

MULTIMODAL INFORMATION RETRIEVAL AND UNCERTAINTY MANAGEMENT IN OPEN-WORLD ENVIRONMENT

Examining Committee:

- Bharat Bhargava (Advisor)
- Chunyi Peng
- Vaneet Aggarwal
- Jianguo Wang
- Xavier Tricoche

KMA Solaiman
Ph.D. Dissertation Defense

July 14, 2023

Collaborations and Acknowledgements

- I am very thankful to Michael Stonebraker from MIT for being a mentor during REALM.
- Thanks to MIT team and Michael Cafarella from University of Michigan during our work in SKOD.
- Jim MacDonald from NGC for leading REALM and always helping us with innovative research directions
- Thanks to Sgt. Greene from West Lafayette Police Department for providing data, discussing use cases, and helping to analyze the problem domain such as missing person, school shooting, etc.
- Novelty Working Group with Mayank Kejriwal, Terry Boult, Josh AlSpector, Eric Kildebeck, Pat Langley, Katarina Doctor.
- Thanks to my colleagues Alina, Palash, Servio, Denis and others from Purdue, and Tao, Aaron, Zack from MIT for our collaborations on building the prototypes.
- Thanks to my colleagues Shafkat and Ruy for our collaborations on open-world novelties to conduct further experiments.



Massachusetts Institute of Technology



Stanford
University

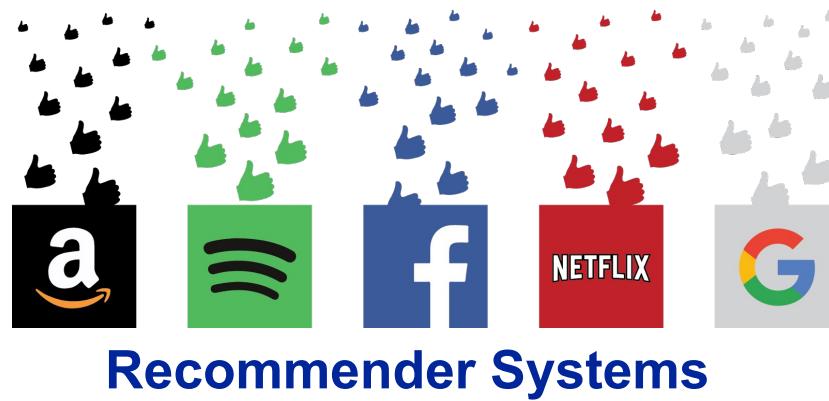
Publications and Artifacts

1. K. Solaiman et al. [***Feature Centric Multi-modal Information Retrieval in Open World Environment \(FemmlR\)***](#), SIGMOD 2023 (positive reviews) (To be submitted in VLDB 2024).
2. S. Islam and K. Solaiman**, R. Oliveira, B. Bhargava, [***Domain Complexity Estimation for Distributed AI Systems in Open-World Perception Domain***](#), Artificial Intelligence (Open-World AI), July 2023. ** Co-first authors.
3. K. Solaiman, T. Sun, A. Nesen, B. Bhargava and M. Stonebraker, "[***Applying Machine Learning and Data Fusion to the Missing Person Problem***](#)" in IEEE Computer, vol. 55, no. 06, pp. 40-55, 2022.
4. K. Solaiman and B. Bhargava, [***Open-Learning Framework for Multi-modal Information Retrieval with Weakly Supervised Joint Embedding***](#), AAAI Spring Symposium on Designing Artificial Intelligence for Open Worlds, March 2022.
5. K. Solaiman and B. Bhargava, [***Measurement of Novelty Difficulty in Monopoly***](#), AAAI Spring Symposium on Designing Artificial Intelligence for Open Worlds, March 2022.
6. A. Nesen, K. Solaiman and B. Bhargava, [***Dataset Augmentation with Generated Novelties***](#), IEEE TransAI, 2021.
7. M. Stonebraker, B. Bhargava, M. Cafarella ,Z. Collins, A. Sipser, T. Sun, J. McClellan, A. Nesen, K. Solaiman, G. Mani, K. Kochpatcharin and P. Angin, J. MacDonald, [***Surveillance Video Querying With A Human-in-the-Loop***](#), In Human-In-the-Loop Data Analytics (HILDA 2020) with SIGMOD conference, 2020.
8. S. Palacios and K. Solaiman** et al. [***SKOD: A Framework for Situational Knowledge on Demand***](#). In Polystores and other Systems for Heterogeneous Data (Poly 2019), at VLDB 2019, LA, California, August 2019. ** Co-first authors.

Artifacts

1. Find-Them, (https://youtu.be/hJ_jtLQUIXo), IEEE Computer 2022.
2. SurvQ Extension, <https://youtu.be/z9iJyGrFBtg>, 2023.
3. SurvQ, <https://github.com/skod-ng/>, SIGMOD 2020.
4. SurvQ Demo (<https://youtu.be/qPO73mGXqds>), SIGMOD 2020.
5. SKOD, <https://github.com/purdue-gask>, VLDB 2019.
6. SKOD Demo (<https://youtu.be/5TqWKzy5Sql>), VLDB 2019.

Background and Motivation



Google search results for "barack obama":

- Did Michelle Obama File for Divorce After Barack Admitted He Was Gay? (Snopes.com)
- A Trump Rescue for the Obama Presidential Center? (The Wall Street Journal)
- Fox New Contributor: Trump Giving Rush Limbaugh Medal Of Freedom Is Same as... (Mediaite)

Evernote sidebar:

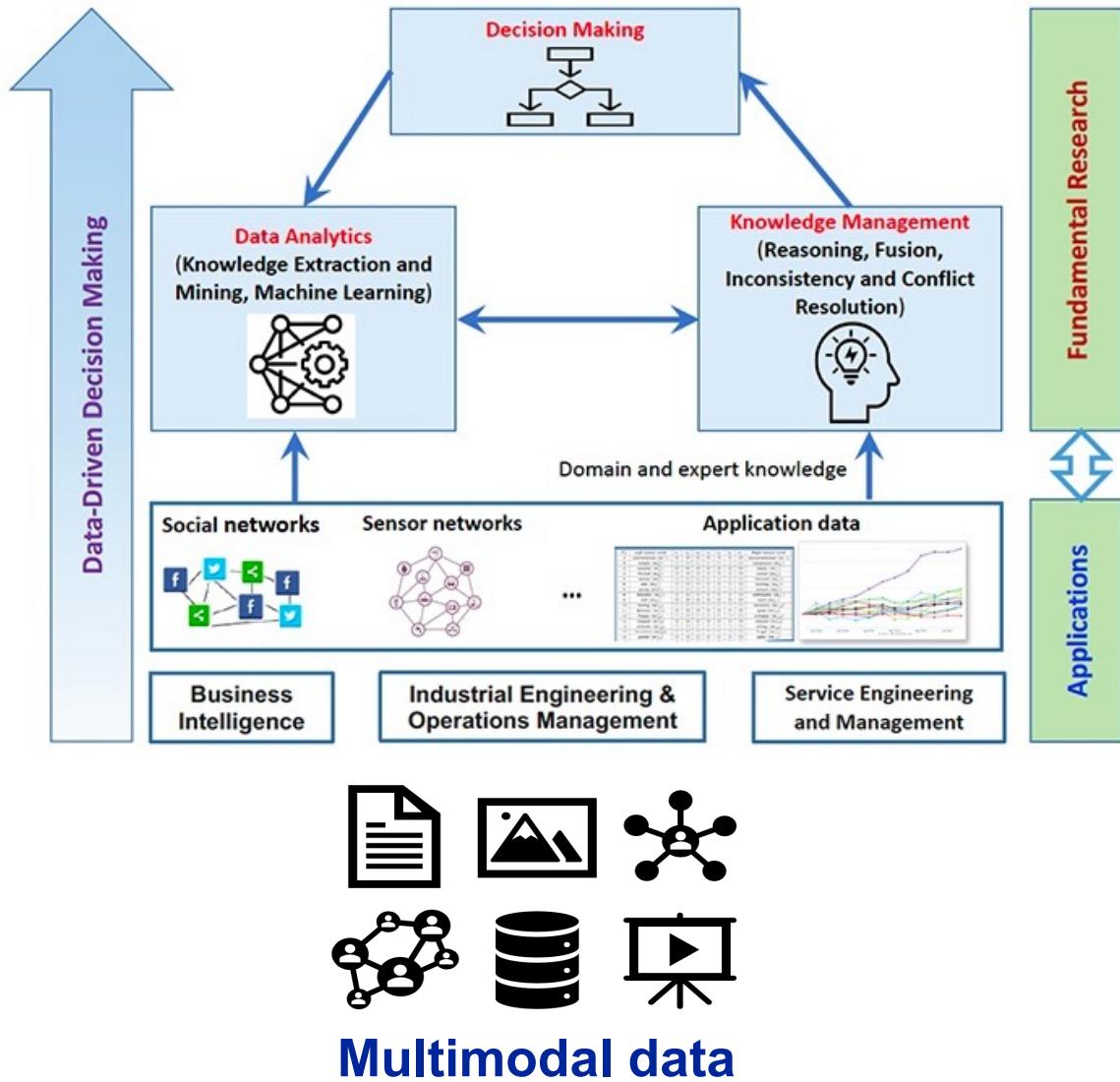
Barack Obama
44th U.S. President
Barack Hussein Obama II is an American attorney and politician who served as the 44th president of the United States from 2009 to 2017. A member of the Democratic Party, he was the first African American president of the United States. [Wikipedia](#)

Presidential term: January 20, 2009 – January 20, 2017 **Trending**

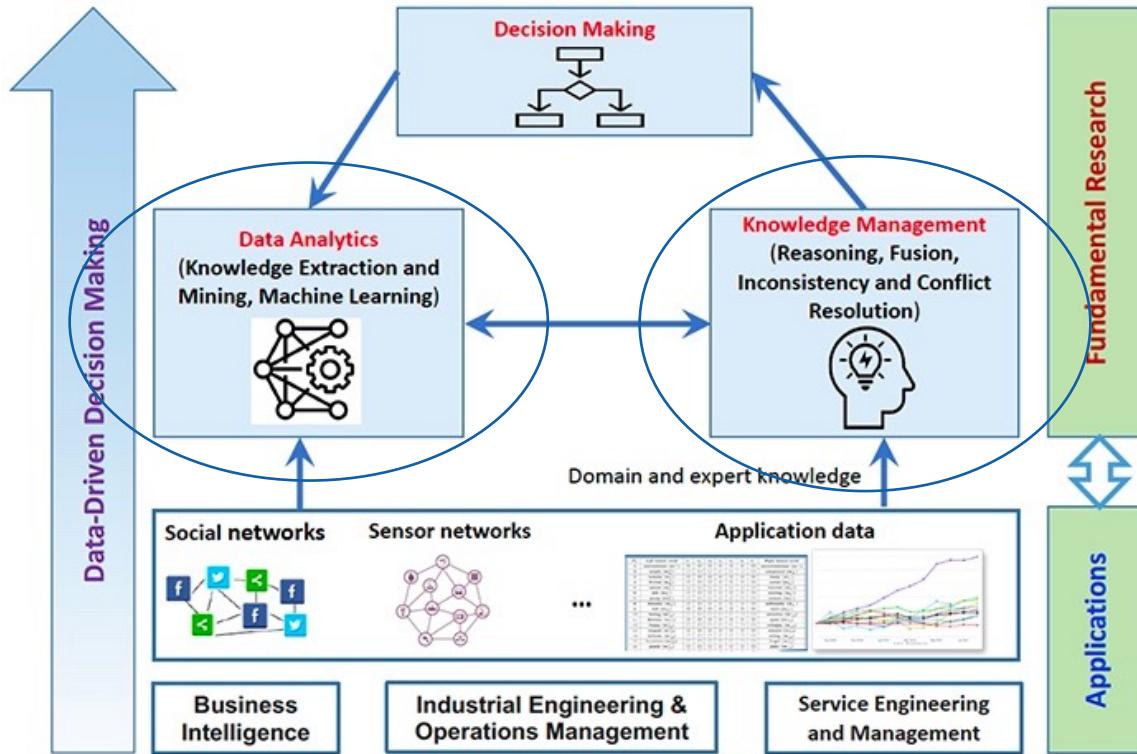
Born: August 4, 1961 (age 58 years), [Kapi'olani](#)

Background and Motivation

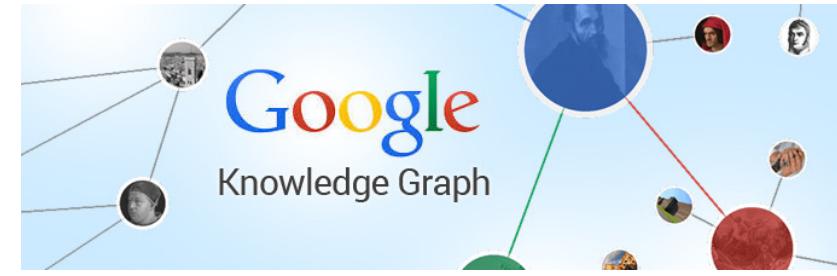
Background and Motivation



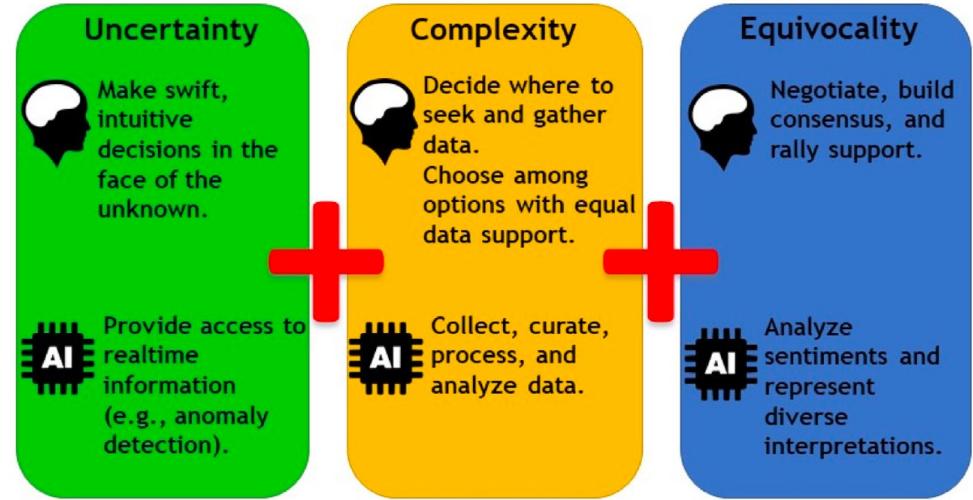
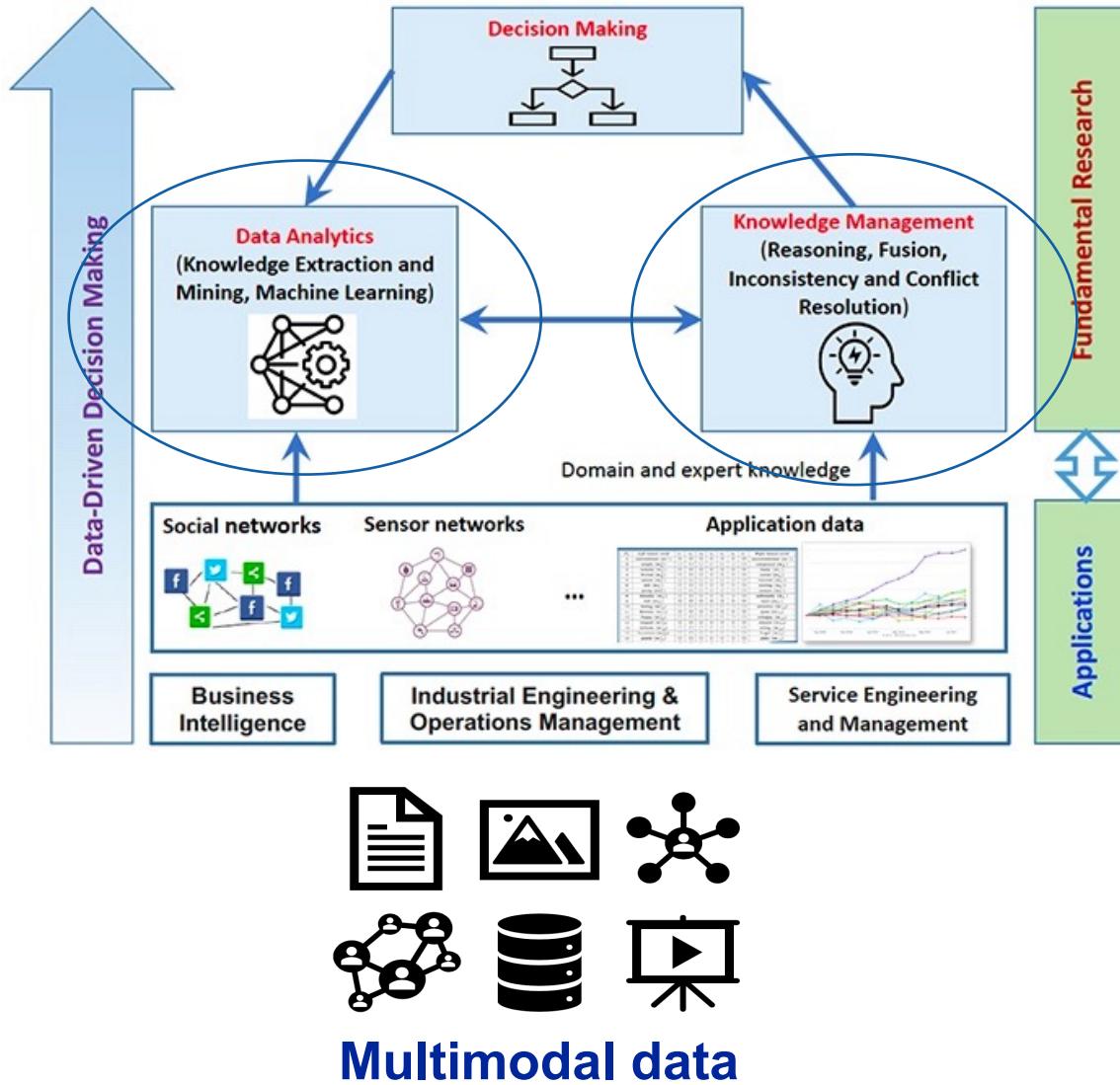
Background and Motivation



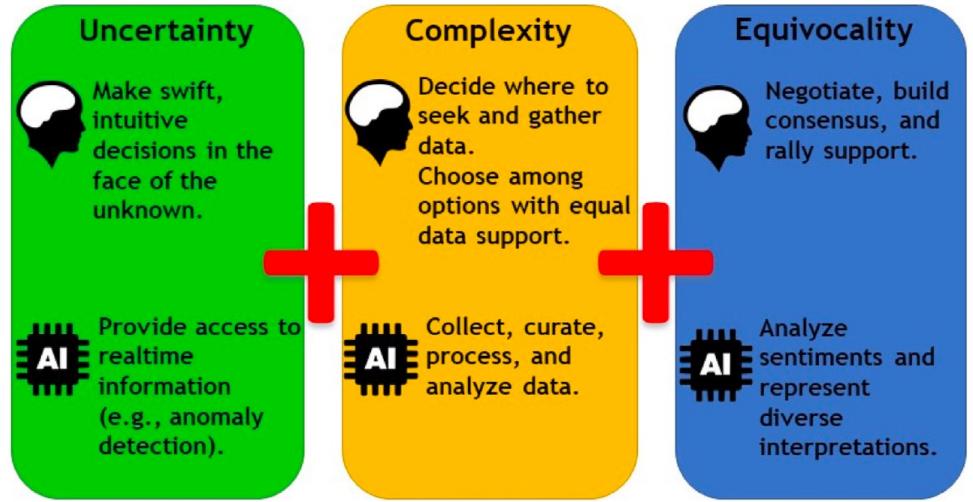
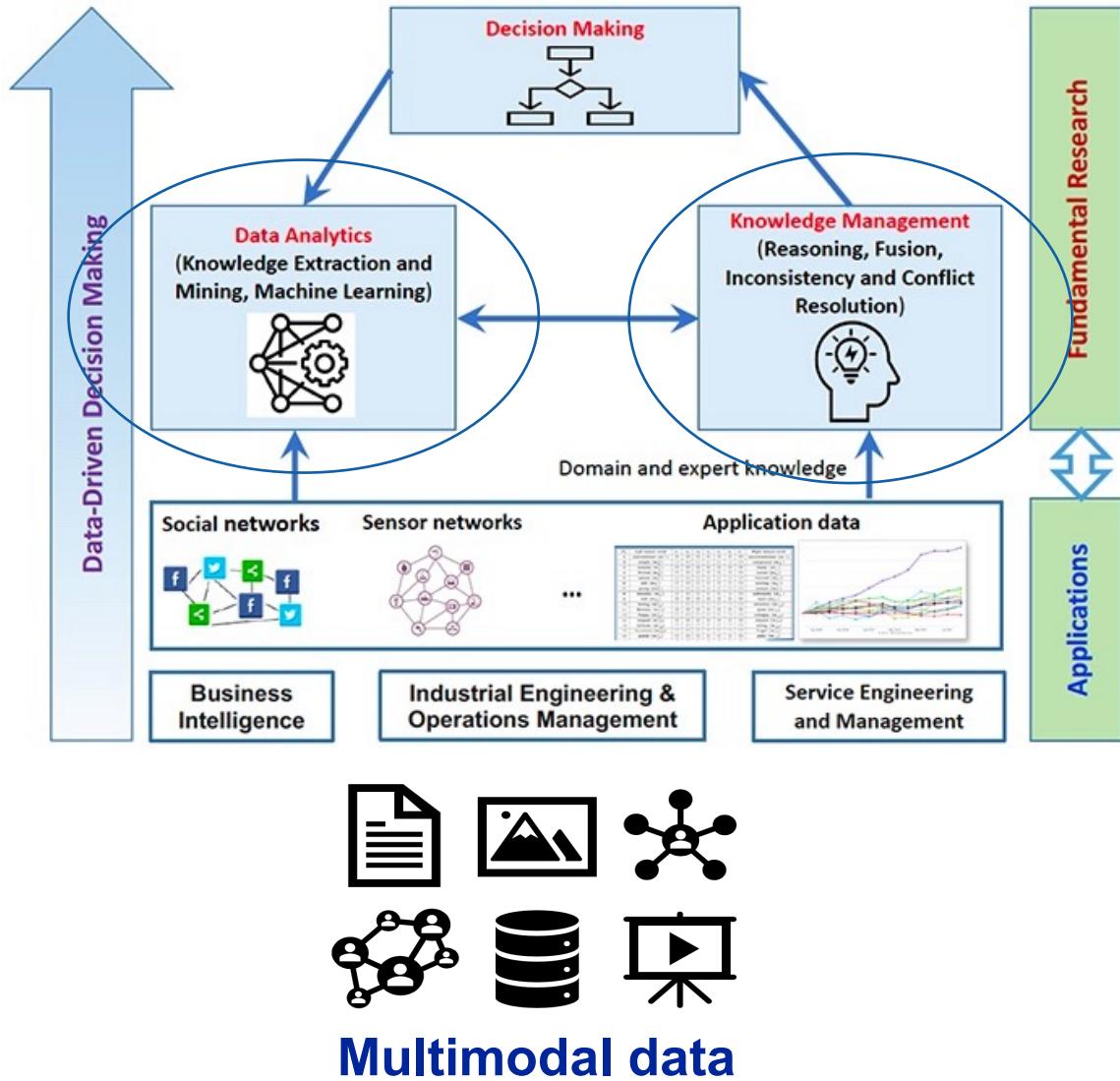
Multimodal data



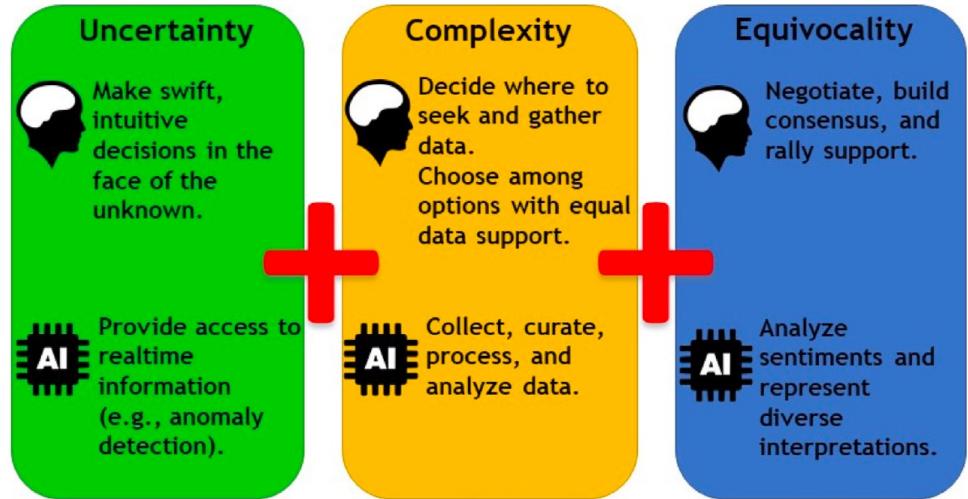
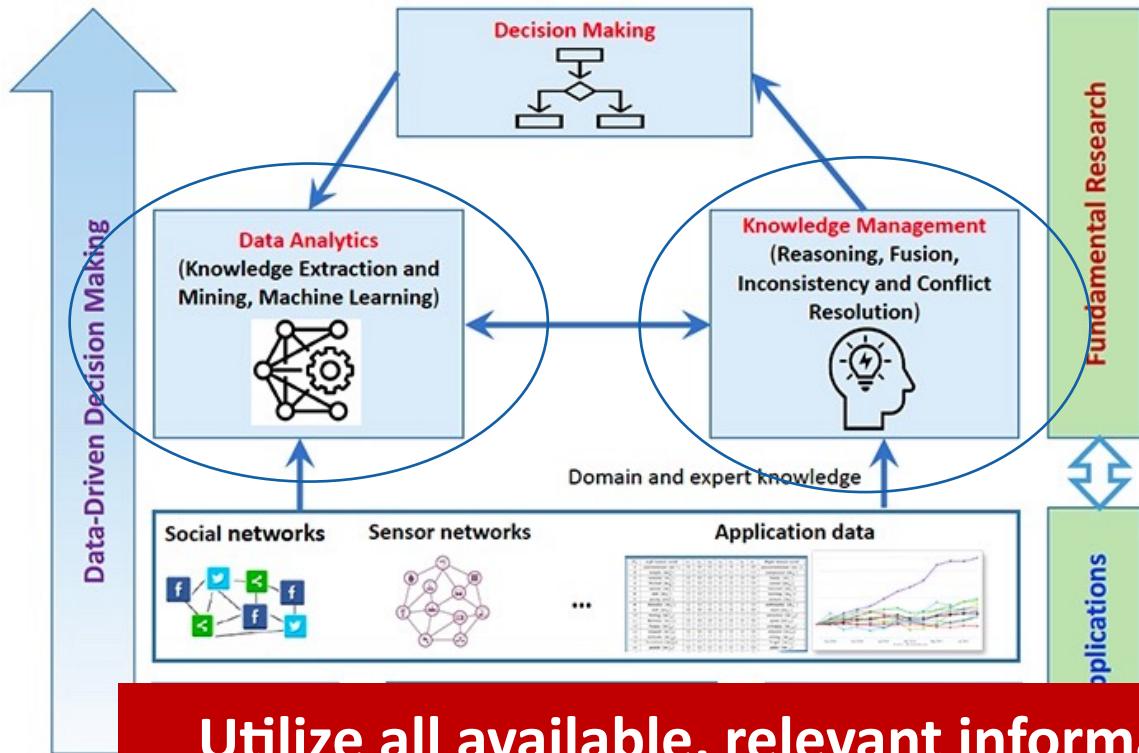
Background and Motivation



Background and Motivation



Background and Motivation



Utilize all available, relevant information obtained from multiple sources to autonomously and continually provide decision making information that is specific to application needs, while being able to adapt to uncertainties in a open-world environment

Multimodal data



Challenges

- Extracting **decision-making information from multimodal data** requires **finding common representation for source and context**, **fulfilling current needs from incomplete data**, and **predicting future needs**.

Solution:

- Exploiting **entity-centric higher-level semantic concepts**, RESTful **indexing**, **relational queries**, and **distributed stream-processing** for building **multimodal query engines**
- Exploiting **inherent semantic and stylistic properties** for **task-specific features**
- Exploiting **encapsulation properties** of view-based data integration to gain scalability

Challenges

- Knowledge management requires **data integration** and **inconsistency resolution**. Data integration approaches need to adapt to exact and approximate matches, novel data sources, and continual learning. This often faces the issue of **lack of annotation data**.

Solution:

- Find **novel sources of supervision** other than relevance labels
- Exploiting **representation learning** to capture the interconnected nature of **semantic features**

Challenges

- Knowledge management requires **data integration** and **inconsistency resolution**. Data integration approaches need to adapt to exact and approximate matches, novel data sources, and continual learning. This often faces the issue of **lack of annotation data**.

Solution:

- Find **novel sources of supervision** other than relevance labels
- Exploiting **representation learning** to capture the interconnected nature of **semantic features**
- Multimodal information retrieval, collaborative learning, or feature extraction models with multimodal data assumes **closed-world fixed models** and needs **re-learning**.

Solution:

- **Characterize, detect, and tackle novelties** in dynamic data-driven applications, **datasets**, and in agent-based systems, such as in planning domains

Contributions

- **Scalable, real-time, feature-centric Situational Knowledge extraction and dissemination framework** with a multi-modal relational knowledge base that can fulfill current and future data needs from incomplete and disaggregate data sources
 - A **novel text attribute detection model** using language representation models and syntactic properties
 - A **Video Querying System** with human-in-the-loop, **Find-Them**, and two **novel MMIR datasets**
 - A **View-based Data Integration** using relational structure of RDBMS and SQL Queries

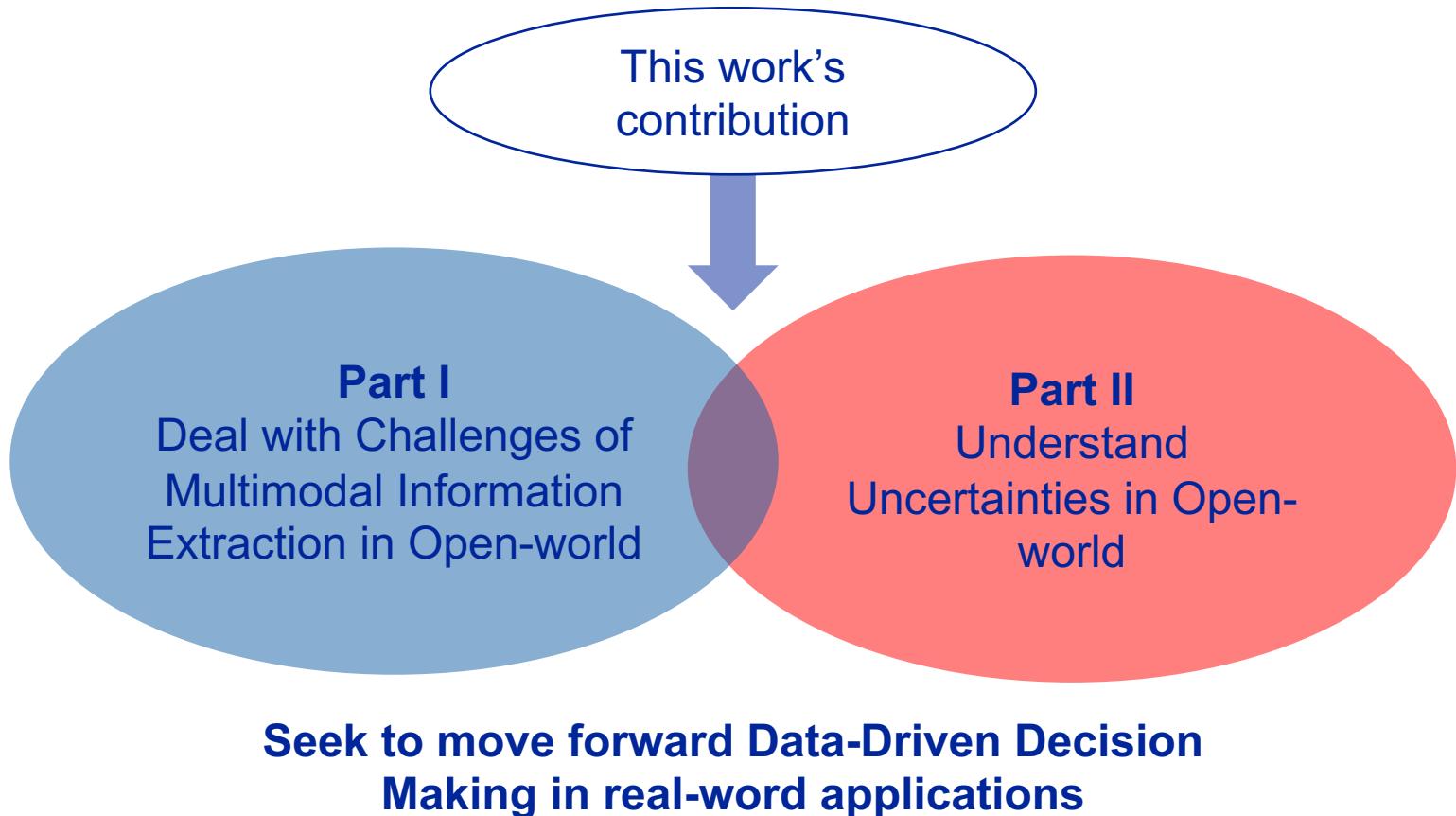
Contributions

- **Scalable, real-time, feature-centric Situational Knowledge extraction and dissemination framework** with a multi-modal relational knowledge base that can fulfill current and future data needs from incomplete and disaggregate data sources
 - A **novel text attribute detection model** using language representation models and syntactic properties
 - A **Video Querying System** with human-in-the-loop, **Find-Them**, and two **novel MMIR datasets**
 - A **View-based Data Integration** using relational structure of RDBMS and SQL Queries
- **Flexible, scalable, representation-invariant learning models** for data integration, relevance-ranking, and similarity matching
 - **Coordinated representation learning** with graph matching and a **novel distance metric** for data objects
 - **Relevance learning using weak supervision**

Contributions

- **Scalable, real-time, feature-centric Situational Knowledge extraction and dissemination framework** with a multi-modal relational knowledge base that can fulfill current and future data needs from incomplete and disaggregate data sources
 - A **novel text attribute detection model** using language representation models and syntactic properties
 - A **Video Querying System** with human-in-the-loop, **Find-Them**, and two **novel MMIR datasets**
 - A **View-based Data Integration** using relational structure of RDBMS and SQL Queries
- **Flexible, scalable, representation-invariant learning models** for data integration, relevance-ranking, and similarity matching
 - **Coordinated representation learning** with graph matching and a **novel distance metric** for data objects
 - **Relevance learning using weak supervision**
- **Novelty Characterization, Detection, and Adaption in Multimodal Information Retrieval**, distributed AI systems, and in planning domains
 - A **novel intrinsic complexity metric** for distributed AI applications in perception domains

Contributions



Contribution #1: Situational Knowledge Extraction on Demand

Use Case 1: Video Querying With A Human-in-the-Loop (SurvQ)

Use Case 2: Applying ML and Data Fusion to the *Missing Person*
Problem (Find-Them)

Problem Statement and Motivation

- Multimodal Information Retrieval (MMIR)
 - Fails to process streaming data as it ingests
 - Limited to modality-specific transformers due to low-level feature extraction
 - Do not consider context information in relevance matching
 - User preference or mission need is not considered, and is not dynamic
- Difference between MMIR benchmark datasets and real-world datasets
 - Noisy, lacks *relevance* labels, heterogenous

Problem Statement and Motivation

- Multimodal Information Retrieval (MMIR)
 - Fails to process streaming data as it ingests
 - Limited to modality-specific transformers due to low-level feature extraction
 - Do not consider context information in relevance matching
 - User preference or mission need is not considered, and is not dynamic
- Difference between MMIR benchmark datasets and real-world datasets
 - Noisy, lacks *relevance* labels, heterogenous
- Use cases: Video Feed Querying and Missing Person Search
 - Involve manually-performed investigations of large amount of multimodal data, especially video data, for a specific mission need

Problem Statement and Motivation

- Multimodal Information Retrieval (MMIR)
 - Fails to process streaming data as it ingests
 - Limited to modality-specific transformers due to low-level feature extraction
 - Do not consider context information in relevance matching
 - User preference or mission need is not considered, and is not dynamic

• Differences between current systems and what is needed

- Noisy, incomplete, and streaming data

• Use cases

- Involve natural language processing, especially for multimodal data, such as images and text.

Determine **relevant** information from *heterogeneous* and **multimodal** data, both at-rest and streaming, either **complete** or **incomplete**, with **scalability** to large amounts of data and based on mission needs and **user provided context**.

Related Works

- Specific domains i.e., sentiment analysis, disaster, healthcare, robotics perform multi-modal data fusion for situational knowledge [1-4]
 - Often structured or just textual data
 - Does not allow user to mention information need
 - Not generalizable across applications
- Visual QA, Missing person search [5-6, 19-20]
 - Limited to visual modality, and only textual augmentation
 - Query cannot be made with image or video data example
 - Does not search from streaming data
- Query-driven approaches for knowledge bases [10-14]
 - Text data driven knowledge base built by queries [10]
 - Inconsistency and missing info [11]
 - Probabilistic methods need training data, fuzzy matching, text and structured data
- Triggers used to deliver matching data [8-9], but rules are fixed

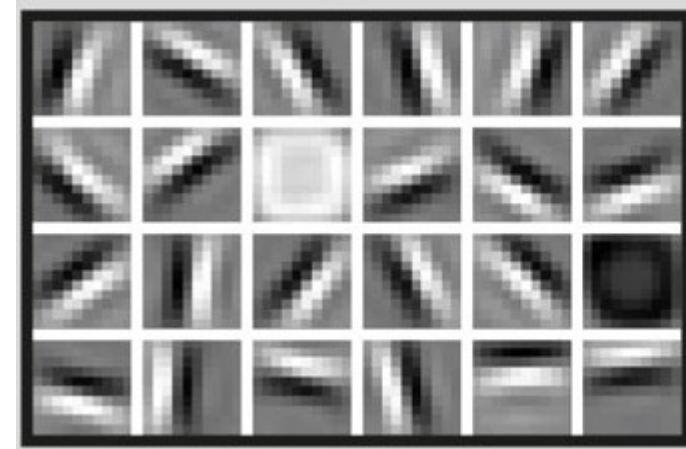
1. S. Poria, et al, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputer*, vol. 174, no. PA, pp. 50–59, Jan. 2016
2. Foresti, G.L et al: Situational awareness in smart environments: socio-mobile and sensor data fusion for emergency response to disasters. *J. Ambient Intelligence and Humanized Computing* 6(2), 239–257 (2015).
3. Meditskos, G. et al: Description logics and rules for multimodal situational awareness in healthcare. In: *MMM (1). Lecture Notes in Computer Science*, vol. 10132, pp. 714–725. Springer (2017)
4. Adjali, O et al: Multimodal fusion, fission and virtual reality simulation for an ambient robotic intelligence. In:*ANT/SEIT. Procedia Computer Science*, vol. 52, pp. 218–225. Elsevier (2015)
5. Y. Zhu et al, "Knowledge acquisition for visual question answering via iterative querying," in *CVPR, IEEE Computer Society*, 2017, pp. 6146–6155.
6. Q. Wu et al, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*
7. Kang, D., Bailis, P., Zaharia, M.: *Blazeit: Fast exploratory video queries using neural networks*. CoRRabs/1805.01046(2018).
8. M. L. Itria, A. Daidone, and A. Ceccarelli, "A complex event processing approach for crisis-management systems," *CoRR*, vol. abs/1404.7551, 2014.
9. G. Cugola and A. Margara, "Processing flows of information: From data stream to complex event processing," *ACM Comput. Surv.*, vol. 44, no. 3, 15:1–15:62, 2012.
10. D. B. Nguyen et al, "Query-driven on-the-fly knowledge base construction," *Proc. VLDB Endow.*, vol. 11, no. 1, pp. 66– 79, Sep. 2017
11. M. Bienvenu et al, "Query-driven repairing of inconsistent dl-lite knowledge bases," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*
12. X. Dong et al., "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *KDD, ACM, 2014*, pp. 601–610.
13. Y. Chen and D. Z. Wang, "Knowledge expansion over probabilistic knowledge bases," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, ser. SIGMOD '14, Snowbird, Utah, USA: ACM, 2014*
14. M. E. Rodriguez, S. Goldberg, and D. Z. Wang, "Sigmakb: Multiple probabilistic knowledge base fusion," *PVLDB*, vol. 9, no. 13, pp. 1577–1580, 2016.
15. D. Kang, J. Emmons, F. Abuzaid, P. Bailis, and M. Zaharia, "Noscope: Optimizing deep cnn-based queries over video streams at scale," *PVLDB*, vol. 10, no. 11, pp. 1586– 1597, 2017.
16. M. R. Anderson, M. J. Cafarella, G. Ros, and T. F. Wenisch, "Physical representation- based predicate optimization for a visual analytics database," *CoRR*, vol. abs/1806.04226, 2018.
17. K. Hsieh et al., "Focus: Querying large video datasets with low latency and low cost," in *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, I. Xarchakos and N. Koudas, "SVQ: streaming video queries," in *SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*,*
18. G. Pearson, M. Gill, S. Antani, L. Neve, and G. Thoma, "People locator: A system for family reunification," *IT Professional*, vol. 14, no. 03, pp. 13–21, May 2012
19. R. S. Ferreira, C. G. de Oliveira, and A. A. Lima, "Myosotis: An information system applied to missing people problem," in *Proceedings of the XIV Brazilian Symposium on Information Systems, 2018*

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation



Layer 2: The computer learns to identify edges and simple shapes.



Layer 4: The computer learns which shapes and objects can be used to define a human face.

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation

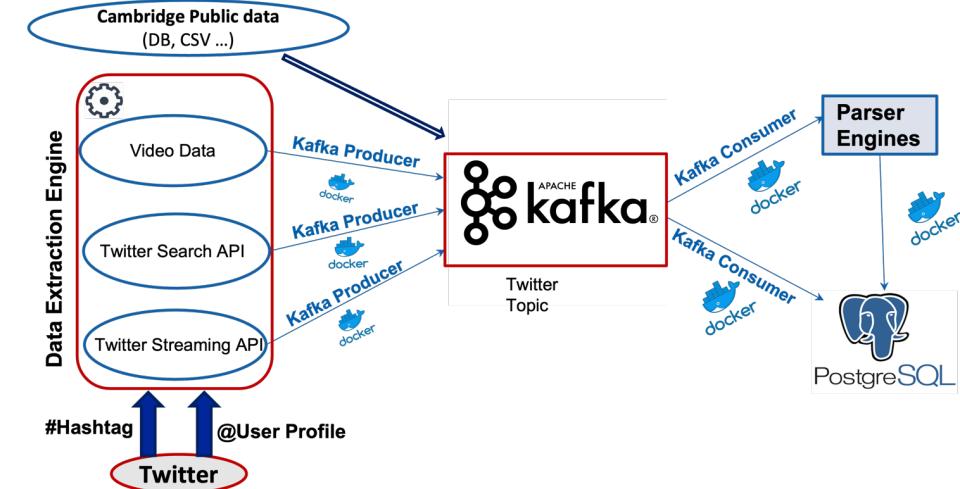
- **High-level Semantic Feature for Data Representation**
 - Interpretability
 - Explainable by interrogation
 - Explainable by design
 - Scalability
 - Compatibility with SOTA models
- **Context Representation**
 - User uploads example or semantic features
 - Query-by-example or Query-by-feature

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data

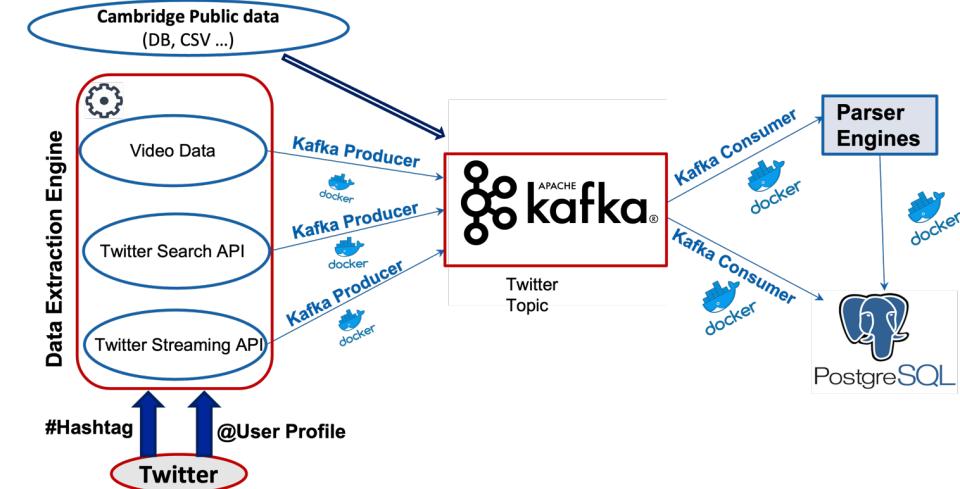
Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data



Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data



**Replicable, Fault tolerant,
Scalable and Continuous**

Solutions

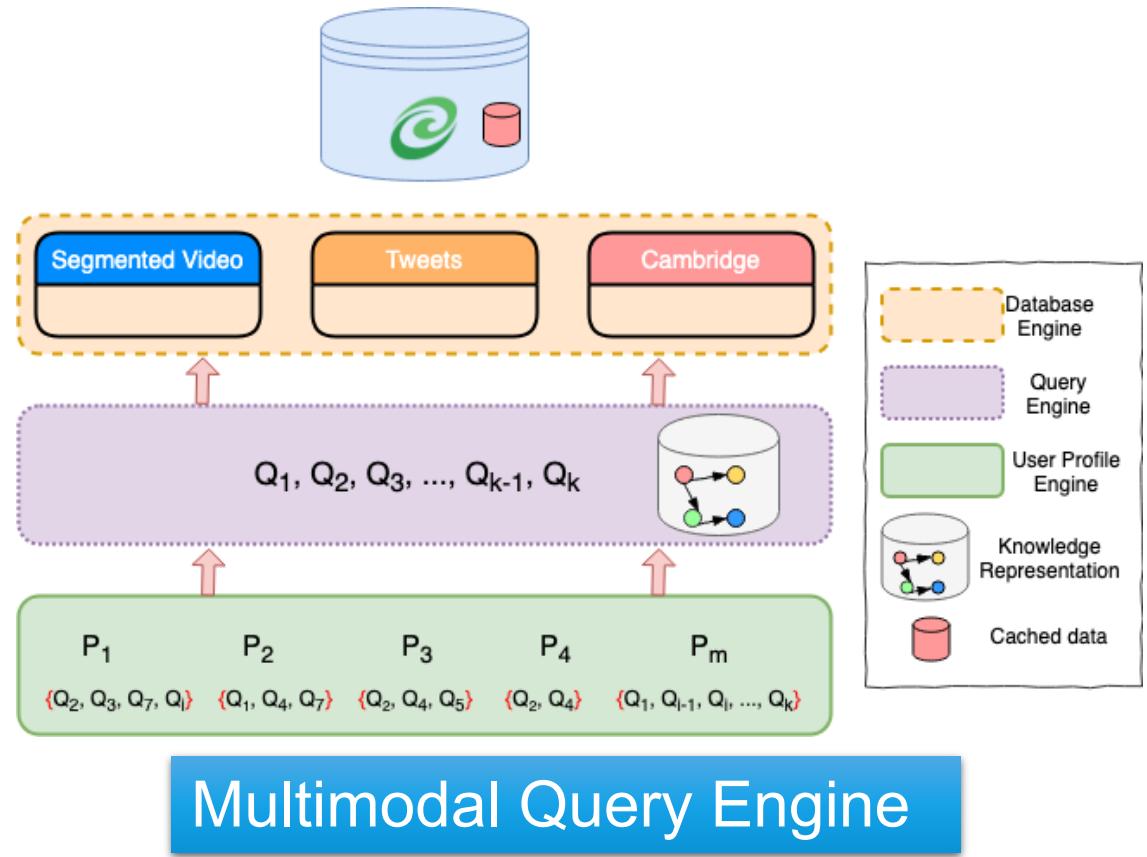
- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data
- Scalable **Data integration** and Data Storage process to petabytes of data
- Creating **adaptable knowledge base** using historical queries and user preference

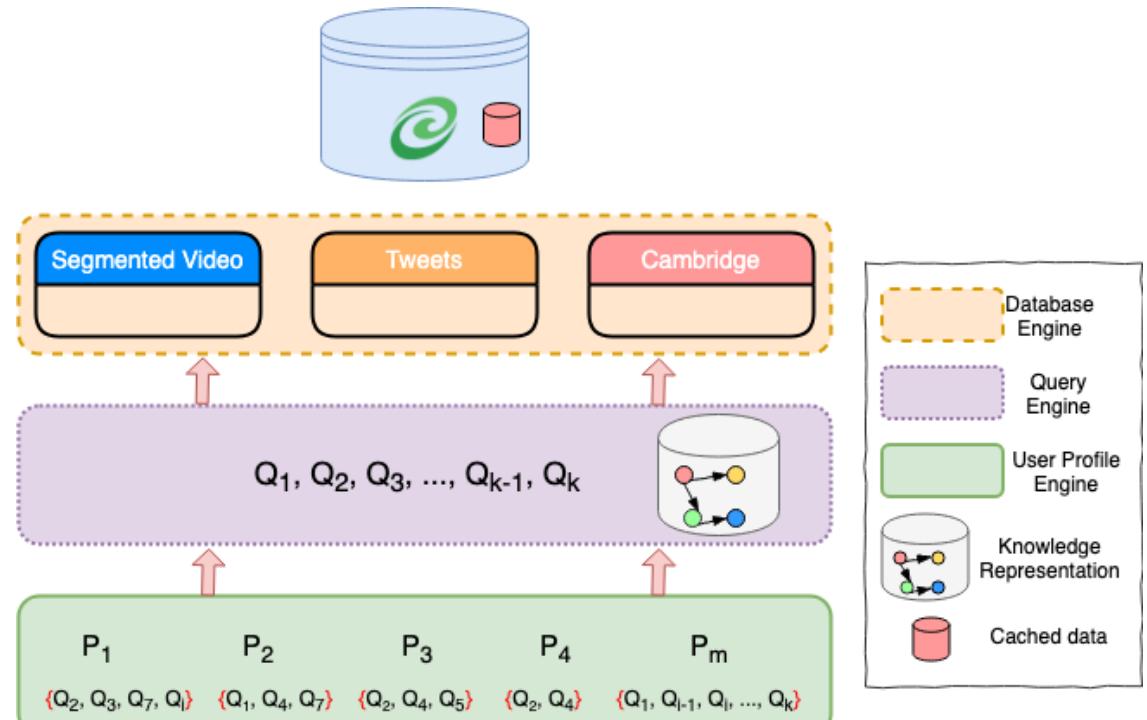
Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data
- Scalable **Data integration** and Data Storage process to petabytes of data
- Creating **adaptable knowledge base** using historical queries and user preference



Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data
- Scalable **Data integration** and Data Storage process to petabytes of data
- Creating **adaptable knowledge base** using historical queries and user preference

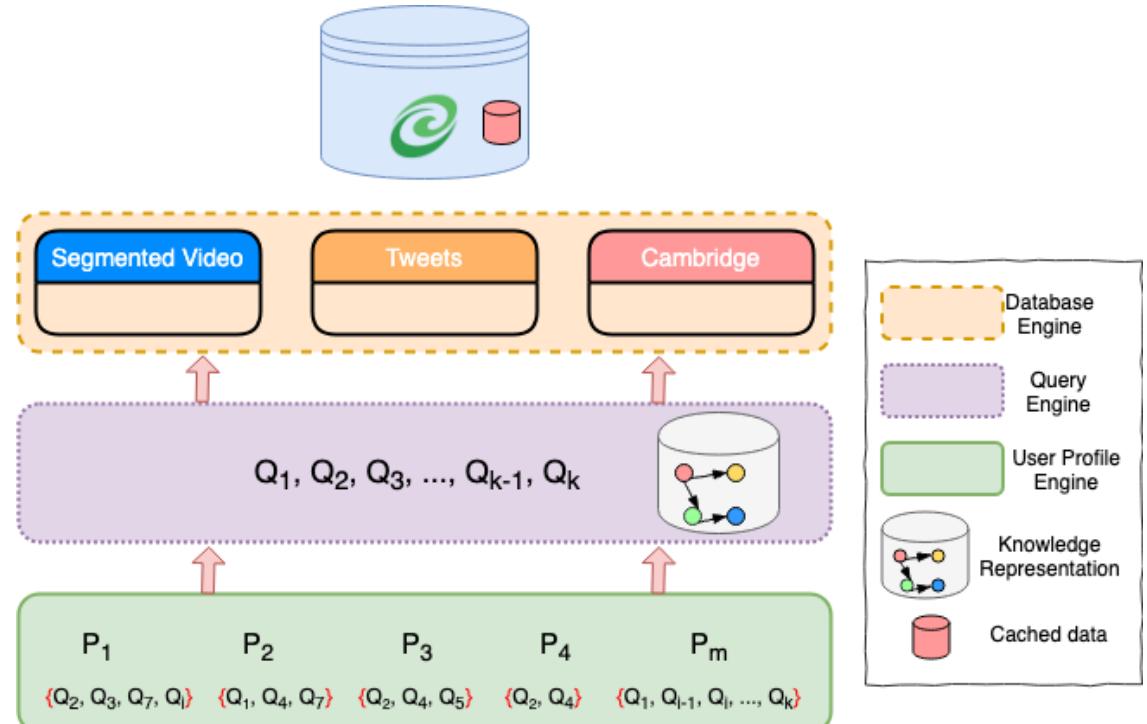


Multimodal Query Engine

- knowledge \equiv relational data and SQL queries on the data
- persist for lifetime of knowledge base and grow with additional user interests

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data
- Scalable **Data integration** and Data Storage process to petabytes of data
- Creating **adaptable knowledge base** using historical queries and user preference
- Methodology to fulfill current needs with existing incomplete modalities and deliver data as it arrives

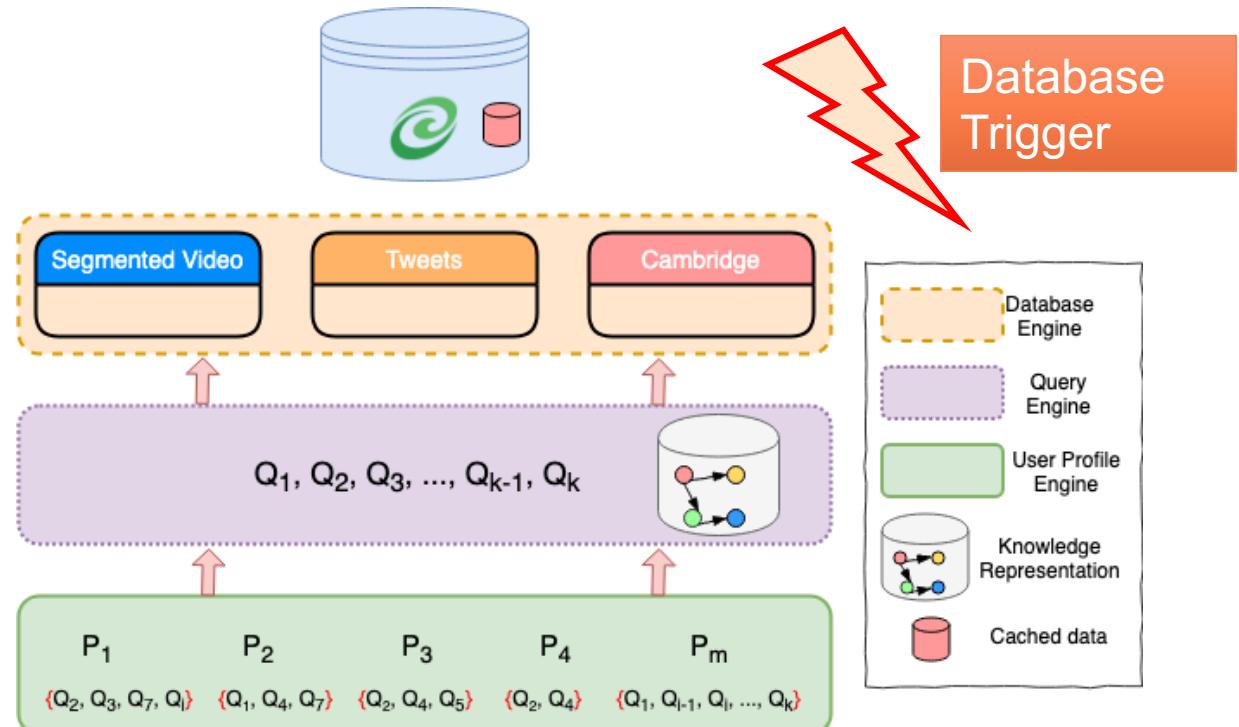


Multimodal Query Engine

- knowledge \equiv relational data and SQL queries on the data
- persist for lifetime of knowledge base and grow with additional user interests

Solutions

- Generalizable representation for content and context
 - Using low level features vs high level semantic features
 - Connect user information need with the representation
- Novel method digesting both streaming and at-rest data and creating a stable ingest process for data
- Scalable **Data integration** and Data Storage process to petabytes of data
- Creating **adaptable knowledge base** using historical queries and user preference
- Methodology to fulfill current needs with existing incomplete modalities and deliver data as it arrives



Multimodal Query Engine

- knowledge \equiv relational data and SQL queries on the data
- persist for lifetime of knowledge base and grow with additional user interests

Knowledge Modeling for Multiple Modalities

Data Integration

- Multimodal Query Engine
 - Feature Extraction
 - Data Integration and Relevance Learning

Knowledge Modeling for Multiple Modalities

Data Integration

- Multimodal Query Engine
 - Feature Extraction
 - Data Integration and Relevance Learning
- Challenges
 - Different Schema / Feature-name in different modalities

Knowledge Modeling for Multiple Modalities

Data Integration

- Multimodal Query Engine
 - Feature Extraction
 - Data Integration and Relevance Learning
- Challenges
 - Different Schema / Feature-name in different modalities

```
Enum cloth_types {  
    shirt  
    pant  
    jeans  
    hat  
    baseball_cap  
    t_shirt  
    jacket  
    hoodie  
}
```

Text Feature Schema

```
Enum body_parts {  
    upper_body  
    lower_body  
    head  
}
```

BodyDetails	
id	int
description	varchar
part	body_parts

Video Feature Schema

Knowledge Modeling for Multiple Modalities

Data Integration

- Multimodal Query Engine
 - Feature Extraction
 - Data Integration and Relevance Learning
- Challenges
 - Different Schema / Feature-name in different modalities
- Solution
 - View-based Data Integration
 - Schema Mapping
 - SQL-JOIN

```
Enum cloth_types {  
    shirt  
    pant  
    jeans  
    hat  
    baseball_cap  
    t_shirt  
    jacket  
    hoodie  
}
```

Text Feature Schema

```
Enum body_parts {  
    upper_body  
    lower_body  
    head  
}  
  


| BodyDetails |            |
|-------------|------------|
| id          | int        |
| description | varchar    |
| part        | body_parts |

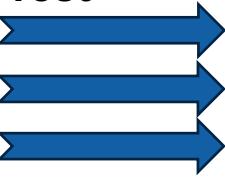


Video Feature Schema

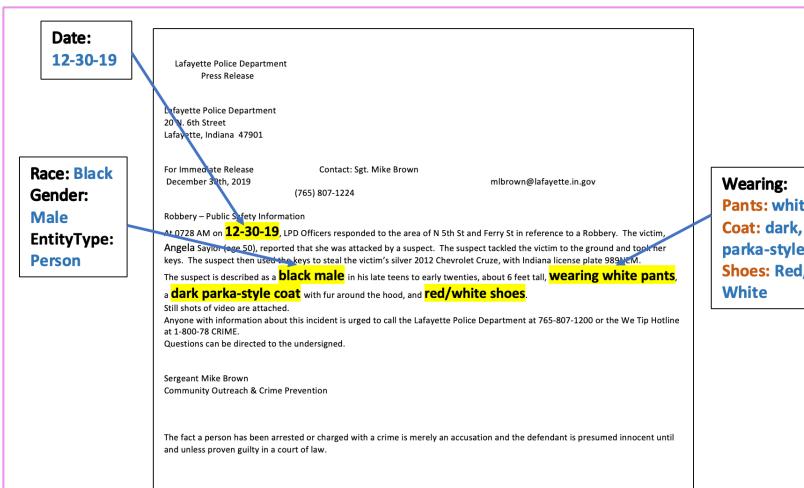

```

Data Fusion and Relevance Matching with EARS

Streaming
Input or
Data-at-
rest

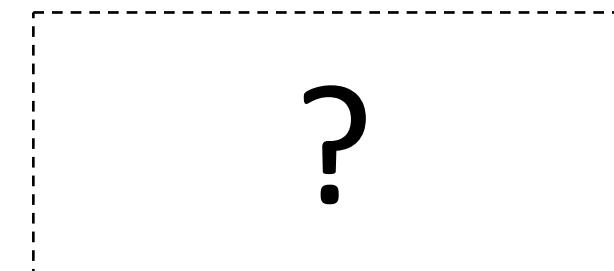


Video
Schema



Text
Schema

Features are extracted
from raw input

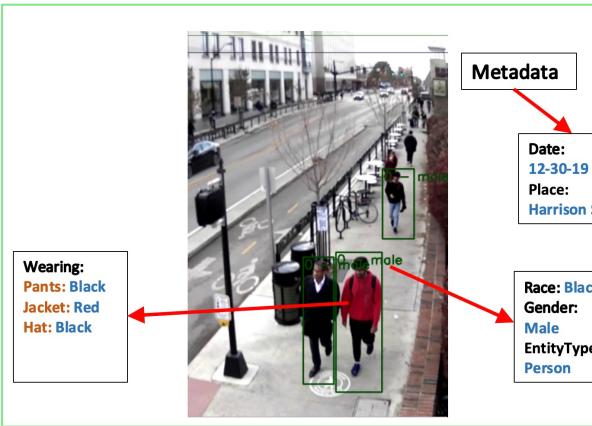
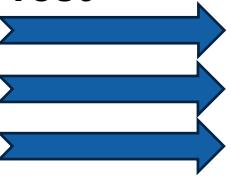


Content Relevance
Discovery

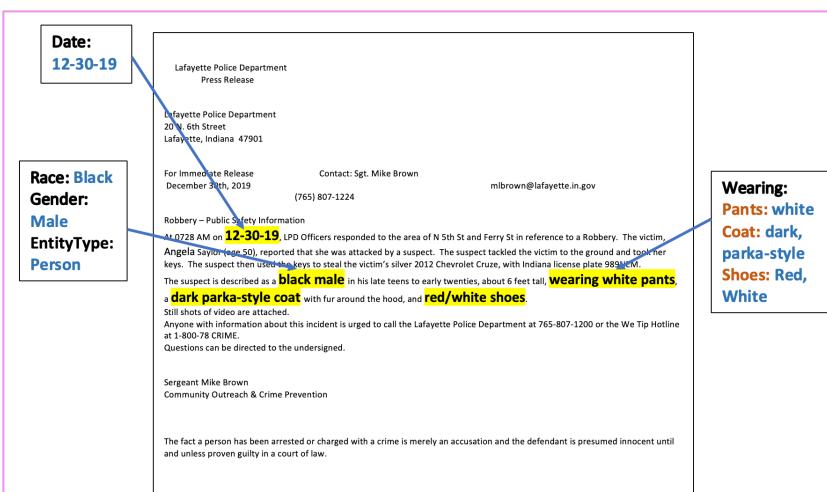


Data Fusion and Relevance Matching with EARS

Streaming
Input or
Data-at-
rest

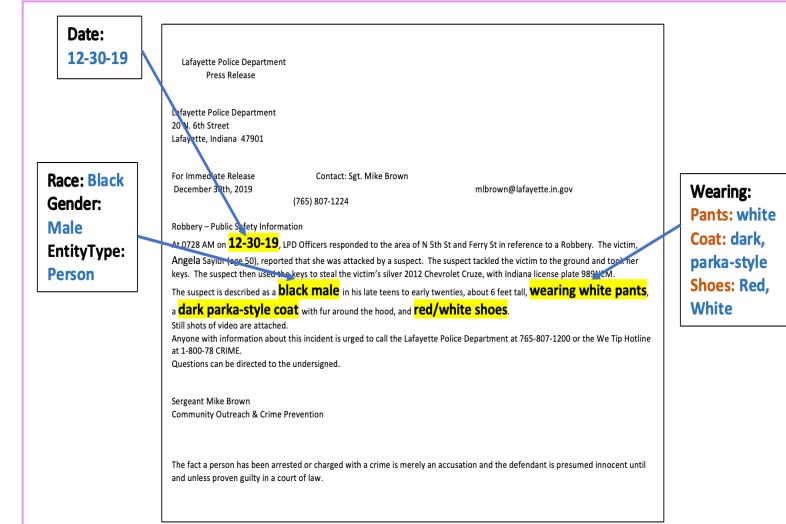


Video
Schema



Text
Schema

Features are extracted
from raw input



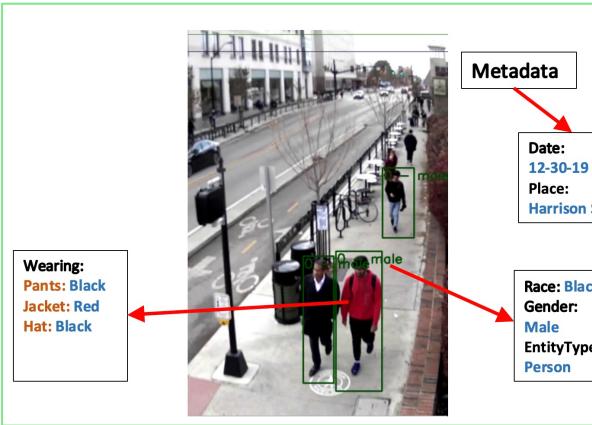
User
Input



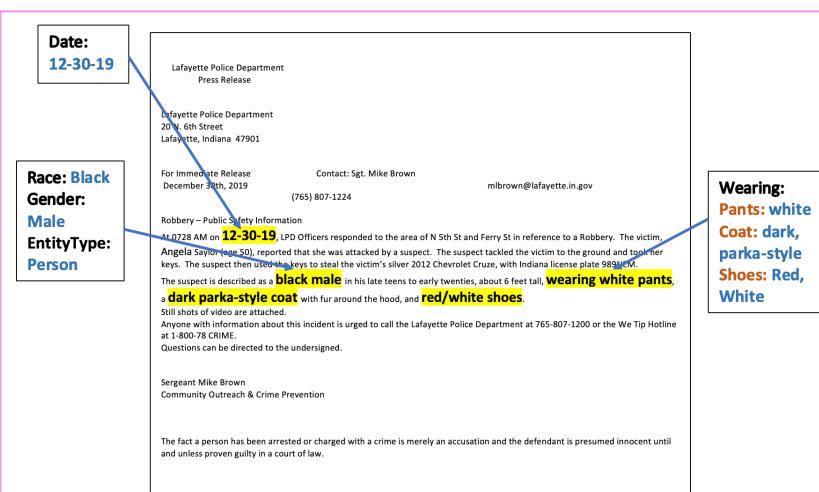
Content Relevance
Discovery

Data Fusion and Relevance Matching with EARS

Streaming
Input or
Data-at-
rest

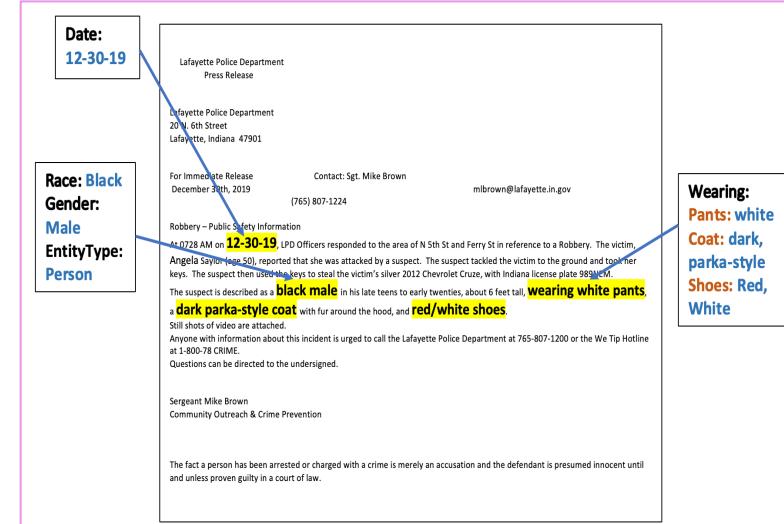


Video Schema



Text Schema

Features are extracted
from raw input



User
Input



Content Relevance
Discovery

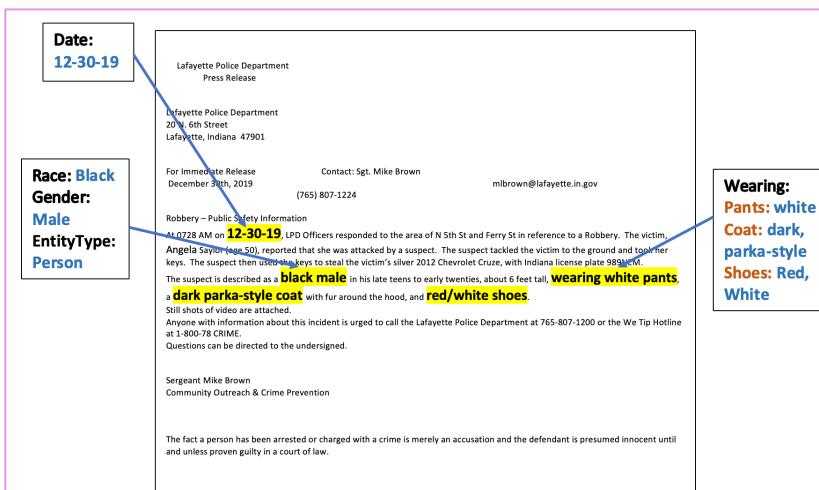
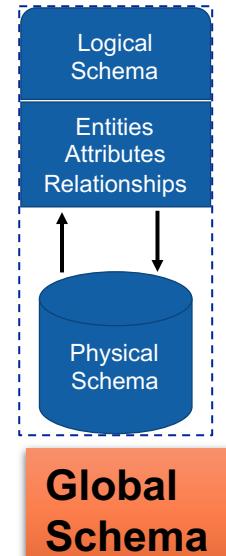
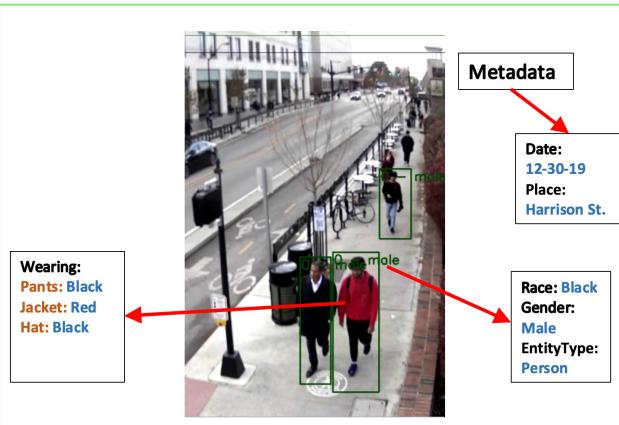
F_2, F_6, F_i

Extracted
Features

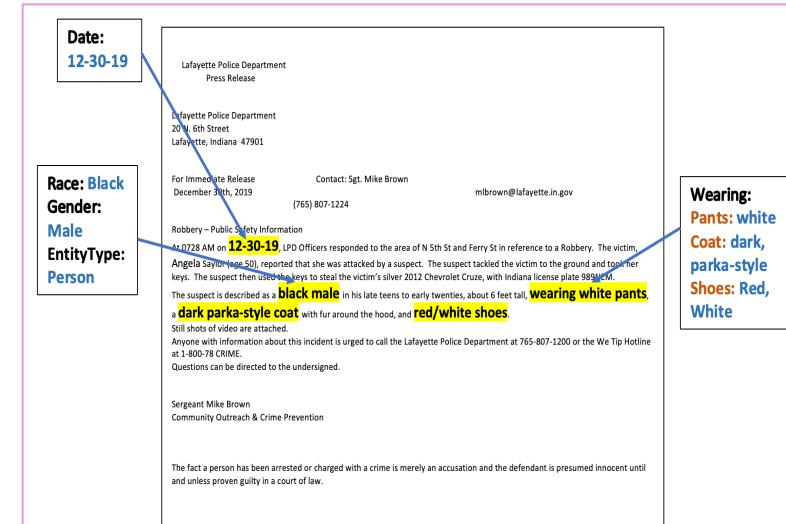


Data Fusion and Relevance Matching with EARS

Streaming
Input or
Data-at-
rest



Features are extracted from raw input



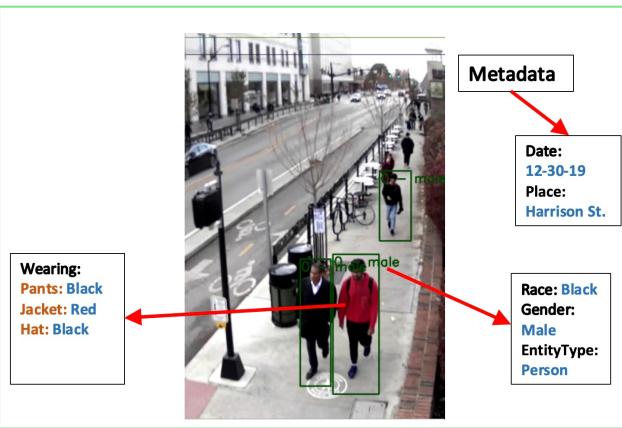
Content Relevance
Discovery

F_2, F_6, F_i

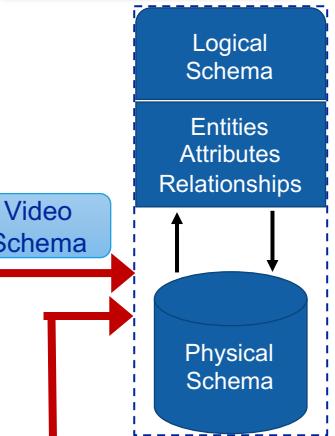
Extracted
Features

Data Fusion and Relevance Matching with EARS

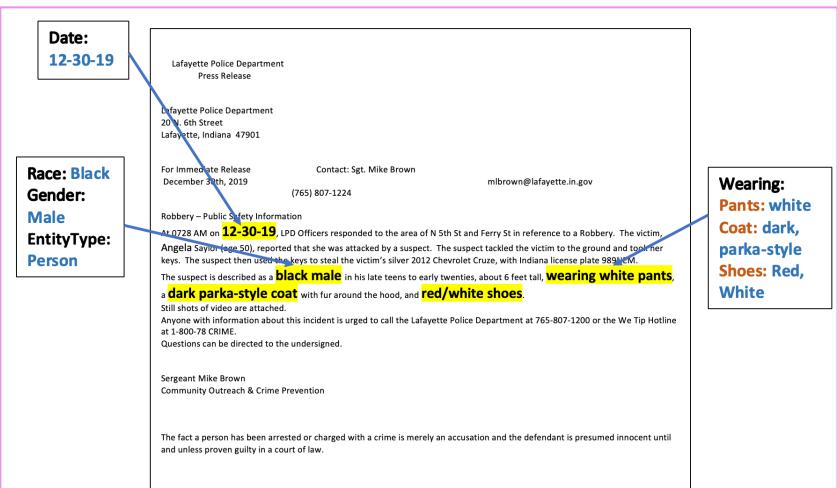
Streaming
Input or
Data-at-
rest



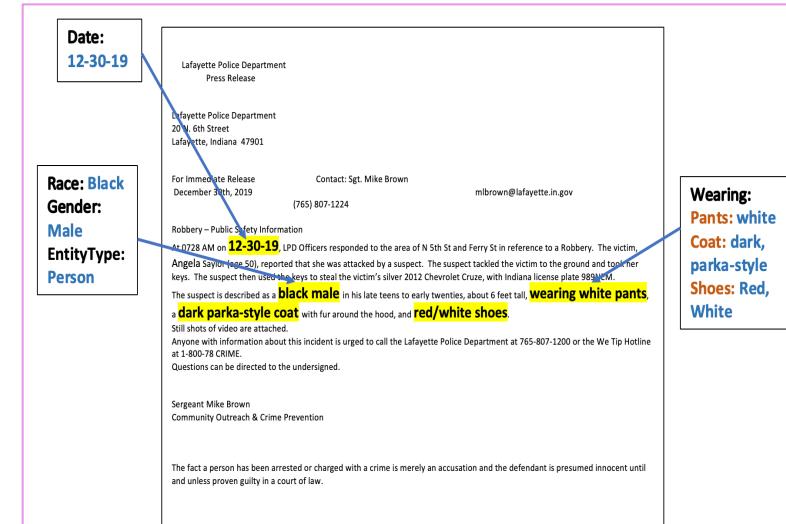
Schema Mapping



Global Schema



Features are extracted from raw input



Content Relevance Discovery

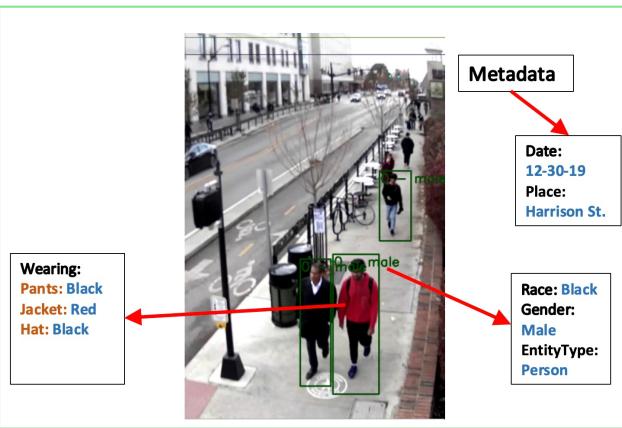
F_2, F_6, F_i

Extracted Features

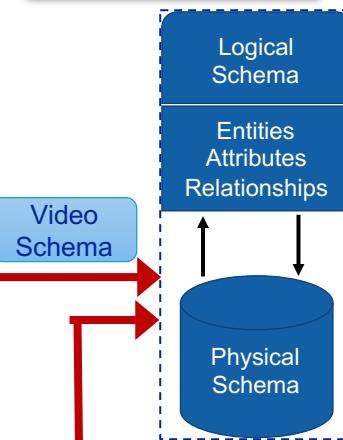
User Input

Data Fusion and Relevance Matching with EARS

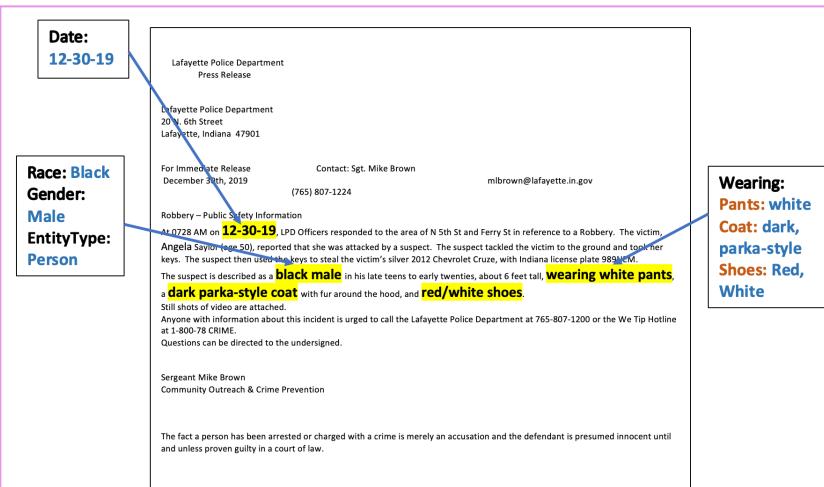
Streaming
Input or
Data-at-
rest



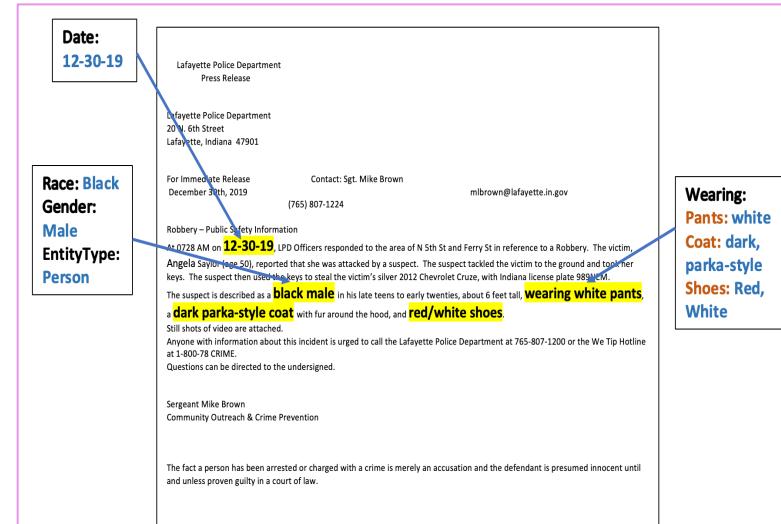
Schema Mapping



Global Schema



Features are extracted
from raw input



Content Relevance Discovery

F_2, F_6, F_i

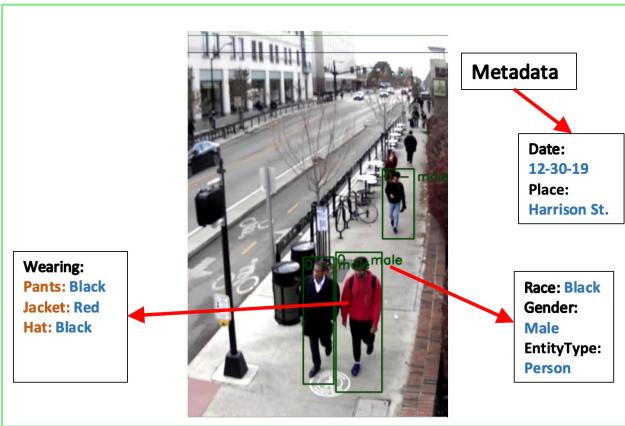
Extracted
Features



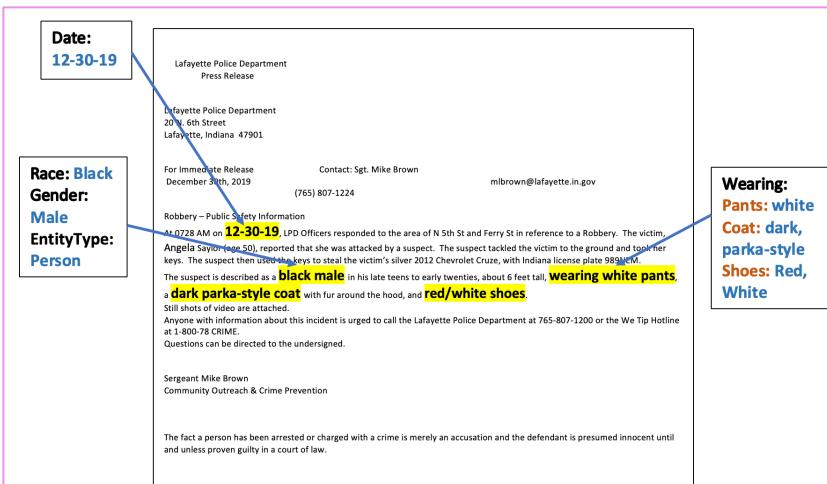
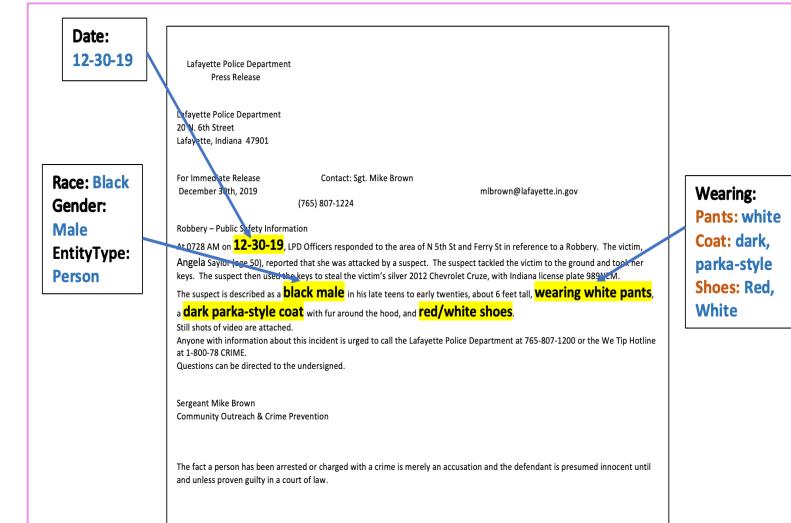
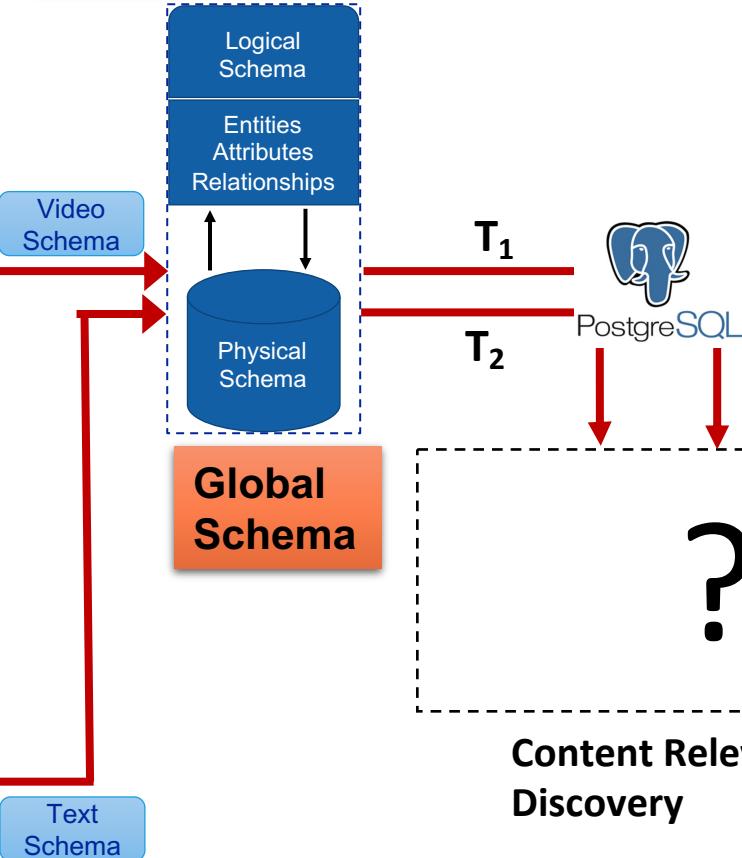
Triggering
Events

Data Fusion and Relevance Matching with EARS

Streaming
Input or
Data-at-
rest

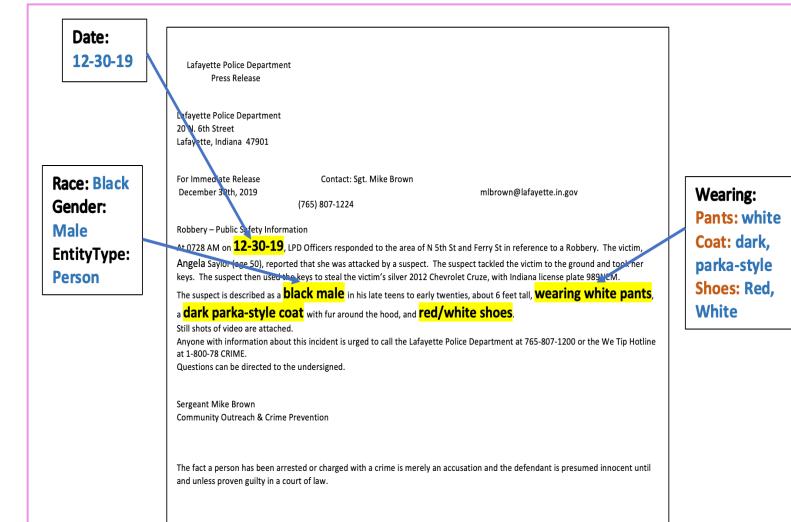
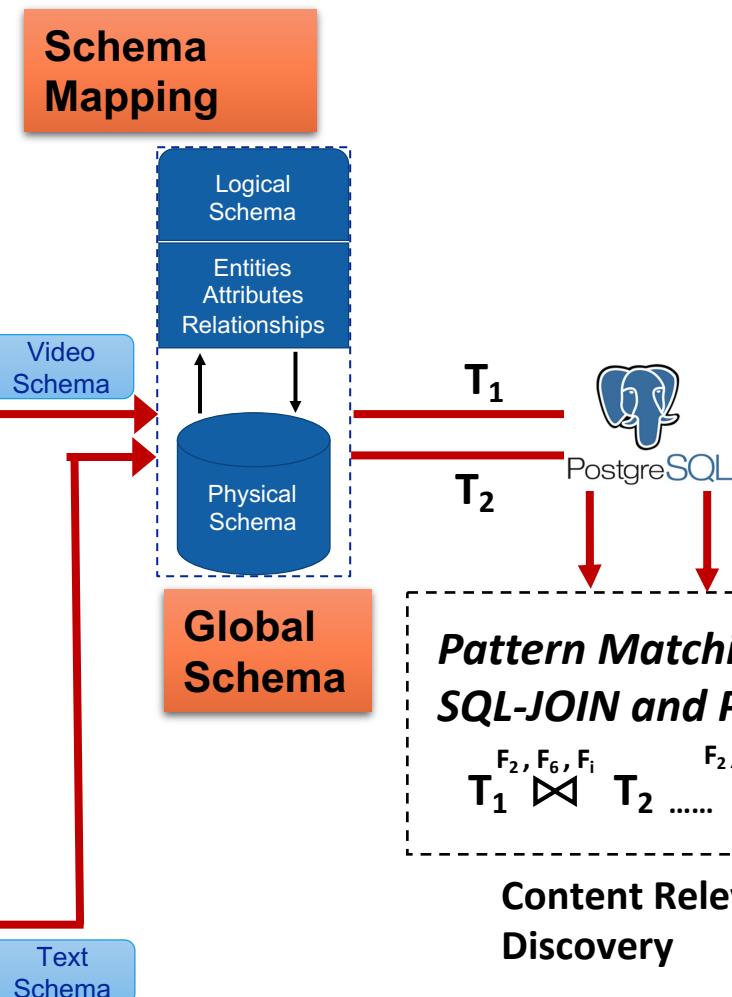
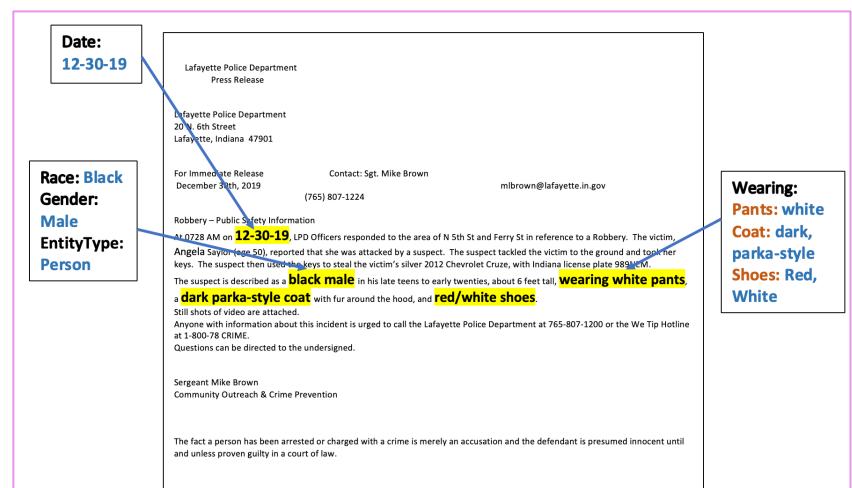
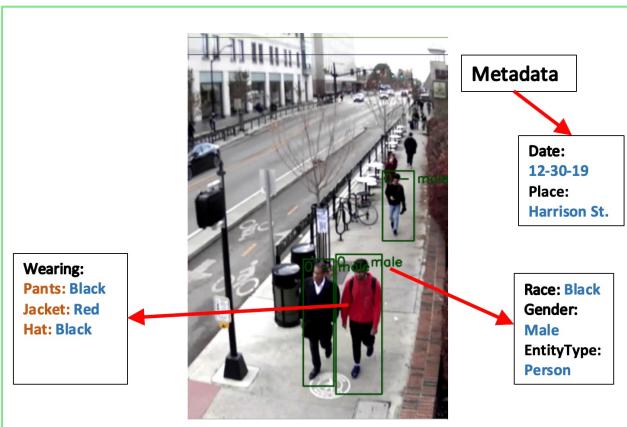


Schema Mapping



Data Fusion and Relevance Matching with EARS

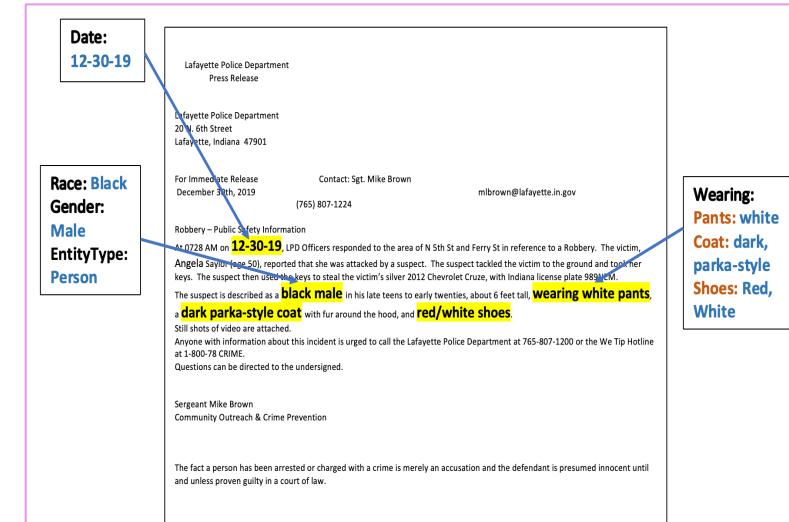
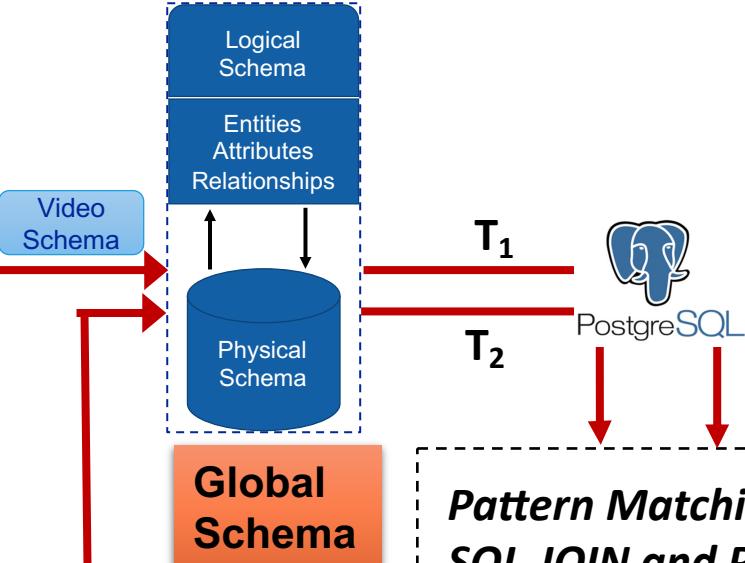
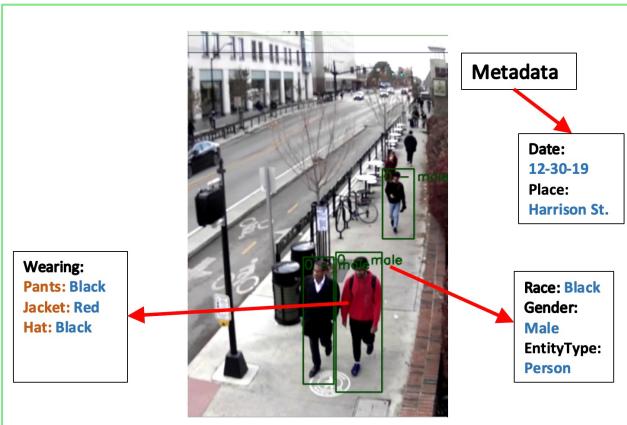
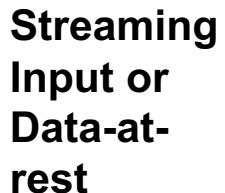
Streaming
Input or
Data-at-
rest



User
Input

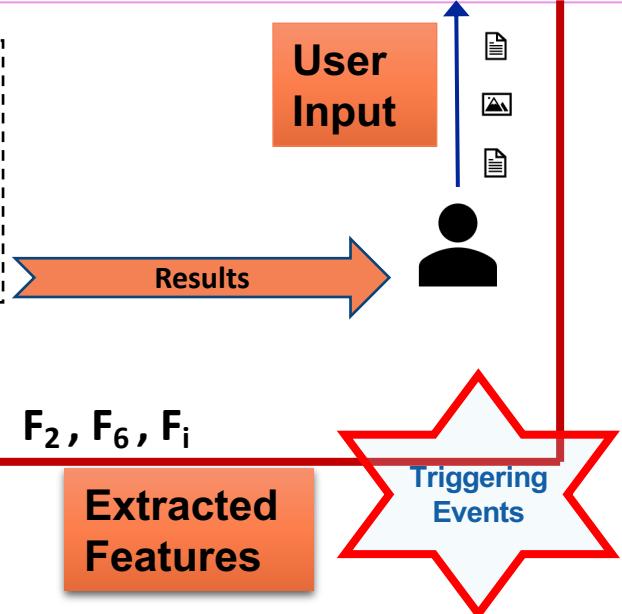


Data Fusion and Relevance Matching with EARS



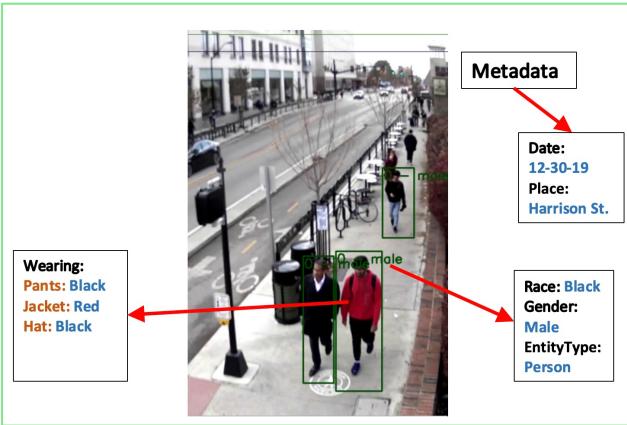
$$T_1 \otimes T_2 \dots \otimes T_k$$

Content Relevance Discovery

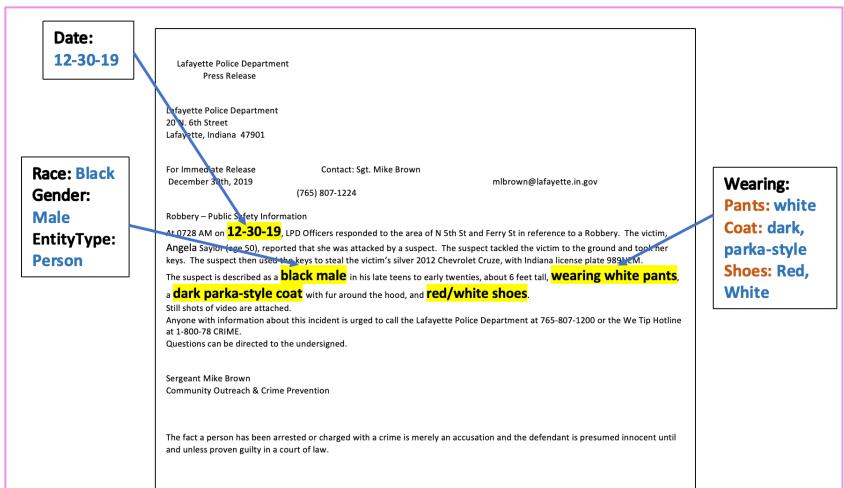
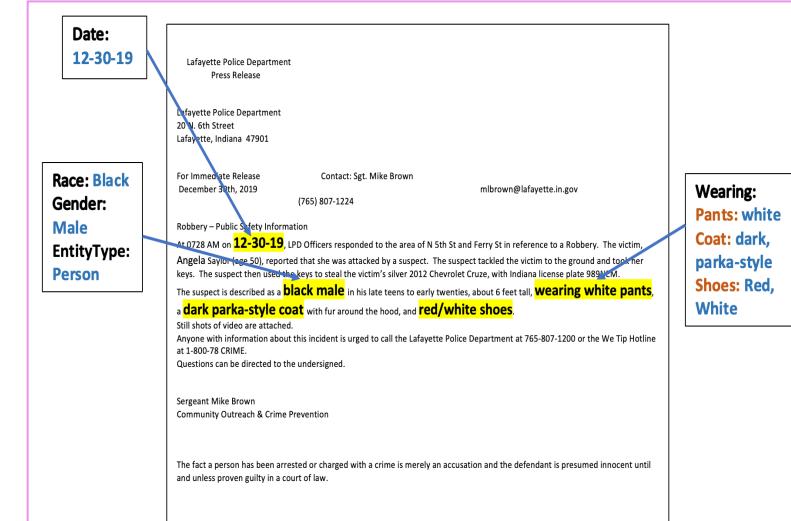
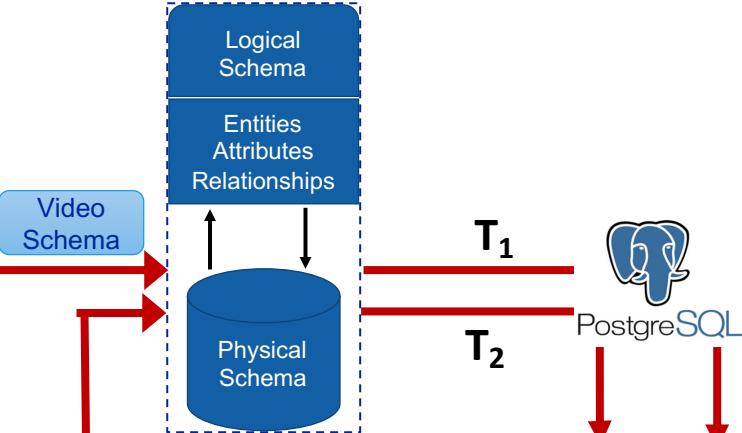


Data Fusion and Relevance Matching with EARS

Streaming Input or Data-at-rest



Schema Mapping



Features are extracted from raw input

Pattern Matching with SQL-JOIN and Predicates

$$T_1 \bowtie T_2 \dots \bowtie T_k$$

Content Relevance Discovery

F_2, F_6, F_i

Extracted Features

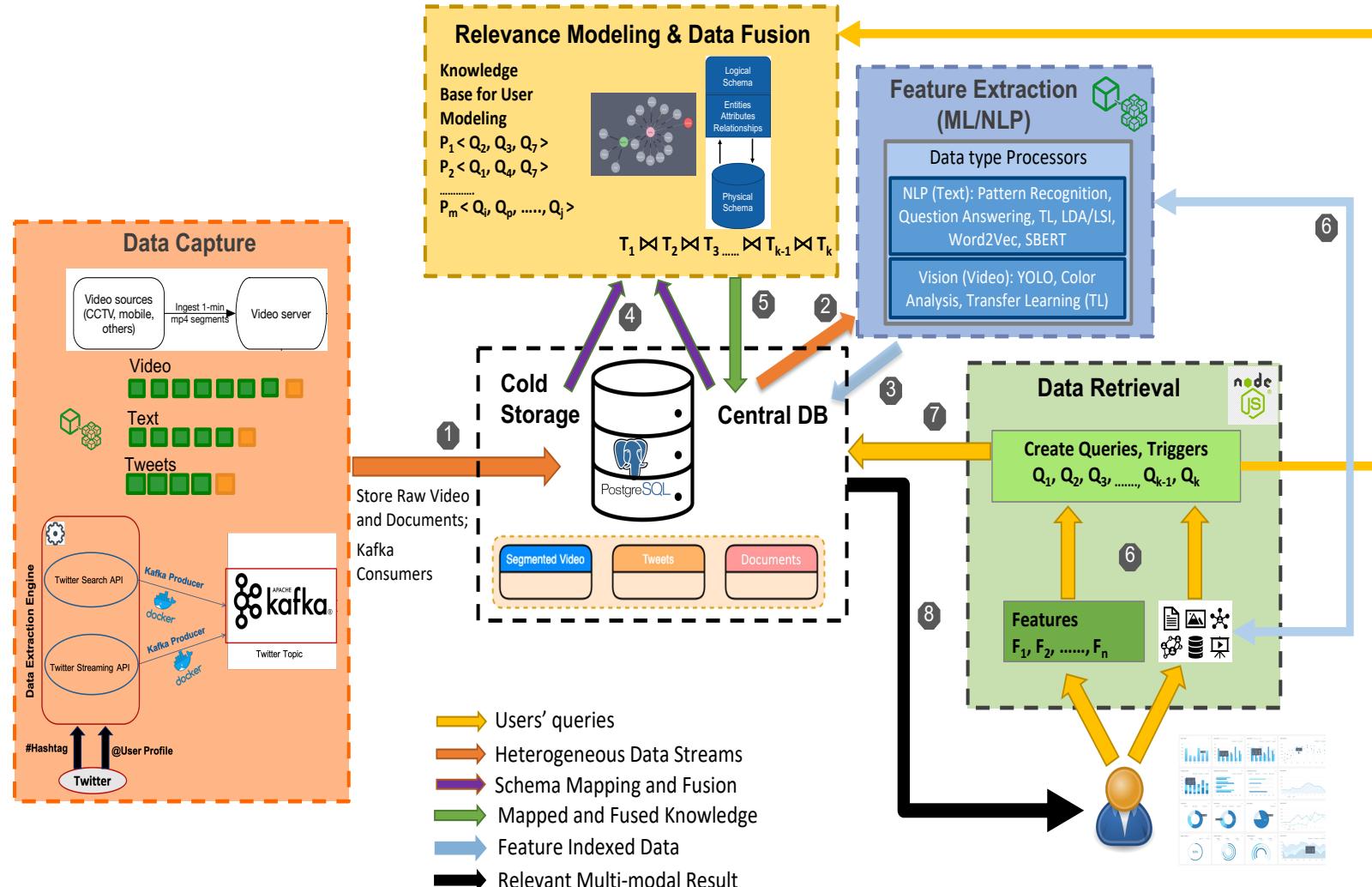
User Input

Results

Triggering Events

SKOD Architecture

- Main Components
 - Data Capture and Ingestion
 - Broker
 - RDBMS data store
 - Multimodal Query Engine
 - Feature Extraction
 - Knowledge Modeling
 - Relevance Learning
 - Data Fusion
 - Data Retrieval
 - Triggers
 - Cache, MRQ



Human Attribute Extraction from Text (HART)

User: TMGREENE	WEST LAFAYETTE POLICE DEPARTMENT	04/23/2020 14:07
Incident / Investigation - Case #: 2015-003151 : Off. Narr		

On August 21, 2015 1, Officer Jeffery Spicer responded to a report of an attempted strong armed robbery initially put out for the area of 201 S Salisbury (later determined to have occurred near the intersection of State Street and Pierce Street). Dispatch advised that three black males attempted to rob a female of her purse, however they failed and fled southbound in a grey Nissan Pathfinder with an Indiana registration of 421MBY. While en route to check south-bound on S River Road, Officer Brewer advised he ran that license plate earlier and it comes back to an older tan (with rust) Pathfinder. As I approached the area of US 231 and State Road 25 I did not observe any vehicles matching that description. As I waited for the light to turn green to go back to meet with the victim, I observed a gold/tan looking Nissan Pathfinder traveling north-bound on US 231 preparing to turn east-bound onto State Road 25. At this time I was able to get behind the vehicle and I noticed that the license plate provided to me by dispatch and Officer Brewer matched that of the one the victim gave to dispatch. I then continued to follow the vehicle east-bound on State Road 25 till we go to the intersection of Old US 231. The vehicle then changed lanes while stopped into the south-bound turn lane for Old US 231 without signaling. Once the vehicle turned south onto Old US 231 it pulled into the CVS parking lot, at which time I activated my emergency lights and stopped the vehicle in the CVS parking lot.

Once the vehicle was stopped I exited my patrol car and held the occupants at gun point until backup units arrived. Once other units arrived on scene we initiated a felony stop on the vehicle and first had the driver exit the vehicle. The driver was later identified as Marquise D LEIGH [REDACTED] and he was detained in handcuffs and placed into a patrol car (LEIGH was wearing a black shirt with long dreadlock style hair). Next we ordered the backseat passenger side occupant out of the vehicle, who was later identified as Kierre D MCCOY [REDACTED] (MCCOY was wearing a light colored white/light blue shirt). After MCCOY was detained in handcuffs the front seat passenger exited the vehicle and was later identified as Derek C SMITH [REDACTED] and he was detained in handcuffs as well (SMITH was wearing a red Adidas track jacket). After the three males were detained and the vehicle cleared of anyone else, Lt. Lord advised that PUPD was going to bring two victims and a witness to my location for an identification show up.

When the three subjects arrived to my location, I had officers bring Kierre MCCOY out of the vehicle and put him up against the wall of the CVS building so that the victims and witness could see him. The male victim, Tyler HO advised that he was not sure on the subject because he was on the ground getting assaulted. The male witness, James ROACH advised that he MCCOY looked familiar when I asked him if he did. I then asked if he was sure and he stated that he was about eighty percent sure. I then had the female, Maggie LENGACHER (who was the victim of the attempted robbery) step out of the police car to look at MCCOY. When LENGACHER stepped out and looked at MCCOY she stated, "ya" and that he was the third one to exit the vehicle when the fight broke out. She also stated she observed him in the front passenger seat and that he exited after the driver and backseat passenger did to fight her friends. LENGACHER also advised that all three of them attempted to take her purse during my video recorded interview with her. When I asked her how certain she was on MCCOY being one of the suspects she advised she was eighty-five percent sure and that she really remembers one wearing a white shirt, one wearing a black shirt, and the other wearing a red shirt.

Next I had LEIGH exit the patrol car and lined him up against the CVS wall, at which time I went over to ROACH and HO and they both advised me that they were one-hundred percent sure that LEIGH was one of the suspects that assaulted HO and attempted to rob LENGACHER. HO also advised that LEIGH was the person that instigated the entire incident. When I went and had LENGACHER look at LEIGH she advised me that she was positive on an identification. When I asked her what role he played she advised that LEIGH was the first person to get out of the car, and he tackled her friend, Eric GABBARD. She also advised that he was the first person to push her to the ground and tackle her. I then had SMITH stand up next to the CVS wall to show the two victims and one witness him. When I went and spoke with HO and ROACH, HO advised he was one-hundred percent sure that SMITH was the one that kicked him in the head. ROACH also stated that he was certain that SMITH was the one who kicked HO in the head. ROACH also advised me that he was certain he saw SMITH hit LENGACHER. I then had LENGACHER look at SMITH and her initial response was, "yes". I then asked her how certain she was and she advised, 95 % sure and that she was not 100 percent sure only because she thought he was wearing a red t-shirt and not a red track jacket. I asked her what role SMITH played in the altercation and she advised that he exited from one of the passenger seats when the driver did and that he was one of

Human Attribute Extraction from Text (HART)

a white male with medium build was seen in Vernon St., wearing white jeans and blue shirt

On August 21, 2015 1, Officer Jeffery Spicer responded to a report of an attempted strong armed robbery initially put out for the area of 201 S Salisbury (later determined to have occurred near the intersection of State Street and Pierce Street). Dispatch advised that three black males attempted to rob a female of her purse, however they failed and fled southbound in a grey Nissan Pathfinder with an Indiana registration of 421IMBY. While en route to check south-bound on S River Road, Officer Brewer advised he ran that license plate earlier and it comes back to an older tan (with rust) Pathfinder. As I approached the area of US 231 and State Road 25 I did not observe any vehicles matching that description. As I waited for the light to turn green to go back to meet with the victim, I observed a gold/tan looking Nissan Pathfinder traveling north-bound on US 231 preparing to turn east-bound onto State Road 25. At this time I was able to get behind the vehicle and I noticed that the license plate provided to me by dispatch and Officer Brewer matched that of the one the victim gave to dispatch. I then continued to follow the vehicle east-bound on State Road 25 till we go to the intersection of Old US 231. The vehicle then changed lanes while stopped into the south-bound turn lane for Old US 231 without signaling. Once the vehicle turned south onto Old US 231 it pulled into the CVS parking lot, at which time I activated my emergency lights and stopped the vehicle in the CVS parking lot.

Once the vehicle was stopped I exited my patrol car and held the occupants at gun point until backup units arrived. Once other units arrived on scene we initiated a felony stop on the vehicle and first had the driver exit the vehicle. The driver was later identified as Marquise D LEIGH [REDACTED] and he was detained in handcuffs and placed into a patrol car (LEIGH was wearing a black shirt with long dreadlock style hair). Next we ordered the backseat passenger side occupant out of the vehicle, who was later identified as Kierre D MCCOY [REDACTED] (MCCOY was wearing a light colored white/light blue shirt). After MCCOY was detained in handcuffs the front seat passenger exited the vehicle and was later identified as Derek C SMITH ([REDACTED]) and he was detained in handcuffs as well (SMITH was wearing a red Adidas track jacket). After the three males were detained and the vehicle cleared of anyone else, Lt. Lord advised that PUPD was going to bring two victims and a witness to my location for an identification show up.

When the three subjects arrived to my location, I had officers bring Kierre MCCOY out of the vehicle and put him up against the wall of the CVS building so that the victims and witness could see him. The male victim, Tyler HO advised that he was not sure on the subject because he was on the ground getting assaulted. The male witness, James ROACH advised that he MCCOY looked familiar when I asked him if he did. I then asked if he was sure and he stated that he was about eighty percent sure. I then had the female, Maggie LENGACHER (who was the victim of the attempted robbery) step out of the police car to look at MCCOY. When LENGACHER stepped out and looked at MCCOY she stated, "ya" and that he was the third one to exit the vehicle when the fight broke out. She also stated she observed him in the front passenger seat and that he exited after the driver and backseat passenger did to fight her friends. LENGACHER also advised that all three of them attempted to take her purse during my video recorded interview with her. When I asked her how certain she was on MCCOY being one of the suspects she advised she was eighty-five percent sure and that she really remembers one wearing a white shirt, one wearing a black shirt, and the other wearing a red shirt.

Next I had LEIGH exit the patrol car and lined him up against the CVS wall, at which time I went over to ROACH and HO and they both advised me that they were one-hundred percent sure that LEIGH was one of the suspects that assaulted HO and attempted to rob LENGACHER. HO also advised that LEIGH was the person that instigated the entire incident. When I went and had LENGACHER look at LEIGH she advised me that she was positive on an identification. When I asked her what role he played she advised that LEIGH was the first person to get out of the car, and he tackled her friend, Eric GABBARD. She also advised that he was the first person to push her to the ground and tackle her. I then had SMITH stand up next to the CVS wall to show the two victims and one witness him. When I went and spoke with HO and ROACH, HO advised he was one-hundred percent sure that SMITH was the one that kicked him in the head. ROACH also stated that he was certain that SMITH was the one who kicked HO in the head. ROACH also advised me that he was certain he saw SMITH hit LENGACHER. I then had LENGACHER look at SMITH and her initial response was, "yes". I then asked her how certain she was and she advised, 95 % sure and that she was not 100 percent sure only because she thought he was wearing a red t-shirt and not a red track jacket. I asked her what role SMITH played in the altercation and she advised that he exited from one of the passenger seats when the driver did and that he was one of

Human Attribute Extraction from Text (HART)

.....
**a white male with medium
build was seen in Vernon St.,
wearing white jeans and blue
shirt**
.....

1. gender = male,
2. race = white,
3. build = medium,
4. *clothes = {jeans, shirt},
5. upper-wear-color= {white},
6. bottom-wear-color = {blue}, and
7. relation = {wearing, †Person, *Clothes}.

Problem 2.1 (Human Attribute Recognition from Text). *Given a large text T with T_s sentences, each with $|w|$ tokens, the problem of human attribute recognition from T is to*

1. *identify the set of sentences $C_s \subset T_s$ that describes properties of a person,*
2. *expose the set of object-properties \mathcal{O}_H from C_s and*
3. *extract the set of values z_p of the identified properties \mathbf{o}_p .*

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”,

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

["clothes", "wearing"]

Key Phrases
describing Attributes

Regex

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

$$SIM(q_H, s) = \max_{w \in s} SIM(q_H, w)$$

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

$$SIM(q_H, s) = \max_{w \in s} SIM(q_H, w)$$

$$SIM(q_H, s) = P(s \text{ is about } q_H \mid B_s, B_{q_H})$$

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

$$SIM(q_H, s) = \max_{w \in s} SIM(q_H, w)$$

Word2Vec

$$SIM(q_H, s) = P(s \text{ is about } q_H \mid B_s, B_{q_H})$$

Wordnet

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

$$SIM(q_H, s) = \max_{w \in s} SIM(q_H, w)$$

$$SIM(q_H, s) = P(s \text{ is about } q_H \mid B_s, B_{q_H})$$

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

$$SIM(q_H, s) = \max_{w \in s} SIM(q_H, w)$$

$$SIM(q_H, s) = P(s \text{ is about } q_H \mid B_s, B_{q_H})$$

Classification Task
On Sentences

Candidate Sentence Extraction

Definition 2.3.3 (Candidate Sentences). *Given a collection of sentences T_s , key-phrase for describing an object in text $q_H \subset Q_H$, and an empirical threshold θ_H , Candidate sentence is*

$$C_s = \{s : s \in T_s, q_H \in Q_H \mid SIM(q_H, s) > \theta_H\} \quad (2.3)$$

“clothes”, “wearing”

Key Phrases
describing Attributes

$$SIM(q_H, s) = \max_{w \in s} SIM(q_H, w)$$

$$SIM(q_H, s) = P(s \text{ is about } q_H \mid B_s, B_{q_H})$$

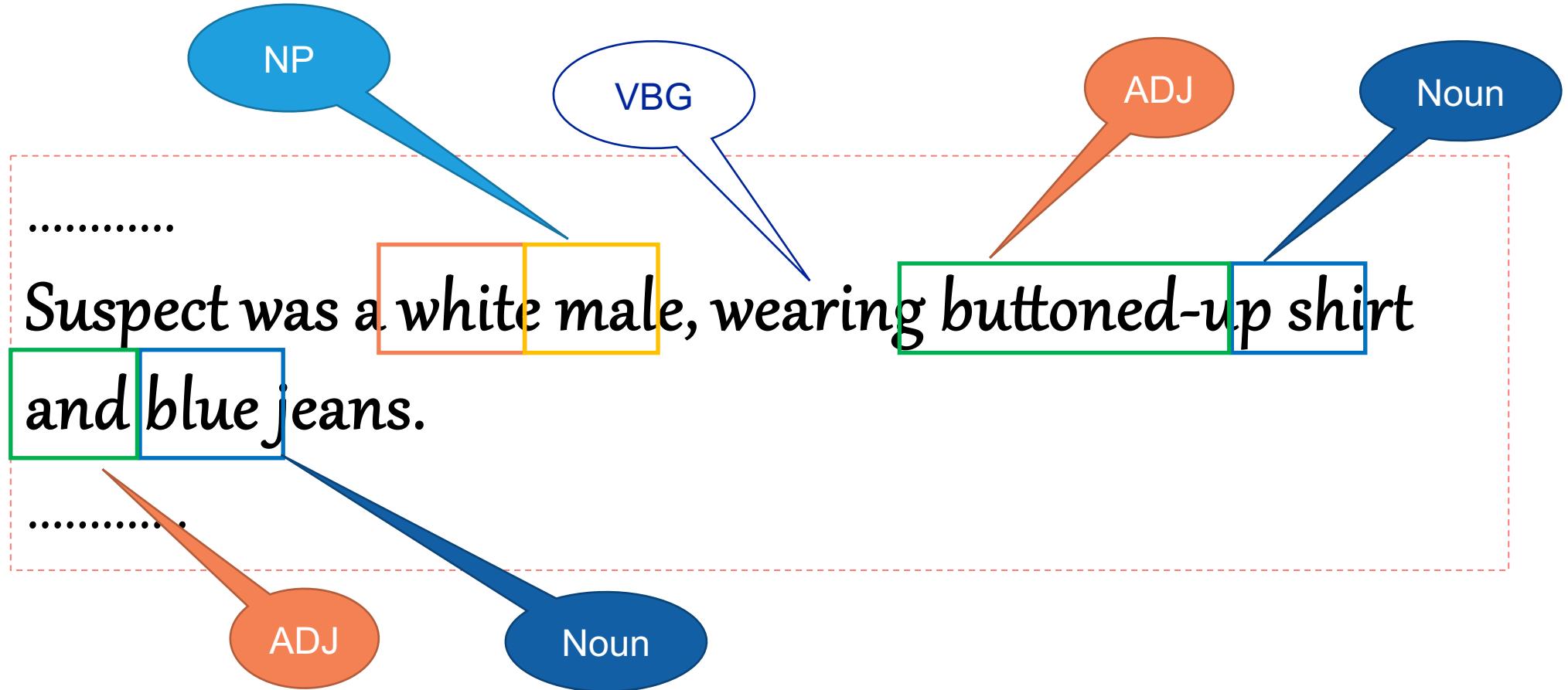
SBERT NLI
Classifier

Classification Task
On Sentences

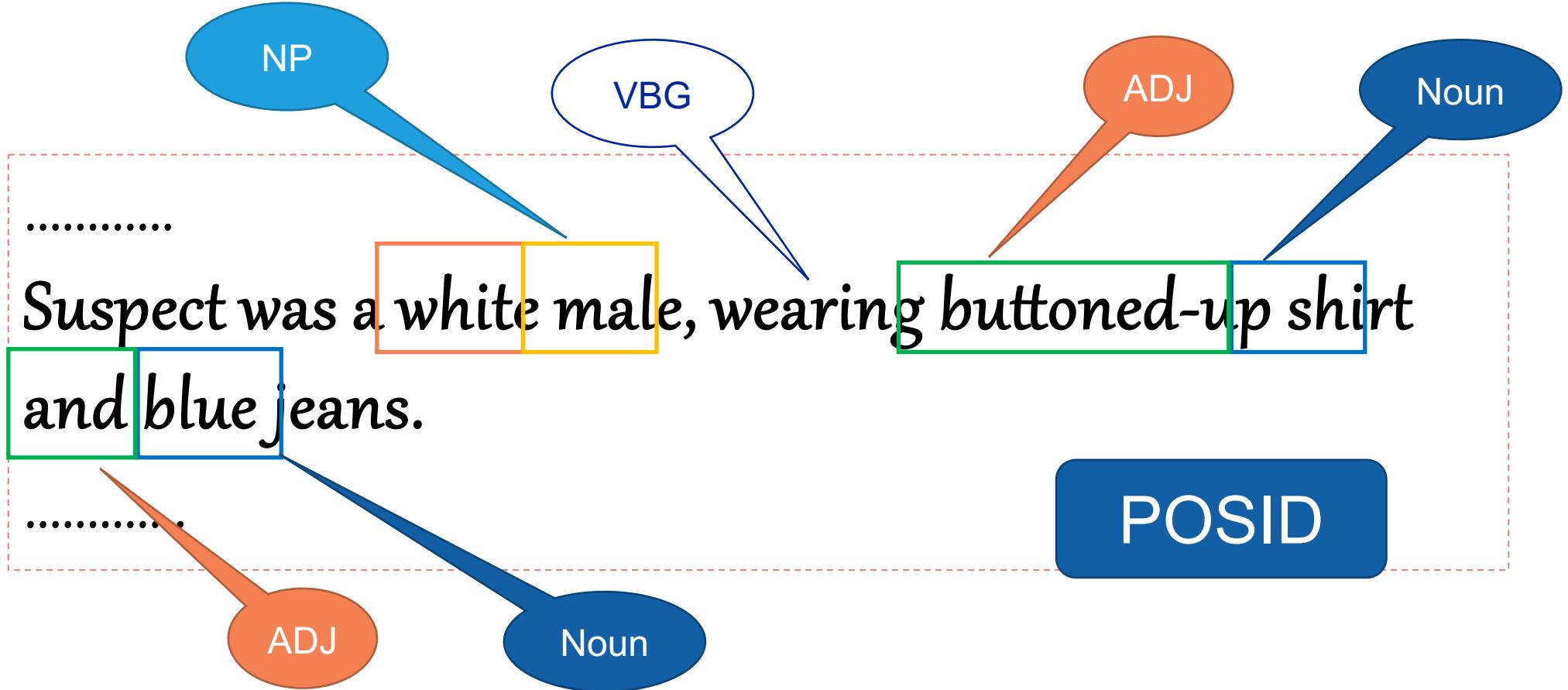
Feature Value Extraction

.....
Suspect was a white male, wearing buttoned-up shirt
and blue jeans.
.....

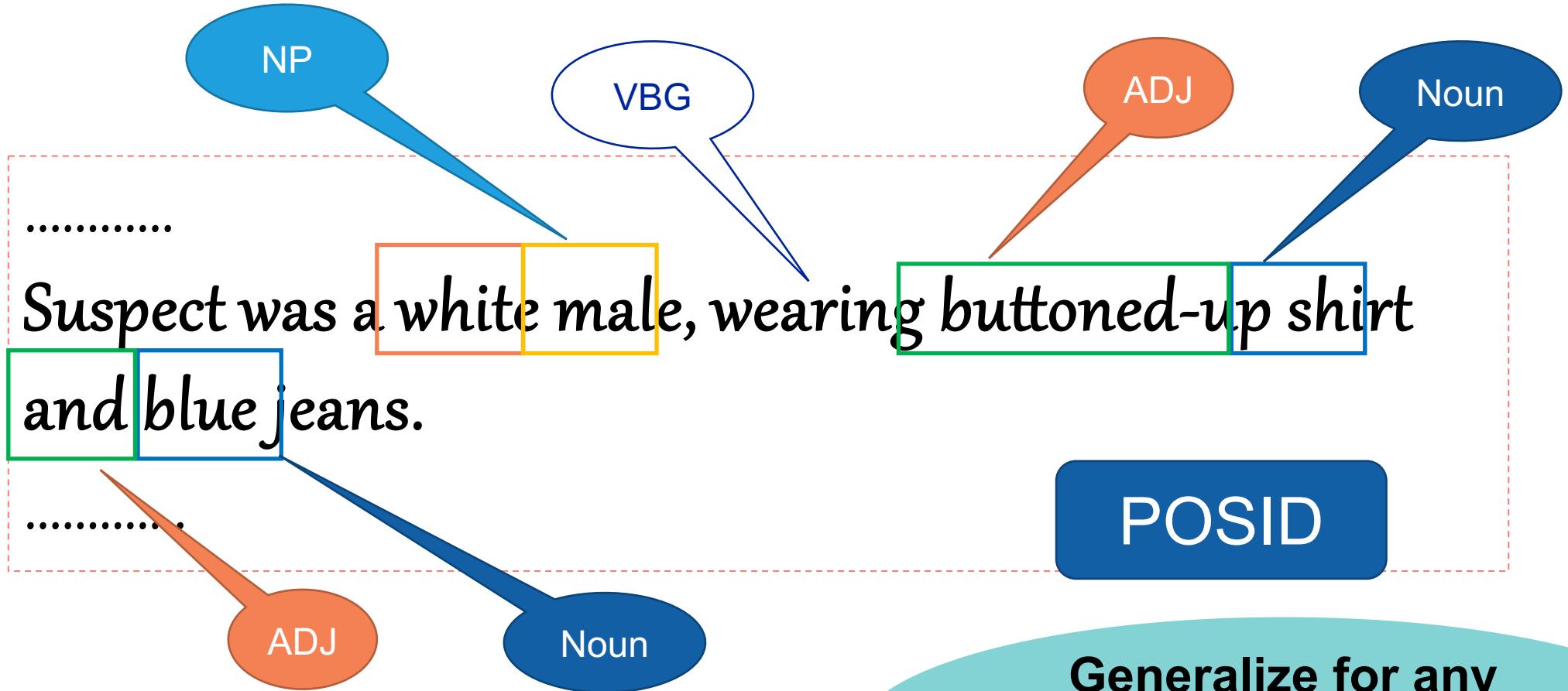
Feature Value Extraction



Feature Value Extraction



Feature Value Extraction



Generalize for any
domain
Based on Key Phrases

Demo + Evaluation Setup

- **Two Use Cases**
 - Video Querying System
 - Finding Missing Persons
- **GCP** for hosting
 - Data Storage (Postgres)
 - Property Identifiers
 - 9 Compute Engines
 - 1 SQL instance, 1 Storage Bucket
- Video Feature Extractor Benchmark
 - OS: Linux Ubuntu 18.04
 - GPU: Tesla K80 with 12GB Memory

Demo + Evaluation Setup

- **Two Use Cases**
 - Video Querying System
 - Finding Missing Persons
- **GCP for hosting**
 - Data Storage (Postgres)
 - Property Identifiers
 - 9 Compute Engines
 - 1 SQL instance, 1 Storage Bucket
- Video Feature Extractor Benchmark
 - OS: Linux Ubuntu 18.04
 - GPU: Tesla K80 with 12GB Memory



Missing from: Lehighton, PA • Date Missing: 04/13/2021 • Issue Date: 04/14/2021



Granvil Lang Jr.

Age: 79
Height: 5'5"
Weight: 180 lbs.
Hair: Brown / Gray
Eyes: Brown

- Lang has a gray beard.
- He is believed to be possibly wearing a flannel shirt, blue jeans and sneakers.



Tom Cunningham

13 years old, white, medium build. Last seen on 17th October 2013 wearing blue jeans, a blue hoody and a sleeveless bubble jacket. If you have seen this boy or know of his whereabouts, please contact us. If you have any information whatsoever, please call Dee Valley Police on this number: 08081 57 0243

DEE VALLEY POLICE

If you know where Tom is or have any information about him please contact Dee Valley Incident Room on 08081 57 0243

Demo + Evaluation Setup

- Two Use Cases
 - Video Querying System
 - Finding Missing Persons
- GCP for hosting
 - Data Storage (Postgres)
 - Property Identifiers
 - 9 Compute Engines
 - 1 SQL instance, 1 Storage Bucket
- Video Feature Extractor Benchmark
 - OS: Linux Ubuntu 18.04
 - GPU: Tesla K80 with 12GB Memory



Missing from: Lehighton, PA • Date Missing: 04/13/2021 • Issue Date: 04/14/2021



Granvil Lang Jr.

Age: 79
Height: 5'5"
Weight: 180 lbs.
Hair: Brown / Gray
Eyes: Brown

- Lang has a gray beard.
- He is believed to be possibly wearing a flannel shirt, blue jeans and sneakers.



Tom Cunningham

13 years old, white, medium build. Last seen on 17th October 2013 wearing blue jeans, a blue hoody and a sleeveless bubble jacket. If you have seen this boy or know of his whereabouts, please contact us. If you have any information whatsoever, please call Dee Valley Police on this number: 08081 57 0243

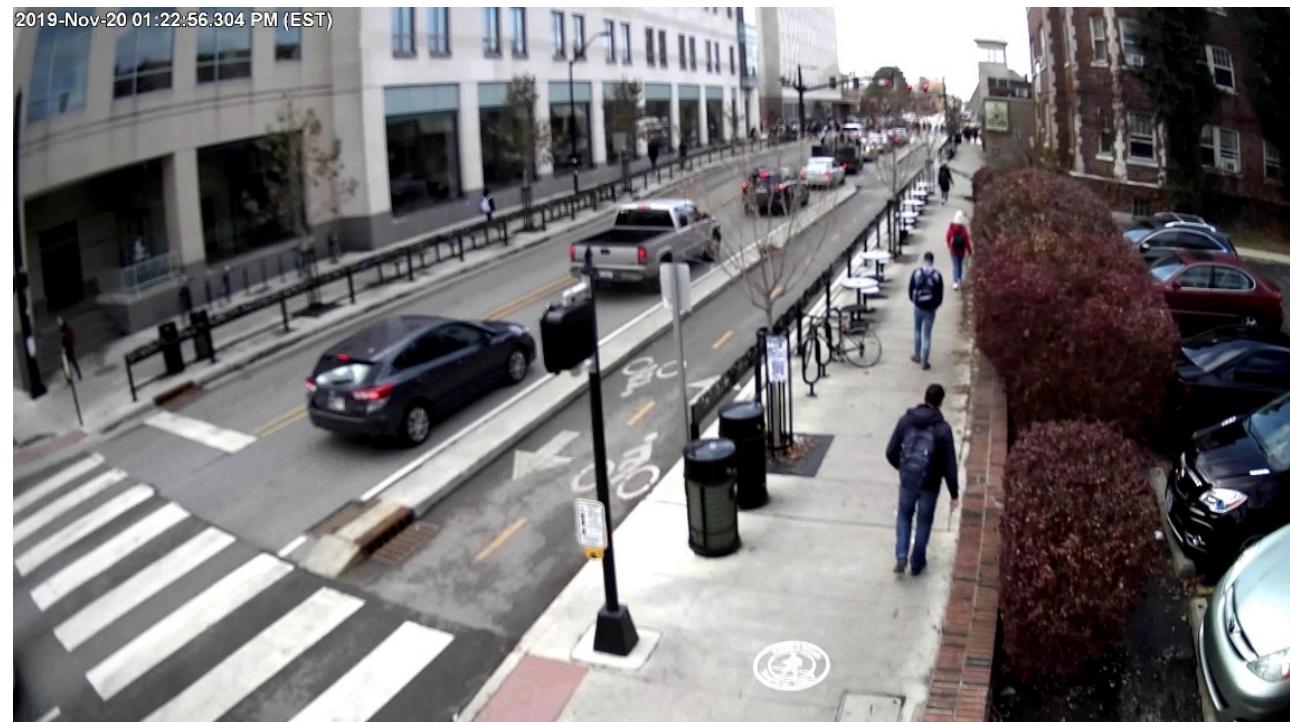
DEE VALLEY POLICE If you know where Tom is or have any information about him please contact Dee Valley Incident Room on 08081 57 0243

The **goal** is to demonstrate the feasibility of the proof of concept with *real* use cases deployed using **SKOD**

Mission needs for Video Querying System & Missing Person Search

Constant Attributes	Changeable Attributes	Other objects
Female	T-shirt	Car
Male	Shorts	Bicycle
White	Jeans	Truck
Black	Pants	Motorcycle
Hispanic	Jacket	Skateboard
Asian	Shoes	Backpack

Activity Recognition	Additional attributes
Smoking	Hair color
Running	Tattoo
Walking	Beard
	Bald
	Tall/short
	Headphones



Mission needs for Video Querying System & Missing Person Search

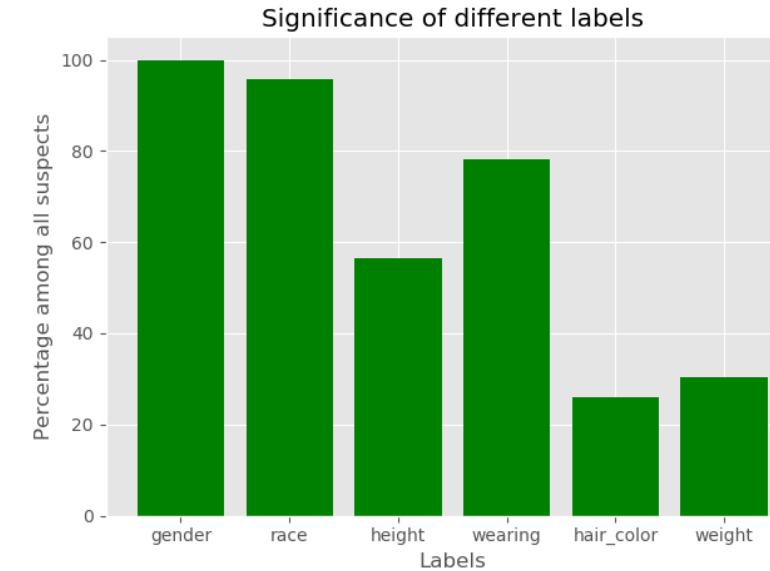
Constant Attributes	Changeable Attributes	Other objects
Female	T-shirt	Car
Male	Shorts	Bicycle
White	Jeans	Truck
Black	Pants	Motorcycle
Hispanic	Jacket	Skateboard
Asian	Shoes	Backpack

Activity Recognition	Additional attributes
Smoking	Hair color
Running	Tattoo
Walking	Beard
	Bald
	Tall/short
	Headphones

Mission needs for Video Querying System & Missing Person Search

Constant Attributes	Changeable Attributes	Other objects
Female	T-shirt	Car
Male	Shorts	Bicycle
White	Jeans	Truck
Black	Pants	Motorcycle
Hispanic	Jacket	Skateboard
Asian	Shoes	Backpack

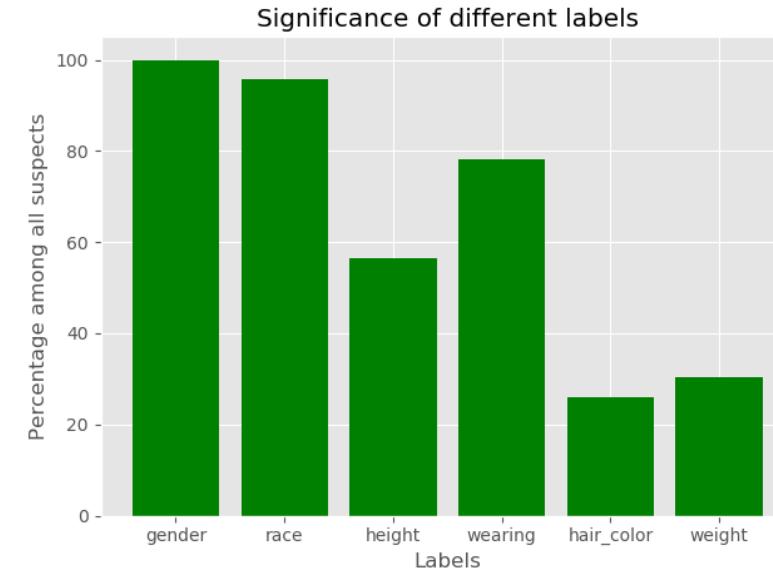
Activity Recognition	Additional attributes
Smoking	Hair color
Running	Tattoo
Walking	Beard
	Bald
	Tall/short
	Headphones



Mission needs for Video Querying System & Missing Person Search

Constant Attributes	Changeable Attributes	Other objects
Female	T-shirt	Car
Male	Shorts	Bicycle
White	Jeans	Truck
Black	Pants	Motorcycle
Hispanic	Jacket	Skateboard
Asian	Shoes	Backpack

Activity Recognition	Additional attributes
Smoking	Hair color
Running	Tattoo
Walking	Beard
	Bald
	Tall/short
	Headphones



- Data Annotation
 - Gender, Race, Age, Hair Color,
 - Clothing (jacket/pants/jeans) and Cloth-details
 - Multiple persons are described in same document and annotated separately

Datasets

Location

- Real World Datasets
- Cambridge, MA
- West Lafayette, IN
- WLPD

Datasets

Location

- Real World Datasets
- Cambridge, MA
- West Lafayette, IN
- WLPD

Types of Data

- Traffic Camera Video
- Dashcam Video
- Incident Reports
- Officer narrative
- Press Release
- Newspaper Articles
- Tweets
- City Management Data
(Structured Data)

Dataset Size

- ~600K tweets
- 100+ hours of dashcam video
- 10+ hours of Traffic Cam Video
- 2000+ pairs Multimodal Data
- WLPD

Experimental Results and Findings

RQ1: Which is a better representation of tokens for finding key phrases – lexical or contextual model?

1. Attribute-Only: #identified clothes
2. Attribute-Value: #details of identified clothes

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^ .]*wear[^ .]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Experimental Results and Findings

RQ1: Which is a better representation of tokens for finding key phrases – lexical or contextual model?

1. Attribute-Only: #identified clothes
2. Attribute-Value: #details of identified clothes

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^.]*wear[^.]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Experimental Results and Findings

RQ1: Which is a better representation of tokens for finding key phrases – lexical or contextual model?

1. Attribute-Only: #identified clothes
2. Attribute-Value: #details of identified clothes

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^.]*wear[^.]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Findings: Wordnet understands similarity in meaning better than word2vec, as we can see we lowered threshold for word2vec quite a lot, but still Wordnet + POS worked better.

Experimental Results and Findings

RQ2: Similarity Search or Classification – which model identifies C_s better?

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^.]*wear[^.]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Experimental Results and Findings

RQ2: Similarity Search or Classification – which model identifies C_s better?

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^ .]*wear[^ .]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Experimental Results and Findings

RQ2: Similarity Search or Classification – which model identifies C_s better?

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^.]*wear[^.]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Findings: Although SBERT identifies lesser number of relevant items than similarity search with WordNet, higher recall indicates SBERT identified lesser number of irrelevant data than WordNet.

Experimental Results and Findings

RQ3: Does the stacked model increase the possibility of finding C_s (Regex Effect)?

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^ .]*wear[^ .]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Experimental Results and Findings

RQ3: Does the stacked model increase the possibility of finding C_s (Regex Effect)?

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^ .]*wear[^ .]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Experimental Results and Findings

RQ3: Does the stacked model increase the possibility of finding C_s (Regex Effect)?

Table 5: Clothes Attribute Extraction Result for Different Candidate Sentence Extraction Models

Models	Attr-Only			Attr-Value			θ	q
	Precision	Recall	F1-Score	Precision	Recall	F1-Score		
Word2Vec + POS	0.83	0.38	0.52	0.85	0.35	0.49	0.5	clothes
RE + POS	0.86	0.82	0.84	0.92	0.82	0.87	X	[^ .]*wear[^ .]*
WordNet + POS	0.93	0.33	0.49	0.89	0.30	0.45	0.9	<i>clothes</i> as noun
SBERT + POS	0.83	0.49	0.62	0.86	0.45	0.59	0.85	clothes
RE + WordNet + POS	0.93	0.65	0.77	0.92	0.87	0.90	0.9	<i>clothes</i> as noun
RE + SBERT + POS	0.87	0.87	0.87	0.92	0.87	0.90	0.85	clothes

Findings: Most definitely Regex search combined with another layer of WordNet/SBERT increases the possibility of finding the correct C_s as unstructured text will most definitely would have noisy texture and style.

Experimental Results and Findings

RQ4: Which one is the best method for video attribute detection in Person Querying System?

* MARS (Motion Analysis and Re-identification Set)

Table 7: Comparisons of recognition accuracy and F1 measure on MARS datasets(%).

Attribute	CNN (Resnet50) ⁶		3D-CNN		CNN-RNN		Temporal Pooling ⁷		Temporal Attention ⁸		Color Sampling	
	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1
top color	75.22	73.98	67.91	65.19	70.54	67.33	74.98	73.13	76.05	74.64	44.65	38.31
bottom color	73.55	54.09	59.77	36.56	67.71	44.44	71.69	47.84	70.15	46.89	45.26	15.88
gender	90.01	89.71	86.49	76.22	90.07	89.62	91.04	90.63	91.82	91.48	-	-
average	79.59	72.59	67.97	59.18	76.11	67.13	79.24	70.53	79.34	71.01	44.96	27.10

Findings:

- If only static features are of interest i.e., gender, CNN-based object detectors are enough.
- If static and dynamic features (have effects from time and space) are of interest, CNN with temporal attention works better.

* K. Solaiman, T. Sun, A. Nesen, B. Bhargava and M. Stonebraker, "Applying Machine Learning and Data Fusion to the Missing Person Problem" in IEEE Computer, vol. 55, no. 06, pp. 40-55, 2022.

Experimental Results and Findings

RQ4: Which one is the best method for video attribute detection in Person Querying System?

* MARS (Motion Analysis and Re-identification Set)

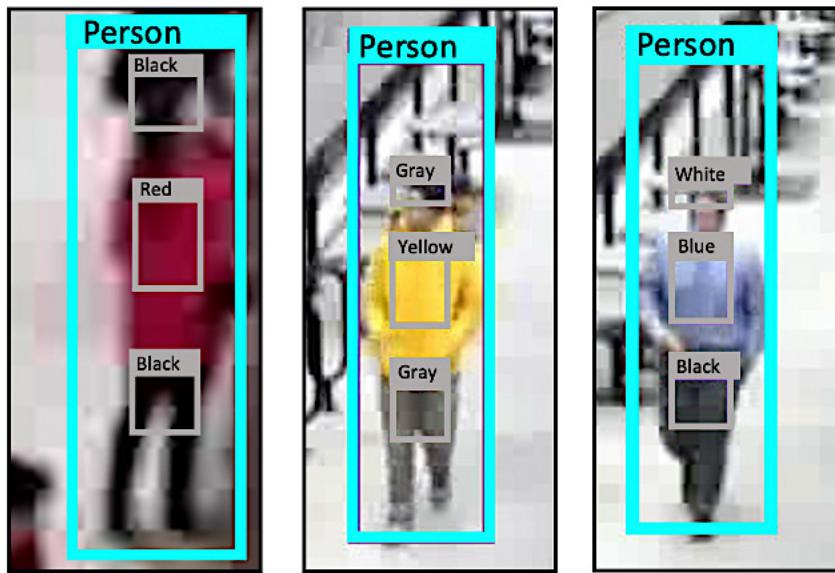
Table 7: Comparisons of recognition accuracy and F1 measure on MARS datasets(%).

Attribute	CNN (Resnet50) ⁶		3D-CNN		CNN-RNN		Temporal Pooling ⁷		Temporal Attention ⁸		Color Sampling	
	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1
top color	75.22	73.98	67.91	65.19	70.54	67.33	74.98	73.13	76.05	74.64	44.65	38.31
bottom color	73.55	54.09	59.77	36.56	67.71	44.44	71.69	47.84	70.15	46.89	45.26	15.88
gender	90.01	89.71	86.49	76.22	90.07	89.62	91.04	90.63	91.82	91.48	-	-
average	79.59	72.59	67.97	59.18	76.11	67.13	79.24	70.53	79.34	71.01	44.96	27.10

* K. Solaiman, T. Sun, A. Nesen, B. Bhargava and M. Stonebraker, "Applying Machine Learning and Data Fusion to the Missing Person Problem" in IEEE Computer, vol. 55, no. 06, pp. 40-55, 2022.

Experimental Results and Findings

RQ4: Which one is the best method for video attribute detection in Person Querying System?



Resnet50 needed 513 minutes and temporal attention model needed 1073 minutes for training

recognition accuracy and F1 measure on MARS datasets(%).

1	CNN-RNN		Temporal Pooling ⁷		Temporal Attention ⁸		Color Sampling	
	acc	F1	acc	F1	acc	F1	acc	F1
5.19	70.54	67.33	74.98	73.13	76.05	74.64	44.65	38.31
6.56	67.71	44.44	71.69	47.84	70.15	46.89	45.26	15.88
6.22	90.07	89.62	91.04	90.63	91.82	91.48	-	-
9.18	76.11	67.13	79.24	70.53	79.34	71.01	44.96	27.10

Findings:

- Color Sampling can be a good solution for cold start problem, but compared to neural network models the performance is weak.

* K. Solaiman, T. Sun, A. Nesen, B. Bhargava and M. Stonebraker, "Applying Machine Learning and Data Fusion to the Missing Person Problem" in IEEE Computer, vol. 55, no. 06, pp. 40-55, 2022.

Experimental Results and Findings

RQ4: Which one is the best method for video attribute detection in Person Querying System?

Resnet50 needed 513 minutes and temporal attention model
needed 1073 minutes for training

Table 7: Comparisons of recognition accuracy and F1 measure on MARS datasets(%).

Attribute	CNN (Resnet50) ⁶		3D-CNN		CNN-RNN		Temporal Pooling ⁷		Temporal Attention ⁸		Color Sampling	
	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1	acc	F1
top color	75.22	73.98	67.91	65.19	70.54	67.33	74.98	73.13	76.05	74.64	44.65	38.31
bottom color	73.55	54.09	59.77	36.56	67.71	44.44	71.69	47.84	70.15	46.89	45.26	15.88
gender	90.01	89.71	86.49	76.22	90.07	89.62	91.04	90.63	91.82	91.48	-	-
average	79.59	72.59	67.97	59.18	76.11	67.13	79.24	70.53	79.34	71.01	44.96	27.10

Findings:

- On average, CNN with temporal attention is a better method for human attribute detection from videos.

Experimental Results and Findings

RQ5: Will SKOD / SurvQ be performant enough to handle an urban camera deployment, including the West Lafayette use case?

- In SurvQ, a trigger function for each incident is invoked every time YOLO results are inserted.
- YOLO runs on 60 frames per minute
- Avg. 5 persons per frame
- At most 200 video sources
- $\sim 60 * 5 * 200 = \mathbf{60000 \text{ inserts per minute}}$
- **Handle under 60 seconds**

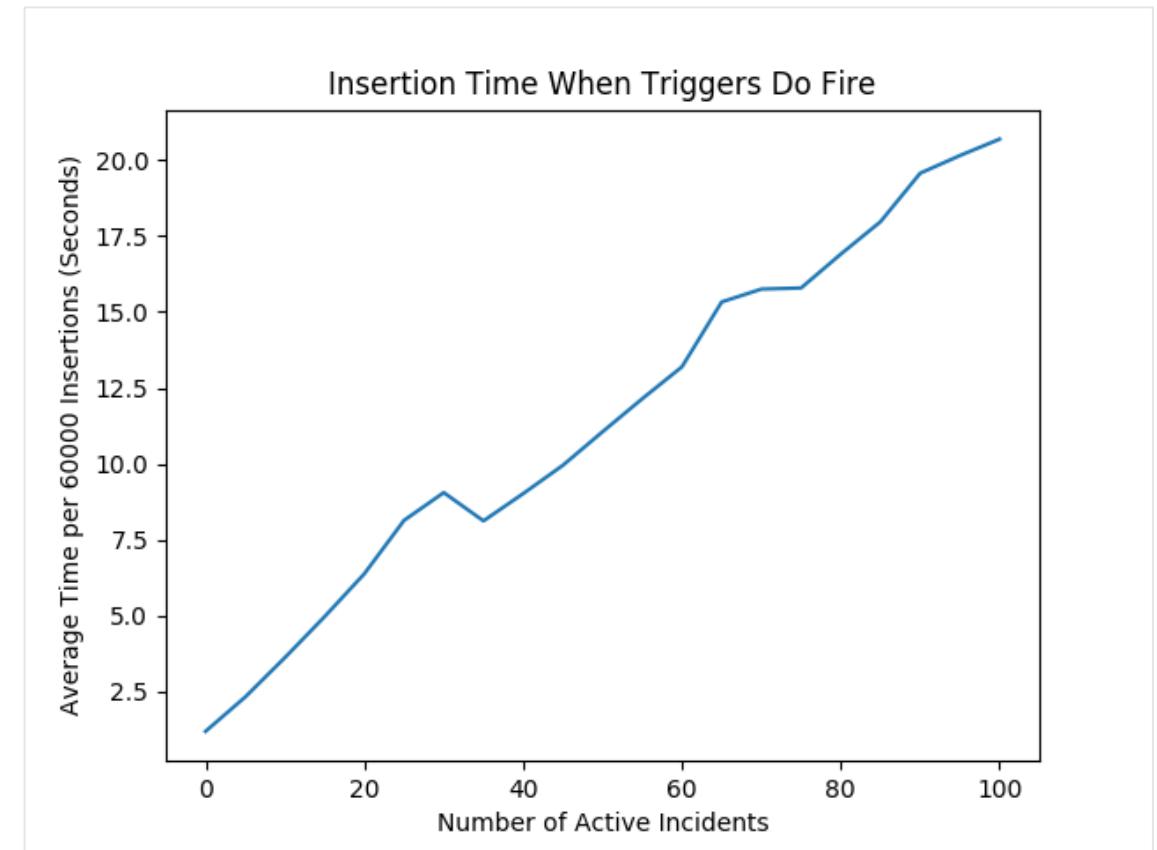


* M. Stonebraker et al, *Surveillance Video Querying with a Human-in-the-loop*, in HILDA, SIGMOD 2020

Experimental Results and Findings

RQ5: Will SKOD / SurvQ be performant enough to handle an urban camera deployment, including the West Lafayette use case?

- In SurvQ, a trigger function for each incident is invoked every time YOLO results are inserted.
- YOLO runs on 60 frames per minute
- Avg. 5 persons per frame
- At most 200 video sources
- $\sim 60 * 5 * 200 = \mathbf{60000 \text{ inserts per minute}}$
- **Handle under 60 seconds**



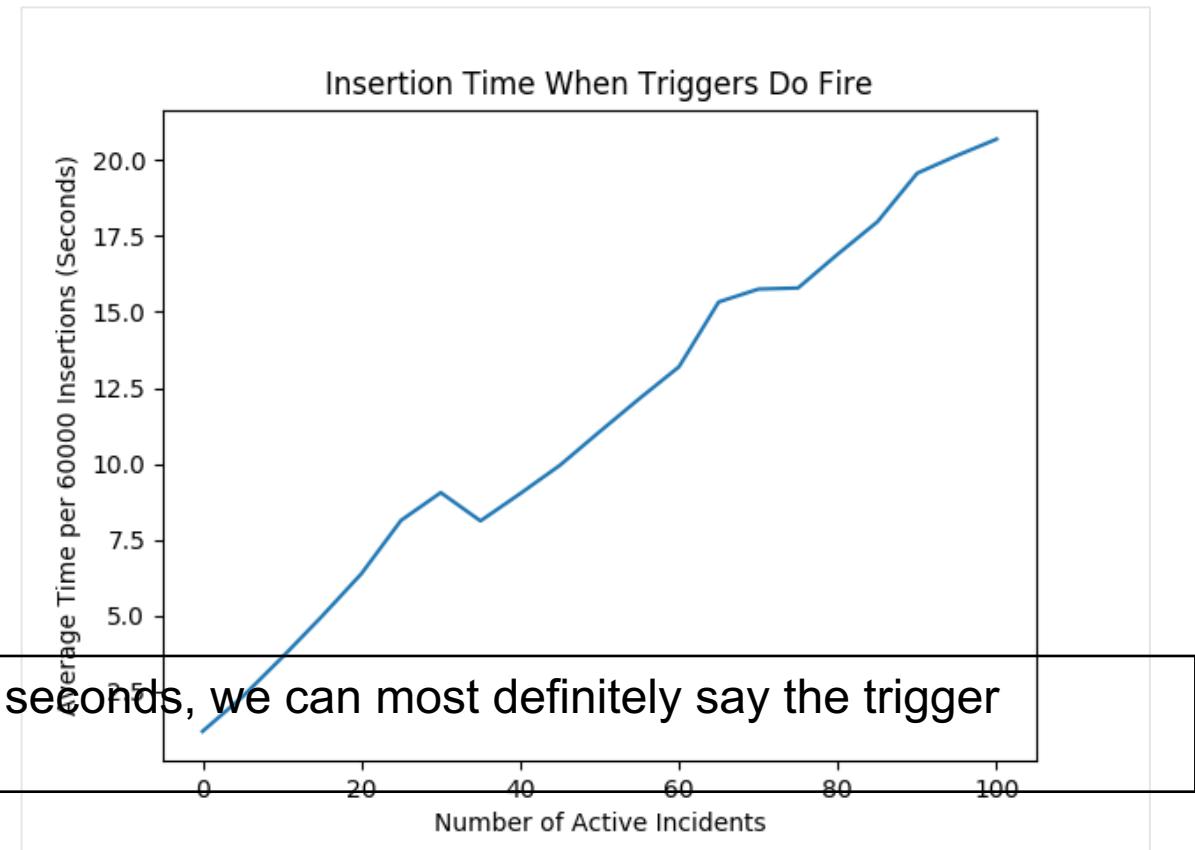
* M. Stonebraker et al, *Surveillance Video Querying with a Human-in-the-loop*, in HILDA, SIGMOD 2020

Experimental Results and Findings

RQ5: Will SKOD / SurvQ be performant enough to handle an urban camera deployment, including the West Lafayette use case?

- In SurvQ, a trigger function for each incident is invoked every time YOLO results are inserted.
- YOLO runs on 60 frames per minute
- Avg. 5 persons per frame
- At most 200 video sources
- $\sim 60 * 5 * 200 = \mathbf{60000 \text{ inserts per minute}}$
- **Handle under 60 seconds**

Findings: Since avg time/60K inserts is well under 60 seconds, we can most definitely say the trigger invocation can easily keep up, and so would SKOD.



* M. Stonebraker et al, *Surveillance Video Querying with a Human-in-the-loop*, in HILDA, SIGMOD 2020

Experimental Results and Findings

RQ6: Can SKOD framework be generalized?



Urban Information
System

A screenshot of a Video Querying System interface. On the left, there is a video player showing a street scene at 0:00 / 1:00. On the right, there is a list of recorded surveillance videos:

- Nov 14th 11:53 pm at 320 W State St, West Lafayette
- Nov 15th 2:14 pm at 200 W State St, West Lafayette
- Nov 15th 2:16 pm at 200 W State St, West Lafayette
- Nov 15th 2:38 pm at 200 W State St, West Lafayette

Video Querying System

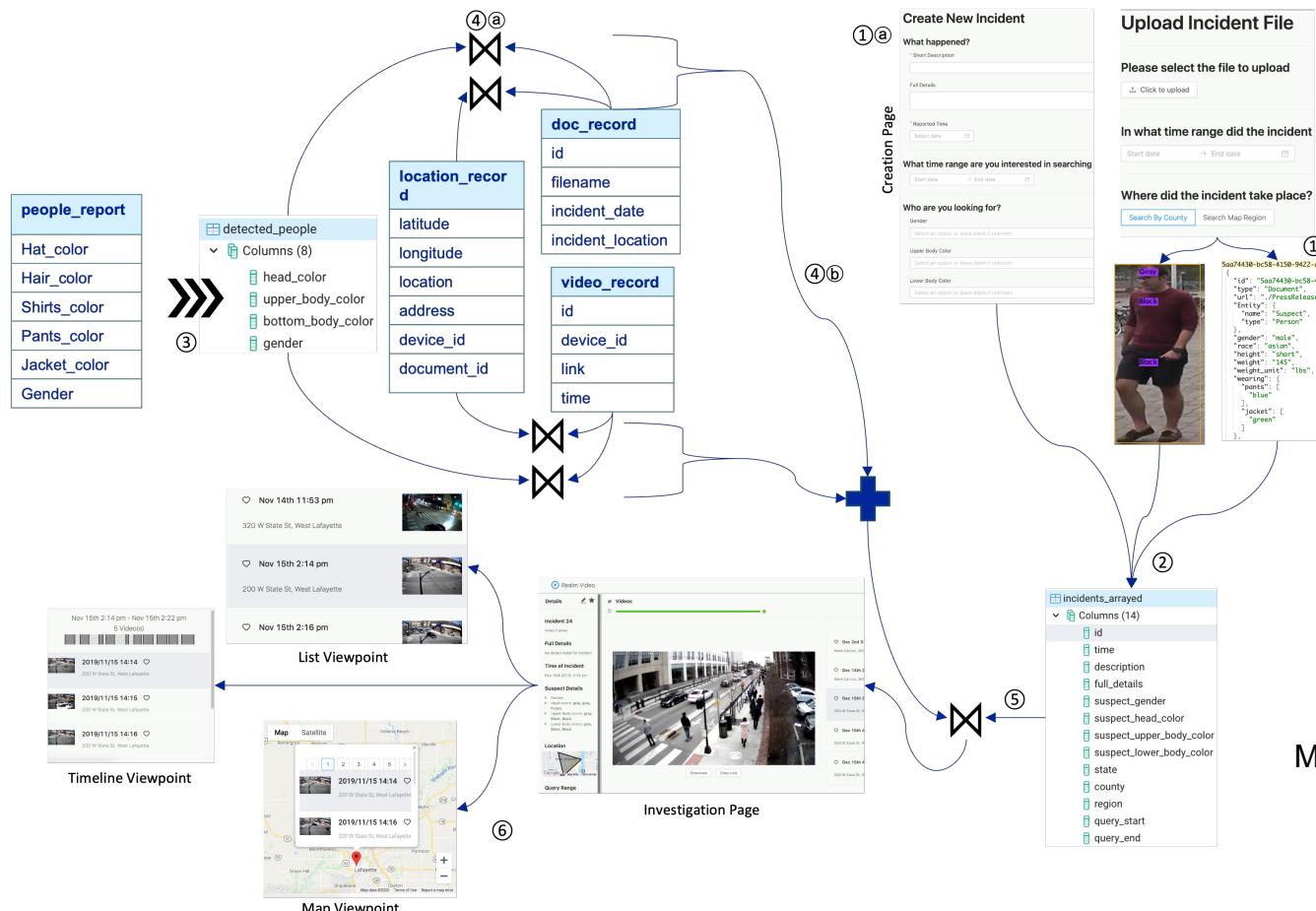
- S. Palacios and K. Solaiman** et al. *SKOD: A Framework for Situational Knowledge on Demand*. In Poly 2019, at VLDB 2019, August 2019.
- M. Stonebraker et al, *Surveillance Video Querying with a Human-in-the-loop*, in HILDA, SIGMOD 2020.
- K. Solaiman, T. Sun, A. Nesen, B. Bhargava and M. Stonebraker, "Applying Machine Learning and Data Fusion to the Missing Person Problem" in IEEE Computer, vol. 55, no. 06, pp. 40-55, 2022.

Experimental Results and Findings

RQ6: Can SKOD framework be generalized?



solving missing person problem + Vt's 1947 cold case



• S. Palacios and K. Solaiman** et al. *SKOD: A Framework for Situational Knowledge on Demand*. In Poly 2019, at VLDB 2019, August 2019.

• M. Stonebraker et al, *Surveillance Video Querying with a Human-in-the-loop*, in HILDA, SIGMOD 2020.

• K. Solaiman, T. Sun, A. Nesen, B. Bhargava and M. Stonebraker, "Applying Machine Learning and Data Fusion to the Missing Person Problem" in IEEE Computer, vol. 55, no. 06, pp. 40-55, 2022.

Research Contributions

- Novel **scalable, real-time** situational knowledge **extraction and dissemination engine** that can process multi-modal knowledge and can deliver information need over time from temporarily absent modalities
- First **attribute recognition model** from unstructured large text
- Novel **real-world datasets** for multimodal information retrieval with image, text and video
- Prototypes with a variety of applications
 - **Surveillance Video Querying**
 - **Finding Missing Persons**
 - Medical Triage, Search and Rescue in Disasters, Classroom teaching

Research Contributions

- Novel **scalable, real-time** situational knowledge **extraction and dissemination engine** that can process multi-modal knowledge and can deliver information need over time from temporarily absent modalities
- First **attribute recognition model** from unstructured large text
- Novel **real-world datasets** for multimodal information retrieval with image, text and video
- Prototypes with a variety of applications
 - **Surveillance Video Querying**
 - **Finding Missing Persons**
 - Medical Triage, Search and Rescue in Disasters, Classroom teaching



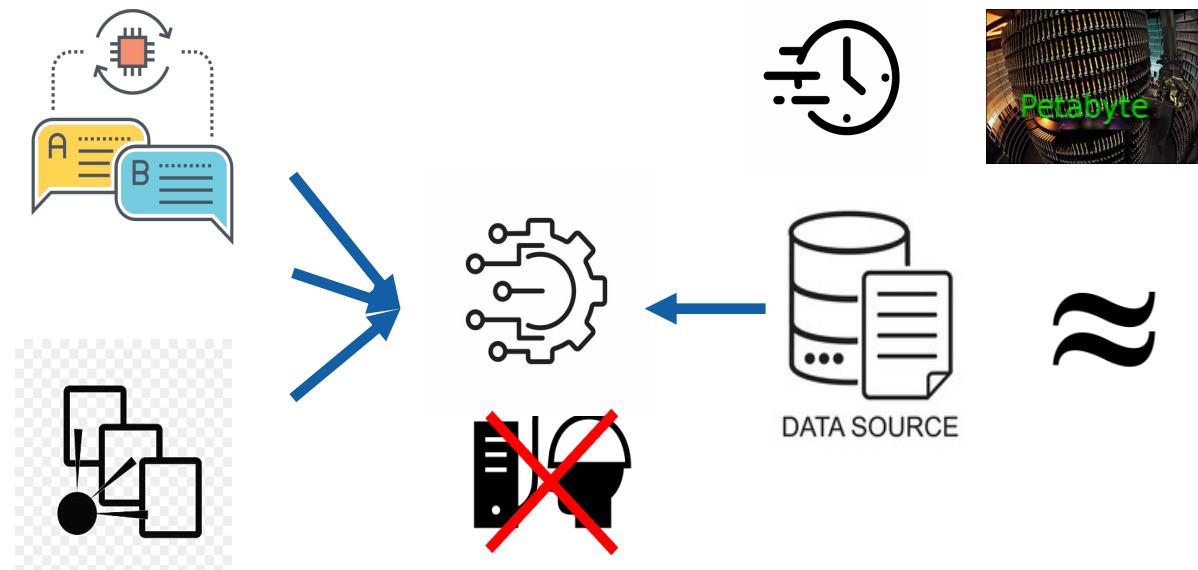
Research Contributions

- Novel **scalable, real-time** situational knowledge **extraction and dissemination engine** that can process multi-modal knowledge and can deliver information need over time from temporarily absent modalities
- First **attribute recognition model** from unstructured large text
- Novel **real-world datasets** for multimodal information retrieval with image, text and video
- Prototypes with a variety of applications
 - **Surveillance Video Querying**
 - **Finding Missing Persons**
 - Medical Triage, Search and Rescue in Disasters, Classroom teaching



Impact: SKOD can **automate** open problems that required **large-scale human endeavor and resources**, including in **data discovery and decision-making**

Contribution #2: Label Independent Data Integration



Problem Statement and Motivation

- Desirable properties for Data Integration in Decision Making
 - Generalizable across modalities
 - Learnable and should use past experiences
 - Scalable

Problem Statement and Motivation

- Desirable properties for Data Integration in Decision Making
 - Generalizable across modalities
 - Learnable and should use past experiences
 - Scalable
- Data Integration for SKOD should
 - Consider context information (query example) in relevance matching
 - Use high-level semantic features to be compatible with SKOD framework
 - Have Approximate matching and Ranking
 - Correlation learning, Metric learning, Encoding Networks fail to do so for situational knowledge, and rely on labels

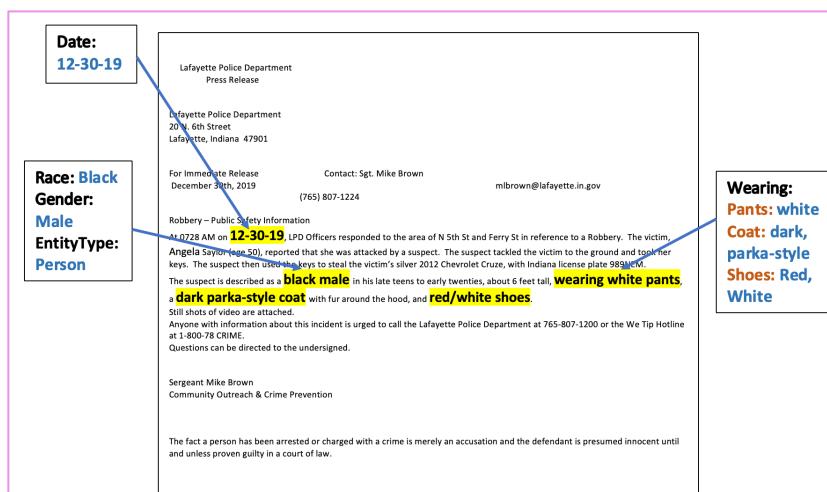
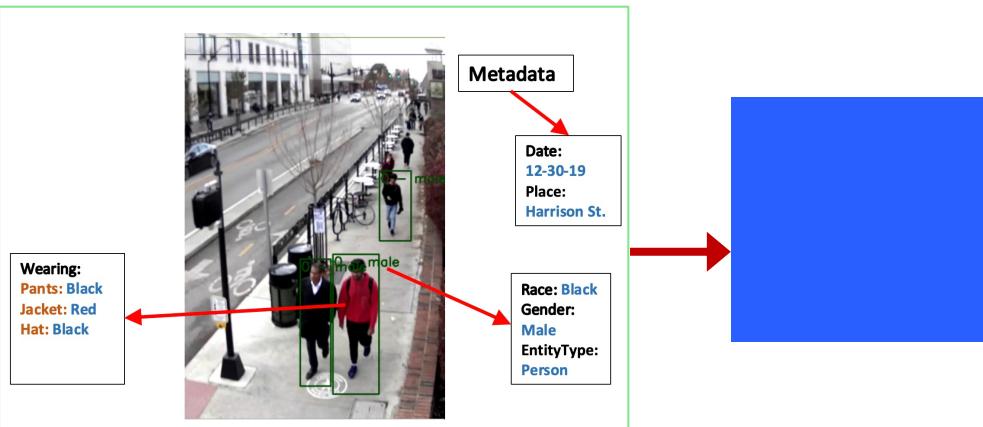
Problem Statement and Motivation

- Desirable properties for Data Integration in Decision Making

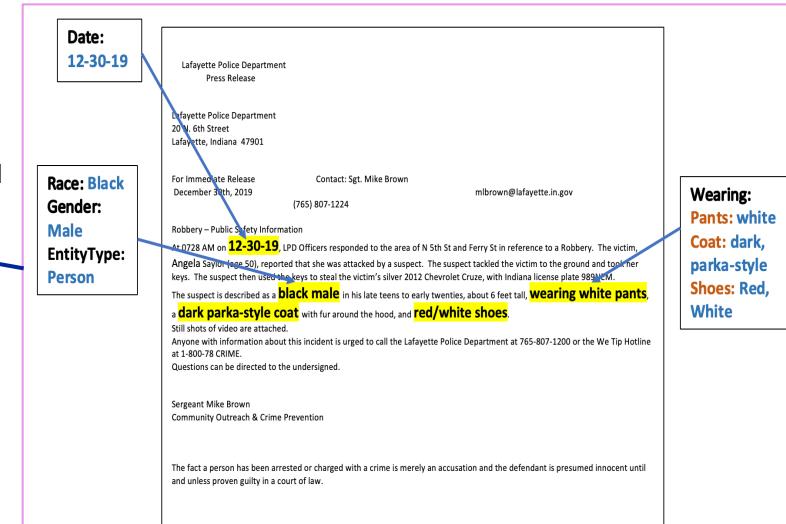
- Retrieve a ranked list, $R = (d_{x_1}, d_{x_2}, \dots d_{x_t})$ of t data-samples from all available modalities satisfying $PROP(d_m)$, where d_m is query data.
 - $PROP(d_j)$ is a relation that maps a data-sample d_j to a set of features.
- Relevance $SIM(d_{x_c}, d_m)$ is scored based on the degree of common features between the data-object d_{x_c} in the ranked list, and the query data d_m , $PROP(d_m) \cap PROP(d_{x_c})$.
- Have Approximate matching and Ranking
- $0 \leq SIM(d_{x_c}, d_m) \leq 1$.

Relevance Matching with Graph Representation Learning

Streaming
Input or
Data-at-
rest

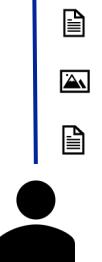


Relevance
Matching



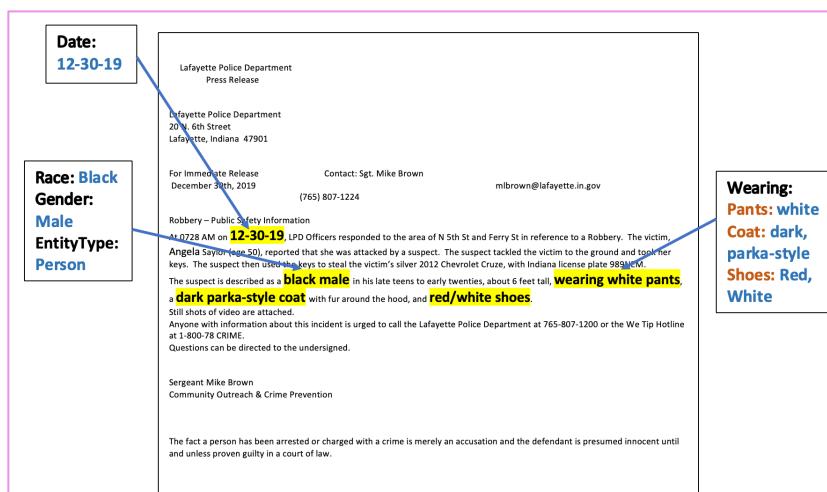
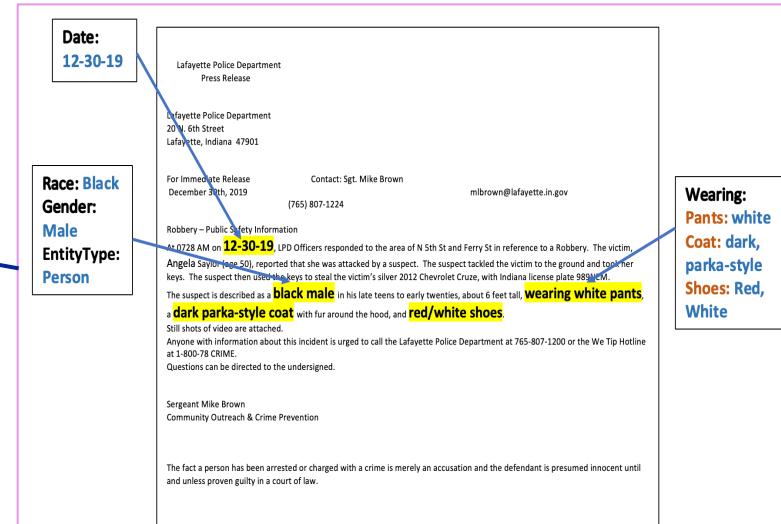
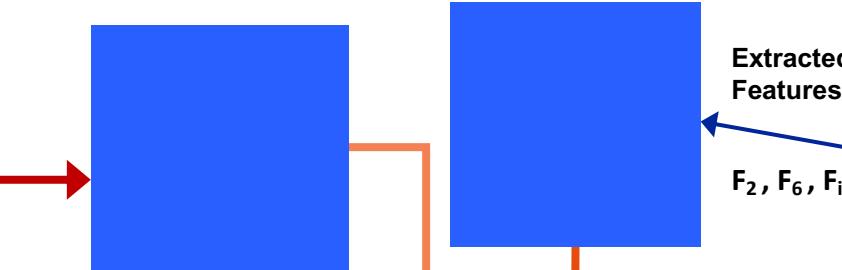
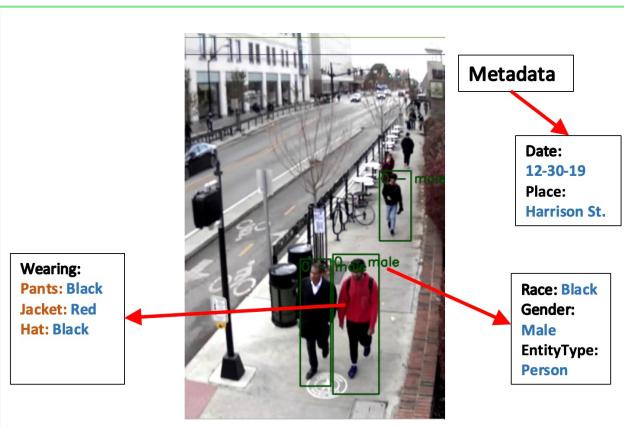
Features are extracted
from raw input

User
Input



Relevance Matching with Graph Representation Learning

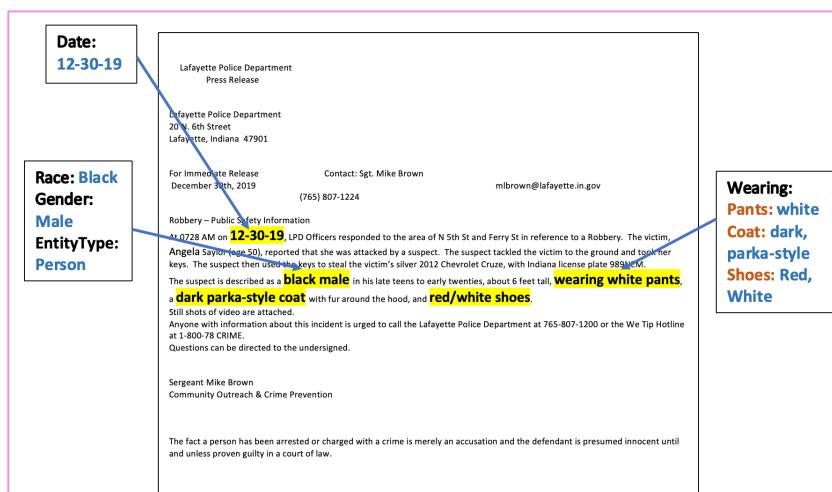
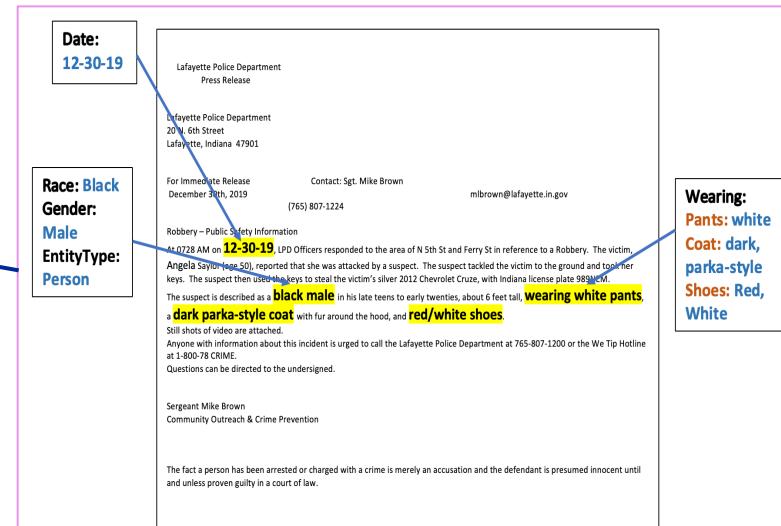
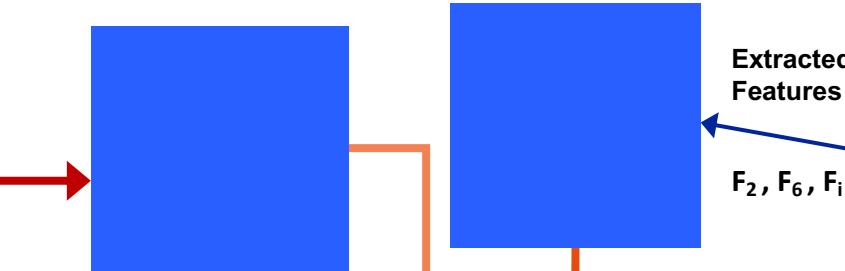
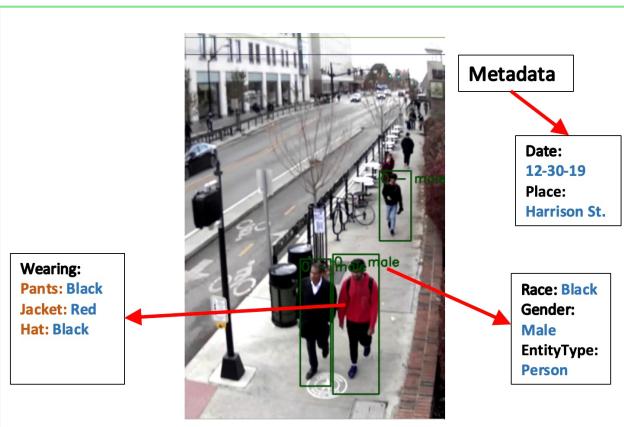
Streaming
Input or
Data-at-
rest



Features are extracted
from raw input

Relevance Matching with Graph Representation Learning

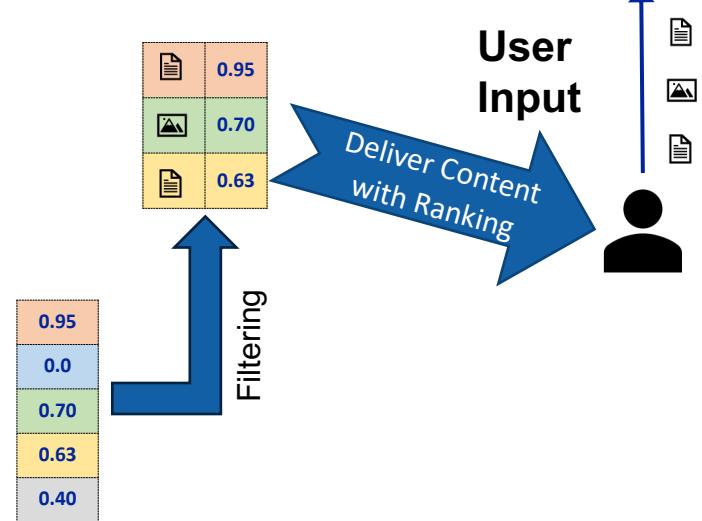
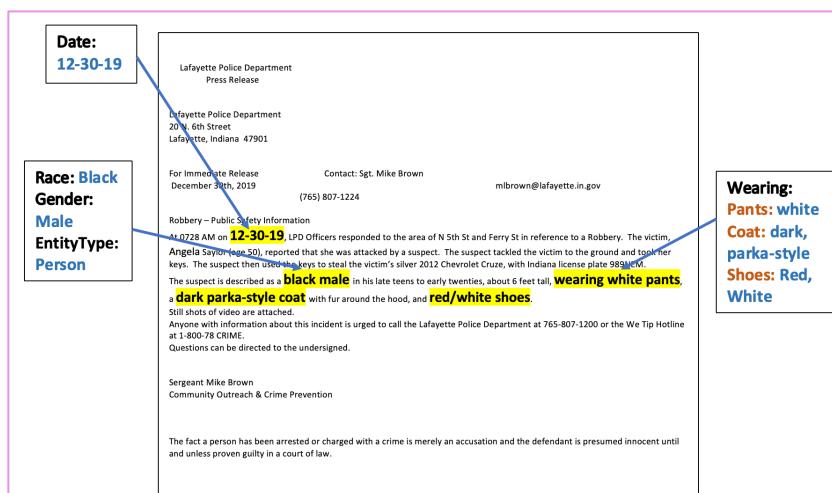
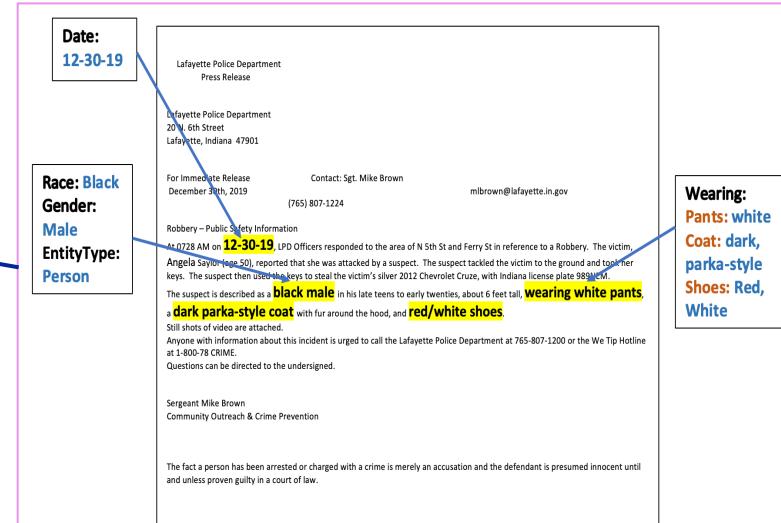
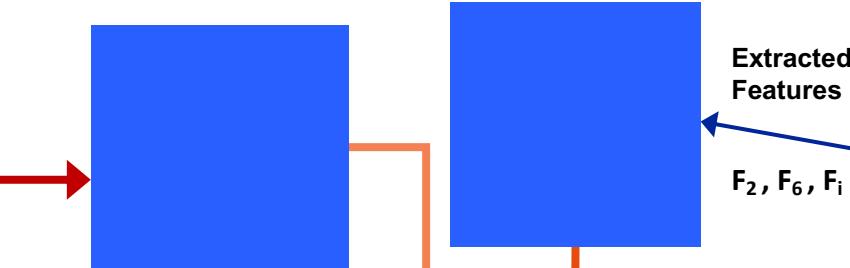
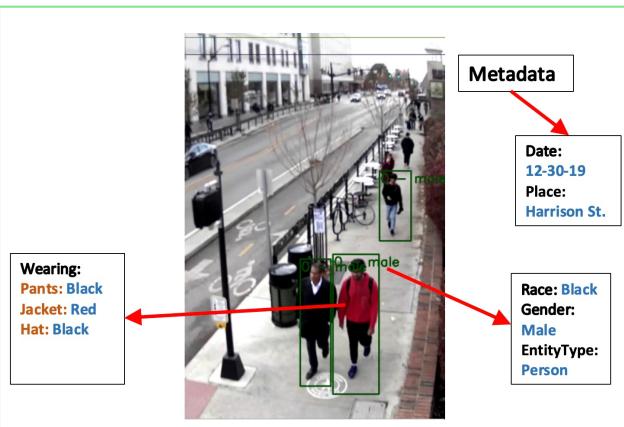
Streaming
Input or
Data-at-
rest



Features are extracted
from raw input

Relevance Matching with Graph Representation Learning

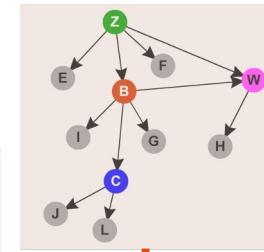
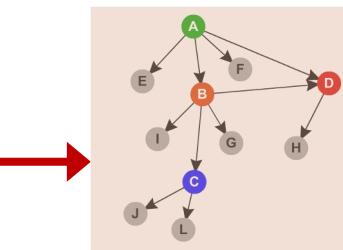
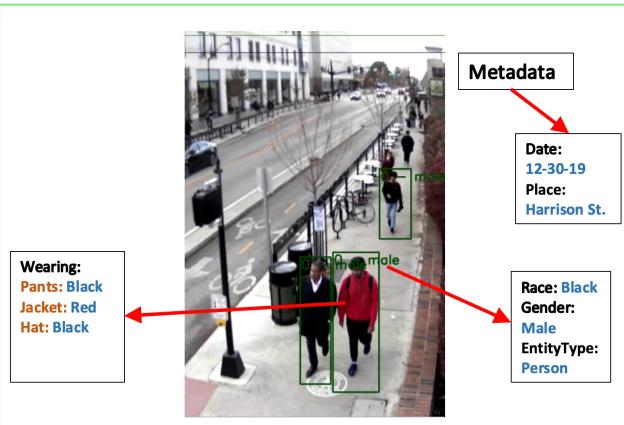
Streaming
Input or
Data-at-
rest



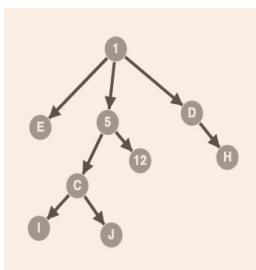
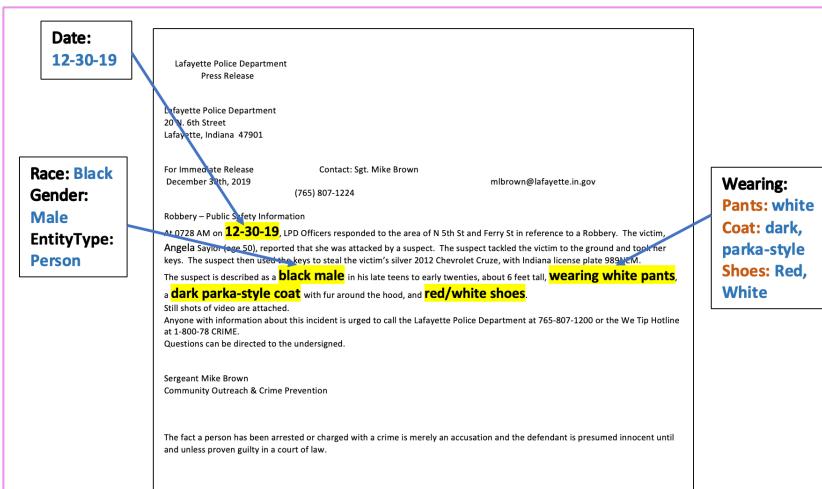
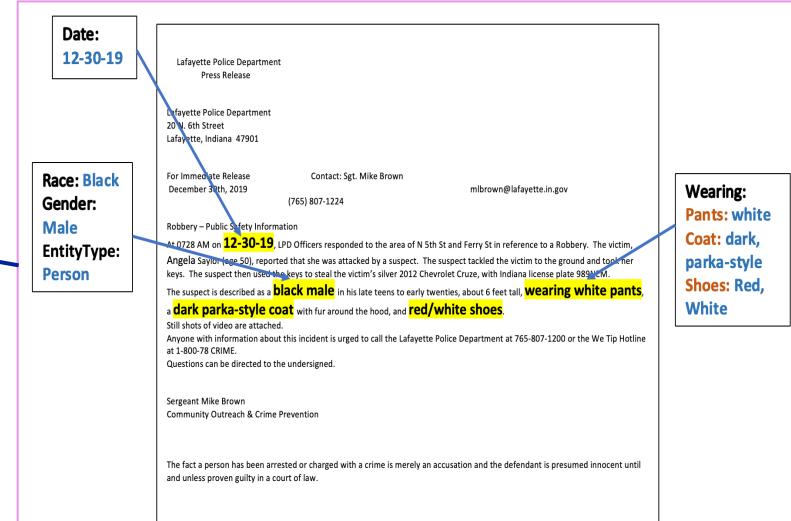
Features are extracted
from raw input

Relevance Matching with Graph Representation Learning

Streaming
Input or
Data-at-
rest

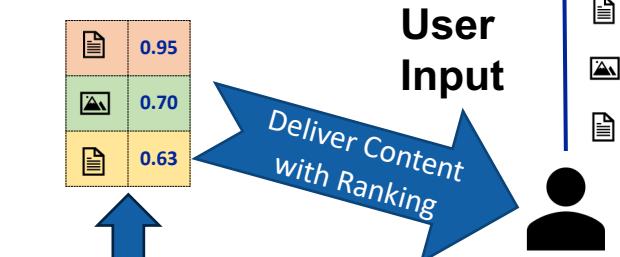


Extracted Features
 F_2, F_6, F_i



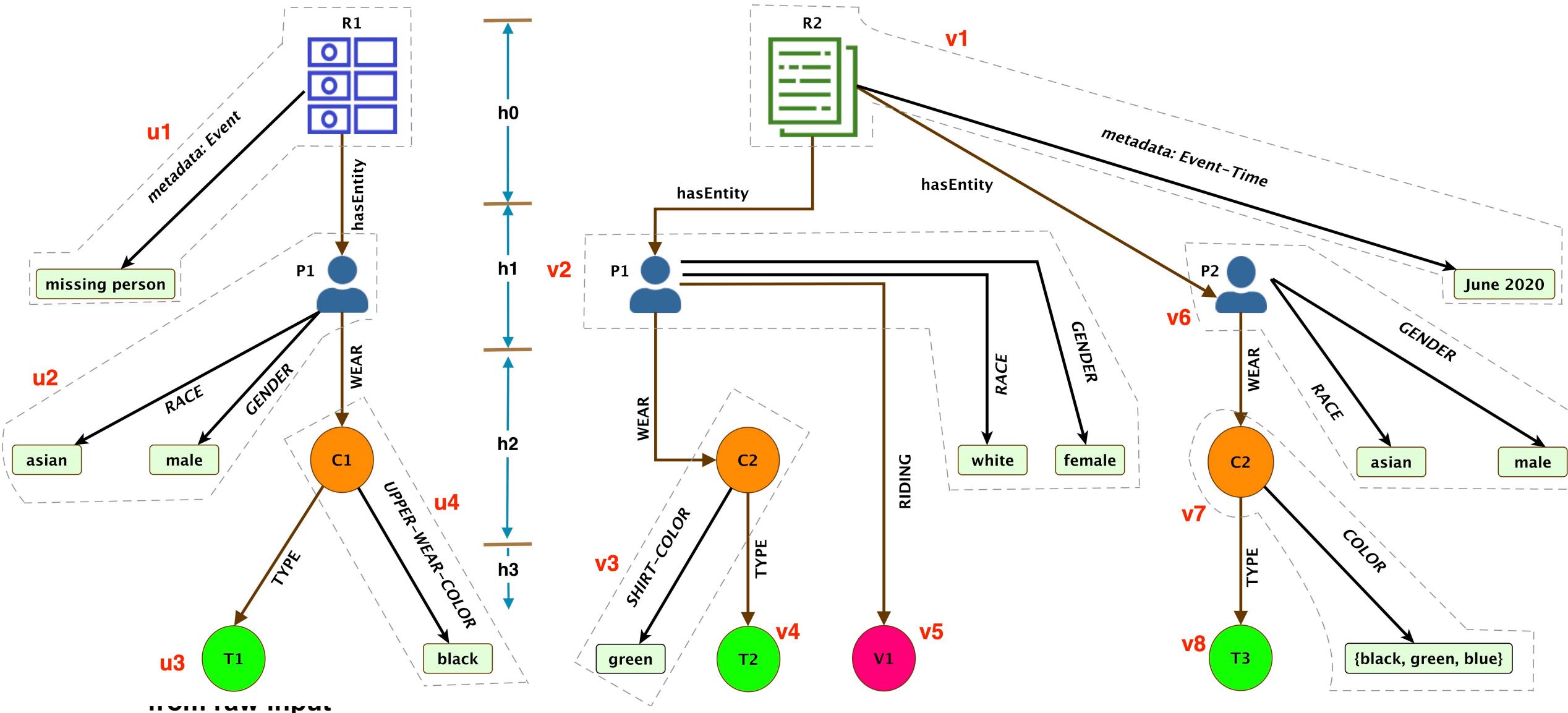
Relevance Matching

0.95
0.0
0.70
0.63
0.40

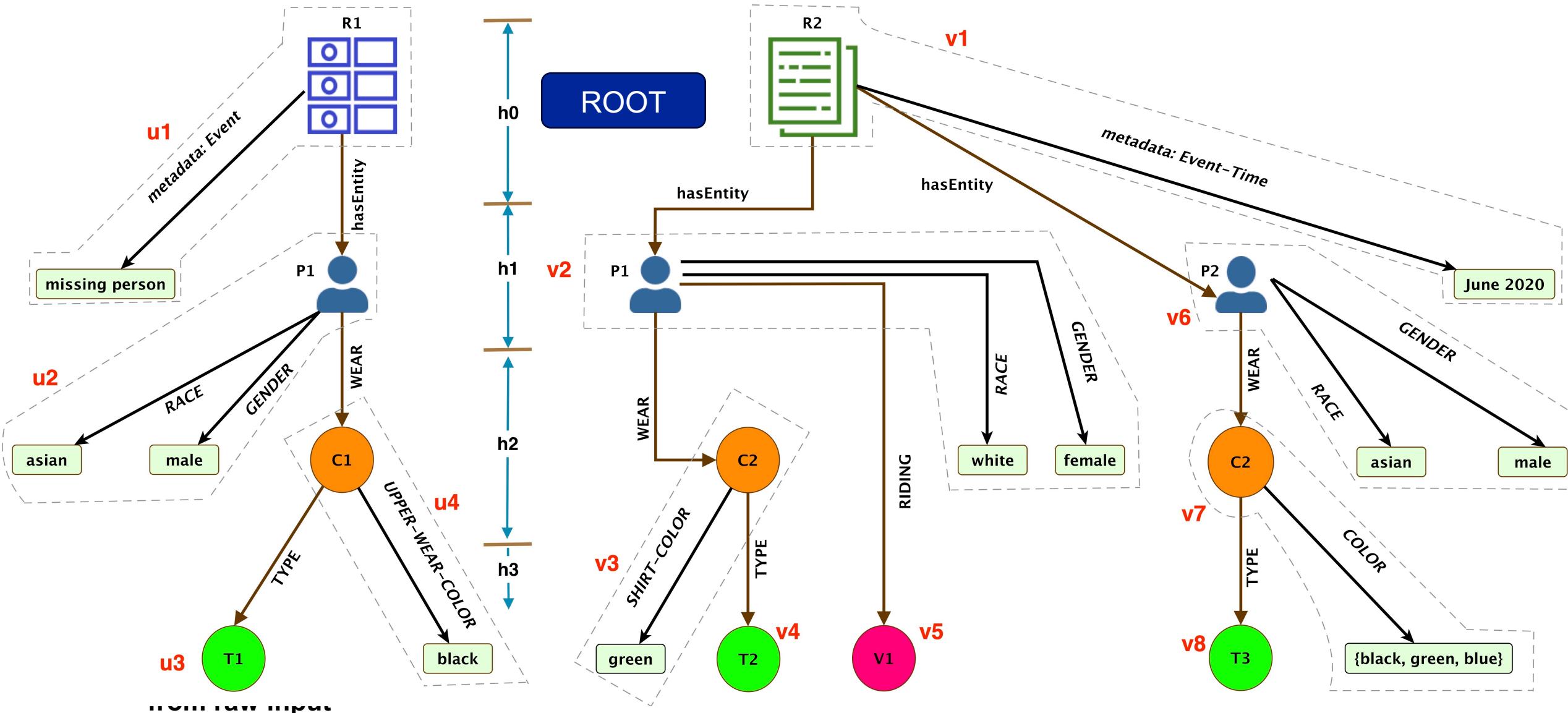


Features are extracted
from raw input

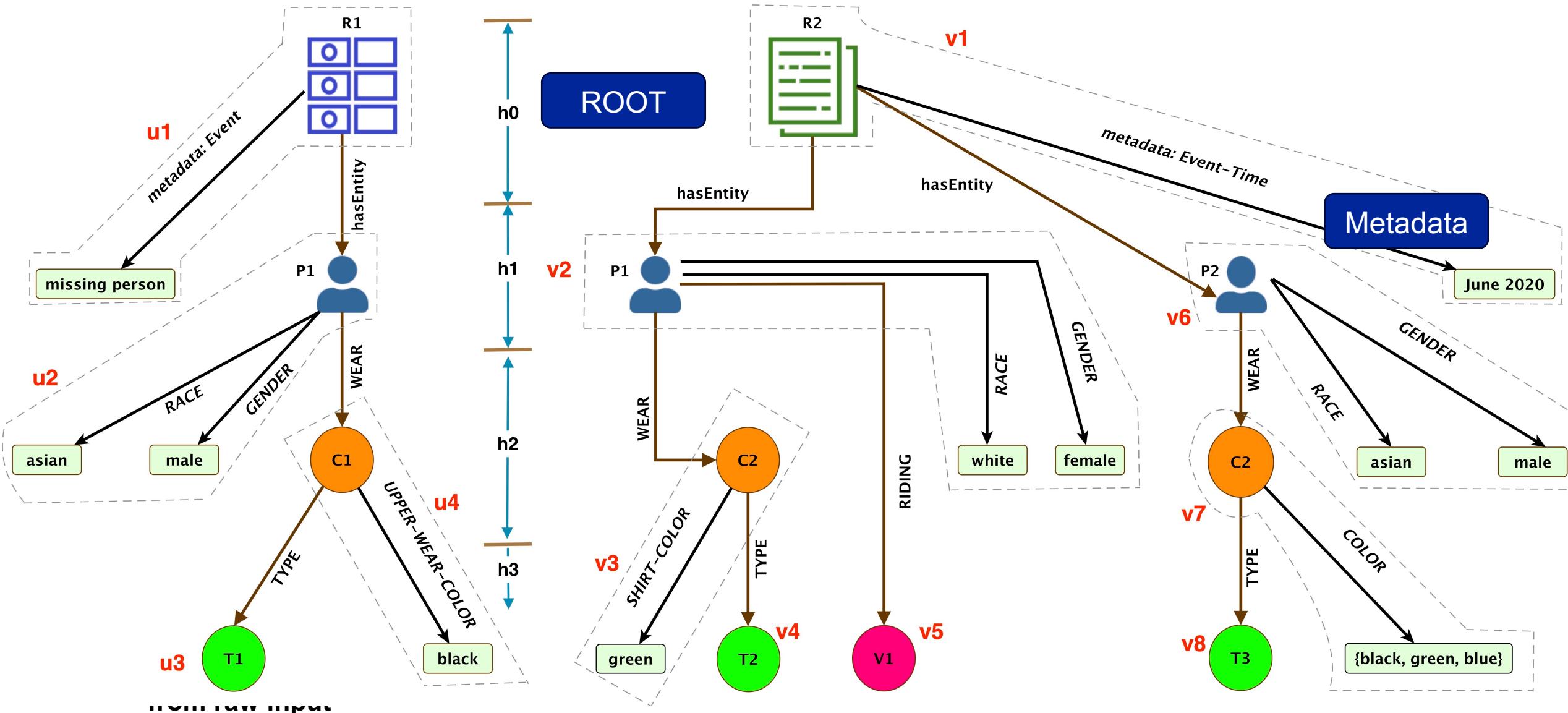
Relevance Matching with Graph Representation Learning



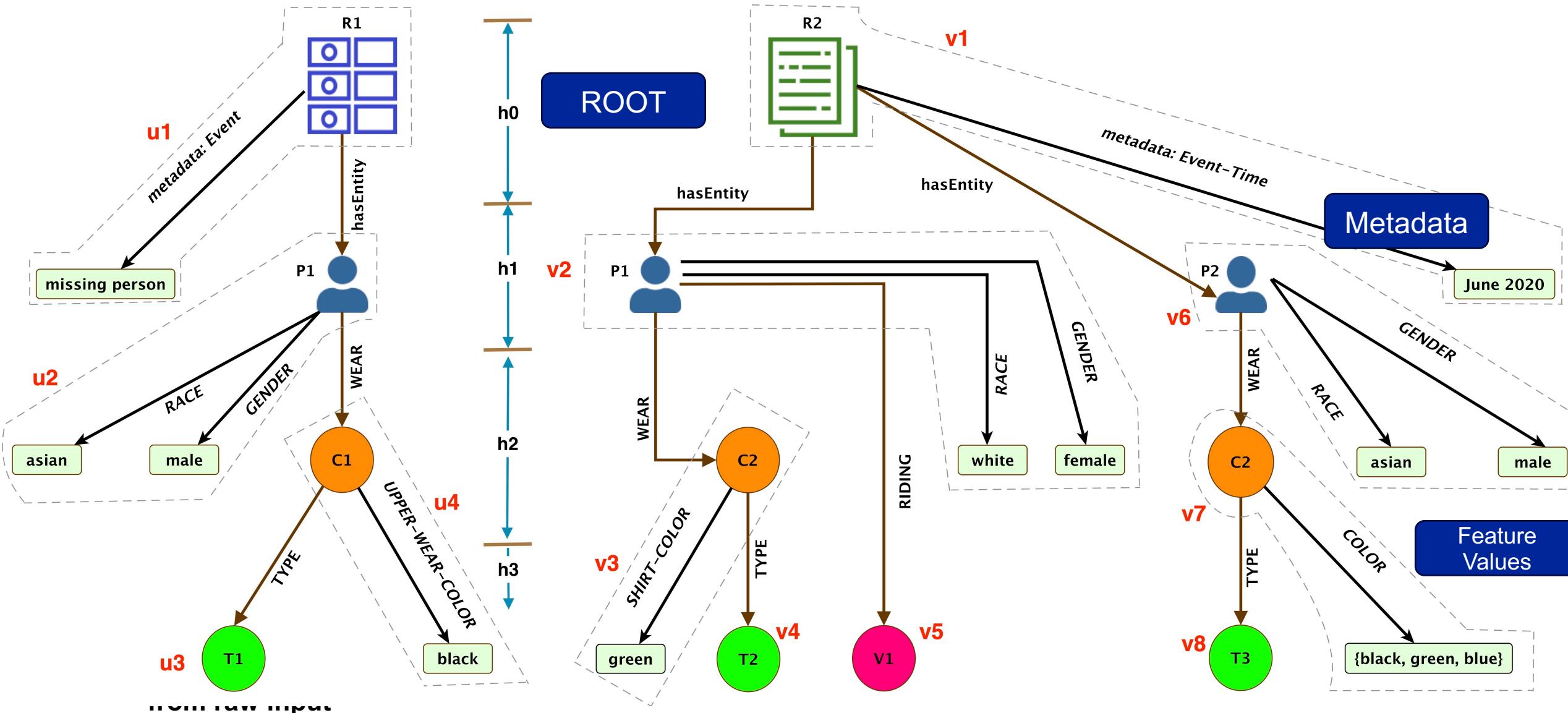
Relevance Matching with Graph Representation Learning



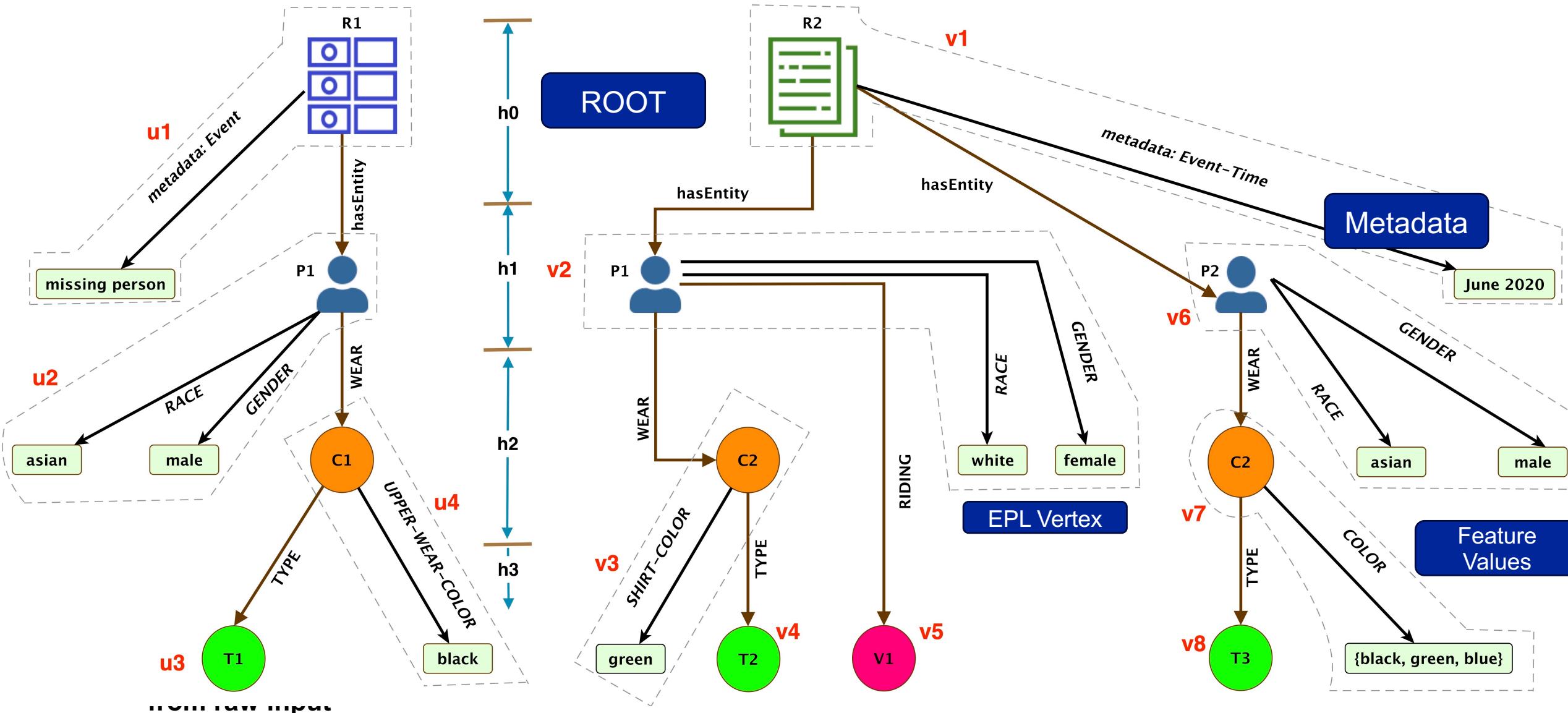
Relevance Matching with Graph Representation Learning



Relevance Matching with Graph Representation Learning

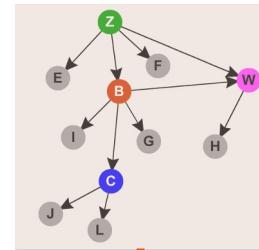
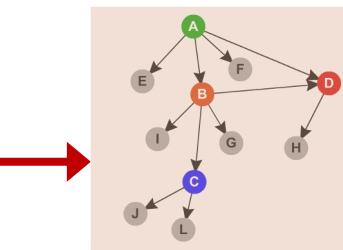
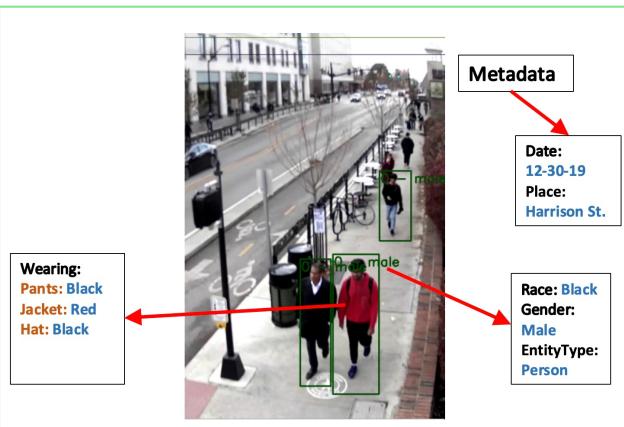


Relevance Matching with Graph Representation Learning

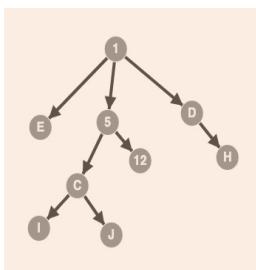
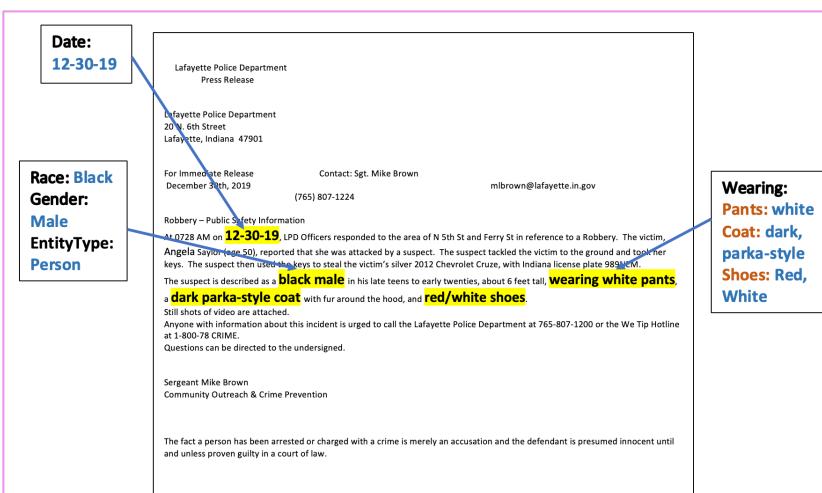
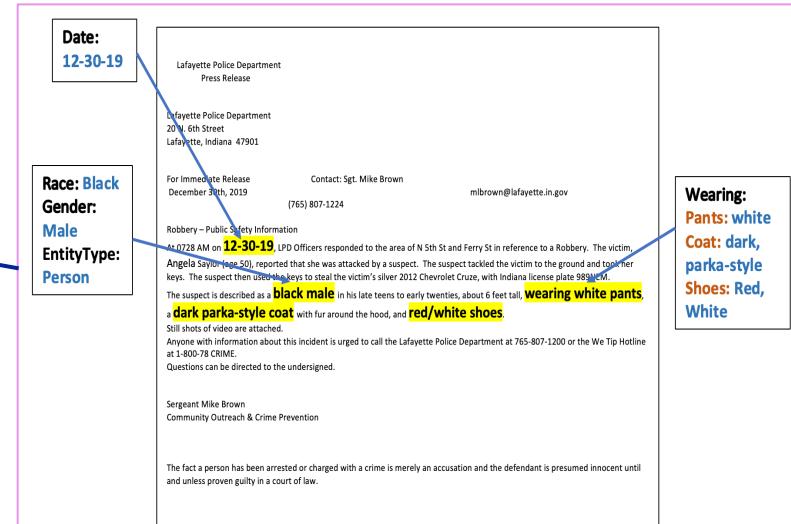


Relevance Matching with Graph Representation Learning

Streaming
Input or
Data-at-
rest



Extracted Features
 F_2, F_6, F_i



Relevance Matching

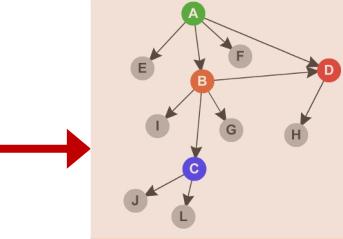
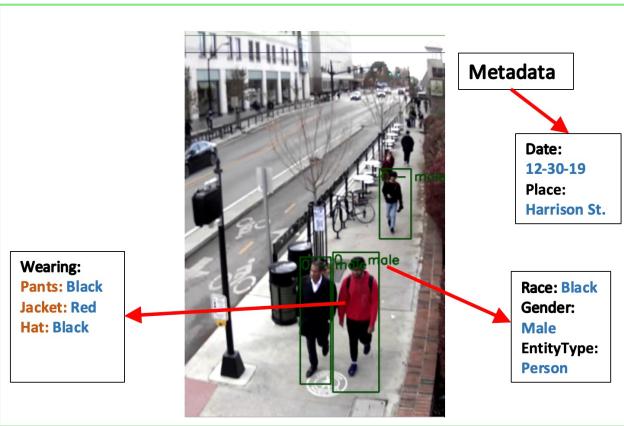
0.95
0.0
0.70
0.63
0.40



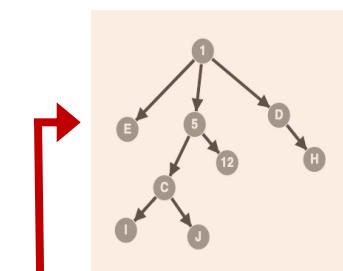
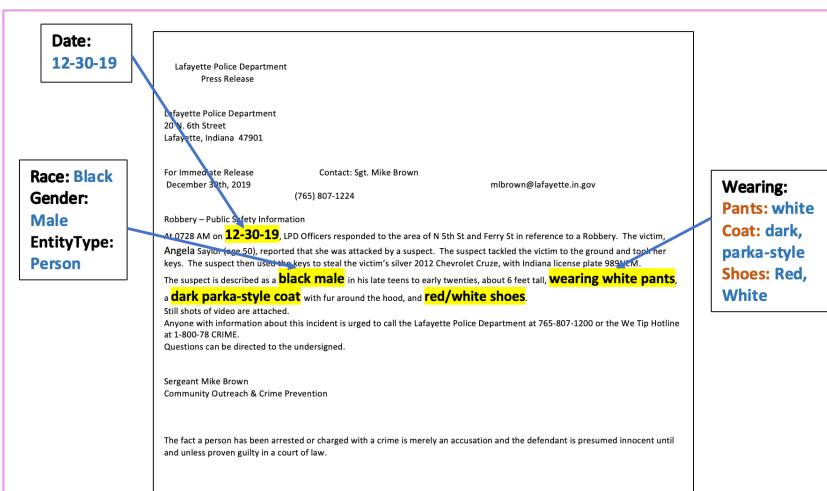
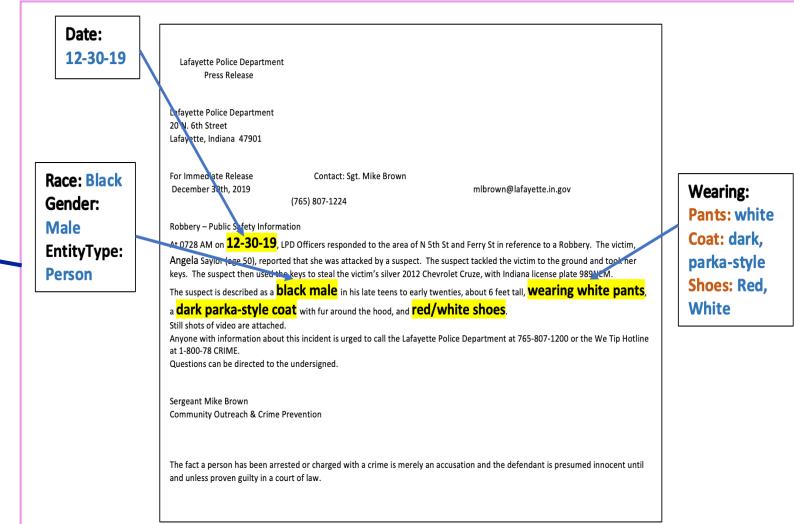
Features are extracted
from raw input

Relevance Matching with Graph Representation Learning

Streaming
Input or
Data-at-
rest

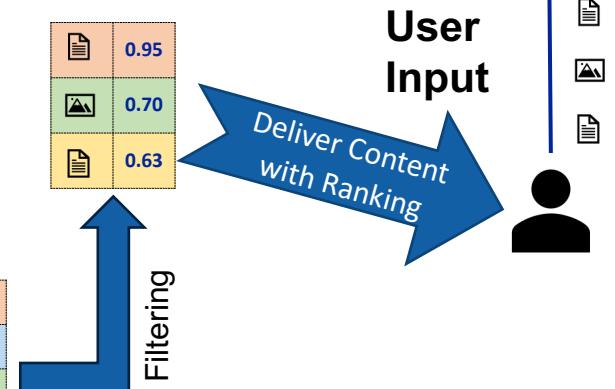


Extracted Features
 F_2, F_6, F_i



Graph Matching
and CED
Approximation

0.95
0.70
0.63
0.40
0.0
0.70
0.63



Features are extracted from raw input

Content Edit Distance (CED)

Different nodes and edges has different change cost

Feature replacement cost is an input

GED calculation algorithms (A*-search, VJ, or Beam) speed increase with #nodes

HARG is variable sized

Edge replacement cost considers both participating nodes

Multi-valued features need change cost, not replacement cost

Content Edit Distance (CED)

Different nodes and edges has different change cost

Feature replacement cost is an input

GED calculation algorithms (A*-search, VJ, or Beam) speed increase with #nodes

HARG is variable sized

Edge replacement cost considers both participating nodes

Multi-valued features need change cost, not replacement cost

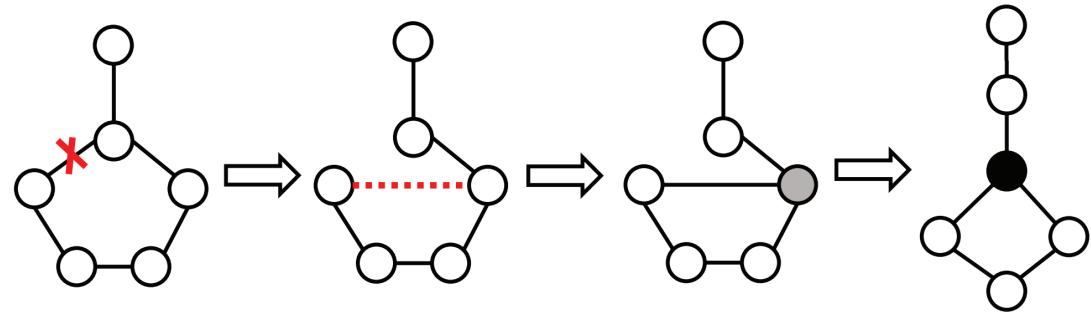


Figure 2: The GED between the graph to the left and the graph to the right is 3, as the transformation needs 3 edit operations: (1) an edge deletion, (2) an edge insertion, and (3) a node relabeling.

Content Edit Distance (CED)

Different nodes and edges has different change cost

Feature replacement cost is an input

GED calculation algorithms (A*-search, VJ, or Beam) speed increase with #nodes

HARG is variable sized

Edge replacement cost considers both participating nodes

Multi-valued features need change cost, not replacement cost

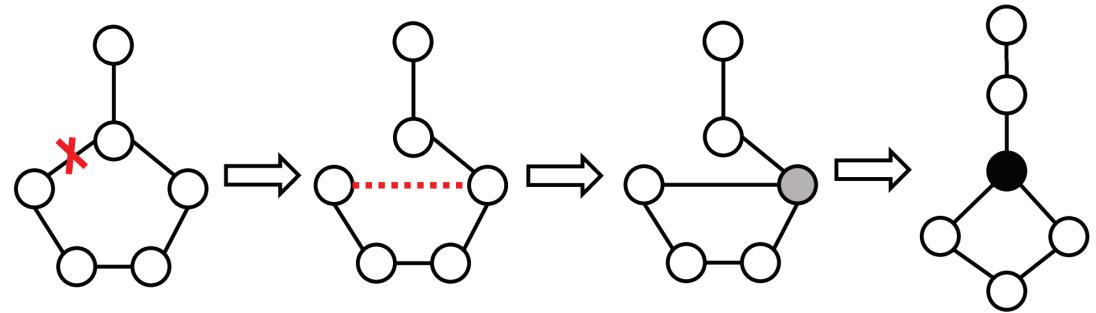
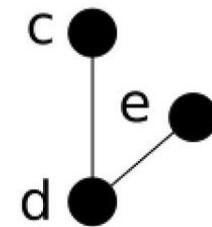


Figure 2: The GED between the graph to the left and the graph to the right is 3, as the transformation needs 3 edit operations: (1) an edge deletion, (2) an edge insertion, and (3) a node relabeling.

Graph A



Graph B



Cost/Edit Matrix

	assign	delete		
add	c	d	e	
a	0	2	4	2
b	2	0	4	∞
c	2	∞	∞	0
d	∞	3	∞	0
e	∞	∞	2	0

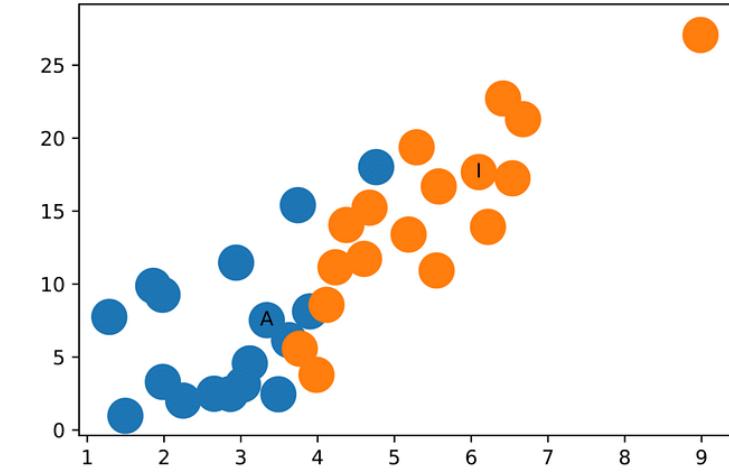
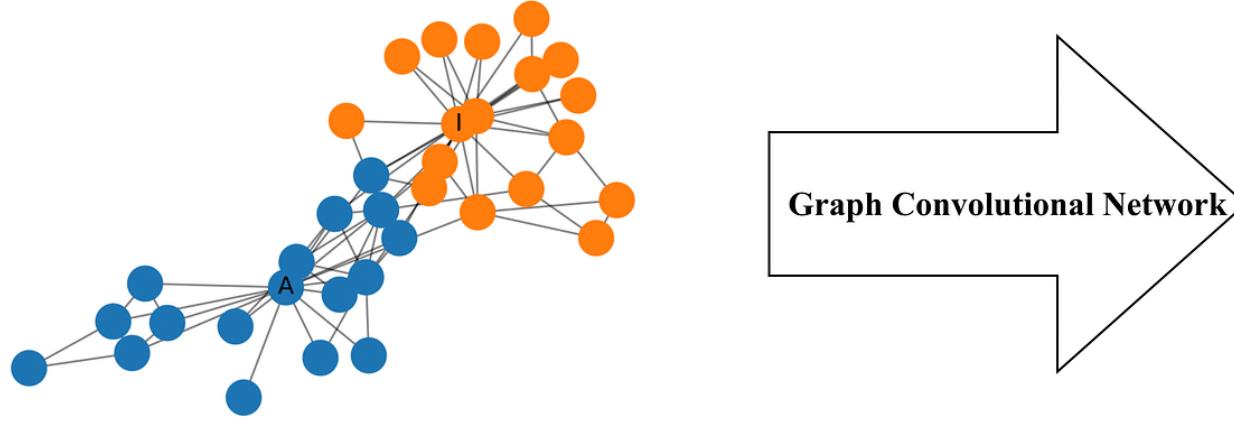
adapted from github.com/Jacobe2169/GMatch4py

Content Edit Distance (CED)

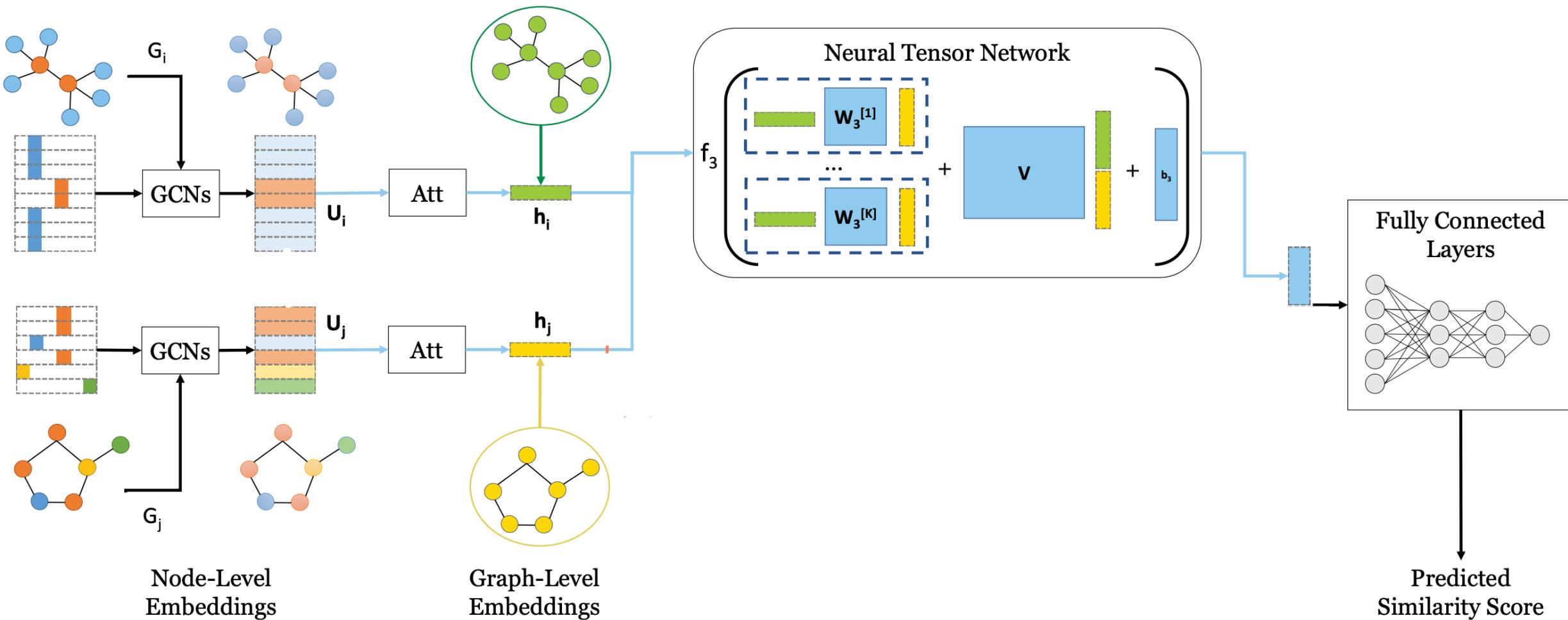
- **Cost-matrix** updates for replacement cost:
 - Different objects on same level: ∞
 - Single object cost:
 - # mismatched feature values of the object
 - Edge replacement cost:
 - **Wu-Palmer distance** between Synsets of two node values
 - **More dissimilar in meaning, more cost**
- **Cumulative Munkres**
 - Each data can contain multiple objects
 - Add that cumulative info to root
 - Add the dependency information into the cost-matrix
 - Levels from HARG
 - Parent EPL-vertices assignment cost are added to child EPL-vertices, starting from EPL-vertices in level-1.

CED Approximation as Function

- Munkres algorithm has a polynomial time complexity.
- We adopt an approximation model for GED
 - Graph Convolutional Network (GCN)
 - representation-invariant
 - Inductive
 - Neural Tensor Network



CED Approximation as Function



Evaluation Setup

- Ground truth
 - Ranked data in ascending order of penalties for mismatched features
 - Penalties for mismatched features:
 - $\text{rcost}(\text{top-color}) = 1$, $\text{rcost}(\text{bottom-color}) = 2$, and $\text{rcost}(\text{gender}) = 3$
 - Higher the penalty, higher the importance
- For Munkres, API from clapper*
- For model architecture,
 - *Hyperparameters*: initial node representation, GCN layer #, NTN layer # (K)
 - We conduct all the experiments on a single machine with Linux Ubuntu 18.04 and one Nvidia Titan GPU.

* <https://software.clapper.org/munkres/api/index.html>

Evaluation Setup

- For evaluation,
 - We consider data samples with $CED < 3$ in comparison to the query object, as relevant for that query. This would return contents where persons only with color mismatches are found.

Evaluation Setup

- For evaluation,
 - We consider data samples with $CED < 3$ in comparison to the query object, as relevant for that query. This would return contents where persons only with color mismatches are found.

Retrieval Task \equiv
Query one modality \rightarrow Get another or any modality

Evaluation Setup

- For evaluation,
 - We consider data samples with $CED < 3$ in comparison to the query object, as relevant for that query. This would return contents where persons only with color mismatches are found.

Retrieval Task \equiv

Query one modality \rightarrow Get another or any modality

$$mAP = \frac{1}{|classes|} \sum_{c \in classes} \frac{\#TP(c)}{\#TP(c) + \#FP(c)}$$

Experimental Results and Findings

RQ1,2: How is FemmIR comparable to EARS? Are the performance of our proposed data integration methods comparable to SOTA methods on different MMIR benchmark datasets?

Query	Target	EARS	FemmIR
Image	Text	0.54	0.40
	Image	0.27	0.27
	Video	0.33	0.29
	All	0.30	0.28
Text	Text	1.0	0.52
	Image	0.37	0.29
	Video	0.46	0.33
	All	0.43	0.31
Video	Text	0.62	0.43
	Image	0.30	0.29
	Video	0.37	0.30
	All	0.34	0.30
Avg		0.44	0.33

Table 4.1. Performance of EARS and FemmIR in mAP(%)

* Solaiman et al., *Feature-centric Multimodal Information Retrieval (FemmIR)*, submitted in SIGMOD 2023

** Hu, Peng, et al. "Scalable deep multimodal learning for cross-modal retrieval." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

Experimental Results and Findings

RQ1,2: How is FemmlR comparable to EARS? Are the performance of our proposed data integration methods comparable to SOTA methods on different MMIR benchmark datasets?

Query	Target	EARS	FemmIR
Image	Text	0.54	0.40
	Image	0.27	0.27
	Video	0.33	0.29
	All	0.30	0.28
Text	Text	1.0	0.52
	Image	0.37	0.29
	Video	0.46	0.33
	All	0.43	0.31
Video	Text	0.62	0.43
	Image	0.30	0.29
	Video	0.37	0.30
	All	0.34	0.30
Avg		0.44	0.33

Table 4.1. Performance of EARS and FemmlR in mAP(%). FemmlR is comparable to EARS, even with the current evaluation setup and no modifications on the training method. It could be improved with different values of K and #GCN layers.

* Solaiman et al., *Feature-centric Multimodal Information Retrieval (FemmlR)*, submitted in SIGMOD 2023

** Hu, Peng, et al. "Scalable deep multimodal learning for cross-modal retrieval." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

Experimental Results and Findings

RQ1,2: How is FemmIR comparable to EARS? Are the performance of our proposed data integration methods comparable to SOTA methods on different MMIR benchmark datasets?

Query	Target	EARS	FemmIR
Image	Text	0.54	0.40
	Image	0.27	0.27
	Video	0.33	0.29
	All	0.30	0.28
Text	Text	1.0	0.52
	Image	0.37	0.29
	Video	0.46	0.33
	All	0.43	0.31
Video	Text	0.62	0.43
	Image	0.30	0.29
	Video	0.37	0.30
	All	0.34	0.30
Avg		0.44	0.33

Table 4.1. Performance of EARS and FemmIR in mAP(%)

* Solaiman et al., *Feature-centric Multimodal Information Retrieval (FemmIR)*, submitted in SIGMOD 2023

** Hu, Peng, et al. "Scalable deep multimodal learning for cross-modal retrieval." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

Experimental Results and Findings

RQ1,2: How is FemmIR comparable to EARS? Are the performance of our proposed data integration methods comparable to SOTA methods on different MMIR benchmark datasets?

Query	Target	EARS	FemmIR
Image	Text	0.54	0.40
	Image	0.27	0.27
	Video	0.33	0.29
	All	0.30	0.28
Text	Text	1.0	0.52
	Image	0.37	0.29
	Video	0.46	0.33
	All	0.43	0.31
Video	Text	0.62	0.43
	Image	0.30	0.29
	Video	0.37	0.30
	All	0.34	0.30
Avg		0.44	0.33

Table 4.1. Performance of EARS and FemmIR in mAP(%)

Table 2: Performance comparison in terms of mAP scores on the PKU XMedia dataset.

Method	Query	Image				Text				Video				Avg.
		Text	Audio	3D	Video	Image	Audio	3D	Video	Image	Text	Audio	3D	
ml-CCA [31]*		0.597	0.241	0.284	0.377	0.613	0.242	0.243	0.355	0.374	0.307	0.166	0.252	0.276
JRL [41]*		0.770	0.296	0.521	0.376	0.788	0.279	0.477	0.348	0.399	0.272	0.172	0.366	0.379
GSS-SL [43]*		0.875	0.360	0.584	0.562	0.878	0.336	0.509	0.527	0.512	0.463	0.207	0.388	0.451
DCCA [2]*		0.869	0.264	0.186	0.463	0.871	0.306	0.221	0.406	0.433	0.369	0.167	0.163	0.312
DCCAE [40]*		0.868	0.278	0.195	0.492	0.878	0.288	0.244	0.450	0.442	0.427	0.185	0.194	0.325
CMPM+CMPC [44]*		0.897	0.544	0.637	0.641	0.896	0.500	0.669	0.675	0.583	0.626	0.385	0.497	0.584
ACMR [38]*		0.882	0.504	0.512	0.559	0.885	0.488	0.483	0.565	0.527	0.523	0.276	0.338	0.480
MCCA [33]		0.128	0.186	0.221	0.140	0.133	0.174	0.177	0.128	0.101	0.079	0.128	0.164	0.148
GMLDA [35]		0.608	0.186	0.513	0.414	0.629	0.170	0.470	0.332	0.368	0.282	0.121	0.329	0.334
MvDA-VC [12]		0.630	0.290	0.550	0.488	0.643	0.264	0.491	0.411	0.435	0.343	0.152	0.353	0.378
SDML		0.899	0.552	0.690	0.659	0.917	0.572	0.722	0.686	0.587	0.604	0.342	0.514	0.609

*These methods are two-modality methods.

* Solaiman et al., *Feature-centric Multimodal Information Retrieval (FemmIR)*, submitted in SIGMOD 2023

** Hu, Peng, et al. "Scalable deep multimodal learning for cross-modal retrieval." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

Experimental Results and Findings

RQ1,2: How is FemmIR comparable to EARS? Are the performance of our proposed data integration methods comparable to SOTA methods on different MMIR benchmark datasets?

Query	Target	EARS	FemmIR
Image	Text	0.54	0.40
	Image	0.27	0.27
	Video	0.33	0.29
	All	0.30	0.28
Text	Text	1.0	0.52
	Image	0.37	0.29
	Video	0.46	0.33
	All	0.43	0.31
Video	Text	0.62	0.43
	Image	0.30	0.29
	Video	0.37	0.30
	All	0.34	0.30
Avg		0.44	0.33

Table 4.1. Performance of EARS and FemmIR in mAP(%)

Table 3: Performance comparison in terms of mAP scores on the Wikipedia dataset.

Method	Image → Text	Text → Image	Average
MCCA [33]	0.202	0.189	0.195
ml-CCA [31]	0.388	0.356	0.372
GMLDA [35]	0.238	0.240	0.239
JRL [41]	0.343	0.376	0.330
MvDA-VC [12]	0.397	0.345	0.387
GSS-SL [43]	0.466	0.413	0.440
DCCA [2]	0.301	0.286	0.294
DCCAE [40]	0.308	0.290	0.299
ACMR [38]	0.479	0.426	0.452
CMPM+CMPC [44]	0.493	0.438	0.466
CCL [27]	0.504	0.457	0.481
CBT [30]	0.516	0.464	0.490
SDML	0.522	0.488	0.505

* Solaiman et al., *Feature-centric Multimodal Information Retrieval (FemmIR)*, submitted in SIGMOD 2023

** Hu, Peng, et al. "Scalable deep multimodal learning for cross-modal retrieval." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

Experimental Results and Findings

RQ1,2: How is FemmlR comparable to EARS? Are the performance of our proposed data integration methods comparable to SOTA methods on different MMIR benchmark datasets?

Query	Target	EARS	FemmIR
Image	Text	0.54	0.40
	Image	0.27	0.27
	Video	0.33	0.29
	All	0.30	0.28
Text	Text	1.0	0.52
	Image	0.37	0.29
	Video	0.46	0.33
	All	0.43	0.31
Video	Text	0.62	0.43
	Image	0.30	0.29
	Video	0.37	0.30
	All	0.34	0.30
Avg		0.44	0.33

Table 4.1. Performance of EARS and FemmlR in mAP(%)

Findings: If we have a feature extractor with a good performance, such as the text feature extractor, EARS and FemmlR can perform similarly or better to other retrieval models.

* Solaiman et al., *Feature-centric Multimodal Information Retrieval (FemmlR)*, submitted in SIGMOD 2023

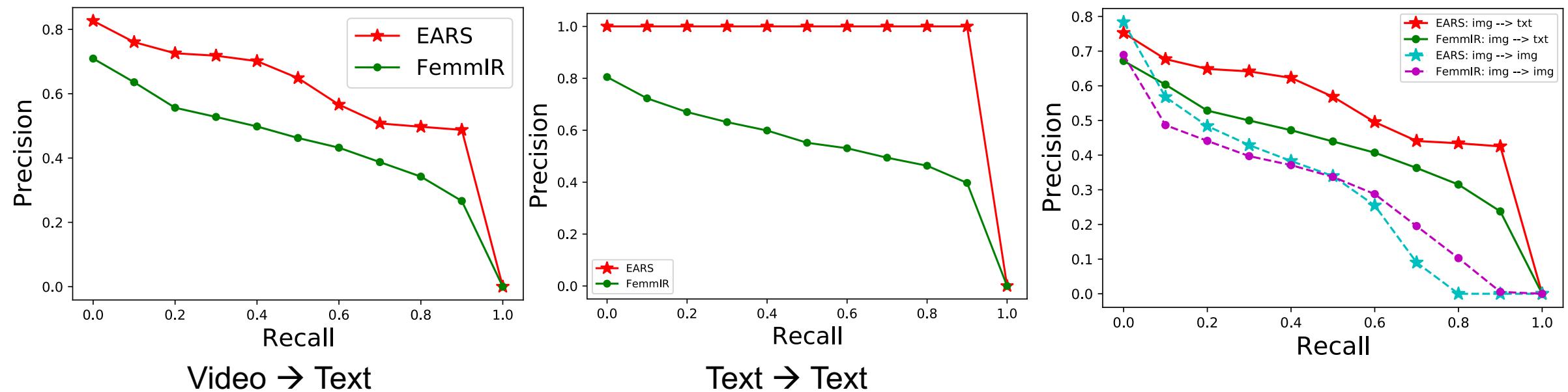
** Hu, Peng, et al. "Scalable deep multimodal learning for cross-modal retrieval." Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval. 2019.

Table 3: Performance comparison in terms of mAP scores on the Wikipedia dataset.

Method	Image → Text	Text → Image	Average
MCCA [33]	0.202	0.189	0.195
ml-CCA [31]	0.388	0.356	0.372
GMLDA [35]	0.238	0.240	0.239
JRL [41]	0.343	0.376	0.330
MvDA-VC [12]	0.397	0.345	0.387
GSS-SL [43]	0.466	0.413	0.440
DCCA [2]	0.301	0.286	0.294
DCCAE [40]	0.308	0.290	0.299
ACMR [38]	0.479	0.426	0.452
CMPM+CMPC [44]	0.493	0.438	0.466
CCL [27]	0.504	0.457	0.481
CBT [30]	0.516	0.464	0.490
SDML	0.522	0.488	0.505

Experimental Results and Findings

- RQ3: Can FemmIR approximate CED closely to an exact method like EARS?



- **Findings:** As the precision-recall curve shows, the results are consistent with our mAP scores. For text modality EARS has near perfect result. For other modalities, FemmIR is performing closely to EARS.

Research Contributions

- Novel **multi-modal information retrieval** approach to retrieve **ranked result** to information need expressed as Query-by-Example and Query-by-Properties.
- Novel **edit distance metric, CED**, to measure the amount of difference between two data samples based on their semantic features.
- **Graph coordinated representation learning** approach leverages a neural-network based graph-matching technique to capture the interactions between the query example and the streaming data features, with a weak supervision from CED.
- Two novel MMIR datasets using MARS, InciText, and PCAM
- Prototype and evaluation

Research Contributions

- Novel **multi-modal information retrieval** approach to retrieve **ranked result** to information need expressed as Query-by-Example and Query-by-Properties.
- Novel **edit distance metric, CED**, to measure the amount of difference between two data samples based on their semantic features.
- **Graph coordinated representation learning** approach leverages a neural-network based graph-matching technique to capture the interactions between the query example and the streaming data features, with a weak supervision from CED.
- Two novel MMIR datasets using MARS, InciText, and PCAM
- Prototype and evaluation

Impact: FemmIR can be used for **data discovery, data fusion**, and Explainable **multimodal data retrieval**

Contribution #3: Intrinsic Domain Complexity Estimation

**Distributed AI Systems in
Open-World Perception Domain**

Problem Statement and Motivation

- Complexities of well-known games
- Intrinsic domain complexity estimation in planning domains
- Distributed Learning Environment [1, 2]
- Novelty Measurement, Characterization, Adaptation and Detection
[Alspector21, Boult22ab, Doctor22, Jafar20, Kim20, Langley20,
Peng21, Molineux21]

1. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)
2. Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B., 2020. Adaptive federated optimization.

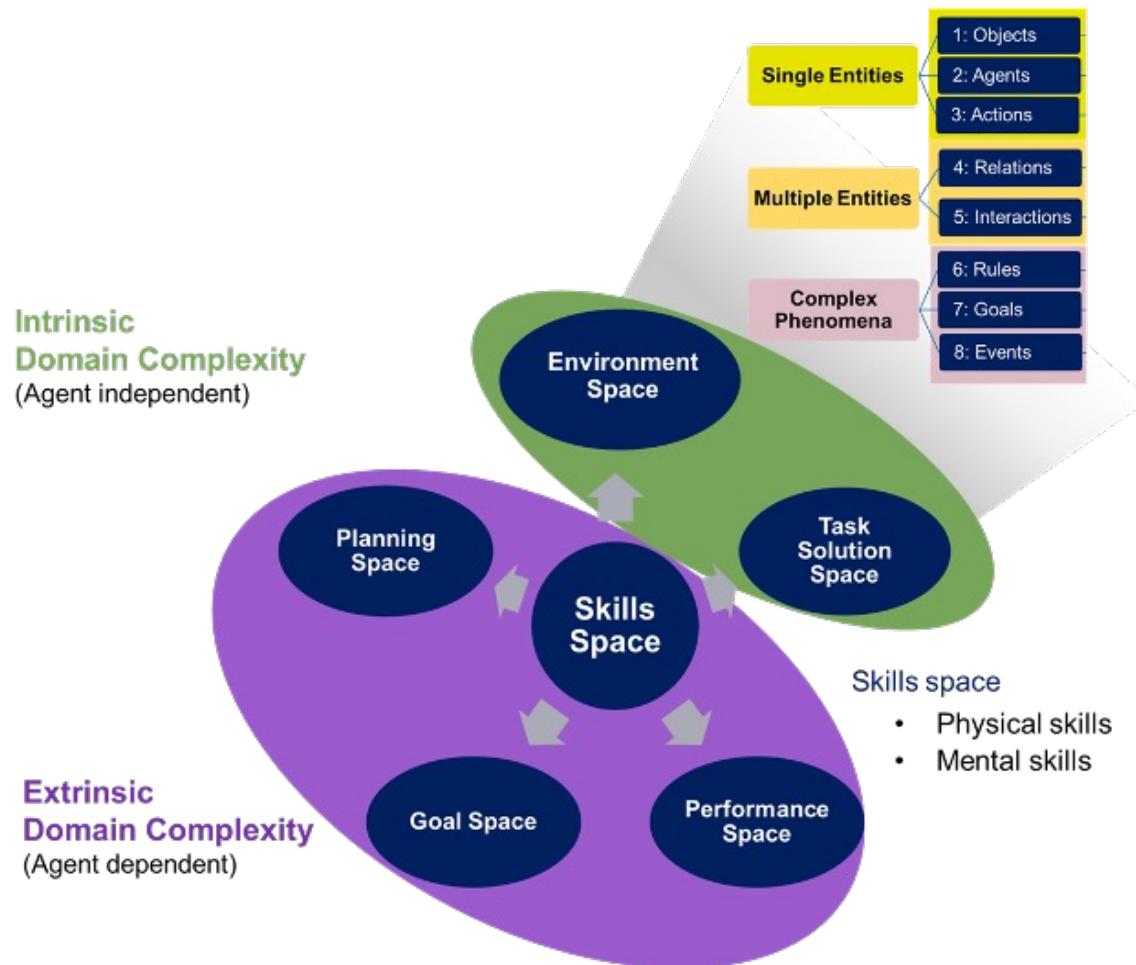
Problem Statement and Motivation

- Complexities of well-known games
- Intrinsic domain complexity estimation in planning domains
- Distributed Learning Environment [1, 2]
- Novelty Measurement, Characterization, Adaptation and Detection
[Alspector21, Boult22ab, Doctor22, Jafar20, Kim20, Langley20, Peng21, Molineux21]

Game	Board size (positions)	State-space complexity (as log to base 10)	Game-tree complexity (as log to base 10)	Average game length (plies)	Branching factor
Tic-tac-toe	9	3	5	9	4
Sim	15	3	8	14	3.7
Pentominoes	64	12	18	10	75

1. McMahan, H.B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A., 2017. Communication-efficient learning of deep networks from decentralized data, in: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)
2. Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., McMahan, H.B., 2020. Adaptive federated optimization.

Problem Statement and Motivation



Intrinsic Perception Domain Complexity



Dimensionality

- 1) Environment complexity (EC)
 - Dataset size * Reduced Dimension Size + # Labels
- 2) Intrinsic dimensionality (ID)



Sparsity

- 1) Environment representation
 - as few components as possible
 - as much information as possible
- 2) Principal Component Analysis (PCA)



Heterogeneity

- Diversity in Dataset
- Entropy
- Shannon Entropy

1. Batty, M., Morphet, R., Masucci, P., Stanilov, K., 2014. Entropy, complexity, and spatial information. *Journal of Geographical Systems* 16, 363–385. URL: <https://doi.org/10.1007/s10109-014-0202-2>. doi:10.1007/s10109-014-0202-2.
2. Tenenbaum, J.B., Silva, V.d., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290, 2319–2323.
3. Wurst, J., Fernández, A.F., Botsch, M., Utschick, W., 2020. An entropy based outlier score and its application to novelty detection for road infrastructure images. *CoRR* abs/2005.13288
4. Mairal, J., Bach, F., Ponce, J., 2014. Sparse modeling for image and vision processing.

Federated Learning Complexity

- FL training environment as Tree-based solution
 - Each path has own complexity
 - Complexity of Federated Learning tree, $F(d, X)$
 - $F(d)$ relies on # distinct entities, d
 - $F(X)$ relies on # participating entities, n
 - Intuition
 - As #participating-entities increases, intrinsic complexity increases, and so the effect on $F(d, X)$
 - More distinct entities in different servers, more complex
 - Federated master selects from multiple different paths in the learning tree based on the availability of the entities.
 - Objective of federated master
- $$\min F_{\pi}(d, X); \quad \forall \pi \in P(G)$$
-
- Communication Rounds ↑
- W,X,Y,Z: Federated Clients
A,B,C,D: Task Accuracy

Evaluation Setup

Datasets

- MNIST handwritten digit [202], Fashion-MNIST [203] and EMNIST-digits [204]
- 70K images, with 60K train, 10K test, 28 X 28 pixels

Windows Intel core i7-8th generation with 16 GB of memory

Probenet as Federated Complexity Benchmark (Shallownet)

- 5 distinct federated clients in the learning environment, where each client contains non-identical local data.
- Identical **shallow ProbeNet** model for each client and the federated server relies on the **FedAvg algorithm** for each global update.
- Effort* represents the number of communication rounds in the federated learning context.
- Local iteration for each client is 1, and each client contains similar amount of data, allowing us to ignore possible data amount disparity.
- For each experiments, we run 100 communication rounds.

Experiment Results and Findings

- RQ1: Is there a complexity order for the MNIST datasets in terms of the proposed metrices in singular environment?

Table 7.1. Heterogeneity, Sparsity, Environment Complexity, and Intrinsic Dimensionality Measurement.

Dataset	Heterogeneity	Sparsity ($r^2 = 80\%$)	Sparsity ($r^2 = 95\%$)	$EC_{upper}(v_\theta = 0)$	$EC_{upper}(v_\theta = 90)$	ID
Handwritten-MNIST	1.60	740	629	717	530	13.368
EMNIST-digits	2.86	751	685	697	557	14.095
Fashion-MNIST	4.11	760	594	784	745	14.547

Experiment Results and Findings

- RQ1: Is there a complexity order for the MNIST datasets in terms of the proposed metrices in singular environment?

Table 7.1. Heterogeneity, Sparsity, Environment Complexity, and Intrinsic Dimensionality Measurement.

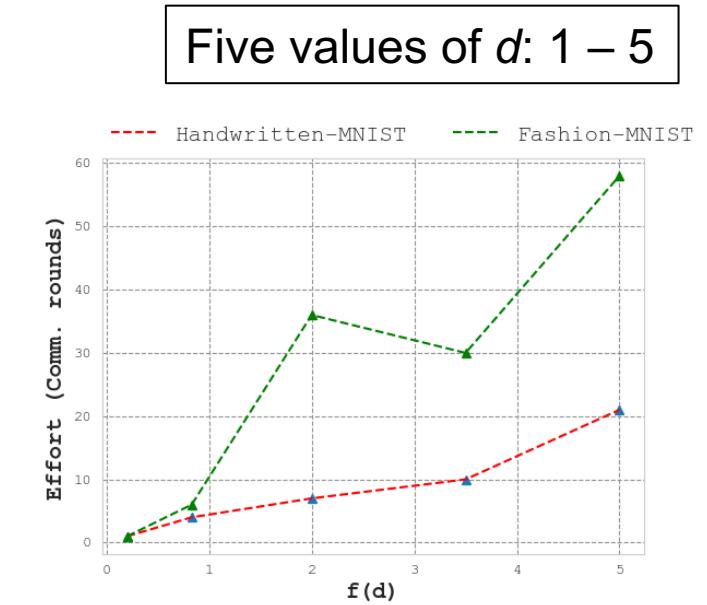
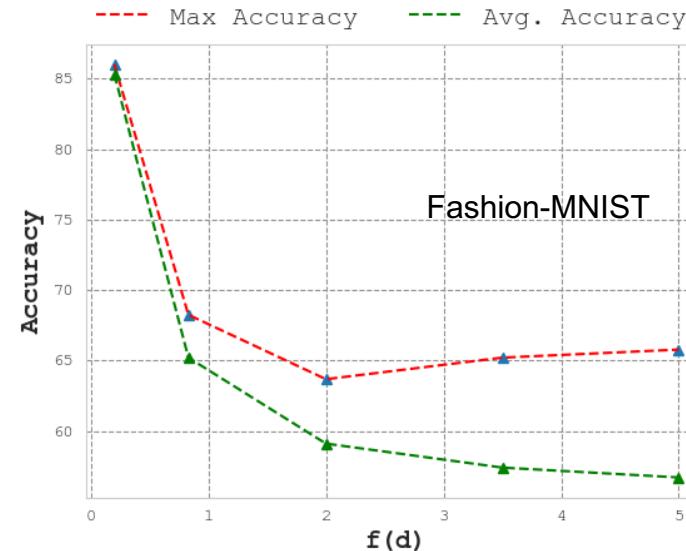
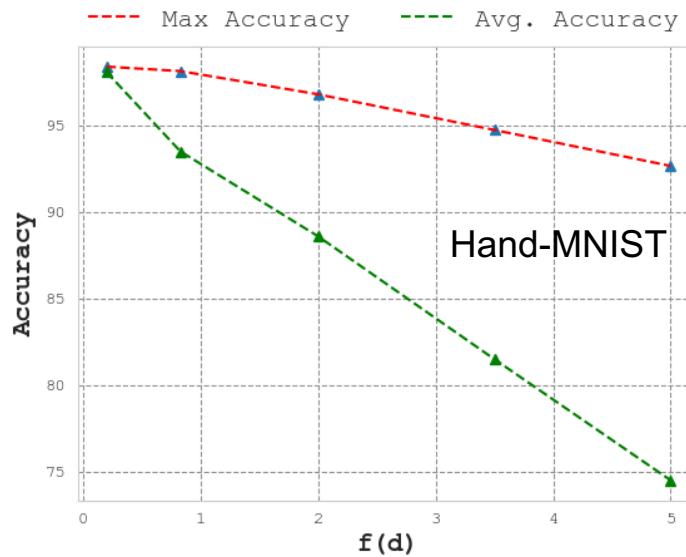
Dataset	Heterogeneity	Sparsity ($r^2 = 80\%$)	Sparsity ($r^2 = 95\%$)	$EC_{upper}(v_\theta = 0)$	$EC_{upper}(v_\theta = 90)$	ID
Handwritten-MNIST	1.60	740	629	717	530	13.368
EMNIST-digits	2.86	751	685	697	557	14.095
Fashion-MNIST	4.11	760	594	784	745	14.547

Findings:

1. Handwritten- MNIST is the least complex and Fashion-MNIST is the most, seen by their variation, and EMNIST-digits is between both.
2. Sparsity at $r^2 = 95\%$ implies that Fashion-MNIST requires a lot of sparse components for explaining variances in between 80% and 95%.
3. At variance threshold = 0, Fashion-MNIST doesn't include many zeros over the data set, whereas if we consider pixel value 90 as threshold, it still has the largest environment complexity.

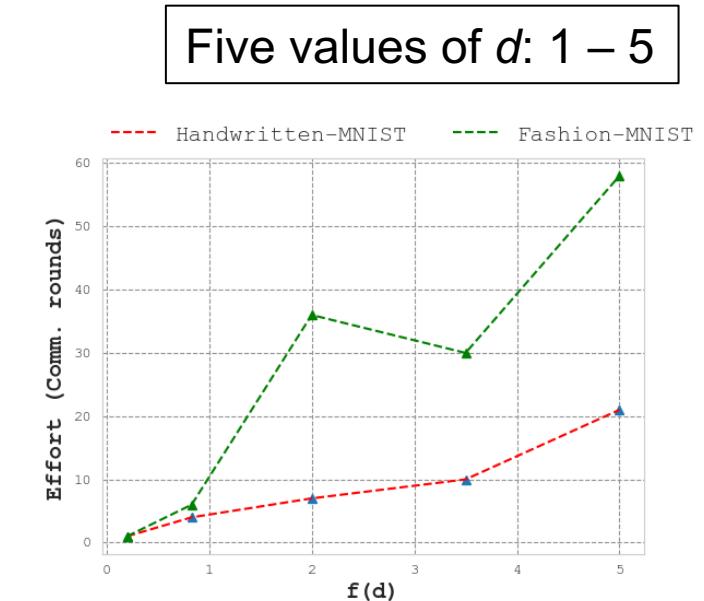
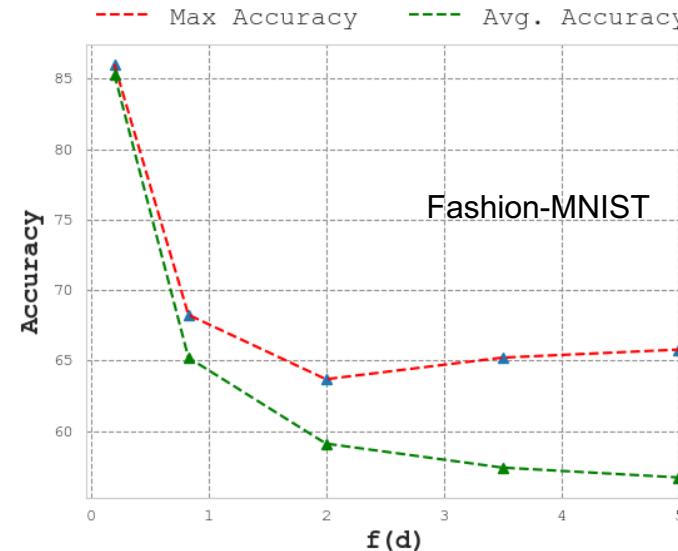
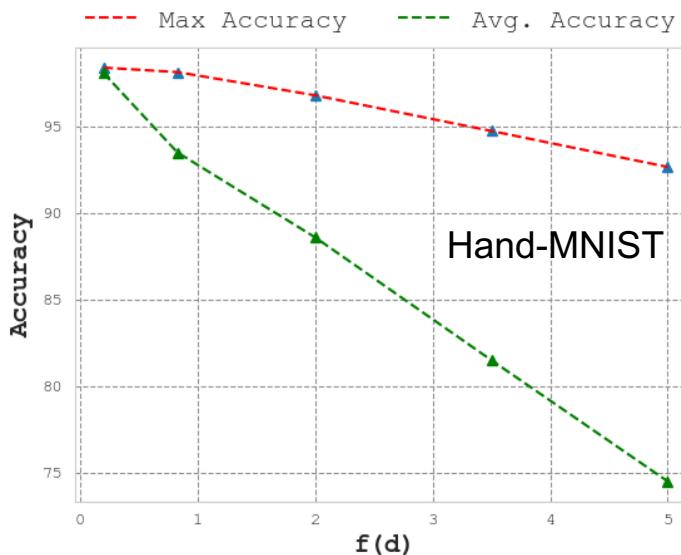
Experiment Results and Findings

- RQ2,3: How does Federated Environment Complexity $f(d)$ effects the overall distributed learning complexity?



Experiment Results and Findings

- RQ2,3: How does Federated Environment Complexity $f(d)$ effects the overall distributed learning complexity?



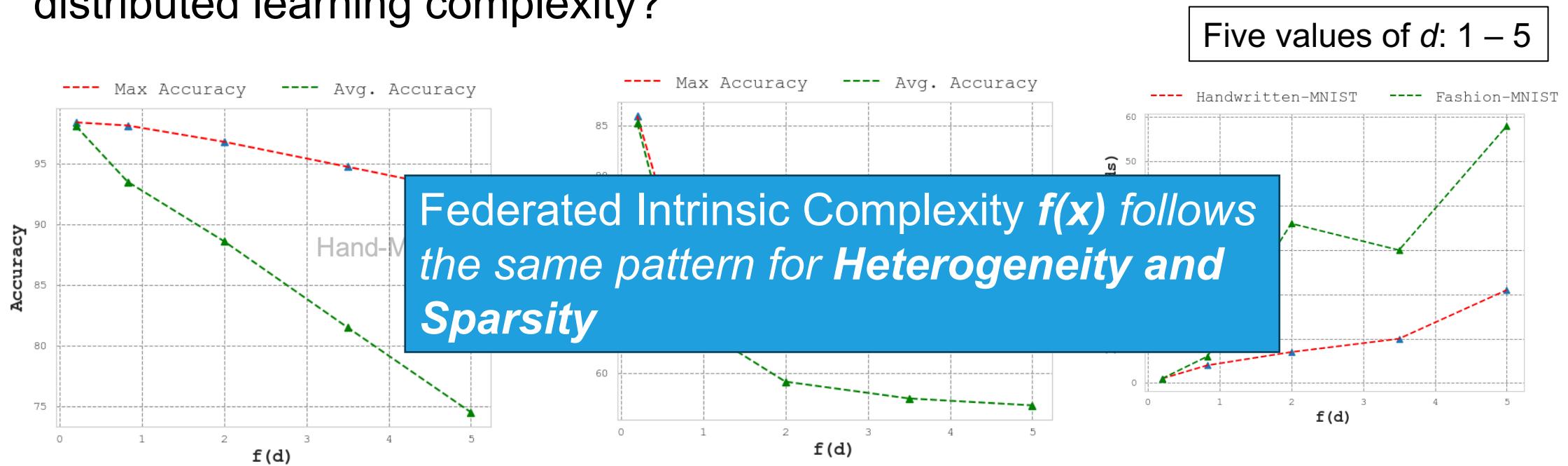
Five values of d : 1 – 5

Findings:

- For easier handwritten-MNIST, accuracy is higher, and effort is lower than fashion-MNIST.
- As the value of $f(d)$ increases, accuracy decreases, in line with the benchmark classifier.
- As $f(d)$ increases, the effort increases, while the intrinsic features (at 60% var. thres.) remains identical.

Experiment Results and Findings

- RQ2,3: How does Federated Environment Complexity $f(d)$ effects the overall distributed learning complexity?

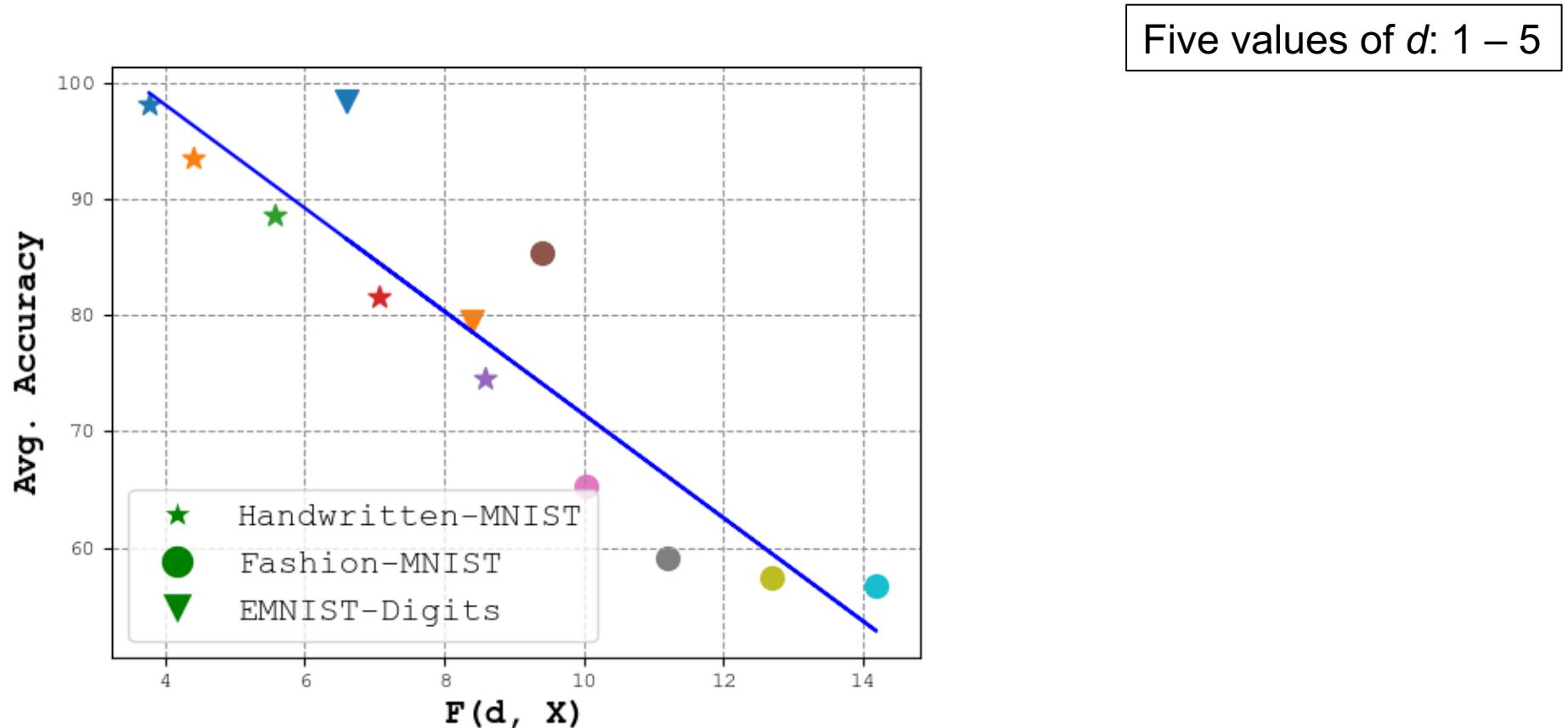


Findings:

- For easier handwritten-MNIST, accuracy is higher, and effort is lower than fashion-MNIST.
- As the value of $f(d)$ increases, accuracy decreases, in line with the benchmark classifier.
- As $f(d)$ increases, the effort increases, while the intrinsic features (at 60% var. thres.) remains identical.

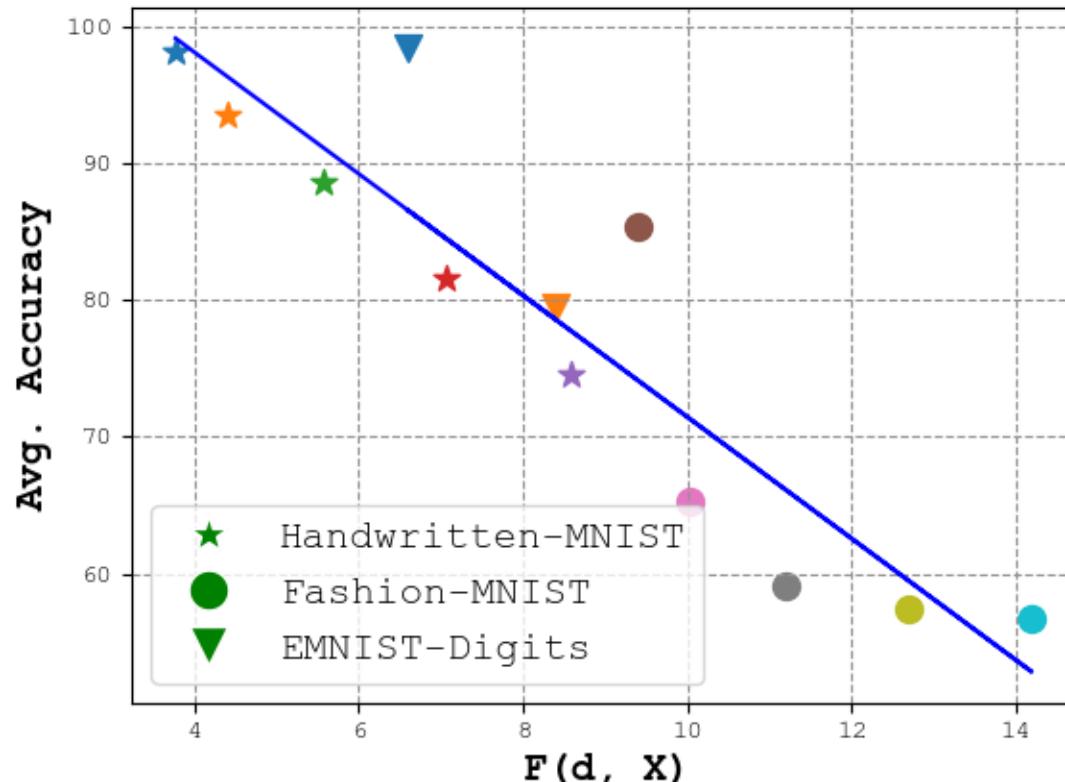
Experiment Results and Findings

- RQ4: Is Federated Complexity Metric correlated to Accuracy?



Experiment Results and Findings

- RQ4: Is Federated Complexity Metric correlated to Accuracy?



Five values of $d: 1 – 5$

Findings:

1. Federated c
2. We can see
3. So, $F(d,X)$ can be considered as a standard metric for evaluating federated learning complexity.

increase the accuracy
federated learning complexity.

Research Contributions

- Novel framework to **measure the inherent complexity of the perception domain** – with upper and lower limits, including non-linearity
- Novel **complexity measurement metric for distributed federated environment** in perception domain
- Extensive experiments on **MNIST, Fashion-MNIST, and EMNIST-digits** in distinct distributed settings and performed ablation study to measure the impact of each components of our proposed metric on the distributed domain complexity

Contribution #4: Novelties in Multimodal Information Retrieval

**Weakly Supervised Joint Embedding for Multimodal
Information Retrieval**

Problem Statement and Background

- How does novelties happen for MMIR?
- How does novelties affect multimodal data?
- Data Integration with Weak Supervision

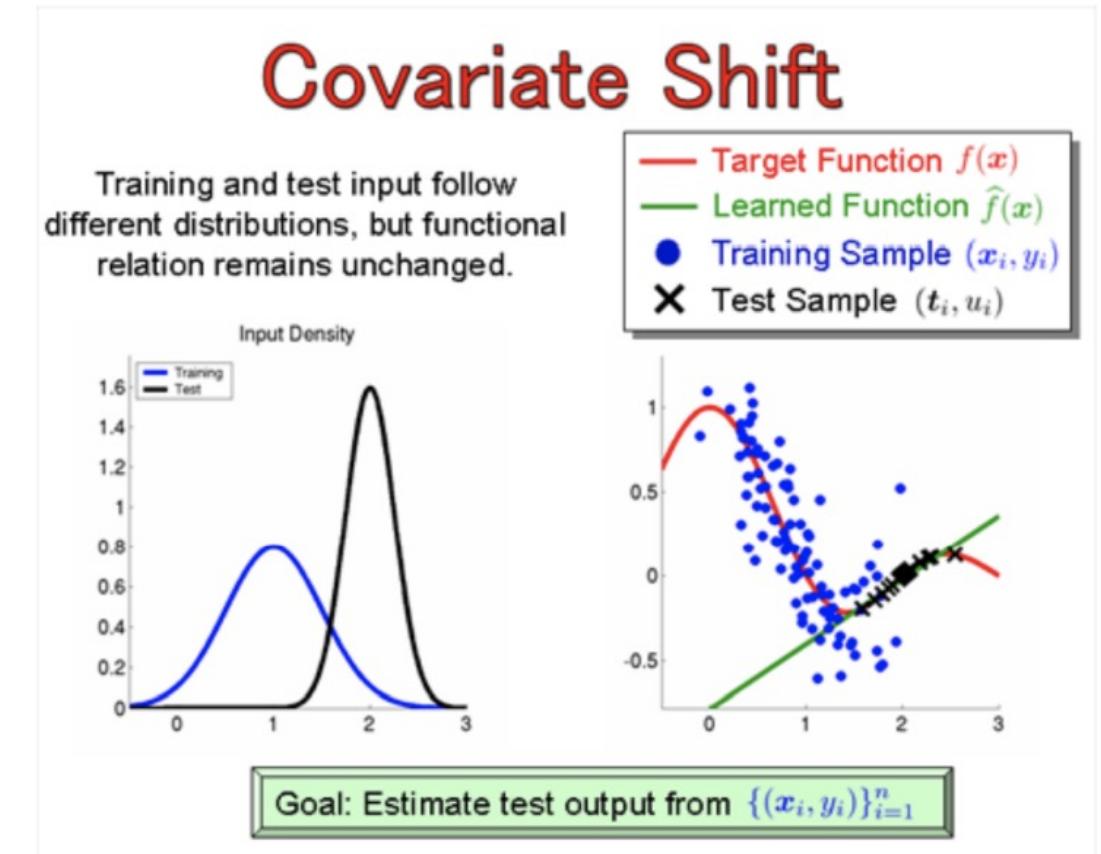
Problem Statement and Background

- How does novelties happen for MMIR?
- How does novelties affect multimodal data?
- Data Integration with Weak Supervision
- Data Shifts

Problem Statement and Background

- How does novelties happen for MMIR?
- How does novelties affect multimodal data?
- Data Integration with Weak Supervision
- Data Shifts

Covariate Shift: *Change in application domain with same modalities, or user writes queries differently.*

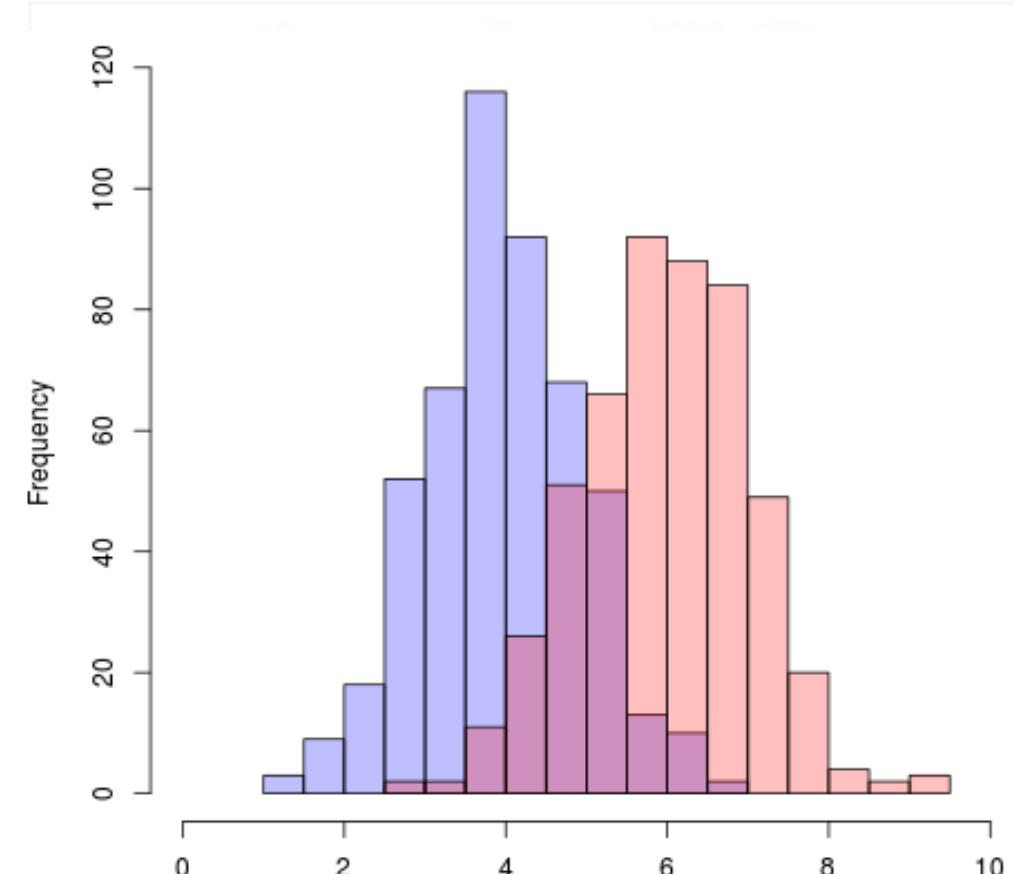


Problem Statement and Background

- How does novelties happen for MMIR?
- How does novelties affect multimodal data?
- Data Integration with Weak Supervision
- Data Shifts

Covariate Shift: *Change in application domain with same modalities, or user writes queries differently.*

Prior Shift: *Distribution change of class-label variable, or relevance-label variable, or weak-feature variable (including no weak feature).*



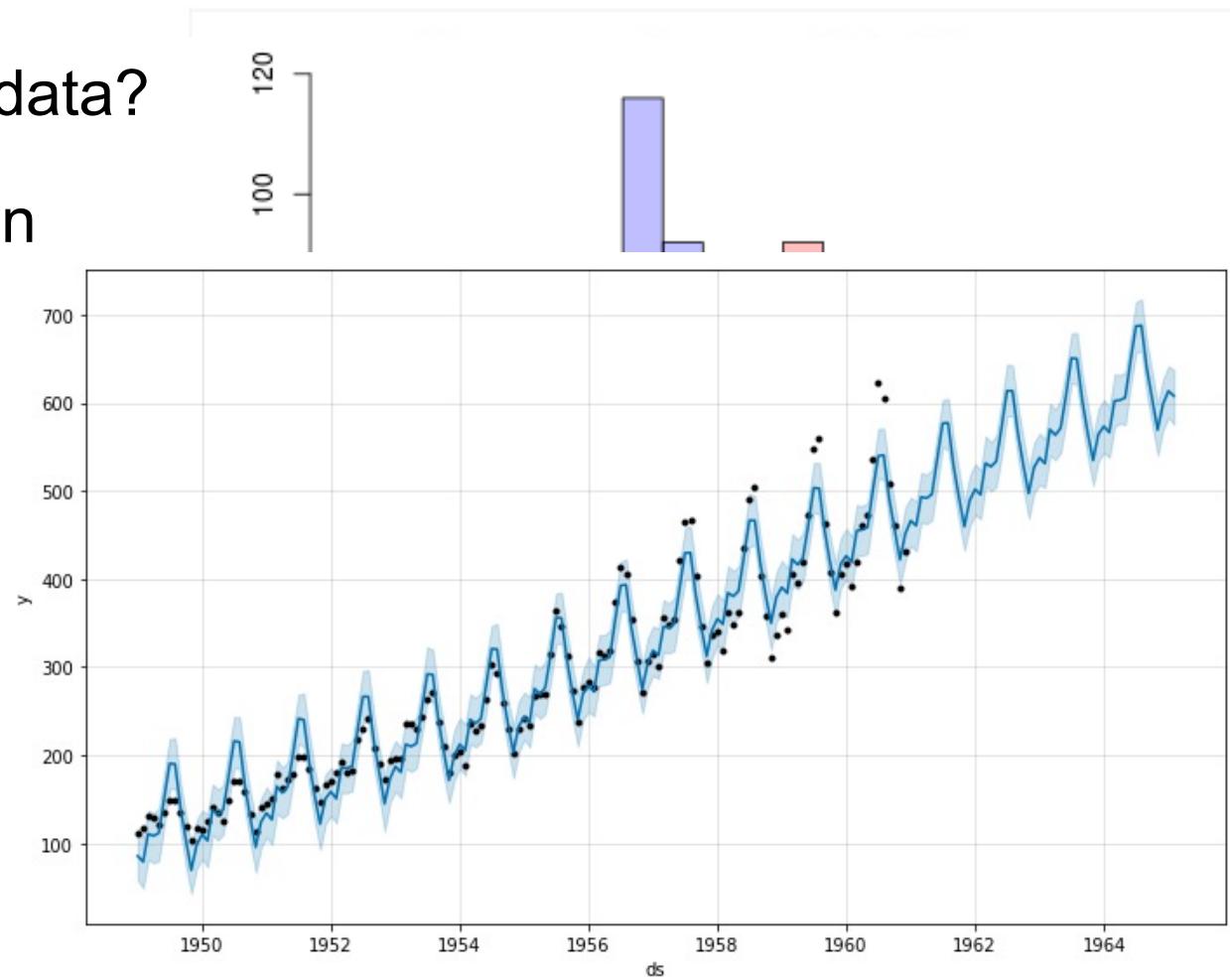
Problem Statement and Background

- How does novelties happen for MMIR?
- How does novelties affect multimodal data?
- Data Integration with Weak Supervision
- Data Shifts

Covariate Shift: *Change in application domain with same modalities, or user writes queries differently.*

Prior Shift: *Distribution change of class-label variable, or relevance-label variable, or weak-feature variable (including no weak feature).*

Concept Drift: *Can be temporal effect or user requirement change over time.*



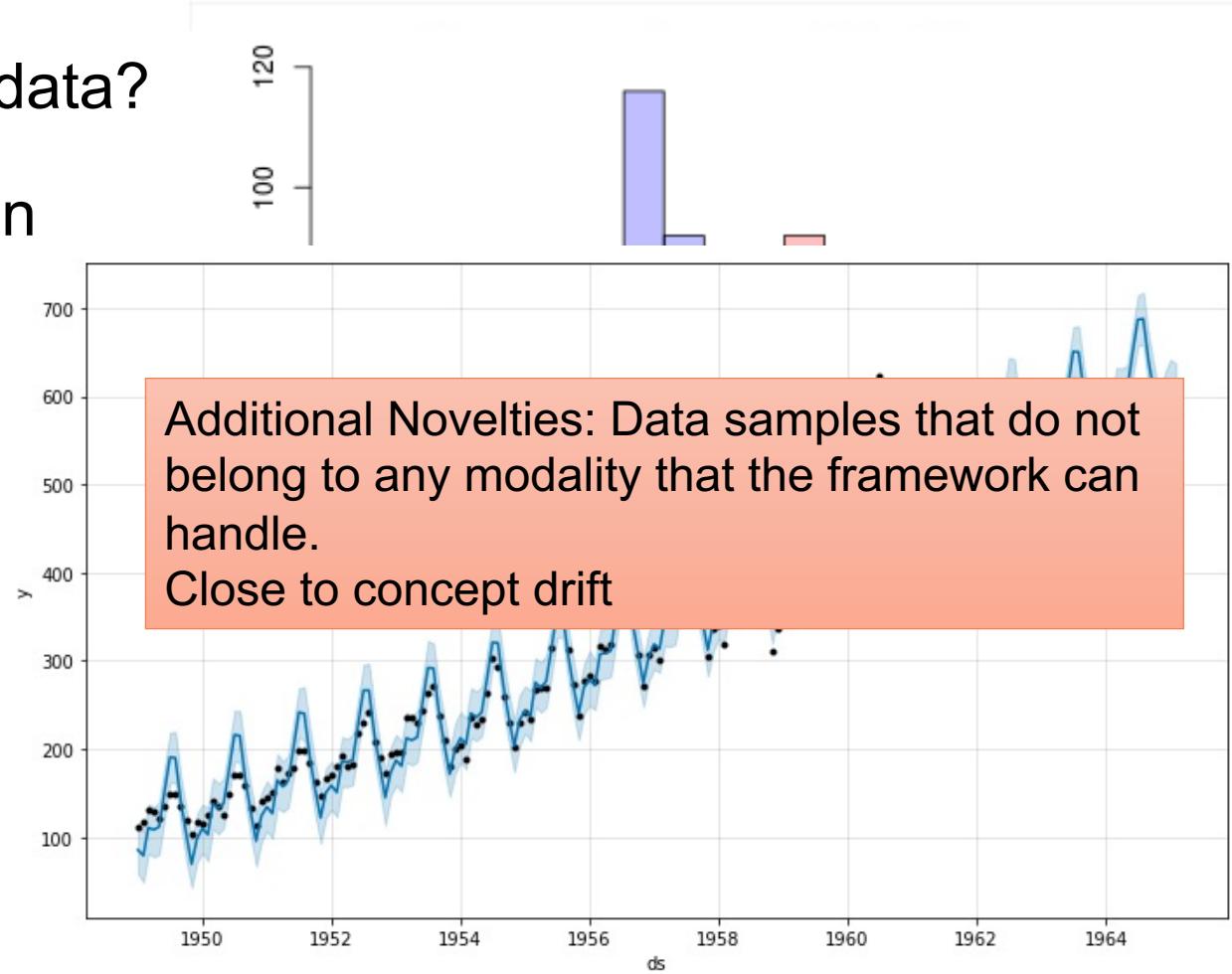
Problem Statement and Background

- How does novelties happen for MMIR?
- How does novelties affect multimodal data?
- Data Integration with Weak Supervision
- Data Shifts

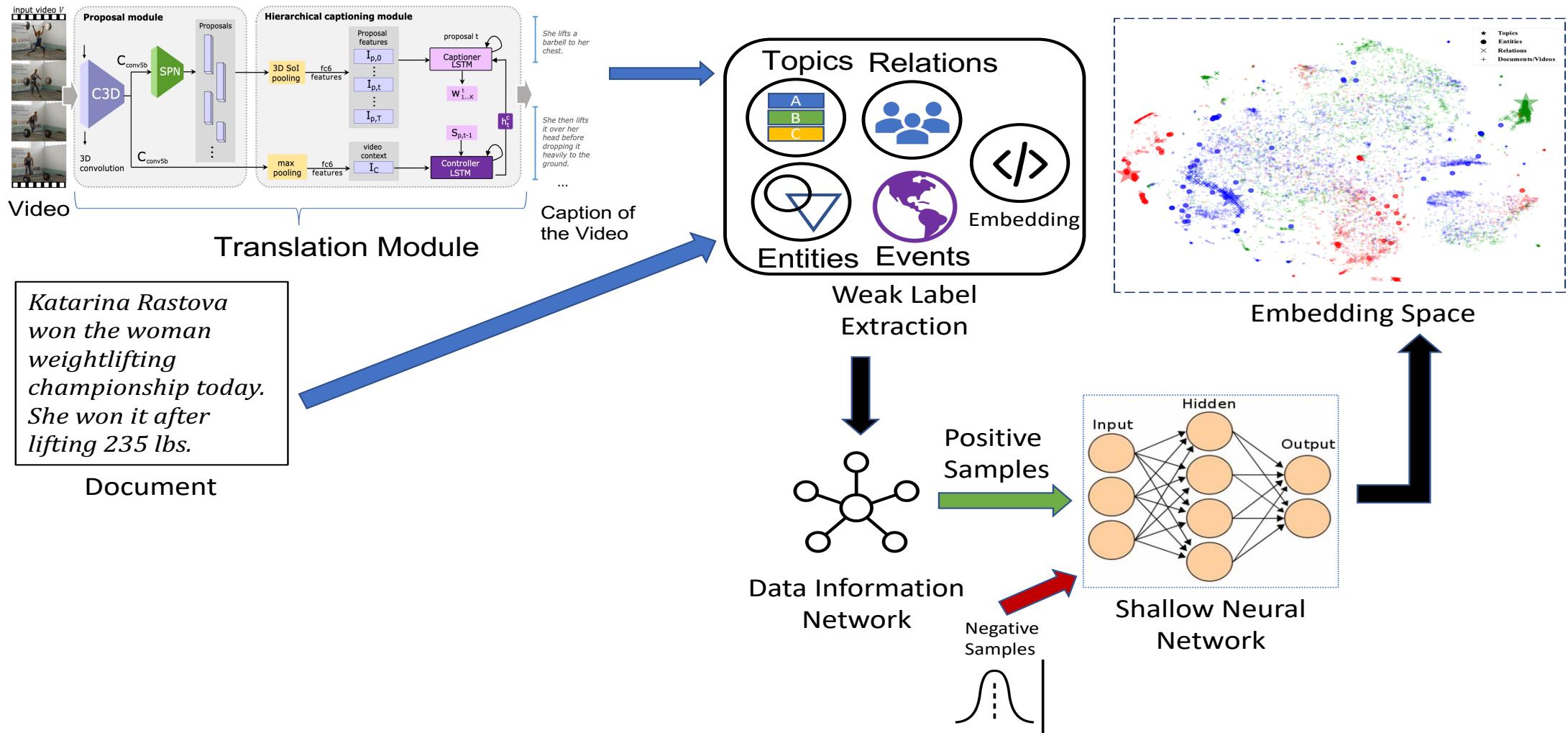
Covariate Shift: *Change in application domain with same modalities, or user writes queries differently.*

Prior Shift: *Distribution change of class-label variable, or relevance-label variable, or weak-feature variable (including no weak feature).*

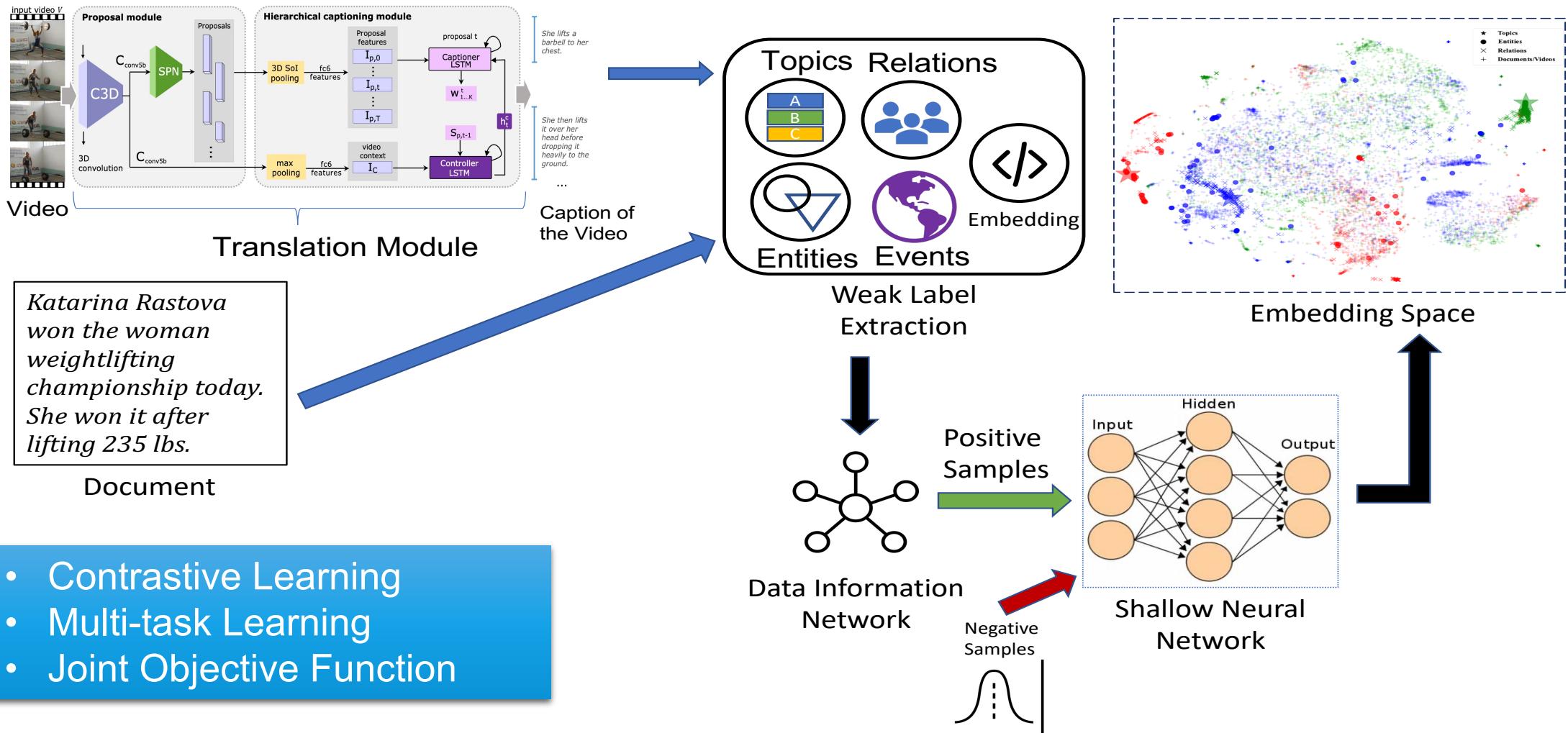
Concept Drift: *Can be temporal effect or user requirement change over time.*



Weakly Supervised Joint Embedding



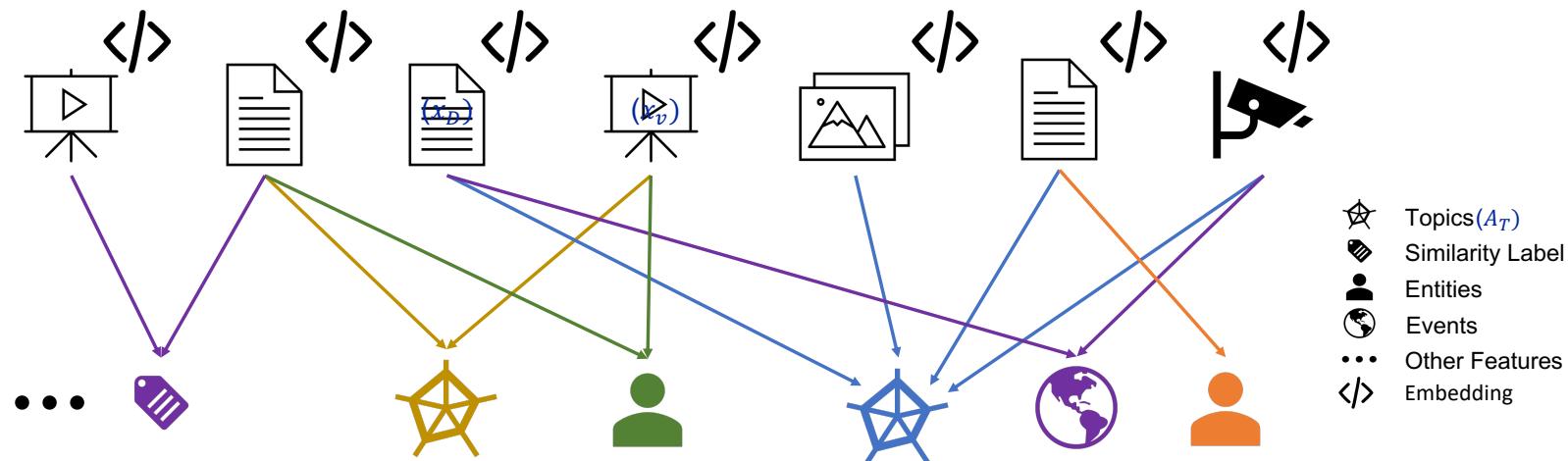
Weakly Supervised Joint Embedding



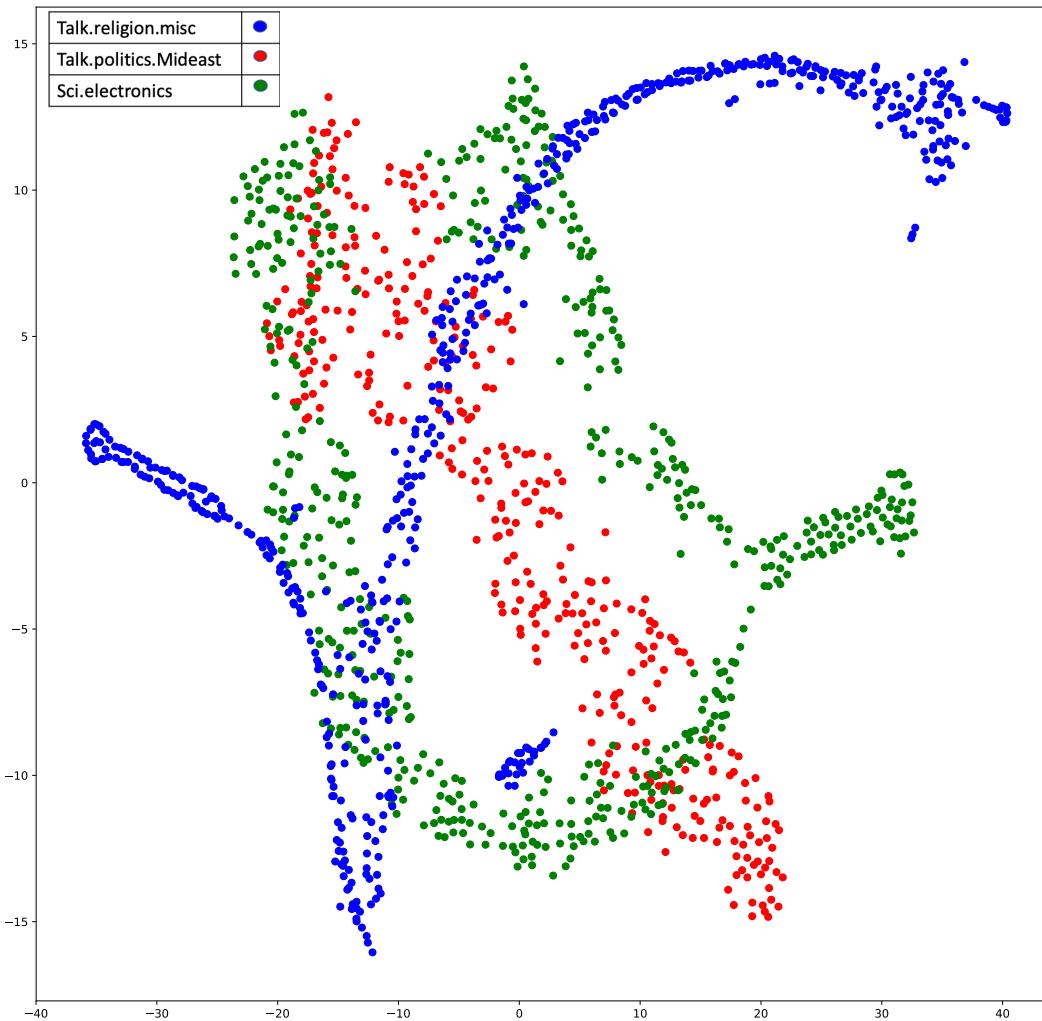
- Contrastive Learning
- Multi-task Learning
- Joint Objective Function

Novelty Detection in WesJeM

- **Data information network** is used to detect the changes during post-novelty inference.
- **Novel Instance.**
 - A test instance x is novel if $G(V_{P_{tr+x}}, E)$ is different from $G(V_{P_{tr}}, E)$.
 - Considering a knowledge base for the weak features during training (A_{tr}), if weak features are absent in A_{tr} during testing, the instance is novel.



Results for Topics-Topics Objective Function

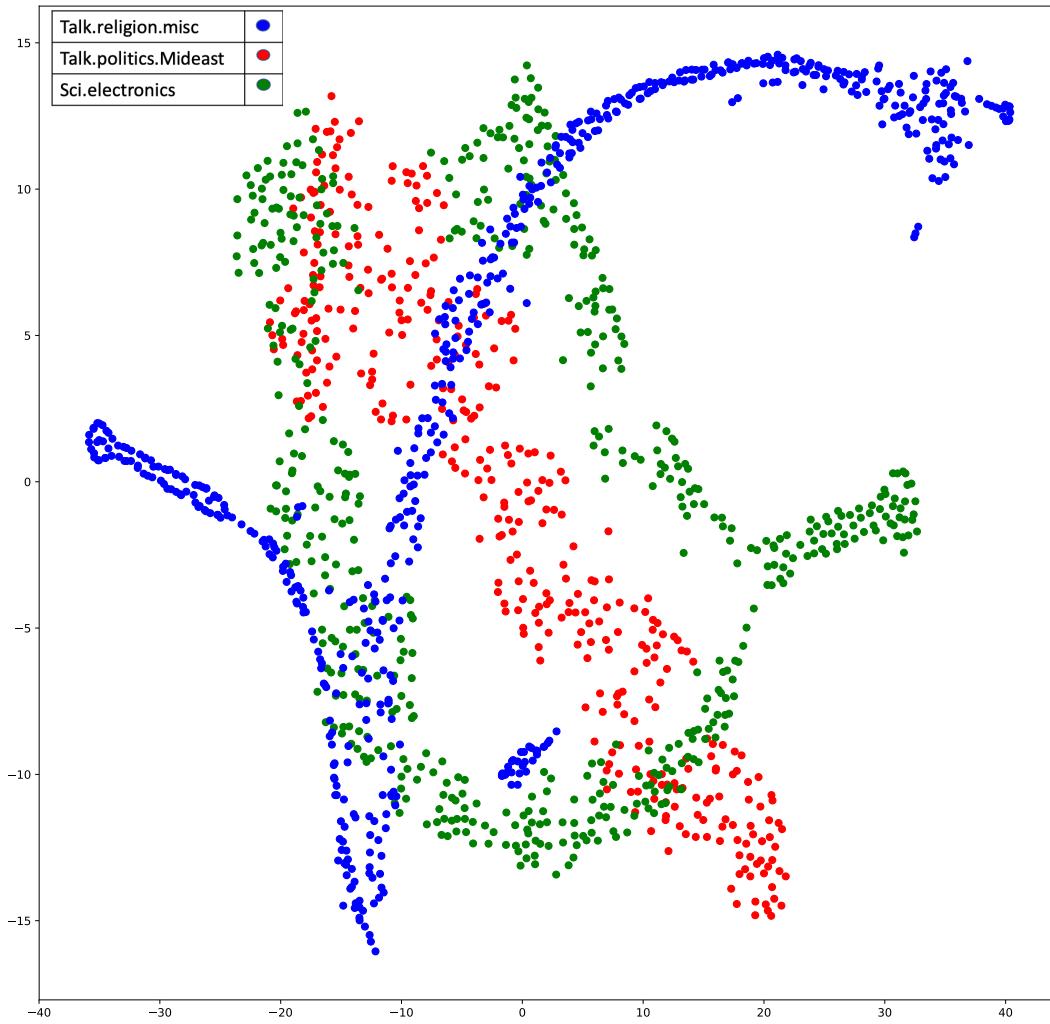


- 20 Newsgroups Dataset
- Baseline: LSA, LDA
- Unimodal: Documents
- Measures inter-similarity and intra-similarity between two halves of different documents
- Different Negative Sampling Heuristics

Table 1: Performance Comparison Results of DT2DVec

	LSA	LDA	DT2DVec
Inter-similarity	0.76	0.66	0.61
Intra-similarity	0.45	0.28	0.047

Results for Topics-Topics Objective Function

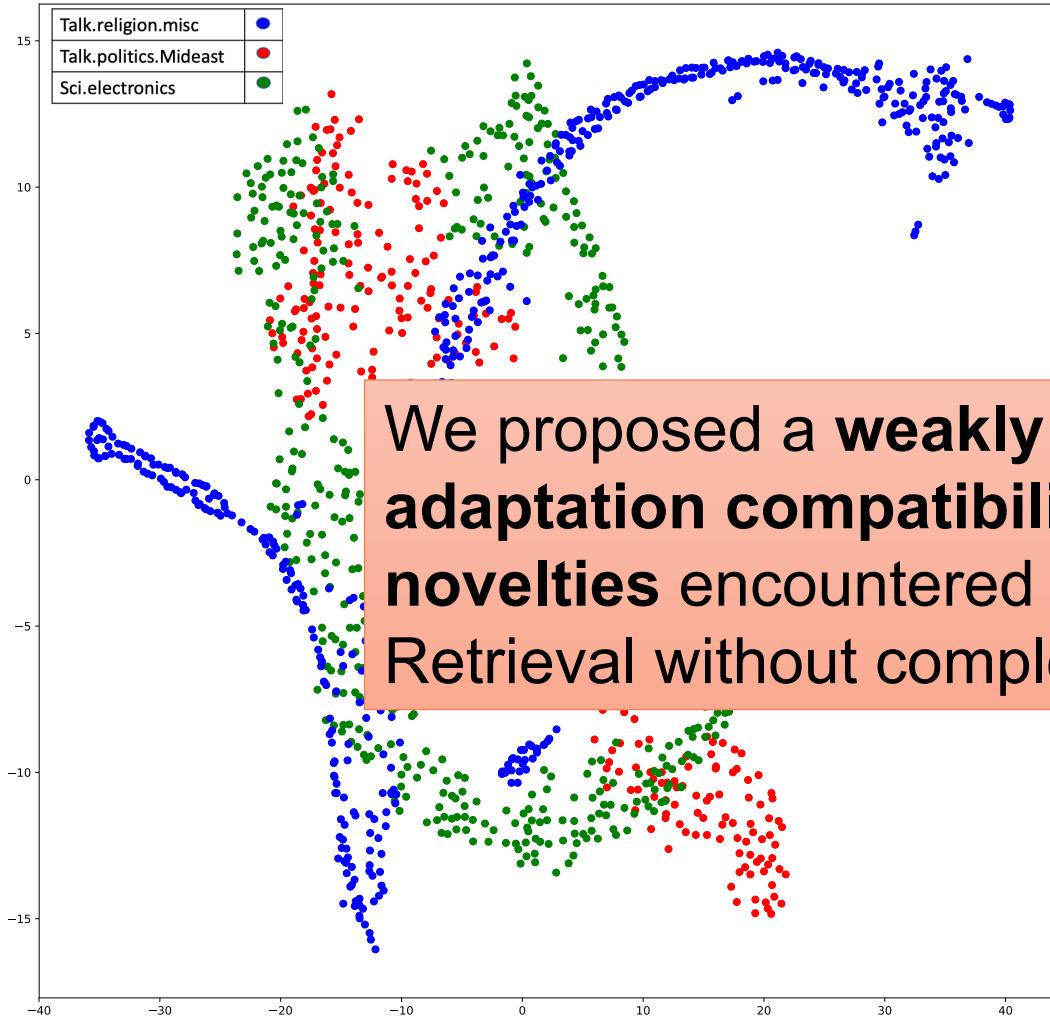


- 20 Newsgroups Dataset
- Baseline: LSA, LDA
- Unimodal: Documents
- Measures inter-similarity and intra-similarity between two halves of different documents
- Different Negative Sampling Heuristics

Table 1: Performance Comparison Results of DT2DVec

	LSA	LDA	DT2DVec
Inter-similarity	0.76	0.66	0.61
Intra-similarity	0.45	0.28	0.047

Results for Topics-Topics Objective Function



- 20 Newsgroups Dataset
- Baseline: LSA, LDA
- Unimodal: Documents
- Measures inter-similarity and intra-similarity between the two halves of the dataset

Table 1: Performance Comparison Results of DT2DVec

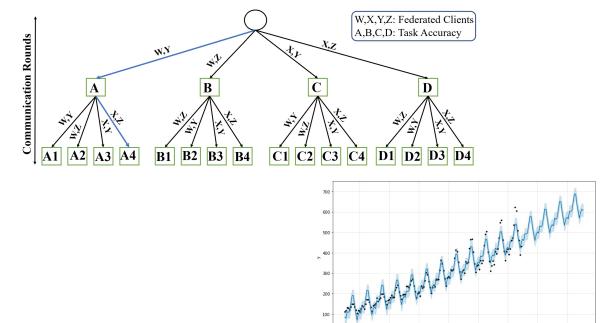
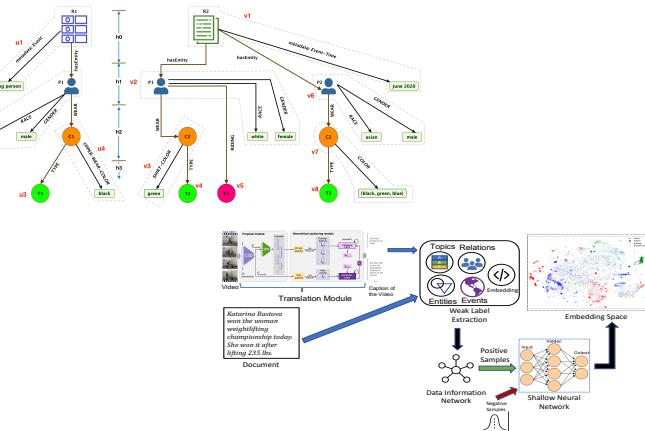
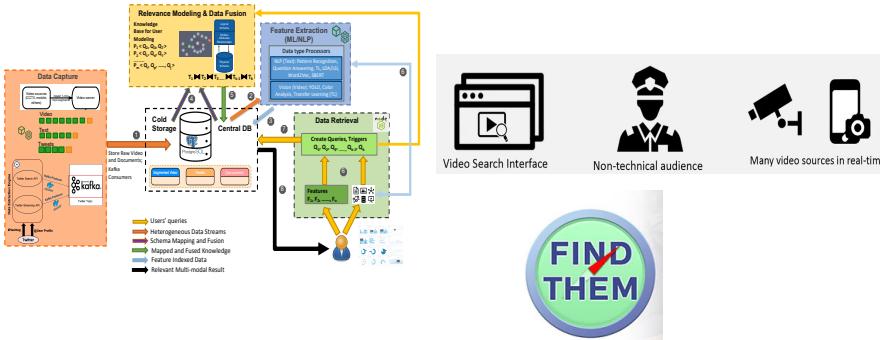
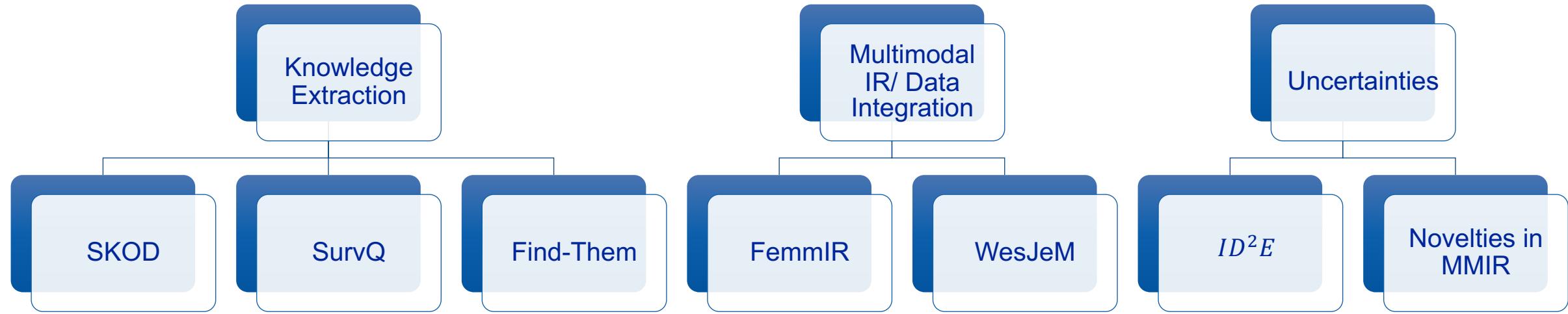
	LSA	LDA	DT2DVec
Inter-similarity	0.76	0.66	0.61
Intra-similarity	0.45	0.28	0.047

Future Works

- User Preference Modeling
 - How to model users' information need in a robust and efficient manner?
 - user requirement is not always obvious or explicitly stated
 - user can be interested in multiple types of events and knowledge bases with varying probabilities
 - Learning algorithms need to adapt to changing user preferences with time
- Trust and Privacy
 - How to avoid bias, increase trust and privacy in recommended results?
- How to build data management strategies that leverages multimodal data as a service (MDaaS)?

Questions?

Accurate data, at the **right place**, and the **right time**.



References

1. S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, “Fusing audio, visual and textual clues for sentiment analysis from multimodal content,” *Neurocomputer*, vol. 174, no. PA, pp. 50–59, Jan. 2016
2. Foresti, G.L., Farinosi, M., Vernier, M.: Situational awareness in smart environments: socio-mobile and sensor data fusion for emergency response to disasters. *J. Ambient Intelligence and Humanized Computing* 6(2), 239–257 (2015).
3. Meditskos, G., Vrochidis, S., Kompatsiaris, I.: Description logics and rules for multimodal situational awareness in healthcare. In: *MMM* (1). *Lecture Notes in Computer Science*, vol. 10132, pp. 714–725. Springer (2017)
4. Adjali, O., Hina, M.D., Dourlens, S., Ramdane-Cherif, A.: Multimodal fusion, fission and virtual reality simulation for an ambient robotic intelligence. In: *ANT/SEIT*. *Procedia Computer Science*, vol. 52, pp. 218–225. Elsevier (2015)
5. Y. Zhu, J. J. Lim, and L. Fei-Fei, “Knowledge acquisition for visual question answering via iterative querying,” in *CVPR*, IEEE Computer Society, 2017, pp. 6146–6155.
706. Q. Wu, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, “Ask me anything: Free-form visual question