



Ontology and Graph Database

The perfect combination for better data
analytics

Dr Peter Tormay





The value of data

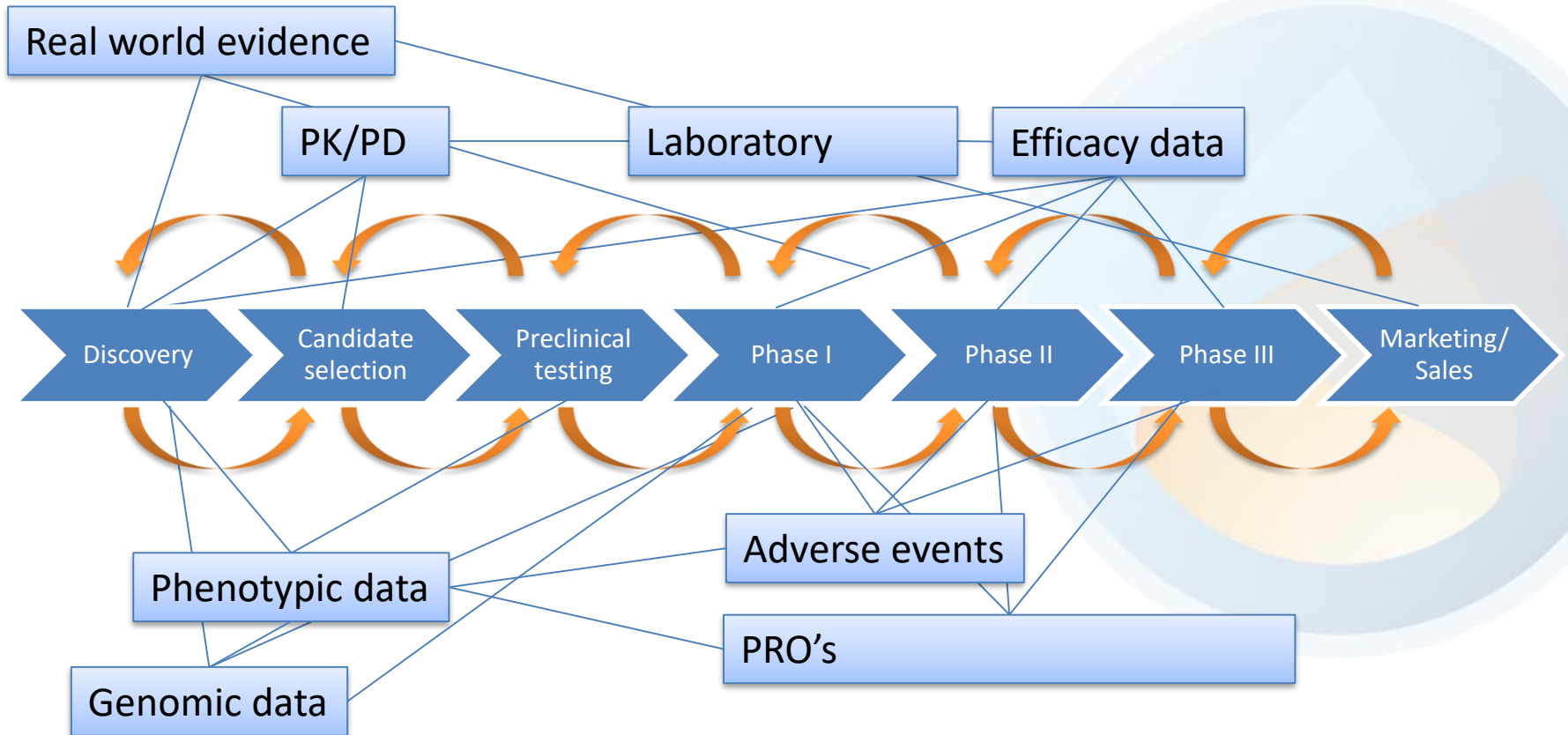
Pharmaceutical value equation

$$P \propto \frac{WIP \cdot p(TS)}{CT \cdot C} \cdot V$$

P	= Productivity (ROI)
WIP	= Work in progress
p(TS)	= probability of technical success
CT	= cycle time
C	= cost
V	= value (Effectiveness)

Paul et al. Nature Reviews Drug Discovery, Vol9:3 203-214

The value of data



Network of data – different data types and sources



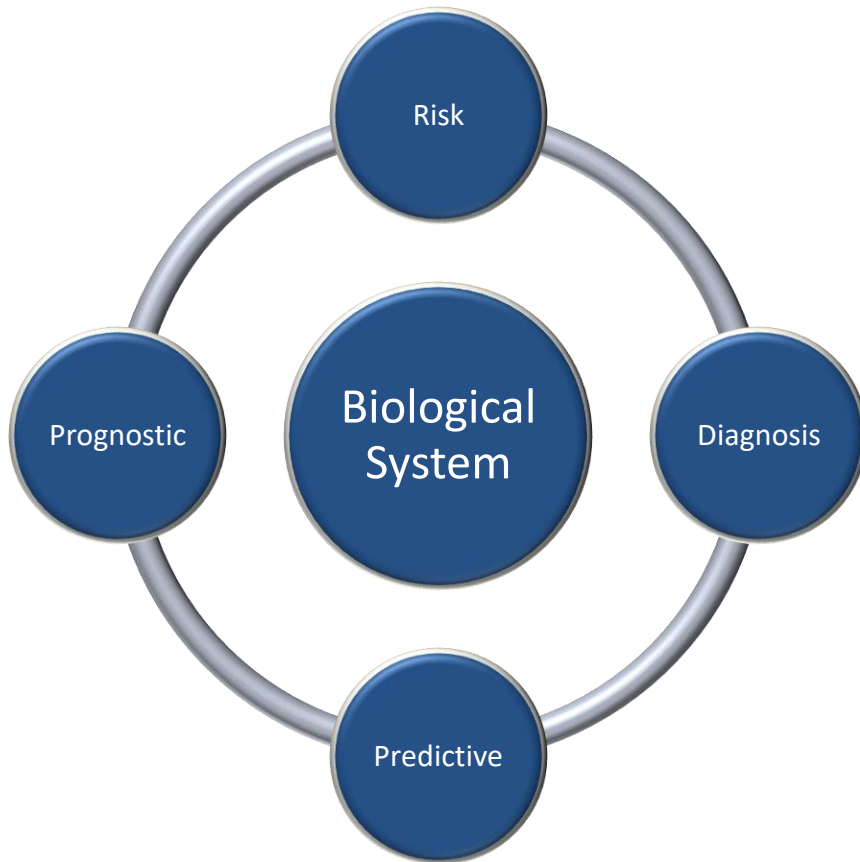
Personalised medicine



Targeted Therapy / Stratified Patient cohort

Biomarkers

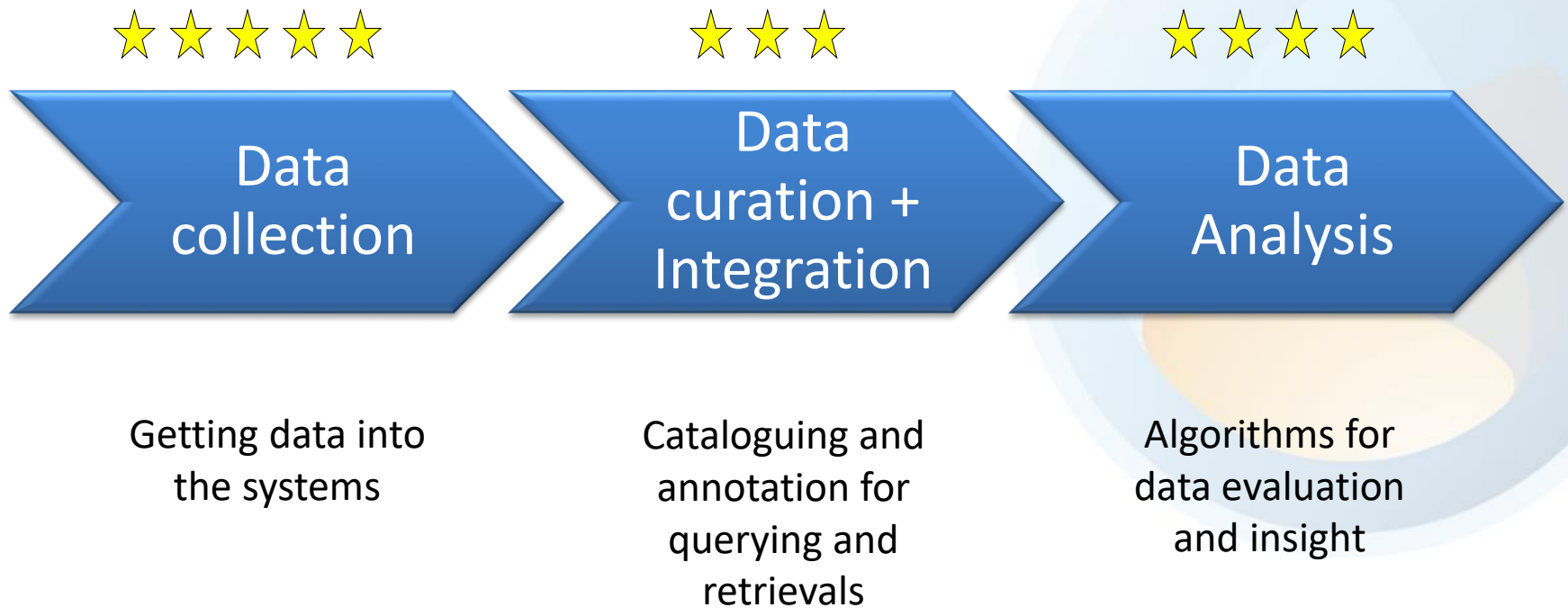
From descriptive to predictive and prescriptive analytics



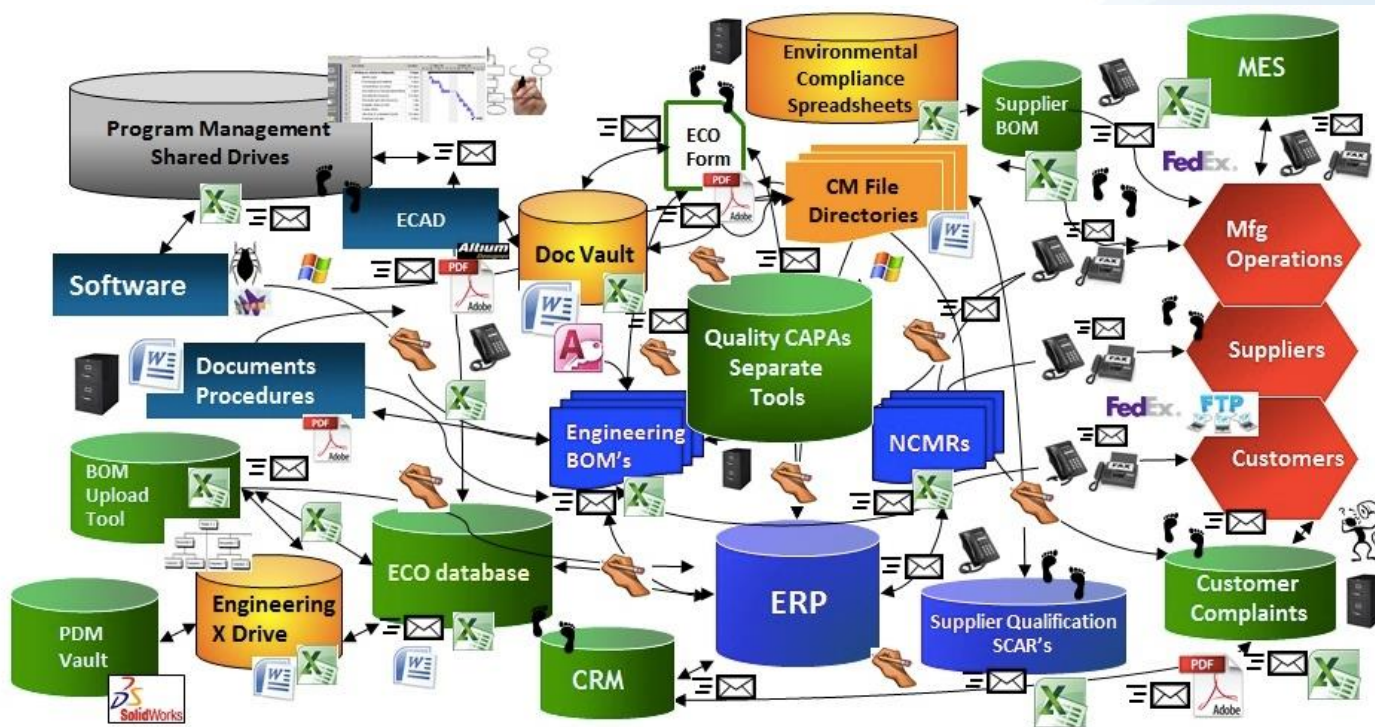
- (Bio)Marker are the core indicator in the life cycle of a drug from target validation to market
- Identification of trends, patterns and correlations



The process of data analytics



Data collection



Unconnected Silos

Relevance



From 2013 to 2020, the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years

EMC Digital Universe with Research & Analysis by IDC

[The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things](#)

April 2014

What proportion of this data is accurate and relevant?



Structured and unstructured data

miss sophie's class were learning about punctuation they knew they needed to remember to use capital letters for proper nouns and full stops at the end of sentences but didn't always bother to put them in please also put new speech in a new paragraph reminded miss sophie before they began new speech has to have a capital letter too doesn't it asked flossie yes but not when speech continues replied miss sophie if the same person is still speaking I often forget punctuation before and after speech admitted one sensible child but I will try my best to get it right first time



Structured and unstructured data

Miss Sophie's class were learning about punctuation. They knew they needed to remember to use capital letters for proper nouns and full stops at the end of sentences, but didn't always bother to put them in. "Please also put new speech in a new paragraph," reminded Miss Sophie, before they began.

"New speech has to have a capital letter too, doesn't it?" asked Flossie.

"Yes but not when speech continues," replied Miss Sophie.

"If the same person is still speaking, I often forget punctuation before and after speech," admitted one sensible child, "but I will try my best to get it right first time."

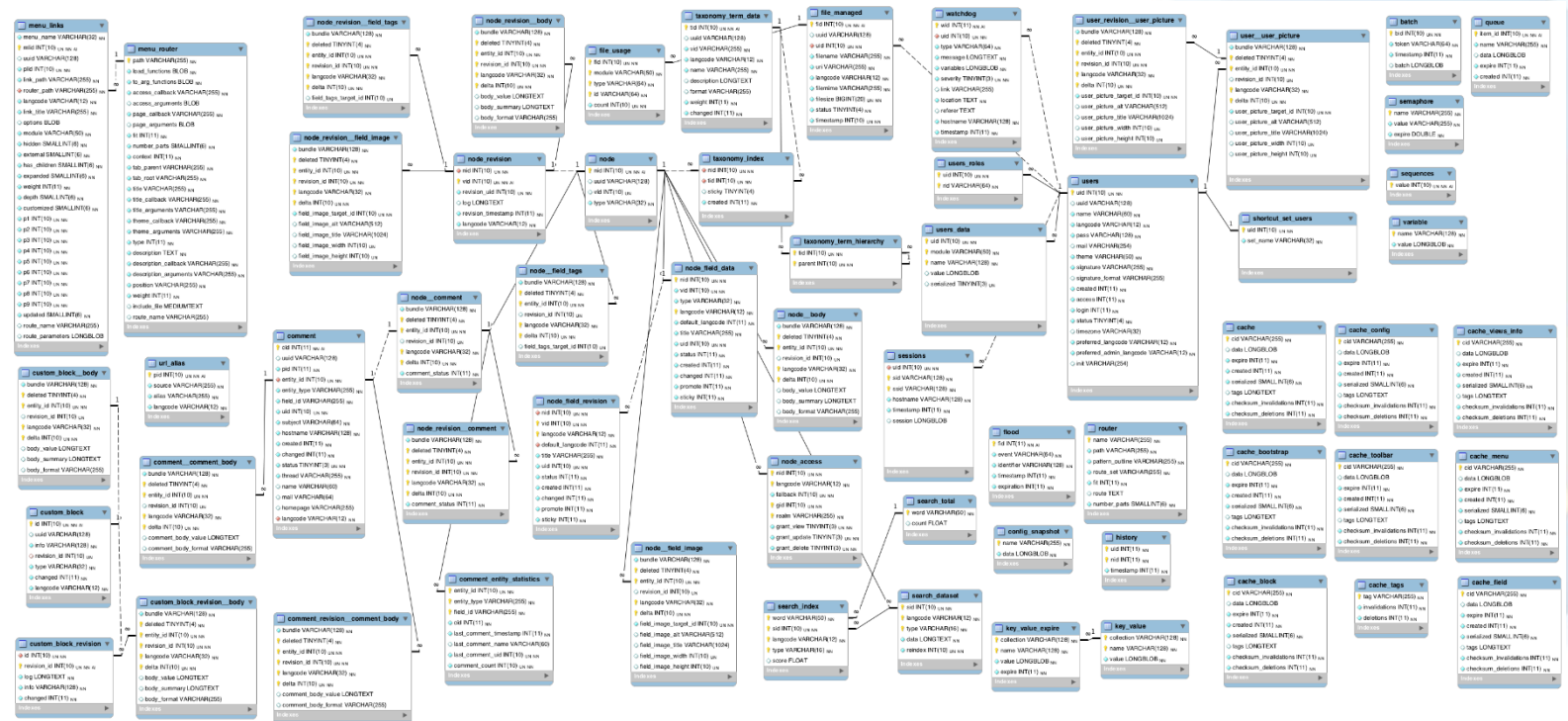
All data is structured
But our systems cannot deal with all these different structures



Scalability

- **Data volume:** Can the system cope with the increase in data volumes, i.e. handle more data of the same type efficiently.
- **Data complexity:** Can the system cope with increasing data complexity, i.e. can the system effectively handle the proliferation of data types? In order to effectively handle complex data, the system not only needs to be able to add these data types into the mix but also needs to be able to connect these different data types with each other.

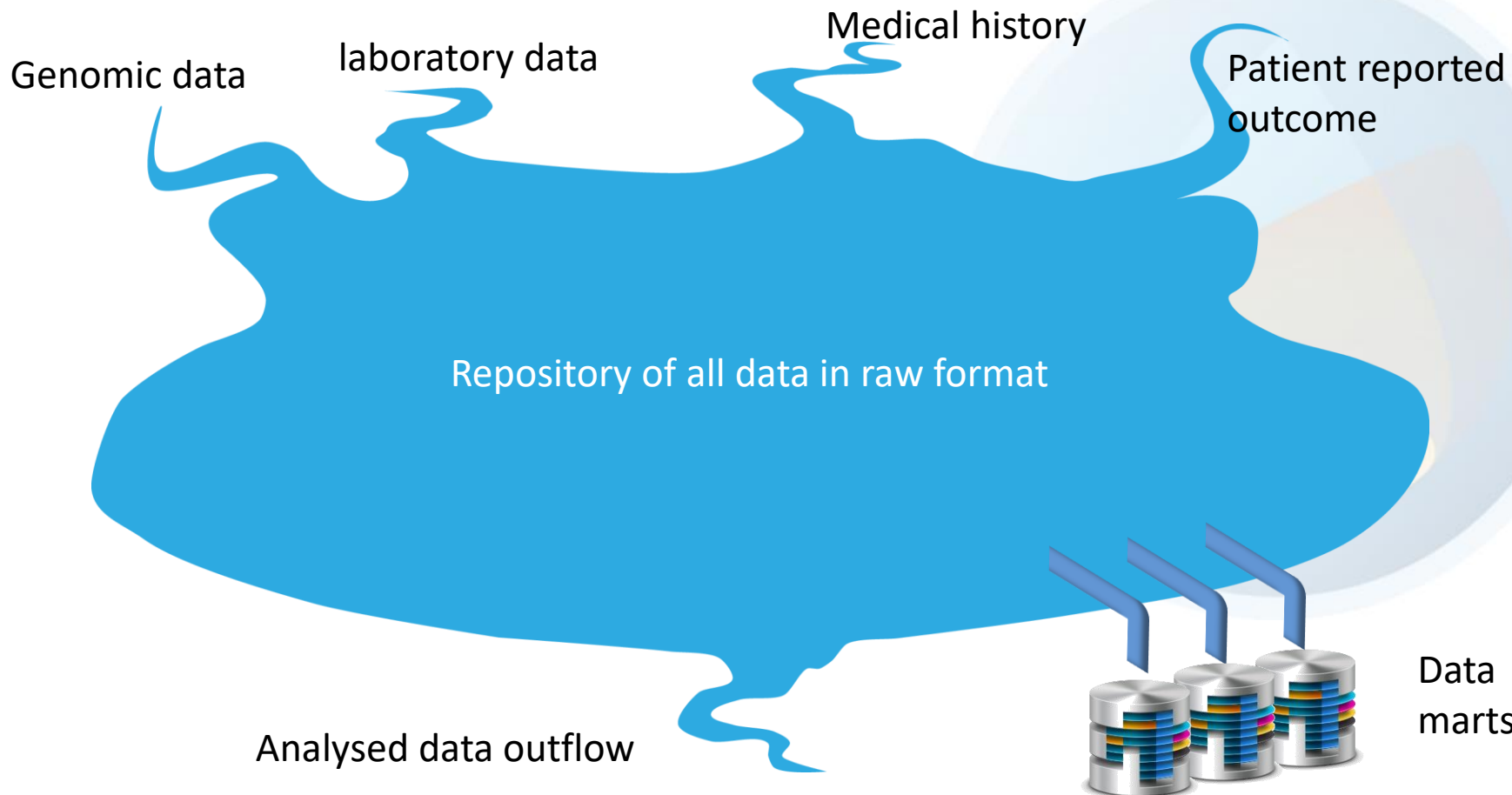
Relational Database





Data lakes

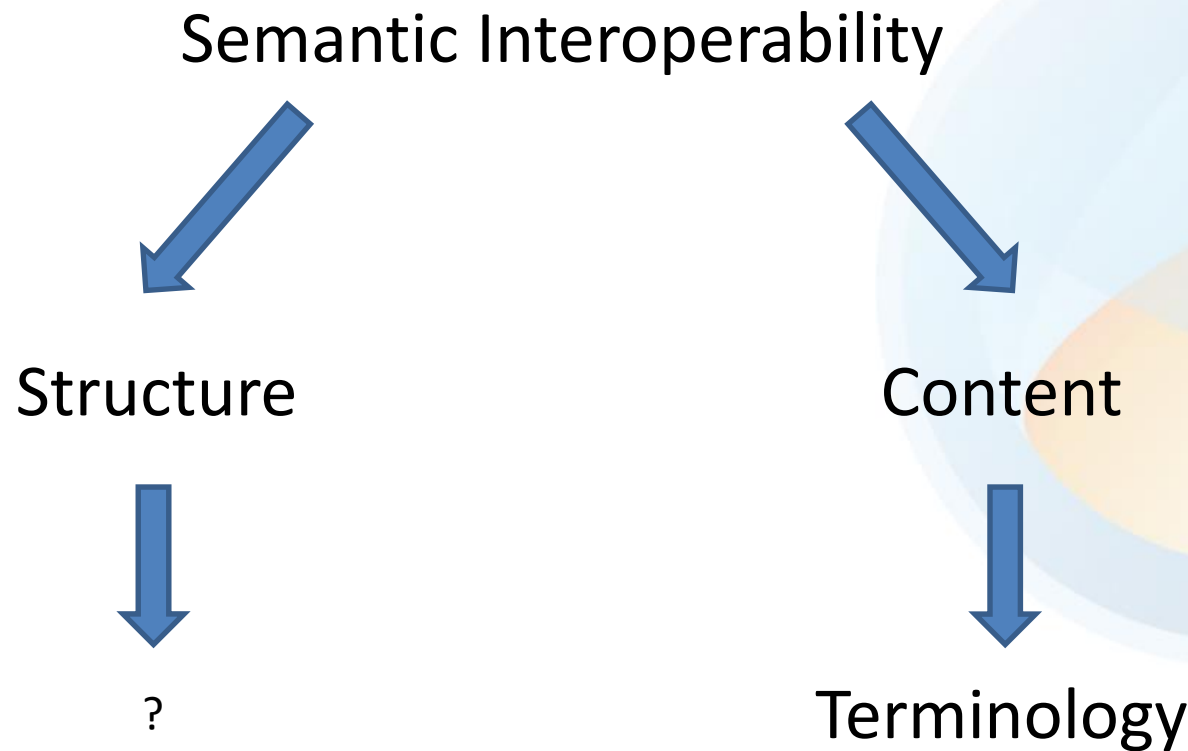
Inflow of multiple data sources in multiple formats





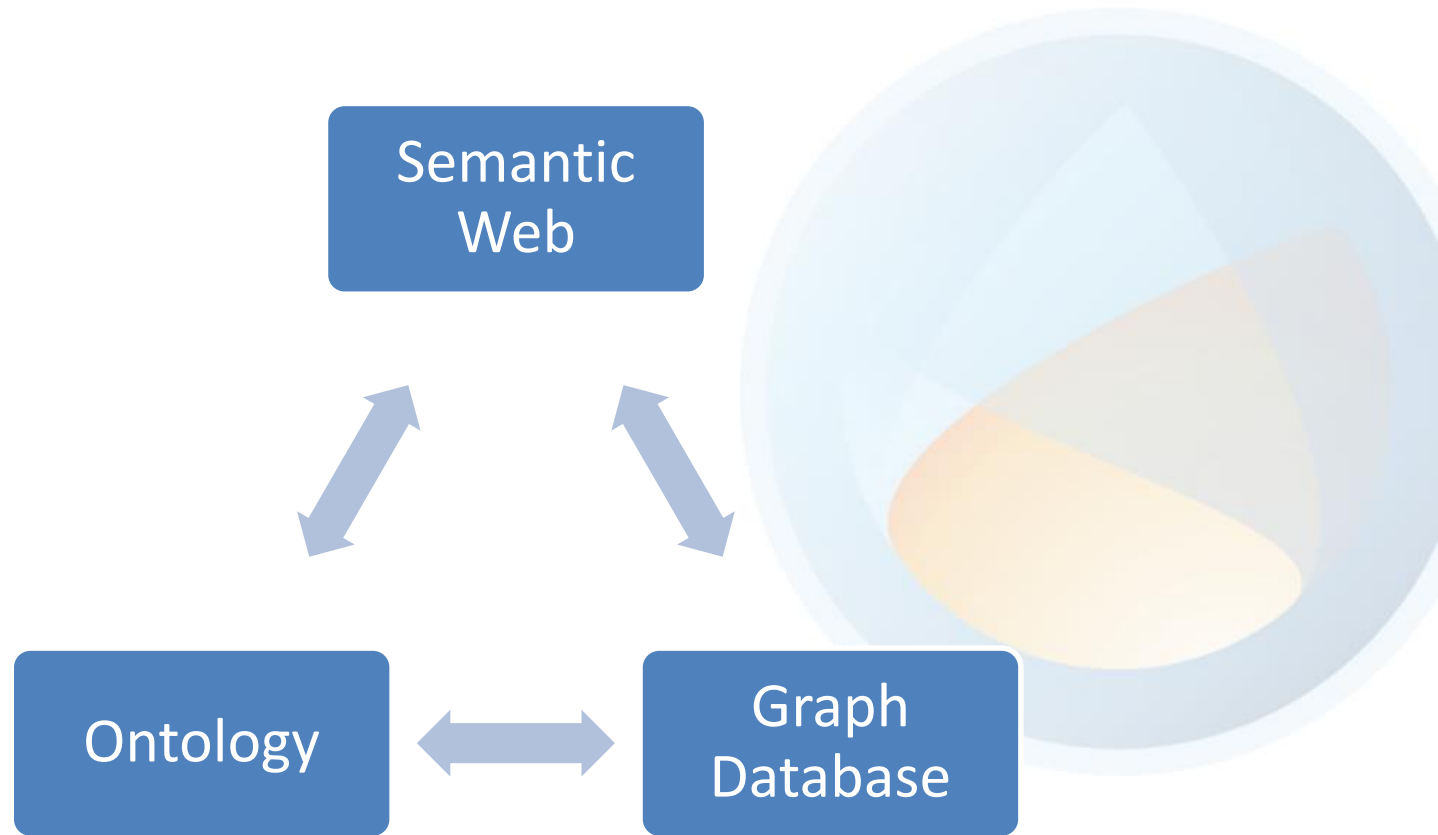


Semantic Interoperability





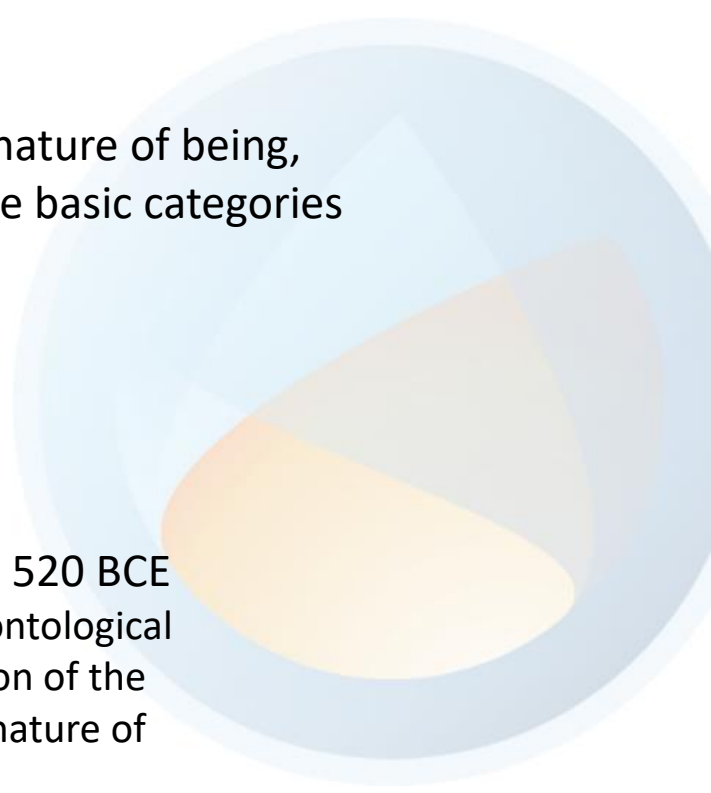
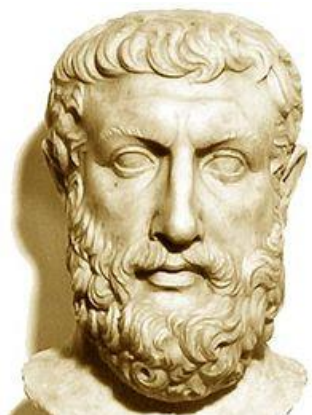
Data model for effective data integration





Ontology (Philosophy)

Ontology is the philosophical study of the nature of being, becoming, existence or reality as well as the basic categories of being and their relations.



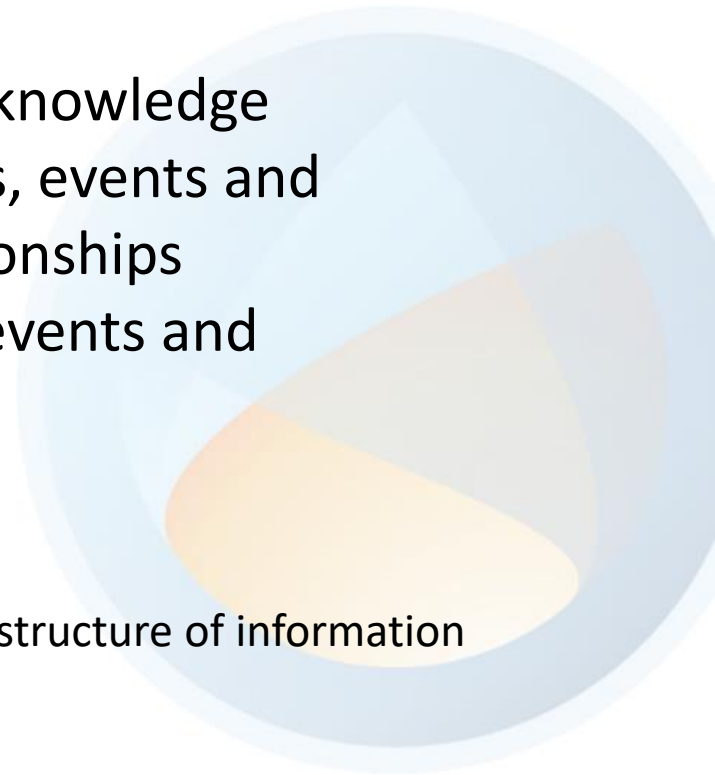
Parmenides 520 BCE
Proposed an ontological characterization of the fundamental nature of reality



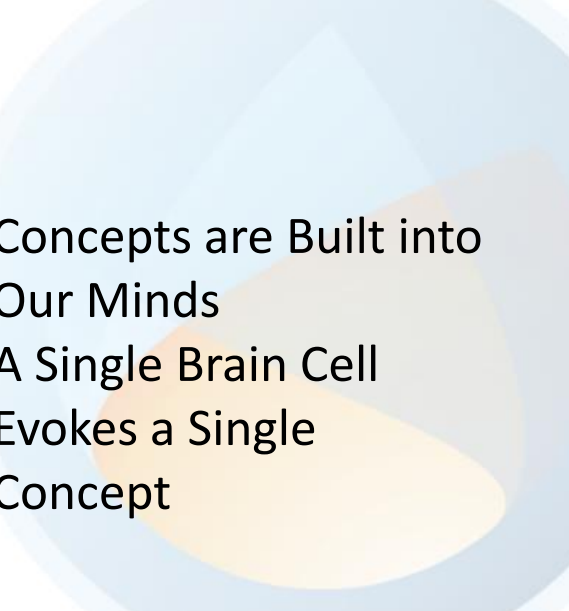
Ontology (IT)

Formal representation of a knowledge domain, describing its entities, events and processes and the relationships connecting these entities, events and processes

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyse domain knowledge



A decorative graphic consisting of a grid of colored squares in shades of blue, orange, and yellow, arranged in a pattern that resembles a stylized letter 'L' or a corner.



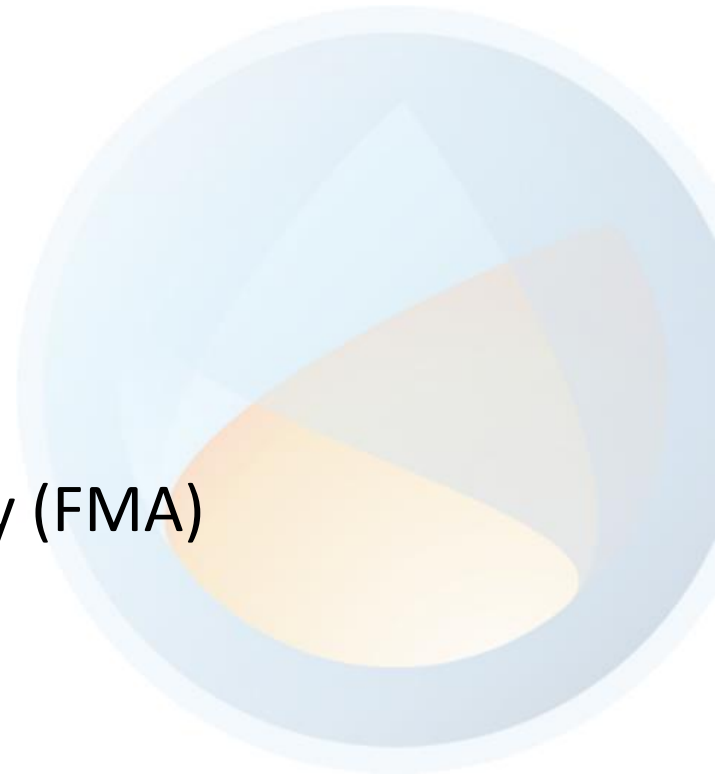
Concepts are Built into
Our Minds
A Single Brain Cell
Evokes a Single
Concept

February 2013



“Ontologies” in Life Sciences

- Snomed CT
- ICD-xx
- MedDRA
- Canonical
- The Foundational Model of Anatomy (FMA)
- Gene Ontology (GO)
- Cell Ontology (CL)
- Protein Ontology (PRO)
- openEHR

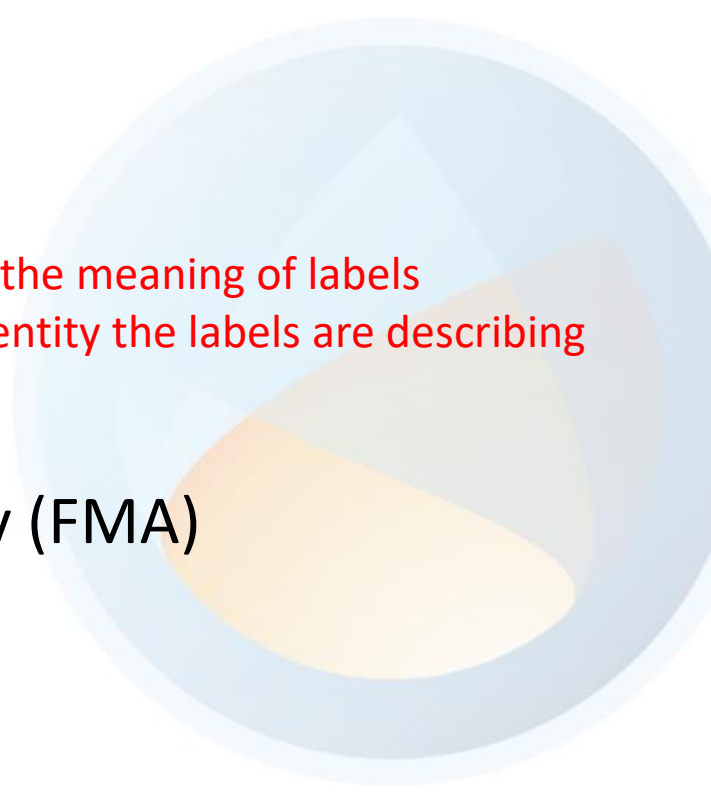




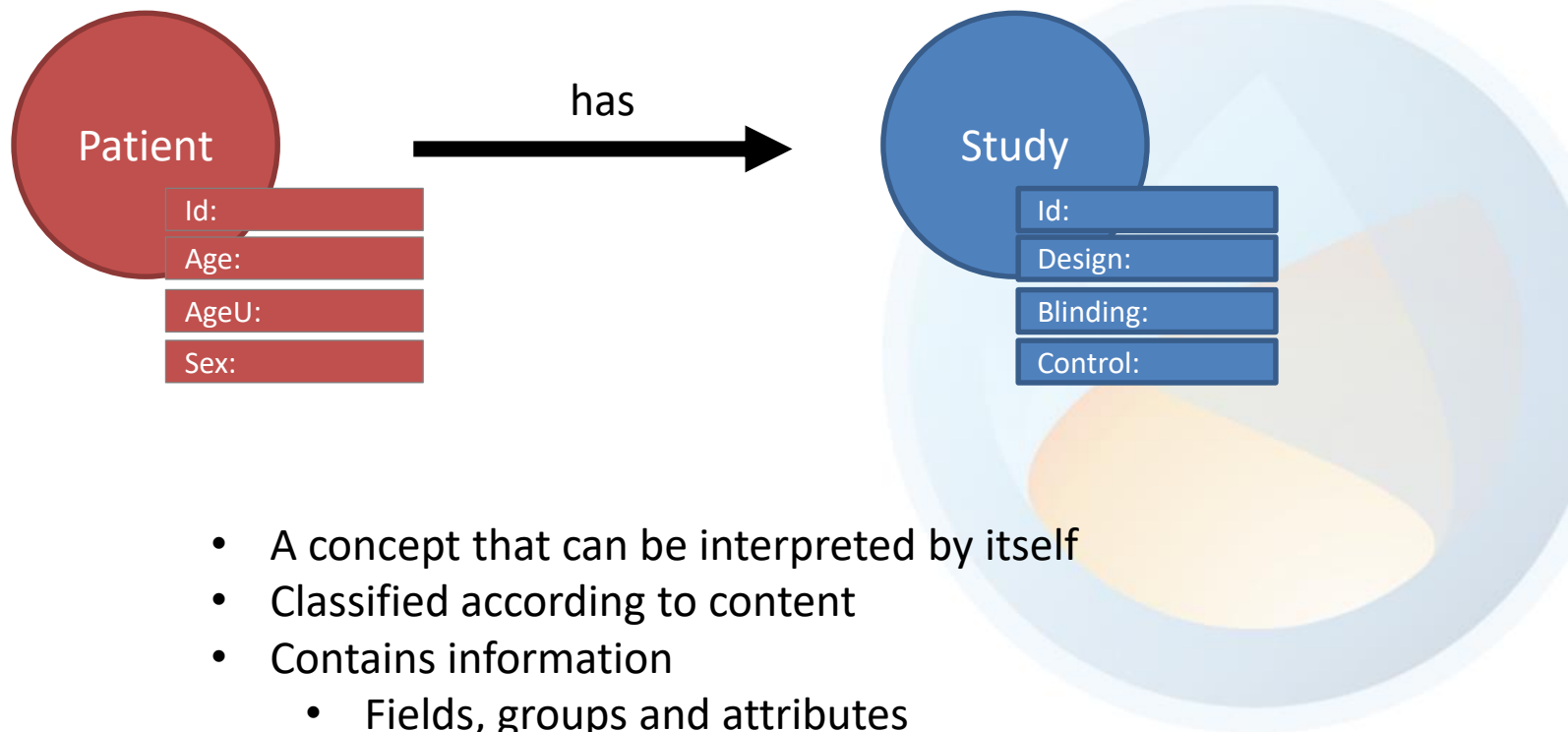
“Ontologies” in Life Sciences

- Snomed CT
- ICD-xx
- MedDRA
- Canonical
- The Foundational Model of Anatomy (FMA)
- Gene Ontology (GO)
- Cell Ontology (CL)
- Protein Ontology (PRO)
- openEHR

Terminologies
Code lists
Concerned with the meaning of labels
rather than the entity the labels are describing



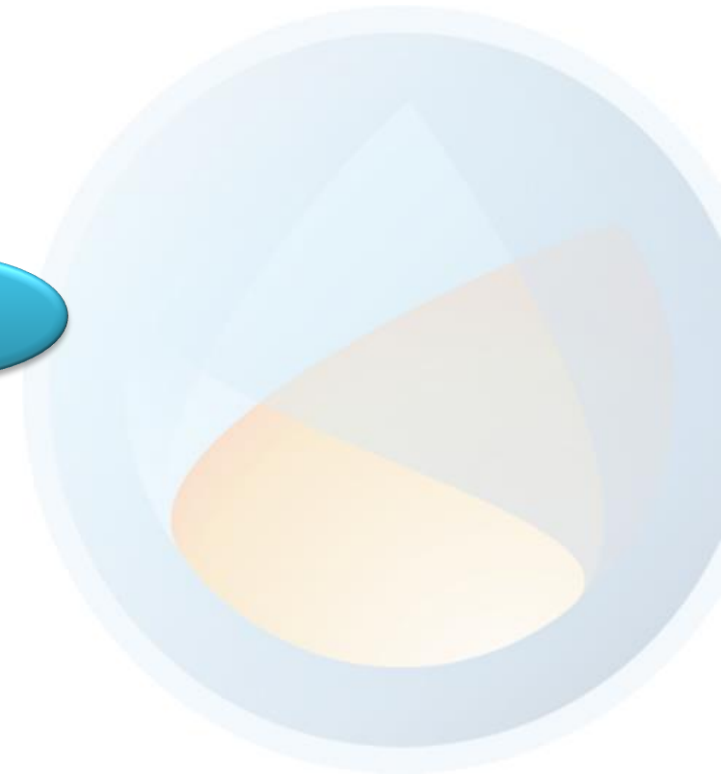
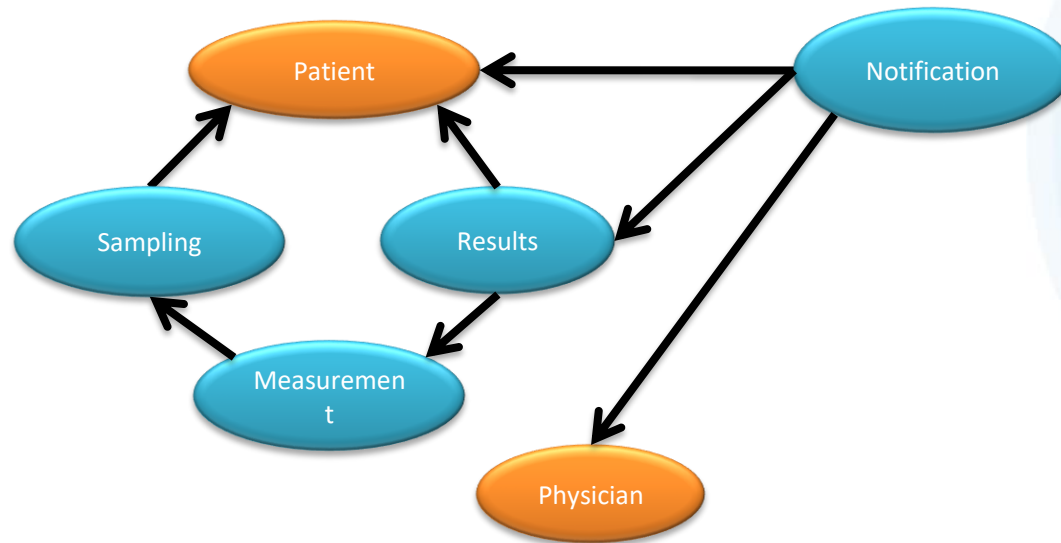
Holons



- A concept that can be interpreted by itself
- Classified according to content
- Contains information
 - Fields, groups and attributes
- Contains relations to other Holons
- Each relation has a specific meaning

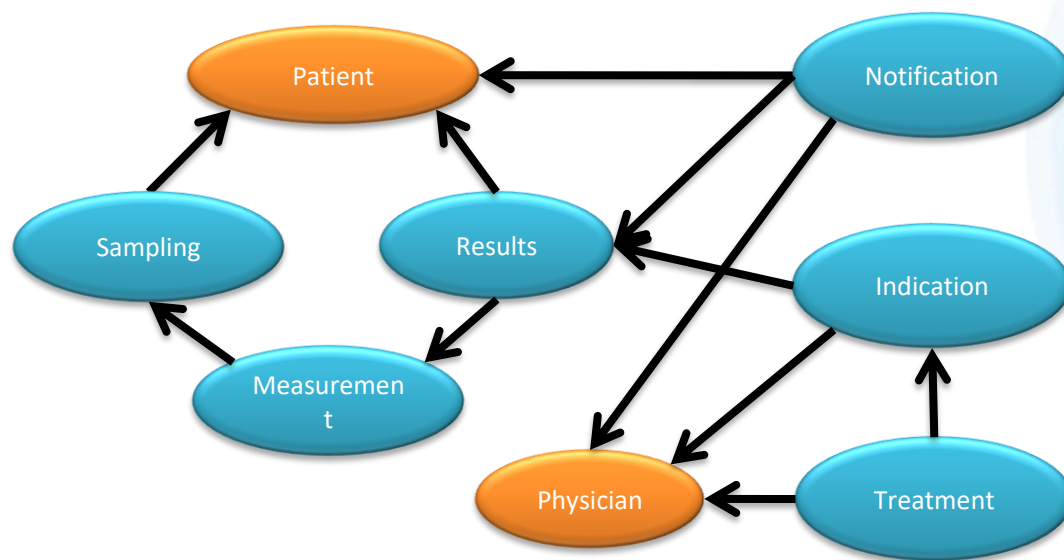


Real World Information Modelling - Using Holons



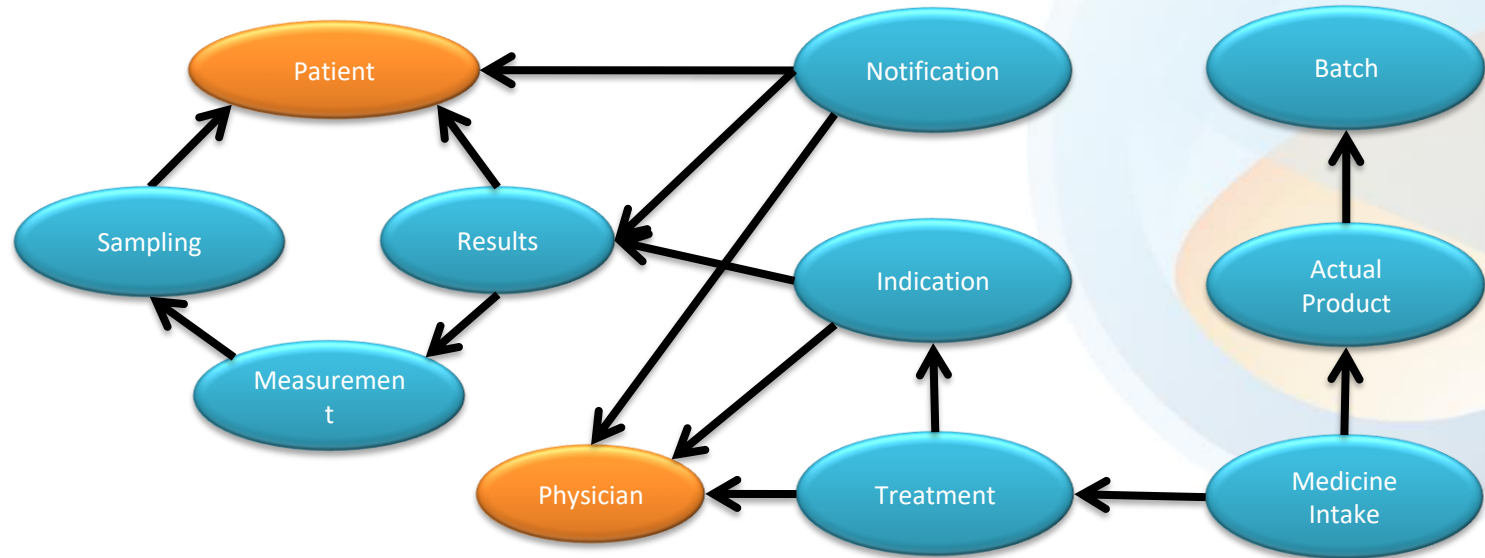


Real World Information Modelling - Using Holons





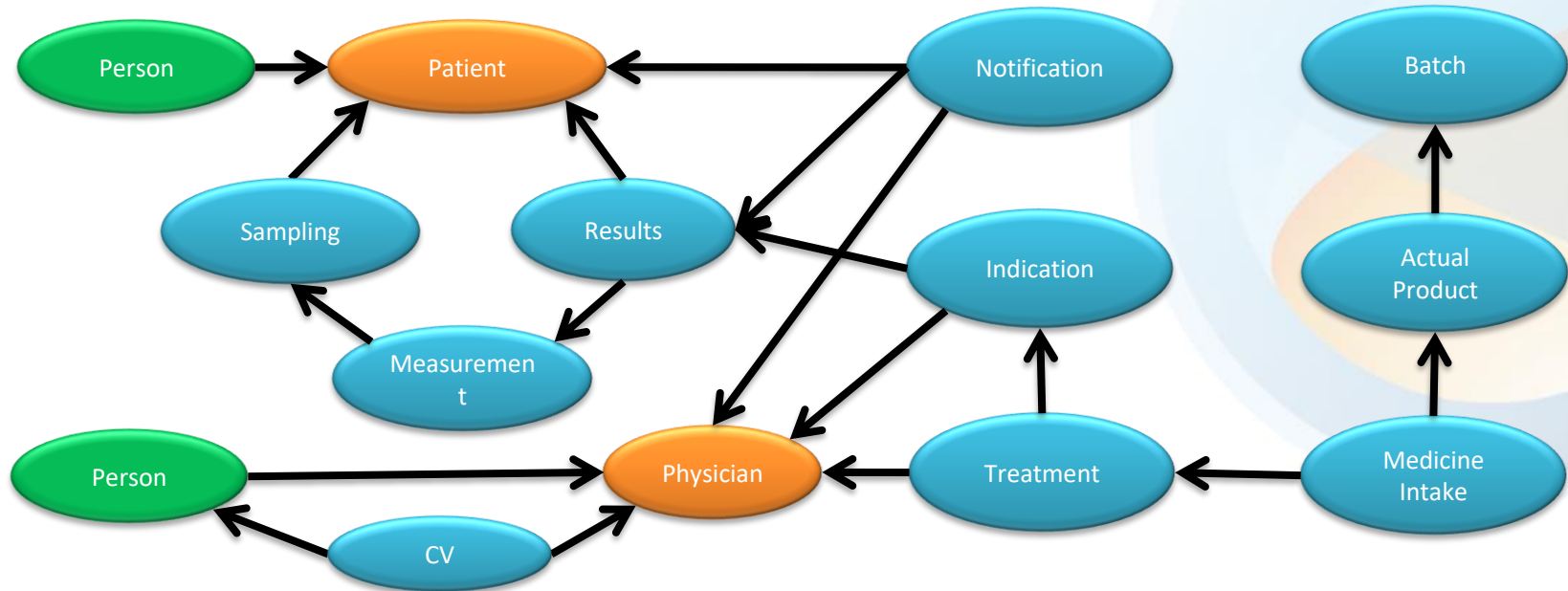
Real World Information Modelling - Using Holons





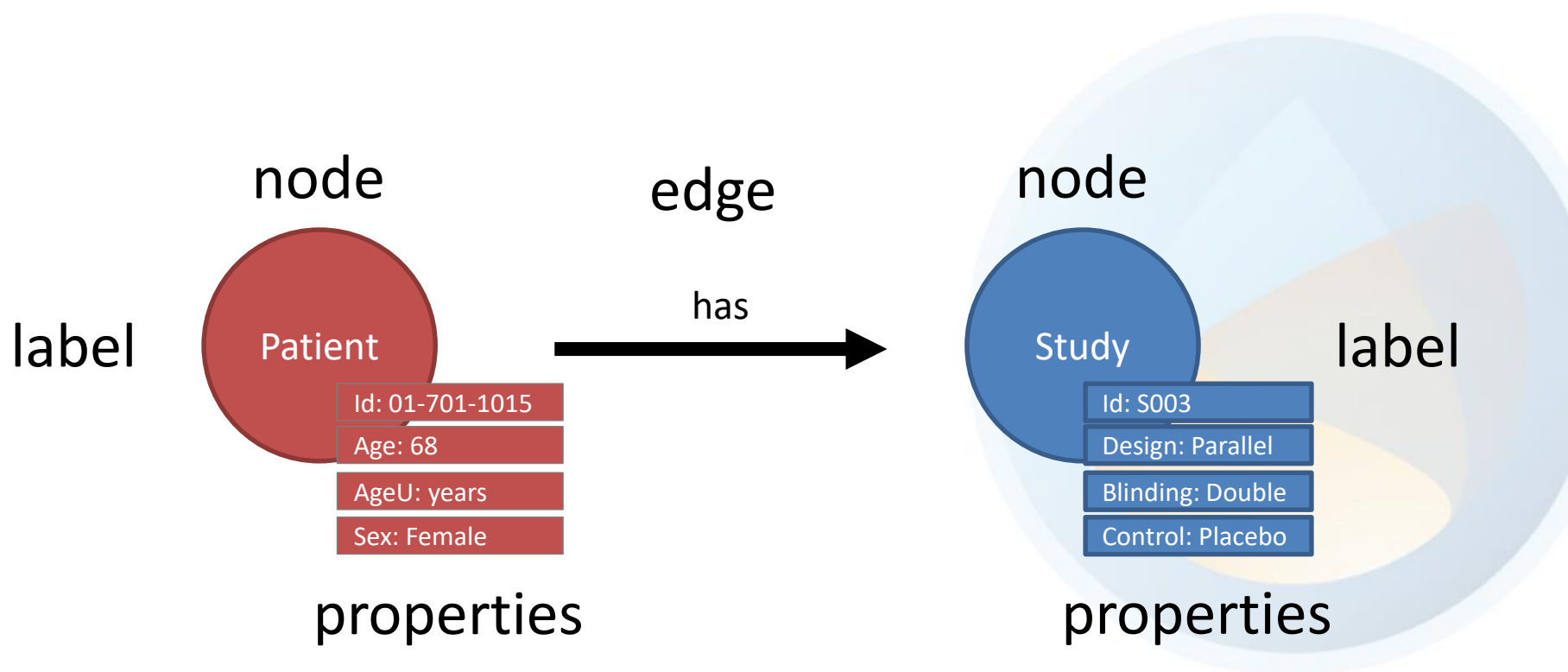
Real World Information Modelling - Using Holons

Building a Conceptual “Mind Map” of Related Holons

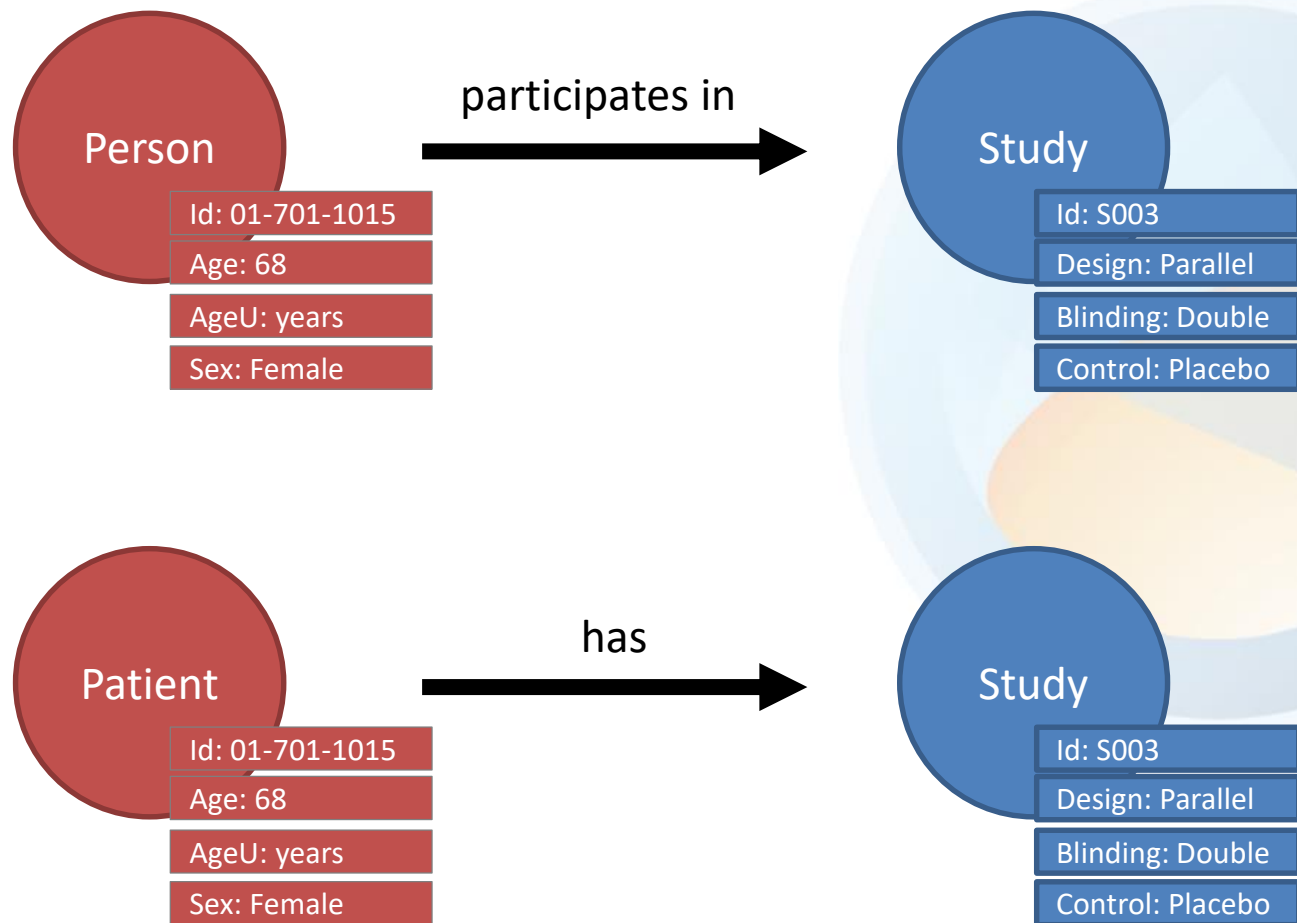




Graph databases



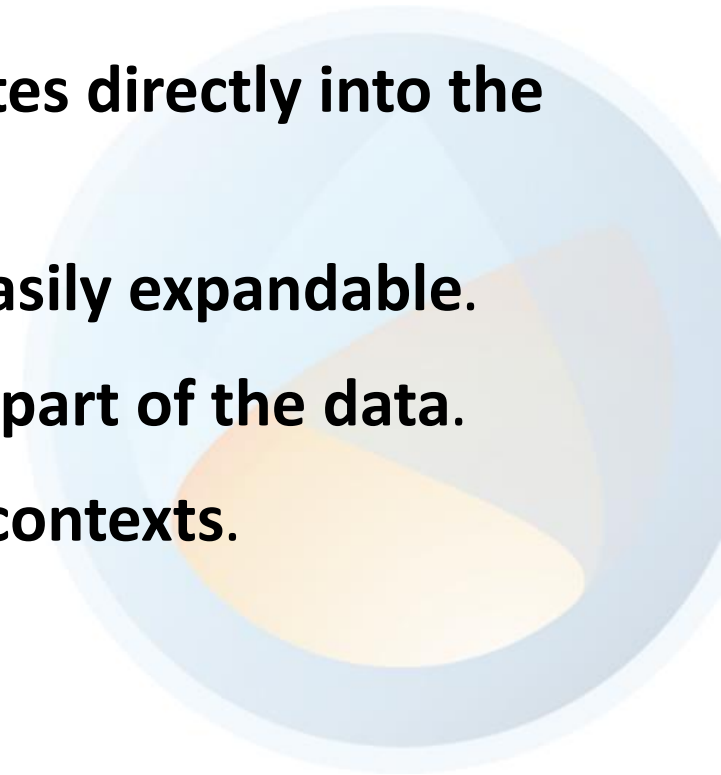
Conceptual modelling



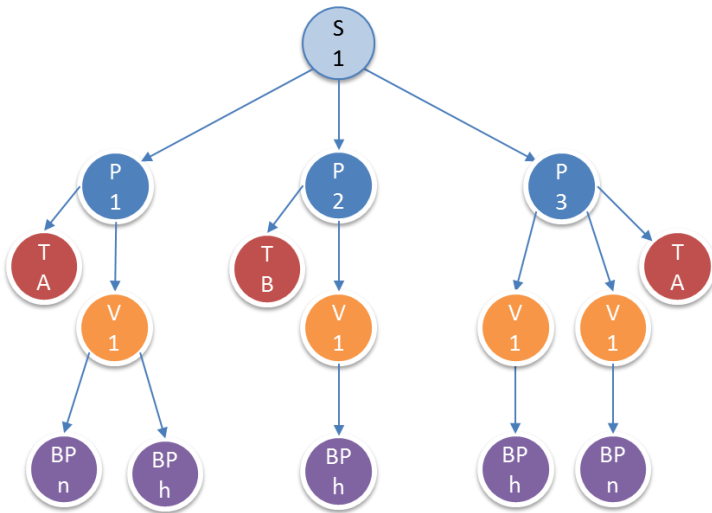


Benefits of graph databases

- **The conceptual data model translates directly into the graph database model.**
- **Graph databases are flexible and easily expandable.**
- **Metadata can be stored directly as part of the data.**
- **Data can be evaluated in different contexts.**



Indices versus graph traversal



- individual nodes (identity index)
- node types (node type identity index)
- property values (property index)
- existence of indirect relationships (relation index)



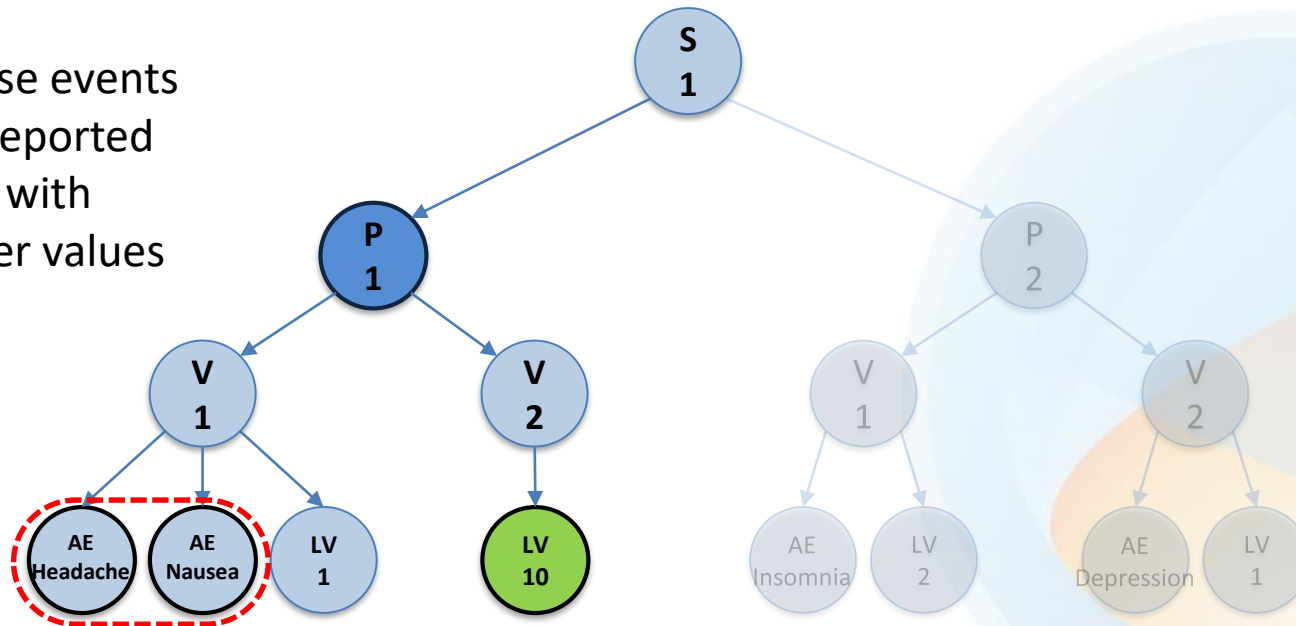
System to answer complex questions

- Find all blood pressure measurements that are considered high and that are connected to patients that were given a dose of the drug A
- What adverse events have been reported for patients with elevated liver values



Querying a graph database

What adverse events
have been reported
for patients with
elevated liver values



Select data:

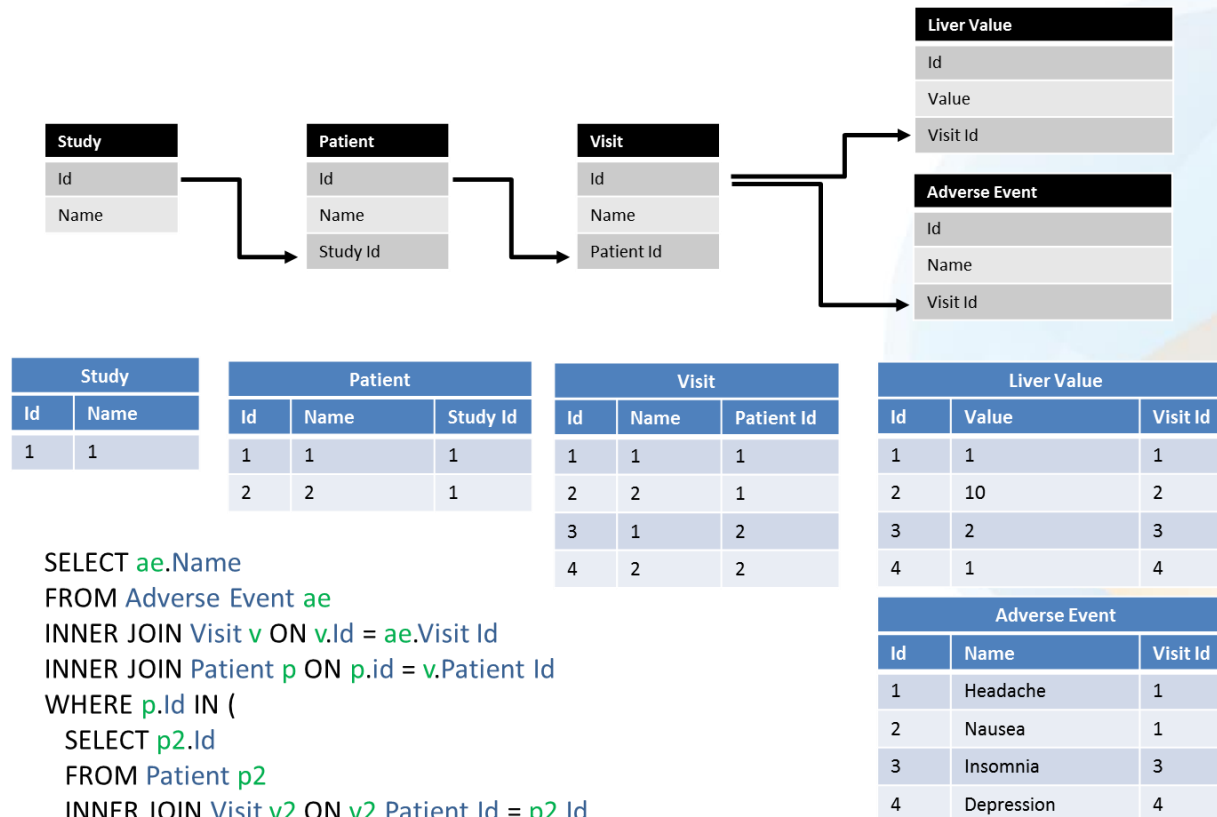
Type of Node: "Patient"

With (Type of Node: "Liver Value", Property: "Value > 5")

Fetch:

Type of Node: "Adverse Event", property: "Name"

Querying a relational database



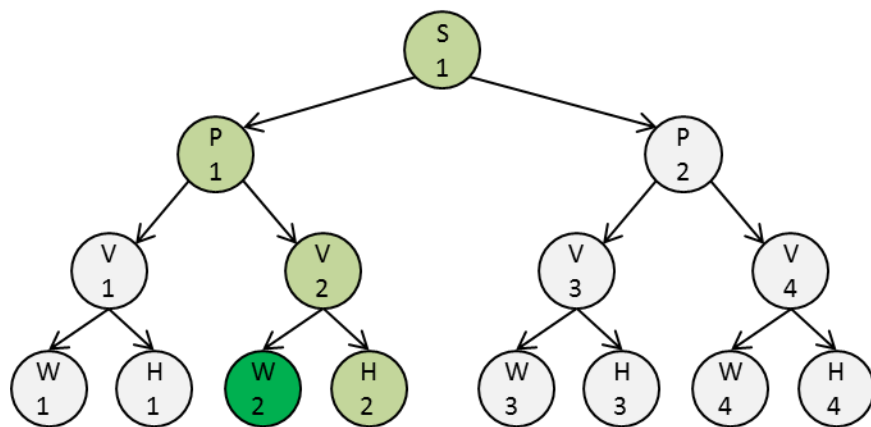
```
SELECT ae.Name
FROM Adverse Event ae
INNER JOIN Visit v ON v.Id = ae.Visit Id
INNER JOIN Patient p ON p.Id = v.Patient Id
WHERE p.Id IN (
  SELECT p2.Id
  FROM Patient p2
  INNER JOIN Visit v2 ON v2.Patient Id = p2.Id
  INNER JOIN Liver Value lv ON lv.Visit Id = v2.Id
  WHERE lv.Value > 5
)
```



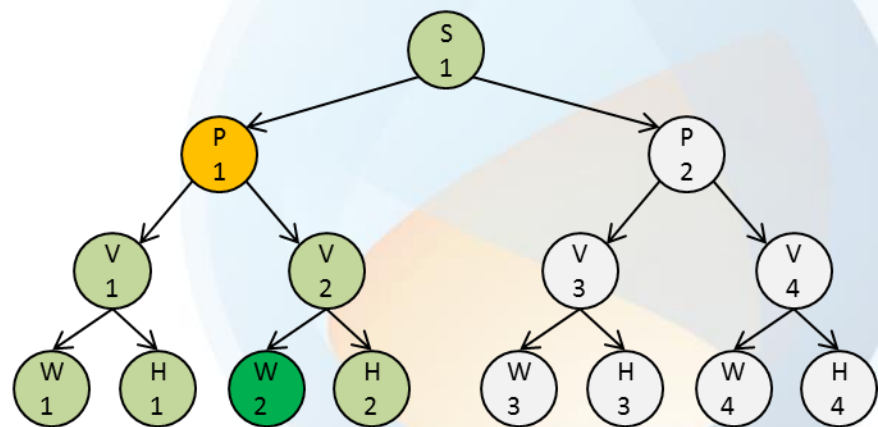
Reflective logic

relate values through a specific type of Holons where no direct relation exist

Without reflection



With reflection in P



- = Holons matching filter
- = Reflection point Holons
- = Holons in result
- = Holons not in result

[illegible]

Patient

Adverse Event

Death

Diagnosis

Employment

Medication

Medical Care Contact.Su...

Study Arm.Randomization

Study Enrollment

Study Patient

Study Termination

Surgical Procedure

Work Influence, Missed Wo...

Tables, List boxes, Panels and ...

Study Basics
1 value

Termination Type
1 value

headache
2 values

Search

Search in: Applied filter

headache

Search results

Diagnosis

Diagnosis Term

Headache

Horton's Headache Or Neuralgia

Diagnosis Term Original

Intermittent headaches

Hortons headache

Adverse Event

Medication

Summary Tables

Study Patient Patient Status AE > Patient AE Info Graph/Ontology

Adverse Event Summary (Reflected in Patient)

Treatm...	Pats	Pats w AE	AE Tot	Pats w SAE	SAE Tot	Death Tot	Pats w AERel	AERel Tot	Pats w SAERel	Pats w SAEOTH	Pats Disc@AE	Pats Disc@AERel
Ezlocil	139	77	163	15	22	0	12	15	4	14	0	0
Kodicillin	139	74	192	17	20	0	5	6	2	14	0	0
Meticillin	150	92	225	18	21	1	8	9	3	15	0	0
Placebo	116	55	120	14	17	0	0	0	0	14	0	0

Current Counts

Patients	544
AE	700
SAE	80

AE System Organ Class >Pat

MedDRA SOC

Infections and infestations	(95) 52
Musculoskeletal and conn...	(76) 43
Nervous system disorders	(67) 39
Gastrointestinal disorders	(56) 34
General disorders and ad...	(56) 32
Respiratory, thoracic and ...	(34) 19

AE High Level Term >Pat

MedDRA HLG

Headaches NEC	(42) 26
Musculoskeletal and conn...	(35) 17
Upper respiratory tract inf...	(30) 14
Oedema NEC	(12) 11
Urinary tract infections	(18) 10
Muscle related signs and ...	(17) 10

AE Preferred Term >Pat

MedDRA PT

Headache	(42) 26
Back pain	(20) 10
Muscle spasms	(17) 10
Pneumonia	(14) 9
Oedema peripheral	(8) 8
Chest pain	(16) 7

AE Low Level Term >Pat

MedDRA LLT

Headache	(42) 26
Pneumonia	(14) 9
Chest pain	(16) 7
Influenza	(12) 6
Coughing	(8) 6

Drug Name >Pat

Medication

Principen	(543) 334
Amoxil	(527) 325
Kodicillin	(146) 92
Ezlocil	(139) 85
Meticillin	(150) 81
Placebo	(116) 79
Paracetamol	(63) 36
Acetylsalicylic acid	(55) 33
Furosemide	(37) 31
Digoxin	(25) 21
Isosorbide mononitrate	(32) 17
Captopril	(21) 17
Influenza virus vaccine po...	(34) 16
Ipratropium bromide	(25) 16
Verapamil hydrochloride	(22) 16
Ibuprofen	(21) 16

AE Seriousness >Pat

Seriousness

Not Serious	(264) 156
Serious	(64) 44

AE Severity >Pat

Severity

Mild	(175) 111
Moderate	(163) 86
Severe	(66) 43

AE Outcome >Pat

Preferred Term

Not recovered	(72) 40
Recovered	(18) 13
Recovered with sequelae	(5) 3
Fatal	(1) 1

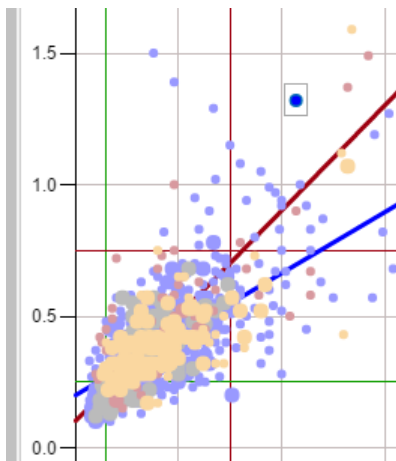
AE Caused by Treatment >...

Causality

Not related	(287) 171
Related	(25) 15
No	(0) 0
Yes	(0) 0

Direct Relation to Holon Source Data

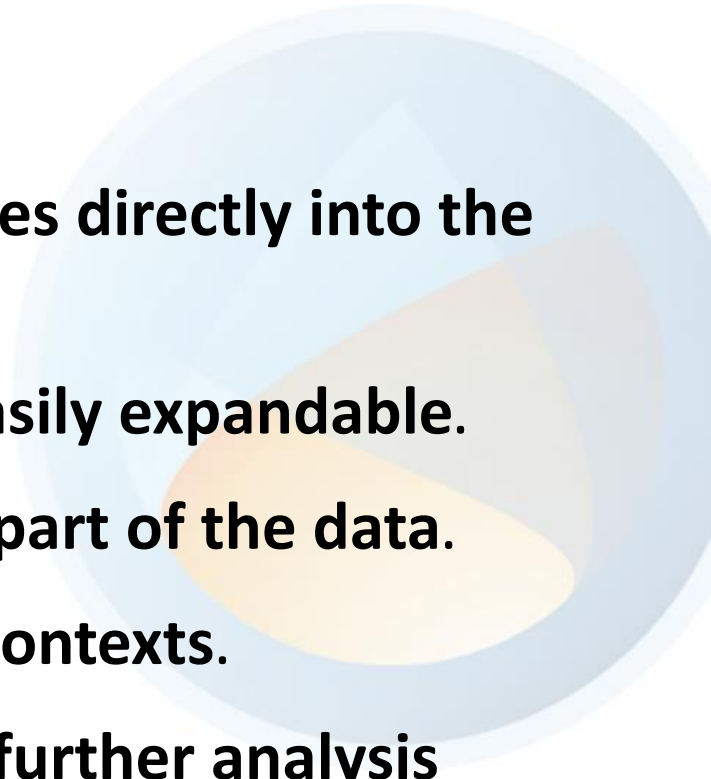
Dealing with Data Veracity



Holon	A23313@SR-003@Capish Alanine Aminotransferase Activity, Serum/Plasma 1.07 2000-03-16
Holon Type	Blood Analysis.Alanine Aminotransferase Activity, Serum/Plasma.Results
Concept	Blood Analysis
Domain	616.07561--dc23
Study Patient Id	A23313@SR-003@Capish
Sample Collection Date	2000-03-16
Alanine Aminotransferase Activity, Serum/Plasma	1.07 ukat/L
Alanine Aminotransferase Activity, Serum/Plasma.Original	64 U/L
Laboratory Id.Code	233
Measurement Type.Code	02101
Normality Statement	
Normality	Abnormal
Normality Specification	Abnormal high
Reference Range Lower Limit	0 ukat/L
Reference Range Source	AM21, AutoCheck
Reference Range Upper Limit	0.83 ukat/L



Benefits of Ontology/graph database approach

- **The conceptual data model translates directly into the graph database model.**
 - **Graph databases are flexible and easily expandable.**
 - **Metadata can be stored directly as part of the data.**
 - **Data can be evaluated in different contexts.**
 - **Easy to query and retrieve data for further analysis**
- 

The curated data lake

Data entry systems



Master Data Sources



Load



Data Lakes

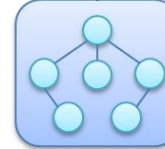


Curation



CXML

CXML



Conceptual
Ontology

Reflect

CQL

Platform

CQL

Other Apps

Users





Thank you

www.capish.com

Peter Tormay
peter.tormay@capish.com

