AUTOMATING KNOWLEDGE WORK WITH LARGE-SCALE KNOWLEDGE GRAPHS

2018 Strata Data Conference, New York

Mike Tung, Founder & CEO
What you’ll learn in this talk

● An architecture for future knowledge work
● What is a Knowledge Graph?
● A Brief History of Knowledge in AI
● Applications of Knowledge Graphs in AI
● The state-of-the-art in Knowledge Graph construction
Knowledge Graphs are coming

Knowledge Graphs have been identified as one of the top 5 emerging technologies that will impact business within the next 5-10 years.

Source: Gartner, Aug 2018
The future of knowledge work is human-AI symbiosis
Why do we need Knowledge in AI?
Exhibit A: “Intelligent” Assistants

Google Assistant

Siri

Assistants can’t answer questions without Knowledge

(Source: n=5000 questions, Stone Temple)
Exhibit B: Object Recognition

That’s not a Frisbee.

She isn’t holding a car.

YOLO is a state-of-the-art deep learning object detection system. (Source: Darknet)
Exhibit C: Product Recommendations

Buy a printer online...

Printer ads “follow” you online for days

Do I need another printer??
Exhibit D: Stock Trading

The Hathaway Effect
Anne Hathaway movie releases are correlated by 98% confidence to rises in Berkshire Hathaway

September 26, 2008 – Passengers opens: BRK.A up 1.43%
October 3, 2008 – Rachel Getting Married Opens: BRK.A up 0.44%
January 5, 2009 – Bride Wars opens: BRK.A up 2.61%
February 8, 2010 – Valentine’s Day opens: BRK.A up 1.01%
March 5, 2010 – Alice in Wonderland opens: BRK.A up 0.74%
November 24, 2010 – Love and Other Drugs opens: BRK.A up 1.62%
November 29, 2010 – Anne announced as co-host of the 83rd Academy Awards: BRK.A up 0.25%
February 28, 2011 – Anne co-hosts the 83rd Academy Awards: BRK.A up 2.94%
Today’s AI systems learn from data, but without knowledge, the results are unstable and non-intuitive.
Not all Bits are Created Equal

Data is a raw stream of symbols. Knowledge is a statement about the world.

DIKW Hierarchy

- Data
  - Raw
  - Fast, ephemeral, transactional
  - Noisy
  - Single-source

- Knowledge
  - Semantic
  - Slow
  - Clean
  - Synthesized over multiple sources
So what is a Knowledge Graph?

- It’s just a kind of database.
- That’s semantic (it stores knowledge).
- Often represented as a set of entities (nodes) and relationships (edges).
Here’s an example:

As a Graph

Mike Tung

- Works
  - Diffbot
- Speaking
  - Strata

- Lives in
  - Mountain View
- Education

- Education
  - Stanford

- Lives in
  - Mountain View
- Headquarters

As Triples

<table>
<thead>
<tr>
<th>Subject</th>
<th>Predicate</th>
<th>Object</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mike Tung</td>
<td>Works</td>
<td>Diffbot</td>
</tr>
<tr>
<td>Mike Tung</td>
<td>Education</td>
<td>Stanford</td>
</tr>
<tr>
<td>Mike Tung</td>
<td>Lives in</td>
<td>Mountain View</td>
</tr>
<tr>
<td>Mike Tung</td>
<td>Speaking</td>
<td>AIConf</td>
</tr>
<tr>
<td>Diffbot</td>
<td>HQ</td>
<td>Mountain View</td>
</tr>
</tbody>
</table>
Why isn’t Knowledge used more in today’s AI systems?
A History of Knowledge in AI

1980: Expert Systems
1990: Cyc
2000: Enterprise Databases
2010: Google Knowledge Graph
Knowledge is expensive to acquire

As each technology cycle reduces the cost of acquiring each fact by roughly 1000X, the size of the possible KG grows exponentially.

What is the next technical breakthrough?

Cost per Fact vs. Size of KG on a log scale
Application: Web Search

- For entity or fact seeking queries
- Summary of the entity/select facts
- Disambiguation
- Mainly “head” entities
Google Knowledge Graph

- Google acquired MetaWeb, a startup developing **Freebase**
- Freebase: Combined Wikipedia + a wiki-style crowd-sourced knowledge base.
- Total of 44M entities, 2.4B Facts
- After 2010 acquisition by Google, Freebase shutdown
- Wikimedia takes up crowd-sourced KG with WikiData project
- Wikipedia editors add ~20,000 new articles per month. ~123k active wikipedia editors

<table>
<thead>
<tr>
<th>Version</th>
<th>DBpedia</th>
<th>YAGO03</th>
<th>YAGO3</th>
<th>Wikidata</th>
<th>YAGO03</th>
<th>OpenCyc</th>
<th>NELL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2016-04</td>
<td>2016-08-01</td>
<td>2016-09-05</td>
<td>2016-09-05</td>
<td>2016-09-05</td>
<td></td>
<td></td>
</tr>
<tr>
<td># instances</td>
<td>5,109,890</td>
<td>5,130,031</td>
<td>17,581,152</td>
<td>118,125</td>
<td>1,974,297</td>
<td></td>
<td></td>
</tr>
<tr>
<td># axioms</td>
<td>397,831,457</td>
<td>1,435,808,056</td>
<td>1,633,309,138</td>
<td>2,413,894</td>
<td>3,402,971</td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. indegree</td>
<td>13.52</td>
<td>17.44</td>
<td>9.83</td>
<td>10.03</td>
<td>5.33</td>
<td></td>
<td></td>
</tr>
<tr>
<td>avg. outdegree</td>
<td>47.55</td>
<td>101.86</td>
<td>41.25</td>
<td>9.23</td>
<td>1.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td># classes</td>
<td>754</td>
<td>576,331</td>
<td>30,765</td>
<td>116,822</td>
<td>290</td>
<td></td>
<td></td>
</tr>
<tr>
<td># relations</td>
<td>3,555</td>
<td>93,659</td>
<td>11,053</td>
<td>165</td>
<td>1,334</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Ringler, 2017
Application: Recommendations

- Netflix moved from conventional similarity methods to knowledge-based recommendations
- Helps explain to user why a Movie was recommended.
- Builds trust in the system
Applications: In the Enterprise

- The large enterprise is a mini-Internet where each business function has its own database. Knowledge is treated as a core IP asset used for decision making.
- Significant human resource (studies indicate 20-30% of knowledge worker’s day) is spent entering and keeping these databases up to date [1].
- Transition from central ERP to SaaS/Cloud => even more fragmentation.

Source: McKinsey
Databases are Knowledge Worker management systems

- All databases have become machine learning problems.
- Automate decisions by predicting attributes of entities (people, accounts, products, inventory, content)

<table>
<thead>
<tr>
<th>Database</th>
<th>AI Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>Lead scoring</td>
</tr>
<tr>
<td>CRM</td>
<td>Churn prediction, credit risk</td>
</tr>
<tr>
<td>HR</td>
<td>Employee performance, sourcing, applicant scoring</td>
</tr>
<tr>
<td>BI</td>
<td>Anomaly detection, Fraud detection, Claims</td>
</tr>
<tr>
<td>Marketing</td>
<td>Smart segmentation, pricing, content personalization, ad buying</td>
</tr>
<tr>
<td>Supply Chain</td>
<td>Inventory forecasting, demand forecast</td>
</tr>
</tbody>
</table>
Application: Text Analysis

KGs can be used to disambiguate meanings of words.

Diffbot Technology resolving entities in a sentence.
We can also resolve the relationships between these entities.

This is a Triple! (subject, object, predicate)

This is a very special application: We can generate Knowledge from documents

Diffbot technology: Relation Extraction
The Next 1000X Leap: Automated Knowledge Base Construction
We can apply AI to generating Knowledge

Diffbot formed as a AI research startup to solve this problem of automated knowledge acquisition

Combining multiple AI disciplines to the task of extracting knowledge from documents:

**Visual layout analysis and Classification**
We render pages in a virtual browser and determine the type of page: article, person, org, image, etc..

**Natural language processing**
We apply multi-lingual NLP to understand the text on the page, the entities, facts, and relations

**Computer Vision**
We analyze the images and videos on the page to determine their content and facts

**Knowledge Fusion**
We fuse facts from records extracted from multiple pages, creating a more accurate and complete view of entities
The Diffbot Knowledge Graph

- We can apply these algorithms to every page on the public web (~50B documents) and build a universal Knowledge Graph that contains all public knowledge.
- Currently adding ~120M entities / month
State of the Art in AKBC

- Linking and fusing the facts extracted from multiple pages
- Estimating the probability of truth of each fact

Diffbot: linked extracted records for George W. Bush
Impacts of Automated Knowledge Acquisition

- Automated Knowledge base construction techniques from "raw" data sources means people will spend less time gathering data.
- Humans focus on analyzing the results and coming up with better questions to ask and new ideas for sources.
- Massive gains in productivity and empowerment.
The future of knowledge work is a human-AI symbiosis.
What the AI system does

The AI system:
- Process inbound inquiries and enhance data using KGs
- Search for new knowledge outside the organization
- Classify and reason, using all available knowledge, how to best handle this case
- Execute the appropriate response
What the human does

The human worker
- No longer spends any time gathering information
- Is out of all high-bandwidth information flows
- Analyzes the output of the AI, offering feedback when necessary
- Specifies how to get information as requirements change
Example: Sales Development

- Inbound lead signs up on website
- Information provided about person, organization, role are enhanced using KG
- ML classifies the enhanced lead (score, use case)
- Personalized response sent back to lead
- Human sales rep specifies qualities of ideal customers (“CIOs at manufacturing companies with 100-200 employees, based in Europe”)
- Query Engine finds all Persons that match criteria and enhance facts with KG
- Personalized outreach message sent to prospect
Example: Bookkeeping

- Each month, transactions such as purchases, sales, receipts, and payments come in
- The KG identifies for each purchase or sale the vendor (company entity) in the KG, the good or service that was purchased (product entity)
- AI classifies the category of the expense or revenue and records it to the accounting system
- KG automatically updates the accounting system with any changes to Vendors (billing contact info, corporate status, name changes)
- System could search the web for cheaper vendors of purchased products
How it works together

The future of knowledge work is a human-AI symbiosis.

The AI system:
- Process inbound inquiries and enhance data using KGs
- Search for new knowledge outside the organization
- Classify and reason, using all available knowledge, how to best handle this case
- Execute the appropriate response

The human worker:
- No longer spends any time gathering information
- Is out of all high-bandwidth information flows
- Analyzes the output of the AI, offering feedback when necessary
- Specifies how to get information as requirements change
Future Directions of KGs in AI

- Wisdom Graphs (generalizability, causality)
- Computational Law, Medicine, Scientific Discovery (deeper)
- Macro-economic, global forecasting (wider)
- One graph capturing knowledge of all cultures (multi-lingual)
- A Universal factoid question Answerer, can answer about any fact that is observable or predictable from data.
Thank you