Opportunity Landscape for Data Scientists in E&P

Philip Lesslar
Data Solutions Consultant

Halliburton STEPS Lecture Series 2018
18th April 2018
Corus Hotel, Kuala Lumpur

Objectives

- Create an awareness of the vast potential for data science applications in E&P

- Show a few examples of what has been done

- Provide some pointers on moving forward

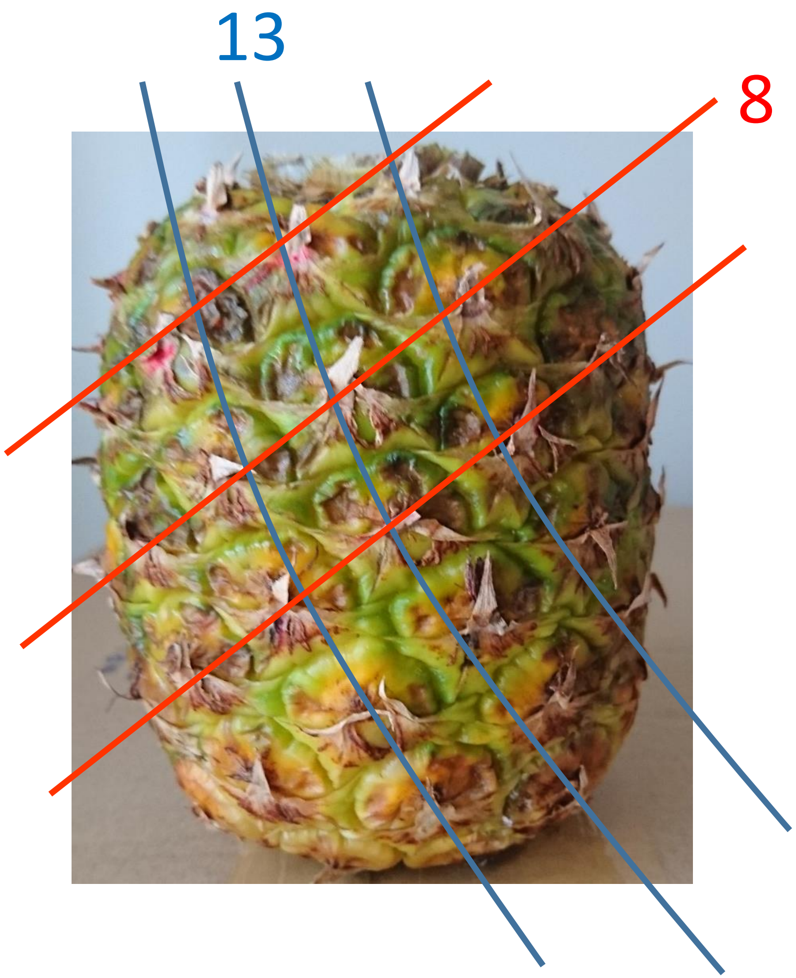
Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

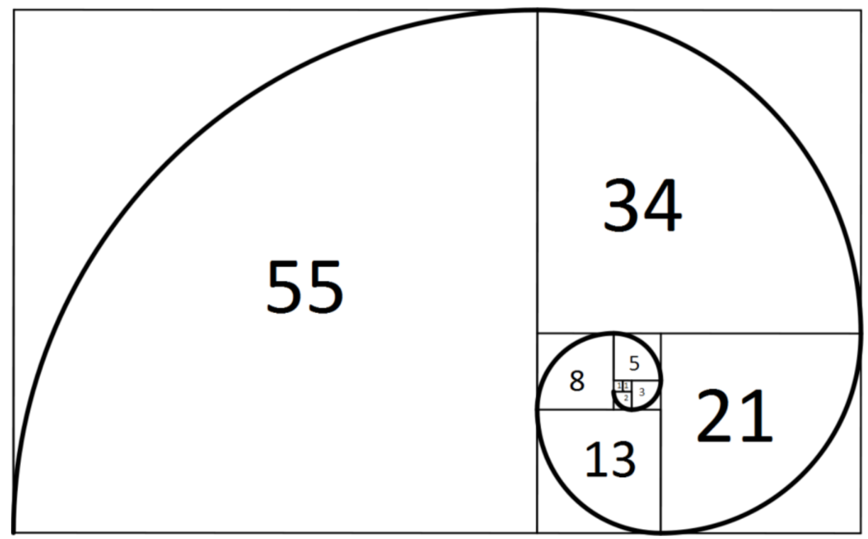
Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

Introduction – 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

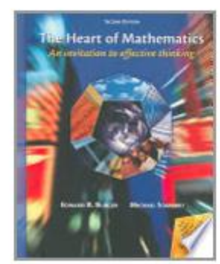
The rule for generating successive numbers in the sequence is to add up the previous two terms. So the next number on the list would be $55 + 89 = 144$.

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

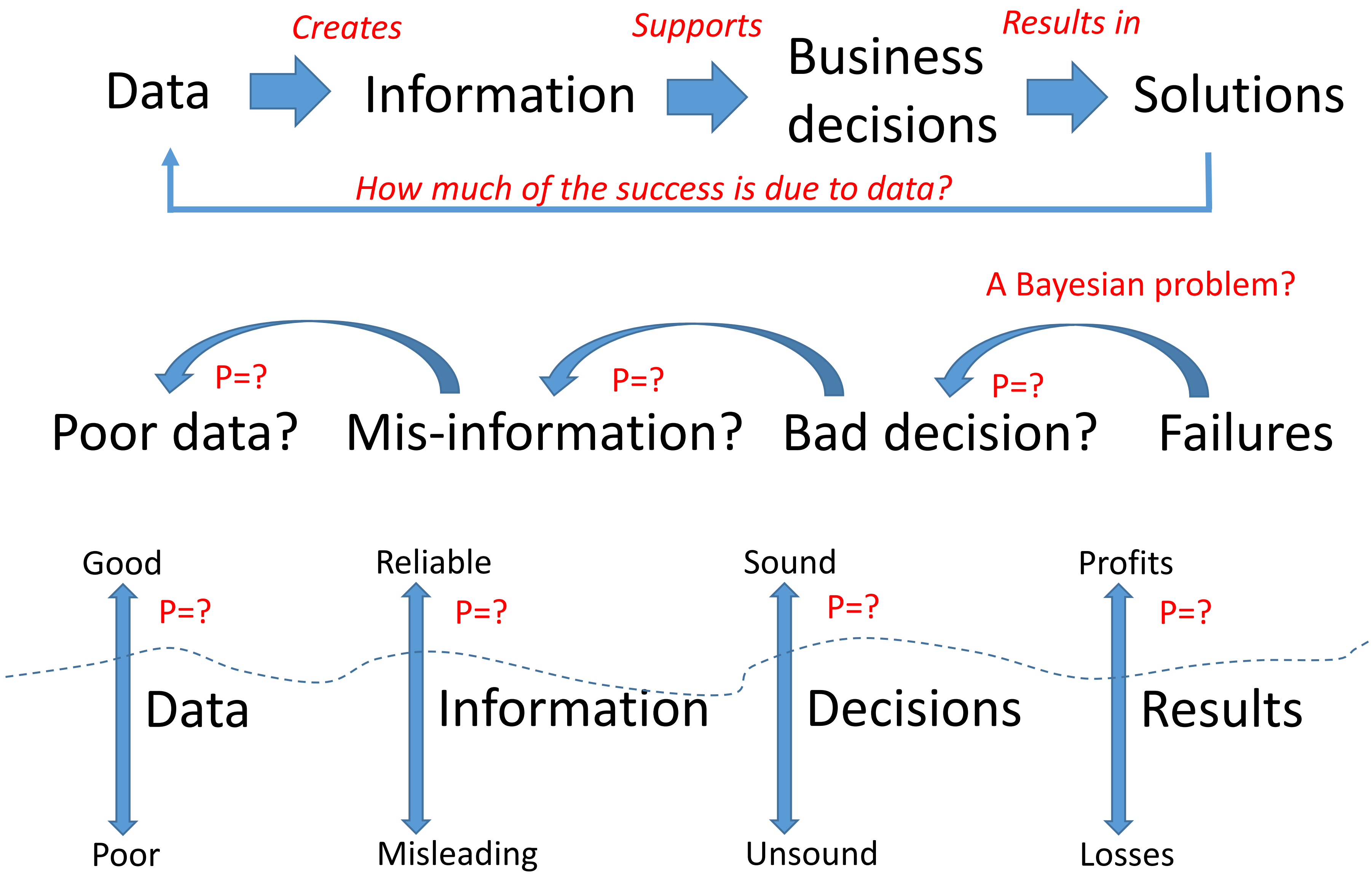
Introduction – Curiosity

A typical cumulus cloud weighs 500,000 kgs or 1.1 million lbs

Peggy LeMone, researcher at National Center for Atmospheric Research
http://mentalfloss.com/article/49786/how-much-does-cloud-weigh

Moral : Ask questions – More likely than not, there are already answers

Introduction - Purpose of data



Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

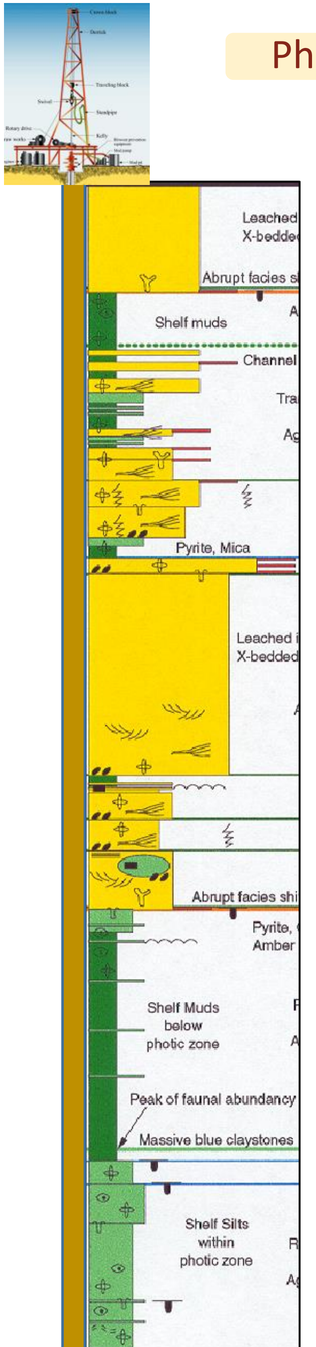
The Upstream Value Chain

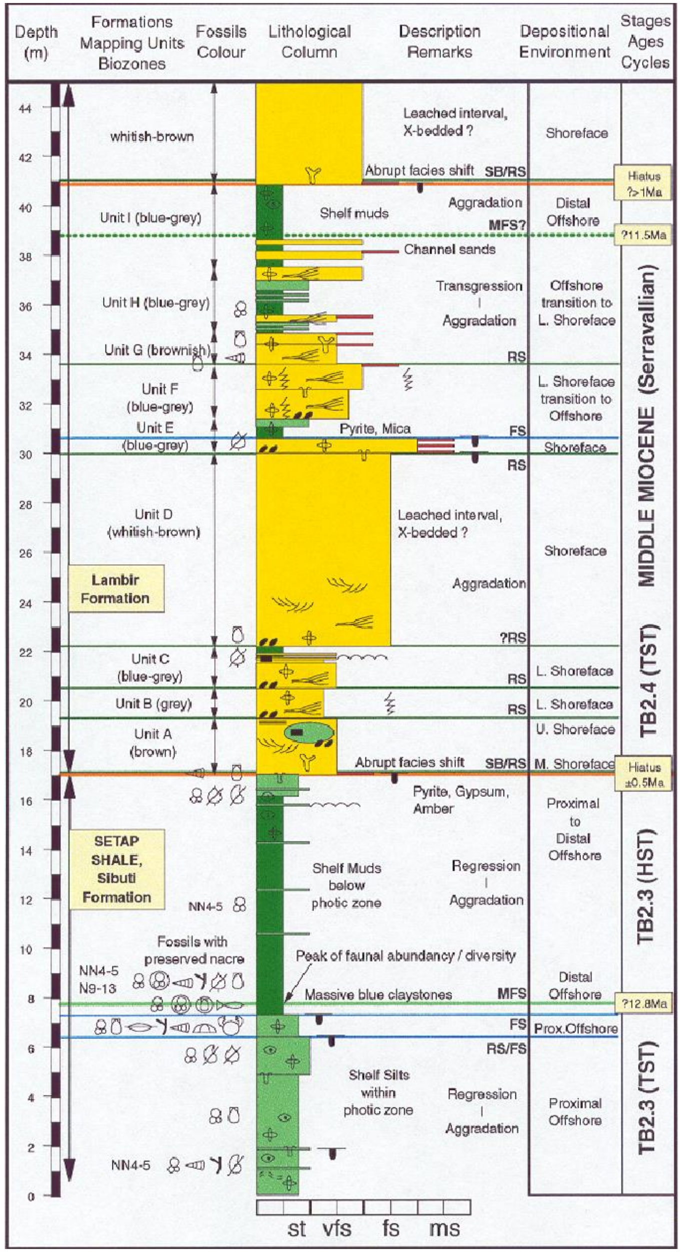
Data aspects

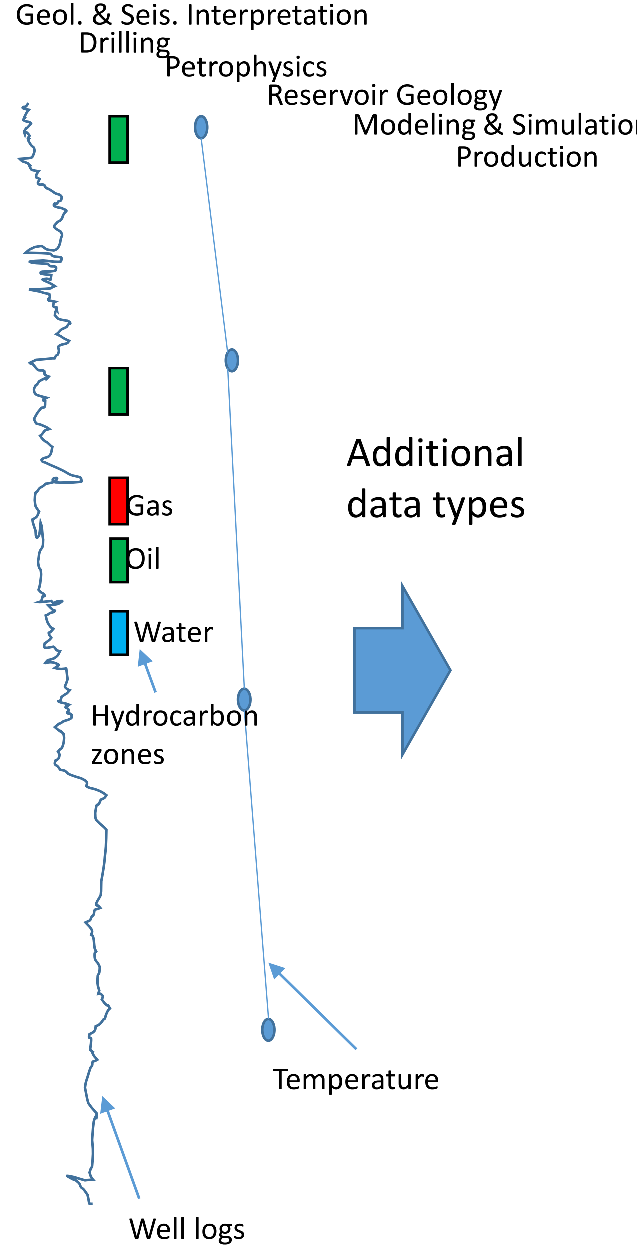
Review	Regional studies, data rooms
Acquire	Acreage, production sharing contracts, seismic (2D, 3D, OBC etc), data purchase, exchanges
Explore	Regional reviews and compilations, play & prospect identification, well locations, well data, correlation
Appraise	Additional well planning & data, detailed studies 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







Additional data types

Typical Problems encountered in E&P Data

Physical 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

Processes & Methodology

- 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

Electronic Data

- 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

- Resource constraints
- Lack of competency
- Lack of people framework
- Lack of proper accountability structure
- Indecision
- Office politics
- Integrity

On a bigger scale…

The Data Underworld

Typical characteristics

- No governance
- Unofficial transactions (data “laundering”)
- Extreme data duplication
- Data hoarding
- Orphan data collections
- Leakage
- No process for data to move to corporate stores
- Includes personal & entertainment media
- Uncontrolled increase in data amounts

Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

Data Types - Upstream

Review → Acquire → Explore → Appraise → Develop → Produce → Abandon

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 -Gravity & Magnetic -Microseismic	Geology – Zones Geology – Markers Faults (Field Extent & Major) Seismic Horizons – Regional Seismic Horizons – Local Velocity Models Structure Maps Time-Depth Curve Gridded Time / Depth Maps Sand Distribution Maps Static Models Dynamic Models Synthetic Seismogram Biostratigraphy – Zones Biostratigraphy – Markers Geology – Zones Geology – Markers	Spill Points Well Logs – Raw Well Logs – Processed & Qced Well Logs – Interpreted Well Logs – Cased Hole Vertical Seismic Profiling Core Analysis including SCAL Formation Pressure (RFT, MDT) Well Tests (DST-Drill Stem Test, FIT-Formation Intake Test) Production Data (Allocated oil/gas/water rates) Production Pressure Data (Well Tubing/Casing Head Pressure) Production Well Test Artificial Lift Fluid Property Fluid Contacts Stimulation Cases Fluid Composition Materials Balance Decline Curve Analysis Volumetrics Reserves and Resources Pressure Maintenance Cases Saturation Height Function Leak Off Test PVT (Pressure-Volume-Temperature)	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 Classification – Digital Data (>100 types in Upstream)

Primary Data			Secondary Data	
Original Format Data	**Reference Data/ Metadata**	**Master Data/ Corporate "Single Source of Truth"**	**Derived Data**	**Data Collections**
Raw Seismic Raw Logs	Units of measure - *Linear measures* - *Pressure*	Static (hard) data - *Well header* - *Deviation* - *Checkshot* - *Temperature* - *Pressure*	Processed data - *Seismic deconvolution* - *Seismic filtering* - *Seismic processing* - *Edited logs* - *Spliced logs*	Composite data - *Completion log* - *Mud log* - *Paleontological composites* - *TRAPIS*
	Abbreviations - *TD, DFE, KB etc*	Interpreted (soft) data - *Geological markers* - *Seismic horizons*	Interpreted data - *Geological markers* - *Seismic horizons*	Data hoards - *Projects en masse* - *Personal stores* - *Team folders*
	Valid Lists			Data archive - *Projects en masse*
	Range indicators			
	Comments			
Requires: - *Official data repository*	Requires: - *Standards* - *Implementation across all impacted tools and databases*	Requires: - *Clear processes, workflows and checkpoints* - *Proper & official repository* - *Management and security processes around repository and data access*	Requires: - *Standard workflows* - *Standard algorithms* - *Standard processes* - *Housekeeping procedures*	Requires: - *Standard display and formatting templates* - *Procedures*

Geology – Data Categories and Types

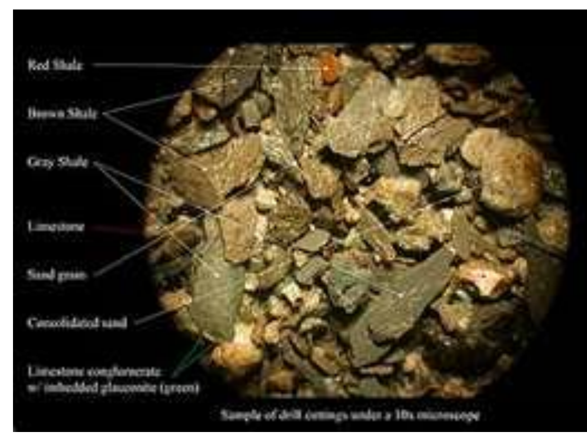
Data type	Control documents	Typical Issues	Corrective actions
-Mud logging data -Chromatography & Hydrocarbon show	None, instrument dependent	Calibration	Contact expert and check calibration
Cuttings Lithology	Litho. Legend, reference templates	Personal experience, inconsistent symbology & interpretation	Review with operations geologist
MWD Formation analysis (non-logs)	None, tools dependent	Hole integrity,	Review with petrophysicist,
Well Core Samples & Analysis	Litho. Legend, reference templates	Recovery factor, orientation, depth corrections	Review with geologist in charge
-Well Pick -Well Fault Observation	Strat scheme, naming convention	Duplication, inconsistent naming	Collate duplication & clarify with team, implement procedures
-Well Fluid Contact	Naming convention	Fluid interpretation	Check with prod geologist
-Well Computed Lithology	Litho legend, color scheme	Parameter choices	Review with petrophysicist
-Well Paleontology	Strat scheme, depositional model, coding system	Wrong codes used, use of arbitrary codes	Check codes against strat scheme, feedback to team
-Well zonation, well interval	Strat scheme, naming convention	Non-use of strat scheme, naming convention not followed	Check codes against scheme & naming convention
-Cross section interpretation -Well correlation	Strat scheme, naming convention	Incorrect ref. datums, availability of required data.	Ensure key datum data are available

1) We do not address here whether the interpretation is correct or not
2) Field geology is also excluded

Geology – drill (Ditch) Cuttings

Shale shaker

Fluoroscope for UV detection of oil
https://www.landseaskyco.com/houston-mud-logging-supplies-hmls-oil-fluoroscope.html



View under the microscope
https://commons.wikimedia.org/wiki/File:Drill_cuttings_-_Annotated_-_2004.jpg

Geology - Data Categories & Types

Data type	Control documents	Typical Issues	Corrective actions
-Mud logging data -Chromatography & Hydrocarbon show	None, instrument dependent	Calibration	Contact expert and check calibration
Cuttings Lithology	Litho. Legend, reference templates	Personal experience, inconsistent symbology & interpretation	Review with operations geologist
MWD Formation analysis (non-logs)	None, tools dependent	Hole integrity,	Review with petrophysicist,
Well Core Samples & Analysis	Litho. Legend, reference templates	Recovery factor, orientation, depth corrections	Review with geologist in charge
-Well Pick -Well Fault Observation	Strat scheme, naming convention	Duplication, inconsistent naming	Collate duplication & clarify with team, implement procedures
-Well Fluid Contact	Naming convention	Fluid interpretation	Check with prod geologist
-Well Computed Lithology	Litho legend, color scheme	Parameter choices	Review with petrophysicist
-Well Paleontology	Strat scheme, depositional model, coding system	Wrong codes used, use of arbitrary codes	Check codes against strat scheme, feedback to team
-Well zonation, well interval	Strat scheme, naming convention	Non-use of strat scheme, naming convention not followed	Check codes against scheme & naming convention
-Cross section interpretation -Well correlation	Strat scheme, naming convention	Incorrect ref. datums, availability of required data.	Ensure key datum data are available

1) We do not address here whether the interpretation is correct or not
2) Field geology is also excluded

Geology – Cores & sidewall samples

Sidewall sample gun array

sws

Recovered sidewall samples

Storage bottle

Cores

Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

Data Science Methods

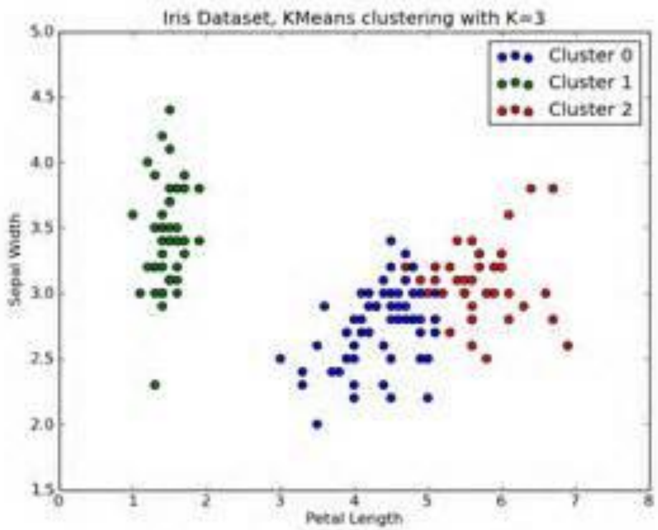
Sequence Analysis	Spatial Analysis	Statistics	Multivariate Data Analysis	Artificial Intelligence	Probabilistic Methods
Markov Chains Runs Test Least Squares & Regression Analysis Splines (polynomial smoothing) Segmented Sequences & Zonation Analysis Auto- and Cross-Correlation SemiVariogram Spectral Analysis	Pattern Analysis (Random, Cluster, Nearest Neighbour) Analysis of directional data Spherical Distributions Fractal Analysis Shape Analysis Contouring, Trend Surfaces & Kriging	Summary Statistics Hypothesis Testing t-Distribution F-Distribution Normal Distribution Chi Square Distribution Chi Square Goodness of fit Regression Analysis of Variance (ANOVA) Non-Parametric Tests - (Mann-Whitney, Kolmogorov-Smirnov, Kruskal-Wallis)	Multiple Regression Discriminant Functions Cluster Analysis Eigenvalues & Eigenvectors Factor Analysis (R & Q Mode) Principal Components Correspondence Analysis MultiDimensional Scaling Canonical Correlations	Classification Natural Language Processing Machine Learning / Deep Learning Text Mining Graph Relationships	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

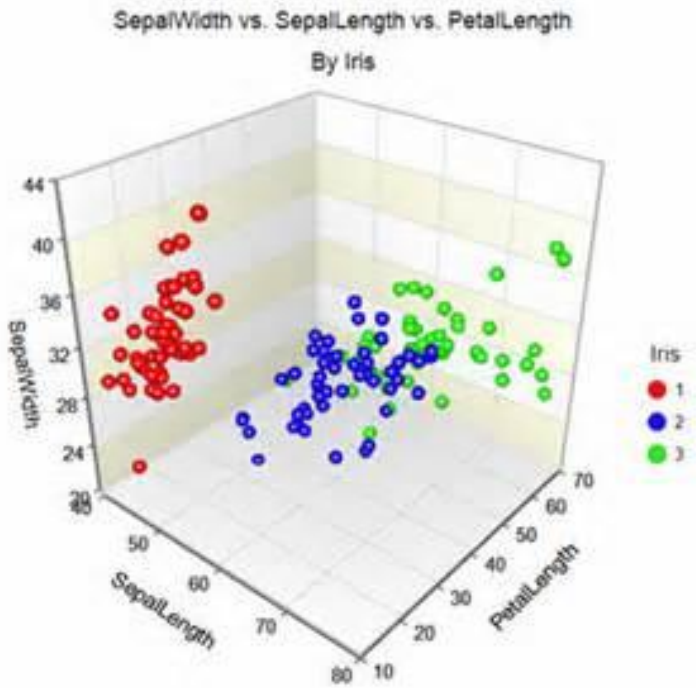
Cluster Analysis – Separating variables in n-dimensions

Visualization

2 dimensions



3 dimensions



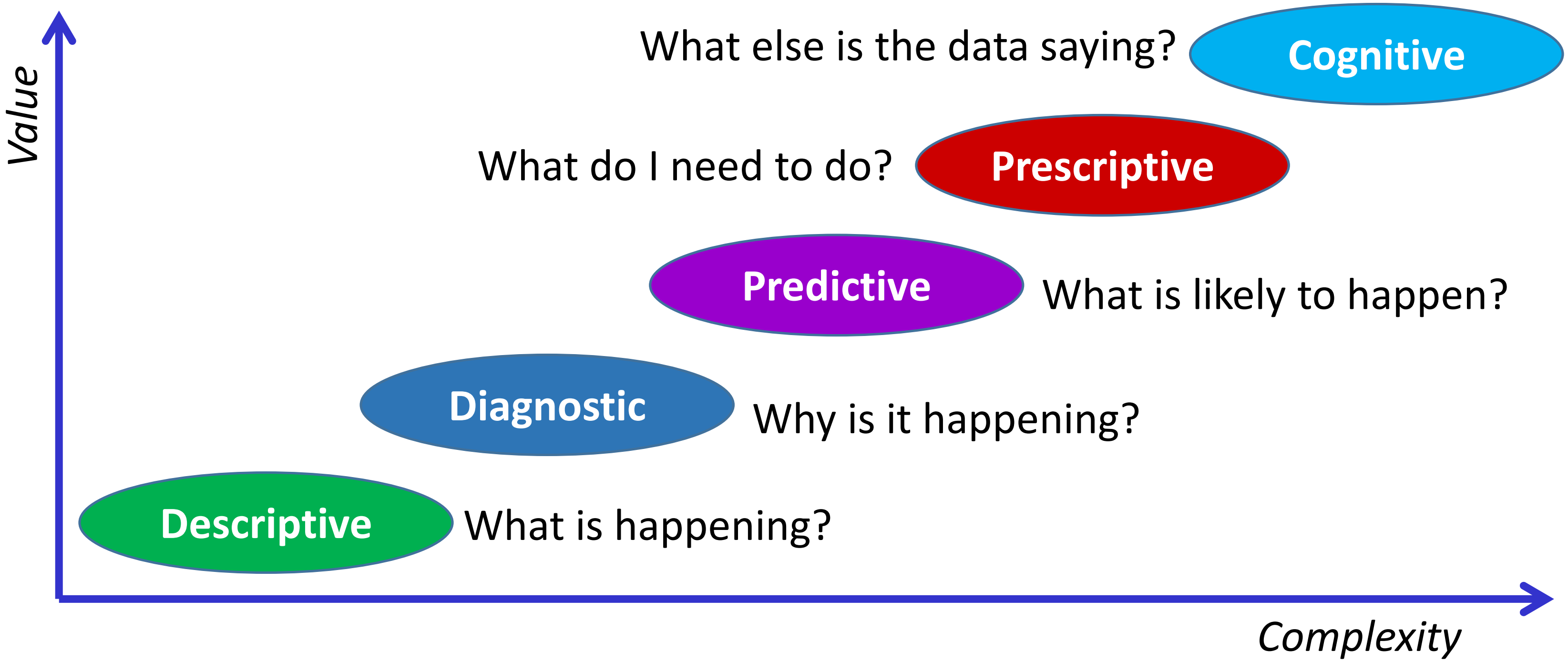
4, 5, ……., n dimensions?

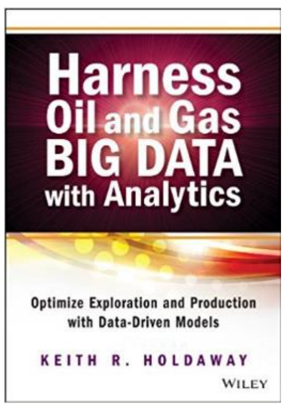
Through the use of dendrograms

Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

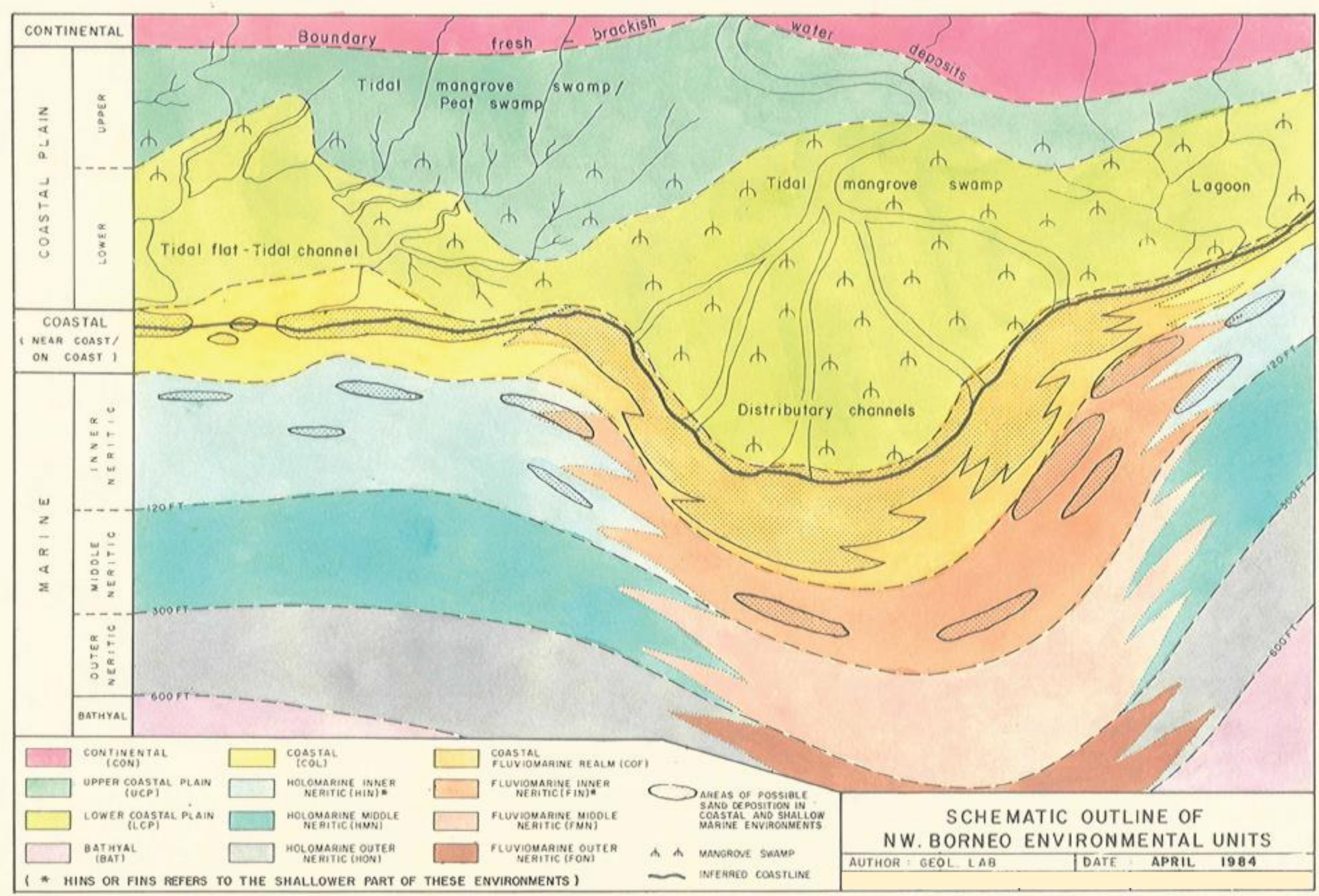
Example 1: Analytics





- Seismic Attribute Analysis
- Reservoir Characterization & Management
- Drilling & Completion Optimization
- Production Forecasting & Optimization
- Examples of Data Analysis

Example 2: Interp. of Depositional Environments - Foraminifera

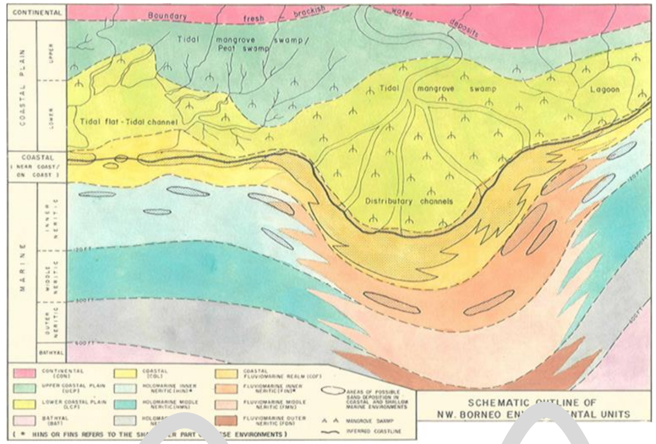


Paleontology - Foraminifera

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

Examples of foraminifera

Example 2: Interp. of Depositional Environments - Foraminifera



Input data

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

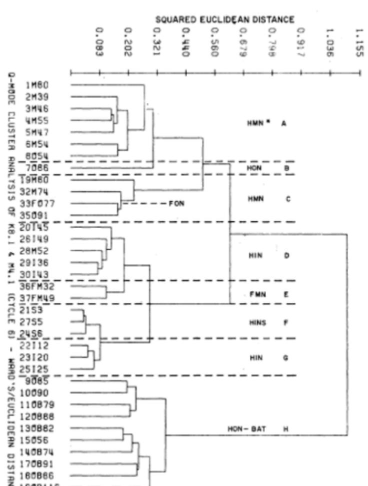
Results

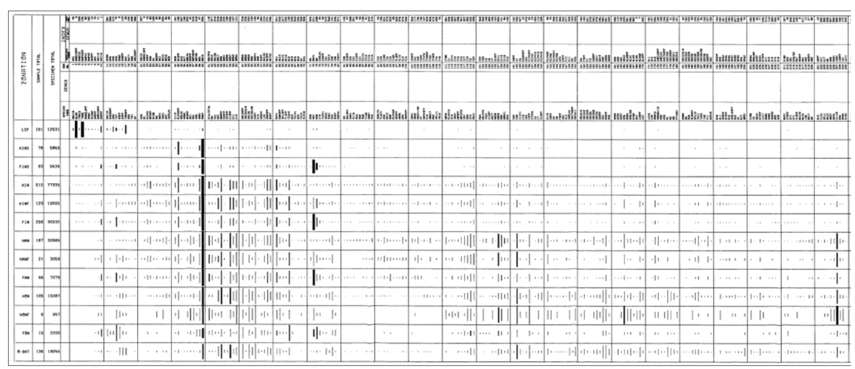
```
SAMPLE =   2638 '          BEST IDENTIFICATION IS .. LCP
===============            CURRENT INTERPRETATION ..
NO.SPECIES = 5          NO.POSITIVE MATCHES WITH IDENT.MATRIX= 5
NO. SPECIMENS =         28          P/B RATIO =   0.00
DIVERSITY INDICES.   YULE-SIMPSON =  3.60, FISHER ALPHA =  1.02

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

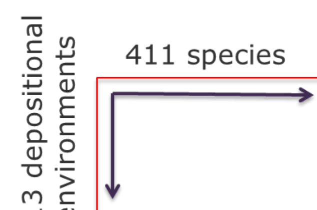
Clustering

Build Probability Matrix of valid clusters

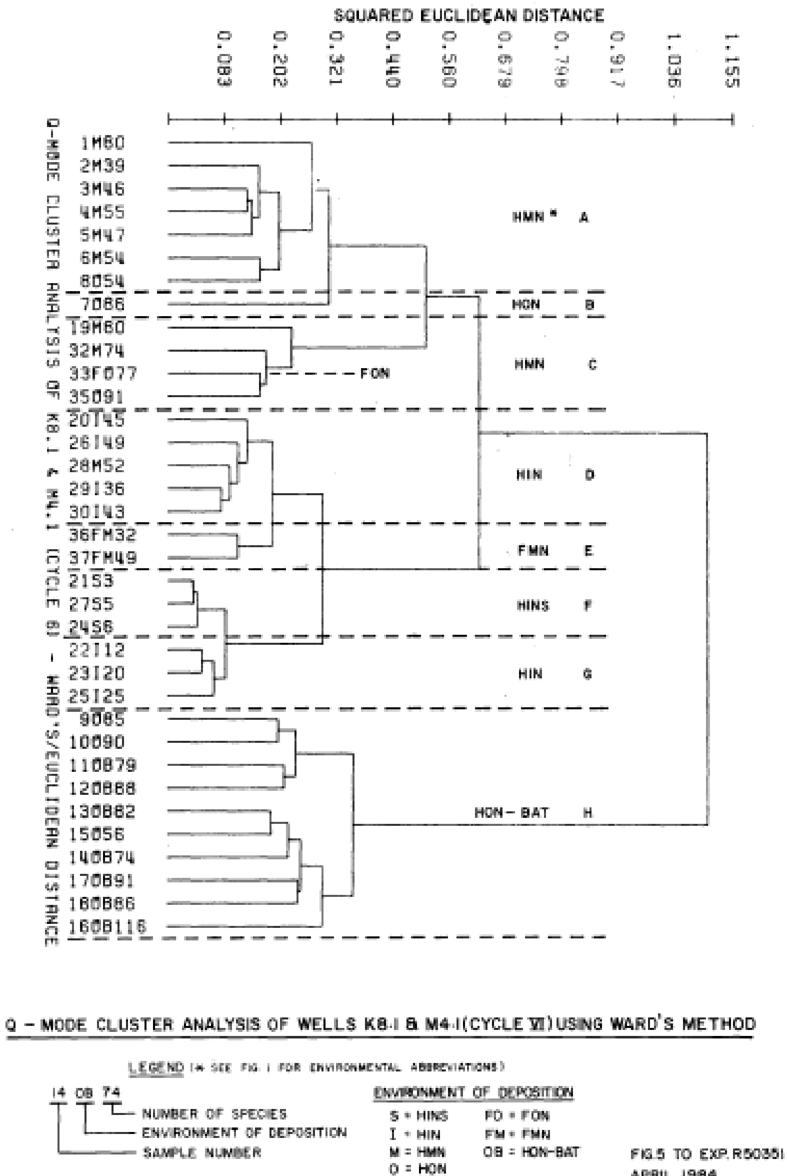
Identification Program - Bayesian Inference (Likelihood Ratio)







Cluster Analysis Example – Environments of Deposition



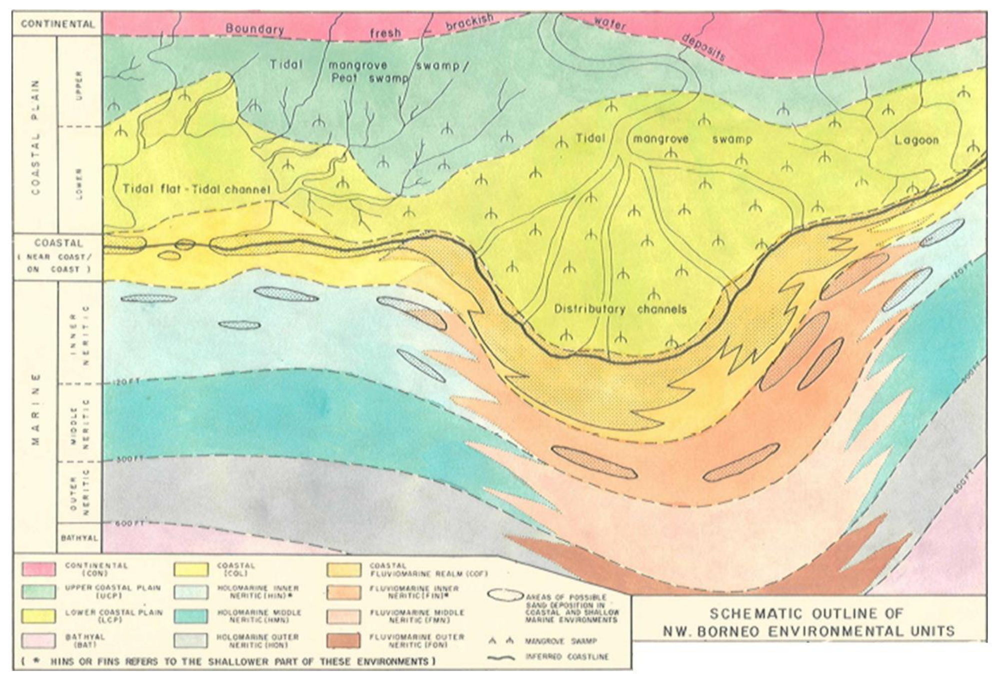
Source:
Computer-assisted interpretation of depositional palaeoenvironments based on foraminifera. Philip Lesslar, Geol. Soc. Malaysia Bulletin 21, December, 1987.

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



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

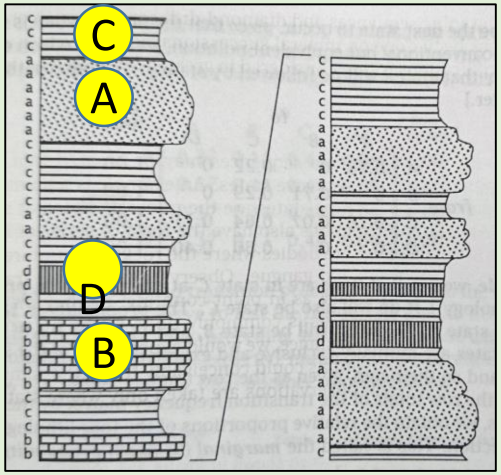
```
PROGRAM FOR IDENTIFICATION OF WELL SAMPLES USING
PRESENCE-ABSENCE DATA AGAINST AN IDENTIFICATION MATRIX
OF PERCENT POSITIVE CHARACTERS OF THE TAXA
==========================================================
BY : P.LESSLAR, XGS/1. MODIFIED FROM SNEATH,1979
DATE : 84/10/23     TIME : 07:43:18

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

Example 3: 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).

OBSERVATIONS

Transition Frequency Matrix

from \ to	A	B	C	D	Row Totals
A	18	0	5	0	23
B	0	5	2	0	7
C	5	2	18	3	28
D	0	0	3	2	5
Column Totals	23	7	28	5	**63**

NORMALIZATION

Probability Matrix

from \ to	A	B	C	D	Row Totals
A	0.78	0	0.22	0	1.00
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

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

A	0.37
B	0.11
C	0.44
D	0.08

Shows the relative proportions of the 4 lithologies in the sequence

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

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)

H_o = Independent states

Expected Transition Probabilities

from \ to	A	B	C	D	Totals	
A	0.37	0.11	0.44	0.08	1.00	X 23 =
B	0.37	0.11	0.44	0.08	1.00	x 7 =
C	0.37	0.11	0.44	0.08	1.00	x 28 =
D	0.37	0.11	0.44	0.08	1.00	x 5 =

EXPECTATIONS

Expected Frequencies

8.5	2.5	10.1	1.8
2.6	0.8	3.1	0.6
10.4	3.1	12.3	2.2
1.9	0.6	2.2	0.4

CHI-SQUARE

Test for Non-randomness

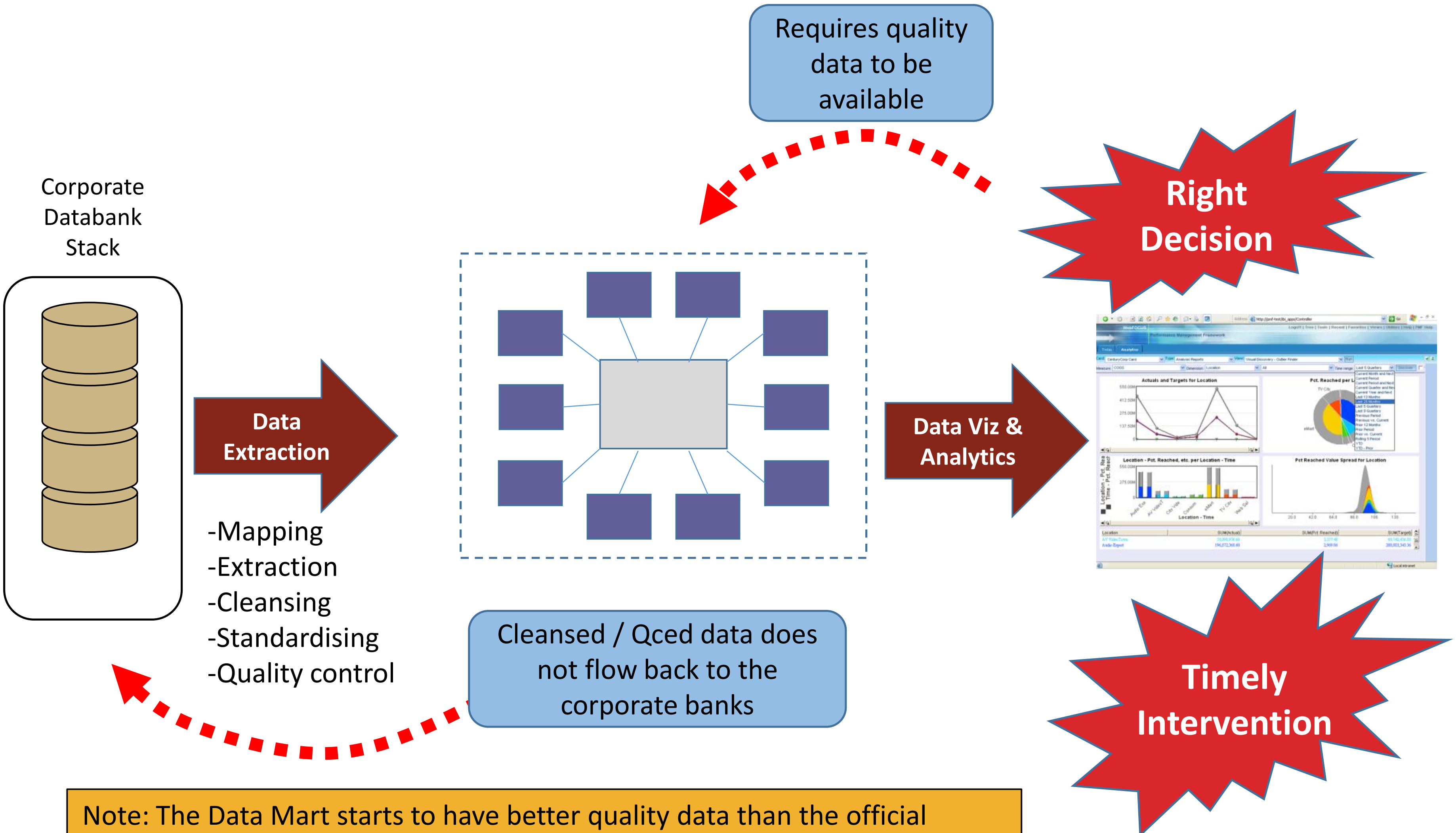
$$\chi^2 = \sum \frac{(O-E)^2}{E} = 20.9$$

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

Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

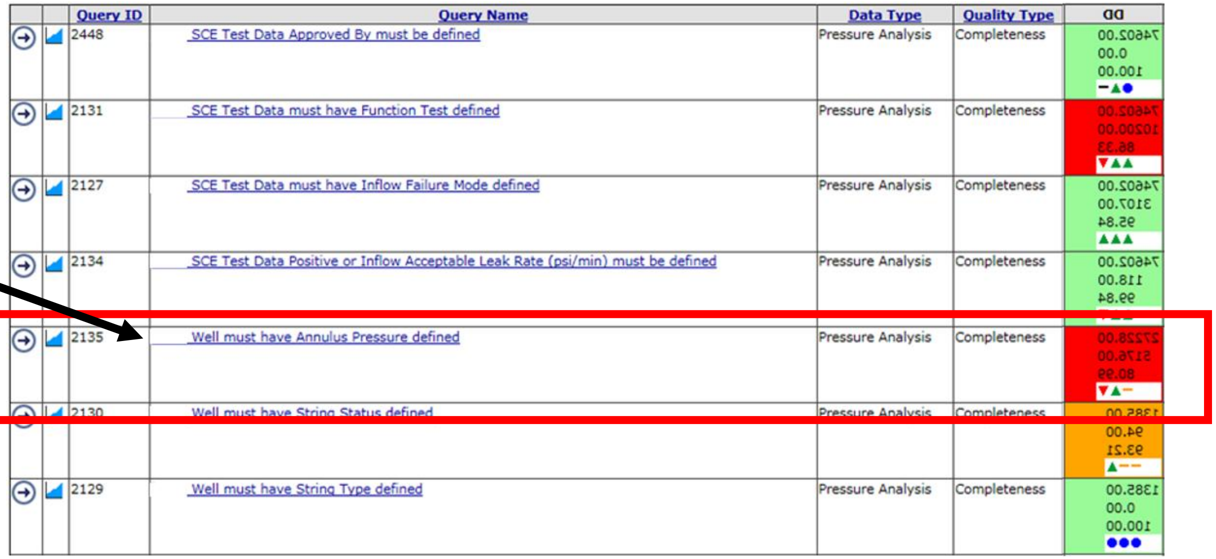
Data Analytics Conceptual Architecture



Note: The Data Mart starts to have better quality data than the official corporate databank

Data Quality Error Persistence

Business Rule: Well must have annulus pressure defined

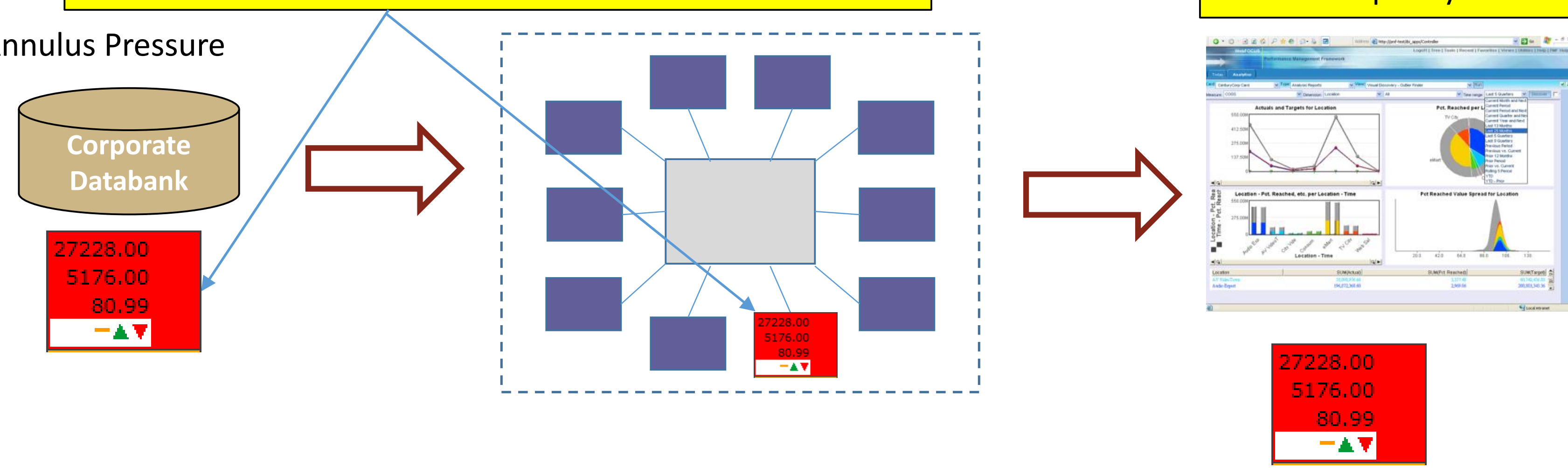


		Query ID	Query Name	Data Type	Quality Type	DQ
		2448	SCE Test Data Approved By must be defined	Pressure Analysis	Completeness	74602.00 0.00 100.00
		2131	SCE Test Data must have Function Test defined	Pressure Analysis	Completeness	74602.00 10200.00 86.33
		2127	SCE Test Data must have Inflow Failure Mode defined	Pressure Analysis	Completeness	74602.00 3107.00 95.84
		2134	SCE Test Data Positive or Inflow Acceptable Leak Rate (psi/min) must be defined	Pressure Analysis	Completeness	74602.00 118.00 99.84
		2135	Well must have Annulus Pressure defined	Pressure Analysis	Completeness	27228.00 5176.00 80.99
		2130	Well must have String Status defined	Pressure Analysis	Completeness	1385.00 94.00 93.21
		2129	Well must have String Type defined	Pressure Analysis	Completeness	1385.00 0.00 100.00

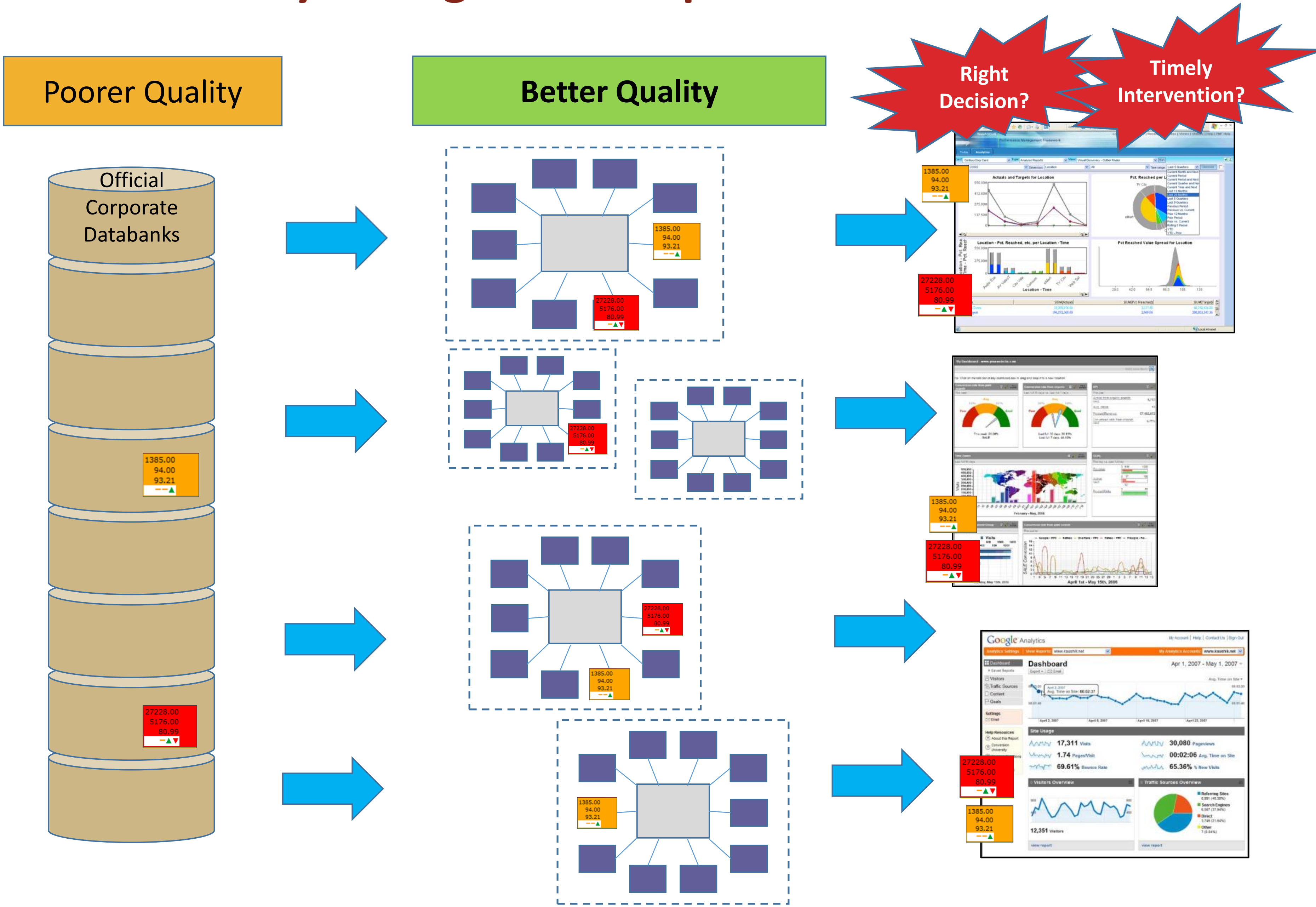
These errors will only be recognised if you are tracking the quality levels in the source databank

The analytics may not indicate quality levels

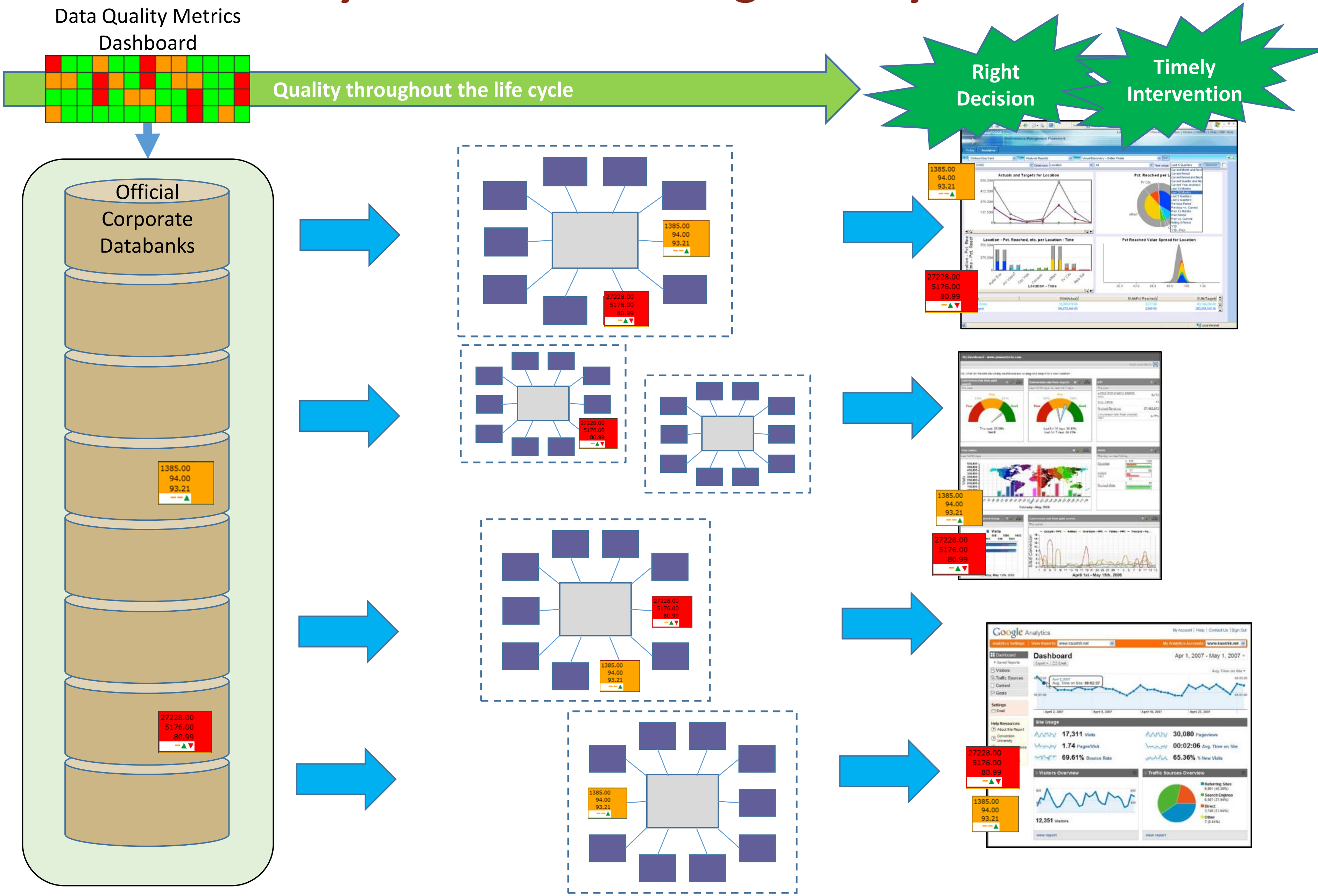
Eg. Annulus Pressure



Data Quality – Progressive Lopsidedness + Hidden Risks



Data Quality Metrics – Tackling Quality at the Source



Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

Data Science opportunities – Paleoenvironmental reconstruction

Structural
-faults
-uplifts
-eustatic
-erosion
-missing sections

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

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

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

Minerals
-glauconite
-siderite
-pyrite
-mica

Data Science opportunities– Source Rocks

Pressure

-Spot readings

-Trends

Temperature

-Sample readings

-Gradients

Surrounding wells

-well data

-Source rock distribution patterns

-maps & trends

Well Logs

-Gamma ray

-Sonic

-Density

-Resistivities

-Caliper

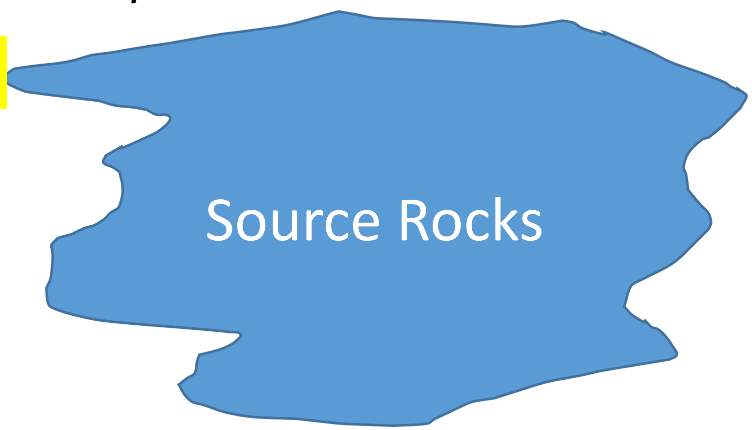
Sedimentary facies

-types

-characteristics

-bedding, dips etc

-log shape interpretation



Burial History

-Sedimentation rates

-Sediment types

-Missing sections

-Palinspastic reconstruction

Computer simulation

-Methods (eg Migration Models

-Probabliistic vs deterministic

Rock properties

-Porosity

-Permeability

-Diagenesis

Macerals

-Organic type (Lip. vs Vit.)

-Kitchen area

-Migration paths

-Maturity levels (DOM, VR/E)

Paleontology

-benthics

-planktonics

-larger forams

-nannofossils

-palynology

-ostracods

Data Science opportunities– Prospect appraisal

Prospect Appraisal

Temperature
-Sample readings
-Gradients

Pressure
-Spot readings
-Trends

Surrounding wells
-Well data
-Correlation
-Maps & trends

Rock properties
-Porosity
-Permeability
-Diagenesis

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

Analogues
-local comparators
-regional
-global

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

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

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

Structural
-faults
-closures
-seals

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

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

Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

Data Science Methods

Sequence Analysis	Spatial Analysis	Statistics	Multivariate Data Analysis	Artificial Intelligence	Probabilistic Methods
Markov Chains	Pattern Analysis (Random, Cluster, Nearest Neighbour)	Summary Statistics	Multiple Regression	Classification	Bayesian & Likelihood Methods
Runs Test	Analysis of directional data	Hypothesis Testing	Discriminant Functions	Natural Language Processing	Ranking & Scaling of Events
Least Squares & Regression Analysis	Spherical Distributions	t-Distribution	Cluster Analysis	Machine Learning / Deep Learning	Markov Chains
Splines (polynomial smoothing)	Fractal Analysis	F-Distribution	Eigenvalues & Eigenvectors	Text Mining	
Segmented Sequences & Zonation Analysis	Shape Analysis	Normal Distribution	Factor Analysis (R & Q Mode)	Graph Relationships	
Auto- and Cross-Correlation	Contouring, Trend Surfaces & Kriging	Chi Square Distribution	Principal Components		
SemiVariogram		Chi Square Goodness of fit	Correspondence Analysis		
Spectral Analysis		Regression	MultiDimensional Scaling		
		Analysis of Variance (ANOVA)	Canonical Correlations		
		Non-Parametric Tests			
		- (Mann-Whitney, Kolmogorov-Smirnov, Kruskal-Wallis)			

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

Data Types - Upstream

Review → Acquire → Explore → Appraise → Develop → Produce → Abandon

Geology & Seismic

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
-Gravity & Magnetic
-Microseismic

Interpretation and Compilations

Geology – Zones
Geology – Markers
Faults (Field Extent & Major)
Seismic Horizons – Regional
Seismic Horizons – Local
Velocity Models
Structure Maps
Time-Depth Curve
Gridded Time / Depth Maps
Sand Distribution Maps
Static Models
Dynamic Models
Synthetic Seismogram
Biostratigraphy – Zones
Biostratigraphy – Markers
Geology – Zones
Geology – Markers

Petroleum Engineering

Spill Points
Well Logs – Raw
Well Logs – Processed & Qced
Well Logs – Interpreted
Well Logs – Cased Hole
Vertical Seismic Profiling
Core Analysis including SCAL
Formation Pressure (RFT, MDT)
Well Tests (DST-Drill Stem Test, FIT-Formation Intake Test)
Production Data (Allocated oil/gas/water rates)
Production Pressure Data (Well Tubing/Casing Head Pressure)
Production Well Test
Artificial Lift
Fluid Property
Fluid Contacts
Stimulation Cases
Fluid Composition
Materials Balance
Decline Curve Analysis
Volumetrics
Reserves and Resources
Pressure Maintenance Cases
Saturation Height Function
Leak Off Test
PVT (Pressure-Volume-Temperature)

Drilling, Engineering & Production Operations

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 Classification – Digital Data (>100 types in Upstream)

← Primary Data → ← Secondary Data →

Original Format Data	Reference Data/ Metadata	Master Data/ Corporate "Single Source of Truth"	Derived Data	Data Collections
Raw Seismic Raw Logs	Units of measure - *Linear measures* - *Pressure*	Static (hard) data - *Well header* - *Deviation* - *Checkshot* - *Temperature* - *Pressure*	Processed data - *Seismic deconvolution* - *Seismic filtering* - *Seismic processing* - *Edited logs* - *Spliced logs*	Composite data - *Completion log* - *Mud log* - *Paleontological composites* - *TRAPIS*
	Abbreviations - *TD, DFE, KB etc*	Interpreted (soft) data - *Geological markers* - *Seismic horizons*	Interpreted data - *Geological markers* - *Seismic horizons*	Data hoards - *Projects en masse* - *Personal stores* - *Team folders*
	Valid Lists			Data archive - *Projects en masse*
	Range indicators			
	Comments			
Requires: - *Official data repository*	Requires: - *Standards* - *Implementation across all impacted tools and databases*	Requires: - *Clear processes, workflows and checkpoints* - *Proper & official repository* - *Management and security processes around repository and data access*	Requires: - *Standard workflows* - *Standard algorithms* - *Standard processes* - *Housekeeping procedures*	Requires: - *Standard display and formatting templates* - *Procedures*

Matching Tools to the Jobs – Emphasis Areas

Data science methods can be applied to all data. Here we look at where their application can be especially pertinent.

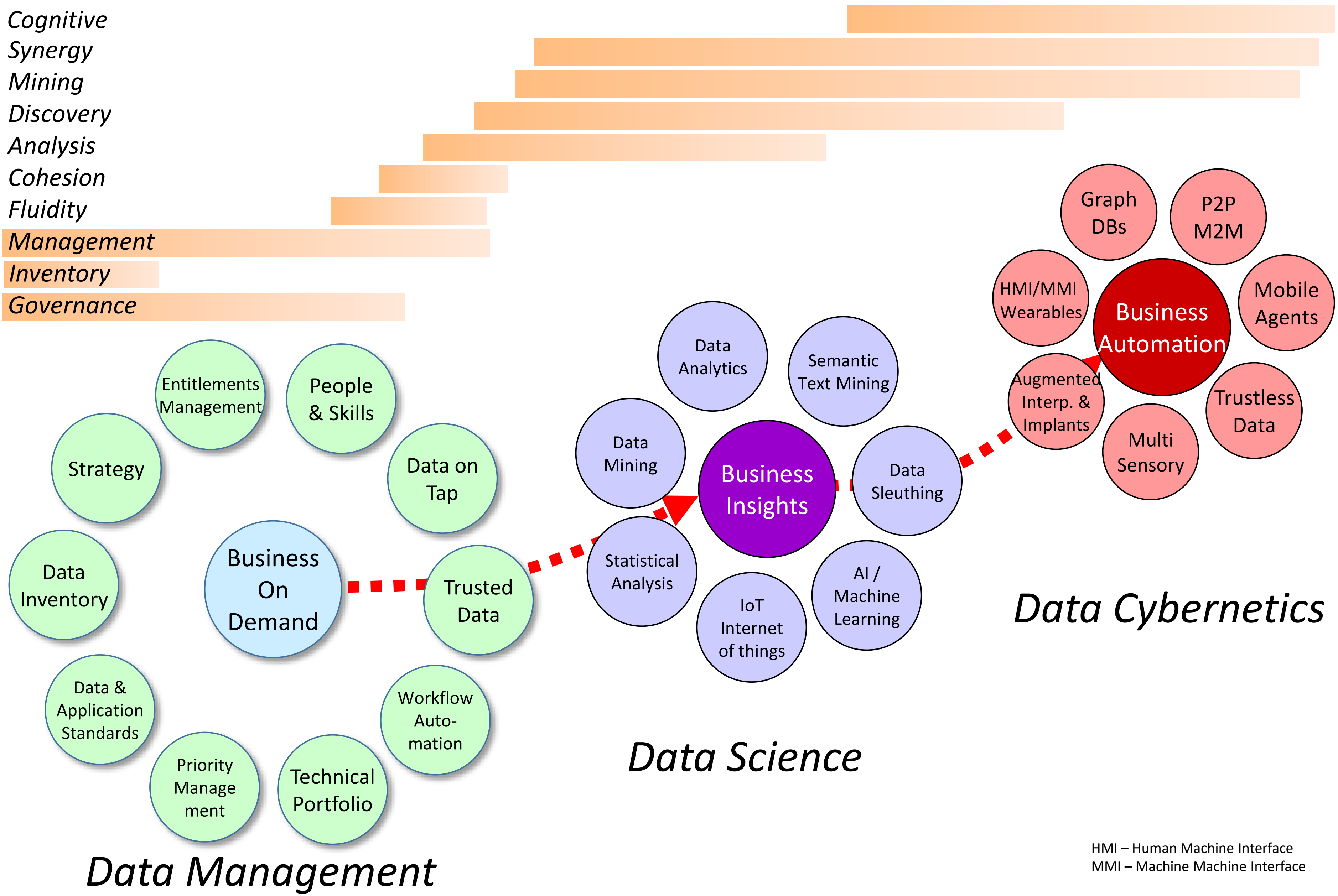
Top 2 areas

	Original Format Data	Reference Data/ Metadata	Master Data/ Corporate "Single Source of Truth"	Derived Data	Data Collections
Spatial Analysis	Original Format Data		Master Data/ Corporate "Single Source of Truth"		
Sequence Analysis			Master Data/ Corporate "Single Source of Truth"	Derived Data	
Statistics			Master Data/ Corporate "Single Source of Truth"	Derived Data	
Multivariate Data Analysis			Master Data/ Corporate "Single Source of Truth"	Derived Data	
Artificial Intelligence		Reference Data/ Metadata			Data Collections
Probabilistic Methods			Master Data/ Corporate "Single Source of Truth"		Data Collections

Structure

1. Introductory thoughts
2. Some typical data problems in EP
3. The data landscape in EP
4. Overview of data science methods
5. Examples
6. Data quality in context
7. Potential opportunity areas
8. More application areas – Tools to Jobs
9. The Future EP Data Driven Organization

The Future Data Driven EP Organization - Components



HMI – Human Machine Interface
MMI – Machine Machine Interface

Some Useful Reading

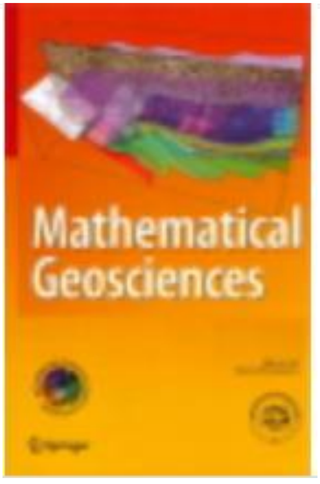
1. Statistics and Data Analysis in Geology. Davis, John C., 3rd Ed. 2002. Wiley
2. Harness Oil & Gas Big Data with Analytics. Holdaway, Keith R., 2014. Wiley
3. Building Expert Systems. Frederick Hayes-Roth, Donald A. Waterman, Douglas B. Lenat, 1983. Addison-Wesley
4. Quantitative Stratigraphy. F.M. 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.

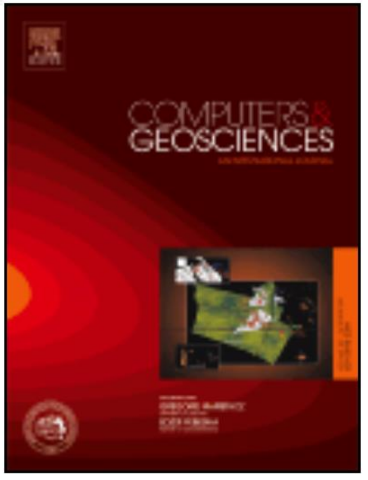
Font size Bigger Reset Smaller

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.





Questions

Philip Lesslar
Data Solutions Consultant

plesslar1@gmail.com