CNR-IMATI Milano

December 14-16, 2004

WAVELETS AND SELF-SIMILARITY: THEORY AND APPLICATIONS

Lecture 3: Self-Similarity: An Appetizer

PLAN

1. **Fourier and Wolf's Numbers**
2. **Hurst and Nile Data**
3. **Ubiquity of Scaling**
4. **Why Data Scale?**

$\{X_t, t \in Z\}$ a real, weakly stationary time series with zero mean and autocovariance function $\gamma(h) = EX(t+h)X(t)$.

■ Spectral Density:

$$f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} \gamma(h) e^{-ih\omega}$$

■ Given the spectral density, the autocovariance function can uniquely be recovered via inverse Fourier transform,

$$\gamma(h) = \int_{-\pi}^{\pi} f(\omega) e^{ih\omega} d\omega, \;\; h = 0, \pm 1, \pm 2, \ldots .$$

■ The periodogram $I(\omega)$, based on a sample $X_0, \ldots, X_{T-1}$ is defined as

$$I(\omega_j) = \frac{1}{2\pi T} \left| \sum_{t=0}^{T-1} X_t e^{-it\omega_j} \right|^2 ,$$

where ω_j is the Fourier frequency $\omega_j = \frac{2\pi j}{T}$, $j = [-T/2] + 1, \ldots, -1, 0, 1, \ldots, [T/2]$.

```
function out = periodogram(ts)


out = abs(fftshift(fft(ts -
mean(ts)))).^2/(2*pi*length(ts));
```

Wolf's Sunspot Number Example

■ The Sun's activity peaks every 11 years, creating storms on the surface of our star that disrupt the Earth's magnetic field. These "solar hurricanes" can cause severe problems for electricity transmission systems.

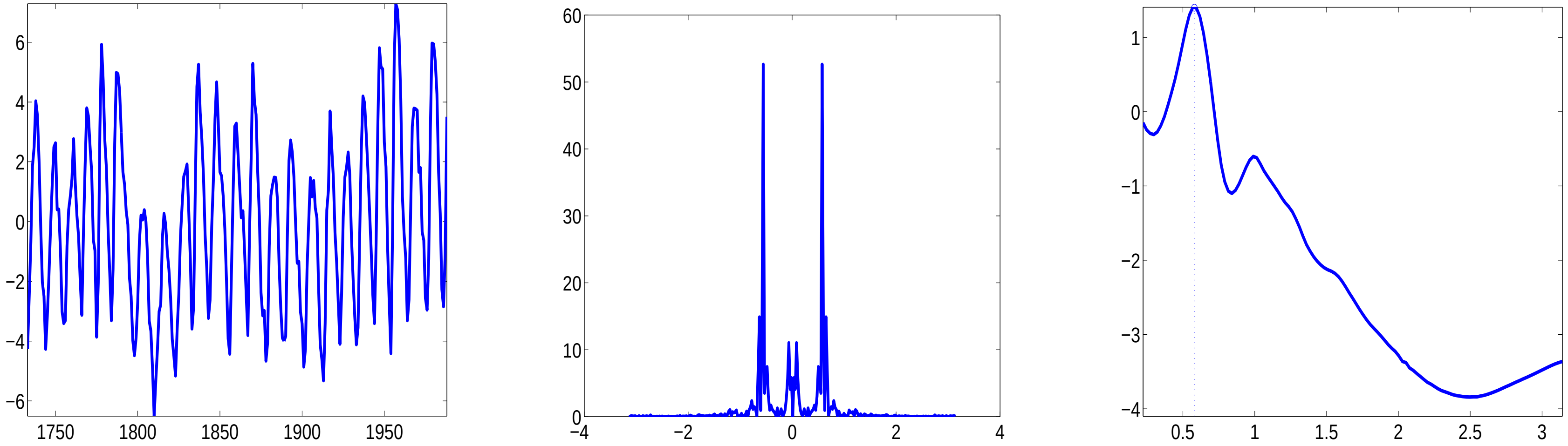
■ An example: 1989 power blackout in the American northeast.

■ Long and rich history starting with Galileo.

■ Estimates of daily activity date back to 1818, monthly averages can be extrapolated back to 1749, and estimates of annual values can be similarly determined back to 1700.

■ Observatory director Rudolph Wolf, who introduced what he called the Universal Sunspot Number as an estimate of the solar activity.

■ Data: The square root of Wolf's yearly sunspot numbers from 1733 till 1998.



The estimator peaks at frequency $\omega^* \approx 0.58$, corresponding to the Schwabe's cycle ranging from 9 to 11.5 (years), with an average of $\frac{2\pi}{0.58} \approx 10.8$ years.

Processes may have:

(i) Long Range Dependence (slowly decaying autocovariances)

(ii) Self-similarity, affinity, fractality, multifractality

(iii) Regular scaling

Joseph Effect: **Genesis 41** ... 17 Then Pharaoh said to Joseph, ”In my dream I was standing on the bank of the Nile, 18 when out of the river there came up seven cows, fat and sleek, and they grazed among the reeds. 19 After them, seven other cows came up–scrawny and very ugly and lean. I had never seen such ugly cows in all the land of Egypt. 20 The lean, ugly cows ate up the seven fat cows that came up first. 21 But even after they ate them, no one could tell that they had done so; they looked just as ugly as before. Then I woke up. 22 ”In my dreams I also saw seven heads of grain, full and good, growing on a single stalk. 23 After them, seven other heads sprouted–withered and thin and scorched by the east wind. 24 The thin heads of grain swallowed up the seven good heads. I told this to the magicians, but none could explain it to me.” 25 Then Joseph said to Pharaoh, ”The dreams of Pharaoh are one and the same. God has revealed to Pharaoh what he is about to do. 26 The seven good cows are seven years, and the seven good heads of grain are seven years; it is one and the same dream. 27 The seven lean, ugly cows that came up afterward are seven years, and so are the seven worthless heads of grain scorched by the east wind: They are seven years of famine.

It Started with Hurst and Nile Data

British hydrologist Harold Edwin Hurst spent 62 years in Egypt working on design and construction of reservoirs along the Nile River. By inspecting historical data on the Nile River flows, Hurst discovered phenomenon (now called Hurst effect).

■ Problem: Optimal reservoir capacity R such that the reservoir holds the river flow in N units of time, $X_1, X_2, \dots X_N$, with a constant withdrawal of $\bar{X}$ per unit time.

■ The optimal volume of the reservoir was given by the so called adjusted range,

$$R = \max_{1 \le k \le N} (X_1 + \cdots + X_k - k\bar{X}) - \min_{1 \le k \le N} (X_1 + \cdots + X_k - k\bar{X})$$

■ Since the records for the waterflow rarely exceeded 100 years Hurst inspected other geophysical data and in order to compare them, he standardized their adjusted ranges R, with sample standard deviation

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2} \; ,$$

and obtained dimensionless ratio R/S - rescaled and adjusted range.

■ On basis of more that 800 records, he found (Hurst, 1951) that quantity R/S scales as N^H, for ranging from 0.46 to 0.93, with mean 0.73 and standard deviation of 0.09.

■ This result was is contrast with the fact that for independent normal random variables H is 1/2 in limit.

■ Feller proved that the limit is 1/2 for independent identically distributed random variables with finite second moment.

■ It was believed that strong Markovian dependence was responsible for this deviation until Barnard (1956) proved that limit $H = 1/2$ holds for the Markovian case.

■ It was the work of Mandelbrot (1975), Mandelbrot and Van Ness (1968), and Mandelbrot and Wallis (1968) who associated the Hurst (or Joseph) phenomenon on the presence of long-memory.

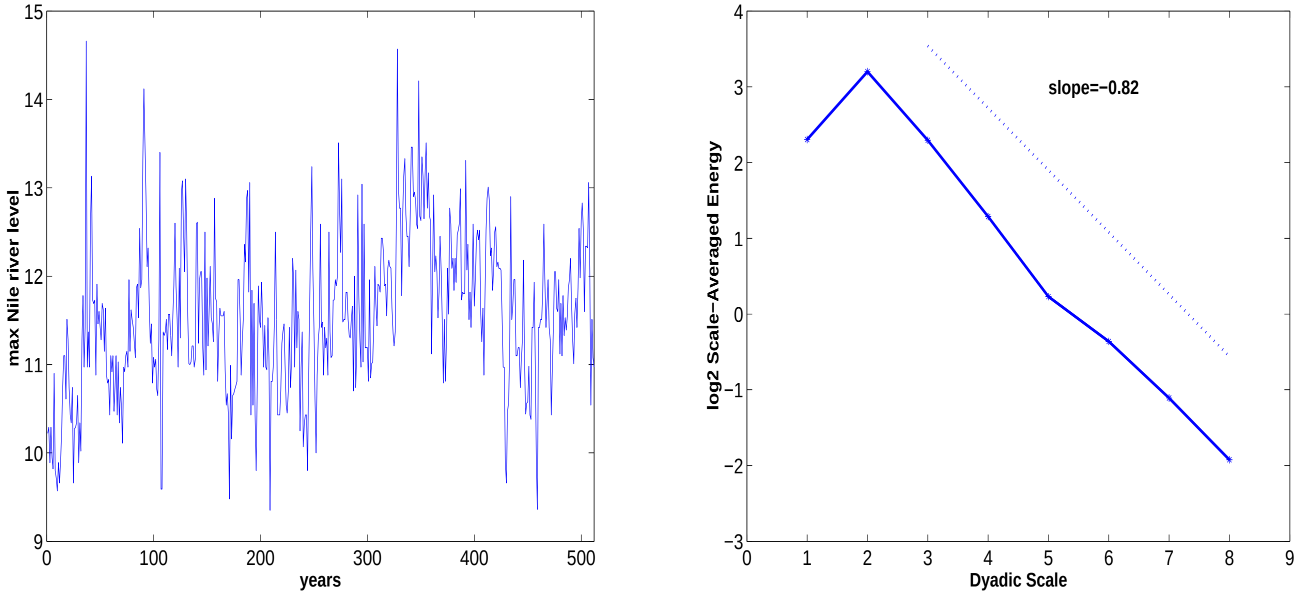


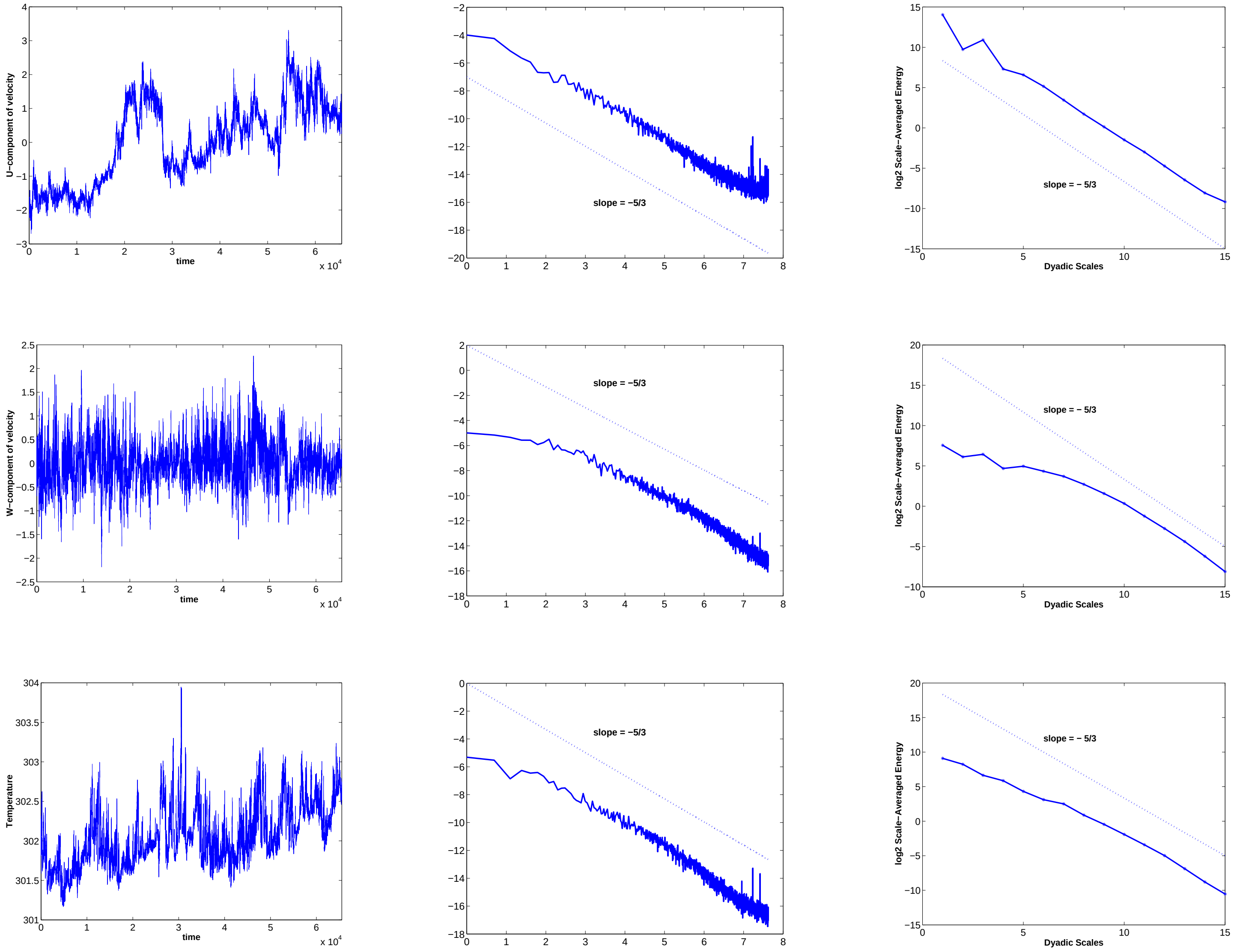
Figure shows gives $n = 512$ consecutive yearly measurements from the famous Nile River Data set for the years 62-1281 A.D. Right panel shows its log-spectra demonstrating the scaling law.

Turbulence

■ The velocity and air temperature measured July 12-16, 1995, at 5.2 m above the ground at the Blackwood division of the Duke Forest in Durham, North Carolina. During the experiment, maximum mean air temperature up to $38^\circ C$ was measured in Durham. The sky condition during these five days was clear with low to moderate winds. The site is a 480 m by 305 m grass-covered forest clearing ($36^\circ 2' N \quad 79^\circ 8' W$, elevation $= 163\ m$)

■ The velocity components (U, V, W) and air temperature T were measured by a triaxial ultrasonic anemometer (Gill Instruments/1012R2).

■ The sampling frequency (f_s) and period (T_p) were 56 Hz and 19.5 minutes, respectively, resulting in $N = 65,536$ measurements per per run.


U−component of velocity
time
x 10^4
4
3
2
1
0
−1
−2
−3
0
1
2
3
4
5
6
7
8
−2
−4
−6
−8
−10
−12
−14
−16
−18
−20
slope = −5/3
log2 Scale−Averaged Energy
Dyadic Scales
15
10
5
0
−5
−10
−15
slope = − 5/3
W−component of velocity
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
Temperature
304
303.5
303
302.5
302
301.5
301
20

■ Kolmogorov developed his theory, often referred to as **K41** theory, for *locally isotropic* turbulence. Let $\boldsymbol{x} = (x_1, x_2, x_3)$ be the position vector and $\boldsymbol{u} = (u_1(\boldsymbol{x}), u_2(\boldsymbol{x}), u_3(\boldsymbol{x}))$ be the velocity components.

■ The probability distribution of the relative velocity differences

$$\Delta \boldsymbol{u}(\boldsymbol{r}) = \boldsymbol{u}(\boldsymbol{x} + \boldsymbol{r}) - u(\boldsymbol{x}),$$

is independent of time, and invariant under translations, reflections, and rotations.

■ The fundamentals in **K41** theory are *structure functions*

$$\langle \Delta \boldsymbol{u}(\boldsymbol{r})^2 \rangle.$$

■ Structure functions are closely related to correlations of two-point velocity differences,

$$\langle \Delta \boldsymbol{u}(\boldsymbol{r})^2 \rangle = 2\sigma_u^2 (1 - \rho_u(r)).$$

■ A (longitudinal) structure function of order p is defined as

$$D_p(r) = \langle ||\Delta \boldsymbol{u}(\boldsymbol{r})||^p \rangle$$

where the angular brackets denote time averaging.

■ A functional description for the moments of velocity differences can be derived using *dimensional analysis* and leads to

$$D_p(r) = C_p[\langle \epsilon \rangle \ r]^{\frac{p}{3}},$$

where C_p is a universal constant.

■ For the third-order structure function, it can be inferred directly from the Navier-Stokes equations that $C_3 = -\frac{4}{5}$.

■ It follows that structure functions possess scaling behavior,

$$D_p(r) \propto r^{\zeta_p}.$$

■ The exponent ζ_p is called the *scaling exponent.* The **K41** theory gives the simple model $\zeta_p = \frac{p}{3}$.

■ Similarly, as for the structure functions, a description of the energy of the turbulent fluctuations per unit of mass of fluid in scales r can be derived from the hypotheses and by dimensional analysis,

$$E_r \propto (r)^{\frac{2}{3}}.$$

■ Via the Fourier transform of E_r, which results in the spectral density $\phi(k)$, the celebrated "$-\frac{5}{3}$ law" for the power spectrum is obtained,

$$E_k = 2R^{-1}k^2\phi(k) \propto k^{-\frac{5}{3}}.$$

DNA Random Walks

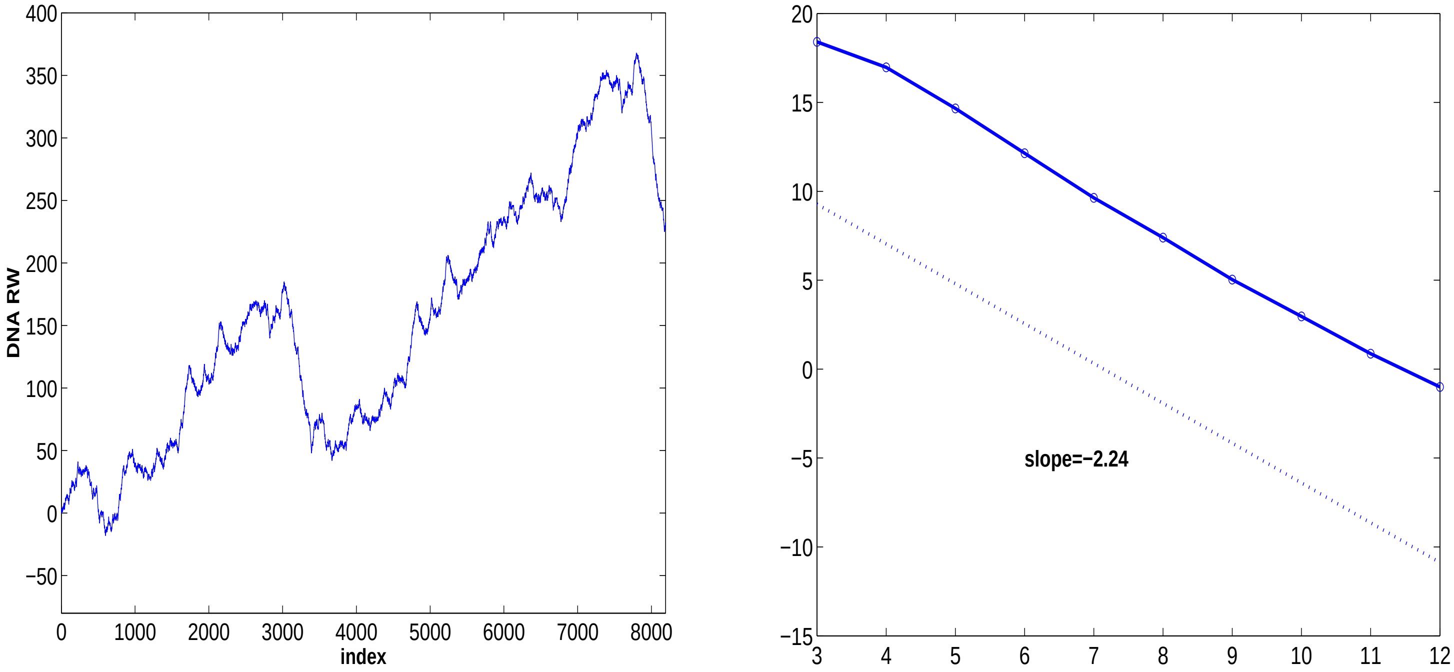
■ In all eucariothic species, a DNA molecule consists of long complementary double helix of purine nucleotides (denoted as A and G) and pyrimidine nucleotides (denoted as C and T).

■ A single strain of this DNA can be represented as a long word that corresponds to a random walk.

■ **A, G** $\longrightarrow x(i) = +1$

■ **C, T**$\longrightarrow x(i) = -1$

■ The random walk is defined as $s(n) = \sum_{i=1}^{n} x(i),\ n \geq 1$ DNA random walks have been first proposed by Peng et al. (1992).



■ A 8196-long DNA random walk for spider, from EMLB Nucleotide sequence alignment DNA database.

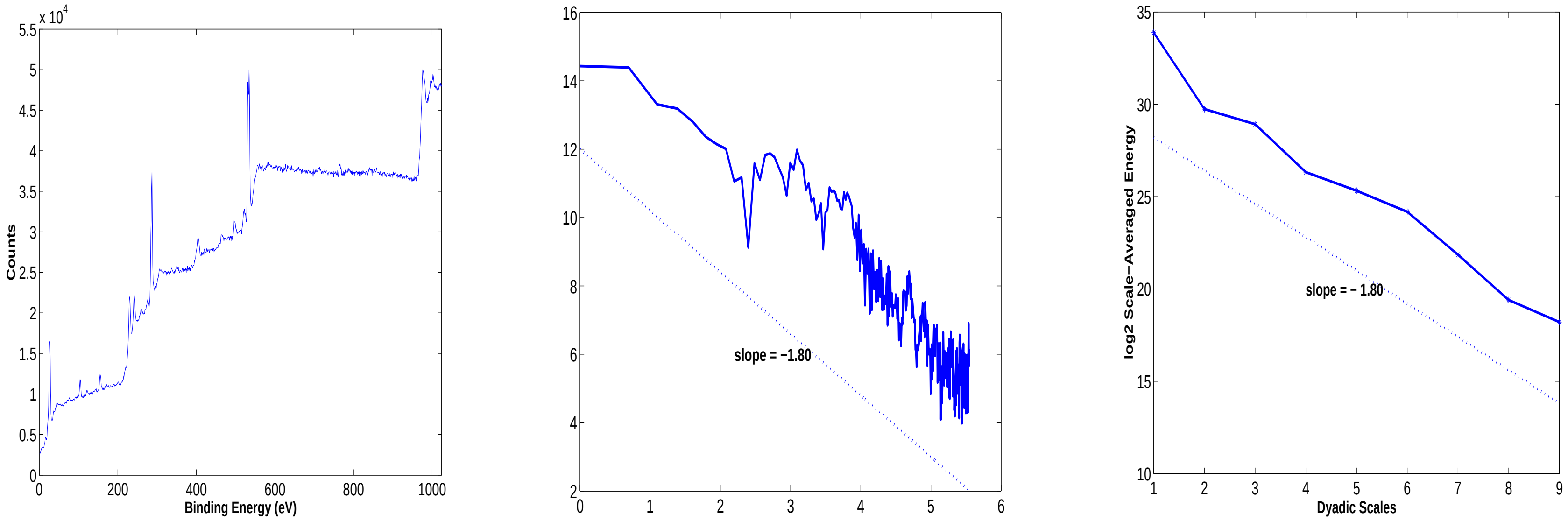
■ Bacry et al. (1995) explored self-similarity and fractality of DNA walks for humans. They find that the Hurst exponent for non-coding sequences (introns) is about 0.6 while for the coding ones (exons) the exponent is close to 0.5.

ESCA Spectrum

■ The ESCA spectrum (J.P. Bibérian, of the Université de Marseille – Luminy). This set is one of the Template data set in WaveLab 802.

■ Electron Spectroscopy for Chemical Analysis (ESCA), also referred to as X-ray Photoelectron Spectroscopy (XPS), irradiates the sample surface with a soft (low energy) X-ray. This X-ray excites the electrons of the sample atoms, and if their binding energy is lower than the X-ray energy, they will be emitted from the parent atom as a photoelectron.

■ Only the photoelectrons at the extreme outer surface (10-100 Angstroms (Å); 1 Å$= 10^{-10}m$) can escape the sample surface, making this methodology a surface analysis technique.



■ An ESCA spectrum consists of a series of peaks corresponding the the binding energies of the photoelectrons that produced these peaks.

■ ESCA analysis not only provides elemental information, but because the technique is detecting the binding energy of emitted electrons, it can also provide some chemical bonding information.

■ Figure shows the ESCA spectrum. Clear power law with the slope of - 1.80 is notable.

Stock Market Prices and Exchange Rates

■ Many economic time series, such as stock market prices, exchange rates and asset returns exhibit scaling laws and long range dependence.

■ This is in empirical contradiction to several economic theories (random walk theory for stock market, perfect markets, etc) and gave rise to several theories and models describing the scaling and LRD (such as ARFIMA, fGn, fBm, GARCH, etc).

■ Coca Cola stock market prices and rates of exchange between Hong Kong Dollar (HKD) and USDollar (USD).

- Coca Cola Stock Market Prices.

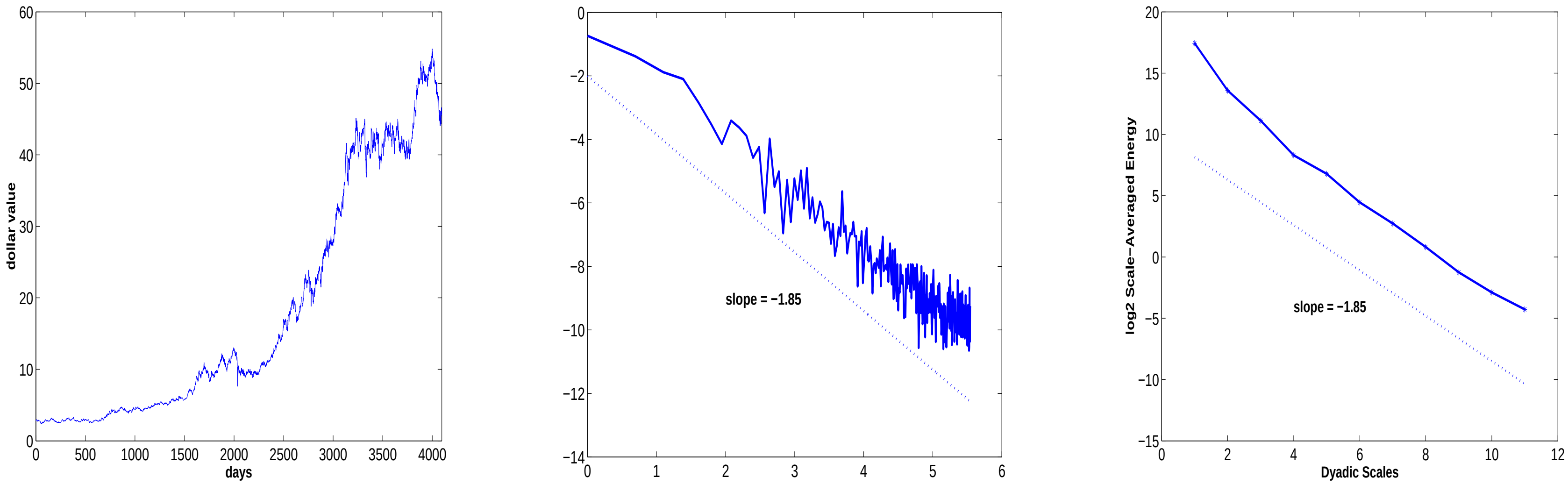


Figure 1: (a) Coke Stock Market Prices; (b) scaling behavior in the Fourier domain, and (c) in the wavelet domain.

- The rates of exchange between Hong Kong Dollar (HKD) and USDollar (USD) as reported by the ONADA Company between 24 March 1995 and 1 November 2000.

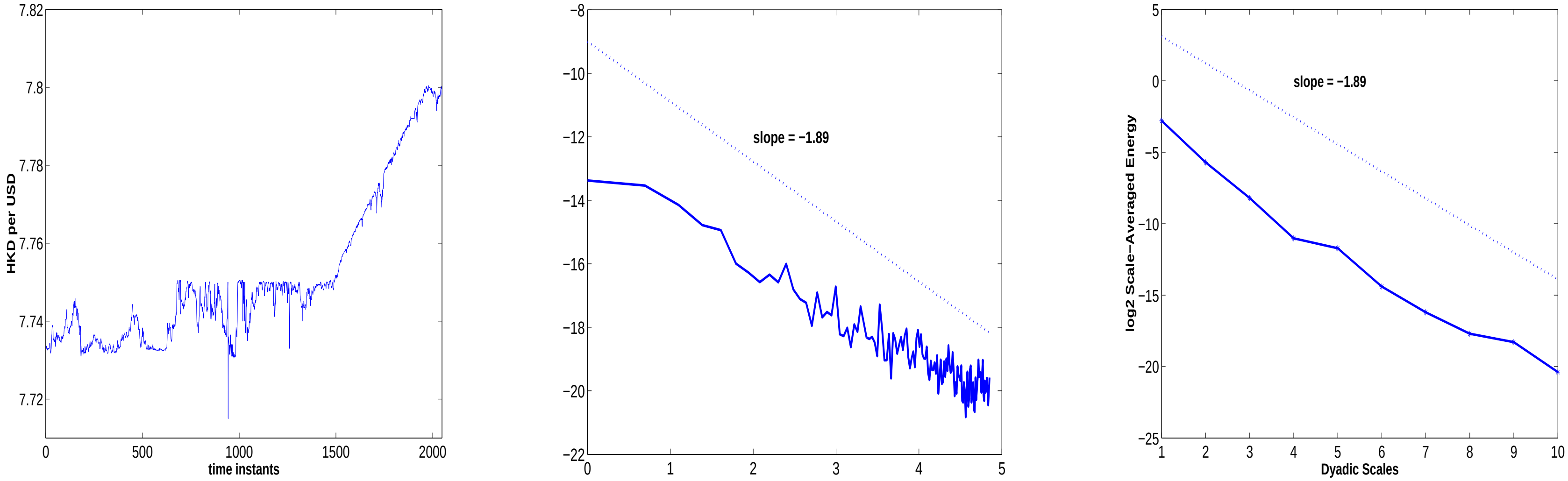


Figure 2: (a) Exchange Rates HKD per US$; (b) scaling behavior in the Fourier domain, and (c) in the wavelet domain.

Gait Data

- Scaling laws were recently detected in the apparently “noisy” variations in the stride interval (duration of the gait cycle) of human walking.

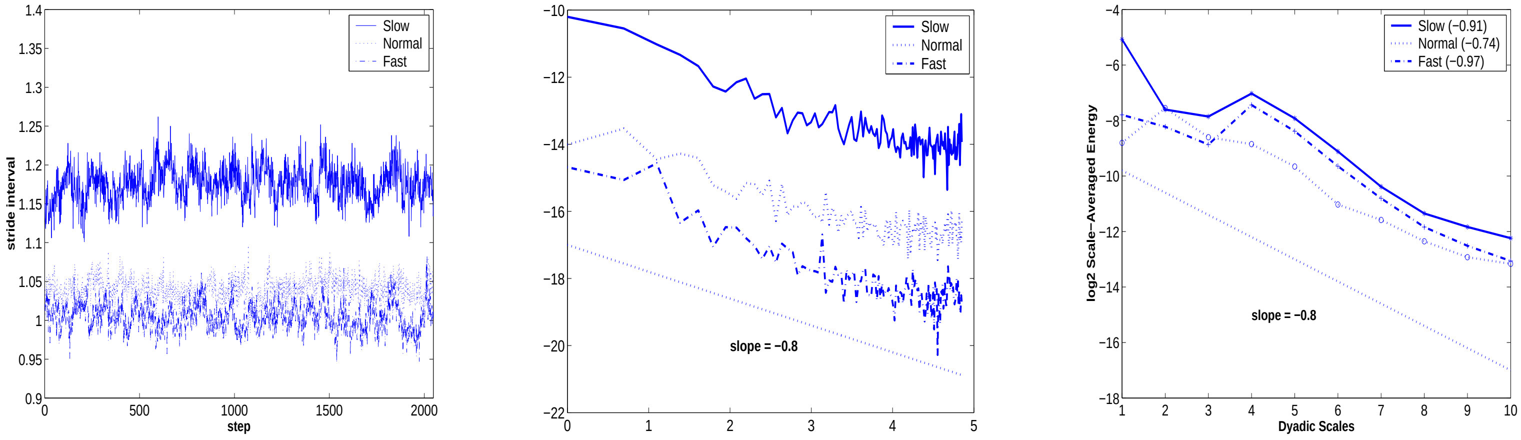
- The experimental data consist of measurements on a healthy subject who walked for 1 hour at his usual, slow and fast paces. The stride interval fluctuations exhibited long-range correlations with power-law decay for up to a thousand strides at all three walking rates.

- It is curious that during metronomically-paced walking, these long-range correlations disappeared; variations in the stride interval were anti-correlated.

■ Participants in this experiment had no history of any neuromuscular, respiratory or cardiovascular disorders, and were taking no medications. Mean age was 21.7 years.

■ Subjects walked continuously on level ground around an obstacle free, long (either 225 or 400 meters), approximately oval path and the stride interval was measured using ultra-thin, force sensitive switches taped inside one shoe.

■ Figure shows 2048 data points for one subject. Slow and fast stride intervals have slopes of -0.91 and -0.97 respectively, and stride intervals for normal walk show scaling with -0.74 slope.

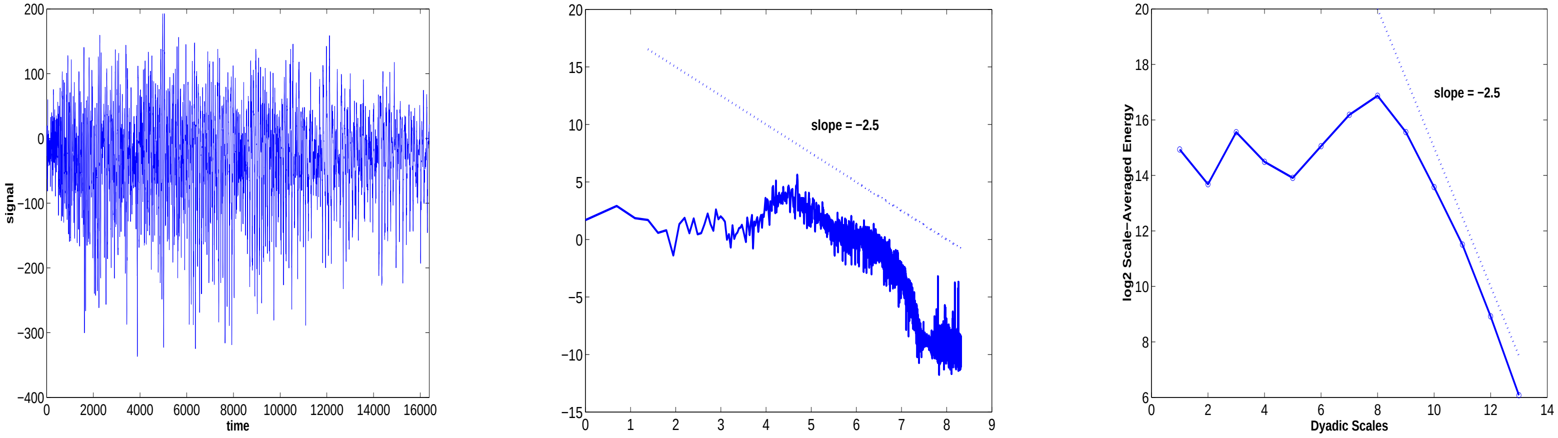


EEG Data

■ This data set represents fluctuations of measured electrical potential (in μV) derived from brain activity of a patient during an epileptic seizure.

■ Recorded in the ECT Lab at Duke University Medical Center (Curtesy of Dr. B. Krystal).

■ A patient undergoing ECT therapy had measuring electrodes in his scalp and this particular time series is one of several “channels.”



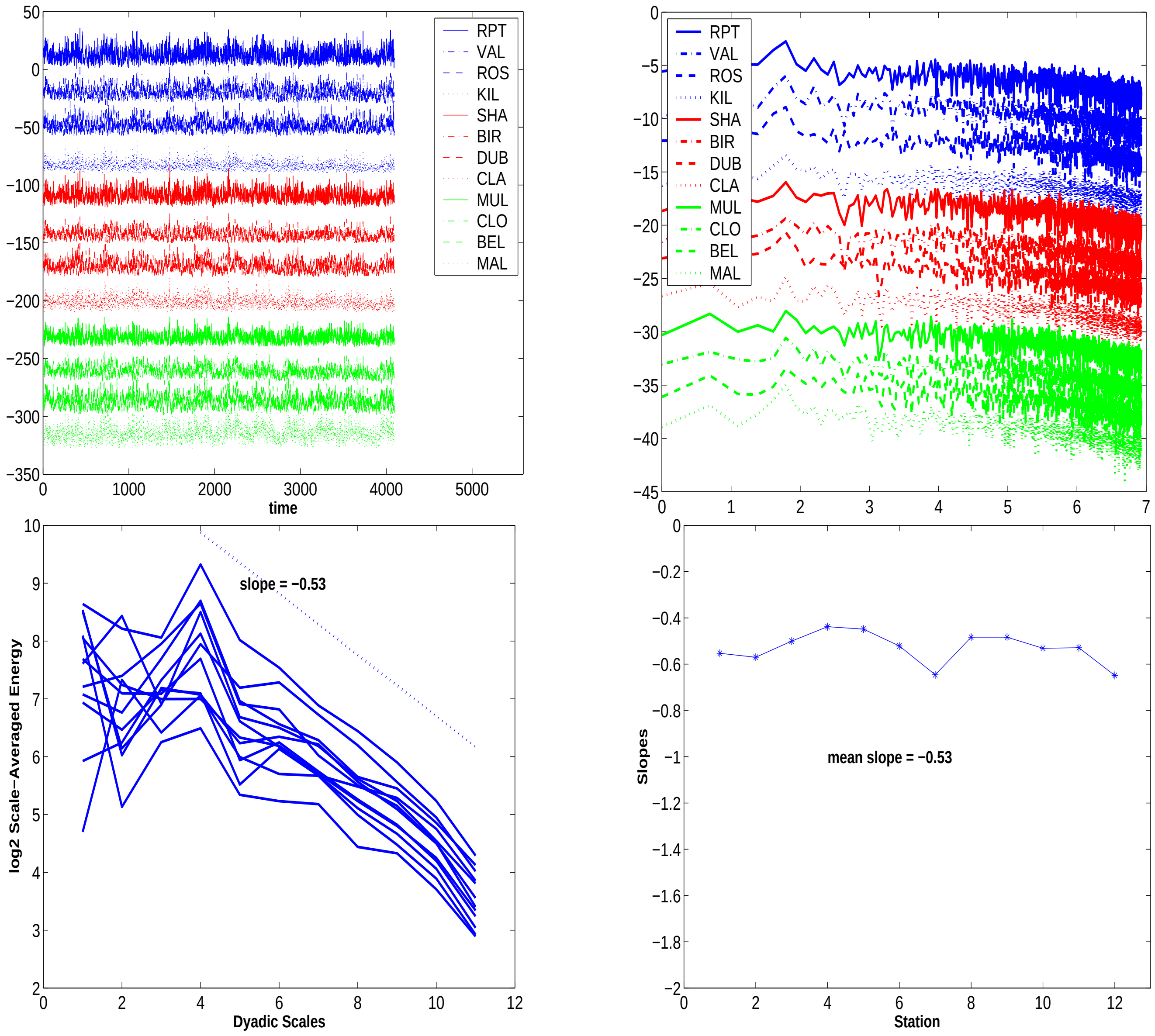
■ Outstanding problems for this kind analysis include the prediction, classification, and space-time localization of seizures, see Benedetto and Colella (1995) for wavelet based diagnostic methodology. The original data set covers a 104-second span at a frequency of 256 observations per second, but for our analysis we took a mid-segment of length 2^{14}.

■ A power law with slope of -2.5 was found only at the end of spectrum (several "binomial decades").

Wind Speed

- The classical data set of Haslett and Raftery (1989) contains daily average wind speeds for 1961-1978 at 12 synoptic meteorological stations in the Republic of Ireland.

- Recorded are square roots of daily wind speeds in knots (1 knot = 0.5148 metres/second). The 12 stations are Roche's Point, Valencia, Rosslare, Kilkenny, Shannon, Birr, Dublin, Mullinger, Claremorris, Clones, Belmullet and Malin Head, (RPT, VAL, ROS, KIL, SHA, BIR, DUB, CLA, MUL, CLO, BEL, MAL) as indicated by the map on page 2 of the Haslett-Raftery paper.

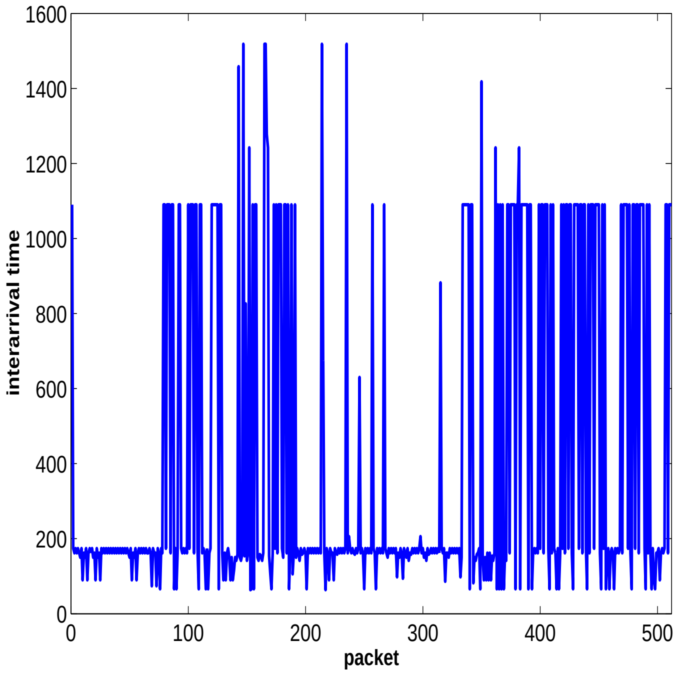

RPT
VAL
ROS
KIL
SHA
BIR
DUB
CLA
MUL
CLO
BEL
MAL
50
0
−50
−100
−150
−200
−250
−300
−350
0
1000
2000
3000
4000
5000
time
0
−5
−10
−15
−20
−25
−30
−35
−40
−45
0
1
2
3
4
5
6
7
slope = −0.53
log2 Scale−Averaged Energy
10
9
8
7
6
5
4
3
2
0
2
4
6
8
10
12
Dyadic Scales
mean slope = −0.53
Slopes
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
−1.4
−1.6
−1.8
−2
0
2
4
6
8
10
12
Station

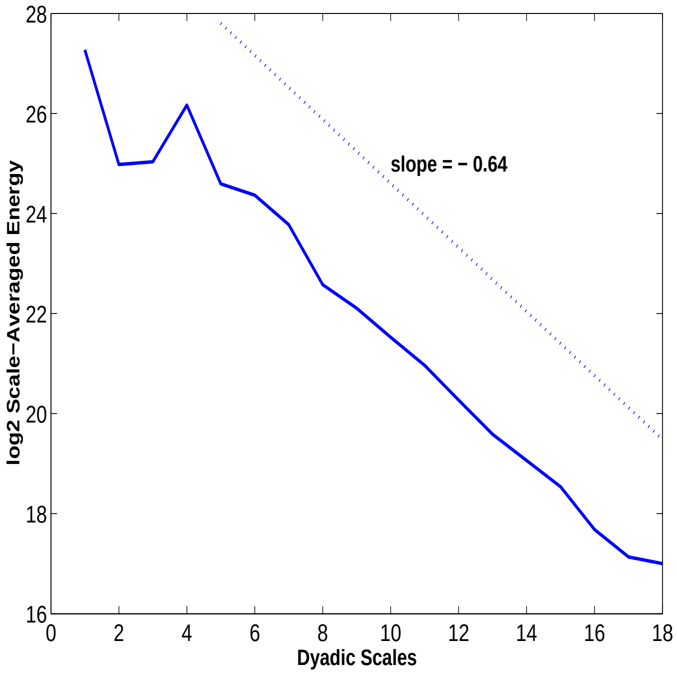
Bellcore Internet Data

- The original file that contains a record of a million packet arrivals on an Ethernet was compiled at the Bellcore Morristown Research and Engineering facility. Each line contains a floating-point time stamp (representing the time in seconds since the start of a trace) and an integer length (representing the Ethernet data length in bytes).

- The hardware clock had an actual resolution of 4 microseconds.

- The trace in file *BC-pAug89*[from http://ita.ee.lbl.gov/html/contrib/BC.html] began at 11:25 on August 29, 1989, and ran for about 3142.82 seconds (until 1,000,000 packets had been captured). In our scaling analysis we used 2^{19} data points and the Figure shows 512 data points (about 2 seconds of data).





■ Due to intrinsic nature of the traffic, the data are clustering into three groups (sizes), at about 150, 1100, and 1500 bits. Clearly, the package size is not described by any nice, a textbook distribution.

■ In our scaling analysis we used 2^{19} data points and the Figure shows 512 data points (about 2 seconds of data). These data were first analyzed by Leland *et al.* (1994) and after, by other researchers who utilized various tools (ARFIMA, Stable Levy Processes, etc).

ON-OFF Data

```
ii=2^9;
obj=[];
mu=[0 1];
k=1;
for i = 1:ii
    jj = floor(rand(1,1)*2^8)+1;
    ind= floor(2*rand)+1;
        for j=1:jj
        obj=[obj mu(ind)];
    end
end
```

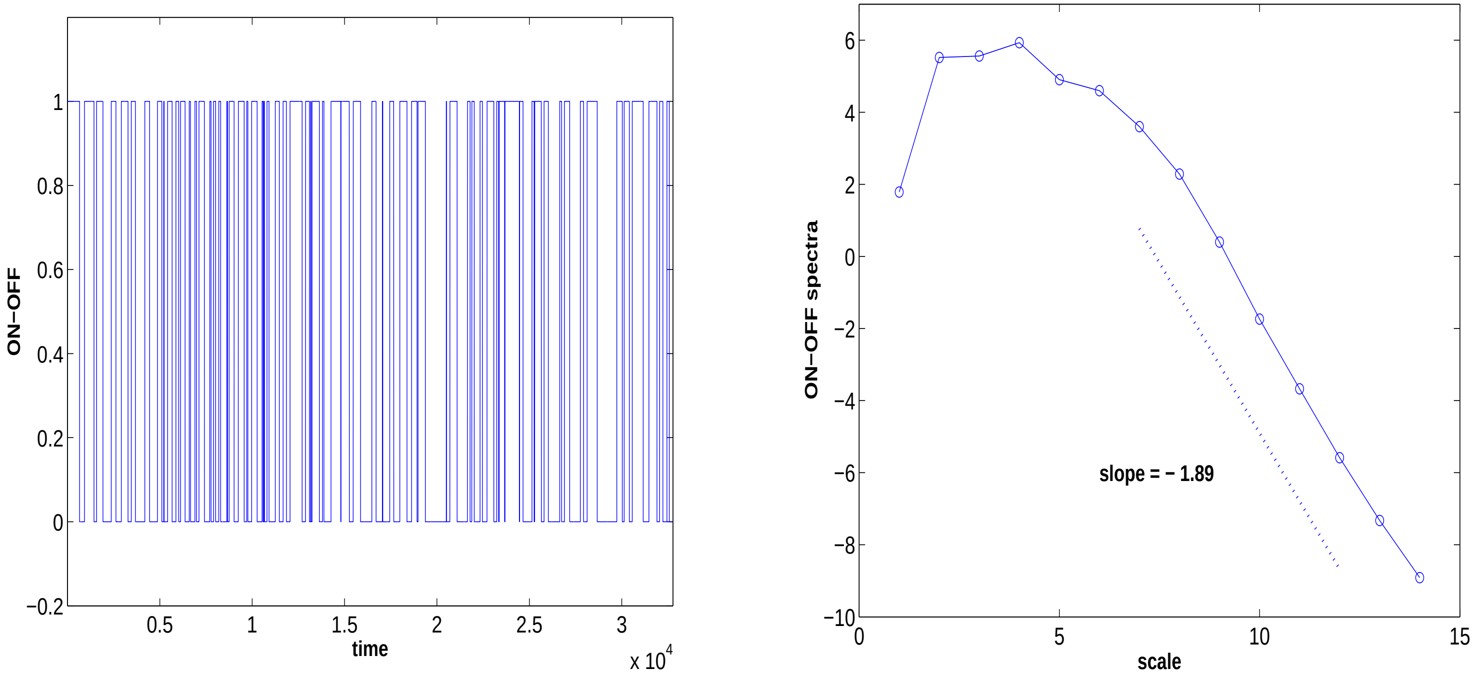


Figure 3: (a) OnOff Data with Uniform Duration; (b) Scaling Behavior in the Wavelet domain.

More Examples...

- High Frequency Measurements of Pupil Diameter (200Hz)
- Industrial Production: Chicken on Kill Line (1 year worth of data at rate 180 per minute).
- Orthosis Data. The data acquired by Dr. Amarantini David and Dr. Martin Luc (Laboratoire Sport et Performance Motrice, EA 597, UFRAPS, Grenoble University, France.
- Georgia Tech Biology Lab Multiple Channel Brain Signals up to 1000Hz.
- Danube level data.
- Etc...