Big data in the Geoscience: A portal to physical properties

Andrew Kingdon, Mark Fellgett and Martin Nayembil

British Geological Survey, Keyworth, Nottingham, NG12 5GG, aki@bgs.ac.uk

Geosciences were early adopters of both computing and digital data; the precursors of the SEG-D and SEG-Y geophysical formats date from as far back as 1967. Data standards, for seismic (SEG-Y, SEG-D) or geophysical log (LAS, DLIS) data simultaneously make interpretation and visualisation of data practicable but also their binary nature makes applying analytical techniques unusually complex. Specialist software is often required to process and interpret different datatypes.

Such problems are exacerbated by historic poor data management practices. Datasets are rarely collated at the end of projects or stored with sufficient metadata to accurately describe them and many strategically useful datasets reach BGS incomplete, unusable or inaccessible. Whether this situation arose through a lack of foresight about the future value of data, poor practise or simply storage space restrictions these problems pose huge challenges to today's geoscientists.

Consequently, there are major problems with applying big data analytics to geoscience. For example, many techniques don't sample geology directly but use proxies needing further interpretation. The use of analytical techniques have commonly been limited by the high proportion of noise incorporated into the datasets with very significant interpretation skills required to identify the signal. Thus far successful applications of "big data" analytics have been limited to closed systems or analyses of very common digital data types.

Significant problems remain, including the lack of data that can be immediately interacted with and difficulties in bringing together multiple datasets about related phenomena. Also the lack of adequate metadata about the data available to understand its context and scope and how to apply and qualify results. Whilst geosciences datasets have all the attributes of big data – volume, veracity, velocity, value and variety – the last two controls are disproportionately significant. The first of these determines the usefulness of the data and the second is the biggest impediment to delivering on the promises that big data offers especially in Earth Sciences.

In order to deliver a standardised platform of data from which individual geological attributes can be identified BGS has invested in the creation of PropBase (Kingdon et al., 2016). This single portal facilitates the collation of datasets supplied in standardised formats. This allows all data from a single point feature (e.g. boreholes) or areas of interest) E.G. to be extracted together in a common format allowing all data to be immediately compared. The existence of PropBase portal allows a researcher to answer the question "What's available at a location?" It has already been used in site characterisation for the UK GeoEnergy Observatories project.

Such initiatives that allow collation of high volumes of data in a single extractable format are a critical step forward to allowing Big Data analytics. Combined with the increasing availability and ever lowering cost of high power computing and analytical routines, the opportunities for big data analytics are ever growing. However, substantial challenges remain and new and more interactions with computer scientists are needed to deliver on this promise.

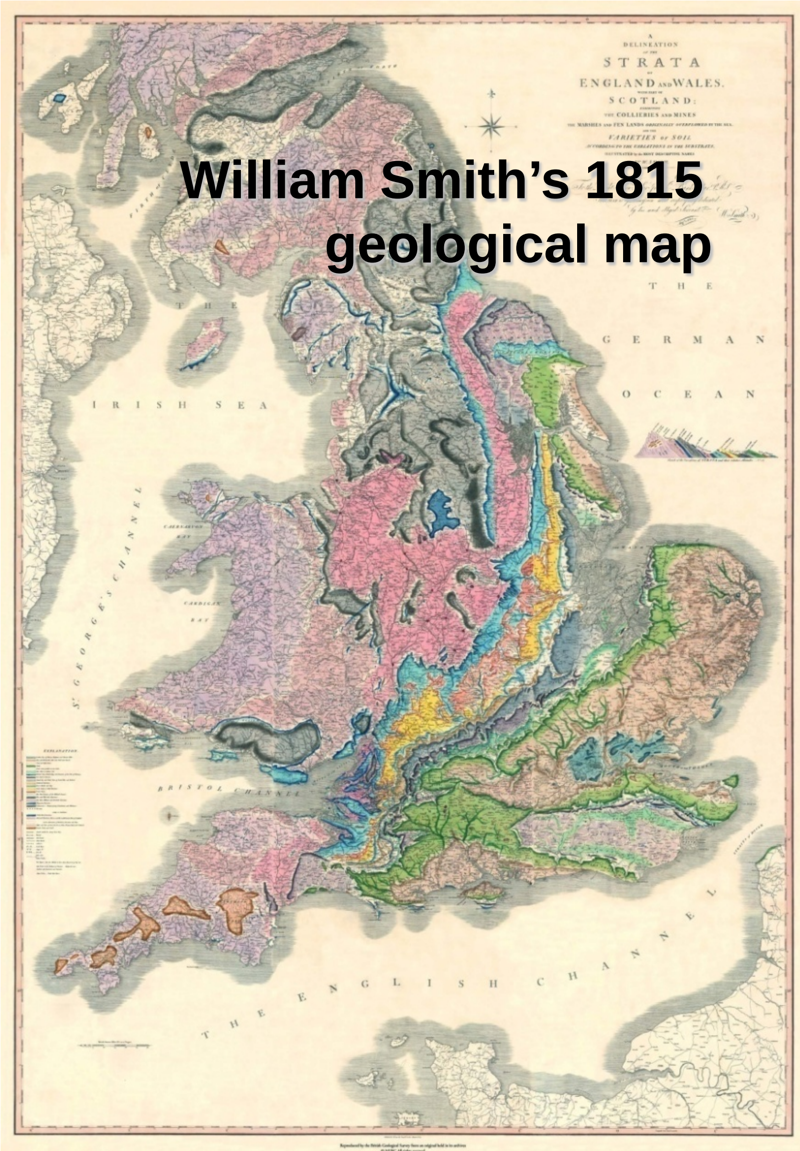
References

Kingdon A., Nayembil, M.L., Richardson A.E., and Smith, A.G., 2016. A geodata warehouse: Using denormalisation techniques as a tool for delivering spatially enabled integrated geological information to geologists. Computers and Geosciences, *96*, 87-97. DOI: 10.1016/j.cageo.2016.07.016.

Big data in the Geoscience: A portal to physical properties

Andrew Kingdon, Martin Nayembil and **Mark Fellgett**
British Geological Survey, Keyworth,
Nottingham, NG12 5GG,
aki@bgs.ac.uk
markf@bgs.ac.uk

Big Data and BGS



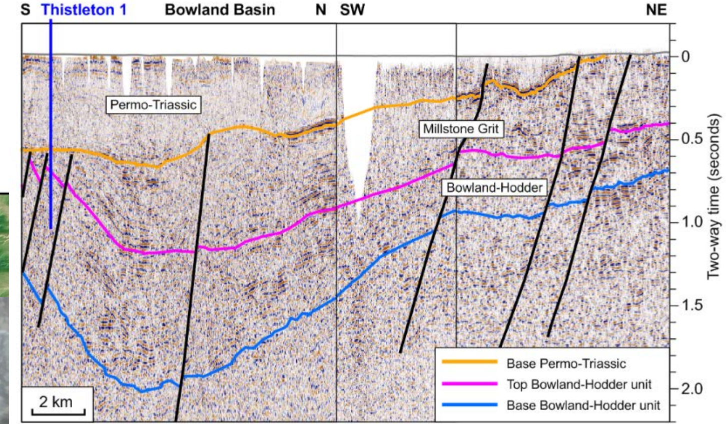
William Smith's 1815 geological map

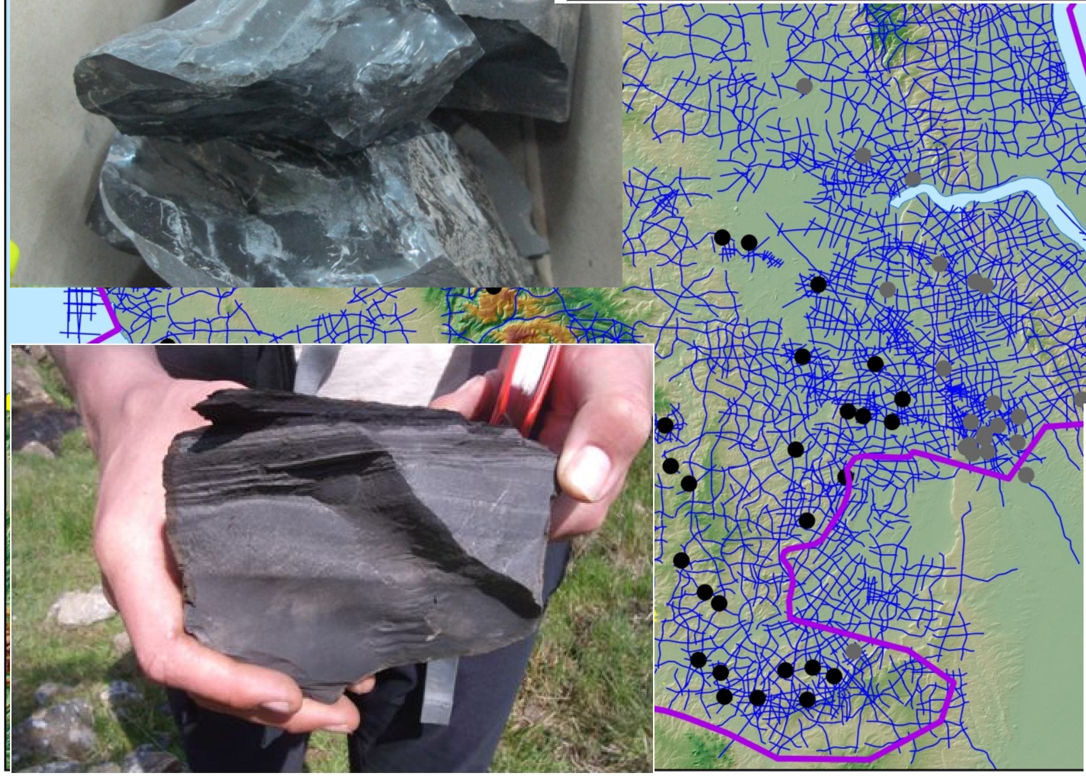
- British Geological Survey (BGS) ***advises*** *government, agencies & public about risks and resources for the UK landmass & UKCS*
- *BGS advice sought on ever more complex & controversial decisions*
- *Delivering this requires better data access, assimilation, analysis and visualisation*
- *Moving from "where is it" questions to "What if…" scenarios*
- *This is not a new challenge…*

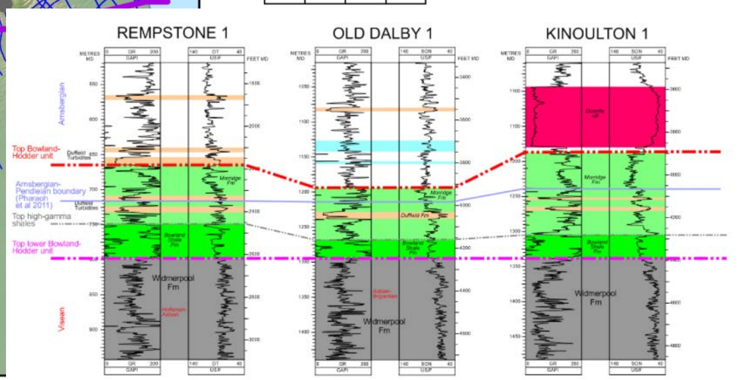
Example: UK shale Gas

- UK shale exploration controversial after 2011 tremors at Preese Hall
- UK government needed:
  - Location & size of reserves
  - Technical constraints on their production
- Typical of the complex questions BGS now has to answer

UK Shale evaluations





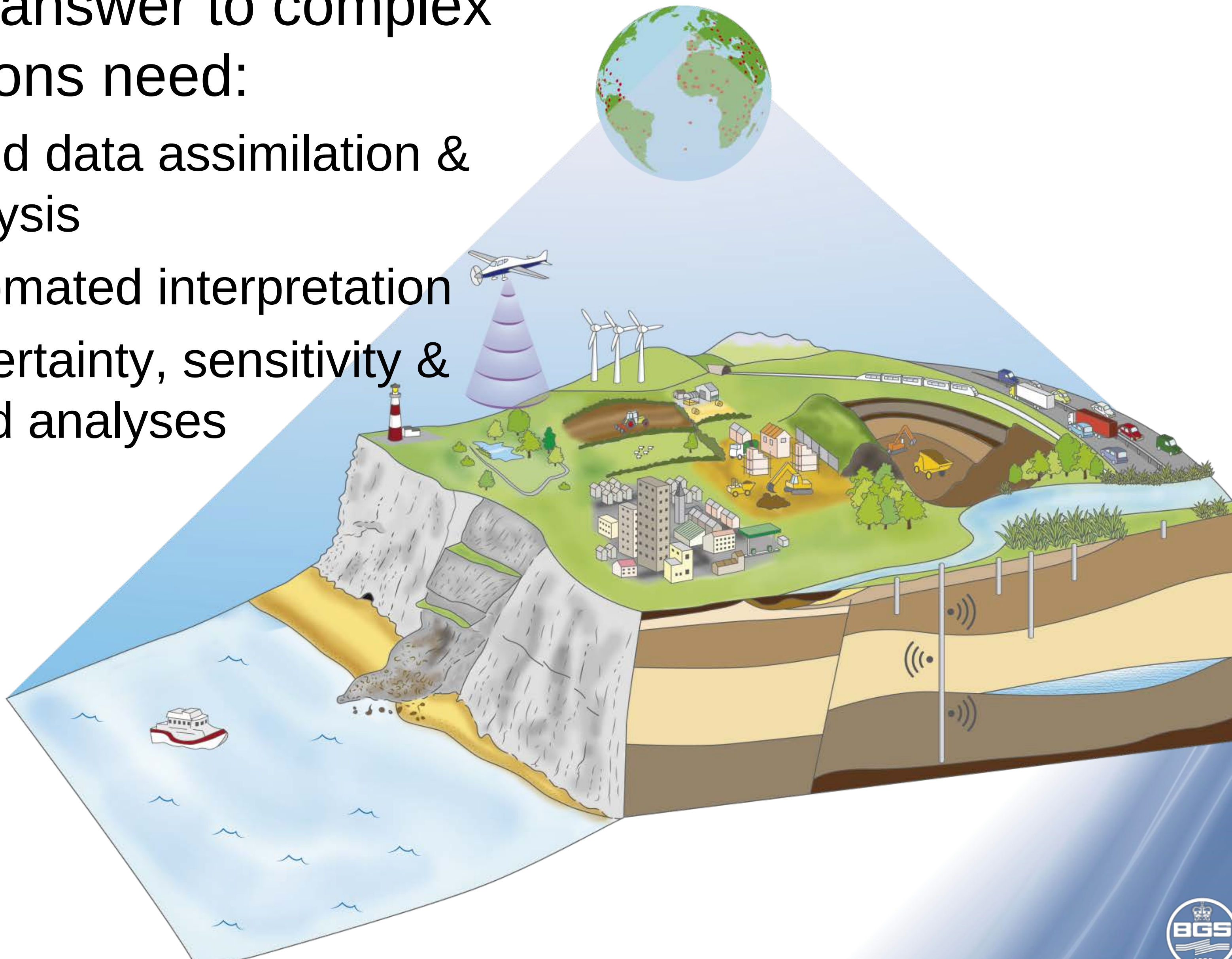


Outcrop studies, 64 key wells, 15,000 miles of seismic data

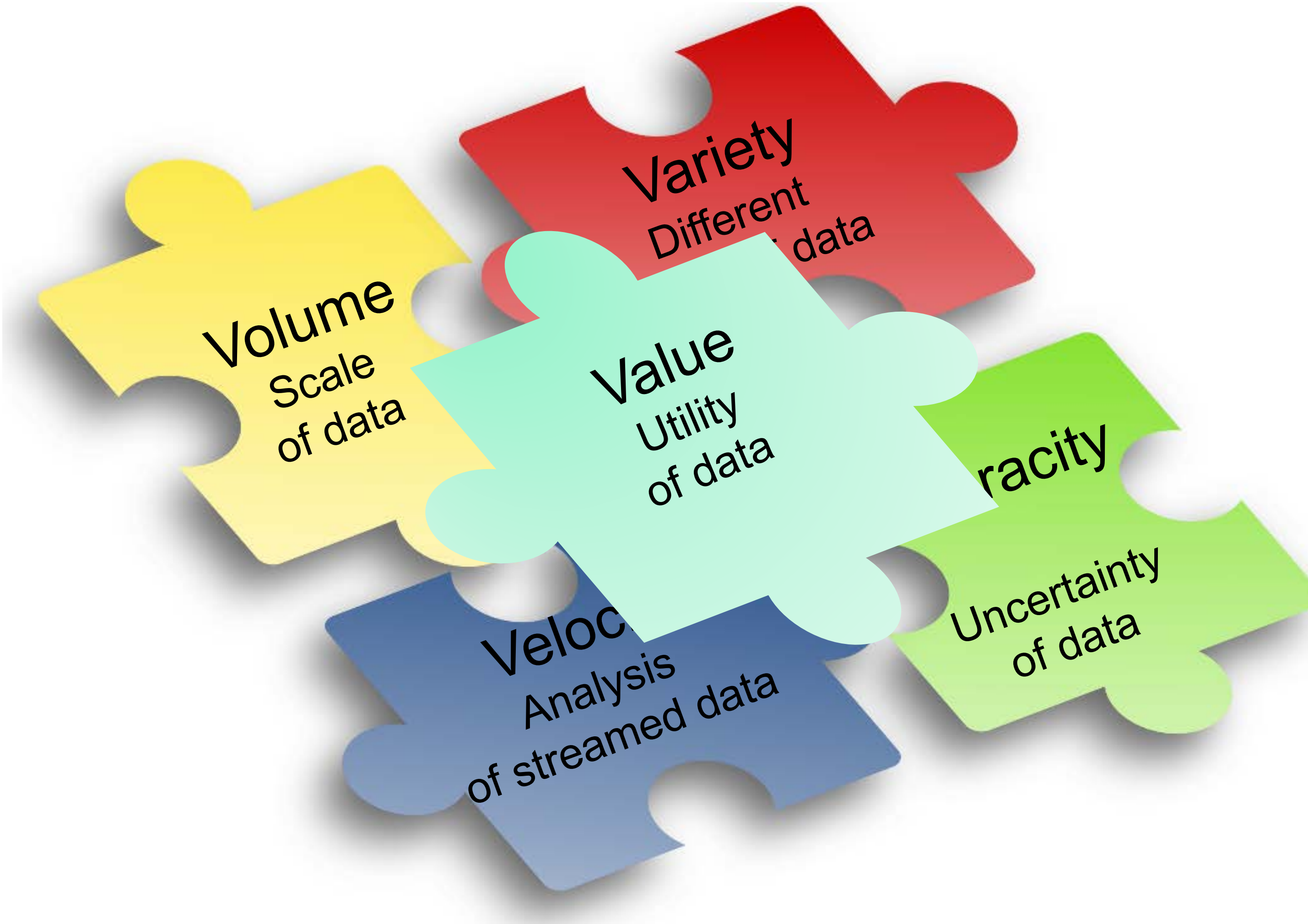
Images: BGS; Andrews, 2013

Opportunities for Geo "Big-Data"

- Quick answer to complex questions need:
  - Rapid data assimilation & analysis
  - Automated interpretation
  - Uncertainty, sensitivity & trend analyses

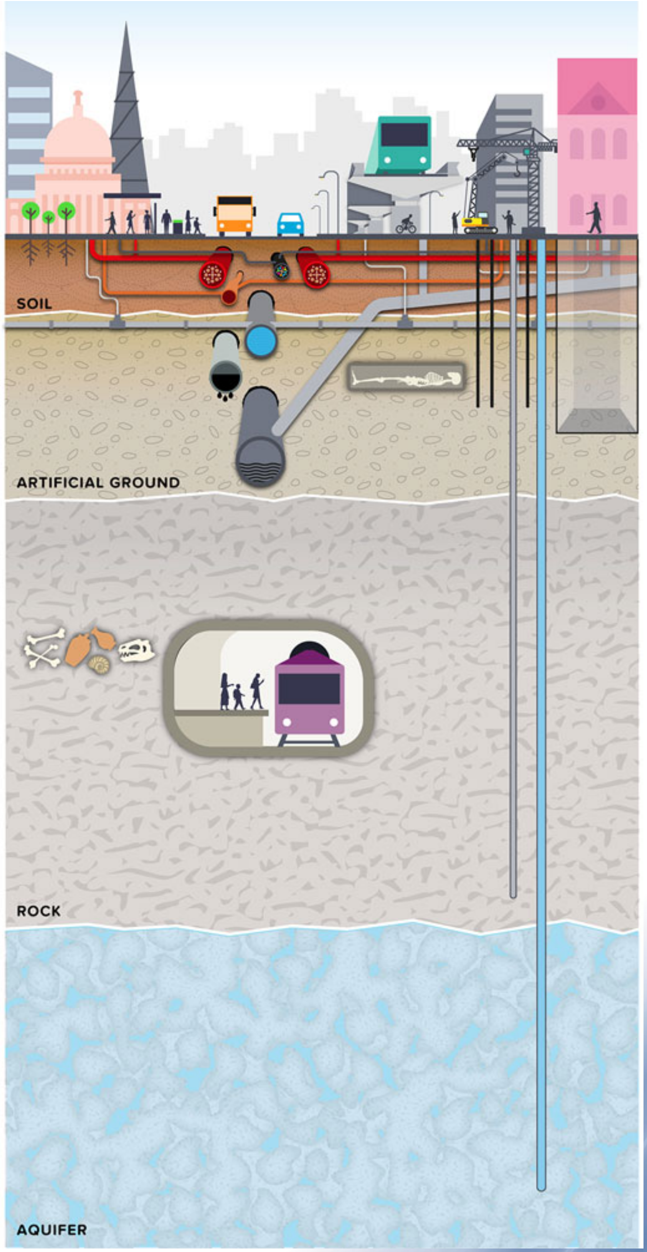


Definitions of Big Data



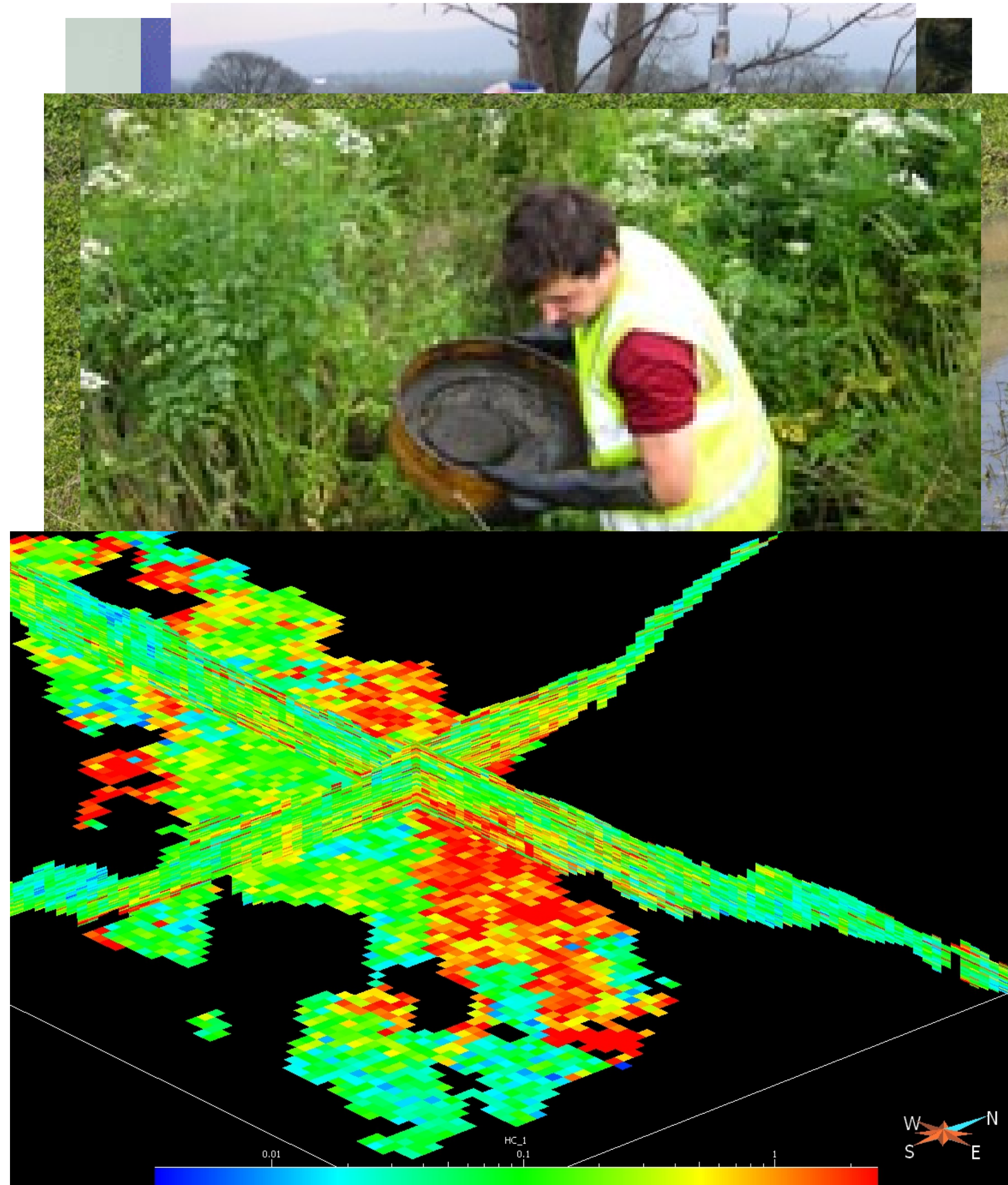
Example geoscience problems

- Complex interactions between geology & anthropogenic processes
  - Large area
  - Multiple datatypes
  - Multiple ages of data
  - Fast turn-around
- Data needs standardising to allow analysis
- BGS has tried phys props mapping in 3D for 20 years
  - Data quality has always prevented it



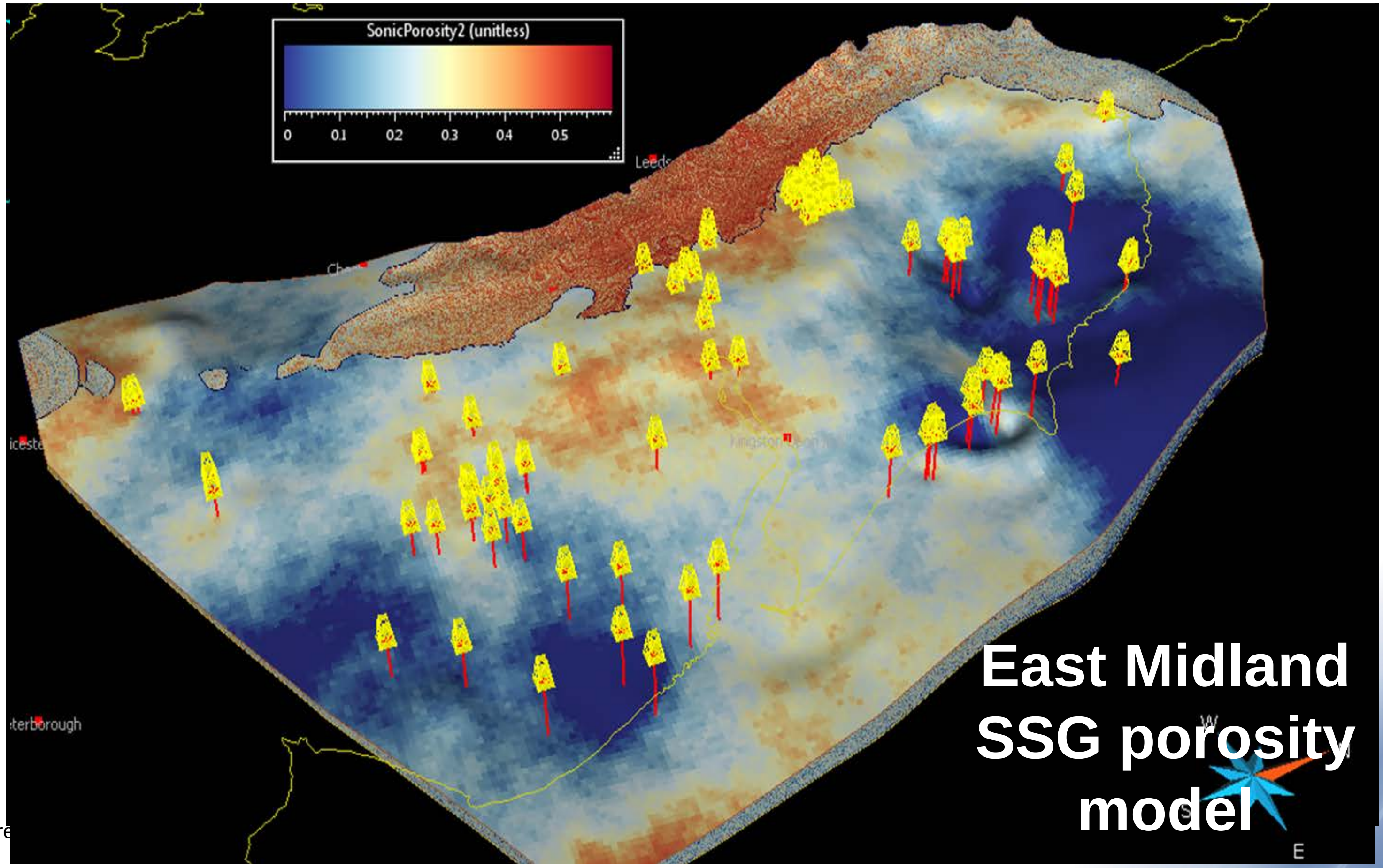
Subsurface Property Data

- Behaviours of rocks is measured in multiple ways
  - Geomechanical properties
  - Geotechnical properties
  - Geophysical properties
  - Groundwater properties
  - Geochemical Properties
  - Etc.
- Needed to attribute 3D geology to model impacts of variability on dynamic processes



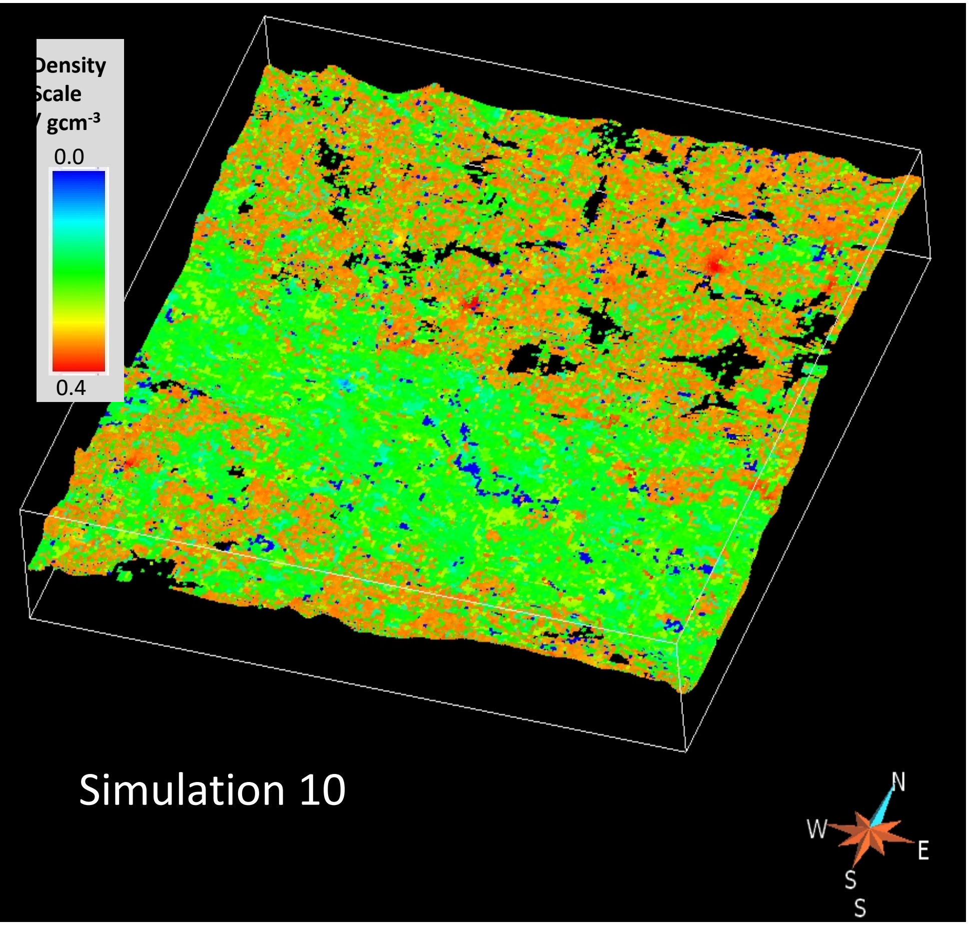
Scientific Challenge: Heterogeneity

- Properties like strength, porosity & permeability affect subsurface processes and are controls on subsurface uses
  - Geologists use complex **terms** to describe this
  - Engineers & Government need data / information in **numbers**



East Midland SSG porosity model

Example: Geology of Glasgow

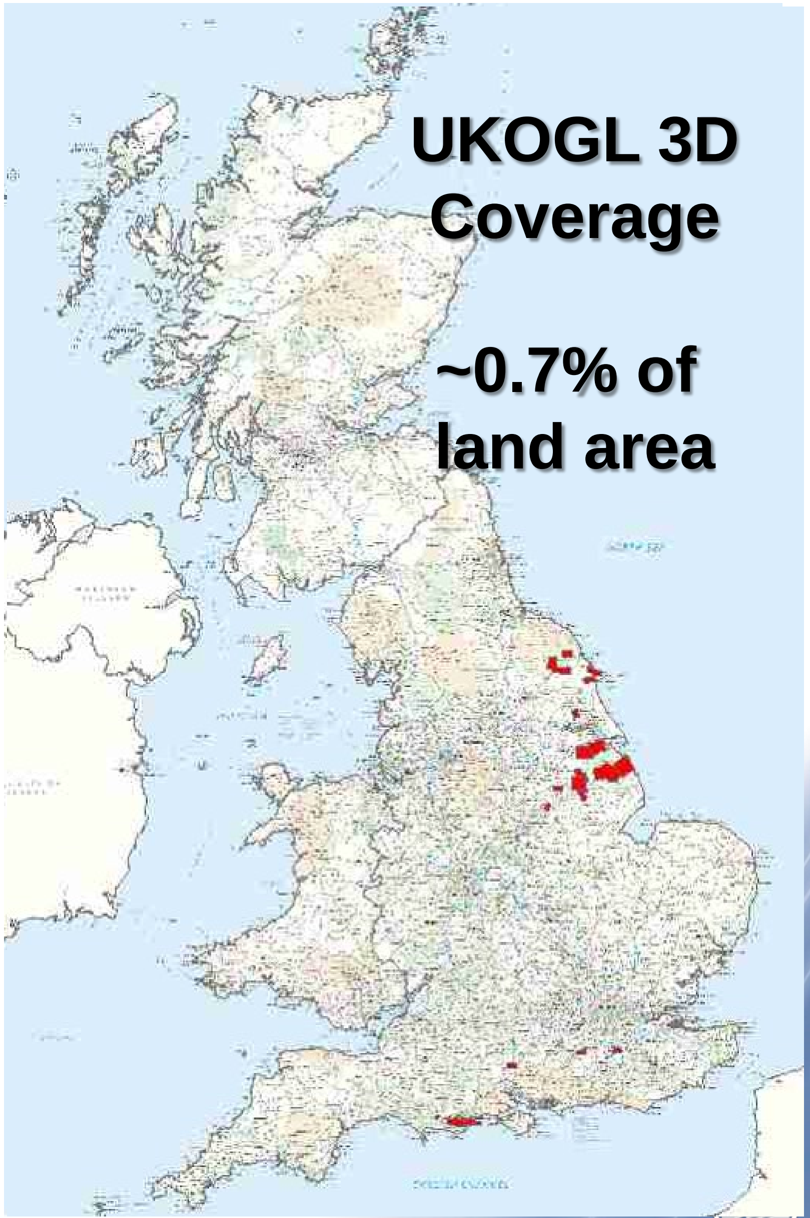


- Heterogeneous mix of clastic lithologies
- Complex packages of fluvial, glacial & marine sediments
- Highly varied physical properties with limited lithostratigraphic control
- Pollution on development sites across the city

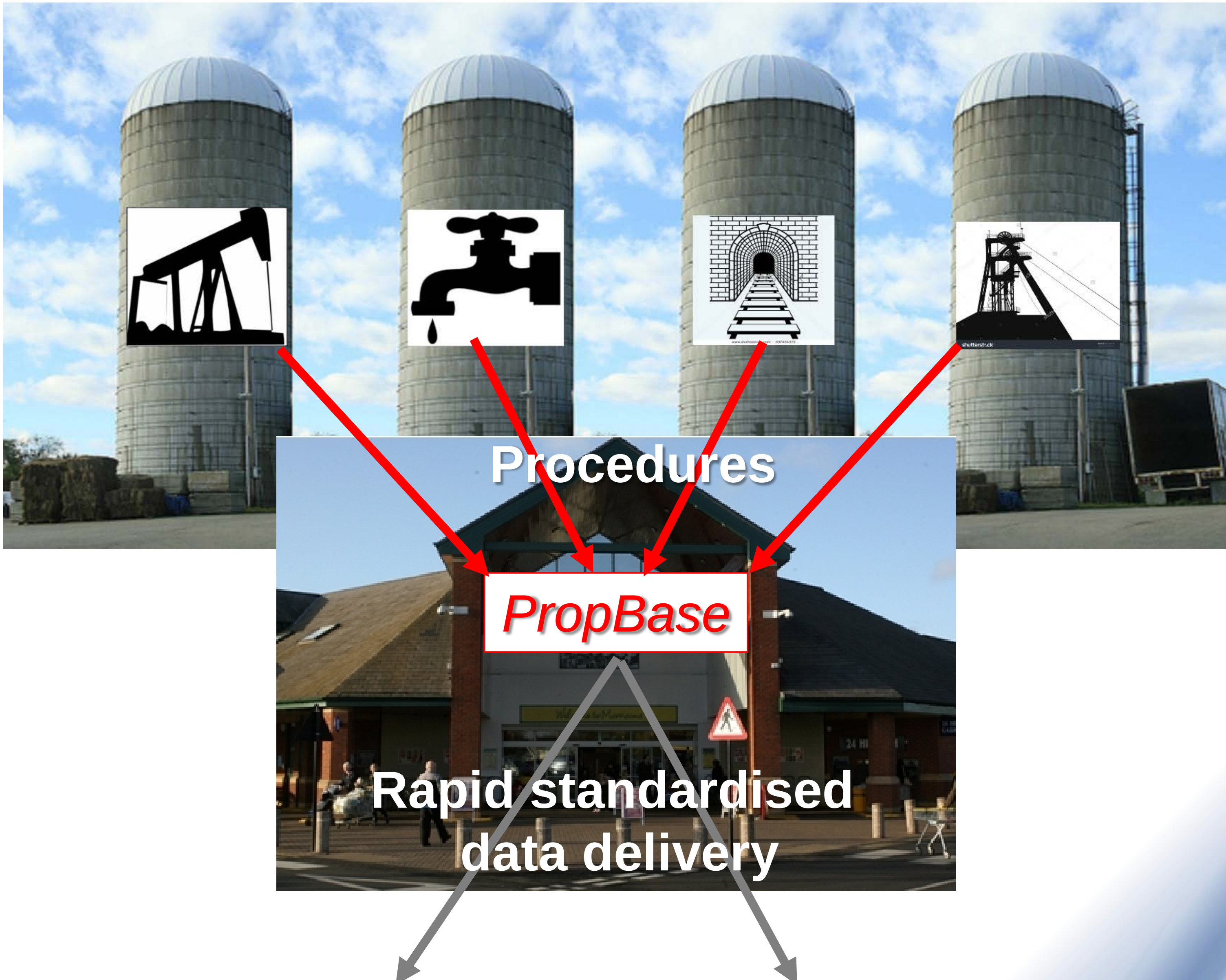
Williams,et al. 2016. Stochastic modelling of hydraulic conductivity derived from geotechnical data; an example applied to Central Glasgow. EARTH ENV SCI T R SO, In Press
Kearsey et al., 2015. Testing the application and limitation of stochastic simulations to predict the lithology of glacial and fluvial deposits in Central Glasgow, UK. DOI:10.1016/j.enggeo.2014.12.017
Bianchi et ,al. 2015. Integrating deterministic lithostratigraphic models in stochastic realizations of subsurface heterogeneity. Impact on predictions of lithology, hydraulic heads and groundwater fluxes. DOI:10.1016/j.jhydrol.2015.10.072

Constraints on "Geo-Big Data"

- Practical Constraints
  - Non-digital data
  - Inadequate metadata
  - Lack of standards
  - Lack of upscaling medium
- Only comparable data can be compared
- "Big-data" analytics need standardised data sets
- Computing power is not the major control

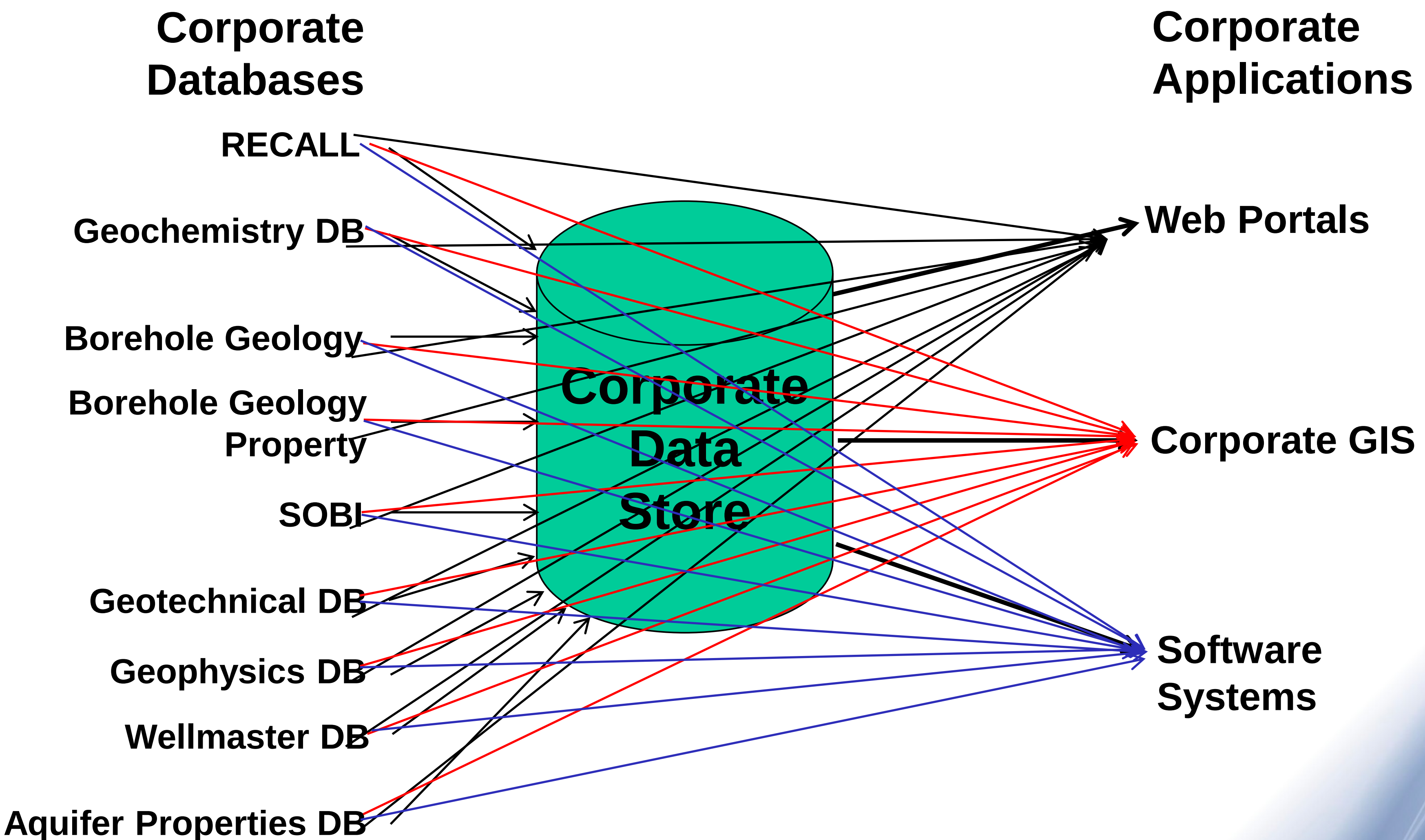


Maximising value from long-tail data

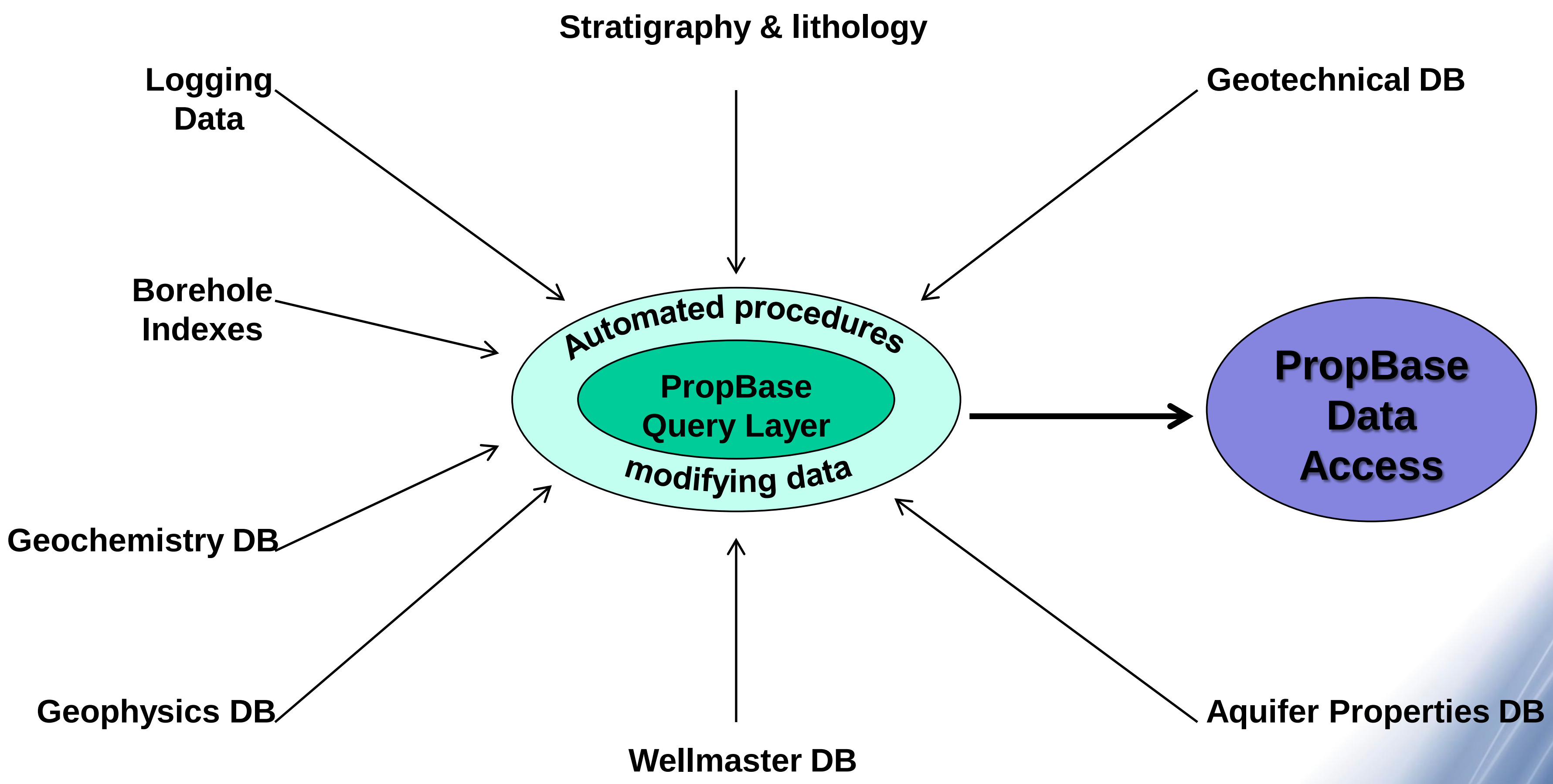


Software tools **Webservices**

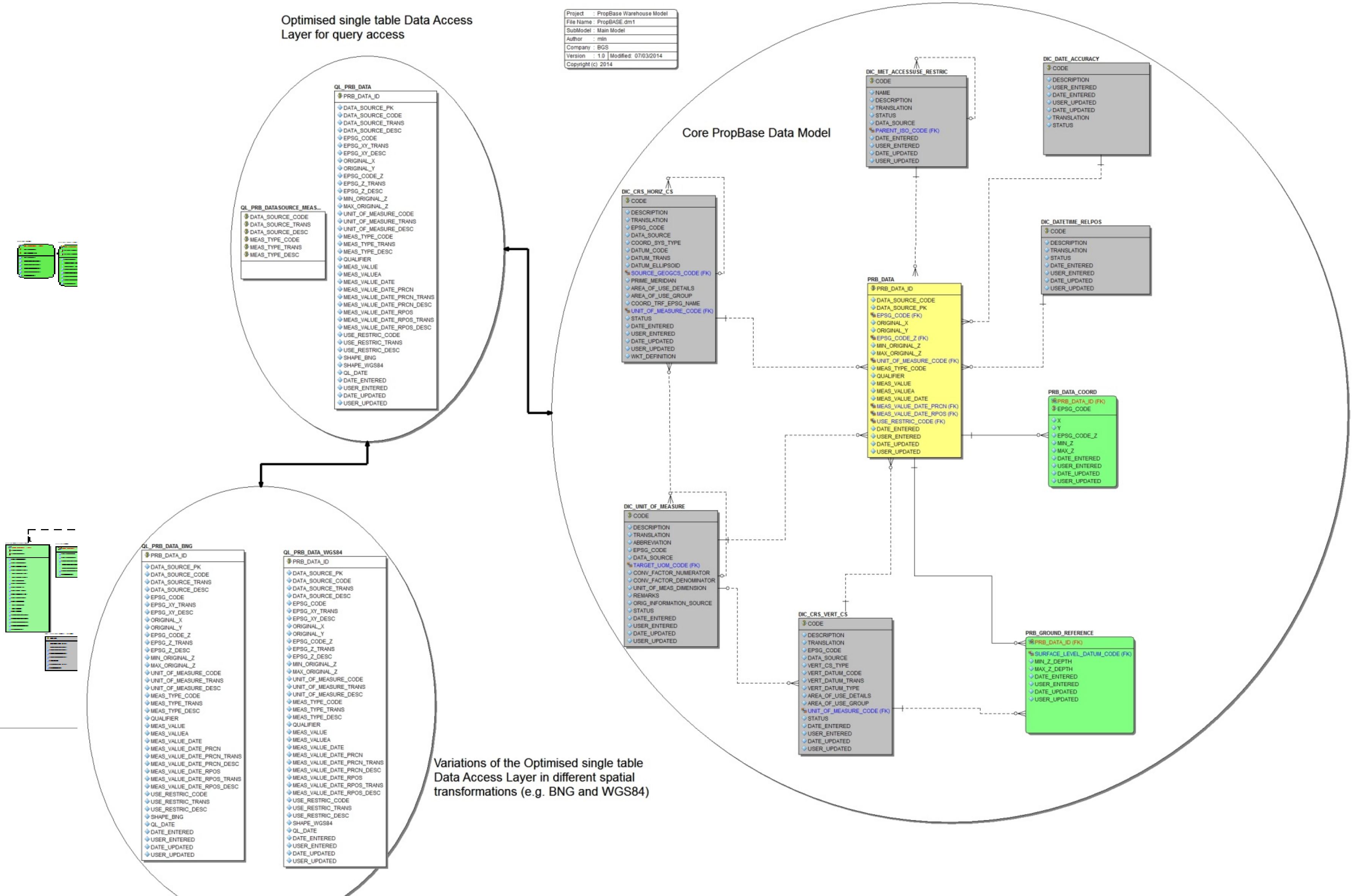
Property data storage



PropBase QueryLayer

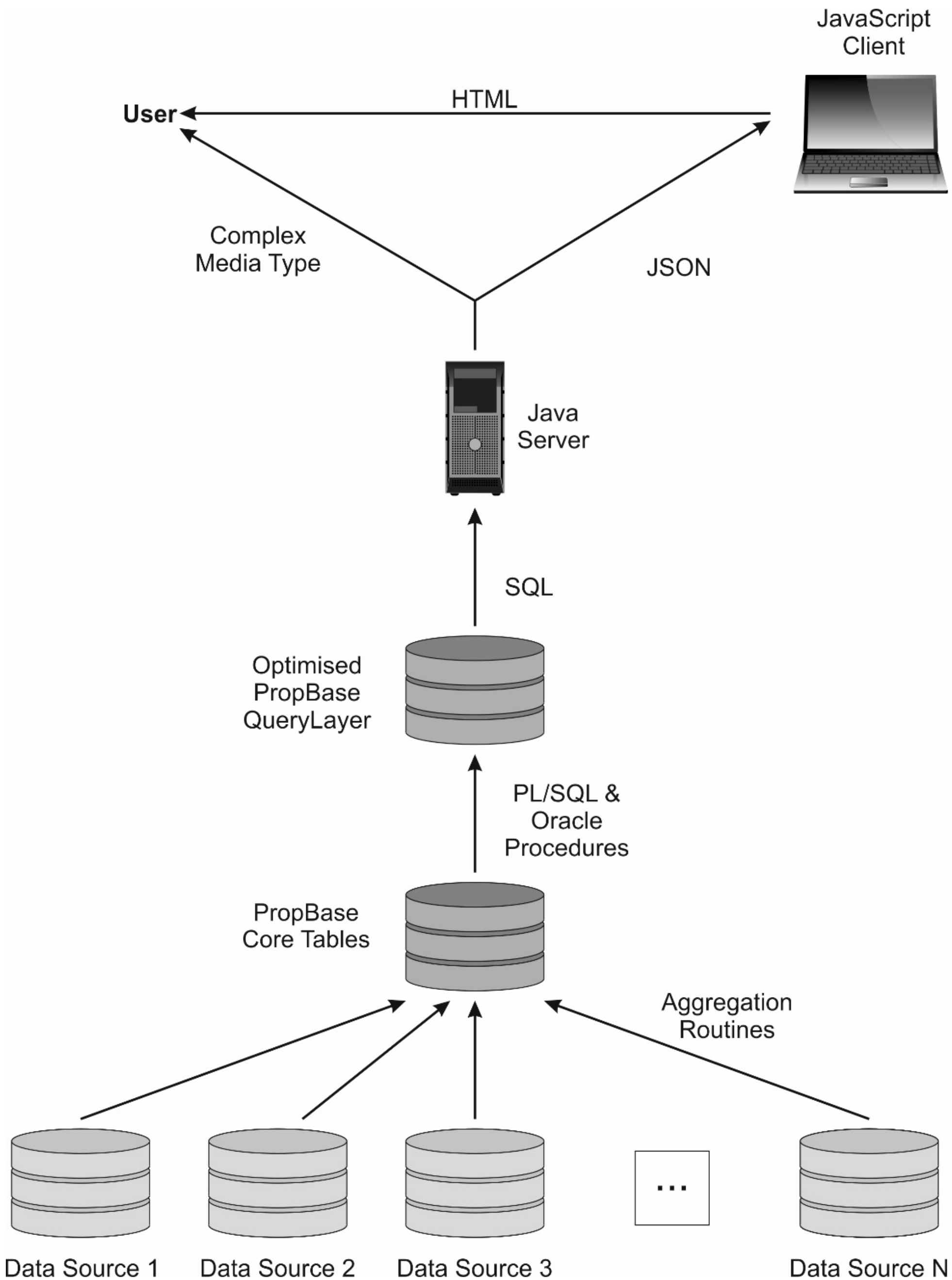


PropBase Data Model



Standard PropBase Output Format

- unique identifier
- data source
- unique id from parent DB (for traceability)
- 3D location (x, y, z)
- property type
- property value
- property units
- necessary qualifiers
- precision information
- audit trail

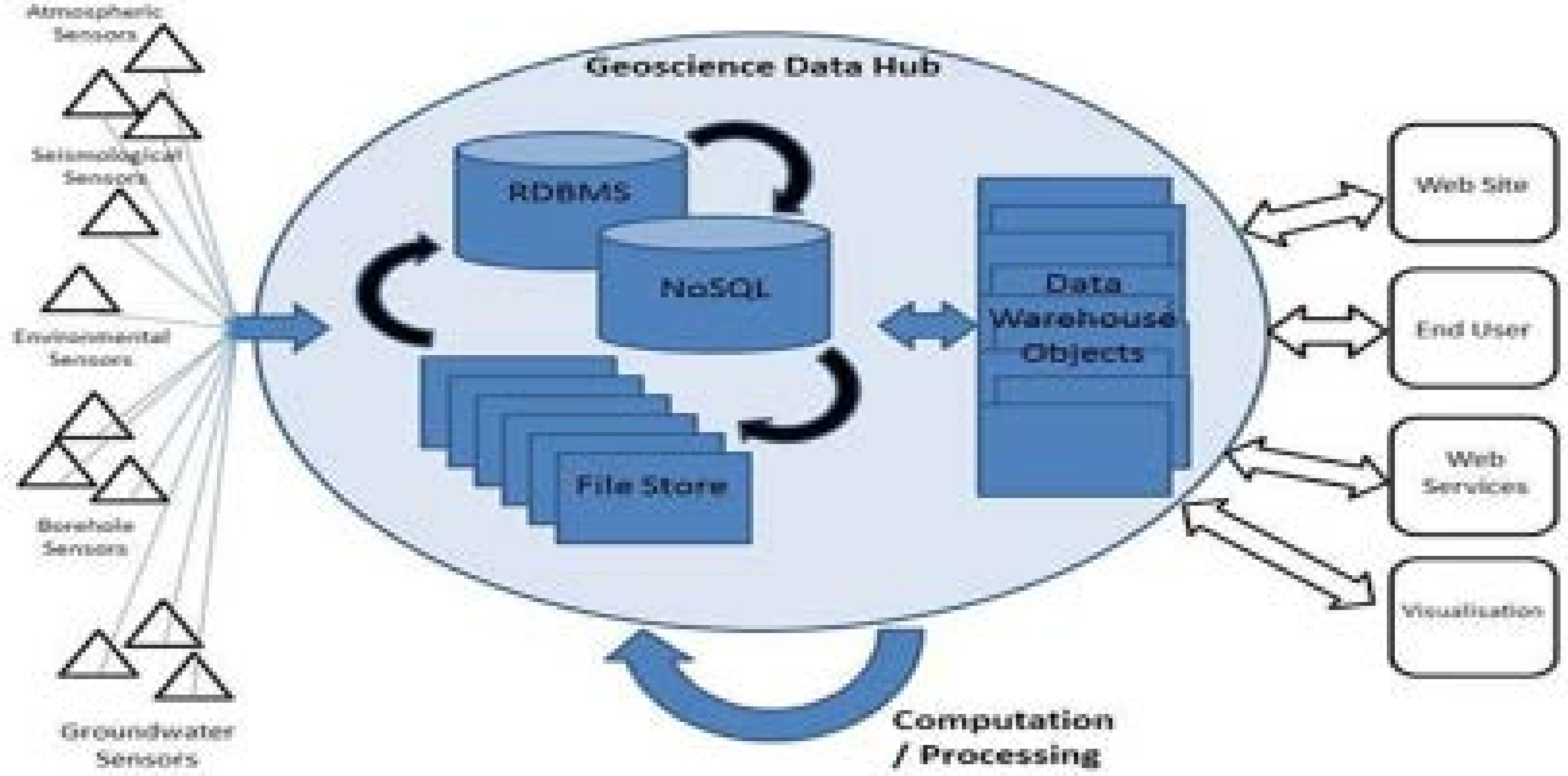


PropBase architecture

How does this solve problems?

- Standardised access allows analysis of:
  - Multiple & large datasets
  - Comparison of many different factors
  - Single input format with toggles between inputs
  - Simplified data analysis
  - Fast response
- Can quickly answer: “what do we know about location X”

Example: SENSORnet

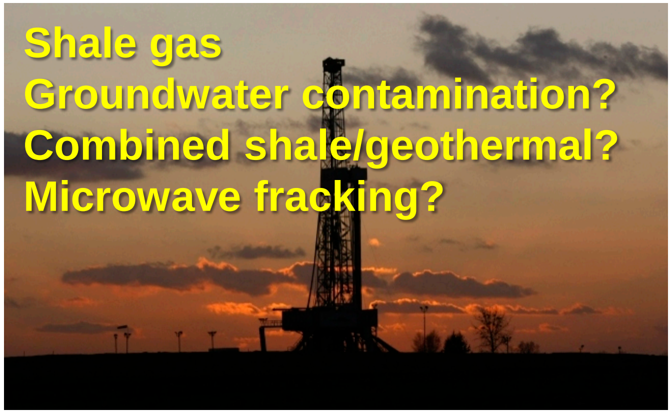


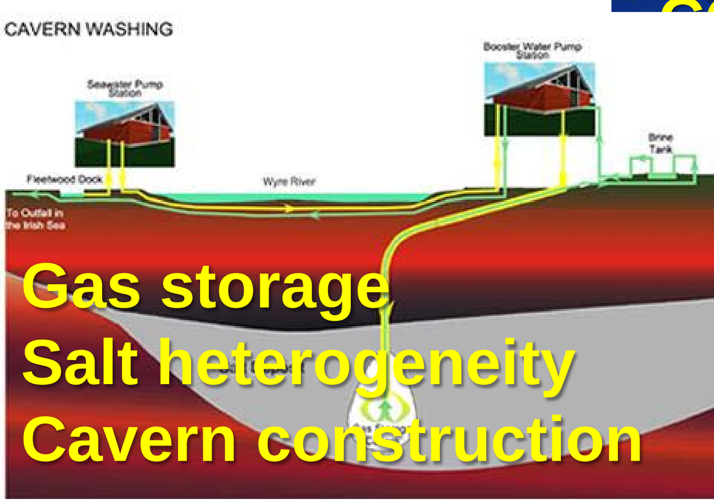
- Near Real-Time sensor data
- Flat file relational database,
- NOSQL databases and
- Data Warehouse objects “Data access layer”

UK Geo Observatories

- BGS's first deep drilling campaign for 30 years
- Studying subsurface operations to understand UK Energy
- Aims to provide best possible datasets to understand how rocks behave in the subsurface
- Data structures and standards must be robust & future-proof

UK Geo Observatories: Science & technology questions in subsurface energy

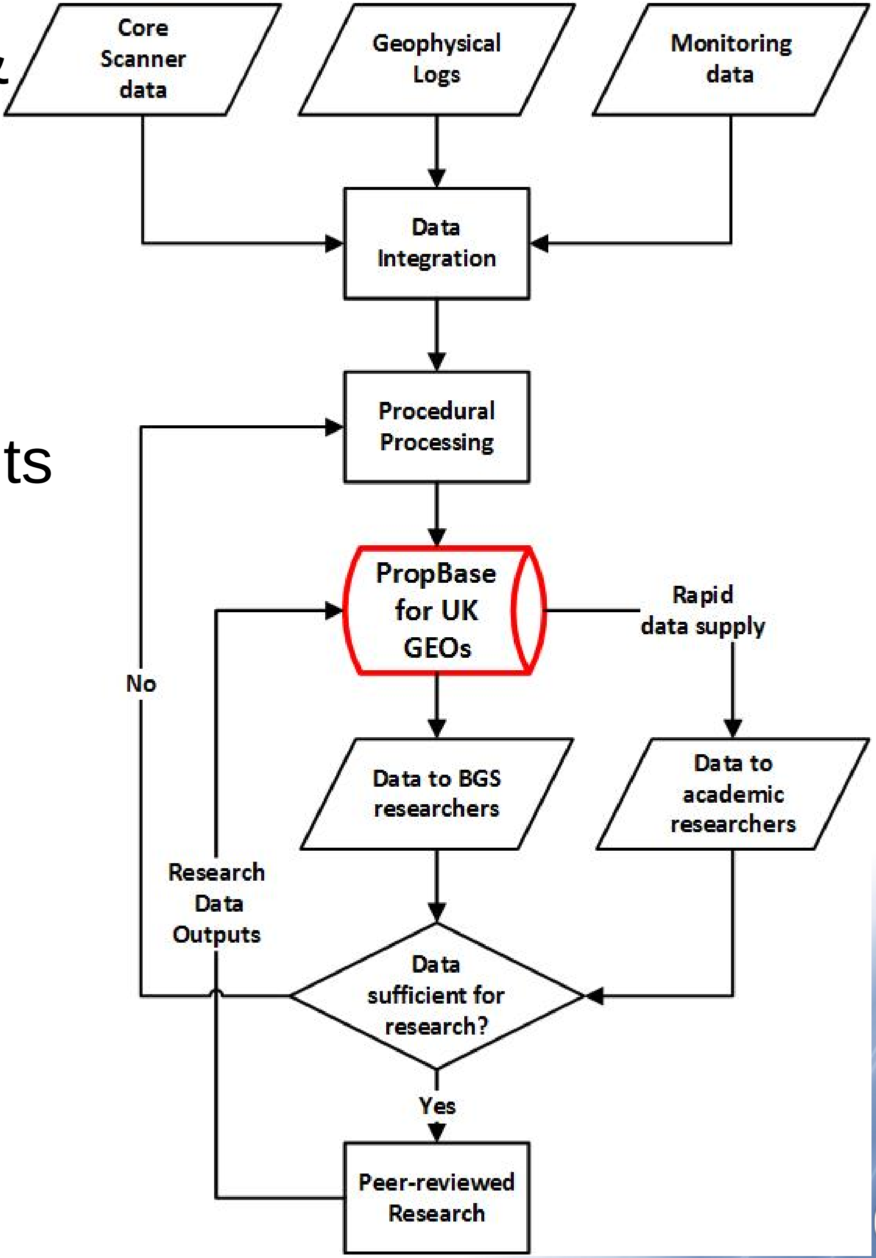




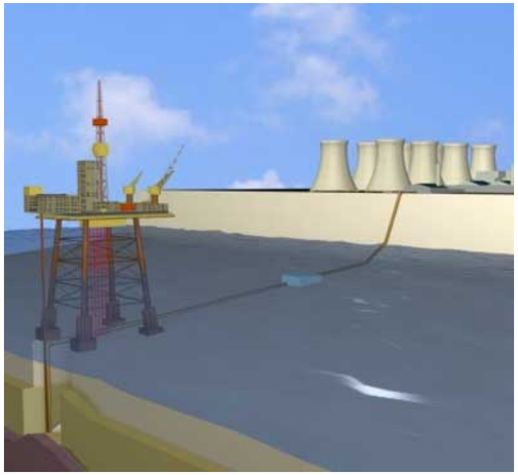
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

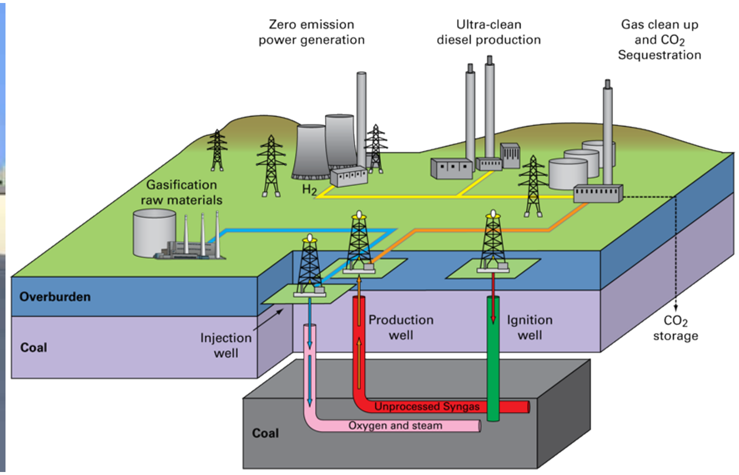
Data to be collected by UK GEOs

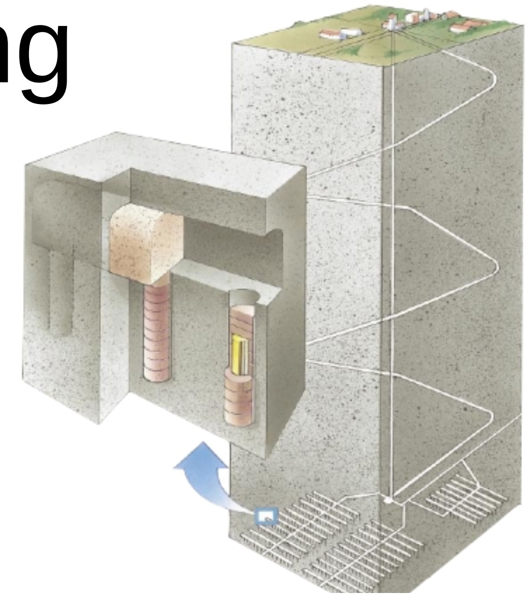
- Baseline geochemistry & seismicity
- Geophysical surveying
- Downhole data inc:
  - Core & core measurements
  - borehole imaging
  - geophysical logging
- Live sensor data
- Fast accurate outputs
  - Quality assured
  - Timestamped
  - Version control



Conclusions: making “Geo-Data” fit for analytical processing







- The challenges of the future are complex and uncertain
- Complex problems need complex analysis to identify trends
- Analytical tools need large & clean datasets
  - Geoscience datasets are dirty with high signal-noise ratio
  - Often measuring proxies not actual parameters
- Data standardisation and QA are essential preconditions to analysis
- PropBase does not undertake analysis, only prepares data
- Effective preparation makes “GeoData” analytics a realistic prospect