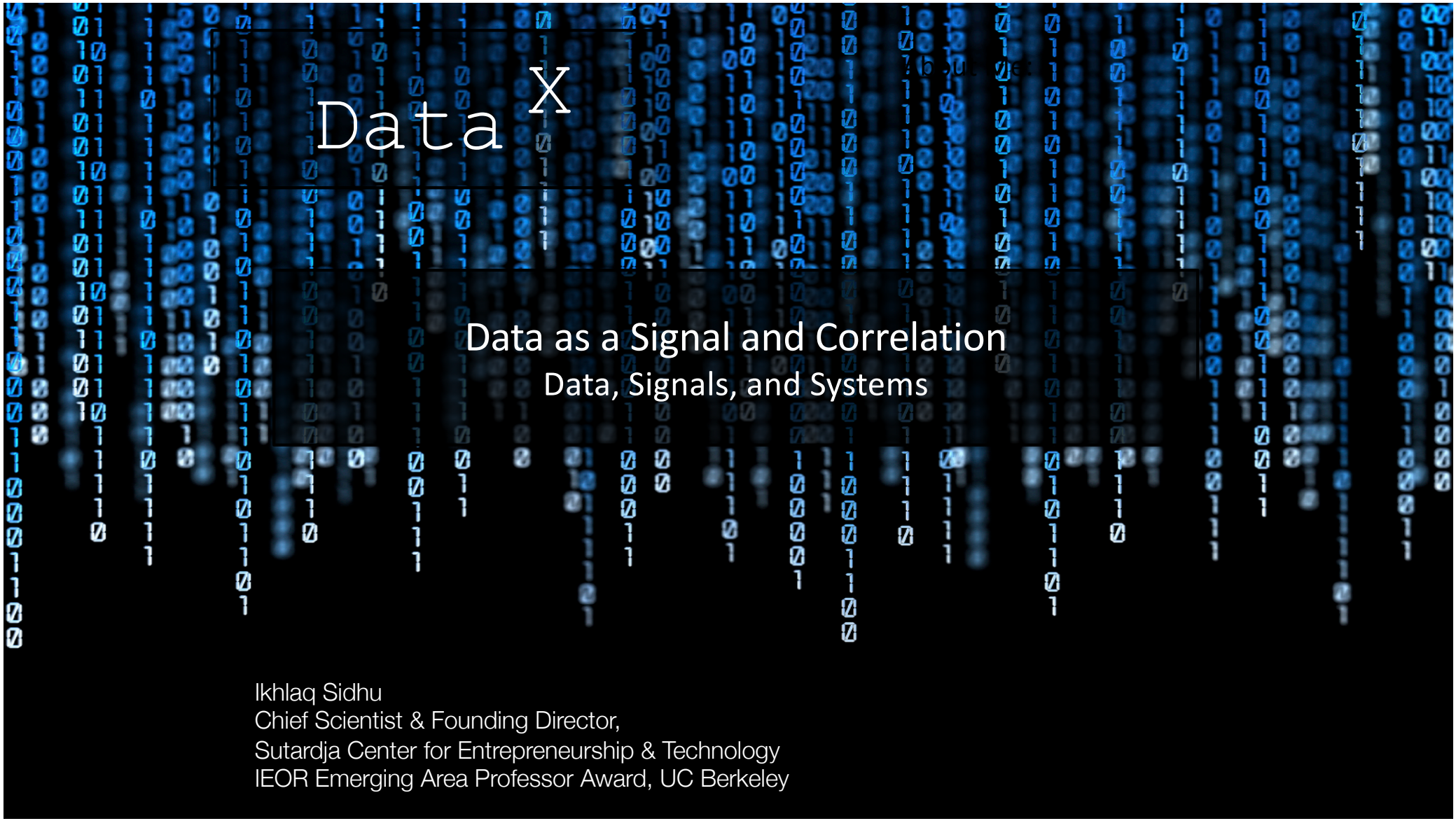


Data as a Signal and Correlation

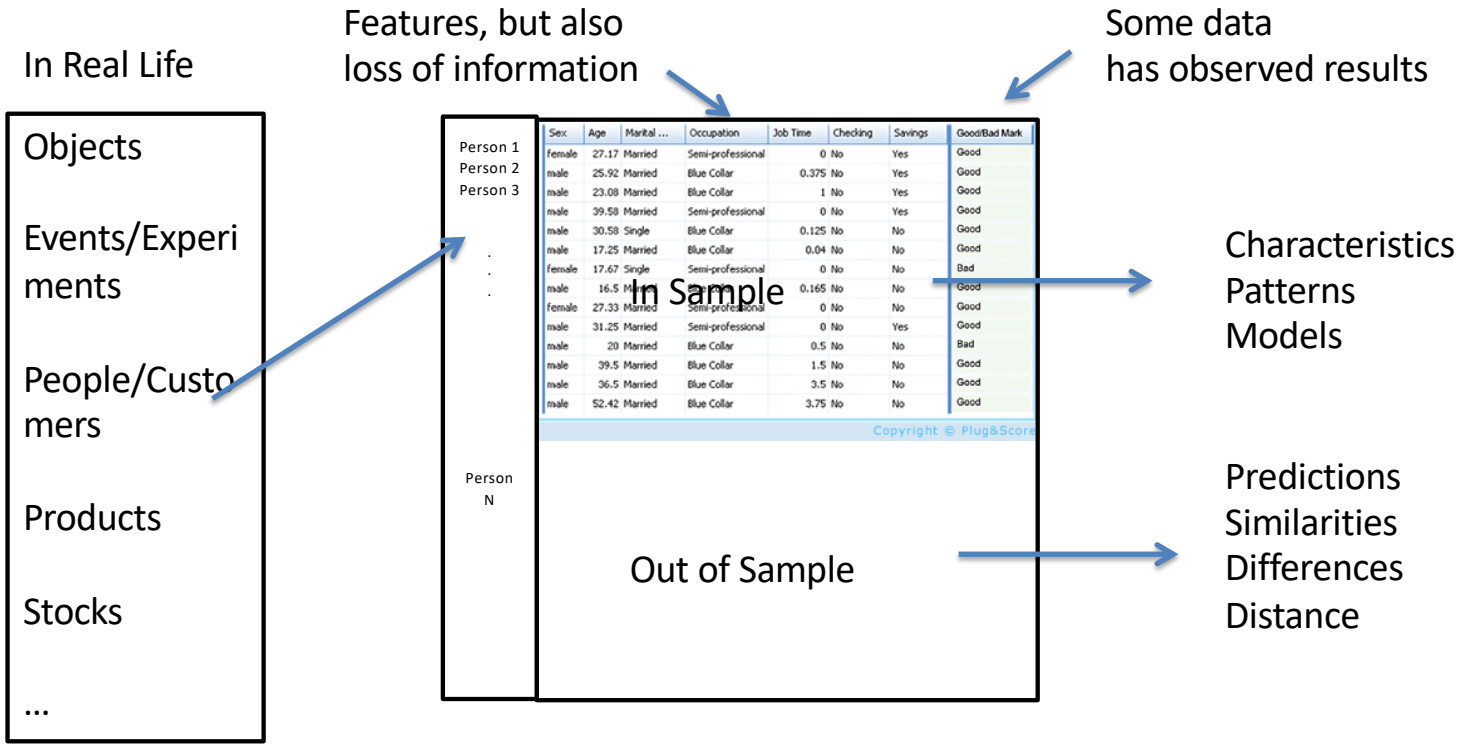
Data, Signals, and Systems

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

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.

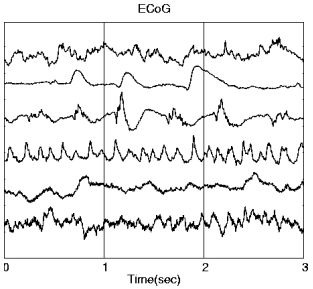
A High Level Framework



Converting From Time Sequence Data to Features

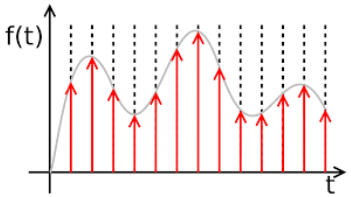
Many Types of data are signals in time

- 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	64.989
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 = x1, x2, x3, …

(might lose time reference)

What we want (for now): features and characteristics

For example:

- Means
- Variances
- Patten matches
- Changes
- accumulation
- Frequency

This leads to tables where the time is an index and each column is a different time sequence (eg stock price, sensor readings, etc)

	A	B	C	D
Date	Ozone ($\mu g/m^3$)	Temperature (°C)	Relative humidity (%)	*n* deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198
7 Jan 2002	4.08	5.2	93.5	180
8 Jan 2002	3.13	3.5	81.5	188
9 Jan 2002	2.05	3.2	88.3	168
10 Jan 2002	5.19	5.3	85.4	194
11 Jan 2002	3.59	3.0	92.6	223
12 Jan 2002	12.87	4.8	94.2	201

Correlation of two rows:

Corr(A,B) is _____

Example rows of time series data from the London dataset showing daily levels of environmental variables

https://www.researchgate.net/figure/Example-rows-of-time-series-data-from-the-London-dataset-showing-daily-levels-of_tbl1_237840362

What is the Correlation of the entire table?

	A	B	C	D
Date	Ozone ($\mu g/m^3$)	Temperature (°C)	Relative humidity (%)	*n* deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198
7 Jan 2002	4.08	5.2	93.5	180
8 Jan 2002	3.13	3.5	81.5	188
9 Jan 2002	2.05	3.2	88.3	168
10 Jan 2002	5.19	5.3	85.4	194
11 Jan 2002	3.59	3.0	92.6	223
12 Jan 2002	12.87	4.8	94.2	201

Example rows of time series data from the London dataset showing daily levels of environmental variables

https://www.researchgate.net/figure/Example-rows-of-time-series-data-from-the-London-dataset-showing-daily-levels-of_tbl1_237840362

What is the Correlation of the table?

	A	B	C	D
Date	Ozone ($\mu g/m^3$)	Temperature (°C)	Relative humidity (%)	*n* deaths
1 Jan 2002	4.59	−0.2	75.7	199
2 Jan 2002	4.88	0.1	77.5	231
3 Jan 2002	4.71	0.9	81.3	210
4 Jan 2002	4.14	0.5	85.4	203
5 Jan 2002	2.01	4.3	93.5	224
6 Jan 2002	2.4	7.1	96.4	198
7 Jan 2002	4.08	5.2	93.5	180
8 Jan 2002	3.13	3.5	81.5	188
9 Jan 2002	2.05	3.2	88.3	168
10 Jan 2002	5.19	5.3	85.4	194
11 Jan 2002	3.59	3.0	92.6	223
12 Jan 2002	12.87	4.8	94.2	201

Leads to question:

What does it mean for one row to be similar to another?

Is what is the Correlation (A, B)

Example rows of time series data from the London dataset showing daily levels of environmental variables

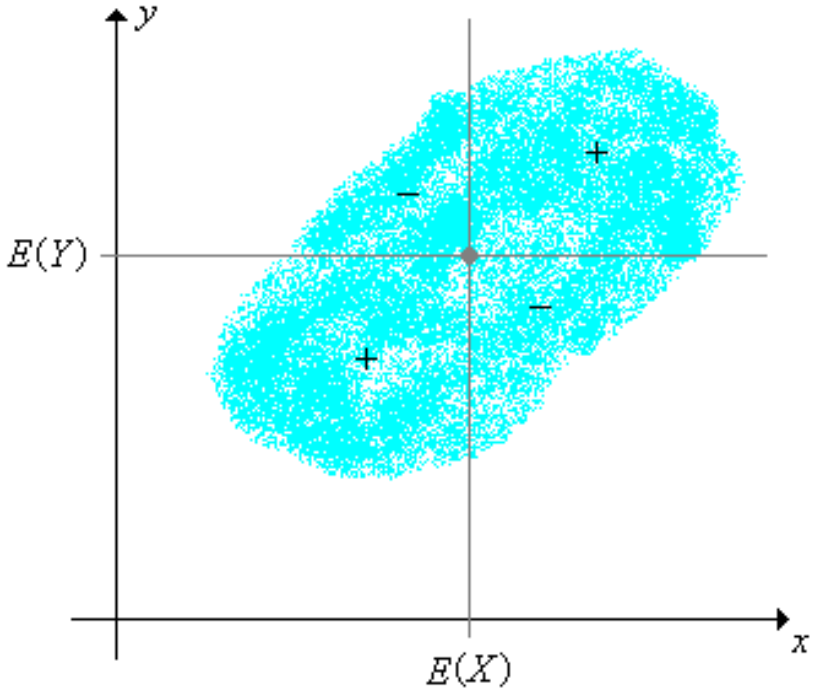
https://www.researchgate.net/figure/Example-rows-of-time-series-data-from-the-London-dataset-showing-daily-levels-of_tbl1_237840362

Correlation and Correlation Matrices

Correlation and Covariance: A practical Example

Data Table

X	Y
5	7
8	10
14	7
15	12
…	…
…	…
…	…
…	…
…	…



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

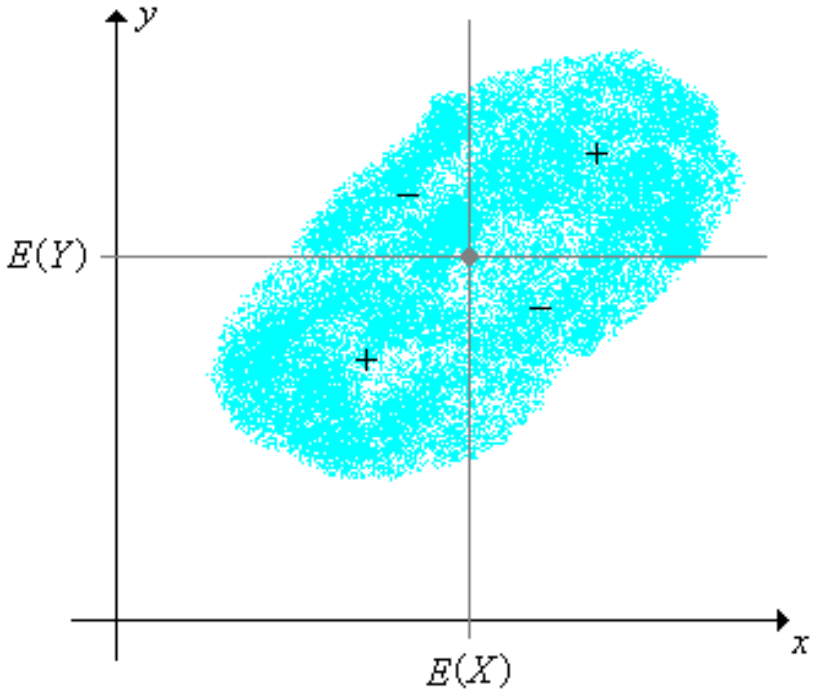
$$\rho_{X,Y} = \operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} =$$

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

Correlation: A practical Example

Data Table

X	Y
5	7
8	10
14	7
15	12
…	…
…	…
…	…
…	…
…	…



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

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

$$|Cov(X,Y)|^2 \le Var(X)Var(Y)$$

$$\therefore |Cov(X,Y)| \le \sqrt{Var(X)Var(Y)}$$

plug this result from the Cauchy-Schwarz

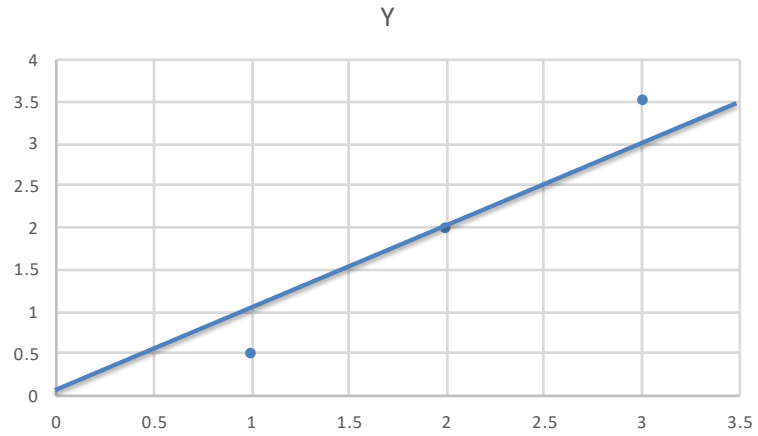
$$|\rho| = \left|\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}\right| \le \frac{\sqrt{Var(X)Var(Y)}}{\sqrt{Var(X)Var(Y)}} = 1$$

Example: What is the correlation of X,Y
How Do we Find It?

X	Y
1	0.5
2	2
3	3.5

Example: What is the correlation of X,Y
How Do we Find It?

X	Y
1	0.5
2	2
3	3.5



Example: What is the correlation of X,Y
How Do we Find It?

	X	Y
	1	0.5
	2	2
	3	3.5
mean	2.00	2.00
stdev	1	1.5
var	1	2.25

Example: What is the correlation of X,Y
One way to do it:

	Example				
	X	Y		X*Y	E[X]E[Y]
	1	0.5		0.5	4
	2	2		4	4
	3	3.5		10.5	4
mean	2.00	2.00		5.00	4.00
stdev	1	1.5			
var	1	2.25			

$$\text{Corr}(X,Y) = \frac{E[XY] - E[X]E[Y]}{\text{stdev}(X) * \text{stdev}(Y)}$$

$$= E[XY] - E[X]E[Y] / 1.5$$

$$= 5 - 4 / 1.5$$

$$= 1 / 1.5 = .67$$

Example: What is the correlation of X,Y
The other way to do it

	X	Y		X-ux	Y-uy	(X-ux)(Y-uy)
	1	0.5		-1	-1.5	1.5
	2	2		0	0	0
	3	3.5		1	1.5	1.5
mean	2.00	2.00		0.00	0.00	1.00
stdev	1	1.5				

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

= E[(X-ux)(Y-uy)] / 1.5

$$= \frac{[(1-2)(0.5-2) + (2-2)(2-2) + (3-2)(3.5-2)]/3}{1.5}$$

= 1.5 + 0 + 1*1.5 / (3* 1.5) = 1/1.5 = .67

Correlation and Covariance

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

Properties

a. $-1 \leq \operatorname{cor}(X,Y) \leq 1$

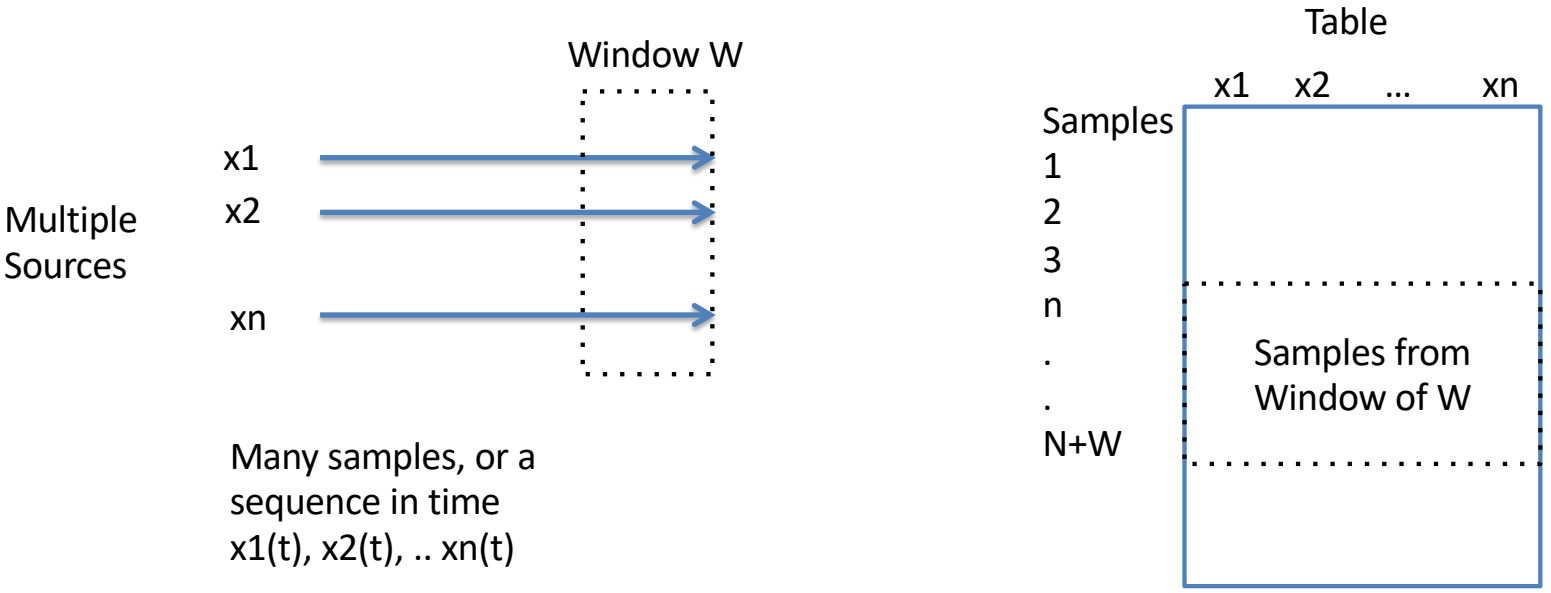
b. $-\operatorname{sd}(X)\operatorname{sd}(Y) \leq \operatorname{cov}(X,Y) \leq \operatorname{sd}(X)\operatorname{sd}(Y)$

c. $\operatorname{cor}(X,Y) = 1$ if and only if Y is a linear function of X with positive sloper.

d. $\operatorname{cor}(X,Y) = -1$ 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

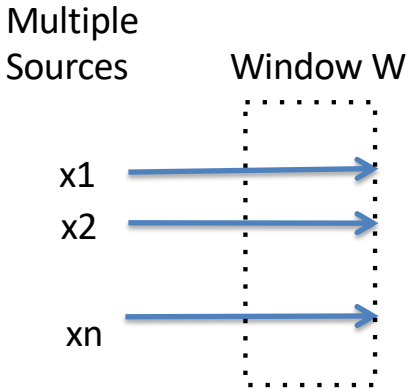


To estimate from data:

a) Use all samples ever collected

b) Use window size of W samples of each to estimate a recent Corr Matrix

Correlation Matrix



$$\begin{pmatrix} \text{Corr}(x1,x1) \ldots\ldots & \text{Corr}(x1,xn) \\ \text{Corr}(xn,x1) & \text{Corr}(xn,xn) \end{pmatrix}$$

Corr(x1,x1) = 1

Corr(x2, x1) = a number from -1 to 1

To estimate from data:

a) Use all samples ever collected
b) 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

	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() method

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

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 pa

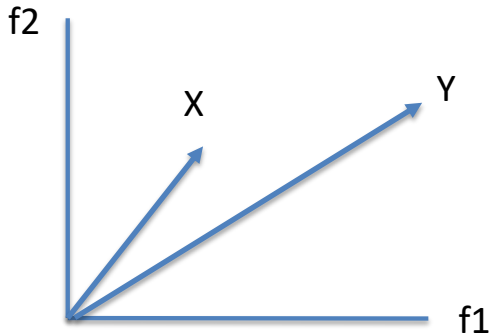
Currently only available for pearson and spearma

Returns:

y : *DataFrame*

Data Table

X	Y
5	7
8	10
14	7
15	12
…	…
…	…
…	…
…	…
…	…



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

Pearson: Understanding
Correlation in a different way

Use n dimensions

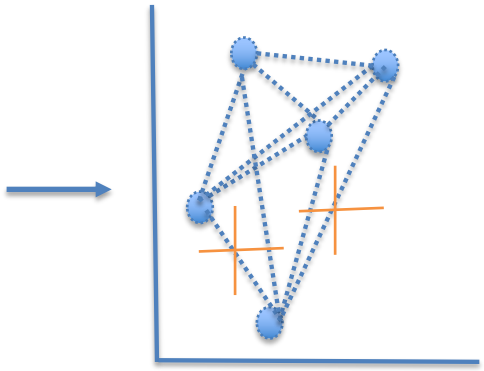
Pandas will create a correlation matrix with “columns”

```
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
```

Kendall Correlation

List of (x,y) points

No	X	Y
1	2	3
2	4	6
3	3	8
4	9	12



N points
N(n-1)/2 pairs of x,y points

Concordant pairs: for (x_i, y_i) and (x_j, y_j), where $i \neq j$,

$x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$

Disconcordant pairs: when the above is not true

if $x_i > x_j$ and $y_i < y_j$

or if $x_i < x_j$ and $y_i > y_j$

The Kendall τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

Spearman Correlation

$$r_s = \rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}$$

where

- ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables.
- $\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)$ is the covariance of the rank variables.
- σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

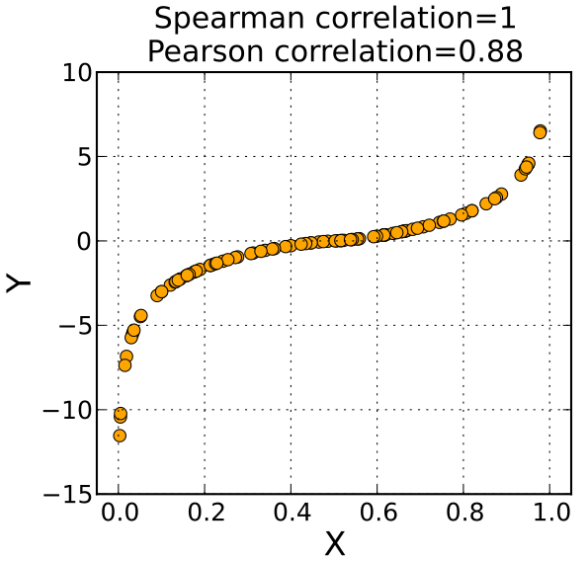
Data (x=IQ,y=TV)

IQ, X_i	Hours of TV per week, Y_i
106	7
86	0
100	27
101	50
99	28
103	29
97	20
113	12
112	6
110	17

Order rows by X and Index X and Y in increasing order

x	y	rgx	rgy		
97	20	2	6	–4	16
99	28	3	8	–5	25
100	27	4	7	–3	9
101	50	5	10	–5	25
103	29	6	9	–3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Then find Pearson Correlation of (rgx,rgy)

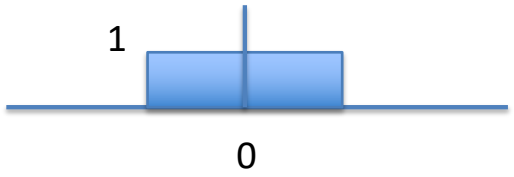


A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data-points with greater x-values than that of a given data-point will have greater y-values as well. In contrast, this does not give a perfect Pearson correlation.

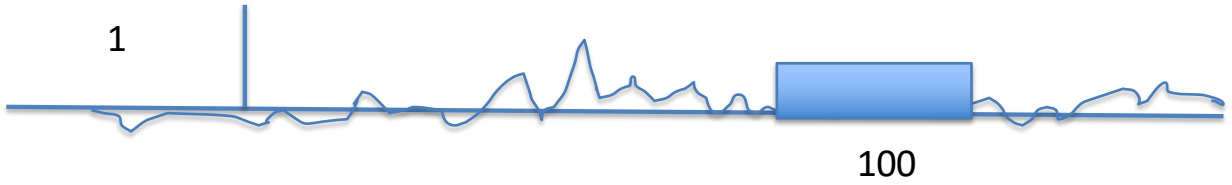
Wikipedia

Correlation Matrix with multiple sources and time segments

Suppose this is x1
as an array of numbers 0 0 1 1..1 0 0 0

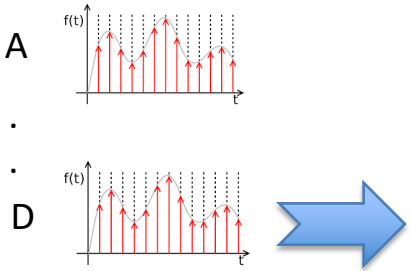


Suppose this is x2



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

Approaches to the Data Sequences from Multiple Sources in Tables



Discrete data for each source
A, B, C..
x_n = x1, x2, x3, …

One row for each source
A, B, C, D..

Eg Numpy arrays

Time ->

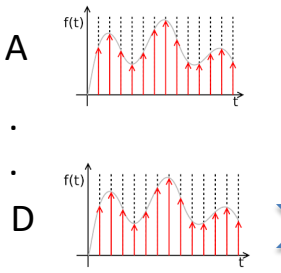
A:	x1, x2, x3, …
B:	x1, x2, x3, …
C:	x1, x2, x3, …
D:	x1, x2, x3, …

One column for each series:
A , B, C, D..

eg Pandas
(add rows with every new time sample)

	A	B	C
1	x1	x1	x1
2	x2	x2	x2
3	x3	x3	x3
.	.	.	.
.	.	.	.

Approaches to the Data Sequences in Tables



Discrete data for each source
A, B, C..
x_n = x1, x2, x3, …

One row for each source
A, B, C, D..

Eg Numpy arrays

Time ->

A:	x1,	x2,	x3,	…
B:	x1,	x2,	x3,	…
C:	x1,	x2,	x3,	…
D:	x1,	x2,	x3,	…

One column for each series:
A , B, C, D..

eg Pandas
(add rows with every new time sample)

	A	B	C
1	x1	x1	x1
2	x2	x2	x2
3	x3	x3	x3
.	.	.	.
.	.	.	.

Preprocess:
Filter/
Convolution
Correlation
Mean, var, stats

Windows of time

Features

If ML:
Table of (X,Y)
for
Prediction
Classification

Data Input and Storage | Preprocessing | ML for Decisions

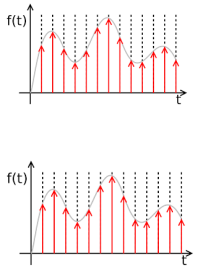
Approaches to the Data Sequences in Tables

A

.

.

D



Discrete data for each source
A, B, C..
x_n = x1, x2, x3, …

One row for each source
A, B, C, D..

Eg Numpy arrays

Time ->

A:	x1,	x2,	x3,	…
B:	x1,	x2,	x3,	…
C:	x1,	x2,	x3,	…
D:	x1,	x2,	x3,	…

One column for each series:
A , B, C, D..

eg Pandas
(add rows with every new time sample)

	A	B	C
1	x1	x1	x1
2	x2	x2	x2
3	x3	x3	x3
.	.	.	.
.	.	.	.

						Labeled
	Cor(w/A)	Cor(w/B)	St.Dev.	Cor(w/A[n-20])		Condition
A	#	#	#	#		Danger
B	#	#	#	#		Safe
C	#	#	#	#		Safe
D	#	#	#	#		Warning
# = number obtained from preprocessing						

Example: pre-processed statistics can be used for in ML predictions

Data Input and Storage

End of Section