Clustering Considerations for Machine Learning
With examples from exploration data

Philip Lesslar

Digital Energy Journal Forum 2019
3rd October 2019
ADAX Center, Bangsar South
Kuala Lumpur, Malaysia

Precision-DM

Key messages

- Focus is only on clustering
- Understand internals to maximise ML effectiveness
- Classification is a big field
- Data analysis is not for the faint-hearted
- Usage with some example exploration data

Machine Learning

Classification:
Creating meaningful groups out of a collection of objects

Build the Model:
Feature extraction to enable effective identification of new objects

Identification:
Use the model to identify new objects to one of the groups

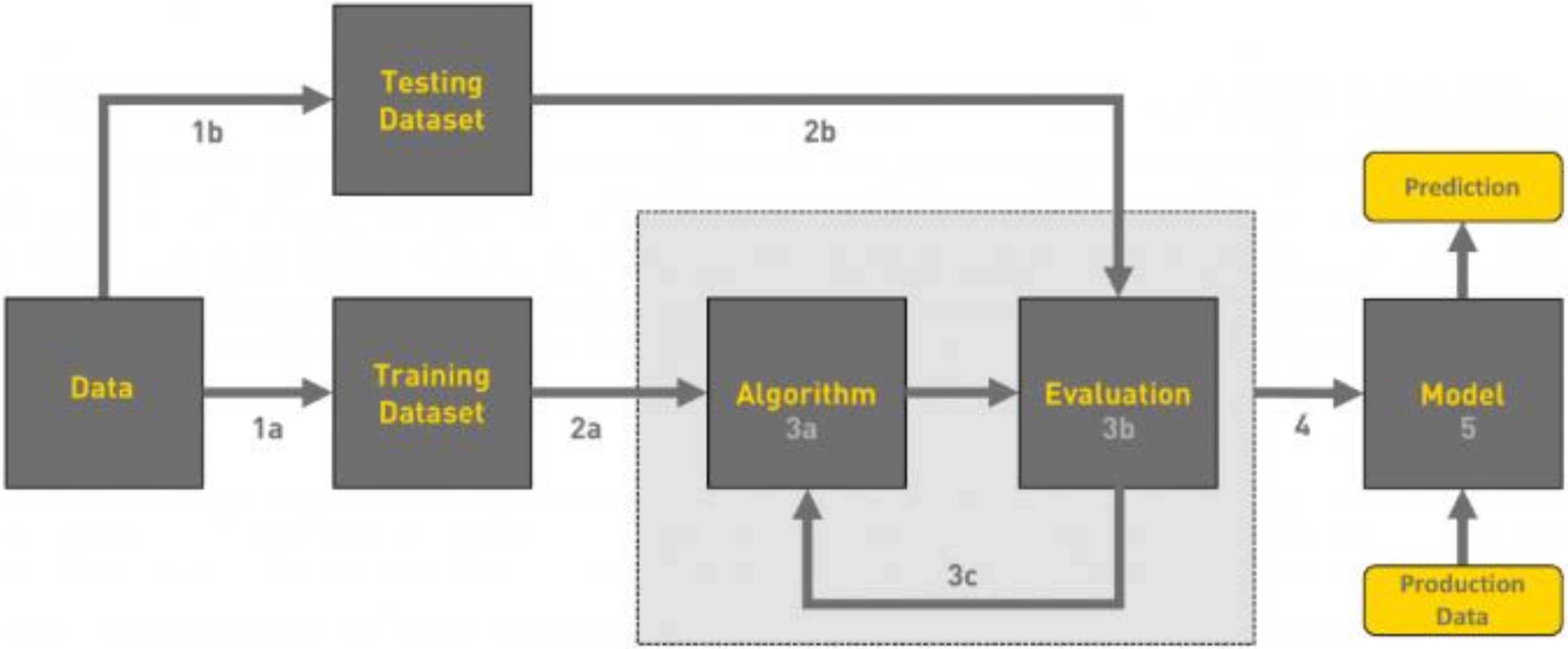
Unsupervised learning

Training
(Model building)

Testing

Supervised learning

The Machine Learning Workflow



https://towardsdatascience.com

Multivariate methods for classification and dimensionality reduction

- Cluster analysis
  - *Finding "natural" or pre-determined groups in datasets*
- Principal components analysis
  - *Reducing the dimensionality of a data set by finding a smaller set of variables that still represents it*
- Factor analysis
  - *For data sets where a large number of observed variables are thought to reflect a smaller number of unobserved/latent variables.*
- Multi dimensional scaling
  - *Technique for visualising the level of similarity of samples transformed onto a 2D plane*
- Linear & Multiple Regression
  - *One or more independent variables are used to predict the value of a dependent variable*

Precision-DM

Sources: "Numerical Taxonomy", Sneath & Sokal 1973. "Cluster Analysis", Everitt et al, 2011, Wikipedia.

Some approaches to Clustering

- K-Means
  - *Iterative computing of distances between points and group means. Requires specification of number of groups.*
- Mean Shift Clustering
  - *Sliding iterative method to find point groups of higher mean density.*
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
  - *Similar to Mean Shift but will identify noise and outliers.*
- Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)
  - *Uses Gaussian approach to define clusters and uses both mean and std deviation unlike K-Means which only uses means. Detects elliptical clusters*
- Agglomerative Hierarchical Clustering
  - *Progressive pairwise clustering until all are merge into one tree in a dendrogram. Not too sensitive to choice of coefficient.*

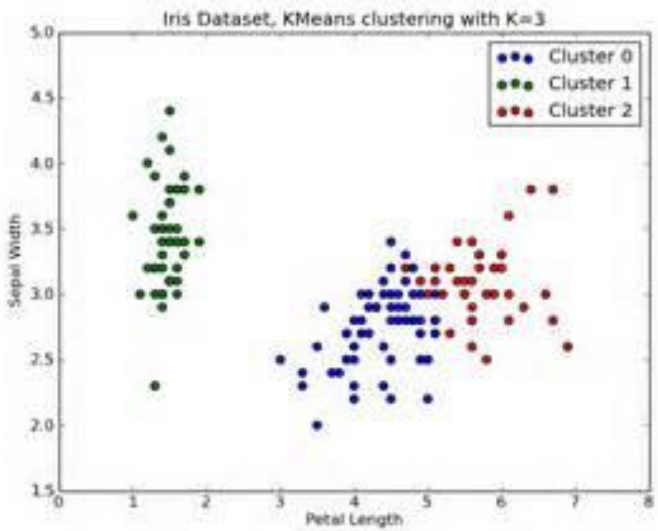
Precision-DM

https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

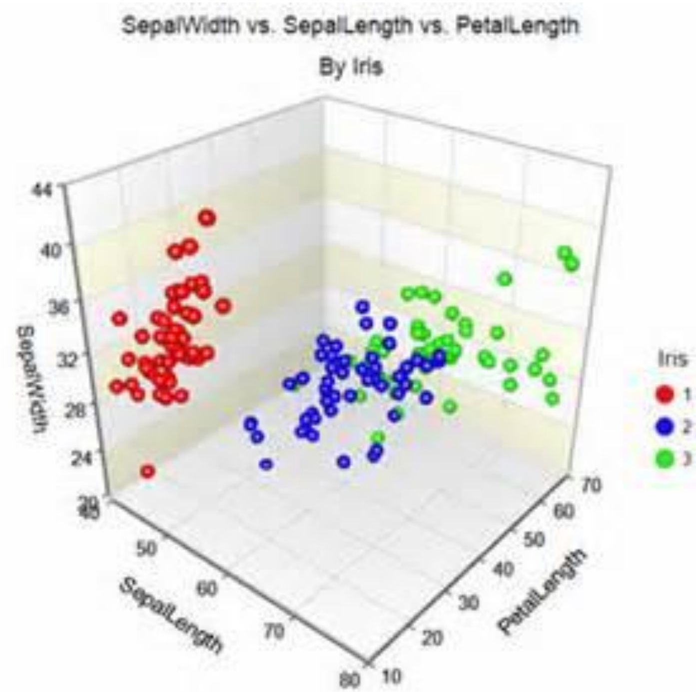
Cluster Analysis – Separating variables in n-dimensions

Visualization

2 dimensions



3 dimensions



4, 5, ……, n dimensions?

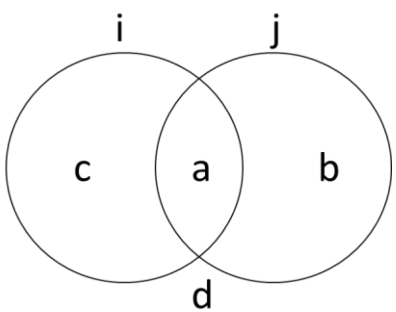
Cluster analysis requires:

1. Measure of pairwise proximities between points
2. Grouping method

Precision-DM

Proximity measures

Data	Measures of Similarity / Dissimilarity (Distance)	
Binary (presence/absence)	Matching coefficient	$S_{ij} = (a + d) / (a + b + c + d)$
	Jaccard coefficient (1908)	$S_{ij} = a / (a + b + c)$
	Rogers & Tanimoto (1960)	$S_{ij} = (a + d) / [a + 2(b + c) + d]$
	Sneath & Sokal (1973)	$S_{ij} = a / [a + 2(b + c)]$
	Gower & Legendre (1986)	$S_{ij} = (a + d) / [a + ½(b + c) + d]$
		$S_{ij} = a / [a + ½(b + c)]$
Continuous	Euclidean Distance *Distance between vectors x & y*	$d(x,y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$
	Canberra Distance *Distance between vectors u & v*	$d(u,v) = \sum_i \frac{\lvert u_i - v_i \rvert}{\lvert u_i \rvert + \lvert v_i \rvert}$

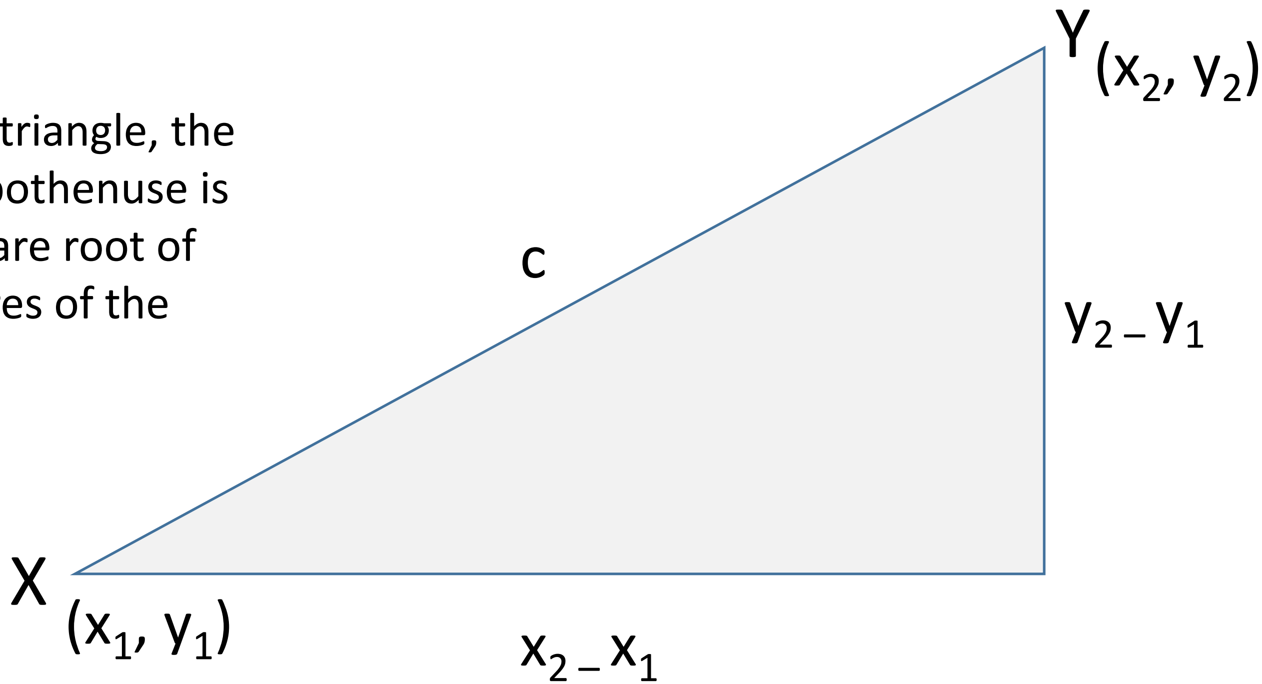


Precision-DM

Source: Cluster Analysis. Everitt, Landau, Leese & Stahl. 5th Edition, Wiley, 2011

Proximity measures - Euclidean Distance – Pythagoras's Theorem

In a right angled triangle, the length of the hypothenuse is equal to the square root of the sum of squares of the other 2 sides



$$C = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The Euclidean Distance $d(x, y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2}$ $= C,\ n = 2$

Examples from Exploration data

1. Prospect Appraisal – Expectation values
2. Well logs – Curve values
3. Micropaleontology – Foraminiferal assemblages

Precision-DM

Exploration Prospect Appraisal

LEAD → PROSPECT → WELL

DATA

Seismic interpretation
Geological picks & zones
Paleontology (incl. palyn, nanno etc)
Lithology & Lithofacies
Environments of deposition
Temperature
etc

→ Prospect Appraisal System →

Probabilistic
- Bootstrap
- Monte Carlo

Expectations

	Cutoffs
POS	
MSV	0 mbbls
HSV	30 mbbls
REC	0 bcf/tcf
STOIIP	
GIIP	

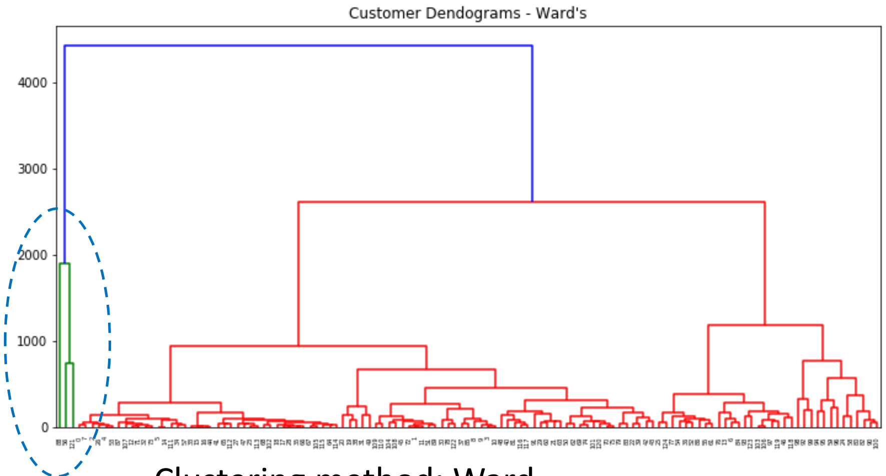
Precision-DM

Exploration Prospect Appraisal – The DATA

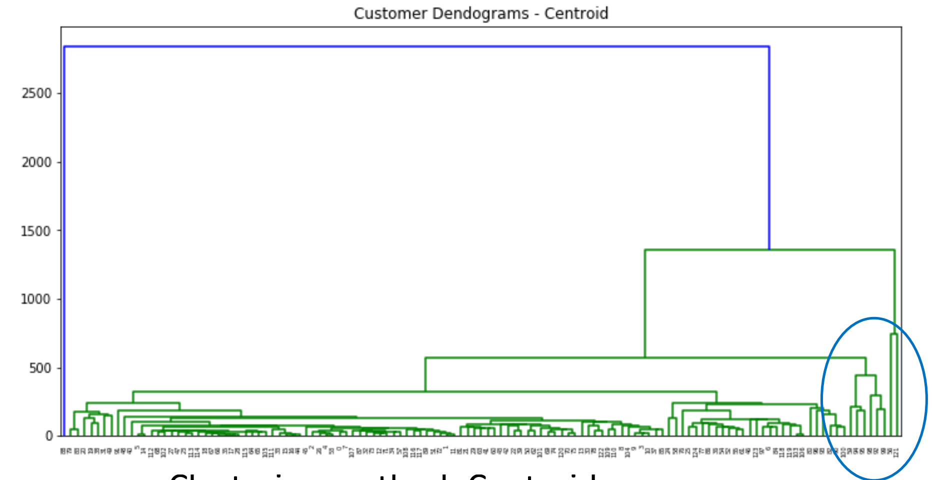
OIL (0 mmbbls cutoff)					OIL (30 mmbbls cutoff)			GAS (0 bscf cutoff)					(values/POS)	
POS	MSV	HSV	Expectation REC.	STOIIP	POS	MSV	HSV	POS	MSV	HSV	Expectation Rec.	GIIP	MSV STOIIP	MSV GIIP
80	6	10	5	24	1	21	0	96	79	133	76	122	30	127
64	11	26	7	23	10	38	60	64	25	57	16	27	36	42
68	11	23	8	31	15	29	38	80	41	90	33	55	46	69
85	5	9	4	27	0	0	0	85	15	32	13	25	32	29
72	7	16	5	22	6	29	40	80	27	64	22	36	31	45
78	3	6	2	11	0	0	0	87	13	30	11	18	14	21
80	4	8	3	11	0	0	0	99	29	49	29	49	14	49
81	11	22	9	43	18	28	36	90	55	114	50	82	53	91
26	8	19	2	10	4	29	36	29	35	75	10	16	38	55
65	4	6	2	12	0	0	0	72	34	59	24	34	18	47
80	2	2	1	5	0	0	0	92	6	12	6	9	6	10
85	22	41	18	73	40	36	52	95	113	219	107	184	86	194
48	2	4	1	5	0	0	0	80	18	33	14	29	10	36
48	2	4	1	5	0	0	0	80	18	33	14	29	10	36
90	18	37	16	76	29	37	56	99	53	109	52	88	84	89
84	20	48	17	81	29	47	75	94	57	135	54	92	96	98
81	11	21	9	37	12	26	31	83	61	110	51	91	46	110
81	11	21	9	37	12	26	31	83	61	110	51	91	46	110
80	12	24	9	46	16	28	37	90	61	125	55	92	58	102
80	12	24	9	46	16	28	37	90	61	125	55	92	58	102
67	6	11	4	17	1	27	34	80	29	61	23	36	25	45

The purpose: Exploring ‘natural’ groups of prospects may trigger ideas

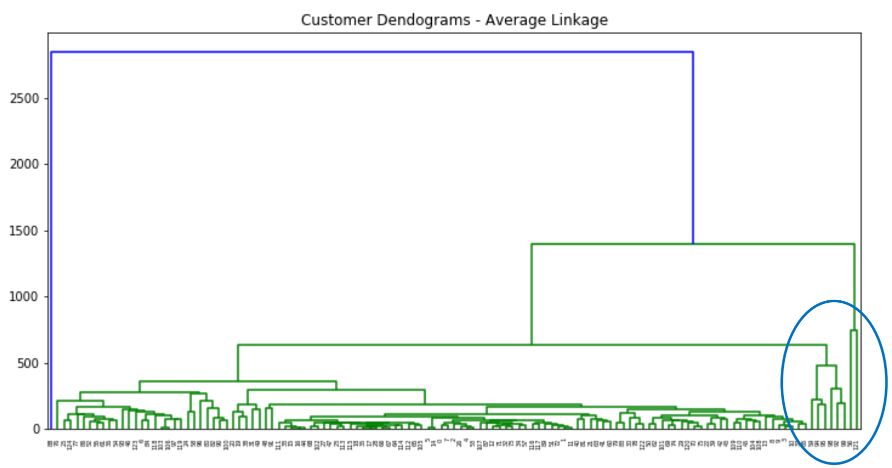
Exploration Prospect Appraisal - Clustering



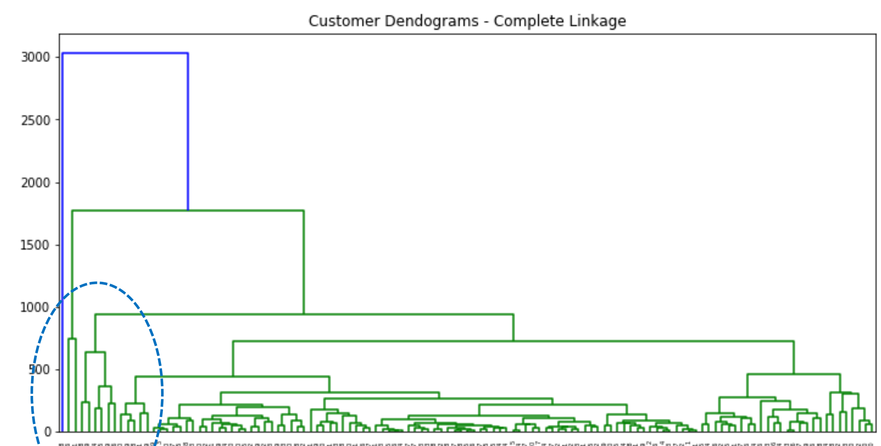
Clustering method: Ward
Coefficient: Squared Euclidean Distance



Clustering method: Centroid
Coefficient: Squared Euclidean Distance



Clustering method: Average Linkage
Coefficient: Squared Euclidean Distance



Clustering method: Complete Linkage
Coefficient: Squared Euclidean Distance

Precision-DM

Cluster analysis using Spyder / Anaconda
Scipy.cluster.hierarchy.dendrogram

1. **Not very distinct clusters**
2. **Review data to remove non-discriminatory data**
3. **Rerun and review**

Well Curves – The DATA

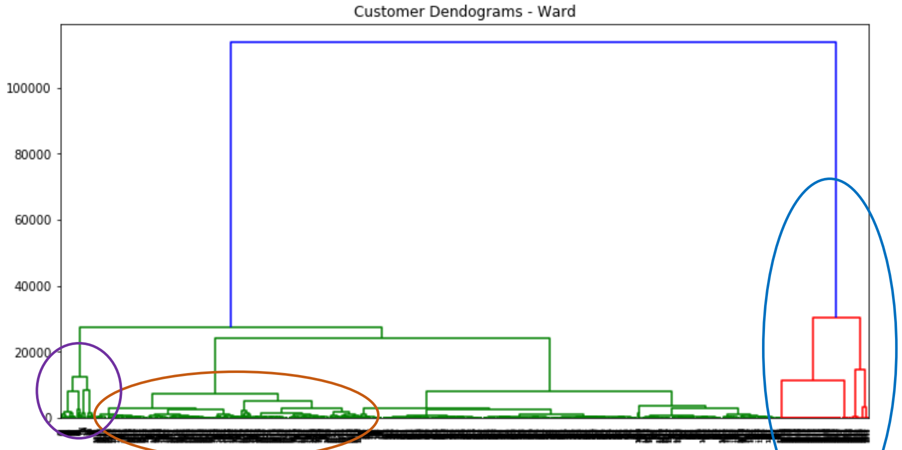
Depth	SGRC	SGRA	SGRB	SEXP	SESP	SEMP	SEDP	SEXC	SESC	SEMC	SEDC	SEDA	STEM	SDDE	SPLF	SNNA	SNFA	SBDC	SCOR	SBD2	SCO2	SNBD	SFBD	SNPE	SHSI
(ft)	(api)	(api)	(api)	(ohmm)	(ohmm)	(ohmm)	(ohmm)	(ohmm)	(ohmm)	(ohmm)	(ohmm)	(ohmm)	(degF)	(ptpf)	(v/v)	(cp30)	(cp30)	(g/cc)	(g/cc)	(g/cc)	(g/cc)	(g/cc)	(g/cc)	(b/e)	(in)
1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
10291	-999.25	-999.25	-999.25	0.06	0.06	120.34	975	0.09	0.09	36.32	194.42	194.42	142.46	-999.25	0.5	2440	475	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10291.5	-999.25	-999.25	-999.25	0.06	0.06	105.39	981.11	0.09	0.09	34.45	193.68	193.68	142.6	-999.25	0.5	2445	475	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10292	-999.25	-999.25	-999.25	0.06	0.06	84.5	986.24	0.09	0.09	31.12	191.57	191.57	142.77	-999.25	0.5	2457	474	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10292.5	-999.25	-999.25	-999.25	0.06	0.06	52.02	952.05	0.09	0.09	28.88	188.31	188.31	142.95	-999.25	0.5	2467	472	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10293	-999.25	-999.25	-999.25	0.06	0.06	32.12	927.16	0.09	0.09	28.85	186.63	186.63	143.16	-999.25	0.5	2470	472	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10293.5	-999.25	-999.25	-999.25	0.06	0.06	27.58	972.77	0.09	0.09	26.99	187.16	187.16	143.84	-999.25	0.5	2470	472	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10294	-999.25	-999.25	-999.25	0.06	0.06	31.96	1047.71	0.09	0.09	21.46	188.23	188.23	144.35	-999.25	0.49	2475	476	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25
10294.5	-999.25	-999.25	-999.25	0.06	0.06	64.08	1005.84	0.09	0.09	17.79	190.26	190.26	144.8	-999.25	0.49	2475	482	-999.25	-999.25	-999.25	-999.25	-999.25	-999.25	11.34	-999.25
10295	-999.25	-999.25	-999.25	0.06	0.06	125.01	886.83	0.09	0.09	13.62	192.59	192.59	145.28	5.66	0.48	2471	487	2.24	-0.29	-999.25	-999.25	-999.25	2.52	11.31	8.5
10295.5	-999.25	-999.25	-999.25	0.06	0.06	241.79	848.52	0.09	0.09	10.03	194.6	194.6	145.93	32.03	0.48	2466	490	2.23	-0.29	-999.25	-999.25	-999.25	2.52	11.32	8.5
10296	-999.25	-999.25	-999.25	0.06	0.06	480.27	976.85	0.09	0.09	5.39	200	200	146.32	51.08	0.48	2465	491	2.23	-0.29	-999.25	-999.25	-999.25	2.52	11.25	8.5
10296.5	-999.25	-999.25	-999.25	0.06	0.06	318.86	885.12	0.09	0.09	0.39	200	200	146.71	48.06	0.48	2462	492	2.23	-0.29	-999.25	-999.25	-999.25	2.52	11.22	8.5
10297	-999.25	-999.25	-999.25	0.06	0.06	188.62	966.88	0.09	0.09	0.45	200	200	147.06	63.25	0.48	2462	494	2.23	-0.29	-999.25	-999.25	-999.25	2.52	11.19	8.5
10297.5	-999.25	-999.25	-999.25	0.06	0.06	110.06	1315.63	0.09	0.09	0.36	200	200	147.62	82.68	0.48	2463	495	2.22	-0.29	-999.25	-999.25	-999.25	2.52	11.15	8.5
10298	-999.25	-999.25	-999.25	0.06	0.06	71.02	1518.89	0.09	0.09	0.32	[illegible]	[illegible]	147.99	46.49	0.48	2465	498	2.22	-0.3	-999.25	-999.25	-999.25	2.52	11.09	8.5
10298.5	-999.25	-999.25	-999.25	0.06	0.06	46.32	1565.17	0.09	0.09	0.29	[illegible]	[illegible]	148.39	26.48	0.47	2467	499	2.22	-0.3	-999.25	-999.25	-999.25	2.51	11.07	8.5

```
Depth,SGRC,SGRA,SGRB,SEXP,SESP,SEMP,SEDP,SEXC,SESC,SEMC,SED[illegible]A,STEM,SDDE,SPLF,SNNA,SNFA,SBDC,SCOR,SBD2,SCO2,SNBD,SFBD,SNPE,SHSI
(ft),(api),(api),(api),(ohmm),(ohmm),(ohmm),(ohmm),(ohmm),(oh[illegible]),(ohmm),(ohmm),(ohmm),(degF),(ptpf),(v/v),(cp30),(cp30),(g/cc),(g/cc),
1,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28
10291,-999.25,-999.25,-999.25,0.06,0.06,120.34,975,0.09,0.09,36.32,194.42,194.42,142.46,-999.25,0.5,2440,475,-999.25,-999.25,-999.25,-
10291.5,-999.25,-999.25,-999.25,0.06,0.06,105.39,981.11,0.09,0.09,34.45,193.68,193.68,142.6,-999.25,0.5,2445,475,-999.25,-999.25,-999
10292,-999.25,-999.25,-999.25,0.06,0.06,84.5,986.24,0.09,0.09,31.12,191.57,191.57,142.77,-999.25,0.5,2457,474,-999.25,-999.25,-999.25,
10292.5,-999.25,-999.25,-999.25,0.06,0.06,52.02,952.05,0.09,0.09,28.88,188.31,188.31,142.95,-999.25,0.5,2467,472,-999.25,-999.25,-999
10293,-999.25,-999.25,-999.25,0.06,0.06,32.12,927.16,0.09,0.09,28.85,186.63,186.63,143.16,-999.25,0.5,2470,472,-999.25,-999.25,-999.2
10293.5,-999.25,-999.25,-999.25,0.06,0.06,27.58,972.77,0.09,0.09,26.99,187.16,187.16,143.84,-999.25,0.5,2470,472,-999.25,-999.25,-999
10294,-999.25,-999.25,-999.25,0.06,0.06,31.96,1047.71,0.09,0.09,21.46,188.23,188.23,144.35,-999.25,0.49,2475,476,-999.25,-999.25,-999
10294.5,-999.25,-999.25,-999.25,0.06,0.06,64.08,1005.84,0.09,0.09,17.79,190.26,190.26,144.8,-999.25,0.49,2475,482,-999.25,-999.25,-99
10295,-999.25,-999.25,-999.25,0.06,0.06,125.01,886.83,0.09,0.09,13.62,192.59,192.59,145.28,5.66,0.48,2471,487,2.24,-0.29,-999.25,-999
10295.5,-999.25,-999.25,-999.25,0.06,0.06,241.79,848.52,0.09,0.09,10.03,194.6,194.6,145.93,32.03,0.48,2466,490,2.23,-0.29,-999.25,-99
10296,-999.25,-999.25,-999.25,0.06,0.06,480.27,976.85,0.09,0.09,5.39,200,200,146.32,51.08,0.48,2465,491,2.23,-0.29,-999.25,-999.25,-9
10296.5,-999.25,-999.25,-999.25,0.06,0.06,318.86,885.12,0.09,0.09,0.39,200,200,146.71,48.06,0.48,2462,492,2.23,-0.29,-999.25,-999.25,-
10297,-999.25,-999.25,-999.25,0.06,0.06,188.62,966.88,0.09,0.09,0.45,200,200,147.06,63.25,0.48,2462,494,2.23,-0.29,-999.25,-999.25,-9
10297.5,-999.25,-999.25,-999.25,0.06,0.06,110.06,1315.63,0.09,0.09,0.36,200,200,147.62,82.68,0.48,2463,495,2.22,-0.29,-999.25,-999.25,
10298,-999.25,-999.25,-999.25,0.06,0.06,71.02,1518.89,0.09,0.09,0.32,200,200,147.99,46.49,0.48,2465,498,2.22,-0.3,-999.25,-999.25,-99
10298.5,-999.25,-999.25,-999.25,0.06,0.06,46.32,1565.17,0.09,0.09,0.29,200,200,148.39,26.48,0.47,2467,499,2.22,-0.3,-999.25,-999.25,-
10299,-999.25,-999.25,-999.25,0.06,0.06,0.85,1502.45,0.09,0.09,3.76,200,200,148.73,22.72,0.48,2460,497,2.21,-0.3,-999.25,-999.25,-999
10299.5,-999.25,-999.25,-999.25,0.06,0.06,0.96,1282.31,0.09,0.09,6.06,200,200,149.6,18.31,0.48,2461,495,2.22,-0.29,-999.25,-999.25,-9
10300,-999.25,-999.25,-999.25,0.06,0.06,1.22,1358.33,0.1,0.09,9.82,200,200,149.57,8.71,0.48,2462,495,2.22,-0.29,-999.25,-999.25,-999.
10300.5,-999.25,-999.25,-999.25,0.06,0.07,1.64,1356.55,0.1,0.09,15.38,200,200,149.55,3.72,0.48,2468,491,2.23,-0.28,-999.25,-999.25,-9
10301,-999.25,-999.25,-999.25,0.06,0.07,2,1166.42,0.1,0.09,25.72,200,200,149.55,0.53,0.48,2473,487,2.23,-0.28,-999.25,-999.25,-999.25,
10301.5,-999.25,-999.25,-999.25,0.06,0.07,1.96,1051.09,0.11,0.09,40.57,200,200,149.53,1.97,0.48,2474,484,2.23,-0.28,-999.25,-999.25,-
10302,-999.25,-999.25,-999.25,0.06,0.07,1.57,1018.14,0.11,0.09,42.86,200,200,149.49,2.73,0.48,2479,482,2.22,-0.28,-999.25,-999.25,-99
10302.5,-999.25,-999.25,-999.25,0.07,0.07,1.07,882.06,0.12,0.08,24.52,200,200,149.42,2.36,0.47,2483,488,2.24,-0.26,-999.25,-999.25,-9
```

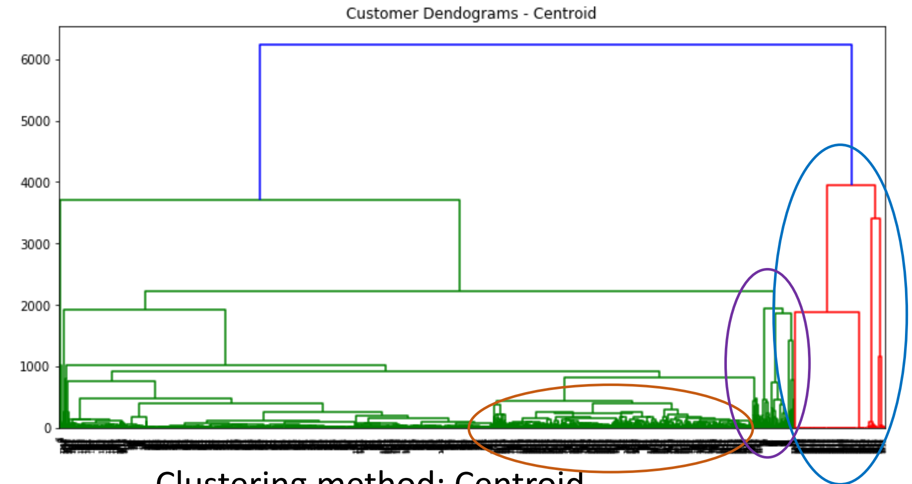
SGRC SGRA SGRB SEXP SESP SEMP SEDP
SEXC SESC SEMC SEDC SEDA STEM SDDE
SPLF SNNA SNFA SBDC SCOR SBD2 SCO2
SNBD SFBD SNPE SHSI

Precision-DM

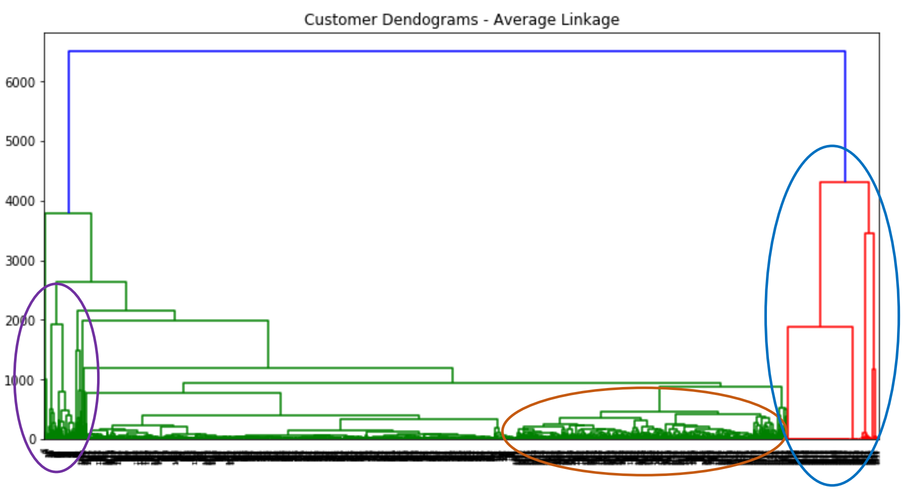
Well Curves – Clustering



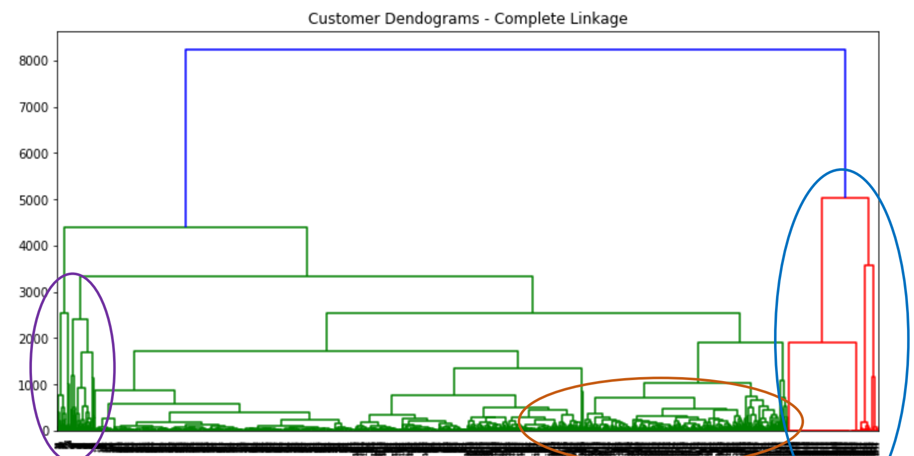
Clustering method: Ward
Coefficient: Squared Euclidean Distance



Clustering method: Centroid
Coefficient: Squared Euclidean Distance



Clustering method: Average Linkage
Coefficient: Squared Euclidean Distance



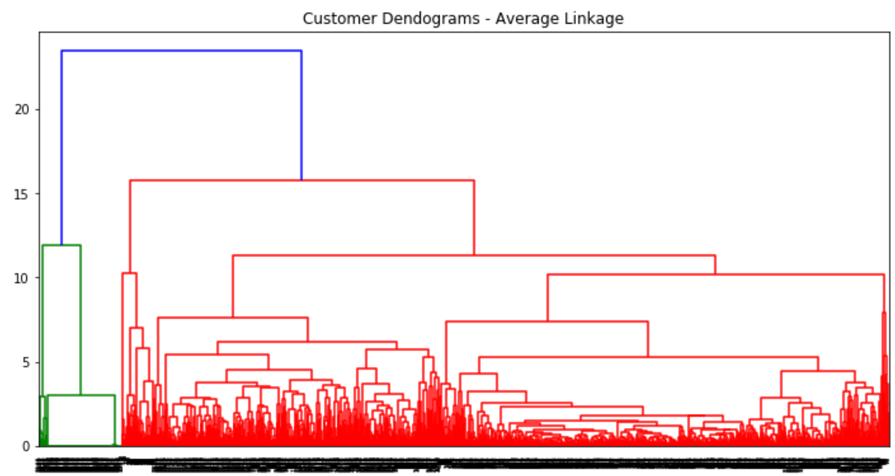
Clustering method: Complete Linkage
Coefficient: Squared Euclidean Distance

Precision-DM

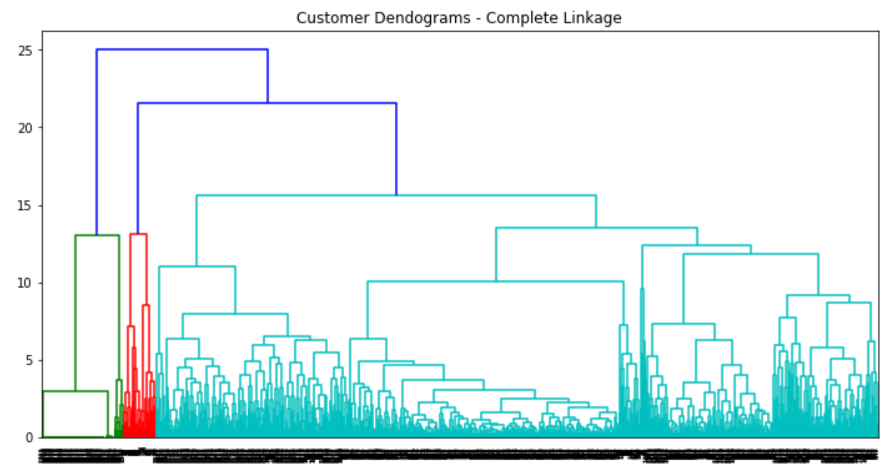
Cluster analysis using Spyder / Anaconda
Scipy.cluster.hierarchy.dendrogram

1. **Some distinct clusters, majority of points are mixed**
2. **Review data to remove non-discriminatory data**
3. **Investigate end points. Rerun and review**

Well Curves – Change of coefficient



Clustering method: Average Linkage
Coefficient: Canberra



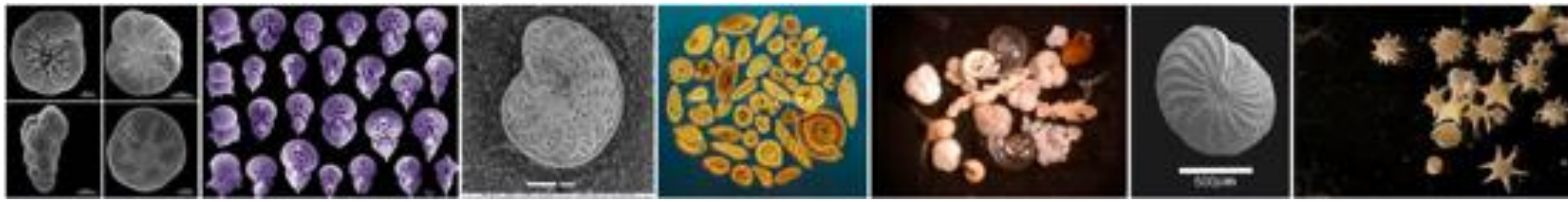
Clustering method: Complete Linkage
Coefficient: Canberra

Precision-DM

Cluster analysis using Spyder / Anaconda
Scipy.cluster.hierarchy.dendrogram

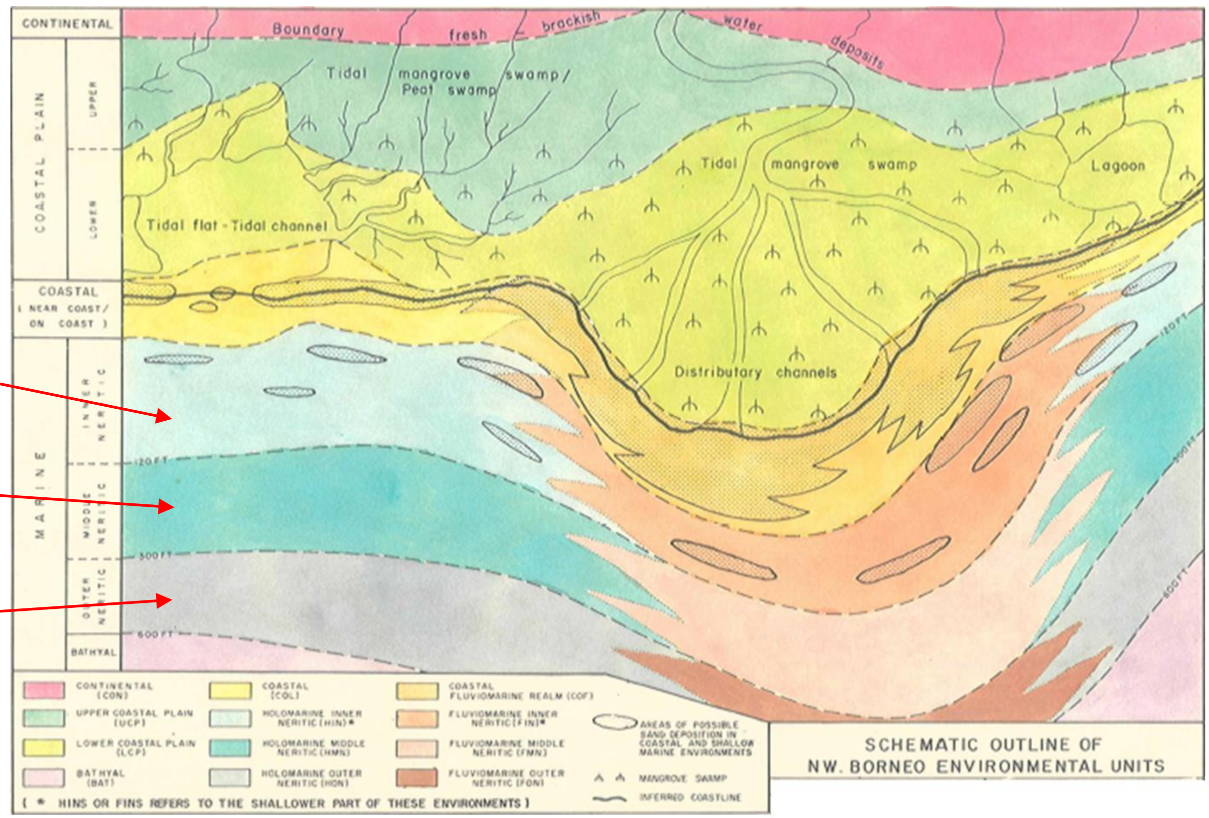
1. **More distinct clusters, easier to differentiate**
2. **Investigate groups for significance**
3. **Review data for noise**

Micropaleontology



Benthonic Foraminifera – Protozoa. Live(d) on the sea bottom. Size ~ 200-2000 microns
Best viewed with binocular microscope at 25x – 80x magnification

North West Borneo Environmental Scheme
(Shell, 1970s)



Holomarine Inner Neritic
0 – 40m water depth

Holomarine Middle Neritic
40 – 100m water depth

Holomarine Middle Neritic
100 – 200m water depth

Fluviomarine realm

Precision-DM

Micropaleontology – The DATA

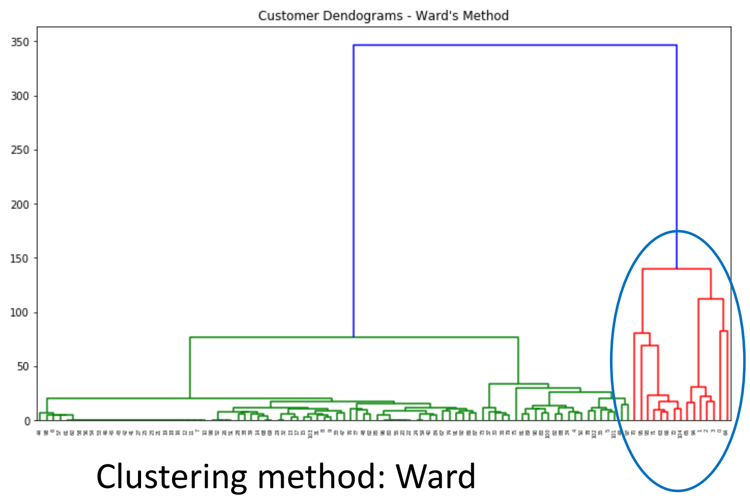
```
03     4933 ART  2         1.00       0
NON1      '      R    'R3         '     C    'R25        '      F    'CI1        '      C    'TX1        '      F    'TX7        '      C    '
BOTA1     '      C    'R2         '     C    'ELPH1      '      C    'BO1        '      C    'NON3       '      R    'END        '
03     4943 ART  2         1.00       0
R25       '      C    'ELPH1      '     C    'R2         '      C    'R3         '      A    'TX1        '      C    'TX7        '      C    '
Q5        '      F    'SGM1       '     F    'GSPP       '      0.0  'AM1        '      C    'SGM3       '      F    'BOTA1      '      C    '
BO1       '      C    'NON1       '     F    'GY1        '      F    'CI1        '      C    'END        '
03     4953 ART  2         1.00       0
TX1       '      C    'TX7        '     F    'Q5         '      F    'R3         '      C    'NON3       '      R    'NON1       '      F    '
GY1       '      R    'R25        '     F    'ELPH1      '      C    'CI1        '      C    'OPSPP      '      F    'SGM5       '      R    '
BOTA1     '      C    'R2         '     F    'END        '
03     4965 ART  2         1.00       0
OPSPP     '      R    'OPLA1      '     R    'ELPH2      '      F    'BO1        '      F    'NON3       '      F    'R2         '      F    '
R3        '      F    'R25        '     F    'ELPH1      '      F    'SGM3       '      R    'BOTA1      '      F    'Q5         '      R    '
CI1       '      F    'END        '
```

```
Depth,NON1,RSPP,R3,R2,AM1,TX7,R4,ELPH1,BO1,NON3,CI1,QSPP,R2V1,GOIDSPP,BO6,GY1,TX1,OPSPP,ELPHSPP,N1,TLSPP,NON2,HM1,SGM5,AM9,GOID8/8A,BO16,SGOID1,AG1,R24,R26V.
504,3.5,3.5,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
600,13,13,60,3.5,60,60,3.5,60,3.5,3.5,3.5,3.5,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
620,13,13,60,3.5,60,3.5,3.5,13,0,3.5,13,3.5,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
630,13,0,60,3.5,60,3.5,3.5,13,3.5,0,3.5,0,3.5,0,3.5,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
640,3.5,0,60,3.5,60,3.5,3.5,3.5,3.5,0,3.5,0,13,0,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
690,0,0,13,0,3.5,0,0,3.5,0,0,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
702,3.5,0,13,3.5,13,3.5,3.5,3.5,0,0,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
850,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
880,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
900,0,3.5,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
990,3.5,3.5,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1010,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1040,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1060,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1070,0,0,0,0,3.5,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
1100,0,0,3.5,0,3.5,0,0,0,0,0,0,0,3.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
```

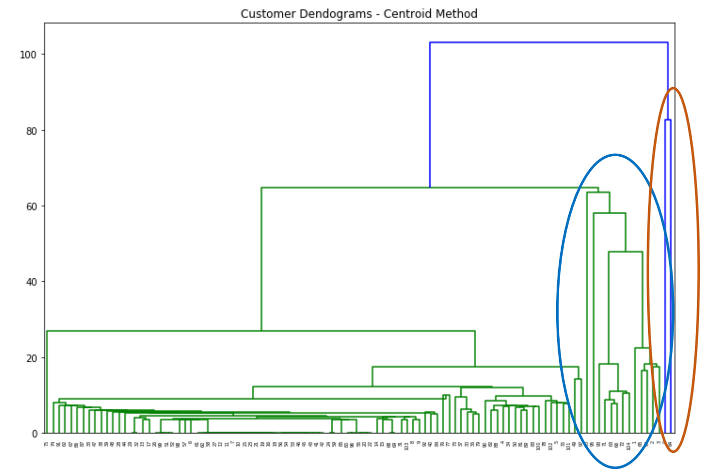
The purpose: Group samples belonging to the same environment of deposition based on species content

Precision-DM

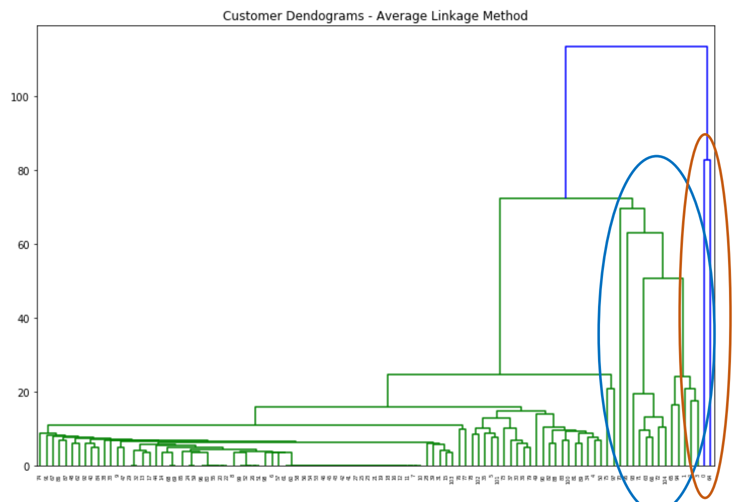
Micropaleontology – Well foraminiferal samples



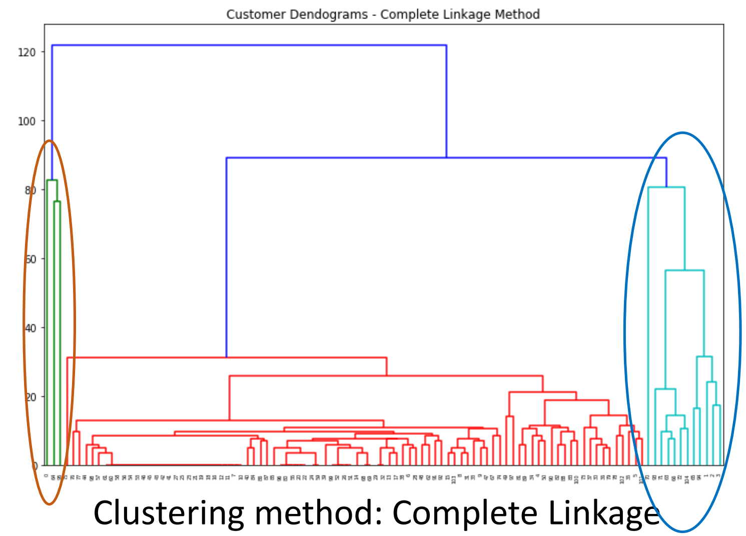
Clustering method: Ward
Coefficient: Squared Euclidean Distance



Clustering method: Centroid
Coefficient: Squared Euclidean Distance



Clustering method: Average Linkage
Coefficient: Squared Euclidean Distance



Clustering method: Complete Linkage
Coefficient: Squared Euclidean Distance

Precision-DM

Cluster analysis using Spyder / Anaconda
Scipy.cluster.hierarchy.dendrogram

1. **Some distinct clusters, mostly mixed**
2. **Investigate groups for significance**
3. **Review data for noise**

Data Science opportunities – Paleoenvironmental reconstruction

Sedimentary facies

-types

-characteristics

-bedding, dips etc

-log shape interpretation

Seismic

-seismic features (seismostrat)

-traces

-Checkshots

-time-depth curve

-Vertical seismic profiling (VSP)

Stratigraphy

-Litho, bio, chrono

-Sea level changes

-flooding surfaces

Paleoenvironments

Structural

-faults

-uplifts

-eustatic

-erosion

-missing sections

Paleontology

-benthics

-planktonics

-larger forams

-nannofossils

-palynology

-ostracods

-trace fossils

Well Logs

-Gamma ray

-Sonic

-Density

-Neutron

-Resistivities

-Caliper

Minerals

-glauconite

-siderite

-pyrite

-mica

Precision-DM

Data Science opportunities– Source Rocks

Pressure
-Spot readings
-Trends

Temperature
-Sample readings
-Gradients

Well Logs
-Gamma ray
-Sonic
-Density
-Resistivities
-Caliper

Sedimentary facies
-types
-characteristics
-bedding, dips etc
-log shape interpretation

Rock properties
-Porosity
-Permeability
-Diagenesis

Macerals
-Organic type (Lip. vs Vit.)
-Kitchen area
-Migration paths
-Maturity levels (DOM, VR/E)

Surrounding wells
-well data
-Source rock distribution patterns
-maps & trends

Source Rocks

Burial History
-Sedimentation rates
-Sediment types
-Missing sections
-Palinspastic reconstruction

Computer simulation
-Methods (eg Migration Models
-Probabliistic vs deterministic

Paleontology
-benthics
-planktonics
-larger forams
-nannofossils
-palynology
-ostracods

Precision-DM

Data Science opportunities– Prospect appraisal

Temperature
-Sample readings
-Gradients

Pressure
-Spot readings
-Trends

Surrounding wells
-Well data
-Correlation
-Maps & trends

Rock properties
-Porosity
-Permeability
-Diagenesis

Analogues
-local comparators
-regional
-global

Sedimentary facies
-Sediment types
-Characteristics
-Bedding, dips etc
-Log shape interpretation

Structural
-faults
-closures
-seals

Prospect Appraisal

Burial History
-Sedimentation rates
-Sediment types
-Missing sections
-Palinspastic reconstruction

Well Logs
-Gamma ray
-Sonic
-Density
-Neutron
-Resistivities
-Caliper

Computer simulation
-Methods (eg Monte carlo)
-Probabliistic vs deterministic

Source Rocks
-Type (lip. vs vit.)
-Kitchen area
-Maturity

Paleontology
-benthics
-planktonics
-larger forams
-nannofossils
-palynology
-ostracods

Precision-DM

Conclusions

- Machine learning is not a black box
- Understand the ML workflow components, behaviors and limitations
- Look at the DATA
- Give importance to feature selection & feature extraction
- Look at the results
- Look at the DATA again

Precision-DM

Questions