Enterprise Knowledge Graphs

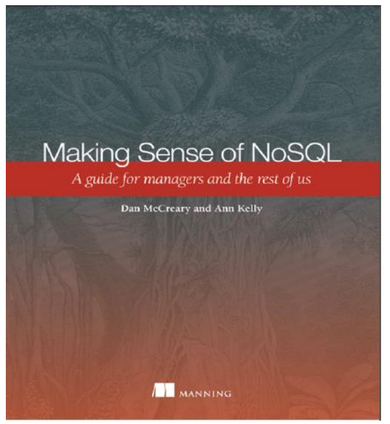
November 7th, 2019

Dan McCreary
Distinguished Engineer

OPTUM® Advanced Technology Collaborative

Hello, my name is Dan

- Distinguished Engineer in AI and Graph Technologies
- Co-founder of "NoSQL Now!" conference
- Author of "Making Sense of NoSQL" (w. Ann Kelly)
- 17+ years of working with NoSQL
- Worked with Brian Kernighan (Bell Labs) and Steve Jobs (NeXT)
- Background in solution architecture, metadata management, NLP, semantics, text analytics and **knowledge representation for AI**
- Active in STEM and robotics volunteering
- Co-founder of AI Racing League



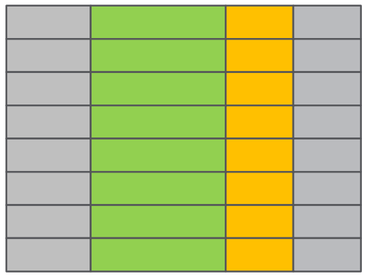
Disclaimer: Opinions are my own and do not represent the views of my employer

dan.mccreary@optum.com

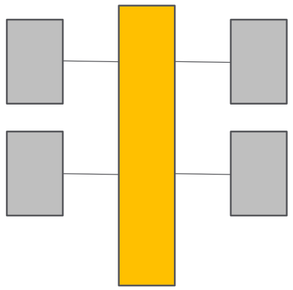
OPTUM®

The World is “Not Only SQL” (a.k.a. NoSQL)

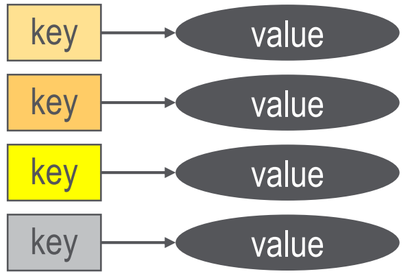
Relational



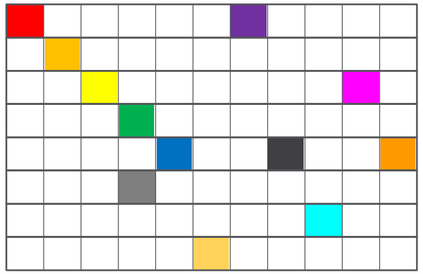
Analytical (OLAP)



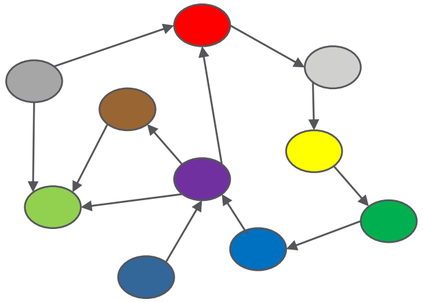
Key-Value



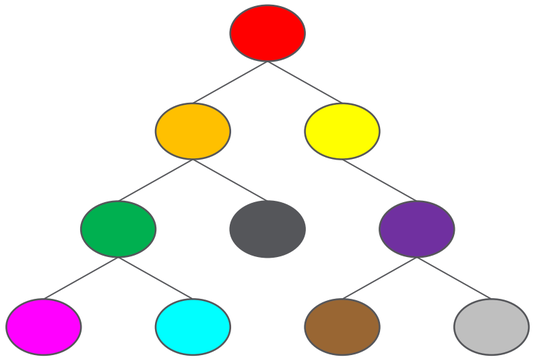
Column-Family



Graph



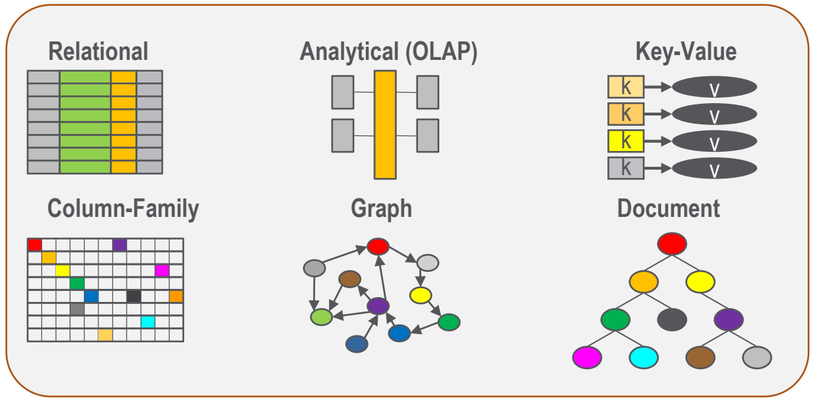
Document



See Chapter 1: https://www.manning.com/books/making-sense-of-nosql

Role of the Solution Architect

Sally Solutions
Title: Solution Architect

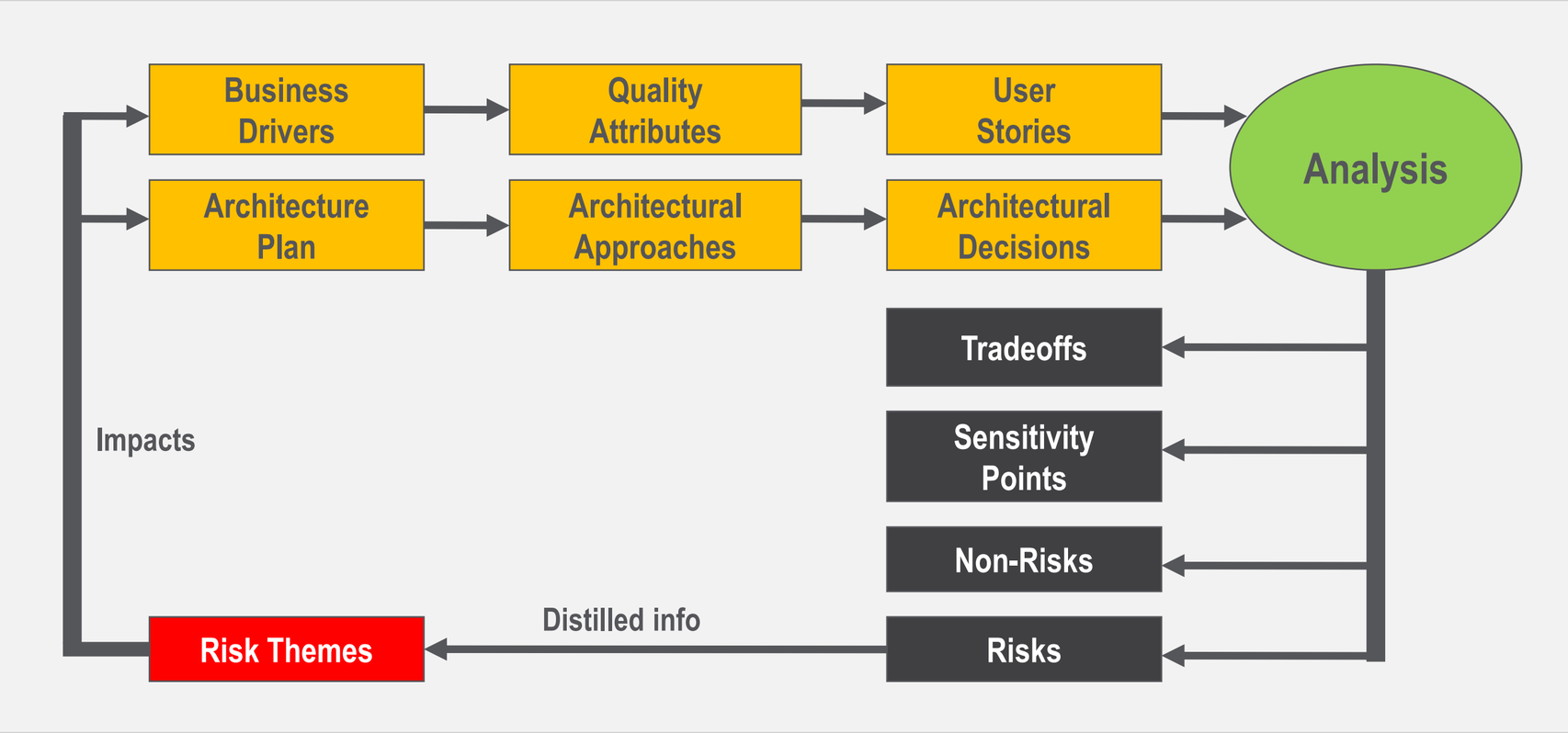


Non-bias matching of business problem to the right data **architecture**

before we begin looking at a specific **products.**

OPTUM®

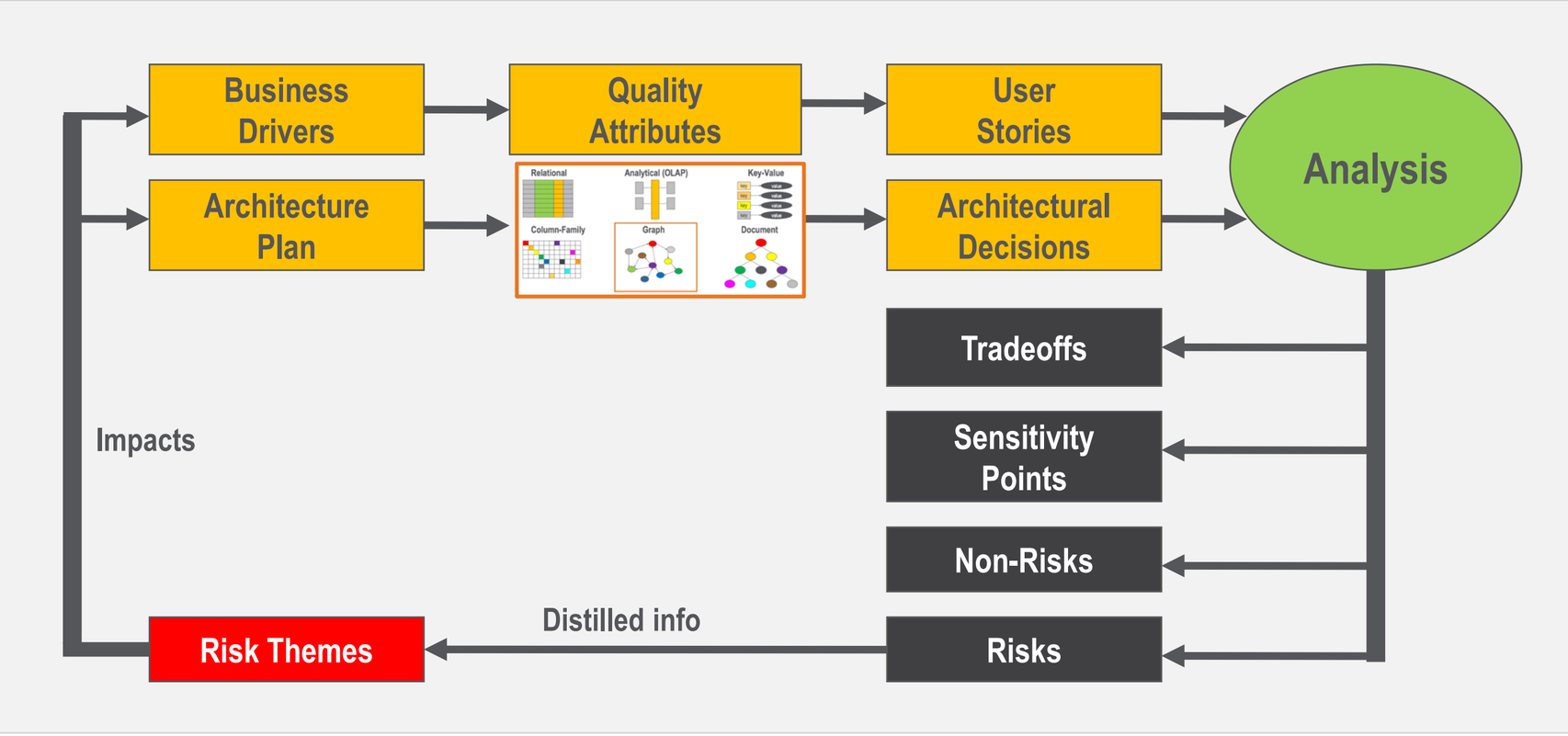
Architectural Tradeoff and Analysis Method (ATAM)



The ATAM process defined by CMU's Software Engineering Institute (SEI)

OPTUM®

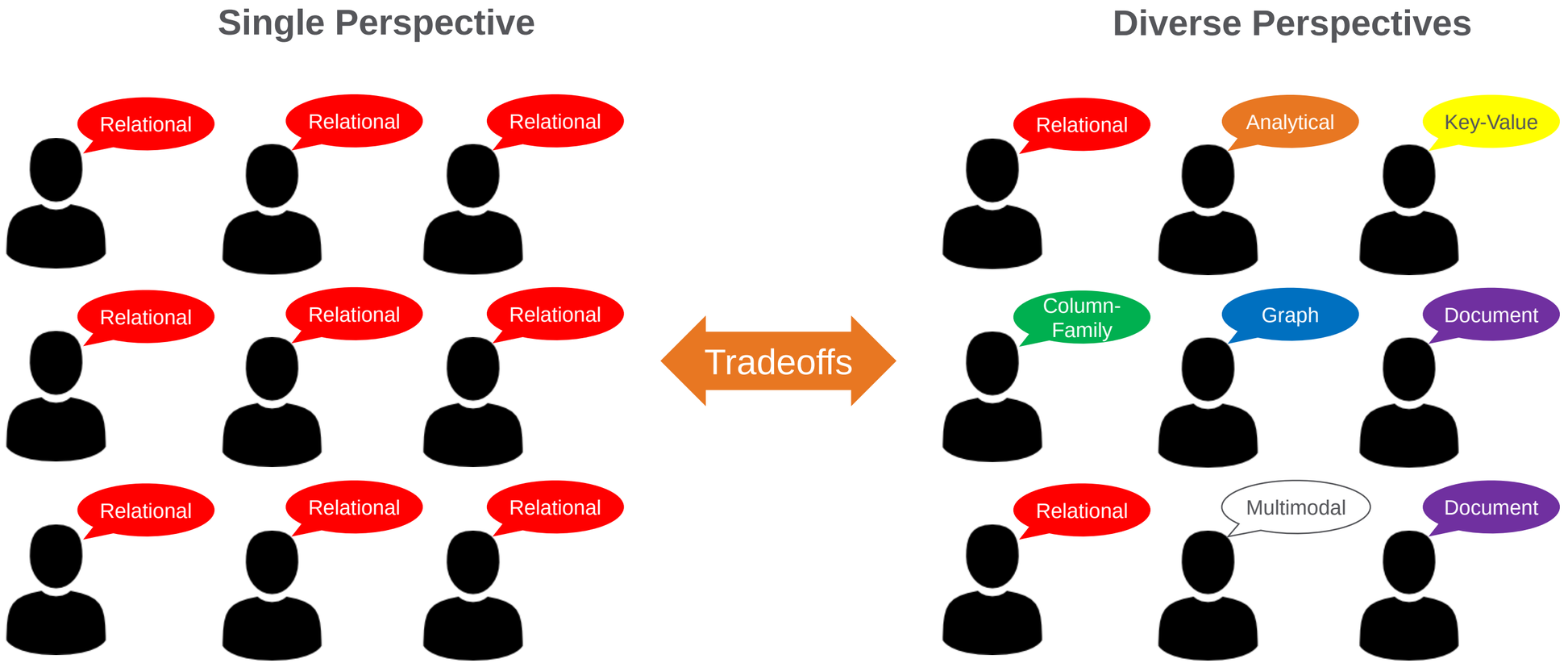
Graph Databases as an Architectural Approach



We want all appropriate architecture options to be considered by Optum/UHC Solution Architects

OPTUM®

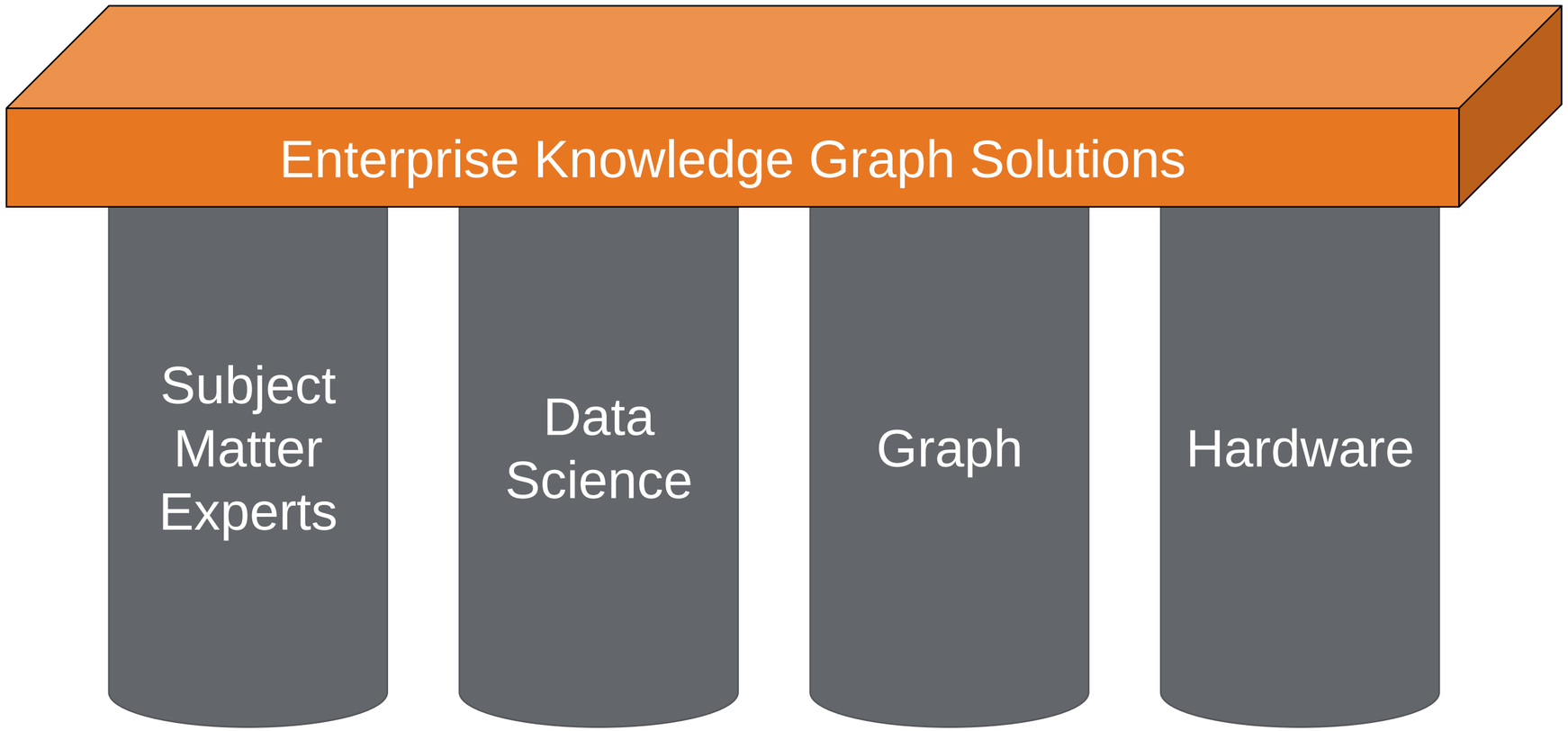
Encouraging Architecturally Diverse Perspectives



- Which perspective will maximize the probability of a low-cost, scalable competitive solution in the marketplace?
- Do we value diversity of opinions as a core value of our organization? Are we biased? What about standards?
- There is no single right answer to this question; only tradeoffs.

OPTUM®

Encourage Systems Thinking



The four pillars of support for enterprise knowledge graphs!
We need specialists **and** systems thinking!

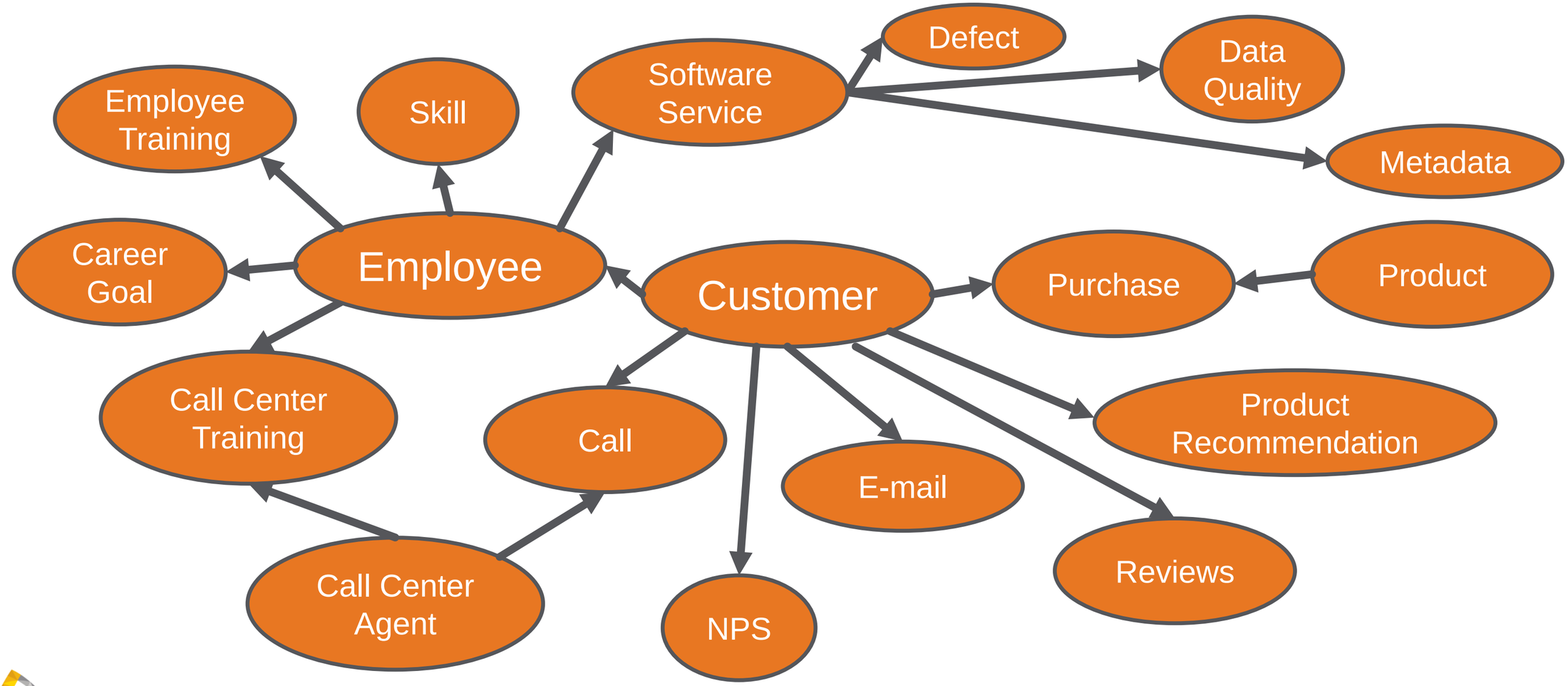
OPTUM®

Agenda

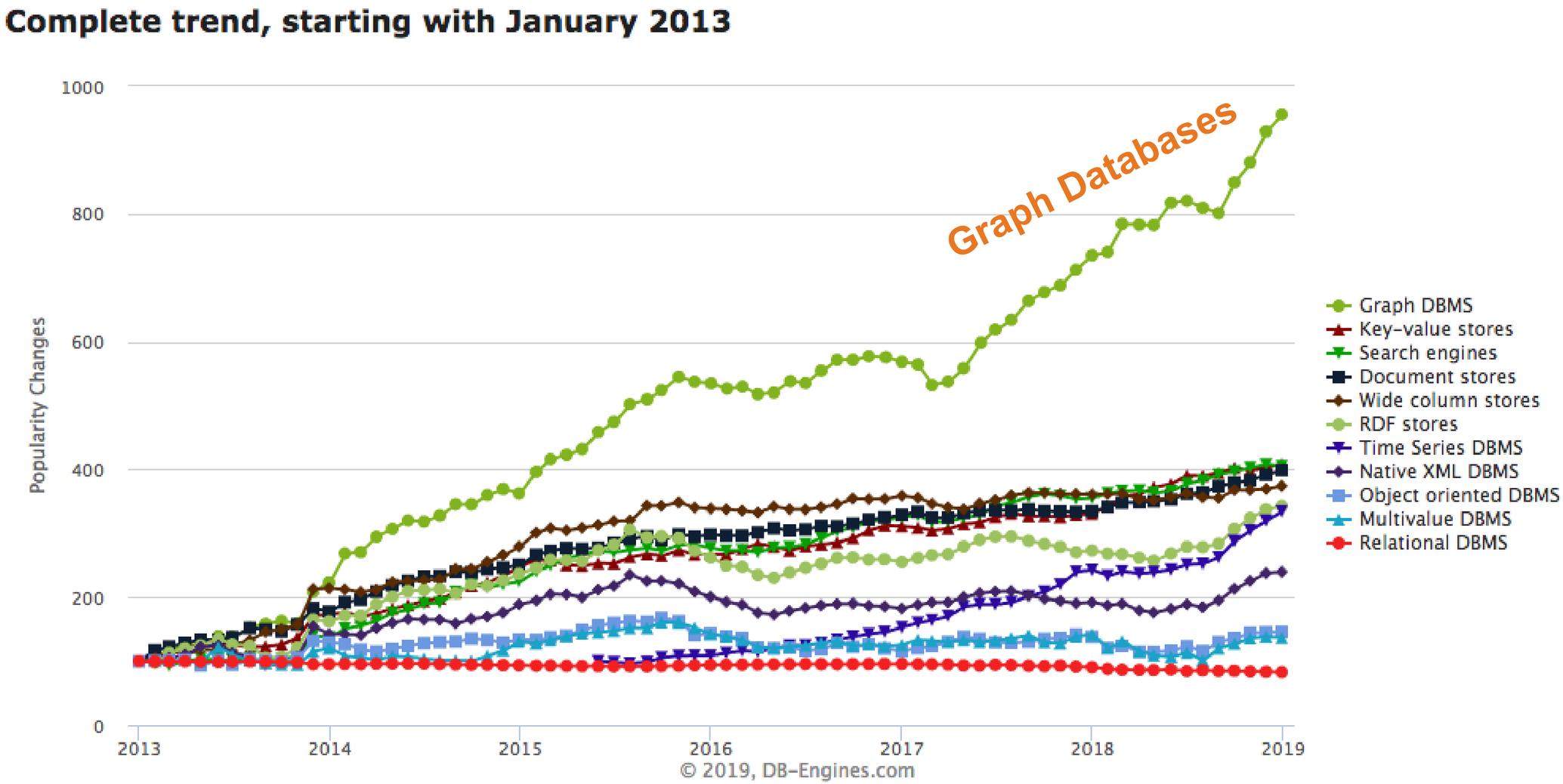
1. What is an Enterprise Knowledge Graph (EKG)?
2. Why are they so popular now?
3. What is a distributed native property graph?
4. What are HTAP systems?
5. Why are graphs closely connected to AI and machine learning?
6. What is graph embedding?
7. What is GQL?
8. How do I explain graphs to my management?
9. Why are current CISC processors inappropriate for cost-effective graph algorithms?
10. The future of Enterprise Knowledge Graphs – Emergence!

Enterprise Knowledge Graph

Definition: A single graph database that connects **all** your organization's information.



Graph Databases are HOT!



https://db-engines.com/en/ranking_categories

Executive Summary

Gartner

*Graph analysis is possibly the **single most effective competitive differentiator** for organizations pursuing data-driven operations and decisions ...”*

https://www.gartner.com/doc/2852717/it-market-clock-database-management

OPTUM®

Gartner on Graph Analytics

Key Analytics Trends for 2019

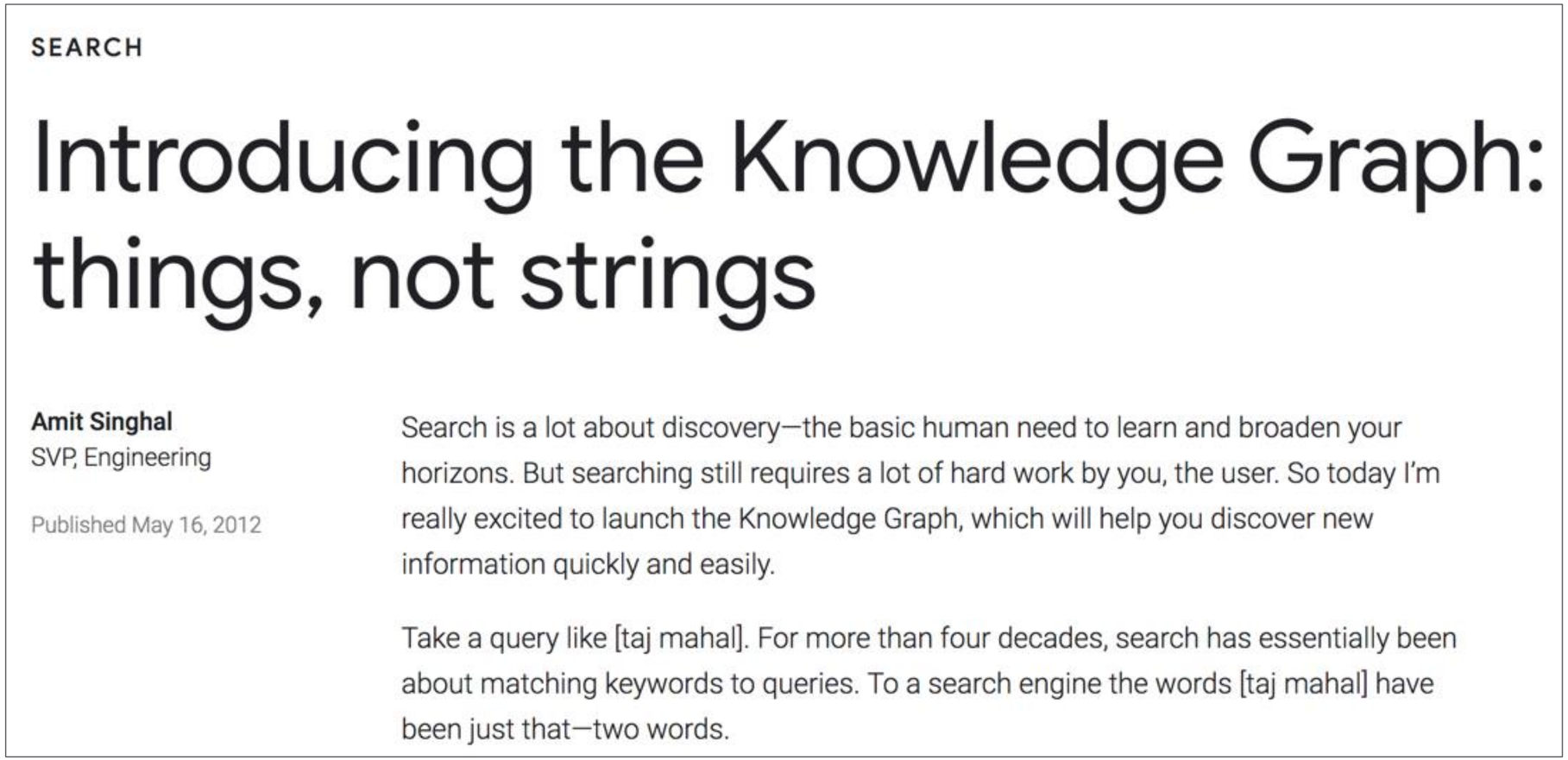
1. Augmented Analytics
2. Augmented Data Management
3. Continuous Intelligence
4. Explainable AI
5. **Graph**
6. Data Fabric
7. NLP/Conversational Analytics
8. Commercial AI/ML
9. Blockchain
10. Persistent Memory Servers

…Graph processing to continuously accelerate data preparation and **enable more complex and adaptive data science**…to efficiently model, explore and query data with complex interrelationships across data silos…the need to ask complex questions across data silos which is not practical or even possible at scale using SQL queries.

Graphs are also related to 4, 6 and 7

OPTUM®

Why the ”Cambrian Explosion” in Graph Innovation since 2014?



SEARCH

Introducing the Knowledge Graph: things, not strings

Amit Singhal
SVP, Engineering

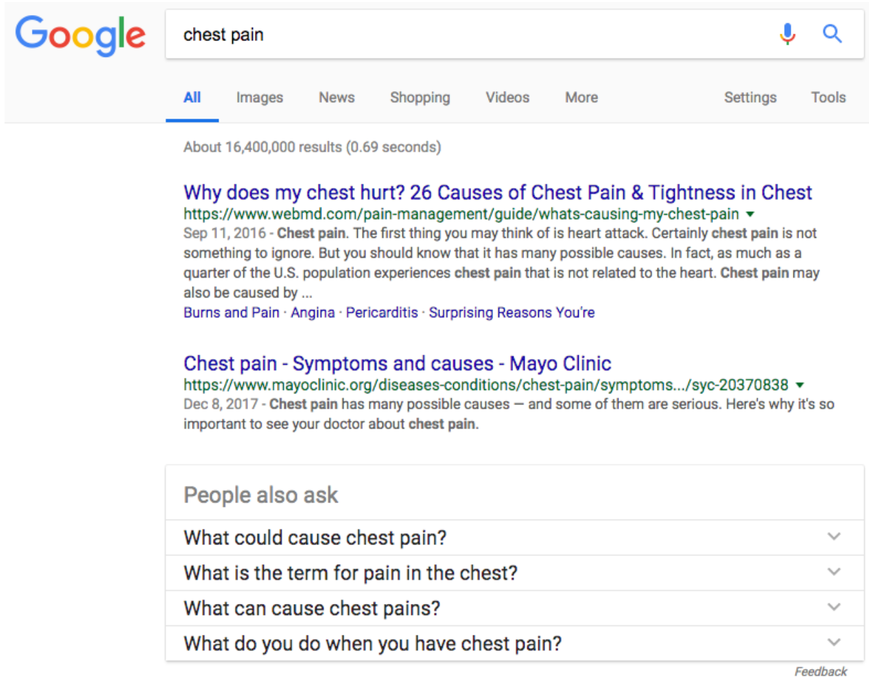
Published May 16, 2012

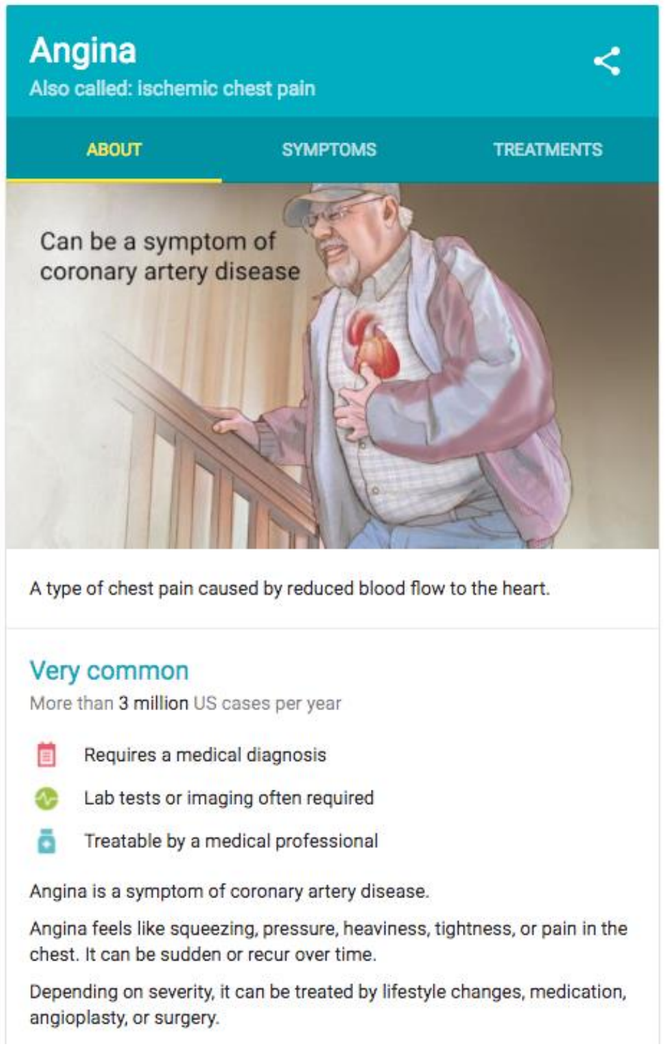
Search is a lot about discovery—the basic human need to learn and broaden your horizons. But searching still requires a lot of hard work by you, the user. So today I'm really excited to launch the Knowledge Graph, which will help you discover new information quickly and easily.

Take a query like [taj mahal]. For more than four decades, search has essentially been about matching keywords to queries. To a search engine the words [taj mahal] have been just that—two words.

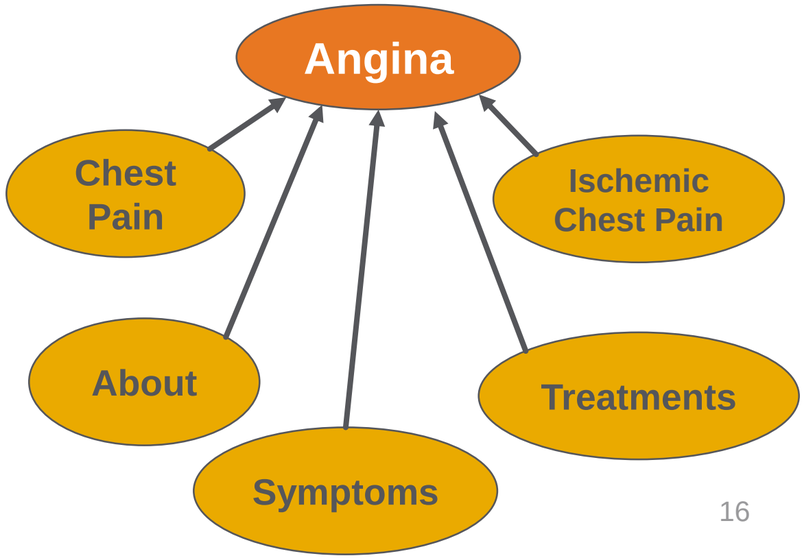
OPTUM®

Google search for “chest pain”



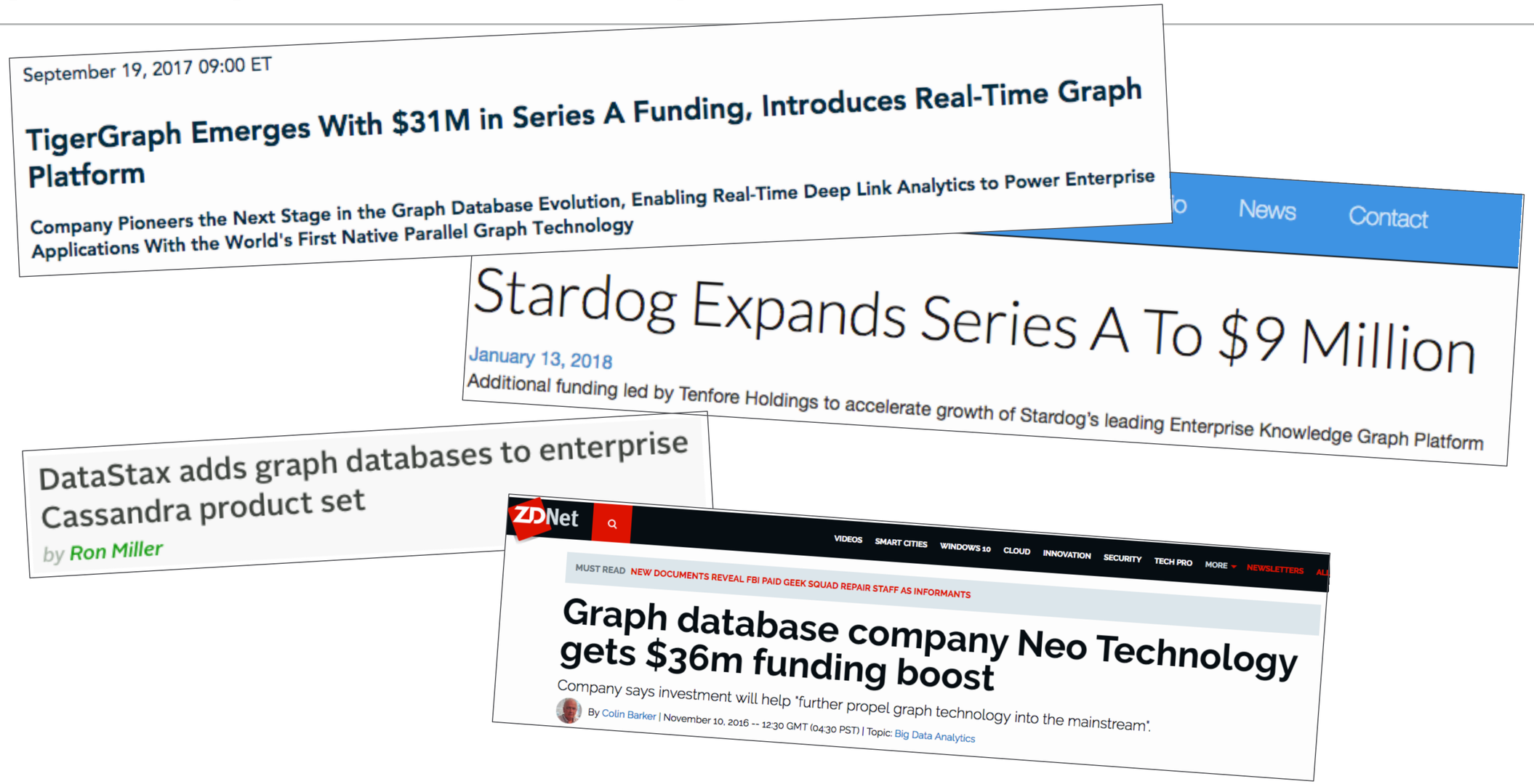


Knowledge Graph Summary



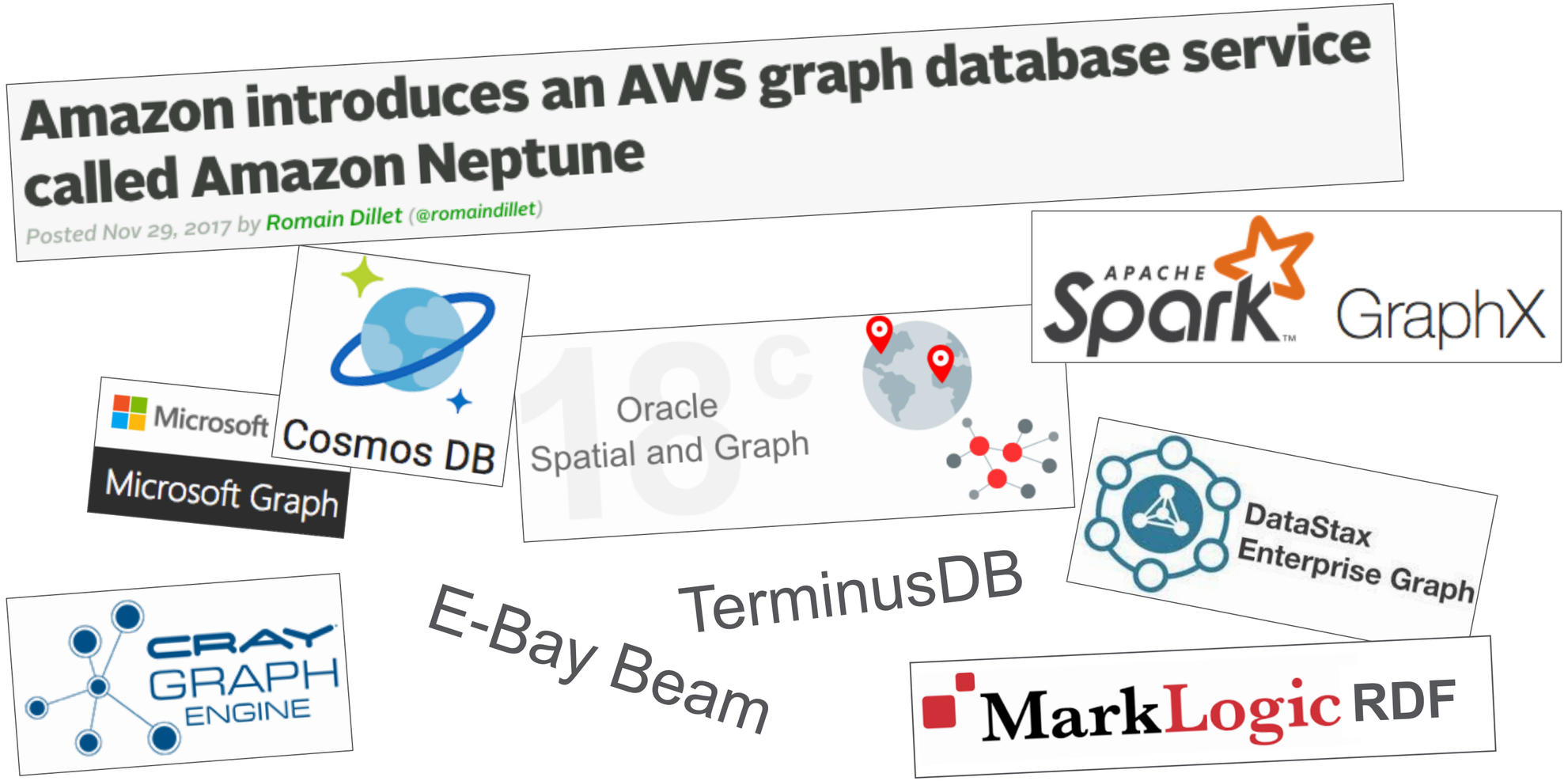
OPTUM®

Graph Startups and Venture Capital



See slide notes for references

New Entrants into the Marketplace



- Big players adding graph capabilities to existing systems
- High Performance Computing (HPC) Graph Custom Hardware

Graph Services and Standards to Support AI

- New data services that share data in the form of graphs
- Adoption of graph standards for sharing information (RDF, OWL, SKOS)
- Upgrading existing standards to support graph interoperability (SNOMED-CT URI)
- Use AI to discover rules – use graphs to store rules
- New in-browser graph visualization libraries

Which of the following organizations use graph databases?

Answer: They all do!

- Every major airline uses a graph database to calculate fares in real-time
- Over half of retailers use graphs for product recommendations

OPTUM®

External Forces Driving Graph Databases

- **Connect to Compete**
  - Data used to live mostly in disparate application “silos” (claims, provider, RX)
  - Now organizations are building 360 views of data: – customer – HR - hardware
  - A Data Lake has added low-cost storage but may not focus on connecting data

Unstructured and Semi-Structured Data

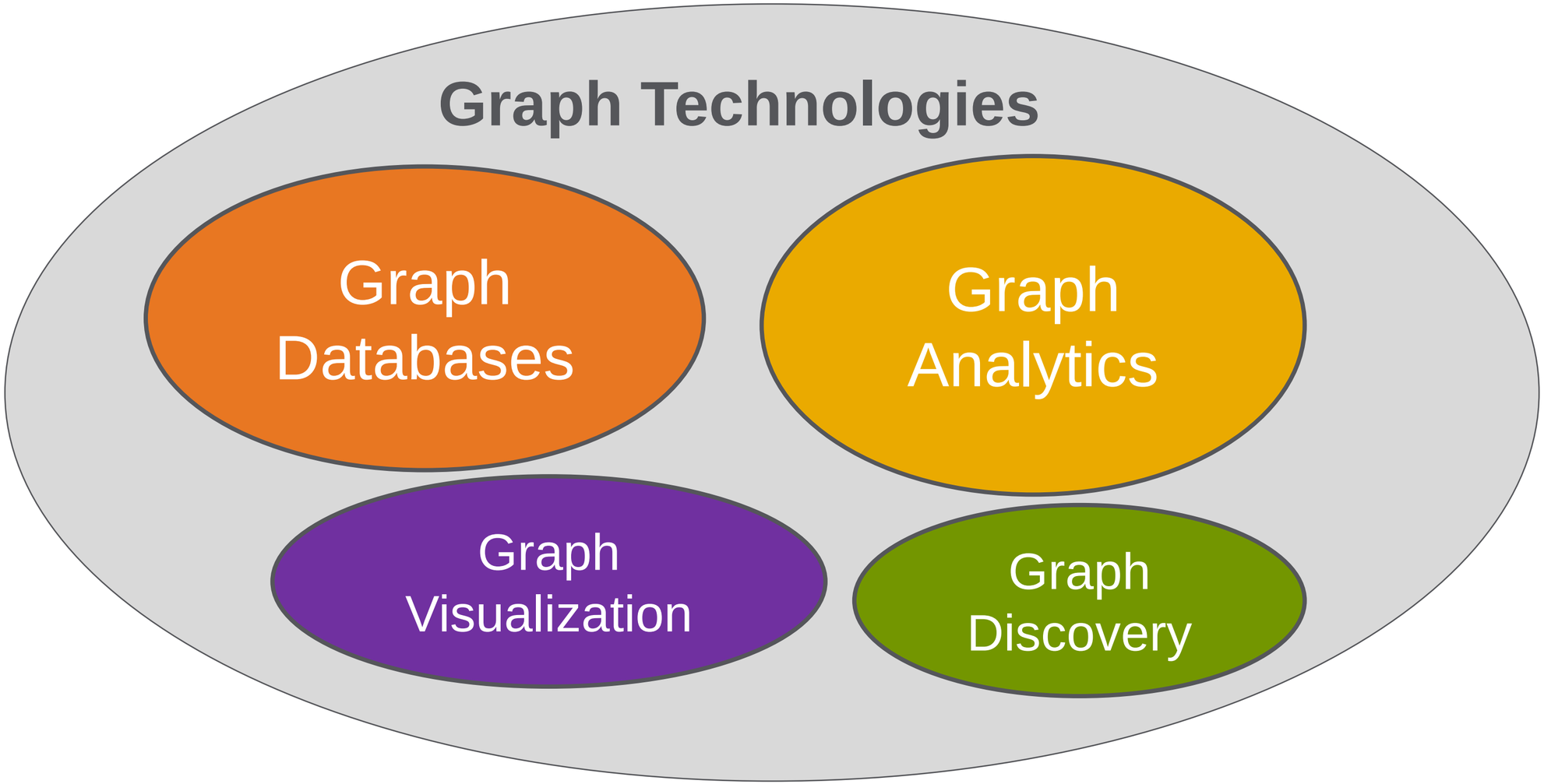
- Movement from keyword-only search to semantic search
- Use of NLP on text, wikis, taxonomies, folksonomies, ontologies

Data Complexity

- AI and inference
- Event log analyses, social networks
- Rules engines, recommendation systems, next best action

OPTUM®

Graph Technologies



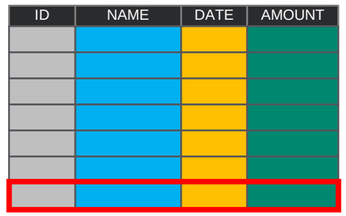
Definition: Graph Databases

*A database that uses verticies (node) and edges (relationships) as atomic units of storage and where **fast graph traversal** is the primary method of accessing information.*

In graph databases, relationships are a **primary** consideration.

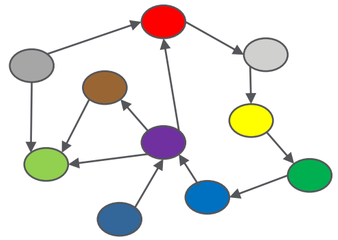
Relational vs. Graph

Relational (row store)



1. Atomic unit of storage is a **row** of a table and data is appended to a table one row at a time
2. All **columns** within a table must have the same structure and no variations within a table are allowed
3. Table structures are **fixed** after design – all rows have the same structure
4. Relationship traversal is done via JOINs at runtime using log(N) **search's** calculated at **query** time
5. Difficult to distribute over a cluster of 100+ nodes

Graph



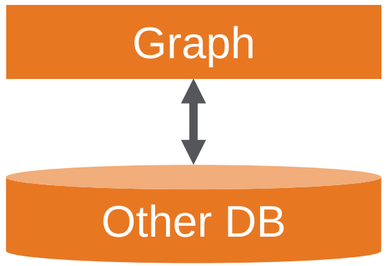
1. Atomic unit of storage are **nodes** and **edges**
2. Each node and edge may have independent properties that are determined at run time (schema agnostic)
3. Joins between nodes and edges are computed at **load** time and are stored as memory pointers
4. Relationships traversal is done using pointer hopping – each core can **evaluate 2M hops per second**
5. Distributed graph products are new

OPTUM®

Two Types of Graph Databases

Native

- Software custom-built for fast graph traversal
- Usually uses a concept called **index-free adjacency**
- Typically runs at 2M hops/second/core

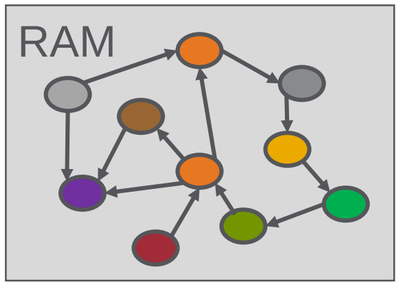


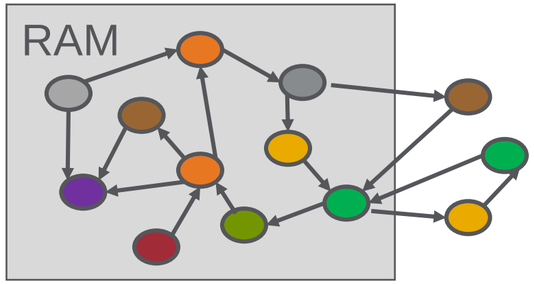
Non-Native

- Graph software built on top of other databases (relational, key-value, column-family or document store)
- Convenient when data is already stored in another database types
- Benefits from the same security policies that govern other databases
- Typically run at 10K hops/second or slower

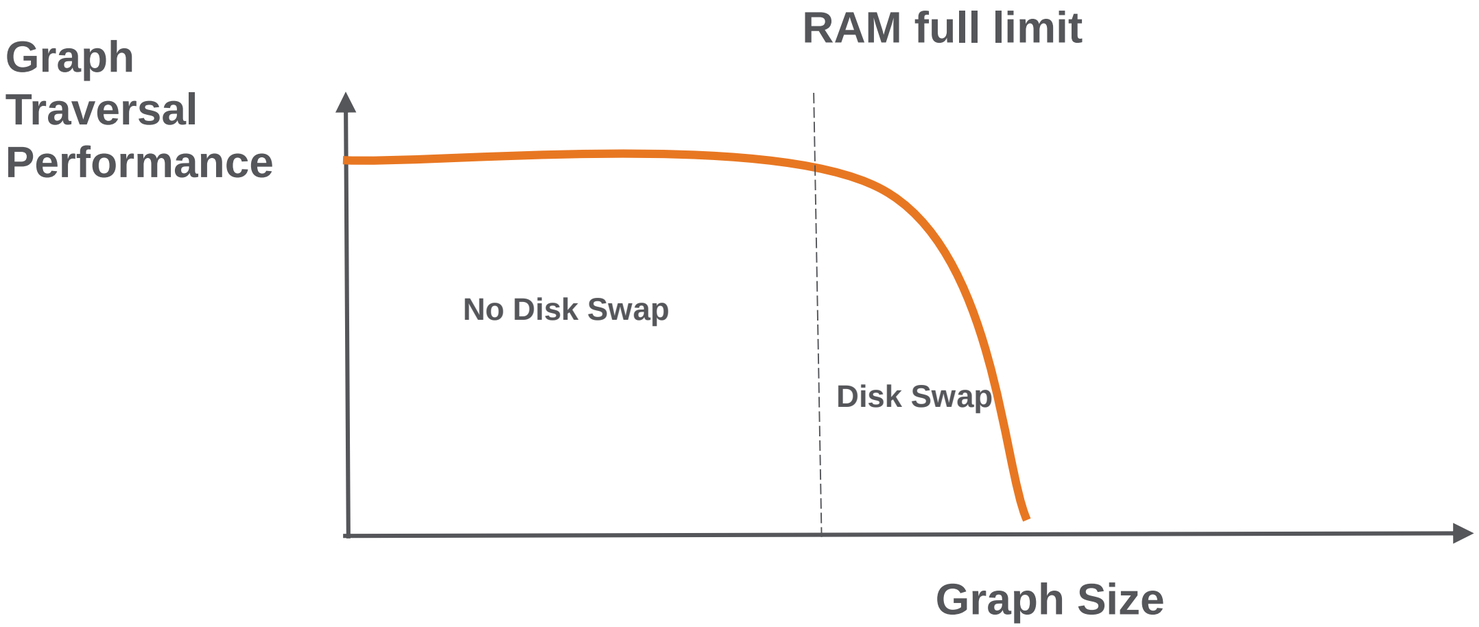
OPTUM®

The “Achilles Heel” of Native Graph Databases





If all nodes and arcs don't fit into RAM memory performance will quickly degrade. A challenge for shared-nothing architectures.



Availability of 12TB Node with 448 CPUs on Amazon EC2

AWS News Blog

Now Available – Amazon EC2 High Memory Instances with 6, 9, and 12 TB of Memory, Perfect for SAP HANA

by Jeff Barr | on 27 SEP 2018 | in Amazon EC2, Launch, SAP | Permalink | Share

0:00 / 0:00

Voiced by Amazon Polly

The Altair 8800 computer that I built in 1977 had just 4 kilobytes of memory. Today I was able to use an EC2 instance with 12 terabytes (12 tebibytes to be exact) of memory, almost 4 billion times as much!

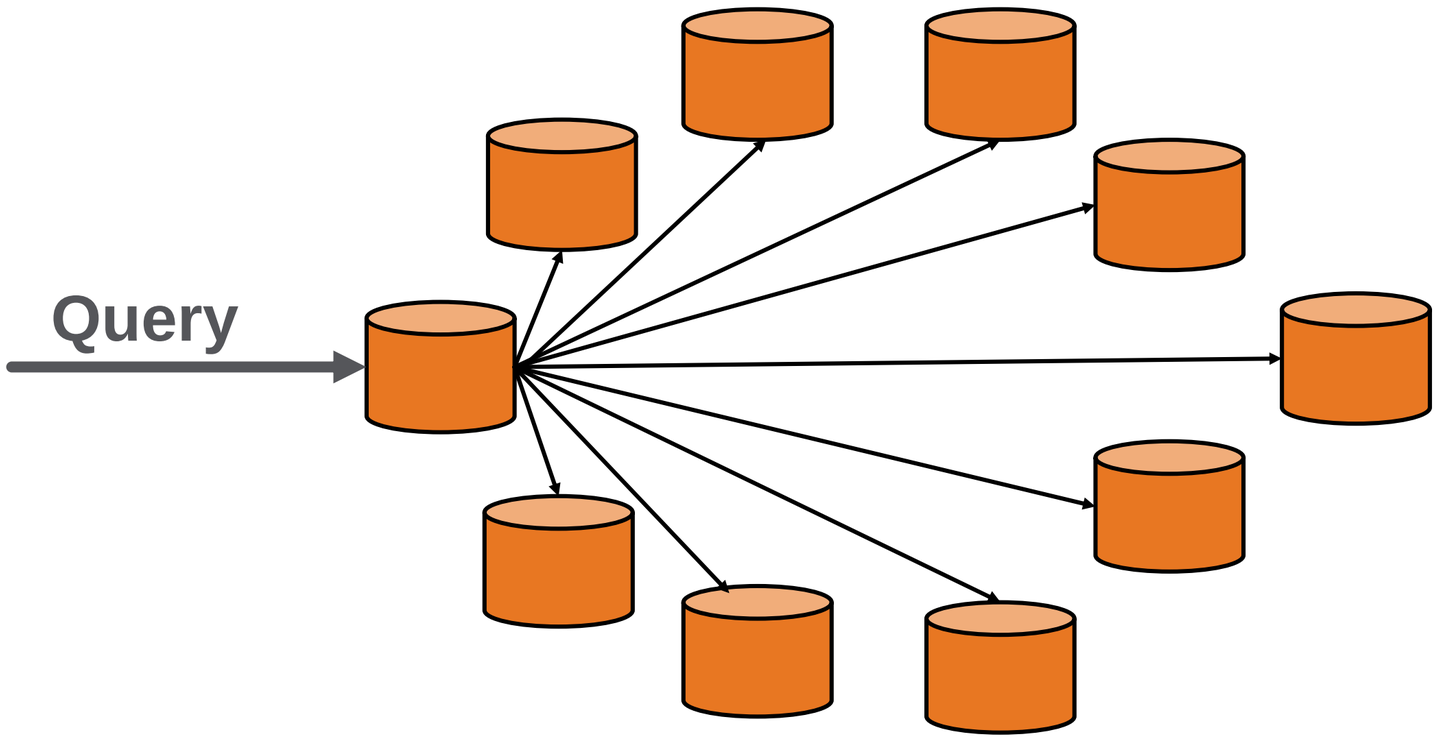
The new Amazon EC2 High Memory Instances let you take advantage of other AWS services including Amazon Elastic Block Store (EBS), Amazon Simple Storage Service (S3), AWS Identity and Access Management (IAM), Amazon CloudWatch, and AWS Config. They are designed to allow AWS customers to run large-scale SAP HANA installations, and can be used to build production systems that provide enterprise-grade data protection and business continuity.

Here are the specs:

Instance Name	Memory	Logical Processors	Dedicated EBS Bandwidth	Network Bandwidth
u-6tb1.metal	6 TiB	448	14 Gbps	25 Gbps
u-9tb1.metal	9 TiB	448	14 Gbps	25 Gbps
u-12tb1.metal	12 TiB	448	14 Gbps	25 Gbps

https://aws.amazon.com/blogs/aws/now-available-amazon-ec2-high-memory-instances-with-6-9-and-12-tb-of-memory-perfect-for-sap-hana/

EKGs must be distributed graphs

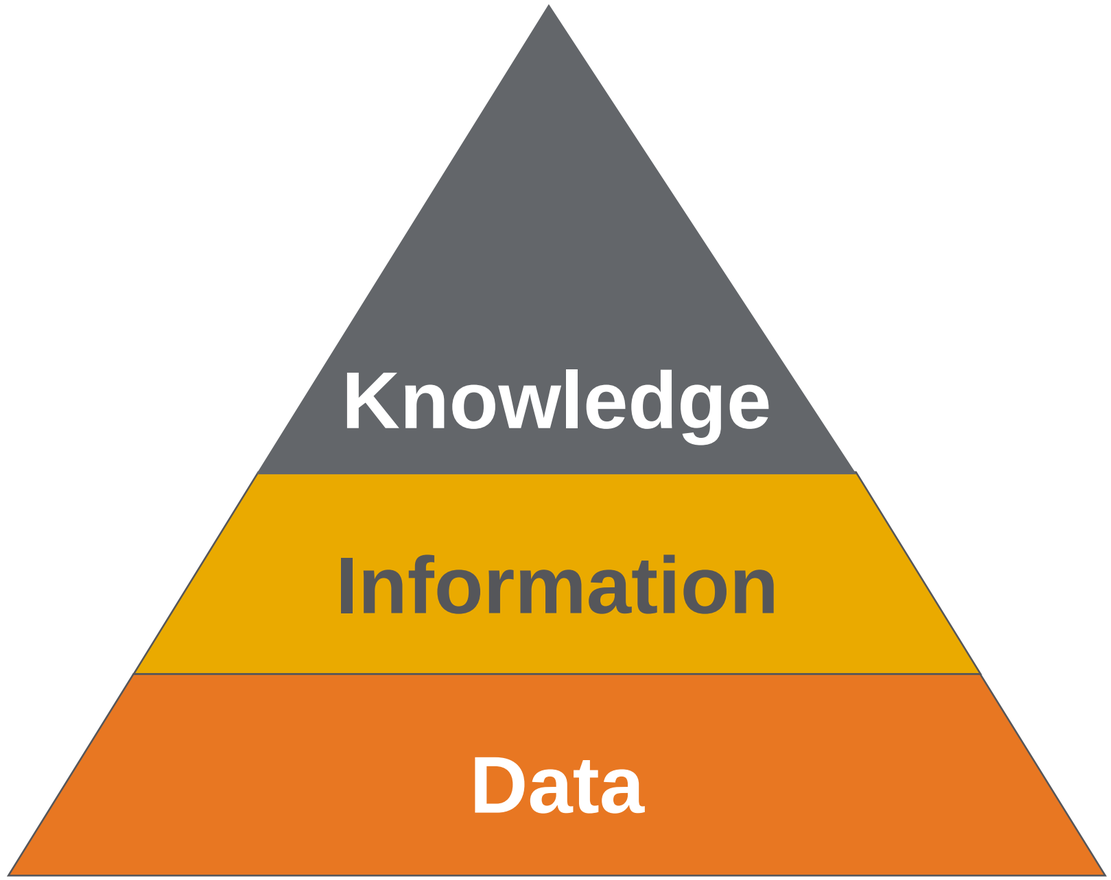


- Graph data is distributed over a large large cluster of nodes
- Graph queries span multiple nodes in the cluster
- Graph queries are distributed to each node and results are returned (just like MapReduce)
- Temporary memory containers (Accumulators) must be transaction safe over the entire cluster

OPTUM®

The Knowledge Triangle

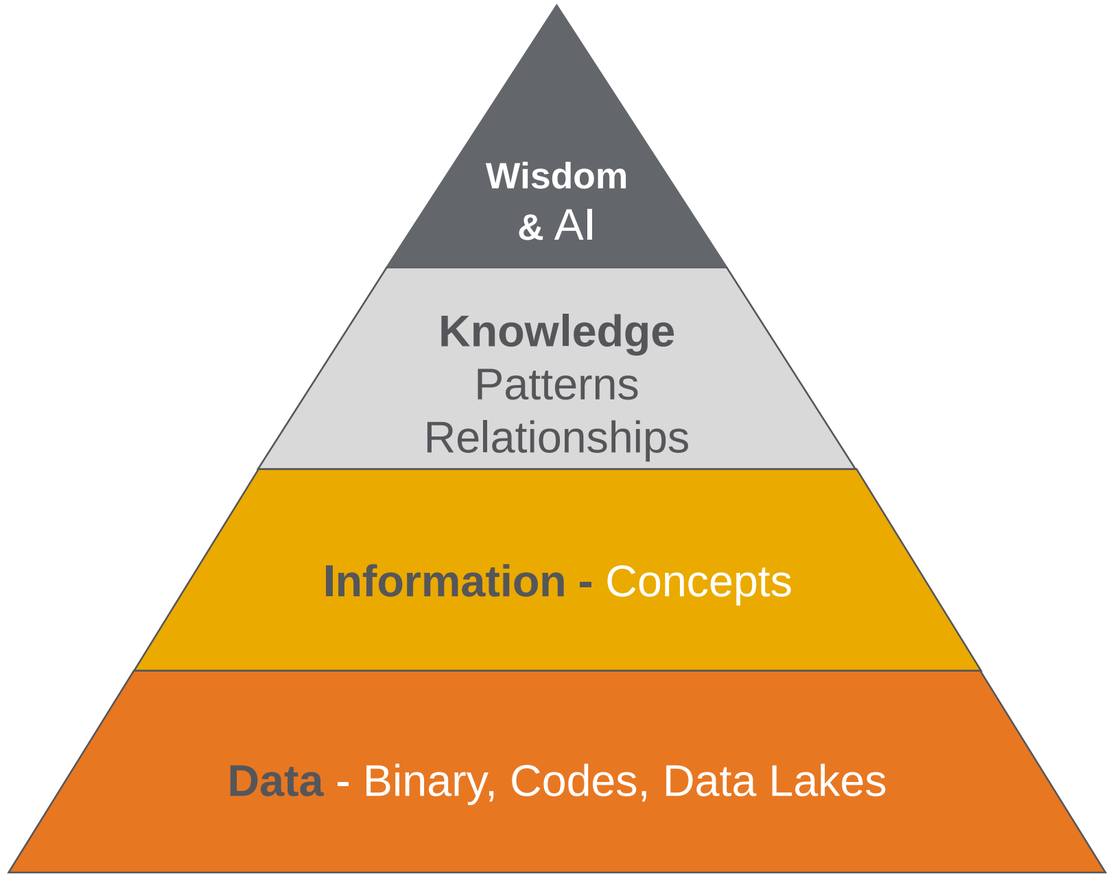
How of graphs capture organizational knowledge.



OPTUM®

The Knowledge Triangle

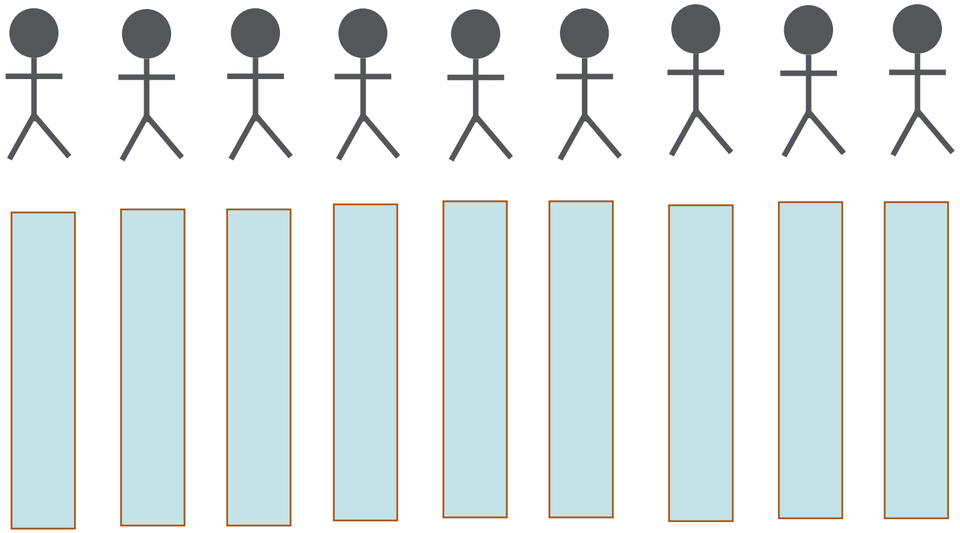
- The Knowledge Triangle or the “DIKW” pyramid
- Diagram for representing the relationships between data, information, knowledge, and wisdom
- Too often we focus on “Big Data” and not enough on connected knowledge and **transferrable** knowledge
- Graphs are connected information concepts
- Wisdom is reusable across multiple context
- Can we capture knowledge in a form that can be reused across multiple domains?



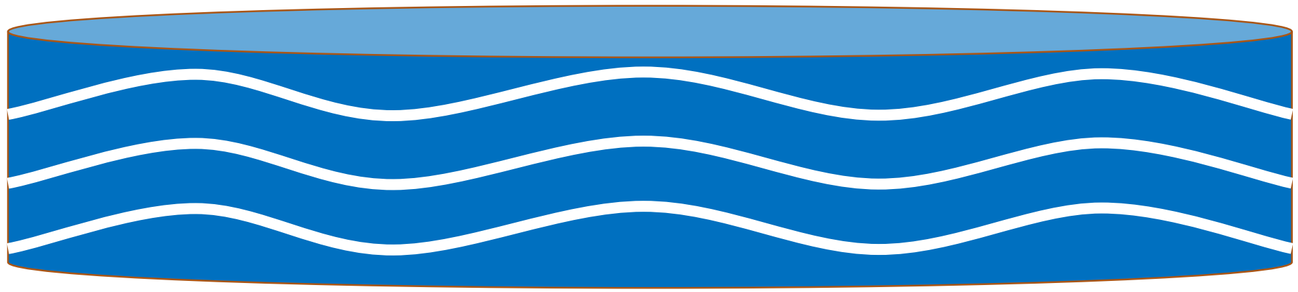
https://en.wikipedia.org/wiki/DIKW_pyramid

OPTUM®

Data Lakes Today



Data Access Code



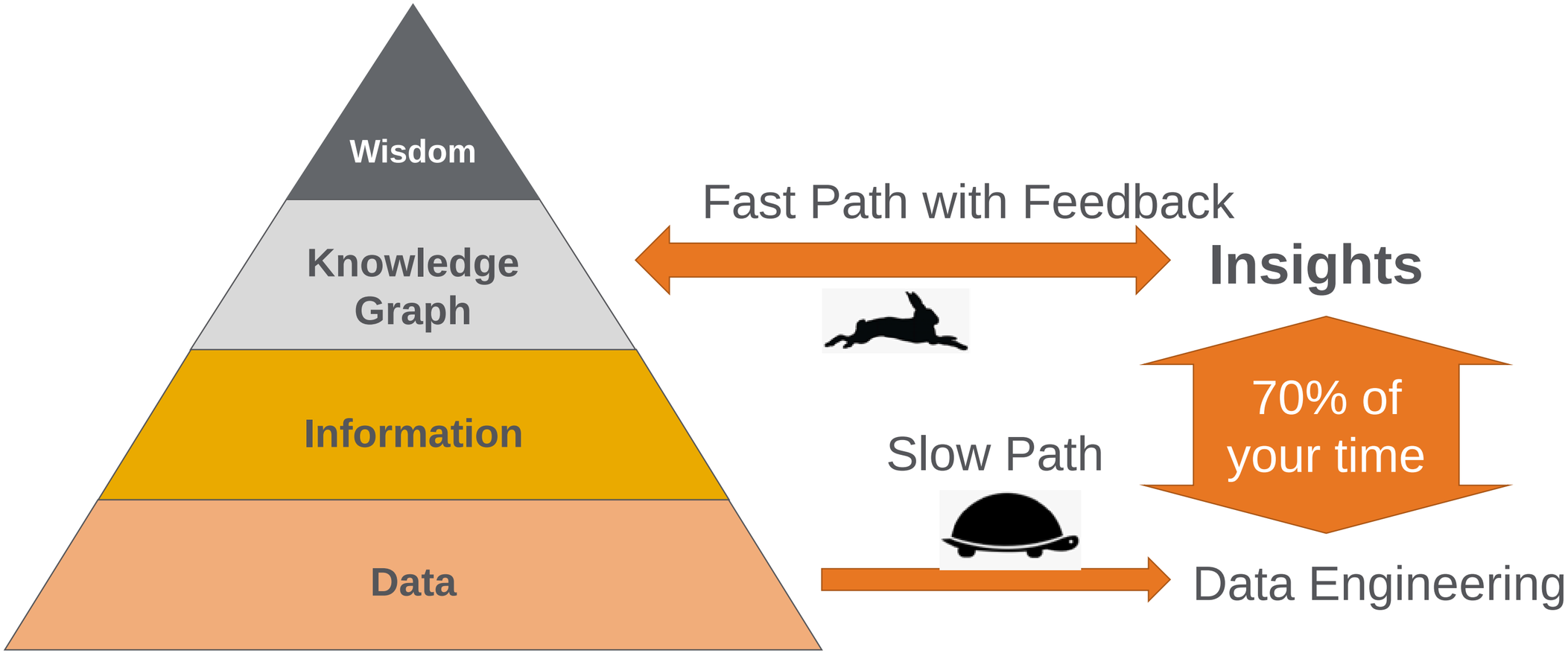
Data Lake

- 100s of Data Scientists
- 100s of R and Python Libraries
- 80% of effort is “Data Engineering”
- Little code reuse
- 20% is real Data Science

No metadata, no semantic layer, no indexing, no search, no ACID transactions

OPTUM®

From Data Scientist to Knowledge Scientist

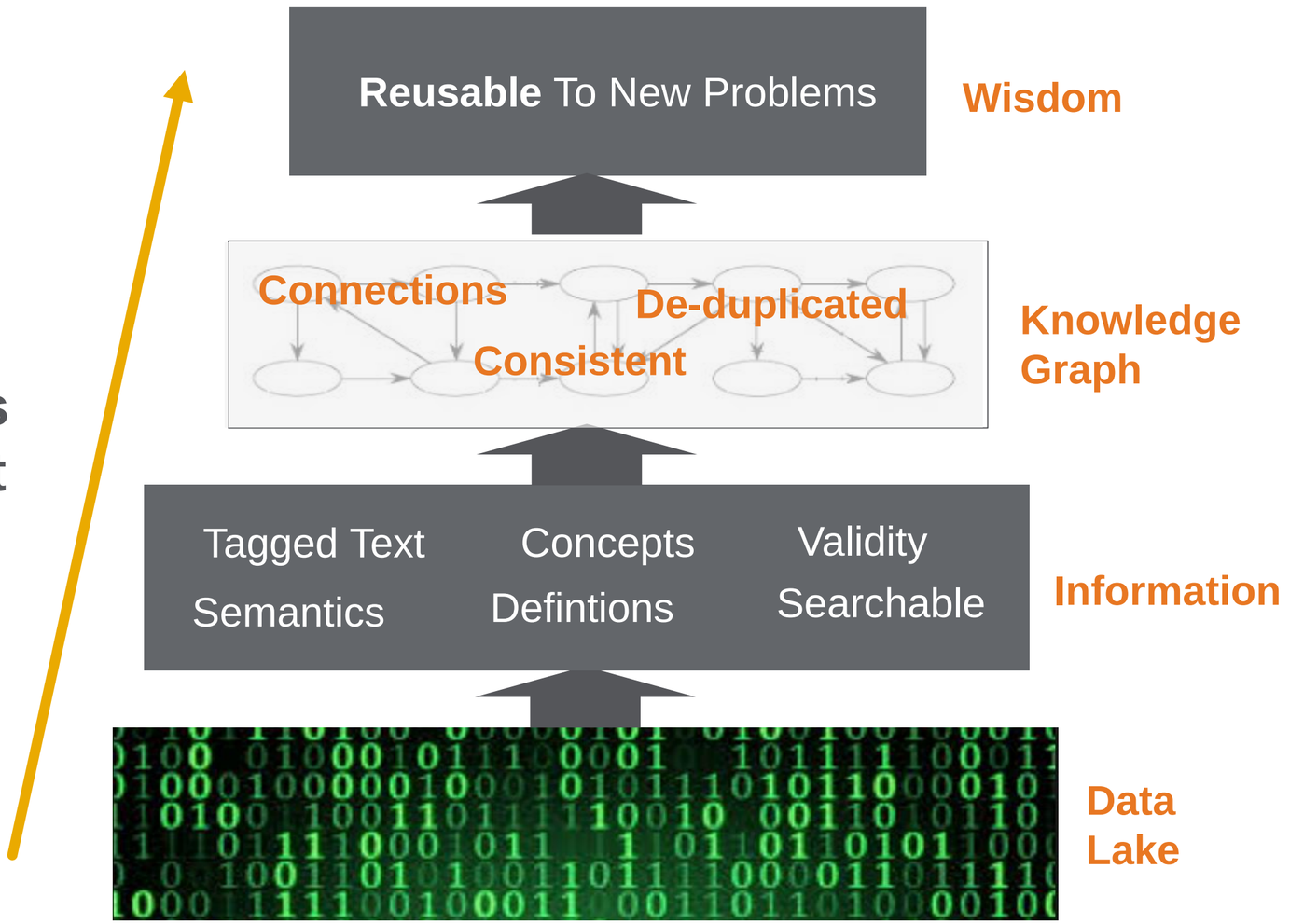


From Raw Data to Wisdom with Continuous Enrichment

Knowledge that can be **reused** across multiple context

Continuous Enrichment

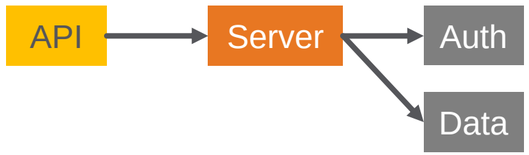
Raw data dumped from a database, log files or documents



Sample Graph Traversal Patterns

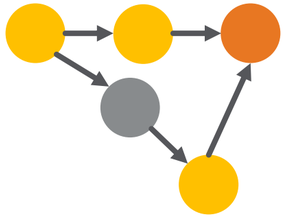
Dependencies

- Failure chains
- Order of operations



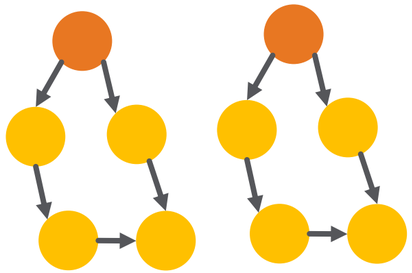
Clustering

- Finding related items
- Friends, fraud networks



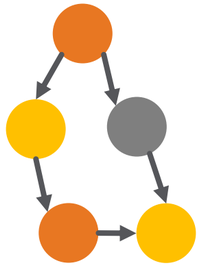
Similarity

- Similar paths and patterns



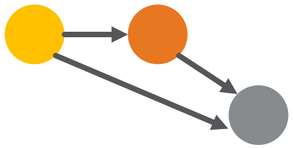
Matching/Categorizing

- Look for and tag specific patterns



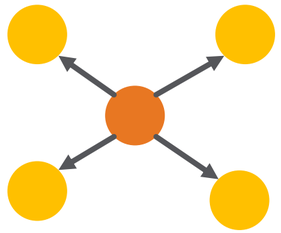
Flow/Cost

- Optimize costs based on routing
- Path optimization



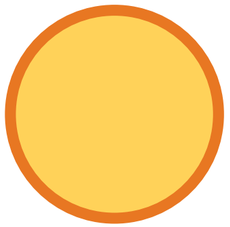
Centrality/Search

- Which nodes are the most connected or relevant?

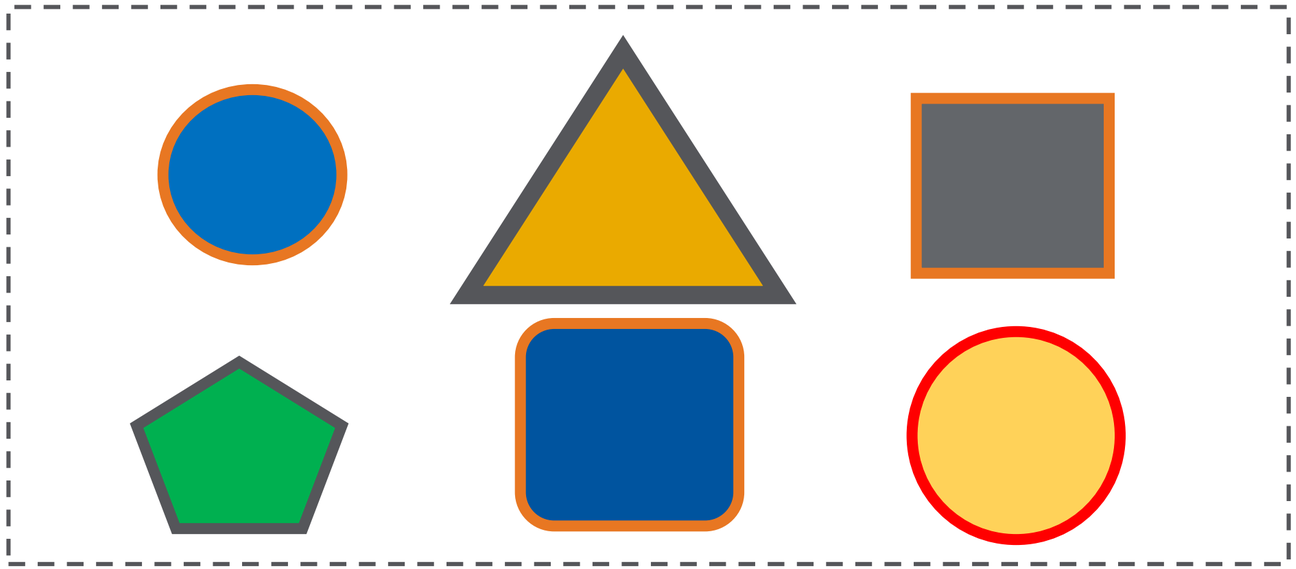


OPTUM®

Property-based Similarity Example



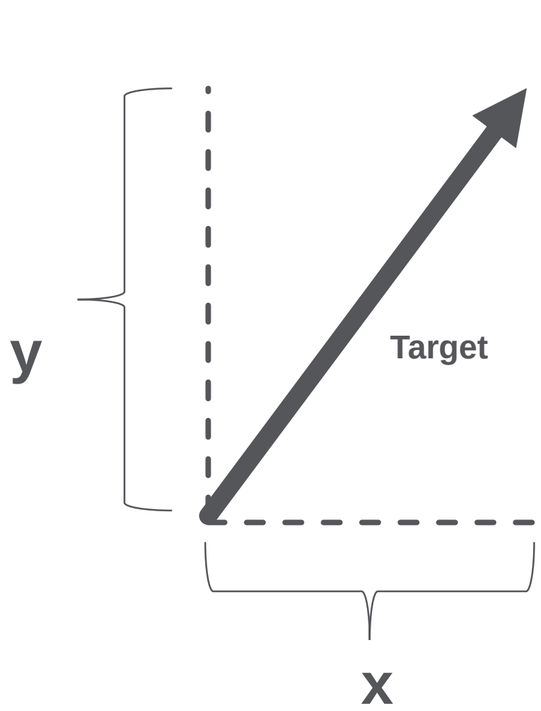
Target

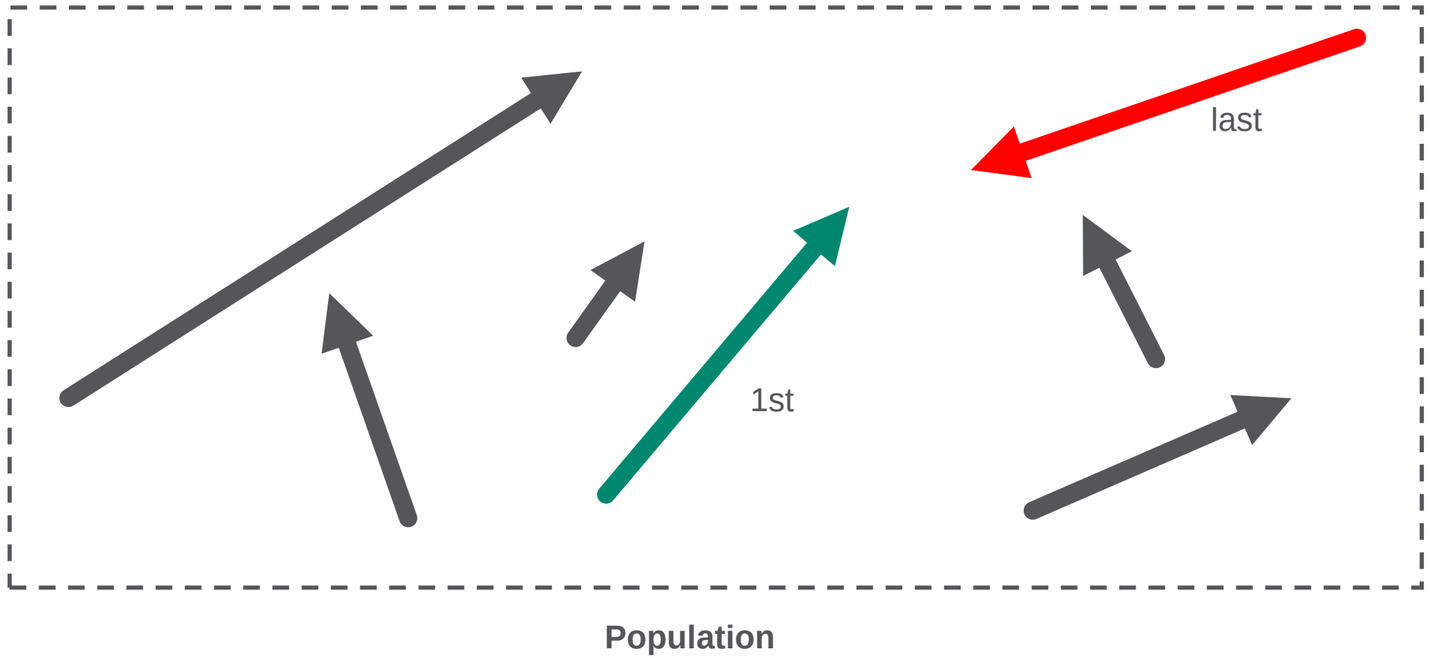


Population

- Used to find the most similar items in a graph by comparing **properties and structure**
- Ideal when you a can compare individual features of an item **numerically**
- Algorithms return a **ranking** of similarity between a target an a population based on the counts and weights of properties that are similar

Vector Similarity





- Vectors are **similar** in two dimensional space if they have the same length and direction
- Compare all the “x” lengths and the “y” lengths and rank by the sum of the totals of the difference

Vector Representation

$$\begin{bmatrix} x=8 \\ y=10 \end{bmatrix}$$

Target Vector

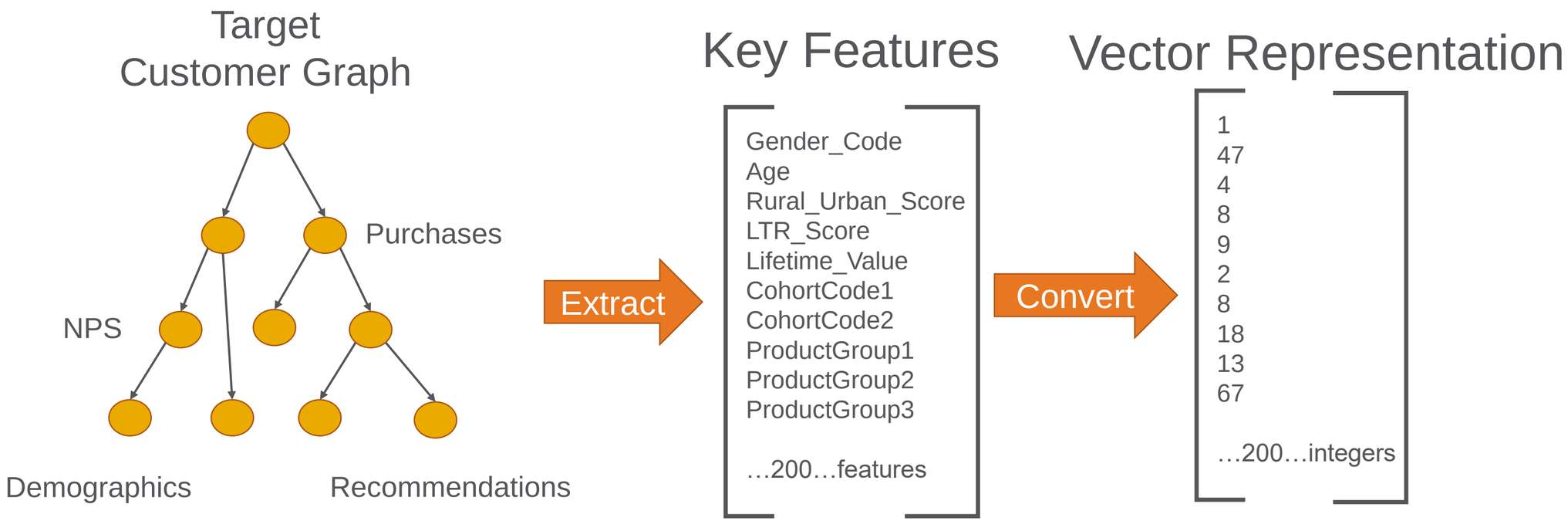
$$\begin{bmatrix} 16 \\ 16 \end{bmatrix} \begin{bmatrix} 8 \\ 6 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 7 \\ 9 \end{bmatrix} \begin{bmatrix} -4 \\ 6 \end{bmatrix} \begin{bmatrix} -4 \\ 6 \end{bmatrix} \begin{bmatrix} -12 \\ -8 \end{bmatrix}$$

Population Vectors

- Each item can be represented by a series of “feature” vectors
- The numbers are scalers

OPTUM®

Graph to Vector



Features should represent both the **properties** and **structure** of your customer's graph

Example: Customer Similarity

Target Customer Vector

0
47
4
8
9
2
8
18
13
67
…
…
…

Cosine similarity

$$\frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Sample Population Vectors

0	1	1
43	45	22
2	4	4
13	8	3
7	9	9
1	3	2
8	8	8
18	15	18
13	13	12
67	45	32
…	…	…
…	…	…
…	…	…

Similarity Score: .80 .85 .92

- Graph **Similarity** algorithms allows graph databases to quickly compare current items with many other items
- For any given customer, we can use a fast, in-memory similarity algorithm to compare the key features of any patients to a larger population
- Cohort codes can be **pre-calculated** to quickly narrow the sample population to a smaller group
- This calculation is considered a “embarrassingly parallel” query and could be accelerated by adding more nodes to a cluster
- Specialized graph hardware such as GPUs and FPGAs can dramatically accelerate these calculations

OPTUM®

The Curse of Dimensionality

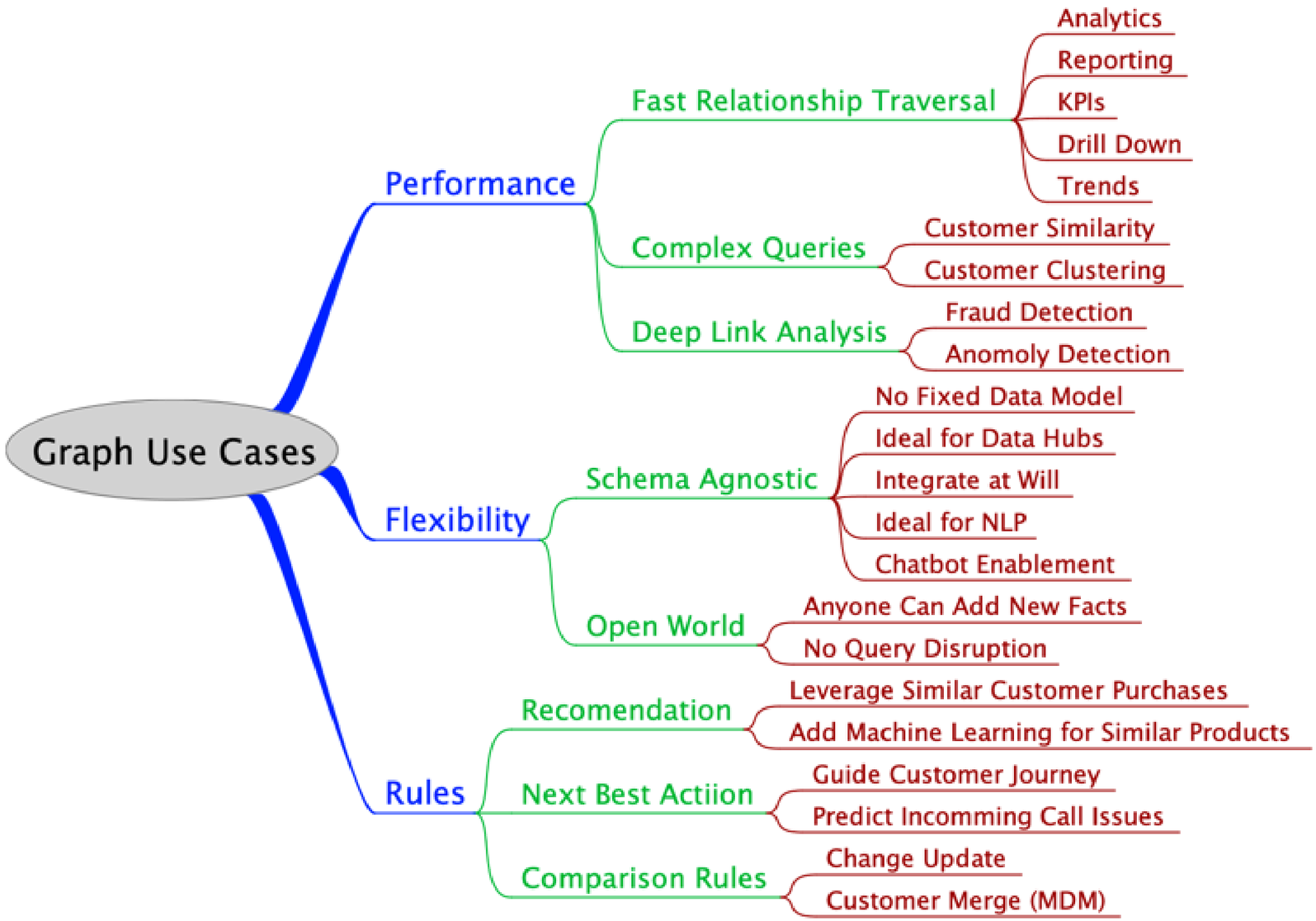
…various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience.

https://en.wikipedia.org/wiki/Curse_of_dimensionality

https://en.wikipedia.org/wiki/Dimensionality_reduction

OPTUM®

When to Use Graph Databases



OPTUM®

Popular Graph Use Cases

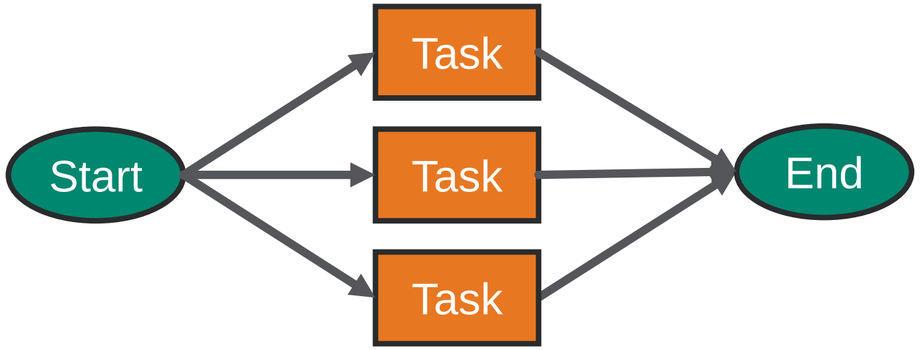
1. Member Journey Graph (graph and an integration data hub)
2. Faceted Search (Health System Explorer)
3. Human Capital Graph (Enterprise People Fabric) Social Networks for staff/skill search
4. Knowledge Graphs (semantic search)
5. Fraud Pattern Detection
6. Computer Network/Service Failure Analysis
   Application resource dependency mapping
7. Rules Engine – Recommendation Engine – Next Best Action
8. Master Data Management – rules for record matching (Entity Resolution)
9. Genomics – Gene Ontologies
10. Ingestion Rules - ETL Replacement (CentriHealth)
11. 360 View of Customer/Products/Servers (Data Hub)
12. Company/Investor graph – competitive analysis
13. Causal Graphs (Bayesian networks)
14. Drug interaction/Drug similarity
15. Map routing - (shortest distance calculations)

Serial vs. Parallel Graph Algorithms

Serial Graph Algorithms

- One task cannot begin before the prior task is complete
- Task order is important
- Serial Algorithms cannot easily be optimized on FPGAs

Parallel Graph Algorithms



- Many tasks can be done independently
- Task order in not relevent
- Tasks can usually be done faster on FPGAs

OPTUM®

Does the Human Brain do Serial or Parallel Computation?

- The following slide photos of two people
- One is a famous actor
- The other is a synthetically generated image of a person (generated by a GAN)
- Shout out “**Left**” or “**Right**” as soon as you can tell which is the **famous actor**

OPTUM®

What just happened

1. The visual cortex received the images as electrical signals from your eyes
2. Your brain identified key **features** of each face from the images - in **parallel**
3. Your brain sent these features as electrical signals to your memories of people's faces
4. Your brain compared these features to **every memory you have of a person's face** – in **parallel**
5. Your brain sent their recognition scores to a control center of your brain
6. Your brain's speech center vocalized the word "right" – in **series**

Answer: The human brain, comprised of around 84B neurons, does **both** parallel and series calculations

Two Key Questions:

1. How does the brain know to pay **attention** to specific features of a face?
2. What portions of real-time recommendation systems can be done cost effectively in **parallel** at low cost?

OPTUM®

The Product Recommendation Challenge

1. A customer comes to your web site
2. You have information on their prior purchases and they view new products
3. How quickly can you generate a real-time recommendation based on 10 products and 1 million product reviews?
4. How much detail can you take into account and return the best match in 100 milliseconds?

How quickly can we compare a given customer and their purchases to 1 million other customers and their purchases?

What comparison tasks can be done in parallel?

The Rise of Automatic Graph Feature Engineering

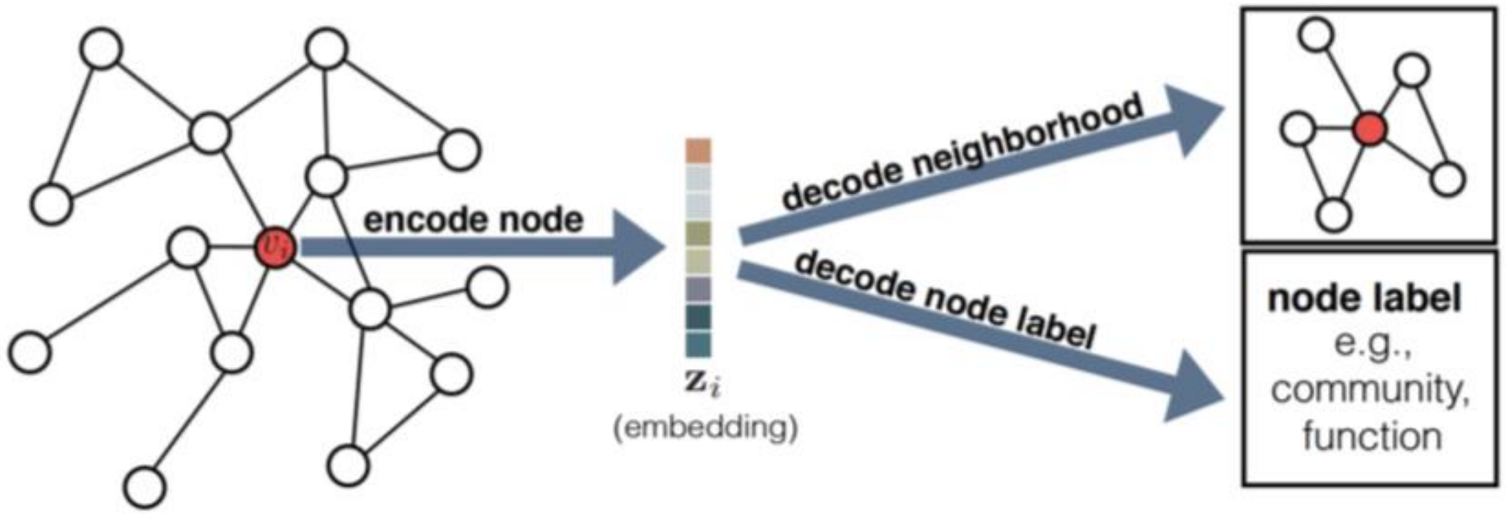
*Recent years have seen a surge in approaches that **automatically** learn to encode graph structure into low-dimensional **embeddings**.*

*The central problem in machine learning on graphs is finding a way to incorporate information about the **structure** of the graph into the machine learning model.*

From Representation Learning on Graphs: Methods and Applications by Hamilton et. El.

OPTUM®

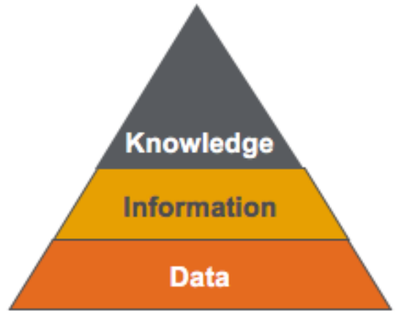
Example of Graph Embedding – Encode and Decode



From Representation Learning on Graphs: Methods and Applications by Hamilton et. El.

Four Graph Stories and Metaphors

The Neighborhood Walk
index free agency, performance



The Knowledge Triangle
data, information and knowledge

The Open World Assumption
graph integration, agility

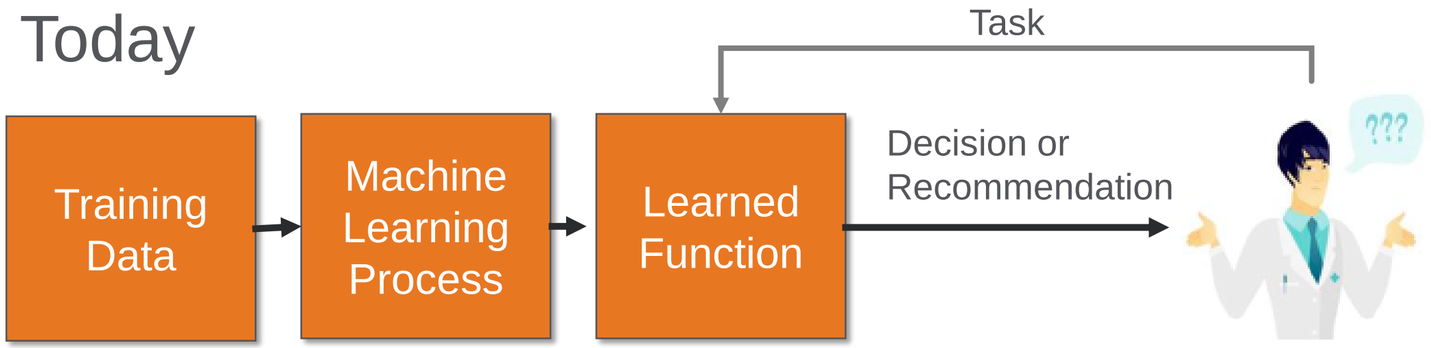
The Jenga Tower
the resilience of graphs to change

OPTUM®

The Deep Learning Bartender

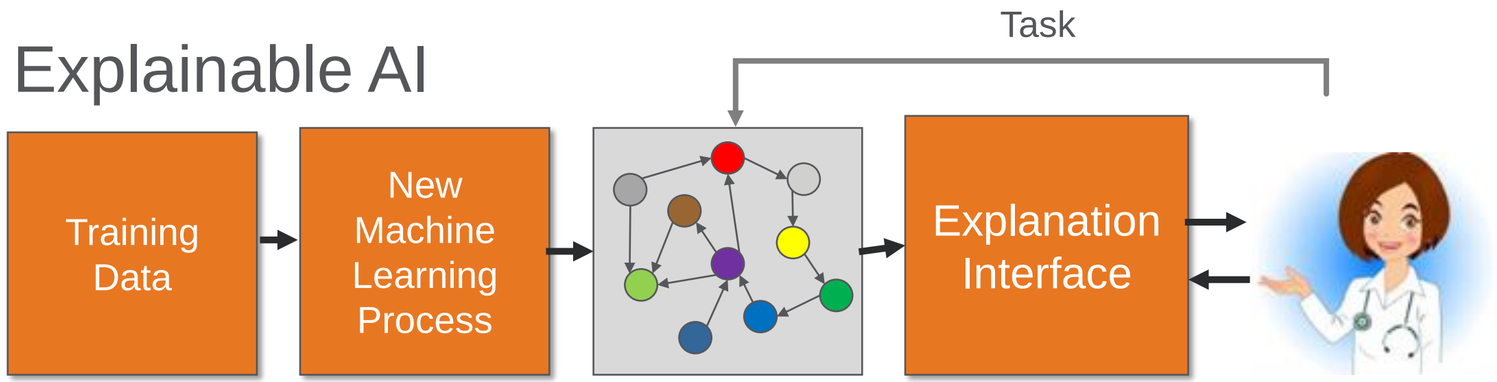
Explainable AI Models Leverage Graphs

Today



- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How to I correct an error?

Explainable AI

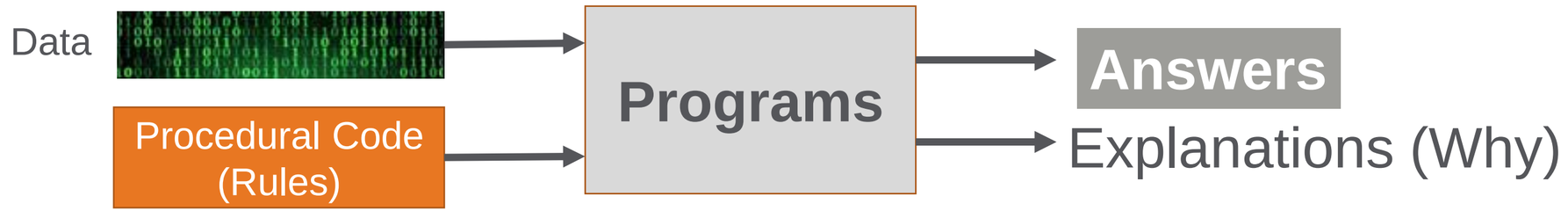


- I understand why
- I understand why not
- I know when you succeed
- I know when you fail
- I know when to trust you
- I know why you erred

https://www.darpa.mil/program/explainable-artificial-intelligence

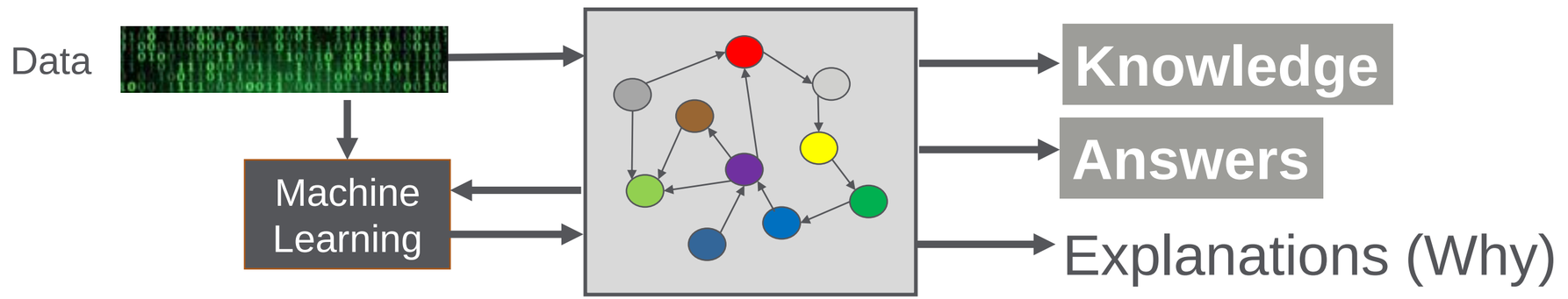
Three Eras of Computing

1) Procedural Era



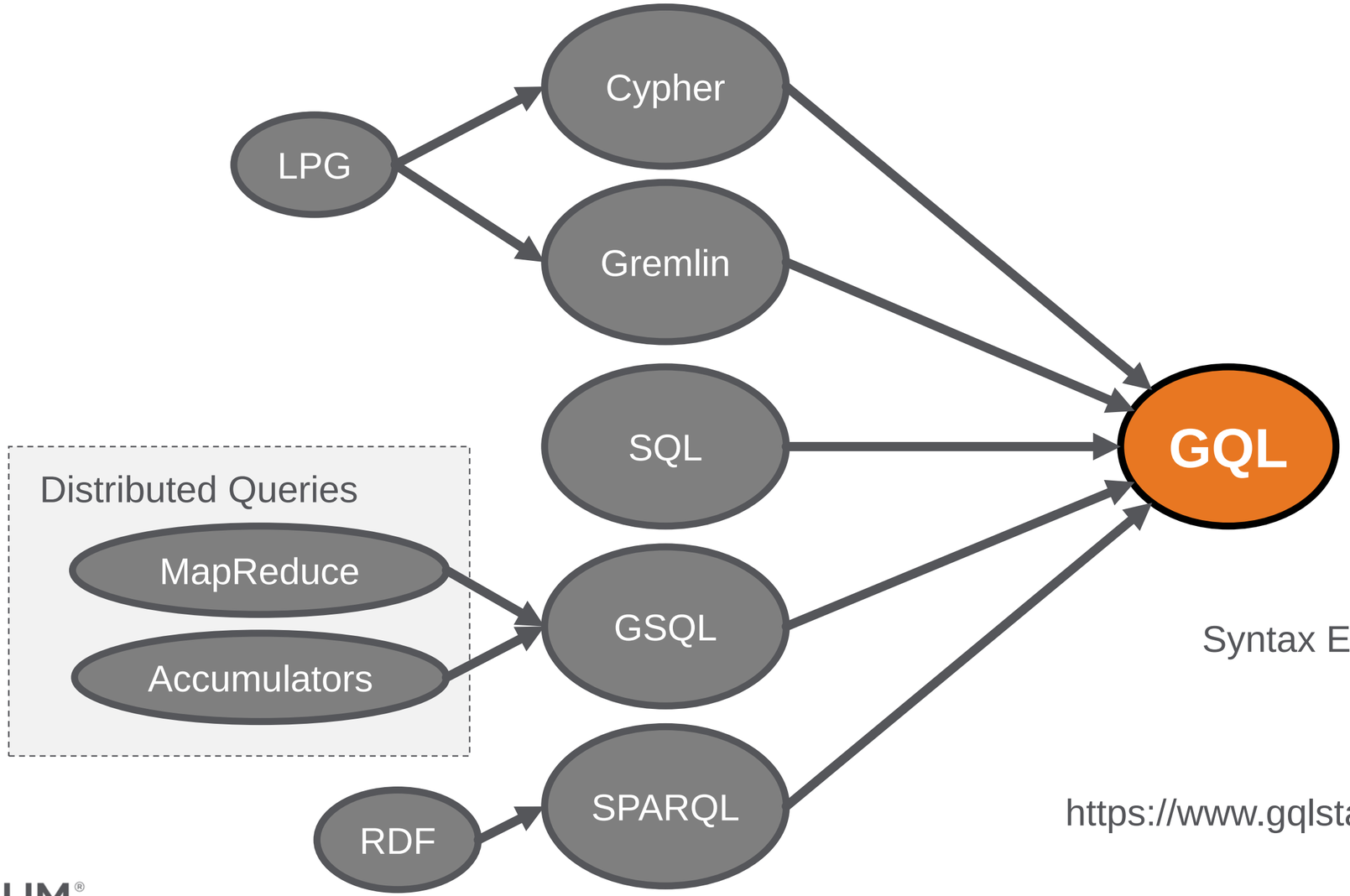
2) Machine Learning Era

3) Graph Era



OPTUM®

Standard ISO Property Graph Query Language (GQL)

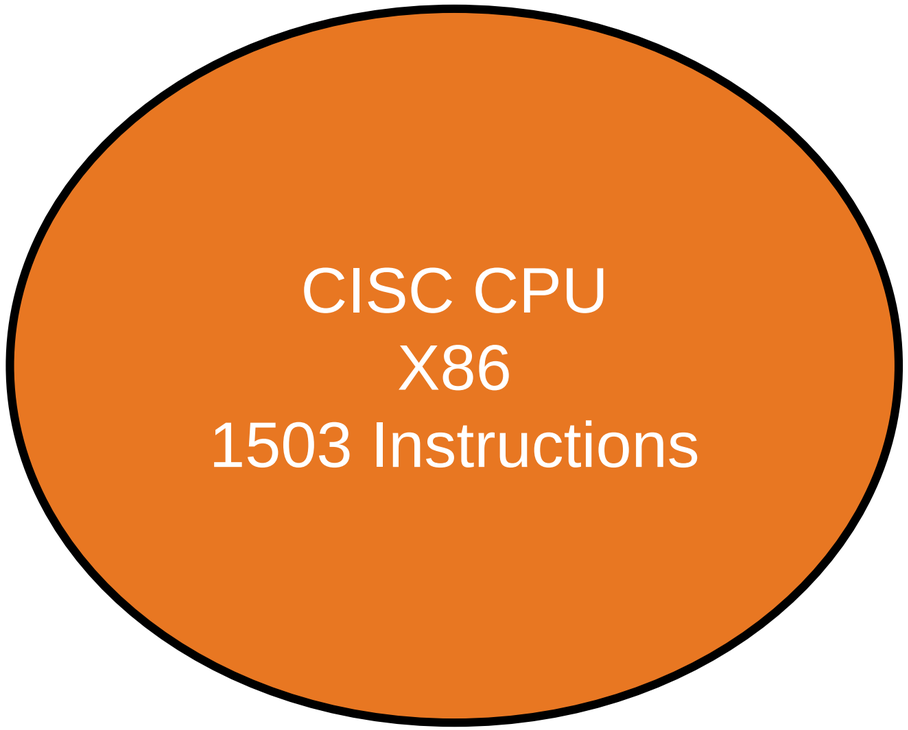


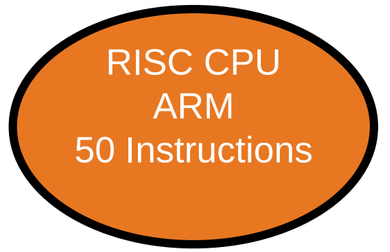
Syntax Expected in Q4 2020

https://www.gqlstandards.org

OPTUM®

CISC vs RISC

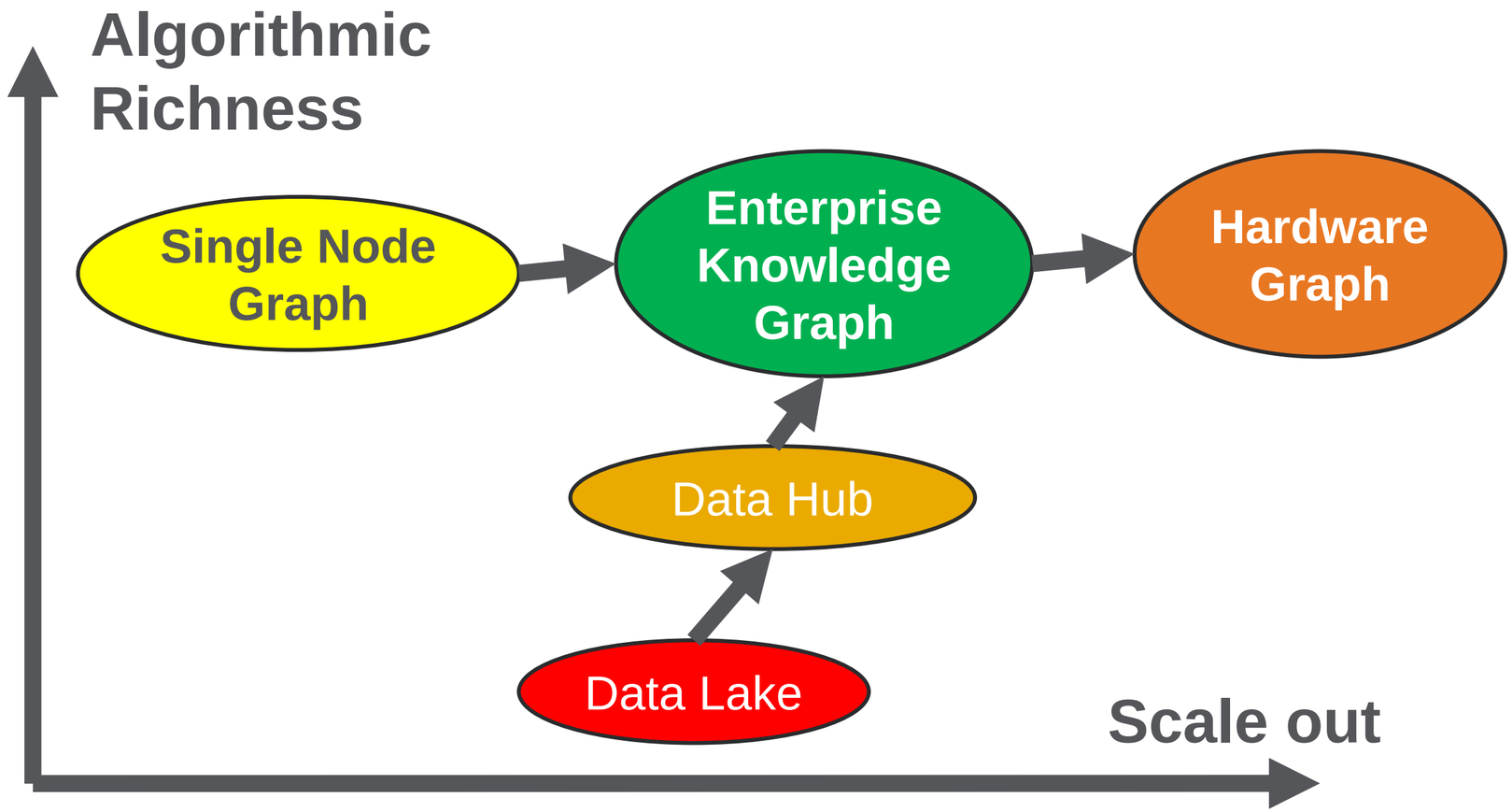




- Most graph traversal algorithms only use simple pointer hopping
- How efficient are CISC and RISC at running graph algorithms?
  - No need for floating point
  - No need for matrix multiplication
  - No instruction for doing encryption / decryption

OPTUM®

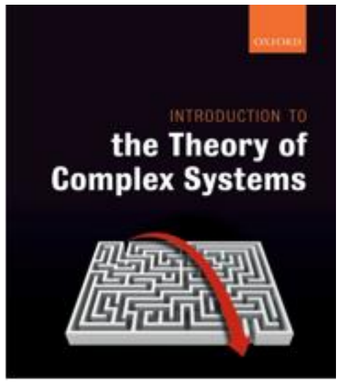
Onward to the Hardware Graph!



Emergence

Can we predict what new insights we will find when we connect systems that have never been connected together before?

What is “the Edge of Chaos”?



Introduction to the Theory of Complex Systems
Peter Klimek, Rudolf Hanel, and Stefan Thurner

OPTUM®

Thank you!

Dan.McCreary@optum.com

https://www.linkedin.com/in/danmccreary

https://medium.com/@dmccreary

@dmccreary

OPTUM®