Upstream Data Management in the 2020s:
New uses for old skills in the world of Big Data

Dr Duncan Irving

DEJ Aberdeen – September 29th, 2015

Our workflows haven't really changed much since the first data started coming back to shore with the oil…

… but now we must run our workflows at the pace of the business processes to ensure maximum value…

©2015 Google

How do you optimise logistics and maintenance to minimise shut-ins?

TERADATA.

… but now we must run our workflows at the pace of the business processes to ensure maximum value…

Data-driven operations

Since 2004, UPS has eliminated millions of miles off delivery routes through advanced analytics. They've:

- Saved 45 million litres of fuel ($3.5M/yr)
- Reduced CO_2 emissions by 100,000 tons, equivalent to 5,300 passenger cars off the road for an entire year.

TERADATA.

…from an operational viewpoint…

Where do you site a well to maximise profitability?

TERADATA.

...from an operational viewpoint...

Data-driven development

Walmart, and many other retailers, routinely integrate functional domains and value chains. New stores are sited based on:

- Potential revenues of product mix
- Cost of supply to store from one of Walmart's 3200 Supercenters
- Daily/Monthly/Seasonal effects
- Long-term demographics

…and from a strategic planning viewpoint

**How do you optimise production
and development on mature fields?**

TERADATA.

…and from a strategic planning viewpoint

Data-driven planning

Since 2007, Daimler has relied on an integrated view of all production line, diagnostic and dealership data for cars and trucks.

They can:

- Identify emerging problems in vehicles within weeks, and correct production processes accordingly.
- Reduce long-term warranty cost and development risk.

Oil & Gas isn't the only industry where everything is more joined up and going faster

Industry 4.0 is changing traditional manufacturing relationships

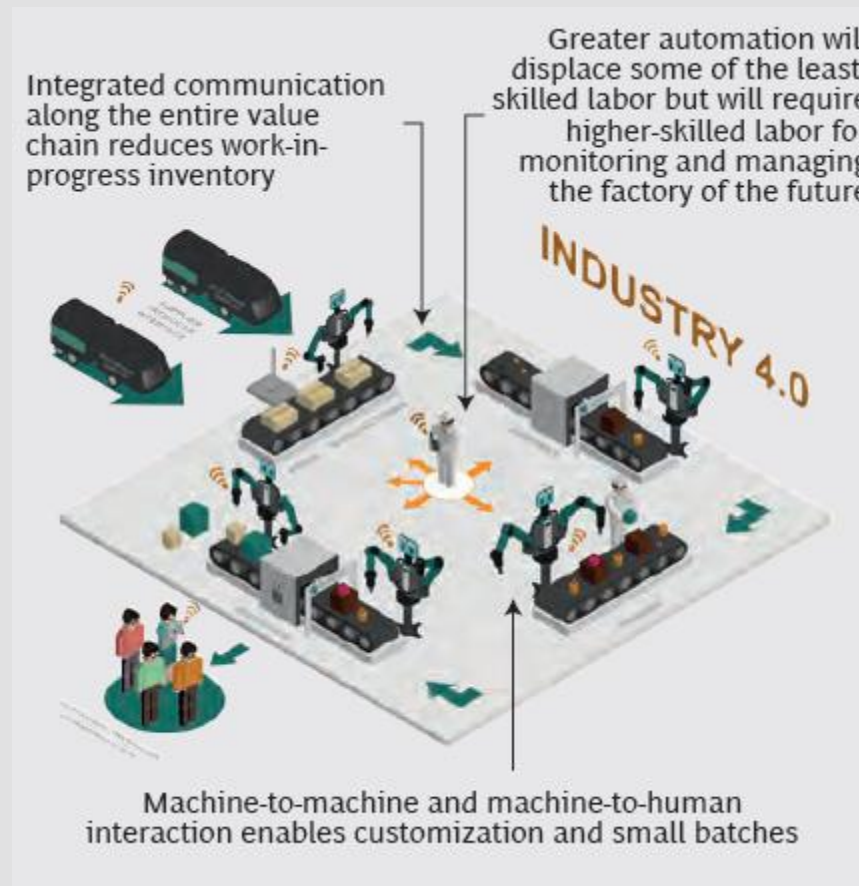From isolated optimized cells…

Source: BCG.

Oil & Gas isn't the only industry where everything is more joined up and going faster

Industry 4.0 is changing traditional manufacturing relationships
    …to fully integrated data and product flows across borders

Integrated communication along the entire value chain reduces work-in-progress inventory

Greater automation will displace some of the least-skilled labor but will require higher-skilled labor for monitoring and managing the factory of the future

INDUSTRY 4.0

Machine-to-machine and machine-to-human interaction enables customization and small batches

Source: BCG

How does this look in E&P? Currently…

Technical Domains

Key Activities

TERADATA.
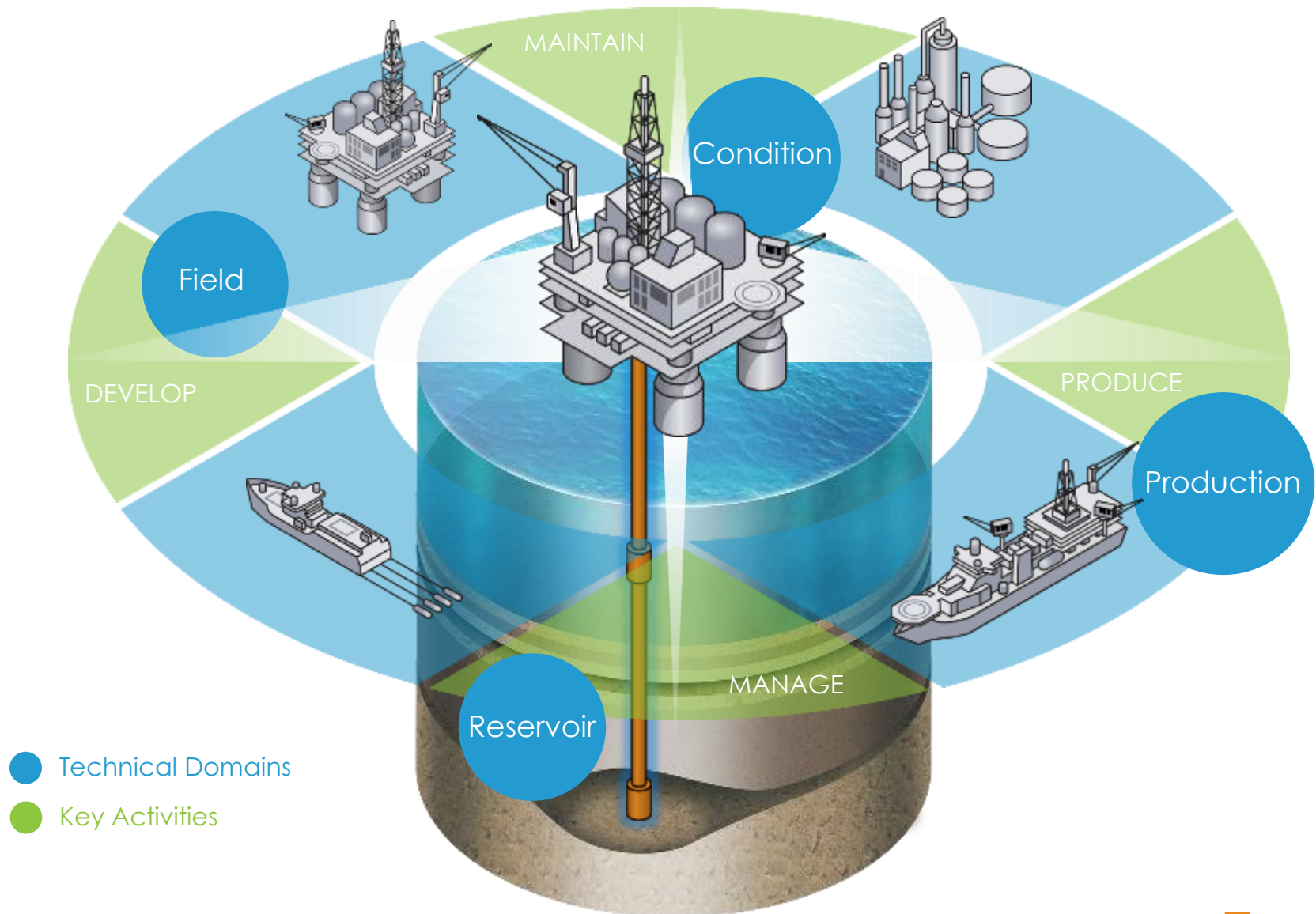
Where we need to get to – the Connected Well

COTS

MAINTAIN

Condition

DEVELOP

PRODUCE

MANAGE

Service Companies

Systems Integrators

Engineering Consultancies

Field

In-house Expertise

Drilling Companies

Systems Integrators

Production

Service Companies

Reservoir

- Technical Domains
- Key Activities

TERADATA
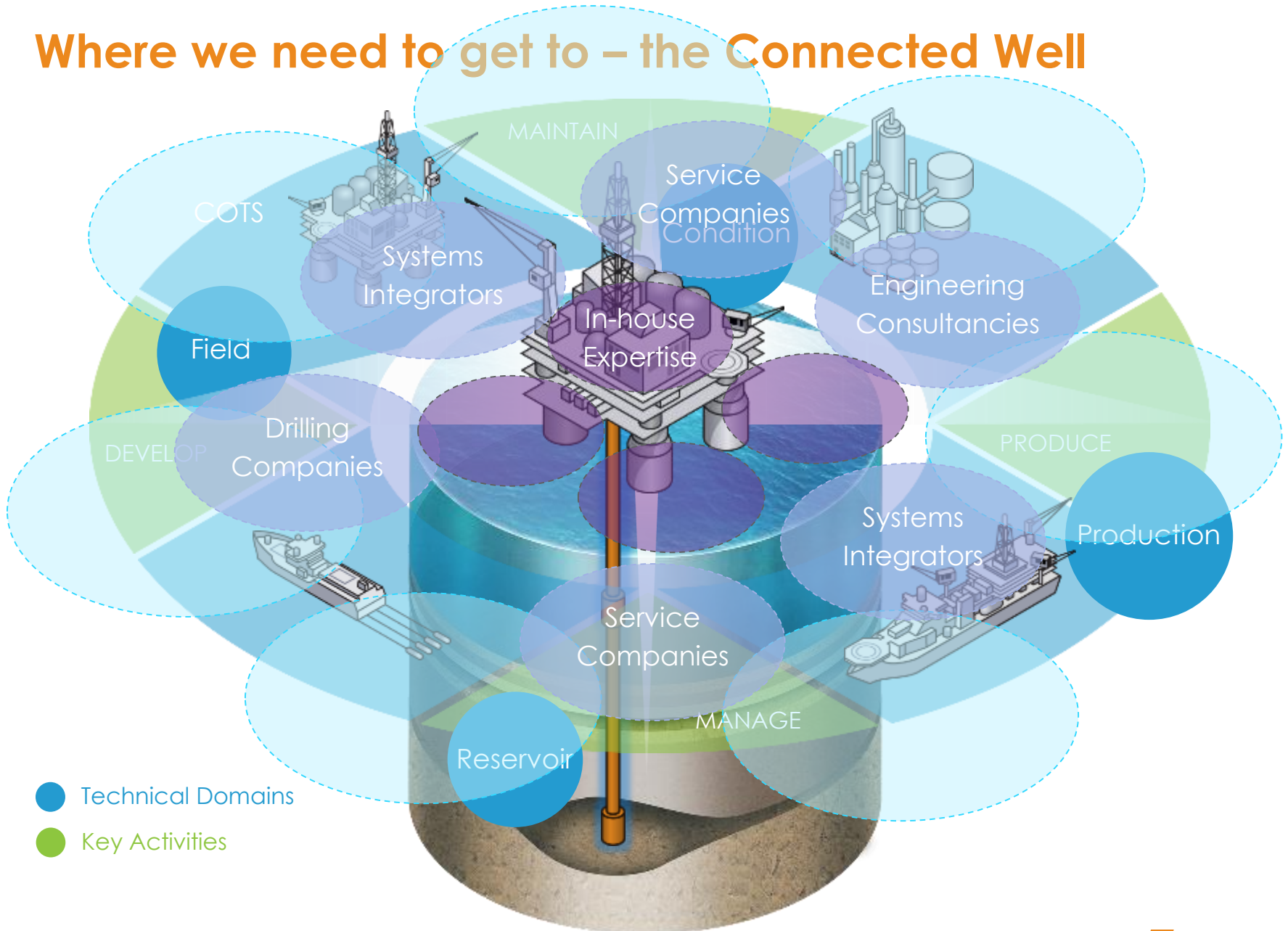
What does this mean for data management?

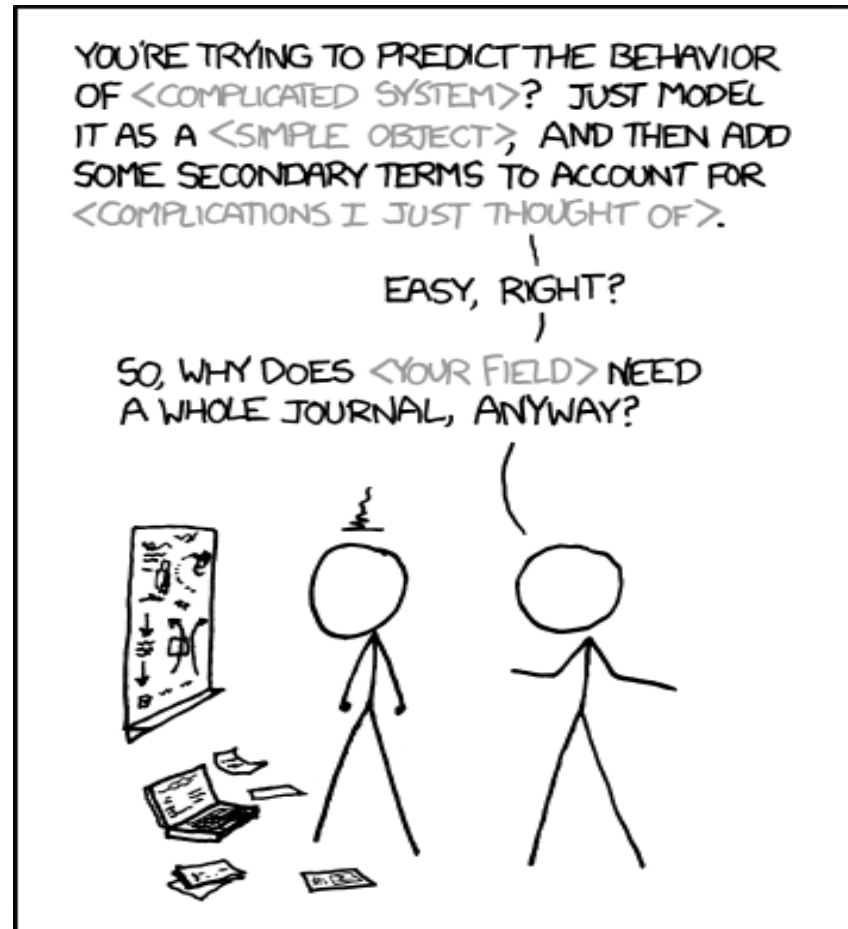The amount of data captured and the speed at which it must be contextualised is our Big Data problem

Upstream IT is the way it is for a reason – **(my six S's)**
- Size
- Science
- Spatial
- Speed
- Sustainability
- Sikkerhet

Why is it so difficult?

Source: xkcd.com

Our data managers are **highly skilled "librarians"** who want to deploy their domain expertise much more than they do

TERADATA

What will the new data manager look like?

How can data managers enable this?

- Use your understanding of the data to enable better re-use for analytics

- Use your knowledge of the business processes to build-in discovery and scalability

- Don't fear the unknown – other industries have travelled down this path too

© 2014 Teradata   You are the answer!   TERADATA

How to experiment with mixing data together

How to experiment with mixing data together

SQL

Business

Unstructured

Multi-structured

Structured

Scientific

Technical

FORTRAN

MATLAB

TERADATA
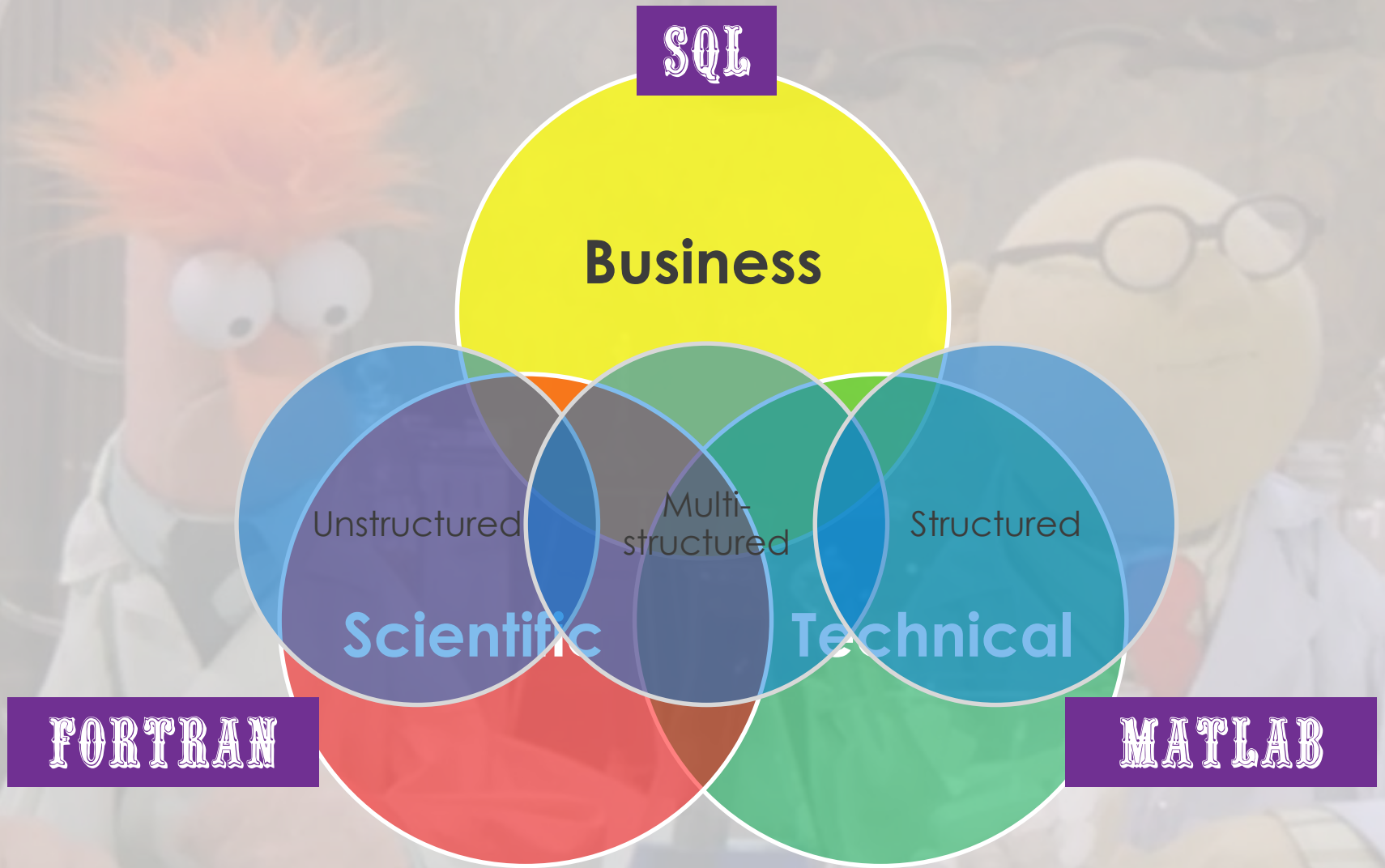
How does a company get started?

TERADATA
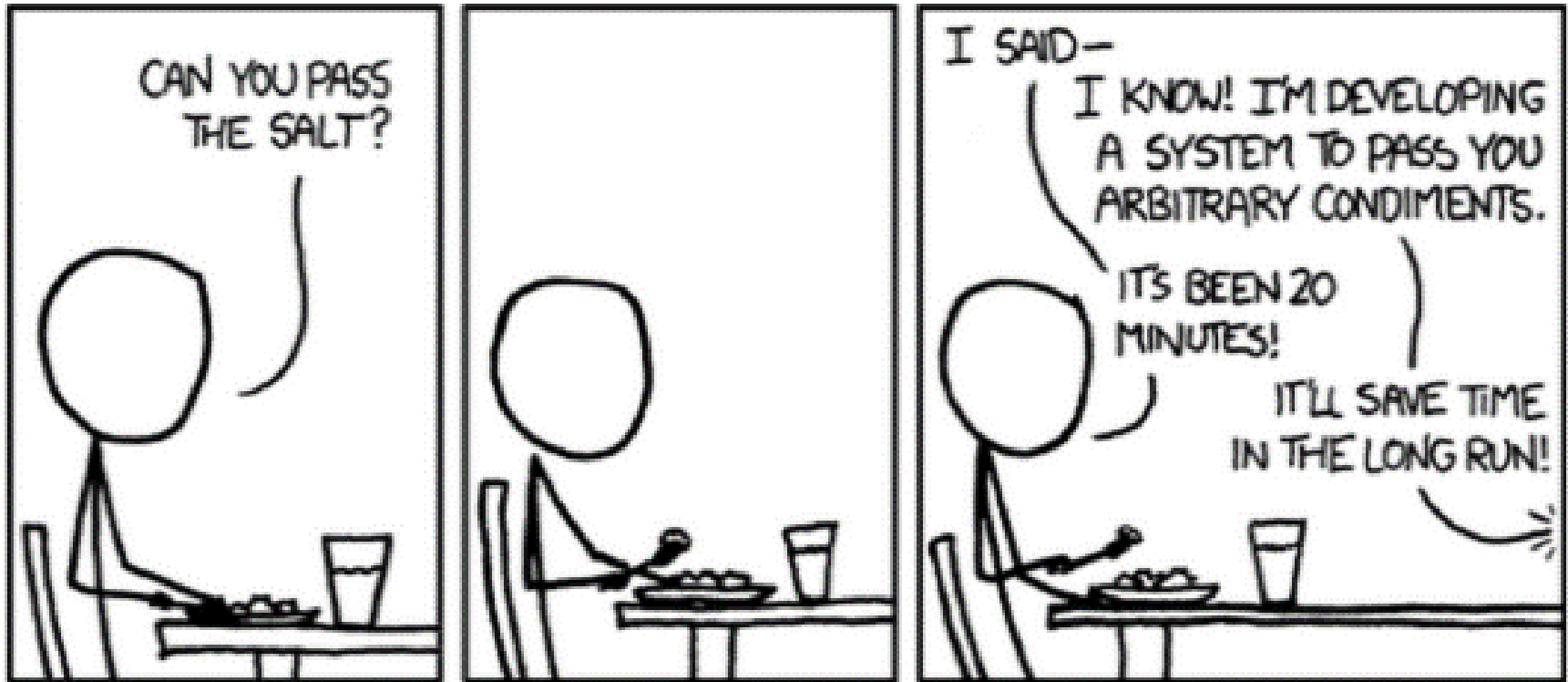
Basin-scale prospectivity analytics

What can we learn from a basin's worth of subsurface data in six weeks?

6 week project undertaken by a MSc student at the University of Manchester with public access data from New Zealand:

- Clean data

- Establish context

- Define targets

- Establish reliability

Supported by Teradata Oil & Gas and Analytics teams

Technical goal: Integrate data

Establishing logical, spatial and stratigraphic relationships across wells

Created pragmatic data model from:

- LAS files
  - 521 wells
  - 25,081 curves
  - 81,141,281 indexed data points
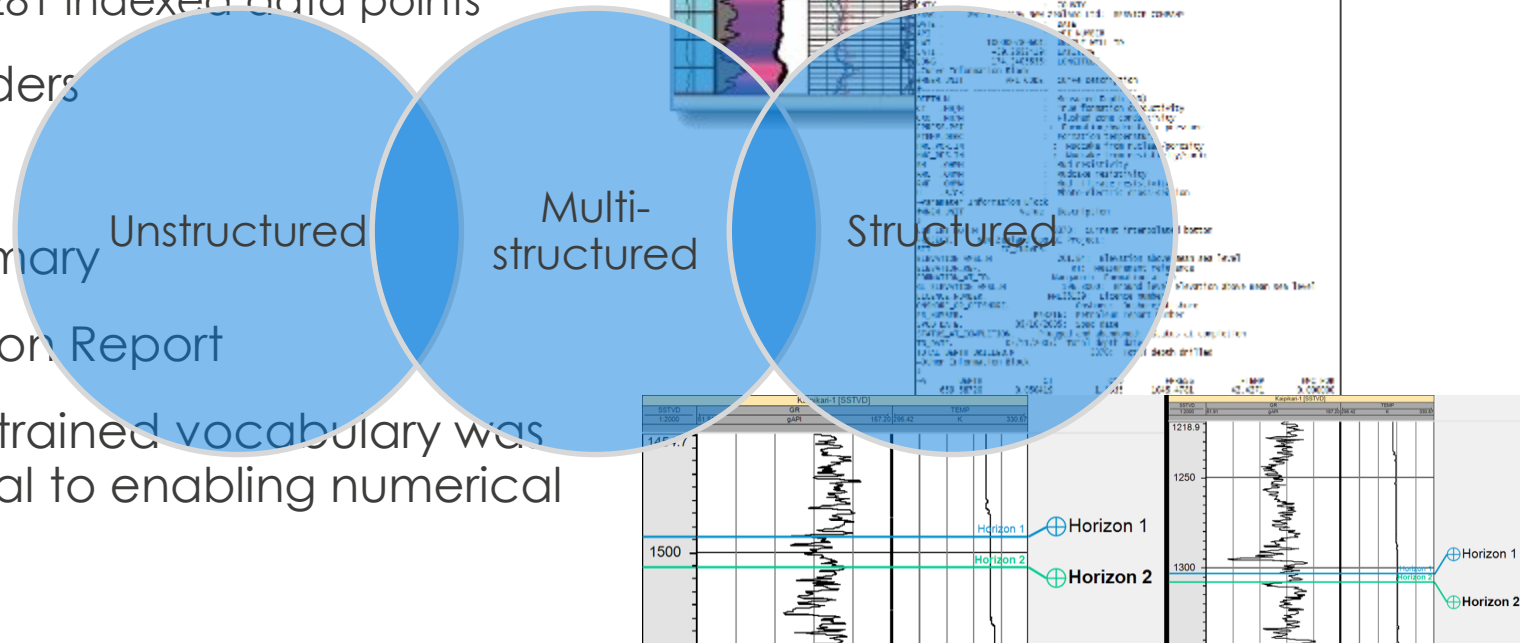- Well headers
- Mud logs
- Well summary
- Completion Report

A well constrained vocabulary was fundamental to enabling numerical analysis

Unstructured

Multi-structured

Structured

TERADATA.

1/ Take wells apart and add geological context

Geological data

Calcareous sandstone interbedded with calcareous siltstone.
Interbedded mudstone, siltstone and sandstone.
Interbedded sandstone, siltstone and mudstone with rare limestone stringers.
Calcareous mudstone.
Limestone.
Argillaceous siltstone and mudstone.
Sandstone, calcareous cement.
Glauconitic siltstone and sandstone.
Calcareous sandstone interbedded with calcareous siltstone.
Interbedded mudstone, siltstone and sandstone.
Interbedded sandstone, siltstone and mudstone with rare limestone stringers.
Calcareous mudstone.
Limestone.
Argillaceous siltstone and mudstone.

LAS files

- Header
- Metadata
- Log curves
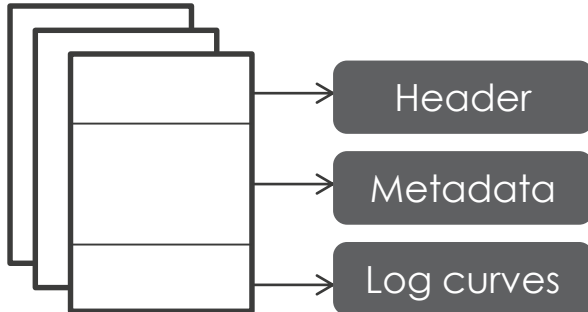
Cleansing

- File version
- Well summary
- Logs taken
- Well parameter
- Other
- ASCII Logs

Geological elements

Wells Table

Log fact Table

Formation and Member Tables
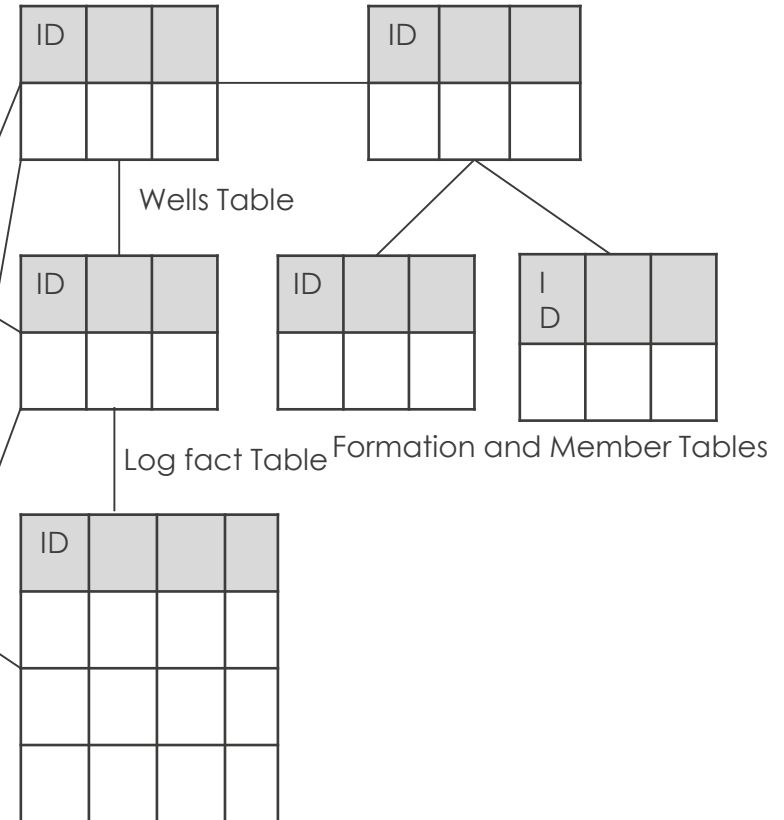
ASCII log Table

Cleaning the text data

Formation name	Member name
Moki formation	Moki
	Moki A
	Moki A Sandstone
	Moki B
	Moki B Sandstone
	Moki B Sandstone Interval
	Moki Equivalent

Formation name	Member name
Moki formation	Moki A Sandstone
	Moki B Sandstone
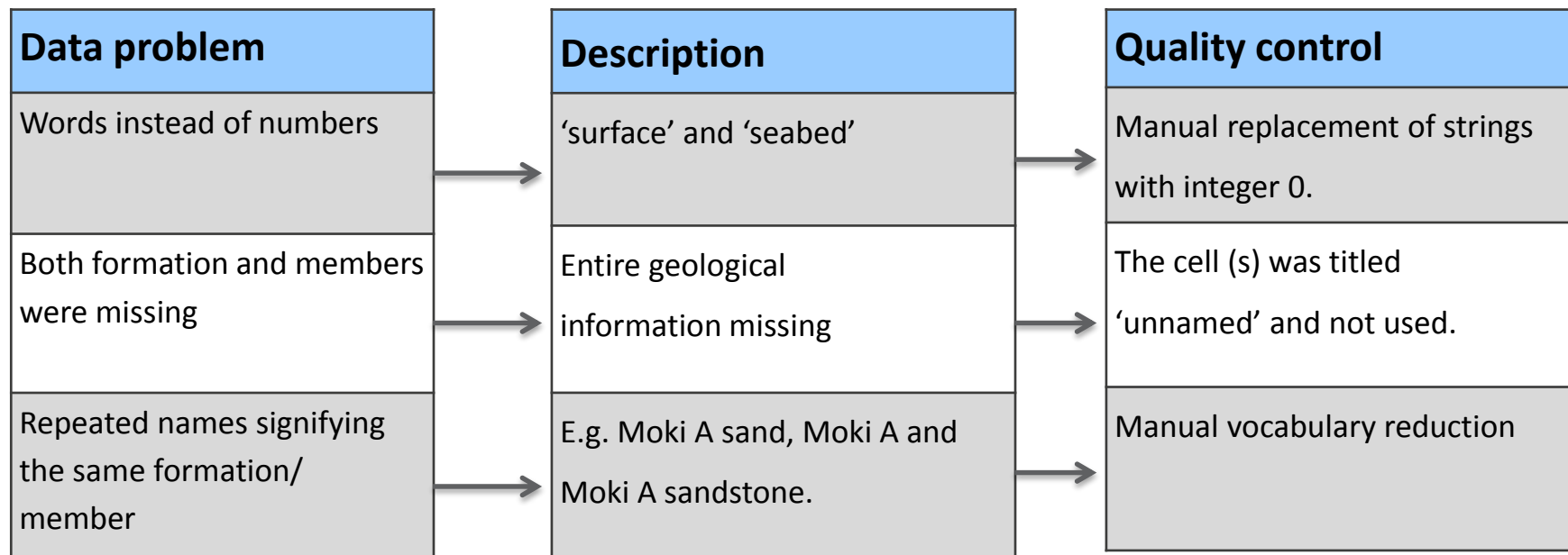
Data problem	Description	Quality control
Words instead of numbers	'surface' and 'seabed'	Manual replacement of strings with integer 0.
Both formation and members were missing	Entire geological information missing	The cell (s) was titled 'unnamed' and not used.
Repeated names signifying the same formation/ member	E.g. Moki A sand, Moki A and Moki A sandstone.	Manual vocabulary reduction

Symbolic Aggregate approXimation

Finding distinct peaks – e.g. thermal "spikes" around hot shales

base dataset → analytical dataset → data demographics → SAXify → match *similar* text → locate

Saxification →

Symbol size = 3
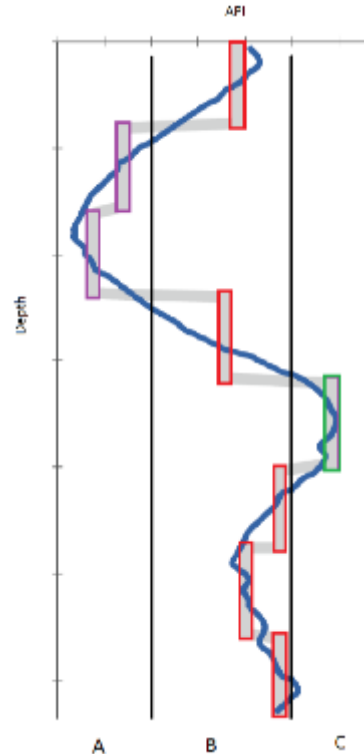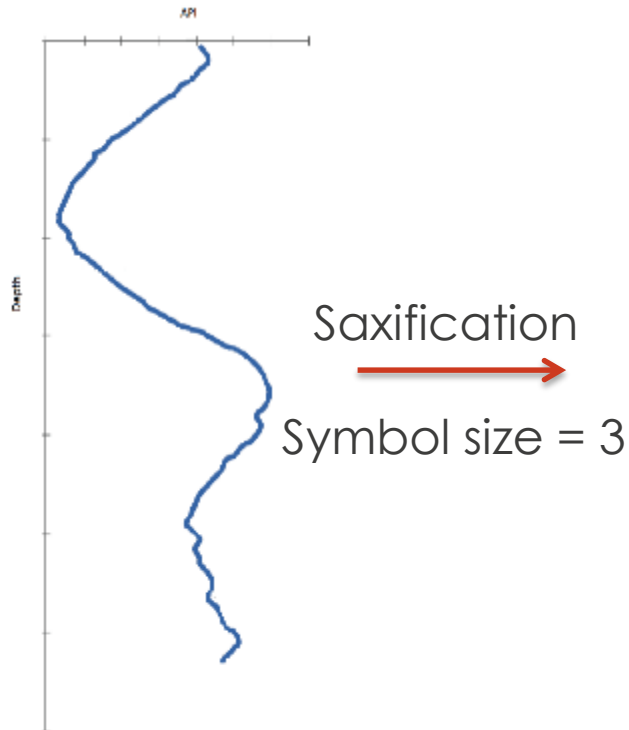
After conversion →

BAABCBBB

Symbolic Aggregate approXimation

Finding distinct peaks – e.g. thermal "spikes" around hot shales

base dataset			
Uwi	Depth	Curvename	curvevalue

analytical dataset				
Uwi	Depth	FTEMP	curvevalue	partitionID

data demographics

Calculate mean and standard deviations for each temperature curve

SAXify

Apply symbolic notation based on data demographics

match *similar* text

Match with standard regular expression notation e.g.
dd*gg
dd*[d-h]

locate

Project matches back to original depths in core data set and validate

---

σ, T$_{mean}$

dddddiggddddddd,
dddddfdehfdddddd,
bbbbbbbbcbbccbbbbb

...ddddddiggdddddddd...,
...ddddddfhfdddddd...,
...bbbbbcddbbccbcbebb...

Dynamic Time Warping

Workflow for classification of interbedded sandstone/mudstone and sandstone/siltstone facies:

base dataset → analytical dataset → nPath → Rebuild and pivot → Apply time warping → locate

Dynamic Time Warping

Workflow for classification of interbedded sandstone/mudstone and sandstone/siltstone facies:

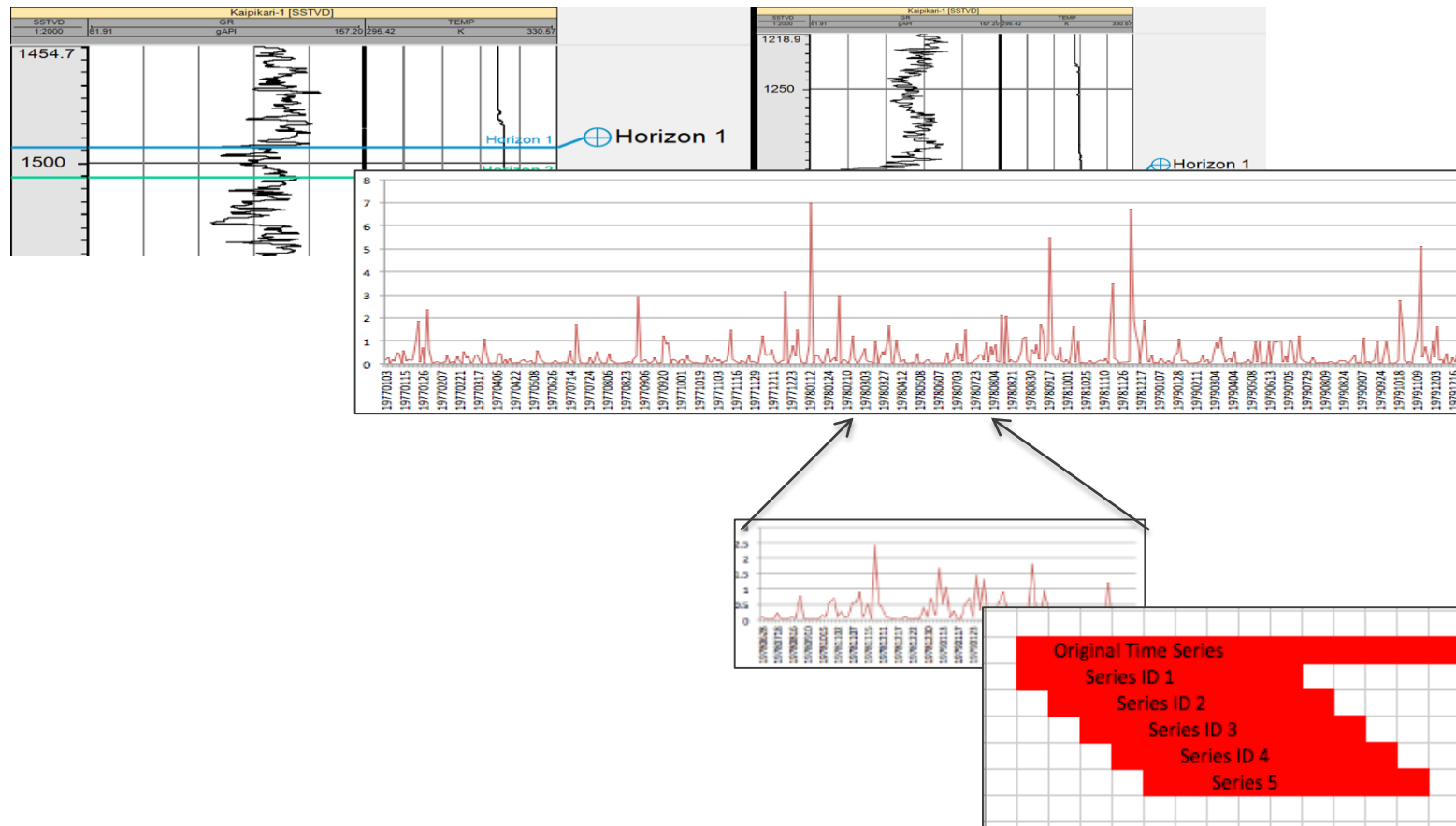**base dataset** → **analytical dataset** → **nPath** → **Rebuild and pivot** → **Apply time warping** → **locate**

Create multiple windows (paths) along

Flip pairs of points into separate key-value pairs and rebuild as

Create *template* and map onto partitions in pre-built table then calculate goodness of fit using DTW

Take top centile of matches and validate

Exploit parallelism to match multiple partitions

Not performed as a closed loop or with any degree of automation or learning

storage and rapid access enables novel access paths by spreading partitions across

Uwi	Depth	Curvename	curvevalue

Uwi	Depth	GR	curvevalu	partitionID

New insights

- A much clearer, simpler model of the reservoir with 62 members in 17 formations

- Identified un-noticed features (hot shales) and re-classified others (interbedded facies)

- Ask any question of the data with spatial, chronological and logical relationships – **at scale**

- An open-ended model to incorporate other data – **(next step: production histories)**

Conclusions

TERADATA.