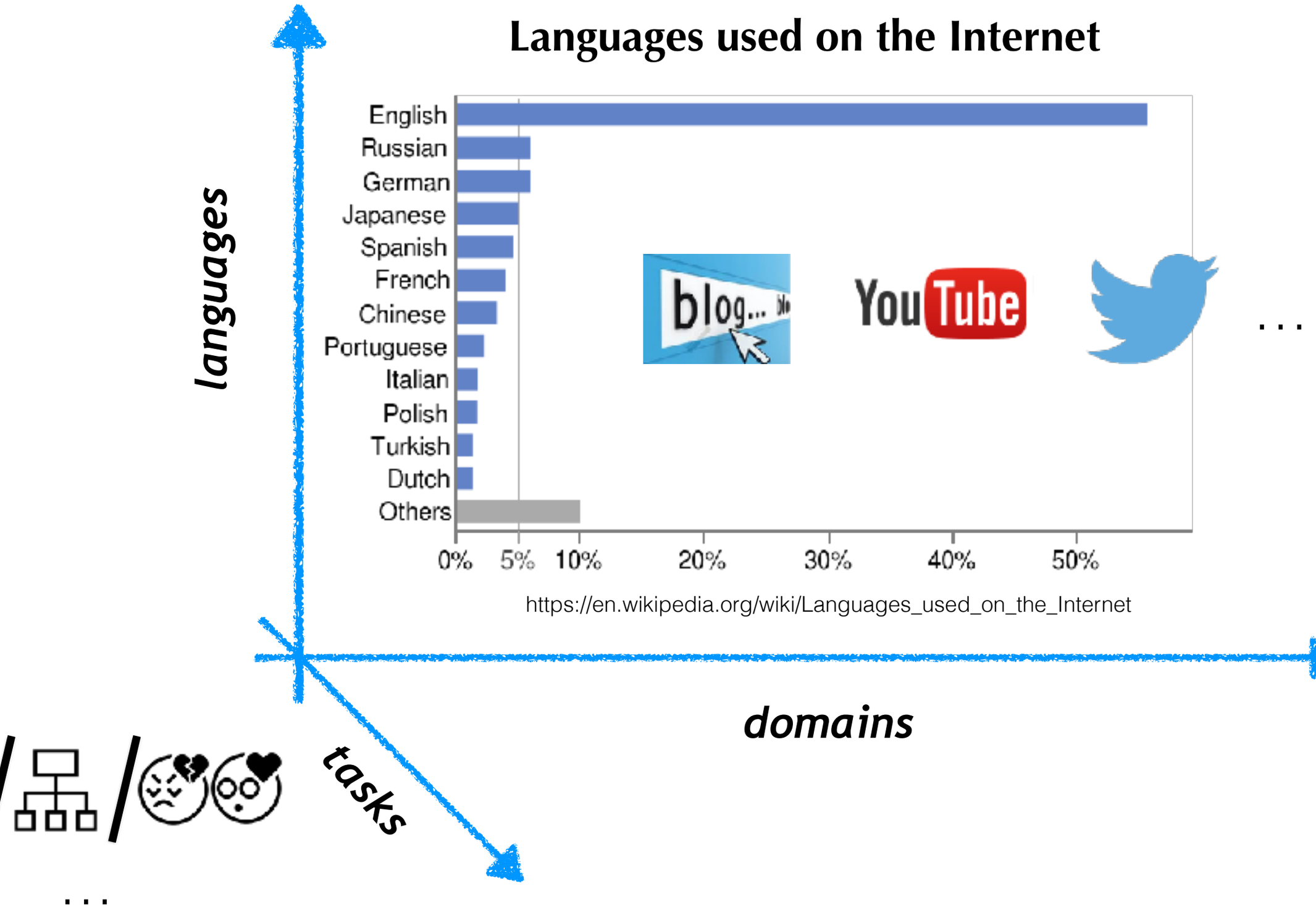
IT UNIVERSITY OF COPENHAGEN

Transfer and Multi-Task Learning in Natural Language Processing

@barbara_plank

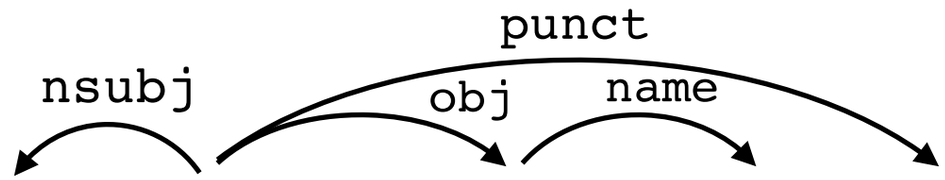
LxMLS, July 10, 2021

Ultimate Goal: NLP for everyone



NLP Tasks: Learning from D_l: {<x,y>}

human-annotated examples

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

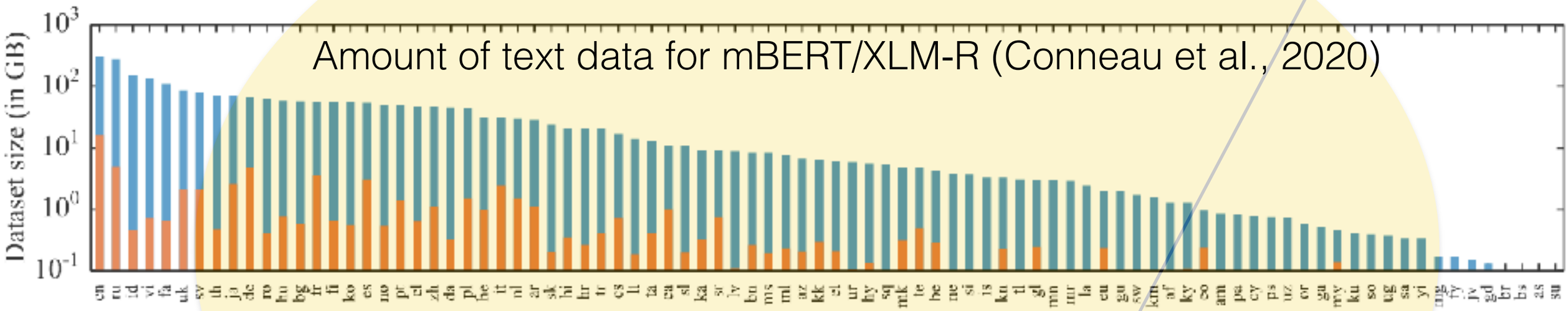
➡ Time-intensive

➡ Expensive

Labeled data is scarce (1/2)

labeled data
D_l: {<x,y>}

Amount of text data for mBERT/XLM-R (Conneau et al., 2020)



Log scale

even X
is scarce

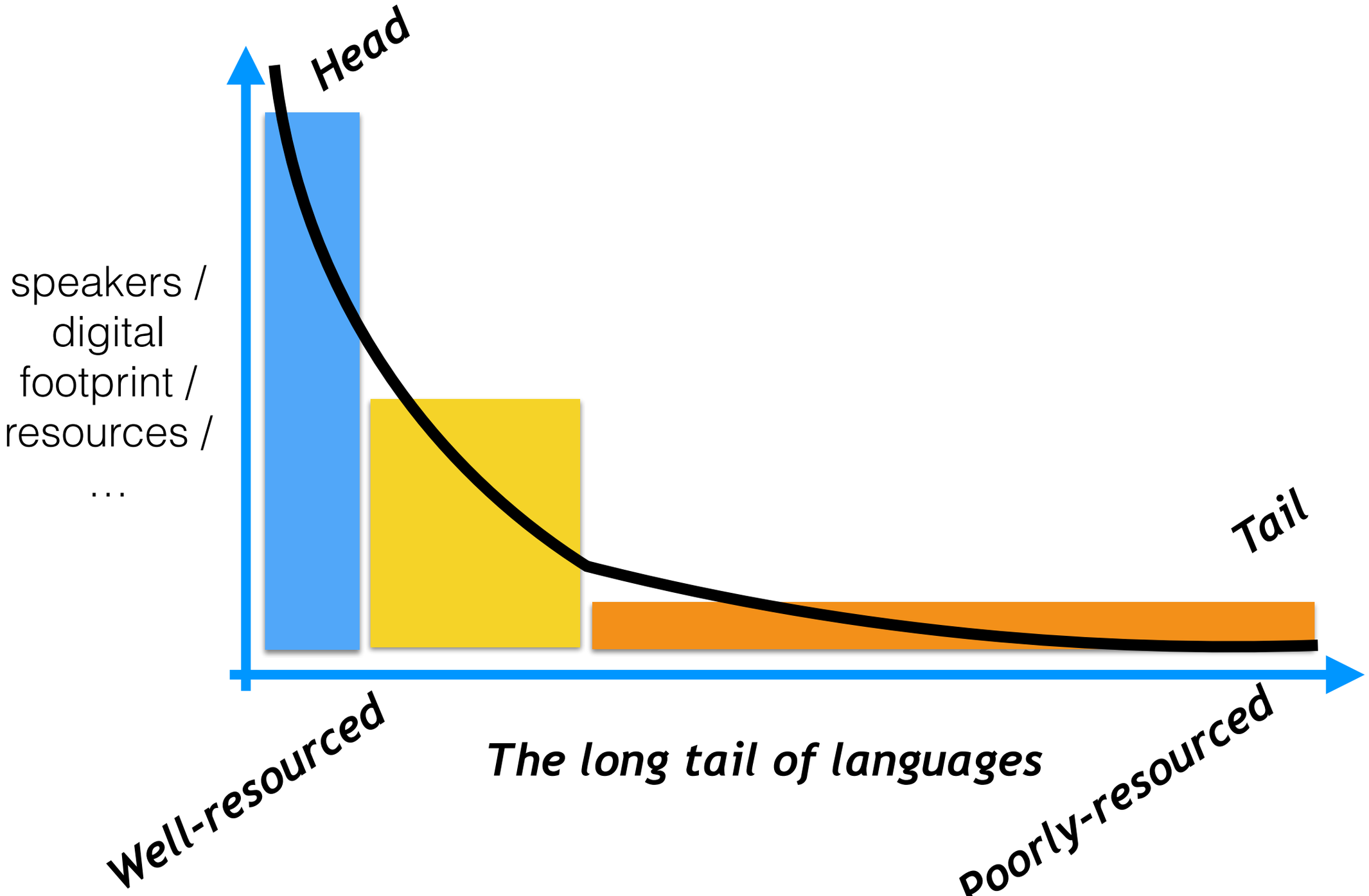
Y sparsity

X sparsity

unlabeled data
D_u: {<x>}

all available data

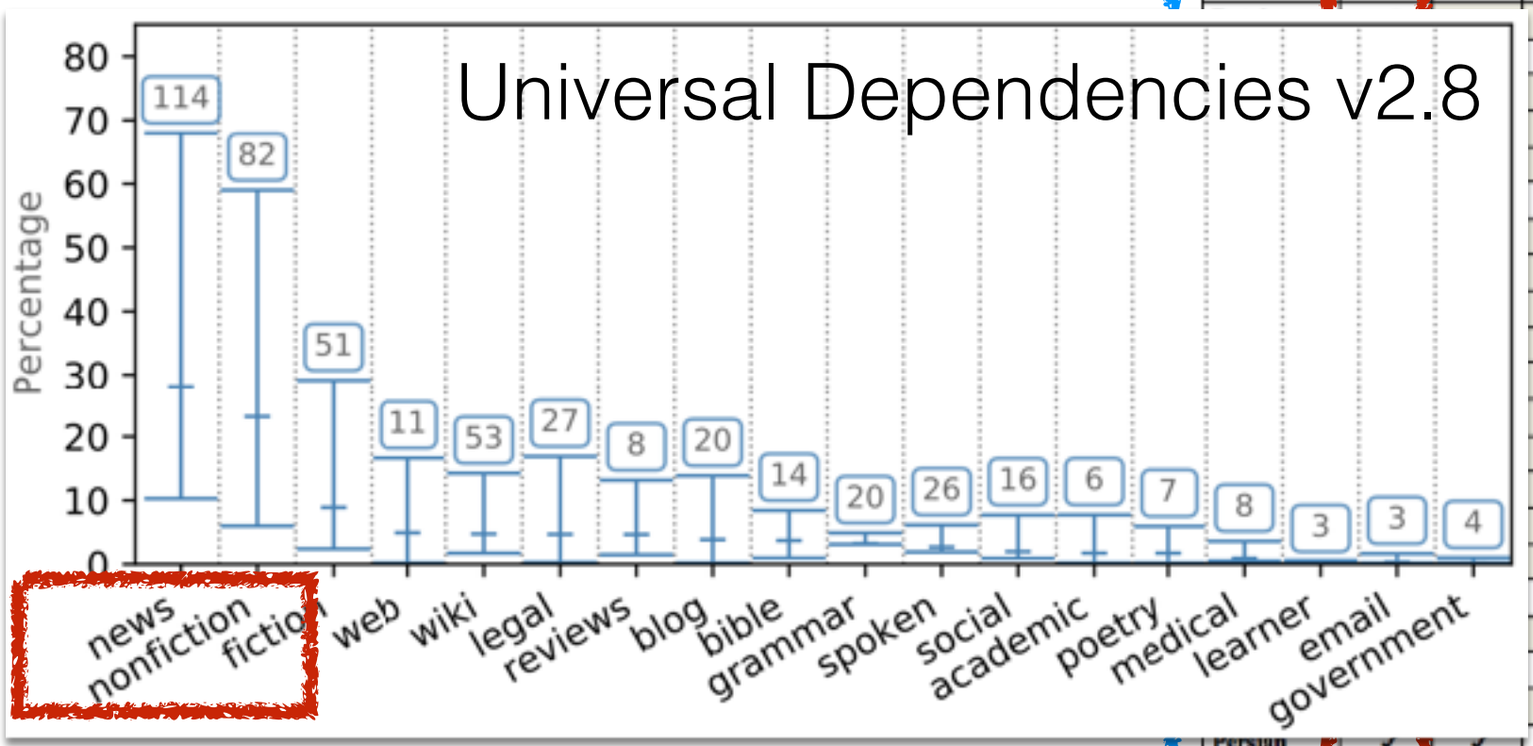
Labeled data is scarce (2/2)



Labeled data is biased (1/2)

Selection bias:

Newswire



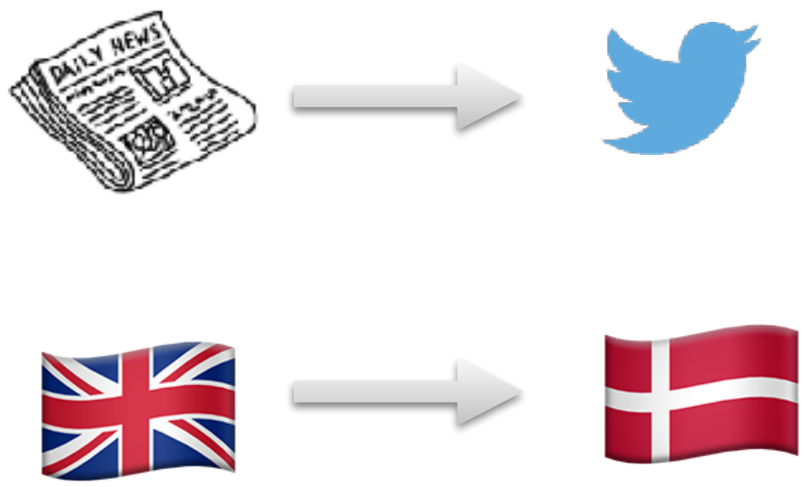
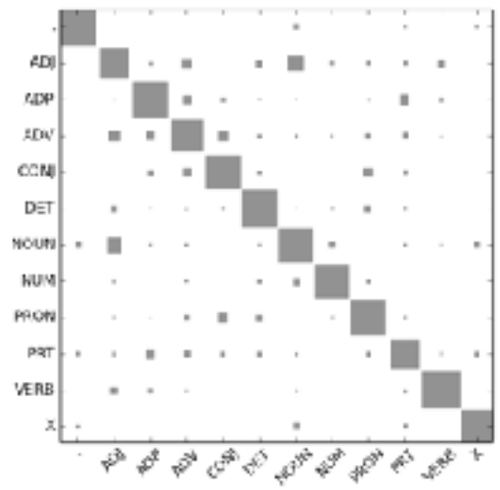
From 40 langs/54 TBs (2016; v1.3)
to 114 langs/202 TBs (2021; v2.8)
Plot by Max Müller-Eberstein (work in progress)

domains

	news	*fiction*	*nonfict.*	*blog*	*bible*	*legal*	*medical*	*social*	*spoken*	*wiki*	*web*	*reviews*
Anc. Greek		✓	✓		✓							
Arabic	✓											
Basque	✓	✓										
Bulgarian	✓	✓				✓						
Catalan	✓											
Chinese										✓		
Croatian	✓									✓		
Czech	✓		✓			✓	✓					✓
Danish	✓	✓	✓						✓			
							✓			✓		
			✓	✓				✓	✓		✓	✓
				✓		✓				✓		
				✓						✓		✓
			✓			✓	✓					
					✓				✓	✓		
				✓								
						✓					✓	
						✓				✓		
						✓				✓		
			✓		✓							
					✓							
			✓	✓								
				✓								
Persian	✓	✓	✓			✓	✓	✓	✓			
Polish	✓	✓	✓									
Portuguese	✓			✓								
Romanian	✓	✓	✓			✓	✓			✓		
Russian	✓	✓	✓							✓		
Slovenian	✓	✓	✓						✓			
Spanish	✓			✓						✓		✓
Swedish	✓	✓	✓						✓			
Tamil	✓											
Turkish	✓		✓									

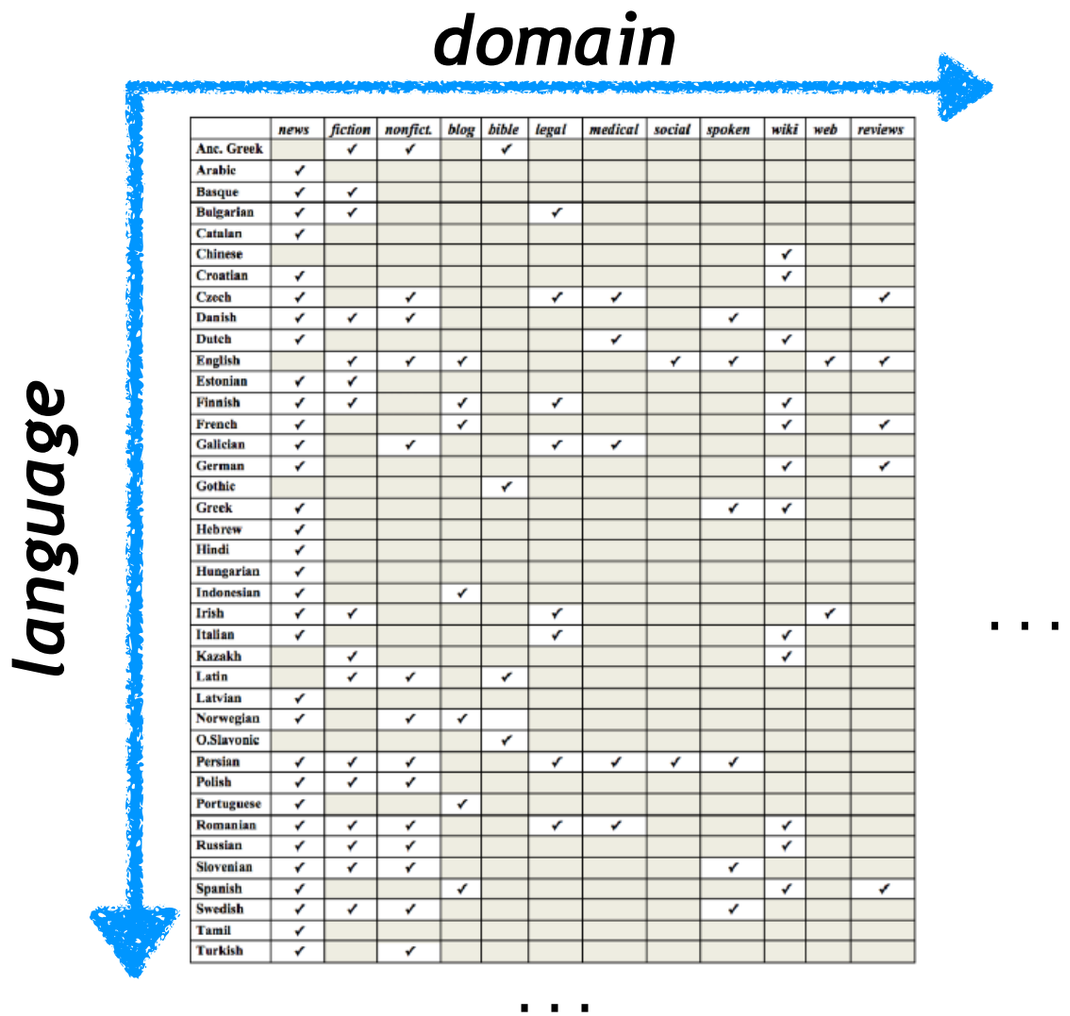
(Plank, 2016, KONVENS)

Labeled data is biased (2/2)

X (input dimension)	Y (output/label dimension)
Input distribution shifts adverse condition for train and evaluation data (not i.i.d.) 	**Single ground truth learning** e.g. disagreement in part-of-speech tags:  (Plank et al., 2014b)

Simple solution: Annotate more?

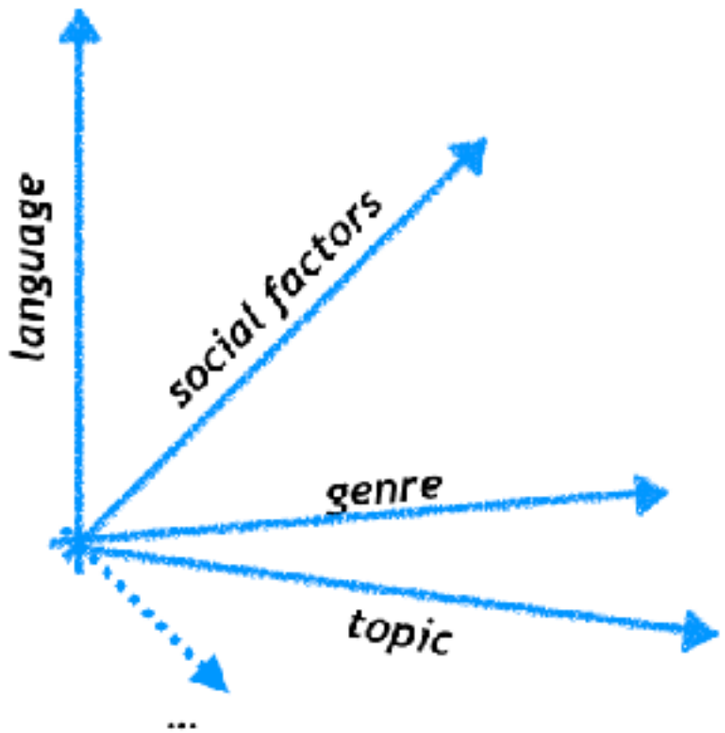
- |*domain*| x |*language*| - **huge space**!



But what's a domain anyway?

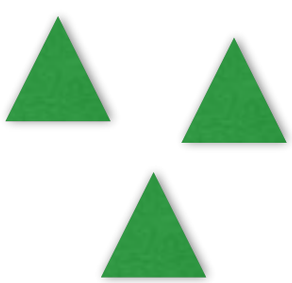
- High-dimensional space, many (unknown) factors
- A **variety** forms a *region in this space* (*subspace*)
  - some members more *prototypical* (Wittgenstein)

The variety space
(Plank, 2016 KONVENS)



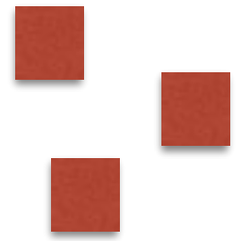
What to do about it?

Typical setup



Traditional ML:

Train & evaluate
on same
domain/task/language



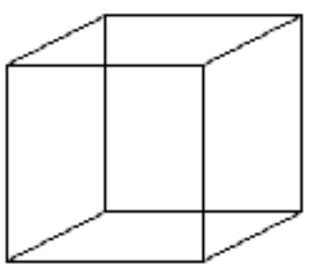
Model A

Model B

Adaptation / Transfer Learning

Source

task A



Data as by-product

Knowledge gained
to help solve
a **related problem**

Model A

Target

task B

task C

Model B

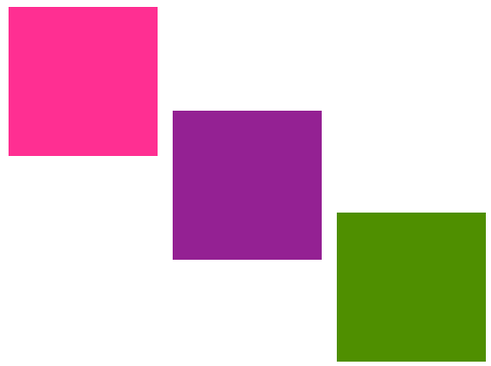
CROSS-DOMAIN

CROSS-LINGUAL

MULTI-TASK

FORTUITOUS

Three views on Transfer Learning (TL)

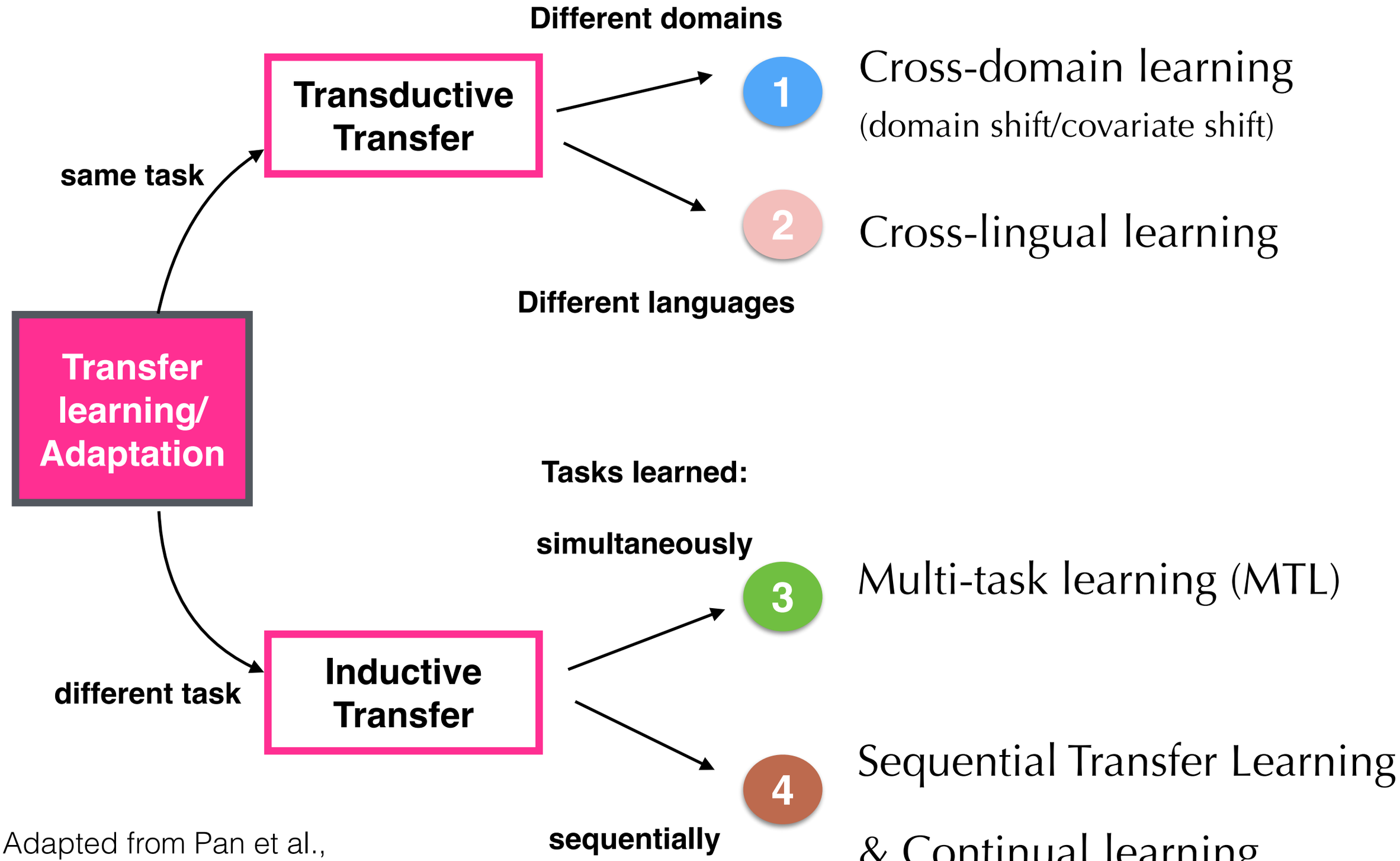


Data domain $\mathcal{D} = \{\mathcal{X}, P(\mathcal{X})\}$
with $\mathcal{X}$ the feature space

~ Notation ~

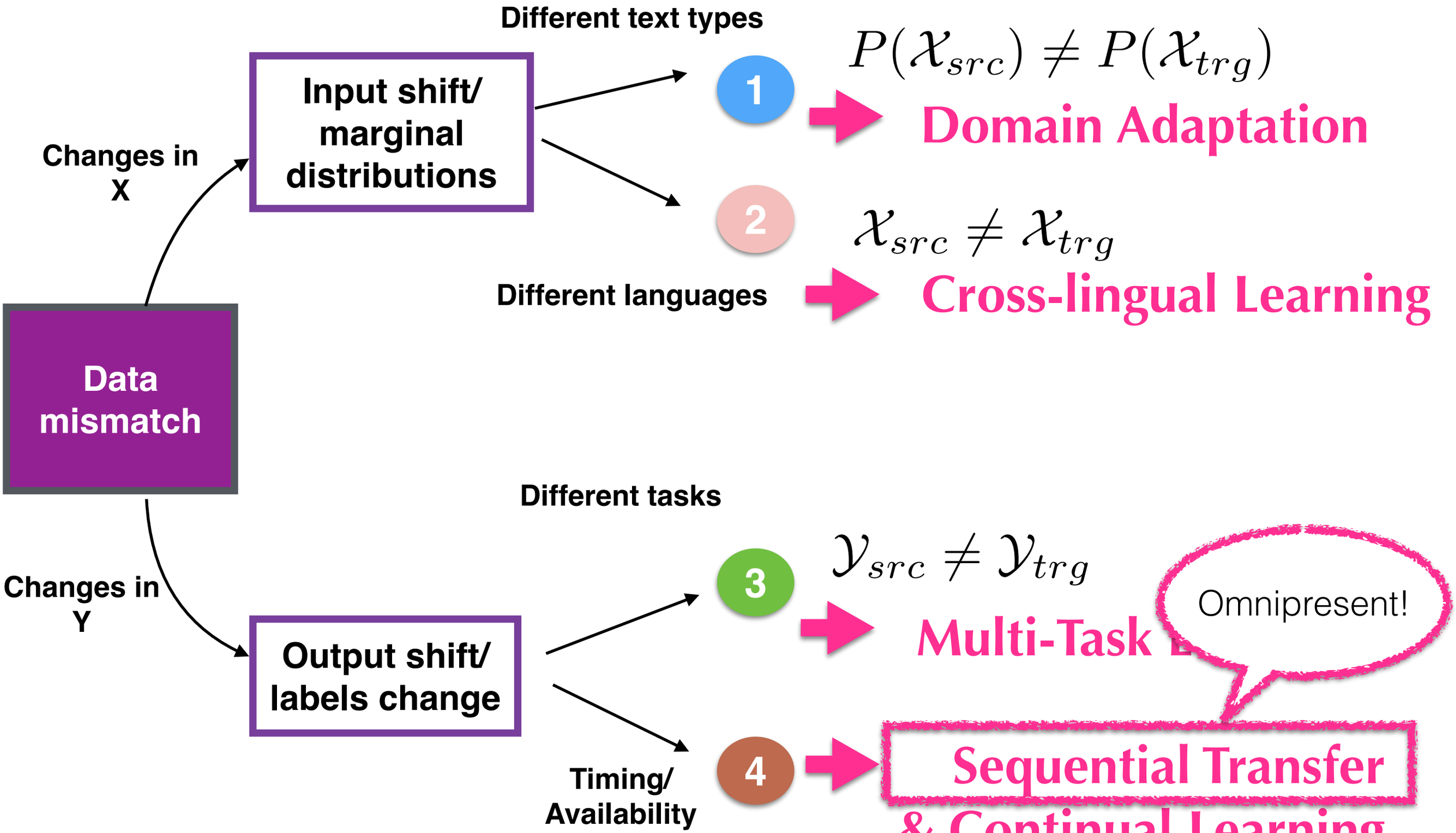
Task $\mathcal{T} = \{\mathcal{Y}, P(\mathcal{Y}|\mathcal{X})\}$
where $\mathcal{Y}$ is the label space

Types of Transfer Learning (1/3)

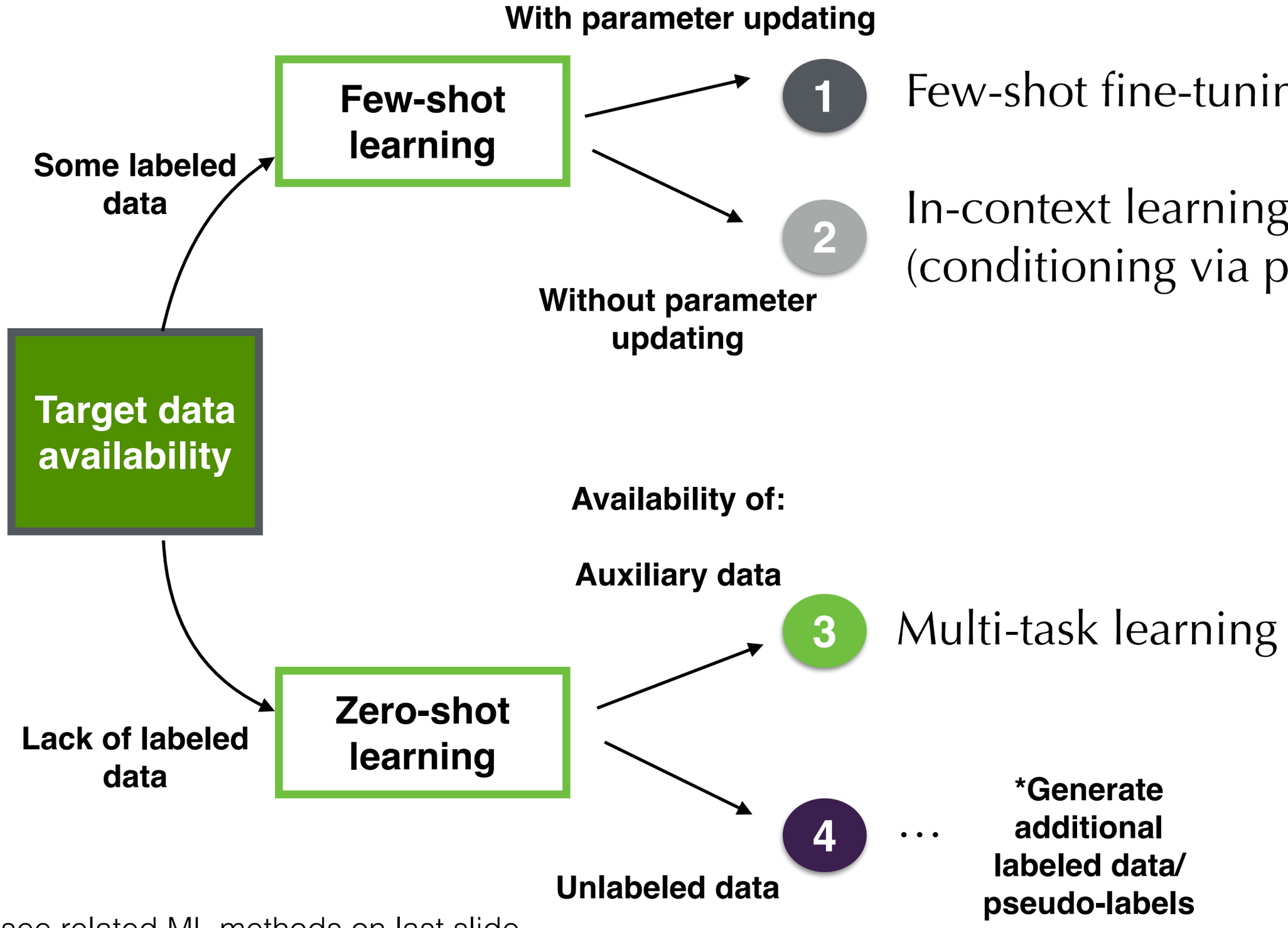


Adapted from Pan et al., (2009) & Ruder (2019)

Types of Data Mismatch (2/3)



Types of Resource Availability (3/3)

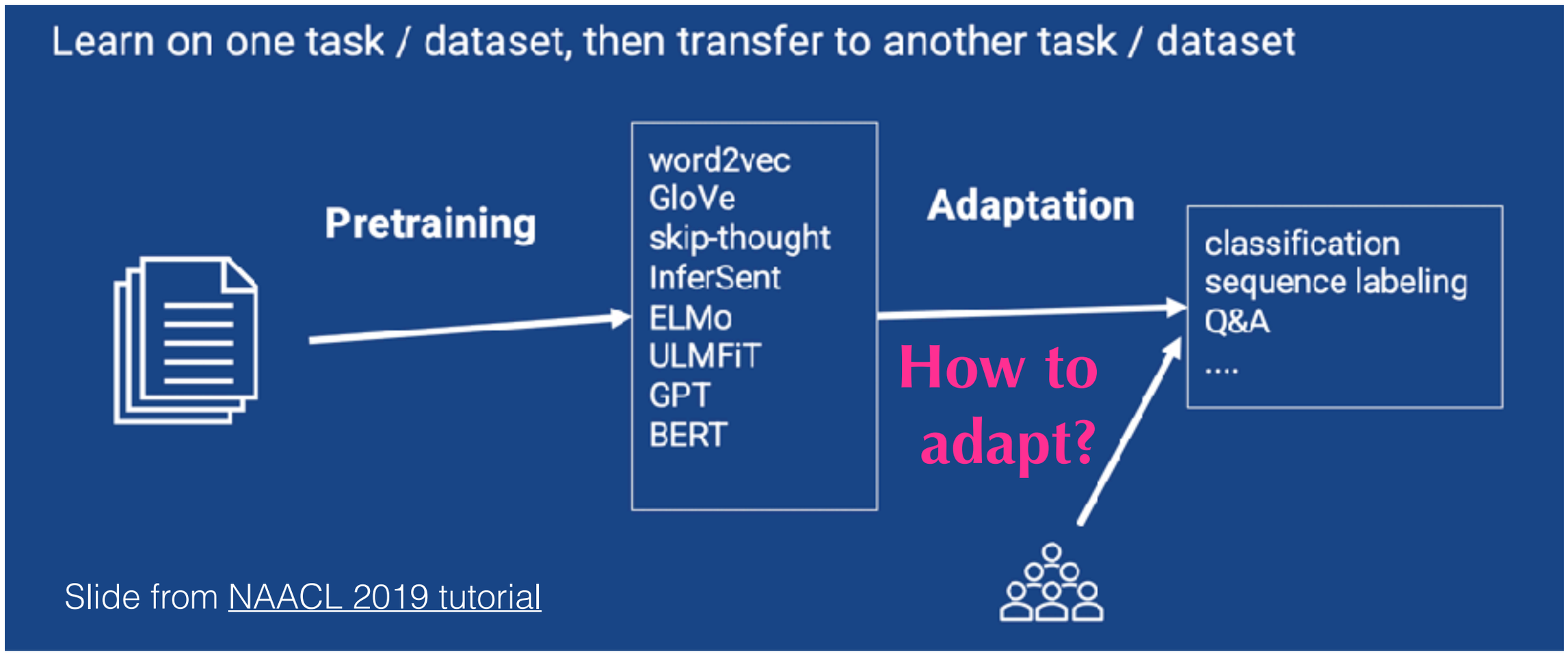


*see related ML methods on last slide

One omnipresent kind of Sequential Transfer Learning

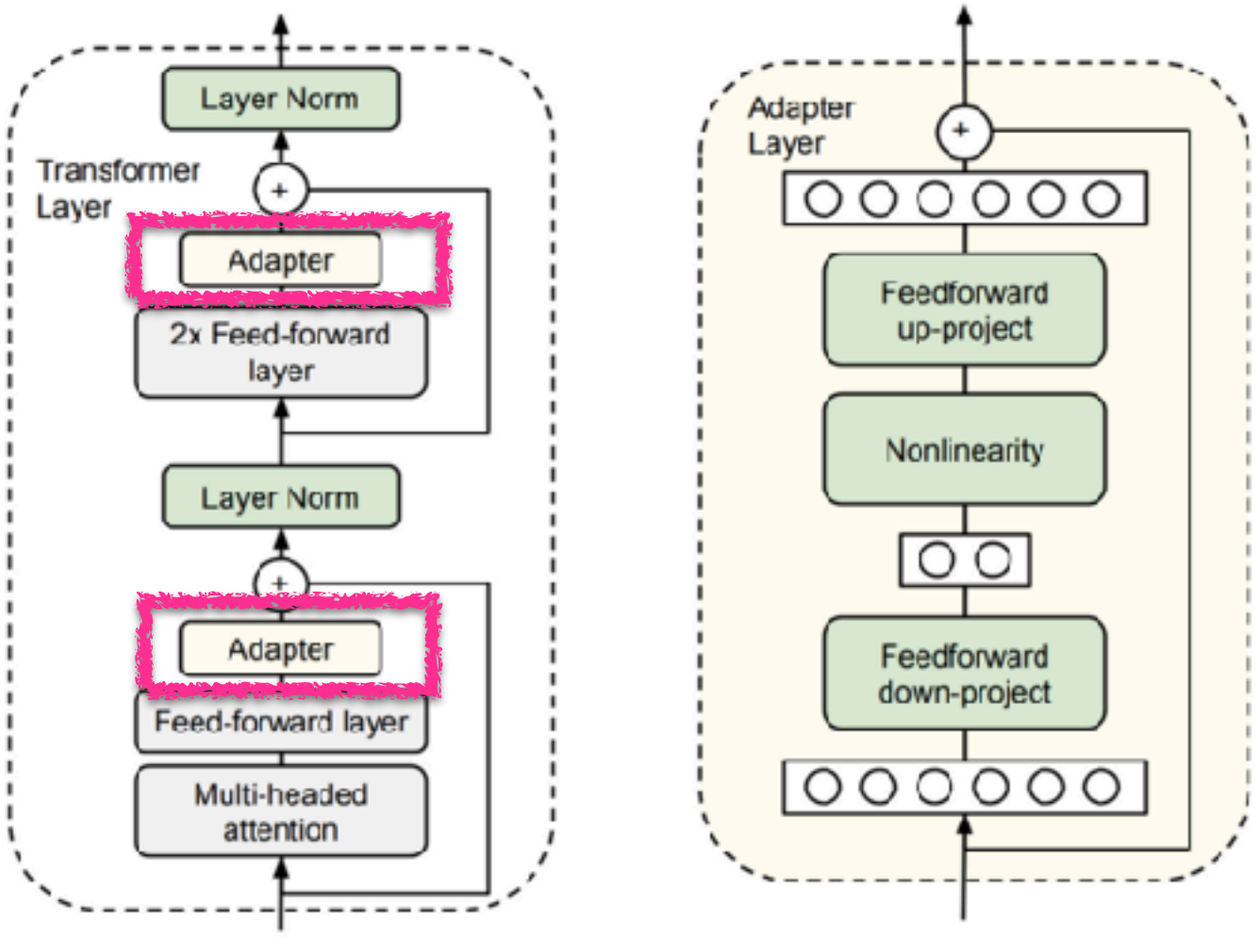
Transfer Learning is broader

- = Largely today's **Pre-train & Fine-tune paradigm**



Sequential TL: Adapters for more efficient fine-tuning

- **Full fine-tuning** vs. **adapters** (Houlsby et al., 2019; Pfeiffer et al., 2020) - see Iryna Gurevich's talk on adapters



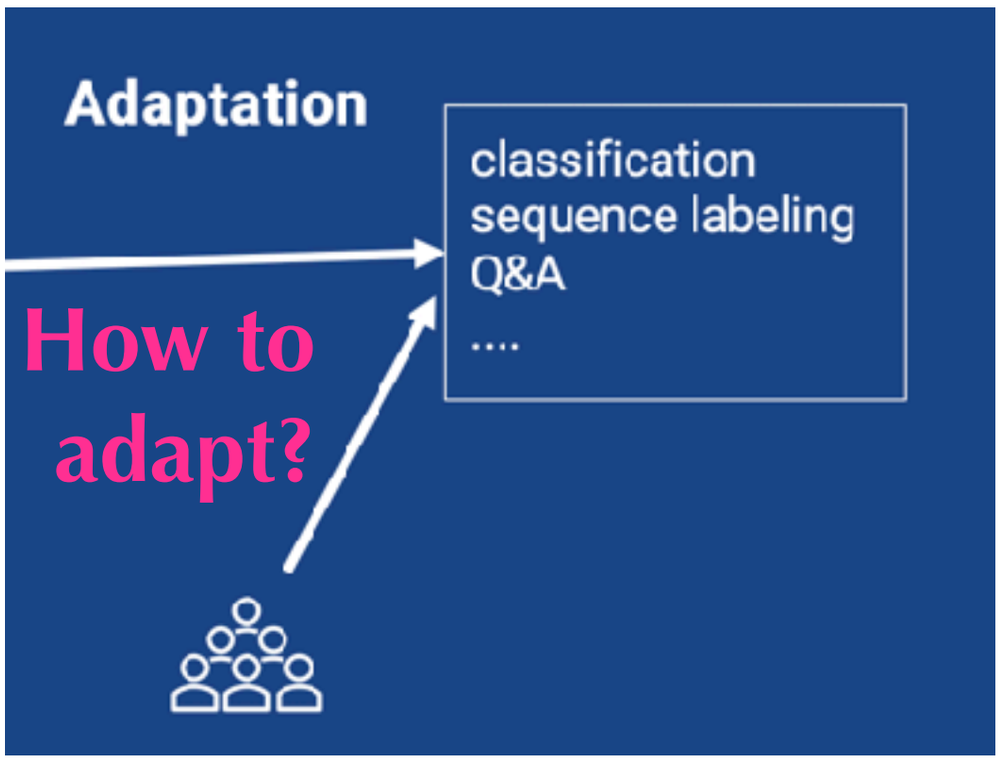
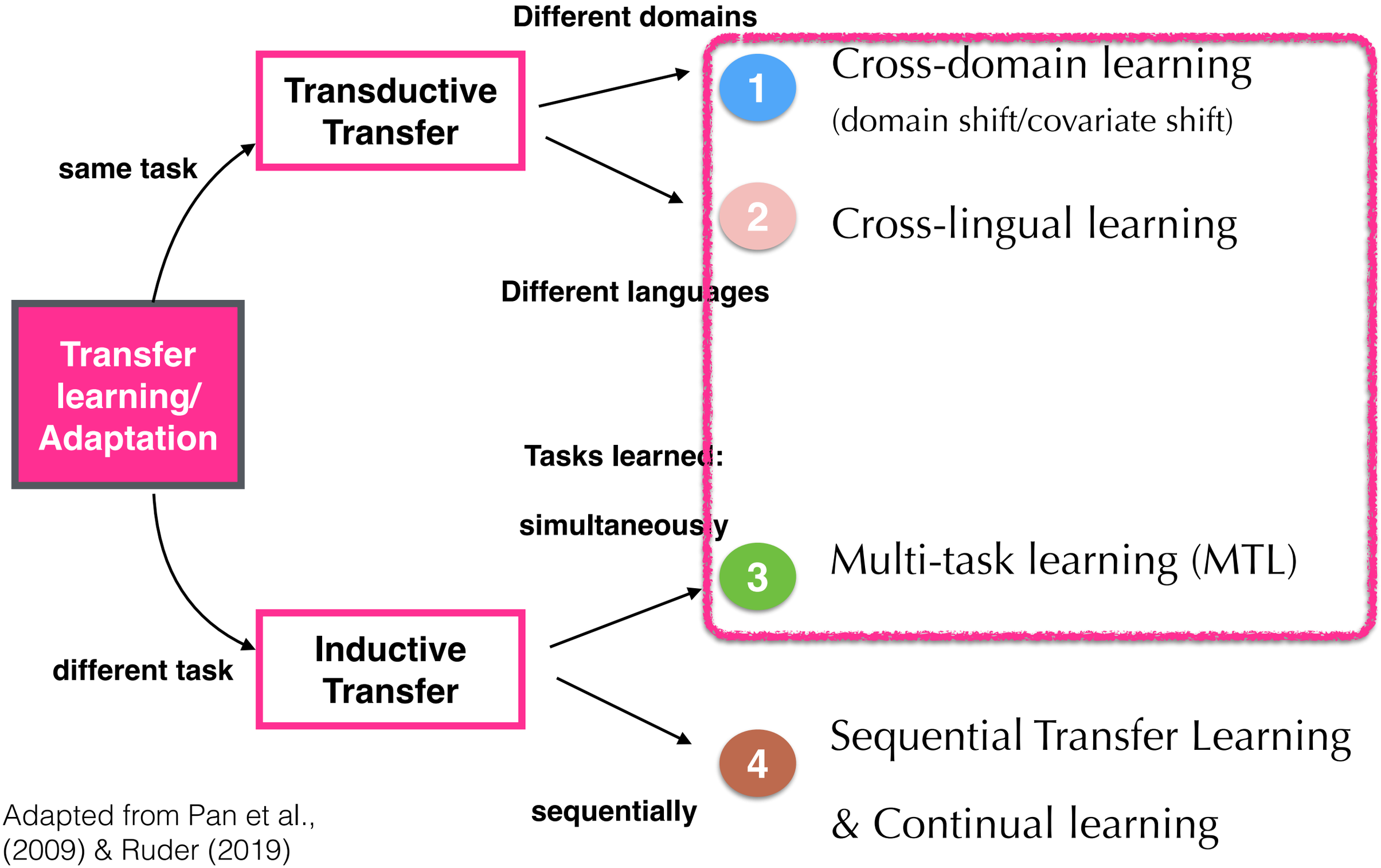


Figure from Houlsby et al., 2019

Roadmap: 3 selected case studies



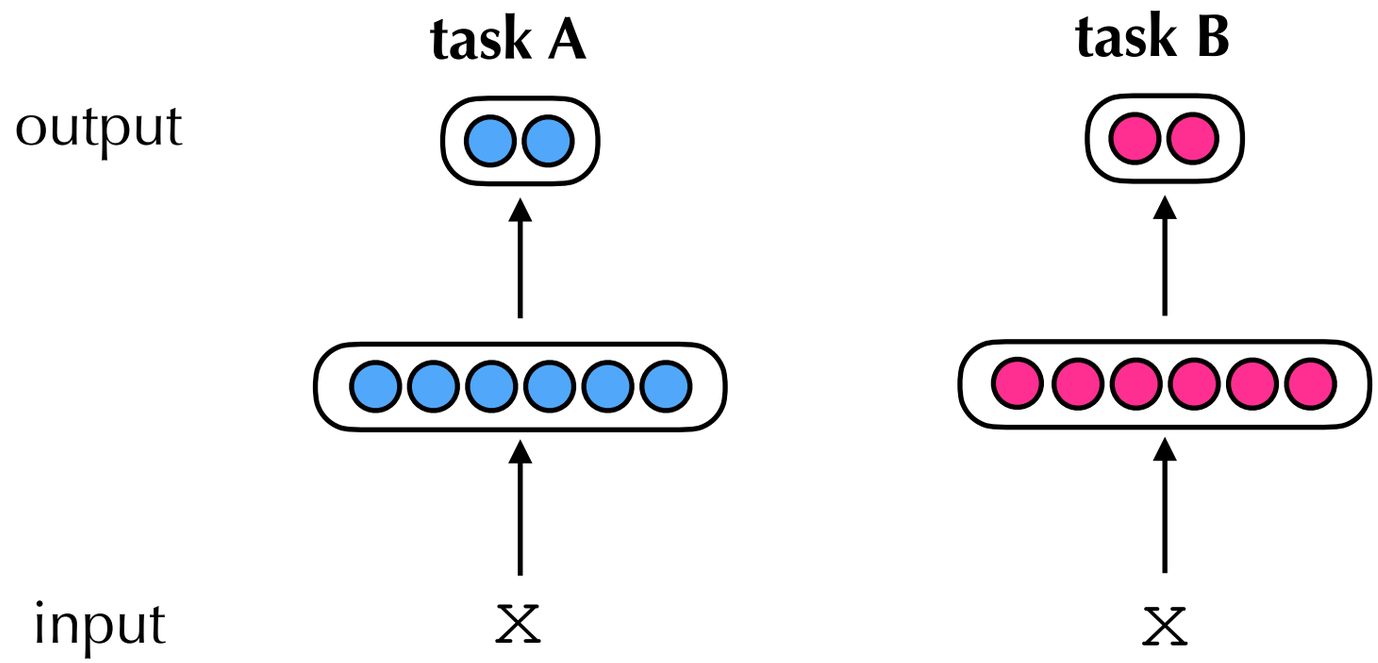
Adapted from Pan et al., (2009) & Ruder (2019)

Overview

- Introduction: The problem of scarce and biased data
- What is Transfer Learning?
  - Three views on Transfer Learning
- **What is Multi-Task Learning?**
  - **Why MTL? Perspectives on MTL**
- **Three selected case studies & Some recent advances**

Multi-task Learning (MTL)

Typical single-task learning



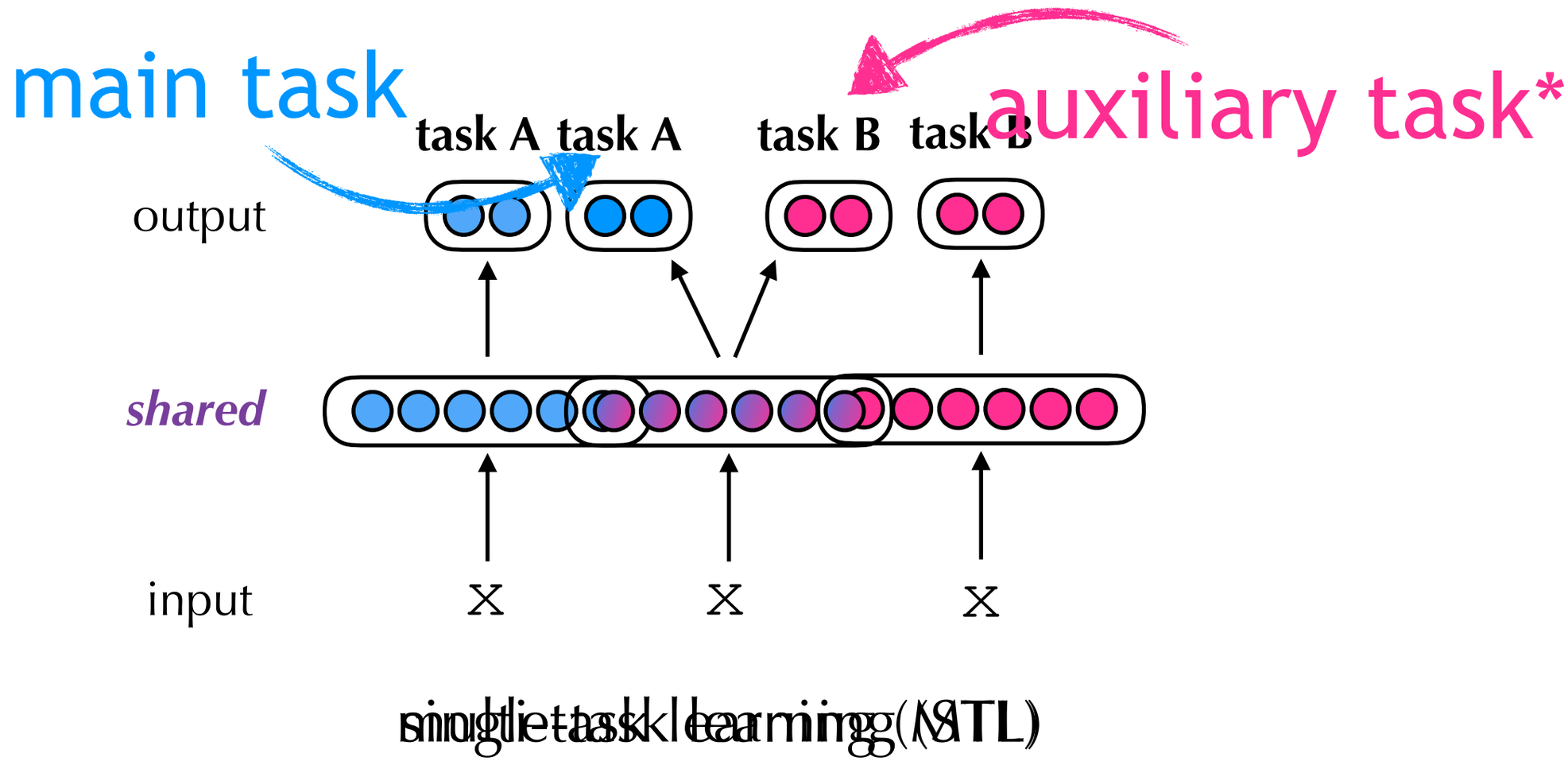
Can we do better?

Example: Learning how to drive a motorbike

main task

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

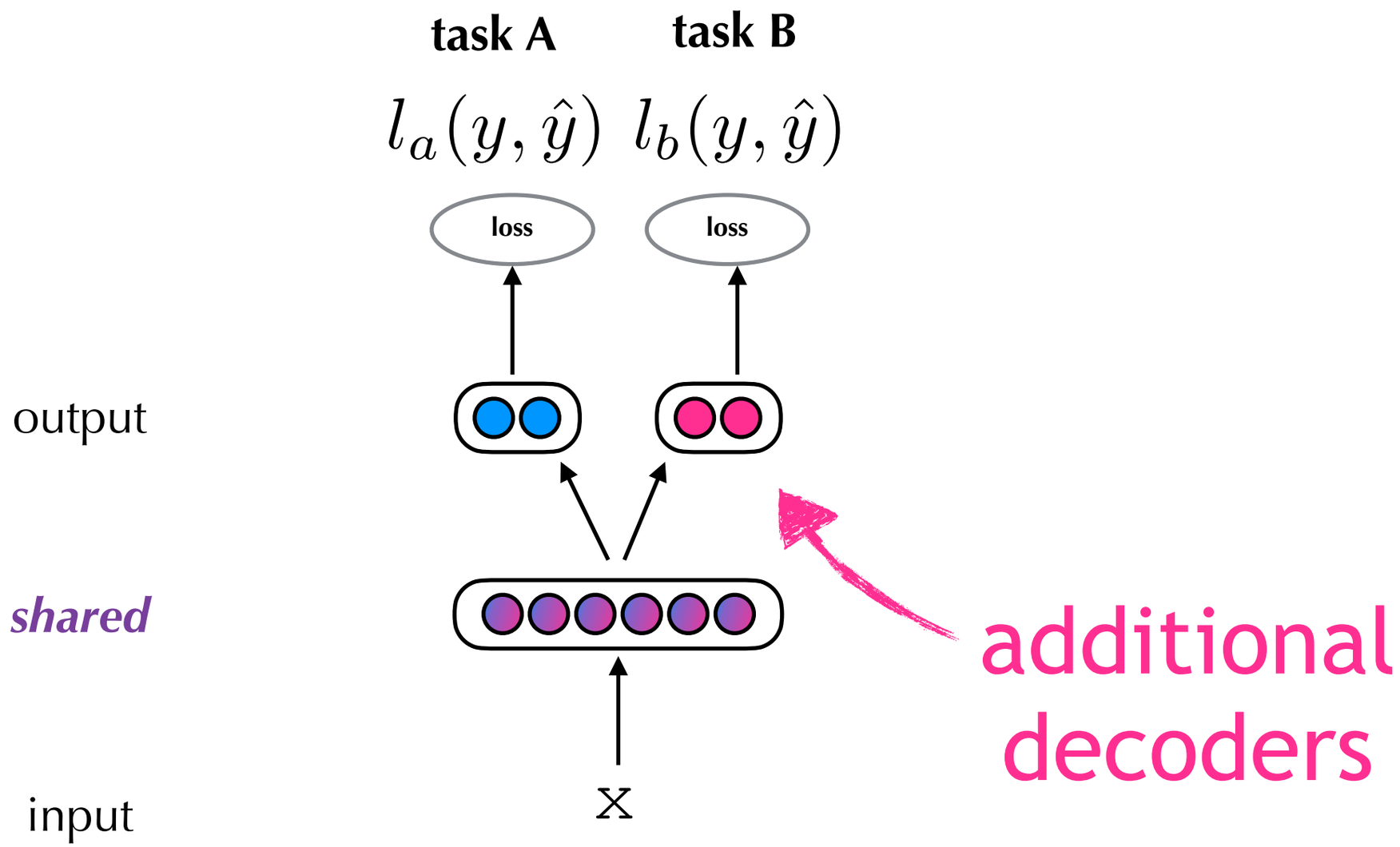
Multi-task Learning (MTL): Key Idea



"[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

MTL in Neural Networks (NNs)



MTL Problem Formulation

Given a set of T tasks $\{\mathcal{D}_\tau\}_{\tau=1}^{T}$

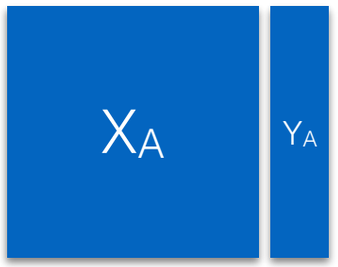
and training data for each task: $\mathcal{D}_\tau = \{(x, y)_i\}_{i=1}^{N_T}$

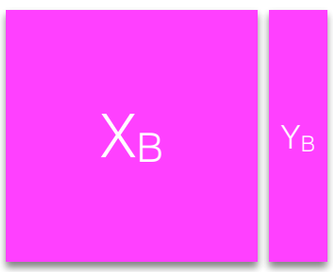
MTL aims to learn a model over all tasks by minimising the loss on the training set:

$$\mathcal{L}(\theta; \{\mathcal{D}_\tau\}_{\tau=1}^{T}) = \sum_{\tau=1}^{T} \sum_{(x,y)_i \in \mathcal{D}_\tau} \lambda_\tau l_\tau(y, \hat{y})$$

MTL Recipe illustrated

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

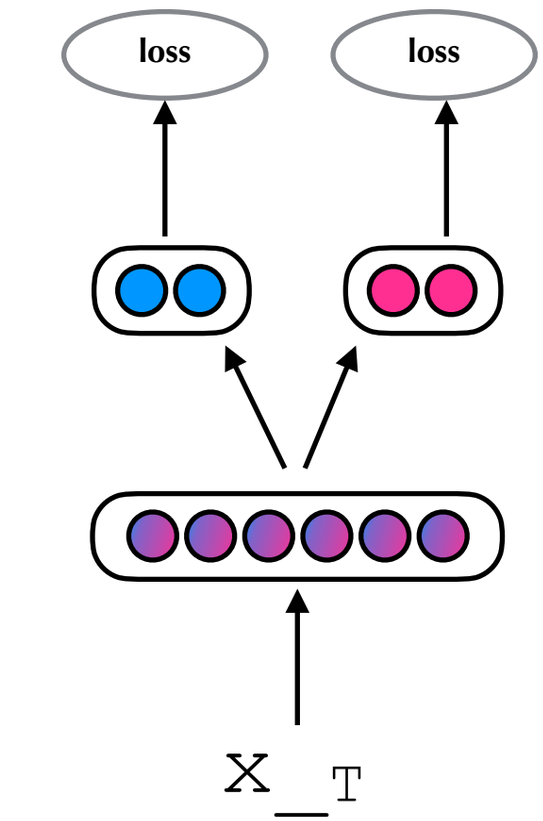




Data

task A **task B**

$$l_a(y, \hat{y}) \quad l_b(y, \hat{y})$$



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

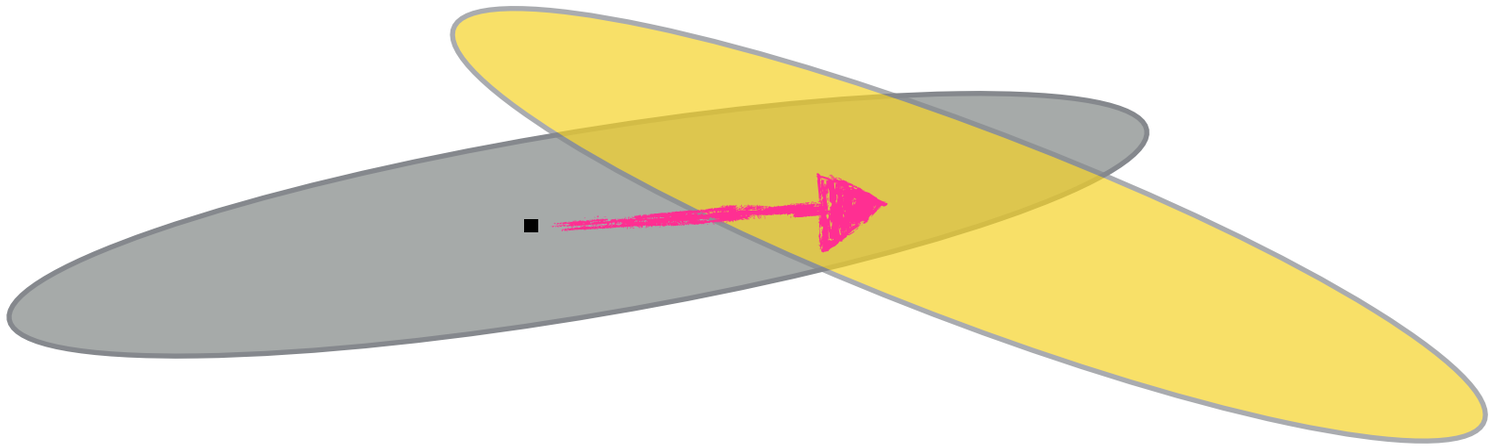
Why does MTL help generalise? (1/4)

- **Attention focusing** (Caruana, 1997): reduced net capacity improves generalisation

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

Why does MTL help generalise? (2/4)

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



Low error for task A

Low error for task B

Why does MTL help generalise? (3/4)

- **Look ahead** - A representation that predicts the future

https://coxlab.github.io/prednet/

Why does MTL help generalise? (4/4)

- **Regularization** (Caruana, 1997): MTL acts as regulariser (Ruder, 2017), reduces the risk of overfitting, particularly on small data.

$$\min_{w} \sum_{i=1}^{n} V(\hat{x}_i \cdot w, \hat{y}_i) + \lambda \|w\|_2^2$$

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

Why does MTL help efficiency? (1/4)

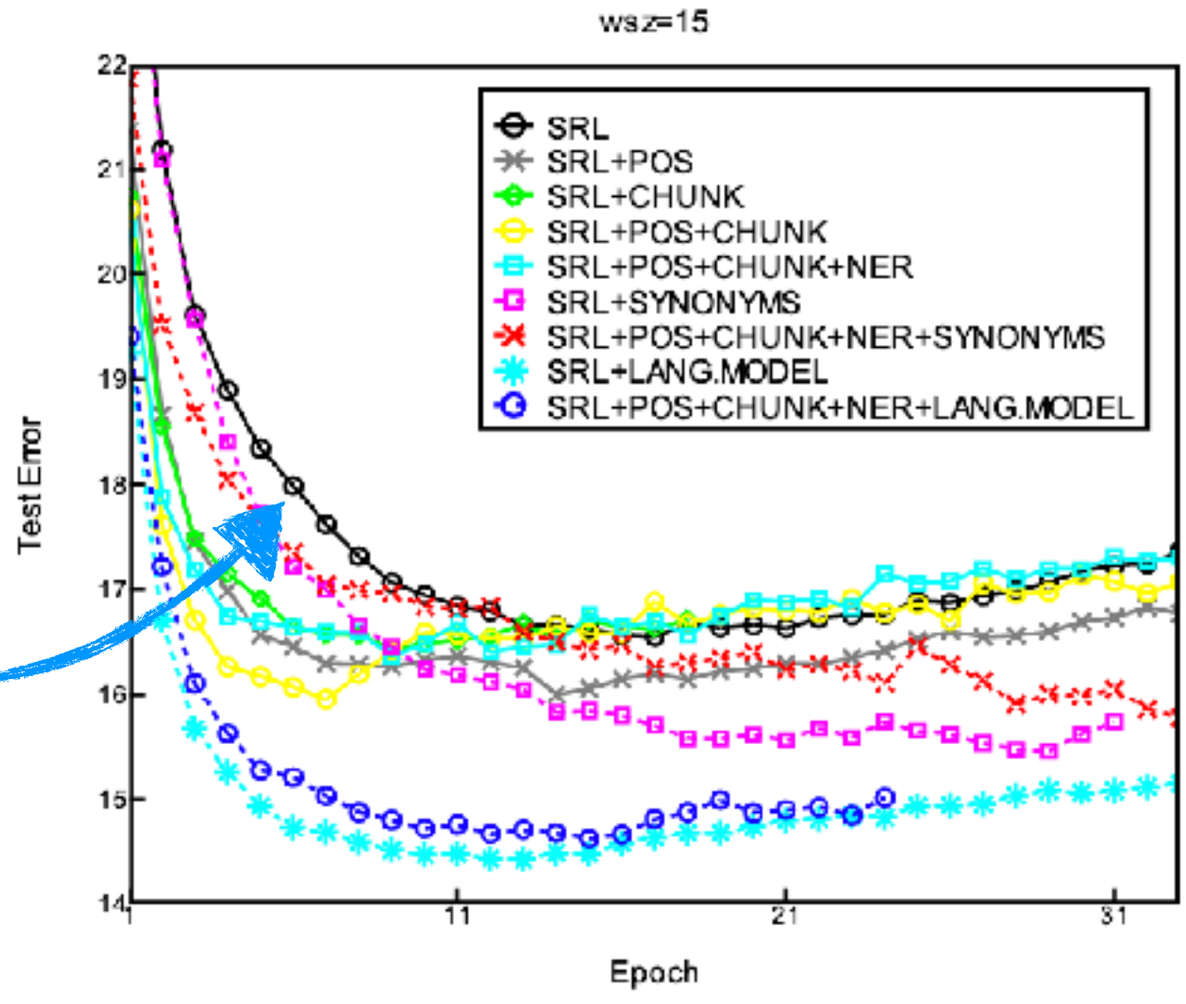
- **Eavesdropping** (Caruana, 1997) - eavedrop on shared representation to learn feature G through task B, which is hard to learn via task A

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? (2/4)

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

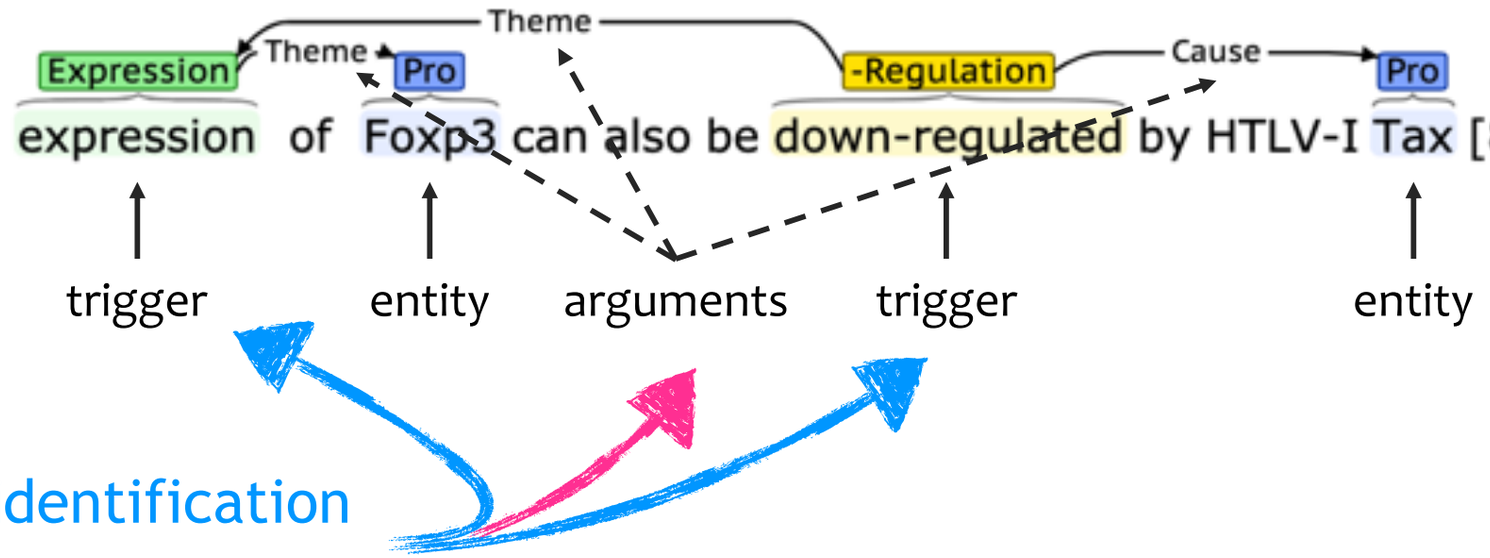
Single task



(Collobert & Weston, 2008, ICML)

Why does MTL help efficiency? (3/4)

- **Replace traditional pipelines** with a single model - Example from biomedical event extraction - Traditional pipeline:



Linearisation + 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

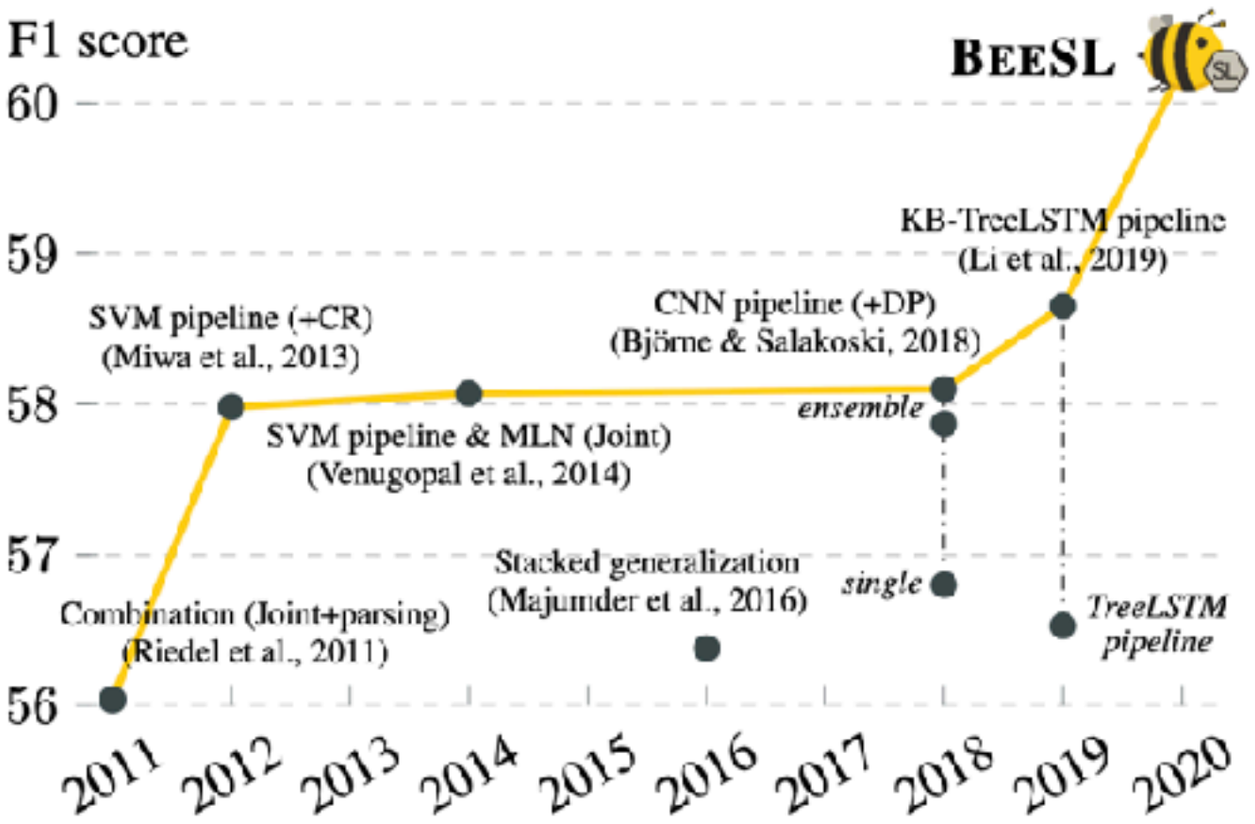


Figure 1: Performance of biomedical event extraction on the BioNLP Genia 2011 test set over time.

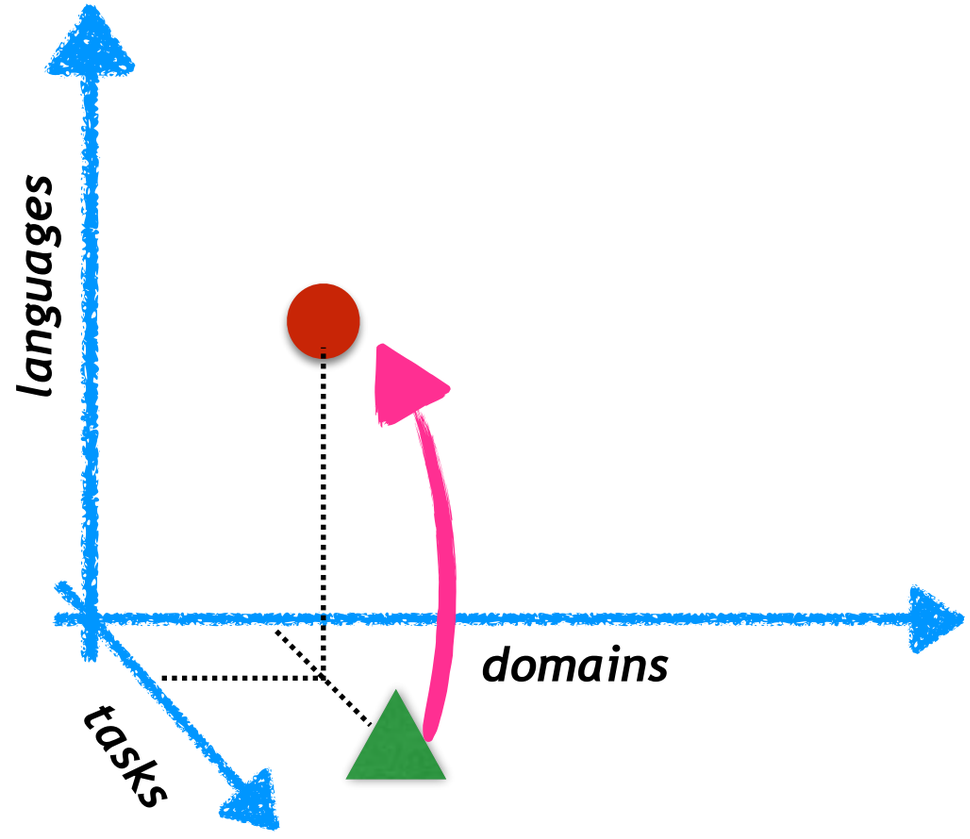
Inference time:
sentences/min

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

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

Why does MTL help efficiency? (4/4)

- **Reduces the need of labeled data** - generalisation via prediction of auxiliary task (e.g. data from other tasks/languages)

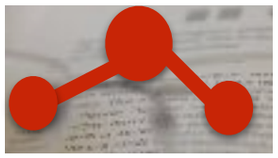


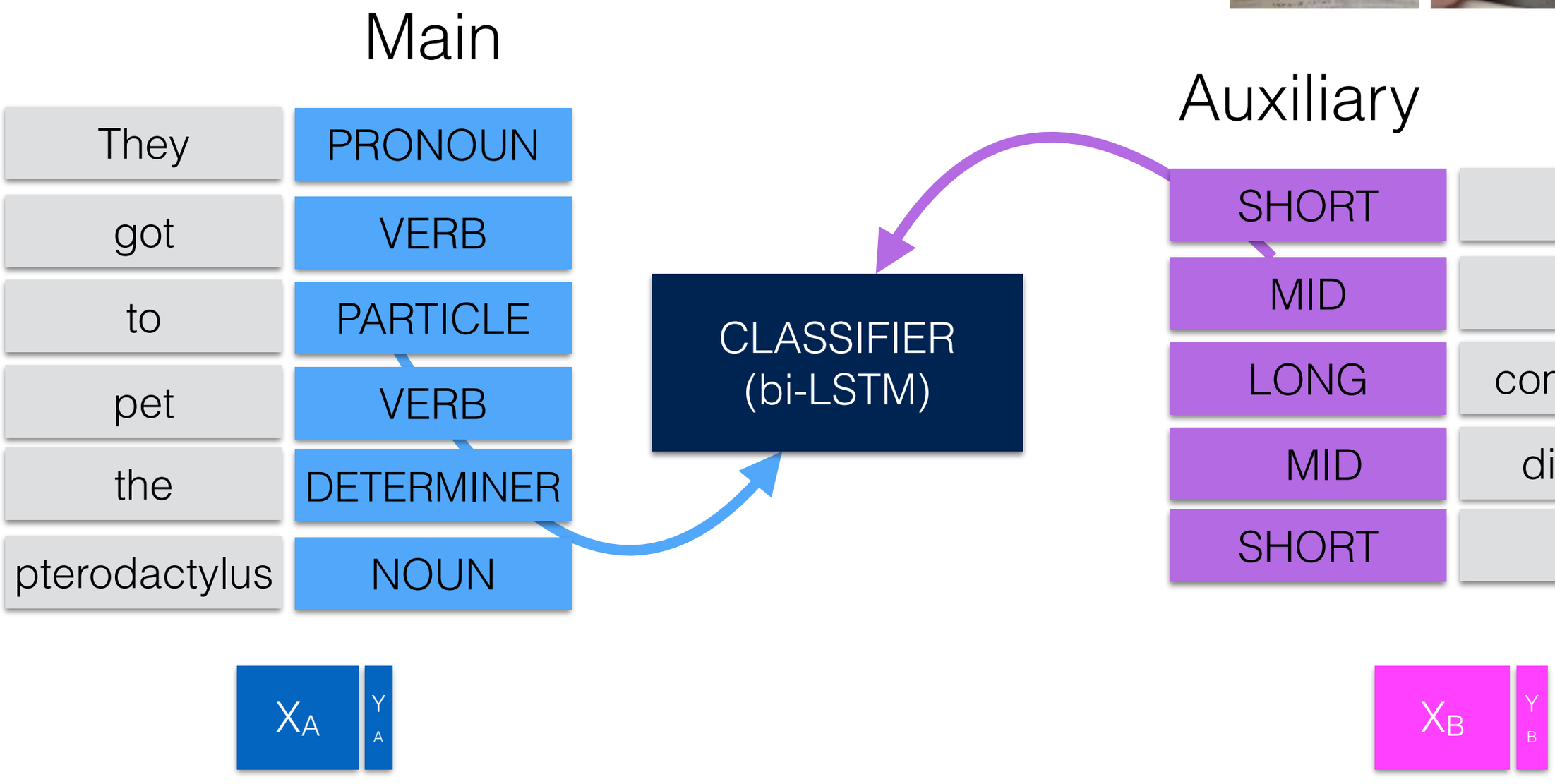
Perspectives on MTL

MTL: learning from distinct sources

e.g., from other languages but also more remote sources like cognitive human data (gaze, keystrokes)

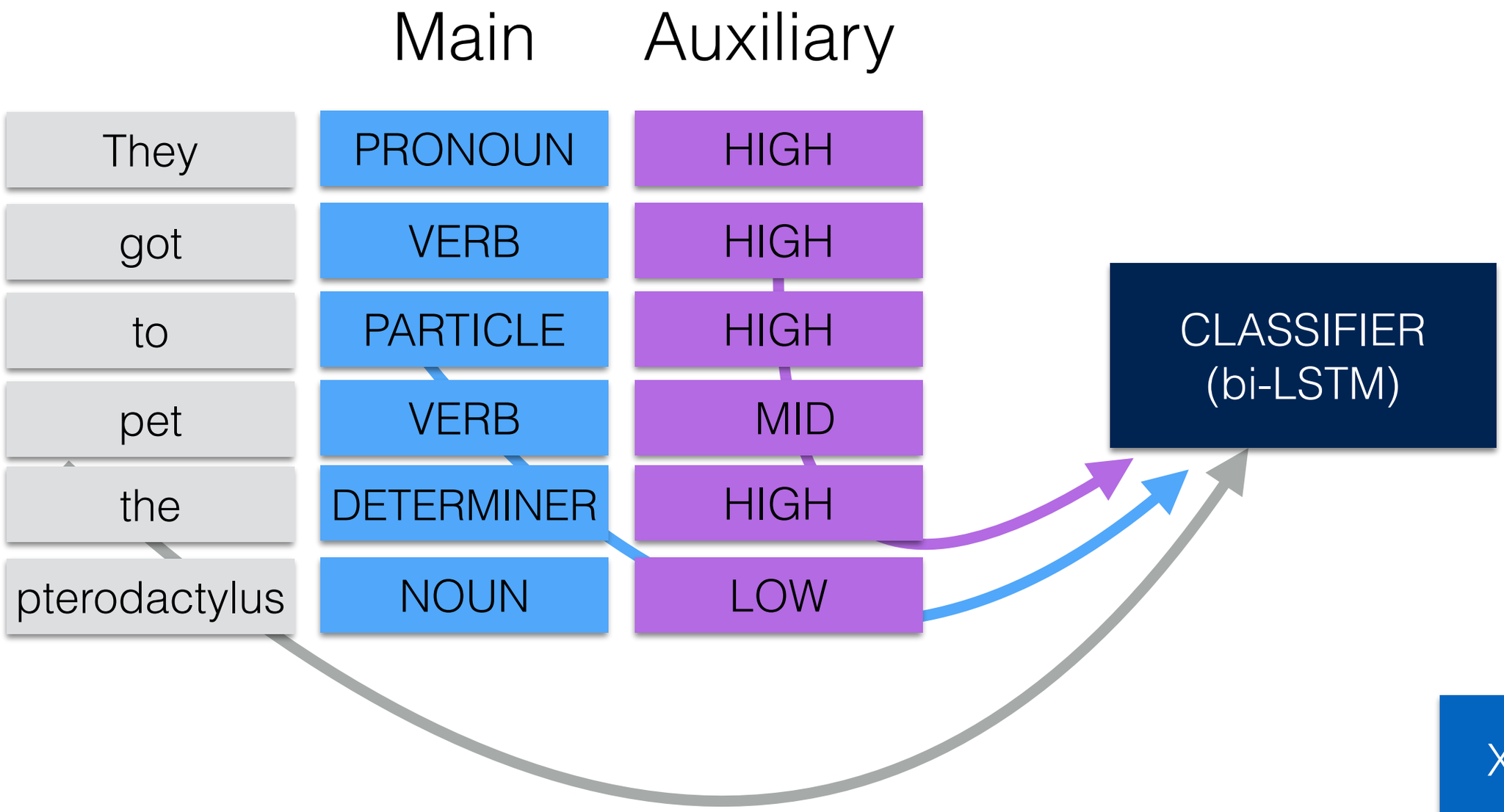
(Klerke et al. 2016 NAACL), (Plank 2016 COLING), (Barrett & Hollenstein, 2020)

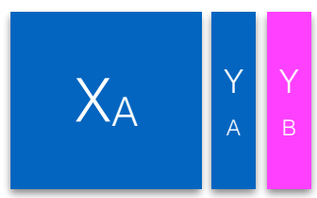


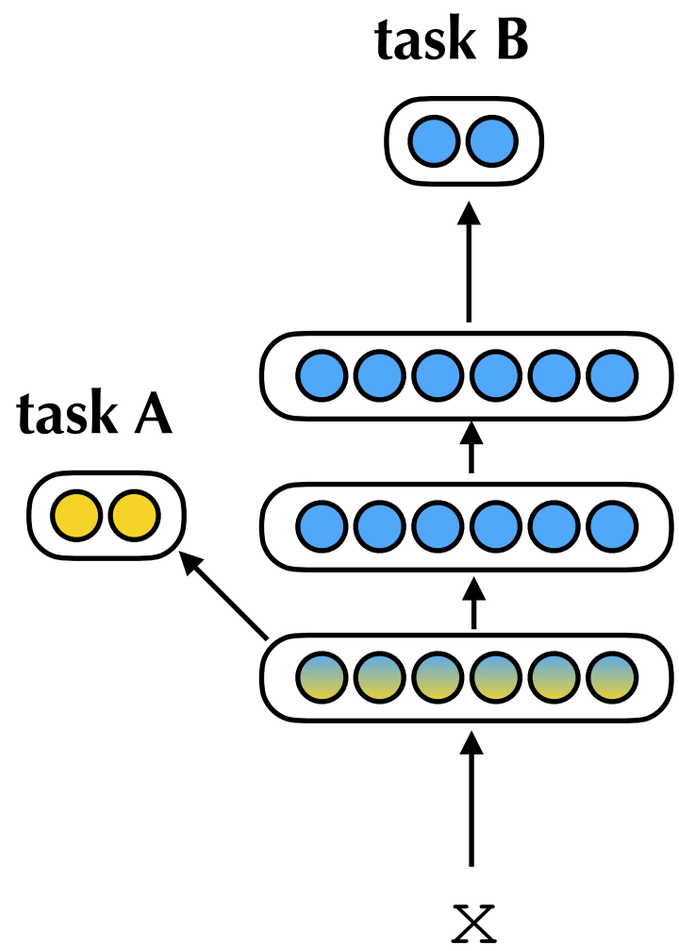
MTL: learning from distinct views

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)





MTL: Learning as selective sharing/ cascading



e.g., (Søgaard & Goldberg 2016 ACL)

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=I39sxwYKIEE

Data-derived auxiliary tasks

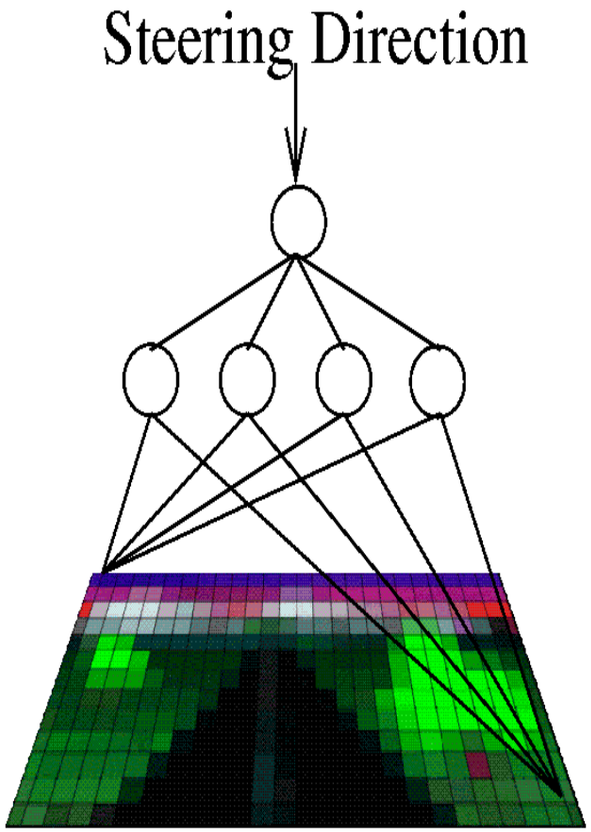
For our MTL experiments, eight additional tasks were used:

- 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)

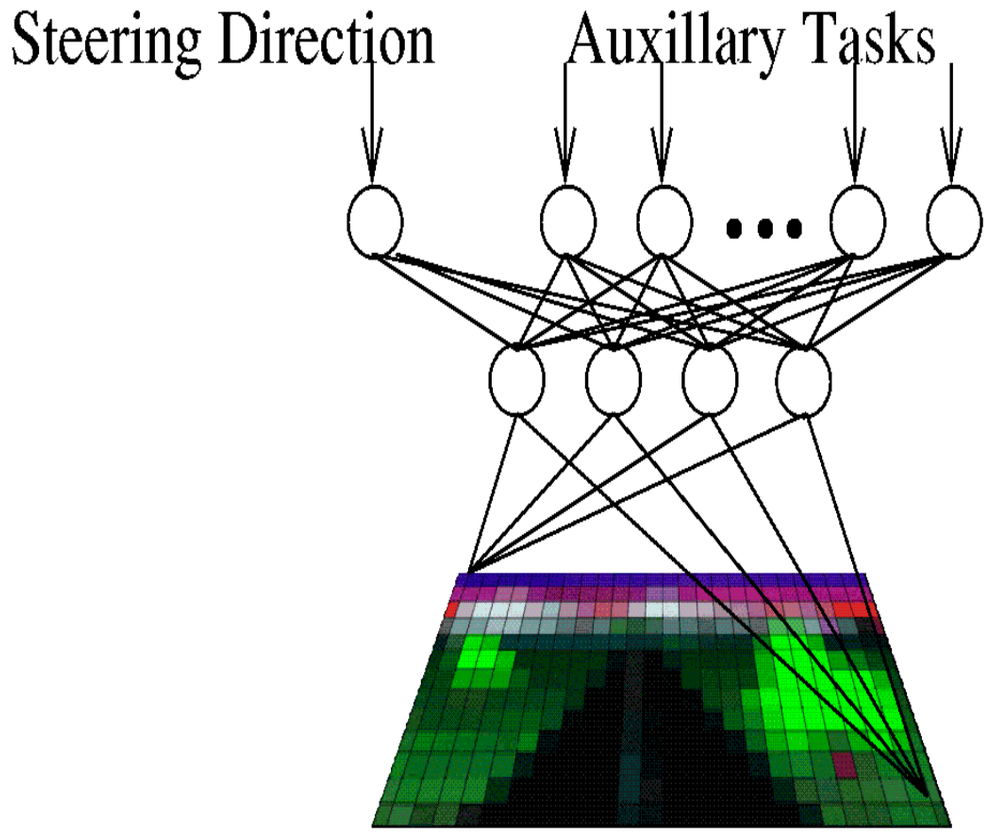
CMU Alvinn MTL (Caruana 1998)

Note: here all task labels computable from data

Alvinn MTL



Single
Task Leaning

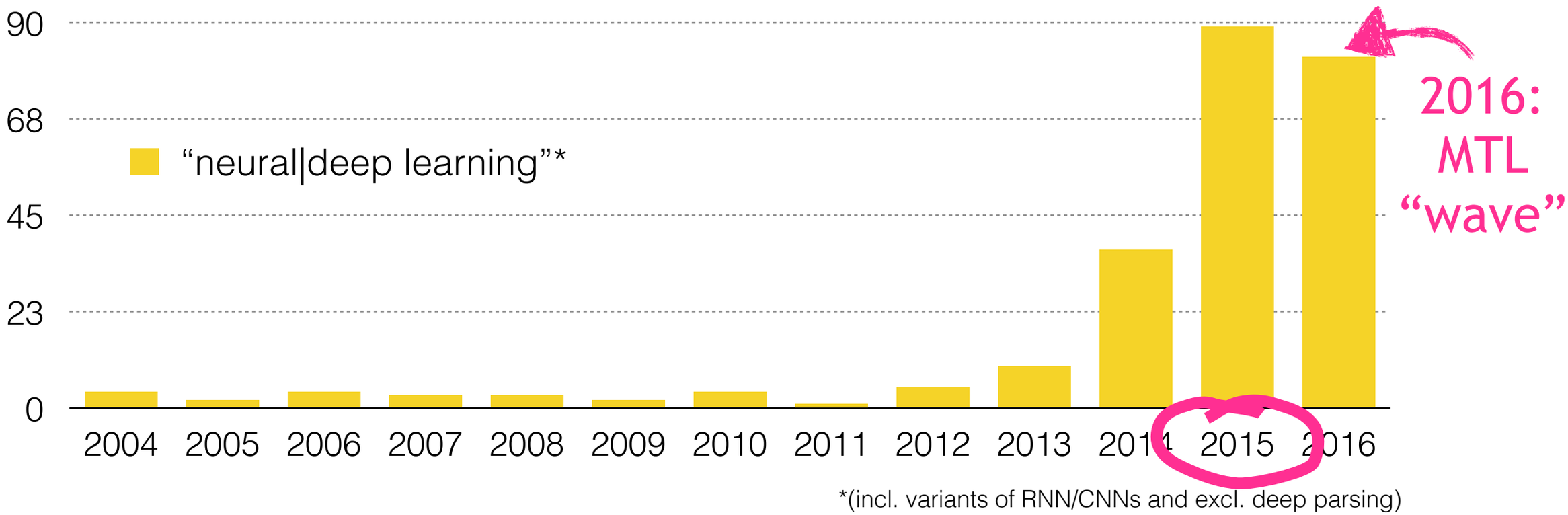


MultiTask Learning

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)



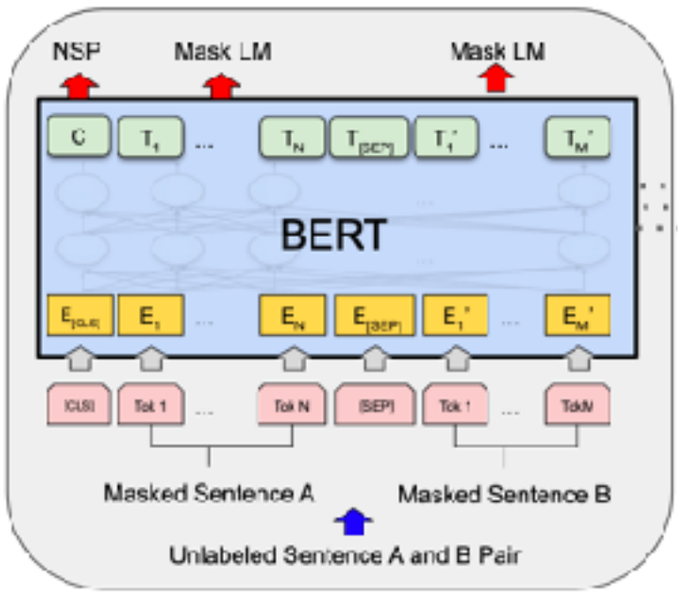
Titles of papers in ACL anthology (from 2004)

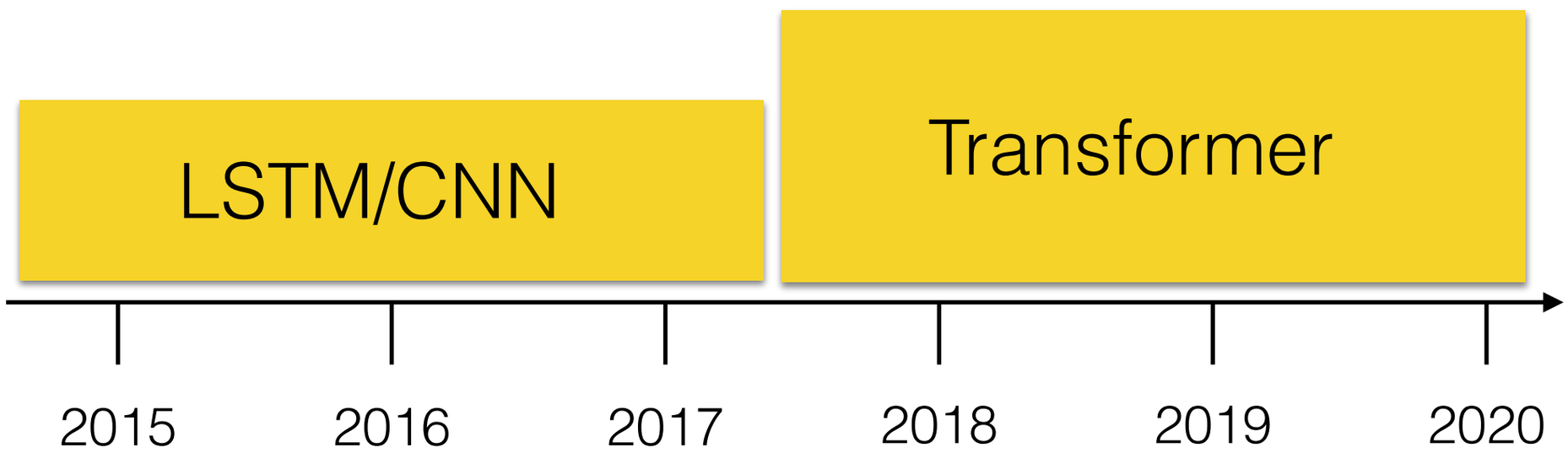
DL "tsunami" (Manning, 2015)

MTL "wave" (Ruder & Plank, 2018)

Today: MTL everywhere!

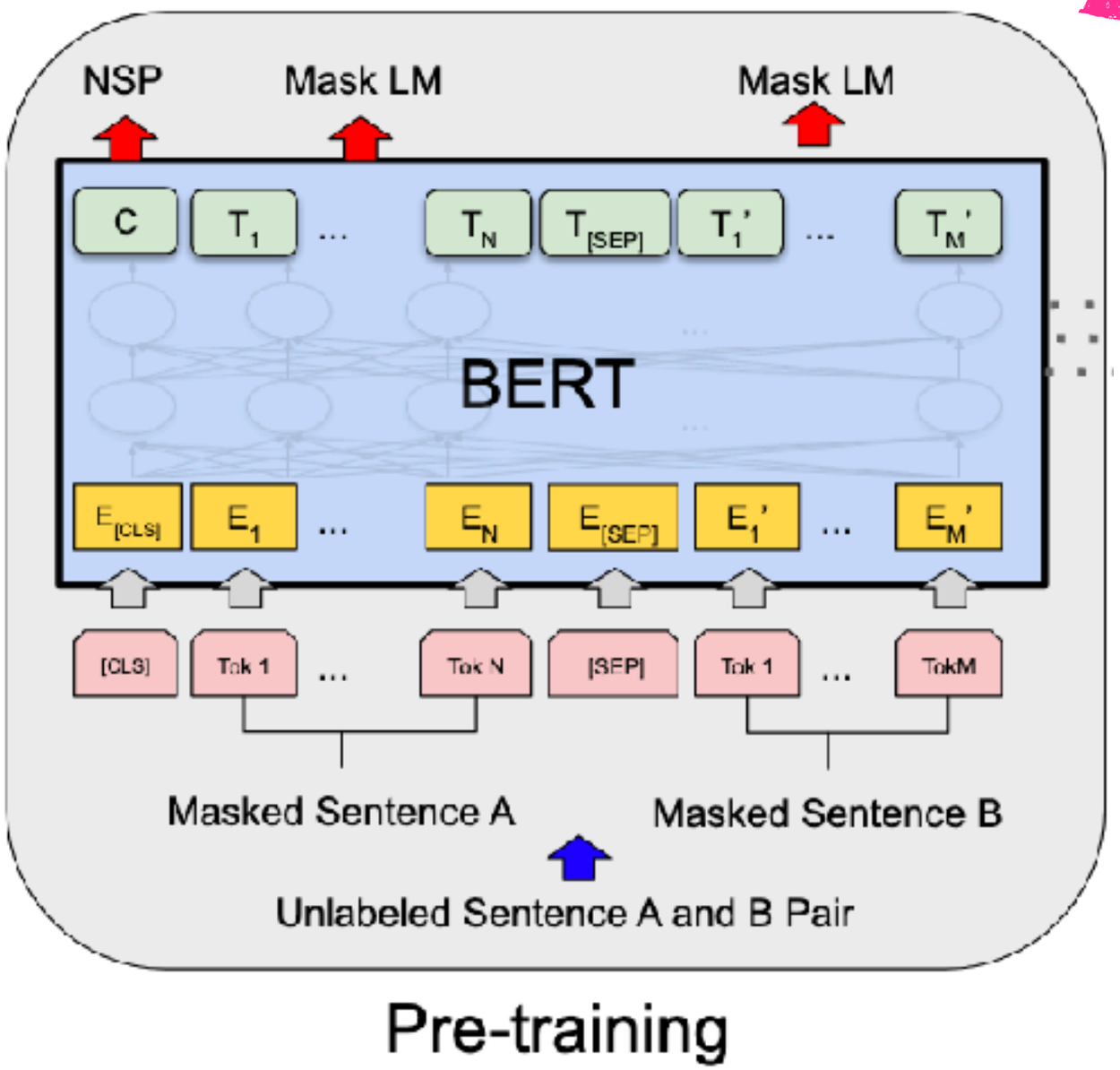
First MTL wave
(2016-2017)





Vaswani et al., 2017 NeurIPS: Attention is all you need

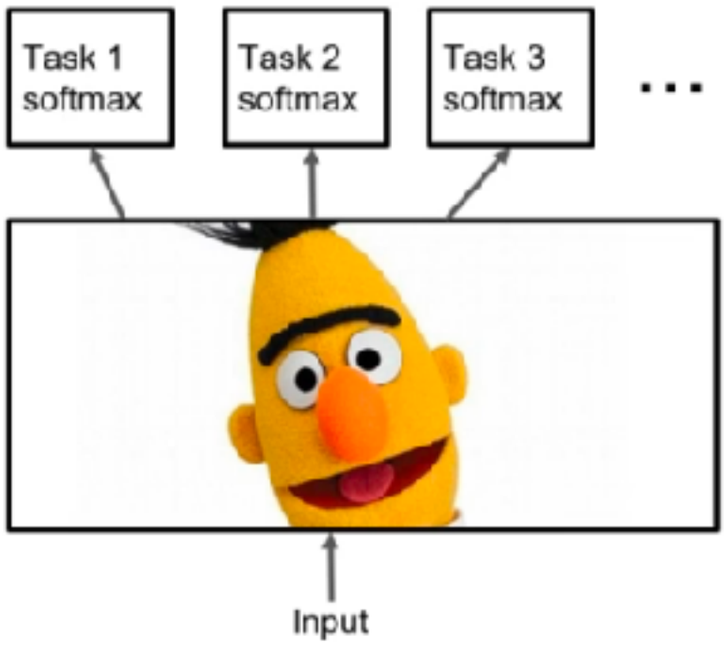
BERT: Multi-task pre-training



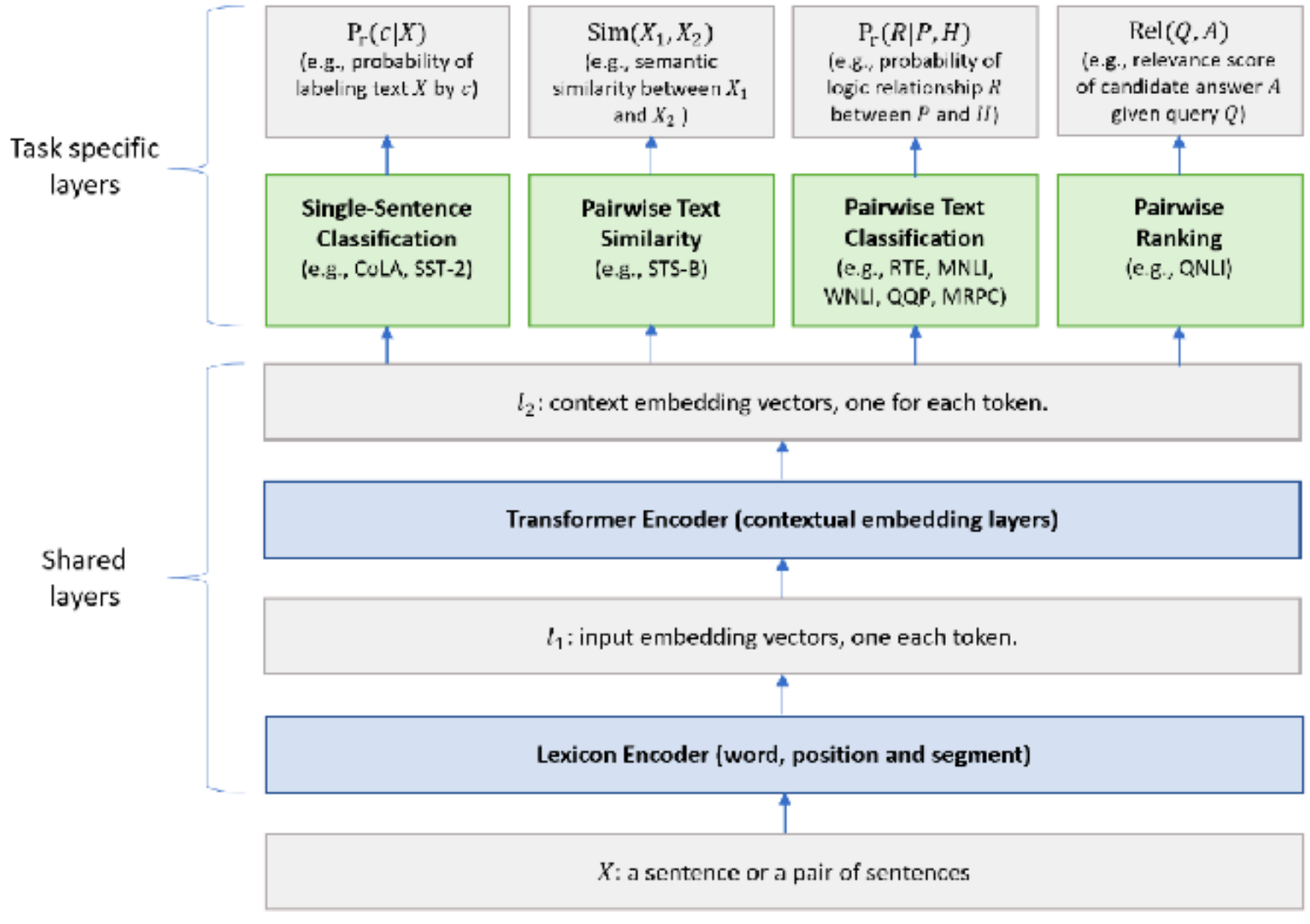
unsupervised MTL objectives: MLM + NSP

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

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



Slide adapted from Clark et al, 2019



MT-DNN by Liu et al., ACL 2019

Roadmap

1. Cross-domain learning
2. Cross-lingual learning
3. Multi-task learning with Fortuitous Data
4. Continual Learning

Learning to select data for transfer learning with Bayesian optimization

Sebastian Ruder and Barbara Plank
EMNLP 2017

Part 1

Data Setup:
Multiple Source Domains

Source domains (labeled)

Idea: Learning to select the most relevant data from multiple sources using data metrics

Motivation

Why? Why don't we just train on all source data?

- **Prevent negative transfer**
- e.g. "predictable" is negative for  , but positive in 

Prior approaches to data selection:

- use a single similarity metric in isolation;
- focus on a single task.

Our approach

Intuition

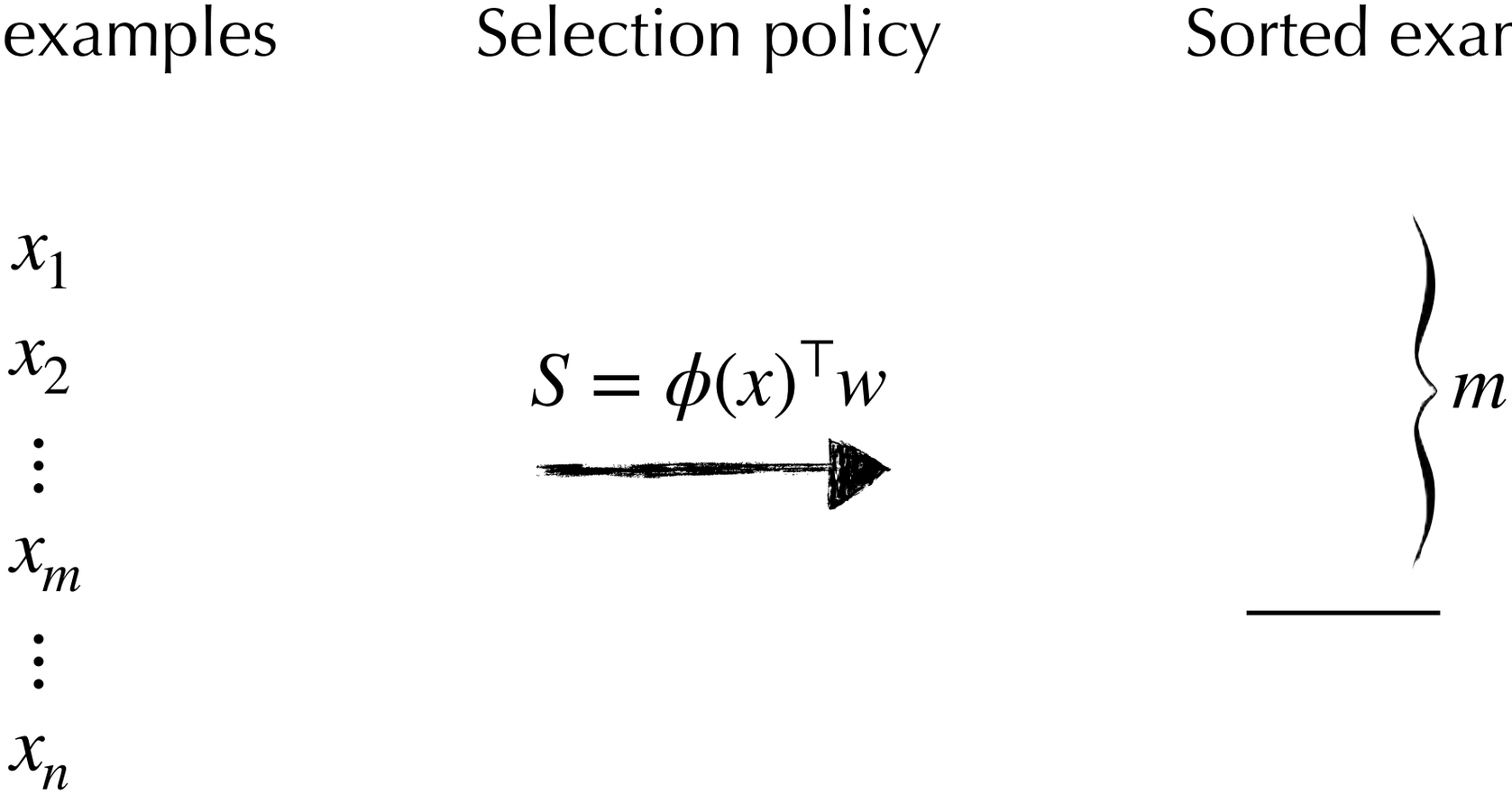
- Different tasks and domains require different notions of similarity.

Idea

- Learn a data selection policy using Bayesian Optimization

- Evaluate on multiple tasks

Our approach



Source domains

- Related: curriculum learning (Tsvetkov et al., 2016)

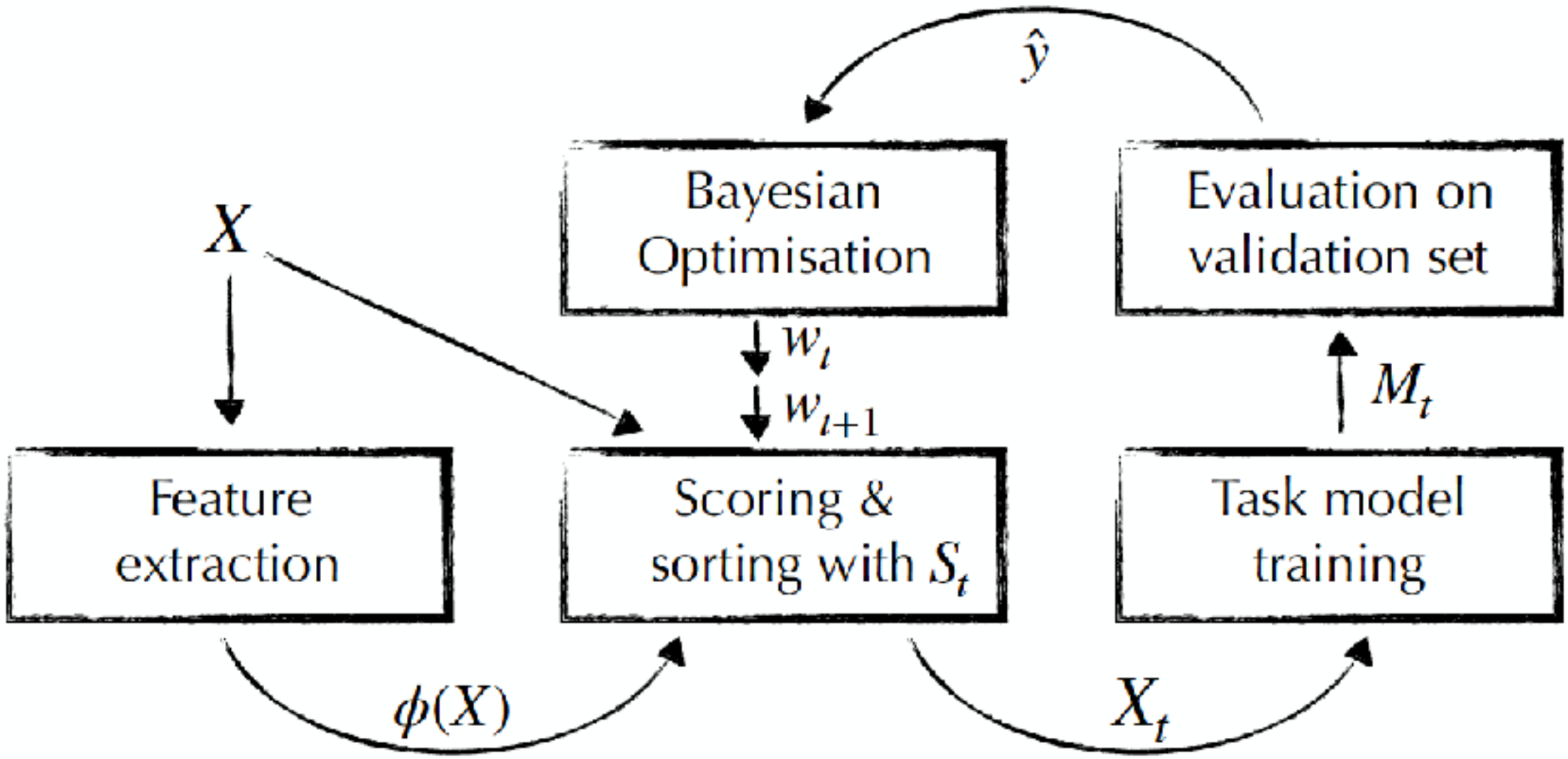
Tsvetkov, Y., Faruqui, M., Ling, W., & Dyer, C. (2016). Learning the Curriculum with Bayesian Optimization for Task-Specific Word Representation Learning. In *Proceedings of ACL 2016*.

Bayesian Data Selection Policy

$$\mathcal{S} = \phi(\mathcal{X}) \cdot w^T$$

learned feature weights

different similarity/diversity features



Features $\phi(X)$

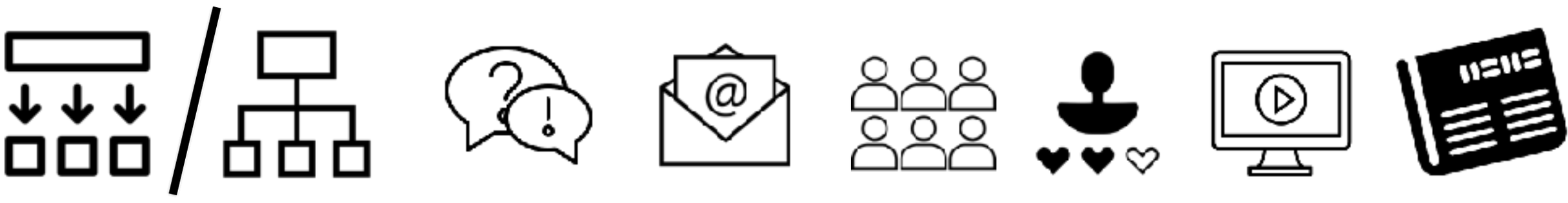
- **Similarity:**
  Jensen-Shannon, Rényi div, Bhattacharyya dist,
  Cosine sim, Euclidean distance, Variational dist
  - **Representations**:
    Term distributions,
    Topic distributions,
    Word embeddings

- **Diversity**: #types, TTR, Entropy, Simpson's index, Rényi entropy, Quadratic entropy

Data & Tasks

Three tasks: **Domains:**

Sentiment analysis on Amazon reviews dataset (Blitzer et al., 2007)

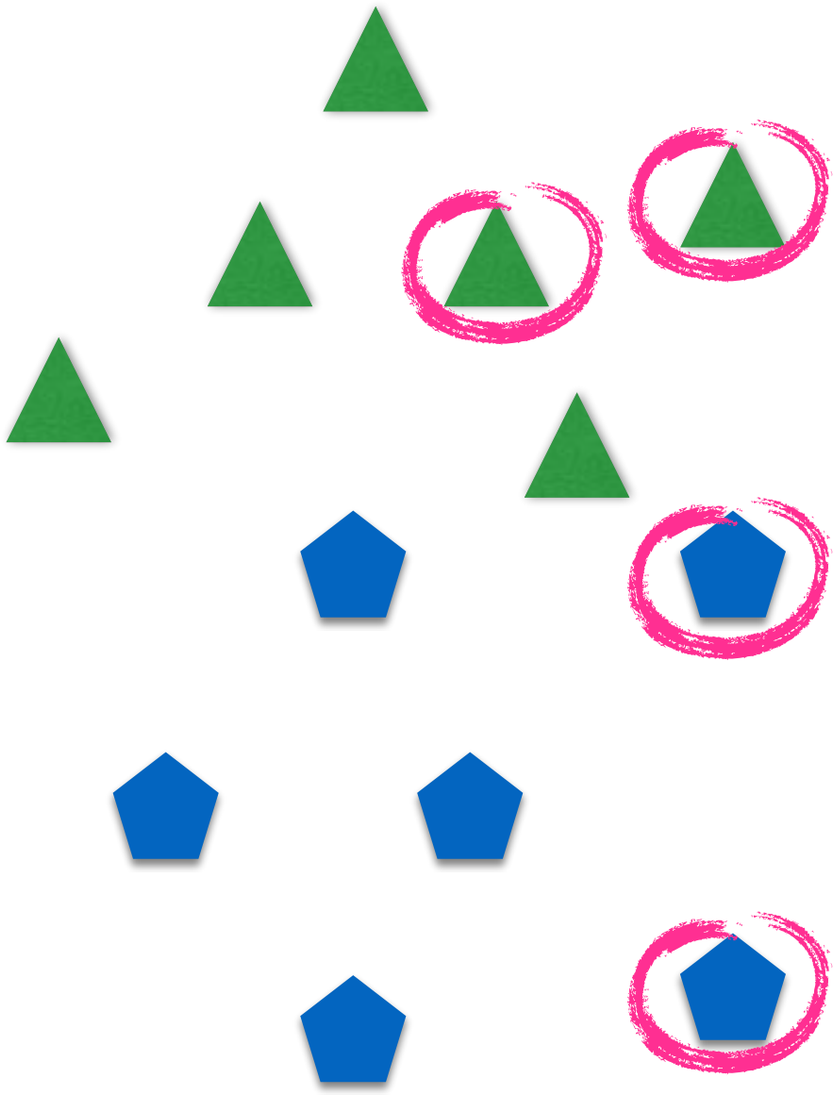


POS tagging and dependency parsing on SANCL 2012 (Petrov and McDonald, 2012)

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*.

Petrov, S., & McDonald, R. (2012). Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*.

Setup & Baselines

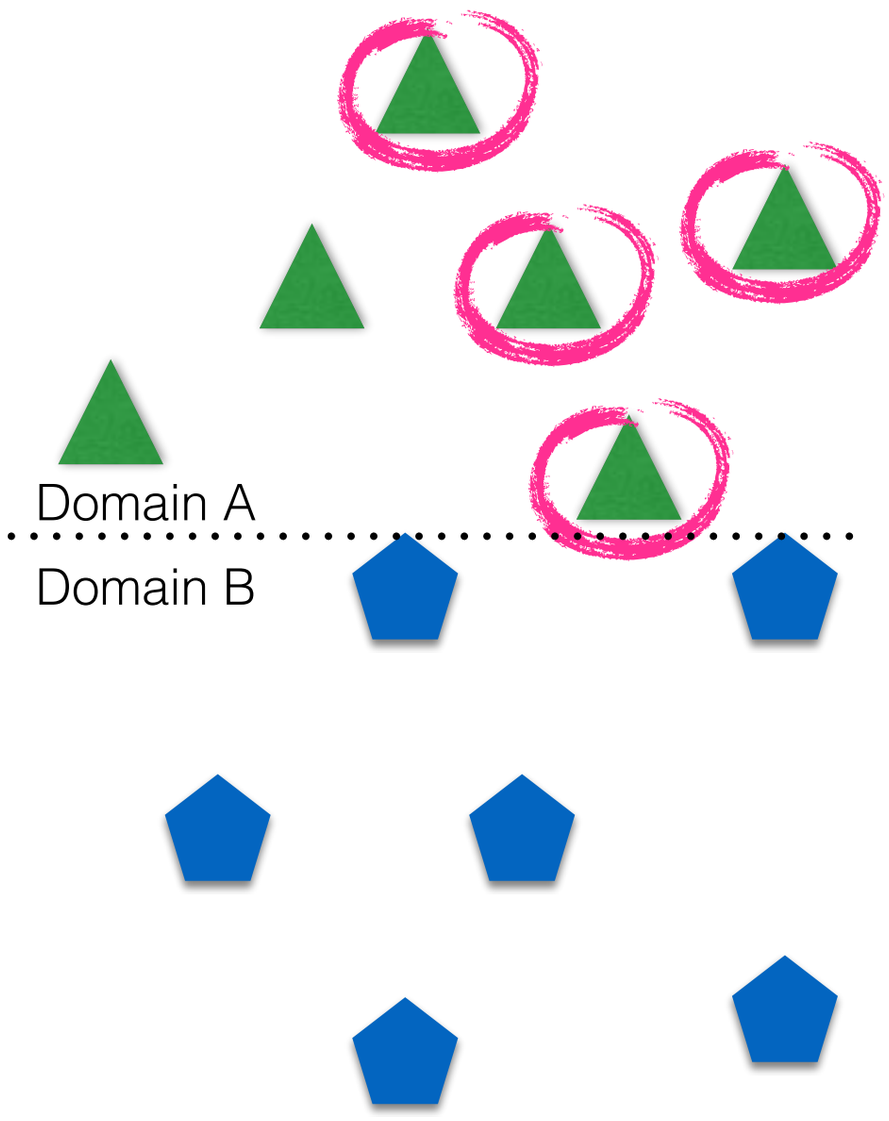


closest *m*
by JS (examples)

Source domains

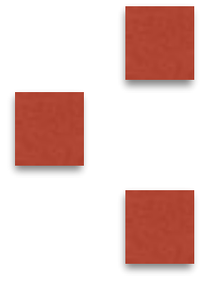
Target domain

Setup & Baselines



Source domains

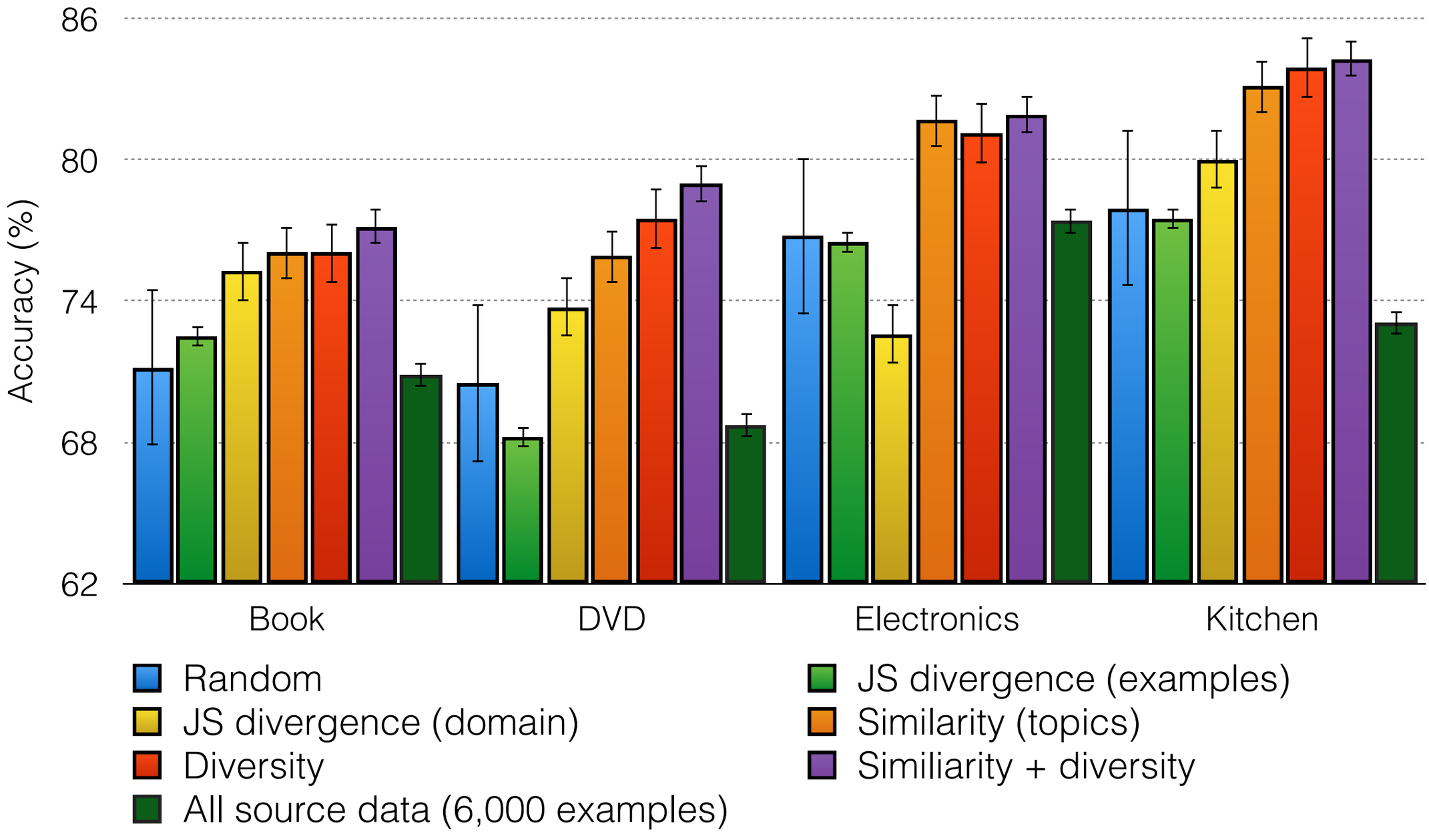
closest *m*
by *JS (domain)*



Target domain

Sentiment Analysis Results

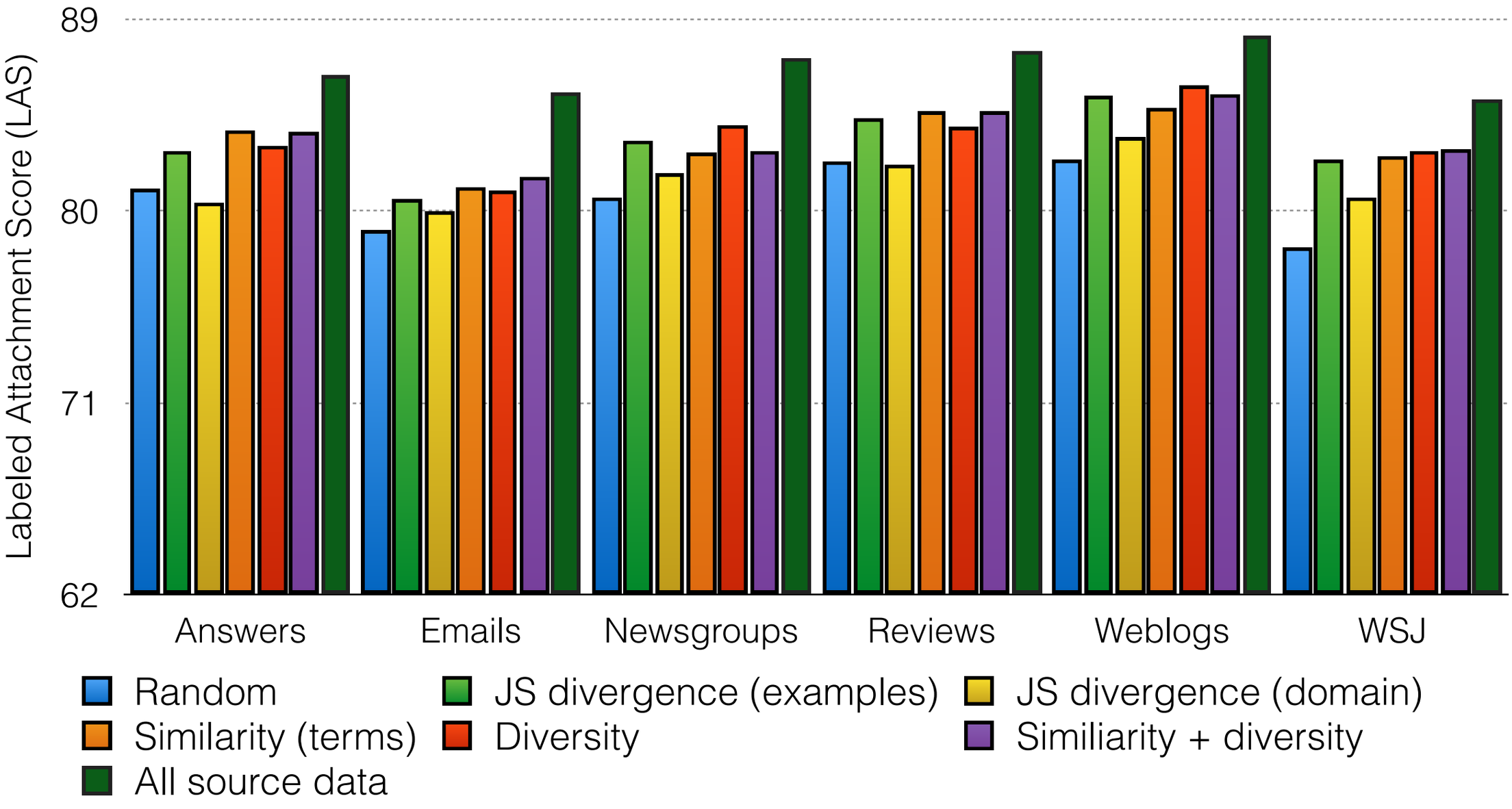
Selecting m=2,000 from 6,000 source domain examples



- Selecting relevant data is useful when domains are very different.

Dependency Parsing Results

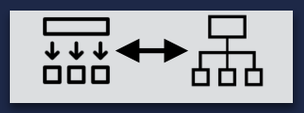
Selecting 2,000 from 14-17.5k source domain examples (POS tagging results in paper, similar trend)



(BIST parser, Kiperwasser & Goldberg, 2016)

Take-aways

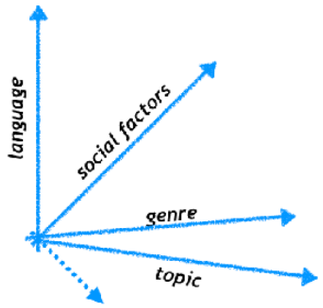
- Domains & tasks have different notions of similarity. Learning a task-specific data selection policy helps
- Can avoid negative transfer (best result for sentiment)



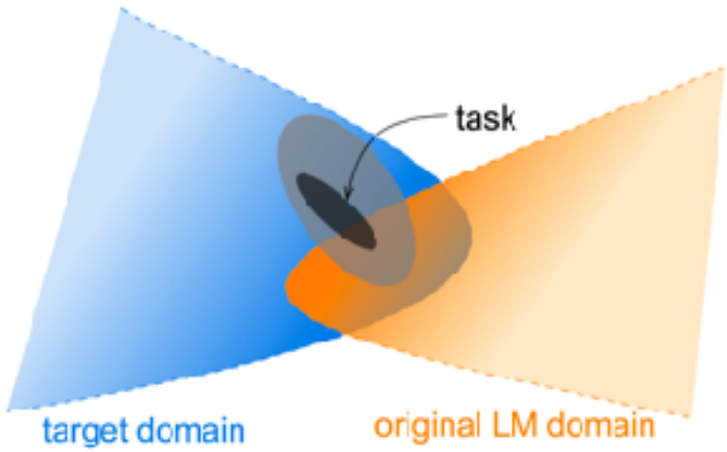
- More results in the paper: The learned policy **transfers** (to some extent) across models, tasks, and domains

Code: https://github.com/sebastianruder/learn-to-select-data

Variety Space & Recent Works on Data Selection



- We saw that domains are not as simplistic and clear-cut as often considered (variety space, Plank 2016)



- Domains are still relevant in BERTology times (Gururangan et al., 2020 ACL)

- Beneficial for data selection in NMT (van der Wees, 2015, 2017; Aharoni & Goldberg, 2020)

Roadmap

1. Cross-domain learning
2. Cross-lingual learning
3. Multi-task learning with Fortuitous Data
4. Continual Learning

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




et al., NAACL 2021

Example: Languages in EU covered by voice assistants

*as of March, 2020

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

Task: Slot and Intent Detection

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

Intent: SearchScreeningEvent

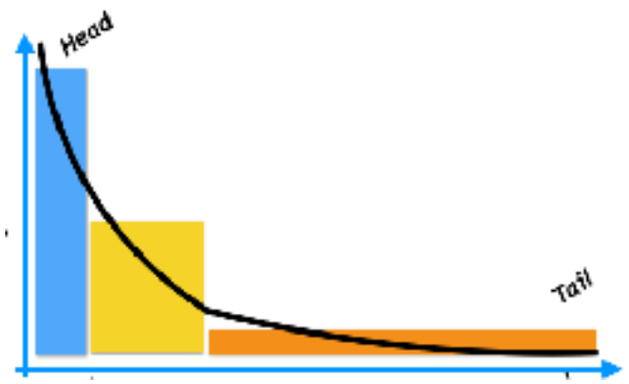
Task: Slot and Intent Detection

Slots:

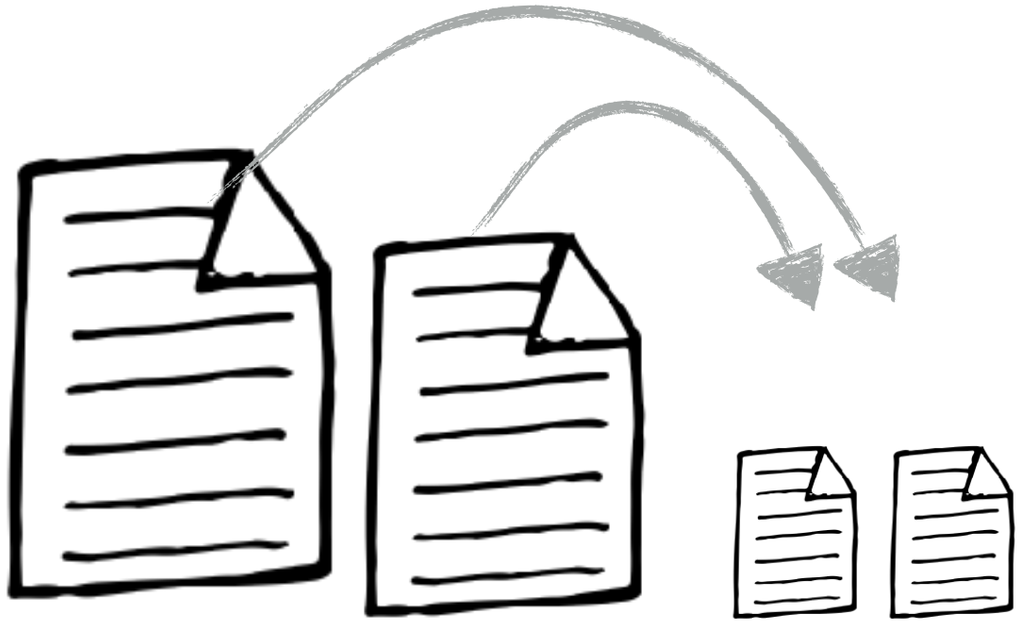
I'd like to see the showtimes for Silly Movie 2.0 at the movie house

Intent: SearchScreeningEvent

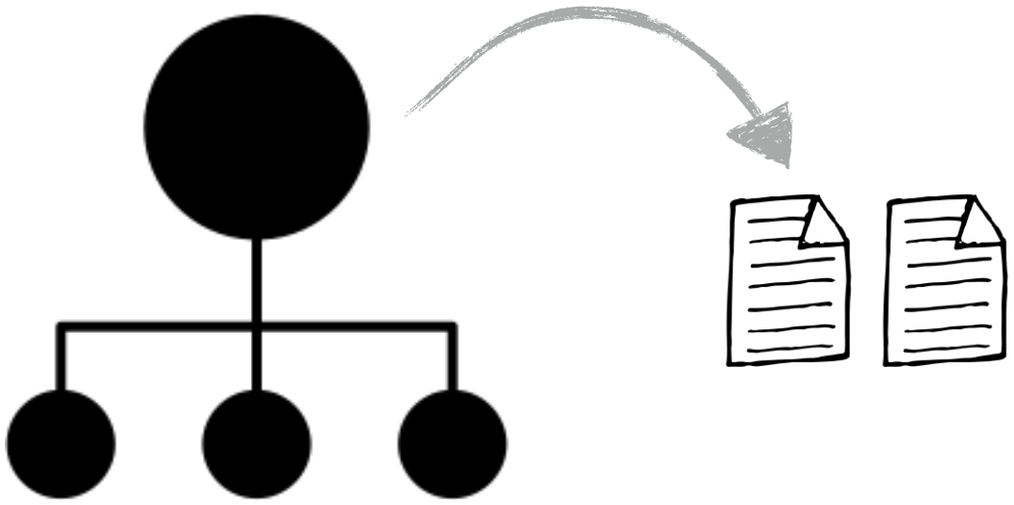
How can we transfer knowledge to low-resource languages?



Approaches

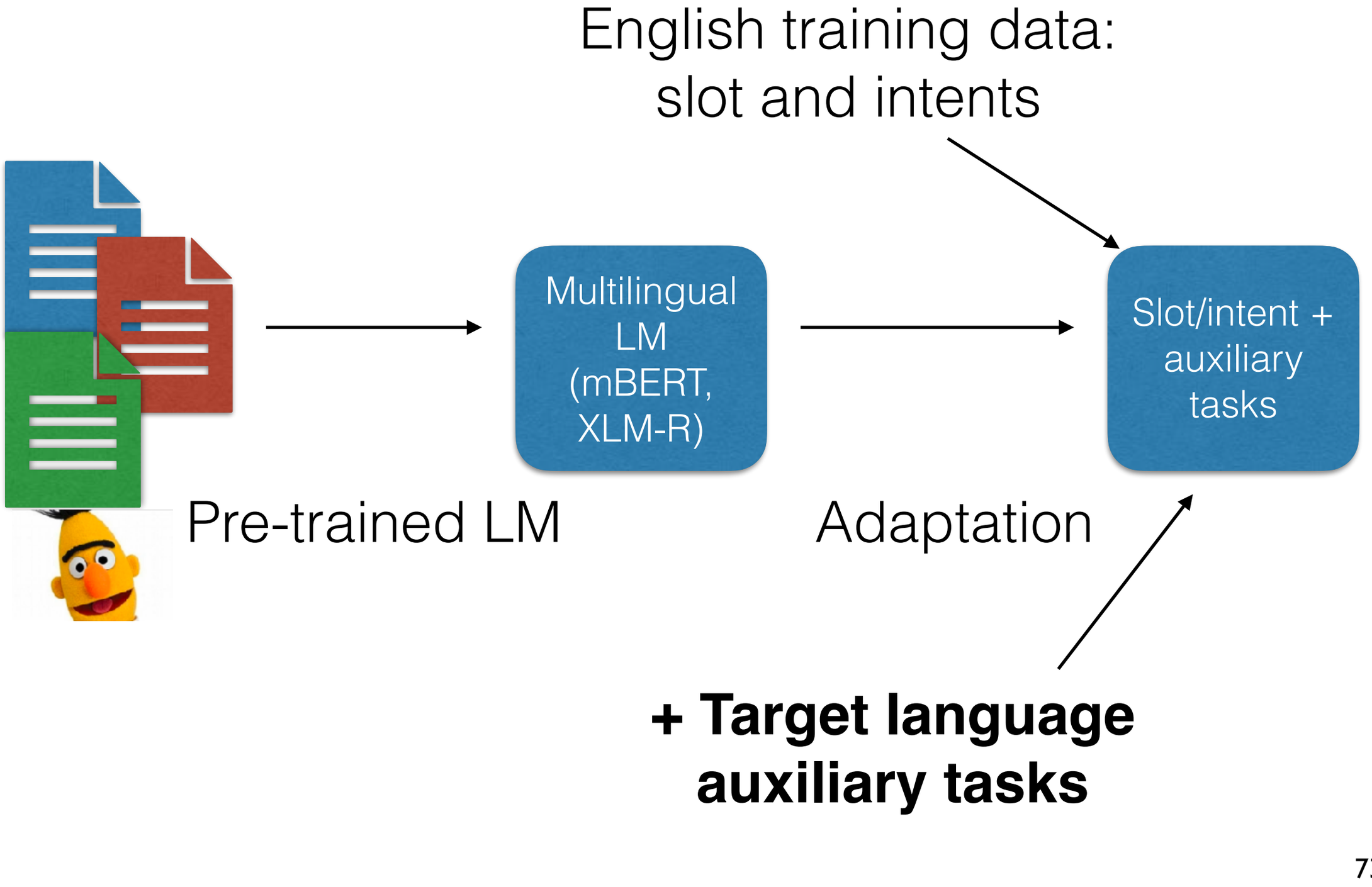


annotation transfer
(e.g. annotation projection, translation)



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

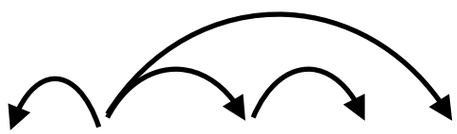
Idea: Non-English Auxiliary Tasks



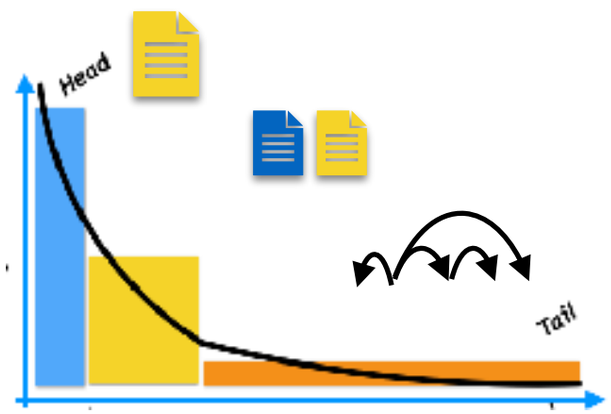
Non-English Auxiliary Tasks

- **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)

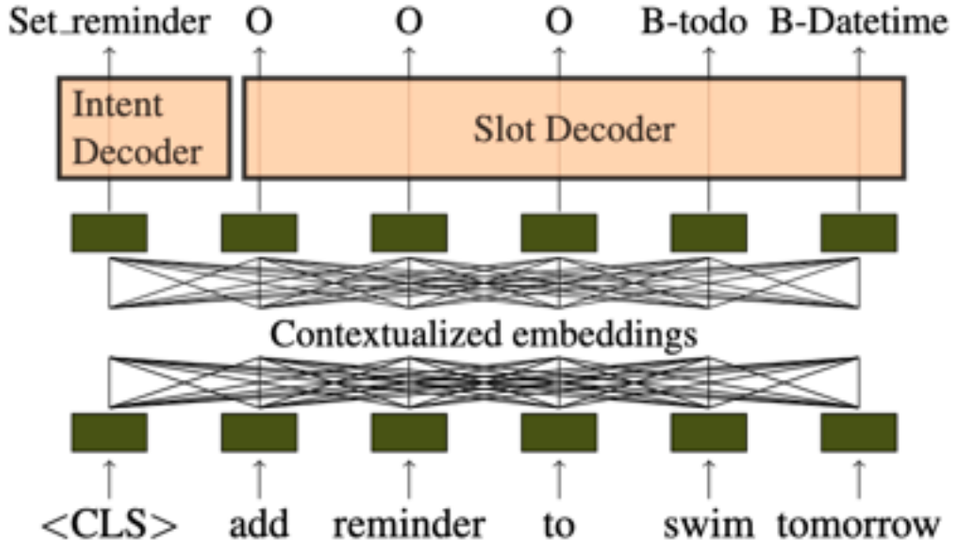


Figure 2: Overview of the baseline model.

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)

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

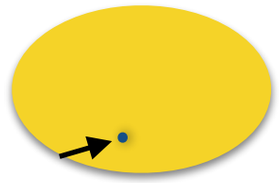
- German dialect (“Südtirolerisch”) spoken by a minority
  - Spoken in the northernmost Italian province of Bozen-Bolzano with ~0.5M inhabitants (~2/3 German dialect speakers)
- No common orthographic standard
- Lexical influence of other official languages (Italian, Ladin)
  - “Hosch is **patent** schun 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

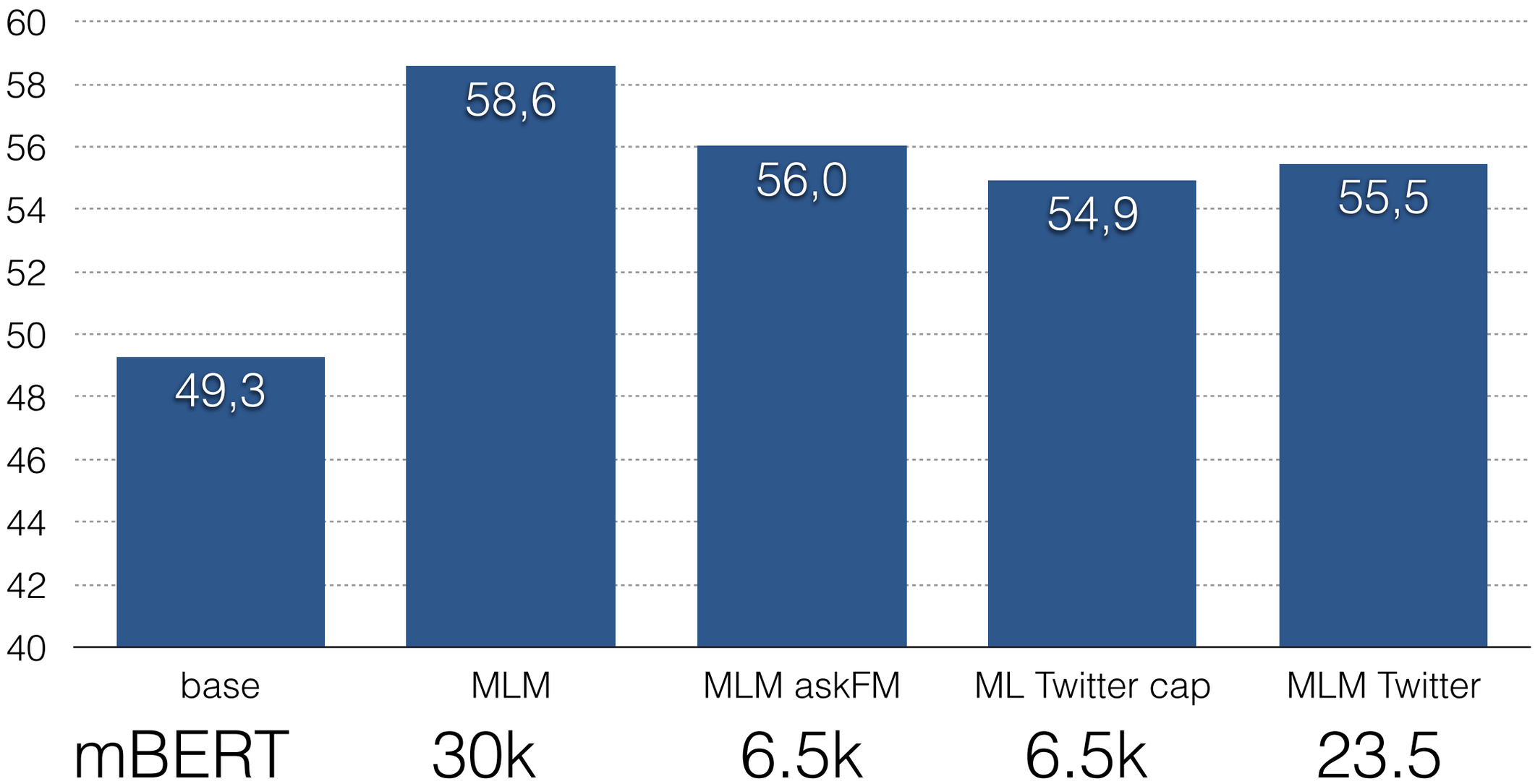


- Very difficult to get access to unlabeled data

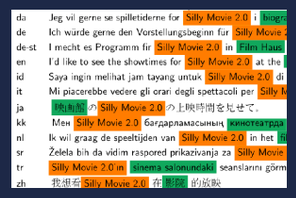
- Social media (Twitter): highly mixed data, switch to “high” languages, no “dialect” identifier exists

- AskFM: short Q&A posts, more dialectal

De-ST: #sentences for MLM



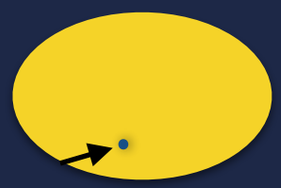
Take-aways



1. xSID is a new multilingual evaluation dataset for intent and slot detection

2. aux-MLM is the most robust auxiliary task

3. First results on 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

Roadmap

1. Cross-domain learning
2. Cross-lingual learning
3. Multi-task learning with Fortuitous Data
4. Continual Learning

Learning with disagreement

joint work with Dirk Hovy, Hector Martinez Alonso, Anders Søgaard, Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Massimo Poesio (EACL 2014, ACL 2014, NAACL 2021)

Disagreement in human annotation is ubiquitous

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**	ADP	NOUN	SYM
VERB	**PRON**	ADP	NOUN	SYM
VERB	**ADV**	ADP	NOUN	SYM

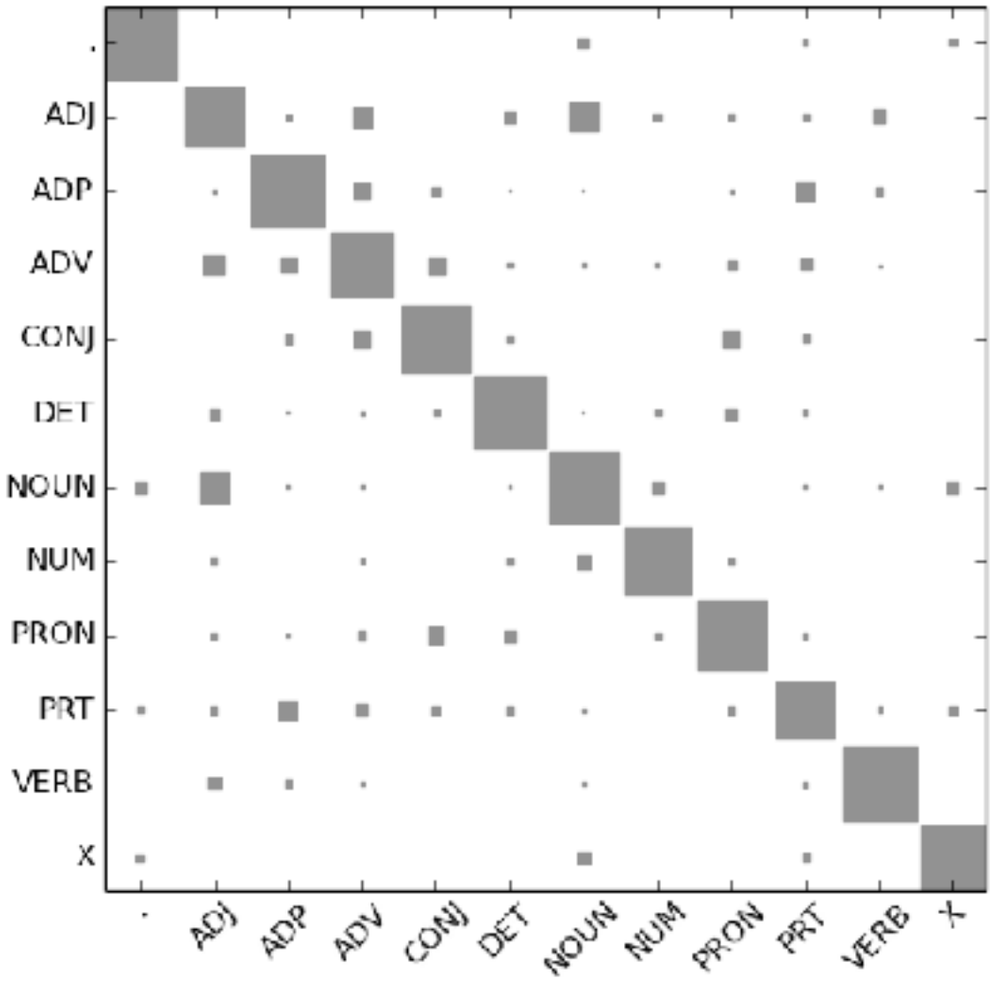
Say Anything with boyfriend :)

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

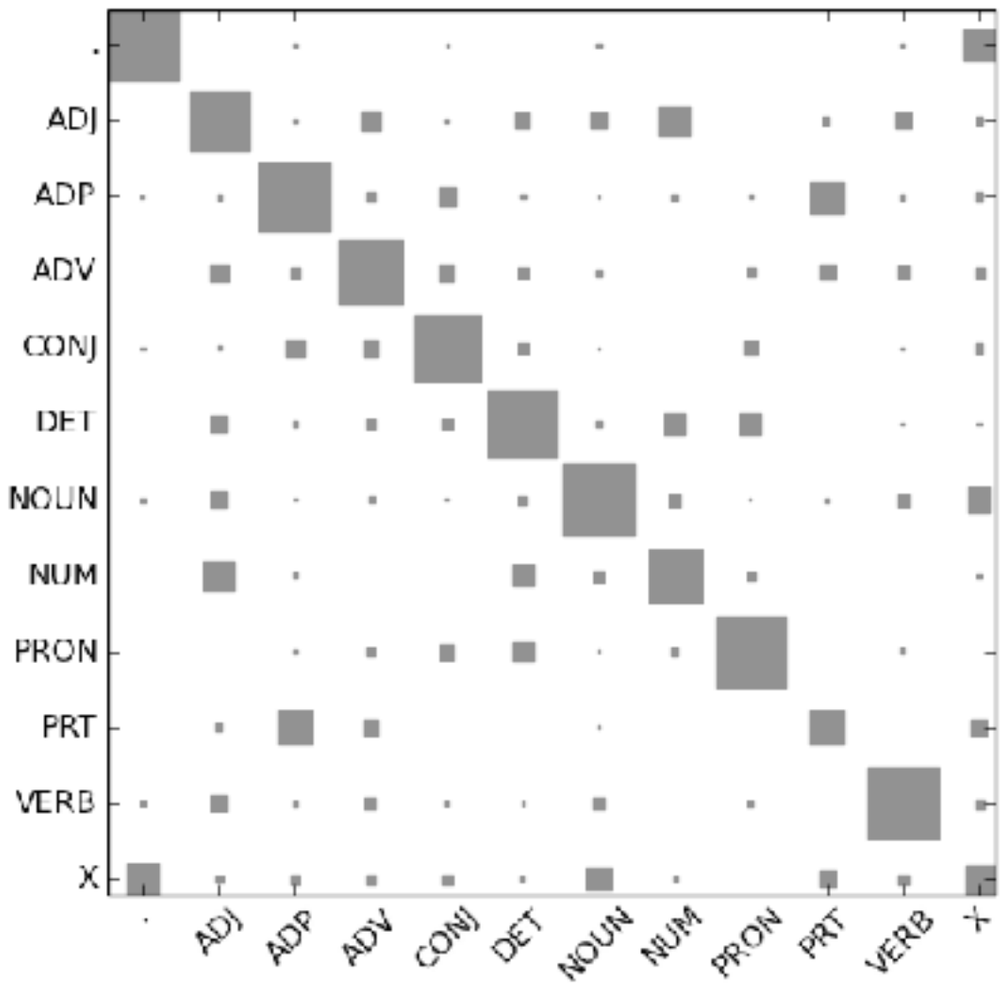
So what can we do?

Are disagreements randomly distributed?

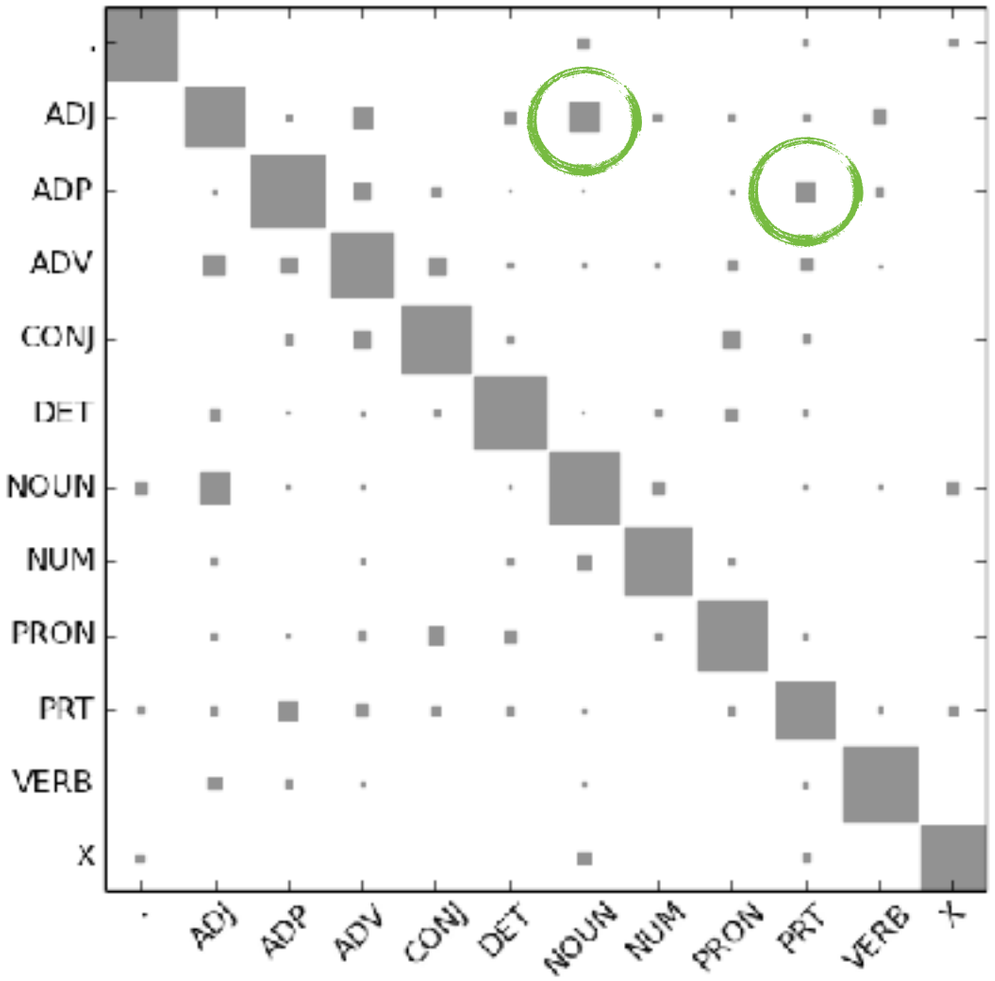
... and can we estimate disagreements from small samples?



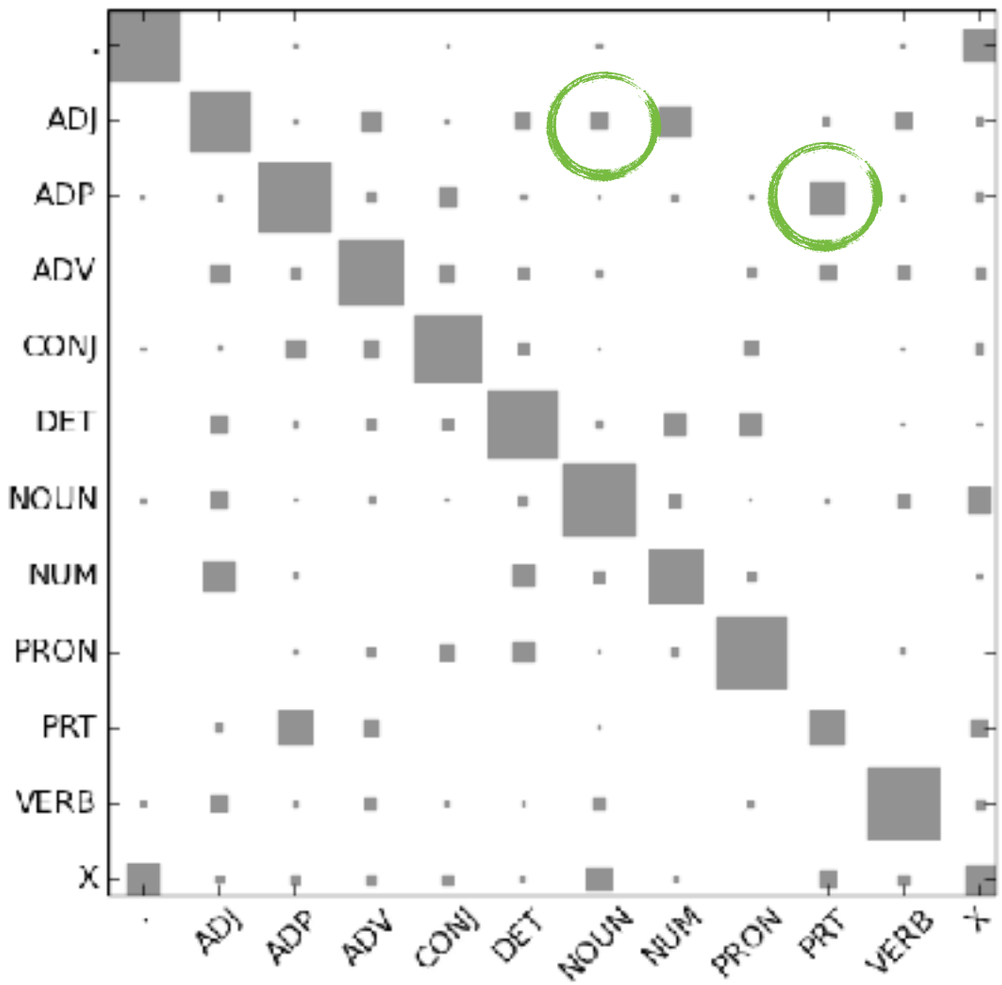
Wall Street Journal PTB-00



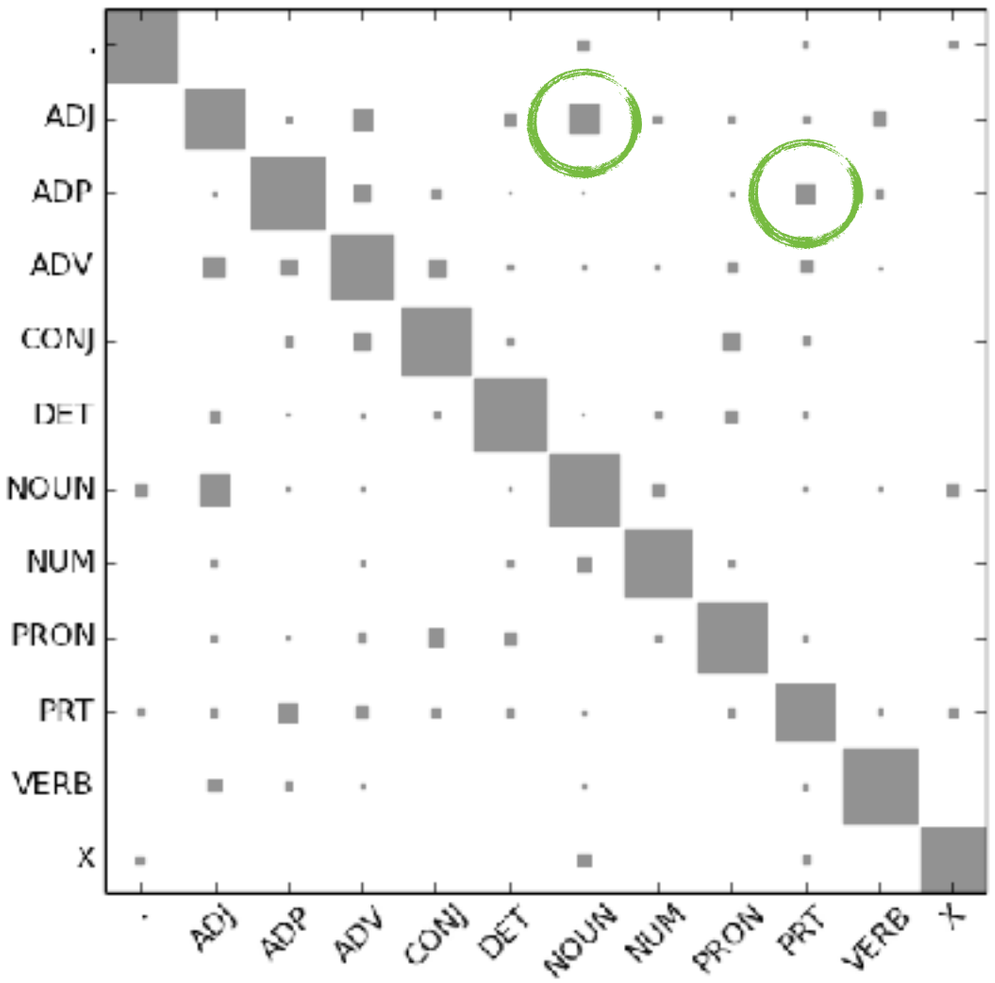
Twitter



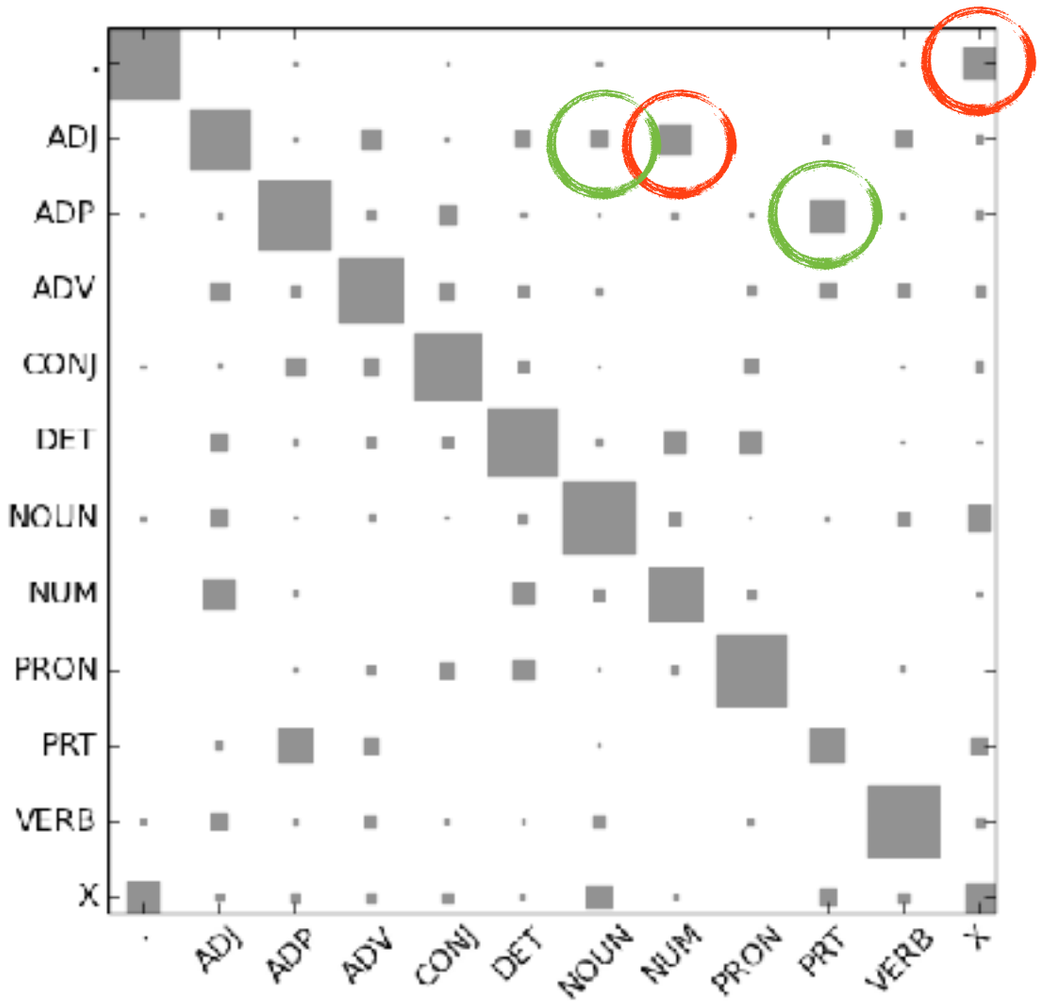
Wall Street Journal PTB-00



Twitter



Wall Street Journal PTB-00



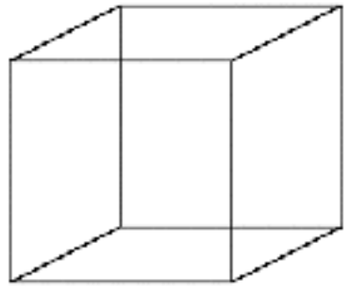
Twitter

How to exploit this information?

Filter data

algorithms

Disagreement is a type of



Fortuitous data

Fortuitous data

- Data out there,

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

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

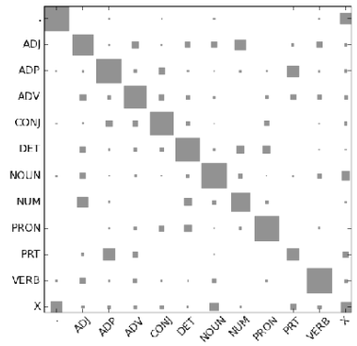
Typology of fortuitous data

Type / Side benefit of	Examples	Availability	Readiness
meta-data	hyperlinks, HTML markup, unlabeled data, symbolic knowledge	+	+
annotation	annotator disagreement	-	+
behavior	cognitive processing data	+	-

How to exploit disagreement during learning?

Survey upcoming (Uma et al., 2021)

CM (confusion matrix)



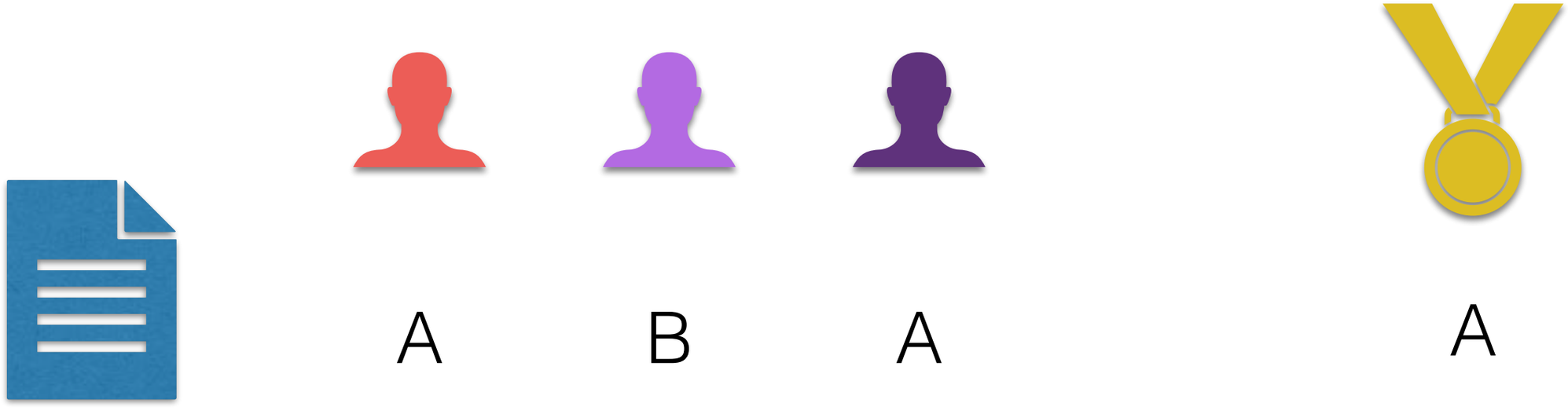
$$\gamma(y_j, y_i)) = 1 - P(\{A_1(X), A_2(X)\} = \{y_j, y_i\})$$

cost-sensitive learning

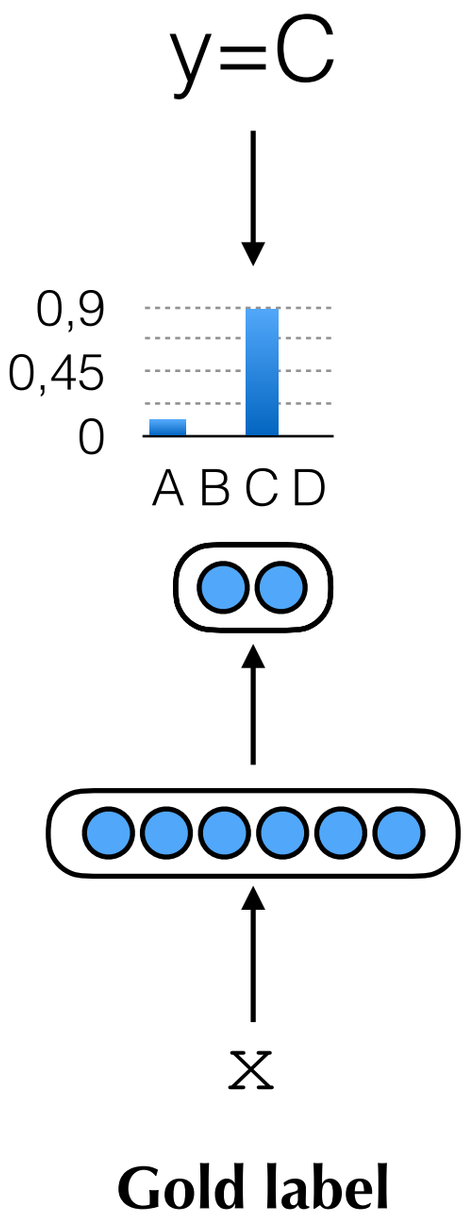
Plank et al., (2014)

Can we exploit MTL?

Multiple human annotations

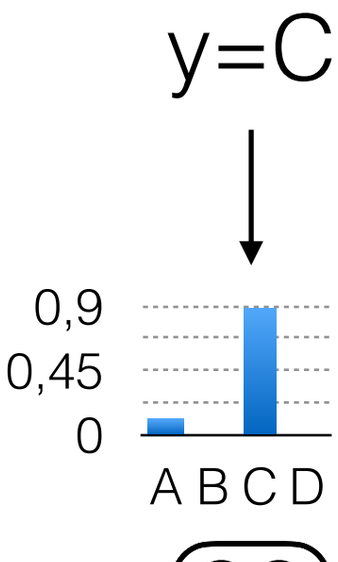


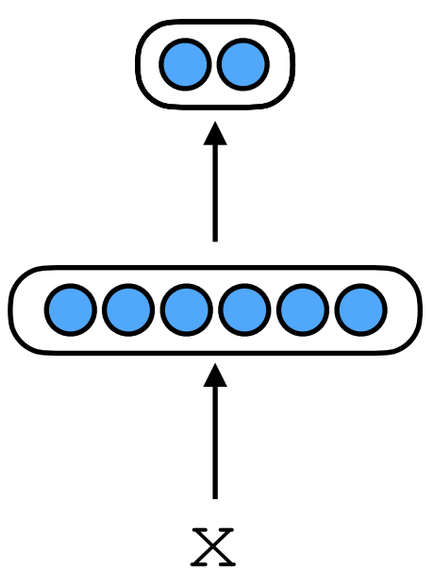
Typical setup



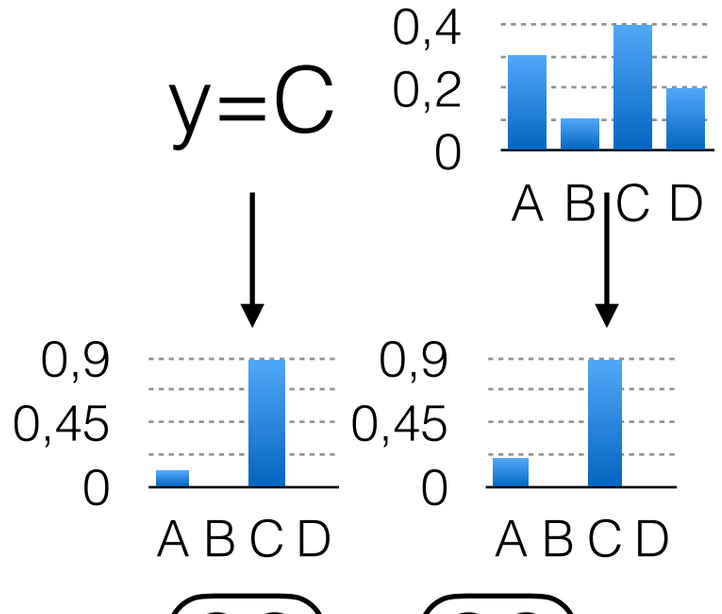
Gold label

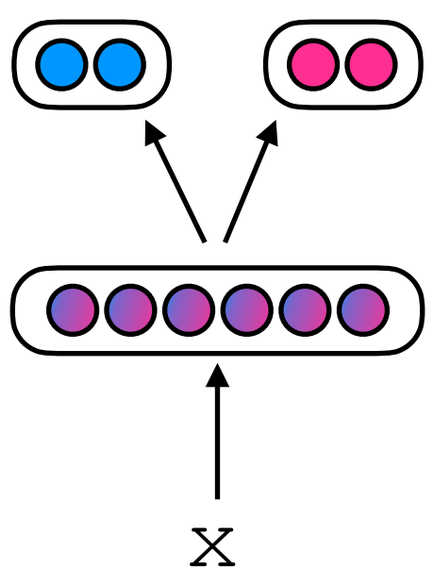
Soft-labels via Multi-Task Learning





Gold label



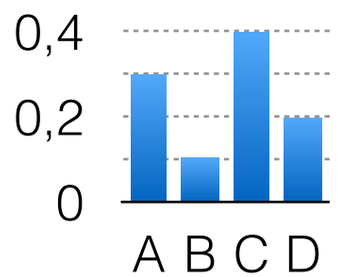


Gold label + Soft label

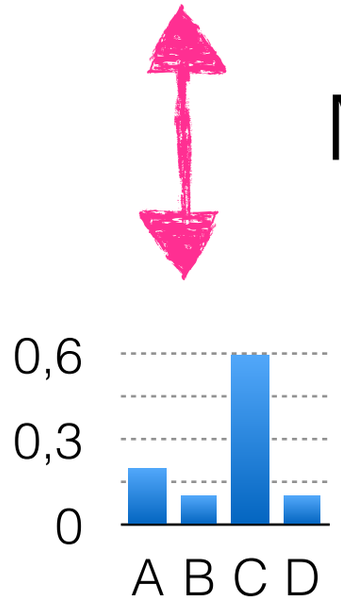
(Fornaciari, Uma, Paul, Plank, Hovy, Poesio 2021 NAACL)

Soft-labels

Annotator distribution P



Predicted softmax Q



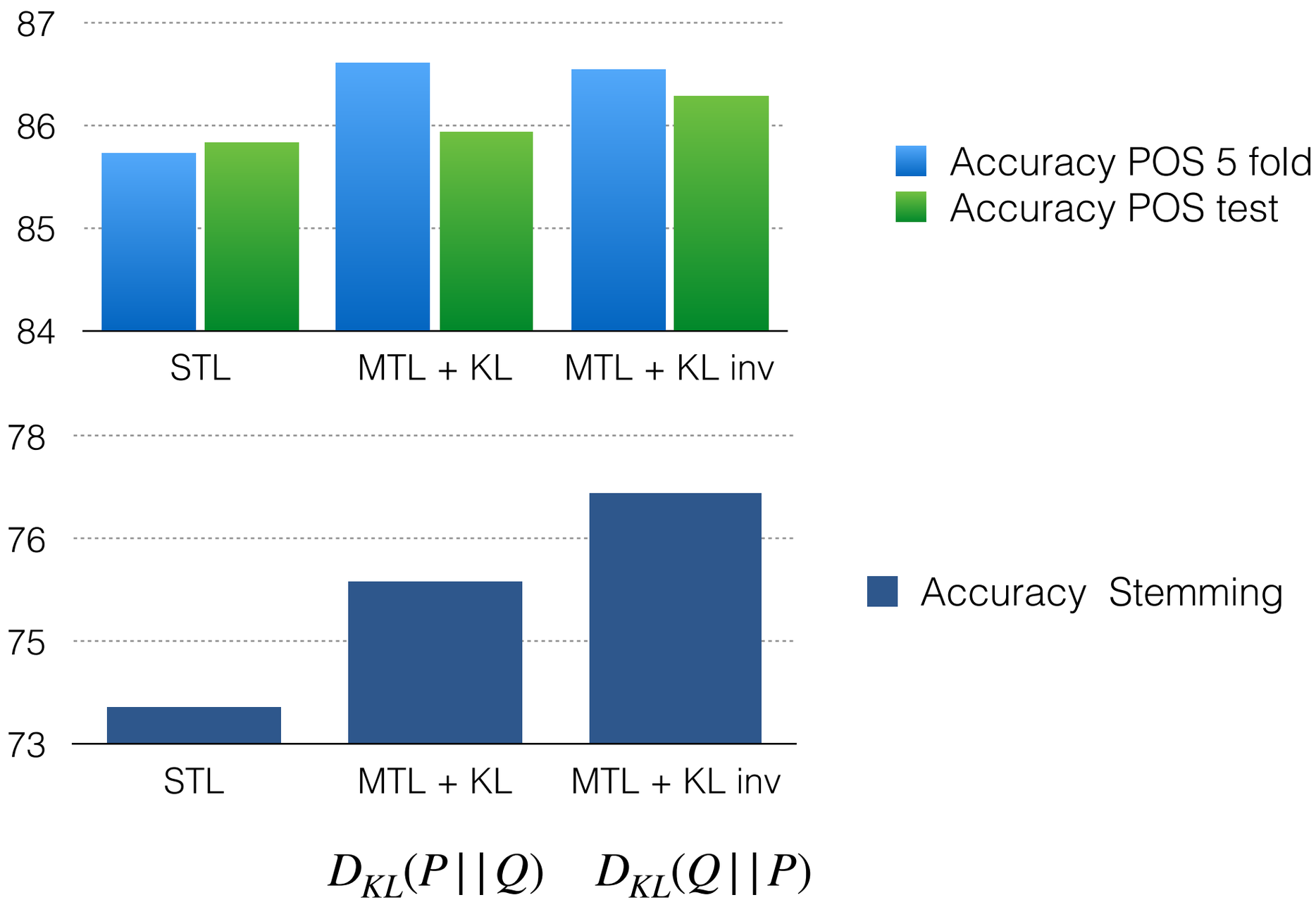
Measure divergence

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

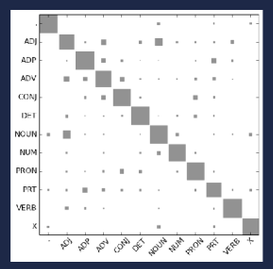
Experiments

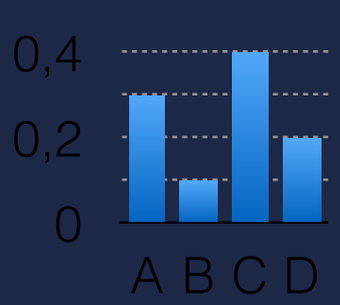
- **Comparison:**
  - Single task learning
  - Multi-task learning (with gold or majority vote)
    - With soft loss
  - Two NLP tasks in this paper: POS and stemming

Results


87
86
85
84
STL
MTL + KL
MTL + KL inv
Accuracy POS 5 fold
Accuracy POS test
78
76
75
73
STL
MTL + KL
MTL + KL inv
Accuracy Stemming
$D_{KL}(P||Q)$
$D_{KL}(Q||P)$

Take-home-message





✓not all disagreement is noise

✓embrace it during learning

➡Consider releasing raw annotations

✦ More work needed to find suitable forms of evaluation and understand different forms of disagreement (Basile et al., 2021)

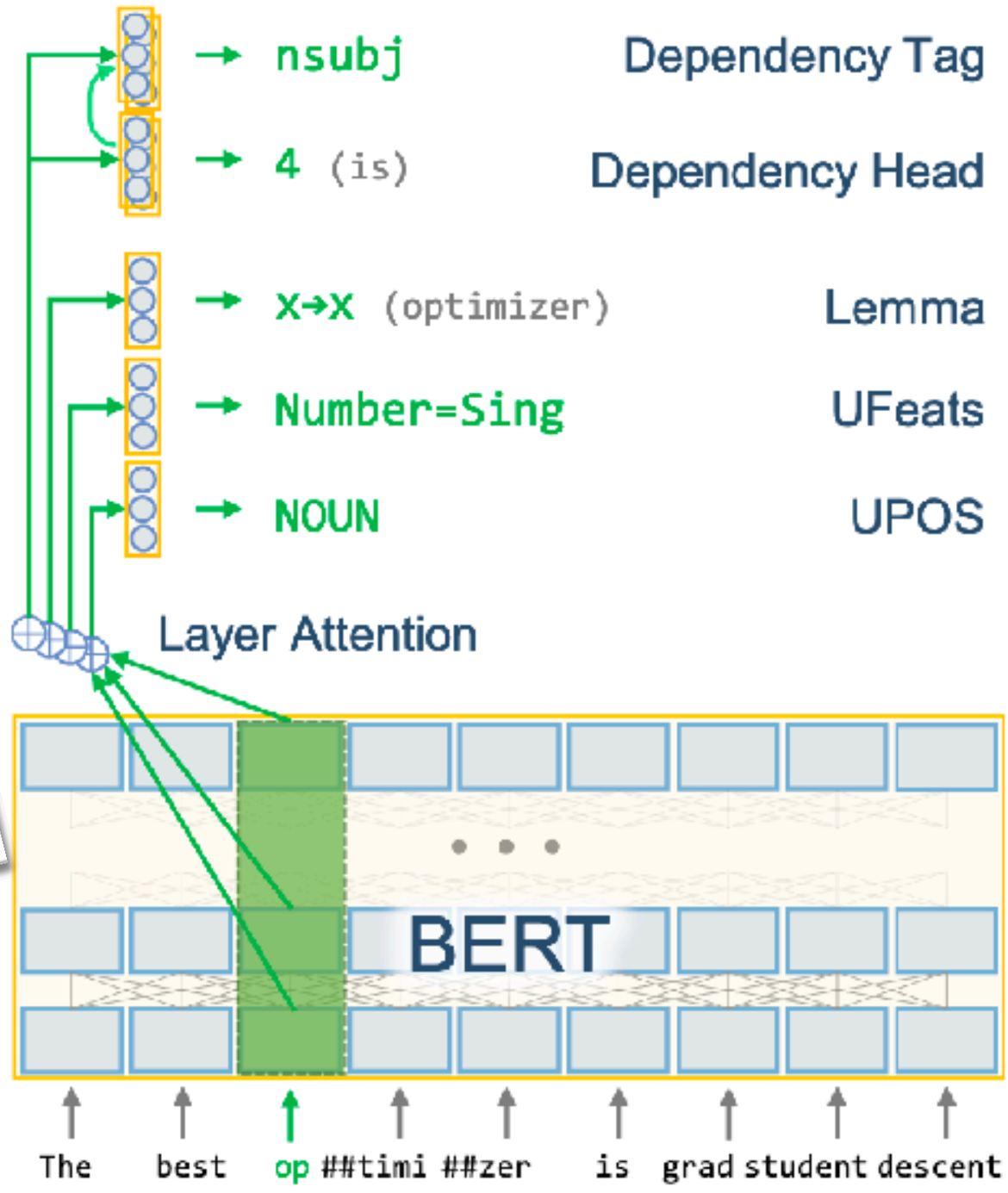
Outro:
MTL Practicalities
When does it work?
Two recent advances

Problem: 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**: sampling of data, weighting of loss

Example: 75 languages, one parser: UDify

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



Smoothing? Parsing UD v2.6 with MaChAmP (van der Goot et al., 2021 EACL)

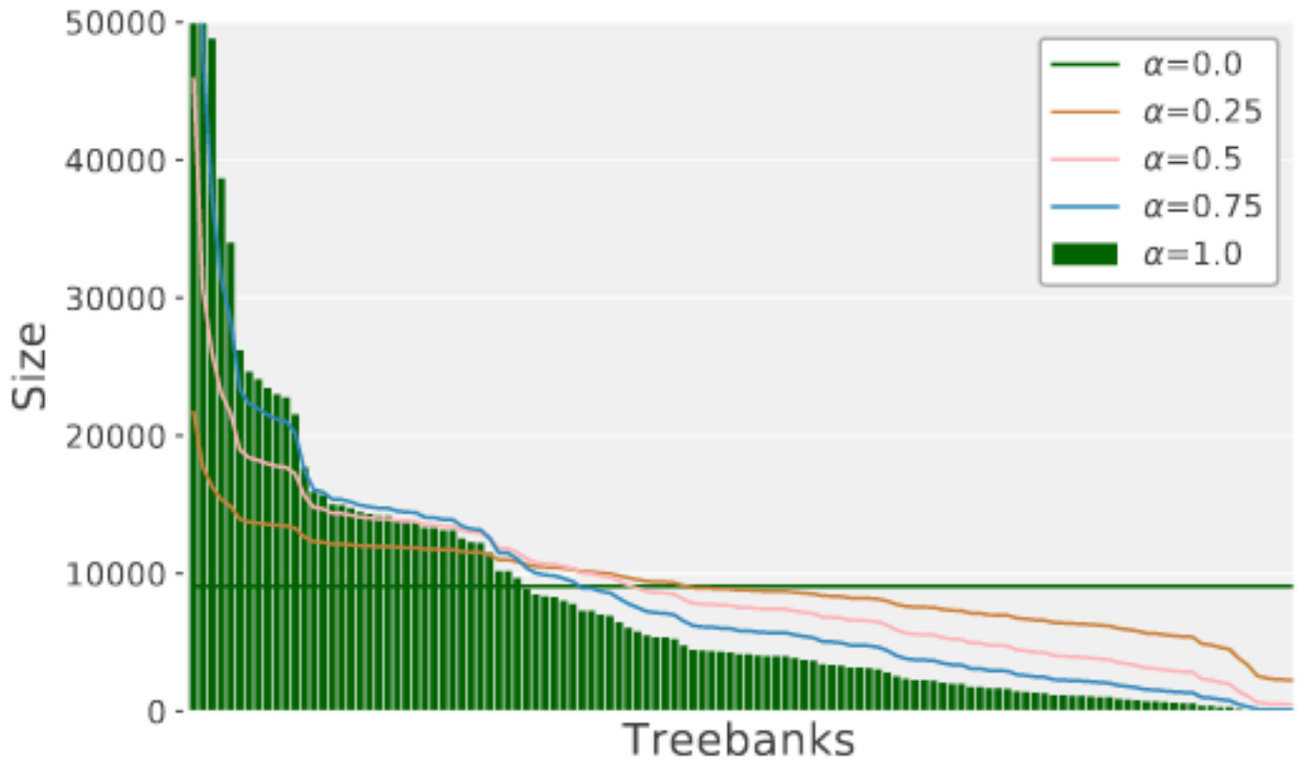
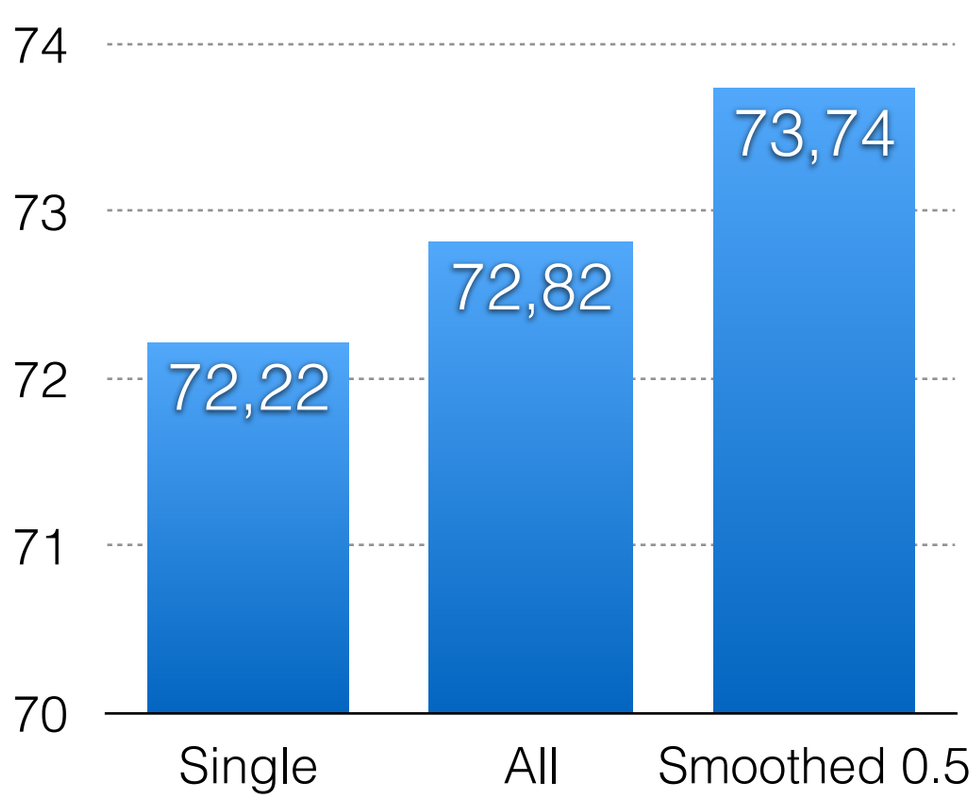
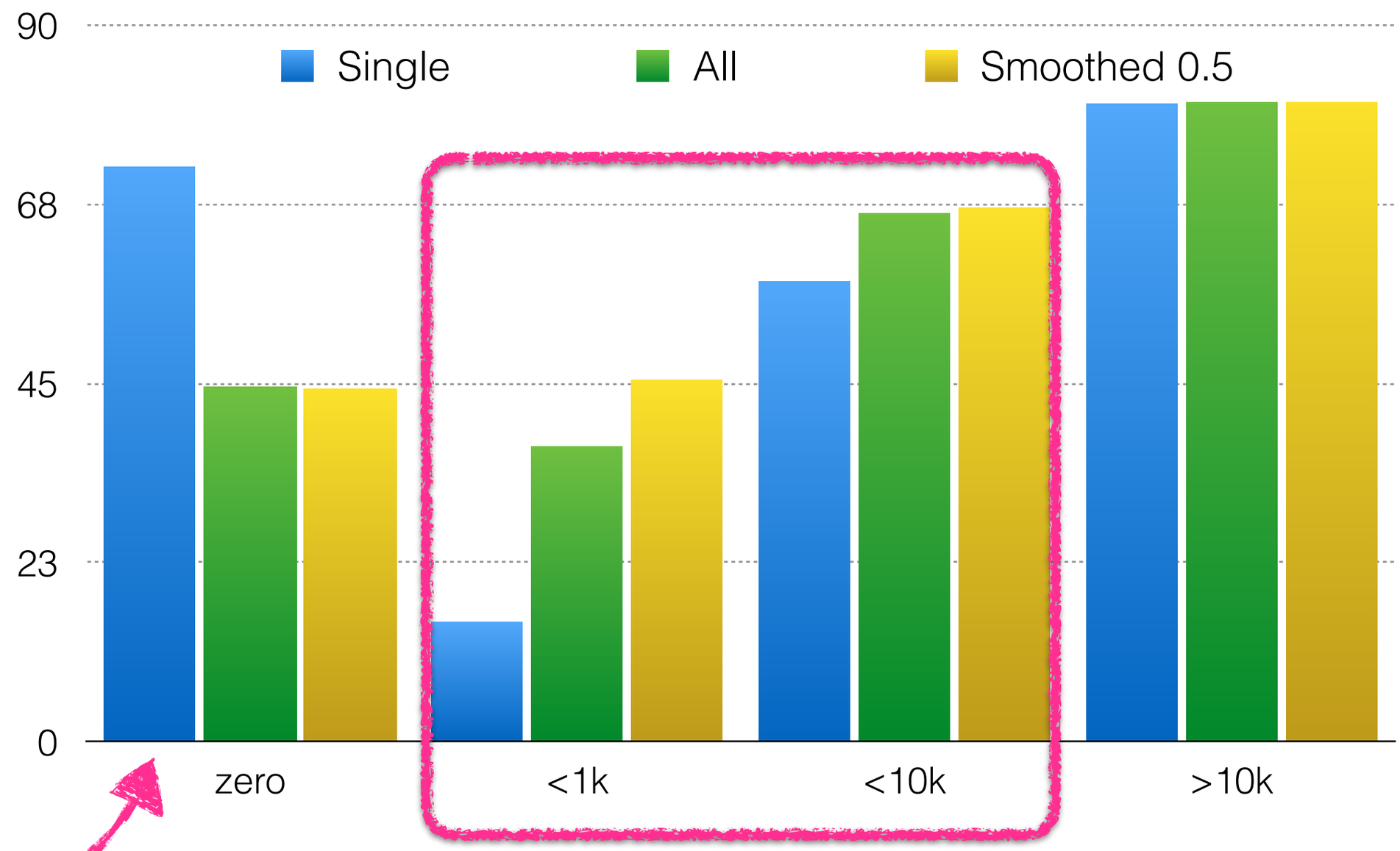


Figure 4: Effect of the sampling parameter α on the training sets of Universal Dependencies 2.6 data.

163 treebanks, 92 languages, released May 15, 2020.



Result by treebank size: Smoothing helps for low-resource languages



Single source selection

Interference of MTL in Parsing

(Grünewald et al., 2021)

Applying Occam's Razor to Transformer-Based Dependency Parsing: What Works, What Doesn't, and What is Really Necessary

Stefan Grünewald[★◆] **Annemarie Friedrich**[◆] **Jonas Kuhn**[★]
[★]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart
[◆]Bosch Center for Artificial Intelligence, Renningen, Germany

- UPOS + UFeats + Parsing as MTL
  - Tagging tasks accuracy increased rapidly and might overwhelm the overall loss, causing the parser to underfit
  - Scaling down tagging losses helps some languages for parsing, but drops tagging performance

Finnish TDT (fi_tdt)	UDPipe+	97.57	95.80	91.66	89.49
	$\text{STEPS}_{\text{dep-only}}$	–	–	**95.69**	**94.36**
	$\text{STEPS}_{\text{MTL}}$	**98.52**	**96.75**	94.62	93.11
	$\text{STEPS}_{\text{MTLscale}}$	98.19	88.70	95.59	94.26
Hindi (hi_hdtb)	UDPipe+	97.58	94.24	95.56	92.50
	$\text{STEPS}_{\text{dep-only}}$	–	–	96.11	93.34
	$\text{STEPS}_{\text{MTL}}$	**98.09**	**94.49**	95.96	93.03
	$\text{STEPS}_{\text{MTLscale}}$	97.51	88.60	**96.18**	**93.39**

So, not all that glitters is gold - When does MTL work?

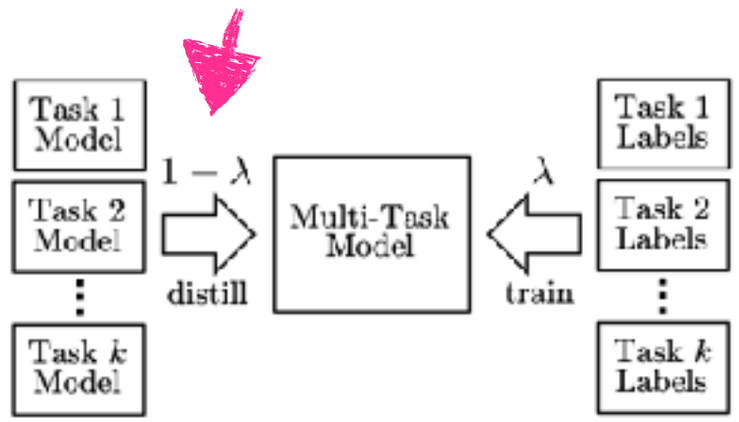
dependent conditions. Our results show that MTL is not always effective, significant improvements are obtained only for 1 out of 5 tasks. When successful, auxil-

- Data properties can give some clue for MTL (Label entropy, Kurtosis) (Alonso & Plank, 2017)
- Learning curve (Bingel & Sogaard, 2017)
- Mutual information (Bjerva, 2017; Schroeder & Biemann, 2020)

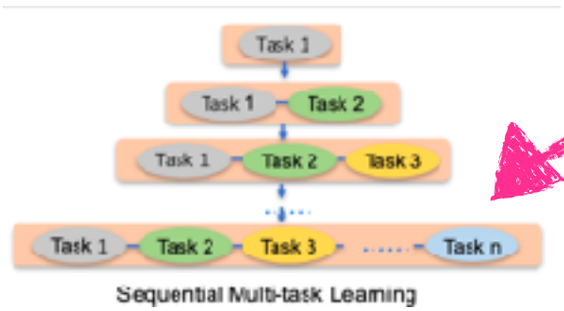
Finally, selected advances in MTL

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

Distillation with teacher annealing



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

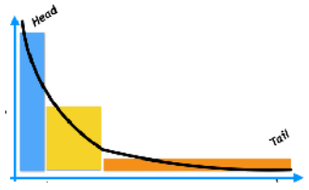


Progressively adding tasks

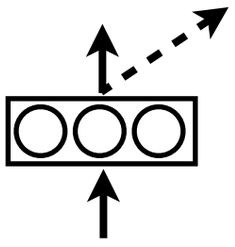
- MTL & *adapters* via shared hypernetworks (Mahabadi et al., 2021 arXiv)

To wrap up…

Key Take-aways

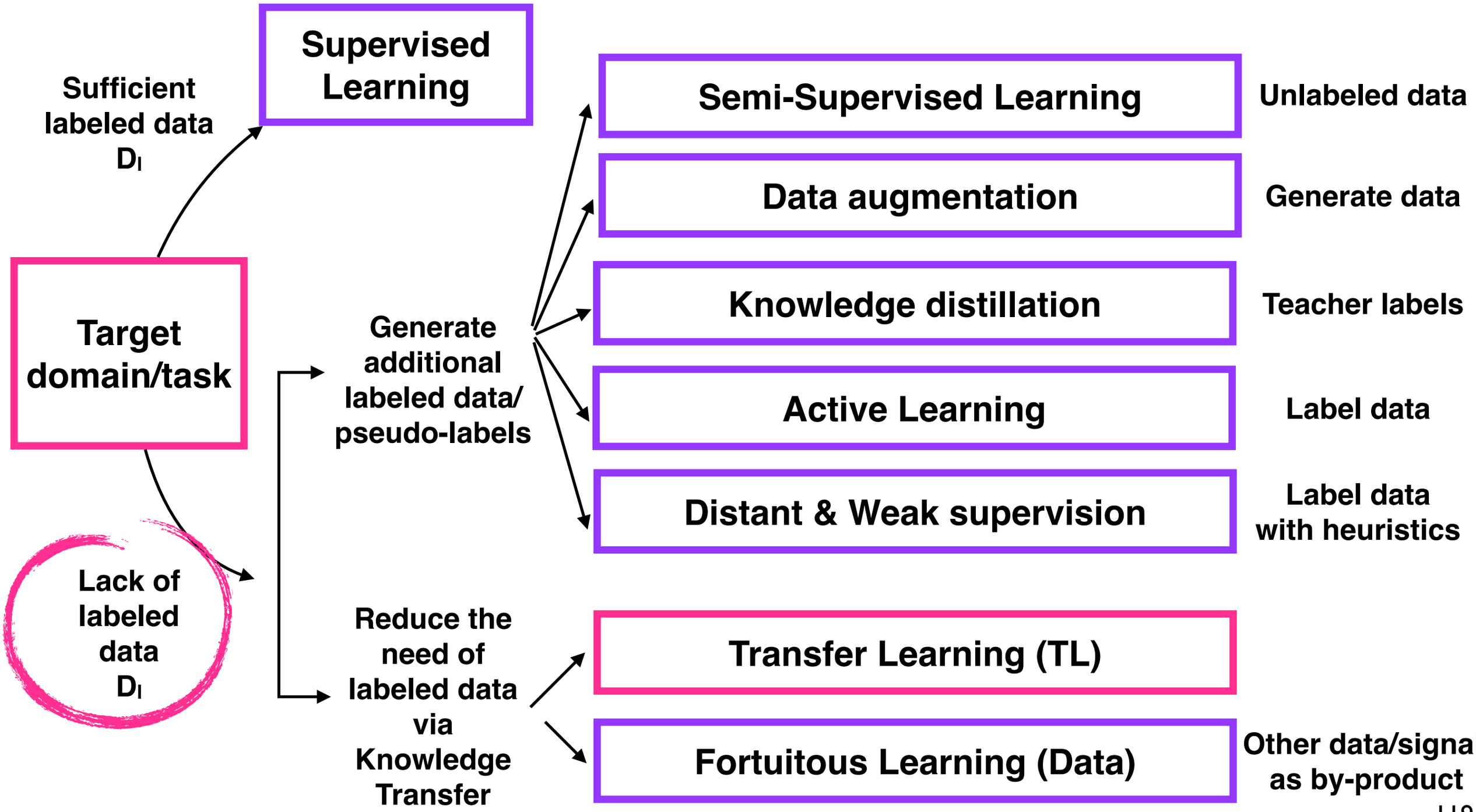


- **Scarce** and **biased** data are ubiquitous
- **Transfer Learning** is broad, STL is one kind

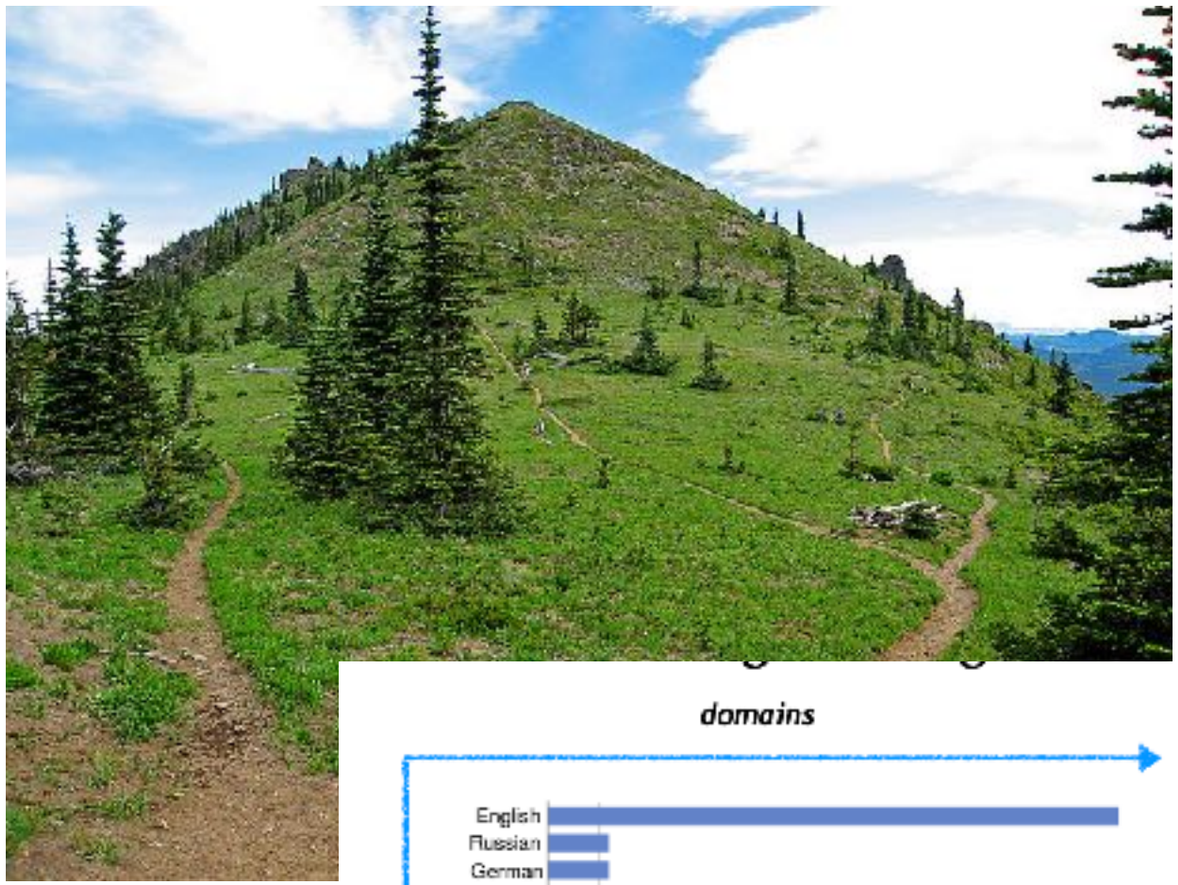


- **Data selection** for transfer learning
- **Multi-Task Learning -** helps in low-resource setups (e.g. aux-MLM); allows use of distinct data sources; interference can be an issue

Relationship to other learning paradigms



It's an exciting time to study how we can best address scarce & biased data conditions



Questions? Thanks!

IT UNIVERSITY OF COPENHAGEN

bplank.github.io
nlpnorth.github.io

Transfer Learning & MTL in NLP

@barbara_plank
bapl@itu.dk

Thanks to the support by:

DANMARKS FRIE FORSKNINGSFOND

amazon

NVIDIA

Papers presented

1. **Learning to select data for transfer learning with Bayesian Optimization**

   (Ruder & Plank, EMNLP 2017): Data selection for transfer learning

2. **From Masked Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-shot Spoken Language Understanding**

   (van der Goot et al., NAACL 2021): xSID and aux-MLM

3. **Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning**

   (Fornaciari et al., NAACL 2021): soft-loss MTL