PETRONAS

Data Science –
A Case Study in Biostratigraphy and Paleo-environmental Interpretation

Philip Lesslar,
Technical Assurance, Technical Data

Digital Energy Conference
5th October 2015
Impiana Hotel, Kuala Lumpur

Objectives

- To illustrate the inter-disciplinary nature of data science.
- To clarify what data science is, using an old case study.
- To show that the idea of data science is not new, just the way that it is now packaged.

Why is Data Science Important to Data Management?

- We deal with a large range of data types.
- Large volumes and increasing significantly.
- For the same data types, resolutions are increasing
- Data is prevalent on servers, PCs, shared areas, etc.

- Commercial hydrocarbons are getting harder to find
- The business needs:
  - Timely access to quality data
  - Better insight into data relationships
  - Faster analysis of data, earlier in the value chain
  - Idea connections that trigger new ideas and concepts

Data Science – Some Interesting Facts

- Data science is an emerging field in the industry
- It is not yet well defined as an academic subject
- At Columbia University, it was first taught as a class only in the fall of 2012
- The interest in formalizing data science is a result of big data

Source: Doing Data Science. Cathy O'Neil & Rachel Schutt

Data Size Table

Value in bytes	Metric	
1000	KB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	zettabyte
1000^8	YB	yottabyte

Big Data - Context

The New York Stock Exchange generates about one terabyte of new trade data per day

Facebook manages a 300 petabyte data warehouse, 600 terabytes per day – 2.5 billion pieces of content.

Ancestry.com, the genealogy site, stores around 2.5 petabytes of data

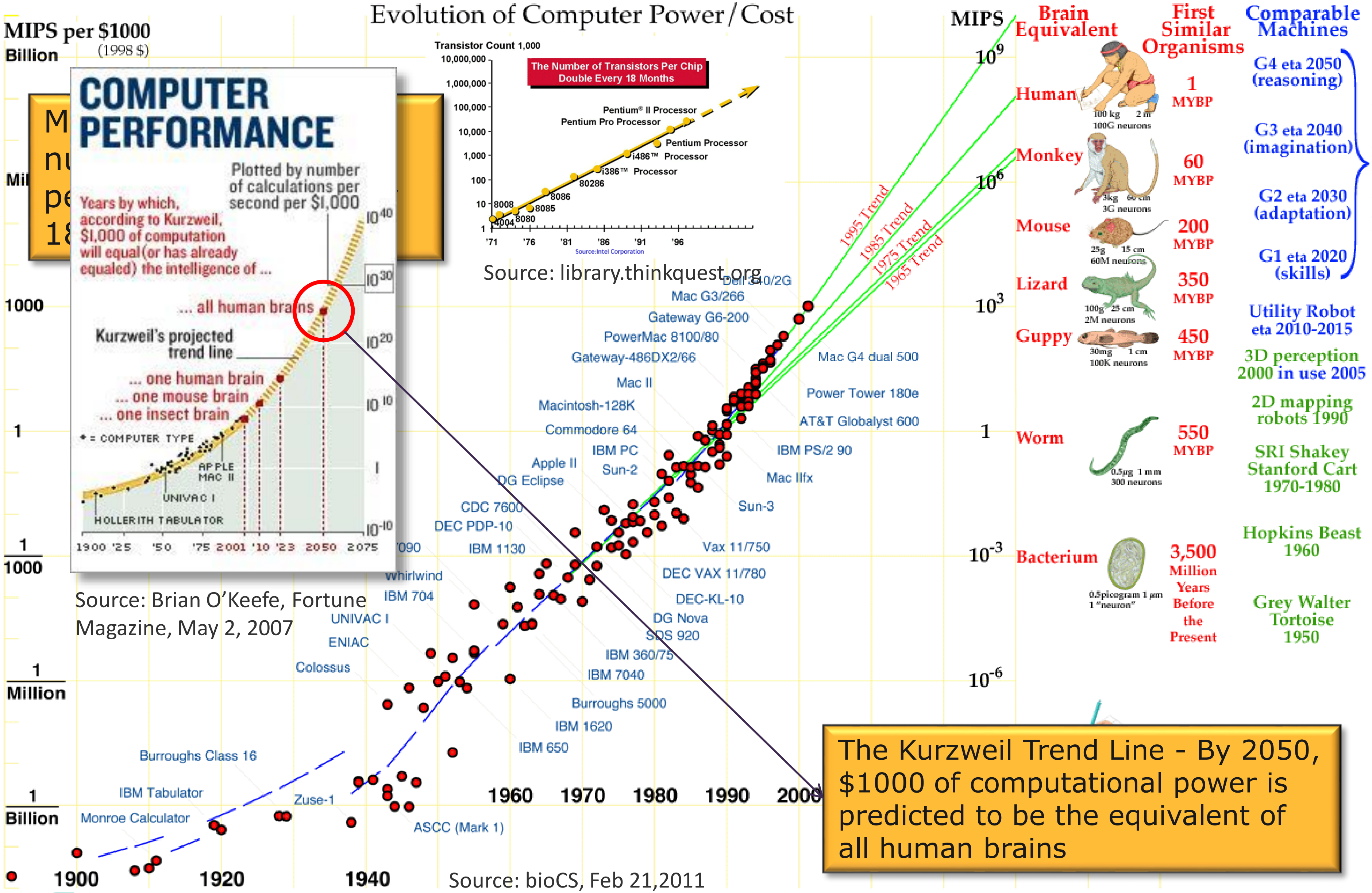
The Internet Archive stores around 2 petabytes of data, growing at the rate of 20 terabytes per month

The Large Hadron Collider near Geneva, Switzerland, produces about 30 petabytes of data per year.

-100 petabytes over the last 20 years, 75 in the last 3
-100 PB ~ 700 years of full HD movies

Source: Hadoop-The Definitive Guide. Tom White

Technology frontiers – Moore's Law & the Kurzweil Trend

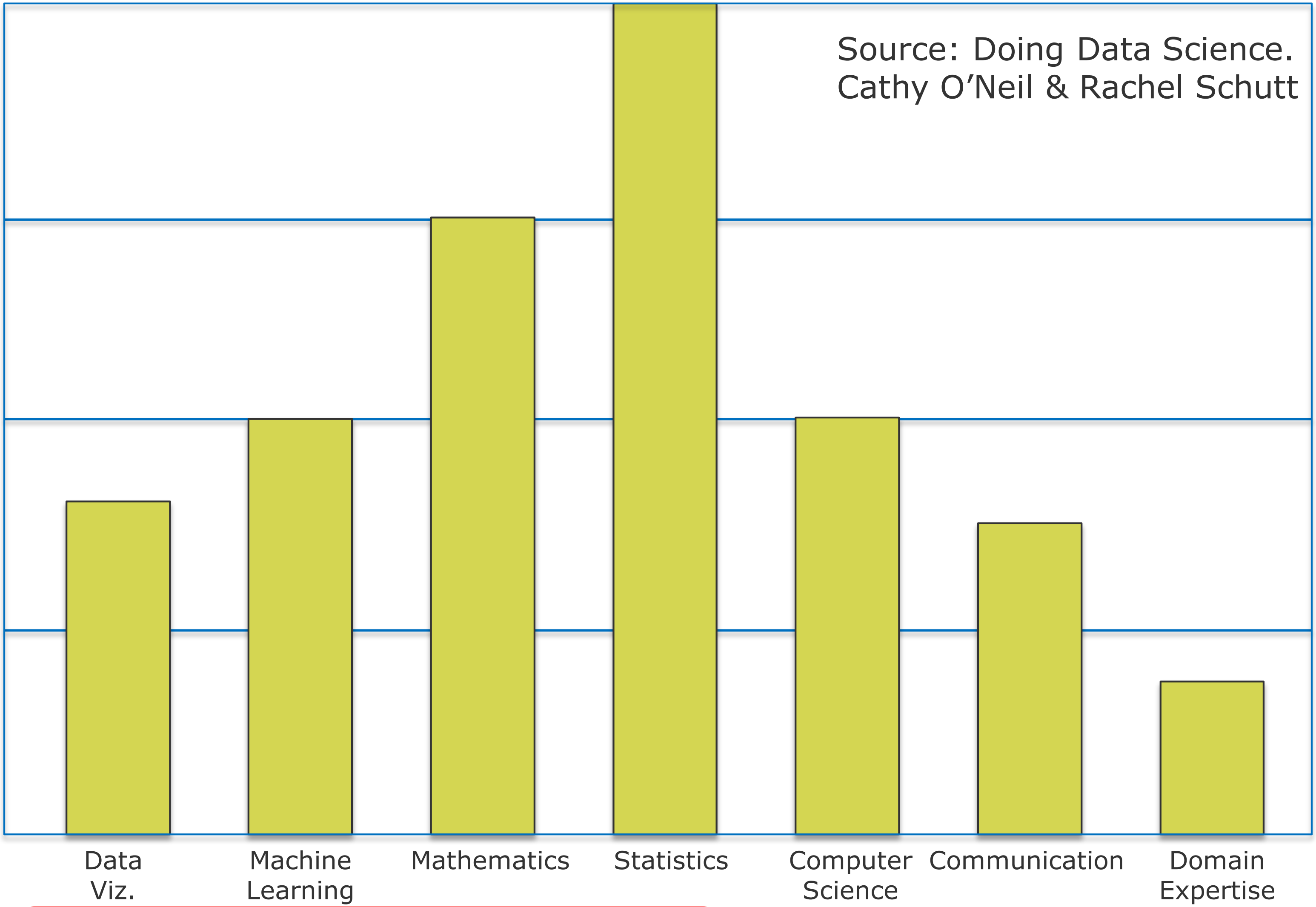


Source: Brian O'Keefe, Fortune Magazine, May 2, 2007

Source: bioCS, Feb 21,2011

The Kurzweil Trend Line - By 2050, $1000 of computational power is predicted to be the equivalent of all human brains

Data Science - Profile



Further elaboration

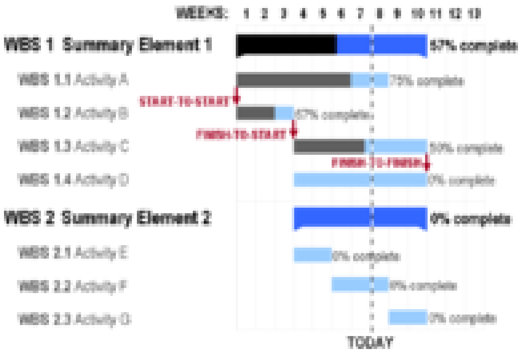
Data Science Topic 1 – Data Visualization

The main goal of data visualization is to communicate information clearly and effectively through graphical means

Typical topics in Data Visualization:
- Exploratory Data Analysis
- Information design
- Descriptive statistics
- Inferential statistics
- Statistical graphics
- Plot graphics
- Data analysis
- Infographics

Examples of application:
- Is there a correlation between carbohydrates and fat?
- Age distribution of shoppers
- Trends in production performance
- Porosities versus depth plots
- Sand distribution in a basin

Data Science Topic 1 – Data Visualization (continued)

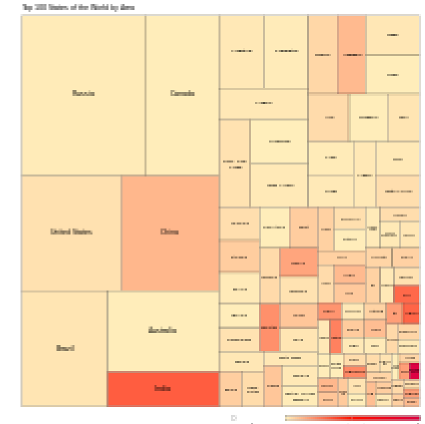


Gantt Chart

Visual dimensions
- Color
- Time (flow)

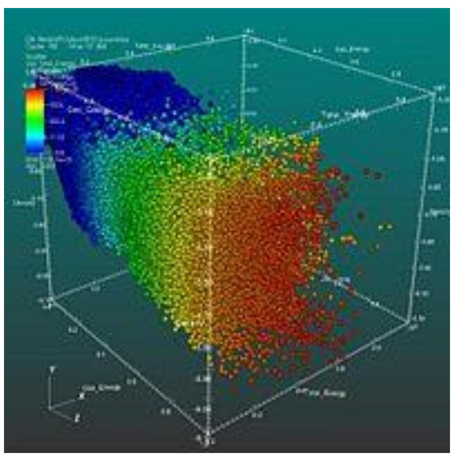
Network

Visual dimensions
- Nodes size
- Nodes color
- Ties thickness
- Ties color
- Spatialization



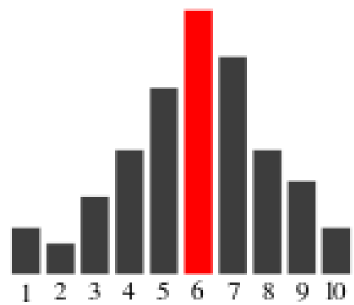
Tree Map

Visual dimensions
- Size
- Color



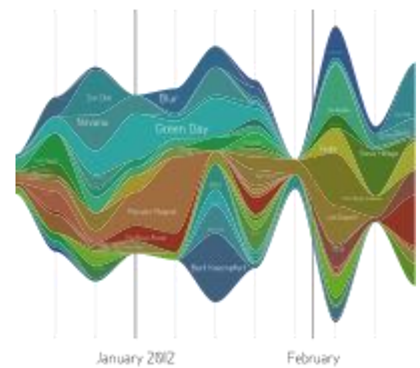
Scatter plot

Visual dimensions
- Position x
- Position y
- Position z
- Color



Bar Chart

Visual dimensions
- Height
- Color



Streamgraph

Visual dimensions
- Width
- Color
- Time (flow)

Data Science Topic 2 – Machine Learning

Machine learning is the study devoted to the development of machines that improve performance with experience.

Typical topics in Machine Learning:
- Classification algorithms
- Splitting dataset - Decision trees
- Probabilistic classification, Bayes
- Regression analysis
- Forecasting & prediction
- Supervised and unsupervised learning
- Principal components analysis
- Matrix algebra
- Big data toolkits (Hadoop, MapReduce)

Examples of application:
- Making sense of diverse data
- Relating apparently unrelated data
- Quantifying concepts such as “maximize profits”, “minimize risk”, “find the best marketing strategy”
- Building autonomous robots

Data Science Topic 3 - Statistics

Statistics is the science of making decisions in the face of uncertainty. It is a branch of applied mathematics.

Typical topics in Statistics:

- Frequency distributions
- Measures of location (mean, median, mode)
- Measures of variation (standard deviation),
- Probability and probability distributions
- Expectations
- Statistical inference
- Analysis of variance
- Nonparametric methods
- Regression
- Correlation

Examples of application:

- Drug testing
- Deciding on which well to drill
- Comparison of the efficiency of 2 production processes
- Election predictions
- Casinos
- Everyday decisions such as whether to bring an umbrella or not
- Which route to take to work

Case Study – Clustering of Foraminiferal groups

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

Examples of foraminifera

Source: Google image search for "foraminifera"

Goals of the Study

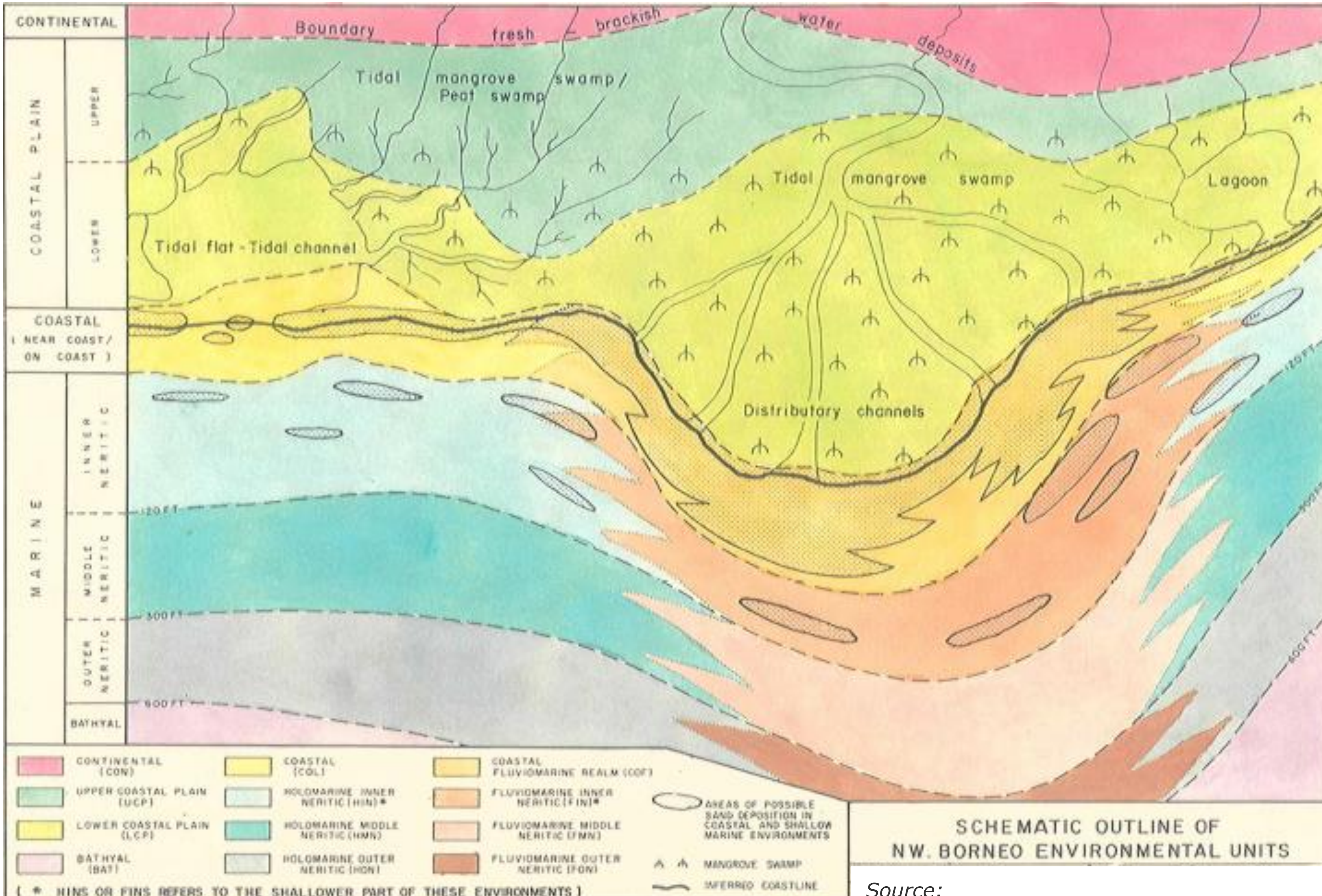
- Develop a quantitative reference matrix of foraminifera occurrences for paleo-environmental classification

- Develop a probabilistic, computer-assisted interpretation system that would remove the inconsistency associated with human interpretation

Context and Problem Statement

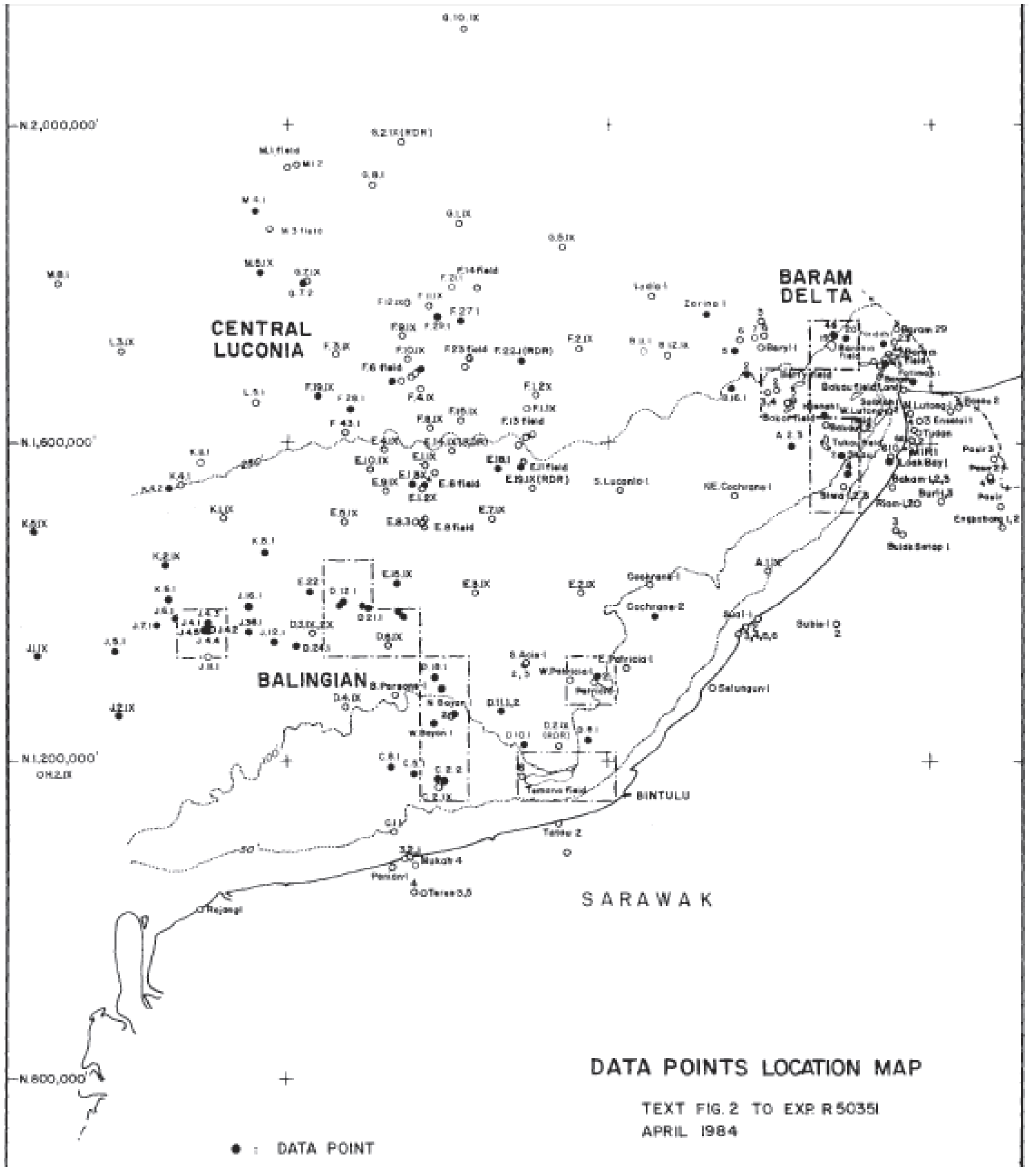
- Benthonic foraminifera are bottom-living forms that are sensitive to environmental conditions. They are therefore good inferential indicators of paleo-environments
- Environmental interpretations based on these forams were done by different investigators, subjective and therefore not always consistent and comparable

Environments of Deposition – The Scheme



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

Location Map – Data Points Used



Data dimensions:

~250 wells
~100 samples per well
20-250 species per sample averaging 120.

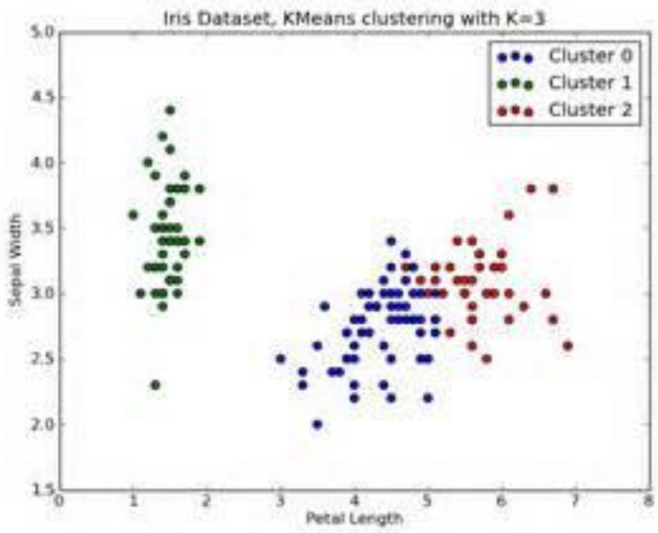
Total species occurrences in play:
~3 million

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

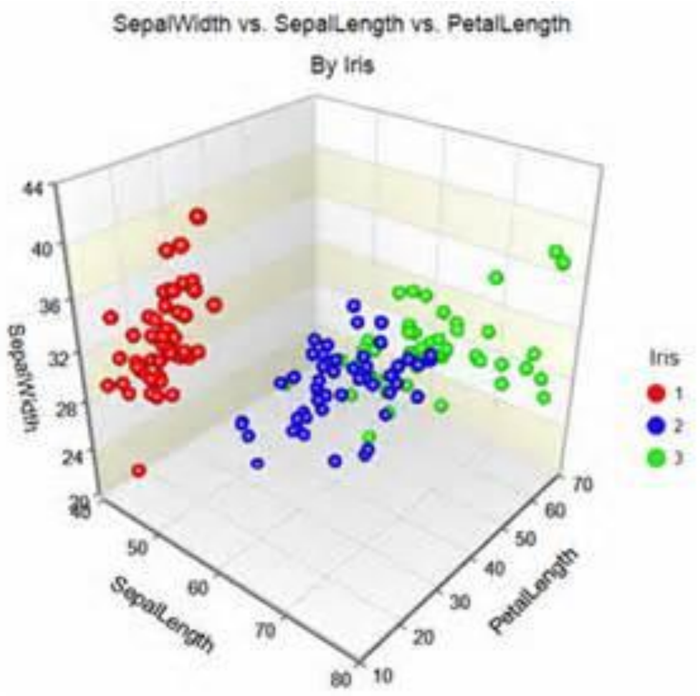
Cluster Analysis – Separating Variables in n-Dimensions

Visualization

2 dimensions



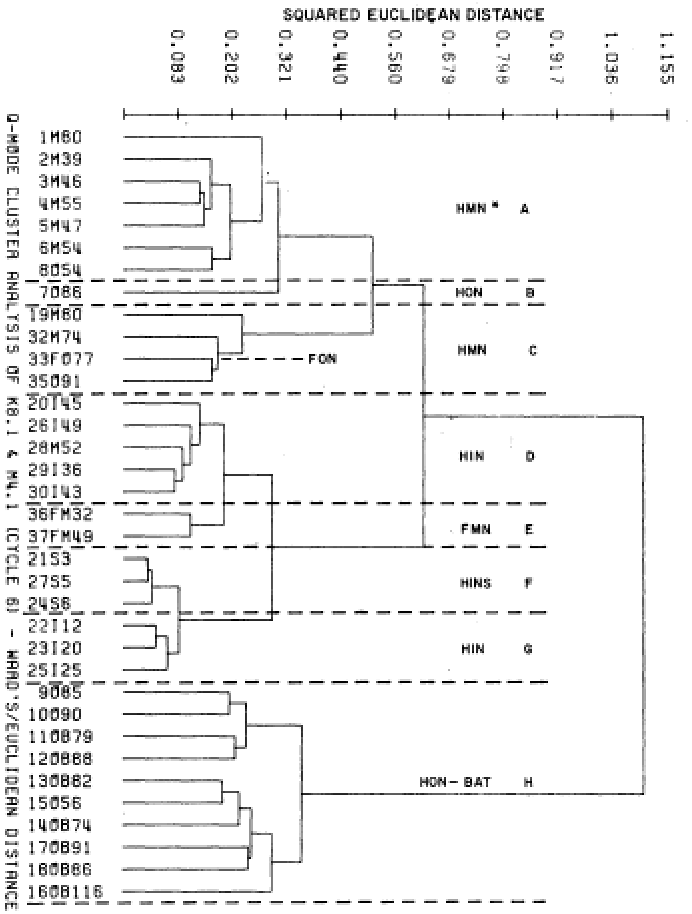
3 dimensions



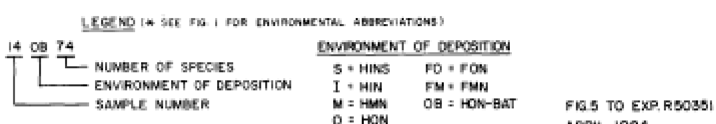
4, 5, ……, n dimensions?

Through the use of dendrograms

Cluster Analysis 1/2



Q – MODE CLUSTER ANALYSIS OF WELLS K8-I & M4-I (CYCLE VI) USING WARD'S METHOD



Dendrogram of samples from 2 wells 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.

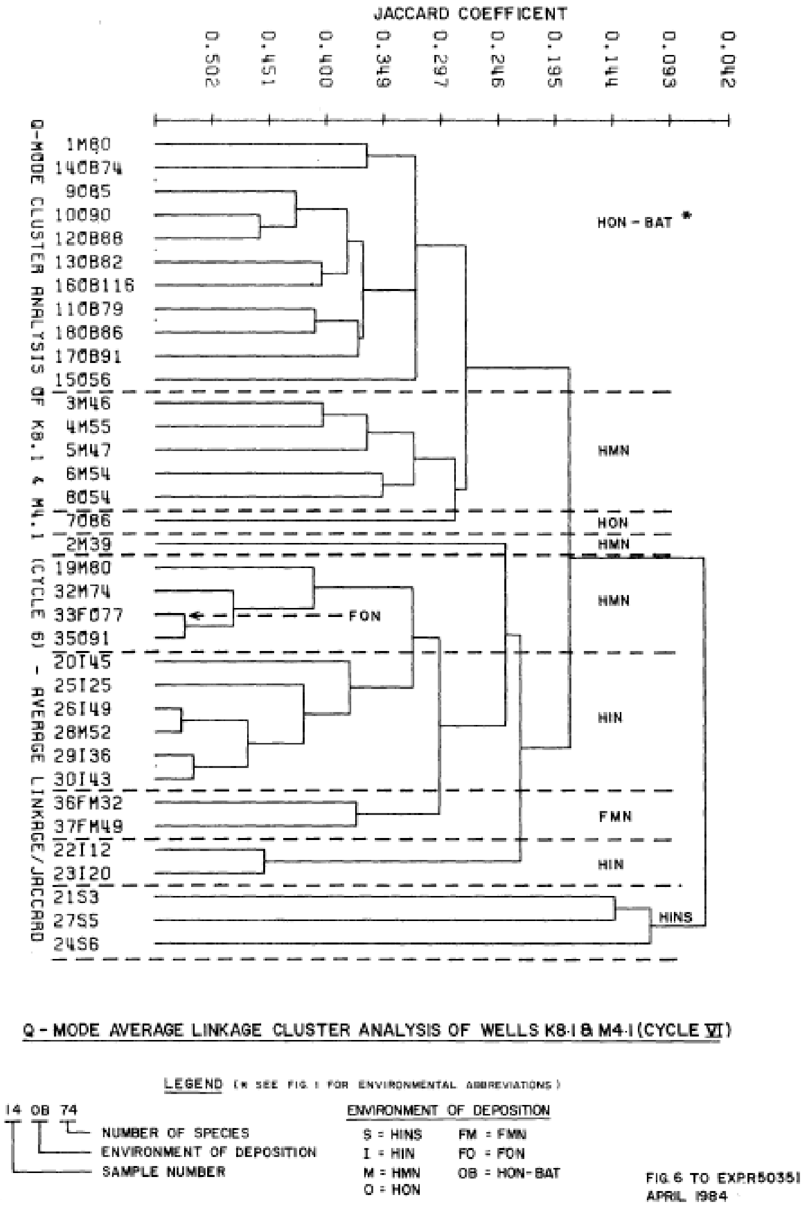
Based on foraminiferal presence/absence data.
The clustering program used was CLUSTAN.

For clustering purposes (Q-mode specifically), each sample can be thought of as a point in
n-dimensional space, where each species represents one dimension. The data of a set of samples can be put in the form of a p x n matrix, where

p = number of samples and
n = total number of species.

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

Cluster Analysis 2/2



Dendrogram of samples from 2 wells using the Average Linkage clustering method and the Jaccard coefficient

This enables the calculation of various coefficients to be done which provide indications of the strength of the relationships between the samples, one of which arises from the concept of distance (Sneath & Sokal, 1973).

The stronger the relationship between two samples points in n-dimensional space, the smaller the distance between them.

Distances between all combinations of p samples are calculated resulting in a p x p matrix, and cluster analysis techniques operate on such a matrix to reveal the inter-relationships between the various points.

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

Next Step – The Identification Matrix

A

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

Schematic of the identification ma

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	8	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		

Environmental zones

Foraminiferal species

Computer listing of the identification matrix

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.

Source:P. Lesslar, Geol. Soc. Malaysia Bulletin 21, Dec 1987.

The Identification Matrix 2/2

411 species

13 depositional environments

Source:P. Lesslar, Geol. Soc. Malaysia Bulletin 21, Dec 1987.

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:

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

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

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
--------------------------------------------------------------
```

Source:P. Lesslar, Geol. Soc. Malaysia Bulletin 21, December, 1987.

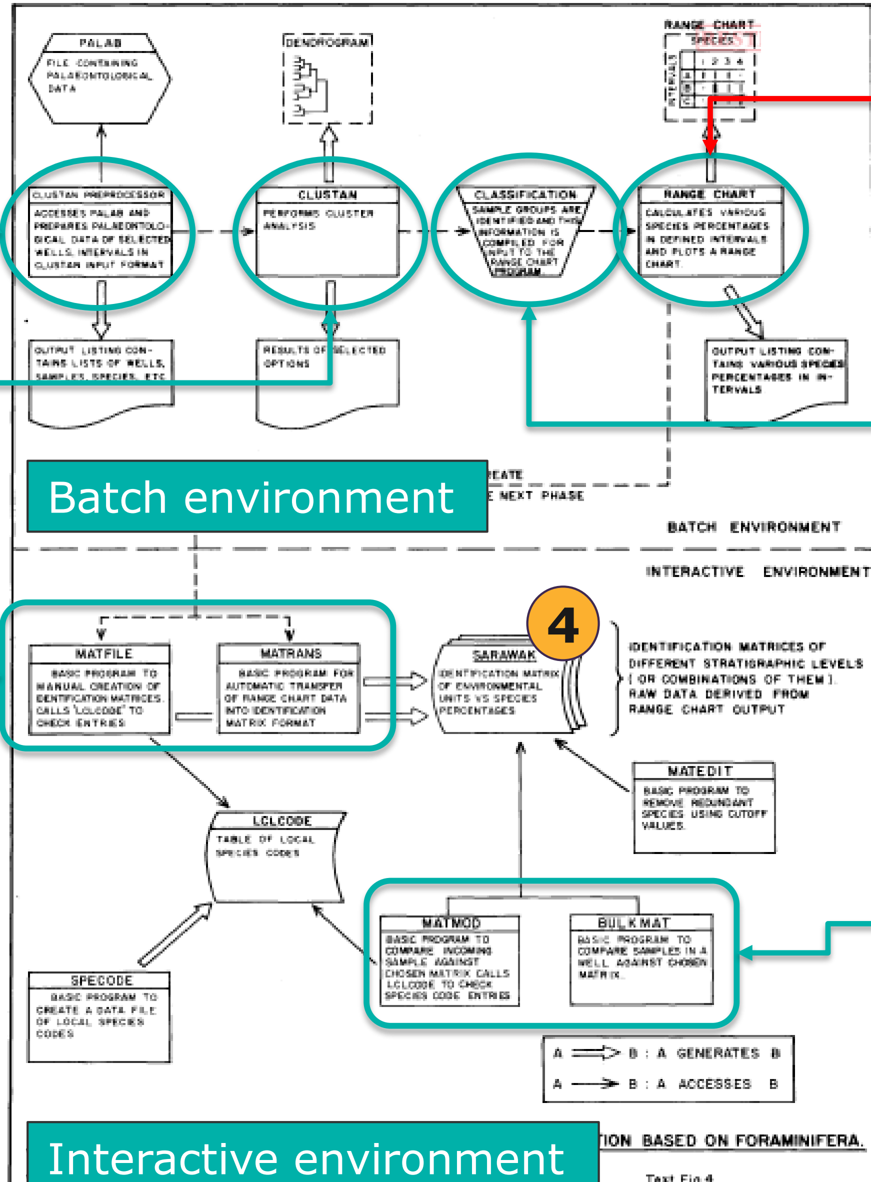
Flow diagram – System & data

1 Data read & preparation from tapes

2 Cluster analysis

3 Sample groupings as input to range chart program

4 Generates probabilistic identification matrix



Batch environment

Interactive environment

5 Paleontological data preparation programs – read data from tapes

4

6 Probabilistic interpretation programs

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

Project Conclusions

- Objectives were successfully proven
  - A quantitative reference matrix for paleo-environmental classification was successfully constructed.
  - A probabilistic, computer-assisted interpretation system that would remove the inconsistency associated with human interpretation was developed.

- Used by the paleontological section to
  - Improve consistency in interpretations
  - Improve interpreter capability

Conclusions

- The case study shared is intended to illustrate the synergic potential of subject areas that is part of data science today.
- Data science embodies subjects that have been around for a long time.
- Advances in computing technology in combination with these subjects open new doors in data analysis and synthesis
- The rate of parallel developments in this new emerging field makes it one of the most exciting aspects of data management

Thank you