A World Leading SFI Research Centre

Science Foundation Ireland SFI For what's next

The Symbiotic Relationship between Data Spaces and AI. Challenges and Trends

Insight

SFI RESEARCH CENTRE FOR DATA ANALYTICS

Prof. Edward Curry
University of Galway
Insight SFI research Centre for Data Analytics

HOST INSTITUTIONS

DCU Dublin City University

Ollscoil na Gaillimhe University of Galway

UCC University College Cork, Ireland

UCD Dublin

PARTNER INSTITUTIONS

Maynooth University

Tyndall National Institute

Trinity College Dublin The University of Dublin

University of Limerick

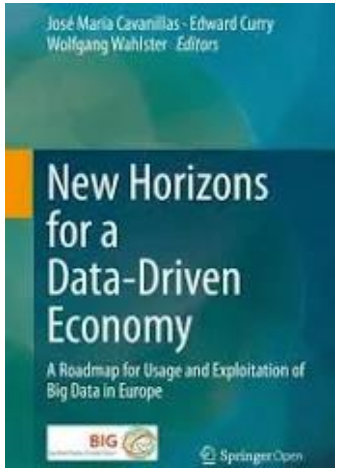
FUNDED BY:

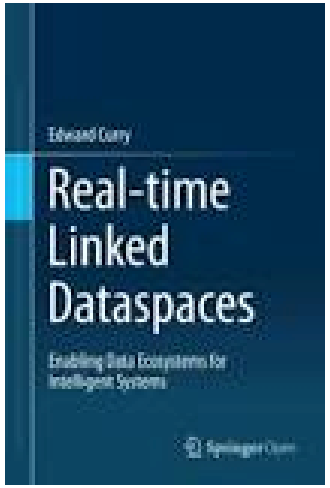
European Union European Regional Development Fund

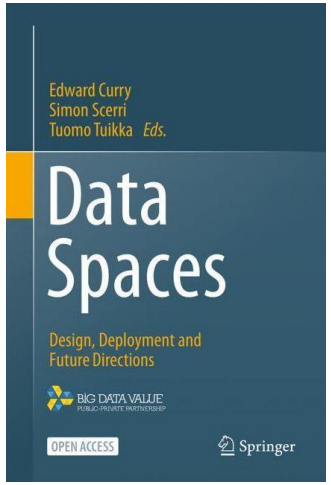
Science Foundation Ireland SFI For what's next

Edward Curry

I have been researching the underlying technology for data spaces for the last decade…







Ollscoil na Gaillimhe
University of Galway

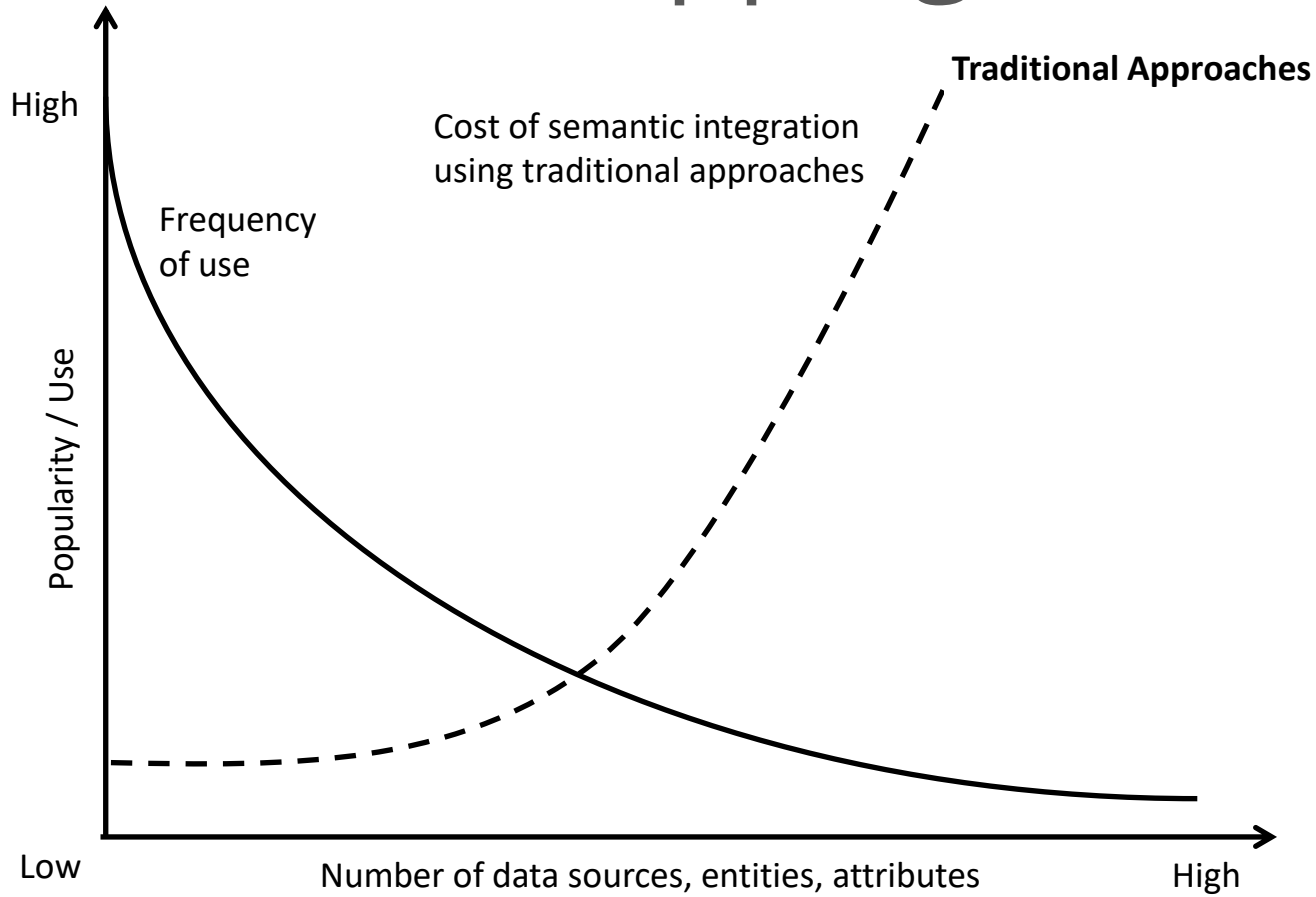
Insight
SFI Research Centre for Data Analytics

BDV
Big Data Value Association

adra
The AI Data Robotics Association

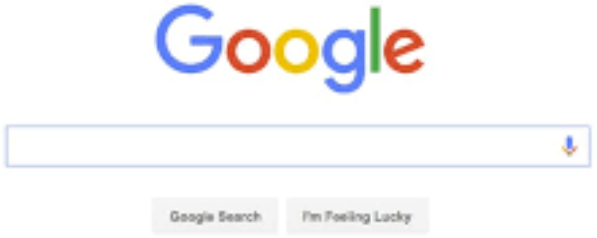
Data Spaces need a data co-existence approach for “Good Enough” Interoperability

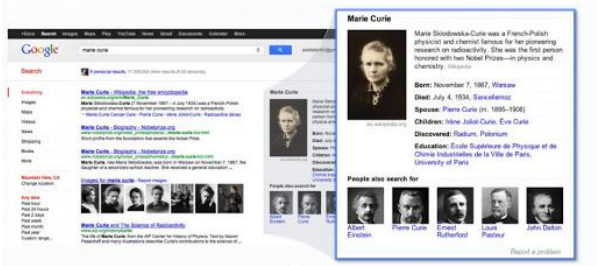
Long Tail Semantics continues to grow… and the cost of mapping increases….



The Long Tail of Data

A new paradigm for human-data Interaction……





From Structure → **to Search** → **to Knowledge Graph** → **to Conversations ?**

~1995
~100K Websites
Exact Results
Human Curated

~1998
~2.4M Websites
Approximate Results
Computed

~2012
~700M
Approximate Results + Exact
Computed + Crowd

~2023
~2B
Approximate Results ?
Content creation?

Insight

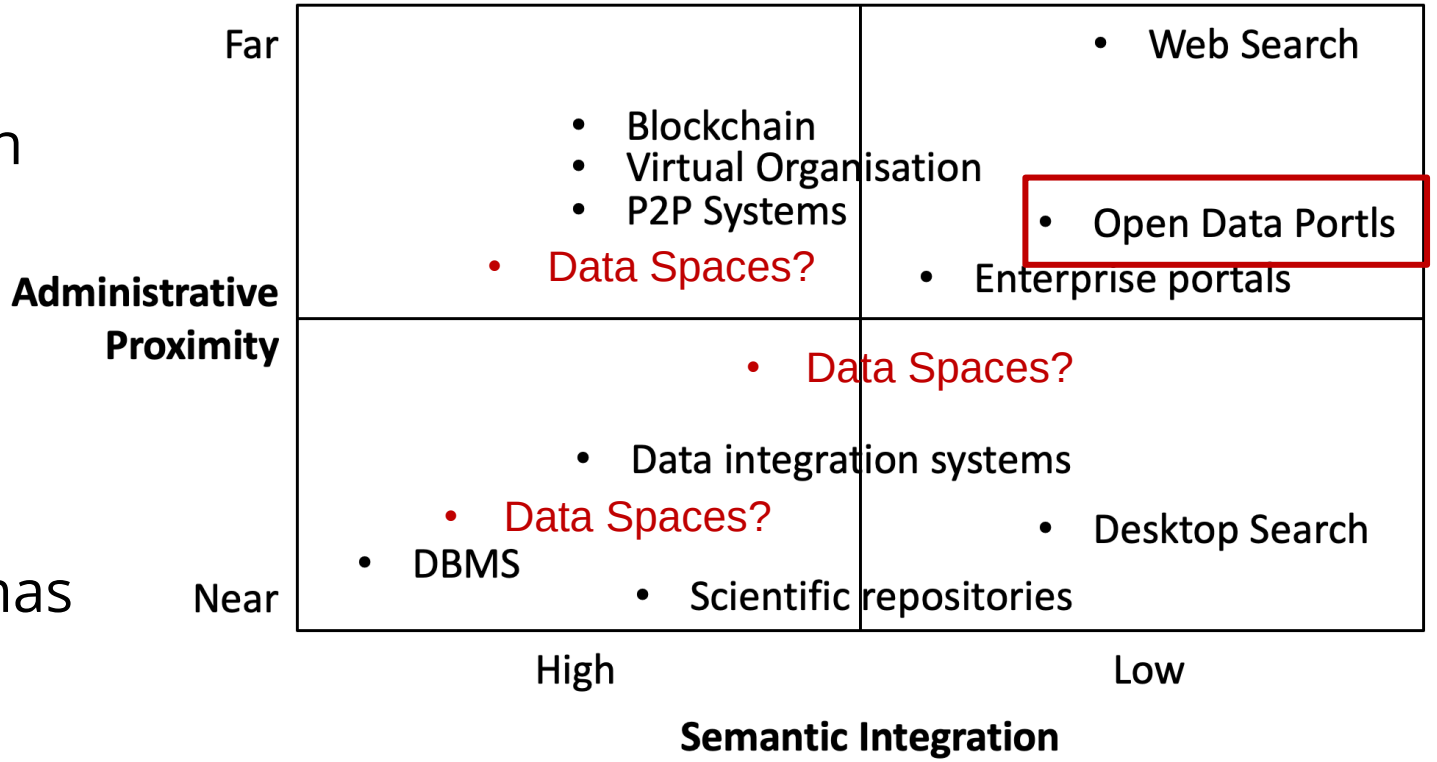
Control and Coordination

Administrative Proximity

- Close vs. Loose Coordination
- Assumptions concerning guarantees such as data, access, quality, and consistency,

Semantic Integration

- Degree to which data schemas are matched up (types, attributes, and names).



Halevy, A., Franklin, M. and Maier, D. 2006. Principles of dataspace systems. *25th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems - PODS '06* (New York, New York, USA, 2006), 1–9.

DOI:10.1145/1953122.1953140

Article development led by acmqueue
queue.acm.org

In today's humongous database systems, clarity may be relaxed, but business needs can still be met.

BY PAT HELLAND

If You Have Too Much Data, then 'Good Enough' Is Good Enough

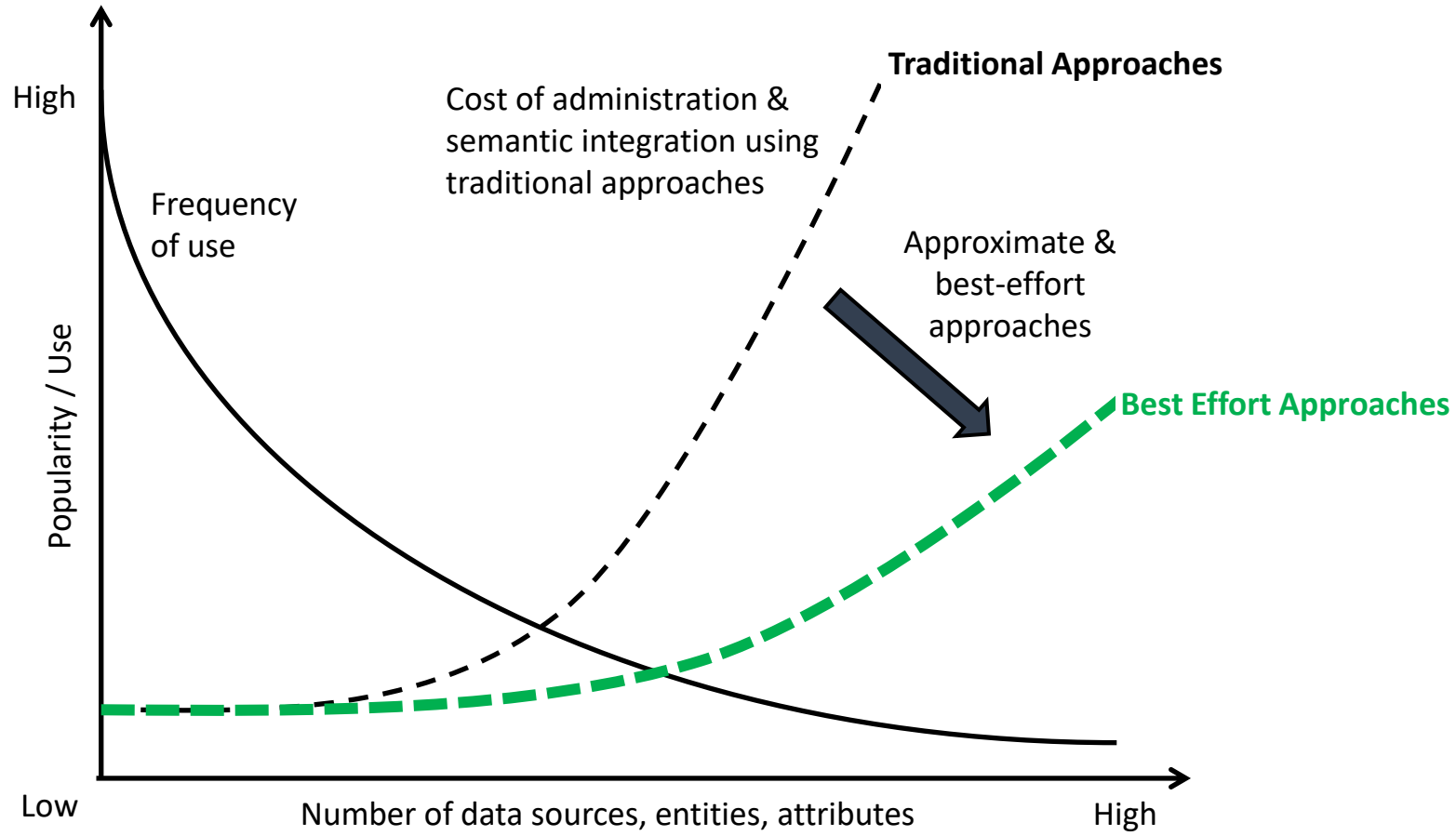
"We can no longer pretend to live in a clean world. SQL and its Data Definition Language (DDL) assume a crisp and clear definition of the data, but that is a subset of the business examples we see in the world around us. It's OK if we have lossy answers—that's frequently what business needs."

What is a Dataspace? (2006)

“Dataspaces are not a data integration approach; rather, they are more of a **data co-existence approach**. The goal of dataspace support is to provide base functionality over all data sources, regardless of how integrated they are.” (Halevy, A., Franklin, M. and Maier, D. 2006.)

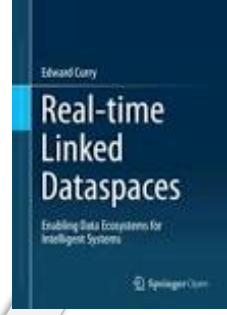
Incrementalism, Approximate, Interactive

Approximate and Best Effort Approaches

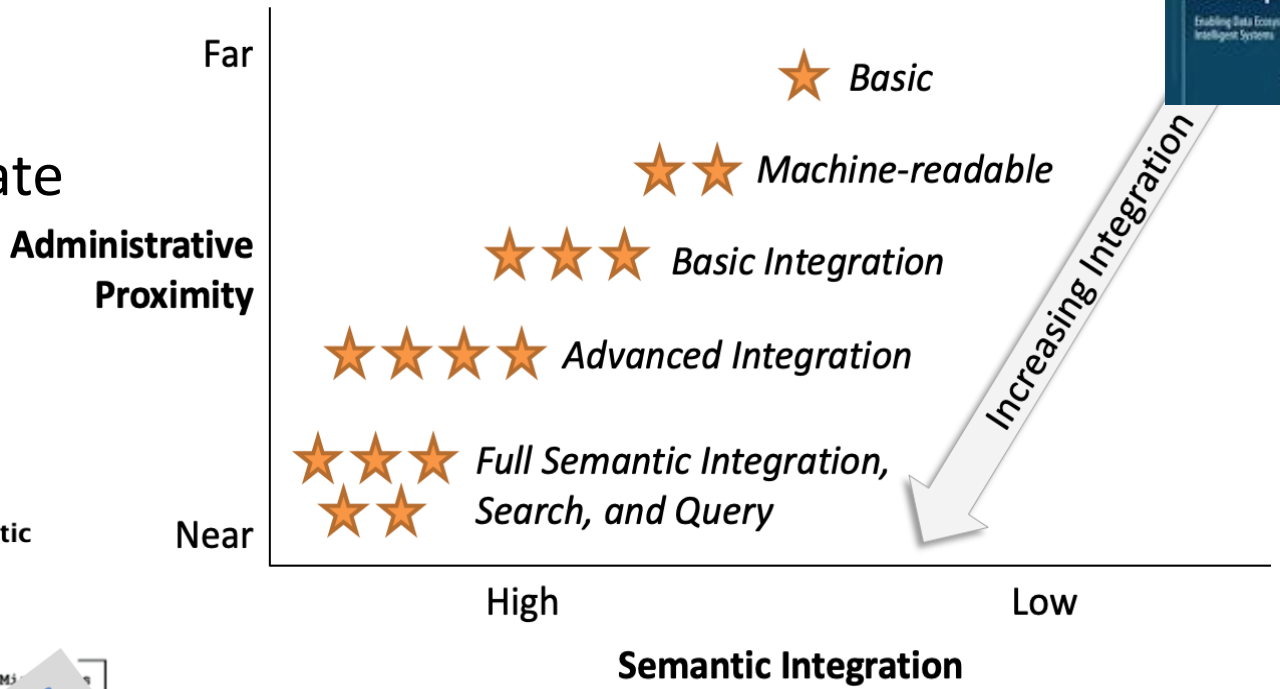


The Long Tail of Data

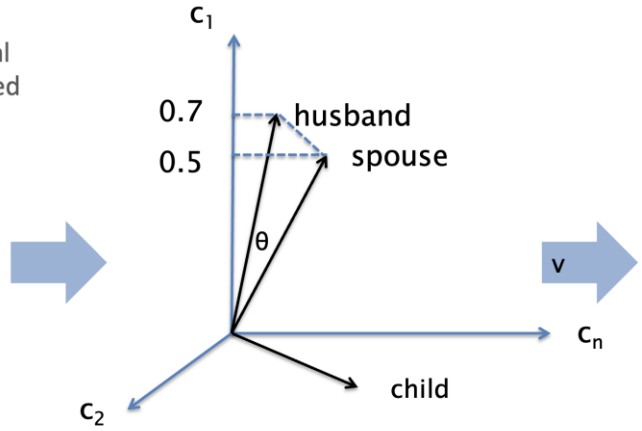
Creating Approximate Services is a Key Challenge



Investigate techniques to enable approximate and best-effort support services for loose administrative proximity and semantic integration



Distributional semantic model: Semantic statistical knowledge extracted from large Web corpora



Works as a semantic ranking function

spouse

Spouses of the Prime Mi... of Canada (0.6558)
Adultery (0.4153)
Widow (0.4095)
Alimony (0.3...
Spousal ab...
Domesti... 3292)
Righ... ities of
m... ited States
...3206)
...marriage (0.2919)
(0.27550)
...ss consort (0.2705)
D...orce (0.2383)
...

"Model for Semantic Interpretation"

E.g. dsm(spouse, widow)= 0.4095
E.g. dsm(spouse, car)= 0.009

Incremental support services

- Catalog
- Entity management
- Query and search
- Data discovery
- Human tasks
- Quality of service
- Complex event processing
- Streams dissemination
- Approximate semantic event matching

LLMs for ”Good Enough” Interoperability

User: As a Semantic Web expert, please extend the SHACL shapes given below such that I can also describe digital versions of paintings with them, using properties from the Integrated Authority File, the GND. For the painter, please use a ‘gndo:firstArtist’ that is a ‘gndo:DifferentiatedPerson’ which has its GND id as a property. For the year of creation, please use ‘gndo:dateOfProduction’.
```turtle ...
```

```turtle
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .

[] a sh:NodeShape ;
  sh:targetClass dcat:Dataset ;
  sh:property [
    sh:path dct:title ;
    sh:maxCount 1 ; ⟷ sh:minCount 1 ;
    sh:message "A dcat:Dataset must have exactly one
    ↪ dct:title." ;
  ] ;
  sh:property [
    sh:path odrl:hasPolicy ;
    sh:maxCount 1 ; ⟷ sh:minCount 1 ;
    sh:message "A dcat:Dataset must have exactly one
    ↪ odrl:hasPolicy." ;
  ] .
```

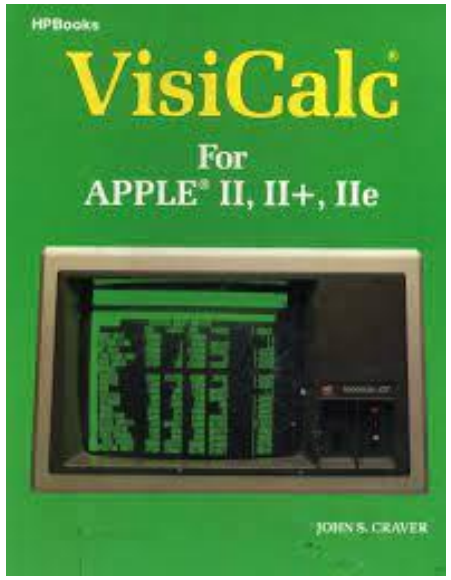
Listing 1: The base SHACL shapes.

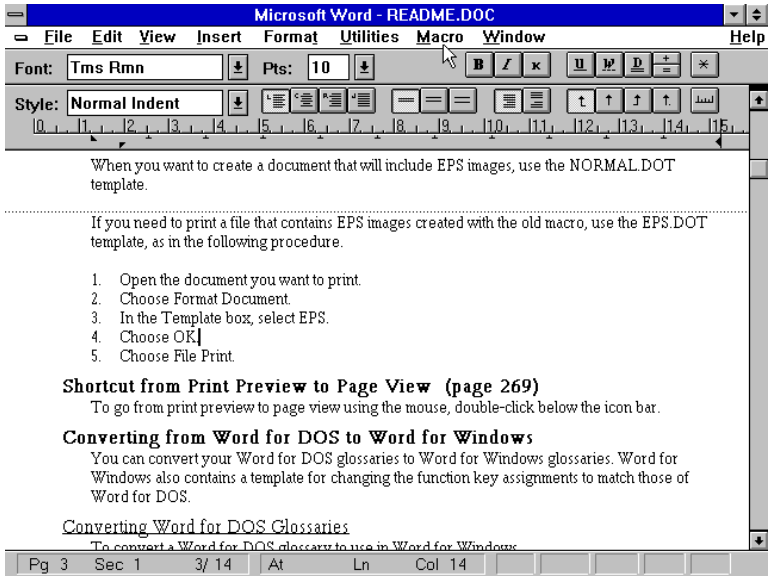

```turtle
@prefix sh: <http://www.w3.org/ns/shacl#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix odrl: <http://www.w3.org/ns/odrl/2/> .
@prefix gndo:
↪ <http://d-nb.info/standards/elementset/gnd#> .

[] a sh:NodeShape ;
  sh:targetClass dcat:Dataset ;

  ... omitted for conciseness ...

  sh:property [
    sh:path gndo:firstArtist ;
    sh:class gndo:DifferentiatedPerson ;
    sh:maxCount 1 ; ⟷ sh:minCount 1 ;
    sh:message "A digital representation of a painting
    ↪ must have exactly one gndo:firstArtist." ;
    sh:node [
      a sh:PropertyShape ;
      sh:path gndo:preferredNameForThePerson ;
      sh:maxCount 1 ; ⟷ sh:minCount 1 ;
      sh:message "Each gndo:DifferentiatedPerson must
      ↪ have exactly one
      ↪ gndo:preferredNameForThePerson." ;
    ] ;
    sh:node [
      a sh:PropertyShape ;
      sh:path gndo:gndIdentifier ;
      sh:maxCount 1 ; ⟷ sh:minCount 1 ;
      sh:message "Each gndo:DifferentiatedPerson must
      ↪ have exactly one gndo:gndIdentifier." ;
    ] ;
  ] ;
  sh:property [
    sh:path gndo:dateOfProduction ;
    sh:maxCount 1 ; ⟷ sh:minCount 1 ;
    sh:datatype xsd:date ;
    sh:message "A digital representation of a painting
    ↪ must have exactly one gndo:dateOfProduction." ;
  ] .
```

Listing 2: The SHACL shapes with extensions related to paintings, by GPT-4.

Data Spaces will require a unified Data and AI Lifecycle if we are to maximise the potential of Generative AI

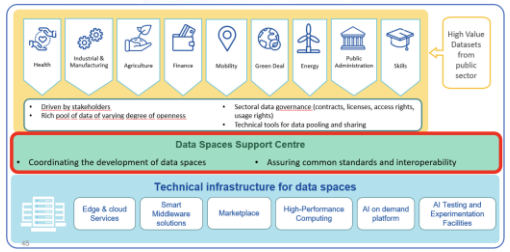
Generative AI and Foundation models will be the Killer App for Data Spaces….



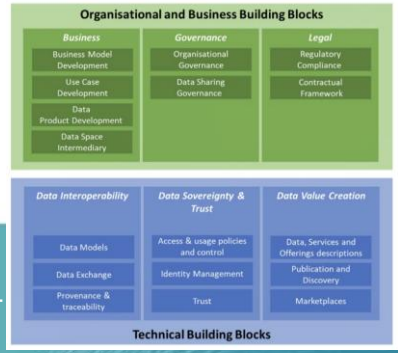


Insight

Symbiotic Relationship between Data Spaces and AI….



DATA SPACES SUPPORT CENTRE



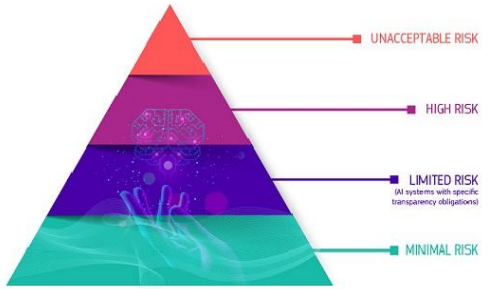
High Quality Training Data
Community of Data Users/Owners
Data Goverance/compliance

Provides training data
(Data Space for AI)

Data Space ⇄ **AI**

Provides models to enable
(AI for Data Space)

Tools for Diverse Data Space Tasks
Faster Value Extraction from Multimodal Data
Knowledge Foundation for Long-Tail Semantics

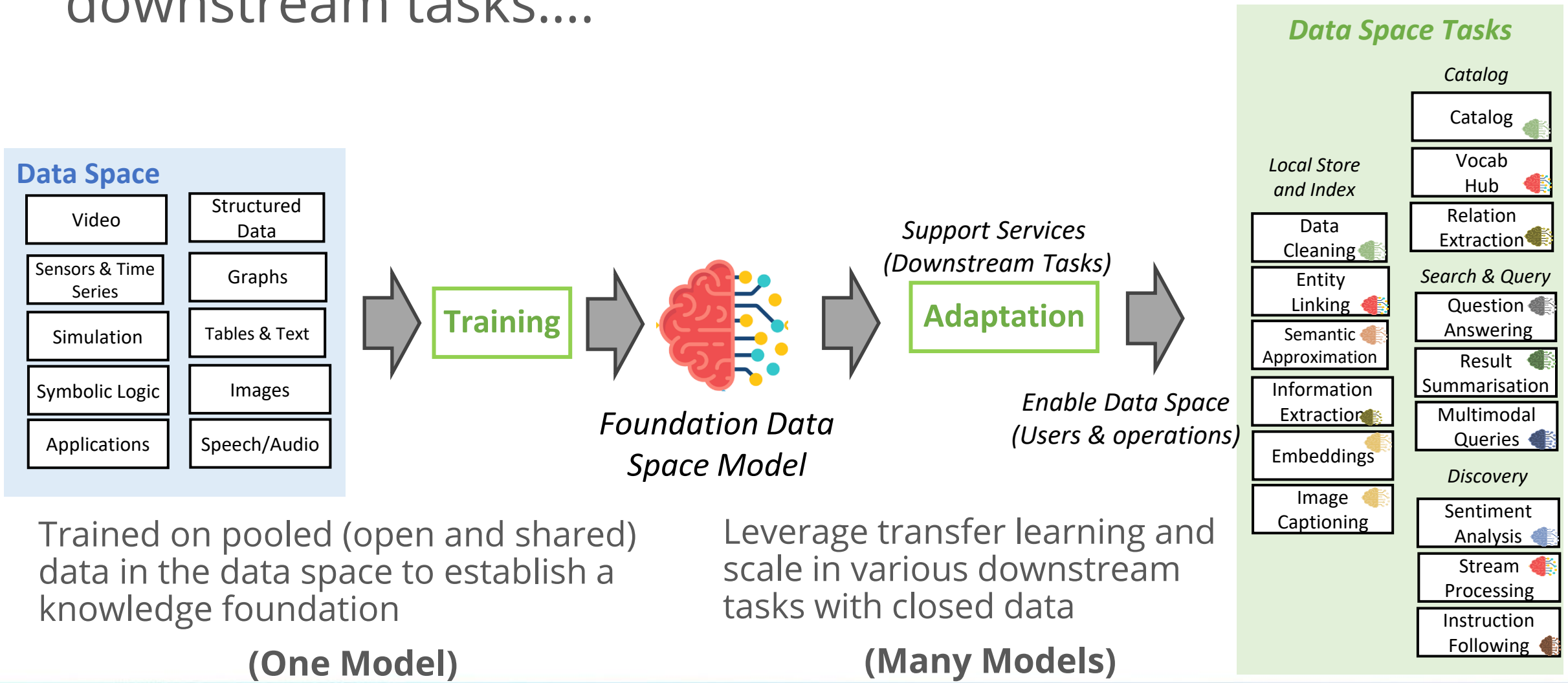


The AI Data Robotics Association

E. Curry, M. Timilsina, T. Zaarour, M. Al-QATF, R. Haque, “Foundational Data Space Models: Bridging the AI and Data Ecosystems (Vision Paper), Proceedings of the *2023 IEEE International Conference on Big Data (Big Data)*, Sorrento, Italy, 2023

Insight

Data Spaces for GenAI: Centralise information from the data space, then adapt to a wide range of downstream tasks….



Trained on pooled (open and shared) data in the data space to establish a knowledge foundation

(One Model)

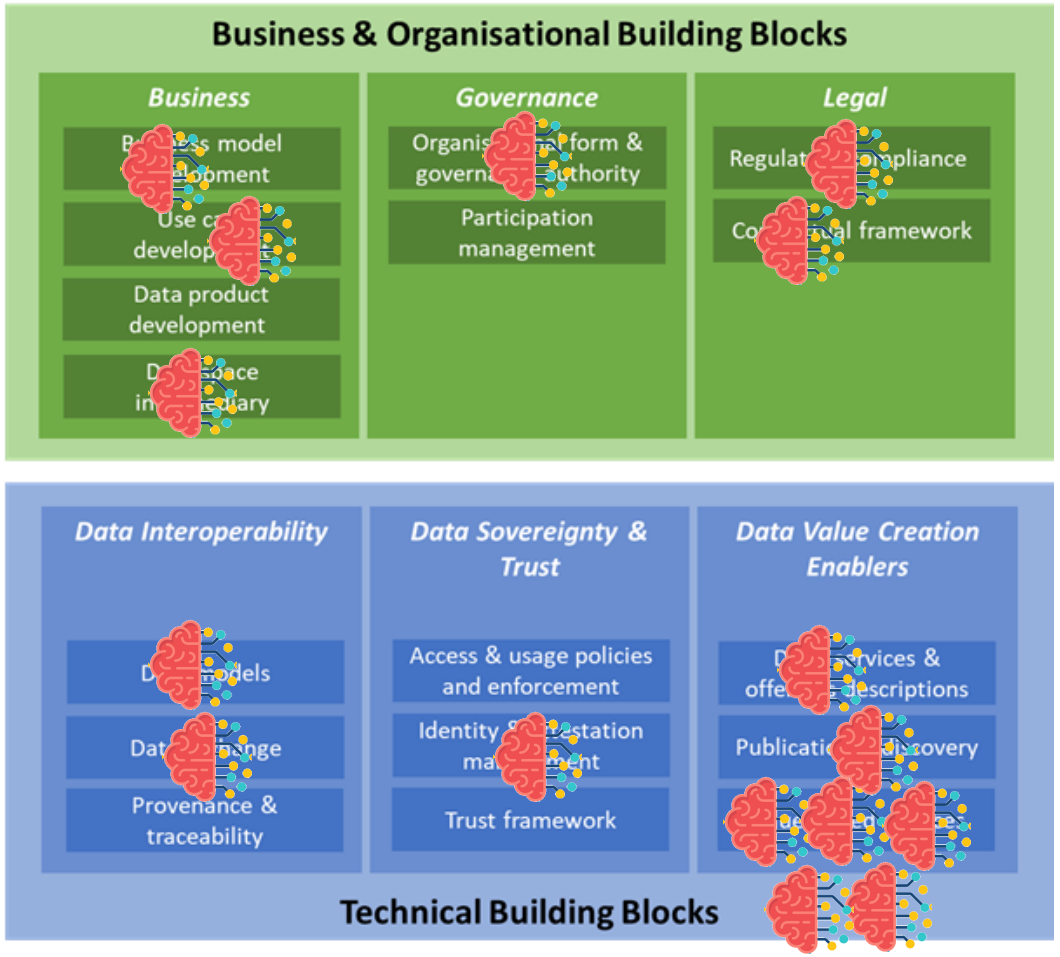
Leverage transfer learning and scale in various downstream tasks with closed data

(Many Models)

E. Curry, M. Timilsina, T. Zaarour, M. Al-QATF, R. Haque, "Foundational Data Space Models: Bridging the AI and Data Ecosystems (Vision Paper), Proceedings of the *2023 IEEE International Conference on Big Data (Big Data)*, Sorrento, Italy, 2023

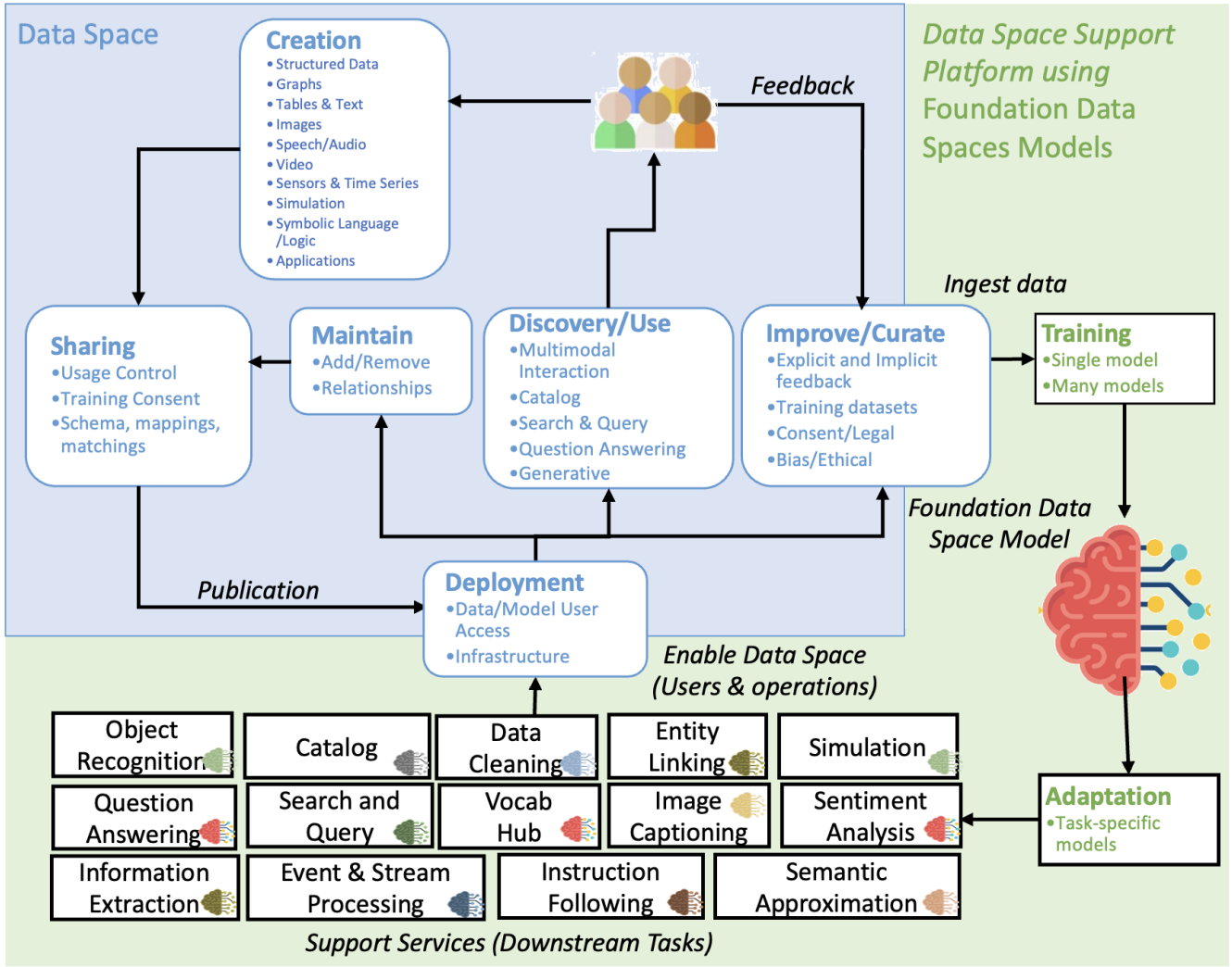
GenAI for Data Spaces: *Supports tasks across the life cycle of the data space…*

DATA SPACES SUPPORT CENTRE



Insight

We need Unified Data and AI Lifecycles



A unified lifecycle for data and AI models recognizes the **symbiotic relationship between both ecosystems** and can serve as the basis to simplify the development, operation, and use of data-intensive AI systems.

E. Curry, M. Timilsina, T. Zaarour, M. Al-QATF, R. Haque, "Foundational Data Space Models: Bridging the AI and Data Ecosystems (Vision Paper), Proceedings of the *2023 IEEE International Conference on Big Data (Big Data)*, Sorrento, Italy, 2023