A World Leading SFI Research Centre

Science Foundation Ireland sfi For what's next

Semantic and AI for Data Management

Prof. Edward Curry

University of Galway

Insight SFI research Centre for Data Analytics

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

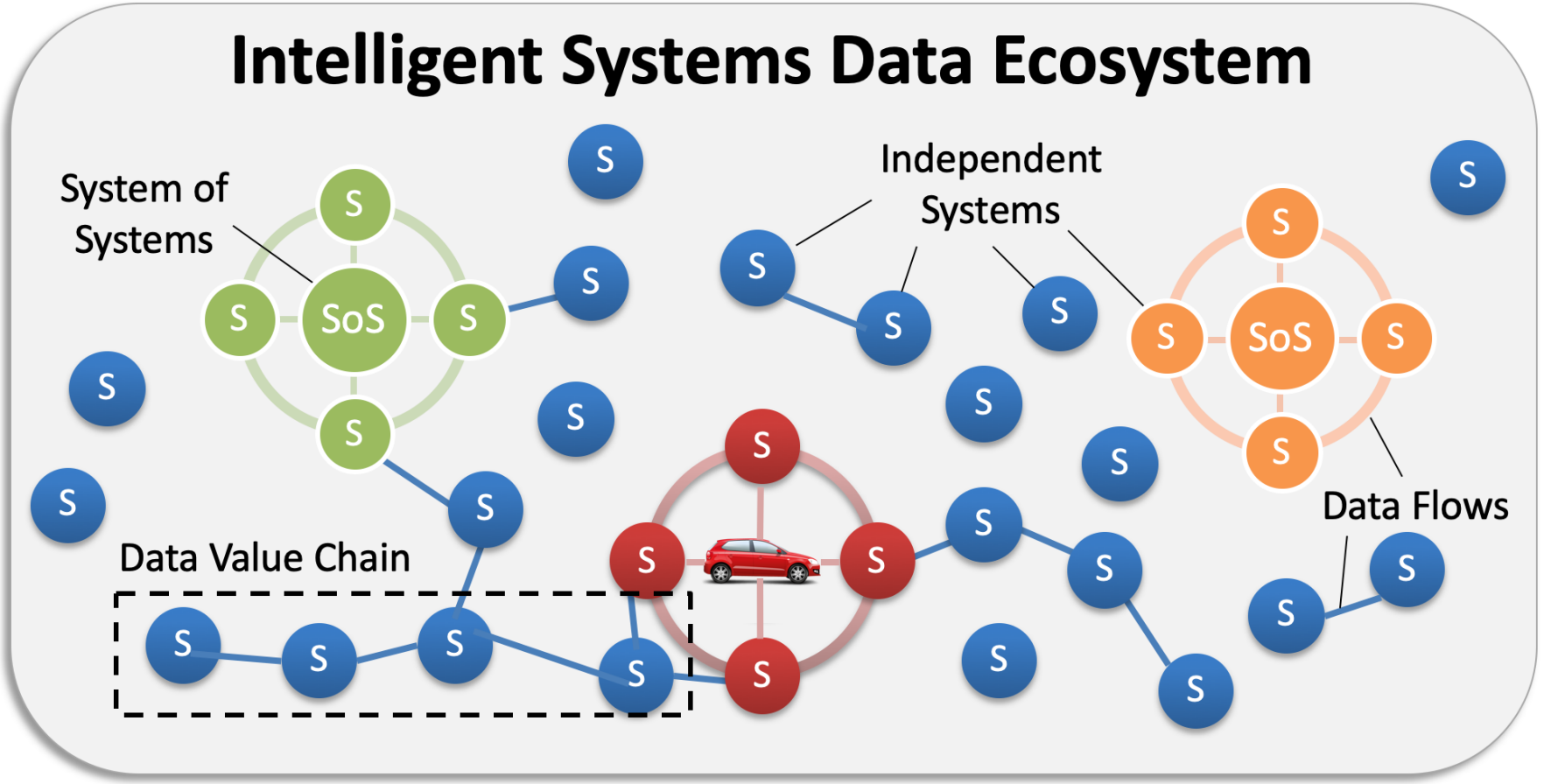
University of Limerick

FUNDED BY:

European Union European Regional Development Fund

Science Foundation Ireland sfi For what's next

Distributed and Decentralised Data Ecosystems



Key Barrier: Interoperability – Protocols and **Semantics**

Curry, E. and Sheth, A. (2018) 'Next-Generation Smart Environments: From System of Systems to Data Ecosystems', *IEEE Intelligent Systems*, 33(3), pp. 69–76. doi: 10.1109/MIS.2018.033001418.

Formal World

Real World

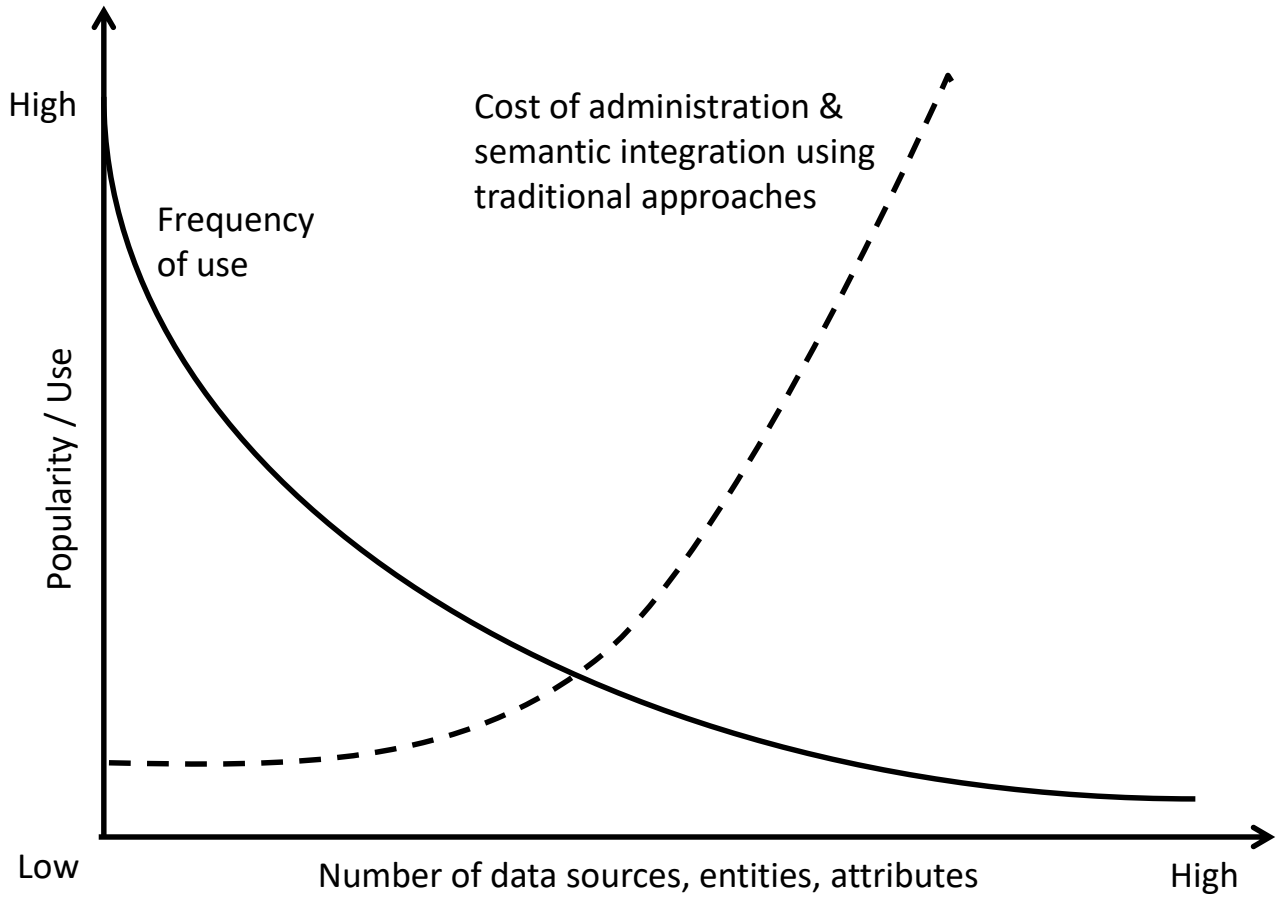
Semantics for a Complex World

- Most semantic models have dealt with particular types of constructions, and have been carried out under very simplifying assumptions under lab conditions.
- If these idealisations are removed it is not clear at all that modern semantics can give a full account of all but the simplest models/statements.

Baroni et al. 2013

Sahlgren, 2013

Traditional Approaches to Data Integration



The Long Tail of Data

http://dataspaces.info

Content Space: From Rigid Schemas to Schema-less.....

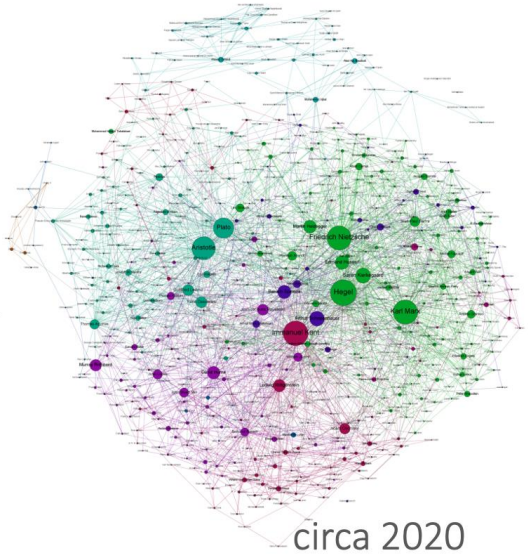
Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

- Heterogeneous, complex and large-scale data
- Very-large and dynamic “schemas”
- Open Environments: distributed, decentralised decoupled data sources, anonymous users, multi-domain, lack of global order of information flow

circa 2000

10s-100s attributes

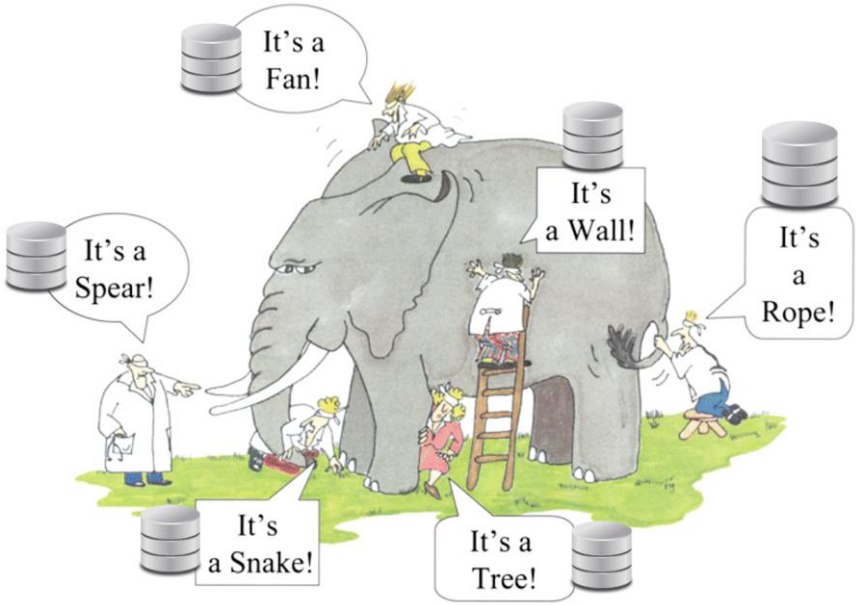
EMP_NO	FIRST_NAME	LAST_NAME	PHONE_EXT	HIRE_DATE	DEPT...	JOB_C...	JOB_GR...	JOB_COUNT...	SALARY	FULL_NAME
2	Robert	Nelson	250	12.28.1988 12:00 am	600	VP	2	USA	105.900,00	Nelson, Robert
4	Bruce	Young	233	12.28.1988 12:00 am	621	Eng	2	USA	97.500,00	Young, Bruce
5	Kim	Lambert	22	02.06.1989 12:00 am	130	Eng	2	USA	102.750,00	Lambert, Kim
8	Leslie	Johnson	410	04.05.1989 12:00 am	180	Mktg	3	USA	64.635,00	Johnson, Leslie
9	Phil	Forest	229	04.17.1989 12:00 am	622	Mngr	3	USA	75.060,00	Forest, Phil
11	K. J.	Weston	34	01.17.1990 12:00 am	130	SRep	4	USA	86.292,94	Weston, K. J.
12	Terri	Lee	256	05.01.1990 12:00 am	000	Admin	4	USA	53.793,00	Lee, Terri
14	Stewart	Hall	227	06.04.1990 12:00 am	900	Finan	3	USA	69.482,63	Hall, Stewart
15	Katherine	Young	231	06.14.1990 12:00 am	623	Mngr	3	USA	67.241,25	Young, Katherine
20	Chris	Papadopoulos	887	01.01.1990 12:00 am	671	Mngr	3	USA	89.655,00	Papadopoulos, Chr
24	Pete	Fisher	888	09.12.1990 12:00 am	671	Eng	3	USA	81.810,19	Fisher, Pete
28	Ann	Bennet	5	02.01.1991 12:00 am	120	Admin	5	England	22.935,00	Bennet, Ann
29	Roger	De Souza	288	02.18.1991 12:00 am	623	Eng	3	USA	69.482,63	De Souza, Roger
34	Janet	Baldwin	2	03.21.1991 12:00 am	110	Sales	3	USA	61.637,81	Baldwin, Janet
36	Roger	Reeves	6	04.25.1991 12:00 am	120	Sales	3	England	33.620,63	Reeves, Roger
37	Willie	Stansbury	7	04.25.1991 12:00 am	120	Eng	4	England	39.224,06	Stansbury, Willie
44	Leslie	Phong	216	06.03.1991 12:00 am	623	Eng	4	USA	56.034,38	Phong, Leslie
45	Ashok	Ramanathan	209	08.01.1991 12:00 am	621	Eng	3	USA	80.689,50	Ramanathan, Ashc
46	Walter	Steadman	210	08.09.1991 12:00 am	900	CFO	1	USA	116.100,00	Steadman, Walter
52	Carol	Nordstrom	420	10.02.1991 12:00 am	180	PRel	4	USA	42.742,50	Nordstrom, Carol
61	Luke	Leung	3	02.18.1992 12:00 am	110	SRep	4	USA	68.805,00	Leung, Luke
65	Sue Anne	O'Brien	877	03.23.1992 12:00 am	670	Admin	5	USA	31.275,00	O'Brien, Sue Anne



circa 2020

1,000s-1,000,000s attributes

- Multiple perspectives (conceptualisations) of the reality.
- Ambiguity, vagueness, inconsistency.



...and Fundamental Decentralisation

Increasing amounts of Subjective Data

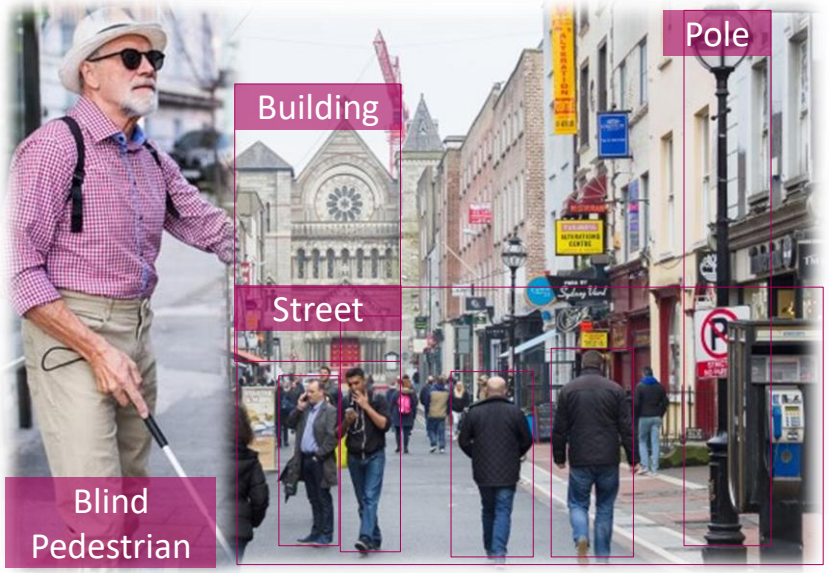
Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

"The very concept of objective truth is fading out of the world."

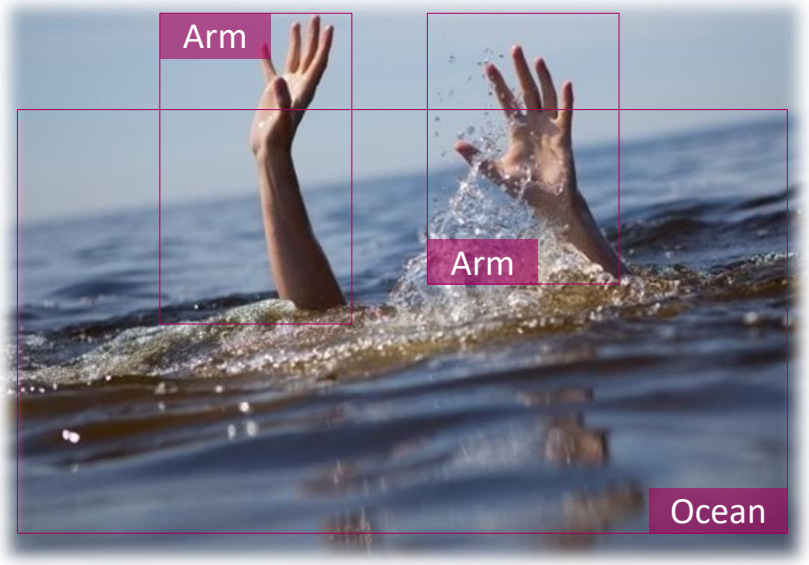
George Orwell

Image Credit: https://www.yesvedanta.com/objectivity-vs-subjectivity-in-light-of-vedanta/

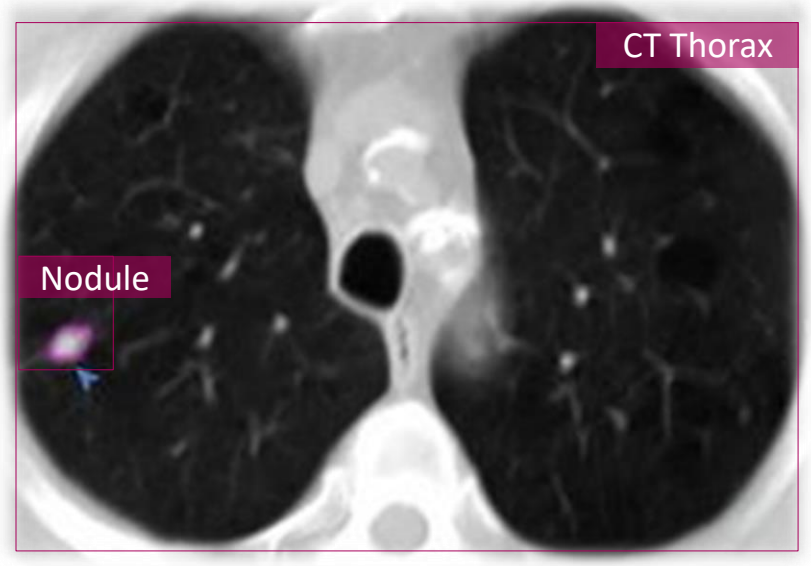
Increasing Amounts of Multimodal Data….



(More than 220k people in Ireland are blind or visually impaired)



(An average of 110 drownings occur in Ireland every year)



(Every 3 minutes in Ireland someone gets a cancer diagnosis; every hour someone dies from cancer)

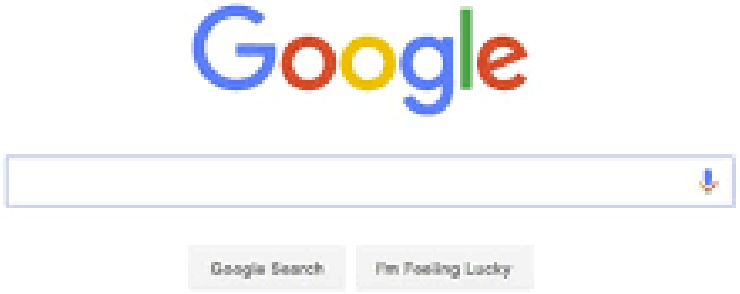
.J. Khan, J.G. Breslin, E. Curry, “Expressive Scene Graph Generation using Commonsense Knowledge Infusion for Visual Understanding and Reasoning,” Extended Semantic Web Conference (ESWC) 2022, Crete, Greece, May-June 2022.

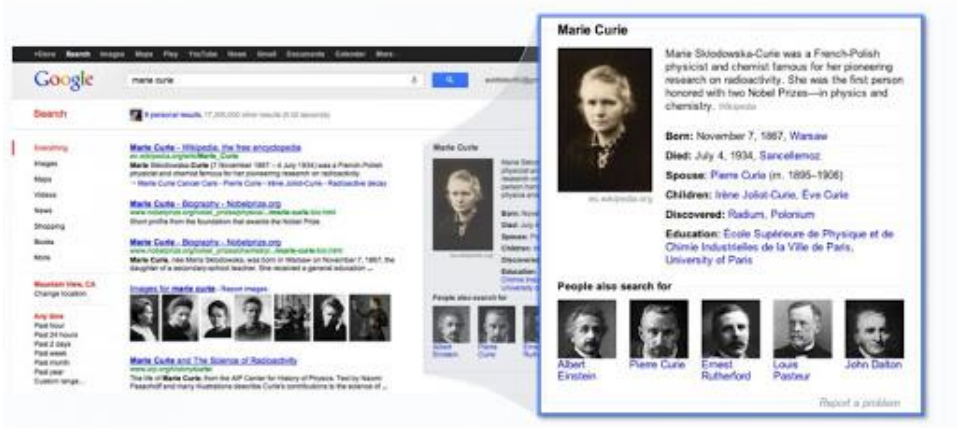
The Red Queen Hypothesis

“It takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!”

Lewis Carroll's Through the Looking-Glass

Human Interactivity: Web Search





From Structure

~1995
~100K Websites
Exact Results
Human Curated

to Search

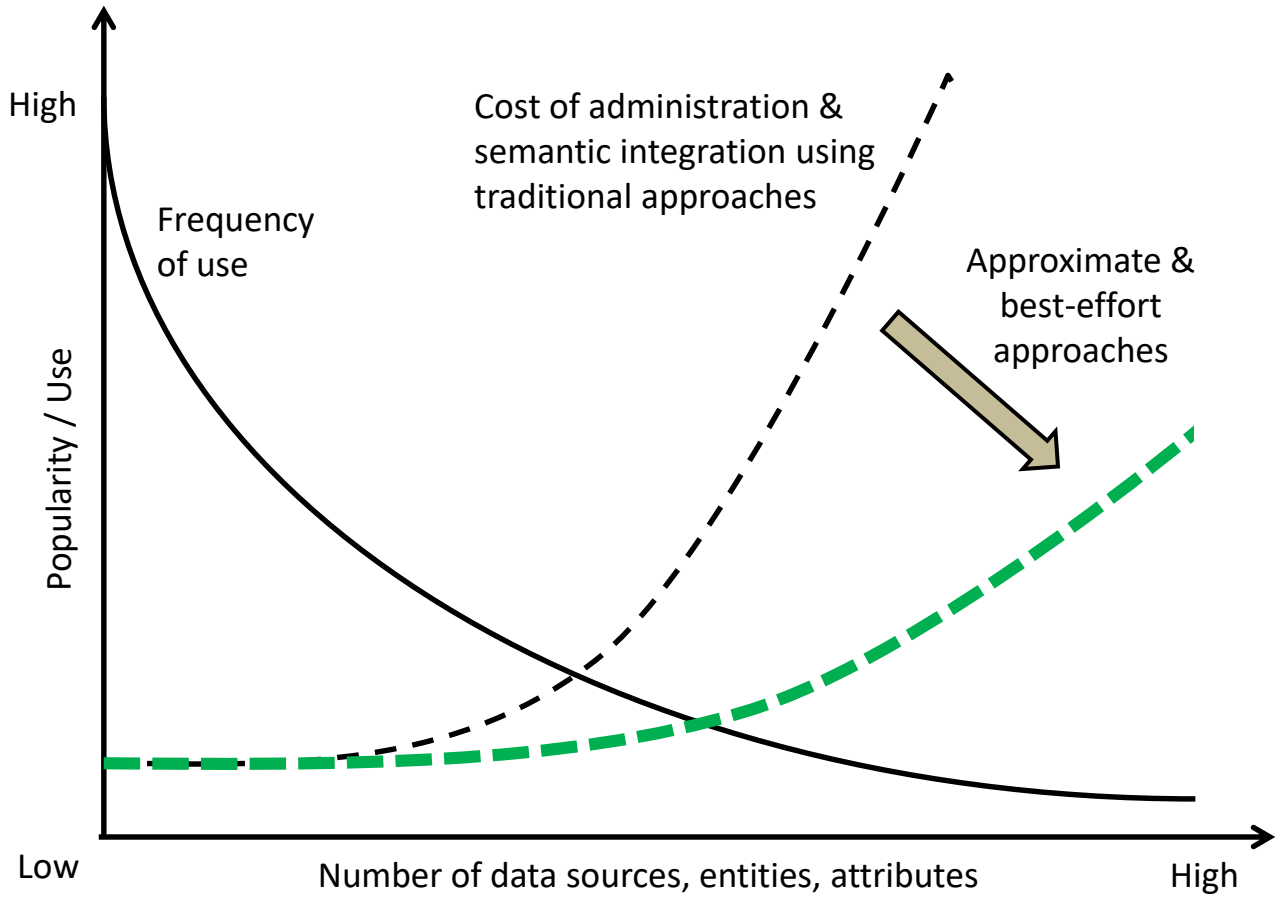
~1998
~2.4M Websites
Approximate Results
Computed

to Knowledge Graph

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

Approximate and Best Effort Approaches

Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

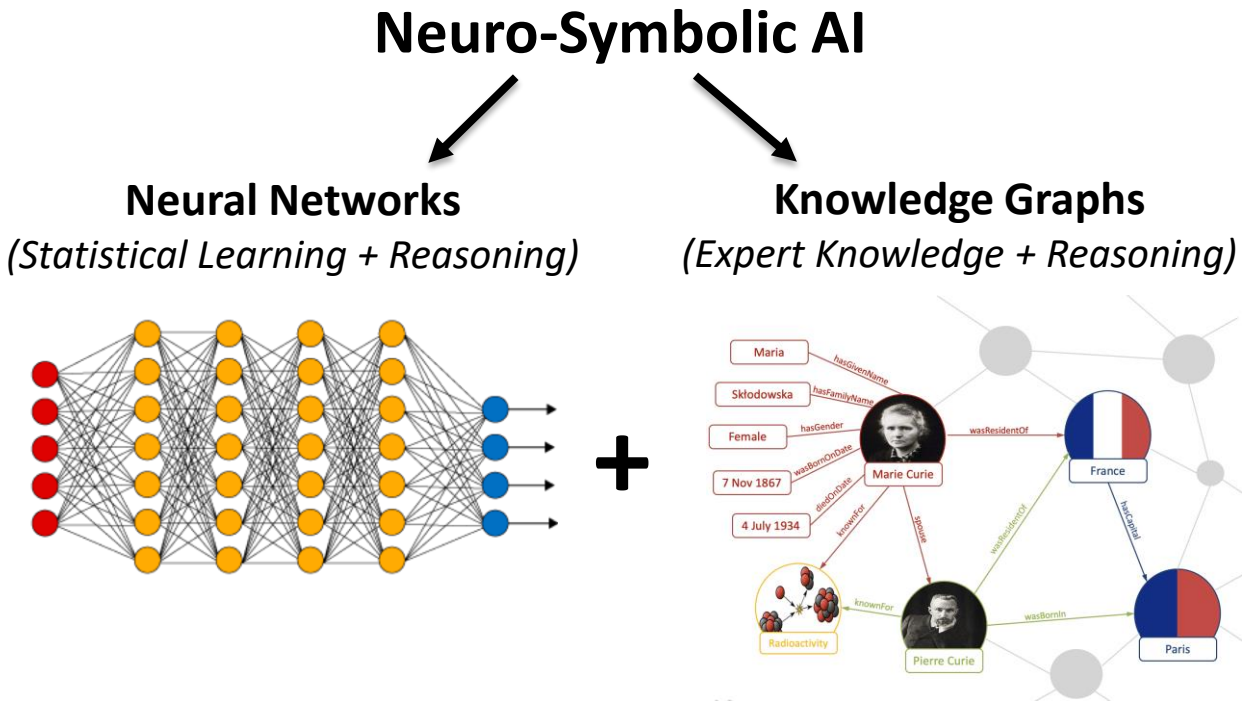


The Long Tail of Data

Hybrid Neuro-Symbolic Approaches

Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

- Combing rules-based AI approaches (Knowledge Graphs) with statistical learning techniques (deep learning)



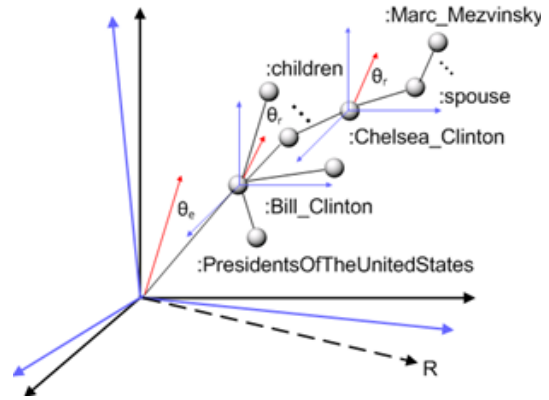
Computational Semantics

Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS

Language Model

Text output

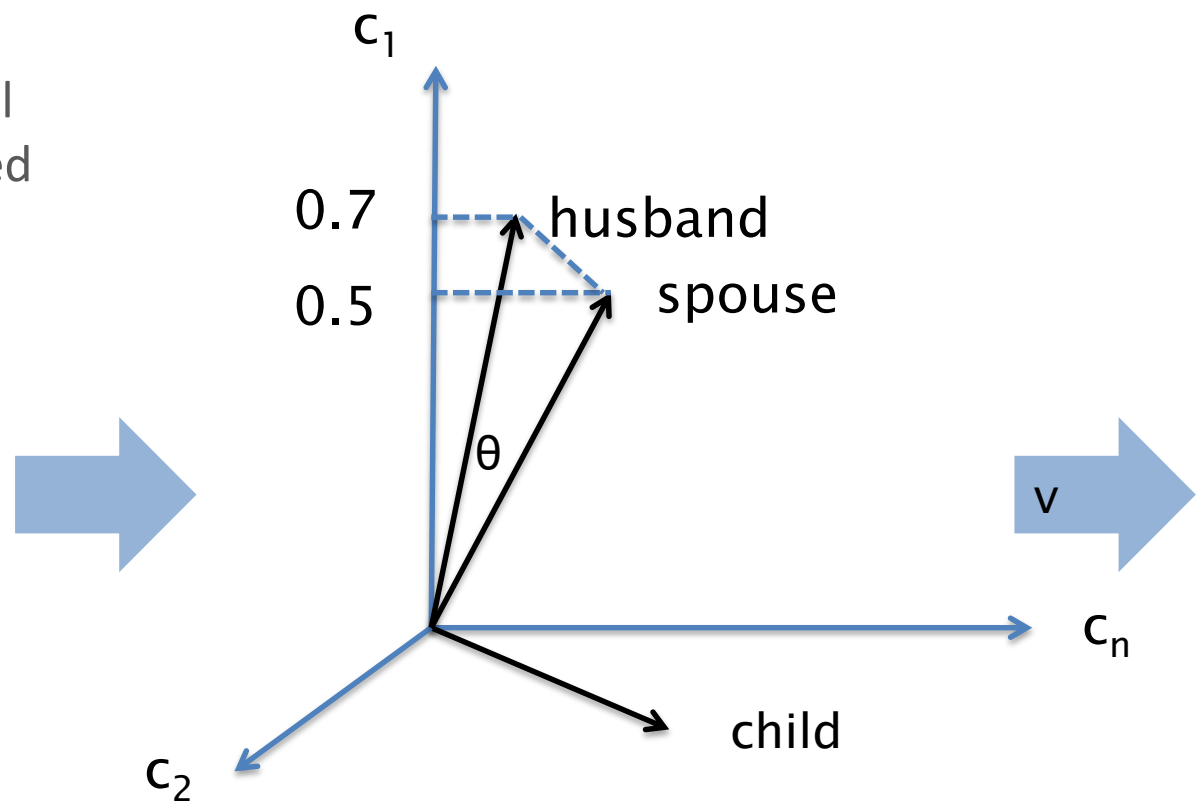
Language Generation



Numeric representation of text

Distributional Semantic Model

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 (0.37...
Spousal ab...
Domesti... 3292)
Righ... ities of
ma... ited States
...3206)
...marriage (0.2919)
...(0.27550)
...ss consort (0.2705)
D...orce (0.2383)
...

"Model for Semantic Interpretation"

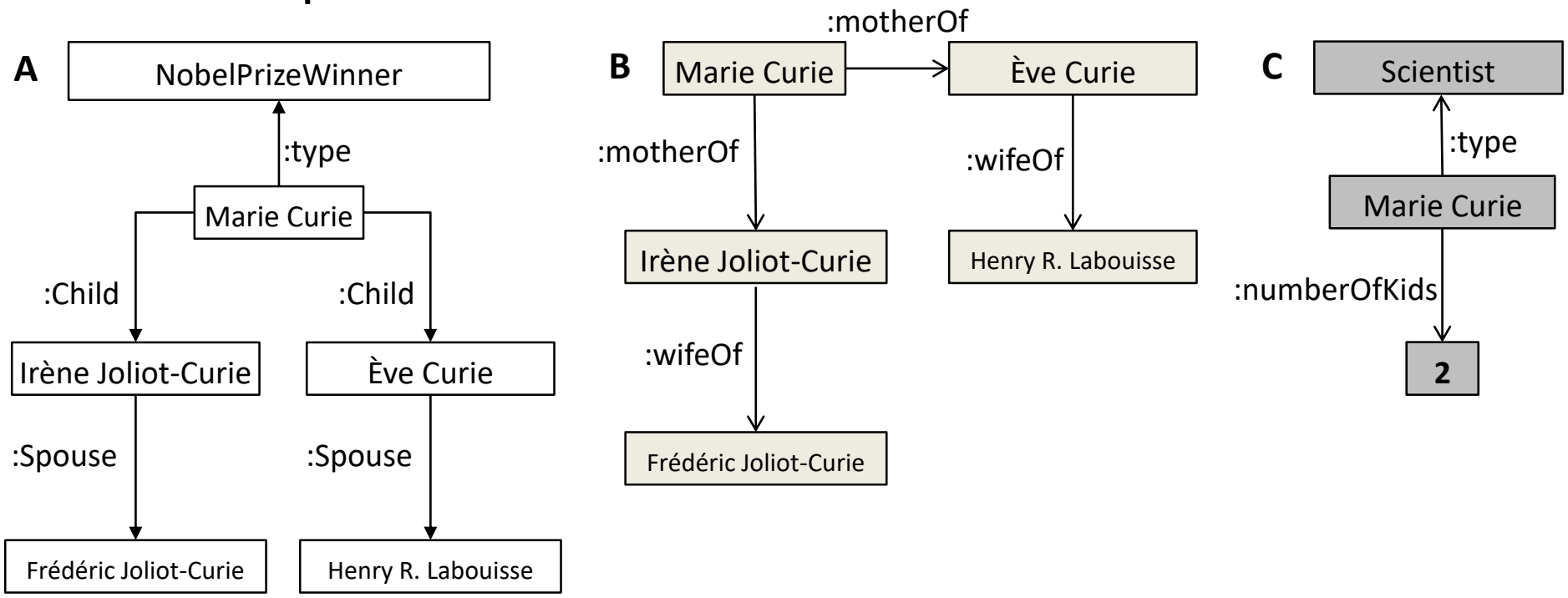
E.g. dsm(spouse, widow)= 0.4095
E.g. dsm(spouse, car)= 0.009

Schema-Agnostic Natural Language Queries

Information Need: Who are the children of Marie Curie married to?

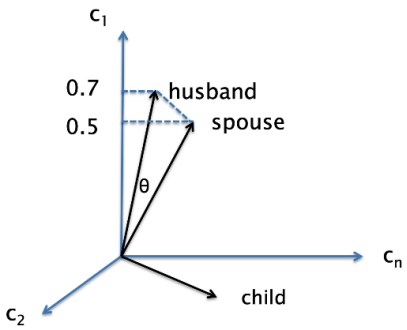
Semantic Gap

Possible Data Representations

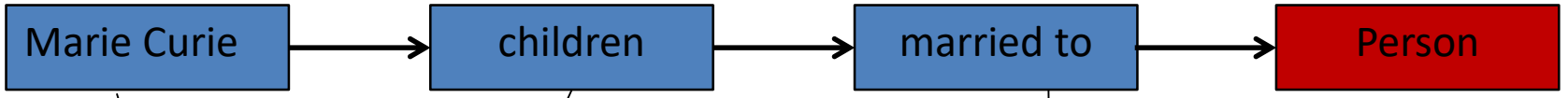


Distributional Semantic Search

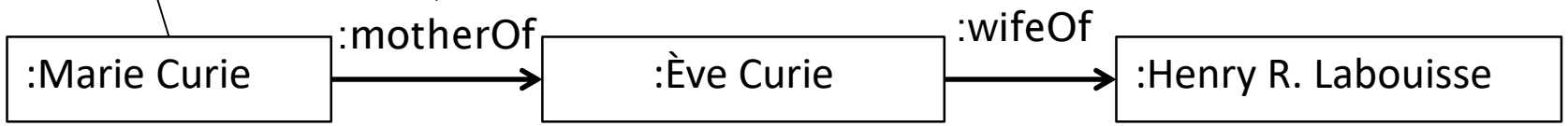
Information Need: Who are the children of Marie Curie married to?



Query:

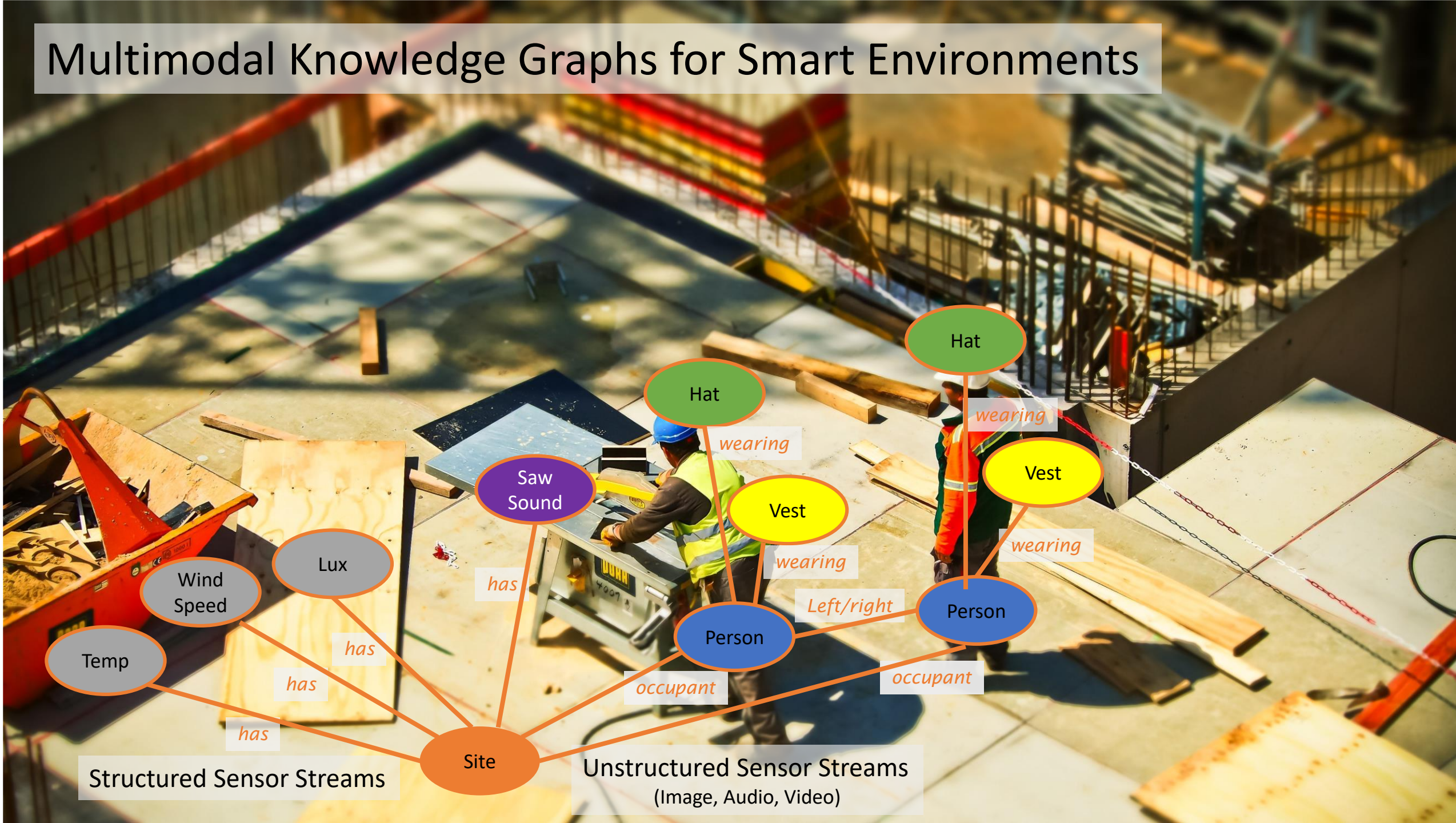


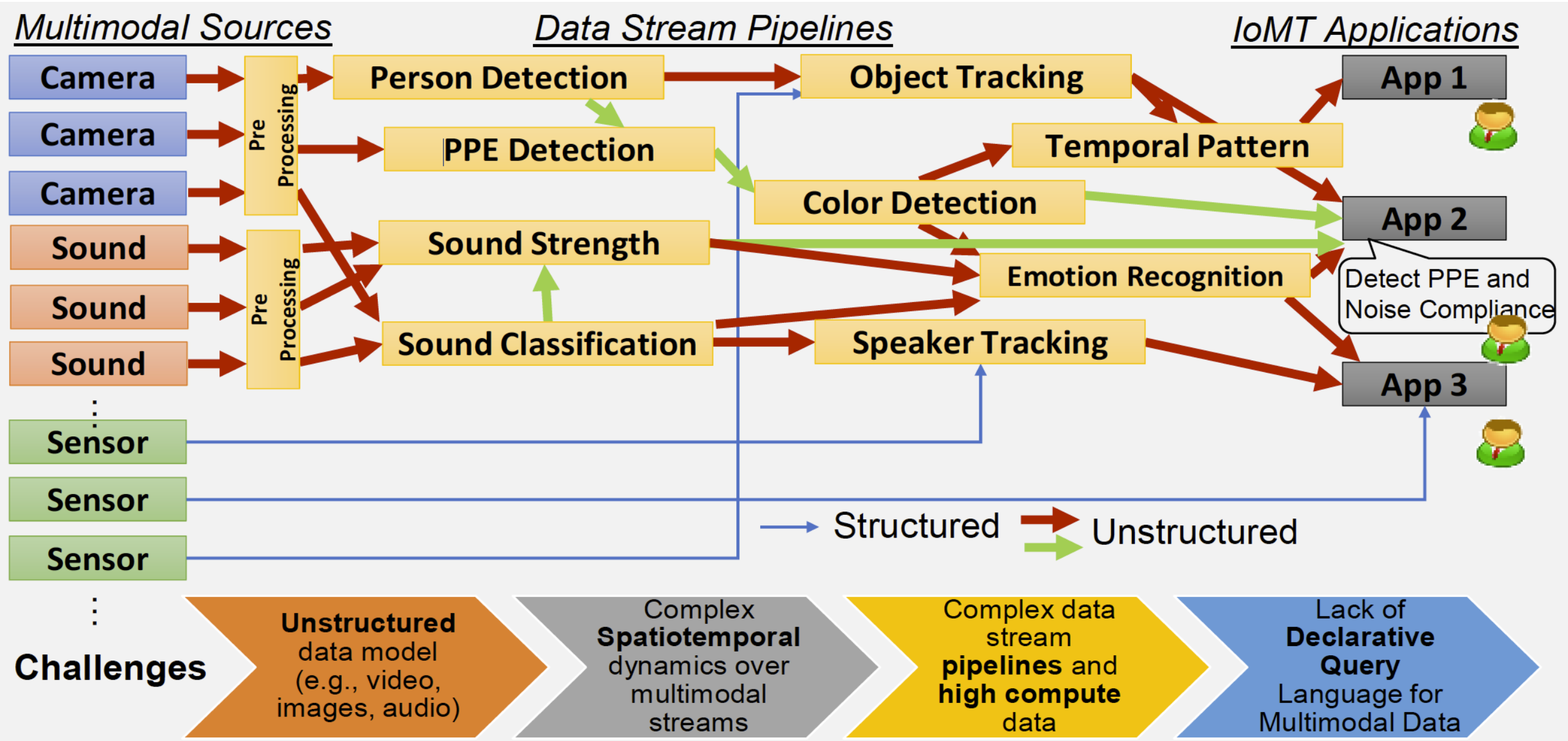
Graph Data:



Semantic model bridges semantic gap between different systems concepts

Multimodal Knowledge Graphs for Smart Environments



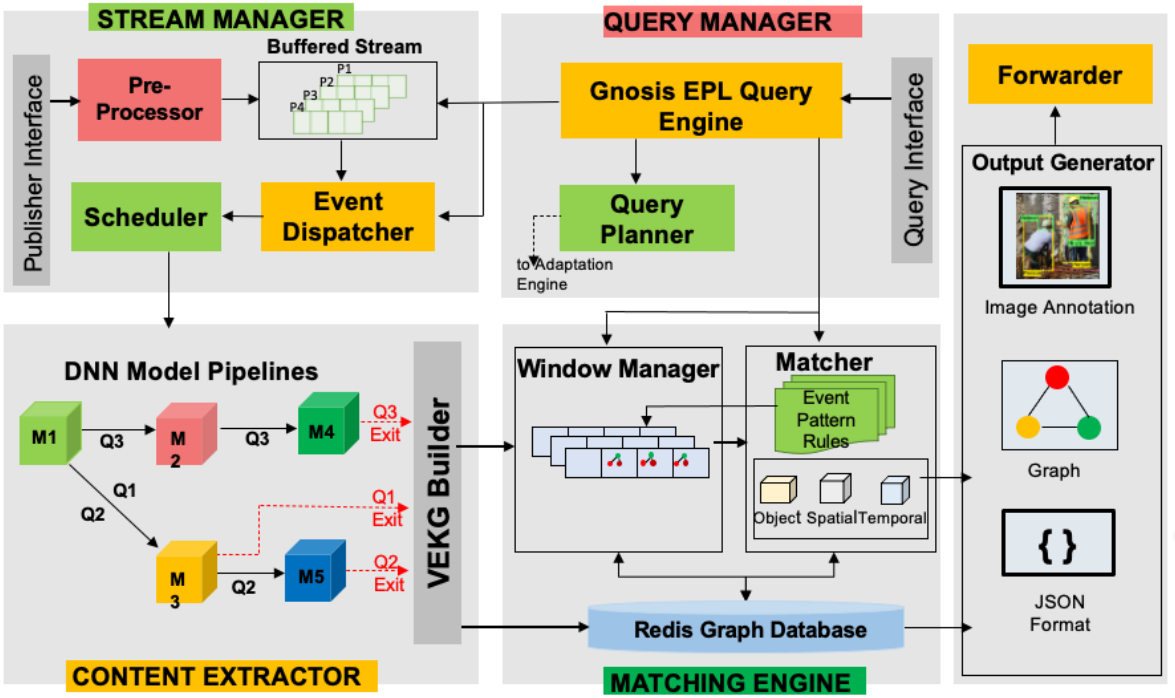

Multimodal Sources
Data Stream Pipelines
IoMT Applications
Camera
Camera
Camera
Sound
Sound
Sound
Sensor
Sensor
Sensor
Pre Processing
Pre Processing
Person Detection
PPE Detection
Sound Strength
Sound Classification
Object Tracking
Temporal Pattern
Color Detection
Emotion Recognition
Speaker Tracking
App 1
App 2
App 3
Detect PPE and Noise Compliance
Structured
Unstructured
Challenges
Unstructured data model (e.g., video, images, audio)
Complex Spatiotemporal dynamics over multimodal streams
Complex data stream pipelines and high compute data
Lack of Declarative Query Language for Multimodal Data

Multimodal Event Processing: A Neural-Symbolic Paradigm for the Internet of Multimedia Things

IoMT Sources

Camera
Camera
Camera
Sound
Sound
Sound
⋮
Sensor
Sensor
Sensor
⋮

Gnosis



IoMT Applications

Query 1

```
Q1: REGISTER QUERY CountHardHatWorkers
OUTPUT ANN_IMAGE_BBOX, ANN_IMAGE_QUERY_OUTPUT
CONTENT PPEDetection
MATCH (worker_hat:HAT)
FROM video.mp4
RETURN
COUNT(worker_hat) as WorkerCount
```

Simple SQL-like query

Query 2

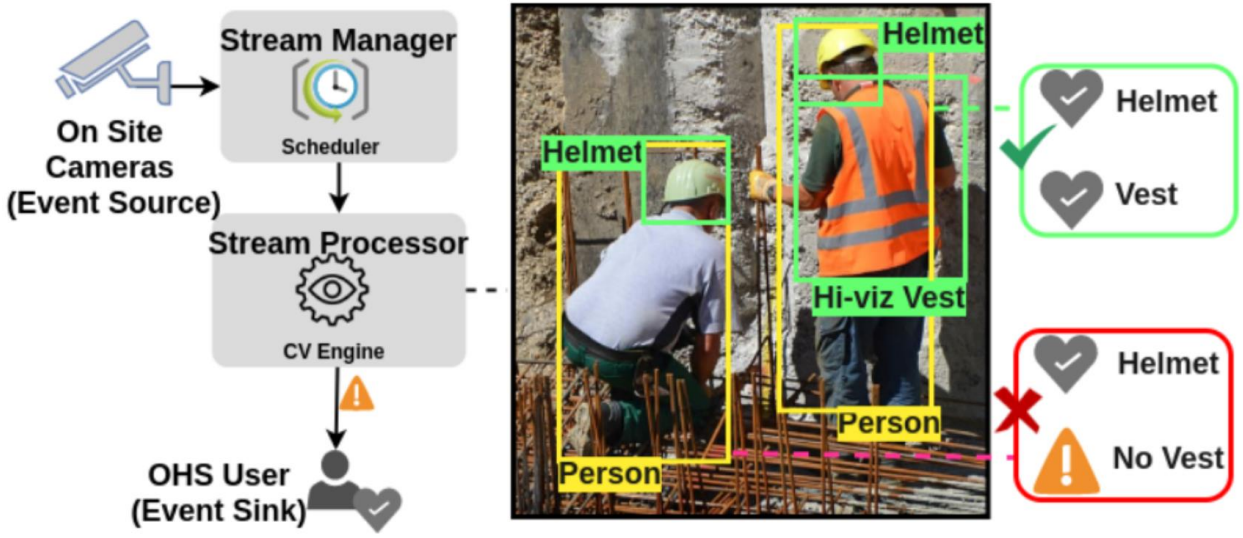
Neuro-Symbolic Processing of Structured and Unstructured Data

E. Curry, D. Salwala, P. Dhingra, F. A. Pontes and P. Yadav, "Multimodal Event Processing: A Neural-Symbolic Paradigm for the Internet of Multimedia Things," in *IEEE Internet of Things Journal*, vol. 9, no. 15, 2022

Structured Flows Multimedia Flows

Reducing Energy usage in Edge AI

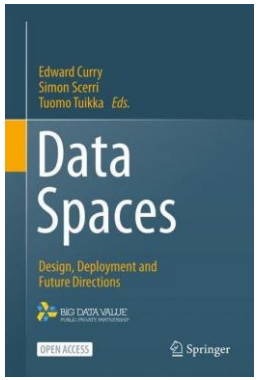
Insight
SFI RESEARCH CENTRE FOR DATA ANALYTICS



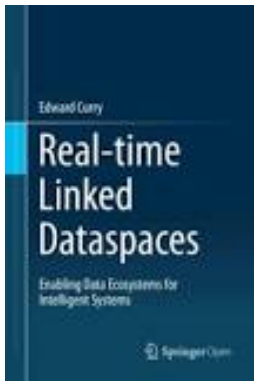
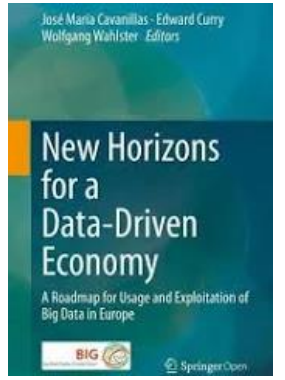
Energy-Speed-Accuracy Trade-off Triangle for the Query Quality of Service
(F.Pontes et. Al. 2023)

Conclusions

- Common European Data Spaces will require new data management paradigms in order to facilitate large-scale data pooling and sharing



- Semantics have never been more challenging
  - Scale, content space, multimodal, subjectivity, trust, cost, …



- Language models and hybrid Neuro-Symbolic approaches provide promising avenues to develop new paradigms
  - Computationally driven
  - Knowledge Graphs + Deep Learning
  - Exact + Approximate best effort results