north

IT UNIVERSITY OF COPENHAGEN

Natural Language Processing: Recent Advances and Challenges

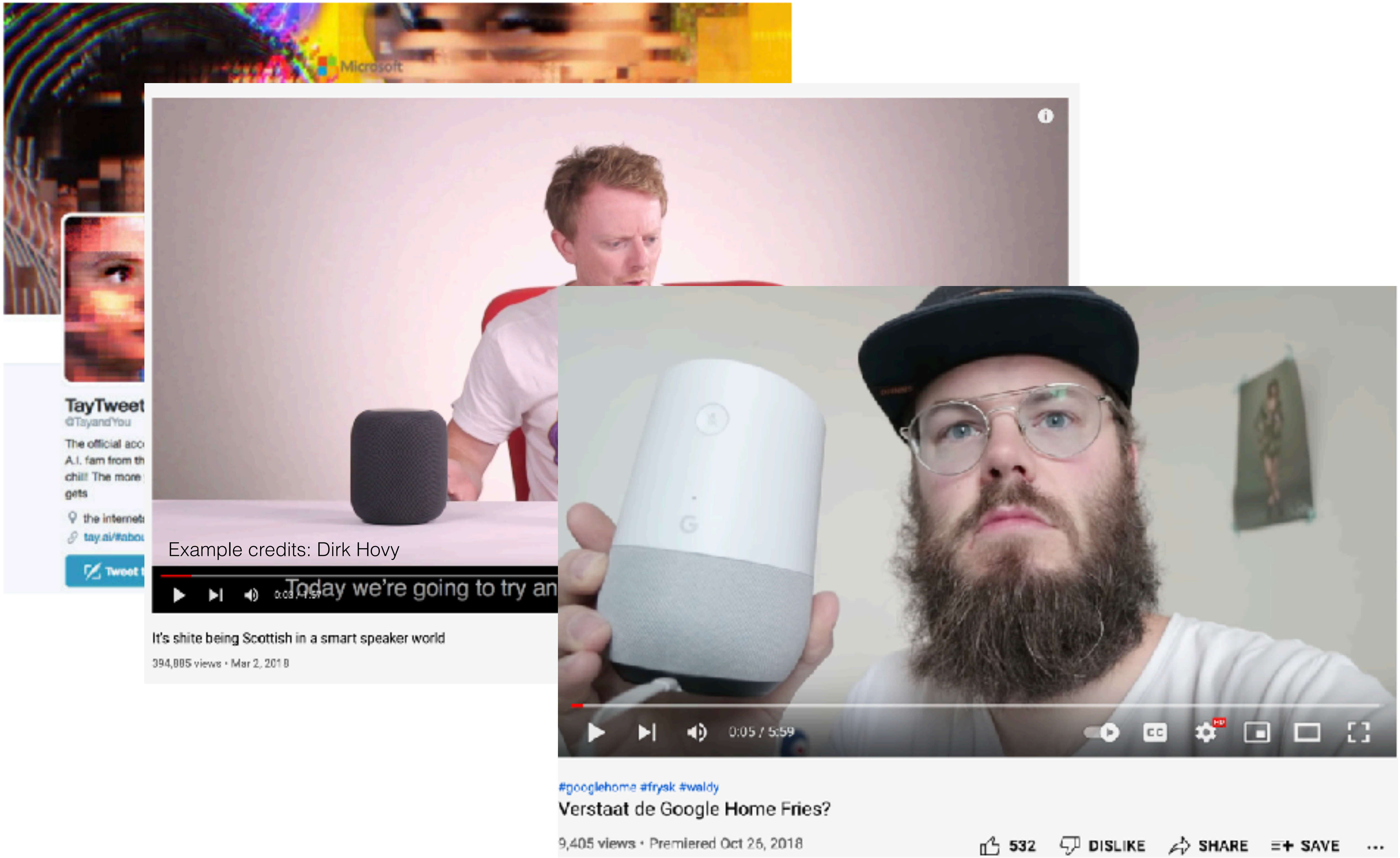
Prof. Barbara Plank
ITU, Copenhagen, Denmark

Sprogteknologisk Konference 2021

Natural Language Processing

Big goal: AI to understand and produce language, just as we humans do

Biased NLP systems & Social Implications



Example credits: Dirk Hovy

It's shite being Scottish in a smart speaker world

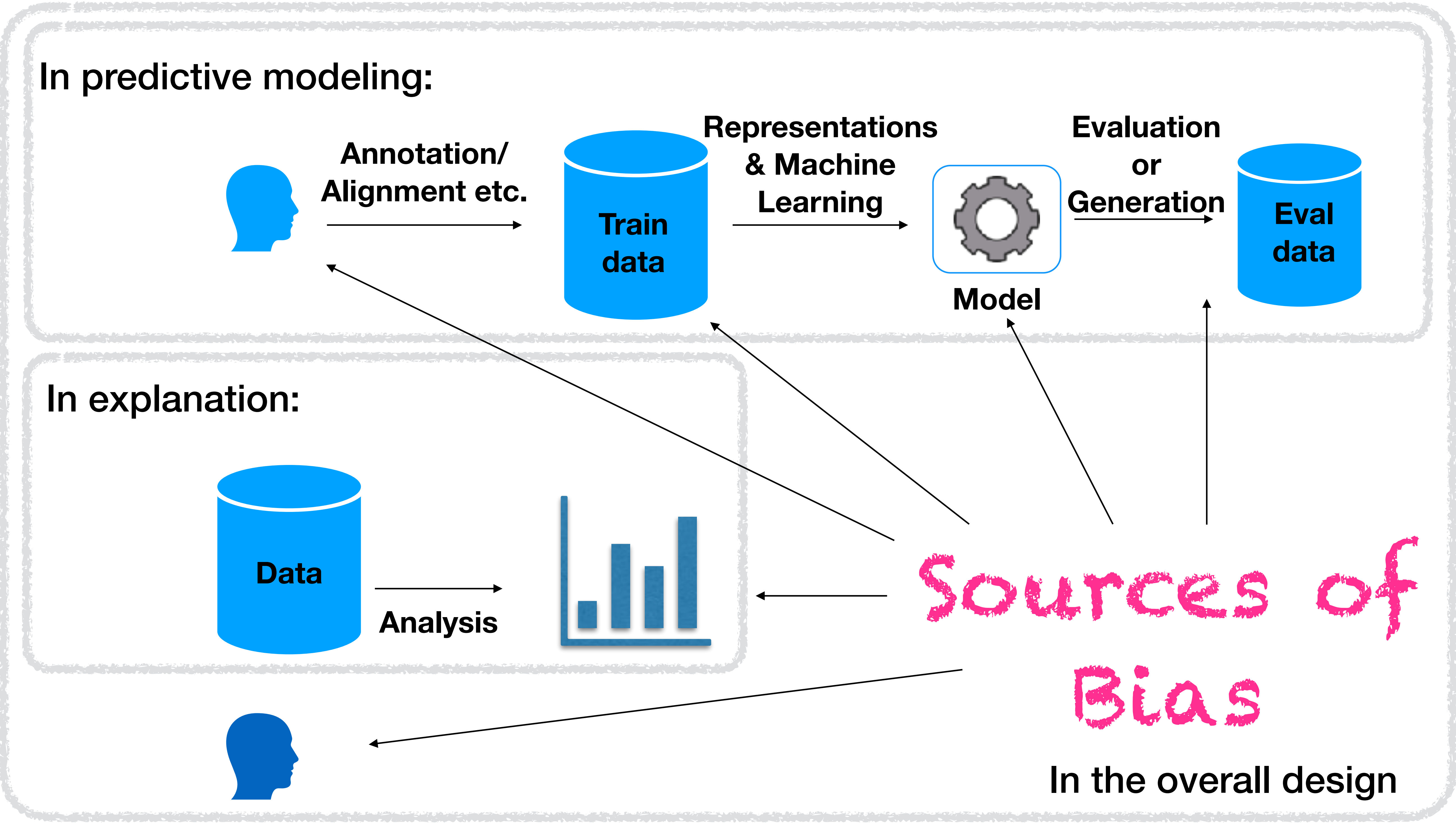
394,885 views · Mar 2, 2018

#googlehome #frysk #waldy

Verstaat de Google Home Fries?

9,405 views · Premiered Oct 26, 2018

BIAS everywhere - Not only in the data



On Bias in NLP: Hovy & Spruit (2016), Hovy & Prabhumoye (2021) & importance of documentation: Data statements (Bender & Friedman, 2018), Data sheets (Gebru et al., 2020)

With big data (*big language models*) comes big responsibility

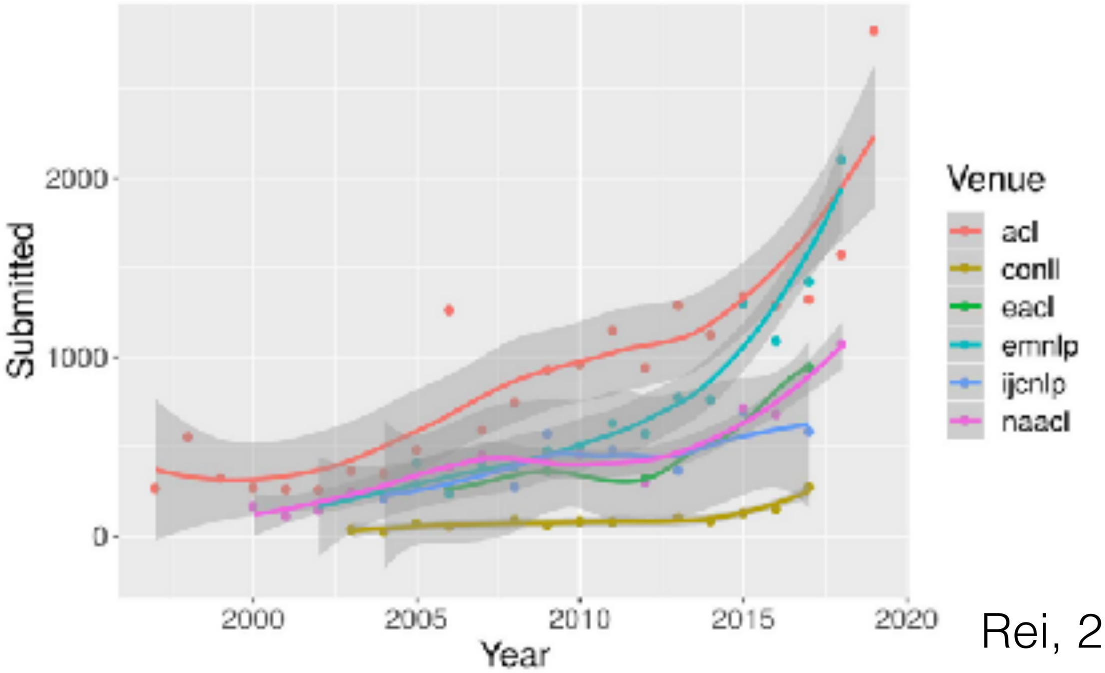
A lot remains to be done, to create inclusive and robust NLP

Where are we now?

A short historical perspective

NLP has grown tremendously in the last decade

- Papers at top NLP conferences:

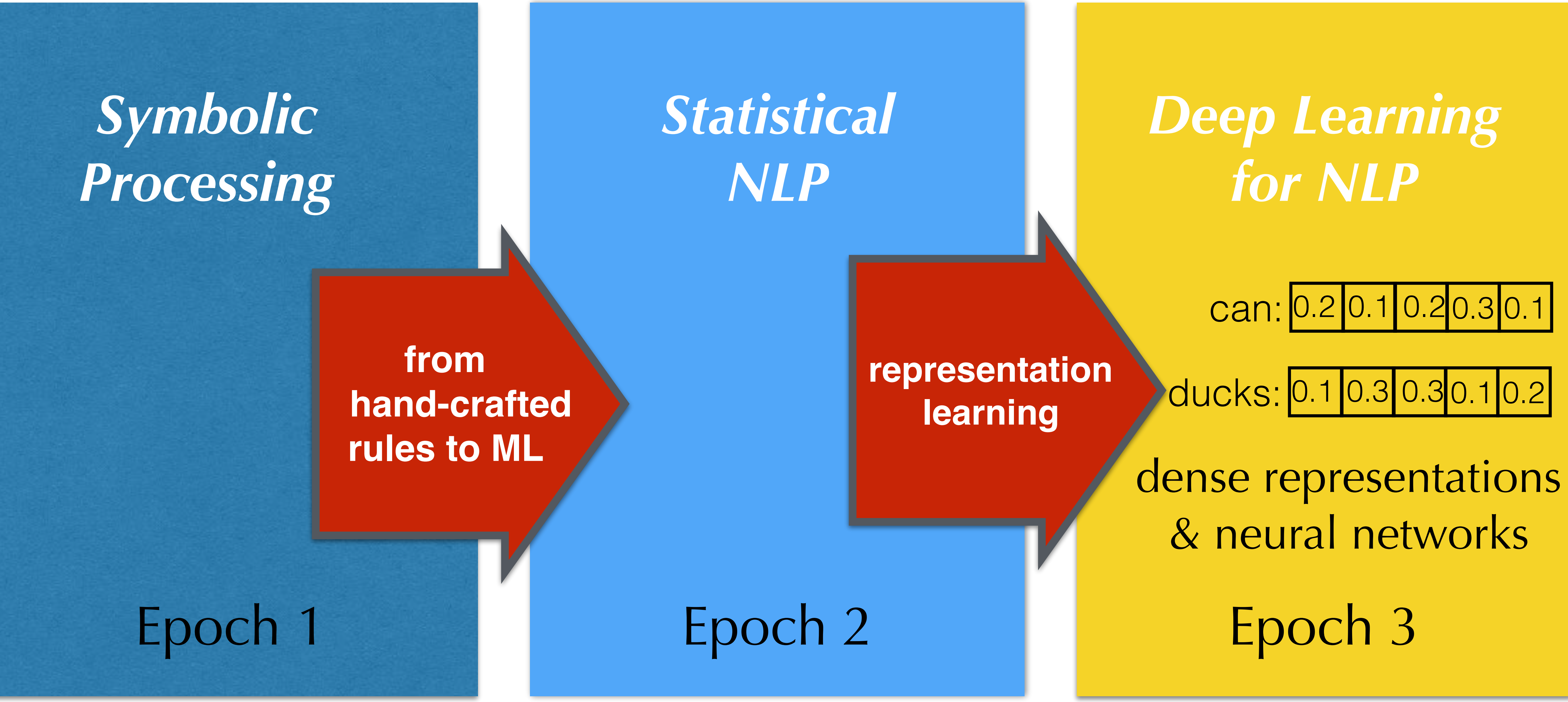


Rei, 2019

- Market size expected to triple from 11.6 Billion USD in 2020 to 35.1 Billion USD by 2026
  - Lower gap to impact on society

http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/

NLP ❤ Deep Learning



Symbolic Processing

from hand-crafted rules to ML

Statistical NLP

representation learning

Deep Learning for NLP

can:	0.2	0.1	0.2	0.3	0.1
ducks:	0.1	0.3	0.3	0.1	0.2

dense representations & neural networks

Epoch 1

Epoch 2

Epoch 3

approx. 1980s

2015

Last 4 years: Large Pre-trained LMs

Deep Learning for NLP

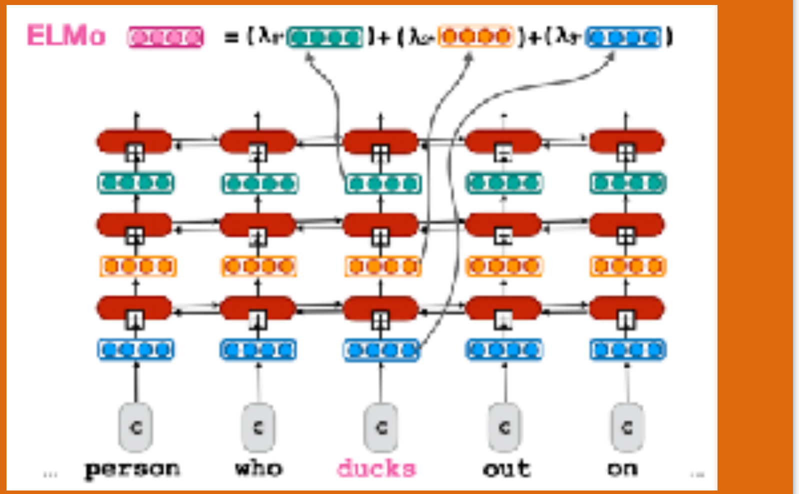
can:	0.2	0.1	0.2	0.3	0.1
ducks:	0.1	0.3	0.3	0.1	0.2

dense representations & neural networks

Epoch 3

Contextualised Embeddings

Pre-trained Language Models



e(ducks) != e(ducks)

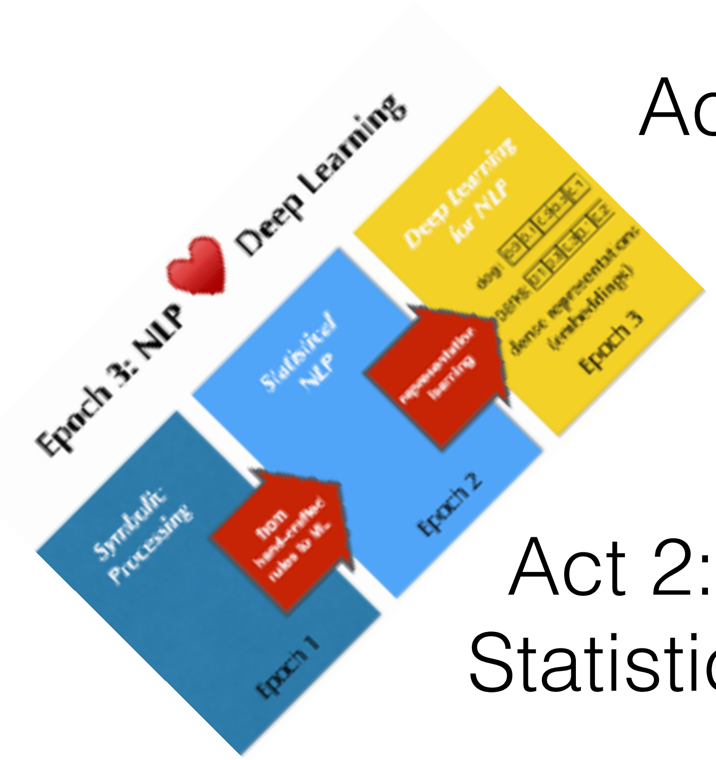
Epoch 4

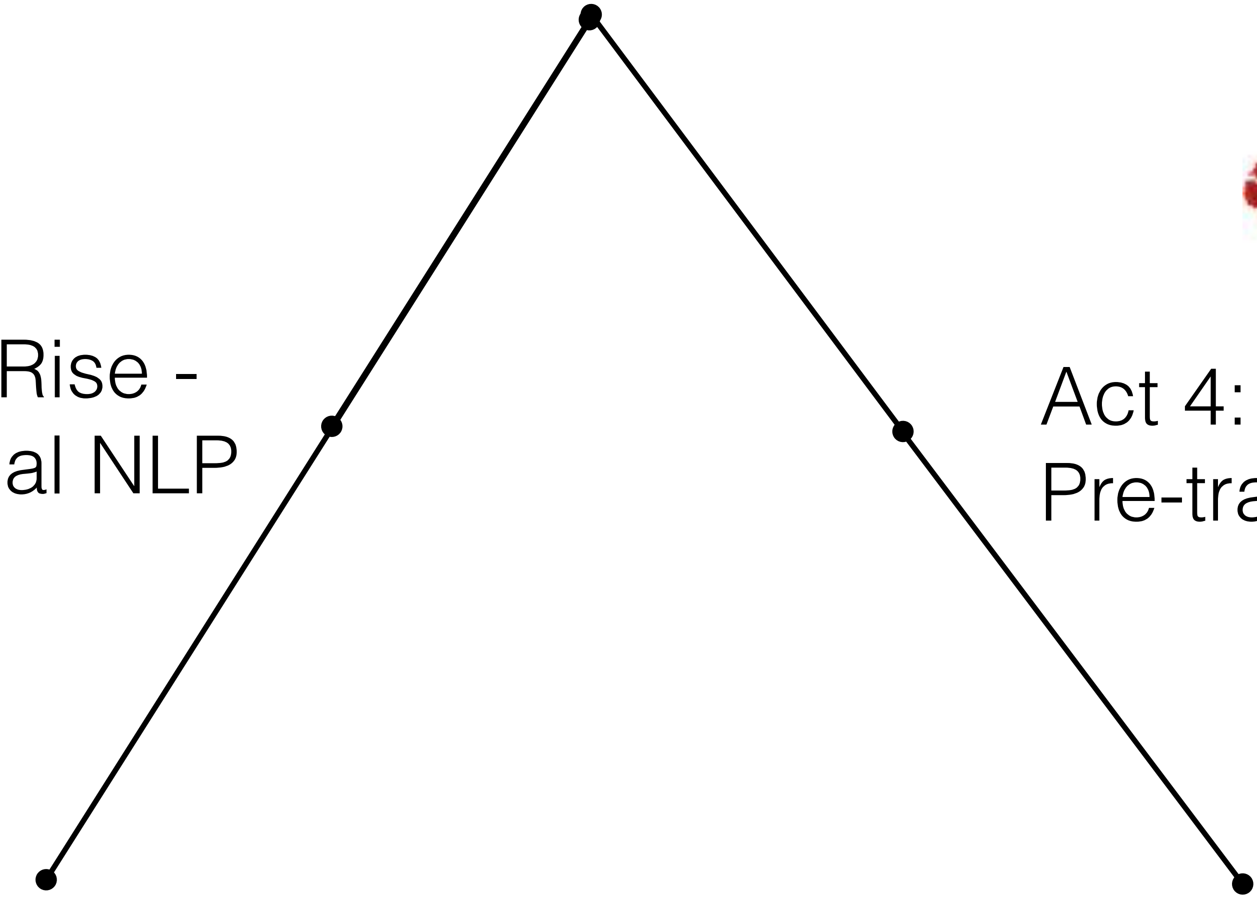
2015

2018

ELMo (Peters et al., 2018)
BERT (Devlin et al., 2019)

NLP History summarized as dramatic act





Act 3: Climax - Deep Learning for NLP

Act 2: Rise -
Statistical NLP

Act 4: return or fall:
Pre-training (2018)

?

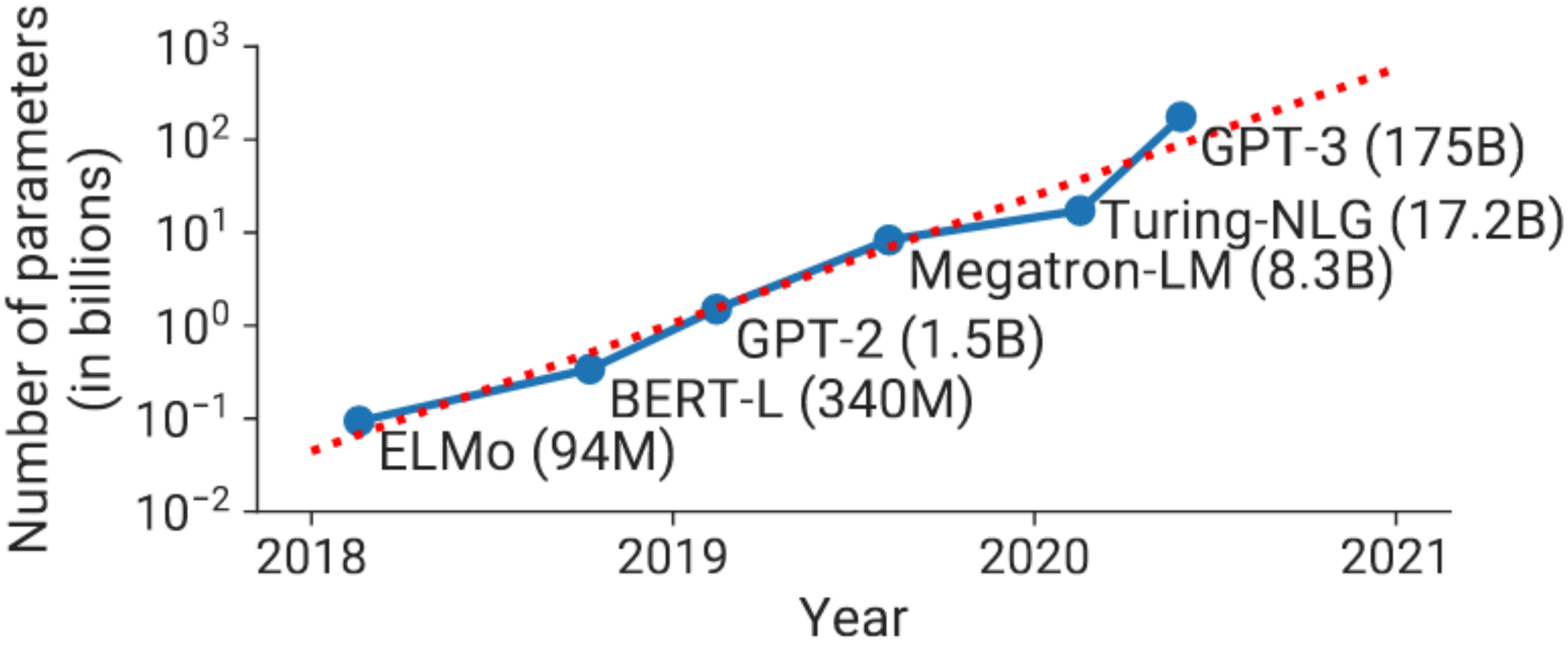
Act 1: Intro -
Rules/Symbolic

Act 5: catastrophe
or happy end?

Slide inspired by a talk by Verena Rieser

Language Models have grown tremendously in the last 4 years

- #parameters are growing at an exponential rate

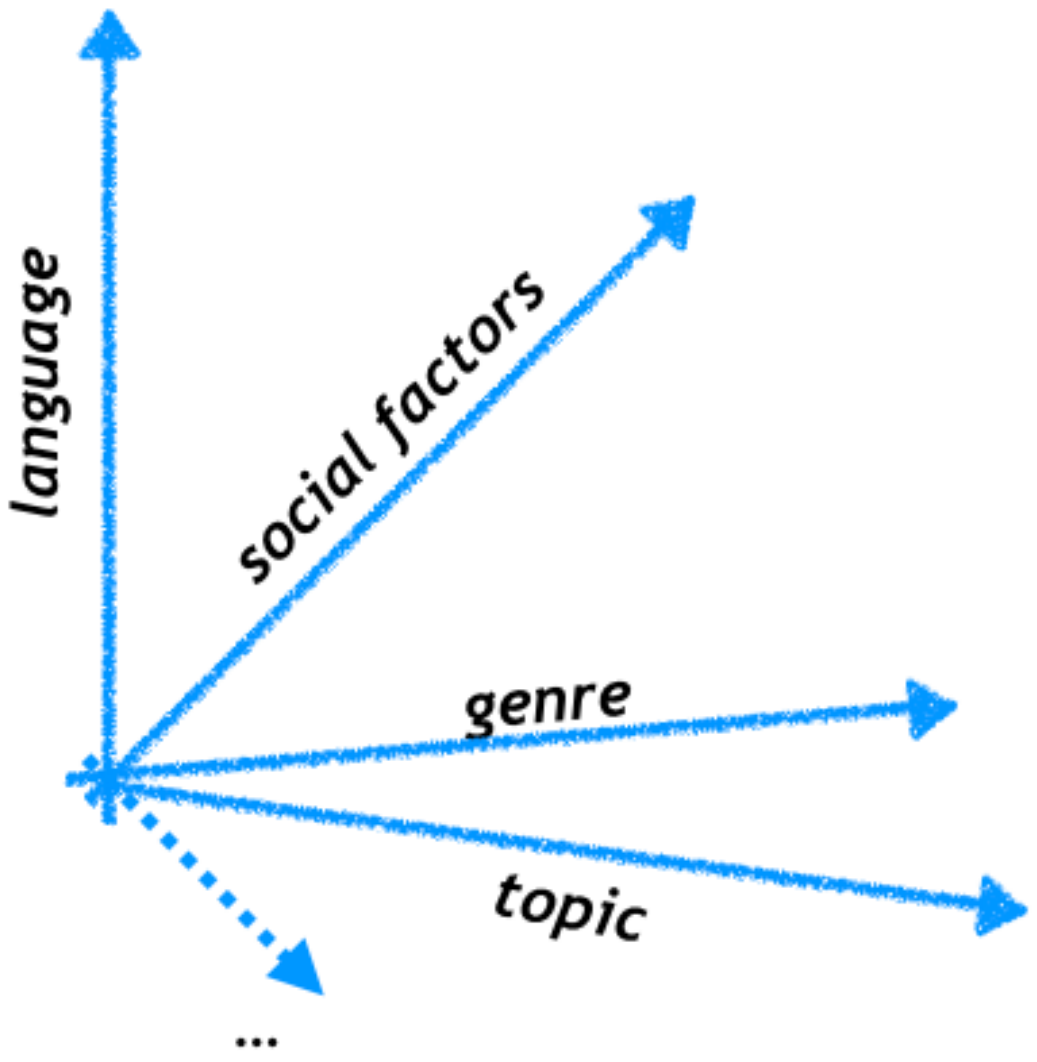


Narayanan et al. 2021

Do we just need to train larger models?

Language varies & is a social phenomenon

Domain shifts happen when collecting language data



variety space (Plank, 2016)

It's raining cats and dogs

Es regnet sehr stark

Es schüttet in Kübeln

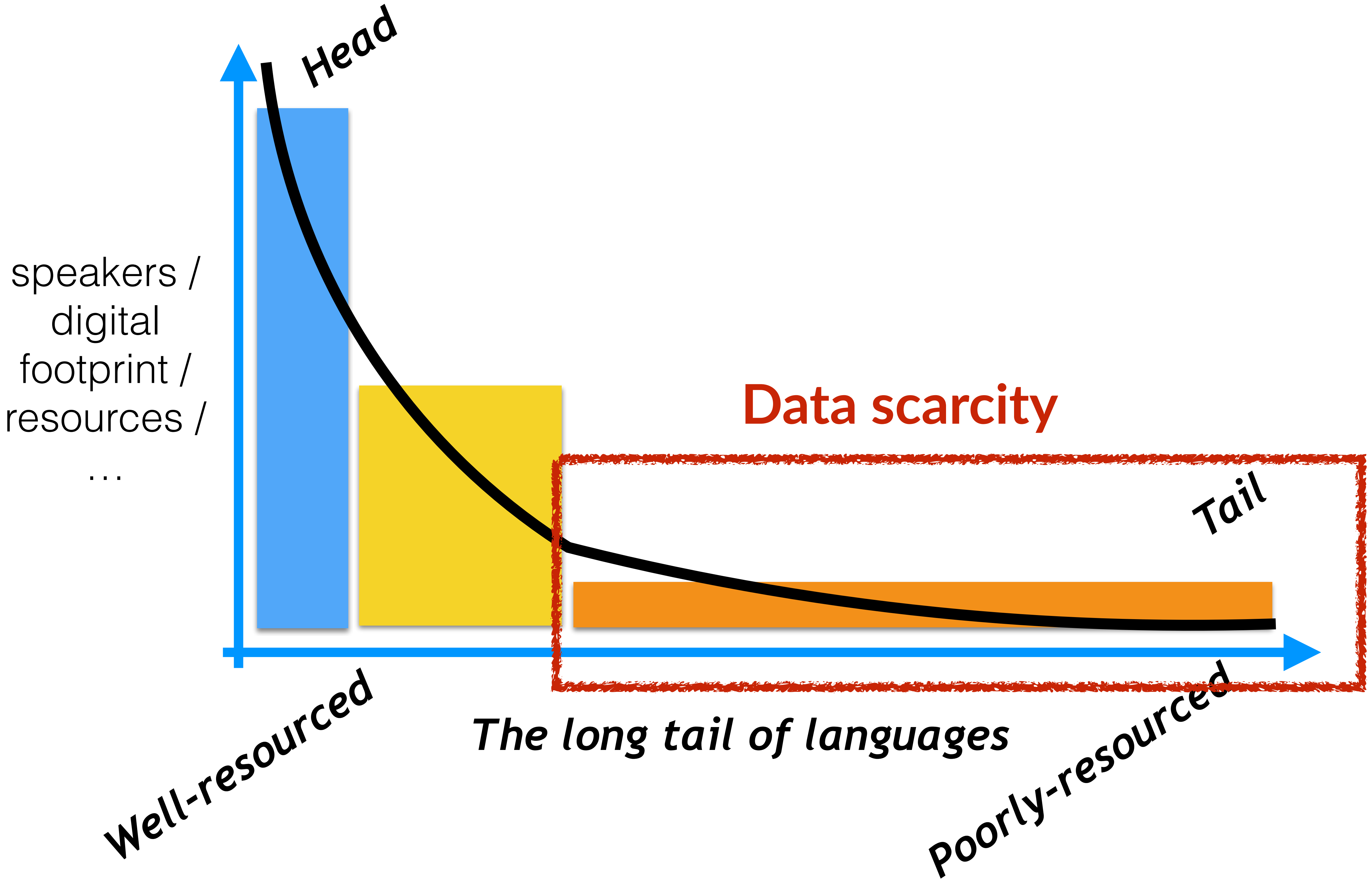
Do we just need to train larger models?

1. **Lack and bias of resources**

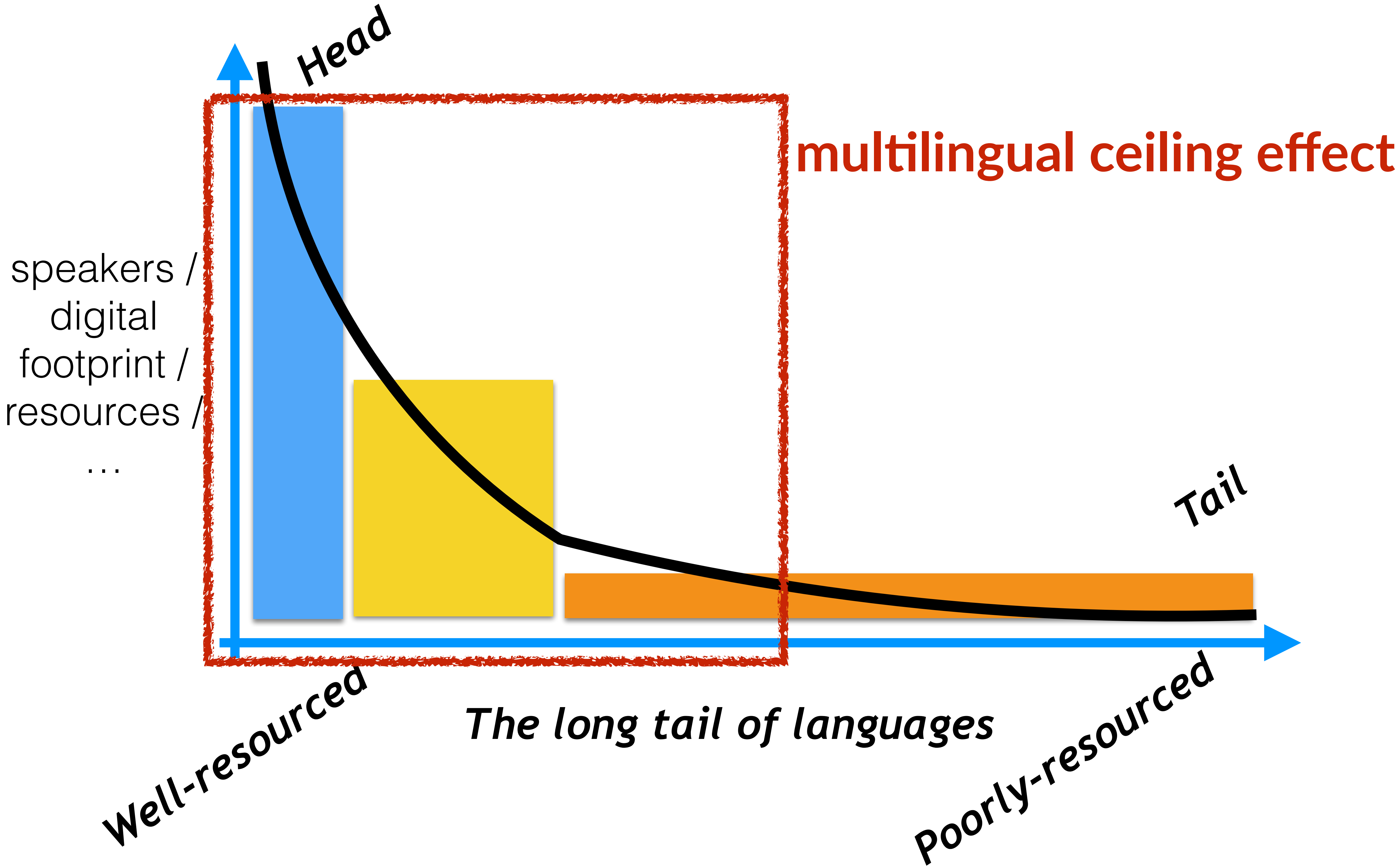
2. Controllability & Safety

3. Scaling

Lack of Resources

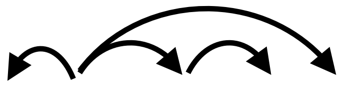

Head
speakers / digital footprint / resources / …
Data scarcity
Tail
Well-resourced
The long tail of languages
Poorly-resourced

“Curse of Multilinguality”

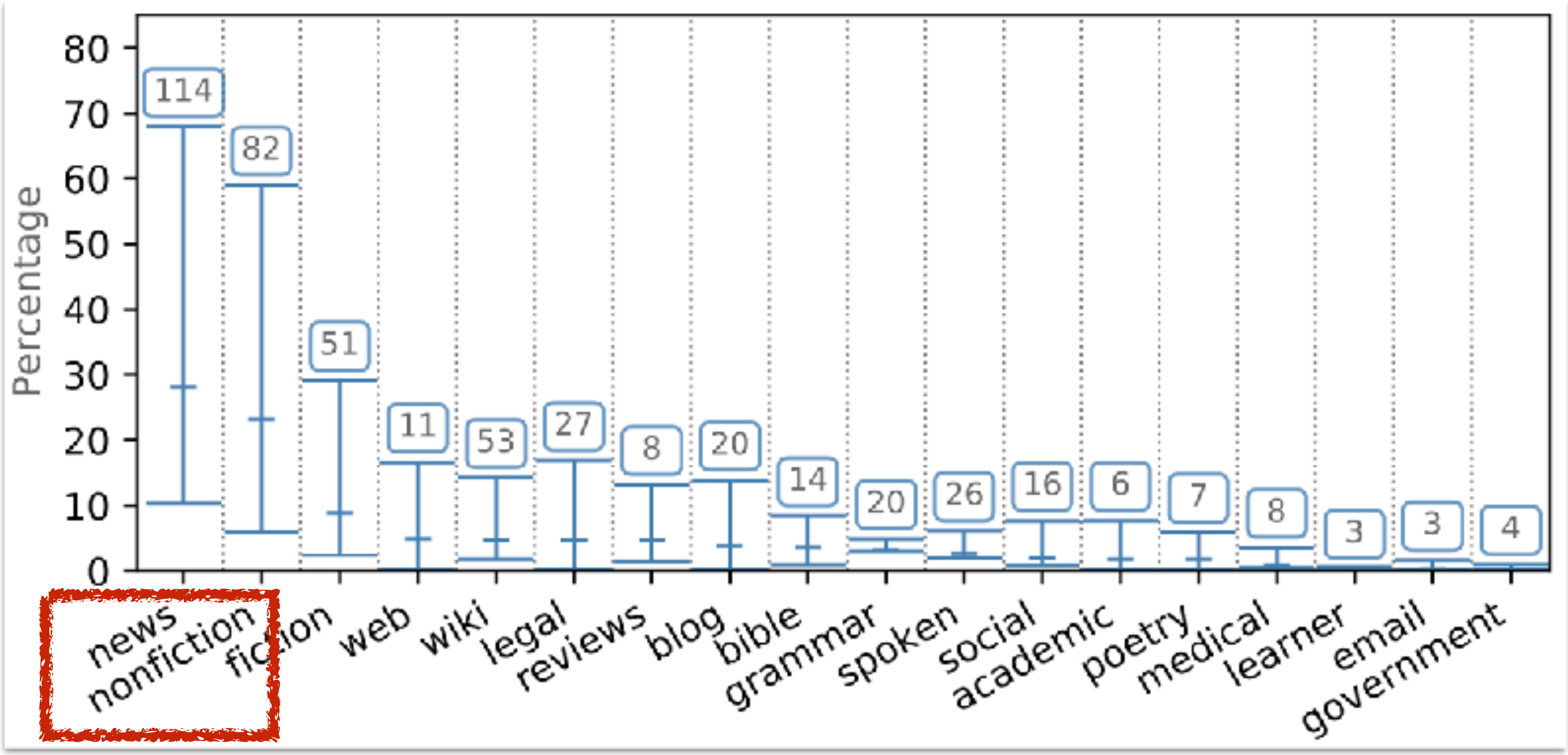

Head
multilingual ceiling effect
speakers / digital footprint / resources / …
Tail
Well-resourced
The long tail of languages
Poorly-resourced

Bias in Resources

Universal Dependencies data



40 languages/54 Treebanks in 2016 (v1.3)
Now: 114 languages/202 Treebank (2021; v2.8)



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

Selection bias: Newswire data is abundant

Do we just need to train larger models?

1. Lack and bias of resources

2. **Controllability & Safety Issues**

3. **Scaling Issues, Costs $$$ & Environment**

Do we just need to train larger models?

No.

Just scaling up LMs is not a (trustworthy) solution.

Ways to go further: Awareness!

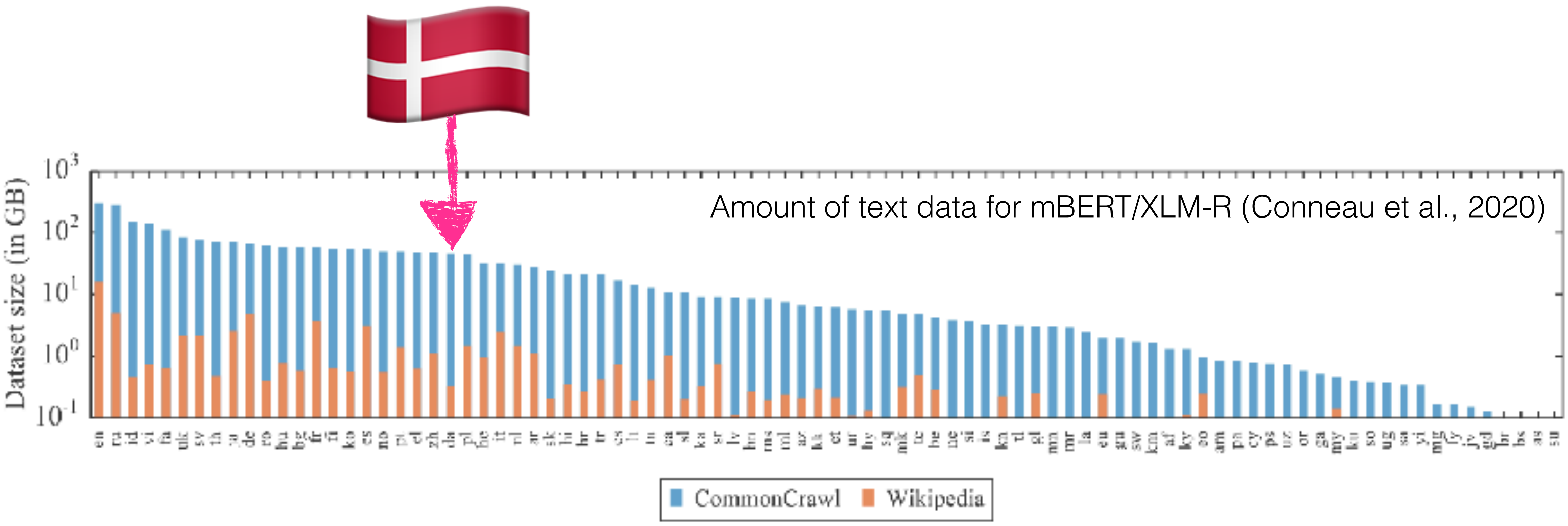
1. Data: open release of resources

2. Modeling: re-use of models, efficient modeling

3. Evaluation: awareness of limitations, embrace users

Importance of research and its larger scope or ecosystem with its implications

Selected research examples to address the lack of resources



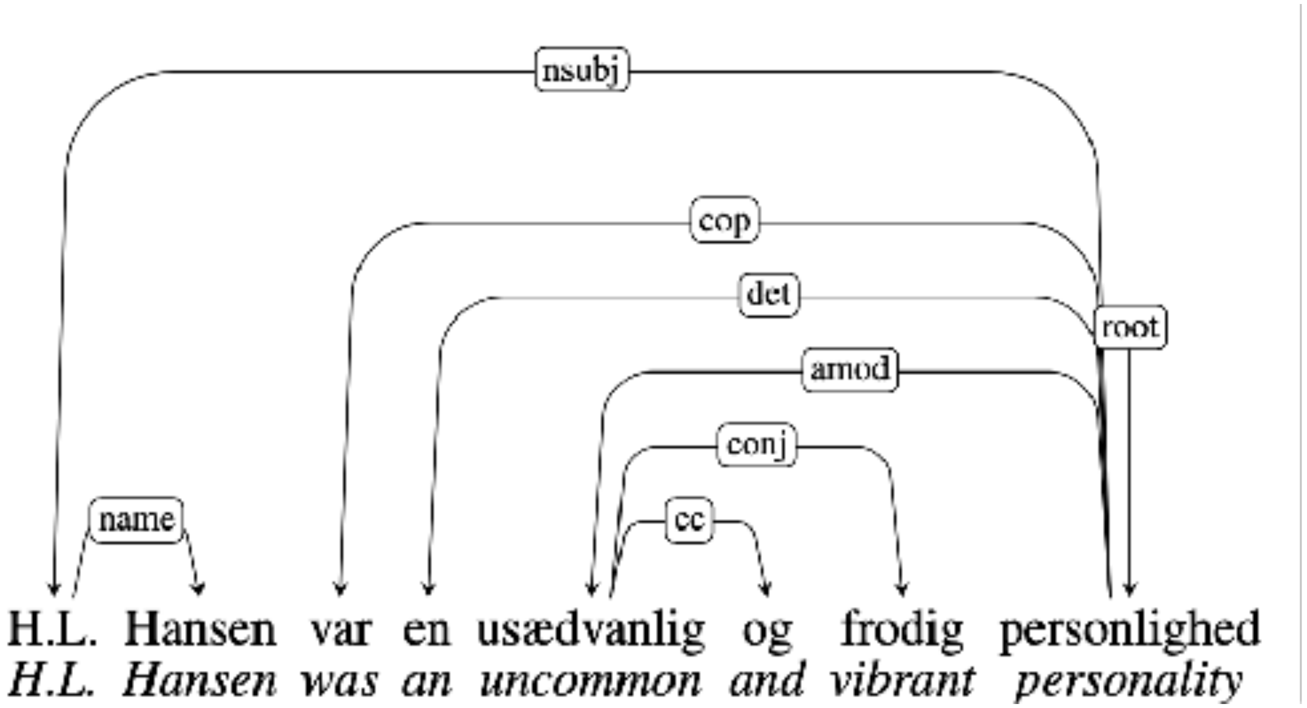
Amount of text data for mBERT/XLM-R (Conneau et al., 2020)

Selected overview of resources for Danish we contributed

🇩🇰

NLP for Danish: Dependency Parsing

- **Universal Dependencies** (UD): Syntactic dependency structure
  - UD for Danish (Johannsen et al., 2015): Conversion of the Copenhagen/Danish Dependency Treebank (Pritt, 1998, Buch-Kromann et al., 2003)

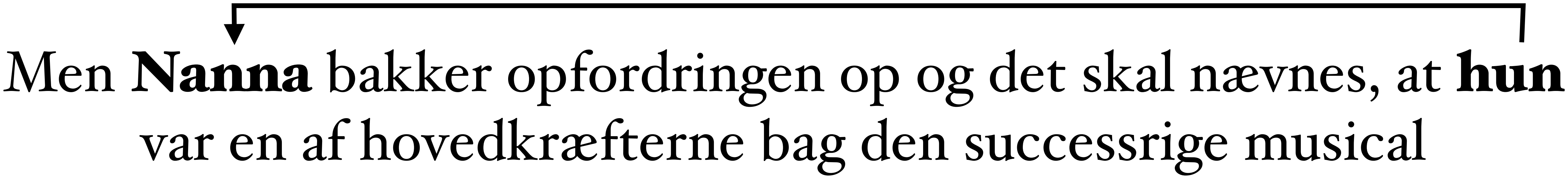

nsubj
cop
det
root
amod
conj
name
cc
H.L. Hansen var en usædvanlig og frodig personlighed
H.L. Hansen was an uncommon and vibrant personality

📒 data: Danish_DDT

📄 paper: (Johannsen, Martinez-Alonso, Plank, 2015)

NLP for Danish: Coreference Resolution

- **Co-reference resolution:** Identification of references to the same entity in text



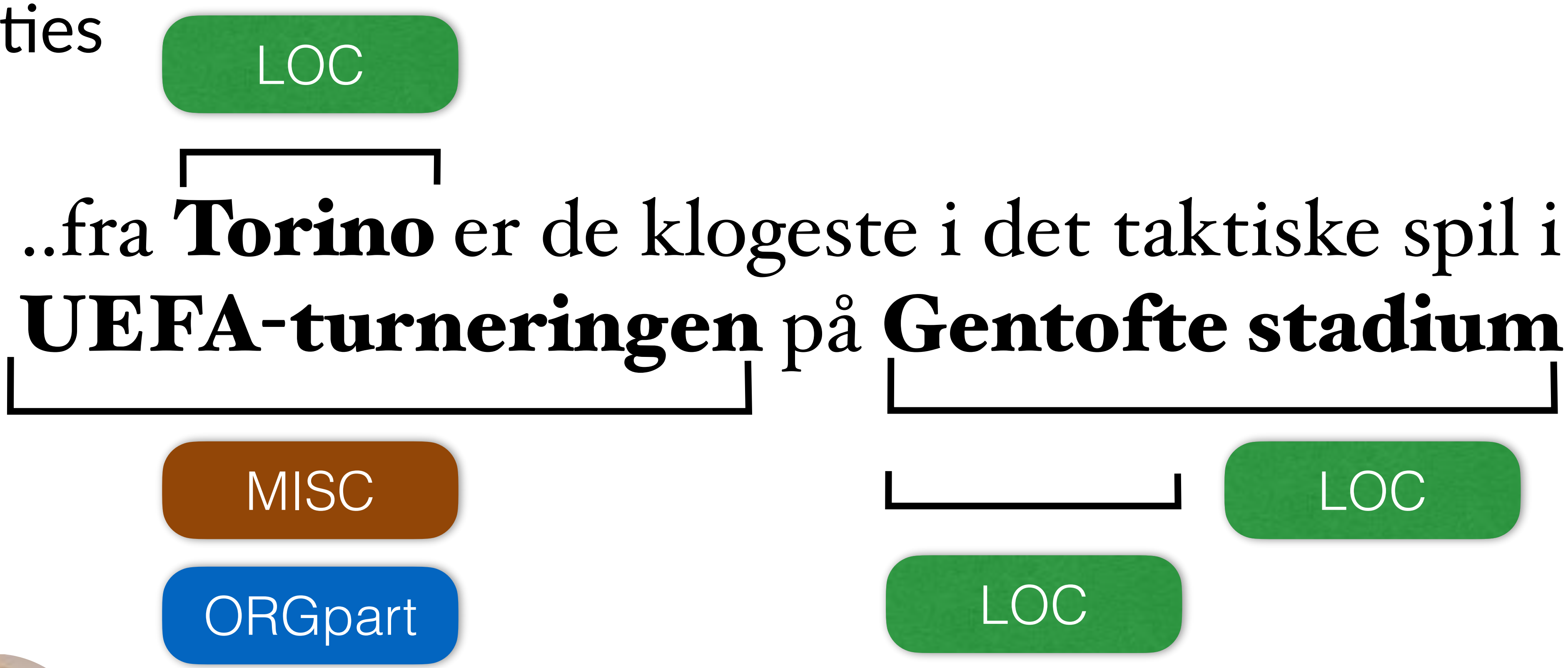
Men **Nanna** bakker opfordringen op og det skal nævnes, at **hun** var en af hovedkræfterne bag den successrige musical

data: https://github.com/alexandrainst/danlp

paper: (Barrett et al., 2021 CRAC)

NLP for Danish: Nested Named Entities

- **NER** to recognise People, Organization, Locations, and other named entities in text
- **DaN+**: **Nested** Named Entity Recognition (NNER) over 4 text varieties



..fra **Torino** er de klogeste i det taktiske spil i **UEFA-turneringen** på **Gentofte stadium**

📒 data: https://github.com/bplank/DaNplus

📄 paper: (Plank et al., 2020 COLING)

NLP for Danish: Lexical Normalization

- **Lexical normalization**: standardisation of non-standard text
- **DaN+:** +lexical normalization evaluation data for 2 domains
  - Part of MultiLexNorm international evaluation campaign

De skarpe lamper gjorde **destromindre ek** bedre

→

De skarpe lamper gjorde destro mindre ikke bedre

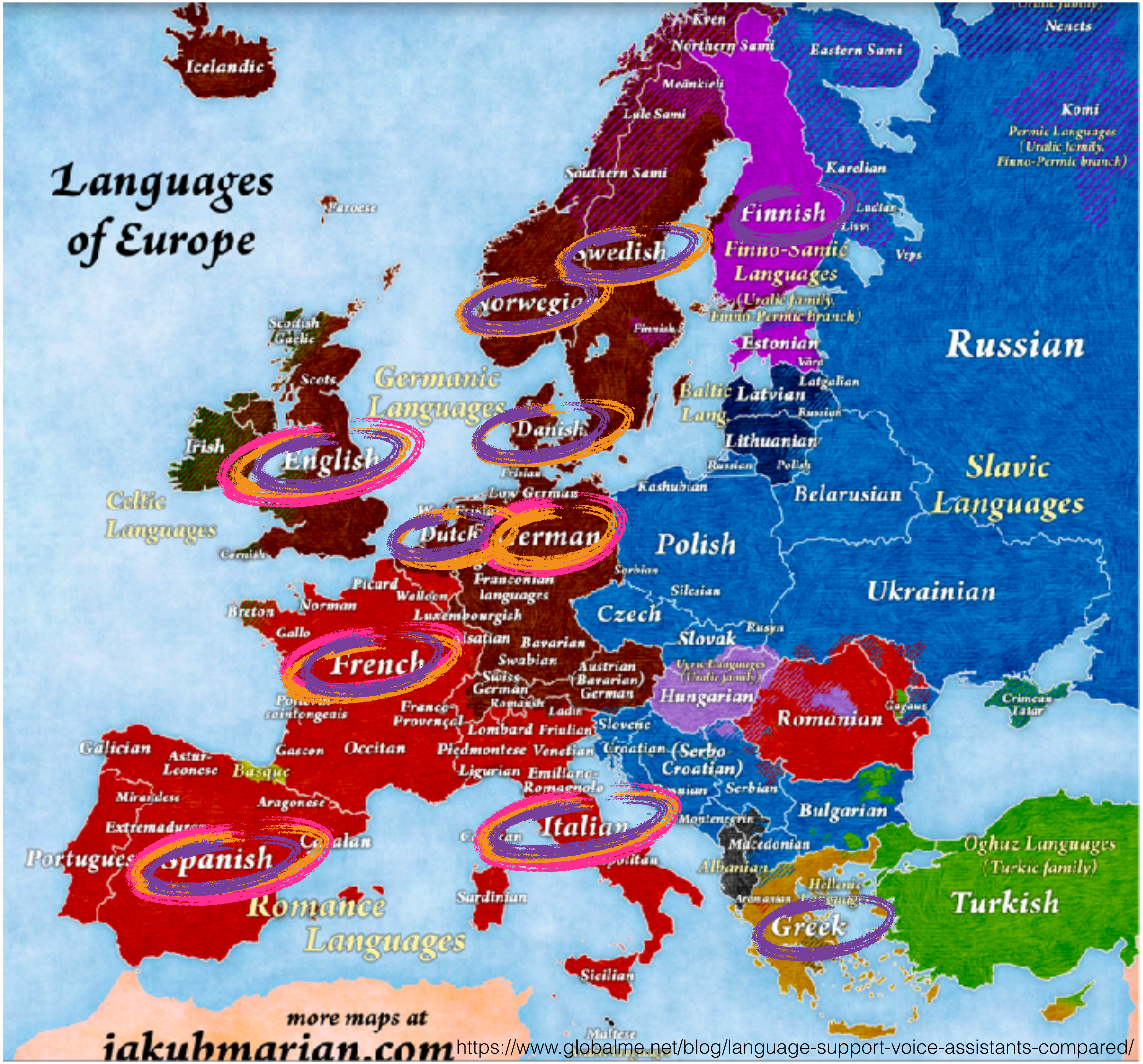
📒 data: https://github.com/bplank/DaNplus

📄 paper: (Plank et al., 2020 COLING)

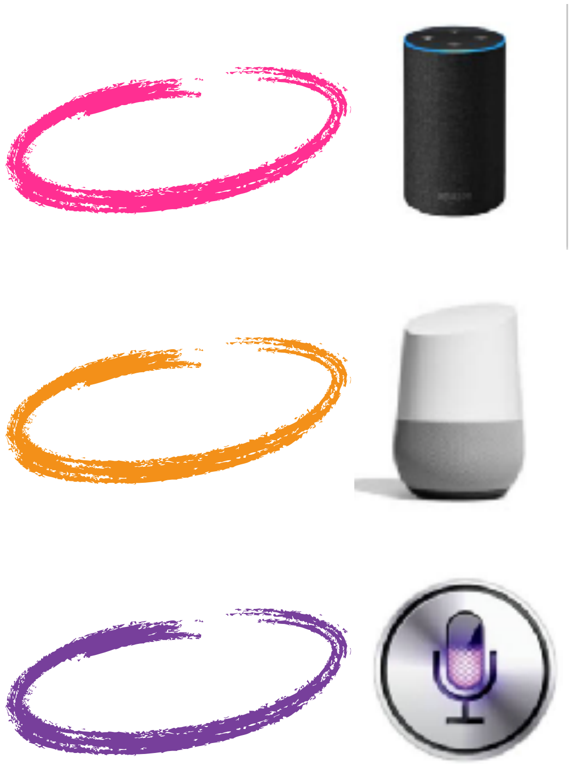
🎯 shared task: http://noisy-text.github.io/2021/multi-lexnorm.html

Example: Languages in EU covered by voice assistants

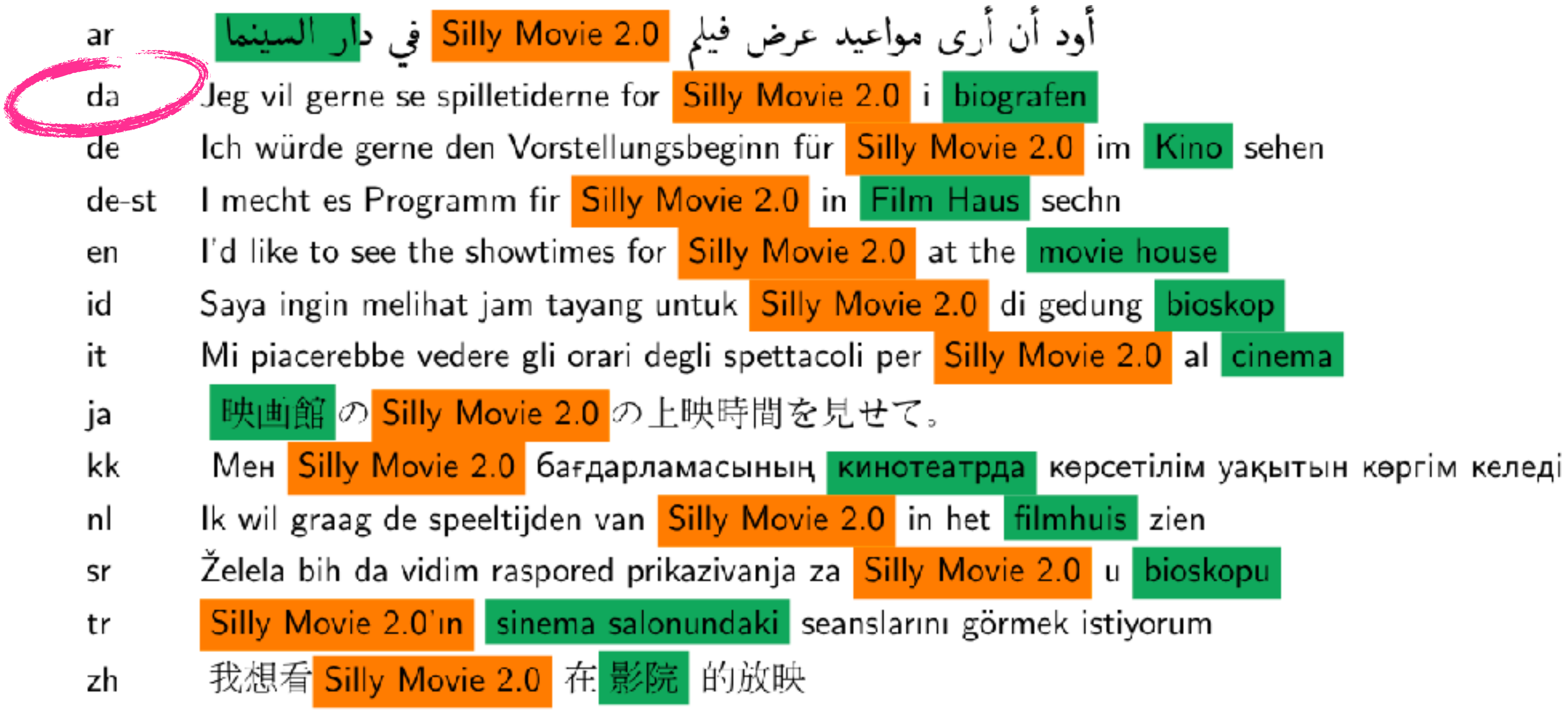
*as of March, 2020



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



NLP for Danish (and 12 more language variants): Slot and Intent detection



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

In part supported by:
amazon
Research Award

data: https://bitbucket.org/robvanderg/xsid

paper: (van der Goot et al., 2021 NAACL)

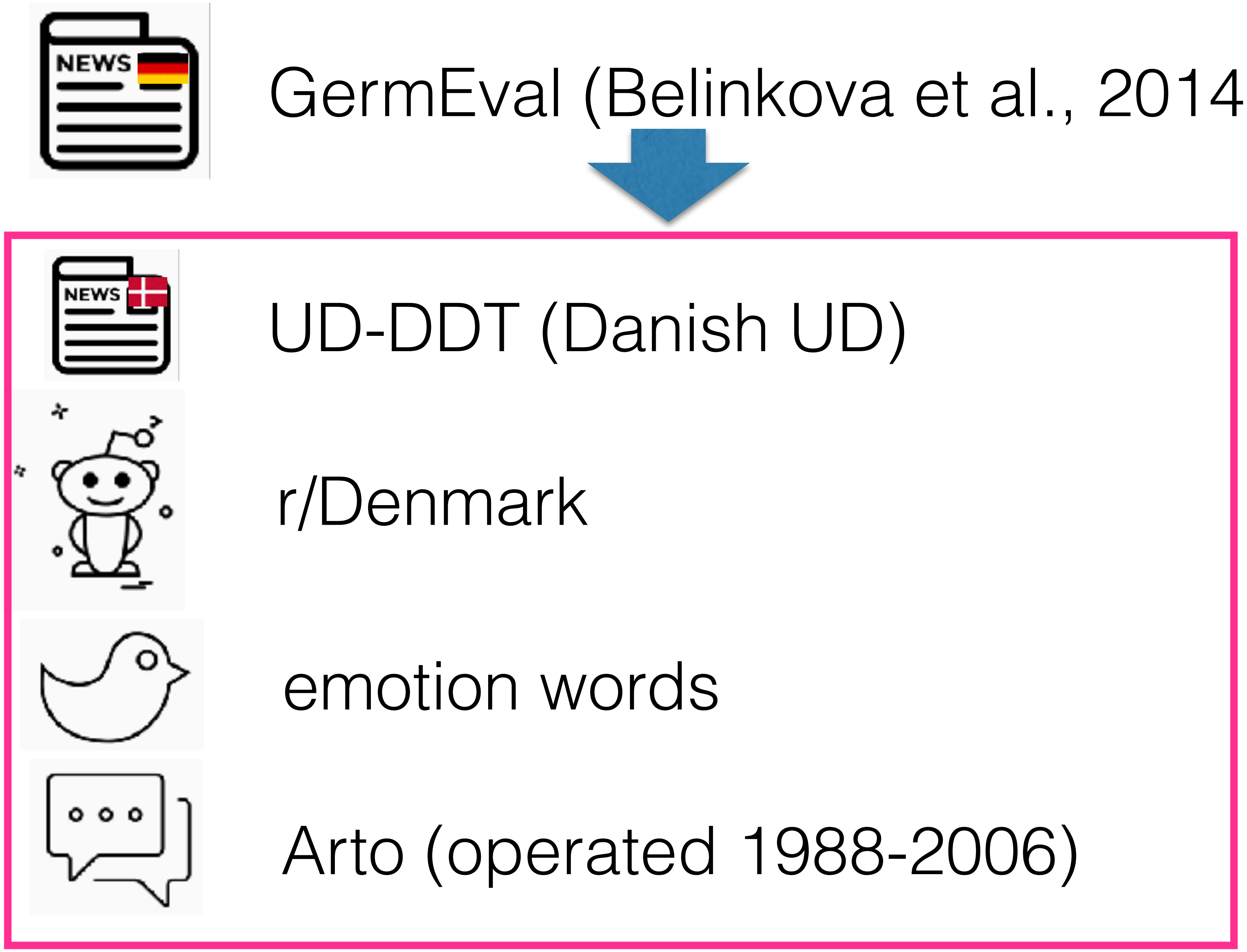
Selected Research towards more Inclusive and Robust NLP

Cross-domain Nested NER - Motivation

- NER studies on Danish **focus on newswire**:
    - First evaluation, part of UD-Danish (Plank, 2019, NoDaLiDa)
    - Annotation of full UD-Danish (Hvingelby et al., 2020 LREC)

- Focus on “flat” named entities and neglect non-noun forms:
    - University of Copenhagen
    - Den tyske ambassade

Danish Nested Named Entities and Normalization (DaN+)



GermEval (Belinkova et al., 2014)

UD-DDT (Danish UD)

r/Denmark

emotion words

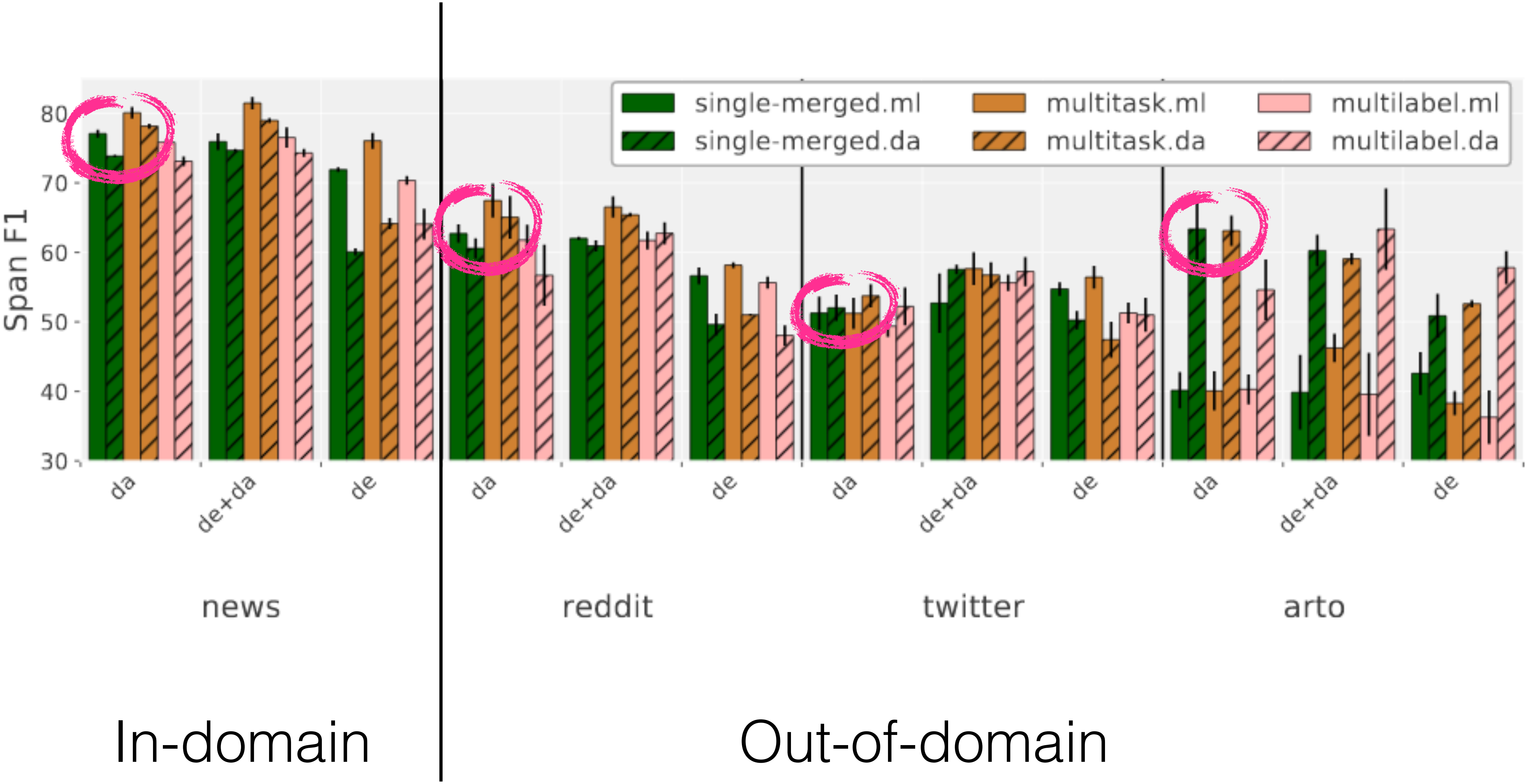
Arto (operated 1988-2006)

DaN+

Setup:

- cross-domain eval.
- 2 layers: single-task learning vs multi-task learning (MTL)
- Danish vs ml-BERT

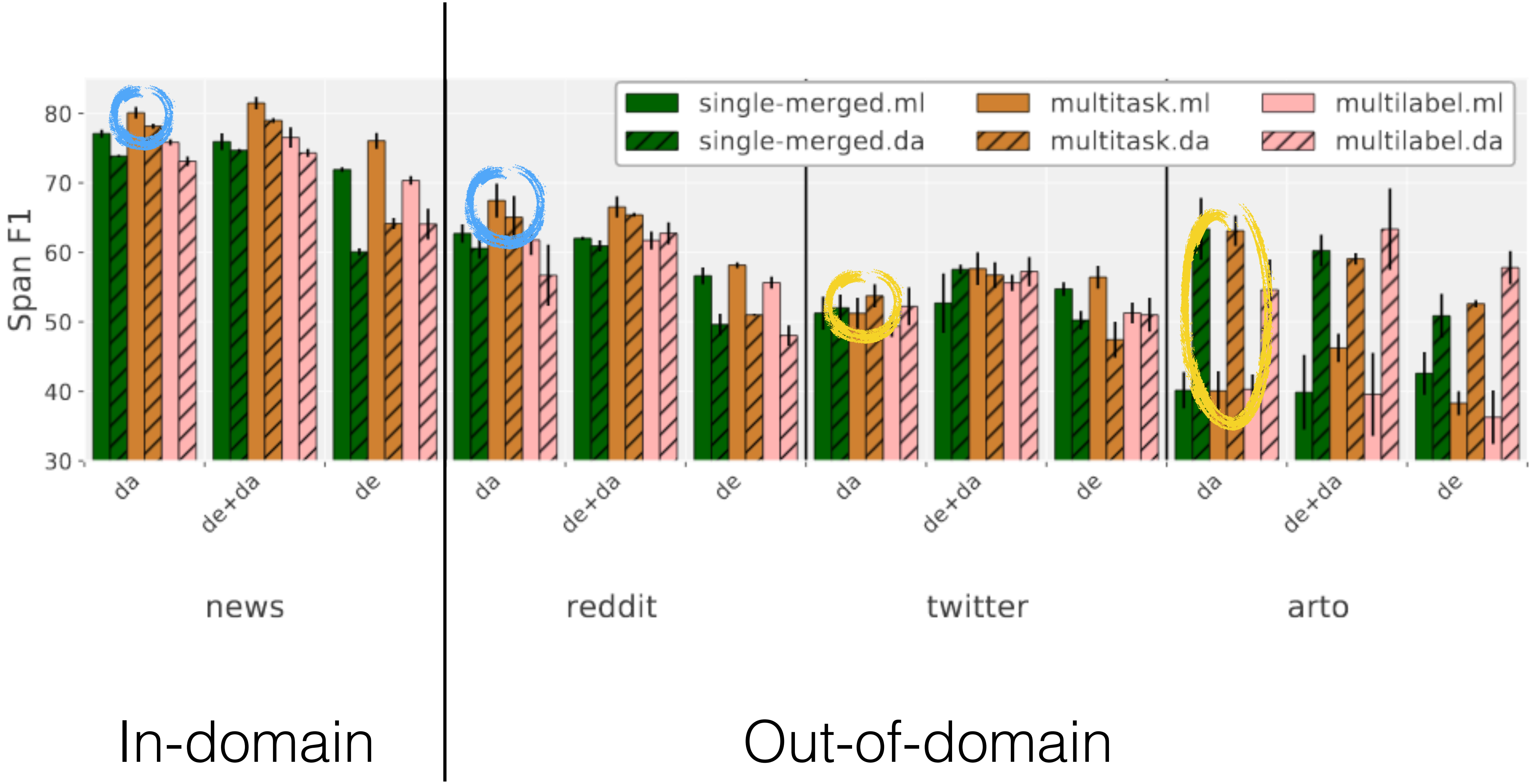
Results for Nested NER: MTL vs STL



In-domain

Out-of-domain

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

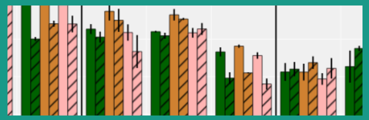


In-domain

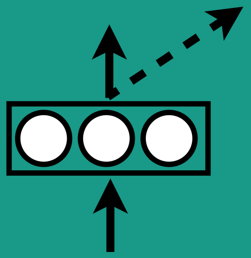
Out-of-domain

Take-aways

1. **DaN+** a new corpus for Danish NER (+ lexical normalisation)



2. Domains shift matters
   No free lunch: no BERT variant best overall

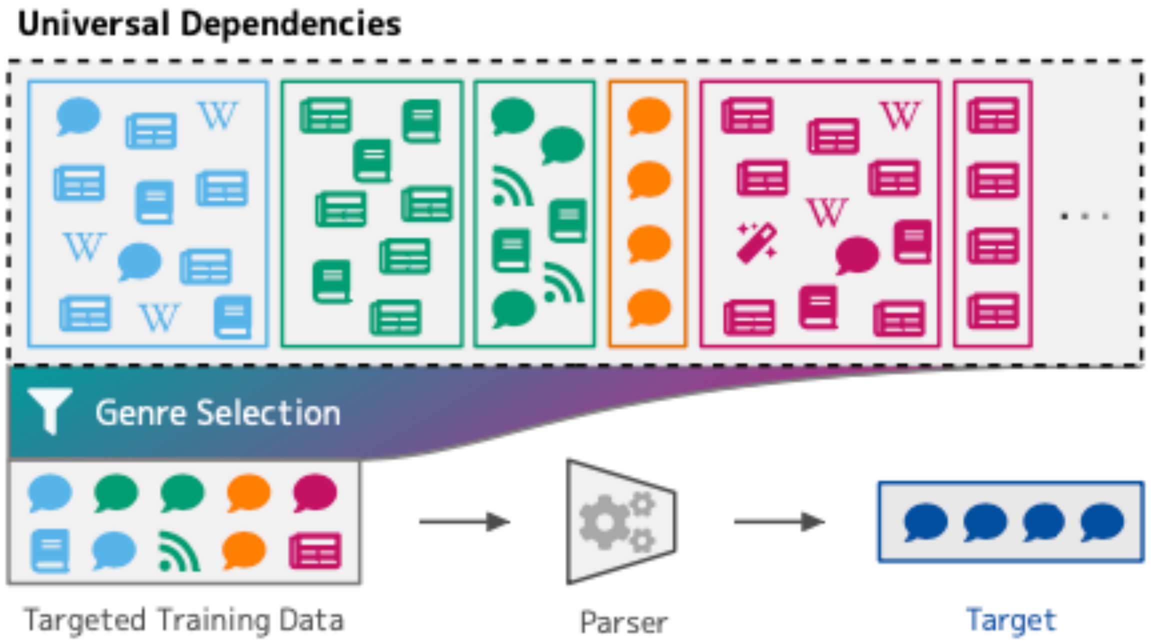


3. Best modelling approach: multi-task learning for nested NER

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

Data Selection for Low-resource parsing

- **Problem:** a single parser trained on 100+ languages is suboptimal and training is inefficient; for a practitioner it is also difficult to choose appropriate training material

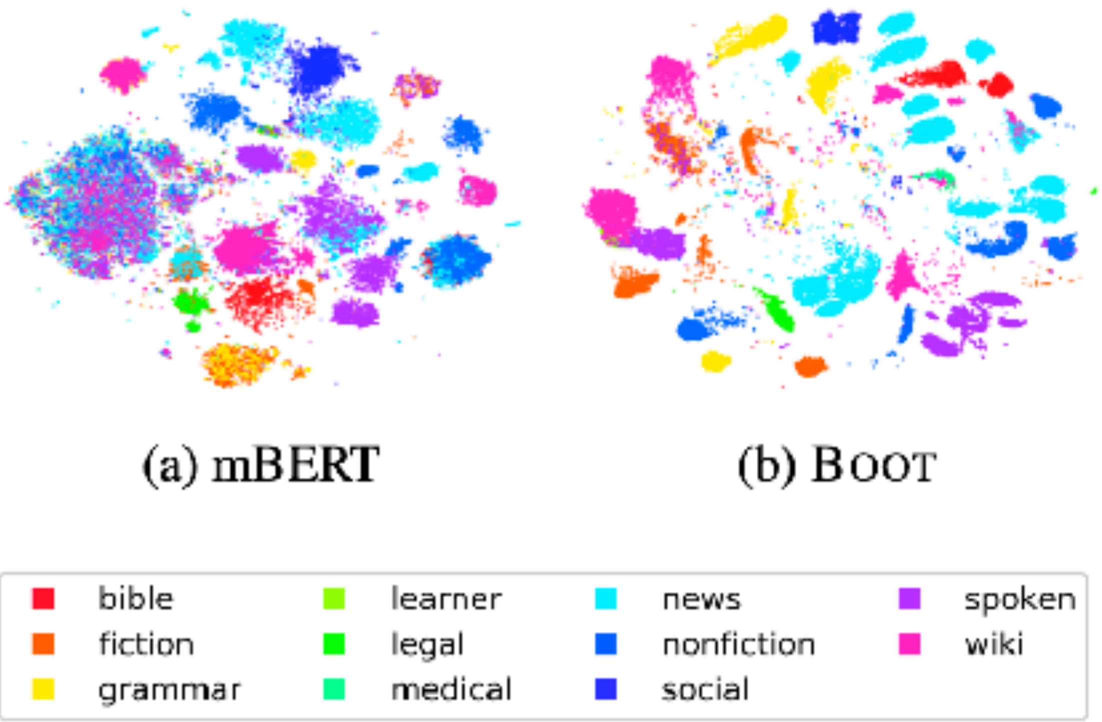


- **Key idea:** we propose **genre** as weak supervision to aid better target data selection for parser training - Is genre inherently captured in multilingual LMs? Can we amplify it?

Müller-Eberstein, van der Goot, Plank (EMNLP 2021) https://arxiv.org/abs/2109.04733

Amplifying genre to improve parsing

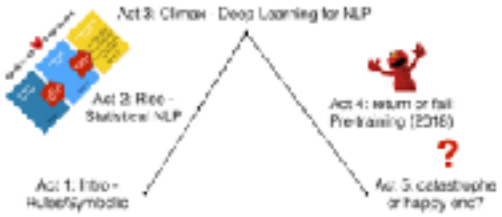
- We propose genre as signal for **weakly**-supervised learning
- **Genre** is captured in large multilingual MLMs


(a) mBERT (b) Boot

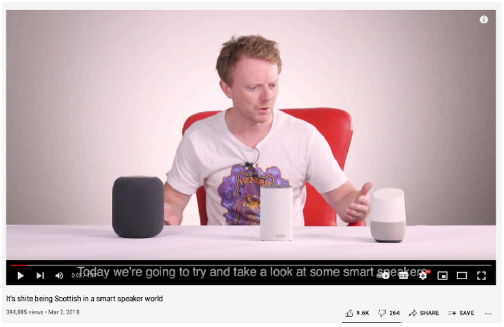
- Amplifying genre improves cross-lingual zero-shot parsing
  - 12 low-resource languages (incl. Faroese: 61 to 68% LAS)
  - Can help to create parsers for new low-resource variants

To wrap up…

Conclusions


Act 3: Climax - Deep Learning for NLP
Act 2: Rise - Statistical NLP
Act 4: return or fall Pre-training (2018)
?
Act 1: Intro - Rules/Symbolic
Act 5: catastrophe or happy end?

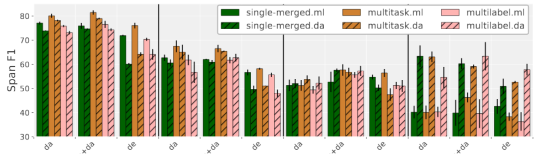
- NLP has grown tremendously


Today we're going to try and take a look at some smart speakers

It's shite being Scottish in a smart speaker world

394,885 views • Mar 2, 2018

- Biases are everywhere, Awareness is key



- Towards more inclusive & robust NLP

IT UNIVERSITY OF COPENHAGEN

Questions? Thanks!

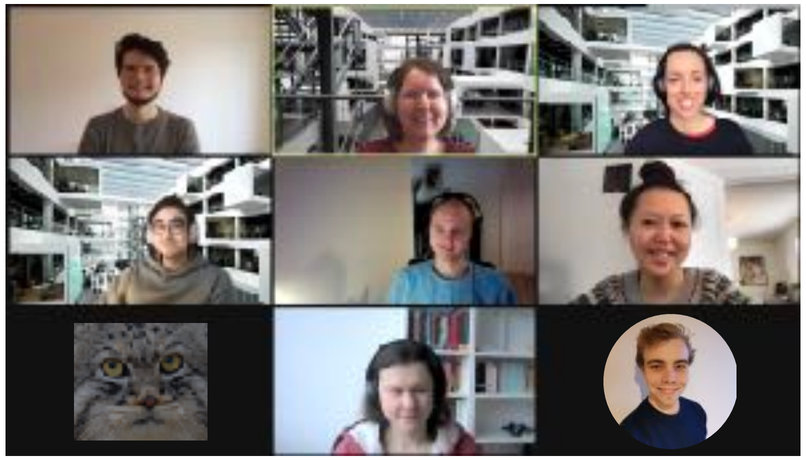
Natural Language Processing: Recent Advances and Challenges

bplank.github.io

More? Come see our posters

(e.g. Entity Disambiguation, CoRef, Information Extraction, de-identification)

nlpnorth.github.io



Research supported by:

DANMARKS FRIE FORSKNINGSFOND

amazon

NVIDIA