Building a Knwoledge Grph Using Meszy Real EsTate Data

John Maiden
Senior Data Scientist
Cherre
Data Council NYC 2019

Our mission is to transform real estate investing and underwriting into a science

Make Better Decisions with a Unified Single Source-Of-Truth

Paid

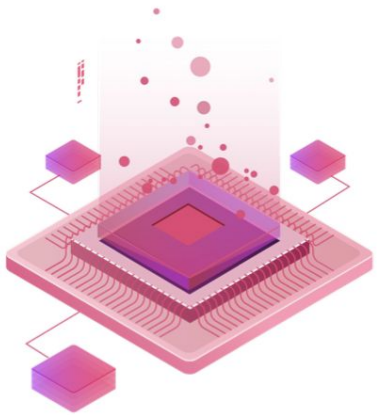
Connect all existing and future paid data feeds

Internal

Connect all internal CRM, reporting and research data

Public

Connect all available public data from any source



Immediate Results
Shave years off business and product roadmaps to generate actionable insights.

Built for Scale
Designed for real-time and end-to-end delivery of high-performance data solutions.

Unparalleled Quality
AI entity resolution and schema mapping provide industry leading source of truth.

Connect all your internal and external data from any source to power your most demanding data architecture needs.

We built the largest real estate knowledge graph in the world, built with unprecedented access to proprietary data

300+ Thousand
Unique Data Sets

2+ Billion
Unified Data Points

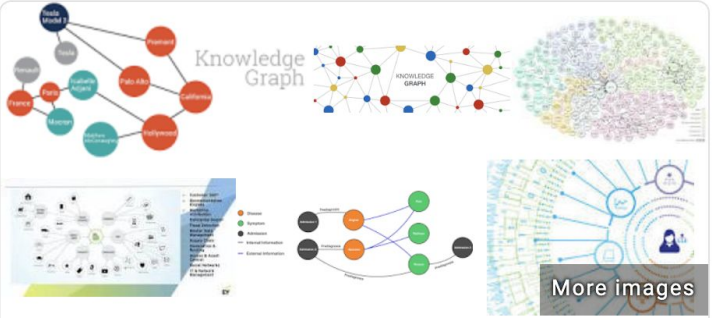
177+ Million
Real Estate Properties

84+ Million
Registered Corporations

Knowledge Graph

What Is A Knowledge Graph?

Google Search #1:



Knowledge Graph

The Knowledge Graph is a knowledge base used by Google and its services to enhance its search engine's results with information gathered from a variety of sources. The information is presented to users in an infobox next to the search results. Wikipedia

cherre

What Is A Knowledge Graph?

Google Search #2:

In computer science and information science, an **ontology** encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many or all domains of discourse.

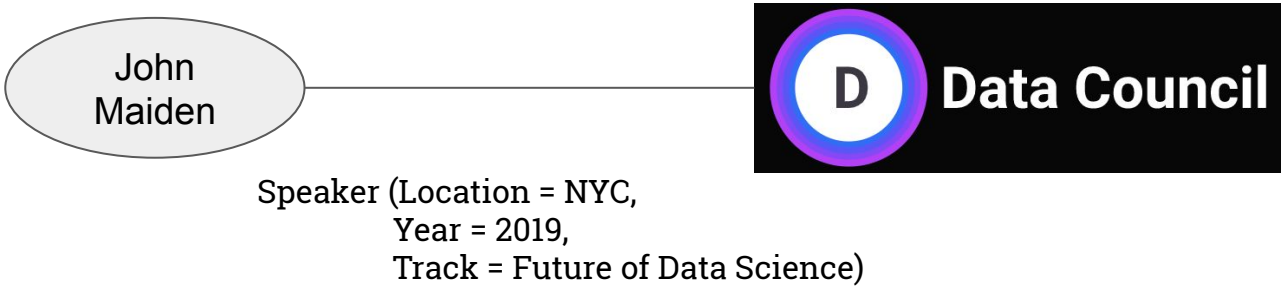
Every field creates ontologies to limit complexity and organize information into data and knowledge. As new ontologies are made, their use hopefully improves problem solving within that domain. Translating research papers within every field is a problem made easier when experts from different countries maintain a controlled vocabulary of jargon between each of their languages.[1]

"Ontology (information science)", Wikipedia, Retrieved October 26, 2019

cherre

Um, So What Is A Knowledge Graph?

It is a **graph** (compared to a knowledge base)



- Easier to visualize
- Relationships are a core component and can be analyzed / measured
- Straightforward to add new connections
- Traversable

"WTF Is a Knowledge Graph", Hackernoon, Retrieved October 26, 2019

cherre

What Questions Do We Want To Answer?

We want to use commercial real estate (CRE) data to answer questions like:

- Who is the property's true owner?
- Which properties has this owner bought and sold in the past five years?
- Which lenders are seeing larger than average number of defaults?

What Questions Do We Want To Answer?

We want to use commercial real estate (CRE) data to answer questions like:

- Who is the property's true owner?
- Which properties has this owner bought and sold in the past five years?
- Which lenders are seeing larger than average number of defaults?

And eventually we want…

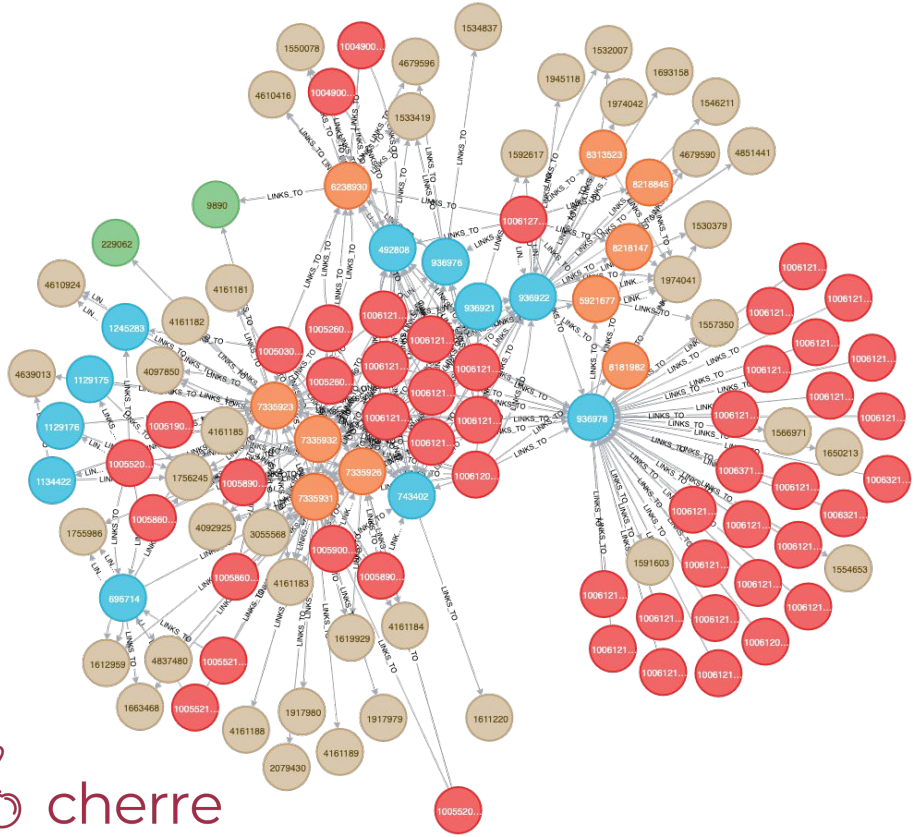
- Owner strategy - what types of properties do they buy?
- Models built from graph data (Comps, Valuation)

cherre

What Can We Do With A Knowledge Graph?

What It Looks Like

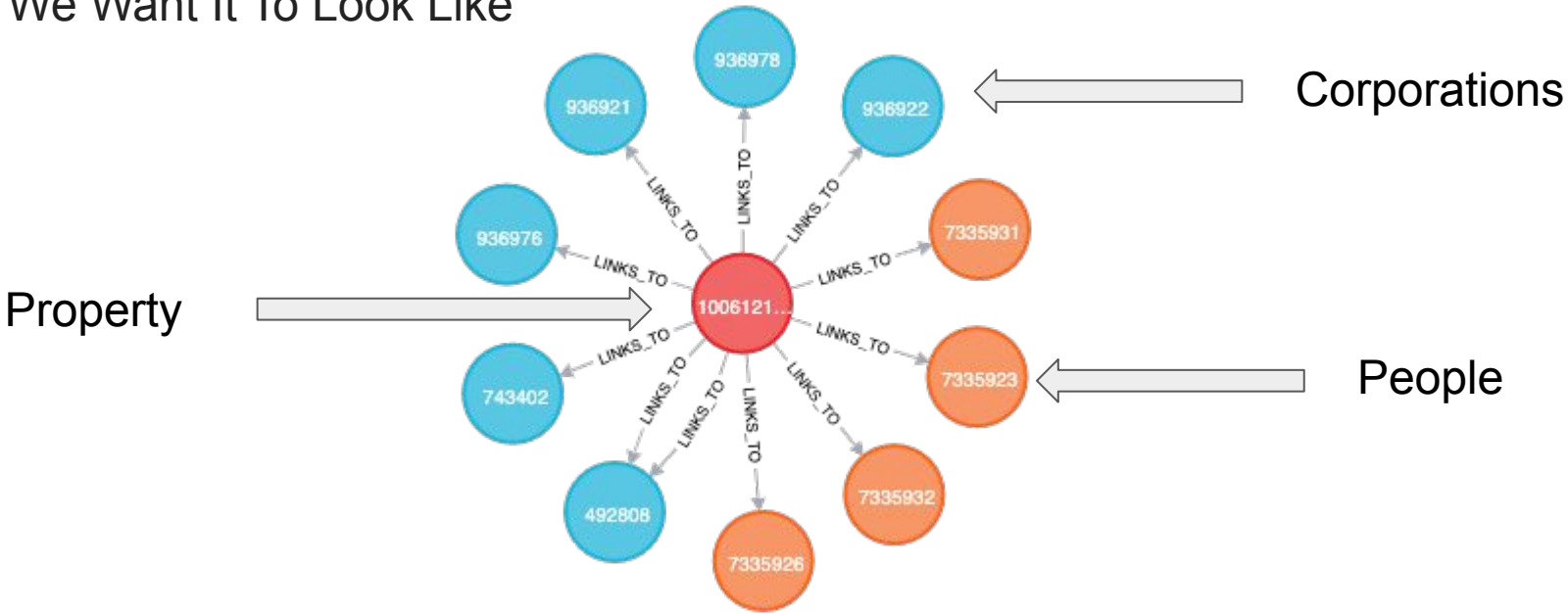
- The NYC Graph alone has millions of edges and nodes!
- Nodes can be properties, people, corporations, or contact info.



cherre

What Can We Do With A Knowledge Graph?

What We Want It To Look Like



cherre

Data

What Goes Into A CRE Knowledge Graph?

https://az505806.vo.msecnd.net/cms/c31664b3-62ce-4b99-9414-de5f8130b27d/545a09fc-d0ba-48da-8237-3be6275eccc9.jpg

cherre

What Goes Into A CRE Knowledge Graph?

Sold to ABC Corp by DEF Corp on 1/23/12

Mortgage lender is Tenth National Bank

Assessed taxes of $145k USD paid on 4/18/19 by 123 Main St LLC

Listed contact phone number on building permit as (111) 111-1111

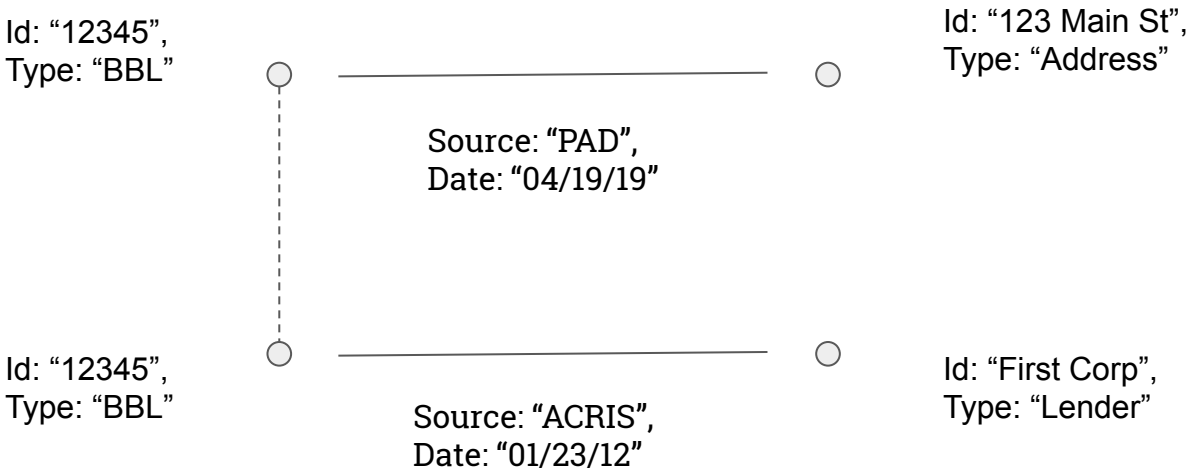
Owned by NYC Dept of Transportation

https://az505806.vo.msecnd.net/cms/c31664b3-62ce-4b99-9414-de5f8130b27d/545a09fc-d0ba-48da-8237-3be6275eccc9.jpg

cherre

NYC Open Data Sources

Translating This To A Graph (NYC)



cherre

Standardization

How Do We Join The Data?

We have three different types of fuzzy join keys:

- *People*
  - “John Maiden” vs “Maiden, John W” vs “The Trust of JW Maiden”
- *Corporations*
  - “Main St LLC” vs “Main Street Advisors LLC”
- *Addresses*
  - “989 6th Ave” vs “989 Sixthe Ave” vs “989 Ave of Americas”

cherre

People / Corporation Standardization

- Names come in multiple formats
  - “John W Maiden” vs “Maiden, J” -> Person

- Categorization is important
  - “The Irrevocable Trust of John Maiden” -> “John Maiden” -> Person
  - “John Maiden LLC” -> Corporation
  - “John King” -> Person, “Burger King” -> Corporation
  - “Grant Herreman” vs “Grant Herrman” vs “GHSK” vs “Grant Herrman Schwartz & Klinger” -> Corporation / Lawyer / Service Provider

- Common Names
  - “John Smith”

cherre

People / Corporation Standardization

How Do We Solve This?

- Regex (re.sub(r ".*TRUST.*", "", …))
- NLP-based classification models (e.g. ngrams + XGBoost)
- Graph + Fuzzy Matching (word1, word2, fuzzy score = 89)
- Good Reference Data

cherre

Address Standardization

- Abbreviations / Alternate Names
  - “989 W 6th Ave” vs “989 West Sixth Avenue” vs “989 Avenue of the Americas”
- Spelling Variations
  - “Gouverneur St” vs “Governor St”
- Obvious Typos / Sticky Components
  - “989 6th St, NYC, NJ”, “123 MAIN STUNIT 7C”
- Embedded Addresses
  - “% John Maiden, 989 6th Ave, NYC, NY”

cherre

Address Standardization

How Do We Solve This?

- *Parse*
- *Standardize*
- *Match*

Address Standardization - Parse

A parser takes an input string and identifies it with its lexical information.

"989 6TH AVE, FL 17, NYC, NY 10018"

Word Tokenization (NLTK)

[('989', 'CD'), ('6TH', 'CD'), ('AVE', 'NNP'), (',', ','), ('FL', 'NNP'), ('17', 'CD'), (',', ','), ('NYC', 'NNP'), (',', ','), ('NY', 'NNP'), ('10018', 'CD')]

Address Tokenization (Cherre)

[('989', 'AddressNumber'), ('6TH', 'StreetName'), ('AVE,', 'StreetNamePostType'), ('FL', 'OccupancyType'), ('17,', 'OccupancyIdentifier'), ('NYC,', 'PlaceName'), ('NY', 'StateName'), ('10018', 'ZipCode')]

cherre

Address Standardization - Standardize

Standarize takes the parsed components and cleans / formats.

Input	989	6TH	AVE,	FL	17,	NYC,	NY	10018
Output	989	*SIXTH*	*AVENUE*	*FLOOR*	17	*NEW YORK*	NY	10018

cherre

Address Standardization - Match

Match takes the cleaned address and matches against an address database.

- SQL Join
  - “123 MAIN STREET, NEW YORK, NY 10001” -> “123 MAIN STREET, NEW YORK, NY 10001”
- SQL Join w/ Business Logic
  - “123 MAIN STREET **APT** 6C, NEW YORK, NY 10001” -> “123 MAIN STREET **SUITE** 6C, NEW YORK, NY 10001”
- Fuzzy Join
  - “**124** MAIN **AVENUE**, NEW YORK, NY, 10001” -> “**123** MAIN **STREET**, NEW YORK, NY 10001”

cherre

Address Standardization - Technology

- *Parse*
  - Regex 😒, Hidden Markov Models, Conditional Random Fields, Neural Network
- *Standardize*
  - Regex, Lookup Tables
- *Match*
  - SQL Join, User Defined Aggregation Functions, Fuzzy Join (e.g. Hashing)

cherre

Standardization - Lessons Learned

- Business Knowledge / Context is Critical
  - Understand your data!
  - Humans are useful!

- Learn to Deal with Scale
  - Standardizing millions of addresses

- Live with Ambiguity 🤷‍♀️

cherre

On the top of the mountain we are all snow leopards.

- **Hunter S. Thompson**