Financial Crime and Compliance — Transforming Text Documents to Graphs

Doga Tekin – Member of Technical Staff, Oracle Labs Zurich

doga.tekin@oracle.com

Helpful Links –

ORACLE LABS:
https://labs.oracle.com/

ORACLE GRAPH:
- https://oracle.com/goto/graph
- https://bit.ly/Graph-LiveLabs

ORACLE FINANCIAL SERVICES COMPLIANCE STUDIO:
https://www.oracle.com/financial-services/aml-financial-crime-compliance/crime-compliance-studio/

ORACLE AI LANGUAGE:
https://www.oracle.com/artificial-intelligence/language/

Future & Past TechCasts:

June 27th

Live! from Kscope

Presented by **Roger Cressey**

July 13th

"Advancing Analytics at Rosendin"

Presented by Cathye Pendley

July 27th

"Data Science Survey Results"

Presented by Karl Rexer & Tim Vlamis

TechCast Archive

Click to see Live TechCast page

2023 | 2022 | 2021 | 2020 | 2019

Date	Title	Presenter(s)	Replay	Download(s)
May 4	Oracle APEX: A Swiss Army Knife Story for Your Analytics	Lucas Hirschegger & Simon Collins	Video	Slides
Apr 20	From Data to Insights with Oracle Analytics	Joel Acha	Video	Slides
Apr 6	Data Platform Migrations – Few Learnings	Sujata Balupala & Sanjay Sabnis	Video	Slides
Mar 14-16	AnDOUC Summit 2023	AnDOUC	– –	– –
Feb 16	Favorite New Features in Jan 2023 OAC	– –	– –	– –
Jan 26	Summit Preview of Presentations	Various Presenters	Video	Slides
Jan 12	Oracle Analytics & Spatial Studio	Wayne Van Sluys and David Lapp	Video	Slides

Submit a topic to share at **https://andouc.org/techcasts/**

Analytics and Data ORACLE USER COMMUNITY

www.andouc.org

Oracle Spatial & Graph SIG

Analytics and Data
ORACLE USER COMMUNITY

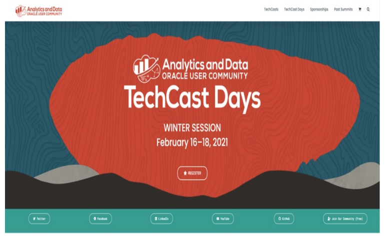
We Have Merch!

Show your "Tech Side" in everything you do!

Visit the AnDOUC Store at ANDOUC.ORG

Analytics and Data
ORACLE USER COMMUNITY

Let’s Connect



Website
http://andouc.org/

Chat with the Experts
https://bit.ly/Join-ANDOUC-Slack

Watch Previous TechCasts
https://bit.ly/3qmGgHN

@AnalyticAndData

https://www.facebook.com/AnDOracleUserCommunity

https://www.linkedin.com/company/analytics-and-data-oracle-user-community

Spatial + Graph SIG
bit.ly/Spatial-Graph-LinkedIn

Analytics and Data
ORACLE USER COMMUNITY

Save the Date!

Analytics and Data Summit 2024

Call for Speakers now open!

March 19-21, 2024
Oracle Conference Center
Redwood Shores, California

www.andouc.org/andsummit2024

Financial Crime and Compliance — Transforming Text Documents to Graphs

Doga Tekin, Member of Technical Staff
@ Oracle Labs

29 June 2023

Oracle Labs

Exploratory Research

- Pursue new ideas within domains relevant to Oracle

Directed Research

- In collaboration with product teams
- Difficult, future-looking problems
- Driven by product requirements

Consulting

- Provide unique expertise
- Small engagement across product organizations

Product Incubation

- Grow new products from Oracle Labs research

This talk will include both ongoing research work and publicly available Oracle features!

Agenda

Introduction

- Knowledge Graphs
- Motivation and Use Cases

Related Work

- Knowledge Graph Construction from Text
- Transformer Models

Approach

- Transformer Pipeline for Text to Graph

Data

- Annotation Requirements

Results & Demo

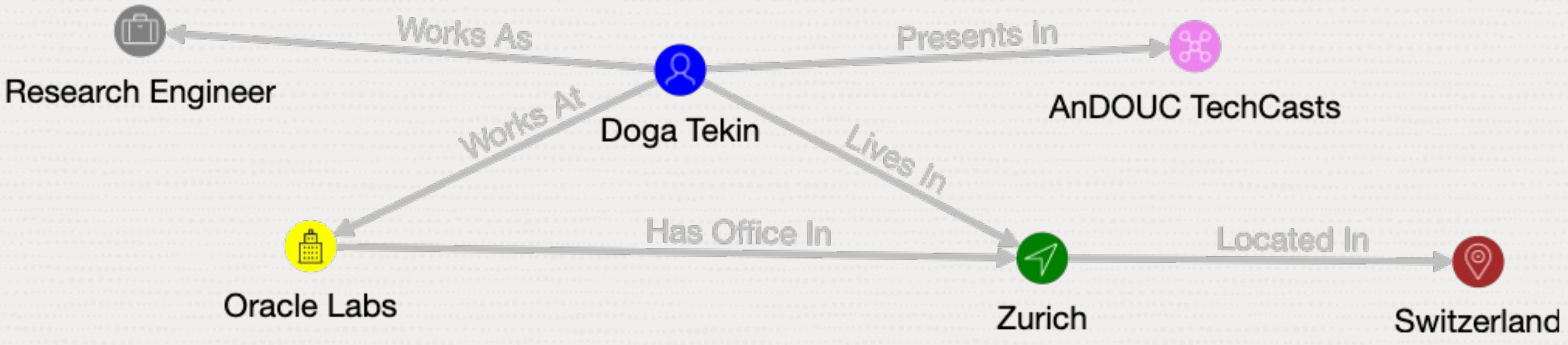
- Experimental Results & Visual Demo

Using the Graphs to Fight Financial Crime

Takeaways

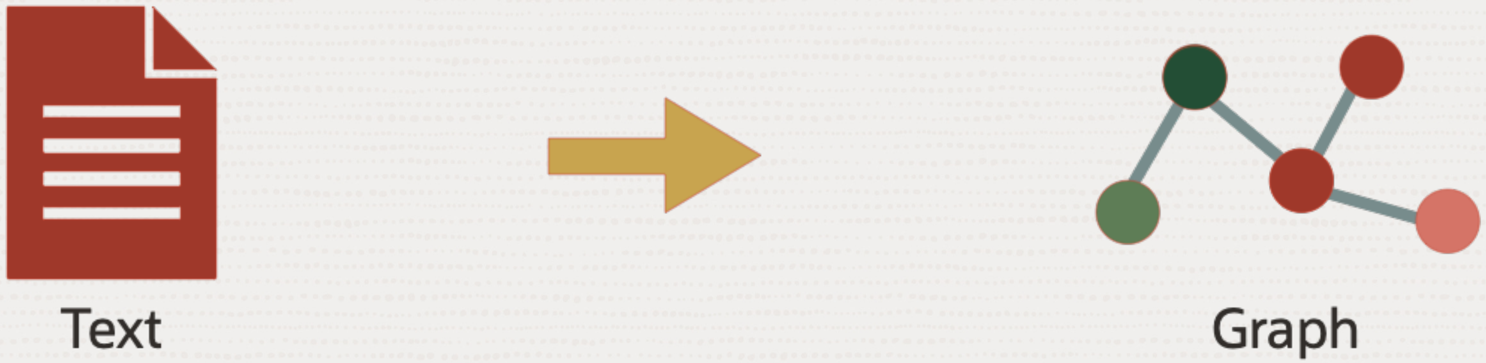
Introduction: Knowledge Graphs

Knowledge graphs are knowledge bases that use a graph structure to represent information about entities and their relations.



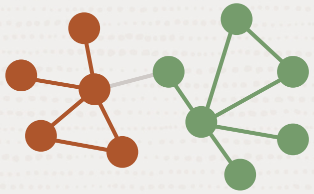
- Useful for many applications, including improving search engines and smart assistants, detecting fraud and analyzing financial crime, and predicting diagnoses in healthcare.
- They are performant & expressive; they allow data integration, unification, and analysis.

Introduction: Motivation



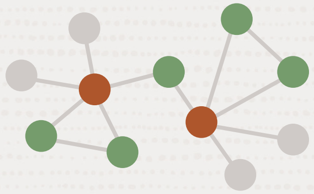
- **Visual inspection:** seeing information presented graphically as connections between nodes can help us see patterns that are easy to miss in plain text.
- **Empower downstream algorithms:** if we obtain a graph representation of relevant information, we can apply many graph algorithms and graph machine learning models to achieve downstream tasks utilizing structural information.

Introduction: Capabilities of Oracle Graph



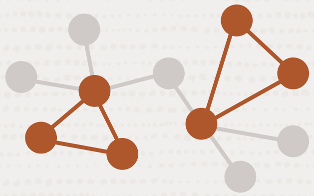
Detecting communities

Strongly Connected Components, Weakly Connected Components, Label Propagation, Louvain, Conductance Minimization, Infomap



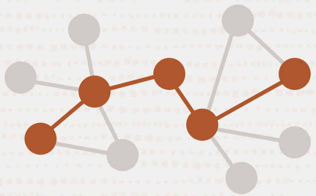
Ranking and walking

PageRank, Personalized PageRank, Degree Centrality, Closeness Centrality, Vertex Betweenness Centrality, Eigenvector Centrality, HITS, Minimum Spanning-Tree (Prim's), BFS, DFS, Random Walk with Restart



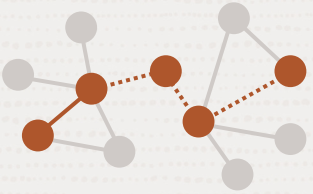
Topology analysis

Conductance, Cycle Detection, Degree Distribution, Eccentricity, K-Core, LCC, Modularity, Reachability Topological Ordering, Triangle Counting, Bipartite Check, Partition conductance



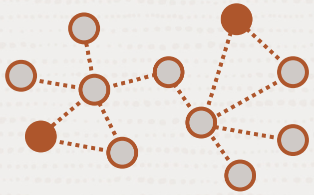
Path-finding

Shortest Path (Bellman-Ford, Dijkstra, Bidirectional Dijkstra), Fattest Path, Compute Distance Index, Enumerate Simple Paths, Filtered and Unfiltered Fast Path Finding, Hop Distance



Link prediction and others

Twitter Whom-to-follow, SALSA, Adamic-Adar Index



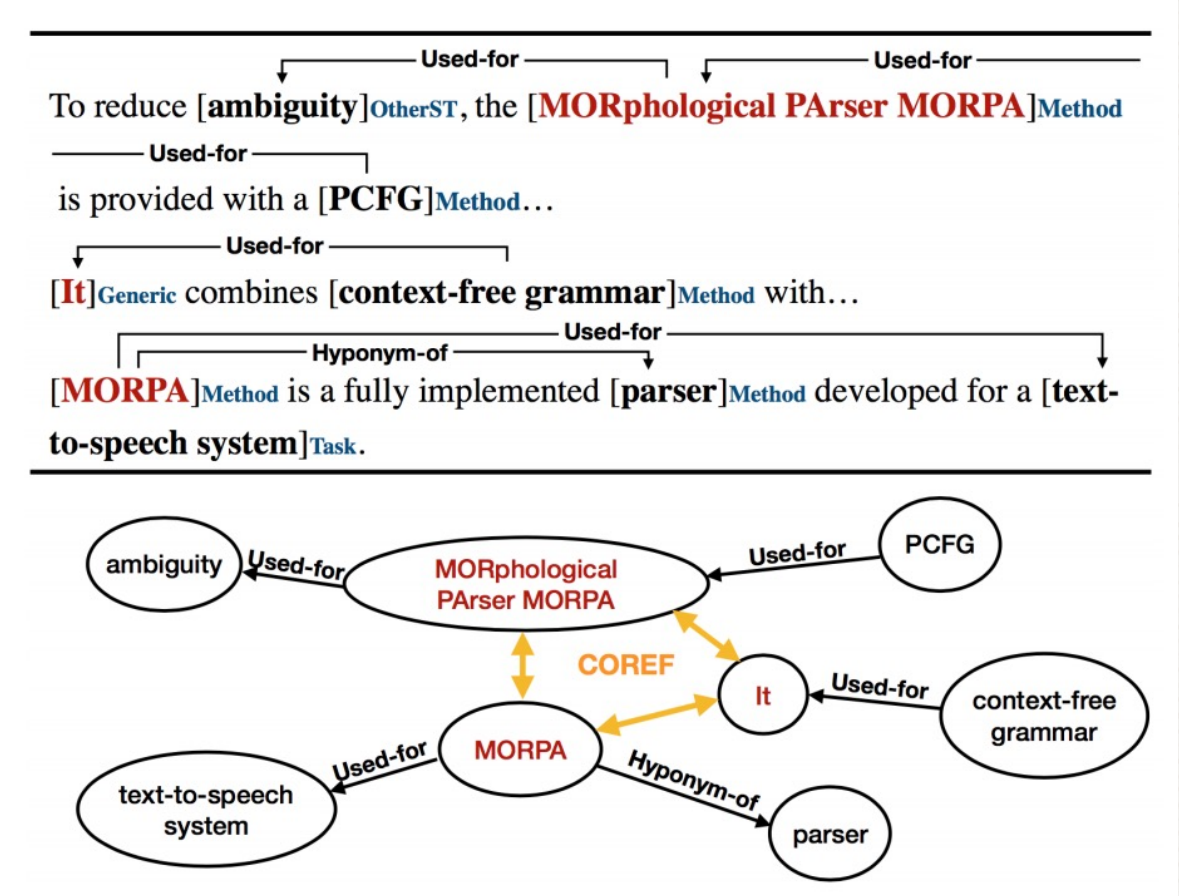
Machine learning

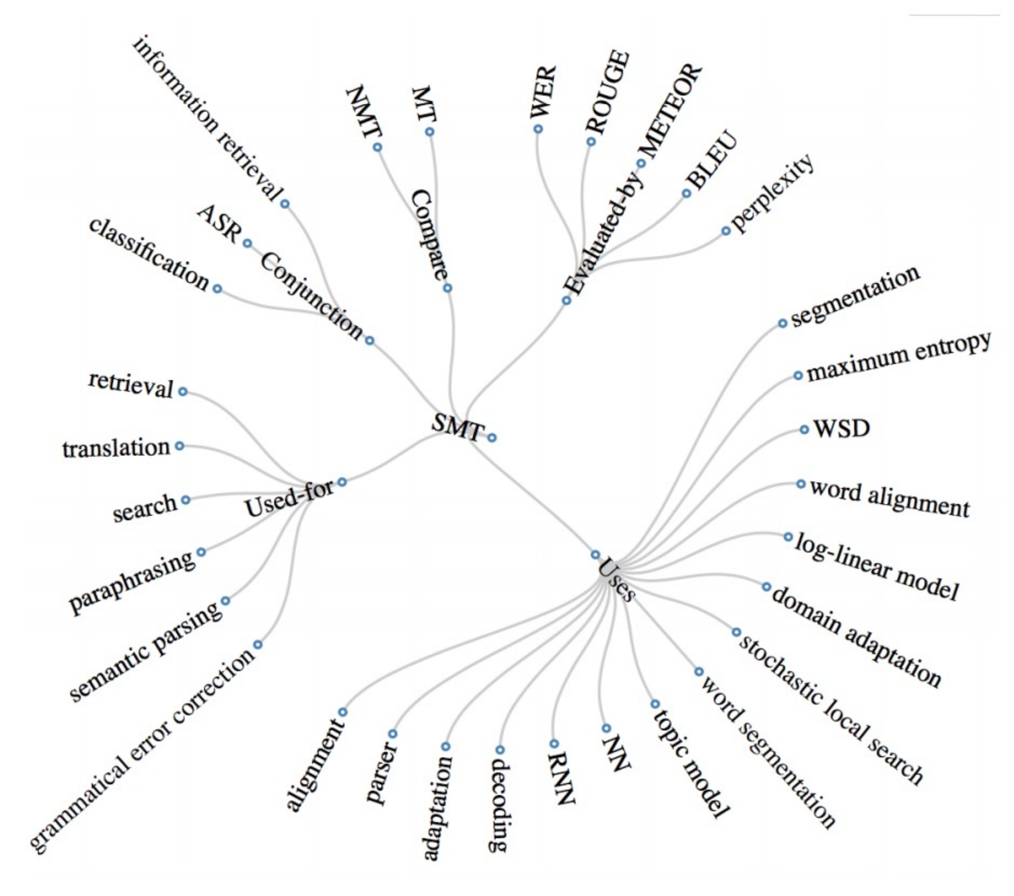
DeepWalk, Supervised GraphWise, Unsupervised GraphWise, Pg2Vec, Matrix Factorization, GNNExplainer

Available in every edition of Oracle Database, including the 23c Free Developer Release!

Introduction: Use Cases

Scientific Knowledge Graph Construction from Paper Abstracts





Source: Luan et al. (2018)

Introduction: Use Cases

Case Graph Construction from Suspicious Activity Reports (SARs)

Introduction:

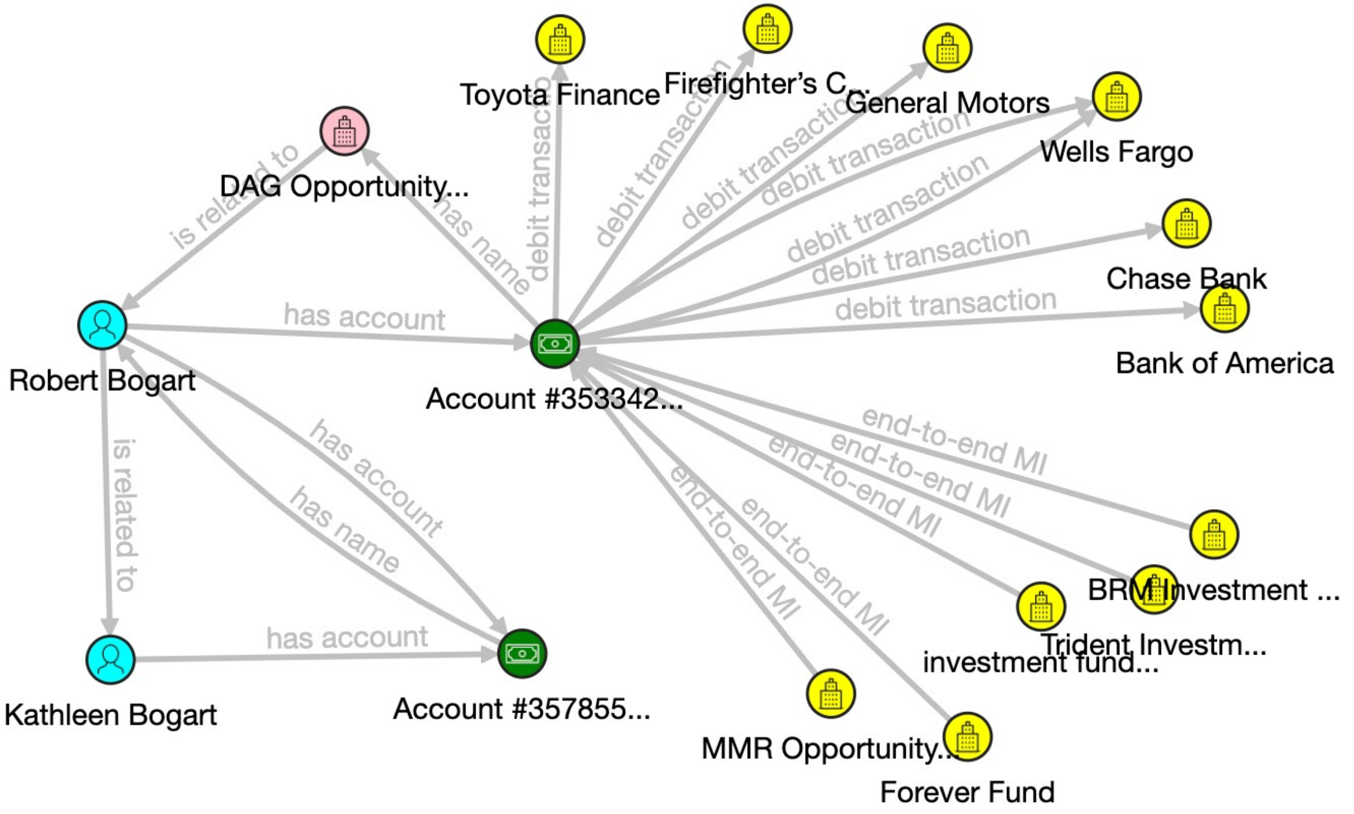
This case was referred for investigation by the MyBank AML Detection unit. The referral identifies potential cash deposit structuring activity in account number 353342287 in the name of DAG Opportunity Fund, LLC.

This investigation, which covers the time period of 1/1/20 through7/6/20 (account closure), revealed suspicious activity totaling $1,399,185.00, occurring between 1/8/20 and 6/18/20.

Details of Investigation:

According to internal bank records, DAG Opportunity Fund, LLC is a pooled | investment account. The account signer, Robert Bogart, age 55, along with his spouse, Kathleen, age 48, are listed as an investment broker and schoolteacher respectively, and maintain the following account relationship with MyBank:

Account #353342287 – DAG Opportunity Fund, LLC - opened 4/16/15 – closed 7/6/20 - signer is Robert Bogart. Credits to this MyBank Business Premier Checking account are cash deposits, and checks drawn against the accounts of various individuals and other investment fund accounts. A considerable number of cash deposits, totaling $96,400.00, appear to be structured to avoid the filing of a Currency Transaction Report (CTR) and are therefore deemed suspicious. A sample of the structuring activity is as follows:



The Most Complete Advanced Analytics Application for Anti-Money Laundering Teams

Open all +

Graph analytics for entity resolution and interactive visualizations of criminal networks

Uncover and explore previously hidden relationships in real time

- Succinctly express complex money patterns, detect multi-hop relationships, and identify hubs and spokes of activity using 30+ supplied graph algorithms and a built-in, SQL-like query language.
- Intuitive, dynamic graph visualizations in Investigation Hub (PDF) boost the speed and accuracy of investigations.
- Drive better modeling through better understanding of the network.
- Enhance detection by applying deep learning to find similar criminal networks.
- Boost efficiency by applying machine learning to graphs to automate case decisions.
- Preconfigured, extensible entity resolution provides a single view of each customer and entity, and ensures that your global graph is accurate.

Click to enlarge

View the Investigation Hub data sheet (PDF)

Source: Oracle Financial Services Compliance Studio

Related Work: Text to Graph

SciERC

- Benchmark dataset for knowledge graph construction from text
- 500 scientific paper abstracts, annotated with entity, relation, and coreference information
- Entities, relations, coreferences must be extracted/understood to obtain a graph

MODEL	NAMED ENTITY RECOGNITION	RELATION EXTRACTION	COREFERENCE RESOLUTION
SciIE (2018)	64.2	39.3	48.2

F1 Scores

Since 2018, there have been improvements in all three tasks but we were unable to find a framework combining these new approaches into an end-to-end solution

Source: Luan et al. (2018)

Related Work: Transformer Models

- Biggest impact on NLP tasks since 2018: **Transformer Models** (BERT, RoBERTa, XLNet, …)

Name	F1
LSTM-CRF (Lample et al., 2016)	91.0
ELMo (Peters et al., 2018)	92.2
BERT (Devlin et al., 2019)	92.8
Akbik et al. (2018)	93.1
Baevski et al. (2019)	93.5
RoBERTa	92.4
LUKE	**94.3**

Named Entity Recognition

Name	F1
BERT (Zhang et al., 2019)	66.0
C-GCN (Zhang et al., 2018b)	66.4
ERNIE (Zhang et al., 2019)	68.0
SpanBERT (Joshi et al., 2020)	70.8
MTB (Baldini Soares et al., 2019)	71.5
KnowBERT (Peters et al., 2019)	71.5
KEPLER (Wang et al., 2019b)	71.7
K-Adapter (Wang et al., 2020)	72.0
RoBERTa (Wang et al., 2020)	71.3
LUKE	**72.7**

Relation Extraction

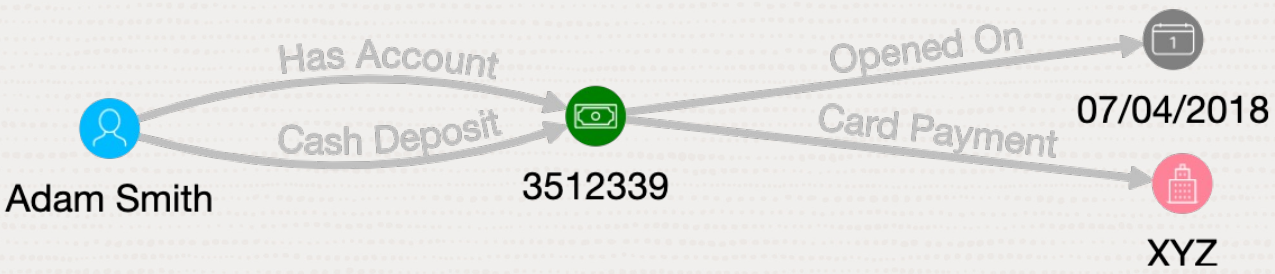
	Avg. F1
e2e-coref(Lee et al., 2017)	67.2
c2f-coref + ELMo (Lee et al., 2018)	73.0
EE + BERT-large (Kantor and Globerson, 2019)	76.6
c2f-coref + BERT-large (Joshi et al., 2019b)	76.9
c2f-coref + SpanBERT-large (Joshi et al., 2019a)	79.6
CorefQA + SpanBERT-base	79.9 (+0.3)
CorefQA + SpanBERT-large	**83.1** (+3.5)

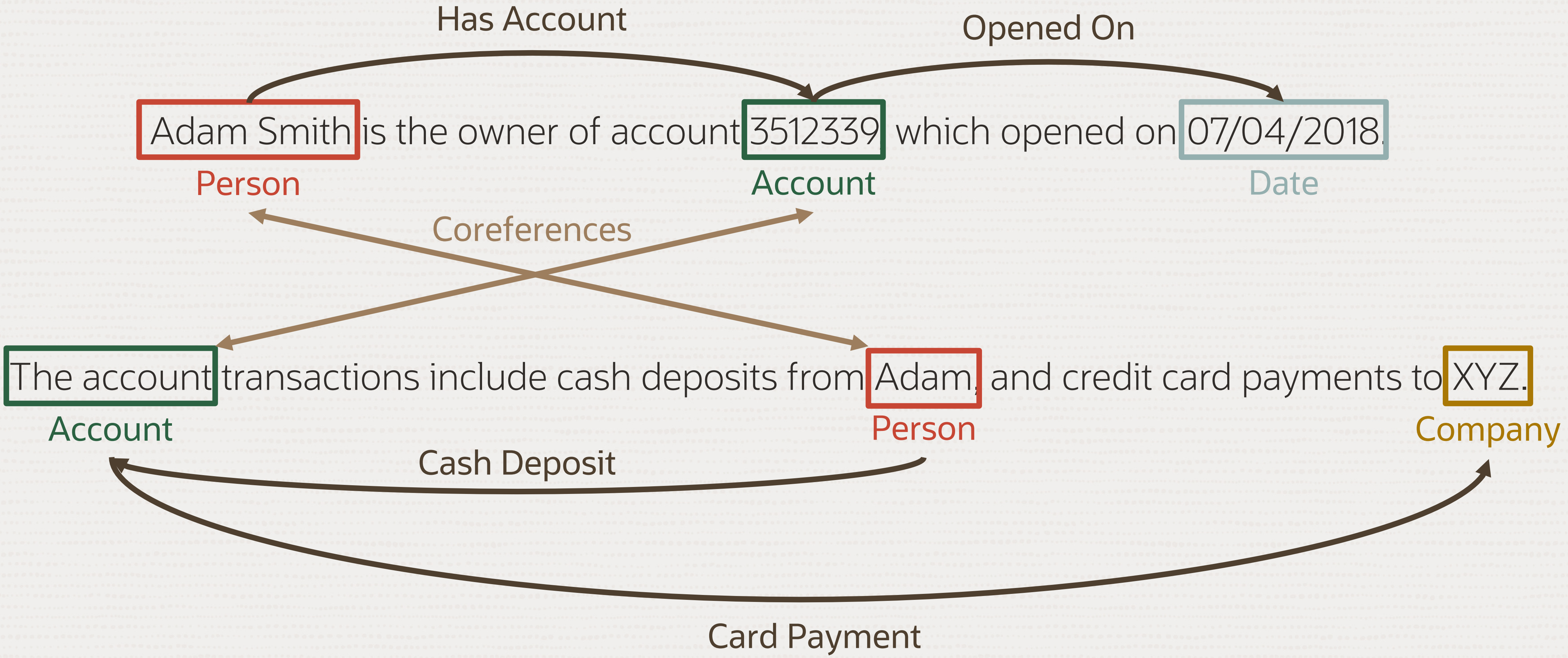
Coreference Resolution

- The strongest models often consist of Large Pretrained Transformer + Simple Classifier.
- We aimed to obtain a state-of-the-art pipeline using these modern approaches.

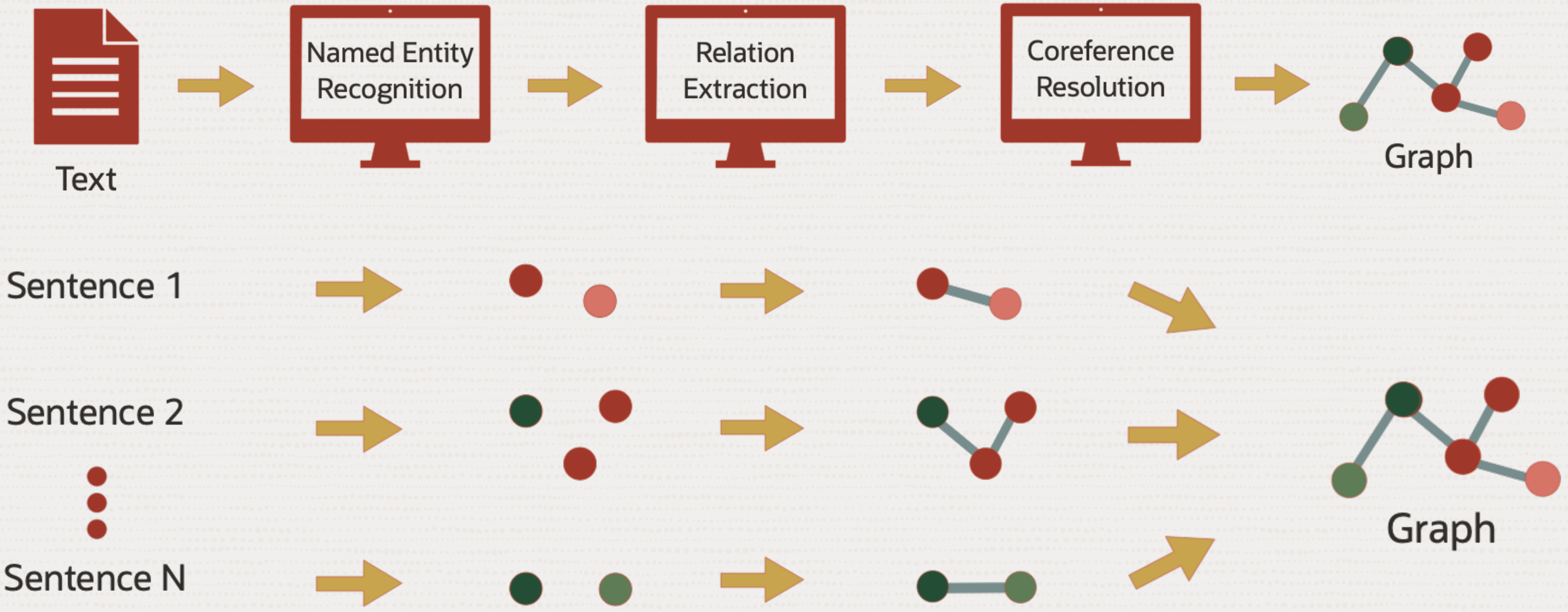
Sources: LUKE, CorefQA

Approach: Example





Approach: High-level Overview



Approach: Named Entity Recognition (NER)

Human Perspective

Adam Smith is the owner of account 3512339 which opened on 07/04/2018

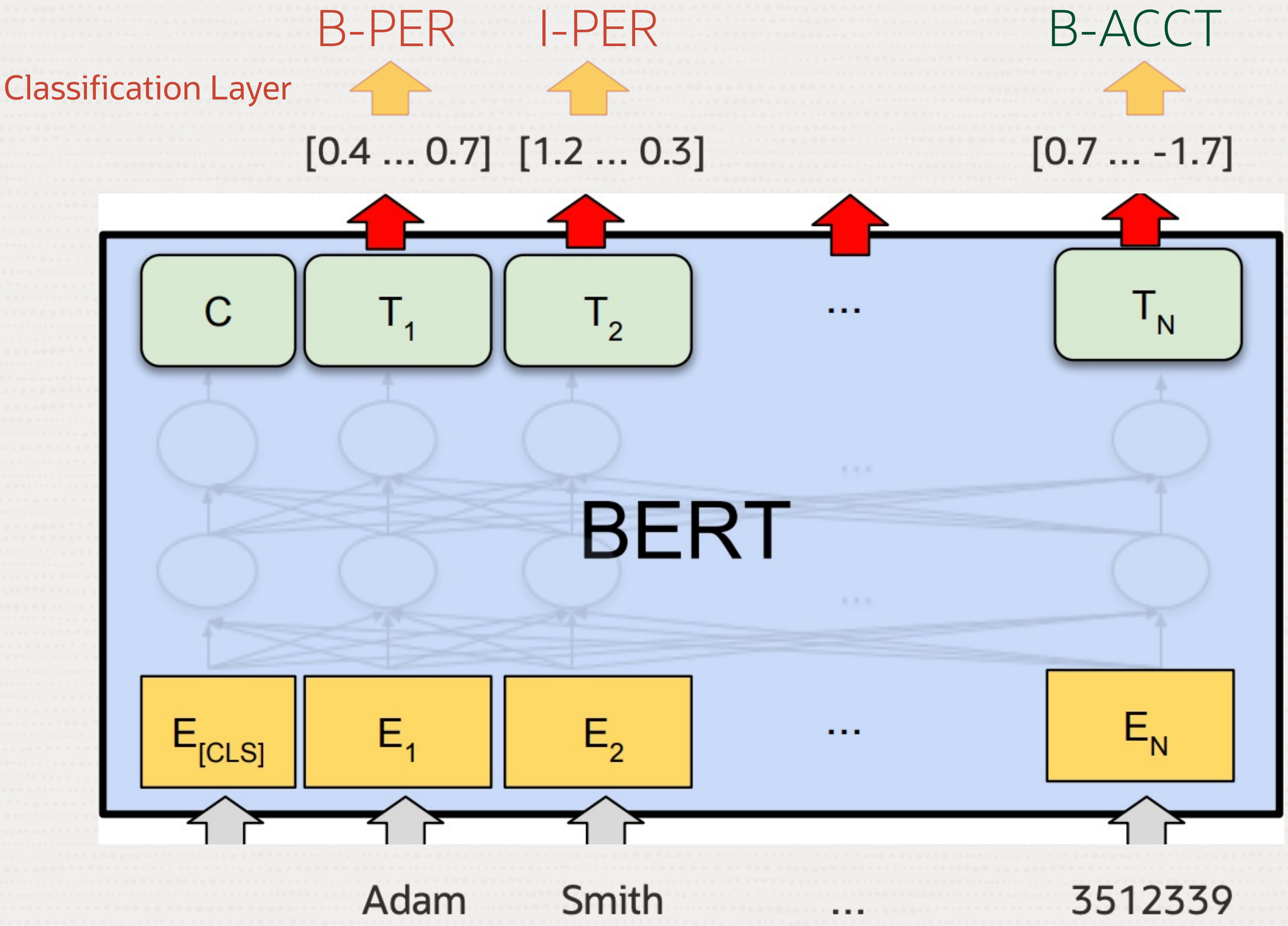
Person **Account** **Date**

Model Perspective

Adam	Smith	is	the	owner	of	account	3512339	,	which	opened	on	07/04/2018	.
B-PER	I-PER	O	O	O	O	O	B-ACCT	O	O	O	O	B-DATE	O

Approach: Named Entity Recognition (NER)

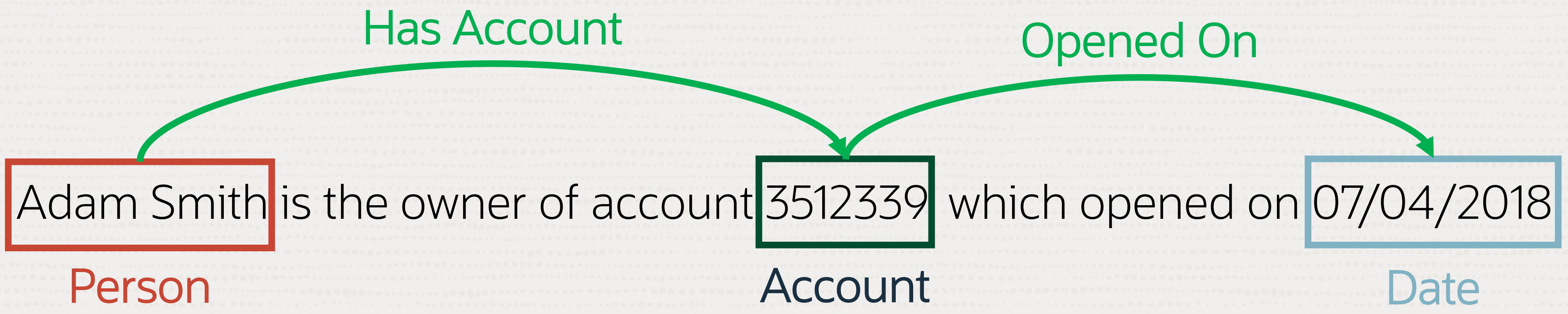
- Transformer models have 100s of millions of parameters
- Pre-trained on a language modeling task using a dataset of billions of words
- They can capture information about syntax and semantics of natural language
- Can be fine-tuned for a given task with labeled data.



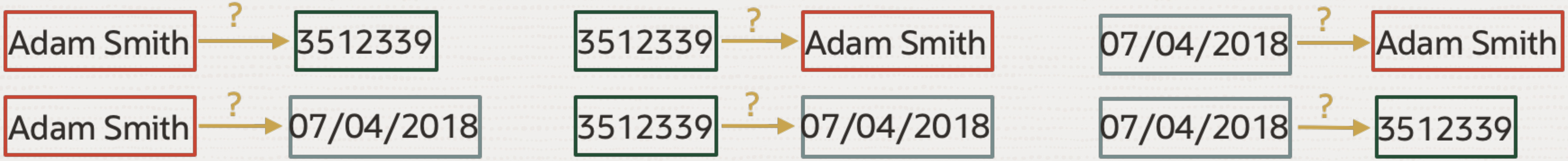
Source: BERT

Approach: Relation Extraction (RE)

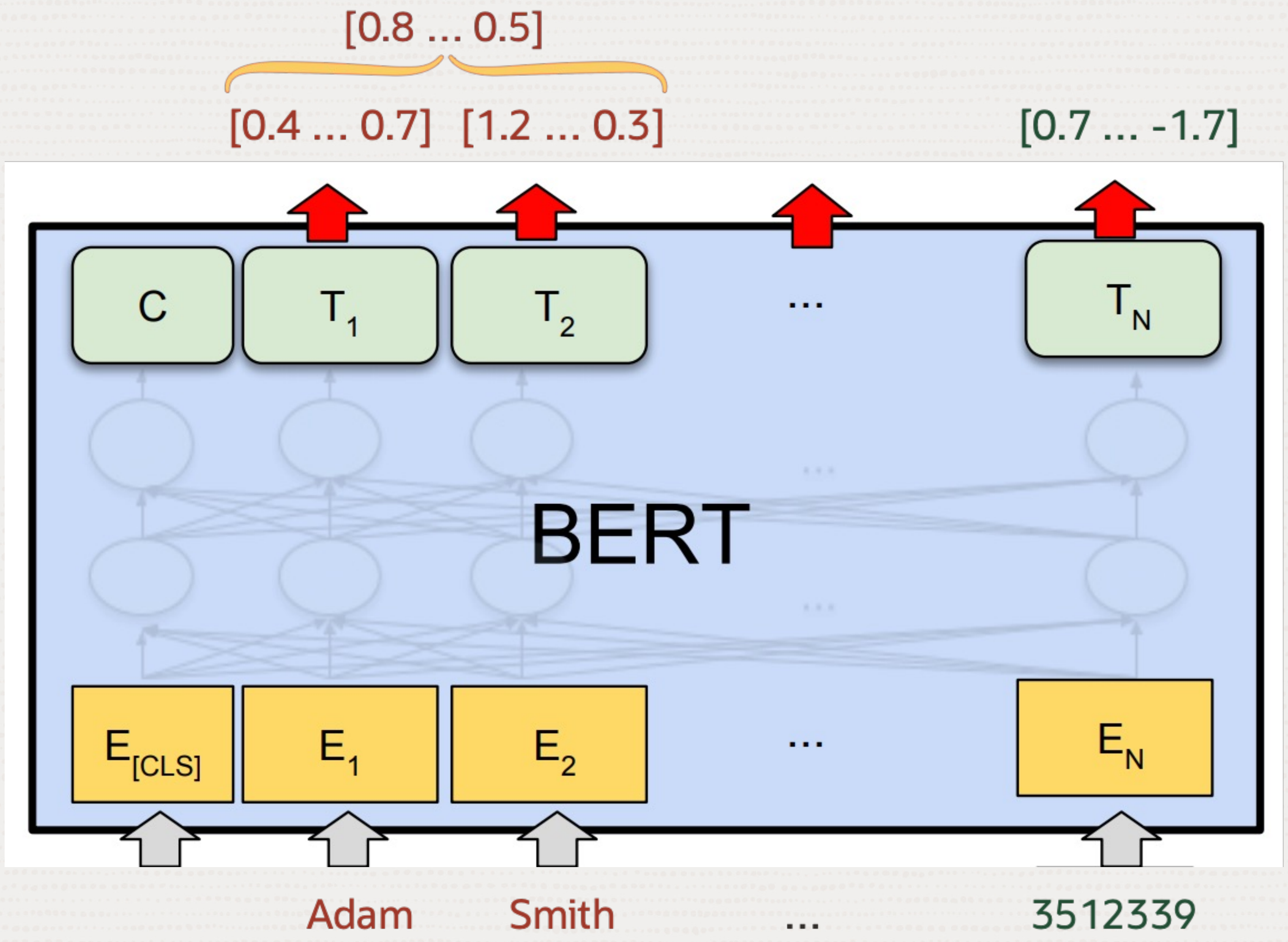
Human Perspective



Model Perspective



Approach: Relation Extraction (RE)



Source: BERT

Approach: Coreference Resolution (CR)

Human Perspective

Adam Smith is the owner of account 3512339 which opened on 07/04/2018

Person Account Date

Coreferences

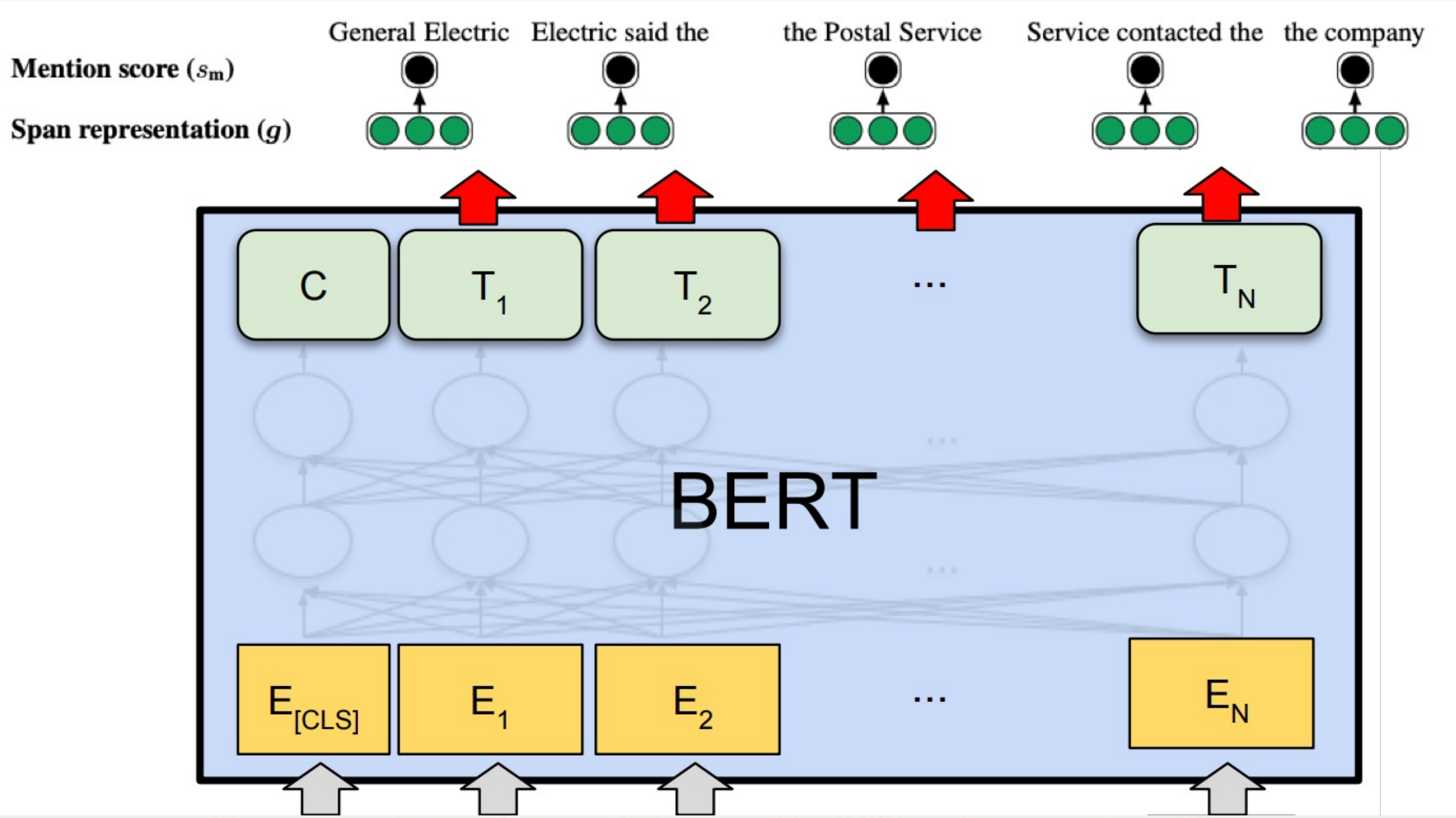
The account transactions include cash deposits from Adam, and credit card payments to XYZ

Account Person Company

Approach: Coreference Resolution (CR)

Model Perspective:

We first need to find “mentions”, spans of words that might be referring to an entity.



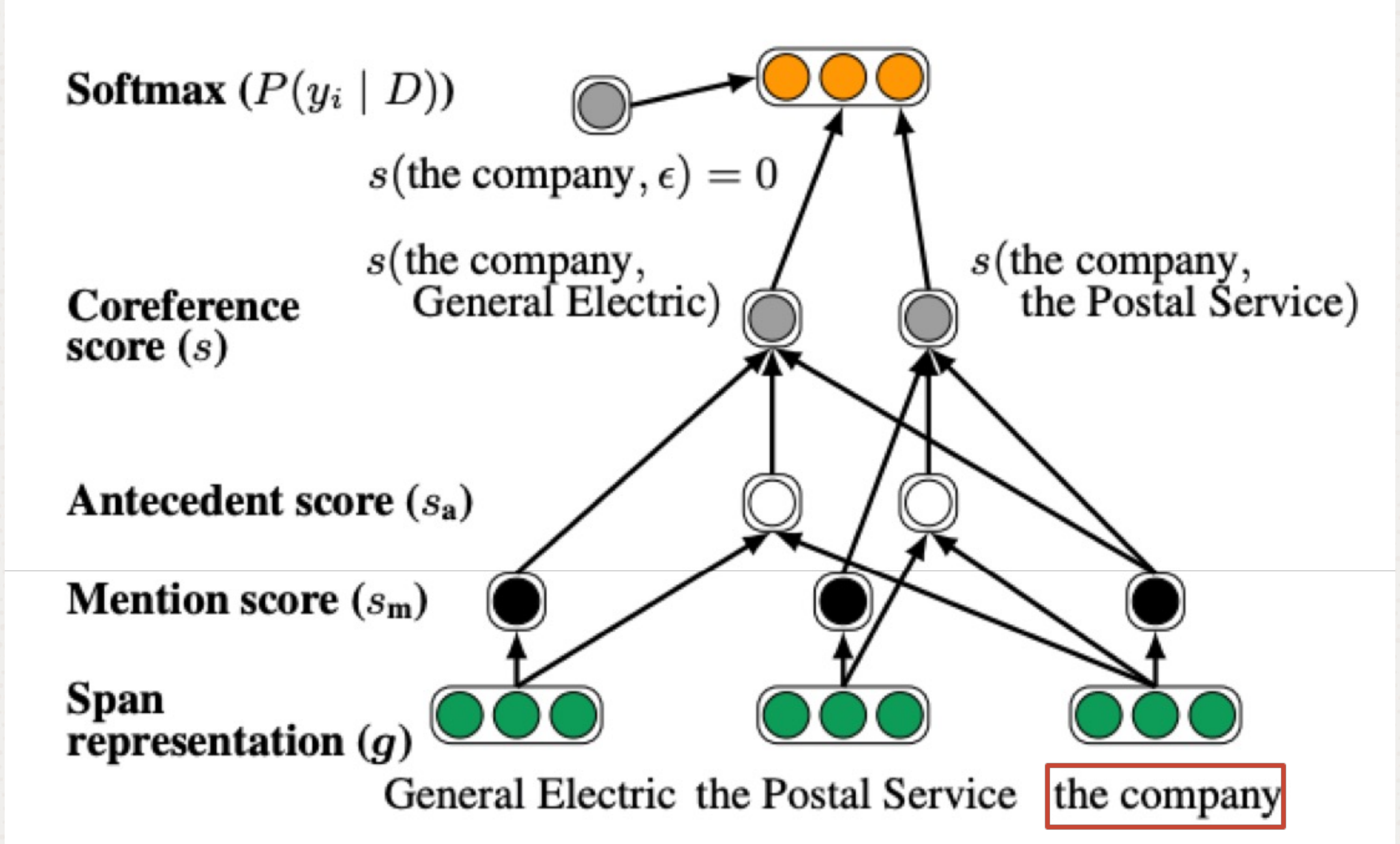
…General Electric said the Postal Service contacted the company…

Source: BERT, e2e-coref

Approach: Coreference Resolution (CR)

Model Perspective:

Then we can try to determine for each mention which earlier mentions are likely referring to the same entity ("antecedents").



What are the antecedents of this span?

Source: e2e-coref

Data: Public Datasets

Several datasets are used to validate the performance of the framework:

- CoNLL-2004: Named Entity and Relation Extraction on news sentences
- CoNLL-2012: Coreference Resolution on news, speech, broadcast, etc.
- SciERC: Scientific knowledge graph construction from scientific paper abstracts

Additionally, we also wanted to test Text to Graph in the financial domain for our use case.

Data: Suspicious Activity Reports (SAR)

Introduction:

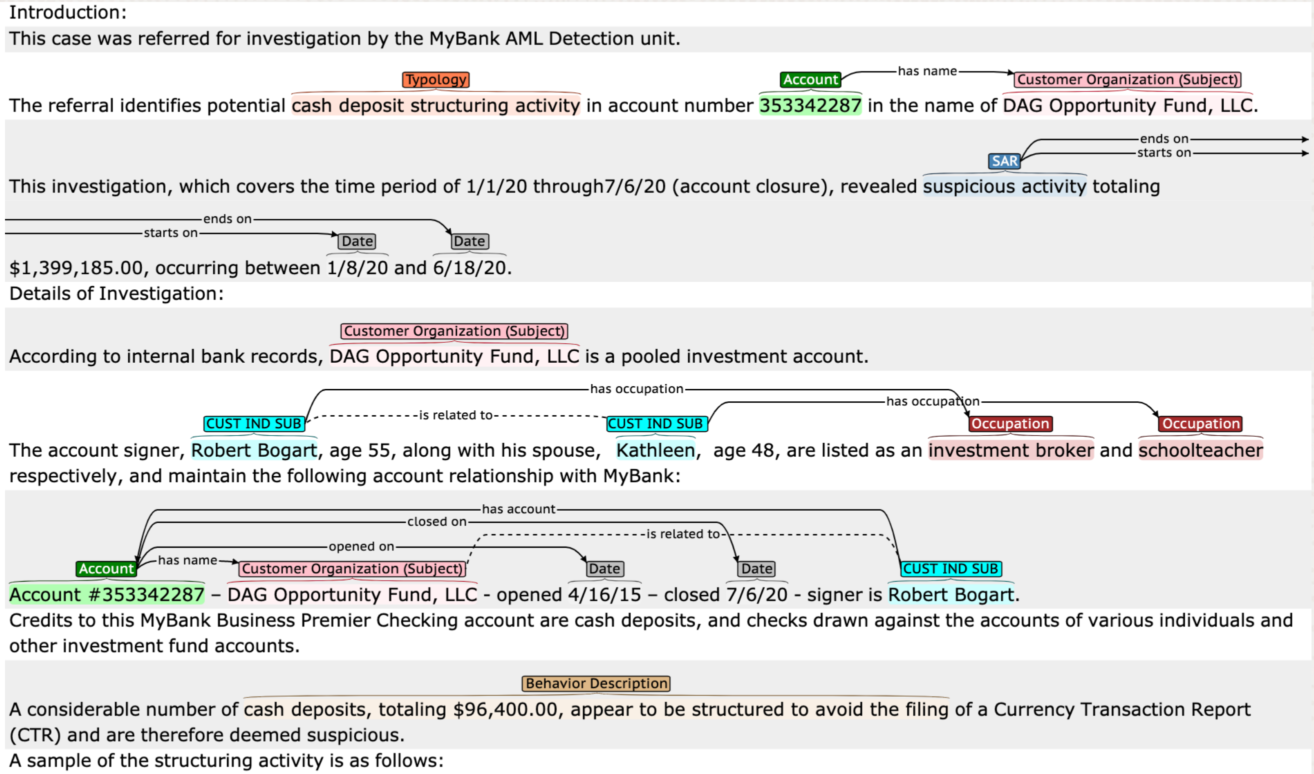
This case was referred for investigation by the MyBank AML Detection unit. The referral identifies potential cash deposit structuring activity in account number 353342287 in the name of DAG Opportunity Fund, LLC.

This investigation, which covers the time period of 1/1/20 through7/6/20 (account closure), revealed suspicious activity totaling $1,399,185.00, occurring between 1/8/20 and 6/18/20.

Details of Investigation:

According to internal bank records, DAG Opportunity Fund, LLC is a pooled investment account. The account signer, Robert Bogart, age 55, along with his spouse, Kathleen, age 48, are listed as an investment broker and schoolteacher respectively, and maintain the following account relationship with MyBank:

Account #353342287 – DAG Opportunity Fund, LLC - opened 4/16/15 – closed 7/6/20 - signer is Robert Bogart. Credits to this MyBank Business Premier Checking account are cash deposits, and checks drawn against the accounts of various individuals and other investment fund accounts. A considerable number of cash deposits, totaling $96,400.00, appear to be structured to avoid the filing of a Currency Transaction Report (CTR) and are therefore deemed suspicious. A sample of the structuring activity is as follows:



105 synthetic SAR documents were annotated by two subject matter experts using the annotation tool brat

Source: brat

Results: SciERC

MODEL	NAMED ENTITY RECOGNITION	RELATION EXTRACTION	COREFERENCE RESOLUTION
SciIE (2018)	64.2	39.3	48.2
Our Framework	70.1	50.5	60.2
Improvement	+5.9	+11.2	+12.0

F1 Scores

There are now better scores published for individual tasks on this dataset, e.g.:

- NER: 71.1 (Jeong and Kim, 2022)
- RE: 51.3 (Santosh et al., 2021)

But the state-of-the-art has not moved too far and among frameworks that can tackle the end-to-end task, these are still competitive results.

Results: SAR

NLP Metrics

NAMED ENTITY RECOGNITION	RELATION EXTRACTION	COREFERENCE RESOLUTION
88.82	81.26	87.06

F1 Scores

Scores are higher in this dataset due to the more structured documents.

Graph Quality Metrics

NODES	EDGES
83.12	82.13

F1 Scores

High NLP performance translates successfully into a high-quality graph.

Demo: End to End

Text to Graph Demo

This research has been done for the Oracle Financial Services Compliance Studio (**OFS Compliance Studio**), which is an integrated, notebook-based platform for financial crime analysis.

The goal is to research/implement/improve methods to transform text documents into knowledge graphs. Why?

- Visual inspection: seeing information presented graphically as connections between nodes can help us see patterns that are easy to miss in plain text.
- Downstream algorithms: it is hard to apply certain algorithms to unstructured text, but if we obtain a graph representation we can apply many rule-based and ML-based graph algorithms to achieve downstream tasks.

To achieve that, we implement a Natural Language Processing (NLP) Machine Learning pipeline with the following stages:

- The NER stage recognizes entities (vertices) from the text document.
- The Relation Extraction stage recognizes links (edges) between entities.
- The Coreference Resolution stage recognizes references in the text that refer to previously seen entities.

All stages use Machine Learning, mainly based on Large Transformers + Simple Classifiers.

Model Setup

Input Text

Text

John Doe works as a research intern at Oracle Labs. John is the current signer of bank account #35301252 – opened on 11/12/20. Credits to this personal student account include payroll deposits from Oracle Labs. Debits to this account consist of card payments to Credit Suisse and Migros Bank.

You can also interact with your graphs in a notebook platform: Graph Studio

Demo: Step 1

Text is split into sentences.

```
%python

text = '''Doga Tekin works as a research intern at Oracle Labs. Doga is the current signer of bank account #35301252
display_sentences(text)
```

Doga Tekin works as a research intern at Oracle Labs.

Doga is the current signer of bank account #35301252 – opened on 11/12/20.

Credits to this personal student account include payroll deposits from Oracle Labs.

Debits to this account consist of card payments to Credit Suisse and Migros Bank.

Demo: Step 2

- Sentences go through NER and RE.
- Document goes through CR.

```
%python
text2graph(text)
```

Named Entity Recognition & Relation Extraction:

Doga Tekin **Individual_Sub** works as a research intern **Occupation** at Oracle Labs **Derived_Entity** .

Doga Tekin -[has_occupation]-> research intern

Doga **Individual_Sub** is the current signer of bank account # 35301252 **Account** – opened on 11/12/20 **Date** .

Doga -[has_account]-> 35301252
35301252 -[opened_on]-> 11/12/20

Credits to this personal student **Account_Type** account **Account** include payroll deposits from Oracle Labs **Derived_Entity** .

account -[has_type]-> personal student
Oracle Labs -[credit_transaction]-> account

Debits to this account **Account** consist of card payments to Credit Suisse **Institution** and Migros Bank **Institution** .

this account -[wire_transaction]-> Credit Suisse
this account -[wire_transaction]-> Migros Bank

Coreference Resolution:

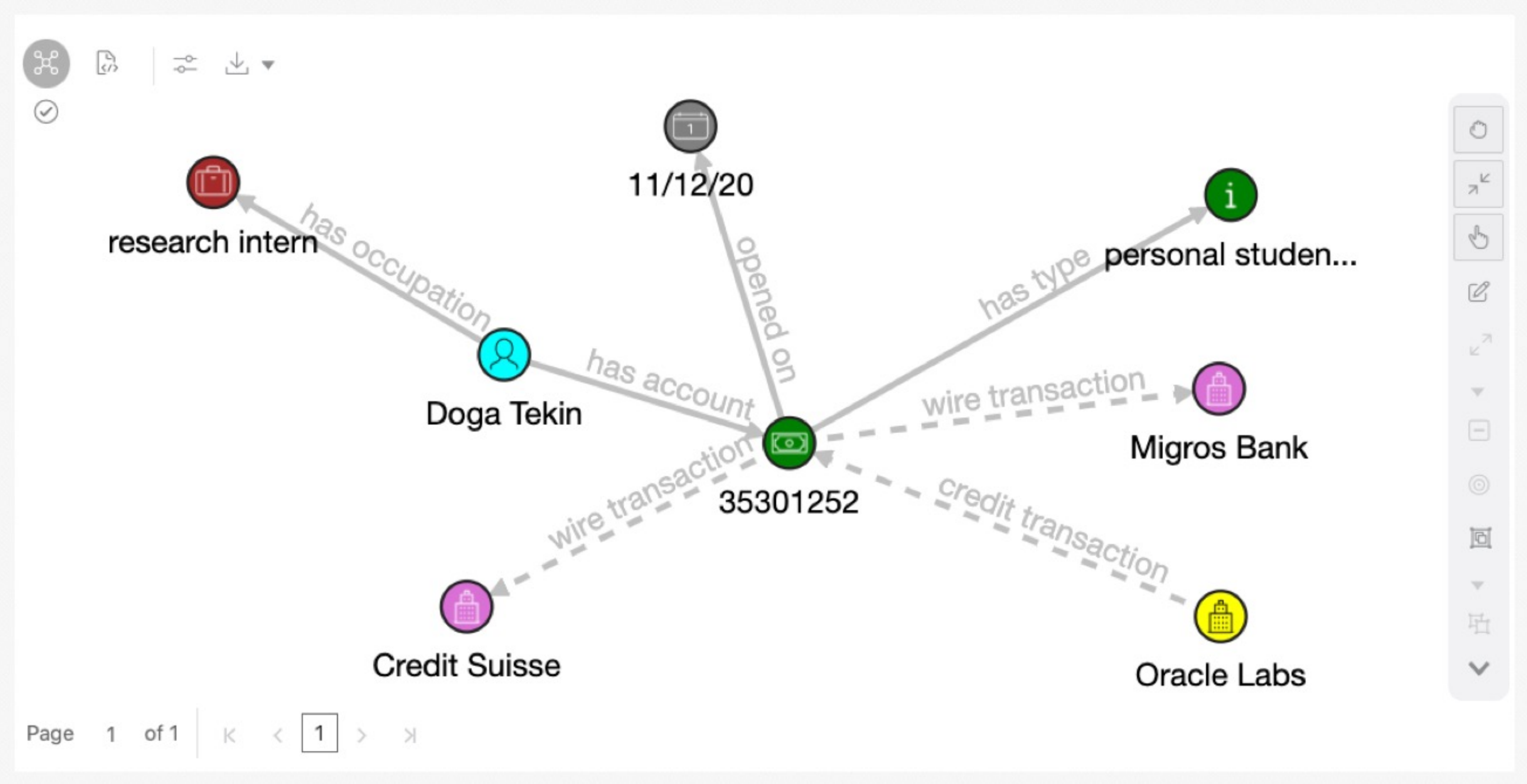
Doga Tekin = Doga

35301252 = account = this account

Oracle Labs = Oracle Labs

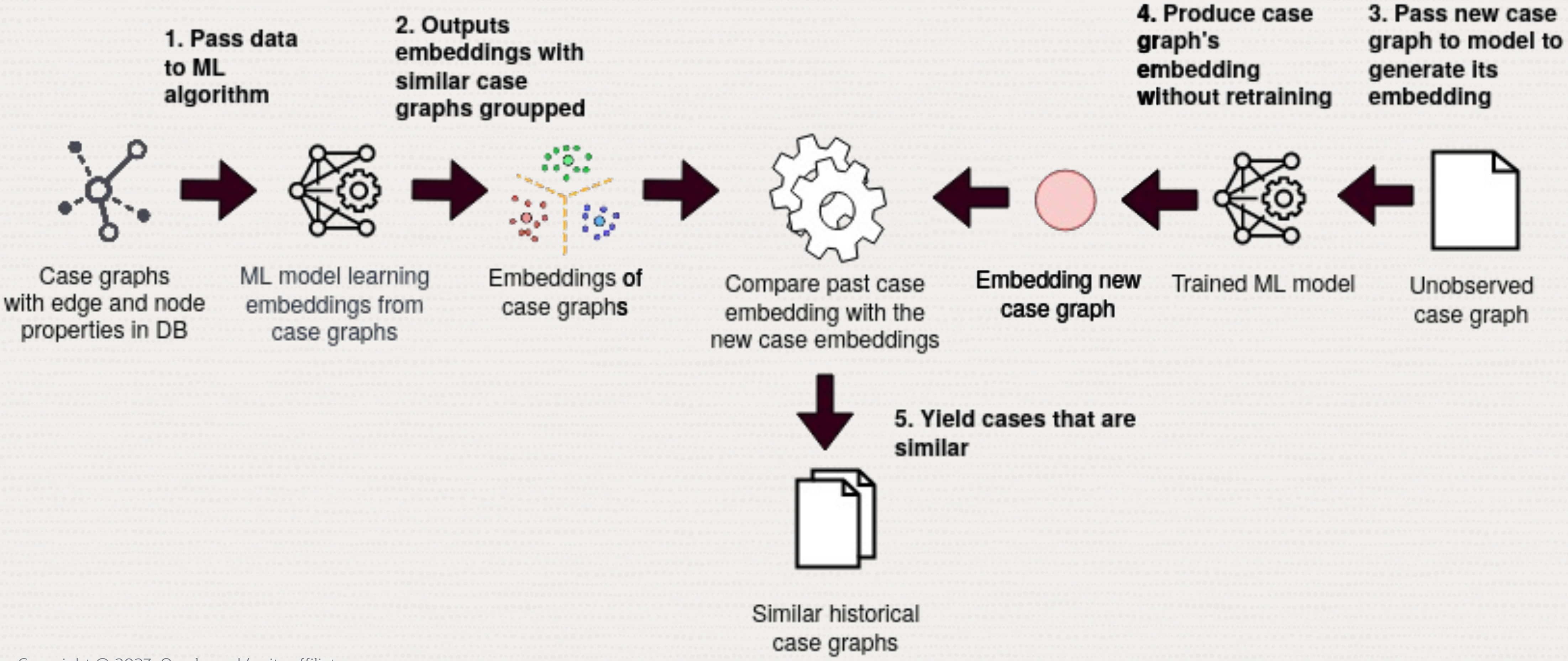
Demo: Step 3

Entities, relations and coreferences are transformed into a graph.



Using the Graphs to Fight Financial Crime

One way to aid investigators is to find historical cases similar to a new investigation graph.

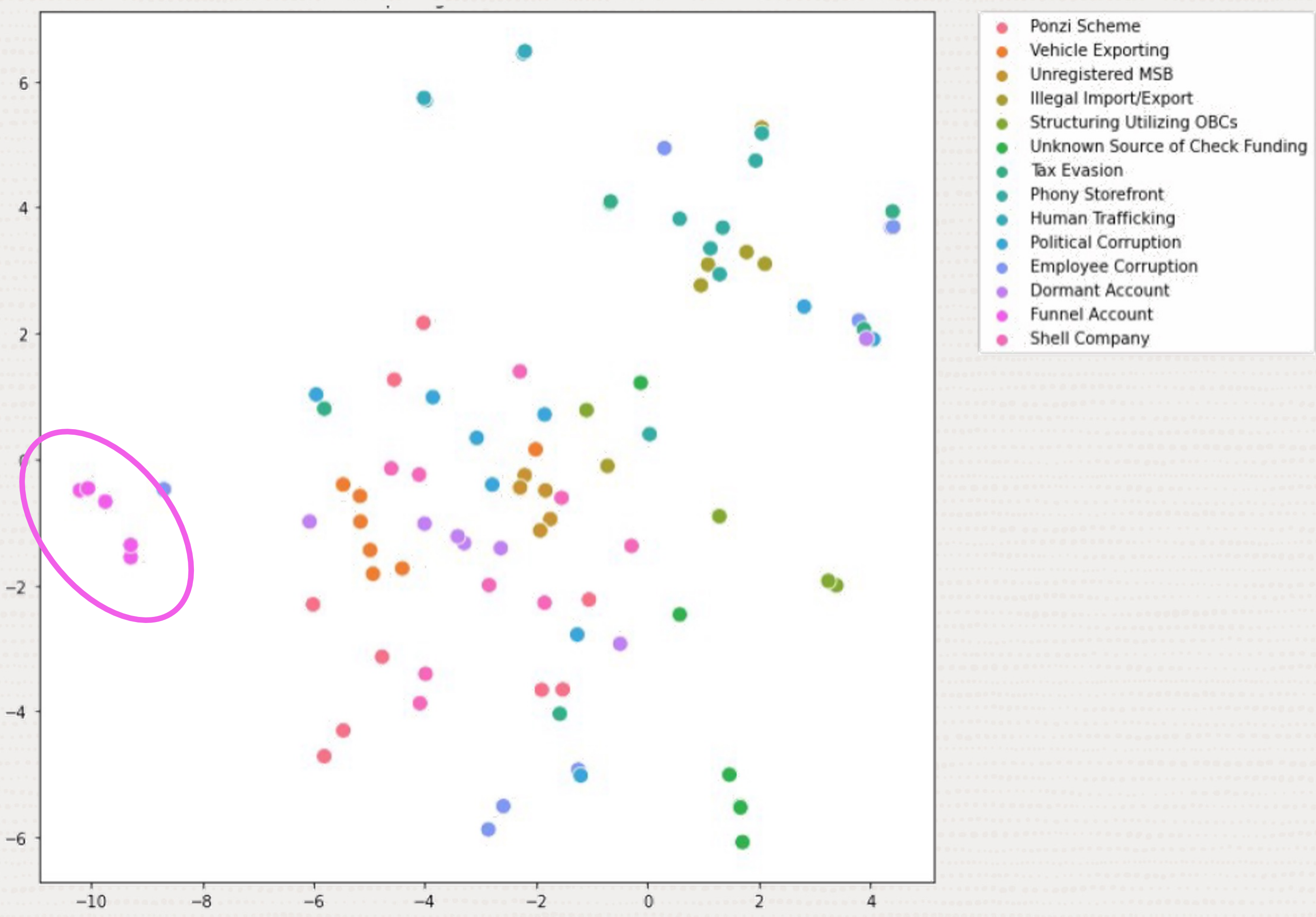


Learn more: [Using the Machine Learning Library (PgxML) for Graphs]

Using the Graphs to Fight Financial Crime

- Graph Machine Learning models can learn representations of case graphs taking into account **both the graph structures and the features of nodes and edges.**
- These representations can be used to cluster similar fraud types together.

Funnel accounts cluster



Takeaways

- Knowledge graphs provide a powerful way to represent and visualize your data, enable novel analytics approaches and uncover useful insights.
  - Oracle Graph already has many features to help you achieve this potential.
  - Oracle Financial Services Compliance Studio uses those features to fight financial crime.
- Graphs can be obtained from text documents automatically by using modern natural language processing techniques such as fine-tuning language models.
  - Work presented here is ongoing research but Oracle already offers some AI Language services publicly.
- Annotating documents is necessary to teach language models your desired graph schema, but pre-trained transformer models reduce the annotation efforts greatly.
  - Important future work to reduce this even further!

Thank you

Feel free to reach out to me with your questions!

Doga Tekin, Member of Technical Staff @ Oracle Labs
doga.tekin@oracle.com