Opportunity Landscape for Data Scientists in E&P

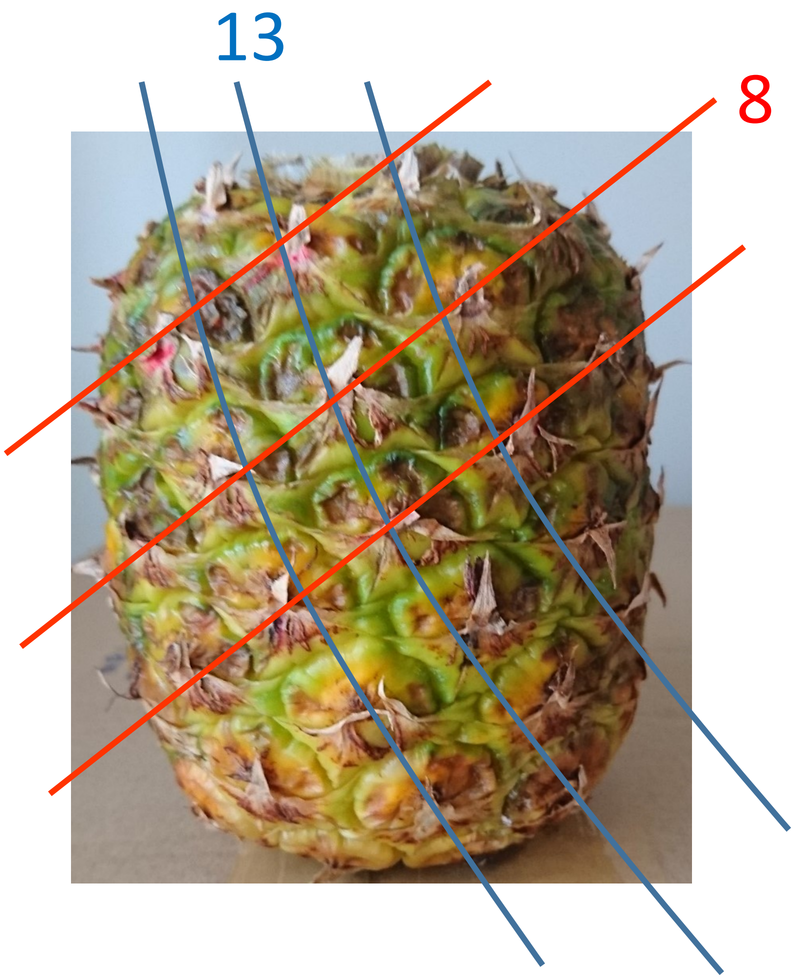
Philip Lesslar
Data Solutions Consultant

Digital Energy Journal Conference
4th October 2017
Impiana Hotel, Kuala Lumpur

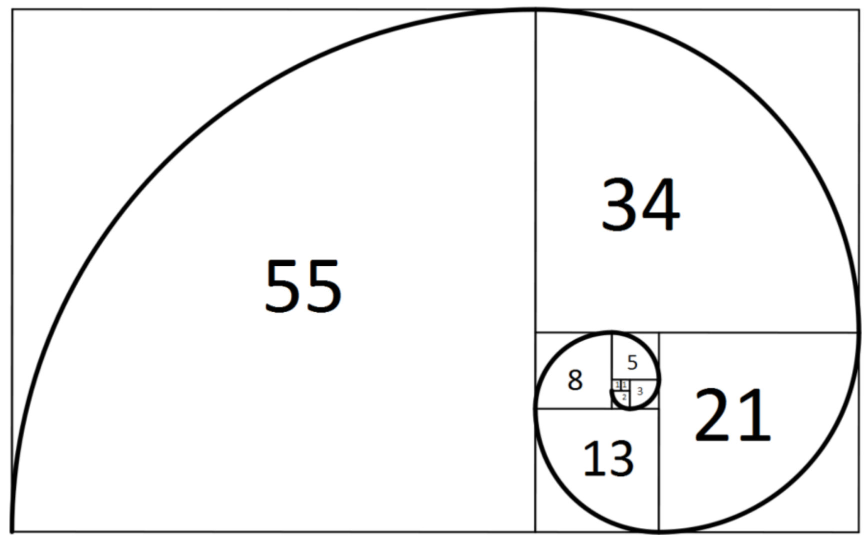
Objectives

- Create an awareness of the vast potential for data science applications in E&P
- Provide some examples of what has been done
- Provide pointers into where the next focus areas could be

Introduction – Looking for patterns



0, 1, 1, 2, 3, 5, 8, 13, 21, 34….



Fibonacci

Look for patterns.

taken a good hard look at what's left once you've finished plucking? A close inspection of the yellow in the middle of the daisy reveals unexpected structure and intrigue. Specifically, the yellow area contains clusters of spirals coiling out from the center. If we examine the flower closely, we see that there are, in fact, two sets of spirals—a clockwise set and a counterclockwise set. These two sets of spirals interlock to produce a hypnotic interplay of helical form.

Interlocking spirals abound in nature. The cone flower and the sunflower both display nature's signature of dual, locking spirals. Flowers are not the only place in nature where spirals occur. A pinecone's exterior is composed of two sets of interlocking spirals. The rough and prickly facade of a pineapple also contains two collections of spirals.

Be Specific: Count

In our observations we should not be content with general impressions. Instead, we move toward the specific. In this case we ponder the quantitative quandary: How many spirals are there? An approximate count is: lots. Is the number of clockwise spirals the same as the number of counterclockwise spirals? You can physically verify that the pinecone has 5 spirals in one direction and 8 in the other. The pineapple has 8 and 13. The daisy and cone flower both have 21 and 34. The sunflower has a staggering 55 and 89. In each case, we observe that the number of spirals in one direction is nearly twice as great as the number of spirals in the opposite direction. Listing all those numbers in order we see

5, 8, 13, 21, 34, 55, 89.

Is there any pattern or structure to these numbers?

Suppose we were given just the first two numbers, 5 and 8, on that list of spiral counts. How could we use these two numbers to build the next number? How can we always generate the next number on our list?

We note that 13 is simply 5 plus 8, whereas 21, in turn, is 8 plus 13. Notice that this pattern continues. What number would come after 89? Given this pattern, what number should come before 5? How about before that? How about before that? And before that?

Leonardo's Legacy: The Fibonacci Sequence

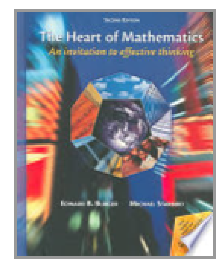
The rule for generating successive numbers in the sequence is to add up the previous two terms. So the next number on the list would be $55 + 89 = 144$.

Through spiral counts, nature appears to be generating a sequence of numbers with a definite pattern that begins

1 1 2 3 5 8 13 21 34 55 89 144 . . . .

Pineapple: 8,13
Daisy: 21,34
Sunflower: 55,89

Leonardo of Pisa, or Fibonacci



G+

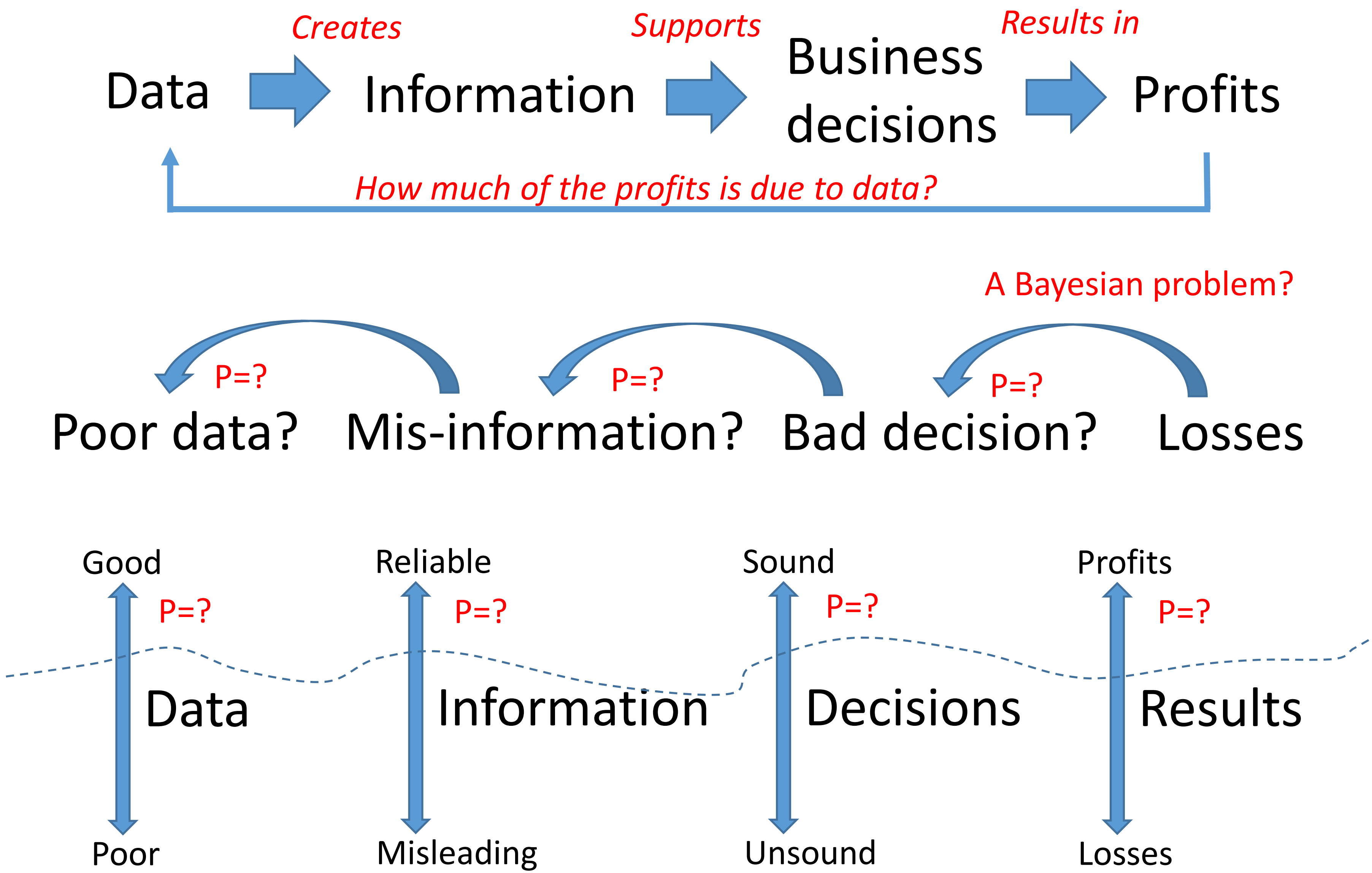
2 Reviews

Write review

The Heart of Mathematics: An invitation to effective thinking

By Edward B. Burger, Michael Starbird

Purpose of data

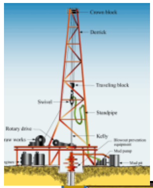


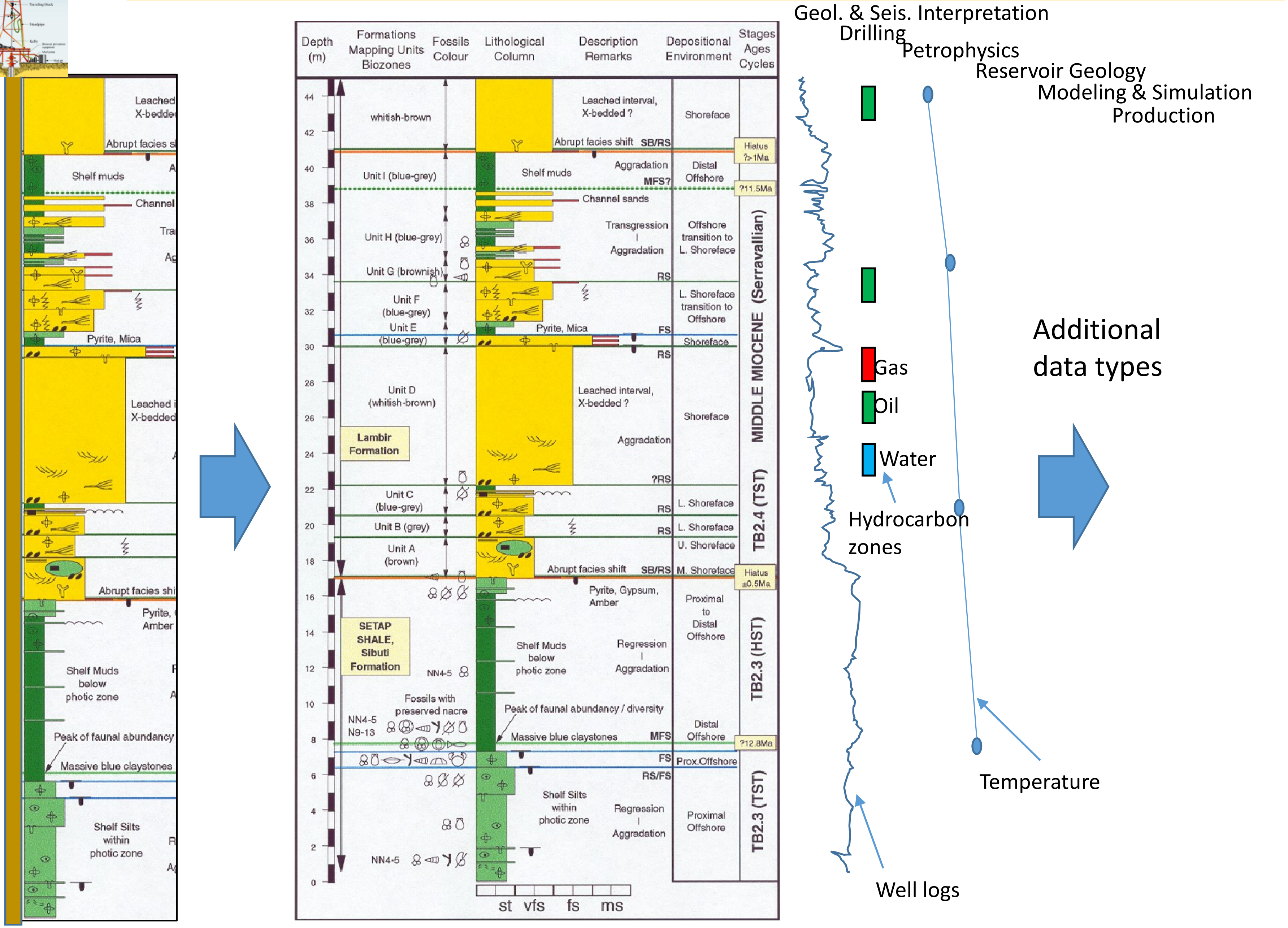
The Upstream Value Chain

	Data aspects
Review	Regional studies, data rooms
Acquire	Acreage, production sharing contracts, seismic (2D, 3D, OBC etc), data purchase, exchanges
Explore	Regional reviews and compilations, play & prospect identification, well locations, well data, correlation
Appraise	Additional well planning & data, detailed studies and correlation, geological modeling, volumetrics,
Develop	Detailed interpretation and analysis, modeling and simulation, real time automation & control.
Produce	Production management, forecasting and economics.
Abandon	Data consolidation & achival.

Data increase through the well life cycle

Phases -> Exploration Appraisal Production





Typical Problems encountered in E&P Data

Physical Data	Electronic Data
• Sampling (accuracy) difficulty due to lack of hole integrity (ditch cuttings) • Contamination of ditch cuttings due to excessive cavings • Poor sample recovery (sidewall samples, cores, fluids) – both % recovery per sample as well as sample loss • Inaccuracy of reading due to inconsistent hole diameters (well logs) • Missing inventory due to poor logistics	• Missing entries • Missing attributes • Inconsistent storage locations in data models • Incorrect values entered • Inconsistent or lack of metadata in entries • Duplication • Large data sets • Distributed or federated data sets and databases • Overlapping data models • Integration challenges • Lack of consistent quality • Data flow breakdowns
People	**Processes & Methodology**
• Resource constraints • Lack of competency • Lack of people framework • Lack of proper accountability structure • Indecision • Office politics	• Lack of governance structure • Lack of standardized workflows • Lack of standards (data, process, systems etc) • Lack of effective data architecture • Lack of transparency • No or loose quantification methodology

Data Types - Upstream

Geology & Seismic	Interpretation and Compilations	Petroleum Engineering	Drilling, Engineering & Production Operations
Well header Info Well Header Spatial Deviation Checkshots Seismic traces (2D & 3D) Mud logs Core description Core Photos Thin Sections / XRD Environments of deposition Prospects & Leads Pore Pressure Temperature – Gradient Temperature – Borehole Geomechanics Geospatial: -Well location Maps -Block Boundaries -Platforms -Pipelines -Geohazards -Site Surveys -Field Outlines -Nett to Gross Thickness Maps -FTG -CSEM -Gravity & Magnetic -Microseismic	Geology – Zones Geology – Markers Faults (Field Extent & Major) Seismic Horizons – Regional Seismic Horizons – Local Velocity Models Structure Maps TZ Curve Gridded Time / Depth Maps Sand Distribution Maps Static Models Dynamic Models Synthetic Seismogram Biostratigraphy – Zones Biostratigraphy – Markers Geology – Zones Geology – Markers	Spill Points (Reqd. by RE) Well Logs – Raw Well Logs – Processed & Qced Well Logs – Interpreted Well Logs – Cased Hole Vertical Seismic Profiling Core Analysis (SCAL RCA, Gamma) Formation Pressure (RFT, MDT) Well Test (DST,FIT) Production Data (Allocated oil/gas/water rates) Production Pressure Data (Well Tubing/Casing Head Pressure) Production Well Test (FBU,PBU,SDS) Artificial Lift Fluid Property Fluid Contacts Stimulation Cases Fluid Composition Material Balance Prosper Models RMS Models Decline Curve Analysis Volumetrics Reserves and Resources EOR Cases Pressure Maintenance Cases Saturation Height Function Leak Off Test PVT	Daily Drilling Data Well Schematics Well Completion Data Well Intervention Data Well Integrity Data Facilities (P&ID, Limit Diagrams) Well design Drilling Fluid Composition Well Completion Cost Casing Data Bit Data BHA (Borehole Analysis) Deviation (Drilling) Well Hydraulics Shallow Hazards Metocean Data eg Climate Facilities As-Built drawings Facilities Info (type, function) Facilities Historical Info Pipeline (flowrate, function) Pipeline (properties) Geotechnical data (general soil, seabed properties)

Data Science Methods

Sequence Analysis
- Markov Chains
- Runs Test
- Least Squares & Regression Analysis
- Splines
- Segmented Sequences & Zonation Analysis
- Auto- and Cross-Correlation
- SemiVariogram
- Spectral Analysis

Spatial Analysis
- Pattern Analysis (Random, Cluster, Nearest Neighbour)
- Analysis of directional data
- Spherical Distributions
- Fractal Analysis
- Shape Analysis
- Contouring, Trend Surfaces & Kriging

Statistics
- Summary Statistics
- Hypothesis Testing
- t-Distribution
- F-Distribution
- Chi Square Distribution
- Chi Square Goodness of fit
- Regression
- Analysis of Variance (ANOVA)
- Non-Parametric Tests
- - (Mann-Whitney, Kolmogorov-Smirnov, Kruskal-Wallis)

Multivariate Data Analysis
- Multiple Regression
- Discriminant Functions
- Cluster Analysis
- Eigenvalues & Eigenvectors
- Factor Analysis (R & Q Mode)
- Principal Components
- Correspondence Analysis
- MultiDimensional Scaling
- Canonical Correlations

Artificial Intelligence
- Classification
- Natural Language Processing
- Machine Learning / Deep Learning
- Text Mining
- Graph Relationships

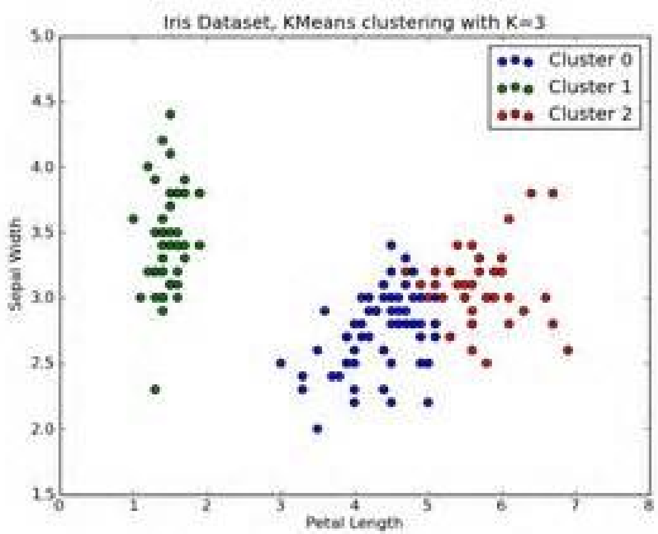
Probabilistic Methods
- Bayesian & Likelihood Methods
- Ranking & Scaling of Events
- Markov Chains

With the possible exception of machine learning / deep learning, all of the above methods have been applied to oil and gas data

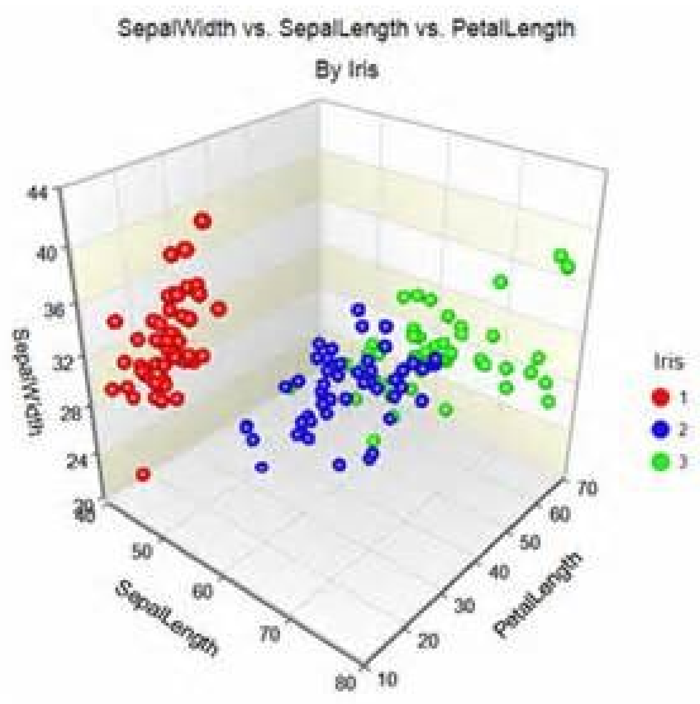
Cluster Analysis – Separating variables in n-dimensions

Visualization

2 dimensions



3 dimensions

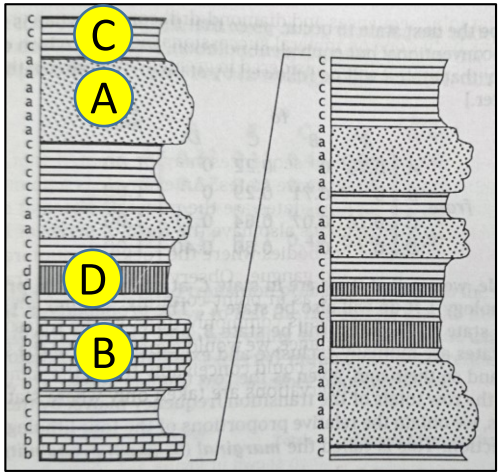


4, 5, ……, n dimensions?

Through the use of dendrograms

Example: Sequence analysis – Non-randomness and layer prediction

Measured stratigraphic section with points measured 1 ft apart



From: Statistics and Data Analysis in Geology, John C. Davis, 2002. Figure 4-5. Measured stratigraphic column in which lithologies have been classified into four mutually exclusive states of sandstones (a), limestones (b), shale ©, and coal (d).

Transition Frequency Matrix

$$
\begin{array}{cc|cccc|c}
 & & & \textit{to} & & & \textit{Row} \\
 & & A & B & C & D & \textit{Totals} \\
\hline
 & A & 18 & 0 & 5 & 0 & 23 \\
\textit{from} & B & 0 & 5 & 2 & 0 & 7 \\
 & C & 5 & 2 & 18 & 3 & 28 \\
 & D & 0 & 0 & 3 & 2 & 5 \\
\hline
\multicolumn{2}{c|}{\textit{Column Totals}} & 23 & 7 & 28 & 5 & 63
\end{array}
$$

Probability Matrix

$$
\begin{array}{cc|cccc|c}
 & & & \textit{to} & & & \textit{Row} \\
 & & A & B & C & D & \textit{Totals} \\
\hline
 & A & 0.78 & 0 & 0.22 & 0 & 1.00 \\
\textit{from} & B & 0 & 0.71 & 0.29 & 0 & 1.00 \\
 & C & 0.18 & 0.07 & 0.64 & 0.11 & 1.00 \\
 & D & 0 & 0 & 0.60 & 0.40 & 1.00
\end{array}
$$

Marginal (or fixed) probability vector obtained by dividing row totals by total number of transitions

$$
\begin{array}{c}
A \\ B \\ C \\ D
\end{array}
\begin{bmatrix}
0.37 \\ 0.11 \\ 0.44 \\ 0.08
\end{bmatrix}
$$

Shows the relative proportions of the 4 lithologies in the sequence

Joint Probability
p(A,B) = p(B|A) p(A)

Therefore, probability that state B will follow, or overlie, state A
P(B|A) = p(B,A) / p(A)

Assuming states are independent:
p(A,B) = p(A) p(B)
And:
P(B|A) = {p(A) p(B)} / p(A) = p(B)

A sequence in which the state at one point is partially dependent, probabilistically, on the previous, is called a **Markov Chain**

Expected Transition Probabilities

$$
\begin{array}{cc|cccc|c|c|cccc}
 & & & \textit{to} & & & & & & & & \\
 & & A & B & C & D & \textit{Totals} & & \multicolumn{4}{c}{\textit{Expected Frequencies}} \\
\hline
 & A & 0.37 & 0.11 & 0.44 & 0.08 & 1.00 & \text{X } 23 = & 8.5 & 2.5 & 10.1 & 1.8 \\
\textit{from} & B & 0.37 & 0.11 & 0.44 & 0.08 & 1.00 & \text{x } 7 = & 2.6 & 0.8 & 3.1 & 0.6 \\
 & C & 0.37 & 0.11 & 0.44 & 0.08 & 1.00 & \text{x } 28 = & 10.4 & 3.1 & 12.3 & 2.2 \\
 & D & 0.37 & 0.11 & 0.44 & 0.08 & 1.00 & \text{x } 5 = & 1.9 & 0.6 & 2.2 & 0.4
\end{array}
$$

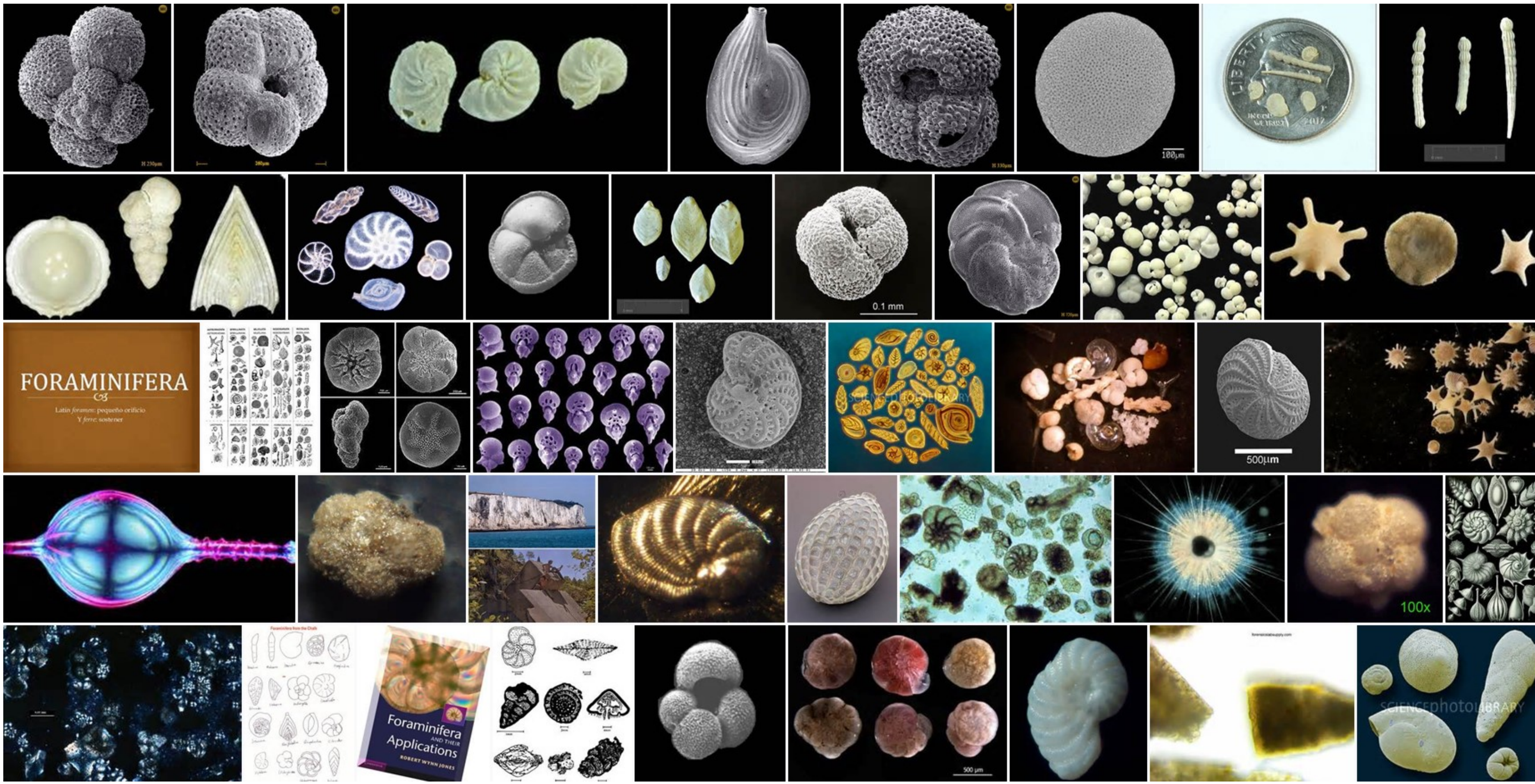
Test for Non-randomness

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 20.9$$

9 deg freedom at 95% significance = 16.92
Conclusion : Sequence is non-random

Example: Interpretation of Depositional Environments - Foraminifera

Foraminifera – Single-celled (Protozoa), marine organisms.
Can be floaters (planktonic) or bottom dwellers (benthonic)

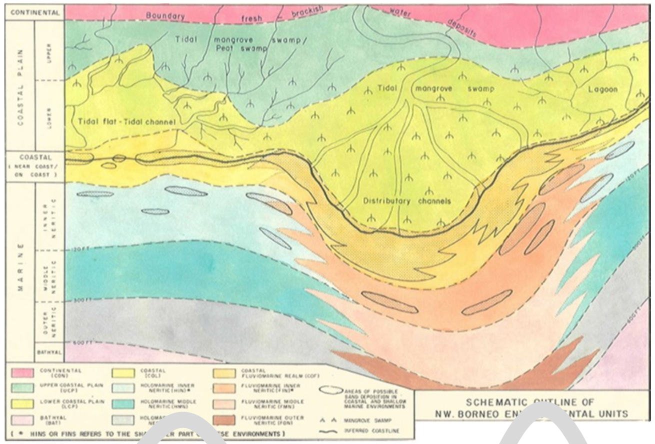


Examples of foraminifera

Example: Interpretation of Depositional Environments - Foraminifera

Input data

- ~2500 samples
- ~1500 species in region
- ~3 million identified specimens in all samples



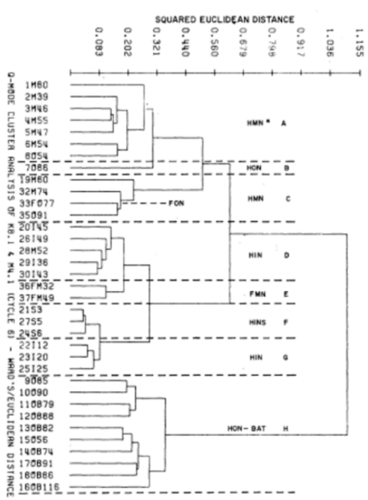
Results

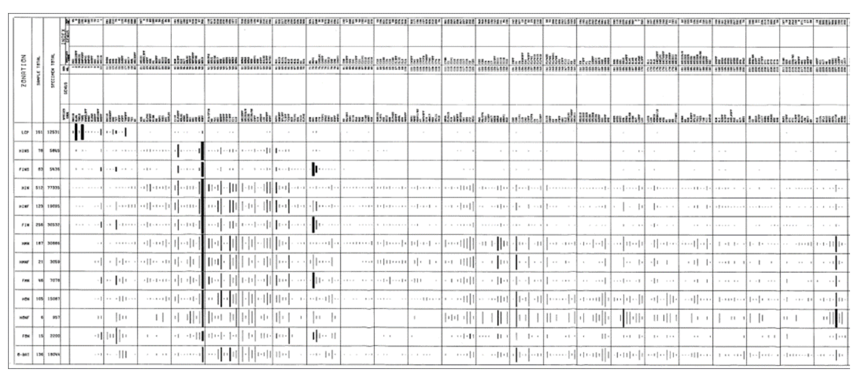
```
SAMPLE =   2638 '        BEST IDENTIFICATION IS .. LCP
=====================    CURRENT INTERPRETATION ..
NO.SPECIES = 5         NO.POSITIVE MATCHES WITH IDENT.MATRIX= 5
NO. SPECIMENS =         28         P/B RATIO =  0.00
DIVERSITY INDICES.  YULE-SIMPSON =  3.60, FISHER ALPHA =  1.02

 TAXA                 WILLCOX PROBABILITY
--------              --------------------
LCP                      1.0000
FINS                     0.0000
HINS                     0.0000
```

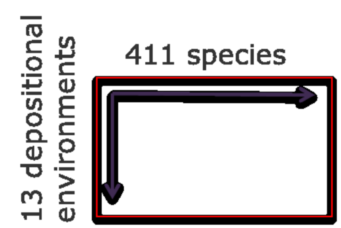
Clustering

Build Probability Matrix of valid clusters

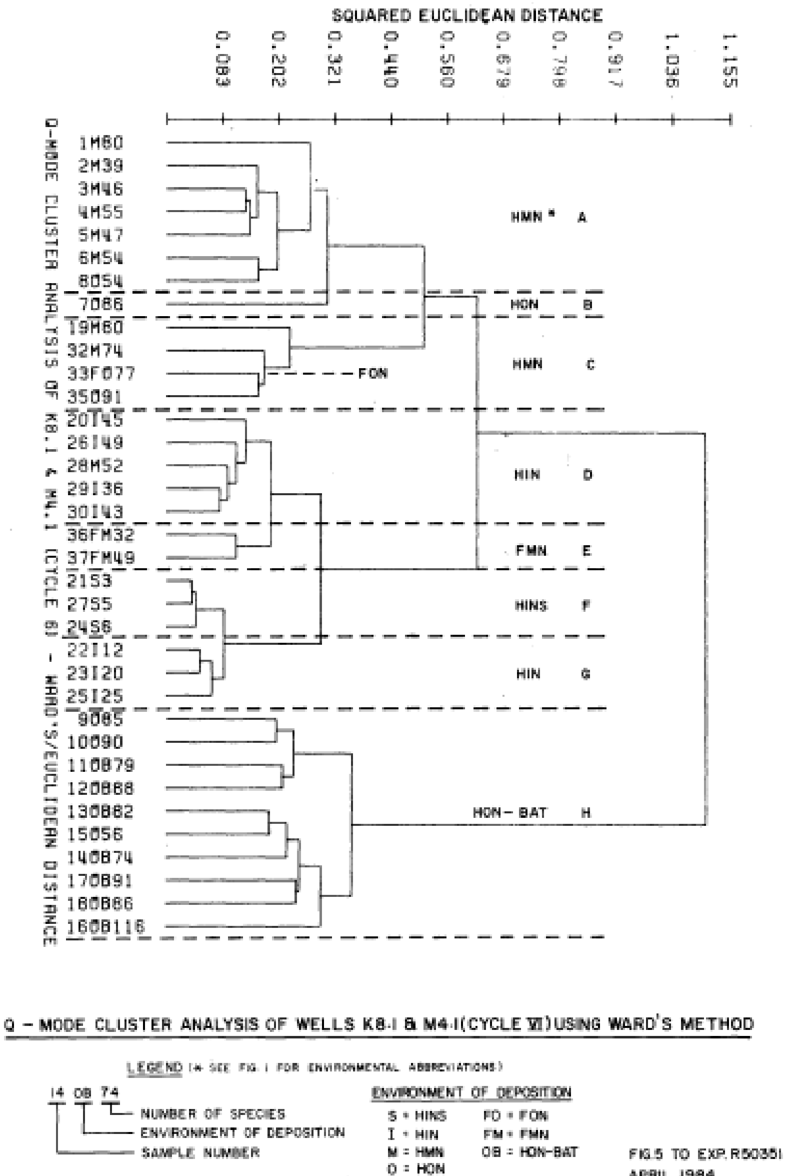
Identification Program - Bayesian Inference (Likelihood Ratio)







Cluster Analysis Example – Environments of Deposition



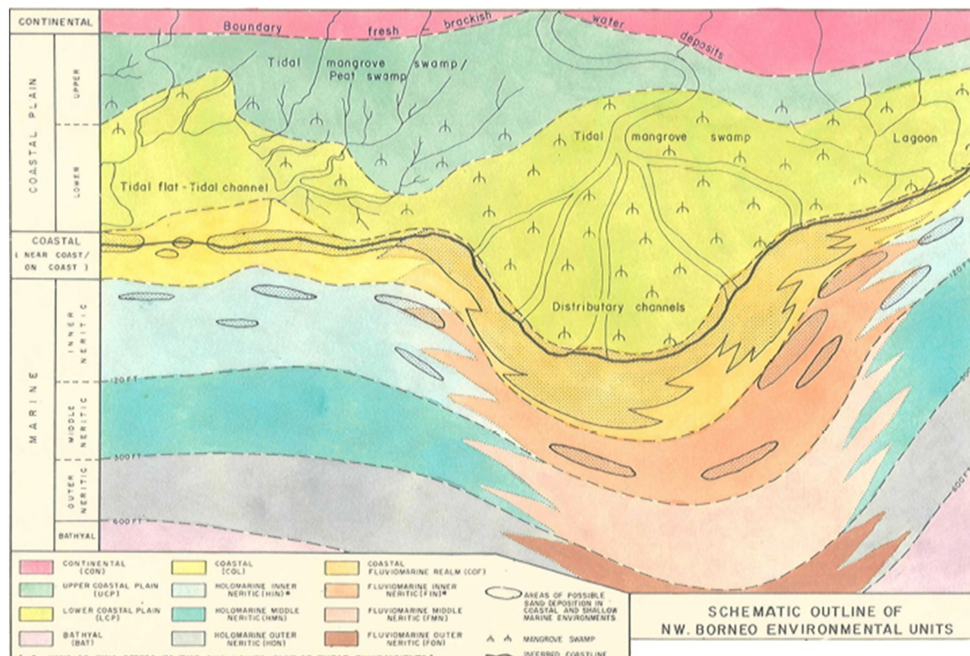
Source:
Computer-assisted interpretation of depositional palaeoenvironments based on foraminifera. Philip Lesslar, Geol. Soc. Malaysia Bulletin 21, December, 1987.

Dendrogram of samples from 1 well using Ward's clustering method and Squared Euclidean Distance coefficient

Cluster analysis is a multivariate technique which allows comparisons and classifications to be done on a set of samples (Q-mode), based on their species content, even when little is known about the structure of the data.

This example is based on foraminiferal presence/absence data.

North West Borneo Environmental Scheme



Next Step – The Identification Matrix

A

SPECIES / ENVIRON.	A	B	. . . . . .	n
1	70	80	. . . . .	
2	60	20	. . . . .	
.	.	.		
.	.	.		
.	.	.		
q	.	.		

Schematic of the identification matrix

B

(16/04/84) RANGE CHART PRINT SARAWAK.MIN=30 SPEC/SAM SPECIES LIST PAGE 1

BSPECIES ACODE GR A	ZONAL CODE	ZONAL SAMPLE COUNT	ZONAL WELL COUNT	ZONAL FORAM SUM	SPECIES SAMPLE COUNT	SPECIES WELL COUNT	ZONAL SPECIMEN COUNT	ZONAL SAMPLE PCNT	ZONAL WELL PCNT	ZONAL SPECIMEN PCNT	SPECIES CODE
BAG1	1	175	33	13891	1	1	2	.6	3.0		AG1
A	2	84	29	6380	9	5	43	9.5	17.2	.7	
A	3	85	31	5608	2	2	5	2.4	6.5	.1	
A	4	552	49	84178	189	36	1493	34.2	73.5	1.8	
A	5	129	40	20194	77	28	774	59.7	70.0	3.8	
A	6	262	46	31674	65	25	477	24.8	54.3	1.5	
A	7	192	29	31961	57	21	336	29.7	72.4	1.1	
A	8	22	10	3263	8	6	34	36.4	60.0	2.6	
A	9	46	12	7078	5	5	47	10.9	41.7	.7	
A	10	105	15	15134	66	11	328	62.9	73.3	2.2	
A	11	6	2	957	1	1	1	16.7	50.0	.1	
A	12	15	5	2200	1	1	6	6.7	20.0	.3	
A	13	136	8	18067	73	7	198	53.7	87.5	1.1	
BAMJIDSPP	1	175	33	13891	4	4	71	2.3	12.1	.5	AMJIDSPP
A	4	552	49	84178	7	7	8	1.3	14.3		
A	5	129	40	20194	3	3	6	2.3	7.5		
A	6	262	46	31674	6	4	19	2.3	8.7	.1	
A	7	192	29	31961	4	2	2	2.1	6.9		
A	8	22	10	3263	2	2	1	9.1	20.0		
A	9	46	12	7078	1	1	1	2.2	8.3		
A	10	105	15	15134	3	3	2	2.9	20.0		
A	11	6	2	957	1	1	3	16.7	50.0	.3	
A	12	15	5	2200	3	3	2	20.0	60.0	.1	
A	13	136	8	18067	8	4	7	5.9	50.0		
BAMJID1	4	552	49	84178	2	2	3	.4	4.1		AMJID1
A	6	262	46	31674	1	1	3	.4	2.2		
A	7	192	29	31961	3	2	1	1.6	6.9		
A	9	46	12	7078	2	2	3	4.3	16.7		
A	10	105	15	15134	2	2	1	1.9	13.3		
A	11	6	2	957	2	2		33.3	100.0		
A	12	15	5	2200	1	1	5	6.7	20.0	.2	
A	13	136	8	18067	9	4	10	6.6	50.0	.1	
BAMDID3	1	175	33	13891	1	1	1	.6	3.0		AMJID3
A	4	552	49	84178	3	2	4	.5	4.1		
A	5	129	40	20194	1	1		.8	2.5		
A	6	262	46	31674	1	1	8	.4	2.2		
A	10	105	15	15134	1	1		1.0	6.7		
A	13	136	8	18067	5	1		3.7	12.5		

Computer listing of the identification matrix

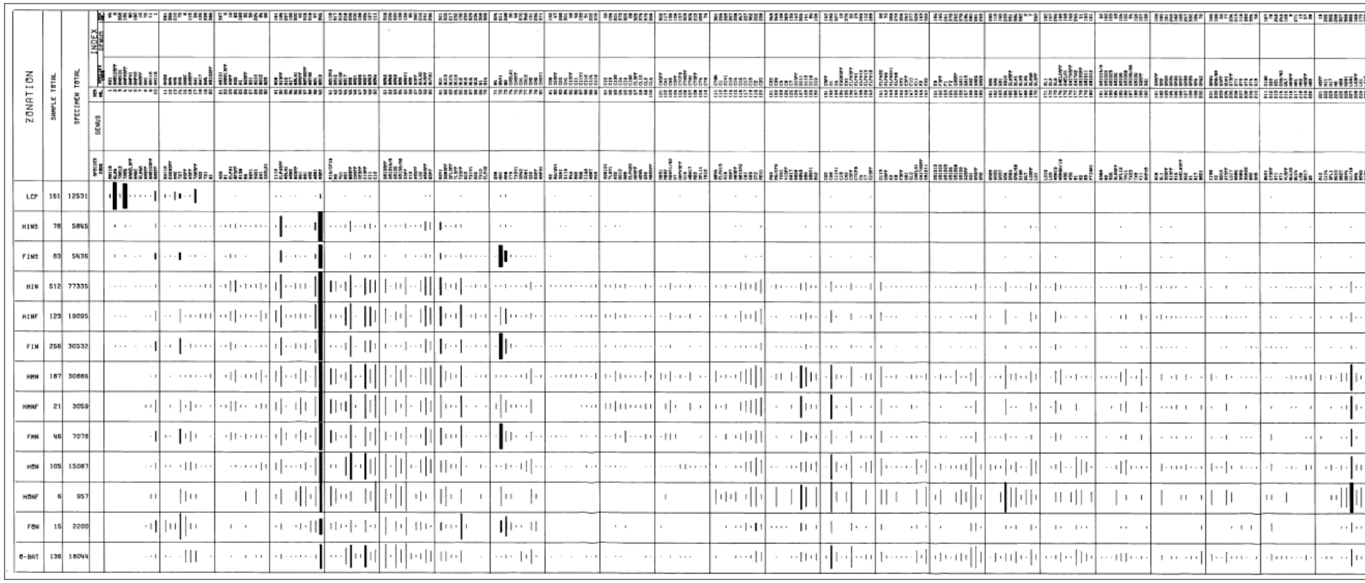
Environmental zones

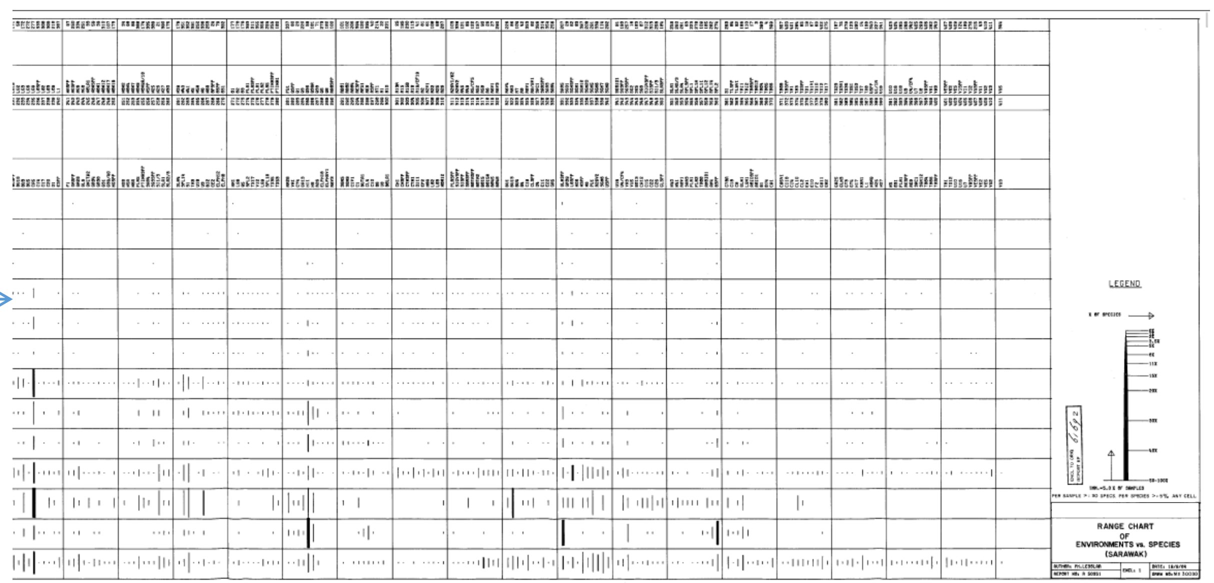
Foraminiferal species

The identification matrix has the form given in Fig.A above where each cell in the q x n matrix contains the percentage of positive occurrence of species in a particular environment.

Incoming samples are mathematically compared against the identification matrix and a set of likelihoods are calculated.

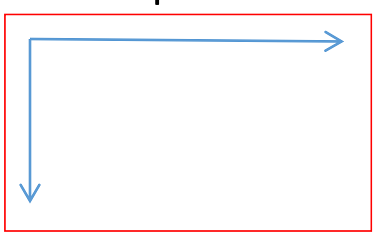
The Identification Matrix (contd)





411 species

13 depositional environments



Probabilistic Approach - Theory

The Willcox Probability is the likelihood of the incoming sample U against environment J divided by the sum of the likelihoods of U against all q environments (Willcox et al, 1973). The likelihood L_{UJ} of U against J is:

$$L_{UJ} = \prod_{i=1}^{n} \left| U_i + P_{ij} - 1 \right|$$

Where U_i represents the i^{th} species in the identification matrix which if present in U is assigned the value 1 otherwise it has the value zero, P_{ij} is the probability of positive occurrence of species i in environment J, and n is the number of species in the identification matrix. When species i in the identification matrix matches up with one in U, then $U_i = 1$ and P_{ij} is used in the calculation. Because the system uses presence-absence species data, the probability of a negative occurrence (species i not present in U) is one minus the probability of a positive occurrence i.e. $(1 - P_{ij})$.

The Willcox Probability of U against J is given by:

$$P_w\ (UJ) = \frac{L_{UJ}}{\sum_{k=1}^{q} L_{UJ_k}}$$

Probabilistic Approach - Results

PROGRAM FOR IDENTIFICATION OF WELL SAMPLES USING
PRESENCE-ABSENCE DATA AGAINST AN IDENTIFICATION MATRIX
OF PERCENT POSITIVE CHARACTERS OF THE TAXA

BY : P.LESSLAR, XGS/1. MODIFIED FROM SNEATH,1979
DATE : 84/10/23 TIME : 07:43:18

THE PROGRAM CALCULATES AND LISTS THE WILLCOX PROBABILITY THAT A GIVEN ASSEMBLAGE BELONGS TO A PARTICULAR TAXON IN THE DATA MATRIX BE IT DEPOSITIONAL ENVIRONMENT, FORAM-BAND OR POLLEN ZONE. DEPENDS ON THE DATA MATRIX USED.

ENTER NAME OF IDENTIFICATION MATRIX TO BE USED

YOUR CHOICES ARE :

- A. CYCLES 1-7 (FORAMS / ENVIRONMENT)
- A1. FAUNAL HORIZONS
- B. BALINGIAN (POLLEN ZONATION)
- C. SARAWAK (POLLEN ZONATION)
- D. SABAH (POLLEN ZONATION)
- E. ARBITRARY (TO BE SPECIFIED YOURSELF)

ENTER A,A1,B,C,D OR E
IDENTIFICATION MATRIX IS : MATBASIC
SPECIES = 411 UNITS = 13
MATBASIC READ IN....
@FORLIST READ IN .....

NAME OF FILE = D9 1
TYPE OF FILE = QUANTITATIVE

TOTAL NUMBER OF SAMPLES = 102 . THEY ARE :

1.	1862	2.	1888	3.	1915	4.	1985
5.	2015	6.	2115	7.	2248	8.	2415
9.	2430	10.	2460	11.	2578	12.	2630
13.	2638	14.	2663	15.	2708	16.	2770
17.	2830	18.	2900	19.	3022	20.	3055
21.	3085	22.	3205	23.	3325	24.	3370
25.	3440	26.	3475	27.	3530	28.	3590
29.	3680	30.	3880	31.	3965	32.	3974
33.	4080	34.	4155	35.	4215	36.	4255
37.	4435	38.	4555	39.	4605	40.	4630
41.	4715	42.	4785	43.	4930	44.	5030
45.	5130	46.	5190	47.	5270	48.	5305
49.	5350	50.	5440	51.	5520	52.	5580
53.	5675	54.	5795	55.	5870	56.	5940
57.	6010	58.	6080	59.	6103	60.	6165
61.	6215	62.	6250	63.	6340	64.	6480
65.	6560	66.	6710	67.	6755	68.	6915
69.	7105	70.	7149	71.	7229	72.	7340
73.	7660	74.	7800	75.	7848	76.	8107
77.	8158	78.	8221	79.	8351	80.	8450
81.	8548	82.	8673	83.	8822	84.	9046
85.	9240	86.	9361	87.	9566	88.	9642
89.	9732	90.	9749	91.	9786	92.	9825
93.	9840	94.	9906	95.	9970	96.	10072
97.	10142	98.	10226	99.	10302	100.	10362
101.	10448	102.	10524	103.		104.	

ANALYSIS BETWEEN SAMPLES 2638 AND 2708

SAMPLE = 2638 ' BEST IDENTIFICATION IS .. LCP
CURRENT INTERPRETATION ..
NO.SPECIES = 5 NO.POSITIVE MATCHES WITH IDENT.MATRIX= 5
NO. SPECIMENS = 28 P/B RATIO = 0.00
DIVERSITY INDICES. YULE-SIMPSON = 3.60, FISHER ALPHA = 1.02

TAXA	WILLCOX PROBABILITY
LCP	1.0000
FINS	0.0000
HINS	0.0000

SPECIES AGAINST ------> LCP

SPECIES	PERCENT IN TAXON	VALUE IN UNKNOWN
AN17	1	+
GLMSPP	9.9	+

SPECIES AGAINST ------> FINS

SPECIES	PERCENT IN TAXON	VALUE IN UNKNOWN
AN17	1	+
GLMSPP	3.6	+
GLM4	9.6	+
TROSPP	7.2	+
TRO5	6	+

SPECIES AGAINST ------> HINS

SPECIES	PERCENT IN TAXON	VALUE IN UNKNOWN
AN17	1	+
GLMSPP	1	+
GLM4	9	+
RSPP	99	-
TROSPP	7.7	+
TRO5	6.4	+

SPECIES	AMT.	SCIENTIFIC NAME
GLMSPP	2	
GLM4	8	MILIAMMINA FUSCA (BRADY)
TROSPP	12	
TRO5	5	TROCHAMMINA MACRESCENS BRADY
AN17	1	

Some Useful Reading

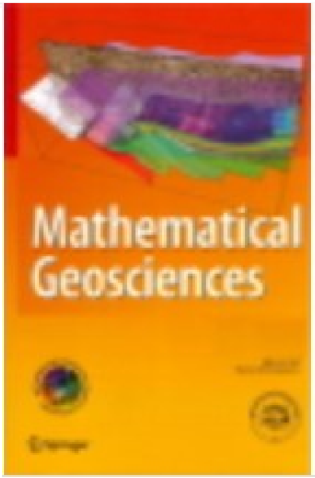
1. Statistics and Data Analysis in Geology. Davis, John C., 3rd Ed. 2002. Wiley
2. Harness Oil & Gas Big Data with Analytics. Holdaway, Keith R., 2014. Wiley
3. Building Expert Systems. Frederick Hayes-Roth, Donald A. Waterman, Douglas B. Lenat, 1983. Addison-Wesley
4. Quantitative Stratigraphy. F.M. Gradstein, F.P.Agterberg, J.C.Brouwer. 1985. Springer
5. Sedimentation Models and Quantitative Stratigraphy. W.Schwarzacher. 1975. Elsevier
6. Cluster Analysis. Brian S. Everitt. 1974. Heinemann Educational Publishers
7. Cluster Analysis 5th Ed. Brian S. Everitt, Sabine Landau, Morven Leese, Danial Stahl. 2011, Wiley
8. Numerical Taxonomy. Peter Sneath, Robert Sokal. 1973. Freeman.

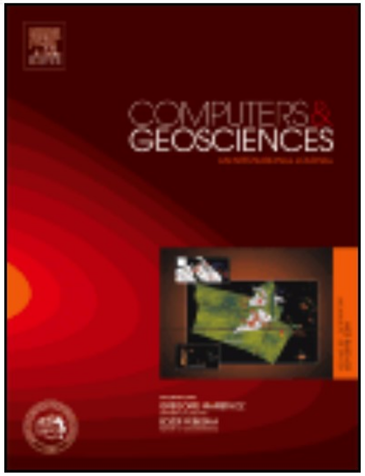
Font size Bigger Reset Smaller

International Association for Mathematical Geosciences

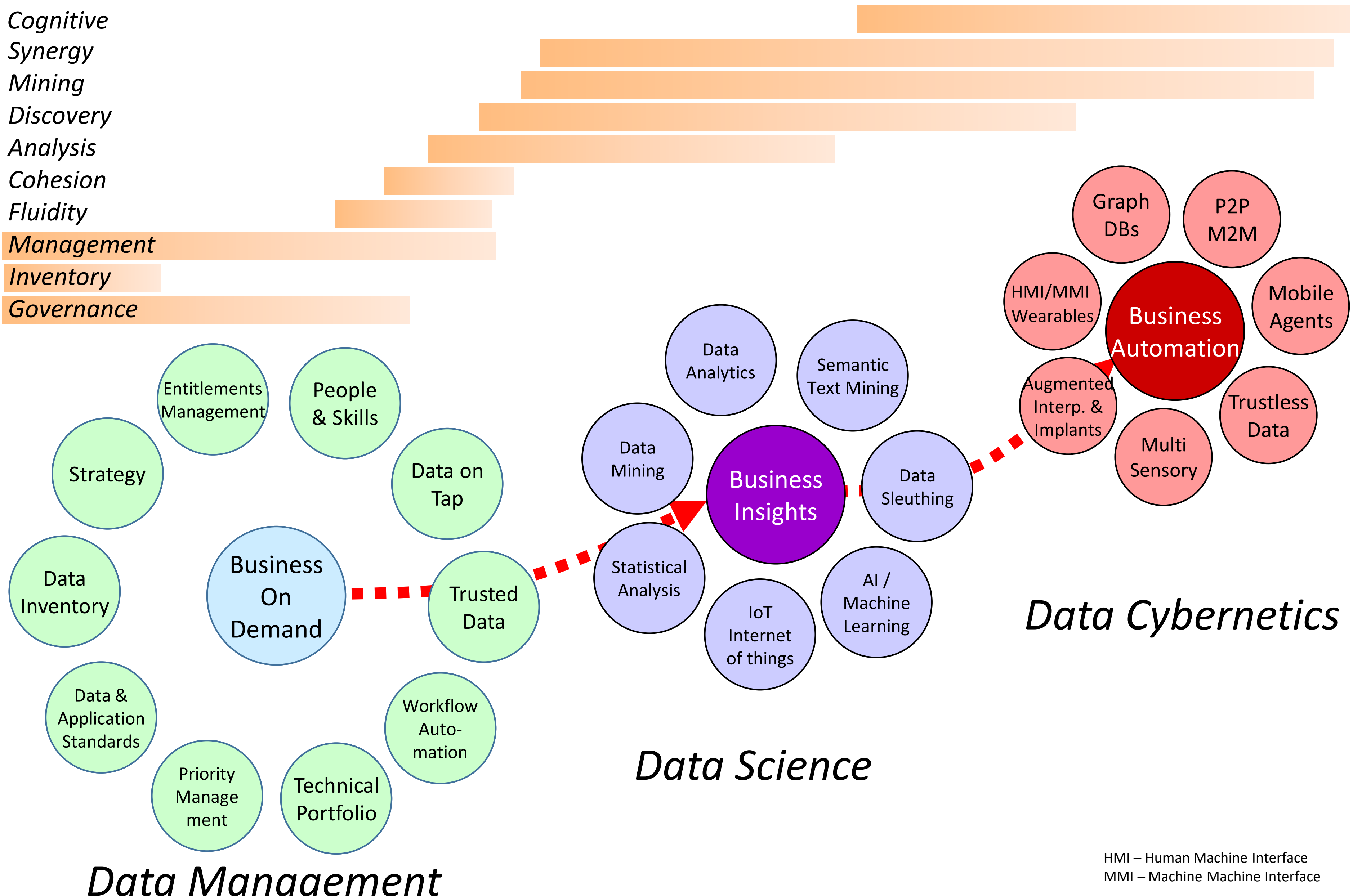
www.iamg.org

The mission of the IAMG is to promote, worldwide, the advancement of mathematics, statistics and informatics in the Geosciences.





The Future Data Driven EP Organization - Components



Thank You