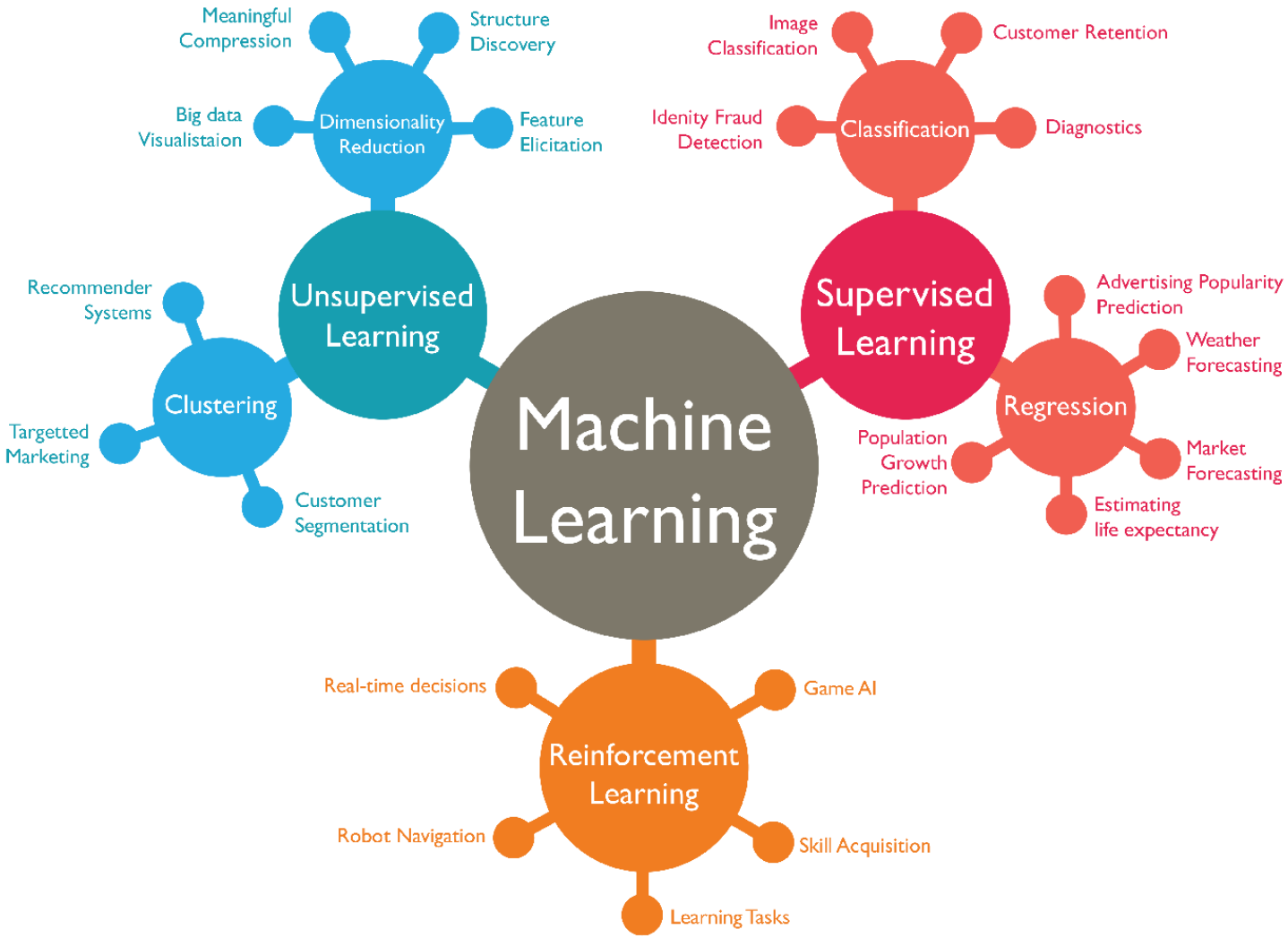
Today Focus !!



INVESTIC

Next

Learn to

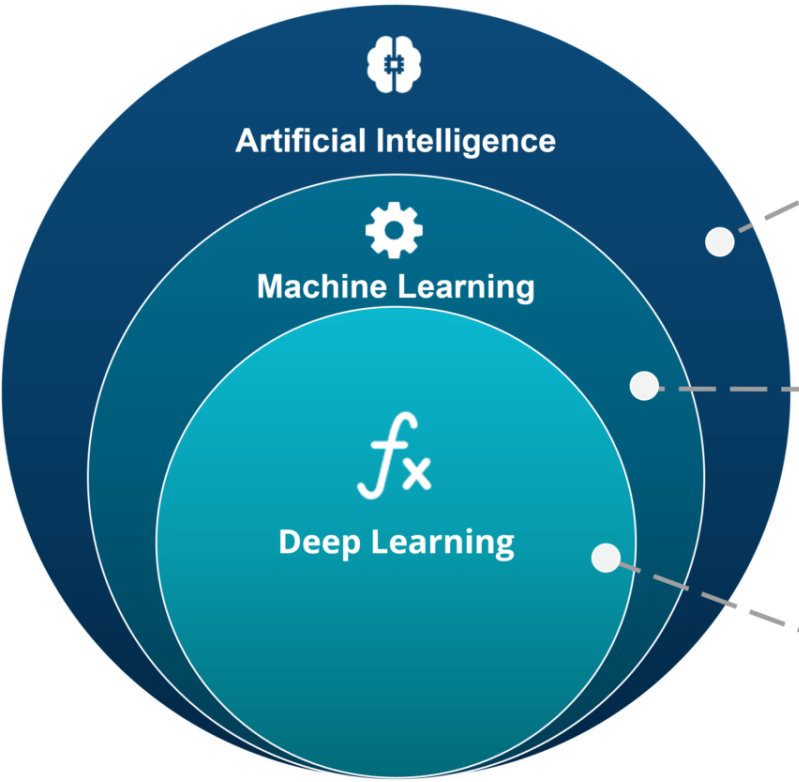
1. Import
2. Data Cleaning
3. Data Wrangling

First Week → EDA part

4. Modeling
    1. Unsupervised
    2. Supervised
5. Evaluate & Backtesting

INVESTIC

What is Machine Learning



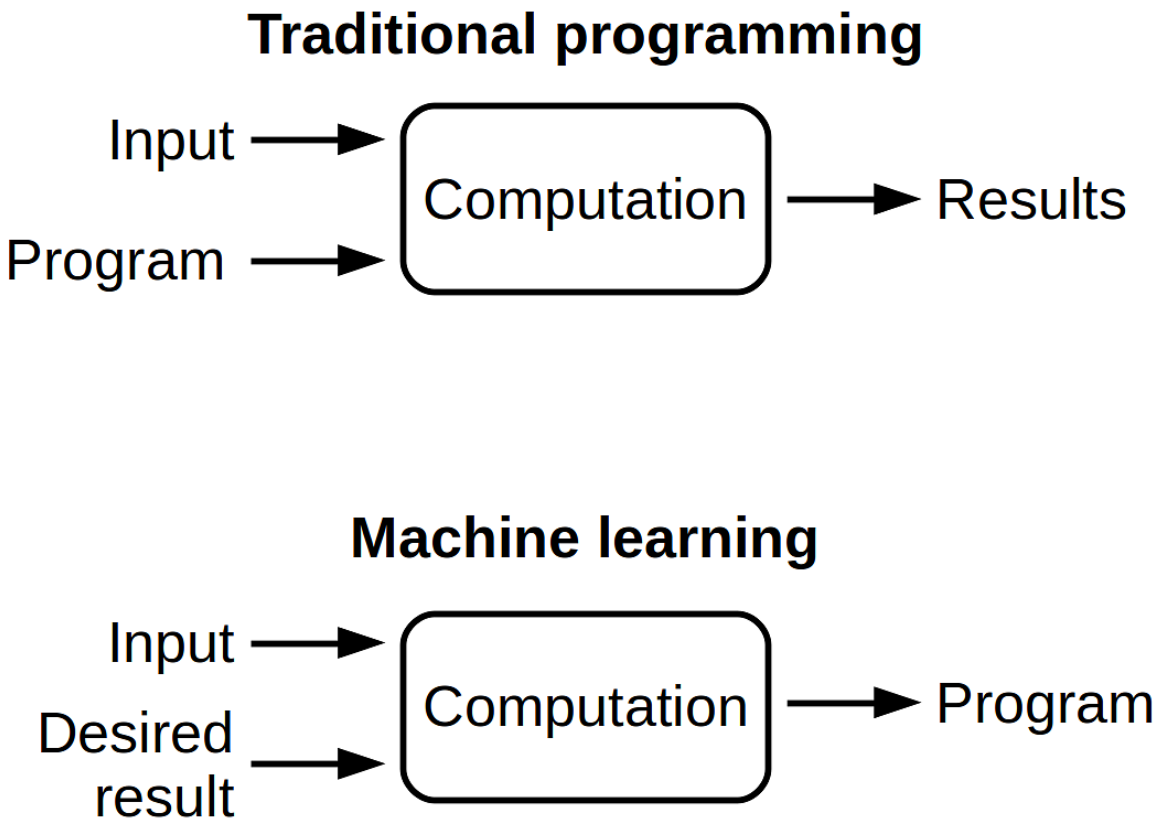
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

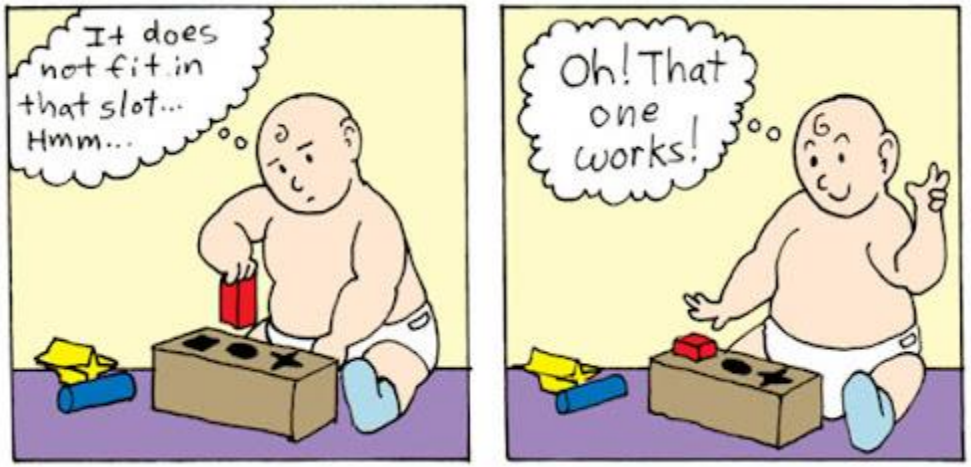
INVESTIC

Traditional Algorithm vs Machine Learning



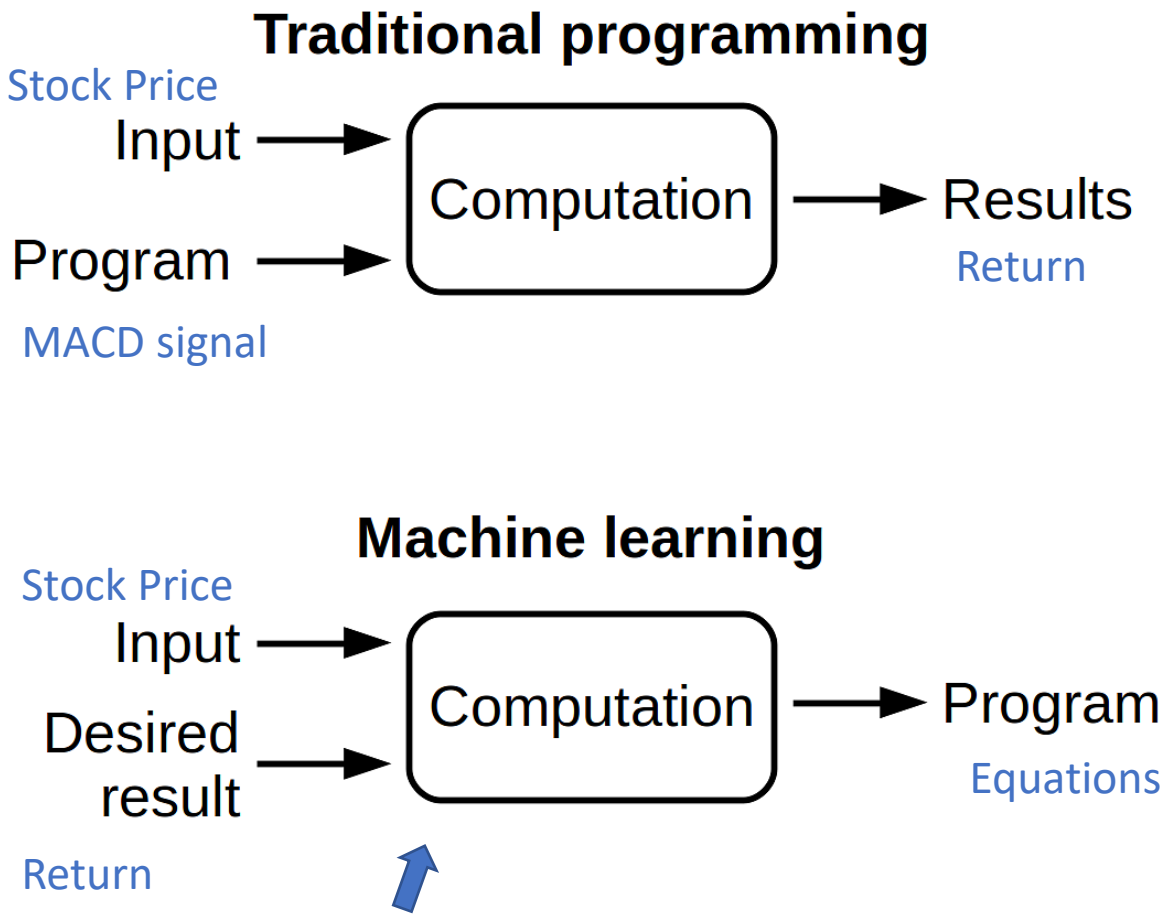
INVESTIC

Machine Learning is process of Trial and Error



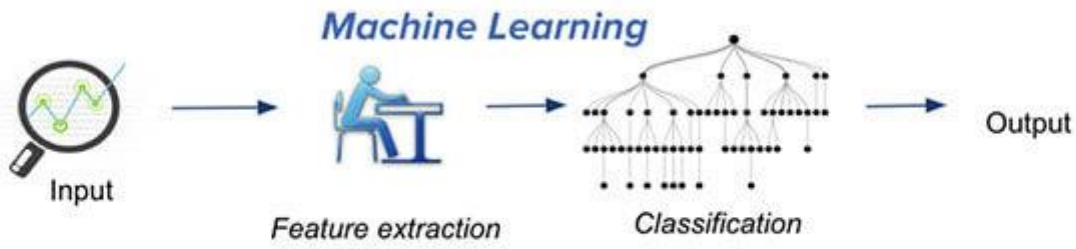
INVESTIC

Traditional Algorithm vs Machine Learning

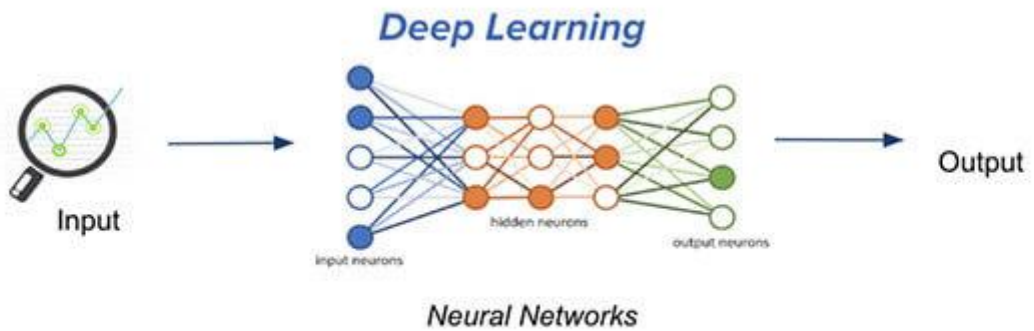


INVESTIC

Machine Learning vs Deep Learning



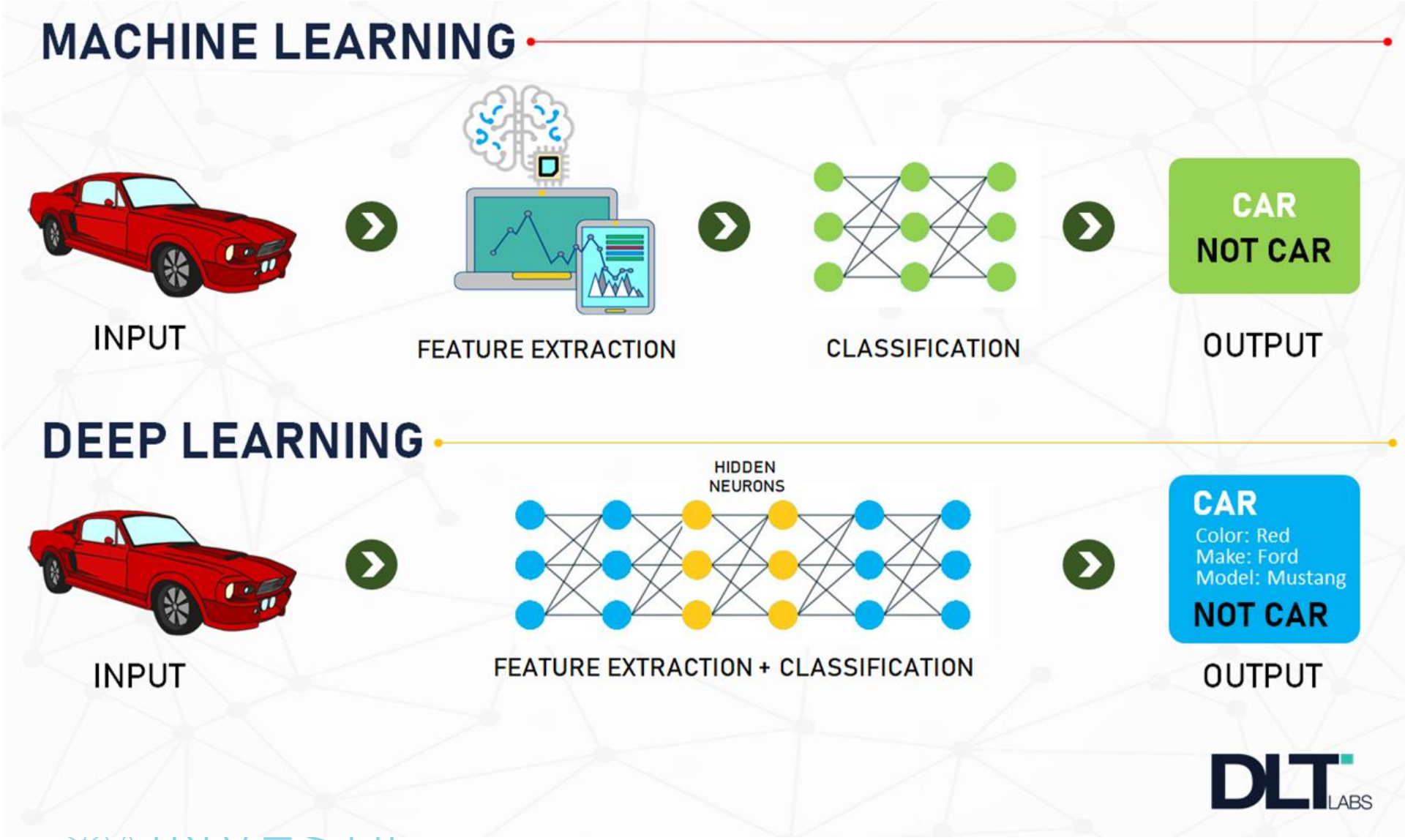
Traditional machine learning uses hand-crafted features, which is tedious and costly to develop.



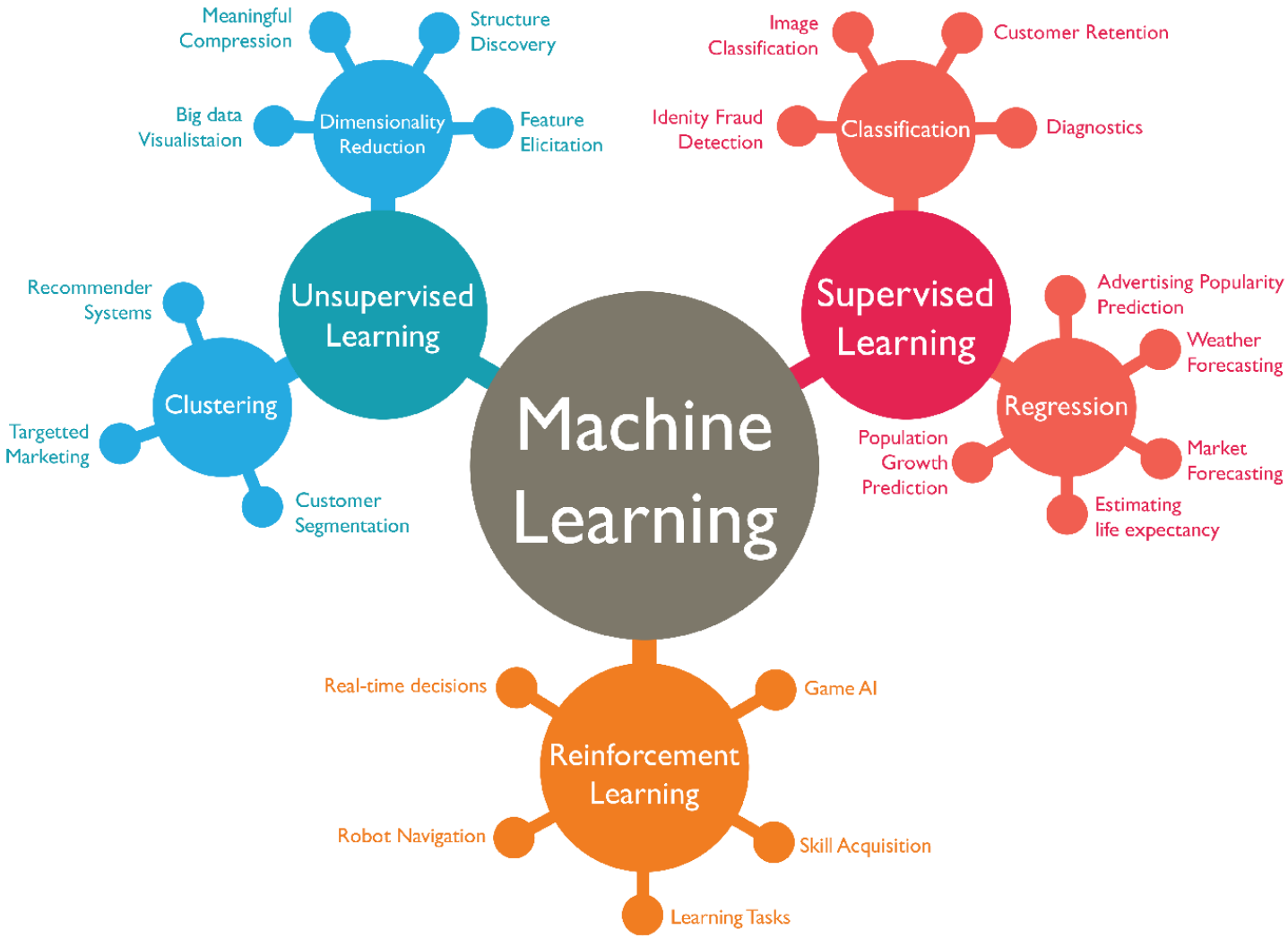
Deep learning learns hierarchical representation from the data itself, and scales with more data.

INVESTIC

Machine Learning vs Deep Learning

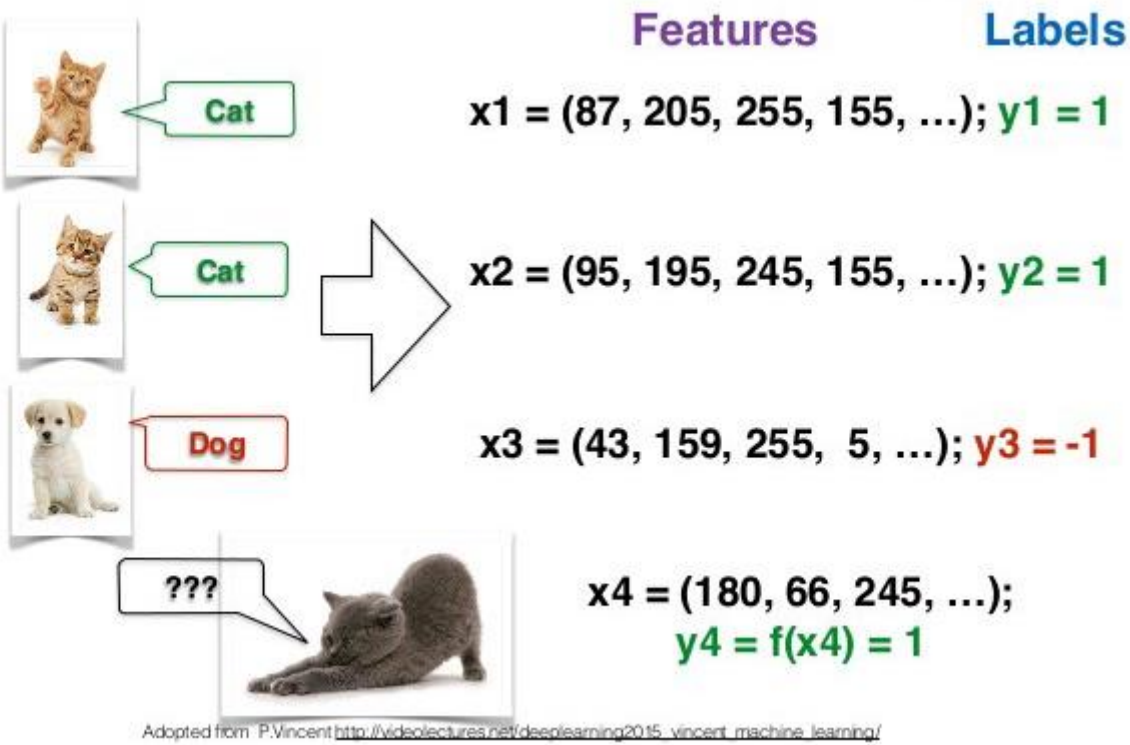


Classified by Type

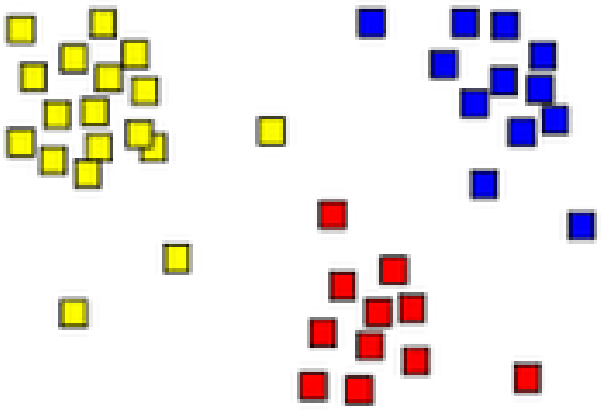


INVESTIC

Supervised Learning



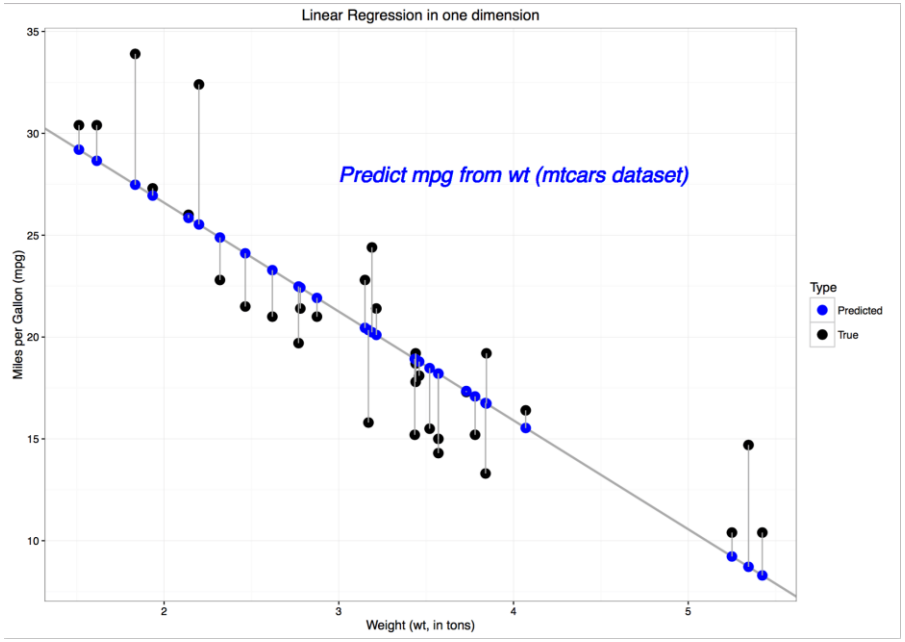
Unsupervised Learning



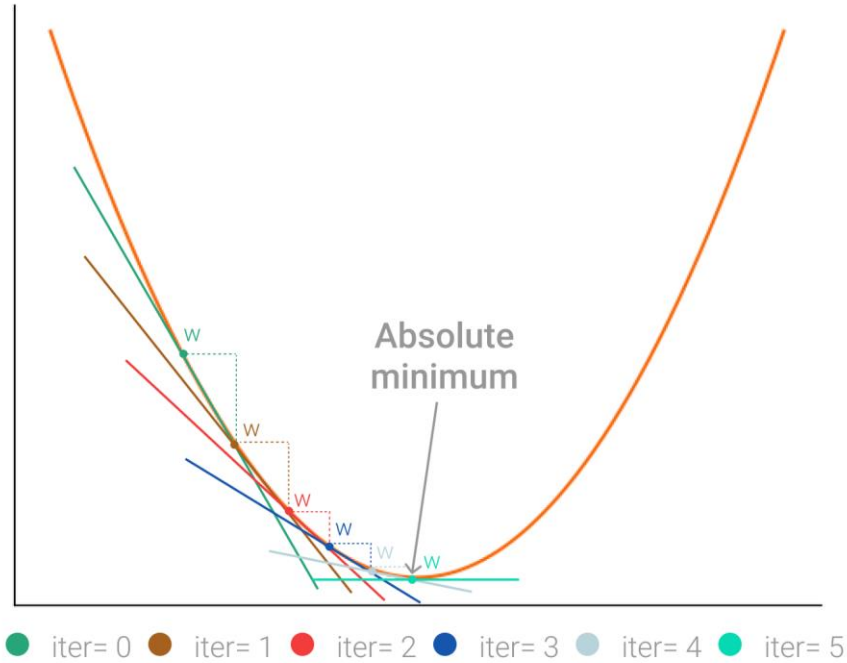
Reinforcement Learning

Second Layer of ML type

Say … Deviation from Mean

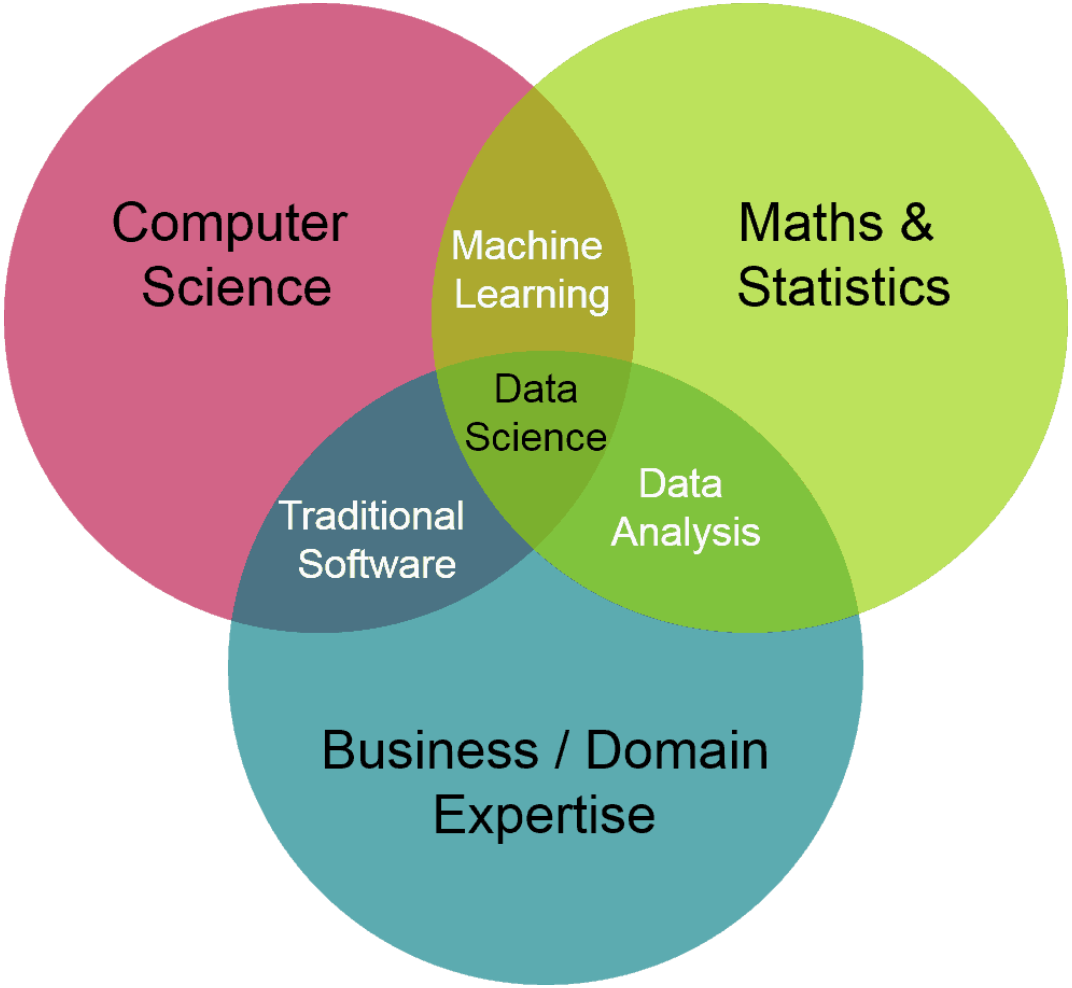


Gradient Descent



INVESTIC

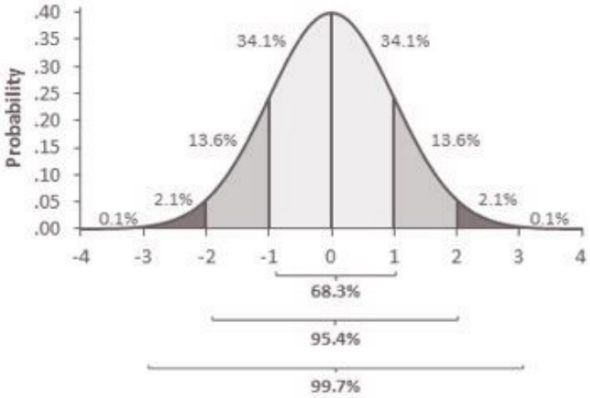
3 Must know



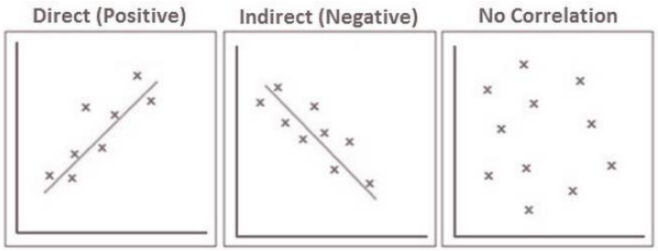
INVESTIC

Math & Stats

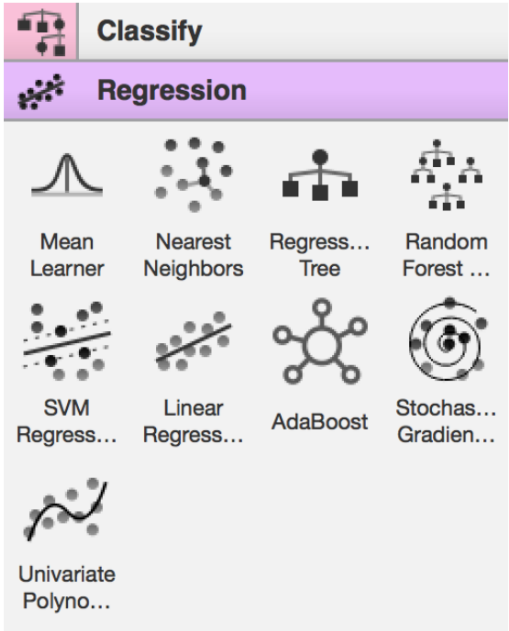
Stats & Probability



Relationship

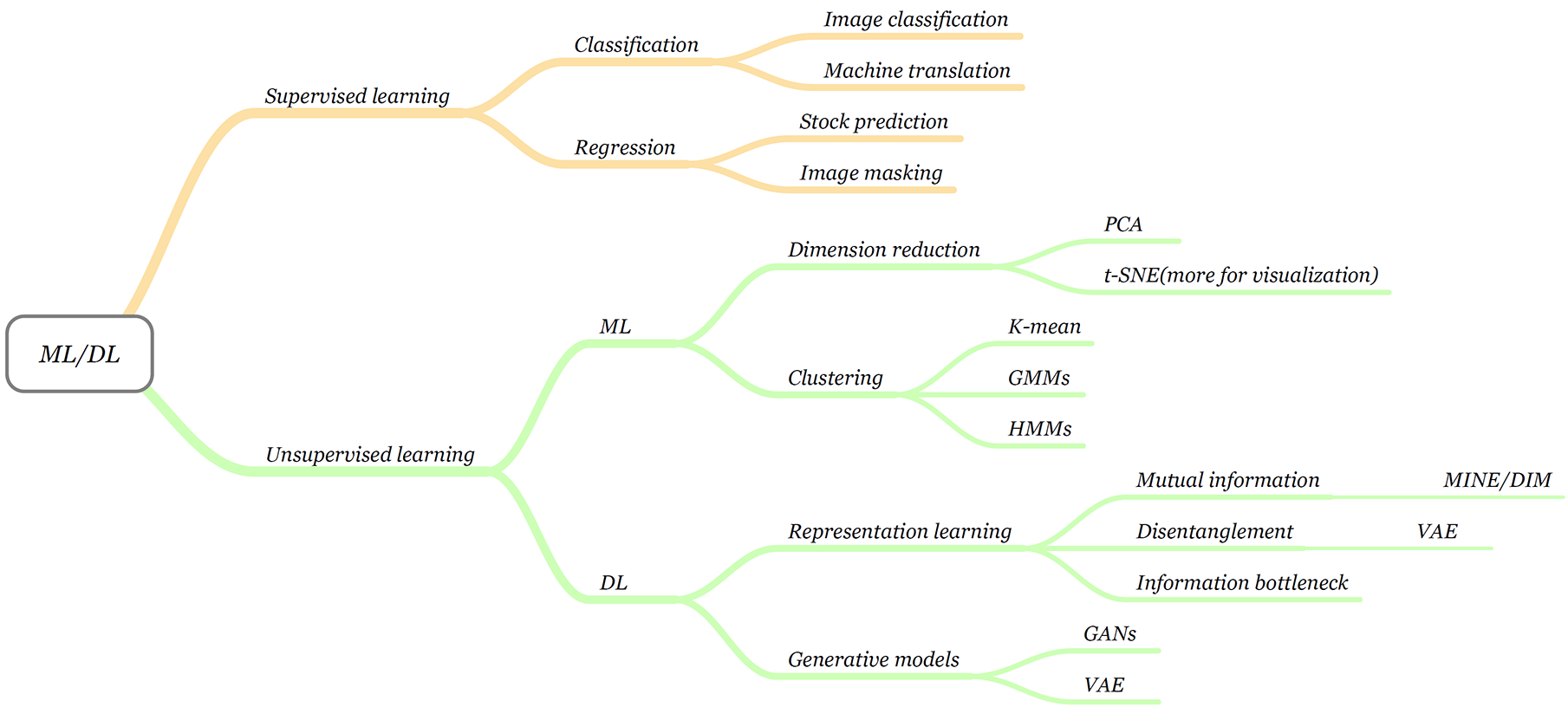


Modeling



INVESTIC

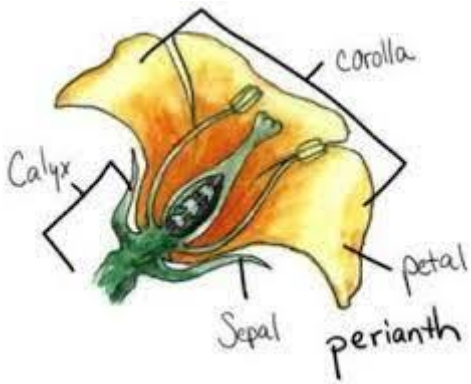
Modeling

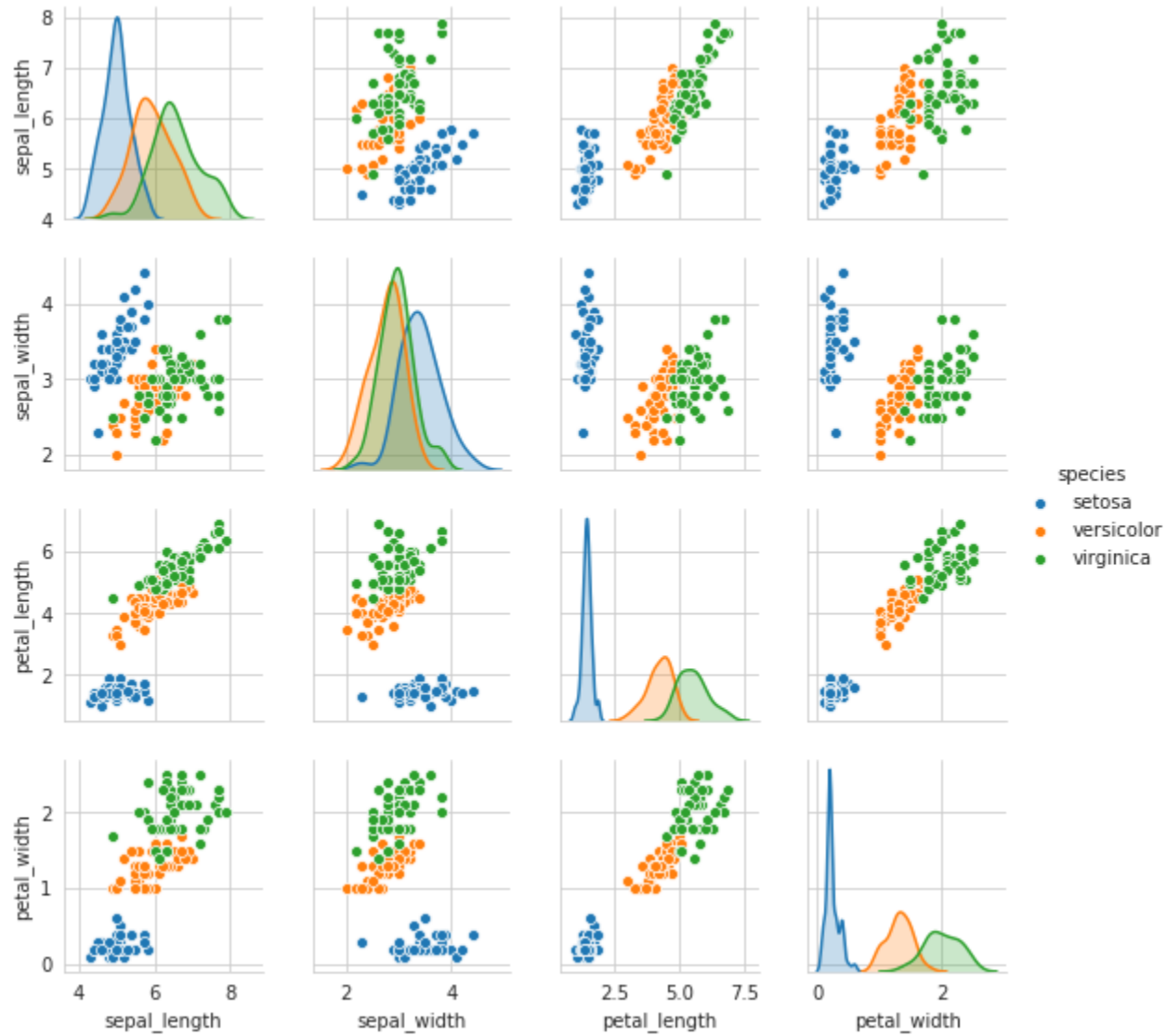

ML/DL
Supervised learning
Classification
Image classification
Machine translation
Regression
Stock prediction
Image masking
Unsupervised learning
ML
Dimension reduction
PCA
t-SNE(more for visualization)
Clustering
K-mean
GMMs
HMMs
DL
Representation learning
Mutual information
MINE/DIM
Disentanglement
VAE
Information bottleneck
Generative models
GANs
VAE

So how to interpret this thing …

กลีบเลี้ยง

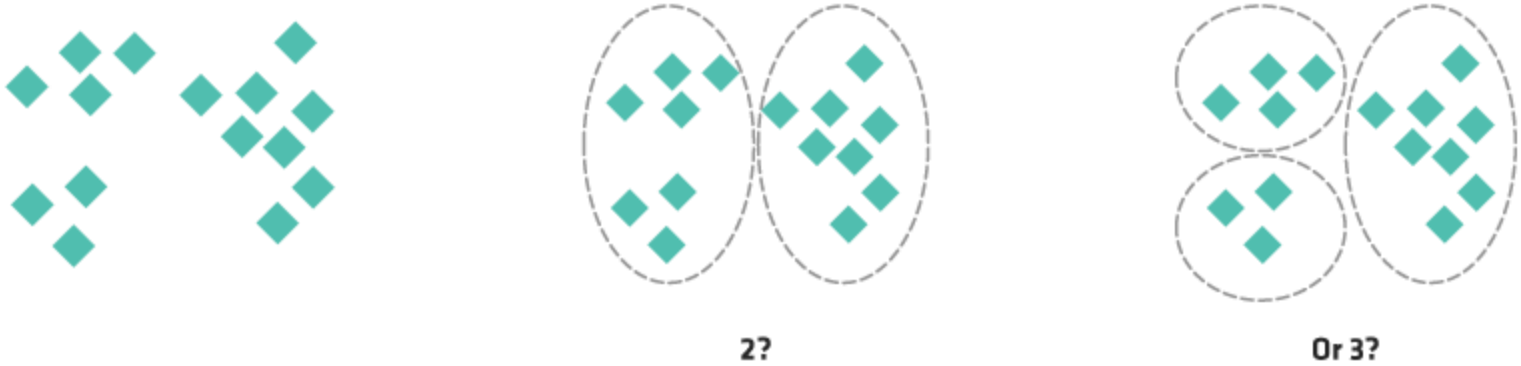
กลีบดอก





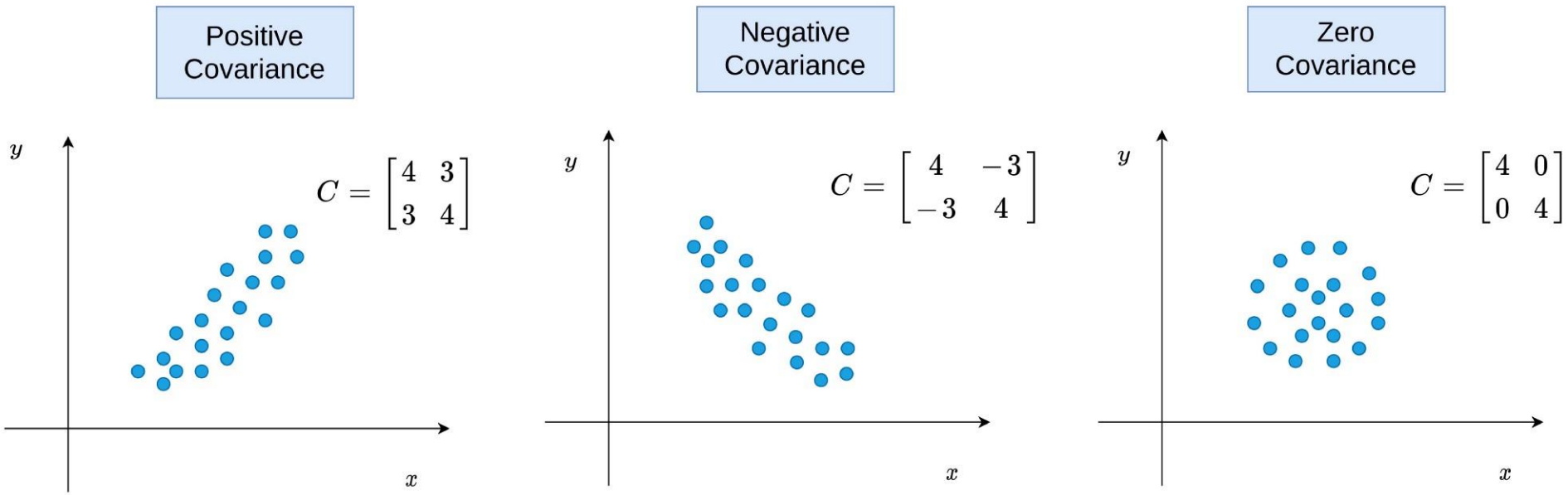
What you see in tutorial

How Many Clusters Do You See?



Source: AllianceBernstein (AB)

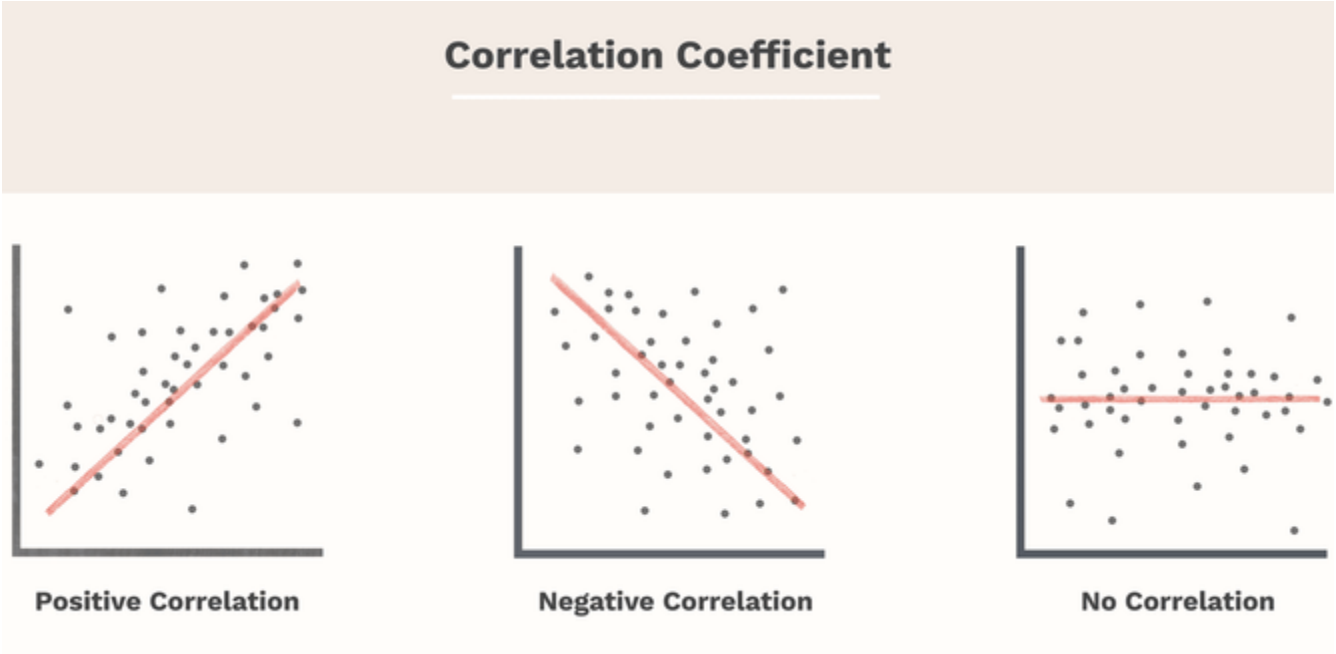
Hello Theory



Hi Covariance Matrix

	Government bond	Muni bond	High yield bond	Real estate	MBS	Investment grade bond	Large cap stock
Government bond	0.031%	0.040%	0.048%	0.037%	0.039%	0.026%	-0.010%
Muni bond	0.040%	0.123%	0.074%	0.070%	0.063%	0.039%	-0.010%
High yield bond	0.048%	0.074%	0.335%	0.107%	0.077%	0.047%	0.002%
Real estate	0.037%	0.070%	0.107%	0.193%	0.091%	0.026%	0.029%
MBS	0.039%	0.063%	0.077%	0.091%	0.112%	0.033%	0.022%
Investment grade bond	0.026%	0.039%	0.047%	0.026%	0.033%	0.032%	-0.007%
Large cap stock	-0.010%	-0.010%	0.002%	0.029%	0.022%	-0.007%	0.099%
Std. dev	1.775%	3.511%	5.789%	4.395%	3.351%	1.799%	3.141%
Skewness	-0.782	-1.374	-0.001	-2.147	-1.133	-0.774	-0.521

So how to interpret this thing …

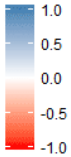


Correlation Matrix

Correlation matrix of real normalized FX carry

based on monthly data from 2000-01-31/2009-08-31 to 2018-04-30/2018-05-31

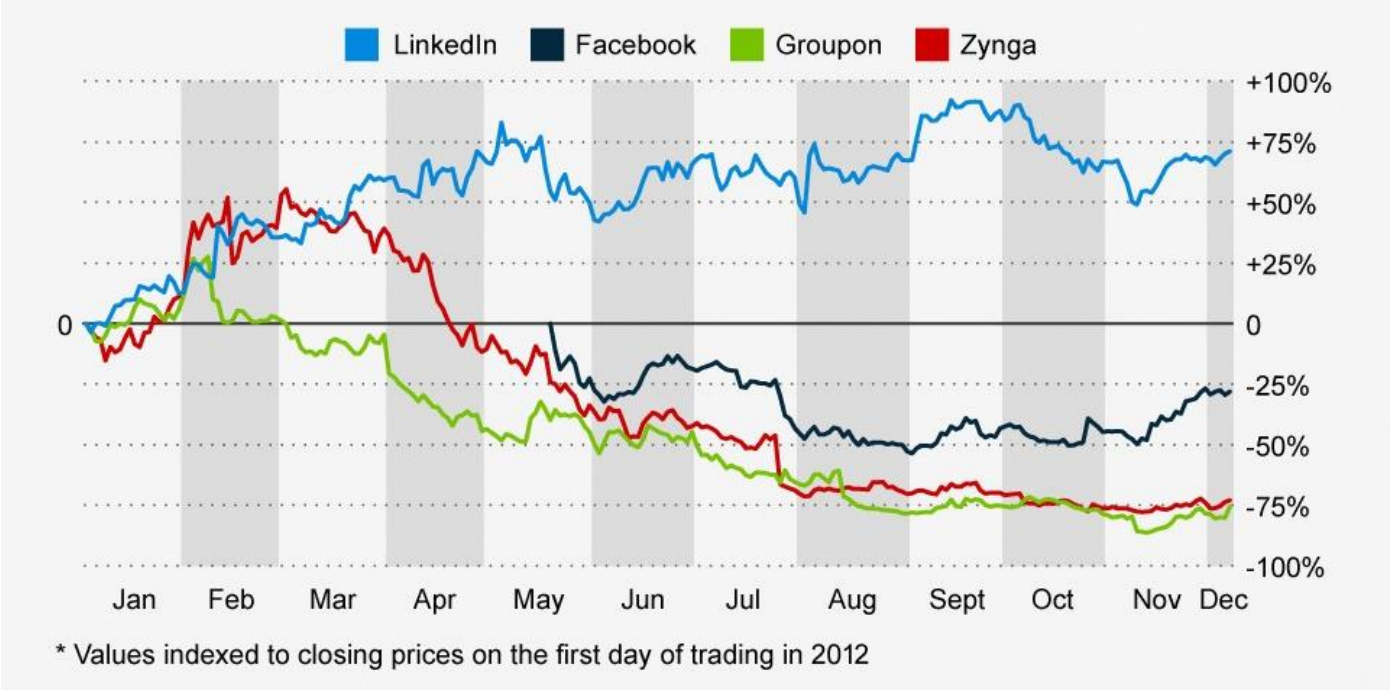
	BRL	CLP	COP	CZK	HUF	IDR	ILS	INR	KRW	MXN	MYR	PEN	PHP	PLN	RON	RUB	THB	TRY	TWD	ZAR
ZAR	0.2	0.1	0.2	0	-0.5	0.5	0.7	0.3	0.5	0.5	0.2	0.2	0.4	0.5	0.6	-0.5	0.3	0.3	0.2	
TWD	0.2	-0.1	0	-0.3	0	0.4	0.2	0.4	0.1	0.5	-0.1	0.4	0.4	0.4	0.2	0.4	0.4	0.1		
TRY	0.5	0.2	0.4	0.1	0.6	0.3	-0.1	0.7	0.2	0.1	0.2	0.6	0.3	0.5	-0.1	0.5	0			
THB	0.4	0.3	0.4	-0.3	0	-0.4	0.6	0.1	0.2	0.1	0.4	0.5	0.4	0.2	0.5	0.4				
RUB	0.1	0.1	0.6	0.5	0.7	-0.1	0	0.5	-0.1	0.2	0.2	0.3	0.3	0.4	0.2					
RON	0.5	0.3	0	0.1	0.2	0	0.2	0.1	0.2	0.3	0.1	0.2	0.2	0.2						
PLN	0.3	0.1	0.2	0.2	0.1	0.5	0.3	0.6	0.1	0.2	0.1	0.2	0.2							
PHP	0.2	0.4	0.1	0.4	0.5	0.2	0.3	0.3	0.3	0.3	0.4	0.4								
PEN	0.5	0.2	0.1	0.4	-0.2	0.5	0.3	0.6	0.3	0.3	0.4									
MYR	0.3	0.2	0	0.4	0.5	0.4	0	0.3	0.3	0.3										
MXN	0.1	0.1	0.2	0.1	0.2	0.5	0.3	0.3	0.1											
KRW	0.5	0.2	0.2	0.3	0.1	0.5	0.6	0.3												
INR	0.2	0.4	0.2	0.1	0.5	0.1	0.3													
ILS	0.3	0.1	0.2	0.2	0.1	0.6														
IDR	0.4	0.3	0.2	0.1	0.3															
HUF	0.2	0.1	0	0.4																
CZK	0	-0.1	0																	
COP	0.2	0.2																		
CLP	0.1																			
BRL																				



© Macrosynergy LLP 2018

INVESTIC

LinkedIn no correlation??



Analysis

$$Corr(x,y) = \frac{\sum_{i=1}^{n}(x_i - x')\,(y_i - y')}{\sqrt{\sum_{i=1}^{n}(x_i - x')^2 \sum_{i=1}^{n}(y_i - y')^2}}$$

$$\mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_x \sigma_y}$$

Covarianced normalized by Standard Deviation

Correlation between X and Y

Standard deviation of X

Standard deviation of Y

Covariance Formula

For Population

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N - 1)}$$

Analysis

POPULATION	SAMPLE
The measurable quality is called a parameter.	The measurable quality is called a statistic.
The population is a complete set.	The sample is a subset of the population.
Reports are a true representation of opinion.	Reports have a margin of error and confidence interval.
It contains all members of a specified group.	It is a subset that represents the entire population.

QuestionPro

Population mean is known

$$var(x) = \frac{\sum_i^n (x_i - \mu)^2}{N}$$

Population mean is unknown

$$var(x) = \frac{\sum_i^n (x_i - \overline{x})^2}{N - 1}$$

Correlation …

FIGURE 3

UK bond-equity correlation

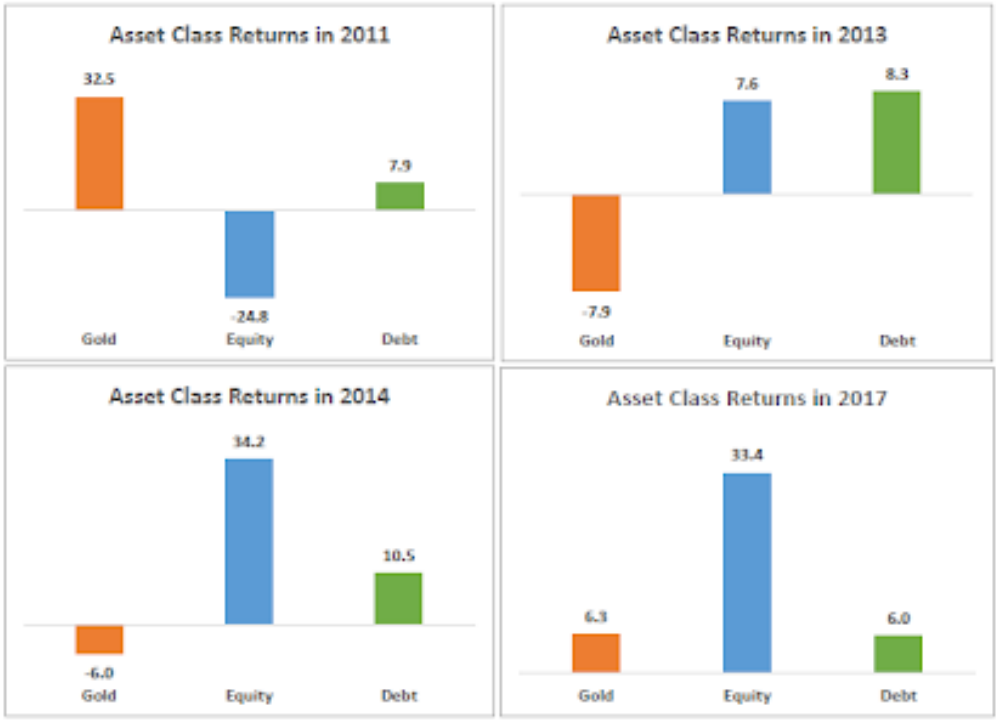
Ten-year trailing correlation of monthly returns on UK equities and consols. | Sources: Thomas and Dimsdale (2016), Bank of England calculations

Past Performance is not necessarily indicative of future results

INVESTIC

Correlation …

Asset Class Returns can be diverse

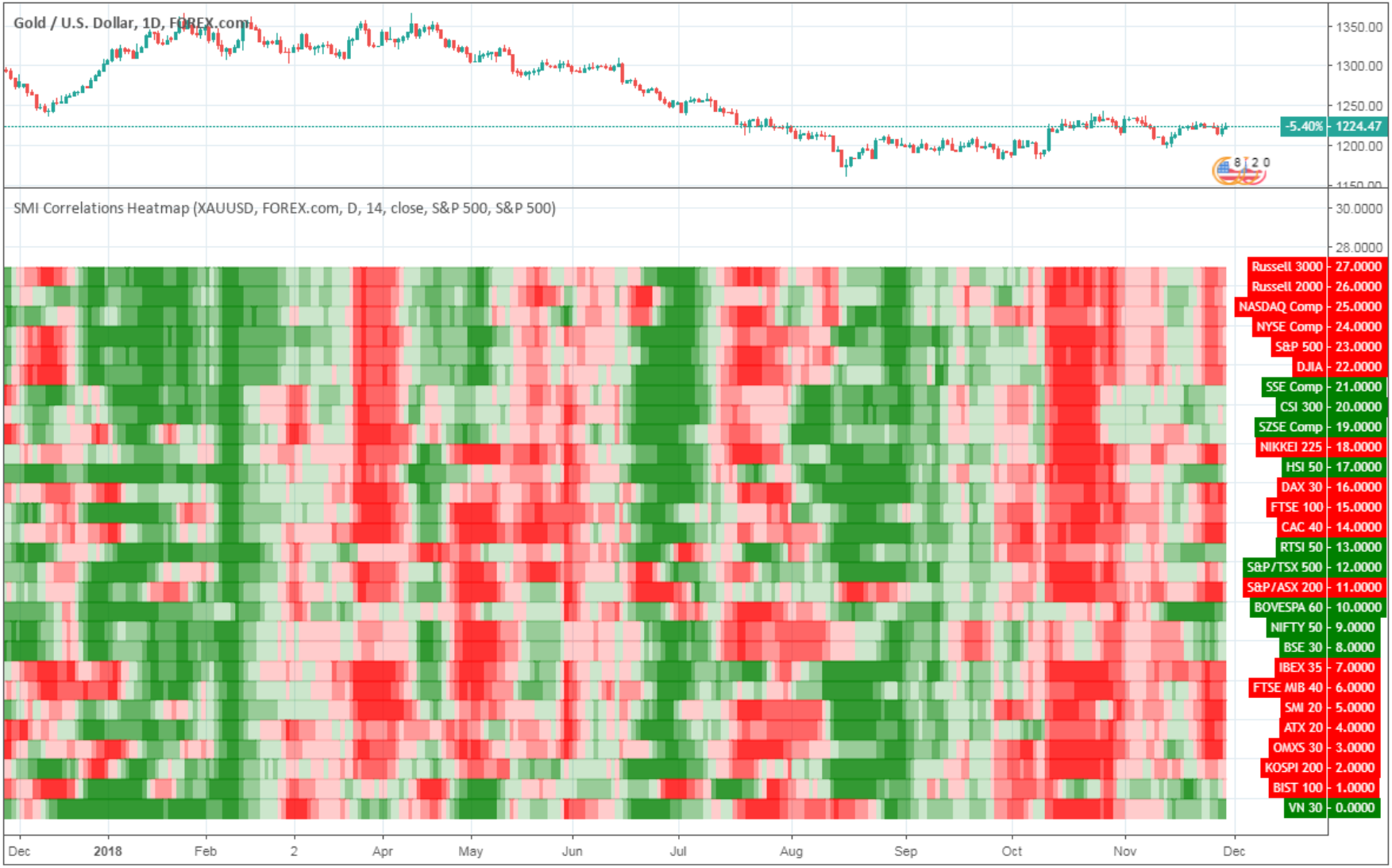


Note: 1) Gold Futures prices from MCX; 2) For Equity, S&P BSE 100 TRI returns are considered; 3) For Debt, CRISIL Short Term Bond Fund Index returns are considered; 4) Source: Bloomberg, MFI Explorer

Correlation …

Correlation …

everget published on TradingView.com, November 29, 2018 22:25 UTC
FOREXCOM:XAUUSD, 1D 1224.47 ▲ +3.40 (+0.28%) **O:**1221.09 **H:**1229.03 **L:**1220.79 **C:**1224.47



Created with TradingView

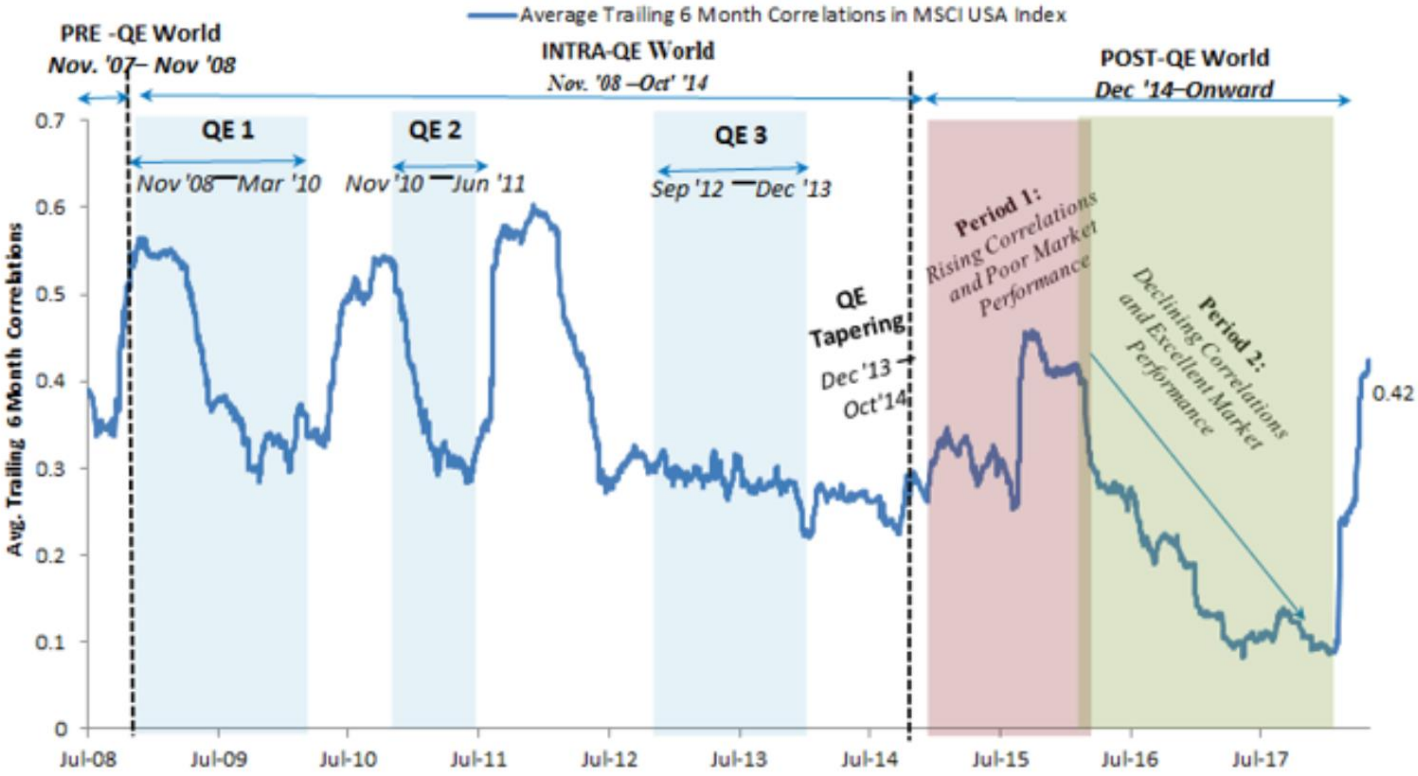
Acceptable Correlation Matrix

Figure 2: Five-Year Weekly Correlation Of The S&P Sectors Represented By SPDR ETFs

Sector	Ticker	XLC	XLY	XLP	XLE	XLF	XLV	XLI	XLB	XLRE	XLK	XLU
Communications Services	XLC	1.00	0.82	0.53	0.69	0.64	0.68	0.67	0.60	0.43	0.81	0.24
Consumer Discretionary	XLY	0.82	1.00	0.62	0.61	0.73	0.71	0.78	0.70	0.47	0.84	0.21
Consumer Staples	XLP	0.53	0.62	1.00	0.39	0.49	0.57	0.58	0.51	0.62	0.64	0.53
Energy	XLE	0.69	0.61	0.39	1.00	0.63	0.49	0.69	0.68	0.22	0.57	0.05
Financials	XLF	0.64	0.73	0.49	0.63	1.00	0.63	0.84	0.75	0.28	0.67	0.08
Health Care	XLV	0.68	0.71	0.57	0.49	0.63	1.00	0.67	0.59	0.46	0.69	0.29
Industrials	XLI	0.67	0.78	0.58	0.69	0.84	0.67	1.00	0.85	0.35	0.76	0.19
Materials	XLB	0.60	0.70	0.51	0.68	0.75	0.59	0.85	1.00	0.36	0.69	0.14
Real Estate	XLRE	0.43	0.47	0.62	0.22	0.28	0.46	0.35	0.36	1.00	0.46	0.67
Technology	XLK	0.81	0.84	0.64	0.57	0.67	0.69	0.76	0.69	0.46	1.00	0.26
Utilities	XLU	0.24	0.21	0.53	0.05	0.08	0.29	0.19	0.14	0.67	0.26	1.00
S&P 500 Index		**0.84**	**0.91**	**0.71**	**0.71**	**0.84**	**0.82**	**0.89**	**0.80**	**0.52**	**0.92**	**0.30**

Analysis

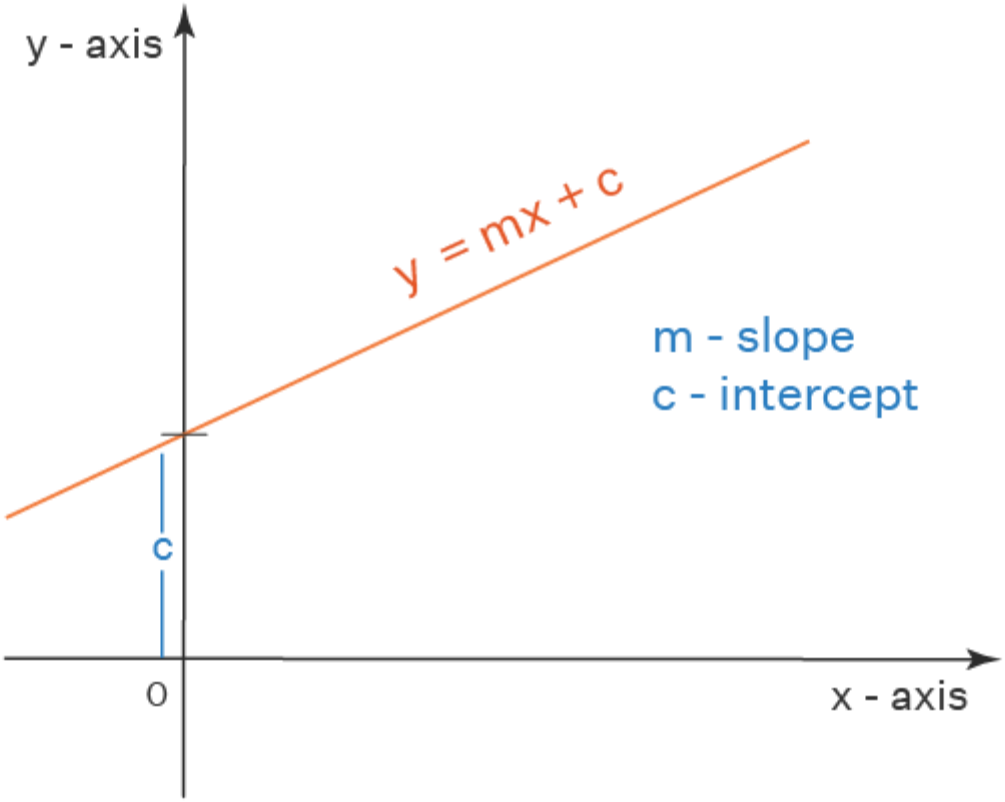
Significant Pickup in Average Stock Correlations



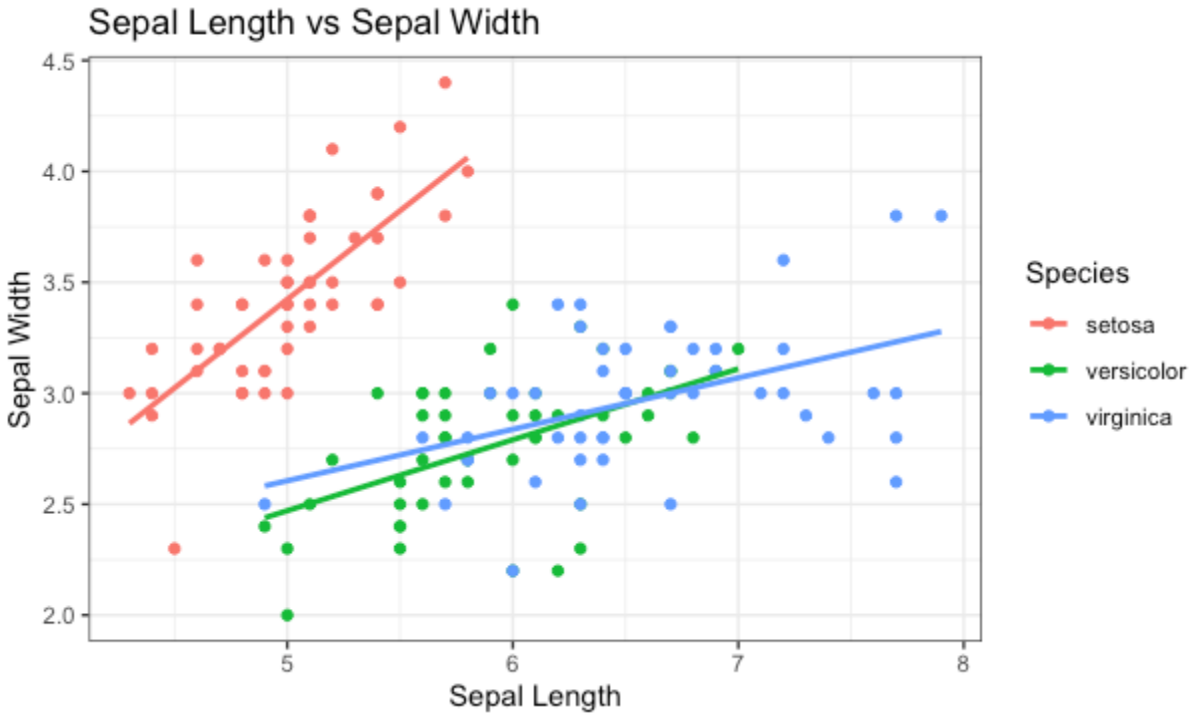
Sources: WisdomTree, Bloomberg, MSCI, as of 4/30/18. Past performance is not indicative of future results. You cannot invest directly in an index.

What you really need to understand

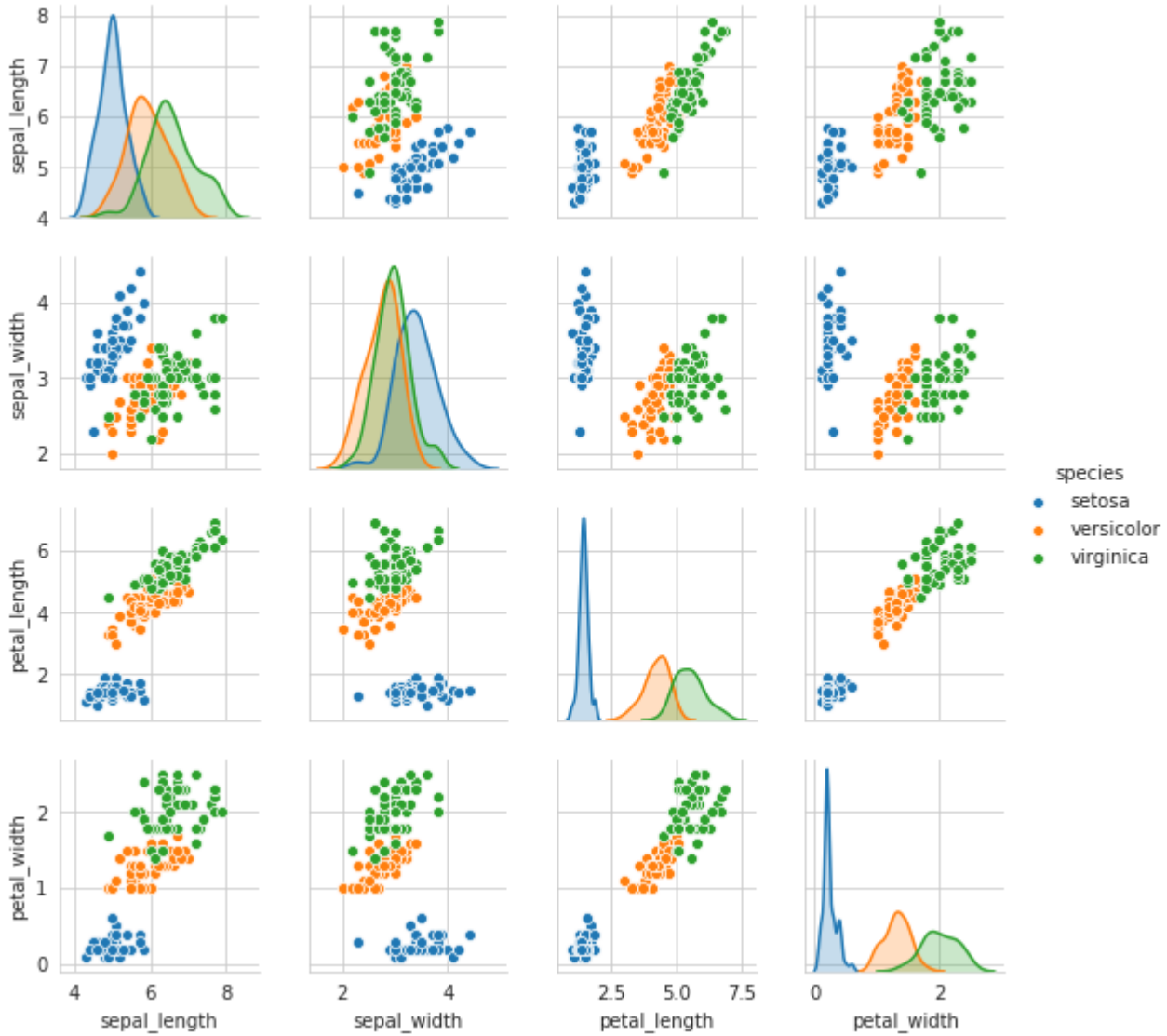
Slope Intercept Form: y = mx + c



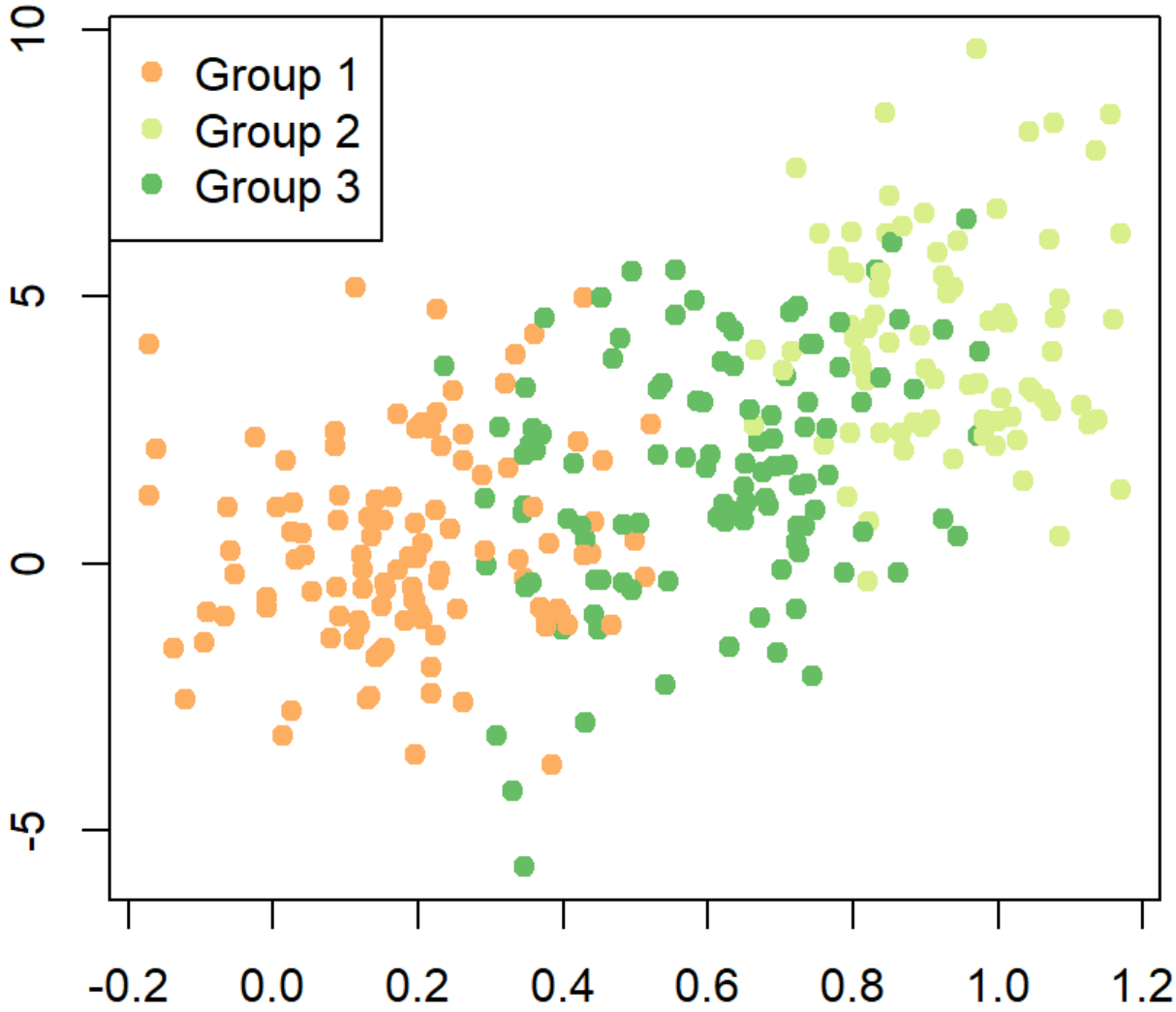
What you really need to understand –



Try Again

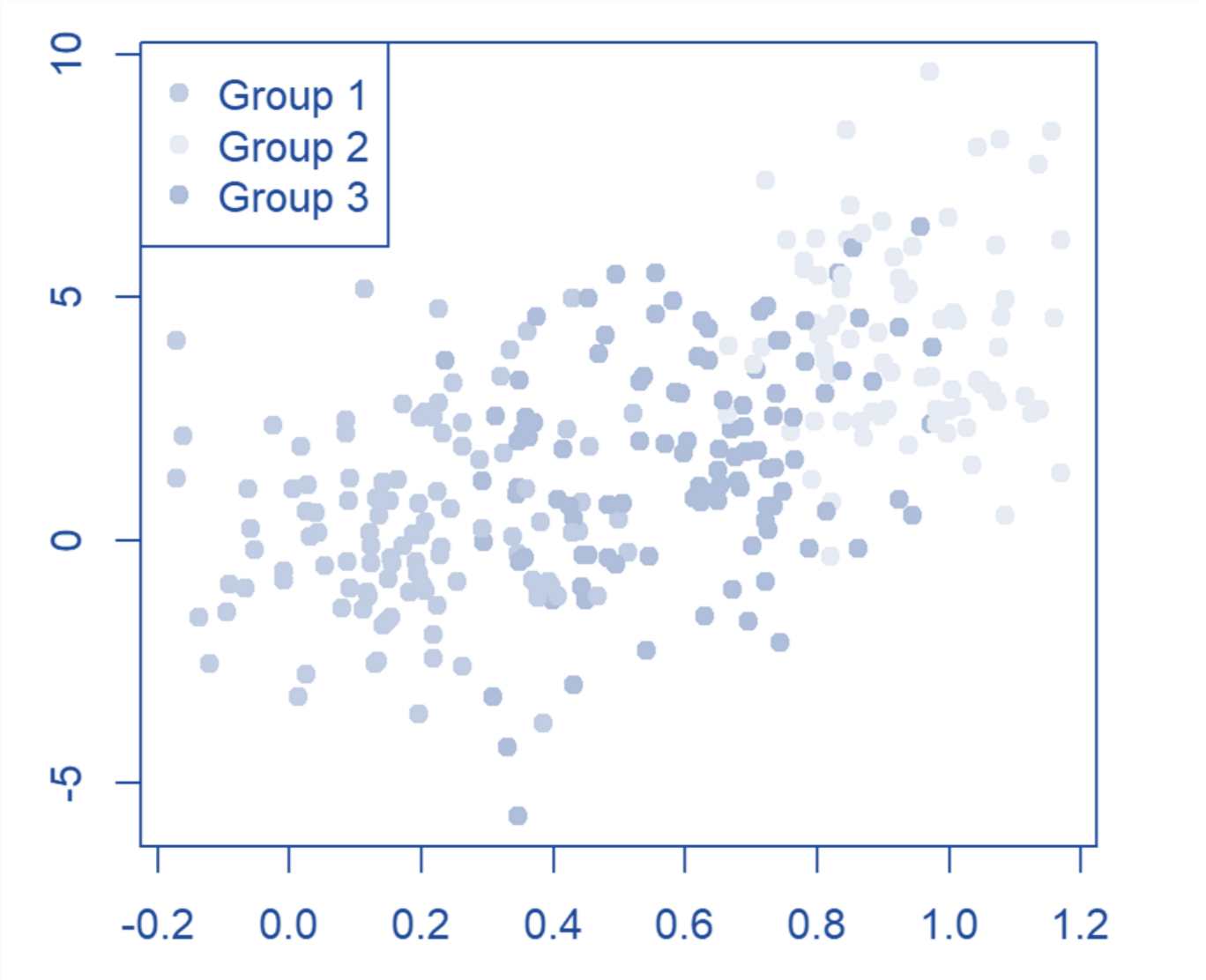

sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica

Easy

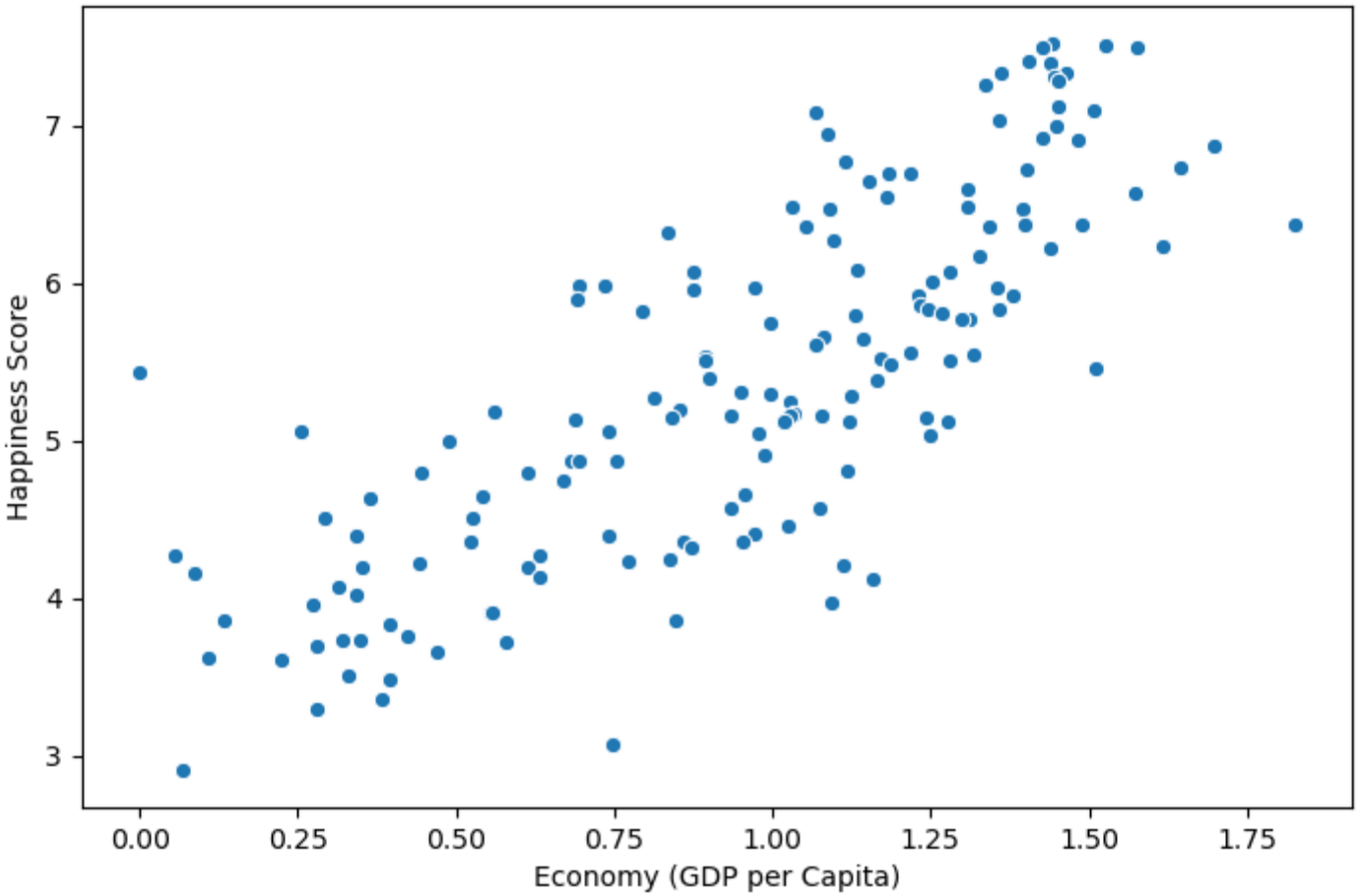


INVESTIC

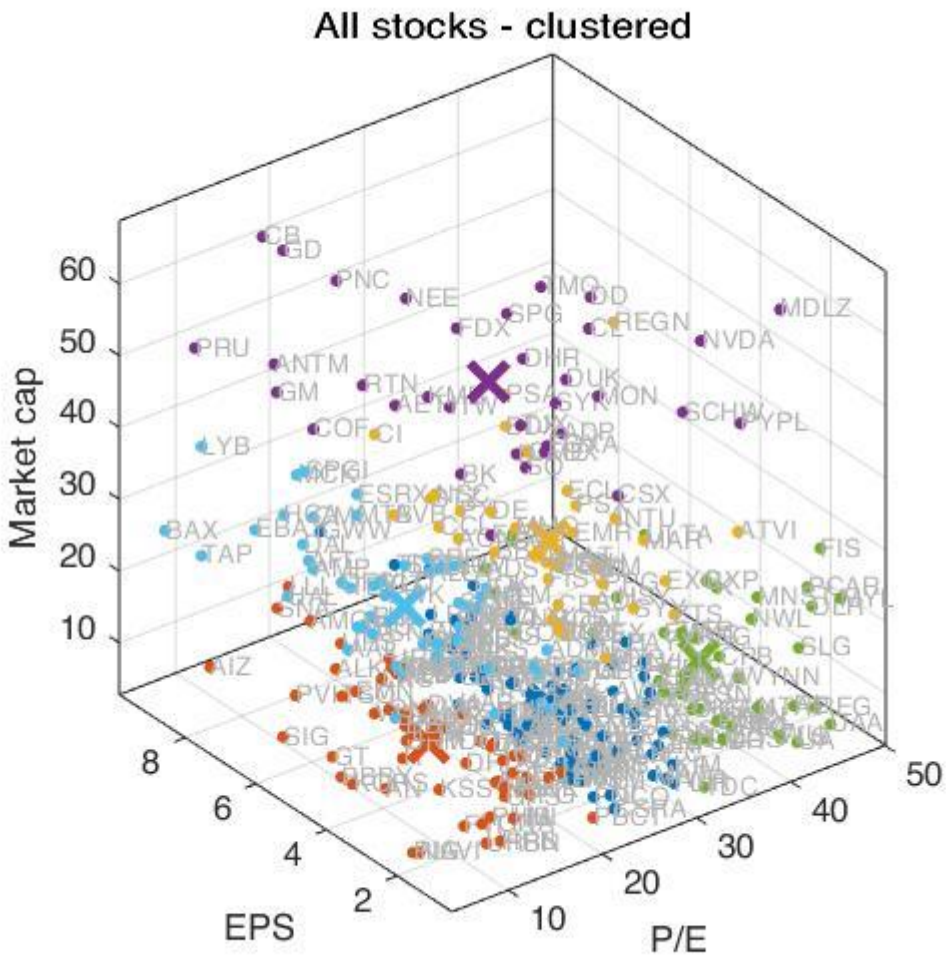
What now?



Say when people are happy?

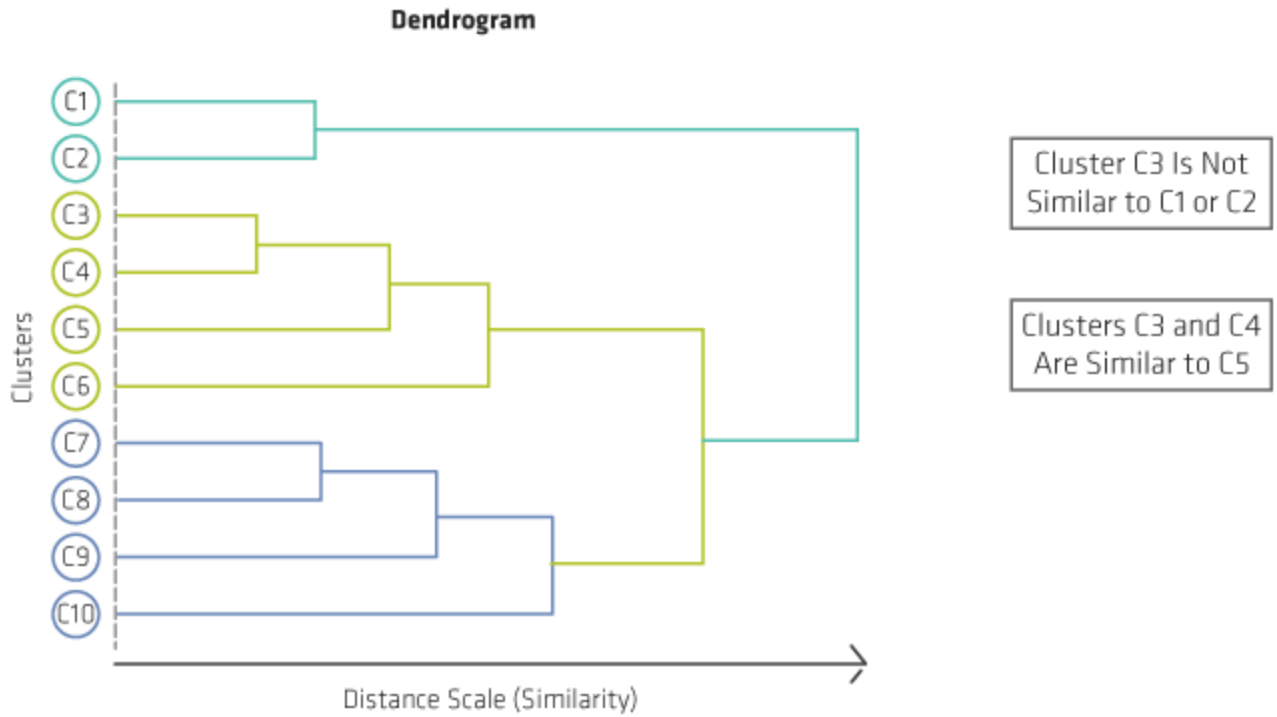


More than 2 Dimension?



Clustering

Hierarchical Clustering Adds Flexibility to Complex Groupings



Source: AllianceBernstein (AB)

Hierarchical Clustering

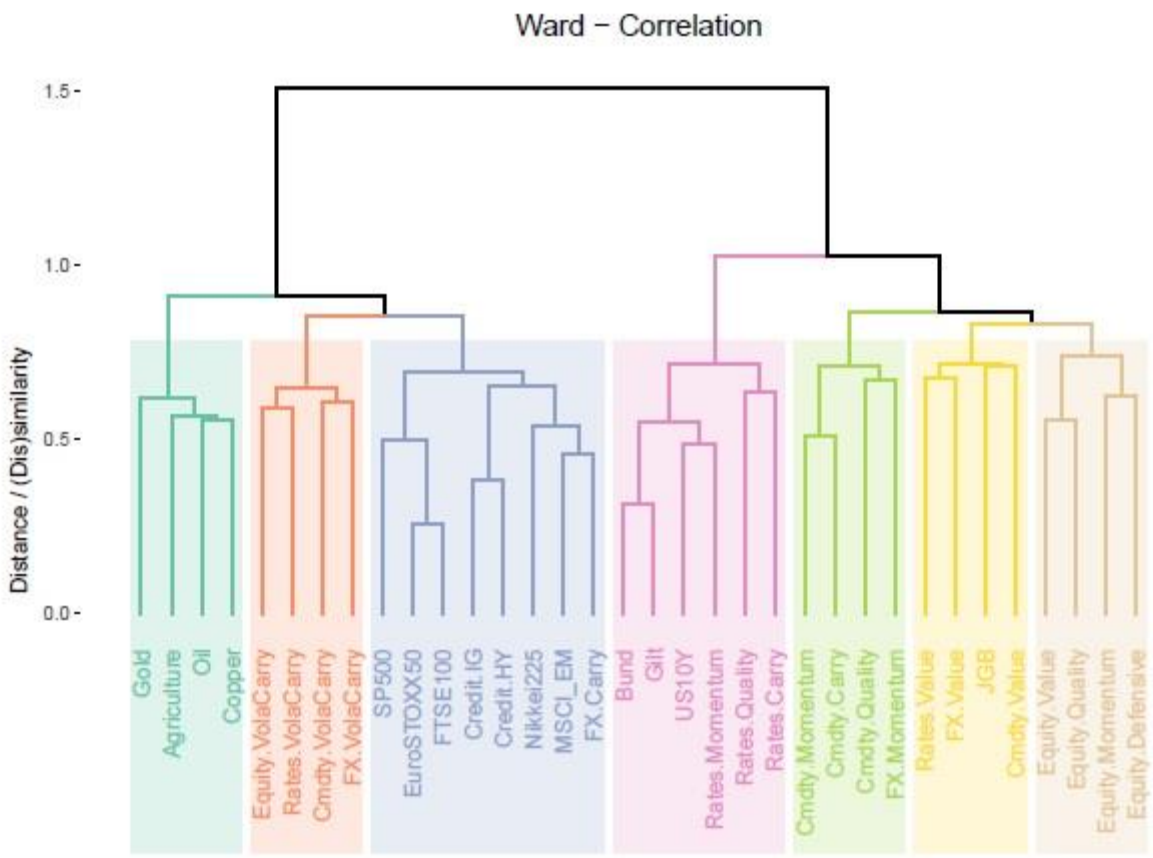
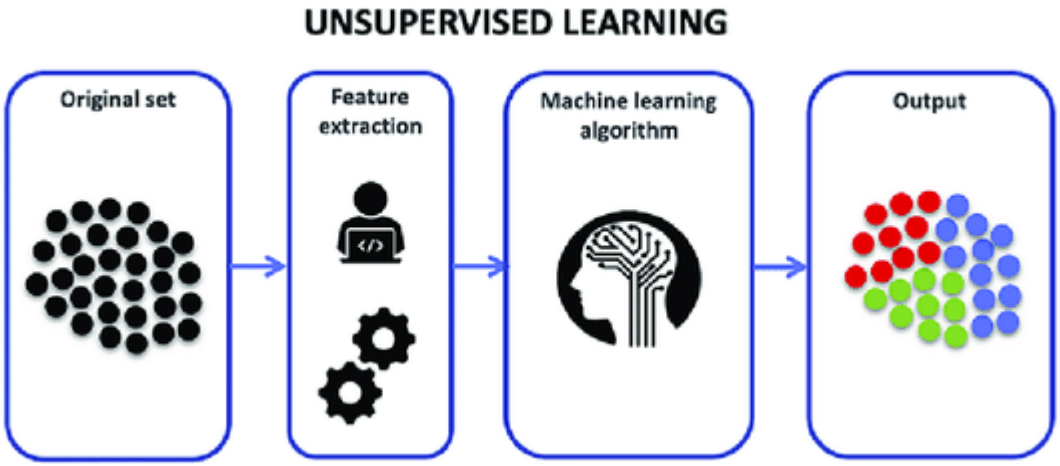
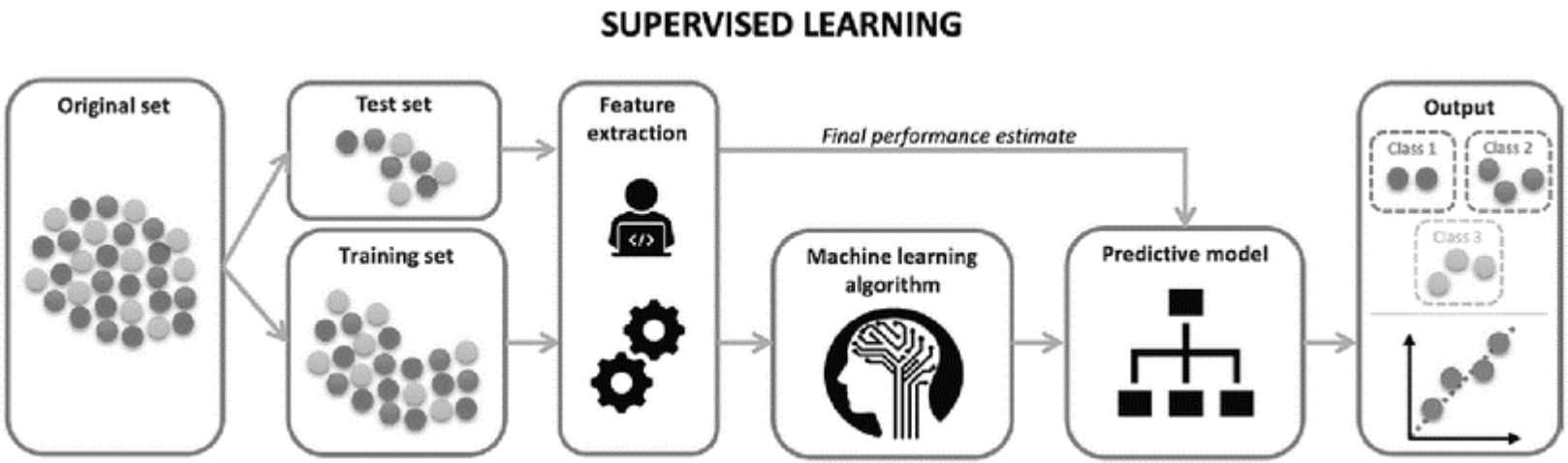


Figure 5: Dendrogram based on Ward's method; cut such that seven clusters obtain

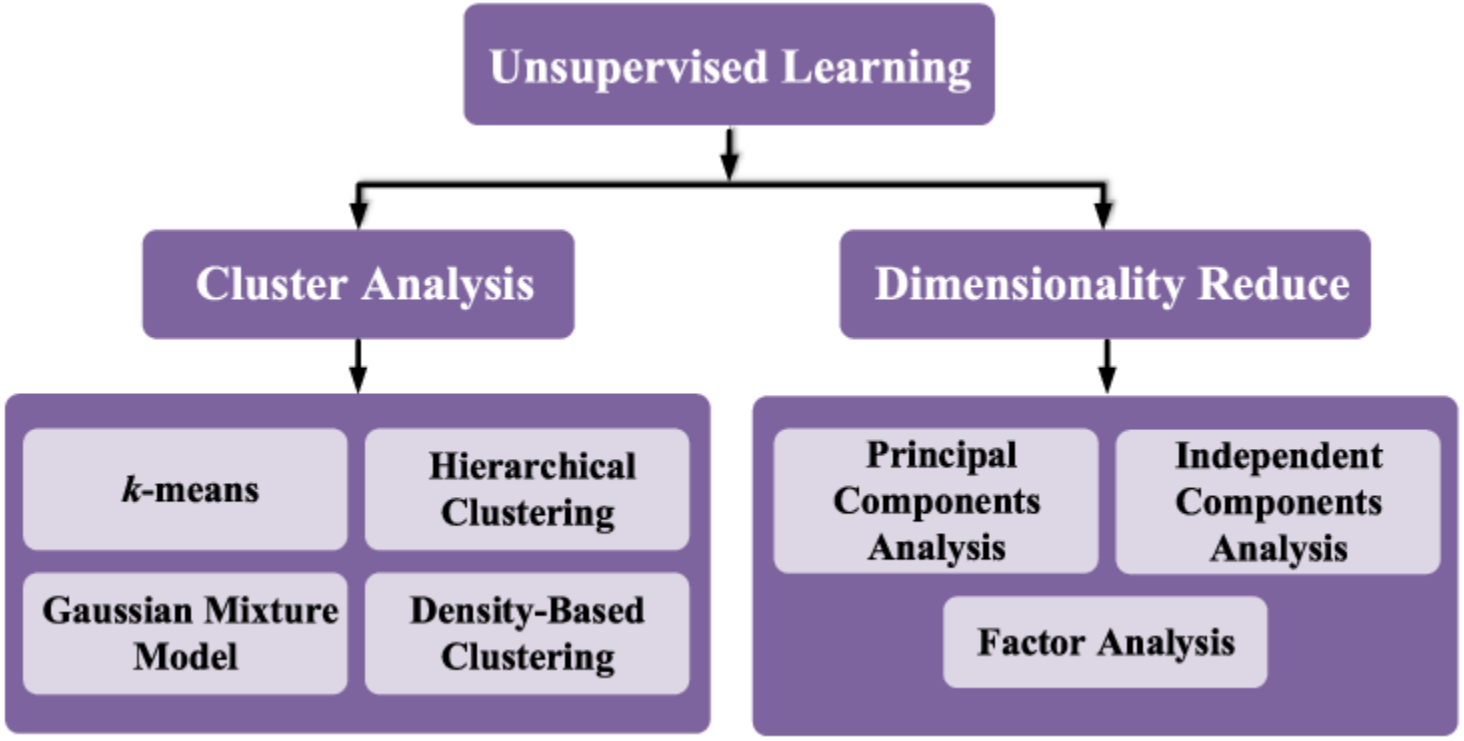
Example

Unsupervised Learning is Easy

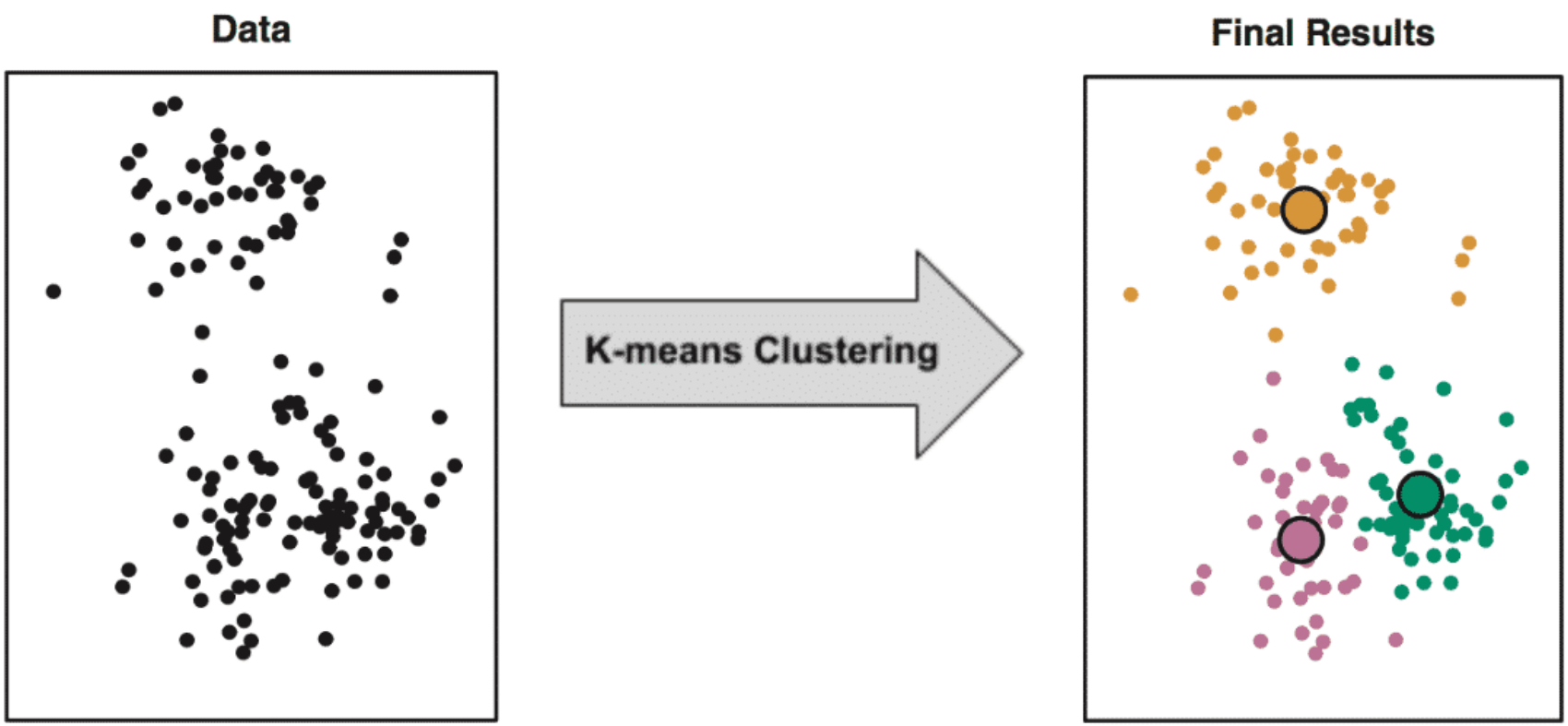




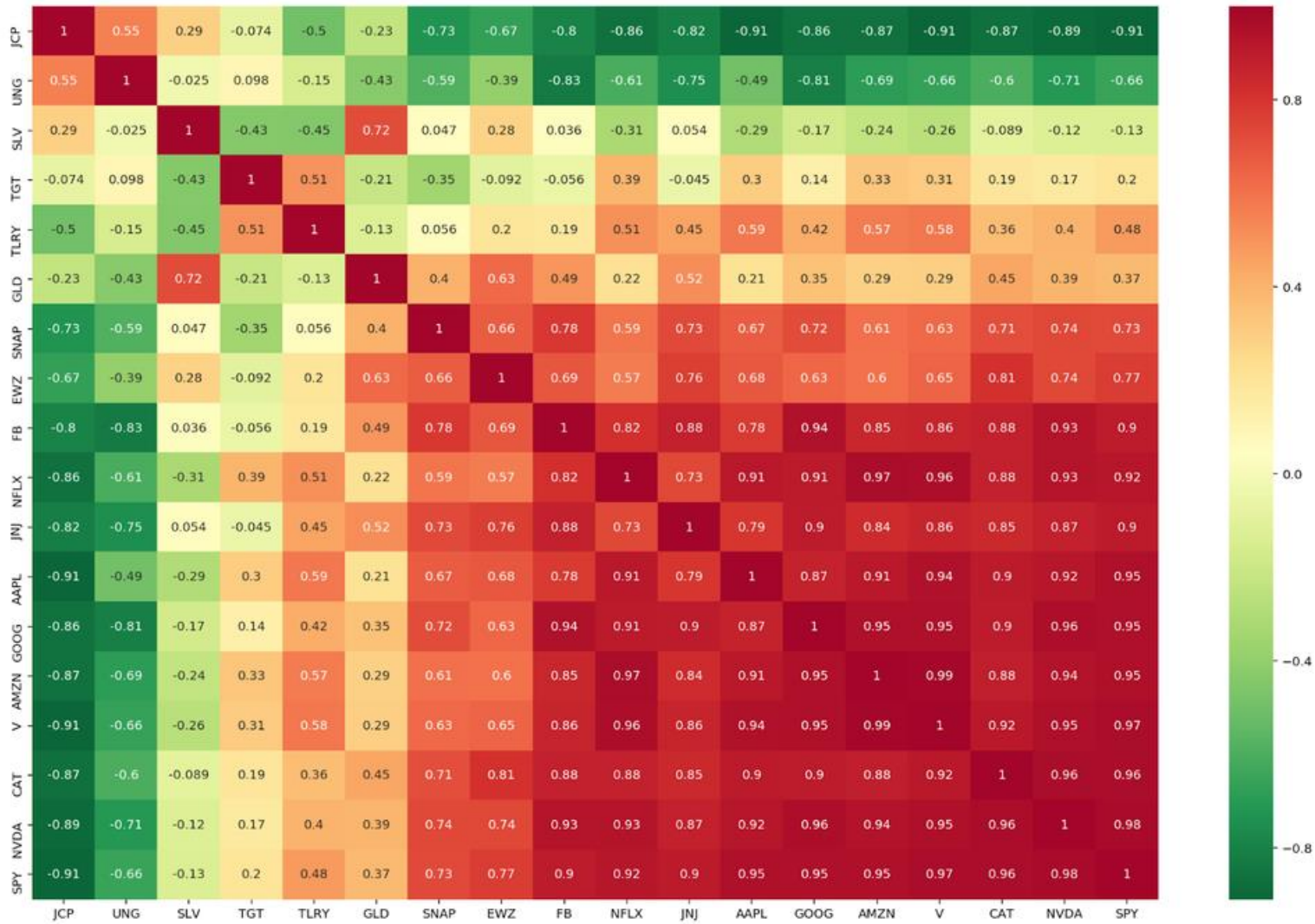
Unsupervised Learning is Easy



Unsupervised Learning Model?

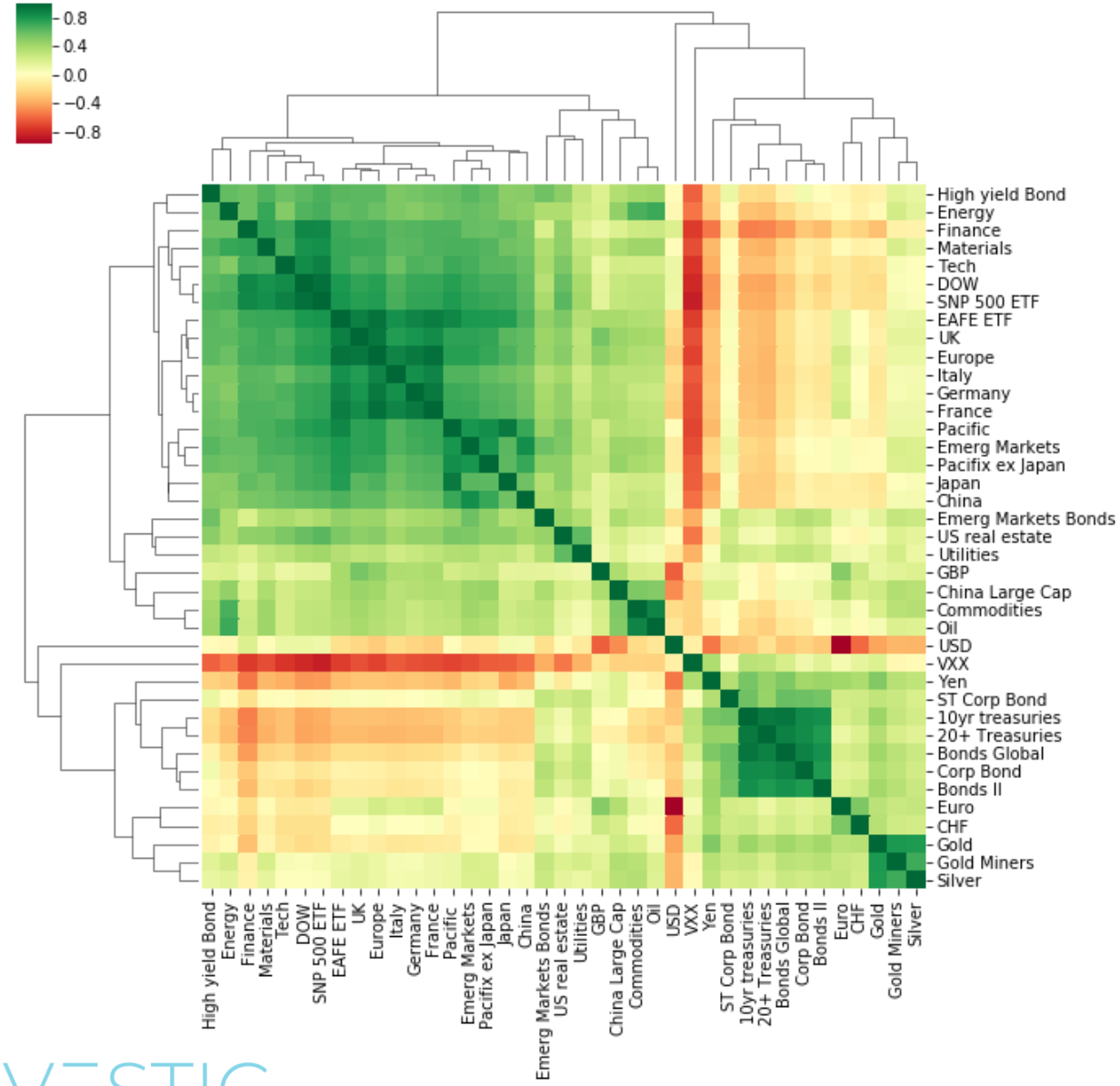


Then for God sake we do this?



INVESTIC

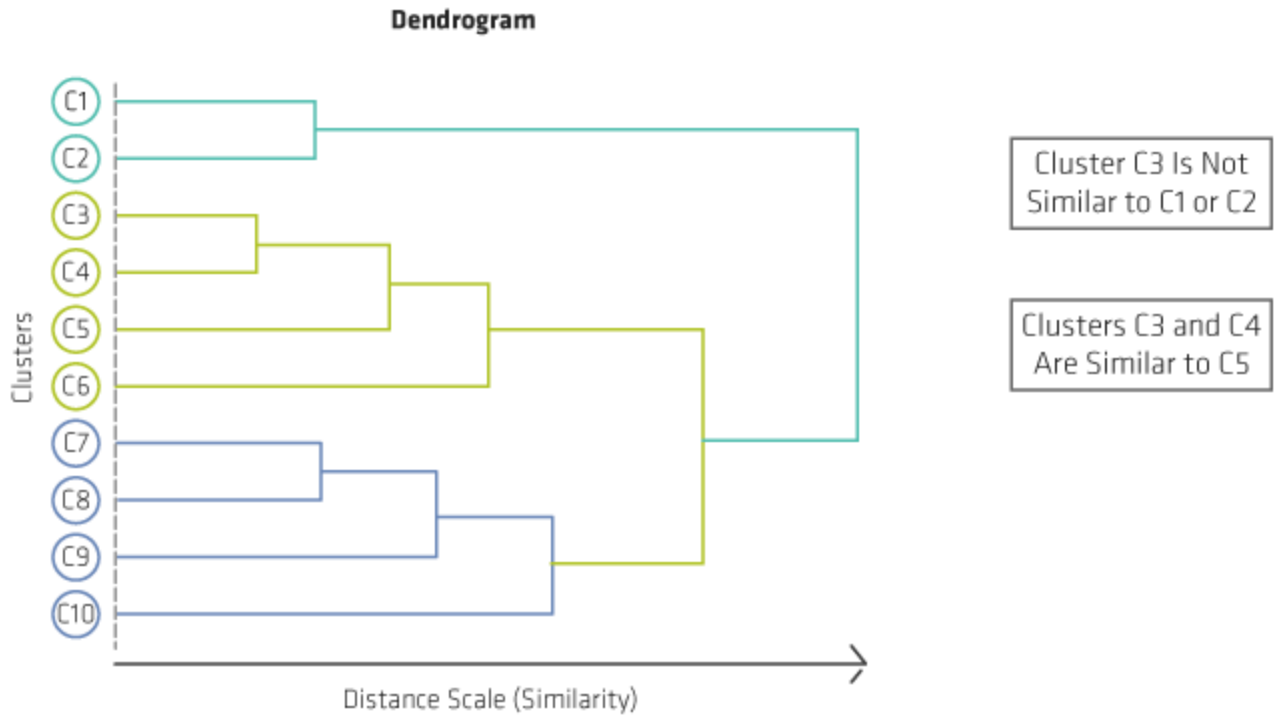
Leverage from correlation concept?



INVESTIC

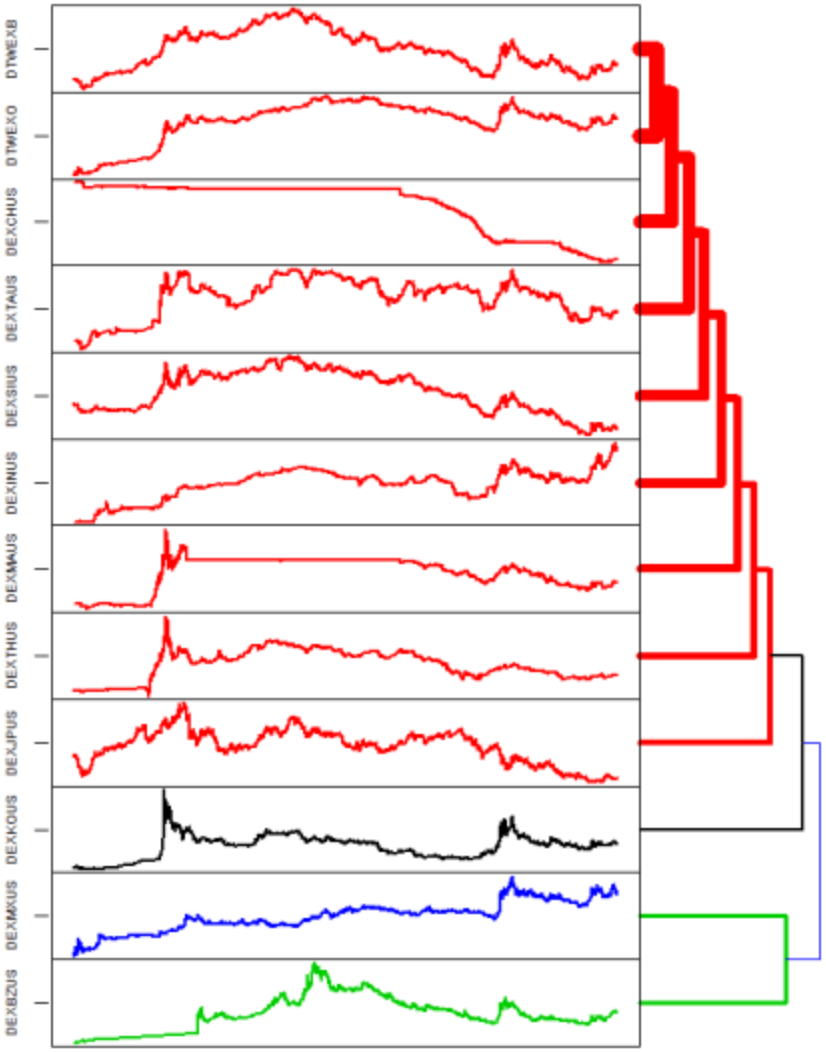
Introducing Hierarchical Clustering

Hierarchical Clustering Adds Flexibility to Complex Groupings

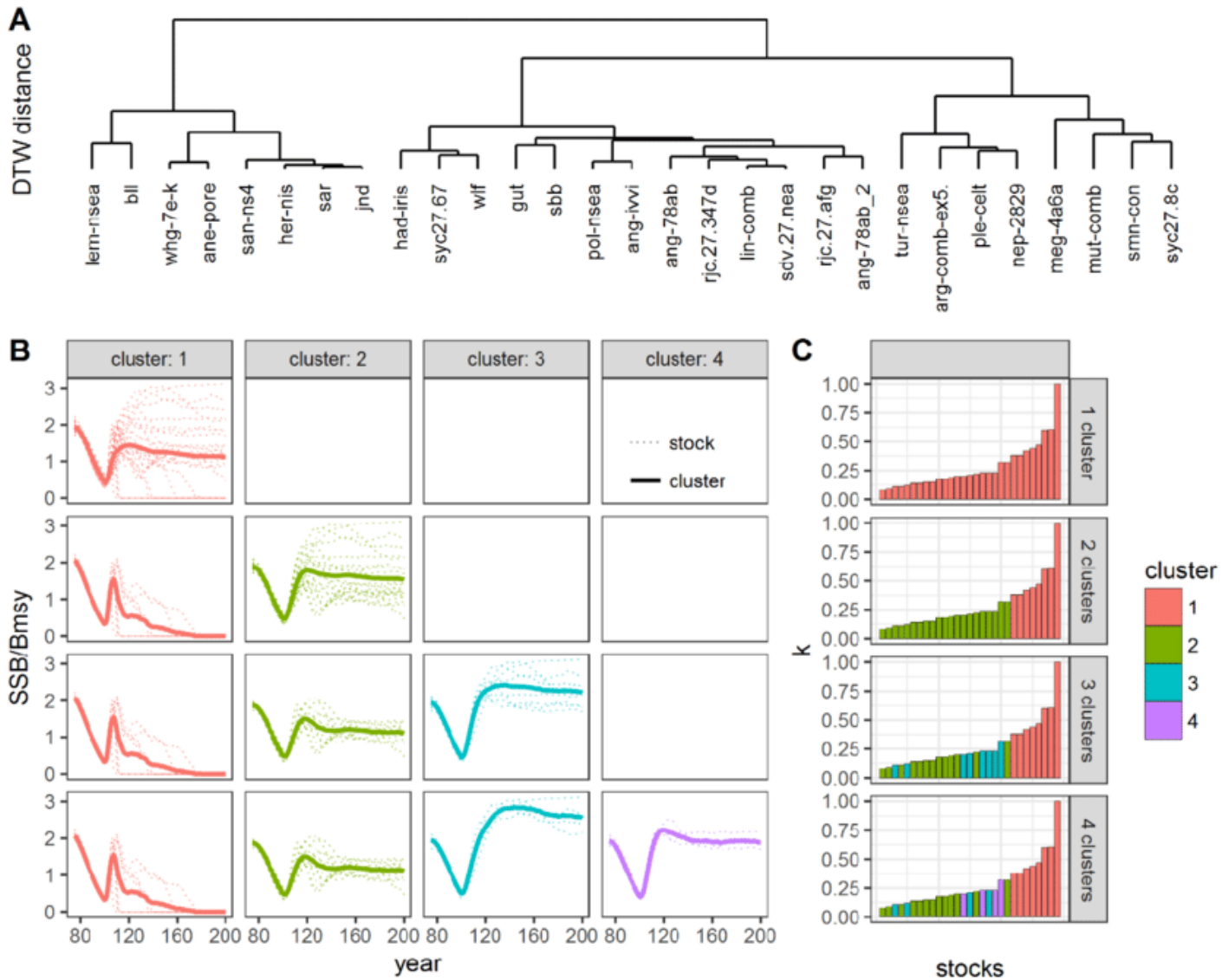


Source: AllianceBernstein (AB)

Introducing Hierarchical Clustering



More in details (example data)



INVESTIC

What you see in tutorial/ paper

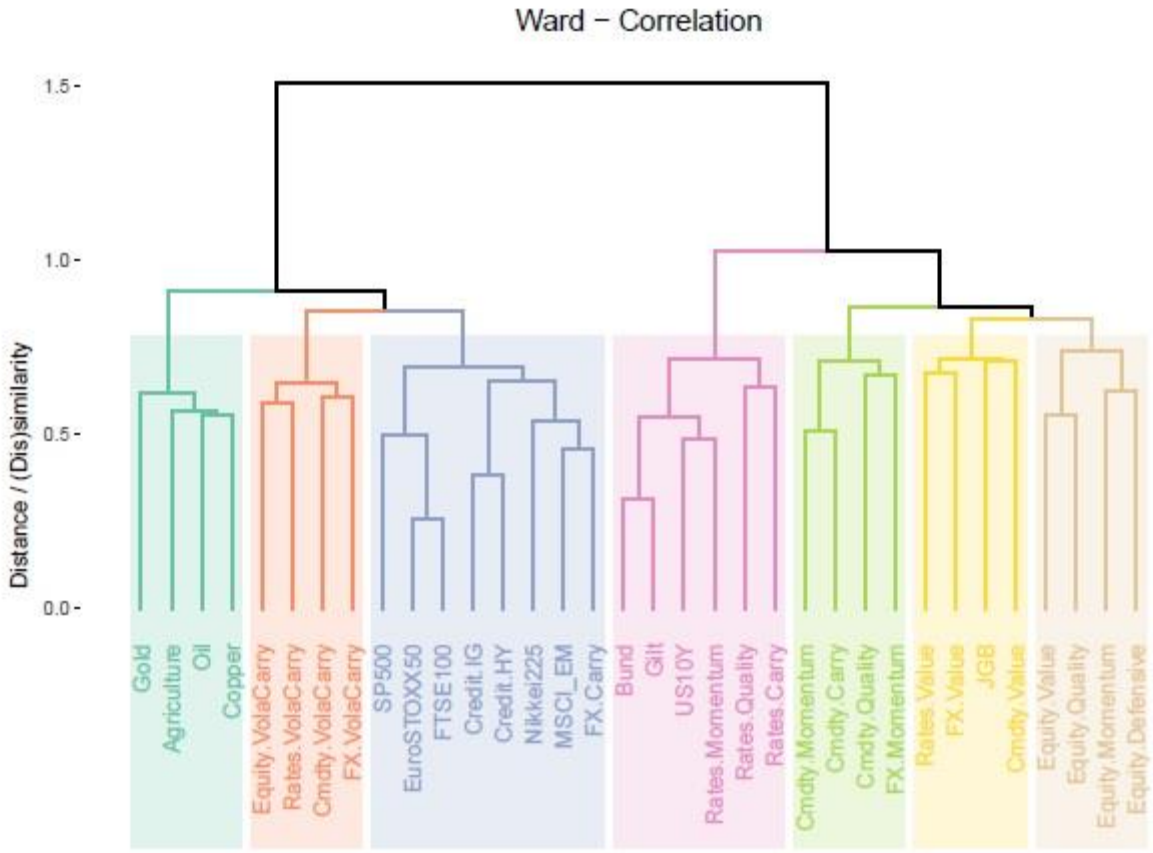
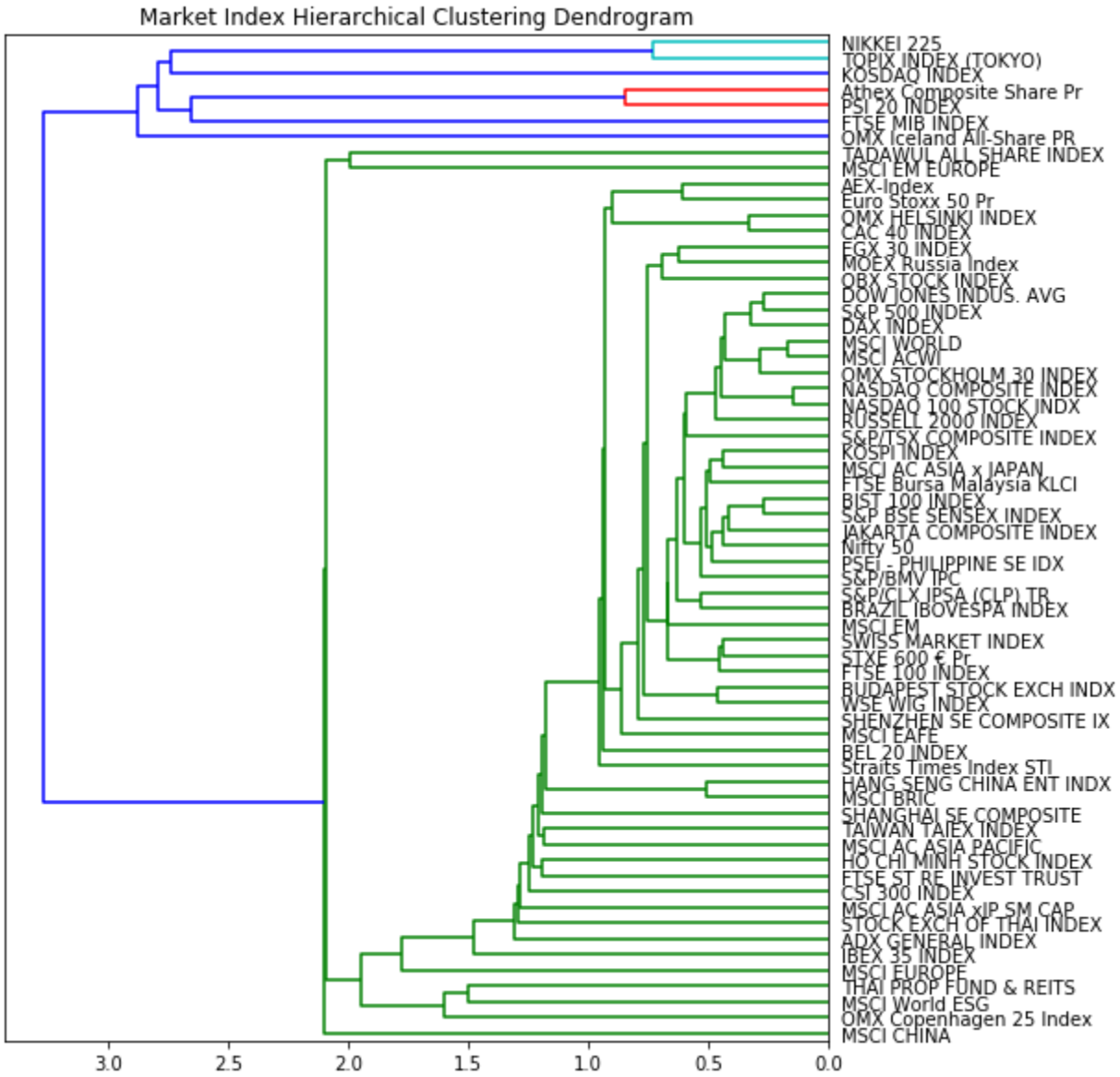


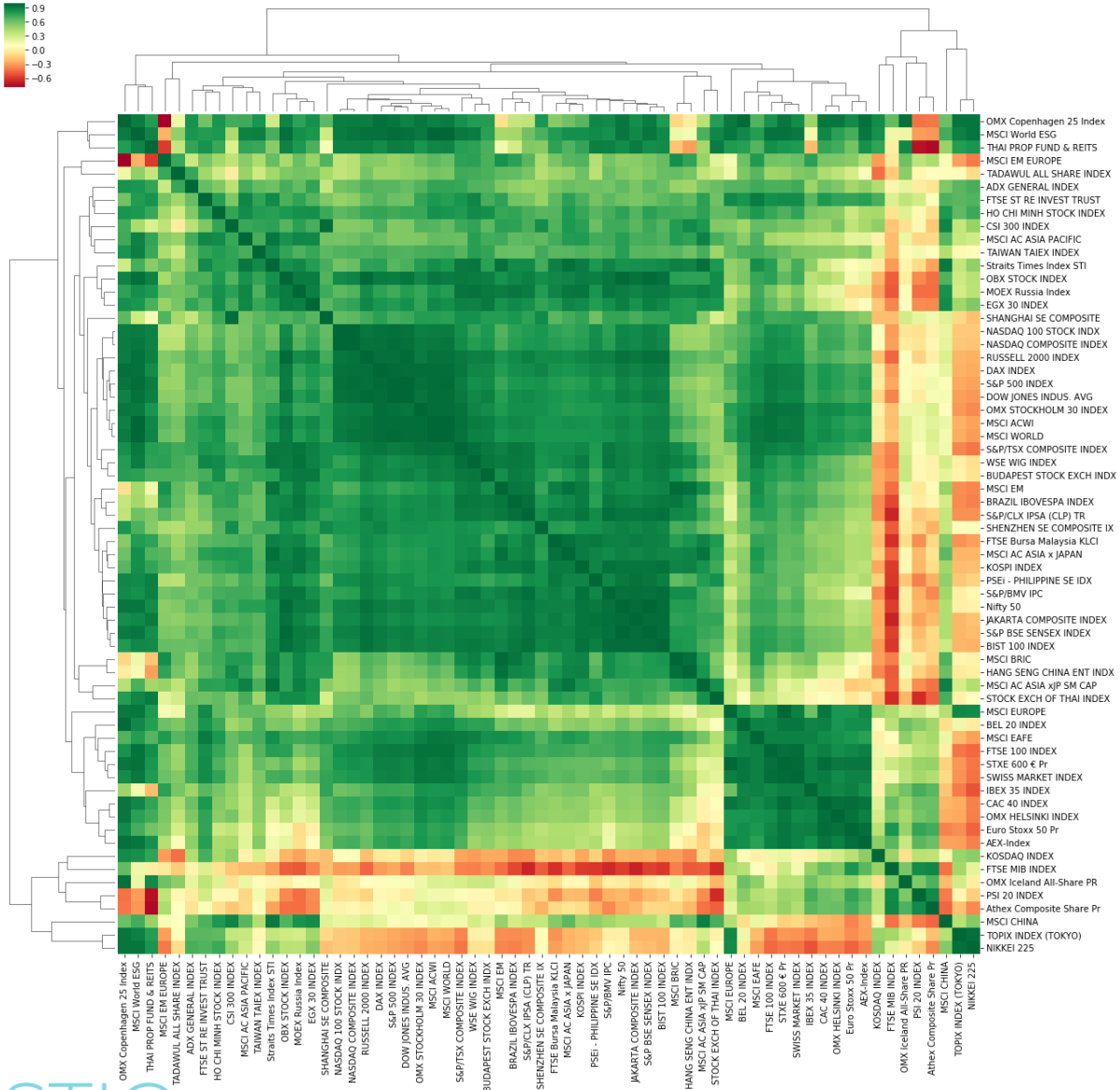
Figure 5: Dendrogram based on Ward's method; cut such that seven clusters obtain

Reality



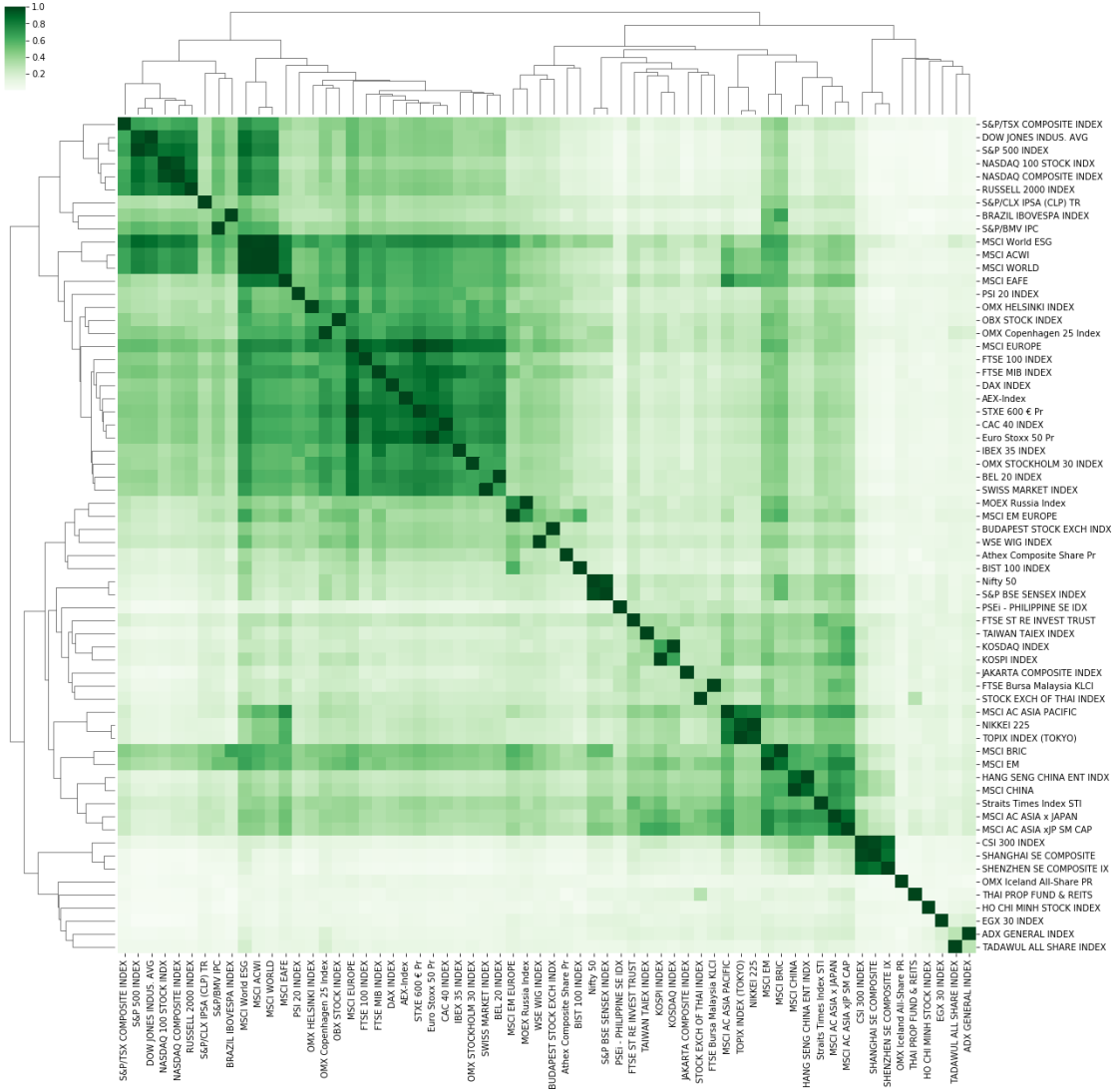
INVESTIC

Reality



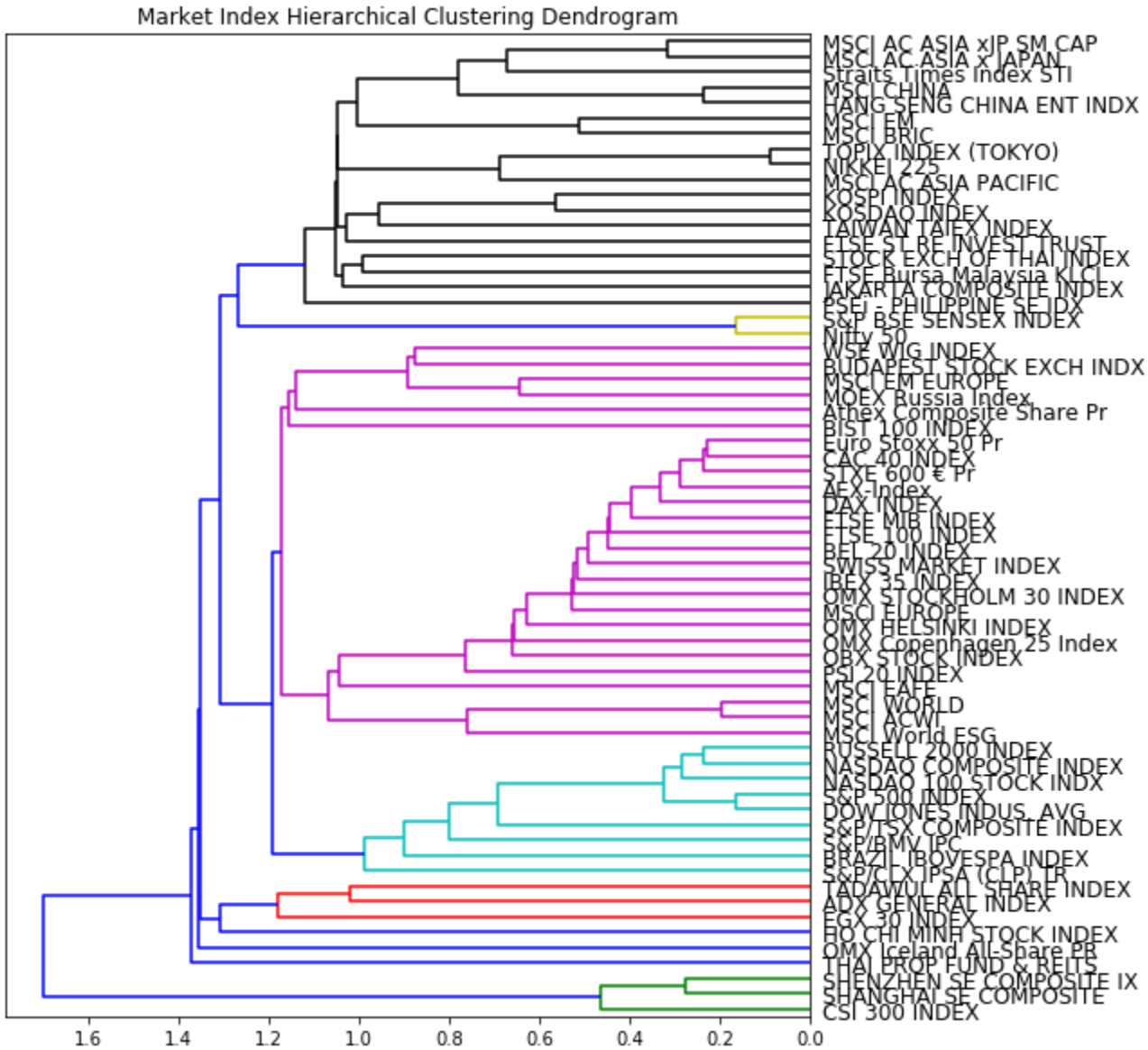
INVESTIC

With some knowledge



INVESTIC

With some knowledge



INVESTIC

Why??

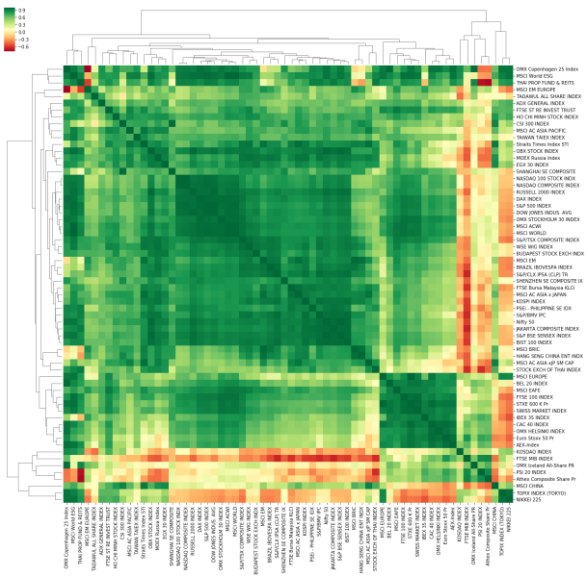
Reality

In [7]:
```
df_weekly = df.resample('W-FRI').last()
df_weekly.tail()
```

Out[7]:

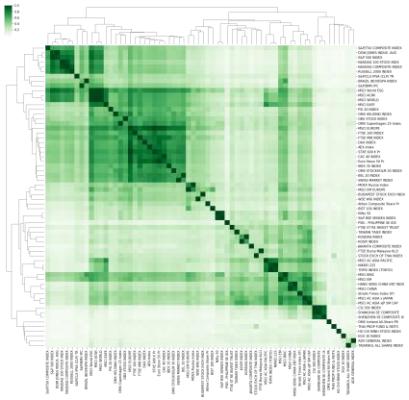
Date	MSCI ACWI	MSCI WORLD	MSCI EM	MSCI BRIC	MSCI AC ASIA x JAPAN	MSCI EUROPE	MSCI EM EUROPE	MSCI CHINA	MSCI AC ASIA xJP SM CAP
2019-01-04	459.94	1904.18	964.97	286.18	588.05	116.20	299.63	70.63	944.32
2019-01-11	473.26	1957.17	1001.11	296.14	612.08	117.96	307.25	74.28	966.98
2019-01-18	483.50	2000.76	1018.00	302.10	622.10	120.57	315.43	76.21	970.26
2019-01-25	484.53	2001.86	1032.34	305.27	631.07	120.72	320.04	77.48	979.52
2019-02-01	484.53	2001.86	1032.34	305.27	631.07	120.72	320.04	77.48	979.52

In [42]:
```
sns.clustermap(df_weekly_corr, figsize = (20,20), cmap = 'RdYlGn')
```

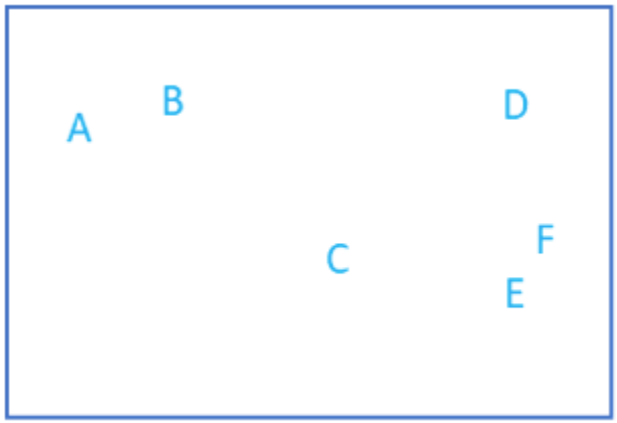


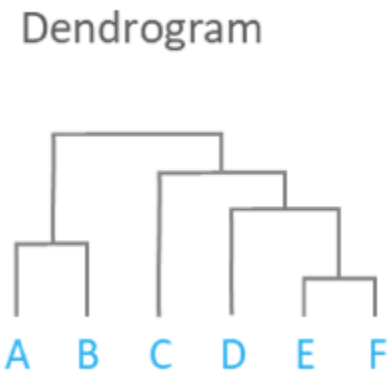
With some knowledge

```
In [9]: df_w_rolling_corr_2y = df_daily_r.rolling(104).corr(df_daily_r,pairwise=True)

In [28]: w_corr = df_w_rolling_corr_2y.loc['1987-01-01':]

In [29]: w_corr = w_corr.groupby(level=[1]).mean()

In [31]: sns.clustermap(w_corr, figsize = (20,20), cmap = 'Greens')
```



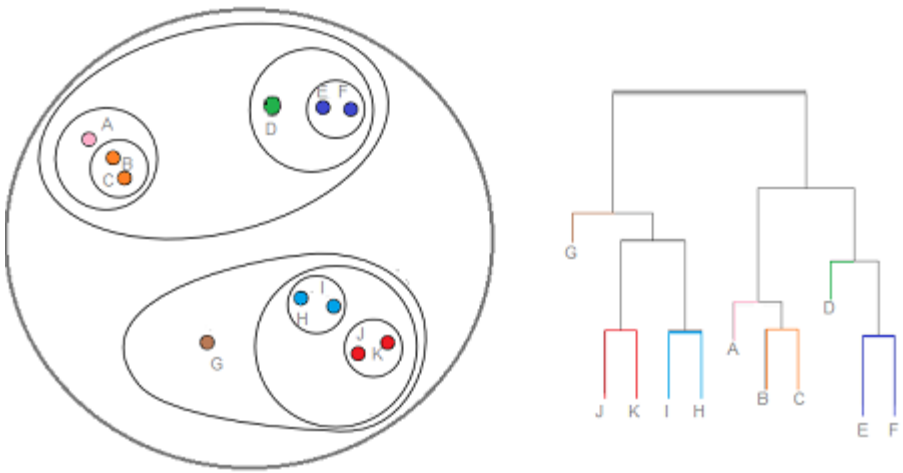
Dendogram





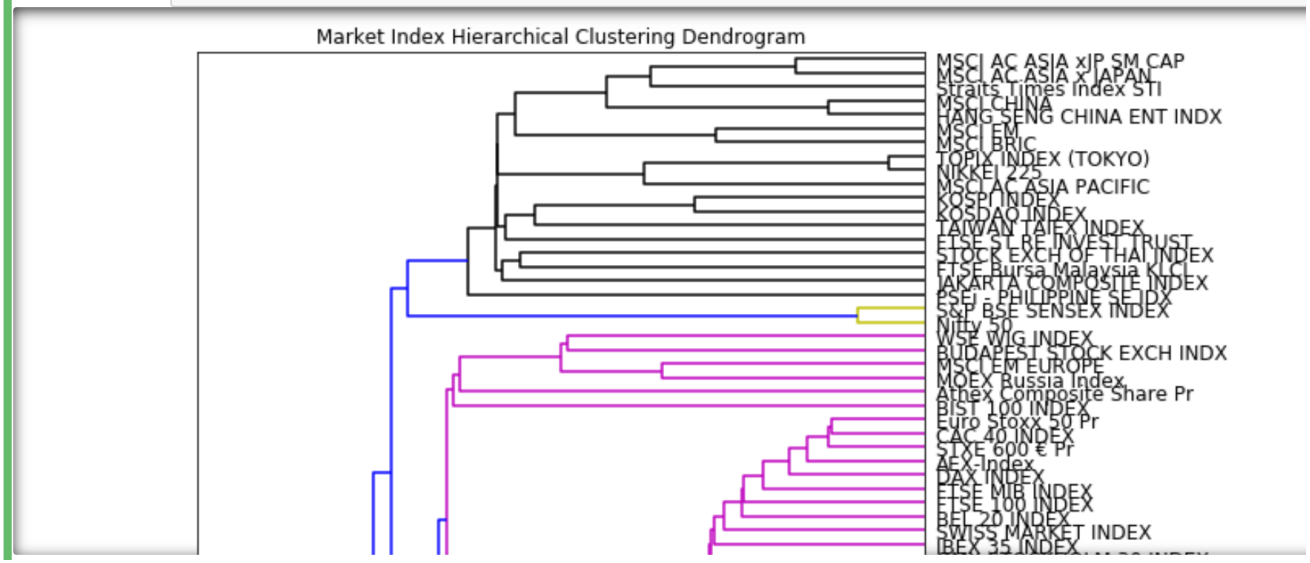
$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Dendogram



Dendogram

```
In [33]: from scipy import stats
         import scipy.cluster.hierarchy as hac

         Z_w = hac.linkage(w_corr, method='single')

         # Plot dendogram
         plt.figure(figsize=(8, 10))
         plt.title('Market Index Hierarchical Clustering Dendrogram')

         hac.dendrogram(
             Z_w,labels = w_corr.index, orientation="left",
             leaf_rotation=0,  # rotates the x axis labels
             leaf_font_size=12.,  # font size for the x axis labels
         )
         plt.show()
```



Dendogram

	Age	Expense($)	Distance(KI
Anne	30	100	
Ben	18	200	
Chad	35	150	
Dave	48	300	
Eric	50	200	

$$d_{(Anne,Ben)} = \sqrt{(Anne_{age} - Ben_{age})^2 + (Anne_{Expense} - Ben_{Expense})^2 + (Anne_{Distance} - Ben_{Distance})^2}$$

If we put all the numbers, it would be.

$$d_{(Anne,Ben)} = \sqrt{(30-18)^2 + (100-200)^2 + (5-2)^2}$$

$$d_{(Anne,Ben)} = \sqrt{(12)^2 + (-100)^2 + (3)^2}$$

$$d_{(Anne,Ben)} = \sqrt{144 + 10000 + 9}$$

$$d_{(Anne,Ben)} = \sqrt{10153}$$

$$d_{(Anne,Ben)} = 100.76$$

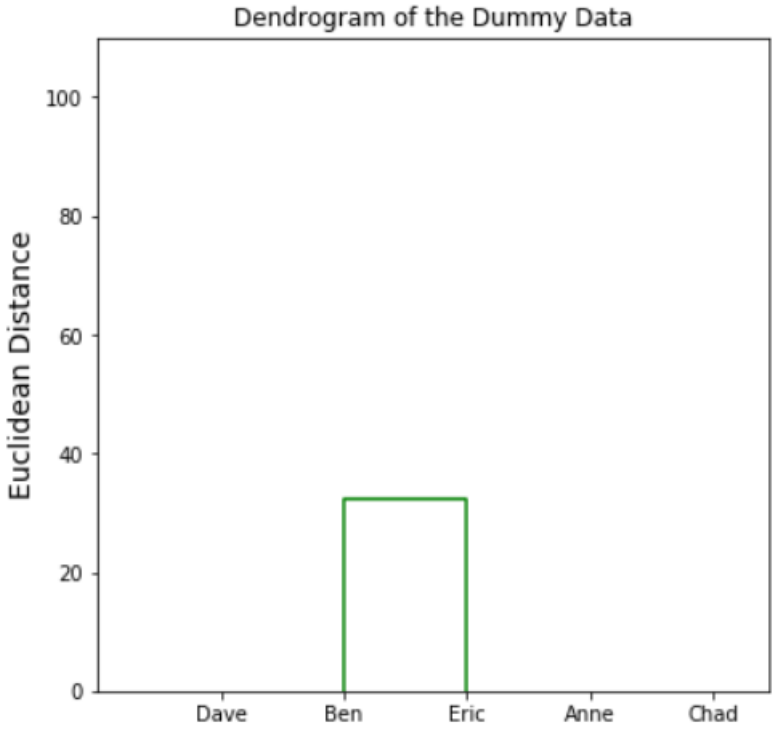
	Anne	Ben	Chad	Dave	Eric
Anne	0.00	100.76	50.29	200.81	101.99
Ben	100.76	0.00	53.05	104.42	32.25
Chad	50.29	53.05	0.00	150.59	52.21
Dave	200.81	104.42	150.59	0.00	100.04
Eric	101.99	32.25	52.21	100.04	0.00

Euclidean distance matrix of the dummy data

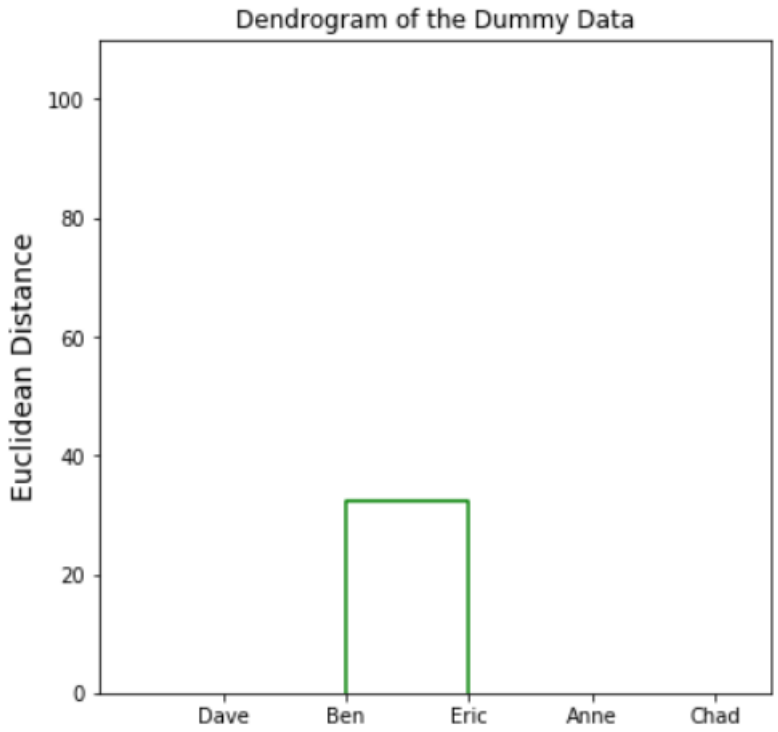
Dendogram

	Anne	Ben	Chad	Dave	Eric
Anne	0.00	100.76	50.29	200.81	101.99
Ben	100.76	0.00	53.05	104.42	32.25
Chad	50.29	53.05	0.00	150.59	52.21
Dave	200.81	104.42	150.59	0.00	100.04
Eric	101.99	32.25	52.21	100.04	0.00

Euclidean distance matrix of the dummy data

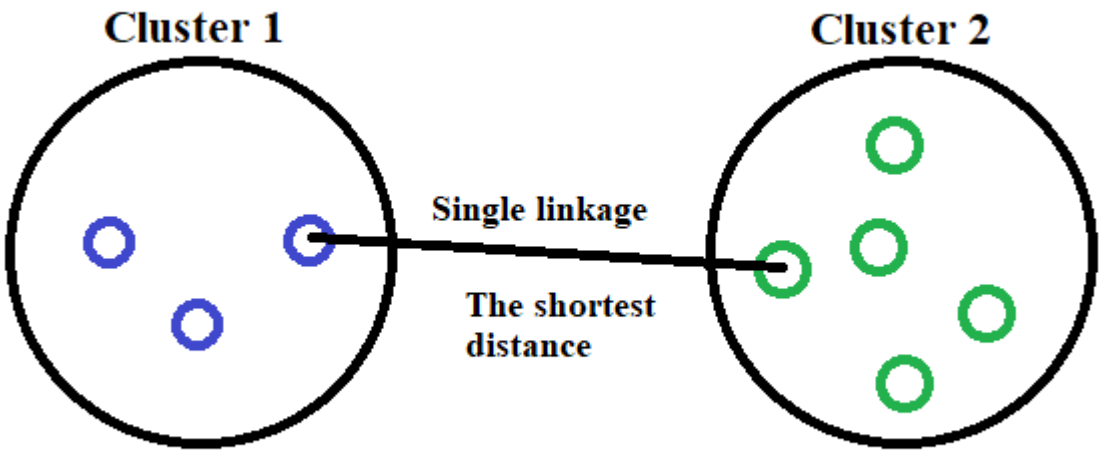


Dendogram



	Anne	Chad	Dave	(Ben,Eric)
Anne	0	50.29	200.81	?
Chad	50.29	0	150.59	?
Dave	200.81	150.59	0	?
(Ben,Eric)	?	?	?	0

Dendogram



In a single linkage criterion we, define our distance as the minimum distance between clusters data point. If we put it in a mathematical formula, it would look like this.

$$D(X, Y) = \min_{x \in X, y \in Y} d_{x,y}$$

Dendogram

$$D(Anne, (Ben, Eric)) = \min(Anne, (Ben, Eric))$$

$$D(Anne, (Ben, Eric)) = \min(dist(Anne, Ben), dist(Anne, Eric))$$

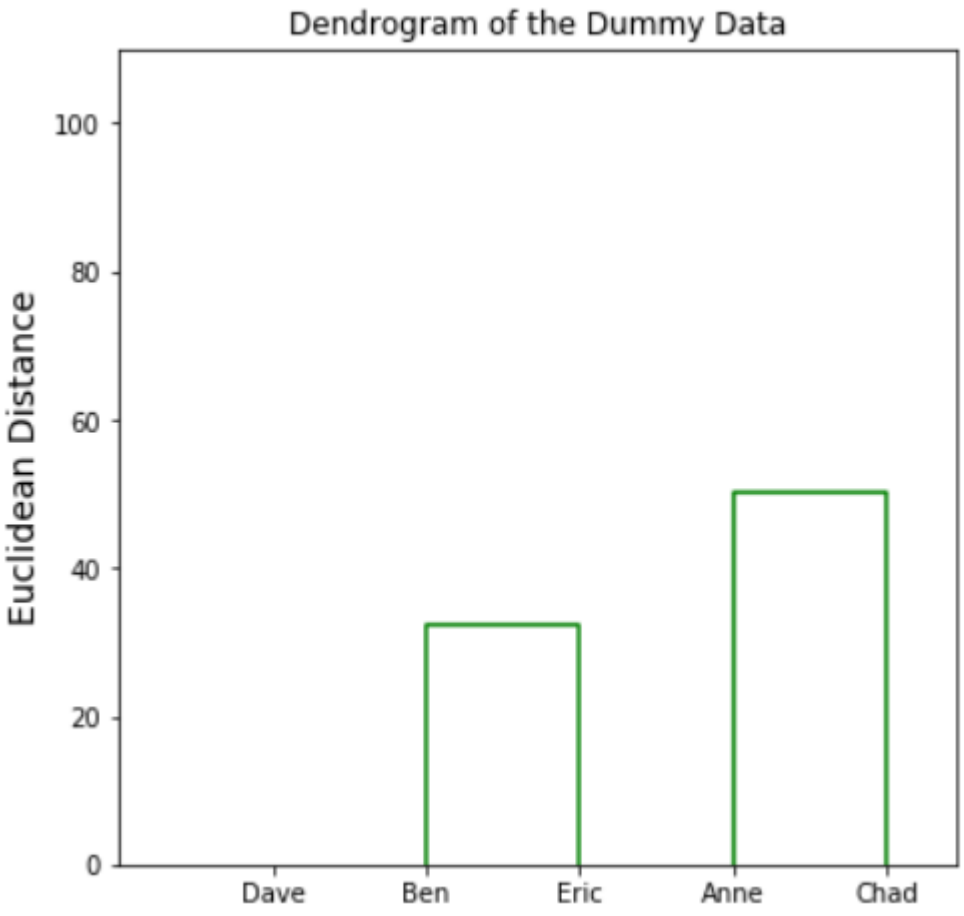
$$D(Anne, (Ben, Eric)) = \min((100.76, 101.99))$$

$$D(Anne, (Ben, Eric)) = 100.76$$

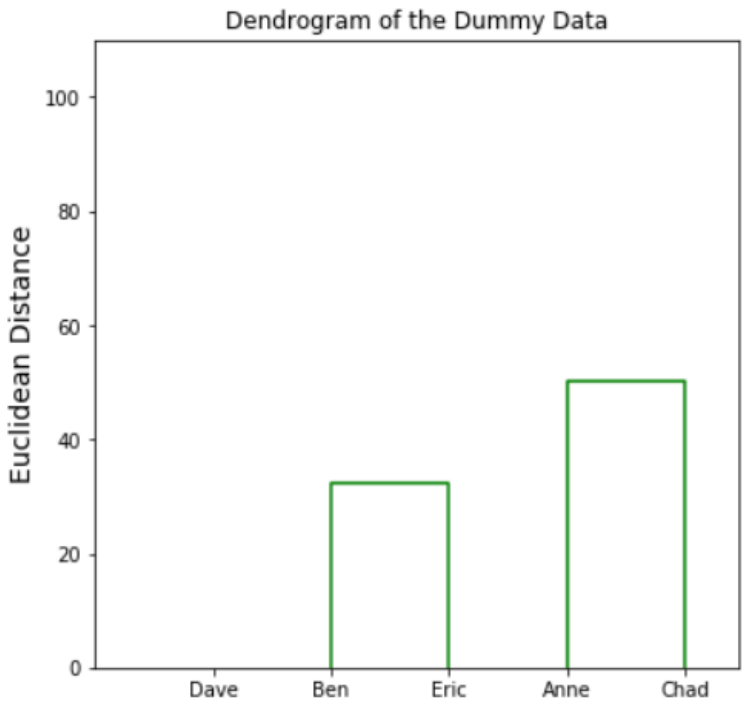
	Anne	Chad	Dave	(Ben,Eric)
Anne	0	50.29	200.81	100.76
Chad	50.29	0	150.59	52.21
Dave	200.81	150.59	0	100.04
(Ben,Eric)	100.76	52.21	100.04	0.00

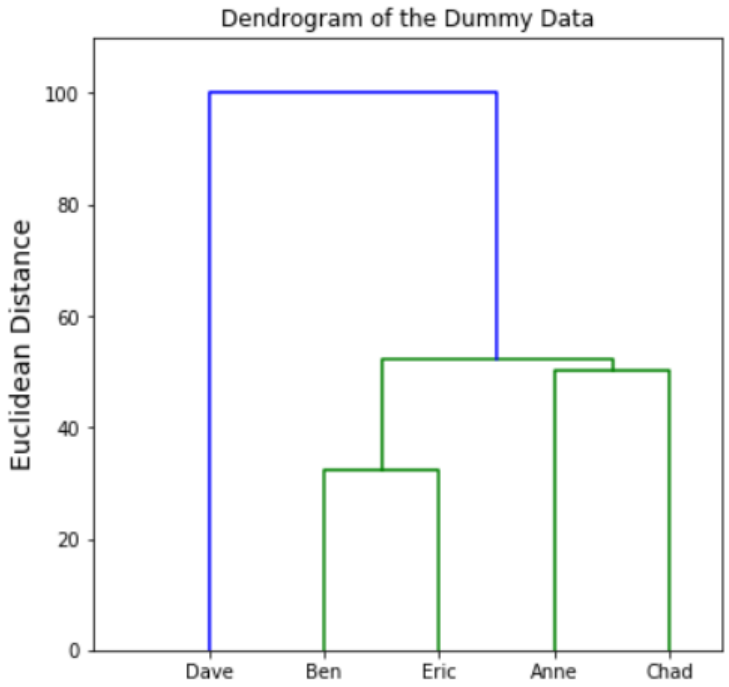
Dendogram

	Anne	Chad	Dave	(Ben,Eric)
Anne	0	50.29	200.81	100.76
Chad	50.29	0	150.59	52.21
Dave	200.81	150.59	0	100.04
(Ben,Eric)	100.76	52.21	100.04	0.00

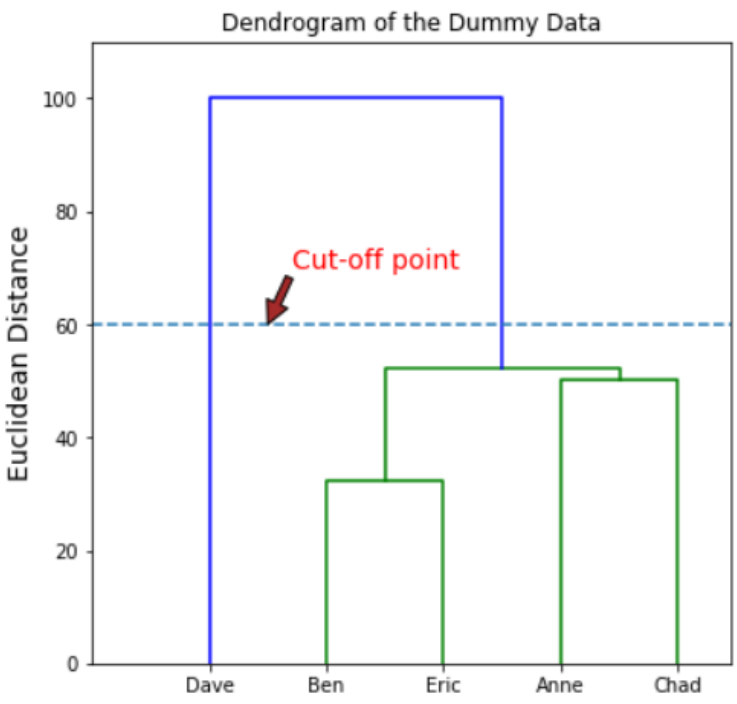


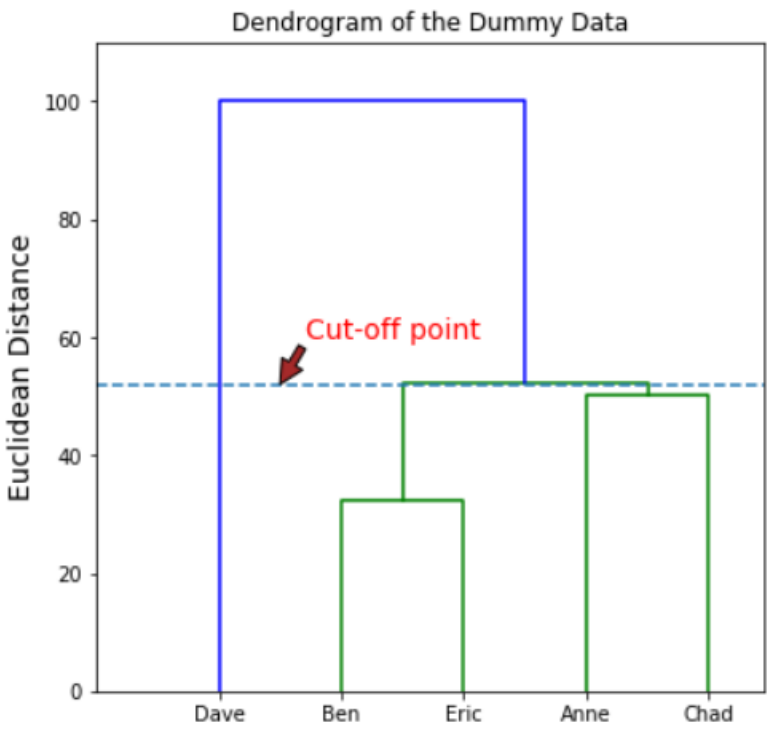
Dendogram


Dendrogram of the Dummy Data
Euclidean Distance
0
20
40
60
80
100
Dave
Ben
Eric
Anne
Chad

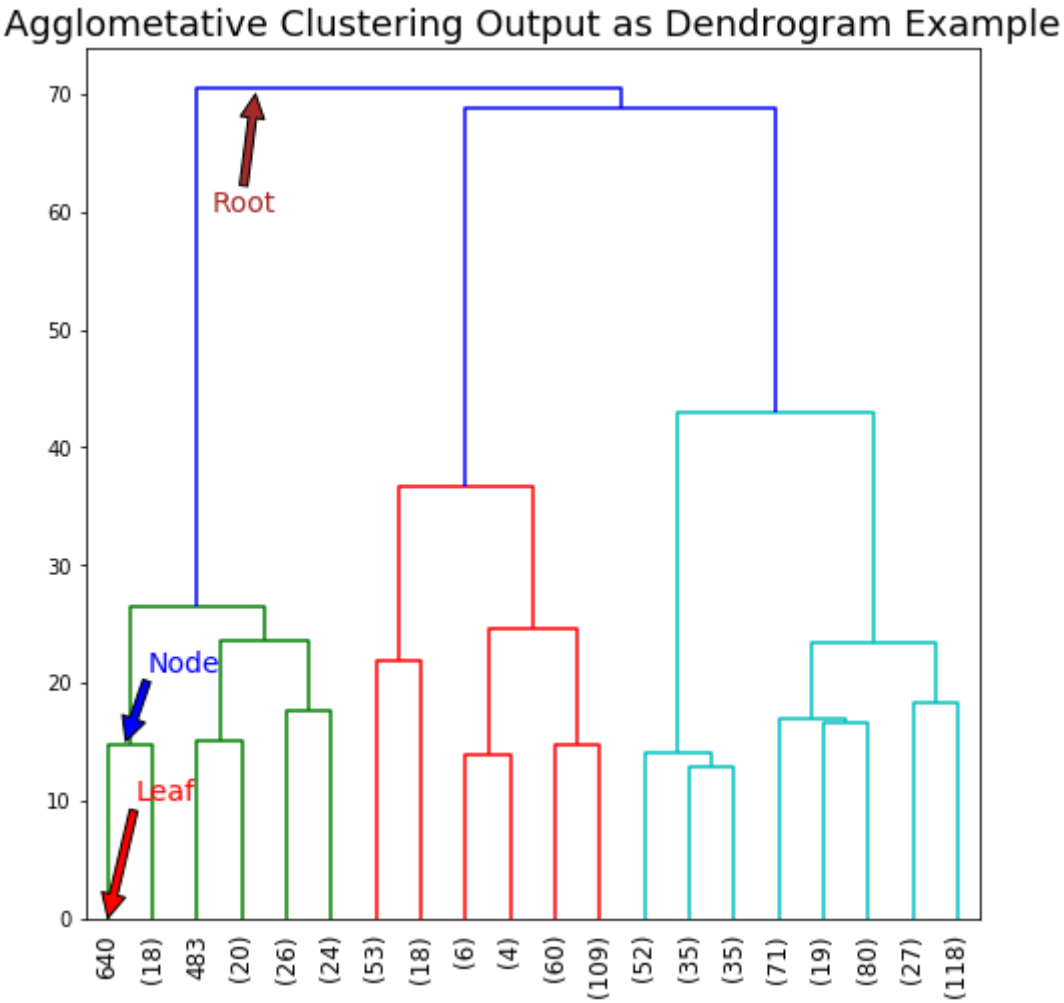

Dendrogram of the Dummy Data
Euclidean Distance
0
20
40
60
80
100
Dave
Ben
Eric
Anne
Chad

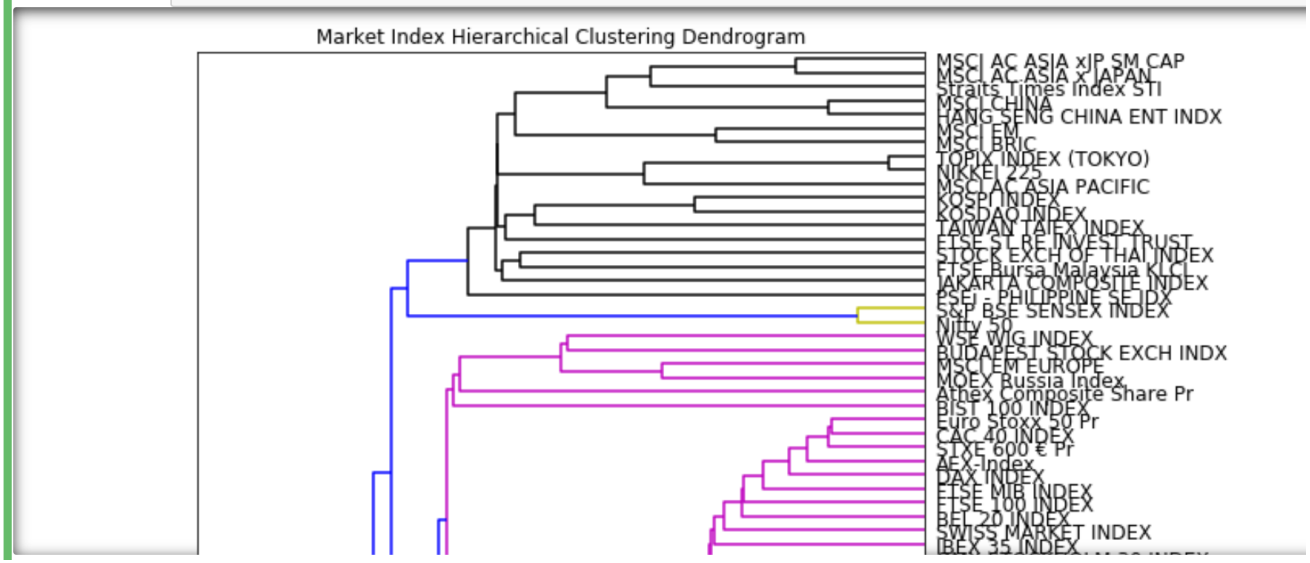
Dendogram


Dendrogram of the Dummy Data
Euclidean Distance
100
80
60
40
20
0
Cut-off point
Dave
Ben
Eric
Anne
Chad


Dendrogram of the Dummy Data
Euclidean Distance
100
80
60
40
20
0
Cut-off point
Dave
Ben
Eric
Anne
Chad

Dendogram


Agglometative Clustering Output as Dendrogram Example
Root
Node
Leaf
70
60
50
40
30
20
10
0
640
(18)
483
(20)
(26)
(24)
(53)
(18)
(6)
(4)
(60)
(109)
(52)
(35)
(35)
(71)
(19)
(80)
(27)
(118)

Dendogram

In [33]:
```
from scipy import stats
import scipy.cluster.hierarchy as hac

Z_w = hac.linkage(w_corr, method='single')

# Plot dendogram
plt.figure(figsize=(8, 10))
plt.title('Market Index Hierarchical Clustering Dendrogram')

hac.dendrogram(
    Z_w,labels = w_corr.index, orientation="left",
    leaf_rotation=0,  # rotates the x axis labels
    leaf_font_size=12.,  # font size for the x axis labels
)
plt.show()
```