

Geological Survey of Queensland

“Data for Discovery”

Unlocking the full value of geoscience
data to enable industry success

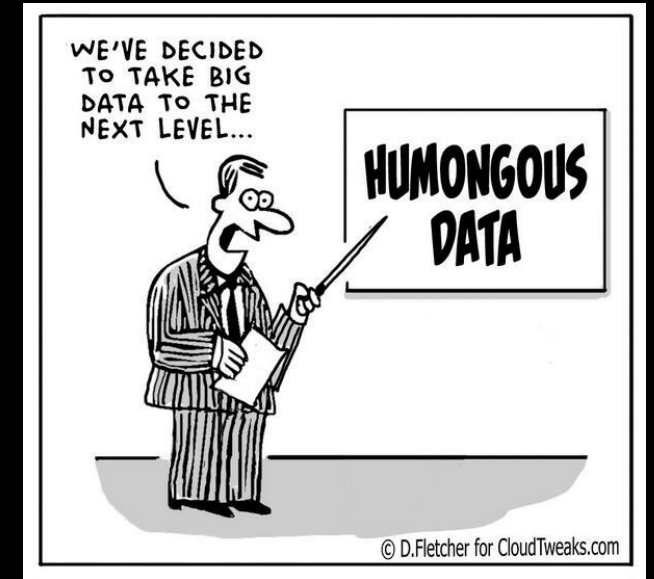


BIG Data Revolution - Global Shift

- Big Data Revolution – fuelling a global change
- We are at the beginning of a new information age
 - Driven by;
 - Cheap internet connectable sensor technology
 - Massive connectivity jump – internet expansion 2005 (1B users) – 2018 (4B users)
 - Organisational culture change to more open sharing of data and information

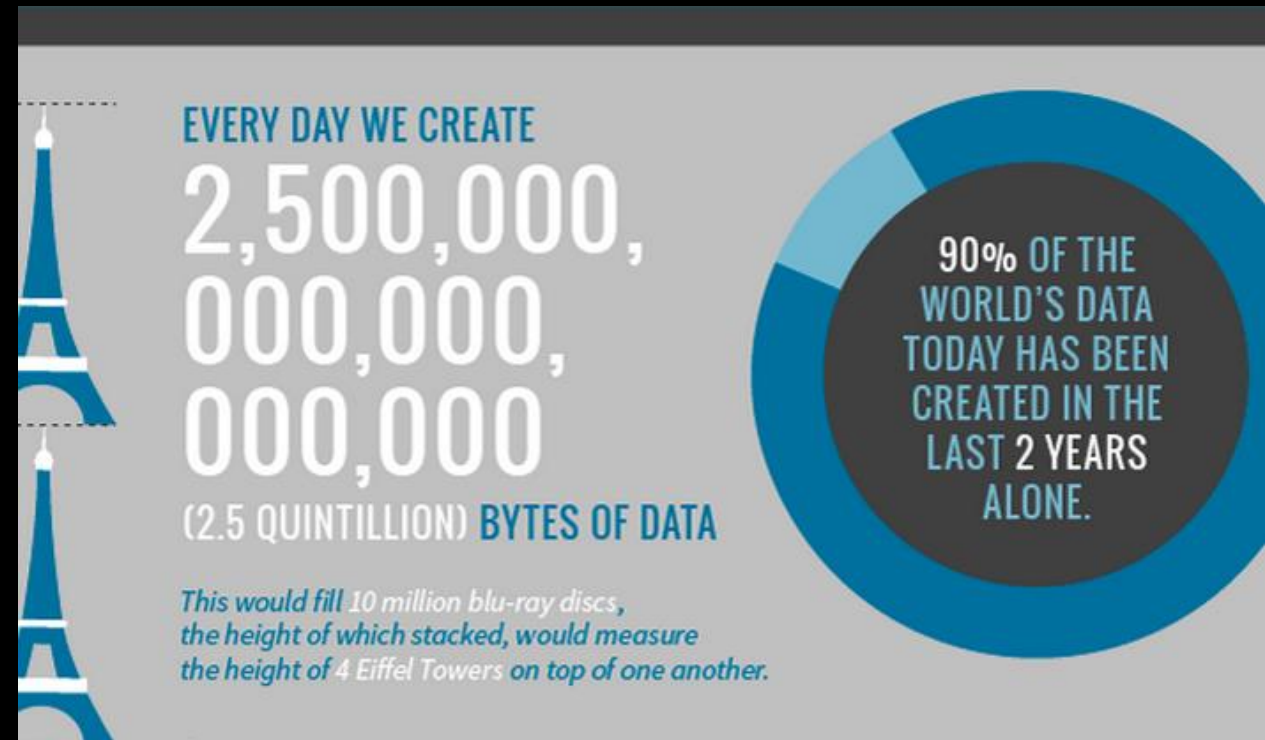
Resulted in access to massive amounts of data –

Big Data!



Big Data Revolution – Some Stats

- 90% of the world's data has been created in the last two years alone.
- Most companies only analyse 12% of the data they have.
- By 2020, there will be more than 50 billion smart connected devices in the world, collecting, analysing and sharing data.
- Bad data costs the US \$3.1 Trillion annually.
- IoT will save consumers and businesses \$1 trillion a year by 2022.



Big Data – The Insight

- Not just about collecting data for data's sake
- Large scale data access supports the capacity for large scale analysis
- It's not all about data resolution
- It's also about finding relationships between different data sources...



Analytics + Data allows us to see NEW patterns.....

Why is this important?

- Access to data and analytical techniques is driving significant improvements in many other domains;

Breast cancer affects 117 out of 1000 Queenslanders every year.

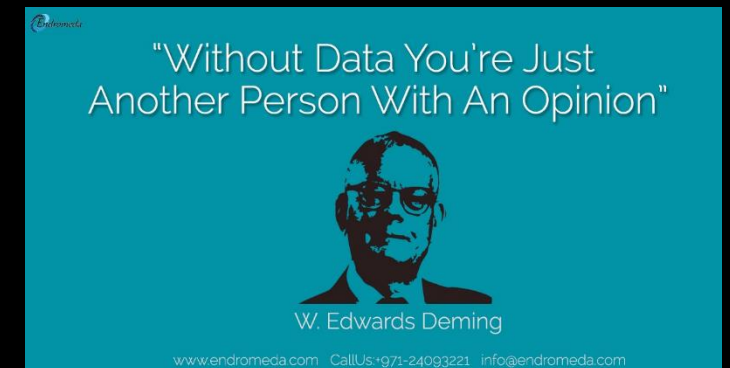
What if I told you we could improve detection rates a further 16% by simply using the same data more effectively...?



What if we could save 10% of Queensland Rail's \$1.1B annual operating cost by just collecting and using data more effectively...?

Some Examples

- Some of the massive opportunities in just using data more effectively...
 - Google Analytics has improved the breast cancer detection rate from 73% to 89% in the US - using biopsy data and machine learning
 - Norfolk Southern (an eastern US rail network provider) optimised it's network and it saved the company 200 million dollars per annum – using sensor data and machine learning.
- It's a big deal in other domains too...



Early Adopters

- Already being embraced in the geoscience world...
 - Using data and analytics GoldSpot Inc. was able to find 86% of the existing gold deposits in the Quebec Abitibi, but only needed 4% of the total surface area to do so.
 - Earth AI – developed machine learning software which generates precise exploration targets in previously unexplored areas.
- Access to both novel data sources and analytical power gives us;
 - Gives us the ability to explore more challenging areas – especially when the potential target is undercover.

Success requires smarter ways in not only redefining how we use data, but how we manage and curate it to ensure it can be used for geoscience problem solving.

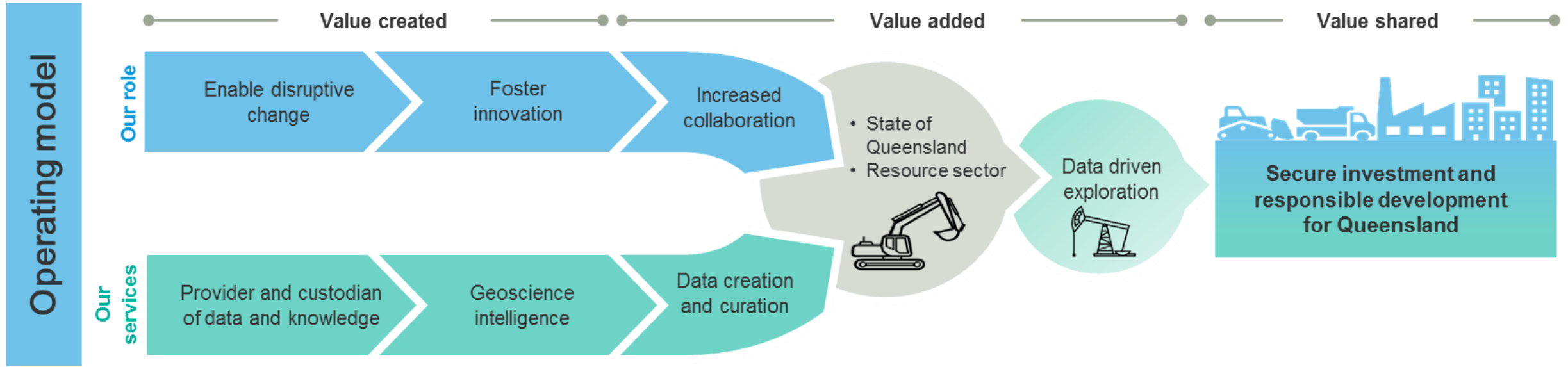
80% of the effort is in data management!

GDMP – Data for Discovery



GDMP will deliver a smarter way to manage and access GSQ custodial data.

GSQ's Strategic Direction



GSQ MISSION :

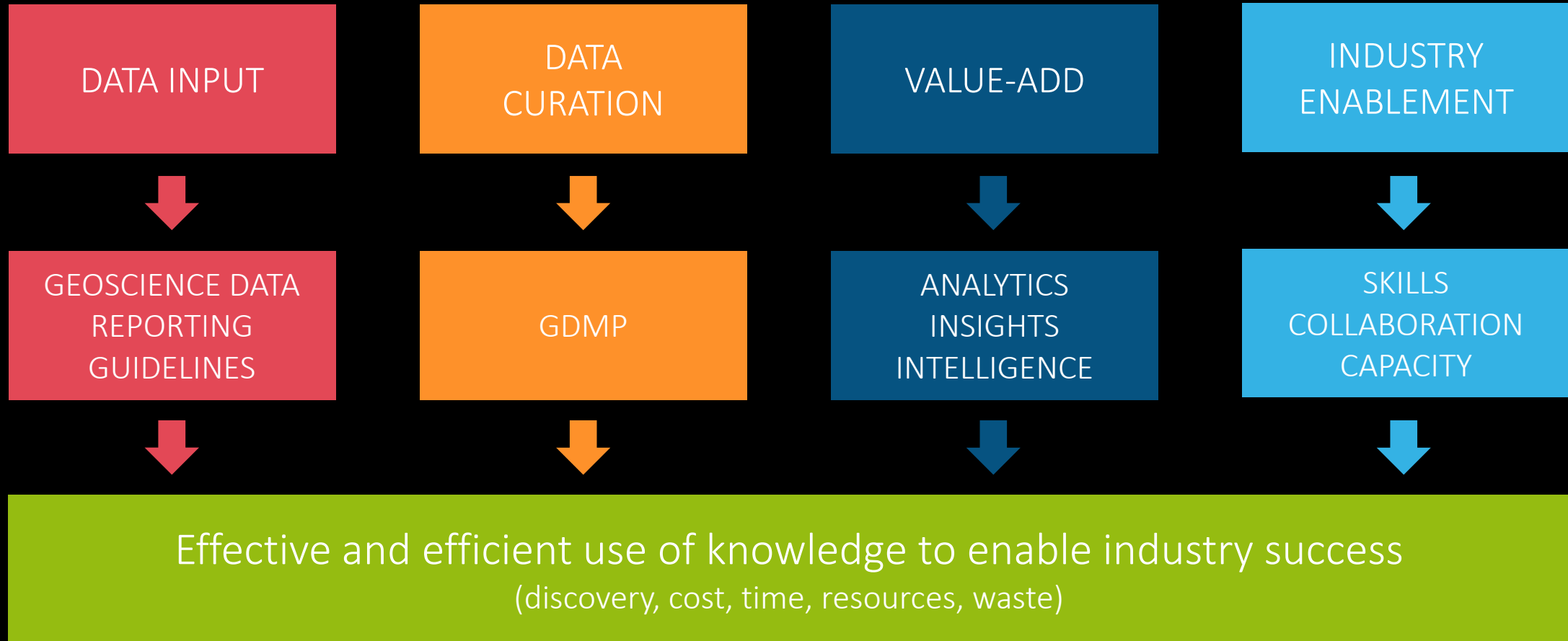
Focus our skills, technology and data to enable industry success.

GDMP – Data for Discovery



GDMP supports the GSQ vision to ‘enable industry success’ by making every possible piece of GSQ custodial data available.

Data-Driven Exploration



What are the current challenges?



DISCOVERABILITY

◁ It is difficult to find all the geoscience information that exists in an area

ACCESSIBILITY

◁ Industry can't access all of GSQ's geoscience data

QUALITY

◁ Data quality is highly variable

USABILITY

◁ We can't unlock the value of our data

COST

◁ Cost of maintaining multiple systems is unsustainable

Data Management Lifecycle

CURRENT STATE

Data unstructured
& locked in PDF



Poor data quality
with minimal QA



Multiple, fragmented
expensive data stores



Multiple data
catalogues



Low volume of data
available to industry



Manual extraction,
variable analysis



RECEIVE/
GENERATE

VALIDATE

STORE

CATALOGUE

DISCOVER

EXTRACT/
ANALYSE



Data itemised &
well structured



Data validated at
point of entry



Single, low-cost,
flexible data store



Single catalogue
of all data



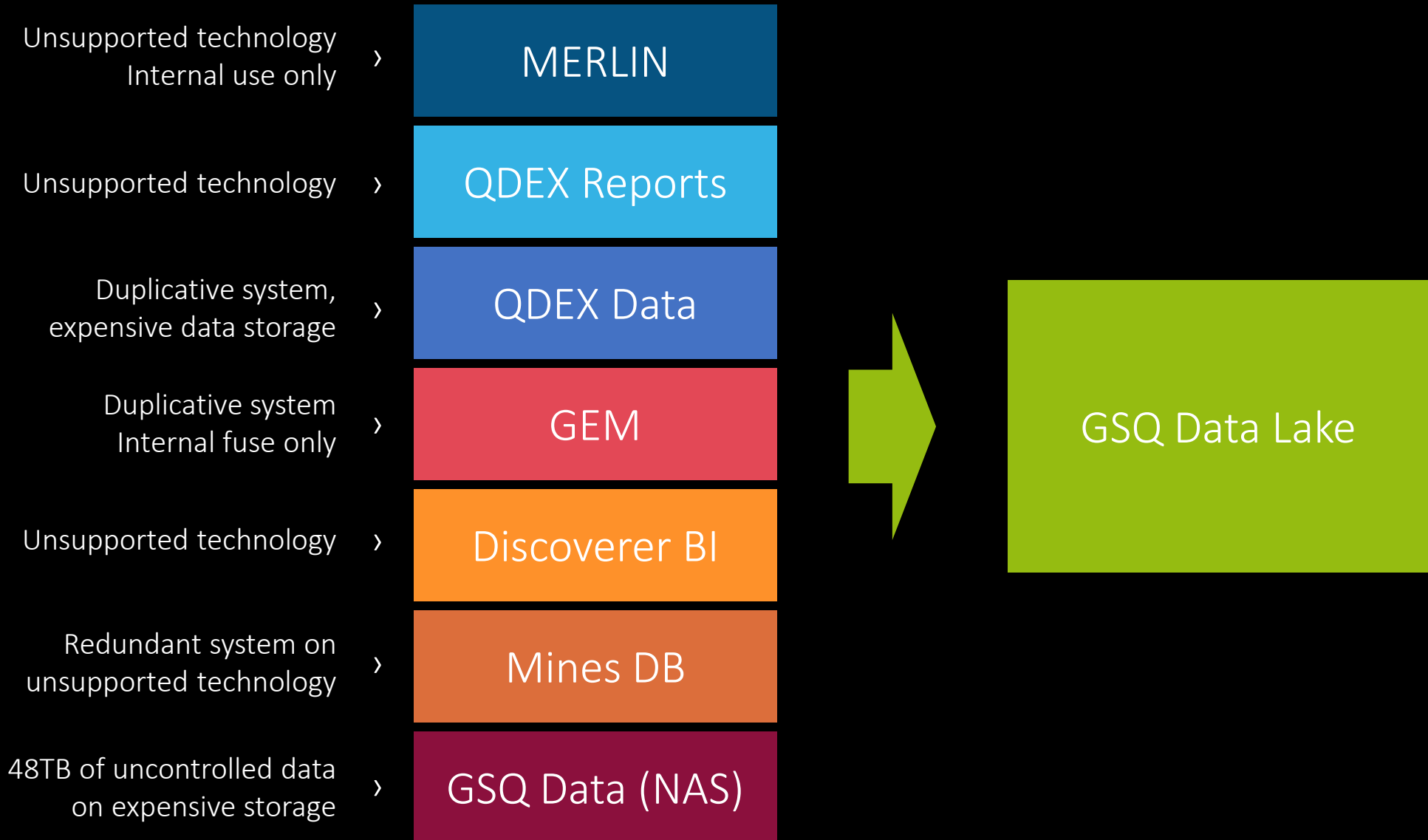
>90% of open data
available to industry



APIs, open integration,
advanced analytics

FUTURE STATE

System risk, cost, duplication



Data Lake Concepts



Find insights in data



Interact with the data via spatial, textual, graph, 3D



DATA VISUALISATION

DATA ANALYTICS

MACHINE LEARNING



AI that learns from data, and identifies patterns & insights

Optimise, enhance, cleanse & curate data



DATA PROCESSING

Index all digital & physical data



DATA CATALOGUE

DATA ACCESS



Human, computer & cloud data access

Store every piece of data as an object



DATA OBJECT STORE

Data Catalogue Concepts

DATA CATALOGUE

A single catalogue of all data – digital, physical, federated

STANDARDISED DATA SCHEMAS

- A DCAT2 master schema to describe all data objects
- Extensible geoscience data schemas
- Schema validations

CONTROLLED VOCABULARIES

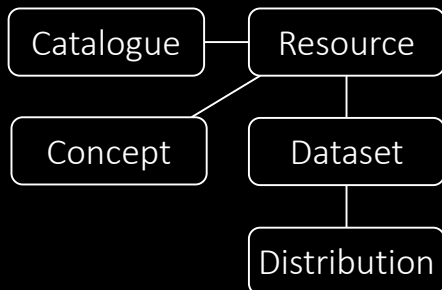
- A standard way to describe objects & their relationships
- Covers reference data, lookup lists, master data
- Preferred and variant terms

PERSISTENT IDENTIFIERS

- Globally unique identifiers
- Auto-provisioned e.g. IGSN for samples
- Alternate identifiers for historical data identities

LINKED DATA

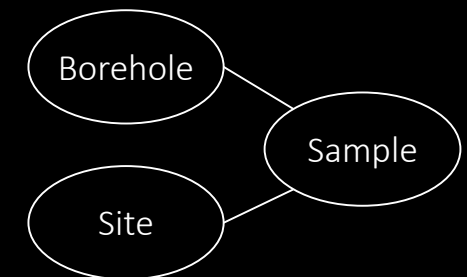
- Connects related data that wasn't previously linked
- Subject-predicate-object
- Human and machine-readable



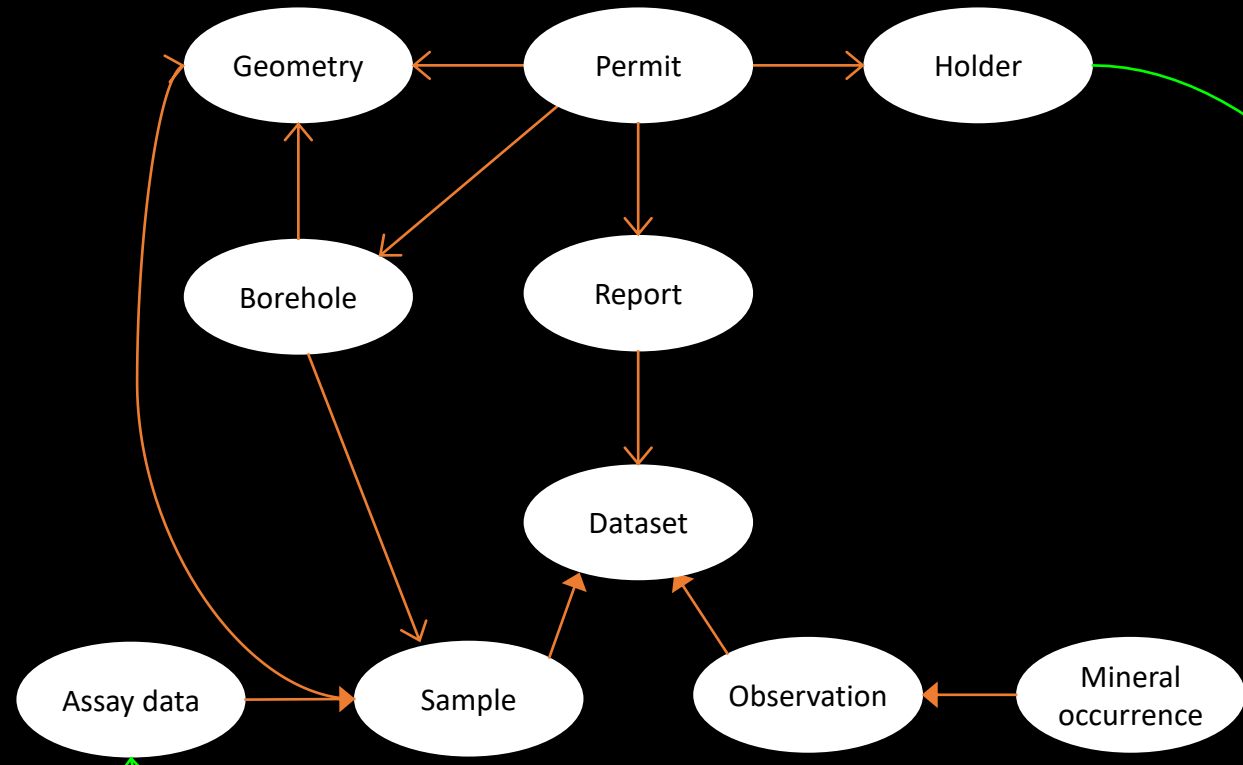
Vocabulary: Rock Types

- igneous
 - Adakite
 - Alkali feldspar granite
 - Basalt
 - 'A'a
 - Pahoehoe

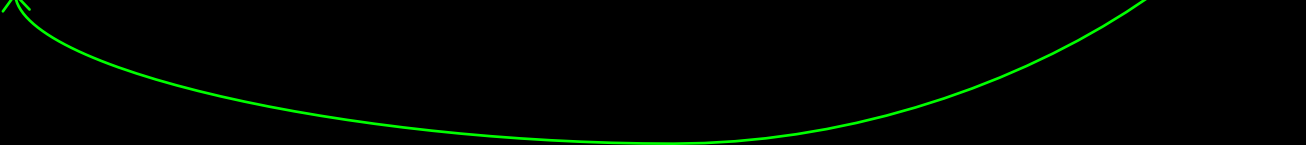
[//pid.geoscience.gov.au/sample/QG123](http://pid.geoscience.gov.au/sample/QG123)



Linked Data



We can infer relationships:
Show me all the assay data
from company x



GDMP Pilot

- GDMP industry pilot will be released June 30th.
- Focussed on seismic and geochemistry data
 - 800+ seismic surveys
 - Complete geochemistry database
- Users will be able to search for and download data
- Great feedback opportunity!

The screenshot displays the website for the Queensland Government Department of Natural Resources, Mines and Energy. The page features a navigation menu with links for Home, Datasets, Reports, Boreholes, Organisations, and About. A search bar is located in the top right corner. The main content area is divided into several sections:

- GSQ Data Lake Pilot:** A featured section with a large image of a rocky landscape and the text "Welcome to the Geological Survey of Queensland".
- Search data:** A search bar with the placeholder text "E.g. geophysics" and a search icon. Below it are "Popular tags" for AGSS, Airborne, and Electromagnetic.
- Quick links to datasets:** A grid of buttons for various data types: Airborne Geophysics, ASTER, Company Reports, Geochemistry, Geophysics, Gravity, Hyperspectral, Magnetotelluric, MnOcc, SEEBASE, Seismic Survey, Stratigraphic Drilling, and Surface Geology.
- GSQ Showcase Products:** A section with a red header containing two product listings:
 - Super ISA Basin Seismic Data Package:** An enhanced and optimised collection of seismic data for the Super ISA Basin. This data has been cleansed and compiled in extended 3D format.
 - Lawn Hill Cobalt Data Signatures Data Package:** A collection of geochemical, geophysical, magnetotelluric, and gravity data with corrected dimensionality showing potential cobalt signatures for Lawn Hill map area.
- GSQ Latest Datasets:** A section with a blue header listing three datasets:
 - 85003 BECKER 2D 1985:** DELHI PETROLEUM 2D Seismic Survey completed in 1985.
 - 95124 SQ01 NORTH NACCOWLAH 3D 2001:** SANTOS 3D Seismic Survey completed in 2001.
 - 95316 CP5AN12C CUISINIER NORTH 3D 2012:** SANTOS 3D Seismic Survey completed in 2012.

Pilot progress to date



EPM report
form prototype

LAS file
validator

Pilot data lake

CKAN data catalogue
Vocabulary manager

Data catalogue

FTP transfer
File gateway



These details
Authorised Holder
Resource Authority
Commodity
Project Name
Summary
Introduction

Automate Geoscience Data Validator (trial)

Upload a LAS file compliant to the LAS 2.0 standard.

This validator will check the metadata and borehole location.

Queensland Government
Department of Natural Resources, Mines and Energy

Datasets Organizations Groups About Search

Filter by location Clear

+ YOR PAP, N. GUI. VANAT N. CAL. AUSTRALIA

Map data © OpenStreetMap contributors
Tiles by Stamen Design (CC BY 3.0)

15 datasets found

Order by: Relevance

PRIVATE GeoTextPDF
Basic PDF with searchable words

AGSS 1359b
Airborne Electromagnetic and Magnetic survey

PRIVATE Land resource areas - Central Highlands
This data set is a Land Resource Areas map (1:500 000) of the central highlands. The central highlands covers 8 645 870 comprising the shires of Emerald, Bauhinia and the...

S3 Bucket

Why is GDMP different?



As-Is

Variety of geoscience subjects
Many different data structures
Diverse terminology
Data quality issues
Data trust issues
Multiple metadata repositories
Multiple data stores
Database primary keys, composite keys
Complex database for metadata & data
Closed system inhibits data sharing
Designed for humans

To-Be

→ Standard representation of concepts
→ Linked data structures – standards based
→ Controlled vocabularies
→ Data quality metrics, quality assurance
→ Data provenance recorded
→ One metadata repository
→ One data store
→ Universally unique persistent identifiers
→ Metadata in database, data in object store
→ Open system, federated search
→ Designed for humans and computers

Why is GDMP technically different?

Database/Data Warehouse

- Expensive data storage
- Technical skill to create data relationships
- Schema on write
- Vendor lock-in system
- Can store structured data

Data Lake

- Low-cost, high volume data storage
- Can infer data relationships
- Schema on read
- Mostly vendor-neutral system
- Can store structured, semi-structured & unstructured data

It's all about data discovery!



Every data object



Is described by metadata



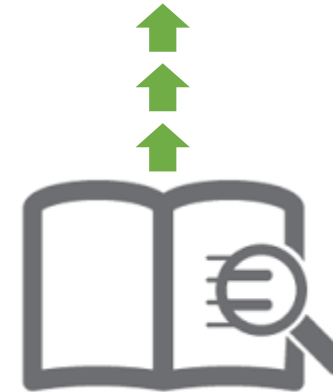
Metadata is validated



Against a controlled vocabulary



For easy data discovery



Resulting in a high quality data catalogue



And a data store with integrity

Opportunity statement



Optimise the geoscience data ecosystem
(people, process and technology)
to unlock the full value of data
to enable exploration success