Transfer and Multi-Task Learning in Natural Language Processing

AILC summer school 2023, Pisa

Barbara Plank
Chair for AI and Computational Linguistics, MaiNLP lab, Center for Information and Language Processing (CIS), LMU München & NLPnorth lab, ITU Copenhagen

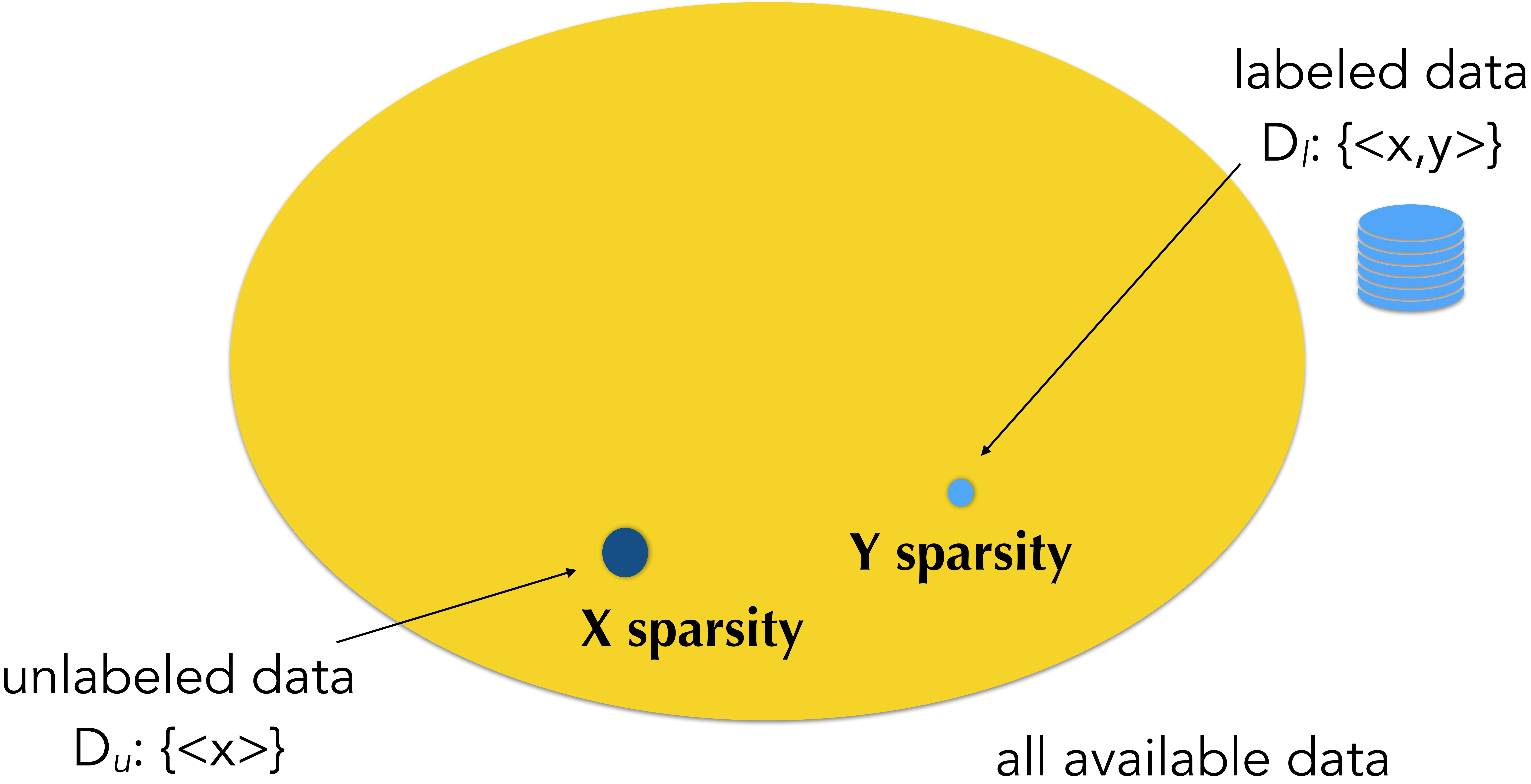
“We have millions of labeled data instances”

Very unlikely the case, especially for NLP.

D_l: {<x,y>}

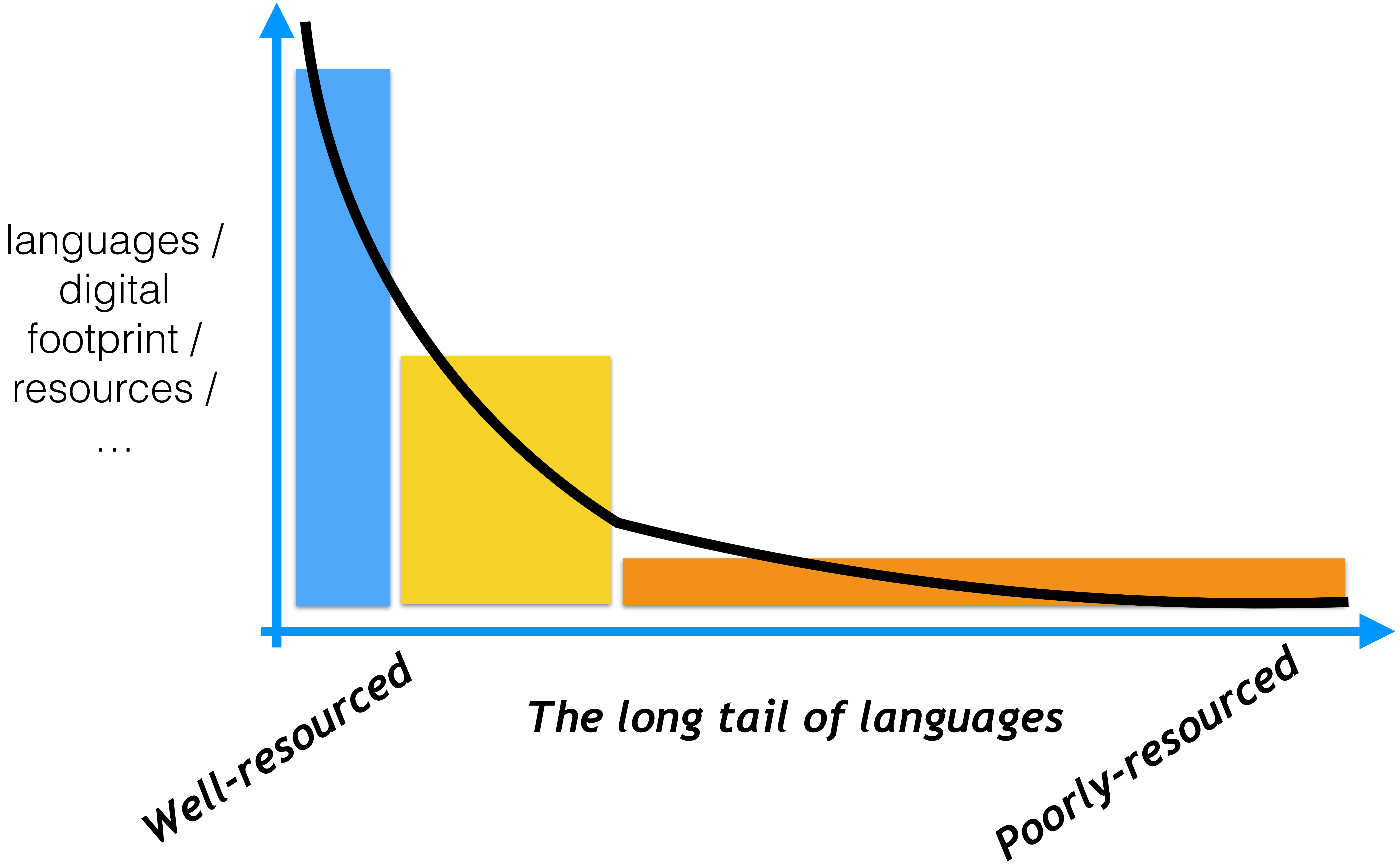
The Motivation: Data scarcity

Learning from limited labeled data

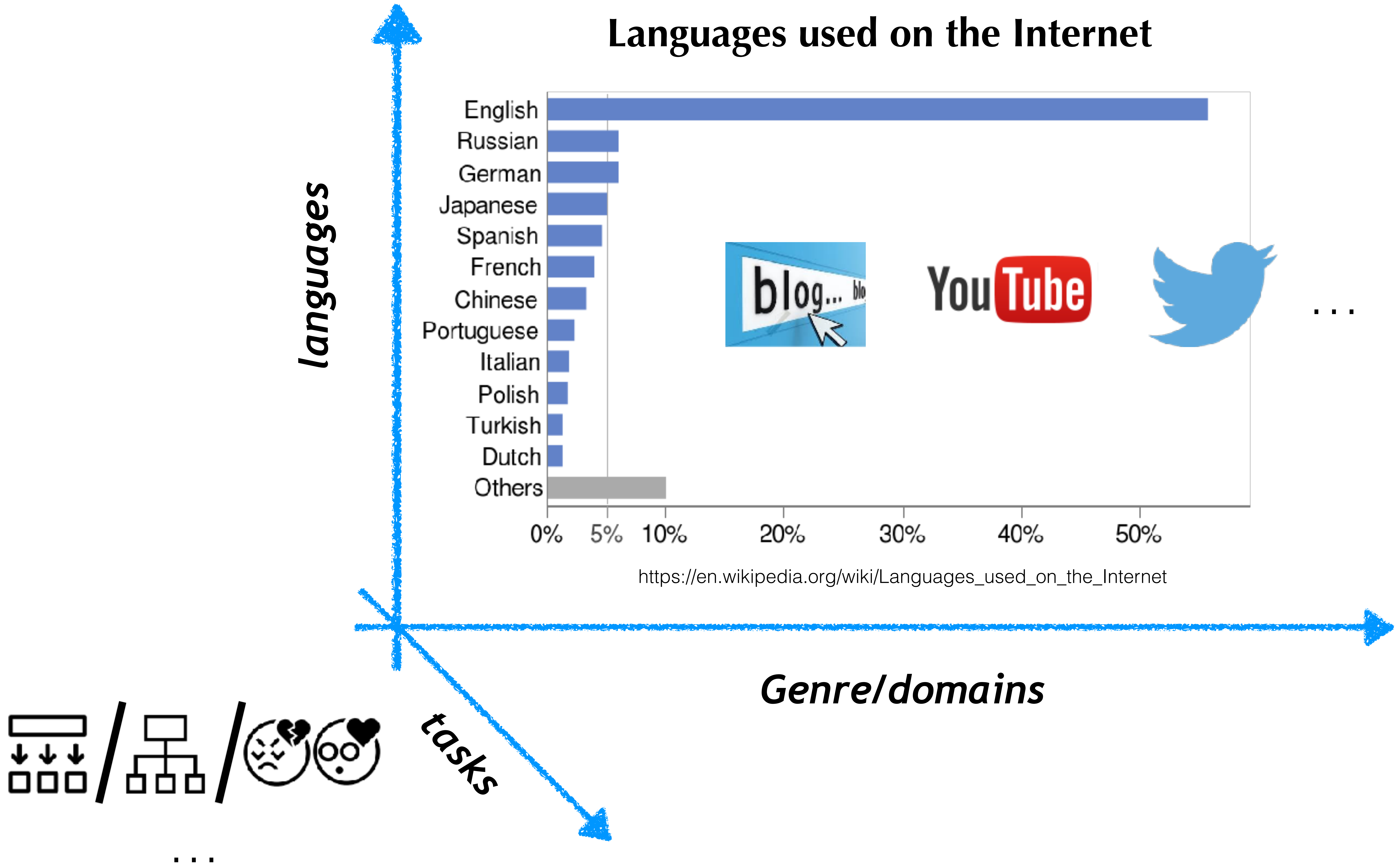


Many languages are poorly resourced

Despite of the richness of language, we constantly face the scarceness of data: Need to tackle the “long tail”



Ultimate Goal: NLP for everyone



https://en.wikipedia.org/wiki/Languages_used_on_the_Internet

What to do about it?

Typical setup: Learn a task at a time

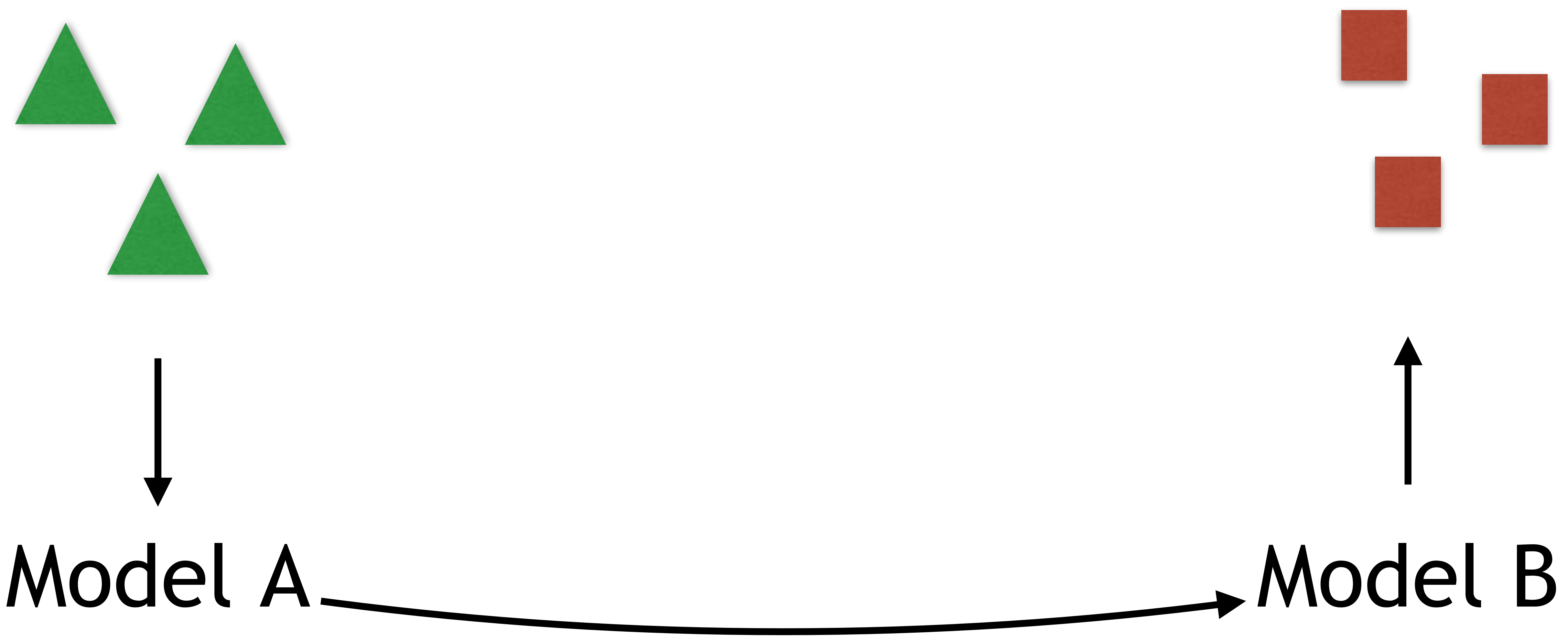
Starting from scratch: No transfer of knowledge

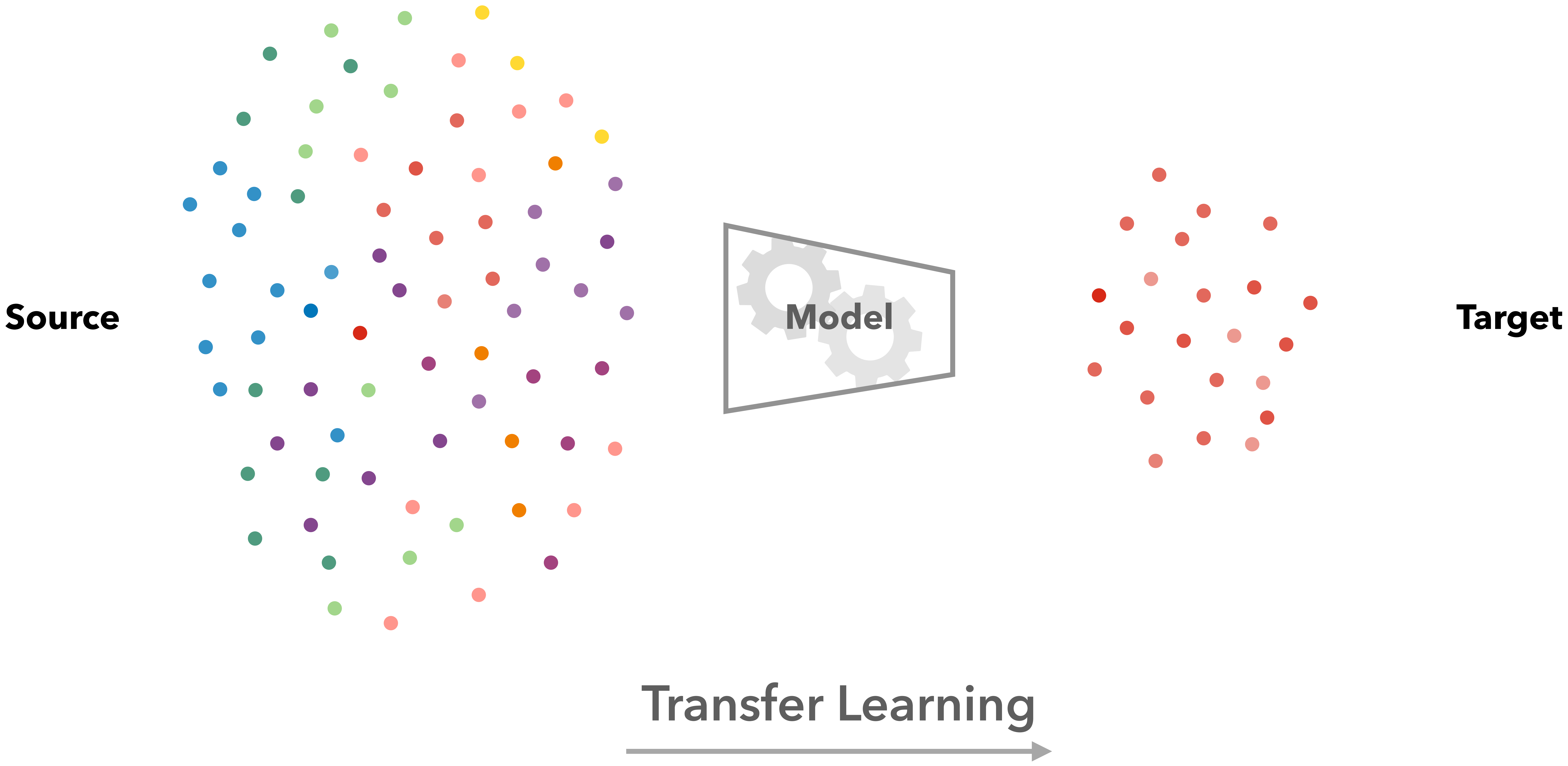
Model A

Model B

Transfer Learning (TL)

Leverage knowledge gained to help solve a related problem





Source

Model

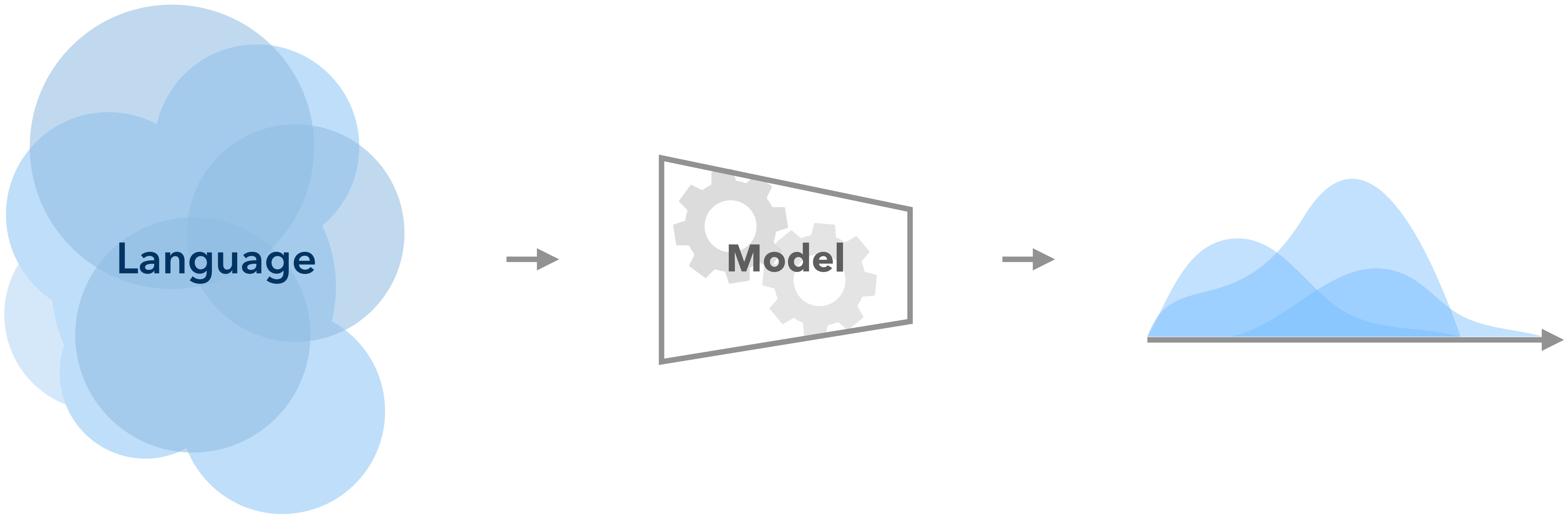
Target

Transfer Learning

Why Transfer Learning?

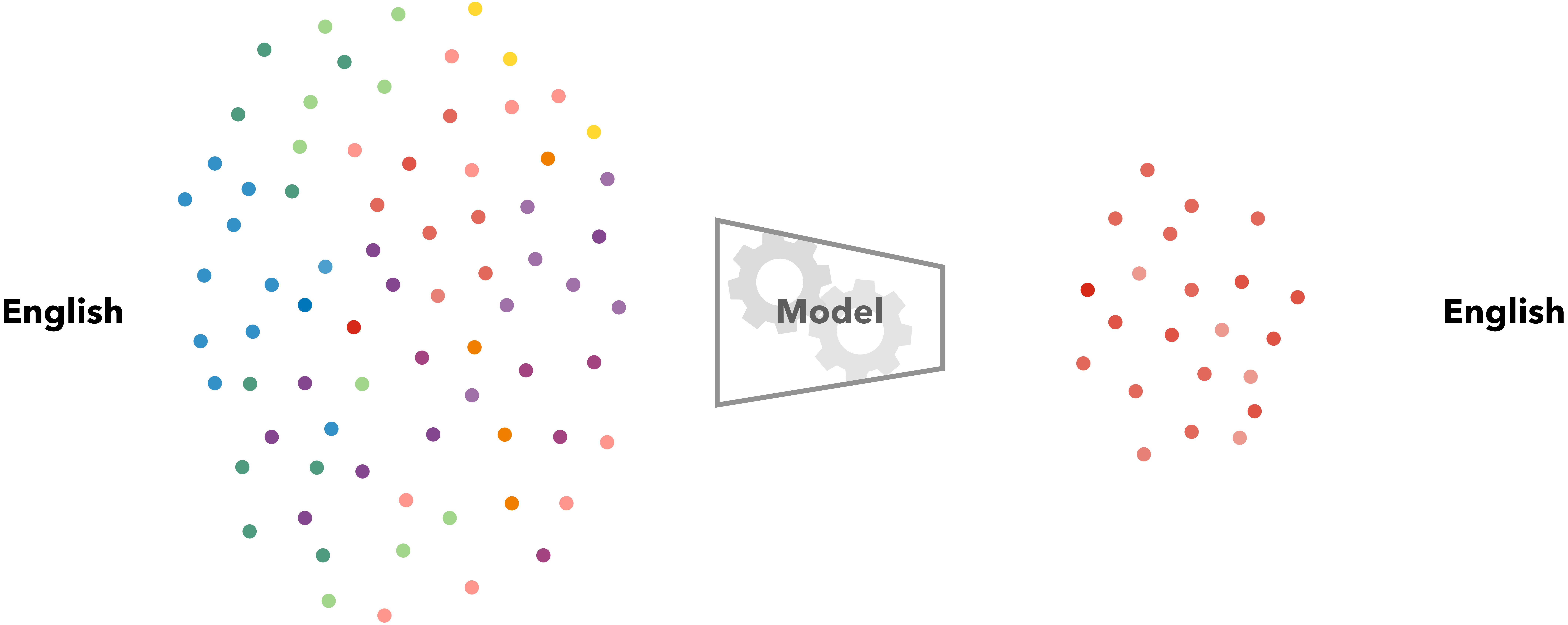
It is all about language variation & out-of distribution learning

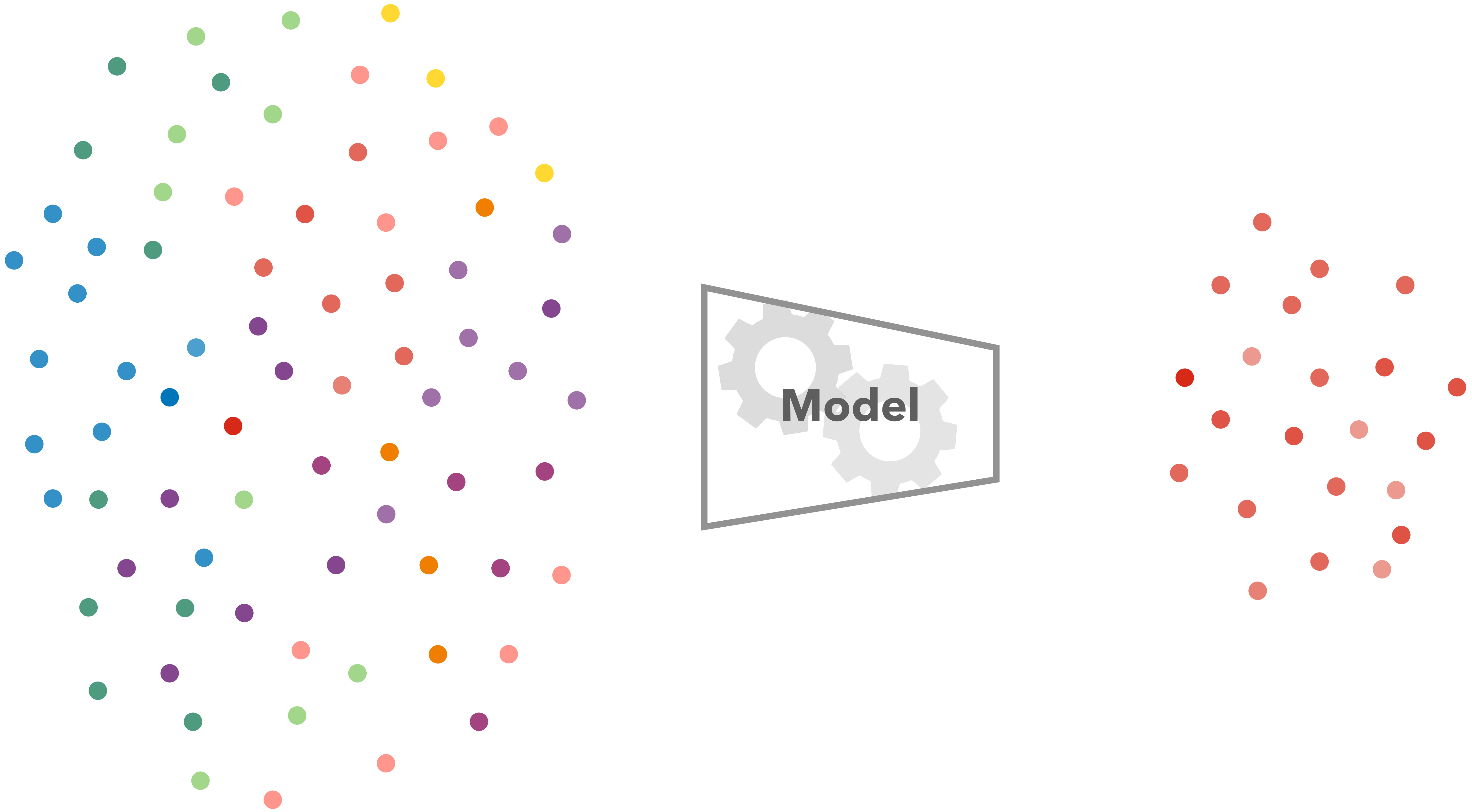
How do we make sure everyone is understood?



Language

Model


English
Model
English

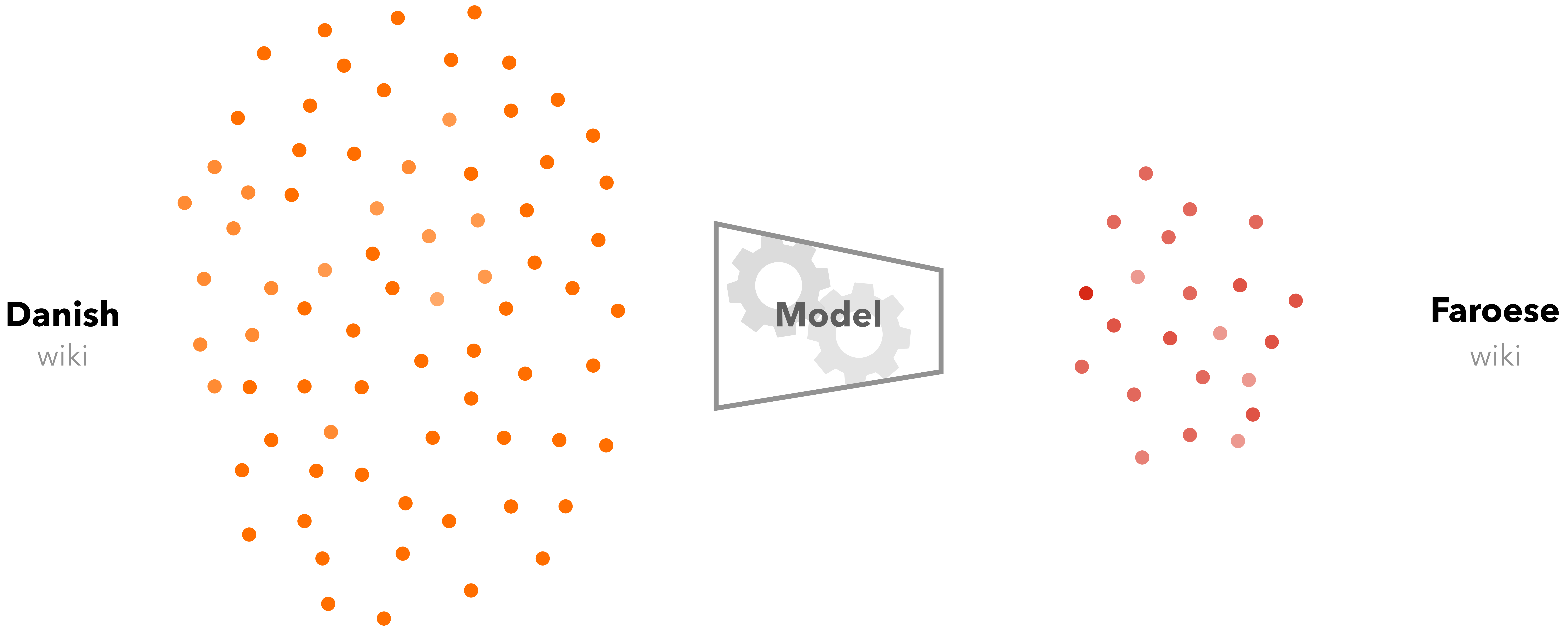


Multilingual

any source

Model

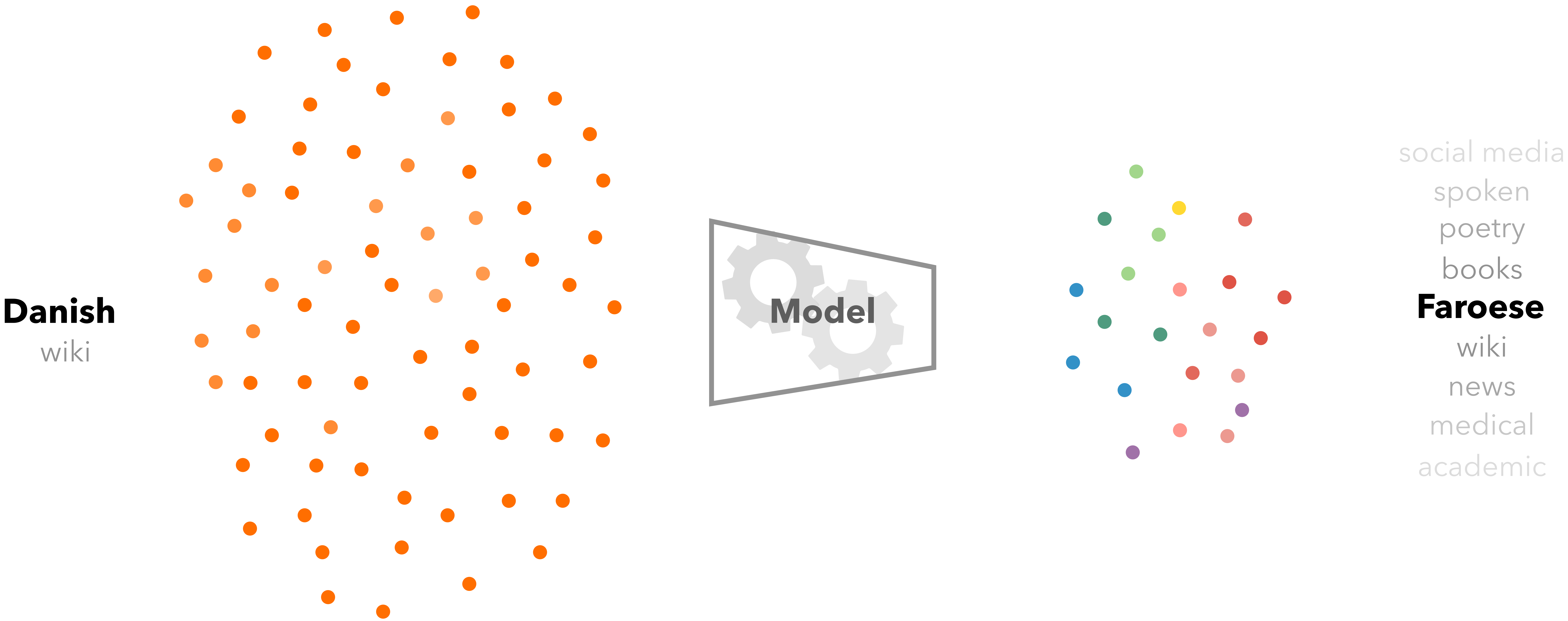
Faroese



Danish
wiki

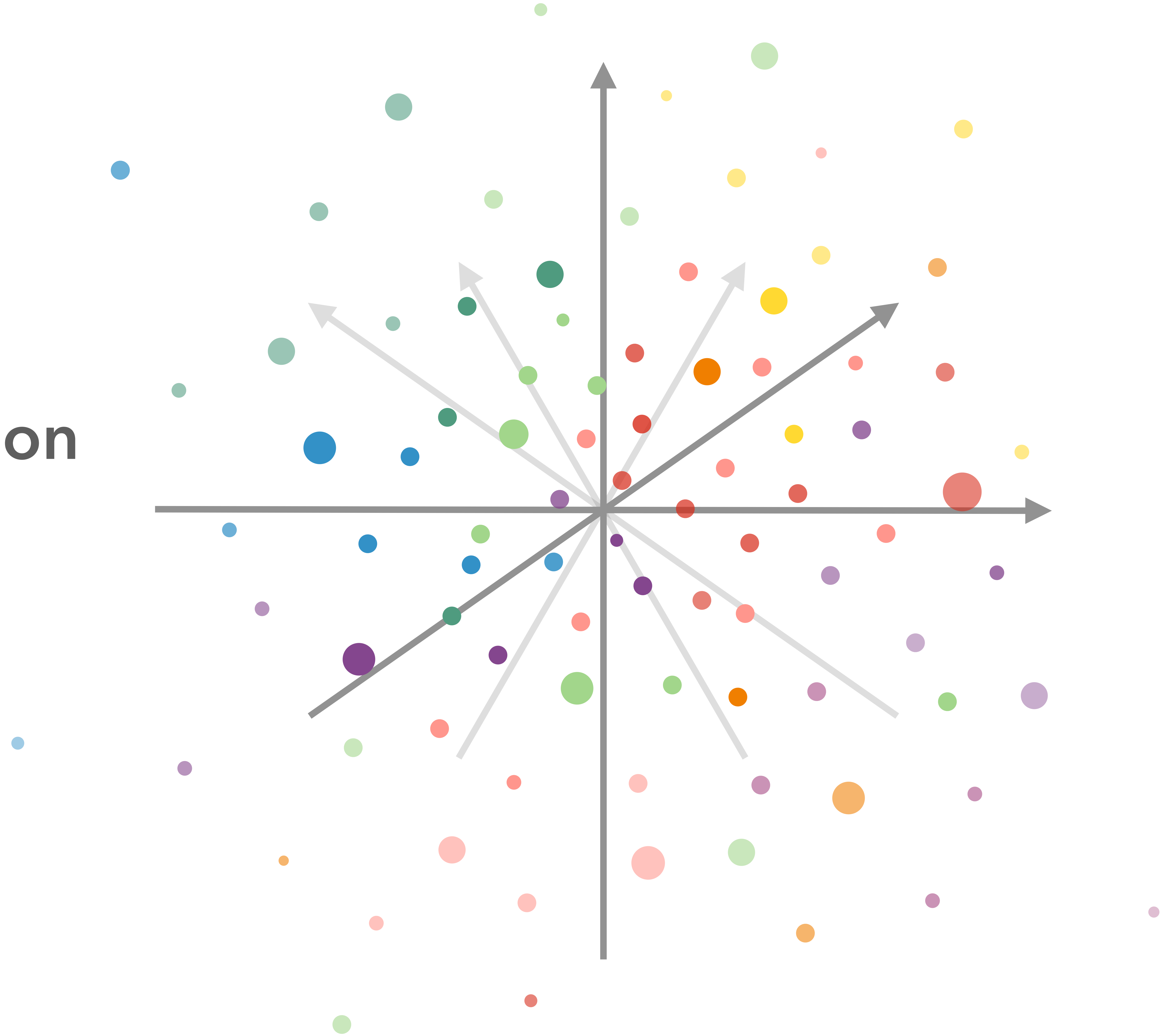
Model

Faroese
wiki



↑ Language Variation

Performance ↓



typology

domain

genre

topic

register

social context

Sekine (1997); Gildea (2001); Plank (2011); Ramesh Kashyap et al. (2021)
Biber (1988); Karlgren and Cutting (1994); Biber (1995); Lee (2001)

What this tutorial is (not) about

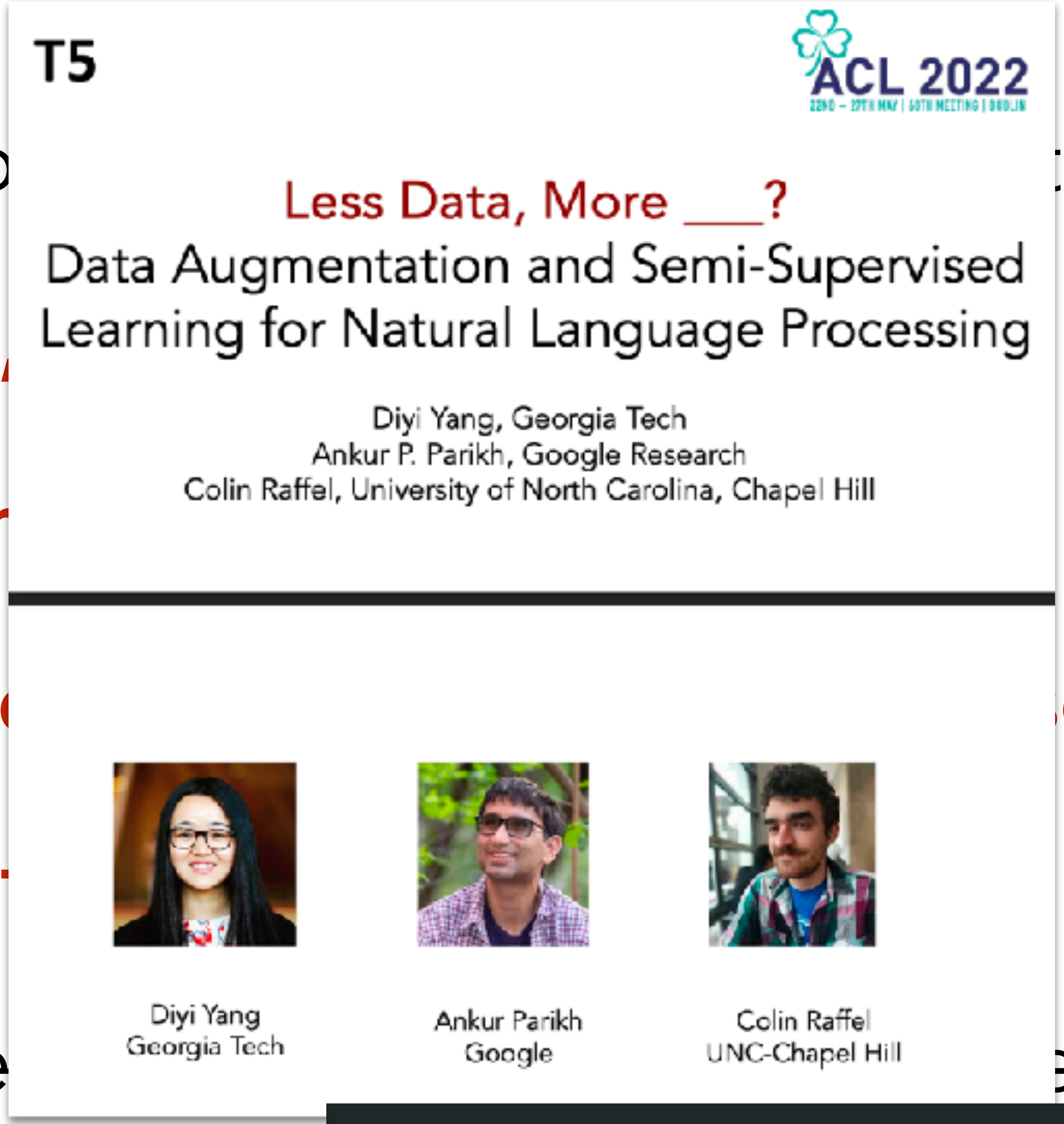
- An overview of early and recent approaches to TL in NLP. This tutorial is *not* exhaustive.
    - Pre-training (vanilla, multilingual, continuous)
    - Data selection (select data that matches the target)
    - Subspaces and Performance Prediction (investigate representations for transfer)
    - Multi-task Learning (use information from other tasks)
    - Data augmentation (modify labeled data to create class-preserving labeled data)
    - Semi-supervised learning (label from labeled and unlabelled data)
    - Zero-shot/few-shot learning (use no/few labeled instances or instructing tuning)
    - Active learning (select data to give to an annotator), Knowledge distillation (use a teacher to label the data), …

What this tutorial is (not) about

- An overview of early and recent approaches to TL in NLP. This tutorial is *not* exhaustive.
  - **Pre-training (vanilla, multilingual, continuous strategies)**
  - **Data selection (select data that matches the target)**
  - **Subspaces and Performance Prediction (investigate representations for transfer)**
  - **Multi-task Learning (use information from other tasks)**
  - ~~Data augmentation (modify labeled data to create class-preserving labeled data)~~
  - ~~Semi-supervised learning (label from labeled and unlabelled data)~~
  - ~~Zero-shot/few-shot learning (use no/few labeled instances or instructing tuning)~~
  - ~~Active learning (select data to give to an annotator), Knowledge distillation (use a teacher to label the data), …~~

What this tutorial is (not) about

- An overview of early and recent app tutorial is *not* exhaustive.
  - **Pre-training (vanilla, multilingual,**
  - **Data selection (select data that r**
  - **Subspaces and Performance Pre entations for transfer)**
  - **Multi-task Learning (use informa**
  - ~~Data augmentation (me sele erving labeled data)~~
  - ~~Semi-supervised learning (label fro label~~
  - ~~Zero-shot/few-shot learning (u ng)~~
  - ~~Active learning (select data to give to an a use a teacher to label the data), …~~



T5

ACL 2022

Less Data, More ___?

Data Augmentation and Semi-Supervised Learning for Natural Language Processing

Diyi Yang, Georgia Tech
Ankur P. Parikh, Google Research
Colin Raffel, University of North Carolina, Chapel Hill

Diyi Yang
Georgia Tech

Ankur Parikh
Google

Colin Raffel
UNC-Chapel Hill

Zero- and Few-Shot NLP with Pretrained Language Models

Iz Beltagy, Arman Cohan, Robert L. Logan IV, Sewon Min, Sameer Singh

Outline

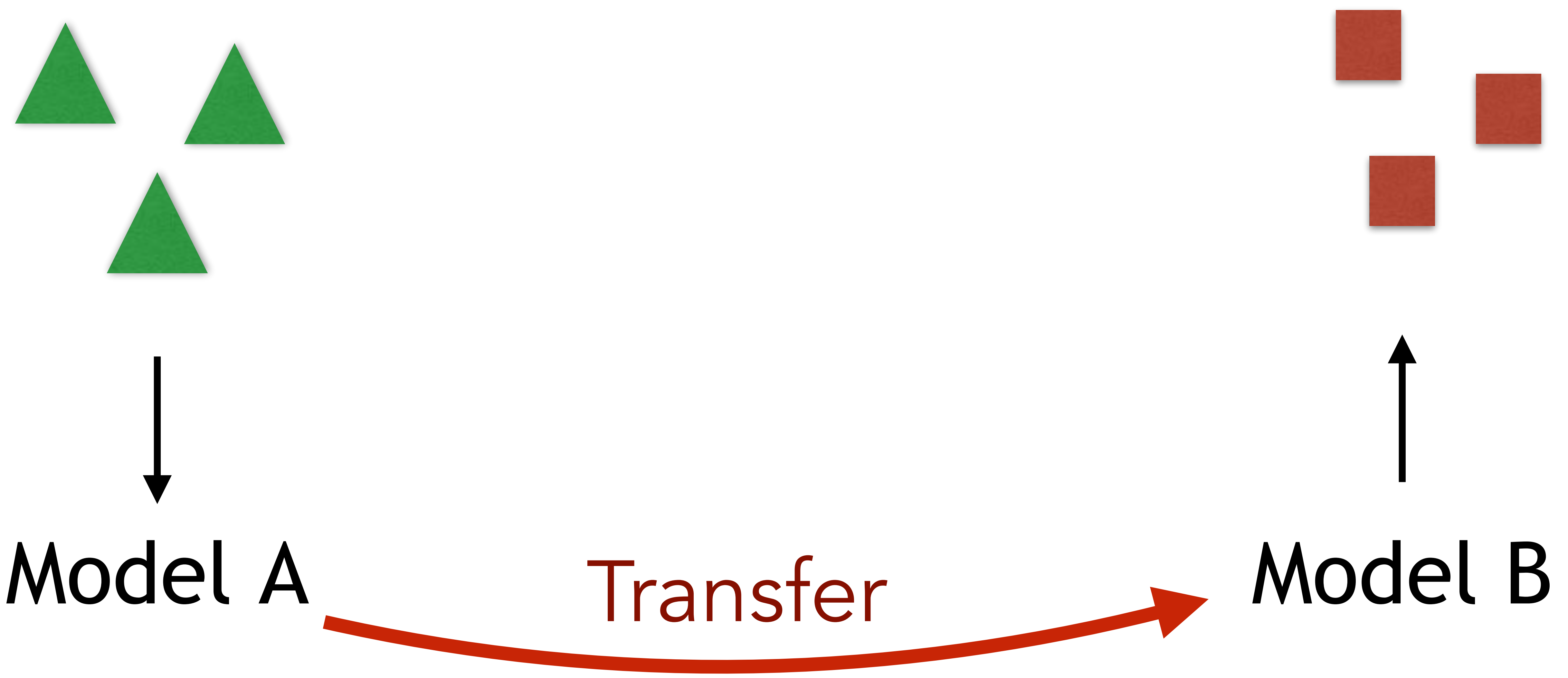
- Introduction: Why Transfer? Dimensions of Language Variation
- Part 1: What is Transfer Learning?
  - Three views on Transfer Learning, Related Learning Strategies
- Part 2: A type of TL: What is Multi-Task Learning?
  - What and Why, Perspectives on MTL
  - Short hands-on tutorial with MaChAmp
- Part 3: Selected Case Studies
  - Applications to Multilinguality, Transferability Estimation, Human Label Variation
- Outro

Part 1: What is Transfer Learning?

Views on Transfer Learning, Related Learning Strategies

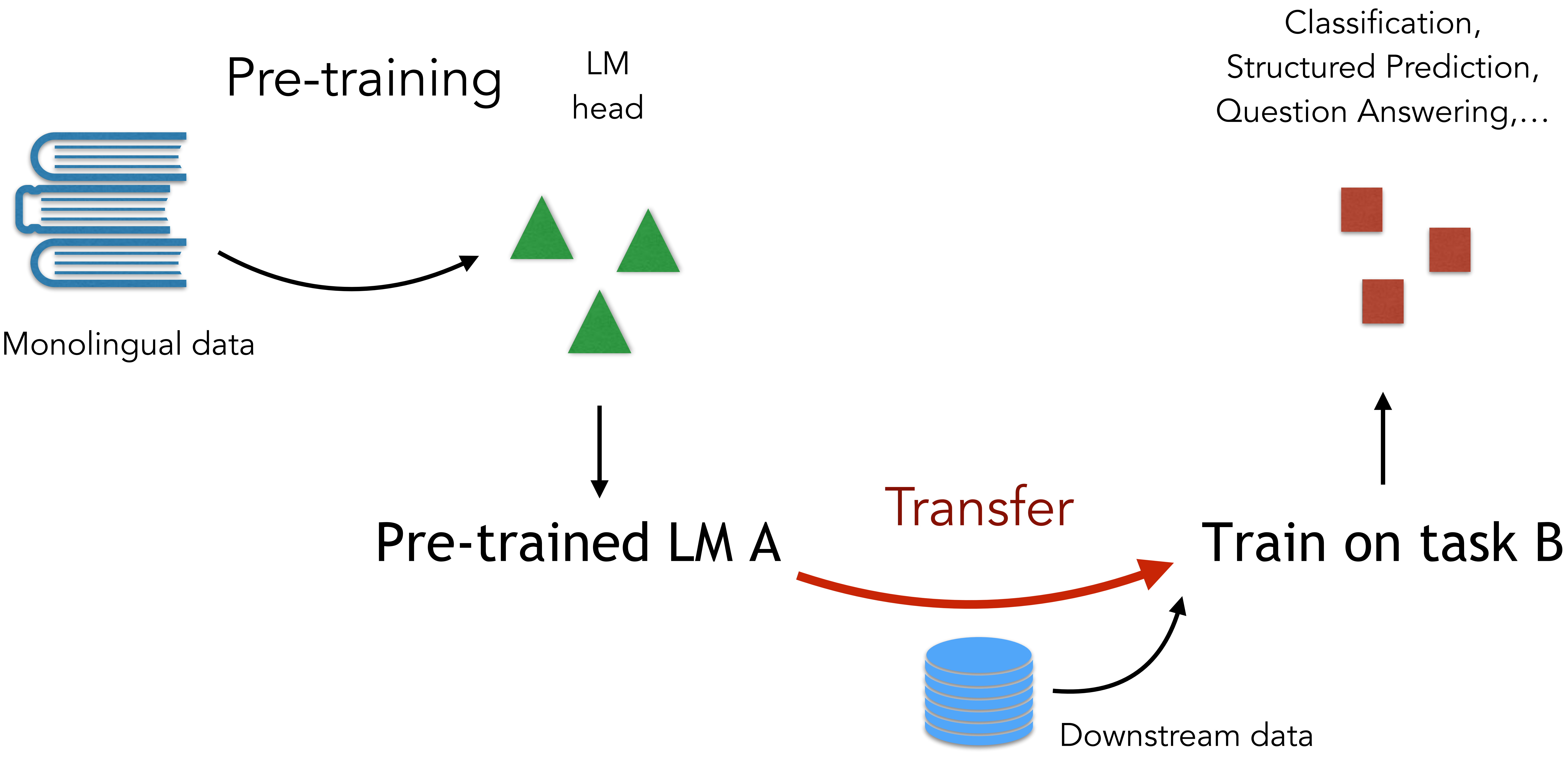
Transfer Learning (TL)

Leverage knowledge gained to help solve a related problem



Today's typical Transfer Learning (TL) setup = Sequential Transfer Learning

Learn on one dataset / task, then transfer to another dataset / task



Pre-training

LM head

Monolingual data

Pre-trained LM A

Transfer

Train on task B

Downstream data

Classification, Structured Prediction, Question Answering,…

Is this all there is to TL? No.

Sequential TL is just one (narrow) view on transfer learning. TL is broader

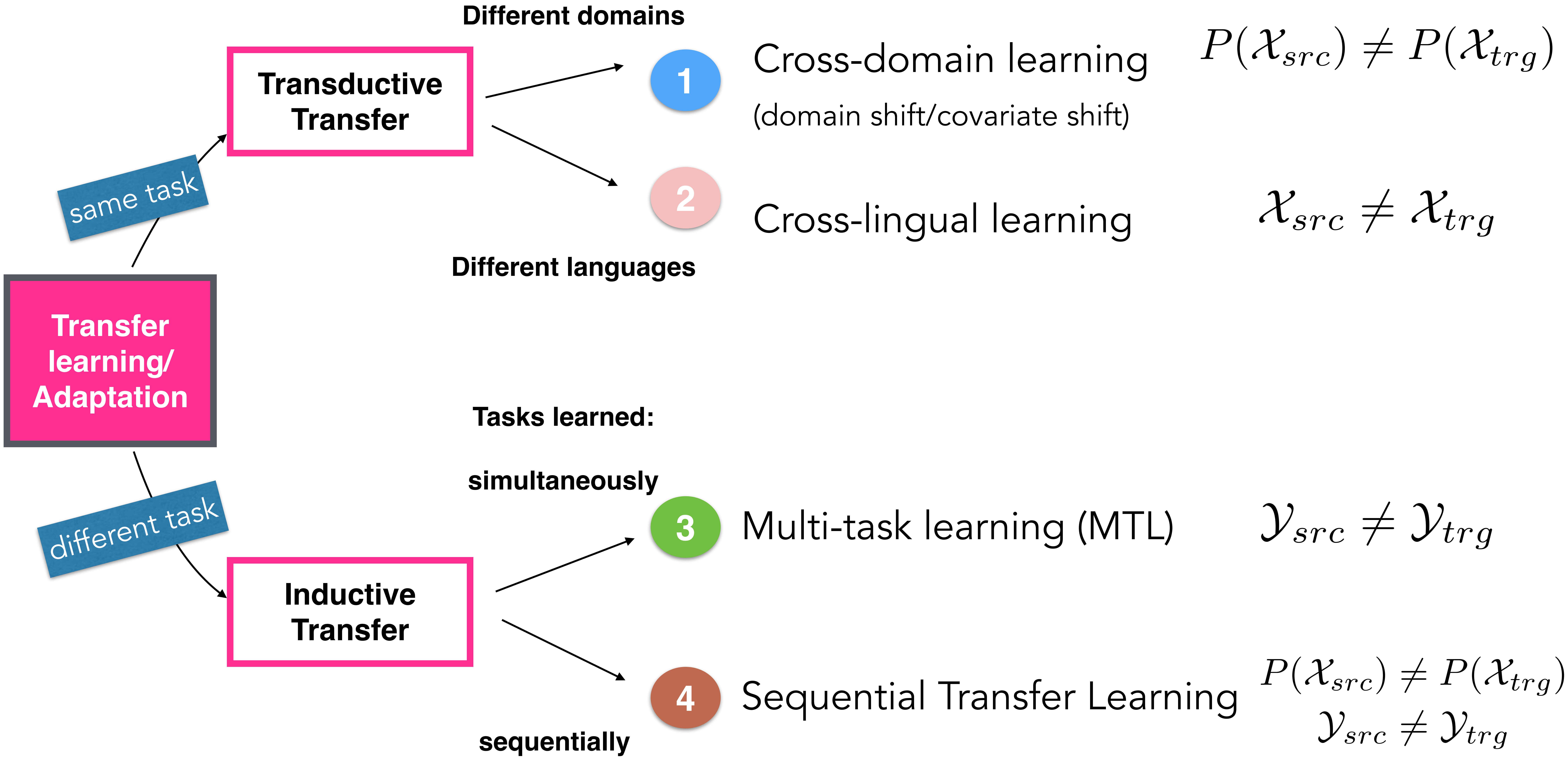
Three views on Transfer Learning

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

Types of Transfer Learning - View 1/3: Kind of tasks, data, timing



Transfer learning/ Adaptation

same task

Transductive Transfer

Different domains

1 Cross-domain learning $P(\mathcal{X}_{src}) \neq P(\mathcal{X}_{trg})$

(domain shift/covariate shift)

2 Cross-lingual learning $\mathcal{X}_{src} \neq \mathcal{X}_{trg}$

Different languages

different task

Inductive Transfer

Tasks learned:

simultaneously

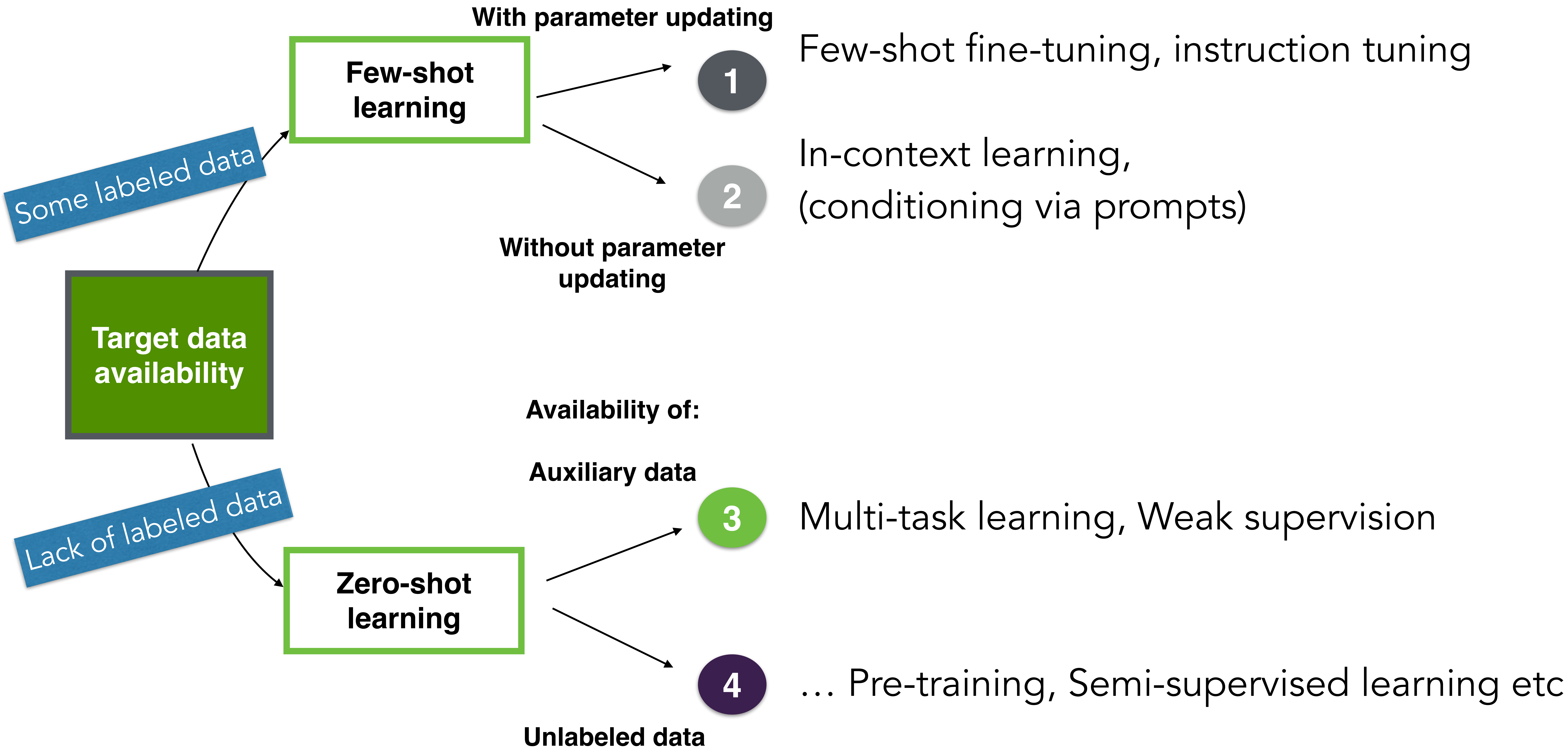
3 Multi-task learning (MTL) $\mathcal{Y}_{src} \neq \mathcal{Y}_{trg}$

4 Sequential Transfer Learning $P(\mathcal{X}_{src}) \neq P(\mathcal{X}_{trg})$ $\mathcal{Y}_{src} \neq \mathcal{Y}_{trg}$

sequentially

Adapted from Pan et al., (2009) & Ruder (2019)

Types of Transfer Learning - View 2/3: Availability of resources



Target data availability

Some labeled data

Few-shot learning

With parameter updating

1 Few-shot fine-tuning, instruction tuning

Without parameter updating

2 In-context learning, (conditioning via prompts)

Lack of labeled data

Zero-shot learning

Availability of:

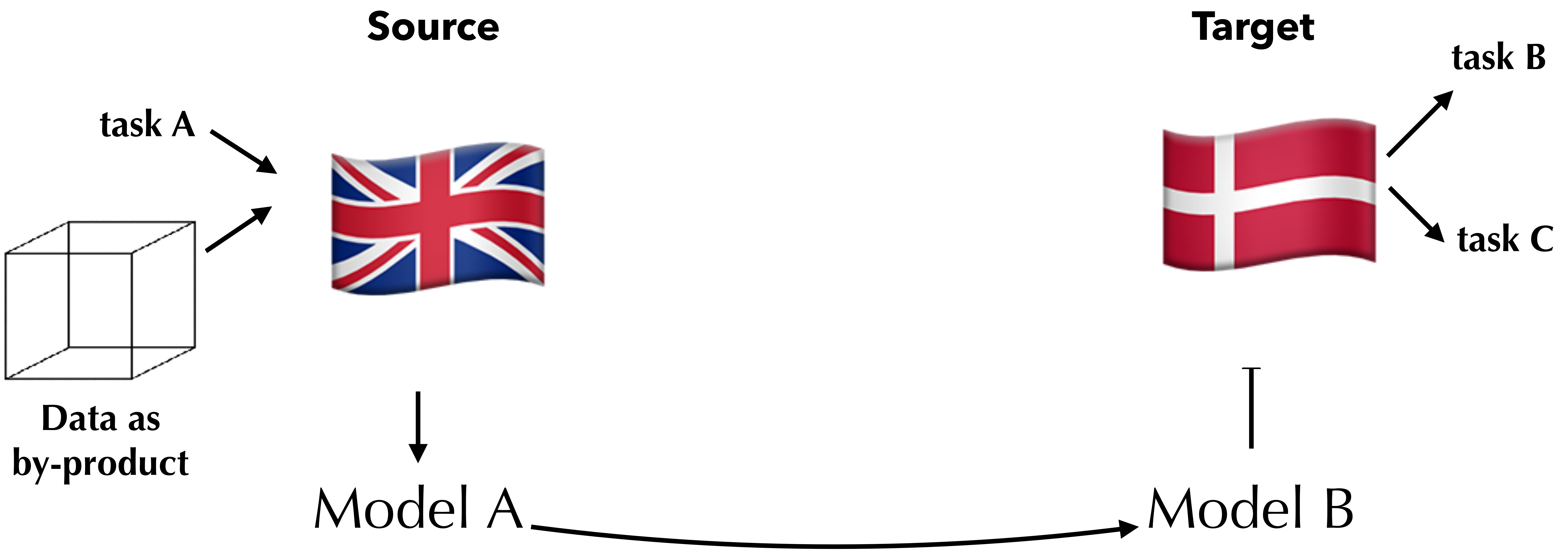
Auxiliary data

3 Multi-task learning, Weak supervision

Unlabeled data

4 … Pre-training, Semi-supervised learning etc

Types of Transfer Learning - View 3/3: How to cross the gulf



CROSS-DOMAIN

CROSS-LINGUAL

MULTI-TASK

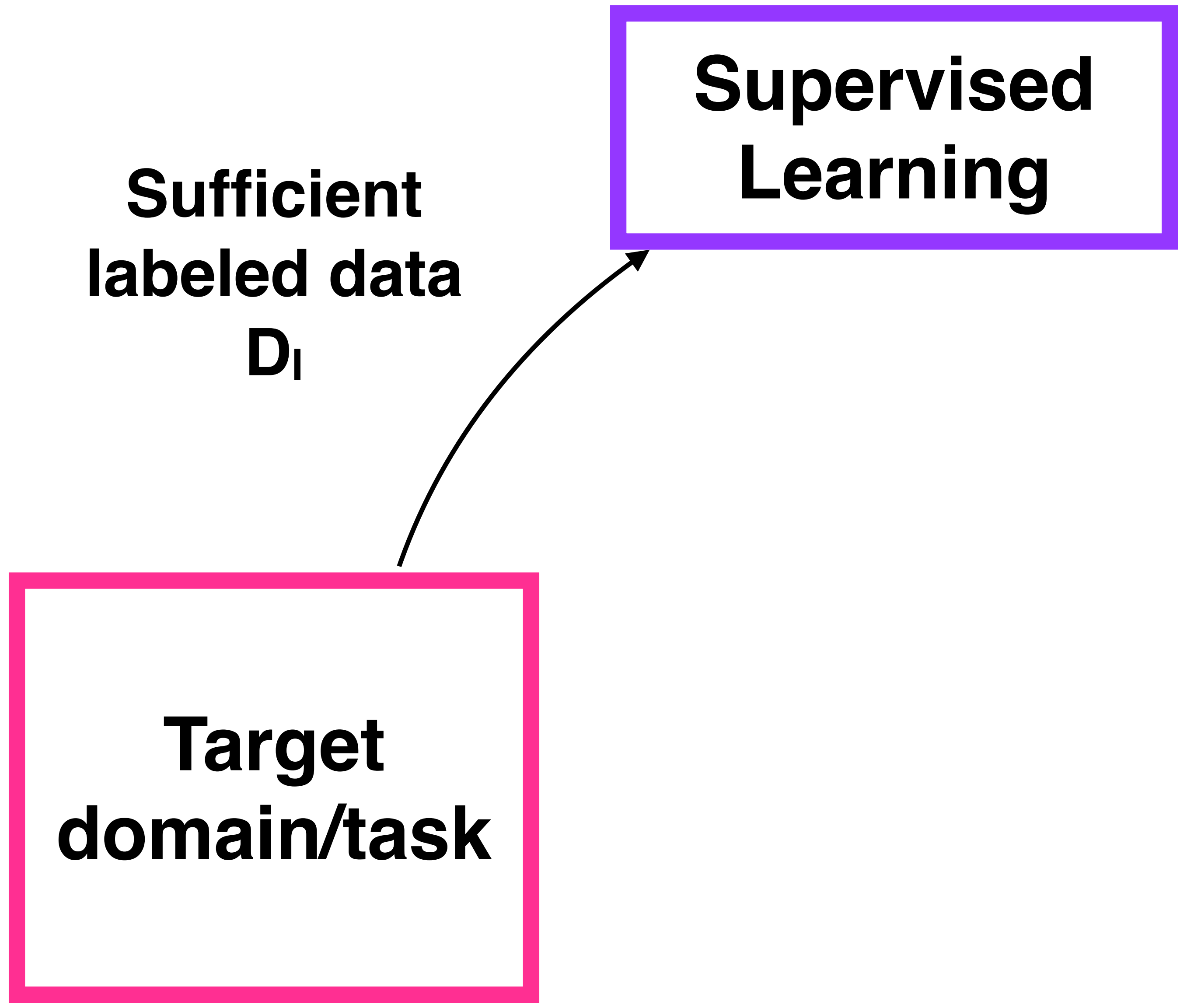
FORTUITOUS/INCIDENTAL SUPERVISION

Effective

Efficient

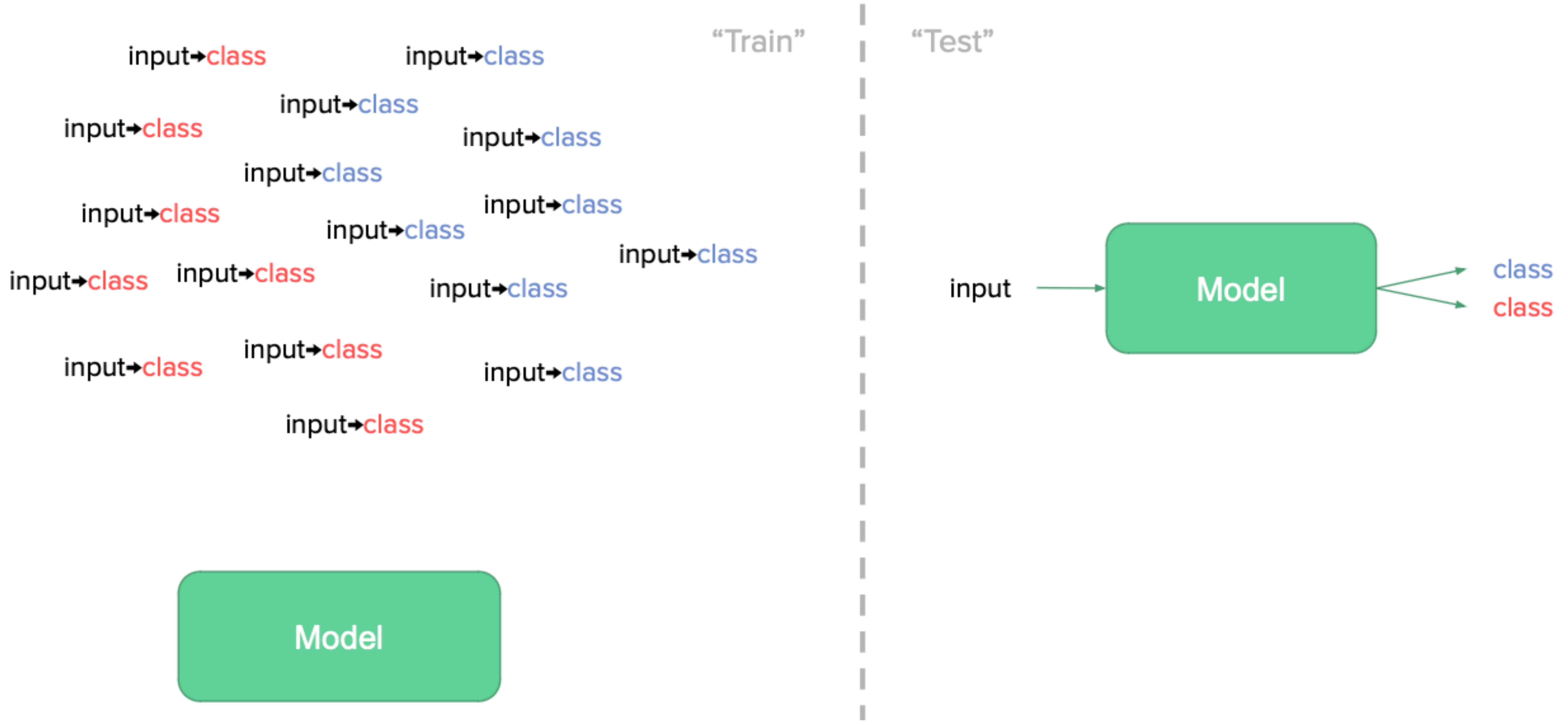
TL is finding smart ways to re-use {knowledge, data, models…} for the purpose of generalisation

Related learning paradigms

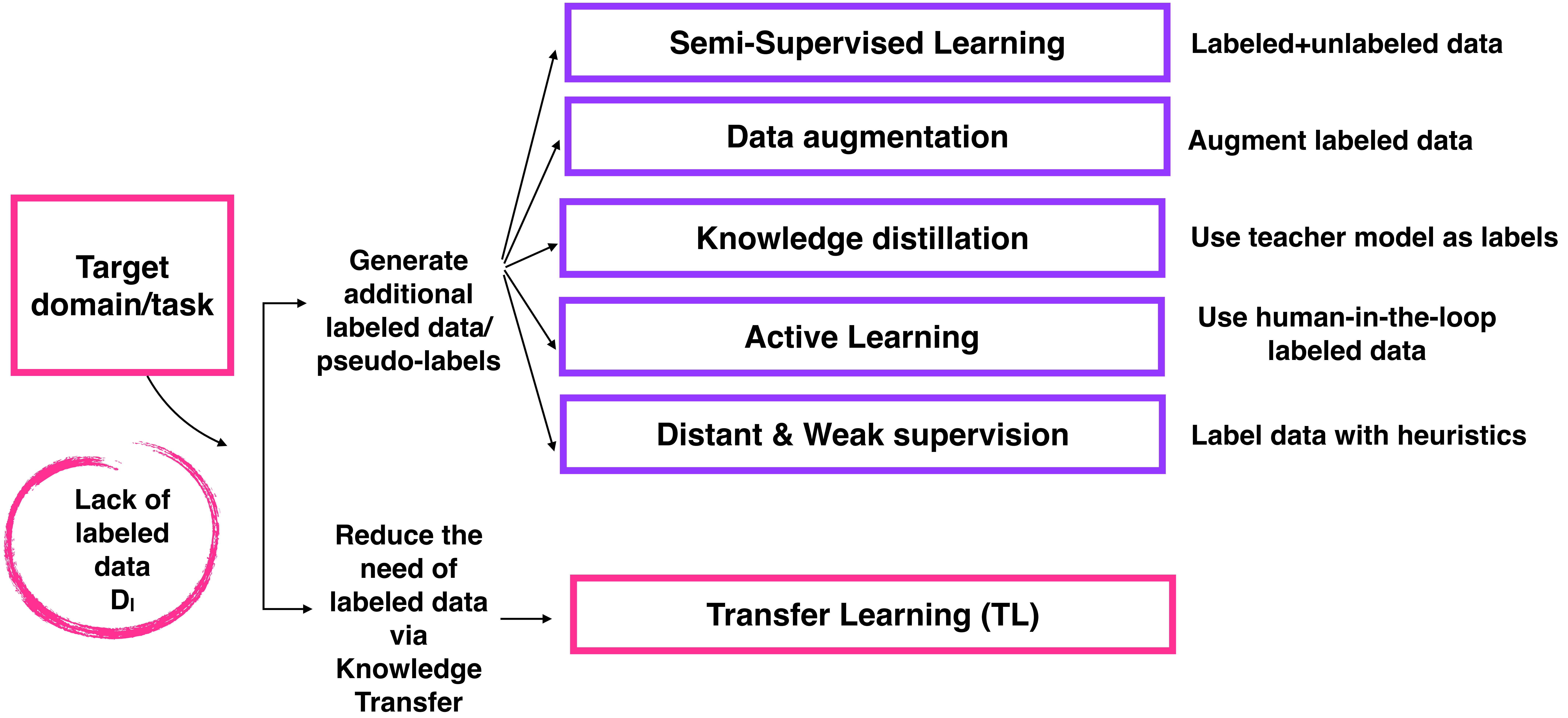


Supervised Learning

Slide by Beltagy et al., ACL 2022 tutorial



Related learning paradigms



Target domain/task

Lack of labeled data D_l

Generate additional labeled data/ pseudo-labels

Semi-Supervised Learning

Labeled+unlabeled data

Data augmentation

Augment labeled data

Knowledge distillation

Use teacher model as labels

Active Learning

Use human-in-the-loop labeled data

Distant & Weak supervision

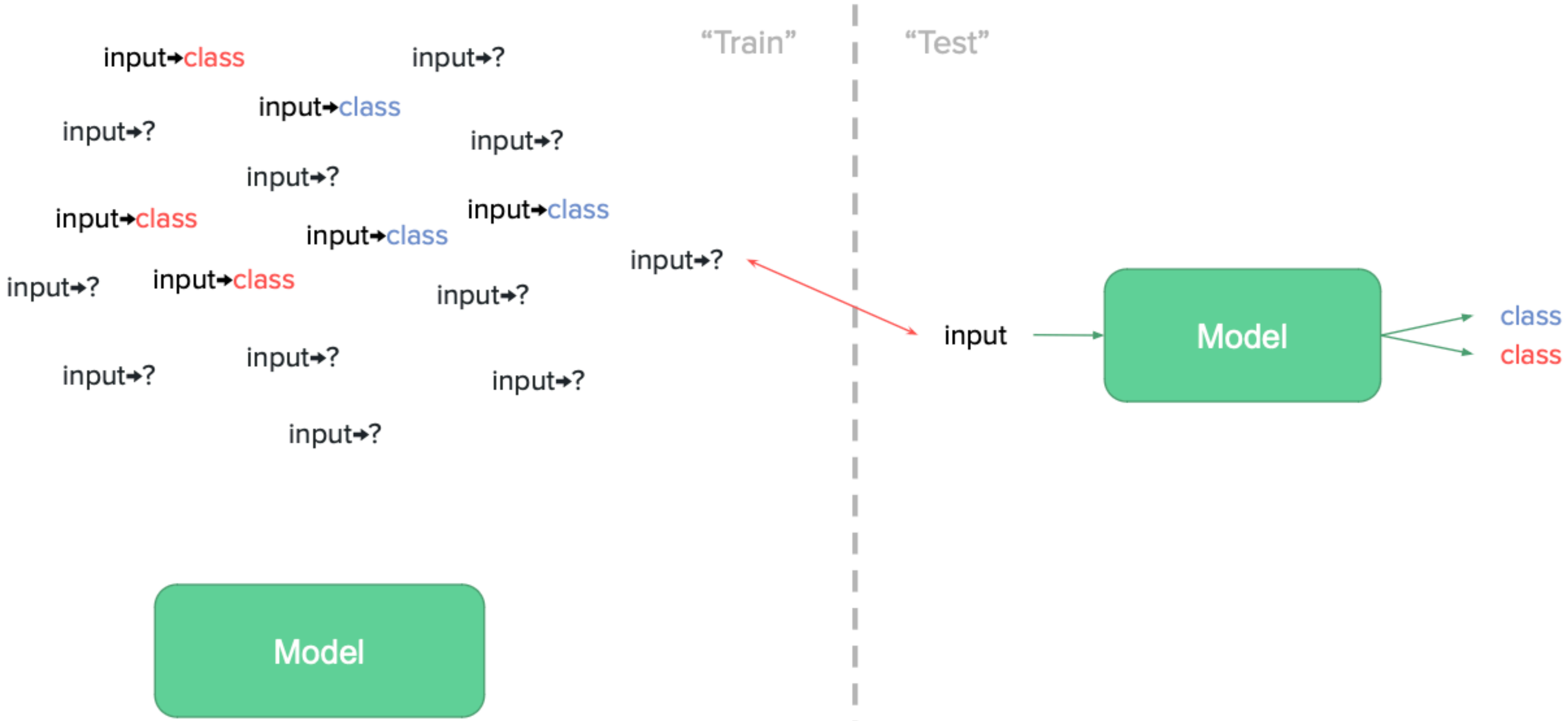
Label data with heuristics

Reduce the need of labeled data via Knowledge Transfer

Transfer Learning (TL)

Semi-Supervised Learning

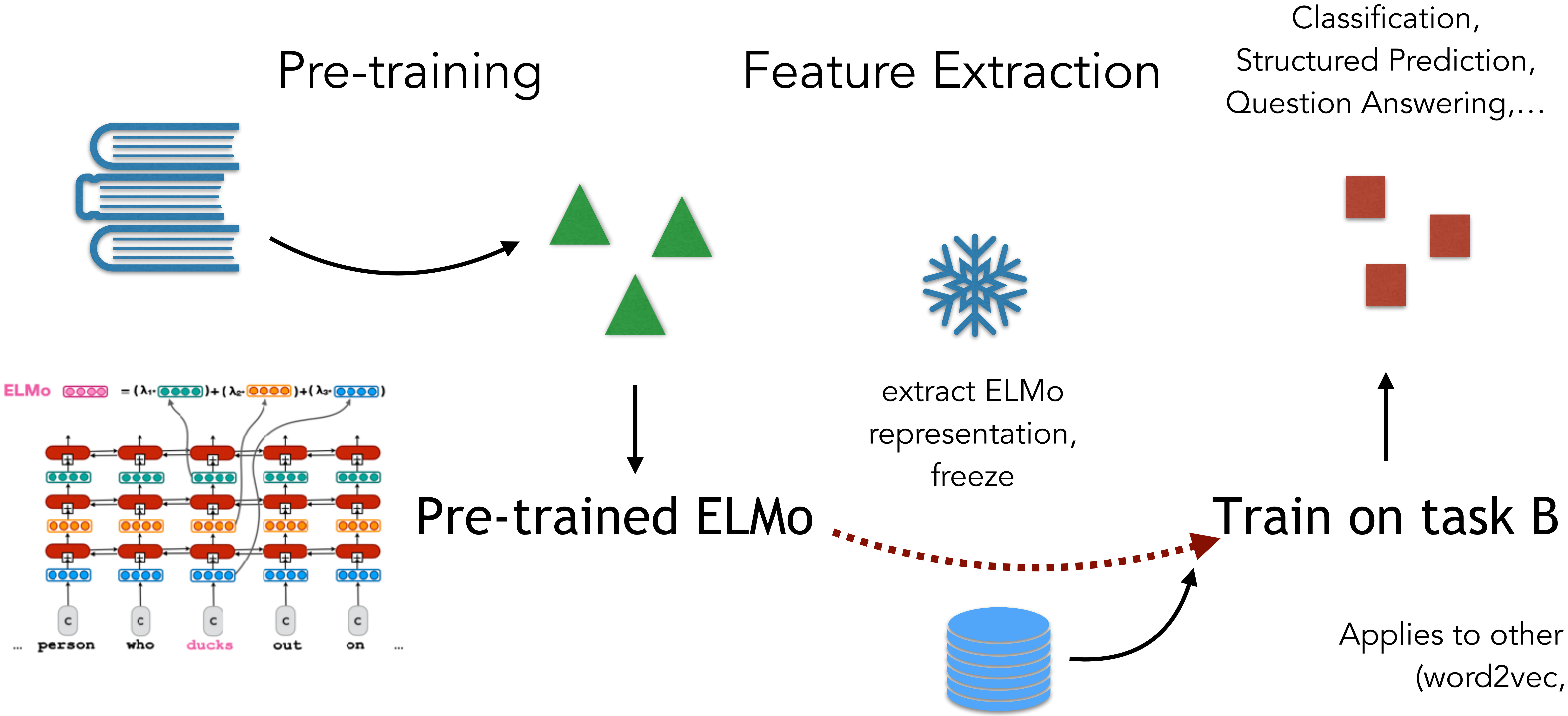
Slide by Beltagy et al., ACL 2022 tutorial



Sequential Transfer Learning - Approaches (incl. a short history)

Transfer Learning (TL) via pre-training I: Feature extraction (e.g. ELMo)

Peters et al. (2018)



Pre-training

Feature Extraction

Classification,
Structured Prediction,
Question Answering,…

Pre-trained ELMo

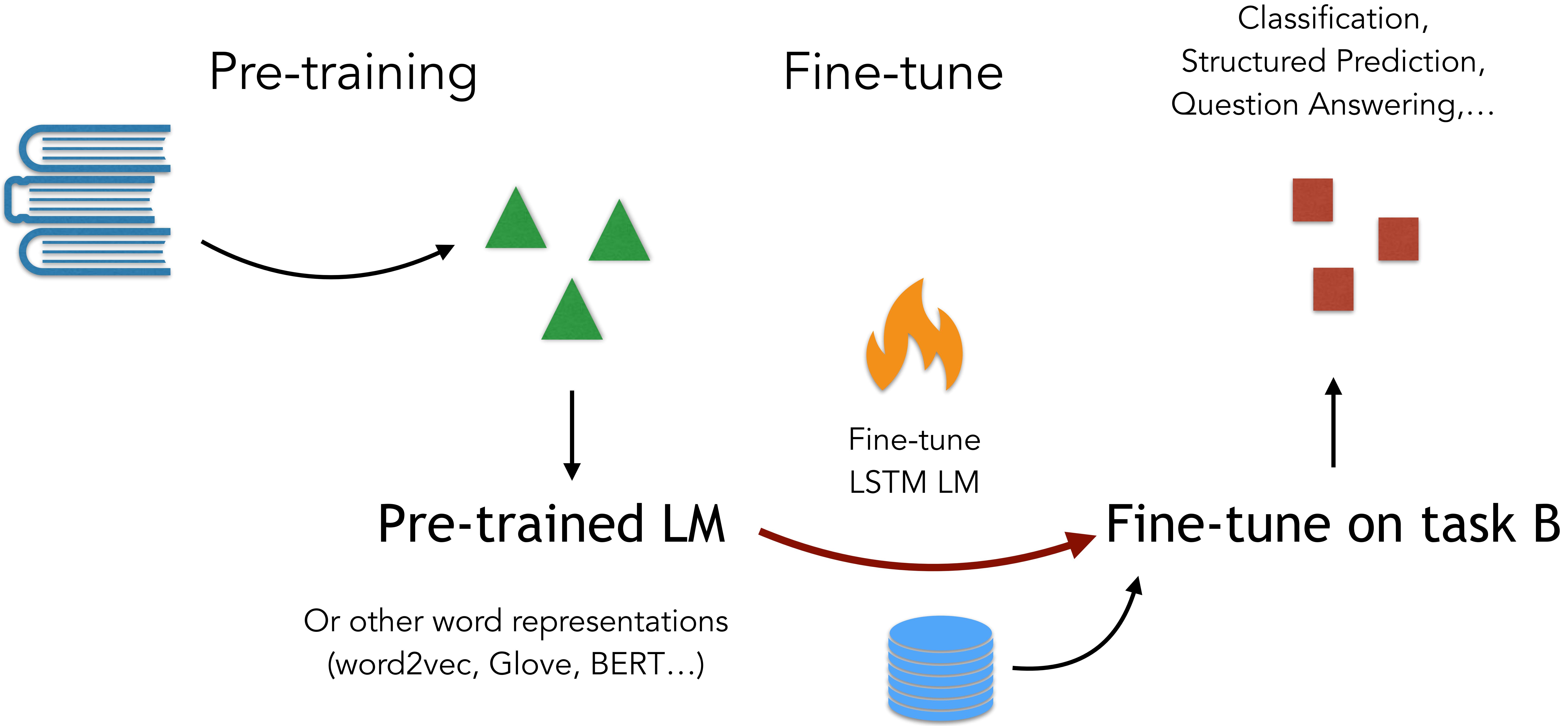
extract ELMo representation, freeze

Train on task B

Applies to other word representations
(word2vec, Glove, BERT…)

Transfer Learning (TL) via pre-training II: Fine Tuning (e.g. ULMFiT, BERT)

Howard & Ruder (2018); Peters et al., (2018)



Pre-training

Fine-tune

Classification,
Structured Prediction,
Question Answering,…

Fine-tune
LSTM LM

Pre-trained LM

Fine-tune on task B

Or other word representations
(word2vec, Glove, BERT…)

Sequential Transfer Learning (TL) - Problems and Solutions

Howard & Ruder (2018), Radford et al. (2018)

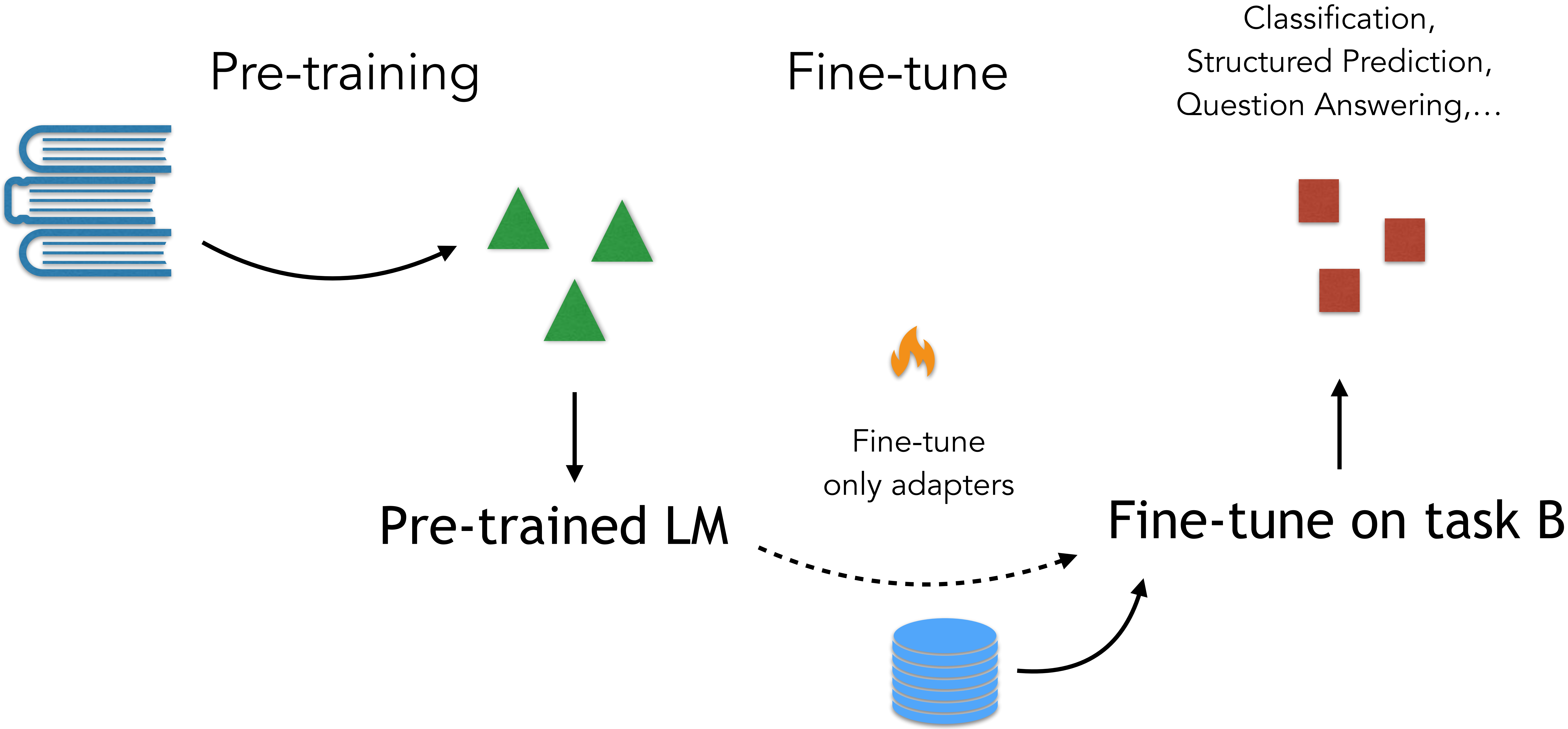
- A common problem of fine-tuning is that retraining the model can mean to loose information about the general pre-training data (**"catastrophic forgetting"**)
  - To address this, in **gradual unfreezing** the model will be trained in steps, starting by the last layer. So all layers are first frozen except the last one. In every step an additional layer is "unfrozen"
- Learning a large model can be **unstable**
  - First increase learning rate, then decrease it (**slanted triangular learning rate**)
- From biLSTMs to **transformers**
  - While first models use LSTMs (Howard & Ruder, 2018), GPT (Radford et al., 2018) used a transformer architecture in early GPT

Full-fine tuning: Further Issues

- Standard fine-tuning updates all LM parameters
  - Prone to overfitting and catastrophic forgetting
  - Practically may be too expensive
- A solution:
  - Modularity - adapters

Full-fine tuning limitations. Solution: Adapters

(Houlsby et al., 2019; Pfeiffer et al., 2020)



Pre-training

Fine-tune

Classification,
Structured Prediction,
Question Answering,…

Fine-tune
only adapters

Pre-trained LM

Fine-tune on task B

Adapters: Modular Adaptation

(Houlsby et al., 2019; Pfeiffer et al., 2020)

- **Adapters**: small modules inserted into transformer layers for efficient fine-tuning

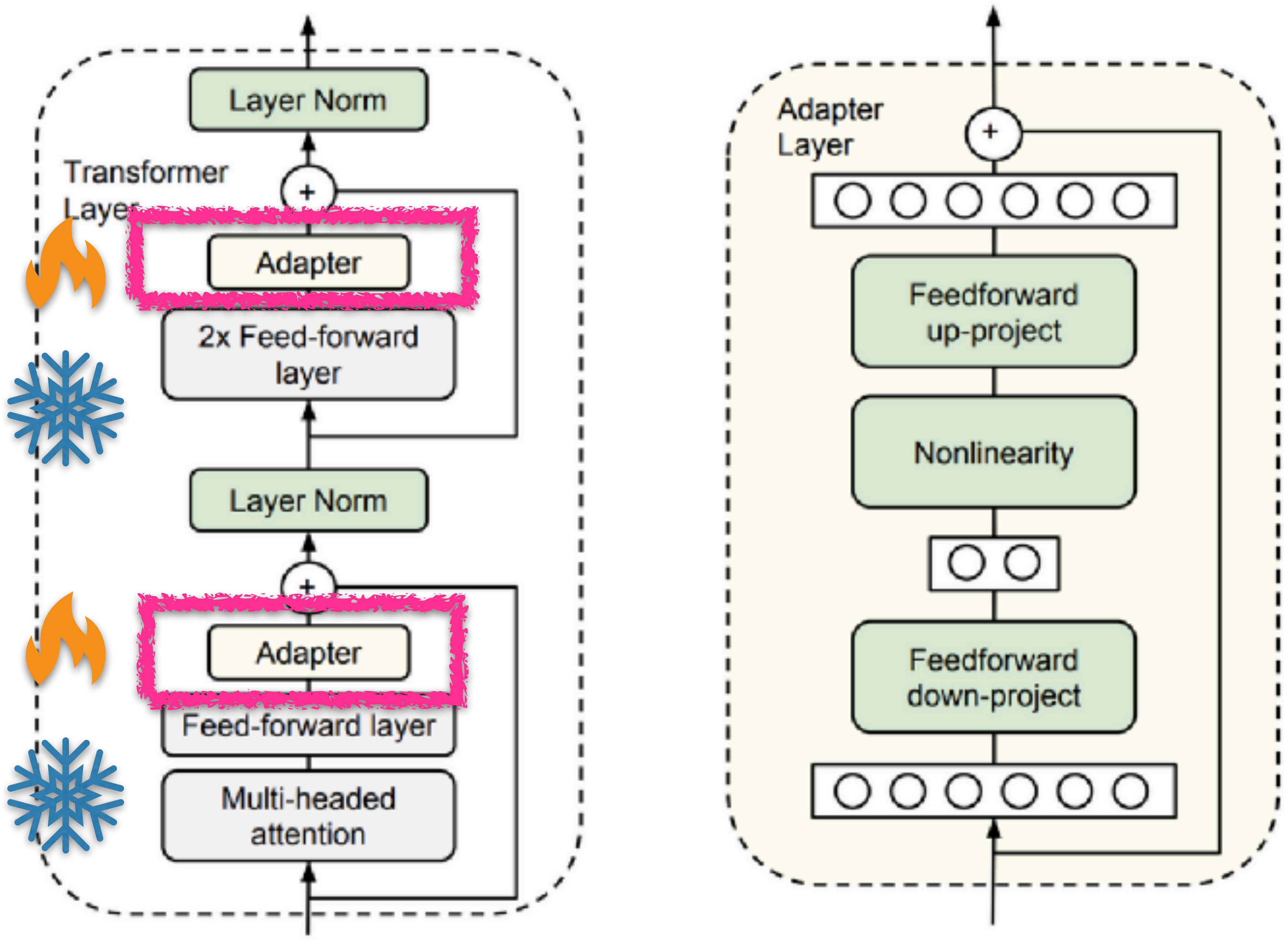
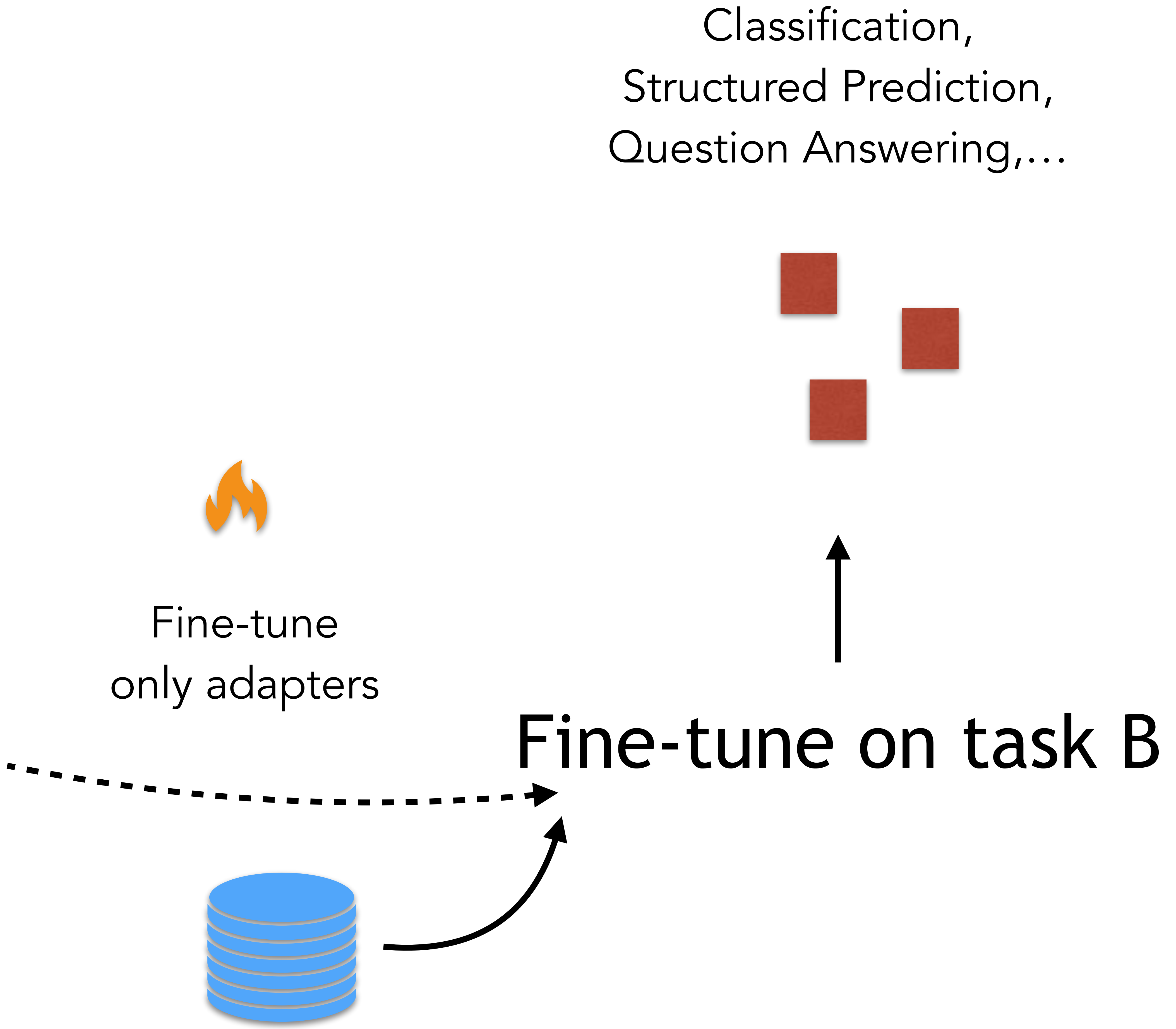


Figure from Houlsby et al., 2019



Classification,
Structured Prediction,
Question Answering,…

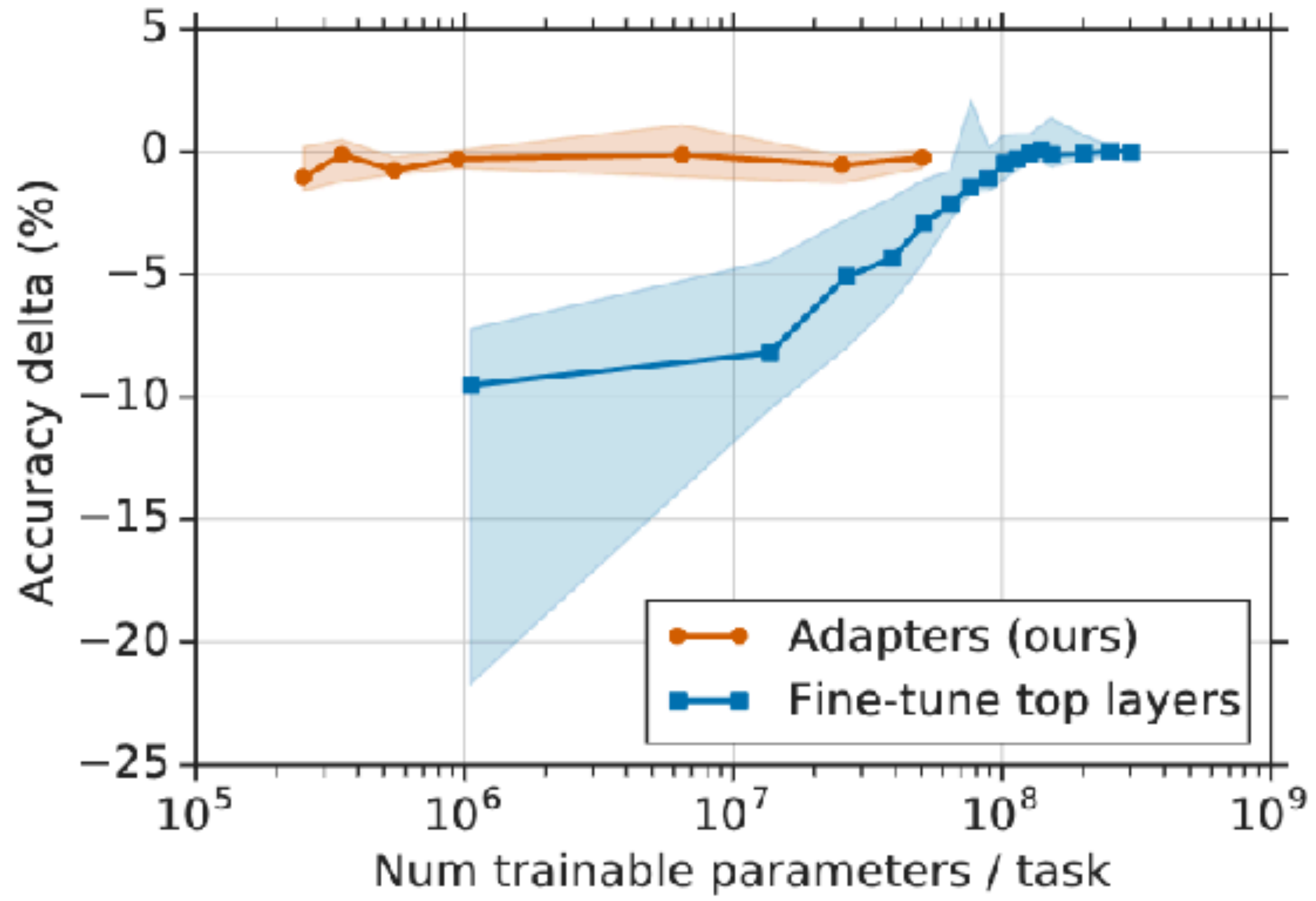
Fine-tune
only adapters

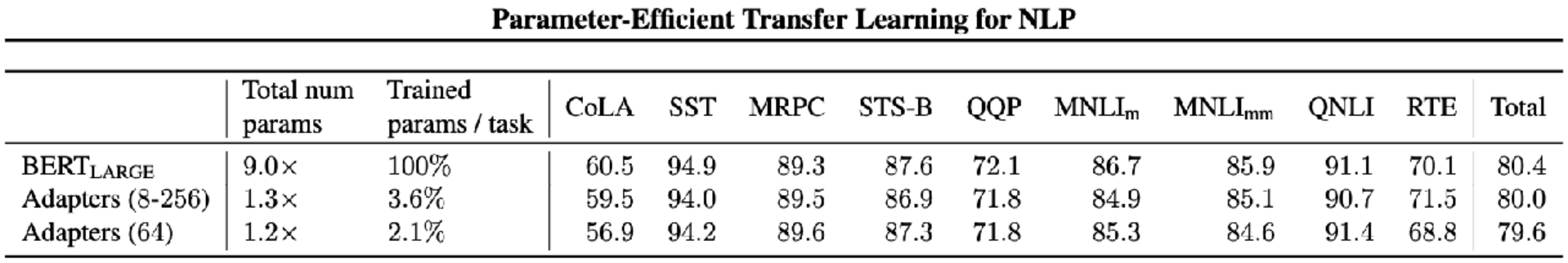
Fine-tune on task B

Adapters: Modular Adaptation

(Houlsby et al., 2019; Pfeiffer et al., 2020; Üstün et al., 2022)

- Adapters learn transformations to adapt a base model to a target task
- Encapsulate knowledge in a modular way
- Do adapters work?




Parameter-Efficient Transfer Learning for NLP

	Total num params	Trained params / task	CoLA	SST	MRPC	STS-B	QQP	MNLI$_m$	MNLI$_{mm}$	QNLI	RTE	Total
BERT$_{LARGE}$	9.0×	100%	60.5	94.9	89.3	87.6	72.1	86.7	85.9	91.1	70.1	80.4
Adapters (8-256)	1.3×	3.6%	59.5	94.0	89.5	86.9	71.8	84.9	85.1	90.7	71.5	80.0
Adapters (64)	1.2×	2.1%	56.9	94.2	89.6	87.3	71.8	85.3	84.6	91.4	68.8	79.6

- Adapters are trained separately. Limitation: **No sharing** between different tasks

A snapshot of NLP history - Act in 4 Epochs

Symbolic Processing

from hand-crafted rules to ML

Epoch 1

Statistical NLP

representations

Epoch 2

Deep Learning for NLP

ducks:

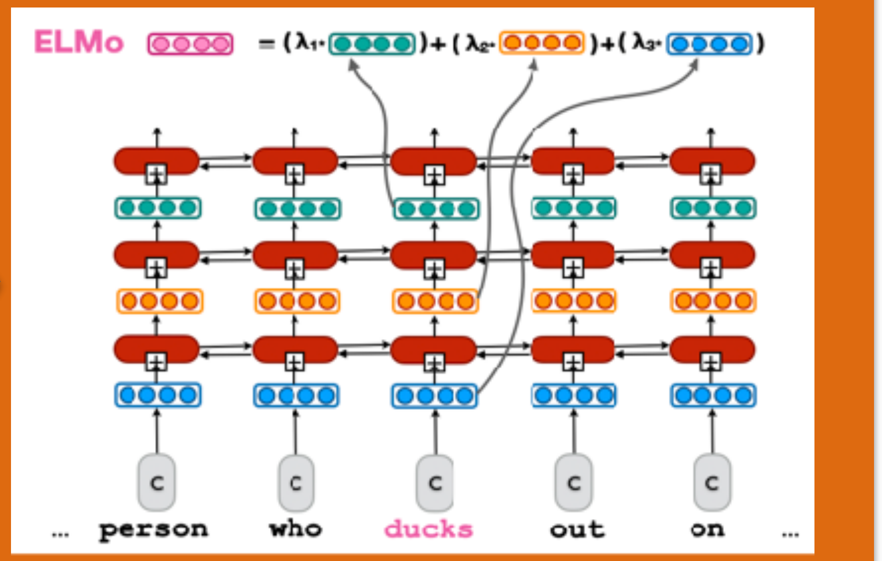
can:

dense representations
& neural networks

Epoch 3

contextualised representations

Large Pre-trained LMs



e(ducks) != e(ducks)

Epoch 4

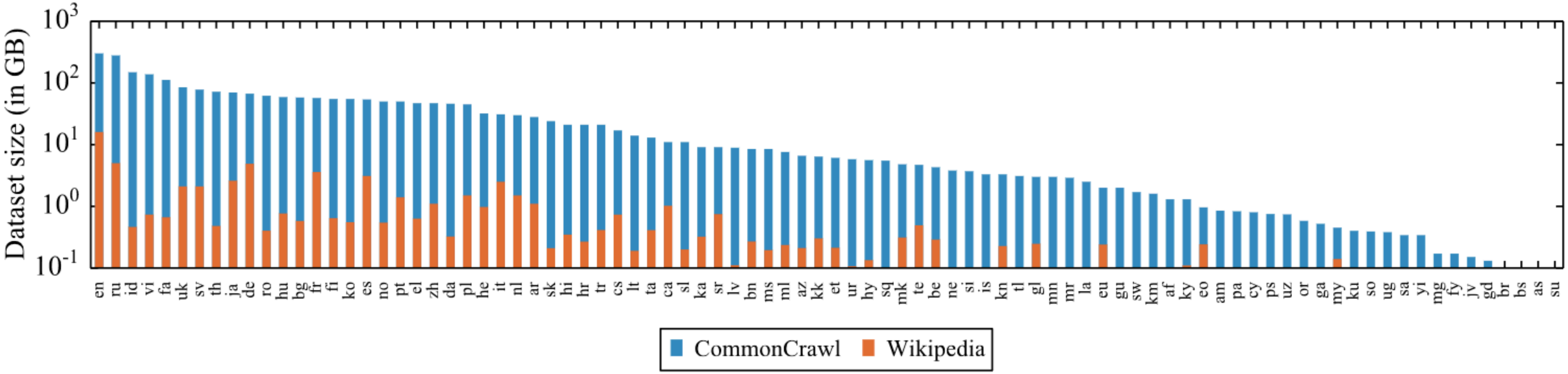
1980s

2018

Are Language Models truly universal?

Languages

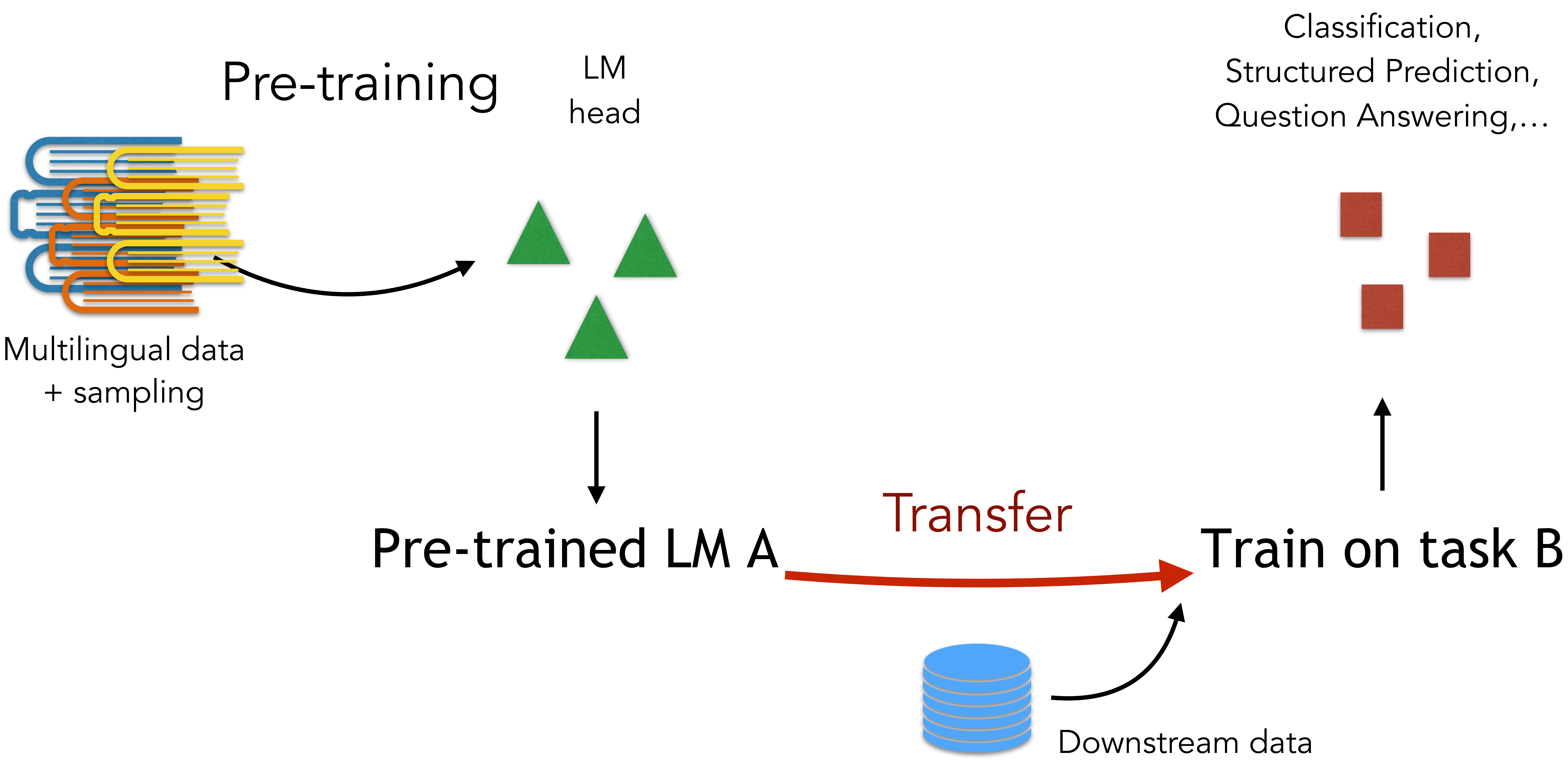
Motivation



Dataset size (in GB)
10^3
10^2
10^1
10^0
10^{-1}
en
ru
id
vi
fa
uk
sv
th
ja
de
ro
hu
bg
fr
fi
ko
es
no
pt
el
zh
da
pl
he
it
nl
ar
sk
hi
hr
tr
cs
lt
ta
ca
sl
ka
sr
lv
bn
ms
ml
az
kk
et
ur
hy
sq
mk
te
be
ne
si
is
kn
tl
gl
mn
mr
la
eu
gu
sw
km
af
ky
eo
am
pa
cy
ps
uz
or
ga
my
ku
so
ug
sa
yi
mg
fy
jv
gd
br
bs
as
su
🟦 CommonCrawl
🟧 Wikipedia

Multilingual Language Models (e.g., mBERT, XLM-R)

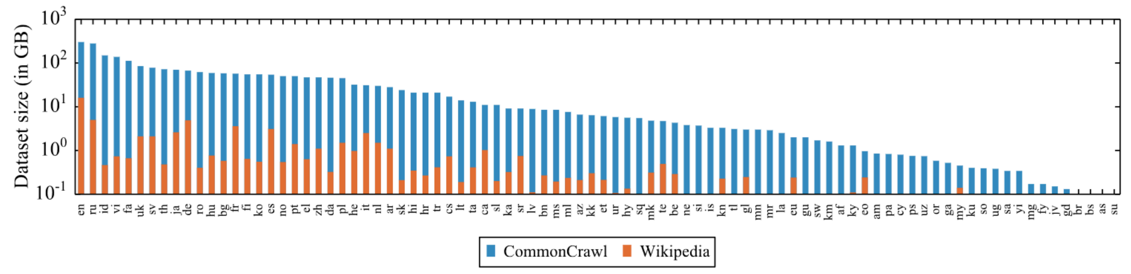
The easiest way to do transfer learning across languages is via the representations

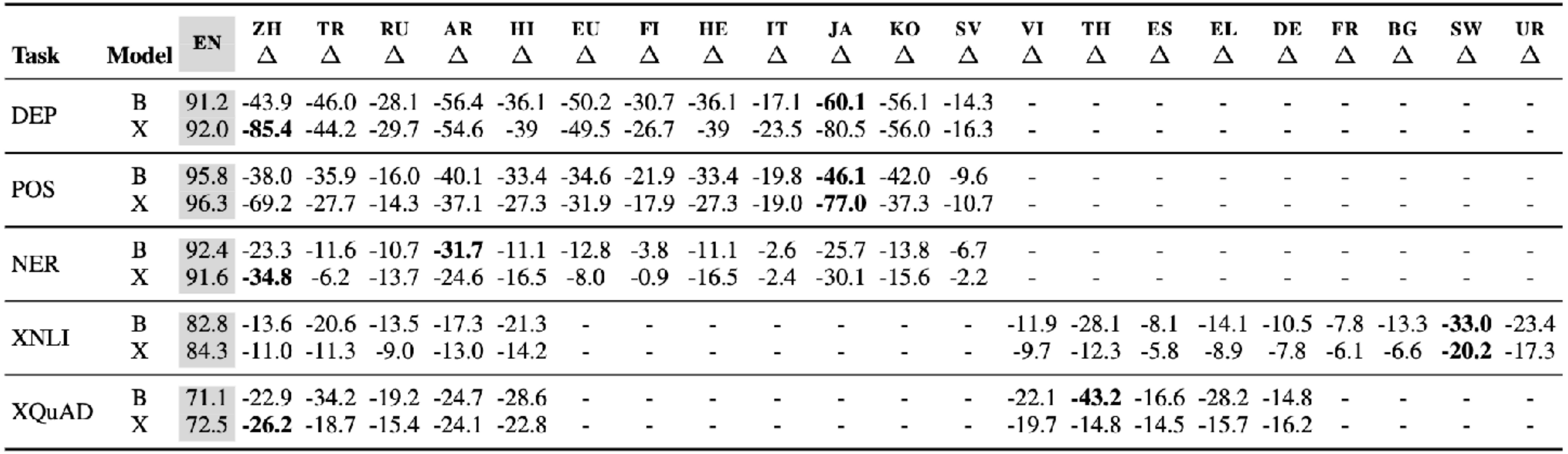


On the limitations of zero-shot TL with Multilingual Transformers

Lauscher et al., 2020; Conneau et al., 2020

- Zero-shot performs poorly to distant languages *and* languages with smaller pre-training corpus sizes





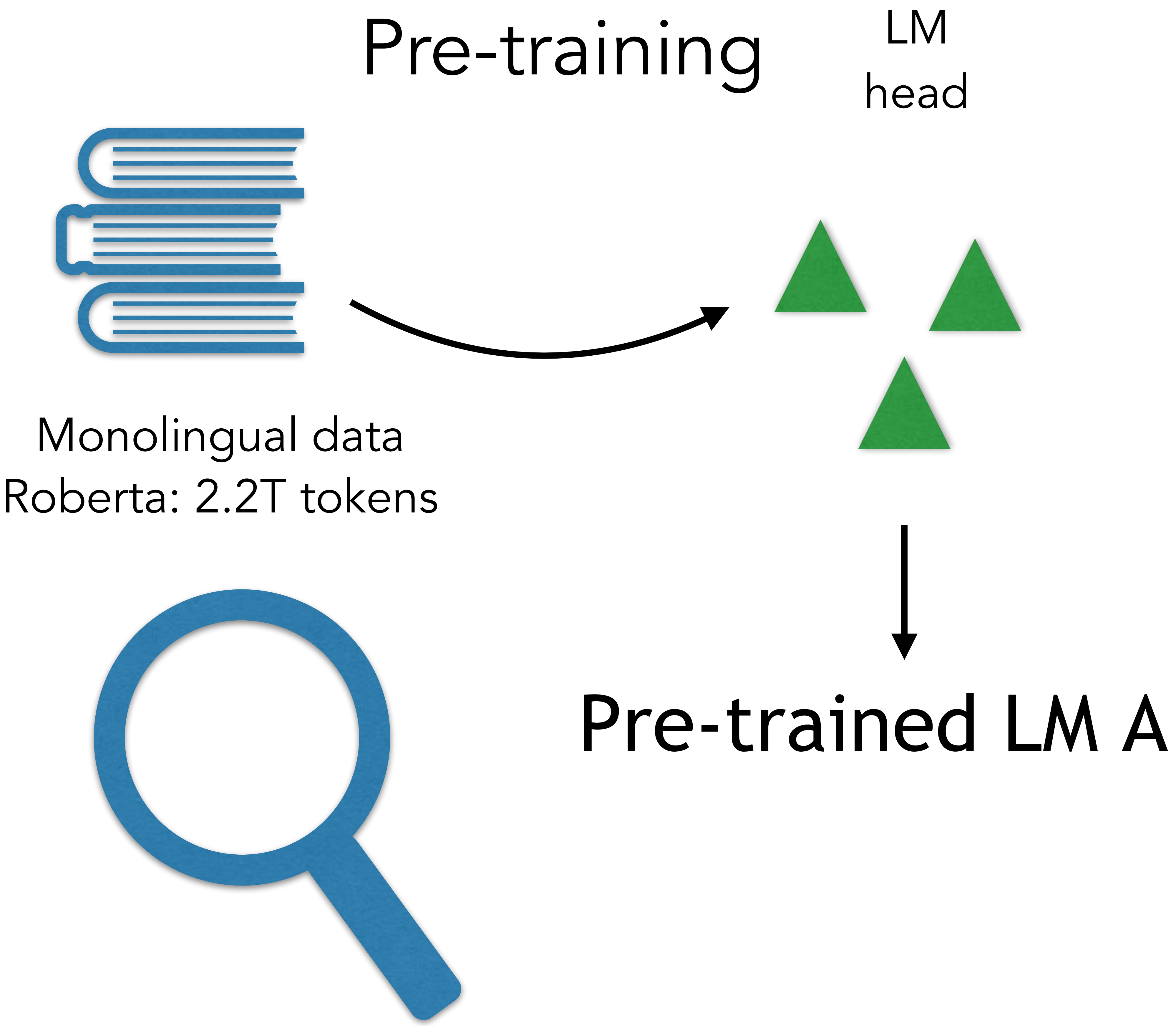
Task	Model	EN	ZH Δ	TR Δ	RU Δ	AR Δ	HI Δ	EU Δ	FI Δ	HE Δ	IT Δ	JA Δ	KO Δ	SV Δ	VI Δ	TH Δ	ES Δ	EL Δ	DE Δ	FR Δ	BG Δ	SW Δ	UR Δ
DEP	B	91.2	-43.9	-46.0	-28.1	-56.4	-36.1	-50.2	-30.7	-36.1	-17.1	**-60.1**	-56.1	-14.3	-	-	-	-	-	-	-	-	-
	X	92.0	**-85.4**	-44.2	-29.7	-54.6	-39	-49.5	-26.7	-39	-23.5	-80.5	-56.0	-16.3	-	-	-	-	-	-	-	-	-
POS	B	95.8	-38.0	-35.9	-16.0	-40.1	-33.4	-34.6	-21.9	-33.4	-19.8	**-46.1**	-42.0	-9.6	-	-	-	-	-	-	-	-	-
	X	96.3	-69.2	-27.7	-14.3	-37.1	-27.3	-31.9	-17.9	-27.3	-19.0	**-77.0**	-37.3	-10.7	-	-	-	-	-	-	-	-	-
NER	B	92.4	-23.3	-11.6	-10.7	**-31.7**	-11.1	-12.8	-3.8	-11.1	-2.6	-25.7	-13.8	-6.7	-	-	-	-	-	-	-	-	-
	X	91.6	**-34.8**	-6.2	-13.7	-24.6	-16.5	-8.0	-0.9	-16.5	-2.4	-30.1	-15.6	-2.2	-	-	-	-	-	-	-	-	-
XNLI	B	82.8	-13.6	-20.6	-13.5	-17.3	-21.3	-	-	-	-	-	-	-	-11.9	-28.1	-8.1	-14.1	-10.5	-7.8	-13.3	**-33.0**	-23.4
	X	84.3	-11.0	-11.3	-9.0	-13.0	-14.2	-	-	-	-	-	-	-	-9.7	-12.3	-5.8	-8.9	-7.8	-6.1	-6.6	**-20.2**	-17.3
XQuAD	B	71.1	-22.9	-34.2	-19.2	-24.7	-28.6	-	-	-	-	-	-	-	-22.1	**-43.2**	-16.6	-28.2	-14.8	-	-	-	-
	X	72.5	**-26.2**	-18.7	-15.4	-24.1	-22.8	-	-	-	-	-	-	-	-19.7	-14.8	-14.5	-15.7	-16.2	-	-	-	-

Table 1: Zero-shot cross-lingual transfer performance on five tasks (DEP, POS, NER, XNLI, and XQuAD) with mBERT (B) and XLM-R (X). We show the monolingual EN performance and report drops in performance relative to EN for all target languages. Numbers in bold indicate the largest zero-shot performance drops for each task.

Domains

Large Language Models and Pre-training Domains

What does training on trillions of tokens afford us in terms of generalisation even within English? (Gururangan et al., 2020)



Pre-training

LM head

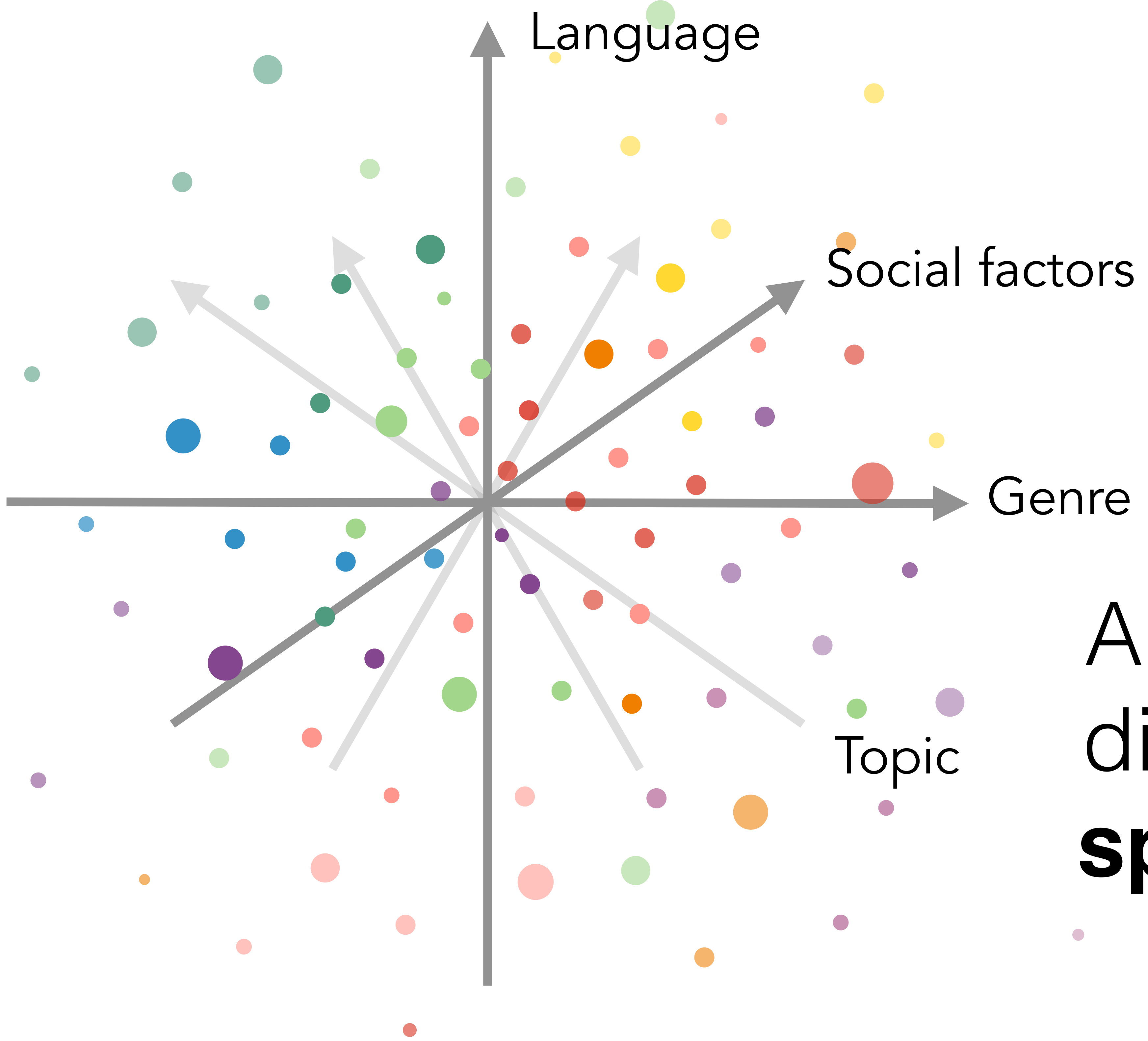
Monolingual data
Roberta: 2.2T tokens

Pre-trained LM A

Large Language Models and Pre-training Domains

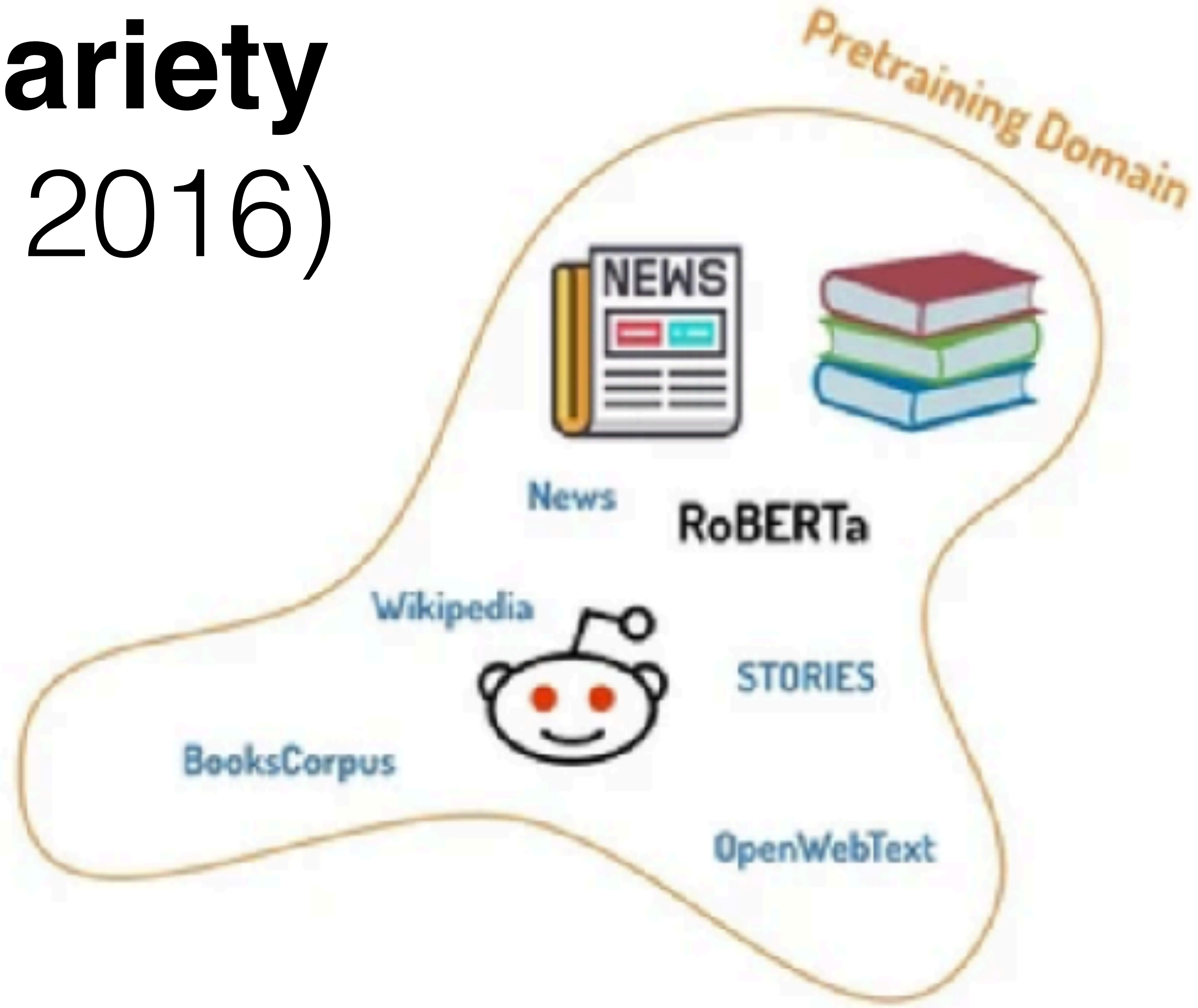
What does training on trillions of tokens afford us in terms of generalisation even within English? (Gururangan et al., 2020)

Monolingual data
Roberta: 2.2T tokens



What is in a domain?

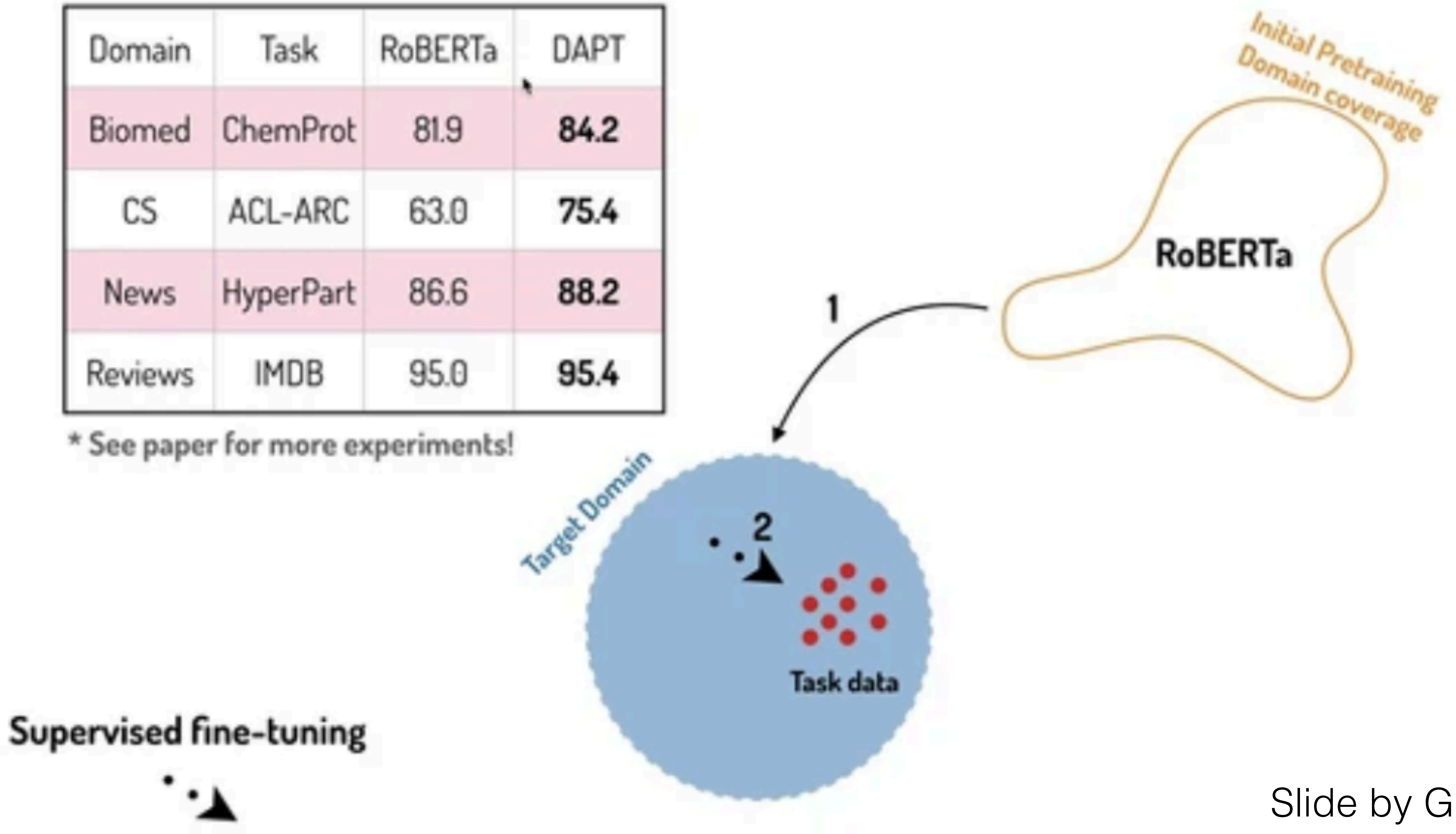
A manifold in a high-dimensional “**variety space**” (Plank, 2016)



Don’t Stop Pre-Training: Adapt Language Models to Domains and Tasks

(Gururangan et al., 2020)

- **Continuous pre-training** on target domain data helps (Domain-adaptive pre-training; DAPT)



Domain	Task	RoBERTa	DAPT
Biomed	ChemProt	81.9	**84.2**
CS	ACL-ARC	63.0	**75.4**
News	HyperPart	86.6	**88.2**
Reviews	IMDB	95.0	**95.4**

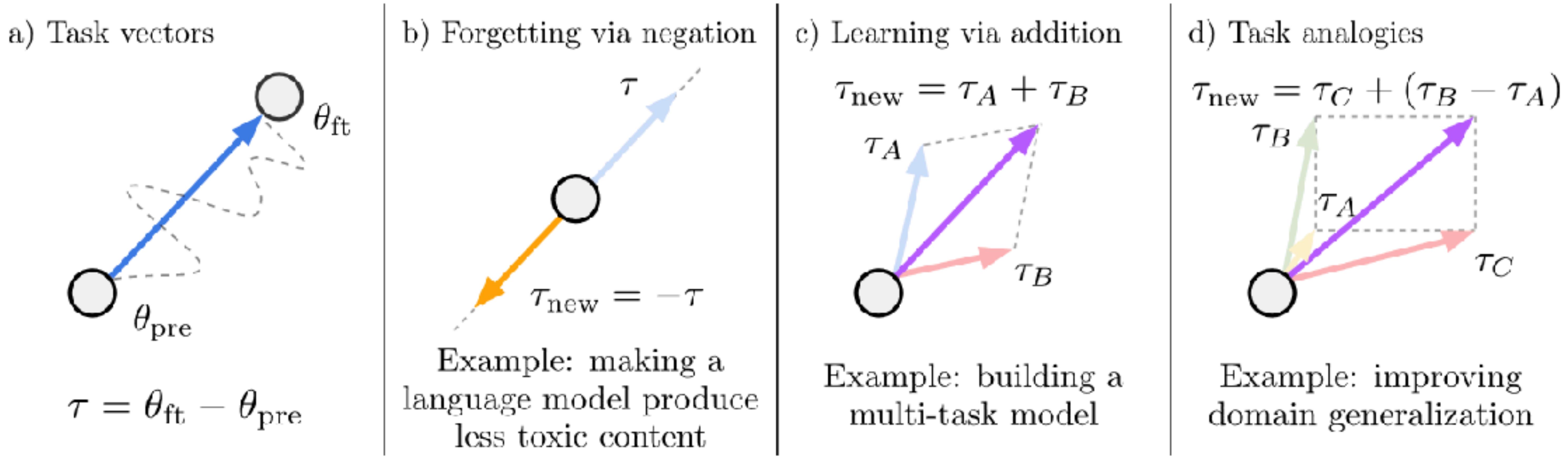
* See paper for more experiments!

Slide by Guruangan et al.

Related recent work: Task Vectors - aka Post-hoc model intervention

(Ilharco, Riberio, Wortsmann, Gururangan et al., 2023)

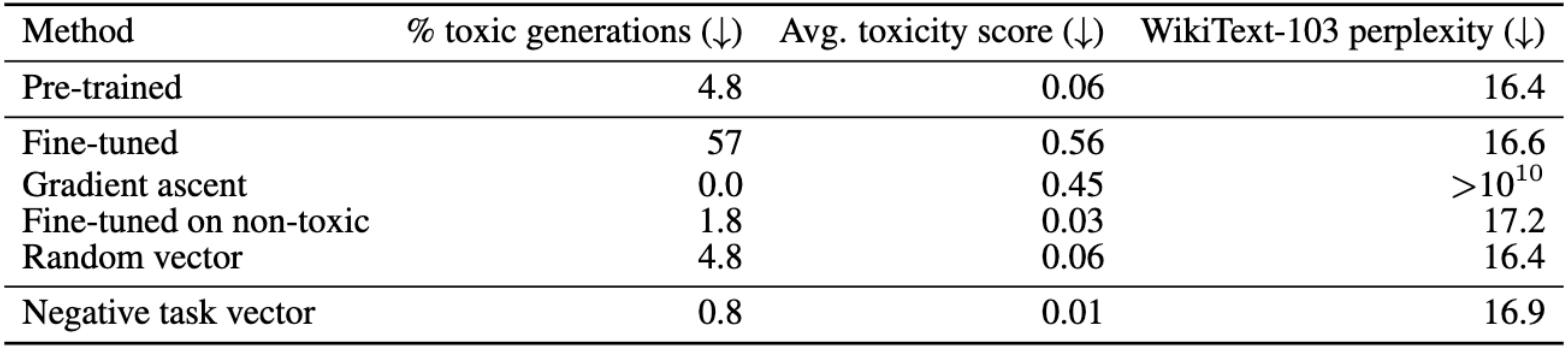
- Motivation: pre-trained models are a commonly used backbone
- In practice, we often want to *edit* the models after pre-training to improve on downstream tasks
- Task vector: difference vector of weights of a model fine-tuned on a task, minus pre-trained weights
  - Allows task arithmetics (negation for forgetting)


a) Task vectors
θ_{ft}
θ_{pre}
$\tau = \theta_{\text{ft}} - \theta_{\text{pre}}$
b) Forgetting via negation
τ
$\tau_{\text{new}} = -\tau$
Example: making a language model produce less toxic content
c) Learning via addition
$\tau_{\text{new}} = \tau_A + \tau_B$
τ_A
τ_B
Example: building a multi-task model
d) Task analogies
$\tau_{\text{new}} = \tau_C + (\tau_B - \tau_A)$
τ_B
τ_A
τ_C
Example: improving domain generalization

Related recent work: Task Vectors - aka Post-hoc model intervention

(Ilharco, Riberio, Wortsmann, Gururangan et al., 2023)

- Example: Making Language Models less toxic



Method	% toxic generations (↓)	Avg. toxicity score (↓)	WikiText-103 perplexity (↓)
Pre-trained	4.8	0.06	16.4
Fine-tuned	57	0.56	16.6
Gradient ascent	0.0	0.45	$>10^{10}$
Fine-tuned on non-toxic	1.8	0.03	17.2
Random vector	4.8	0.06	16.4
Negative task vector	0.8	0.01	16.9

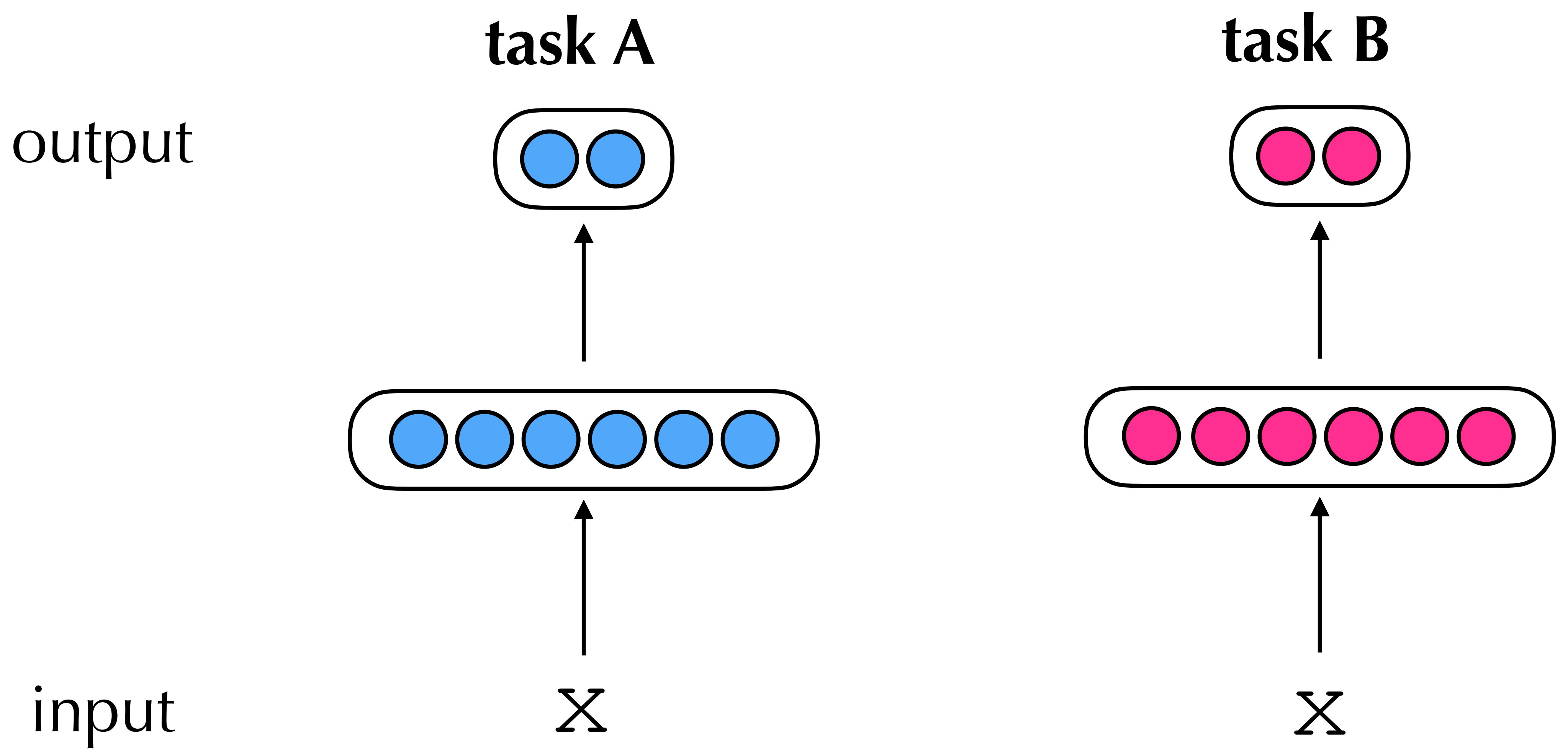
Outline

- Introduction: Why Transfer? Dimensions of Language Variation
- Part 1: What is Transfer Learning?
  - Three views on Transfer Learning, Related Learning Strategies
- Part 2: A type of TL: What is Multi-Task Learning?
  - What and Why, Perspectives on MTL
  - Short hands-on tutorial with MaChAmp
- Part 3: Selected Case Studies
  - Applications to Multilinguality, Transferability Estimation, Human Label Variation
- Outro

Part 2: What is Multi-Task Learning (MTL)?

Views on MTL and Why

Typical single-task learning

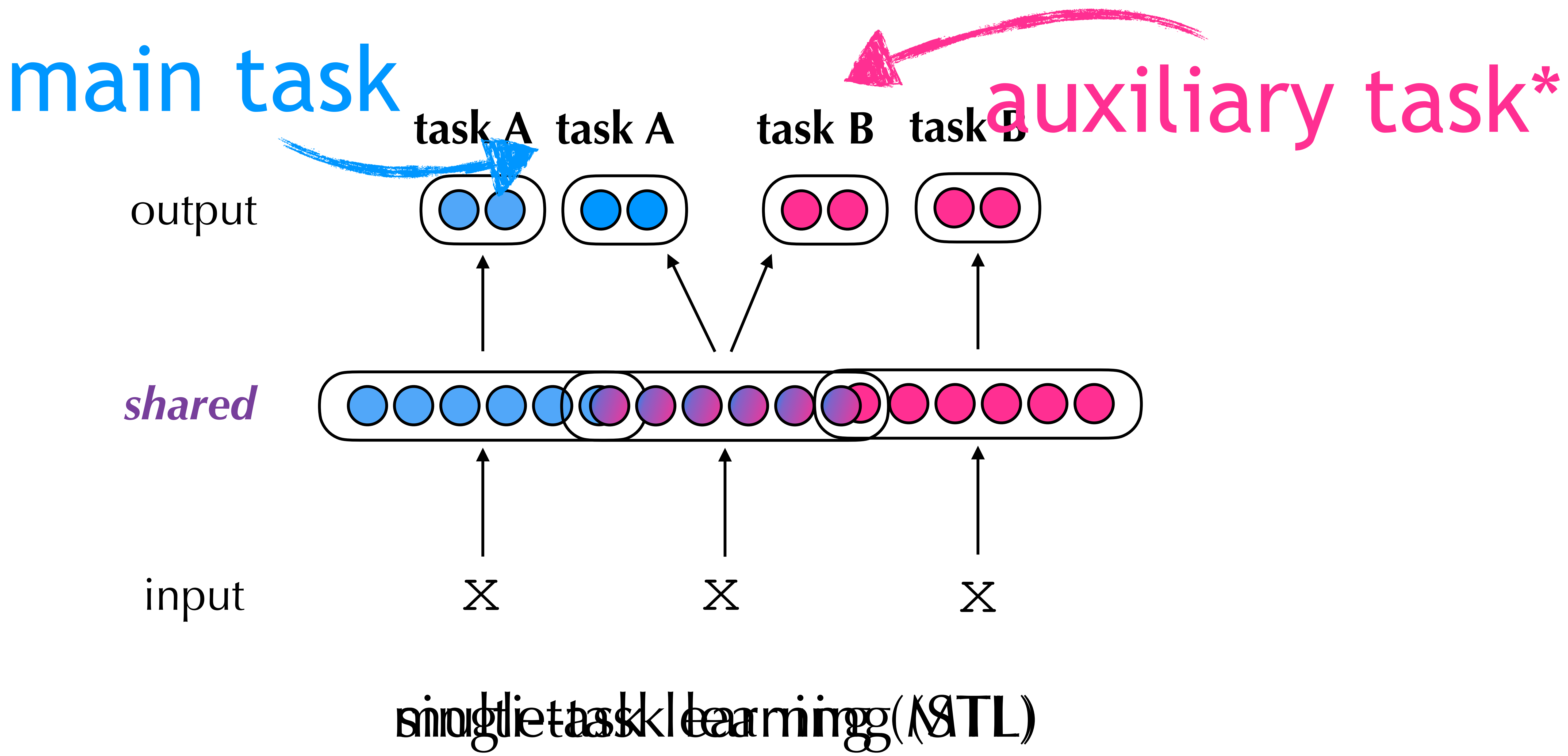


Can we do better?

Example: Learning how to drive a motorbike

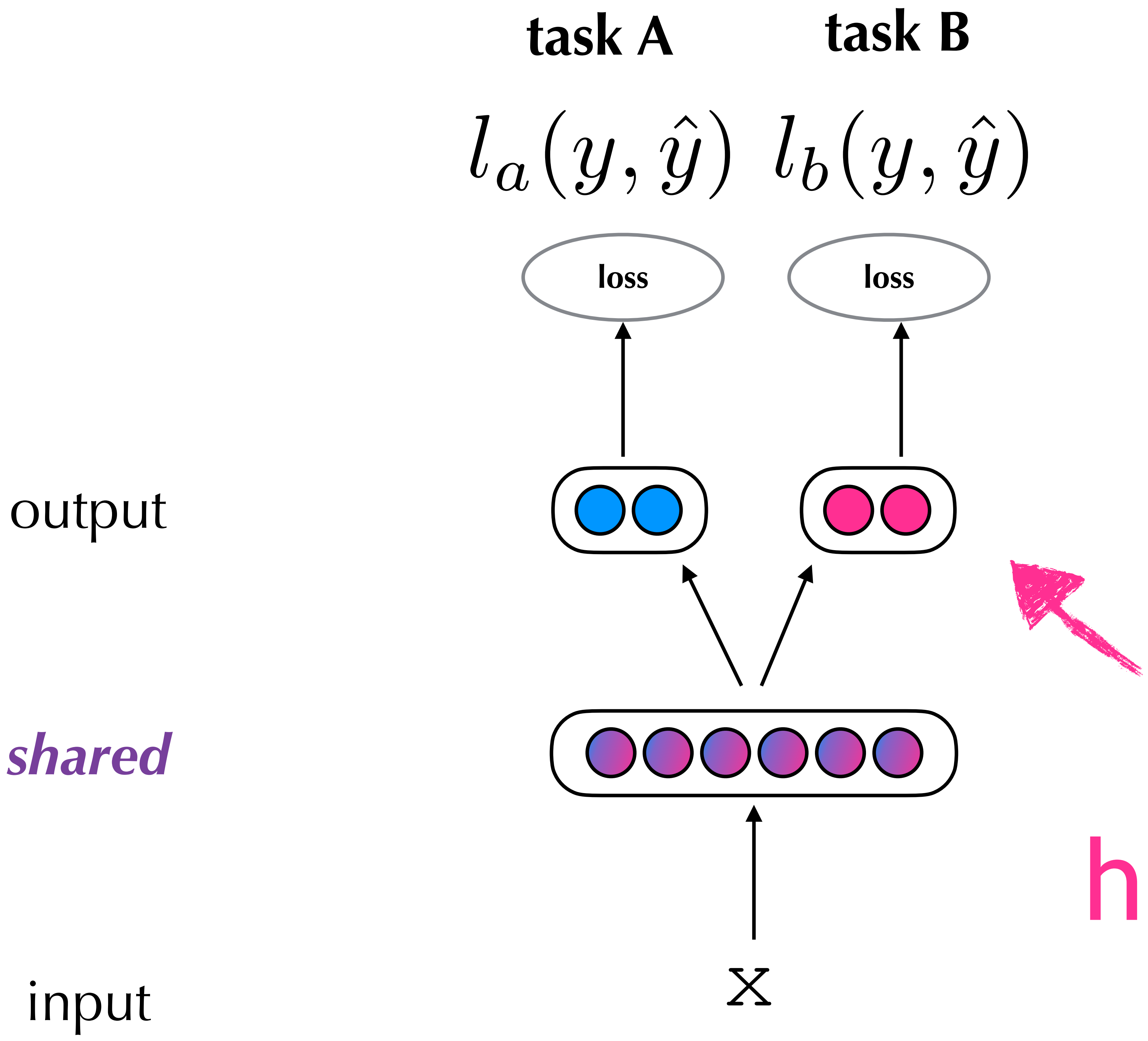
https://www.pexels.com/photo/bike-motorbike-motorcycle-21170/

Multi-task Learning (MTL): Key Idea



* sometimes auxiliary task might be equally important

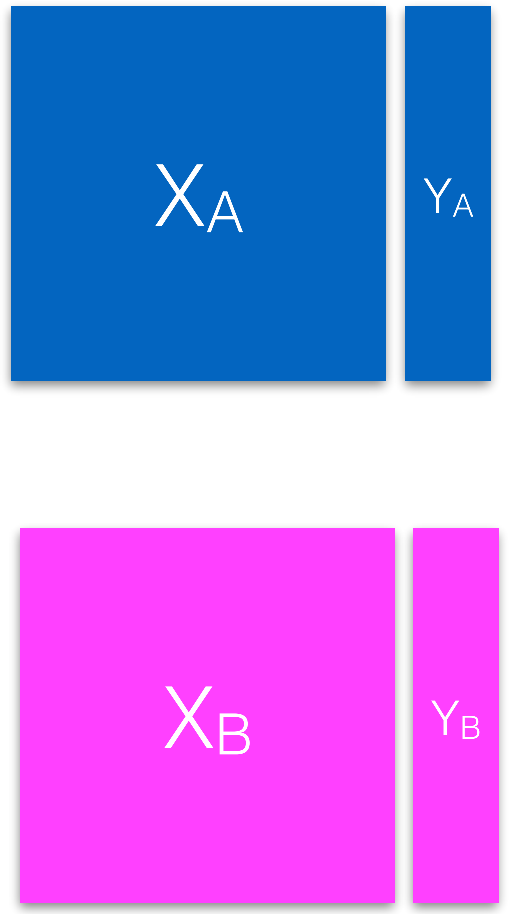
MTL in Neural Networks (NNs): shared encoder, task-specific heads



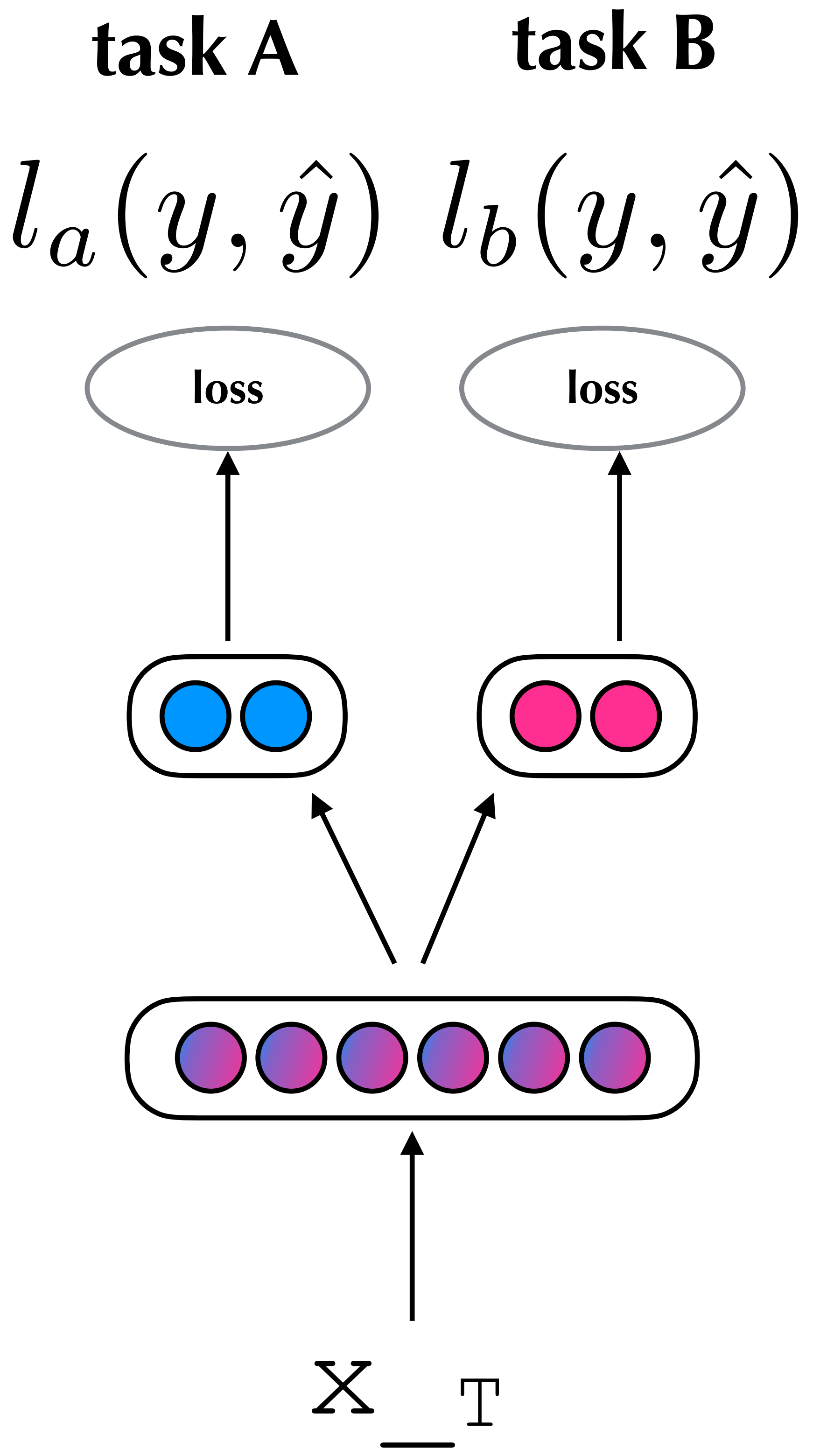
Task-specific heads (decoders)

MTL Recipe illustrated

$$\{\mathcal{D}_\tau\}_{\tau=1}^{T}$$

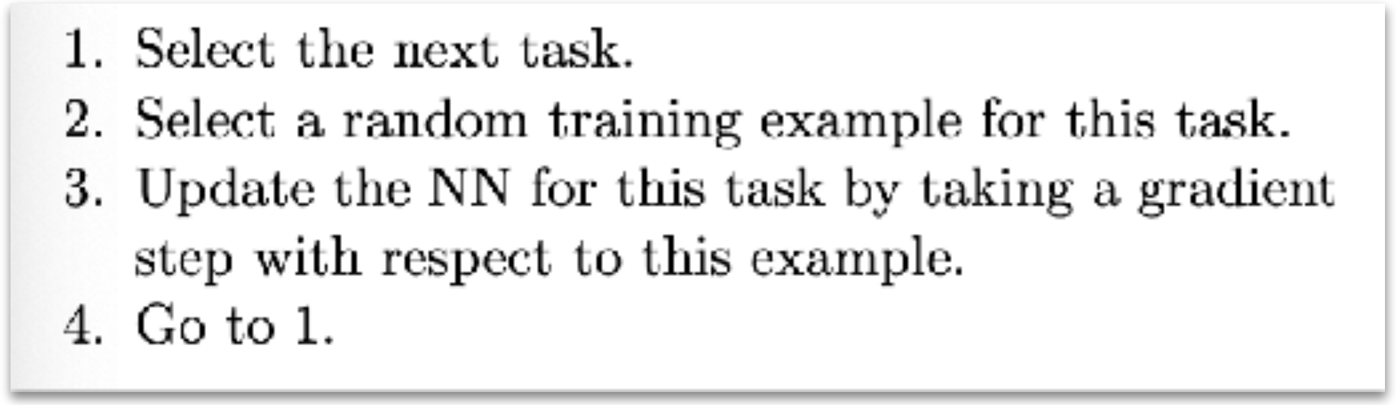


Data



Architecture

Sample task:



1. Select the next task.
2. Select a random training example for this task.
3. Update the NN for this task by taking a gradient step with respect to this example.
4. Go to 1.

(Collobert & Weston, 2008, ICML)

Training

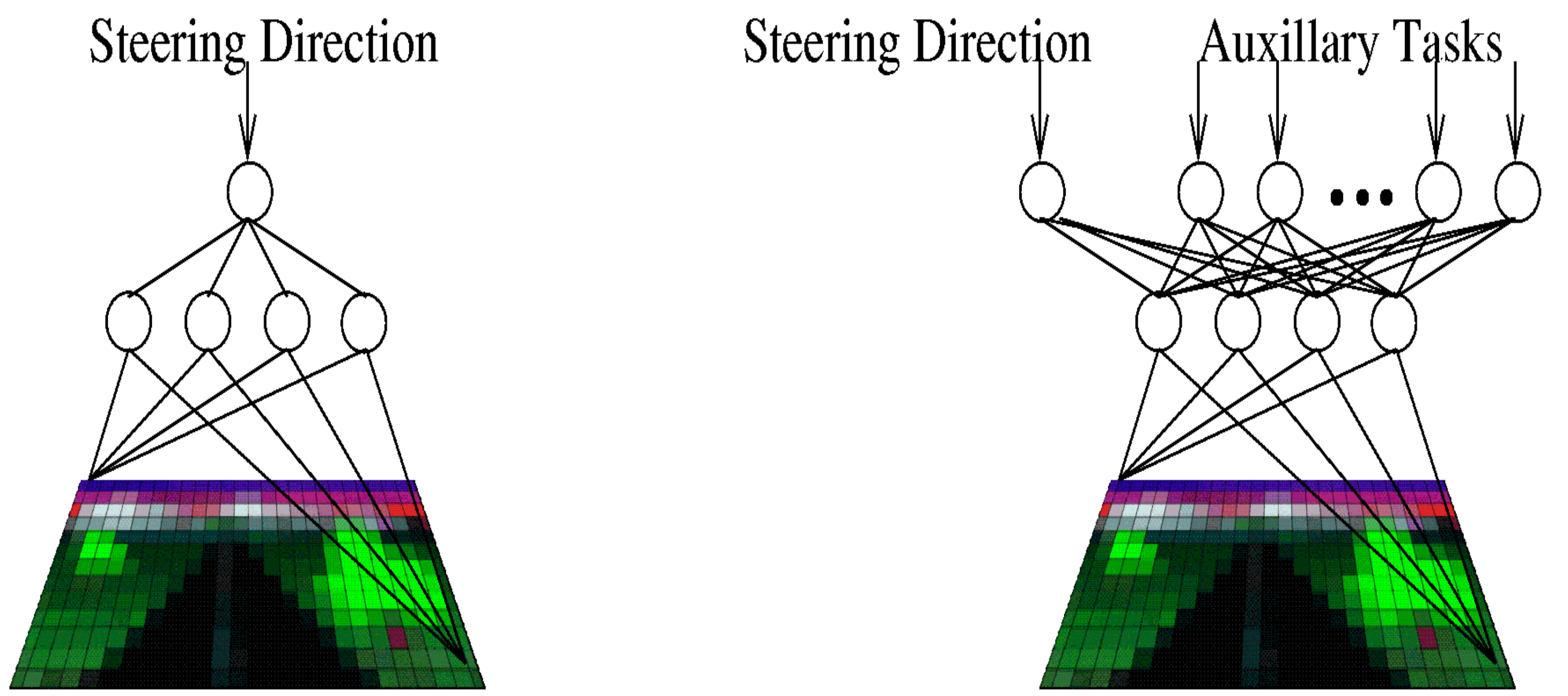
Why MTL?

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

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

Why does MTL help generalise? (1/2)

- **Attention focusing** (Caruana, 1997): reduced net capacity improves generalisation
- Example: ALVINN self-driving car



Single
Task Leaning

MultiTask Learning

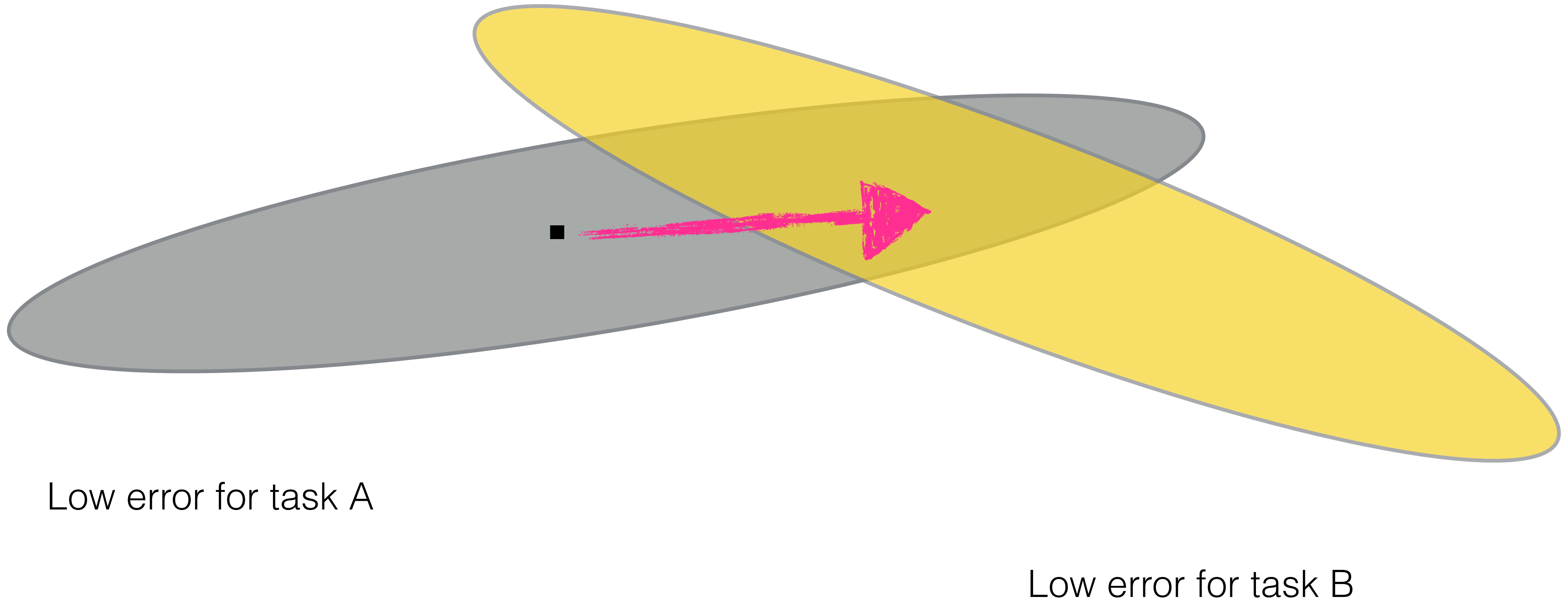
https://commons.wikimedia.org/wiki/File:Edible_fungi_in_basket_2012_G1.jpg

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

CMU Alvinn MTL (Caruana 1998)

Why does MTL help generalise? (2/2)

- **Representation bias** (Caruana, 1997) - MTL prefers solutions which other tasks prefer, acts as a **regulariser**



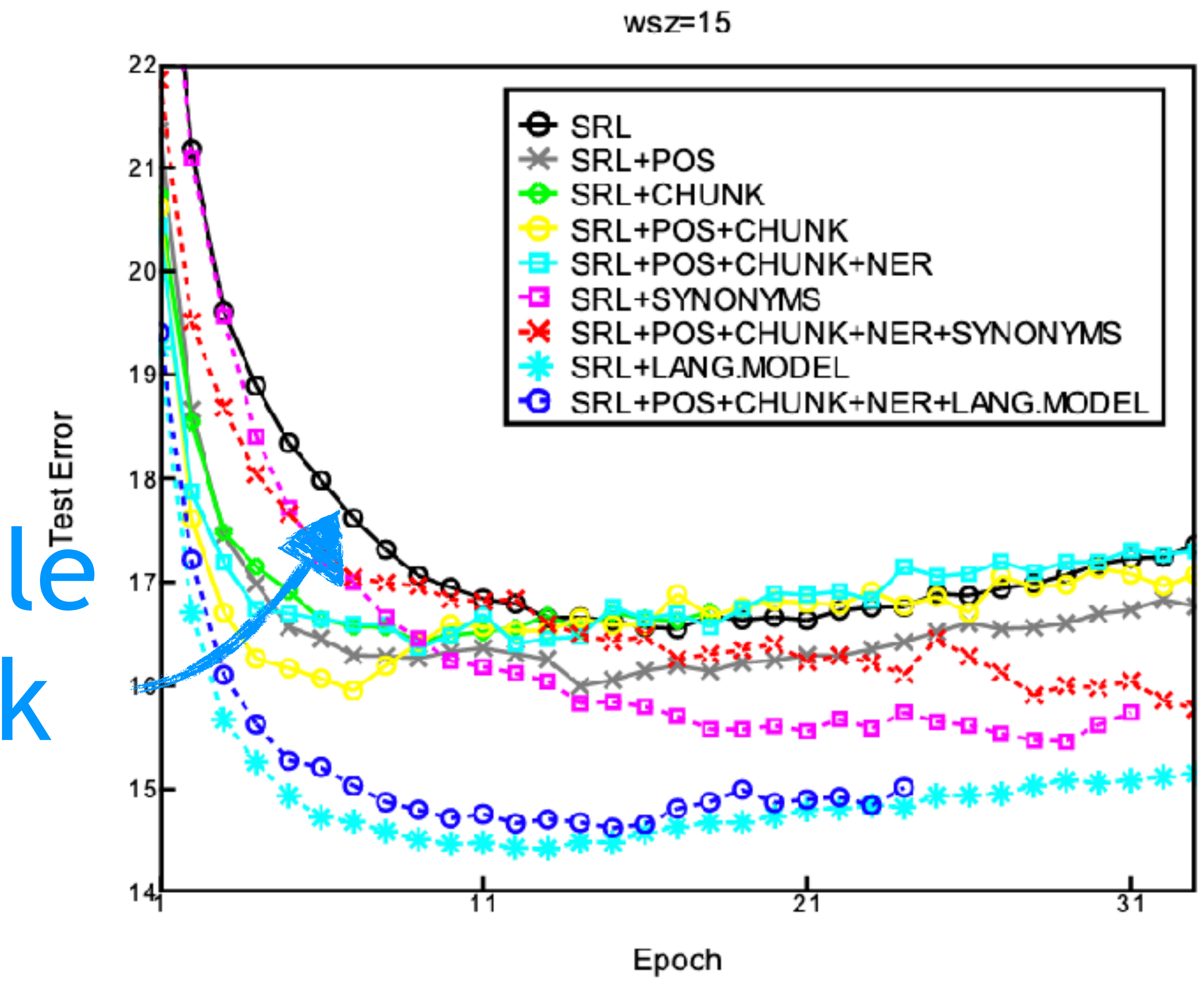
Why does MTL help efficiency? (1/3)

- **Eavesdropping** (Caruana, 1997) - eavedrop on shared representation to learn feature G through task B, which is hard to learn via task A

https://upload.wikimedia.org/wikipedia/commons/thumb/7/7b/Henri_Adolphe_Laissement_Kardin%C3%A4le_im_Vorzimmer_1895.jpg/1920px-Henri_Adolphe_Laissement_Kardin%C3%A4le_im_Vorzimmer_1895.jpg

Why does MTL help efficiency? (2/3)

- **Faster convergence** through learning tasks in parallel

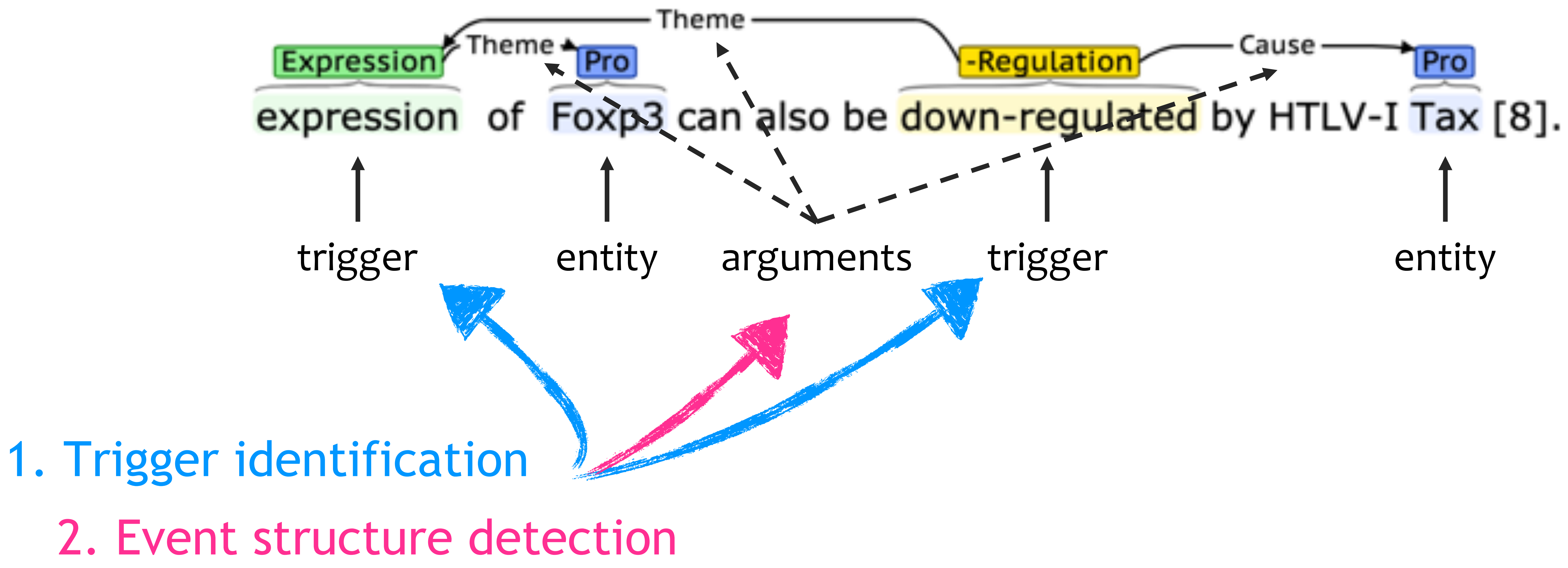


Single task

(Collobert & Weston, 2008, ICML)

Why does MTL help efficiency? (3/3)

- Replaces traditional pipelines with a single model for **faster inference** - Example from biomedical event extraction - Traditional pipeline:



1. Trigger identification

2. Event structure detection

Linearisation (cast as seq. labelling problem) + MTL = **BeeSL**

Biomedical **E**vent **E**xtraction as **S**equence **L**abeling

(Ramponi, van der Goot, Lombardo, Plank, EMNLP, 2020)

BeeSL: gains in accuracy + speed

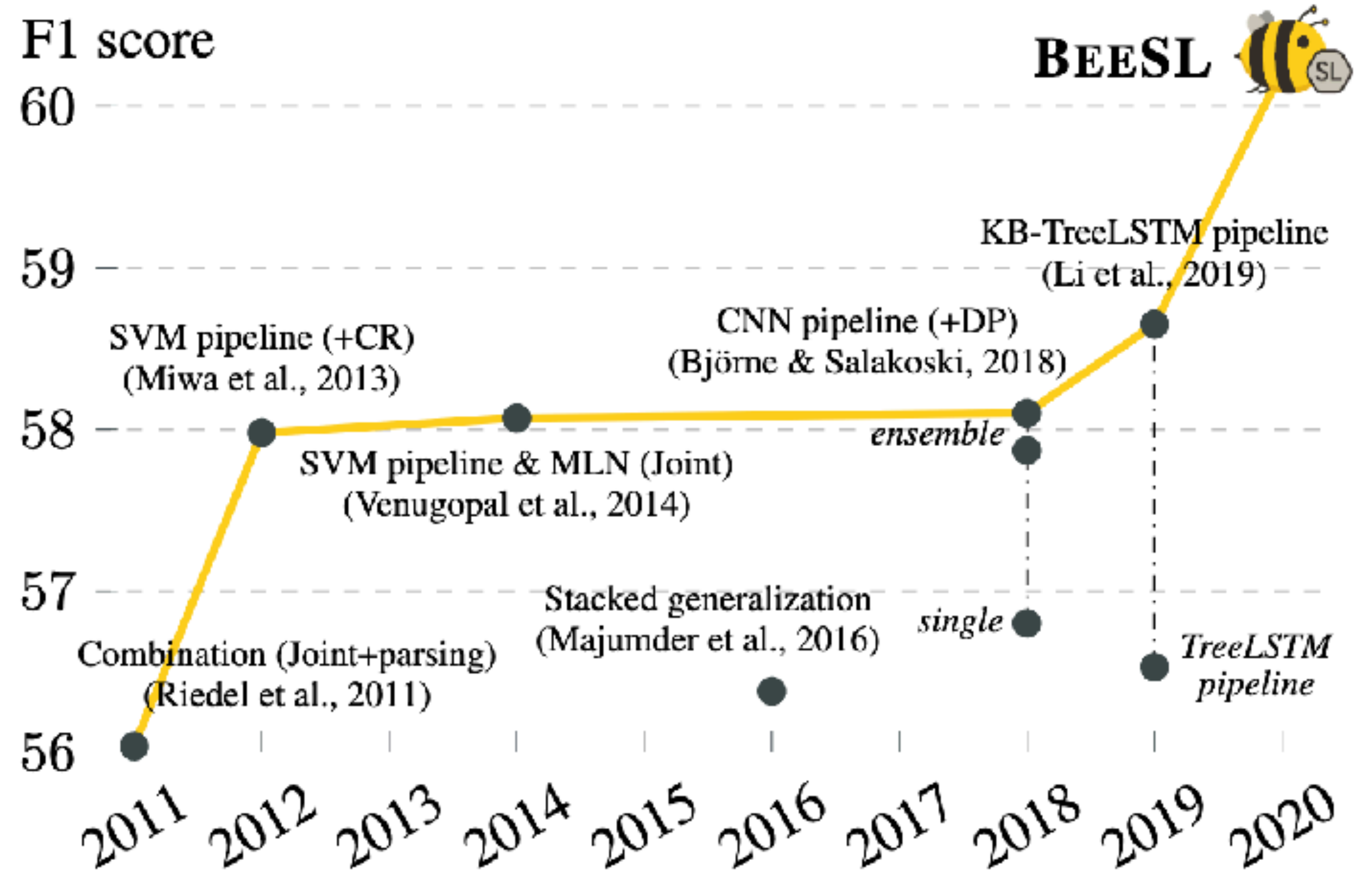
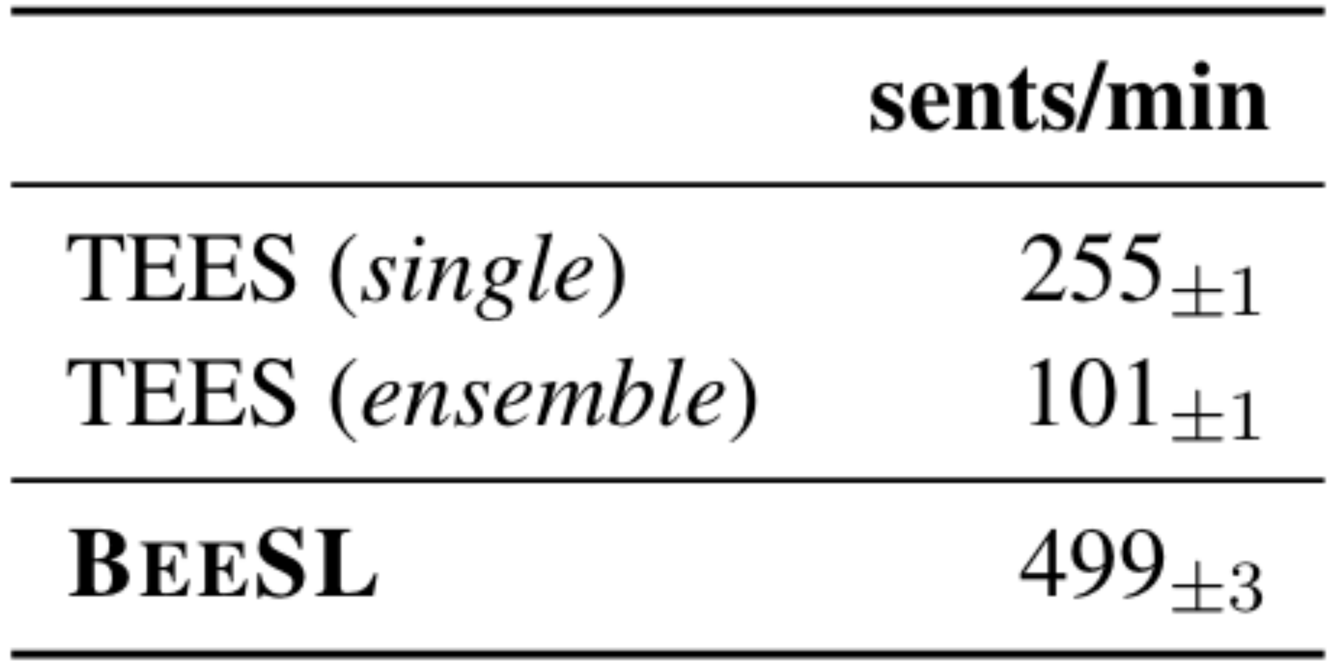


Figure 1: Performance of biomedical event extraction on the BioNLP Genia 2011 test set over time.

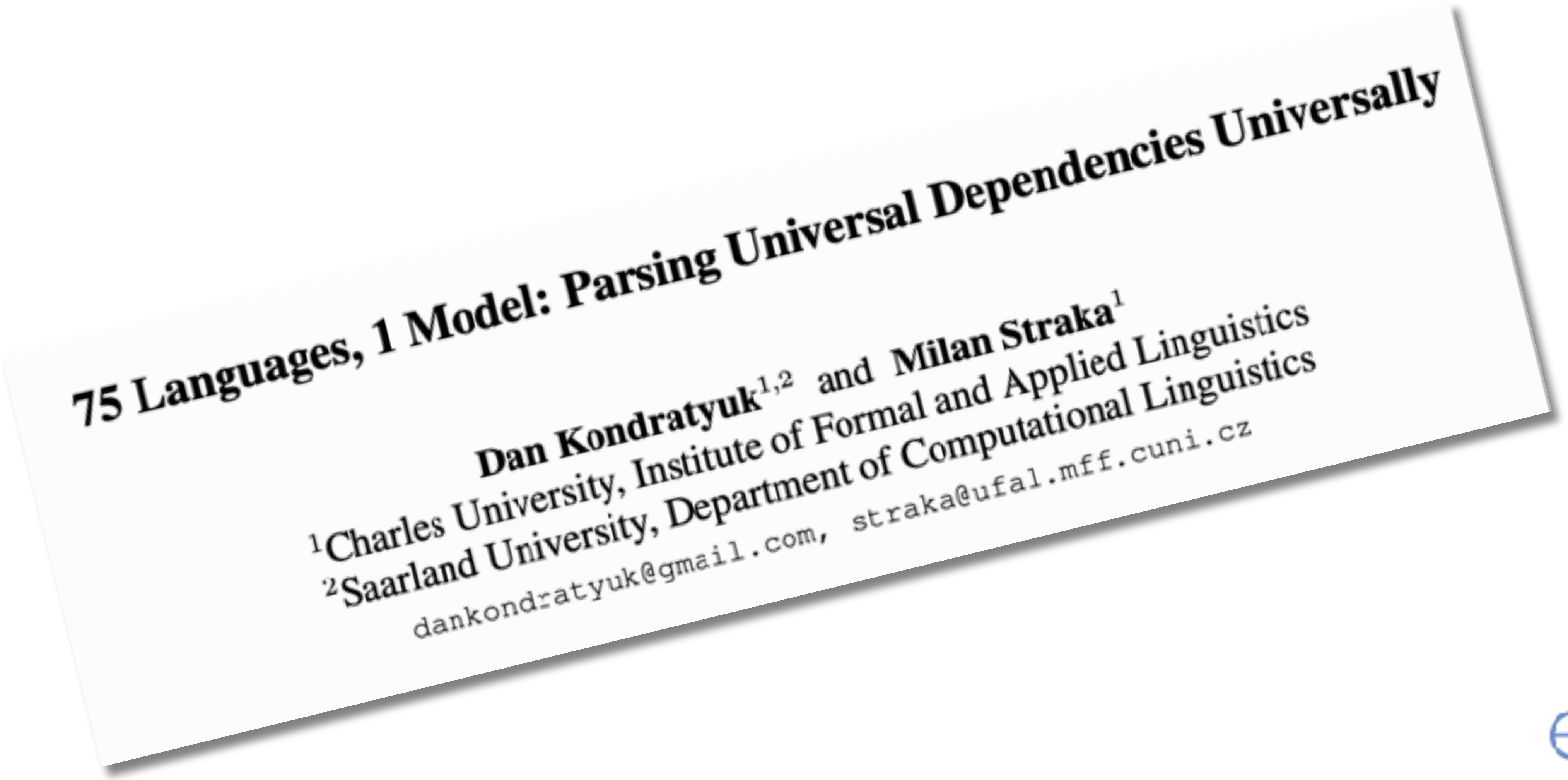
Inference time:
sentences/min



	sents/min
TEES (*single*)	$255_{\pm 1}$
TEES (*ensemble*)	$101_{\pm 1}$
BEESL	$499_{\pm 3}$

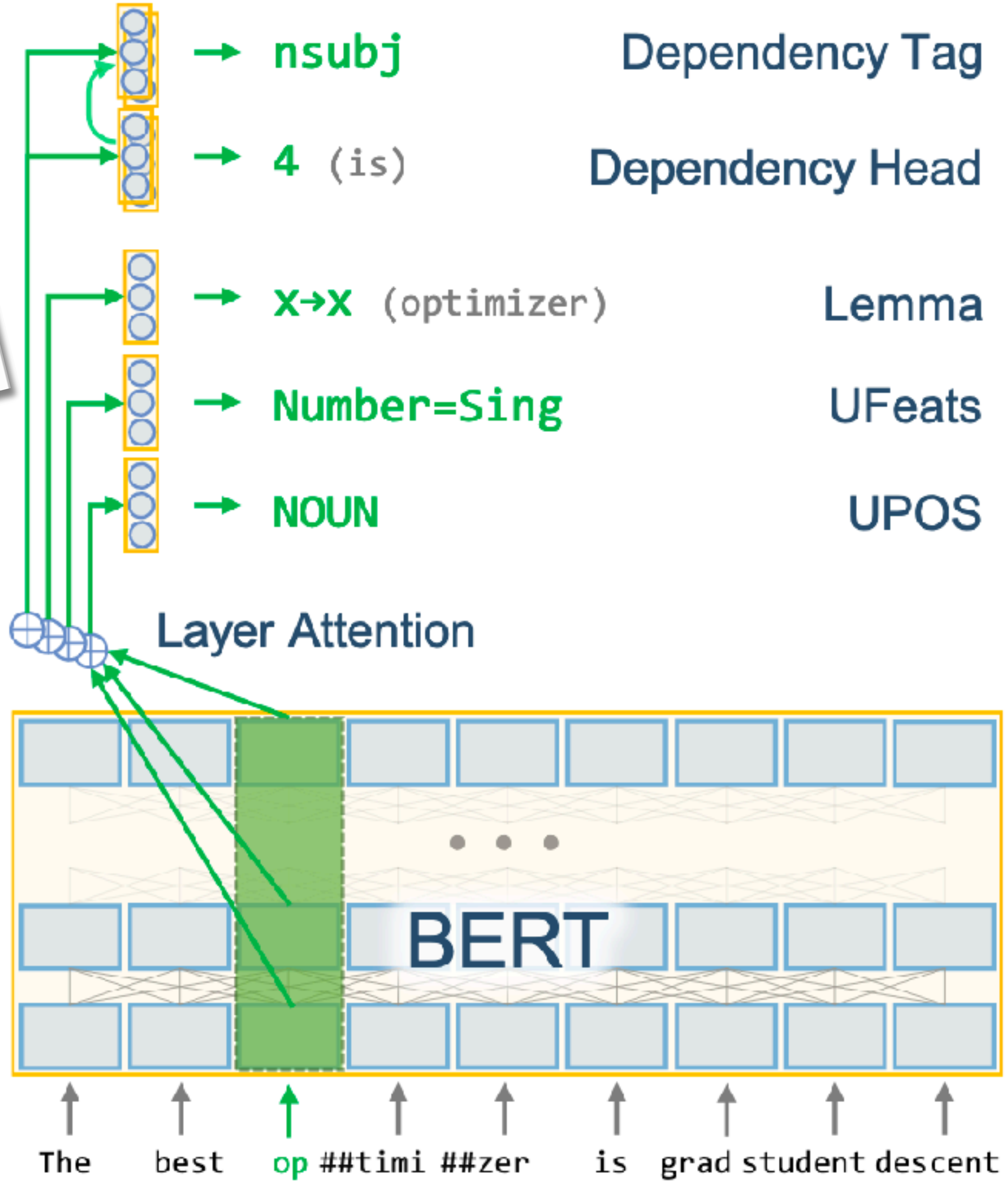
(Ramponi, van der Goot, Lombardo, Plank, EMNLP, 2020)

Example MTL Dependency Parser: 75 languages, 4 tasks, one model: UDify



75 Languages, 1 Model: Parsing Universal Dependencies Universally

Dan Kondratyuk[1,2] and Milan Straka[1]
[1]Charles University, Institute of Formal and Applied Linguistics
[2]Saarland University, Department of Computational Linguistics
dankondratyuk@gmail.com, straka@ufal.mff.cuni.cz



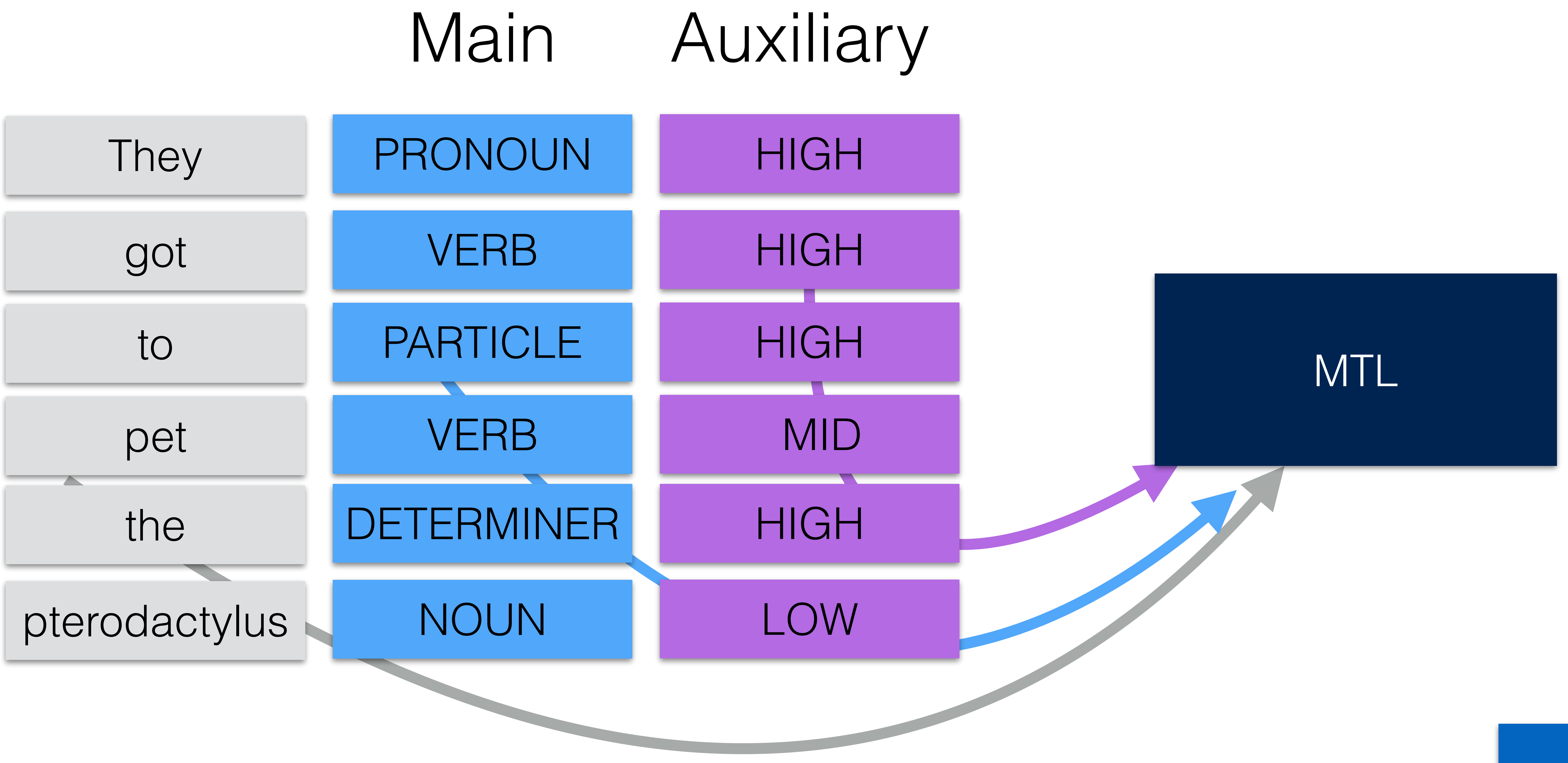
https://arxiv.org/pdf/1904.02099v2.pdf

EMNLP, 2019

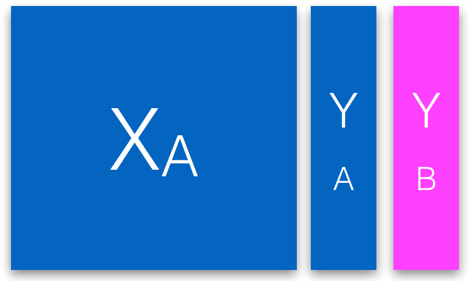
Perspectives on MTL

MTL: learning from distinct views

e.g., predict data properties (Plank et al., 2016 ACL),
predict other data views like discourse tree views (Braud et al. 2016 CoNLL),
predict other layers like syntax tree layers (Kondratuk & Straka, 2019 EMNLP)

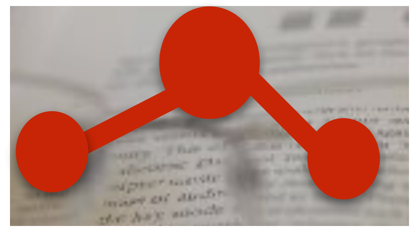


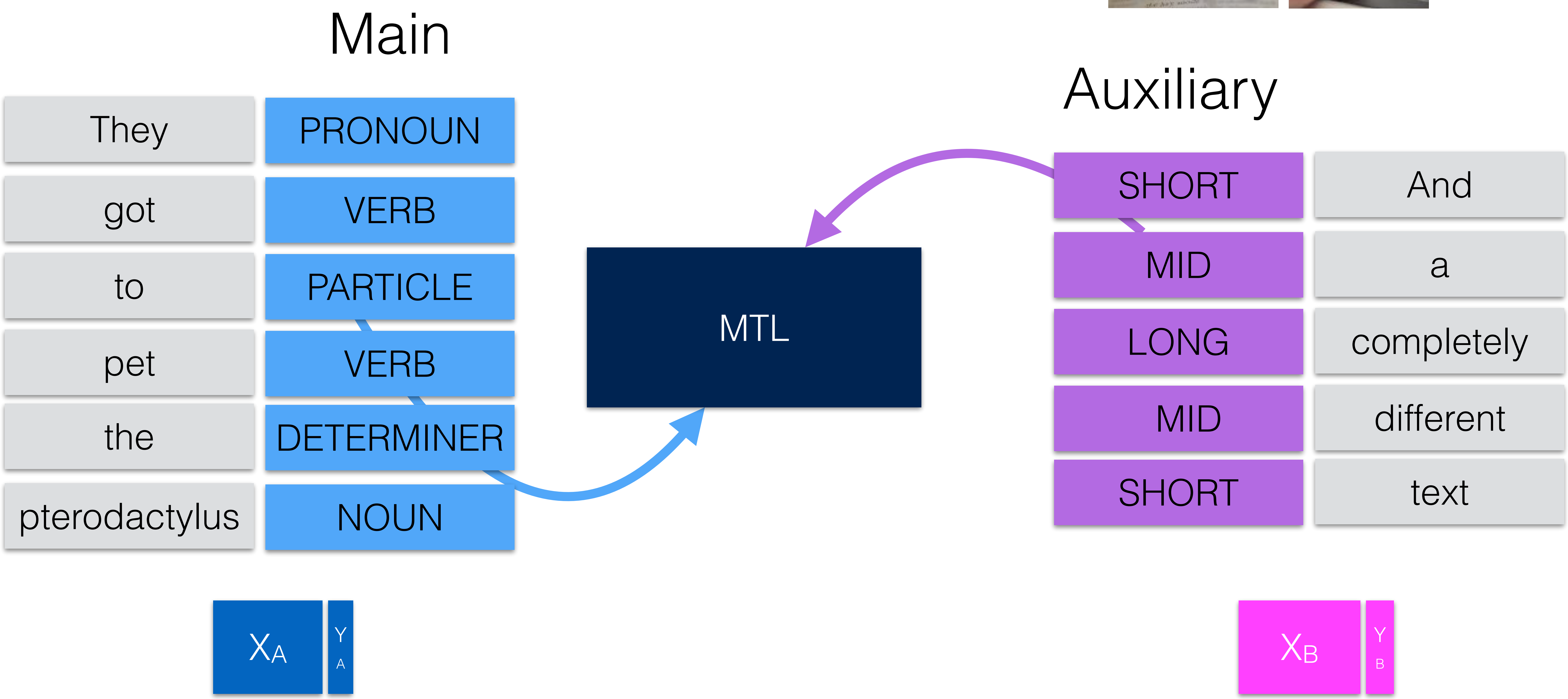
	Main	Auxiliary
They	PRONOUN	HIGH
got	VERB	HIGH
to	PARTICLE	HIGH
pet	VERB	MID
the	DETERMINER	HIGH
pterodactylus	NOUN	LOW



MTL: learning from distinct sources

e.g., from other languages but also more remote sources like
cognitive human data (gaze, keystrokes)
(Klerke et al. 2016 NAACL), (Plank 2016 COLING), (Barrett & Hollenstein, 2020)





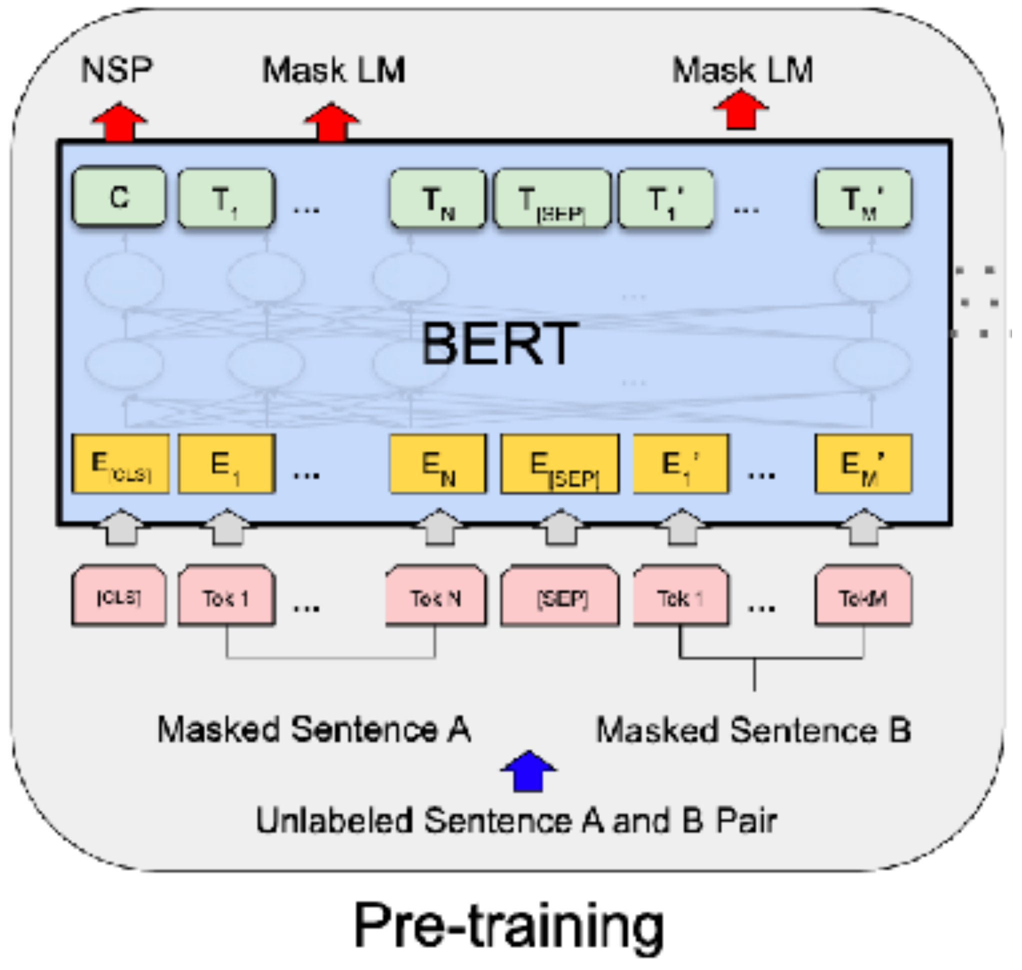
Main

They	PRONOUN
got	VERB
to	PARTICLE
pet	VERB
the	DETERMINER
pterodactylus	NOUN

Auxiliary

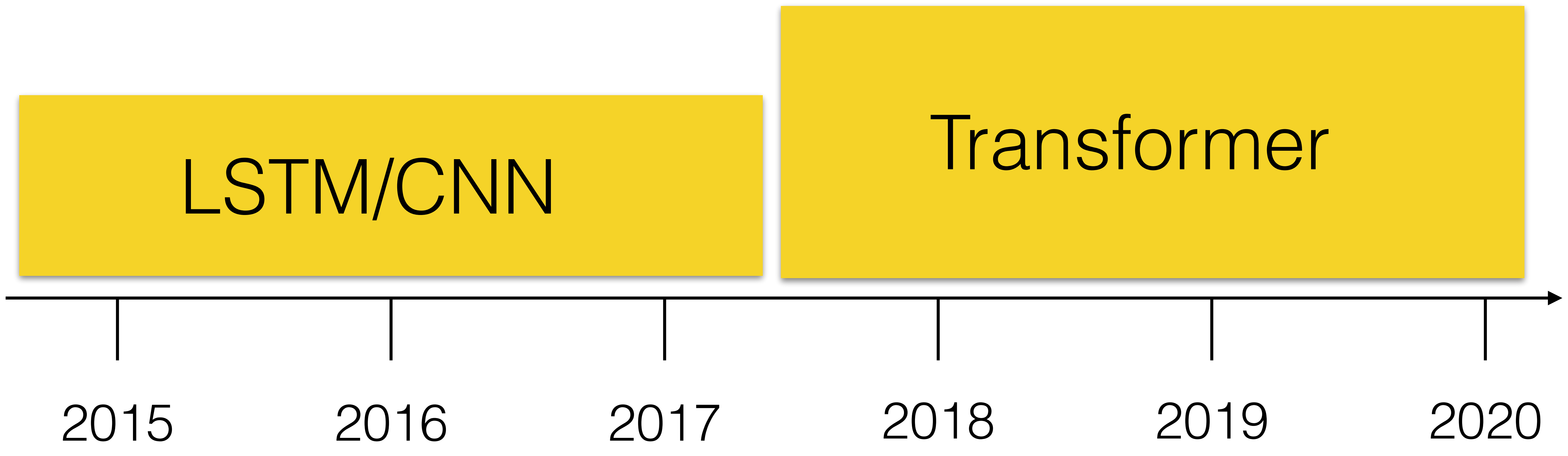
SHORT	And
MID	a
LONG	completely
MID	different
SHORT	text

Today: MTL everywhere!



Self-supervised MTL objectives: MLM + NSP

First MTL wave (2016-2017)



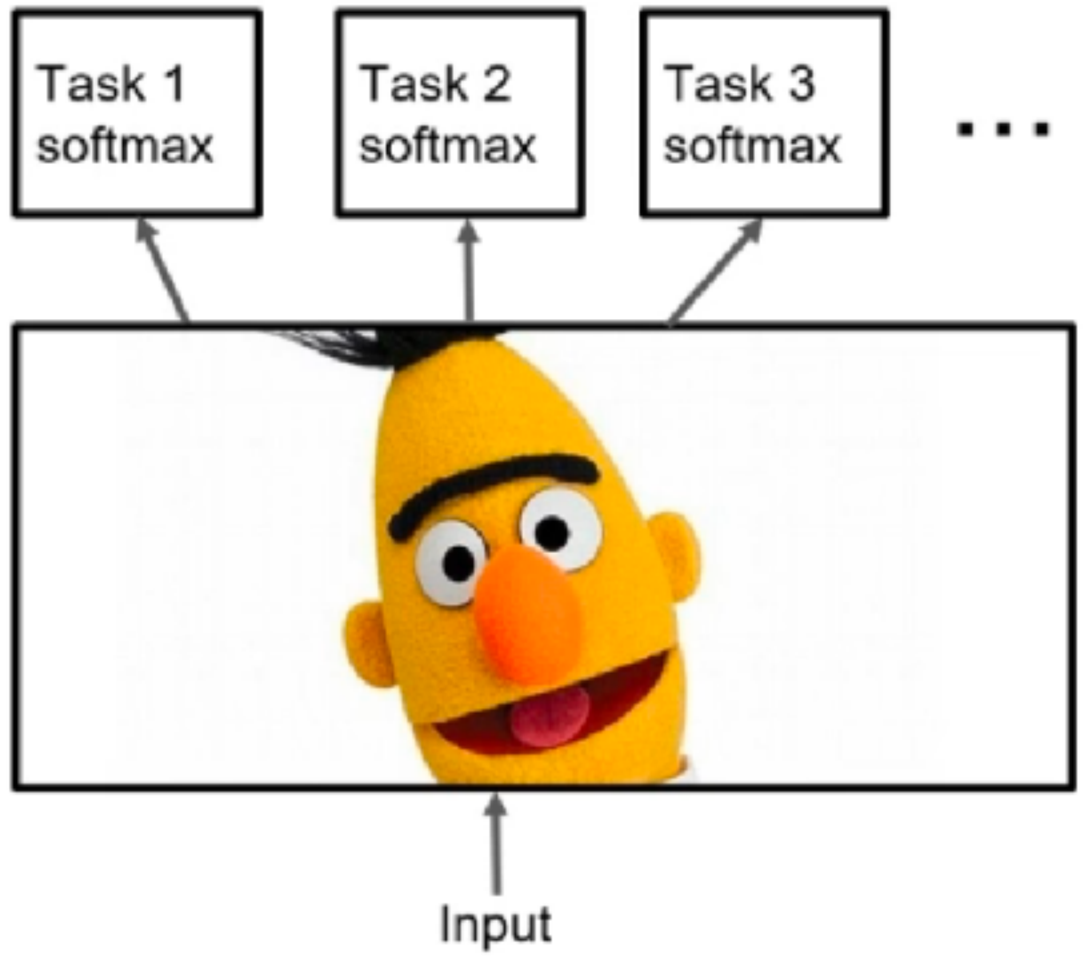
LSTM/CNN

Transformer

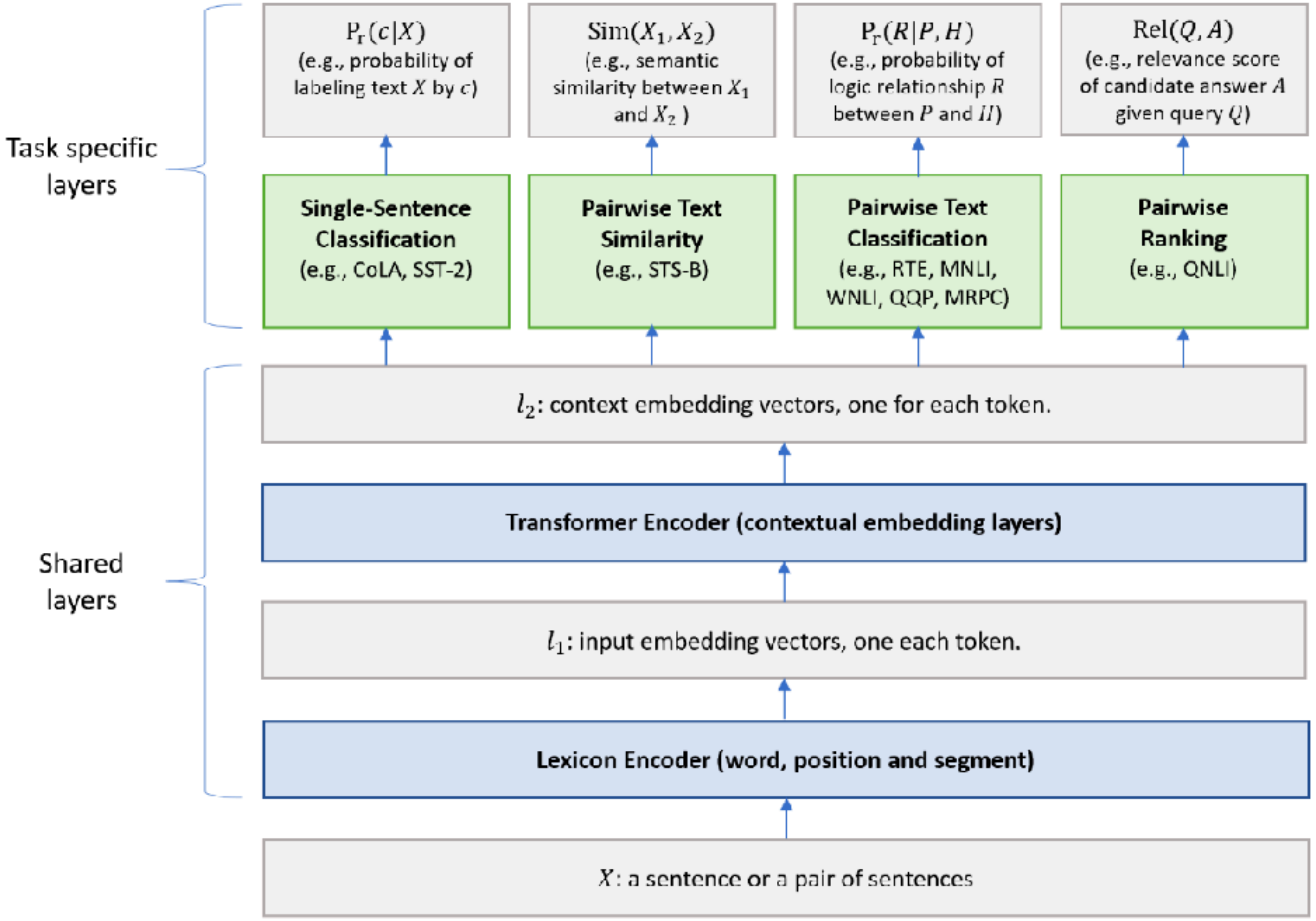
2015 2016 2017 2018 2019 2020

Vaswani et al., 2017; Peters et al., 2018

… and Multi-task Fine-Tuning using BERT & co



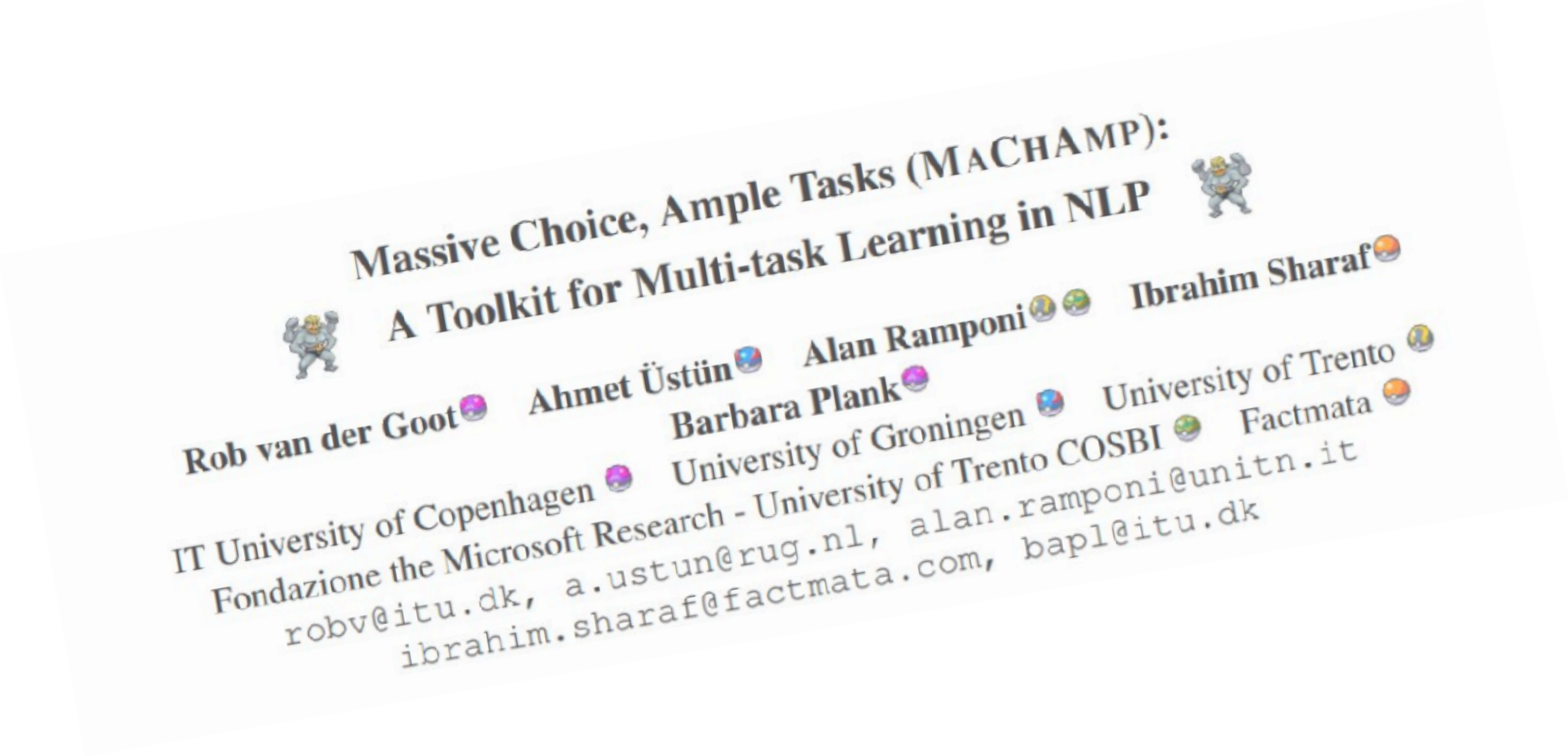
Slide adapted from Clark et al, 2019


Task specific layers
Shared layers
$P_r(c|X)$ (e.g., probability of labeling text X by c)
$\mathrm{Sim}(X_1, X_2)$ (e.g., semantic similarity between X_1 and X_2)
$P_r(R|P,H)$ (e.g., probability of logic relationship R between P and H)
$\mathrm{Rel}(Q, A)$ (e.g., relevance score of candidate answer A given query Q)
Single-Sentence Classification (e.g., CoLA, SST-2)
Pairwise Text Similarity (e.g., STS-B)
Pairwise Text Classification (e.g., RTE, MNLI, WNLI, QQP, MRPC)
Pairwise Ranking (e.g., QNLI)
l_2: context embedding vectors, one for each token.
Transformer Encoder (contextual embedding layers)
l_1: input embedding vectors, one each token.
Lexicon Encoder (word, position and segment)
X: a sentence or a pair of sentences

MT-DNN by Liu et al., ACL 2019

Outline

- Introduction: Why Transfer? Dimensions of Language Variation
- Part 1: What is Transfer Learning?
  - Three views on Transfer Learning, Related Learning Strategies
- Part 2: A type of TL: What is Multi-Task Learning?
  - What and Why, Perspectives on MTL
  - Short hands-on tutorial with MaChAmp
- Part 3: Selected Case Studies
  - Applications to Multilinguality, Transferability Estimation, Human Label Variation
- Outro



Massive Choice, Ample Tasks (MACHAMP):
A Toolkit for Multi-task Learning in NLP

Rob van der Goot Ahmet Üstün Alan Ramponi Ibrahim Sharaf
Barbara Plank
IT University of Copenhagen University of Groningen University of Trento
Fondazione the Microsoft Research - University of Trento COSBI Factmata
robv@itu.dk, a.ustun@rug.nl, alan.ramponi@unitn.it
ibrahim.sharaf@factmata.com, bapl@itu.dk

Massive Choice, Ample Tasks: MaChAmp

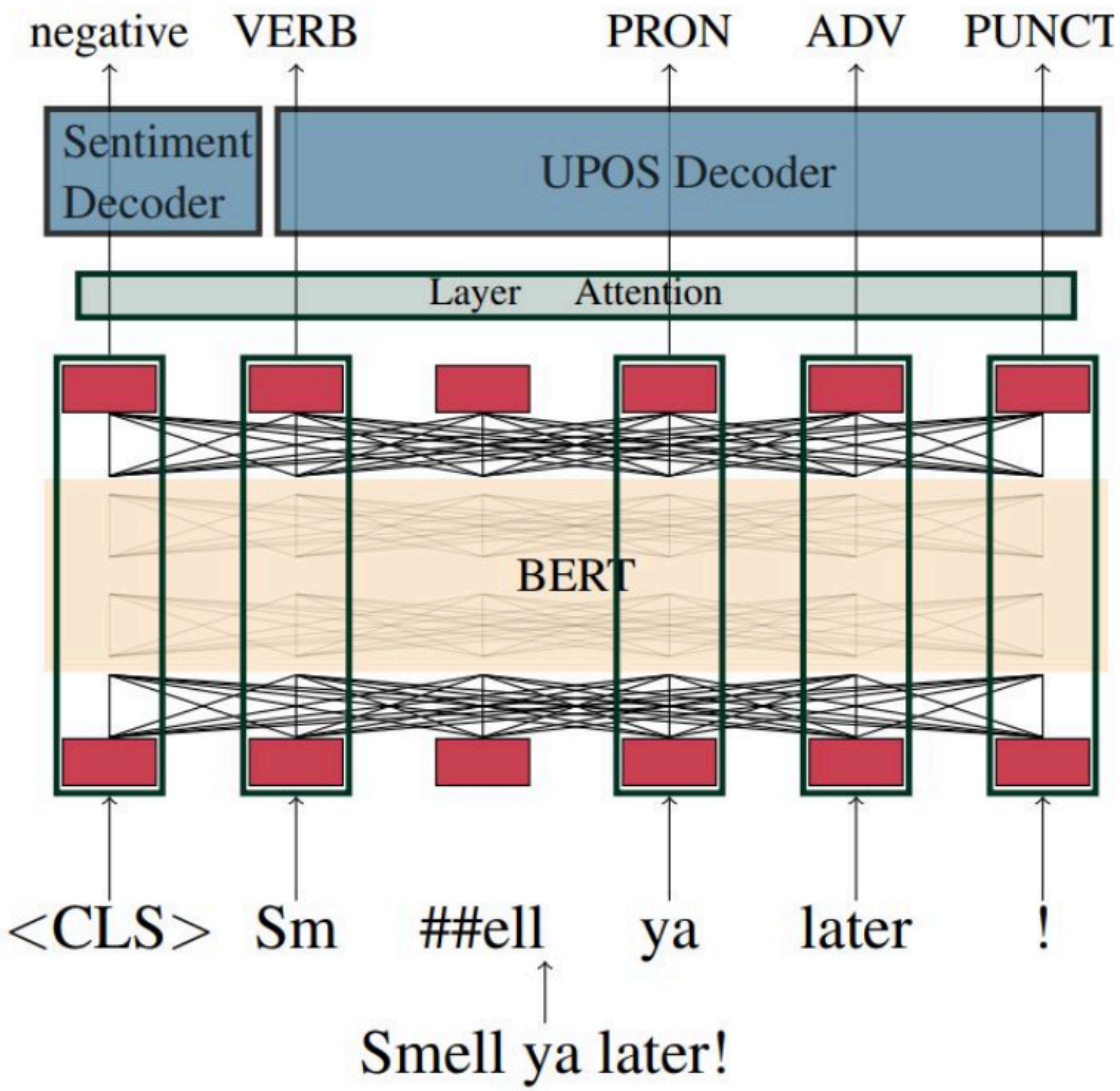
An easy-to-use (MTL) toolkit

MaChAmp

- Ease of use (all based on simple configuration files)
- Support many tasks (classification, sequence labelling, pairwise sentence classification, dependency parsing..)
- Ease of switching underlying LM encoder
- Multi-task learning via configuration files

One arm alone can move mountains.

Architecture



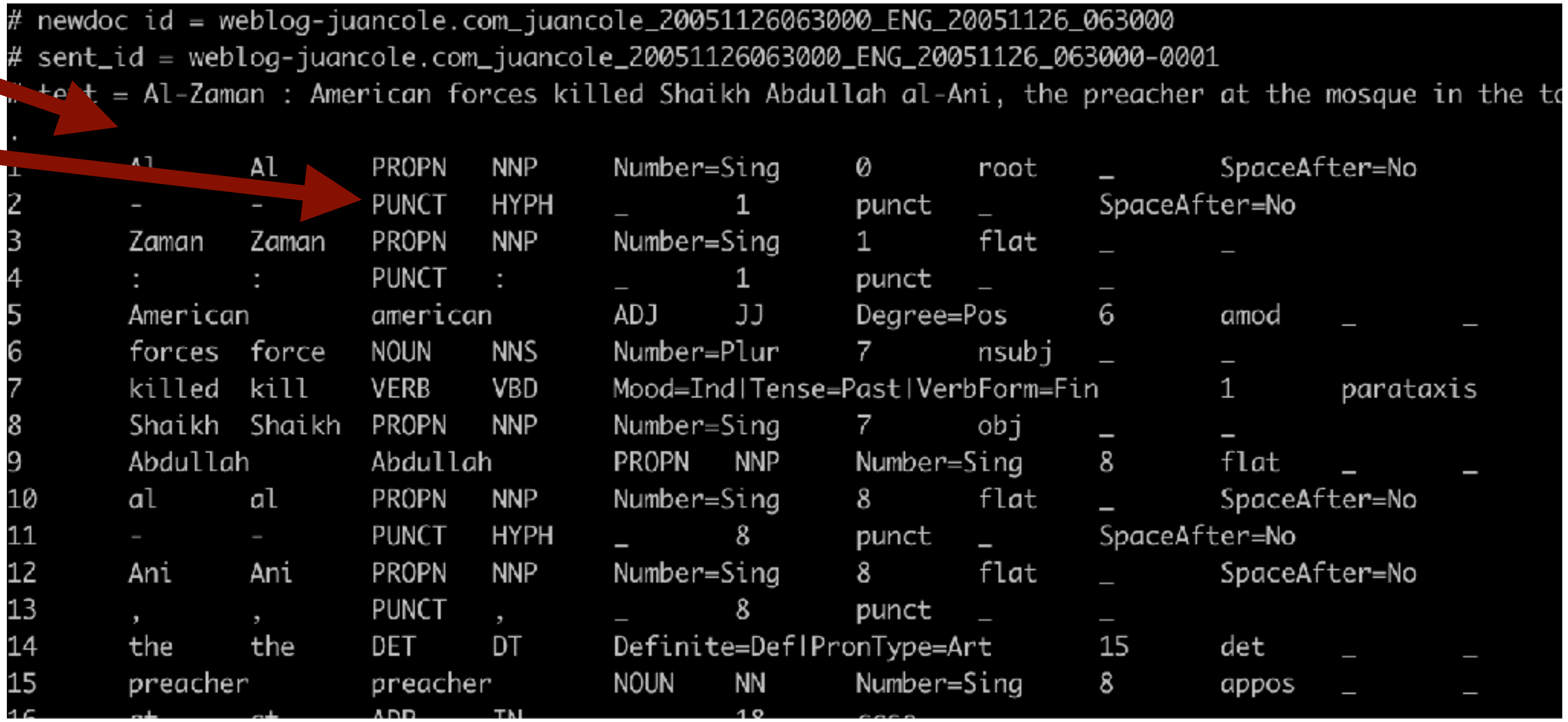
Configuration and Training of a single task

- Configuration file:

```
{
    "UD": {
        "train_data_path": "data/ewt.train",
        "validation_data_path": "data/ewt.dev",
        "word_idx": 1,
        "tasks": {
            "upos": {
                "task_type": "seq",
                "column_idx": 3
            }
        }
    }
}
```

- Training:

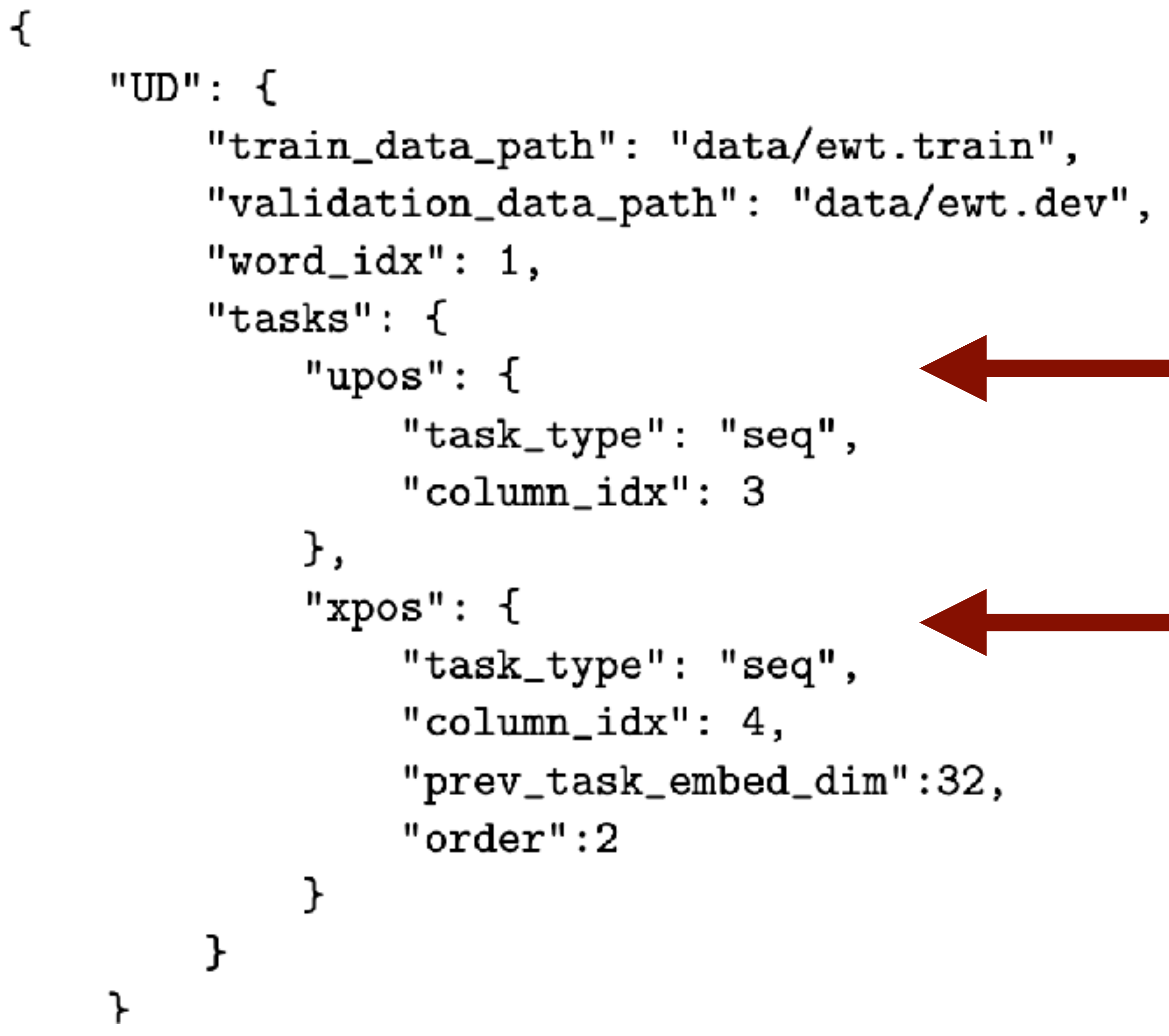
```
python3 train.py --dataset_config upos.json
```



```
# newdoc id = weblog-juancole.com_juancole_20051126063000_ENG_20051126_063000
# sent_id = weblog-juancole.com_juancole_20051126063000_ENG_20051126_063000-0001
# text = Al-Zaman : American forces killed Shaikh Abdullah al-Ani, the preacher at the mosque in the to
.
1       Al      Al      PROPN   NNP     Number=Sing     0       root    _       SpaceAfter=No
2       -       -       PUNCT   HYPH    _       1       punct   _       SpaceAfter=No
3       Zaman   Zaman   PROPN   NNP     Number=Sing     1       flat    _       _
4       :       :       PUNCT   :       _       1       punct   _       _
5       American        american        ADJ     JJ      Degree=Pos      6       amod    _       _
6       forces  force   NOUN    NNS     Number=Plur     7       nsubj   _       _
7       killed  kill    VERB    VBD     Mood=Ind|Tense=Past|VerbForm=Fin        1       parataxis
8       Shaikh  Shaikh  PROPN   NNP     Number=Sing     7       obj     _       _
9       Abdullah        Abdullah        PROPN   NNP     Number=Sing     8       flat    _       _
10      al      al      PROPN   NNP     Number=Sing     8       flat    _       SpaceAfter=No
11      -       -       PUNCT   HYPH    _       8       punct   _       SpaceAfter=No
12      Ani     Ani     PROPN   NNP     Number=Sing     8       flat    _       SpaceAfter=No
13      ,       ,       PUNCT   ,       _       8       punct   _       _
14      the     the     DET     DT      Definite=Def|PronType=Art       15      det     _       _
15      preacher        preacher        NOUN    NN      Number=Sing     8       appos   _       _
16      at      at      ADP     IN      _       18      case
```

Configuration and Training of two tasks (e.g. coarse and fine POS)

- Configuration file:



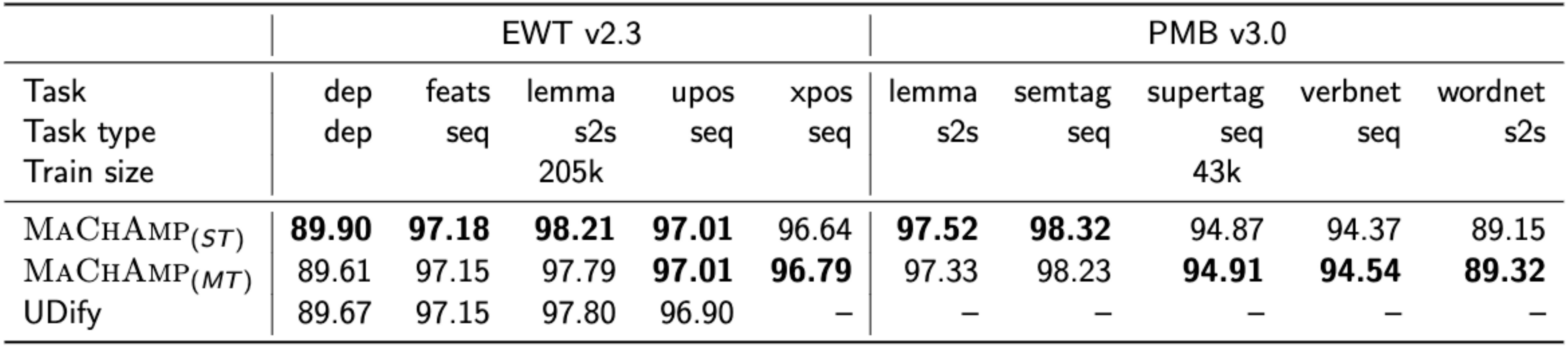
```
{
    "UD": {
        "train_data_path": "data/ewt.train",
        "validation_data_path": "data/ewt.dev",
        "word_idx": 1,
        "tasks": {
            "upos": {
                "task_type": "seq",
                "column_idx": 3
            },
            "xpos": {
                "task_type": "seq",
                "column_idx": 4,
                "prev_task_embed_dim":32,
                "order":2
            }
        }
    }
```

Task types:

- seq: standard sequence labeling.
- string2string: same as sequence labeling, but learns a conversion from the original word to the instance, and uses that as label (useful for lemmatization).
- seq_bio: a masked CRF decoder enforcing complying with the BIO-scheme.
- multiseq: a multilabel version of seq: multilabel classification on the word level
- multiclass: a multilabel version of classification: multilabel classification on the utterance level.
- dependency: dependency parsing.
- classification: sentence classification, predicts a label for N utterances of text.
- mlm: masked language modeling.
- regression: to predict (floating point) numbers

Results to Udify

- More details in van der Goot et al., 2021 EACL



	EWT v2.3					PMB v3.0				
Task	dep	feats	lemma	upos	xpos	lemma	semtag	supertag	verbnet	wordnet
Task type	dep	seq	s2s	seq	seq	s2s	seq	seq	seq	s2s
Train size			205k					43k		
MaChAmp$_{(ST)}$	**89.90**	**97.18**	**98.21**	**97.01**	96.64	**97.52**	**98.32**	94.87	94.37	89.15
MaChAmp$_{(MT)}$	89.61	97.15	97.79	**97.01**	**96.79**	97.33	98.23	**94.91**	**94.54**	**89.32**
UDify	89.67	97.15	97.80	96.90	–	–	–	–	–	–

More info on MaChAmp

- Website with code, documentation: https://machamp-nlp.github.io/
- MaChAmp Colab tutorial (short, check out the documentation above): https://colab.research.google.com/drive/1zkowQPeiQMgKnEmKITjccTRvtfdpGfEH
- Slack channel and GitHub issues, see website for more information

Outline

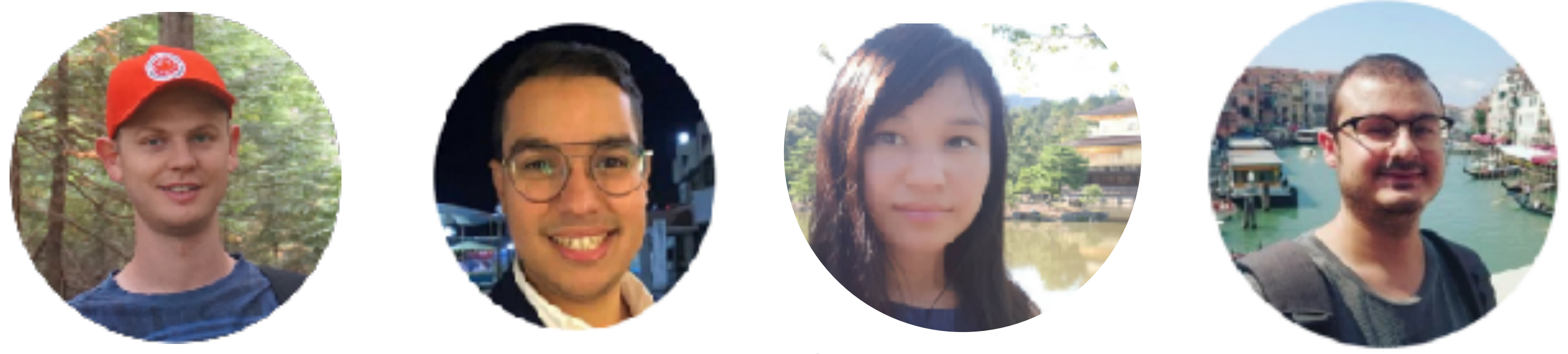
- Introduction: Why Transfer? Dimensions of Language Variation
- Part 1: What is Transfer Learning?
  - Three views on Transfer Learning, Related Learning Strategies
- Part 2: A type of TL: What is Multi-Task Learning?
  - What and Why, Perspectives on MTL
  - Short hands-on tutorial with MaChAmp
- Part 3: Selected Case Studies
  - Applications to Multilinguality, Transferability Estimation, Human Label Variation
- Outro

Applications to Multilinguality

Selected Case Studies

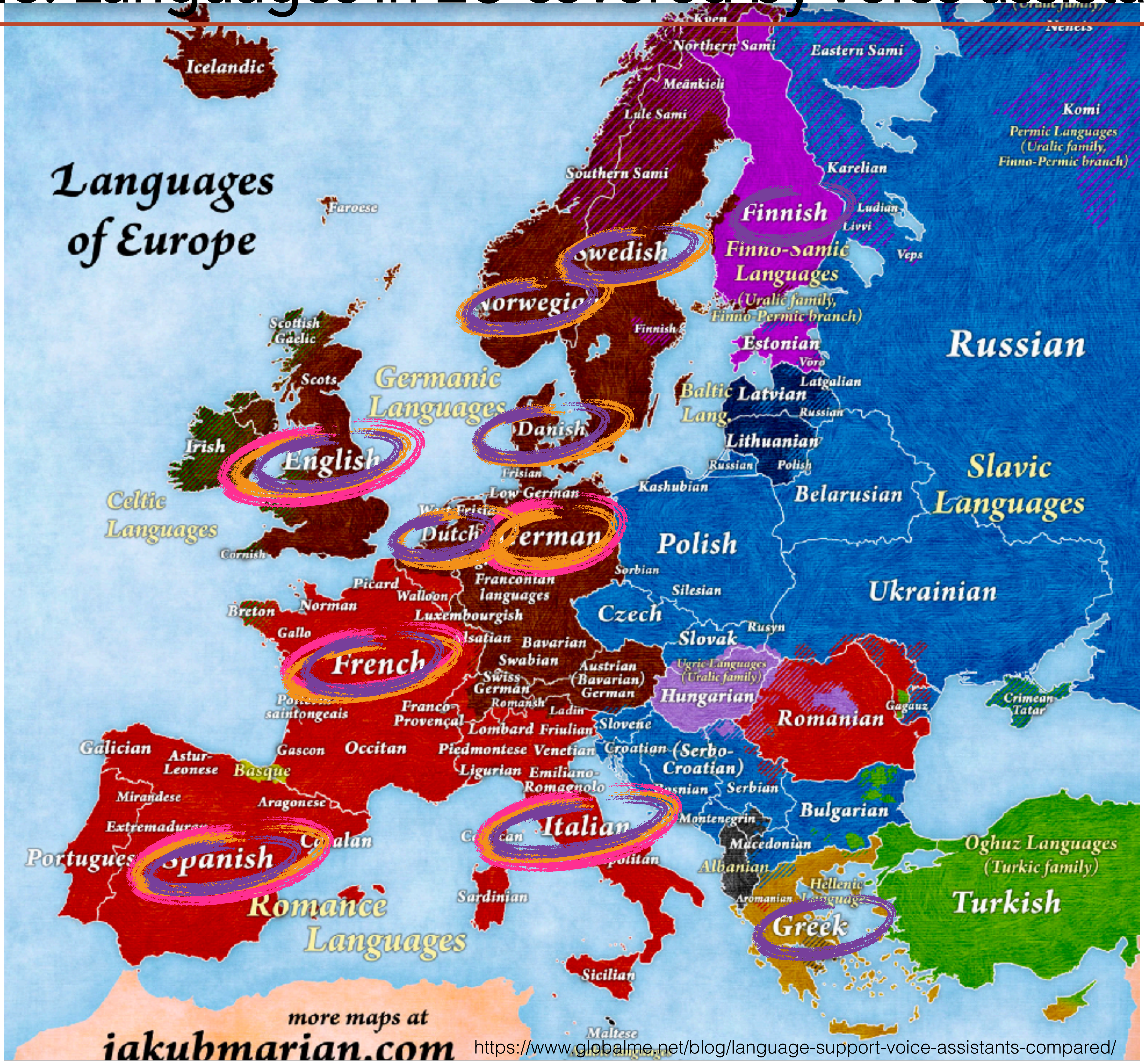
From Masked-Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-Shot Spoken Language Understanding

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



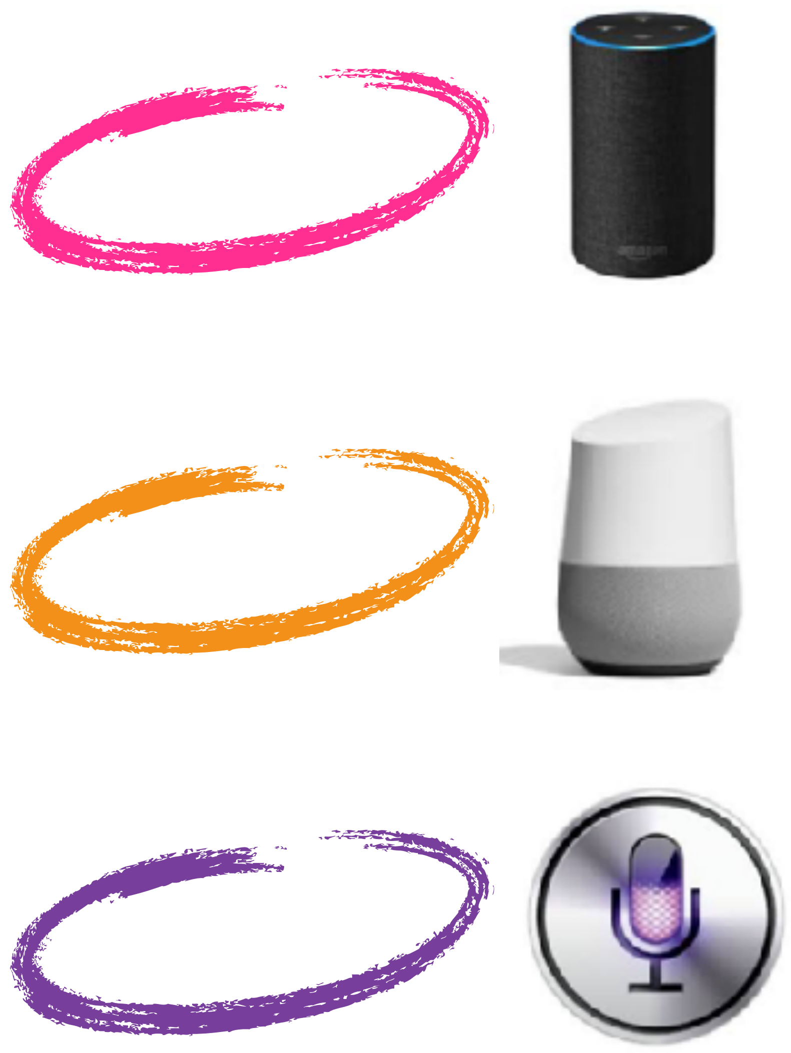
et al., NAACL 2021

Example: Languages in EU covered by voice assistants

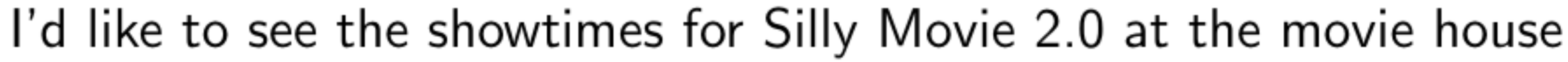


https://www.globalme.net/blog/language-support-voice-assistants-compared/

*as of March, 2020



Task: Slot and Intent Detection

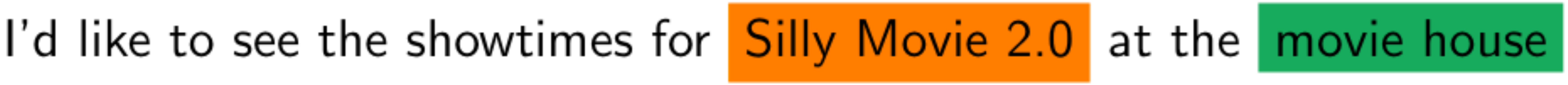


I’d like to see the showtimes for Silly Movie 2.0 at the movie house

Intent: SearchScreeningEvent

Task: Slot and Intent Detection

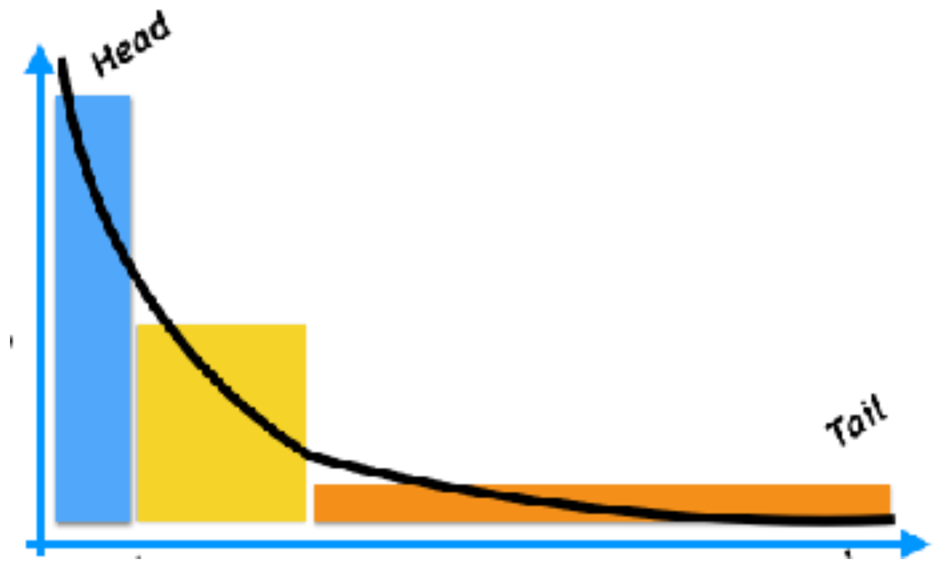
Slots:



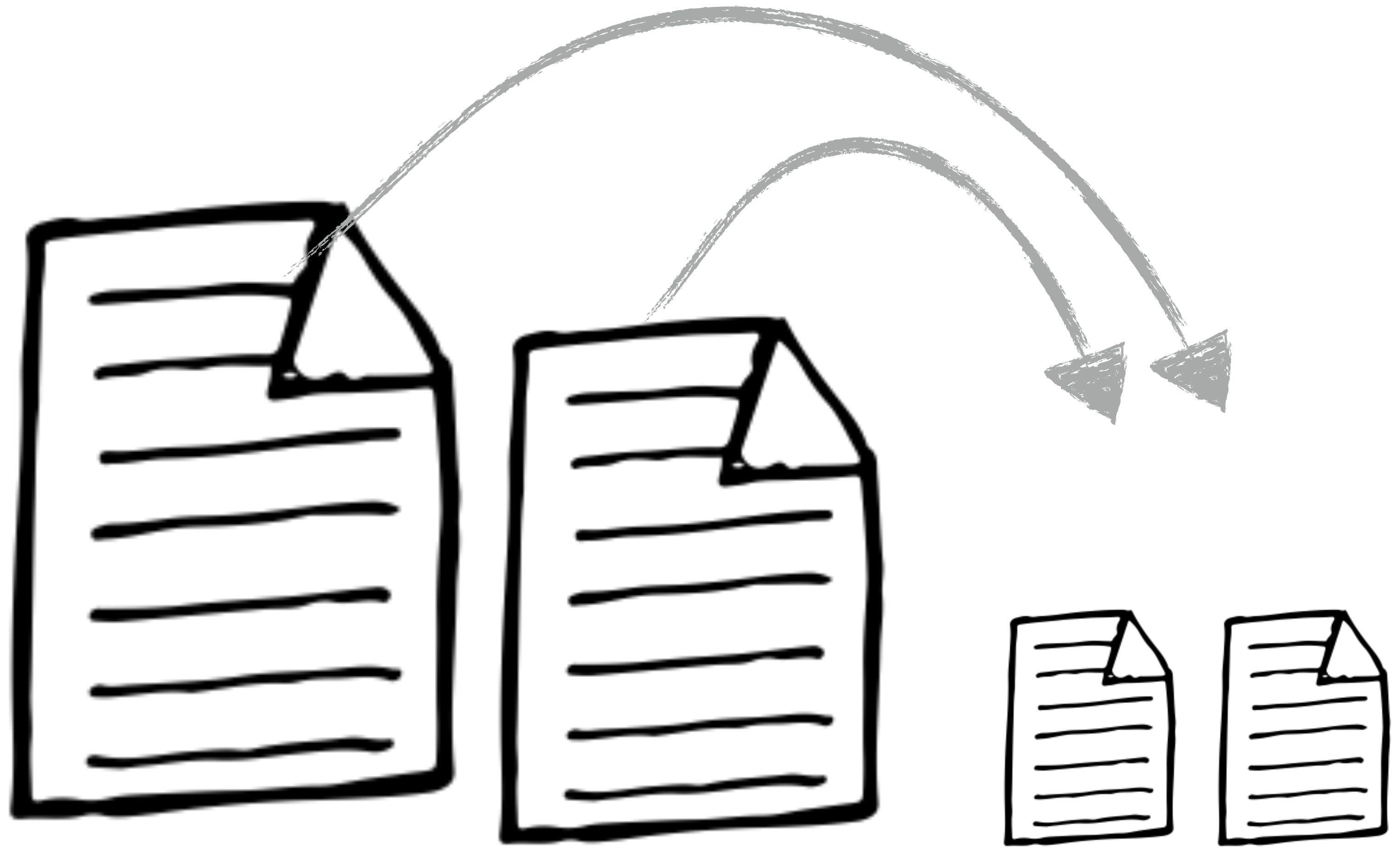
I’d like to see the showtimes for Silly Movie 2.0 at the movie house

Intent: SearchScreeningEvent

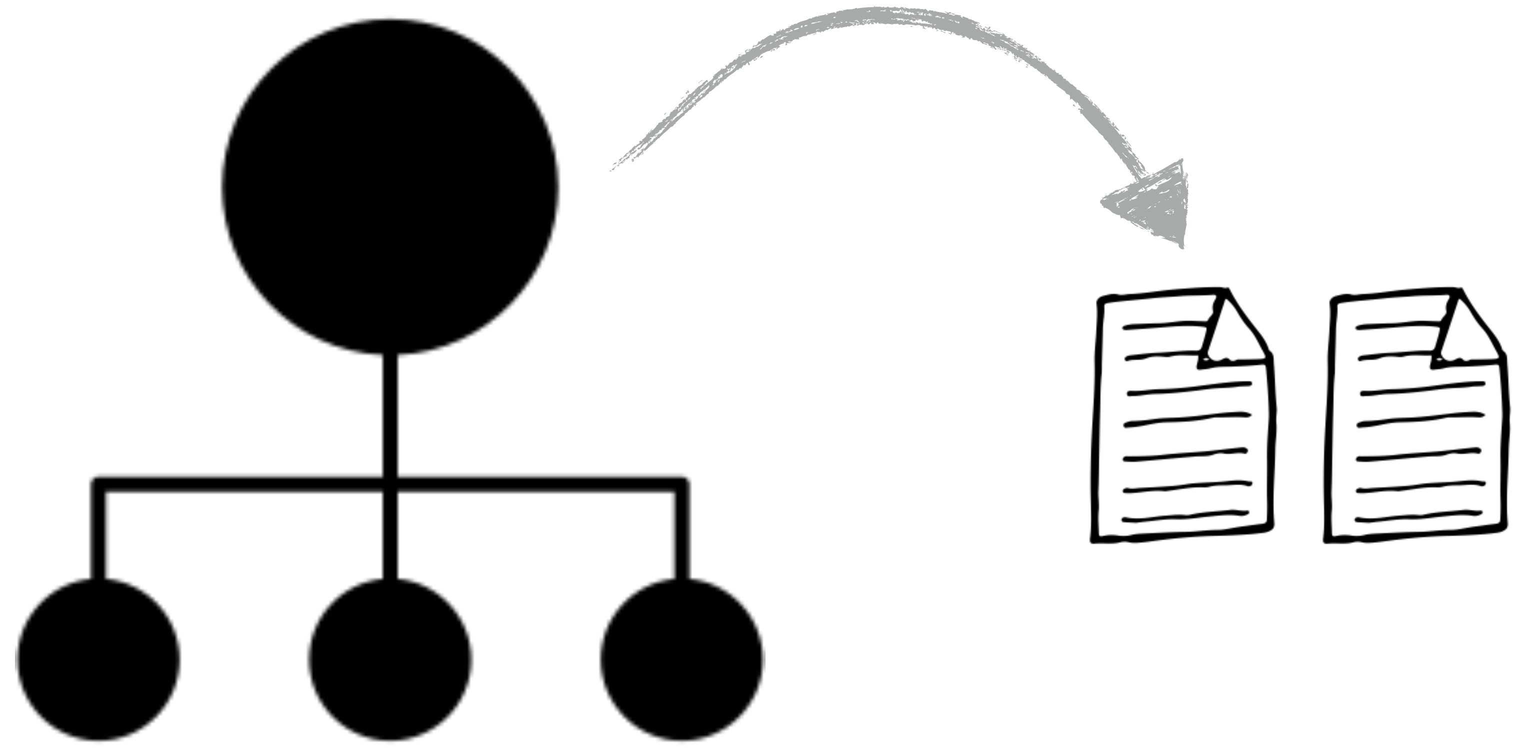
How can we transfer knowledge to low-resource languages?



Cross-lingual transfer: Two kinds of approaches

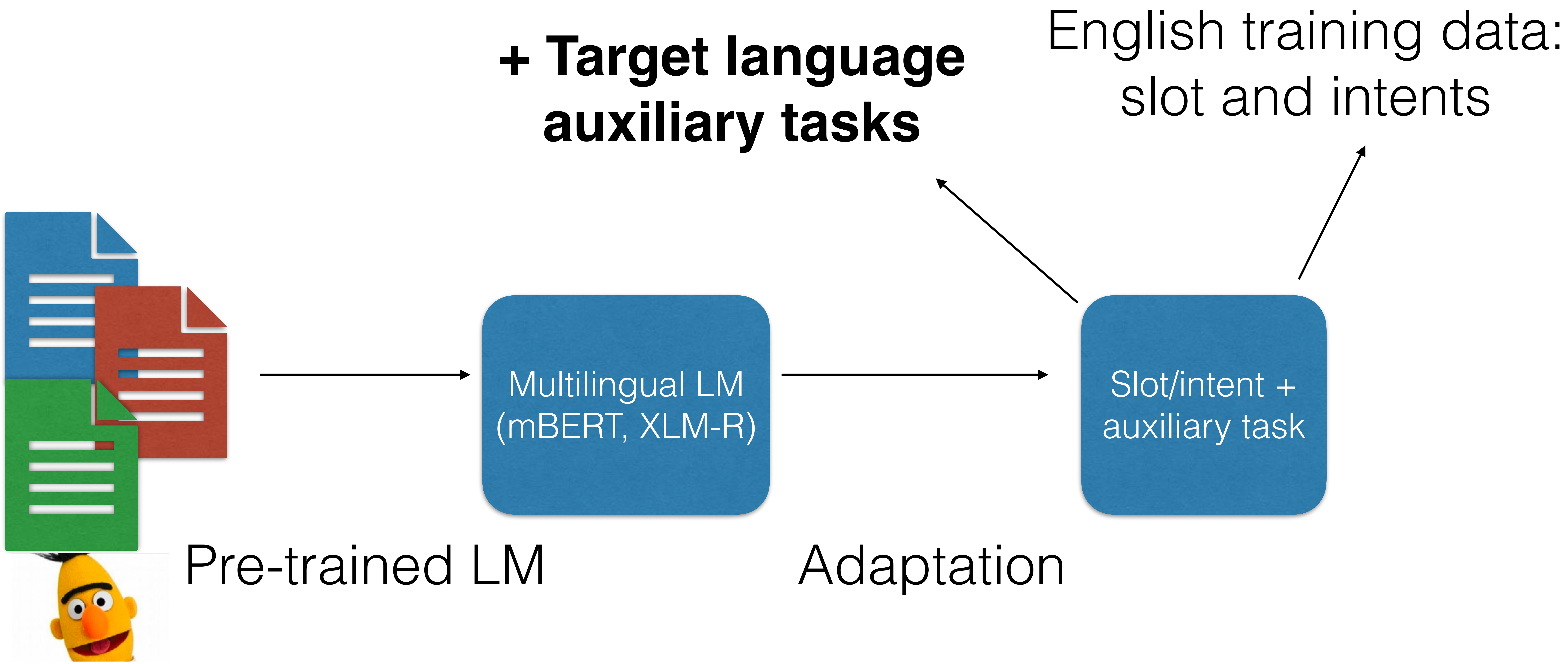


annotation transfer
(e.g. annotation projection, translation)



model transfer
(e.g. representation transfer like multilingual embeddings, delexicalization)

Idea: Non-English Auxiliary Tasks



+ Target language auxiliary tasks

English training data: slot and intents

Multilingual LM (mBERT, XLM-R)

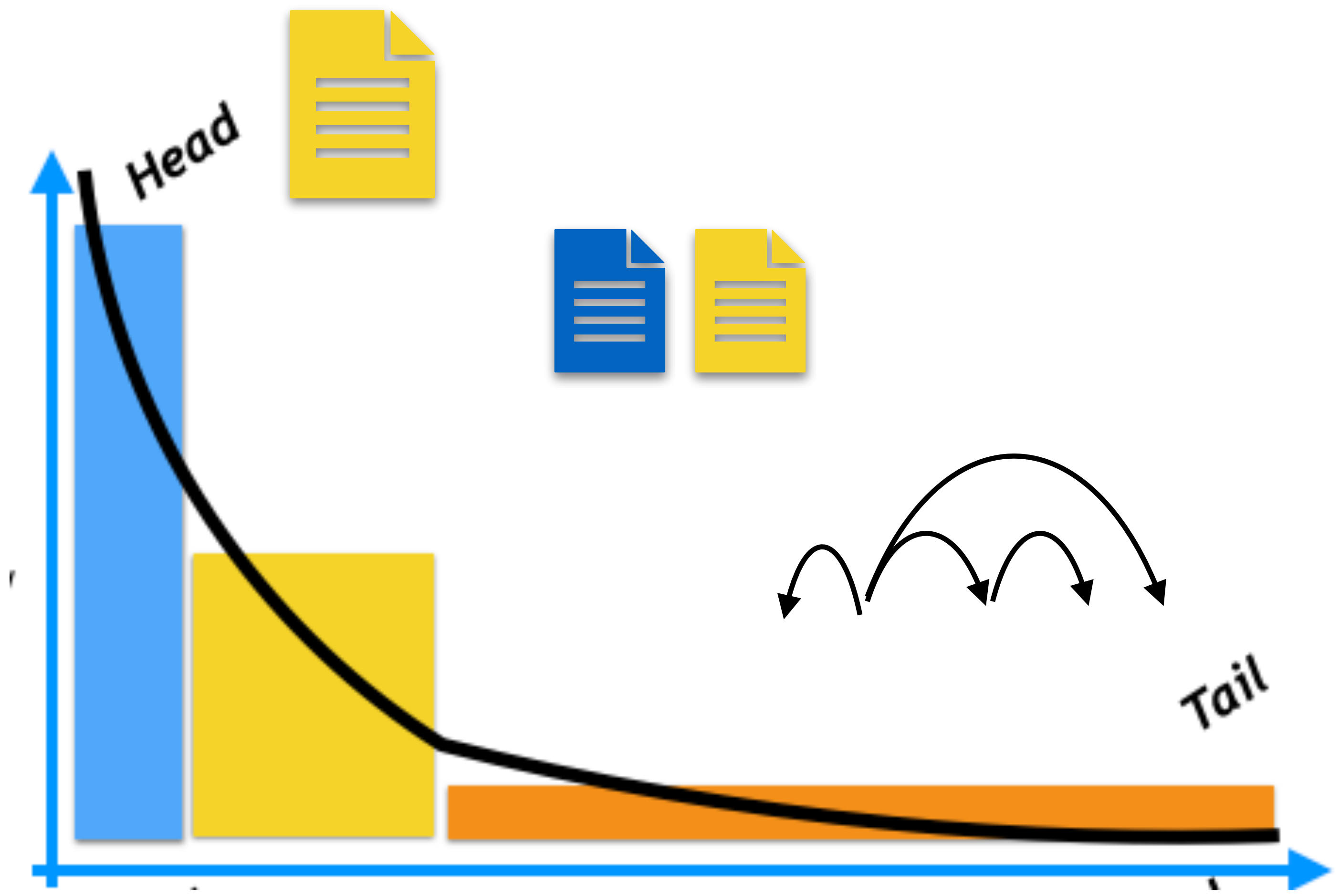
Slot/intent + auxiliary task

Pre-trained LM

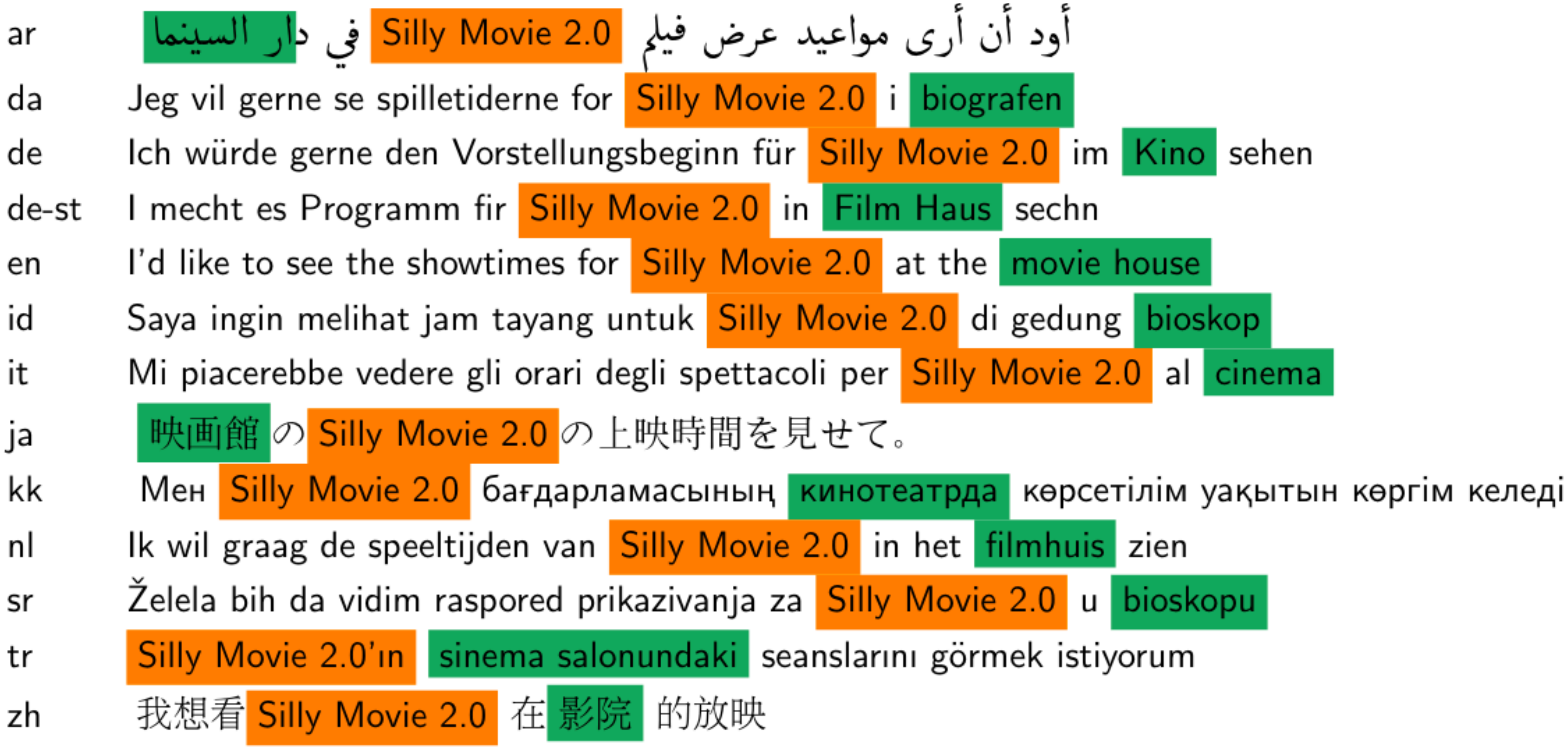
Adaptation

Non-English Auxiliary Tasks

- **Raw data:** Masked language modelling (aux-mlm)
- **Parallel data:** Neural machine translation (aux-nmt)
- **Parsing data:** UD parsing (aux-ud)



New 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

Experiments

- Baselines:
  - Baseline (mBERT): joint intent + slot prediction (MaChAmP, van der Goot et al., 2021)
  - Strong baseline (nmt-transfer): NTM (translate training data to target language) + annotation projection (map slots with attention)

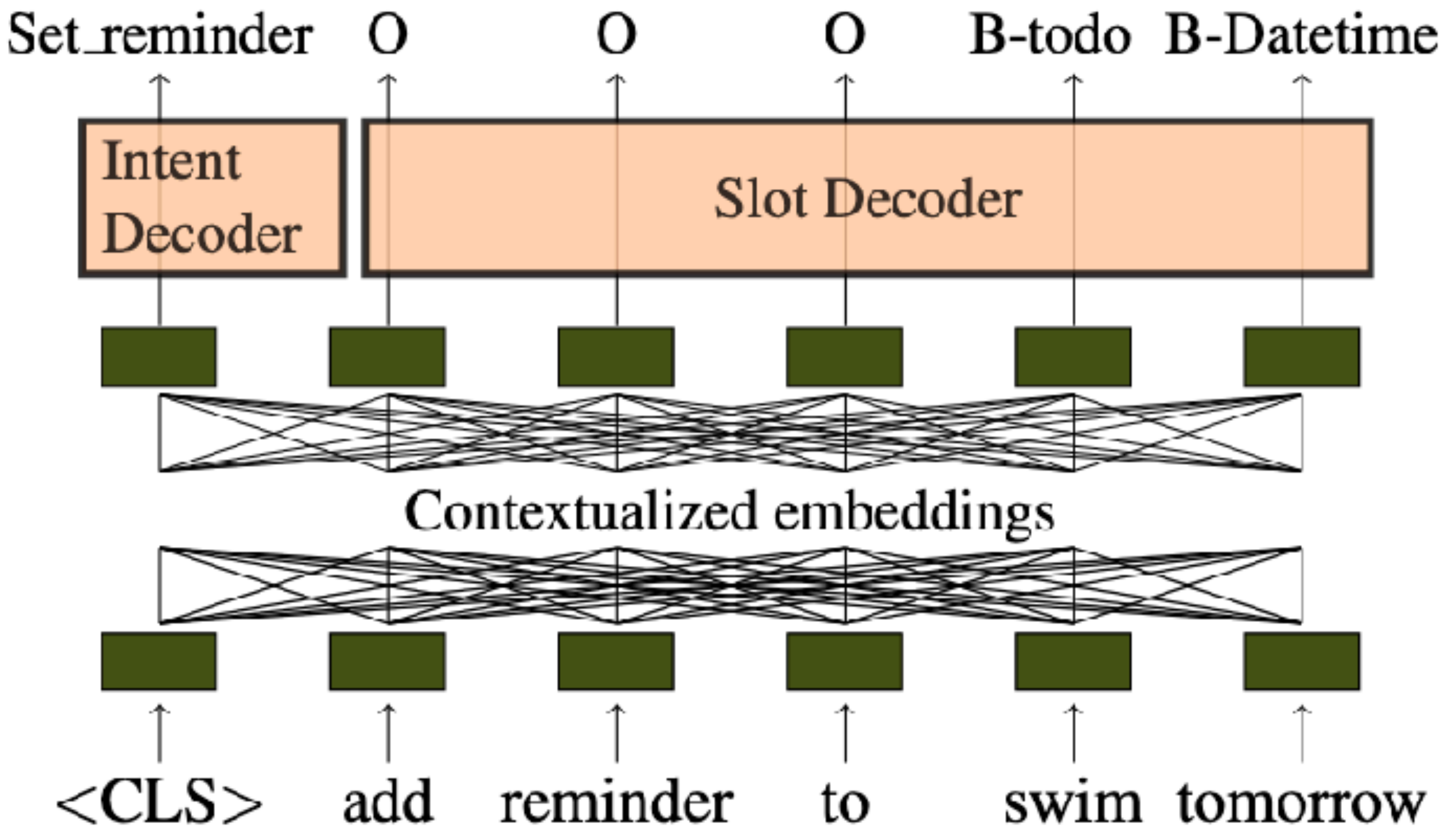
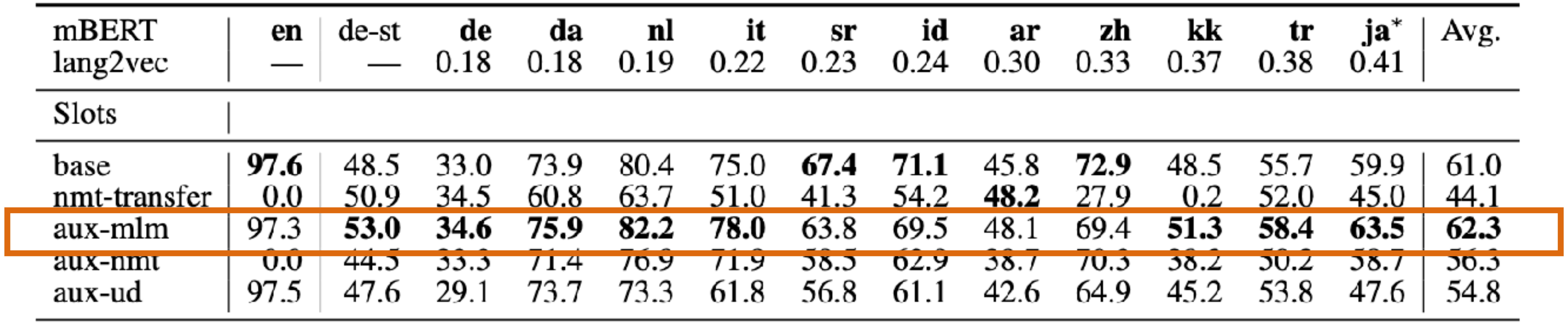


Figure 2: Overview of the baseline model.

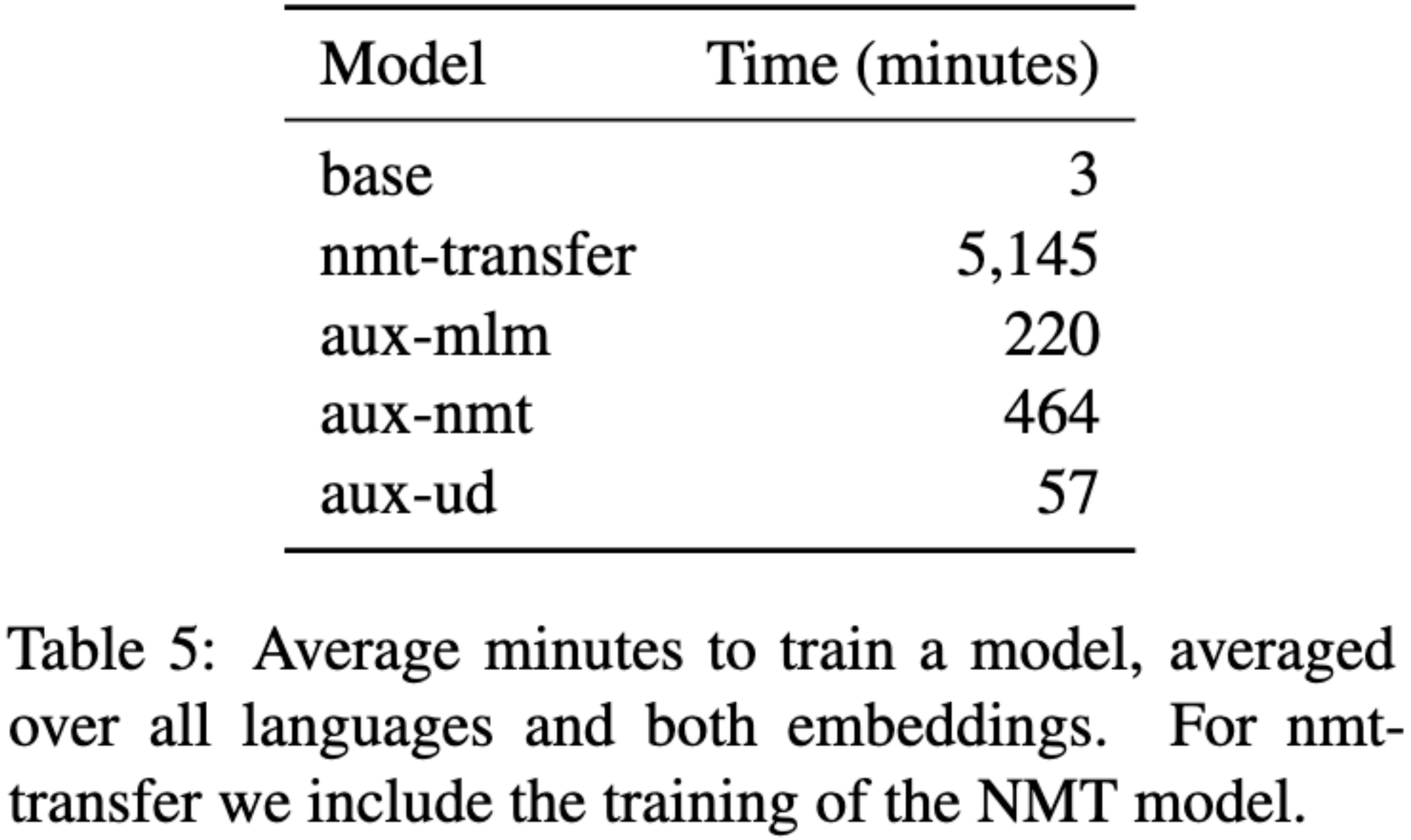
Results on Slots - Main take-away



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)

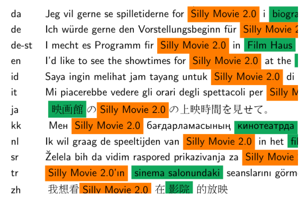
How much training resources (time)?



Model	Time (minutes)
base	3
nmt-transfer	5,145
aux-mlm	220
aux-nmt	464
aux-ud	57

Table 5: Average minutes to train a model, averaged over all languages and both embeddings. For nmt-transfer we include the training of the NMT model.

Take-aways



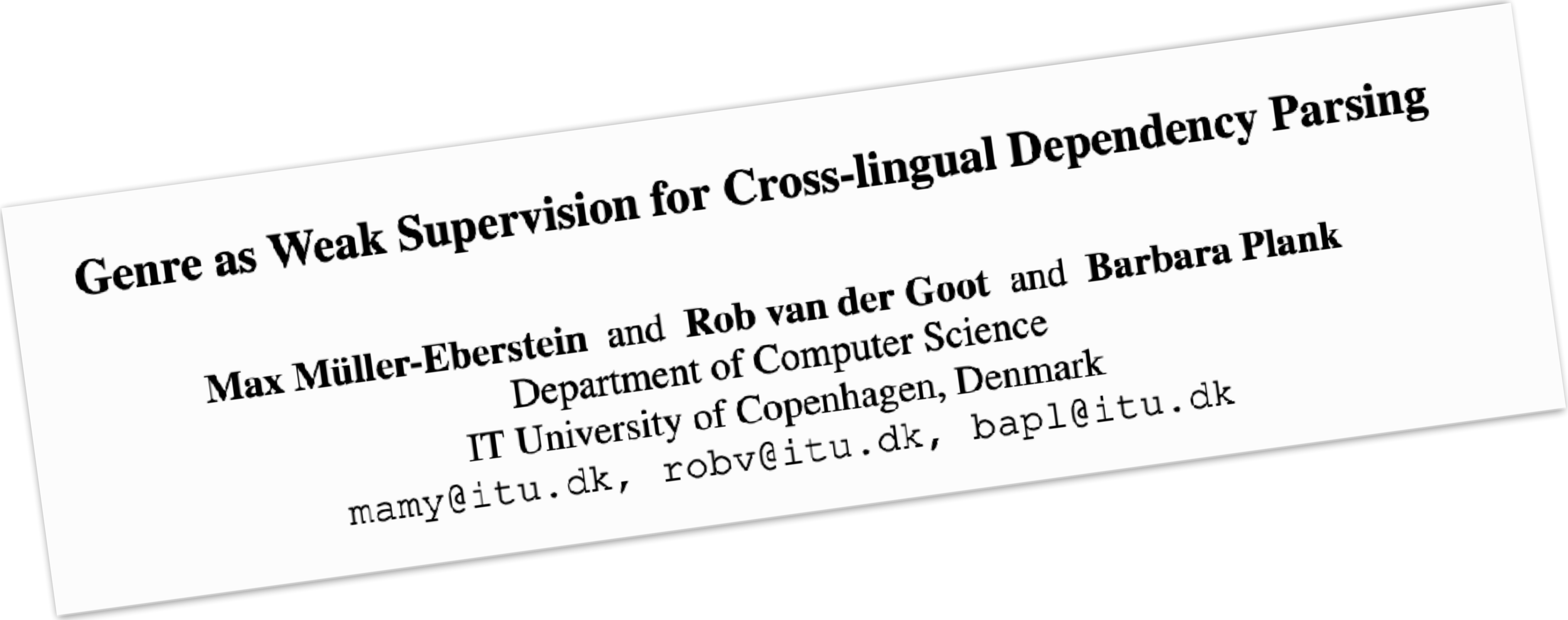
da	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biogra
de	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2
de-st	I mecht es Programm fir Silly Movie 2.0 in Film Haus
en	I'd like to see the showtimes for Silly Movie 2.0 at the
id	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di
it	Mi piacerebbe vedere gli orari degli spettacoli per Silly M
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
nl	Ik wil graag de speeltijden van Silly Movie 2.0 in het fil
sr	Želela bih da vidim raspored prikazivanja za Silly Movie
tr	Silly Movie 2.0'ın sinema salonundaki seanslarını görm
zh	我想看 Silly Movie 2.0 在影院的放映

- Slot and Intent Detection dataset (xSID) and annotation guidelines released, xSID is growing: Bernese Swiss German and Neapolitan added in VarDial (Aepli et al. 2023)

  ⭐ Let us know if you would like to contribute a new language variant!

- MLM auxiliary task was most robust (similar to DAPT but across languages), and help particularly for a low-resource dialect (South Tyrolean)

- Limitation: sharing via MTL helped only in limiting degrees

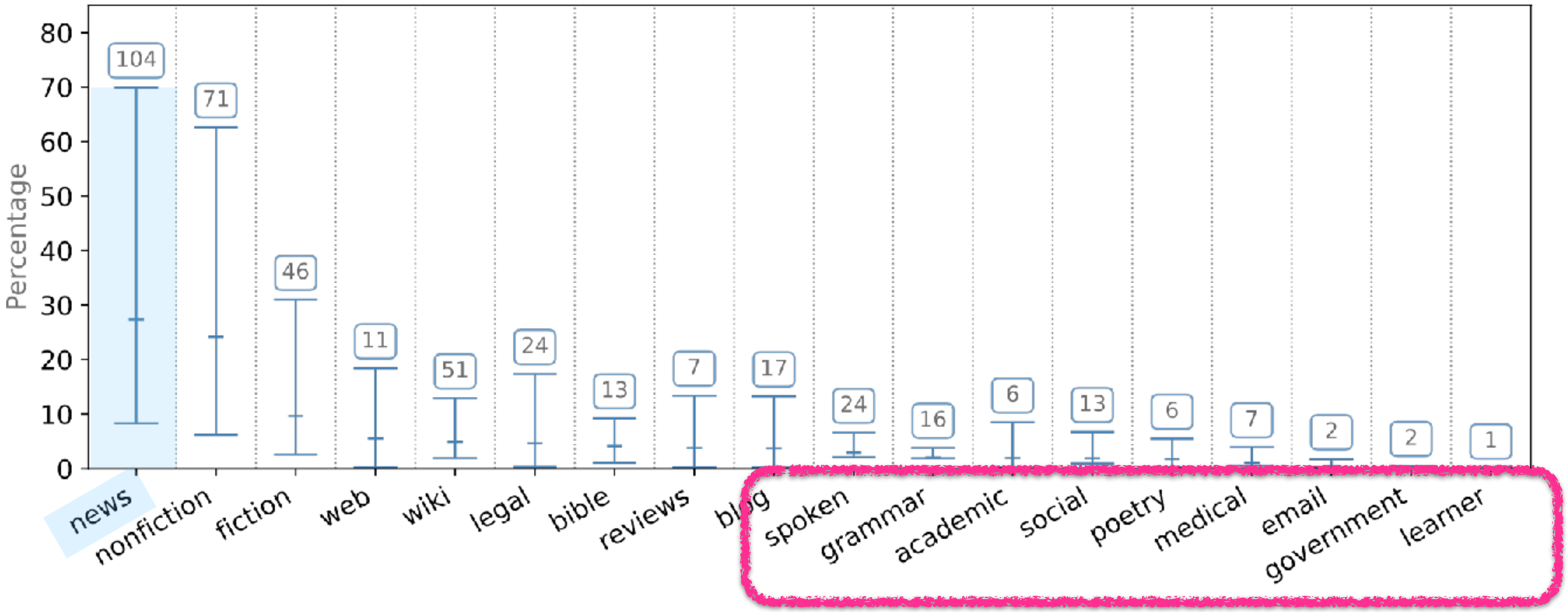


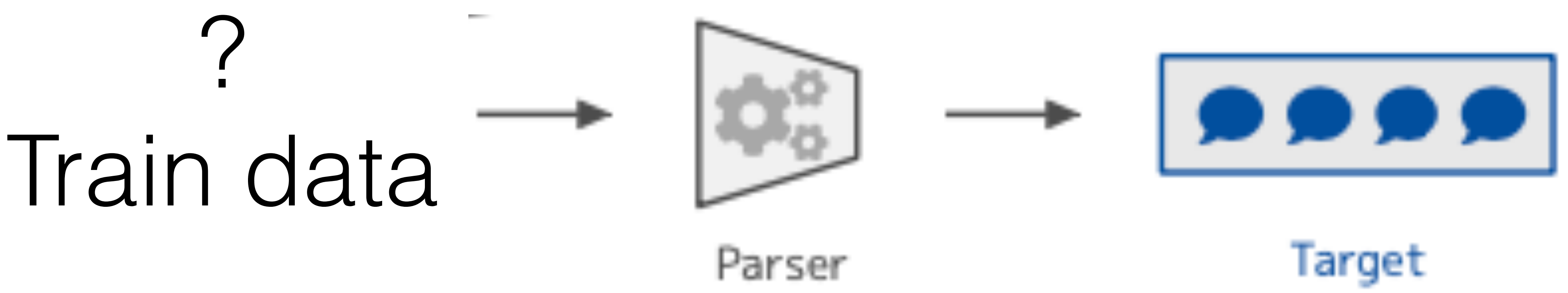
Genre as Weak Supervision for Cross-lingual Dependency Parsing

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

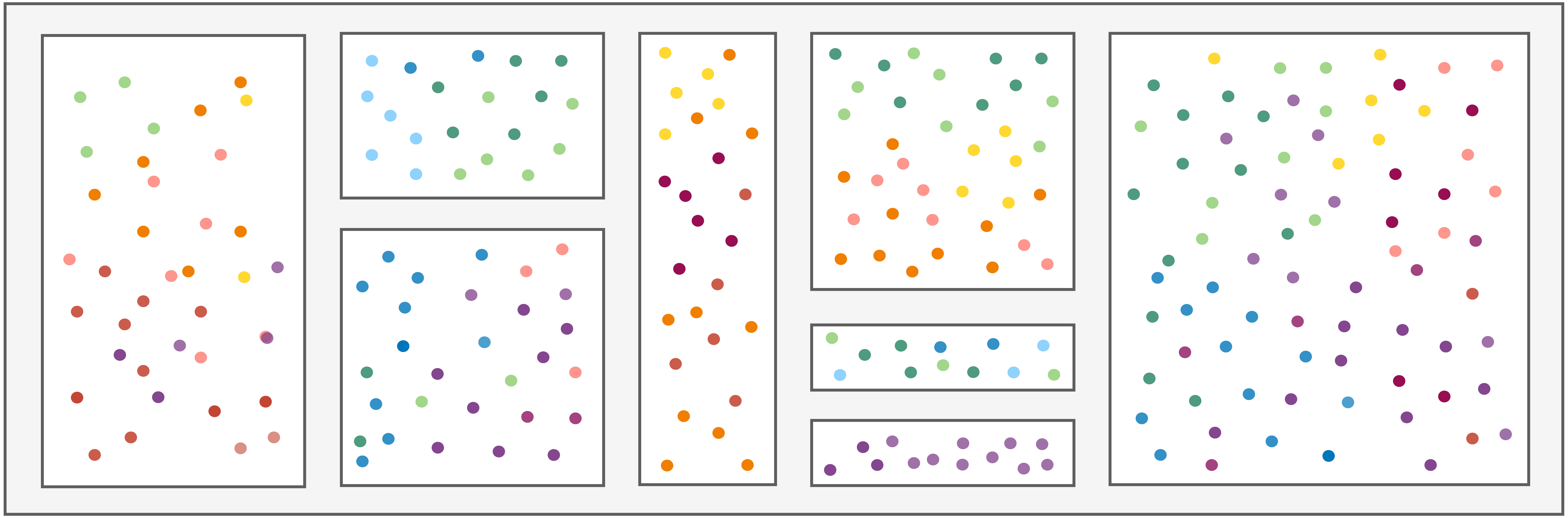
Genre Distribution in Universal Dependencies (UD)




?
Train data

Universal Dependencies

Müller-Eberstein, van der Goot, and Plank (2021b)



200 TREEBANKS **114** LANGUAGES **1.51M** SENTENCES

Nivre et al. (2020); Statistics as of version 2.8

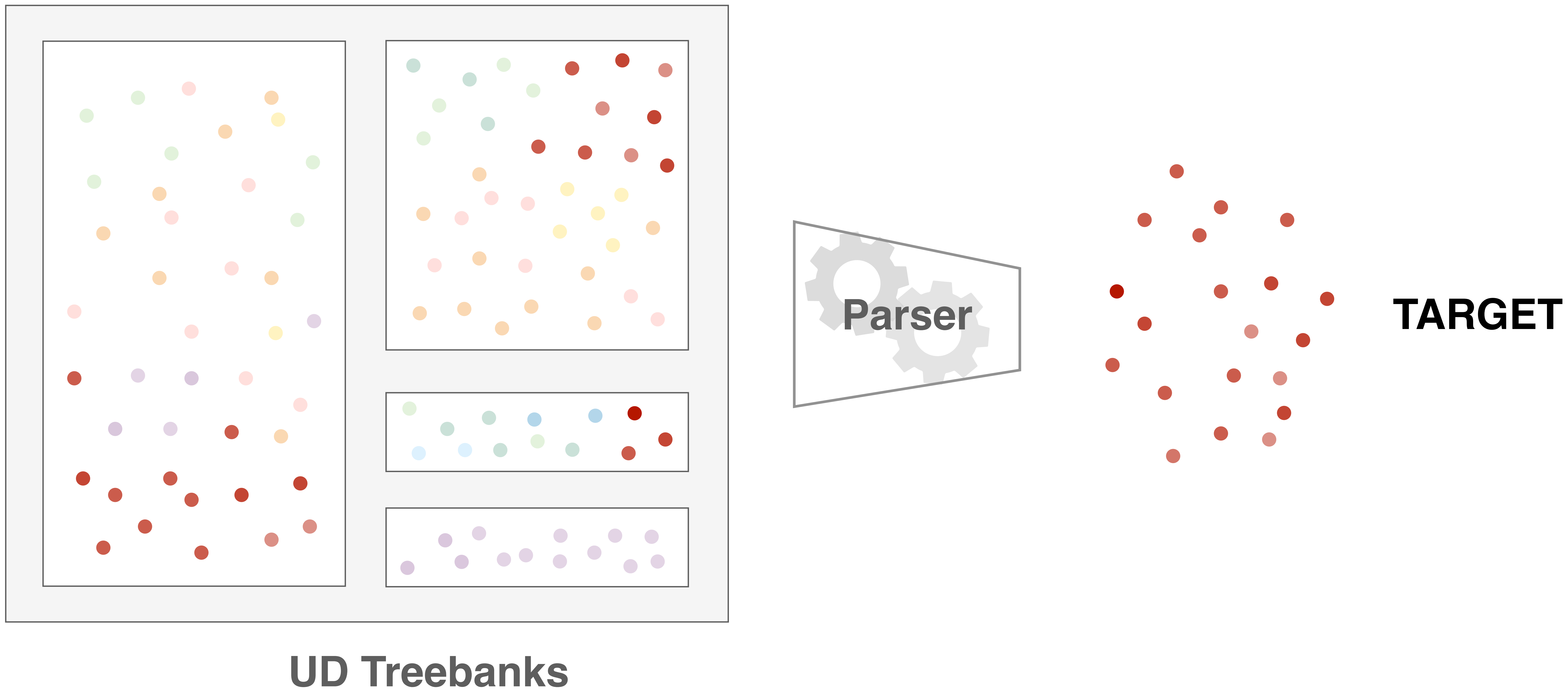
Universal Dependencies Genre Meta-data

What's (not) in a corpus?

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

18 GENRES

Nivre et al. (2020); Statistics as of version 2.8



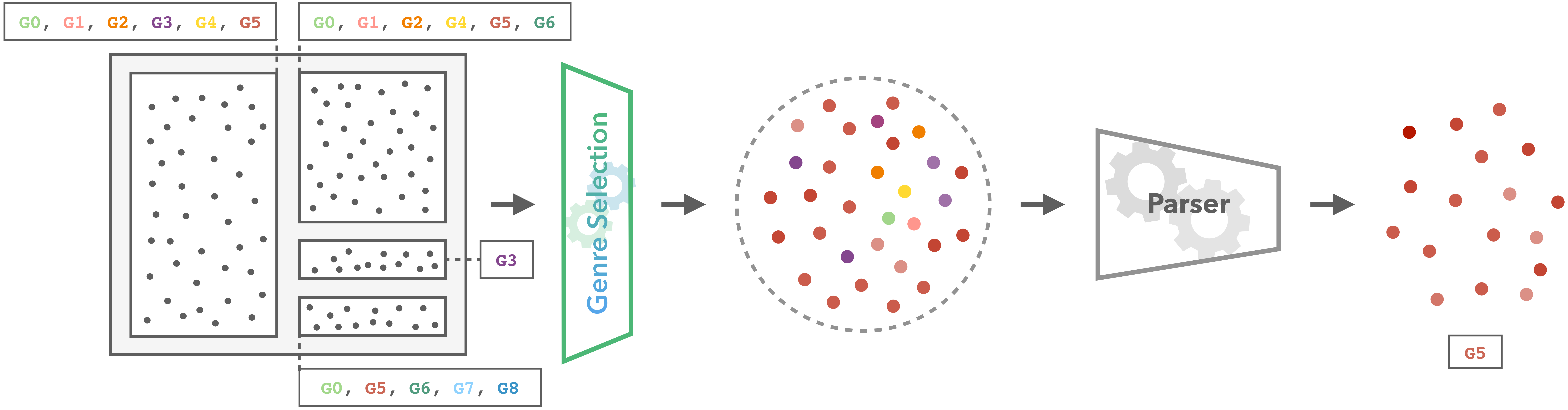
UD Treebanks

Parser

TARGET

Genre as Weak Supervision for Cross-lingual Dependency Parsing

Müller-Eberstein, van der Goot, and Plank (2021a)



Universal Dependencies
(no instance genre labels)

Proxy Data
(weakly genre-labelled)

Target Data
(zero-shot language)

Dozat and Manning (2017)

Genre as Weak Supervision for Cross-lingual Dependency Parsing

Sort by size (lowest first).

TARGET   SWL 💬   SA 📕   KPV 📕   TA 📰   GL 📰   YUE 💬   CKT 💬   FO W   TE 🪄   MYV 📕   QHE ⚟   QTD 💬

Data Selection Results

Less is more.

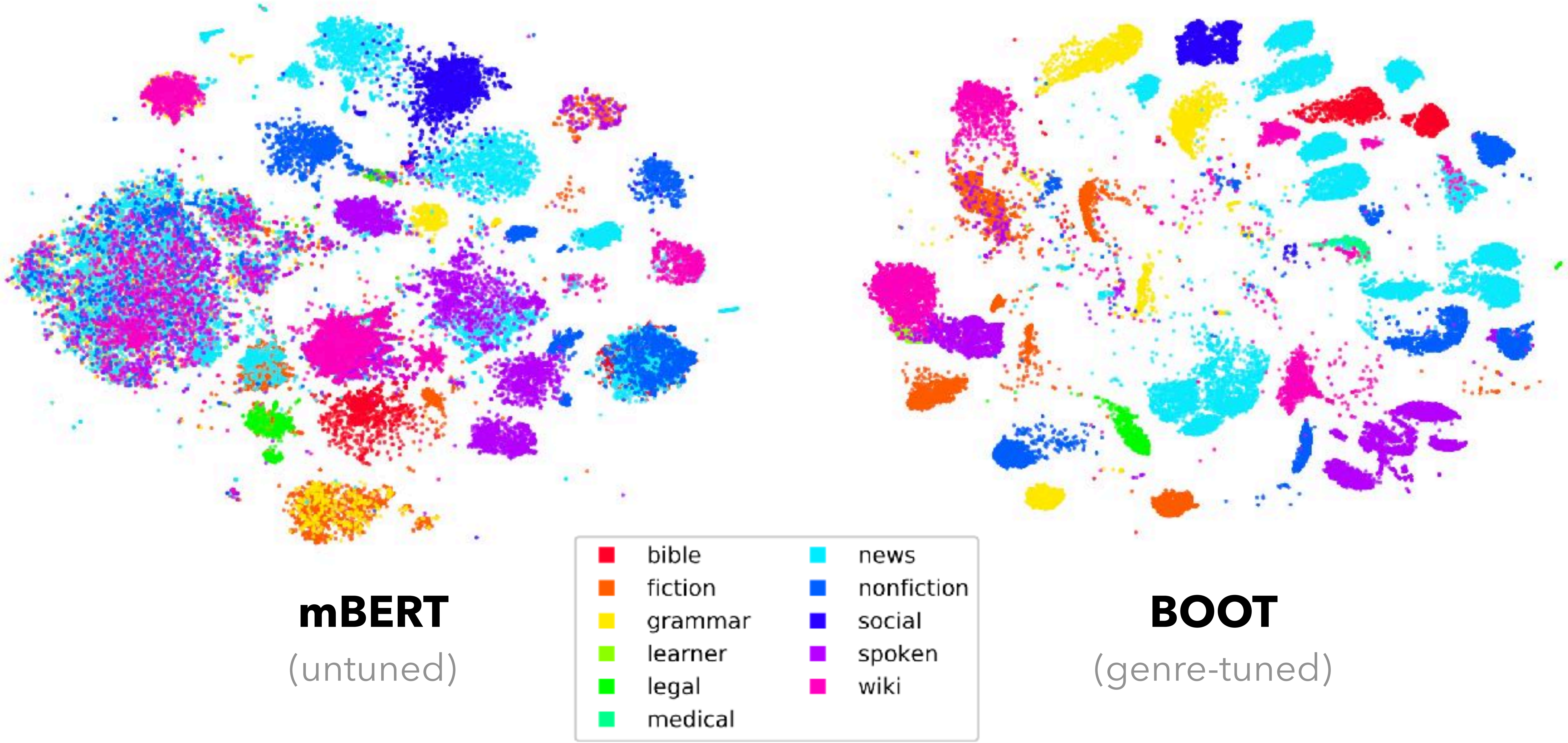
8x more data

Projected instance genre for best adaption

TARGET	META	RAND	SENT	BOOT	GMM LDA
50.3	34.1	36.5	36.8	37.7	**38.7**

Genre as Weak Supervision for Cross-lingual Dependency Parsing

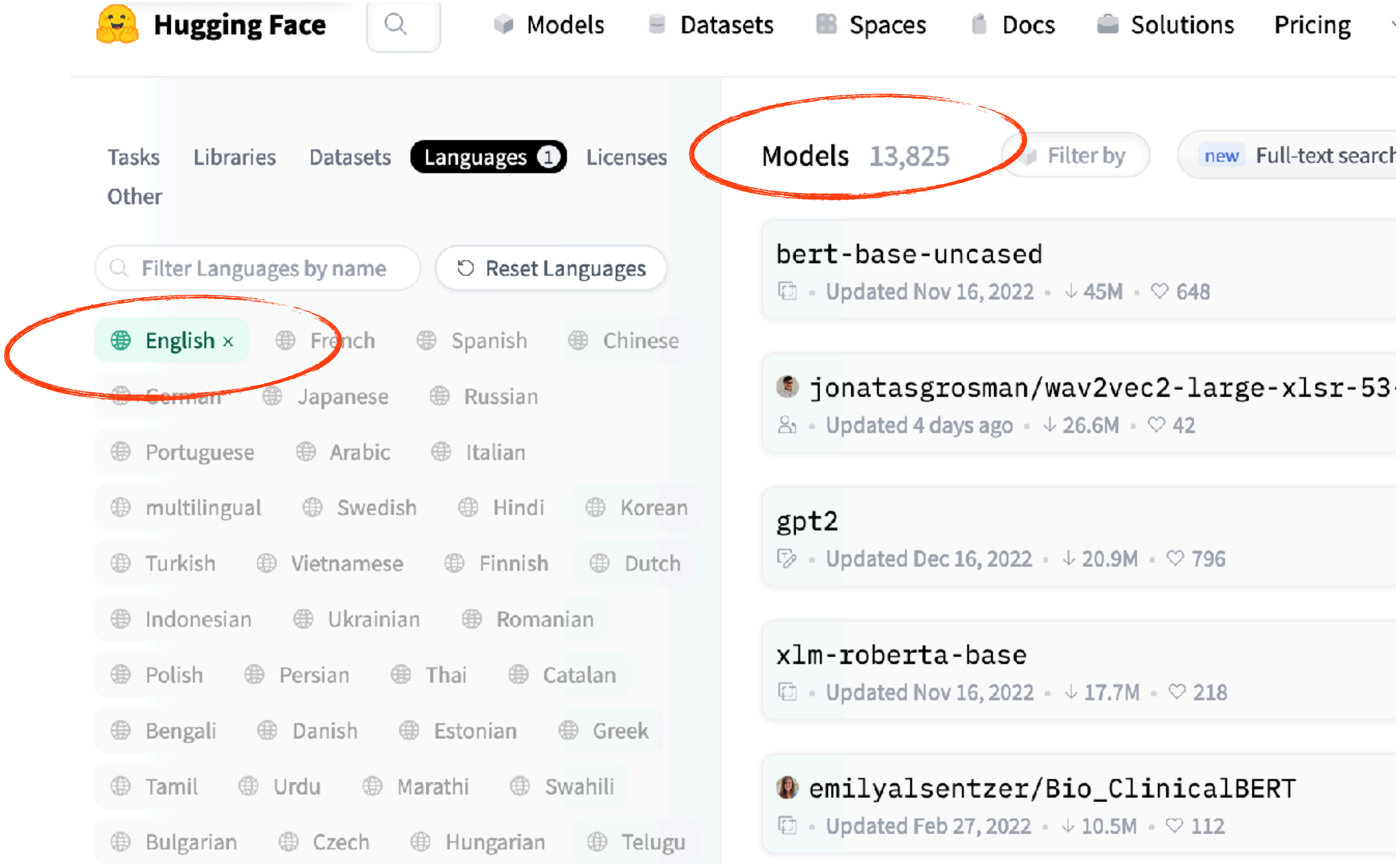
Left: genre in mBERT. Right: genre-tuned mBERT via weak supervision.



Applications to Transferability Estimation

Subspaces for Performance Prediction

Which Large Pre-Trained LM to pick?



Models 13,825

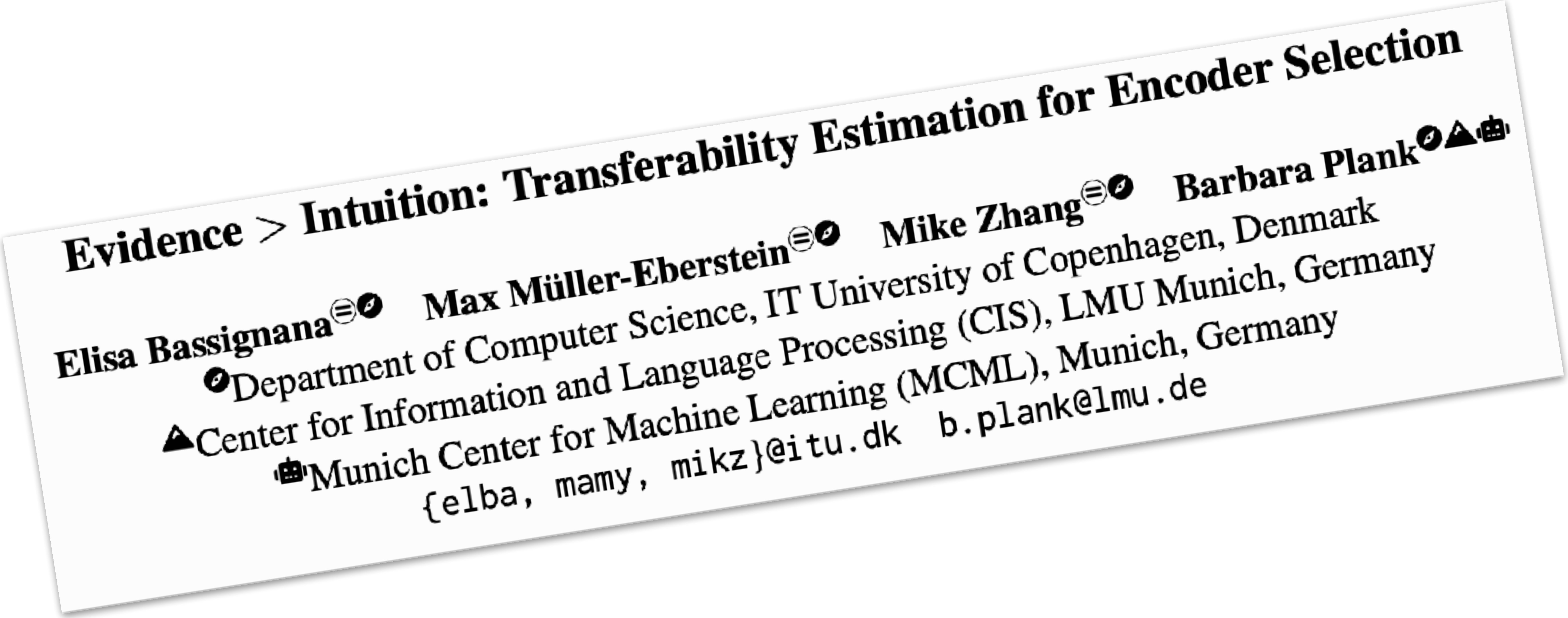
bert-base-uncased
Updated Nov 16, 2022 • 45M • 648

jonatasgrosman/wav2vec2-large-xlsr-53
Updated 4 days ago • 26.6M • 42

gpt2
Updated Dec 16, 2022 • 20.9M • 796

xlm-roberta-base
Updated Nov 16, 2022 • 17.7M • 218

emilyalsentzer/Bio_ClinicalBERT
Updated Feb 27, 2022 • 10.5M • 112



Evidence > Intuition: Transferability Estimation for Encoder Selection

Elisa Bassignana⊜● Max Müller-Eberstein⊜● Mike Zhang⊜● Barbara Plank●▲🤖

●Department of Computer Science, IT University of Copenhagen, Denmark
▲Center for Information and Language Processing (CIS), LMU Munich, Germany
🤖Munich Center for Machine Learning (MCML), Munich, Germany
`{elba, mamy, mikz}@itu.dk` `b.plank@lmu.de`

EMNLP, 2022

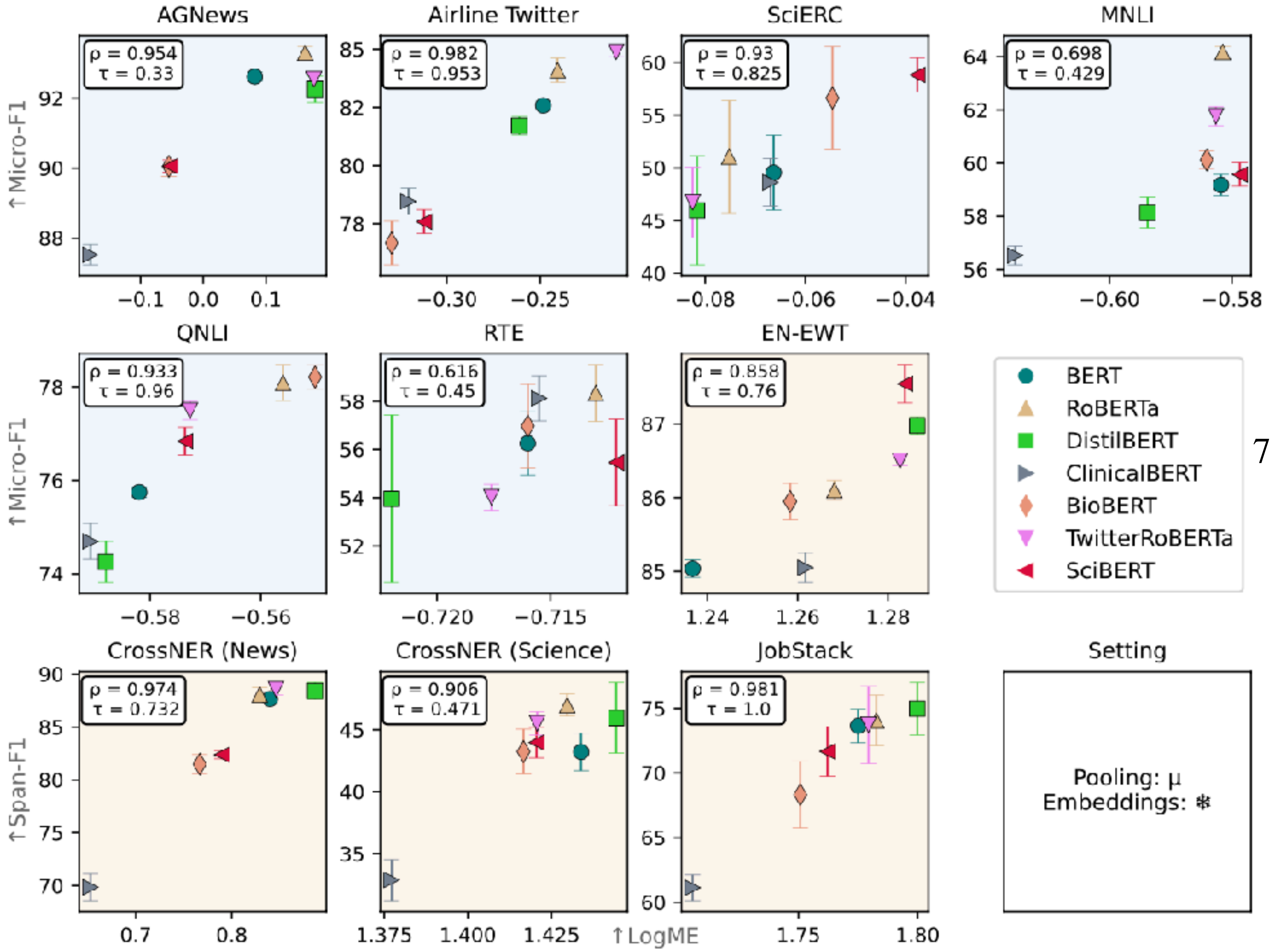
Which Large Pre-Trained LM to pick?

- Problem: LLMs are appearing at an incredible pace. It becomes increasingly difficult to pick a pre-trained LM
    - Fine-tuning with all is infeasible (and not sustainable)
    - Today's LM choice is largely based on heuristics
- **Question**: Given an NLP task, to what extent can we estimate the transferability of pre-trained LMs to specific NLP tasks, a-priori (without fine-tuning?)
- Prior work on this in NLP is limited; Some distantly related work on performance prediction not on LLM choice though (e.g. Xia et al., 2020; Ye et al., 2021)

Transferability Estimation

- Problem setup: Given L pre-trained language models and a dataset D, estimate a score for each language model without fine-tuning on D
  - Use the obtained rank to select the best LLM encoder
  - As ranking function, we use the LogMe framework proposed in Computer Vision (You et al., 2021) - an iterative process that draws lightly parametrised Gaussian distributions to estimate the fit of the LM to the dataset D
- We evaluate model ranking across 10 tasks of two kinds (classification, structured prediction) using 4 setups and 7 LLMs (general, domain-specific)
- We compare it to human experts (12 NLP researchers)

Transferability Estimation: Results



7 LMs

X-axis: LogMe score

Y-axis: Task performance

Blue: classification tasks, Orange: sequence labelling tasks

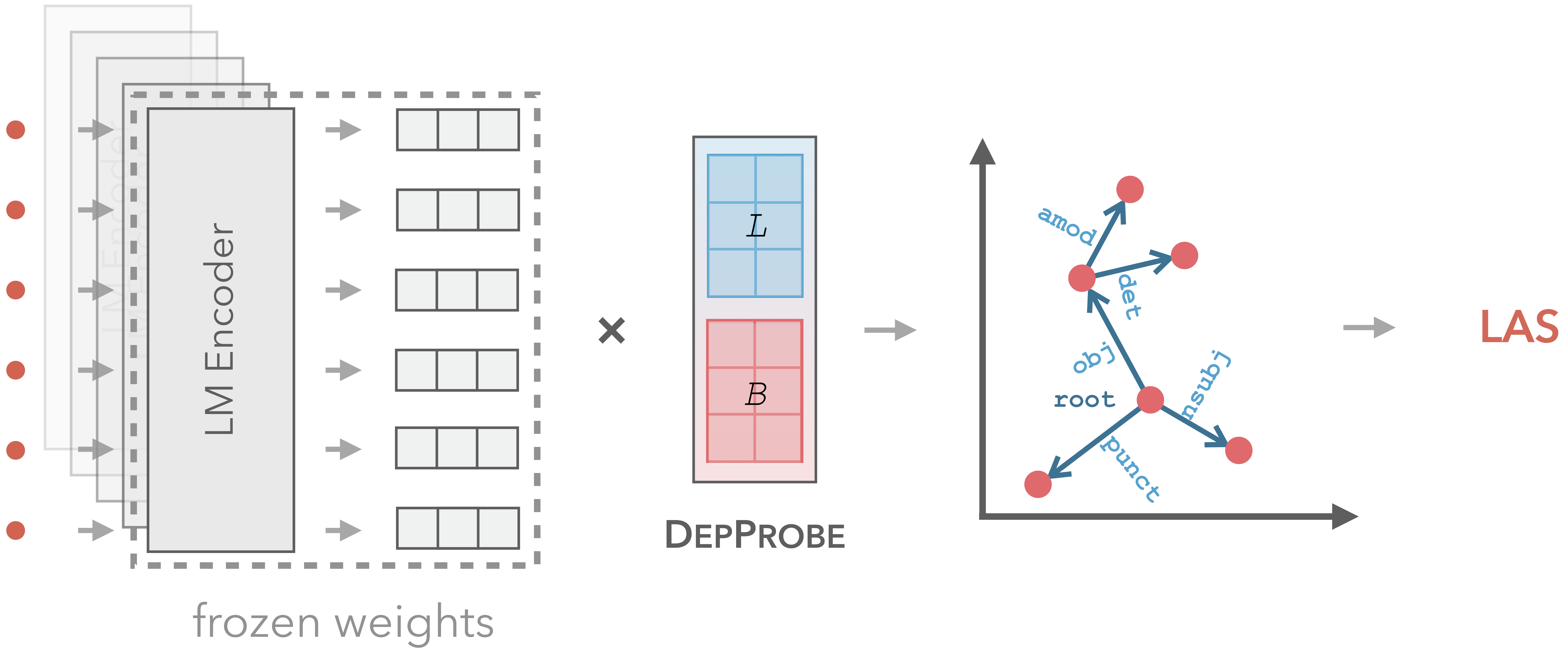
Vs Human Performance

- Task turns out to be difficult for humans
  - No single participant was the expert in all setups
- Wider range of correlation:
  - LogME range of τ is in [−0.20; 1.00]; Human rankings fall into a wider range of [−0.54; 1.00], higher uncertainty.
- Benefit of LogMe: provides a continuous scale, humans ranks offer no indication of relative performance differences
- Take-Away: Evidence > human-intuition for a-priori LM ranking
- Limitation: limited (12) human rankings, generalisability beyond the task sample?

What about dependency parsing?

Probing for labeled dependency trees

Müller-Eberstein, van der Goot, and Plank (2022b)



Sort by Structure: Language Model Ranking as Dependency Probing

9 languages, 22 LMs, 46 setups.

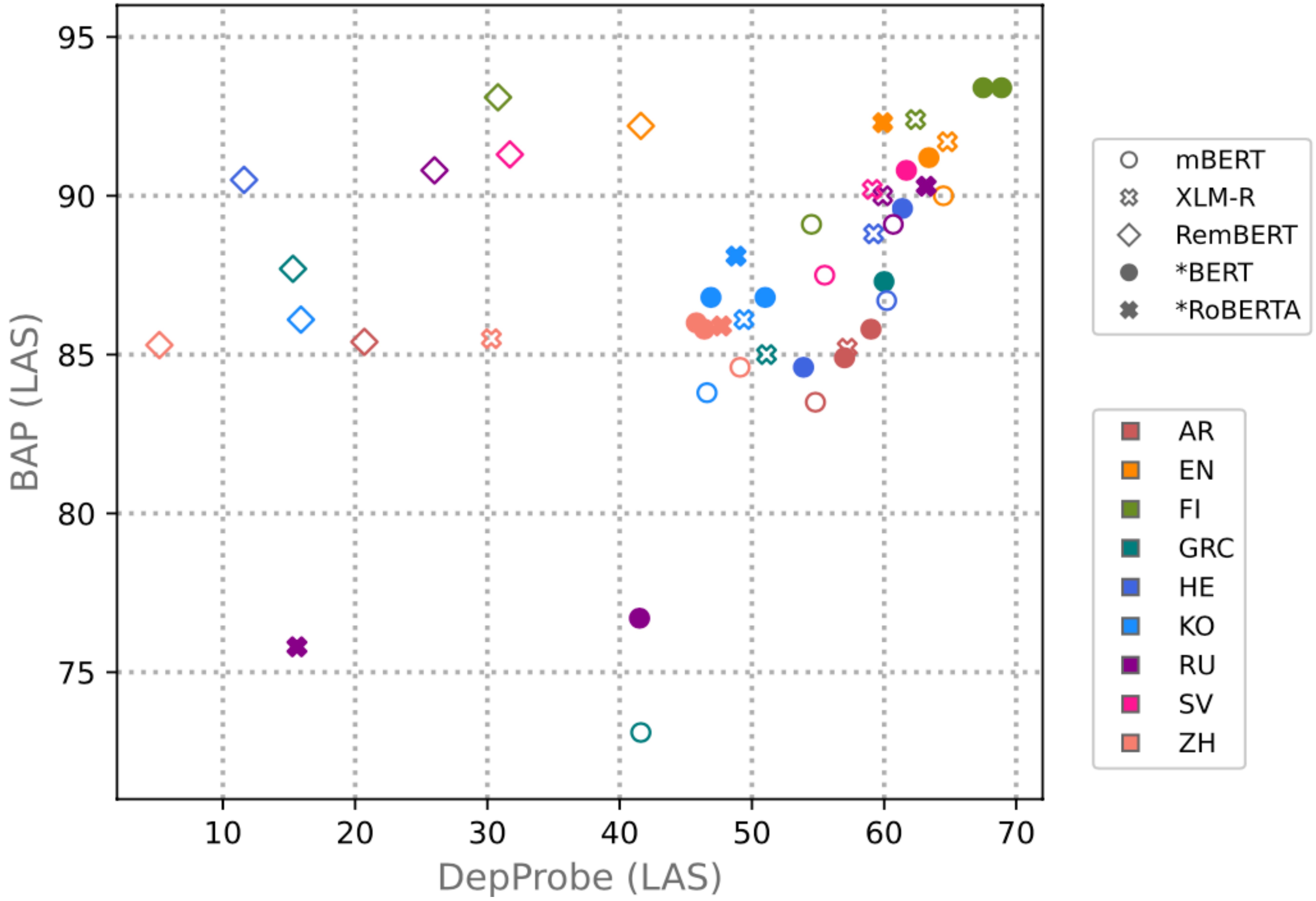
Arabic	English	Finnish	Anc. Greek	Hebrew	Korean	Russian	Swedish	Chinese
mBERT	mBERT	mBERT	mBERT	mBERT	mBERT	mBERT	mBERT	mBERT
XLM-R	XLM-R	XLM-R	XLM-R	XLM-R	XLM-R	XLM-R	XLM-R	XLM-R
RemBERT	RemBERT	RemBERT	RemBERT	RemBERT	RemBERT	RemBERT	RemBERT	RemBERT
AraBERT	BERT	BERT-FI	BERT-GRC	א-BERT	BERT-KO	RuBERT	BERT-SV	BERT-ZH
BERT-AR	RoBERTA	BERT-fi	BERT-EL		RoBERTA-KO	RuBERTa		BERT-ZH WWM
					BERT-KOR	RoBERTA-RU		RoBERTA-ZH WWM

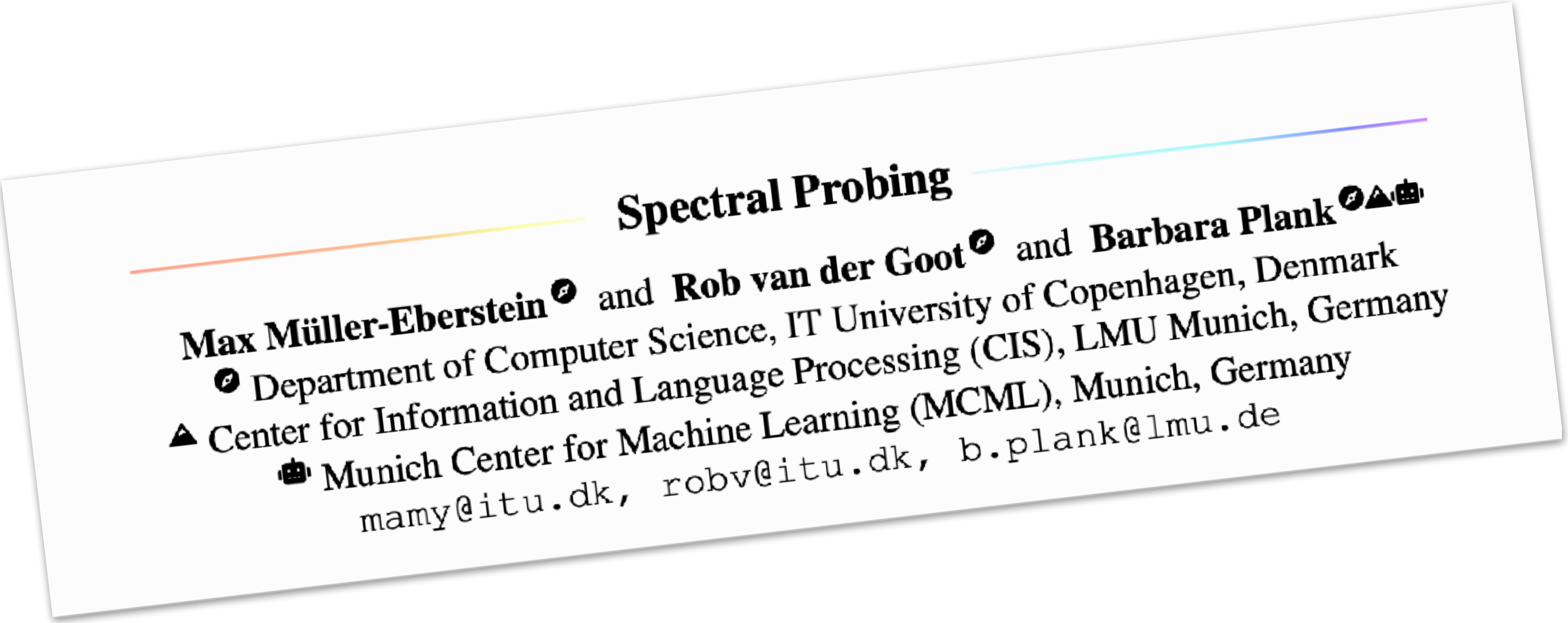
Sort by Structure: Language Model Ranking as Dependency Probing

LAS of DEPPROBE in relation to BAP

Predictive Power

τ_w .58 → 79%





Spectral 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

🤖 Munich Center for Machine Learning (MCML), Munich, Germany

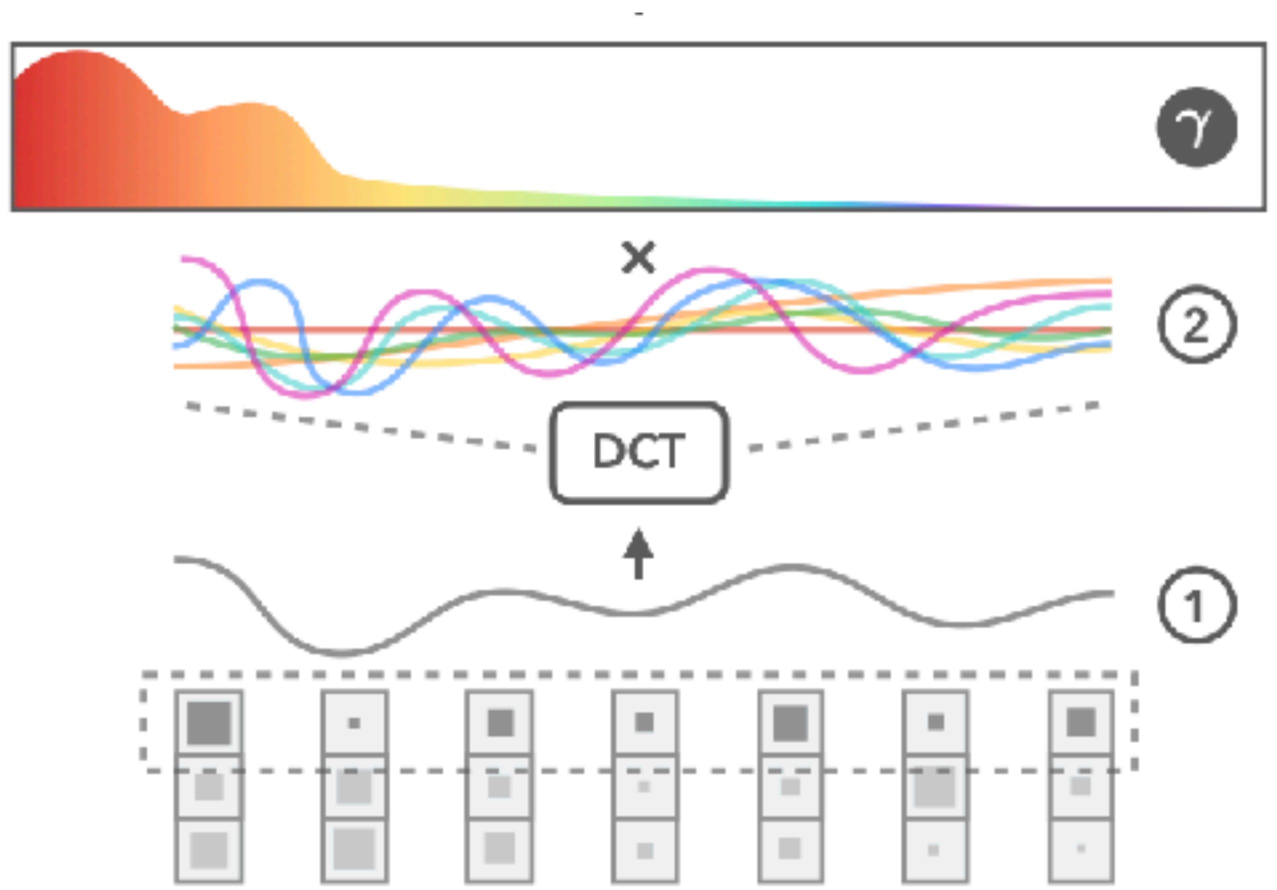
mamy@itu.dk, robv@itu.dk, b.plank@lmu.de

EMNLP, 2022

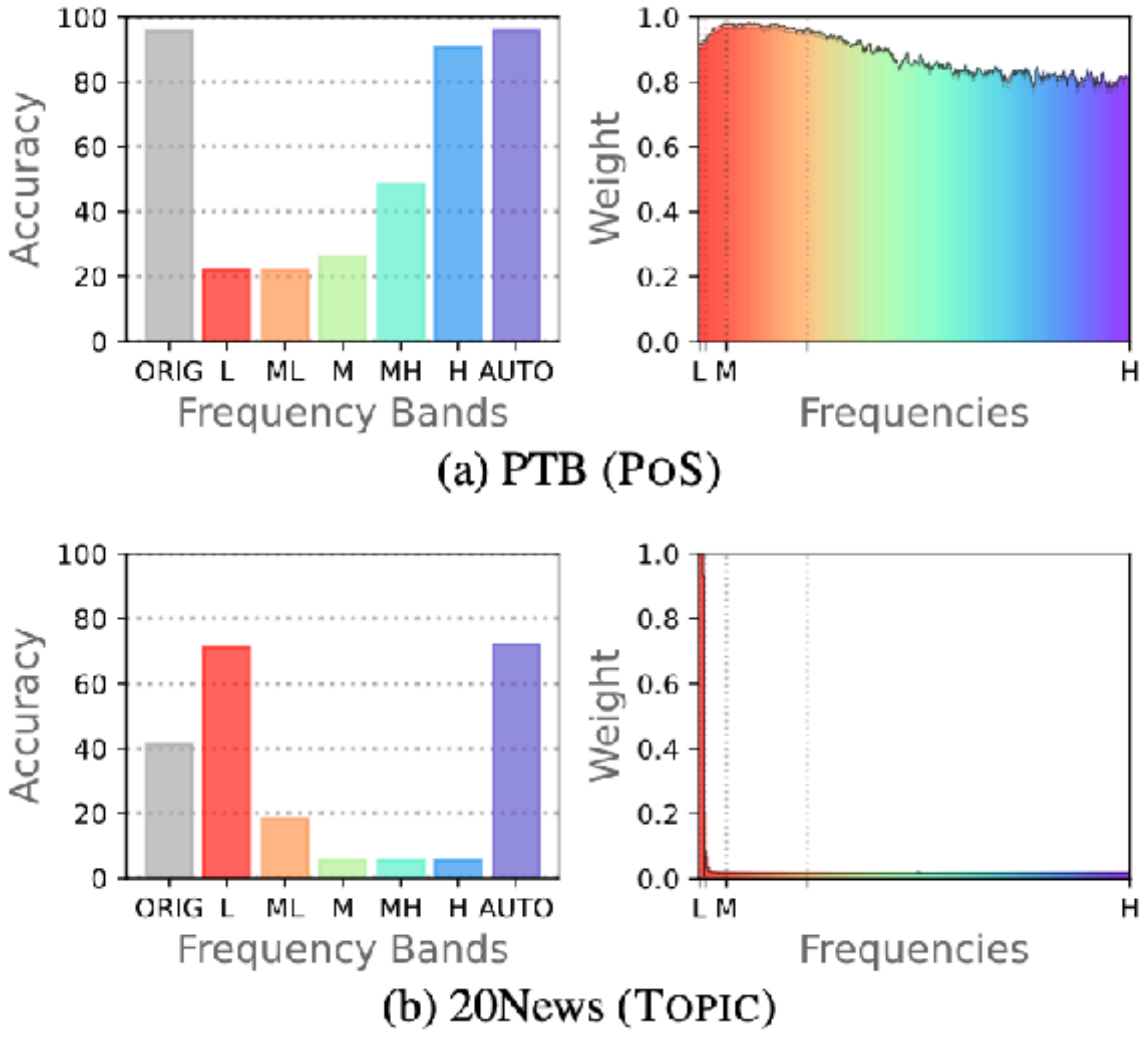
Introspection: What is captured in contextualised embeddings?

- **Probing** has developed into a widely-used toolkit (e.g. Conneau et al., 2018; Hewitt & Manning, 2019; Tamkin et al., 2020)
- **Linguistic information** is encoded at **varying timescales** (subwords, phrases etc) and levels (syntax, semantics etc).
- **Question**: To what extent do multilingual representations capture linguistic properties at different time-scales?

—> Spectral Probing as a 🔍 into large LLMs



Take-Away: Spectral probes rediscover the linguistic hierarchy



(a) PTB (PoS)

(b) 20News (Topic)

Figure 2: **Monolingual Results on PTB and 20News.** Acc of unfiltered (Orig), low (L), mid-low (ML), mid (M), mid-high (MH), high (H), and the spectral probe's automatic filters (Auto) with frequency weightings.

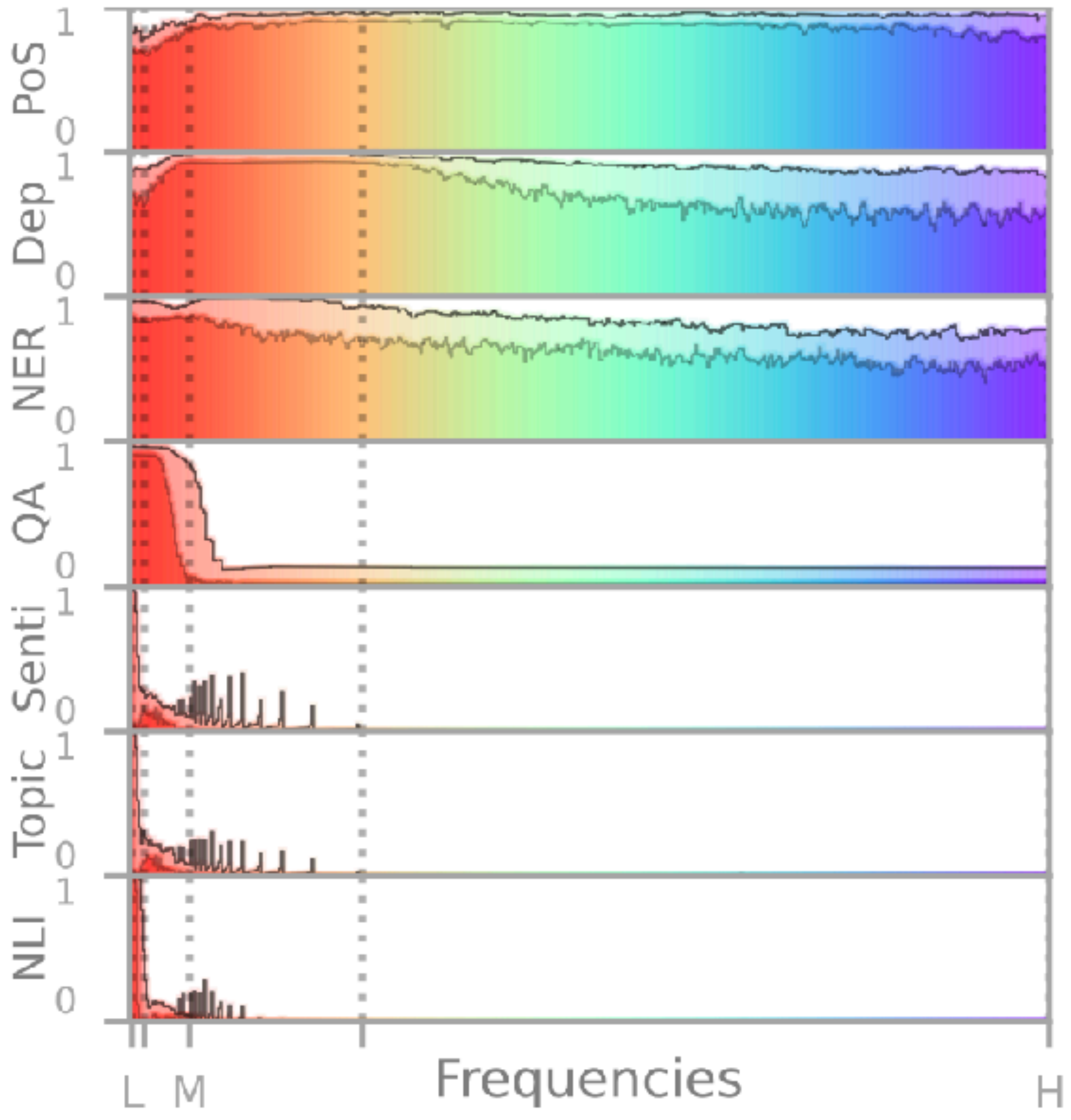
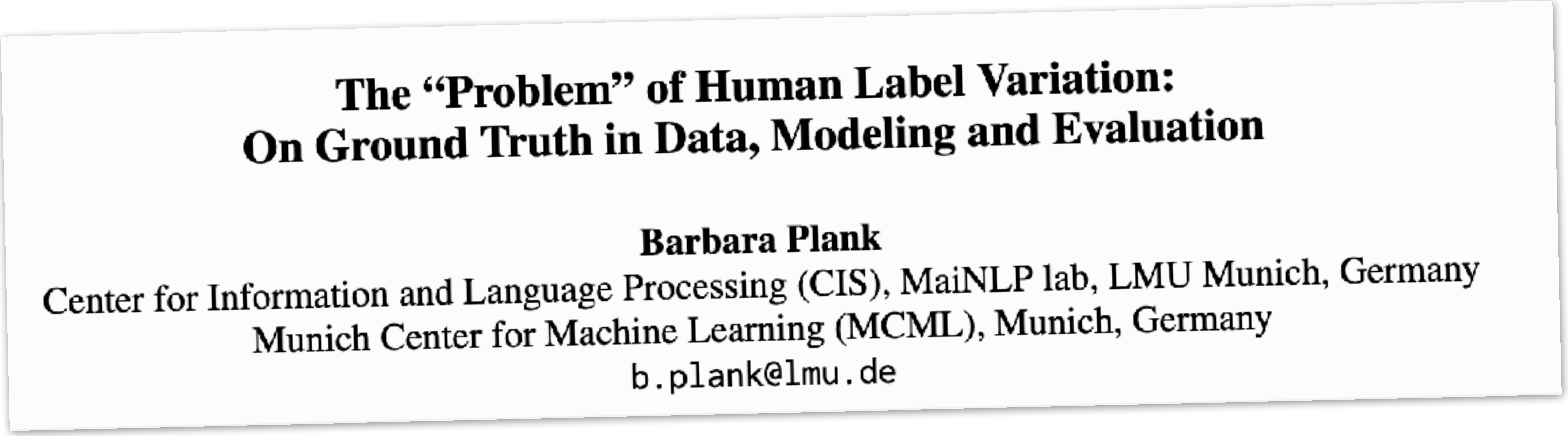


Figure 3: **Spectral Profiles** of all tasks (weight per frequency), with lower and upper bounds across languages.

Applications to Human Label Variability

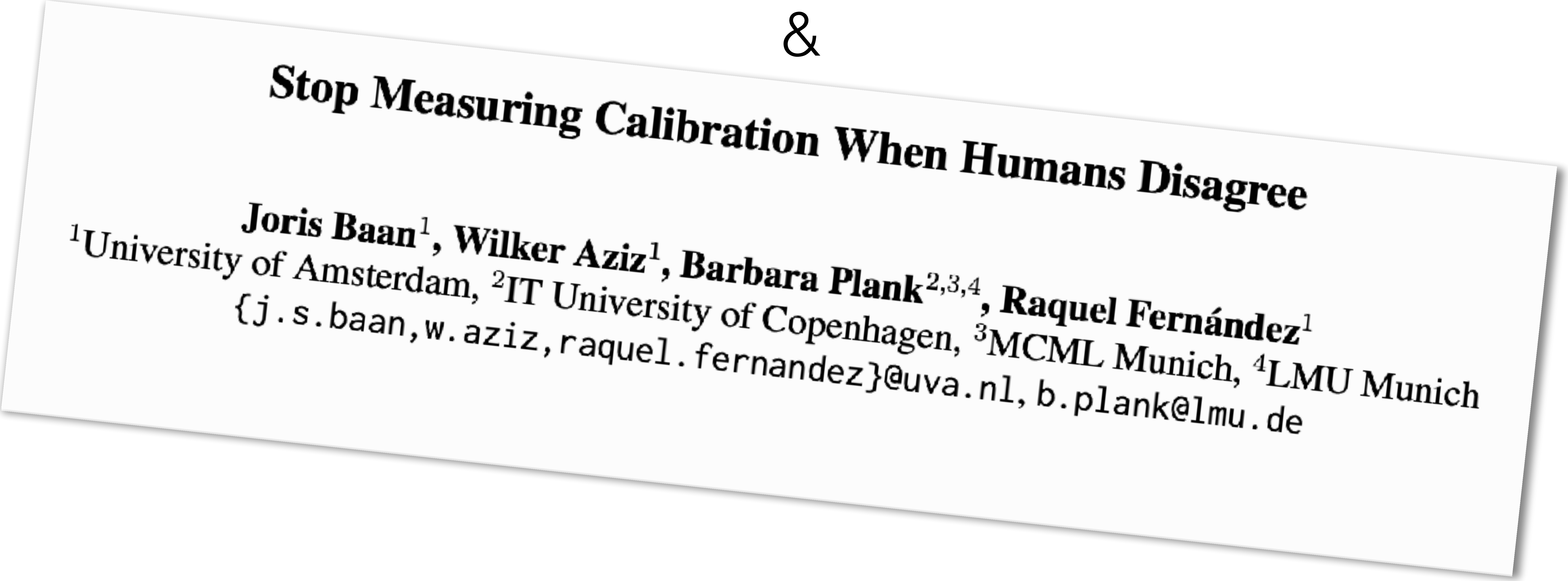
Often there exists no ground truth



The “Problem” of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation

Barbara Plank
Center for Information and Language Processing (CIS), MaiNLP lab, LMU Munich, Germany
Munich Center for Machine Learning (MCML), Munich, Germany
b.plank@lmu.de

&

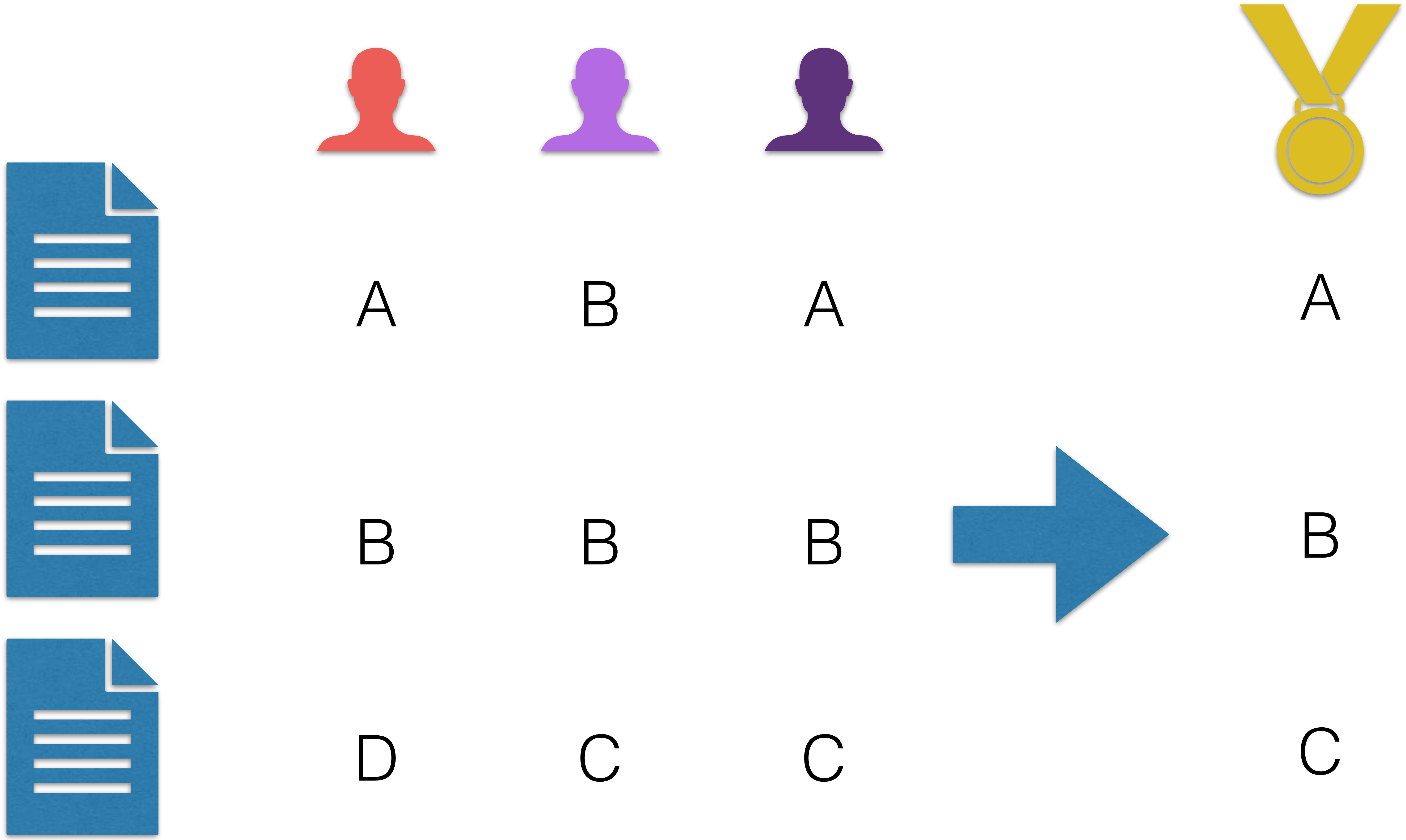


Stop Measuring Calibration When Humans Disagree

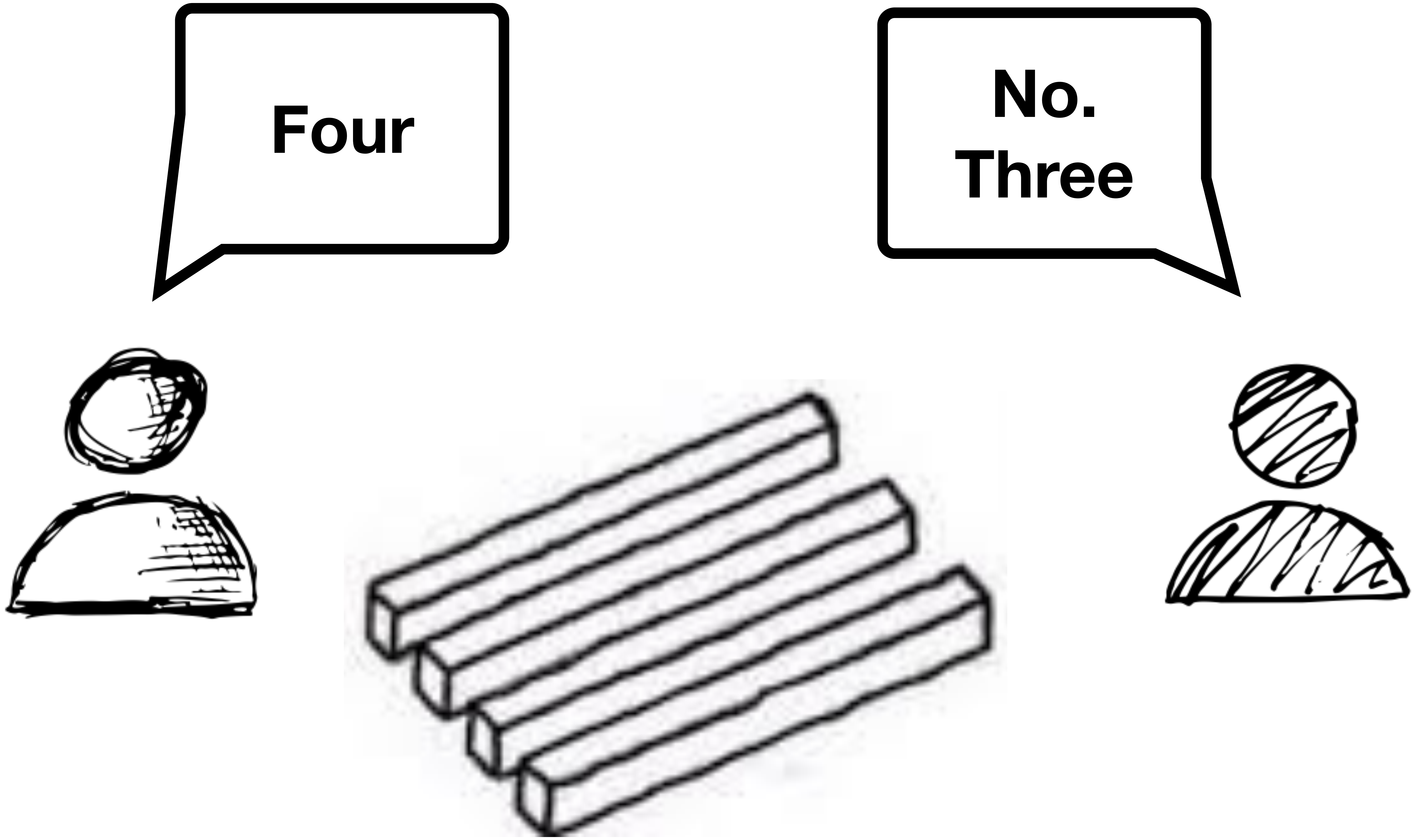
Joris Baan[1], **Wilker Aziz**[1], **Barbara Plank**[2,3,4], **Raquel Fernández**[1]
[1]University of Amsterdam, [2]IT University of Copenhagen, [3]MCML Munich, [4]LMU Munich
{j.s.baan,w.aziz,raquel.fernandez}@uva.nl, b.plank@lmu.de

EMNLP, 2022

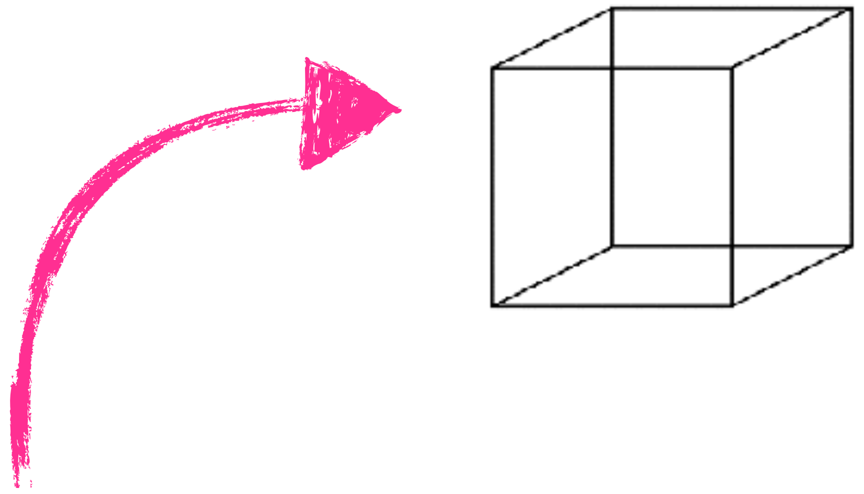
Multiple human annotations



A	B	A		A
B	B	B	→	B
D	C	C		C


Four
No.
Three

Can we turn disagreement into *advantage*?

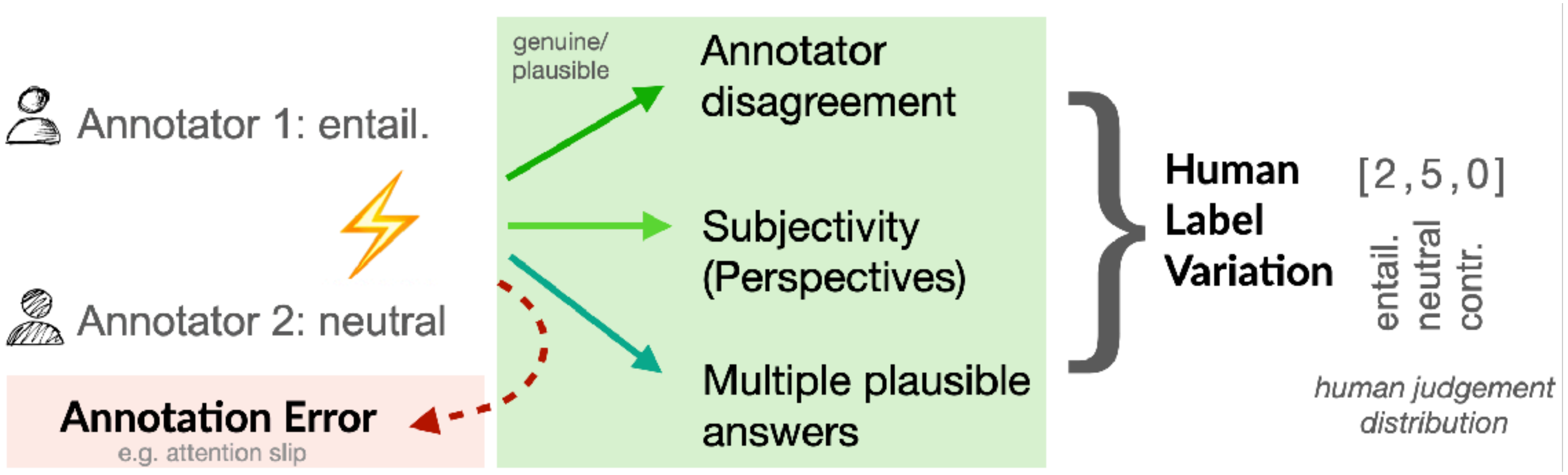


Fortuitous data

Disagreement in human annotation is ubiquitous

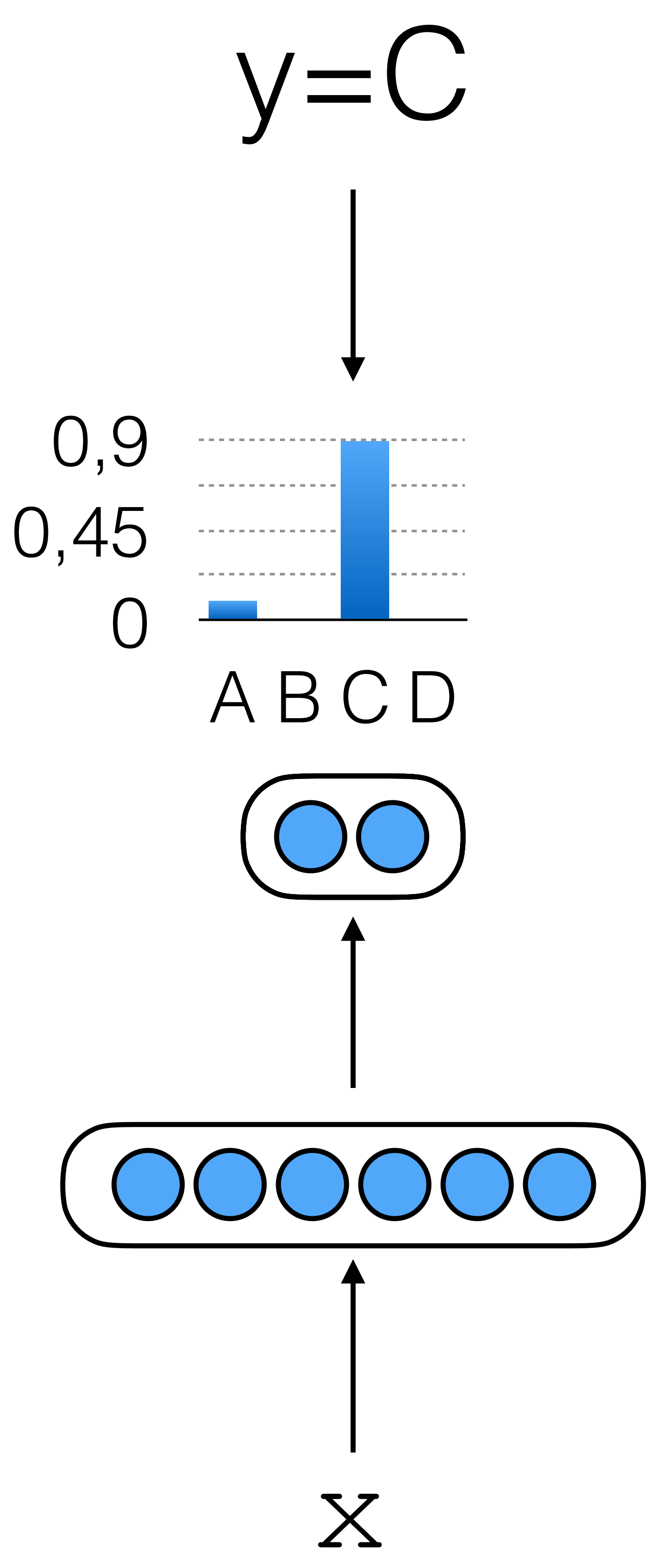
— This impacts all 3 stages of the NLP pipeline.
— Human disagreement is one important form of uncertainty.

Disagreement or variation?

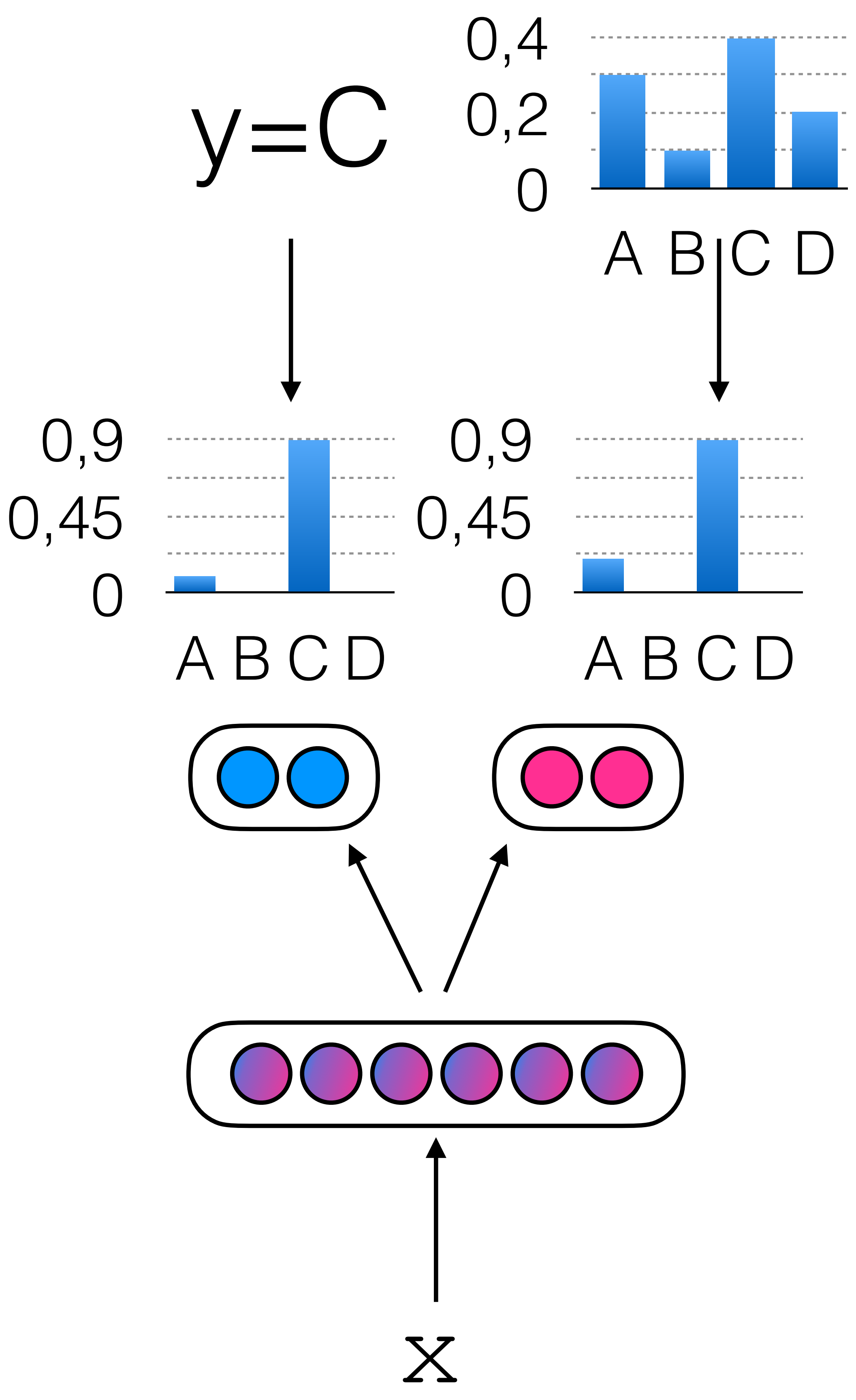


- I propose to call it **Human label variation (HLV)** = plausible variation in annotation (Plank, 2022 EMNLP)
  - Preferred over ‘disagreement’ as that implies two or more views cannot all hold
  - To reconcile different notions in the literature (‘human uncertainty’, ‘perspectives’, ‘hard cases’, ‘disagreement’ etc)
- In contrast: annotation errors

Soft-labels via Multi-Task Learning: Auxiliary task for "human distribution"



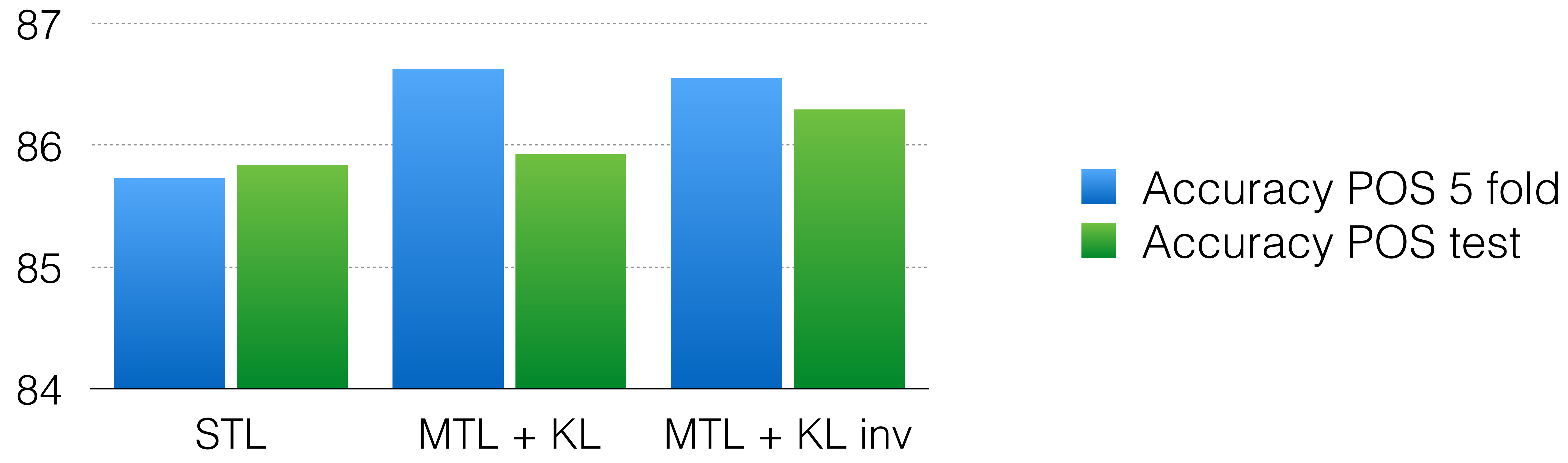
Gold label

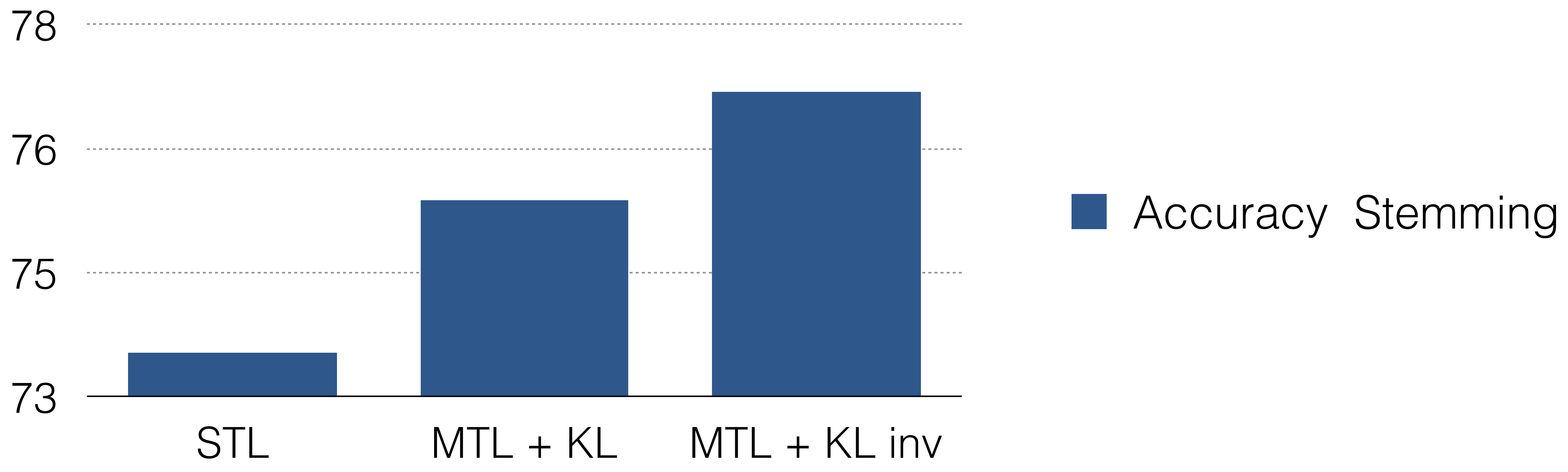


Gold label + Soft label

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

Results





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

Learning with Human Label Variation

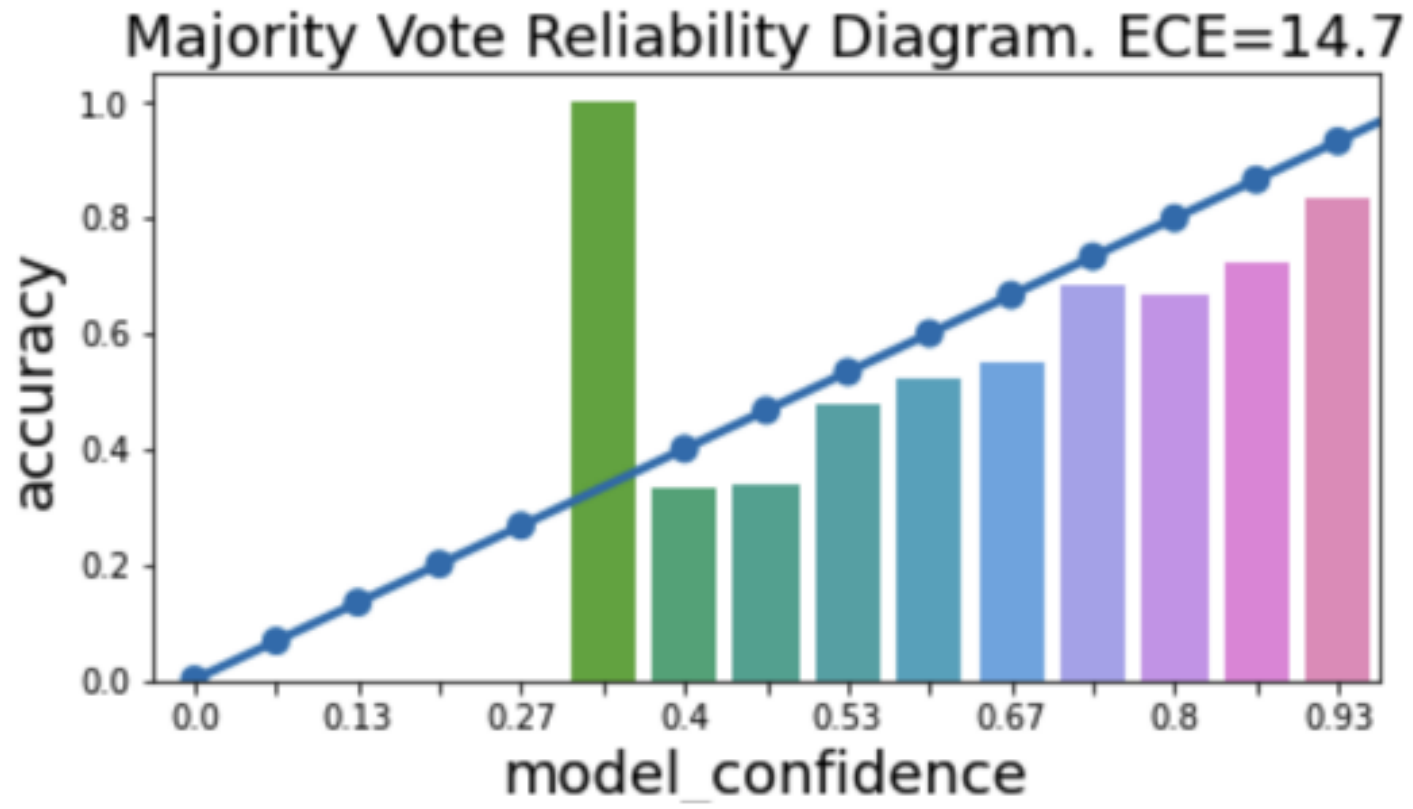
- Soft-label MTL is only one way to use MTL
- Alternative: Davani et al. (2021) who model each annotator separately as output head in a MTL model (instead of one head with the “human distribution”)
- Many more approaches to learn with Human Label Variation (see survey in Uma et al., 2021 JAIR)

Is *Human Label Variation* So Bad? No.

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

More trustworthy models: Calibration & Model Uncertainty

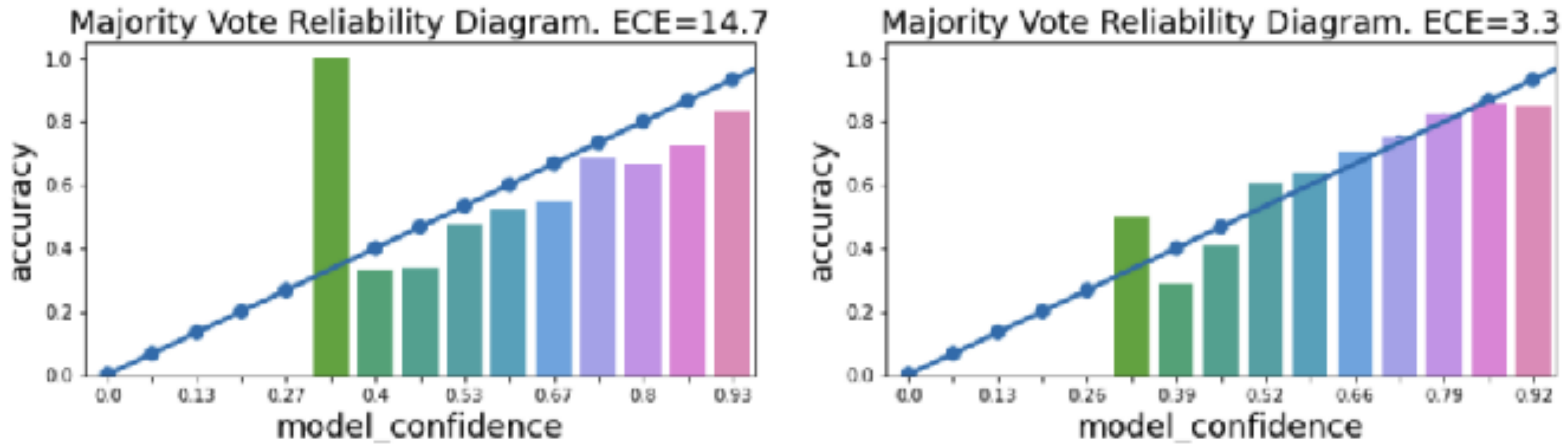
- Calibration is a popular framework to evaluate whether a classifier knows when it does not know



- However, calibration assumes there exists a ground truth
- We examine calibration under the lens of human label variation

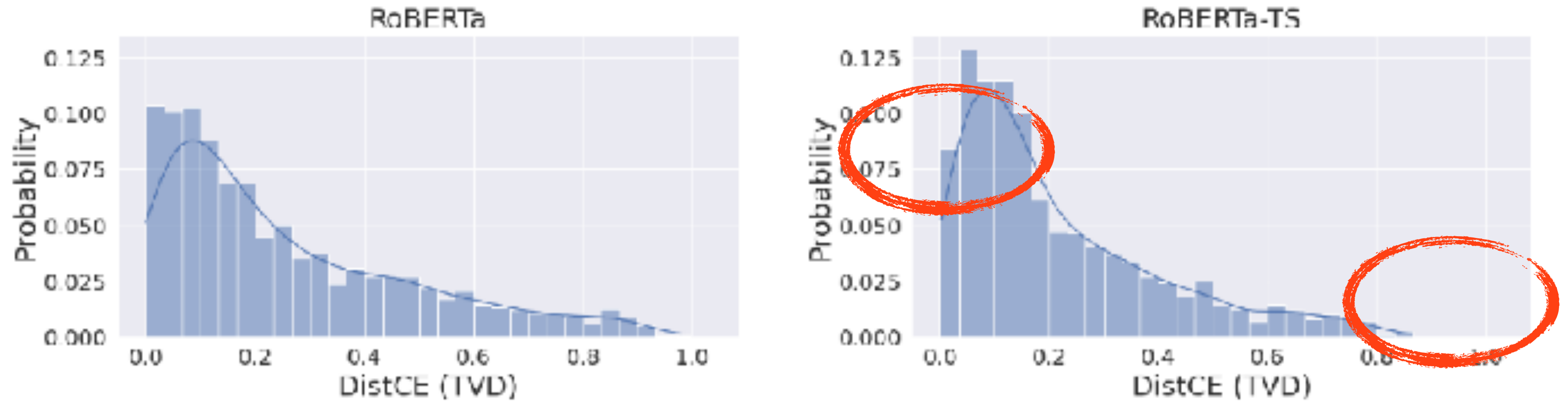
Calibration to majority is flawed

- Temperature Scaling improves typical calibration measures (ECE). But what does that mean?



(c) ECE: Vanilla (d) ECE: Temp Scaling

- With instance-level distributions we get a more fine-grained view on model calibration (TVD distance; Baan et al., 2022)



(a) DistCE: Vanilla (b) DistCE: Temp Scaling

Fewer extremely miscalibrated

BUT even fewer perfectly calibrated instances!

(Baan, Aziz, Plank, Fernandez, 2022 EMNLP) https://arxiv.org/abs/2210.16133

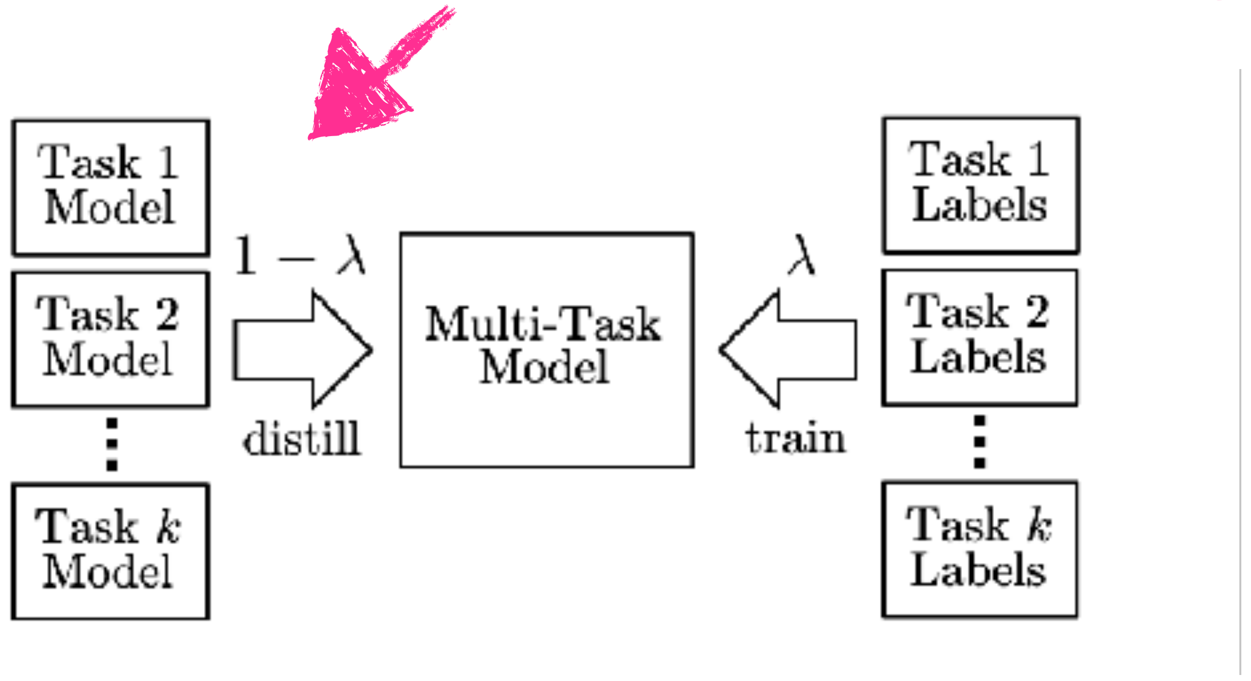
Outline

- Introduction: Why Transfer? Dimensions of Language Variation
- Part 1: What is Transfer Learning?
  - Three views on Transfer Learning, Related Learning Strategies
- Part 2: A type of TL: What is Multi-Task Learning?
  - What and Why, Perspectives on MTL
  - Short hands-on tutorial with MaChAmp
- Part 3: Selected Case Studies
  - Applications to Multilinguality, Transferability Estimation, Human Label Variation
- Outro

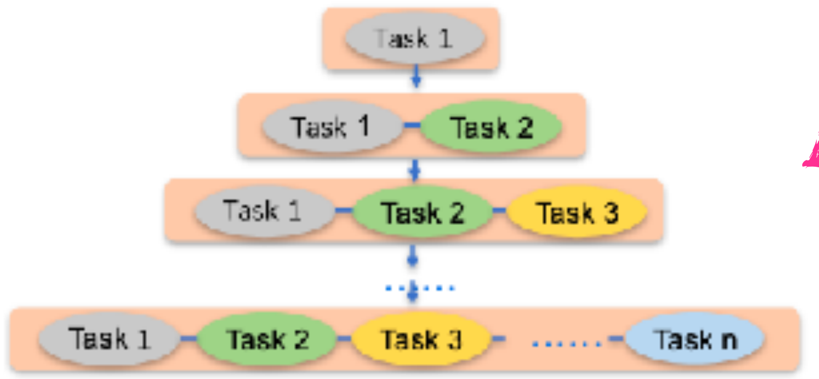
Some selected advances in traditional MTL

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

Distillation with teacher annealing



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



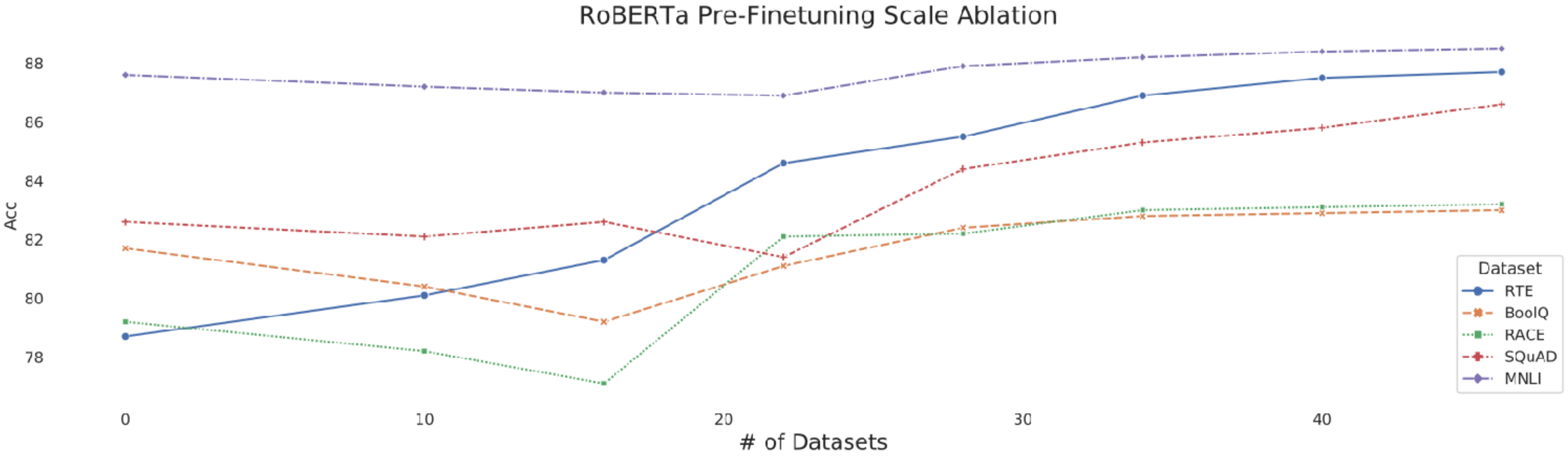
Sequential Multi-task Learning

Progressively adding tasks

- MTL & *adapters* via shared ***hypernetworks*** (Mahabadi et al., 2021, Üstün et al., 2022)

Generates adapter parameters

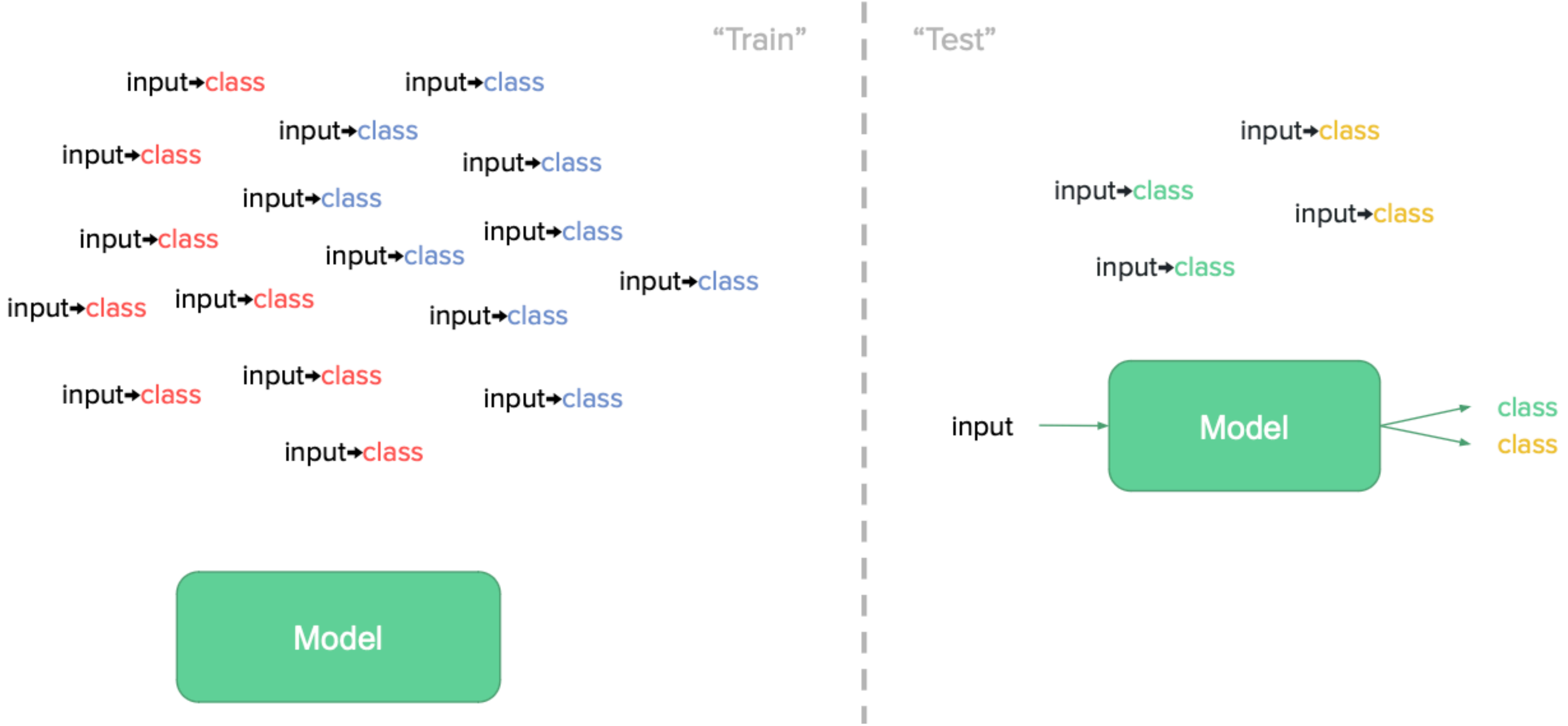
Scaling up seems one key finding for MTL

- Pre-Fine tuning (Aghajanyan et al., 2021): MTL between pre-training and fine-tuning
- Scaling up and using many in-between tasks was key
- 

Multi-task learning in light of T5, ChatGPT etc: Not just approaches and models change, also our terminology!

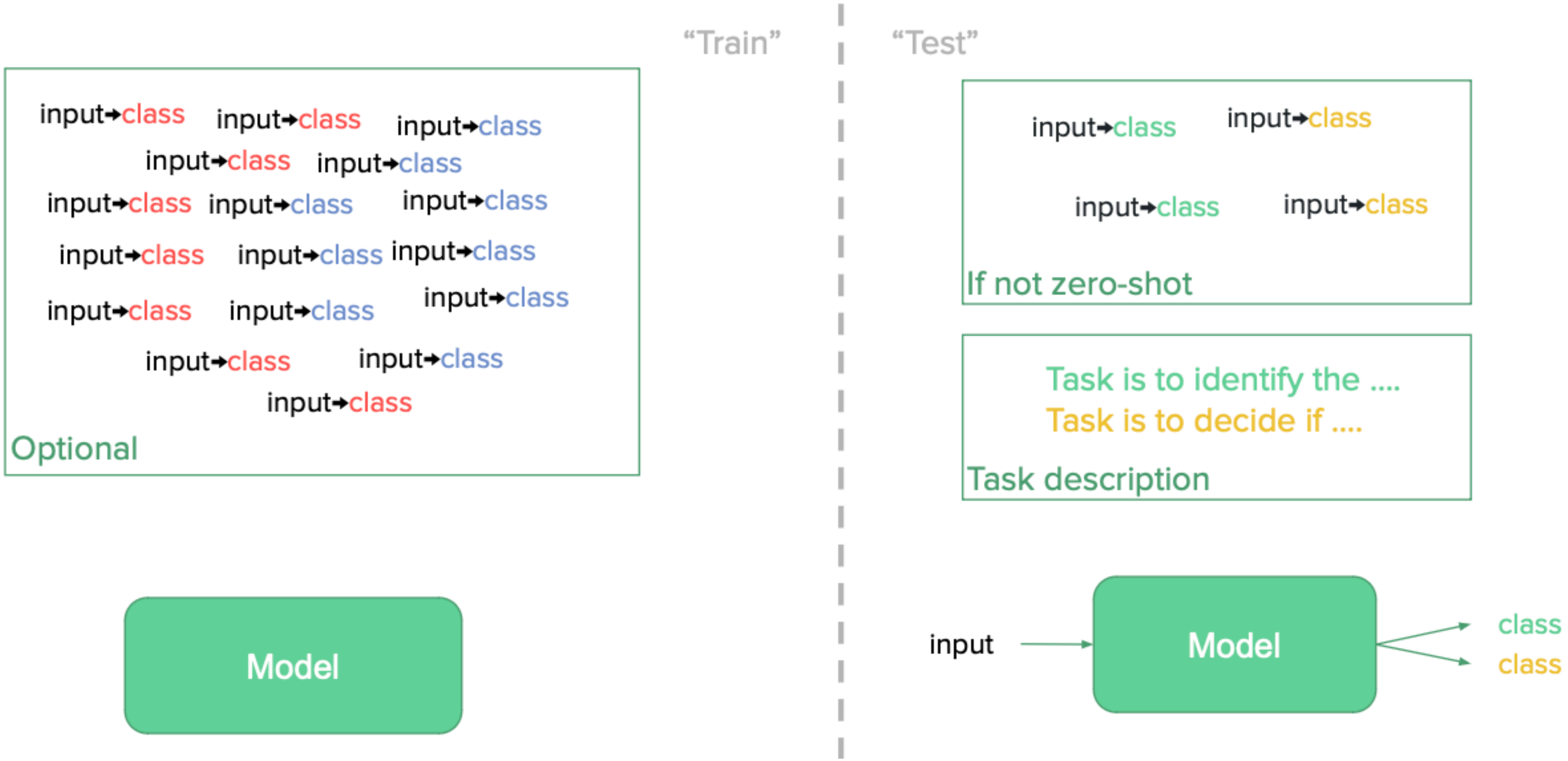
Traditional notion of few-shot learning

Slide by Beltagy et al., ACL 2022 tutorial



Recent notion of few-shot learning

Slide by Beltagy et al., ACL 2022 tutorial



"Train"
"Test"
input→class input→class input→class
input→class input→class
input→class input→class input→class
input→class input→class input→class
input→class input→class input→class
input→class input→class
input→class
Optional
Model
input→class input→class
input→class input→class
If not zero-shot
Task is to identify the ....
Task is to decide if ....
Task description
input
Model
class
class

“Learning many tasks” takes on new meanings, too (e.g. FLAN)

Wei et al., 2022

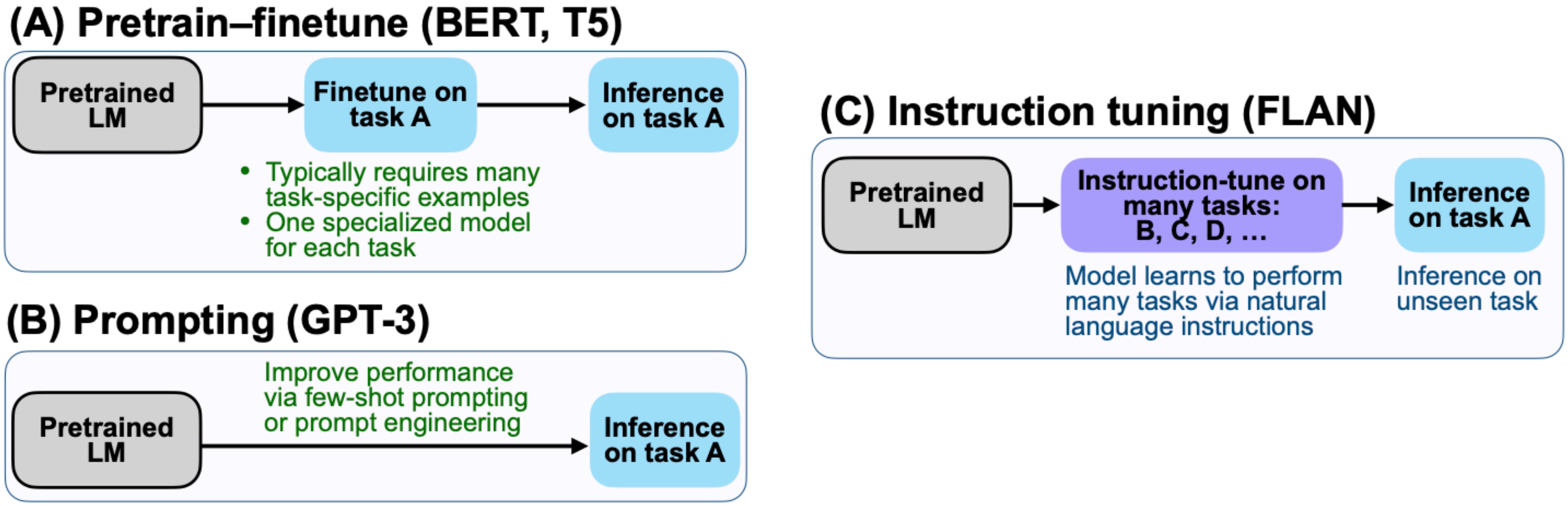
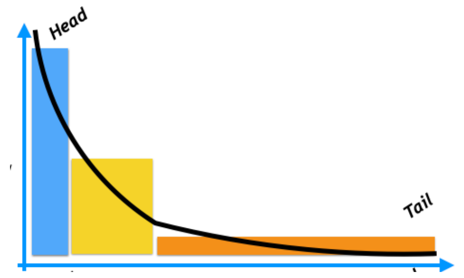

(A) Pretrain–finetune (BERT, T5)
Pretrained LM
Finetune on task A
Inference on task A
• Typically requires many task-specific examples
• One specialized model for each task
(B) Prompting (GPT-3)
Pretrained LM
Improve performance via few-shot prompting or prompt engineering
Inference on task A
(C) Instruction tuning (FLAN)
Pretrained LM
Instruction-tune on many tasks: B, C, D, …
Inference on task A
Model learns to perform many tasks via natural language instructions
Inference on unseen task

Figure 2: Comparing instruction tuning with pretrain–finetune and prompting.

What is a task?

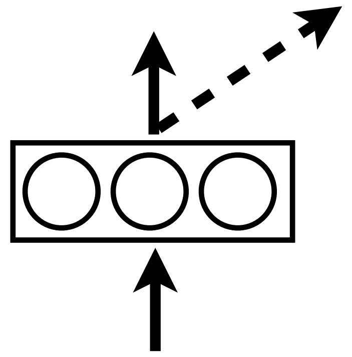
To wrap up…

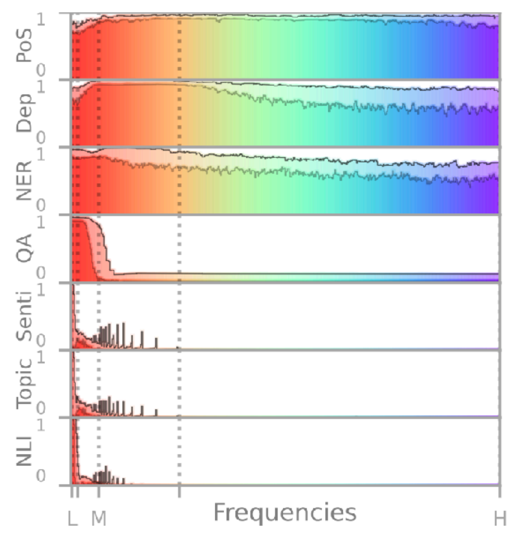
To wrap up

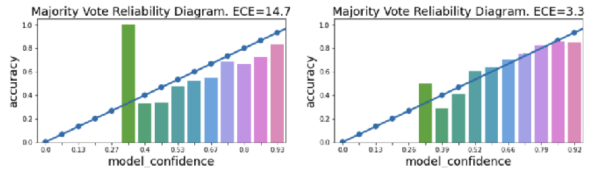


- Scarce and biased data are ubiquitous
- Transfer Learning is broad! Sequential TL (pre-training) is just one kind

- We have seen applications of:
  - Data selection
  - Multi-Task Learning
  - Probing
  - Performance Prediction
  - Human Label Variation and Calibration







(c) ECE: Vanilla

(d) ECE: Temp Scaling

N 48° 9′ 8.676″
E 11° 34′ 49.332″

Thanks to my team and collaborators

nlp

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

north

IT UNIVERSITY OF COPENHAGEN

Questions? Thanks!

@barbara_plank

b.plank@lmu.de

Research support by:

erc

DANMARKS FRIE FORSKNINGSFOND

MCML

ellis unit | AMSTERDAM