Tackling scarce & biased data for more inclusive NLP

Barbara Plank

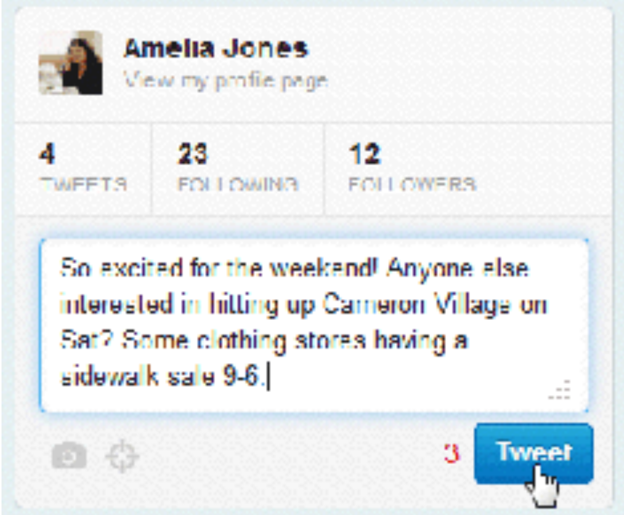
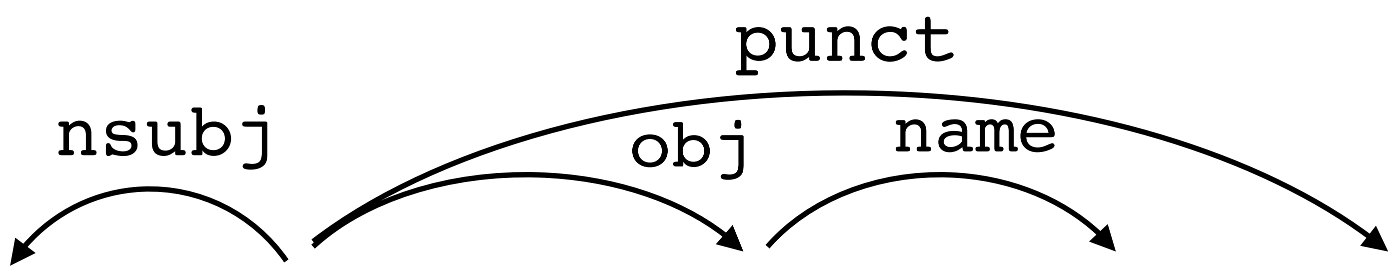
(collaborators and lab member contributions highlighted throughout)

The 19th Annual Workshop of the
Australasian Language Technology Association
December 9, 2021

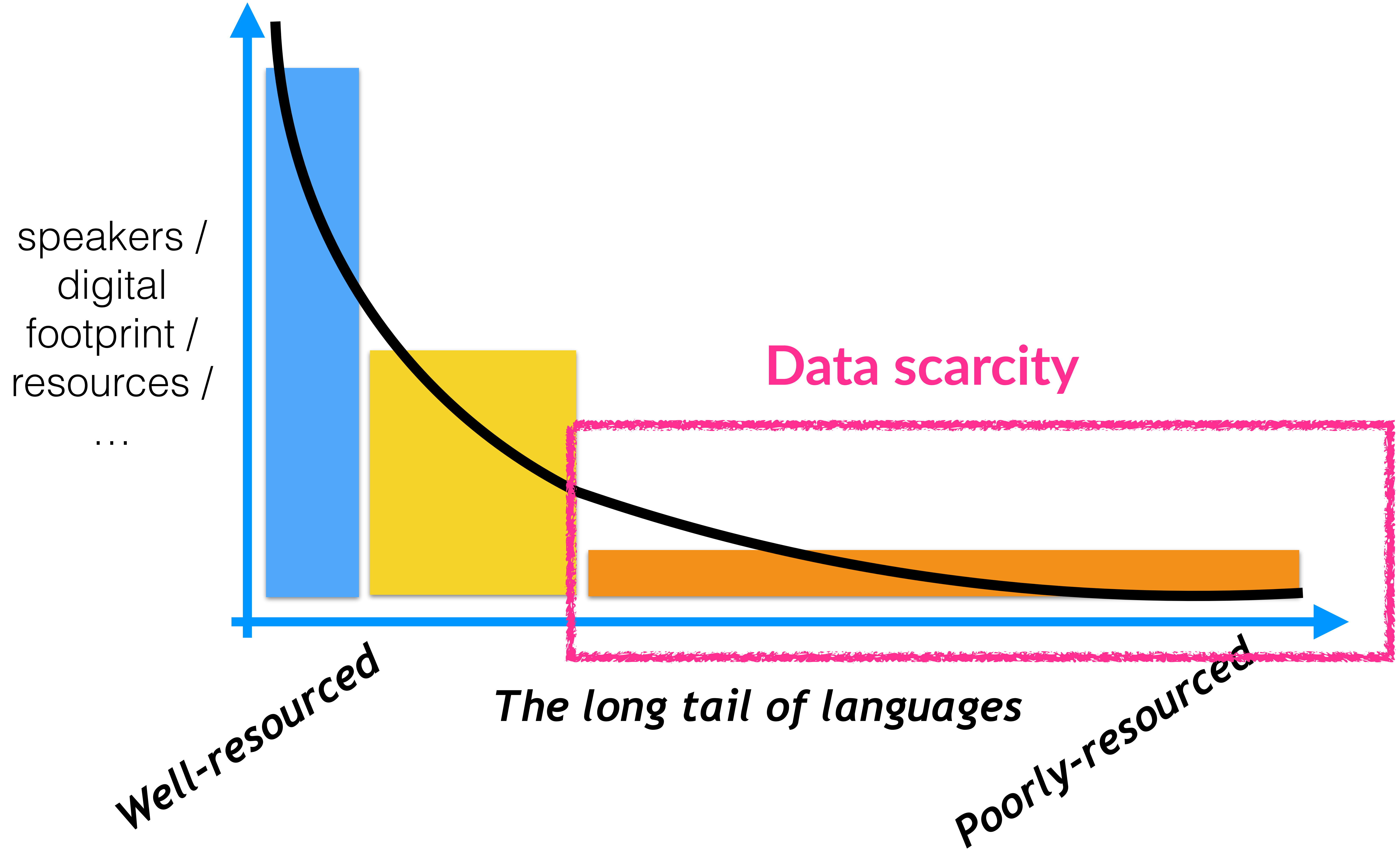
IT-UNIVERSITETET I KØBENHAVN / north

NLP Tasks: Learning from <X,Y>

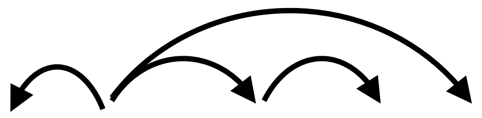
Human-annotated examples ➡ Time-intensive
➡ Expensive

X (input)	Y (output)
	
I like Vince Gilligan .	
Citigroup has taken over EMI,	**CompanyAcquired(Citigroup, EMI)**

Lack of Resources

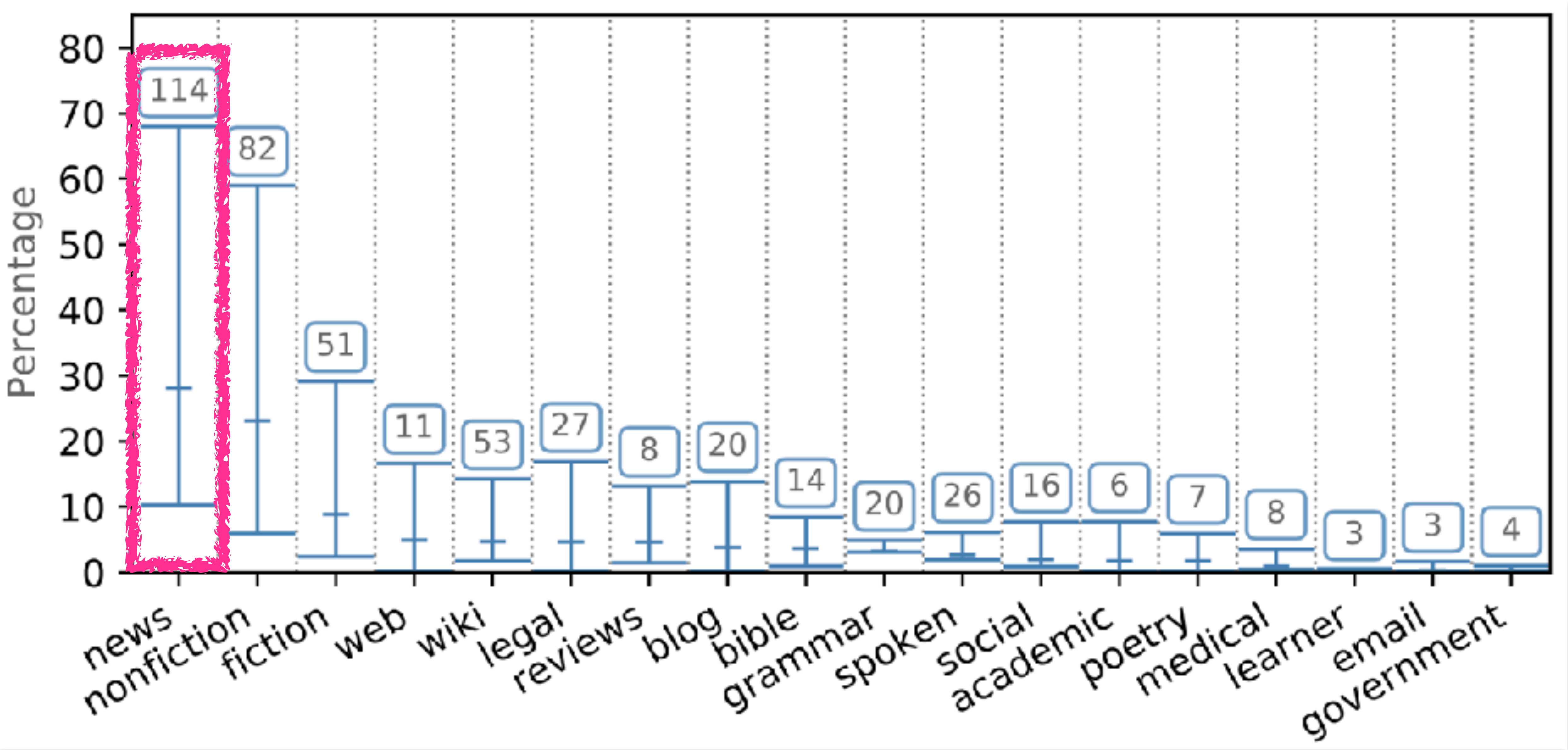


Bias of Resources



Universal Dependencies (UD)

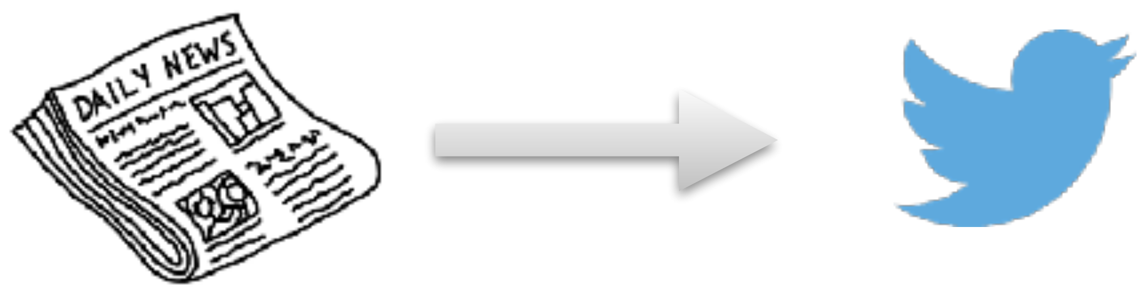
Selection bias:
Newswire

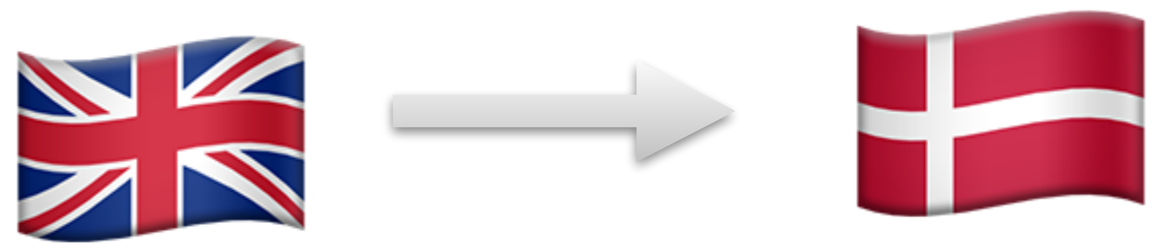


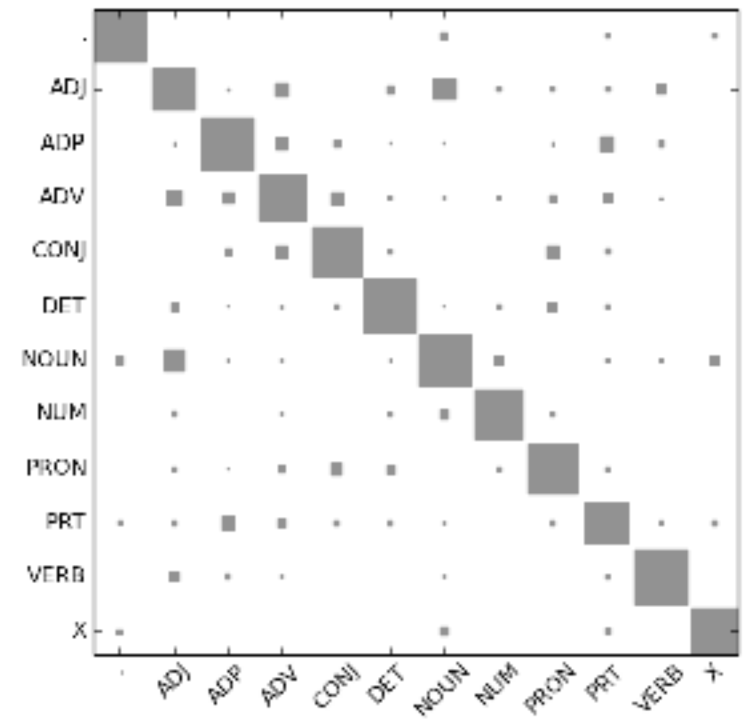
Müller-Eberstein, van der Goot, Plank (EMNLP 2021)

Challenges in Inputs, and Outputs

X (input space)	Y (output space)
Input distribution shifts Data changes (e.g. across genres, across languages)	**Inherent disagreement** Humans often do not agree on what's the correct label







(Plank et al., ,2014; Pavlick & Kwiatkowski 2019)

There's heavy rain

It's raining heavily

It's raining cats and dogs

It's a frog strangler

Heavy precipitation in this area

Starkregen in dieser Region

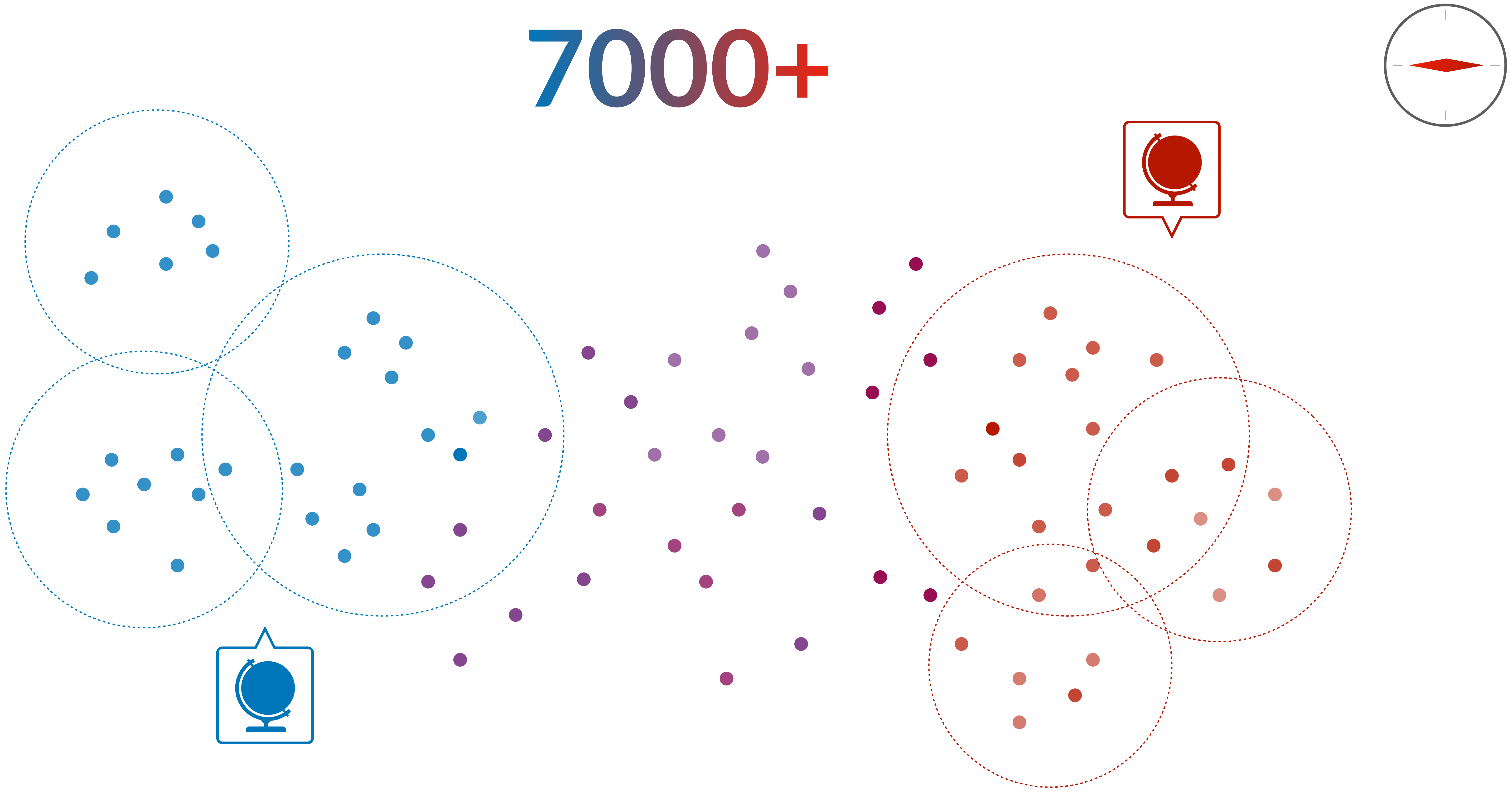
Es regnet sehr stark

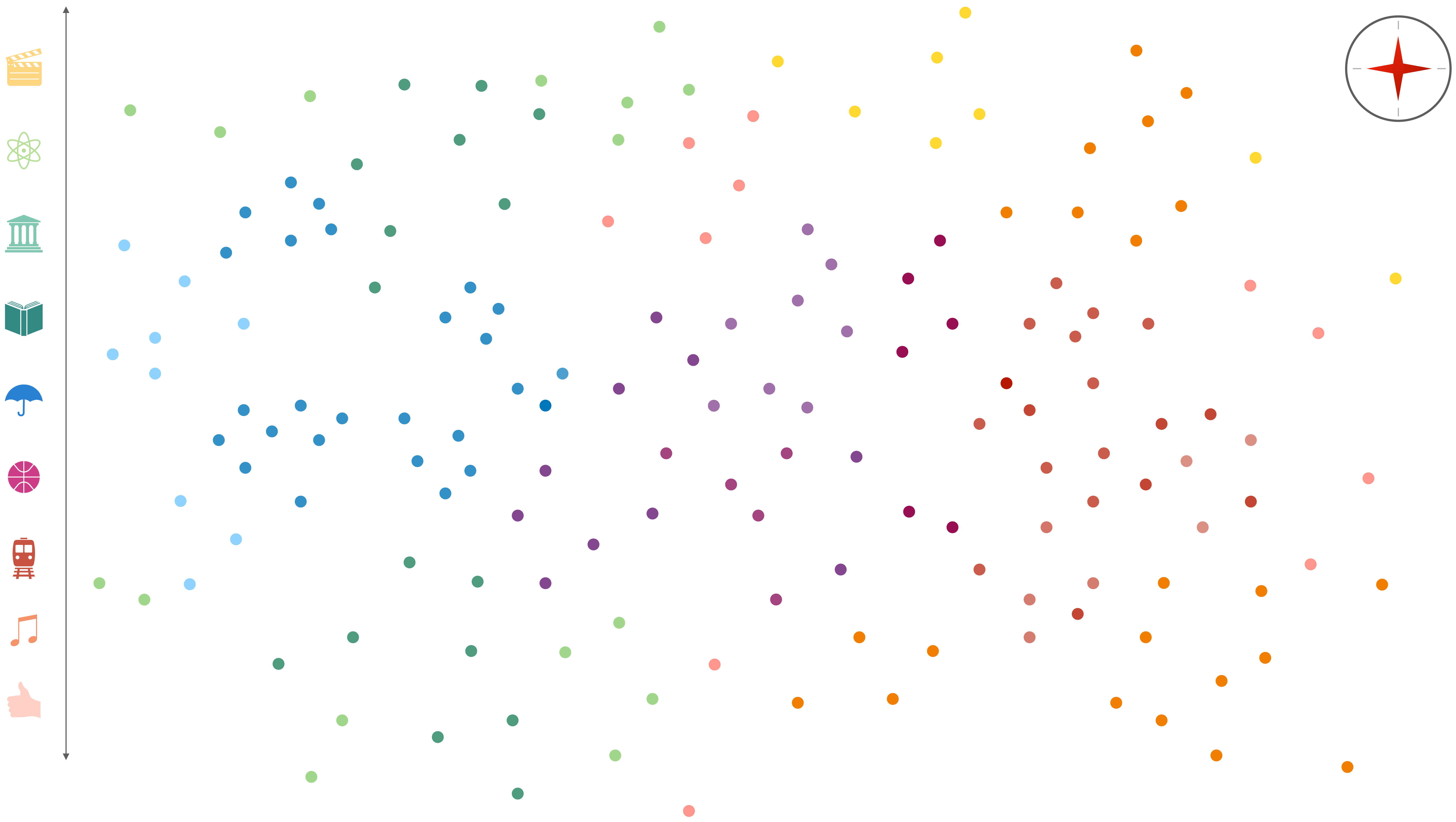
Es schüttet aus Kübeln

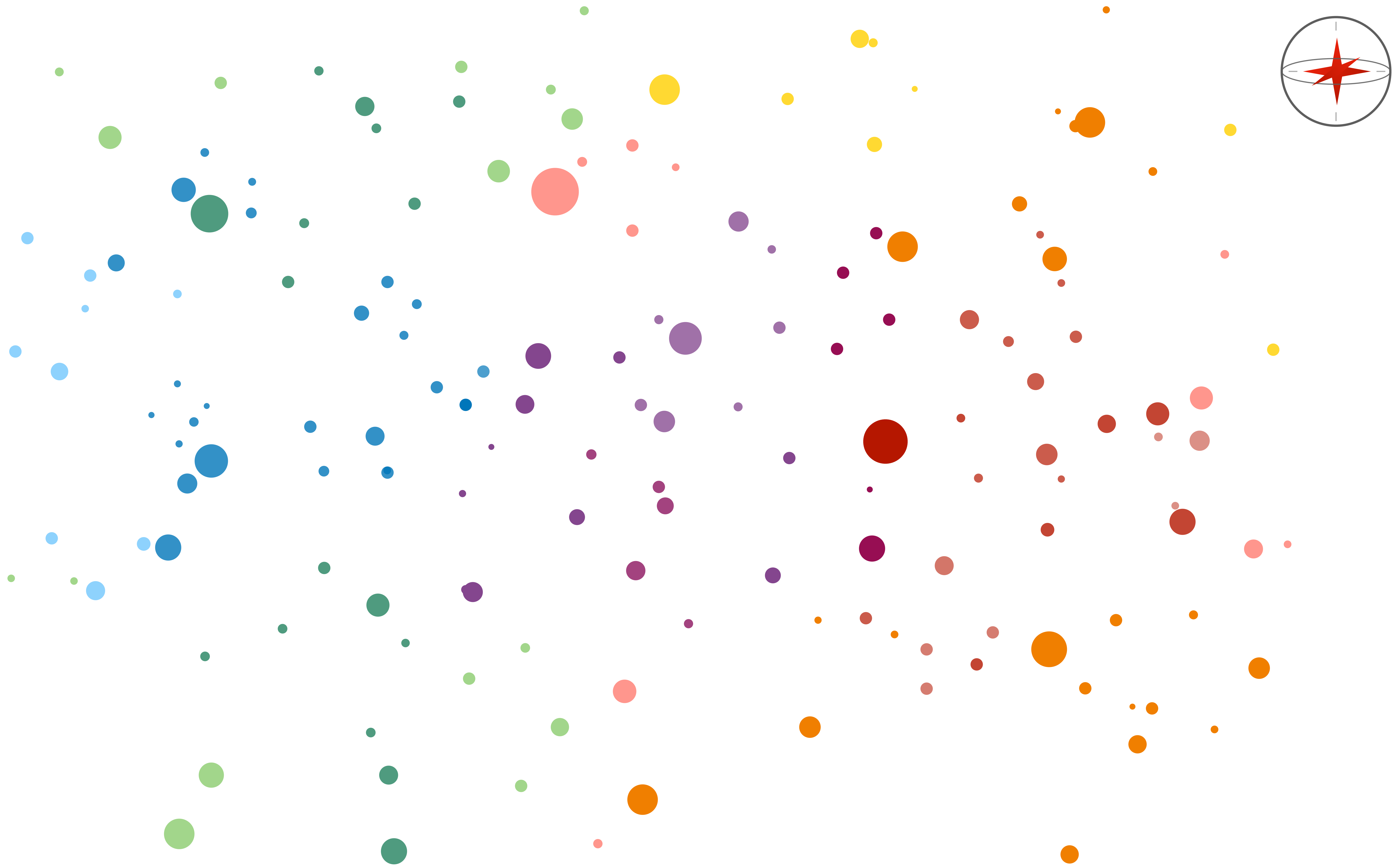
Es schüttet aus Eimern

Was für ein Wolkenbruch

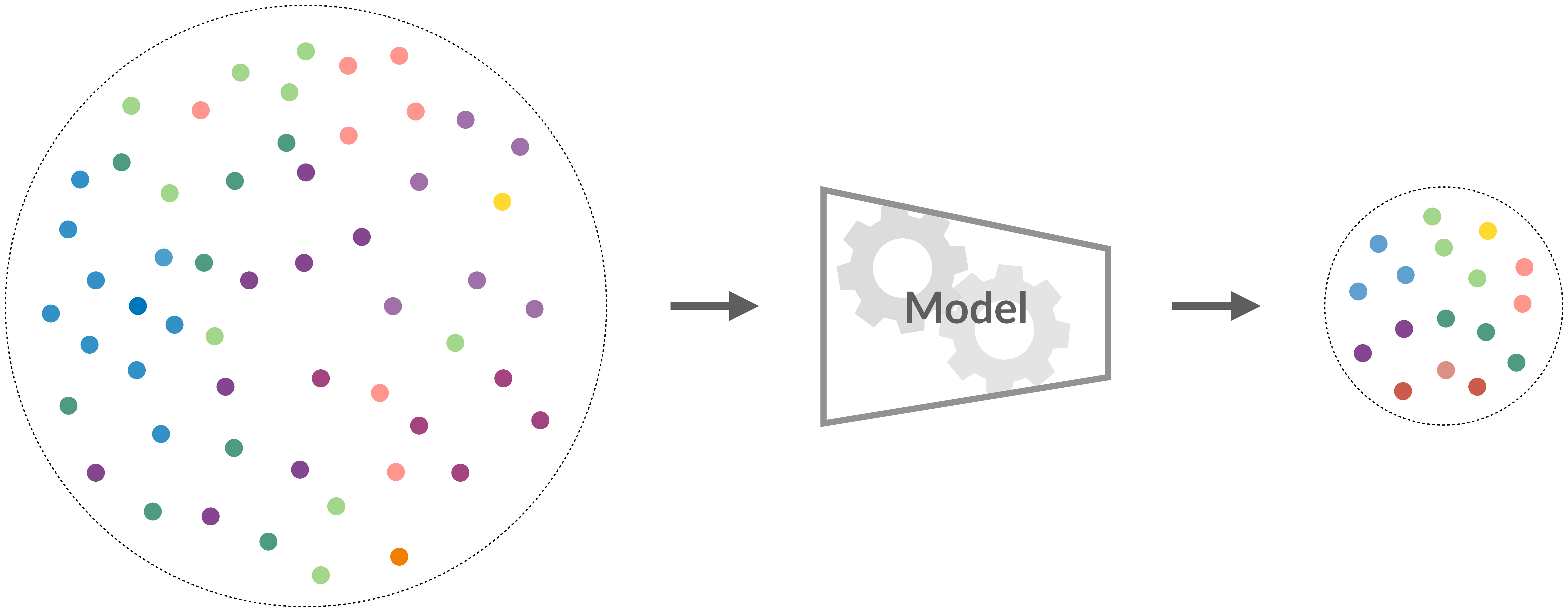
7000+



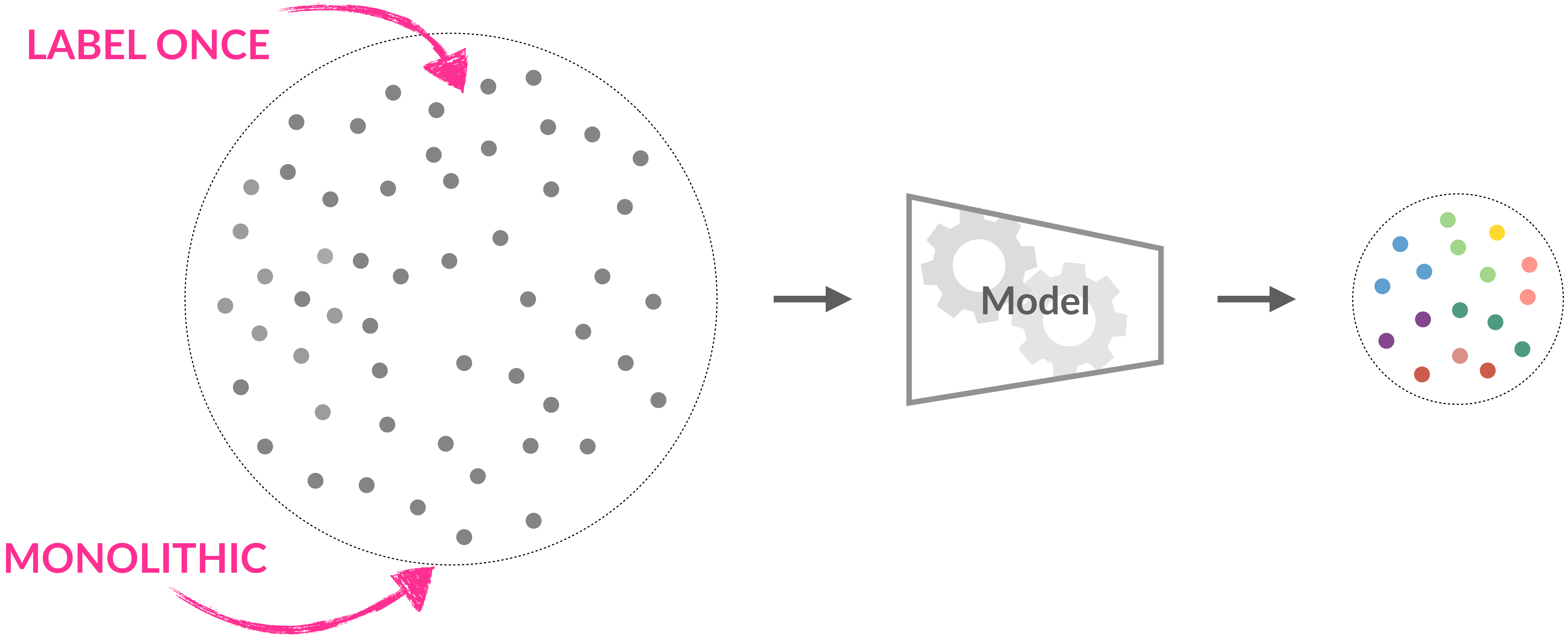




Variety Space


Model

NLP today is often "monolithic processing"



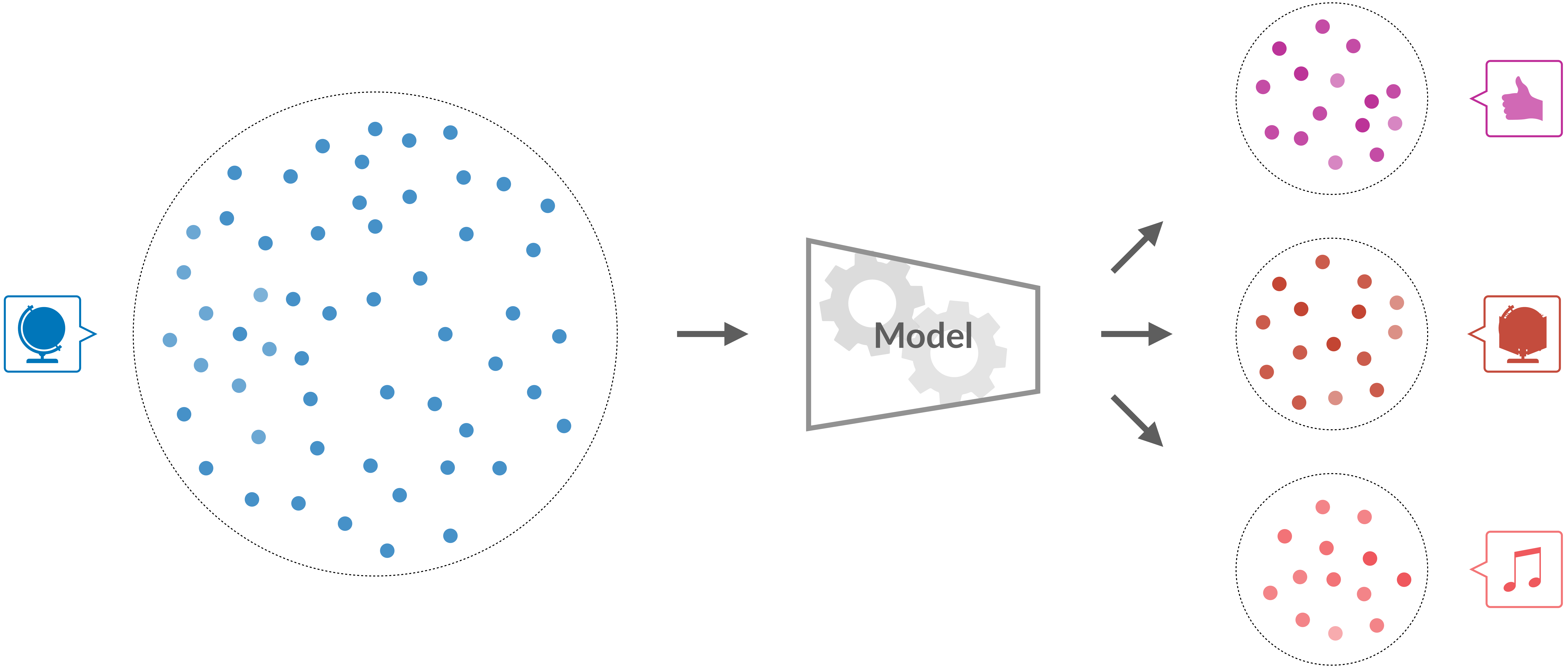
A lot remains to be done, to create inclusive and robust NLP

Roadmap for the Rest of the Talk

- Introduction
  - Scarce and Biased Data in NLP
  - The Variety Space
- 3 broad research goals and selected case studies

How can we create more inclusive NLP?

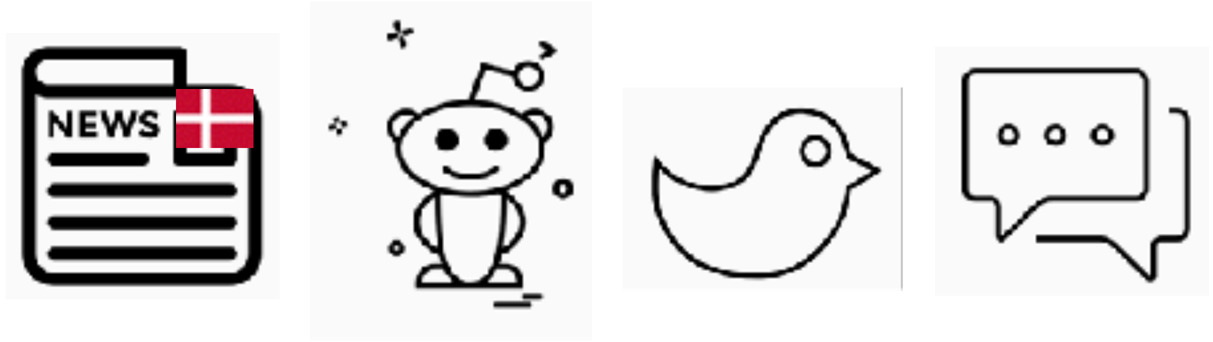
- Creation of dedicated in-language resources
- Transfer from better-resourced languages


Model

DaN+

DaN+ corpus

- **Da**nish **N**ested **N**amed Entities and Lexical **N**ormalization
  (Plank, Nørgaard Jensen, van der Goot, 2020 COLING)
  - **Nested** NER corpus for Danish
    - [[Danmarks]**LOC** Radio]**ORG** (nested, genitive)
    - De [københavnske]**LOCderiv** gader (location adjective)
    - [pro-hongkong]**LOCpart** (parts of tokens)
  - Over **multiple target domains**



Paper, Data, Code: https://www.aclweb.org/anthology/2020.coling-main.583.pdf

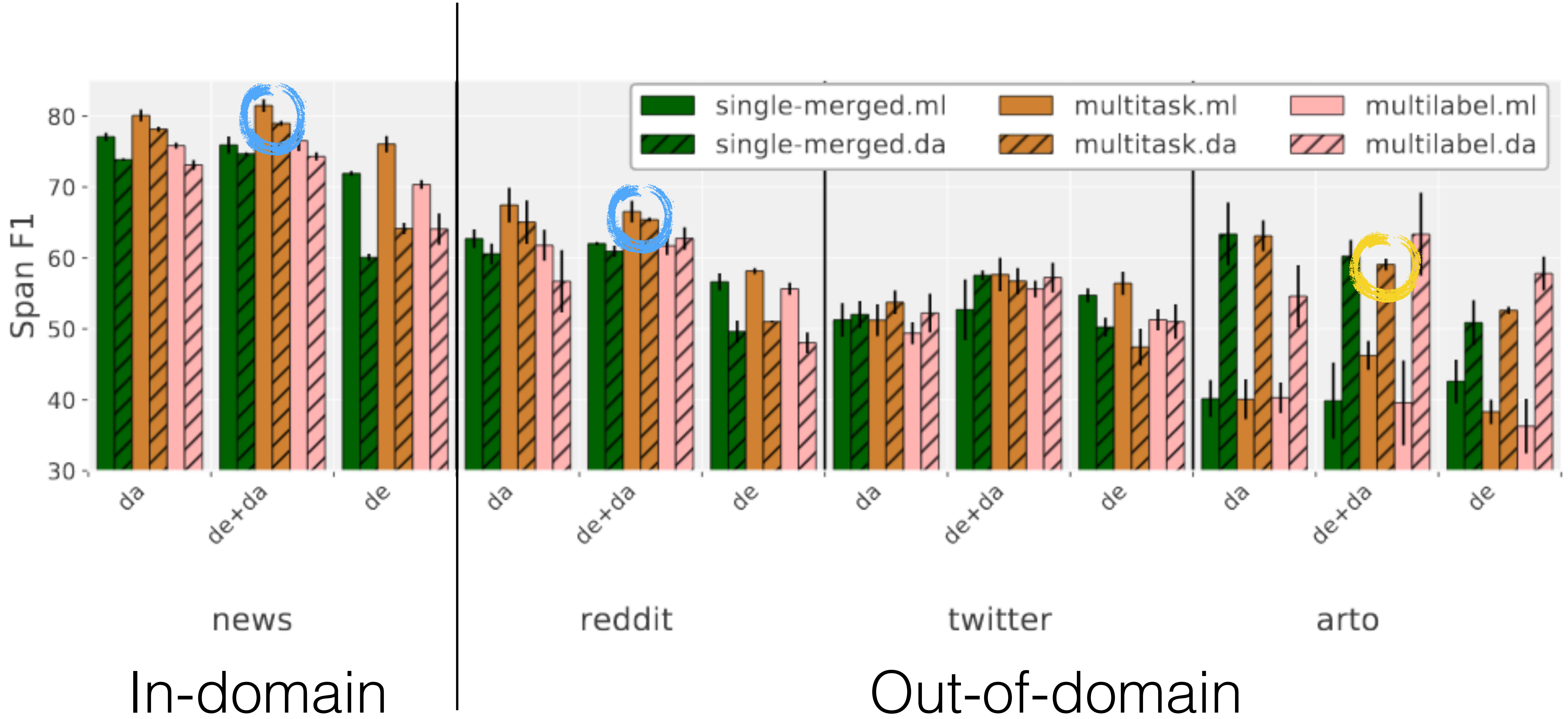
Danish Nested Named Entities and Normalization (DaN+)

GermEval (Belinkova et al., 2014)

DaN+

- UD-DDT (Danish UD)
- r/Denmark
- emotion words
- Arto (operated 1988-2006)

Results for Nested NER: Danish Bert (da) vs multilingual BERT (ml)



Takeaway: Domains shift matters & No free lunch - no best overall BERT variant

xSID

Languages in EU covered by voice assistants

https://www.globalme.net/blog/language-support-voice-assistants-compared/

*as of March, 2020

From Masked-Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-Shot Spoken Language Understanding

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanovic,
Alan Ramponi, Siti Orzya Khairunnisa, Mamoru Komachi, Barbara Plank






et al., NAACL 2021

Task: Slot and Intent Detection

I'd like to see the showtimes for Silly Movie 2.0 at the movie house

Intent: SearchScreeningEvent

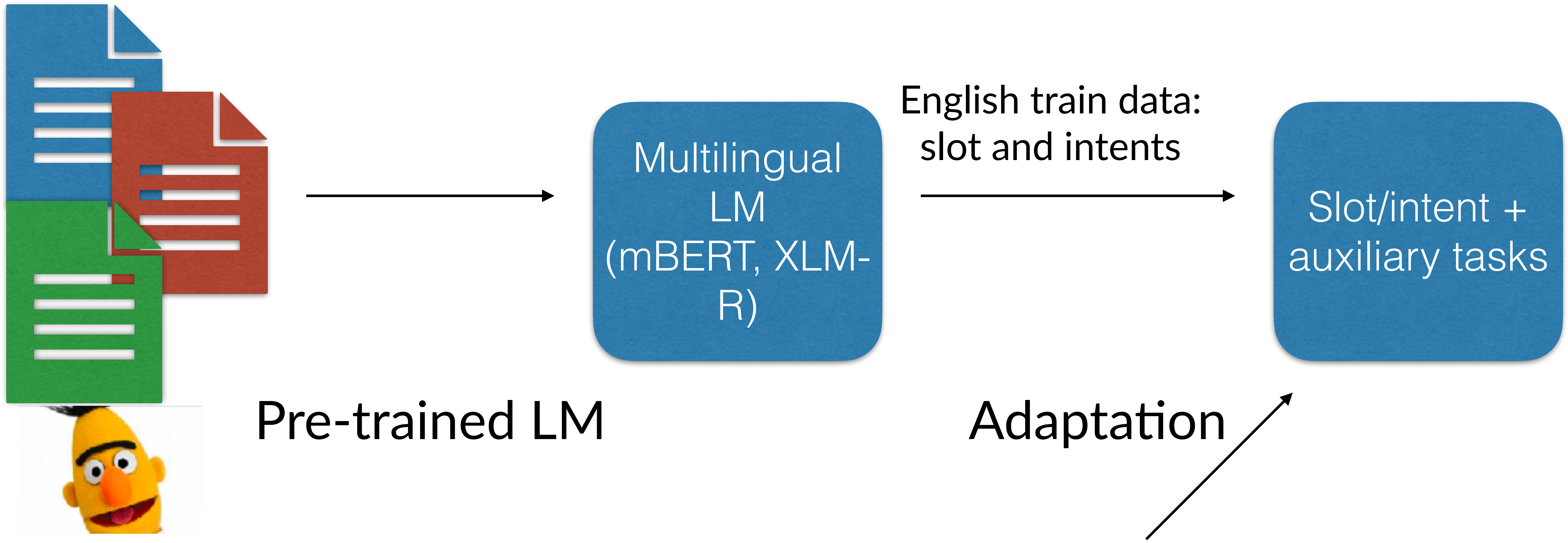
Task: Slot and Intent Detection

Slots:

I'd like to see the showtimes for Silly Movie 2.0 at the movie house

Intent: SearchScreeningEvent

Non-English Auxiliary Tasks



Can we improve zero-shot performance with auxiliary data from target languages? (3 tested: MT, Parsing, MLM)

Evaluation dataset: xSID

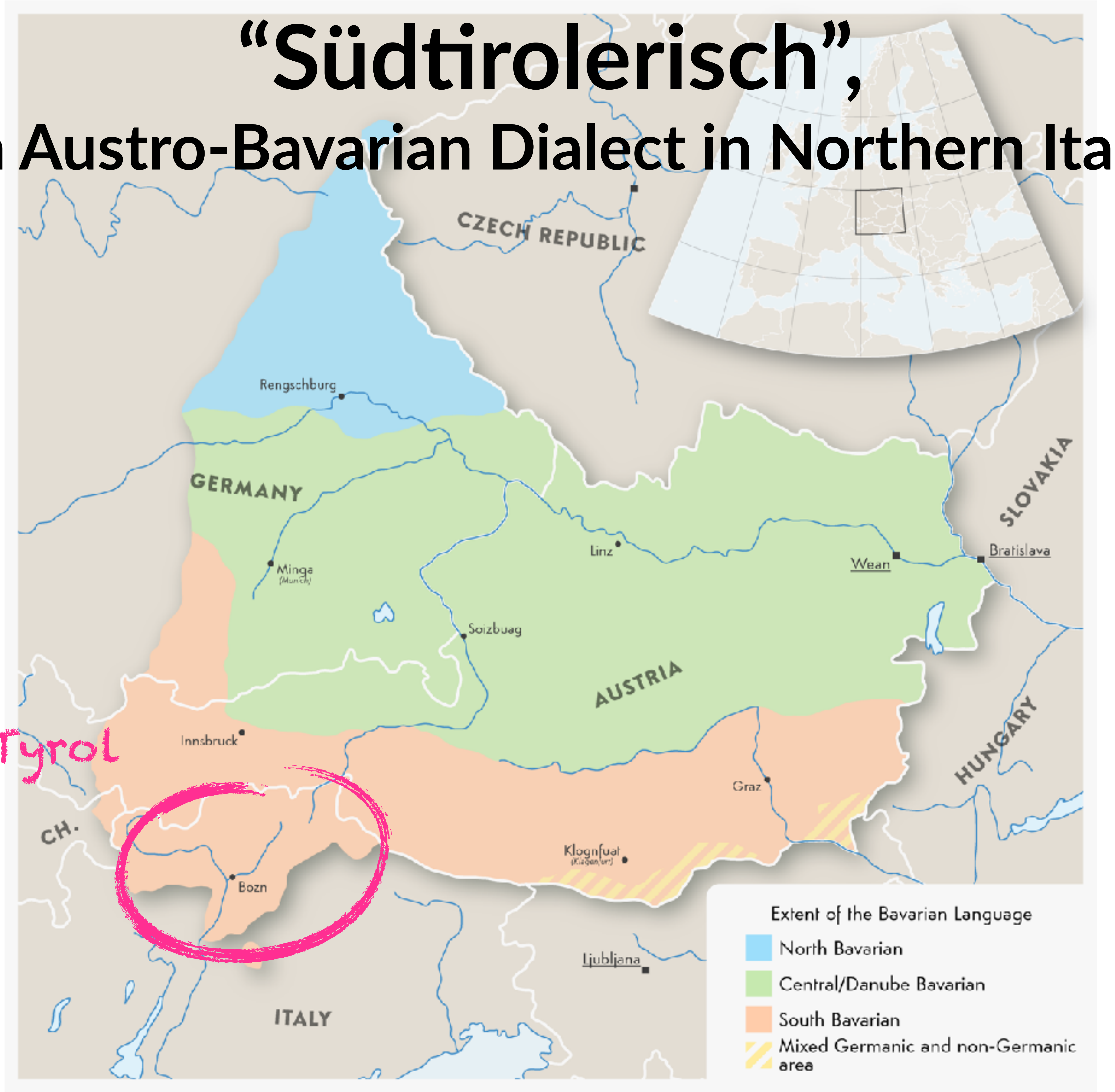
ar	أود أن أرى مواعيد عرض فيلم Silly Movie 2.0 في دار السينما
da	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biografen
de	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2.0 im Kino sehen
de-st	I mecht es Programm fir Silly Movie 2.0 in Film Haus sechn
en	I'd like to see the showtimes for Silly Movie 2.0 at the movie house
id	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di gedung bioskop
it	Mi piacerebbe vedere gli orari degli spettacoli per Silly Movie 2.0 al cinema
ja	映画館の Silly Movie 2.0 の上映時間を見せて。
kk	Мен Silly Movie 2.0 бағдарламасының кинотеатрда көрсетілім уақытын көргім келеді
nl	Ik wil graag de speeltijden van Silly Movie 2.0 in het filmhuis zien
sr	Želela bih da vidim raspored prikazivanja za Silly Movie 2.0 u bioskopu
tr	Silly Movie 2.0'ın sinema salonundaki seanslarını görmek istiyorum
zh	我想看 Silly Movie 2.0 在影院的放映

★ Data, code: https://bitbucket.org/robvanderg/xsid

Short-cut: MLM aux task was best for slots

A closer look at a low-resource German dialect

“Südtirolerisch”,
an Austro-Bavarian Dialect in Northern Italy



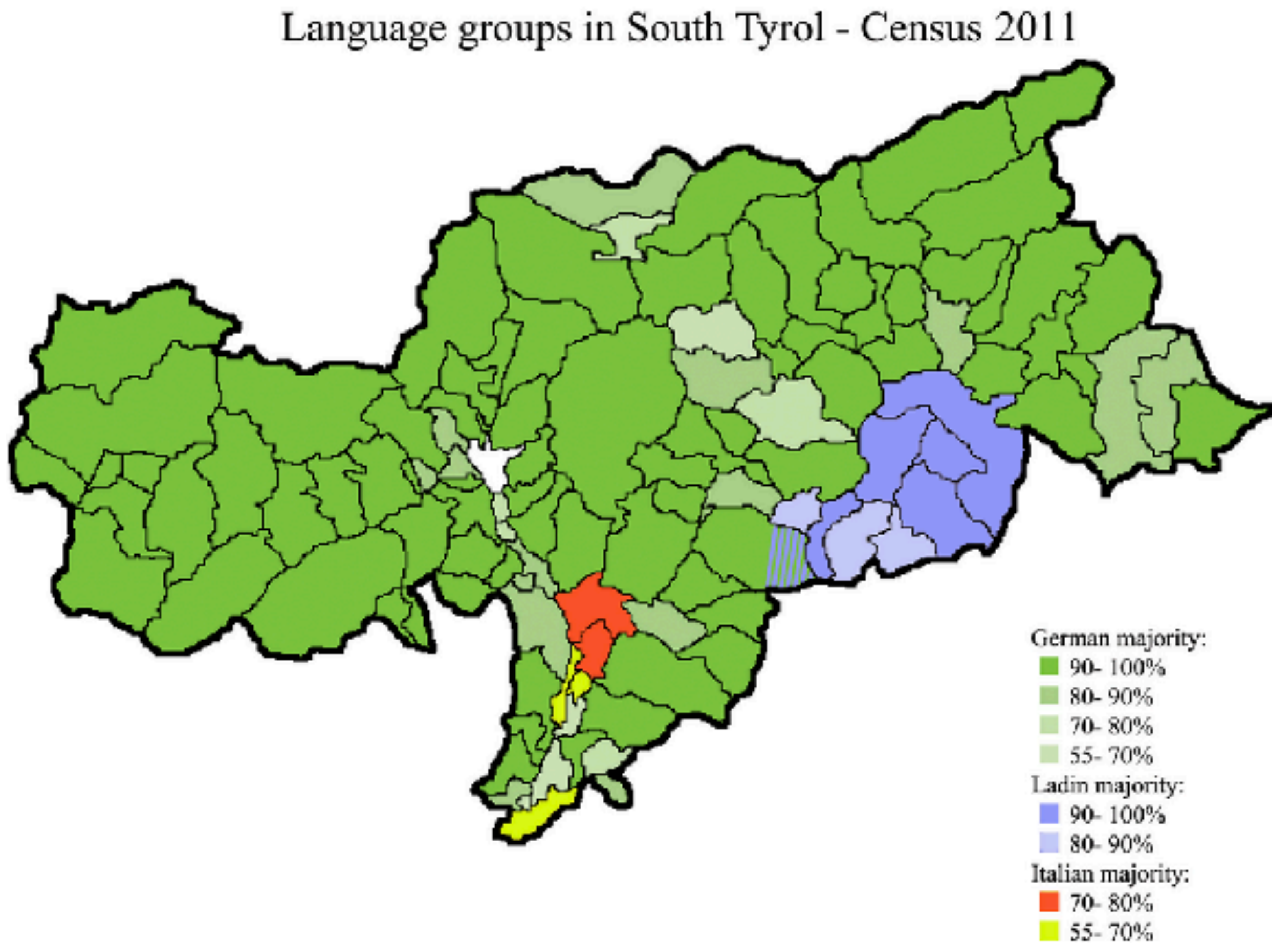
South Tyrol

https://upload.wikimedia.org/wikipedia/commons/thumb/a/ac/Austro_Bavarian_Languages-01.png/1024px-Austro_Bavarian_Languages-01.png

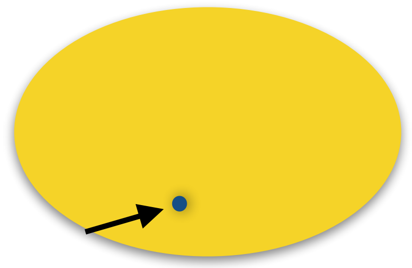
Languages in South Tyrol

- German dialect ("Südtirolerisch"), northernmost Italian province of Bozen-Bolzano with ~0.5M inhabitants
- approx. 62% German speakers, 24% Italian, 4% Ladin, 10% other native languages
- No common orthographic standard
- Lexical influence of other official languages (Italian, Ladin)
  - Example: "Hosch is **patent** schun gemocht?"

    [patent (neut.)=
    ital. la patente (fem.),
    dt. der Führerschein (masc.),
    eng. driver's license]

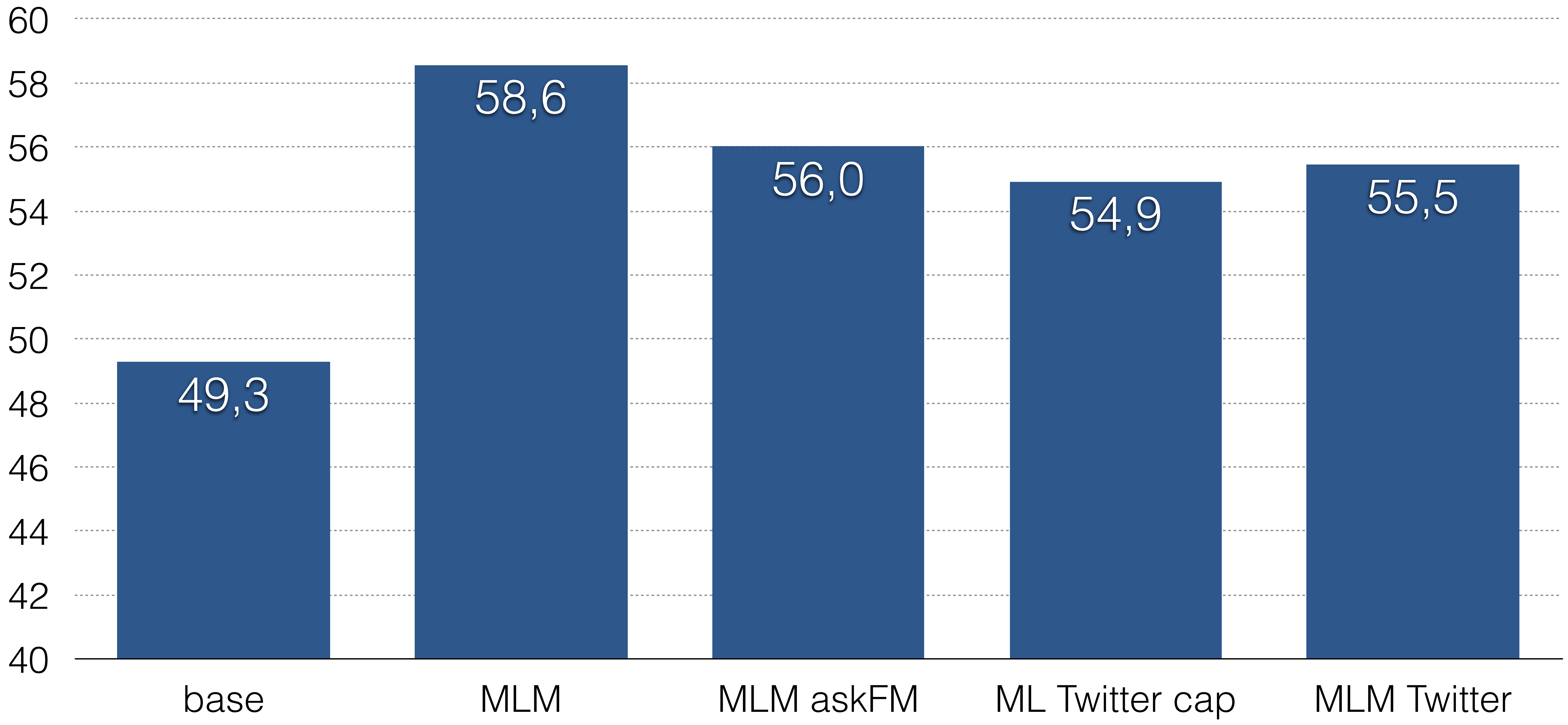


X Sparsity



- Hard to get access to unlabeled data
  - Social media (Twitter): highly mixed data, switch to “high” languages, no “dialect” identifier exists
  - AskFM: short Q&A posts, more dialectal
- Are small amounts of unlabeled data still useful to improve zero-shot performance?

De-ST: #sentences for MLM



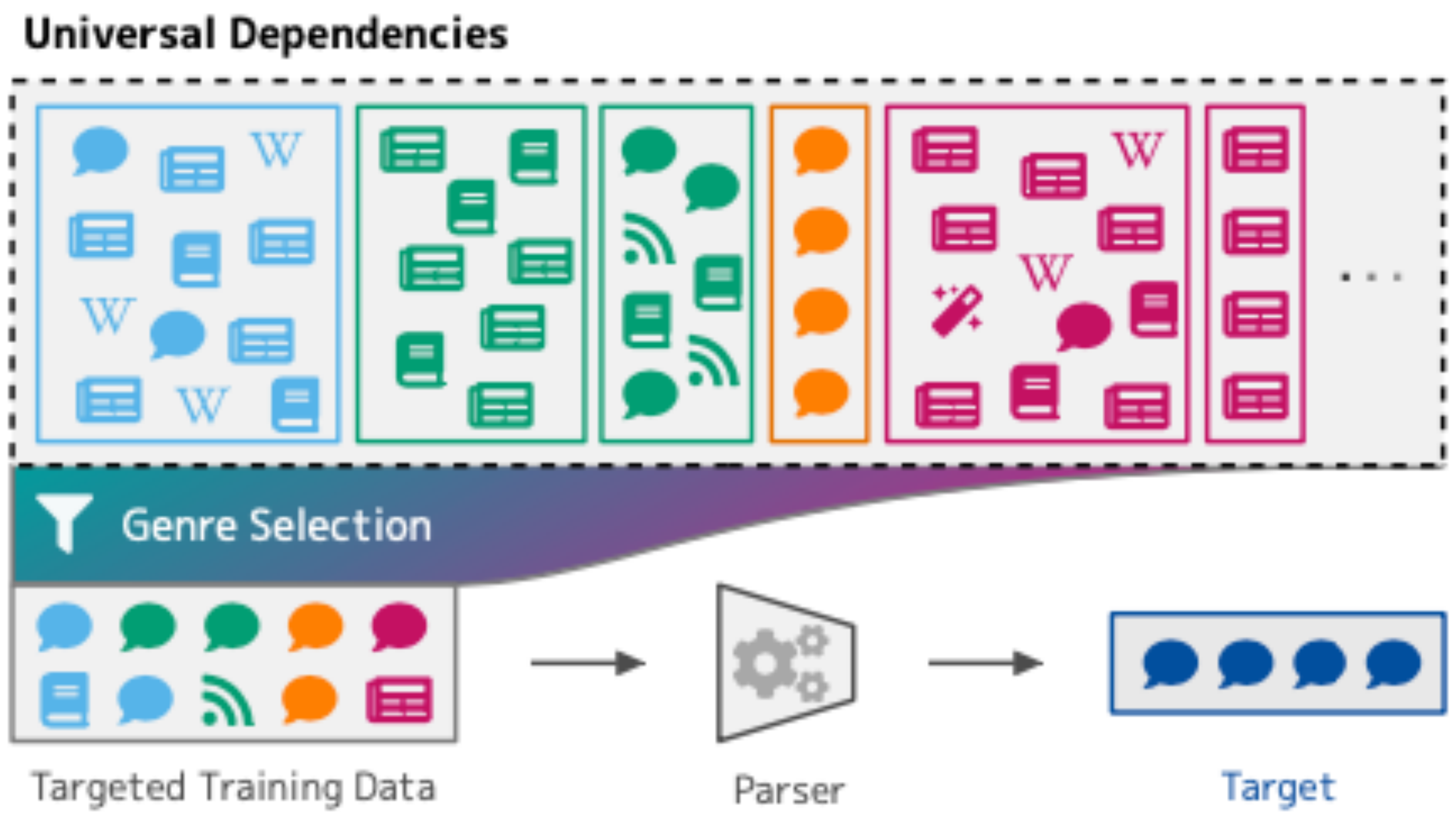
mBERT 30k 6.5k 6.5k 23.5

How can we create more efficient NLP systems?

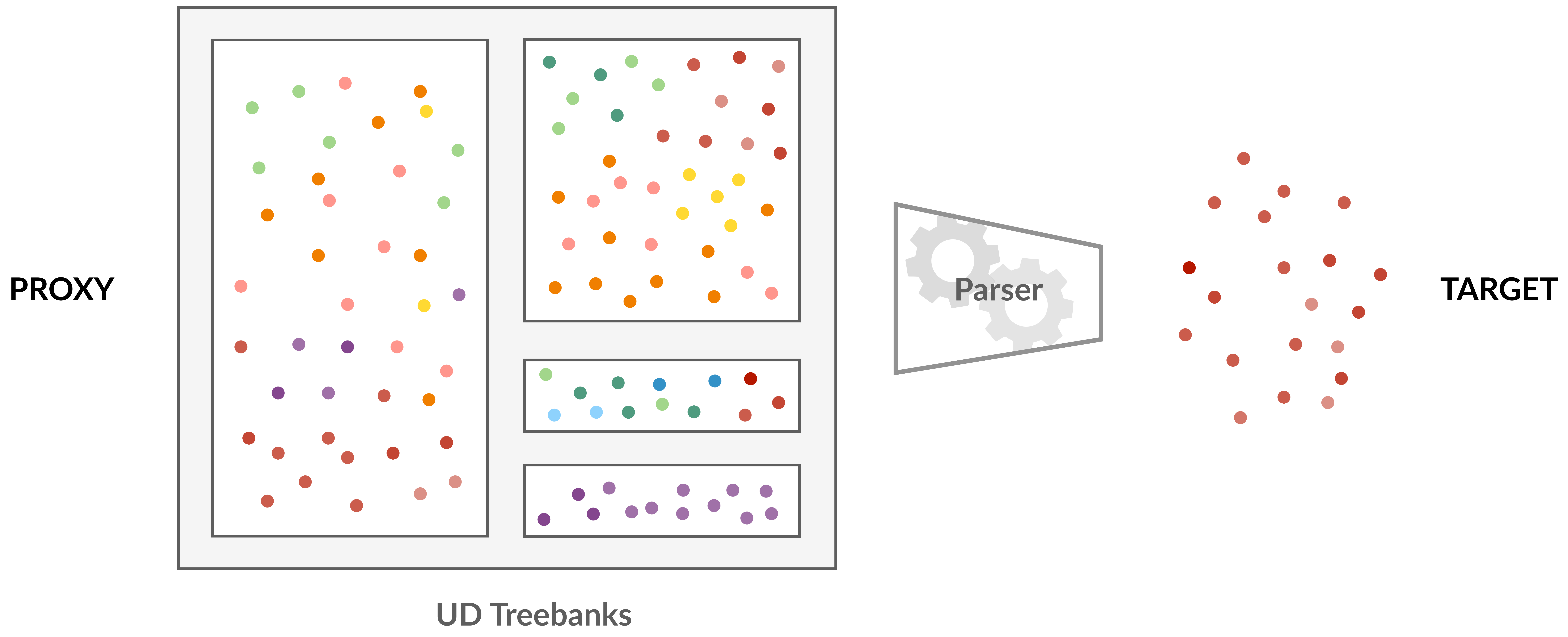
- Data Selection
- Weak Supervision

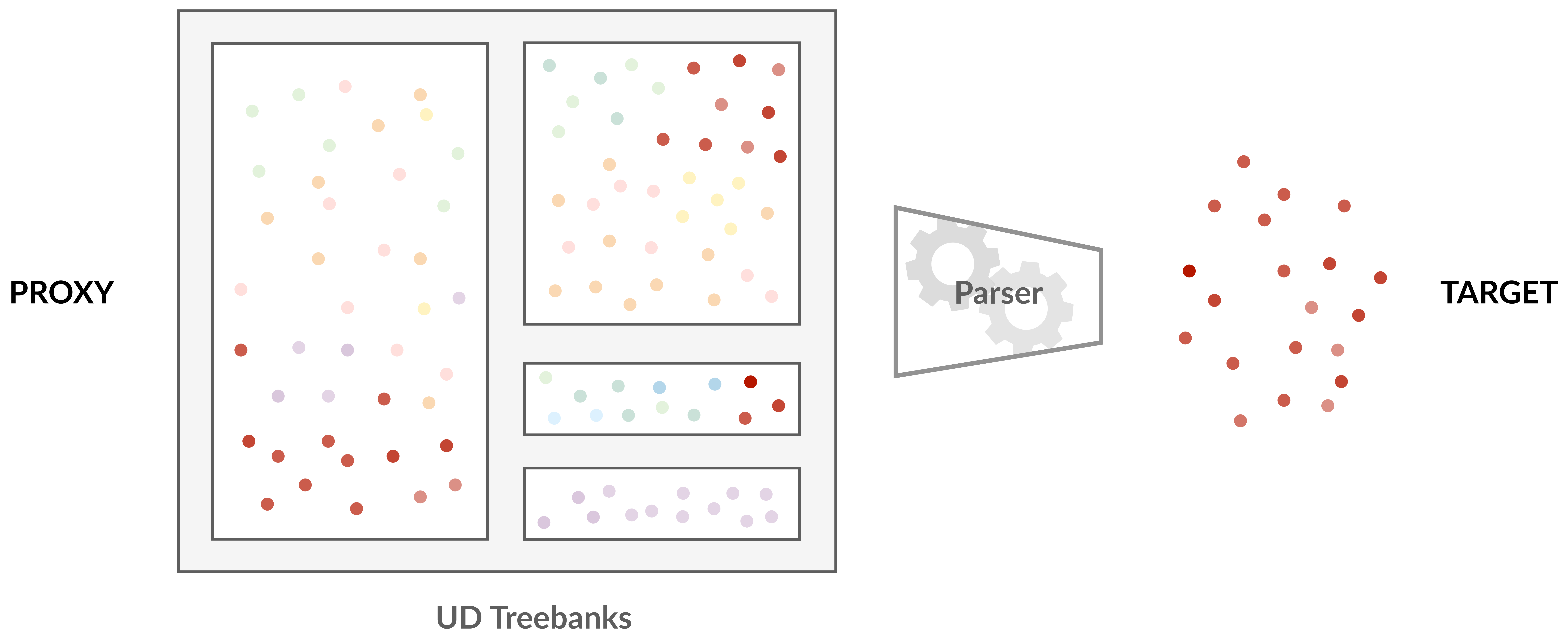
Data Selection for Low-resource Parsing

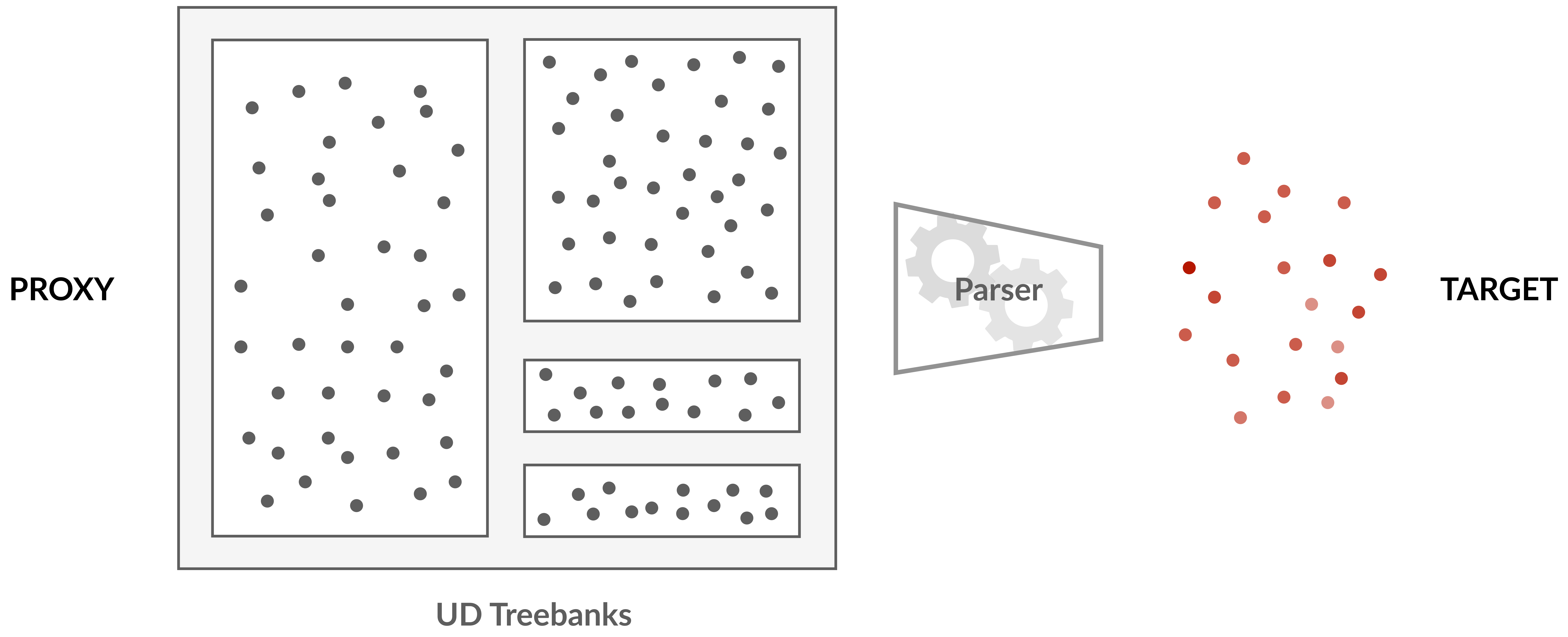
- **Problem:** a single parser trained on 100+ languages
  - suboptimal ("curse of multilinguality")
  - training is inefficient
  - practitioner: difficulty of choosing appropriate training material



Müller-Eberstein, van der Goot, Plank (EMNLP 2021)


PROXY
UD Treebanks
Parser
TARGET


PROXY
UD Treebanks
Parser
TARGET


PROXY
UD Treebanks
Parser
TARGET

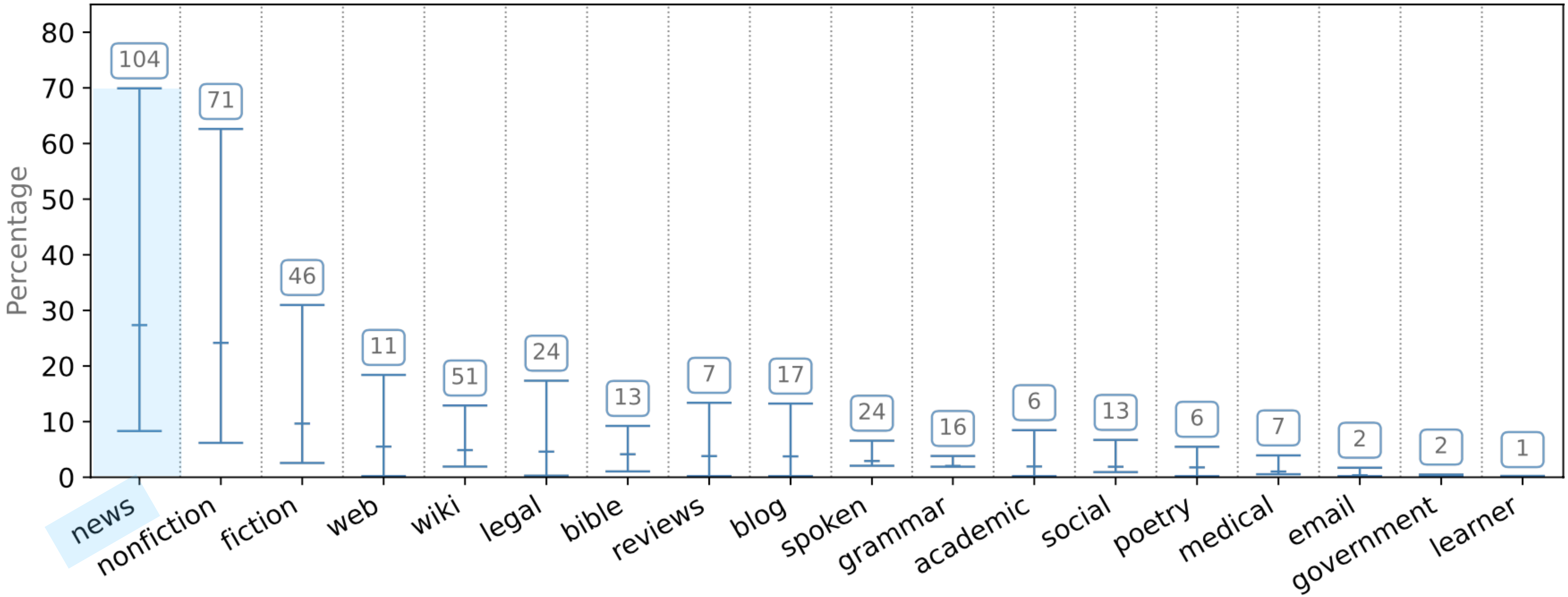
Genre as Weak Supervision

Domain $\boxed{\textbf{Genre}}$ Register

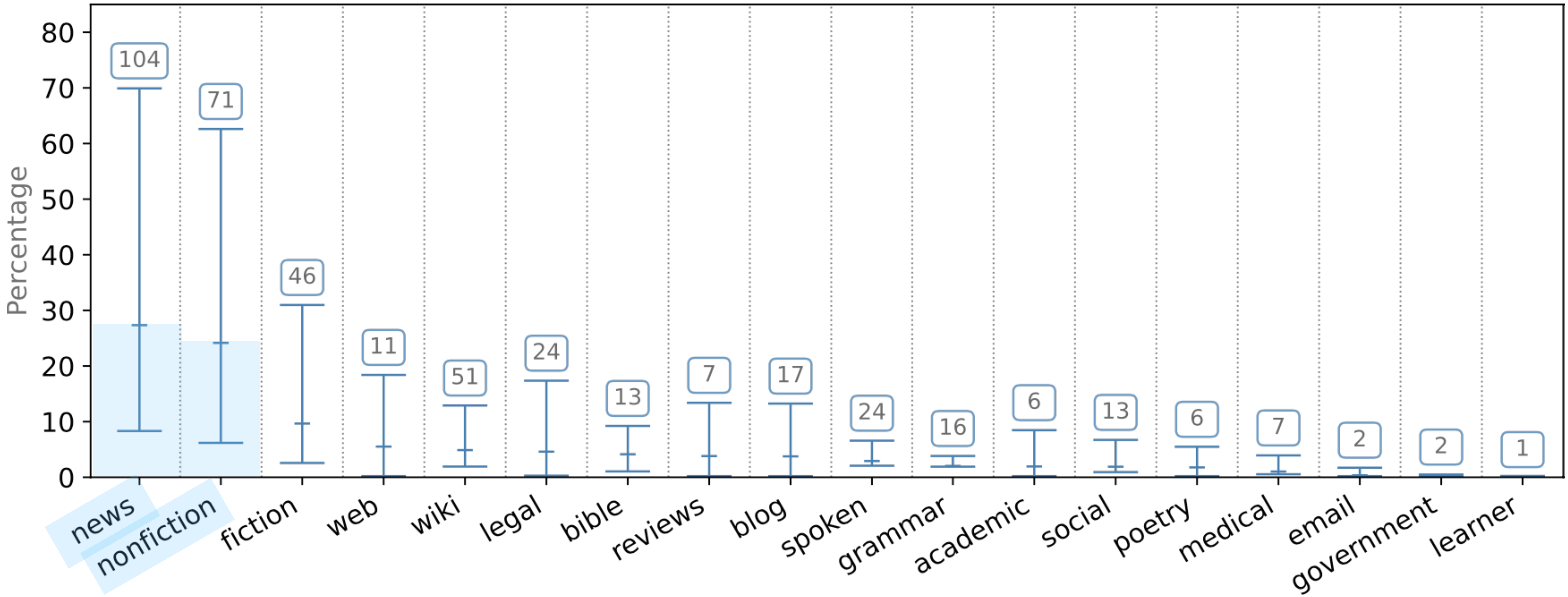
Kessler et al. (1997); Lee (2001); Webber (2009); Plank (2011)

18 community-provided categories in UD

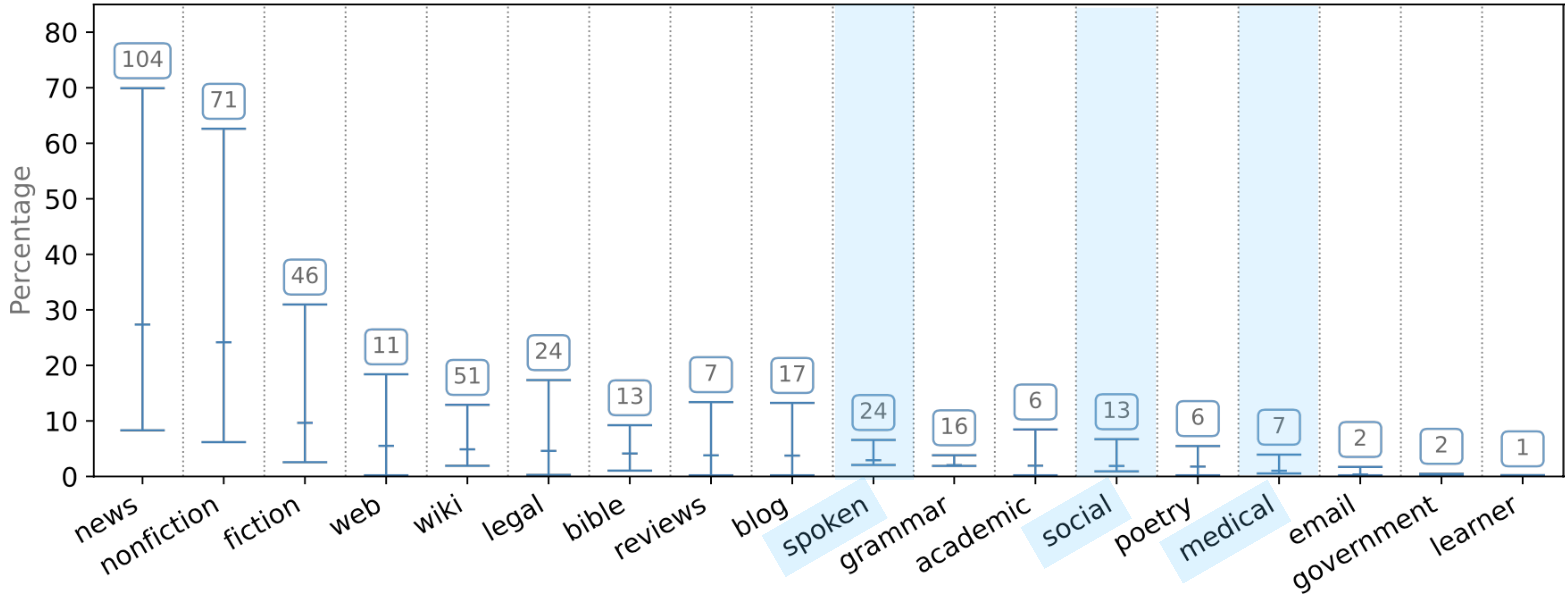
Genre Distribution in UD



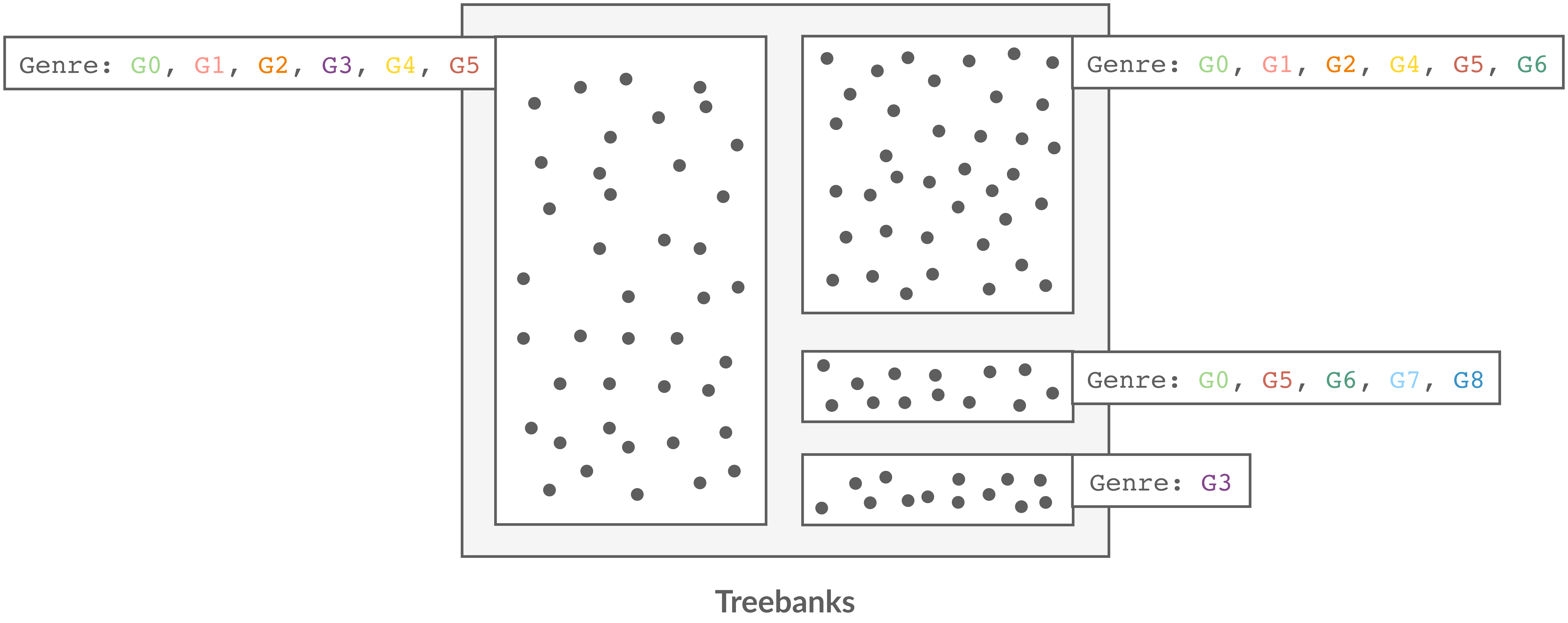
Genre Distribution in UD



Genre Distribution in UD

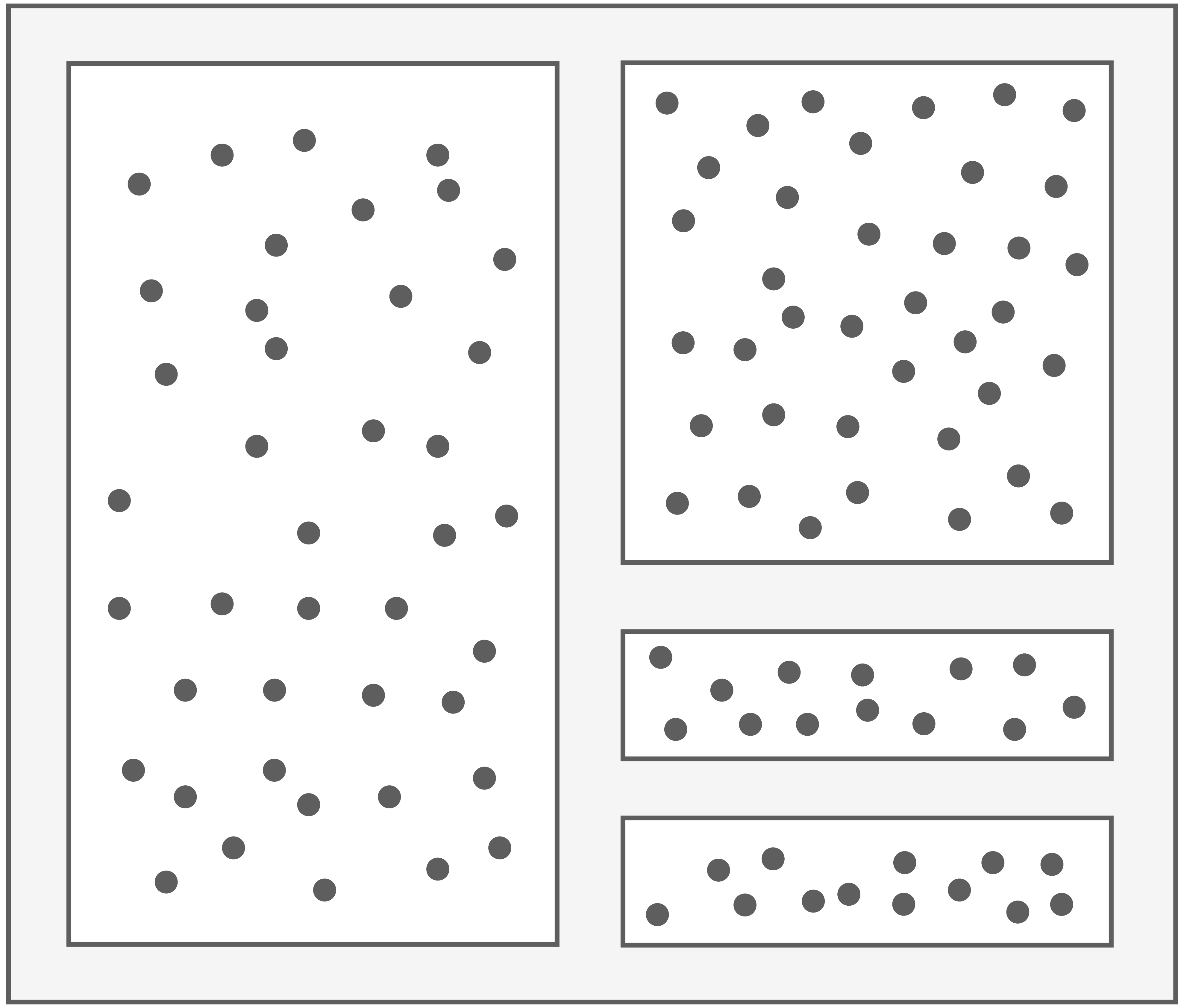


Targeted Data Selection

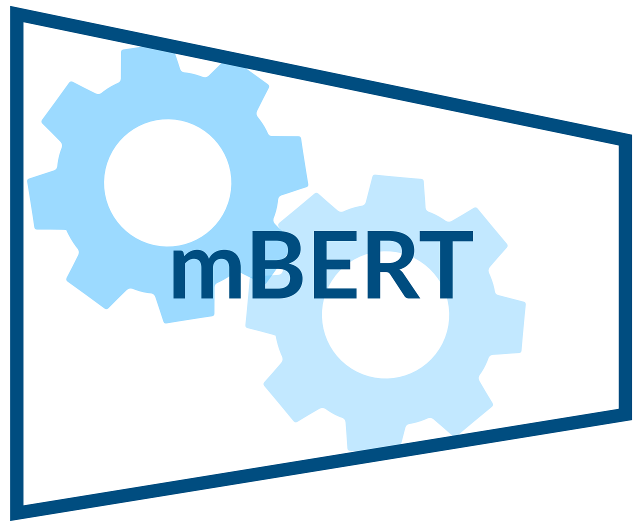


Genre: G0, G1, G2, G3, G4, G5
Genre: G0, G1, G2, G4, G5, G6
Genre: G0, G5, G6, G7, G8
Genre: G3
Treebanks

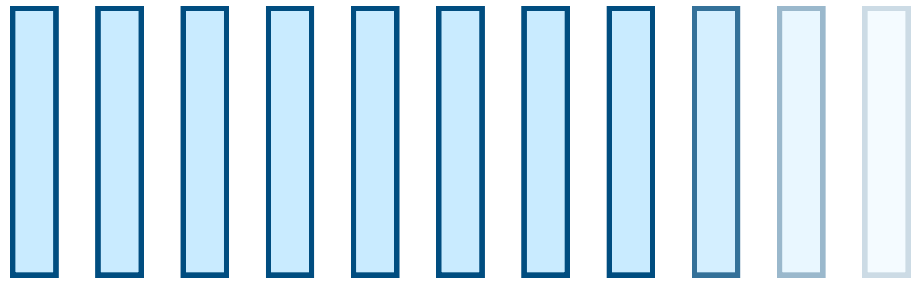
	MODEL	GENRES	LANGS
This Work	mBERT	18	104
Aharoni & Goldberg (2020)	BERT	5	1



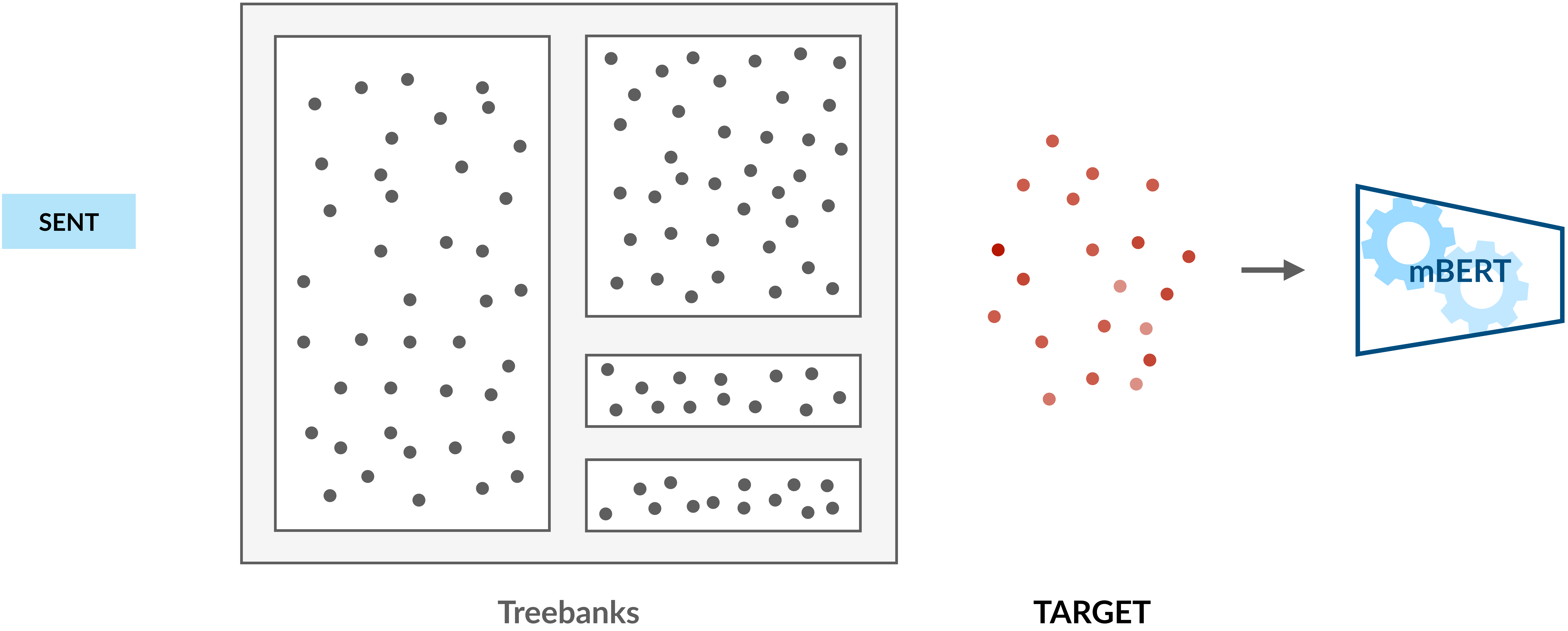
Treebanks



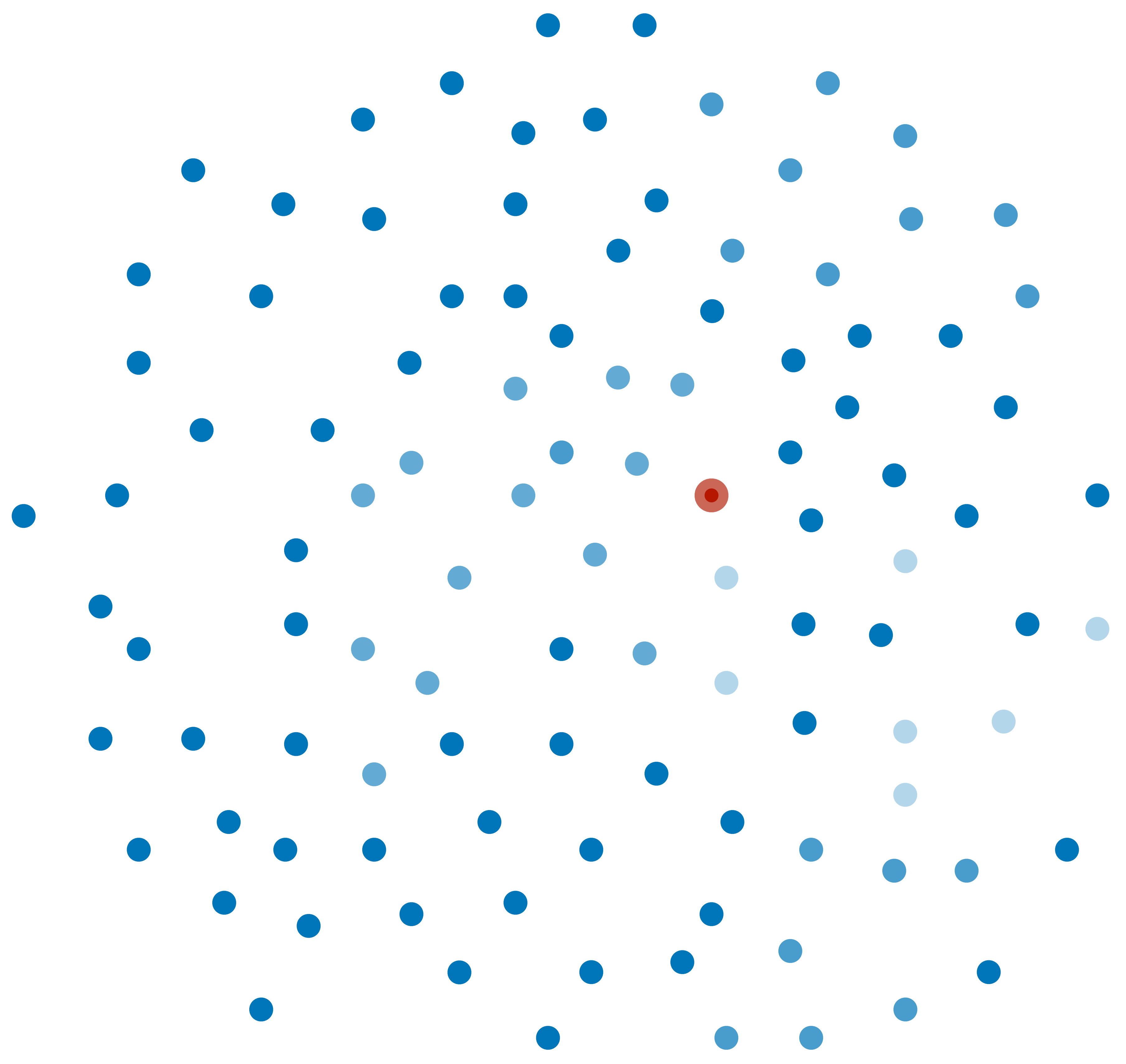
Devlin et al. (2019)



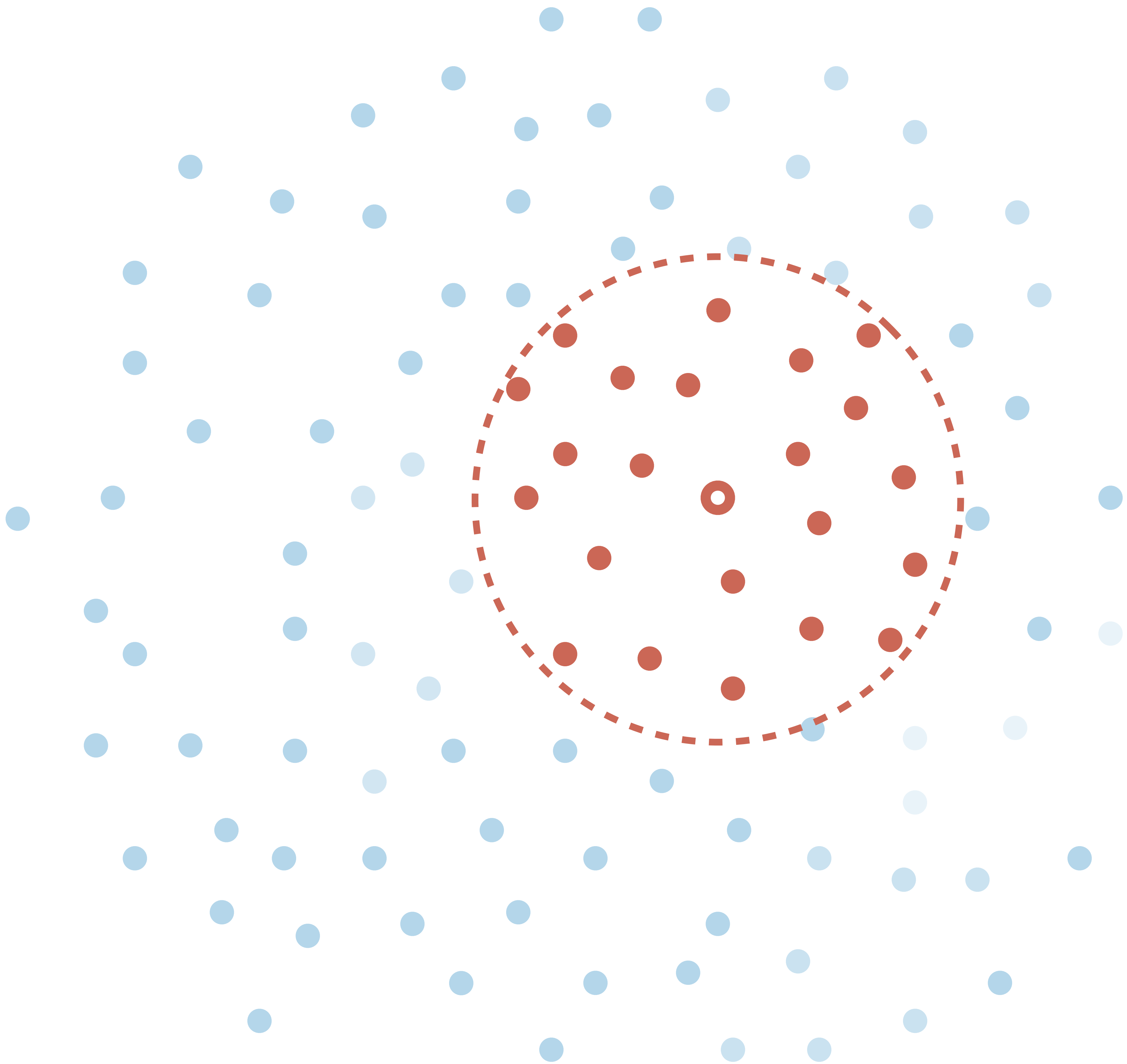
SENT


SENT
Treebanks
TARGET
mBERT

SENT



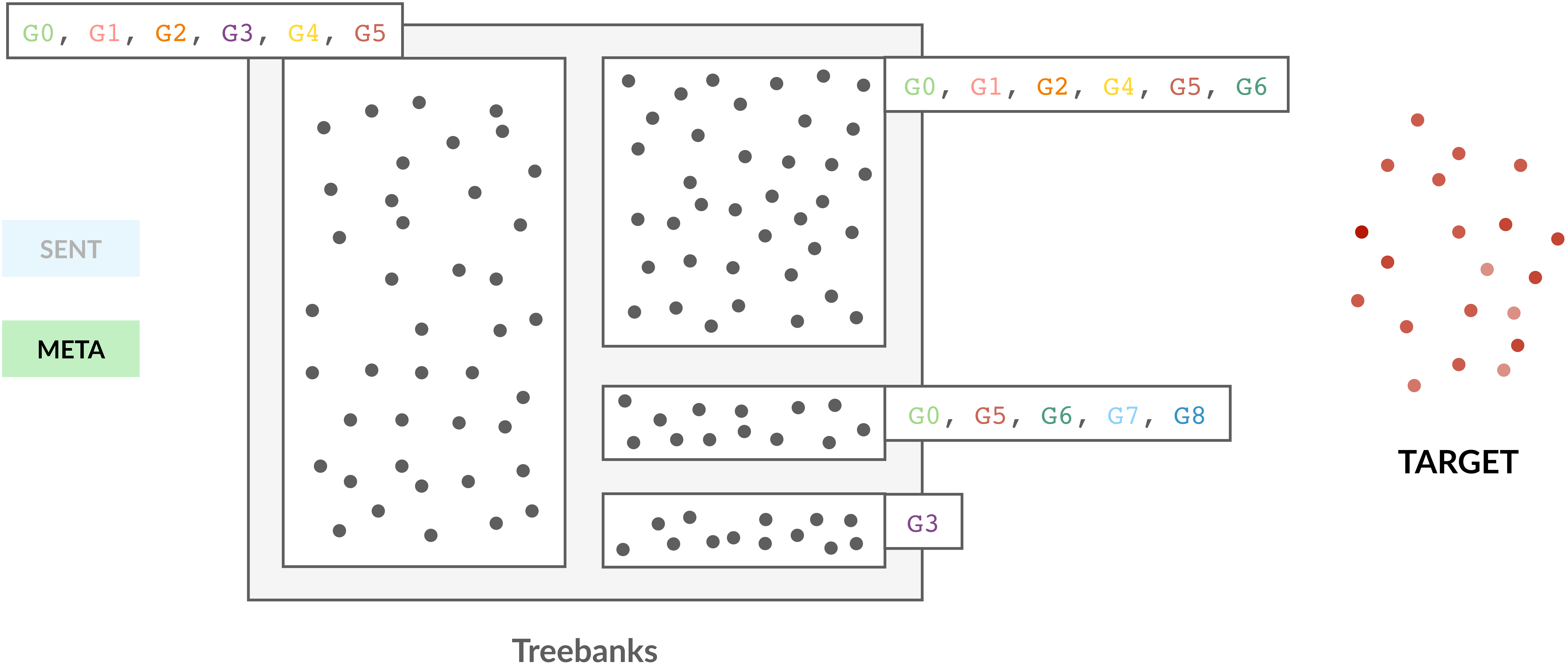
SENT



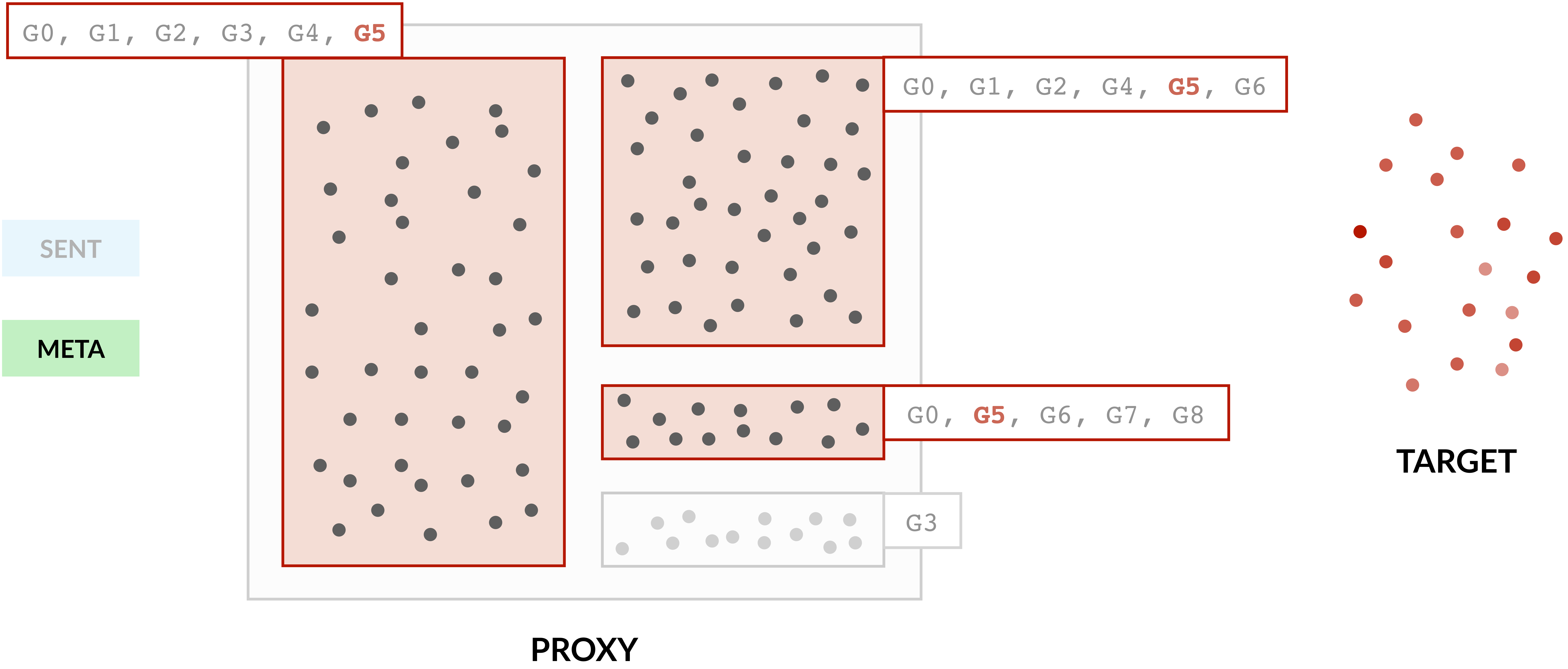
PROXY

SENT

META



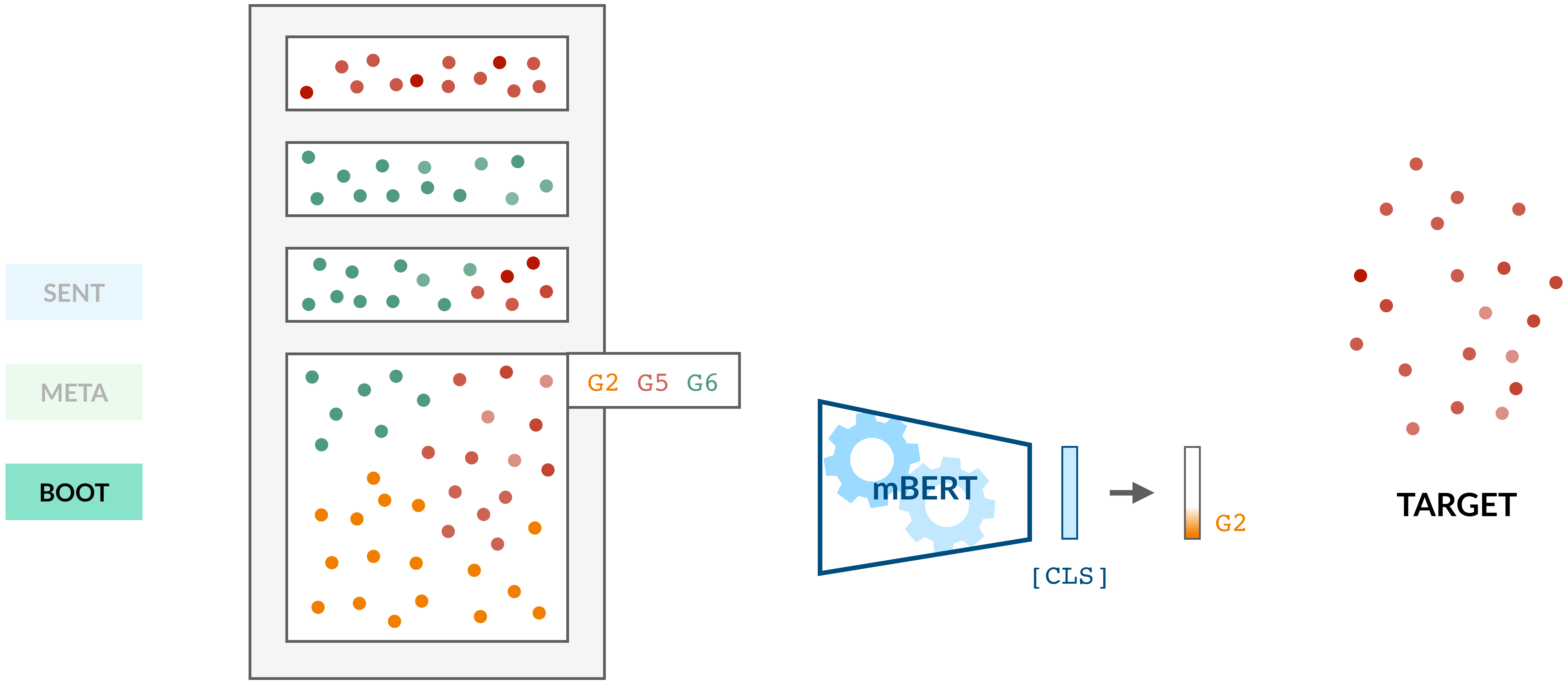
G0, G1, G2, G3, G4, G5
G0, G1, G2, G4, G5, G6
G0, G5, G6, G7, G8
G3
SENT
META
Treebanks
TARGET

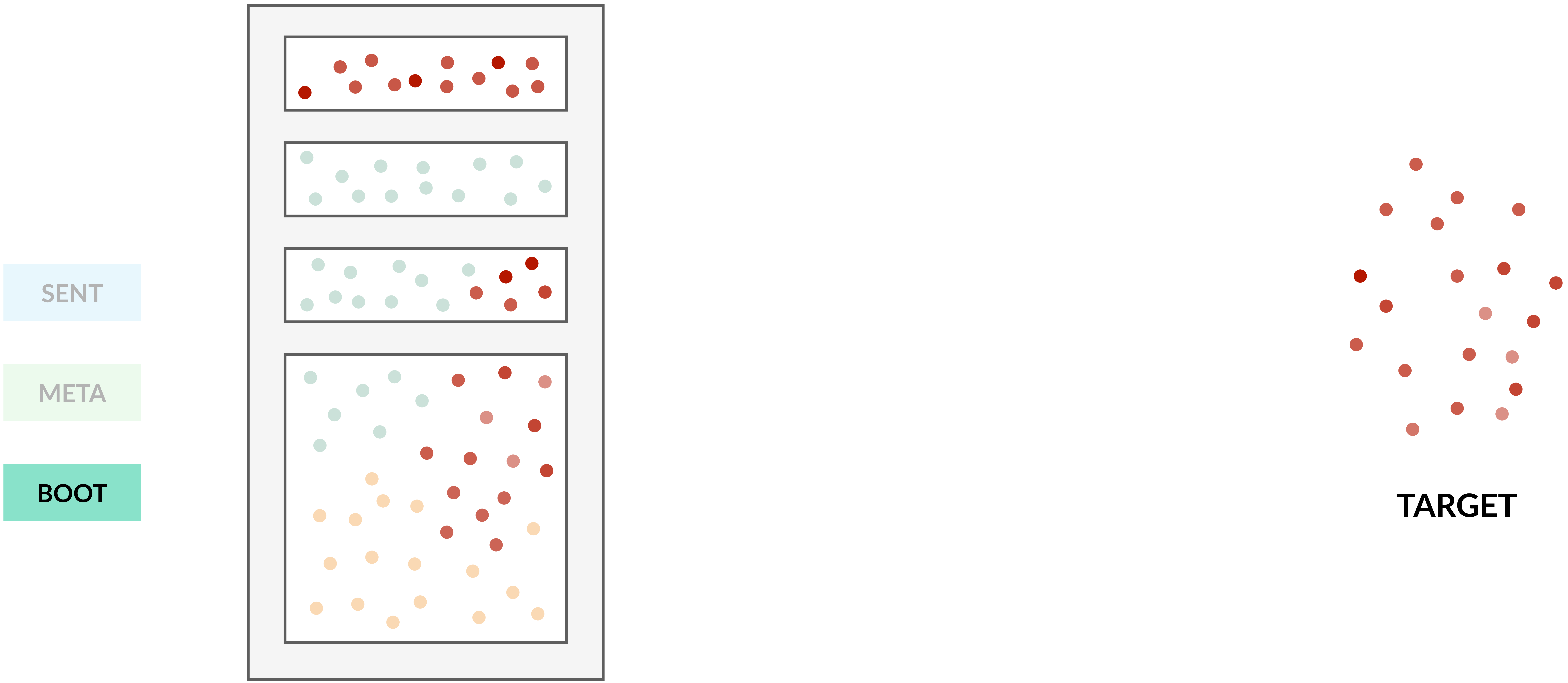

G0, G1, G2, G3, G4, G5
G0, G1, G2, G4, G5, G6
G0, G5, G6, G7, G8
G3
SENT
META
PROXY
TARGET

SENT

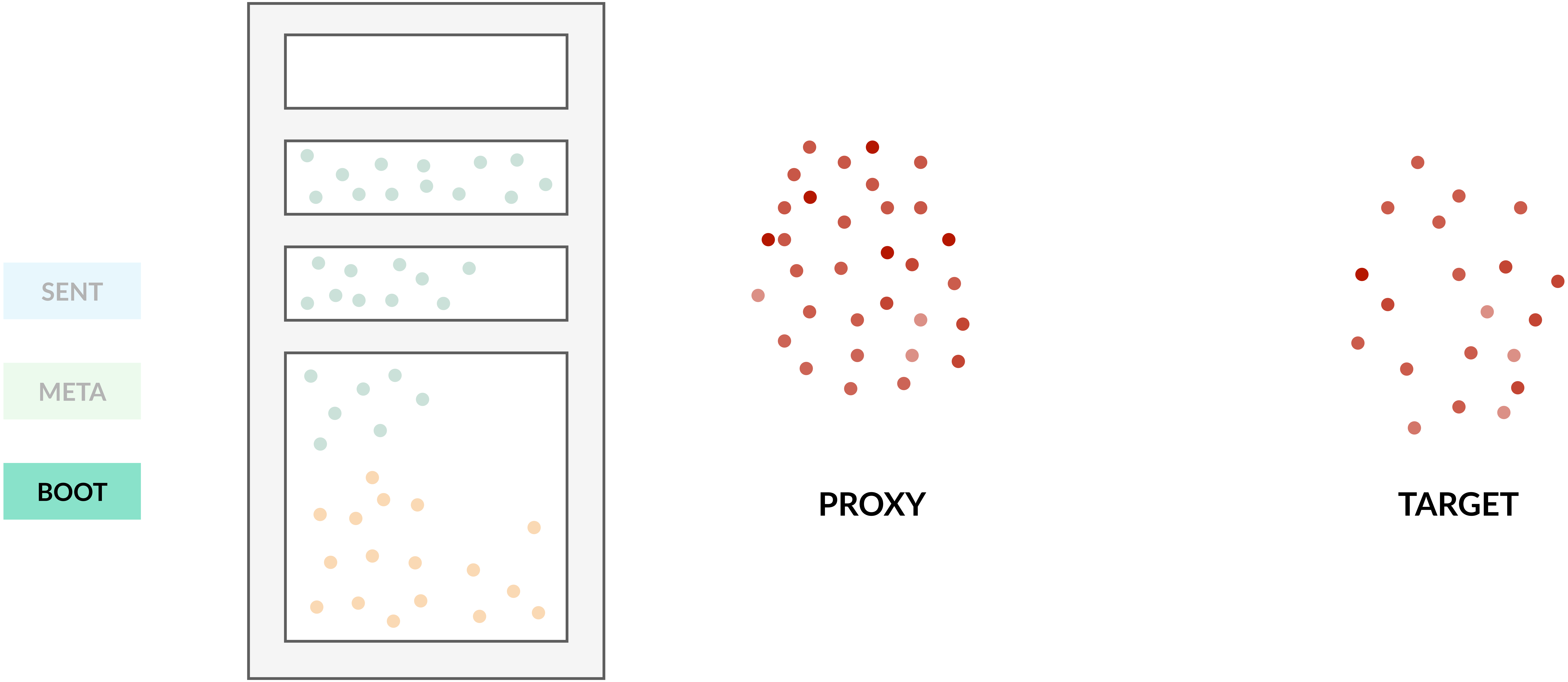
META

BOOT


SENT
META
BOOT
G2 G5 G6
mBERT
[CLS]
G2
TARGET



SENT
META
BOOT
TARGET


SENT
META
BOOT
PROXY
TARGET

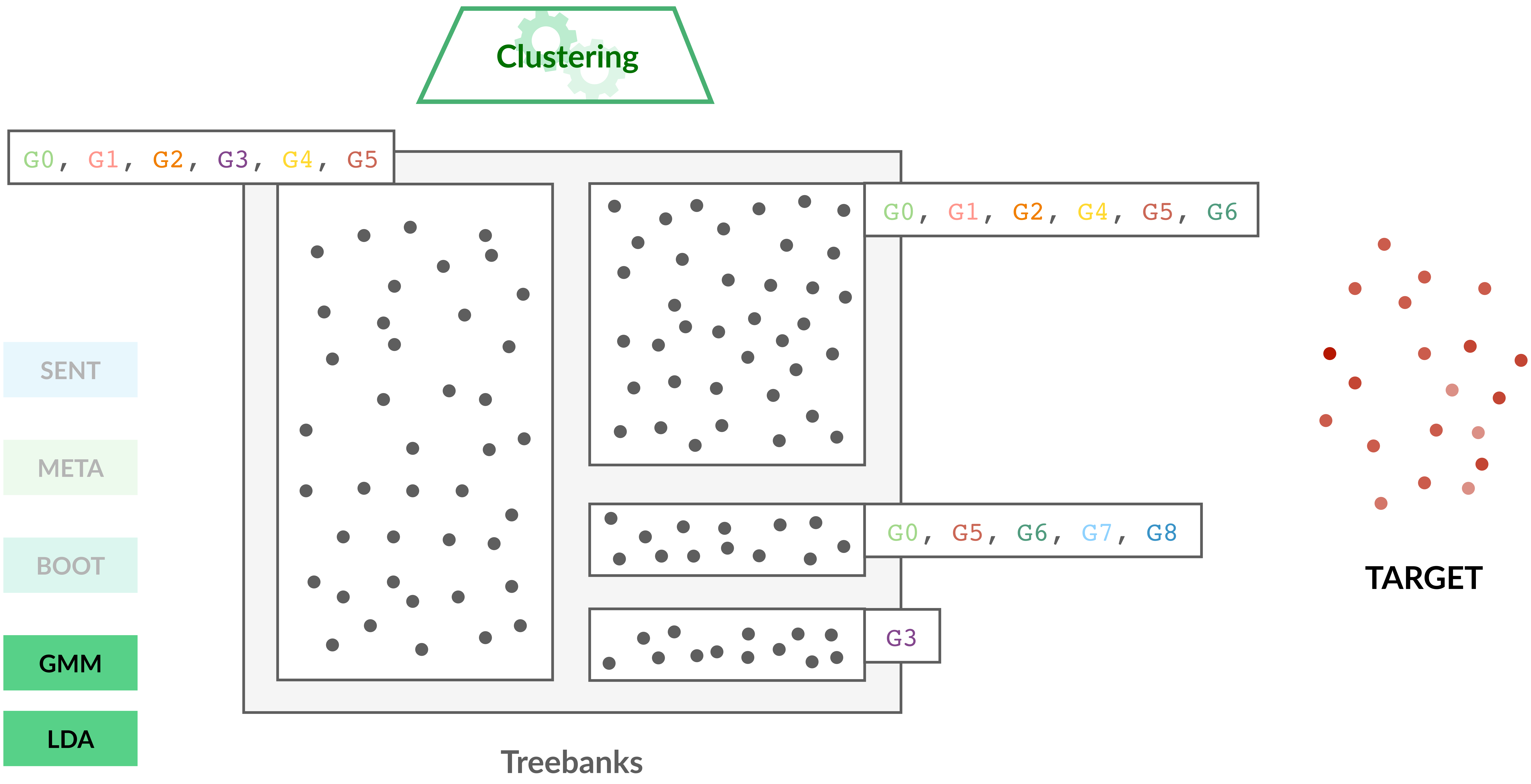
SENT

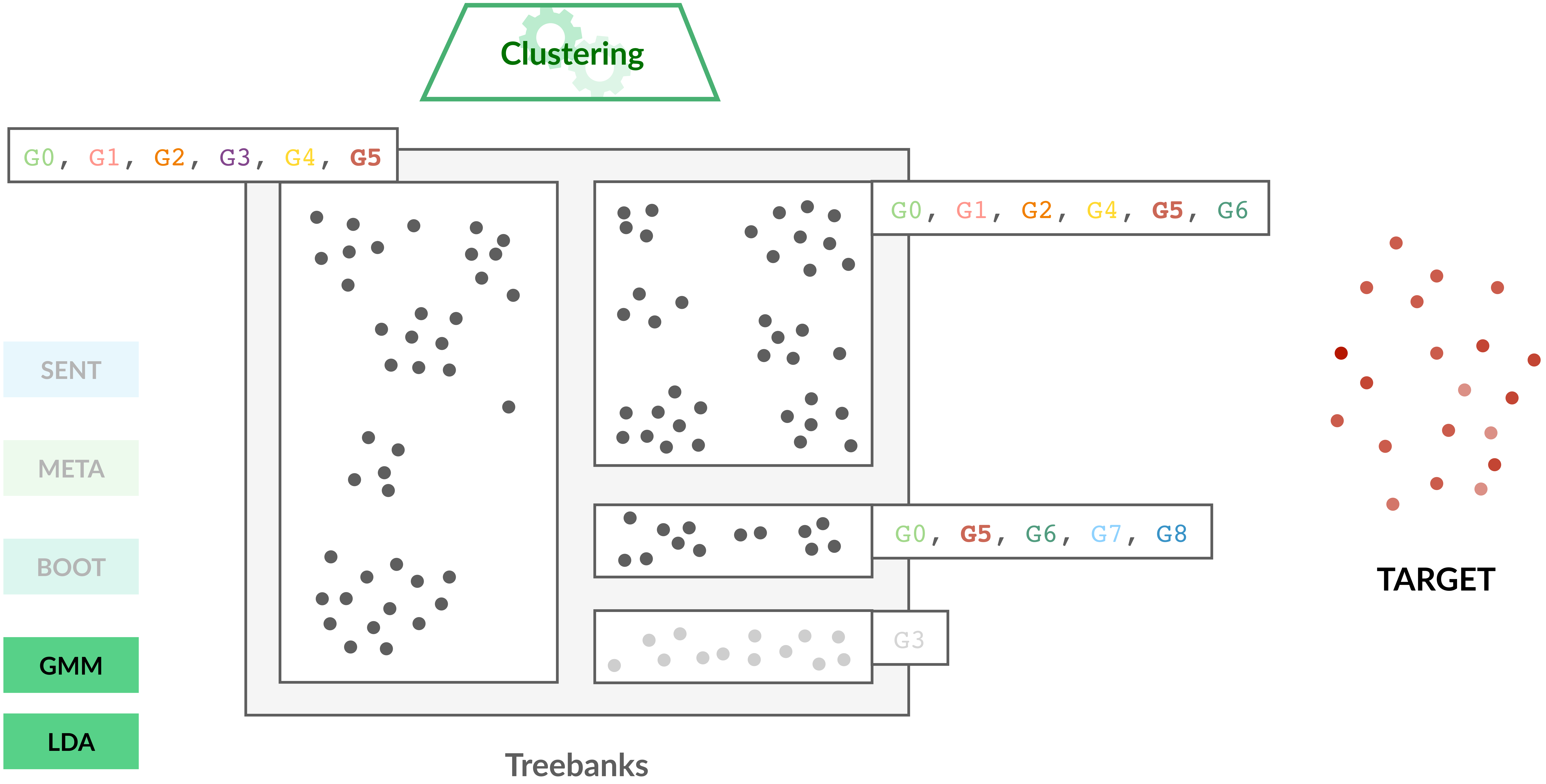
META

BOOT

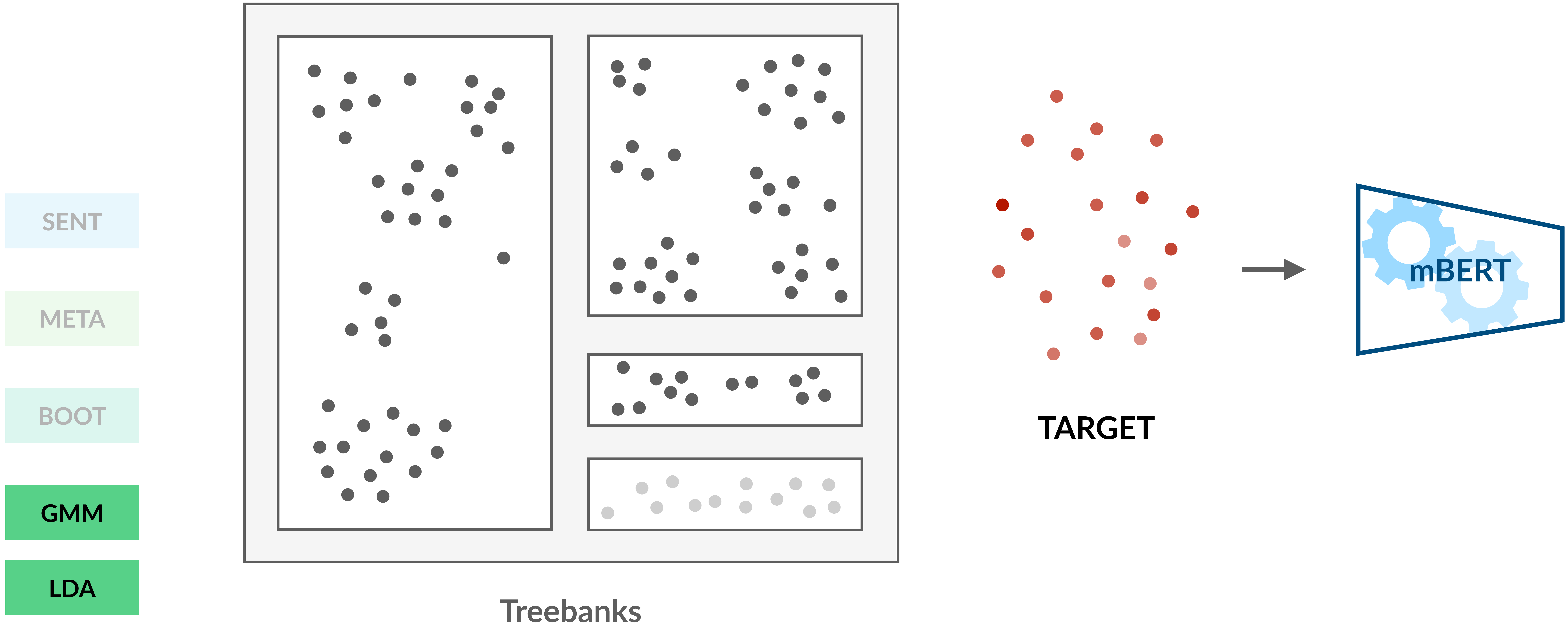
GMM

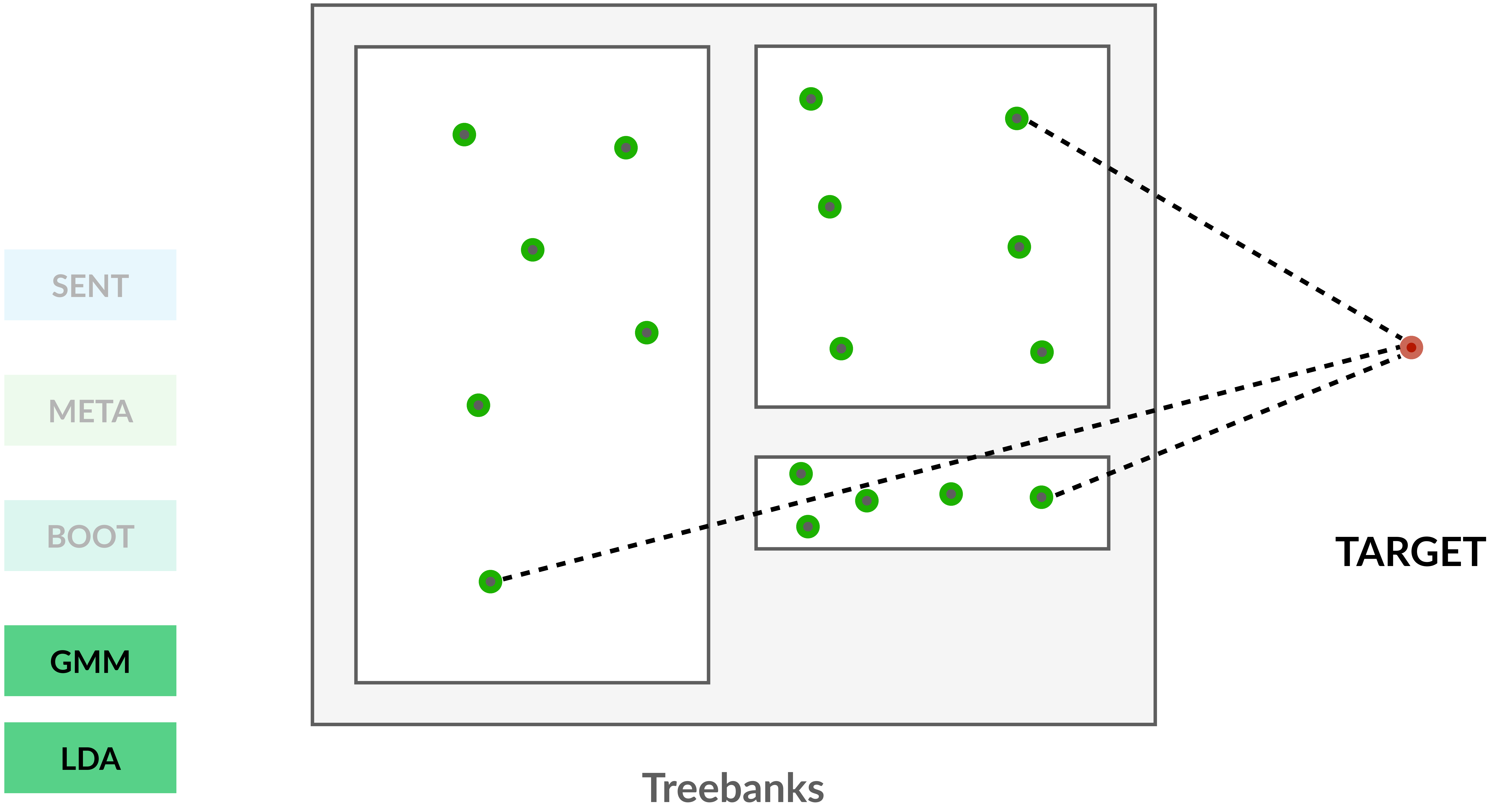
LDA


Clustering
G0, G1, G2, G3, G4, G5
G0, G1, G2, G4, G5, G6
G0, G5, G6, G7, G8
G3
SENT
META
BOOT
GMM
LDA
TARGET
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
TARGET
Treebanks


SENT
META
BOOT
GMM
LDA
Treebanks
TARGET
mBERT


SENT
META
BOOT
GMM
LDA
TARGET
Treebanks

SENT
META
BOOT
GMM
LDA
PROXY
TARGET

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	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 TE MYV QHE QTD

SENT

META

BOOT

GMM

LDA

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	TE	MYV	QHE	QTD
TARGET	✓	~	~	✓	✓	✗	✗	~	✓	✗	✓	✓
SENT												
META												
BOOT												
GMM												
LDA												

SWL SA KPV TA GL YUE CKT FO TE MYV QHE QTD

TARGET

RAND

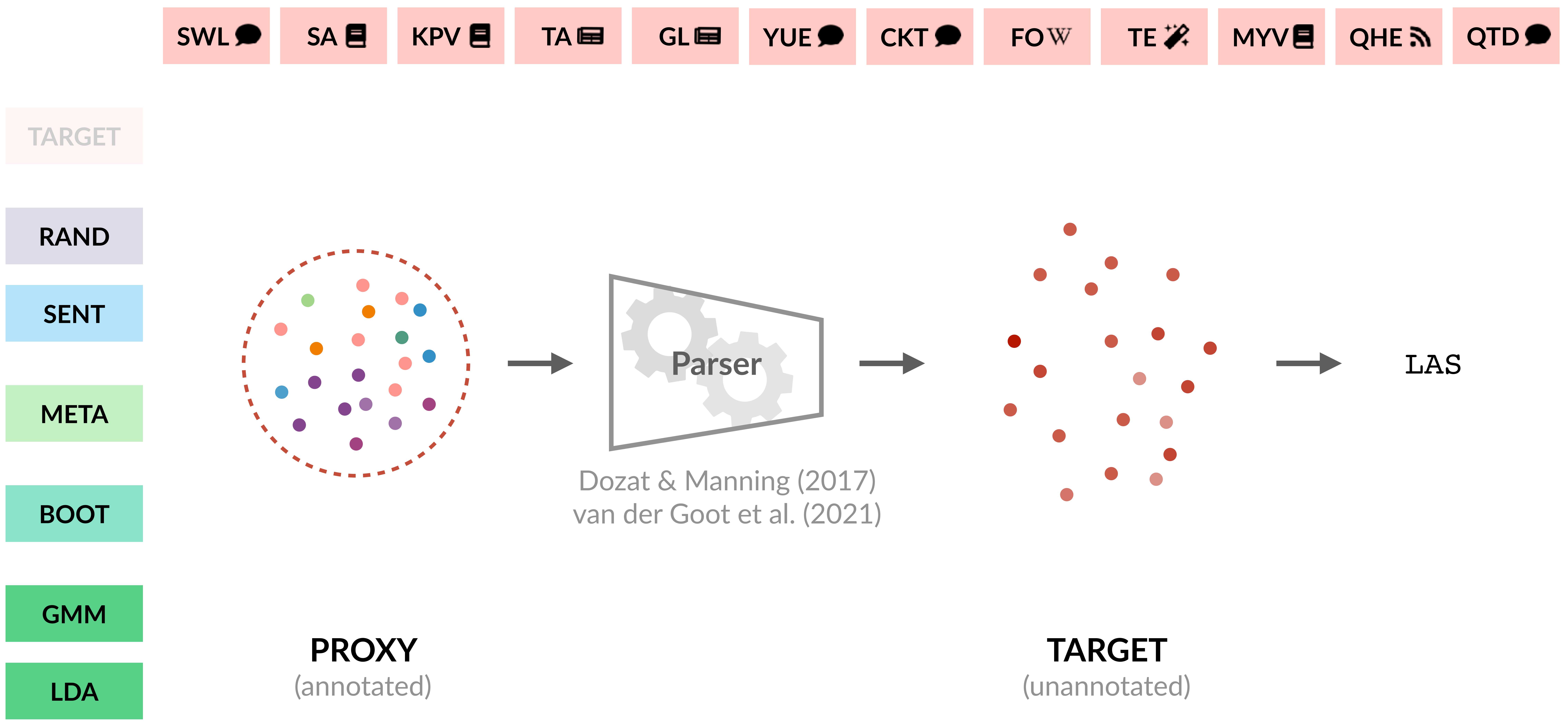
SENT

META

BOOT

GMM

LDA


SWL
SA
KPV
TA
GL
YUE
CKT
FO
TE
MYV
QHE
QTD
TARGET
RAND
SENT
META
BOOT
GMM
LDA
Parser
Dozat & Manning (2017)
van der Goot et al. (2021)
LAS
PROXY
(annotated)
TARGET
(unannotated)

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	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													
SENT													
META													
BOOT													
GMM													
LDA													

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	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													
SENT													
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													
GMM													
LDA													

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	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													
GMM													
LDA													

	SWL	SA	KPV	TA	GL	YUE	CKT	FO	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	TE	MYV	QHE	QTD	∅

TARGET

RAND

SENT

META

BOOT

GMM

LDA

mBERT
(untuned)

BOOT
(genre-tuned)

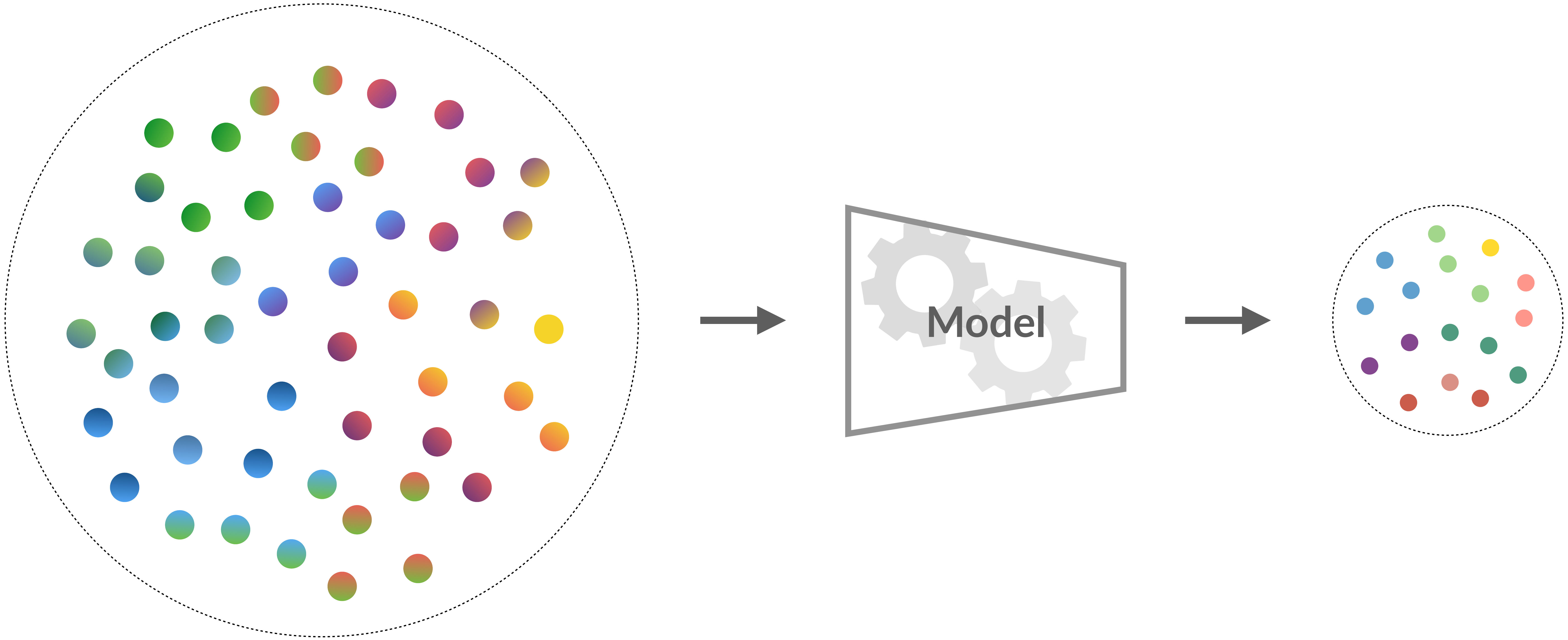
Conclusion

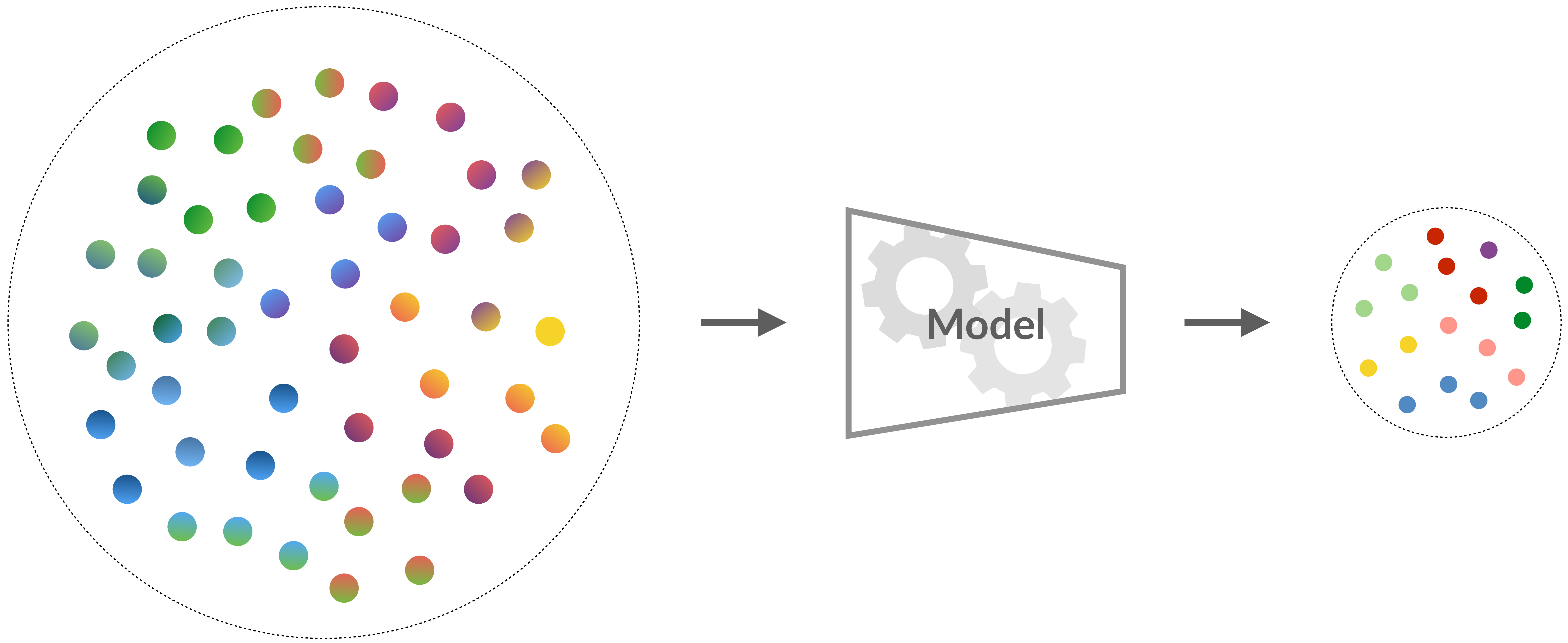
BOOT GMM LDA

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

How can we create more human-centred NLP?

- Learn from human disagreement
- Learn with humans in the loop

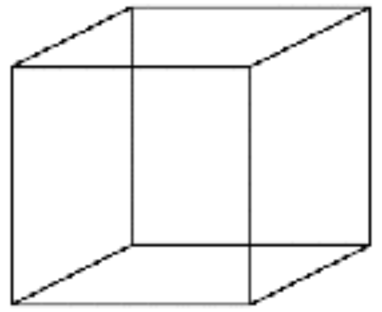

Model


Model

Disagreement in human annotation is ubiquitous

Side benefit of annotation - fortuitous data:

Disagreement as a source of information?



there are linguistically hard cases, even for POS tagging

e.g. Manning (2011). *Part-of-Speech tagging from 97% to 100%. Is It Time for Some Linguistics?*

Part-of-Speech (POS)

VERB	**NOUN**	ADP	NOUN	SYM
VERB	**PRON**	ADP	NOUN	SYM
VERB	**ADV**	ADP	NOUN	SYM
Say	Anything	with	boyfriend	:)

Example from Twitter crowdsourced dataset (Gimpel et al., 2012; Hovy et al., 2015)

Understanding Indirect Answers

Q: Hey. Everything ok?
A: I'm just mad at my agent

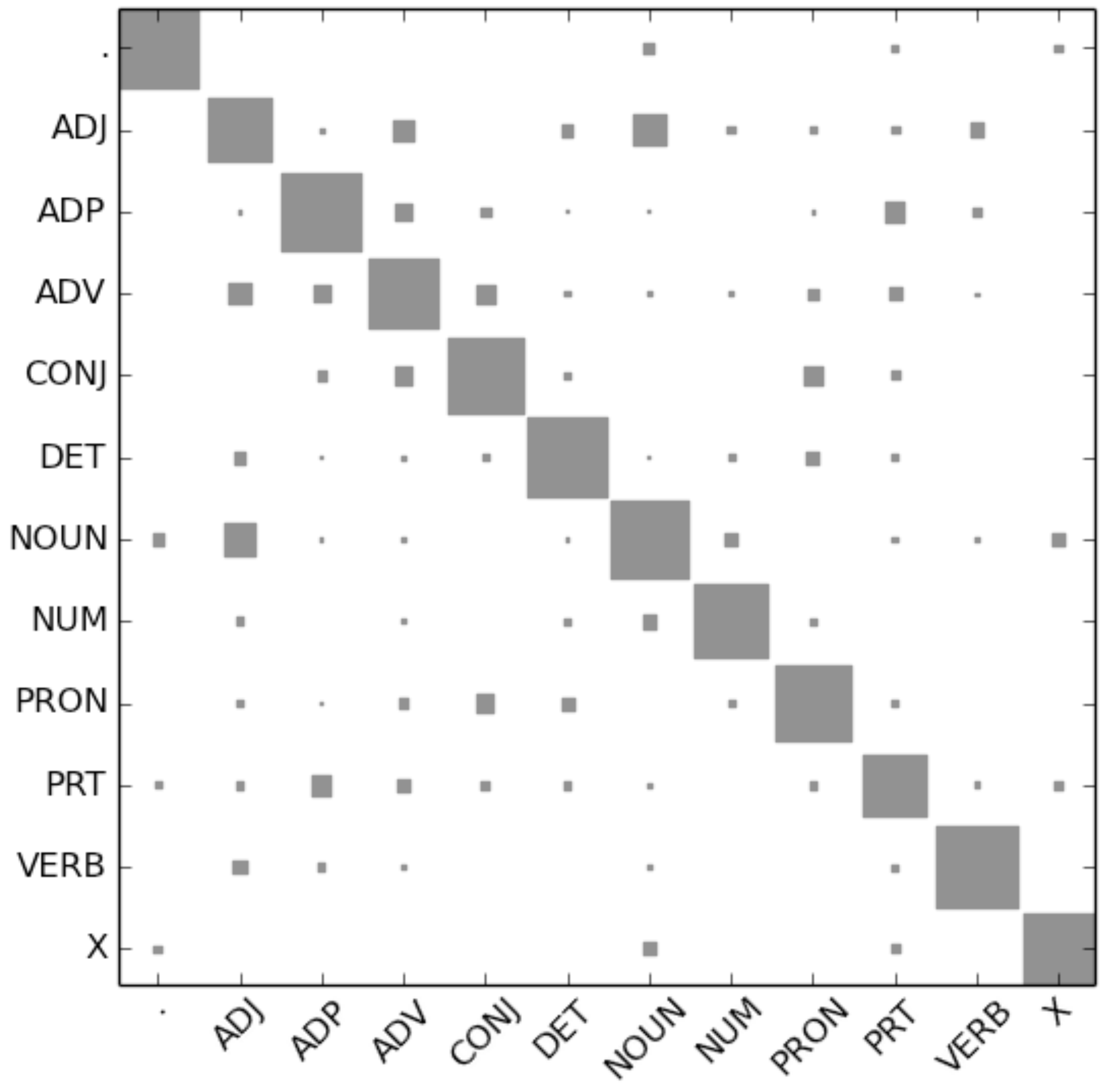
Yes
No
Yes, subject to some condition
Neither Yes nor no
Other
N/A

All agree	75.02%
Two agree	23.39%
All disagree	1.59%

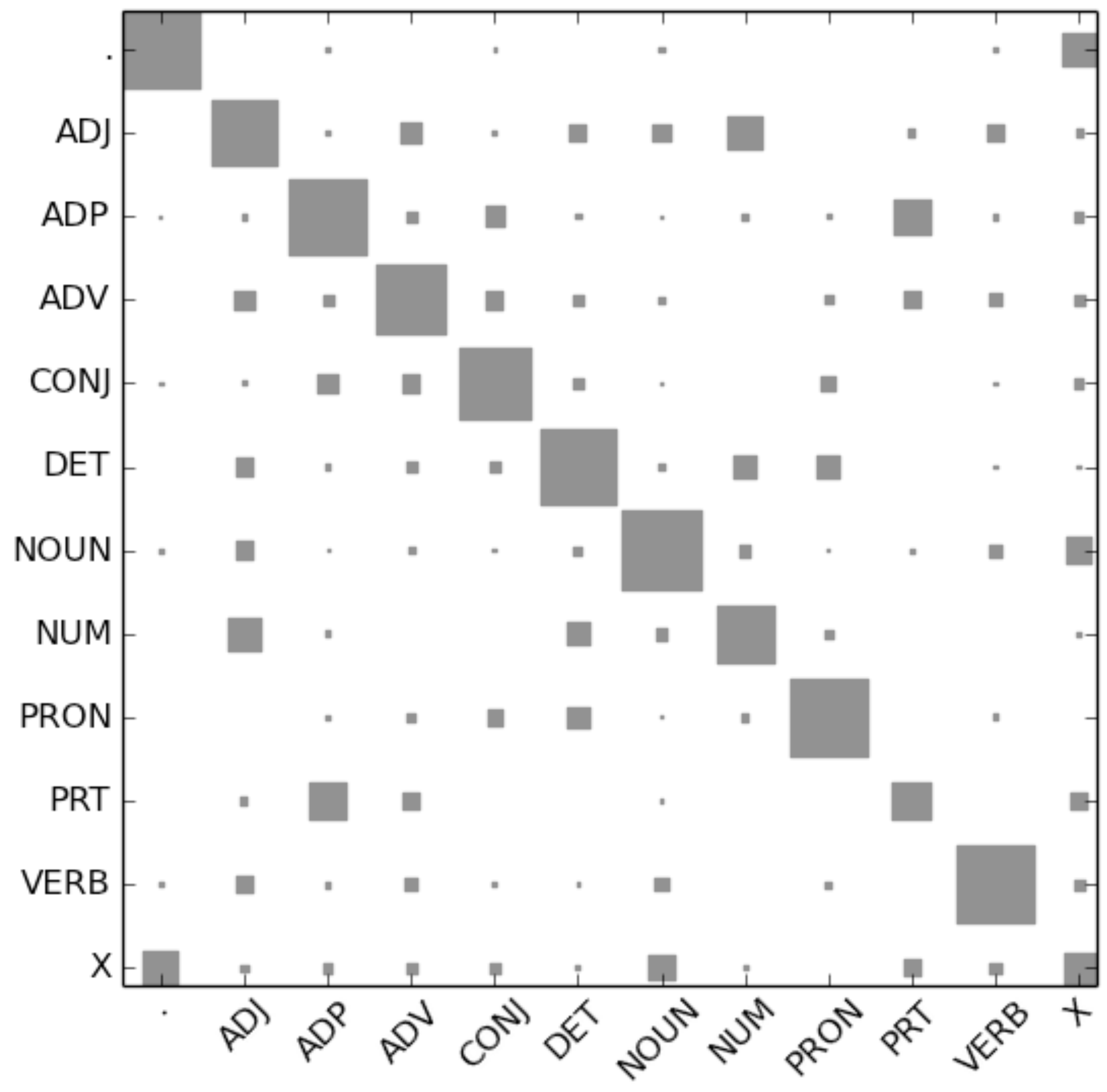
Damgaard, Toborek, Eriksen & Plank (2021)

Are disagreements randomly distributed?

... and can we estimate disagreements from small samples?

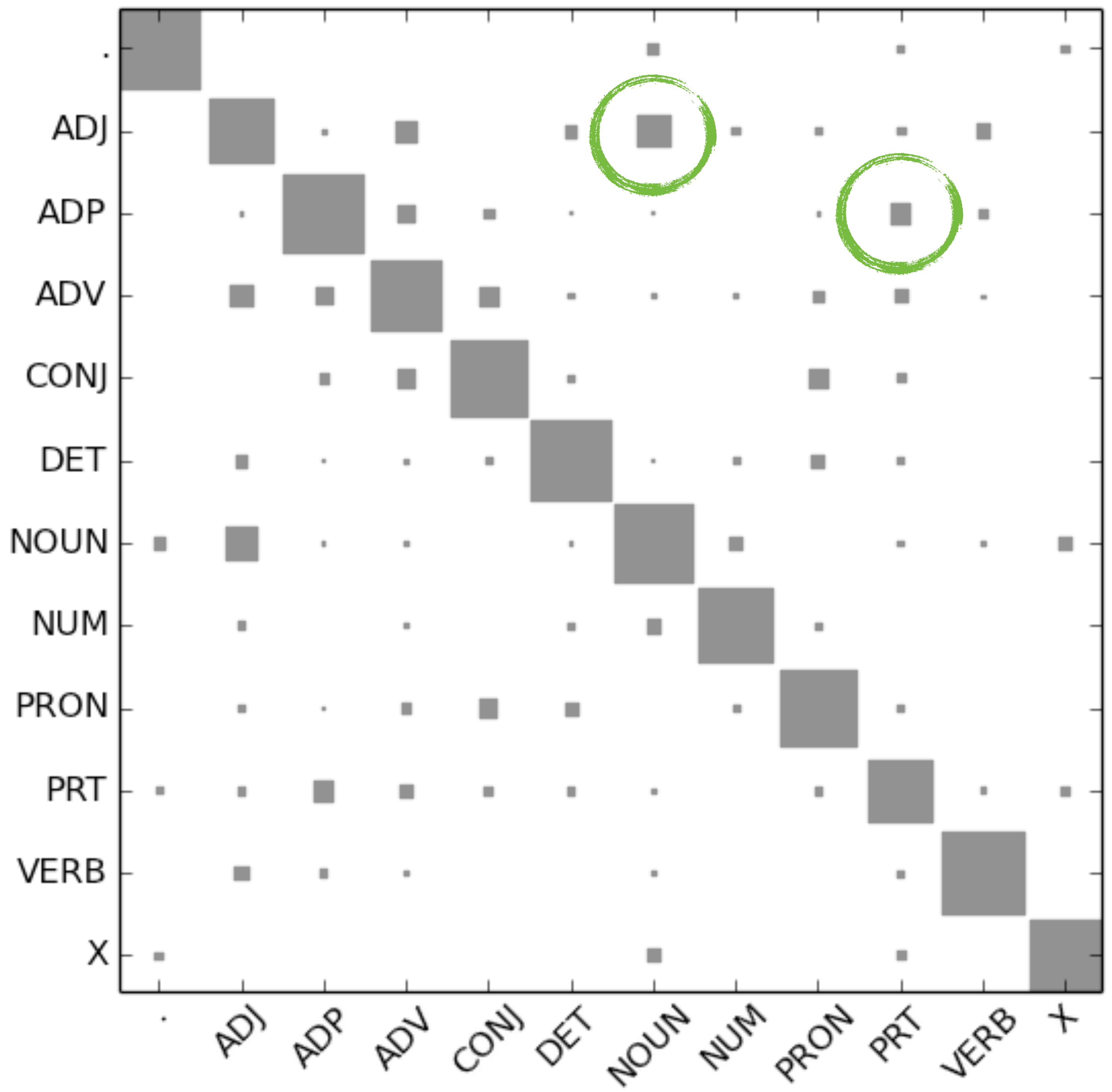


Wall Street Journal PTB-00

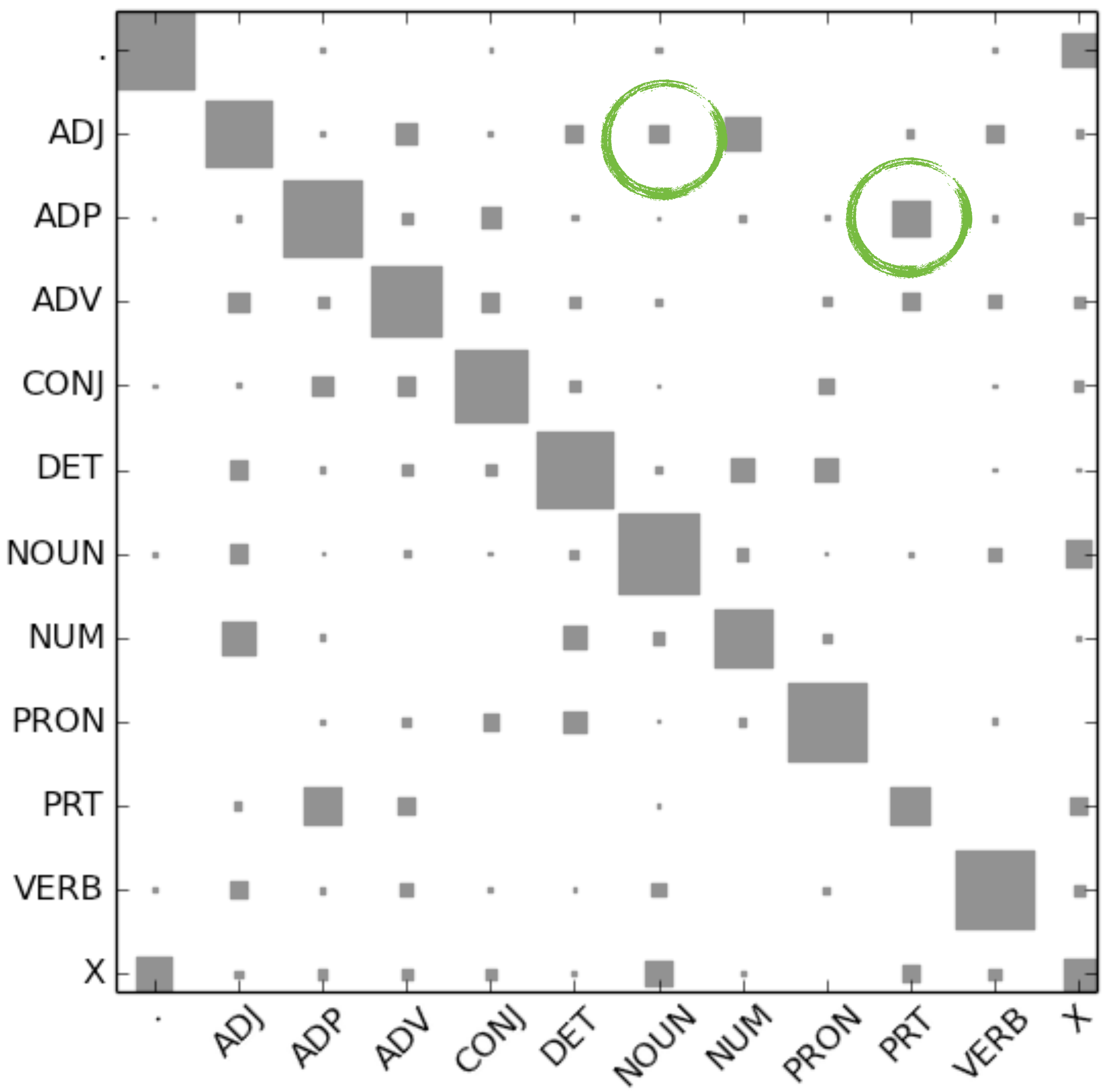


Twitter

(Plank et al., 2014)

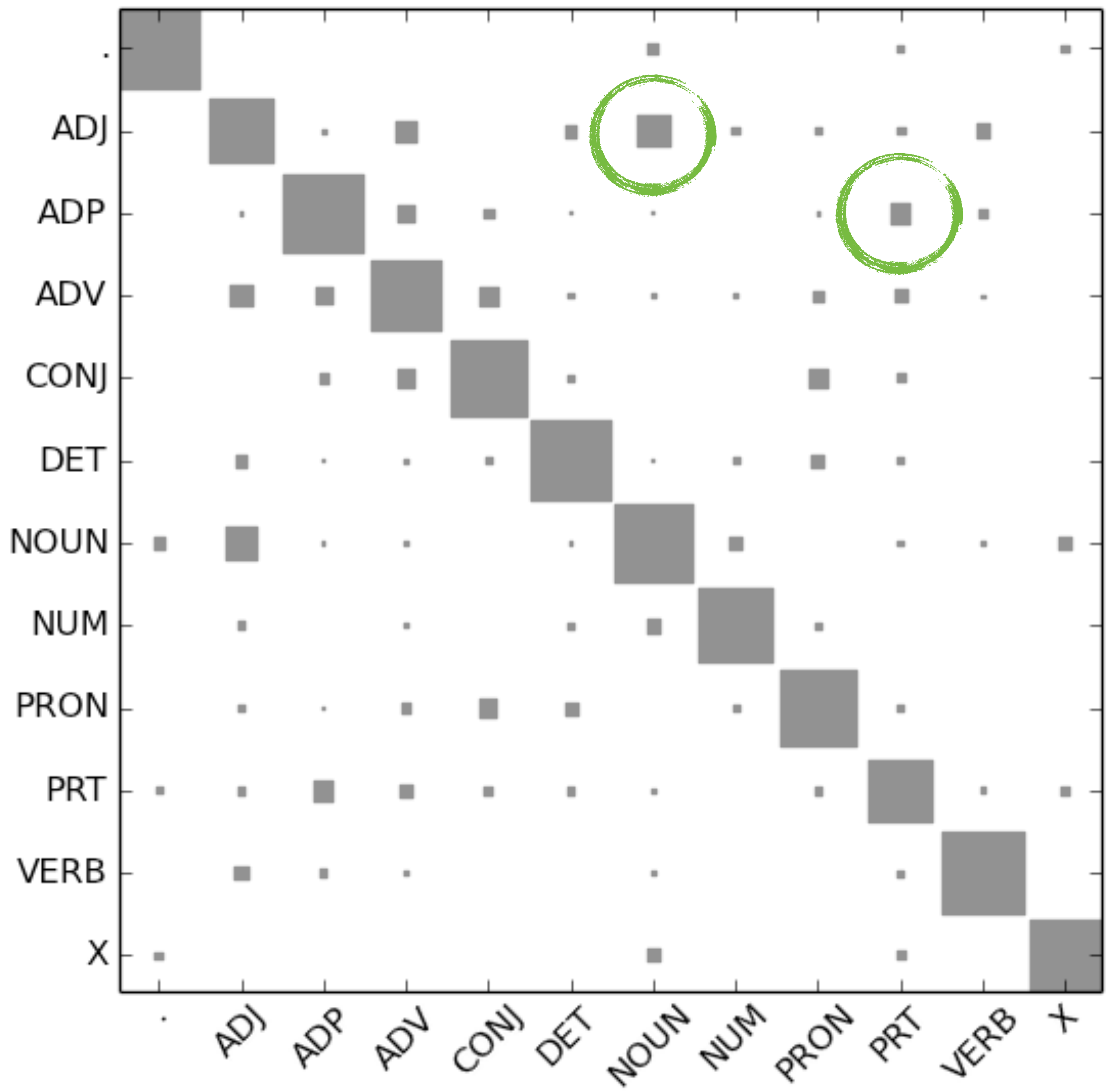


Wall Street Journal PTB-00

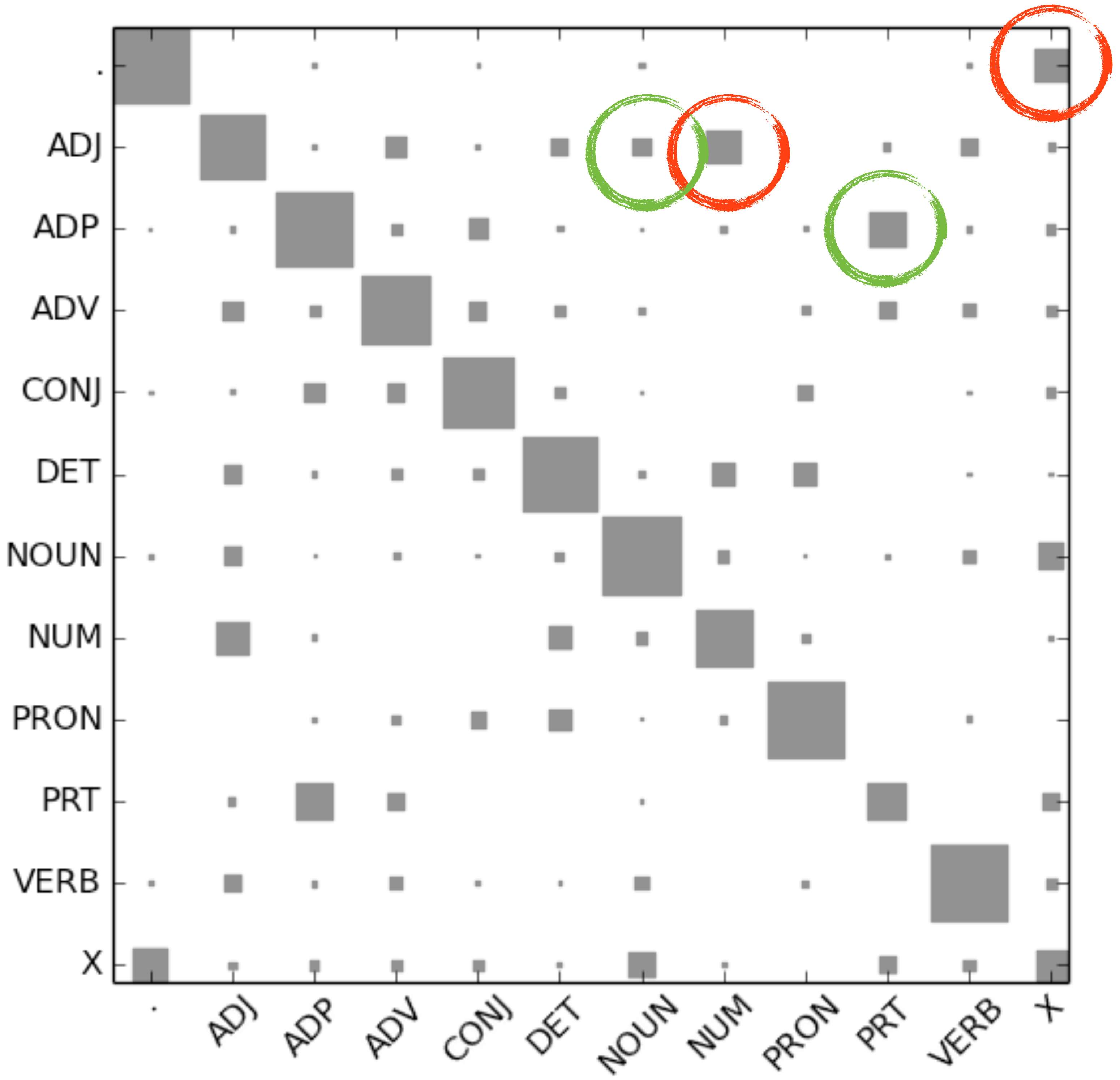


Twitter

(Plank et al., 2014)



Wall Street Journal PTB-00



Twitter

(Plank et al., 2014)

Are disagreements randomly distributed? No.

... and can we estimate disagreements from small samples? Yes!

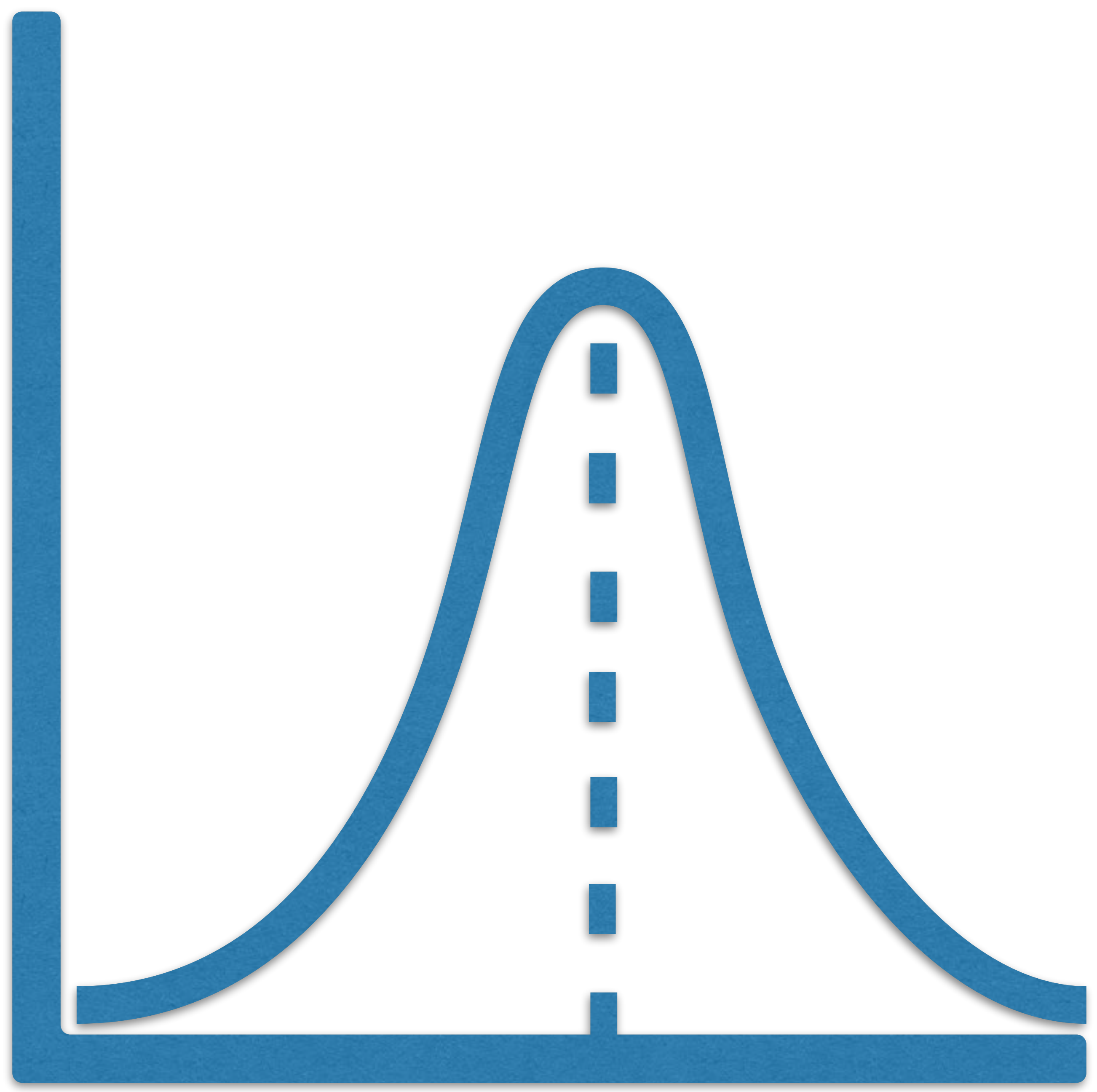
(Plank et al., 2014)

Are disagreement distributions unimodal?

… do they contain inherent disagreement signal?

(Pavlick & Kwiatkovski, 2019)

Is Unimodal (= Single Truth) Enough?

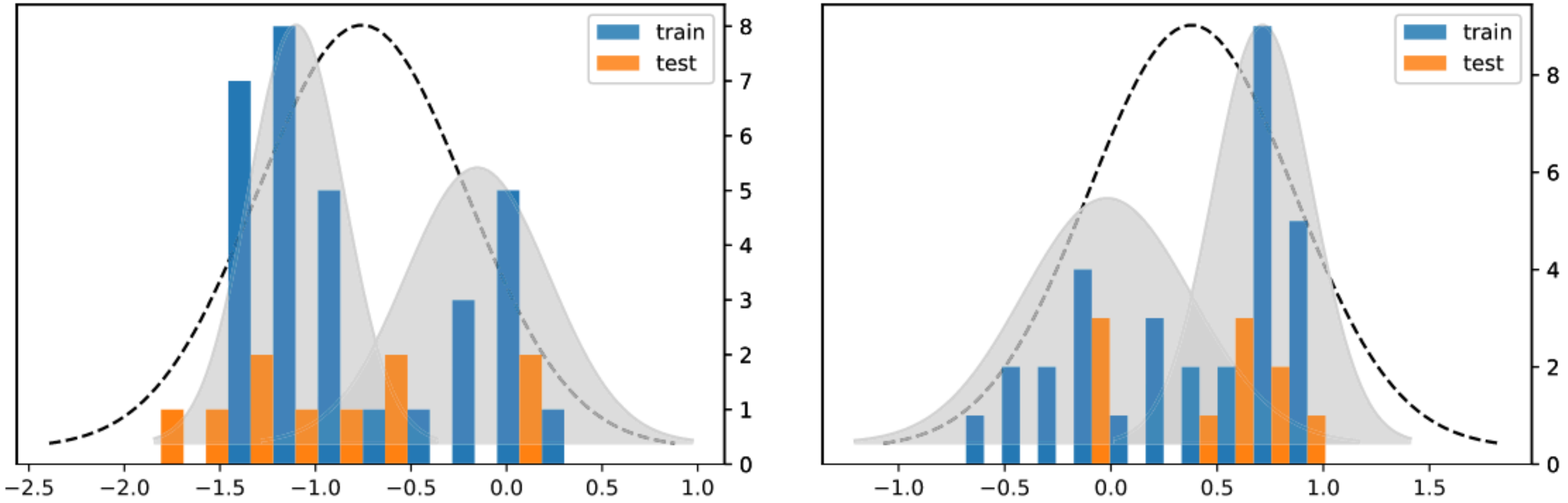


(Pavlick & Kwiatkovski, 2019)

Examples with bi-modal human judgements

p: A homeless man being observed by a man in business attire.
h: Two men are sleeping in a hotel.

p: Paula swatted the fly.
h: The swatting happened in a forceful manner.



Does p->h?

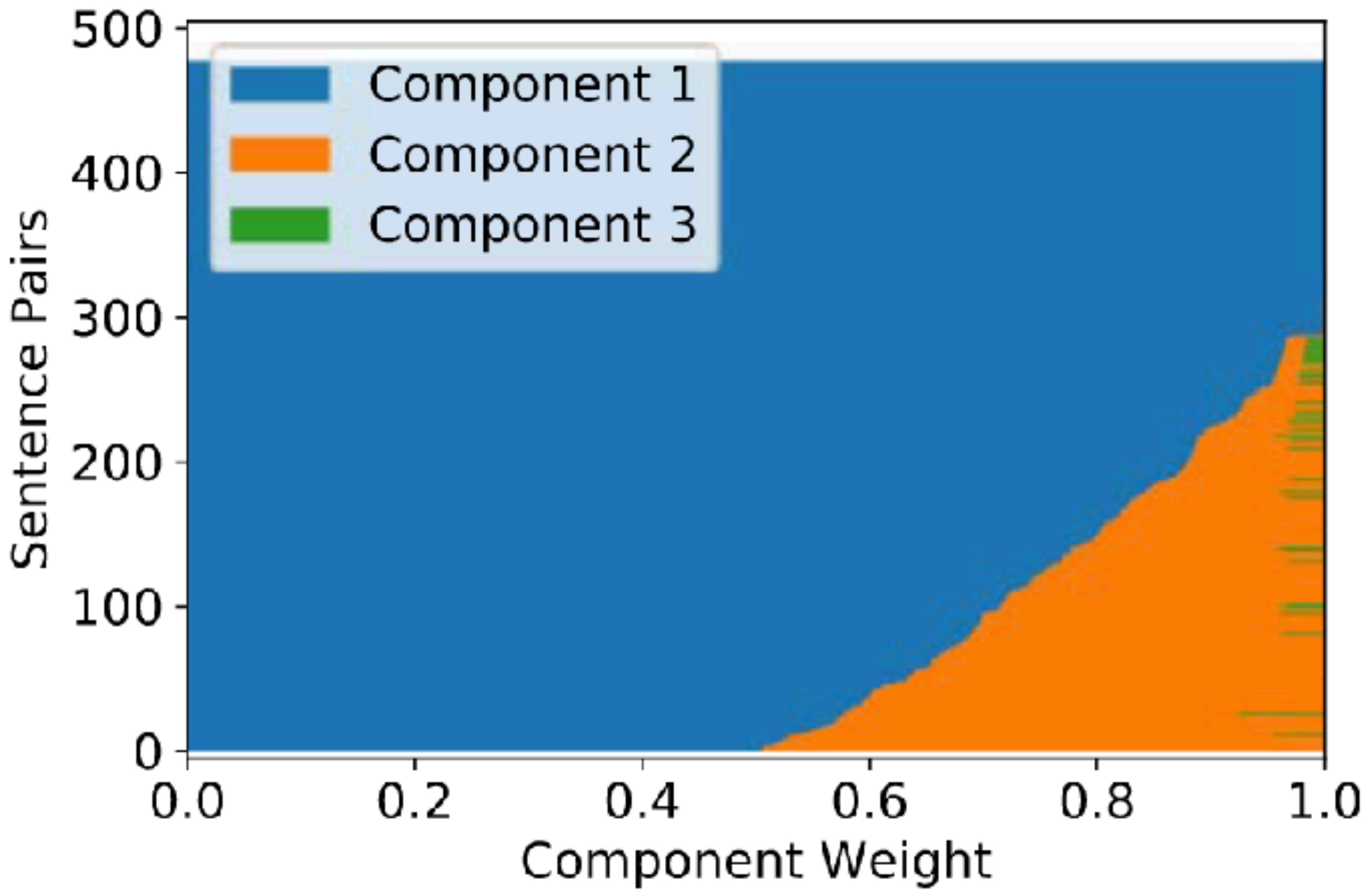
GMM with 1 *component* vs *k components*

(Pavlick & Kwiatkovski, 2019)

contradiction ~neutral entailment

Recognising Textual Entailment (RTE)
Analysis of re-crowdsourced data

"For 20% of the sentence pairs, there is a non-trivial second component"



(Pavlick & Kwiatkovski, 2019)

Are disagreement distributions unimodal? No.

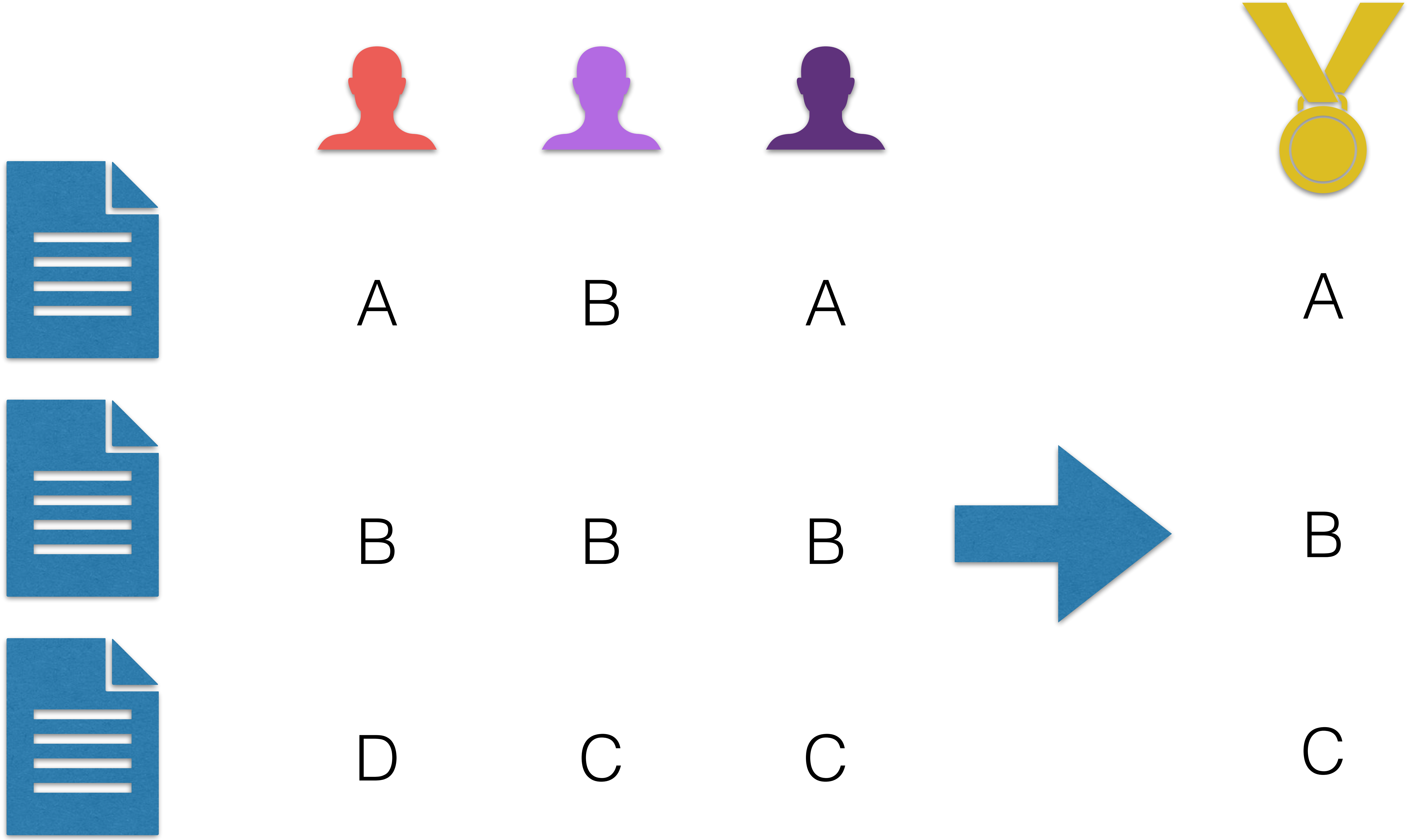
… do they contain inherent disagreement signal? Yes!

(Pavlick & Kwiatkovski, 2019)

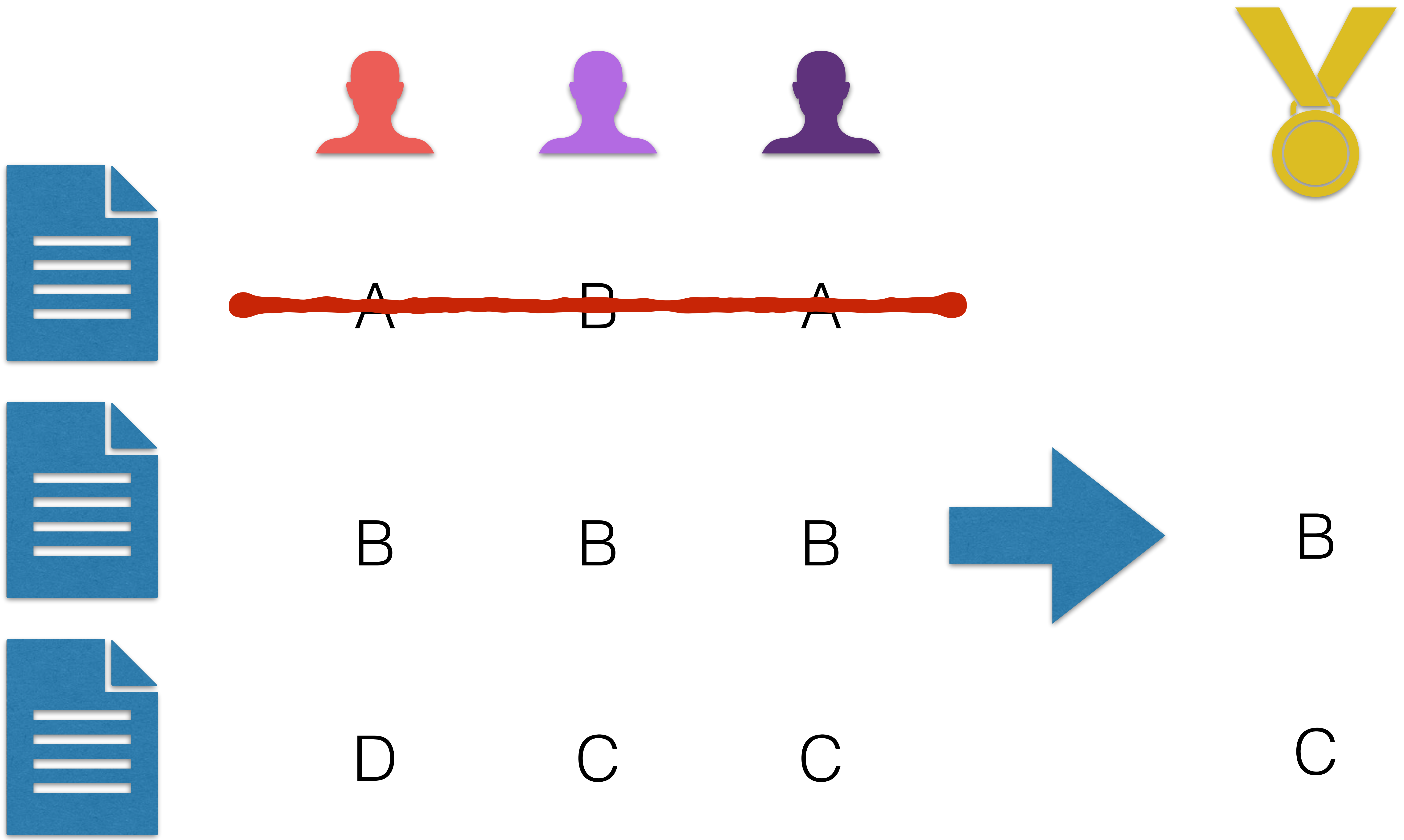
Roadmap

1. Data: Is disagreement random noise?
2. Modeling: How can we leverage disagreement?

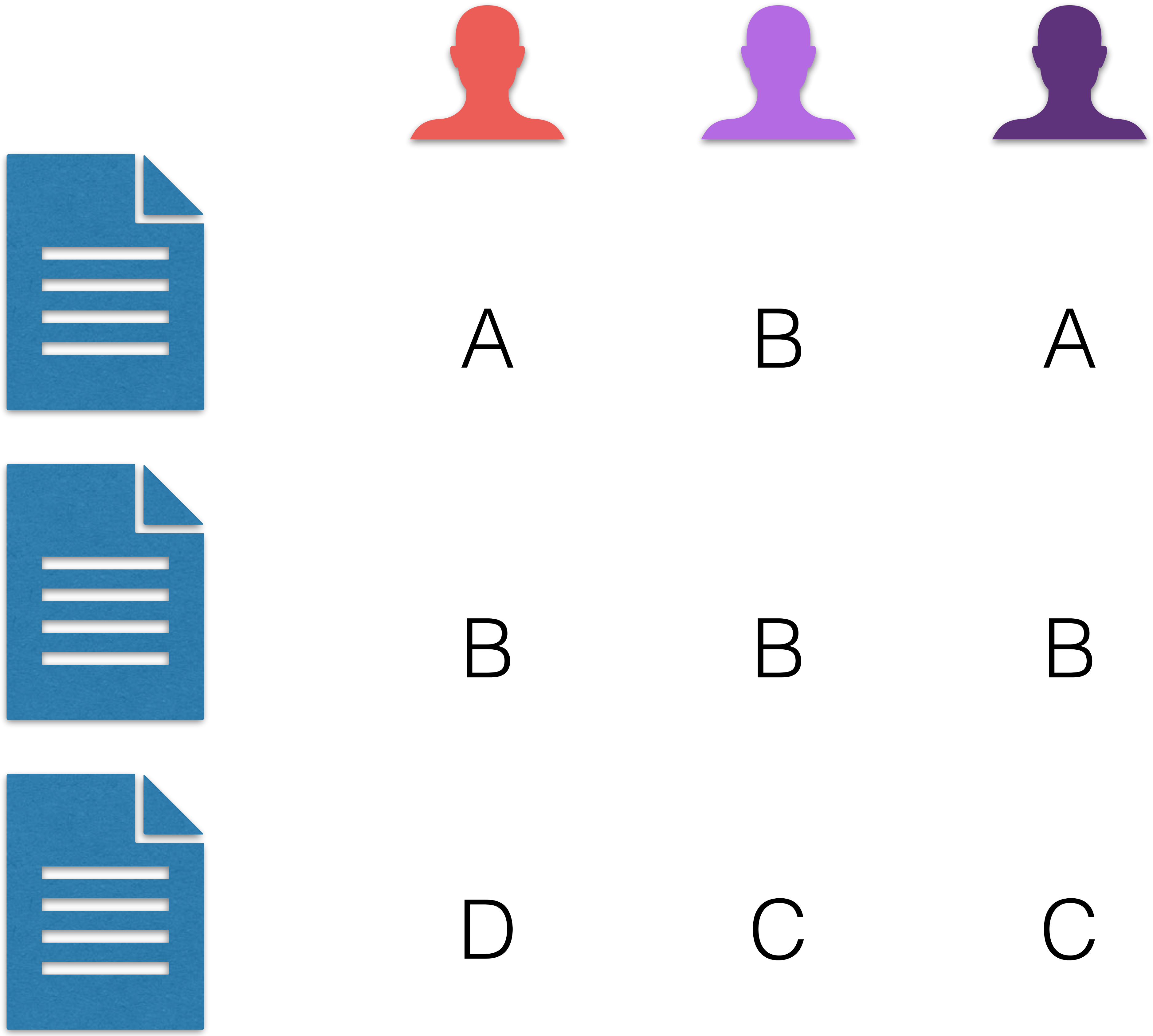
1 Aggregation



2 Filter



3 Learn directly from Raw Annotations

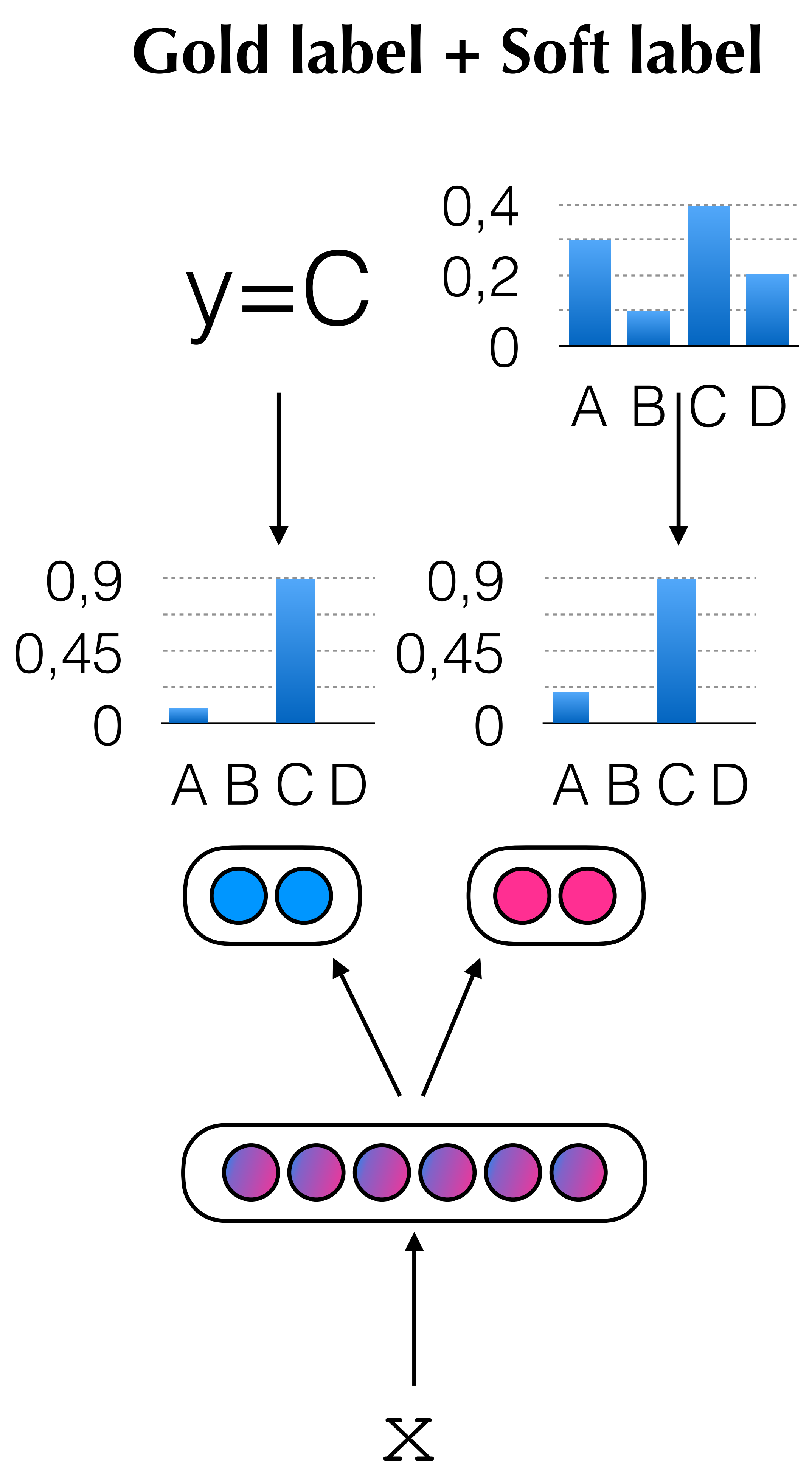


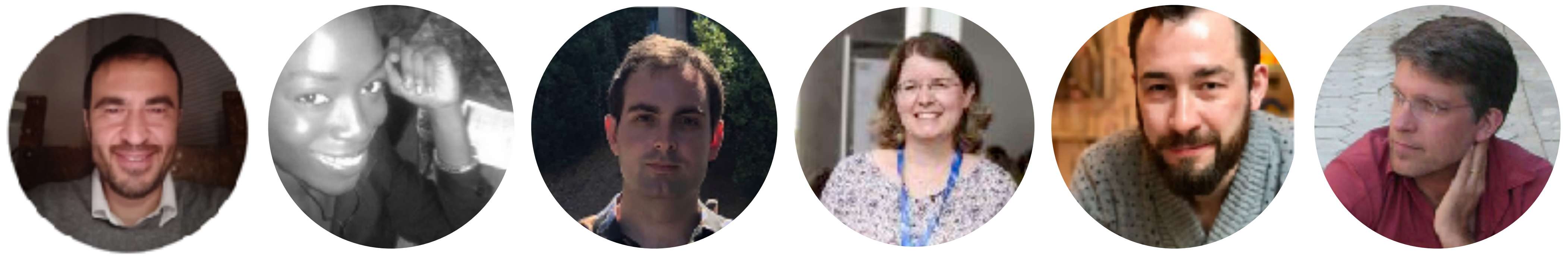
4

Augment Gold with Disagreement

A	B	A	A
B	B	B	B
D	C	C	C

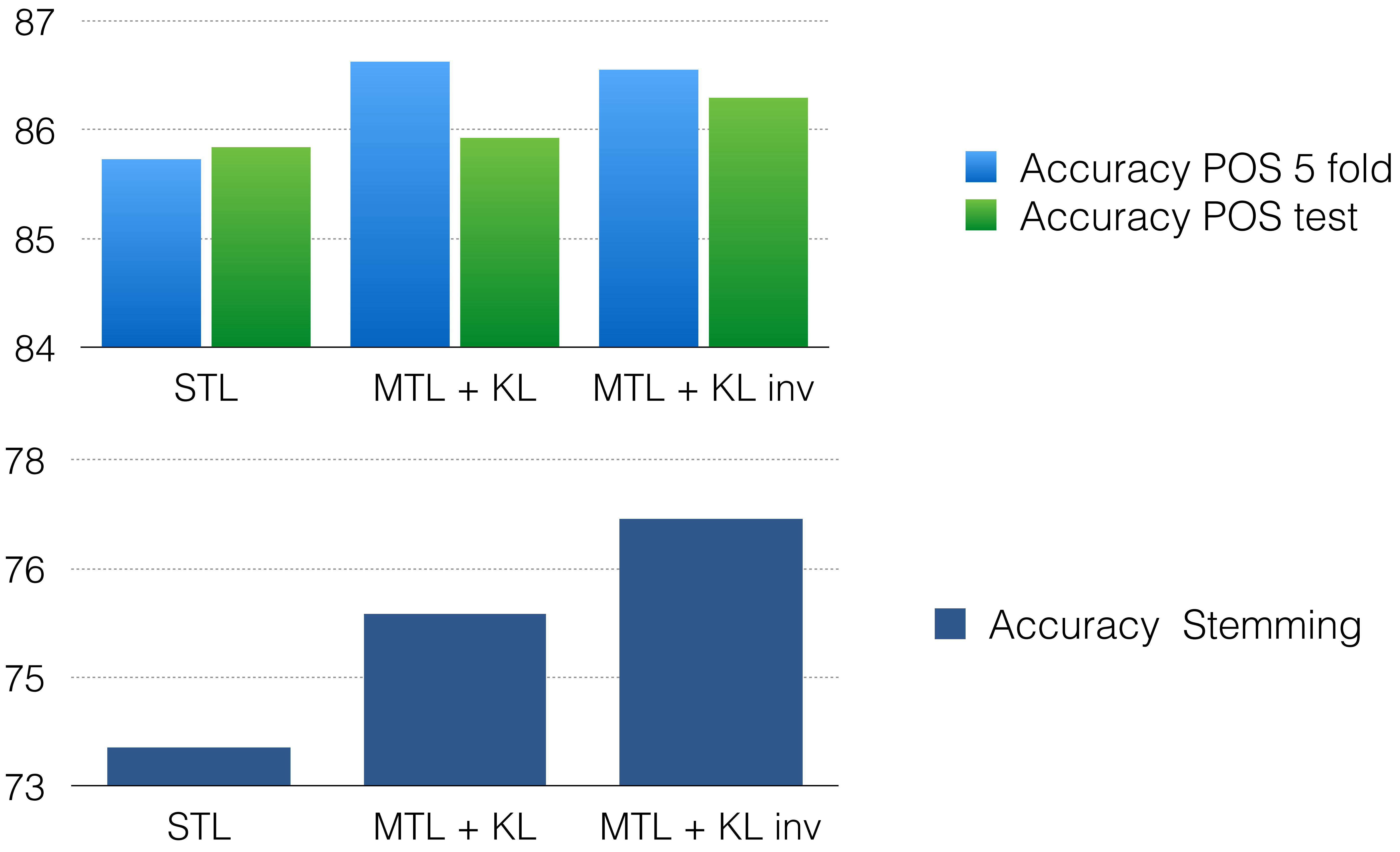
4 Soft-labels via Multi-Task Learning





(Tommaso Fornaciari, Alexandra Uma, Silviu Paul, Barbara Plank, Dirk Hovy, Massimo Poesio 2021 NAACL)

Results



$D_{KL}(P||Q)$ $D_{KL}(Q||P)$

Understanding Indirect Questions

- **Problem:** Humans often reply to polar questions w/o explicit use of Yes/No clues

  Q: Do you wanna crash on the couch?
  A: I gotta go home sometime

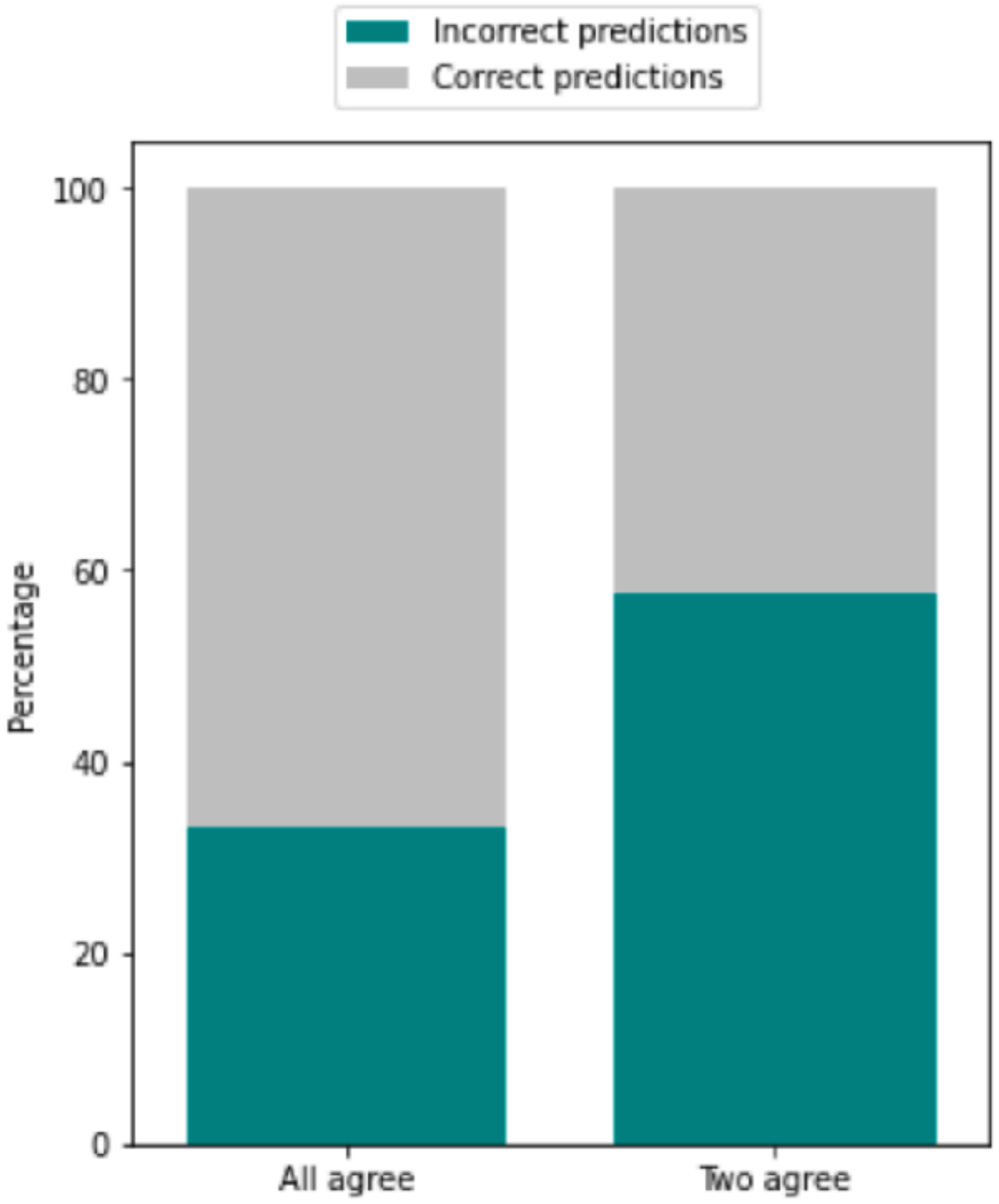
- **Dataset:** Friends-QIA dataset

Dataset	FRIENDS-QIA
All	5,930
Train	4,744
Dev	593
Test	593

(Damgaard, Toborek, Eriksen & Plank, 2021 CODI@EMNLP)

Most incorrect predictions on instances humans did not agree on

	Accuracy	F1-score
Majority baseline	49.07	16.46
Train on FRIENDS-QIA:		
CNN with BERT	**61.33**	45.65
CNN with BERT, multi-input	61.10	45.53



Correct and incorrect predictions of CNN with BERT vs. annotator agreement.

Does it help to embrace human disagreement?

	Accuracy	F1-score
Majority baseline	49.07	16.46
Train on FRIENDS-QIA:		
CNN with BERT	**61.33**	45.65
CNN with BERT, multi-input	61.10	45.53
CNN with BERT + crowd layer	60.32	**47.89**

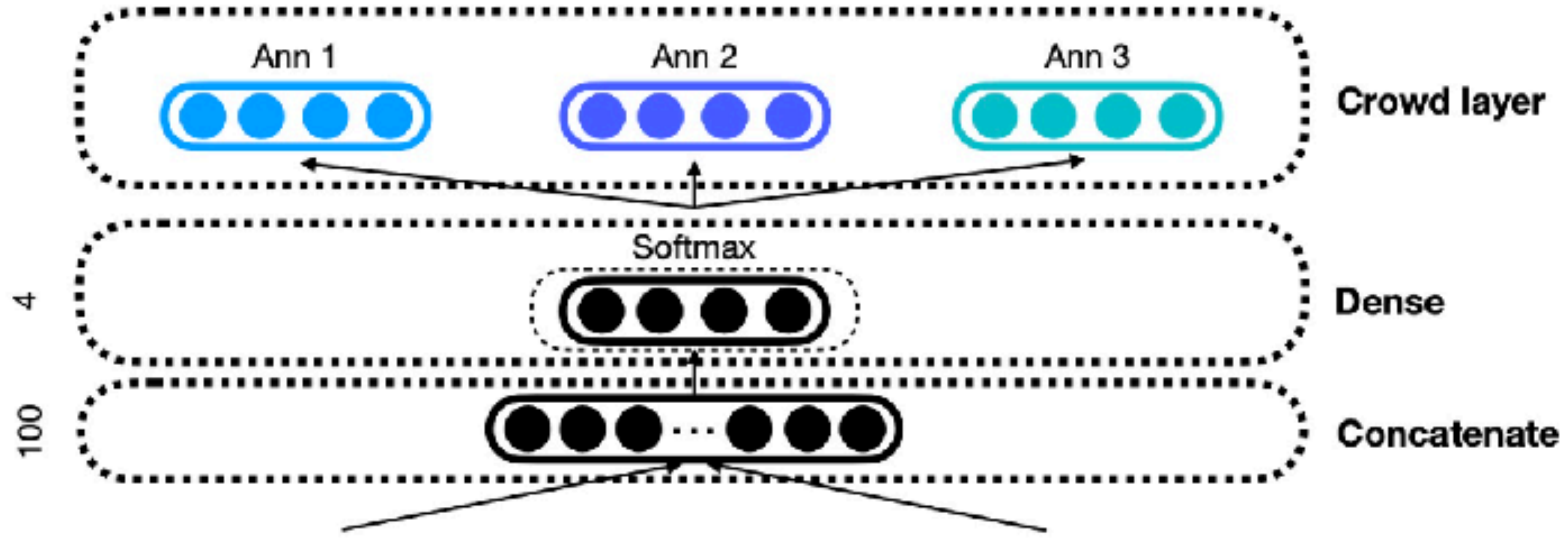
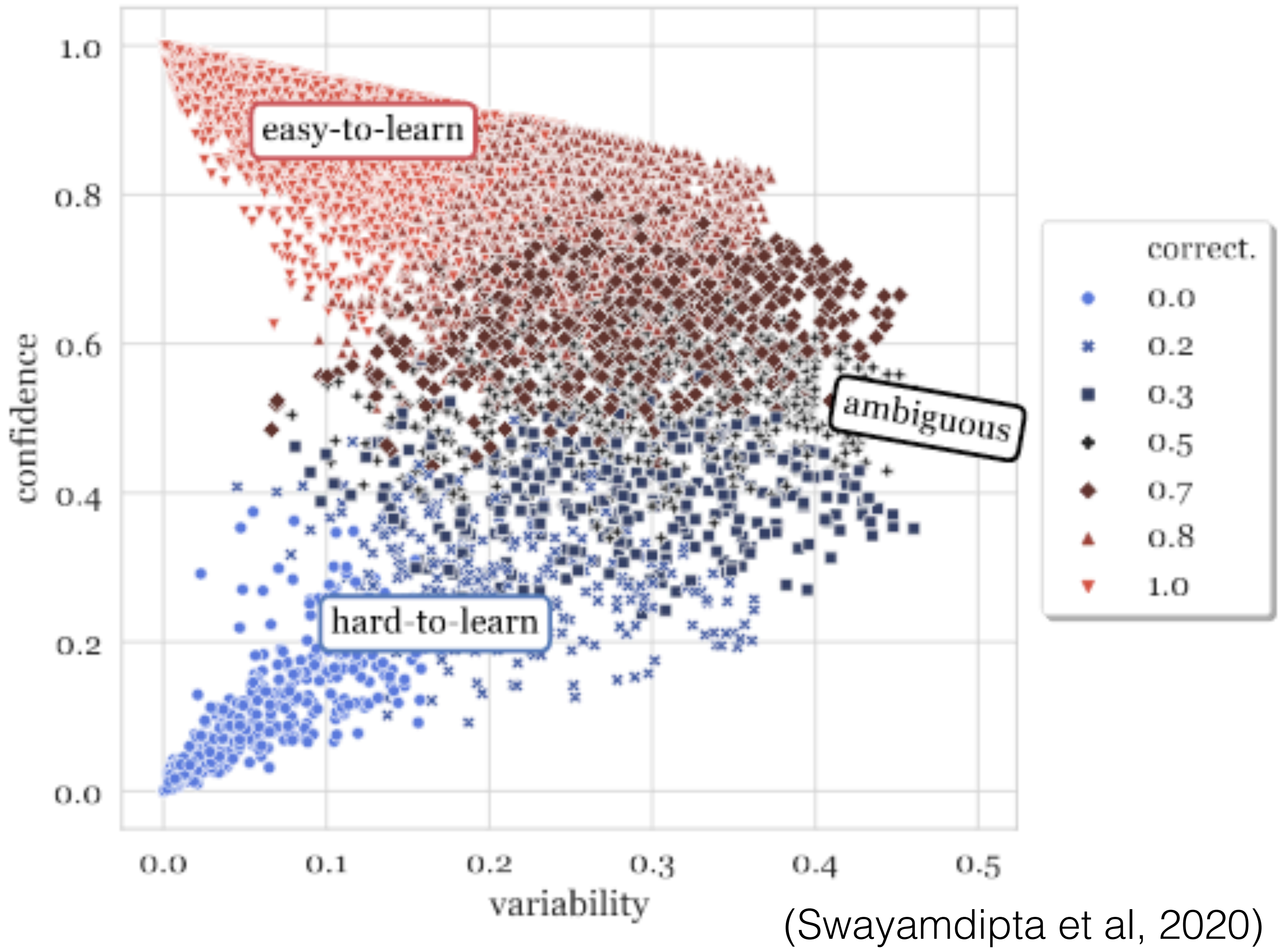


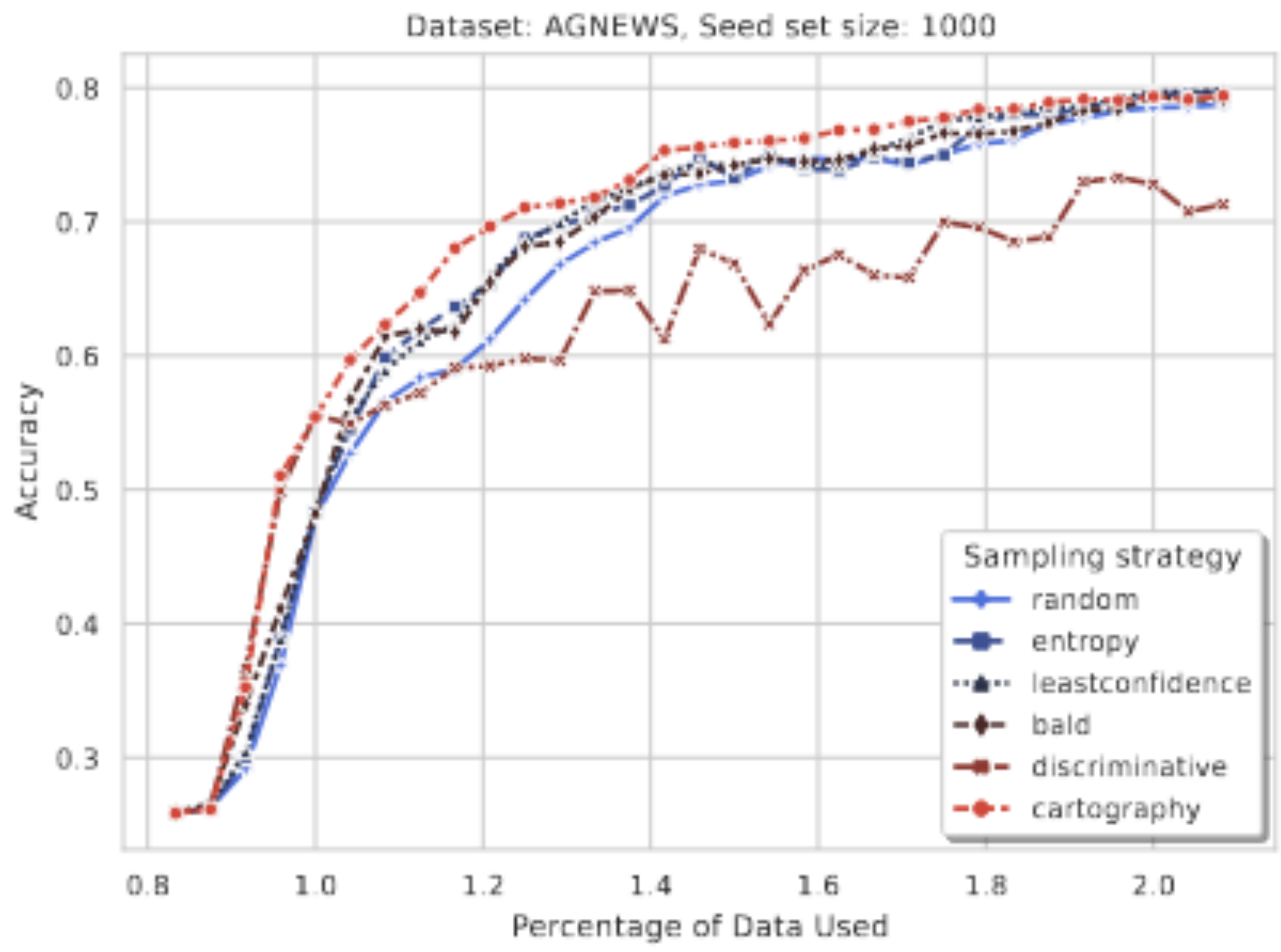
Figure 3: Illustration of deep learning from crowds proposed by Rodrigues and Pereira (2017).

Can we use the model uncertainty for selecting better data?

CAL: Learning with data maps & humans-in-the-loop

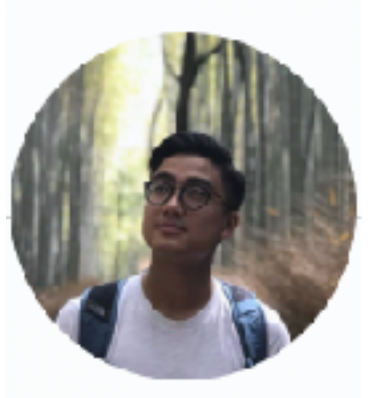
- **Problem:** Labeling data is costly. Can we find a better way to select effective data to give to a human annotator?





(Swayamdipta et al, 2020)

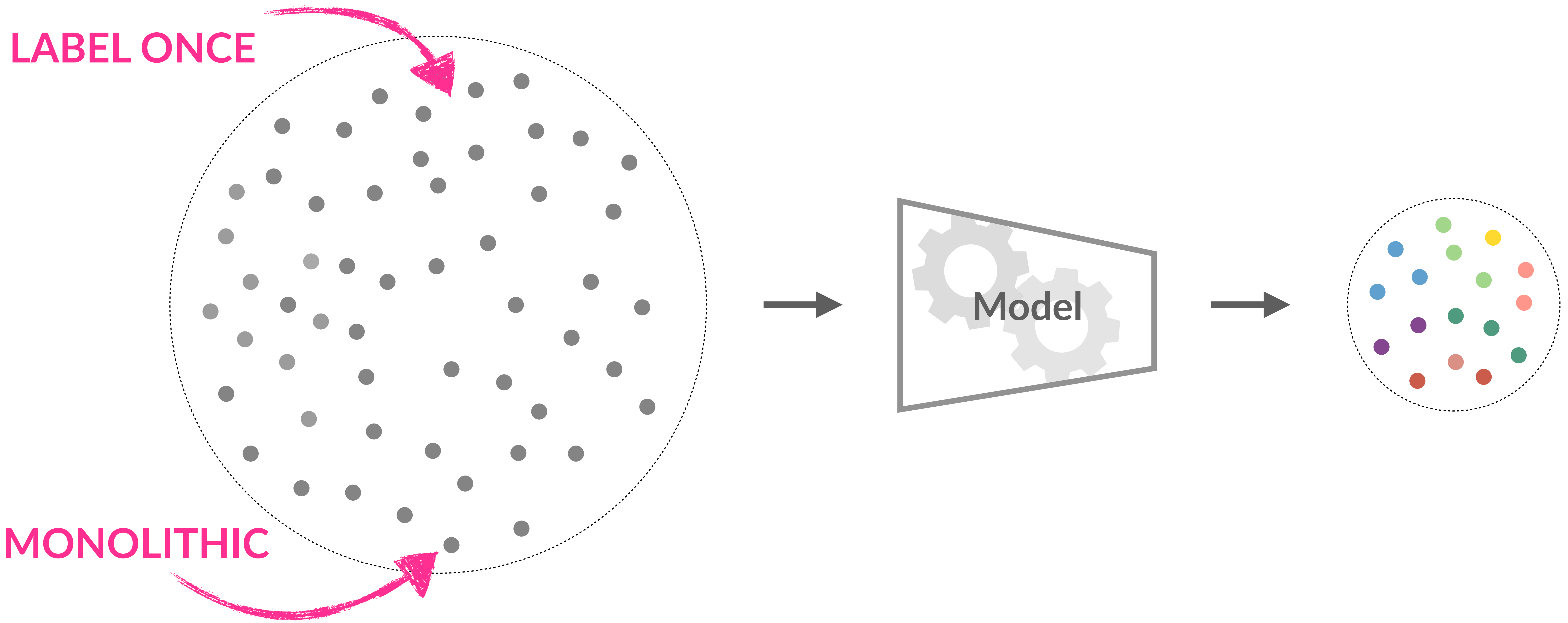
- **Key idea**: Data maps provide insights into training dynamics. We propose data maps for more effective active learning.

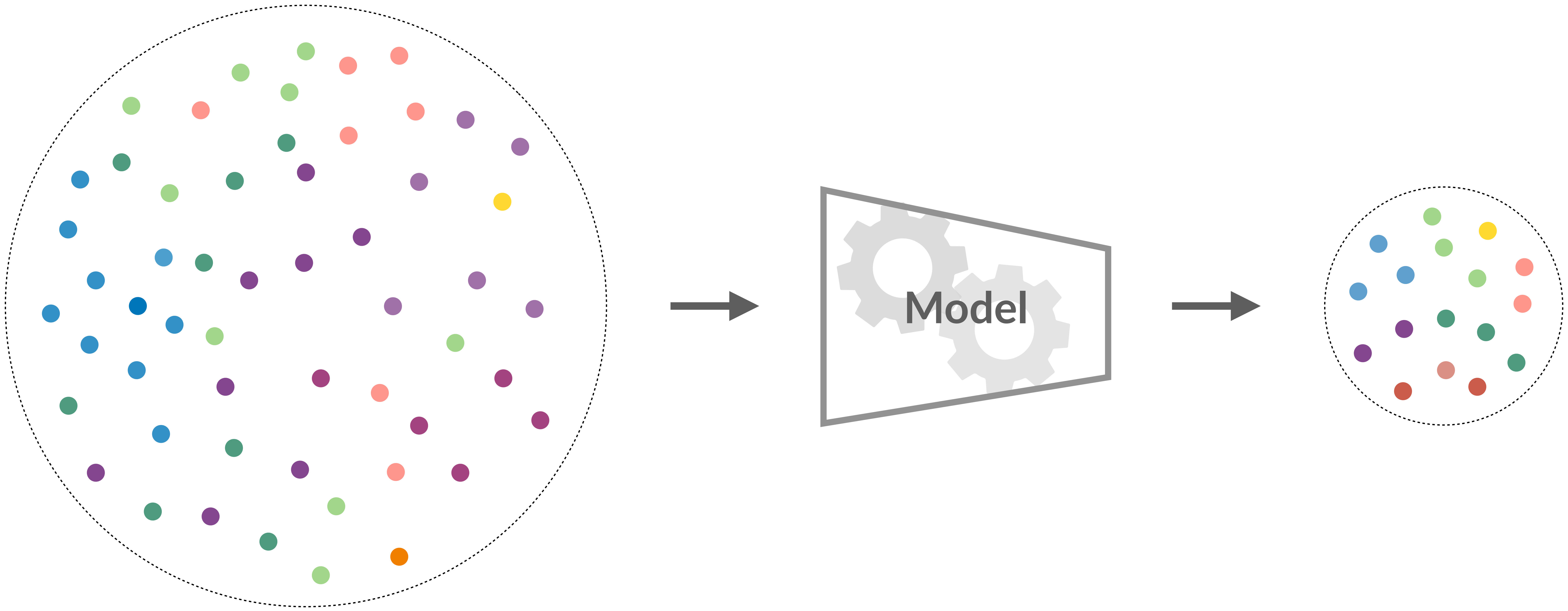


Zhang & Plank (EMNLP 2021 Findings). Cartography Active Learning

To wrap up…

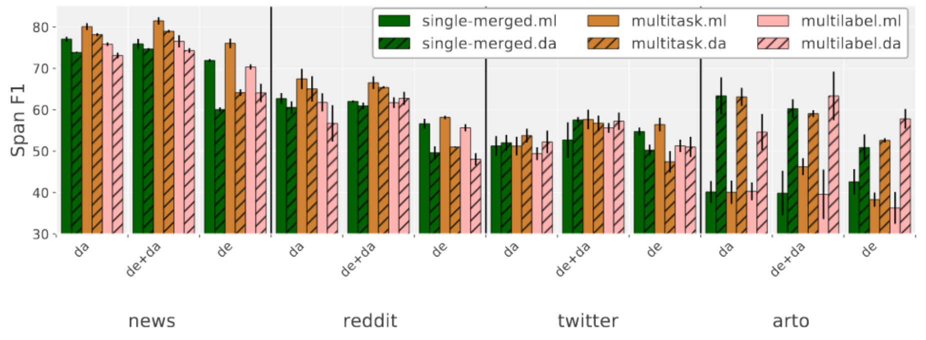
NLP today is often “monolithic processing”




Model

How can we create more inclusive NLP?

- Creation of dedicated in-language resources (data, modeling, evaluation)
- Transfer from better-resourced languages



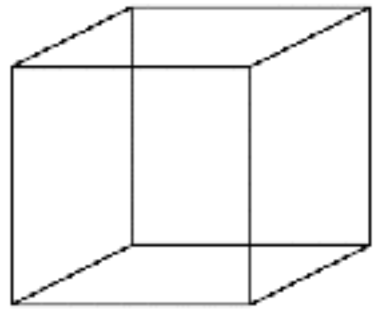
How can we create more efficient NLP systems?

- Data Selection
- Weak Supervision

How can we create more human-centred NLP?

- Learn from human disagreement

- Learn with humans in the loop



IT-UNIVERSITETET I KØBENHAVN

Thank You

Have a great ALTA 2021 !

Tackling scarce & biased data for more inclusive NLP

Barbara Plank, IT University of Copenhagen, NLPnorth

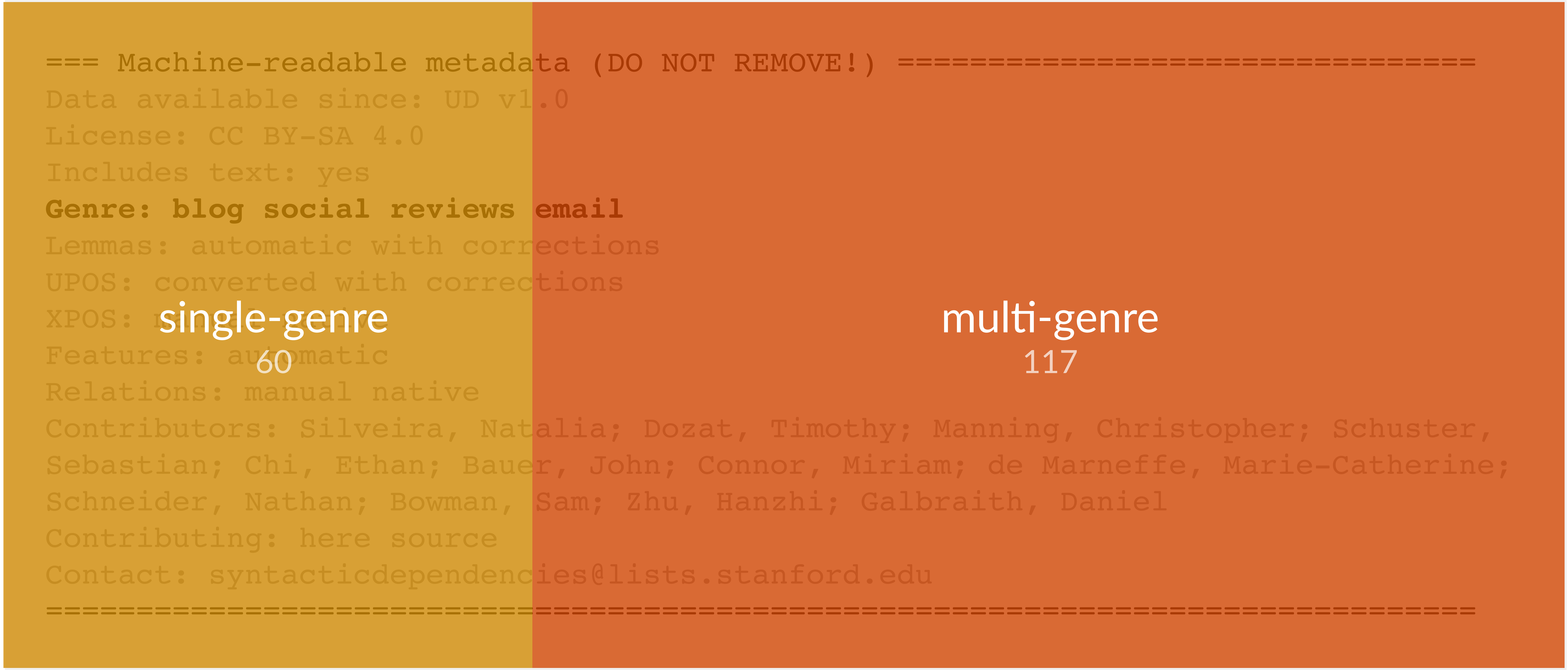
Supported by: DANMARKS FRIE FORSKNINGSFOND amazon NVIDIA

Appendix

Follow-up: Nested NER for English

- Back then, we had German data with 2-level annotation
- New: Ringland et al., 2019 ACL: up to 6 layers, Penn TB WSJ
- Plank 2021 ACL Findings: **English Web TB**, GermEval style entities (4 coarse types) over 12k sentences and **5 domains**

Figure 2: In-language cross-domain evaluation.


=== 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:
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
==
single-genre
60
multi-genre
117