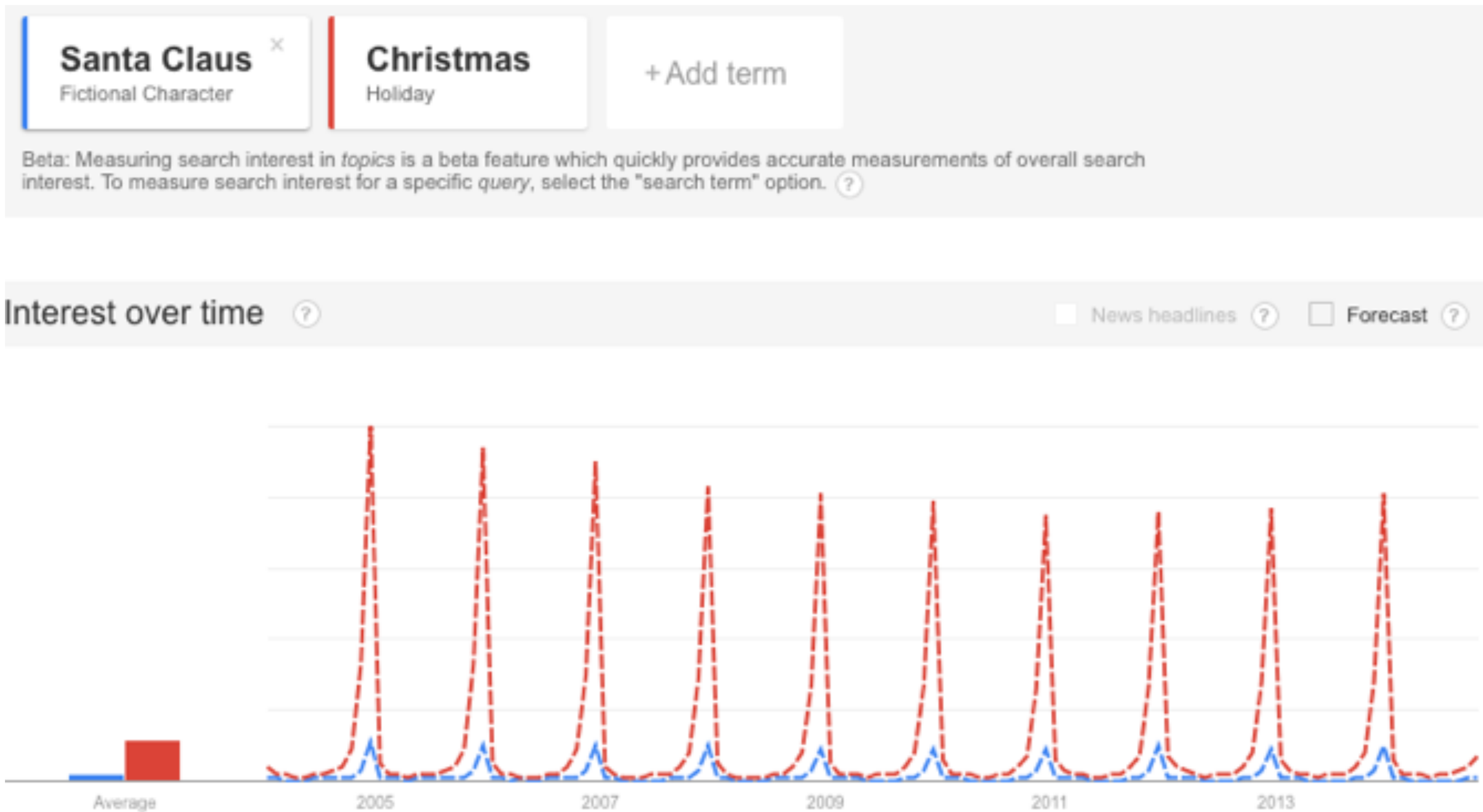
Foundations - 2

Periodicity Detection, Time-series Correlation, Burst Detection

Time Series

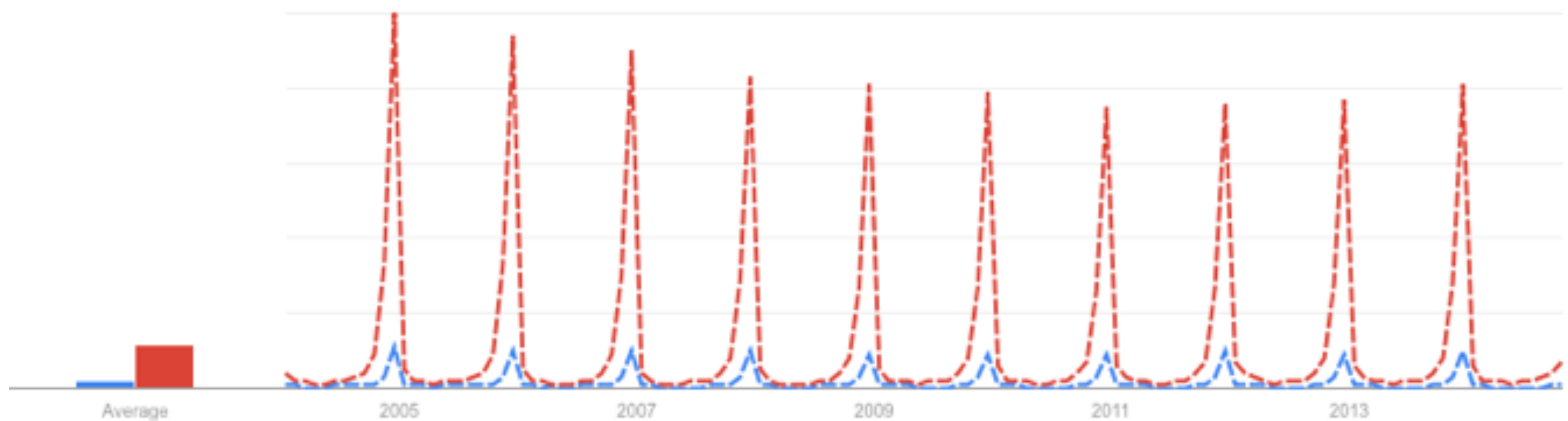
- An ordered sequence of values (data points) of variables at equally spaced time intervals

Periodicity Detection



- How does one identify periodic values

Periodicity Detection

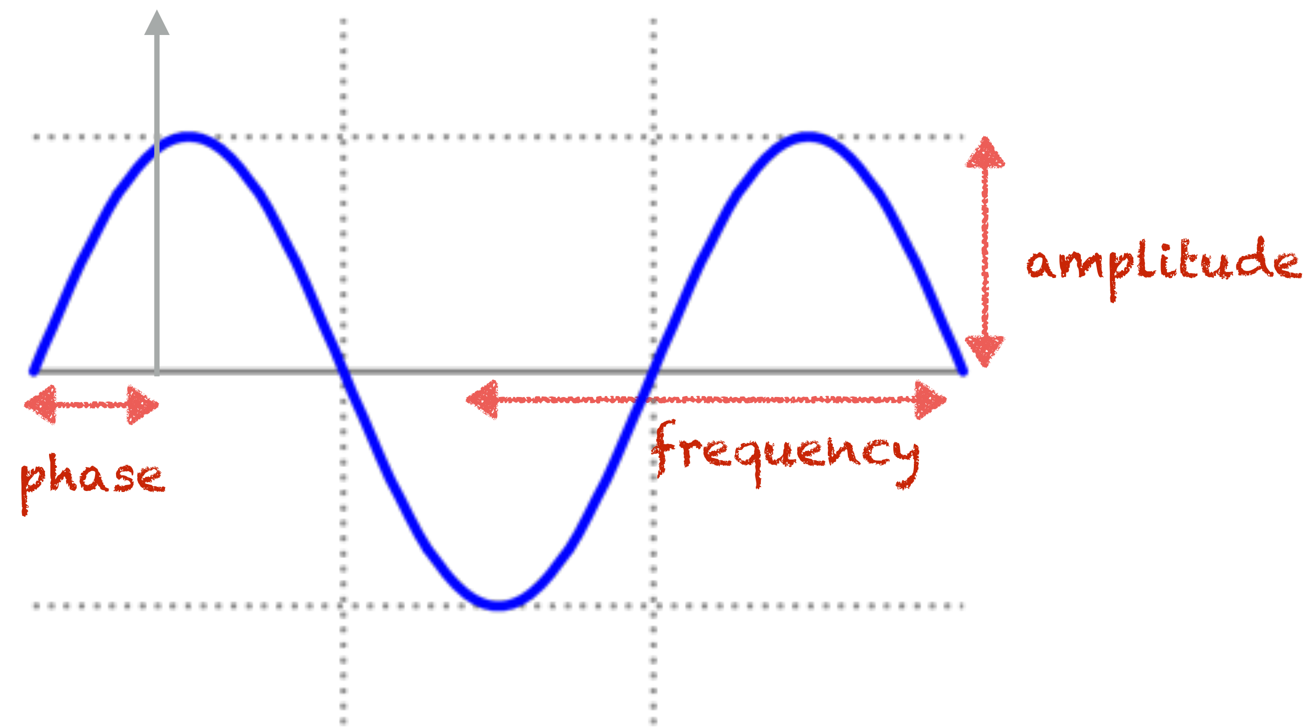


- Time-series is in the time domain
- Method1 (DFT): Identify the underlying periodic patterns by transforming into the frequency domain
- Method 2 (Autocorrelation) Correlate the signal with itself

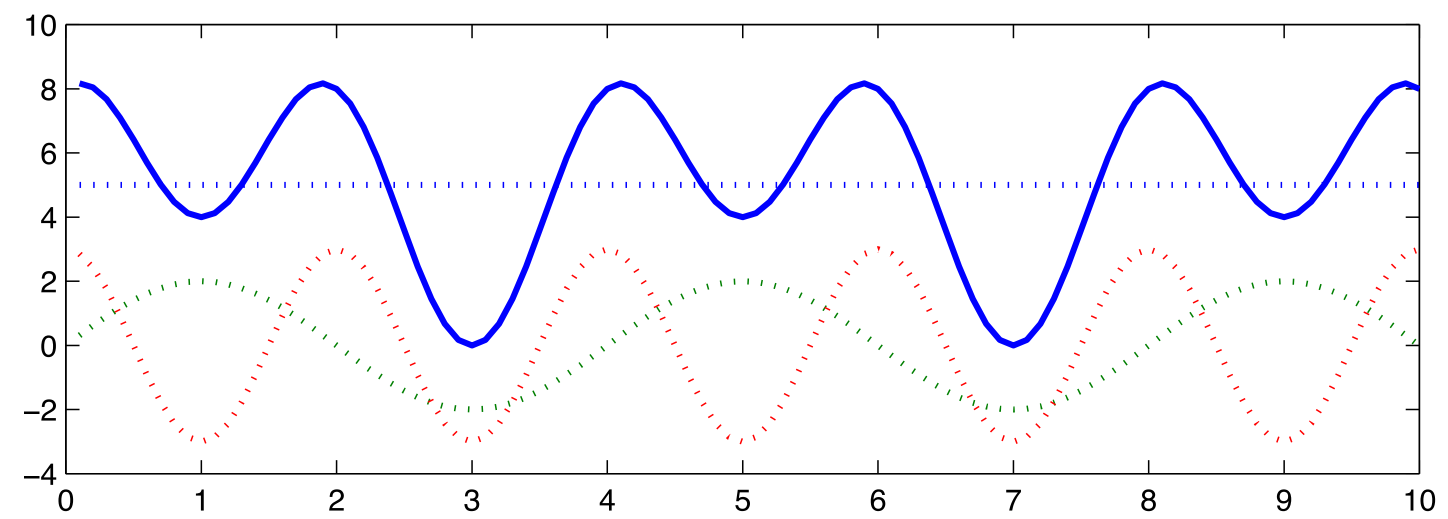
Find dominant frequencies

Fourier Transform

- A signal has an amplitude (strength), frequency (periodicity) and phase (offset)



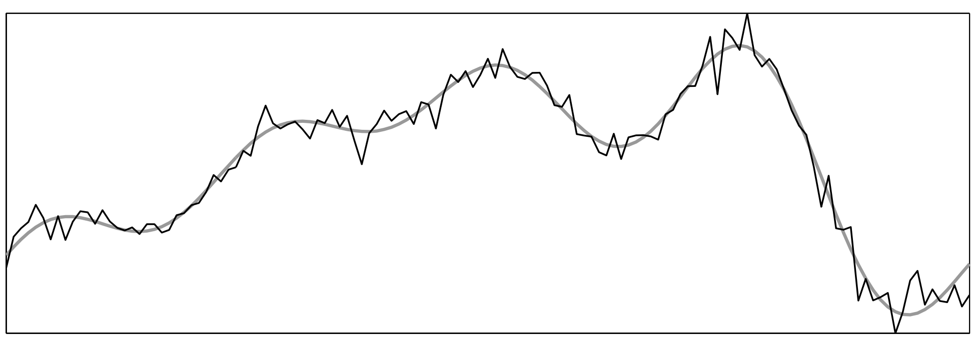
- Fourier Transform converts a signal from the time domain to the frequency domain



Discrete Fourier Transform (DFT)

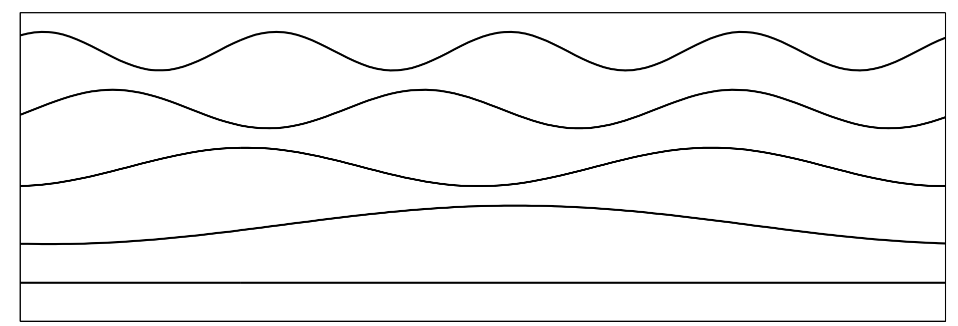
- A Fourier analysis is a method for expressing a function as a sum of periodic components, and for recovering the function from those components.

- When both the function and its Fourier transform are replaced with discretized counterparts, it is called the discrete Fourier transform (DFT).



fourier transform

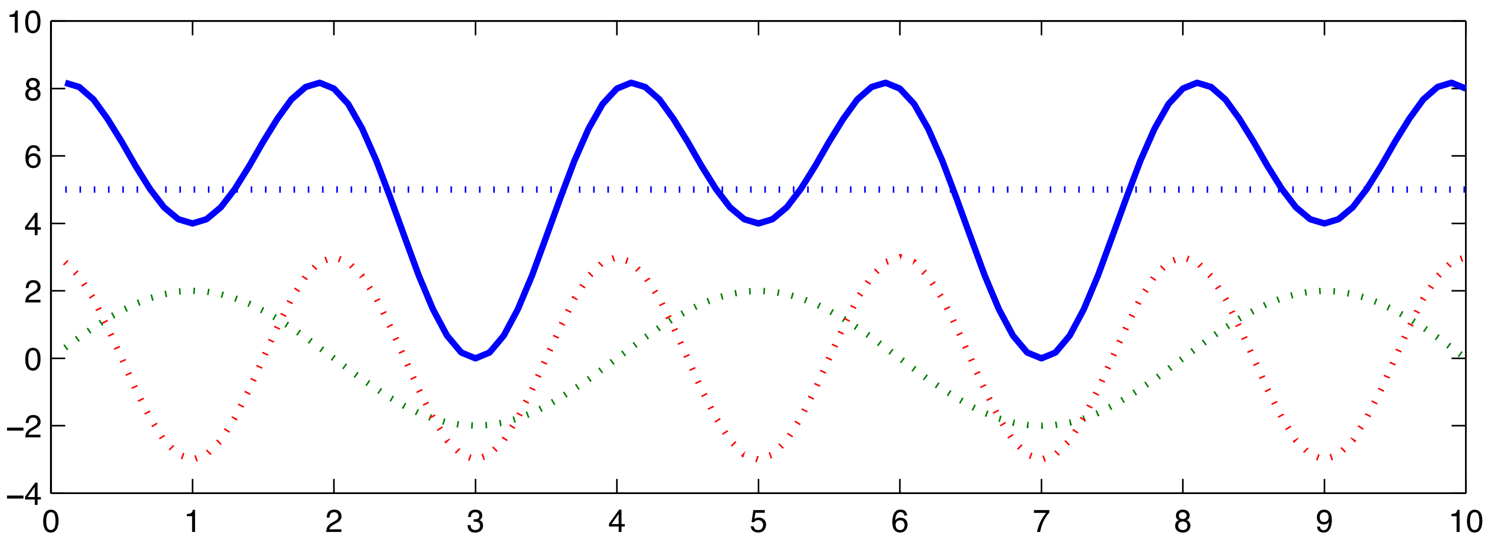
inv. fourier transform



Advantages of DFT apart from periodicity detection ?

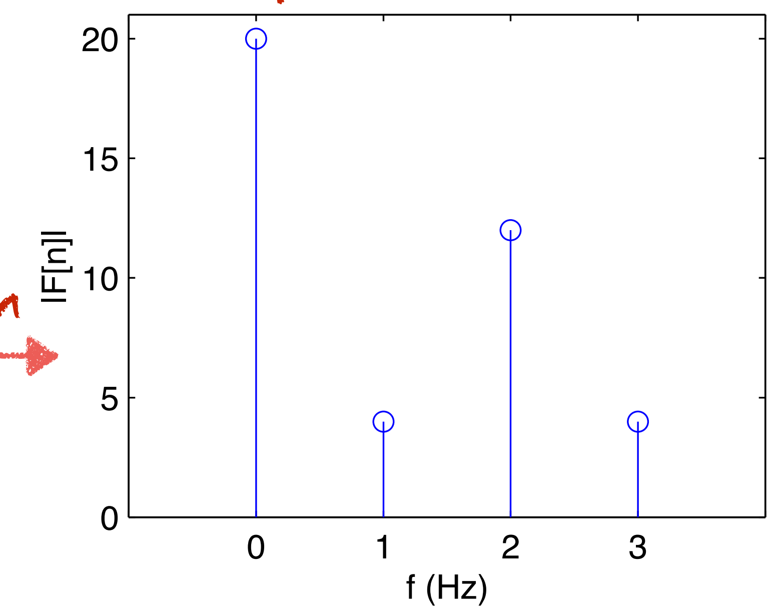
denoising, compression

Discrete Fourier Transform (DFT)



fourier transform

periodogram



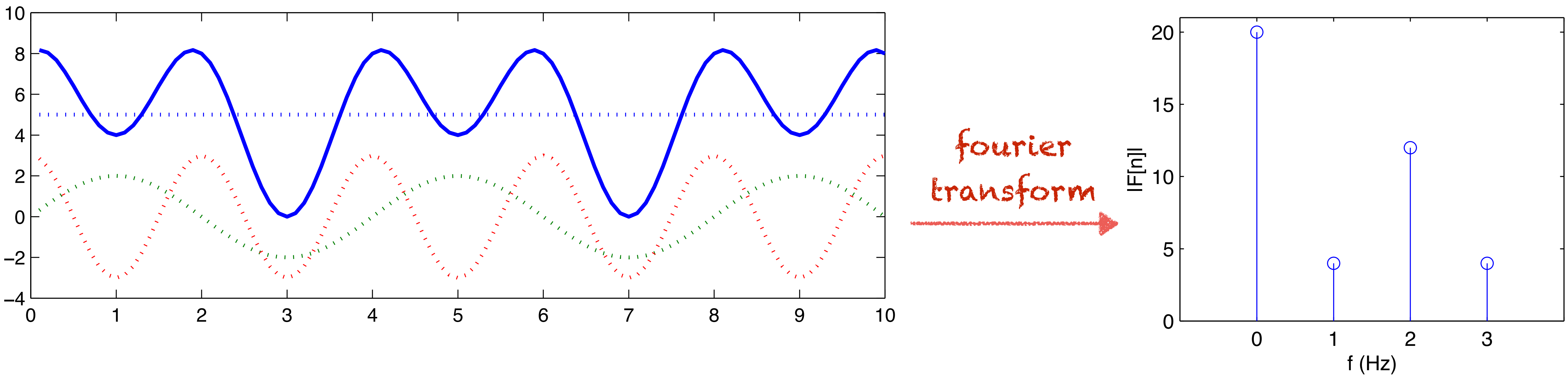
sinusoid

fourier coefficients

$$X(f_{k/N}) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x(n)\ e^{\frac{-j2\pi kn}{N}}$$

- The fourier coefficients encode both the amplitude and phase

Power Spectral Density (PSD) Estimation

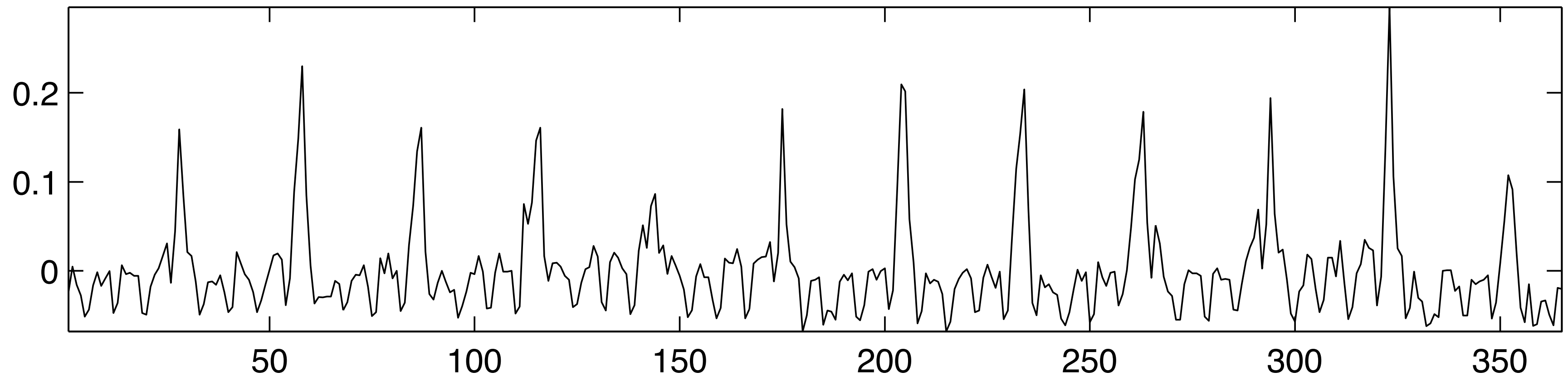


- To find out the dominant frequency we need to find the power at each frequency

- **Periodogram** encodes the strength at a given frequency

$$\mathcal{P}(f_{k/N}) = \|X(f_{k/N})\|^2 \quad k = 0, 1 \ldots \lceil \tfrac{N-1}{2} \rceil$$

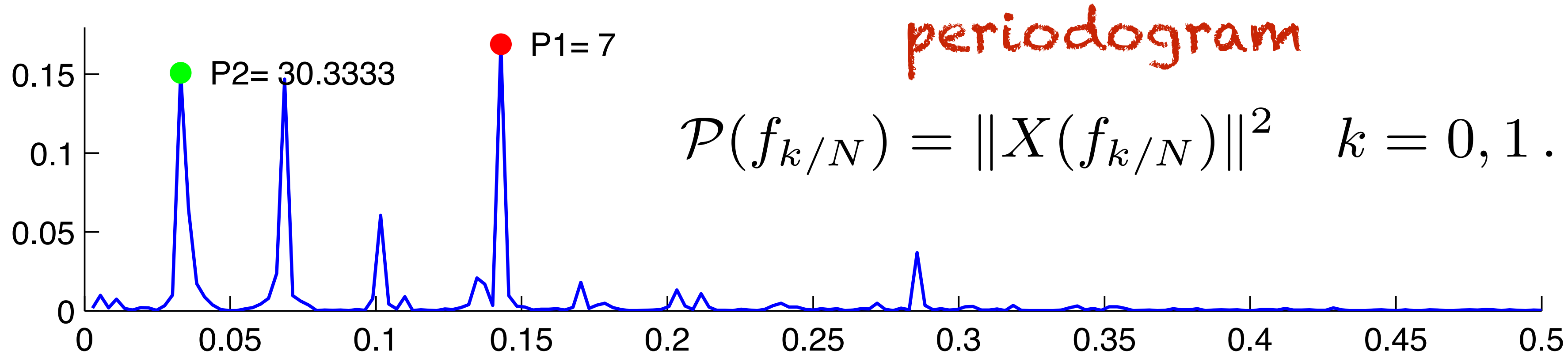
PSD estimation using Periodogram



time series data or signal

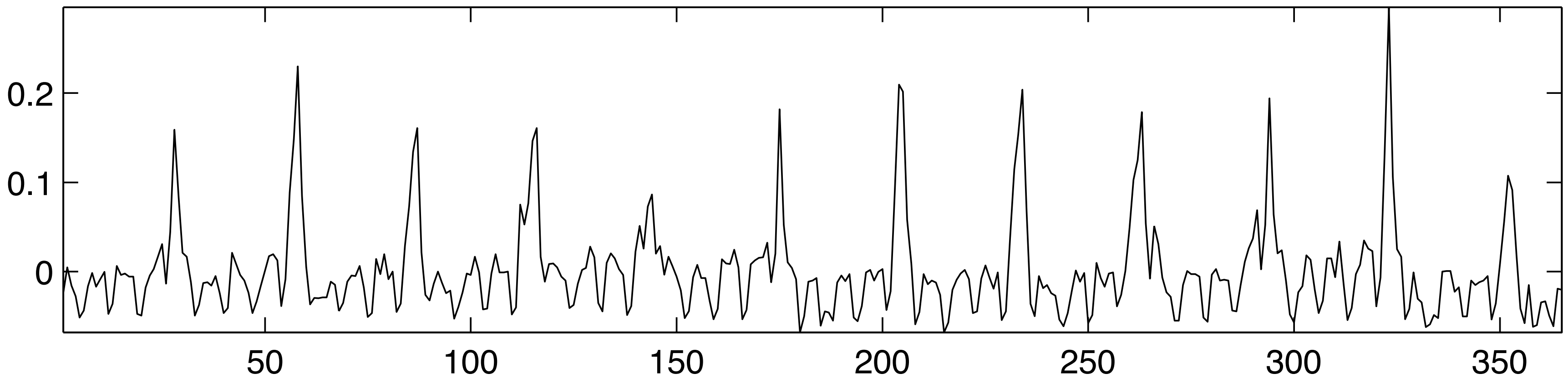
periodogram

$$\mathcal{P}(f_{k/N}) = \|X(f_{k/N})\|^2 \quad k = 0, 1 \ldots \lceil \tfrac{N-1}{2} \rceil$$

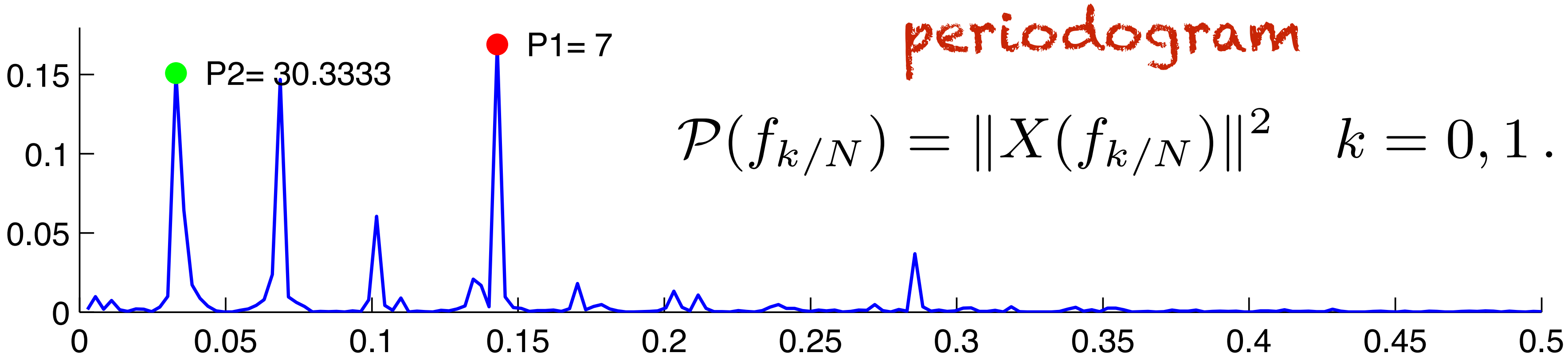


- To find the dominant frequencies choose the top-k dominant frequencies

Disadvantages of the Periodogram



time series data or signal



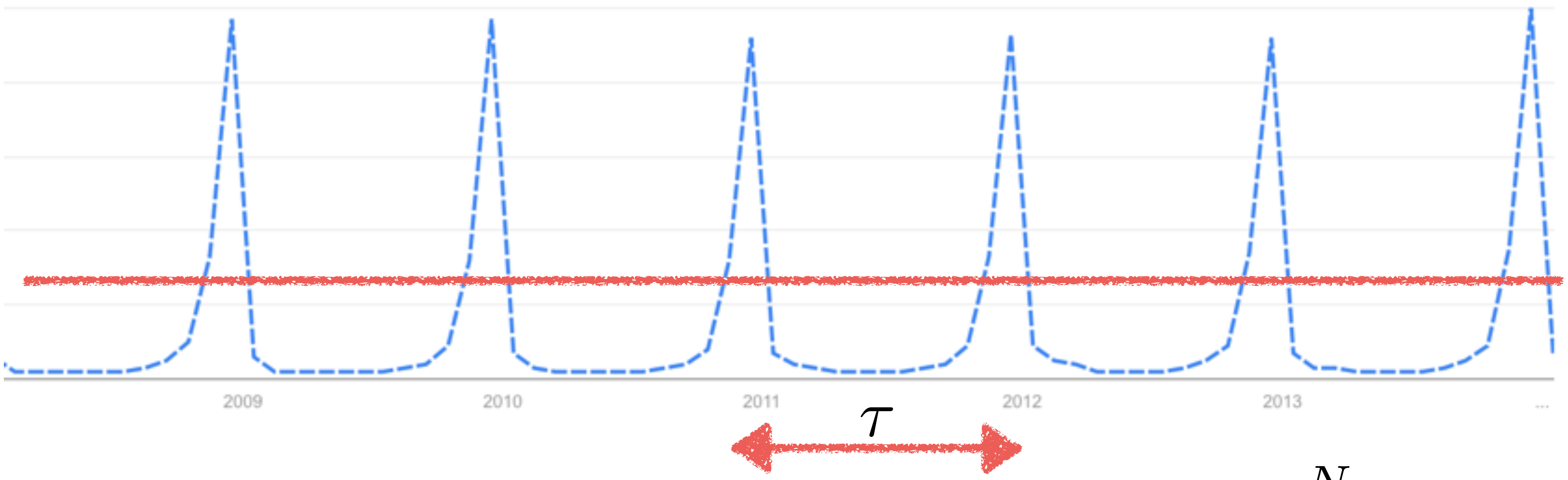
periodogram

$$\mathcal{P}(f_{k/N}) = \|X(f_{k/N})\|^2 \quad k = 0, 1 \ldots \lceil \frac{N-1}{2} \rceil$$

- Good only for short and medium periodicities
- Spectral leakage - frequencies not integral multiples of the DFT bin spread over other bins — false alarms

Autocorrelation

- Correlate the time series with itself

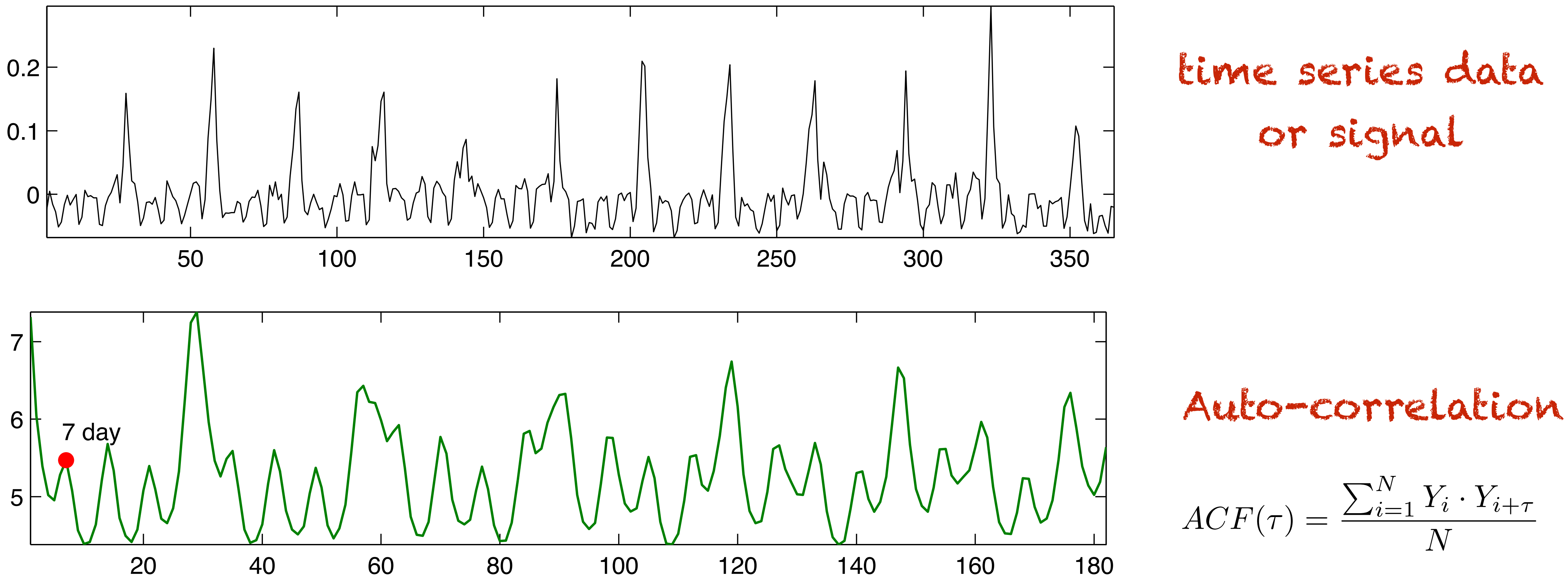


- Peaks get amplified

$$ACF(\tau) = \frac{\sum_{i=1}^{N} Y_i \cdot Y_{i+\tau}}{N}$$

- Fine-grained periodicity detector

Autocorrelation

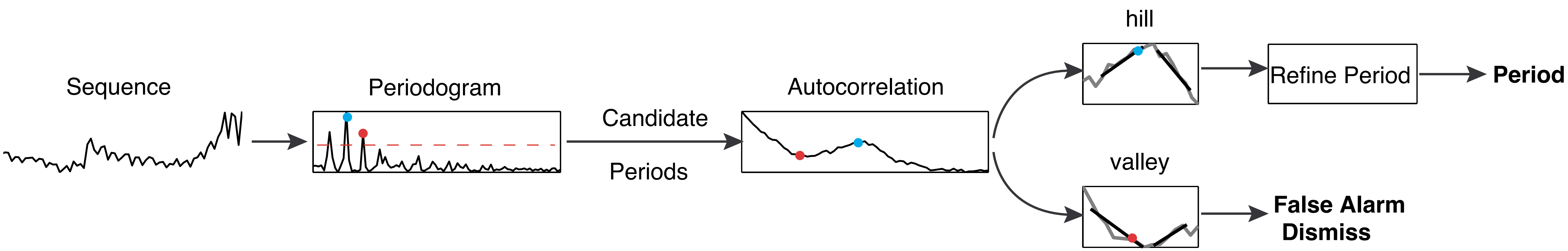


$$ACF(\tau) = \frac{\sum_{i=1}^{N} Y_i \cdot Y_{i+\tau}}{N}$$

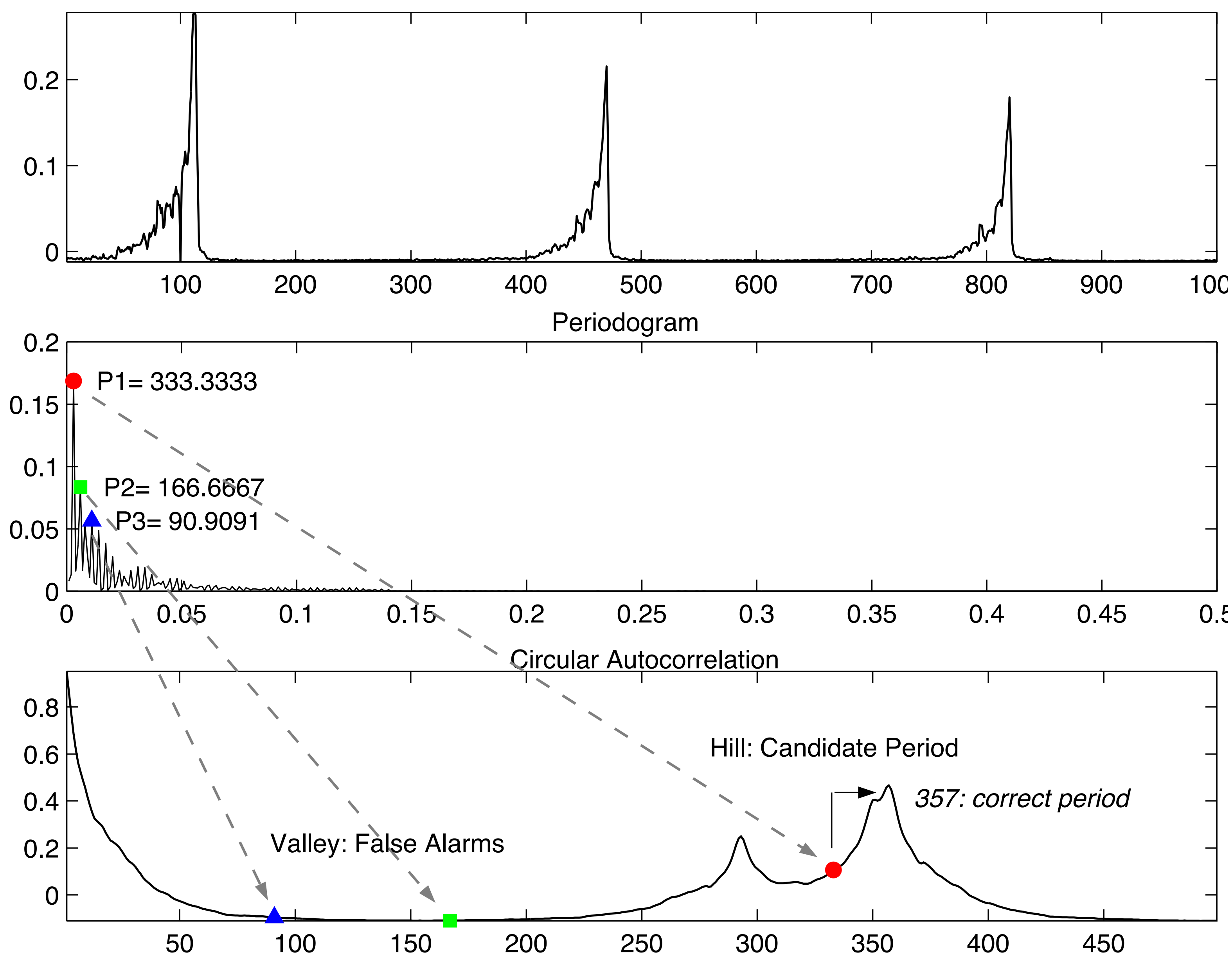
- To determine dominant period significance threshold needs to be specified
- Multiples of the same period are also peaks — needs post processing

Auto-Period

- **Auto-correlation :** Good for large periods but difficult to automatically determine periods
- **Periodogram :** Easy to threshold but not accurate for short periods
- **Idea:** Get candidate periods from Periodogram and validate false alarms using Auto-correlation



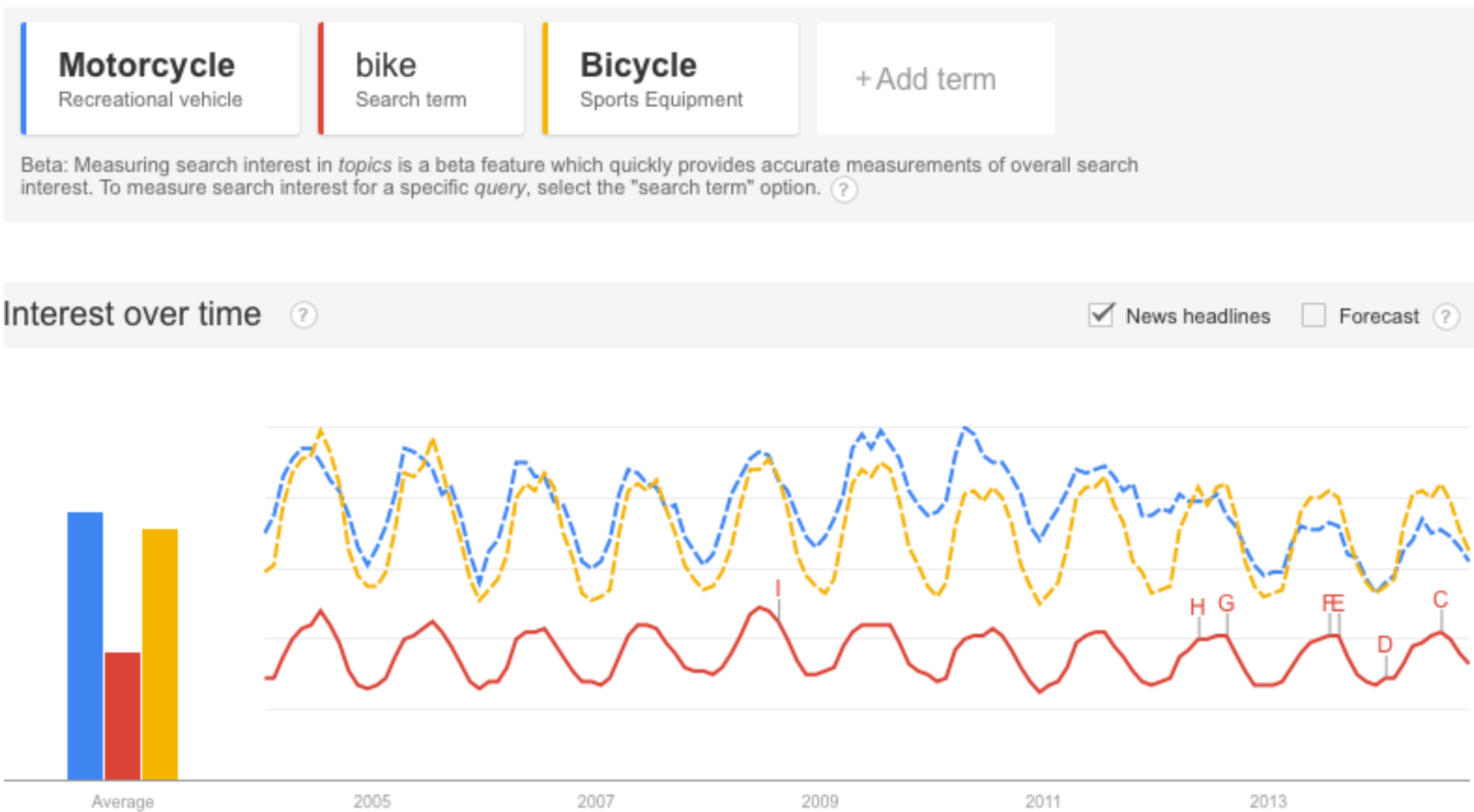
Auto-Period



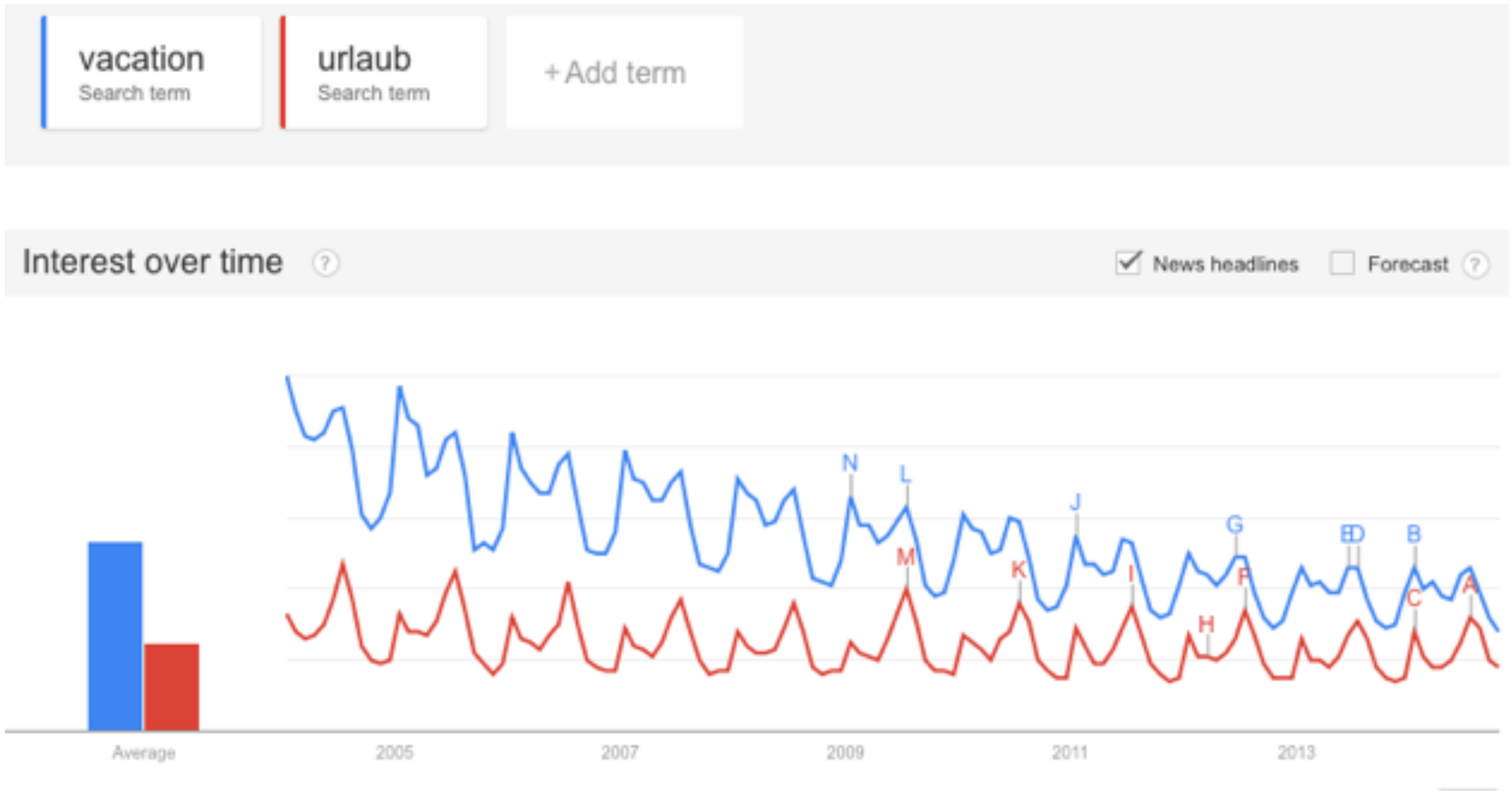
0.2
0.1
0
100
200
300
400
500
600
700
800
900
100
Periodogram
0.2
0.15
0.1
0.05
0
P1= 333.3333
P2= 166.6667
P3= 90.9091
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
Circular Autocorrelation
0.8
0.6
0.4
0.2
0
Hill: Candidate Period
357: correct period
Valley: False Alarms
50
100
150
200
250
300
350
400
450

Matching Time Series

- Similar time series suggest similar things



Matching Time Series



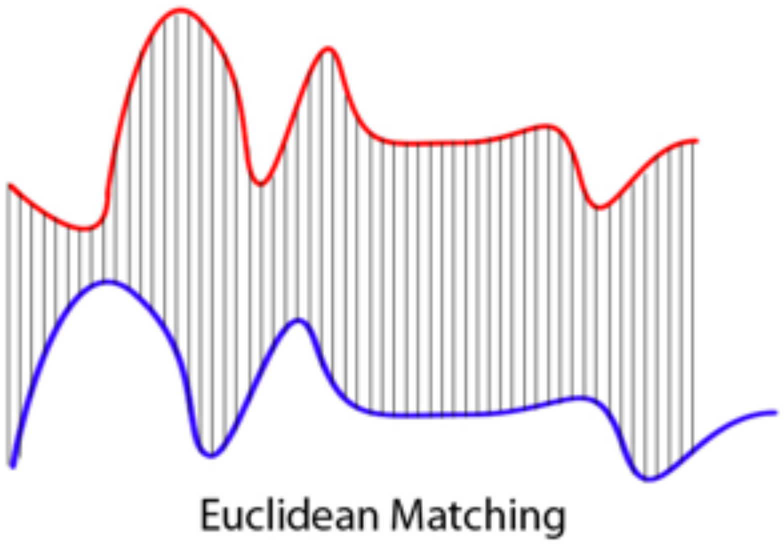
- Correlating time series used for clustering, classification, anomaly detection, speech recognition etc.

Matching Time Series

What measure would you use to match two time series ?

$$d = \sum_t |y_t - x_t|$$

Euclidean Distance



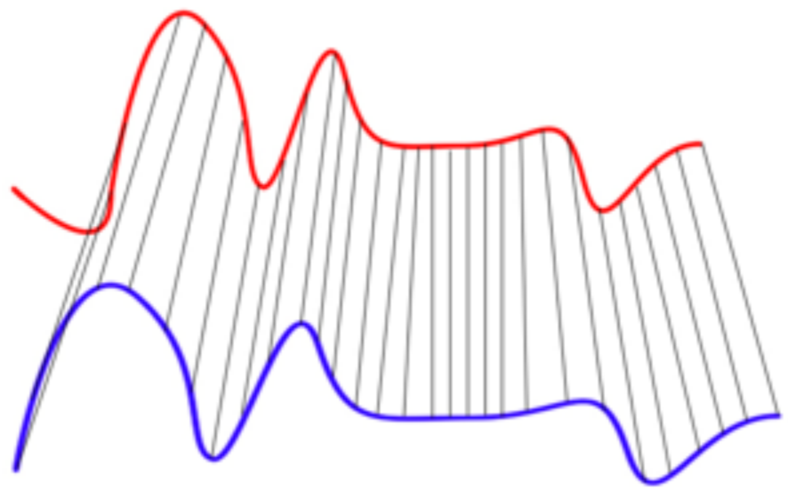
Why is Euclidean matching not good enough ?

Dynamic Time Warping

Time series might be shifted

Time series might be compressed at some point in time

Random noise at some points



Dynamic Time Warping Matching

Dynamic time warping measures the distance between two sequences under certain restrictions.

Not a metric. Triangle inequality doesn't hold

Detour - Edit Distance

- Edit distance measures how many steps it takes to convert a string to another based on restrictions

- Restrictions define cost function — insertion, deletion, replacement

	f	o	x
f	0	1	2
a	1	2	3
x	2	3	2

insertions and deletions

	f	o	x
f	0	1	2
a	1	1	2
x	2	2	1

insertions, deletions and replacements

Edit Distance

	f	o	x
f	0	1	2
a	1	2	3
x	2	3	2

insertions and deletions

	f	o	x
f	0	1	2
a	1	1	2
x	2	2	1

insertions, deletions and replacements

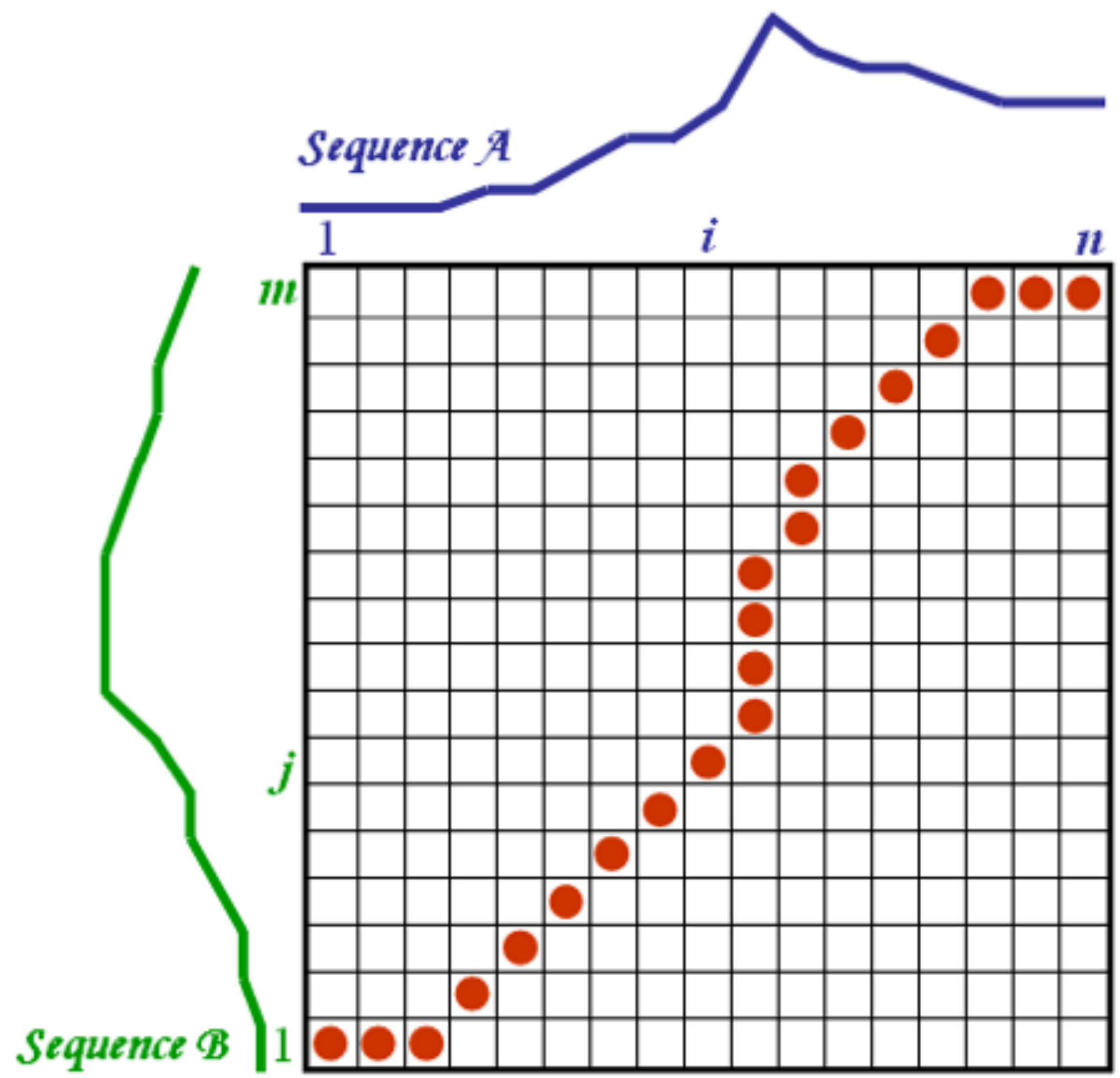
$$d_{ij} = \begin{cases} d_{i-1,j-1} & a_j = b_i \\ \min \begin{cases} d_{i-1,j} + w_{\mathrm{del}}(b_i) \longrightarrow 1 \\ d_{i,j-1} + w_{\mathrm{ins}}(a_j) \longrightarrow 1 \\ d_{i-1,j-1} + w_{\mathrm{sub}}(a_j, b_i) \longrightarrow 1 \text{ or } 2 \end{cases} & a_j \neq b_i \end{cases}, \quad for\ 1 \leq i \leq m, 1 \leq j \leq n.$$

Dynamic Time Warping

- DTW aligns two sequences of feature vectors by warping the time axis iteratively until an optimal match (according to a suitable metrics) between the two sequences is found.

$$x_1, x_2, \ldots x_n \qquad\qquad y_1, y_2, \ldots y_n$$

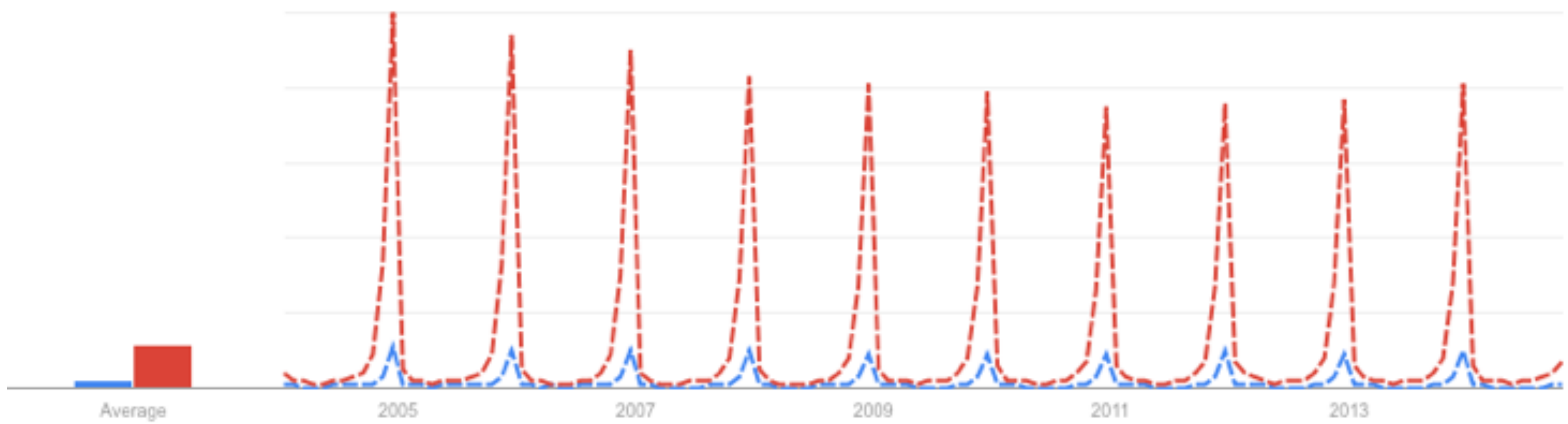
$$d(i,j) = c(i,j) + min \begin{cases} d(i-1,j), \\ d(i-1,j-1), \\ d(i,j-1) \end{cases}$$



Burst Detection

- Bursts are rare but extremely beneficial in time-series
- Used in number of applications
  - Twitter: Trending topics
  - Stock markets: Trending Stocks
  - Text Mining: finding important time periods
- Elastic burst detection:
  - Stream of data
  - Quadratic computations not allowed

Burst Detection



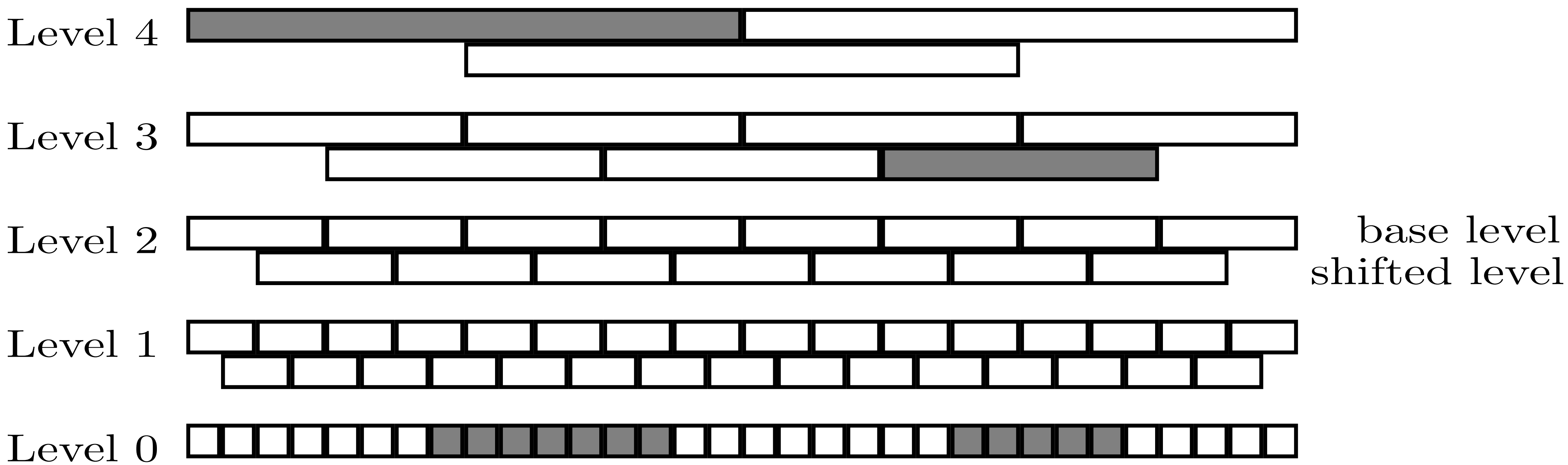
- Global Average
- Moving Average
- Damped Average

Elastic Burst Detection

- Given a time-series {xi}
- A set of window sizes W
- A monotonic, associative aggregation function A which maps a sequence of values to a number. E.g. Average, Max
- and Thresholds associated with each window size w, f(w)
- Find all pairs (t,w) such that t time a time point and w is a window size in W

$$A[x_t \cdots x_{t+w-1}] \geq f(w)$$

Burst Detection - Shifted Binary Tree



Whenever more than $f(2 + 2^{i-1})$ events are found in a window of size 2^{i+1}, then a detailed search must be performed to check if some subwindow of size w, $2+2^{i-1} \le w \le 1+2^{i}$, has $f(w)$ events.

Summary

- Periodicity of Events
  - Auto-correlation, Periodograms and their combinations
- Burst Detection Techniques and elastic detection
- Matching of time series
  - Euclidean matching
  - Dynamic Time Warping

References

- **“On Periodicity Detection and Structural Periodic Similarity”, 2005.**
    - Michail Vlachos, Philip Yu, Vittorio Castelli
- **Burst Detection in Hierarchical Streams.**
    - Jon Kleinberg
- **Everything you know about Dynamic Time Warping is wrong**
    - Eamonn Keogh

Projects

http://www.l3s.de/~anand/tir14/projects.html

- **Temporal and Phrase-based Indexing** - Avishek(anand@l3s.de)

- **Temporal Retrieval Models** - Jaspreet (singh@l3s.de)

- **Temporal Query Autocompletion**- Avishek

- **Crawling for Temporal Collections** - Gerhard (gossen@l3s.de)

- **Temporal Query Suggestions** - Helge (holzmann@l3s.de)