Why Language Variety Matters and How To Embrace it in Our Models

Barbara Plank
MaiNLP lab @ Center for Information and Language Processing (CIS)
LMU Munich
&
NLPnorth lab @ Department of Computer Science
IT University of Copenhagen

July 29, 2022 - Lisbon

LxMLS 2022

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

IT UNIVERSITY OF COPENHAGEN

Example: NLP for well-resourced languages

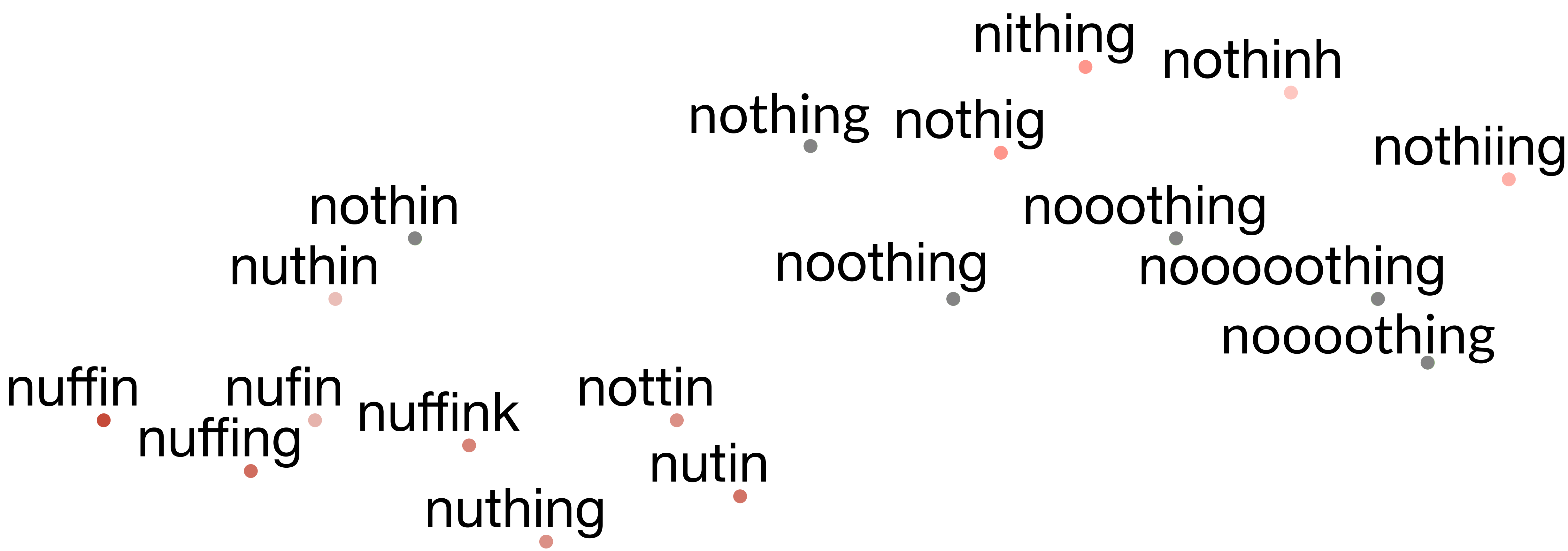
Language Coverage of Digital Assistants in 2021 (Europe)

https://slatorlalingua.com/language-technology/language-support-voice-assistants-compared/

Src: Galician, ELE project example

Natural Language is characterised by High Variability

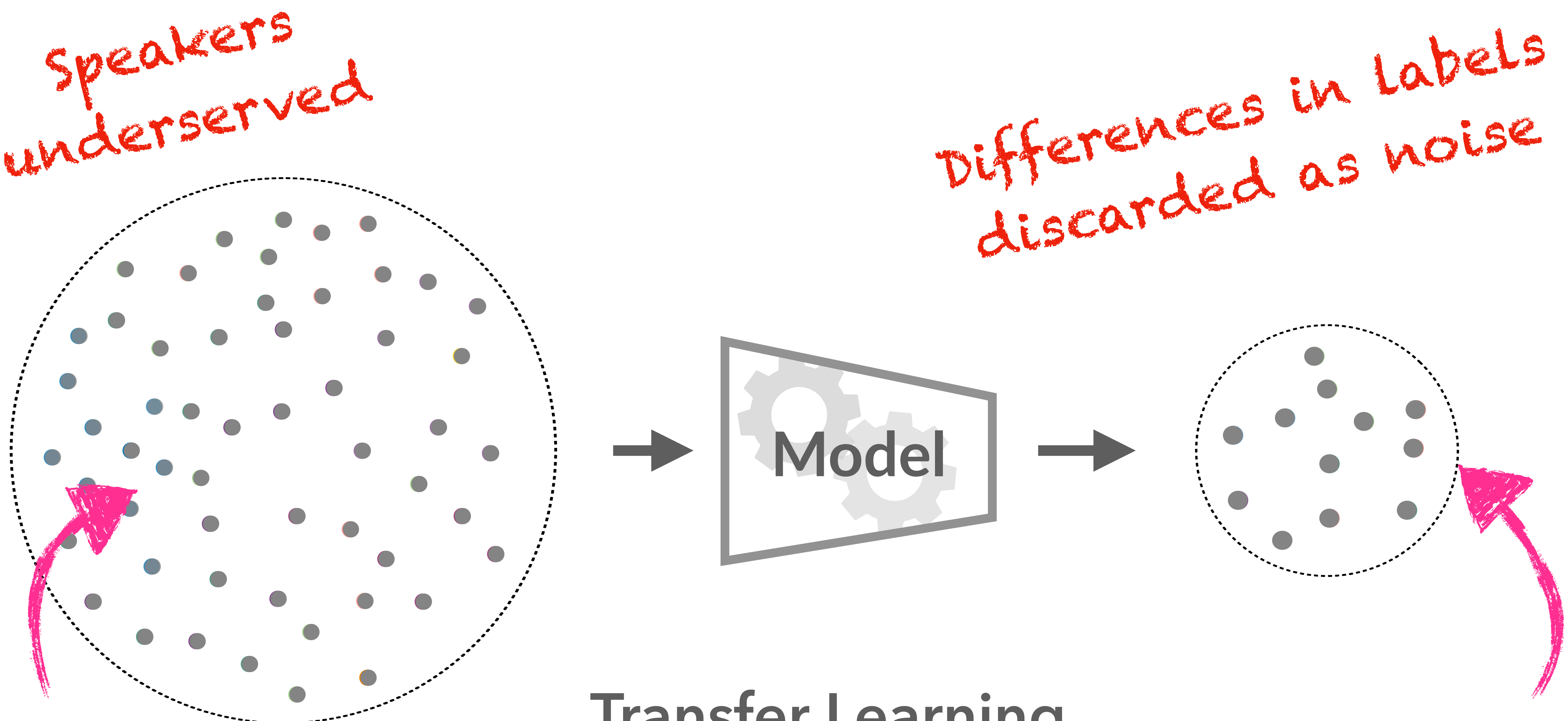
- The way we express a message **carries social meaning**



- **Limitation**: More **variation** → **higher error rates** in NLP

Nothing and **spelling choices** in Reddit (Nguyen & Grieve, 2020)

Need To Account for Language Variability



Transfer Learning

Learning w/ Human Label Variation

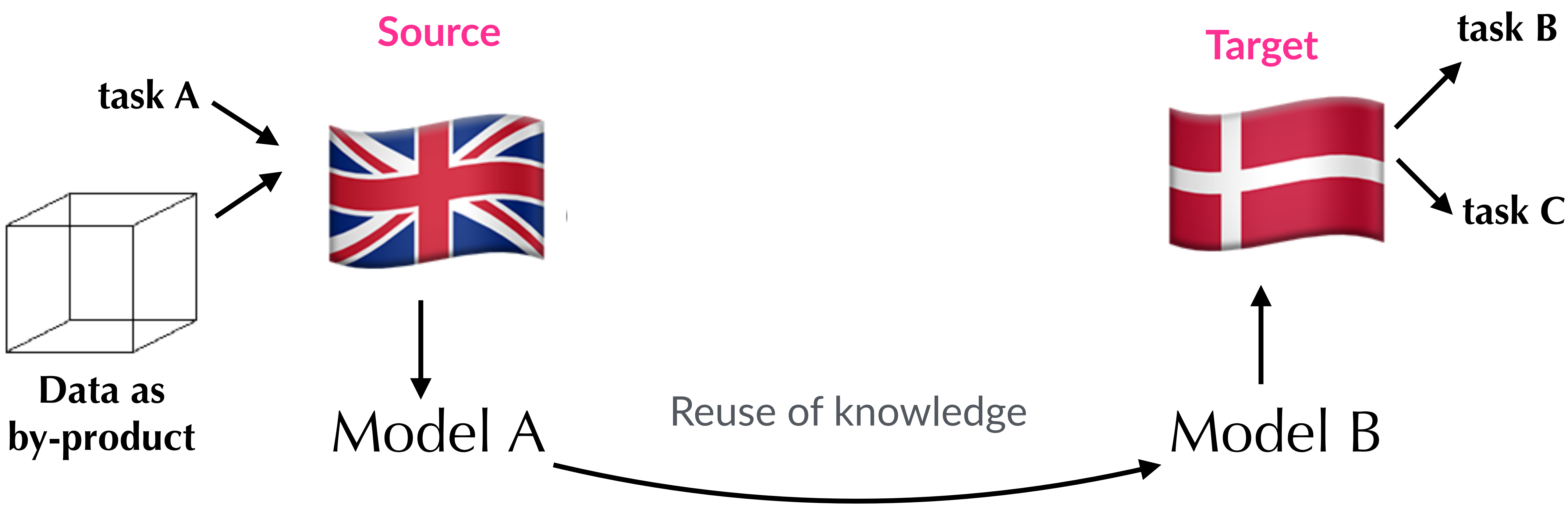
Input: language variability not recognised

Output: only standardised categories accepted

Outline

- Transfer Learning Overview
- Three selected case studies
  - [Paper 1] Data Selection
  - [Paper 2] Multi-task Learning
  - [Paper 3] Learning with Human Label Variation
- Conclusion: Outro & Moving Forward

Transfer Learning (TL): Crossing the Gulf



CROSS-DOMAIN: Generalize to new text variety

CROSS-LINGUAL: Generalize to new language variety

MULTI-TASK: Leverage information from different tasks in learning

FORTUITOUS: Leverage other data/by-products as signal (incidental supervision)

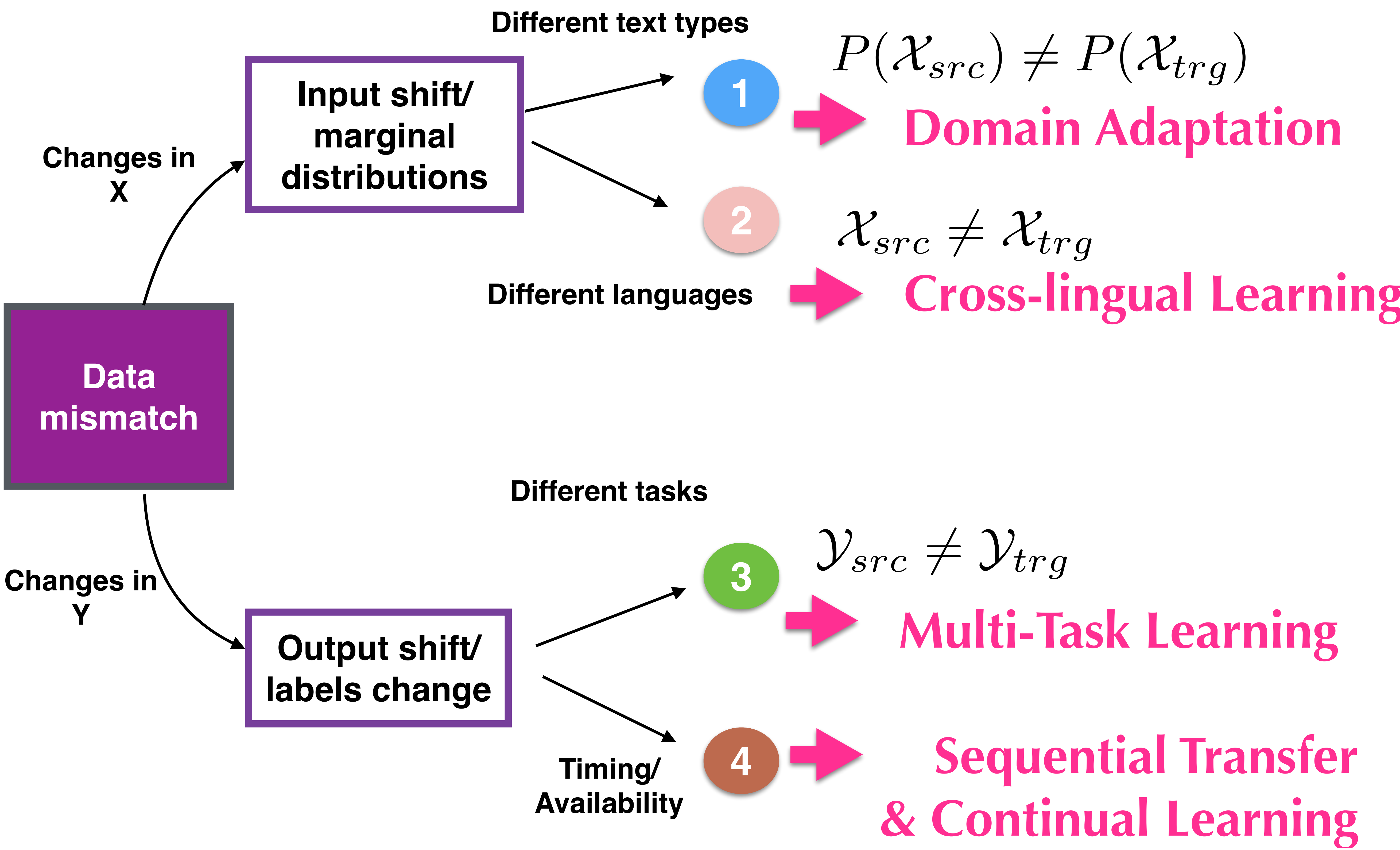
Dimensions of TL

Data domain $\mathcal{D} = \{\mathcal{X}, P(\mathcal{X})\}$
with $\mathcal{X}$ the feature space

~ Notation ~

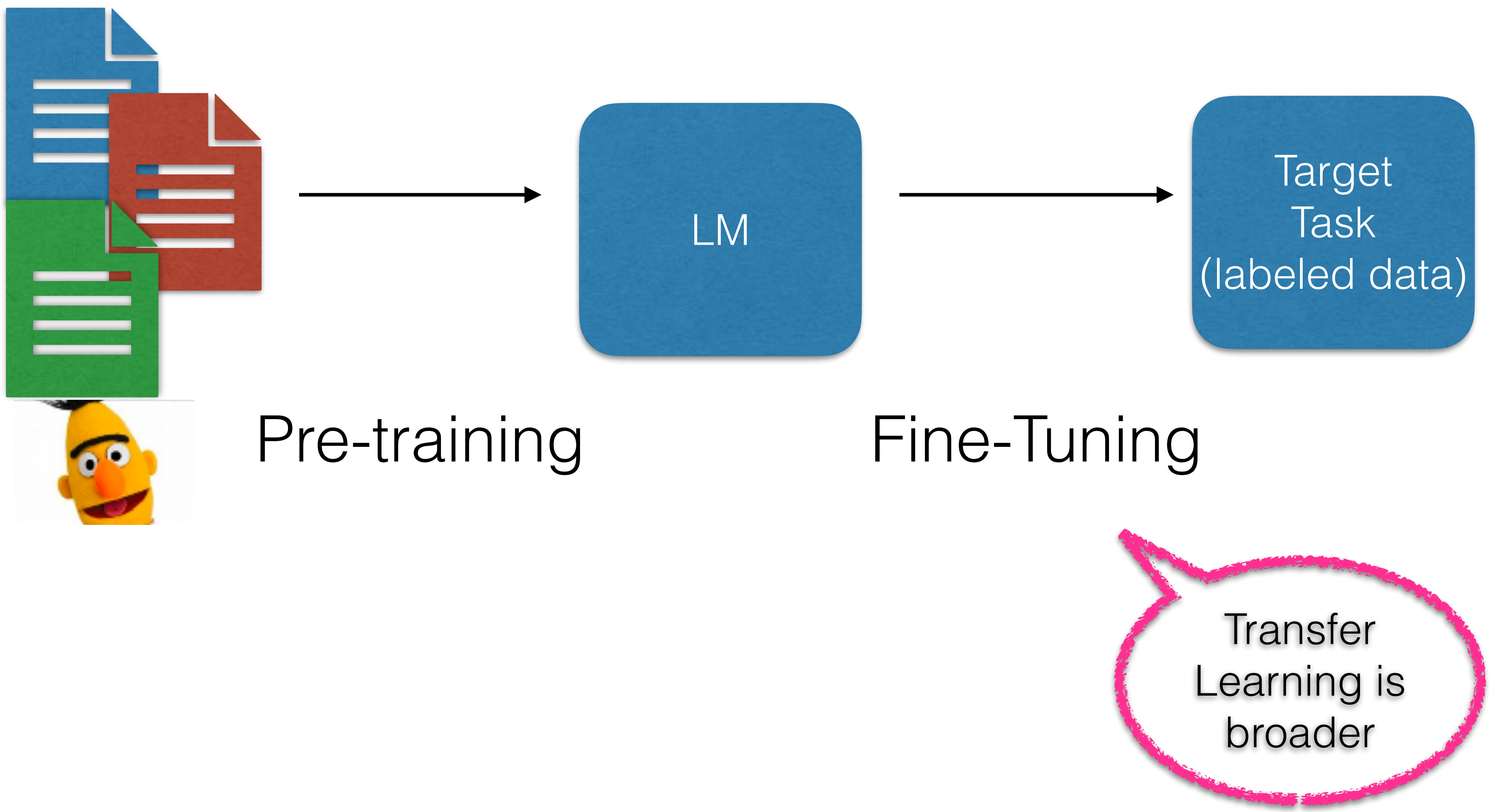
Task $\mathcal{T} = \{\mathcal{Y}, P(\mathcal{Y}|\mathcal{X})\}$
where $\mathcal{Y}$ is the label space

What Type of Data Mismatch (1/2)

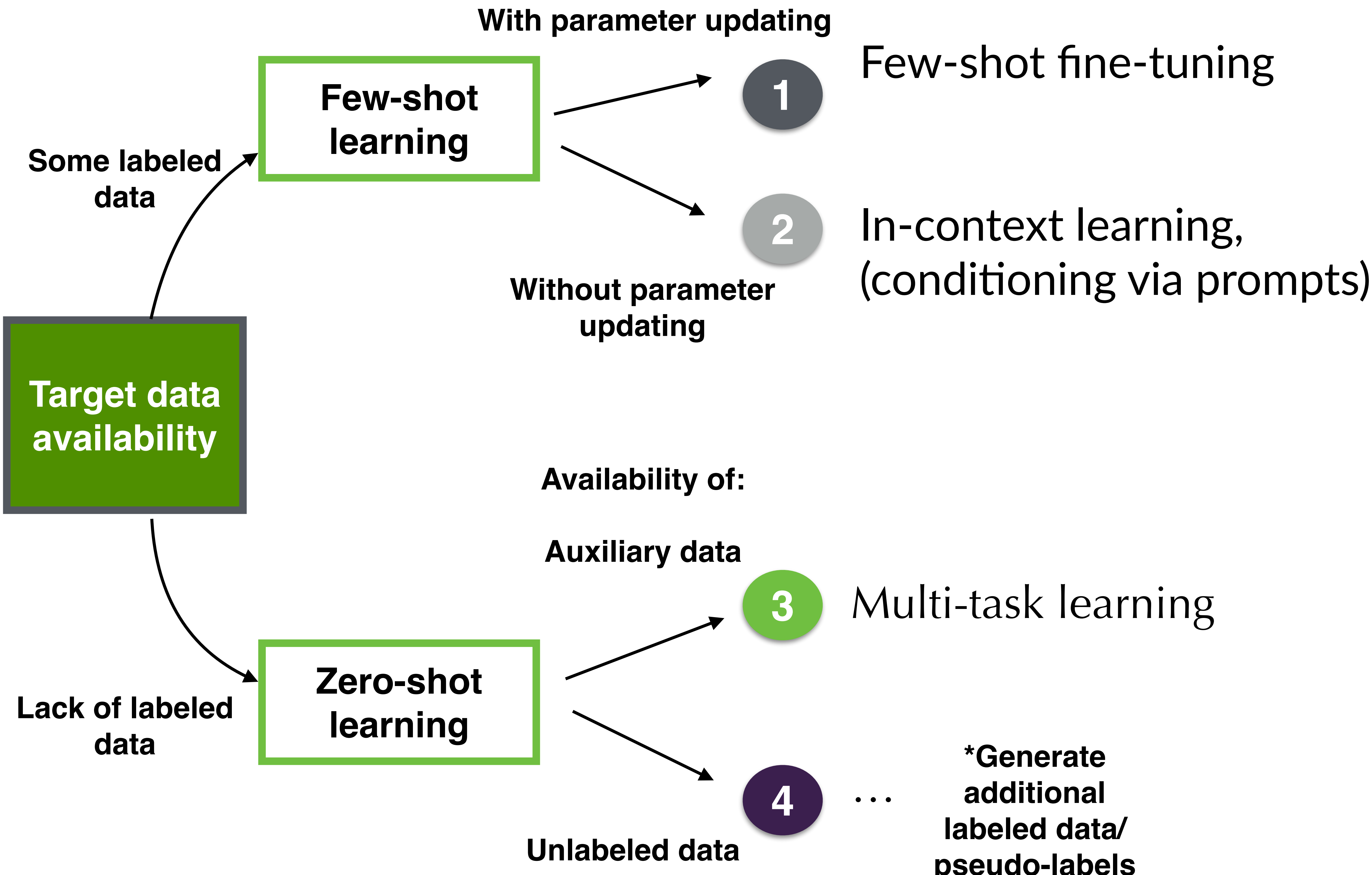


Myopic View: (Sequential) Transfer Learning = Fine-tuning

- = Largely today's omnipresent **Pre-train & Fine-tune paradigm (aka sequential transfer)**

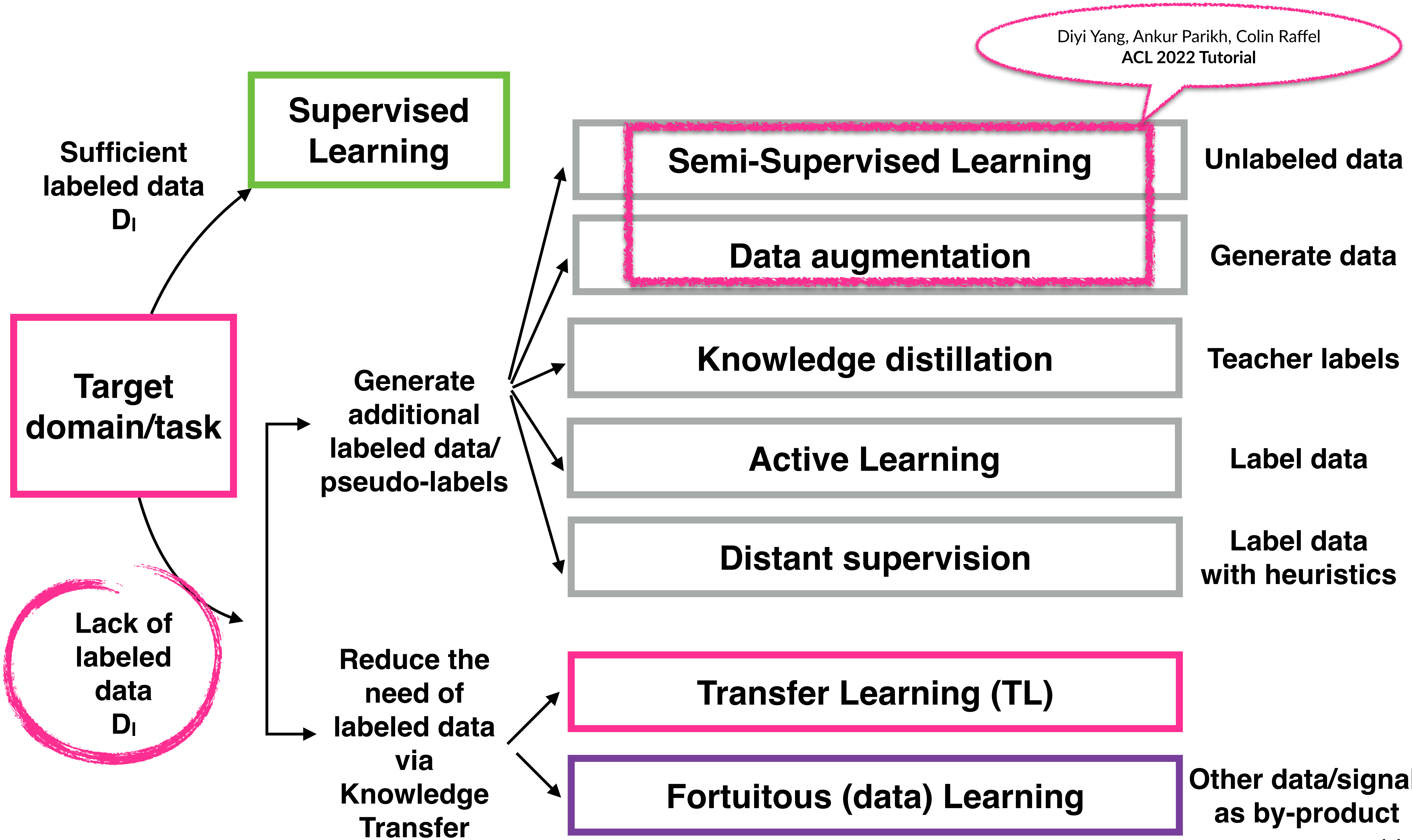


What is the Resource Availability (2/2)



*see related ML methods - next slide

Relationship to other learning paradigms

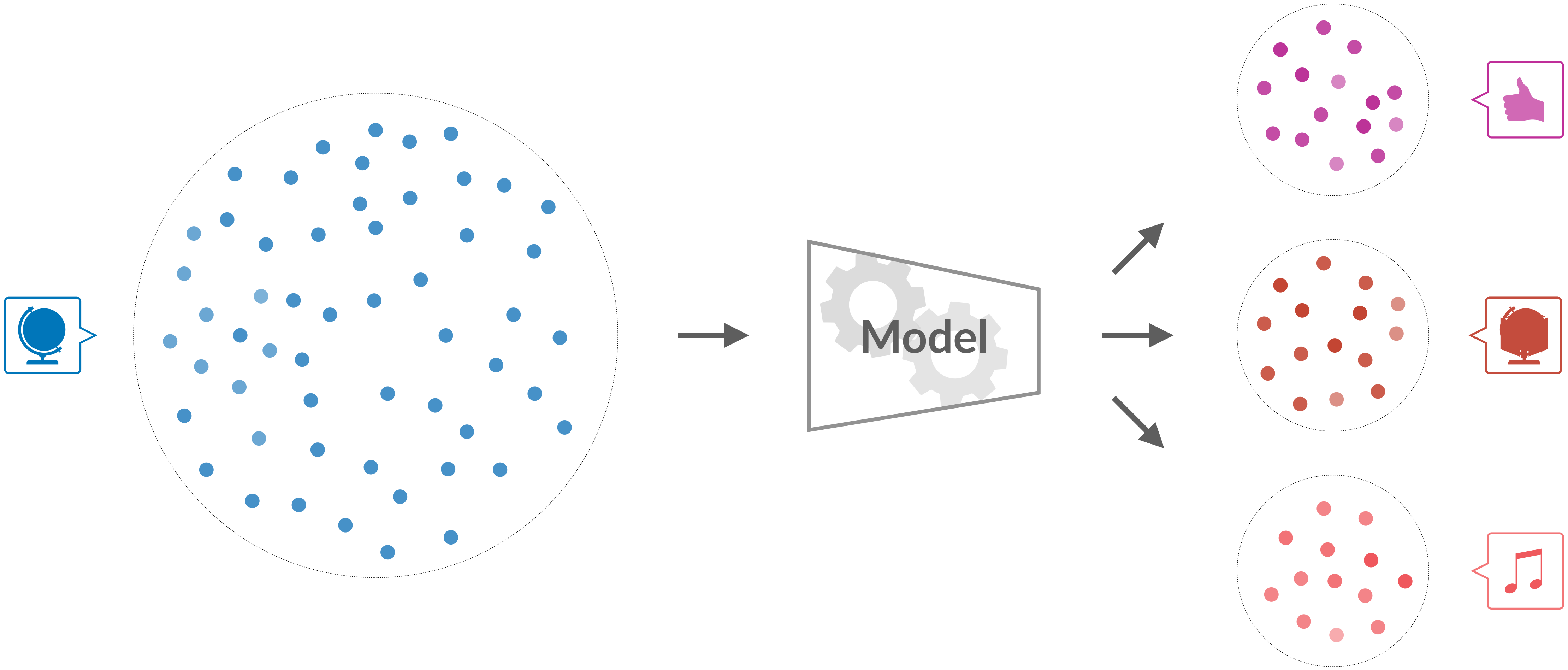


Outline

- Transfer Learning Overview
- Three selected case studies
    - [Paper 1] Data Selection
    - [Paper 2] Multi-task Learning
    - [Paper 3] Learning with Human Label Variation
- Conclusion: Outro & Moving Forward

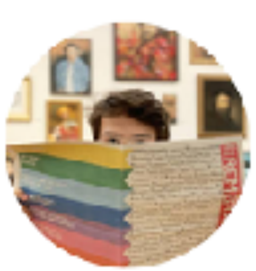
Roadmap for the Three Use Cases

1. How useful is (fortuitous) meta-data for low-res parsing?
2. How effective are non-English auxiliary tasks for transfer?
3. How can we integrate human label variation in NLP?


Model

Genre as Weak Supervision for Cross-lingual Dependency Parsing

Max Müller-Eberstein and **Rob van der Goot** and **Barbara Plank**
Department of Computer Science
IT University of Copenhagen, Denmark
mamy@itu.dk, robv@itu.dk, bapl@itu.dk

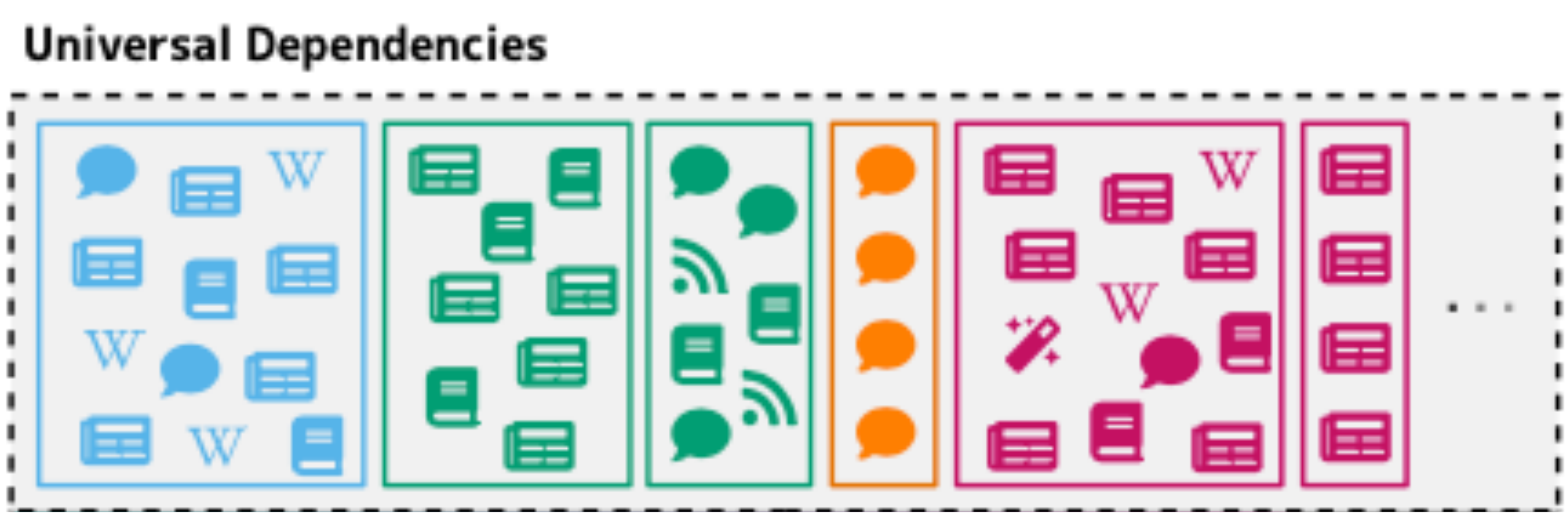



EMNLP, 2021

Part 1

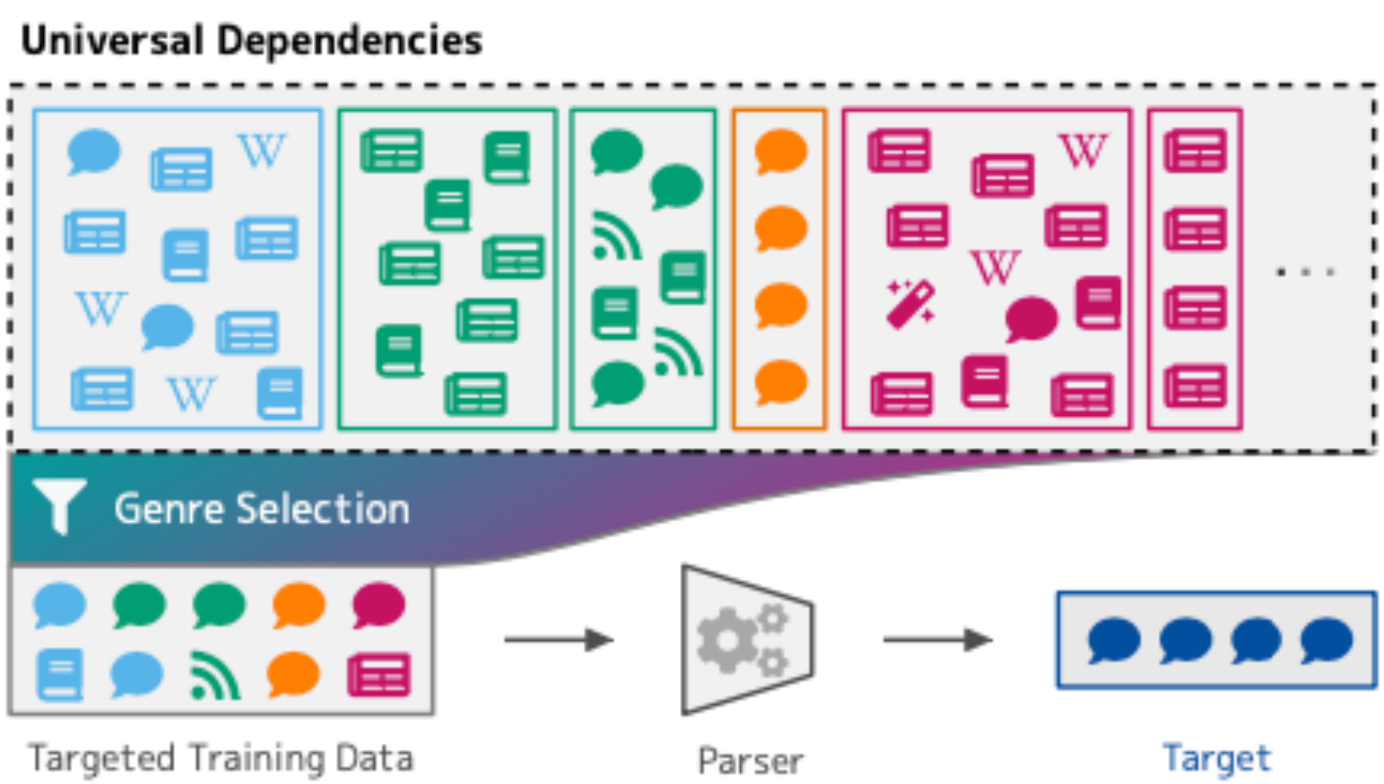
Data Selection: How to Find Task-Specific Data?

- **Problem & Motivation:**
  - A single parser trained on 100+ languages is suboptimal (training time, accuracy); also: for a practitioner it is difficult to choose appropriate training material.
  - Given Universal Dependencies (over 200 languages), how can we find better targeted training data?
    - **Less is more?**



Key Idea: Genre as Fortuitous treebank-level meta-data

- **Research Questions:**
  - RQ1: To what extent does **genre** aid better proxy target data?
  - RQ2: Is genre **inherently** captured in multilingual LMs?



Domain **Genre** Register

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

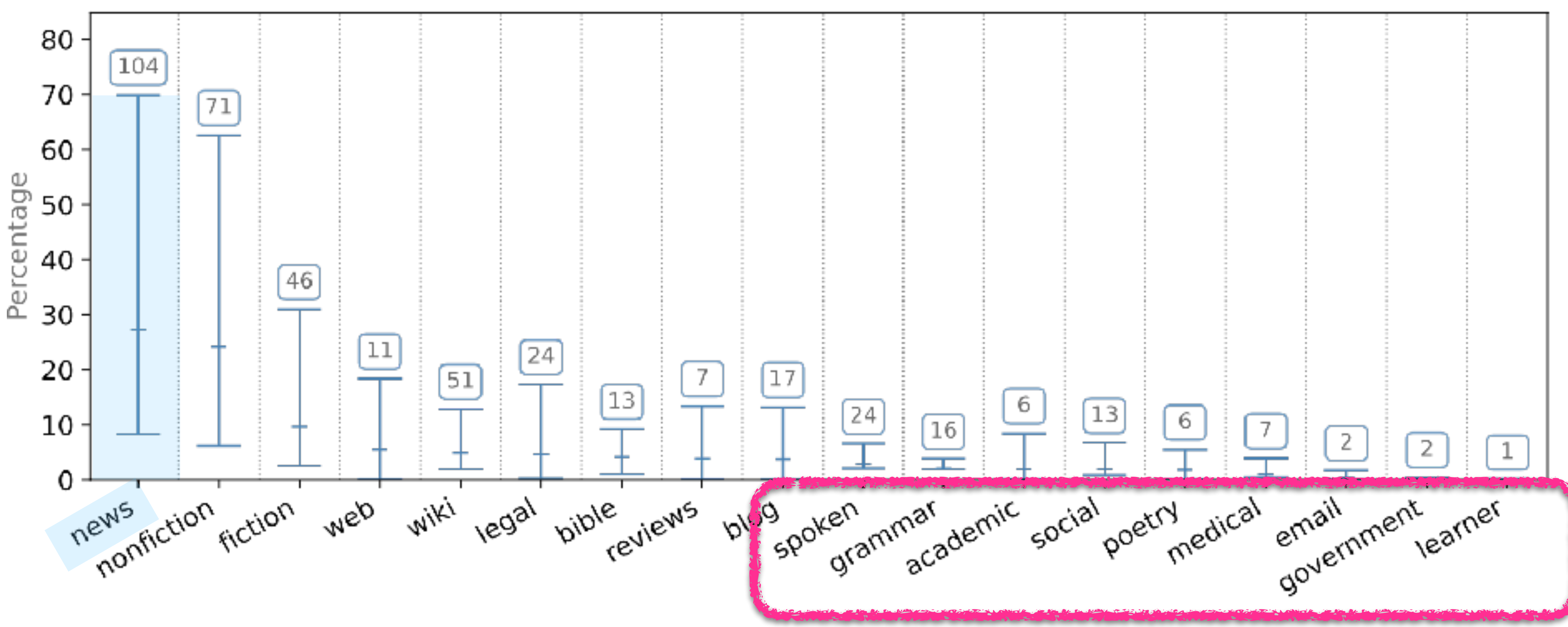
18 community-provided categories in UD

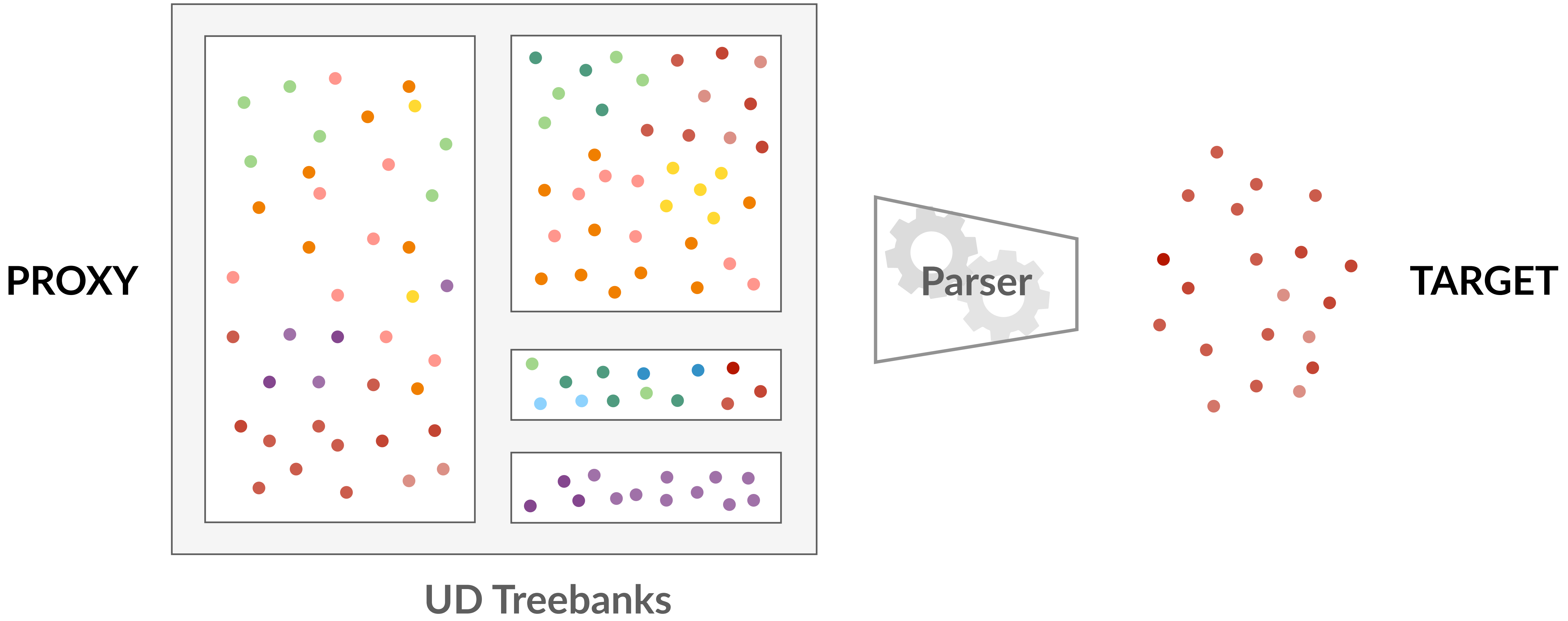
⚡ Meta-data "Failure"? No, Opportunity!

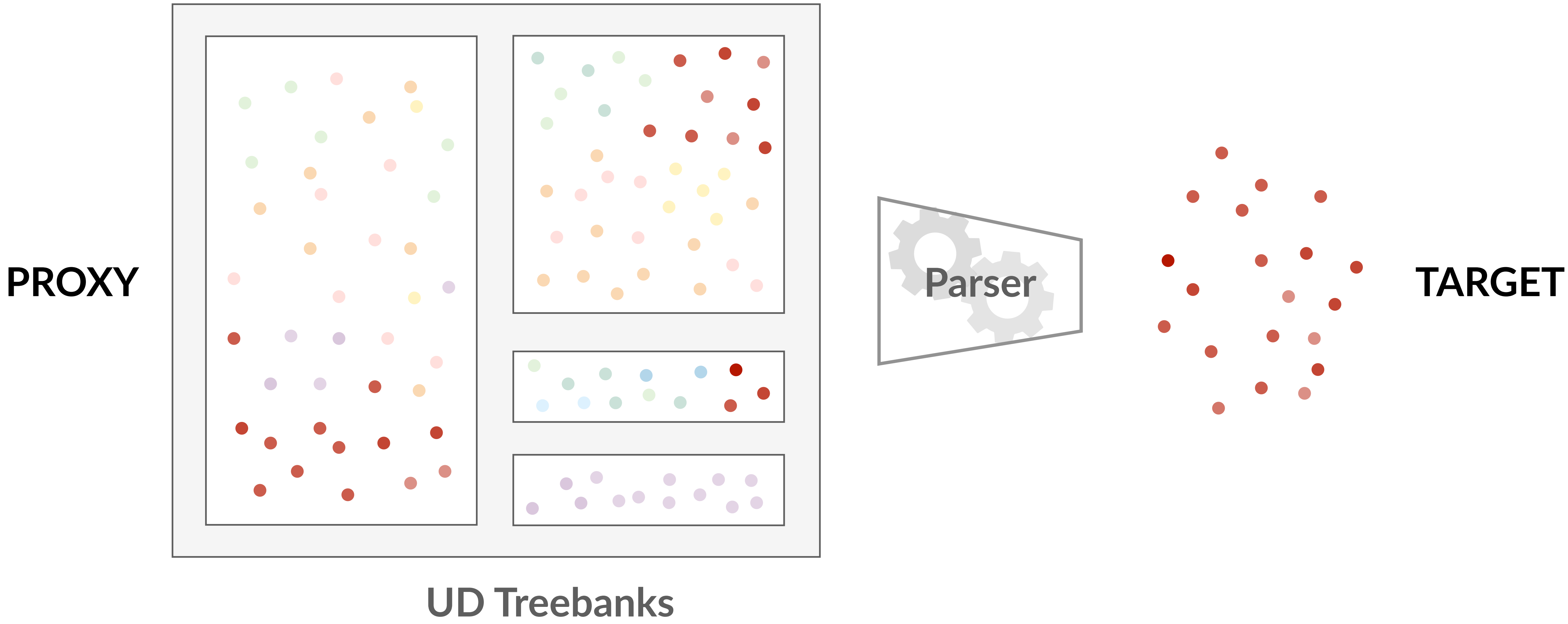
The treebanks in UD v2.5 are also heterogeneous with respect to the type of text (or spoken data) annotated. A very coarse-grained picture of this variation can be gathered from Table 5, which specifies the number of treebanks that contain some amount of data from different "genres", as reported by each treebank provider in the treebank documentation. The categories in this classification are neither mutually exclusive nor based on homogeneous criteria, but it is currently the best documentation that can be obtained.

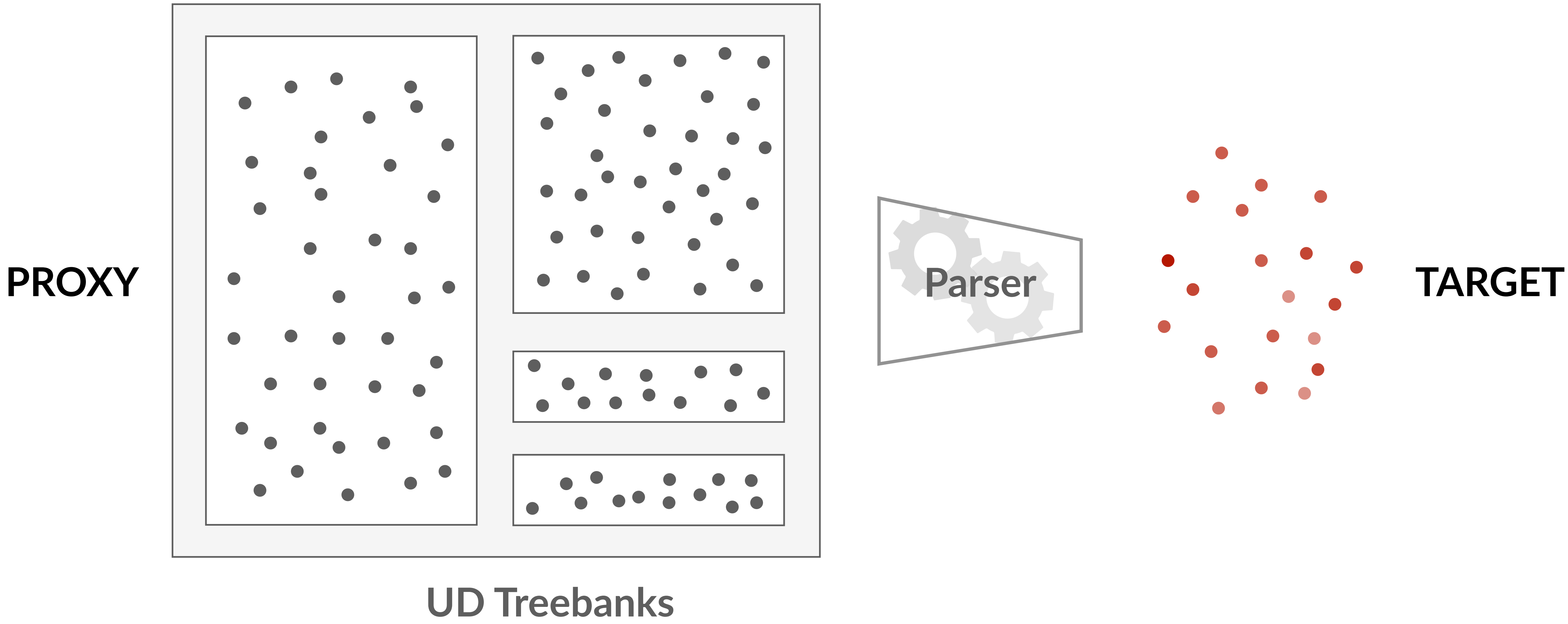
Nivre et al., (2020), https://aclanthology.org/2020.lrec-1.497.pdf

Genre Distribution in UD

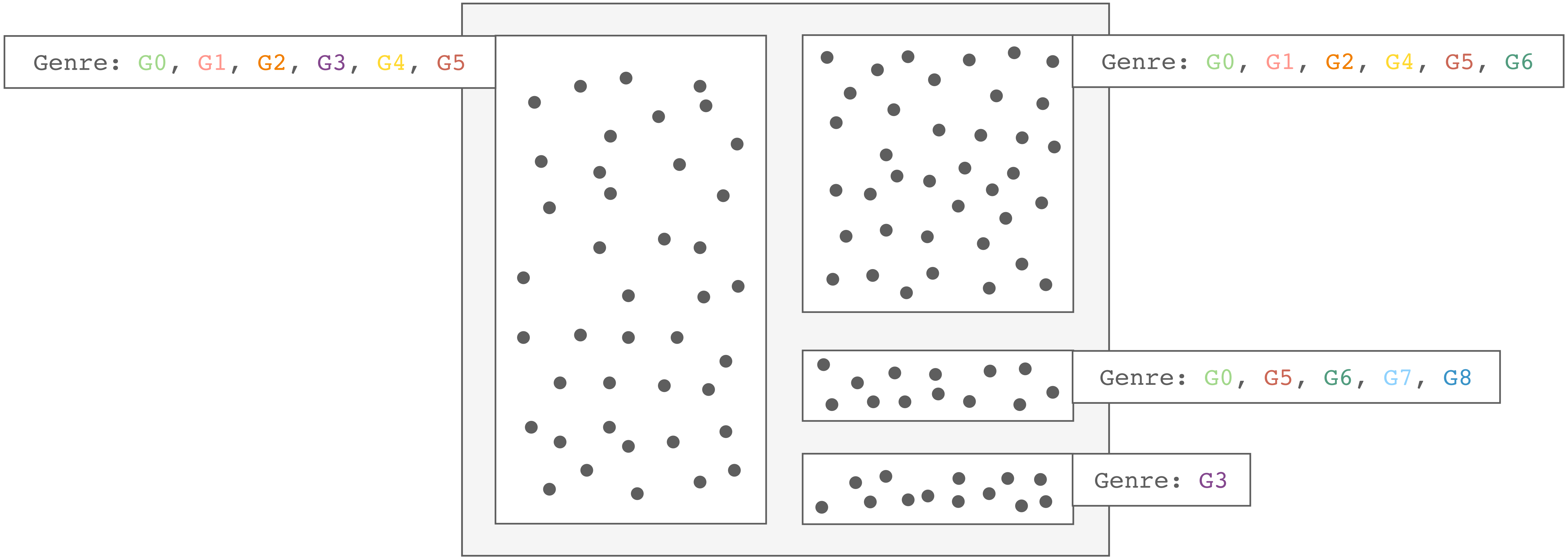



PROXY
UD Treebanks
Parser
TARGET


PROXY
Parser
TARGET
UD Treebanks

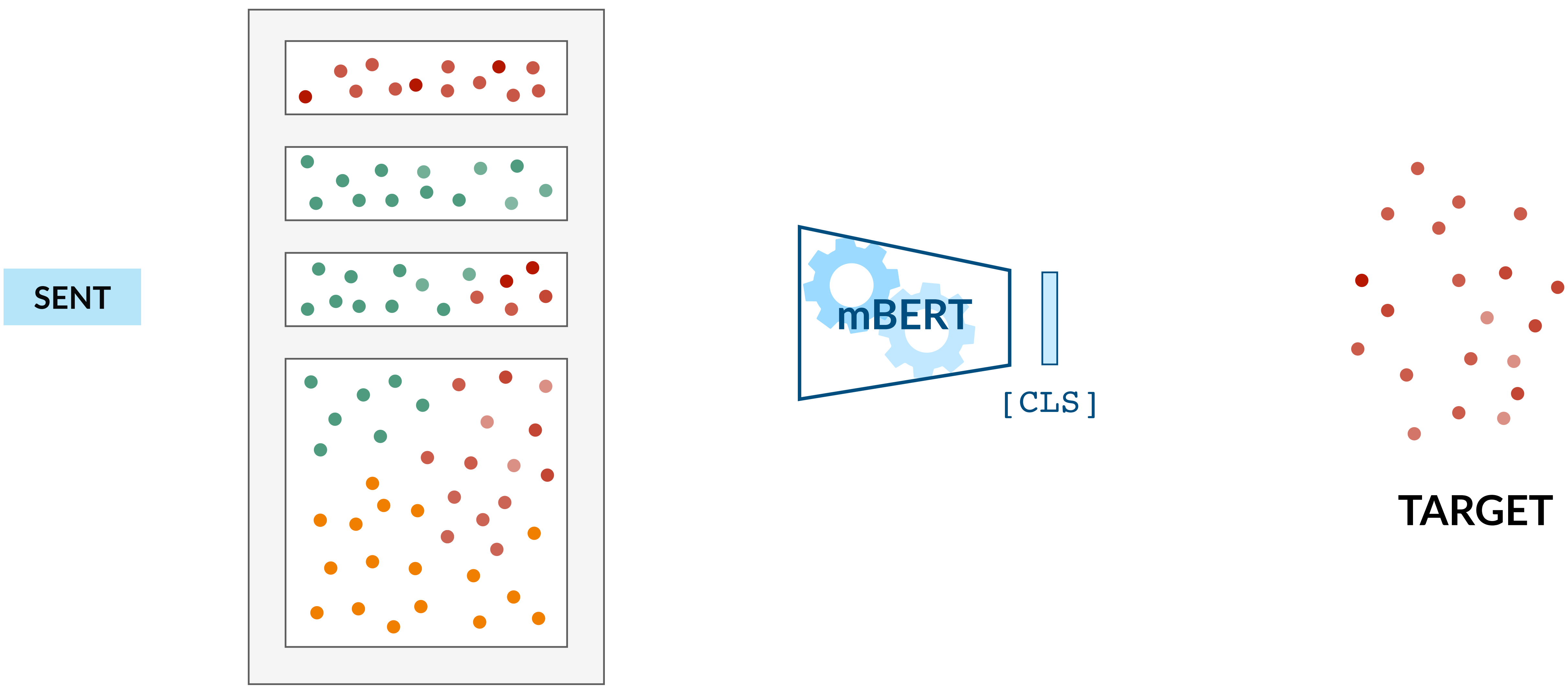

PROXY
UD Treebanks
Parser
TARGET

Targeted Data Selection

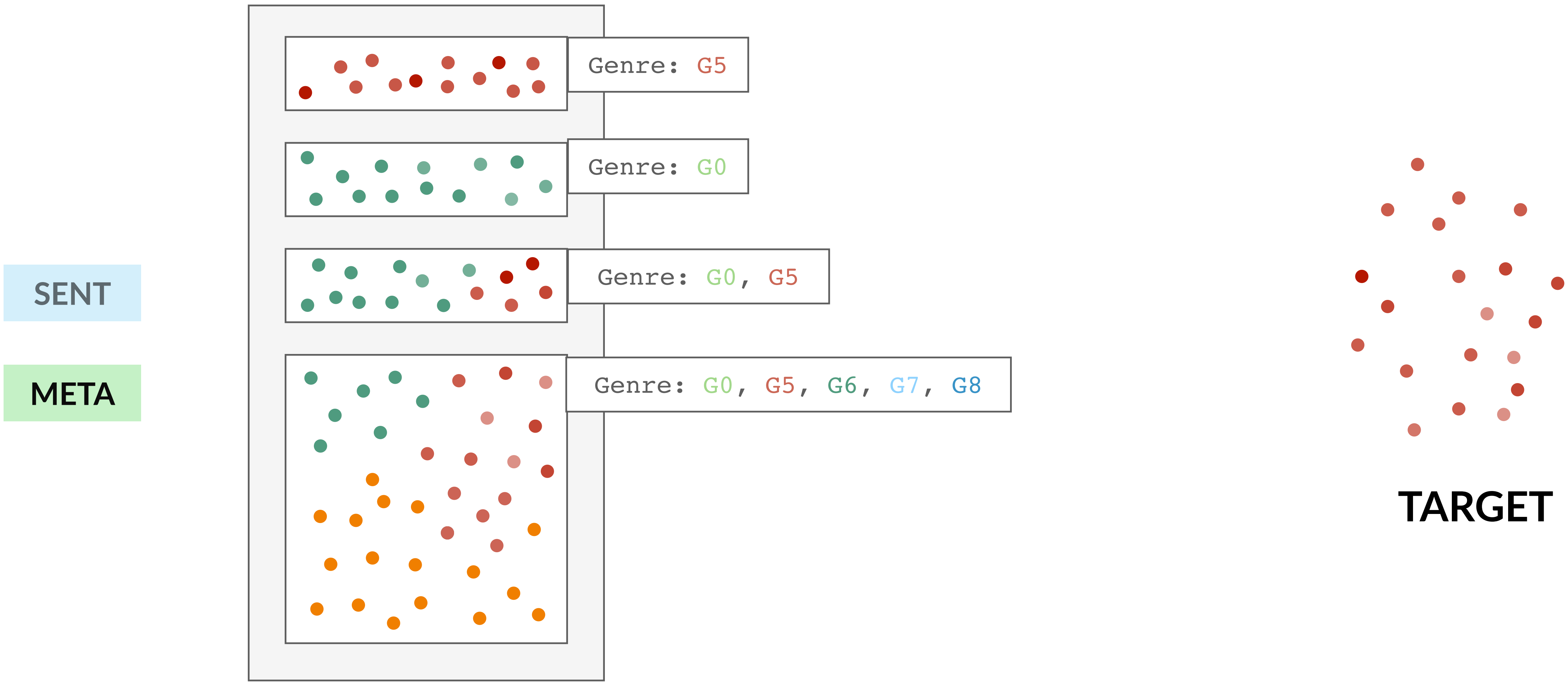


Genre: G0, G1, G2, G3, G4, G5
Genre: G0, G1, G2, G4, G5, G6
Genre: G0, G5, G6, G7, G8
Genre: G3

Treebanks

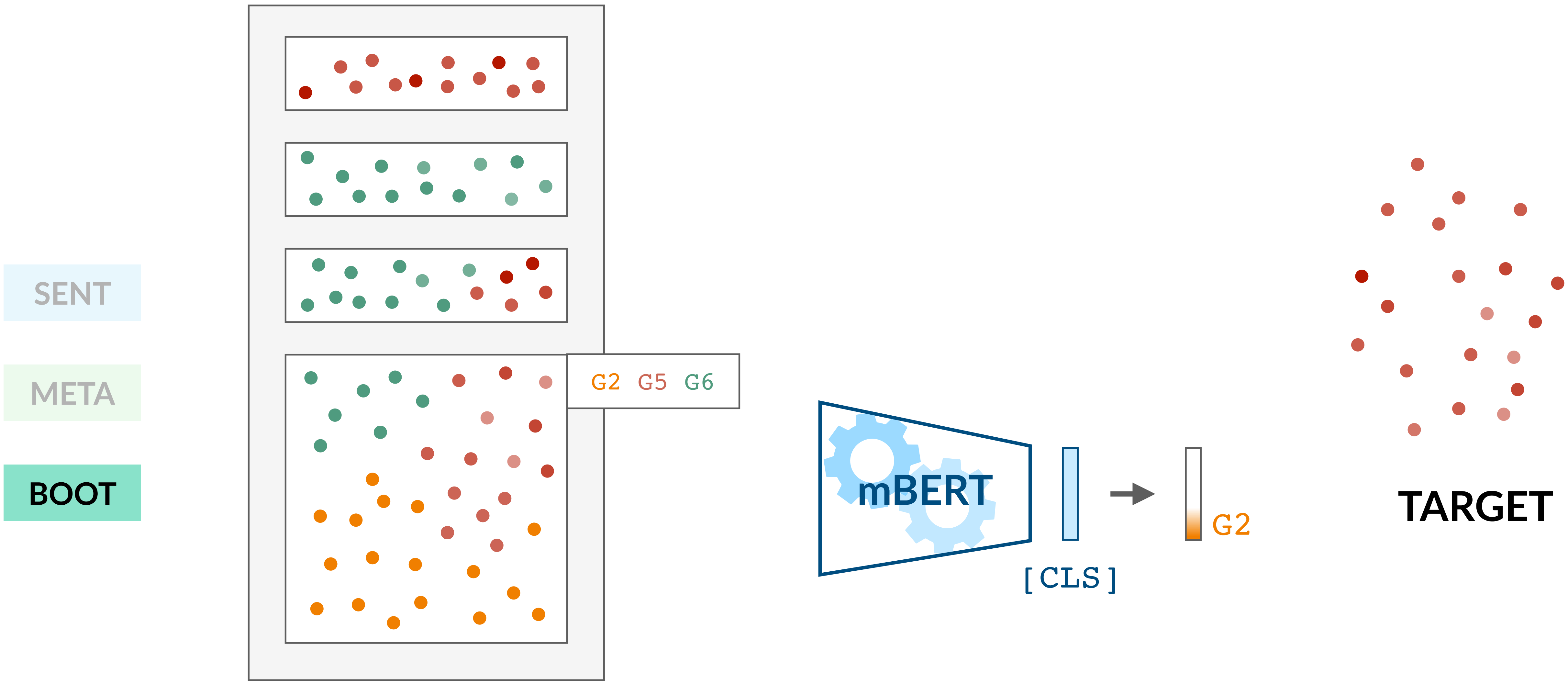


SENT: Closest cosine distance (Aharoni & Goldberg, 2020)

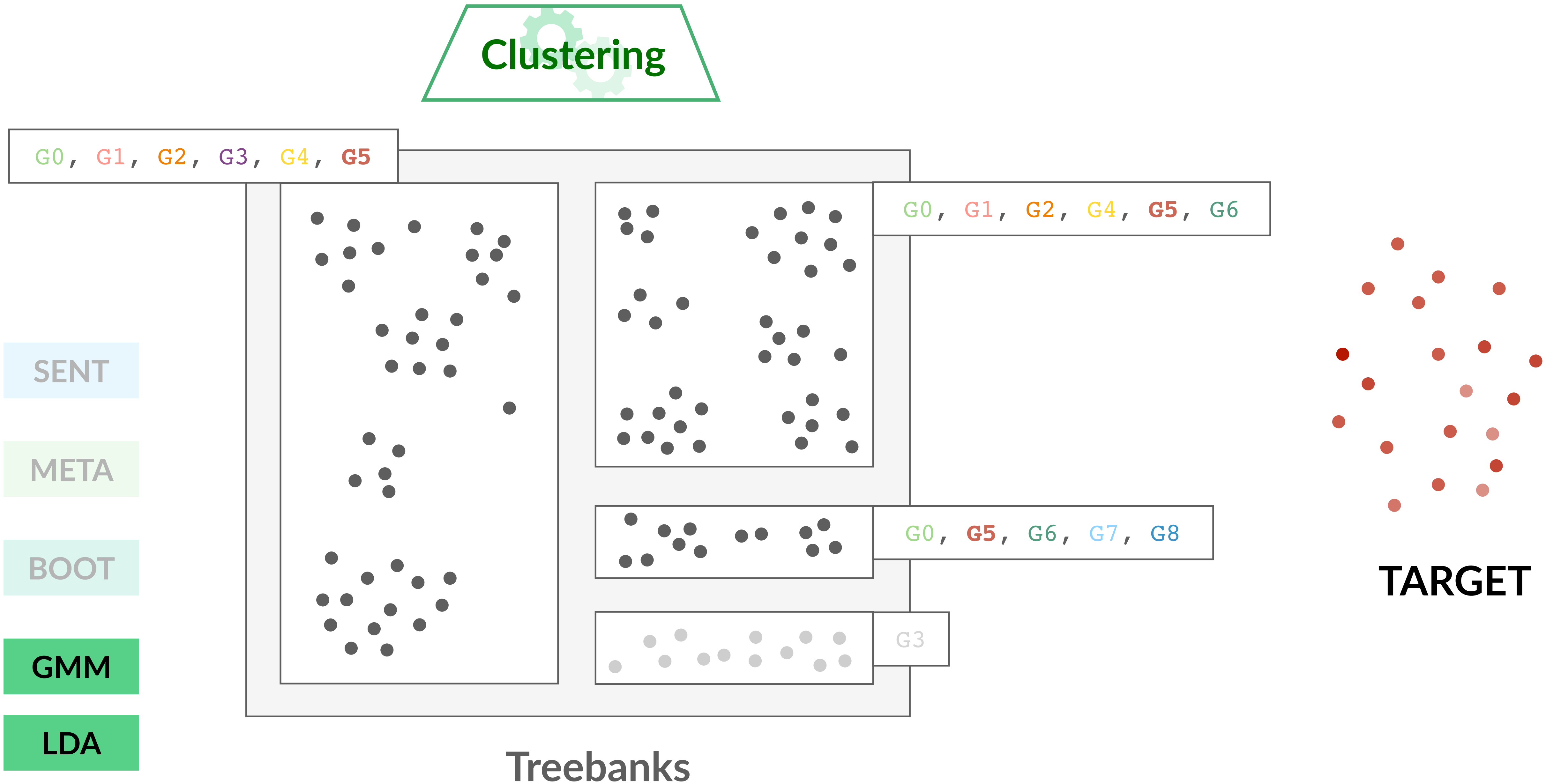


SENT: Closest cosine distance (Aharoni & Goldberg, 2020)

META: practitioner’s choice based on meta-data



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



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

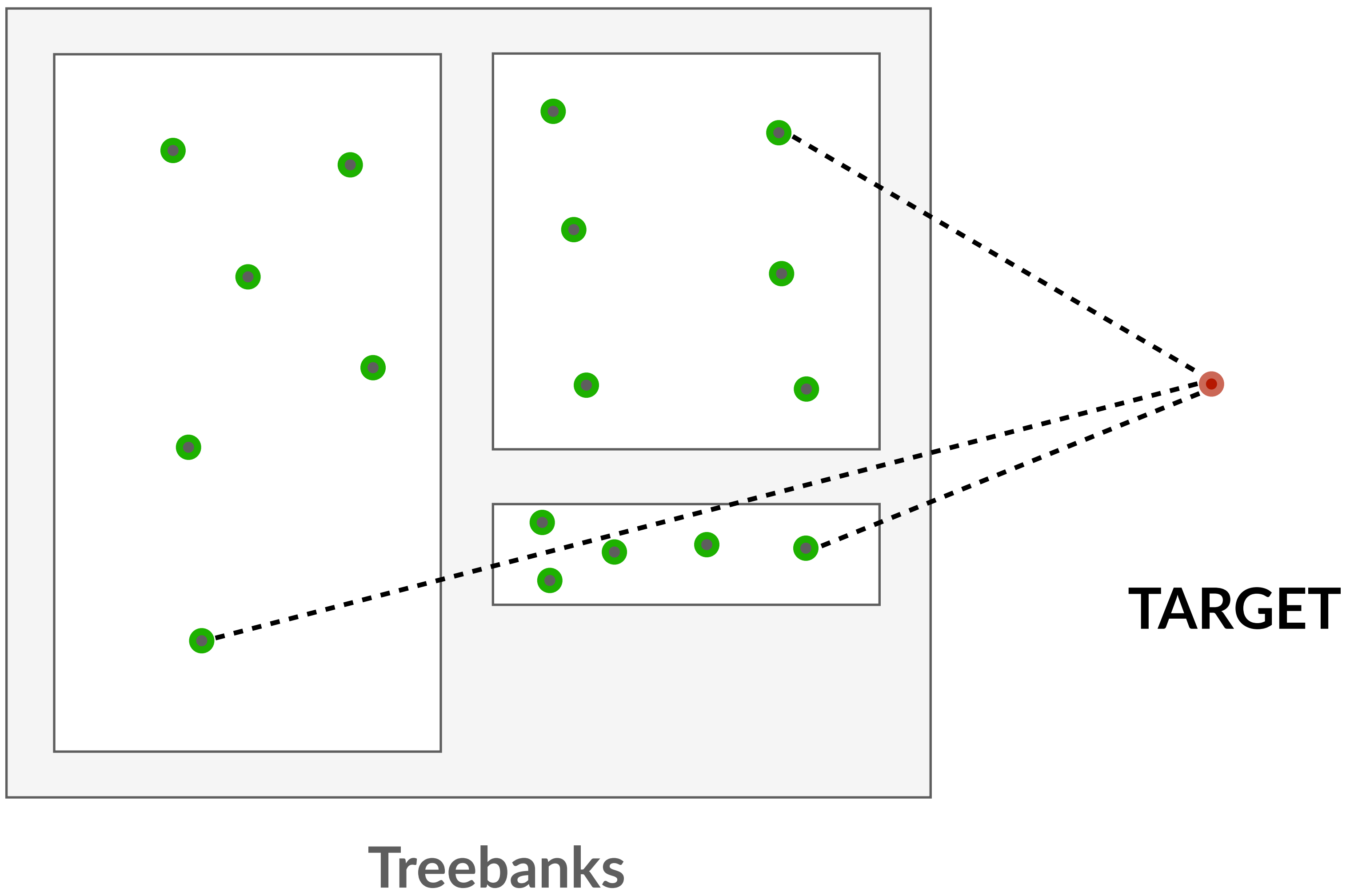
SENT

META

BOOT

GMM

LDA



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 & 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

TARGET

RAND

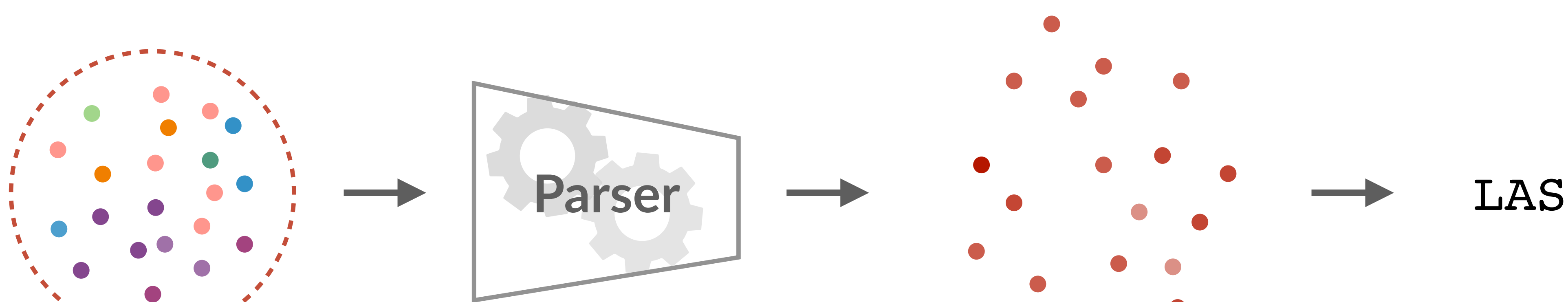
SENT

META

BOOT

GMM

LDA

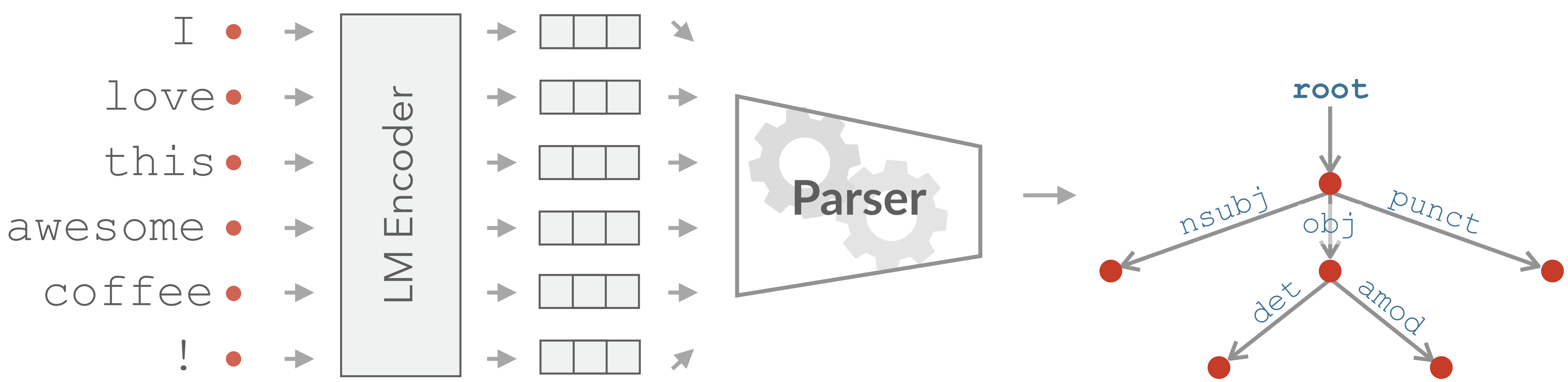


Dozat & Manning (2017)
van der Goot et al. (2021)

PROXY
(annotated)

TARGET
(unannotated)

Dependency Parsing Setup



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

34

	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	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	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**

- RQ2: Is genre **inherently** captured in multilingual LMs?

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

TARGET

RAND

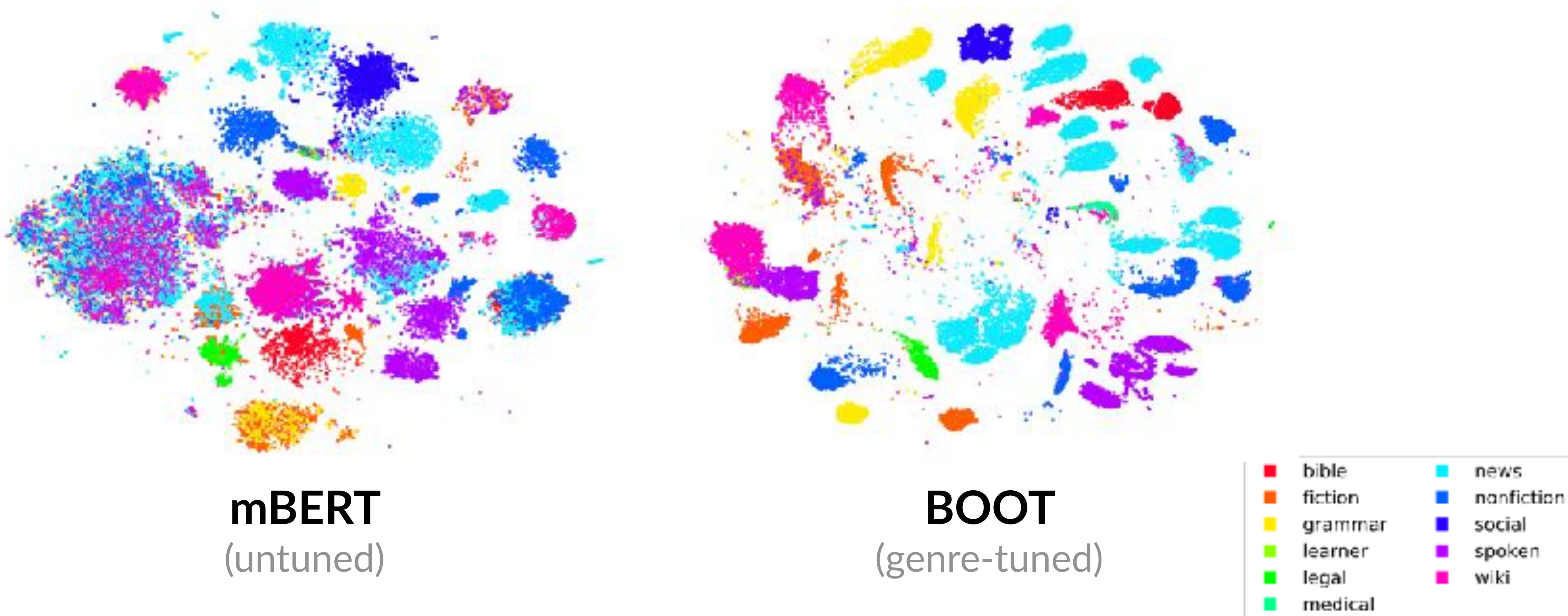
SENT

META

BOOT

GMM

LDA



Take-Aways

BOOT

GMM

LDA

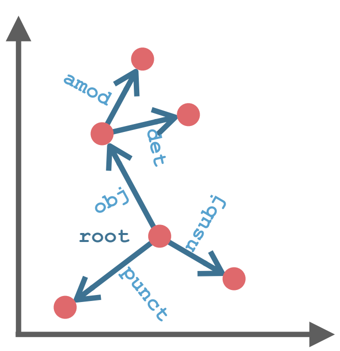
RQ1: Genre is a valuable signal for parsing unseen, low-resource targets

RQ2: Genre is inherently captured in multilingual LMs and amplifying it helps to improve parsing performance

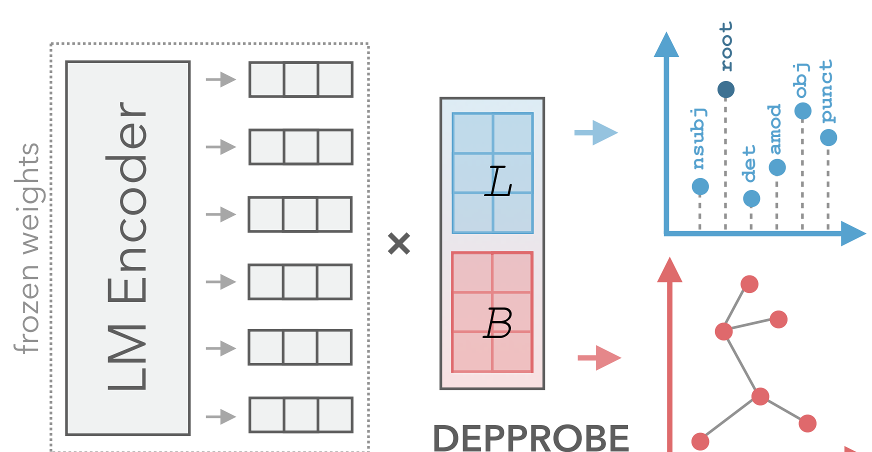
Related Follow-Up Work (1/2)

Genre: G3

What is in UD? How well can we predict genre?
An in-depth analysis of genre in UD and an instance-level genre prediction evaluation
(Müller-Eberstein et al., 2021 SyntaxFest)

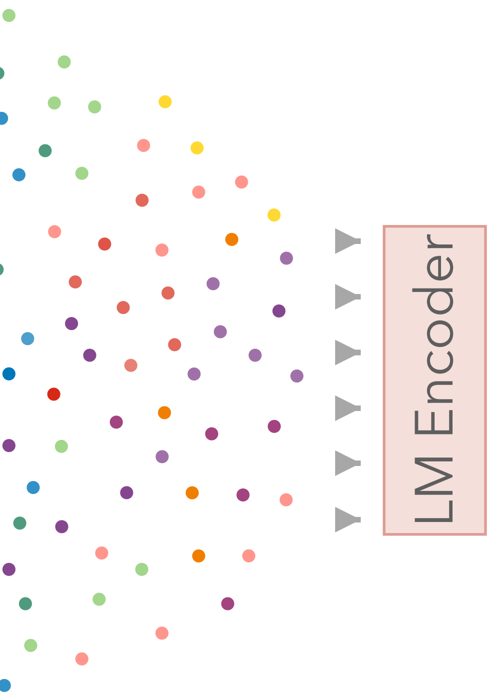


Can we efficiently probe for fully labeled trees?
DepProbe: A light-weight probe to extract labelled dependency trees from frozen LM embeddings
(Müller-Eberstein et al., 2022 ACL)



126k vs 183M
Parameters

Related Follow-Up Work (2/2)



Which language model encoder should we choose?
Language Model Ranking as Dependency Probing
(Müller-Eberstein et al., 2022 NAACL)

Sort by Structure: Language Model Ranking as Dependency Probing

Max Müller-Eberstein and Rob van der Goot and Barbara Plank
Department of Computer Science, IT University of Copenhagen, Denmark
Center for Information and Language Processing (CIS), LMU Munich, Germany
mamy@itu.dk, robv@itu.dk, bplank@cis.lmu.de

Roadmap for the Three Use Cases

1. How useful is (fortuitous) meta-data for low-res parsing?
2. How effective are non-English auxiliary tasks for transfer?
3. How can we integrate human label variation in NLP?

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




Part 2

van der Goot et al., 2021 NAACL

Task: Slot and Intent Detection (SID)

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

Intent: SearchScreeningEvent

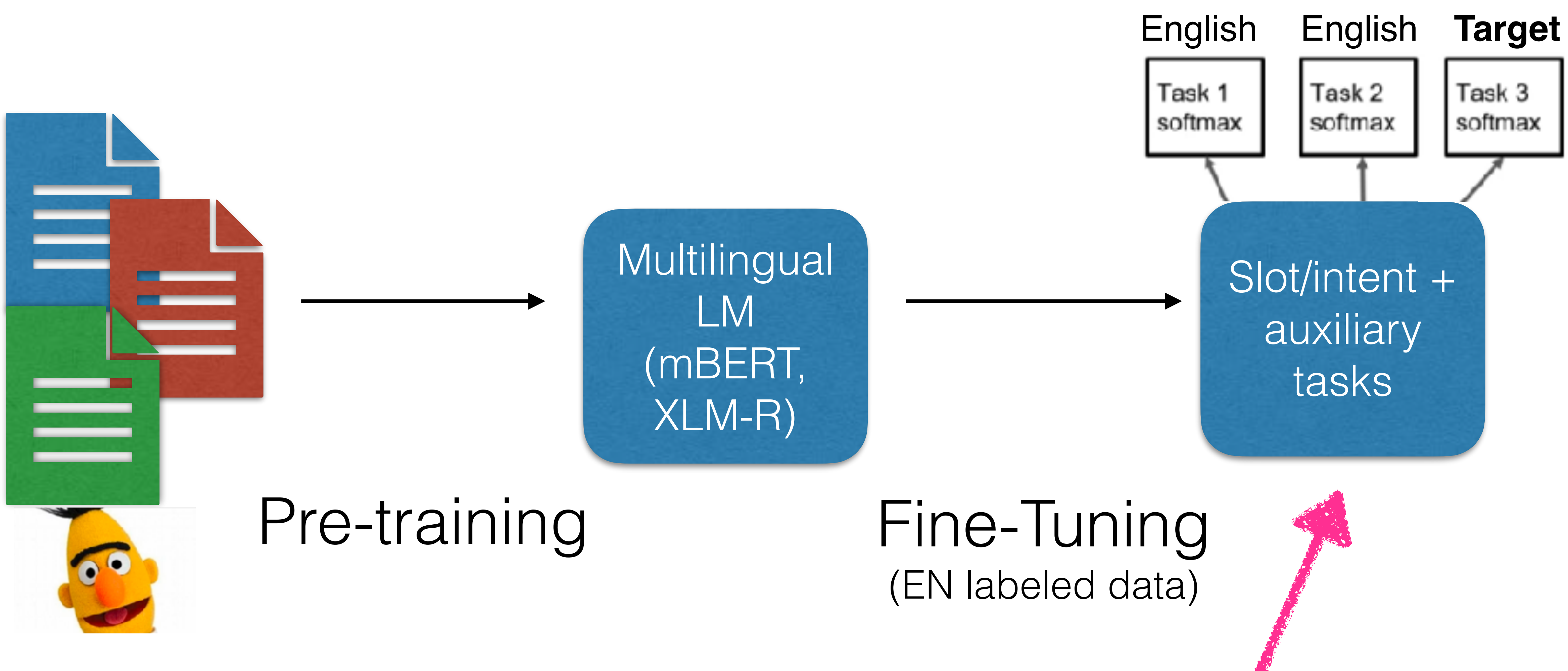
Task: Slot and Intent Detection (SID)

Slots:

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

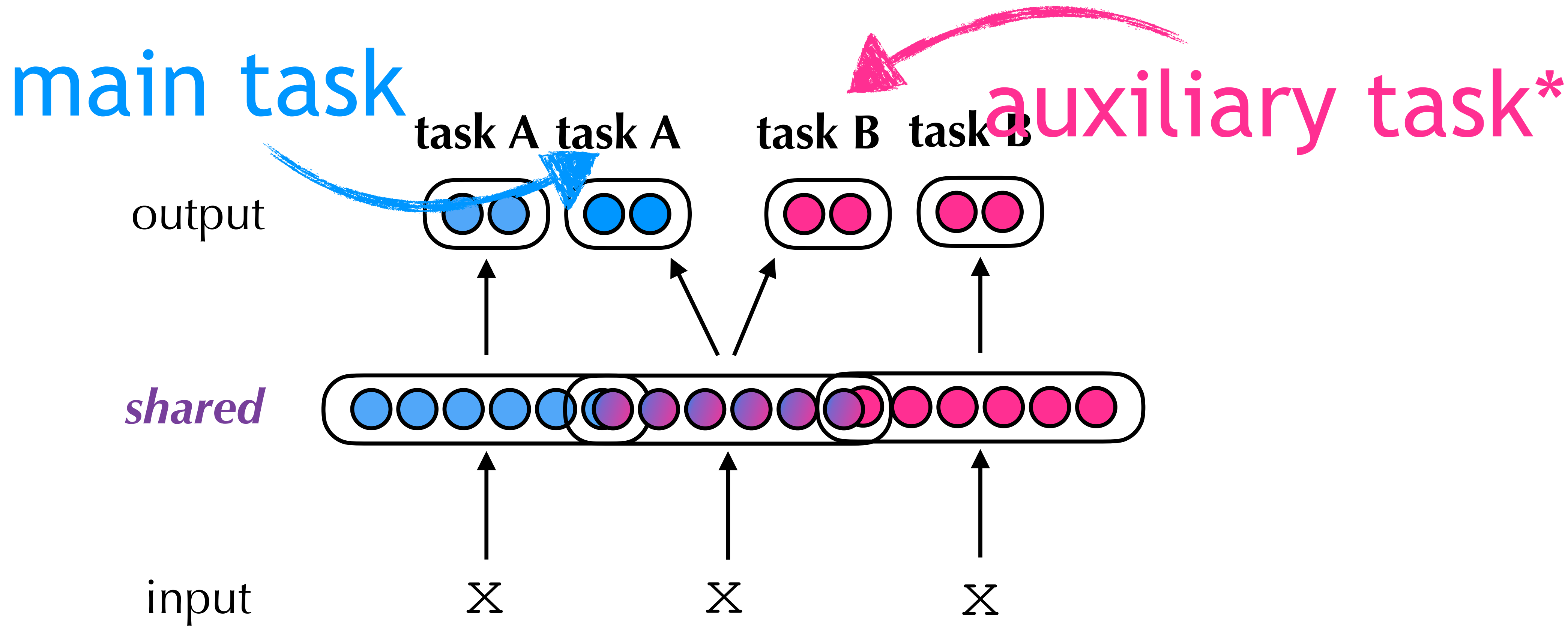
Intent: SearchScreeningEvent

Idea: Non-English Auxiliary Tasks



+ Target language auxiliary task via MTL (multi-task learning)

Multi-task Learning (MTL): Key Idea



single-task learning (STL)
multi-task learning (MTL)

"[MTL] is an approach for **inductive transfer** that improves **generalisation** by using the domain information contained in the training signal of related tasks as an inductive bias. It does this by **learning tasks in parallel** while using a **shared representation**; what is learned for each task **can help other tasks be learned better**" (Caruana, 1997)

* sometimes auxiliary task might be equally important

Why MTL? (1/2)

- **Scientific view:** jointly solving related problems to work towards more general language understanding

- **Practical view:** simpler model able to handle multiple tasks, which generalizes better and is more efficient in learning

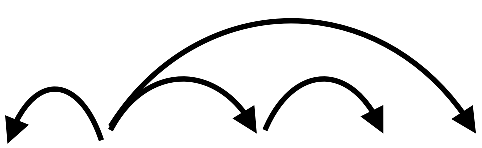
Why MTL? (2/2)

- **Attention focusing** (Caruana, 1997): reduced net capacity can improve generalisation
- **Representation bias** (Caruana, 1997) - MTL prefers solutions which other tasks prefer
- **Regularization** (Caruana, 1997): MTL acts as regularizer (Ruder, 2017), reduces the risk of overfitting, particularly on small data.
- **Reduces the need of labeled data** - generalisation via prediction of auxiliary task(s) - early work in NLP by Collobert & Weston (2008)

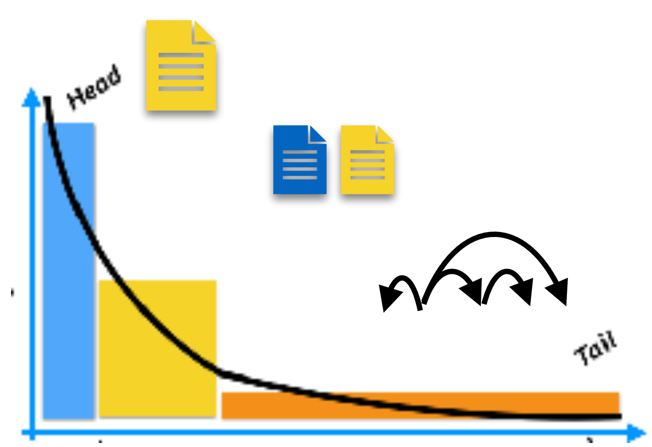
Non-English Auxiliary Tasks - Sorted by Data Availability

- **Raw data:** Masked language modelling (aux-mlm)

- **Parallel data:** Neural machine translation (aux-nmt)



- **Parsing data:** UD parsing (aux-ud)



New 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 seansları görmek istiyorum
zh	我想看 Silly Movie 2.0 在影院的放映

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

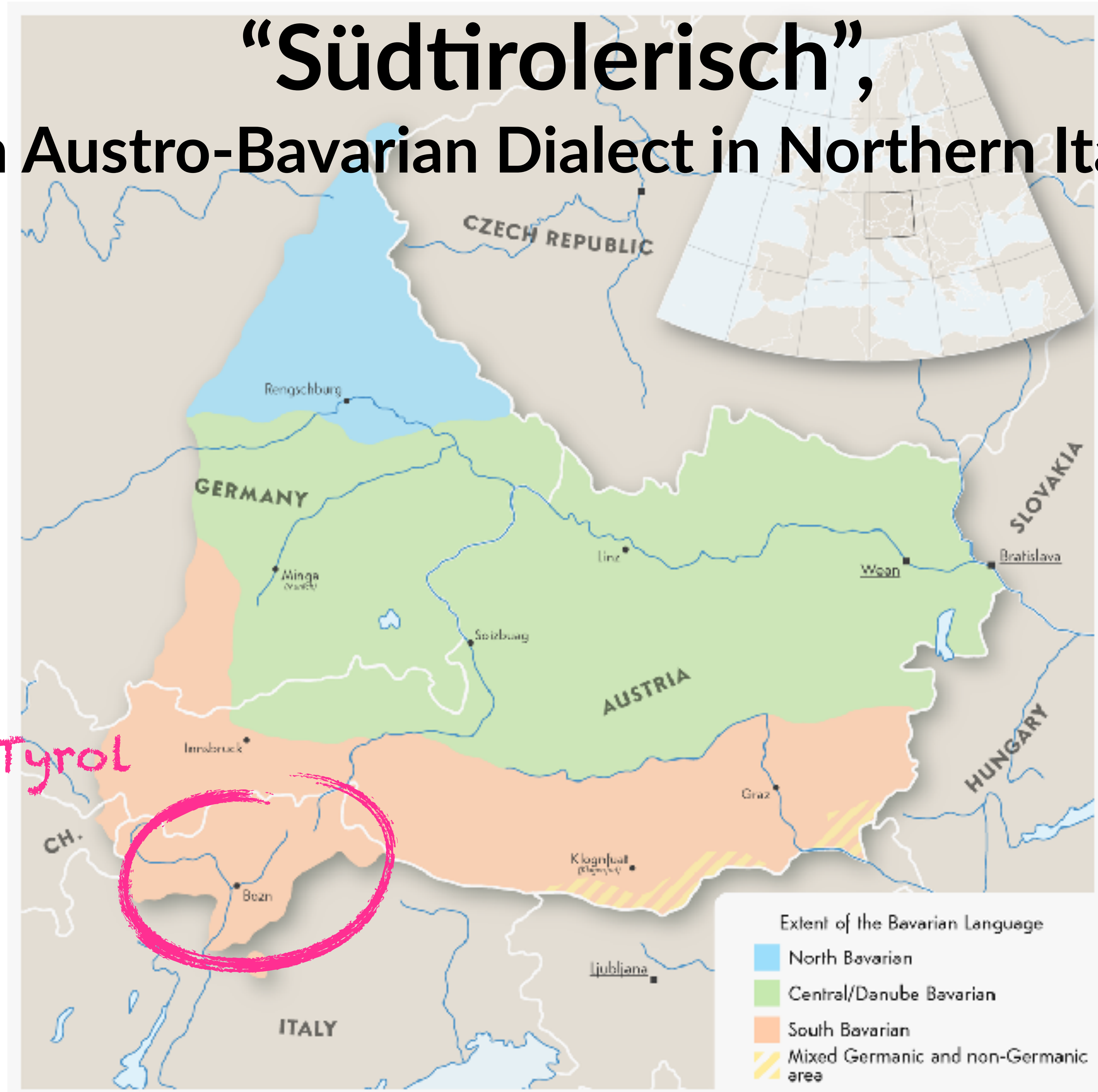
Results on Slots - Main take-away

mBERT lang2vec	**en** —	de-st —	**de** 0.18	**da** 0.18	**nl** 0.19	**it** 0.22	**sr** 0.23	**id** 0.24	**ar** 0.30	**zh** 0.33	**kk** 0.37	**tr** 0.38	**ja*** 0.41	Avg.
Slots														
base	**97.6**	48.5	33.0	73.9	80.4	75.0	**67.4**	**71.1**	45.8	**72.9**	48.5	55.7	59.9	61.0
nmt-transfer	0.0	50.9	34.5	60.8	63.7	51.0	41.3	54.2	**48.2**	27.9	0.2	52.0	45.0	44.1
aux-mlm	97.3	**53.0**	**34.6**	**75.9**	**82.2**	**78.0**	63.8	69.5	48.1	69.4	**51.3**	**58.4**	**63.5**	**62.3**
aux-nmt	0.0	44.5	33.3	71.4	76.9	71.9	58.5	62.9	38.7	70.3	38.2	50.2	58.7	56.3
aux-ud	97.5	47.6	29.1	73.7	73.3	61.8	56.8	61.1	42.6	64.9	45.2	53.8	47.6	54.8

(More results in the paper)

A closer look at a 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

South Tyrolean

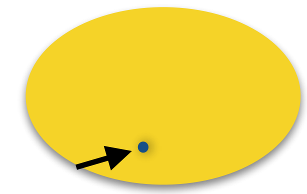
- German dialect (“Südtirolerisch”) spoken by a minority
  - Spoken in the northernmost Italian province of Bozen-Bolzano with ~0.5M inhabitants (~2/3 German dialect speakers)
  - No common orthographic standard
  - Lexical influence of other official languages (Italian, Ladin)
    - “Hosch is **patent** schun gemocht?”
      [patent (neut.)=
      ital. la patente (fem.),
      dt. der Führerschein (masc.),
      eng. driver’s license]

Example

text-en: Is it going to rain **today**?
text: Regnts **heinte**?

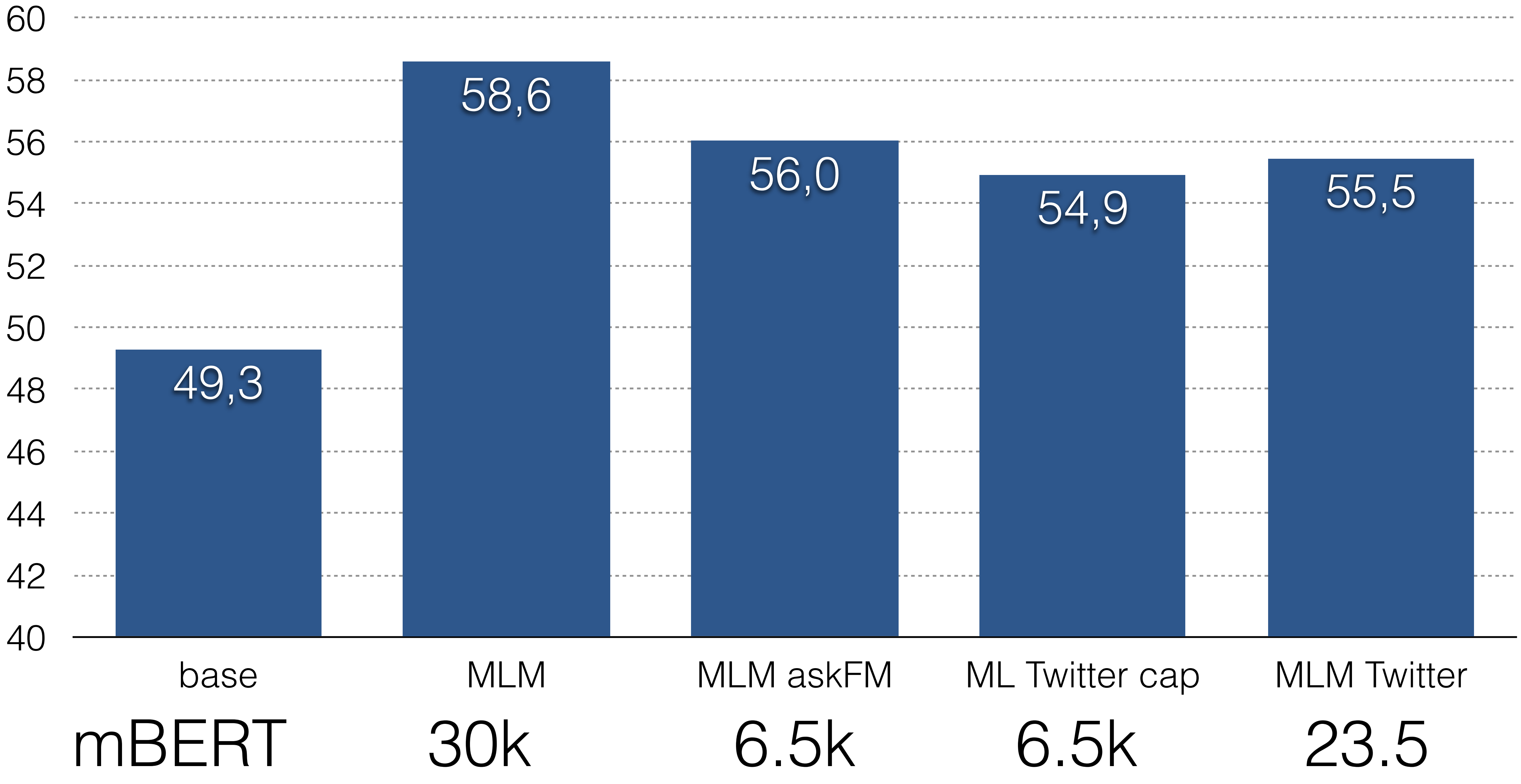
text-en: Will it be sunny **today**?
text: Wearts **heint** sunnig?

X Sparsity

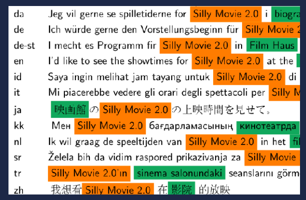


- Very difficult to get access to unlabeled data
    - Social media (Twitter): highly mixed data, switch to “high” languages, no “dialect” identifier exists
    - AskFM: short Q&A posts, more dialectal

De-ST: #sentences for MLM

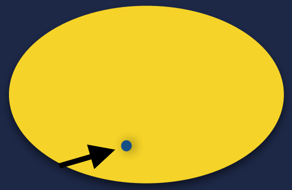


Take-aways



1. xSID is a new multilingual evaluation dataset for intent and slot detection
   —> see Razumovskaia et al. 2022 JAIR survey for more emerging multilingual SID datasets

2. We found aux-MLM the most robust auxiliary task



3. First results on DE-ST, a very-low resource German dialect (X sparsity)

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

★ Video: https://www.youtube.com/watch?v=DH0C-n_p6h0

A short detour: Is MTL new? No.

Successful Multi-task learning

in early ML

One of the early self-driving cars

Figure 4: NAVLAB, the CMU autonomous navigation test vehicle.

CMU Alvinn MTL (Caruana 1998)

First autonomous car: Ernst Dickmann's VaMoRs Mercedes (1986)
Src: https://www.youtube.com/watch?v=I39sxwYKlEE

Data-derived auxiliary tasks

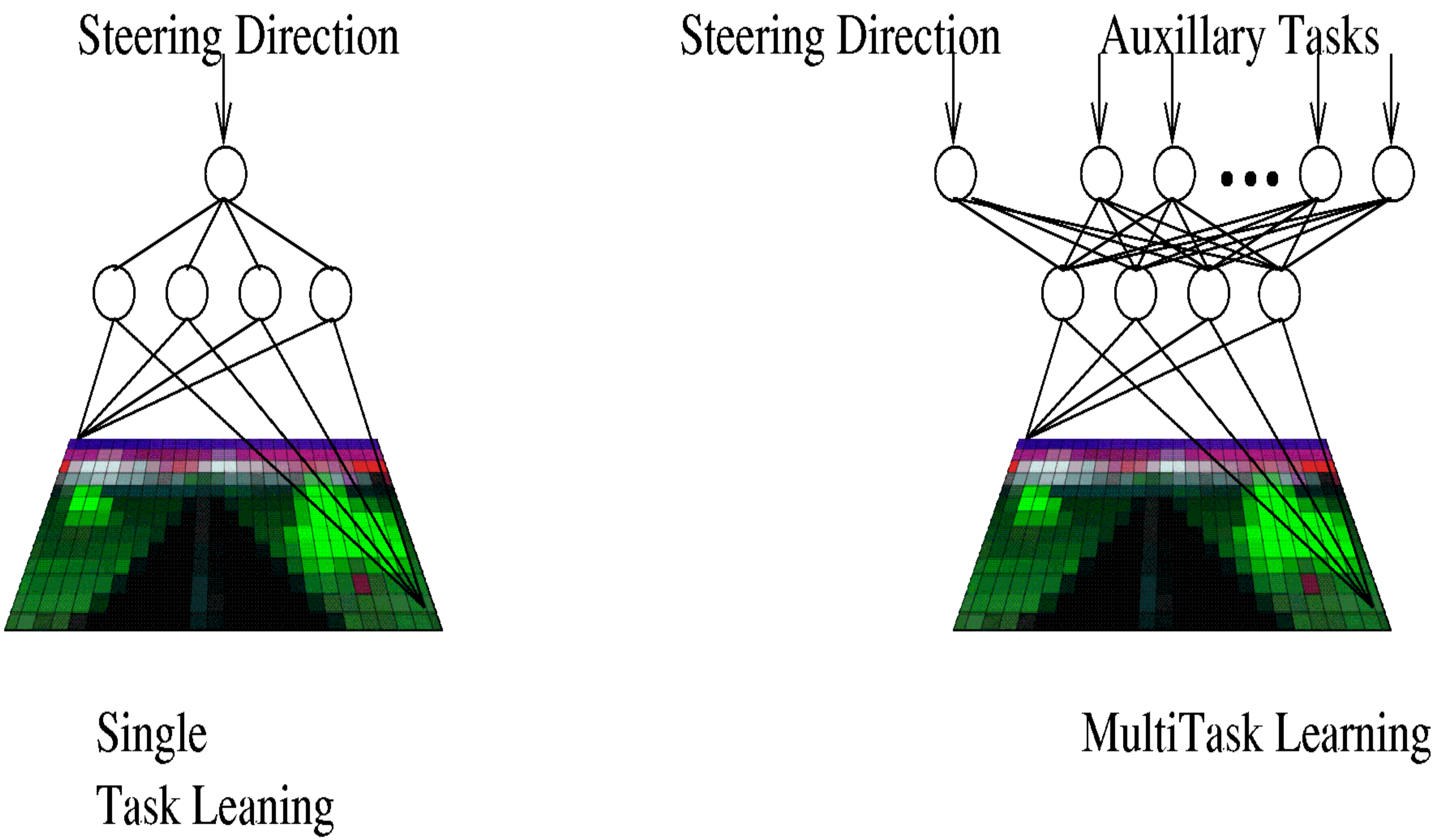
For our MTL experiments, eight additional tasks were used:

- whether the road is one or two lanes
- location of left edge of road
- location of road center
- intensity of region bordering road
- location of centerline (2-lane roads only)
- location of right edge of road
- intensity of road surface
- intensity of centerline (2-lane roads only)

CMU Alvinn MTL (Caruana 1998)

Note: here all task labels computable from data

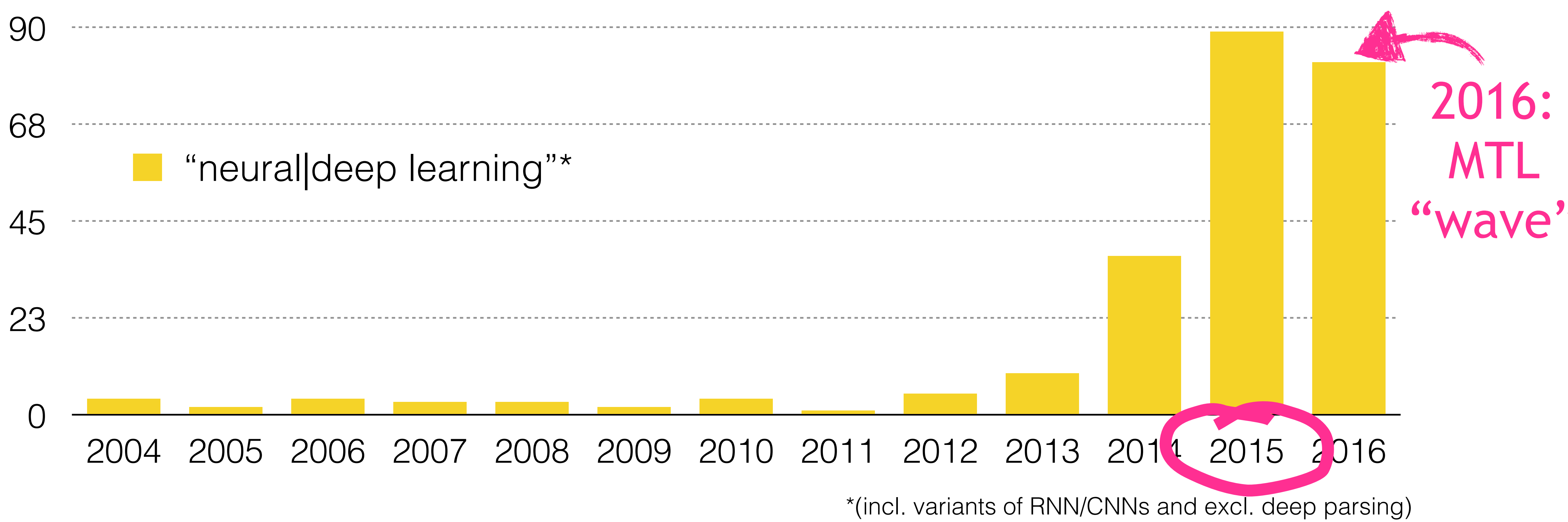
Alvinn MTL



Focus of Attention

Deep learning & MTL in NLP

"2015 seems like the year when the full force of the tsunami hit the major NLP conferences"
—Chris Manning (2015)



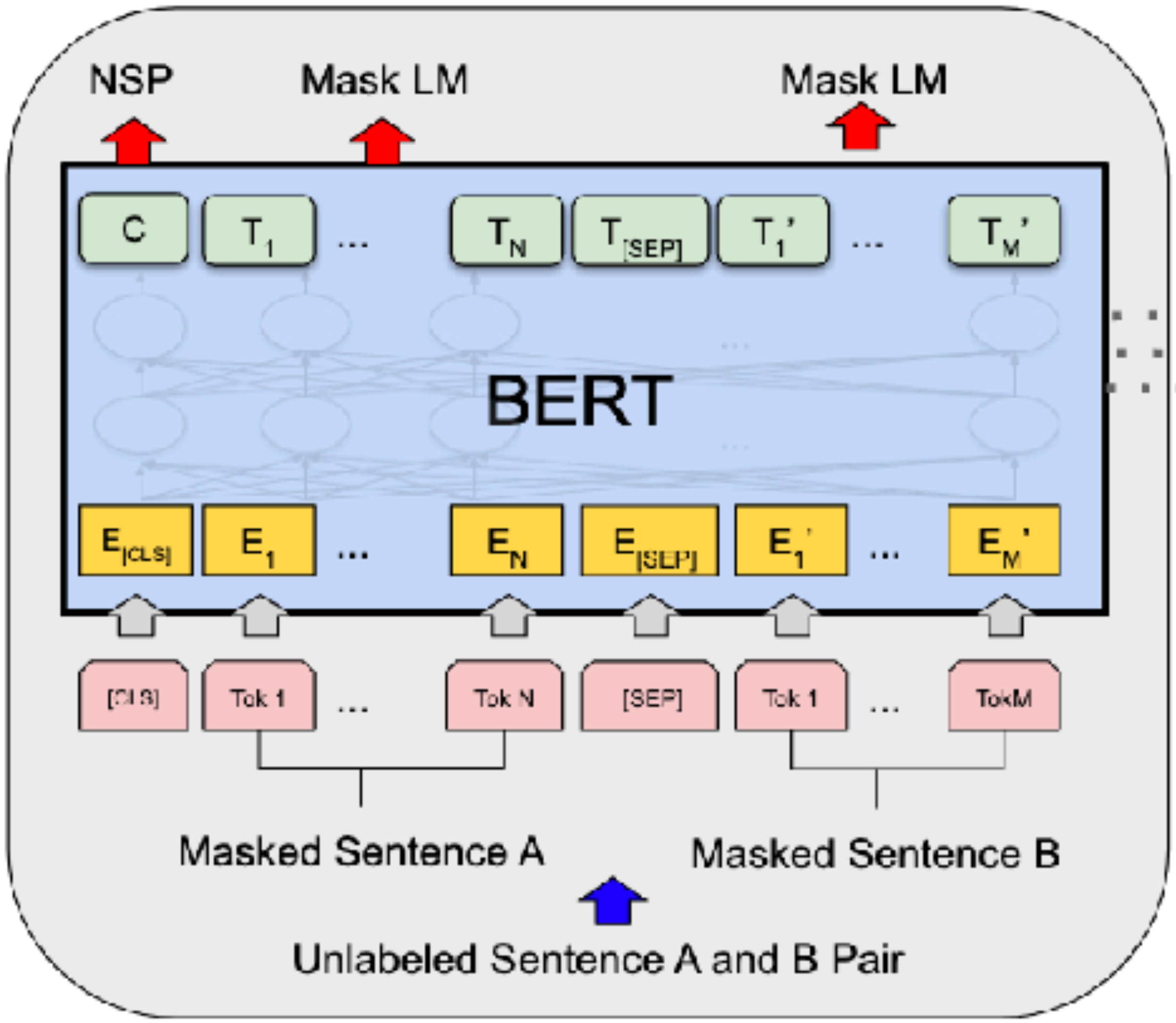
Titles of papers in ACL anthology (from 2004)

DL "tsunami" (Manning, 2015)

MTL "wave" (Ruder & Plank, 2018)

MTL is nowadays ubiquitous in NLP

Multi-task Pre-Training



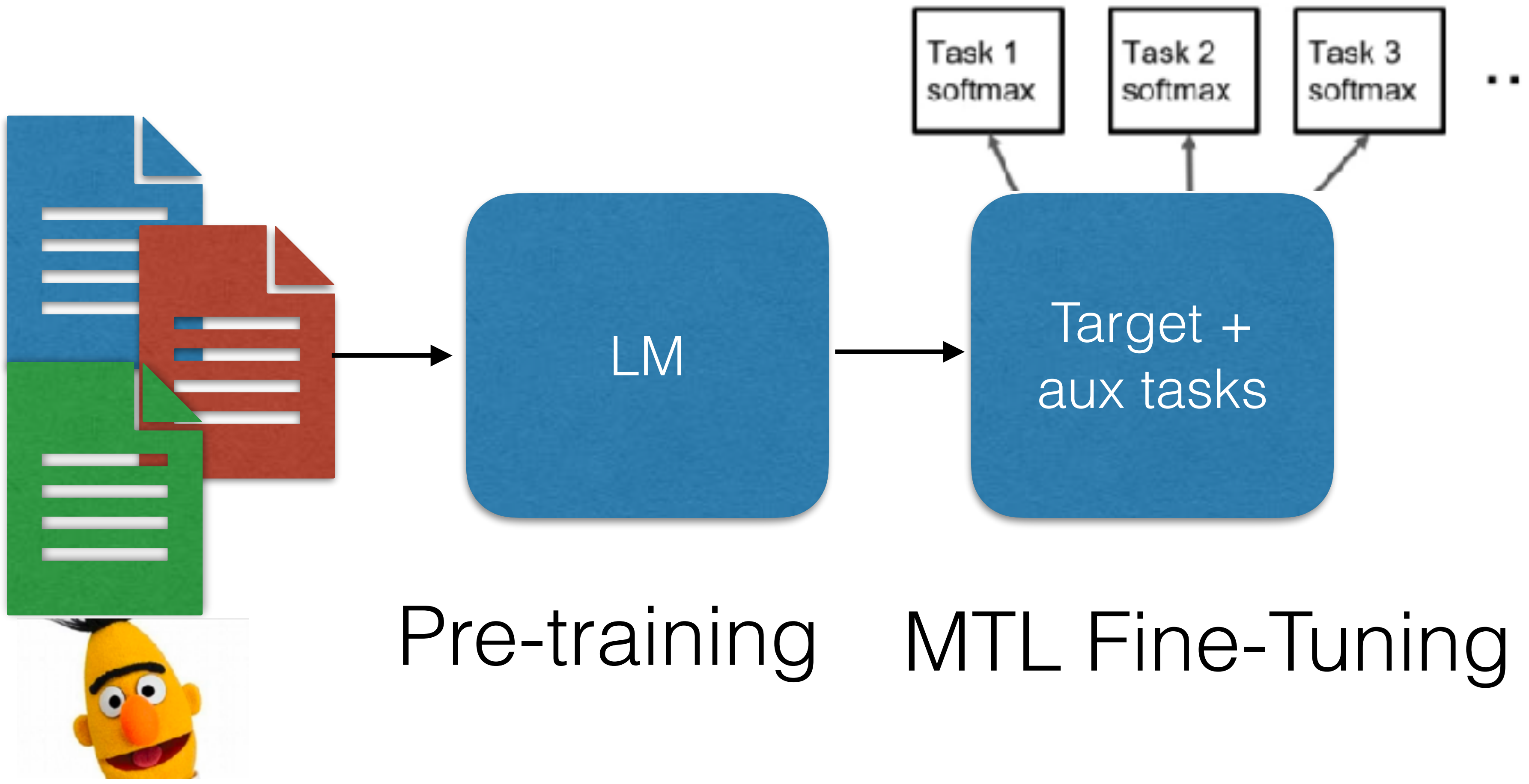
self-supervised objectives:

MLM + NSP

e.g. Devlin et al., (2019), Raffel et al. (2020)

Slide adapted from Clark et al, 2019

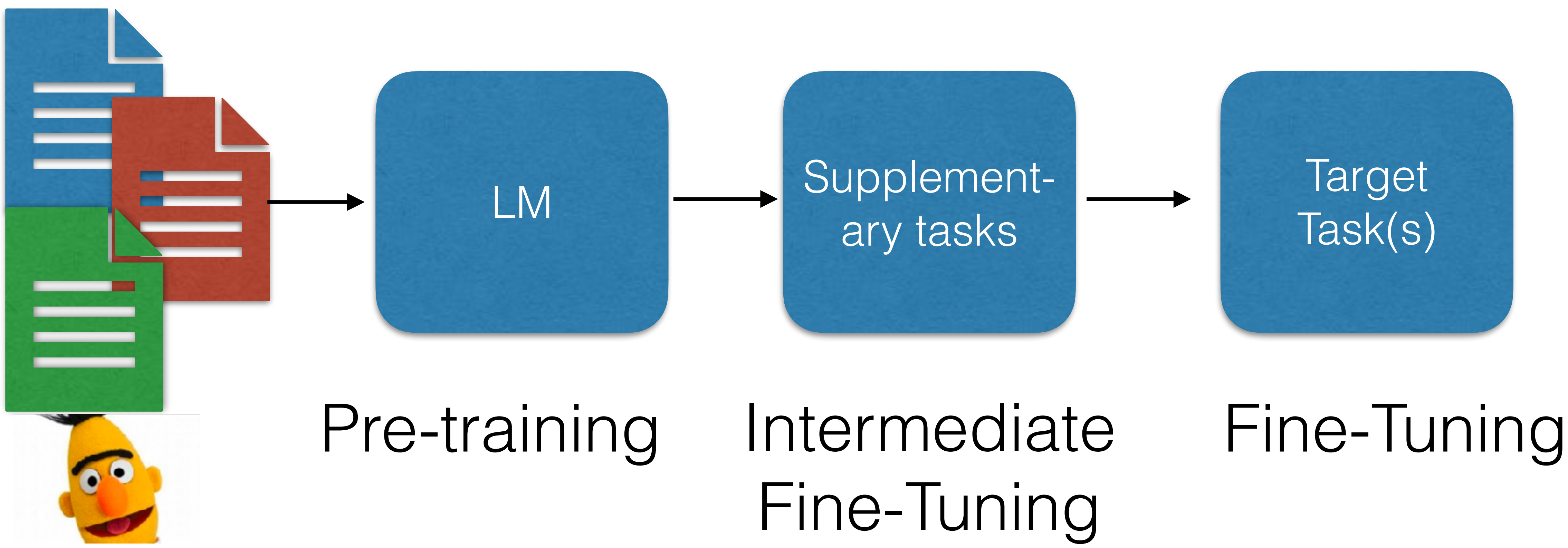
Multi-task Fine-Tuning



e.g. MT-DNN by Liu et al., (2019), van der Goot et al., (2021)

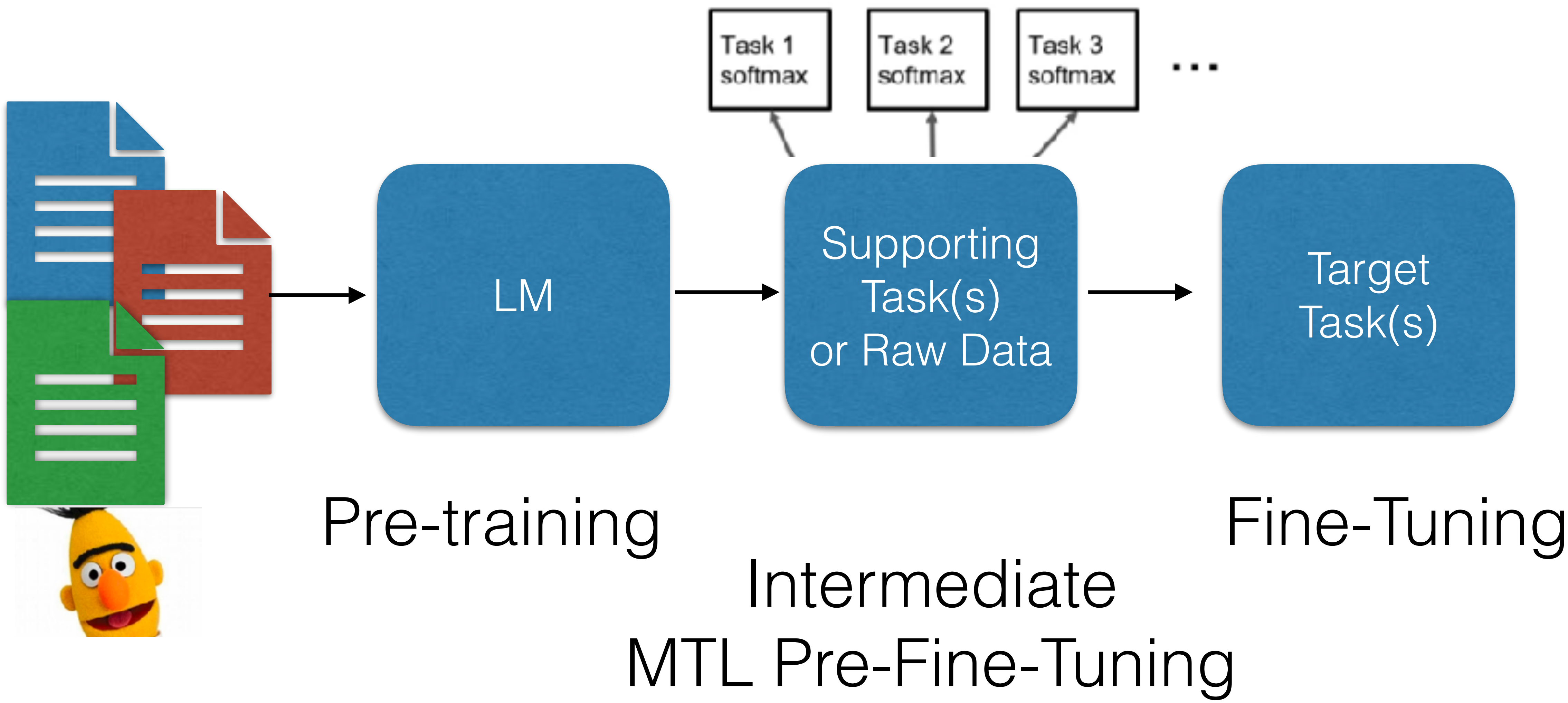
Slide adapted from Clark et al, 2019

Supplementary Training on Intermediate Tasks (STILTs)



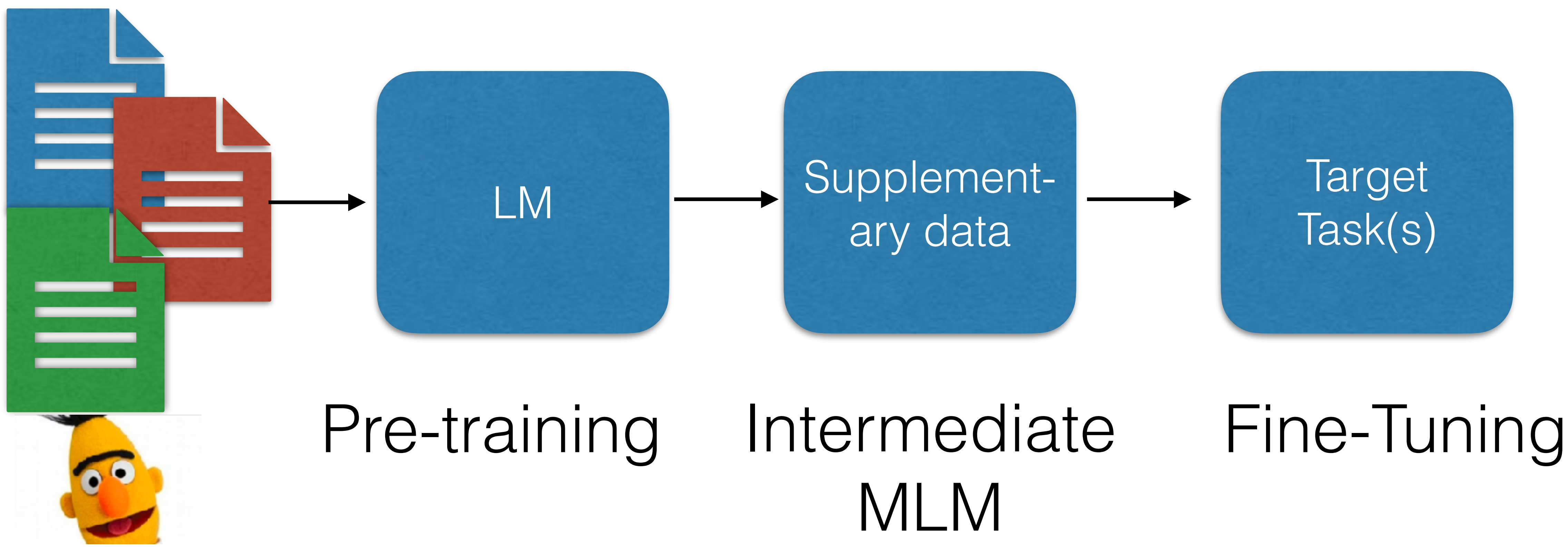
e.g. Phang et al., 2019 (STILTs) - labeled data

Multi-task Pre-Finetuning



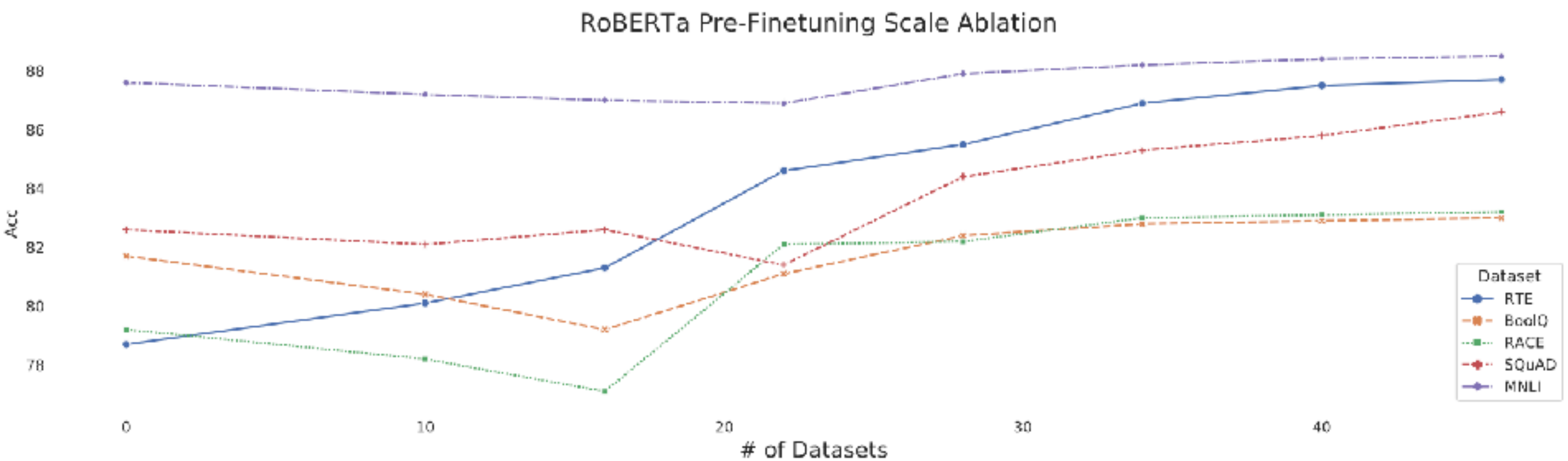
Aghajanyan et al. 2021 (MUPPET); Weller et al.,. 2022 (intermediate multi-task learning)

Domain Adaptive Pre-Training



e.g. Gururangan et al., 2020 (DAPT, TAPT) - sequential MLM pre-training

Multi-task Pre-Finetuning: Importance of Scale (See also Slav & Sasha's talks)



MUPPET paper (Aghajanyan et al., 2021 EMNLP)
[they use task-specific heads, loss scaling, and large-scale MTL with 15+ tasks]

… to Extreme Text-to-Text Tasks
Multi-task Pre-Training

(Aribandi et al., 2022 ICLR)

[they recast tasks to text-to-text training, i.e. MTL as seq2seq w/o specific heads]

Published as a conference paper at ICLR 2022

EXT5: TOWARDS EXTREME MULTI-TASK SCALING FOR TRANSFER LEARNING

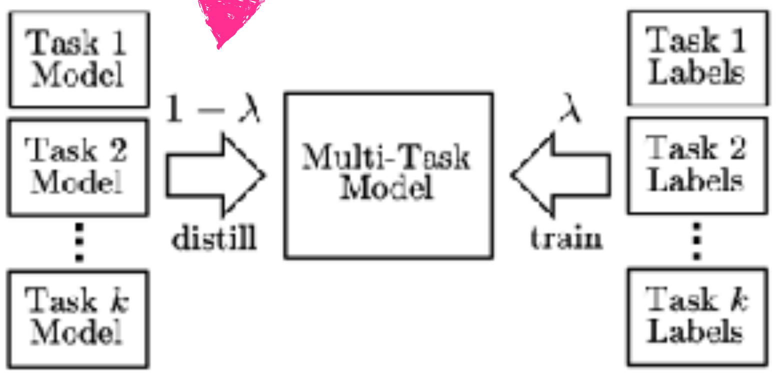
Vamsi Aribandi[*†], **Yi Tay**[†], **Tal Schuster, Jinf**
Sanket Vaibhav Mehta, Honglei Zhuang, Vin
Jai Gupta, Kai Hui, Sebastian Ruder[♣], **Dona**
Google Research, ♣DeepMind
{aribandi, yitay}@google.com

Despite the recent success of multi-task learning and transfer learning for natural language processing (NLP), few works have systematically studied the effect of scaling up the number of tasks during pre-training. Towards this goal, this paper introduces EXMIX (**Ex**treme **Mix**ture): a massive collection of 107 supervised NLP tasks across diverse domains and task-families. Using EXMIX, we study the effect of multi-task pre-training at the largest scale to date, and analyze co-training transfer amongst common families of tasks. Through this analysis, we show that manually curating an ideal set of tasks for multi-task pre-training is not straightforward, and that multi-task scaling can vastly improve models on its own. Finally, we propose EXT5: a model pre-trained using a multi-task objective of self-supervised span denoising and supervised EXMIX. Via extensive experiments, we show that EXT5 outperforms strong T5 baselines on SuperGLUE, GEM, Rainbow, Closed-Book QA tasks, and several tasks outside of EXMIX. EXT5 also significantly improves sample efficiency while pre-training.

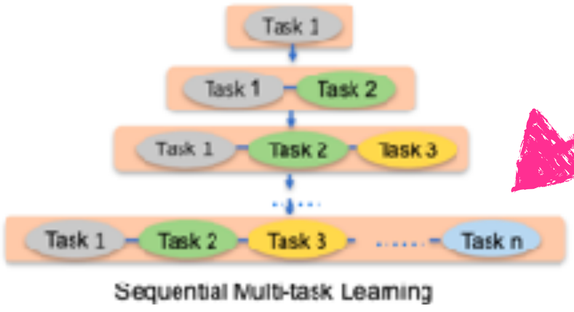
Selected advances in MTL related to Efficiency & Timing

- MTL & *knowledge distillation* (Clark et al., 2019)

Distillation with teacher annealing



- MTL & *continual learning* (Sanh et al. 2019; Sun et al., 2020)



Progressively adding tasks

- MTL & *adapters* via shared hypernetworks (Mahabadi et al., 2021 arXiv)

Intermediate Take-Aways

Large-scale Multi-Task Pre-Fine-Tuning (e.g. Muppet) > pairwise MTL fine-tuning > MTL_all fine-tuning

Intermediate Task Training (STILT) vs MTL Pre-Fine-Tuning:

- STILT better if aux data is small
- If aux data is large MTL Pre-Finetuning better

Task/Data relationships and MTL/TL success still an on-going research question

To wrap up this MTL detour

MTL Benefits: Flexibility & Reuse

- Flexible, easy-to-use method
- Shown to work particularly well in low-data scenarios
- Allows the re-use of very different kinds of data (incl. distinct data sources)

MTL Issues: Catastrophic Forgetting & Interference

- **Sharing parameters** across tasks might lead to a **deterioration** of performance

  - Not all tasks might be equally useful

- Training data from one task might **swamp** learning

- **Possible solutions**: data sampling (e.g. Sanh et al., 2019, van der Goot et al., 2021), loss weighting (e.g. Aghajanyan et al., 2022; Lin et al., 2021), heterogeneous batches (e.g. Aghajanyan et al., 2022), moving to adapters to avoid interference (e.g. Houlsby et al., 2019; Pfeiffer et al., 2020)

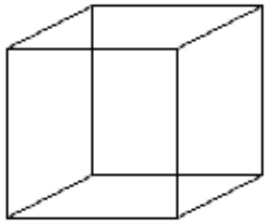
Roadmap

1. How useful is (fortuitous) meta-data for low-res parsing?
2. How effective are non-English auxiliary tasks for transfer?
3. How can we integrate human label variation in NLP?

Disagreement in human annotation is ubiquitous

Side benefit of annotation - fortuitous data:

Disagreement as a source of information?



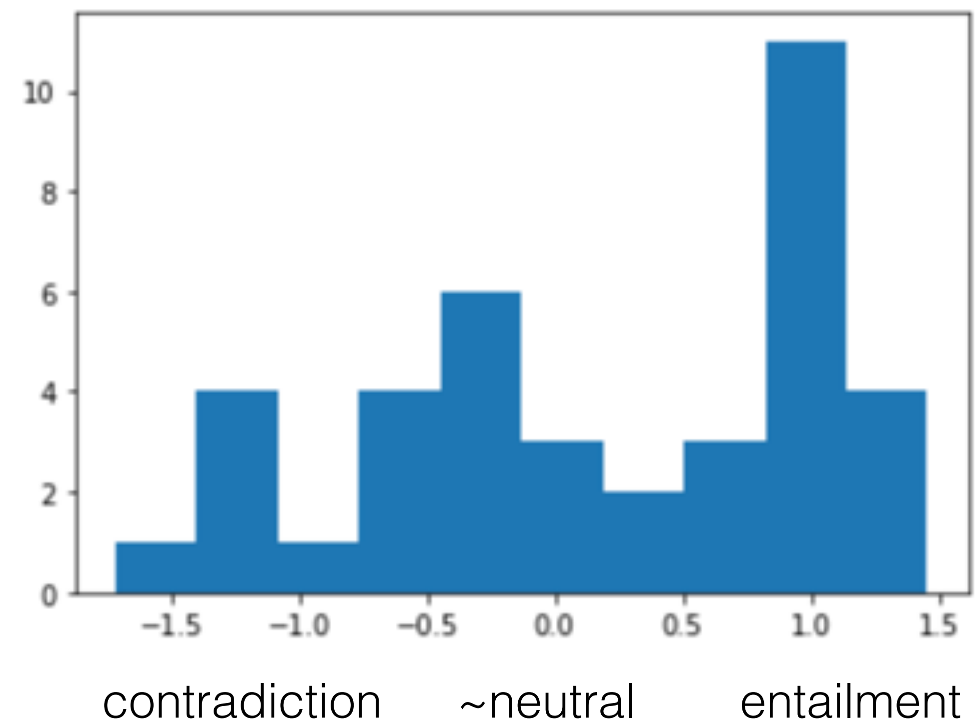
Note on **naming**: I'm calling it human disagreement, but I will return to this name in the end
(Assumption is: not plain noise, but implicit/genuine disagree.)

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?*

VERB	NOUN	**NOUN**	ADP	NOUN
VERB	NOUN	**VERB**	ADP	NOUN
luv	paper	presenting	at	#lxlms

Recognising Textual Entailment (RTE)



Premise *p*: Amanda carried the package from home .
Hypothesis *h*: Amanda moved .

Does p->h?
RTE original-dataset-label: entailed

Data with 50 annotators by Pavlick & Kwiatkowski (2019)
Newer ChaosNLI with 100 a. by Nie, Zhou, Bansal (2020)

More examples (selected)

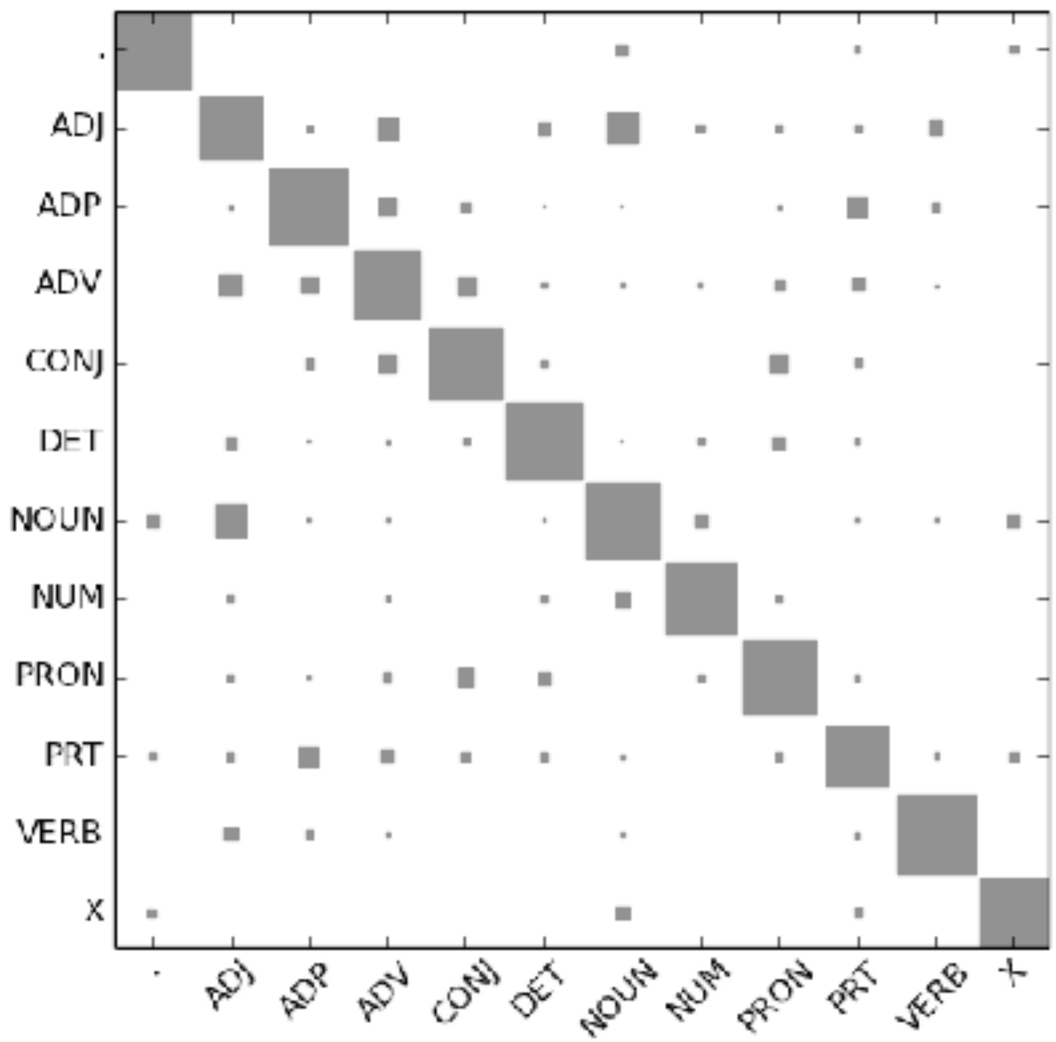
- Relation Extraction (Aroyo & Welty, 2013)
- Abusive & offensive language (Akhtar et al, 2021; Leonardelli et al., 2021; Ceras Curry et al., 2021)
- Dependency Parsing (Martinez Alonso et al., 2015; Liu et al., 2018)
- Visual Question Answering (Jolly et al., 2021)

Figure 1: One image from VQA2.0 with two questions and the answers by 10 annotators. Frequency of each unique answer (e.g., *plaid* : 4) and EASE values of the samples (the higher, the easier) are reported.

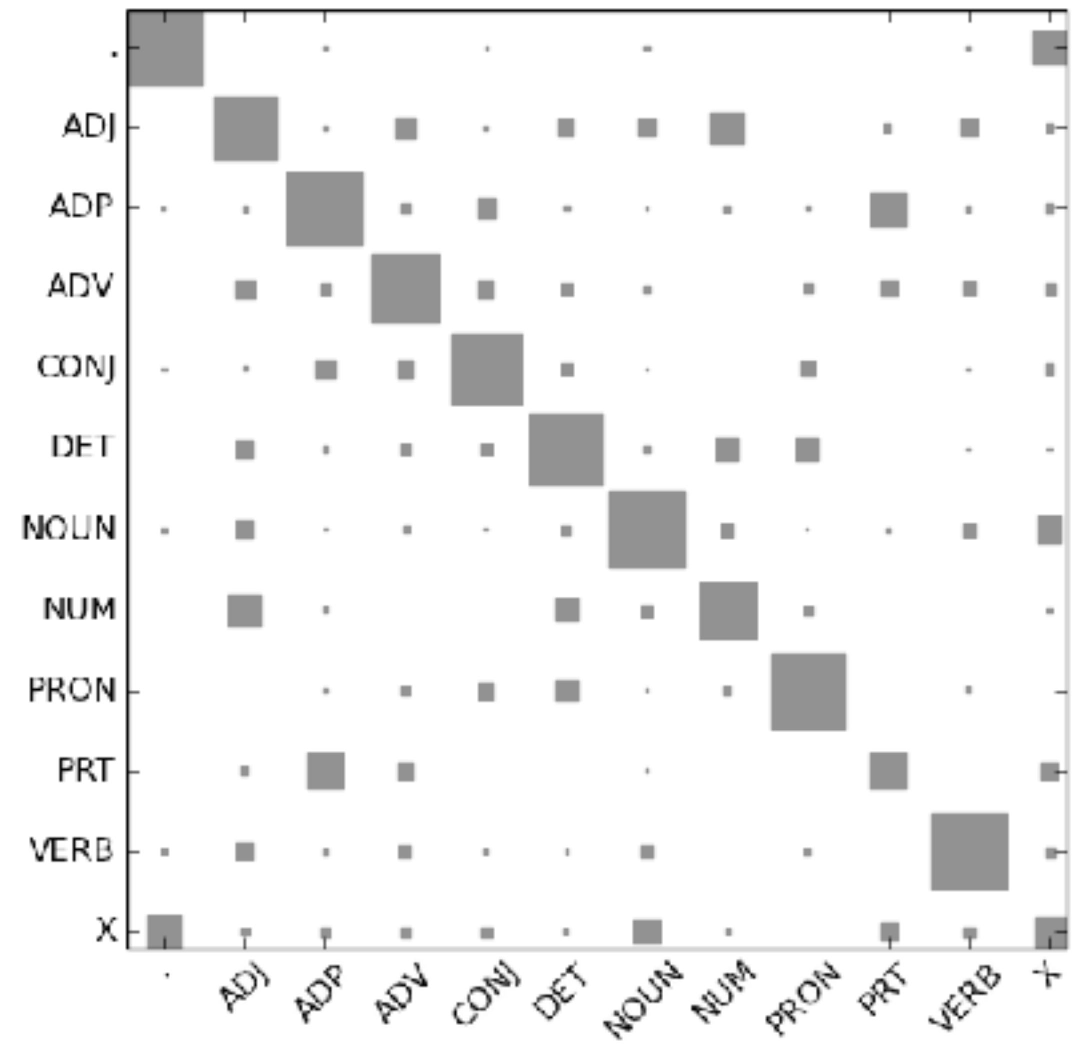
Are disagreements randomly distributed?

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

(Plank et al., 2014)

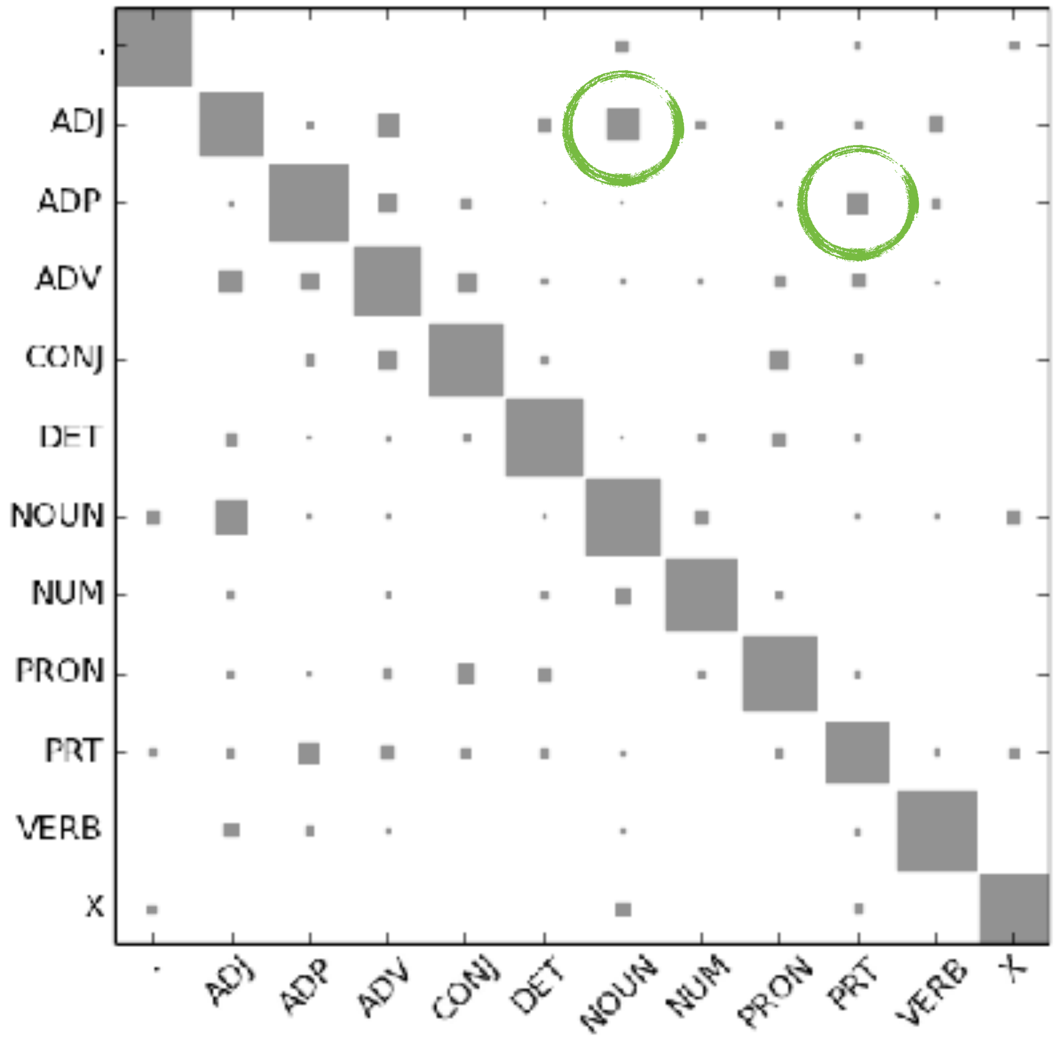


Wall Street Journal PTB-00

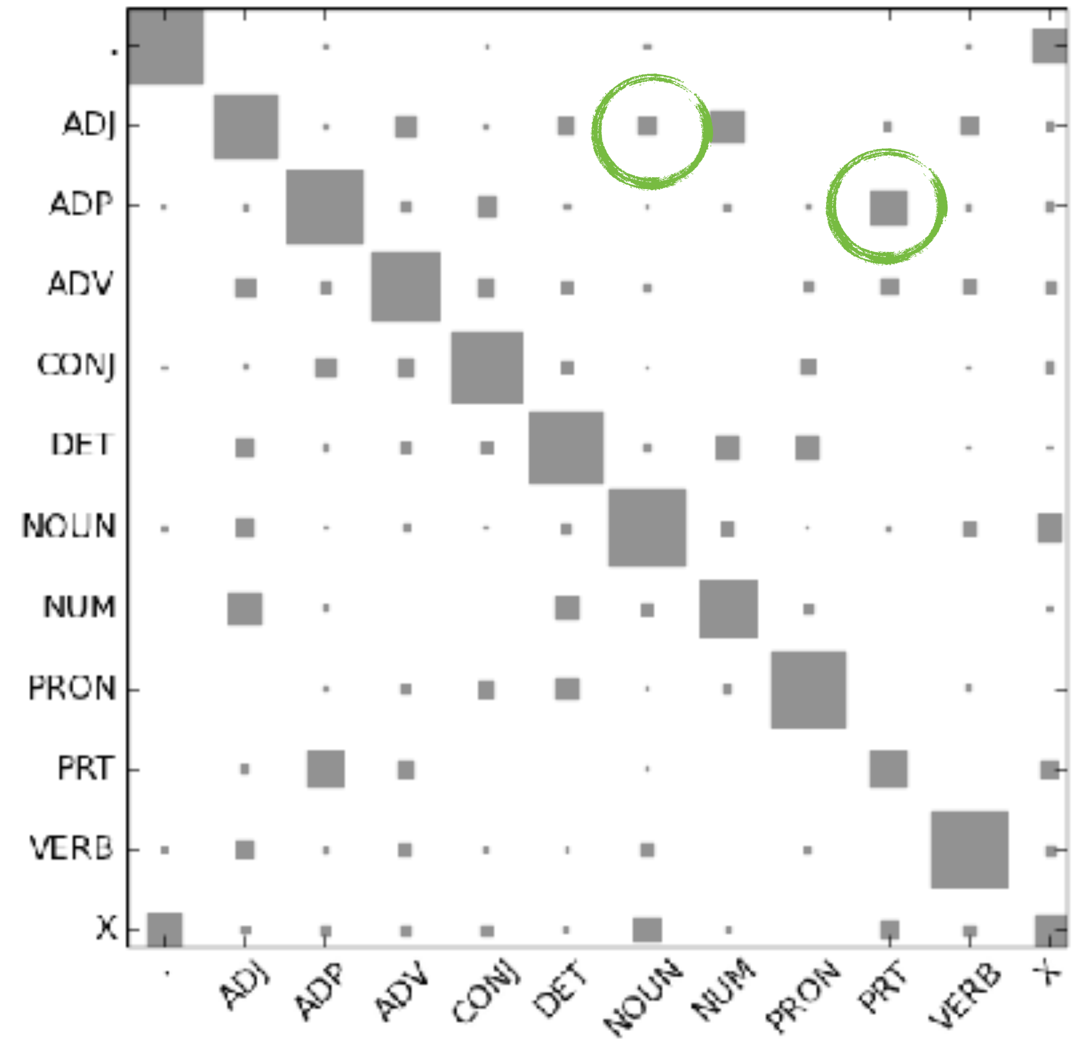


Twitter

(Plank et al., 2014)

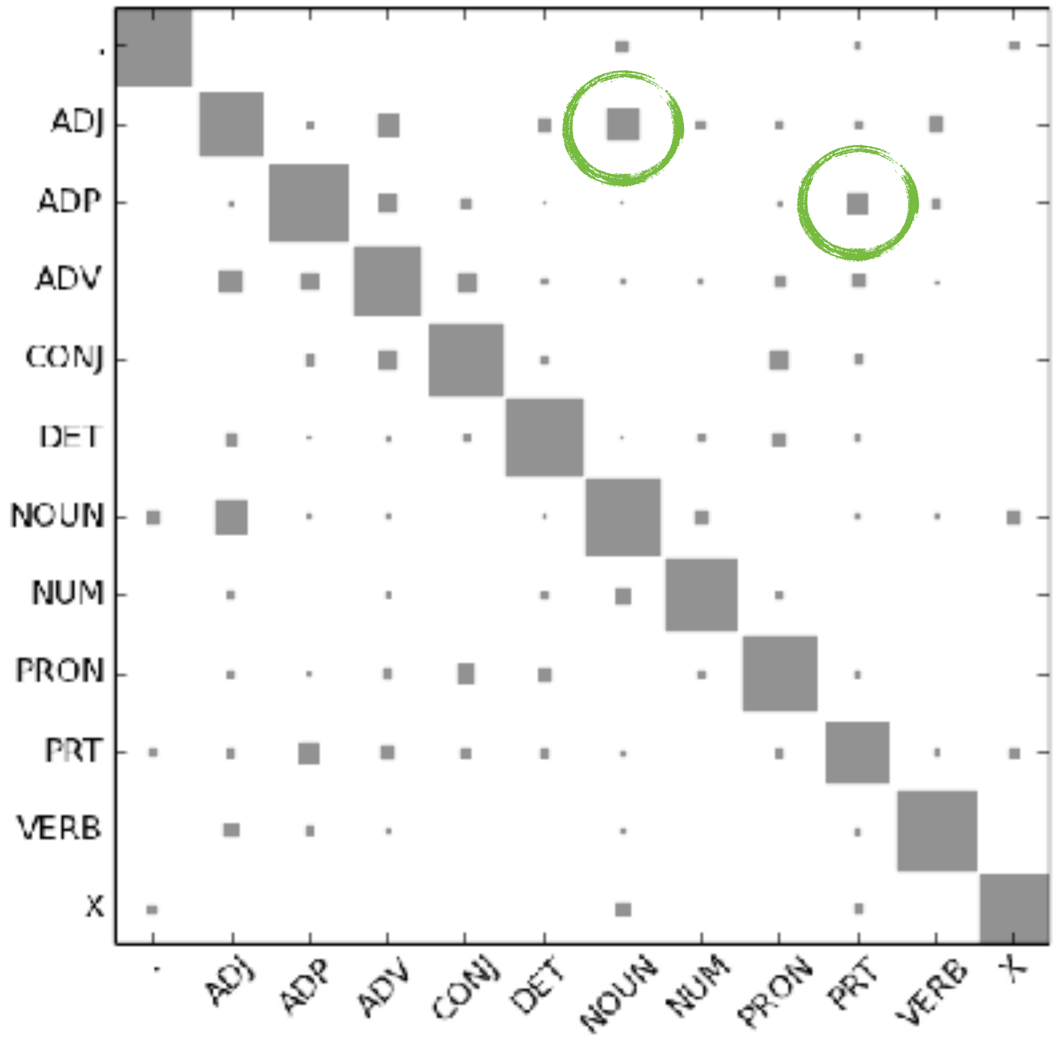


Wall Street Journal PTB-00

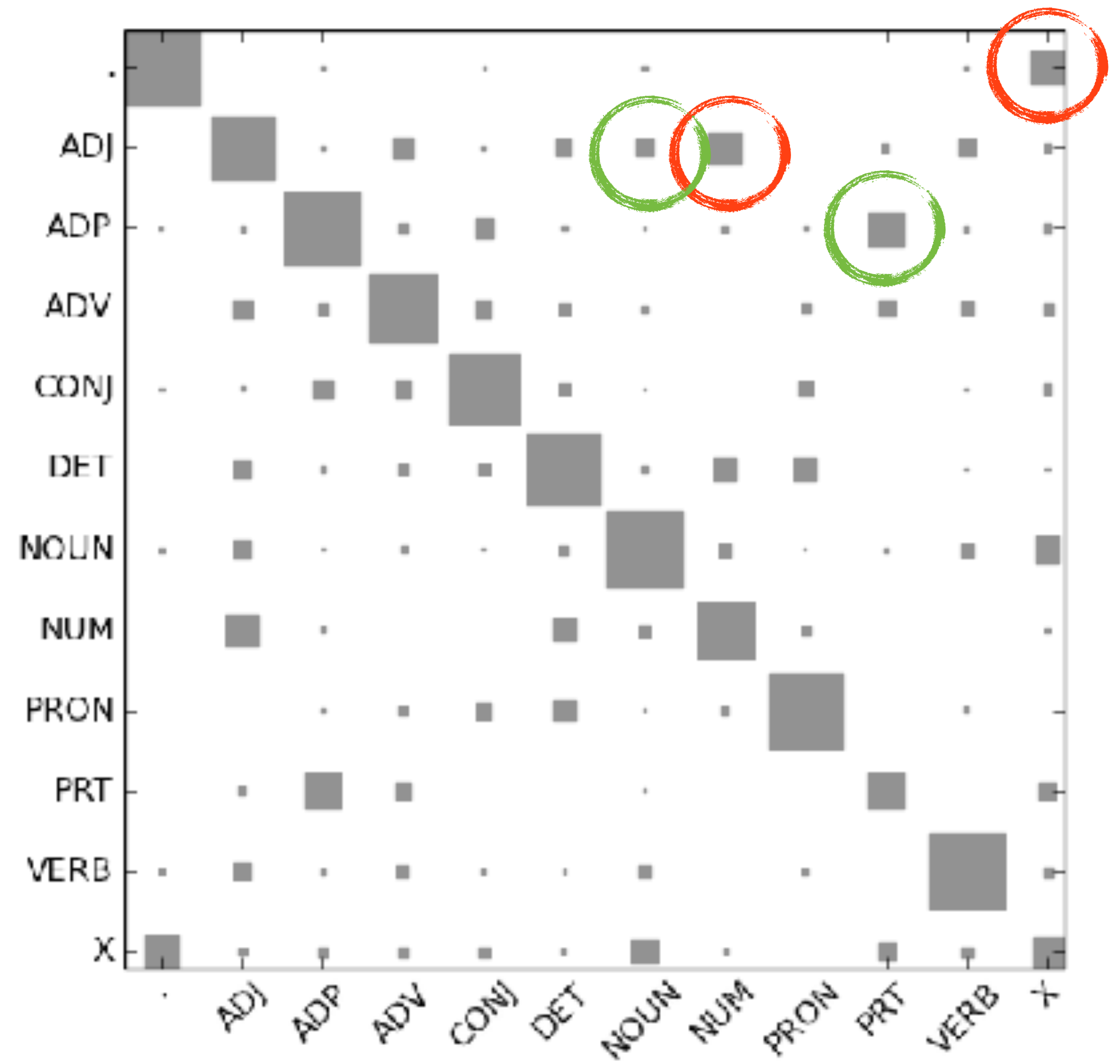


Twitter

(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)

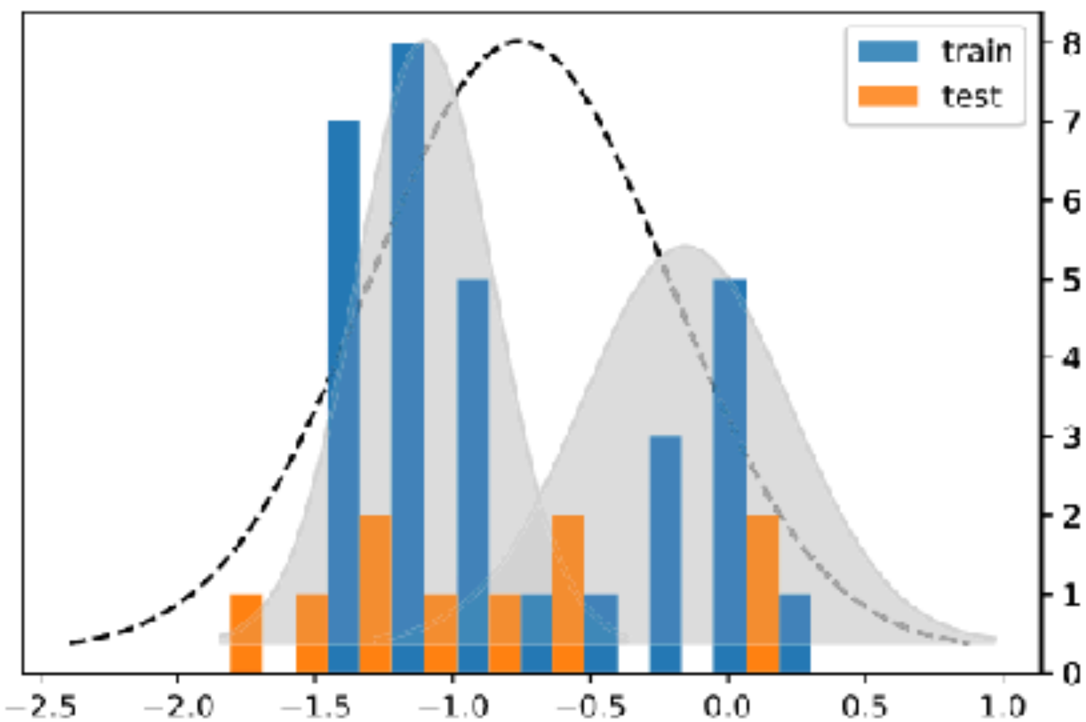
Are disagreement distributions unimodal?

… or do they contain inherent disagreement signal?

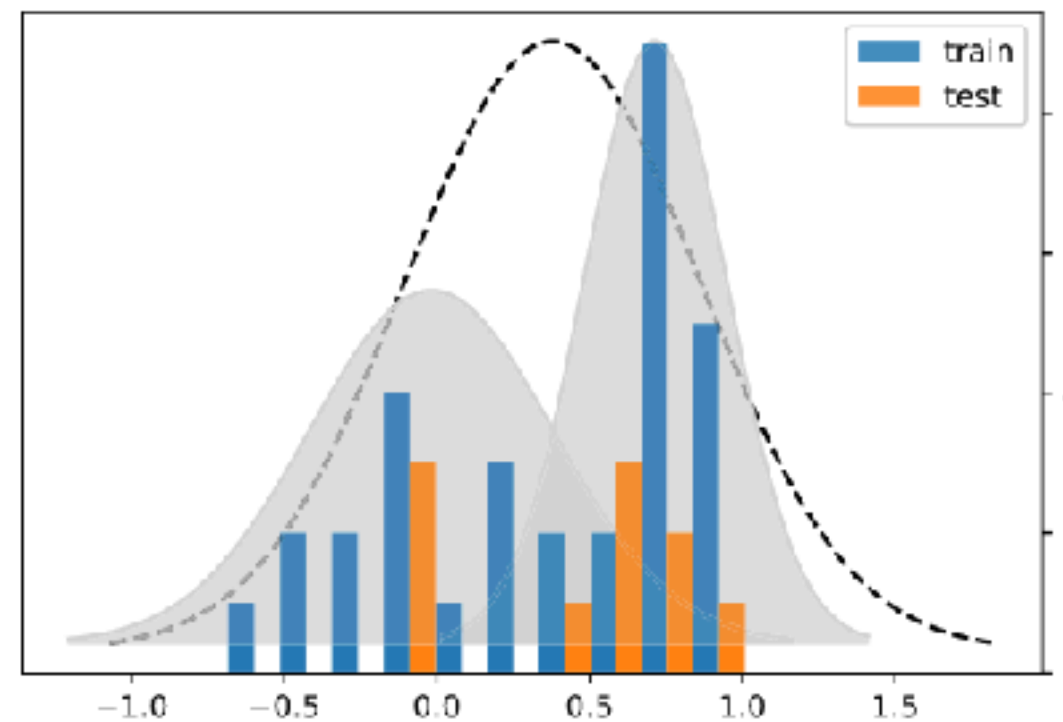
(Pavlick & Kwiatkovski, 2019)

Examples with bi-modal human judgement distributions

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.

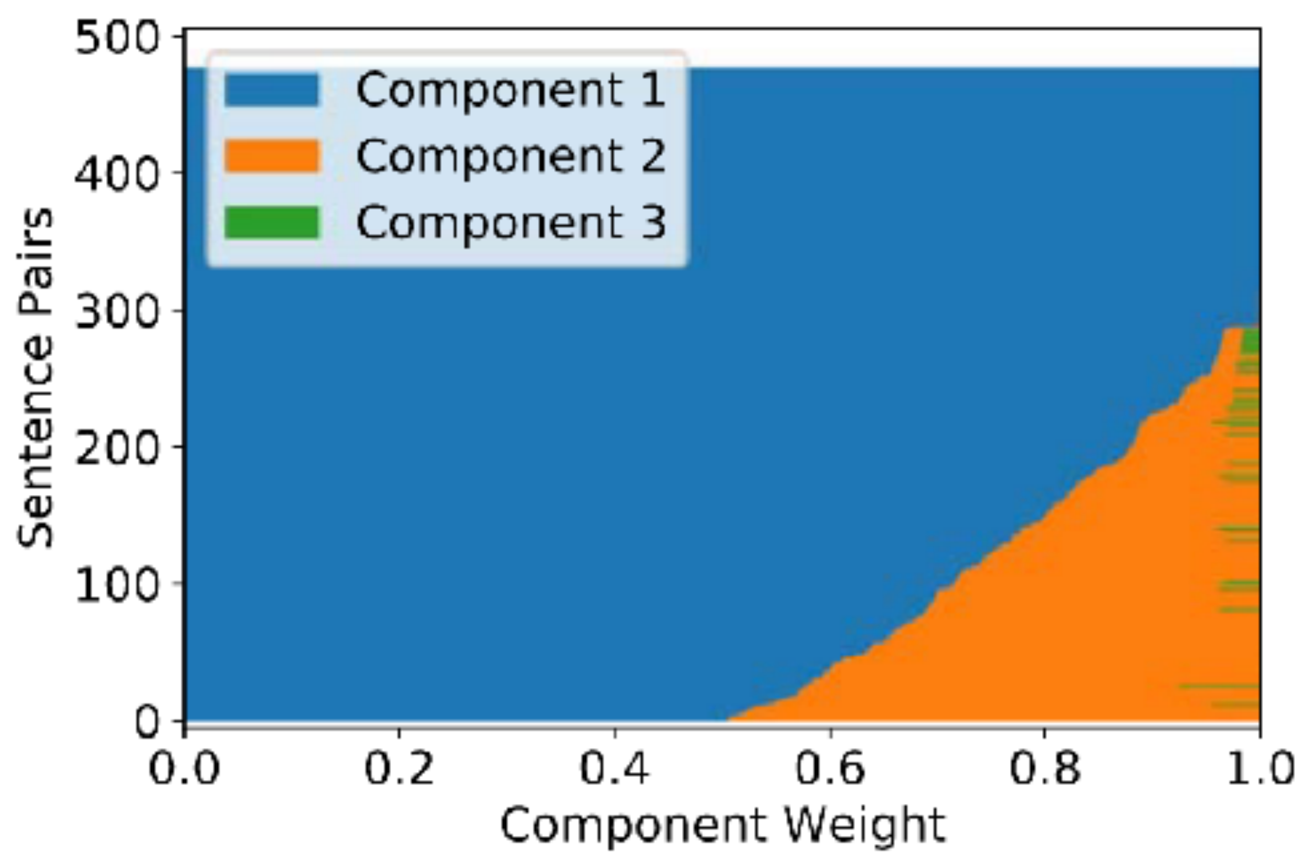


GMM with 1 *component* vs *k components*

(Pavlick & Kwiatkovski, 2019)

RTE Re-Annotation Analysis

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



(Pavlick & Kwiatkovski, 2019)

Are disagreement distributions unimodal? No.

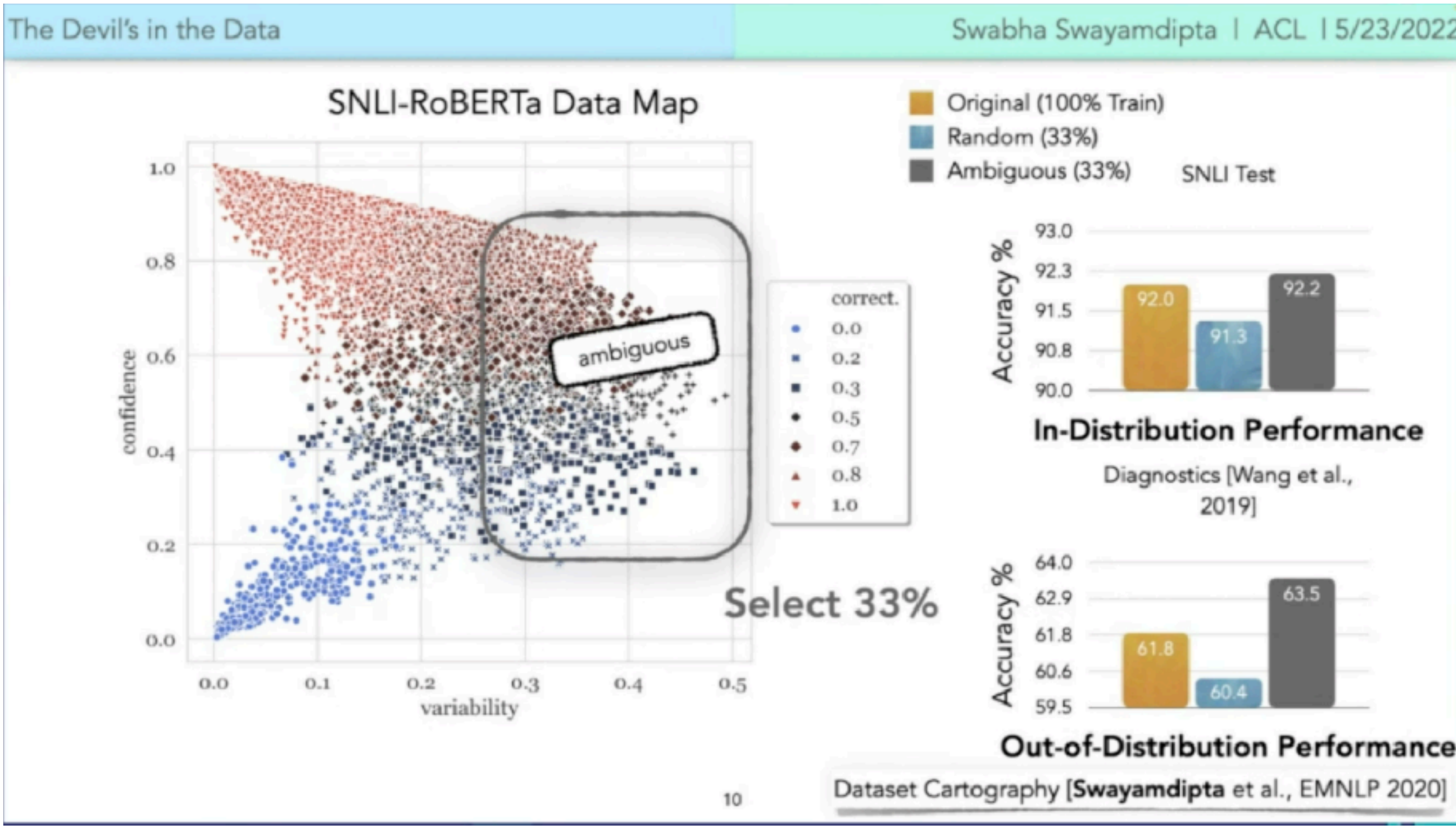
… do they contain inherent disagreement signal? Yes!

(Pavlick & Kwiatkovski, 2019)

Disagreement in human labeling is signal.

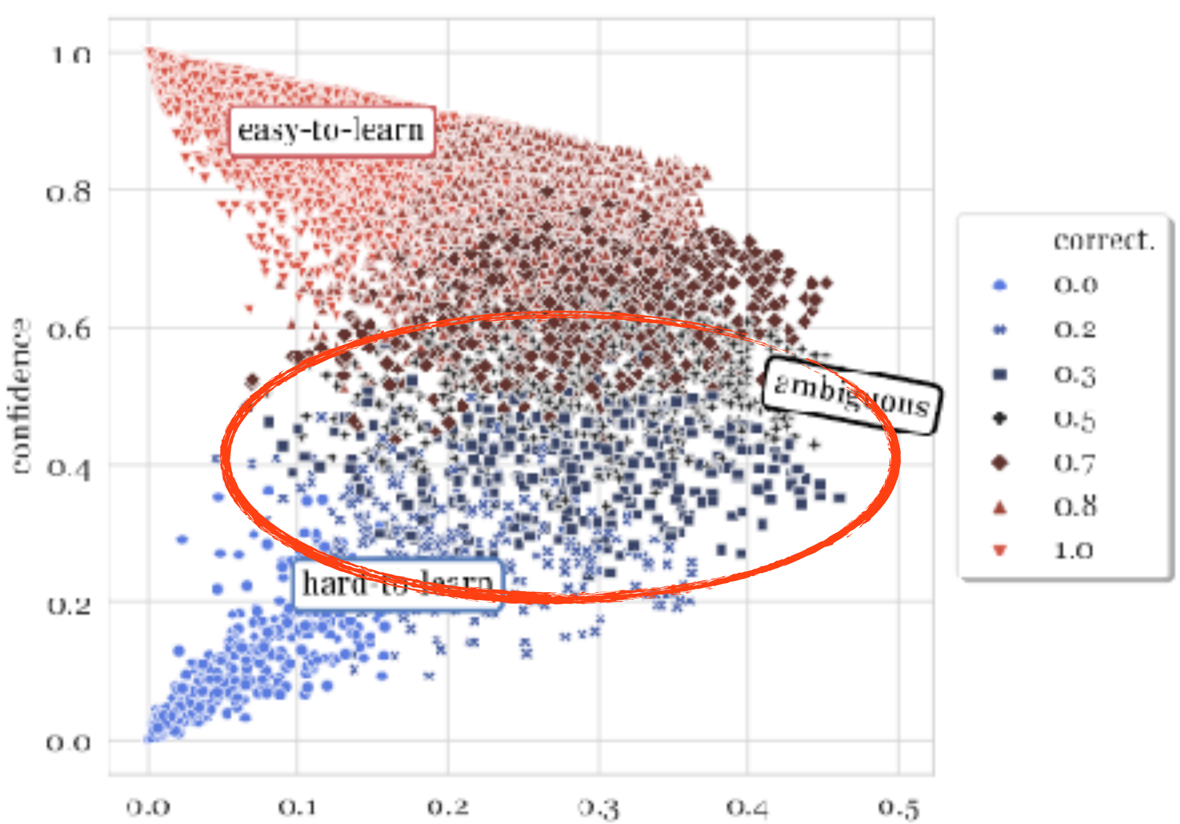
Further evidence: Ambiguous Instances help OOD generalisation

(Swabha Swayamdipta's ACL 2022 STIR talk)



Further evidence: Ambiguous Instances help active learning

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

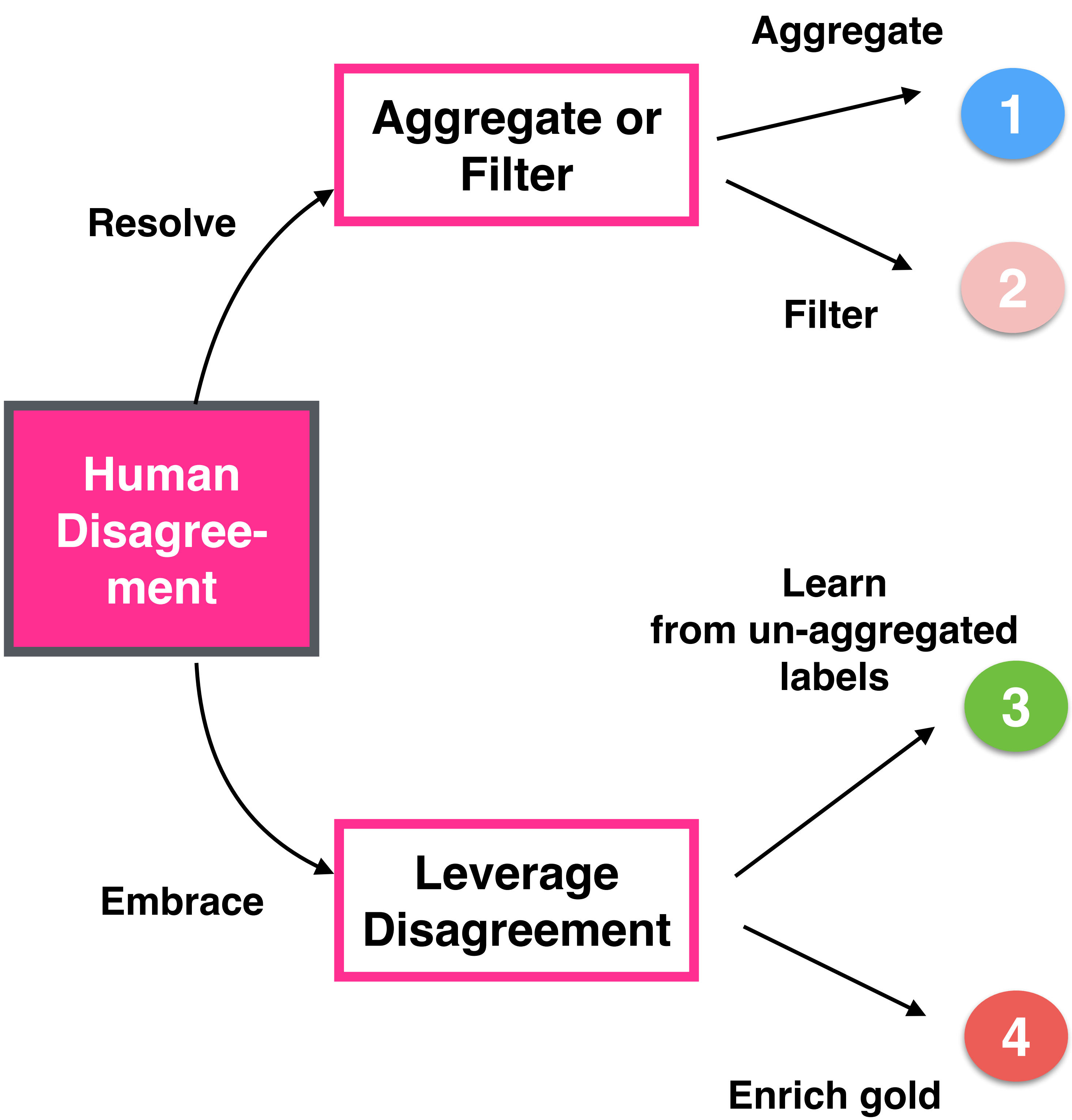


(Swayamdipta et al, 2020)

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

How can we leverage disagreement?

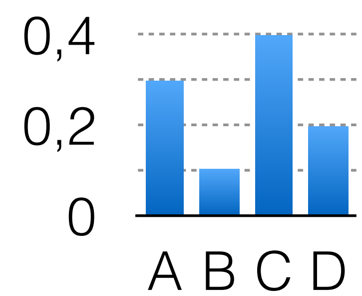
Learning with Disagreement



1. Aggregation methods
   (e.g. Dawid & Skene, 1979; Hovy et al., 2013; Paun, Artstein, Poesio, Morgan & Claypool book)
2. Filter data
   (e.g. hard filtering by Reidsema & van den Akker, 2008; Beigman-Klebanov et al.)
3. Repeated Labeling; Soft-labels, CrowdTruth, CrowdLayer, MTL
   (e.g. Peterson et al., 2019, Uma et al. 2020; Aroyo & Welty, 2014; Rodrigues & Peireia, 2017; Sheng et al., 2008; Specia & Cohn, 2013; Davani et al., 2021 TACL)
4. Weighting, Multi-task Soft Loss
   (e.g. Plank et al., 2014; Fornaciari et al., 2021)

Soft-labels

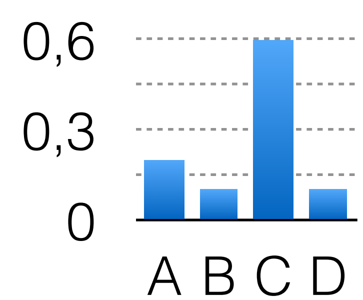
Annotator distribution P



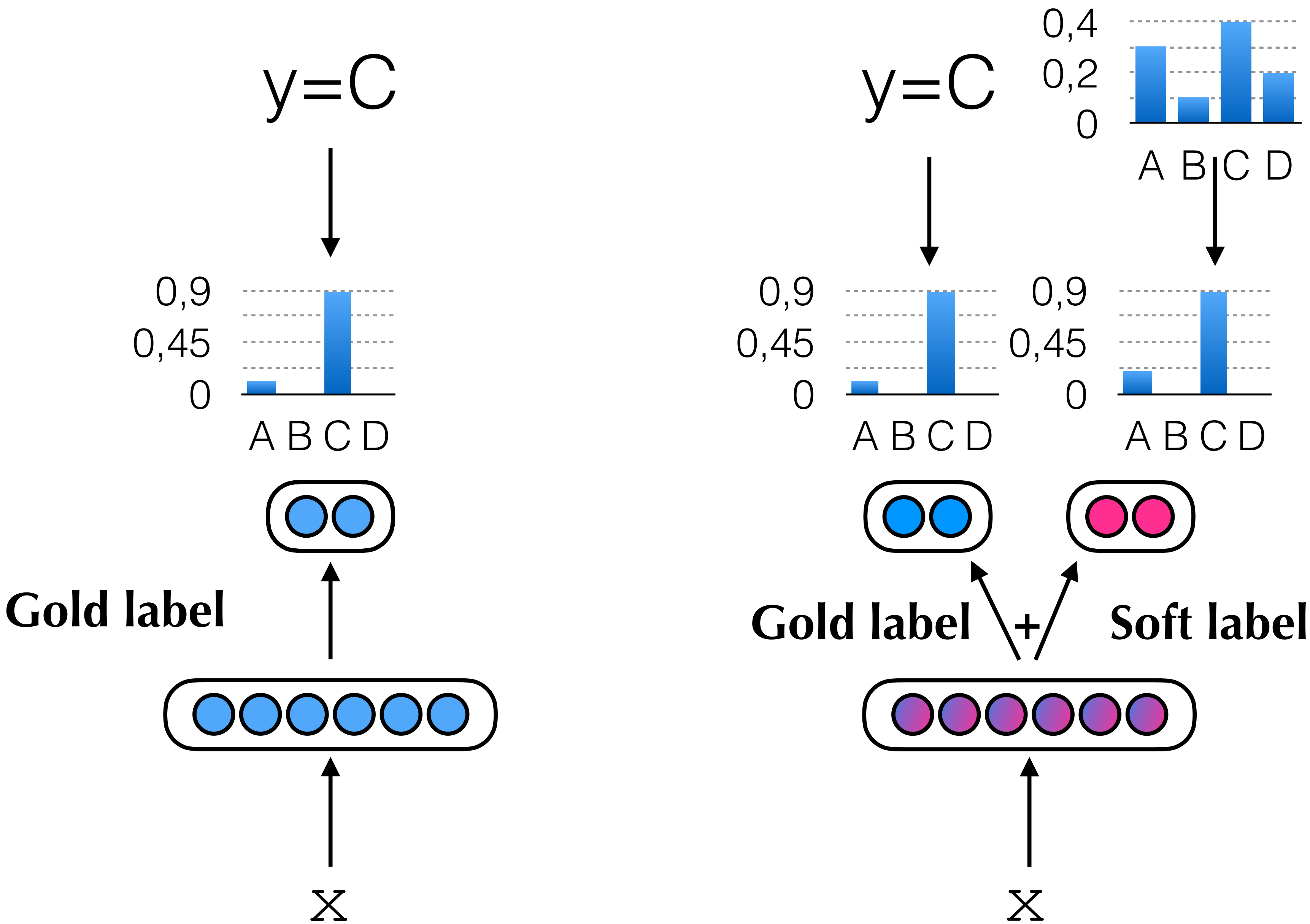
Measure divergence

$$D_{KL}(P||Q) = \sum_i P(i)\, log_2 \left(\frac{P(i)}{Q(i)}\right)$$

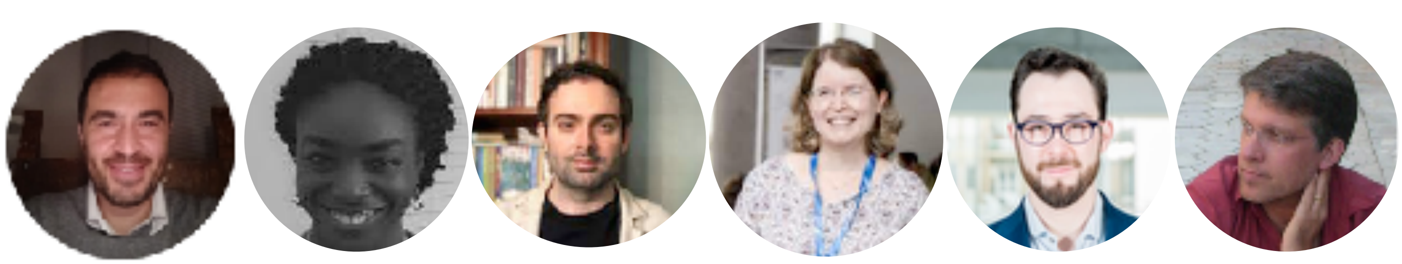
Predicted softmax Q



Soft-label Multi-Task Learning



- **Needs one auxiliary head** (instead of one per annotator as proposed by Specia & Cohn, 2013 and Davani et al., 2021)
- **Good results across tasks** (Uma et al., 2021 JAIR survey)



(Fornaciari, Uma, Paun, Plank, Hovy, Poesio 2021 NAACL)

— Human Disagreement in Labeling impacts all 3 stages of the NLP pipeline:

1) Data 2) Modelling 3) Evaluation

Is Human Label Variation So Bad? No.

It provides opportunities for more trustworthy, human-facing AI.

Ways Forward

Ways Forward (1/3): Data

- **Data:** collect & release **annotator-level labels & more meta-data**

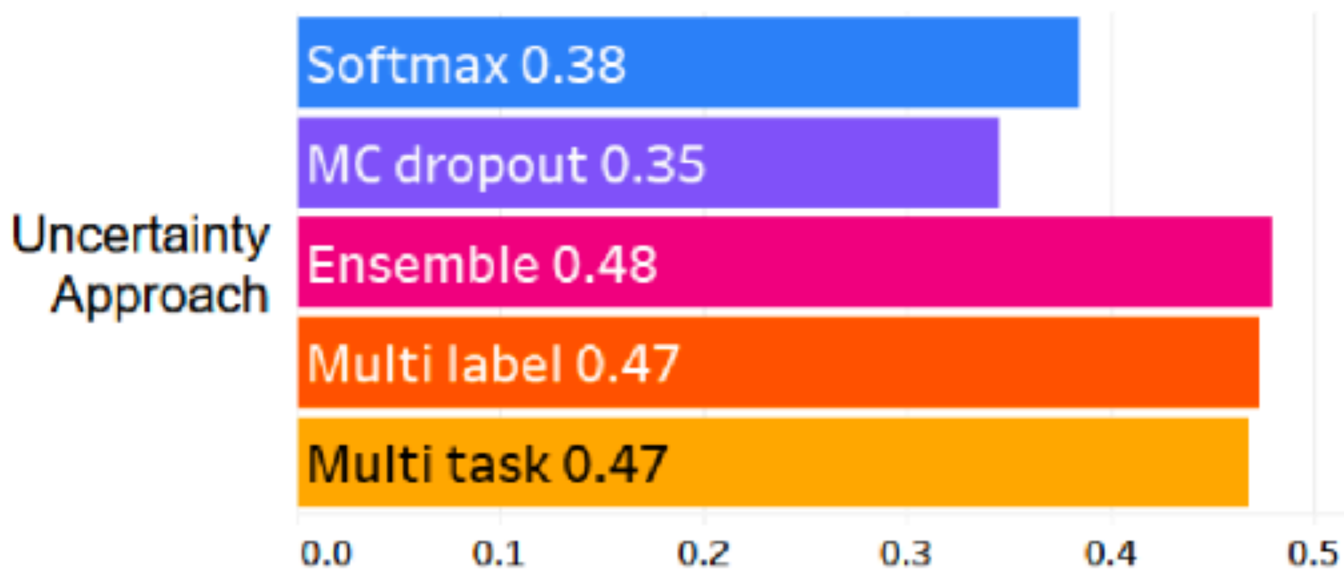
https://aclanthology.org/2021.law-1.14/

THE PERSPECTIVIST DATA MANIFESTO

pdai.info

Ways Forward (2/3): Modeling

- Human disagreement and **correlation** to **model uncertainty**



(Davani et al., 2021)

Ways Forward (3/3): Evaluation & Learning

- Rethink evaluation and the way we collect data
- Categories exist, but they are fluid; Let's not throw away signal!

Noise vs.

A range of

Human label variation

— To model Human Perspectives
— Provide highly-informative examples
(less but more informative data)

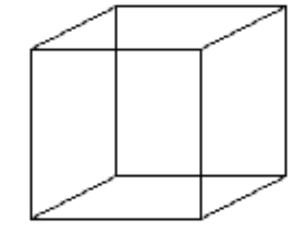
Fortuitous data

- Data out there,

  that waits to be harvested (**availability**),

  and can be used (relatively) easily (**readiness**)

Typology of fortuitous data



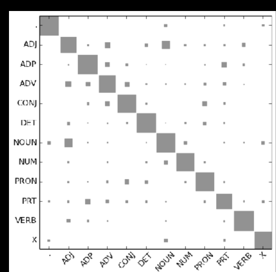
Plank (2016)

Type / Side benefit of	Examples	Availability	Readiness
meta-data	hyperlinks, HTML markup, genre labels, symbolic knowledge..	**+**	**+**
annotation	Human label variation (annotator disagreement)	**-**	**+**
behavior	cognitive processing data	**+**	**-**

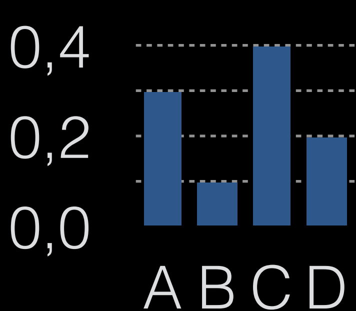
- Ways to use (non-standard) fortuitous data, related to ideas on “Incidental” supervision by Dan Roth

Take-home message

✓ not all **human label variation** is noise



✓ embrace it during **learning** / Let's not continue to model only the "mode", but the collective human label variation!



✓ embrace it during **evaluation**

✦ Research opportunities in this space

✦ Plug: Upcoming SemEval 2023 shared task

Elisa Leondardelli, Gavin Abercrombie, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, Massimo Poesio

To Sum Up

A ⚡“Genre” tags in UD are not perfect.

👉 Making **meta-data** count as weak supervision signal.

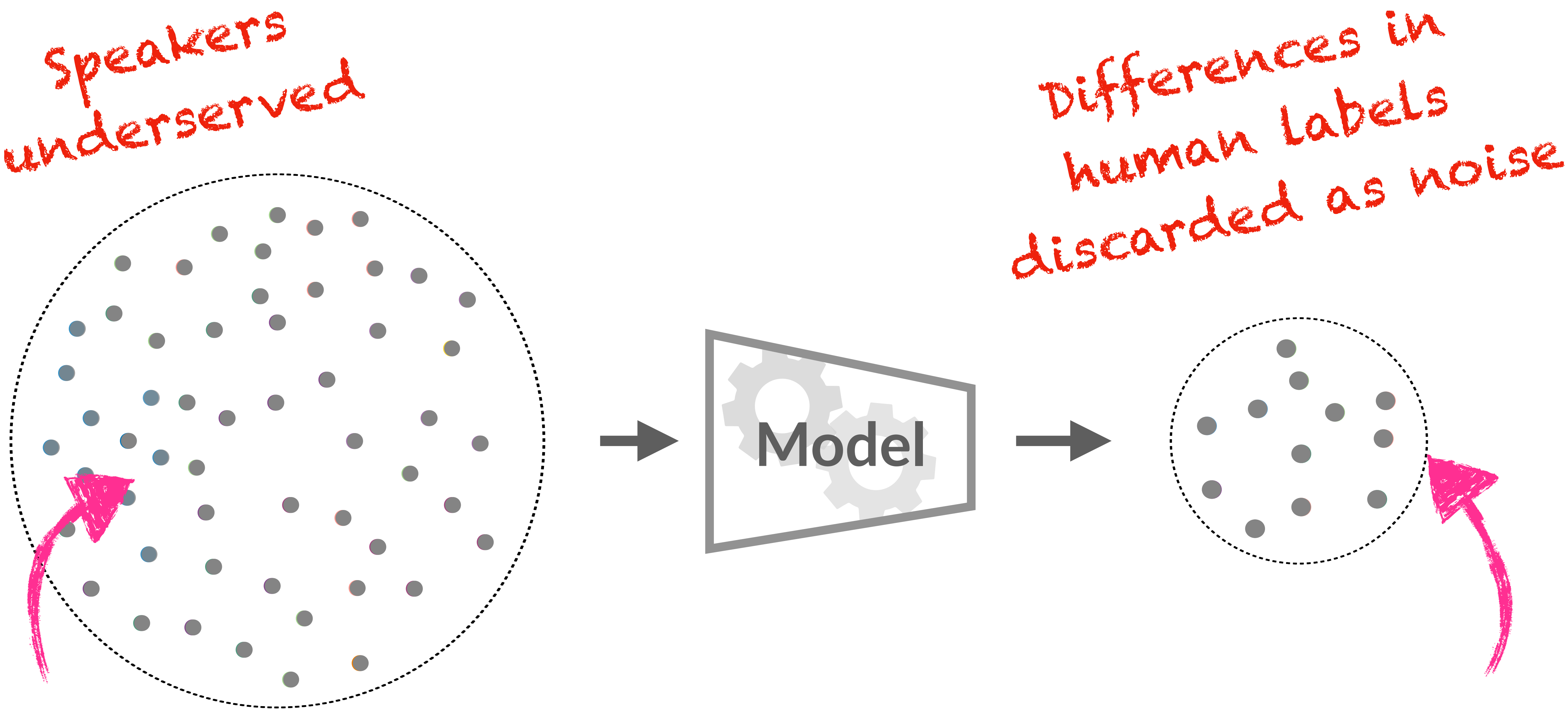
B ⚡ Choosing a good auxiliary task for transfer is difficult.

👉 Raw data via aux-MLM as effective, simple transfer method.

C ⚡ Humans disagreement in labels is noise.

👉 Making **human label variation** count in all steps of modelling.

Need To Account for Language Variability



Input: language variability not recognised

Output: only standardised categories accepted

Questions? Thanks!

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

@barbara_plank

B.Plank@lmu.de

Interested?
I'm hiring PhDs!

north

IT UNIVERSITY OF COPENHAGEN

erc

Research supported by:

DANMARKS FRIE FORSKNINGSFOND

MCML amazon

Thanks to all students, lab members and collaborators.