Generations of Knowledge Graphs: The Crazy Ideas and The Business

Xin Luna Dong

1/2023

This talk does not represent the company’s point of view
Theme I. Three Generations of Knowledge Graphs

1. Entity-Based KGs
2. Text-Rich KGs
3. Media-Rich Time-Based KGs

Google Generic KGs
Amazon Product KGs
Personal KGs
Theme II. The Recipe from Innovation to Practice

Feasibility
A prototype, an experiment, to show it is possible

Quality
Production quality, E2E MVP experiences

Repeatability
E2E pipelines, automations to allow for extensions to more domains

Scalability
Significant cost reduction for scale ups

Ubiquity
High coverage, long-tail use cases, assumption removal, to next cycle of inventions
From Roofshots to Moonshots

Feasibility | Quality | Repeatability | Scalability | Ubiquity
---|---|---|---|---

Roofshots | Moonshots
Generation #1: Entity-Based Knowledge Graphs
Entity-Based KG Example

“Shake it off”
- Type: Song
- Artist: Taylor Alison Swift
- Genre: Country pop

“Love Story”
- Type: Song
- Artist: Taylor Swift
- Genre: Pop

Birth Date: 12/13/1989
Entity-Based KGs

Characteristics of Entity-Based KGs
- Ontology (types, relationships) manually defined w. clear semantics
- Entities are named-entities, w. no overlap

Crazy Idea
Create a graph of entities and relationships to represent the world
Transforming Wikipedia to A Knowledge Graph

Feasibility

A prototype, an experiment, to show it is possible

First pot of gold: Transform Wikipedia Infoboxes to knowledge entities and relationships

Swift in 2022

Taylor Swift

Born
Taylor Alison Swift
December 13, 1989 (age 33)
West Reading, Pennsylvania, U.S.

Other names
Nils Sjöberg

Occupations
Singer-songwriter · producer · director · actress · businesswoman

mid128

name

“Taylor Alison Swift”

name

“Taylor Swift”

birth_date

12/13/1989
Transforming Wikipedia to A Knowledge Graph

High quality of Wikipedia data guarantees production quality. Common practice in industry

Well-known Examples:

1K

Quality

Production quality, E2E MVP experiences

Integrating Data from Different Sources

Tools for more data sources, long-tail domains

Repeatability
E2E pipelines, automations to allow for extensions to more domains

Are they the same person? → **Entity Linkage**
Are “Born” and “date of birth” the same? → **Schema Alignment**
Why “May 14, 1982” vs “7 November 1983”? → **Data fusion**

Heterogeneity, Heterogeneity, Heterogeneity
Integrating Data from Different Sources

Biggest Challenge: Entity Linkage

- Random forest on attribute-wise similarity
- Results between Freebase and IMDb movies

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie</td>
<td>99.0%</td>
<td>98.7%</td>
</tr>
<tr>
<td>People</td>
<td>99.3%</td>
<td>99.6%</td>
</tr>
</tbody>
</table>

1.5M labels
Integrating Data from Different Sources

Biggest Challenge: Entity Linkage

- Apply active learning to minimize #labels

For 99% precision and recall, active learning reduces #labels by 2 orders of magnitude

With 15K labels we get prec=99% and rec=\~95% (30 labelers for 1 week!)

Reaching prec=99% and rec=\~99% requires 1.5M labels

Zhu et al., Collective multi-type entity alignment between knowledge graphs, WebConf 2020.
Zhang et al., AutoBlock: A hands-off blocking framework for entity matching, WSDM 2020.
Zhang et al., OpenKI: Integrating open information extraction and knowledge bases with relation inference, NAACL 2019.
Trivedi et al., LinkNBed: Multi-Graph representation learning with entity linkage, ACL 2018.
Extracting Data from Semi-Structured Websites

Scalability

Significant cost reduction for scale ups

Two-stage extraction based on distant supervision

- Identify subject
- Identify (attribute, value) pairs

Lockard et al., Ceres: Distantly supervised relation extraction from the semi-structured web. VLDB, 2018.
Extracting Data from Semi-Structured Websites

Latest progress: Extraction on 30+ long-tail movie websites

Scalability
Significant cost reduction for scale ups

Production ready

Lockard et al., Ceres: Distantly supervised relation extraction from the semi-structured web. VLDB, 2018.

Lockard et al., OpenCeres: When open information extraction meets the semi-structured web. NAACL, 2019.

Lockard et al., ZeroShotCeres: Zero-shot relation extraction from semi-structured webpages, ACL 2020.
Web Knowledge Extractions & Fusion

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Solution: Extract knowledge from different types of web sources, and apply knowledge fusion to remove noises and generate probabilistic knowledge facts

Dong et al., Knowledge Vault: A Web-scale approach to probabilistic knowledge fusion. SIGKDD2014.
Web Knowledge Extractions & Integration

Why NOT in Google Knowledge Graph?

- Didn’t reach **production quality**
  - Accuracy=0.7 is far less than the production requirement (Accuracy=0.99)

- Didn’t find an **E2E MVP experience**
  - The 0.3B facts (vs. 70B in KG) are very “long-tail” to support meaningful use cases

- But underlying techs applied in long-tail knowledge collection etc.

<table>
<thead>
<tr>
<th>#Triples</th>
<th>3.2B</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.3B w. pr&gt;=0.7)</td>
<td></td>
</tr>
<tr>
<td>#URLs</td>
<td>2.5B</td>
</tr>
<tr>
<td>(28M Websites)</td>
<td></td>
</tr>
<tr>
<td>#Extractors</td>
<td>16</td>
</tr>
</tbody>
</table>
Entity-Based KG Summary

Crazy idea: A graph of entities and relationships to represent the world

Main challenges: Heterogeneous data everywhere

Framework:

Crazy idea not successful yet: Web-scale Extraction & Fusion
What About the Product Domain?

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Can we do the same for products?
Generation #2: Text-Rich Knowledge Graphs
Text-Rich KG Example

Product KG

Taxonomy

hasType

flavor

synonym

color

flavor

Cappuccino

Coffee

Gold

Prod. 1

Prod. 2

Prod. 3
Text-Rich KGs

Characteristics of Text-Rich KGs

- Ontology (types, relationships) very complex with overlaps and ambiguities; E.g., millions of product types
- Entities may not be named-entities, such as products
  E.g., “Onus 2 Colors Highlighter Stick, Shimmer Cream Powder Waterproof Light Face Cosmetics, creamy Self Sharpening Crayon STick Highlighter” vs. “Xin Luna Dong”
- Attribute values are oftentimes texts, with overlaps and ambiguities
  E.g., “Coffee” vs “Cappuccino” as icecream flavors

Crazy Idea:
Finding structure and modeling ambiguity from text sources
Use Case 1: Providing Information

Brand: Cetaphil
Ingredients: Water, Cetyl Alcohol, Propylene Glycol, Iodopropynyl Butylcarbamate, 2-Bromo-2-Nitropropane-1, 3-Diol, Sodium Lauryl Sulfate, Stearyl Alcohol, Methylparaben, Propylparaben, Sodium Citrate, Butylparaben, Allantoin, Zinc Gluconate.

Scent: Fragrance free
Additional Item Information: Non-Comedogenic, Fragrance-free, Natural

Skin Type: Sensitive

About this item:
- Gentle for everyday use; Cetaphil gentle skin cleansing cloths will leave your skin feeling clean, refreshed and balanced after every use.
- Removes makeup & dirt: Thoroughly remove makeup and dirt, leaving skin clean.
- Mild & non-irritating: Soap-free formulation won't strip skin of its natural protective oils and emollients.
Use Case II: Providing Choices
Use Case III: Improving Search
Use Case III: Improving Search
Use Case III: Improving Search
Use Case IV: Improving Recommendation
Do We Need Different Techniques?

Different challenges: *Unstructured* and *Noisy* product data
AutoKnow: Self-Driving Product Knowledge Collection

---

### Taxonomy

- Grocery
- Snacks
- Drinks
- Candy

### Catalog

<table>
<thead>
<tr>
<th>Product</th>
<th>Type</th>
<th>Flavor</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 1</td>
<td>Snacks</td>
<td>Cherry</td>
<td></td>
</tr>
<tr>
<td>Product 2</td>
<td>Candy</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Product 3</td>
<td>Candy</td>
<td>Choc.</td>
<td>Gold</td>
</tr>
</tbody>
</table>

---

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
AutoKnow: Self-Driving Product Knowledge Collection

<table>
<thead>
<tr>
<th>Product</th>
<th>Type</th>
<th>Flavor</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 1</td>
<td>Snacks</td>
<td>Cherry</td>
<td></td>
</tr>
<tr>
<td>Product 2</td>
<td>Candy</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>Product 3</td>
<td>Candy</td>
<td>Choc.</td>
<td>Gold</td>
</tr>
</tbody>
</table>

Dong et al., AutoKnow: Self-driving knowledge collection for products of thousands of types, SigKDD, 2020.
## Extracting Product Attribute Values

### Feasibility

A prototype, an experiment, to show it is possible
Bill Gates founded Microsoft in 1975.
Generic KE v.s. Product-Specific KE

Bill Gates founded Microsoft in 1975.

Solution: *OpenTag*—Applying deep tagging to identify structured attributes from product titles, descriptions, and bullets.

Objects are often not entities.

Subject is given.

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.
OpenTag Extraction from Product Profiles

- B: Beginning of attribute value
- I: Inside of attribute value
- O: Outside of attribute value
- E: End of attribute value

Input: $x = \{w_1, w_2, ..., w_n\}$ input sequence

Output: $y = \{t_1, t_2, ..., t_n\}$ tagging decision

Flavor Extractions:
- beef meal
- ranch raise lamb

The diagram shows the mapping of input words to output tags, with BIOE tagging scheme.
OpenTag Extraction from Product Profiles

**Bi-LSTM**
Captures sequence info

**Word Embedding**
Captures semantics of each token

**CRF**
Captures correlations between BIOE tags

**Attention**
Identifies important terms leading to attribute values

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.
OpenTag Extraction from Product Profiles

BiLSTM+CRF+Attention obtains best results

Extraction on new values is comparable to already known values

Zheng et al., OpenTag: Open attribute value extraction from product profiles, KDD 2018.
Building a Product Knowledge Extraction Pipeline

Quality goal: 90% accuracy for product attributes
Building a Product Knowledge Extraction Pipeline

**Repeatability**

E2E pipelines, automations to allow for extensions to more domains

1M

- Understand domain and attributes, and generate LOTS OF training data
- Train and fine-tune models
- Postprocess extraction results to further improve data quality
- Pre-publish evaluation as gatekeeper to guarantee high quality data
Building a Product Knowledge Extraction Pipeline

**Repeatability**

E2E pipelines, automations to allow for extensions to more domains

**Automatic Training Data Generation**

Weak Learning

Train and fine-tune models

OpenTag

Postprocess extraction results to further improve data quality

Benchmarking

Pre-publish evaluation as gatekeeper to guarantee high quality data
Building a Product Knowledge Extraction Pipeline

**1M**

**Repeatability**
E2E pipelines, automations to allow for extensions to more domains

**Automatic Training Data Generation**

**Train and fine-tune models**

**OpenTag**

**Deep Learning Data Cleaning**

**Pre-publish evaluation as gatekeeper to guarantee high quality data**

**Weak Learning**

**Benchmarking**

**Postprocess extraction results to further improve data quality**

**E2E pipelines, automations to allow for extensions to more domains**
Building a Product Knowledge Extraction Pipeline

1M

Repeatability

E2E pipelines, automations to allow for extensions to more domains

Automatic Training Data Generation

Train and fine-tune models

OpenTag

Deep Learning Data Cleaning

Scale-up pre-publish evaluation w. lower labeling needs

Weak Learning

Benchmarking

Postprocess extraction results to further improve data quality

Building a Product Knowledge Extraction Pipeline
Building a Product Knowledge Extraction Pipeline

1M

Repeatability
E2E pipelines, automations to allow for extensions to more domains

AutoML
Automatic Training Data Generation
Weak Learning

OpenTag

Deep Learning Data Cleaning

Benchmarking

Scale-up pre-publish evaluation w. lower labeling needs
Transformer-Based Anomaly Detection

Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag
Brand: Love of Candy
Price: $84.99 (1.50 / ounce) + $16.92 shipping
Pay $14.17/month for 6 months, Interest-free with your Amazon Prime Rewards Visa Card

Flavor Name: Pink
Choose from: Blue, Green, Orange, Pastel Assortment, Pink, Red, White, Yellow

Size: 6 Pound
- 1 Pound
- 2 Pound
- 3 Pound
- 4 Pound
- 5 Pound
- 6 Pound
- 7 Pound
- 8 Pound
- 9 Pound
- 10 Pound

• Love of Candy's huge selection of bulk candy now includes Premium Mint Chocolate Lentils in a variety of bold & striking colors. Available in small to large sizes ranging from 1 to 10 lb bags. These beautiful, chocolate morsels feature gourmet, dairy free dark chocolate coated in a crispy and crunchy mint candy shell. Similar to M&M's, these mint chocolate candy lentils are fun, bite-sized snacks that can be enjoyed during any occasion.
• Sourced from the most esteemed candy makers from around the world, we’ve put together an extremely broad collection of wholesale candy to fulfill your every need. Whether you're in need of candy for vending machines, pillars or candy buffets, you can trust that Love of Candy’s got you covered. Our consistent product quality and unmatched customer satisfaction have quickly made Love of Candy the market’s most trusted source of high quality, wholesale bulk candy.
Transformer-Based Anomaly Detection

Is the flavor “Pink”?
Transformer-Based Anomaly Detection
Transformer-Based Anomaly Detection

Category as input for model training
Transformer-Based Anomaly Detection

- Identify **1.77MM** incorrect values for Flavor and Scent for Consumables with **90% precision**

<table>
<thead>
<tr>
<th>Product</th>
<th>Attr</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Love of Candy Bulk Candy - Pink Mint Chocolate Lentils - 6lb Bag</td>
<td>Flavor</td>
<td>Pink</td>
</tr>
<tr>
<td>Scott's Cakes Dark Chocolate Fruit &amp; Nut Cream Filling Candies with Burgandy Foils in a 1 Pound Snowflake Box</td>
<td>Flavor</td>
<td>1 lb. snowflake box</td>
</tr>
<tr>
<td>Lucky Baby - Baby Blanket Envelope Swaddle Winter Wrap Coral Fleece Newborn Blanket Sleeper Infant Stroller Wrap Toddlers Baby Sleeping Bag (color 1)</td>
<td>Flavor</td>
<td>color 1</td>
</tr>
<tr>
<td>ASUTRA Himalayan Sea Salt Body Scrub Exfoliator + Body Brush (Vitamin C), 12 oz</td>
<td>Scent</td>
<td>vitamin c body scrub - 12oz &amp; body brush</td>
</tr>
<tr>
<td>Folgers Simply Smooth Ground Coffee, 2 Count (Medium Roast), 31.1 Ounce</td>
<td>Scent</td>
<td>2 Packages (Breakfast Blend, 31.1 oz)</td>
</tr>
</tbody>
</table>
Scaling Up Product Knowledge Extraction

**Solution**: One-size-fits-all models

- **Scalability**: Significant cost reduction for scale ups
- **Millions of categories**
- **Thousands of attributes**
- **Hundreds of languages**

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Scaling Up for Millions of Categories

**Option 1. Train a single model?**  
Train/Test Distribution shift -> Invalid predictions

- **Samsung UN58RU7100FXZA Flat 58-Inch 4K UHD 7 Series Ultra HD Smart TV with HDR and Alexa Compatibility (2019 Model)**
- **Taylors of Harrogate Classic Tea Variety Box, 48 Count (Pack of 1)**
- **Caribou Coffee Caribou Blend, Medium Roast Ground Coffee, 20 Ounce Bag, Rainforest Alliance Certified**
Scaling Up for Millions of Categories

Option 1. Train a single model?  
Train/Test Distribution shift -> Invalid predictions

Option 2. Train a model for each category?

Store/orchestrate 100K+ OpenTag

Most categories are very sparse

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Scaling Up for Millions of Categories

Scalability
Significant cost reduction for scale ups

Figure 2: Our TXtract architecture for hierarchical multi-task learning.
Scaling Up for Millions of Categories

Scalability

Significant cost reduction for scale ups

Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Attention conditioned on category representation

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Scaling Up for Millions of Categories

Scalability

Significant cost reduction for scale ups

Multi-task Learning

Figure 2: Our TXtract architecture for hierarchical multi-task learning.

Karamanolakis et al., TXtract: Taxonomy-aware knowledge extraction for thousands of product categories, ACL 2020.
Train one model on 4K categories, and improve state-of-the-art by 10.4% in F1, and by 11.7% in coverage

<table>
<thead>
<tr>
<th>Title</th>
<th>OpenTag</th>
<th>TXtract</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Controlled Labs Purple Wraath 90 Servings - <strong>Purple Lemonade</strong></td>
<td>flavor: -</td>
<td>flavor: purple lemonade</td>
</tr>
<tr>
<td>2 Click - Espresso Protein Drink <strong>Vanilla Latte</strong> - 16 oz.</td>
<td>flavor: espresso</td>
<td>flavor: vanilla latte</td>
</tr>
<tr>
<td>3 Mason Vitamins Melatonin 500 mcg Fast Meltz Tablets, <strong>Fruit</strong>, 60 Count</td>
<td>flavor: -</td>
<td>flavor: fruit</td>
</tr>
<tr>
<td>4 Fashion Glitter Matte Eye Shadow Powder <strong>Palette</strong> Single Shimmer Eyeshadow</td>
<td>scent: palette</td>
<td>scent: -</td>
</tr>
<tr>
<td>5 Baby car seat cover, Nursing covers <strong>Breastfeeding cover carseat canopy (Style5)</strong></td>
<td>scent: style5</td>
<td>scent: -</td>
</tr>
</tbody>
</table>
Product kno(wl)edg(e) extraction from images, reviews, etc.

Ubiquity
High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Product Knowl. Extraction from Broader Sources

1T

Ubiquity

High coverage, long-tail use cases, assumption removal, to next cycle of inventions

Multi-modal knowledge extraction

<table>
<thead>
<tr>
<th>Models</th>
<th>P(%)</th>
<th>R(%)</th>
<th>F1(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM w/o text</td>
<td>79.9</td>
<td>63.4</td>
<td>70.7</td>
</tr>
<tr>
<td>PAM w/o image</td>
<td>88.7</td>
<td>72.1</td>
<td>79.5</td>
</tr>
<tr>
<td>PAM w/o OCR</td>
<td>82.0</td>
<td>69.4</td>
<td>75.1</td>
</tr>
<tr>
<td>PAM</td>
<td>91.3</td>
<td>75.3</td>
<td>82.5</td>
</tr>
</tbody>
</table>

Text still plays the most important role

Lin et al., PAM: Understanding product images in cross product category attribute extraction, KDD 2021.
Text-Rich KG Summary

Crazy idea: Finding structure and modeling ambiguity from text-rich sources

Main challenges: **Sparse & Noisy** structured data everywhere

Framework:

Unmentioned crazy idea: **Automatic taxonomy extraction & construction**

- Zhang et al., Minimally supervised structure-rich text categorization via learning on text-rich networks.
- Zhang et al., OA-Mine: Open-World attribute mining for e-Commerce products with weak supervision. Webconf, 2022
Generation #3: Media-Rich Time-Based KGs
Background: Virtual Intelligent Assistant

Respond to commands

“Hey Siri, set a timer to 7pm”

“Ok, added to today’s reminders”
Background: Virtual Intelligent Assistant

Control devices

“Hey Alexa, turn off bedroom lights”
Background: Virtual Intelligent Assistant

Provide information

“Hey, Google, when did winter start?”

“Winter started on Wed, December 21”
Meta’s Assistant

Empowering connection to people and experiences in your life

Meta Quest 2

“Hey Facebook” (double press the button on your controller)
“Who’s online?”—meet up with friends
“Open Beat Saber”—jump straight in the game, and more.

Ray-Ban Stories

“Hey Facebook, take a picture”—capture moments hands-free
“Hey Facebook”—call friends on Messenger, manage device settings, and more.
What is An Ideal Virtual Intelligent Assistant?

An *intelligent assistant* should be an agent that *knows you and the world*, can *receive your requests* or *predict your needs*, and provide you the *right services at the right time* with your permission.
What is An Ideal Virtual Intelligent Assistant?

An intelligent assistant should be an agent that knows you and the world, can receive your requests or predict your needs, and provide you the right services at the right time with your permission.

Public KGs

Personal KGs
Evolution of Intelligent Assistant

**Chatbot**
- Text input

**Voice Asst**
- Voice input

**AR/VR Asst**
- Voice + Visual + Context
What Is Different for An AR/VR Assistant?

You see through it

You wear it everywhere

May not have connection

You wear for a long time
From Voice-Only to Multi-Modal

“How tall is Empire State Building?”

“What's the name of this building and how tall is it?”
From Context-Agnostic to Context-Aware

“Show my shopping list”

“Remember to buy apples and bananas at the grocery store around the corner”
“What’s the weather today?”

“Today is sunny, 70 degree. Would you like to play your favorite morning music?”
From Server-Side to On-Device

May not have connection
Two Sides of One Coin (1):
Great Vehicle for Life Recording

MEMEX (MEMory & EXpansion)
by Vannevar Bush (1945)
Two Sides of One Coin (2): Great Vehicle for Personal Assistant Recommendation

- Reactive conversational recommendation
- Proactive service recommendation
- Personal knowledge (preferences, routines, etc.)
Personal KG and Memovoir

Utility
Where did I put my key?
I must have seen this lady before but when and where?

Memoir
[Image: MY STORY]

Inspiration
At Lyon you can visit the statue of Saint-Exupéry with the Little Prince by his side. You read that book in 2018 and loved it.
Time-Based KGs & Media-Rich Memovoir

Characteristics of a Personal KG & Memovoir
- Rich audio/video, associated w. time, location, etc.
- Modeling
  - entities at an abstract level; e.g., latte art coffee (from different stores), my key for front door
  - activities; e.g., dancing, watching ballet
- Historical—each activity is associated with a timestamp

Crazy Idea
Trace and abstract one’s life from rich audios/videos, and use it for life experience Q&A, life journal creation, and life recommendation.
Research Problems for Personal KG and Memovoir

- How to record one’s life with hardware (memory, battery) constraints?
- How to extract personal knowledge from the recordings?
- What is the best frequency, granularity, and domain richness to capture one’s life?
- How to leverage the Personal KG and Memovoir for utility, memoir, and inspiration applications?
- How to leverage Personal KGs to best understand contexts?
- How to combine public and personal KGs for context-aware recommendation?
<table>
<thead>
<tr>
<th>Take-Aways</th>
<th>1. Entity-Based KGs</th>
<th>2. Text-Rich KGs</th>
<th>3. Media-Rich Time-Based KGs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resolving</strong> heterogeneity with <strong>entity linkage</strong> and <strong>web knowledge extraction</strong></td>
<td><strong>Extractions</strong> and <strong>cleanings</strong> from sparse and noisy source data, and handling semantics ambiguities</td>
<td><strong>Many new challenges</strong> for knowledge collection and applications</td>
<td></td>
</tr>
</tbody>
</table>
Take-Away 2. From Roofshots to Moonshots

1,000,000,000,000,000

Feasibility  Quality  Repeatability  Scalability  Ubiquity

START

Roofshots

Moonshots
Shameless Advertisements

Book (2021)

Two benchmark datasets

- DI2KG Challenge: http://di2kg.dia.uniroma3.it/#challenge
- Extended SWDE benchmark: https://homes.cs.washington.edu/lockardc/expanded_swde.html
Thank You

Q&A?