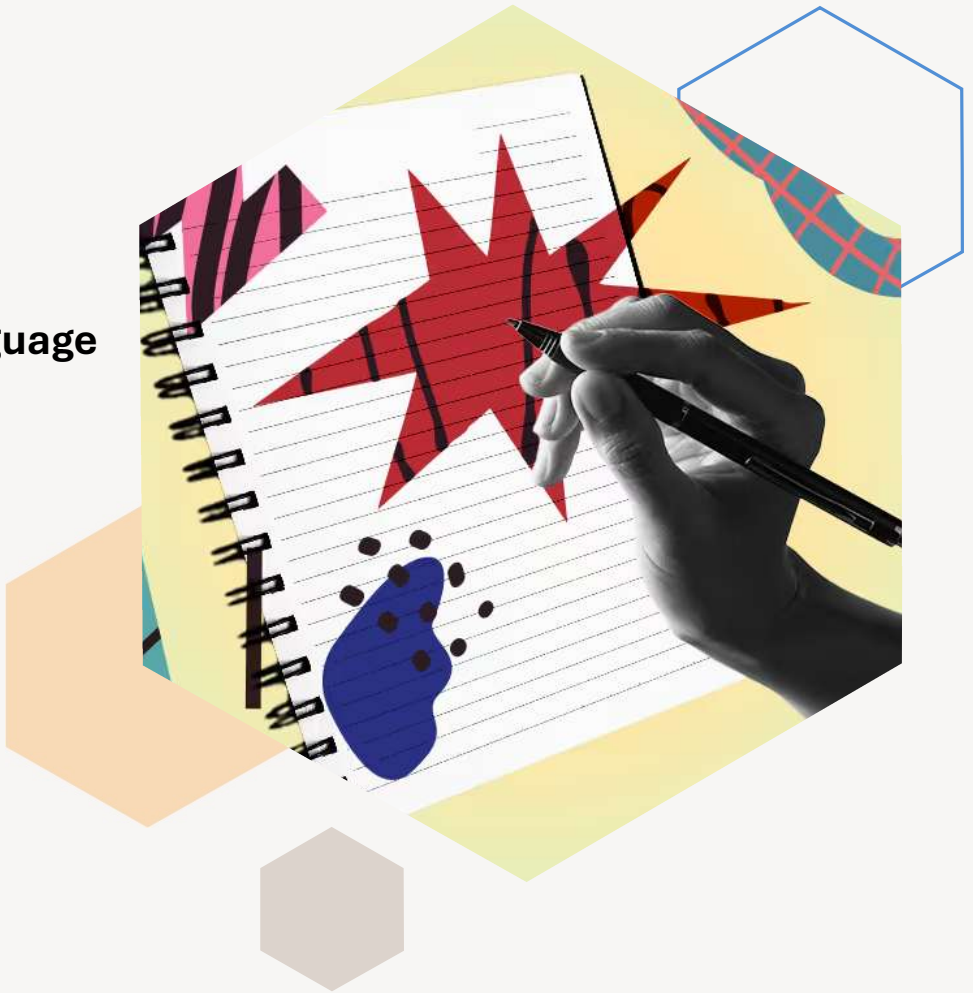
WEEK 9

Detecting & Defining Specialized Language in Text

Dr. Santu Karmaker, Dr. Song Wang, Dr. Mubarak Shah, Adam Bawatneh

Valerie
Lopez

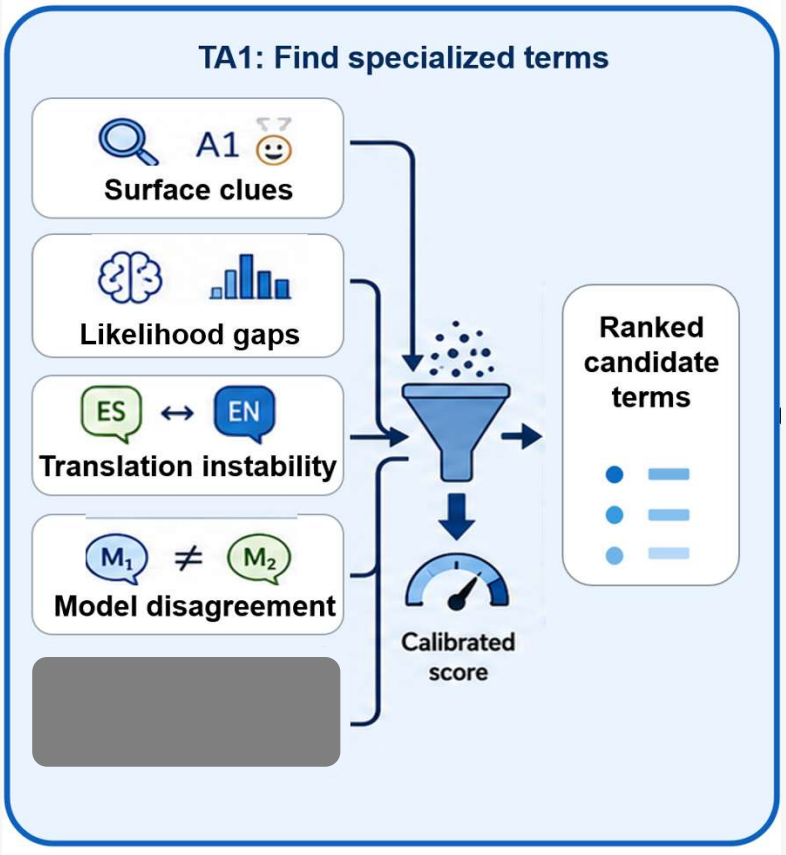


Project Recap

- Identify **specialized language** from text
- Apply **4** detection **methods**

This Week:

- Compared MT versus MT-first pipeline
- Experimented on multiple datasets



MT Recap

MT-First Pipeline

1. Run MT-only word detection.
2. Build consecutive candidates as phrases.
3. Run SC, LG, MA on proposed MT word and phrase candidates.
    1. Each method applies phrase picker.
4. Apply Logistic Fusion and final phrase picker.

Experiment

- Last week's results demonstrated MT-first pipeline does not improve greatly on MT-only
- We ran the full MT-first pipeline on various datasets and compared initial versus final results
- **Do we need the full pipeline**, or is MT-only enough?

Datasets

- GenZ-A – English Internet slang
- GenZ-B – English Internet slang and acronyms
- LM-Lex – English Internet slang
- MedSB – English Medical slang
- CryptoS – English Cryptocurrency slang
- SilentSig – English **dogwhistle dataset**, similar to code words
- PolS – Polish slang
- SDU-DA –Danish acronyms
- SDU-PE – Persian acronyms
- SDU-VI – Vietnamese acronyms

Set Up

- Datasets were normalized into 1 schema

sentence	gold_terms	term_type	source_dataset	is_single_word	is_multiword

- Driver and child slurms were built to run any normalized dataset through the entire pipeline
- Debugging to allow MT to run on multilingual datasets

Metrics

Precision

	GenZ-A	GenZ-B	LM-Lex	MedSB	CryptoS	SilentSig	PolS	SDU-DA	SDU-PE	SDU-VI
MT-Only	0.662	0.215	0.138	0.148	0.170	0.019	0.018	0.004	0.010	0.030
MT-First	**0.785**	**0.797**	**0.529**	**0.446**	**0.560**	**0.173**	**0.150**	**0.090**	**0.919**	**0.427**

Recall

	GenZ-A	GenZ-B	LM-Lex	MedSB	CryptoS	SilentSig	PolS	SDU-DA	SDU-PE	SDU-VI
MT-Only	**0.863**	**0.579**	**0.773**	**0.815**	**0.582**	**0.348**	**0.626**	**0.206**	**0.234**	**0.202**
MT-First	0.827	0.513	0.624	0.467	0.236	0.239	0.145	0.083	0.127	0.157

F1

	GenZ-A	GenZ-B	LM-Lex	MedSB	CryptoS	SilentSig	PolS	SDU-DA	SDU-PE	SDU-VI
MT-Only	0.749	0.314	0.234	0.250	0.264	0.037	0.035	0.009	0.019	0.053
MT-First	**0.805**	**0.624**	**0.573**	**0.456**	**0.332**	**0.201**	**0.148**	**0.086**	**0.224**	**0.229**

Metrics Observations

- Precision improves greatly from the pipeline in all datasets
- Recall and F1 are limited to MT initial performance
- MT-first pipeline cannot help if initial recall is poor

- Our approach performs best with English Slang datasets
- MT recall drops greatly with multilingual datasets

Precision

	GenZ-A	GenZ-B	LM-Lex	MedSB	CryptoS	SilentSig	PolS	SDU-DA	SDU-PE	SDU-VI
MT-Only	0.662	0.215	0.138	0.148	0.170	0.019	0.018	0.004	0.010	0.030
MT-First	**0.785**	0.797	**0.529**	0.446	**0.560**	0.173	0.150	0.090	**0.919**	**0.427**
Llama	0.684	0.907	0.392	0.452	0.369	**0.196**	**0.268**	**0.094**	0.312	0.307
Qwen	0.607	0.911	0.452	**0.642**	0.417	0.166	0.236	0.052	0.235	0.237
Mistral	0.624	**0.922**	0.433	0.473	0.442	0.164	0.188	0.069	0.319	0.228

- MT-first pipeline consistently performs at or above the level of an LLM

Recall

	GenZ-A	GenZ-B	LM-Lex	MedSB	CryptoS	SilentSig	PolS	SDU-DA	SDU-PE	SDU-VI
MT-Only	**0.863**	0.579	**0.773**	**0.815**	**0.582**	0.348	**0.626**	**0.206**	0.234	0.202
MT-First	0.827	0.513	0.624	0.467	0.236	0.239	0.145	0.083	0.127	0.157
Llama	0.598	**0.933**	0.613	0.643	0.522	**0.493**	0.399	0.122	0.596	**0.632**
Qwen	0.442	0.832	0.621	0.684	0.466	0.405	0.396	0.085	0.579	0.526
Mistral	0.458	0.915	0.697	0.727	0.517	0.492	0.439	0.096	**0.607**	0.537

- LLMs perform well on internet slang datasets due to previous exposure and training
- MT-first pipeline can never perform better than MT-only, resulting in recall drops

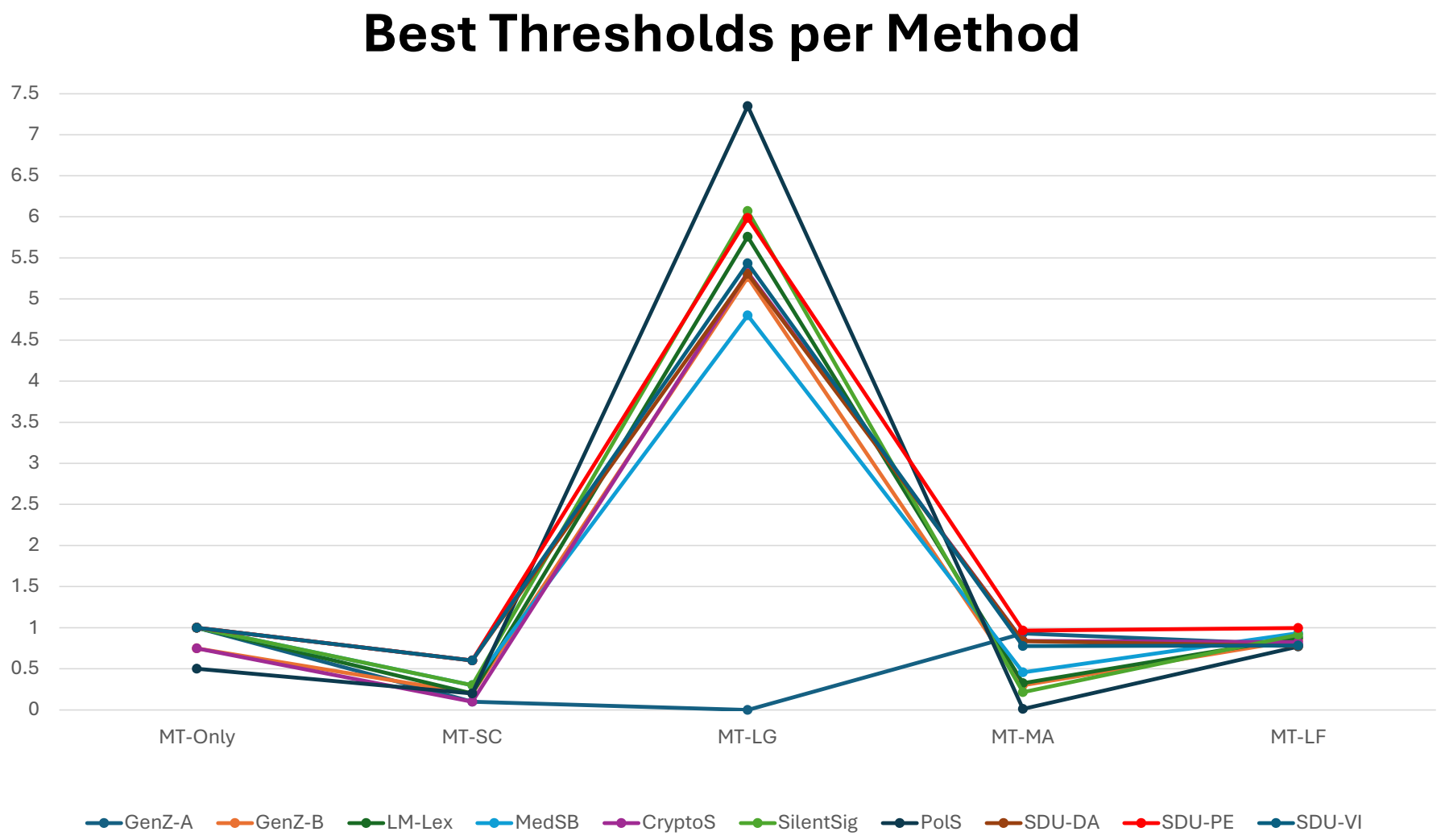
F1-Score

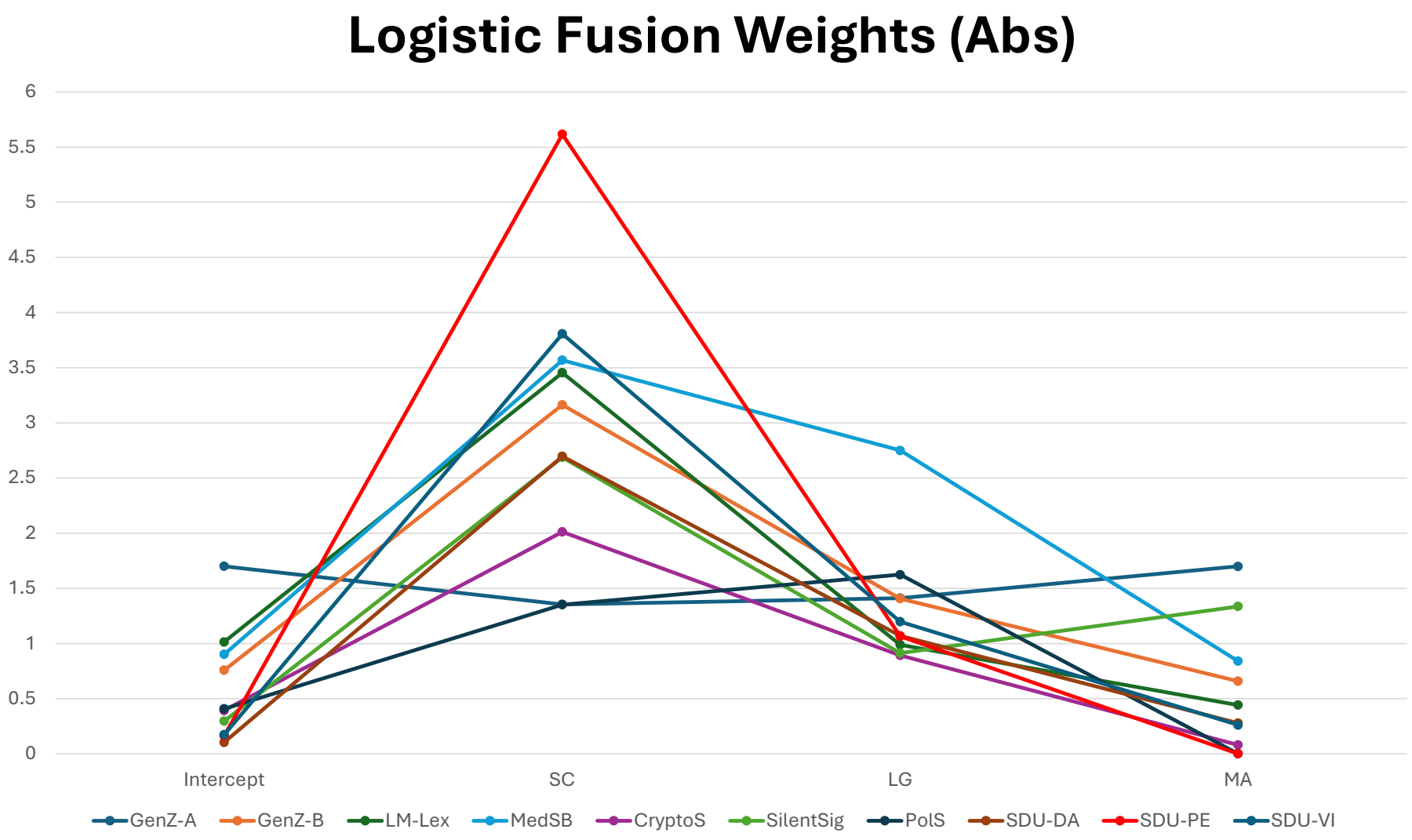
	GenZ-A	GenZ-B	LM-Lex	MedSB	CryptoS	SilentSig	PolS	SDU-DA	SDU-PE	SDU-VI
MT-Only	0.749	0.314	0.234	0.250	0.264	0.037	0.035	0.009	0.019	0.053
MT-First	**0.805**	0.624	**0.573**	0.456	0.332	0.201	0.148	0.086	0.224	0.229
Llama	0.638	**0.920**	0.478	0.531	0.432	**0.281**	**0.321**	**0.106**	0.410	**0.414**
Qwen	0.512	0.870	0.524	**0.662**	0.440	0.236	0.295	0.065	0.334	0.326
Mistral	0.528	0.918	0.534	0.573	**0.476**	0.246	0.264	0.080	**0.418**	0.320

- Recall limitations hold back F1 scores for MT-first pipeline

Runtime Notes:

- MT-First complete pipeline runtime depends on initial dataset size and MT-only results
- MA often responsible for slow runtime
- LLM runtimes span from 10 to 45 min


Best Thresholds per Method
7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
MT-Only
MT-SC
MT-LG
MT-MA
MT-LF
GenZ-A
GenZ-B
LM-Lex
MedSB
CryptoS
SilentSig
PolS
SDU-DA
SDU-PE
SDU-VI


Logistic Fusion Weights (Abs)
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
Intercept
SC
LG
MA
GenZ-A
GenZ-B
LM-Lex
MedSB
CryptoS
SilentSig
PolS
SDU-DA
SDU-PE
SDU-VI

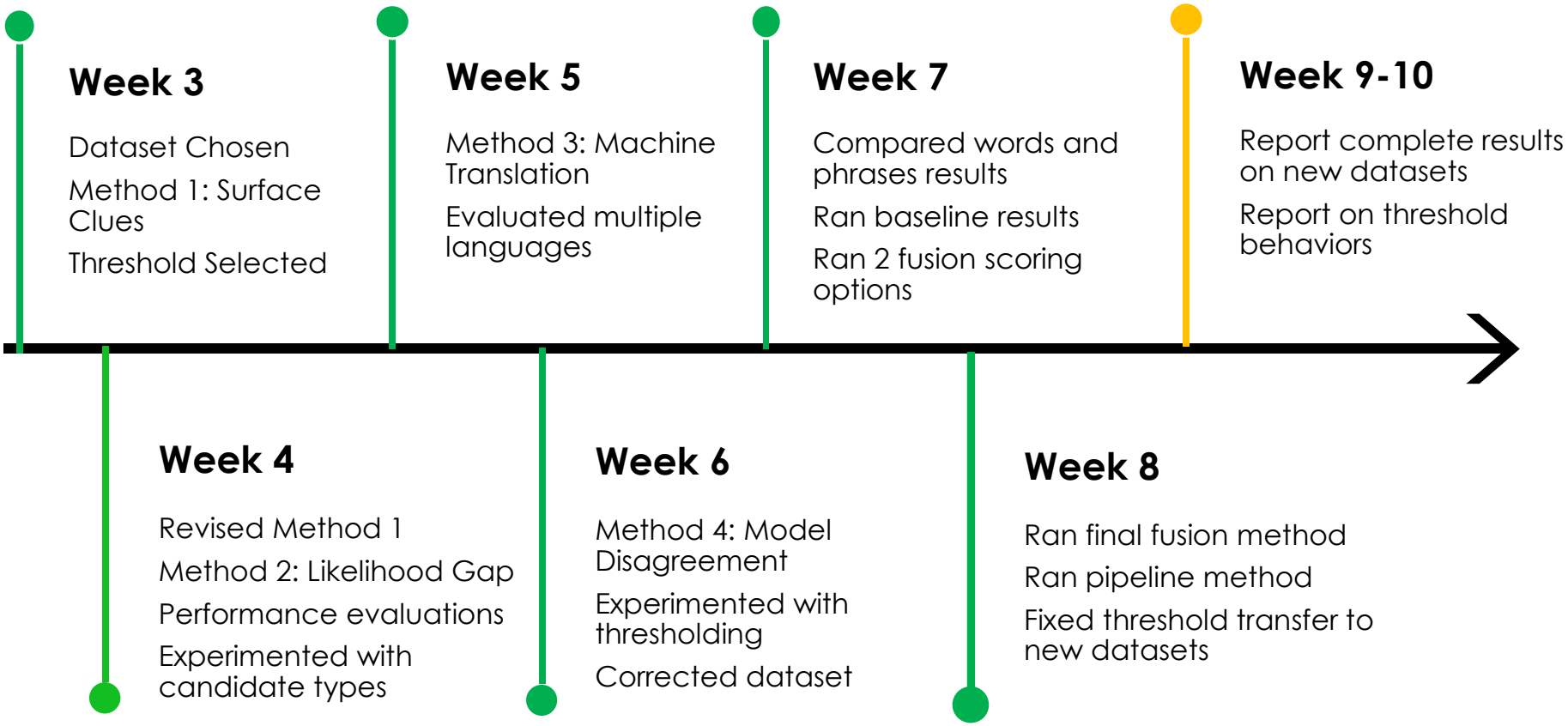
Method Weights Observations

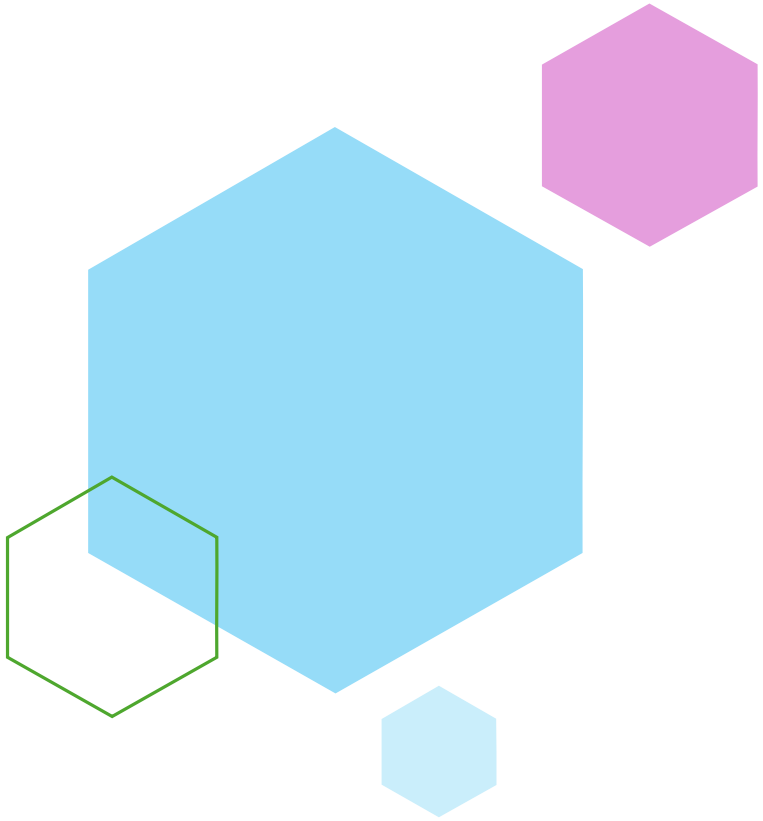
- SC is given more weight compared to LG and MA, indicating it is reliable and precise
- MA shows inconsistent pattern, sometimes becoming inconclusive data for the pipeline

Next Steps

- ❑ Continue writing REU report and poster
- ❑ Compare pipeline results with the removal of MA
- ❑ Run further studies on SilentSignals dataset

Updated Timeline





Any Questions?

Valerie Lopez