Využití zlomkových stochastických procesů pro analýzu signálu a časových řad

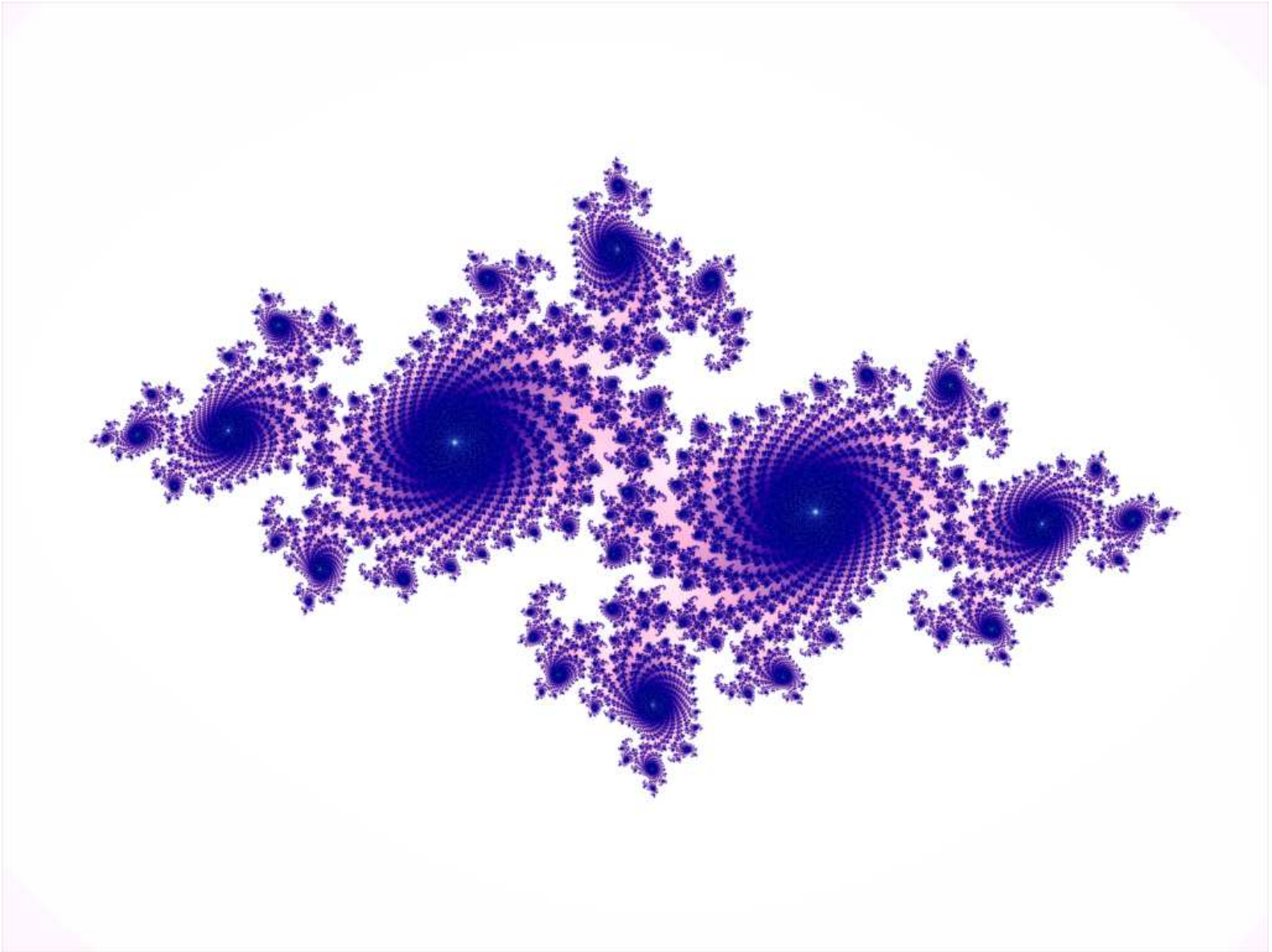
Seminář strojového učení a modelování

Martin Dlask (KSI FJFI)

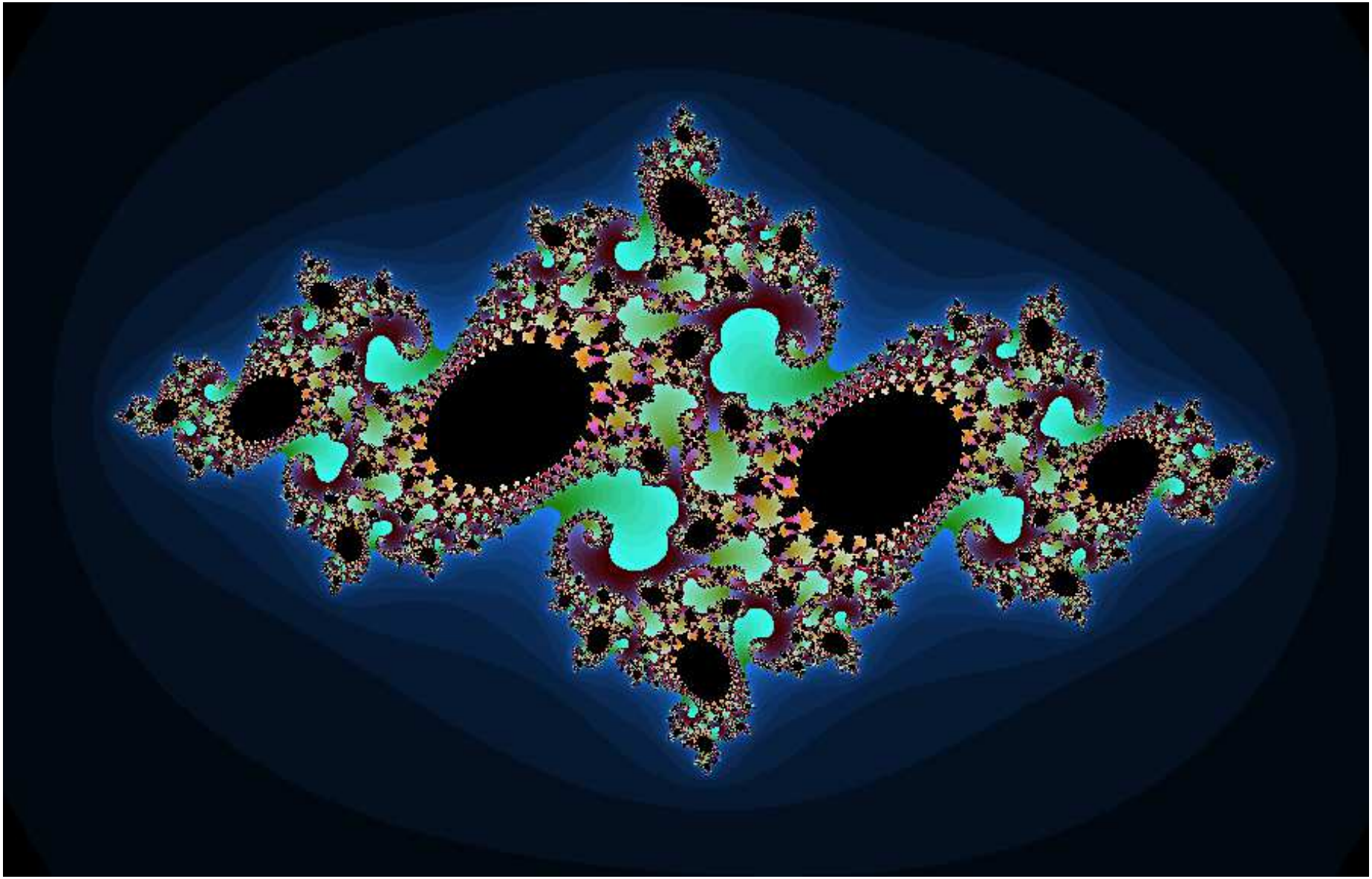
http://people.fjfi.cvut.cz/dlaskma1/

3. března 2016

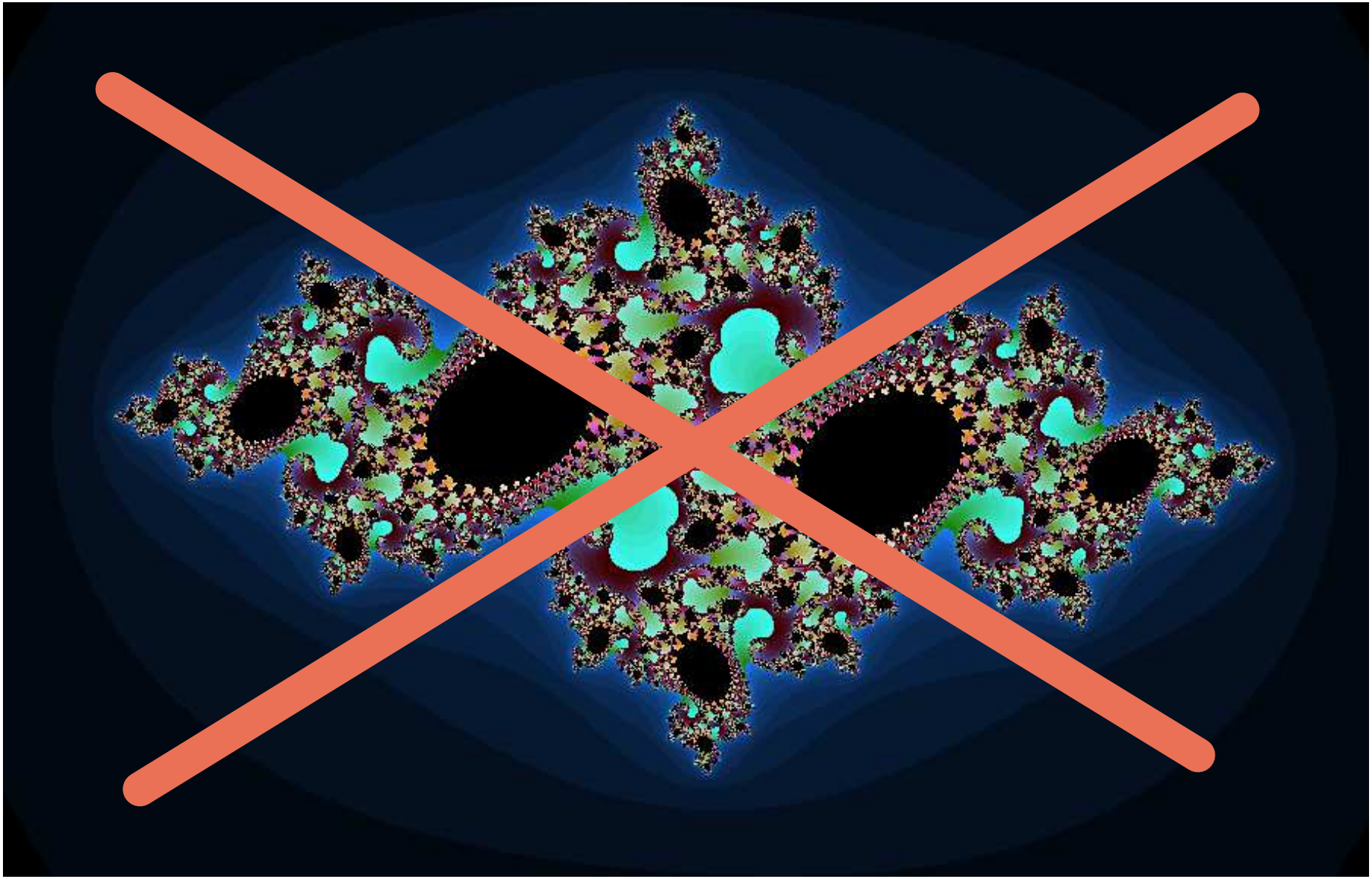
Juliova množina



Juliova množina



Juliova množina



Fraktální množiny

- Tak co jsou vlastně fraktály? (pokud ne jen barevné obrázky)

Fraktální množiny

- Tak co jsou vlastně fraktály? (pokud ne jen barevné obrázky)

  $\rightarrow$ množiny s neceločíselnou dimenzí

Fraktální množiny

- Tak co jsou vlastně fraktály? (pokud ne jen barevné obrázky)

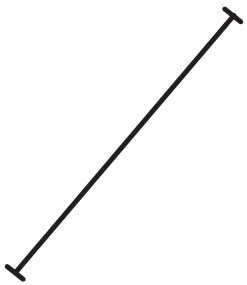
  $\rightarrow$ množiny s neceločíselnou dimenzí

Fraktální množiny

- Tak co jsou vlastně fraktály? (pokud ne jen barevné obrázky)

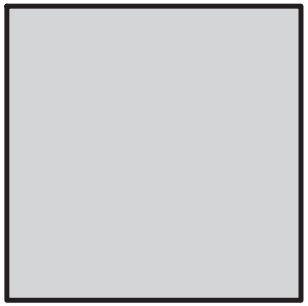
  $\rightarrow$ množiny s neceločíselnou dimenzí

úsečka v $\mathbf{R^2}$



dimenze = 1

čtverec v $\mathbf{R^2}$



dimenze = 2

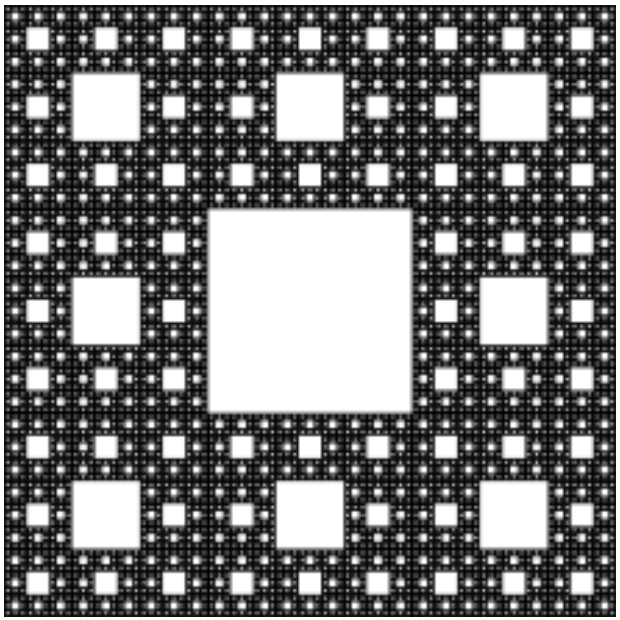
Fraktální množiny

- Tak co jsou vlastně fraktály? (pokud ne jen barevné obrázky)

  $\rightarrow$ množiny s neceločíselnou dimenzí

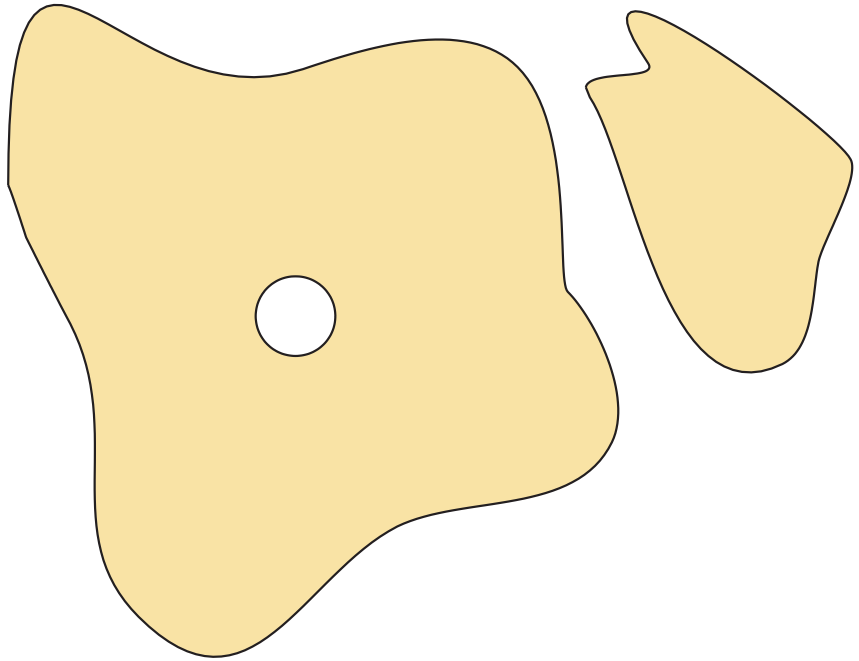
Fraktální množiny

- množina je charakterizována svojí
  - mírou
  - fraktální dimenzí - Hausdorffova dimenze



Fraktální množiny

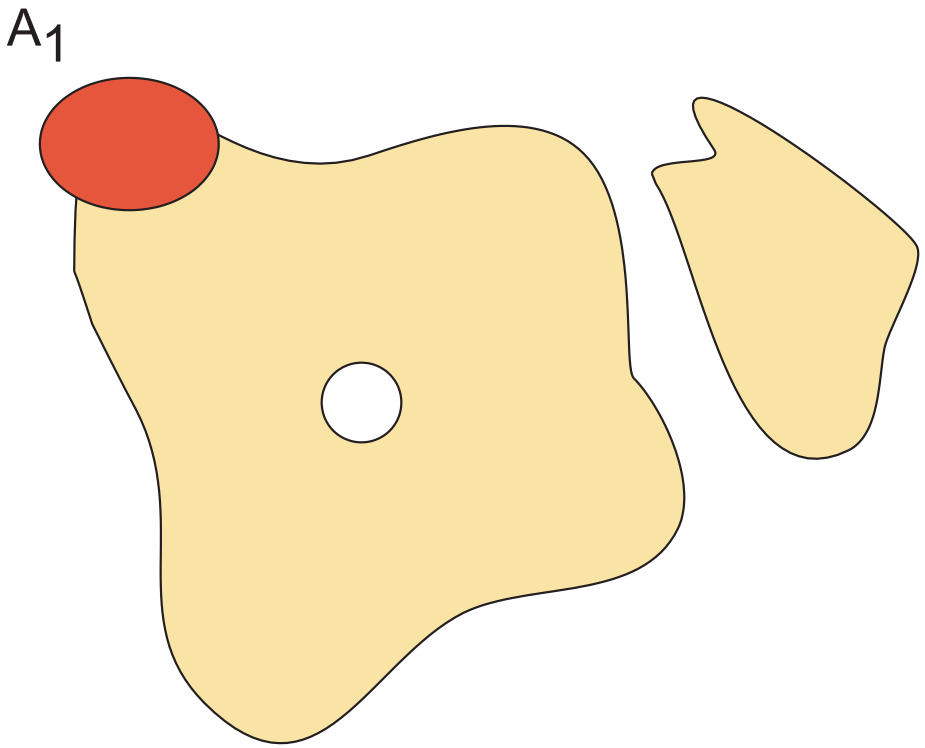
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

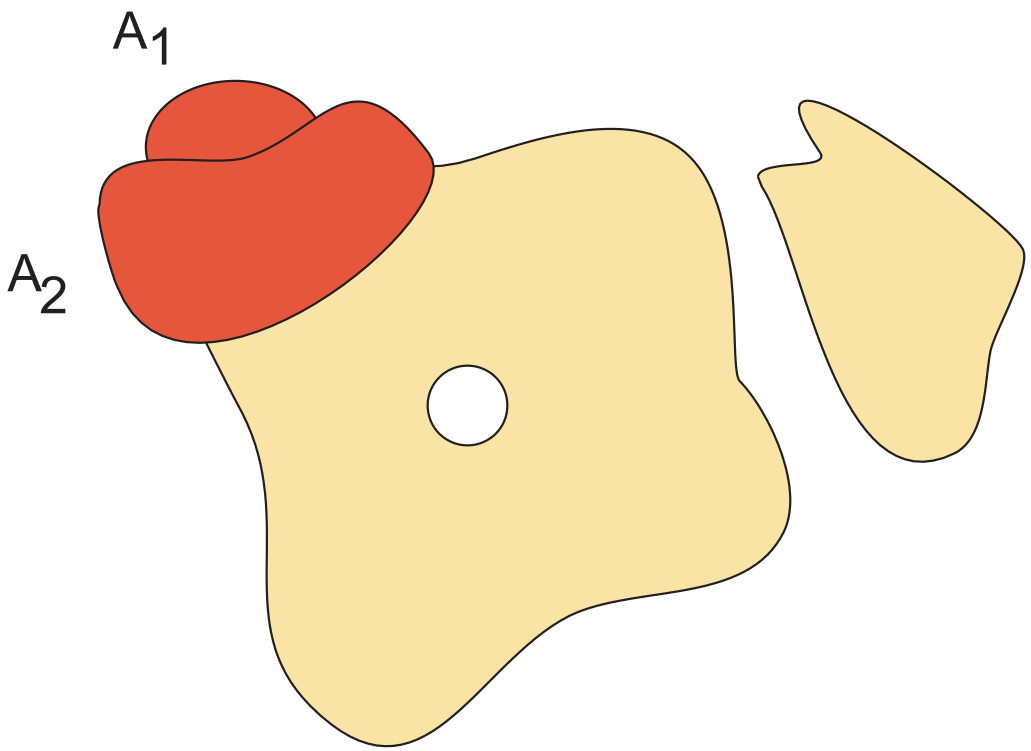
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

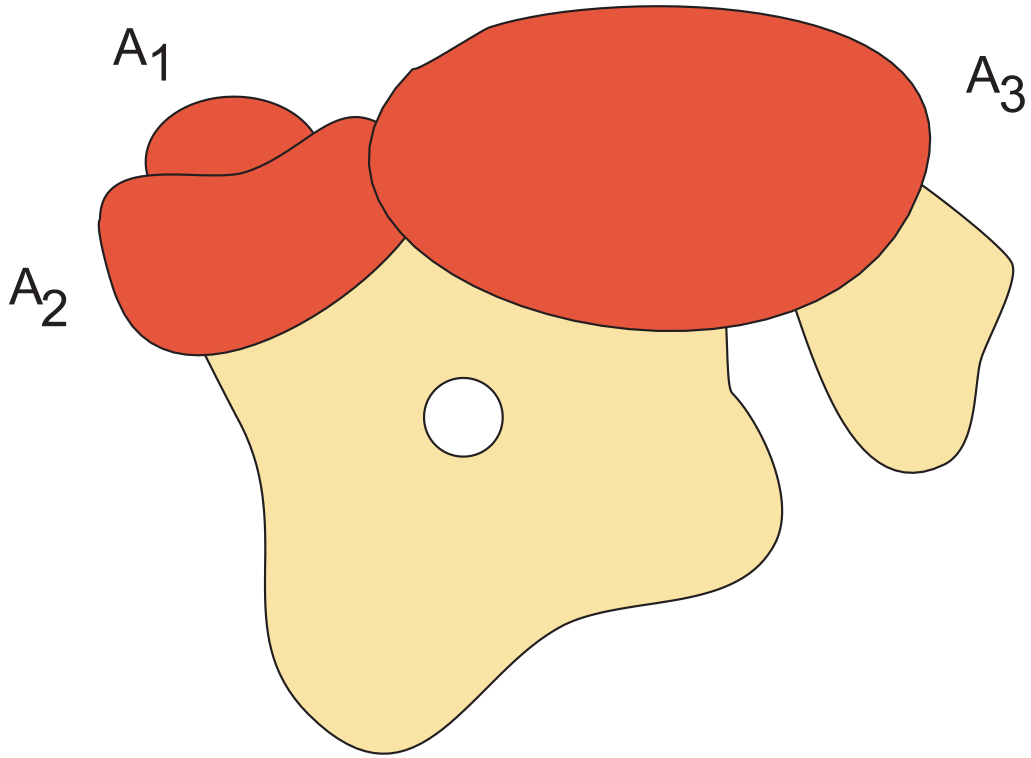
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

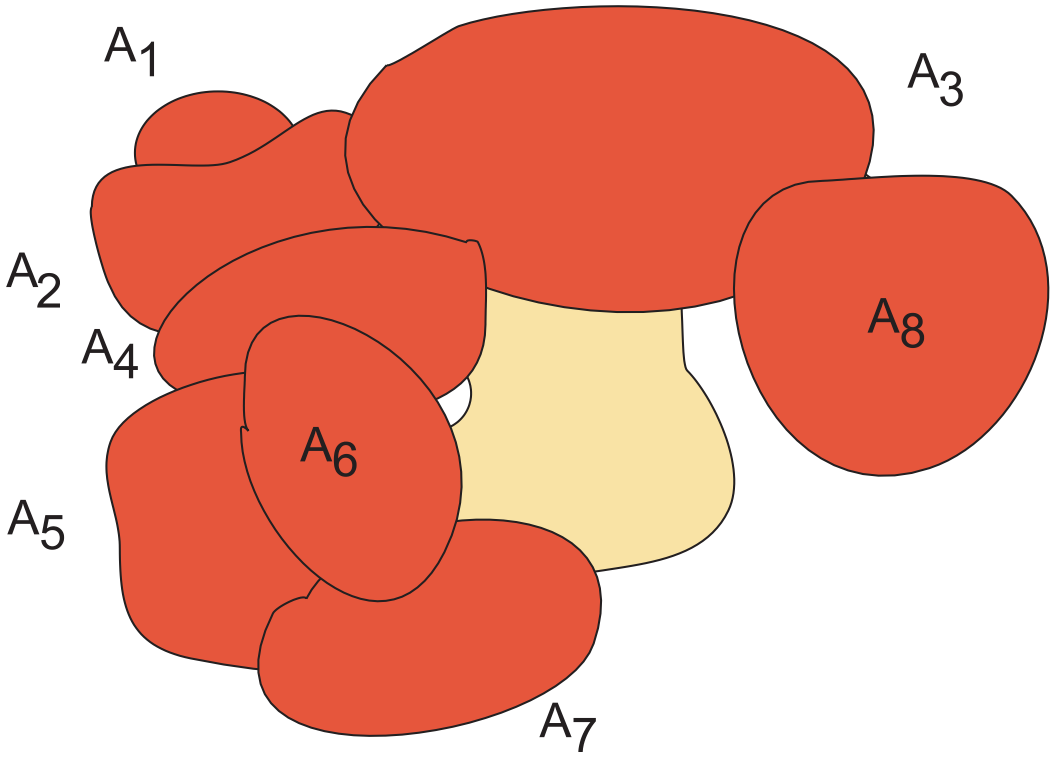
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

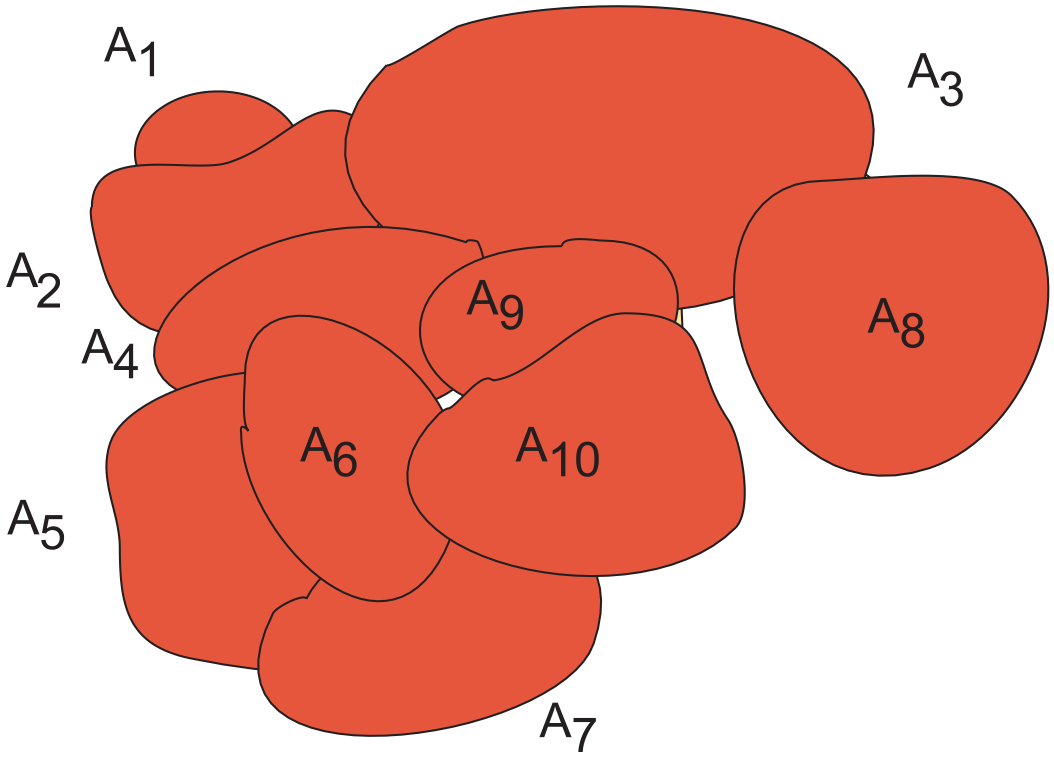
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

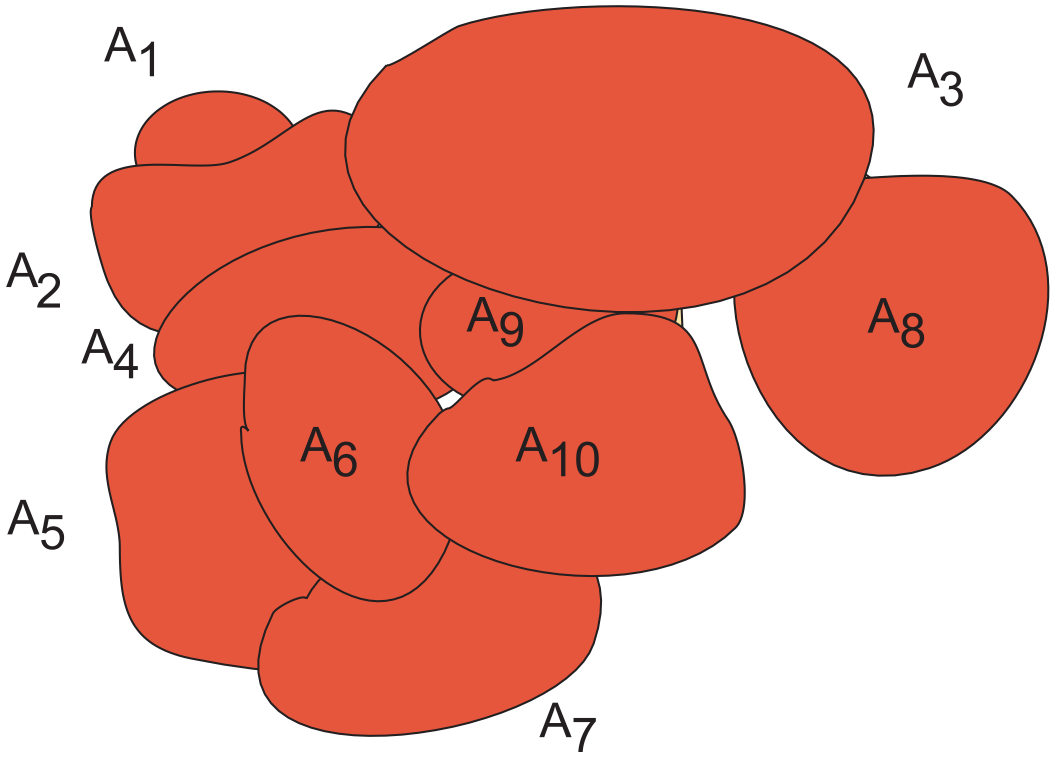
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

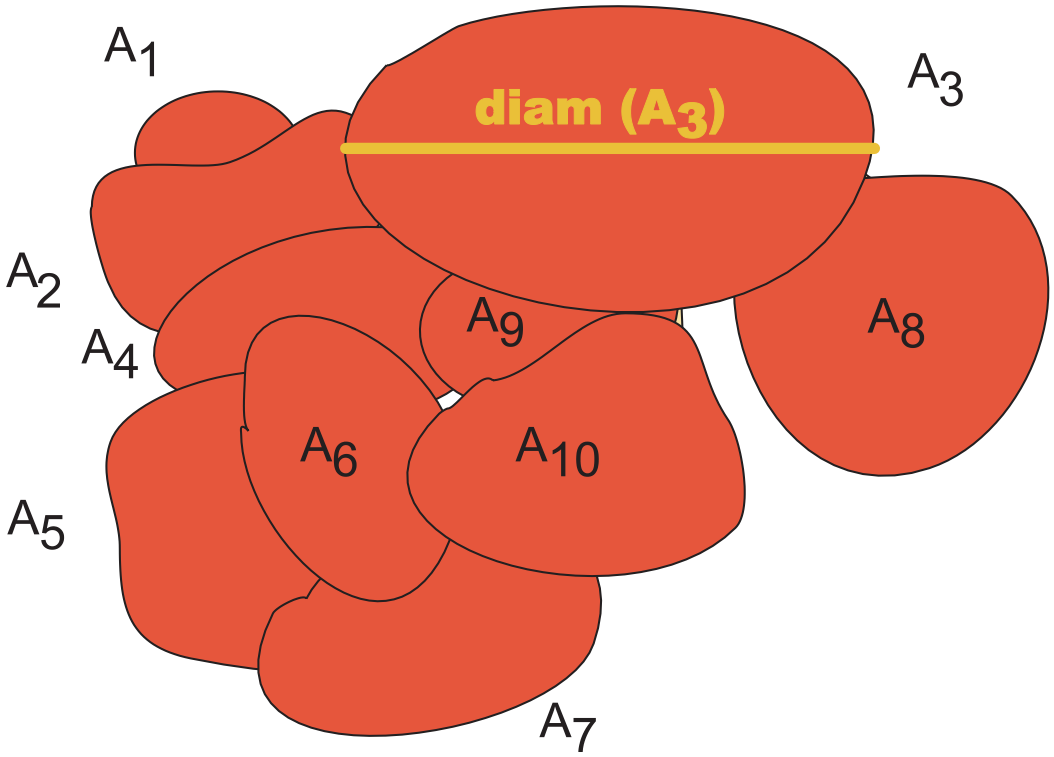
- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

Fraktální množiny

- δ-pokrytí množiny $F \subset \mathbf{R}^n$



množina v $\mathbf{R}^2$

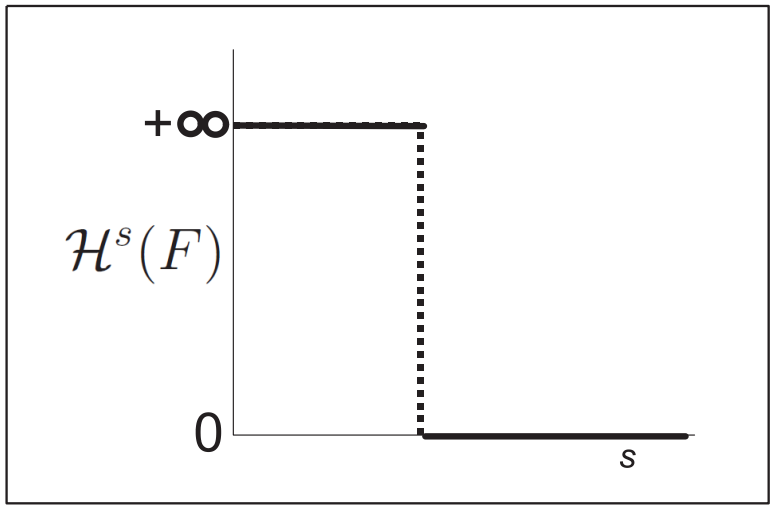
Hausdorffova míra a dimenze

- Hausdorffova s-rozměrná míra množiny $F \subset \mathbf{R}^n$

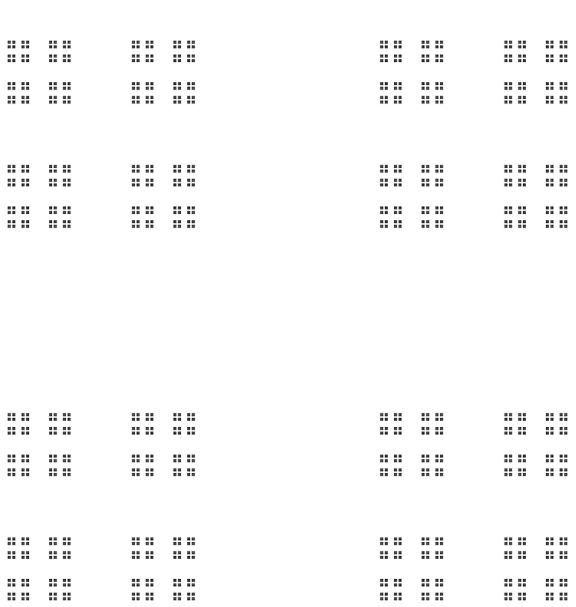
$$\mathcal{H}^s(F) = \lim_{\delta \to 0} \inf \left\{ \sum_{i=1}^{+\infty} (\mathrm{diam}(A_i))^s : (A_i) \text{ je } \delta\text{-pokrytí množiny} F \right\}$$

- Hausdorffova dimenze

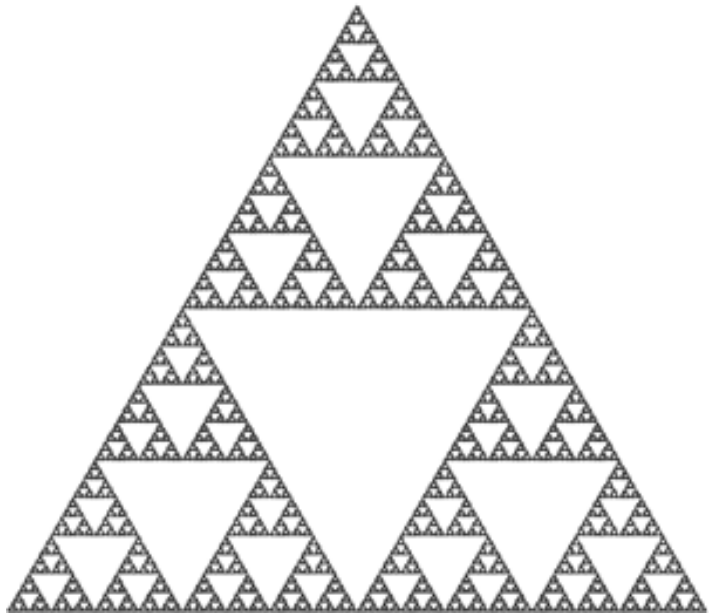
$$\dim_H(F) = \sup \{ s : \mathcal{H}^s(F) = +\infty \}$$



Fraktální množiny



Cantorův prach



Sierpinského trojúhelník

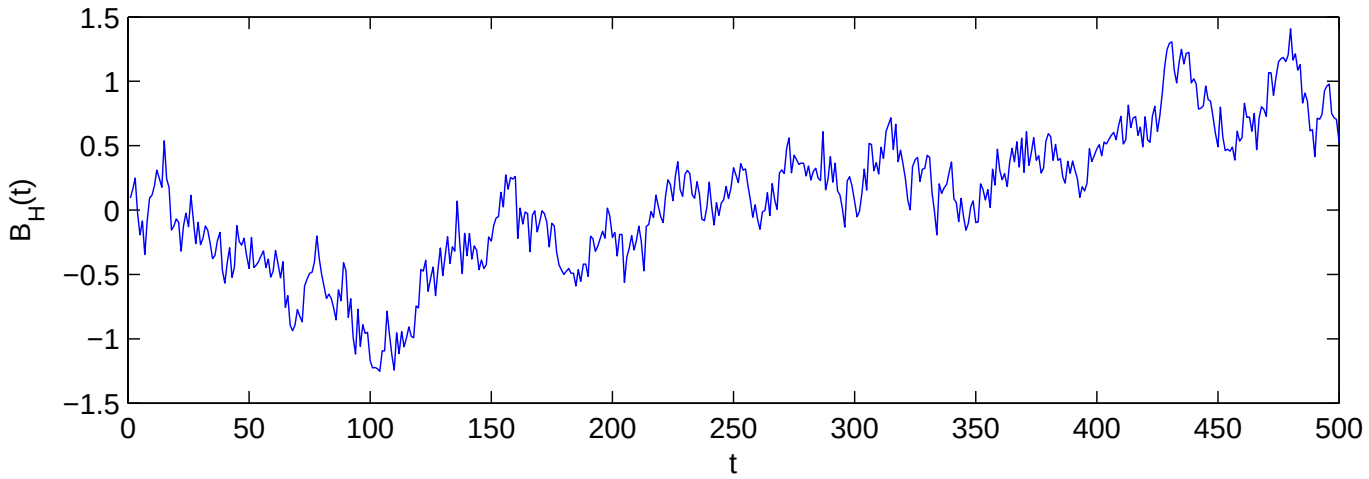
$$\dim_H(C) = \log(4)/\log(3)$$

$$\dim_H(S) = \log(3)/\log(2)$$

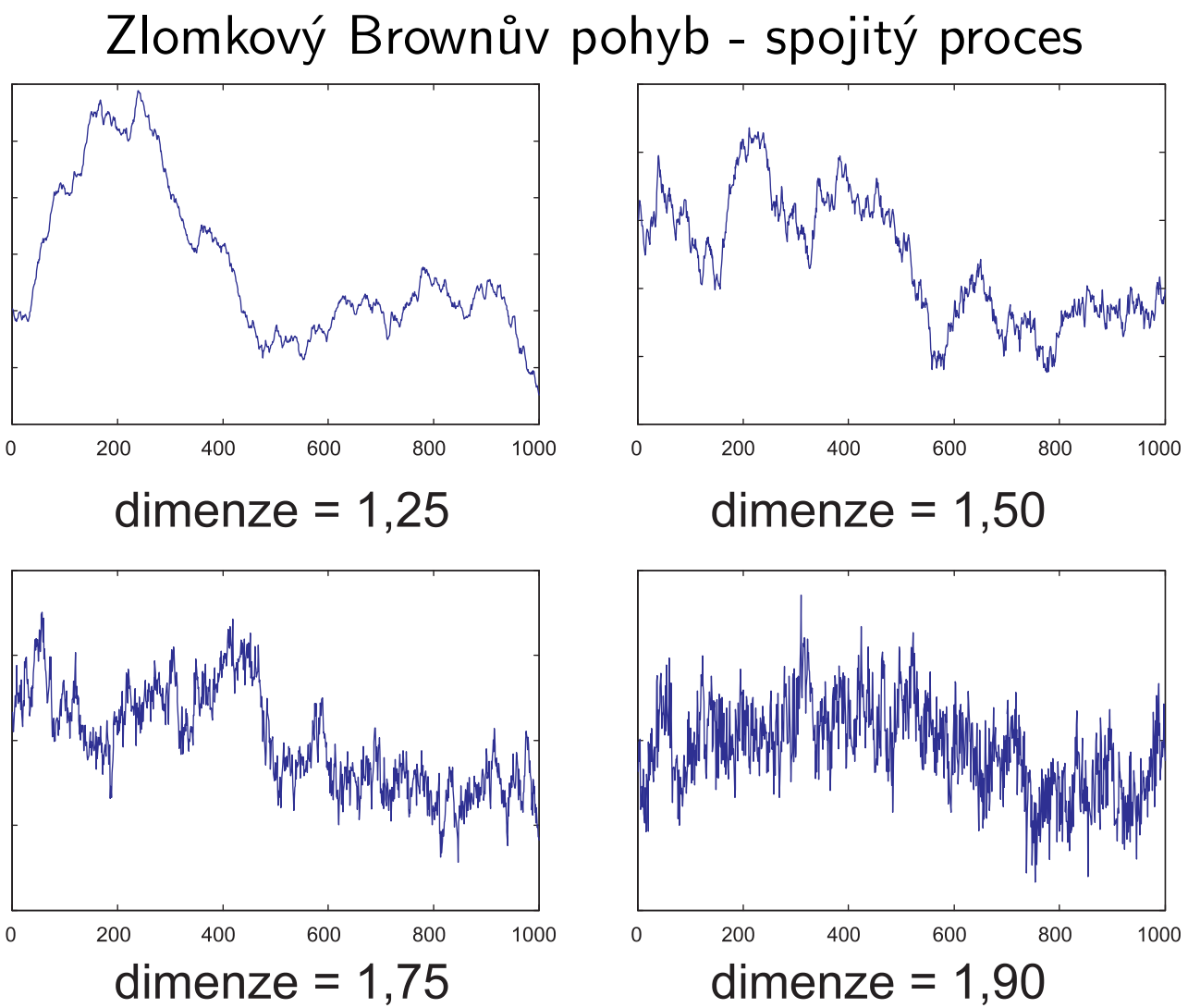
Zlomkový Brownův pohyb

- zlomkový Brownův pohyb $B_H(t)$ - spojitý gaussovský proces
- definovaný na intervalu $\langle 0; +\infty)$
- je závislý na parametru $H \in (0; 1)$
- $\dim_H(B_H(t)) = 2 - H$
- autokovarianční funkce

$$\operatorname{cov}(B_H(t), B_H(s)) = \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} + |t-s|^{2H}\right)$$



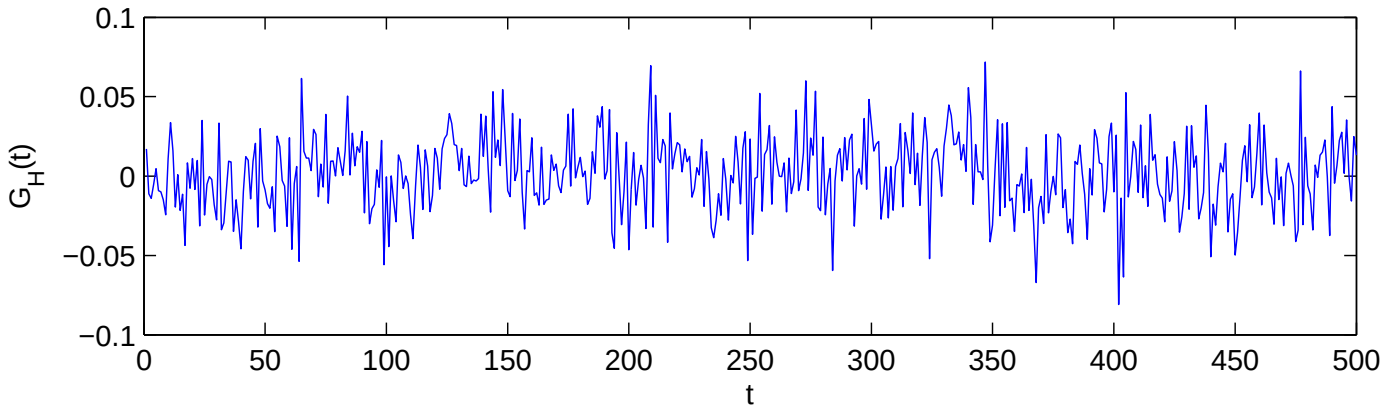
Zlomkový Brownův pohyb



Zlomkový Gaussův šum

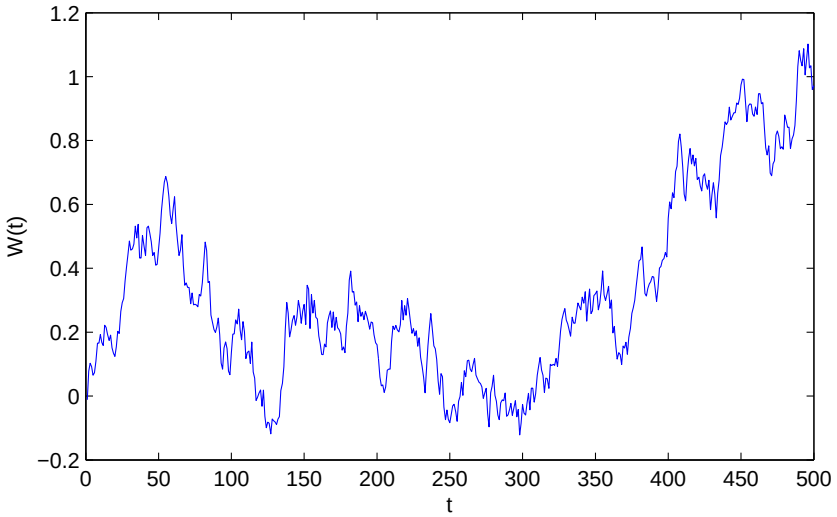
- zlomkový Gaussův šum $G_H(t)$ - spojitý gaussovský proces
- definovaný na intervalu $\langle 0; +\infty)$
- je závislý na parametru $H \in (0; 1)$
- $\dim_H(G_H(t)) = 2$
- autokovarianční funkce

$$\operatorname{cov}(G_H(t), G_H(t+k)) = \frac{\sigma^2}{2}\left(|k+1|^{2H} - 2\cdot|k|^{2H} + |k-1|^{2H}\right)$$

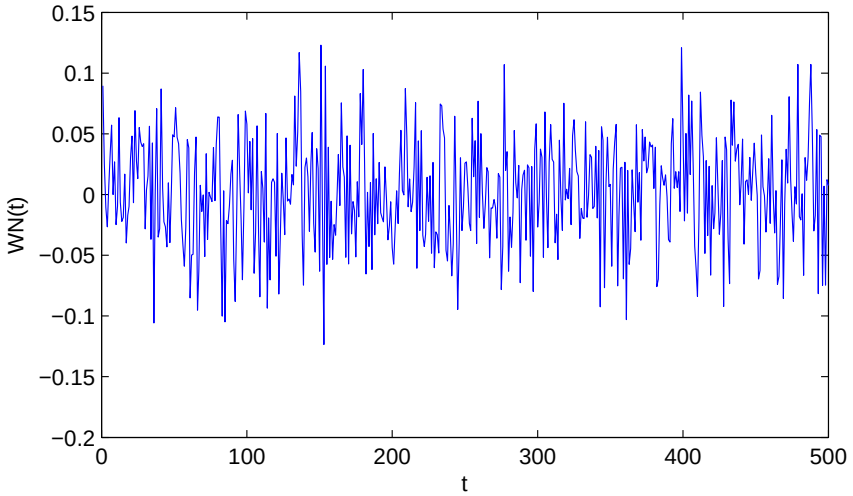


Speciální případy pro H=0,5

- zlomkový Brownův pohyb pro $H = 0,5 \rightarrow$ Wienerův proces

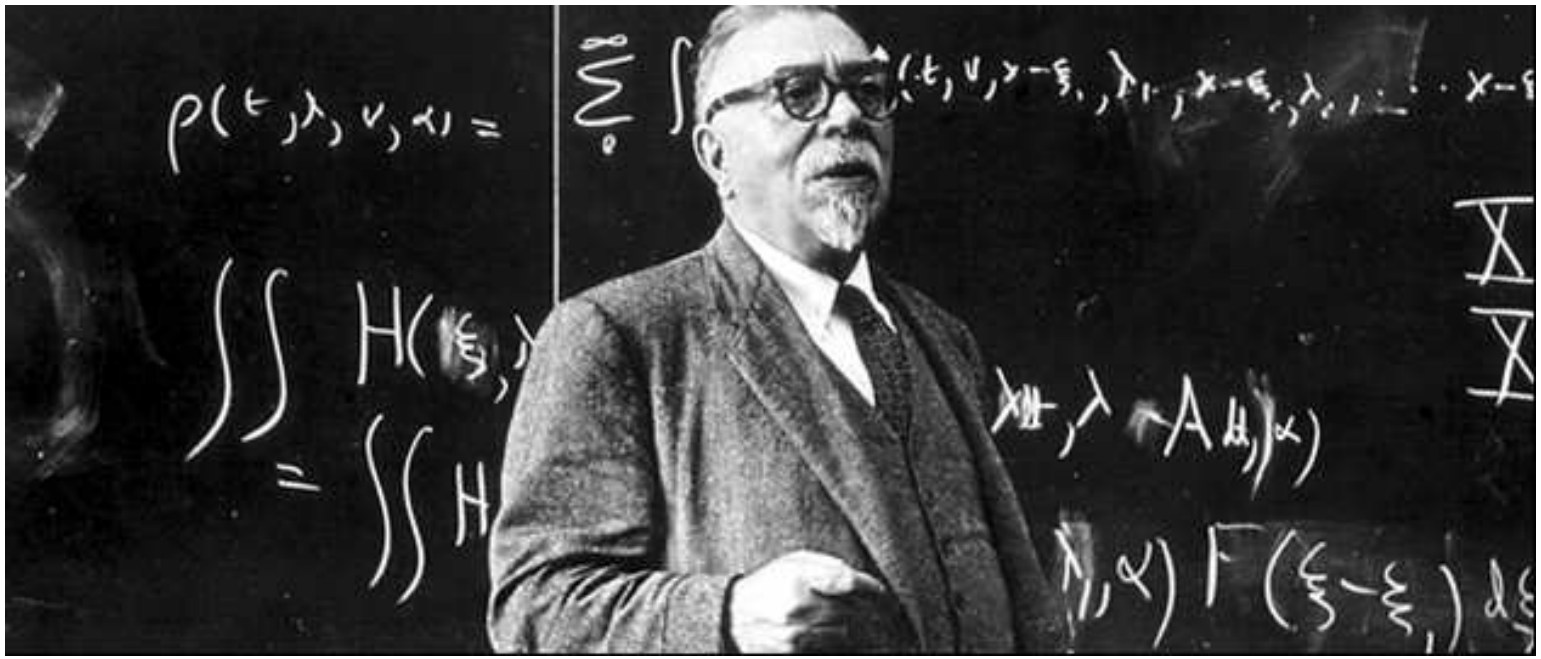


- zlomkový Gaussův šum pro $H = 0,5 \rightarrow$ bílý šum



Speciální případy pro H=0,5

- Wienerův proces
- jeho diference - bílý šum (short memory)



Norbert Wiener

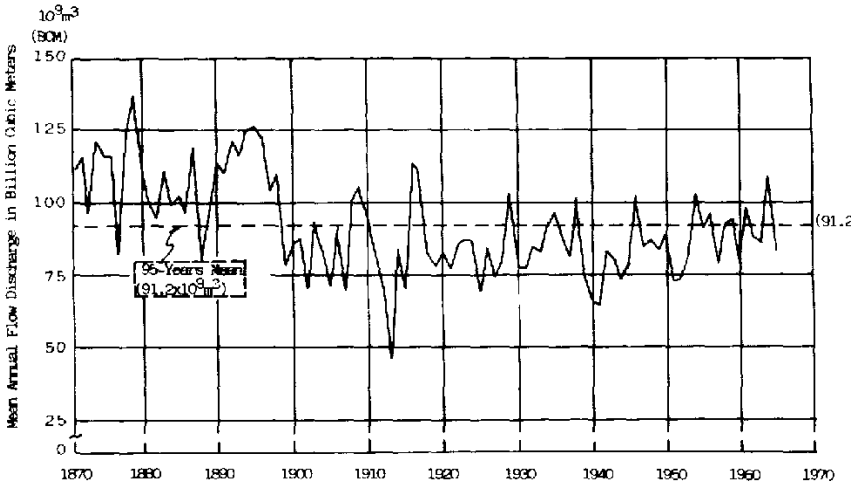
Vlastnosti zlomkového Brownova pohybu

parametr H	**fraktál**	**závislost přírůstků**
$H = 1$	ne	ano
$H = 0{,}5$	ano	ne
$H = 0$	ano	ano

Hurstův exponent

$$D = 2 - H$$

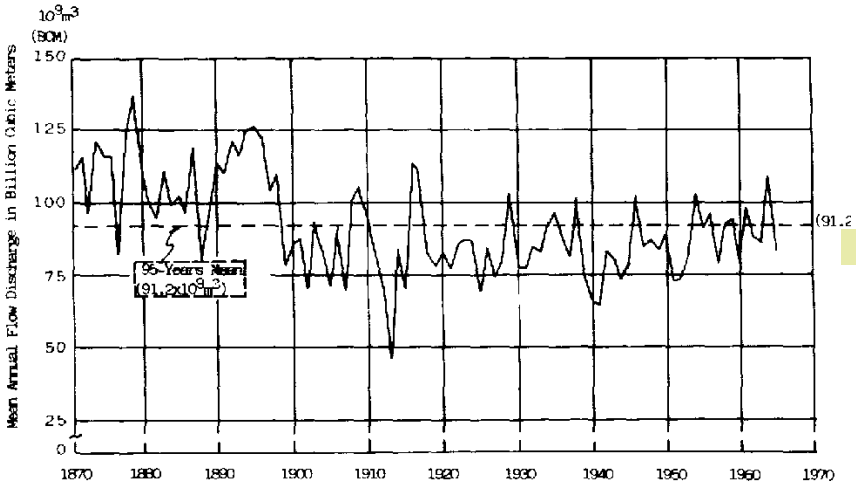
D ... Hausdorffova dimenze
H ... Hurstův exponent

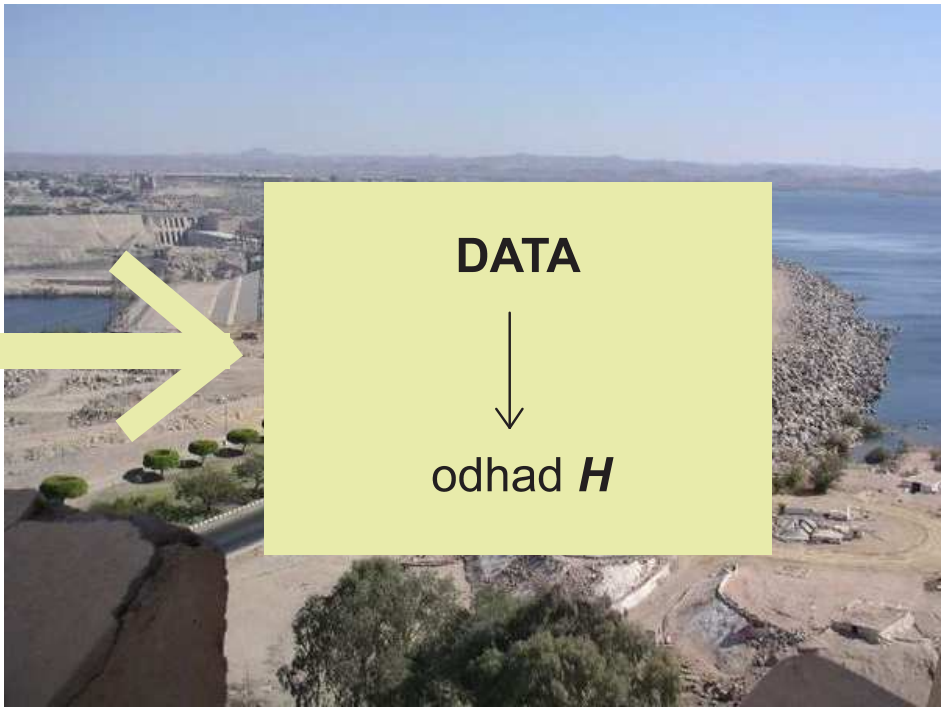


Hurstův exponent

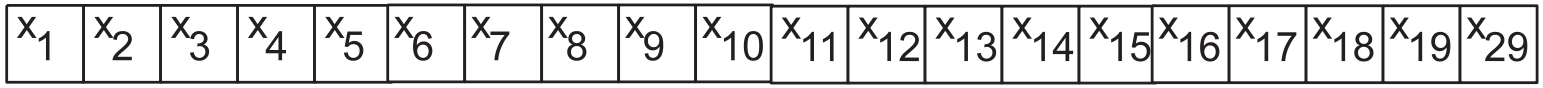
$$D = 2 - H$$

D ... Hausdorffova dimenze
H ... Hurstův exponent





Metoda R/S



- R ... rozsah (range)
- S ... směrodatná odchylka (std)
- m ... velikost segmentu

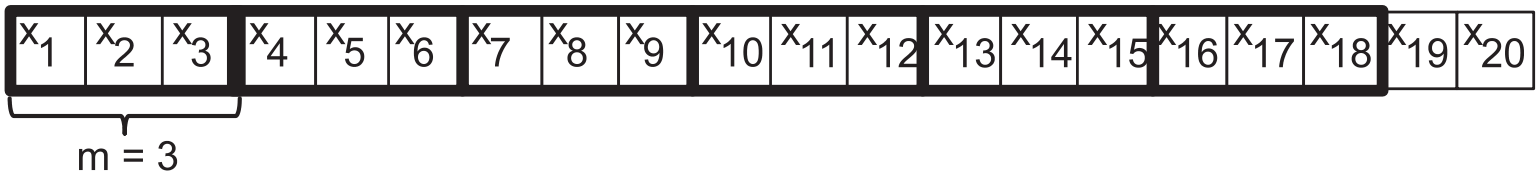
1. výpočet střední hodnoty R/S v každém segmentu

$$(R/S)_m = \frac{1}{r} \sum_{j=1}^{r} \frac{R_j}{S_j},$$

2. použití regresního modelu

$$\mathrm{E}\left[(R/S)\right] \propto m^H.$$

Metoda R/S



- R ... rozsah (range)
- S ... směrodatná odchylka (std)
- m ... velikost segmentu

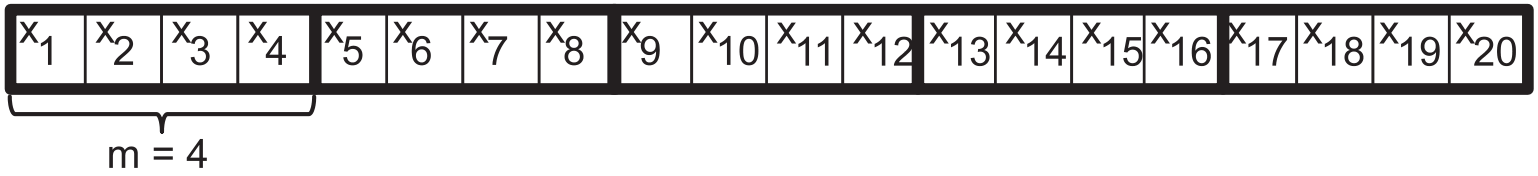
1. výpočet střední hodnoty R/S v každém segmentu

$$(R/S)_m = \frac{1}{r} \sum_{j=1}^{r} \frac{R_j}{S_j},$$

2. použití regresního modelu

$$\mathrm{E}\left[(R/S)\right] \propto m^H.$$

Metoda R/S



- R ... rozsah (range)
- S ... směrodatná odchylka (std)
- m ... velikost segmentu

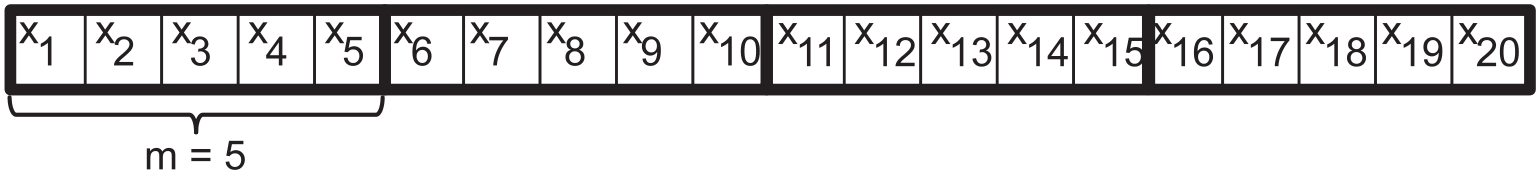
1. výpočet střední hodnoty R/S v každém segmentu

$$(R/S)_m = \frac{1}{r} \sum_{j=1}^{r} \frac{R_j}{S_j},$$

2. použití regresního modelu

$$\mathrm{E}\left[(R/S)\right] \propto m^H.$$

Metoda R/S



- R ... rozsah (range)
- S ... směrodatná odchylka (std)
- m ... velikost segmentu

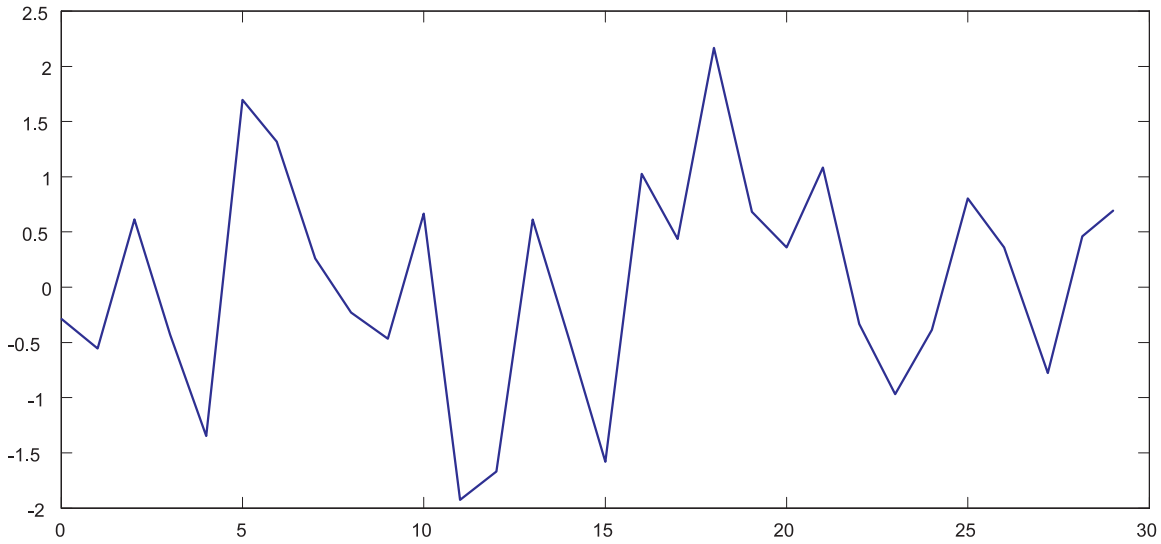
1. výpočet střední hodnoty R/S v každém segmentu

$$(R/S)_m = \frac{1}{r} \sum_{j=1}^{r} \frac{R_j}{S_j},$$

2. použití regresního modelu

$$\mathrm{E}\left[(R/S)\right] \propto m^H.$$

Metoda průchodů nulou (zero-crossing)



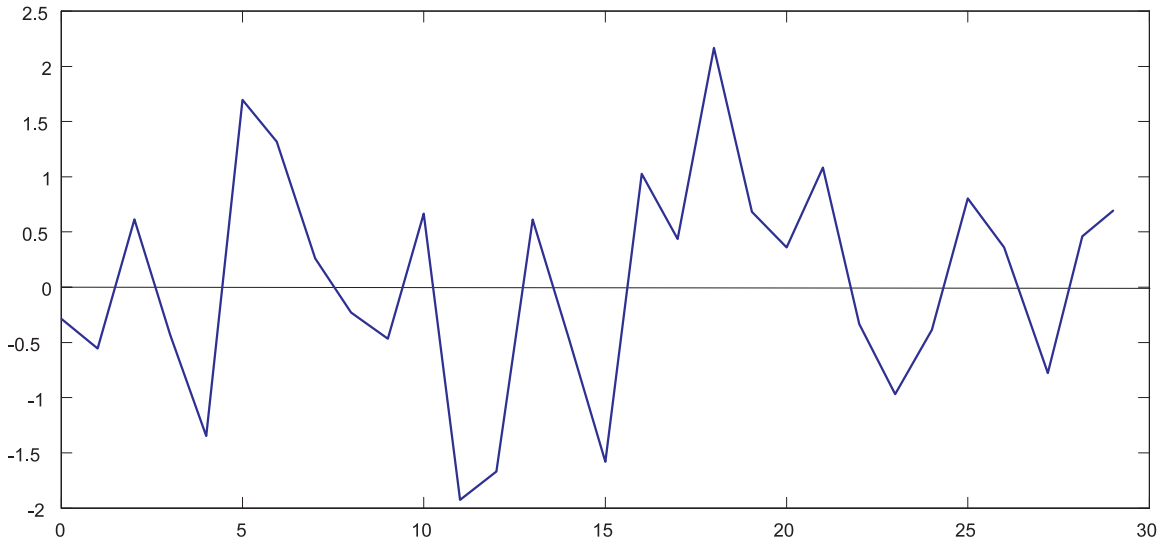
1. výpočet pravděpodobnosti průchodu nulou

$$p^* = \frac{\text{počet průsečíků}}{\text{počet dat}}$$

2. výpočet bodového odhadu H

$$H = 1 + \log_2 \cos \frac{\pi p^*}{2}$$

Metoda průchodů nulou (zero-crossing)



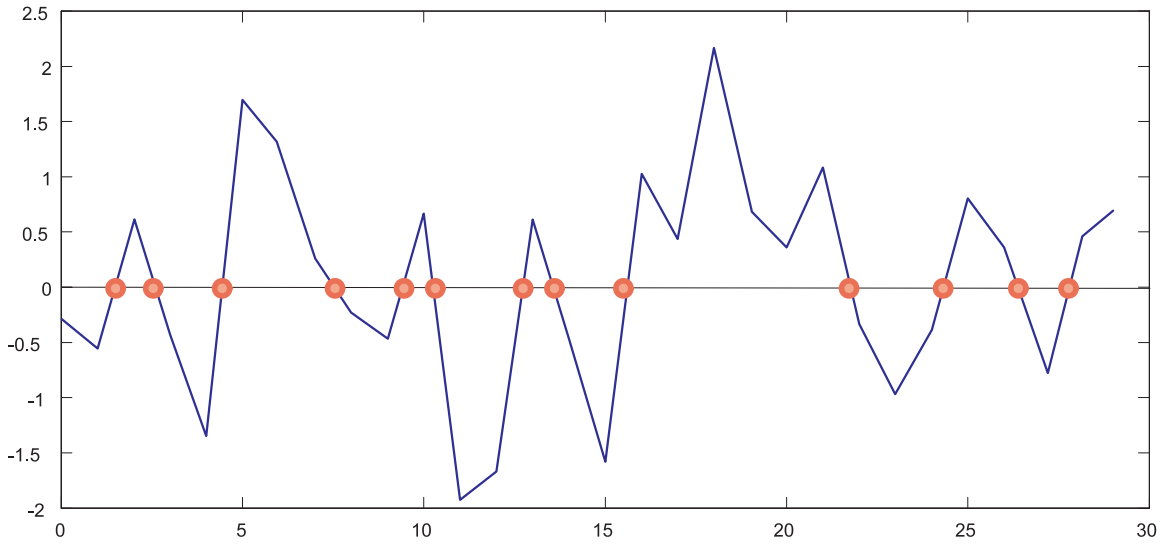
1. výpočet pravděpodobnosti průchodu nulou

$$p^* = \frac{\text{počet průsečíků}}{\text{počet dat}}$$

2. výpočet bodového odhadu H

$$H = 1 + \log_2 \cos \frac{\pi p^*}{2}$$

Metoda průchodů nulou (zero-crossing)



1. výpočet pravděpodobnosti průchodu nulou

$$p^* = \frac{\text{počet průsečíků}}{\text{počet dat}}$$

2. výpočet bodového odhadu H

$$H = 1 + \log_2 \cos \frac{\pi p^*}{2}$$

Vylepšená metoda průchodů nulou (zero-crossing revisited)

- bodový odhad Hurstova exponentu H za předpokladu známé pravděpodobnosti p^*

$$H = 1 + \log_2 \cos \frac{\pi p^*}{2}$$

- počet průchodů nulou Z ve vzorku délky N se řídí binomickým rozdělením

$$f(Z|p^*) = \binom{N}{Z} (p^*)^Z \cdot (1 - p^*)^{N-Z}$$

- po použití Bayesovské inverze má pravděpodobnost průchodů nulou Beta-rozdělení

$$f_{\mathrm{POST}}(p^*|Z) = \frac{(p^*)^{Z-\alpha}(1 - p^*)^{N-Z-\alpha}}{\mathrm{B}(Z + 1 - \alpha, N - Z + 1 - \alpha)}$$

Vylepšená metoda průchodů nulou (zero-crossing revisited)

- po rozdělení časové řady na L částí (v každé části je Z_k průchodů nulou) můžeme vypočítat agregovanou hustotu pravděpodobnosti

$$f_L(p) = \frac{1}{L} \sum_{k=1}^{L} \frac{p^{Z_k - \alpha}(1-p)^{N-\alpha-Z_k}}{\mathrm{B}(Z_k + 1 - \alpha, N - Z_k + 1 - \alpha)}$$

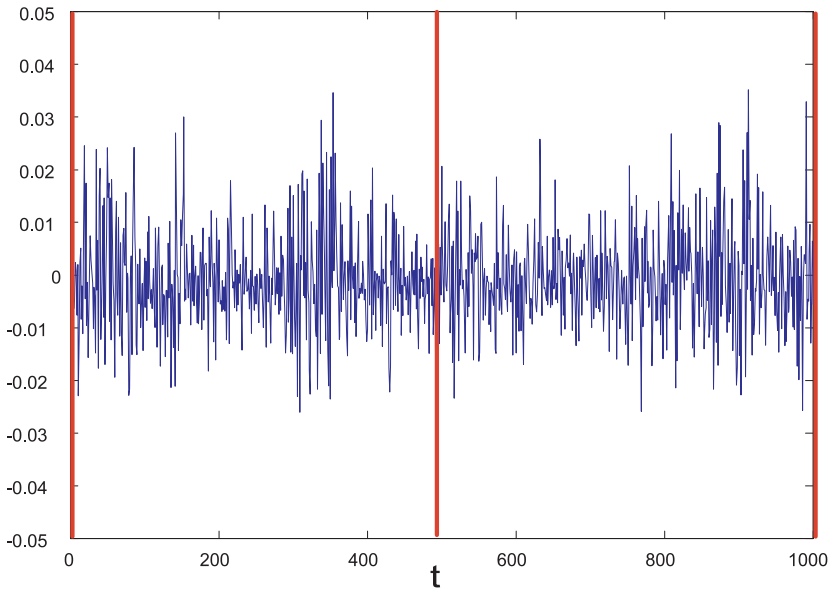
- ideální rozdělení (**optimální segmentace**) je na L^* dílů

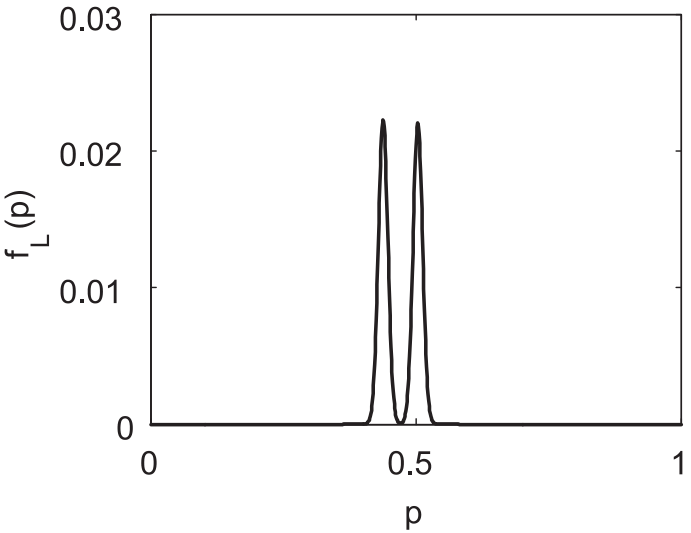
$$L^* = \begin{cases} 1, \text{ když } f_L(p) \text{ je jednovrcholová pro všechna } L \\ \min\{L > 1 : f_L(p) \text{ je jednovrcholová } \}, \text{jinak} \end{cases}$$

- $\rightarrow$ dělíme signál na jemnější segmenty do té doby, dokud není agregovaná hustota pravděpodobnosti $f_L(p)$ jednovrcholová

Vizualizace optimální segmentace

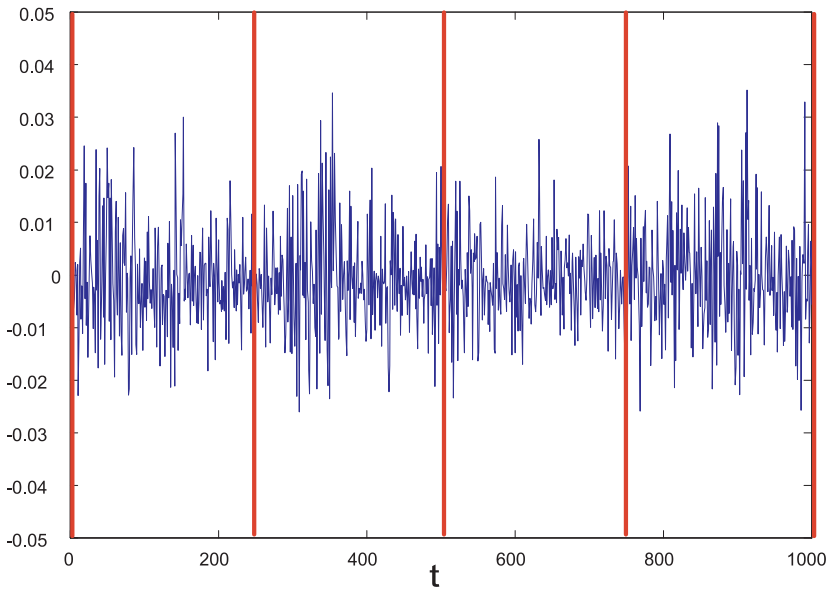
L = 2

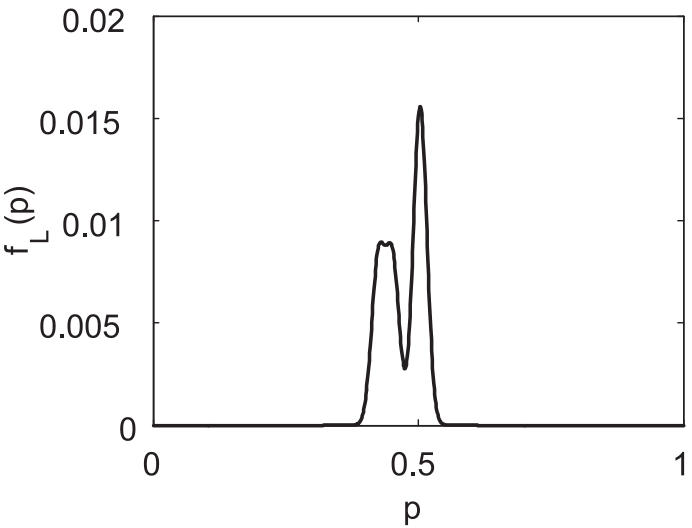




Vizualizace optimální segmentace

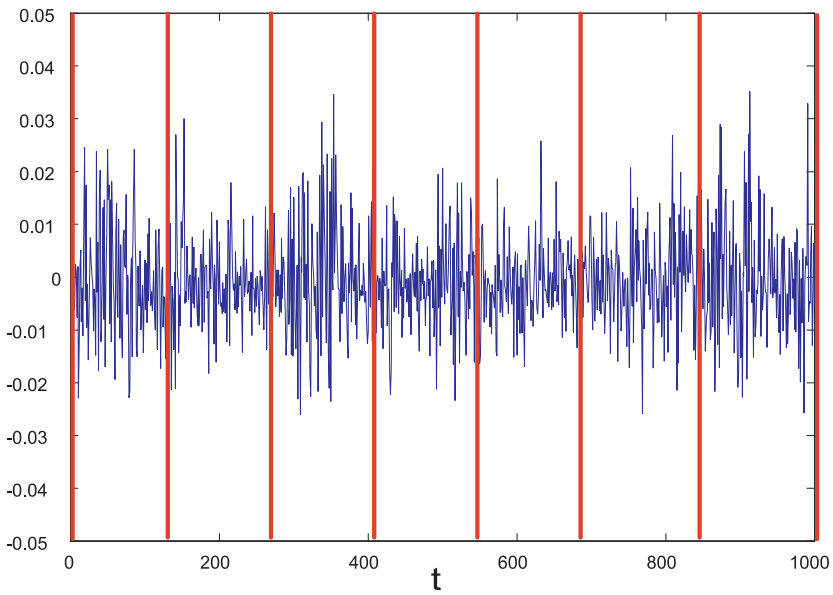
L = 4

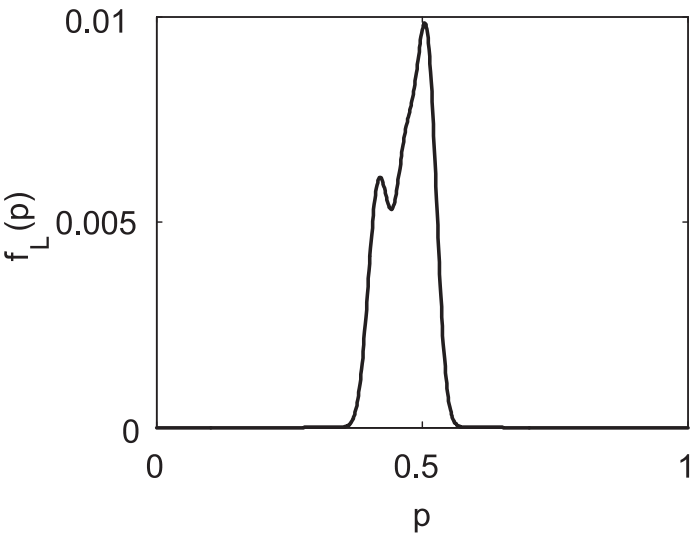




Vizualizace optimální segmentace

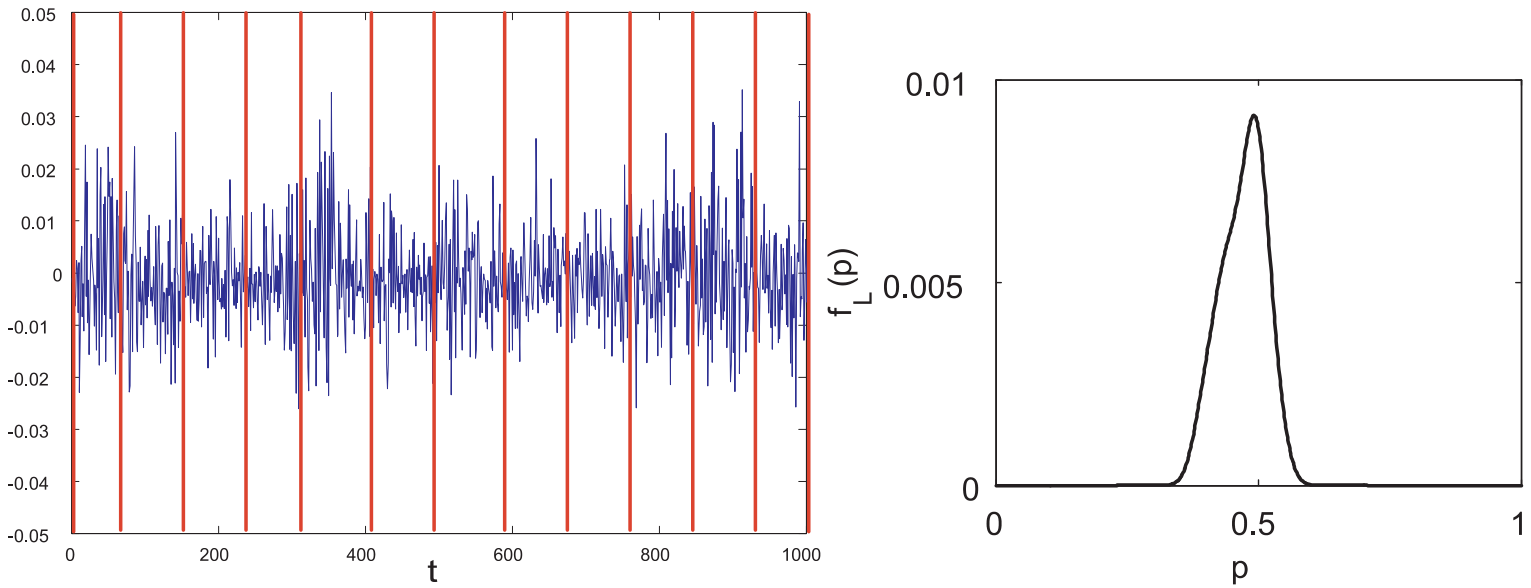
L = 7





Vizualizace optimální segmentace

L = 12



Vylepšená metoda průchodů nulou (zero-crossing revisited)

- výpočet Hurstova exponentu

$$\mathrm{E}(H) = 1 + \int_0^1 f_{L*}(p) \log_2 \cos \frac{p\pi}{2} \mathrm{d}p$$

tradiční metoda	**nová metoda**
bodový odhad H	střední hodnota a konfidenční intervaly
pravděpodobnost průchodu nulou je nahrazena relativní četností	pravděpodobnost průchodů nulou má Beta-rozdělení
pojímá signál jako celek	využívá bayesovského přístupu a segmentace signálu
absence konfidenčního intervalu	realistický konfidenční interval


Aplikace
fraktální
teorie

Aplikace fraktální teorie

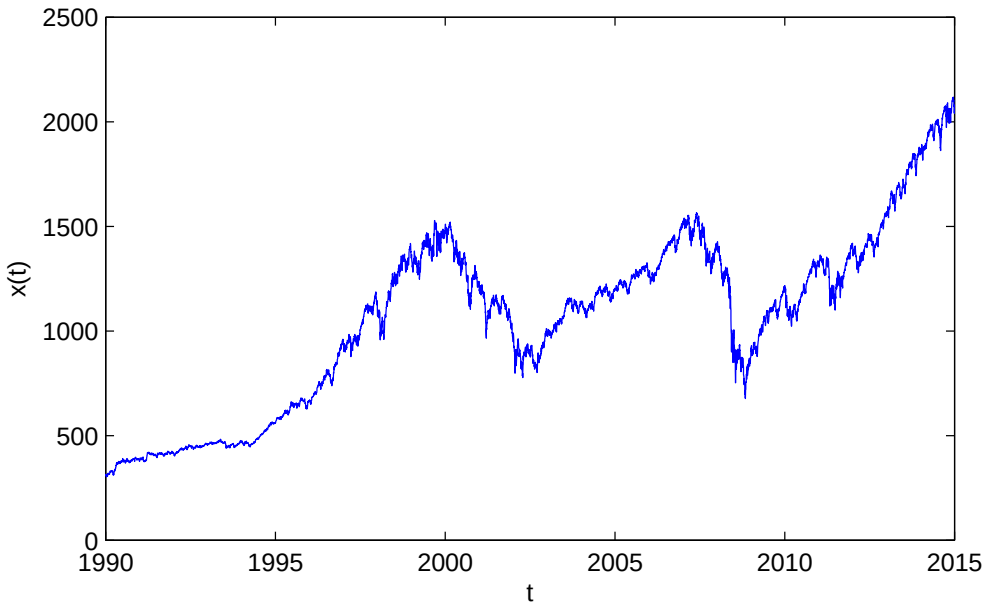
Použití Hurstova exponentu:

1. test generátoru náhodných čísel - nejjednodušší použití
2. ekonomie - závislost vývoje časových řad (indexy akciových trhů)
3. biomedicína - detekce Alzheimerovy choroby ze signálu EEG
4. IT - vytížení počítačových sítí z hlediska množství dat
5. hydrologie - analýza průtoků vodních toků
6. hudba - klasifikace hudebních žánrů

Ekonomie - měření závislosti a predikování časových řad

Ekonomie - měření závislosti a predikování časových řad

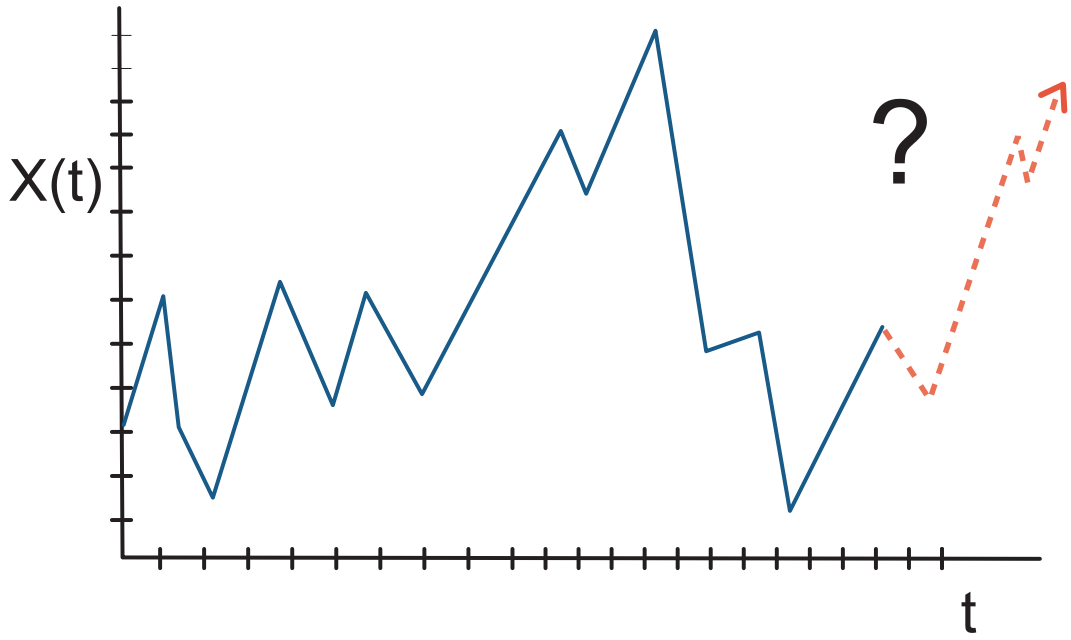
- použití nové metodiky pro odhad Hurstova exponentu
- data z období 1991 - 2015
- zkoumané indexy akciových trhů:
  - Evropa - CAC40, DAX, FTSE, SMI
  - Asie - HSI, NIKKEI
  - Severní Amerika - SP500, NASDAQ, TSX
- zkoumány jejich logaritmické diference (za předpokladu fGn)



SP500

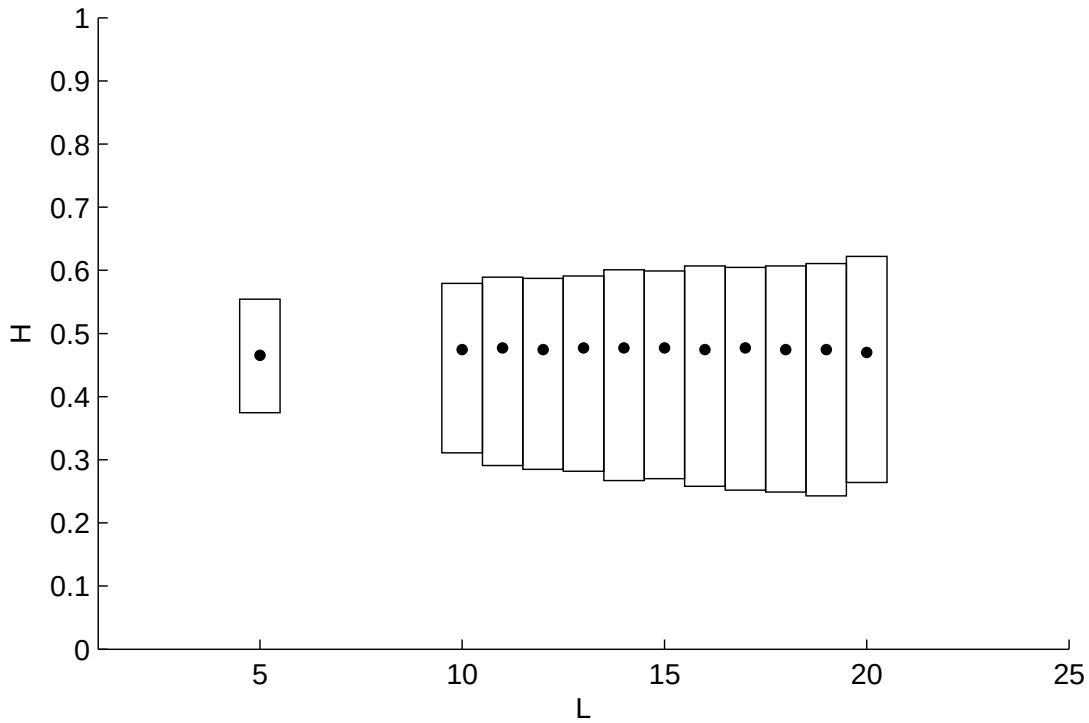
Ekonomie - měření závislosti a predikování časových řad

- míra závislosti časové řady je dána Hurstovým exponentem
- časová řada je predikovatelná, když je Hurstův exponent větší než 0,5
- použití nástroje pro zlomkové modelování časových řad (ARFIMA)



Ekonomie - měření závislosti a predikování časových řad

- ekonomické časové řady jsou obtížně predikovatelné
- většinou je Hurstův exponent blízký 0,5
- u více závislých časových řad je zapotřebí malý počet dělení



SP500

Ekonomie - měření závislosti a predikování časových řad

index	L^*	EH	95% CI
CAC40	5	0.4652	(0.3746;0.5545)
DAX	1	0.4790	(0.4556;0.5023)
FTSE	3	0.4863	(0.4361;0.5482)
HSI	1	0.4974	(0.4746;0.5201)
NASDAQ	**14**	**0.5418**	**(0.4008;0.7136)**
NIKKEI	3	0.4532	(0.4034;0.5068)
SMI	1	0.5087	(0.4863;0.5310)
SP500	6	0.4460	(0.3225;0.5628)
TSX	**12**	**0.5607**	**(0.3957;0.6992)**

Optimální segmentace a konfidenční intervaly

ARFIMA modely

- B . . . operátor posunutí

$$B^i(X_t) = X_{t-i}$$

- ARMA(p,q) model

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) X_t = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) e_t$$

- ARIMA(p,d,q) model, kde p $\in \mathbf{N}$

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) (1 - B)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) e_t$$

ARFIMA modely

- ARFIMA(p,d,q) model, kde d $\in (0, 1)$

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right) (1 - B)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i B^i\right) e_t$$

- má stejný zápis jako ARIMA s rozdílem, že

$$(1 - B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = 1 - d \cdot B + \frac{d(d-1)}{2!} B^2 - \cdots$$

- parametr d vypočteme z Hurstova exponentu H jako

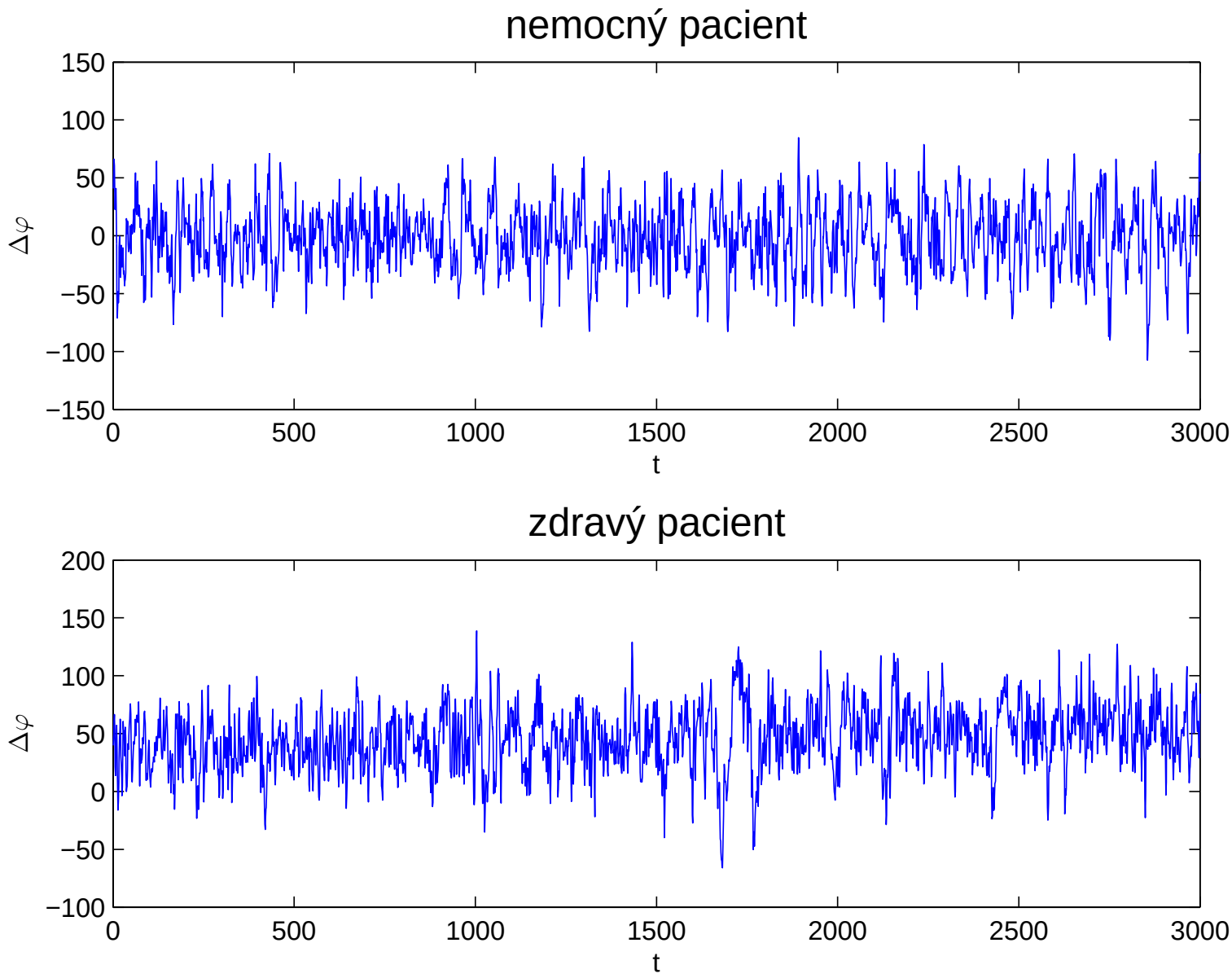
$$d = H - 1/2$$

Biomedicína - diagnostika Alzheimerovy choroby z EEG

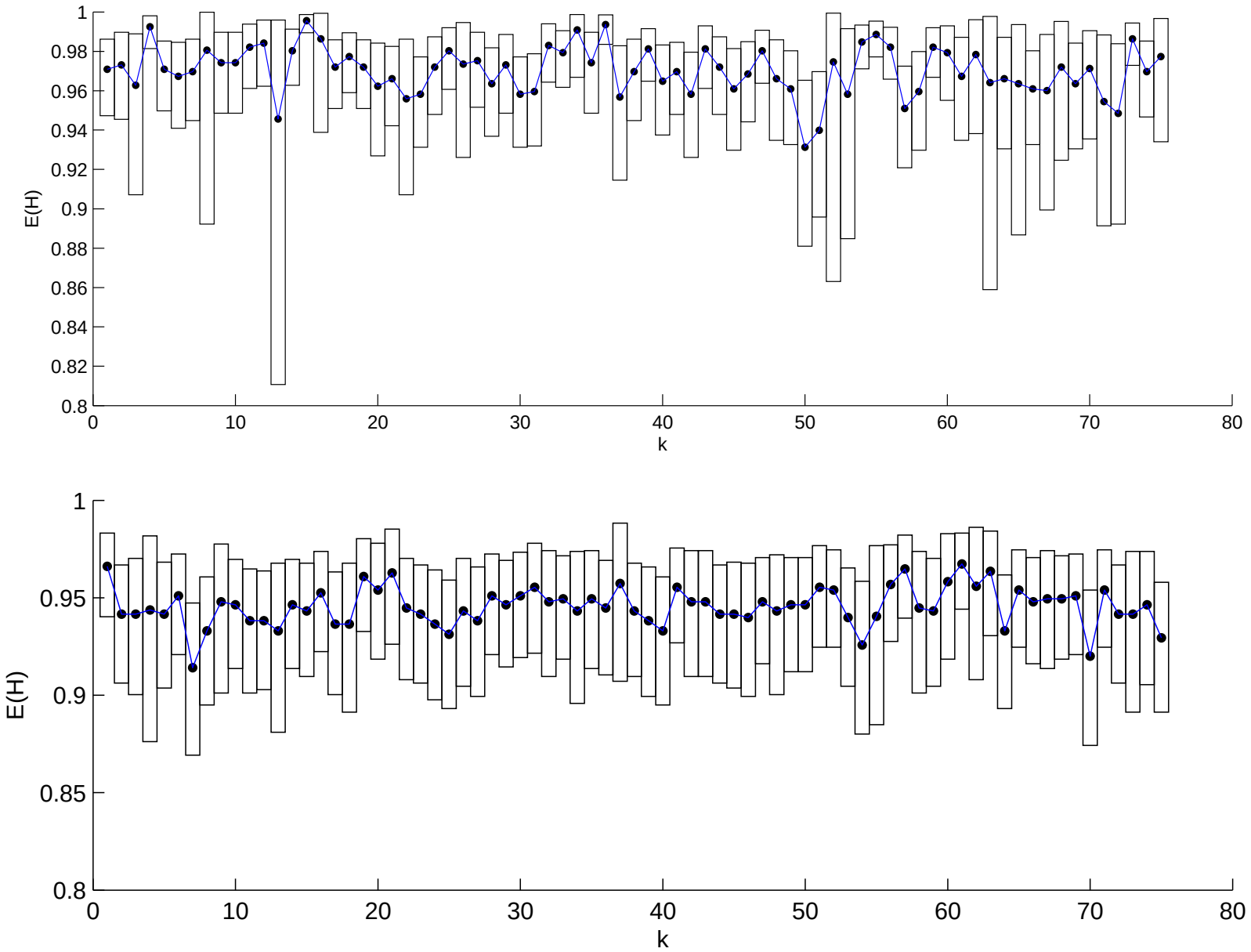
Biomedicína - diagnostika Alzheimerovy choroby z EEG

- analýza 19 kanálů EEG
- rozdílný Hurstův exponent zdravých subjektů a pacientů trpících Alzheimerovou demencí
  - zdravý člověk (CN) - nižší hodnoty H
  - nemocný člověk (AD) - vyšší hodnoty H
- nalezeny významné rozdíly mezi oběma skupinami
- nemoc se u každého projevuje různě
- metodika zatím neumožňuje přesně určit, jestli konkrétní pacient nemocí trpí

Biomedicína - diagnostika Alzheimerovy choroby z EEG

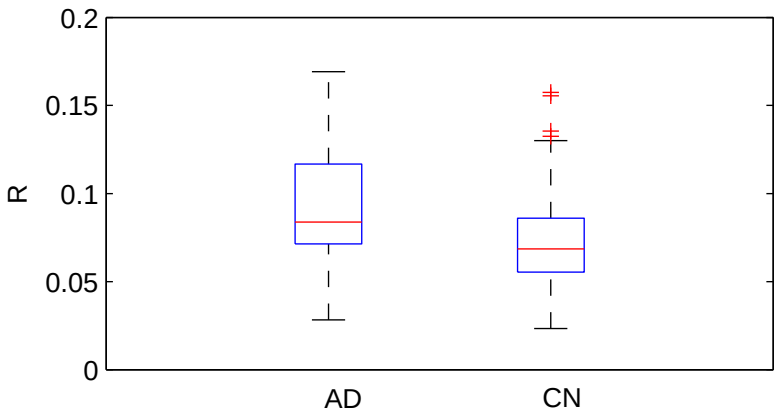


Biomedicína - diagnostika Alzheimerovy choroby z EEG

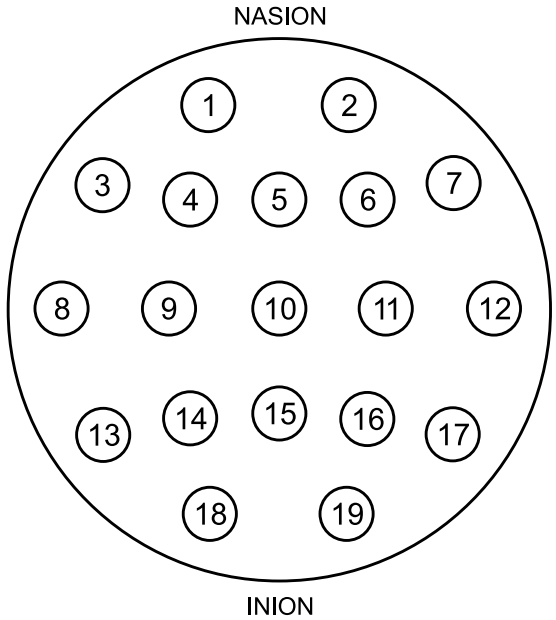


Biomedicína - diagnostika Alzheimerovy choroby z EEG

rozsah (*range*) Hurstova exponentu u nemocných (AD) a zdravých (CN) pacientů na druhém kanálu



největší rozdíly většinou nastávají na druhém, třetím a sedmém kanálu

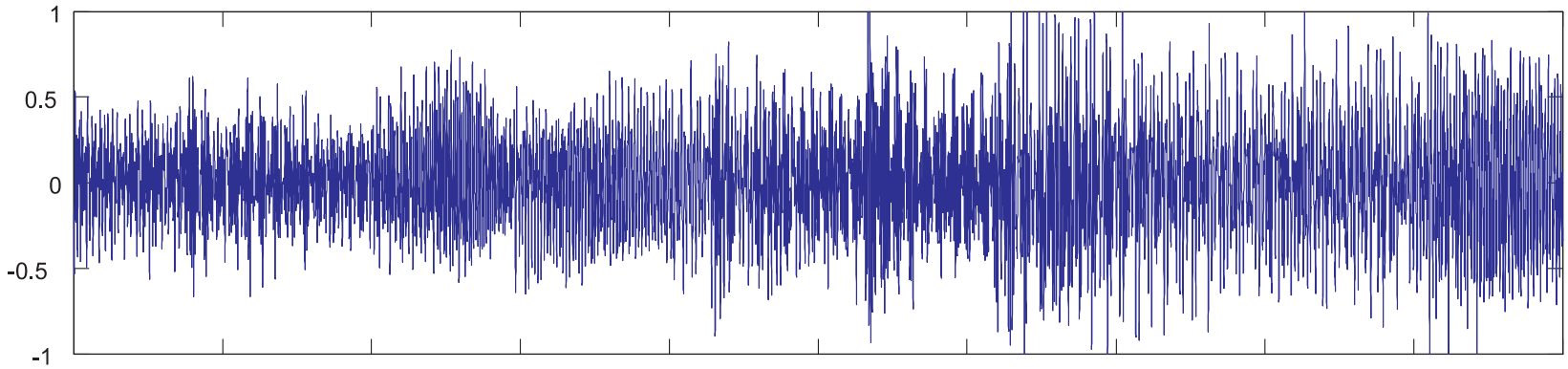


Aplikace na hudební záznam

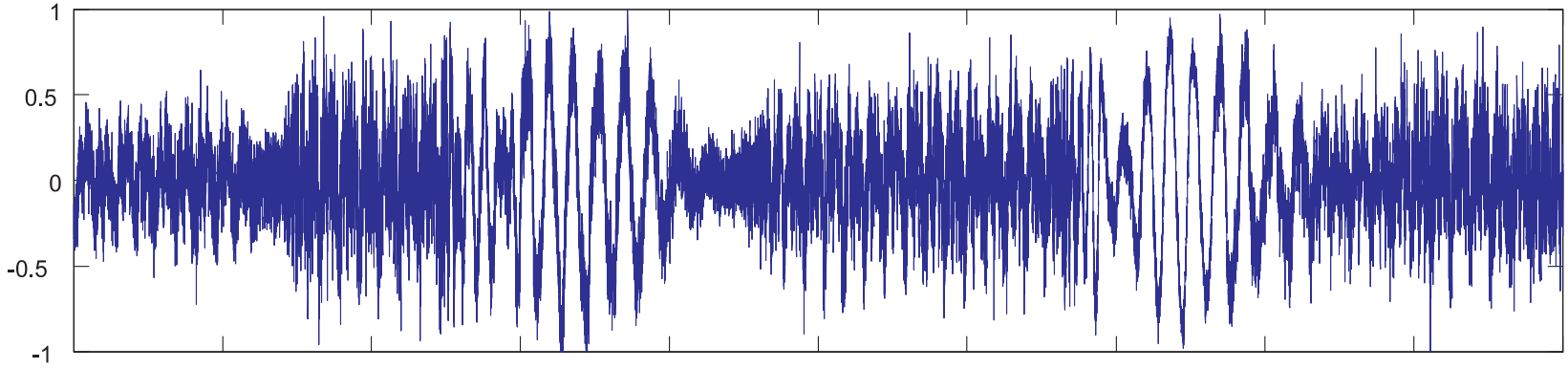
Aplikace na hudební záznam

- jaký hudební styl reprezentuje každý ze zvukových signálů?
- jaký je jejich Hurstův exponent?

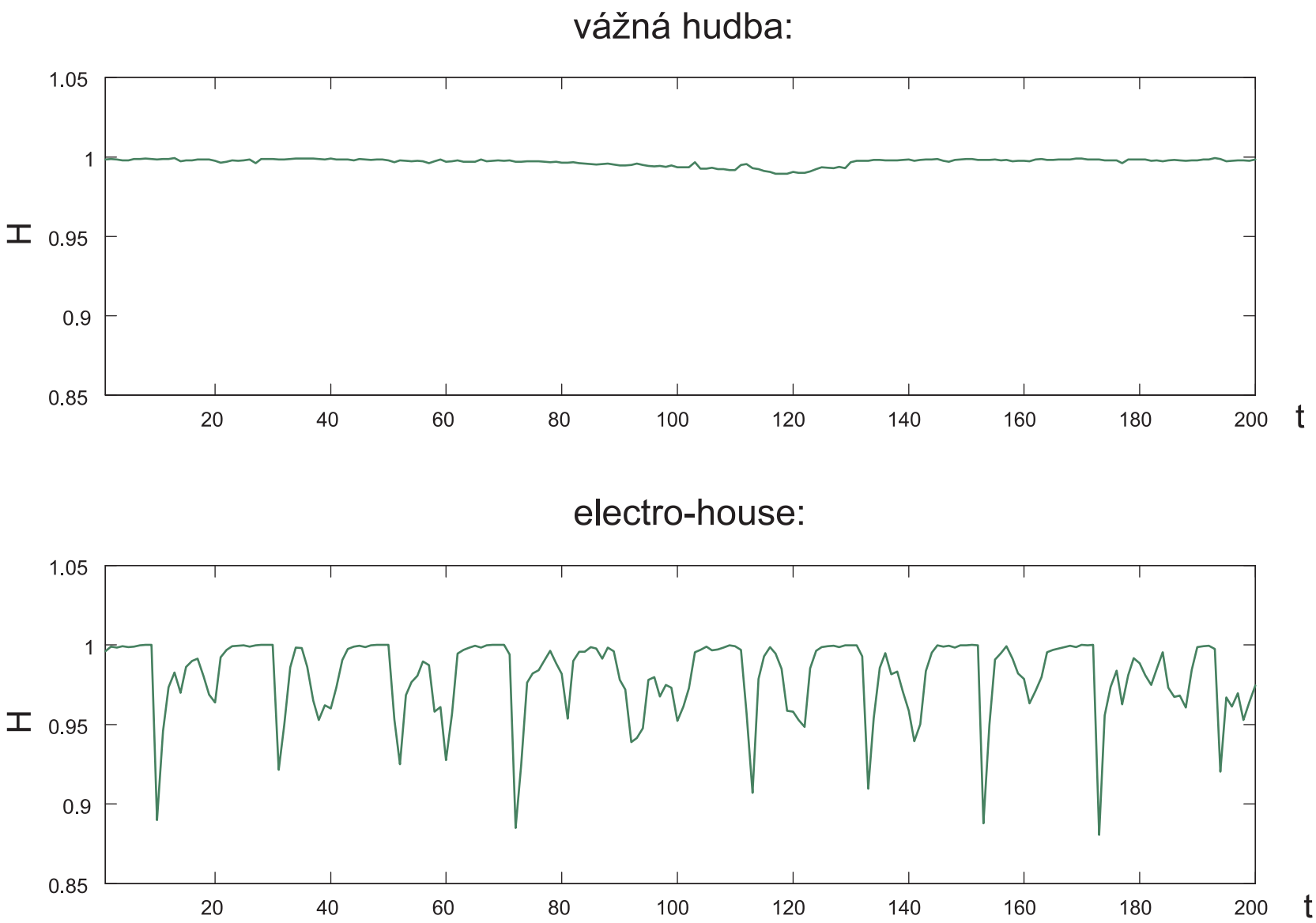
ukázka 1:



ukázka 2:

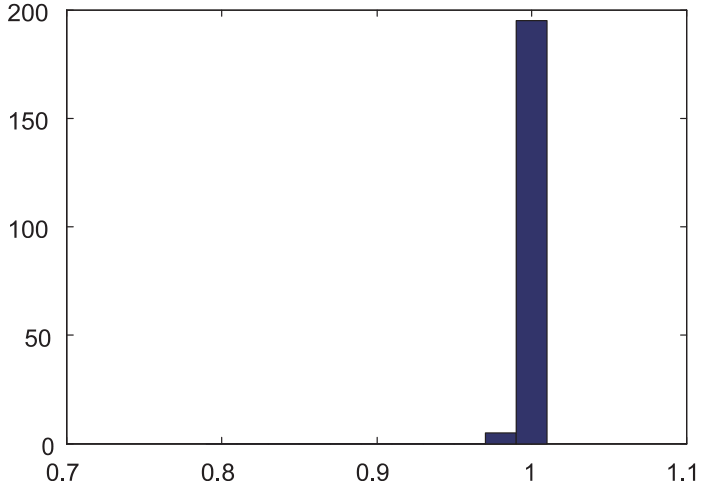


Aplikace na hudební záznam

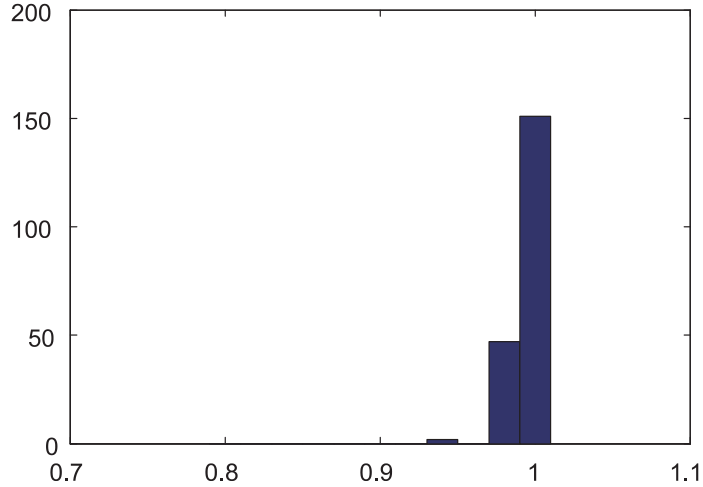


Aplikace na hudební záznam

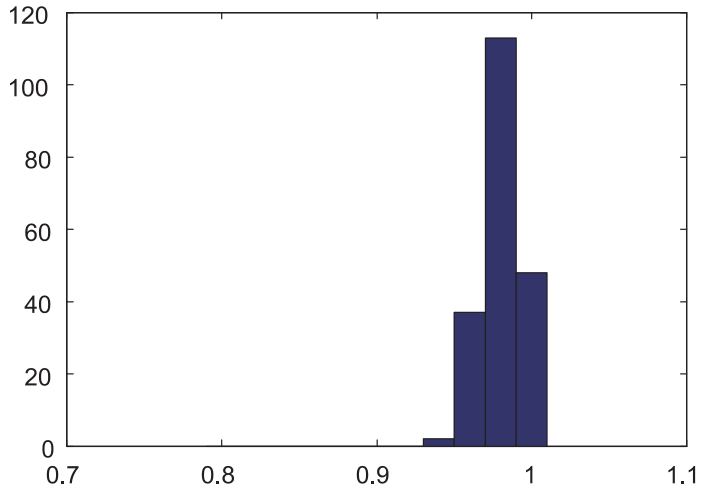
vážná hudba:



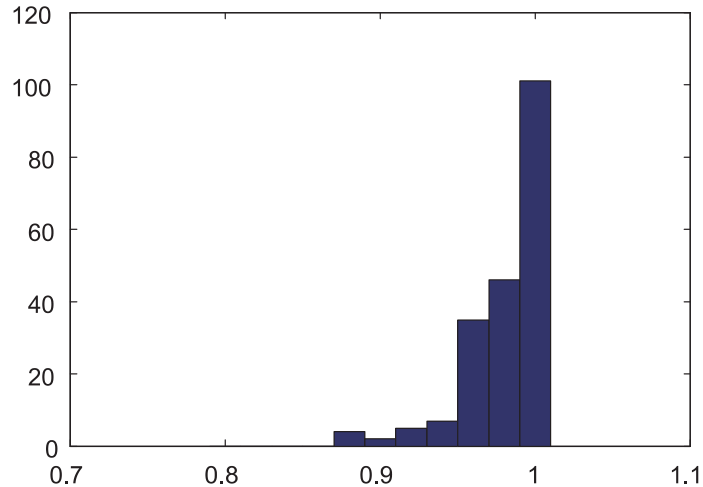
jazz:



rock:



electro-house:



Závěr

vs.

Díky za pozornost! ;)