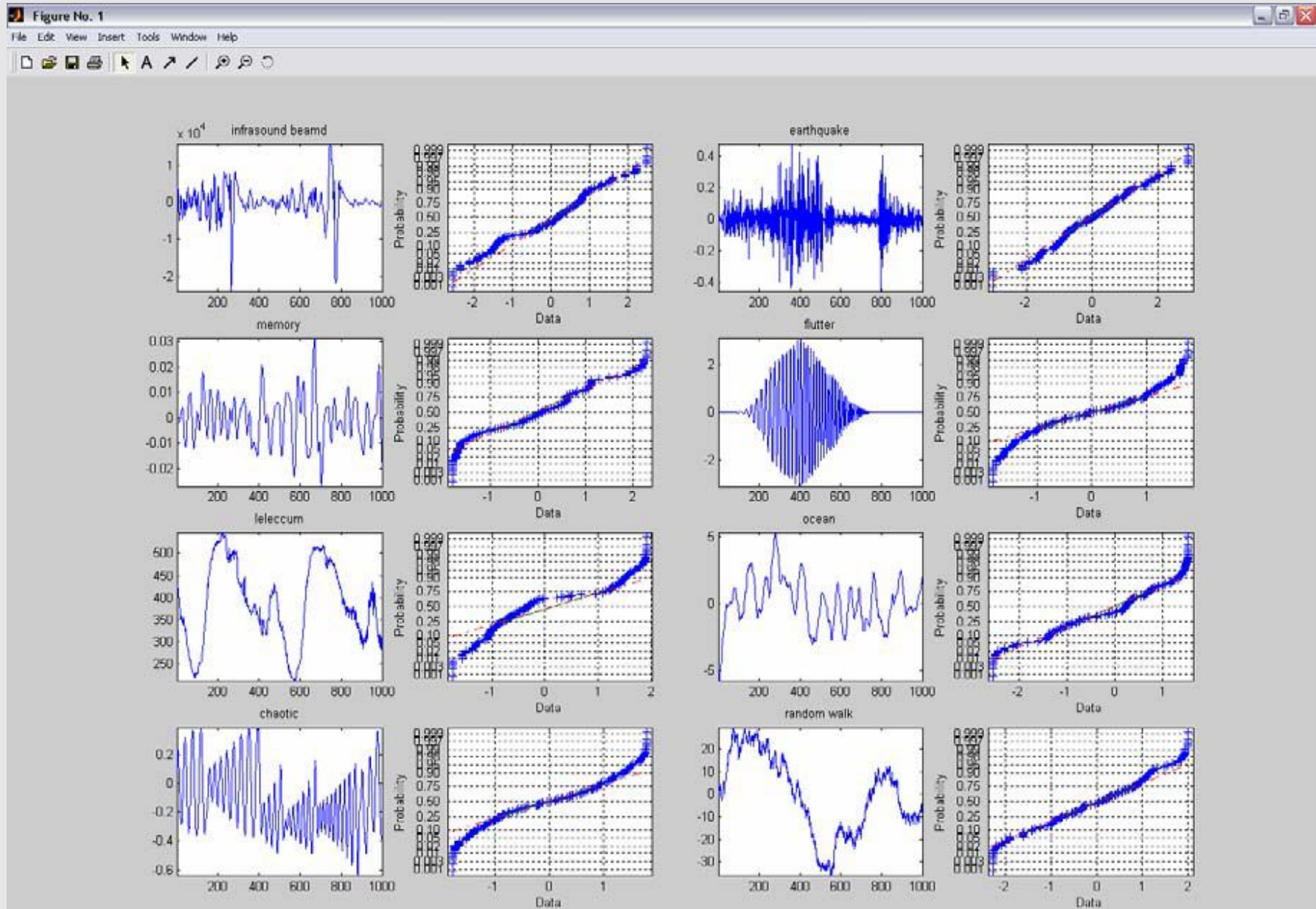


Dimenzióredukció PAA (**P**iecewise **A**ggregate **A**pproximation)



$$\bar{C}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j$$

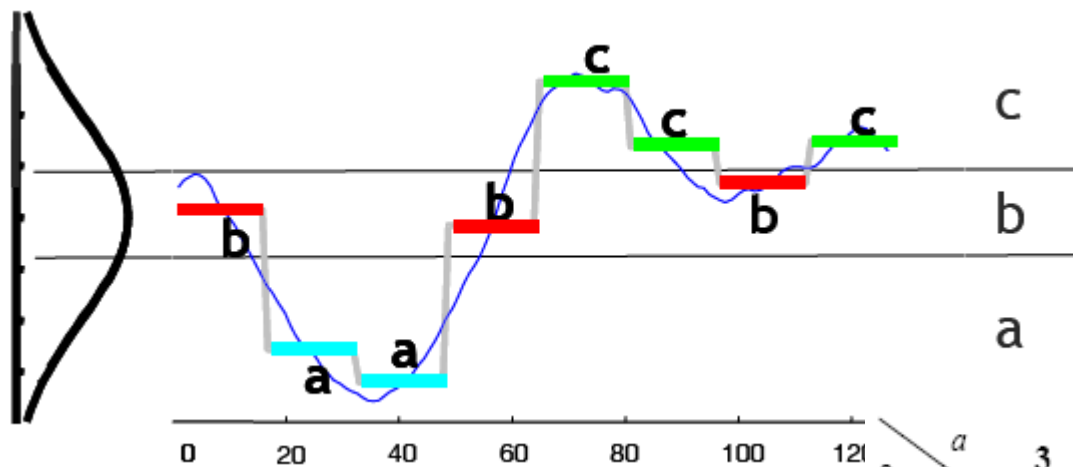
Gyakorlati jelek eloszlása – Gauss hipotézis



Diszkretizálás

1. amplitudó eloszlás – Gauss (feltételezés)
2. $\beta_k - \beta_{k+1}$ között azonos az amplitudó valószínűsége = $1/a$

$$\hat{c}_i = \alpha_j, \quad \text{iif} \quad \beta_{j-1} \leq \bar{c}_i < \beta_j$$



β_i	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

És most a távolságok (hasonlóságok)

Euklideszi távolság

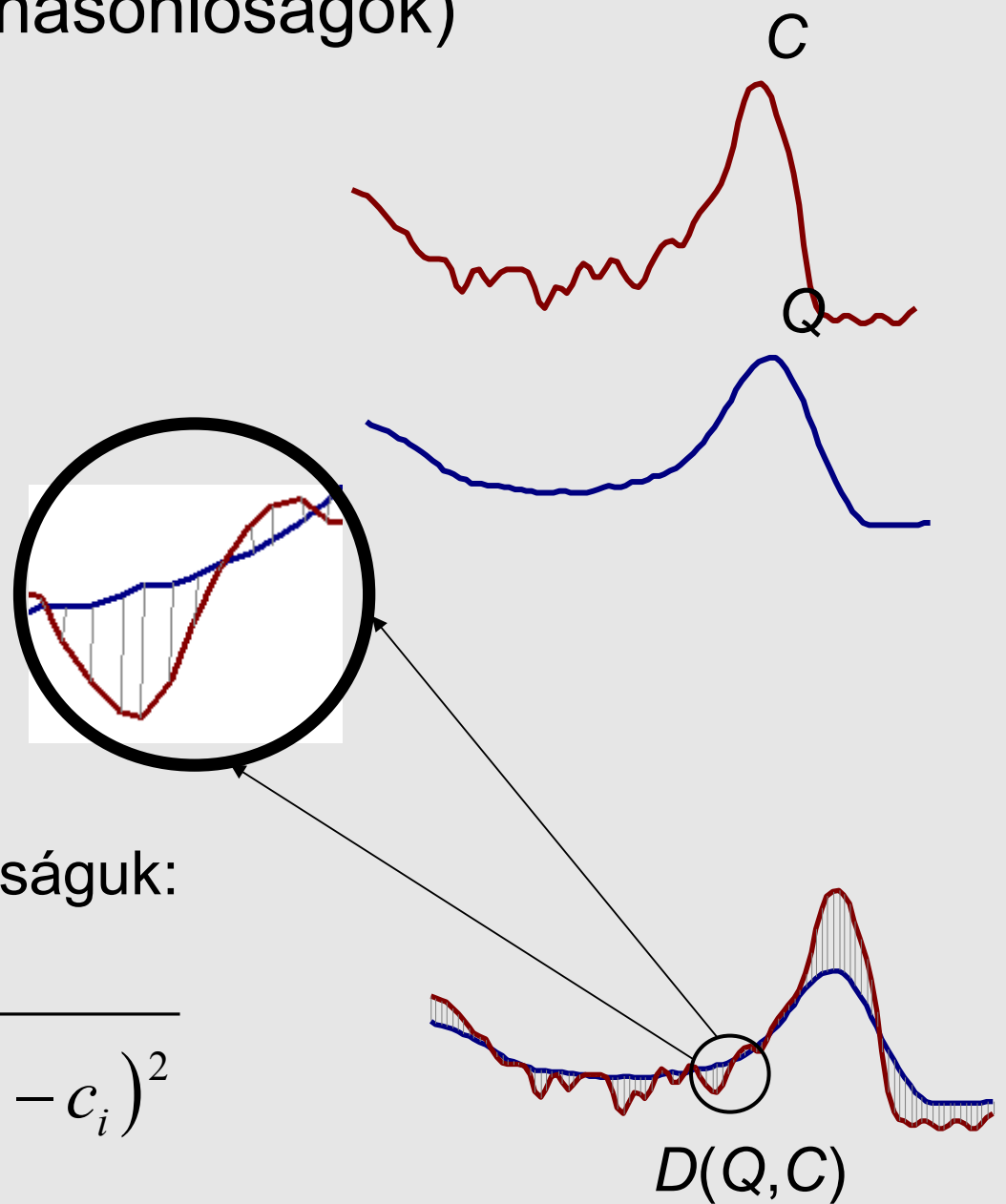
Két idősor

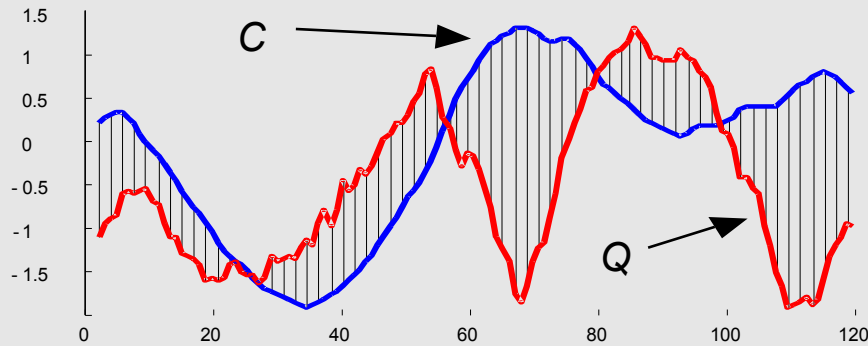
$$Q = q_1 \dots q_n$$

$$C = c_1 \dots c_n$$

esetén az euklideszi távolságuk:

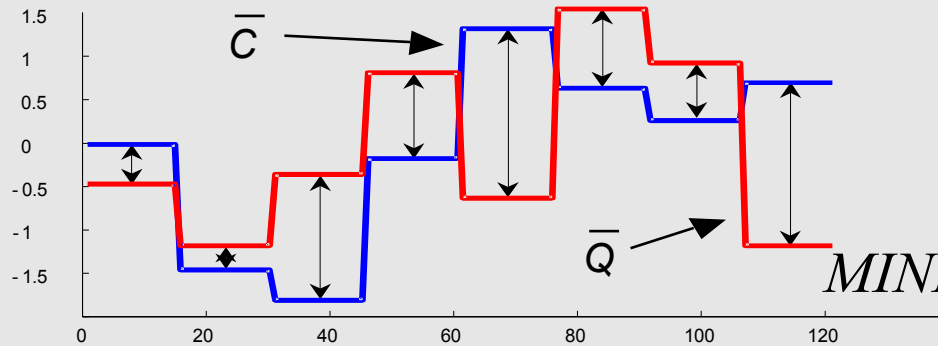
$$D(Q, C) \equiv \sqrt{\sum_{i=1}^n (q_i - c_i)^2}$$





$$D(Q, C) \equiv \sqrt{\sum_{i=1}^n (q_i - c_i)^2}$$

Euklideszi távolság



$$DR(\bar{Q}, \bar{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w (\bar{q}_i - \bar{c}_i)^2}$$

PAA távolság

$$MINDIST(\hat{Q}, \hat{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w (dist(\hat{q}_i, \hat{c}_i))^2}$$

szimb. távolság

$\hat{C}$ = **baacbcda**



$\hat{Q}$ = **babcacca**

dist()

	a	b	c	d
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

$$cell_{r,c} = \begin{cases} 0, & \text{if } |r - c| \leq 1 \\ \beta_{\max(r,c)-1} - \beta_{\min(r,c)}, & \text{otherwise} \end{cases}$$

Alulról korlátos (Lower Bound) távolsági fogalom

Alulról korlátos azt jelenti, hogy minden Q és S esetén:

$$D(\hat{C}, \hat{S}) \leq D(C, S)$$

PAA távolság alulról közelíti az euklideszi távolságot

$$\sqrt{\sum_{i=1}^N (q_i - c_i)^2} \geq \sqrt{\frac{N}{w}} \sqrt{\sum_{i=1}^w (\bar{q}_i - \bar{c}_i)^2}$$

$N = n \times w$ teljes jelhossz, n db PAA szakasz

Egy szegmenst vizsgáljuk ($w = 1$),
a teljes képlet a szegmensek szerinti
összegzéssel nyerhető

$$\sqrt{\sum_{i=1}^n (q_i - c_i)^2} \geq \sqrt{n} \sqrt{(\bar{Q} - \bar{C})^2}$$

$$\sum_{i=1}^n (q_i - c_i)^2 \geq n (\bar{Q} - \bar{C})^2$$

$$q_i = \bar{Q} - \Delta q_i$$

$$\Delta q_i = \bar{Q} - q_i$$

$$\sum_{i=1}^n ((\bar{Q} - \Delta q_i) - (\bar{C} - \Delta c_i))^2 \geq n (\bar{Q} - \bar{C})^2$$

$$\sum_{i=1}^n ((\bar{Q} - \bar{C}) - (\Delta q_i - \Delta c_i))^2 \geq n (\bar{Q} - \bar{C})^2$$

$$\sum_{i=1}^n ((\bar{Q} - \bar{C})^2 - 2(\bar{Q} - \bar{C})(\Delta q_i - \Delta c_i) + (\Delta q_i - \Delta c_i)^2) \geq n (\bar{Q} - \bar{C})^2$$

$$\sum_{i=1}^n (\bar{Q} - \bar{C})^2 - \sum_{i=1}^n 2(\bar{Q} - \bar{C})(\Delta q_i - \Delta c_i) + \sum_{i=1}^n (\Delta q_i - \Delta c_i)^2 \geq n (\bar{Q} - \bar{C})^2$$

$$n (\bar{Q} - \bar{C})^2 - 2(\bar{Q} - \bar{C}) \sum_{i=1}^n (\Delta q_i - \Delta c_i) + \sum_{i=1}^n (\Delta q_i - \Delta c_i)^2 \geq n (\bar{Q} - \bar{C})^2$$



$$\begin{aligned} \sum_{i=1}^n (\Delta q_i - \Delta c_i) &= \sum_{i=1}^n ((\bar{Q} - q_i) - (\bar{C} - c_i)) \\ &= \left(\sum_{i=1}^n \bar{Q} - \sum_{i=1}^n q_i \right) - \left(\sum_{i=1}^n \bar{C} - \sum_{i=1}^n c_i \right) \\ &= (n\bar{Q} - \sum_{i=1}^n q_i) - (n\bar{C} - \sum_{i=1}^n c_i) \\ &= \left(\sum_{i=1}^n q_i - \sum_{i=1}^n q_i \right) - \left(\sum_{i=1}^n c_i - \sum_{i=1}^n c_i \right) \\ &= 0 - 0 \\ &= 0 \end{aligned}$$

$$n (\bar{Q} - \bar{C})^2 - 0 + \sum_{i=1}^n (\Delta q_i - \Delta c_i)^2 \geq n (\bar{Q} - \bar{C})^2$$



$$\sum_{i=1}^n (\Delta q_i - \Delta c_i)^2 \geq 0$$

Szimb. távolság alulról közelíti a PAA távolságot

$$n(\bar{Q} - \bar{C})^2 \geq n(\text{dist}(\hat{Q}, \hat{C}))^2$$

$$\text{dist}('c', 'a') = \beta_2 - \beta_1$$

$$\hat{c}_i = \alpha_j, \quad \text{iif } \beta_{j-1} \leq \bar{c}_i < \beta_j$$

Case 1: $|(\hat{Q} - \hat{C})| \leq 1$.

Ilyenkor 0, nincs mit tenni

Case 2: $|(\hat{Q} - \hat{C})| > 1$



$$\text{dist}(\hat{Q}, \hat{C}) = \beta_{\hat{Q}-1} - \beta_{\hat{C}}$$

$$\begin{cases} \beta_{\hat{Q}-1} \leq \bar{Q} < \beta_{\hat{Q}} \\ \beta_{\hat{C}-1} \leq \bar{C} < \beta_{\hat{C}} \end{cases}$$

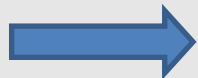


$$n(\bar{Q} - \bar{C})^2 \geq n(\beta_{\hat{Q}-1} - \beta_{\hat{C}})^2$$

$$|\bar{Q} - \bar{C}| \geq |\beta_{\hat{Q}-1} - \beta_{\hat{C}}|$$



$$|(\hat{Q} - \hat{C})| > 1 \text{ and } \hat{Q} > \hat{C}$$



$$\bar{Q} - \bar{C} \geq \beta_{\hat{Q}-1} - \beta_{\hat{C}}$$

$$\bar{Q} - \beta_{\hat{Q}-1} \geq \bar{C} - \beta_{\hat{C}}$$

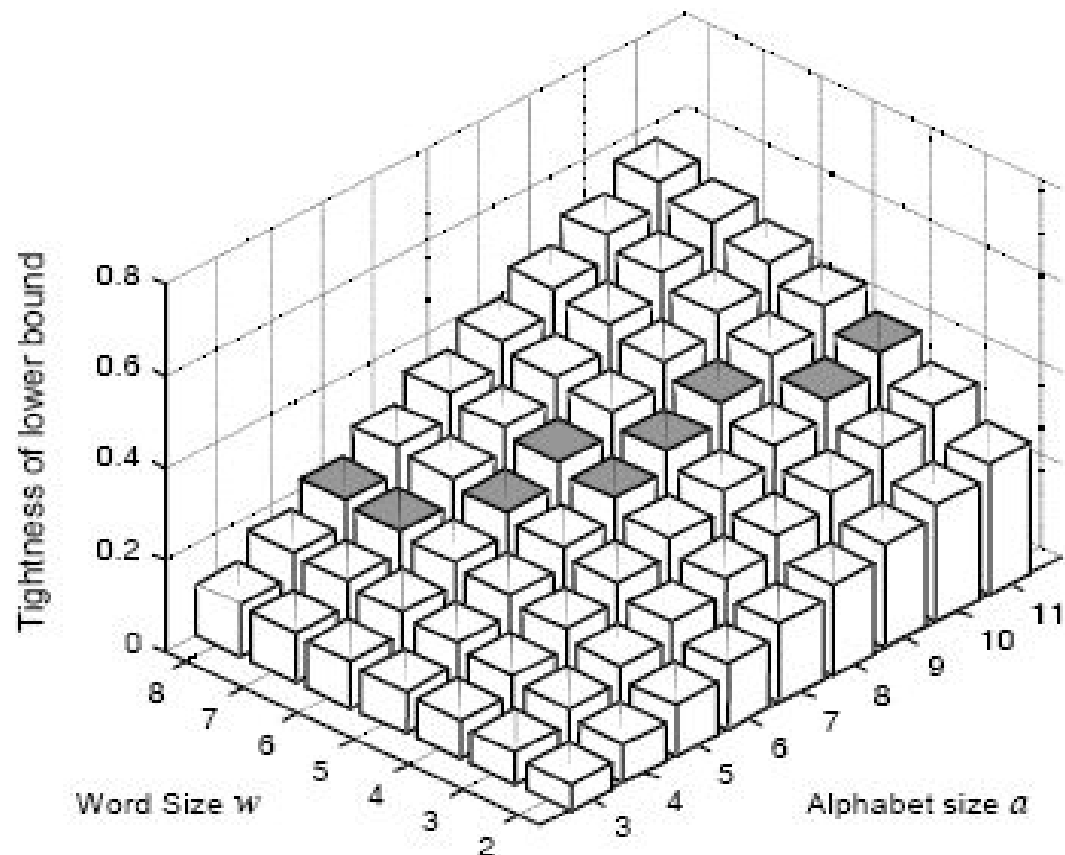


$$\begin{cases} \bar{Q} - \beta_{\hat{Q}-1} \geq 0 \\ \bar{C} - \beta_{\hat{C}} < 0 \end{cases}$$

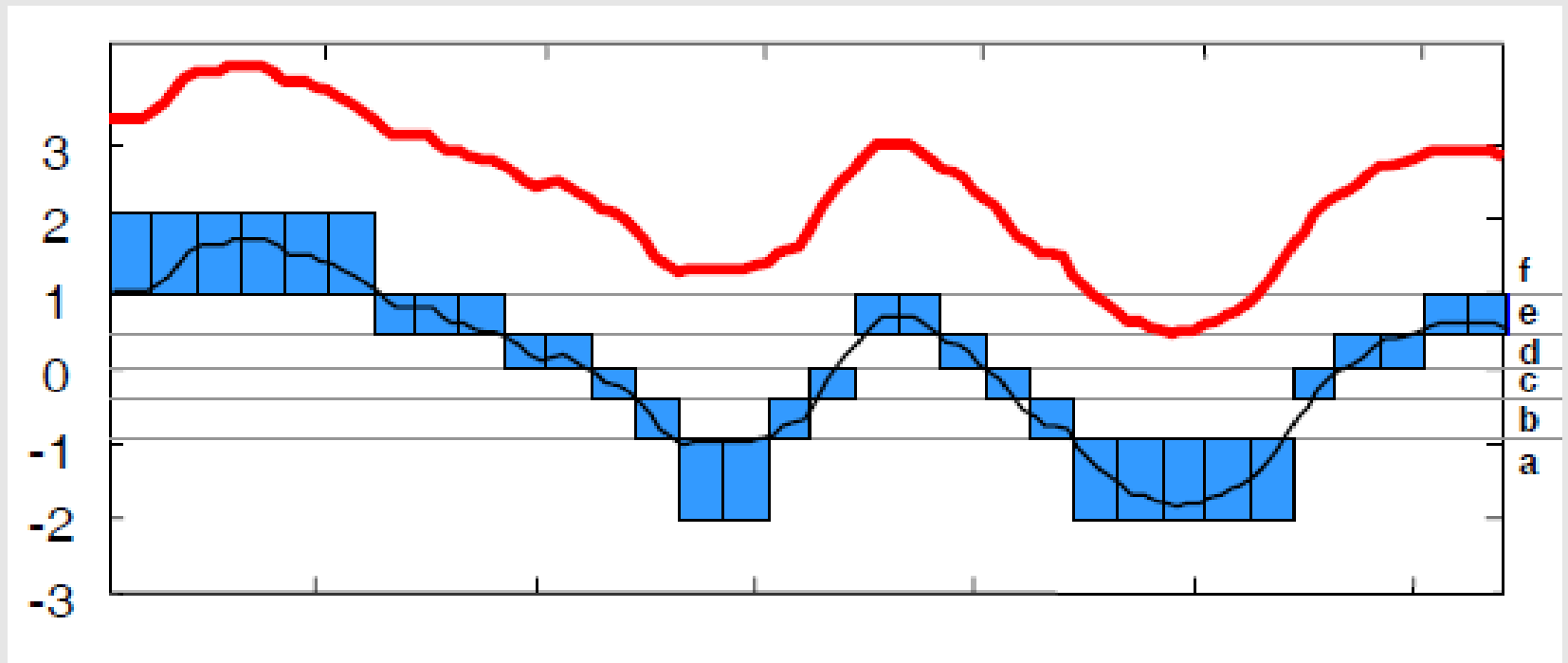
Avagy az egyenlőtlenség fennáll

Az alulról korlátozás szigorúsága adatredukció függvényében

$$\text{Tightness of Lower Bound} = \frac{MINDIST(\hat{Q}, \hat{C})}{D(Q, C)}$$



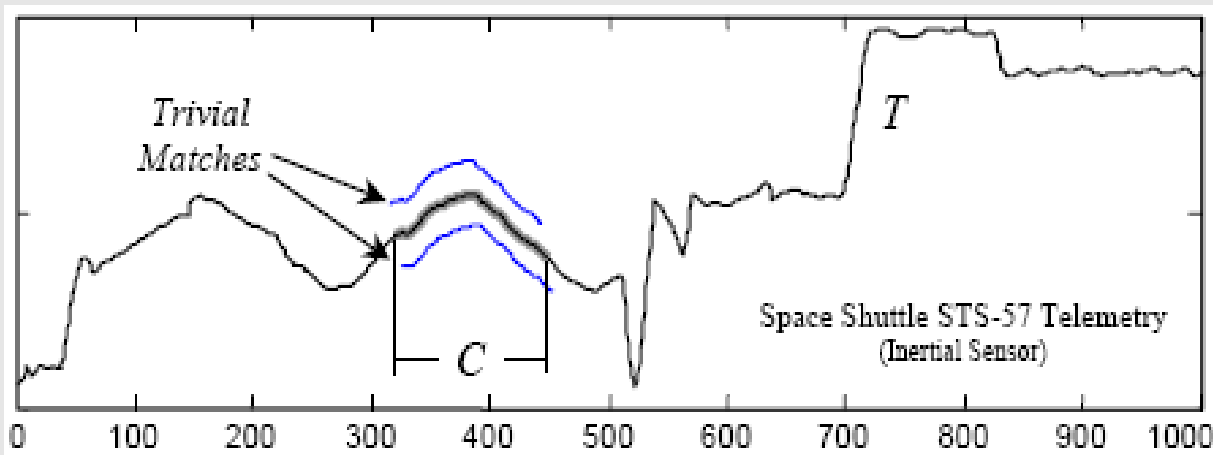
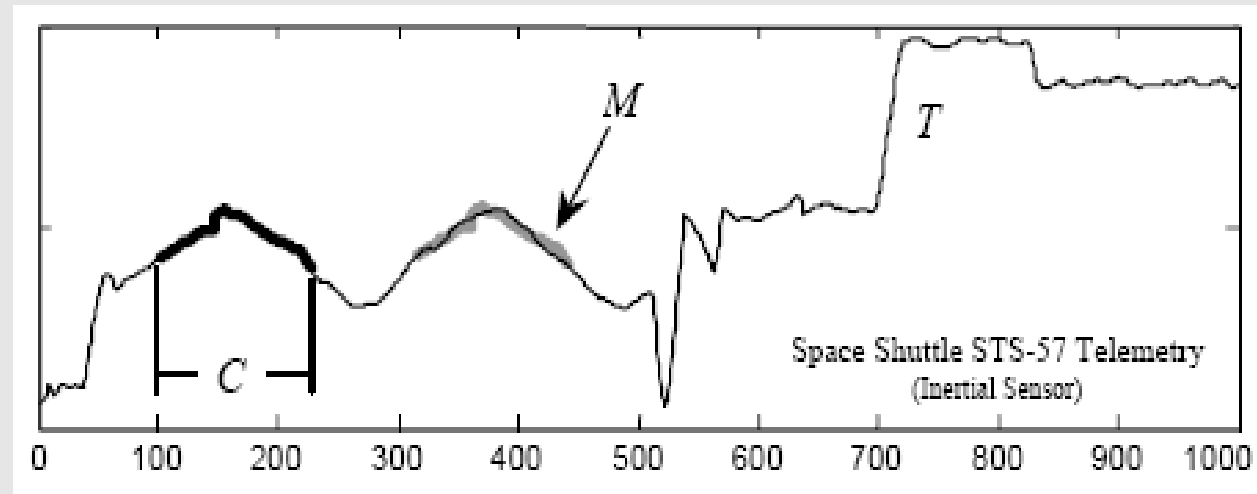
SAX reprezentáció, példa



És most az anomália fogalma felé

Definíciók:

non-self-match =
„igazán” hasonló
jelszakasz, idősor
anomália, idősor
k-adik anomália



és a definíciók
mögötti megfontolások

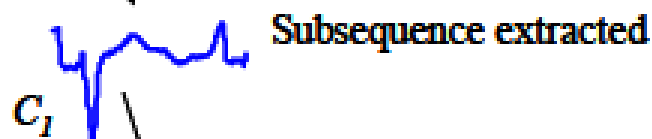
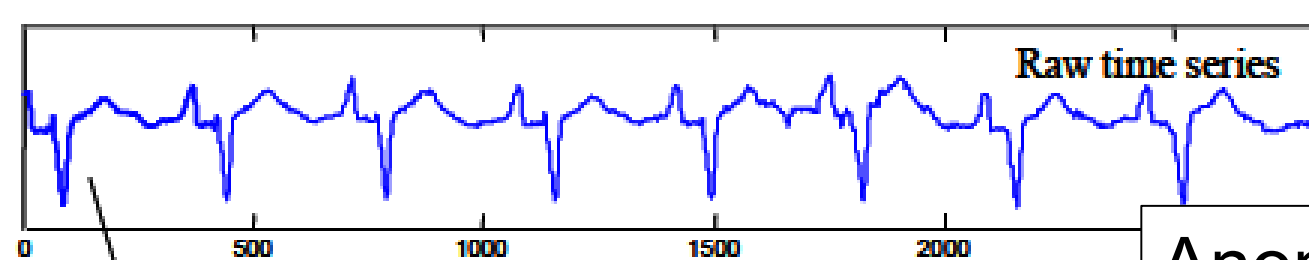
Tovább:

BIR-6e-SAX-anomalia-detektalas-2a-2013

...

...

BIR-6e-SAX-anomalia-detektalas-2f-2013



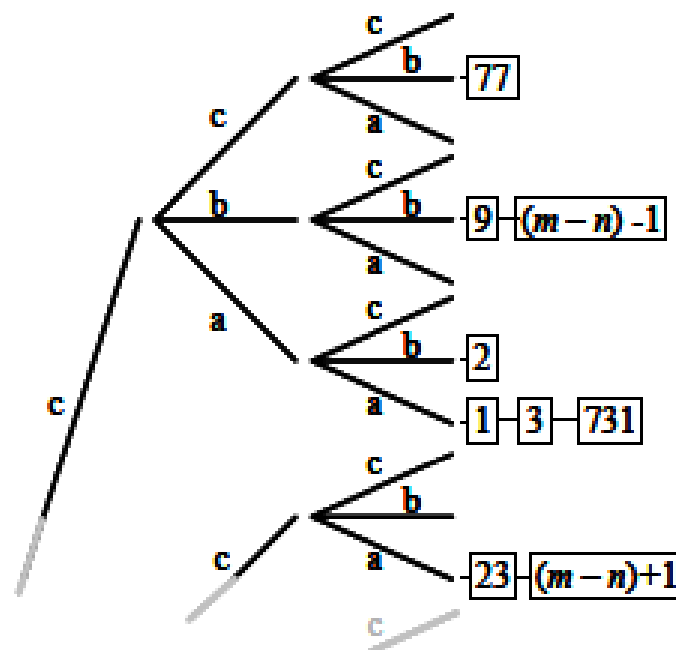
Converted to SAX

$\hat{C}_1$ **c a a**

Inserted into array

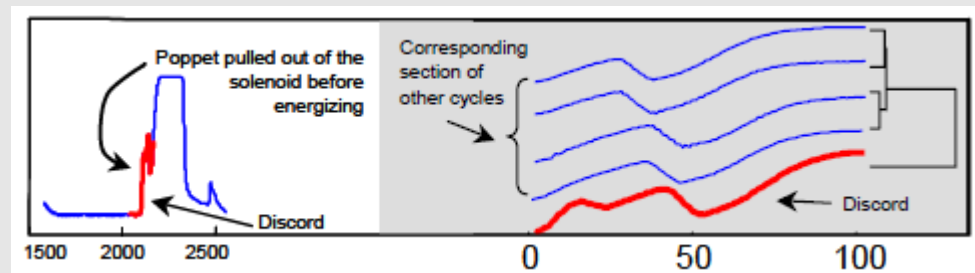
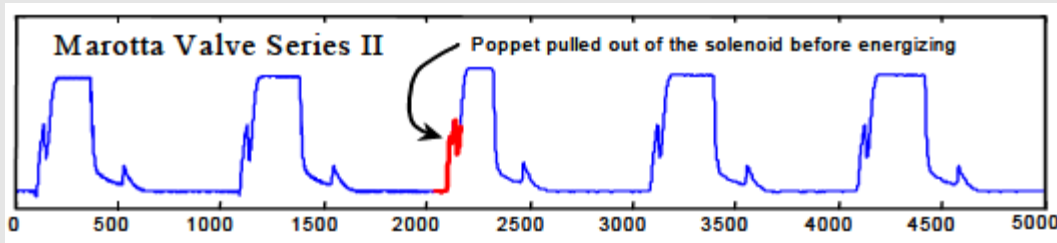
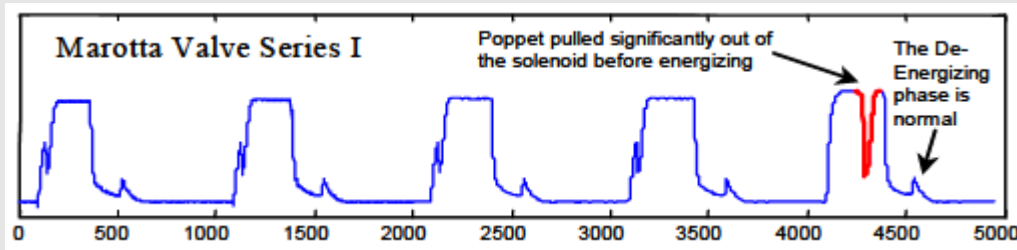
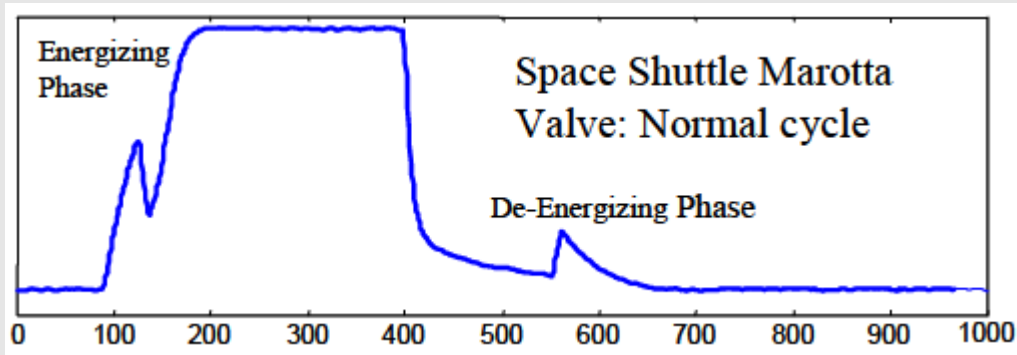
1	c	a	a	3
2	c	a	b	1
3	c	a	a	3
::	::	::	::	::
::	::	::	::	::
$(m-n)-1$	c	b	b	2
$(m-n)$	a	c	b	1
$(m-n)+1$	b	c	a	2

Augmented Trie

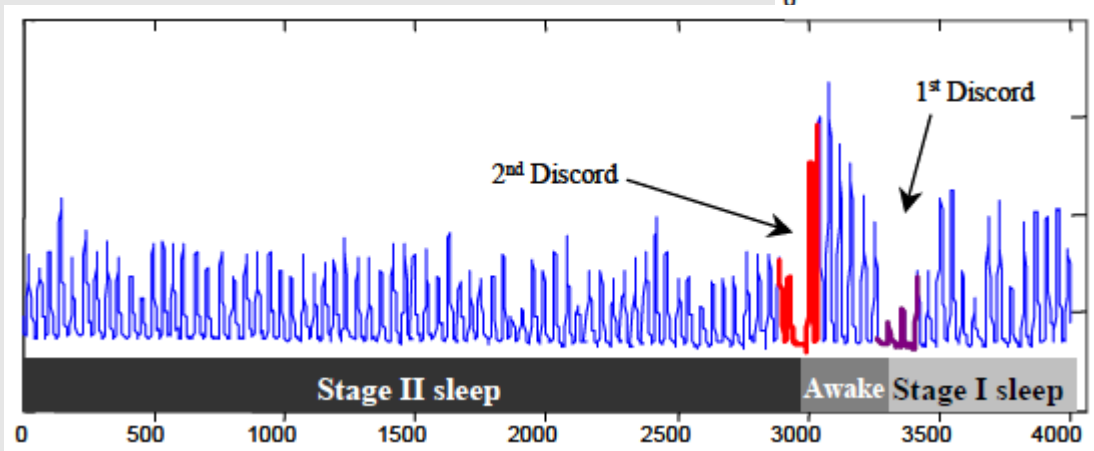
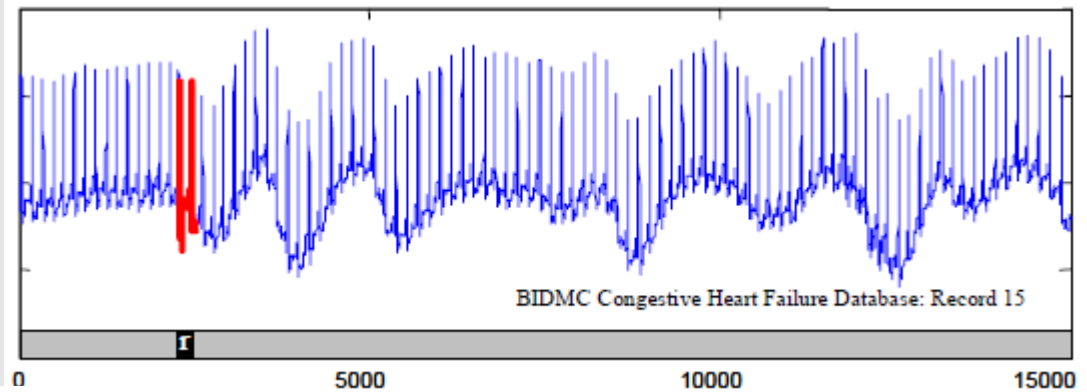
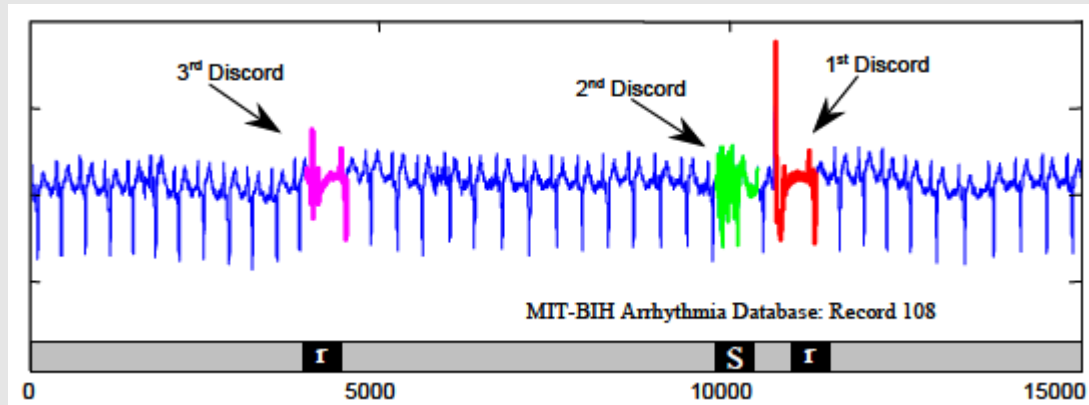


Anomáliadetektálás
– SAX struktúrák
heurisztika
számításához

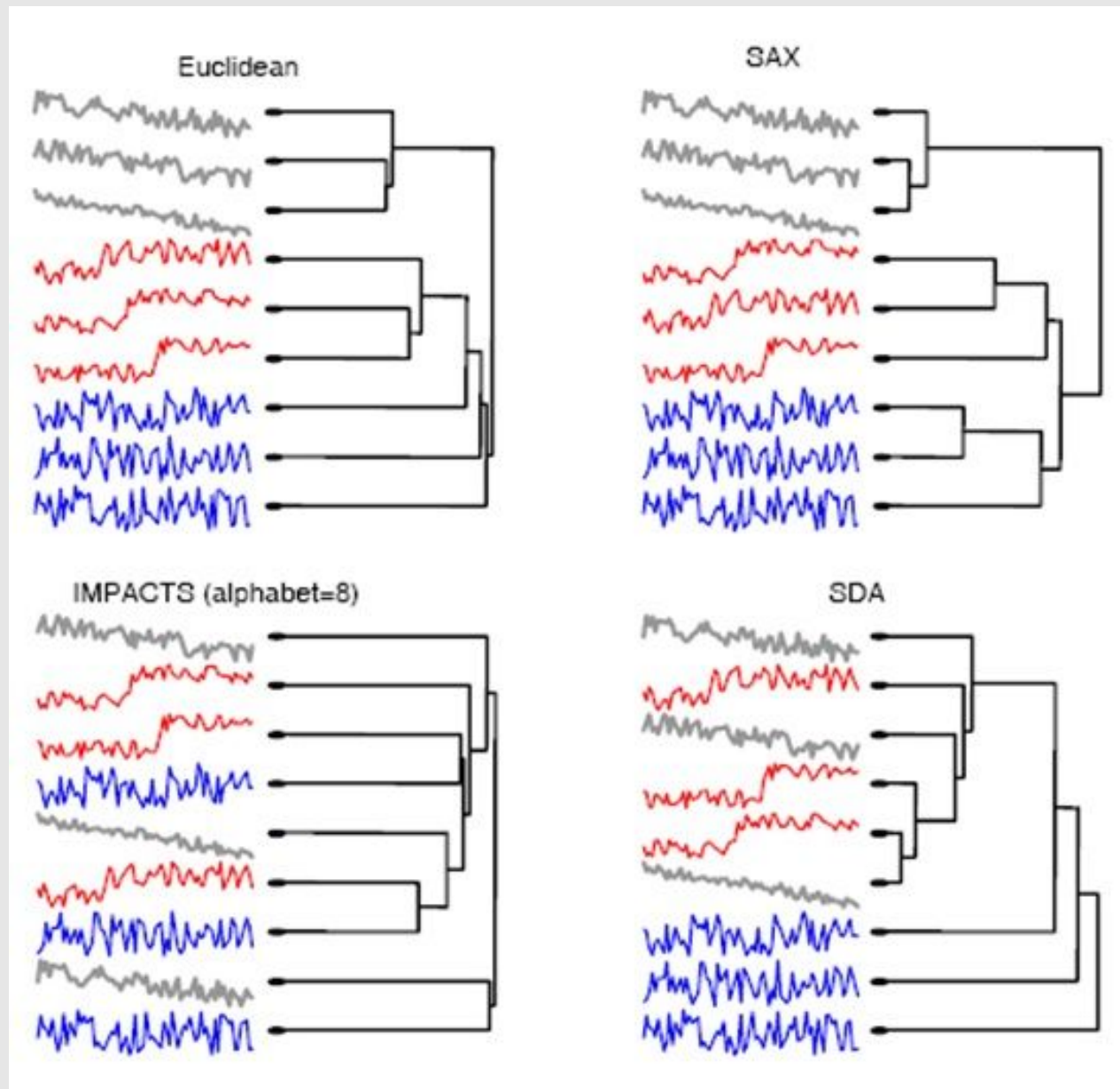
Anomáliadetektálás – űrsikló diagnosztika



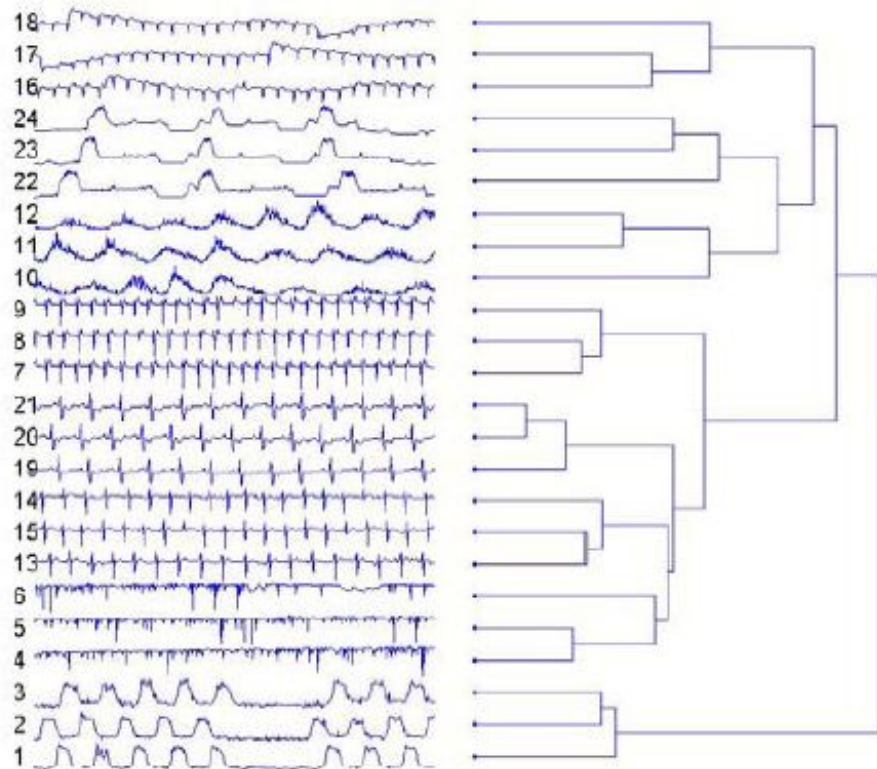
Anomáliadetektálás – EKG, légzés



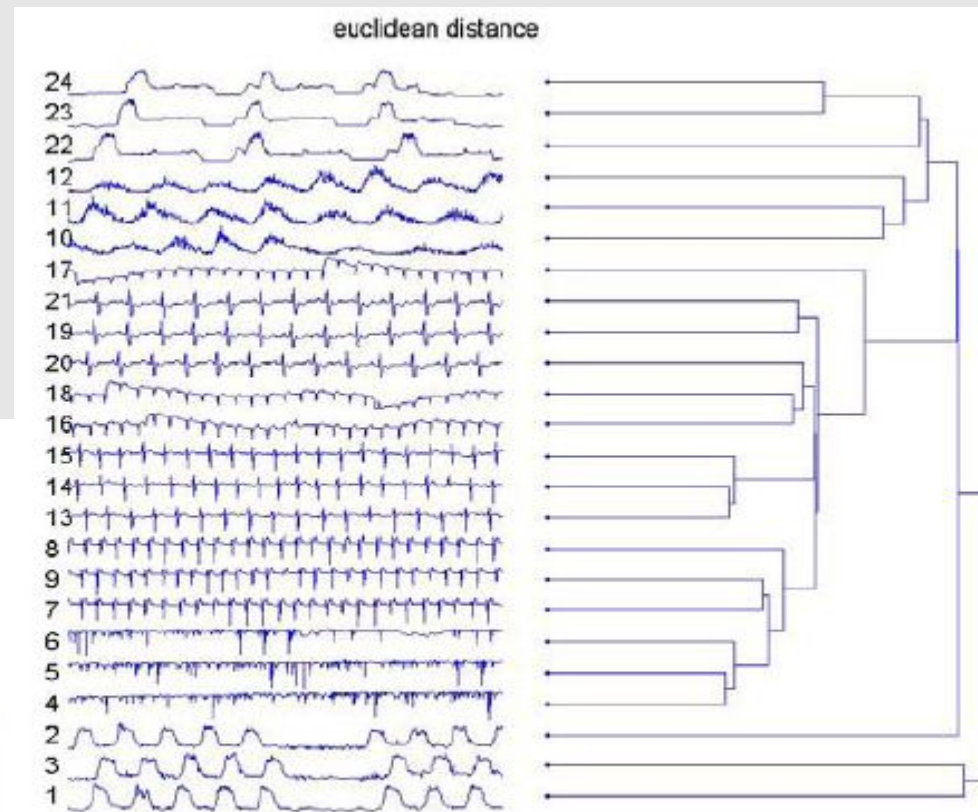
Anomáliadetektálás – klaszterezés



Anomáliadetektálás – klaszterezés



Clustering result using the (dis)similarity coefficient



Clustering result using Euclidean distance

Anomáliadetektálás – motívum felfedezés

