GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 21:TIME SERIES MODELING

Key Concepts - What is a Time Series

- A time series is just a sequence of points where each point has an associated timestamp, which usually reflects the time of measurement

- There is an ordering, based upon time

- In general analysis is conducted when the points are equally spaced in time

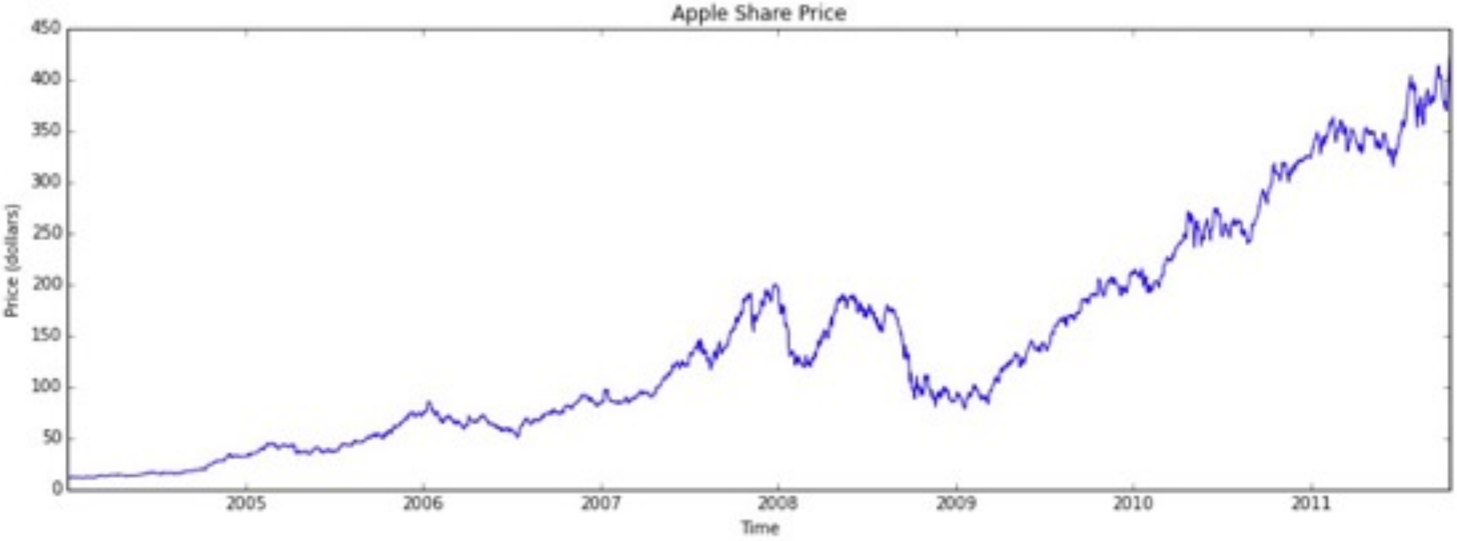
KEY CONCEPTS - TIME SERIES ANALYSIS

- Time series analysis consists of:
  - Descriptive - The extraction of meaningful statistics and characteristics from the time series
  - Forecasting - using a predictive model to predict future values of the time series based upon previous values
  - Classification - time series characteristics are used to identify an underlying state

- Time series analysis has significant overlap with the field of signal processing or digital signal processing (DSP)

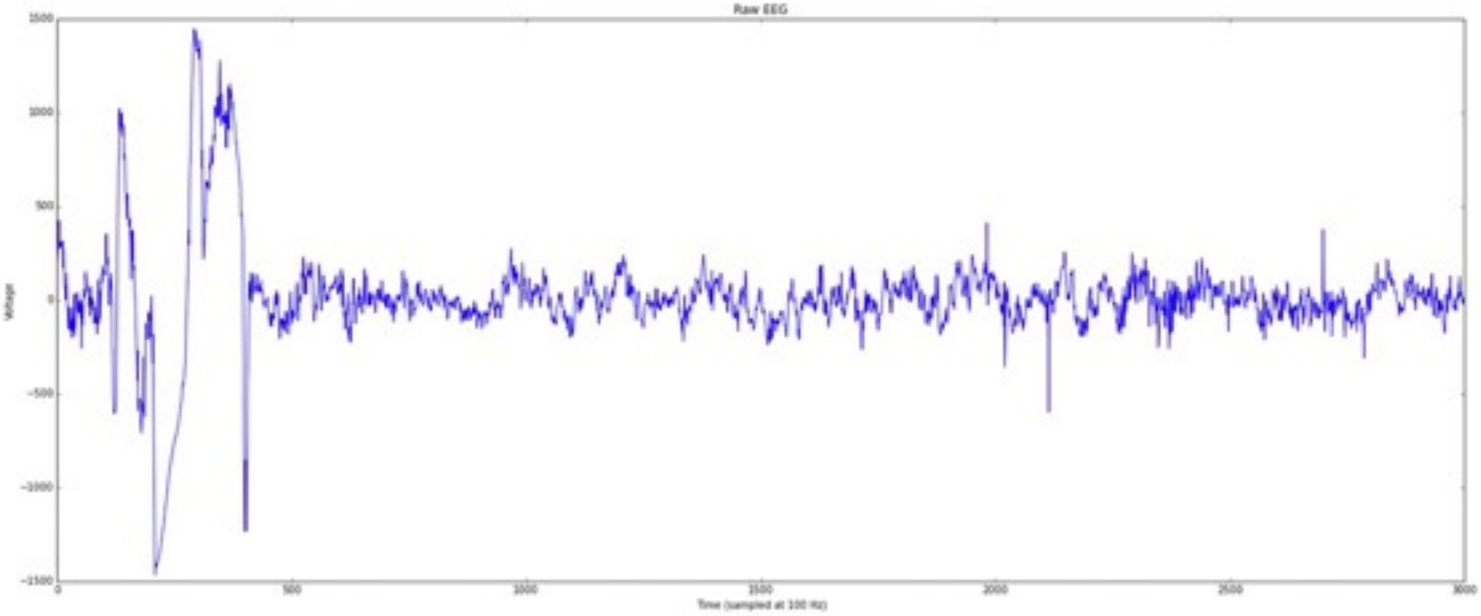
KEY CONCEPTS - TIME SERIES ANALYSIS - EXAMPLES

- Financial data



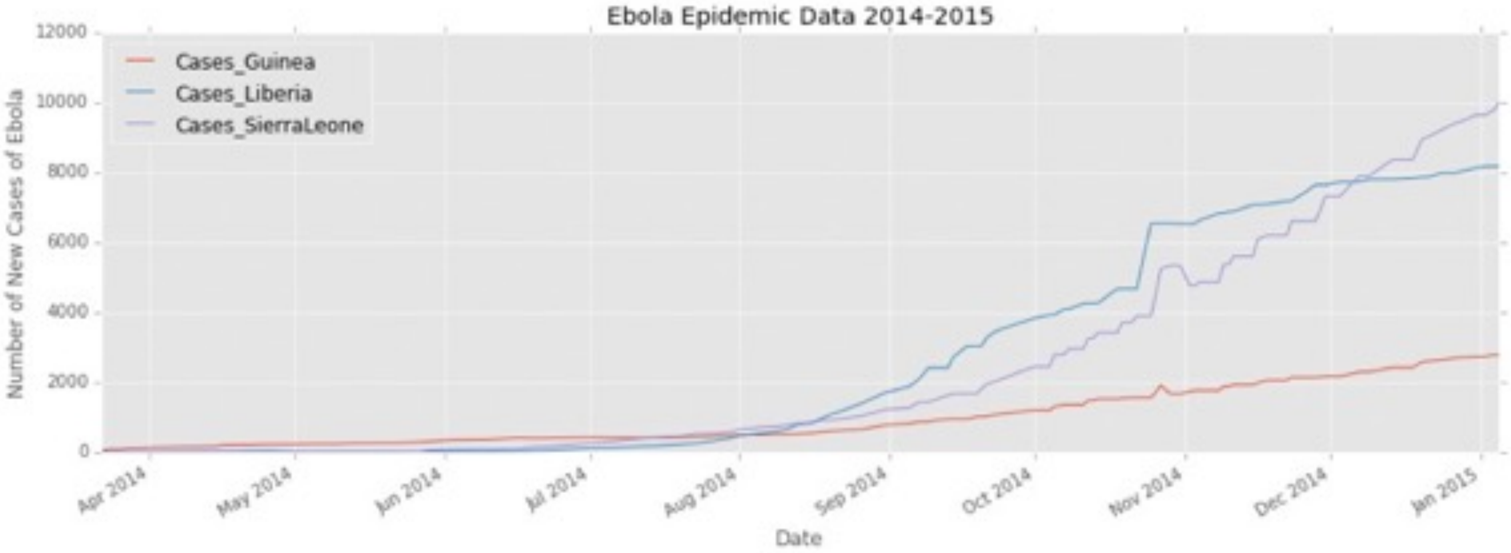
KEY CONCEPTS - TIME SERIES ANALYSIS - EXAMPLES

- Neurophysiological Data



KEY CONCEPTS - TIME SERIES ANALYSIS - EXAMPLES

- Epidemiological Data



Key Concepts - Time Series Analysis - Examples

What all of these have in common is the x-axis represents a time sequence

BASIC SIGNAL PROCESSING

KEY CONCEPTS - SAMPLING

“Real World” Signals

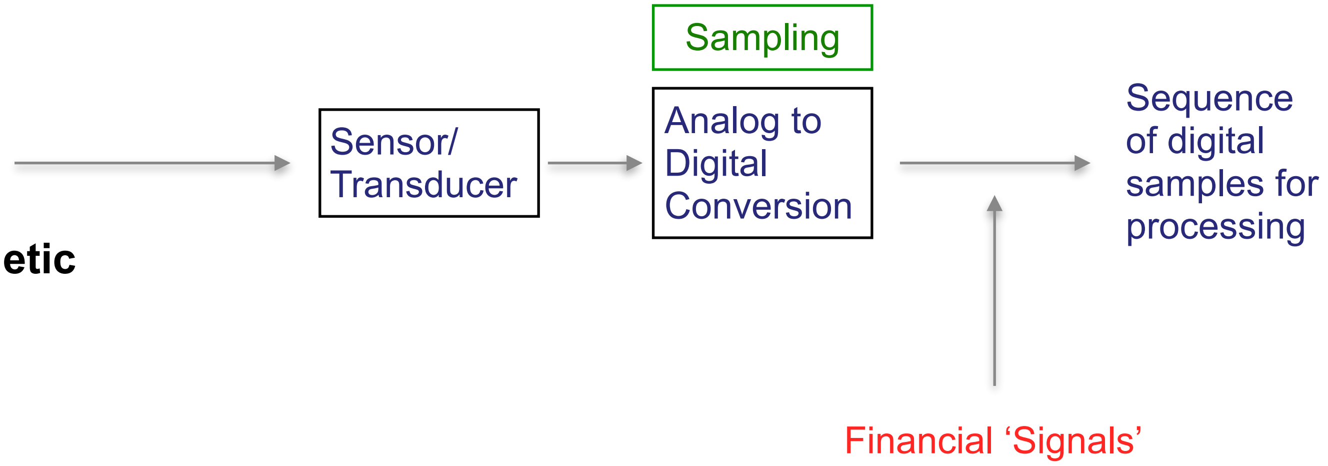
Pressure:
- Speech
- Music
- SONAR

Bio-electric:
- EEG
- EKG

Electromagnetic
- Radio
- RADAR

Images
- Camera
- XRAY, Ultrasound, MRI

Other
- Seismic



Sensor/ Transducer

Sampling

Analog to Digital Conversion

Sequence of digital samples for processing

Financial ‘Signals’

KEY CONCEPTS - TIME SERIES

- Financial and epidemiological data are not like typical digital signals

- Financial data is aggregated based upon a time period of interest
  - daily stock prices
  - but not really daily, we don’t trade on weekends and holidays
  - we also don’t trade a regular, specified intervals

- Trading activity generates data, which is then aggregated into regularly spaced time intervals dictated by the user
  - the ‘zoom’ factor very much affects the time series you see

KEY CONCEPTS - TIME SERIES

- The signals generated from electronics, however, generate regularly spaced numbers representing a ‘signal’

- The sensor in a digital camera counts photons.

- The number of photons counted in each pixel of the sensor becomes a representation of the blackness or whiteness of that pixel

- When recording movement (i.e. as a movie) the capture rate is at a steady 15, or 30 frames (images) per second

Key Concepts - Nyqvist - Shannon Sampling Theorem

- Sampling Circuits or Analog to Digital Converters effectively convert electrical voltage waveform into a number
  - Sampling is done at a fixed interval
  - Sampling frequency is the number times this process occurs every second - i.e. how often do we sample the signal

- The accuracy of the sampling depends on the number of bits used in the A/D converter, and the sampling frequency (f_S)

- The aim of sampling is to capture the signal, which means capturing all the component sinusoids (if you have sampled the signal properly it should be possible for you to, almost, reconstruct the signal using the samples)

KEY CONCEPTS - TIME SERIES

Sinusoids

One key to understanding time series analysis/DSP

KEY CONCEPTS - BASIC SIGNAL PROCESSING

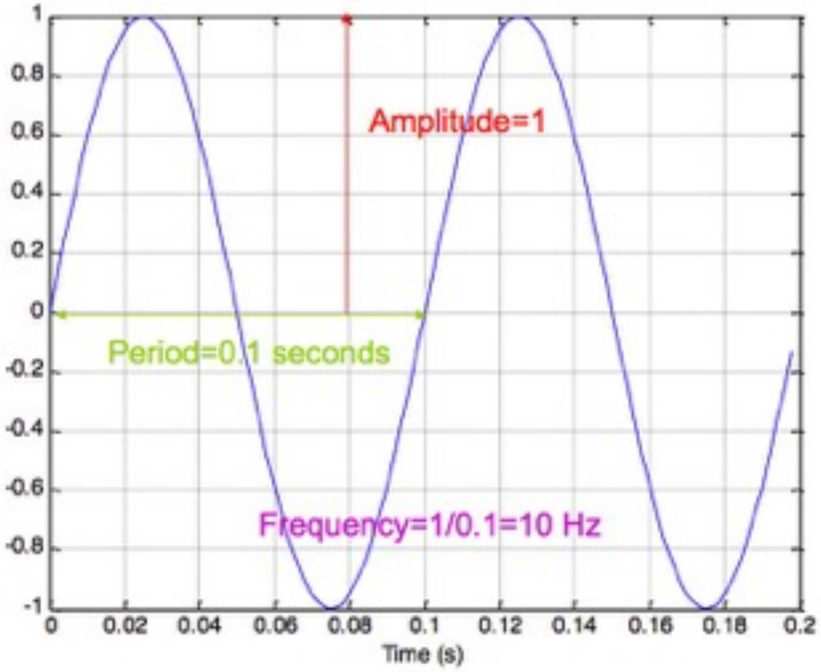
- Consider a simple sinusoid in the Time domain

- 3 basic parameters that define it's behavior:

1. Amplitude

2. Wavelength/Frequency
   - F = 1/W

3. Phase



KEY CONCEPTS - BASIC SIGNAL PROCESSING - PHASE

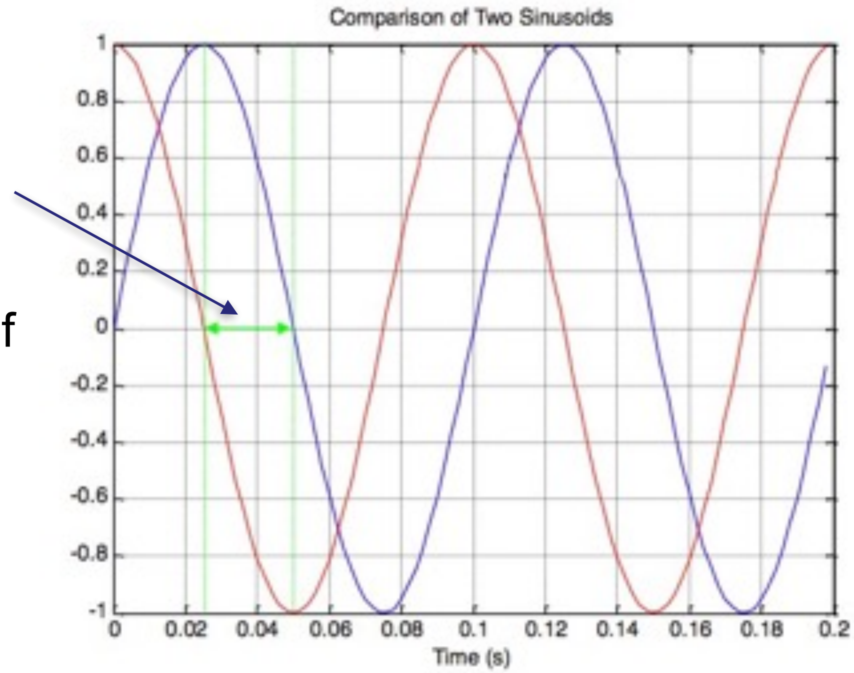
- Phase refers to ‘where’ the signal is, in terms of it’s cycle

- Here are 2 sinusoids separated by 0.025 seconds

- Since each has a period or wavelength of 0.1 seconds phase can be measured

- A single oscillation is considered 360 degrees (or 2π radians)

- Phase = 0.025/0.1   * 360(degrees)
  = 90 degrees



KEY CONCEPTS - WHY IS THE STUDY OF SINUSOIDS RELEVANT?

- Principle of Superposition
  - Any real-world time series can be constructed by adding together a combination of individual sinusoids of the correct frequencies and phases
  - A signal is merely the sum of it’s component simple sinusoidal waves!
  - The converse applies, in that we can break any real-world signal into it’s component sinusoids!

KEY CONCEPTS - WHY IS THE STUDY OF SINUSOIDS RELEVANT?

- There are, of course, some assumptions:
  - the complex signal must have a period, or an oscillation

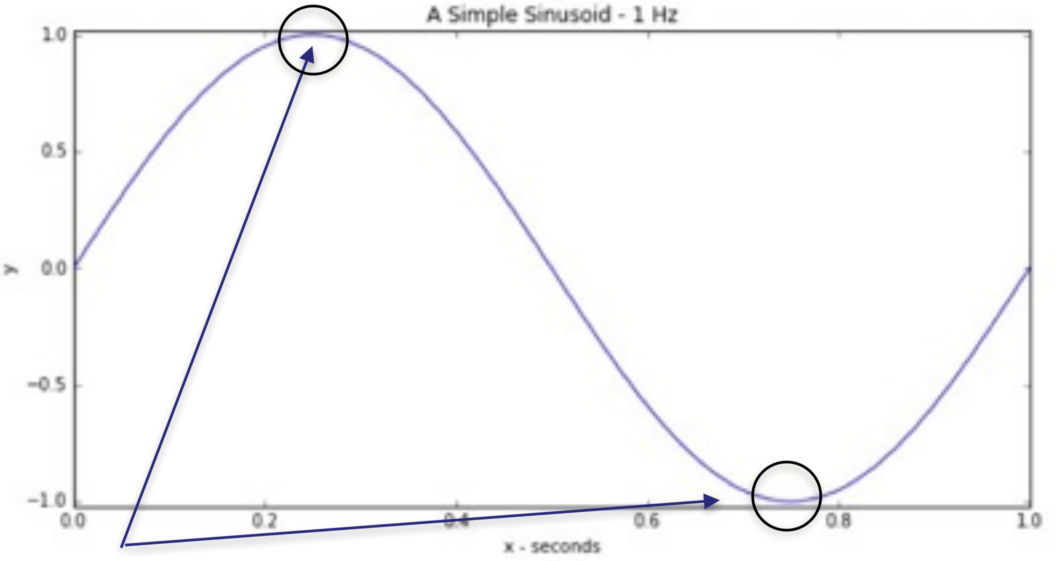
KEY CONCEPTS - NYQUIST - SHANNON SAMPLING THEOREM

- fs MUST be at least twice the highest frequency sinusoid present in the signal you are trying to capture

Here is a 1 Hz sinusoid

How many samples do I need to take such that I can reconstruct the signal exactly?

Answer: 2



You must capture these 2 points in order to recreate the sinusoid

KEY CONCEPTS - NYQUIST - SHANNON SAMPLING THEOREM

- If you don’t meet the Nyquist criterion and sample at a lesser rate then your samples will not accurately represent the original signal
- This is referred to as aliasing
- Once sampling has occurred you end up with a vector of numbers, which are the samples
- This vector is ordered in time
- The sequence of samples is equally spaced
- You end up with a time series

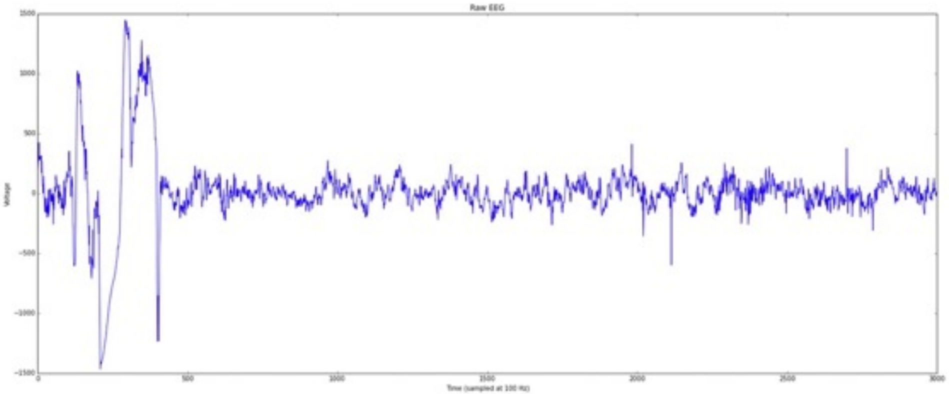
KEY CONCEPTS - SAMPLING - EXAMPLES

- Here is our bio-electric signal

- It has been sampled at 100Hz, or 100 times a second

- The maximum frequency that can be truthfully captured, is therefore, 50 Hz

- This is fine for EEG as we know that brain activity operates at frequencies of between 0.5 and 30Hz



Key Concepts - Descriptive Analysis of Time Series

Things to observe in a time series:

- Trends
  - Increase or decrease over time

- Periodicity or Frequency
  - Is there an observable cycle, e.g. seasonal variation

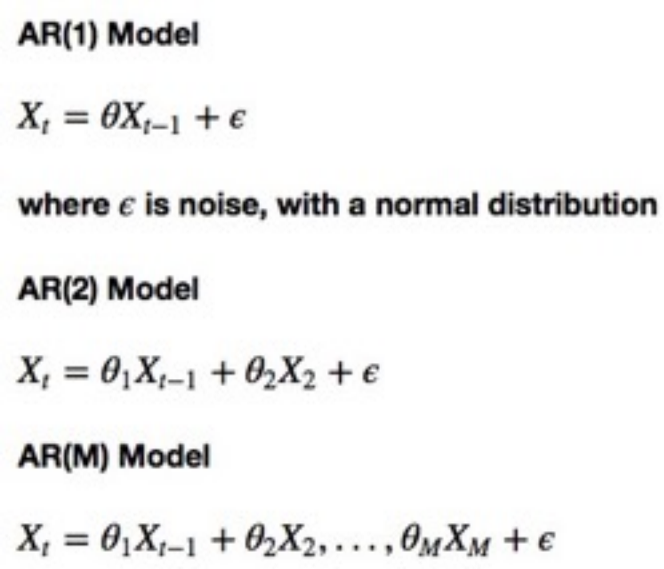
- Relationships (correlation) to other time series

The ipython notebook Introduction to Pandas Date and Time tools provides a brief summary of some of the date time functionality that Pandas provides

See the ipython notebook on Time Series Basics for examples of measuring trends

KEY CONCEPTS - TIME SERIES FORECAST PREDICTION

The Auto-regressive model



AR(1) Model

$$X_t = \theta X_{t-1} + \epsilon$$

where ϵ is noise, with a normal distribution

AR(2) Model

$$X_t = \theta_1 X_{t-1} + \theta_2 X_2 + \epsilon$$

AR(M) Model

$$X_t = \theta_1 X_{t-1} + \theta_2 X_2, \ldots, \theta_M X_M + \epsilon$$

KEY CONCEPTS - TIME SERIES FORECAST PREDICTION

- The Auto-regressive model is based upon linear regression

- Therefore, to model the time series well the linear regression assumptions become important

We need:

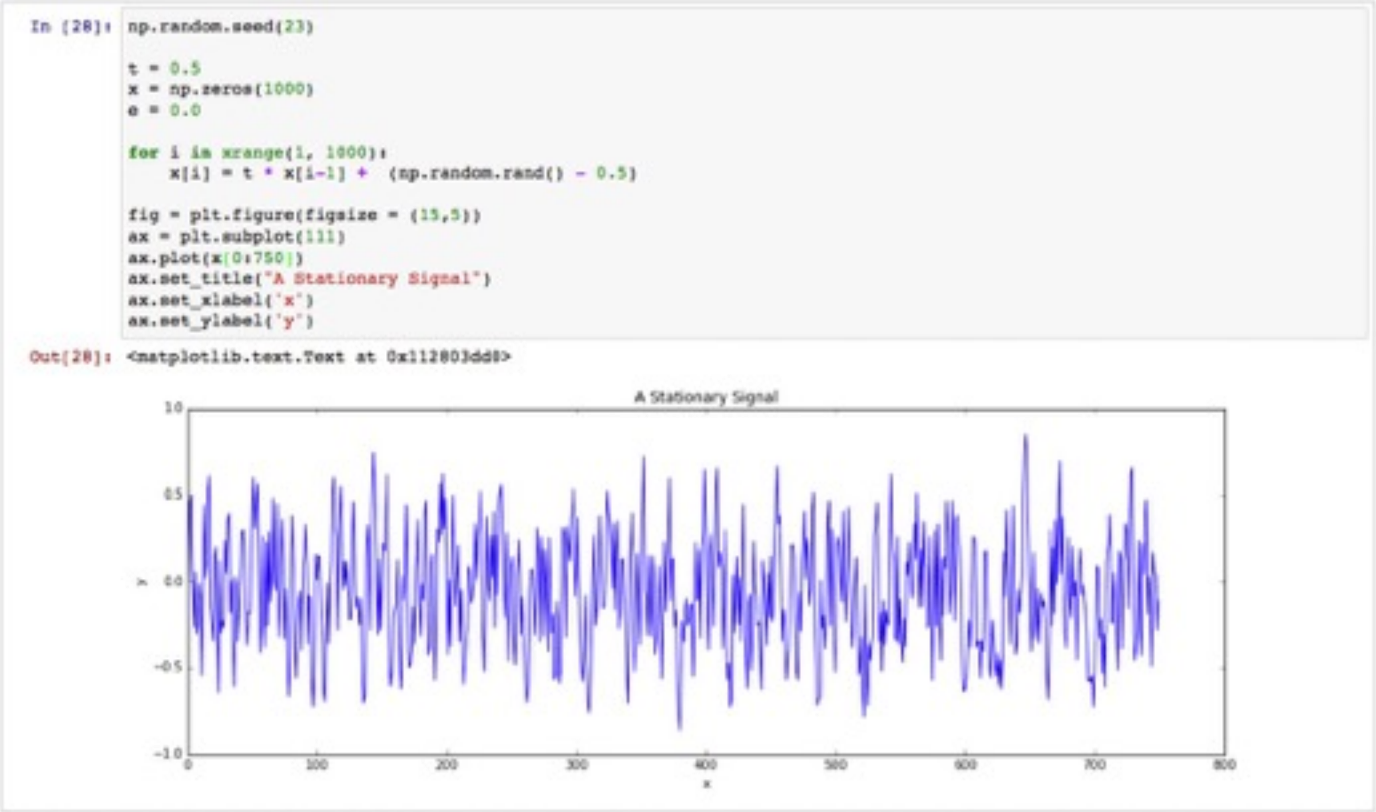
1. Linear relationship between the output and the features
2. The features are identically distributed
3. The features are independent of each other
4. The residual errors should be uncorrelated
5. The variance in the residuals is ‘constant’ - the residuals exhibit homoscedasticity

KEY CONCEPTS - TIME SERIES FORECAST PREDICTION

- This leads into the concept of signal stationarity
- Stationarity means the statistics of the signal do not change over time
- White noise is a stationary process
- A cymbal clashing is not, because after the initial power of the clash the noise diminishes over time

KEY CONCEPTS - TIME SERIES STATIONARITY

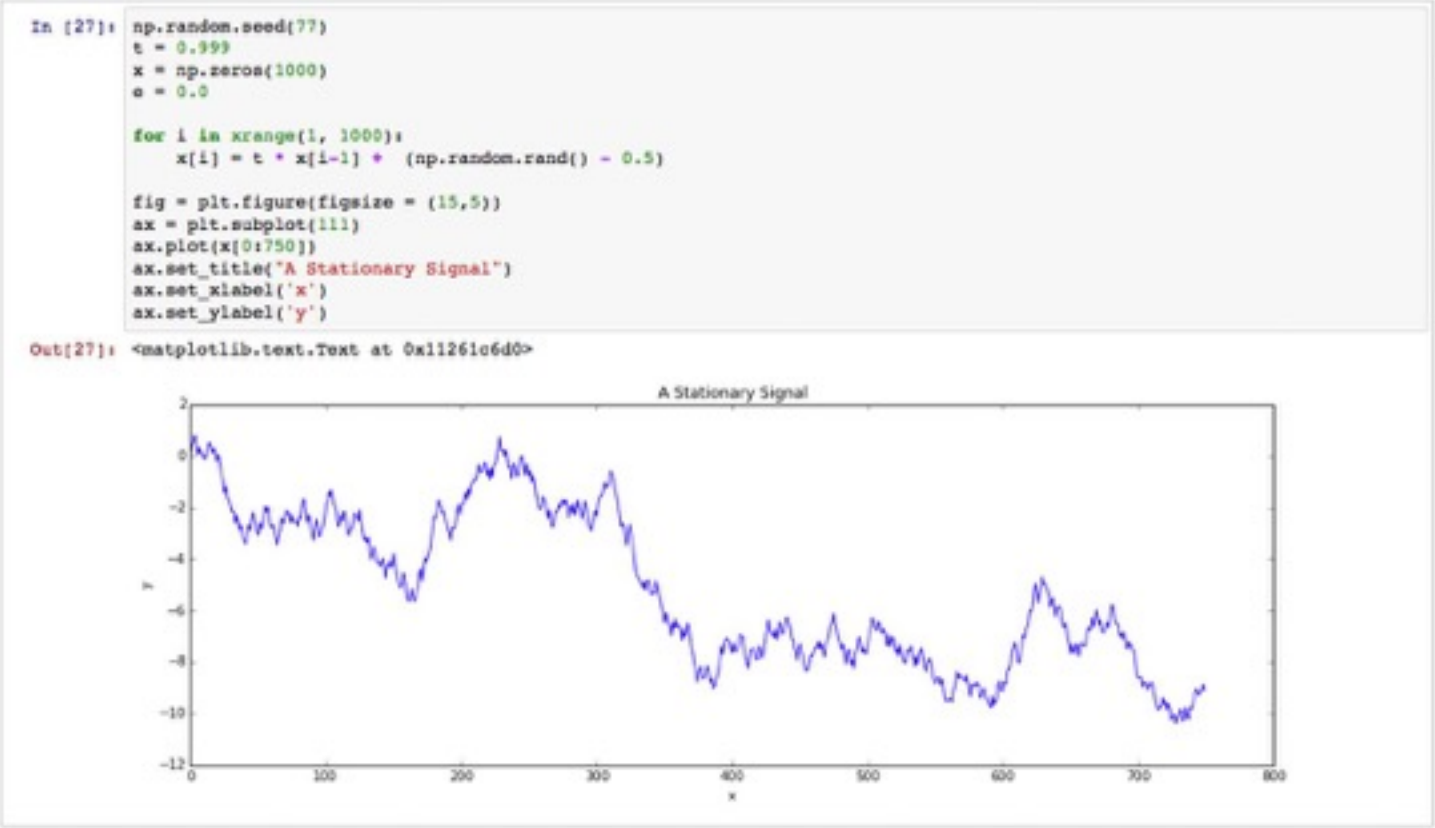
Generated by using an AR(1) model with theta = 0.5



```
In [28]: np.random.seed(23)

         t = 0.5
         x = np.zeros(1000)
         e = 0.0

         for i in xrange(1, 1000):
             x[i] = t * x[i-1] +  (np.random.rand() - 0.5)

         fig = plt.figure(figsize = (15,5))
         ax = plt.subplot(111)
         ax.plot(x[0:750])
         ax.set_title("A Stationary Signal")
         ax.set_xlabel('x')
         ax.set_ylabel('y')

Out[28]: <matplotlib.text.Text at 0x112803dd8>
```

KEY CONCEPTS - TIME SERIES STATIONARITY

Generated by using an AR(1) model with theta = 0.999



```
In [27]: np.random.seed(77)
         t = 0.999
         x = np.zeros(1000)
         e = 0.0

         for i in xrange(1, 1000):
             x[i] = t * x[i-1] +  (np.random.rand() - 0.5)

         fig = plt.figure(figsize = (15,5))
         ax = plt.subplot(111)
         ax.plot(x[0:750])
         ax.set_title("A Stationary Signal")
         ax.set_xlabel('x')
         ax.set_ylabel('y')

Out[27]: <matplotlib.text.Text at 0x11261c6d0>
```

Key Concepts - Time Series Stationarity

- You can still analyze such signals, but you may need to de-trend, or zero-mean
- scipy signal processing library has de-trending algorithms
- See the ipython notebook on Time Series Modeling for a very light introduction to AR modeling
- statsmodels python library is the place to go for AR and more complicated time series modeling and tools

KEY CONCEPTS - TIME SERIES CLASSIFICATION

- In this scenario we associate properties of a ‘window’ of the time series with a known underlying state that is responsible for generating the time series

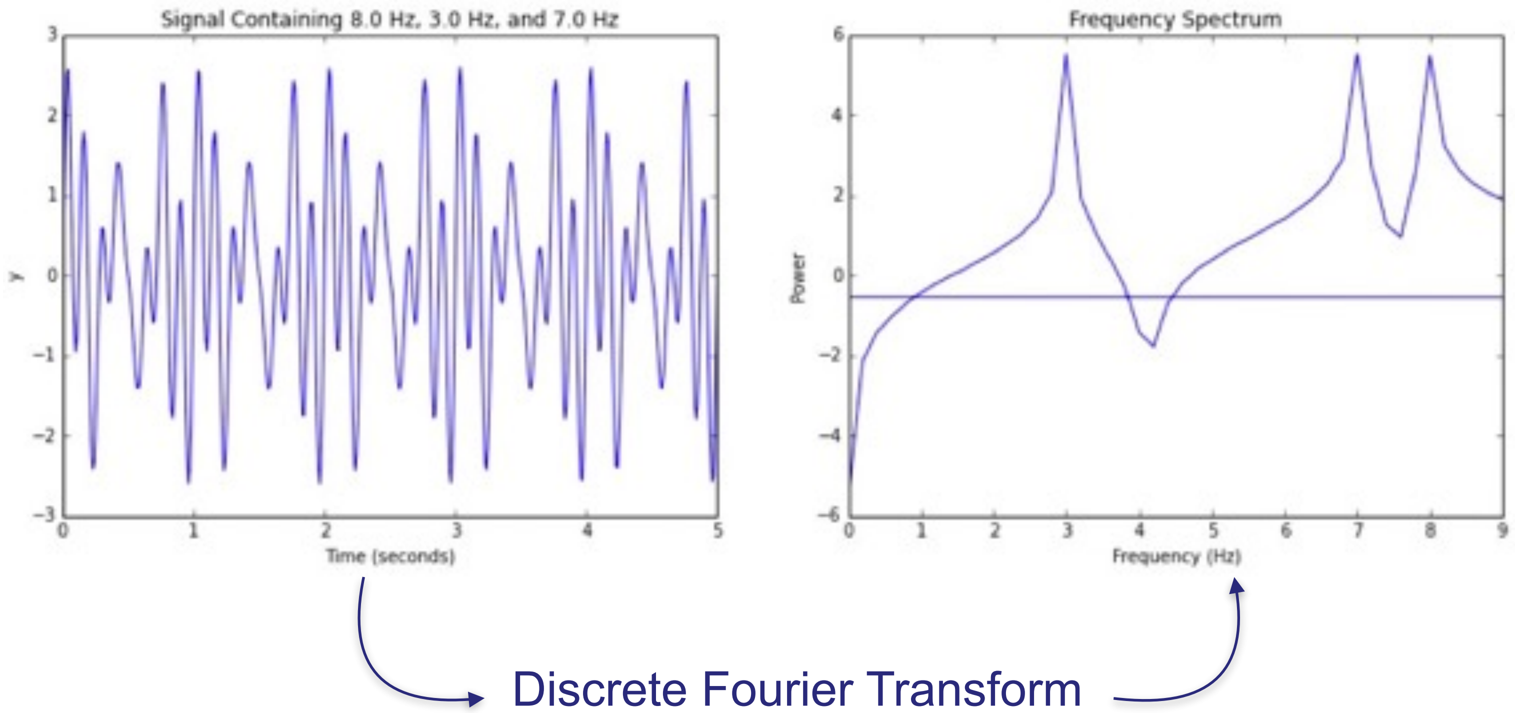
- Often the signal will need filtering, in order to reduce noise and fix baseline wander secondary to sensor drift

- scipy signal processing library is good place to look at filtering

- Generally move to the frequency domain to do this kind of work

See the ipython notebook on EEG Time Series for an example

KEY CONCEPTS - TIME DOMAIN VS FREQUENCY DOMAIN



Discrete Fourier Transform

KEY CONCEPTS - ADVANCED TOPICS

Scipy Signal Processing Library

http://docs.scipy.org/doc/scipy/reference/signal.html

If you are really interested in getting into AR modeling head to stats models

http://statsmodels.sourceforge.net/