Understanding Customer Requirements
An Enterprise Knowledge Graph Approach


Basel Shbita

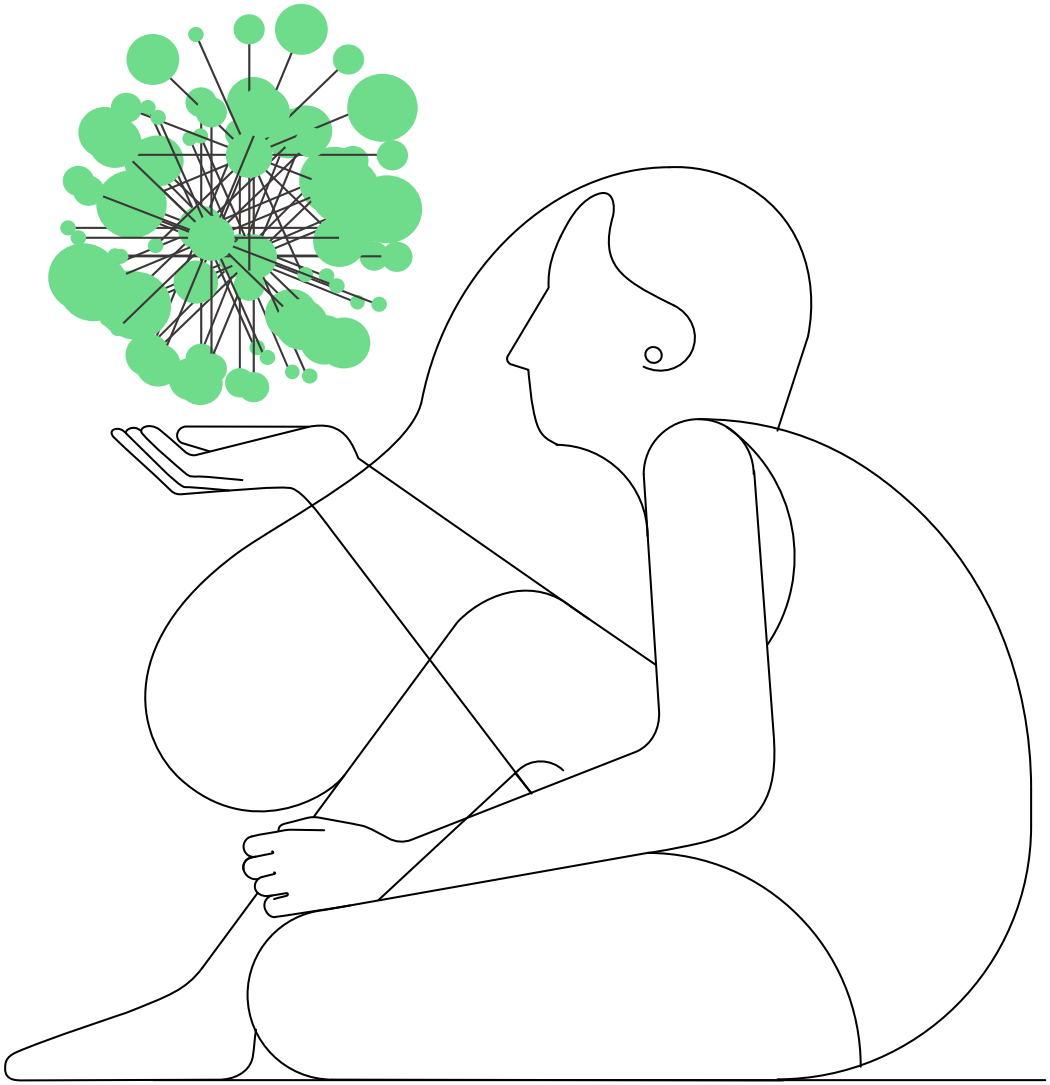

Anna Lisa Gentile


Pengyuan Li

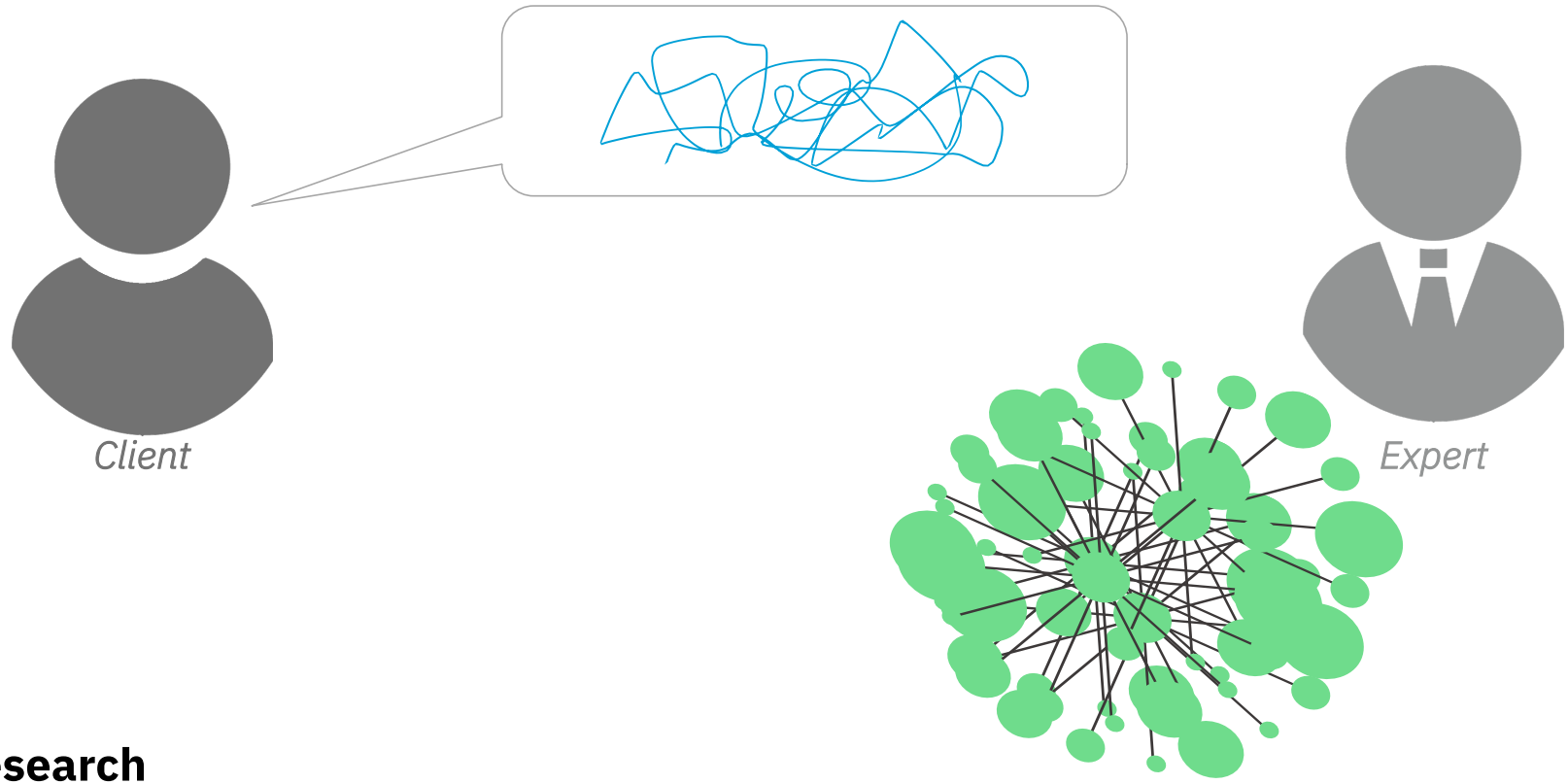

Chad DeLuca


Guang-Jie Ren

IBM Offerings Ecosystem: Enterprise Knowledge Graph (EKG)



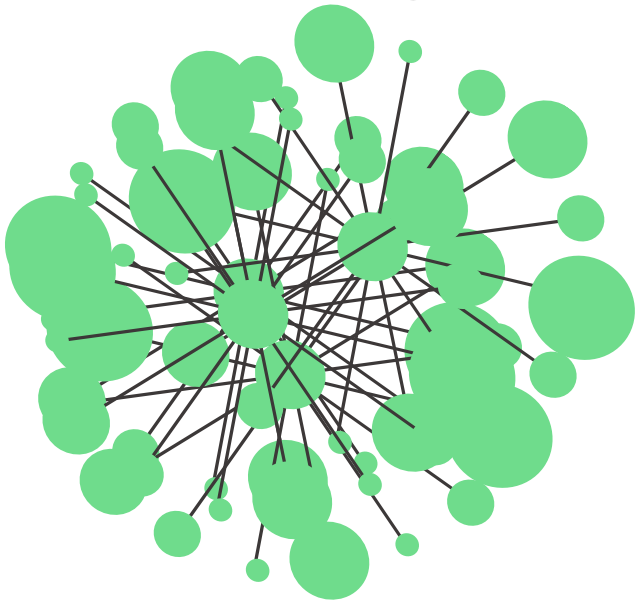
IBM **Research**

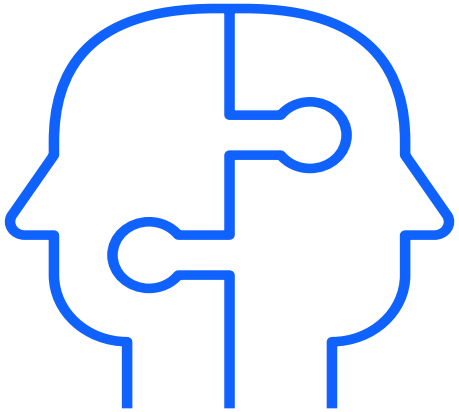


IBM **Research**

Initial EKG – IBM Offerings Ecosystem:

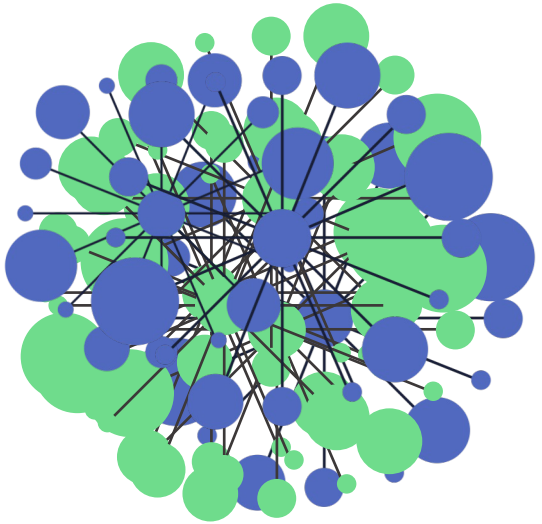
- Business Units
- Topical Verticals
- Business Offerings/Assets





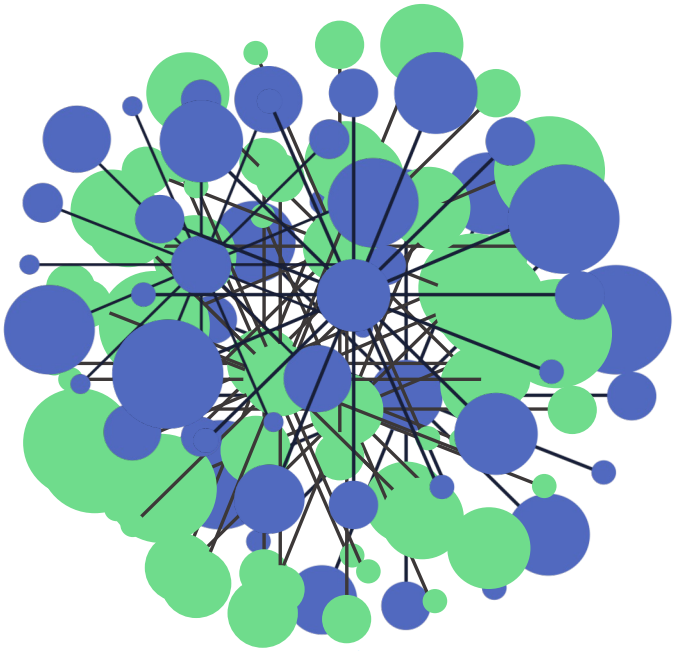
Augmented EKG:

- + Client requirements
- + Client pain points
- + Client needs



IBM **Research**

Augmented EKG



Internal Unstructured Data
(Client Stories)

Open Knowledge
from the Web

Domain Expertise

Contributions

Orchestration of

NLP techniques + Semantic Resources + Human Expertise

(i) retrieve, select, and distill instrumental **knowledge from the Web** to build the EKG
(ii) extract structured information from technical text
(iii) user-driven mechanism to encapsulate the extracted knowledge within the EKG

Context

Open IE

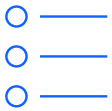
(i) Machine reading
(ii) Information represented without explicit semantics (often numerical vectors)
(iii) No reasoning mechanism
(iv) When semantics is explicit → no specific domain ontology

Targeted IE

(i) Ontology population from text
(ii) Finding entities belonging to predefined classes
(iii) KG-based Named Entity Recognition methods

Client Stories

Strategic goals



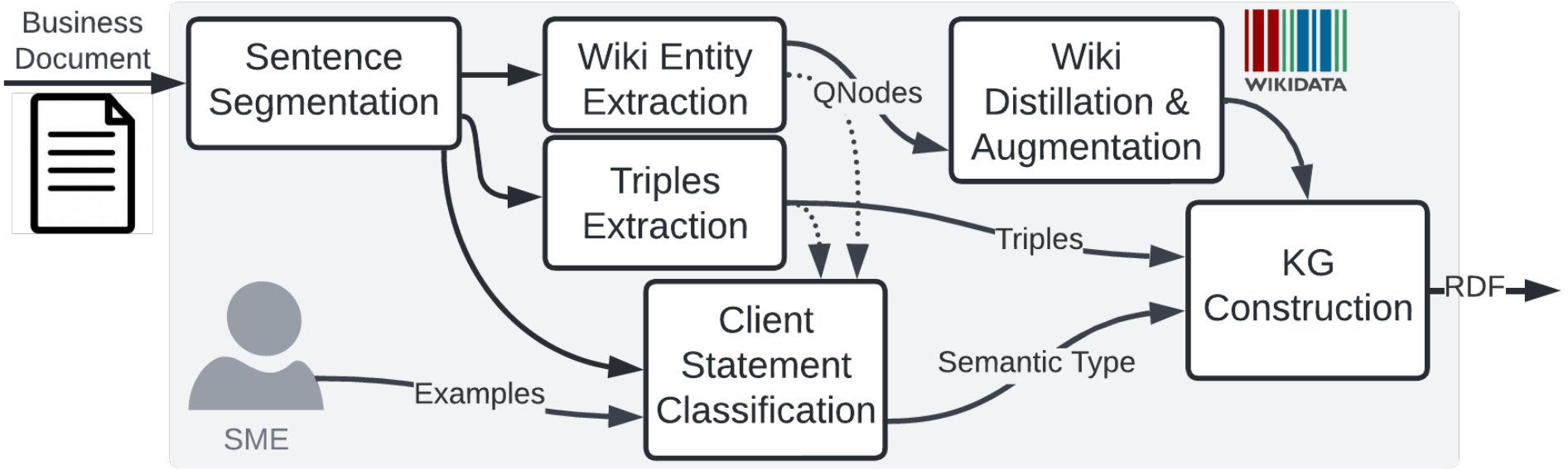
Pain points

Targeted areas for improvement

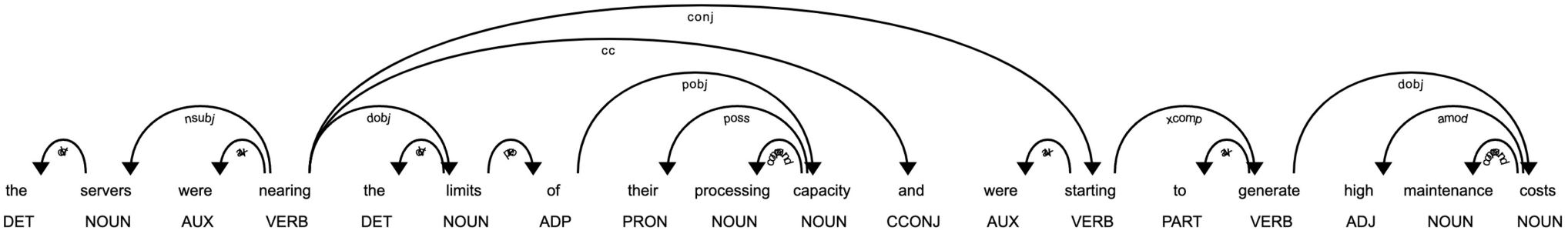
<COMPANY NAME> uses a Microsoft Distributed File System to enable employees to access data and files residing on different servers, usually to create financial reports. Previously, <COMPANY NAME> was supporting its Microsoft Distributed File System with older IBM Power 570 servers. Although the incumbent technology had performed satisfactorily, the servers were nearing the limits of their processing capacity and were starting to generate high maintenance costs. In addition, <COMPANY NAME> recognized the opportunity to increase storage capacity and streamline administration tasks within its storage area network (SAN). Thus, <COMPANY NAME> sought a reliable IT provider to help it refresh its server and storage technology and boost the performance of its Microsoft Distributed File System.

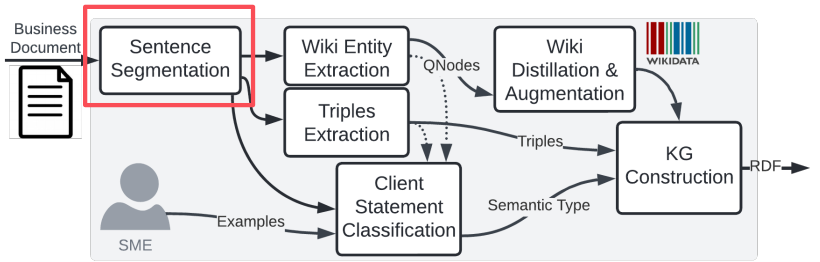
Business Document Example

EKG augmentation Pipeline



Sentence Extraction





IBM **Research**

Triple Extraction

<COMPANY NAME> uses a Microsoft Distributed File System to enable employees to access data and files residing on different servers, usually to create financial reports.

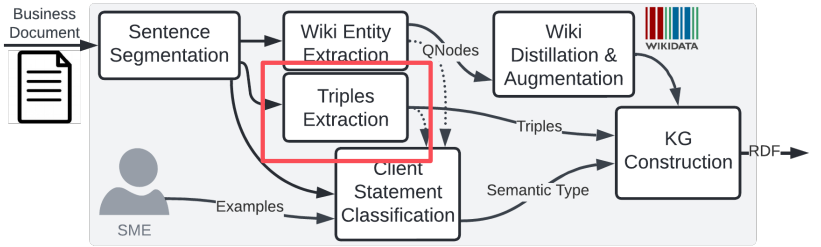
Previously, <COMPANY NAME> was supporting its Microsoft Distributed File System with older IBM Power 570 servers. Although the incumbent technology had performed satisfactorily, the servers were nearing the limits of their processing capacity and were starting to generate high maintenance costs.

In addition, <COMPANY NAME> recognized the opportunity to increase storage capacity and streamline administration tasks within its storage area network (SAN).

Thus, <COMPANY NAME> sought a reliable IT provider to help it refresh its server and storage technology and boost the performance of its Microsoft Distributed File System.

subject
predicate(s)
object(s)

QNode(s)



IBM **Research**

Wiki Entity Extraction

Q1229600 [Distributed File System (Microsoft)]/ class: **Q339678** (distributed data store)

<COMPANY NAME> uses a Microsoft Distributed File System to enable employees to access data and files residing on different servers, usually to create financial reports.

Previously, <COMPANY NAME> was supporting its Microsoft Distributed File System with older IBM Power 570 servers. Although the incumbent technology had performed satisfactorily, the servers were nearing the limits of their processing capacity and were starting to generate high maintenance costs.

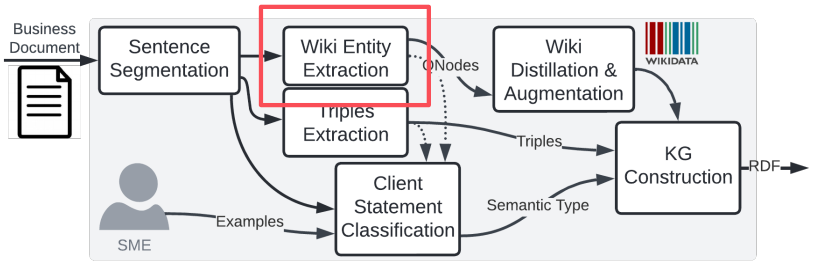
In addition, <COMPANY NAME> recognized the opportunity to increase storage capacity and streamline administration tasks within its storage area network (SAN).

Q237576 [Storage area network] / class: **Q994895** (technical system)

Thus, <COMPANY NAME> sought a reliable IT provider to help it refresh its server and storage technology and boost the performance of its Microsoft Distributed File System.

subject
predicate(s)
object(s)

QNode(s)



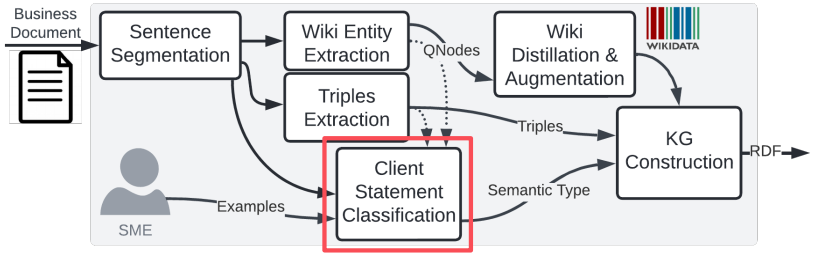
IBM **Research**

User Assisted Client Statement Classification

Statement type	Description
profileStatement	General attributes about the entity
attemptStatement	Attempt or historical actions the entity has taken
needStatement	Needs or desired outcomes the entity requires
painPointStatement	Pain points or obstacles the entity finds critical

```yaml
# org entity
---
type: entity
name: org
automatically_extensible: yes
values:
  - 3M Co.
  - [HP, Hewlett-Packard]
  - IBM
  - Intel
```

```yaml
# profileStatement intent
---
type: intent
name: profileStatement
slots:
  - name: orgEntity
    entity: org
utterances:
  - "[orgEntity] aims to be a personal bank in the digital age"
  - "[orgEntity] is one of the largest financial services companies in the United States"
```



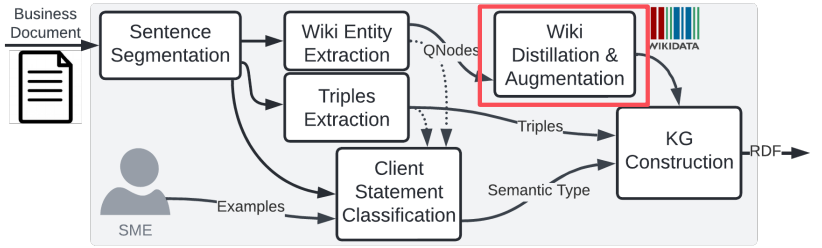
Subgraph Retrieval

Algorithm 1: Distilled Wikidata graph construction algorithm

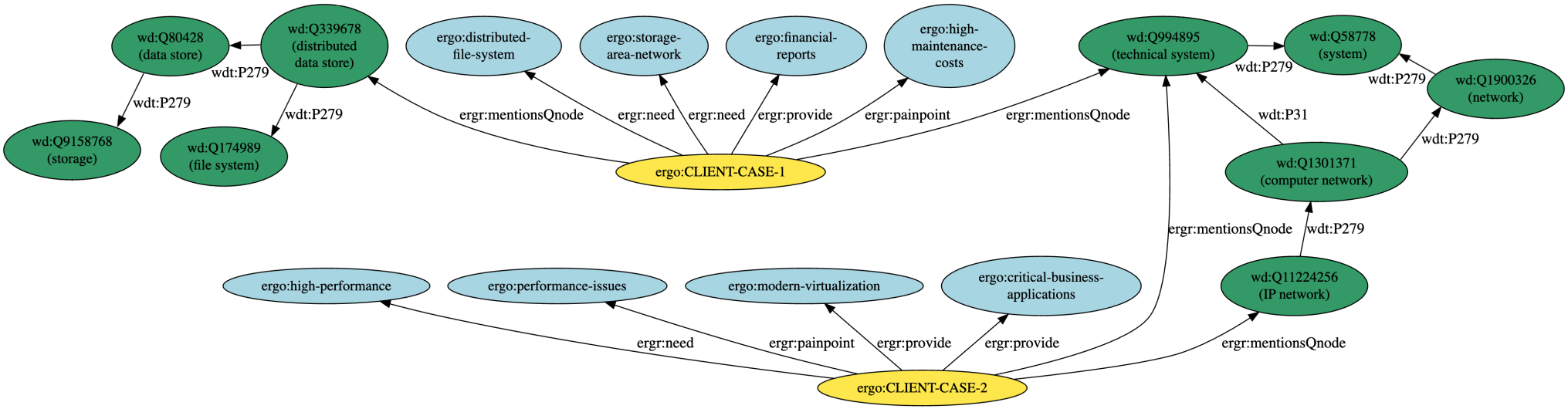
Data: a set $\mathcal{Q}$ of QNodes
Result: a directed acyclic graph $\mathcal{G}$ of QNodes

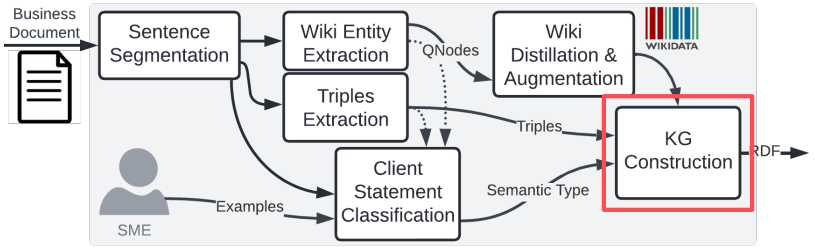
1 **foreach** $q \in \mathcal{Q}$ **do**
2 $\mathcal{G}$.add(q); // add all QNodes to graph
3 **while** *True* **do**
4 $\mathcal{L}$ = list of nodes in $\mathcal{G}$ with no outgoing P31/P279 edges;
5 $\mathcal{Q}^* = \texttt{getP31P279}(\mathcal{L}) \setminus \mathcal{Q}$;
6 **foreach** $q^* \in \mathcal{Q}^*$ **do**
7 $\mathcal{G}$.add($\bigcup_{q_j^* \in P31P279(q)} q \mapsto q_j^*$); // add new nodes & edges
8 **if** $|\mathcal{Q}^*| == 0$ **then**
9 break;
10 $\mathcal{Q} = \mathcal{Q}^* \bigcup \mathcal{Q}$

```
SELECT ?c ?cLabel WHERE {
  wd:Q994895 wdt:P31?/wdt:P279* ?c .
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en". } }
```



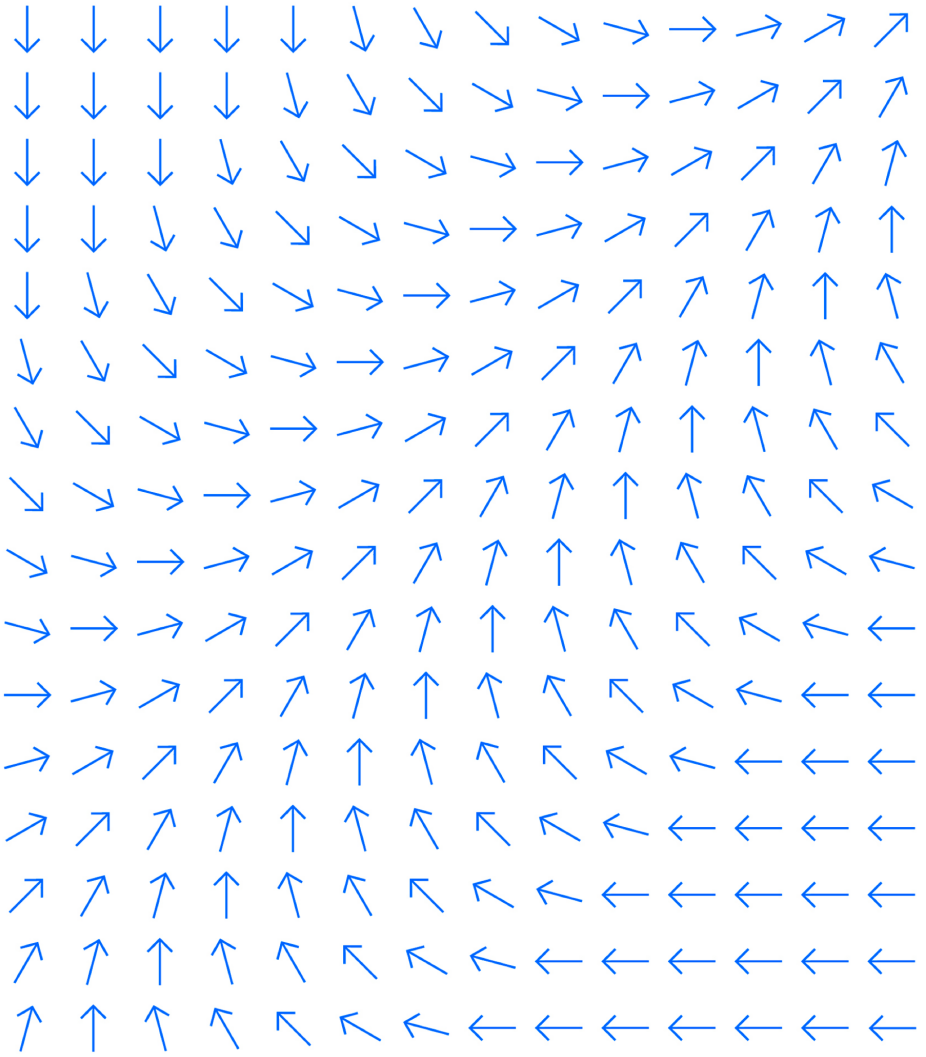
Augmented EKG





IBM **Research**

Experiments



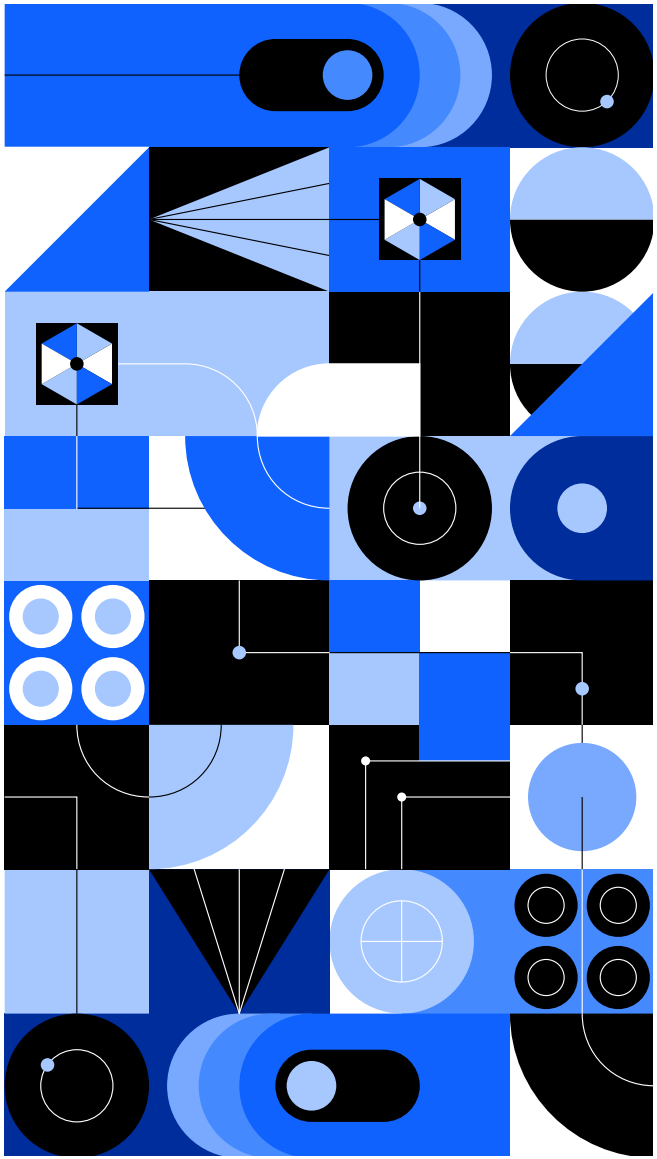
Dataset

10 years of historical data
successful collaborations with IBM customers

24,180 customer stories
customer requirements
corresponding sold assets and services
assigned set of business-unit

average document length: 127 words

business-unit labels
1 to a maximum of 9 (avg 2.4)
distinct labels: 13



Multi-label Classification Task

Classifying an unseen document with one or more business units

Baseline

- text based features only
- **Doc2Vec** embedding
  - document → feature vector
  - dimension: 100
  - simple and fast model
  - useful choice for a medium-sized dataset
- train a One-vs-Rest (OvR) multi-label classification model
  - Support Vector Machines (SVM) with a linear kernel

KG settings

- document → graph
  - with KG variations
- graph → KG embedding
  - **ComplEx**
  - low-dimensional representation of entities and relations in the KG
  - vector representation for each node in the EKG, including those representing client documents
- concatenate:
  - **Doc2Vec + ComplEx**
  - dimension: 100 (50 +50)

KG Variations

- no augmentation
  Doc2Vec + KG
- only **SME** semantic types
  Doc2Vec + Augmented KG (SME)
- only **Wikidata** entities
  Doc2Vec + Augmented KG (Wiki)
- **fully augmented** KG
  Doc2Vec + Augmented KG (Full)

Results

Method	Precision	Recall	F_1
Doc2Vec	0.730	0.590	0.653
Doc2Vec + KG	0.741	0.605	0.666
Doc2Vec + Augmented KG (SME)	0.754	0.608	0.673
Doc2Vec + Augmented KG (Wiki)	**0.761**	0.621	0.684
Doc2Vec + Augmented KG (Full)	**0.761**	**0.638**	**0.694**

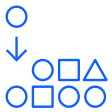
Conclusions

Goal

Helping Sales People to navigate and browse company offerings

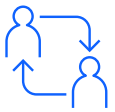
EKG

Offerings
services
Client Needs



Downstream Task Evaluation

Business Unit Classification



Customer data

10 years of client success stories

Understanding Customer Requirements
An Enterprise Knowledge Graph Approach


Basel Shbita


Anna Lisa Gentile


Pengyuan Li


Chad DeLuca


Guang-Jie Ren