How Data Management can put the Science into Data Science

Dr Duncan Irving, Lead Consultant Oil & Gas

Digital Energy Journal event, KLCC 2016

Big Data and Data Science: disruption and innovation

How we understand and interact with each other

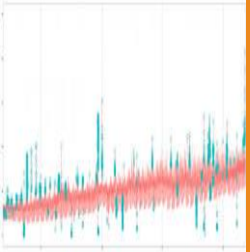
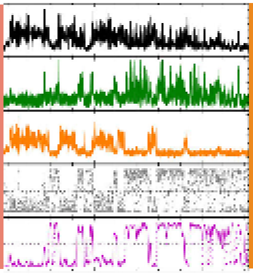
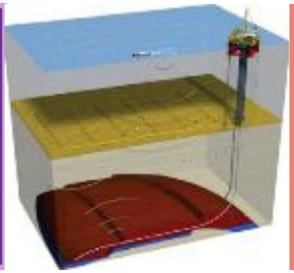
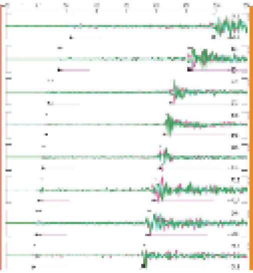
How organisations understand and interact with us

How we interact with technology and services

How we exploit knowledge…
at scale and pace

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

“New data” comes in three flavours

It comes from		It can contain…		It has impact
Fleets: from lots and lots of similar things		**Outliers:** Which of my things are behaving differently?		“Fleet-wide” 24/7 for holistic management
Systems: across the same big “thing”		**Emergent behaviour:** Is my system changing to a new state?		High-level KPIs at business units and facilities level
Collectors: “big models” or monitoring		**Events:** are there hidden signals?		Performed at sub-second level and data kept for decades

…but that looks a lot like the old data!

Yes, but the KPIs are different

- Business related
- Business budgets, not IT (Low Capex / spend from Opex)
- Show business value – early, and continuously

Our data managers are **highly skilled "librarians"**

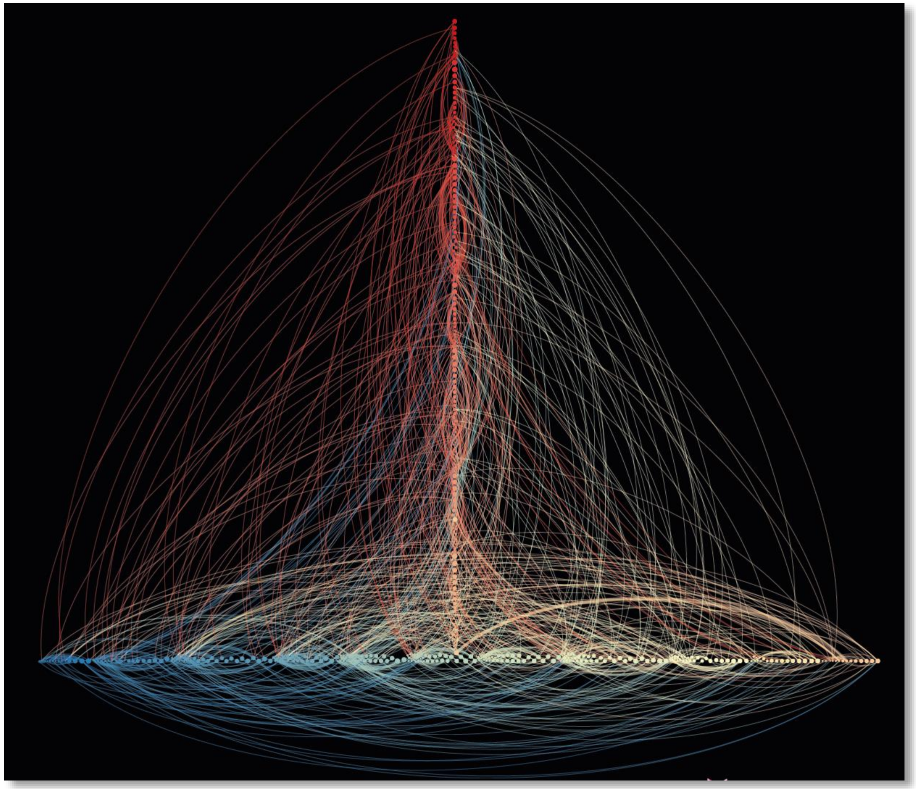
- curate measurement data
- Ad hoc management of interp
- "work to spec"

…but want to deploy their domain expertise much more!

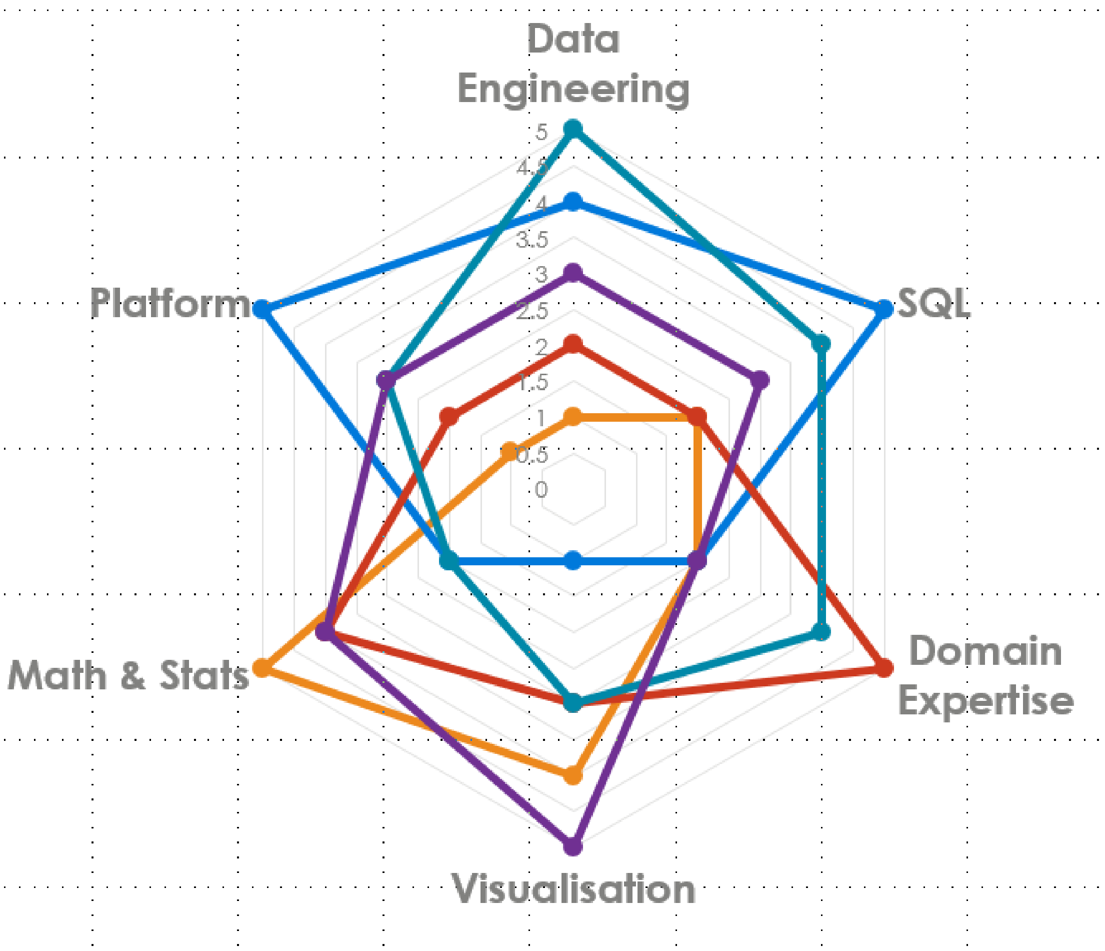
TERADATA.

What is data science?

- Finding relationships with complex data sets
- Characterising behaviour and understanding the **demographics** of data
- It can be applied to:
  - Data profiling and QC
  - Data preparation
  - Data mining
  - Operational processes
  - Data art



What does a data scientist look like?



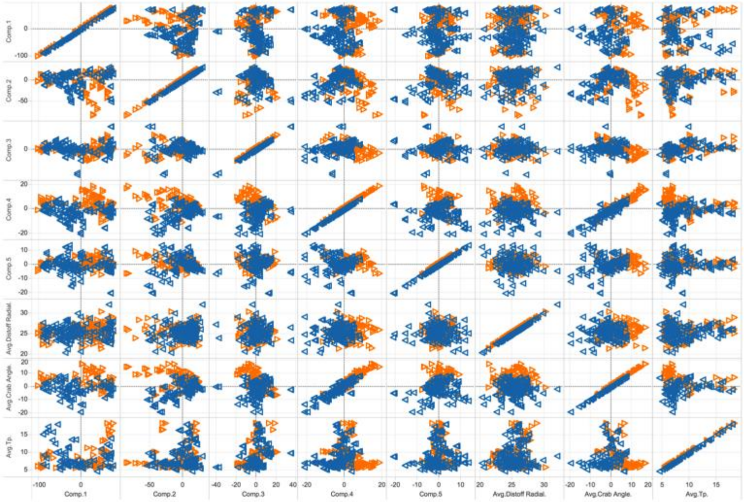
- No such thing as a perfect data scientist
- You need outstanding data management and data engineering skills (and culture)
- For sustainability and deployment you need platform expertise

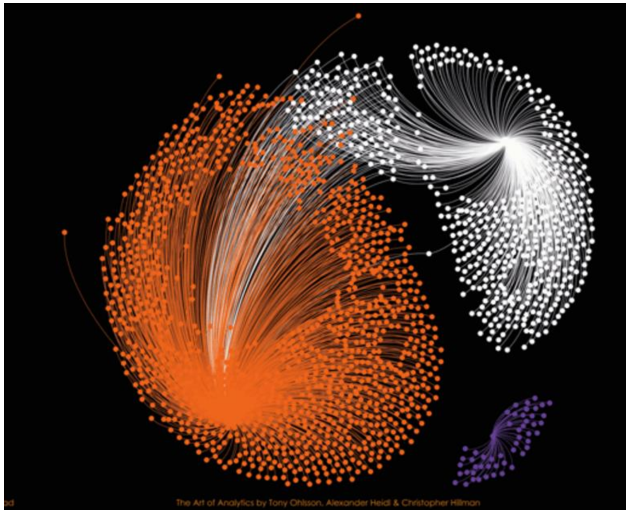
So where does the SCIENCE come in to it?

- Let's widen it out to STEM: Science, Technology, Engineering and Maths
- In upstream this is:
  - (mainly) the physical sciences
  - Spatial relationships and geospatial measurements
  - Lots of time series
  - Engineering concepts
  - Operational science
- It's not like banking and retail – they can do this because they've evolved with analytics and BI over the decades and their mindset is already data-driven
- The applications have grown around the scientific questions and the mathematical algorithms and many were baked-in or black-boxed years ago

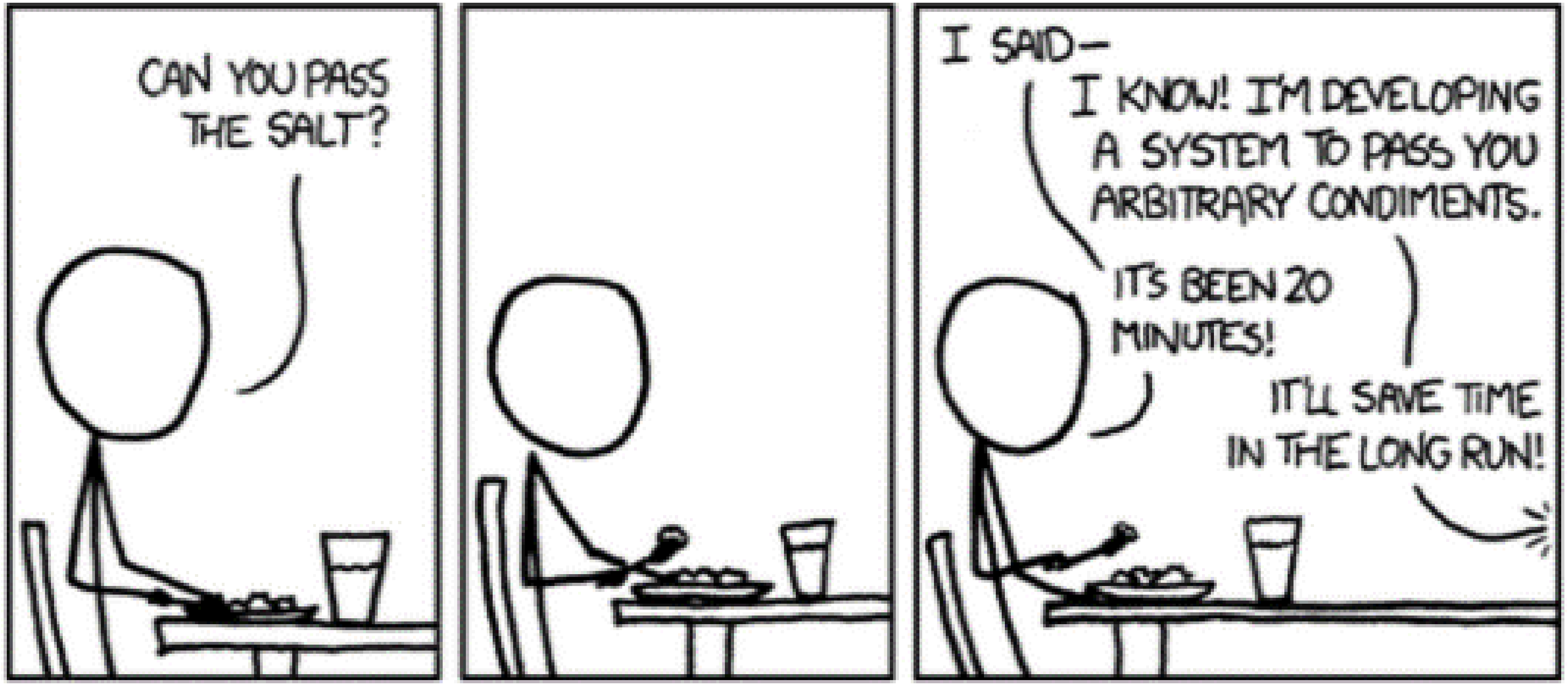
Let’s recap

- Other industries achieve high value from their data (even their digital exhaust)
- They use statistical approaches to great effect
- We’ve got some very smart people in our own industry but
  - They can’t access the data
  - They didn’t read that part of the maths book at university
  - No one knows what tools to use, or how to use them (see above!)
  - No one trusts the use cases because they’re not Oil & Gas




The Art of Analytics by Tony Ohlsson, Alexander Heidl & Christopher Hillman

How do we move forward?


CAN YOU PASS THE SALT?
I SAID–
I KNOW! I'M DEVELOPING A SYSTEM TO PASS YOU ARBITRARY CONDIMENTS.
IT'S BEEN 20 MINUTES!
IT'LL SAVE TIME IN THE LONG RUN!

E&P data management – stuck in the 90s

Curation and custodianship

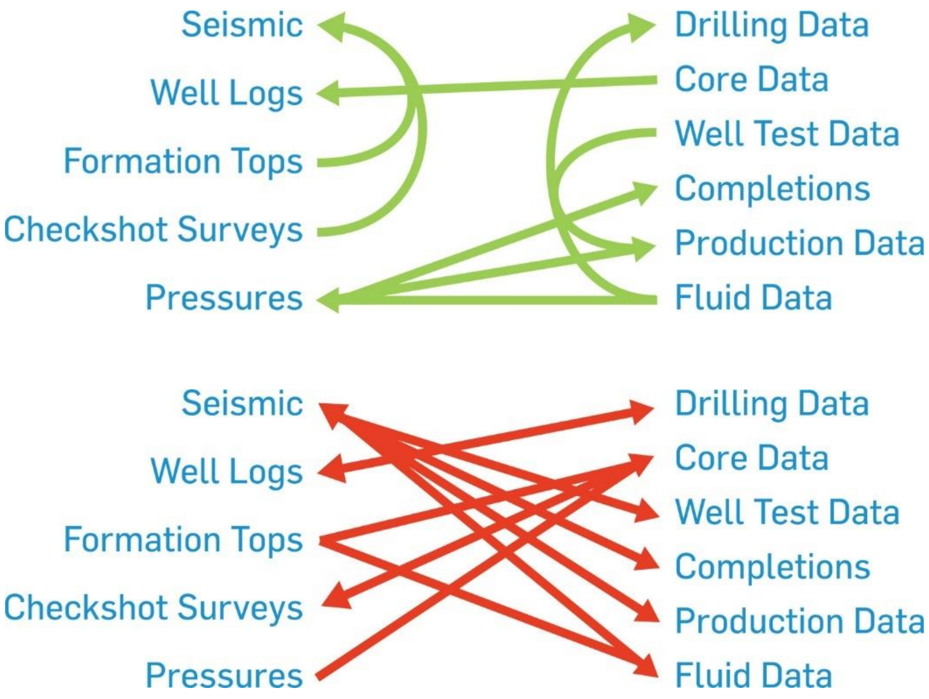
Self-describing file formats

The E&P data challenge

Data Management problems in existing workflows

- "Knowledge development" applications come with import filters for specific file types and specific tasks
- Data is modelled logically for well-defined (and hence brittle) processes that may not reflect all (or even any!) use cases
- Only "perfect" data can be imported into applications or schemas

New data types, or new combinations challenge all of this



- Kerry Blinston, CGG, ECIM 2014

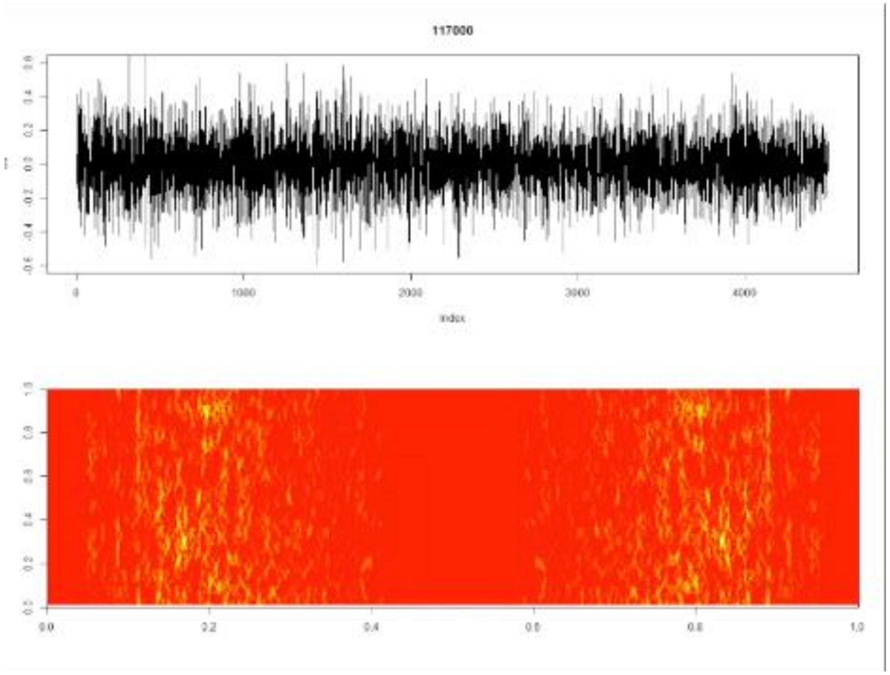
What should data look like?

Don’t be a data hoarder!

Why not store data at a granularity good enough to extract value?

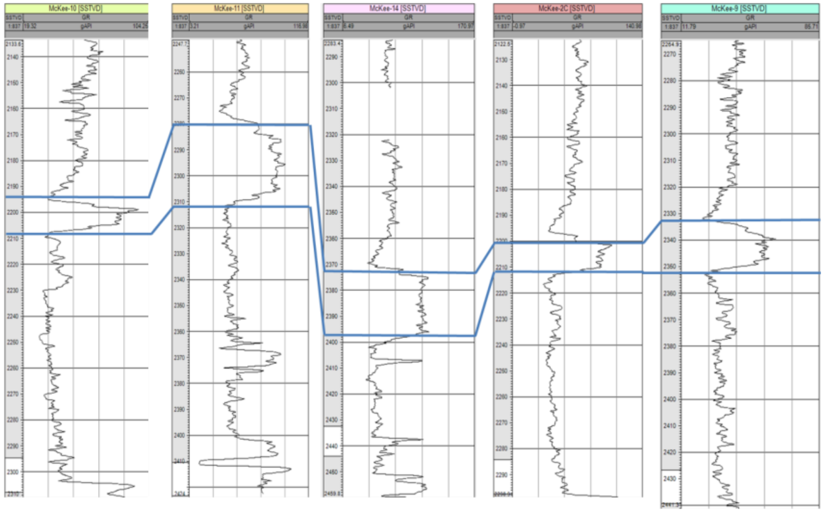
- Granular enough
- Dimensioned (time, space) enough
- Resample, interpolate, aggregate

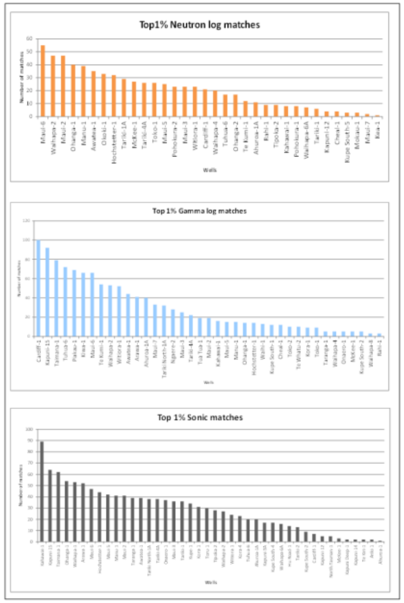
GOOD ENOUGH!



“Difficult file formats” (Multi-structured data)

- Parse out the measurement data
- Link it through time and space
- Relate using metadata and master data
- Form a view on whether a hypothesis is worth developing





GOOD ENOUGH!

Dealing with unstructured data

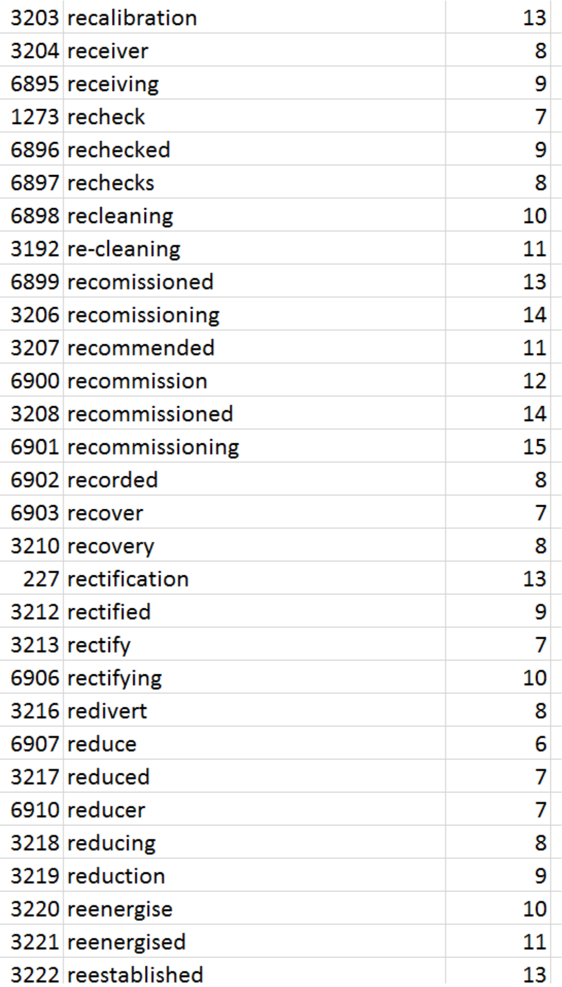
Text

- Language
- Typos
- Consistency
- Quality

Use simple characterisation tools to understand what is in the data

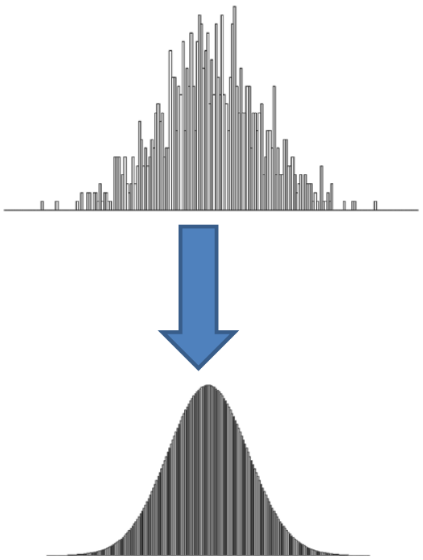
Don’t try to build a whole text input and cleansing framework

GOOD ENOUGH!



3203	recalibration	13
3204	receiver	8
6895	receiving	9
1273	recheck	7
6896	rechecked	9
6897	rechecks	8
6898	recleaning	10
3192	re-cleaning	11
6899	recomissioned	13
3206	recomissioning	14
3207	recommended	11
6900	recommission	12
3208	recommissioned	14
6901	recommissioning	15
6902	recorded	8
6903	recover	7
3210	recovery	8
227	rectification	13
3212	rectified	9
3213	rectify	7
6906	rectifying	10
3216	redivert	8
6907	reduce	6
3217	reduced	7
6910	reducer	7
3218	reducing	8
3219	reduction	9
3220	reenergise	10
3221	reenergised	11
3222	reestablished	13

Profiling data

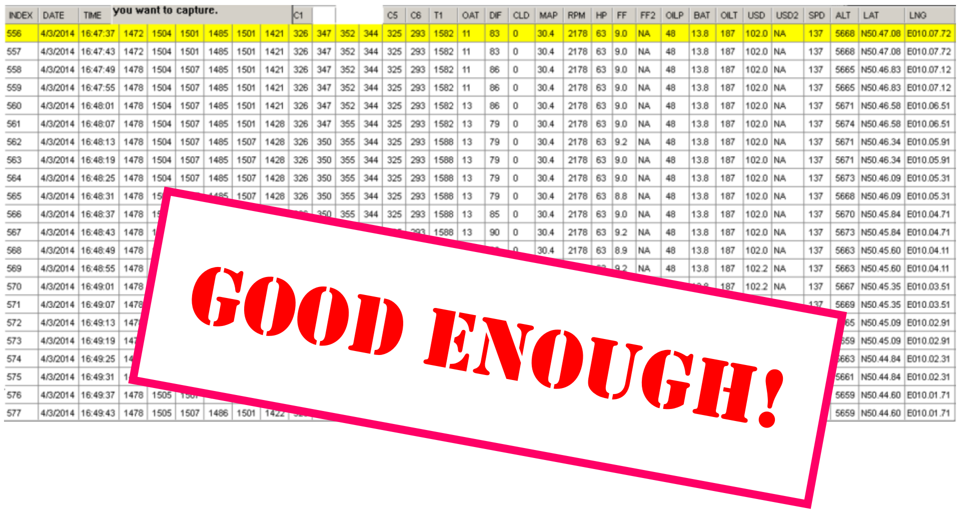


Historically in E&P

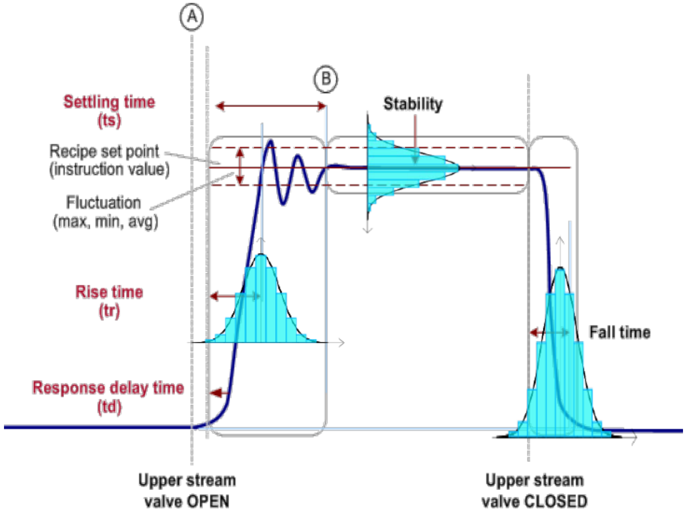
- We've stripped all the context away from each measure and observation for the sake of more storage

But now, in 2016:

- Storage is cheap
- If the data is still to large to handle then profile and decimate (it's better than never using it!)



INDEX	DATE	TIME	you want to capture.						C1				C5	C6	T1	OAT	DIF	CLD	MAP	RPM	HP	FF	FF2	OILP	BAT	OILT	USD	USD2	SPD	ALT	LAT	LNG
556	4/3/2014	16:47:37	1472	1504	1501	1485	1501	1421	326	347	352	344	325	293	1582	11	83	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5668	N50.47.08	E010.07.72
557	4/3/2014	16:47:43	1472	1504	1501	1485	1501	1421	326	347	352	344	325	293	1582	11	83	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5668	N50.47.08	E010.07.72
558	4/3/2014	16:47:49	1478	1504	1507	1485	1501	1421	326	347	352	344	325	293	1582	11	86	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5665	N50.46.83	E010.07.12
559	4/3/2014	16:47:55	1478	1504	1507	1485	1501	1421	326	347	352	344	325	293	1582	11	86	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5665	N50.46.83	E010.07.12
560	4/3/2014	16:48:01	1478	1504	1507	1485	1501	1421	326	347	352	344	325	293	1582	13	86	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5671	N50.46.58	E010.06.51
561	4/3/2014	16:48:07	1478	1504	1507	1485	1501	1428	326	347	355	344	325	293	1582	13	79	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5674	N50.46.58	E010.06.51
562	4/3/2014	16:48:13	1478	1504	1507	1485	1507	1428	326	350	355	344	325	293	1588	13	79	0	30.4	2178	63	9.2	NA	48	13.8	187	102.0	NA	137	5671	N50.46.34	E010.05.91
563	4/3/2014	16:48:19	1478	1504	1507	1485	1507	1428	326	350	355	344	325	293	1588	13	79	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5671	N50.46.34	E010.05.91
564	4/3/2014	16:48:25	1478	1504	1507	1485	1507	1428	326	350	355	344	325	293	1588	13	79	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5673	N50.46.09	E010.05.31
565	4/3/2014	16:48:31	1478			1485	1507	1428	326	350	355	344	325	293	1588	13	79	0	30.4	2178	63	8.8	NA	48	13.8	187	102.0	NA	137	5668	N50.46.09	E010.05.31
566	4/3/2014	16:48:37	1478							350	355	344	325	293	1588	13	85	0	30.4	2178	63	9.0	NA	48	13.8	187	102.0	NA	137	5670	N50.45.84	E010.04.71
567	4/3/2014	16:48:43	1478												1588	13	90	0	30.4	2178	63	9.2	NA	48	13.8	187	102.0	NA	137	5673	N50.45.84	E010.04.71
568	4/3/2014	16:48:49	1478																30.4	2178	63	8.9	NA	48	13.8	187	102.0	NA	137	5663	N50.45.60	E010.04.11
569	4/3/2014	16:48:55	1478																			9.2	NA	48	13.8	187	102.2	NA	137	5663	N50.45.60	E010.04.11
570	4/3/2014	16:49:01	1478																							187	102.2	NA	137	5667	N50.45.35	E010.03.51
571	4/3/2014	16:49:07	1478																										137	5669	N50.45.35	E010.03.51
572	4/3/2014	16:49:13																													N50.45.09	E010.02.91
573	4/3/2014	16:49:19																													N50.45.09	E010.02.91
574	4/3/2014	16:49:25																													N50.44.84	E010.02.31
575	4/3/2014	16:49:31																												5661	N50.44.84	E010.02.31
576	4/3/2014	16:49:37	1478	1505																										5659	N50.44.60	E010.01.71
577	4/3/2014	16:49:43	1478	1505	1507	1486	1501	1422																						5659	N50.44.60	E010.01.71



Democratise your data

Give people access

- Political?
- Physical?
- Semantic?
- At an appropriate granularity?
- In context?

This is not some high-minded principal… it enables “good enough” access for the people who make operational, tactical and strategic decisions and strips out IT complexity and time.

Data Engineering: putting it all together

Data engineering should be "good enough" to decide whether it is worth caring about, before investing in a more rigorous approach.

Text	Spatial, Chronological, logical	Relationships	Operationalisation
Technical	Resampling, profiling	Populations and outliers	Scaling
Measurement	Aggregate statistics	Behaviour and states	Impact

"Data only has value when someone asks to use it"

- Create demand
- Facilitate access to the data

The data science challenge

- Give the context back to each measurement:
  - Context within population (data set scale problem)
  - Context across domains (use time and space dimensions)
- Let the data speak for itself
  - Use statistical techniques first
  - Apply domain expertise to validate and guide
- If we care about it then find a better way to enable access
- Always have a view to:
  - Business value
  - Operationalisation
  - Wider data domains

The Bigger picture

- Data Science approaches uncover:
  - patterns and trends in behaviour
  - Outliers in populations
- This leads to an understanding of why something is happening
- Once we have the “why”, we can drive optimisations:
  - What leads to effective drilling for a given formation and well plan?
  - Quantifying the repeatability of 4D seismic data to de-risk reservoir decisions
  - Where is hidden pay likely to be found in badly interpreted logs?