NECINA
2016 AI & Machine Learning Conference

Machine Learning with GraphDB

10/29/2016

Presented by: Albert Ma

Chief Innovation Officer, InSigma Hengtian

Agenda

- Background
- Challenges of Preparing Data for Analytics
- Graph DB – An Effective Alternative for Data Mining
- Use Cases
- Q & A

The AI Impact – Employment to Population Rate

"According to our estimate, 47 percent of total US employment is in the high risk category, meaning that associated occupations are potentially automatable "

CARL BENEDIKT FREY AND MICHAEL A. OSBORNE, Oxford Martin School & Faculty of Philosophy, UK

Image: Oxford Martin School & Faculty of Philosophy

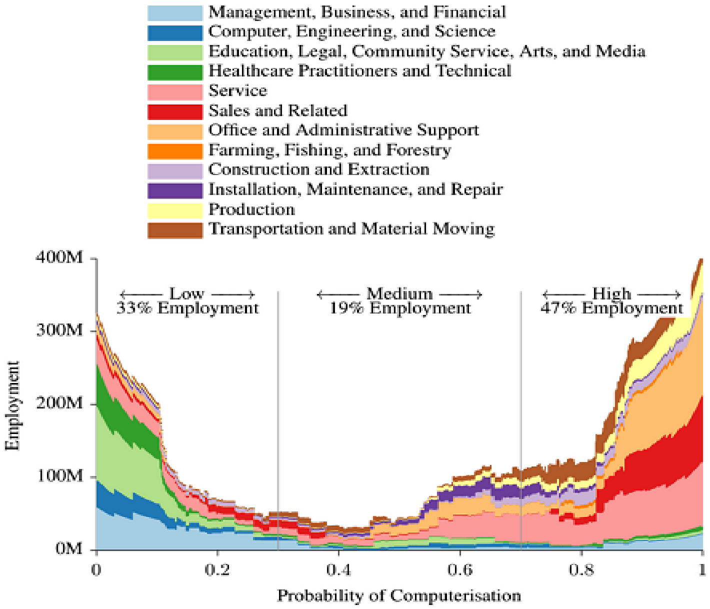


Figure 1. Employment Affected by Computerisation.

Source: http://www.techrepublic.com/article/ai-is-destroying-more-jobs-than-it-creates-what-it-means-and-how-we-can-stop-it/

Rise of China in AI

The Washington Post:- Brian Fung 10/13/2016
China has now eclipsed USA in AI research

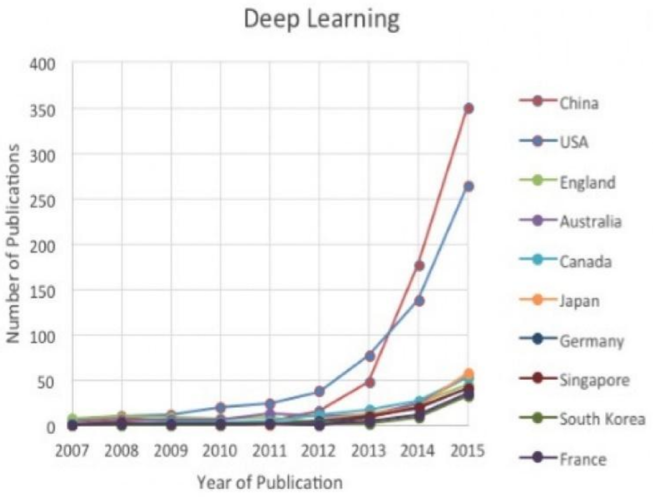


Figure 1: Journal articles mentioning "deep learning" or "deep neural network", by nation.[62]

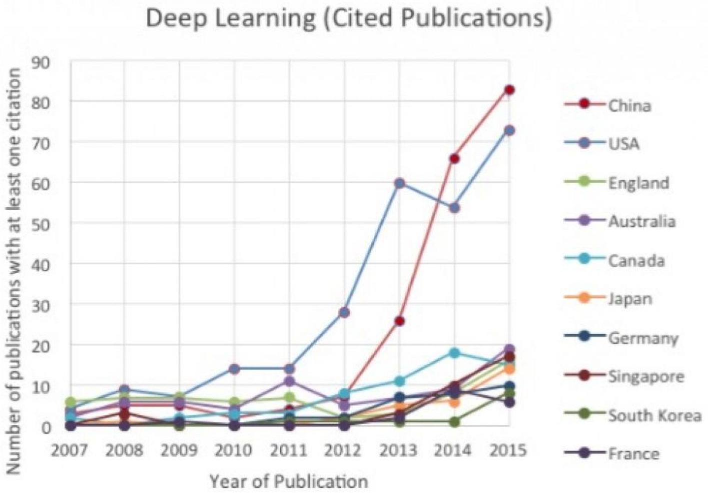


Figure 2: Journal articles cited at least once, mentioning "deep learning" or "deep neural network", by nation.[63]

Challenges of Preparing Data Sets

Researches indicated that 80% of effort were spent on preparing data with the remaining 20% on modeling

Key Challenges

1. Poor data quality
2. Don’ t understand the data deep enough
3. Factors like data variety and velocity prolong times in data preparation
4. Dealing with huge data sets
5. Don’ t know what business problems to solve

Why GraphDB ? Some Quotes:-

“Machine Learning is really the umbrella and graph technology is a way of representing data when using machine learning.” - *Claus Jepsen, Chief Architect, R&D of Unit4*

“Graph technology can be considered a type or technique of machine learning, or, at a minimum, aspects of graph technology have strong application to machine learning.” - *David Thompson, Sr. Director of Production Management of LightCyber*

“Automatically assembling a graph of connectivity between data points is a powerful addition to learning.” - *Simon Crosby, co-founder & CTO of Bromium*

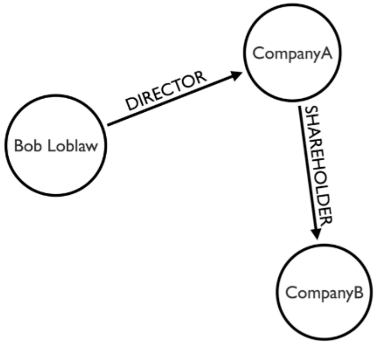
“Machine learning algorithms help data scientists discover meaning in data sets, and these insights can be expressed as relationships between nodes in a graph. Graph databases enable efficient storage and traversal of information about relationships. Therefore, graph data can either be the input or the output of machine learning processing.” - *Jim Webber, Chief Scientist of Neo Technology*

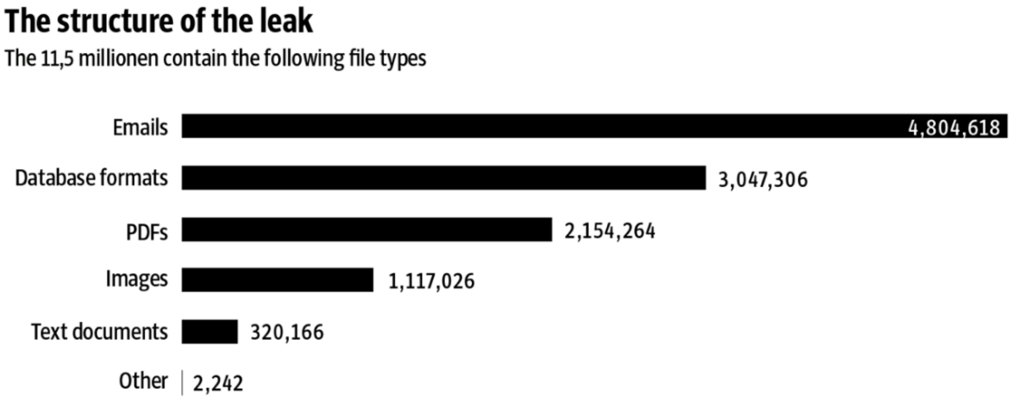
Sample Use Case: Fraud Detection

The Panama Papers - 11 Million records in over one terabyte data.

Steps:

1. Acquire documents
2. Classify documents
   a) Scan / OCR
   b) Extract document metadata
3. Whiteboard domain
   a) Determine entities and their relationships
   b) Determine potential entity and relationship properties
   c) Determine sources for those entities and their properties
4. Work out analyzers, rules, parsers and named entity recognition for documents
5. Parse and store document metadata and document and entity relationships
   a) Parse by author, named entities, dates, sources and classification
6. Infer entity relationships
7. Compute similarities, transitive cover and triangles
8. Analyze data using graph queries and visualizations

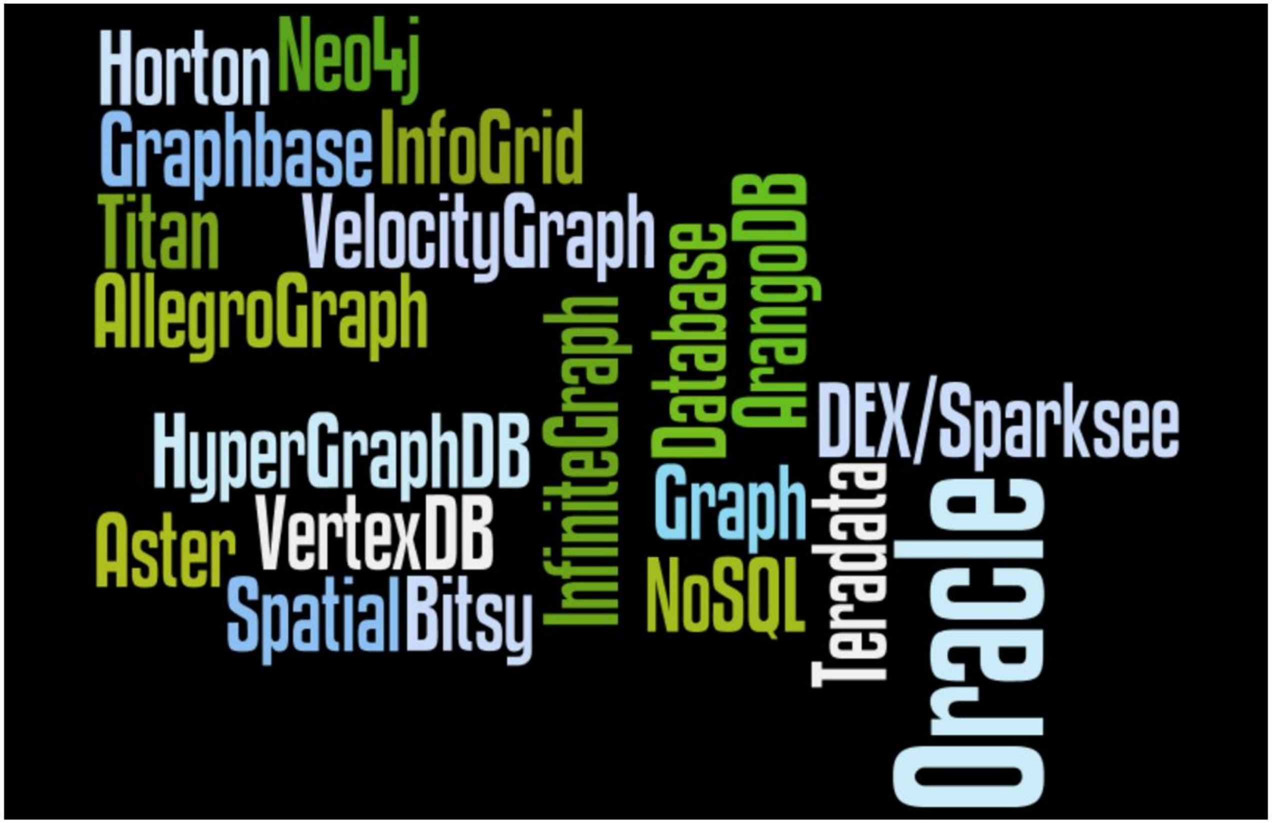




Source: https://neo4j.com/blog/analyzing-panama-papers-neo4j/

Top GraphDB Vendors

TOP 31 GRAPH DATABASES



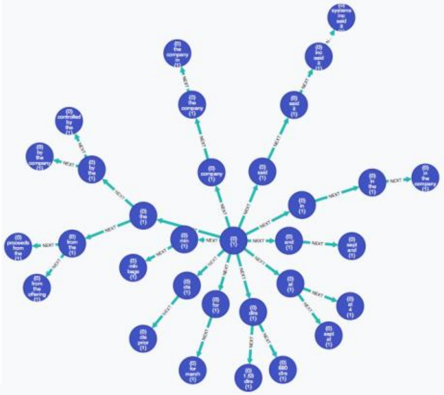
Source: http://www.predictiveanalyticstoday.com/top-graph-databases/

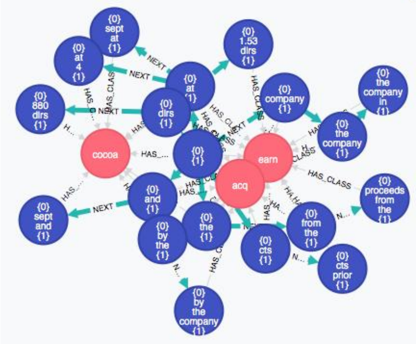
Use Case:- Using a Graph Database for Deep Learning Text Classification

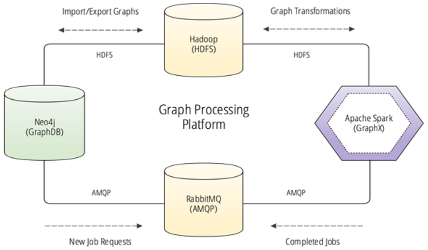
Graphify (Quotes):-

When training a model to recognize the meaning of a text, you can send an article of text with a provided set of labels that describe the nature of the text. Over time the natural language parsing model in Neo4j will grow to identify those features that optimally disambiguate a text to a set of classes.

This kind of deep learning doesn't require a neural network because of the nature of Neo4j's property graph data model, providing a way to generate a vector space model of extracted features and relate them to feature vectors by means of cosine similarity of the classes which are mapped to a subset of feature nodes within the hierarchy.





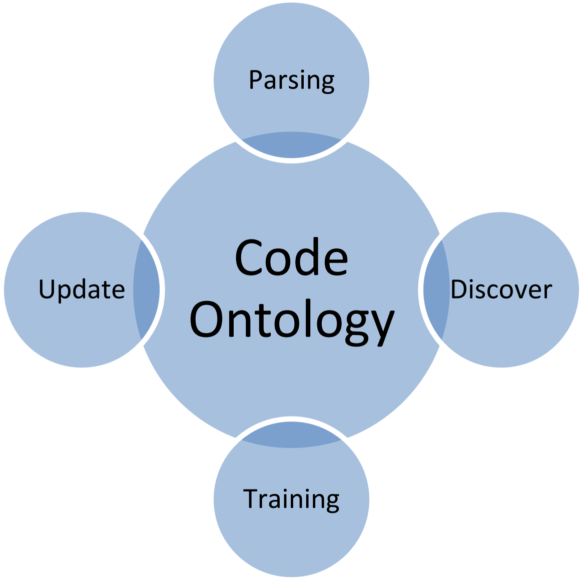


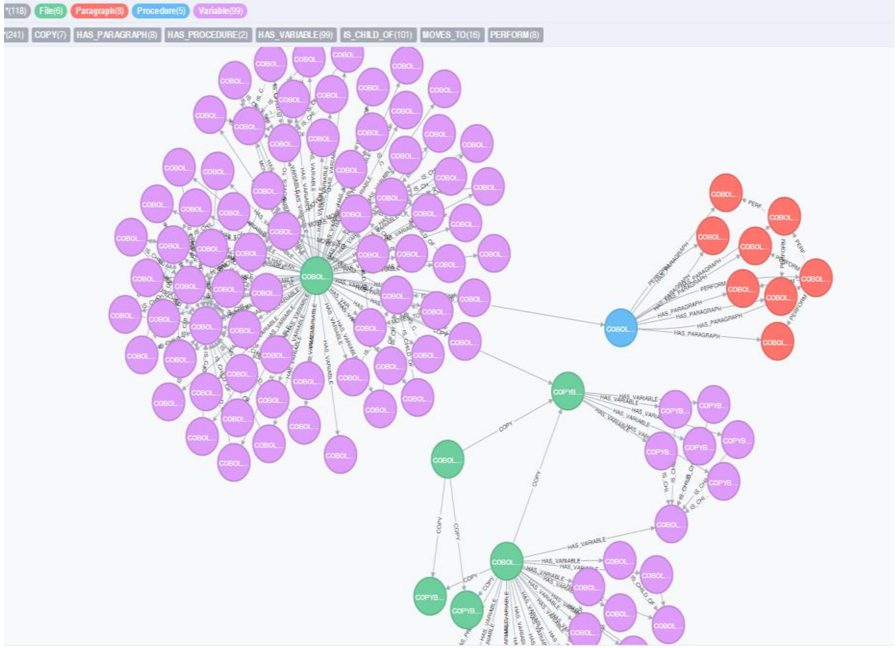
Source: http://www.kennybastani.com/2014/08/using-graph-database-for-deep-learning-text-classification.html

Use Case – Code Ontology with BlueMorpho

Code Signature (Vector)

Proc: Proc A -> Proc B -> Proc C
BizDef: Private Bank Account Opening
SysDef: PBAO Main
Loop: Proc A - > Loop A-> Condition (VarA < VarB)-> PassedValue (VarA)
Dataflow: VarA ->Proc A (???) -> Proc B (???); VarB -> ProcC(???) -> ProcA(???)





http://www.bluemorpho-tech.com

GraphDB – Pros and Cons

Pros	Cons
• Highly applicable for solving relationship-driven business problems, like fraud detection, money laundering, consumer behavior etc. • Strong visualization capability provides better insight to the data • Finding relationships amongst data is the easiest way to qualify data sets • Persist results in GraphDB provides high flexibility and extensibility of the ontological view • Most products support sophisticated SQL- liked scripting languages • Linkage to other data types with semantic RDF	• Not all business cases are applicable to use GraphDB (e.g. image recognition) • May take time and effort to design and discover the data relationships • Learning curve with a new programming language • Major machine / deep learning libraries are not natively working with GraphDB yet • Better integration to the Big Data infrastructure is expected

Reference

- http://www.idgconnect.com/abstract/18124/the-wave-disruption-graph-machine-learning, posted by Kathryn Cave on July 12, 2016
- http://www.idgconnect.com/abstract/15609/database-helped-reporters-follow-panama-papers-money
- http://info.salford-systems.com/blog/bid/305673/9-Data-Mining-Challenges-From-Data-Scientists-Like-You
- http://www.cs.uvm.edu/~icdm/10Problems/10Problems-06.pdf
- http://www.kdd.org/exploration_files/V13-02-09-Munson.pdf

Questions/Comments?