Scalable Access to Big Data

Martin Giese

UiO : University of Oslo

Optique™

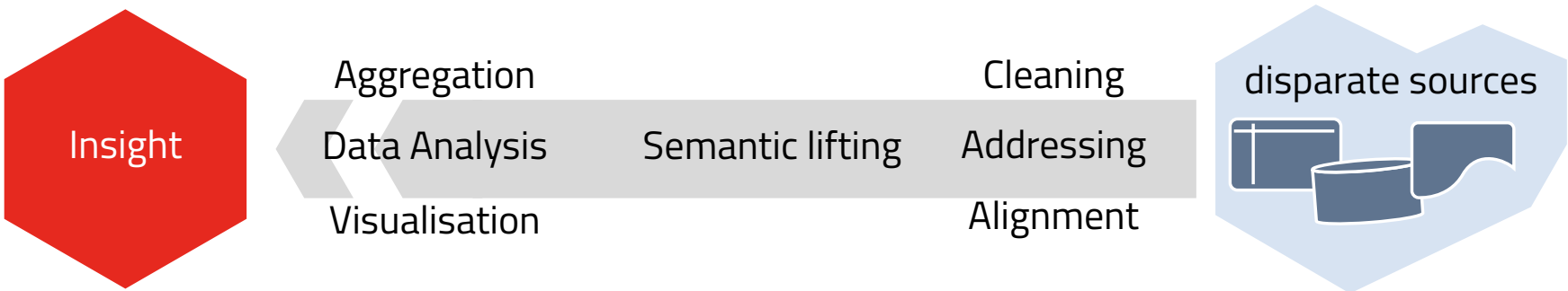
SEVENTH FRAMEWORK PROGRAMME

SIRIUS

sfi Centre for Research-based Innovation
The Research Council of Norway

The Research Council of Norway

The Data Science **Continuum**



The Data Science **Continuum**

Engineer

IT-expert

Insight

Aggregation
Data Analysis
Visualisation

Semantic lifting

Cleaning
Addressing
Alignment

disparate sources

The Data Science **Continuum**

Engineer

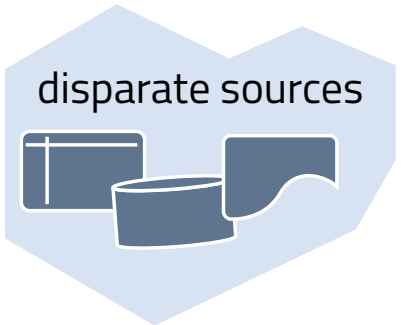
Data Scientist

IT-expert

Aggregation
Data Analysis
Visualisation

Semantic lifting

Cleaning
Addressing
Alignment



The Data Science **Continuum**

Engineer

Data Scientist

IT-expert

Insight

Aggregation
Data Analysis
Visualisation

Semantic lifting

Cleaning
Addressing
Alignment

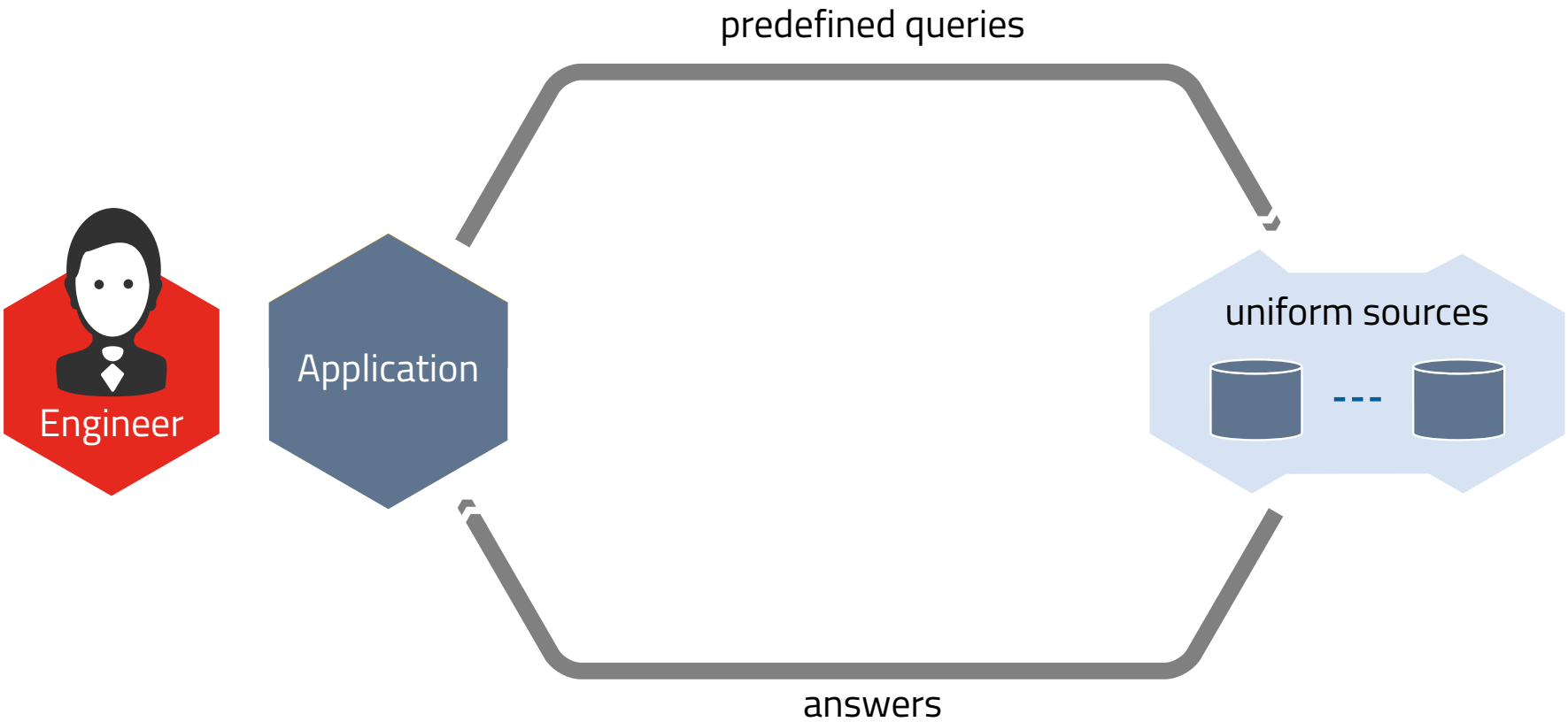
disparate sources

Existing Technology

'Data Access'

Optique

The Problem of **Data Access**



When does this **Go Wrong?**

I need to find all rock samples where my Company had at least a 30% share of the licence at the time the sample was taken. I'm sure the information is there but there are so many concepts involved that I can't find it in the application.

I need all wellbores with a pore pressure of over 14ppg, but lower than 12ppg further down the hole. I can't say this to the application.

I need to find all rock samples for this oil field, including the ones in this Excel sheet from Dinoco. The application doesn't know about this data.

What then happens?

- *Where is this information stored, and what is it called?*
- *Can you hand-craft a query for my information need?*
- *Can you include data from this spreadsheet in the db?*

- May take **weeks** to respond
- Takes several years to master data stores *and* user needs

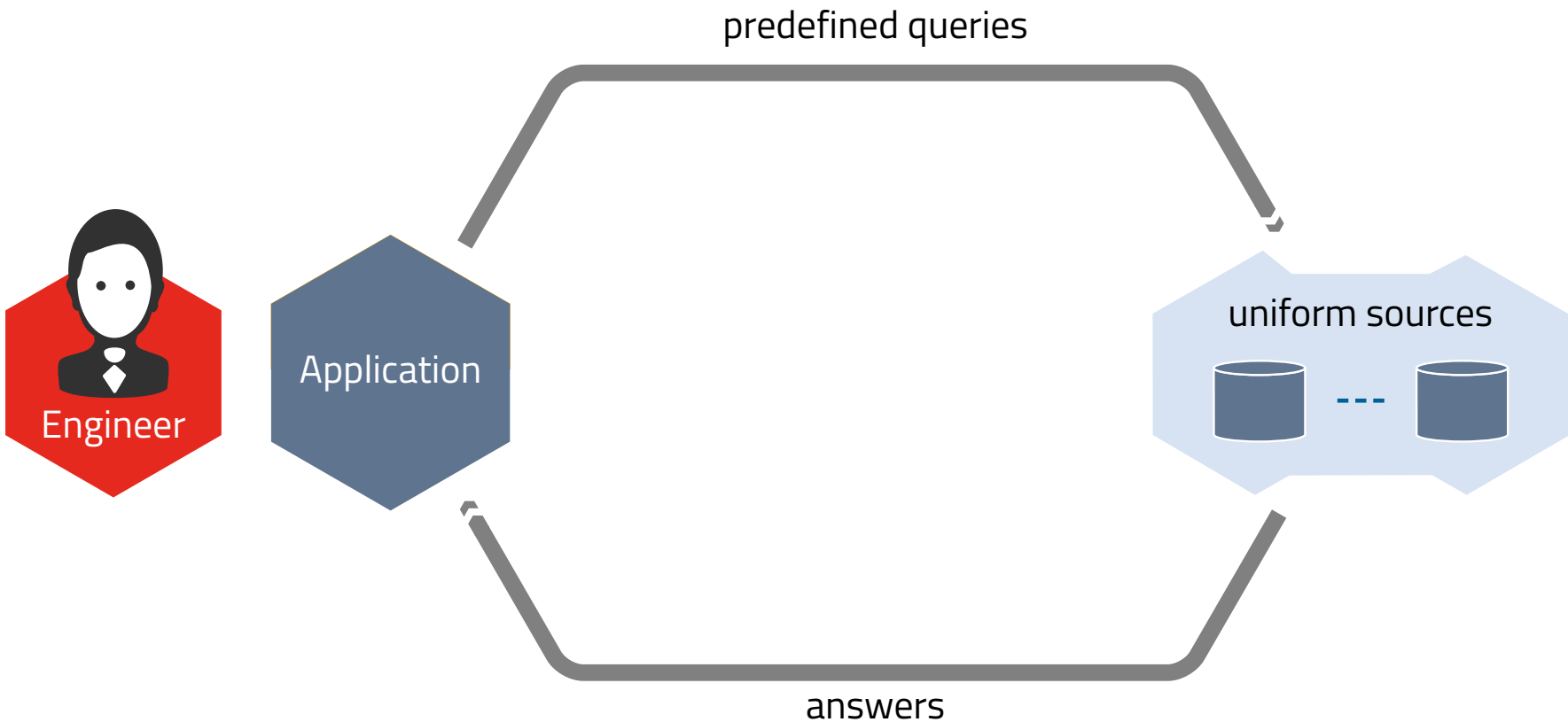
What then happens?

- *Where is this information stored, and what is it called?*
- *Can you hand-craft a query for my information need?*
- *Can you include data from this spreadsheet in the db?*

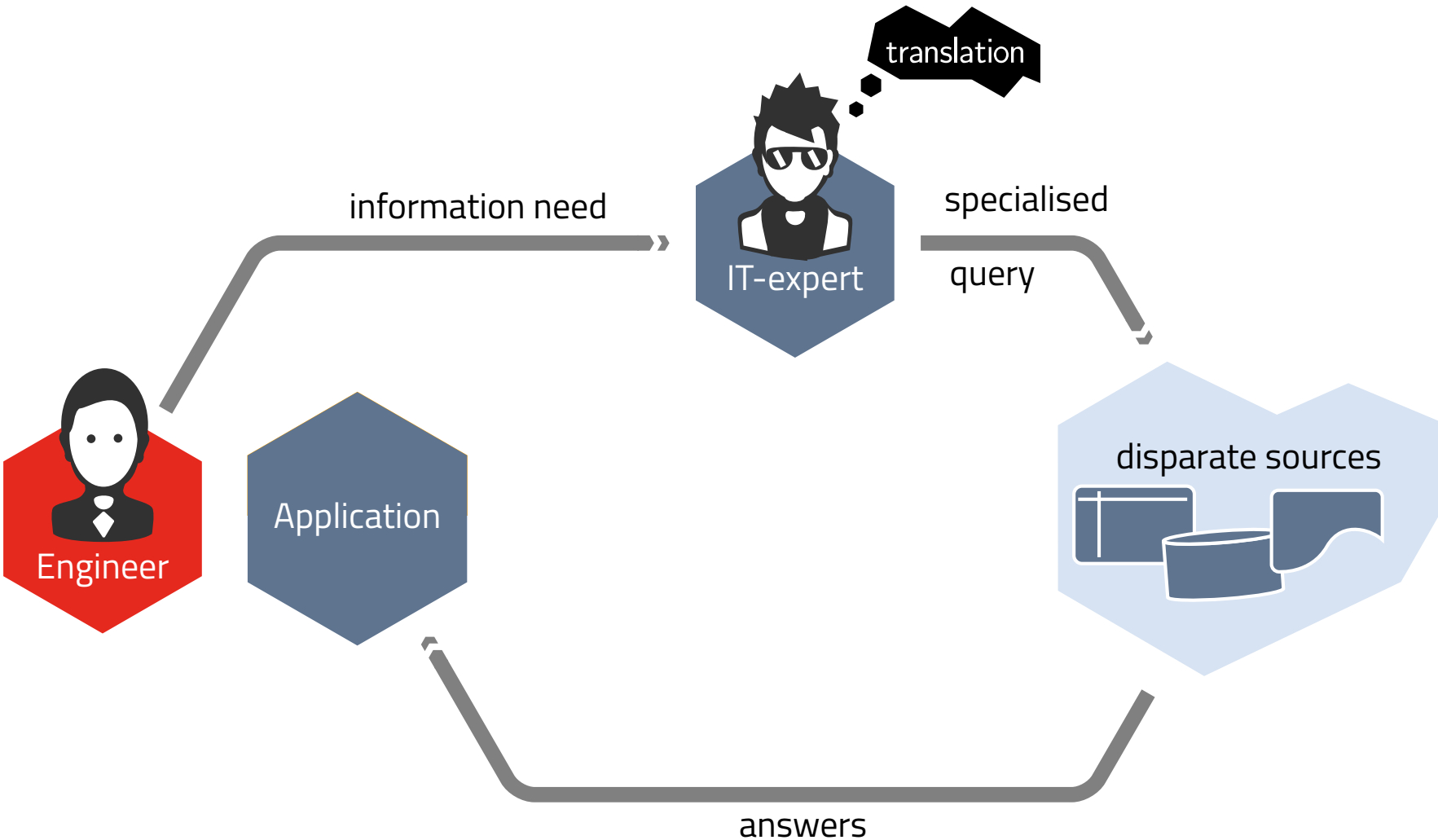
- May take **weeks** to respond
- Takes several years to master data stores *and* user needs

30–70% of domain expert time spent looking for and assessing the quality of the data found

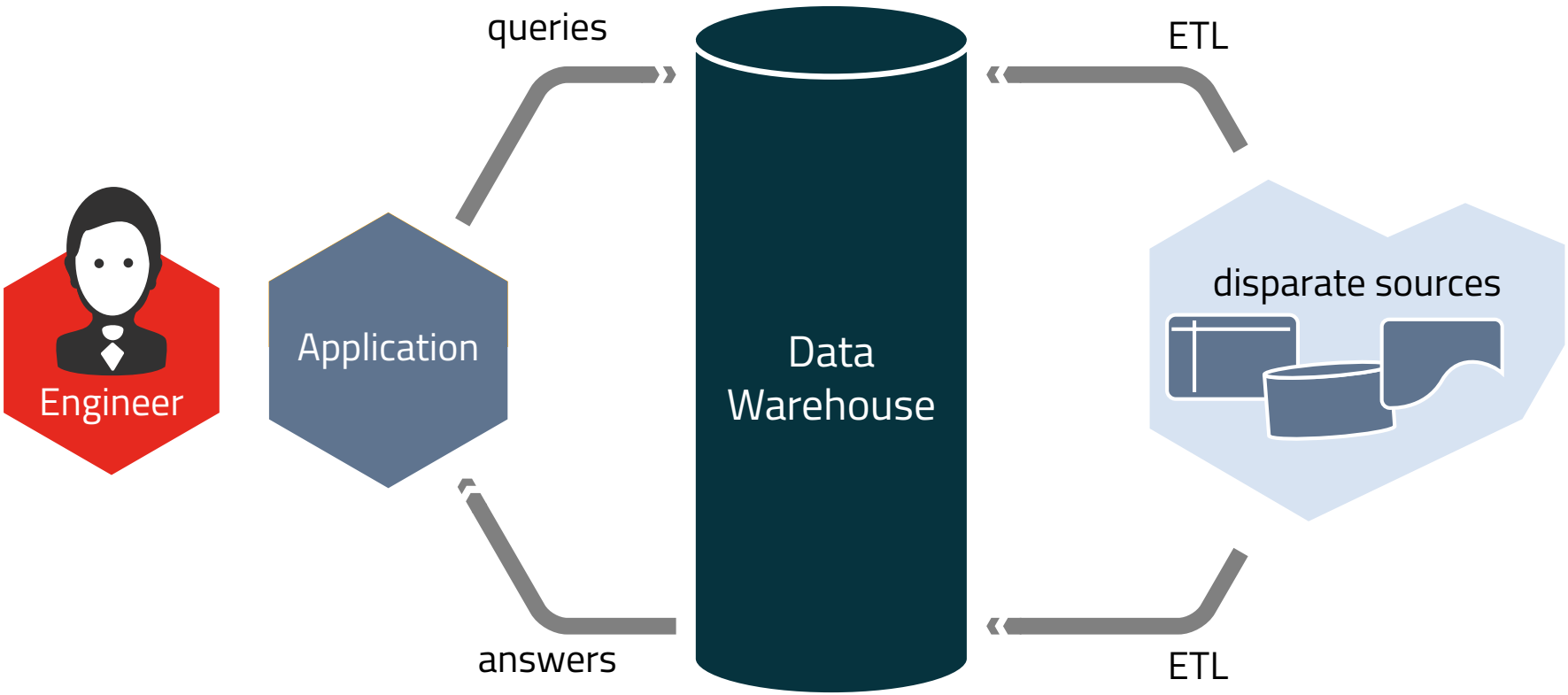
The Problem of **Data Access**



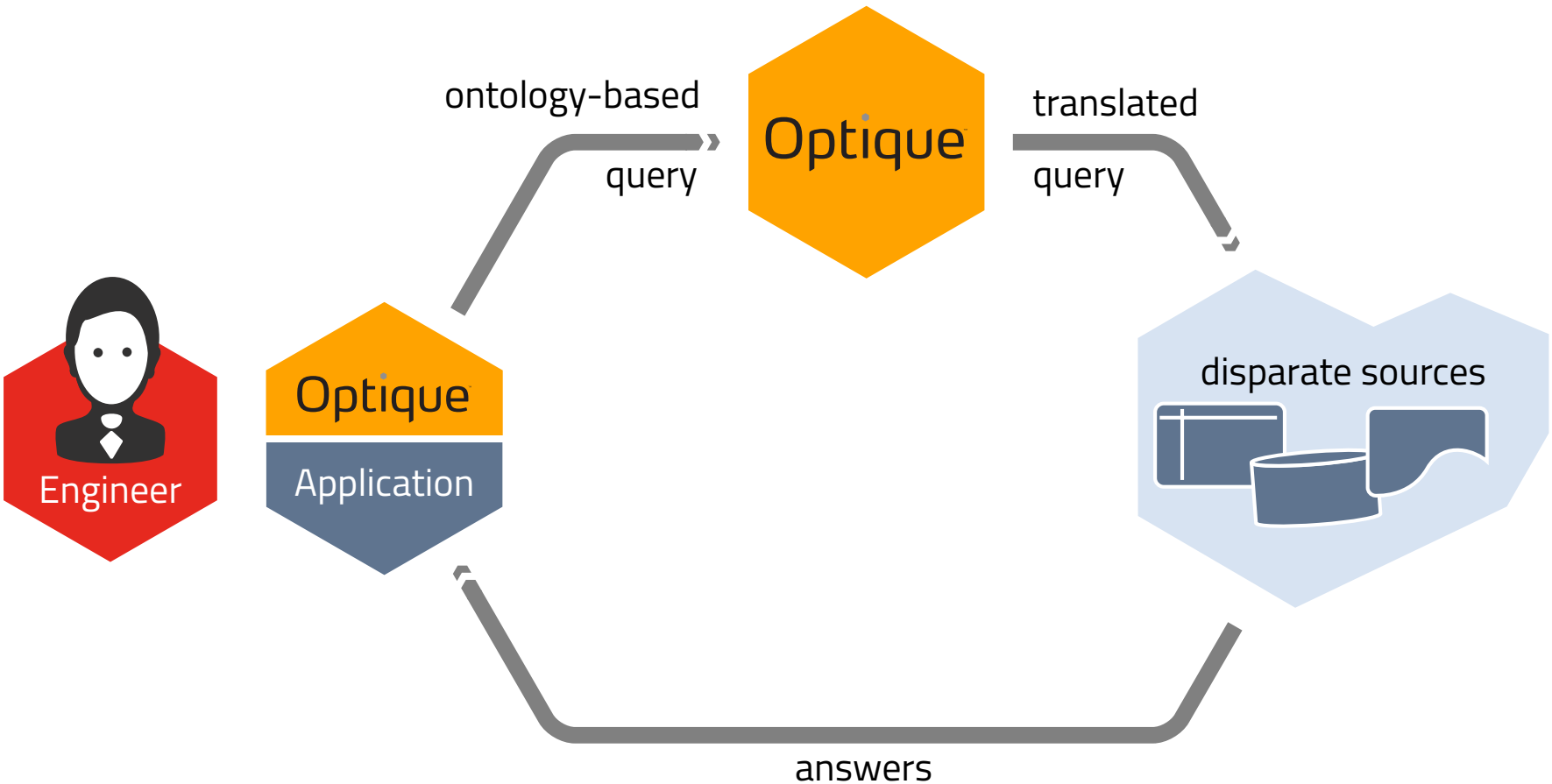
The Problem of **Data Access**



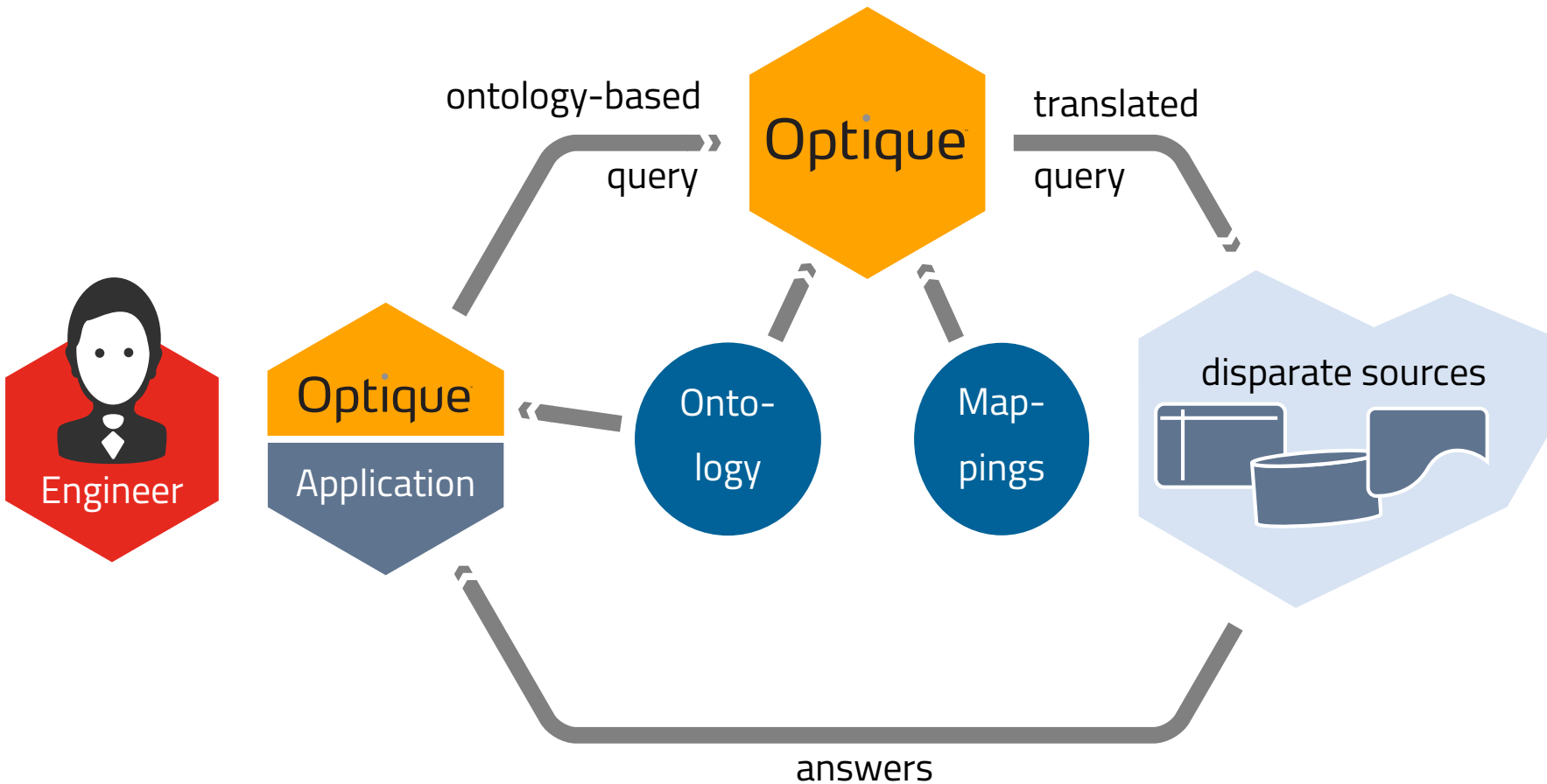
Data Access, with a **Data Warehouse**



Data Access: The **Optique Solution**



Data Access: The **Optique Solution**



Ontology-based Data Access

- Capture End-user vocabulary in an "Ontology"
  - $\approx$ Domain model
  - Classes and relations known to end-users
  - Some minimal domain knowledge
- Mappings that relate Ontology with data sources
  - 'Column "Type" is "T" in row *x* of table "Sensors" if sensor Nr. *x* is a Temperature Sensor'
- Automatically translate queries in End-user language to queries over data sources.

  In: 'List all temperature sensors.'

  Out: 'Print "Sensor Nr. *x*" for all rows *x* in "Sensors" table where "Type" column is "T."'

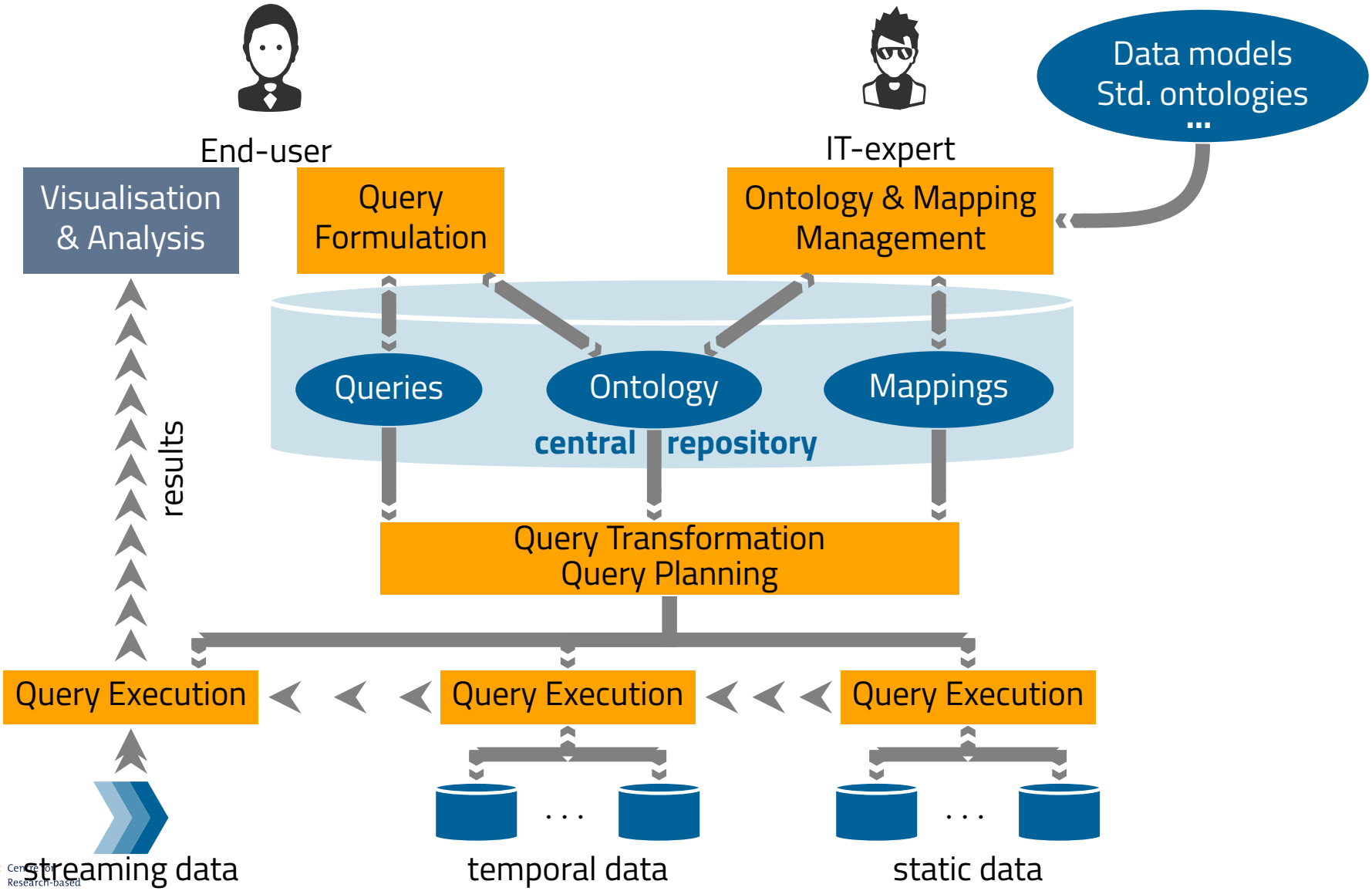
Optique **Focus Areas**

Basic principles of OBDA predate Optique

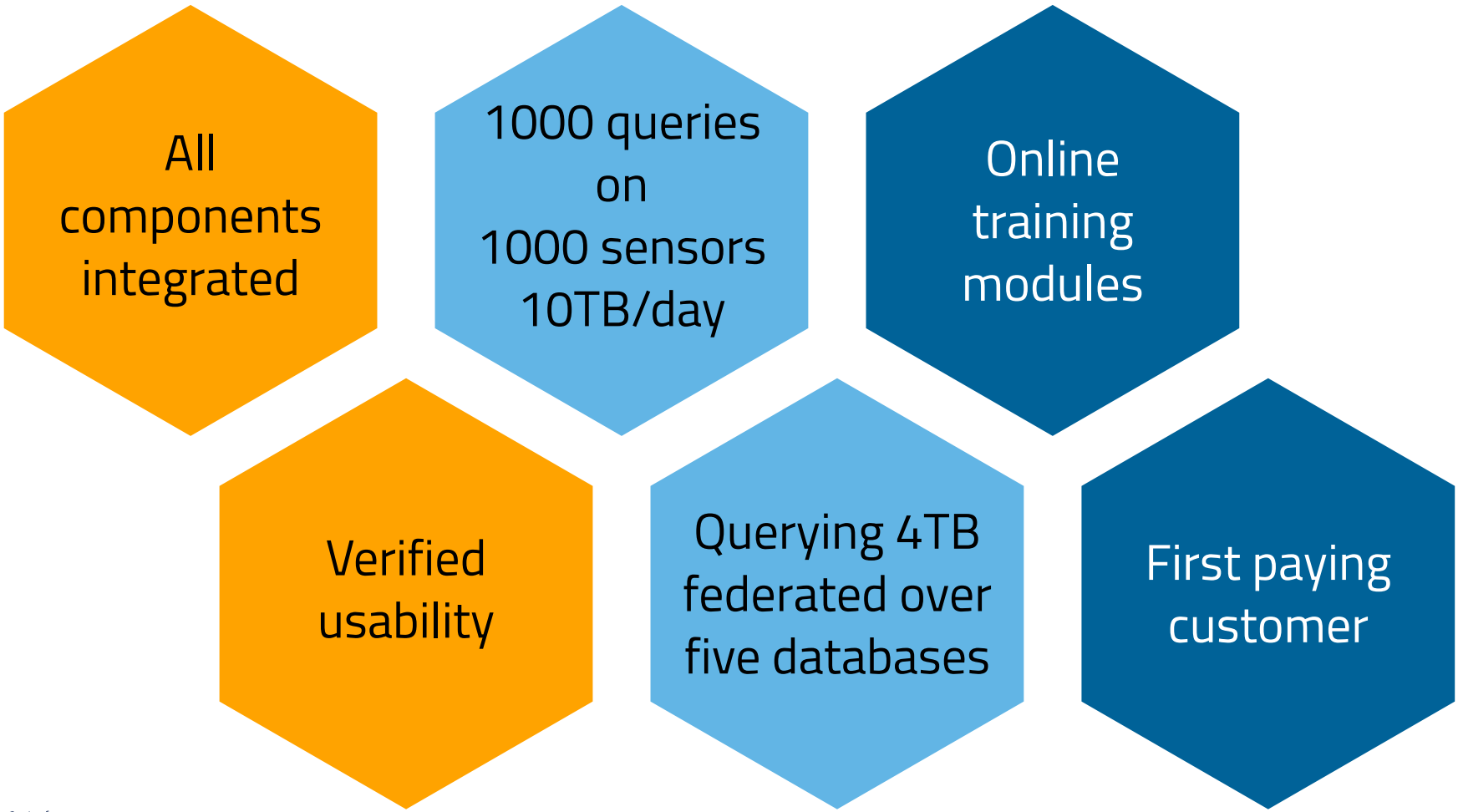
Optique focussed on practical issues:

- Usability
  - How do end-users formulate queries? In first-order logic?
  - Need a user interface for 'query formulation'
- Scope
  - What about queries with time? Or geology? Or chemistry?
  - Need to extend bare-bones query rewriting
  - Plug-in architecture for special domains
- Prerequisites
  - Where do the ontology and mappings come from?
  - How do you maintain them?
- Efficiency
  - SQL databases not good at queries from OBDA
  - Big Data is maybe not best stored in an SQL database
  - Optimize rewritten queries *and* storage layer

Optique **Architecture**



Status at end of project



Optique is Dead **Long Live...**

Centre For Scalable Data Access In The Oil & Gas Industry

- Optique style Big Data Access
- Natural Language Processing
- Cloud Computing
- HPC hardware

http://www.sirius-labs.no/

Data Science @ University of **Oslo**

- Optique, SIRIUS: Data Access
- BigInsight: Data Analysis, Statistics, Machine Learning

- UiO aspires to become a leading institution in Data Science

The Data Science **Continuum**

Engineer

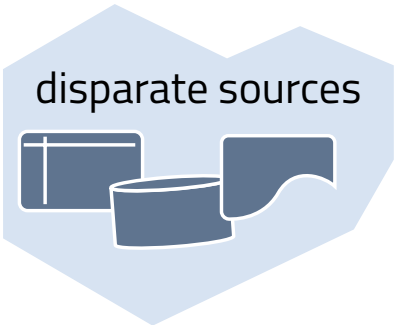
Data Scientist

IT-expert

Aggregation
Data Analysis
Visualisation

Semantic lifting

Cleaning
Addressing
Alignment



BigInsight

SIRIUS

The Data Science **Continuum**

Engineer

Data Scientist

IT-expert

Insight

Aggregation
Data Analysis
Visualisation

Semantic lifting

Cleaning
Addressing
Alignment

disparate sources

BigInsight

SIRIUS

Data Science @ UiO

Analytics Aware OBDA

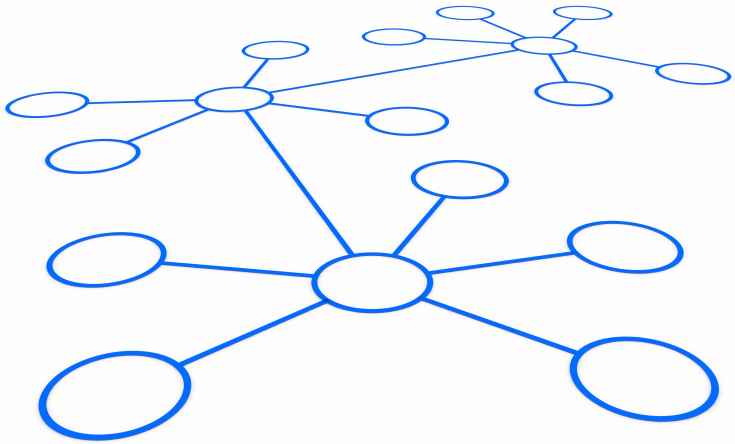
- Analytics operations as part of query language
- Execute Aggregation and Analysis near the sources
- Novel possibilities for optimisation and distribution
  - Fog Computing
- ISWC 2016 paper:
  "Towards Analytics Aware Ontology Based Access to Static and Streaming Data"
  E. Kharlamov et al.
  `https://arxiv.org/abs/1607.05351`

Semantics-based **Data Cleaning**

- Ontologies describe the *world*
- … how the data *should be*
- Basis for detecting problems
  - Data inconsistent with ontology
  - Need more information to fix
- Combine with statistical model of what happened to data!
  - Clean data
  - Repair data
- Next step: Information in missing data!
  - Extend ontology, add mappings
  - Can statistical methods help to do this?

Statistical Methods for **Graph Data**

- Ontologies describe the world in terms of
  - Classes (types)
  - Relations (object-to-object)
  - Properties (object-to-value)
- Compared to other data:
  - Richer than vector data
  - Richer than pure graphs/networks
- In Oil & Gas
  - Description of installations
  - Stratigraphy
  - …
- Combine probabilistic and ontological models:
  - Cleaning and repairing graph data
  - Machine learning, prediction on graph data




SIRIUS
Center for Scalable Data Access
in the Oil and Gas Domain