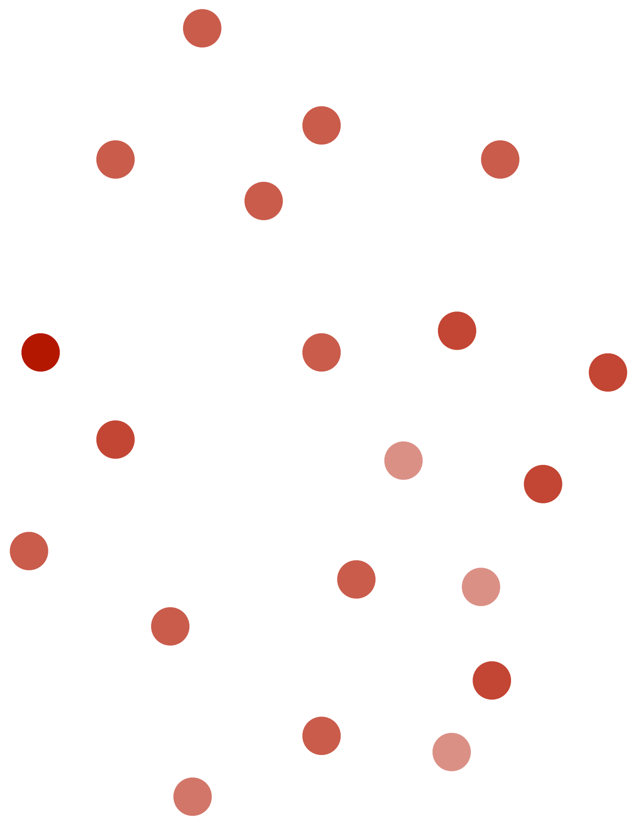
Genre as Weak Supervision for Cross-lingual Dependency Parsing

Max Müller-Eberstein, Rob van der Goot and Barbara Plank

EMNLP 2021

IT-UNIVERSITETET I KØBENHAVN / north

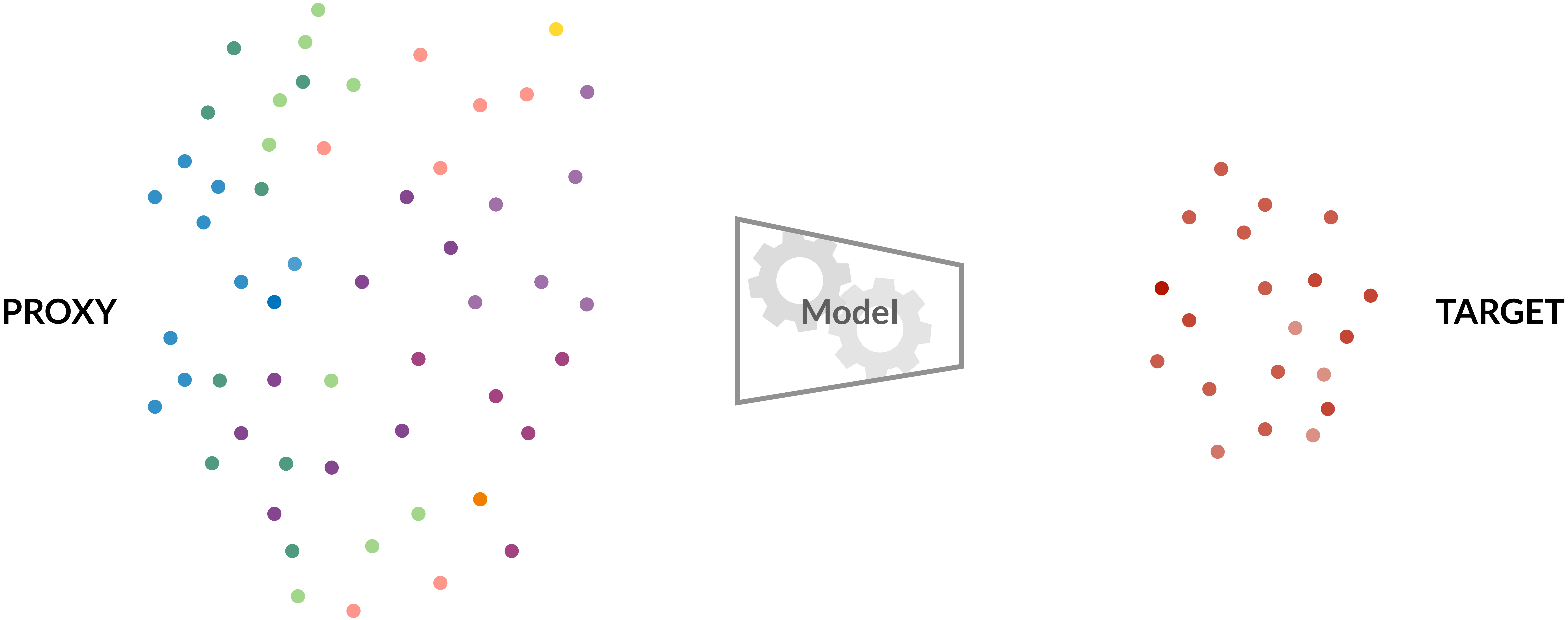


TARGET

unannotated

language

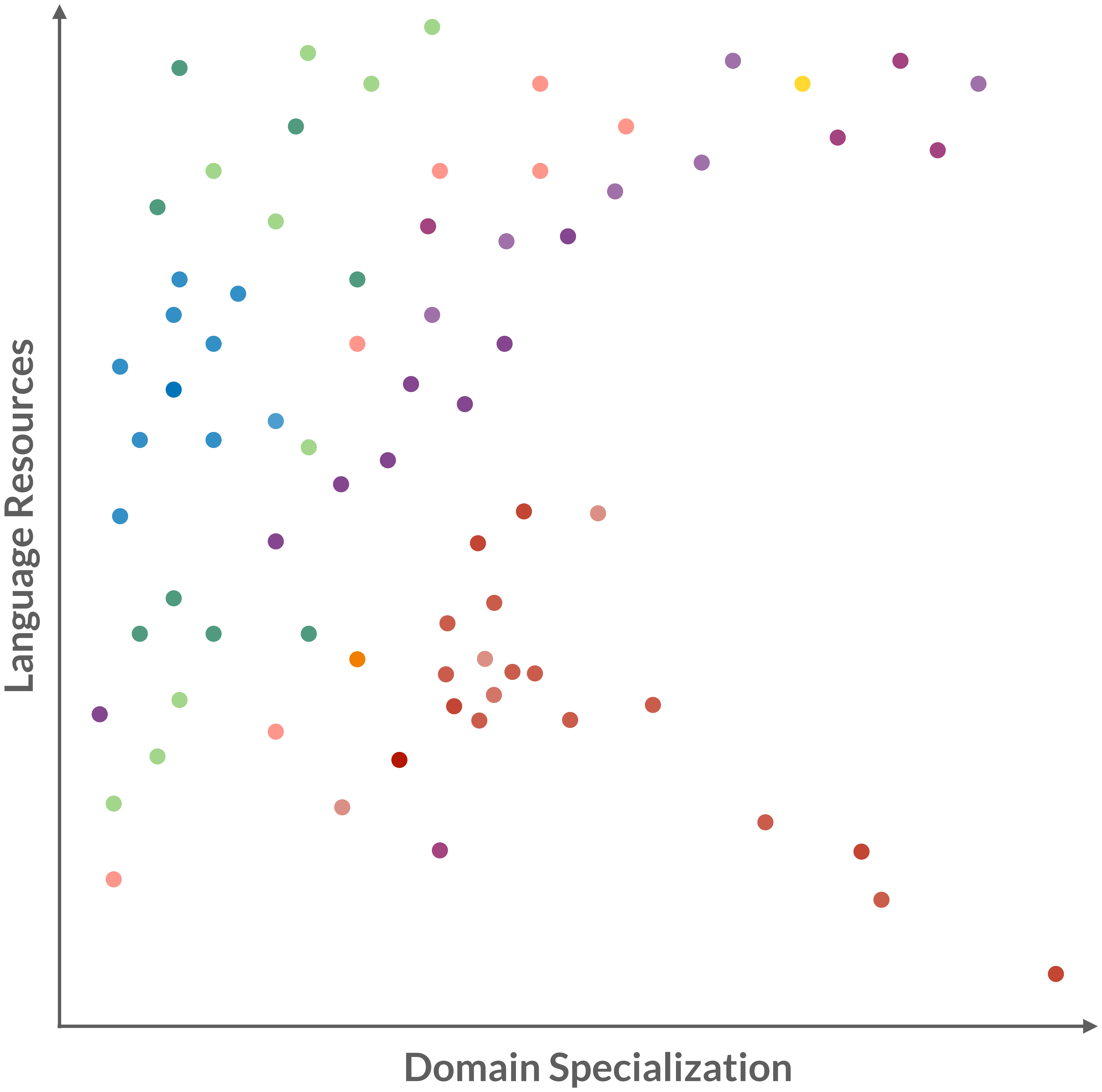
domain

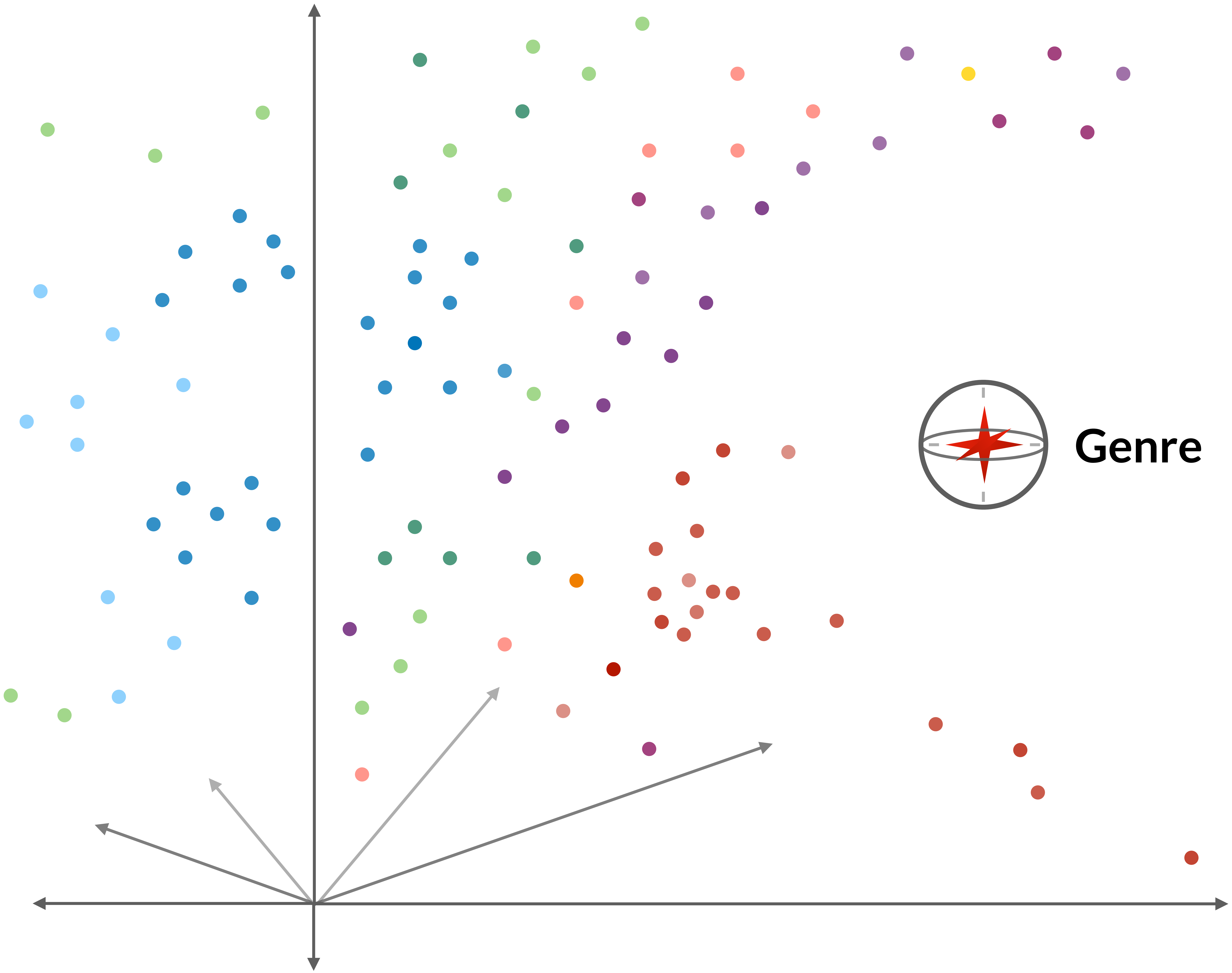


PROXY

Model

TARGET


Language Resources
Domain Specialization

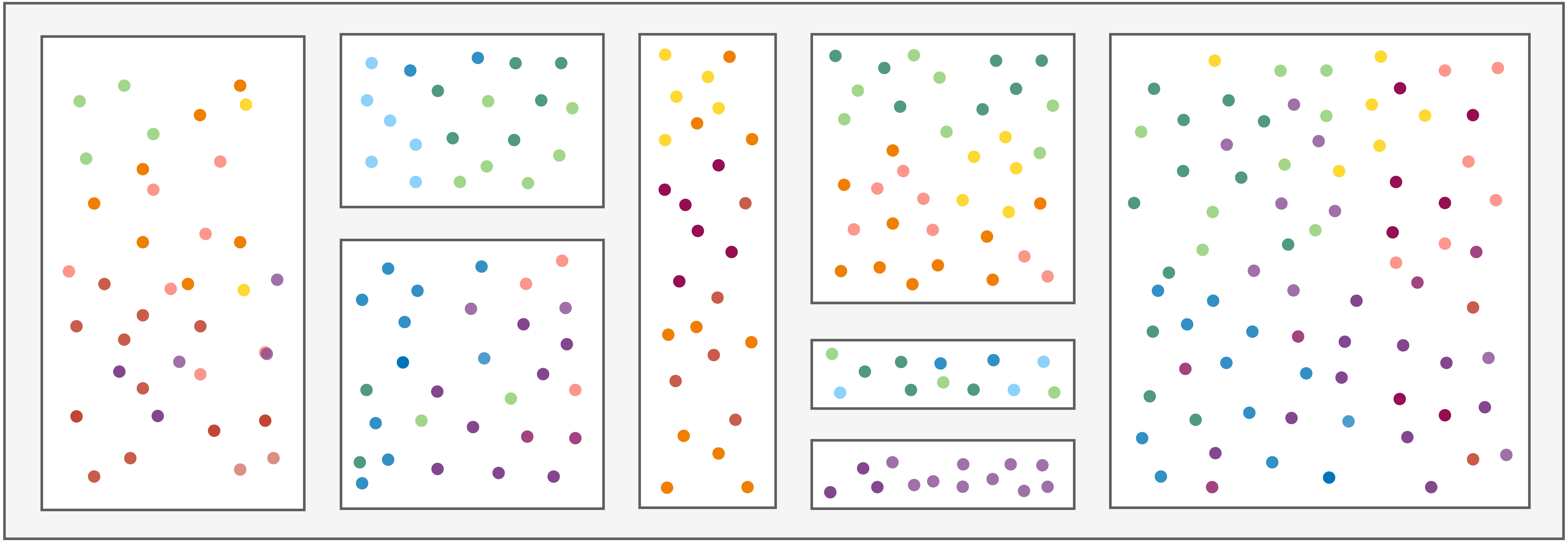


Genre

> *If our goal is to develop a parser for an* ***unseen language*** *with a* ***known domain****, can a signal such as* ***genre*** *guide our selection of cross-lingual proxy training data?*

Universal Dependencies v2.7

Zeman et al., 2020



177 TREEBANKS

104 LANGUAGES

1.38M SENTENCES

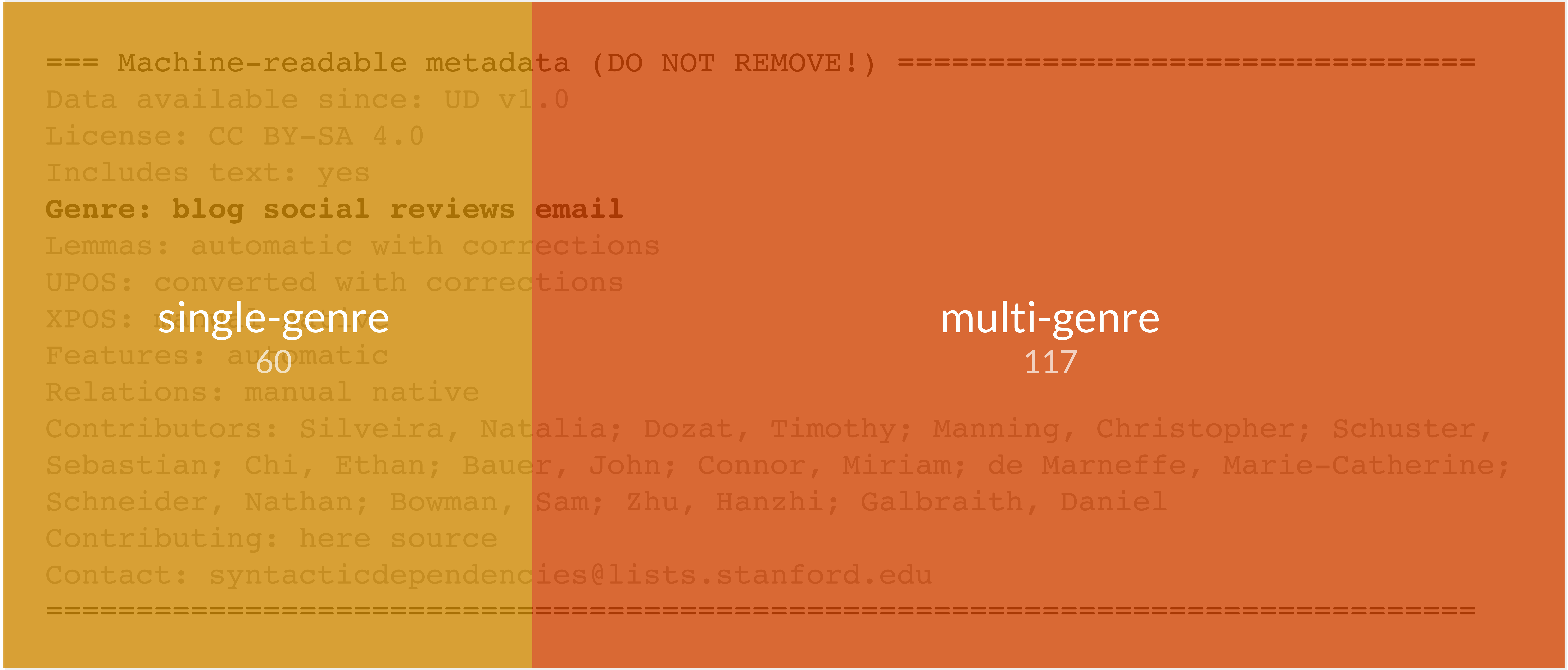
Genre as Weak Supervision

Domain **Genre** Register

Kessler et al. (1997); Lee (2001); Webber (2009); Plank (2011)

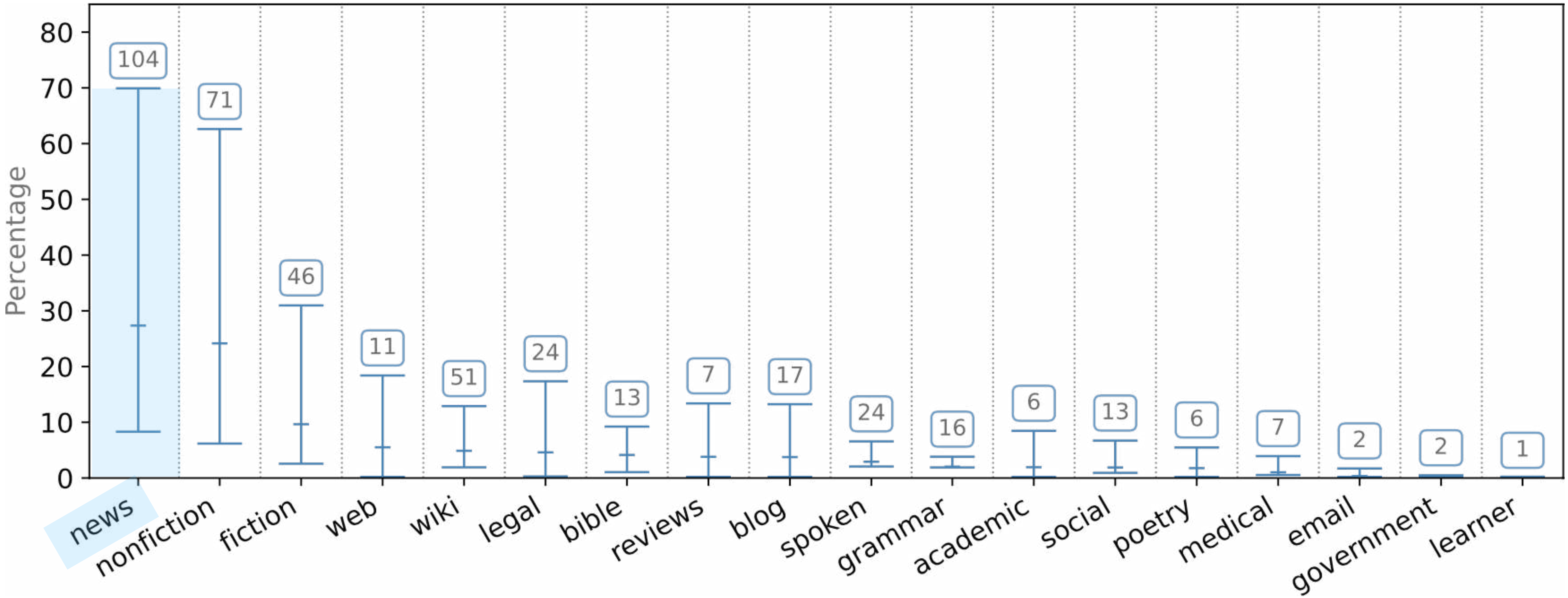
18 community-provided categories in UD

```
=== Machine-readable metadata (DO NOT REMOVE!) ================================
Data available since: UD v2.7
License: CC BY-SA 4.0
Includes text: yes
Genre: spoken
Lemmas: not available
UPOS: converted with corrections
XPOS: not available
Features: not available
Relations: manual native
Contributors: Tyers, Francis; Mischenkova, Karina
Contributing: elsewhere
Contact: ftyers@iu.edu
===============================================================================
```

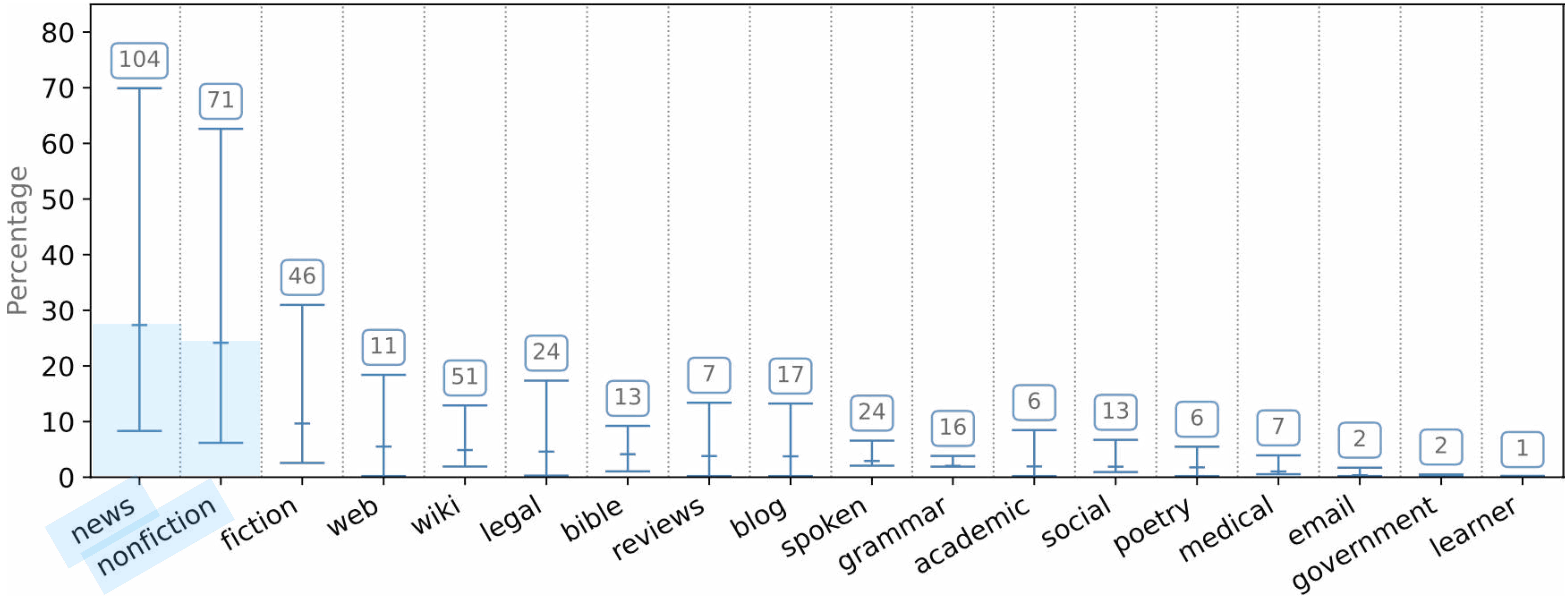


```
=== Machine-readable metadata (DO NOT REMOVE!) ================================
Data available since: UD v1.0
License: CC BY-SA 4.0
Includes text: yes
Genre: blog social reviews email
Lemmas: automatic with corrections
UPOS: converted with corrections
XPOS: manual native
Features: automatic
Relations: manual native
Contributors: Silveira, Natalia; Dozat, Timothy; Manning, Christopher; Schuster,
Sebastian; Chi, Ethan; Bauer, John; Connor, Miriam; de Marneffe, Marie-Catherine;
Schneider, Nathan; Bowman, Sam; Zhu, Hanzhi; Galbraith, Daniel
Contributing: here source
Contact: syntacticdependencies@lists.stanford.edu
===============================================================================
```

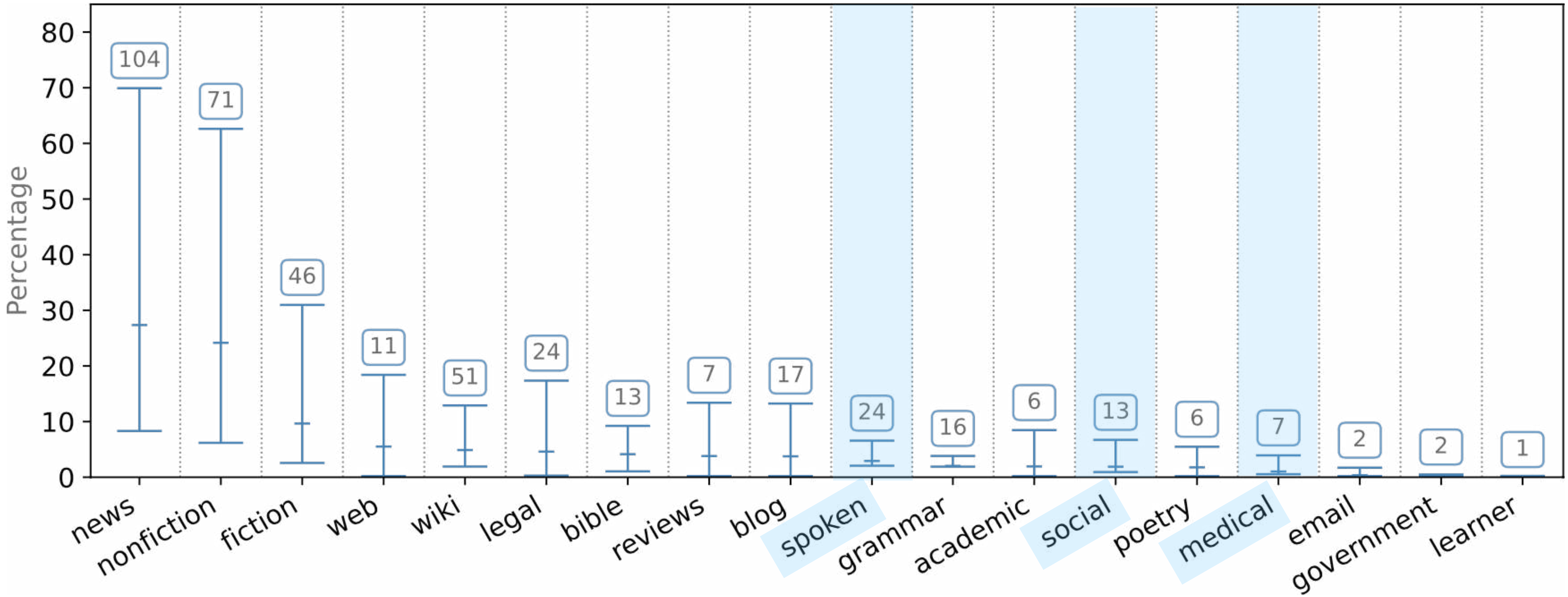
Genre Distribution in UD

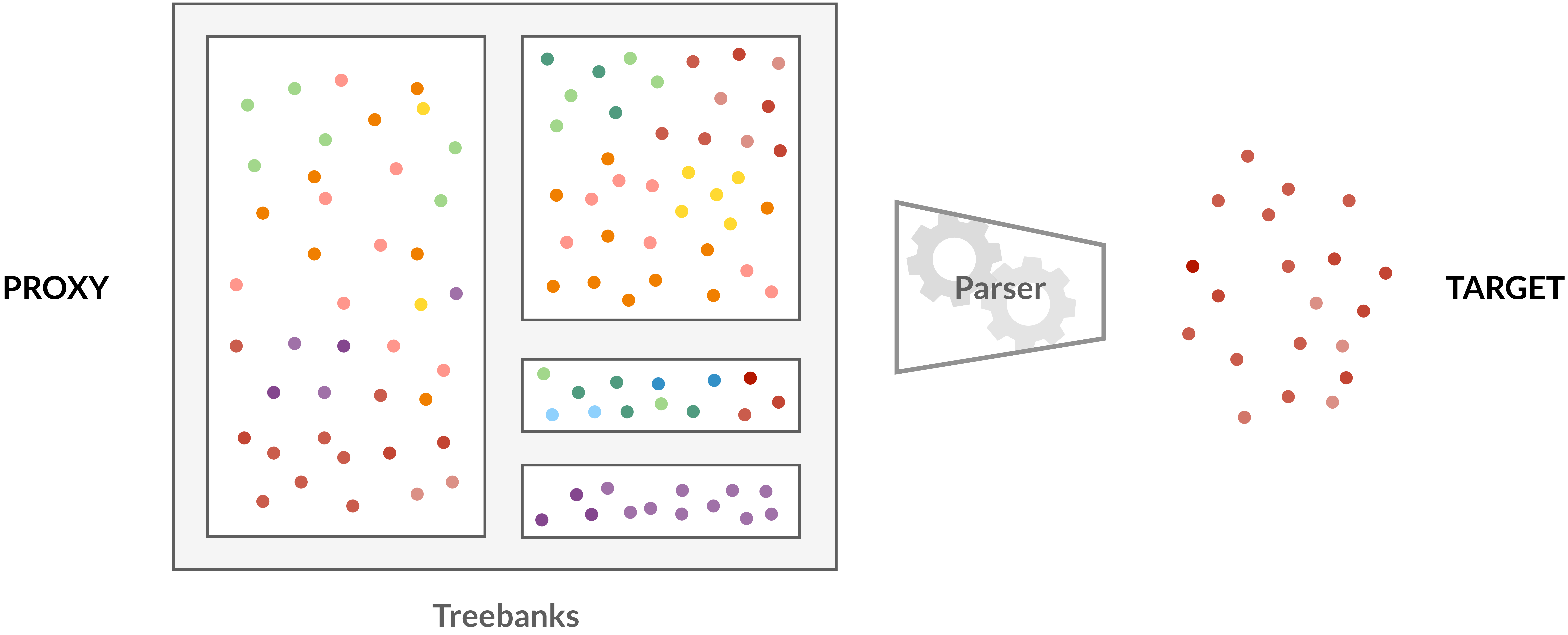


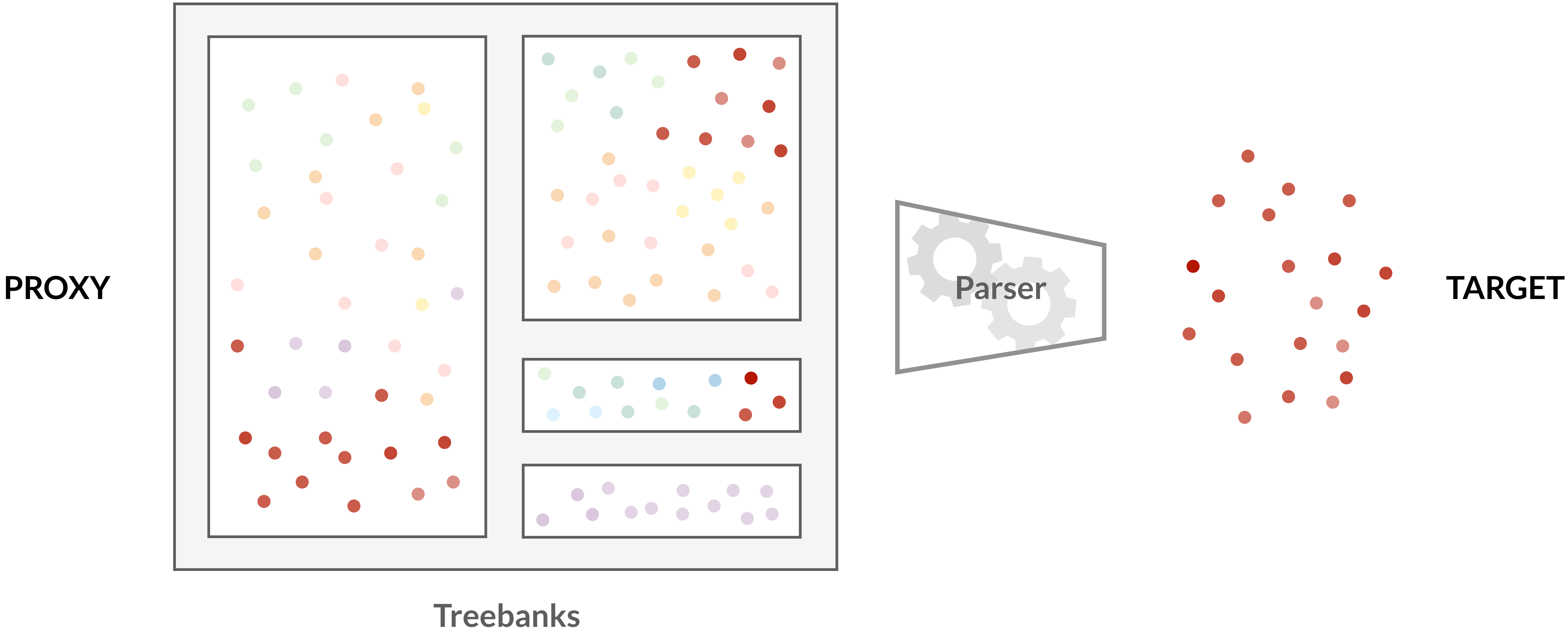
Genre Distribution in UD



Genre Distribution in UD




PROXY
Treebanks
Parser
TARGET



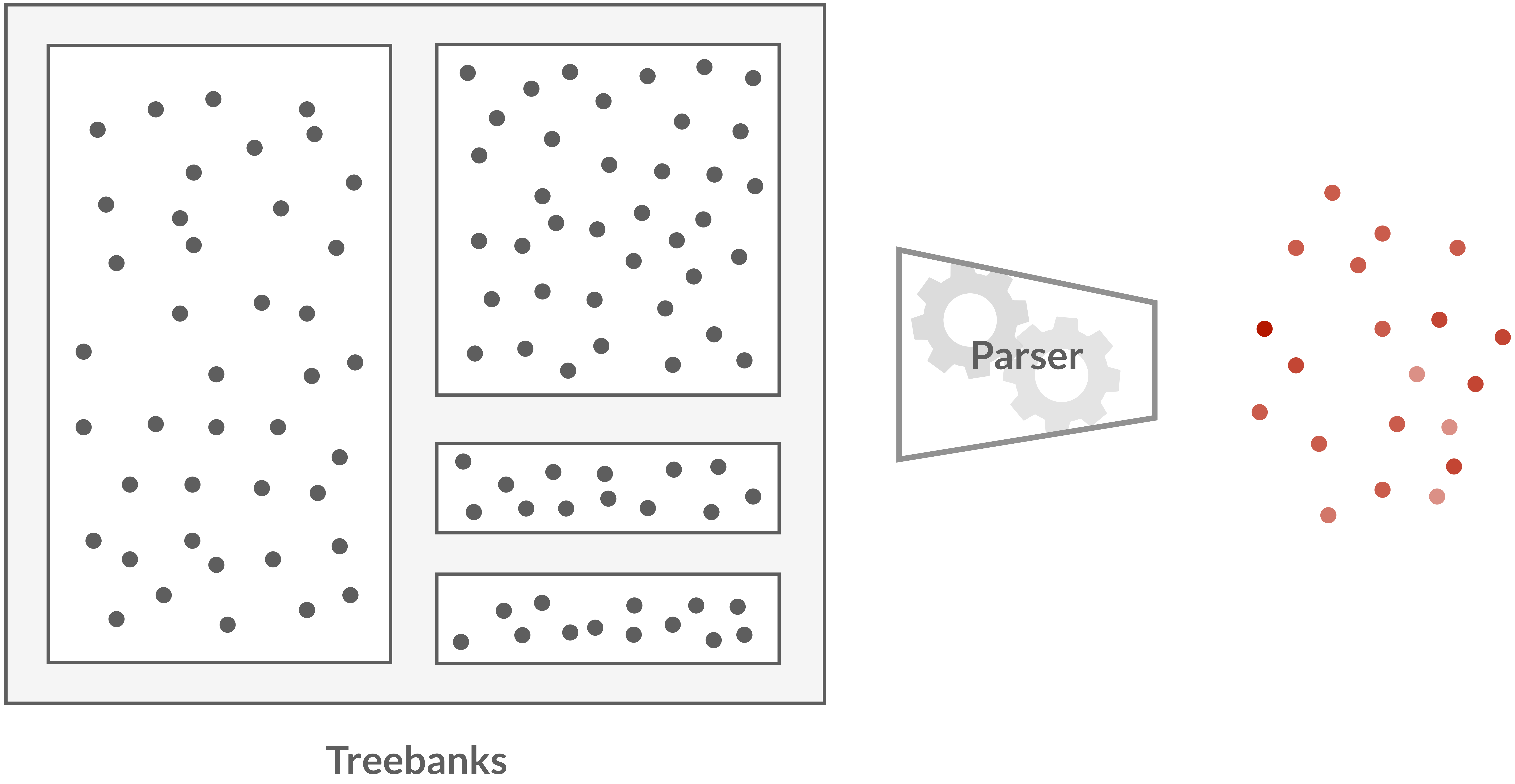
PROXY

Treebanks

Parser

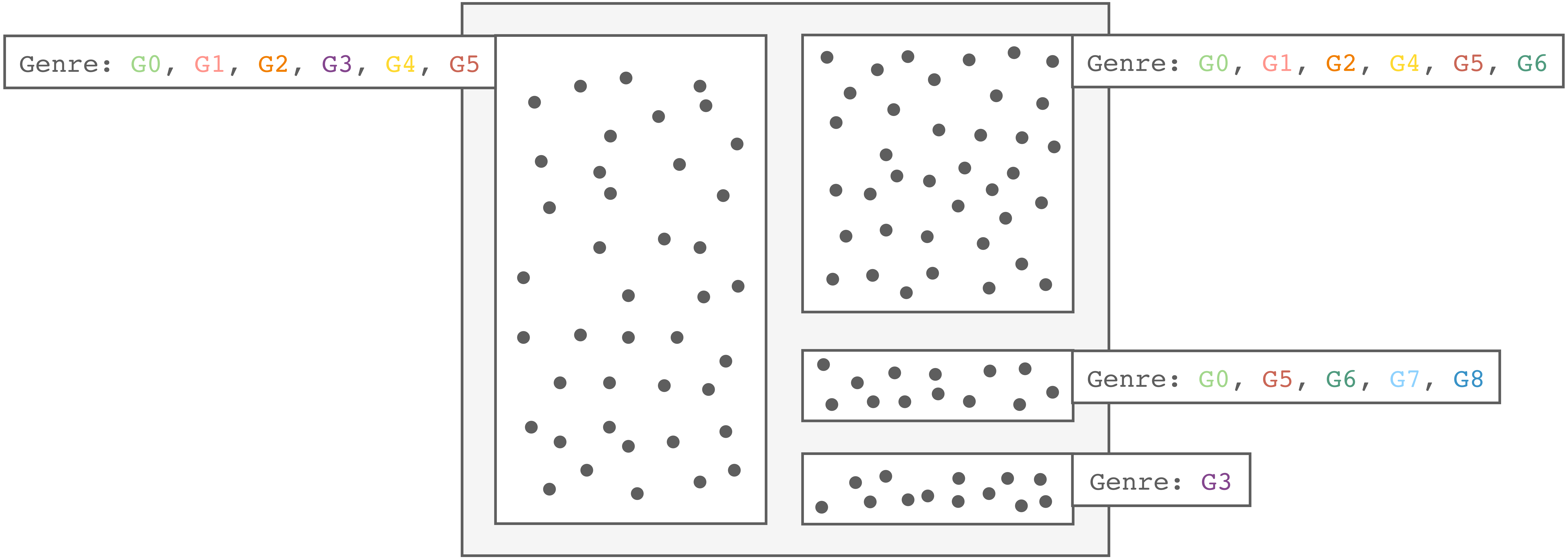
TARGET

PROXY



TARGET

Targeted Data Selection



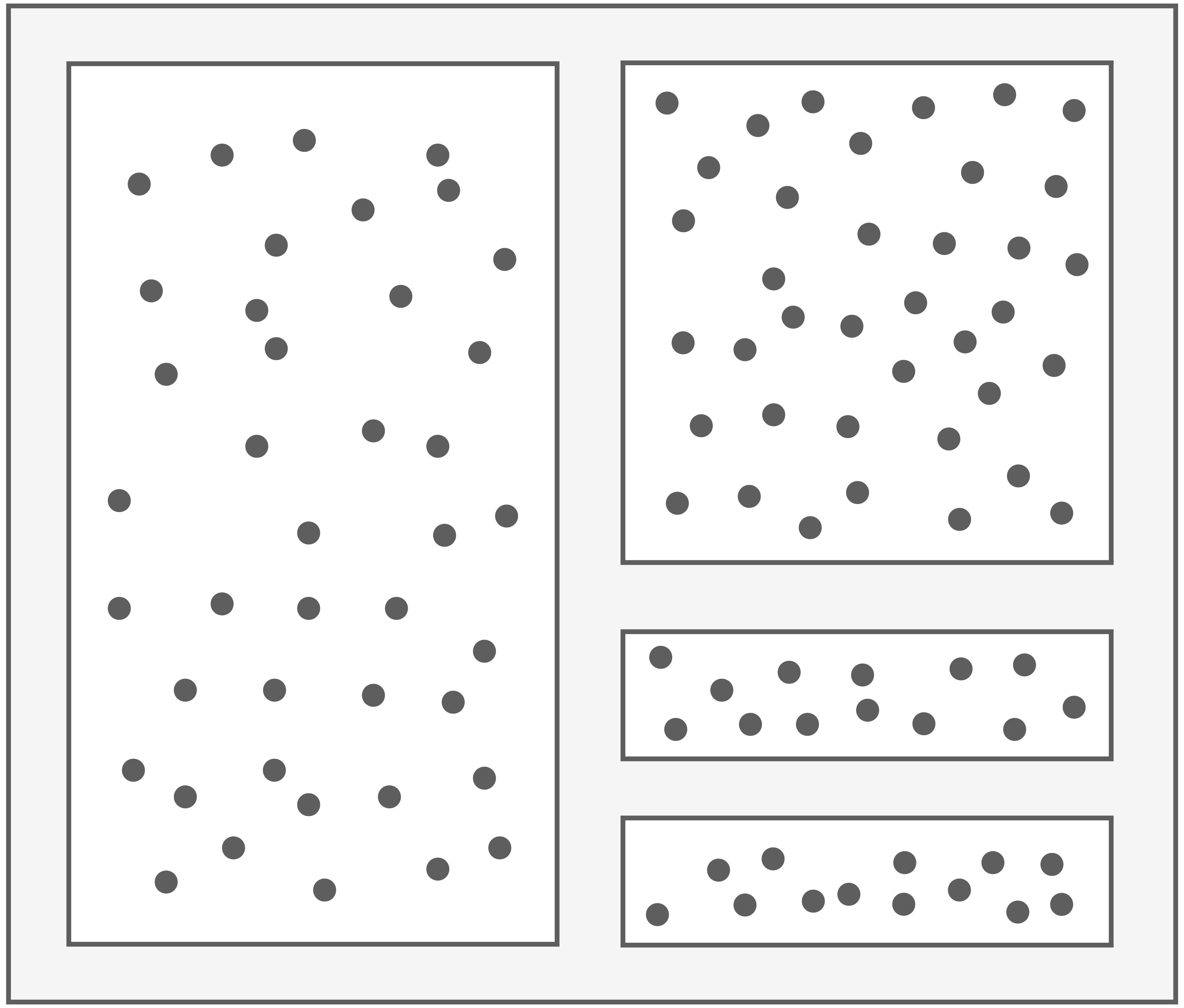
Genre: G0, G1, G2, G3, G4, G5

Genre: G0, G1, G2, G4, G5, G6

Genre: G0, G5, G6, G7, G8

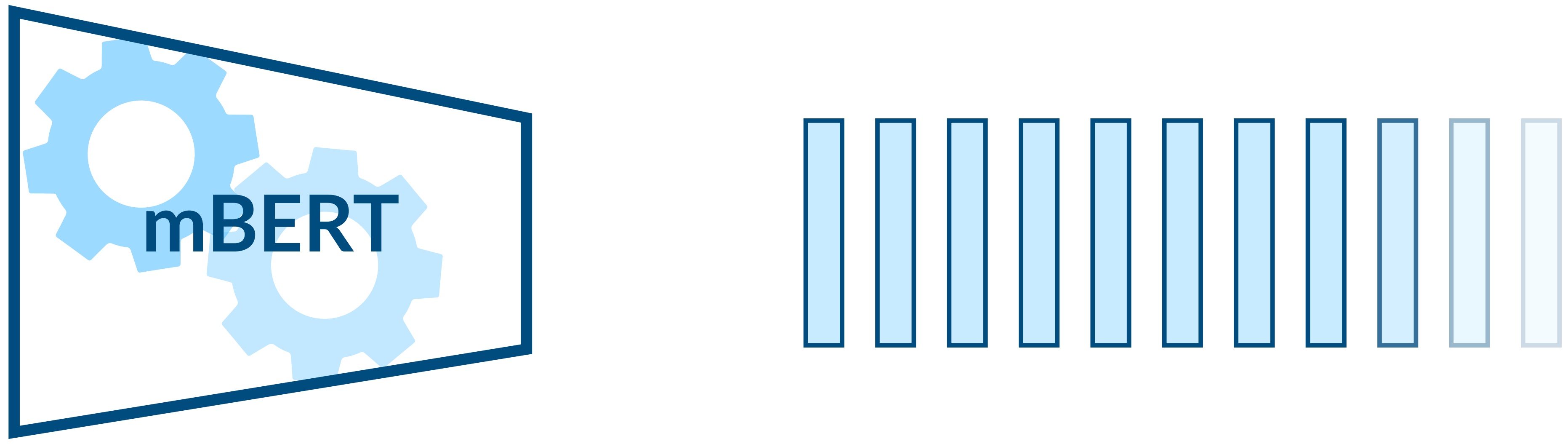
Genre: G3

Treebanks

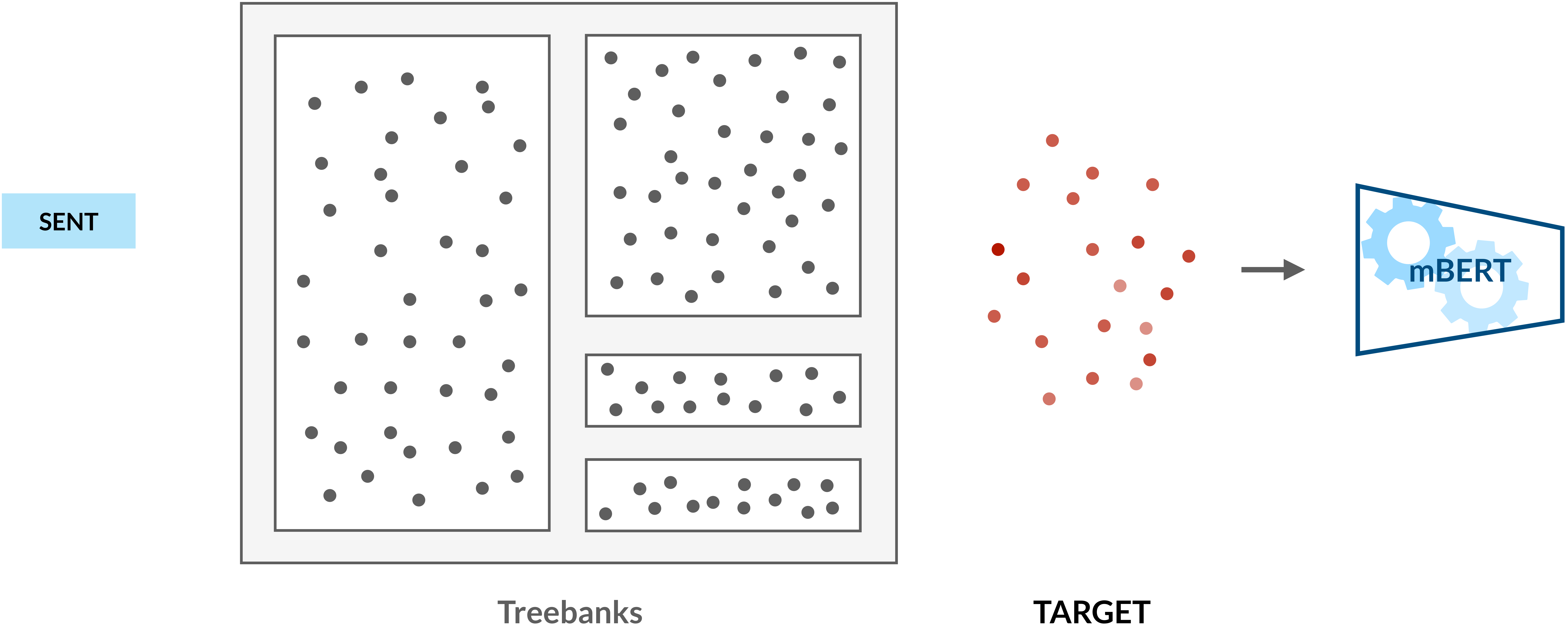


Treebanks

	MODEL	GENRES	LANGS
This Work	mBERT	18	104
Aharoni & Goldberg (2020)	BERT	5	1



Devlin et al. (2019)



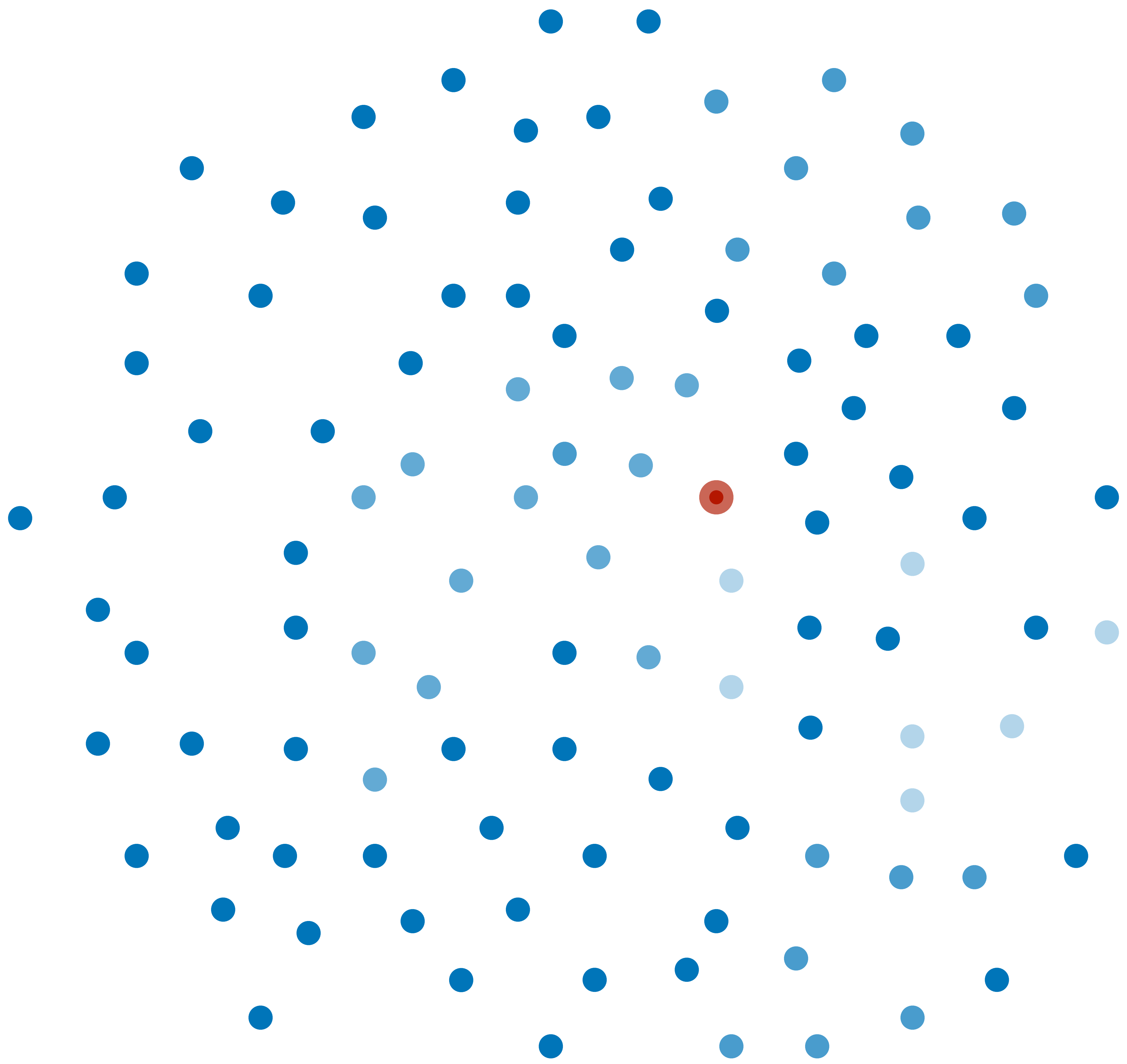
SENT

Treebanks

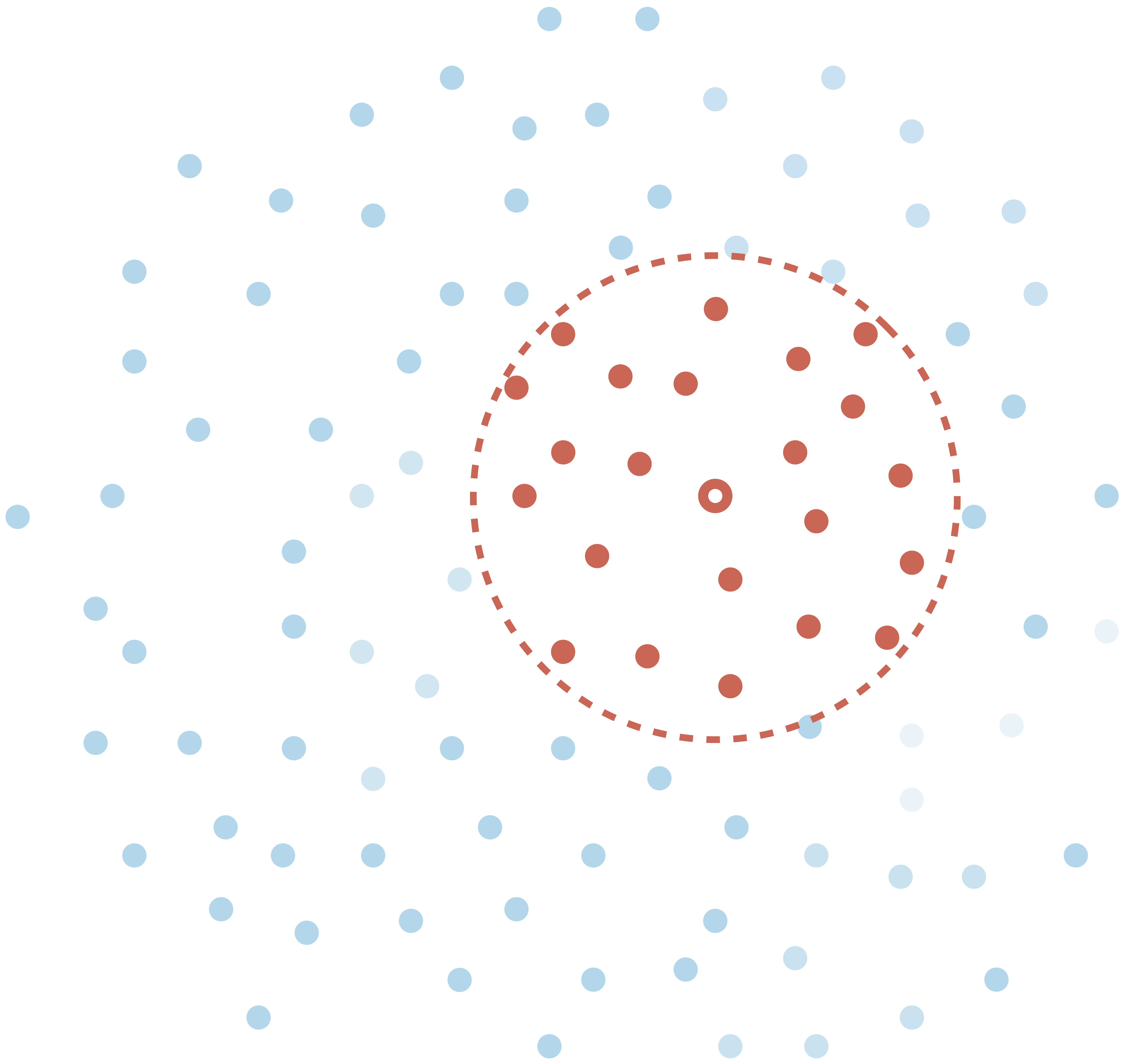
TARGET

mBERT

SENT

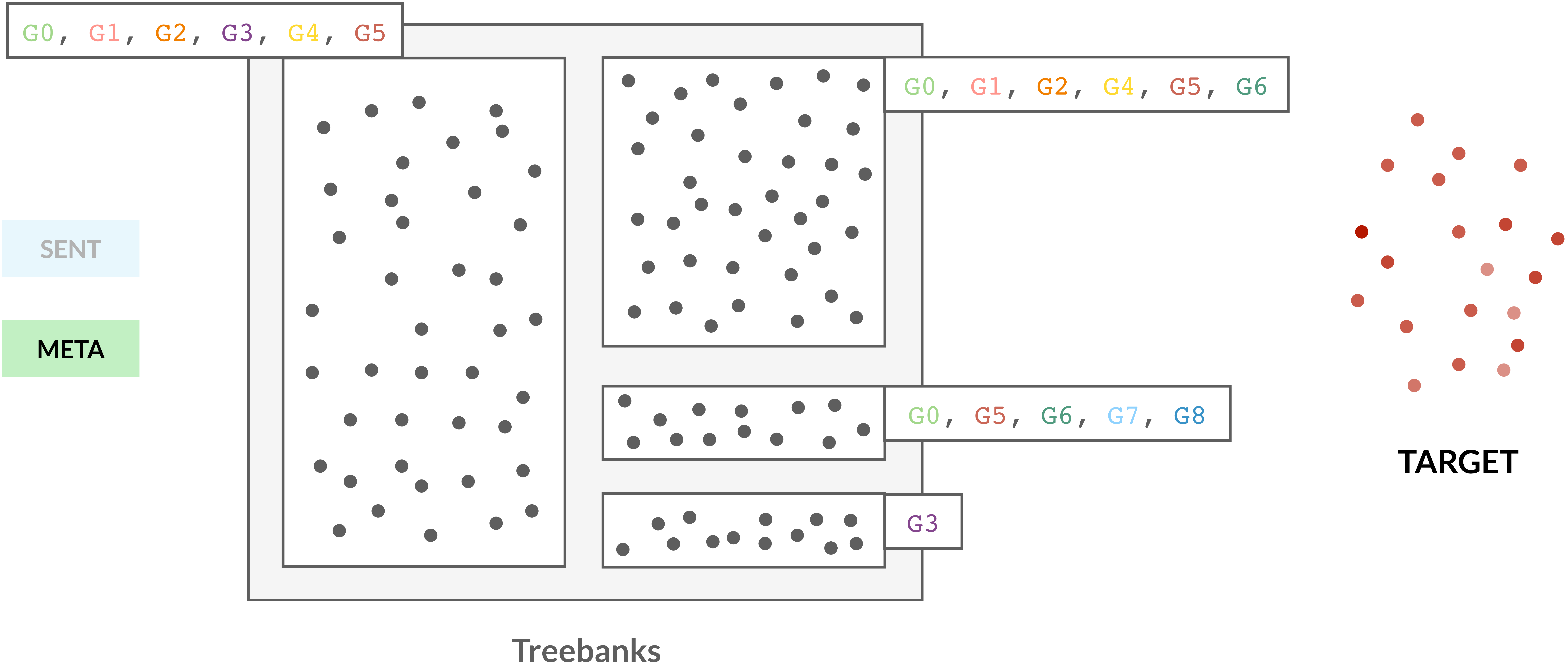


SENT



SENT

META



G0, G1, G2, G3, G4, G5

G0, G1, G2, G4, G5, G6

G0, G5, G6, G7, G8

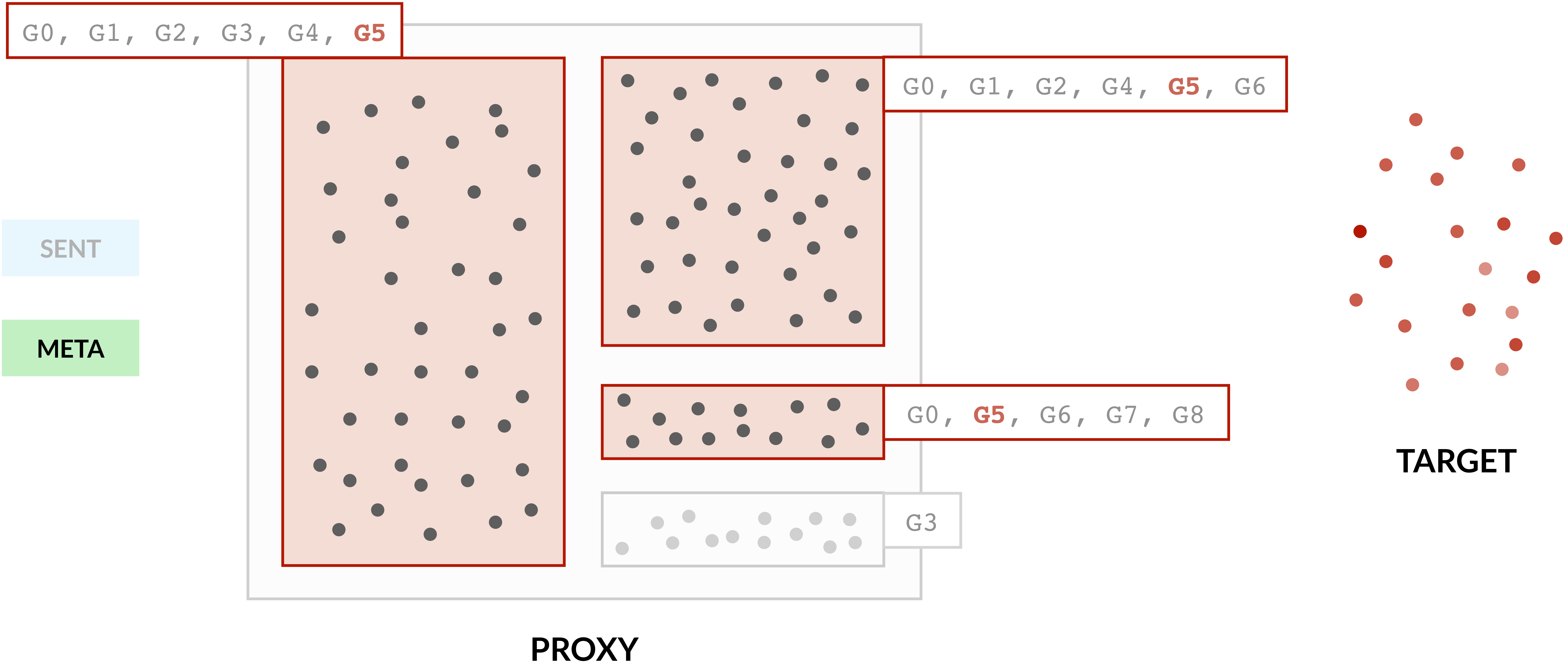
G3

SENT

META

TARGET

Treebanks



G0, G1, G2, G3, G4, G5
G0, G1, G2, G4, G5, G6
G0, G5, G6, G7, G8
G3
SENT
META
PROXY
TARGET

SENT

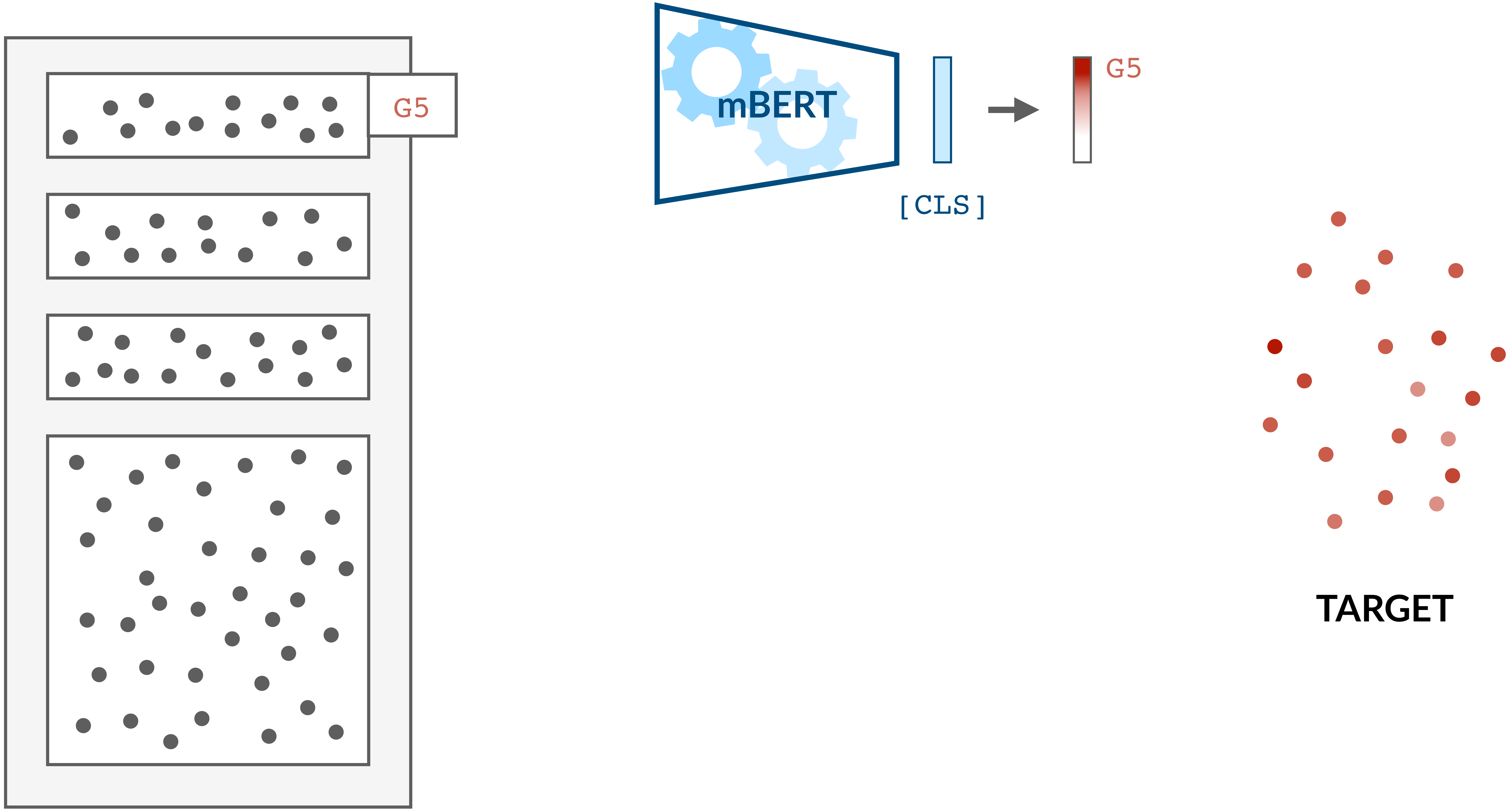
META

BOOT

SENT

META

BOOT



G5

mBERT

[CLS]

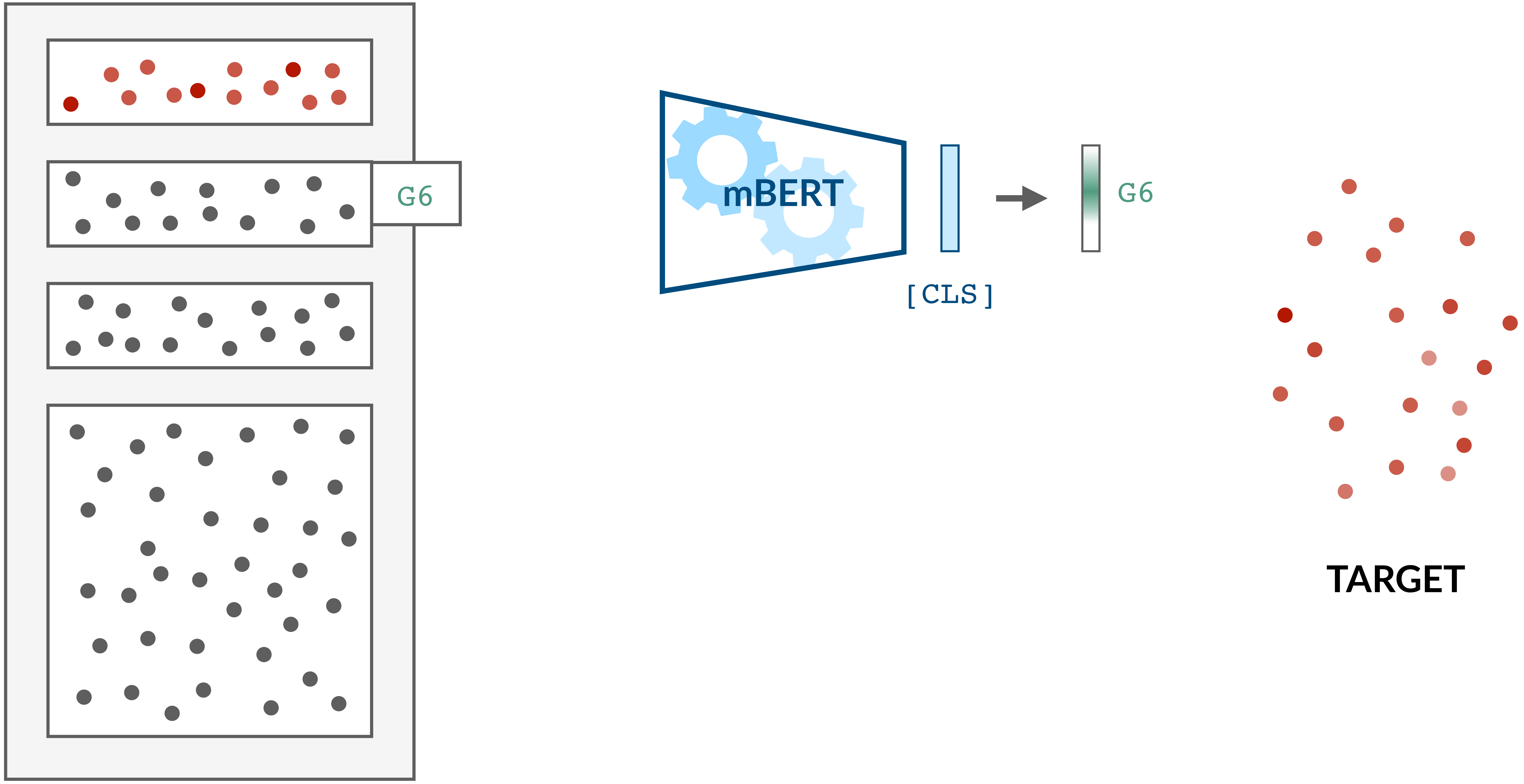
G5

TARGET

SENT

META

BOOT



G6

mBERT

[CLS]

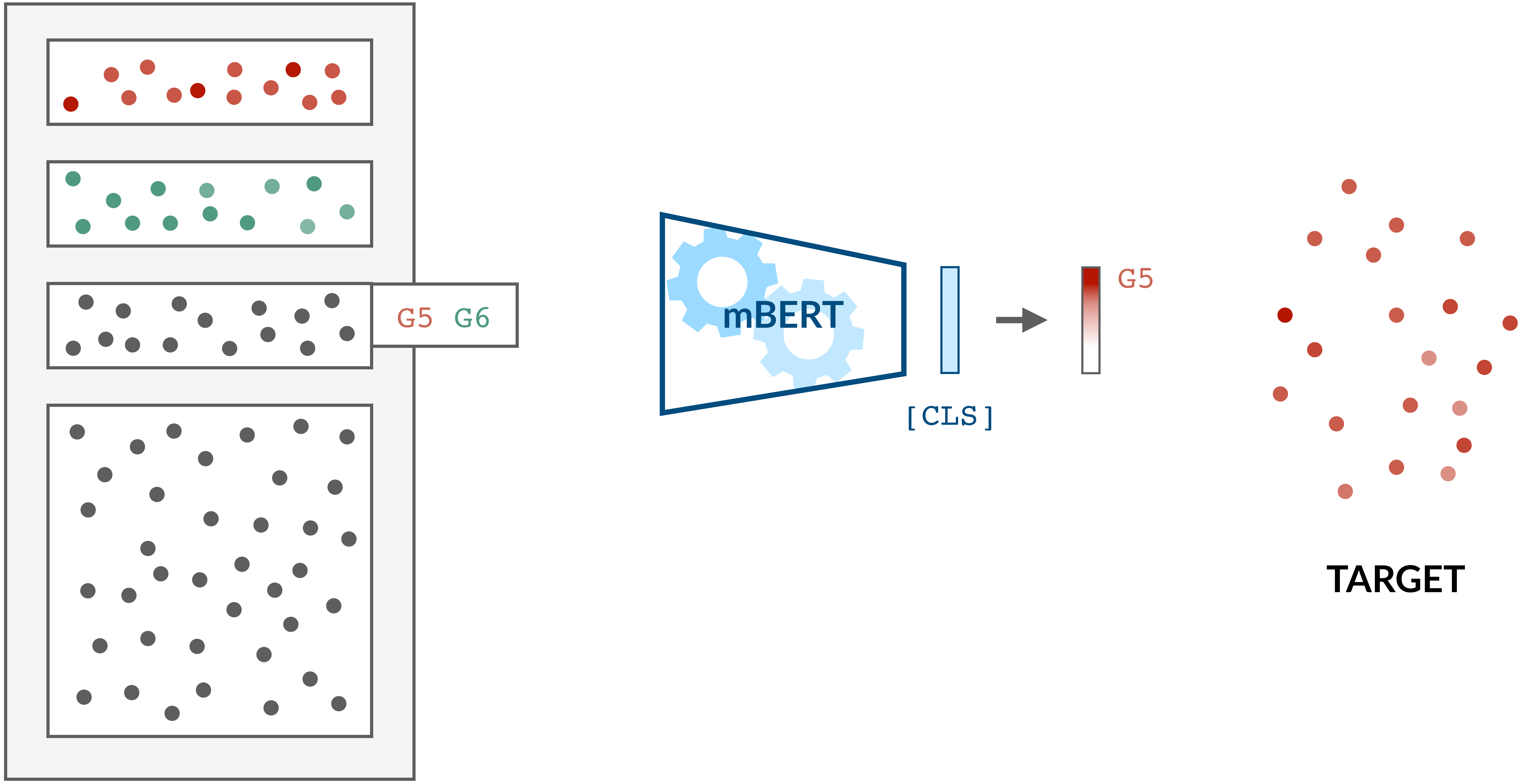
G6

TARGET

SENT

META

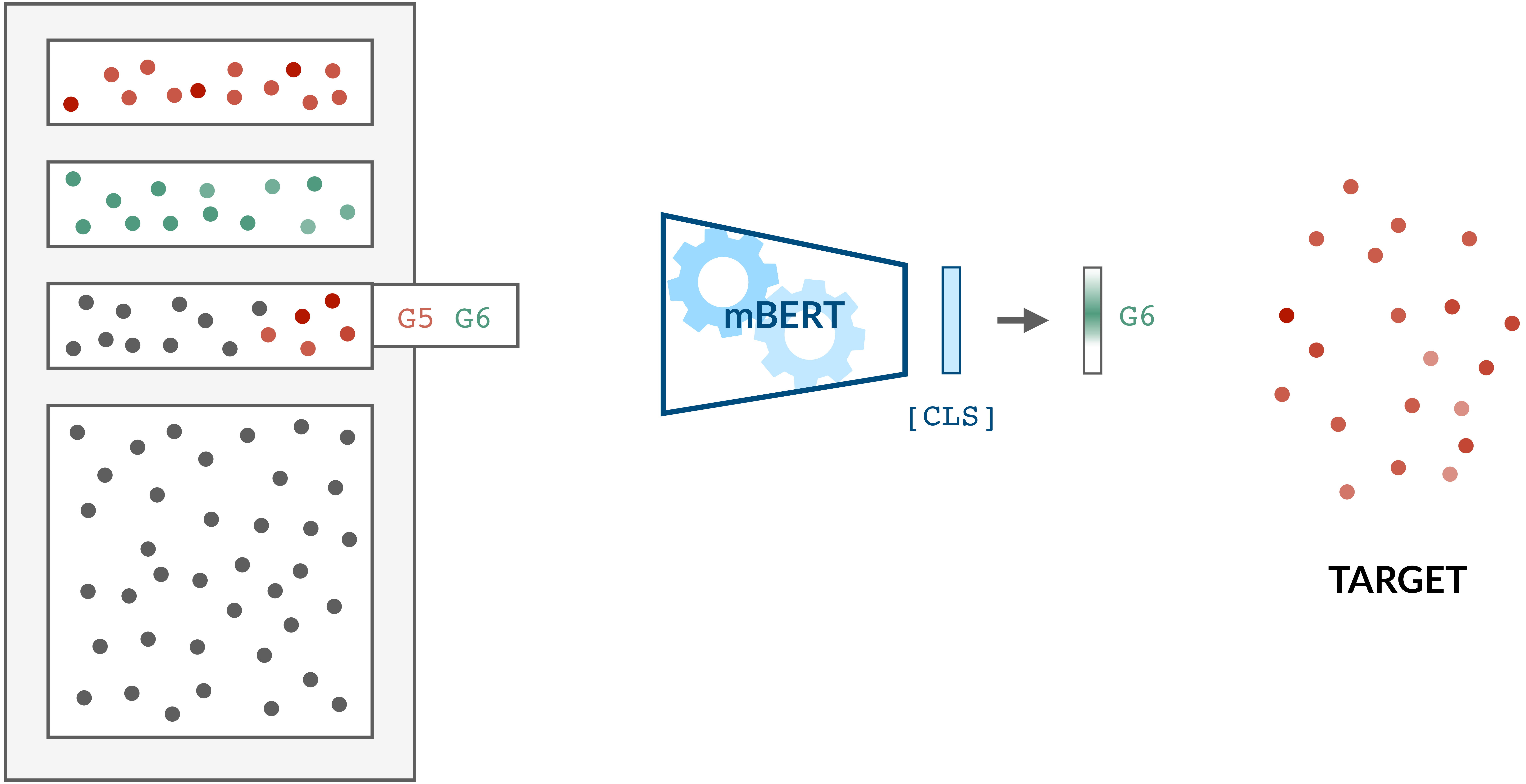
BOOT



SENT

META

BOOT



G5 G6

mBERT

[CLS]

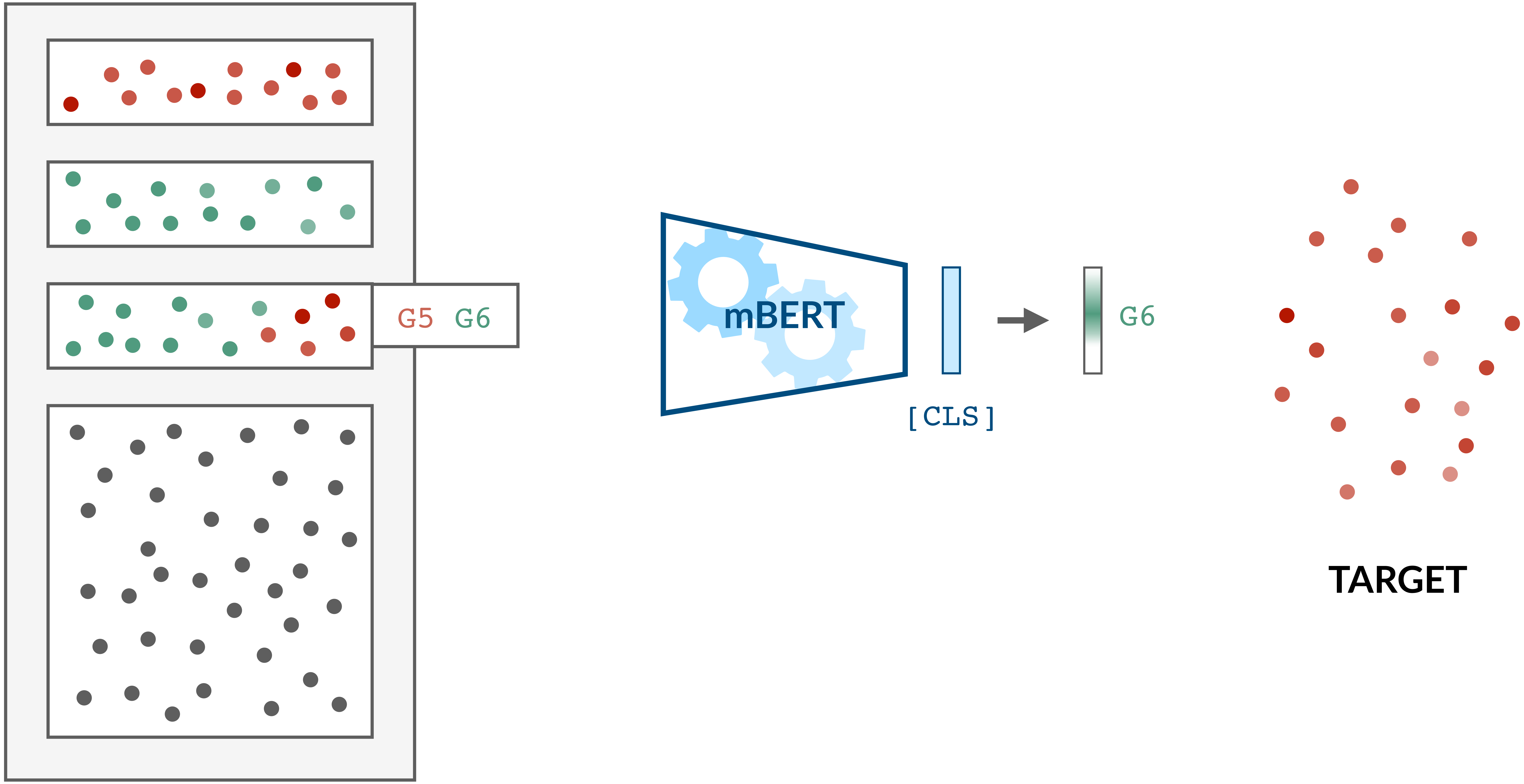
G6

TARGET

SENT

META

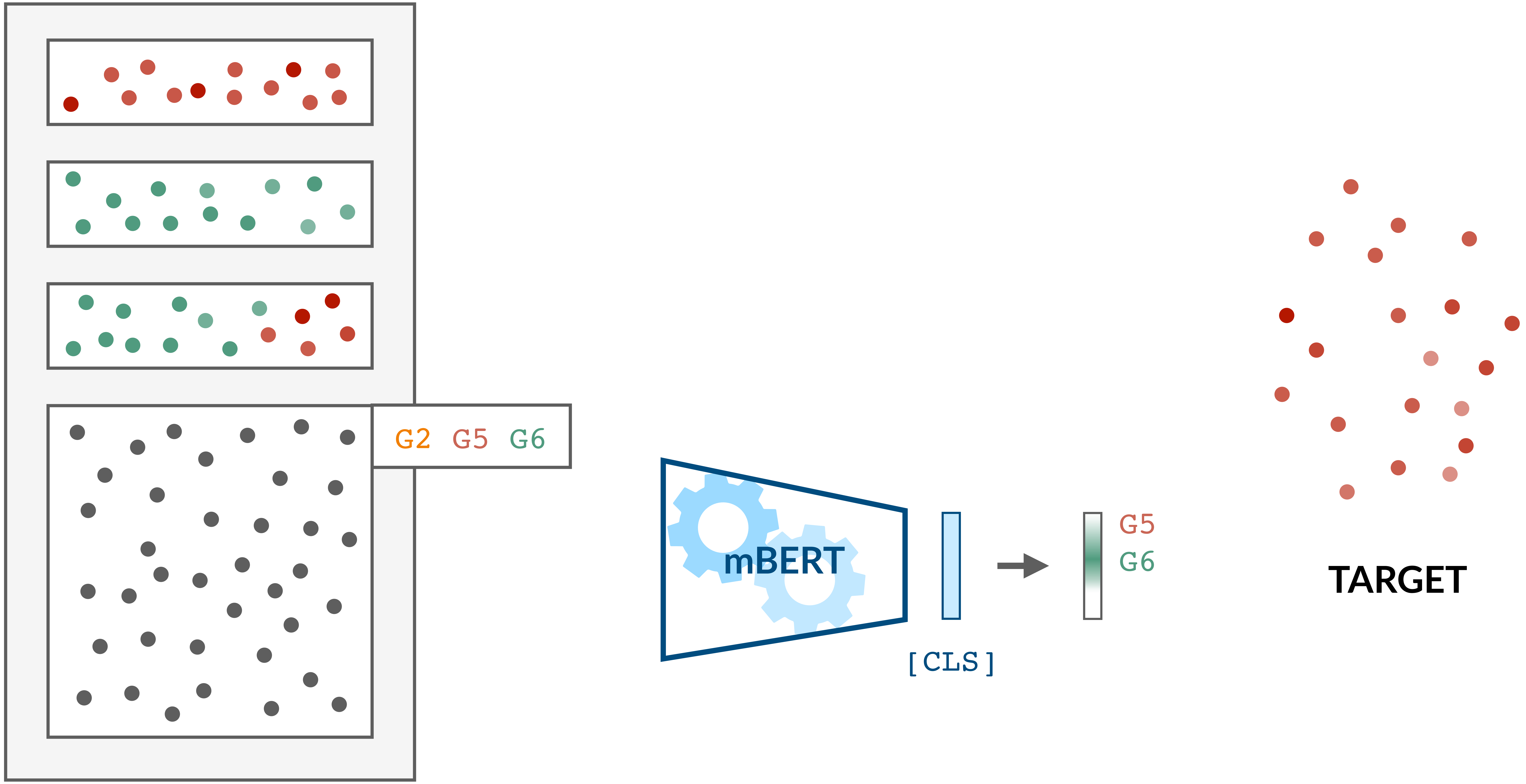
BOOT



SENT

META

BOOT



G2 G5 G6

mBERT

[CLS]

G5

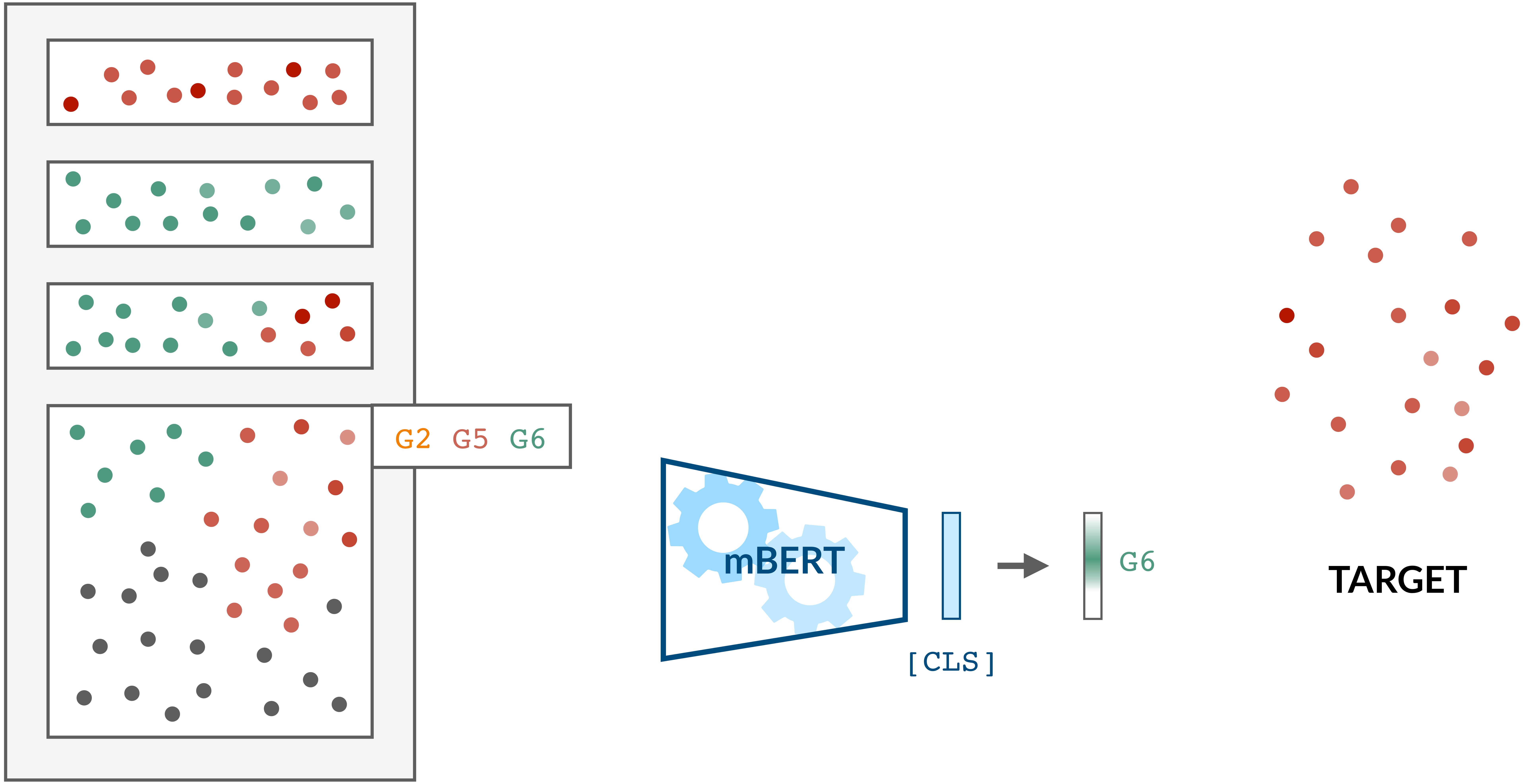
G6

TARGET

SENT

META

BOOT



G2 G5 G6

mBERT

[CLS]

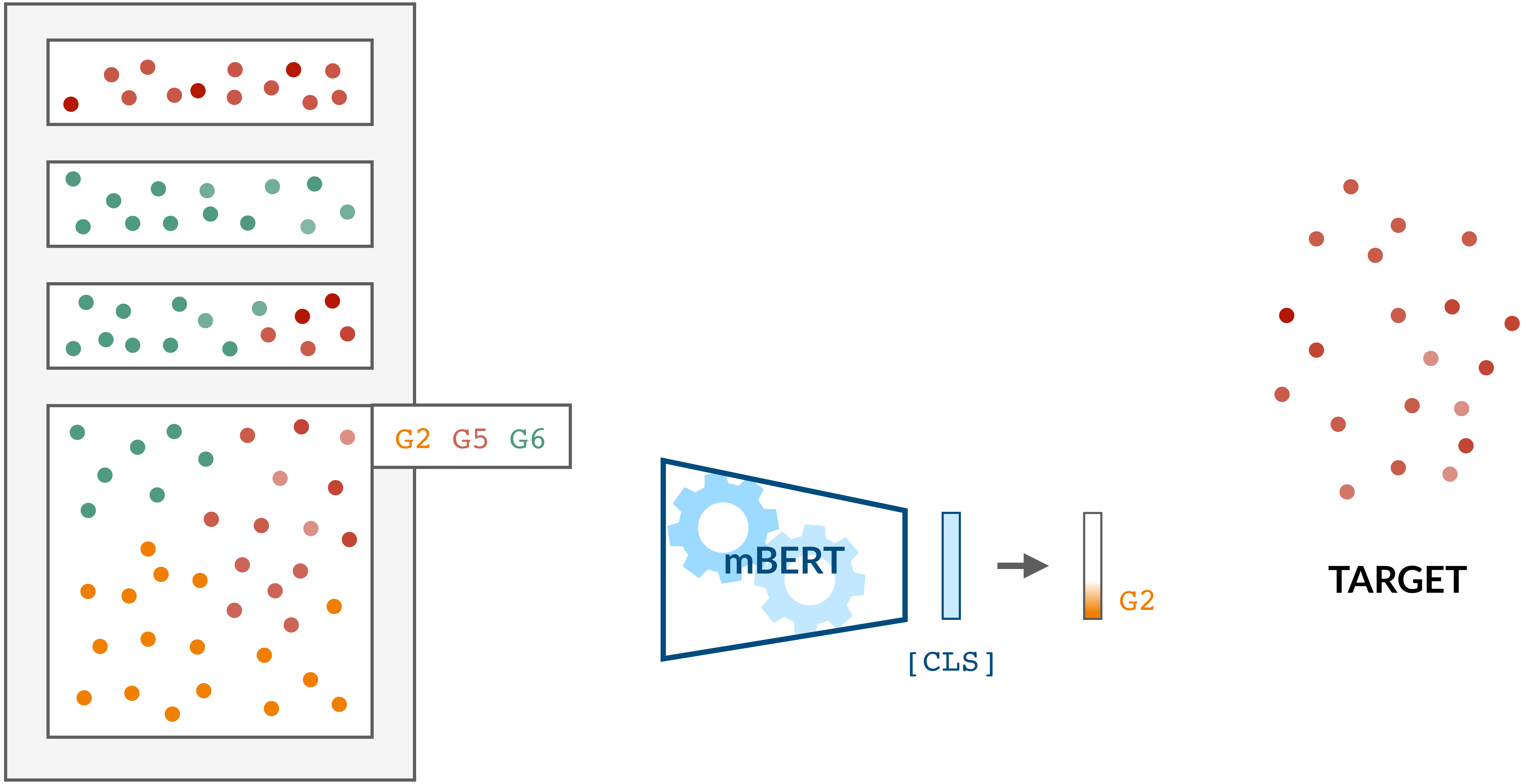
G6

TARGET

SENT

META

BOOT



G2 G5 G6

mBERT

[CLS]

G2

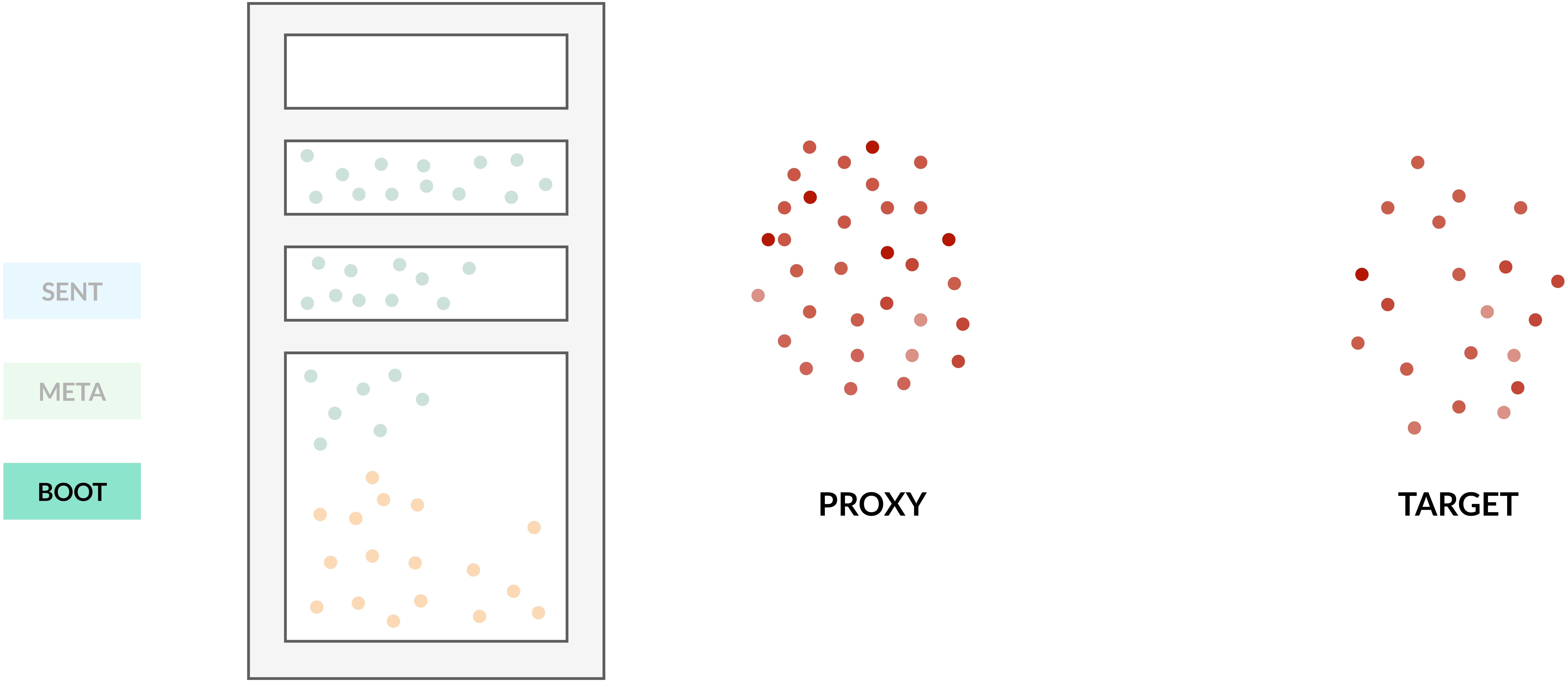
TARGET

SENT

META

BOOT

TARGET



SENT

META

BOOT

PROXY

TARGET

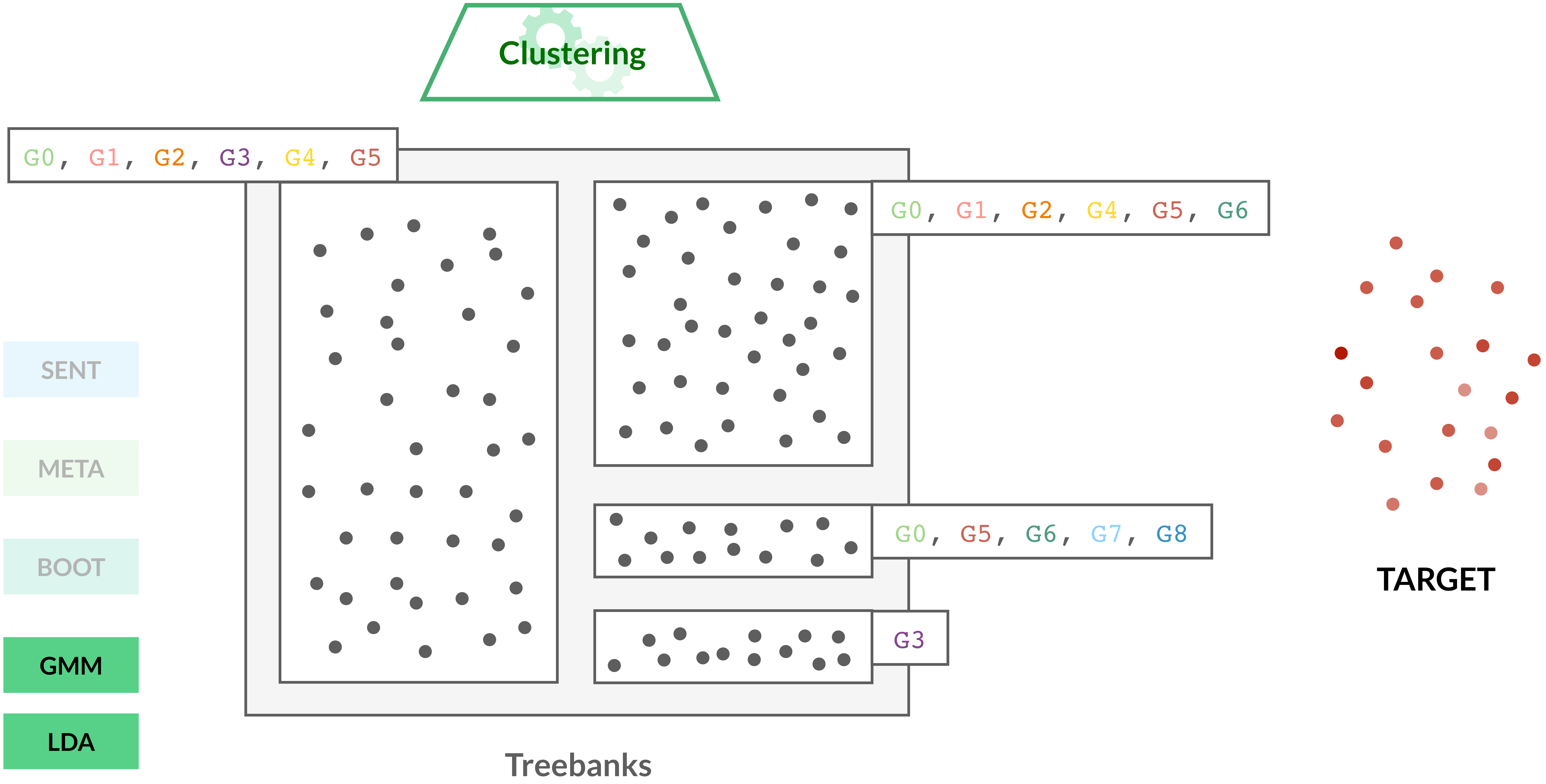
SENT

META

BOOT

GMM

LDA



Clustering

G0, G1, G2, G3, G4, G5

G0, G1, G2, G4, G5, G6

G0, G5, G6, G7, G8

G3

TARGET

SENT

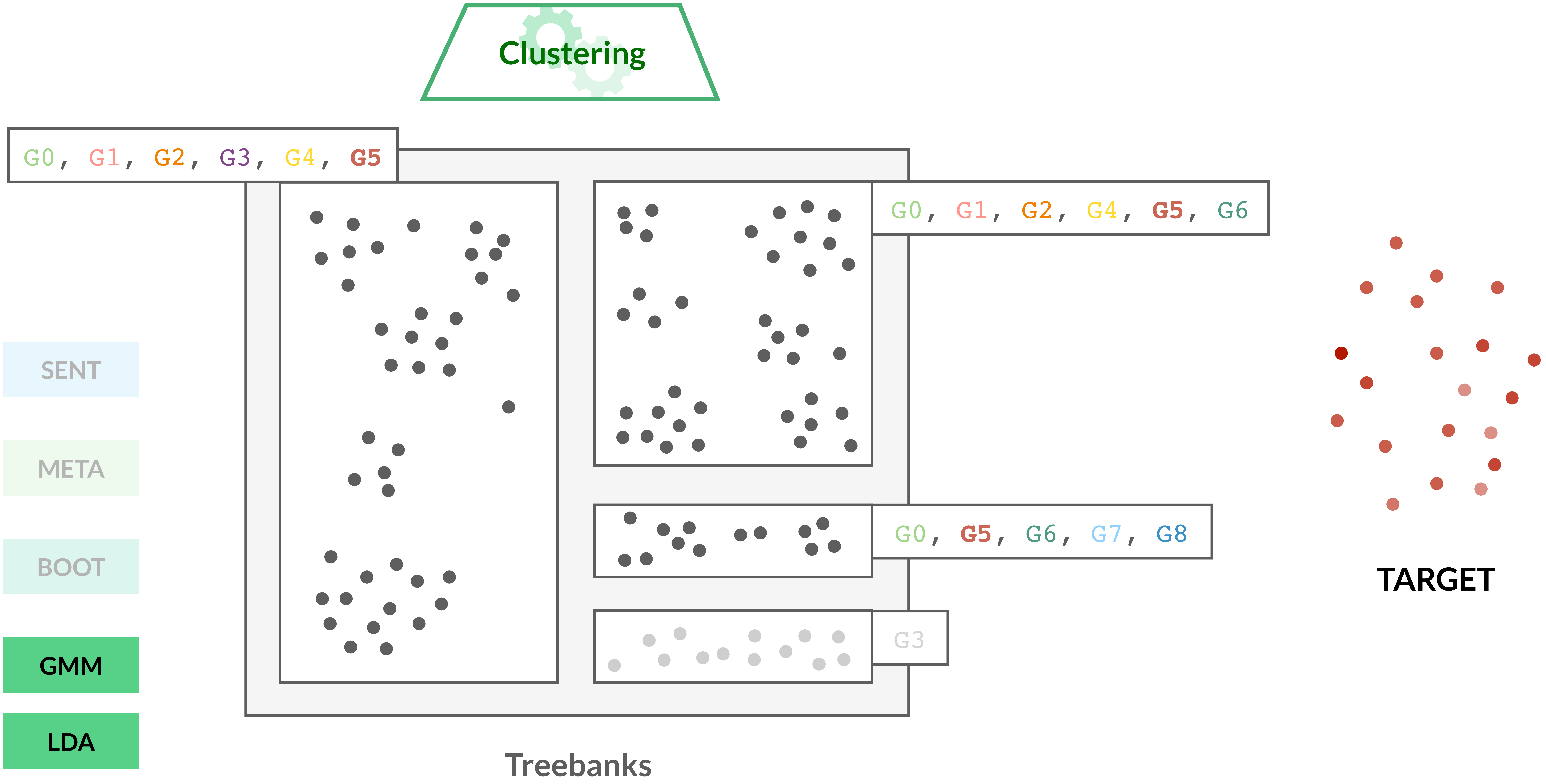
META

BOOT

GMM

LDA

Treebanks



Clustering

G0, G1, G2, G3, G4, G5

G0, G1, G2, G4, G5, G6

G0, G5, G6, G7, G8

G3

SENT

META

BOOT

GMM

LDA

Treebanks

TARGET

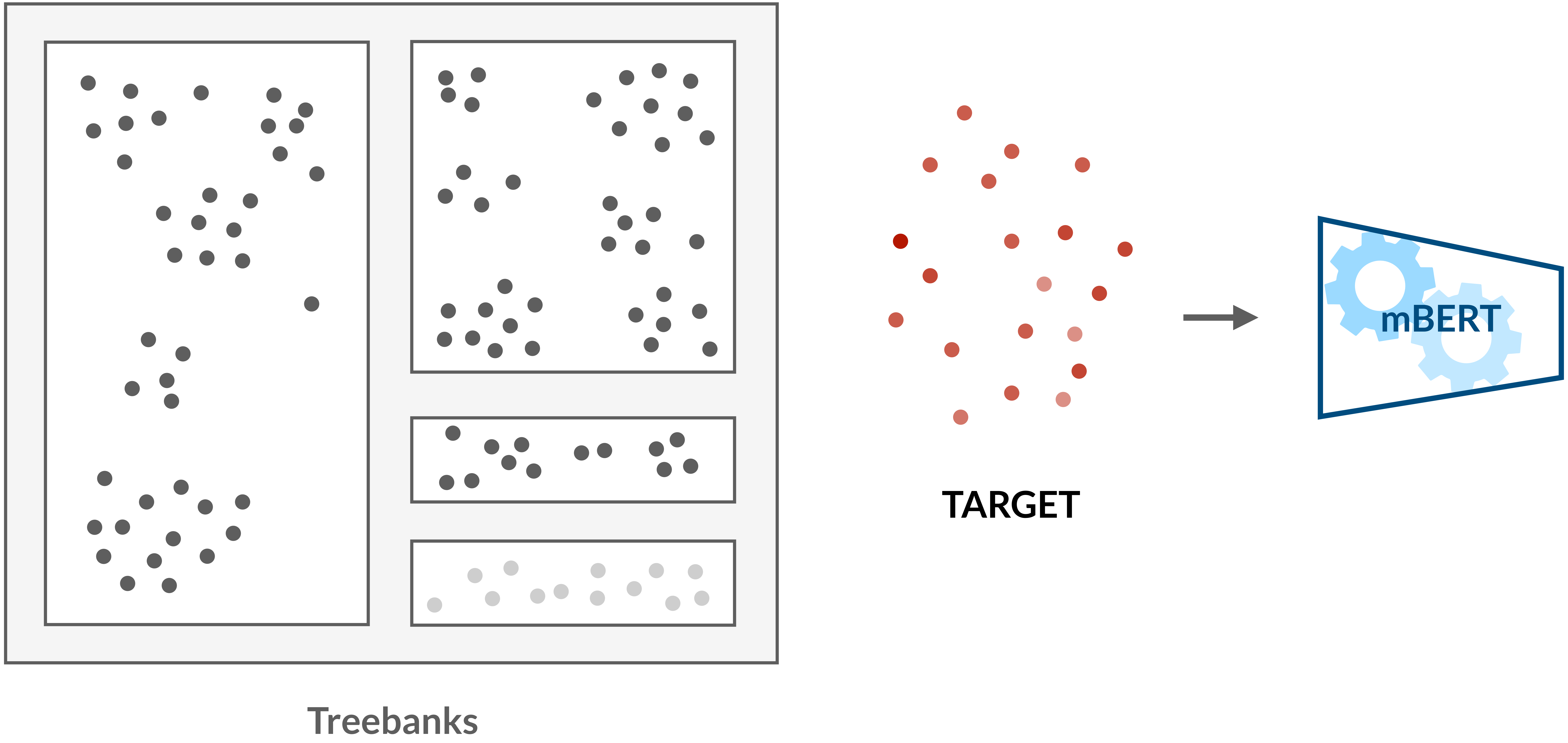
SENT

META

BOOT

GMM

LDA



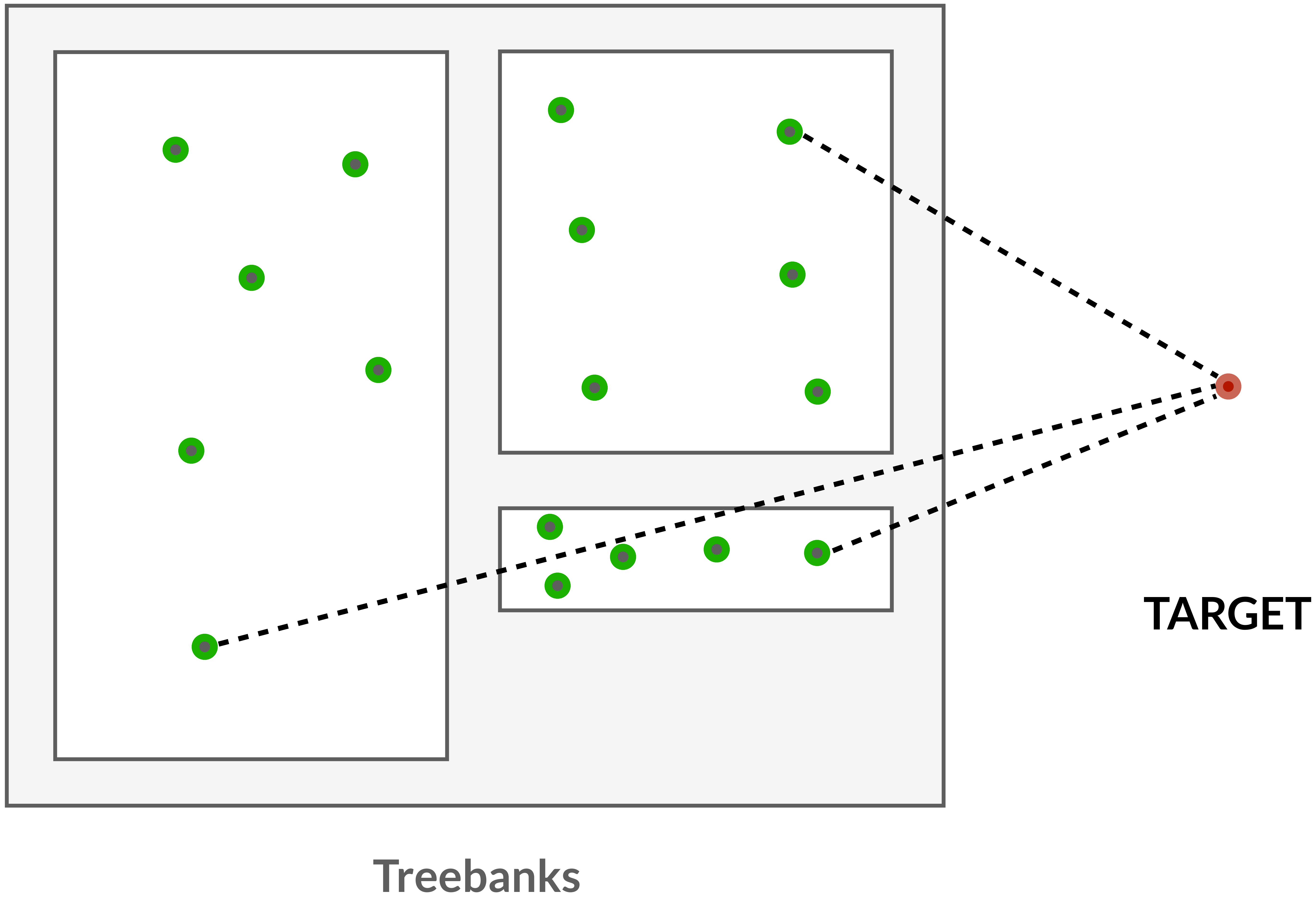
SENT

META

BOOT

GMM

LDA



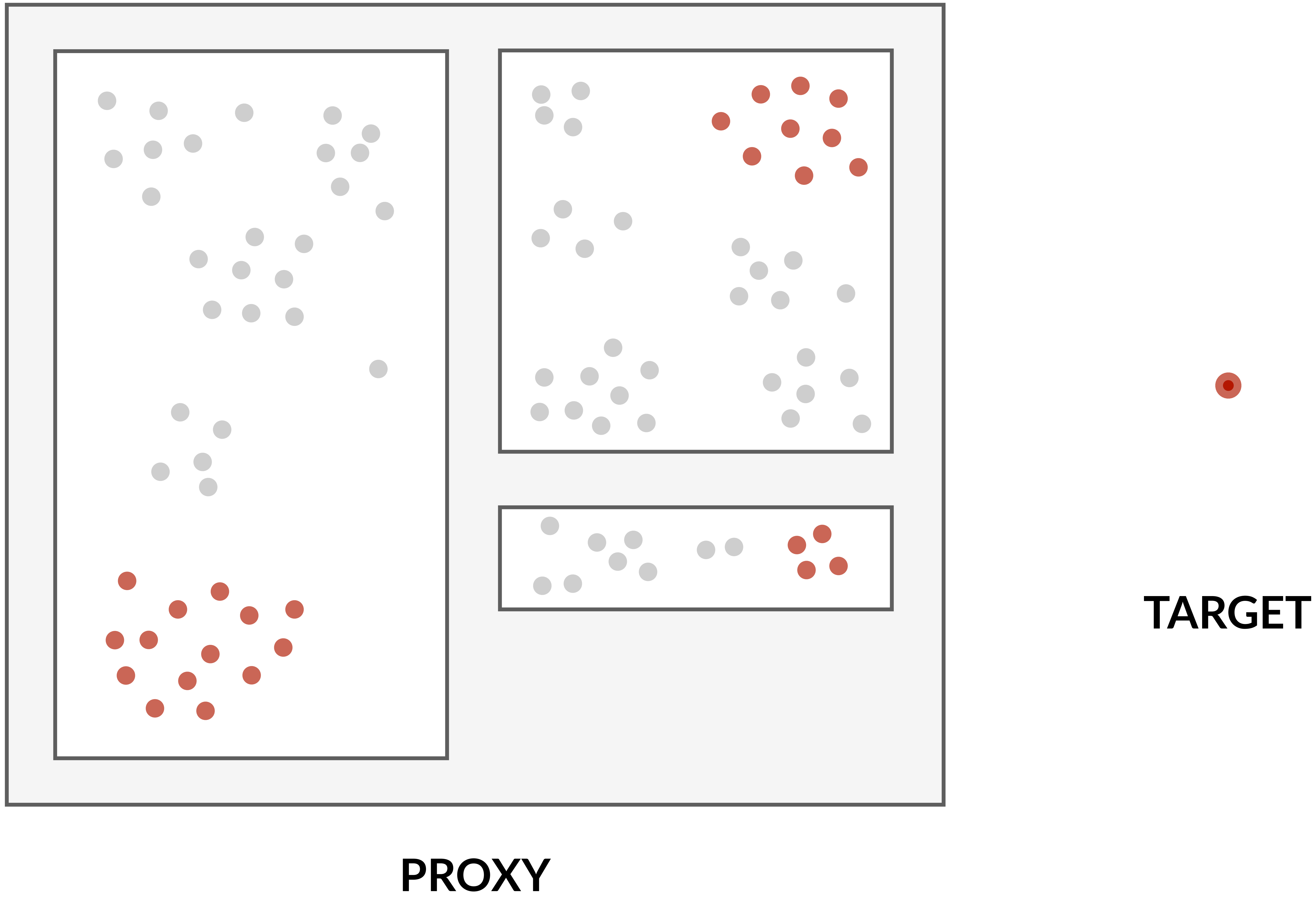
SENT

META

BOOT

GMM

LDA



Experiments

Target		Authors	Language	#Sentences	mBERT	Genre
SWL 💬	SSLC	Östling et al. (2017)	Swedish Sign Language	203	✗	spoken
SA 📕	UFAL	Dwivedi and Easha (2017)	Sanskrit	230	✗	fiction
KPV 📕	Lattice	Partanen et al. (2018)	Komi Zyrian	435	✗	fiction
TA 📰	TTB	Ramasamy and Žabokrtský (2012)	Tamil	600	✓	news
GL 📰	TreeGal	Garcia (2016)	Galician	1,000	✓	news
YUE 💬	HK	Wong et al. (2017)	Cantonese	1,004	✗	spoken
CKT 💬	HSE	Tyers and Mishchenkova (2020)	Chukchi	1,004	✗	spoken
FO W	OFT	Tyers et al. (2018)	Faroese	1,208	✗	wiki
TE 🪄	MTG	Rama and Vajjala (2017)	Telugu	1,328	✓	grammar
MYV 📕	JR	Rueter and Tyers (2018)	Erzya	1,690	✗	fiction
QHE 📡	HIENCS	Bhat et al. (2018)	Hindi-English	1,800	~	social
QTD 💬	SAGT	Çetinoğlu and Çöltekin (2019)	Turkish-German	1,891	~	spoken

	SWL 💬	SA 📕	KPV 📕	TA 📰	GL 📰	YUE 💬	CKT 💬	FO W	TE 🪄	MYV 📕	QHE 📡	QTD 💬
TARGET	✓	~	~	✓	✓	✗	✗	~	✓	✗	✓	✓
SENT												
META												
BOOT												
GMM												
LDA												

	SWL 💬	SA 📕	KPV 📕	TA 📰	GL 📰	YUE 💬	CKT 💬	FO W	TE 🪄	MYV 📕	QHE 📶	QTD 💬
TARGET												
RAND												
SENT												
META												
BOOT												
GMM												
LDA												

SWL 💬 | SA 📖 | KPV 📖 | TA 📰 | GL 📰 | YUE 💬 | CKT 💬 | FO W | TE 🪄 | MYV 📖 | QHE 📶 | QTD 💬

TARGET

RAND

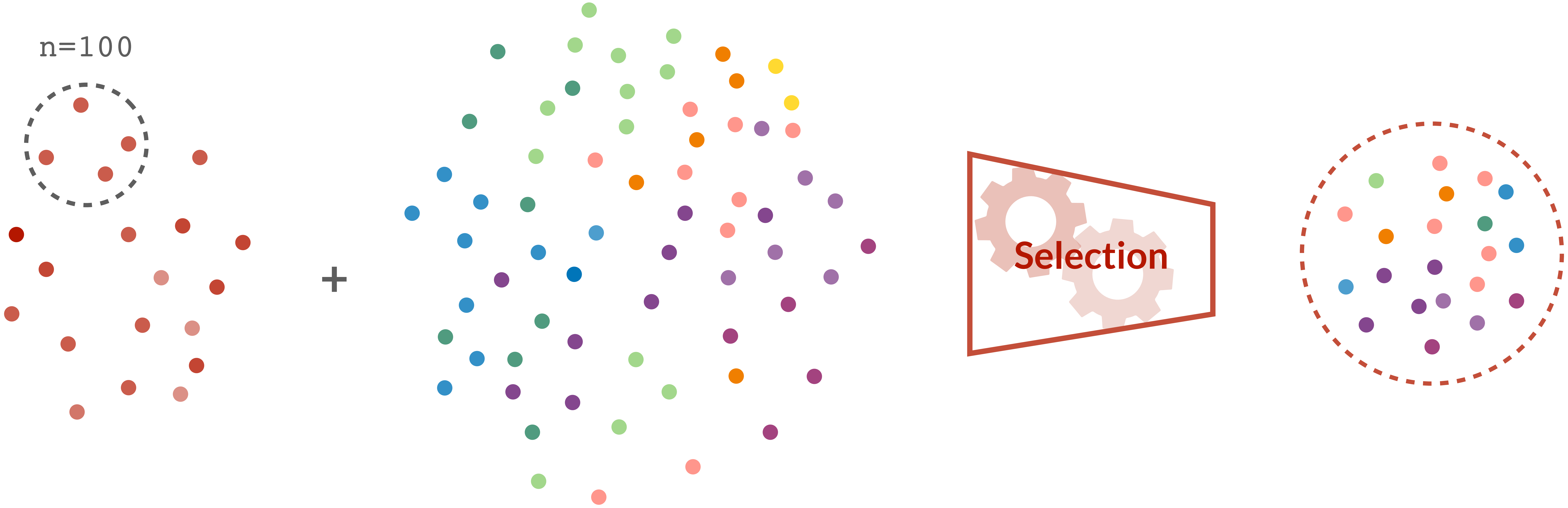
SENT

META

BOOT

GMM

LDA



TARGET
(unannotated)

NON-TARGET
(annotated)

PROXY
(annotated)

SWL 💬 | SA 📖 | KPV 📖 | TA 📰 | GL 📰 | YUE 💬 | CKT 💬 | FO W | TE 🪄 | MYV 📖 | QHE 📡 | QTD 💬

TARGET

RAND

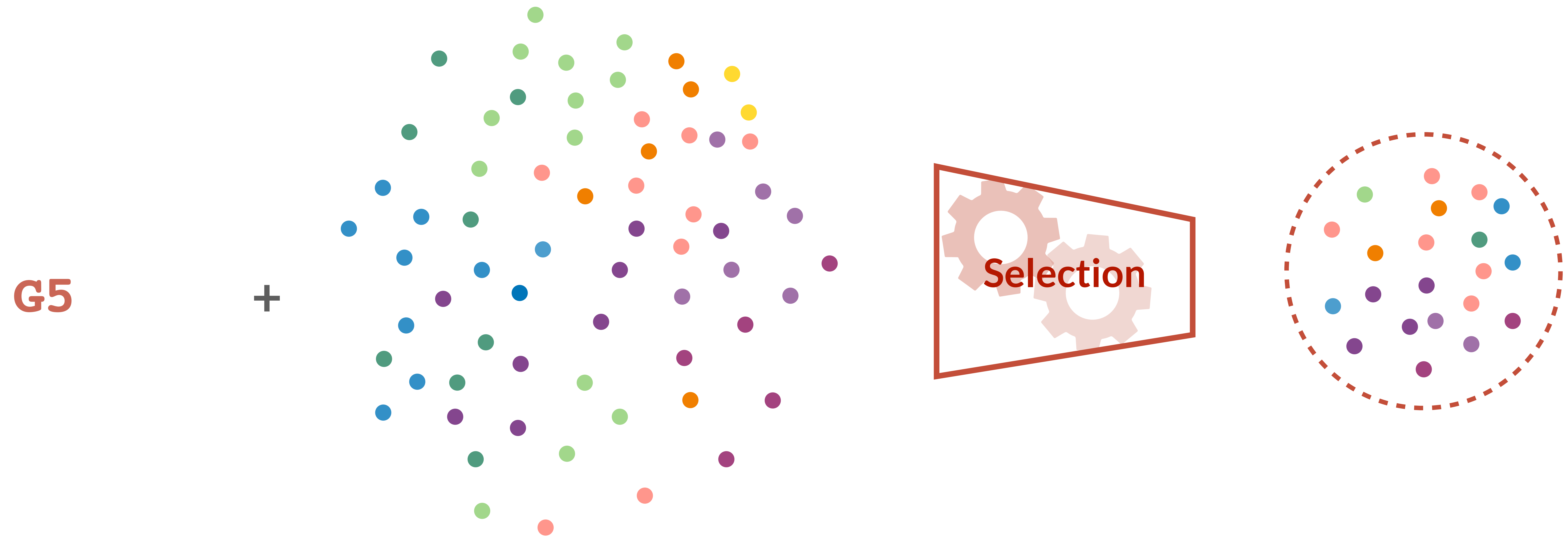
SENT

META

BOOT

GMM

LDA



TARGET
(no data)

NON-TARGET
(annotated)

PROXY
(annotated)

SWL 💬 SA 📖 KPV 📖 TA 📰 GL 📰 YUE 💬 CKT 💬 FO W TE 🪄 MYV 📖 QHE 📡 QTD 💬

TARGET

RAND

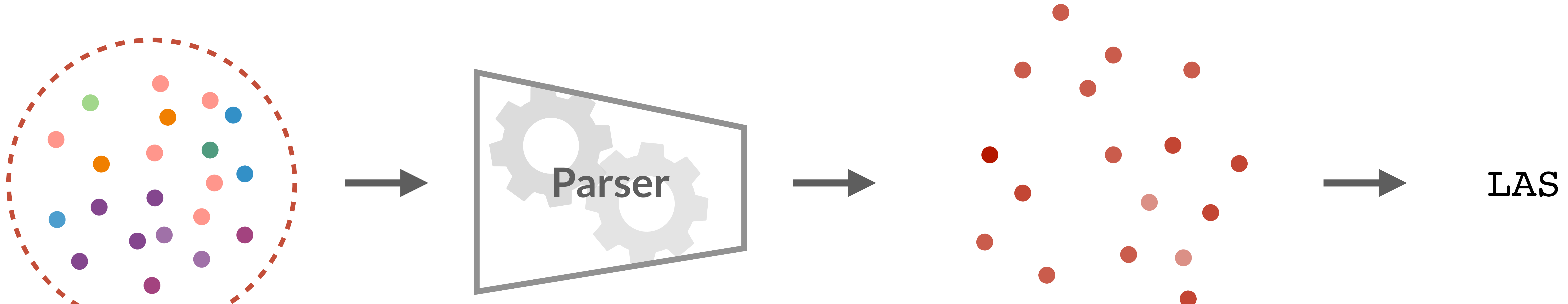
SENT

META

BOOT

GMM

LDA



Dozat & Manning (2017)
van der Goot et al. (2021)

PROXY
(annotated)

TARGET
(unannotated)

	SWL 💬	SA 📕	KPV 📕	TA 📰	GL 📰	YUE 💬	CKT 💬	FO W	TE 🪄	MYV 📕	QHE 📶	QTD 💬	Ø
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND	3.7	**24.8**	10.9	50.7	77.7	33.3	15.5	61.9	67.7	20.0	**27.0**	44.6	36.5
SENT	3.6	23.7	13.7	47.9	77.6	35.8	16.4	62.5	68.1	**22.9**	26.5	42.8	36.8
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1
BOOT	5.2	21.8	*21.1	49.4	76.7	*49.9	18.4	*66.3	65.6	19.5	14.8	43.8	37.7
GMM	4.9	22.9	*20.9	* **51.5**	**77.8**	* **49.9**	* **19.8**	*68.3	67.9	20.2	15.1	**45.4**	**38.7**
LDA	**6.6**	23.7	* **22.3**	49.2	77.0	*49.4	*19.1	* **68.3**	* **68.6**	20.5	15.1	44.7	**38.7**

	SWL 💬	SA 📖	KPV 📖	TA 📰	GL 📰	YUE 💬	CKT 💬	FO W	TE 🪄	MYV 📖	QHE ⌁	QTD 💬	⌀
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND	3.7	**24.8**	10.9	50.7	77.7	33.3	15.5	61.9	67.7	20.0	**27.0**	44.6	36.5
SENT	3.6	23.7	13.7	47.9	77.6	35.8	16.4	62.5	68.1	**22.9**	26.5	42.8	36.8
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1
BOOT	5.2	21.8	*21.1	49.4	76.7	*49.9	18.4	*66.3	65.6	19.5	14.8	43.8	37.7
GMM	4.9	22.9	*20.9	***51.5**	**77.8**	***49.9**	***19.8**	*68.3	67.9	20.2	15.1	**45.4**	**38.7**
LDA	**6.6**	23.7	***22.3**	49.2	77.0	*49.4	*19.1	***68.3**	***68.6**	20.5	15.1	44.7	**38.7**

SWL 💬	SA 📖	KPV 📖	TA 📰	GL 📰	YUE 💬	CKT 💬	FO W	TE 🪄	MYV 📖	QHE 📶	QTD 💬	Ø

TARGET

RAND

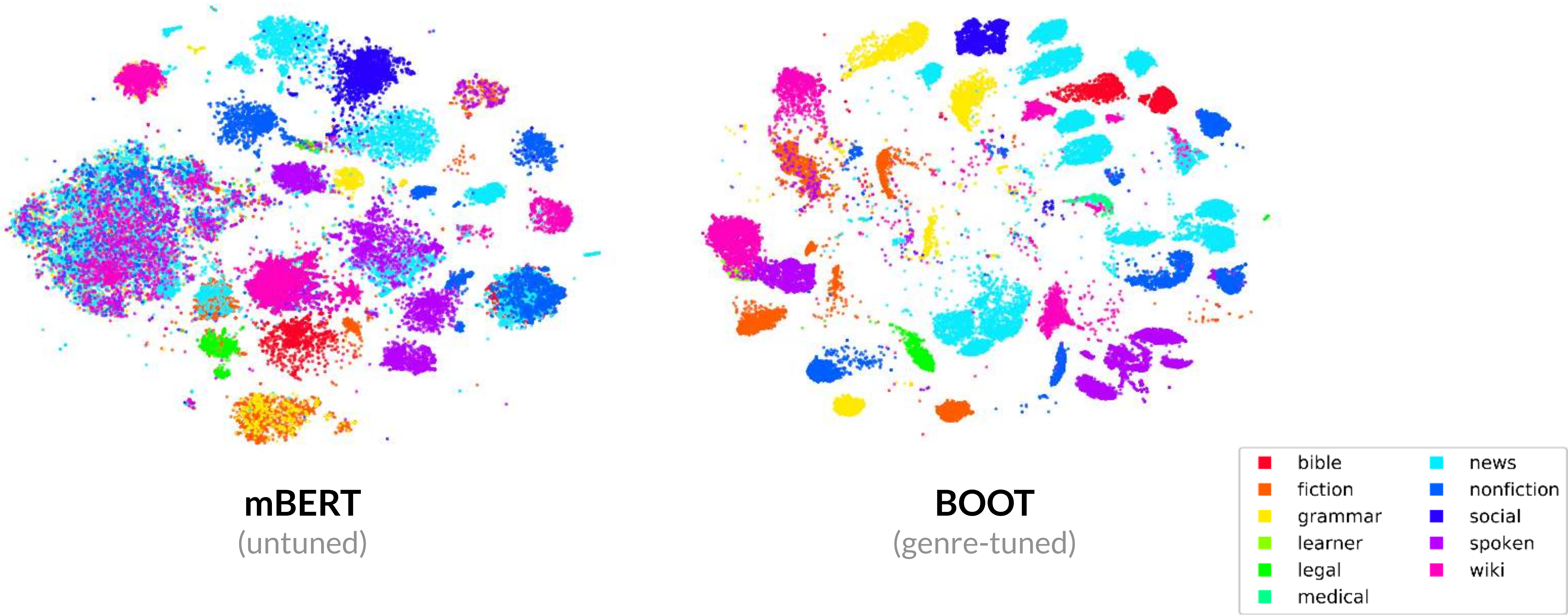
SENT

META

BOOT

GMM

LDA



mBERT
(untuned)

BOOT
(genre-tuned)

	SWL 💬	SA 📕	KPV 📕	TA 📰	GL 📰	YUE 💬	CKT 💬	FO W	TE 🪄	MYV 📕	QHE 📶	QTD 💬	Ø
TARGET	28.0	15.7	13.4	64.1	80.9	—	—	49.6	83.6	—	62.7	55.0	50.3
RAND	3.7	**24.8**	10.9	50.7	77.7	33.3	15.5	61.9	67.7	20.0	**27.0**	44.6	36.5
SENT	3.6	23.7	13.7	47.9	77.6	35.8	16.4	62.5	68.1	**22.9**	26.5	42.8	36.8
META	6.5	24.3	10.2	50.4	76.6	31.2	11.6	61.2	64.9	20.4	9.42	42.6	34.1
BOOT	5.2	21.8	*21.1	49.4	76.7	*49.9	18.4	*66.3	65.6	19.5	14.8	43.8	37.7
GMM	4.9	22.9	*20.9	* **51.5**	**77.8**	* **49.9**	* **19.8**	*68.3	67.9	20.2	15.1	**45.4**	**38.7**
LDA	**6.6**	23.7	* **22.3**	49.2	77.0	*49.4	*19.1	* **68.3**	* **68.6**	20.5	15.1	44.7	**38.7**
van der Goot et al. (2021)		16.5	11.7			32.7	15.3	62.7					

Conclusion

BOOT GMM LDA

Genre is a valuable signal for parsing unseen, low-resource targets

Thank You

Have a great EMNLP!

IT-UNIVERSITETET I KØBENHAVN / north