A data-ready approach to knowledge management

Ben Gardner

Data Standards, Interoperability and Governance, Data Office, Data Science & Artificial Intelligence, R&D, AstraZeneca, Cambridge, United Kingdom

10th October 2024

Overview

- A bit of context
- IA Before AI
- Controlled Vocabularies
- Knowledge Graphs
- Bringing it back to AI

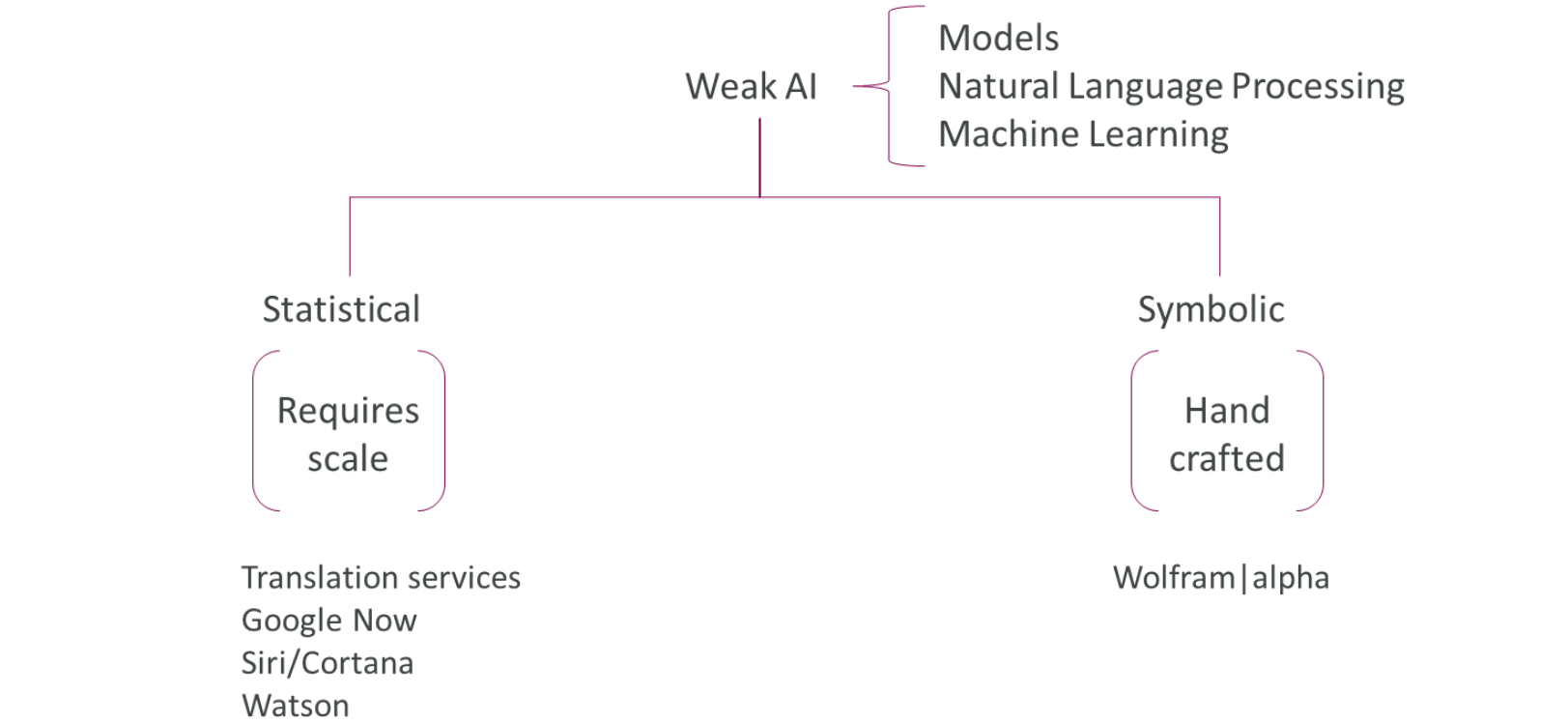
A bit of context

What are we talking about and why do we have problems?

Image credit https://alienencyclopedia.fandom.com/wiki/Deep_Thought?file=Deep-Thought-11krlic.png

What is AI?

"Weak AI - non-sentient computer intelligence or AI that focuses on one narrow task" - Wikipedia



Replicating what individual human brains do

Vs

Replicating the knowledge of our civilization

Stephen Wolfram

Large Language Models are statistical AI

The added value of LLMs

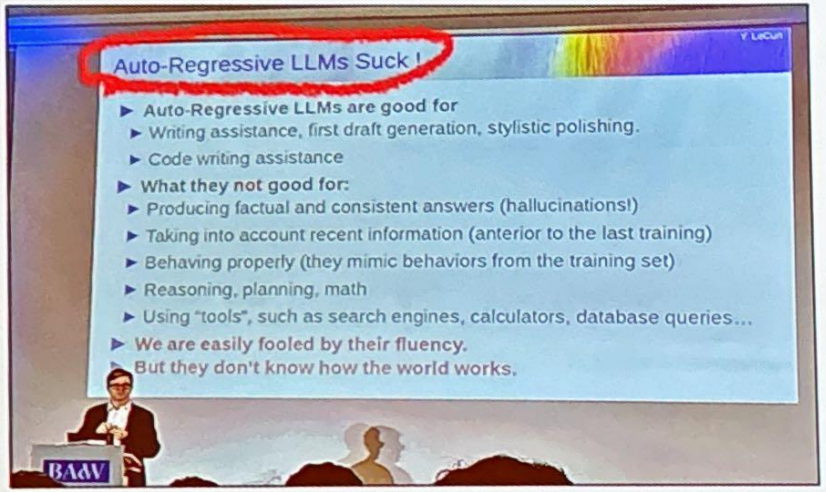
The Economy of Promises

Roche

LLMs limitations and risks

Is your business/activities are not bound to quality, veracity, validation, regulation, (customer) trust?

- *"No"* then no problem, go ahead, use LLMs at will
- *"Yes"*: wait a second, people are *"fooled by their fluency but LLMs don't understand how the world works"* [1]. What they outputs are probabilities of word/pixel association based on training data



How to improve and check/trust LLMs' output?

Improve models

- **Pre-train you own model:** very expensive and computing-intensive; only the big digital can do it
- **Fine-tuning:** need for high-quality annotated data

Improve process around models

- **In-context learning**: aka prompt engineering including Retrieval Augmented Generation (RAG)

Topics of this presentation

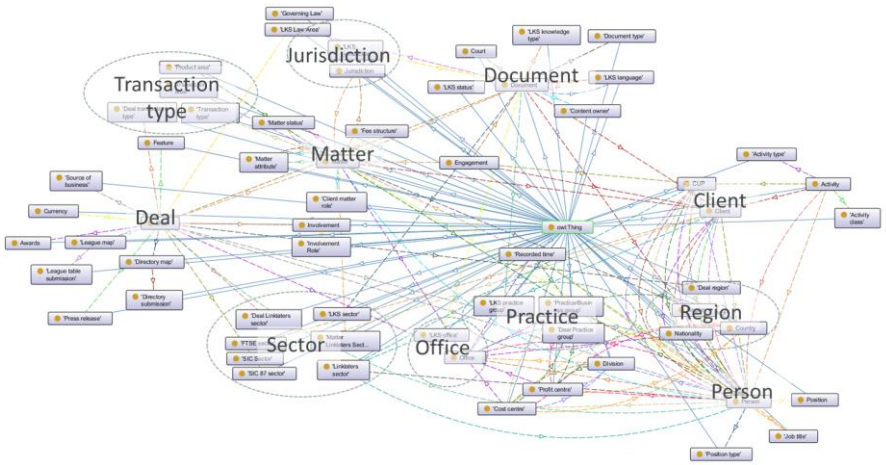
SEMANTICS @ROCHE

1) https://time.com/collection/time100-ai/6309052/yann-lecun/

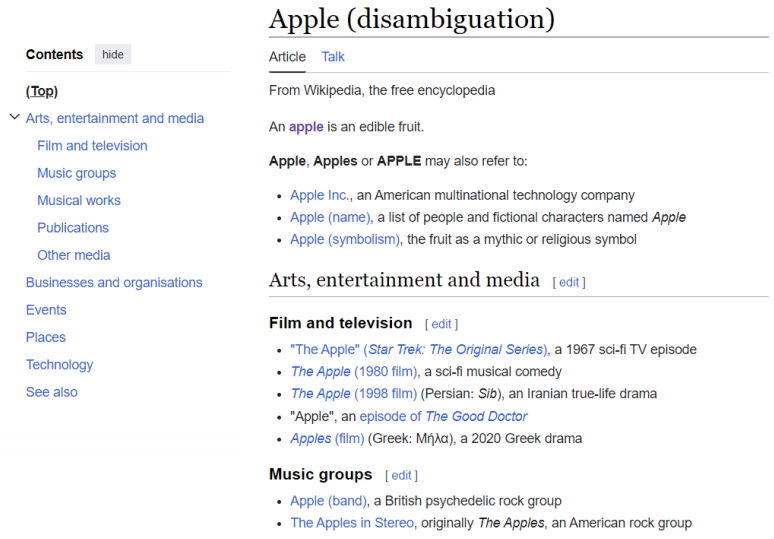
Credit to Dr Cedric Berger, Roche

Ambiguity is the enemy

Fragmentation



Disambiguation



Contents hide

(Top)

Arts, entertainment and media

Film and television

Music groups

Musical works

Publications

Other media

Businesses and organisations

Events

Places

Technology

See also

Apple (disambiguation)

Article Talk

From Wikipedia, the free encyclopedia

An **apple** is an edible fruit.

Apple, **Apples** or **APPLE** may also refer to:

- Apple Inc., an American multinational technology company
- Apple (name), a list of people and fictional characters named *Apple*
- Apple (symbolism), the fruit as a mythic or religious symbol

Arts, entertainment and media [edit]

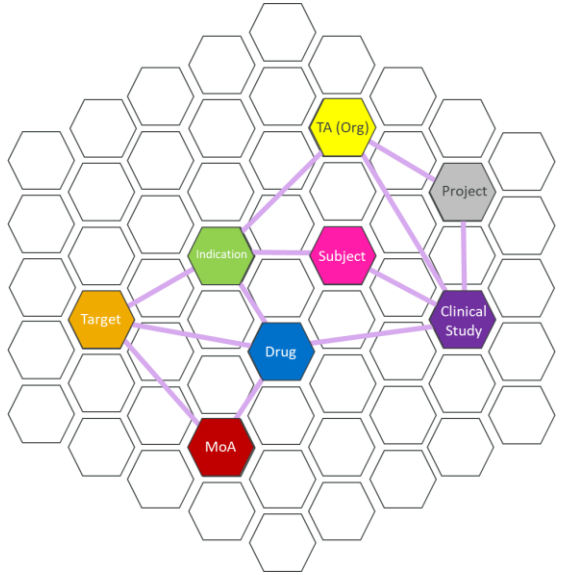
Film and television [edit]

- "The Apple" (*Star Trek: The Original Series*), a 1967 sci-fi TV episode
- *The Apple* (1980 film), a sci-fi musical comedy
- *The Apple* (1998 film) (Persian: *Sib*), an Iranian true-life drama
- "Apple", an episode of *The Good Doctor*
- *Apples* (film) (Greek: Μήλα), a 2020 Greek drama

Music groups [edit]

- Apple (band), a British psychedelic rock group
- The Apples in Stereo, originally *The Apples*, an American rock group

Context



Dirty data

Knowledge Graph

IA Before AI

Information Architecture
before Artificial intelligence

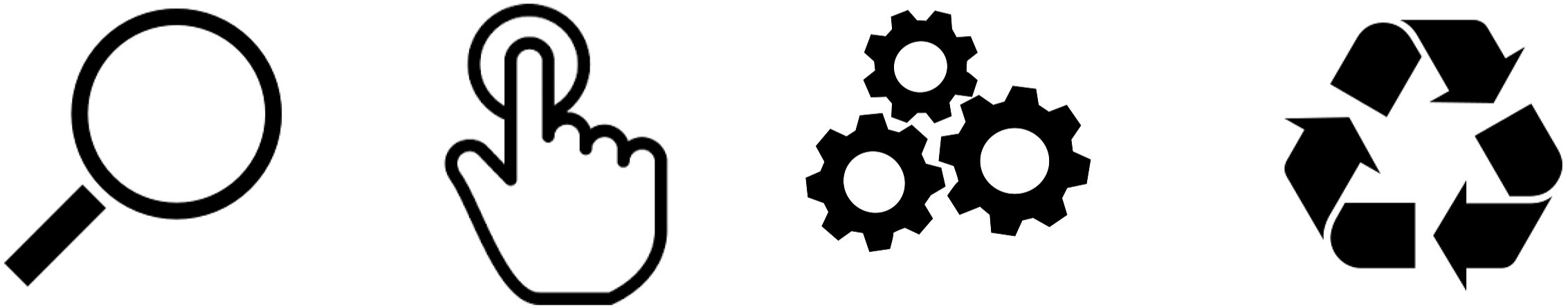
What is Data-Centricity?

Data-Centricity puts **data at the centre of the enterprise**.

Applications are optional visitors to the data. (Data-centric manifesto)

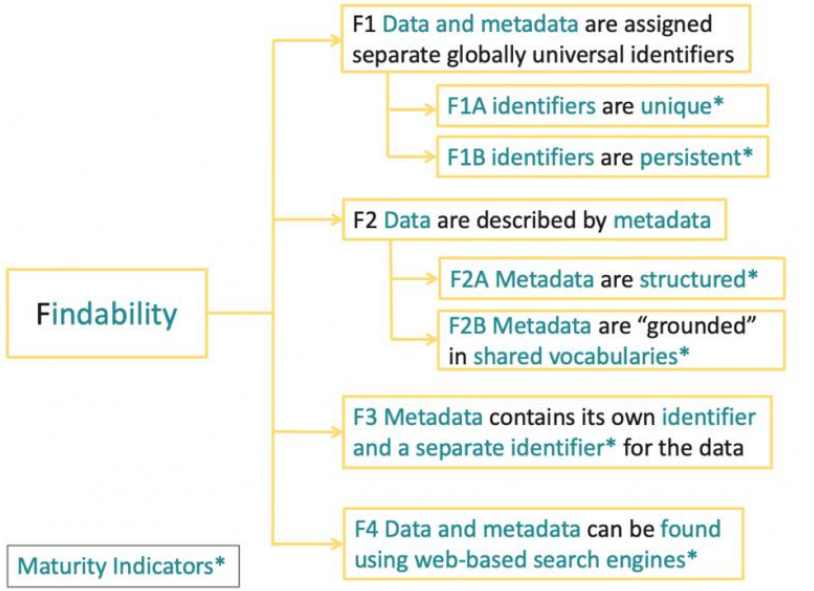
Data-centricity involves structuring our **data around the science** that we do **rather than the systems** that we use. It promotes data reusability over system-centric design.

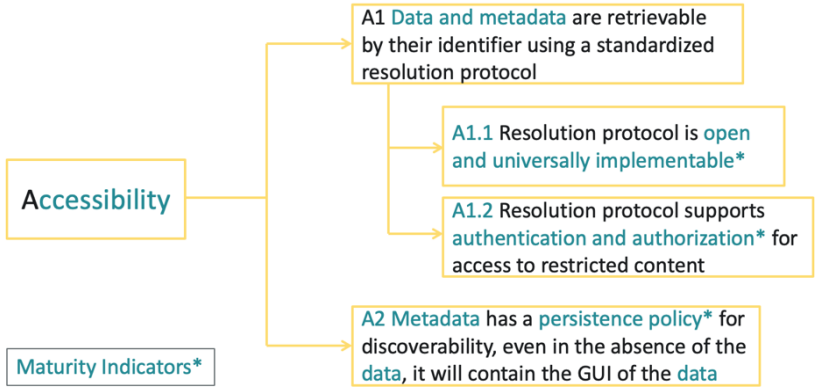
Findable Accessible Interoperable Reusable

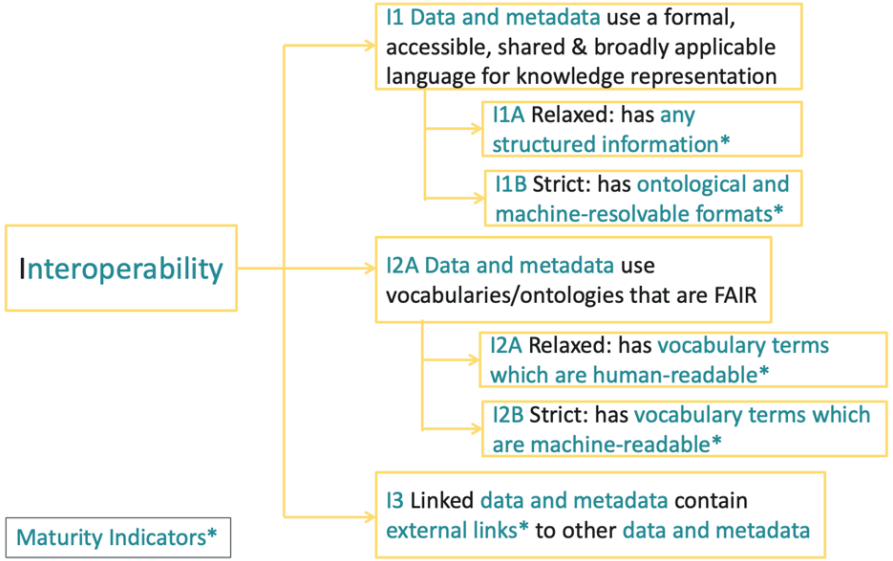


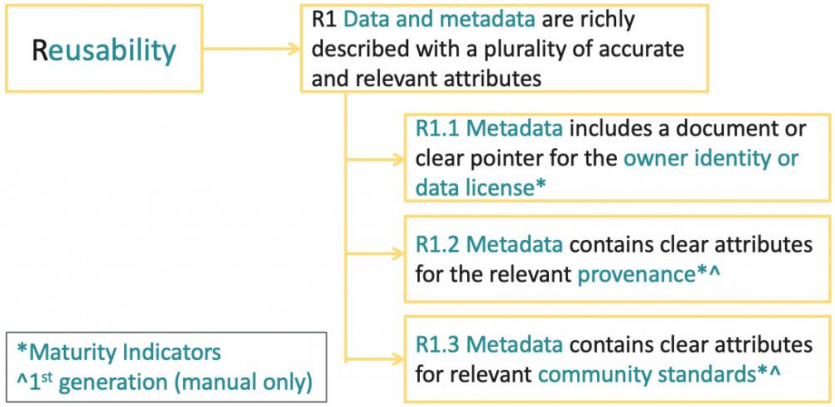
FAIR more than an acronym

Ten principles and 13 sub-principles









see https://fairtoolkit.pistoiaalliance.org/methods/

FAIR data journey as an enabler

From system centric to data centric

Think about your shopping experience….

Junk Shop

Car Boot Sale

Moroccan Street Market

Local Market

Hyper Market

Amazon Prime

How FAIR is your data?

Findable
- My colleagues can discover my dataset/s

Accessible
- My colleagues are able to access data they need

Interoperable
- My data can by easily combined with other data

Reusable
- My data can be used for research or by other processes.

FAIR Maturity Matrix - https://pistoiaalliance.atlassian.net/wiki/spaces/PUB/pages/3308126211/FAIR+Maturity+Matrix

Controlled Vocabularies

Fixing the problem of dirty data

Invest in Controlled Vocabularies

Standardising terms and creating Unique Reference Identifiers (URIs)

Preferred Labels

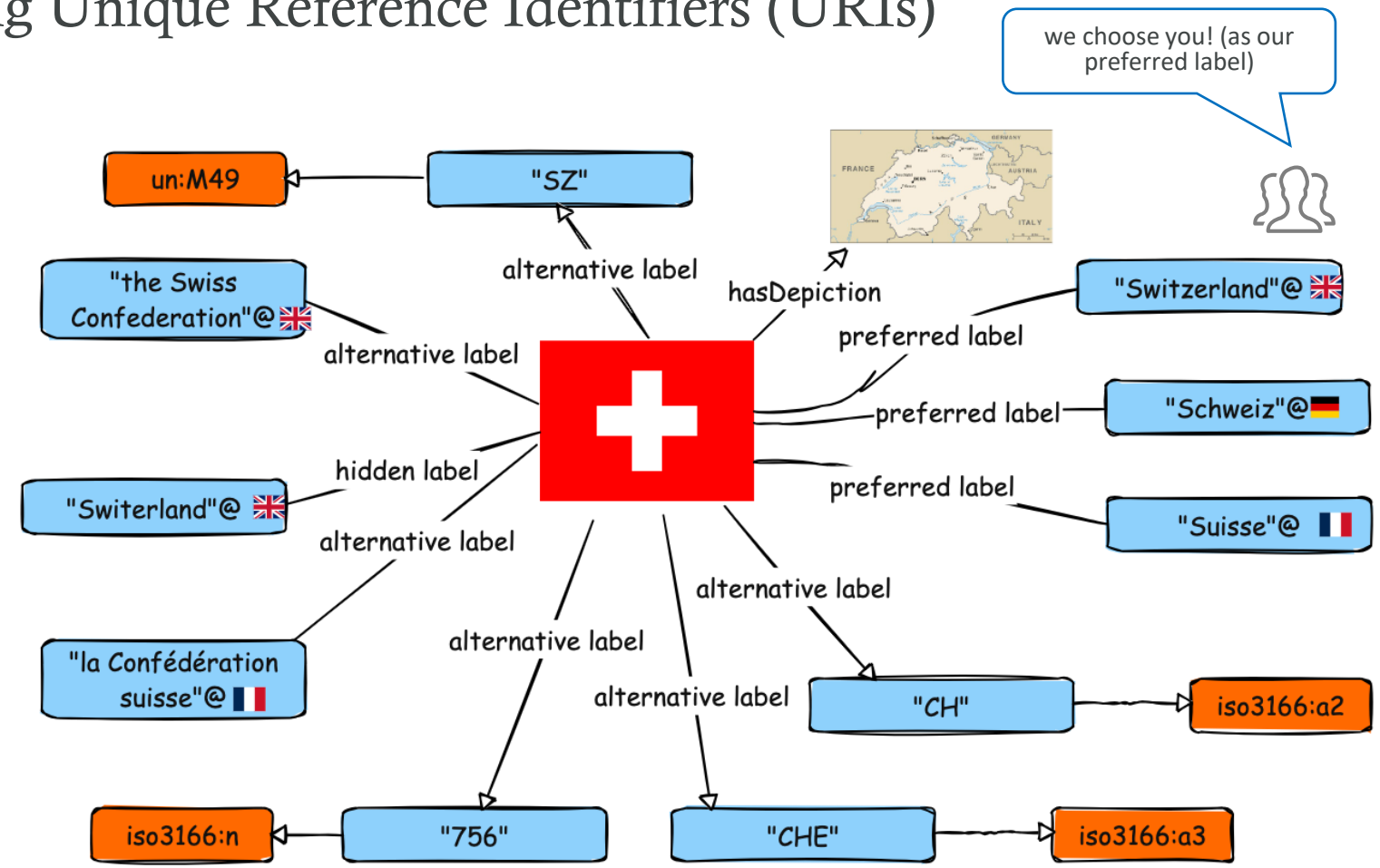
The preferred label for AstraZeneca, that resonates with the business vernacular

Alternative Labels

Well defined non-case variant, alternative labels that are used for this concept. – some may call these “synonyms”

Hidden Labels

Common mis-representations (spelling mistakes, etc) of the concept that exist and we don’t want used by humans. Often used to support NLP and AI activity



A pool of lexical labels exist for each concept. They are common use OR attributed to systems and vocabularies.
AZ curators decide which one will be preferred (for AZ) and whether other labels will be alternative or hidden.
Each label should be further characterized by a signifier.

Credit to Nathalie Conte, Derek Scuffell & Jon Ison – OWG Summer project team

Applying Controlled Vocabularies

Removing ambiguity from structured data

Dirty data

Study	Indication	Drug
D1234C00001	Non small cell lung cancer	Tagrisso
ADORA	NSCLC	Osimertinib
CP11278-CMA33G	Diabetes type 2	Forxiga

- Inconsistent identifiers & terms
- Column values can be concatenated
- etc

Data Transform

Term

prefLabel & URI

Interoperable data

Study_ID	Study_ID__URI	Indication	Indication__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857	Forxiga	https://pid.astrazeneca.com/Product/853584

prefLabel URI

- Shared Controlled Vocabularies
  - Enrich with preferred Label and URI

Controlled Vocabularies
(Master & Reference Data)

Interoperable data benefit all

Inclusion of URIs simplifies data integration irrespective of target data model

Interoperable data sets

Study_ID	Study_ID__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Forxiga	https://pid.astrazeneca.com/Product/853584

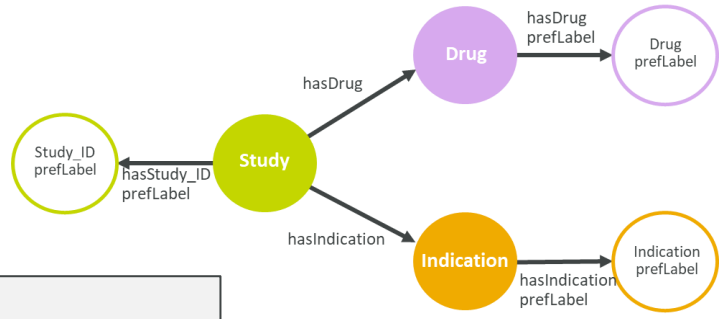
+

Study_ID	Study_ID__URI	Indication	Indication__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857

Relational

Study_ID	Study_ID__URI	Indication	Indication__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857	Forxiga	https://pid.astrazeneca.com/Product/853584

Graph



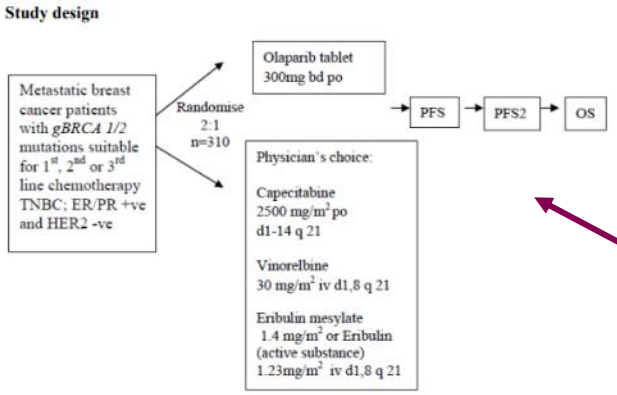
Value

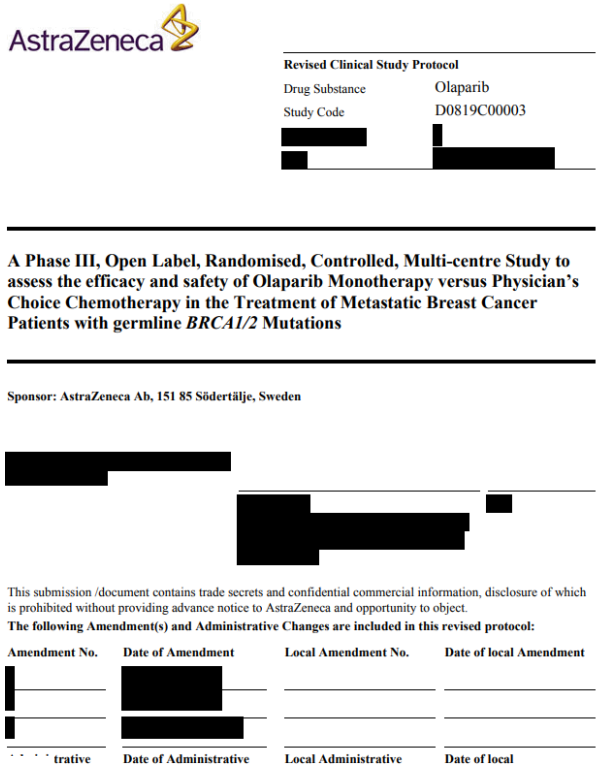
- 80% of a data scientists time is spent wrangling data
- 60% of IT costs are spend on data integration issues

Applying Controlled Vocabularies

Removing ambiguity from unstructured content

Images and tables




AstraZeneca

Revised Clinical Study Protocol

Drug Substance Olaparib

Study Code D0819C00003

[redacted]

A Phase III, Open Label, Randomised, Controlled, Multi-centre Study to assess the efficacy and safety of Olaparib Monotherapy versus Physician's Choice Chemotherapy in the Treatment of Metastatic Breast Cancer Patients with germline *BRCA1/2* Mutations

Sponsor: AstraZeneca Ab, 151 85 Södertälje, Sweden

[redacted]

This submission /document contains trade secrets and confidential commercial information, disclosure of which is prohibited without providing advance notice to AstraZeneca and opportunity to object.

The following Amendment(s) and Administrative Changes are included in this revised protocol:

Amendment No.	Date of Amendment	Local Amendment No.	Date of local Amendment
[redacted]			

trative | Date of Administrative | Local Administrative | Date of local

Sections

Primary Objective

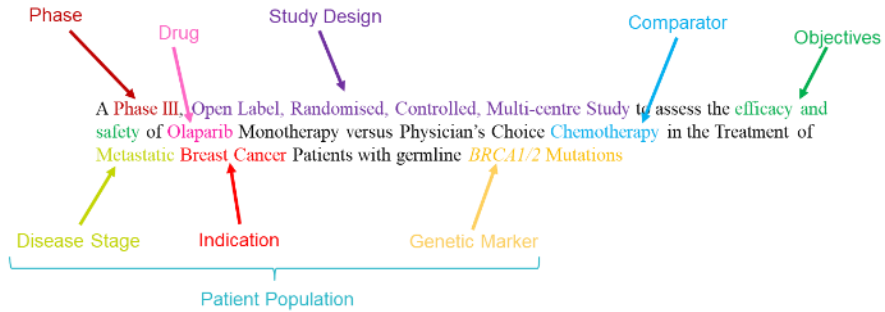
To determine the efficacy of single agent olaparib vs physician's choice chemotherapy (capecitabine, vinorelbine or eribulin) by progression-free survival (PFS) using blinded independent central review (BICR) data assessed by Response Evaluation Criteria in Solid Tumours (RECIST 1.1).

- **Genetic selection**: Documented germline mutation in *BRCA*1 or *BRCA*2 that is predicted to be deleterious or suspected deleterious (known or predicted to be detrimental/lead to loss of function). Patients with *BRCA*1 and/or *BRCA*2 mutations that are considered to be non detrimental (eg, "Variants of uncertain clinical significance" or "Variant of unknown significance" or "Variant, favor polymorphism" or "benign polymorphism," etc) will not be eligible for the study.

Insights

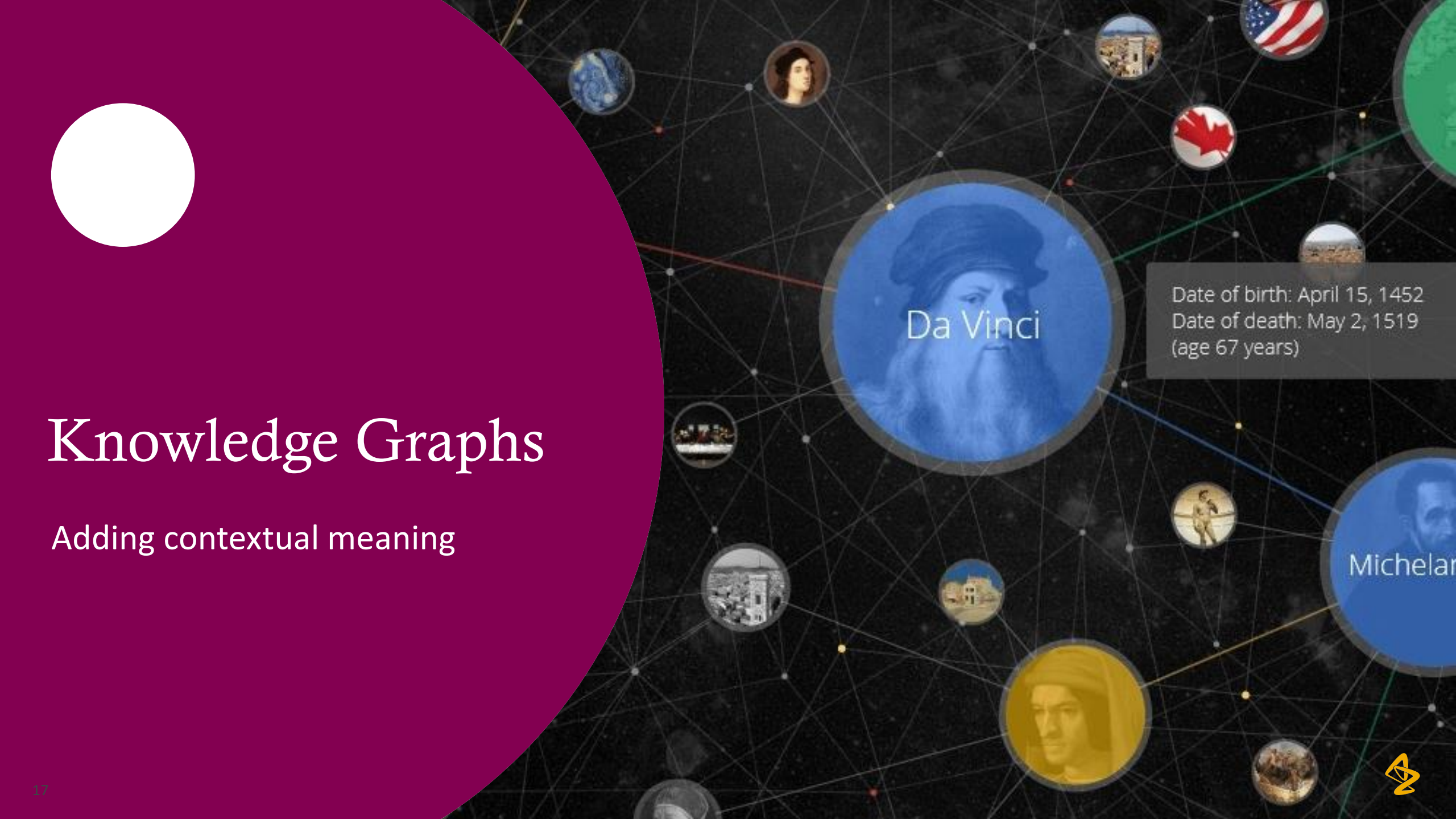
Facts placed in context of the text surrounding it i.e. In a clinical study the implications of a type of measurement can be contextual to the indication i.e. Six minute walk in Asthma measures time to attack, while in COPD it can be the distance walked.

Facts



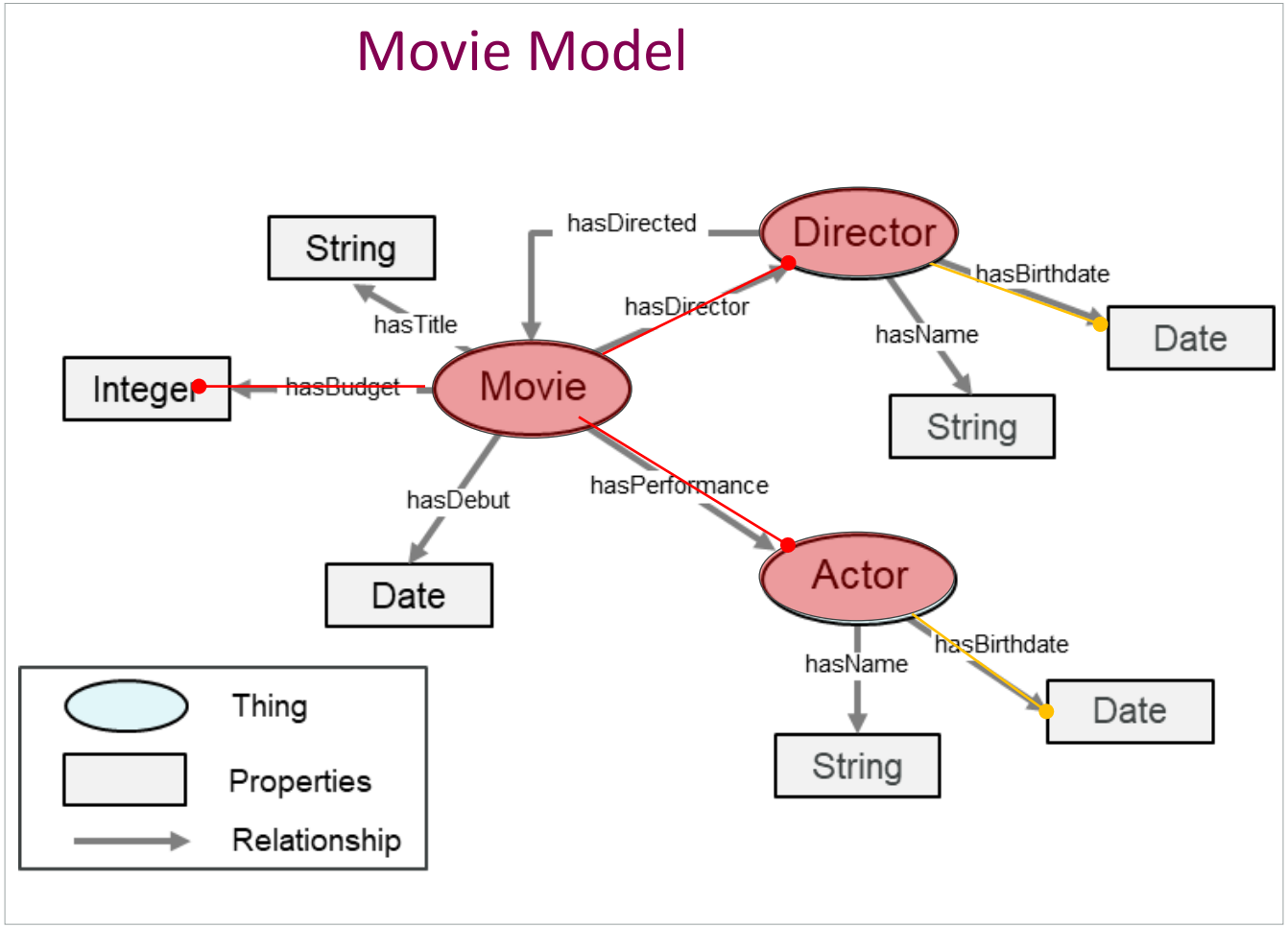
Knowledge Graphs

Adding contextual meaning



Knowledge Graph as a simple concept

A knowledge graph is a model of a domain of knowledge – in this case "Movie"

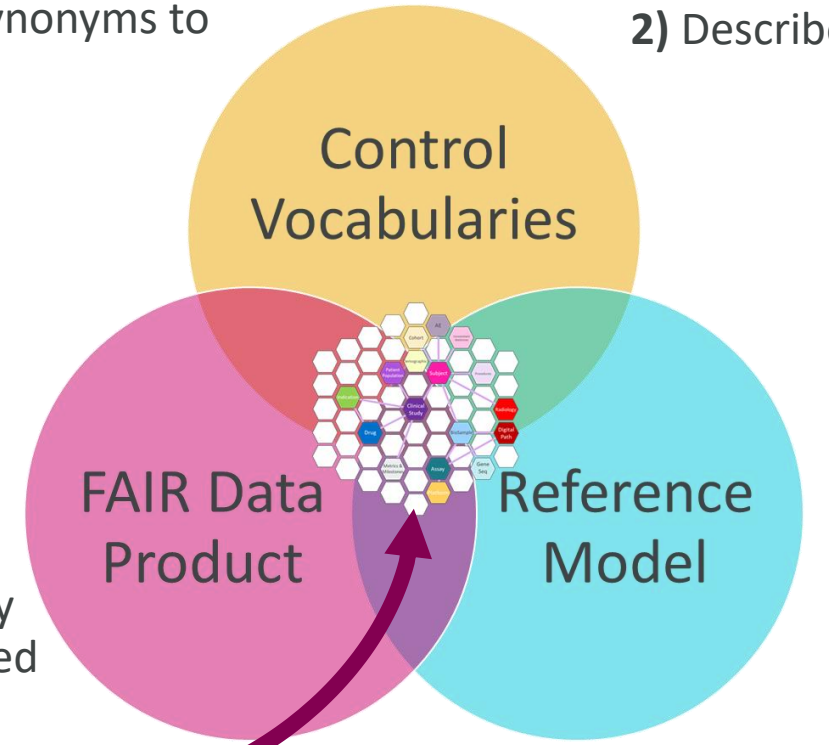


- A knowledge graph consists of **things** (Director, Movie, Actor), the **relationships** between things (hasDirector, hasPerformance) and **information** about a thing (date, title, name, etc)
- This model can be used to **discover** how remote parts of a domain relate to each other providing **insights** that might not be initially obvious.
- **Knowledge = Data + Meaning** (Linked), the richer the linkages, the higher the knowledge value.
- In general we make a thing a **node** if it will have **multiple relationships** to other **things** i.e. Movie, Director or Actor

Which movies that had a budget >$1M have performances from actors born before 1960 and were directed by people born before 1950?

Building knowledge graphs

Start with meaning



1) Start with meaning by identifying the things that matter

2) Describe the things identified

3) Map terms and synonyms to unique identifier

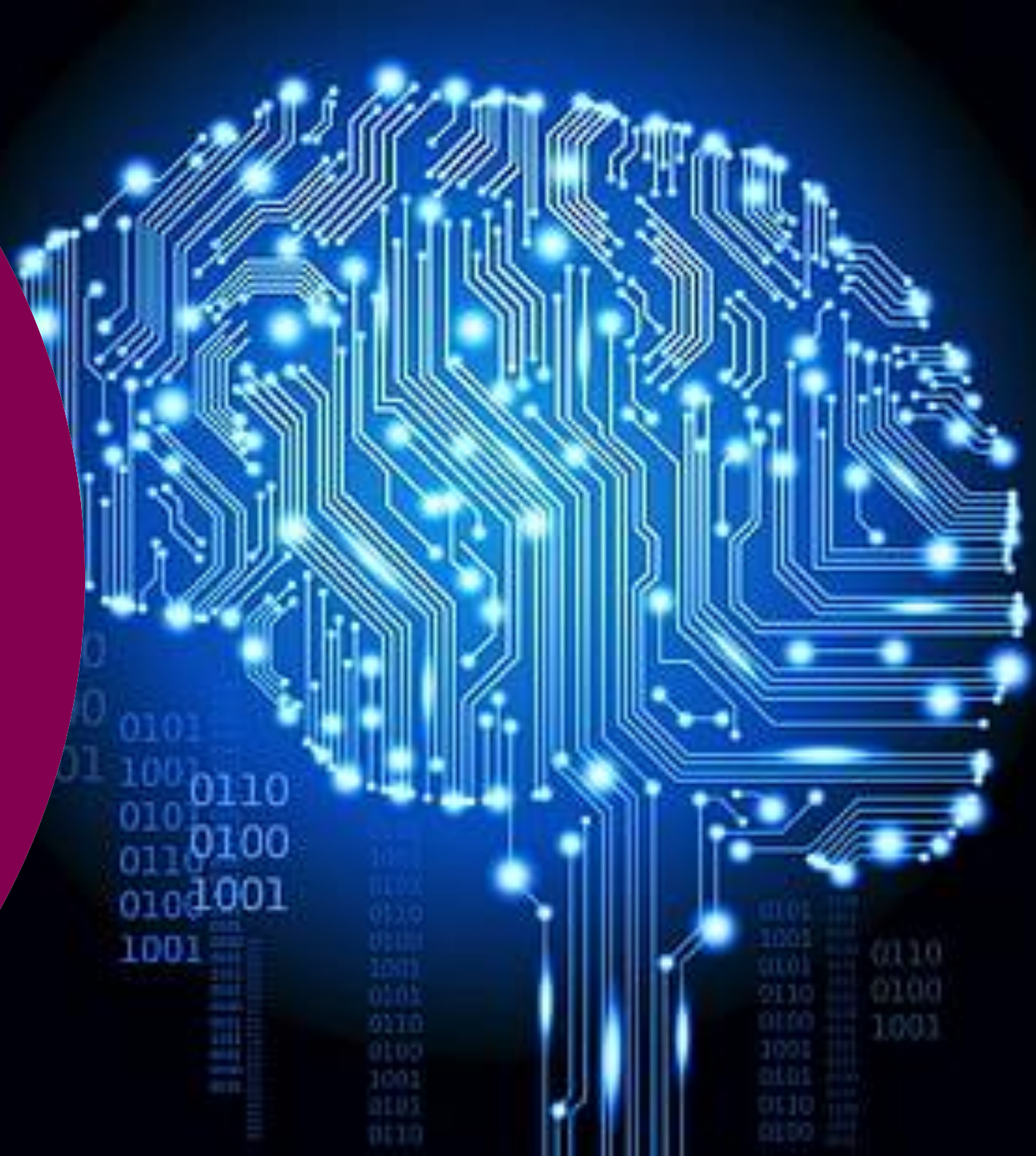
4) Create interoperability by enriching data with preferred term and unique identifier

5) Populate the knowledge graph

Reuse data

Bringing it back to AI

Improving accuracy using
FAIR Data-centric data



Minimising ambiguity in the output from LLMs

“Weak AI - non-sentient computer intelligence or AI that focuses on one narrow task” - Wikipedia

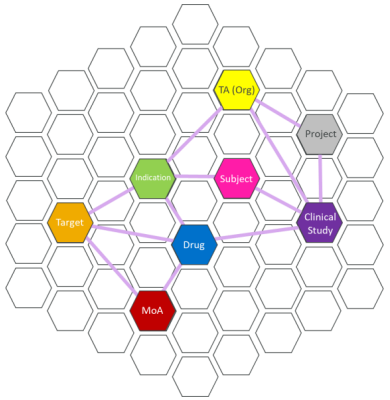
Weak AI
- Models
- Natural Language Processing
- Machine Learning

Statistical

Replicating what individual human brains do

Vs

Symbolic



Replicating the knowledge of our civilization

Stephen Wolfram

Retrieval Augmented Generation (RAG AI)

Constrain generative AI

Ask

Semantic search

Question (prompts) plus contextual information

Generative output

Generative answer based on contextually relevant and disambiguated data

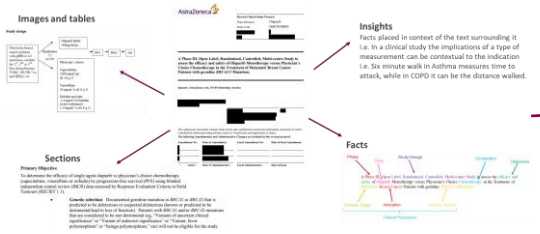
Contextually relevant information

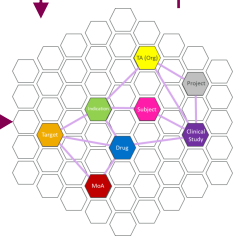
FAIR Data

Structured

Study_ID	Study_ID__URI	Indication	Indication__URI	Drug	Drug__URI
D1234C00001	https://pid.astrazeneca.com/1/12345	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D1234C00012	https://pid.astrazeneca.com/1/48373	Non small cell lung cancer	https://pid.astrazeneca.com/Indication/23456	Tagrisso	https://pid.astrazeneca.com/Product/965723
D4568L00007	https://pid.astrazeneca.com/1/97538	Diabetes type 2	https://pid.astrazeneca.com/Indication/9857	Forxiga	https://pid.astrazeneca.com/Product/853584

Un- & Semi-Structured





Knowledge graph

Take home message - Invest in your data

- Reduce ambiguity by applying controlled vocabularies
- Increase contextual relevance through knowledge graphs

- Minimises hallucinations and increase accuracy of generative answer