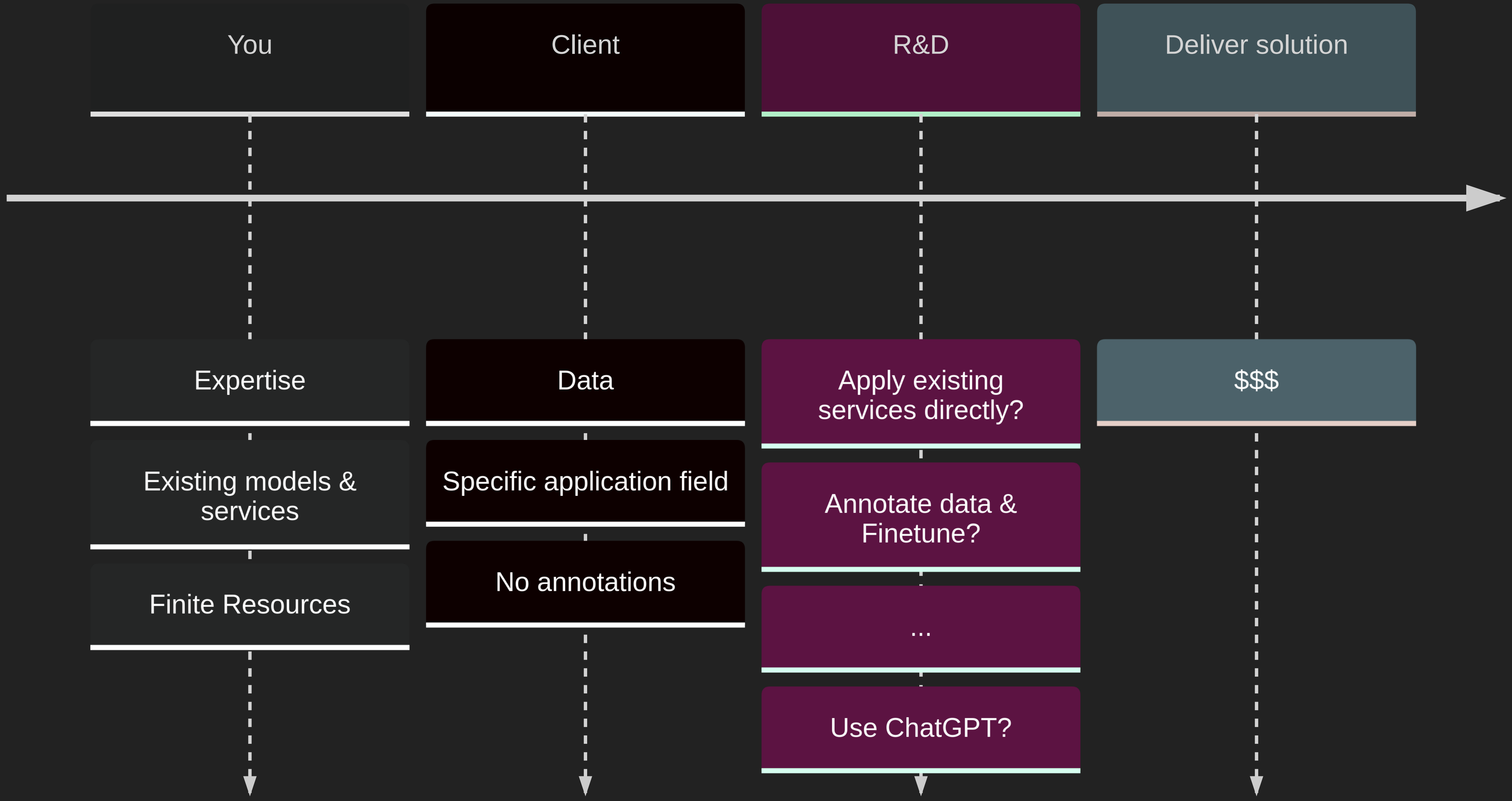
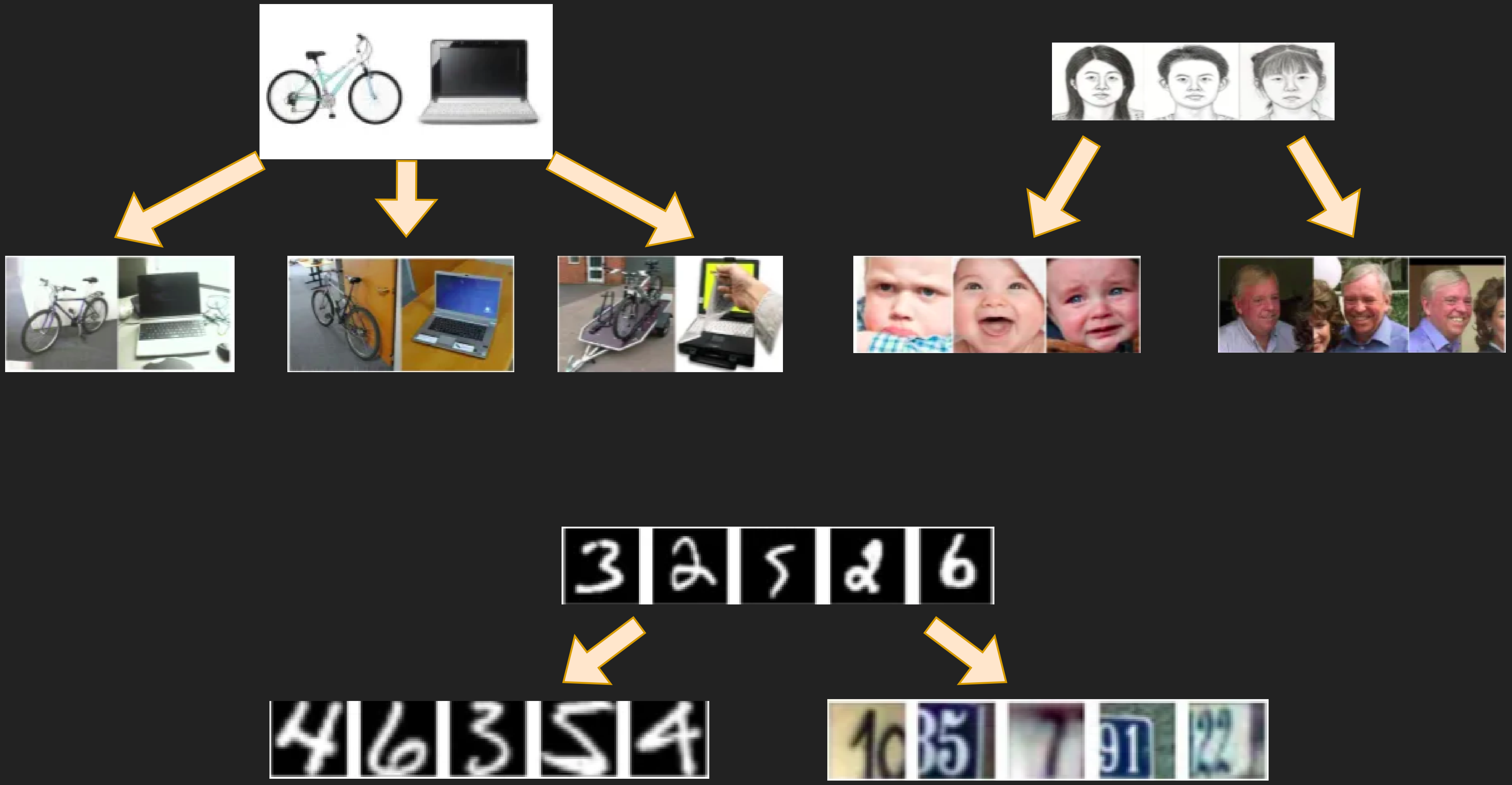
Sample-Efficient Domain Adaptation

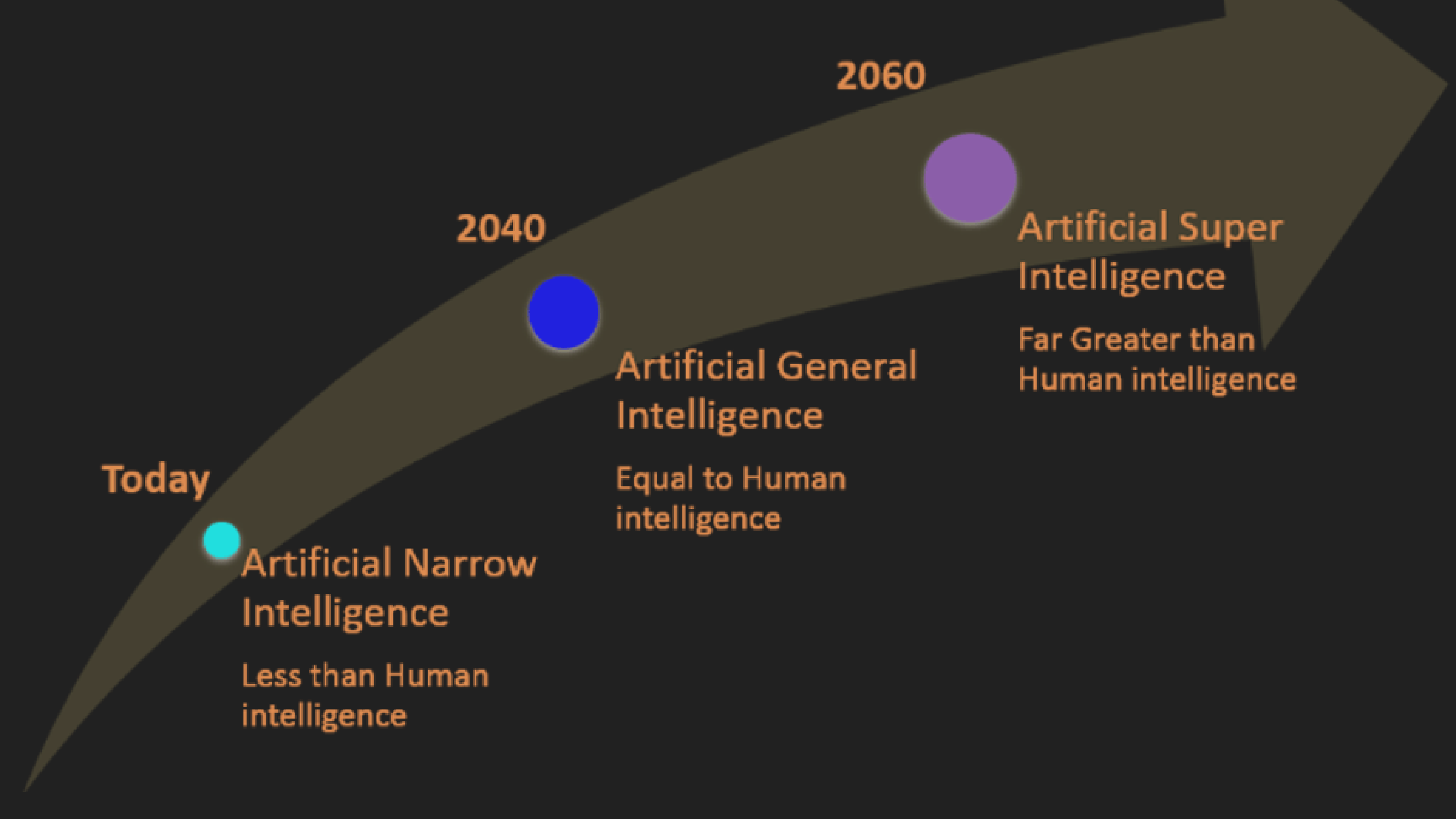
Giorgos Paraskevopoulos

ML projects

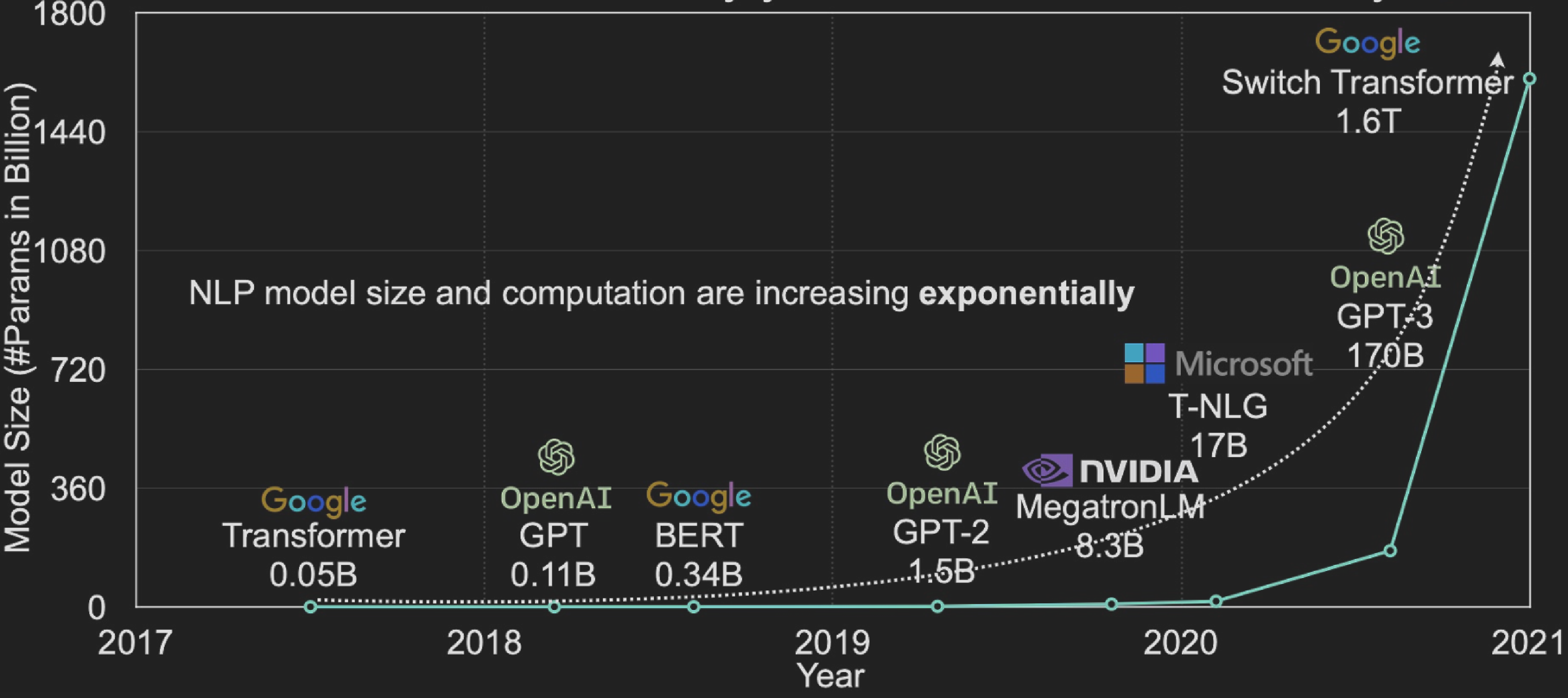

You
Client
R&D
Deliver solution
Expertise
Existing models & services
Finite Resources
Data
Specific application field
No annotations
Apply existing services directly?
Annotate data & Finetune?
...
Use ChatGPT?
$$$

One solution: Fast Domain Adaptation




Today
Artificial Narrow Intelligence
Less than Human intelligence
2040
Artificial General Intelligence
Equal to Human intelligence
2060
Artificial Super Intelligence
Far Greater than Human intelligence

NLP's Moore's Law: Every year model size increases by 10x



In this presentation

Domain adaptation to improve:

- small-ish models (300M parameters)
- with few in-domain data
- for new application settings

In this presentation

Domain adaptation to improve:

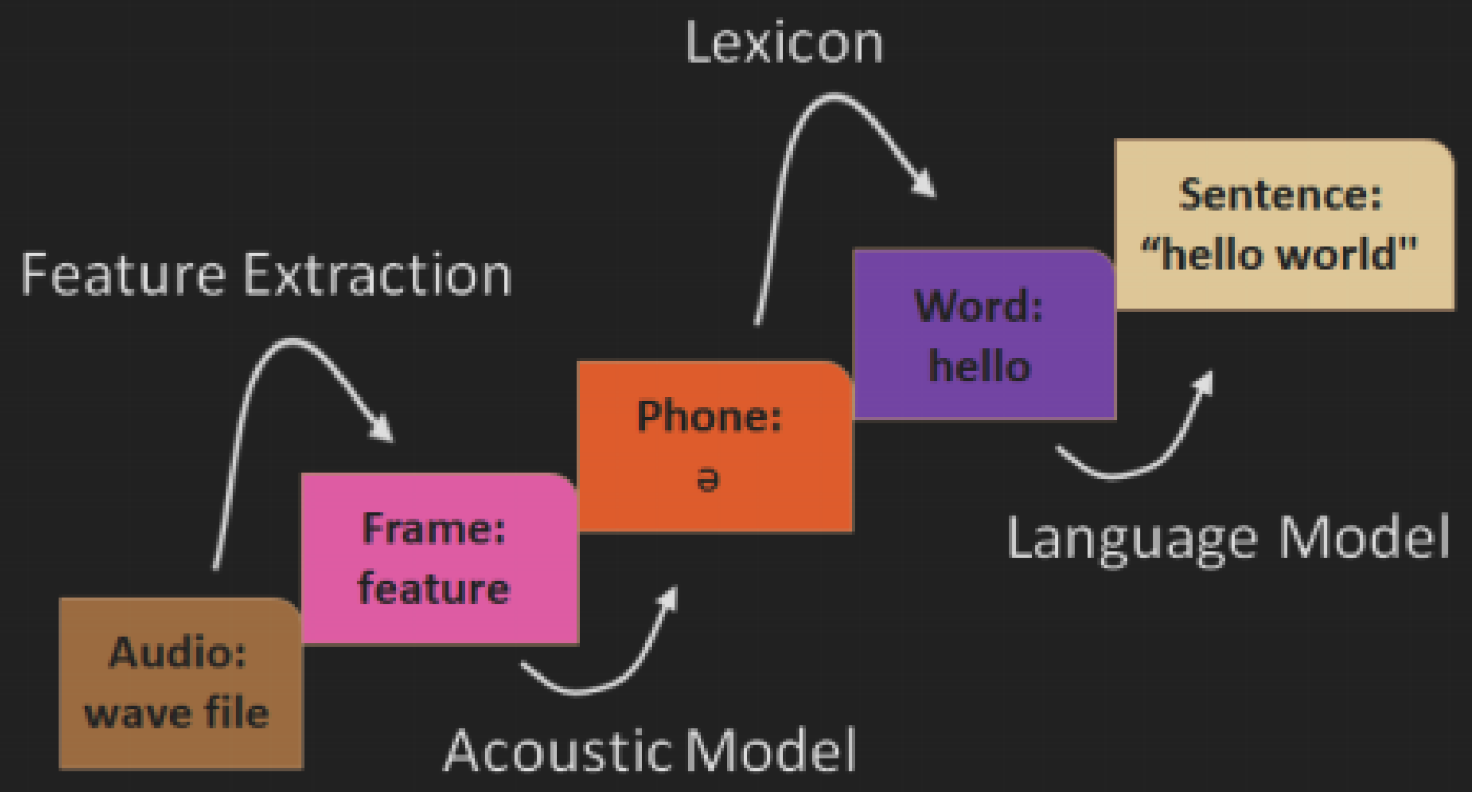
- small-ish models (300M parameters)
- with few in-domain data
- for new application settings

Use-cases:

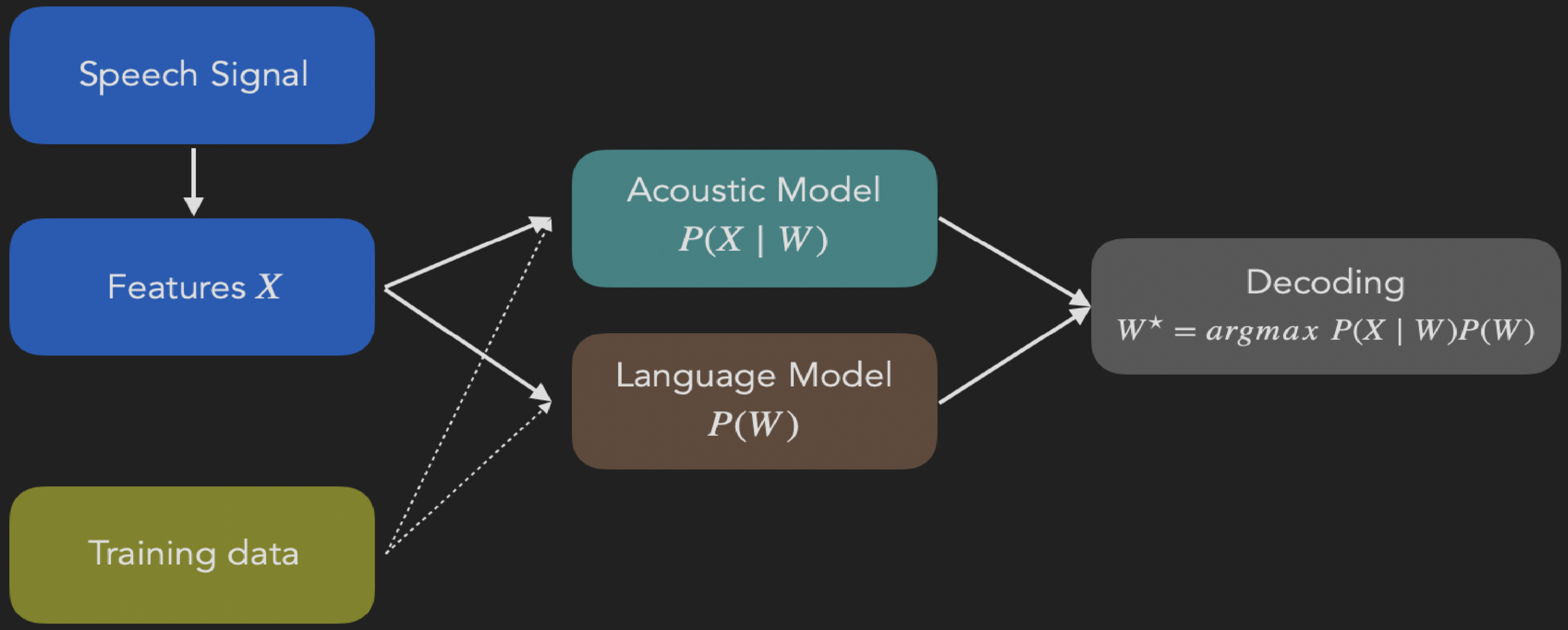
- Modular systems:
  - Automatic speech regognition (Language adaptation)
- End-to-end systems
  - Text classification (Sentiment analysis)
  - Automatic speech recognition (Acoustic adaptation)

Language adaptation for ASR systems

Modular ASR systems



Modular ASR systems



1. Collect in-domain text data

1. Collect in-domain text data

2. Add new terminology to lexicon

1. Collect in-domain text data
2. Add new terminology to lexicon
3. Adapt $P(W)$

1. Collect in-domain text data

2. Add new terminology to lexicon

3. **Adapt $P(W)$**

Very cheap: Train new n-gram LM and swap or interpolate the old one

Automated recipe: Trivial to apply for new application domains

Projects using this technique

Plan-V

Greek aphasic speech transcription and error detection

NLP-Theater Results

Speech Recognition

Model	WER
Google	32.4
Ours unadapted	39.2
Ours adapted	**15.7**

Subtitle Synchronization

Model	Error (mean)	Error (median)
Google	10.30	3.78
Ours adapted	**4.81**	4.43

- Error measured in seconds
- Avoids extreme errors

> G. Bastas, et al. "Towards a DHH Accessible Theater: Real-Time Synchronization of Subtitles and Sign Language Videos with ASR and NLP Solutions." PETRA. 2022.

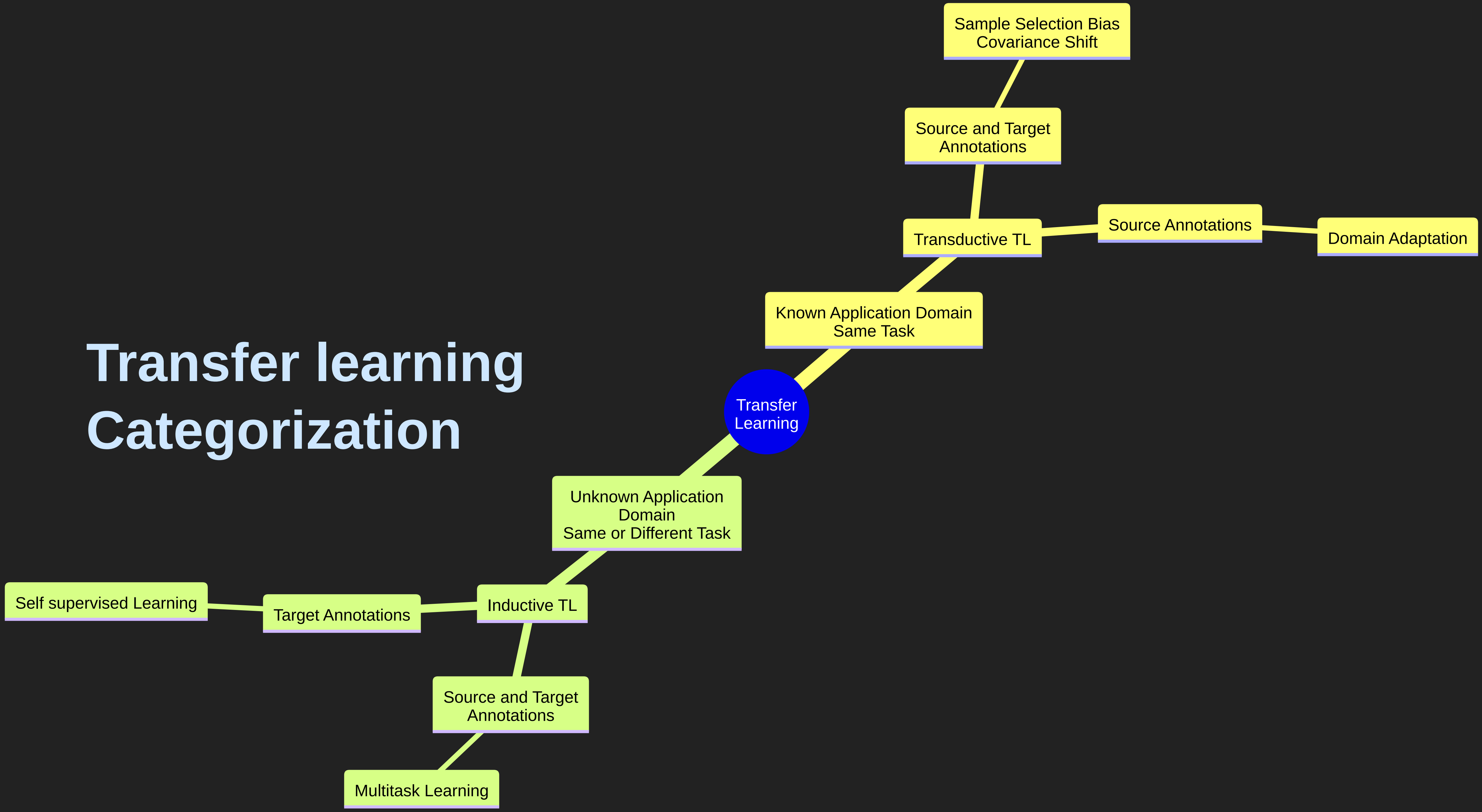
Plan-V Results

Lev. Distance	Percentage (%)
0	82.0
2	7.0
1	4.5
> 3	6.5

- Adapted model necessary
- Include mispronounced versions of words in lexicon
- Measure Levenshtein distance between transcribed word and ground truth
- Example: "καλοριθέρ" $\rightarrow$ "καλοριφέρ"

Domain adaptation for end-to-end models

Transfer learning Categorization



Popular techniques for end to end models

Pseudolabeling

- Train model on labeled out-of-domain data
- Use model to annotate unlabeled in-domain data
- Reduce the task to supervised learning on generated labels

Adversarial Training

- Manipulate the latent space so that the extracted features are domain-invariant
- Use an adversarial cost so that the network can't predict the domain based on the latent features

Self-supervision

- Use pretext tasks to gradually adapt the model to the target domain data distribution
- Learn the task on the source domain

Popular techniques for end to end models

Pseudolabeling

- Train model on labeled out-of-domain data
- Use model to annotate unlabeled in-domain data
- Reduce the task to supervised learning on generated labels

Adversarial Training

- Manipulate the latent space so that the extracted features are domain-invariant
- Use an adversarial cost so that the network can't predict the domain based on the latent features

Self-supervision

- Use pretext tasks to gradually adapt the model to the target domain data distribution
- Learn the task on the source domain

Popular techniques for end to end models

Pseudolabeling

- Train model on labeled out-of-domain data
- Use model to annotate unlabeled in-domain data
- Reduce the task to supervised learning on generated labels

Adversarial Training

- Manipulate the latent space so that the extracted features are domain-invariant
- Use an adversarial cost so that the network can't predict the domain based on the latent features

Self-supervision

- Use pretext tasks to gradually adapt the model to the target domain data distribution
- Learn the task on the source domain

Popular techniques for end to end models

Pseudolabeling

- Pro: Straightforward
- Pro: Well explored in the literature
- Con: Error propagation

Adversarial Training

- Pro: Theoretical background
- Pro: Truly e2e approach
- Con: Convergence can be challenging

Self-supervision

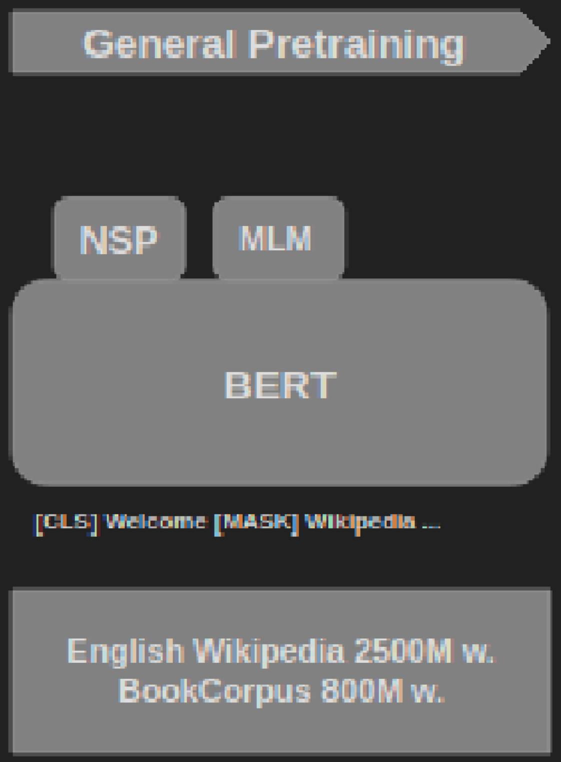
- Pro: Easy to apply
- Pro: In-domain sample-efficiency
- Con: Computationally more expensive

Use case 1: Sentiment Analysis

C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Step 1

- Our approach is based on BERT
- We start from a pretrained model



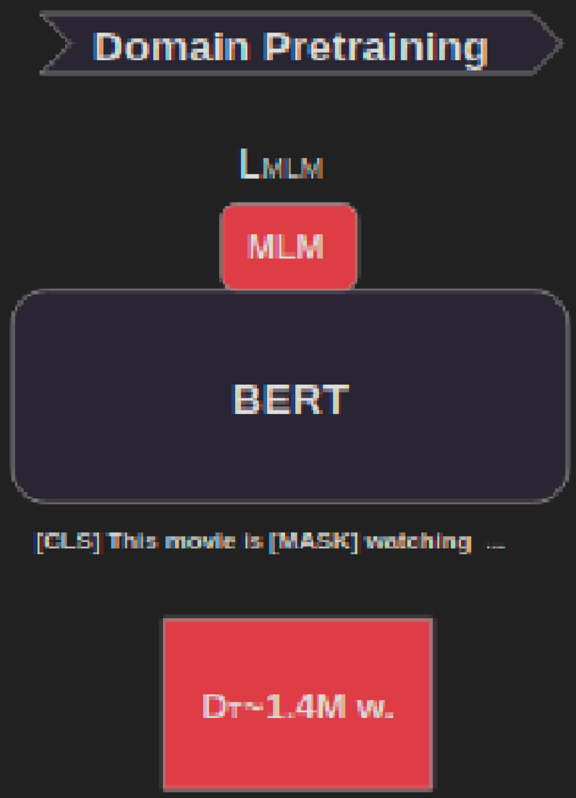
C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Step 2

- Continue pretraining

- On unlabeled target data
- With MLM task

- Makes BERT aware of target domain



C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

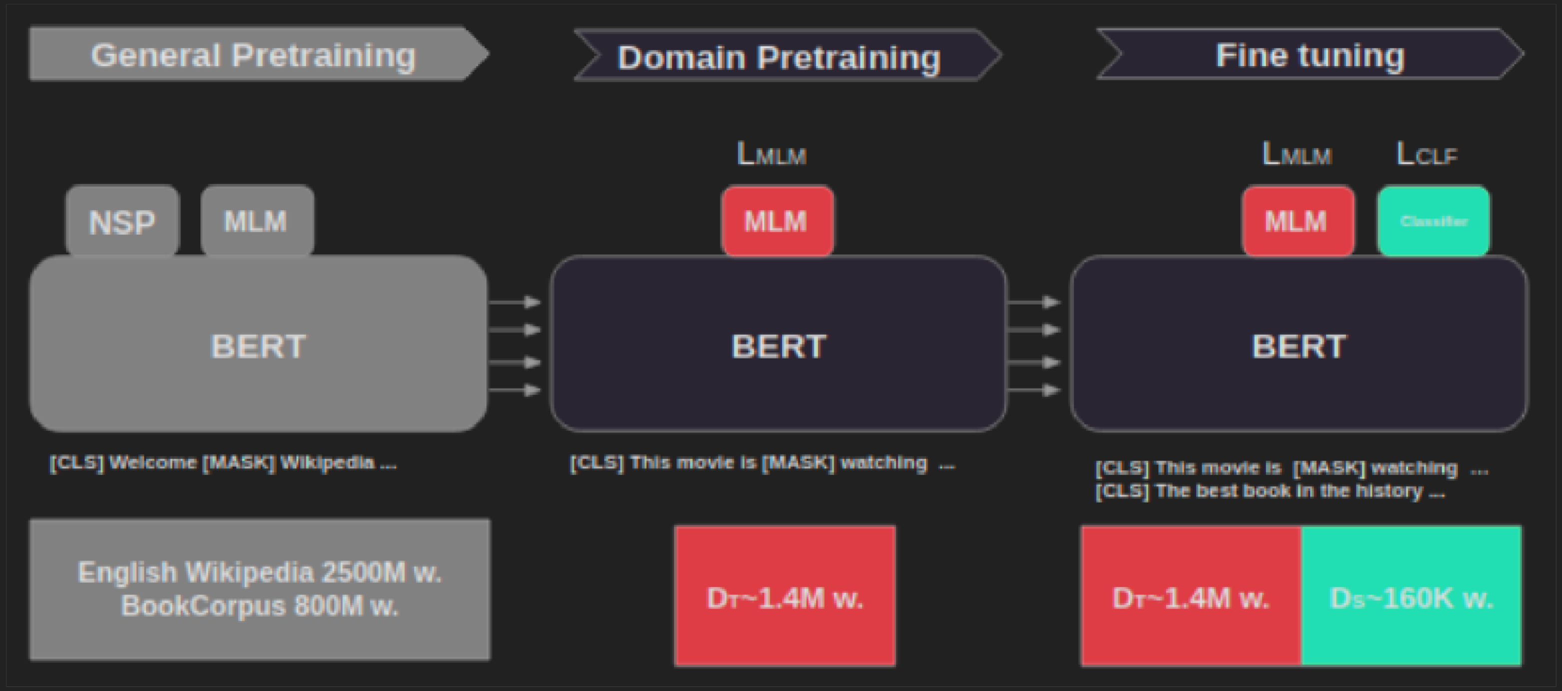
Step 3

- Add a sentiment classifier
- Keep MLM in a multi-tasking manner
- Mixed batches of source and target data
- Labeled source data for classification
- Unlabeled target data for MLM

$$L(\mathbf{s},\mathbf{t}) = \lambda L_{CLF}(\mathbf{s}) + (1-\lambda) L_{MLM}(\mathbf{t})$$

C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Overview



> C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Dataset: Amazon reviews

- Standard benchmark dataset for domain adaptation.
- Binary sentiment classification task.
- Domains: Books (**B**), DVDs (**D**), Electronics (**E**), Kitchen appliances (**K**)
- 12 adaptation scenarios of source-target domain pairs (e.g. **B** → **D**).
- 2,000 labeled reviews per domain.
- 19,809 **B**, 19,798 **D**, 19,937 **E** and 17,805 **K** unlabeled reviews.

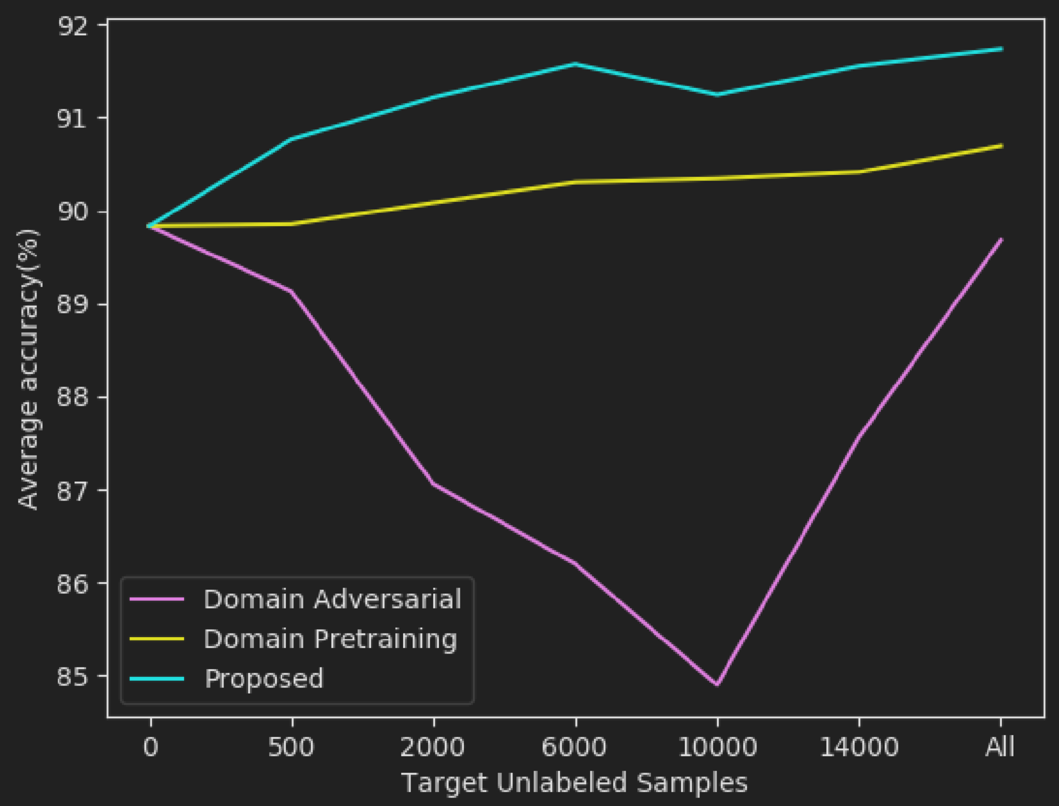
C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Results

	R-PERL	DAAT	p+CFd	Source Only	Adv.	Domain PT	Proposed
$B \to D$	87.8%	90.9%	87.7%	90.5%	90.7%	90.7%	**91.3%**
$B \to E$	87.2%	88.9%	**91.3%**	**91.3%**	91.1%	90.9%	91.2%
$B \to K$	90.2%	88.0%	92.5%	91.6%	92.8%	92.3%	**92.9%**
$D \to B$	85.6%	89.7%	**91.5%**	90.2%	90.6%	90.5%	91.4%
$D \to E$	89.3%	90.1%	91.6%	88.5%	88.8%	91.7%	**92.9%**
$D \to K$	90.4%	88.8%	92.5%	90.5%	92.0%	92.0%	**94.3%**
$E \to B$	90.2%	89.6%	88.7%	87.8%	89.4%	88.3%	**90.6%**
$E \to D$	84.8%	**89.3%**	88.2%	87.2%	86.5%	87.3%	88.4%
$E \to K$	91.2%	91.7%	93.6%	92.8%	94.6%	94.1%	**94.8%**
$K \to B$	83.0%	**90.8%**	89.8%	88.6%	83.6%	89.4%	89.4%
$K \to D$	85.6%	**90.5%**	87.8%	87.1%	83.6%	88.0%	89.2%
$K \to E$	91.2%	93.2%	92.6%	91.9%	92.4%	93.1%	**94.3%**
Average	87.50%	90.12%	90.63%	89.83%	89.68%	90.69%	**91.73%**

C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Sample-Efficiency



C. Karouzos, G. Paraskevopoulos, A. Potamianos. "UDALM: Unsupervised Domain Adaptation through Language Modeling." NAACL 2021.

Use case 2: Adaptation of XLSR-53 for ASR

G. Paraskevopoulos et al., Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

Key take-aways

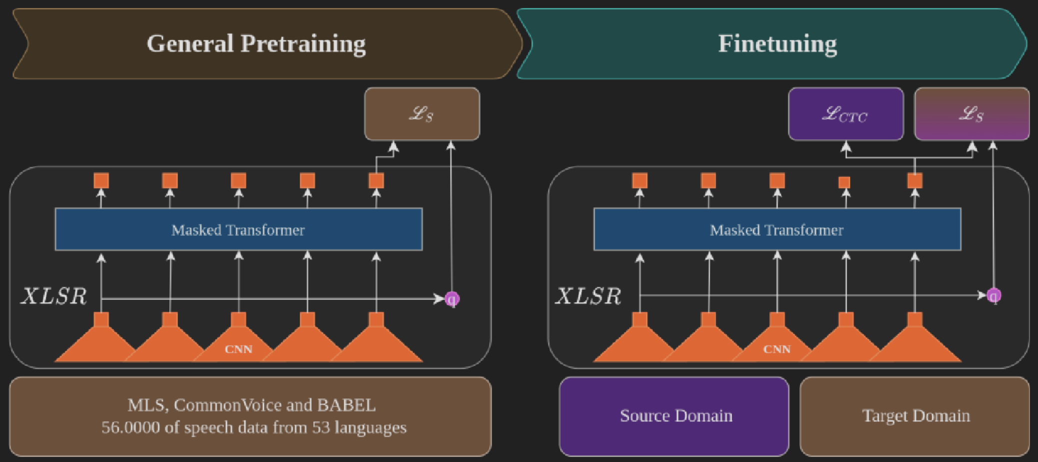
- Apply ideas from UDALM for acoustic adaptation of XLSR-53
    - Multi-domain instead of in-domain self-supervision
- Combine with simple language adaptation techniques
- Apply for cross-corpus speech recognition in Greek
- Demonstrate sample-efficiency
- New corpus: Hellenic Parliament recordings (120 hours)

G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

M2DS2: Mixed Multi-domain Self-Supervision

Method similar with UDALM

- No Continual Pretraining step
- Add source domain self-supervision in multitask loss
- Avoid mode-collapse of discrete code-vectors



$$L = L_{CTC}(\text{source samples}) + \alpha \cdot L_{SS}(\text{source samples}) + \beta \cdot L_{SS}(\text{target samples})$$

G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

Corpora

Dataset	Domain	Speakers	Train	Dev	Test	Total Duration
HParl	Public (political) speech	387	99:31:41	9:03:33	11:12:28	119:47:42
CV	Crowd-sourced speech	325	12:16:17	1:57:44	1:59:19	16:13:20
Logotypografia	News casts	125	51:58:45	9:08:35	8:59:22	70:06:42
Total	-	713	163:46:43	20:09:52	22:11:44	206:08:19

- 6 adaptation scenarios between Logotypografia (LG), Common Voice (CV) and HParl (HP)

> G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

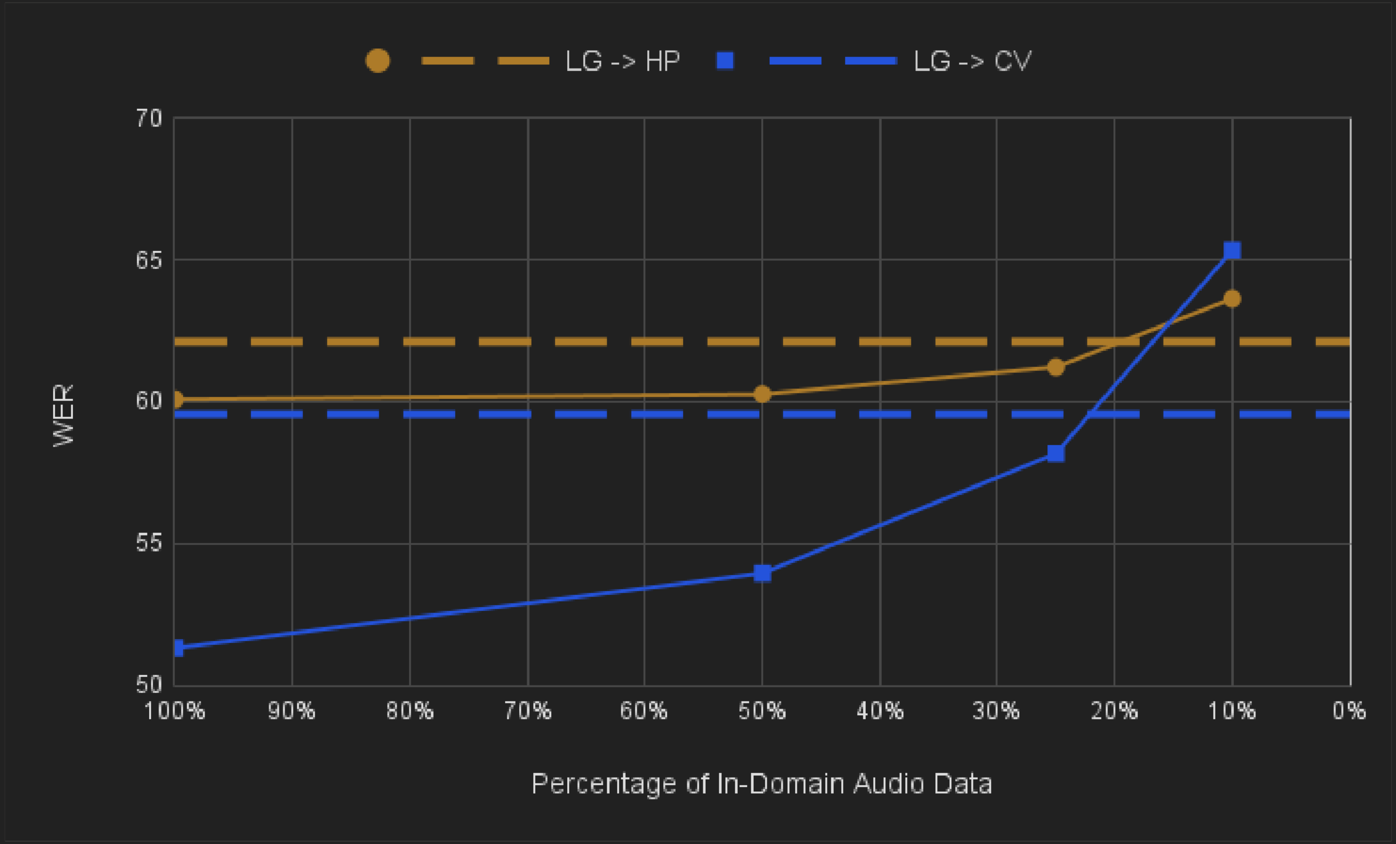
Results

Method Setting	SO (G) WER	CPT (G) WER	CPT (G) WRR	PSL (G) WER	PSL (G) WRR	M2DS2 (G) WER	M2DS2 (G) WRR	SO (LM) WER	CPT (LM) WER	CPT (LM) WRR	PSL (LM) WER	PSL (LM) WRR	M2DS2 (LM) WER	M2DS2 (LM) WRR
HP $\rightarrow$ CV	55.90	54.80	4.1	53.48	9.1	**52.95**	**11.1**	25.26	23.26	12.7	24.34	5.9	**18.35**	**43.9**
HP $\rightarrow$ LG	48.65	47.99	4.0	51.75	−18.6	**46.47**	**12.5**	30.34	33.88	−91.0	31.92	−40.6	**29.56**	**20.1**
LG $\rightarrow$ CV	59.57	60.81	−4.1	63.28	−12.3	**51.31**	**27.3**	25.96	29.10	−19.1	23.46	15.2	**17.30**	**52.7**
LG $\rightarrow$ HP	62.13	60.60	4.3	66.60	−12.4	**60.09**	**5.7**	31.48	31.54	−0.4	39.15	−48.4	**31.36**	**0.8**
CV $\rightarrow$ LG	69.55	68.98	1.5	68.29	3.4	**63.40**	**16.4**	50.80	47.61	13.1	42.53	34.0	**36.93**	**57.0**
CV $\rightarrow$ HP	70.72	71.79	−2.4	69.68	2.3	**68.70**	**4.5**	52.09	48.14	10.8	53.8	−4.7	**41.88**	**28.0**

- WER $\rightarrow$ Word Error Rate WRR $\rightarrow$ Relative adaptation improvement (%)
- G $\rightarrow$ Greedy decoding LM $\rightarrow$ Generic LM reweighting

G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

Sample Efficiency



G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

Combine with LM adaptation

- Biased LM:
  - Train N-gram LM on in-domain data
- Augmented LM:
  - Train N-gram LM on in-domain data
  - Use in-domain LM to filter lines with low perplexity from large corpus
  - Tran N-gram LM on augmented data

	Biased LM	Augmented LM
100%	11.22	12.84
50%	15.13	15.05
25%	20.84	16.64
10%	27.75	18.47
5%	33.04	19.31
Baseline (M2DS2 + Generic LM)	20.7	

G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

What we gain overall?

Method	#Audio (h)	#Tokens	LM	WER
SO (U)	-	-	N/A	59.57
M2DS2 (U)	3	-	N/A	57.31
M2DS2 (U)	12	-	N/A	51.31
SO (U)	-	-	Generic	25.96
SO (U)	-	38,632	Augmented	24.67
SO (U)	-	751,953	Augmented	20.46
M2DS2 (U)	3	-	Generic	20.7
M2DS2 (U)	12	-	Generic	17.3
M2DS2 (W)	3	38,632	Augmented	19.31
M2DS2 (W)	12	38,632	Augmented	16.29
M2DS2 (W)	3	751,953	Augmented	12.84
M2DS2 (W)	12	751,953	Augmented	10.61
Supervised	12	751,953	Generic	9.52
Supervised	12	751,953	Augmented	7.94

> G. Paraskevopoulos, T. Kouzelis, G. Rouvalis, A. Katsamanis, V. Katsouros, A. Potamianos, Sample-Efficient Unsupervised Domain Adaptation of Speech Recognition Systems: A case study for Modern Greek, under revision IEEE/ACM TASL-P

Conclusions

Under some closed world assumptions (known application domain) we can improve performance with

1. few in-domain data without annotations
2. in-domain self-supervision to avoid catastrophic forgetting

The techniques presented are:

1. simple to implement
2. evaluated in diverse settings
3. domain-invariant (no domain expertise needed)