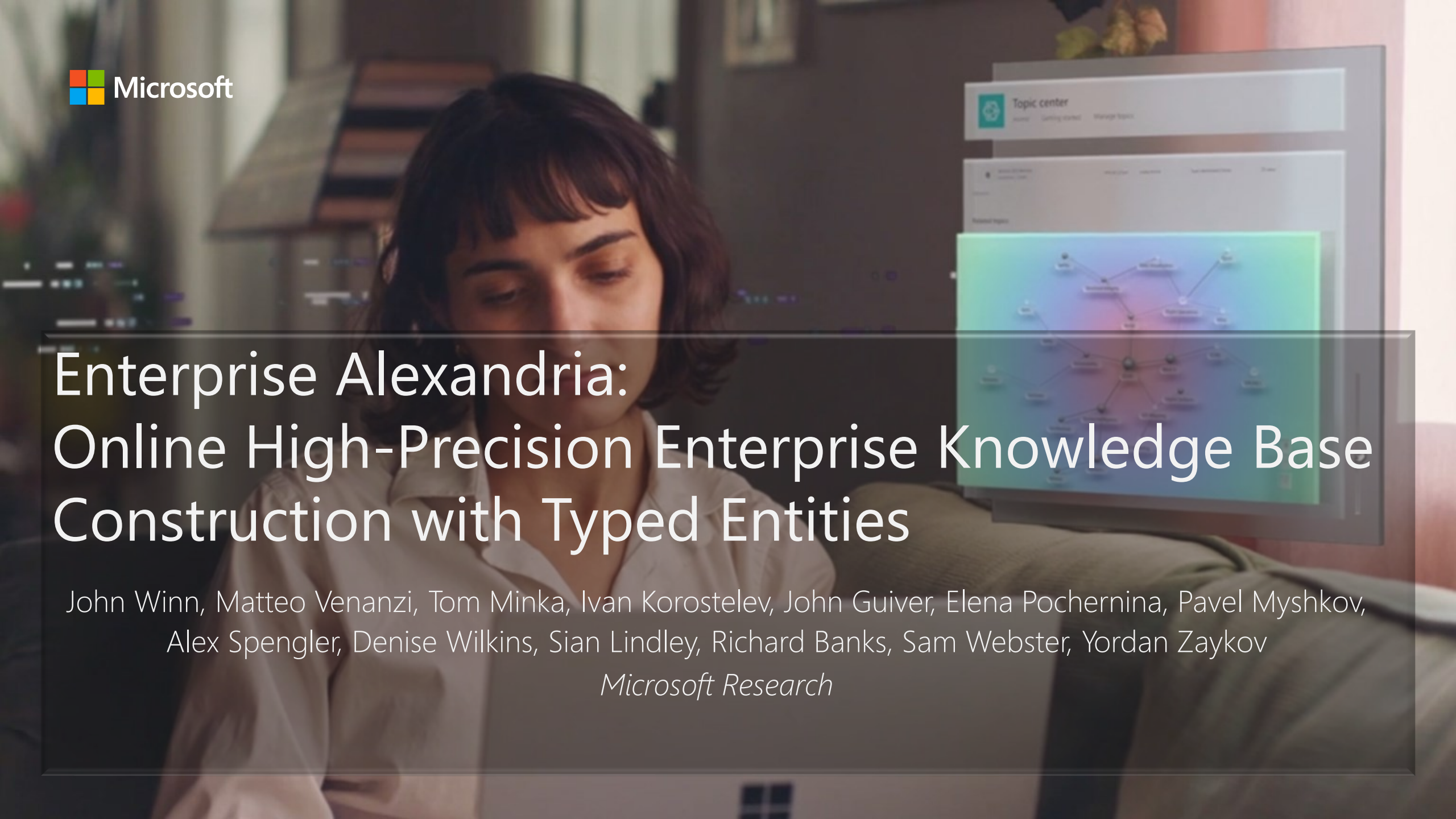


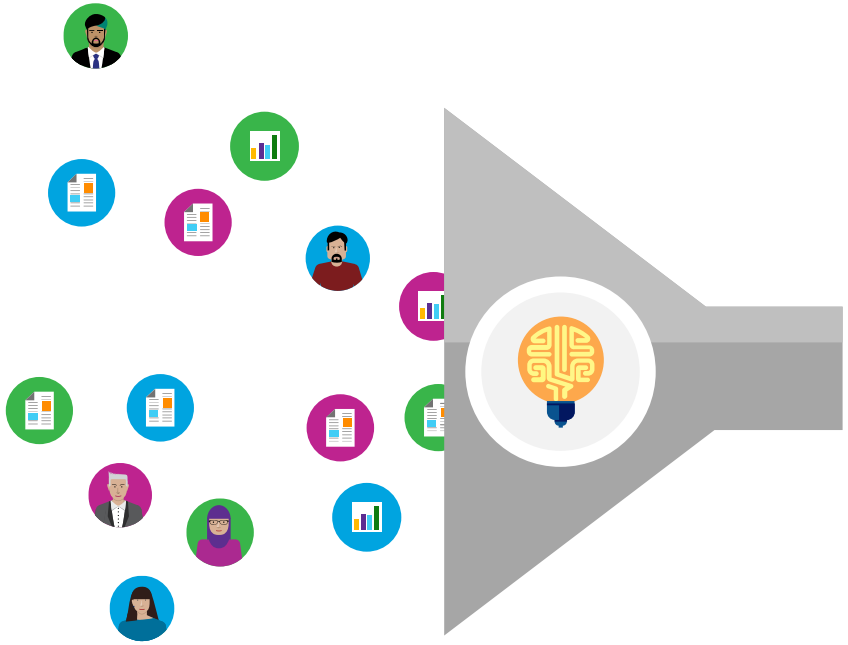
Microsoft

Enterprise Alexandria: Online High-Precision Enterprise Knowledge Base Construction with Typed Entities

John Winn, Matteo Venanzi, Tom Minka, Ivan Korostelev, John Guiver, Elena Pochernina, Pavel Myshkov, Alex Spengler, Denise Wilkins, Sian Lindley, Richard Banks, Sam Webster, Yordan Zaykov

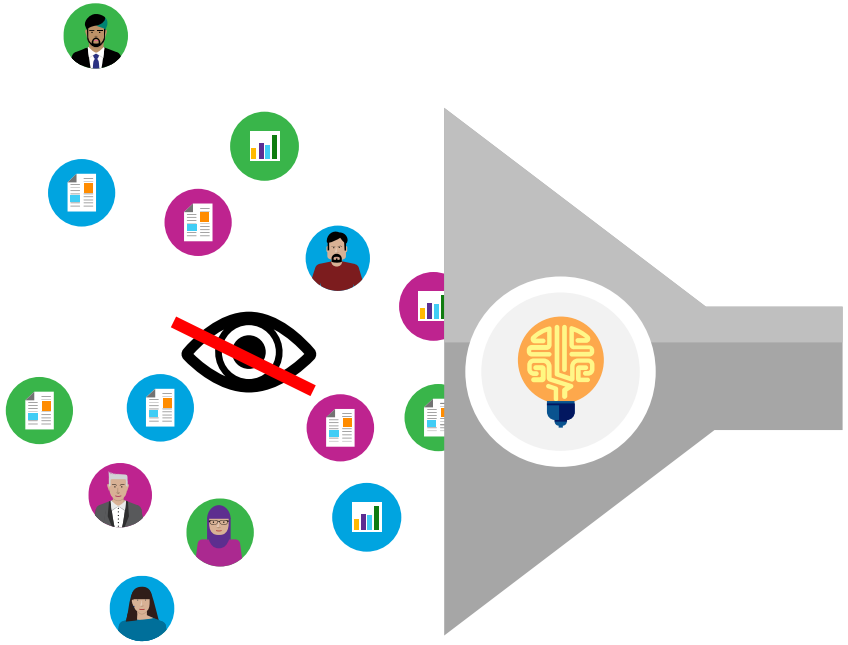
Microsoft Research

Automatic Knowledge Base Construction in the Enterprise



	‘Open Skies’	‘Live Faster’	‘Mini Adventure’
Client:	ACME Ltd.	ACME Ltd.	Coningsby
Lead:			
Sales Manager:			
Proposal:			
Budget:	$70,000	$150,000	$80,000
Pitch:			

Automatic Knowledge Base Construction in the Enterprise



	‘Open Skies’	‘Live Faster’	‘Mini Adventure’
Client:	ACME Ltd.	ACME Ltd.	Coningsby
Lead:			
Sales Manager:			
Proposal:			
Budget:	$70,000	$150,000	$80,000
Pitch:			

A Generative Model for Enterprise Documents

Name	Member	Type	Start da
Open Skies, OS, O-Skies, OpenSkies	Name: John Smith, Email: jsmith@contoso.com	Project, Team	Day: 1 Month: Year: 2(

Open Skies | J. Smith | Project

{Member} work on the {Name} {Type}

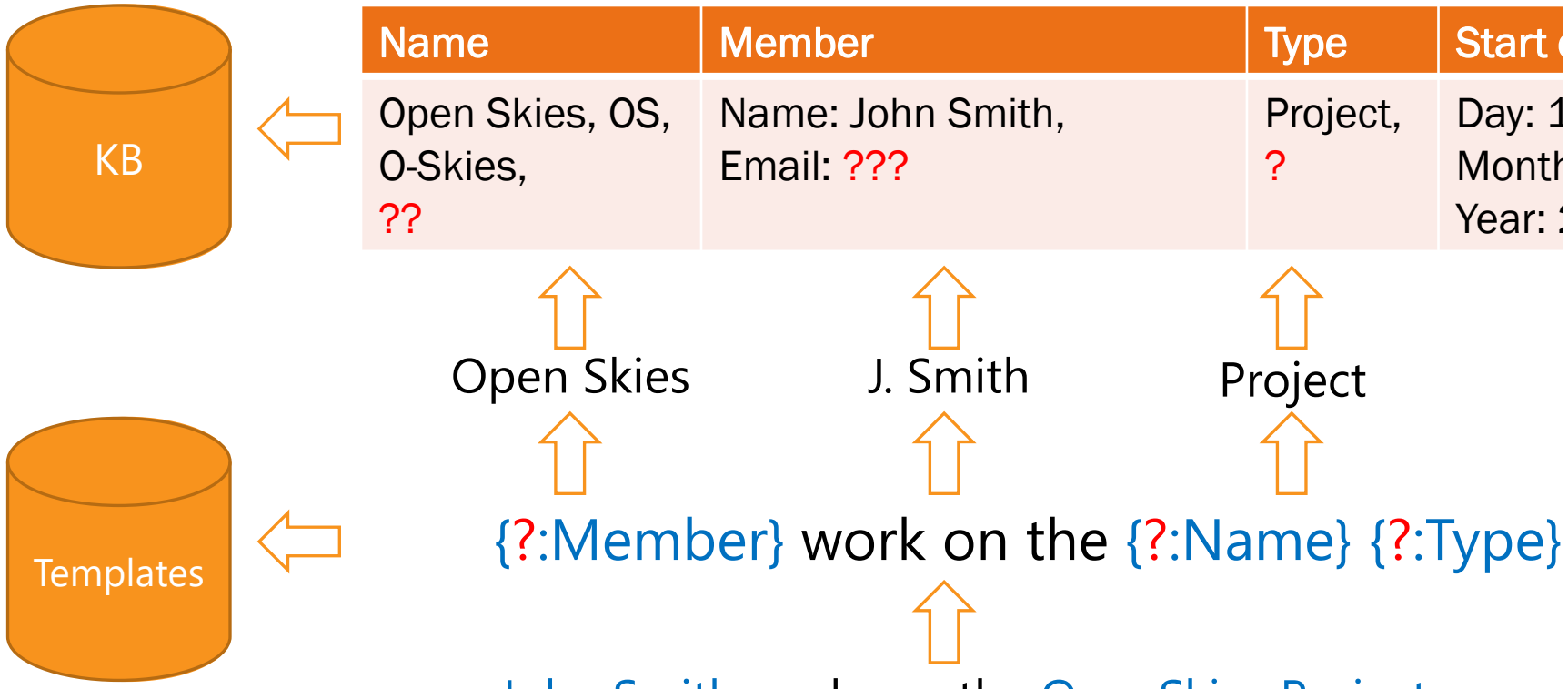
John Smith works on the OpenSkies Project

Email: The OS team, led by Mark J. and Eric W., is going to work with various partners to deliver…

Document: Mark Jones is a founding member of O-Skies, a project started in April 2019…

Meeting: You are invited to attend the quarterly review for Project "Live Faster" …

A Generative Model for Enterprise Documents



Email: The OS team, led by Mark J. and Eric W., is going to work with various partners to deliver…

Document: Mark Jones is a founding member of O-Skies, a project started in April 2019…

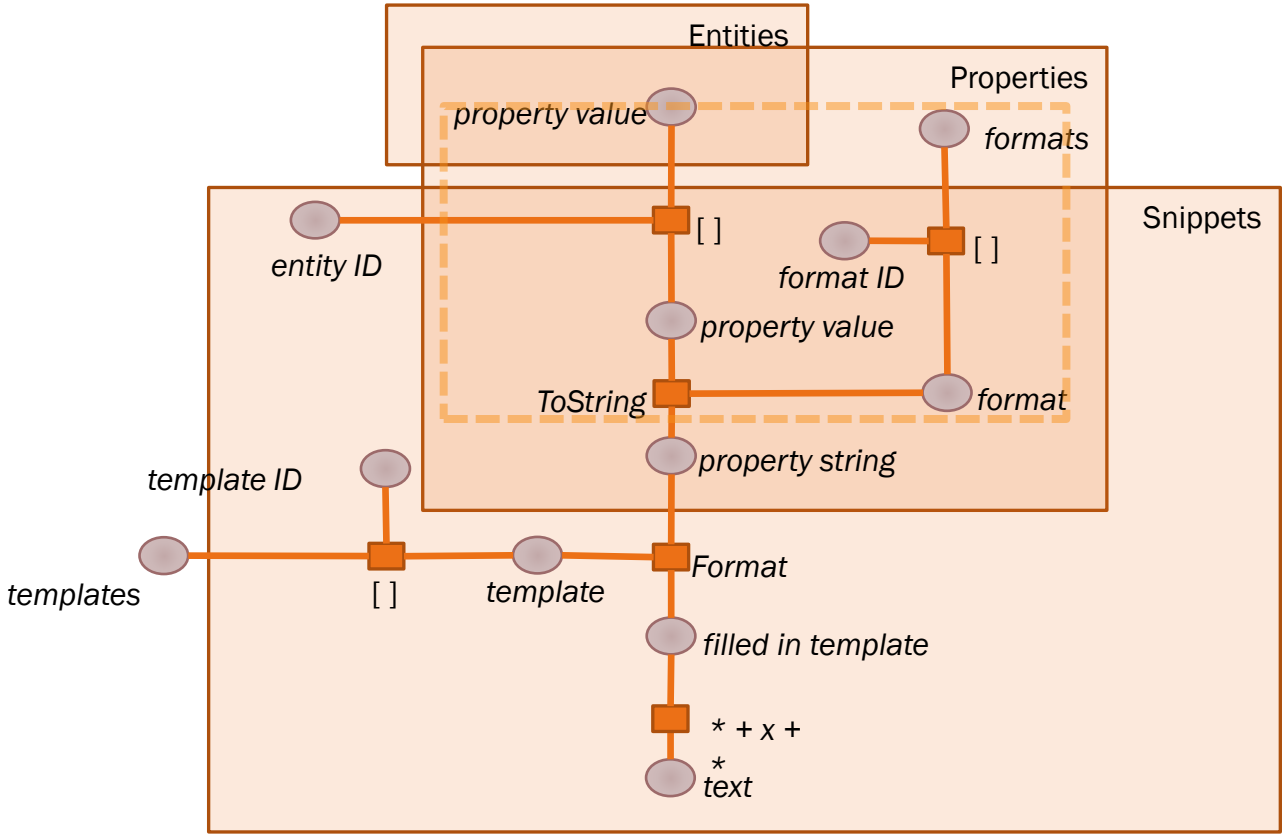
Meeting: You are invited to attend the quarterly review for Project "Live Faster" …

⋮ ⋮ ⋮ ⋮

Enterprise Alexandria: KB to Structured Text

Enterprise Alexandria is built upon the core probabilistic model of "Alexandria" (Winn et al, AKBC 2019)

- *Knowledge base* contains entities
  - Each entity has a type
- *Entity Type*
  - E.g. Person, Company, Event
  - Defined by schema
- *Schema*: Set of typed properties
  - E.g. Name, Height, DoB, …
- *Property* defined by
  - Alexandria Type + Prior
- *Alexandria Type*
  - Value type
  - Distribution type
  - Formats
  - ToString factor



Enterprise Alexandria: Key Contributions

1) Structured Templates for modelling Document Metadata

Entity
- **Name**: "Tokyo"
- **Types**: Project, System
- **Members**: Alice, Bob, John

Latent KB Entity

+

CalendarEvent
- **Name**: "{Name} {Types} sync"
- **Organizers**: Subset(Members)
- **Attendees**: Subset(Members)
- **IsRecurring**: True

Structured Template

=

CalendarEvent
- **Name**: "Tokyo project sync"
- **Organizers**: John
- **Attendees**: Alice, Bob
- **IsRecurring**: True

Generated Data Item

Enterprise Alexandria: Key Contributions

1) Structured Templates for modelling Document Metadata

Entity
- **Name**: "Tokyo"
- **Types**: Project, System
- **Members**: Alice, Bob, John

Latent KB Entity

+

CalendarEvent
- **Name**: "{Name} {Types} sync"
- **Organizers**: Subset(Members)
- **Attendees**: Subset(Members)
- **IsRecurring**: True

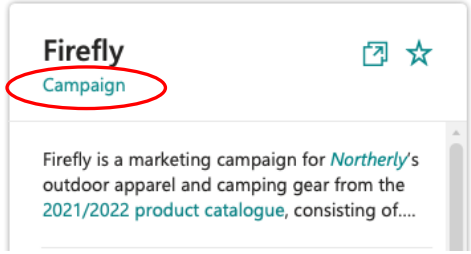
Structured Template

=

CalendarEvent
- **Name**: "Tokyo project sync"
- **Organizers**: John
- **Attendees**: Alice, Bob
- **IsRecurring**: True

Generated Data Item

2) Type Discovery



Enterprise Alexandria: Key Contributions

1) Structured Templates for modelling Document Metadata

Entity
- **Name**: "Tokyo"
- **Types**: Project, System
- **Members**: Alice, Bob, John

Latent KB Entity

+

CalendarEvent
- **Name**: "{Name} {Types} sync"
- **Organizers**: Subset(Members)
- **Attendees**: Subset(Members)
- **IsRecurring**: True

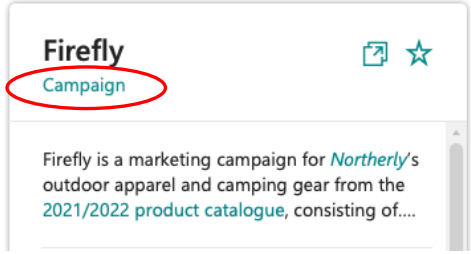
Structured Template

=

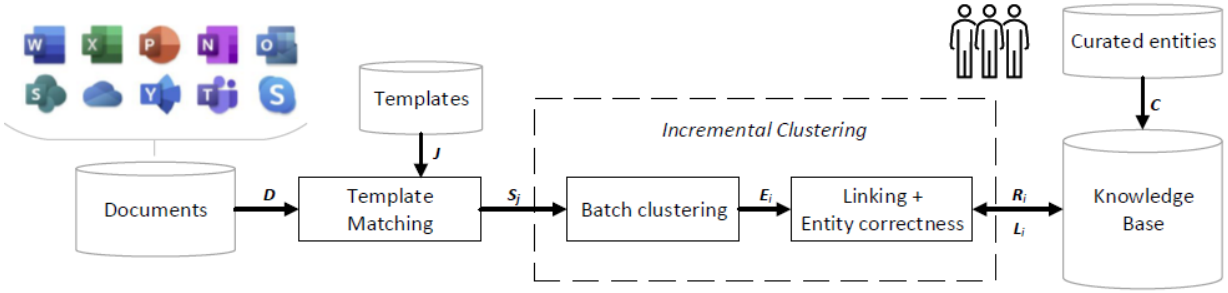
CalendarEvent
- **Name**: "Tokyo project sync"
- **Organizers**: John
- **Attendees**: Alice, Bob
- **IsRecurring**: True

Generated Data Item

2) Type Discovery



3) Incremental Clustering with Human Curation



Enterprise Alexandria: Key Contributions

1) Structured Templates for modelling Document Metadata

Entity
- **Name**: "Tokyo"
- **Types**: Project, System
- **Members**: Alice, Bob, John

Latent KB Entity

+

CalendarEvent
- **Name**: "{Name} {Types} sync"
- **Organizers**: Subset(Members)
- **Attendees**: Subset(Members)
- **IsRecurring**: True

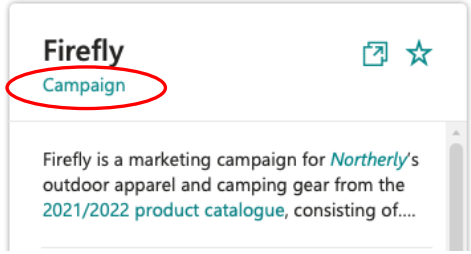
Structured Template

=

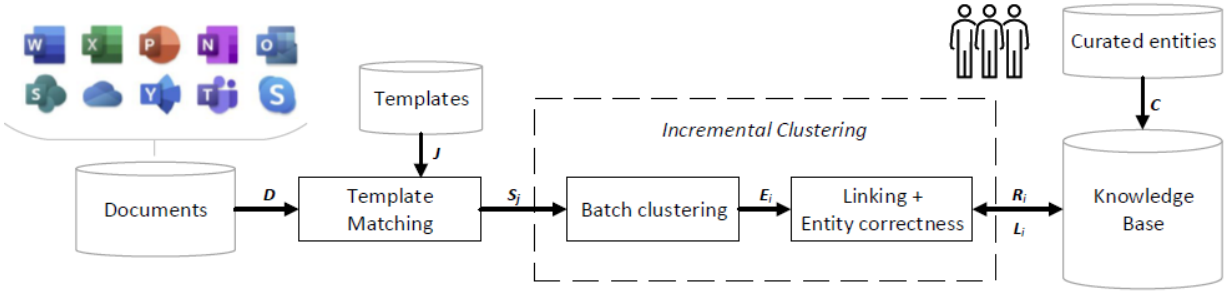
CalendarEvent
- **Name**: "Tokyo project sync"
- **Organizers**: John
- **Attendees**: Alice, Bob
- **IsRecurring**: True

Generated Data Item

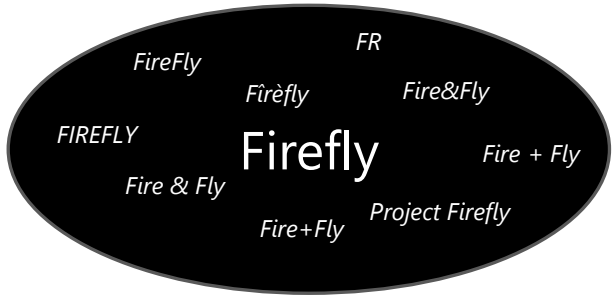
2) Type Discovery



3) Incremental Clustering with Human Curation



4) Entity Names and Variants



Experimental Results

- Significantly higher precision on the “Enron” and “Microsoft” dataset compared to neural baselines.

	Enron				*Microsoft Eyes-on*			
Method	Entities	Types	Prec.	Rel. Cov.	Entities	Types	Prec.	Rel. Cov.
BERT F	2,764	3	0.08 ± 0.04	1.33 ± 0.03	990	3	0.44	0.33
BERT MF	5,789	3	0.13 ± 0.03	7.86 ± 0.04	2,610	3	0.39	0.77
BERT LF	2,296	16	0.18 ± 0.04	4.31 ± 0.04	867	16	0.58	0.38
RoBERTa LF	5,988	3	0.13 ± 0.03	8.13 ± 0.04	2,542	3	0.49	0.94
$\text{EA}_{\text{size}>100}$	449	**152**	0.37 ± 0.05	1.73	1,029	**141**	0.71	0.55
EA	104	**152**	$\mathbf{0.92} \pm 0.03$	1	1,591	**141**	**0.83**	1

- On large-scale runs, it discovers 675k entities from 1M documents on a standard machine.
- Automatically learns types for each entity.

Enron

Microsoft Eyes-on

Experimental Results

Now in Microsoft Viva Topics

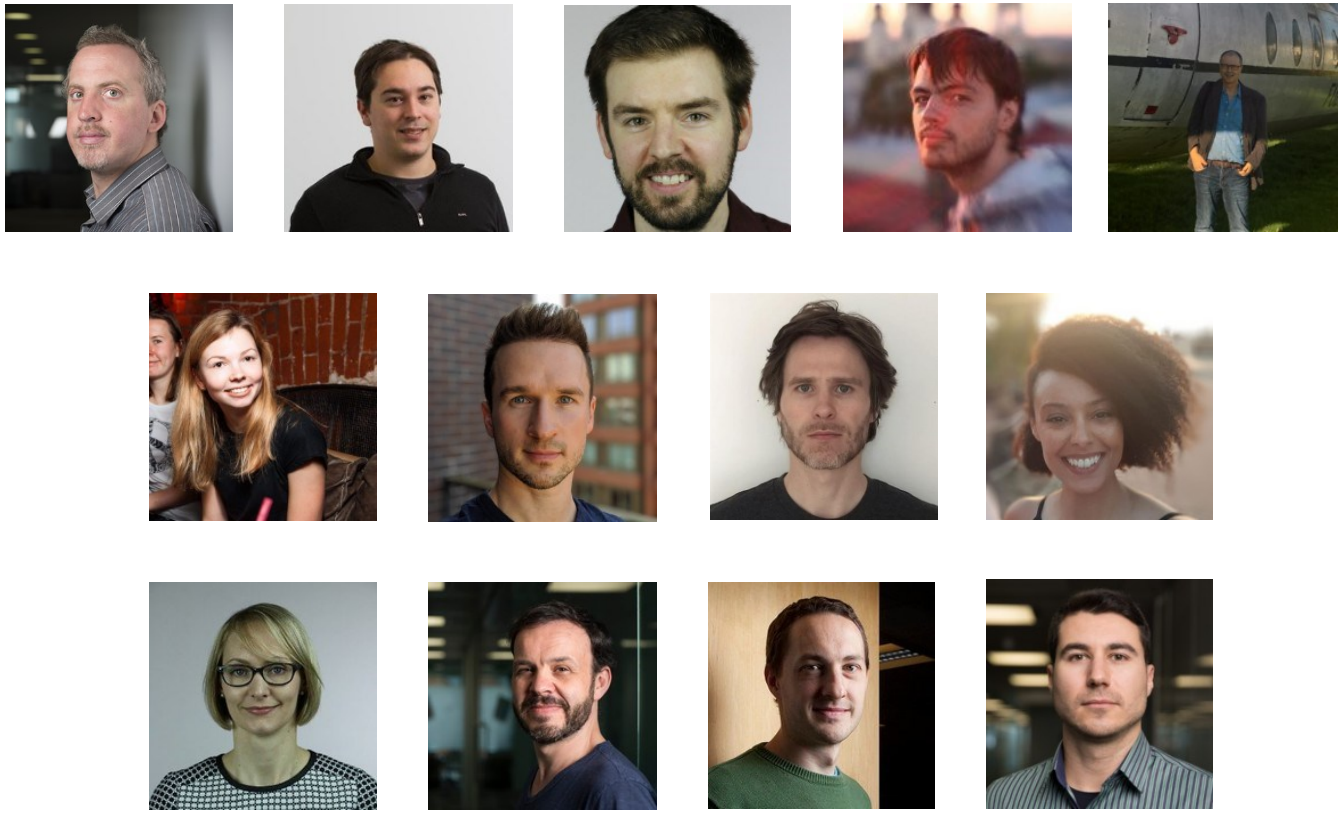
- Significantly higher precision on the “Enron” and “Microsoft” dataset compared to neural baselines.

	Enron				*Microsoft Eyes-on*			
Method	Entities	Types	Prec.	Rel. Cov.	Entities	Types	Prec.	Rel. Cov.
BERT F	2,764	3	0.08 ± 0.04	1.33 ± 0.03	990	3	0.44	0.33
BERT MF	5,789	3	0.13 ± 0.03	7.86 ± 0.04	2,610	3	0.39	0.77
BERT LF	2,296	16	0.18 ± 0.04	4.31 ± 0.04	867	16	0.58	0.38
RoBERTa LF	5,988	3	0.13 ± 0.03	8.13 ± 0.04	2,542	3	0.49	0.94
$EA_{size>100}$	449	**152**	0.37 ± 0.05	1.73	1,029	**141**	0.71	0.55
EA	104	**152**	$\mathbf{0.92} \pm 0.03$	1	1,591	**141**	**0.83**	1

- On large-scale runs, it discovers 675k entities from 1M documents on a standard machine.
- Automatically learns types for each entity.

Enron

Microsoft Eyes-on



Thanks