Overview of Time Series Analysis

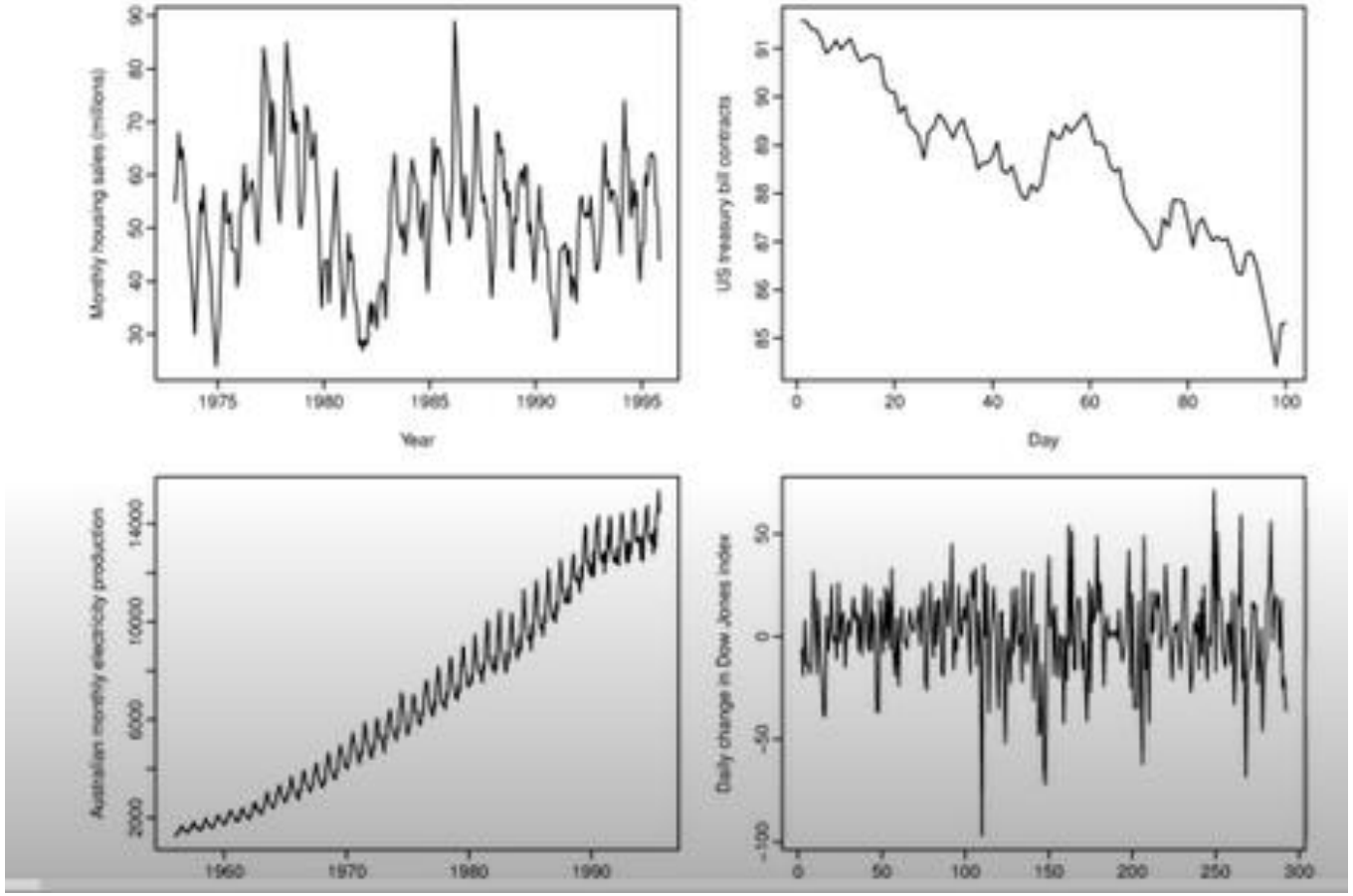
by Alex Dance

Data Time series can broken out to

- Yearly
- Monthly
- Weekly
- Daily
- Hourly
- By Minute

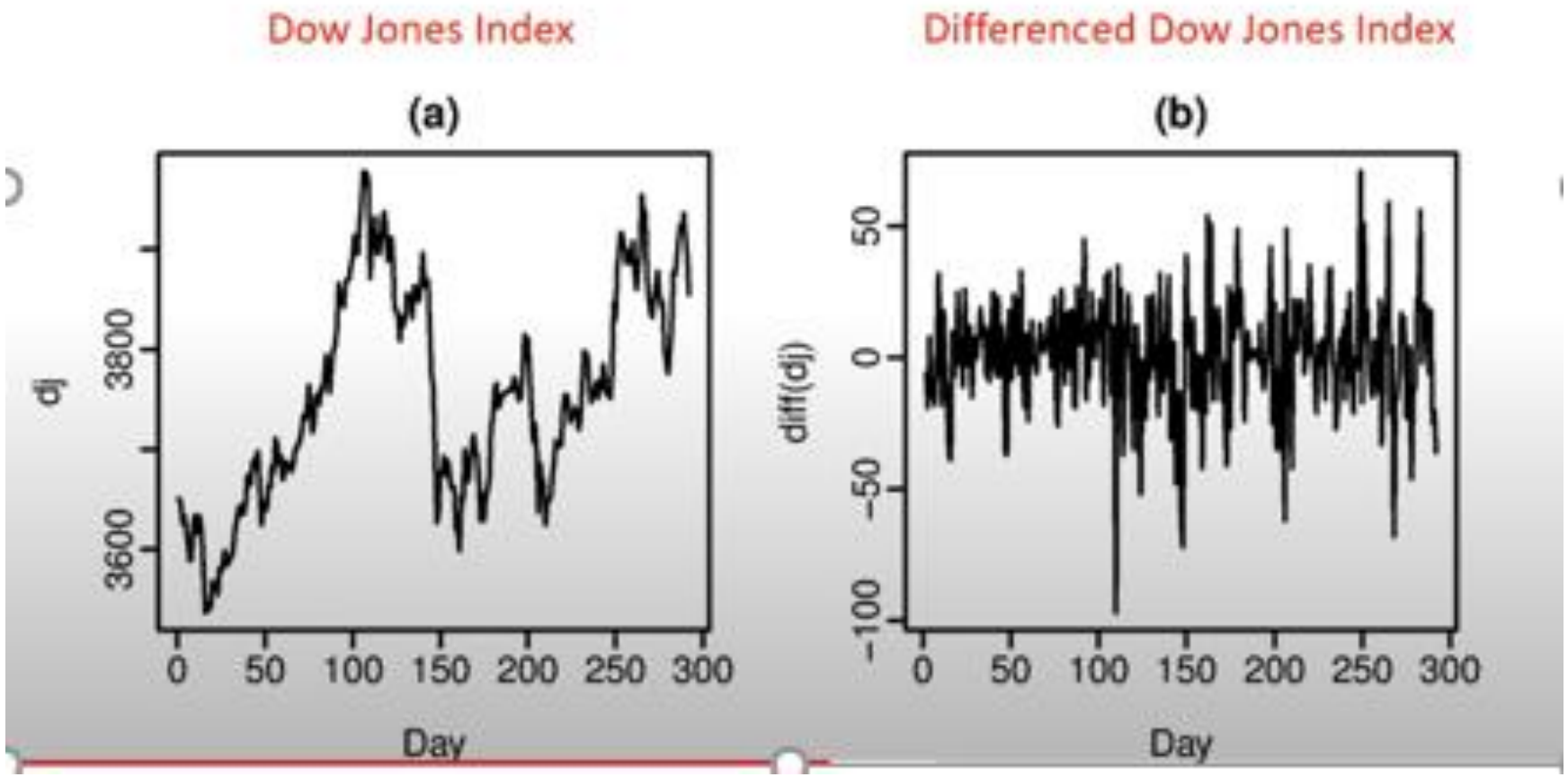
And lesser + greater time frames

Time series Definition: An ordered sequence of values and variables equally spaced over time

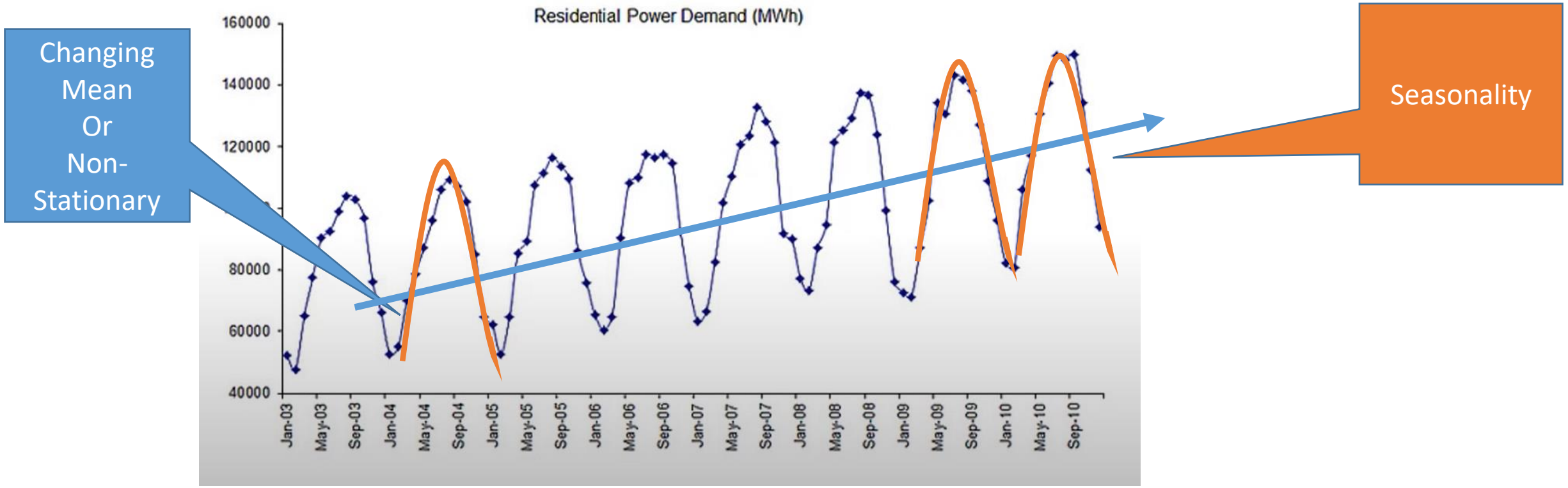


Differences over time can be a better way of looking at the data

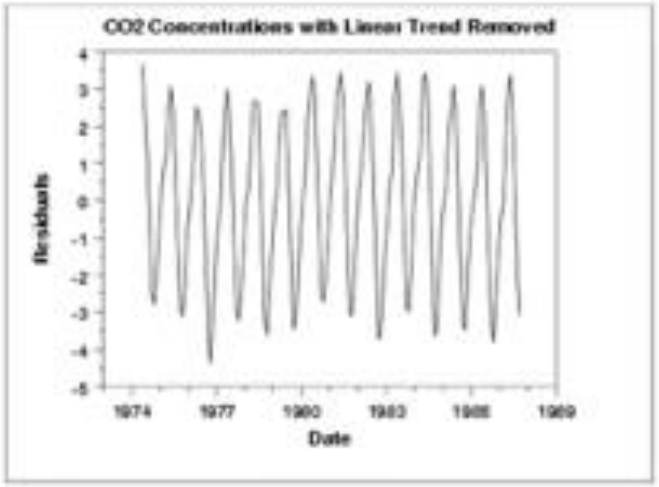
The first difference of a time series is the series of changes from one period to the next. If Y_t denotes the value of the time series **Y** at period **t**, then the first difference of **Y** at period **t** is equal to $Y_t - Y_{t-1}$

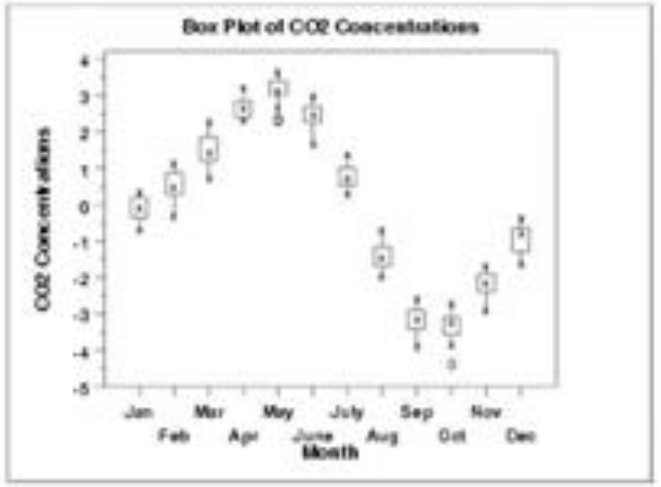


We can predict easier by eliminating factors: 2 of which are:



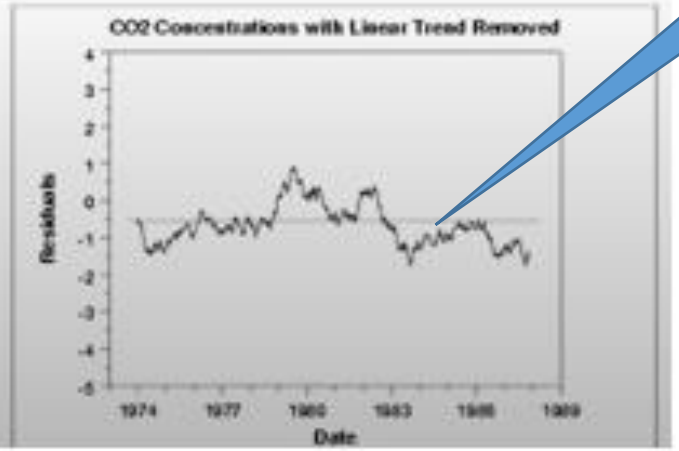
We can subtract seasonal or periodic trend from the data





We might get something that looks like this:

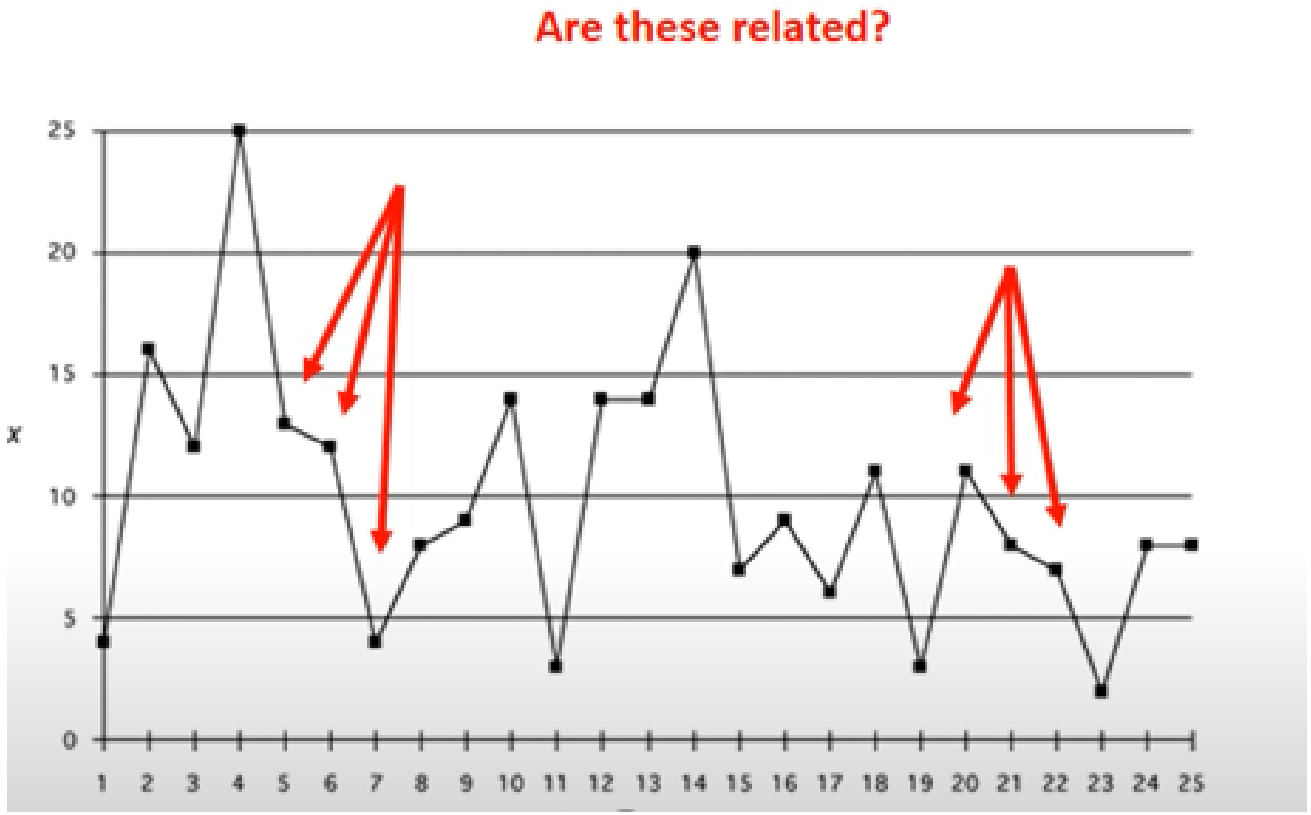
Is this a random process? (can we replicate it by just sampling from a known distribution?)



Not entirely random

Auto correlation and trends over time

There can be trends between things over time. The value of x(t) can depend on the value of x(t-1)

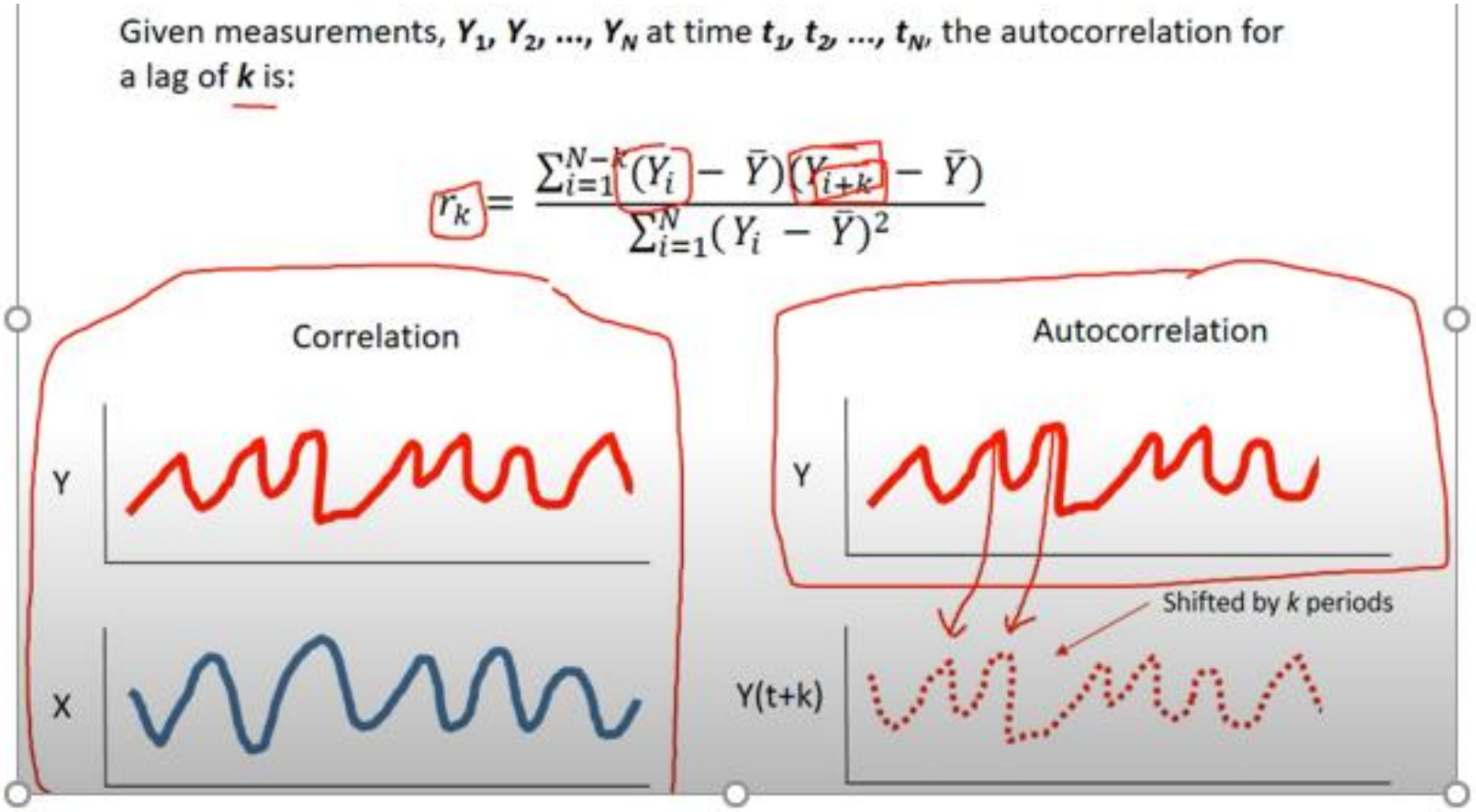


Autocorrelation is the “memory” of past time frames

The autocorrelation function can be used for the following two purposes:

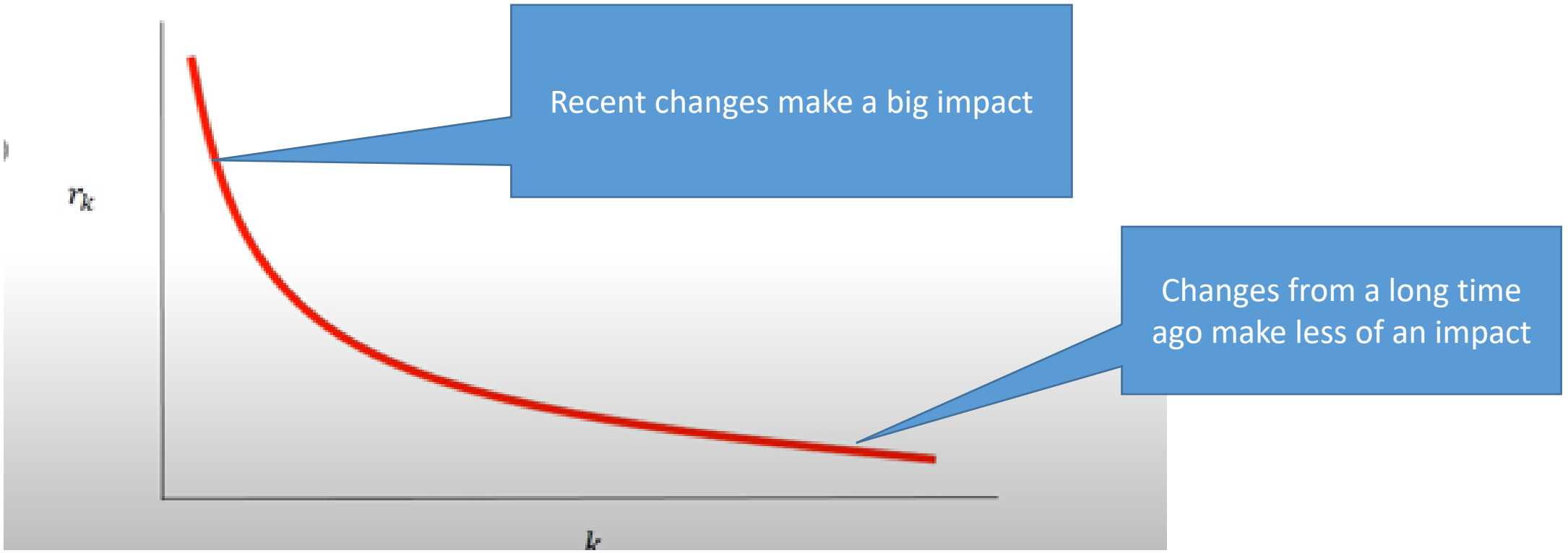
- To detect non-randomness in data
- To identify an appropriate time series model if the data are not random

This different “lags” (shifts in time) can be measured

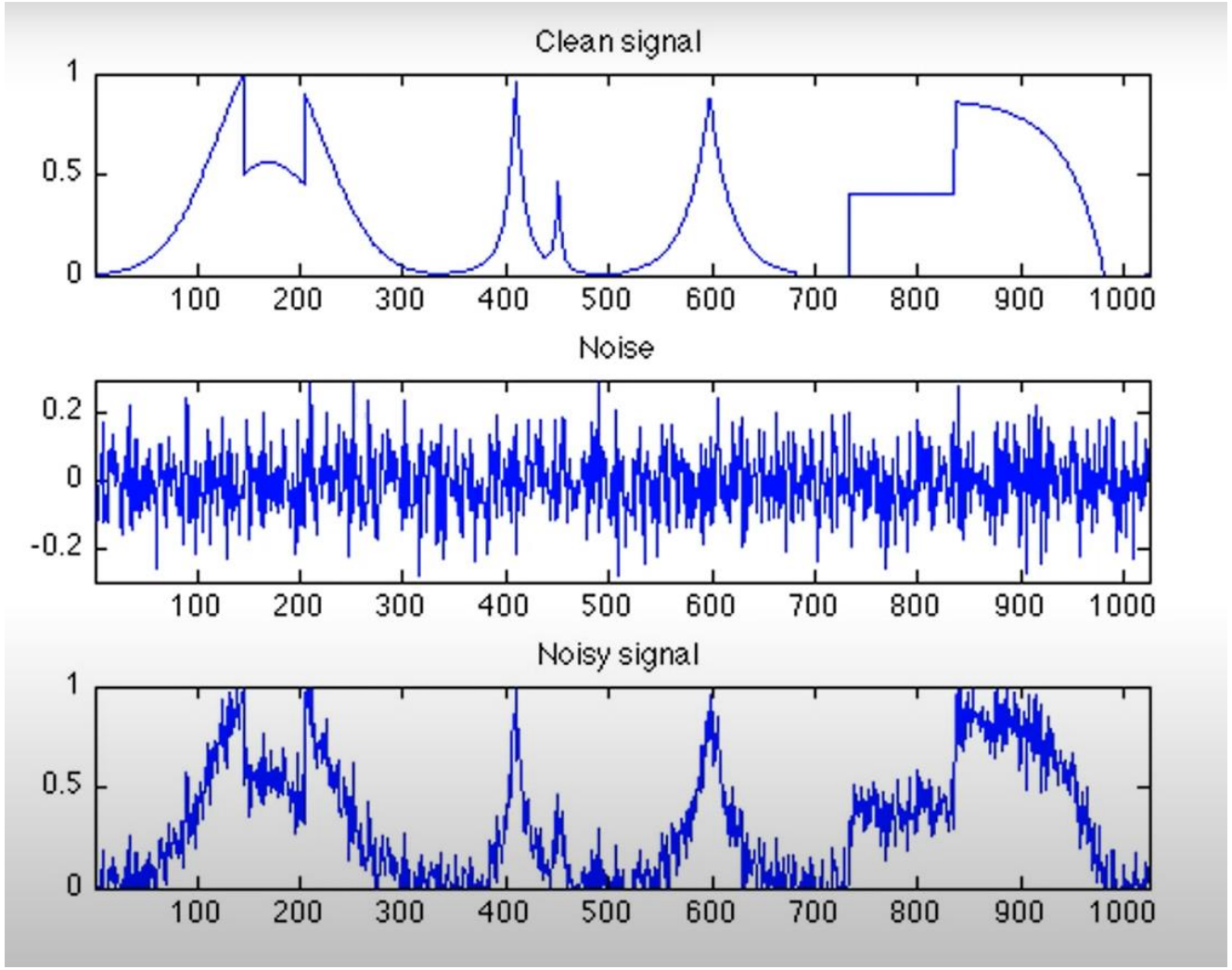


The difference in time periods can be measured and plotted

What does this tell us?



There are noise factors to consider.
Noise is randomness.

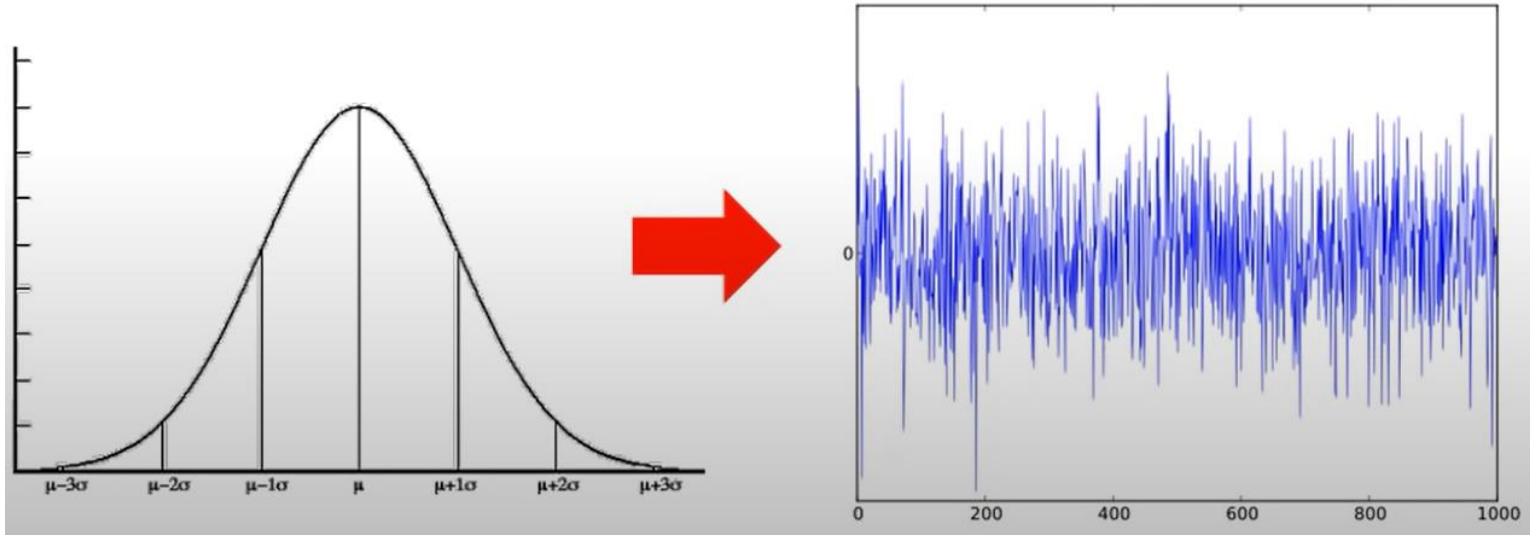


White noise is the left over which can be measured – giving us more certainty of our forecast

If we eliminate all elements of "signal" (trends, periodicity, autocorrelation), what are we left with? **White Noise**.

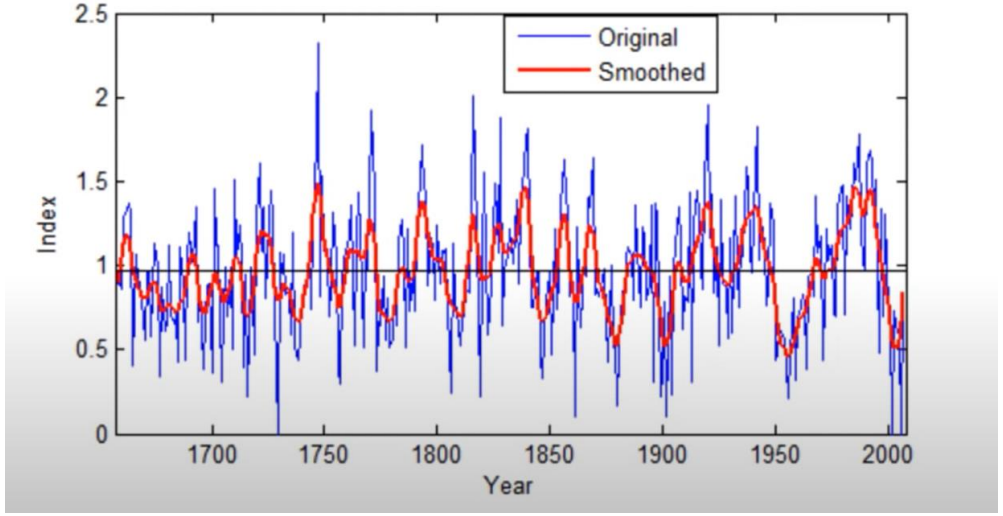
White noise is a random process, whose samples are regarded as a sequence of serially uncorrelated random variables with zero mean and finite variance.

That means we can replicate its nature by simply sampling from an appropriate statistical distribution with replacement.



Smoothing

Smoothing –eg moving averages and other options



Smoothing can be a series of weighs

A statistical filter, or digital filter, is a series of weights that when cumulatively multiplied by consecutive values of a time series gives the filtered series. The series of weights is sometimes called the filtering function, or simply the filter.

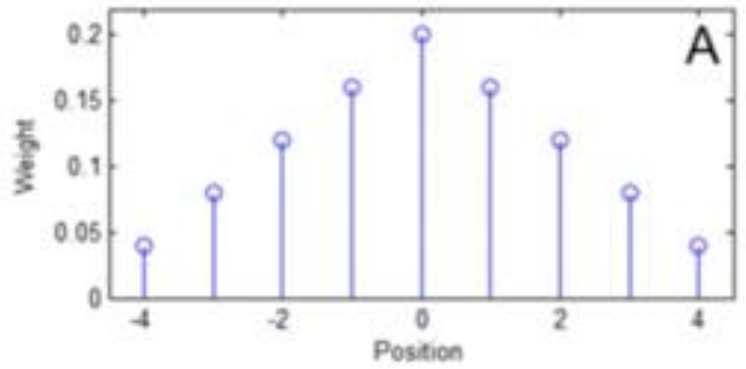
The operation of filtering is illustrated in Table 1

Table 1. Filtering

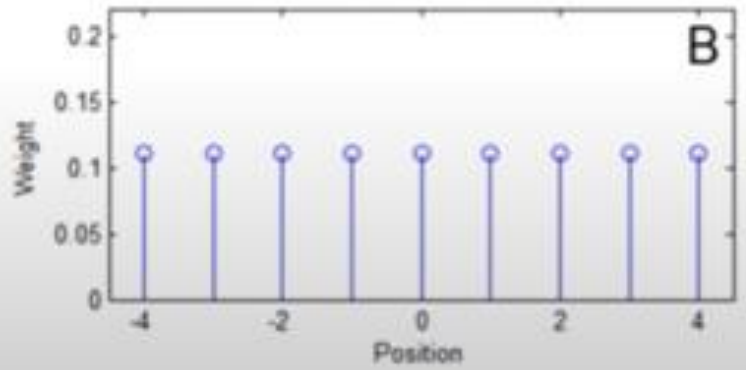
Year	Filter	Time Series	Filtered Values
1		12	
2	.25 x	17	14.00
3	.50 x	10	14.75
4	.25 x	22	17.25
5		15	15.75
6		11	13.75
7		18	18.50
8		27	21.50
9		14	

Filter proceeds by sliding alongside the time series one value at a time, each time computing a cumulative product.

Smoothing Factors can be designed in multiple ways

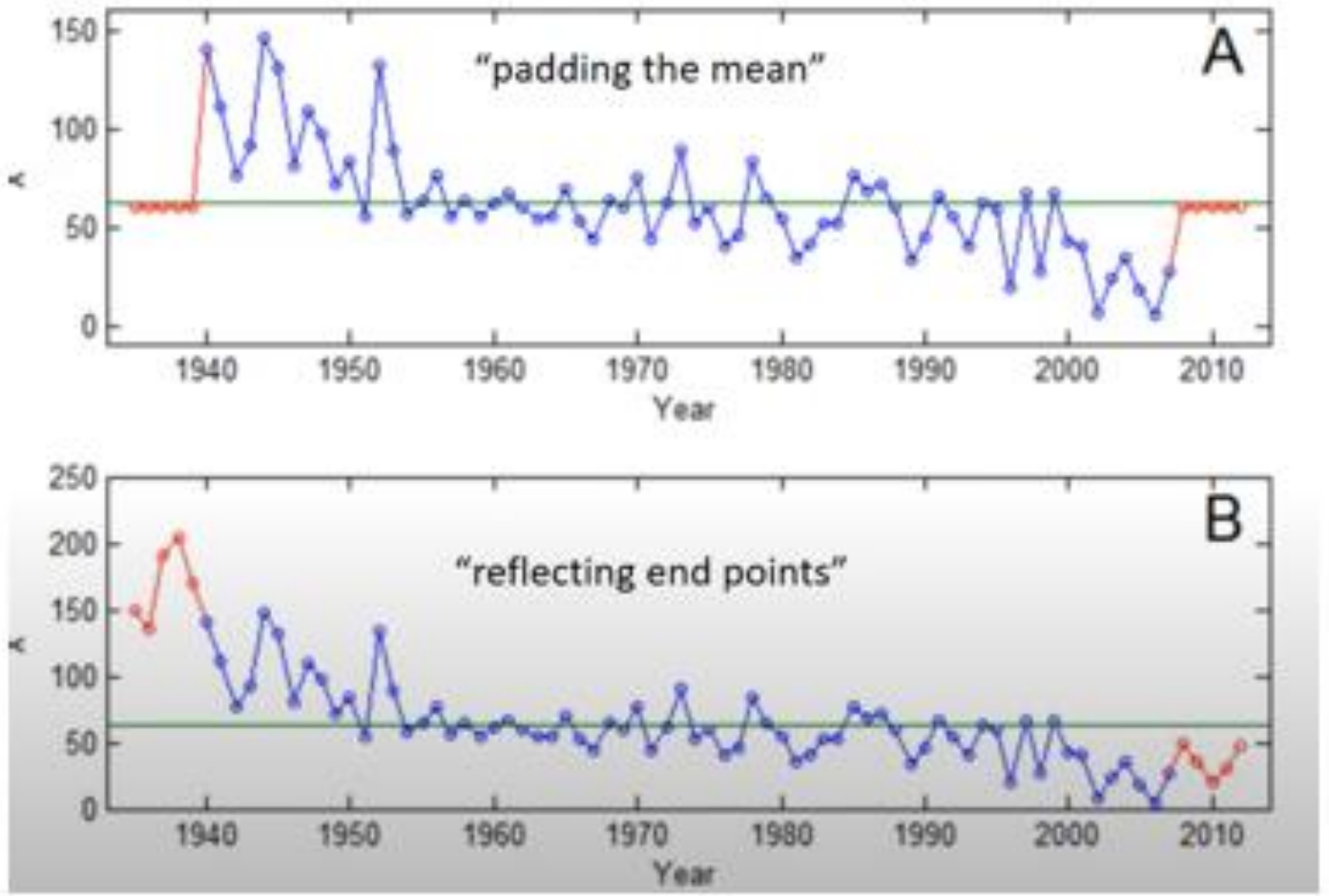


Linear decay.

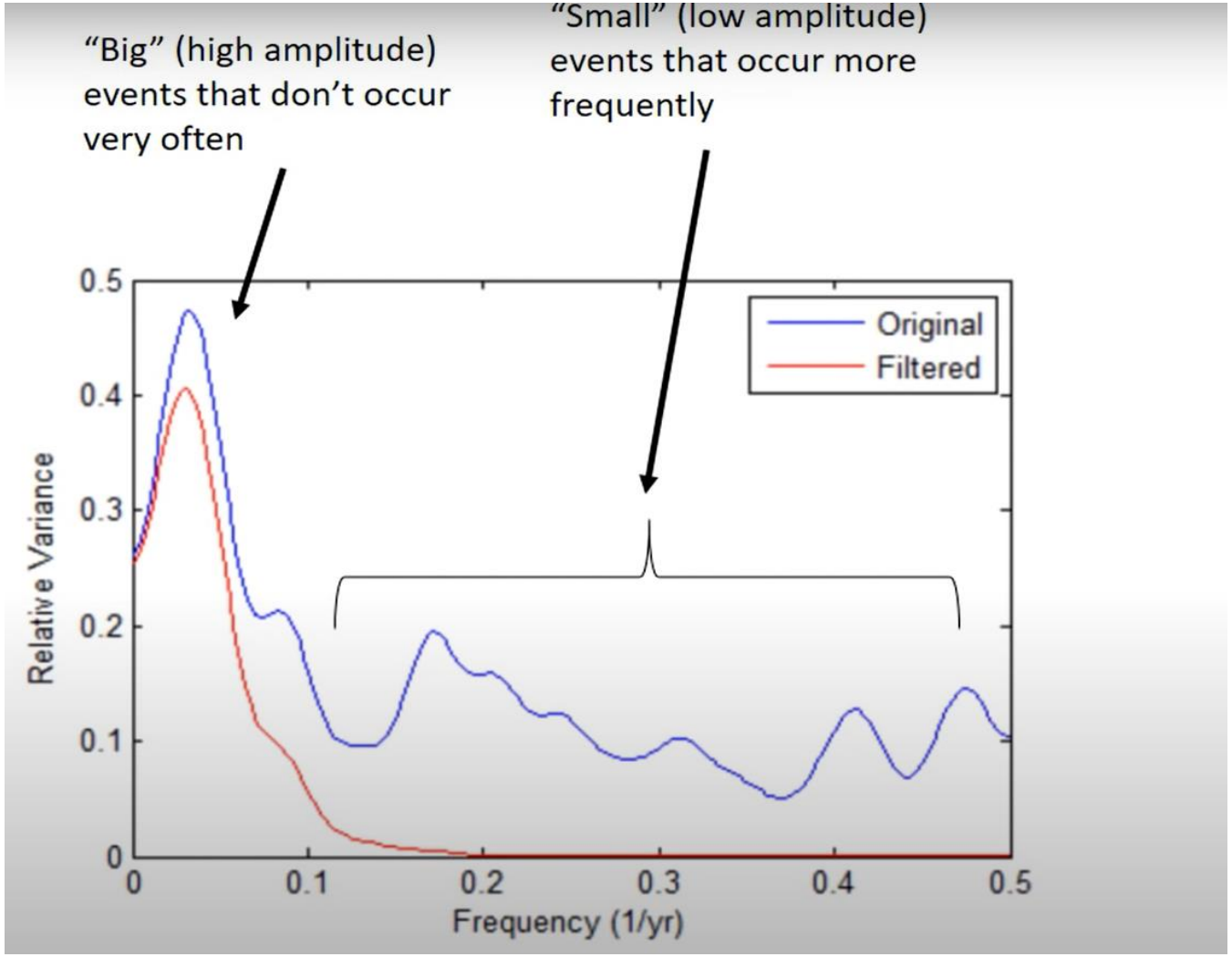


Moving average.

With smoothing you lose data



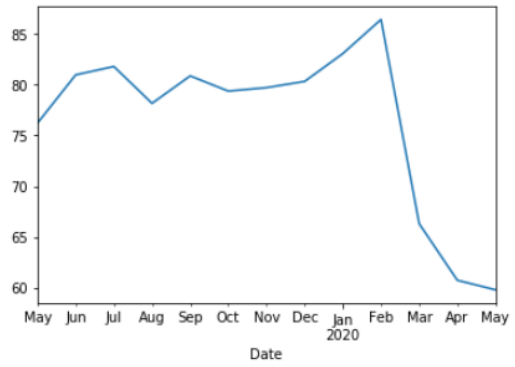
Big and small events can be measured and presented



I then did some basic work with real data

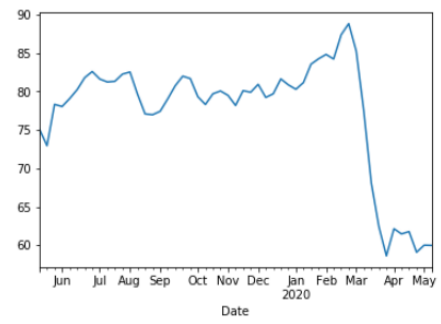
Looked at the CBA share price – in Python

- Downloaded
- Imported
- Made the date the index
- Looked at the mean over a period

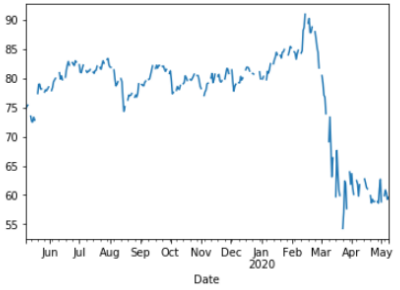



```
df.Close.resample('W').mean().plot()
```

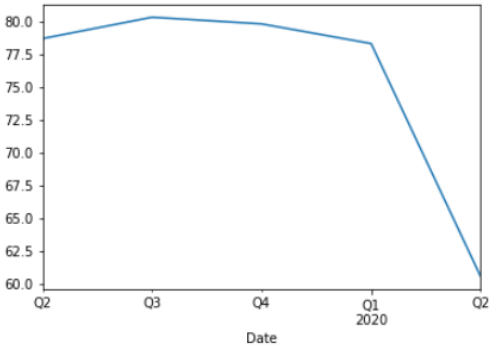
```
<matplotlib.axes._subplots.AxesSubplot at 0x15bc41413c8>
```




```
In [14]: df.Close.resample('D').mean().plot() #Daily
Out[14]: <matplotlib.axes._subplots.AxesSubplot at 0x15bc4283948>
```

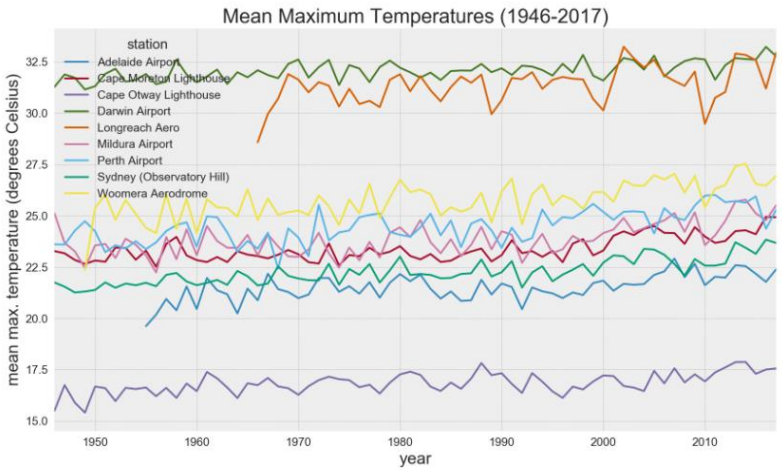


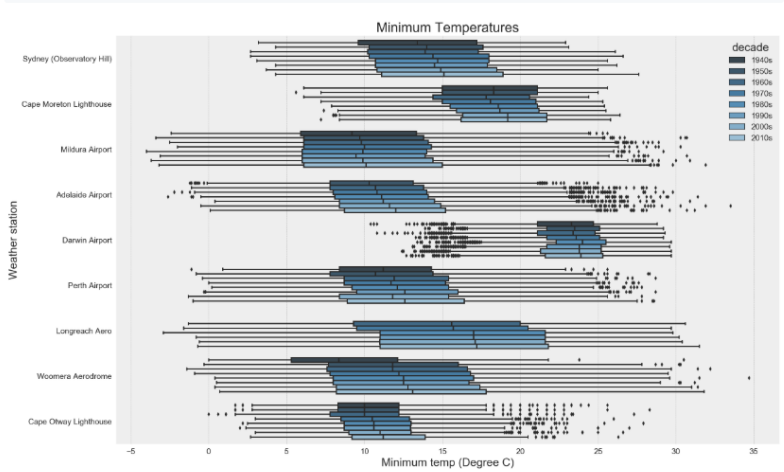

```
<matplotlib.axes._subplots.AxesSubplot at 0x15bc4365b08>
```



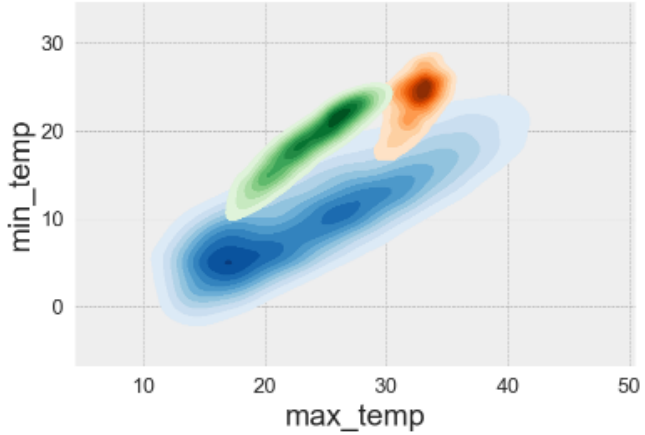
Impressive Sydney weather analysis from

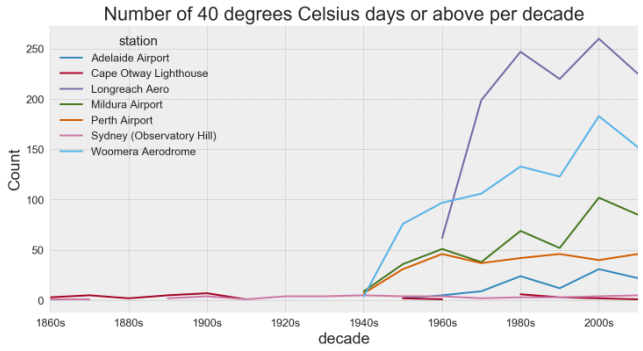
https://gist.github.com/chalg/8644dd03a90fd9ffb72698f77edc980c

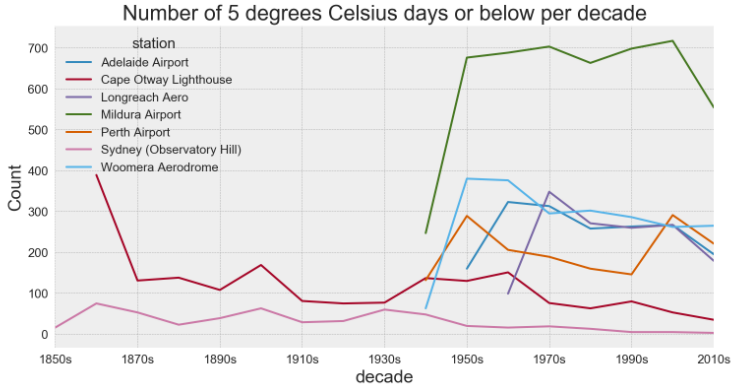




In the below **kernel density estimation plots,** we can see the broader distribution of temperatures in the temperate zone at Mildura Airport (blues), compared to Cape Morten Lighthouse (greens) and Darwin Airport (oranges).







Looked at BOM data

- Found data
- Imported data by day from 1900
- Changed from integer to date
- Made date the index

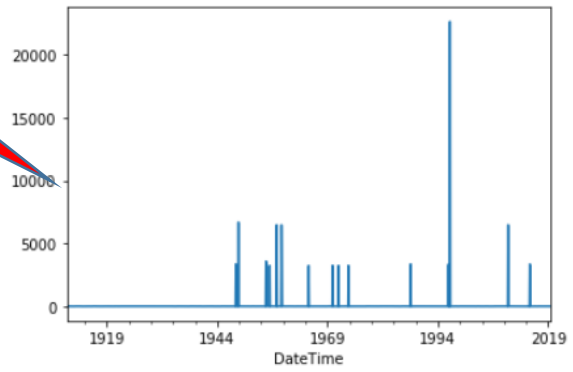
	date	TEMP	DateTime
0	19100102	25.3	1910-01-02

I had obviously bad outliers

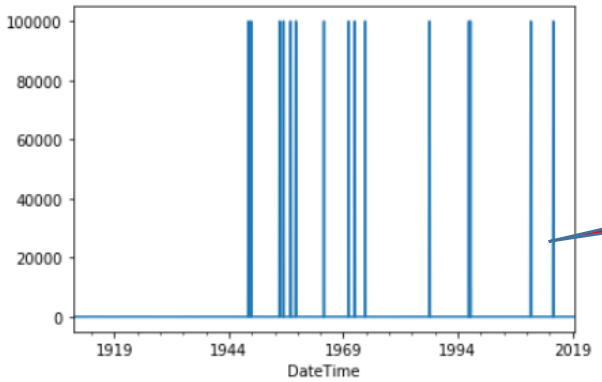
or I did something wrong

Average bad

```
Out[38]: <matplotlib.axes._subplots.AxesSubplot at 0x235ecf62488>
```

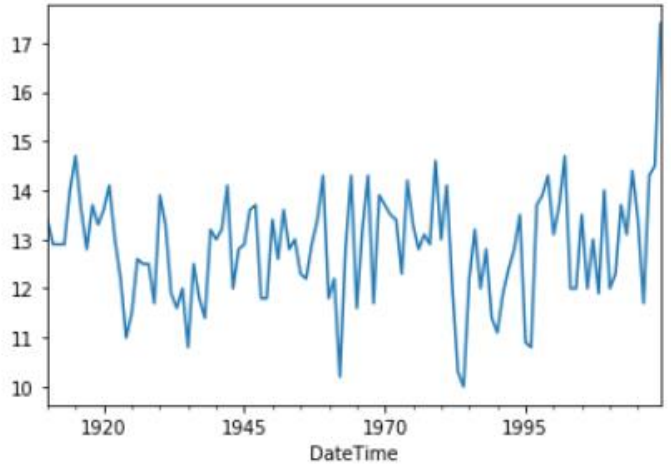



```
In [44]: bom.TEMP.resample('M').max().plot()
Out[44]: <matplotlib.axes._subplots.AxesSubplot at 0x235edf9d548>
```



Max bad

```
In [47]: bom.TEMP.resample('Y').min().plot()
Out[47]: <matplotlib.axes._subplots.AxesSubplot at 0x235ee036bc8>
```



Min good

pandas.Timestamp is useful and you can work out the time now

```
In [1]: import datetime

In [2]: print(datetime.datetime.now())
2020-05-09 15:11:02.981902
```

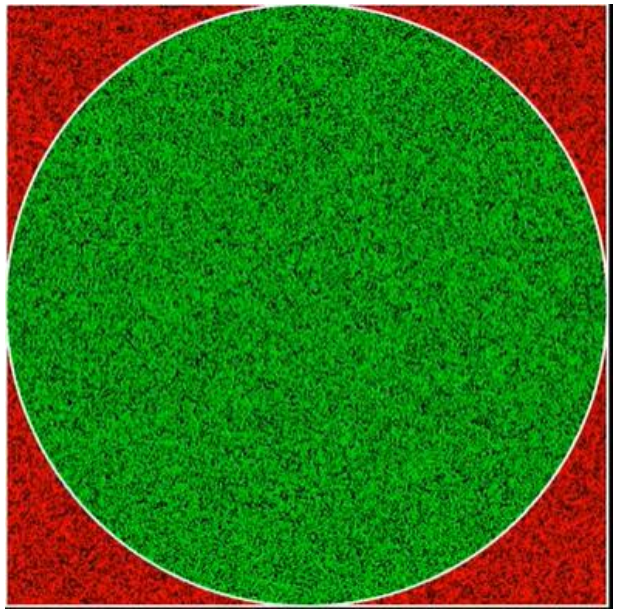
dayofweek	Return day of the week.
dayofyear	Return the day of the year.
days_in_month	Return the number of days in the month.
daysinmonth	Return the number of days in the month.
freqstr	Return the total number of days in the month.
is_leap_year	Return True if year is a leap year.
is_month_end	Return True if date is last day of month.
is_month_start	Return True if date is first day of month.
is_quarter_end	Return True if date is last day of the quarter.
is_quarter_start	Return True if date is first day of the quarter.
is_year_end	Return True if date is last day of the year.
is_year_start	Return True if date is first day of the year.
quarter	Return the quarter of the year.

Some examples from pandas.Timestamp

A few key takeaways

Monte Carlo, or random sampling, is a common technique used

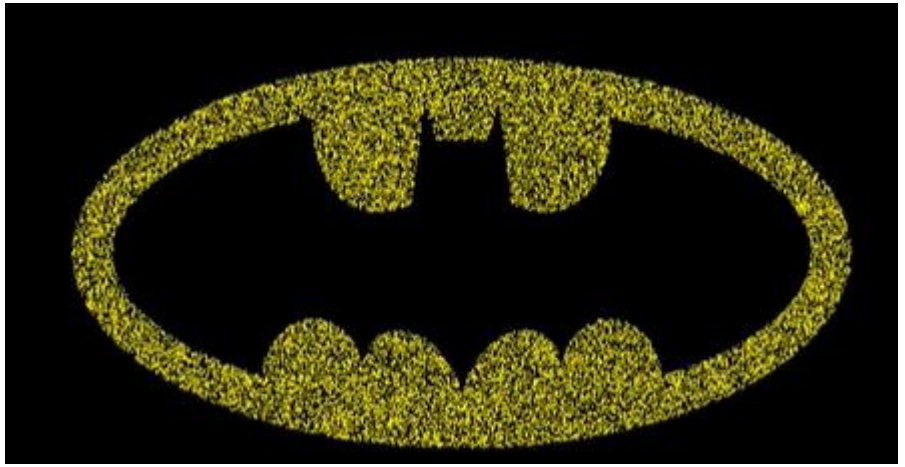
- From Python at Data Camp
- Bird Comparison
- Get random Sample (plug it in) get results
- See Results



Popular in financial planning

- What at 40 ->65 ->80
- Amount will spend scenarios
- Income scenarios
- Return on investments scenarios

All scenarios with a likelihood



Time series is different to regression analysis. In time series we look for trends and model the process

But where regression aims to quantify the specific impacts of specific underlying independent variables, of the form:

$$Y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

Time series modeling allows us to replicate every element of the process by decomposing the mathematical process into a combination of signals (e.g., year-on-year growth in electricity demand, seasonal variability, etc.) and noise (random probabilistic processes), without necessarily knowing the underlying causes for each.

Next Steps

- Find libraries available on Amazon, Google, Facebook and others
- Do Comparison of CBA share price – run model and then see if works in 2018, 2017 etc – learn how
- Fix weather data

Thanks

Alex Dance

Background

- Maths / statistics degree
- Background in big data, strategy, analytics
- Worked at Optus, Salmat, Reuters, Pathfinder Solutions

Copy of This Presentation and code

https://github.com/alexdance2468/

Plus other data science projects completed

Contact Details

www.linkedin.com/in/alex-dance/

Thanks

Sources:

https://www.youtube.com/watch?v=Prpu_U5tKkE - main presentation source
https://www.youtube.com/watch?v=r0s4slGHwzE - timeseries in panda for analysing share price
https://au.finance.yahoo.com/quote/CBA.AX/history/ - daily share prices
https://gist.github.com/chalg/8644dd03a90fd9ffb72698f77edc980c – very impressive on Sydney weather data
http://www.bom.gov.au/climate/data/ - Australian weather data