

# Data Quality Metrics

Status of the National Data Repository (NDR) work  
for Regulators - 2014-2017

Philip Lesslar

Digital Energy Journal Conference  
3<sup>rd</sup> October 2017  
Kuala Lumpur

# Summary of **NDR 2017**, 6-8 June, Stavanger, Norway

Sponsored by:



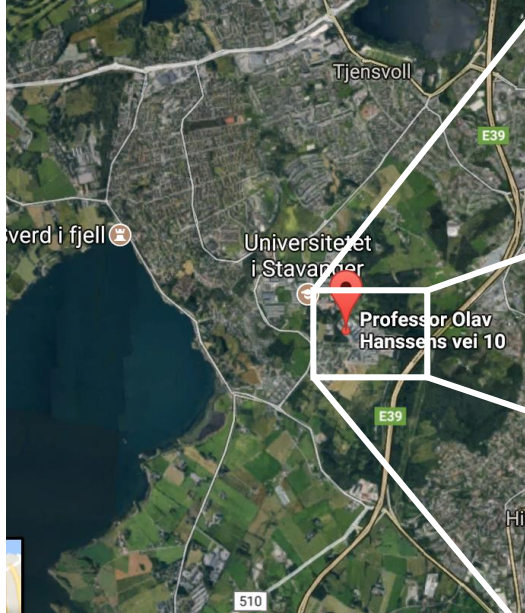
NORWEGIAN PETROLEUM DIRECTORATE



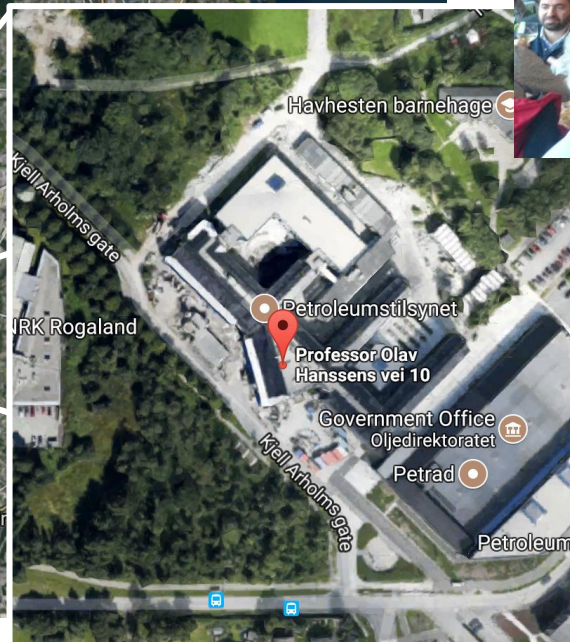
Organized by:



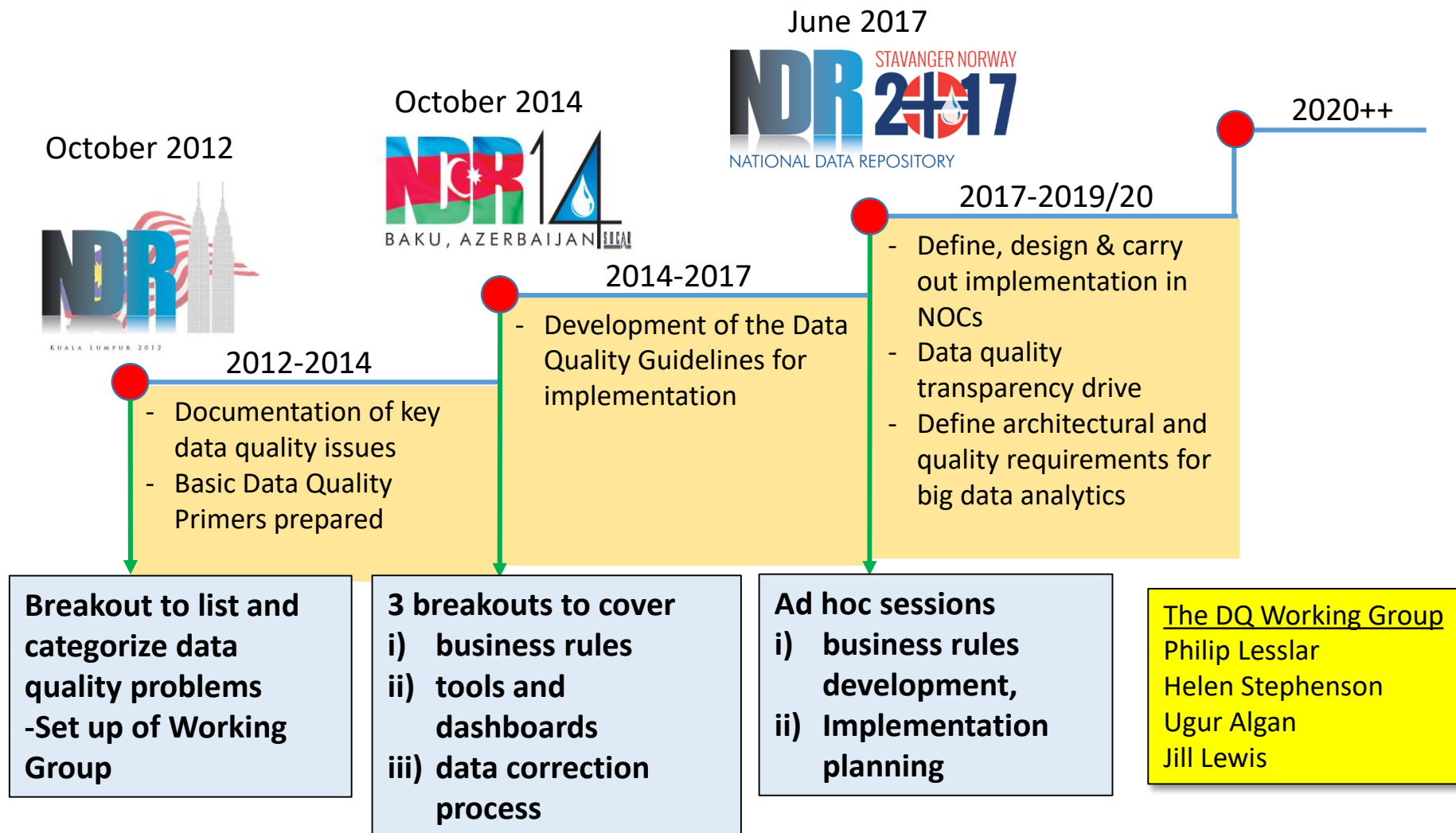
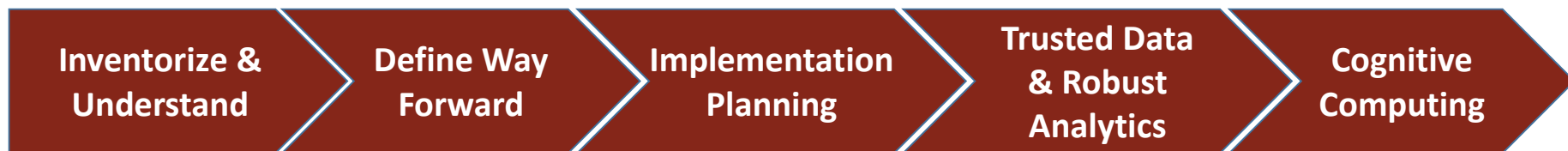
- 165 participants
- Representing 30 countries
- 20 sponsors



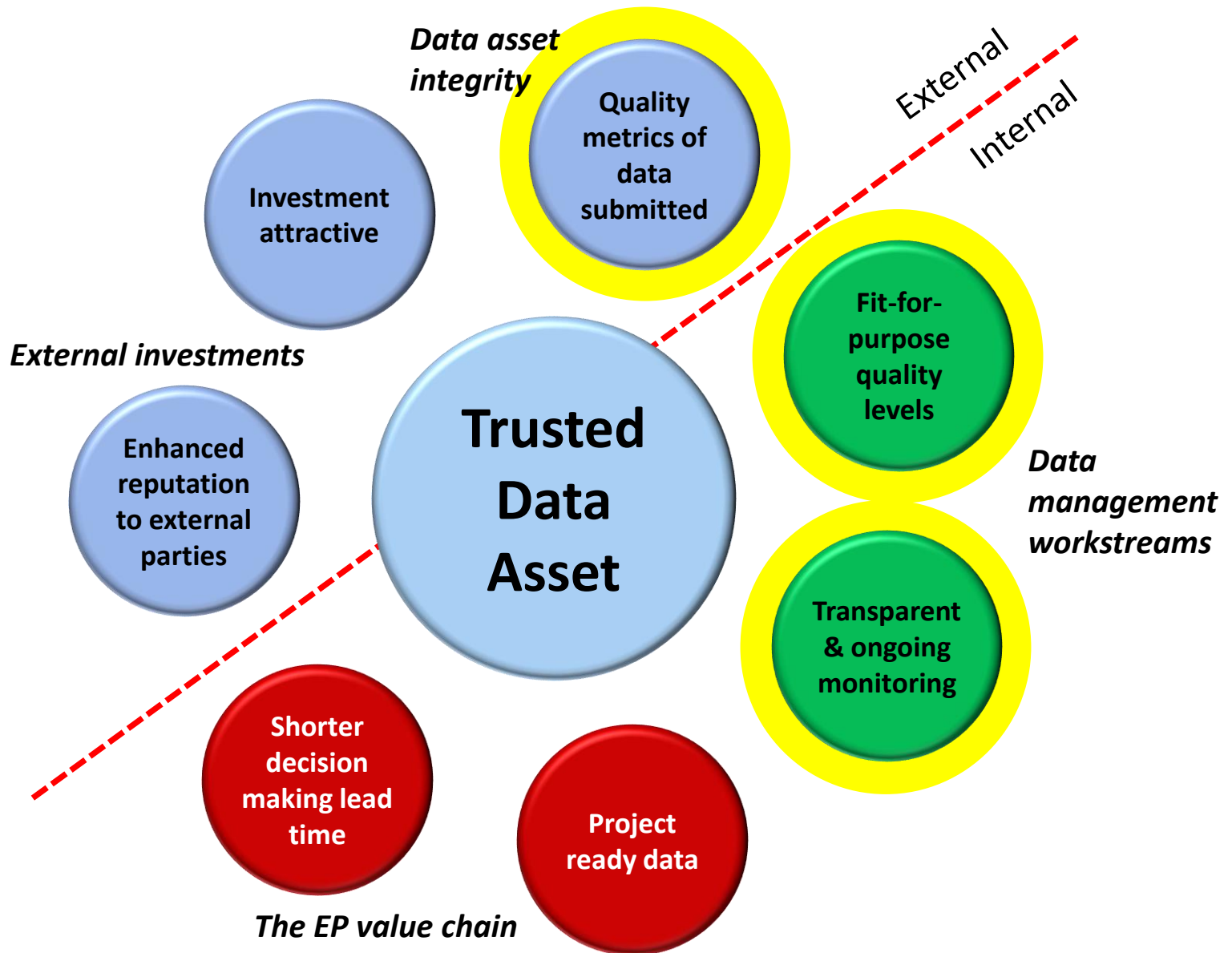
Google Maps



# Data Quality Metrics – The Journey



# Data Quality in the context of NDRs





# NDR11 Data Quality Workgroup Team Members

	Name
1	Helen Stephenson
2	Andrew Ochan
3	Marco Cota
4	Fanny Herawati
5	Melissa Amstelveen
6	Sarah Spinoccia
7	Jess Kozman
8	Ugur Algan
9	Richard Wylde
10	Cyril Dzreke
11	Lim TeckHuat
12	Hairel Dean
13	Deano Maling
14	Choo Chuan Heng
15	Iman Al-Farsi
16	Jill Lewis
17	Henri Blondelle
18	Armando Gomez
19	Kapil Joneja
20	Giuseppe Vitobello
21	Gareth Wright
22	Chan Kok Wah
23	Ali Alyahyaee (Scribe)
24	Philip Lesslar (Facilitator)



# NDR2014 Data Quality Workgroup Team Members

## 2 TOOLS & DASHBOARDS

1	Kapil Jonjega
2	Jack Walten
3	Johanda du Toit
4	Gustavo Tinoco
5	Ferdinand Aniwa
6	David Atta-Peters
7	Angus Craig
8	Natalia Rakhmanina
9	Glab Khanuntin
10	Edem Mawuko
11	Alexander Kosolapov
12	Daniel Arthur
13	Eric Toogood
14	Tatiana Vassilieva
15	Henri Blondelle
16	Marianne Hansen
17	Jill Lewis
18	Mikhail Leypunsky
19	Aygun Mamedova
20	Rena Huseyn-zade
21	Irada Huseynova
22	Philip Lesslar



## 1 BUSINESS RULES

1	Helen Stephenson
2	Richard Salway
3	Abraham Oseng
4	Malcolm Flowers
5	Uffe Larsen
6	Calisto Nhatugues
7	Sylvester Nguessan
8	Gianluca Monachese
9	Jan Adolfssen
10	Lee Allison

## 3 DATA CORRECTION WORKFLOW

1	Mehman Yusufov
2	Vahid Jafarov
3	Aleksa Shchorlich
4	Ngwako Maguai
5	Joseph Justin Soosai
6	Samit Sencurta
7	Julian Pickering
8	Ugur Algan

# What we did : 2014-2017

## The National Data Repository Data Quality Metrics Workstream

### Part 1: Background and Case for Change

Prepared by:  
The NDR Data Quality Working Group

Philip Lesslar  
Helen Stephenson  
Ugur Algan  
Jill Lewis



### Part 1: Background and Case for Change

Context and justification  
to Management

## The National Data Repository Data Quality Metrics Workstream

### Part 2: Business Rules Fundamentals

Prepared by:  
The NDR Data Quality Working Group

Philip Lesslar  
Helen Stephenson  
Ugur Algan  
Jill Lewis



### Part 2: Business Rules Fundamentals

Data quality dimensions,  
key concepts around  
business rules, 18 data  
types, 241 rules

## The National Data Repository Data Quality Metrics Workstream

### Part 3 : Implementation

Prepared by:  
The NDR Data Quality Working Group

Philip Lesslar  
Helen Stephenson  
Ugur Algan  
Jill Lewis



### Part 3: Implementation

Metrics, dashboards,  
implementing rules as  
queries, understanding  
results, getting the  
program going

# Why implement data quality metrics?

- Without metrics, we cannot measure the quality of the data we have
- Consequently, we cannot show how much quality, fit-for-purpose data there is...



Streamlines the business and its workflows

Increases data asset value and investor confidence

Builds essential data condition for effective use of new technologies

Perspective

Business

Enabler for improving data efficiency by up to 90%

NDR

Data Management

Data Science & Analytics



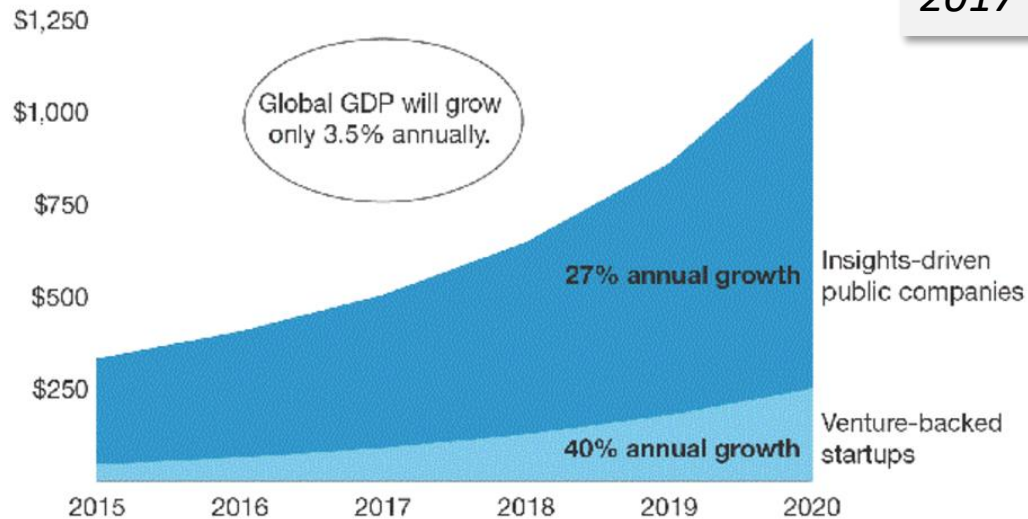
# Investment Trends

FORRESTER RESEARCH

## Insights-Driven Businesses Will Steal \$1.2 Trillion Annually By 2020

*Predictions 2017: Artificial Intelligence Will Drive The Insights Revolution*

Revenue forecast of insights-driven businesses  
(\$ billions)



Across all businesses, there will be a *greater than 300% increase in investment in artificial intelligence in 2017 compared with 2016.*

Note: The data point for public companies in 2015 is actual revenue; all other data points shown are estimates or projected figures.

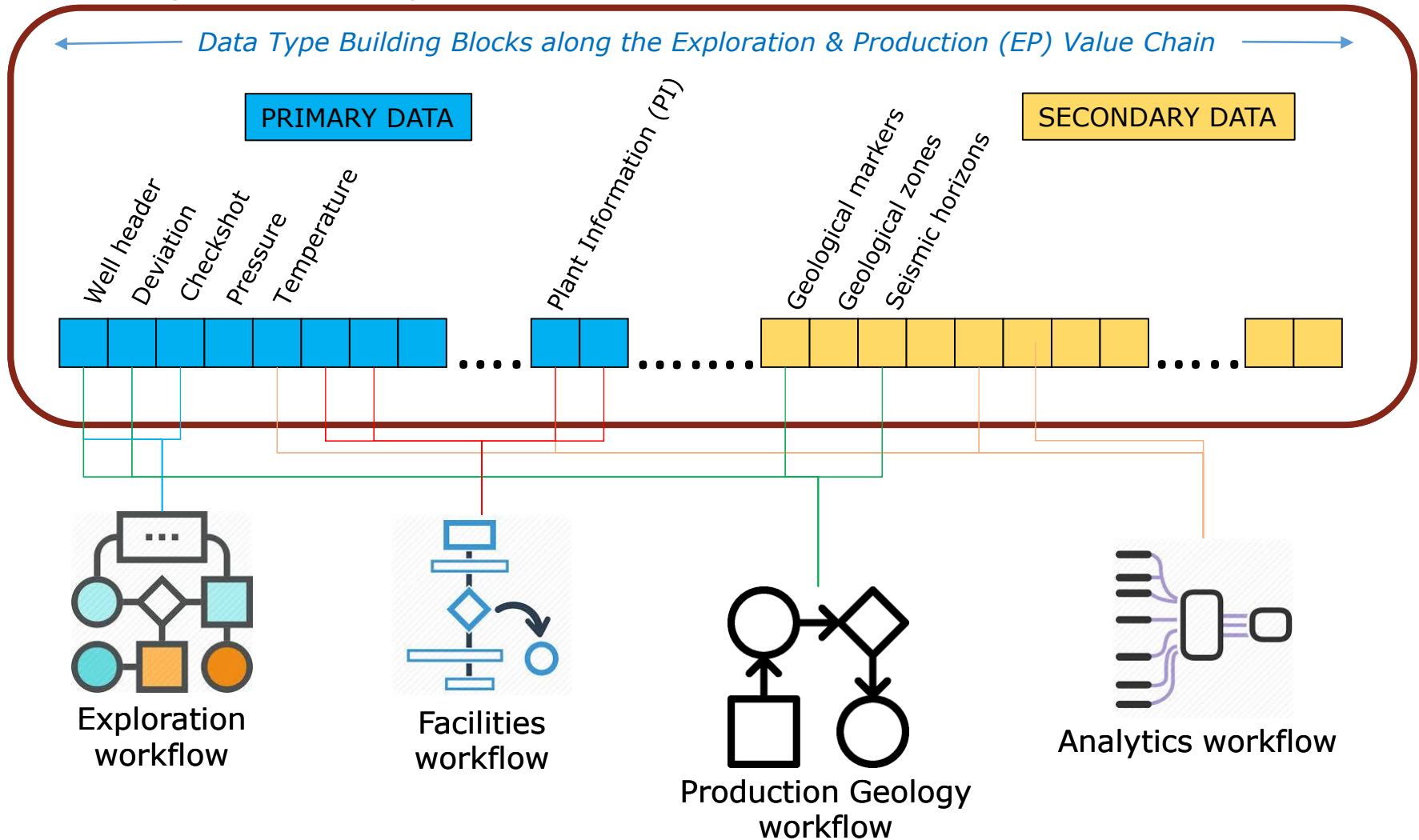
Source: Economic Intelligence Unit, Morningstar, and PitchBook Data

# Data Classification – Digital Data (>100 types in Upstream)

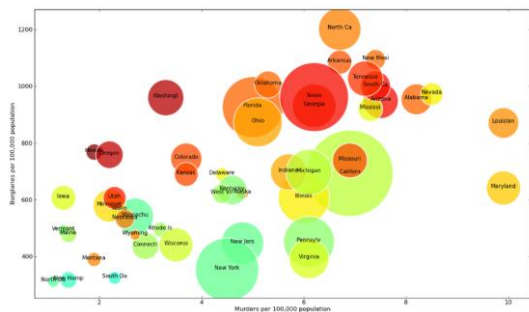
Primary Data ←			Secondary Data →	
Original Format Data	Reference Data/ Metadata	Master Data/ Corporate "Single Source of Truth"	Derived Data	Data Collections
Raw Seismic Raw Logs	Units of measure - <i>Linear measures</i> - <i>Pressure</i>	Static (hard) data - <i>Well header</i> - <i>Deviation</i> -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

# Packaging Quality Data – The Building Blocks

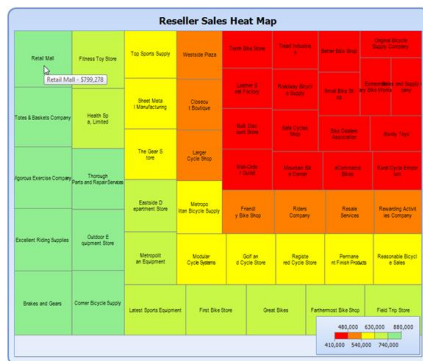
## Quality Data Envelope



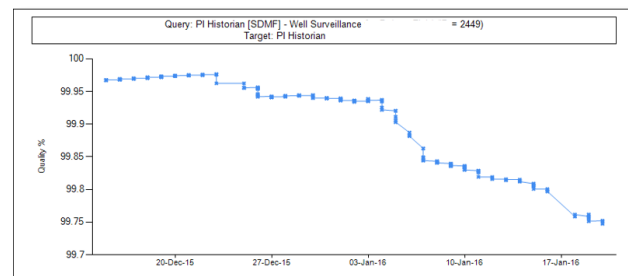
# Data Science / Analytics – Typical Deliverables



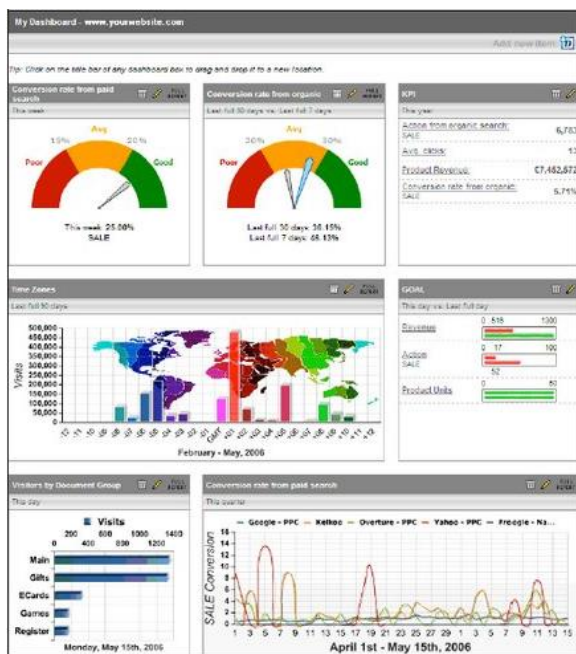
Bubble Plots



Heat Map



Trend Plot (IQM)



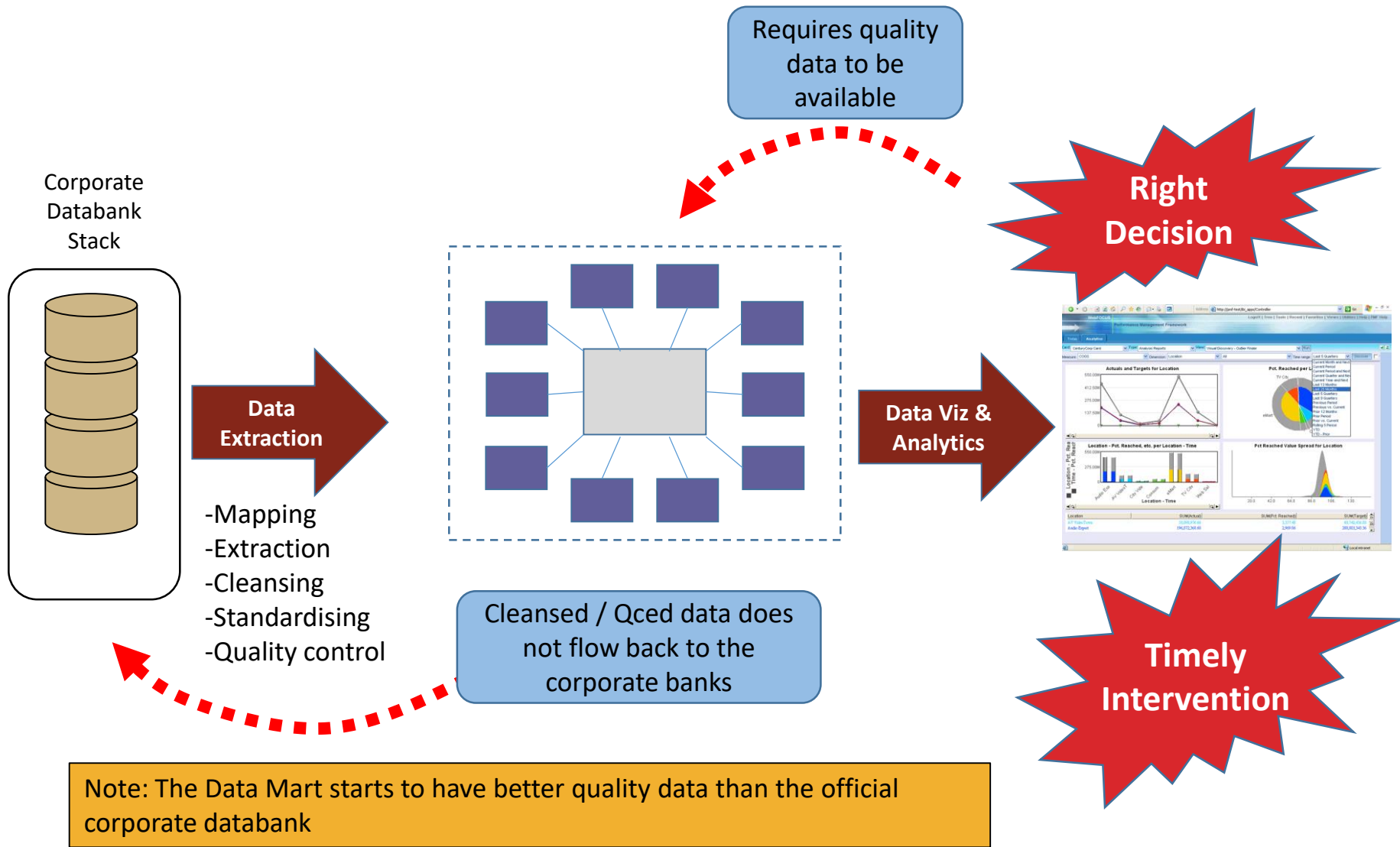
Yahoo Web Analytics



Google Analytics



# Data Analytics Conceptual Architecture



# Data Quality Error Persistence

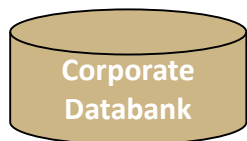
**Business Rule:**  
Well must have  
annulus pressure  
defined

Query ID	Query Name	Data Type	Quality Type	QD
2448	<a href="#">_SCE Test Data Approved By must be defined</a>	Pressure Analysis	Completeness	00.5000 00.0 00.001 ▲●●
2131	<a href="#">_SCE Test Data must have Function Test defined</a>	Pressure Analysis	Completeness	00.5000 00.0000 00.00 ▼▲▲
2127	<a href="#">_SCE Test Data must have Inflow Failure Mode defined</a>	Pressure Analysis	Completeness	00.5000 00.0010 00.00 ▲▲▲
2134	<a href="#">_SCE Test Data Positive or Inflow Acceptable Leak Rate (psi/min) must be defined</a>	Pressure Analysis	Completeness	00.5000 00.0010 00.00 ▲▲▲
2135	<a href="#">_Well must have Annulus Pressure defined</a>	Pressure Analysis	Completeness	00.0000 00.0010 00.00 ▼▲▲
2130	<a href="#">_Well must have String Status defined</a>	Pressure Analysis	Completeness	00.0000 00.00 00.00 ▲▲▲
2129	<a href="#">_Well must have String Type defined</a>	Pressure Analysis	Completeness	00.0000 00.0 00.001 ▲▲▲

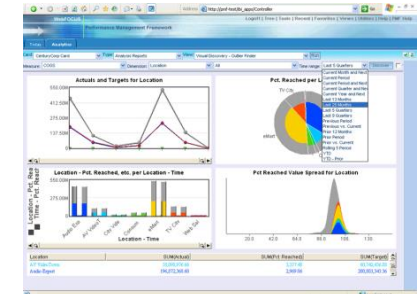
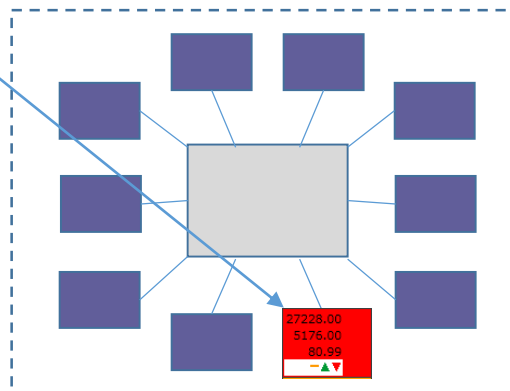
These errors will only be recognised if you are tracking the quality levels in the source databank

The analytics may not indicate quality levels

Eg. Annulus Pressure



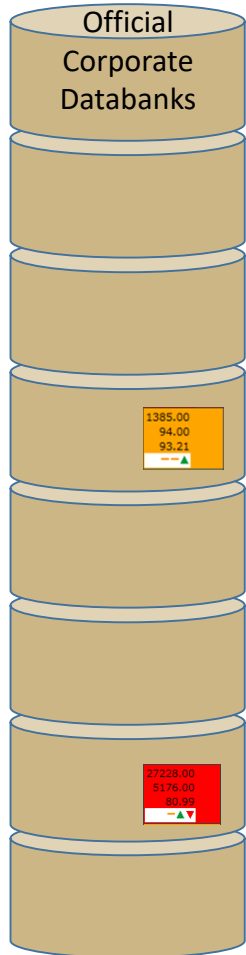
27228.00  
5176.00  
80.99  
▲●●



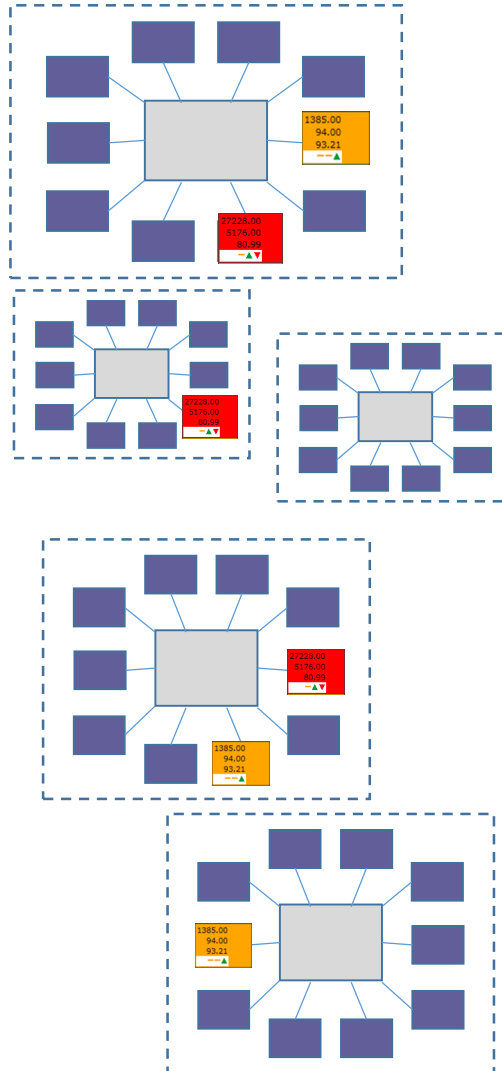
27228.00  
5176.00  
80.99  
▲●●

# Data Quality – Progressive Lopsidedness + Hidden Risks

Poorer Quality

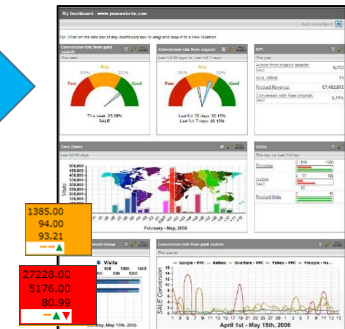


Better Quality



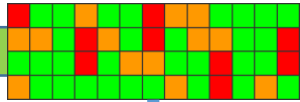
Right Decision?

Timely Intervention?

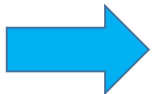
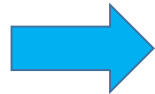
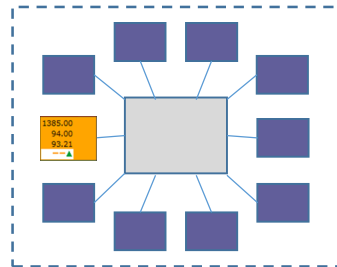
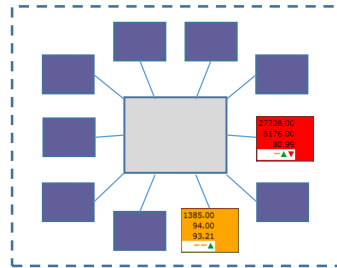
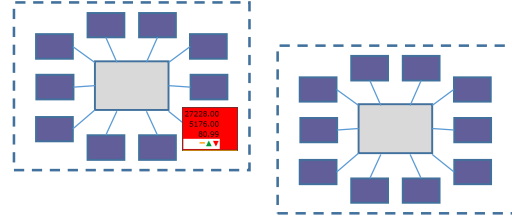
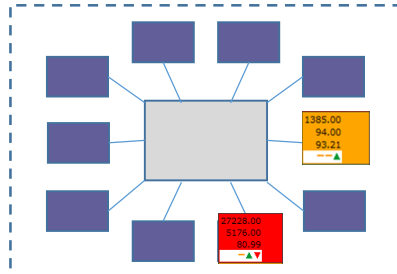
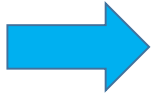
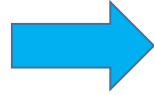
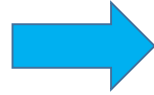
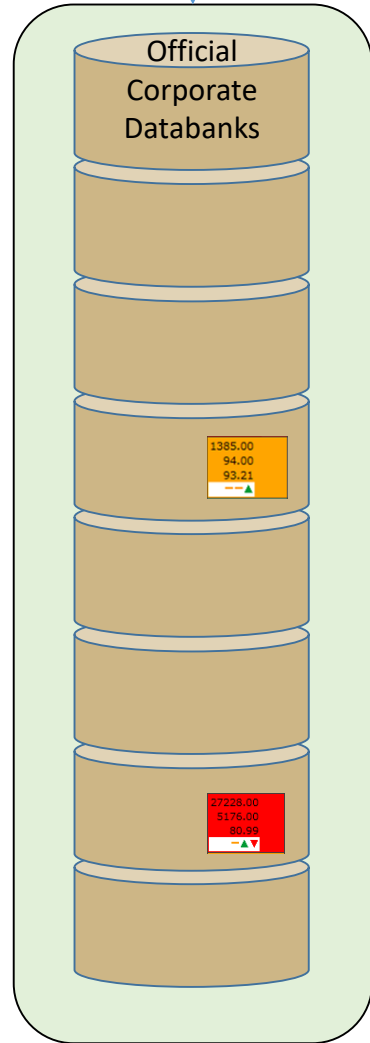


# Data Quality Metrics – Tackling Quality at the Source

Data Quality Metrics  
Dashboard

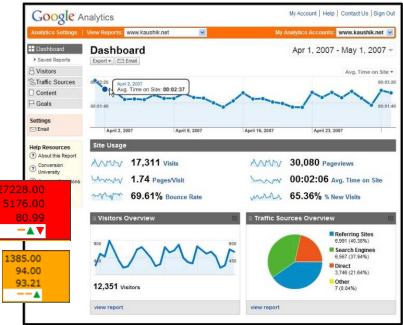
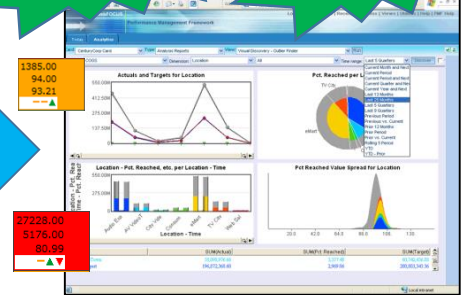


Quality throughout the life cycle



Right  
Decision

Timely  
Intervention



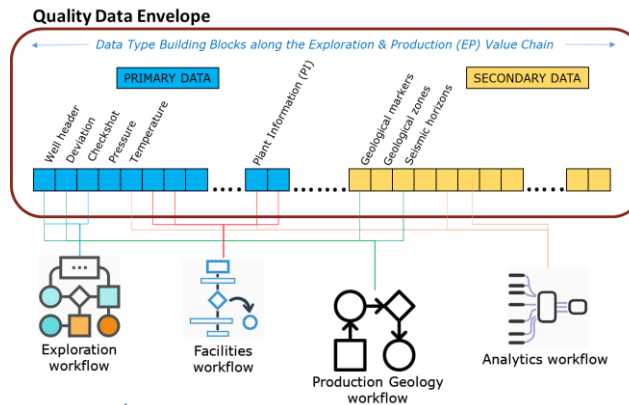
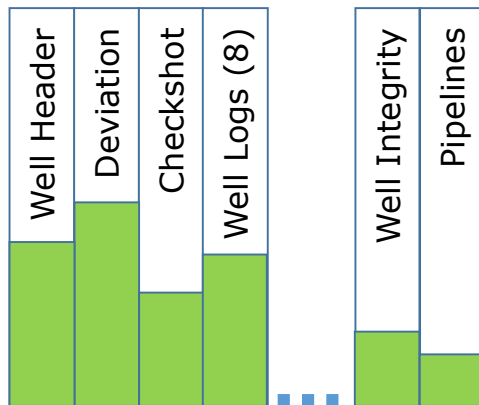


# Concluding Remarks

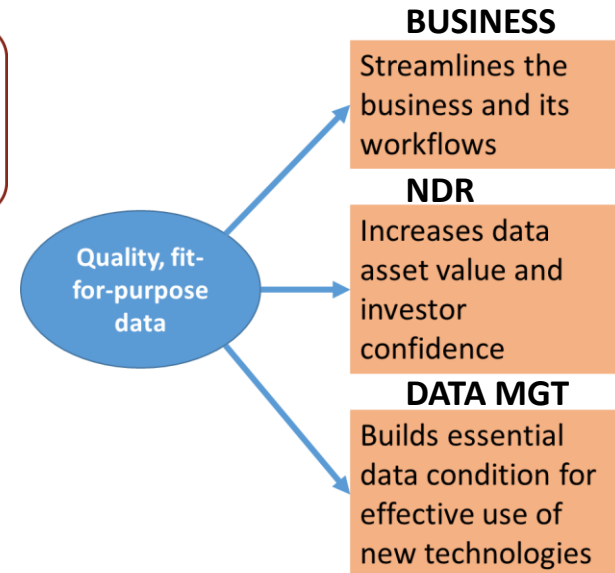
- Understand our DATA INVENTORY
- Implement METRICS to improve QUALITY

- Address data types as building blocks across all 100+ EP types

- We solve business problems and create new opportunities
- While measuring and knowing where we are at all times



- Measure and KNOW how much FIT-FOR-PURPOSE data there is



*Towards data science and big data analytics,  
by putting science into data management*

Thank You