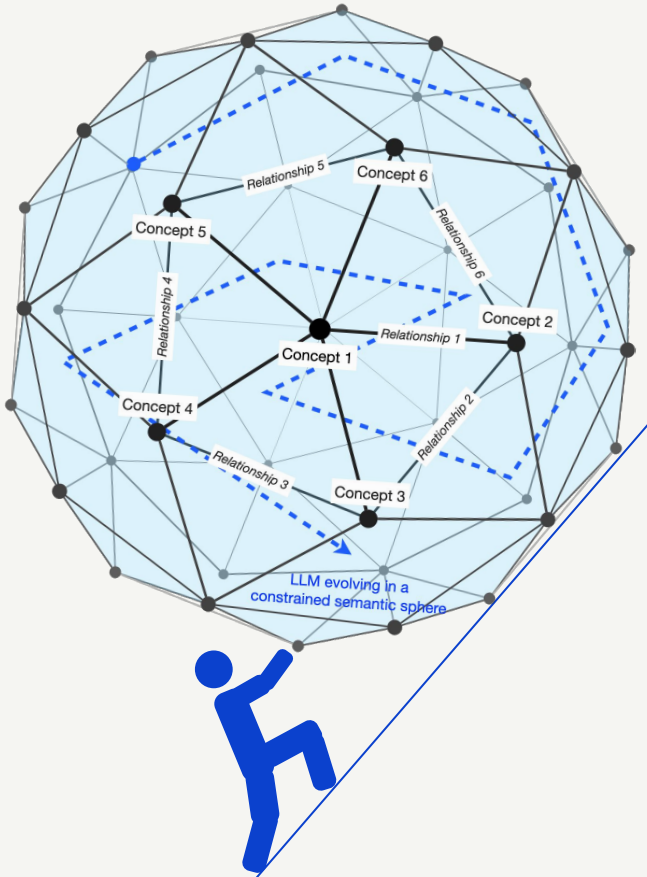
Harnessing Digital Complexity and Enabling AI

Semantic Knowledge Graph as Universal Enablers



Knowledge Graph Forum 2024 - Roche Basel

Cedric Berger, Head of Knowledge Extraction and Integration
Roche, Technical Operations

Table of contents

Context

Complexity Everywhere

At what point is it too much for human brains?

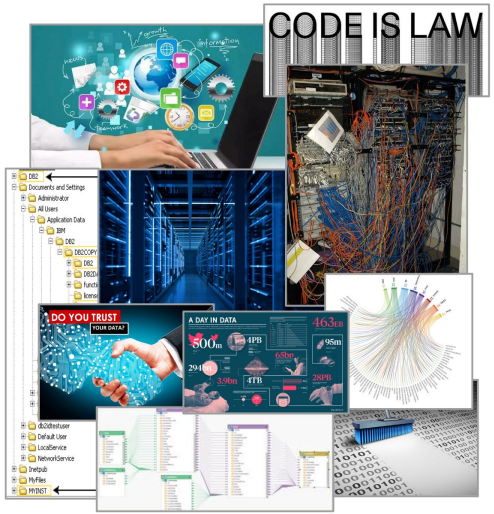
Did you notice?

The world is becoming more complex and we understand less of it everyday.

Global complexity

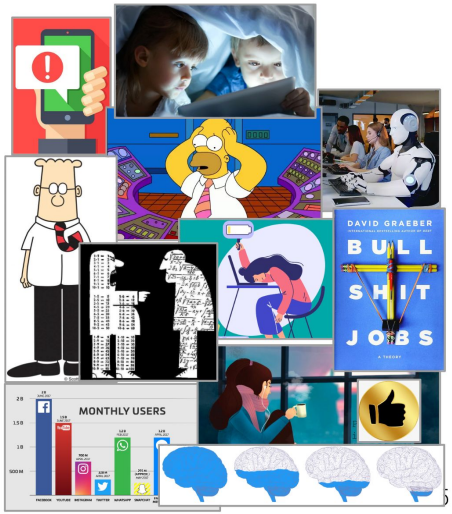
$\in$

Digital complexity



$\Rightarrow$

Individual issues



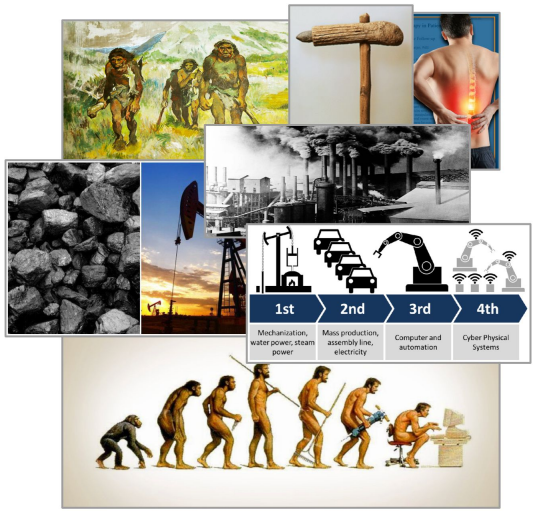
Digital Complexity

To whom benefits the situation?

Did you notice?

We are overwhelmed by a data/information deluge due to the proliferation of IT systems (soft) and infrastructure (hard).

Techno-messianism

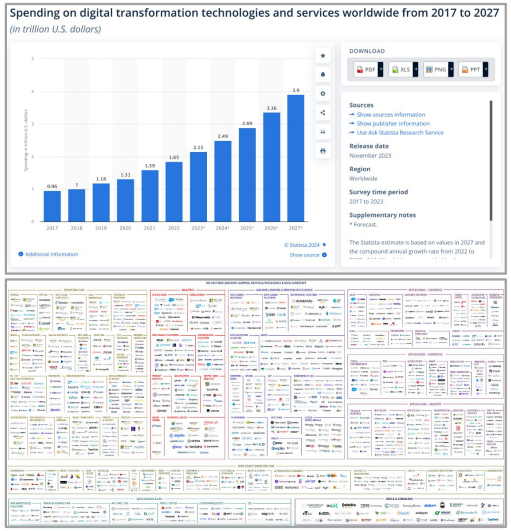


+

Economic liberalism

=

Lucrative digital transformation business



If Data is the New Oil, why Focusing on Pumps and Pipes?

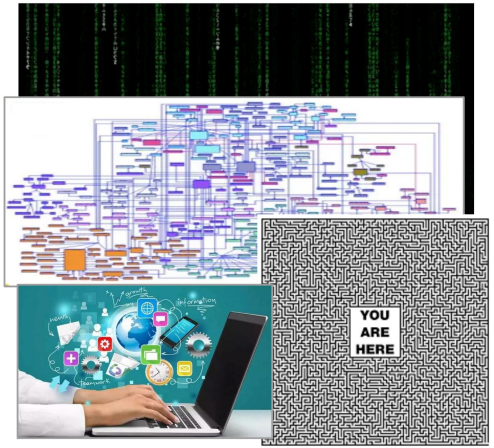
Data mess, not mesh

Did you notice?

In the 90's, nobody got fired for buying IBM, in the 2000's eCommerce, in the 2010's cloud, nowadays AI. Where is the data?

System-centricity

Data-complexity



Business-perplexity

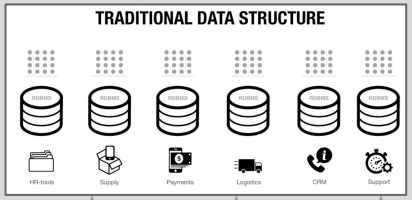
Harnessing Digital Complexity

WHAT: Knowledge Graphs to Capture and Link Anything

Universal linkers

Effective or operational layer - The hands

- Where daily data operations are performed
- Un-, semi- and -structured data
- Your main platforms, tools for transactional, analysis, reporting…



Descriptive or reference layer - The brain

- Where daily data operations are described
- Graph data
- Governance (prescription)
- (Meta) analysis (prediction)

Knowledge graph

Design time Run time

WHY: Knowledge Graphs are Universal Enablers

Knowledge foundation

If you have the right knowledge, you'll do the right thing

A knowledge base is richer than a "data" base. It contains unique knowledge: data context and purpose. It gives sense to your job.

Relationships are stored

- In a highly siloed world, all you need is relationships
- No need for compute time/resource to query relationships
- Not limited by SQL join queries (logically and technically)

Business purpose-driven design

- Based and linking back to the "why"
- Not driven by technical requirements (RDBMS optimization)
- Owned by business experts: only them know when data makes sense
- Schema-free
    - modelling flexibility
    - encompass fast business changes

One place for shared DIK

- Data and semantic (what it means)
- Shared common good

HOW: DIK Extraction and Integration

Three main sources, not more: brains, tables, documents

Brains

Knowledge elicitation

- Business descriptions from experts
- Problem statements, needs, requirements, blockers…
- Know-how, tacit knowledge
- Business processes
- FAQs
- Business resources identification

Tables

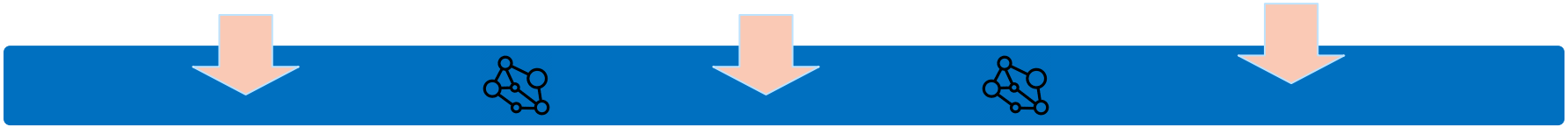
Data discovery, profiling, ETL/ELT

- Data models (theory and reality)
- Architecture handbook
- System diagrams
- Data flow diagrams
- Profiling metadata
- Data ETL/ELT pipelines (for selected data only)

Documents

Text retrieval and mining

- Document types
- Document concepts
  - Structure
  - Text
- Document metadata
  - Documentum system
  - Document format
- Business context metadata



Use Case 1:
Eight-years of Graph-Based Problem Solving

Overview of Use Cases I Addressed of the Last 8 Years

Covering drug research, development and manufacturing domains

1. Data Governance

Standards, data quality, FAIRification, provenance, RACI,...
(See next slides)

2. Modelling and Linguistics: master and reference data management

Concept and vocabularies integration across domains

3. Business architecture

Business and IT ecosystem description and integration

4. Portfolio management

Molecules, studies, projects, program, applications, collaborations...

5. Inventories and cataloging

Any business and TI assets

6. Clinical data mapping and integration

25 years of legacy clinical study variables mapped with 20'000 semantic rules

7. Data productization

Data set, domains, roles, systems, customers...

8. Raw material 360

Material, grade, packaging, supplier, trade name, unit price

9. Batch release digitalization

Raw material, API, bulk, semi-finished, pack., finished products

10. Monitoring business value delivery

Projects, success, KPIs, roles and responsibilities
(See next slides)

Use Case 2:
Data Governance 4.0

Use Case Summary

Bootstrapping the Novartis Data42 program

Mission

First Clinical Development and then Enterprise Data Governance

Observation

We can only govern what we know

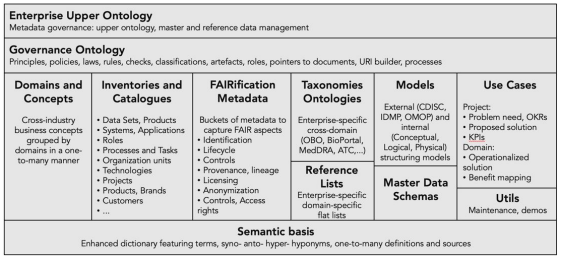
Guiding principle

Know the data and the “why” of governance to prevent governance constraints to prime over its benefits.

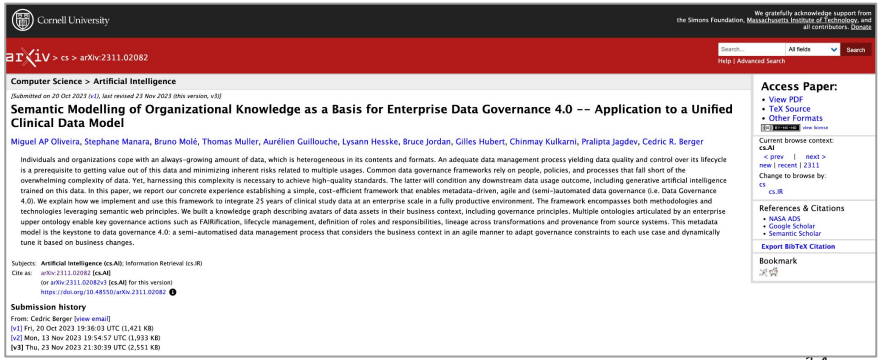
How to achieve that?

Create a knowledge graph of data landscape and governance scope with all relevant information/knowledge.

KG content



Story: https://arxiv.org/abs/2311.02082

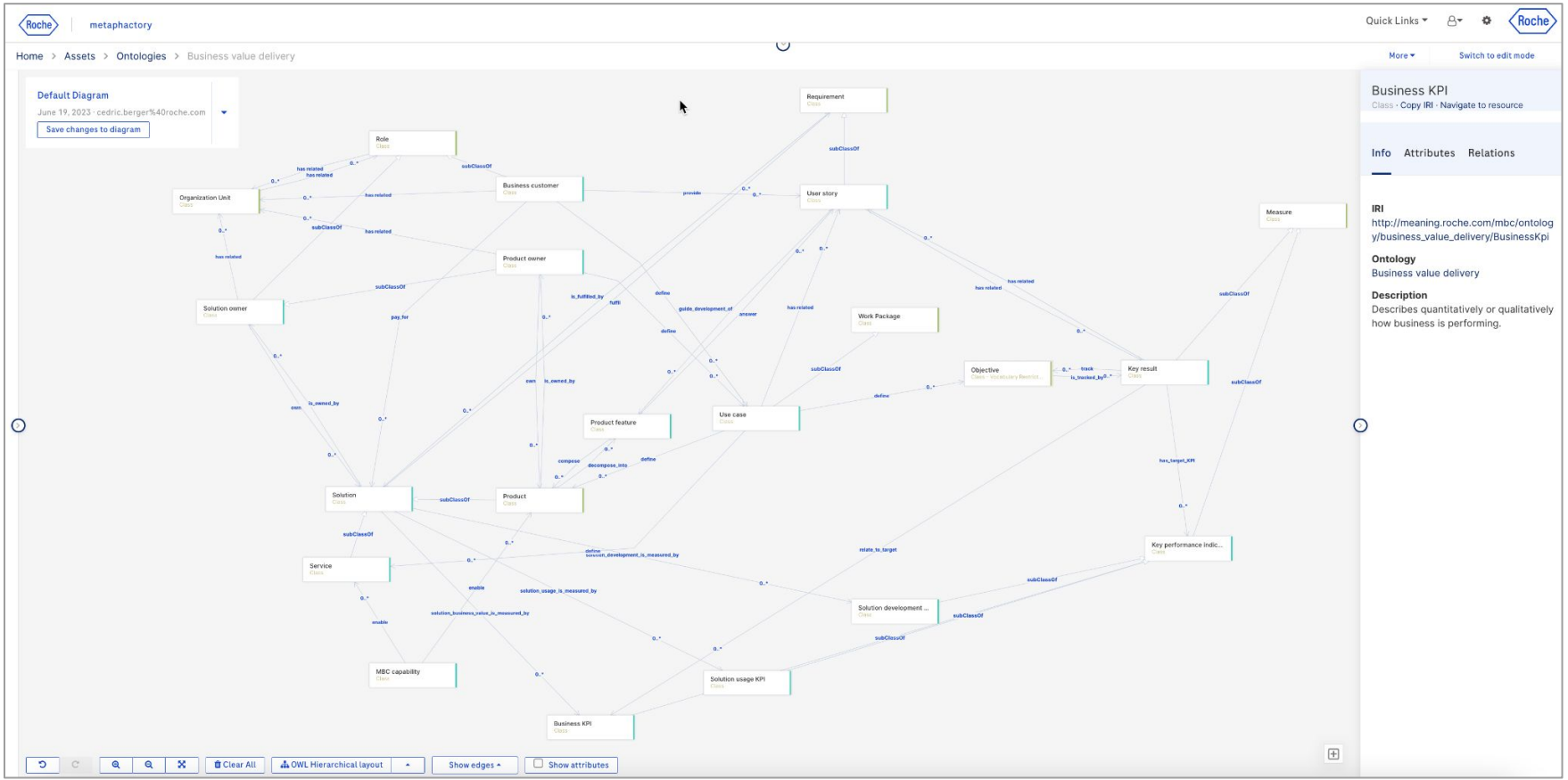


Most recent update

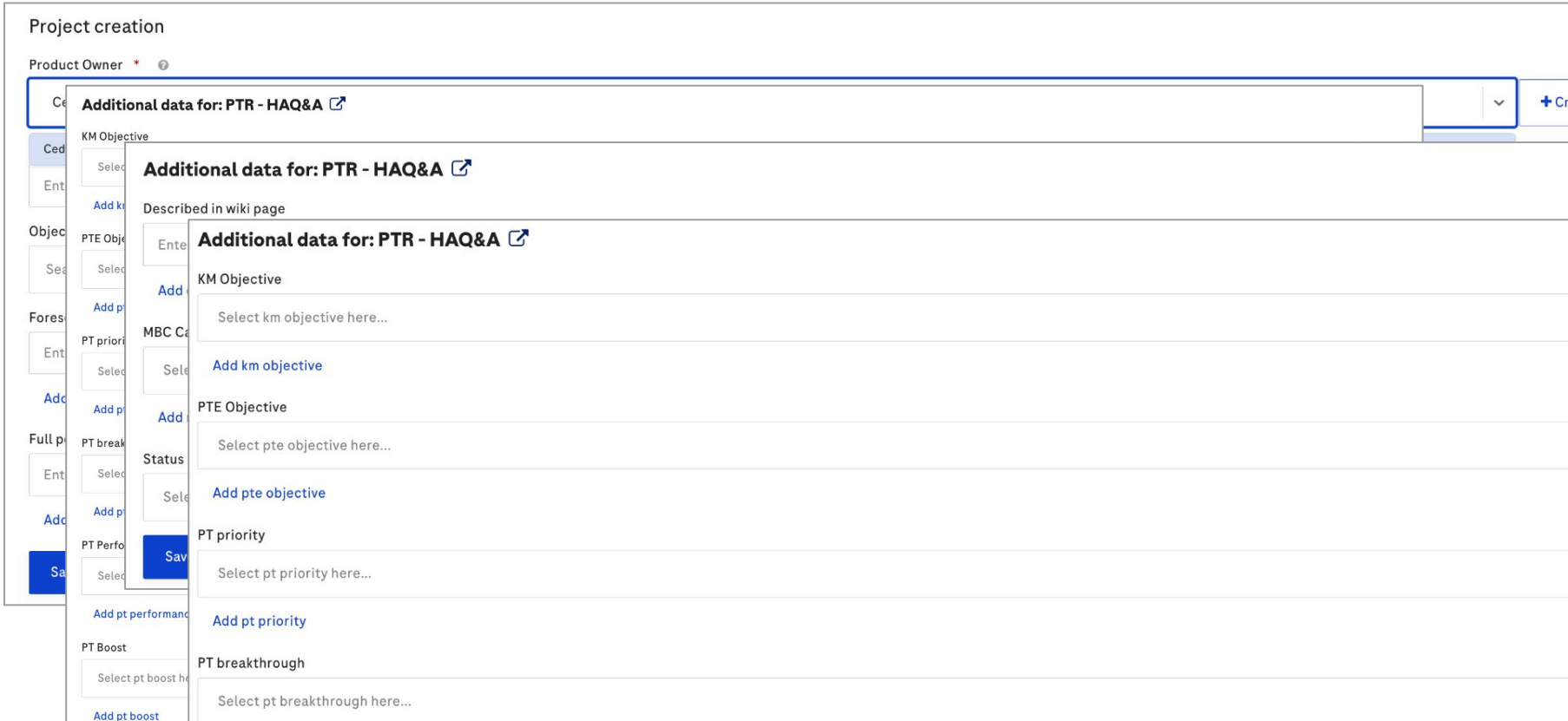
- Not used anymore
- To be decommissioned with the last knowledgeable person going in pension
- management interest in Pistoia ClinOps-O

Use Case 3:
Monitoring Business Value Delivery

Linking all you need to capture business value delivery



Model-driven (use case) data capture forms



Model-driven report (for each use case)

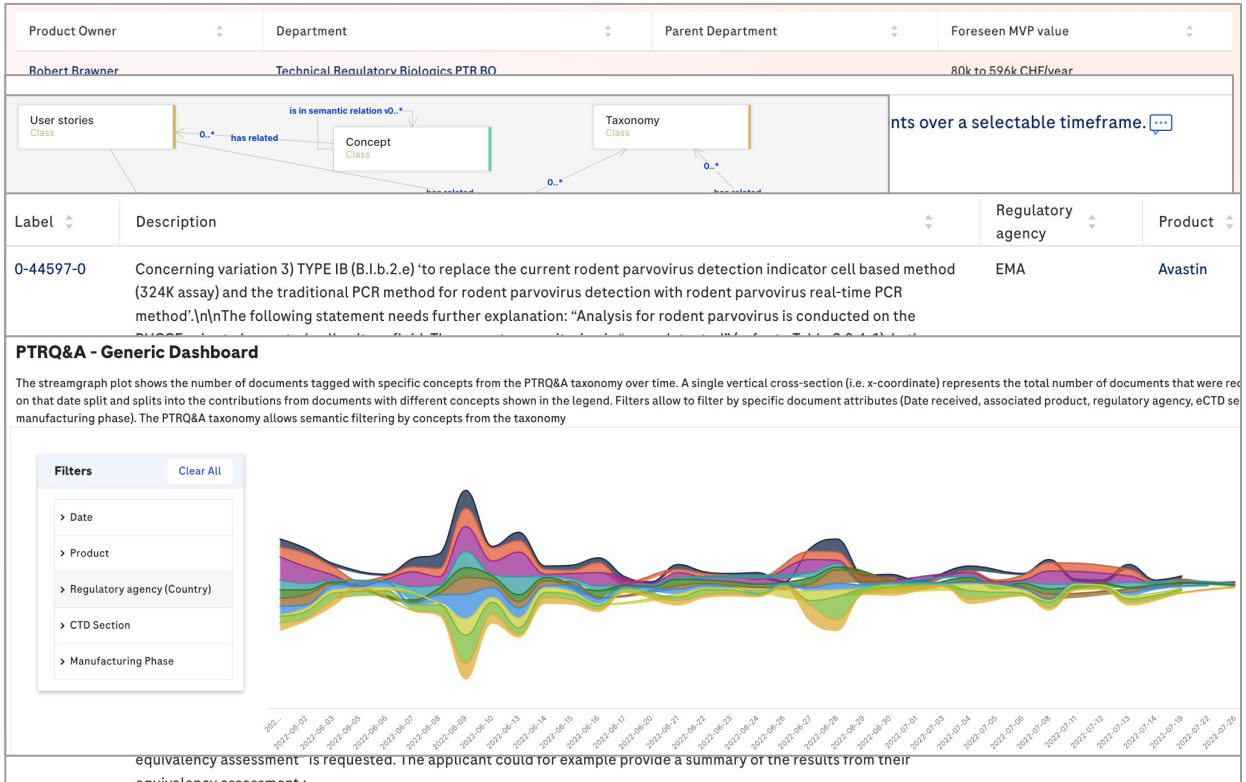
Executive summary

Requirements/user stories

Business concepts

Data sample

Digital solutions



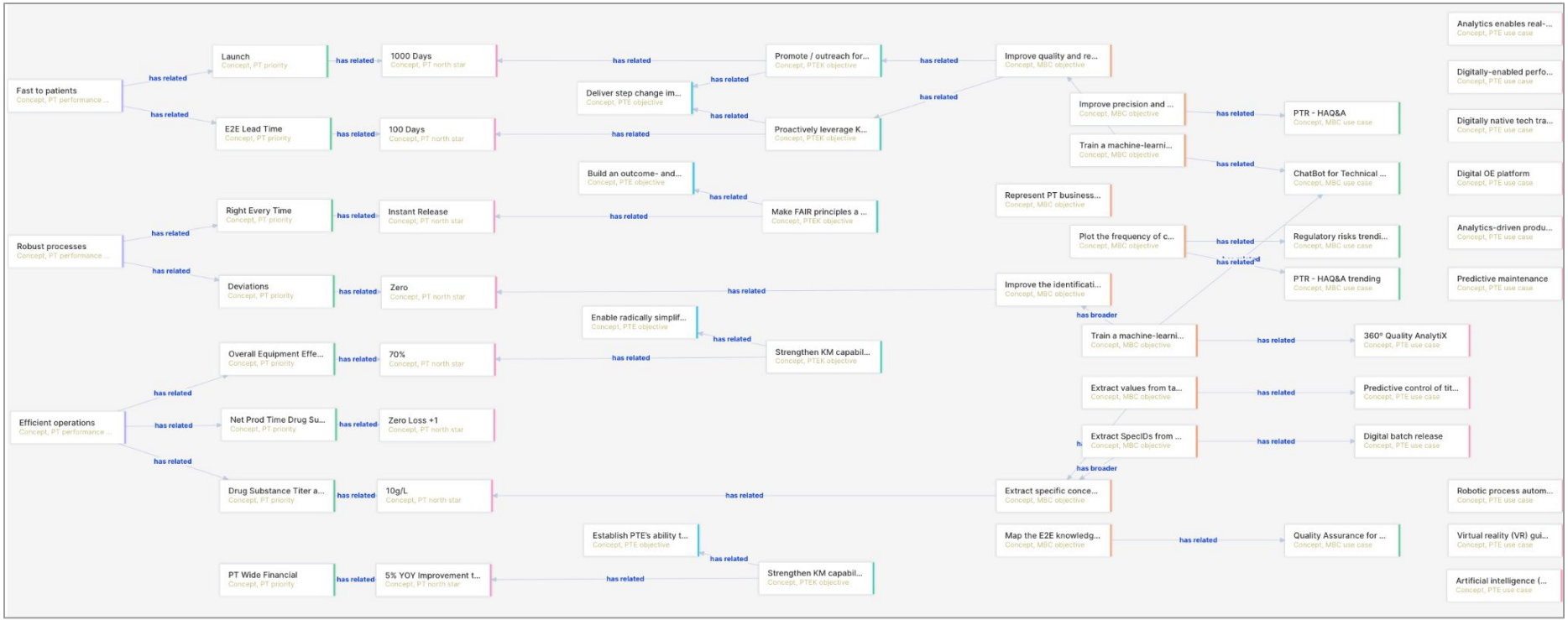
Model-driven integrated report

Quick search

Project	Department	Objectives	Impact	Status
PTR - HAQ&A	Technical Regulatory Biologics PTR BO	Improve quality and response time to HAs questions by improving search and retrieve of legacy Q&A and enabling trending	80k to 596k CHF/year	Under discussion
ChatBot for Technical Transfer	Europe Biologics PTD EU	Have a chat bot that serves as a senior SME explaining and consulting about the drug product manufacturing process across the manufacturing network and product phases	49k - 20 Mio CHF/year	Pilot
Quality Assurance for Quality Control	QA for QC	Reduce time spent finding knowledge along the supply chain process	58K to 189 CHF/year	Under discussion
Predictive Yield/Titer	Digital Strategy & Value Realization	Automatically extract and process knowledge contained in tables in PDFs documents from raw material suppliers		
Digital Batch Release	Tech Reg OBE Business Systems	Automatically extract and process knowledge out of DS and DP specification document (Knowledge extraction from S.4.1 and P.5.1)		
AI scouting and landscape	Pharma Global Technical Operations	Use automatic text-based document processing to extract AI technology information and report it		
Deviation trending	Pharma Global Technical Operations	Create a dashboard to automatically identify, characterise and trend GxP process deviations		
Expert search	Pharma Global Technical Operations	Characterise experts based on artefacts they generate by extracting automatically skills		
Semantic IDMP	Pharma Global Technical Operations	Raise awareness about and increase adoption of IDMP ontology		
Integration of PT systems inventories	Pharma Global Technical Operations	Create a model-based integration pipelines for the different inventories of PT systems		

1 2 » Show 10 rows

Linking deliverables at different levels of the organization

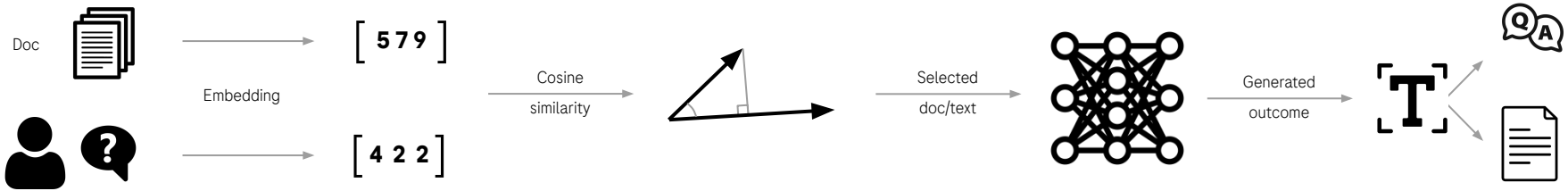


Use Case 4:
Enabling Generative AI

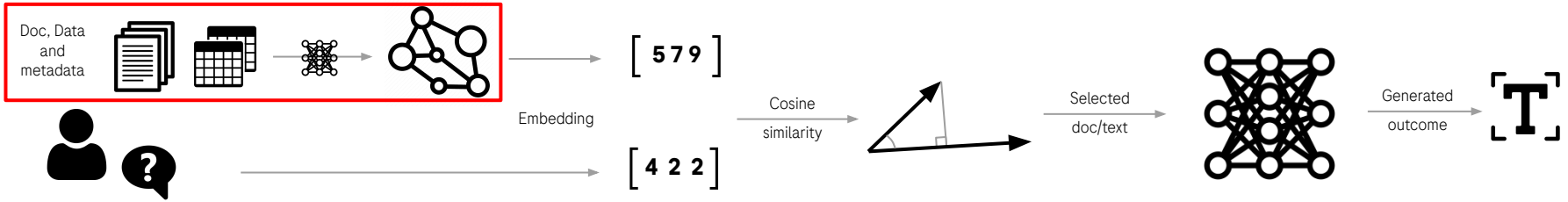
On-demand Fit-for-purpose Graph out of Data

Increasing signal-to-noise ratio inbound LLMs

Current common RAG pipeline



Proposed Graph-RAG pipeline



Why Graph-RAG?

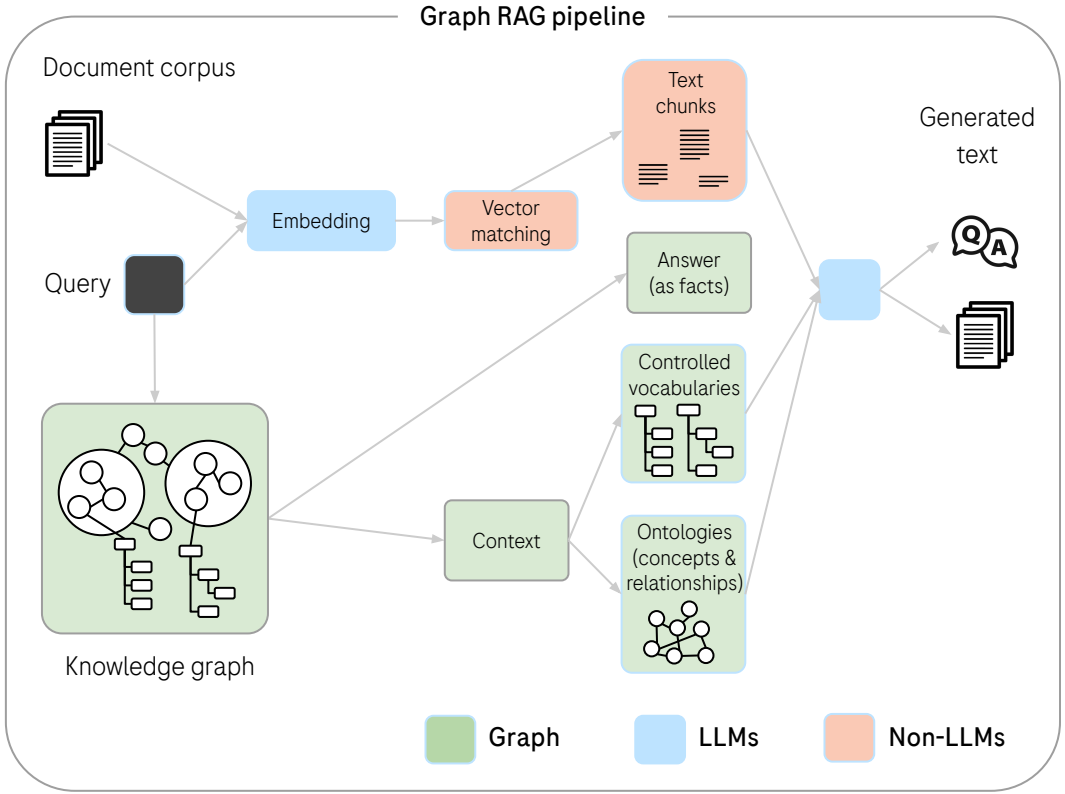
One option among adaptive RAG

What is graph RAG

Using knowledge graph as part of LLM feed

How g-RAG improves LLMs outcome

- **Densified**: only facts (triple); richer outcome
- **Personalized**: your taxonomies, your answer
- **Fit-for-purpose**: your use case, your user stories
- **Reduces hallucinations**: ground LLMs with facts



LLM Selection

What Existing LLM is Best at Creating RDF Triples Out-Of-The-Box?

In scope

Model	Creator	Context	Publish Time
GPT-3.5 Turbo	OpenAI	4k	2023.03
Llama2-7b-chat	Meta	4k	2023.07
Llama2-13b-chat	Meta	4k	2023.07
Mistral-7B-Instruct-v0.1	MistralAI	4k	2023.09
Mistral-7B-Instruct-v0.2	MistralAI	32k	2023.12
mpt-7b-8k-instruct	MosaicML	8k	2023.07
Yi-6B-Chat	01-AI	4k	2023.11
InternLM2-chat-7B	InternLM	200k	2024.01
Qwen1.5-7B-chat	Qwen	32k	2024.02

Error analysis

- Factual error
- Not in ground truth
- Not adhering to output format
- Hallucinations

Results

Mistral-7B-Instruct-v0.2 was selected

Performances

With PharmText2KG-Text2KG

Models	Fact Extraction			OC	Hallucinations				
	Precision	Recall	F1	OC	SH	OH	AH*	RH	Avg H
GPT 3.5	**0.56**	**0.57**	**0.56**	**0.98**	0.01	0.13	0.08	0.01	**0.0575**
Vicuna-13b (Llama)*	0.34	0.27	0.30	0.93	0.12	0.38	-	0.07	-
Llama2-7b-chat	0.20	0.26	0.22	0.80	0.30	0.50	0.16	0.18	0.285
Llama2-13b-chat	0.41	0.43	0.41	0.90	0.07	0.26	0.13	0.09	0.1375
Mistral-7B-Instruct-v0.1	0.32	0.34	0.32	0.81	0.08	0.31	0.18	0.13	0.175
Mistral-7B-Instruct-v0.2	**0.48**	**0.52**	**0.49**	**0.94**	0.04	0.18	0.17	0.05	**0.11**
mpt-7b-8k-instruct	0.16	0.19	0.16	0.50	0.17	0.34	0.24	0.19	0.235
Yi-6B-Chat	0.12	0.23	0.14	0.54	0.27	0.52	0.22	0.40	0.3525
InternLM2-chat-7B	0.16	0.47	0.23	0.62	0.31	0.50	0.17	0.31	0.3225
Qwen1.5-7B-chat	**0.48**	0.49	0.48	0.91	0.08	0.20	0.16	0.08	0.13

With PharmText2KG-BioRED

Models	Fact Extraction			OC	Hallucinations				
	Precision	Recall	F1	OC	SH	OH	AH*	RH	Avg H
GPT 3.5	**0.093**	**0.132**	**0.109**	**0.62**	0.31	0.5	0.17	0.37	0.338
Llama2-7b-chat	0.014	0.017	0.015	0.35	0.51	0.78	0.29	0.47	0.513
Llama2-13b-chat	0.046	0.072	0.056	0.49	0.41	0.69	0.23	0.45	0.445
Mistral-7B-Instruct-v0.1	0.017	0.022	0.019	0.31	0.58	0.74	0.34	0.54	0.550
Mistral-7B-Instruct-v0.2	**0.055**	0.078	**0.065**	0.51	0.38	0.62	0.21	0.42	0.408
Mpt-7b-8k-instruct	0.026	0.037	0.031	0.28	0.47	0.72	0.51	0.58	0.570
Yi-6B-Chat	0.017	0.042	0.024	0.32	0.57	0.81	0.49	0.51	0.595
InternLM2-chat-7B	0.016	0.031	0.021	0.26	0.49	0.78	0.43	0.56	0.565
Qwen1.5-7B-chat	0.043	**0.097**	0.060	**0.55**	0.29	0.53	0.19	0.39	**0.35**

Deliverable 1: A New Training Data Set

Dedicated to Transform Pharma Documents into Graphs

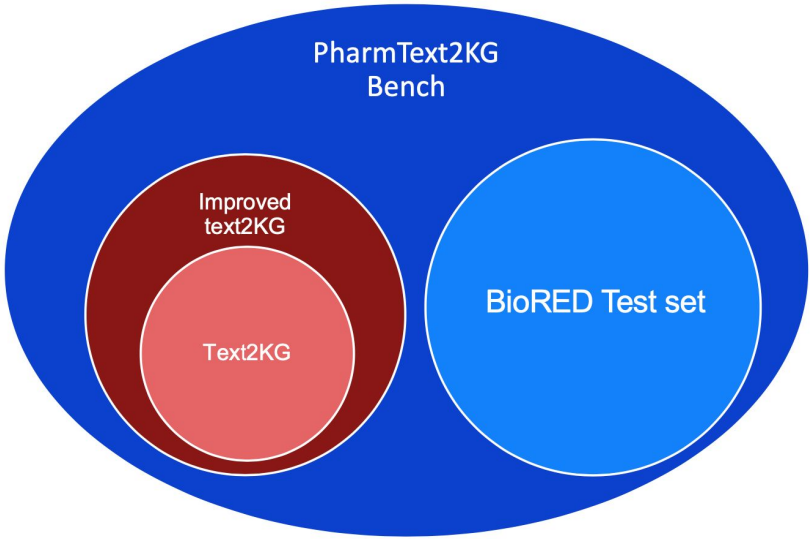
PharmText2KG

Composed of

- **Text2KG**
  - Wikidata-TekGen with 10 ontology and 13,474 sentences
  - DBpedia-WebNLG with 19 ontologies and 4,86 sentences

- **Improved Text2KG:** subset merging, error fixing, paragram formation (30%)

- **BioRED:** Biomedical Relation Extraction Dataset (BioRED) contains labels with multiple entity types (e.g. gene/protein, disease, chemical) and relation pairs (e.g. gene- disease; chemical–chemical) at the document level, on a set of 600 PubMed titles and abstracts (400 train, 100 validation, 100 test).

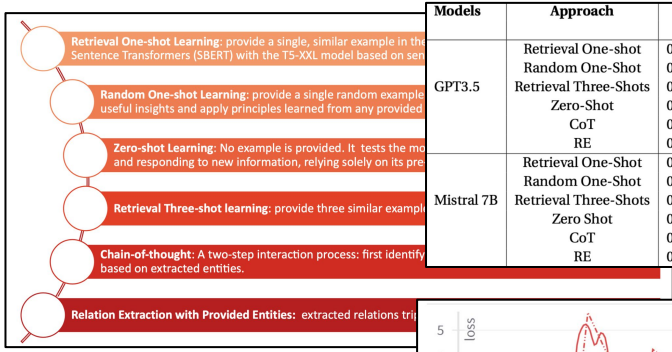


Deliverable 2: A Fine Tuned LLM

Dedicated for Pharma

Why fine tuning?

- We tested In-Context Learning methods
- Results were disappointing



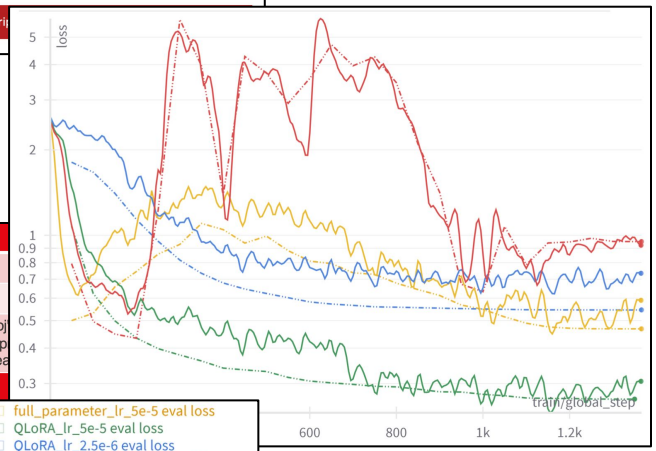
Models	Approach	Fact Extraction			OC	Hallucinations				
		P	R	F1	OC	SH	OH	AH	RH	Avg H
GPT3.5	Retrieval One-shot	0.56	0.57	0.56	0.98	0.01	0.13	0.08	0.01	0.05
	Random One-Shot	0.53	0.55	0.53	0.94	0.01	0.03	0.07	0.05	0.04
	Retrieval Three-Shots	0.58	0.56	0.57	0.99	0.01	0.04	0.04	0.06	0.04
	Zero-Shot	0.29	0.3	0.29	0.97	0.01	0.36	0.34	0.02	0.18
	CoT	0.37	0.4	0.38	0.96	0.01	0.05	0.02	0.03	0.03
	RE	0.59	0.62	0.6	0.98	0.02	0.01	0.01	0.04	0.02
Mistral 7B	Retrieval One-Shot	0.48	0.52	0.49	0.94	0.04	0.18	0.17	0.05	0.11
	Random One-Shot	0.41	0.43	0.41	0.91	0.01	0.08	0.39	0.08	0.14
	Retrieval Three-Shots	0.49	0.55	0.52	0.96	0.02	0.08	0.09	0.01	0.05
	Zero Shot	0.19	0.19	0.18	0.91	0.02	0.55	0.51	0.08	0.29
	CoT	0.26	0.2	0.12	0.65	0.09	0.38	0.04	0.34	0.22
	RE	0.46	0.49	0.47	0.9	0.02	0.14	0.6	0.09	0.21

How did we fine tune the model?

Quantized Low-Rank Adaptation (QLoRA)

- **Why LoRA:** the number of parameters required to effectively perform a given downstream task is much smaller than the total number of parameters in the model
- **Why quantized:** require less computational resources

QLoRA Hyper-parameter	
Rank	16
lora_alpha	32
target_modules	["q_proj "gate_p "lm_hea
General Hyper-parameter	
epochs	2
Batch size	32 (1
Learning rate	5e-5
Lr scheduler	cosi
Warmup ratio	0.2



Training data

8199 sentence-level records from the Text2KG training set, in addition to 500 paragraph-level records from the BioRED training set.

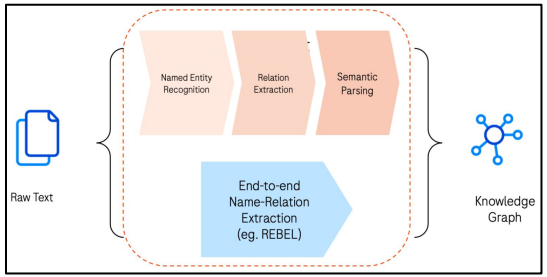
Deliverable 3: An Innovative RAG pipeline

From Noisy General Text to Concentrated Fit-for-Purpose Knowledge Graph

Legacy

Concepts and relationship extraction but

- domain complexity
- data dependence
- limited adaptability
- multilingual and customization



LLM-based approach*

Input:

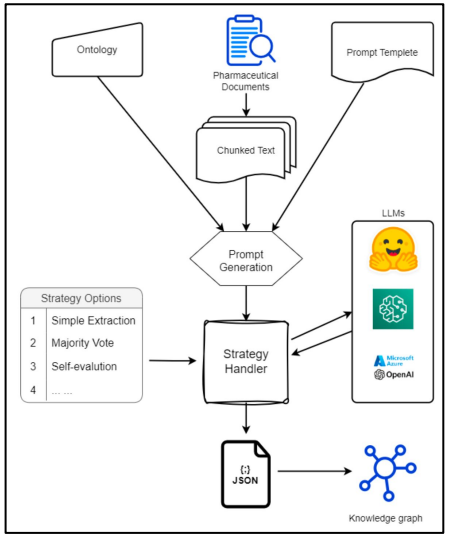
- Raw text/doc corpus ← Generic
- Ontology
- Query/instructions
- Example
- Target data

Prompt

Use case specific

Output:

- Knowledge graph triples in specific format (e.g. TTL)



* Trajanoska et al. investigated the use of ChatGPT for KG construction of unstructured text in the context of sustainability topics; Zhu et al. assessed GPT-4's performance and created a "Virtual Knowledge Extraction" task and proved that the information extraction ability of LLMs comes from their potent generalization capabilities facilitated by instruction tuning, instead of their intrinsic knowledge; Caufield et al. introduce SPIRES, a novel approach for leveraging LLMs to recursive extraction information and populate KGs without the need for extensive training data; Text2KGBench: a benchmark designed to evaluate the efficacy of LLMs in KG generation from textual inputs.

Lessons learned

Are KGs really Universal Enabler?

Still limited understanding and adoption

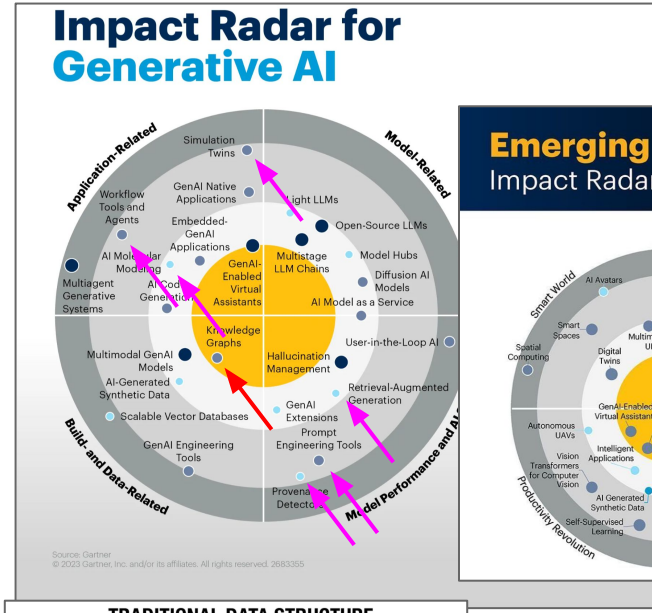
Most advanced GenAI enabler →

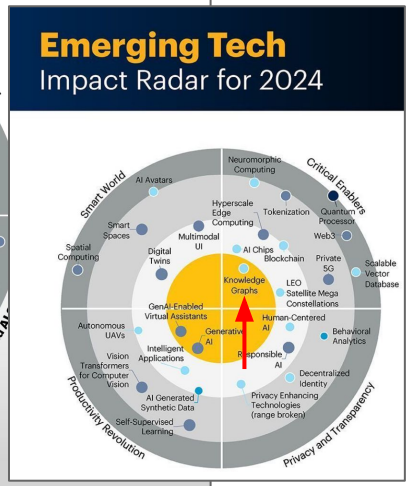
Also contributes to →

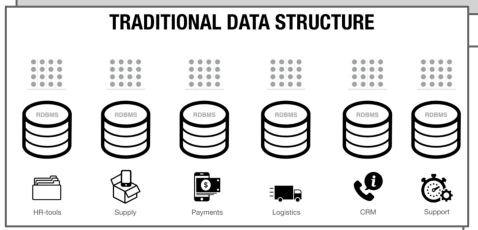
- Hallucination management, RAG, Prompt engineering, Provenance detector in sector "**Model performance and AI safety**" and Simulation twins, Workflow tools and agents, AI molecular modeling, in sector "**AI-enabled applications**"
- DIK management applications as a descriptive layer

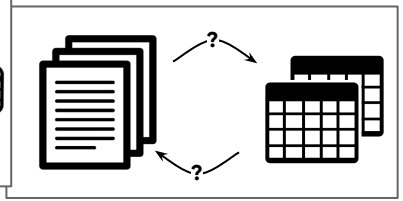
But still not widely adopted

- Most of enterprise systems are based on relational/SQL databases
- There is little connections between structured and unstructured data









SQL join across 1000 of tables

NLP matches across 10^6 of doc

Excel mapping between SQL join and NLP samples

Q: *“Why is adoption so low?”*

A: *Technology has some limitations (like any) but it excels with unique capabilities. Technical limitations do not justify low adoption.*

Q: *“What then?”*

A: *It is just about people, as usual.*

First, people (decision-makers) have to recognizing there is a problem with DIK* management

The first step to the cure is to recognize being sick

Personal go/no-go

Do we have a problem with DIK* management in our organization?

1. **I don't care/ I don't know:** digital is not our core business
2. **No we don't**: we single-shot fix issues by throwing money, technology, people at it
3. **Yes we do…**
   a. **… but we won't do anything about it**: business is still profitable, why changing?
   b. **… we can try to do something but** changing current ways of working have poor chances to succeed because it is resource demanding for individuals and the organization (self-preservation, complacency, politics, lack of understanding or skills)
   c. **… yes I feel the pain and I will do everything we can to improve it**

Single thread ownership

* https://en.wikipedia.org/wiki/DIKW_pyramid

Do we have a problem with DIK* management in our organization?

Why is the answer often “no”?

People answering don’t have the skills, knowledge or experience

- “Let me organize a meetings, presentations, workshops to explain”?
  - I/we Don’t have time
  - Learning curve is steep, it is too much of an effort
  - I have other more important priorities
- Then you have to trust me

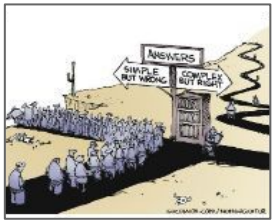
This is where it stops because:

- it questions the expertise of the decision maker/manager
- it is a risk and a risk-averse business/companies only selects risk-averse managers

People answering understand your proposal and see it as…

A unrealistic person, proposing difficult and long-lasting deliverables

- Not aligned with personal success
- Not aligned with short-term business performances



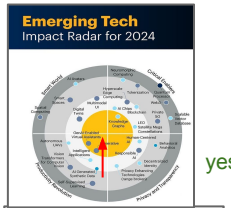
A threat to established order/system

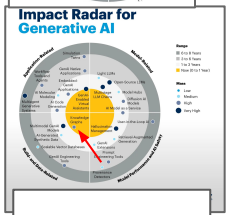
- That they understand
- That ensure their status of expert
- That they exploit to keep they privileges

* https://en.wikipedia.org/wiki/DIKW_pyramid

Accepting Solutions One Doesn't Understand

We don't know what we don't know





Do you trust Gartner's expertise?

yes

Do you understand how KGs can add value to our business?

yes

Will you implement yourself a framework to leverage KGs

yes

Your organization shall benefit from KGs

no

Would you dedicate time and effort to understand?

yes

Would you trust a KG expert to do the job?

yes

Do you understand that implementing technology is not enough

yes

Would you commit to break-free from corporate IT?

yes

Would you commit to be the single thread owner for this initiative?

yes

Would you commit to long-term support this initiative?

yes

no

Your organization shall not benefit from KGs

no

no

no

no

no

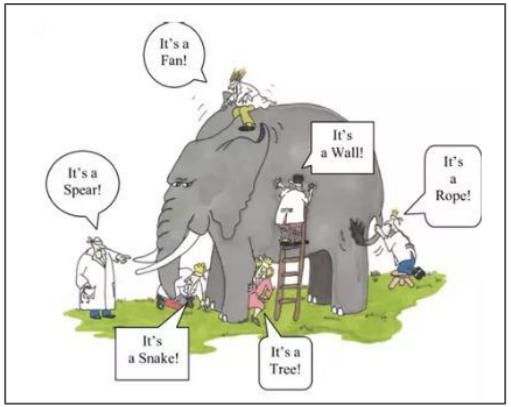
no

System-thinking

Rare cognitive abilities

Most people don't think in system

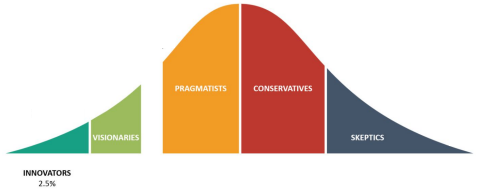
- There is a long chain of factors and logic between raw data and P&L financial statement
- This chain is siloed and managed by silo experts: there might be siloed solutions but no one has the full picture/understanding



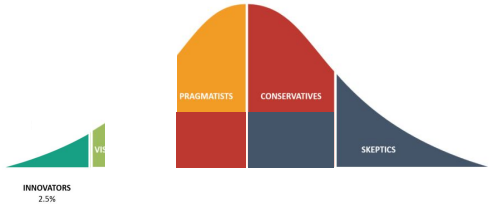
Most people don't want to improve the system

Innovation adoption curve: while the proportion of innovators does not change, the rest varies based on intersubjective stress and trust

Normal times



Stressful times



Second, people (change-drivers) have to recognizing that explaining is not enough

Many smokers continue to smoke despite understanding it might kill them

Hard skills are not enough

- Explaining technical casualties is not enough to convince people to change

Additional soft skills needed

- Cognitive sciences, biases, rejection, denial
- Psychology, sociology
- Rhetoric, story telling
- Negotiation, diplomatic techniques
- Strategy, influencing tactics
- Game theory
- Marketing of ideas: how to make new ideas stick
- Change management

Changing paradigm (unlearning) is a tremendous effort for human brains

For change drivers

1. Not only presenting, explaining the problem and the solutions
2. But also providing a framework to trigger and roll out changes that includes
   a. Reward loops
   b. Notion of common good sharing for cont. and collective improvement

For change effectors

- need to feel pain of current paradigm
- need to feel pain relief (or reward) due to change

How to force decision-makers to take risks?

Increase pressure

Bottom-up pressure/influence from employees

- Week: most probably just innovators (2.5%); story of my life…

Top-down pressure/influence from management/shareholders

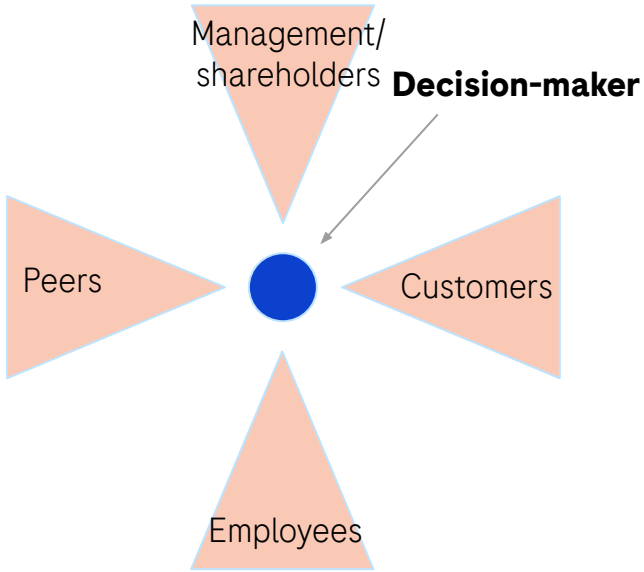
- Extremely rare: main interest/knowledge is/about short-term money-making and self-preservation

Orthogonal pressure/influence from peers

- Punctual/bucket examples
- Cross-pharma initiatives (IMI, Pistoia, Transcelerate…)
- FOMO

Orthogonal pressure/influence from customers

- Rare and indirect: need for digital-aware customers finding the link between digital and pharma business



Conclusion

Less Tech more Influencing Skills

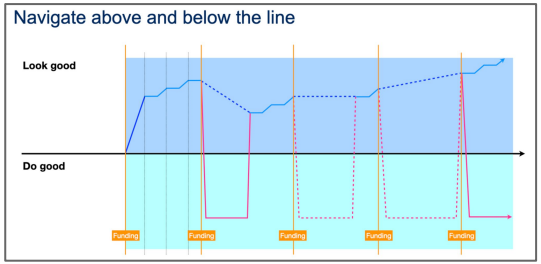
Stop showing graphs

Ideal prerequisites to engage in a graph-based DIK management program

- Decision-makers feel the pain of as-is kludgy DIK assets and management
- Commitment to improve as proven by actions (not just speeches)
- Single-threaded ownership
- Clear success definition and measurement framework based on business requirements

What a good start looks like

- Either get IT support or split up with IT
- Identify small scale, low-hanging fruits, select 3 of them, focus on look-good, show results within 3 months, capitalize on results to to-good



Do to inspire and create pockets of change beyond the aficionados

- Present, influence different people/groups than usual preach to the choir

Doing now what patients need next