GEORGE MASON UNIVERSITY

System Architecture for Recombinant AI (SARAI)

AI4SE & SE4AI Research and Application Workshop 2024

Dr. Joshua Poore

Associate Research Scientist with the Applied Research Laboratory for Intelligence and Security (ARLIS)

GEORGE MASON UNIVERSITY

Dr. Ali Raz

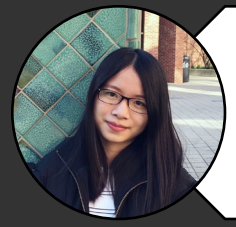
Assist. Prof. Systems Engineering and Operations Research

(Expertise: Complex Systems and System of Systems)

Dr. Paulo Costa

Professor – Cyber Security Engineering

(Expertise: Information Fusion and Cyber Security)



Dr. Ziyu Yao

Assist. Prof. Computer Science

(Expertise: Natural Language Processing)

Dr. Shou Matsumoto

Research Assist. Prof. C4I and Cyber Center

(Expertise: Probabilistic Programming and Ontology)

Saurabh Srivastava

PhD Student

Nikita Patil

MS Student

Gaurav Singh

MS Student

Undergrad. Student Annotators

The Project Team

System Architecture for Recombinant AI (SARAI) Objectives

Provide reusable technical assets to expedite and inspire future research into SE4AI

- Quick **prototyping**, **combination**, and **reuse** of data and AI software assets

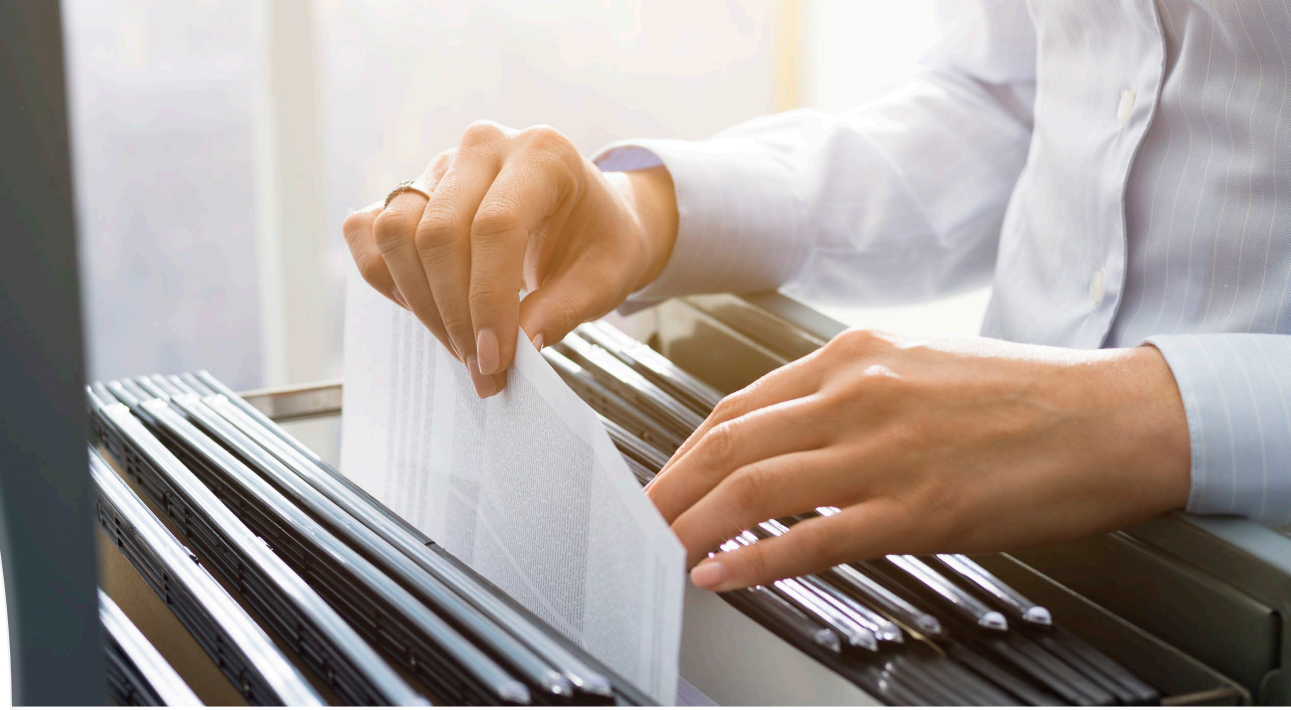
Proof-of-concept experimentation to demonstrate critical capabilities that shape how AI is acquired and evaluated

- Address the **reuse**, **repurpose**, **obsolescence** of current and historical AI applications
  - acquisitions, machine learning, autonomy, advanced analytics…

SARAI Tasks/Works

- GMU/C5I center worked towards:
  - development and integration of a **pipeline of AI solutions** within the scope of **document processing and triage**.

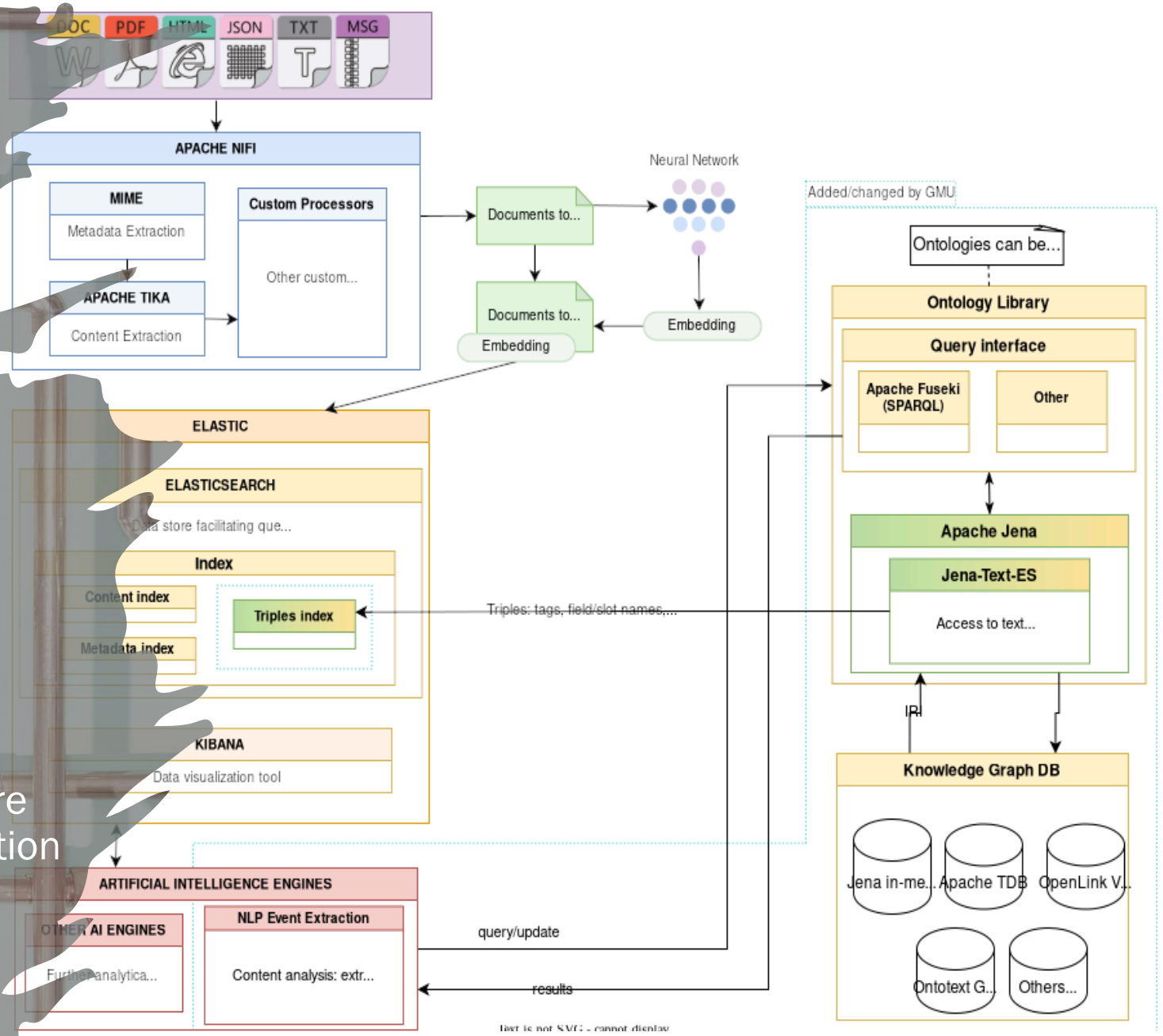
- This work exercises/provides:
  - basis for demonstrating how **parallel lines of AI development efforts** and **legacy works** can benefit from underlying **system architecture**
    - **Methodically integrated** various AI solutions



Methodology

Synopsis:

Our recombinant AI engineering **pipeline** develops a system engineering architecture with a focused **ontology** to provide integration of separately developed AI tools...

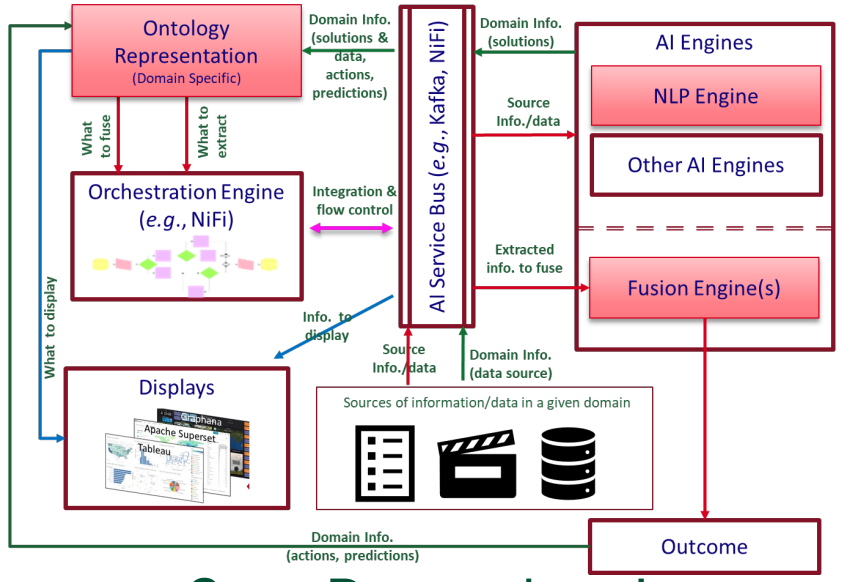


Event Extraction form Ontological Framework and NLP

- **Objective:**
  - Complement data ingest pipeline to enable event extraction via an ontological framework
- **Solution:**
  - Combine Natural Language Processing (NLP) with a domain-specific Ontology to identify key entities of interest and how these entities are interlinked

Discriminating Features

- Knowledge base construction and population via Machine Learning
- Semi-automatic ontology development (Machine Learning + “curators”)
- (Natural language-based) user interface
- Fact verification & aggregation
- **Domain-specific applications, such as email triage**



Core Dependencies

Ontology:

- Apache Jena, Fuseki2, and TDB2
- Protégé (GUI)
- OWLGrEd (data visualization)

NLP tools:

- Hugging Face Hub/API/Widget
- PyTorch

Probabilistic reasoning (future):

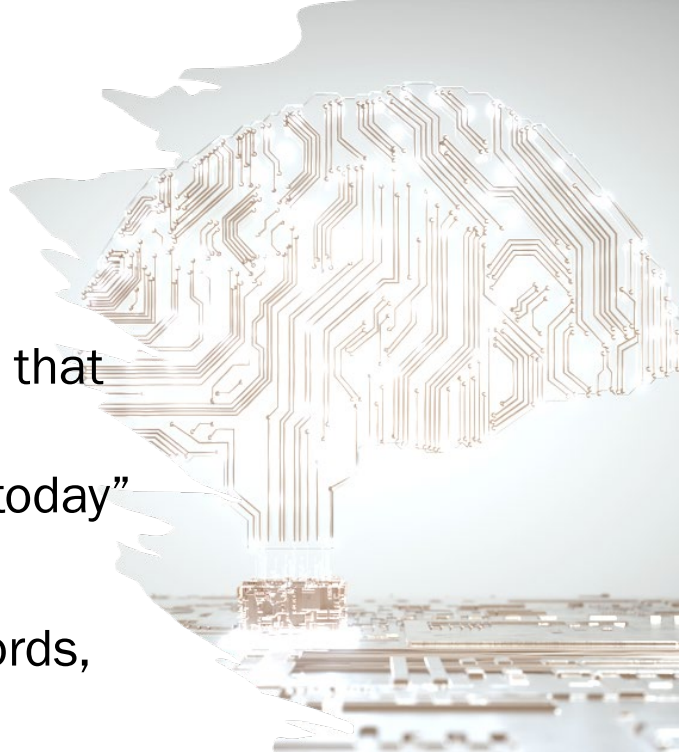
- UnBBayes framework

C5I Center

Ontology and NLP-based framework sets the foundations to exploit high-level data fusion towards inferring root causes and/or potential courses of action.

Neural Network NLP Modules for Email Event Extraction

- A three-stage pipeline

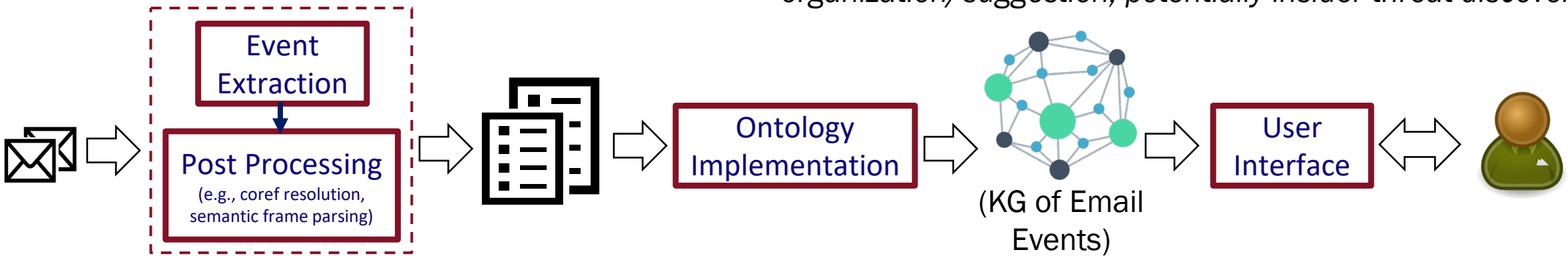


1) Trigger identification
   - Given an email thread, extract words that signal the types of event
   - E.g., “please send me the summary today”
2) Event type prediction
   - Given an email thread and trigger words, predict the type of event
   - E.g., “send me the summary” → Request Data
3) Argument extraction
   - Given an email thread, the trigger words, and the event type, extract word spans as arguments/roles for the event
   - E.g., “the summary” is the requested data, “today” is the requested date

Srivastava, Saurabh, Gaurav Singh, Shou Matsumoto, Ali Raz, Paulo Costa, Joshua Poore, and Ziyu Yao. 2023. “MailEx: Email Event and Argument Extraction.” In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 12964–87. Singapore: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.801.

Instantiation: Constructing a Domain-specific Knowledge Graph

Domain: Email data (for email triage, intelligent task organization/suggestion, potentially insider threat discovery, etc.)



EmailID3

```
Hi Andrew, the attached is a PDF of the summary report -- Alice
```

EmailID10

```
How about meeting this Wed to discuss the report from Alice?
```

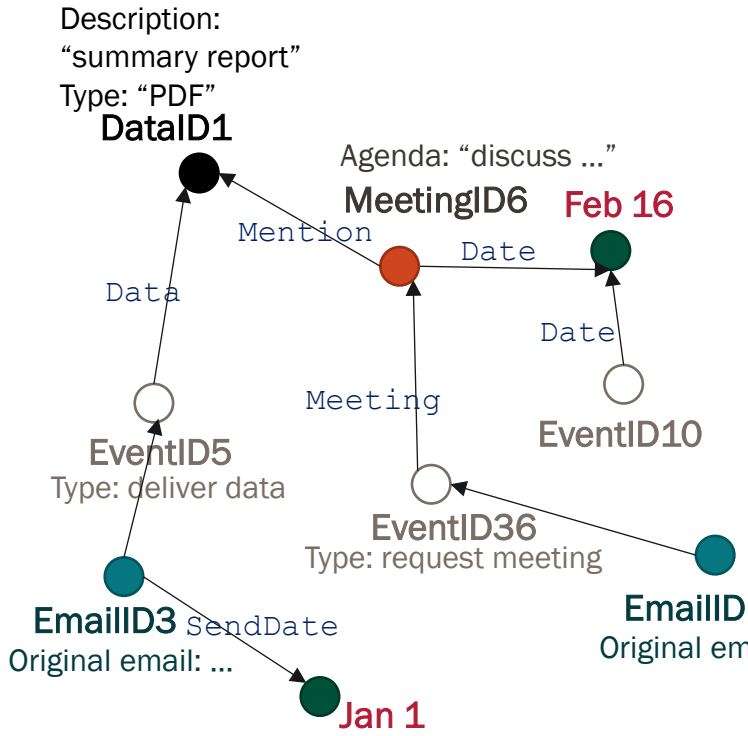
EventID5

```
EventType: Deliver Data
Data: DataID1
|- Type: "PDF"
|- Description: "summary report"
```

EventID36

```
EventType: Request Meeting
Meeting: MeetingID6
|- Agenda: "discuss the report from Alice" → DataID1
|- Date: "this Wed"
```

Feb 16



Scenario 1: SPARQL query

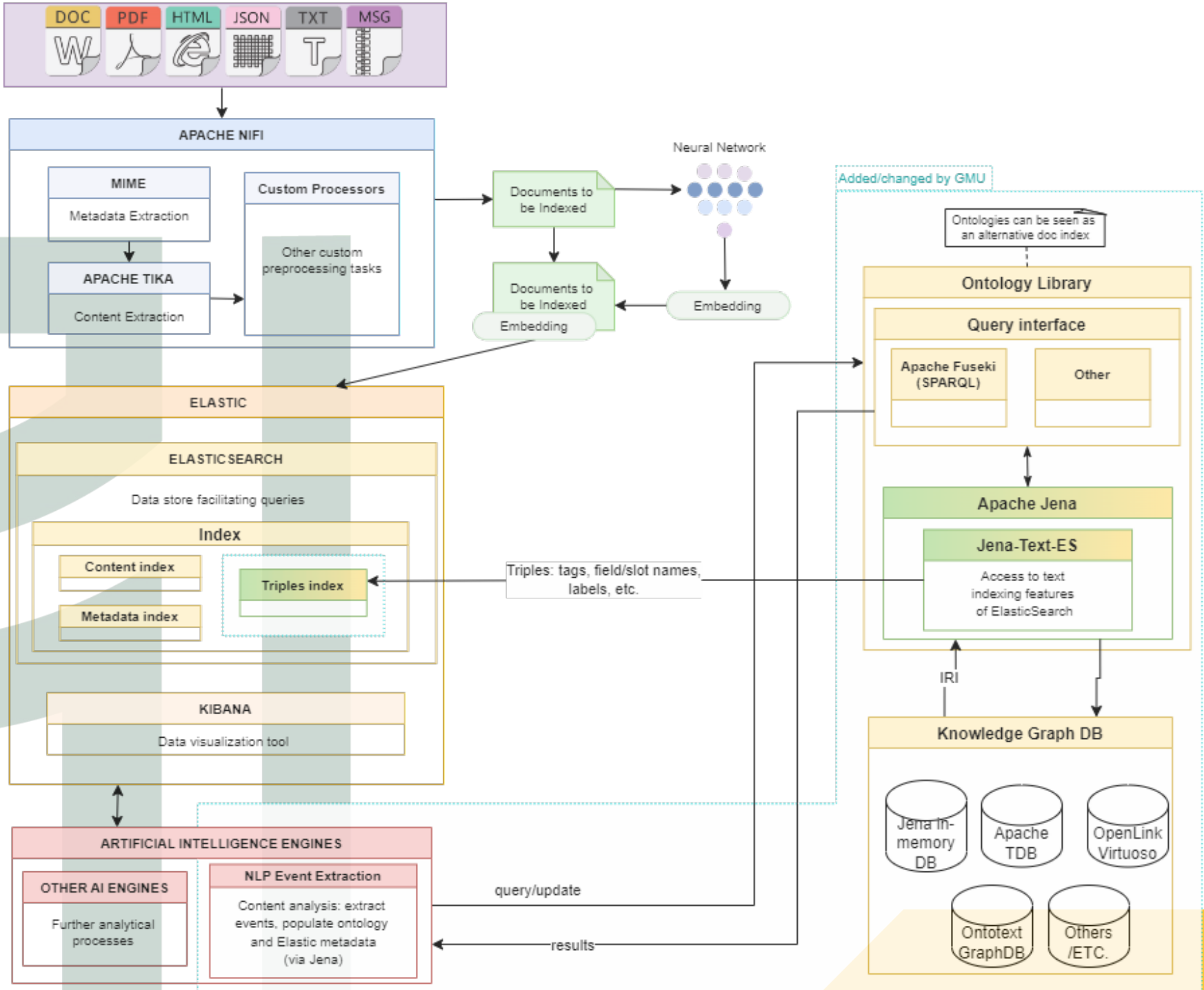
Scenario 2: Semantic search via natural language (e.g., "show me a description of the data we will discuss in the Feb 16 meeting")

And more: reasoning, pattern learning for anomaly detection, etc.

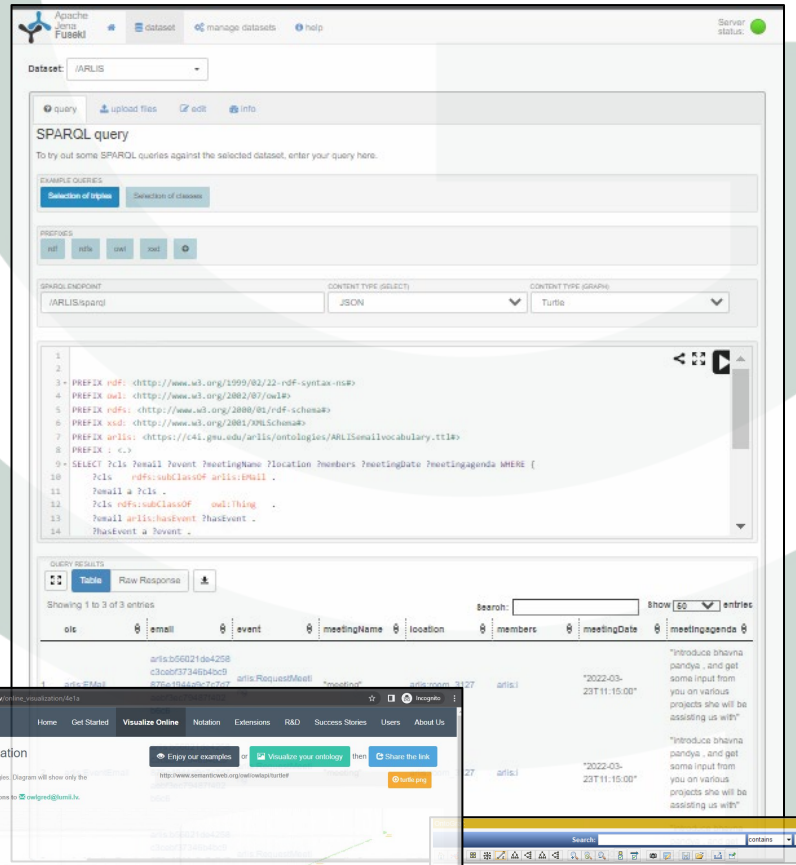
C5I Center

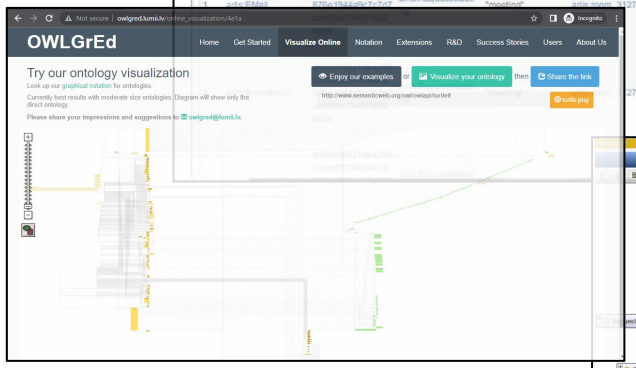
GMU-ARLIS System Architecture

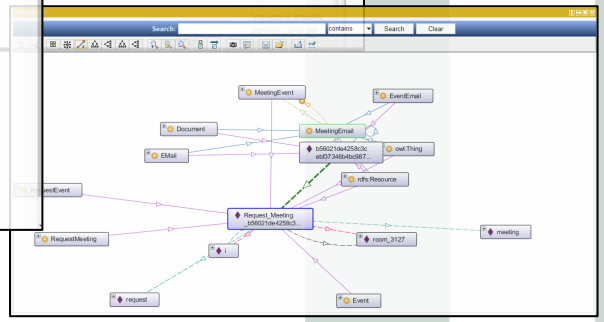
Components' collaboration model

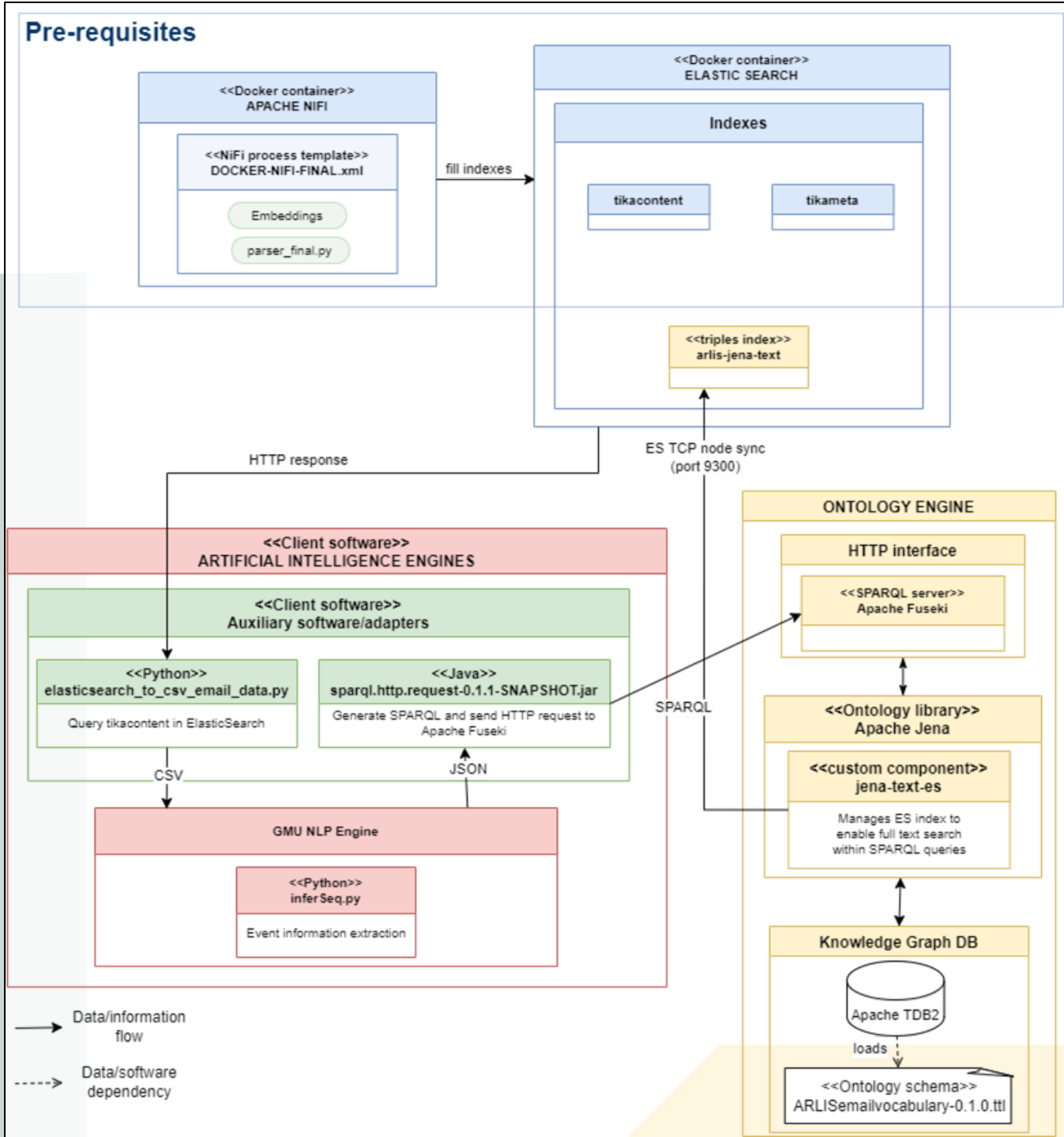


Prototype/Demo Implementation









Prototype/demo implementation – raw data (Enron emails)

Formatted/filtered for better visualization

	A	C	F	G	H	S	T	U	V	W	X	Y	Z	AA	AB	AW	AX	AZ	BA	BB
1	Body	_id	_type	content_SHA-256	content_SHA-256.keyword	email.headers.from.0	email.headers.from.0.keyword	email.headers.message-id	email.headers.message-id.keyword	email.headers.sent_date	email.headers.sent_date.keyword	email.headers.subject	email.headers.subject.keyword	email.headers.to.0	email.headers.to.0.keyword	file.creationTime	file.group	file.lastAccessTime	file.lastModifiedTime	file.permissions
5	I would support Matt Lenhart's promotion to the next level.	zepfLoABNLR	_doc	cfb4bc3fdb	cfb4bc3fdb	phillip.allen@	phillip.allen@e	<5361157.107!	<5361157.107!	Wed Dec 06 12	Wed Dec 06 12	Re: Associates	Re: Associates	andrea.richarc	andrea.richarc	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
6	---------- Forwarded by Phillip K Allen/HOU/ECT on 09/08/2000 12:28 PM ----------	cepfLoABNLR	_doc	af8927cda!	af8927cda!	phillip.allen@	phillip.allen@e	<27210125.10	<27210125.10	Fri Sep 08 12:2	Fri Sep 08 12:2	Westgate Profo	Westgate Profo	pallen70@hot	pallen70@hot	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
7	---------- Forwarded by Phillip K Allen/HOU/ECT on 09/08/2000 12:29 PM ----------	eepfLoABNLR	_doc	6bc452eba	6bc452eba	phillip.allen@	phillip.allen@e	<5509840.107!	<5509840.107!	Fri Sep 08 12:3	Fri Sep 08 12:3	Sagewood Tow	Sagewood Tow	pallen70@hot	pallen70@hot	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
8	Kathy, Regarding the guest password for	TepfLoABNLR	_doc	f33c5c7475	f33c5c7475	phillip.allen@	phillip.allen@e	<20726955.10	<20726955.10	Fri Sep 22 07:0	Fri Sep 22 07:0	-	-	kathy.moore@	kathy.moore@	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
9	---------- Forwarded by Phillip K Allen/HOU/ECT on 09/25/2000 02:00 PM ----------	T-pfLoABNLR	_doc	d4709970a	d4709970a	phillip.allen@	phillip.allen@e	<22224290.10	<22224290.10	Mon Sep 25 14	Mon Sep 25 14	-	-	ina.rangel@er	ina.rangel@er	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
10	9/8 9/7 diff Socal 36,600 37,200 -600	VOpfLoABNLF	_doc	14a8686c3	14a8686c3	phillip.allen@	phillip.allen@e	<32465930.10	<32465930.10	Mon Sep 11 15	Mon Sep 11 15	-	-	john.lavorato@	john.lavorato@	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
11	Jeff, I received the rent roll. I am going to	WOpfLoABNL	_doc	346873121	346873121	phillip.allen@	phillip.allen@e	<5725770.107!	<5725770.107!	Mon Sep 11 16	Mon Sep 11 16	Re: Chelsea Vi	Re: Chelsea Vi	jsmith@austin	jsmith@austin	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
12	Lucy, I want to have an accurate rent roll as	YOpfLoABNLF	_doc	987dfe0f14	987dfe0f14	phillip.allen@	phillip.allen@e	<25391613.10	<25391613.10	Wed Sep 13 13	Wed Sep 13 13	-	-	stagecoachma	stagecoachma	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
13	---------- Forwarded by Phillip K Allen/HOU/ECT on 09/19/2000 04:35 PM ----------	fupfLoABNLRI	_doc	a0785f42d(	a0785f42d(	phillip.allen@	phillip.allen@e	<8579795.107!	<8579795.107!	Tue Sep 19 16:	Tue Sep 19 16:	Westgate Profo	Westgate Profo	pallen70@hot	pallen70@hot	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
14	John, Denver's short rockies position	f-pfLoABNLRI	_doc	89d7e875b	89d7e875b	phillip.allen@	phillip.allen@e	<29388698.10	<29388698.10	Wed Sep 20 15	Wed Sep 20 15	-	-	john.lavorato@	john.lavorato@	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
15	Chris, What is the latest with PG&E? We	gepfLoABNLR	_doc	6d3097aa1	6d3097aa1	phillip.allen@	phillip.allen@e	<29638176.10	<29638176.10	Mon Sep 25 14	Mon Sep 25 14	-	-	christopher.ca	christopher.ca	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
16	http://www.hearme.com/vc2/?chnlOwnr=pallen@enron.com	hepfLoABNLR	_doc	50d3fc0fb6	50d3fc0fb6	phillip.allen@	phillip.allen@e	<15270601.10	<15270601.10	Tue Sep 12 07:	Tue Sep 12 07:	-	-	moshuffle@hc	moshuffle@hc	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
17	Jeff, I need to see the site plan for Burnet.	iOpfLoABNLR	_doc	afbf89e9c9	afbf89e9c9	phillip.allen@	phillip.allen@e	<32270225.10	<32270225.10	Mon Sep 18 09	Mon Sep 18 09	Re: burnet	Re: burnet	jsmith@austin	jsmith@austin	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
18	Jeff, What is up with Burnet?	nepfLoABNLR	_doc	cbe4ee97c	cbe4ee97c	phillip.allen@	phillip.allen@e	<28267740.10	<28267740.10	Tue Sep 19 10:	Tue Sep 19 10:	-	-	jsmith@austin	jsmith@austin	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
19	George, Here sales numbers from Reagan:	pOpfLoABNLF	_doc	648c9280b	648c9280b	phillip.allen@	phillip.allen@e	<17231095.10	<17231095.10	Tue Sep 19 14:	Tue Sep 19 14:	-	-	cbpres@austi	cbpres@austi	Apr 15, 2022 (	root	Apr 15, 2022 (	Apr 15, 2022 (	rwxrwxrwx
	George,																			

Prototype/demo implementation – classification/inference

Simplified/filtered for better visualization. Names are redacted.

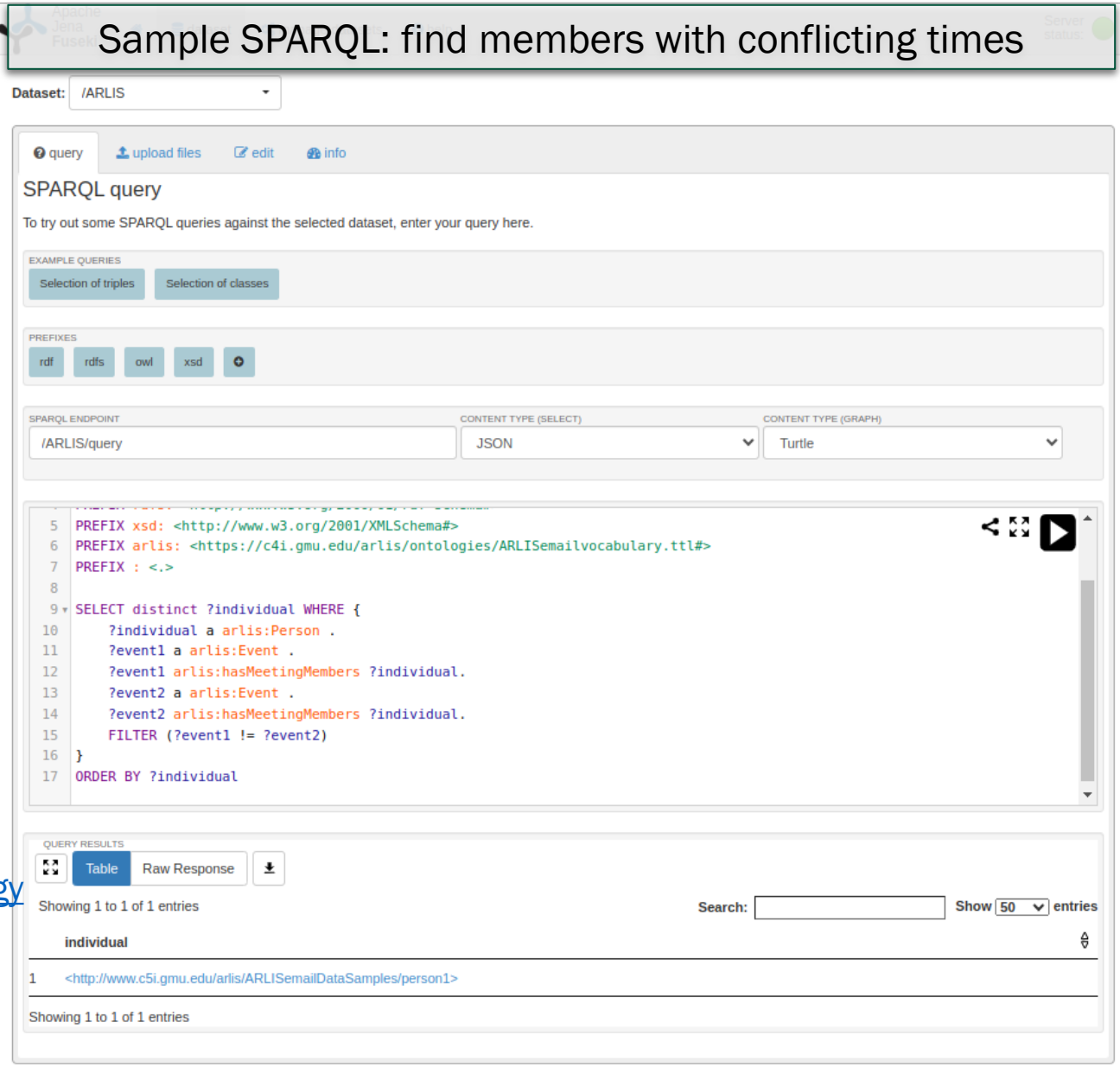
NLP event classification/extraction


```
{
  "ID":
"b56021de4258c3cebf37346b4bc9876e1944a9c7c7d7
aebf3ec79487f402b6c6",
  "Request_Meeting": [ {
      "Meeting Name": "meeting",
      "Meeting Date": "2022-03-23T11:15:00",
      "Meeting Location": "room 3127",
      "Meeting Members": [ "[redacted]", "[redacted]" ],
      "Meeting Agenda": "introduce [redacted] and
get some input on various projects she will
be assisting us with"
  } ]
}
```


Some common concepts were reused from BFO/CCO

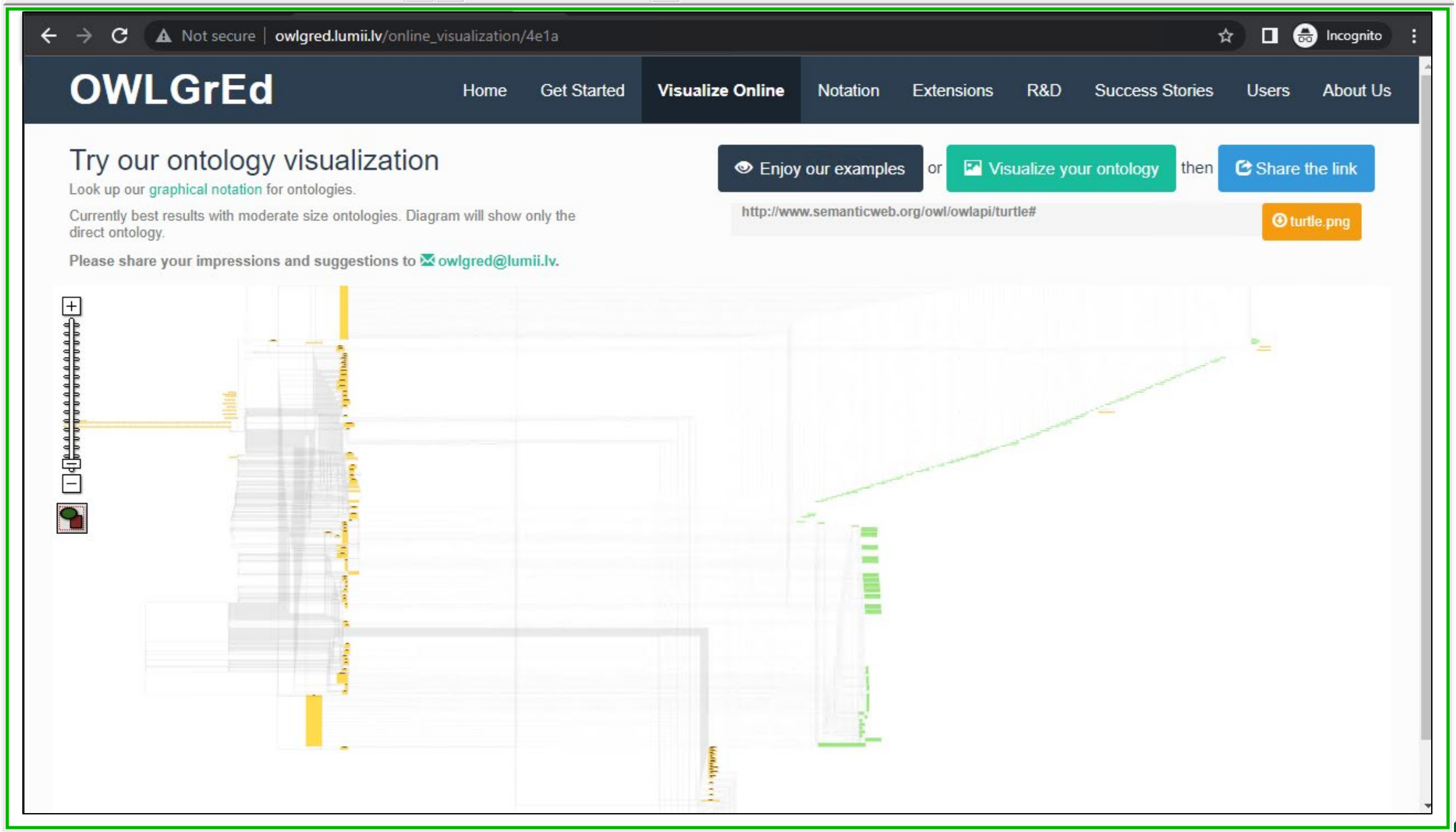
- Basic Formal Ontology (BFO): https://basic-formal-ontology.org/
- Common Core Ontologies (CCO): github.com/CommonCoreOntology

Sample SPARQL: find members with conflicting times



C5I Center

Prototype/demo implementation – visualization tools

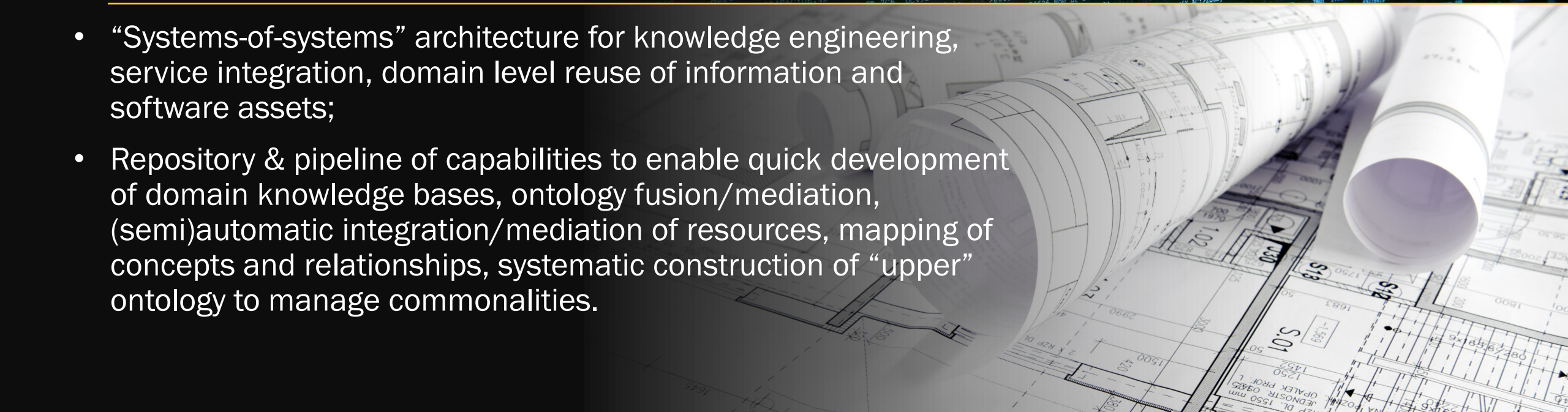


C5I Center

Results/outcomes

Synopsis:

- “Systems-of-systems” architecture for knowledge engineering, service integration, domain level reuse of information and software assets;
- Repository & pipeline of capabilities to enable quick development of domain knowledge bases, ontology fusion/mediation, (semi)automatic integration/mediation of resources, mapping of concepts and relationships, systematic construction of “upper” ontology to manage commonalities.



Conclusion and Key Contributions

- System Architecture & ontology development
  - Use cases of systems integration & semantic technologies
  - Assists reuse, prevents obsolesce of AI solutions
  - Agile/faster prototyping, development, V&V of AI pipelines
- NLP components that produce ontological knowledge base for document triage
  - Framework for parallel data processing, enterprise level data indexing, data curation…
- Technical products disseminated as open-source assets
  - Assist in scaffolding technically similar efforts
  - Provide AI Engineering examples for a larger community

C5I Center

Future Research

Domain adaptation

- Currently, information extraction is on email data only
- Many other domains: science, medicine, cybersecurity, etc.
- Can we build a similar system using as fewer annotations as possible?

Data fusion across different knowledge sources

- Emails, X (social media), news articles, policy documents…
- Text, tables, images, etc.

Language interface to intelligence assistants

- Interpretability & interaction
- Analyst Query

C5I Center

C5I CENTER

Thank you for your attention

GEORGE MASON UNIVERSITY

Backup slides

Recombinant AI Seedling Objectives

- Perform proof of concept experimentation to demonstrate critical capabilities that will shape how artificial intelligence is acquired and evaluated;
- Provide reusable technical assets to expedite and inspire future research into artificial intelligence;
  - Explore reusable and recombinant AI/ML through flexible data engineering pipelines and efficient Data Service architectures
- Provide examples of engineering methodology and process to scaffold the AI Engineering discipline and develop the ODNI's current and future workforce.

GMU Team Sow/Tasks

- Develop analytical frameworks (e.g., *ontological framework*) to enable heterogenous data exploitation and fusion in support of Recombinant AI objectives.

- Building on the outcomes of *Natural Language Processing (NLP)* applied to varies documents in a given domain (e.g., data extraction, indexing, and translation etc.), an ontology framework, for example, will identify key entities of interest in that domain and how these entities are interlinked towards inferring root causes or potential future courses of action.

- Such frameworks set foundations for exploiting and fusing heterogeneous data to overcome limitations of missing data and/or extract new information from disparate and siloed data sets.

C5I Center

Dataset: MailEx

- A conversational email-domain event extraction dataset
  - Source: the Enron email corpus
  - ~1,200 email threads
  - ~3,400 individual emails
  - ~4,500 events
  - (numbers after handling annotator disagreement)

- 11 event types (with arguments)

Event Type	%	Frequent Argument Roles
Request Data	8.91	Data IdString (72%), Request Members (23%), Request Date (2%)
Request Action	20.22	Action Description (54%), Action Members (35%), Action Date (6%)
Request Meeting	5.02	Meeting Members (31%), Meeting Agenda (21%), Meeting Date (18%)
Request Action Data	2.39	Action Description (51%), Action Members (38%), Request Members (8%)
Request Meeting Data	0.71	Meeting Members (31%), Meeting Agenda (21%), Meeting Date (18%)
Deliver Data	24.32	Data IdString (48%), Data Value (39%), Deliver Members (10%)
Deliver Action Data	28.72	Action Description (46%), Action Members (41%), Action Date (9%)
Deliver Meeting Data	6.21	Meeting Members (34%), Meeting Date (19%), Meeting Time (12%)
Amend Data	2.22	Amend Members (26%), (Context) Data IdString (25%), (Revision) Data Value (25%)
Amend Meeting Data	1.27	(Revision) Meeting Time (22%), (Revision) Meeting Date (19%), (Context) Meeting Name (16%)

Table 3: Distributions of event types (in percentage) and frequent argument roles in MAILEX.

C5I CENTER

Intelligent ML/NLP Component

- A neural network model for event extraction on email threads

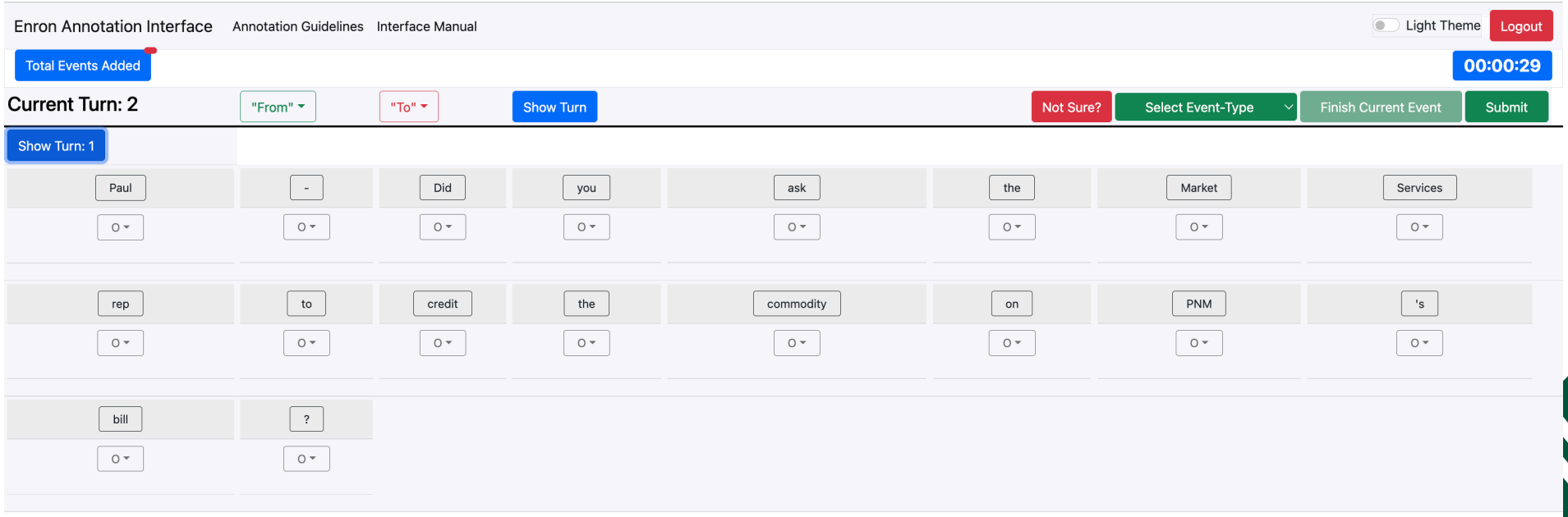
Hugging Face

PYTORCH

Email Event Definition
(Ontology schema design, literature study, intensive expertise)

Email Data Annotation
(Python Libraries: Flask for backend, SQLAlchemy for DB, Jinja2 for Jscript & Flask)

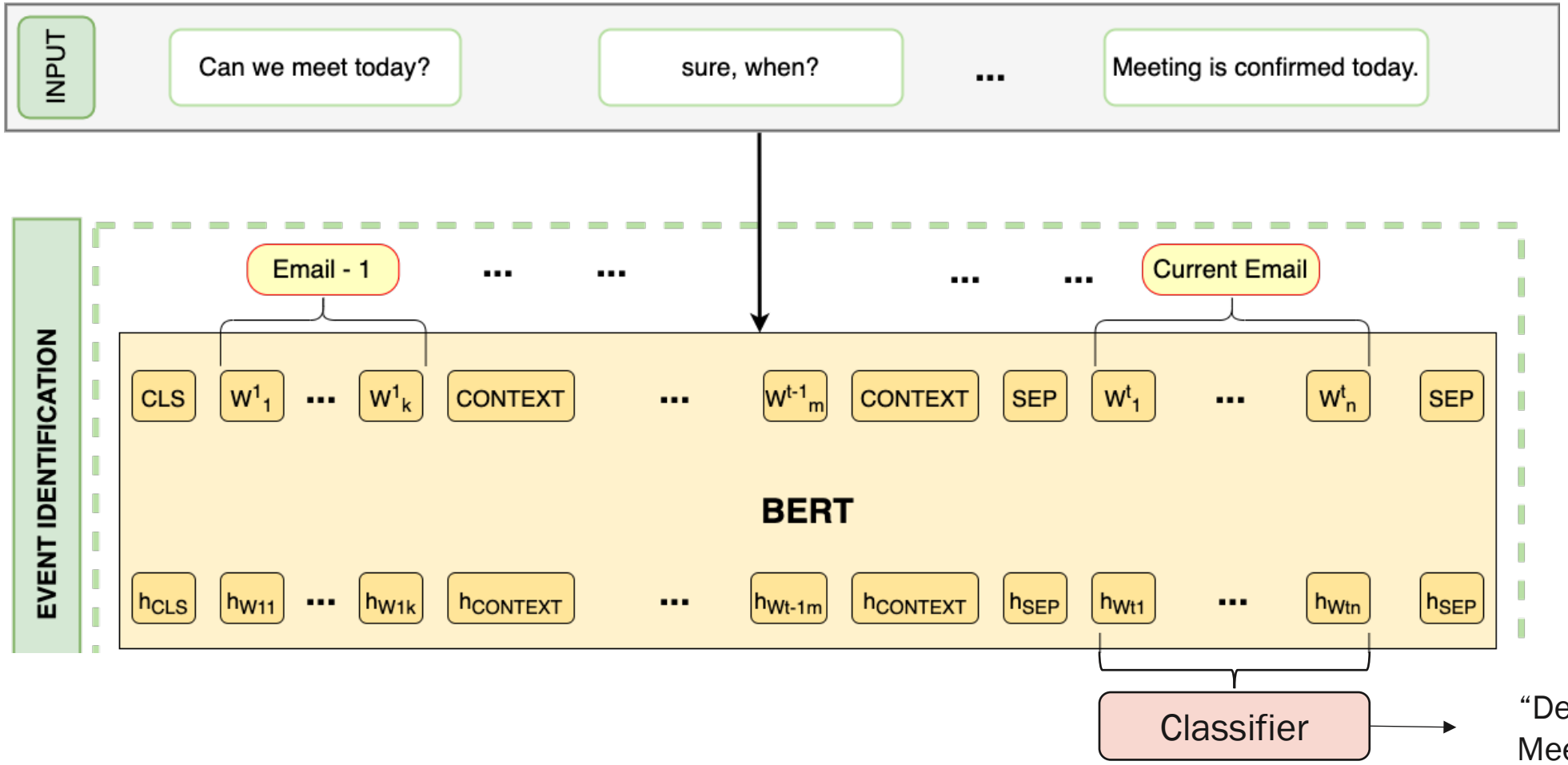
Neural Model Construction
(Implement with Python Hugging Face & PyTorch library)



C5I Center

Neural Network Modules for Email Event Extraction

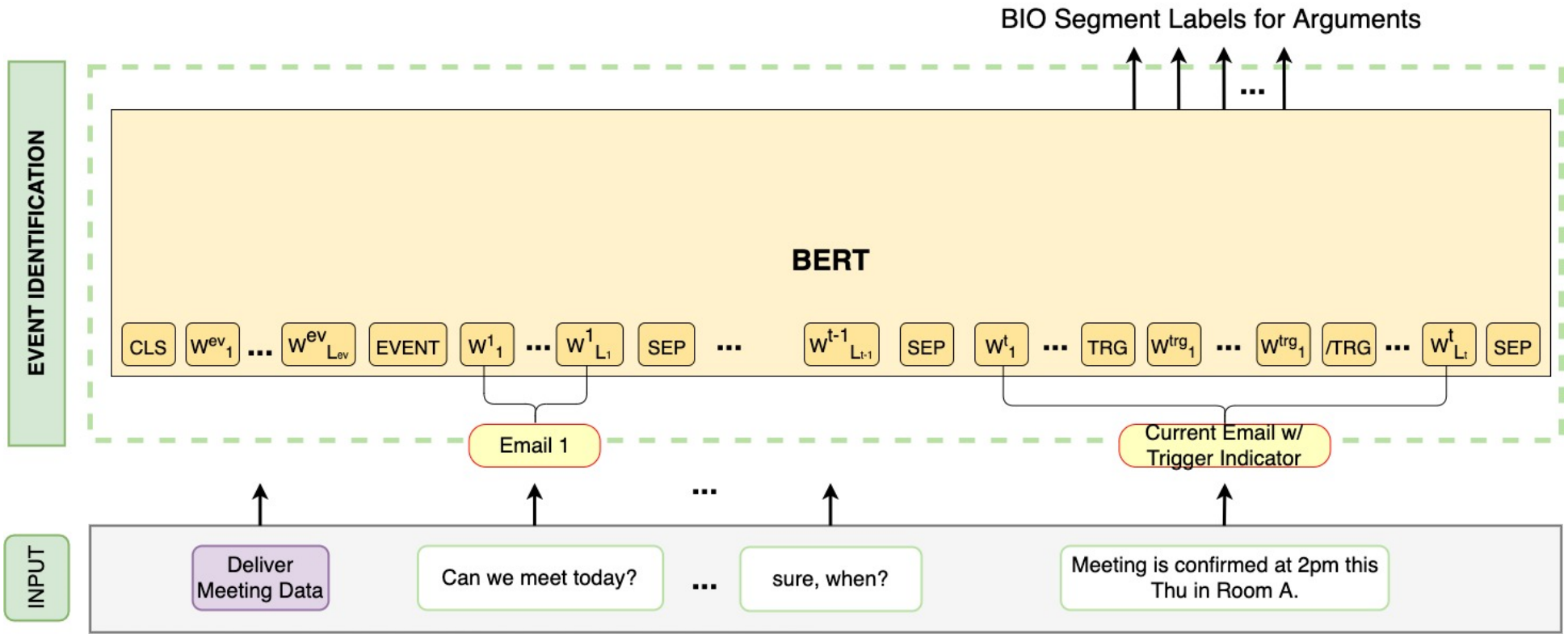
- Event Identification
  - Given the current email and the thread history, identify event types in the email



C5I Center

Neural Network Modules for Email Event Extraction

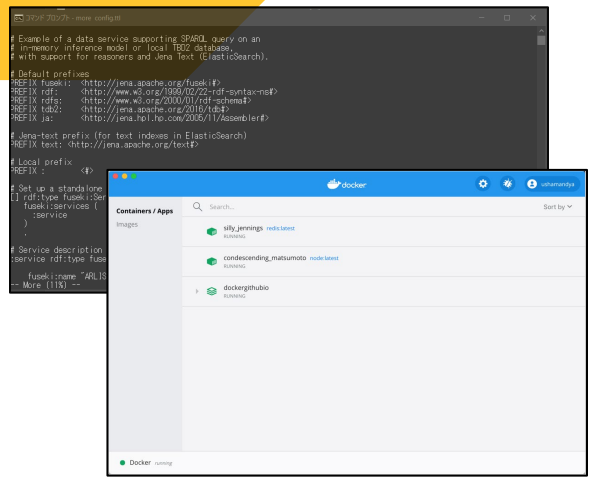
- Argument Extraction



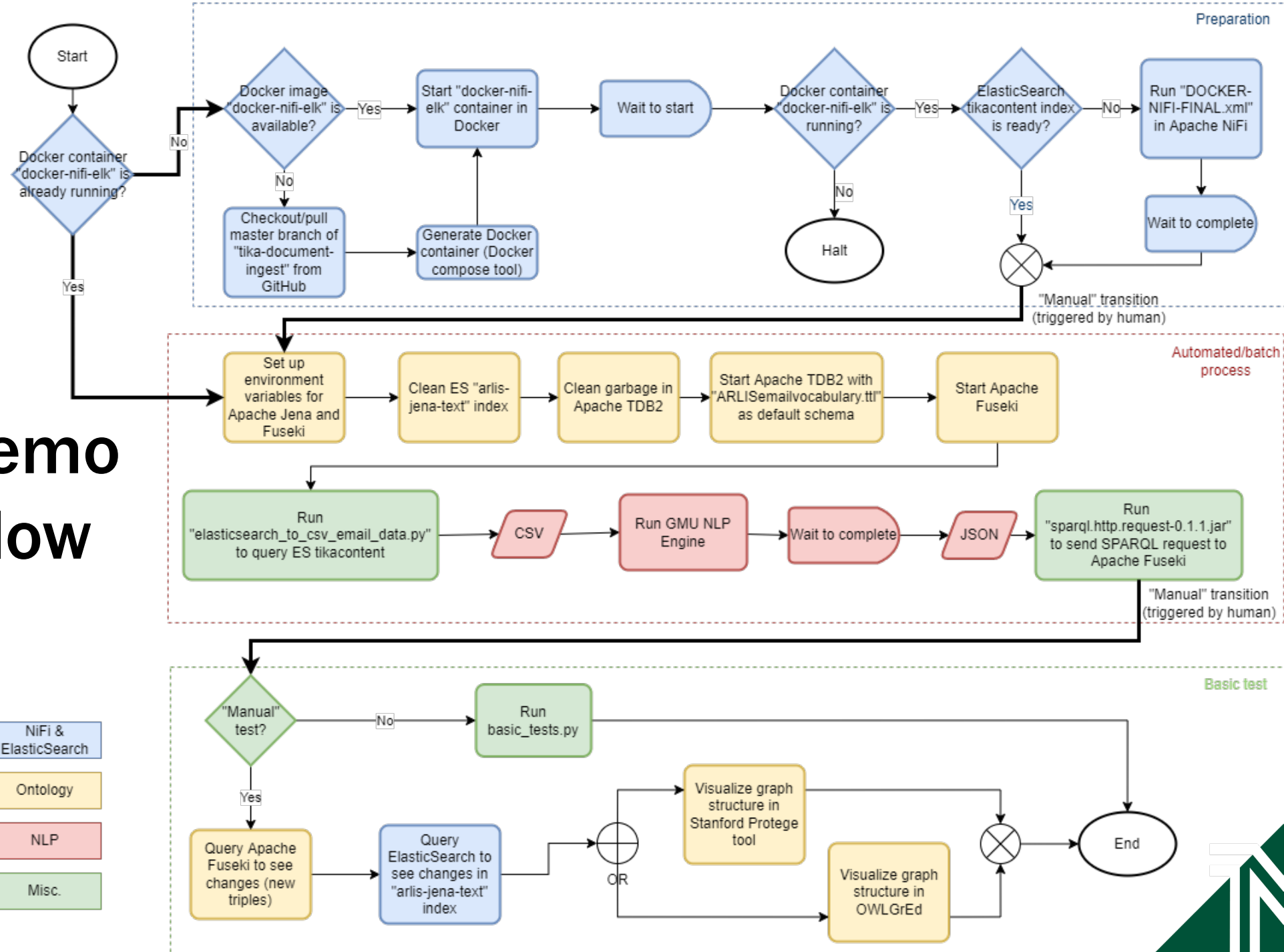
e.g.,
[B-DeliverMeet:Date, I-DeliverMeet:Date] gives a “Date” span

Deliver Meeting:
- Time: 2pm
- Date: this Thu
- Location: Room A

C5I Center



Prototype/Demo Execution Flow



NiFi & ElasticSearch

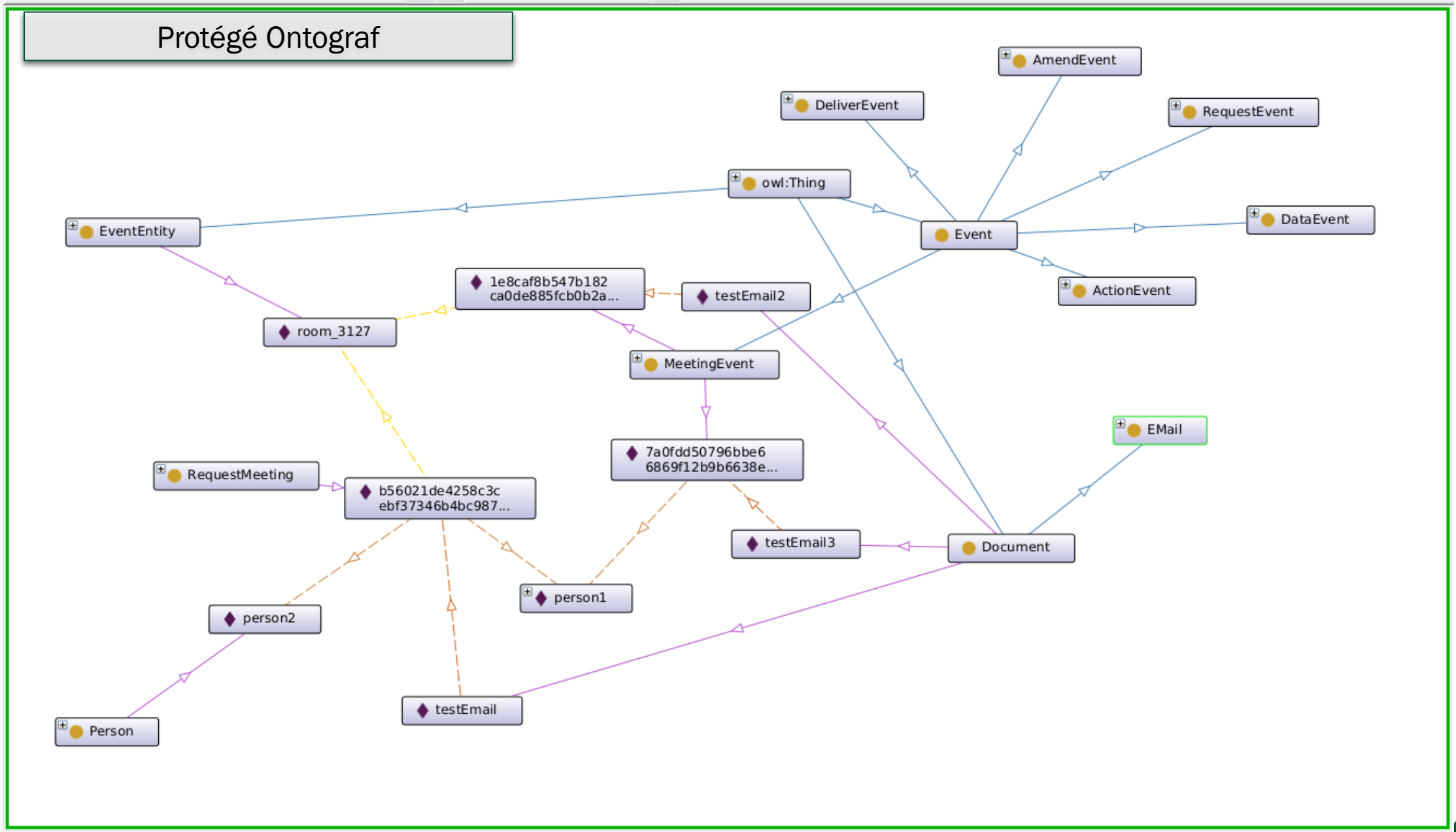
Ontology

NLP

Misc.

C5I Center

Prototype/demo implementation – visualization tools



Domain Adaptation of Event Extraction

- Research Objective: Enabling fast development of domain-specific KG from data
  - E.g., adapting from open domain to emails (our own dataset), AI Incident Database [1], COVID tweets [2], cybersecurity documents [3,4], BioNLP [5], scientific research [6]
  - Some topics to explore:
    - Pre-training, data augmentation, etc.
    - Uncertainty
      - Analyze and explore/exploit epistemic (ignorance, lack of data, low comprehensiveness) VS aleatoric (ubiquitous, natural in domain) uncertainty.
        - Outlier, out-of-distribution analysis, etc.
    - Identify events/information summaries to report as AI Incident

[1] https://incidentdatabase.ai/

[2]Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. Extracting a knowledge base ofcovid-19 events from social media. arXiv preprint arXiv:2006.02567, 2020.

[2] Taneeya Satyapanich, Francis Ferraro, and Tim Finin. Casie: Extracting cybersecurity event information from text. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 8749– 8757, 2020.

[3] Hieu Man Duc Trong, Duc-Trong Le, Amir Pouran Ben Veyseh, Thut Nguyn, and Thien Huu Nguyen. Introducing a new dataset for event detection in cybersecurity texts. In Proceedings of the 2020Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5381–5390, 2020.

[4] Jin-Dong Kim, Tomoko Ohta, Sampo Pyysalo, Yoshinobu Kano, and Jun'ichi Tsujii. Overview of bionlp'09 shared task on event extraction. In Proceedings of the BioNLP 2009 workshop companion volume for shared task, pages 1–9, 2009.

[5] Luan, Yi, Mari Ostendorf, and Hannaneh Hajishirzi. "Scientific Information Extraction with Semi-supervised Neural Tagging." Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017.

C5I Center

Cross-domain knowledge/data fusion

- Research Objective: Fusing the cross-domain knowledge/data for applications
  - Fusion across modalities for higher-level inference
    - Knowledge graph + Text (and more in the long term)
    - Applications: recommendation, question answering, search, etc.
  - Potential Application Domains with domain knowledge graphs
    - COVID/new virus discoveries
      - News articles reporting the unusual/novel pathogen activity (e.g., COVID cases)
      - CDC websites describing the disease symptoms and prevention
    - Emerging AI/ML Research
      - AI/ML papers and blogs
      - Social media discussion on AI/ML applications
    - Fact verification/misinformation detection
      - Web-pages, tables, tweets with miss information
  - Techniques:
    - Leveraging our domain adaptation techniques
    - With humans in the loop for reliability
    - Explainable AI techniques, probabilistic ontology/reasoning…
    - …

C5I Center