DATA GOVERNANCE AND MASTER DATA MANAGEMENT CONFERENCE EUROPE

11 - 14 March 2024 | London, UK

Please score and comment on this session and speaker in the event mobile app

IRMUK

HOW TO ACHIEVE DATA FABRIC AND DATA MANAGEMENT AT SCALE

Through Operationalising Semantic Discovery

INTRODUCTIONS

Open Knowledge Graph SIG Steer Co

Open Knowledge Graph INNOVATION LAB

DCAM

EDMCouncil DCAM CERTIFIED TRAINER

CDMC Cloud Data Management Capabilities

EDMCouncil

EDMCouncil CDMC CERTIFIED v1 TRAINER

CDMC+

Father and Husband

Endurance Cyclist

UK / London

Enthusiast & Advocate of the Hero's Journey

Podcast

Experience & Interests

- Data Management
- Data Architecture
- Knowledge Graph
- Business Architecture & Analysis
- Client Relationship Management
- Financial Crime Intelligence

Roles

- Head of Information Architecture
- Global Head of Business Data Architecture
- Global Business Architect
- Global Head of Governance, Financial Intelligence Unit

SME Data Speaker

Google Cloud Next '22

DataVision

DataManagement Summit London an A-Team Group event

ortecha DATA LEADERSHIP SERIES JUNE – DECEMBER 2023

Gartner Data & Analytics Summit 22 – 24 May 2023 | London, U.K.

NewScientist

cdo CHIEF DATA OFFICER MAGAZINE

POOLPARTY SUMMIT 2023

IRM UK Data Governance Conference & Master Data Management Summit

IRM UK Enterprise Architecture & Business Process Management Conference

BT Group

HSBC

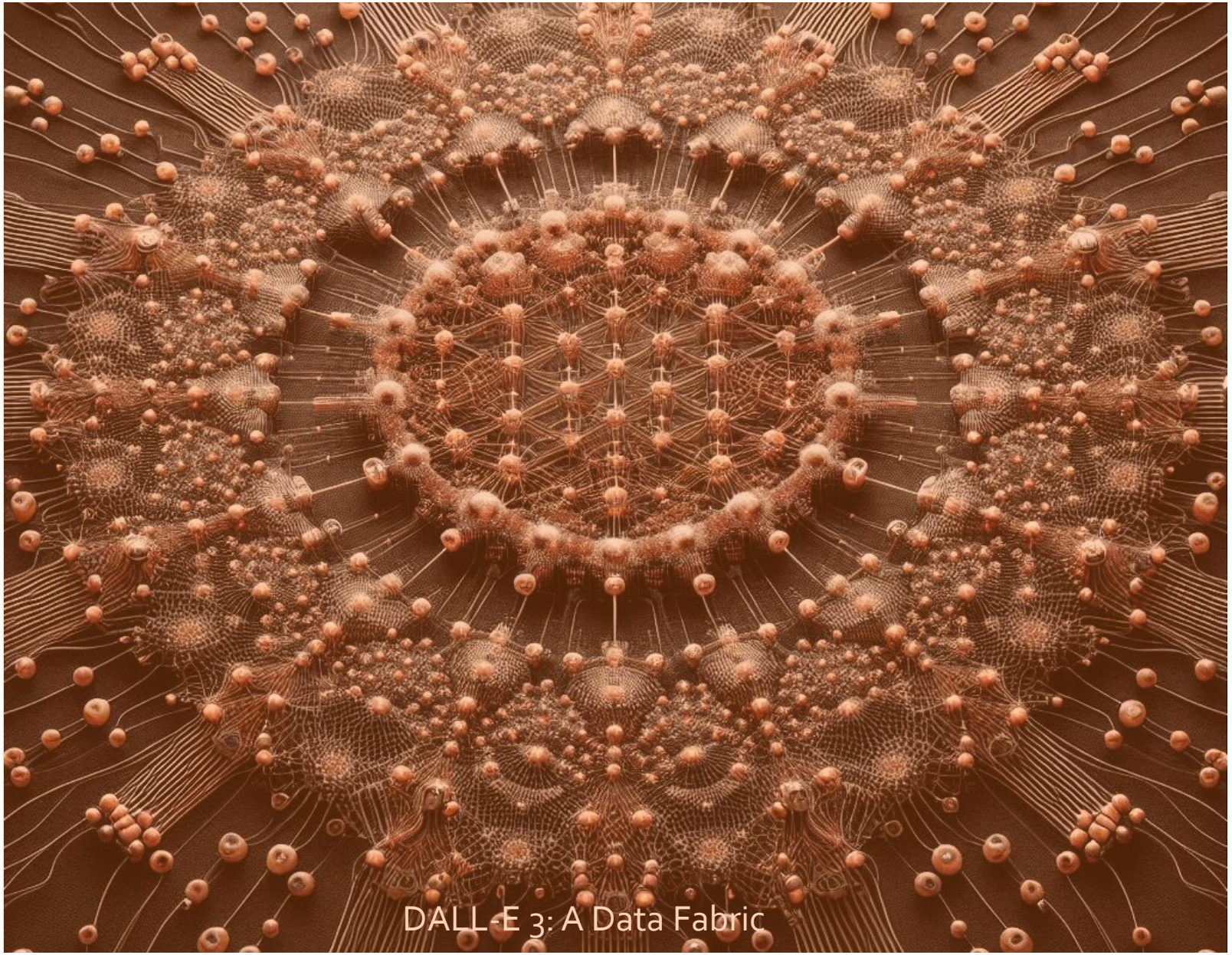
citi

RBS

UBS

OVERVIEW

1. **What is a Data Fabric?**
2. Why metadata is at the heart of the Fabric
3. What metadata do you need?
4. How to find the metadata?
5. Economies of scale and precautionary tales

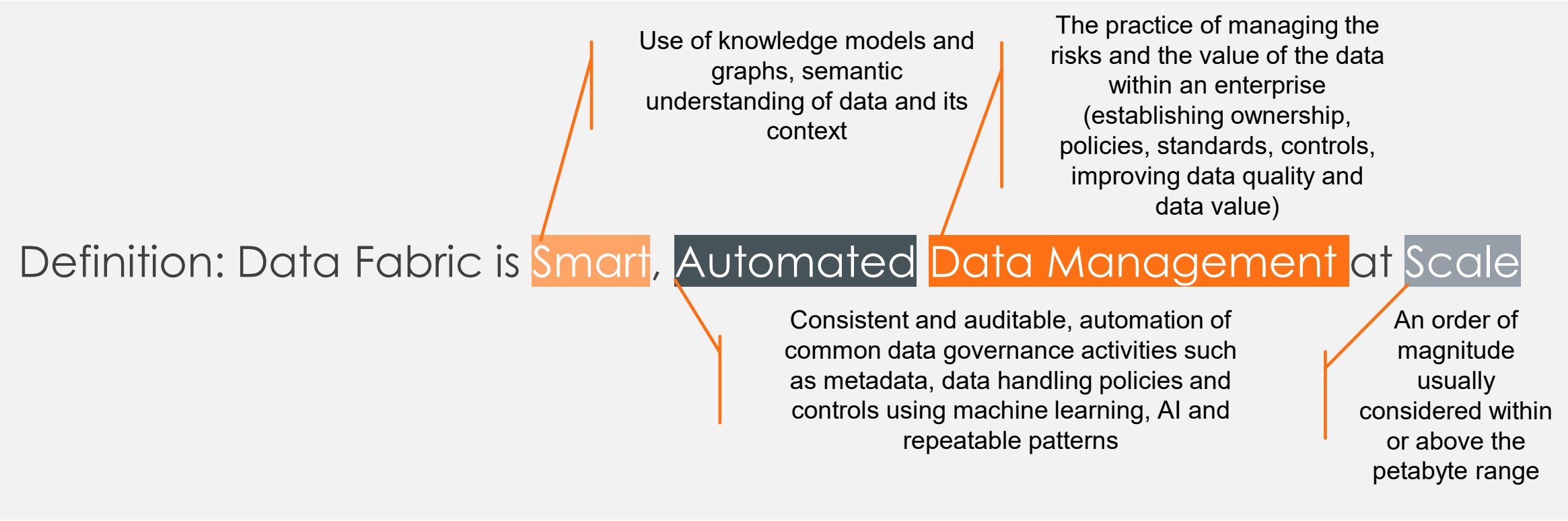


DALL-E 3: A Data Fabric

WHAT IS A DATA FABRIC

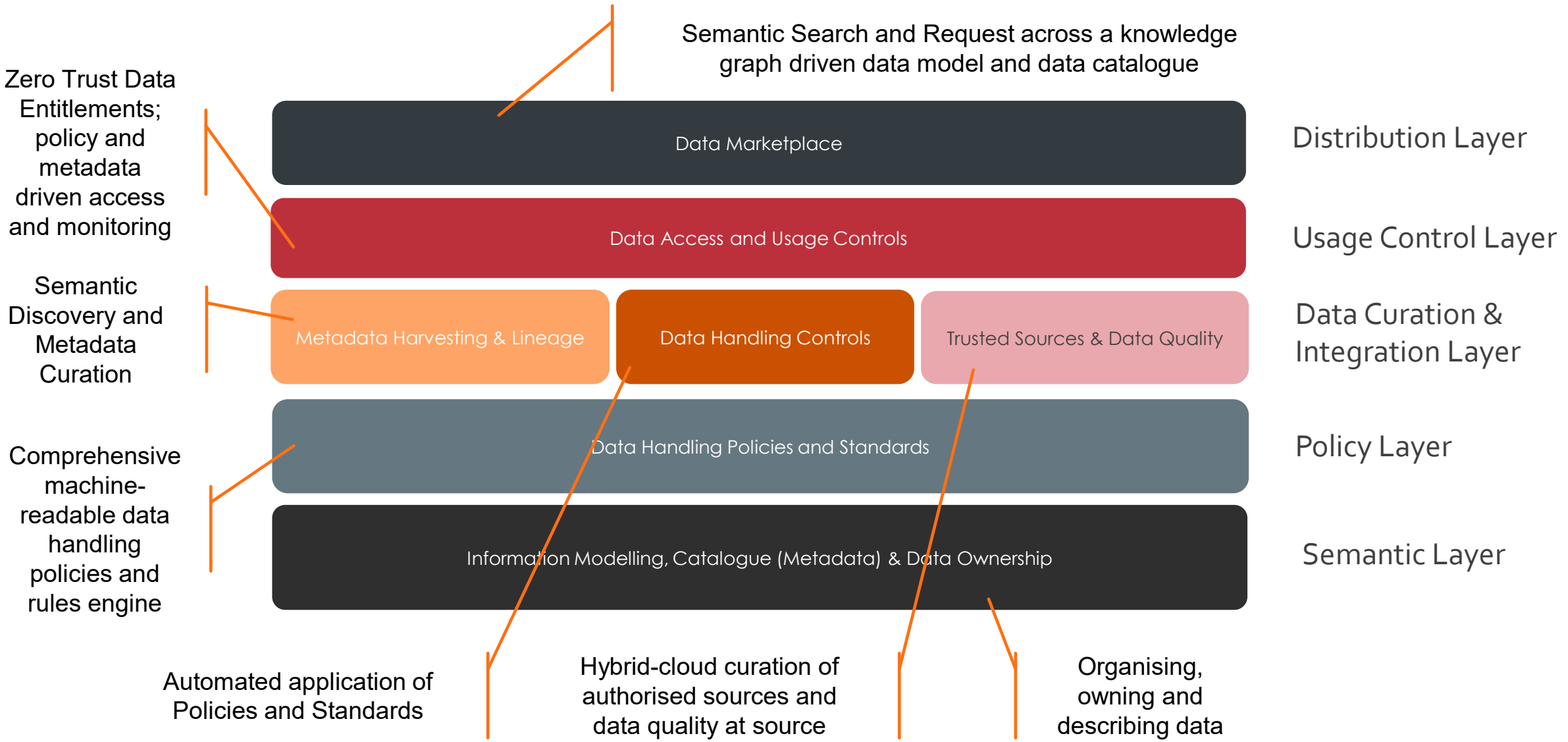
...and why you need one!

DATA FABRIC: A WORKABLE DEFINITION

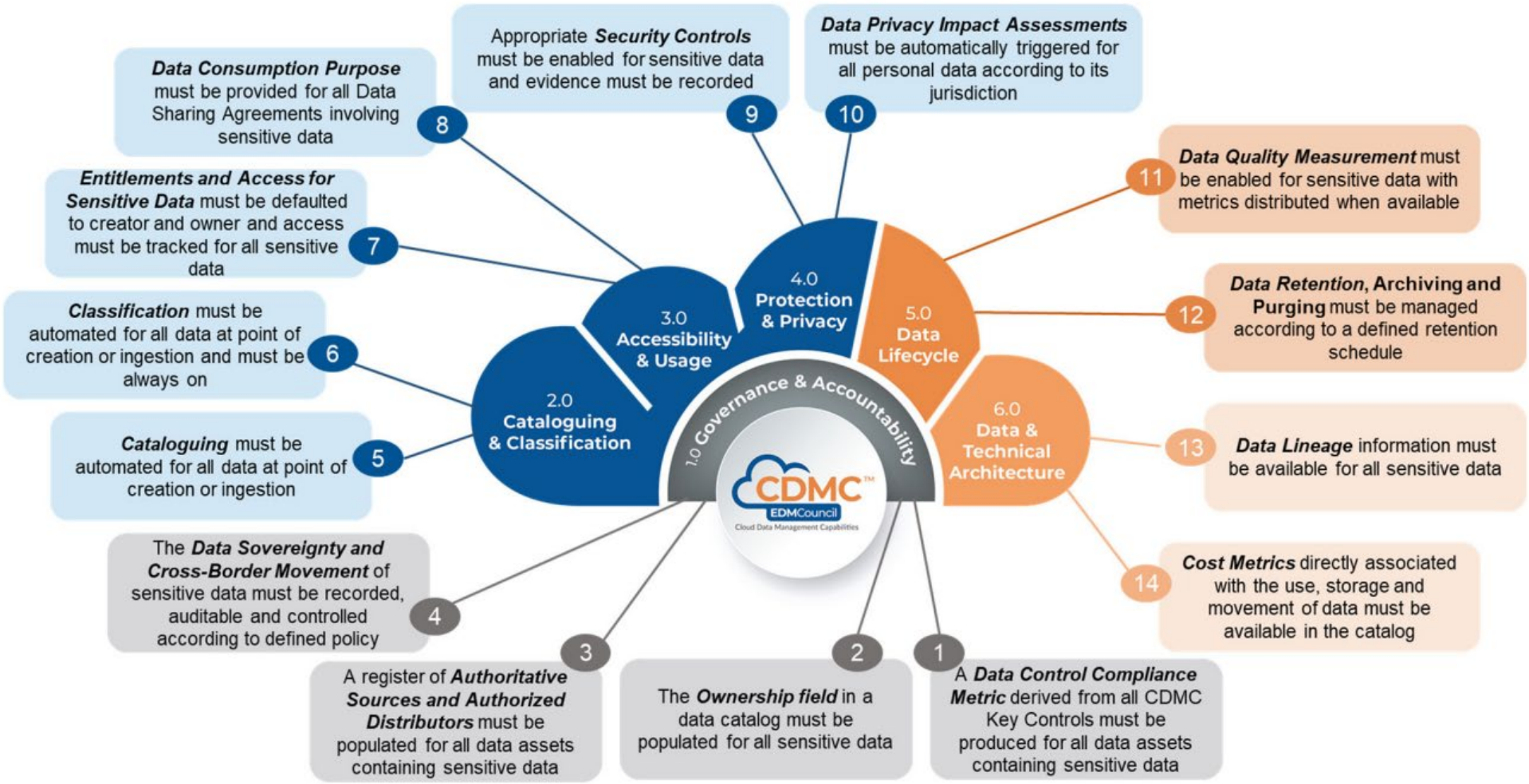


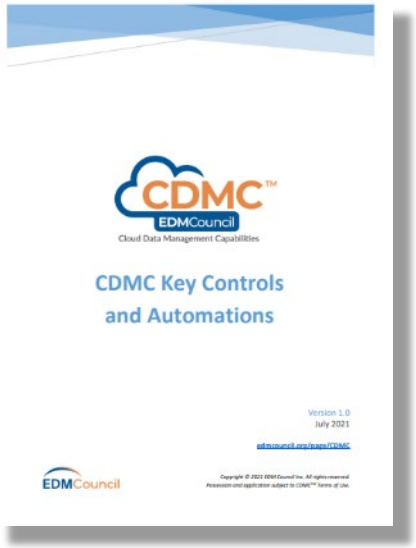
Point to Note: Data Fabric doesn't replace the need for data management; it automates it

THE ANATOMY OF A DATA FABRIC



CLOUD DATA MANAGEMENT CAPABILITIES – DESCRIBING A FABRIC





https://edmcouncil.org/frameworks/cdmc/

WHY METADATA IS AT THE HEART OF THE FABRIC

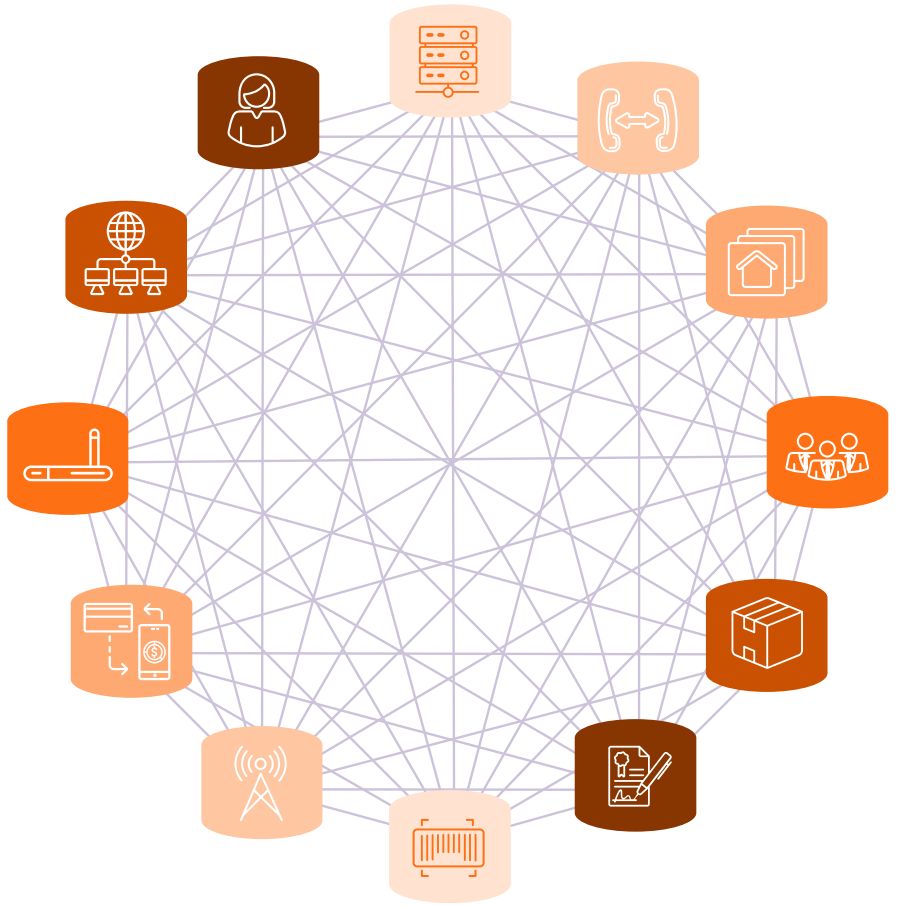
and flexible architecture

VISION FOR THE INFORMATION MODEL

"The Information Model is for our data what the organisation chart is for our people.

In a digital company data is one of our most valuable assets.

The information model needs to be as pervasive in our everyday thinking and activities as the org chart."



The Information Model provides Semantic Discovery with a rich set of semantic meaning that can be used to inform entity resolution and automate controls while simplifying the way we manage data.

WHAT METADATA DO YOU NEED?

Encoding the business rules

METADATA MATTERS – DATA MUST HAVE CONTEXT TO BE USEFUL

No Metadata

It could be anything.

Technical Metadata

May be true but how useful is it?

Business and Technical Metadata

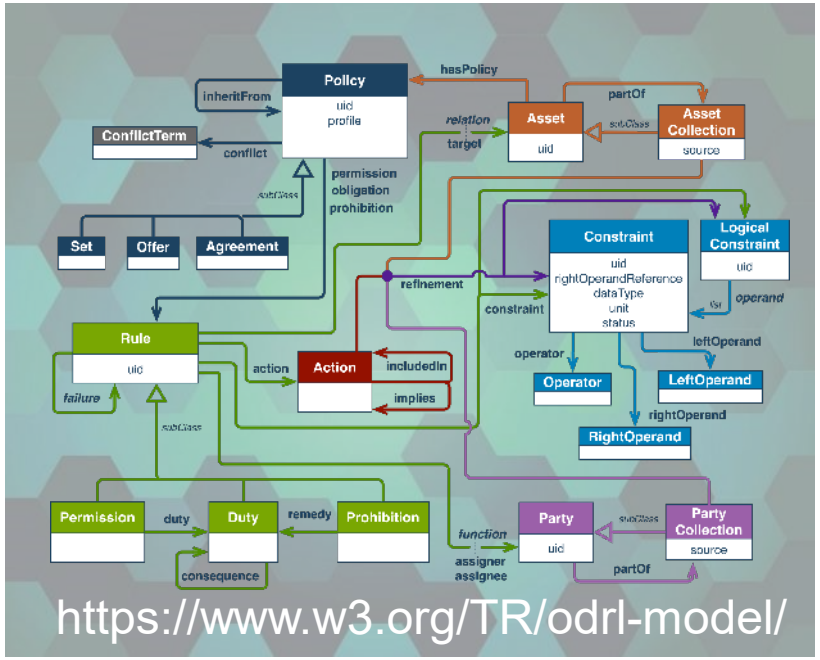
Labelled to be found, understood and used

 Special Credit to my friend, Jonathan Sunderland, at Ab Initio for the analogy that I like to riff on.

Machine-Readable Policy Examples

If this… then…. that… according to…

If (this) metadata then rule + action (that) according to policy source



Information Classification
If this is classified as "Highly Confidential" then it must be encrypted in motion and compute according to policy 34325

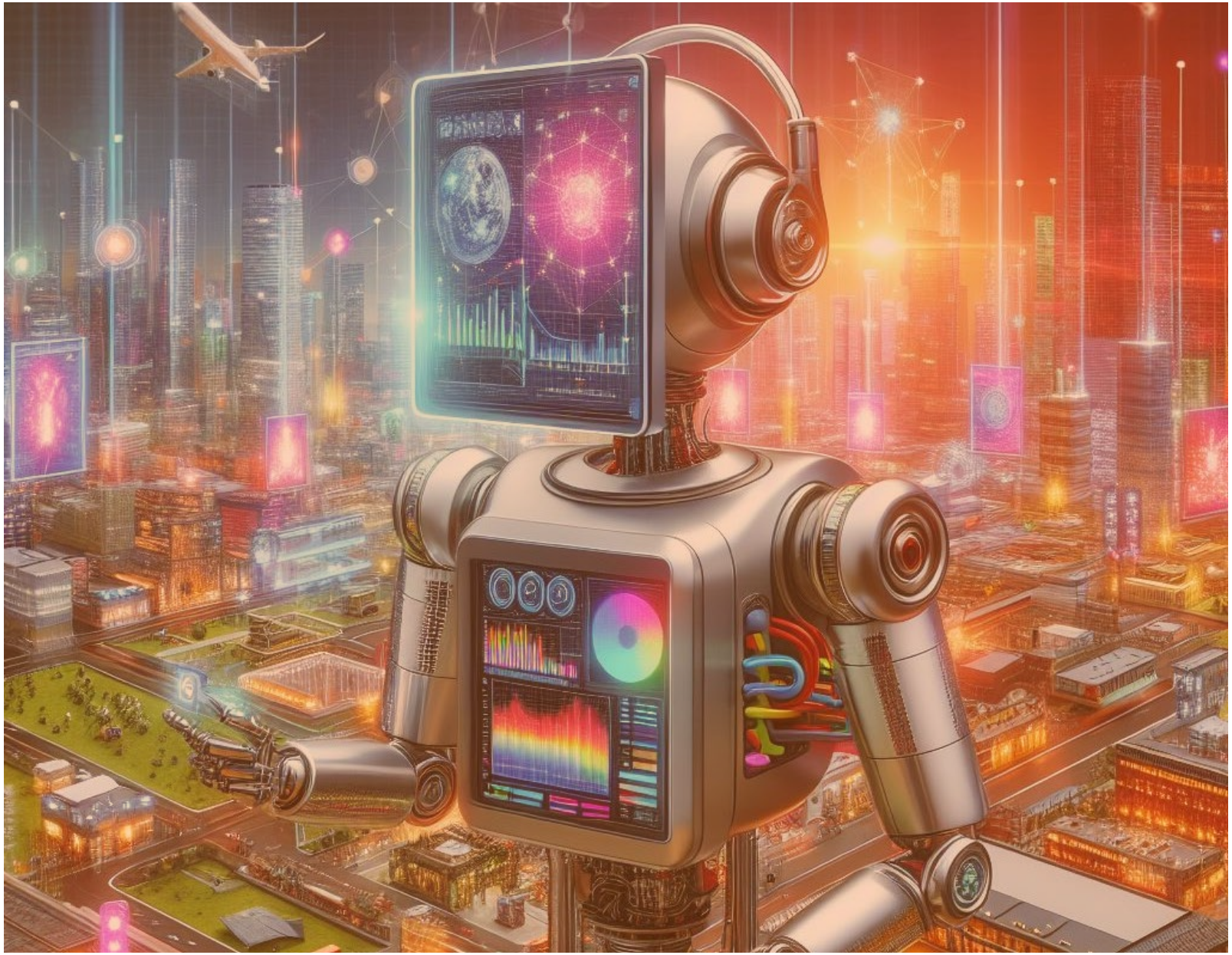
Record Retention (minimum)
If this is Payroll data it must be stored in the system of record for a minimum of 3 years from the end of the relevant tax year according to Gov.uk

Privacy
If this is classified as Personally Identifiable Information (PII) then it must be tokenised in standard storage public cloud data assets according to policy 13245

Data Sovereignty
If this is German Citizen Data without specific consent, then it must not be transferred outside Germany according to BDSG

A worked example of Metadata

	Data Handling Policy	Business Metadata	Technical Metadata	Data Handling Control
Roles	Policy Defined by Policy Owner -> Interpreted and entered into Rules Engine by Metadata ops	Business metadata interpreted by metadata ops and confirmed by Policy owner	Accountability for tagging of system data by system owner – can be delegated or automated (semantic discovery)	Accountability for applying controls of system data is on system owner but can be delegated or autoamted
Description	Data Handling controls recorded in business rules engine.	Business Terms and required metadata recorded against the control in the catalogue	Technical metadata and data dictionaries linked at field level to relevant business terms	Controls applied at a system level (where metadata is detected) and implemented based on the Enterprise-wide rules
Example	If this data is classified as Personally Identifiable Information (PII) then it must be tokenised in standard storage public cloud data assets according to policy 13245	• Customer Name is classified as PII • Definition of PII and Customer Name classifications stored in Catalogue as maintainable, versioned metadata	System X is in Public cloud and System X; table Y: Fname = Customer Name System X; table Y: Lname = Customer Name Sysemt X: table Y: Initial = Customer Name	Apply tokenisation to System X; Table Y; Fields Fname, Initial, Lname



HOW TO FIND THE METADATA?

Automation and Metadata Ops

WAIT A MOMENT…..CAN'T ARTIFICIAL INTELLIGENCE DO THIS?

- YES and NO
- Cost and proportionality
- This is AI (deductive AI) and can be augmented with generative AI
- At its heart the same principles apply

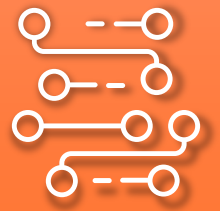
Metadata Ops and Where to Capture the Metadata

01

Capture at Source

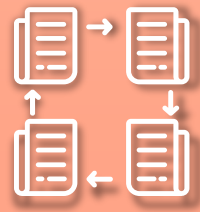
02

Capture in transit (Code, ETL)



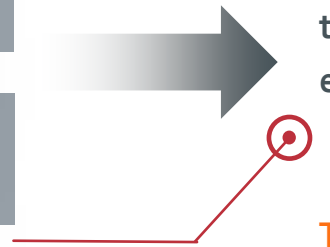
03

Propagate with the data



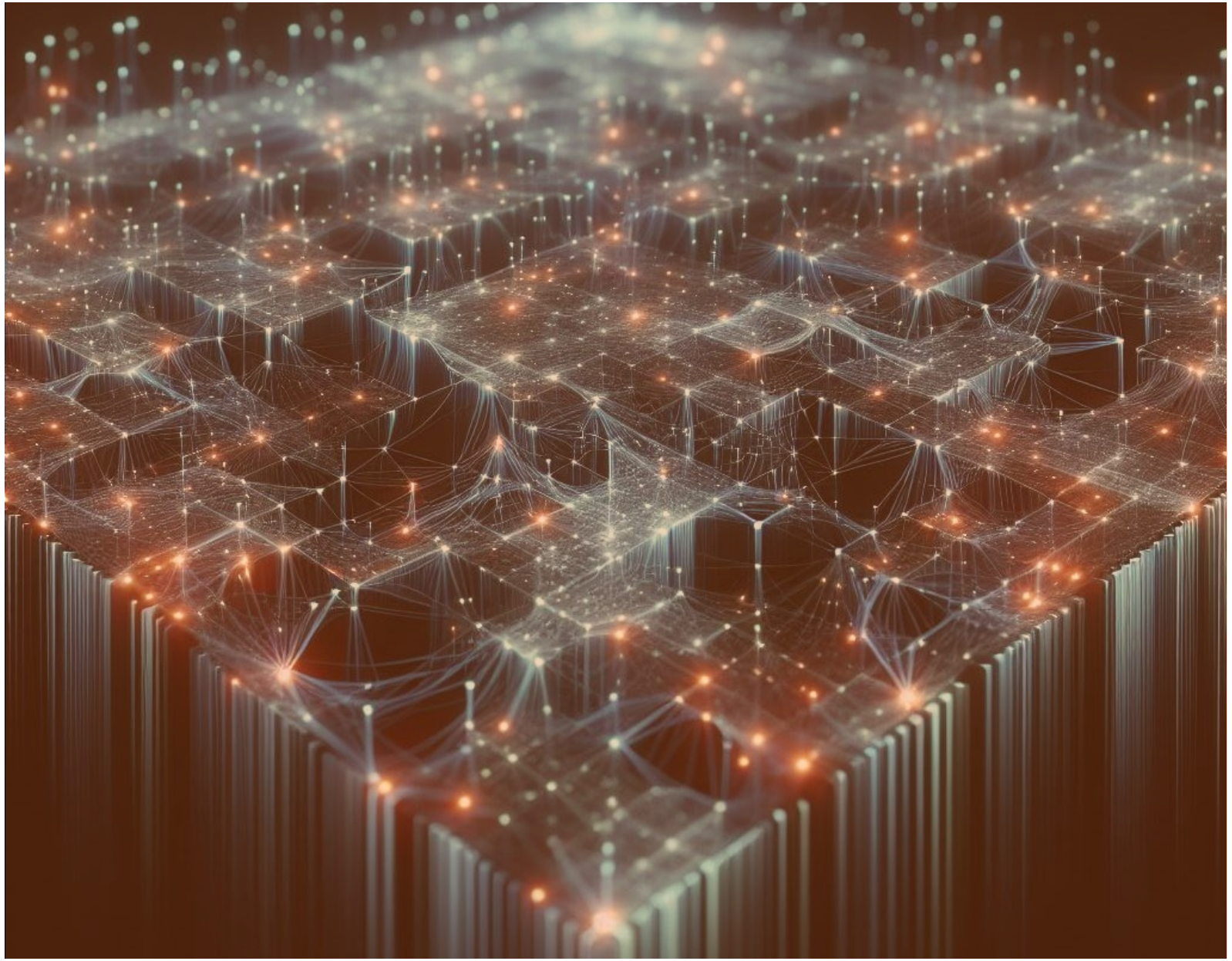
SEMANTIC DISCOVERY PATTERN EXAMPLES

Is the data a familiar shape?	nn/nn/nnnn -> looks like a date	Regex (data shapes)
Does the value look familiar?	Terry, Jon, Fred -> Looks like a name	Common value lists
Are there any clues in the system that is storing it?	How is the system, table, field described? Is it described as a customer management system, table or field. Is it an abbreviation?	Common Abbreviations, Metadata Harvesting
Does the range tell me something?	19/04/2021 – may be too young to be customer d.o.b. but could be a transaction	Applied Logic
Does the dataset it belong to match others I have seen?	Although these are all 8 digit numbers, 70% of them match exactly to a data set I have already catalogued as account number	Data Set Comparison / Fingerprinting
Does the data context tell you anything?	The value is a single letter but in the context of a table where first and last name are present this is most likely an initial.	Grouping / Contextual Characteristics



Combined to form a level of confidence that is confirmed by end users.

The end user confirmation provides a stronger feedback loop for training the algorithm.



ECONOMIES OF SCALE

And precautionary tales

KEY CONSIDERATIONS FOR IMPLEMENTATION

Beware the "Free Text"

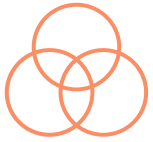
Prioritise for scale and impact of discovery

Roll-back and History

Test for what you know

Humans in the loop (Metadata Ops)

The dark side of scalability



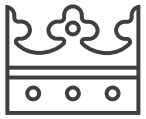
Utilise delta testing versus rescan

Read data once and run analysis on the profile (at rest)

Federate metadata ops creation to domain experts

Deploy patterns via central configuration

Monitor and Govern



Recognise that manual tagging may be required

SEMANTIC INFLEXIBILITY – WHAT TO AVOID

We can't begin until the data model is done

But what if it changes?

We need a quote to manually tag the metadata

Erm.. We have new products & new data needs + regulation has changed on privacy and ethics

Actions

Data Model is "done" & metadata requirements are "complete"

Manually tag the data in target systems

You can't change requirements

Change the reqs at time and cost

We can't tag until we are aware of ALL metadata we will need & it won't change

We need to tag all data in every environment to control the data

We need to manually retag everything all over again.

Outcomes

Long delays based on waiting for a full and final set of requirements

Long delays, regrettable spend & human-error from manual tagging of data in multiple locations

A brittle framework unwelcoming of iterative improvements or change

SEMANTIC FLEXIBILITY

What do we know that we need so far?

We have these needs so far…

Let's build and test a pattern

Erm.. We have new products & new data needs + regulation has changed on privacy and ethics

Start building new meta data automation based on known requirements

Let's deploy once and measure everywhere

Let's build a new pattern and test it

Let's build and test a pattern and only manually tag as a last resort

Let's deploy and monitor the detection pattern

What more have you got for me?

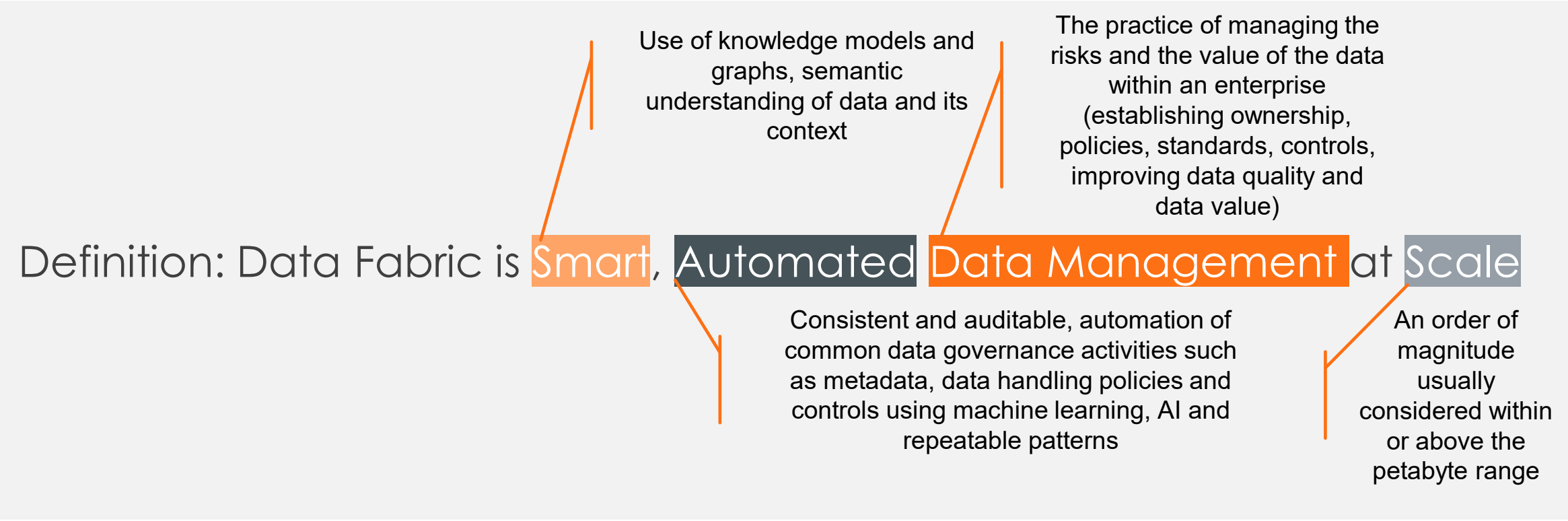
Outcomes

Can get started from the very first requirement.

Every semantic discovery design is iterative + can be tested once and deployed everywhere

The approach is flexible, iterative and grows with your organisation's evolving needs

DATA FABRIC: A WORKABLE DEFINITION



Point to Note: Data Fabric doesn't replace the need for data management; it automates it

LET'S STAY IN TOUCH

LinkedIn

DATA GOVERNANCE AND MASTER DATA MANAGEMENT CONFERENCE EUROPE

11 - 14 March 2024 | London, UK

Please score and comment on this session and speaker in the event mobile app

IRMUK