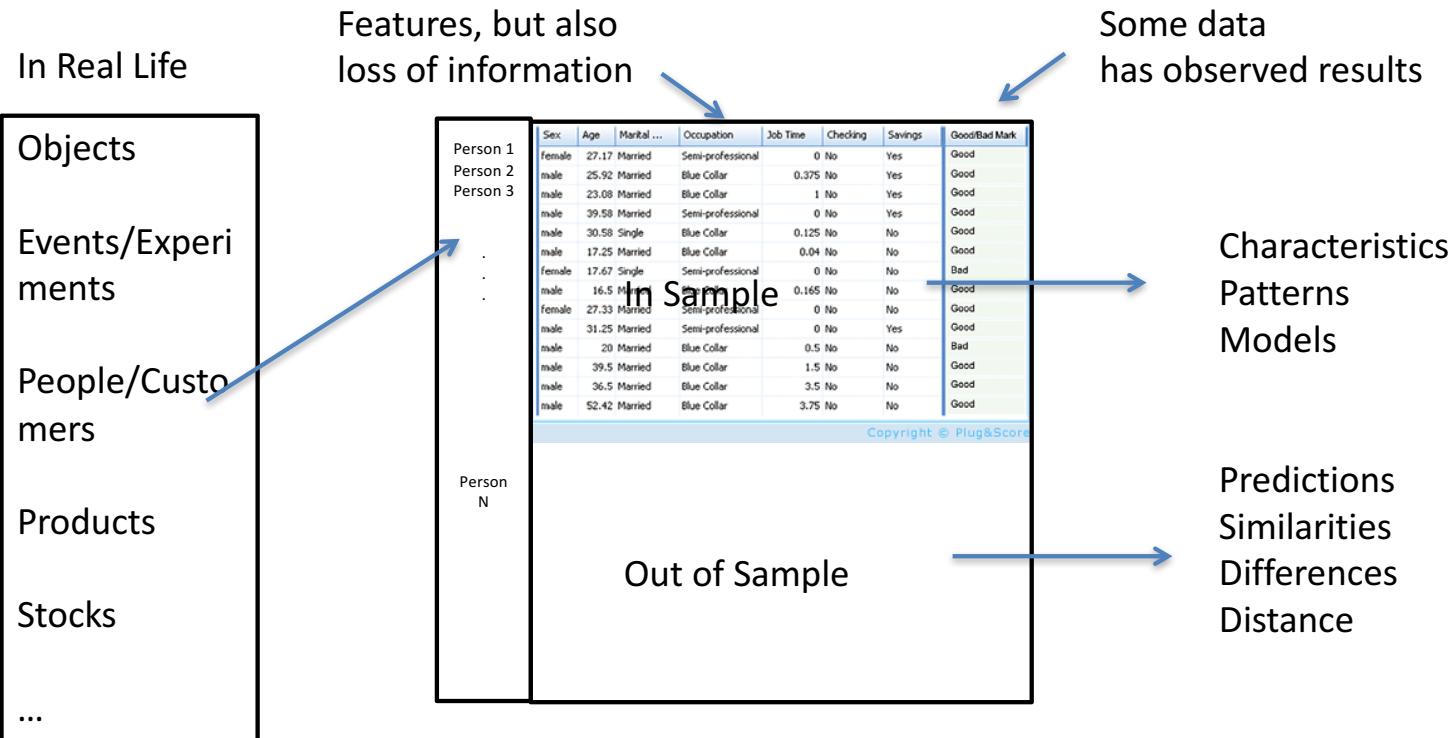


Data X

Data as a Signal and Correlation  
Data X: A Course on Data, Signals, and Systems

Ikhlaq Sidhu  
Chief Scientist & Founding Director,  
Sutardja Center for Entrepreneurship & Technology  
IEOR Emerging Area Professor Award, UC Berkeley

# A High Level Framework



## Converting From Time Sequence Data to Features

Of course, not all data has a time property, but lets start with this type.

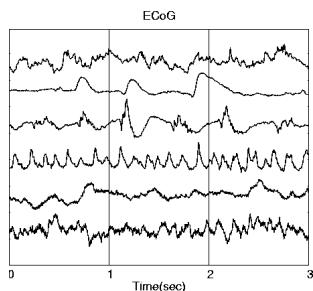
For example( key1, value 1),( key 2, value 2)... in this case, the keys are indexed by time.

Data
X

# Converting From Time Sequence Data to Features

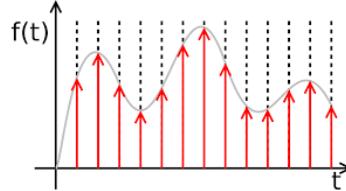
Many Types of data  
are signals in time

- Stock market
- Temperature
- Instrument readings



Continuous signals  
 $x(t)$

Sometimes we  
sample them,  
record at intervals  
of  $T$



Sampled signals (data)  
 $x(nT)$

We get a  
list in a table,  
array, or vector

Rec	Observed
1	60.323
2	61.122
3	60.171
4	61.187
5	63.221
6	63.639
7	64.999
8	63.761
9	66.019
10	67.857
11	68.169
12	66.513
13	68.655
14	69.564
15	69.331
16	70.551

Discrete data  
 $x_n = x_1, x_2, x_3, \dots$



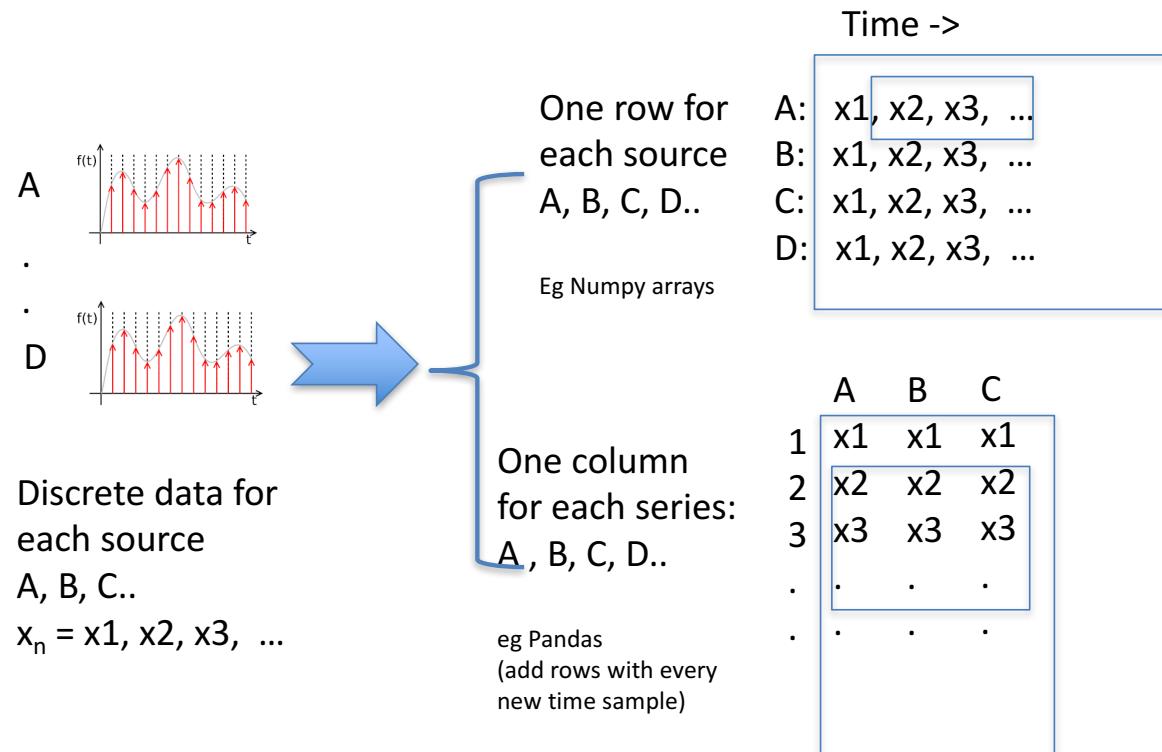
What we want  
(for now):  
features and  
characteristics

- For example:
- Means
  - Variances
  - Pattern matches
  - Changes
  - accumulation
  - Frequency

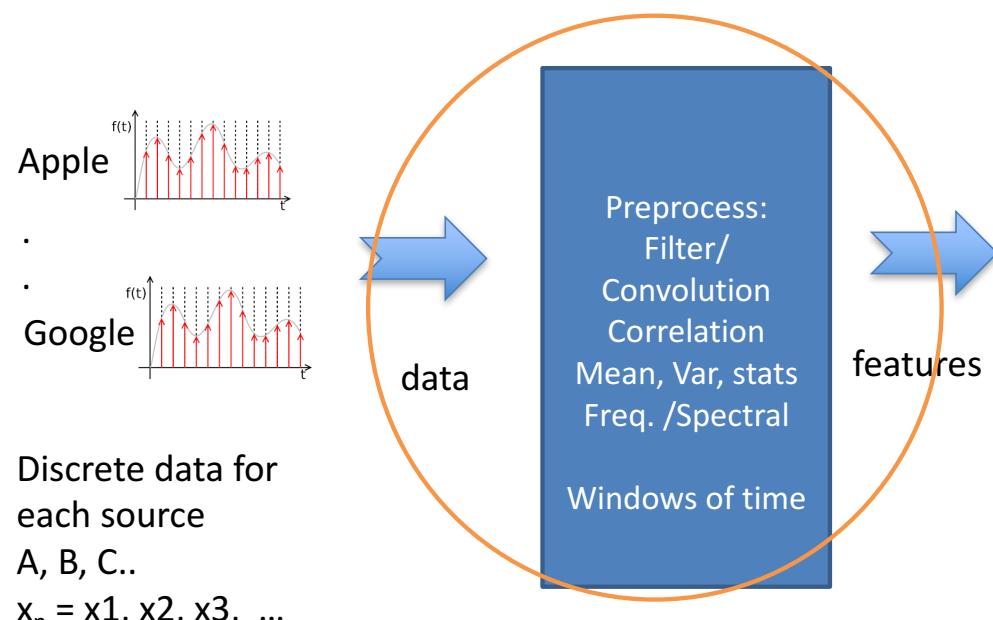
(might lose time reference)



## Approaches to the Data Sequences from Multiple Sources in Tables



## Data Sequence in Tables Example



	Price	Price[n-20]	20 day MA	1 year average	Expected Price?
APPL	appl[n]	appl[n-20]	mva(appl, 20)	mva(appl, 200)	
FB	fb[n]	fb[n-20]	mva(fb, 20)	mva(fb, 200)	
GOOG	goog[n]	goog[n-20]	mva(goog, 20)	mva(goog, 200)	

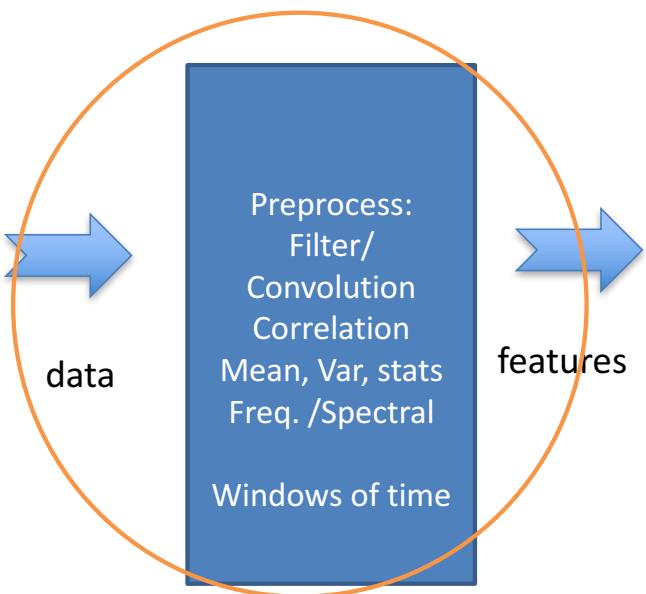
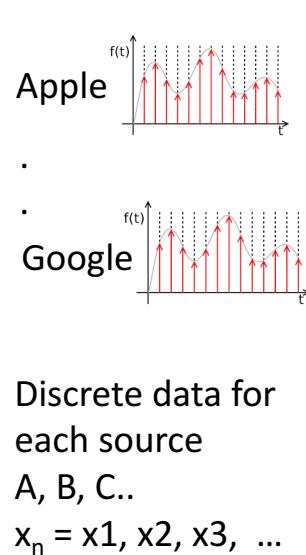
Data Input and  
Temp Storage

Preprocess  
(and lose  
some information)

ML for Decisions / Predictions



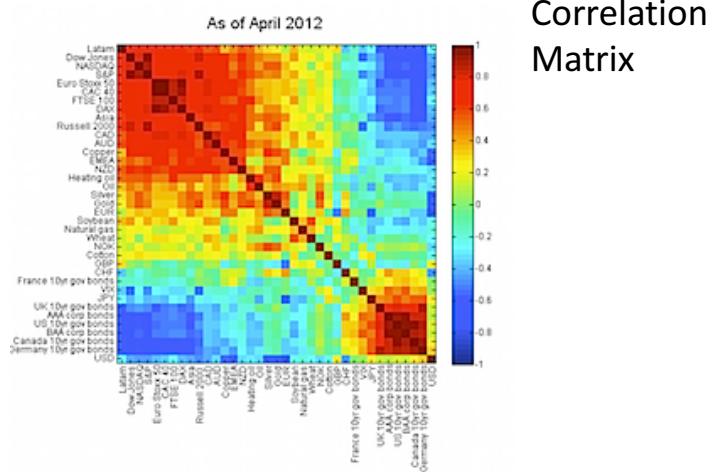
# Data Sequence in Tables Example



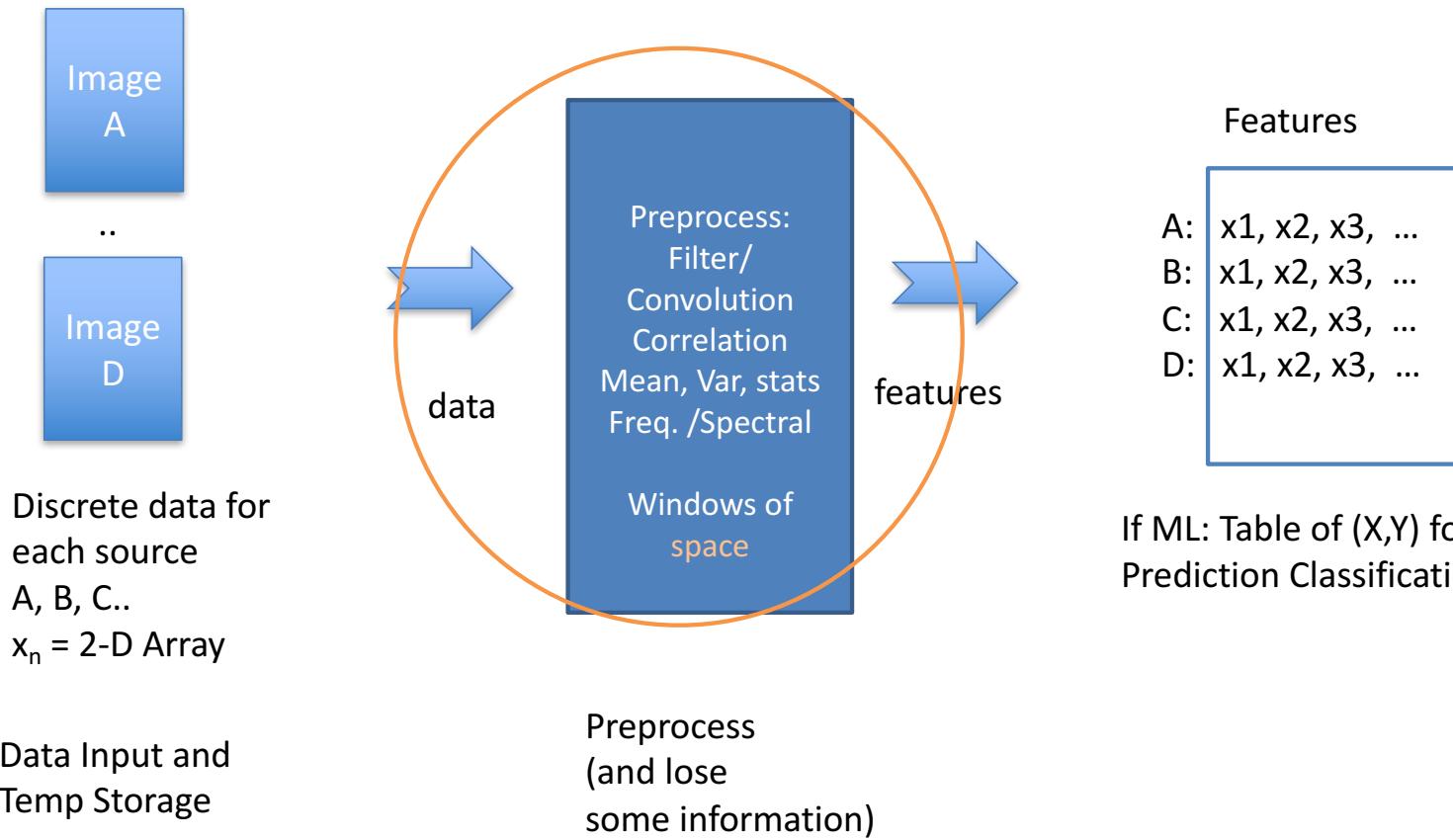
Data Input and  
Temp Storage

Preprocess  
(and lose  
some information)

Correlation Matrix: Daily % Change Correlation Over Last Ten Years																
Ticker	S&P 500	C. Disc.	C. Stap.	Energy	Finan.	H Care	Indust.	Mater.	Tech	Telcom	Utilit.	Oil	Gold	Dollar	L Bnd	
S&P 500	1.00	0.94	0.85	0.81	0.88	0.86	0.94	0.89	0.90	0.78	0.76	0.24	-0.01	-0.17	-0.37	
Cons. Disc.	0.94	1.00	0.81	0.68	0.82	0.78	0.90	0.82	0.85	0.72	0.67	0.16	-0.07	-0.12	-0.34	
Cons. Stap.	0.85	0.81	1.00	0.65	0.67	0.81	0.78	0.71	0.70	0.69	0.71	0.12	-0.06	-0.10	-0.28	
Energy	0.81	0.68	0.65	1.00	0.62	0.66	0.73	0.81	0.66	0.59	0.70	0.48	0.16	-0.28	-0.30	
Financials	0.88	0.82	0.67	0.62	1.00	0.68	0.82	0.73	0.72	0.64	0.57	0.16	-0.06	-0.13	-0.31	
H Care	0.86	0.78	0.81	0.66	0.68	1.00	0.78	0.71	0.73	0.68	0.69	0.14	-0.04	-0.12	-0.29	
Industrials	0.94	0.90	0.78	0.73	0.82	0.78	1.00	0.87	0.84	0.70	0.68	0.21	-0.01	-0.18	-0.37	
Materials	0.89	0.82	0.71	0.81	0.73	0.71	0.87	1.00	0.79	0.79	0.65	0.67	0.29	0.15	-0.27	-0.35
Technology	0.90	0.85	0.70	0.66	0.72	0.73	0.84	0.79	1.00	0.71	0.63	0.16	-0.04	-0.10	-0.35	
Telecom	0.78	0.72	0.69	0.59	0.64	0.68	0.70	0.65	0.71	1.00	0.63	0.12	-0.05	-0.10	-0.25	
Utilities	0.76	0.67	0.71	0.70	0.57	0.69	0.68	0.67	0.63	0.63	1.00	0.19	0.02	-0.15	-0.22	
Oil	0.24	0.16	0.12	0.48	0.16	0.14	0.21	0.29	0.16	0.12	0.19	1.00	0.29	-0.30	-0.22	
Gold	-0.01	-0.07	-0.06	0.16	-0.06	-0.04	-0.01	0.15	-0.04	-0.05	0.02	0.29	1.00	-0.43	0.07	
Dollar	-0.17	-0.12	-0.10	-0.28	-0.13	-0.12	-0.18	-0.27	-0.10	-0.10	-0.15	-0.30	-0.43	1.00	-0.05	
Long Bond	-0.37	-0.34	-0.28	-0.30	-0.31	-0.29	-0.37	-0.35	-0.35	-0.25	-0.22	-0.22	0.07	-0.05	1.00	



## Data Sequence in Tables Example



## Correlation Matrices

Data X

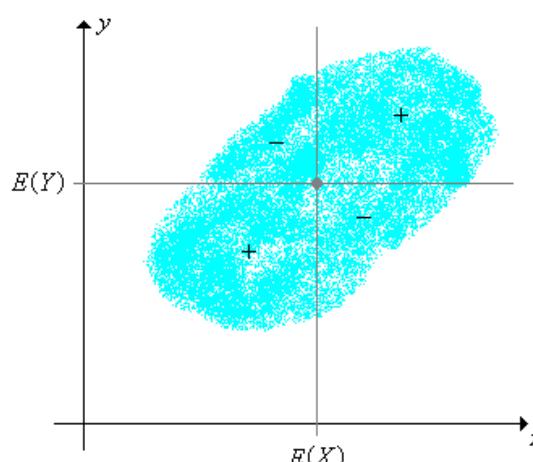


# Correlation and Covariance

1. The *covariance* of  $(X, Y)$  is defined by

$$\text{cov}(X, Y) = \mathbb{E}([X - \mathbb{E}(X)][Y - \mathbb{E}(Y)])$$

and, assuming the variances are positive, the *correlation* of  $(X, Y)$  is defined by



$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} :$$

$$= \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

<http://www.math.uah.edu/stat/expect/Covariance.html>



## Correlation and Covariance

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

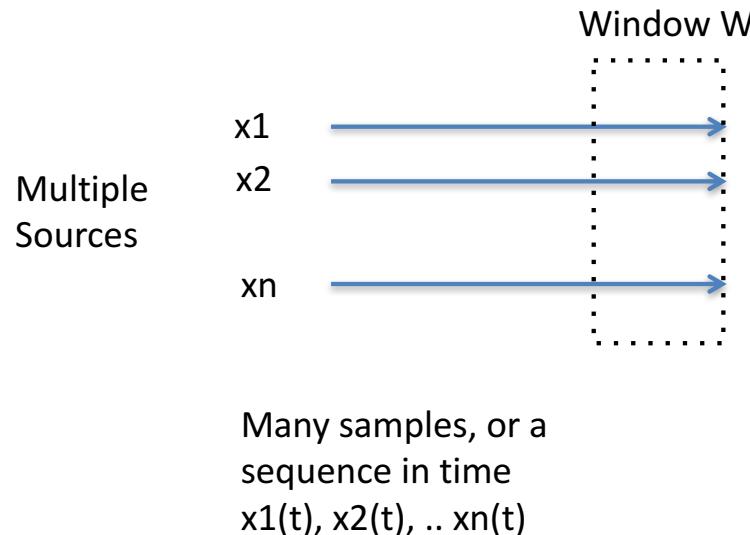
### Properties

- a.  $-1 \leq \text{cor}(X, Y) \leq 1$
- b.  $-\text{sd}(X)\text{sd}(Y) \leq \text{cov}(X, Y) \leq \text{sd}(X)\text{sd}(Y)$
- c.  $\text{cor}(X, Y) = 1$  if and only if  $Y$  is a linear function of  $X$  with positive slope.
- d.  $\text{cor}(X, Y) = -1$  if and only if  $Y$  is a linear function of  $X$  with negative slope.

<http://www.math.uah.edu/stat/expect/Covariance.html>



## Correlation Matrix



Table

Samples	$x_1$	$x_2$	$\dots$	$x_n$
1				
2				
3				
$n$				
.				
$N+W$				

Samples from Window of  $W$

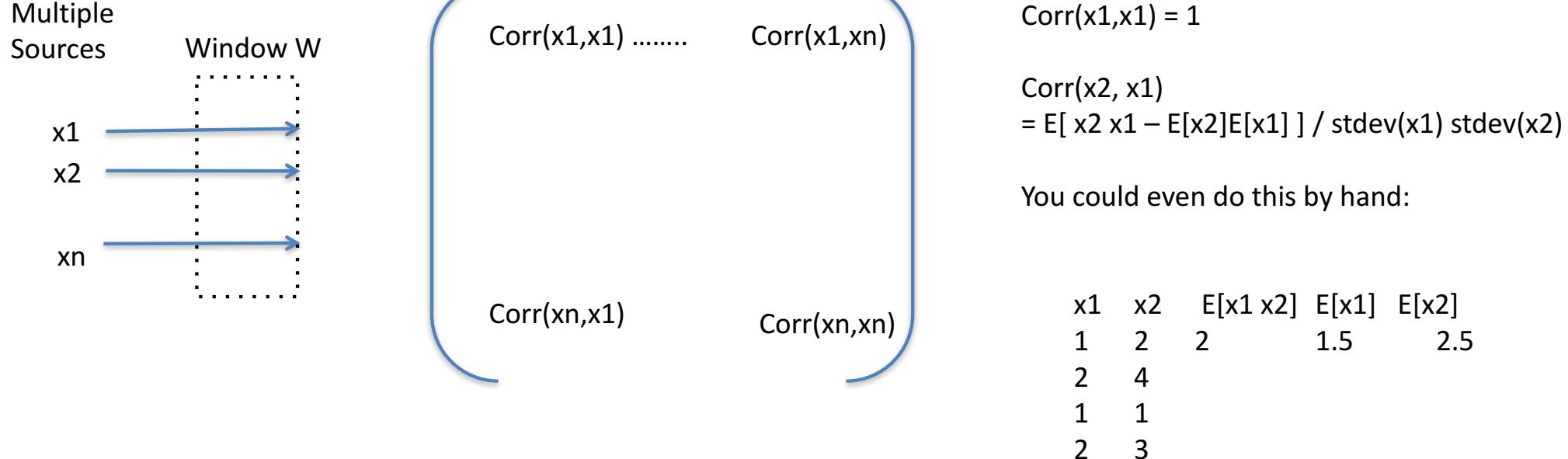
The diagram shows a table representing a correlation matrix. The columns are labeled  $x_1$ ,  $x_2$ ,  $\dots$ ,  $x_n$ . The rows are labeled "Samples" followed by indices 1, 2, 3,  $n$ ,  $\cdot$ ,  $N+W$ . A dashed rectangle highlights the "Samples from Window of  $W$ ", which corresponds to the samples within the window  $W$  shown in the previous diagram.

To estimate from data:

- a) Use all samples ever collected
- b) Use window size of  $W$  samples of each to estimate a recent Corr Matrix



## Correlation Matrix



To estimate from data:

- Use all samples ever collected
- Use window size of  $W$  samples of each to estimate recent Corr Matrix



## Code Examples: Correlation of Rows with NumPy

```
Import numpy as np

# ignore line formatting
x = np.array([
    [[0.1, .32, .2, 0.4, 0.8],
     [.23, .18, .56, .61, .12],
     [.9, .3, .6, .5, .3],
     [.34, .75, .91, .19, .21]]]

np.corrcoef(x)
Out[4]: array([
    [ 1.          , -0.35153114, -0.74736506, -0.48917666],
    [-0.35153114,  1.          ,  0.23810227,  0.15958285],
    [-0.74736506,  0.23810227,  1.          , -0.03960706],
    [-0.48917666,  0.15958285, -0.03960706,  1.          ]
])
```

Here each row is a vector of length 5  
There are 4 vectors

Correlation matrix is 4 x 4

If you want the correlation of the columns,  
just use transpose

`np.corrcoef ( np.transpose(x) )`

For a window, use a slice:  
`window = x[0:4,3:5]` for the last  
two columns



# Correlation of Features from Different Sources

Mazda RX4  
Mazda RX4 Wag  
Datsun 710  
Hornet 4 Drive  
Hornet Sportabout  
Valiant

	mpg	disp	hp	drat	wt	qsec
Mazda RX4	21.0	160	110	3.90	2.620	16.46
Mazda RX4 Wag	21.0	160	110	3.90	2.875	17.02
Datsun 710	22.8	108	93	3.85	2.320	18.61
Hornet 4 Drive	21.4	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	360	175	3.15	3.440	17.02
Valiant	18.1	225	105	2.76	3.460	20.22

Pandas Table  
Use corr()

like dataframe.corr()

## pandas.DataFrame.corr

DataFrame.corr(method='pearson', min\_periods=1)

[source]

Compute pairwise correlation of columns, excluding NA/null values

Parameters:

- method : {'pearson', 'kendall', 'spearman'}
  - pearson : standard correlation coefficient
  - kendall : Kendall Tau correlation coefficient
  - spearman : Spearman rank correlation

min\_periods : int, optional

Minimum number of observations required per pair of columns to have a valid result.  
Currently only available for pearson and spearman correlation

Returns:

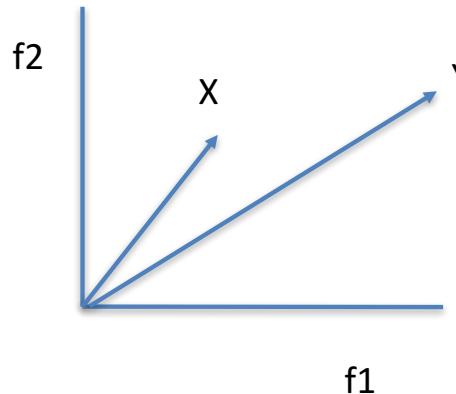
y : DataFrame

	mpg	disp	hp	drat	wt	qsec
mpg	1.00	-0.85	-0.78	0.68	-0.87	0.42
disp	-0.85	1.00	0.79	-0.71	0.89	-0.43
hp	-0.78	0.79	1.00	-0.45	0.66	-0.71
drat	0.68	-0.71	-0.45	1.00	-0.71	0.09
wt	-0.87	0.89	0.66	-0.71	1.00	-0.17
qsec	0.42	-0.43	-0.71	0.09	-0.17	1.00



# Correlation Types: Pearson Kendal, Spearman

Understanding Correlation in a different way



$$X \cdot Y = |X| |Y| \cos \Theta$$

## pandas.DataFrame.corr

`DataFrame.corr(method='pearson', min_periods=1)`

Compute pairwise correlation of columns, excluding NA/null values

### Parameters:

- method : {'pearson', 'kendall', 'spearman'}
  - pearson : standard correlation coefficient
  - kendall : Kendall Tau correlation coefficient
  - spearman : Spearman rank correlation

### min\_periods : int, optional

Minimum number of observations required per pair. Currently only available for pearson and spearman.

### Returns:

y : DataFrame

	mpg	disp	hp	drat	wt	qsec
mpg	1.00	-0.85	-0.78	0.68	-0.87	0.42
disp	-0.85	1.00	0.79	-0.71	0.89	-0.43
hp	-0.78	0.79	1.00	-0.45	0.66	-0.71
drat	0.68	-0.71	-0.45	1.00	-0.71	0.09
wt	-0.87	0.89	0.66	-0.71	1.00	-0.17
qsec	0.42	-0.43	-0.71	0.09	-0.17	1.00



## Pandas will create a correlation matrix with “columns”

```
In [15]: frame = pd.DataFrame(np.random.randn(1000, 5), columns=['a', 'b', 'c', 'd', 'e'])

In [16]: frame.ix[::2] = np.nan

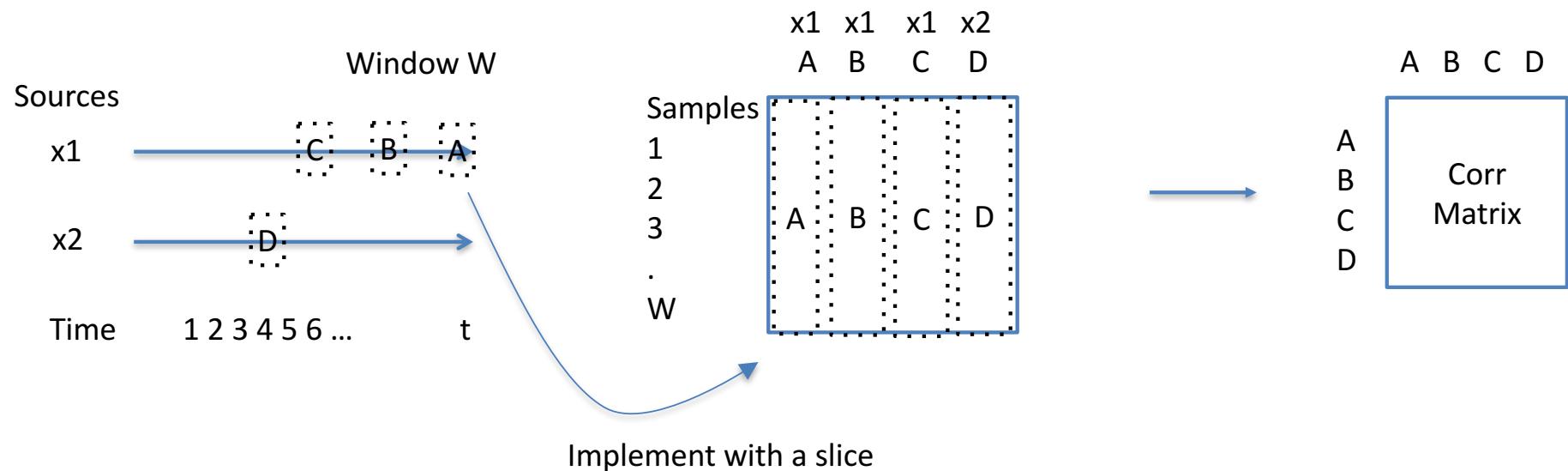
# Series with Series
In [17]: frame['a'].corr(frame['b'])
Out[17]: 0.013479040400098775

In [18]: frame['a'].corr(frame['b'], method='spearman')
Out[18]: -0.0072898851595406371

# Pairwise correlation of DataFrame columns
In [19]: frame.corr()
Out[19]:
      a          b          c          d          e
a  1.000000  0.013479 -0.049269 -0.042239 -0.028525
b  0.013479  1.000000 -0.020433 -0.011139  0.005654
c -0.049269 -0.020433  1.000000  0.018587 -0.054269
d -0.042239 -0.011139  0.018587  1.000000 -0.017060
e -0.028525  0.005654 -0.054269 -0.017060  1.000000
```



## Correlation Matrix with multiple sources and time segments

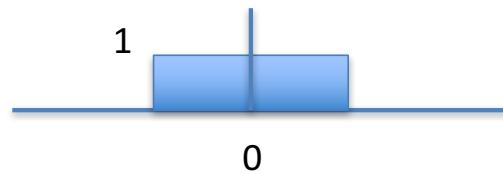


Why would you want a correlation matrix like this?

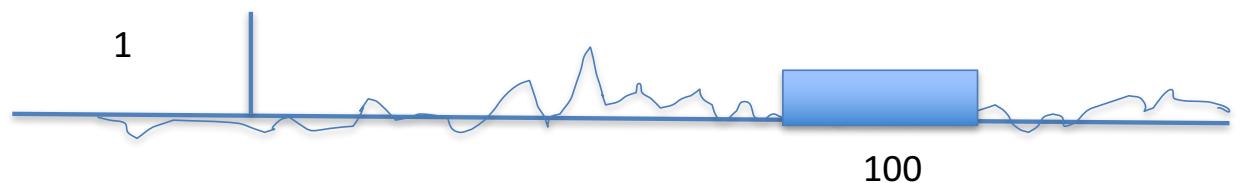


## Correlation Matrix with multiple sources and time segments

Suppose this is  $x_1$   
as an array of numbers 0 0 1 1..1 0 0 0



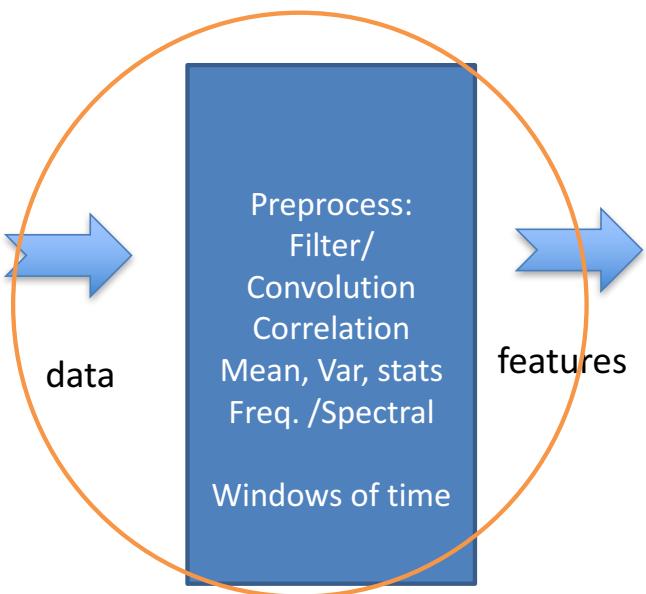
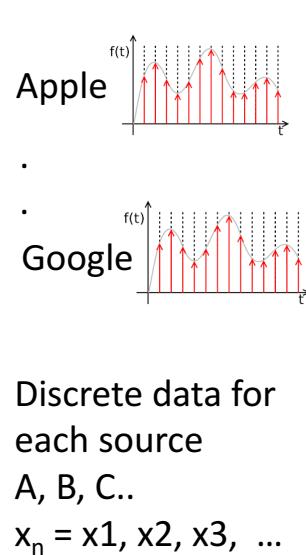
Suppose this is  $x_2$



What is `np.corr(x1,x2[n:n+w])`?



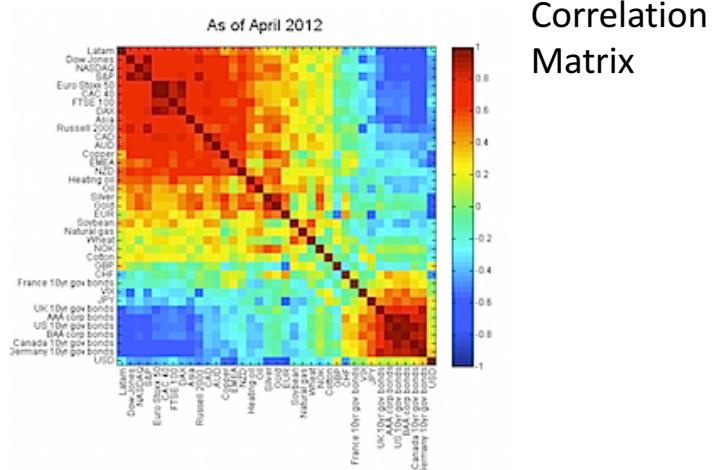
# Data Sequence in Tables Example



Data Input and  
Temp Storage

Preprocess  
(and lose  
some information)

Correlation Matrix: Daily % Change Correlation Over Last Ten Years																
Ticker	S&P 500	C. Disc.	C. Stap.	Energy	Finan.	H Care	Indust.	Mater.	Tech	Telcom	Utilit.	Oil	Gold	Dollar	L Bnd	
S&P 500	1.00	0.94	0.85	0.81	0.88	0.86	0.94	0.89	0.90	0.78	0.76	0.24	-0.01	-0.17	-0.37	
Cons. Disc.	0.94	1.00	0.81	0.68	0.82	0.78	0.90	0.82	0.85	0.72	0.67	0.16	-0.07	-0.12	-0.34	
Cons. Stap.	0.85	0.81	1.00	0.65	0.67	0.81	0.78	0.71	0.70	0.69	0.71	0.12	-0.06	-0.10	-0.28	
Energy	0.81	0.68	0.65	1.00	0.62	0.66	0.73	0.81	0.66	0.59	0.70	0.48	0.16	-0.28	-0.30	
Financials	0.88	0.82	0.67	0.62	1.00	0.68	0.82	0.73	0.72	0.64	0.57	0.16	-0.06	-0.13	-0.31	
H Care	0.86	0.78	0.81	0.66	0.68	1.00	0.78	0.71	0.73	0.68	0.69	0.14	-0.04	-0.12	-0.29	
Industrials	0.94	0.90	0.78	0.73	0.82	0.78	1.00	0.87	0.84	0.70	0.68	0.21	-0.01	-0.18	-0.37	
Materials	0.89	0.82	0.71	0.81	0.73	0.71	0.87	1.00	0.79	0.79	0.65	0.67	0.29	0.15	-0.27	-0.35
Technology	0.90	0.85	0.70	0.66	0.72	0.73	0.84	0.79	1.00	0.71	0.63	0.16	-0.04	-0.10	-0.35	
Telecom	0.78	0.72	0.69	0.59	0.64	0.68	0.70	0.65	0.71	1.00	0.63	0.12	-0.05	-0.10	-0.25	
Utilities	0.76	0.67	0.71	0.70	0.57	0.69	0.68	0.67	0.63	0.63	1.00	0.19	0.02	-0.15	-0.22	
Oil	0.24	0.16	0.12	0.48	0.16	0.14	0.21	0.29	0.16	0.12	0.19	1.00	0.29	-0.30	-0.22	
Gold	-0.01	-0.07	-0.06	0.16	-0.06	-0.04	-0.01	0.15	-0.04	-0.05	0.02	0.29	1.00	-0.43	0.07	
Dollar	-0.17	-0.12	-0.10	-0.28	-0.13	-0.12	-0.18	-0.27	-0.10	-0.10	-0.15	-0.30	-0.43	1.00	-0.05	
Long Bond	-0.37	-0.34	-0.28	-0.30	-0.31	-0.29	-0.37	-0.35	-0.35	-0.25	-0.22	-0.22	0.07	-0.05	1.00	



Descriptive Statistics



End of Section

0 0 0 1 0 1 0 1 0 1 1 1 0 0 0 0 0 0 1 0 0 1 0 1 0 1 1 1 0 0  
1 0 1 1 X 1 1 0 0 1 0 1 0 0 1 0 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1 0 0  
1 Data 0 0 1 0 1 0 1 0 1 0 0 1 0 1 0 1 0 1 1 1 1 1 0 1 0 1 1 1 0 0