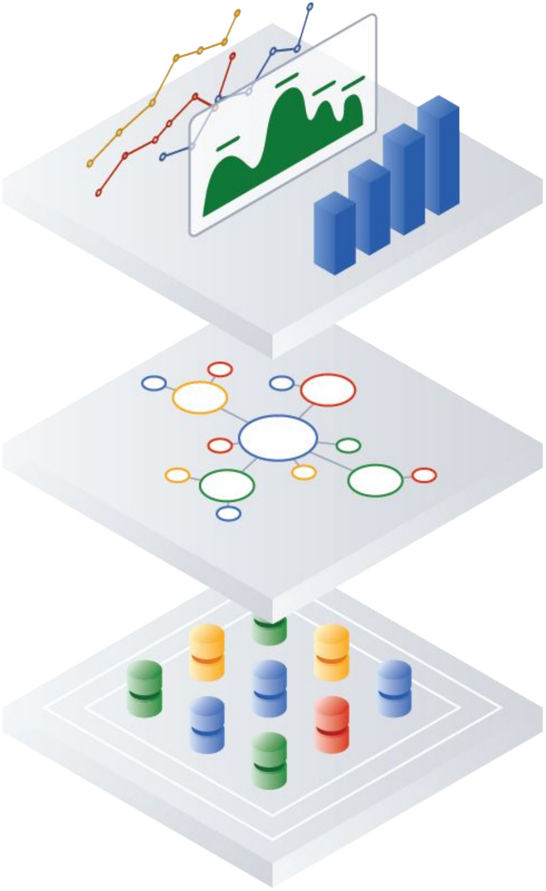
Google Cloud

Knowledge Graphs for Security

January 2022
09:45 - 10:15 MST

Dr Scott Mongeau CSci CAP
Solutions Engineer
Public Sector
Google Cloud
smongeau@google.com

All opinions are his my own and not those of my employer.
All images and references used are purely for educational purposes.



Knowledge Graphs for Security

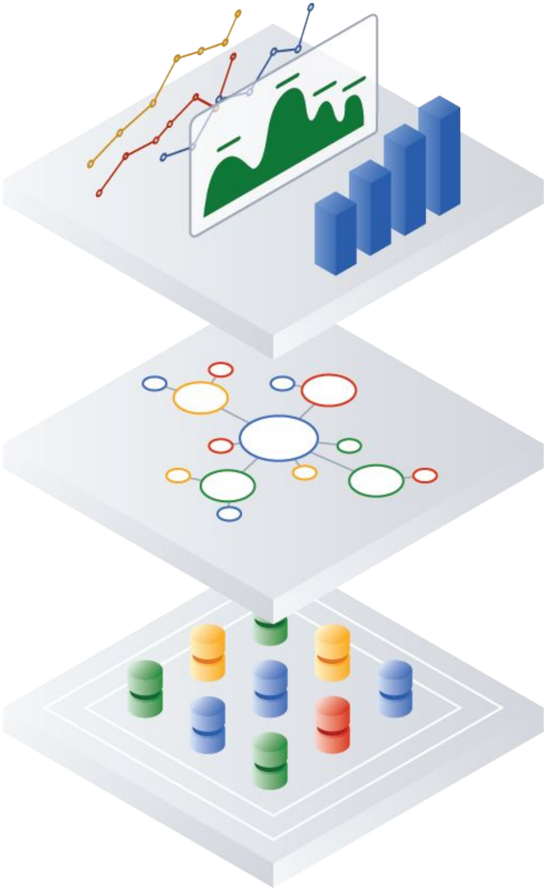
I. **WHAT?** Knowledge Graphs (KGs)

II. **WHY?** KG Use Cases

III. **HOW & WHO?** Implementing KGs

IV. **WHERE TO?** Future Prospects

V. Questions & Discussion

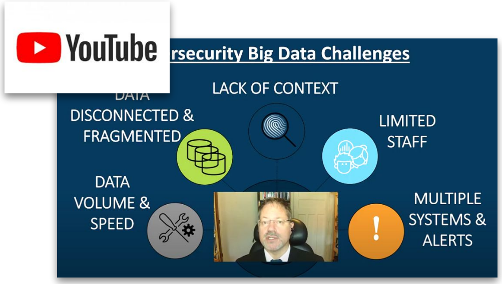


Google Cloud

AMSTERDAM

NYENRODE BUSINESS UNIVERSITEIT

RSM Erasmus



Cybersecurity Analytics (YouTube)

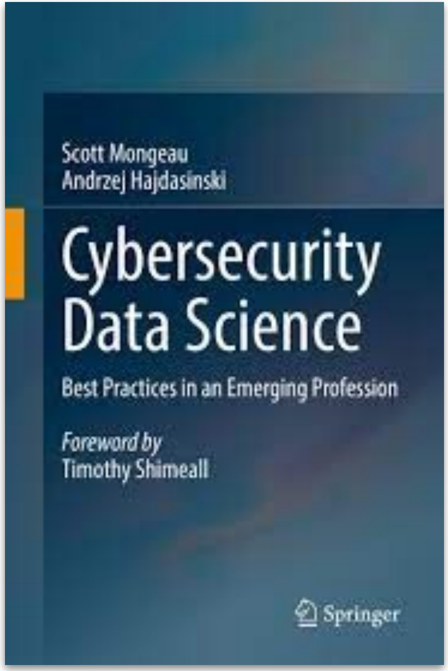
Dr Scott Mongeau CSci CAP

Solutions Engineer
Public Sector
Google Cloud
smongeau@google.com

linkedin.com/in/smongeau

CSci Chartered Scientist

CAP



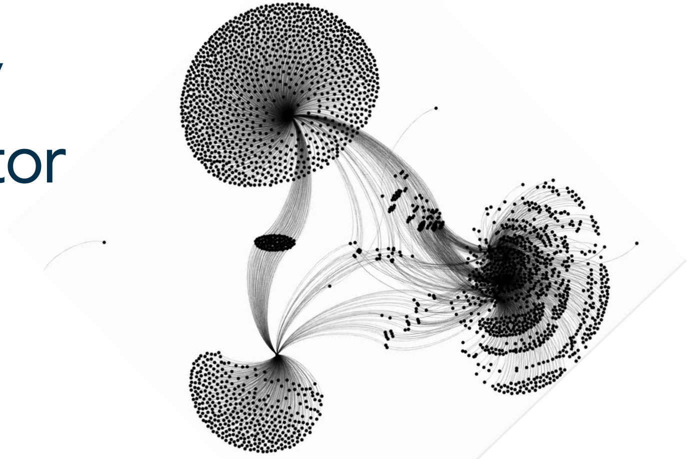
Book (Springer)
Research Portfolio
www.sark7.com/csds

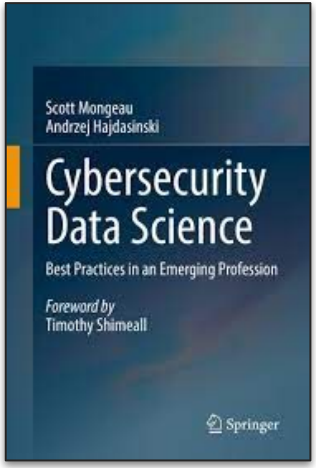
Flocon 2021 **Emerging Trends** YouTube | slides

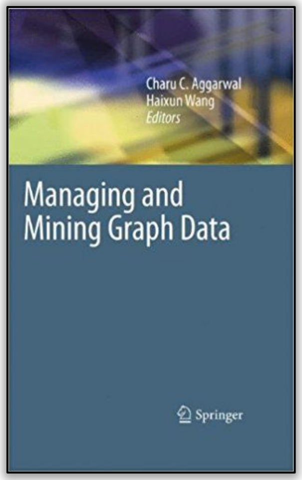
Network Graph Analytics

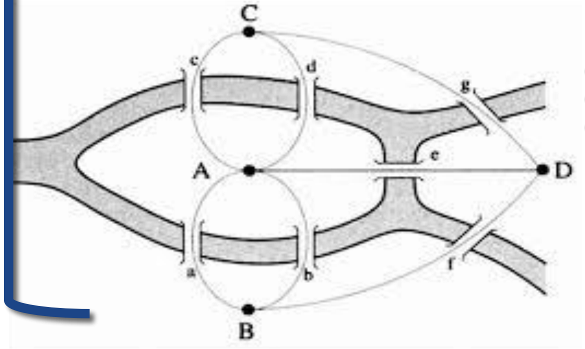
METHODS
Deep learning – computer vision & acoustics
Network graph analytics
Natural language, semantic & knowledge engineering
Forecasting and time series analysis

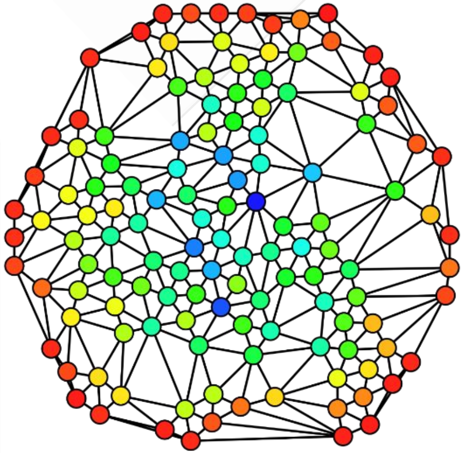
- Centrality
- Eigenvector
- Density
- Reach
- Strength
- Reciprocity







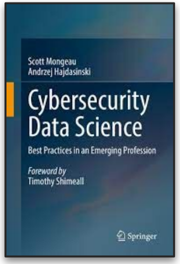


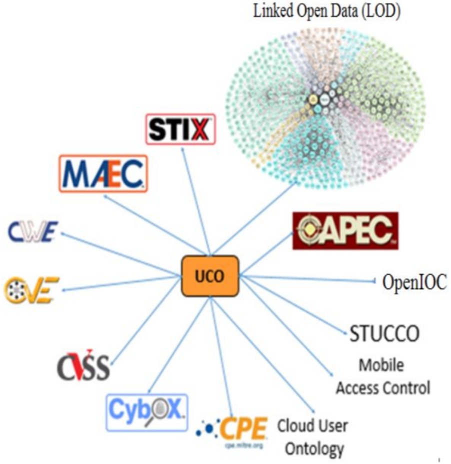


Flocon 2021 **Emerging Trends** YouTube **|** slides

NLP & Semantic Engineering

METHODS
Deep learning – computer vision & acoustics
Network graph analytics
Natural language, semantic & knowledge engineering
Forecasting and time series analysis





FRAMEWORKS

- MITRE Cyber Observable eXpression
- NIST Cybersecurity Framework
- Intrusion Kill Chain (Lockheed Martin)

ONTOLOGIES

- **DFAX** Digital Forensic Analysis eXpression
- **CVE** Cyber Intelligence Ontology
- **ICAS** Information Security (example)
- **UCO / UCO (OWL)**

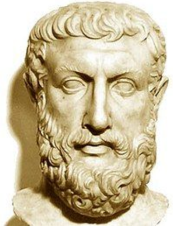
  Unified Cybersecurity Ontology

www.us-cert.gov/Information-Sharing-Specifications-Cybersecurity

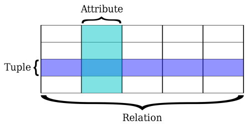
Knowledge Graphs

Two+ Millennia of Semantic Engineering

Analytic Philosophy, Classical Logic, Computer Science



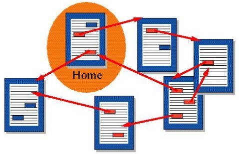
Classical Metaphysics & Ancient Logic



1970

Relational Databases

Structured and normalized data



1989

World Wide Web

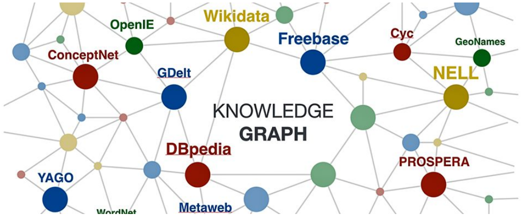
Document + link

Semantic Web

1999

Semantic Web

Ontological engineering



2006

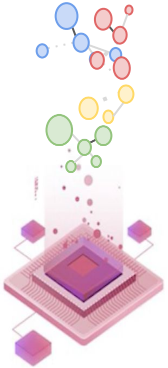
Interlinked Data

Graph-based context

2012

Google Knowledge Graph

Entity + graph links



2022

Semantic Engineering

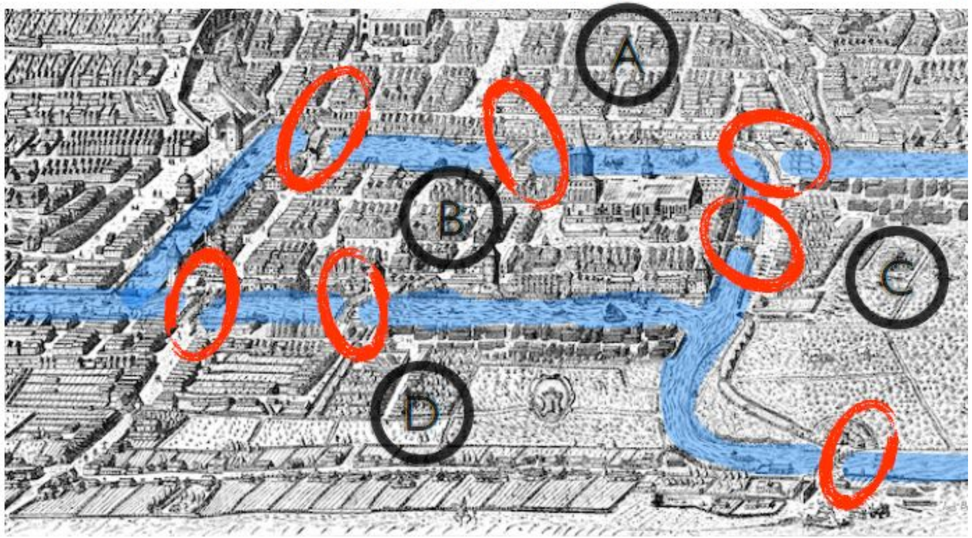
Hybrid human-machine facilitation

Neuro-Symbolic ML

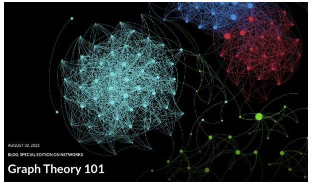
Graph Analytics... ~300 year old innovation

Leonhard Euler 1707-1783

Königsberg (Prussia) - 1736



Seven Bridges of Königsberg (Euler 1736)



- Harvard University Graph Theory 101 - Networks in everyday life
- BulitIn An Introduction to Graph Theory
- Google PageRank algorithm

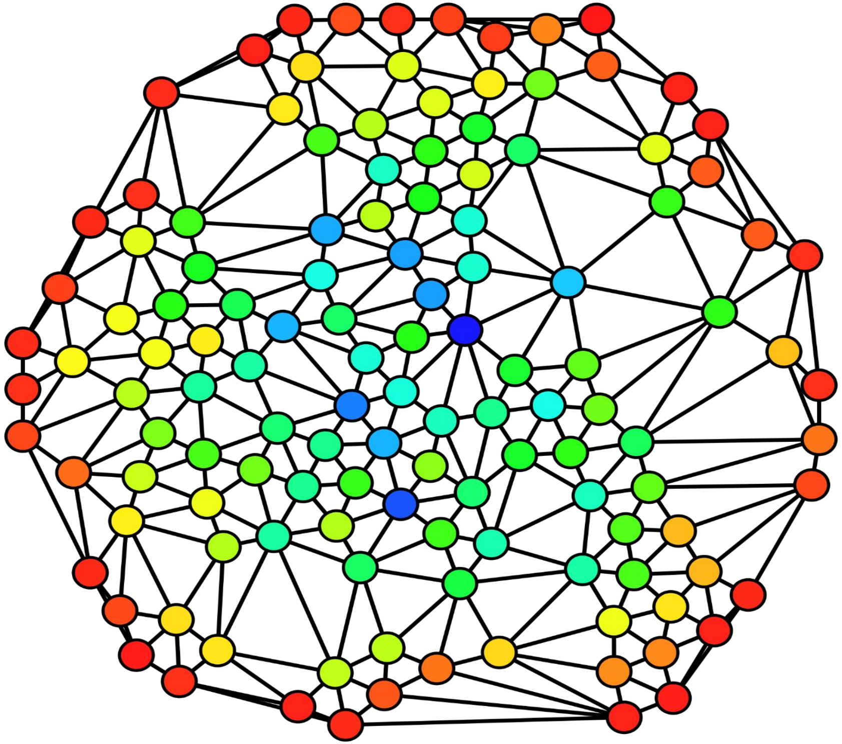
Quantify 'structural' statistical patterns

- Bridges
- Outliers
- Centrality
- Eigenvector

- Density
- Reach

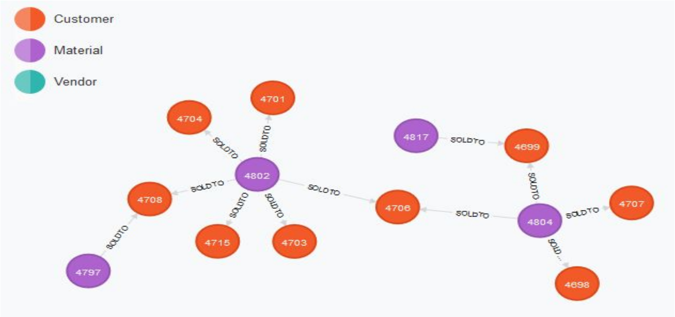
- Strength
- Reciprocity

- Modularity
- Community
- Subgraphs (matching)

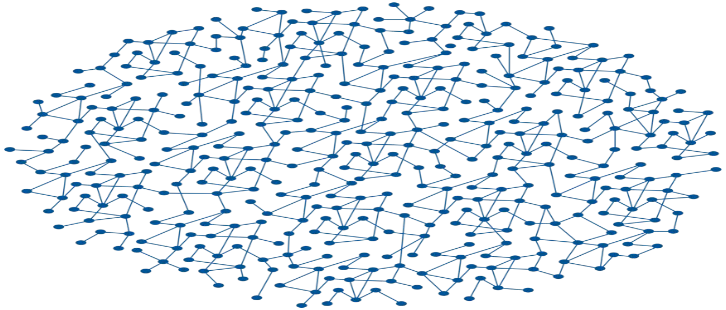


Discover patterns in networked ‘big data’

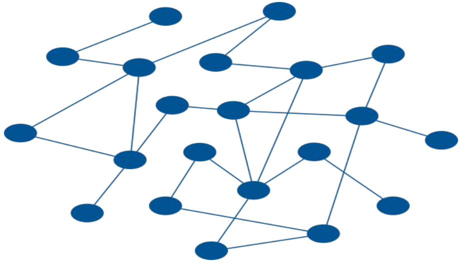
1. Build graph of connected entities



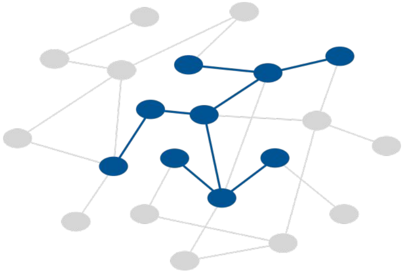
2. Generate and visualize network (e.g. people, transactions, property)



3. Identify ‘normal’ and unusual clusters



4. Identify patterns and chains

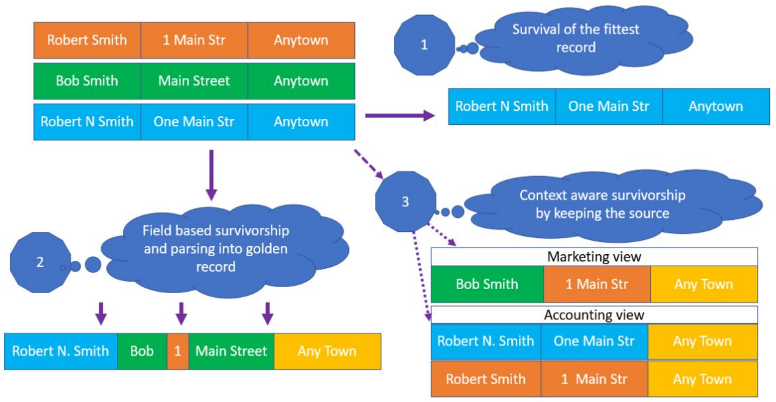


5. Search for known and new patterns / rules

Graphs for structured data storage & retreival

Relational databases

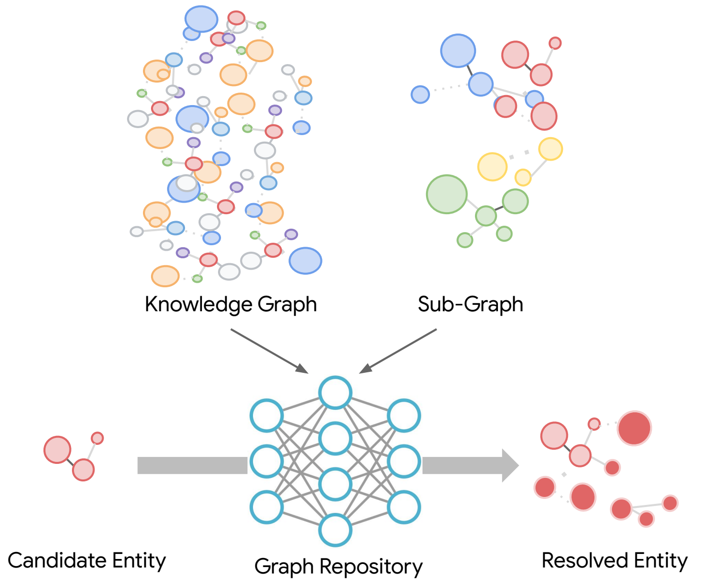
Strictly defined linear relationships & structured entities (brittle, limited extensibility)



Source mdmlist.com/2019/08/22/three-master-data-survivorship-approaches

Graph datastores

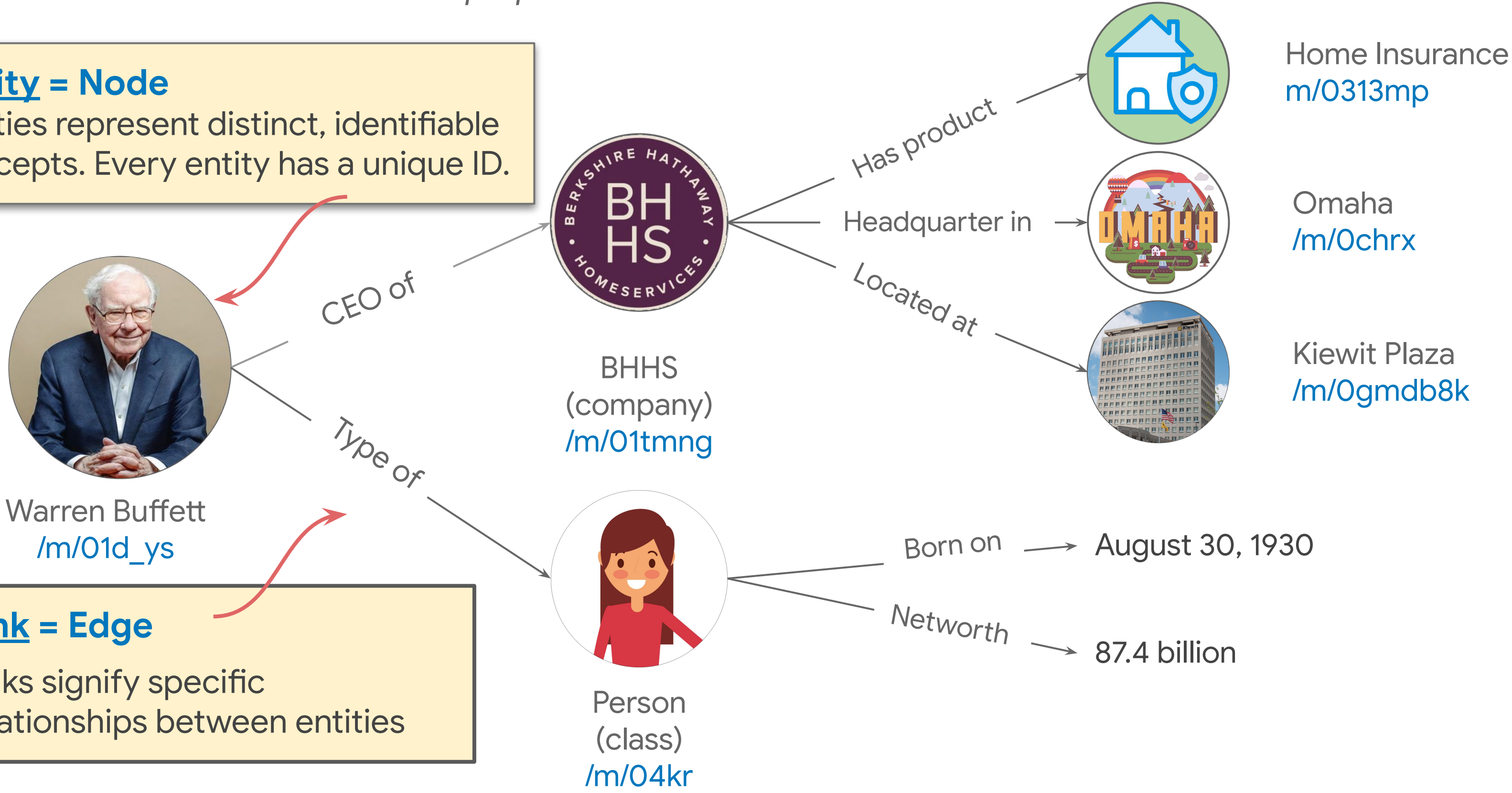
Graph-based data structures support flexible representations of multi-contextual domains



How a Knowledge Graph represents data

Links connect Entities with properties

Entity = Node
Entities represent distinct, identifiable concepts. Every entity has a unique ID.



Link = Edge

Links signify specific relationships between entities

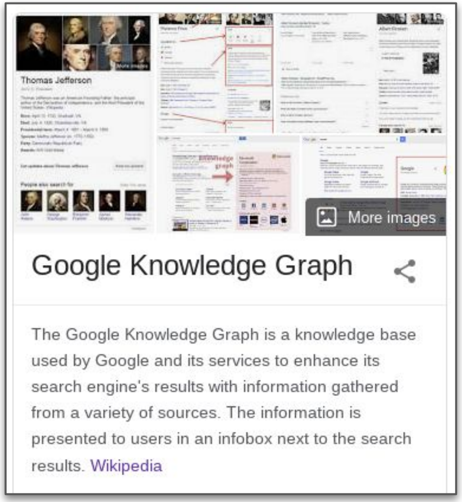
Google Knowledge Graph

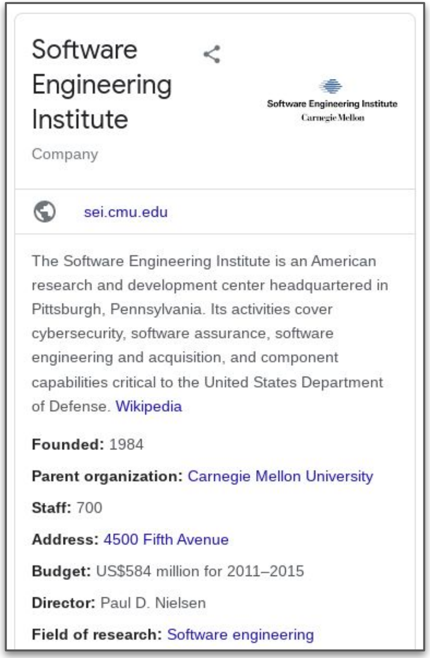
Massive collection of structured data about the world:

- **entities** (people, places, things)
- factual **attributes** of entities
- **relationships** between entities

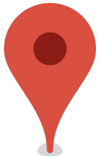
Google Knowledge Graph

- **BLOG** Things, not strings
- **BLOG** Knowledge Graph Panels



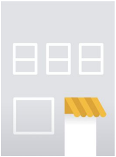


Places



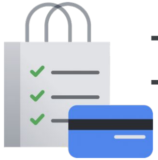
- ➔ >200M points of interest
- ➔ >100M local guides
- ➔ ~20M updates / day
- ➔ >1B monthly Map users

Businesses



- ➔ Mil's of organizations
- ➔ Bil's interactions / month
- ➔ Mil's fraudulent profiles detected & removed / yr

Products



- ➔ B+ entities
- ➔ M+ new entities / day

Property Graph Less rigorous than an ontology

Less formal structure (flexible)

Entities / nodes

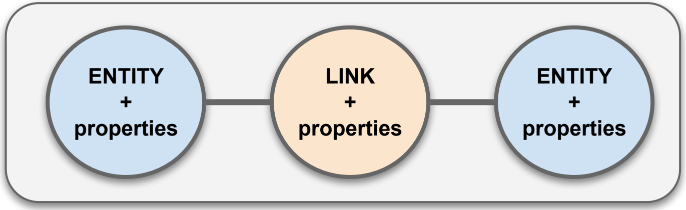
Links / relationships

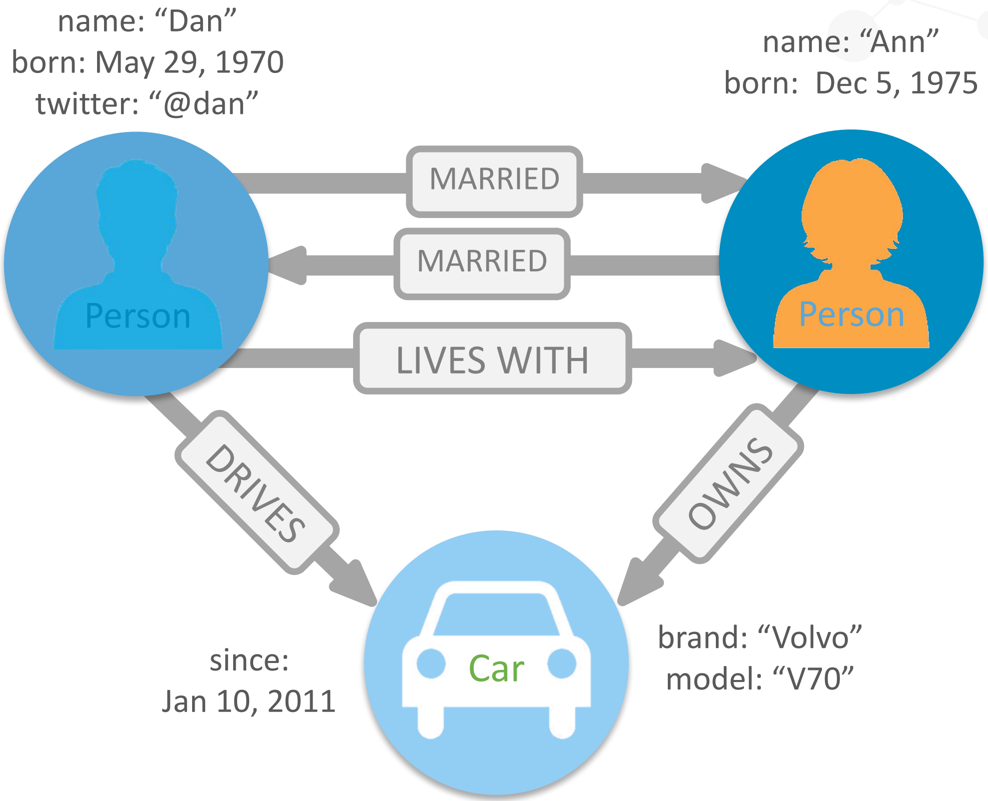
- Relate nodes by *type* & *direction*

Properties

- On both entities & links

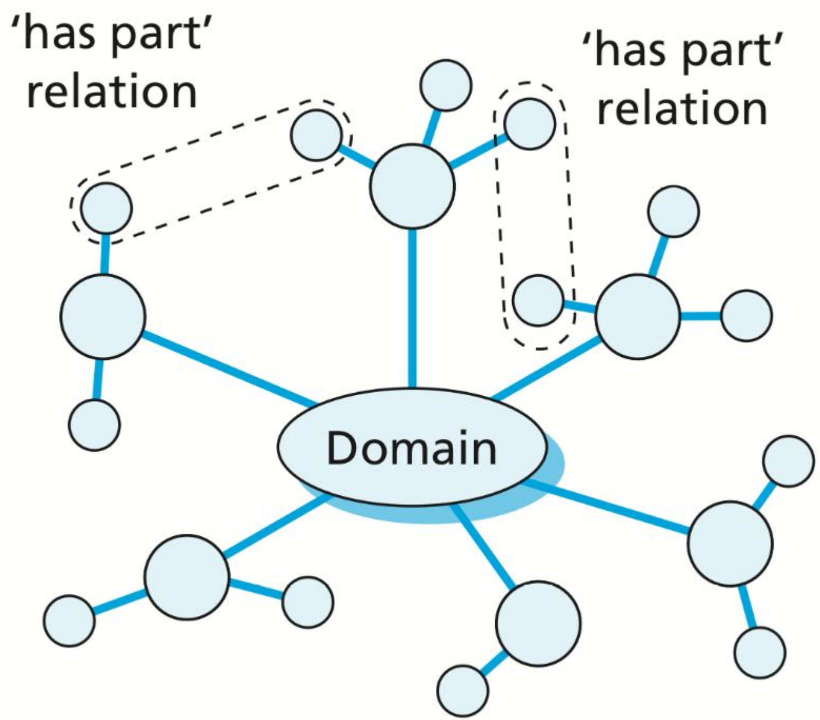
Semantic context inferred in structure





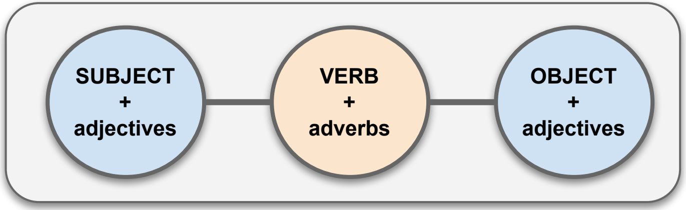
neo4j

Ontology More powerful - more effort required



- Explicit formal description of entities, properties, and relations between elements in a conceptual domain (Gruber 1993)
- Computer-readable *(support for logical reasoning & inference)*

'Triple': logical linguistic factual assertion



More formal structure (strict)	***Support for automated reasoning***	***More effort to deploy***

UCO
Unified Cybersecurity Ontology

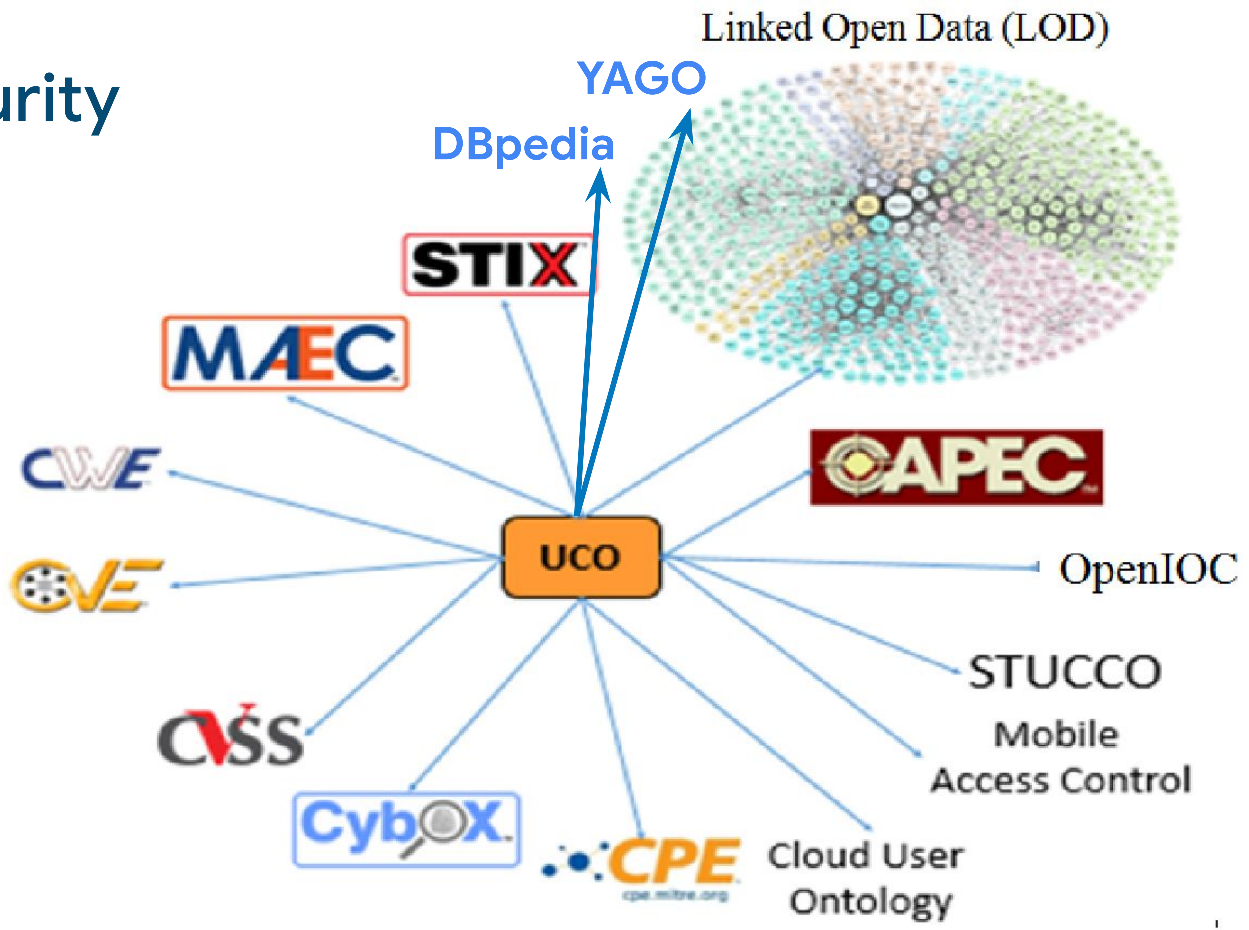
Other examples

DFAX - Digital Forensic Analysis eXpression

CVE - Cyber Intelligence Ontology

Ontologies can describe

- Technical
- Economic
- Behavioral
- Semantic



SOURCE UCO: A unified cybersecurity ontology ebiquity.umbc.edu/_file_directory_/papers/781.pdf

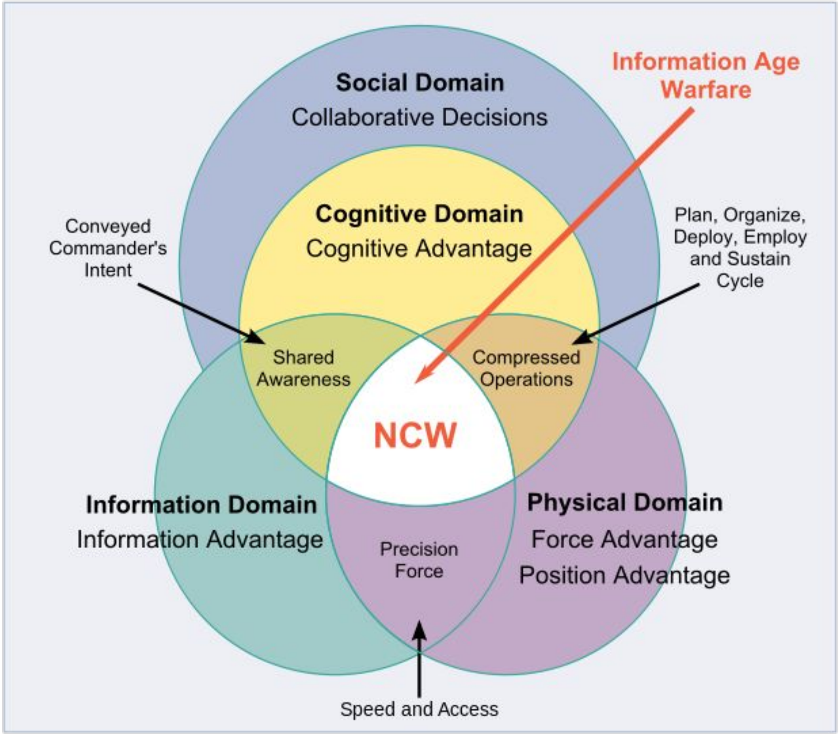
KG Use Cases

Monitoring / Understanding Complex Cyber Infrastructure

- Dr. Steven Noel et al. (MITRE)
  ‘DeCypher: NLP Interface for Cyber Situational Understanding from Graph Knowledge Bases’

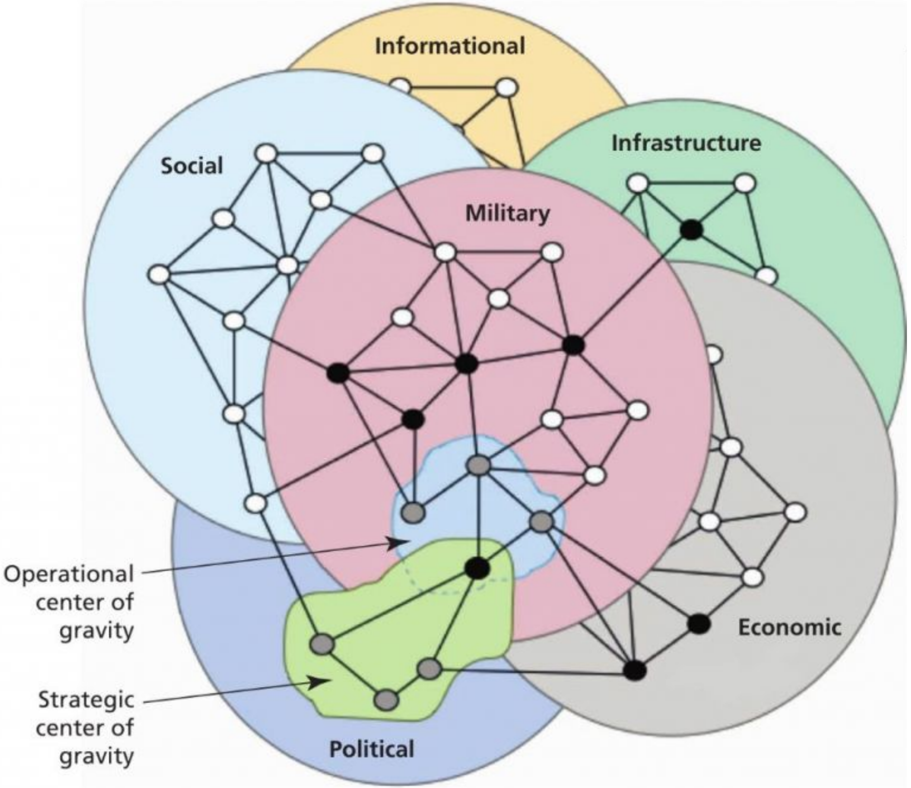
- BLOG **Neo4j** Manage and Monitor Complex Networks with Real-Time Insight

Security & defense...



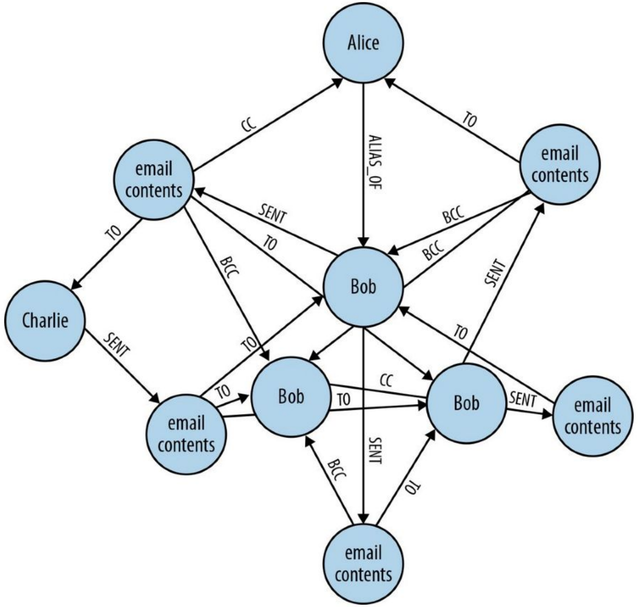
Source [The Implementation of Network-Centric Warfare (Office of Force Transformation, DoD, 2004](#)

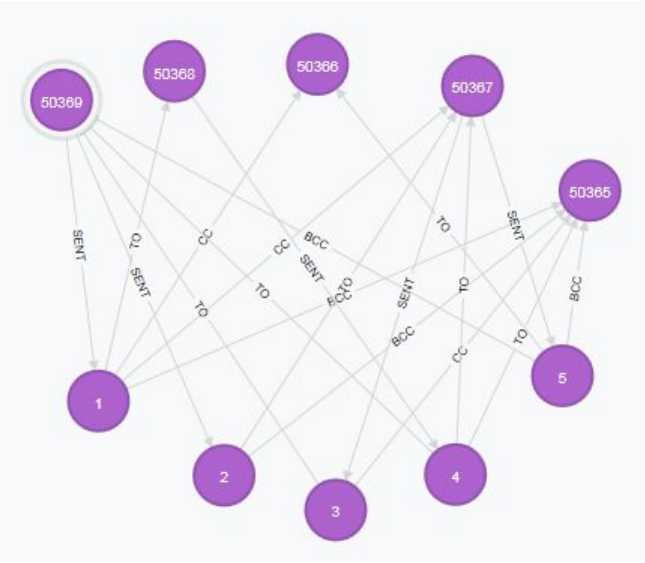
as a semantic graph challenge



Source [U.S. Joint Chiefs of Staff, 2011a, p. IV-5, Figure IV-2](#)
Ref [Virtual War-Information Supremacy on the Virtual Battlefield-Wavell Room](#)

Detecting patterns in suspect communications





- Suspicious communication patterns (e.g., use of emails, social media posts)
- Combine communication network analysis with sentiment scoring (i.e. use of negative terms) and/or semantic analysis (content analysis)
- Can combine other entities to enhance analysis (e.g., financial transactions, geolocation, org membership)

Misinformation monitoring

CONTEXT

- Domain entities
- Lexicon of key terms
- Codes & abbreviations
- Synonyms
- Sentiment

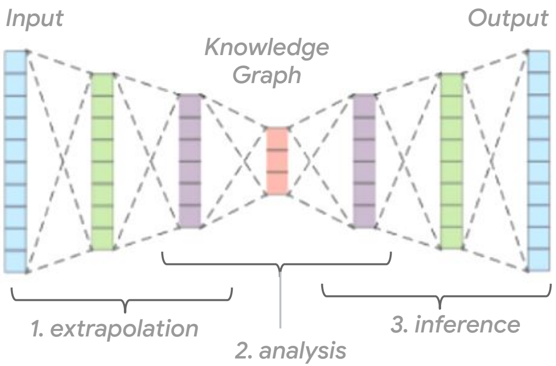
NEWS MEDIA

- Sources & reputation
- News feeds
- Region & industry
- Entity extraction
- Structured & unstructured

SOCIO-ECONOMIC CONTEXT

- Lexicon of key terms
- Entities and codes
- Links

KG Data Pipeline



AUTOMATION

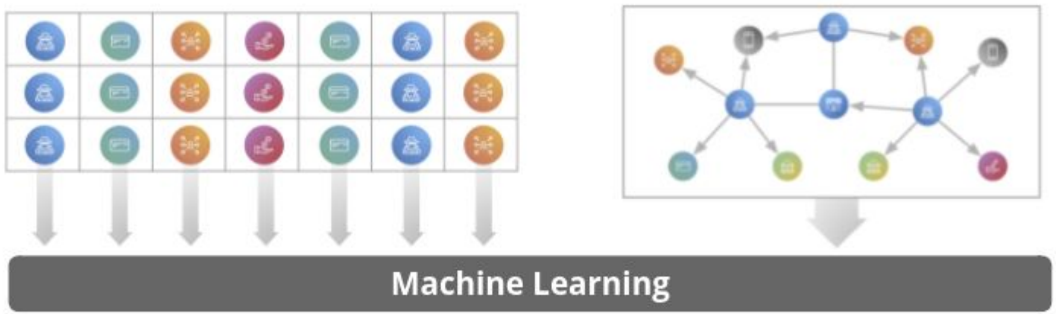
ANALYST

Machine learning

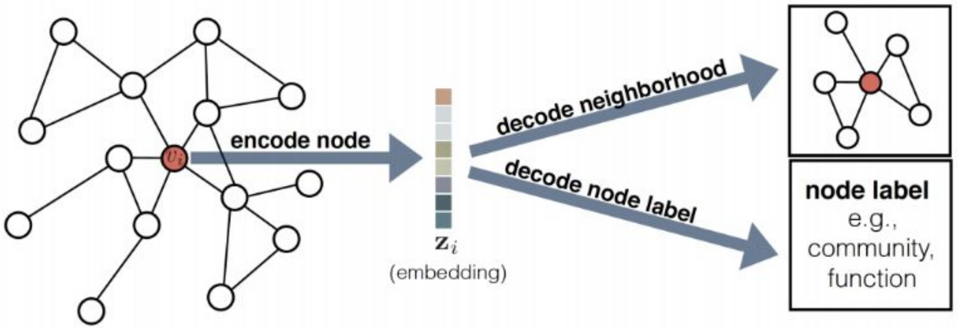
An enabler in both *populating to* and *inferring from* a knowledge graph

Graph Data as Machine Learning Features

- Clustering (unsupervised)
- Graph measures
- Identify statistically uncommon patterns (e.g. temporal sequences)



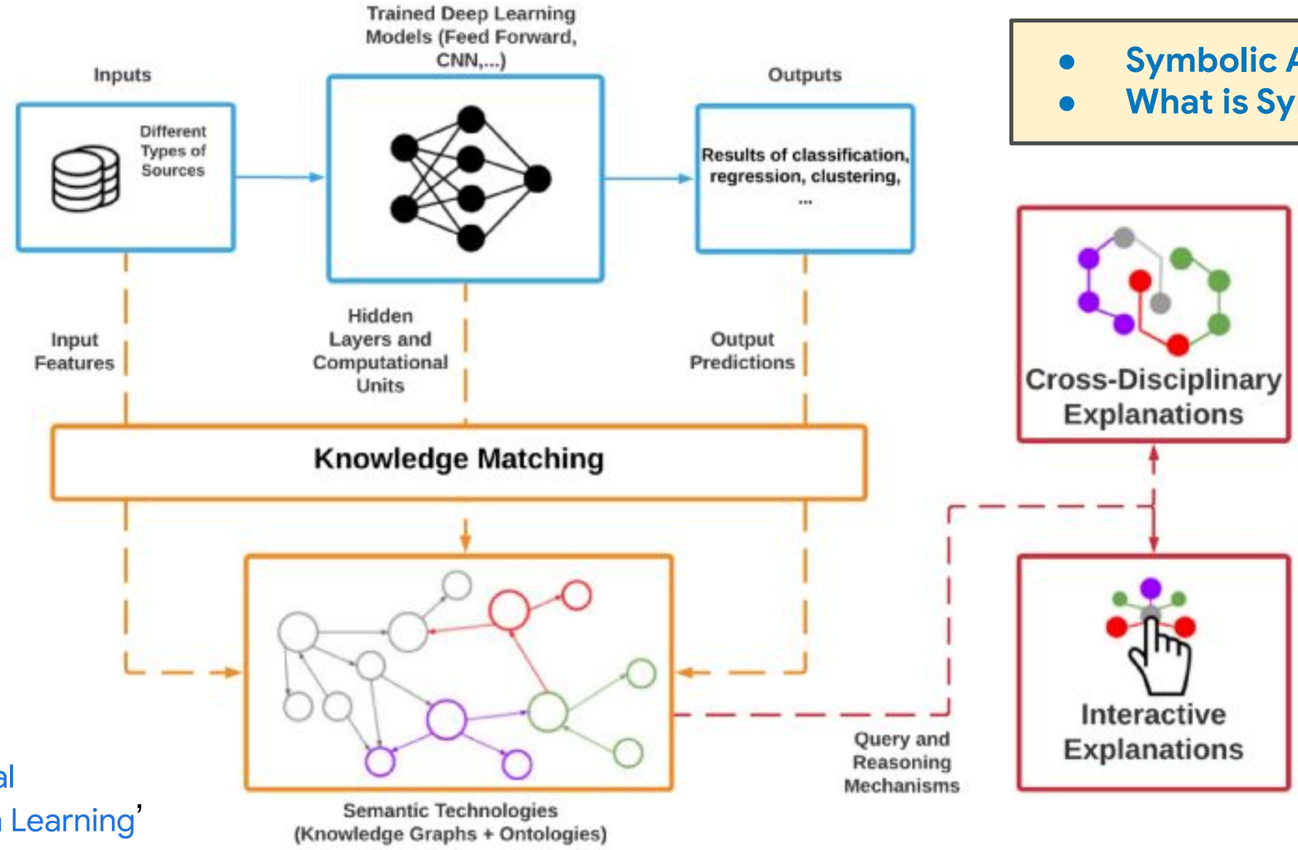
- Applied process mining
- Graph embeddings for ML
- Ontologies (see Bloehdorn, S. & Hotho, A. 'Ontologies for machine learning')



Knowledge Graphs for Explainable AI

Semantic context to improve ML results and model explainability

- Symbolic AI
- What is Symbolic AI?



'Towards Causal Representation Learning'

III Implementing KGs

Knowledge Graphs (KGs) three elements of success

Technology

KG as a step in a chain of technologies and tools

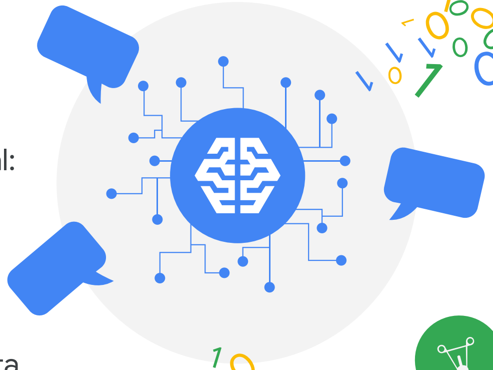
Process

KG as a step in a clearly defined functional process

People

Right resources with right skills are aligned to build, maintain, and utilize KGs

Knowledge Graphs (KGs) three elements of success



Technology

- Big data platform
- Data storage/retrieval:
  - Graph database
  - Triplestore
  - RMDB
- ETL
- ML/AI (NLP)
- Data mgmt / metadata repository

Process

- **Defining why you need a KG / how it will be used**
- Scoping knowledge domain
- Building a schema / ontology
- Populating initial KG
- Ongoing care & feeding (curation)

People

- Ontologist / information management specialist
- Data engineer
- Data scientist
- IT systems (security & integration)
- Domain specialist(s)

Steps for Building & Maintaining a Knowledge Graph

Iterative, non-linear process

1. Clarify business & data requirements
2. Gather and analyse relevant data
3. Clean data & ensure quality

I. Build corpus

4. Create semantic data model
5. Integrate data with ETL and/or virtualization
6. Harmonize: reconciliation, fusion, alignment (ML aided)

II. Structure & populate

7. Enable data access & search layer
8. Augment with reasoning, analytics, ML, NLP

III. Enable interfaces

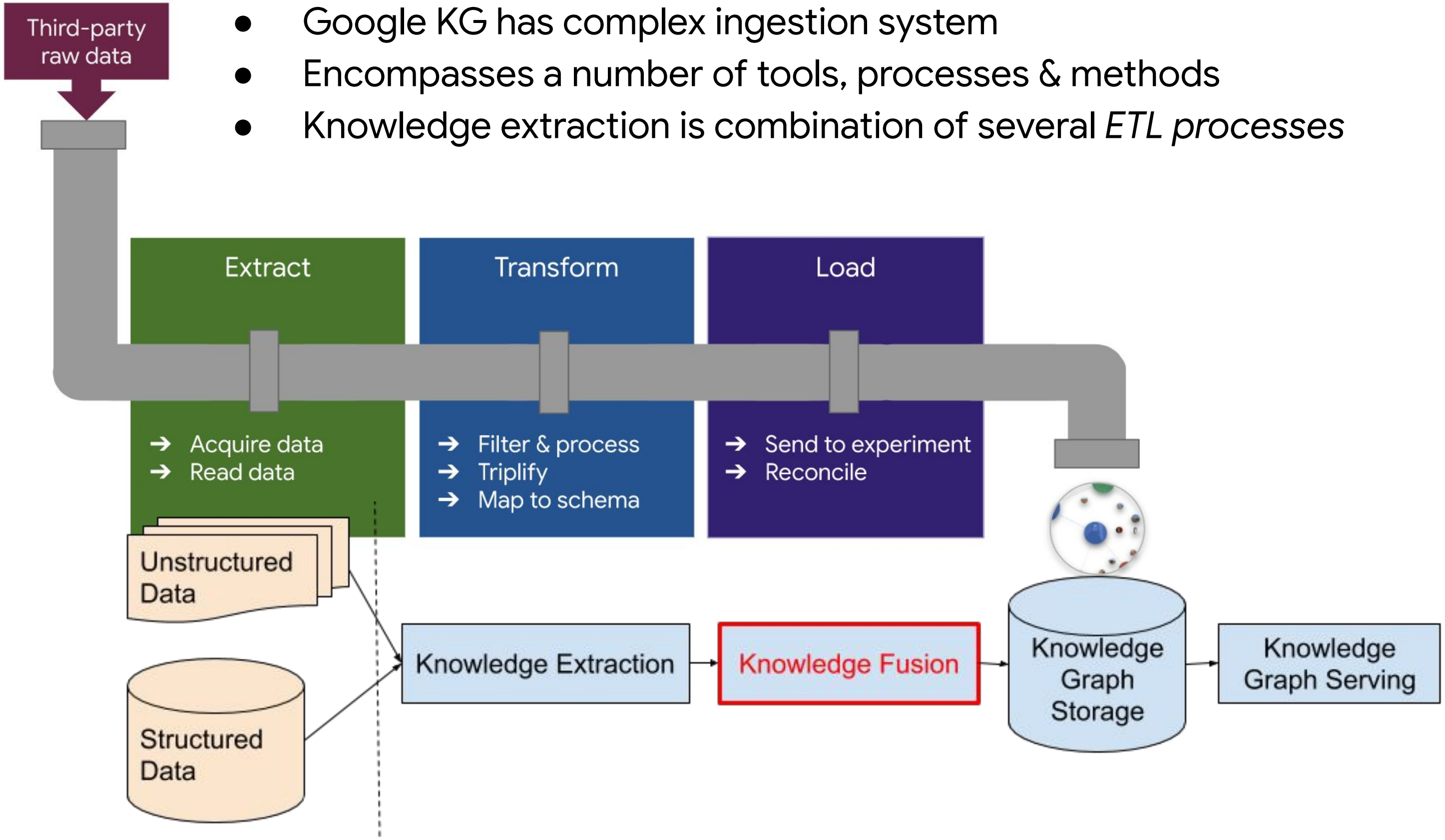
9. Semi-automate quality, maintenance, re-population
10. Enable tools for ease of human curation

IV. Maintenance / curation

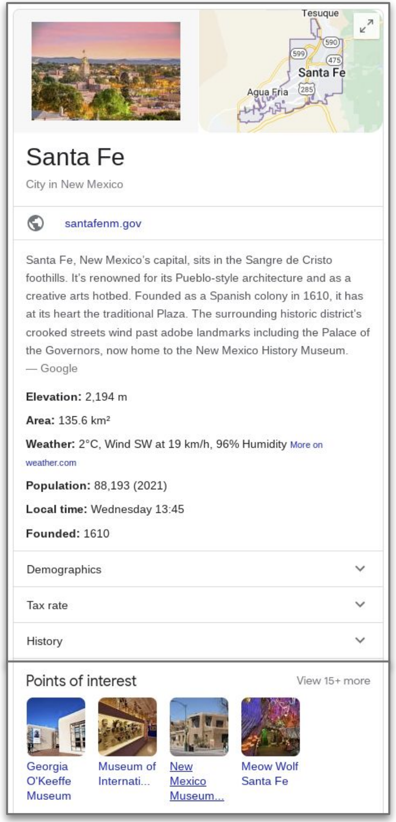
'Creatio ex nihilo'

Google KG - Knowledge Extraction

- Google KG has complex ingestion system
- Encompasses a number of tools, processes & methods
- Knowledge extraction is combination of several *ETL processes*

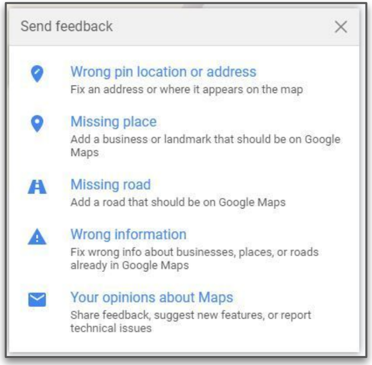


Feeding the Google Knowledge Graph

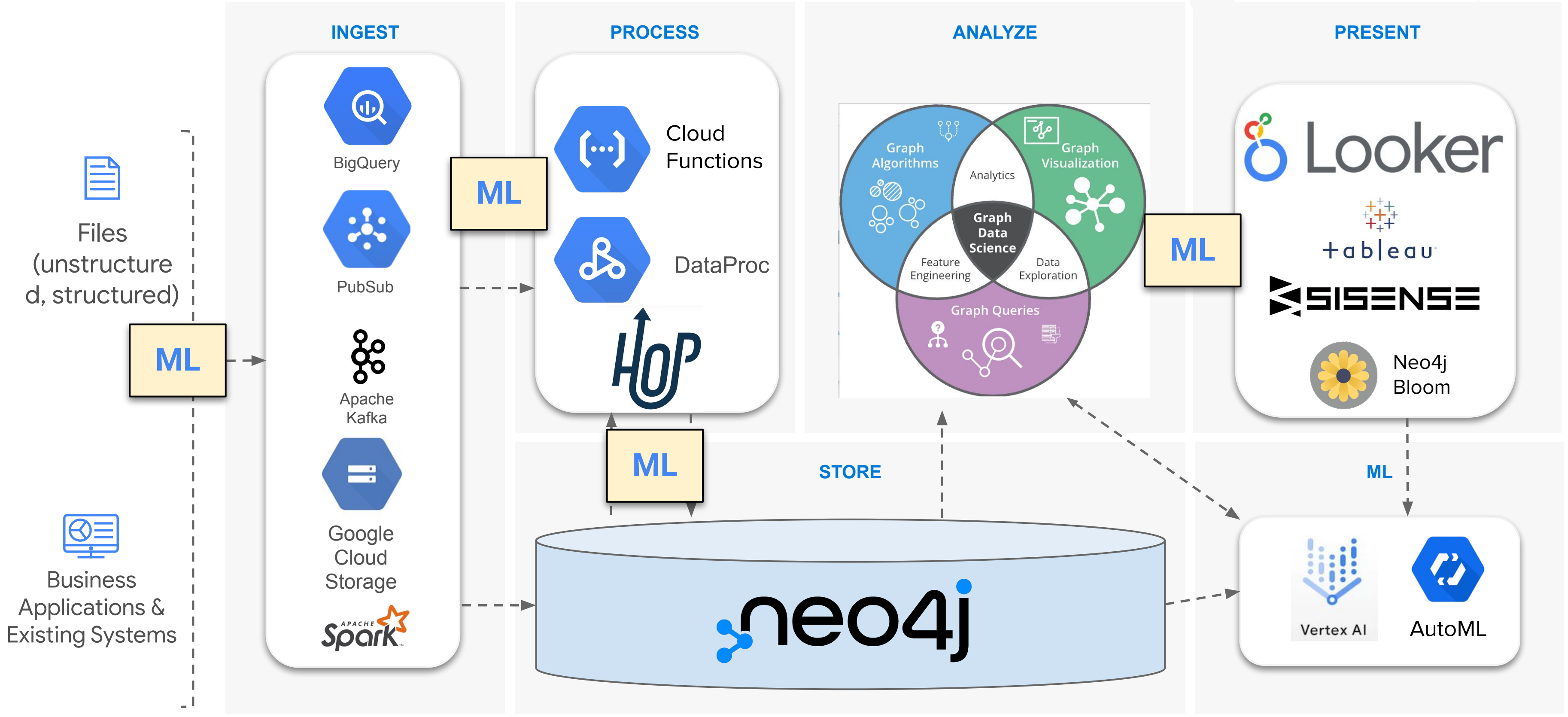


KG Data is created, compiled & maintained across 100s of types of sources

- Direct data (Ads, Maps, "Own this"...)
- Managed web data (Wikipedia...)
- Unmanaged web data
- Licensed data (IMDB, Stocks, Satellites...)
- Crowdsourced ("Suggest an edit")
- Humans (>1000 Employees / Contractors)
- User feedback and reviews

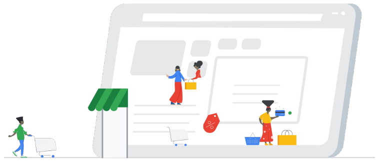


Graph Datastore Google Cloud implementation example



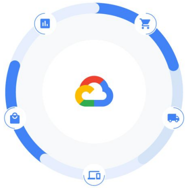
Best Practices Maintaining & Utilizing Knowledge Graphs

Aligning people, processes and technologies



AI/ML to realize efficiencies

Building & implementing is iterative



Form follows function (carefully define use)

IV Future Prospects

Autonomic Security Operations

BLOG Modernizing the SOC

Iman Ghanizada
Global Head of Autonomic Security Operations

Anton Chuvakin
Security Solution Strategy, Google Cloud

Autonomic Security Operations
A combination of philosophies, practices, and tools that improve an organization's ability to withstand security attacks through an adaptive, agile, and highly automated approach to threat management.

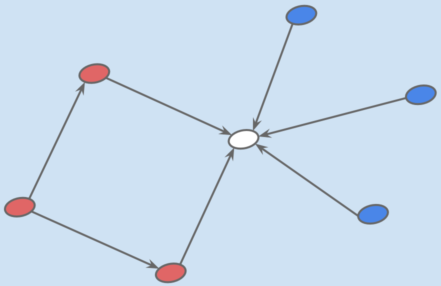
Graphs and Safety Spam, Fraud and Abuse Detection

By leveraging structure and relationships, graph-based learning allows for greater inferences to **prevent spam, fraud, and abuse** across all Google offerings: e.g. Ads, YouTube, Search, Play/Android, Payments, Cloud, etc.

Semi-supervised

e.g. label propagation, GNNs

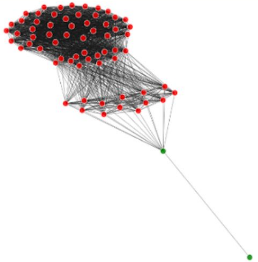
Starts with known bad actors, and use the graph structure to identify nearby neighbors that may also be suspicious.



Unsupervised

e.g. anomaly/trend detection, clustering

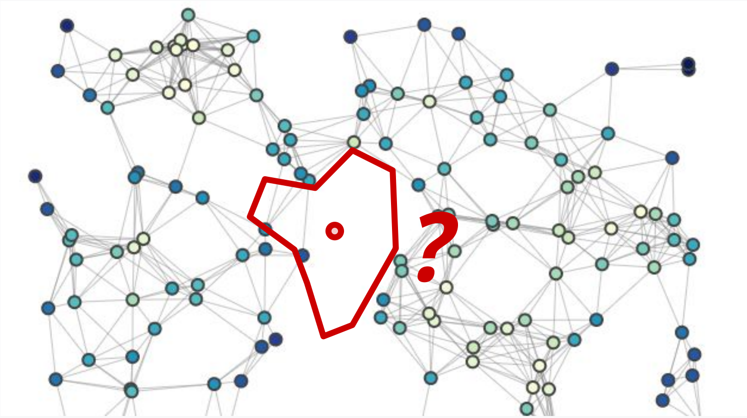
Statistically unlikely dense clusters and other structures correlate highly with malicious behavior.

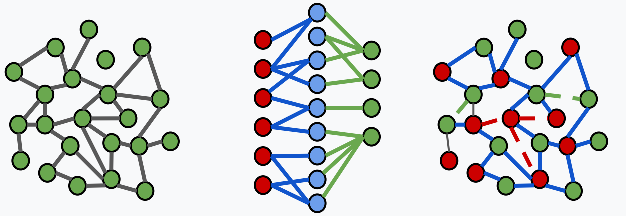


Graph Neural Networks (GNNs)

- A class of **neural network ML** for processing graph data
- Used for **node, link, subgraph,** or **whole graph classification** and **ranking**
- Can be used for **node, link,** or **whole graph prediction**

Distill A Gentle Introduction to Graph Neural Networks





Google AI Graph Mining team research

Blog GraphWorld graph benchmarking

- Provides insights into how GNN models perform on datasets with drastically different structure
- Can generate millions of different graphs, vary their properties, and benchmark models against them.
- Draw insights about the types of graphs that different models perform best on
- Example: are we overfitting on citation datasets?

Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion

www.cs.ubc.ca/~murphyk/Papers/kv-kdd14.pdf

Grale: Designing Networks for Graph Learning

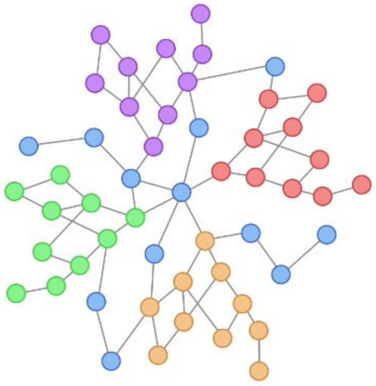
dl.acm.org/doi/pdf/10.1145/3394486.3403302

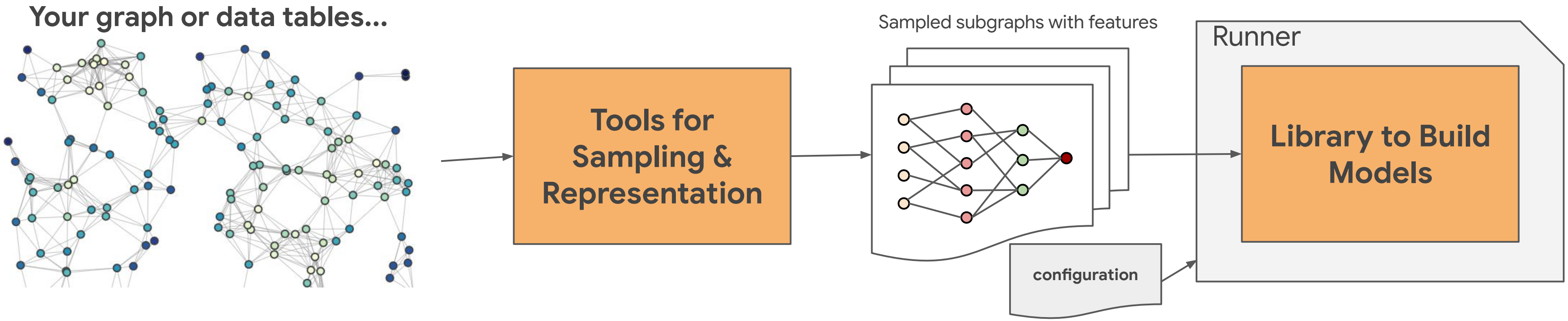
Graph Neural Networks with TensorFlow Graph

TensorFlow Graph Neural Networks library for working with graph data using TensorFlow (on GitHub)

- Port of Google GNN internal library
- Efficient graph manipulation functionality
- Descriptive schema to declare & validate topology
- Pooling operations
- Library of convolutions
- Keras-style API to create GNN models
- API interfaces to expose services to DevOps
- Can be used with other graph mining tools

BLOG Robust GNNs || **BLOG** TensorFlow GNNs





Large language Models (LLMs)

Supercharged NLP: feeding & care of KMs (BLOG)

- LLMs to populate and maintain ontologies
- LLMs as enriched by ontologies
- An aid to inference

Reasoning and agents

- Training cyber response / playbooks on LLMs
- Adversarial benefits: malicious code generation => countering
- Research Jure Leskovec Stanford University

Misinformation as a growing aspect of cyber

- Generating / detecting misinformation
- For instance, building a 'malicious generator' trained on misinformation to aid detection improve ML models

The race to understand the exhilarating, dangerous world of language AI
MIT Technology Review

...and in conclusion

- KGs tie to graph theory and ontologies
  - Well grounded, long-standing domains
  - Although only recently has tech caught-up
- KGs provide (structured) context
- KGs cross (and can connect) security use cases (e.g., monitoring, analytics, threat/risk assessment, compliance)
- Require an investment in people
- Fast-growing, particularly where KGs & ML overlap (i.e., GNNs, automated reasoning)

V Questions & Discussion

Thank you!

Cyber Ontologies context and cases

- Building an Ontology of Cyber Security http://ceur-ws.org/Vol-1304/STIDS2014_T08_OltramariEtAl.pdf
- Science of cybersecurity: Developing scientific foundations for the operational cybersecurity ecosystem
  - http://www.slideshare.net/shawnriley2/cscss-science-of-security-developing-scientific-foundations-for-the-operational-cybersecurity-ecosystem
- Ontological Representation of Networks for IDS in Cyber-Physical Systems:
  - http://rd.springer.com/chapter/10.1007/978-3-319-26123-2_40
- Mission Impact of Cyber Events: Scenarios and Ontology to Express the Relationships Between Cyber Assets, Missions and Users:
  - http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA517410
- Modeling Cyber-Physical Systems: https://www.researchgate.net/publication/220473317_Modeling_Cyber-Physical_Systems
- Ontological Approach toward Cybersecurity in Cloud Computing: https://arxiv.org/pdf/1405.6169.pdf
- The Essential Features of an Ontology for Cyberwarfare: http://www.crcnetbase.com/doi/abs/10.1201/b15253-7
- An insider threat indicator ontology: http://resources.sei.cmu.edu/asset_files/TechnicalReport/2016_005_001_454627.pdf
- Overview on cybersecurity semantic operationalization: http://www.slideshare.net/shawnriley2/cscss-science-of-security-developing-scientific-foundations-for-the-operational-cybersecurity-ecosystem
- Modeling cyber-physical systems: https://www.researchgate.net/publication/220473317_Modeling_Cyber-Physical_Systems

Cyber Ontologies technologies & implementation

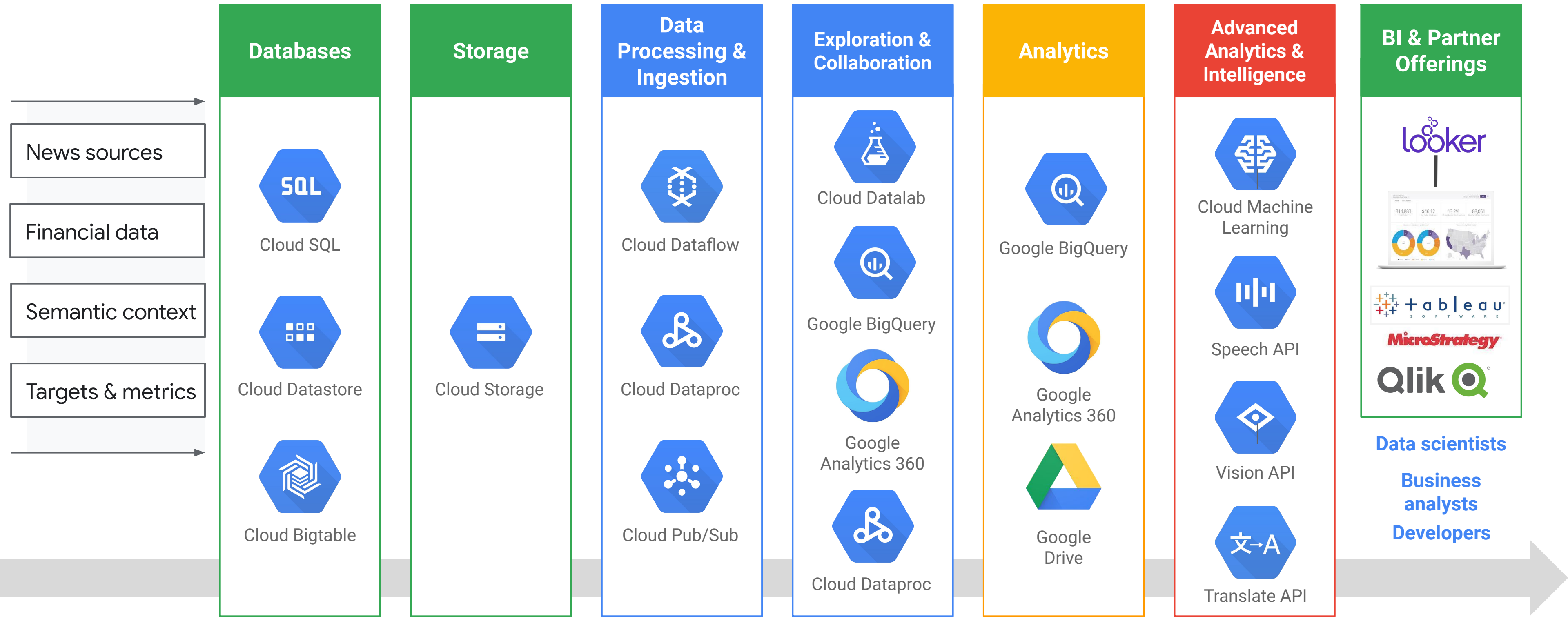
- Example storage technologies:
  - Apache Jena (RDF): https://jena.apache.org/
  - Apache Spark GraphX (graphs): http://spark.apache.org/graphx/
  - AllegroGraph: http://franz.com/agraph/allegrograph/
  - Neo4J (graph DB): https://neo4j.com/ - storing and querying RDF in Neo4J: http://www.snee.com/bobdc.blog/2014/01/storing-and-querying-rdf-in-ne.html
  - CumulusRDF: https://www.w3.org/2001/sw/wiki/CumulusRDF
  - NOTE: RDF / graph / triplestore databases are not mutually exclusive, but some graph DBs are not RDF compliant and some triplestores are less friendly to looser specifications and are storage and computationally demanding, so there are implementation and performance considerations for each approach:
- NOSQL Databases for RDF: An Empirical Evaluation: http://ribs.csres.utexas.edu/nosqlrdf/nosqlrdf_iswc2013.pdf
- Lengthily listing of triplestore DBs: https://www.w3.org/2001/sw/wiki/Category:Triple_Store
- Related blog post on RDF databases: http://blog.datagraph.org/2010/04/rdf-nosql-diff
- Research article evaluating performance of several implementations - ‘NOSQL Databases for RDF’ (2013): http://ribs.csres.utexas.edu/nosqlrdf/nosqlrdf_iswc2013.pdf

> Concerning querying / retrieval: SPARQL: RDF query language: https://en.wikipedia.org/wiki/SPARQL

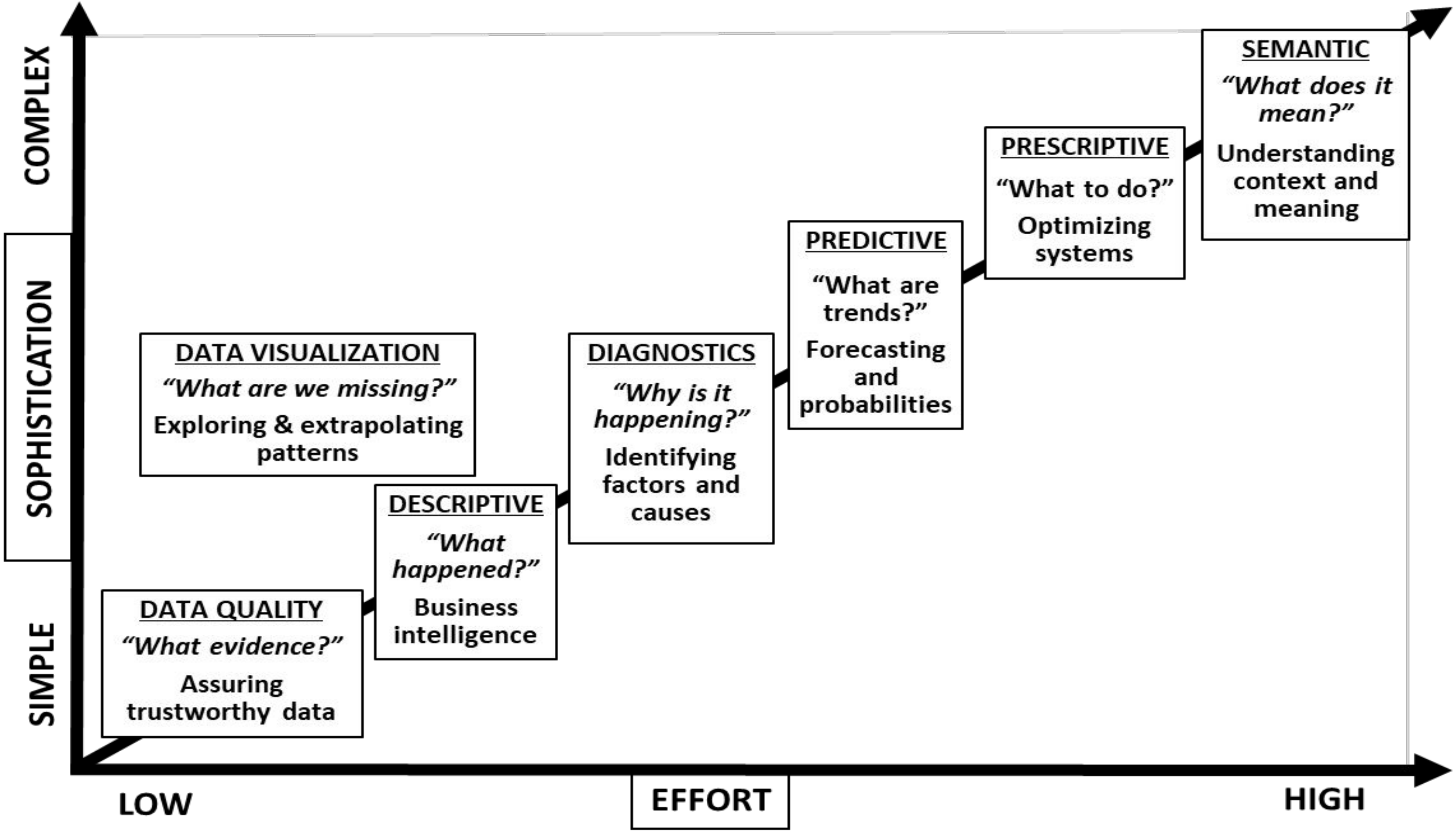
> Concerning structuring / maintaining / editing / managing ontologies:

- Cognitum FluentEditor: http://www.cognitum.eu/semantics/FluentEditor/
- Protégé: http://protege.stanford.edu/

Tools Google Cloud Data Platform



Analytics Methods Cumulative Value of Context



Ontology formalization

Semantic web standards for structuring ontologies

RDF is a framework for representing resources information in a graph

RDFS describes taxonomies of classes and properties and creates lightweight ontologies

OWL is an ontology language derived from description logics, offering more constructs over RDFS

RIF | SWRL provide rules beyond the constructs available from OWL

SHACL Shapes Constraint Language validates RDF graphs against a set of conditions