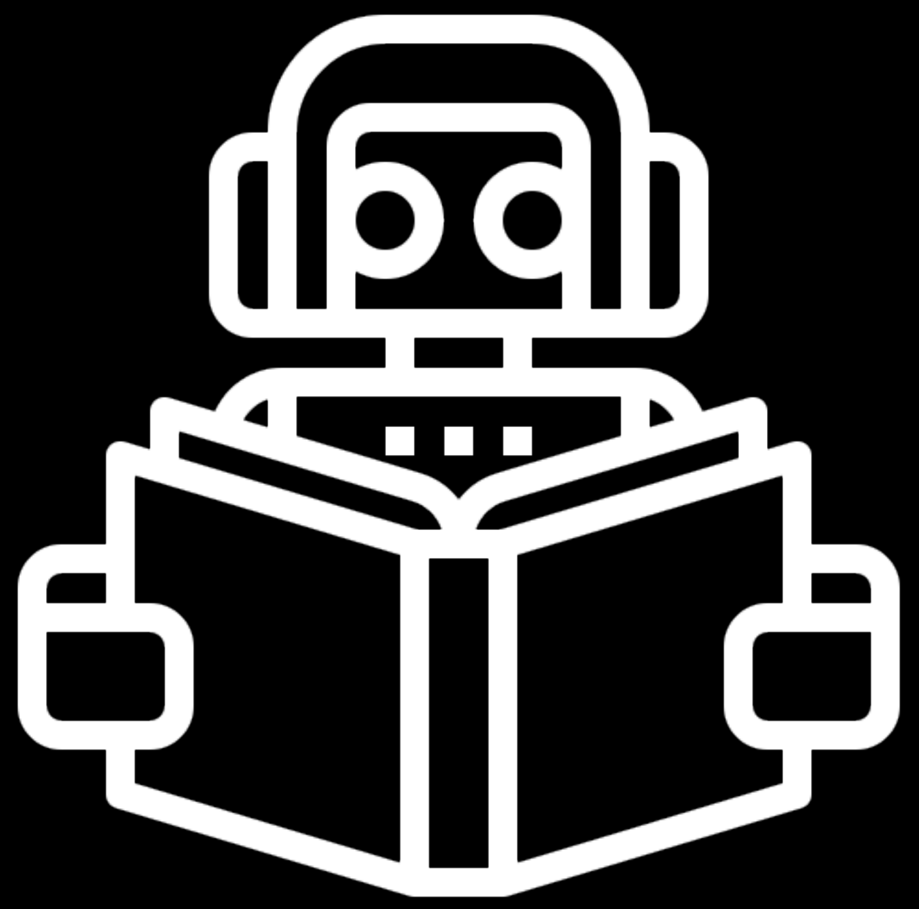
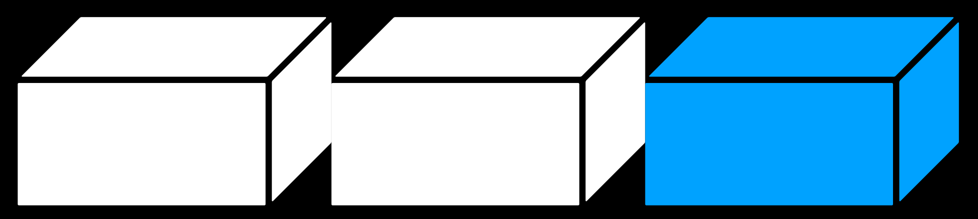
Structure in Language Learning Systems

Lukas Galke, 2022-10-20

Three Aspects of Structure
Outline

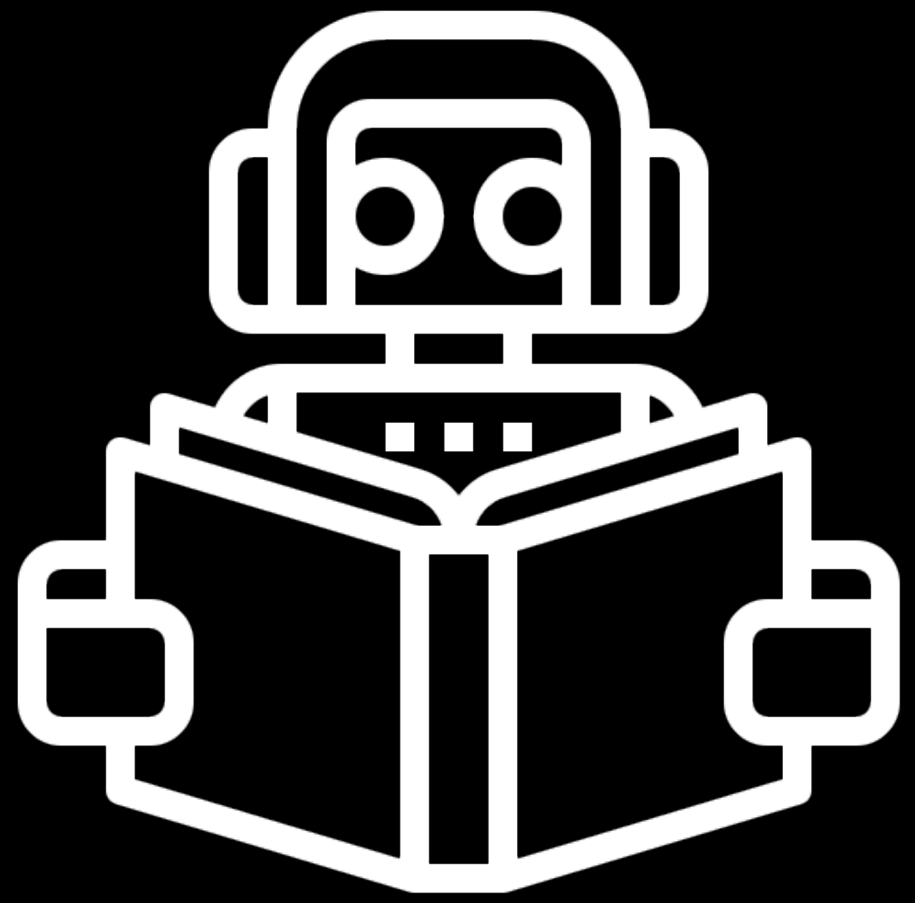
- Induced structure by the models
  - Bag-of-words vs. sequence vs. graph for text classification
- External structure in side information
  - Lifelong learning on graphs
- Internal/compositional structure of language
  - Does structure help neural nets?





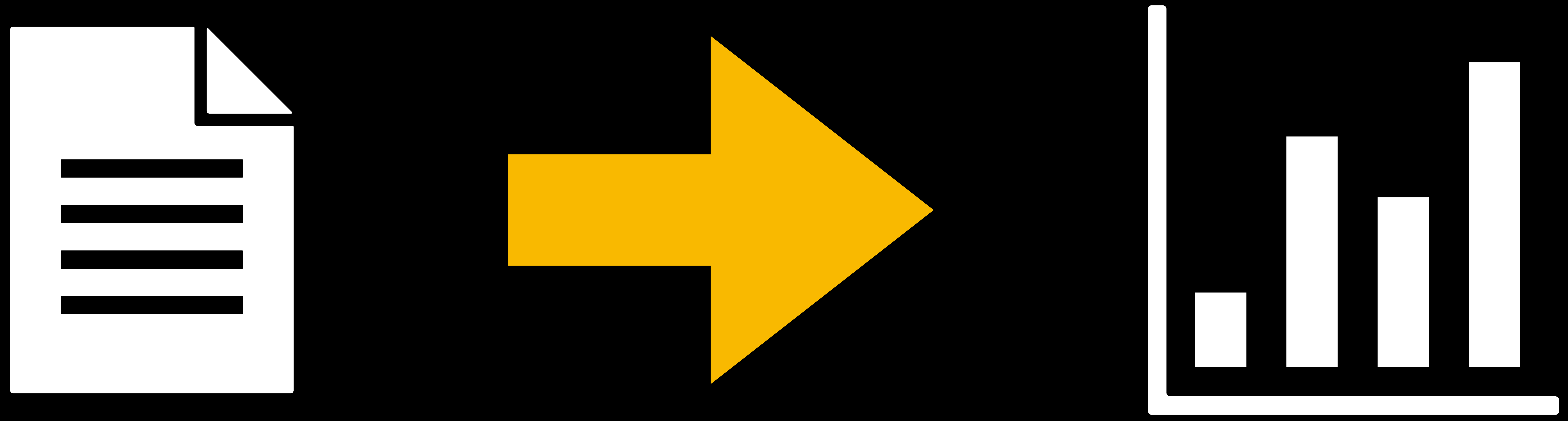
Three Aspects of Structure

Outline



- **Induced structure**
- External structure
- Internal/compositional structure

Text Classification



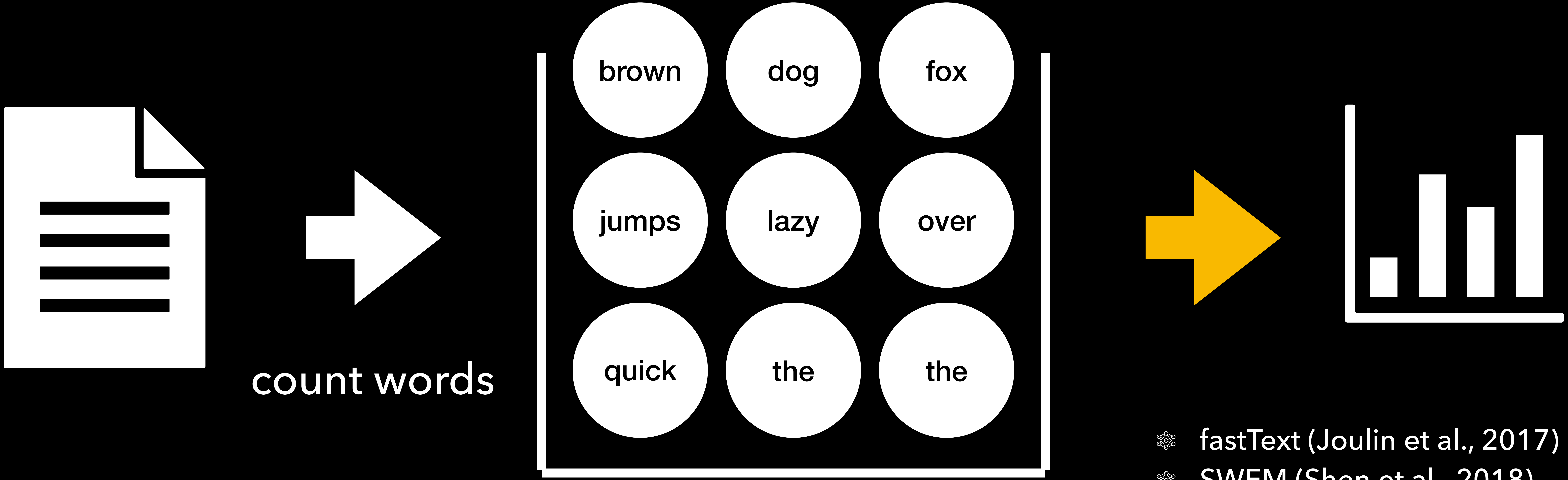
Text

Class label

Bag-of-words model?
Graph-based model?
Sequence model?

Galke & Scherp (ACL 2022)

Bag-of-Words Model Family

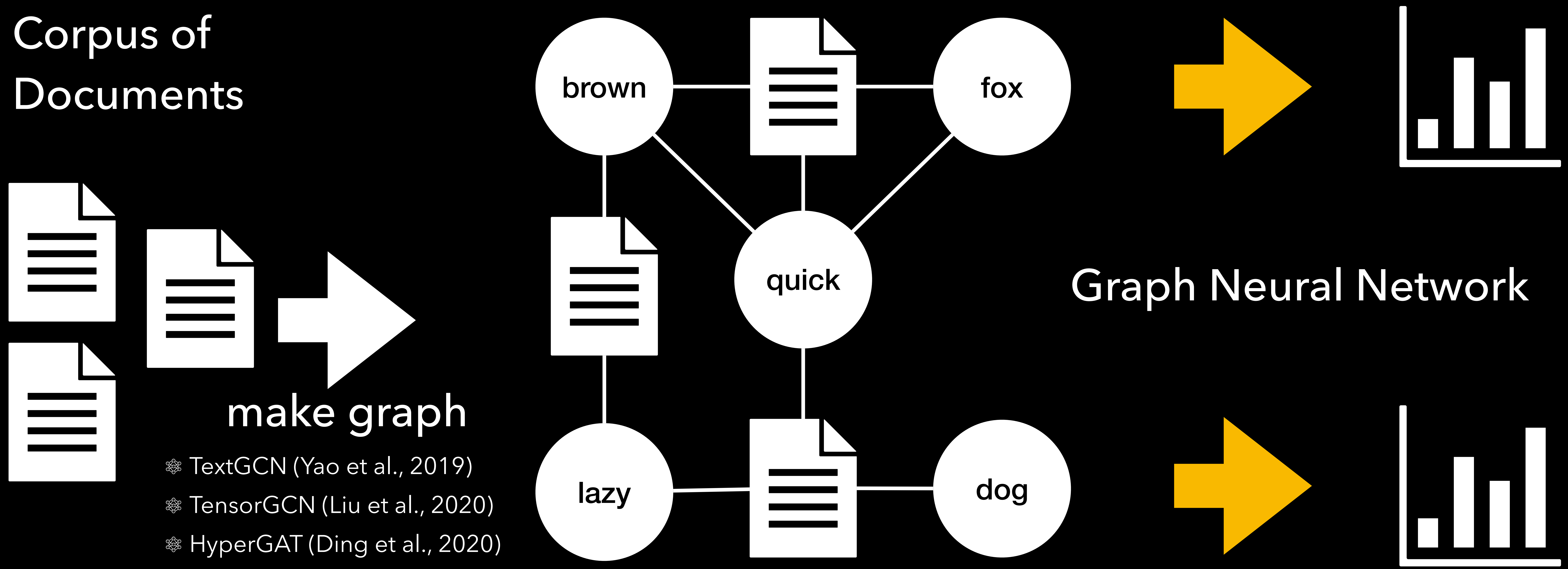


- fastText (Joulin et al., 2017)
- SWEM (Shen et al., 2018)
- **Multilayer perceptron**

Galke & Scherp (ACL 2022)

Graph-Based Model Family

Corpus of Documents

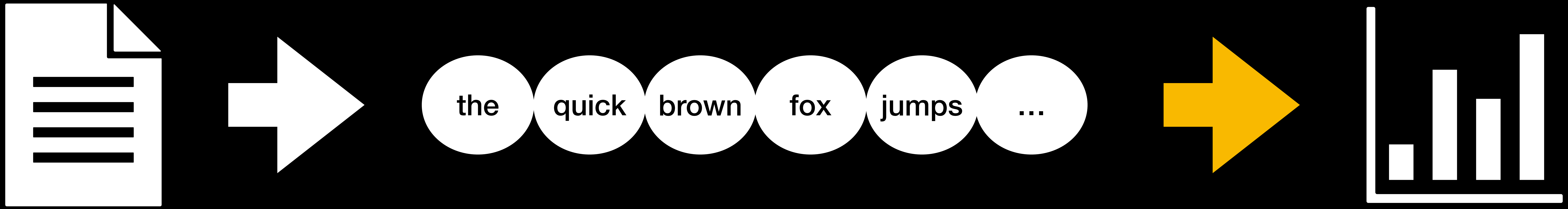


- TextGCN (Yao et al., 2019)
- TensorGCN (Liu et al., 2020)
- HyperGAT (Ding et al., 2020)
- DADGNN (Liu et al., 2021)
- HeteGCN (Ragesh et al., 2021)

Synthetic Graph

Galke & Scherp (ACL 2022)

Sequence Model Family



- Transformer (BERT, DistilBERT)
- RNN (LSTM, GRU)
- 1D-CNN

Galke & Scherp (ACL 2022)

Conceptual Considerations

Bag-of-words models

- Not sensitive to word order

Transformer-based sequence models

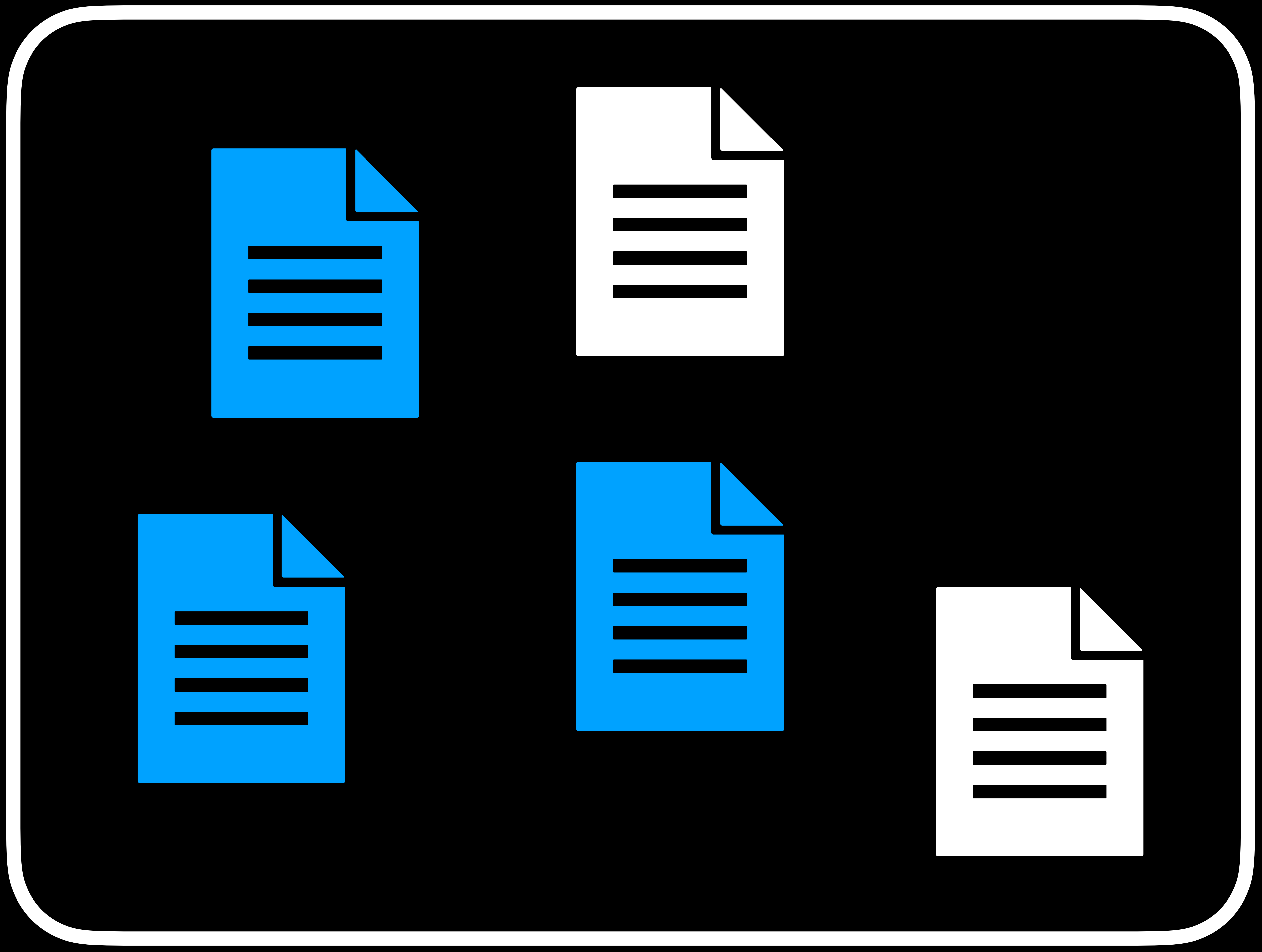
- Quadratic in sequence length
- Maximum sequence length

Graph-based models

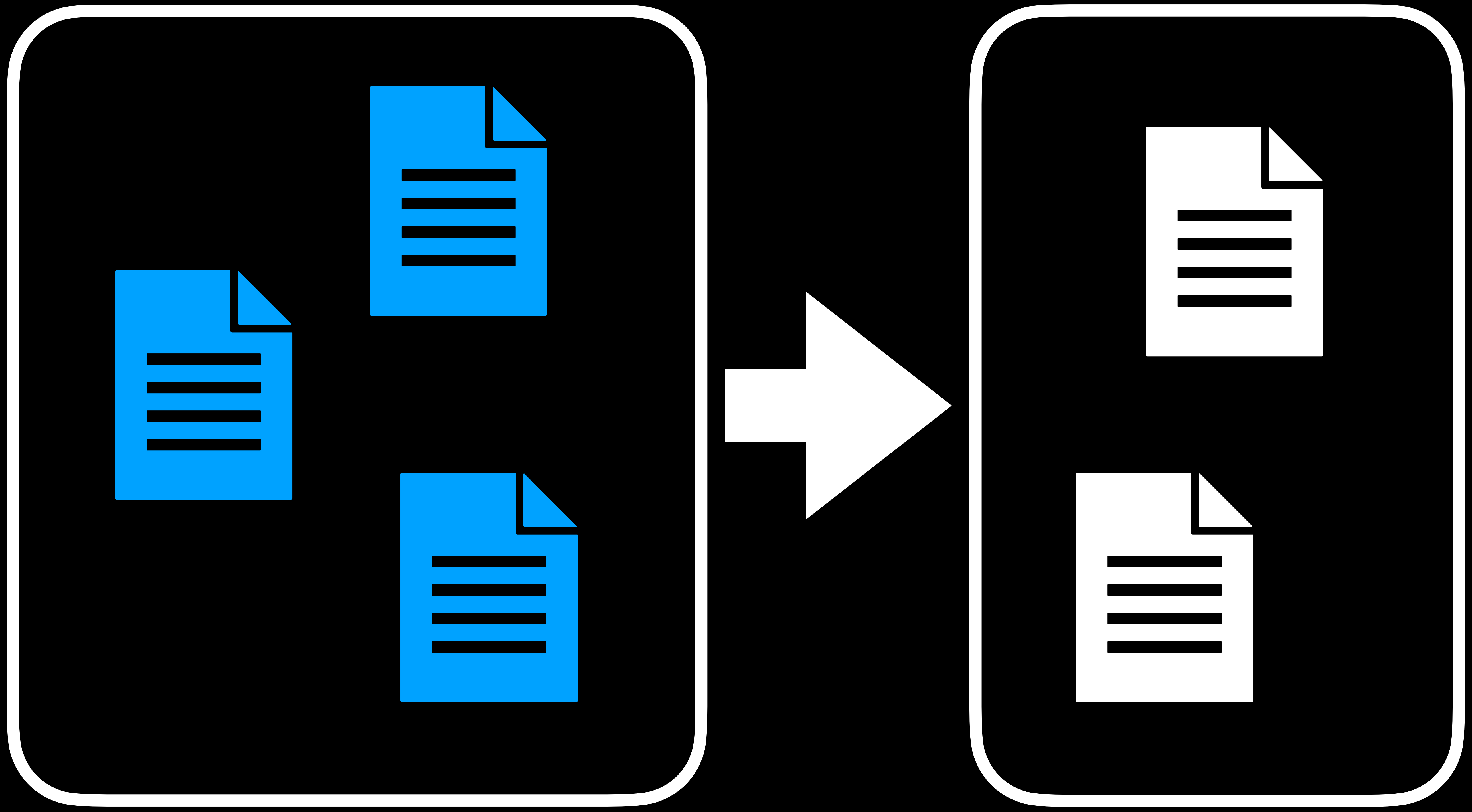
- Set up a large graph (number of documents plus vocabulary size)
- Graph neural networks are difficult to scale
- Inductive learning not trivial
- Not sensitive to word order

Galke & Scherp (ACL 2022)

Transductive and Inductive Settings



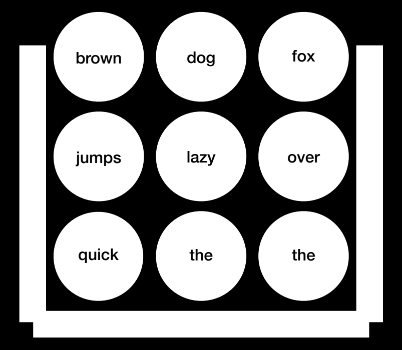
Transductive: All examples visible during training



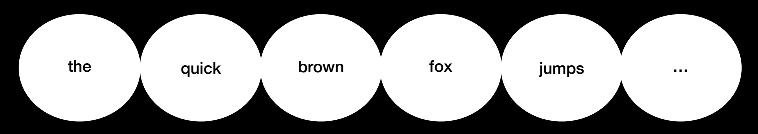
Inductive: Test examples not visible during training

Galke & Scherp (ACL 2022)

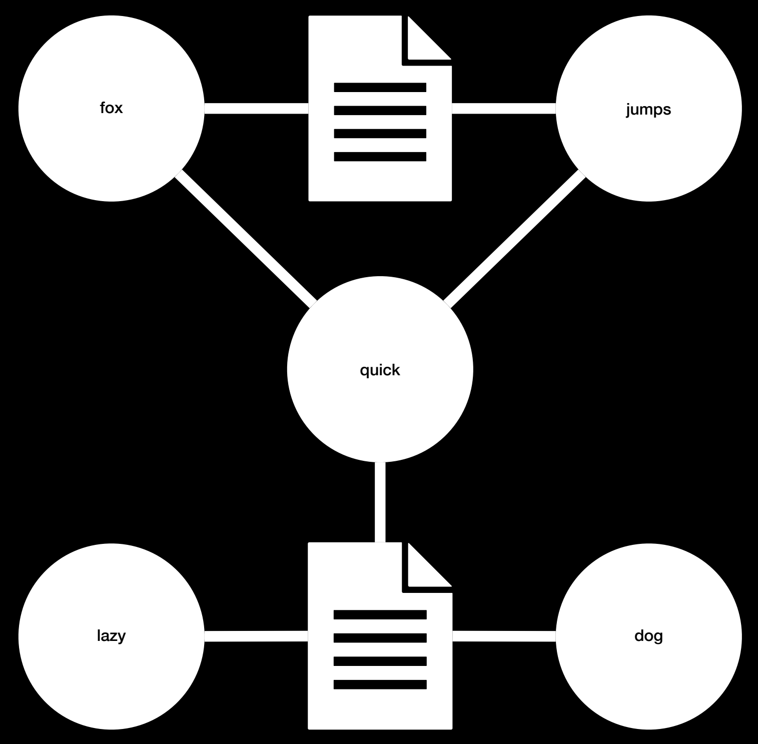
Three Model Types for Text Classification



- Bag-of-Words models: Classic methods, CBOW methods



- Sequence models: Pretrained Language Models



- Recently: Graph-based models via synthetic text-graphs
  - TextGCN (AAAI 2019)
  - TensorGCN (AAAI 2020)
  - HyperGAT (EMNLP 2020)
  - DADGNN (EMNLP 2021)
  - HeteGCN (WSDM 2021)

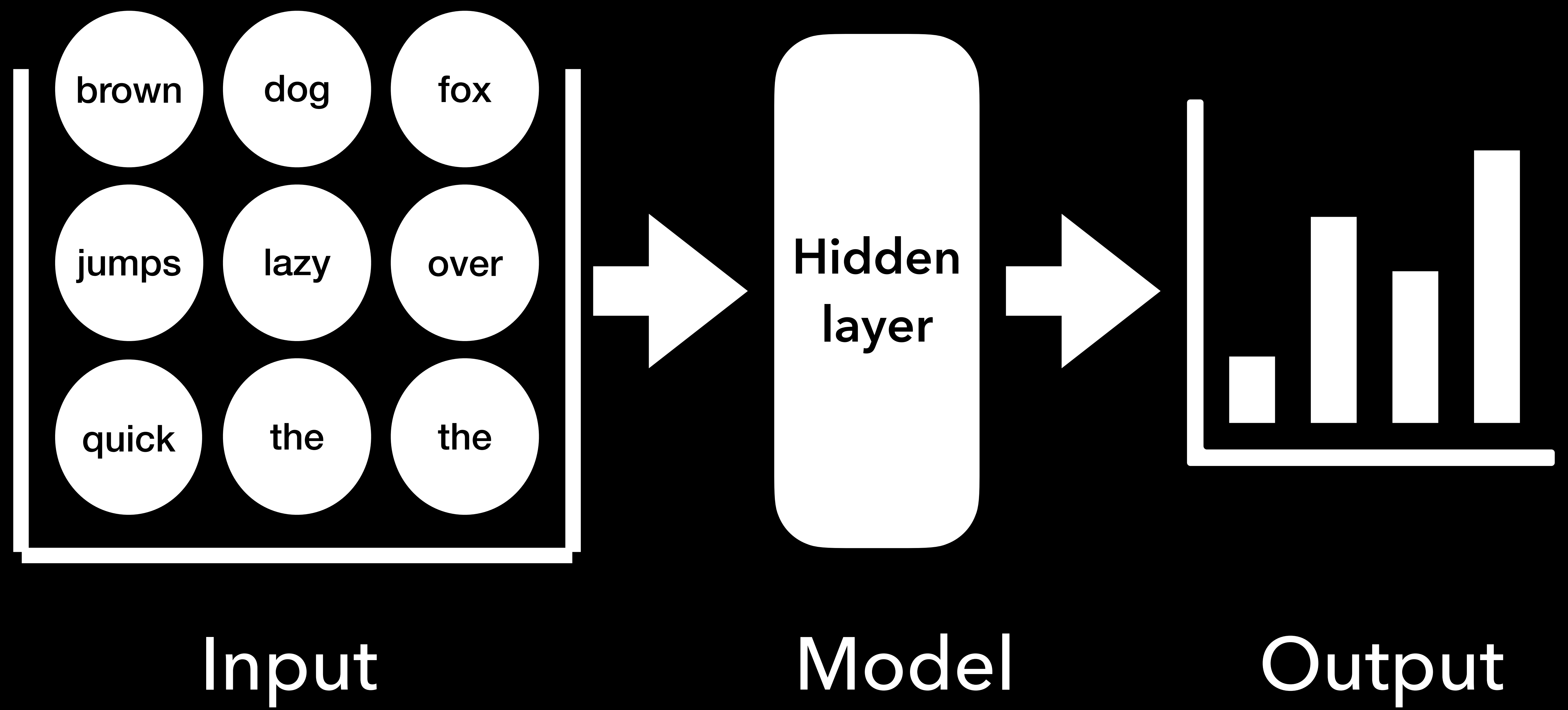
So what's best?

Galke & Scherp (ACL 2022)

Wide Multilayer Perceptron

Revisiting a decades old technique

- Bag-of-Words input repr.
- single **wide** hidden layer
- No pretrained word embeddings
- ReLU activation, high dropout, long training



Galke & Scherp (ACL 2022)

Datasets

	#documents	#classes	Avg. length ± SD
20ng	18.846	20	551 ± 2,047
R8	7.674	8	119 ± 128
R52	9.100	52	126 ± 133
ohsumed	7.400	23	285 ± 123
Movie Reviews	10.662	2	25 ± 11

Topical Classification: 20ng, R8, R52, ohsumed

Sentiment Analysis: Movie Reviews

Very long texts: 20ng

Long texts: R8, R52, ohsumed

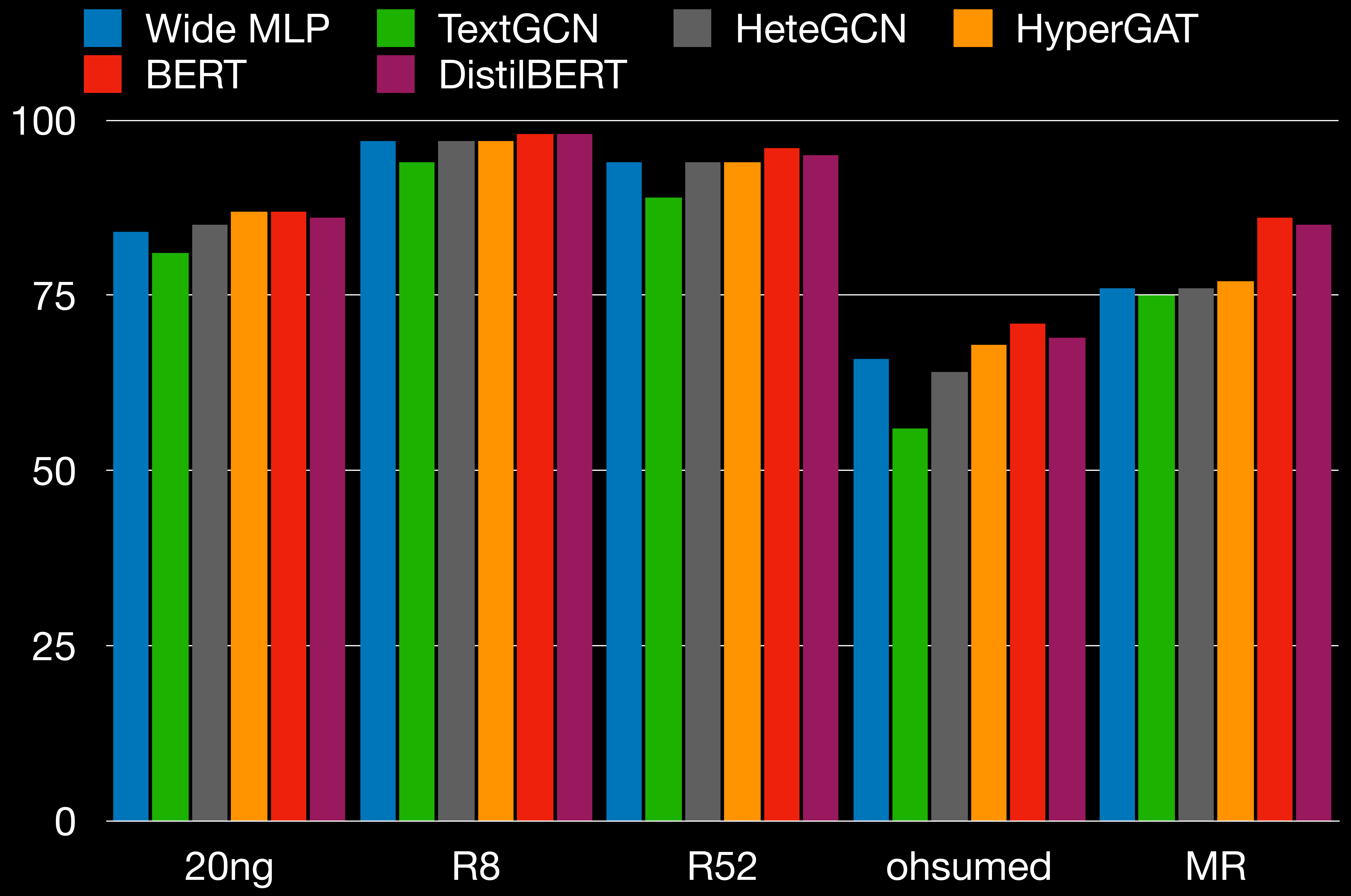
Short texts: Movie Reviews

Galke & Scherp (ACL 2022)

Results

Inductive Setting

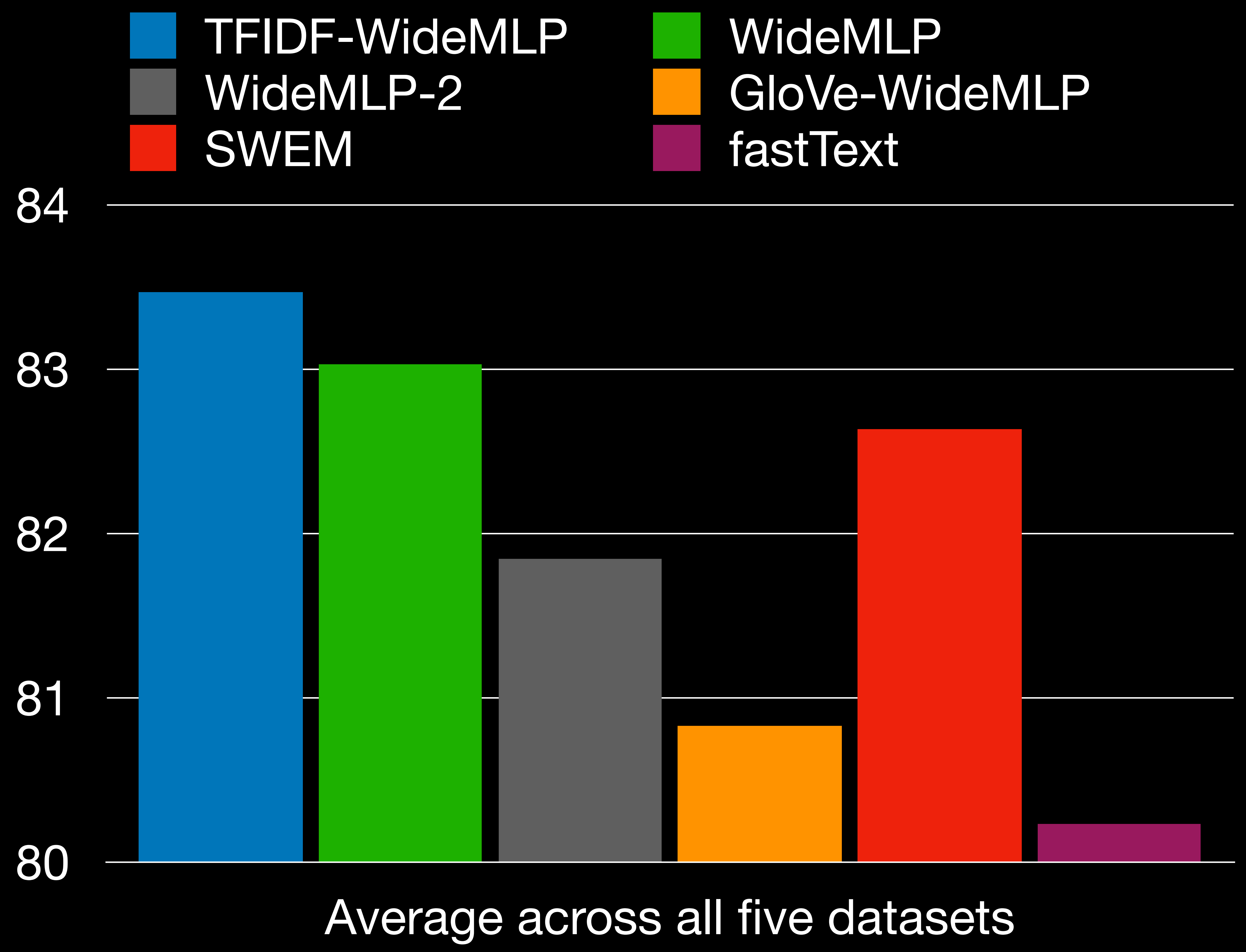
- WideMLP better than TextGCN, on par with HeteGCN, HyperGAT
- BERT best, closely followed by DistilBERT



Galke & Scherp (ACL 2022)

Results of MLP Variations

- Wide hidden layer better than pretrained word embeddings (GloVe+MLP, SWEM, fastText)
- Single hidden layer better than two hidden layers
- TF-IDF weighting is better than unweighted average

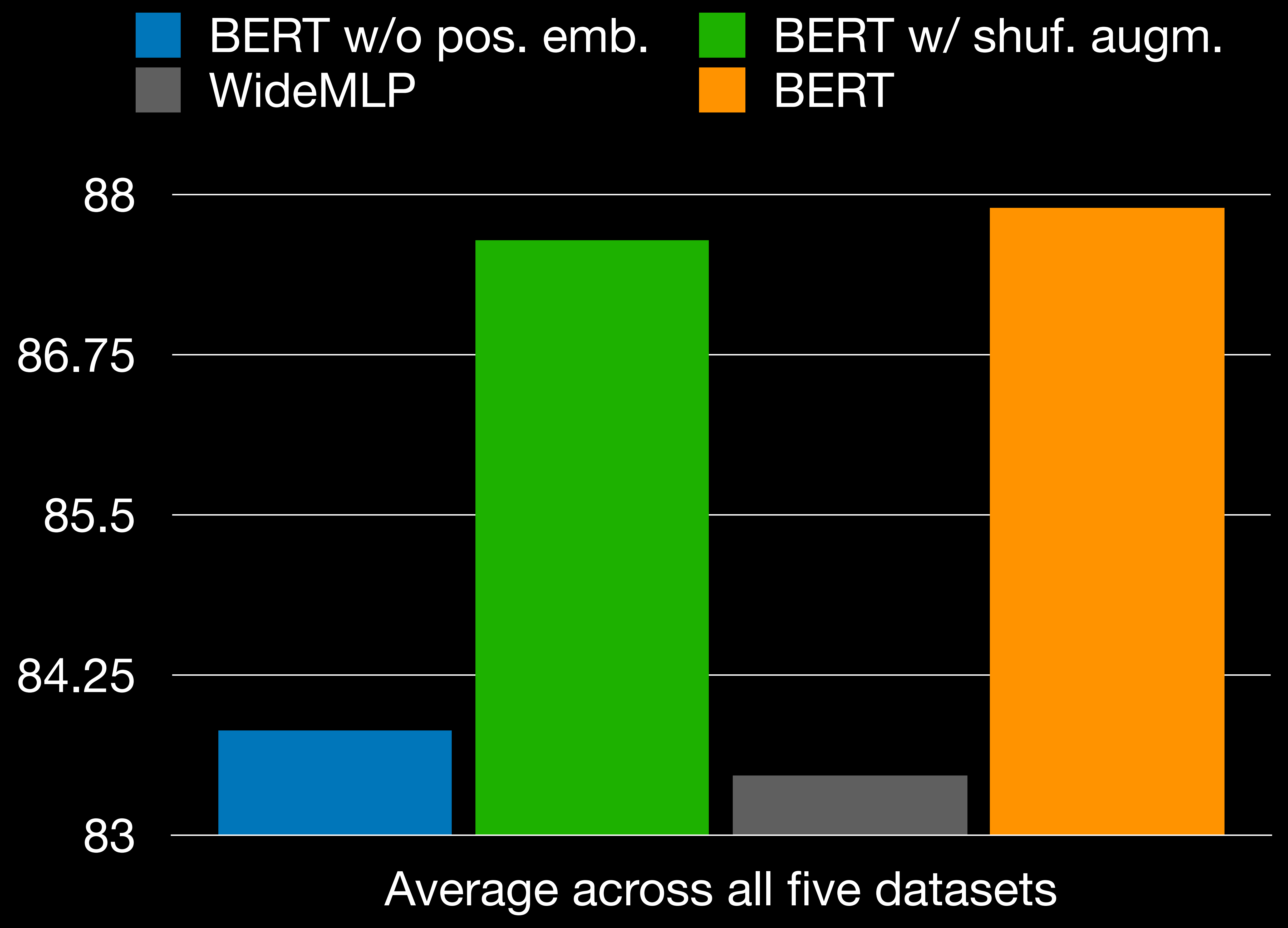


Galke & Scherp (ACL 2022)

Importance of Word Order in BERT

- Removing position embeddings in BERT leads to a notable decrease

- Augmenting the training data with shuffled sequences does not help



Galke & Scherp (ACL 2022)

Parameter Count & Training Time

- Bag-of-words MLP has relatively few parameters
- First layer can be implemented efficiently as an embedding bag
- MLP is fast

	Number of parameters	Runtime/epoch (20ng)
Wide MLP	**31M**	**5s**
DistilBERT	66M	48s
BERT	110M	90s

Galke & Scherp (ACL 2022)

Induced Structure — Summary

- A **wide** MLP on a bag-of-words is a surprisingly strong and fast text classifier
- Pretrained language models best
- Text-graphs seem **not** necessary

Code: GitHub.com/lgalke/text-clf-baselines

See also: Extension with Multi-Label Classification: arxiv.org/abs/2204.03954

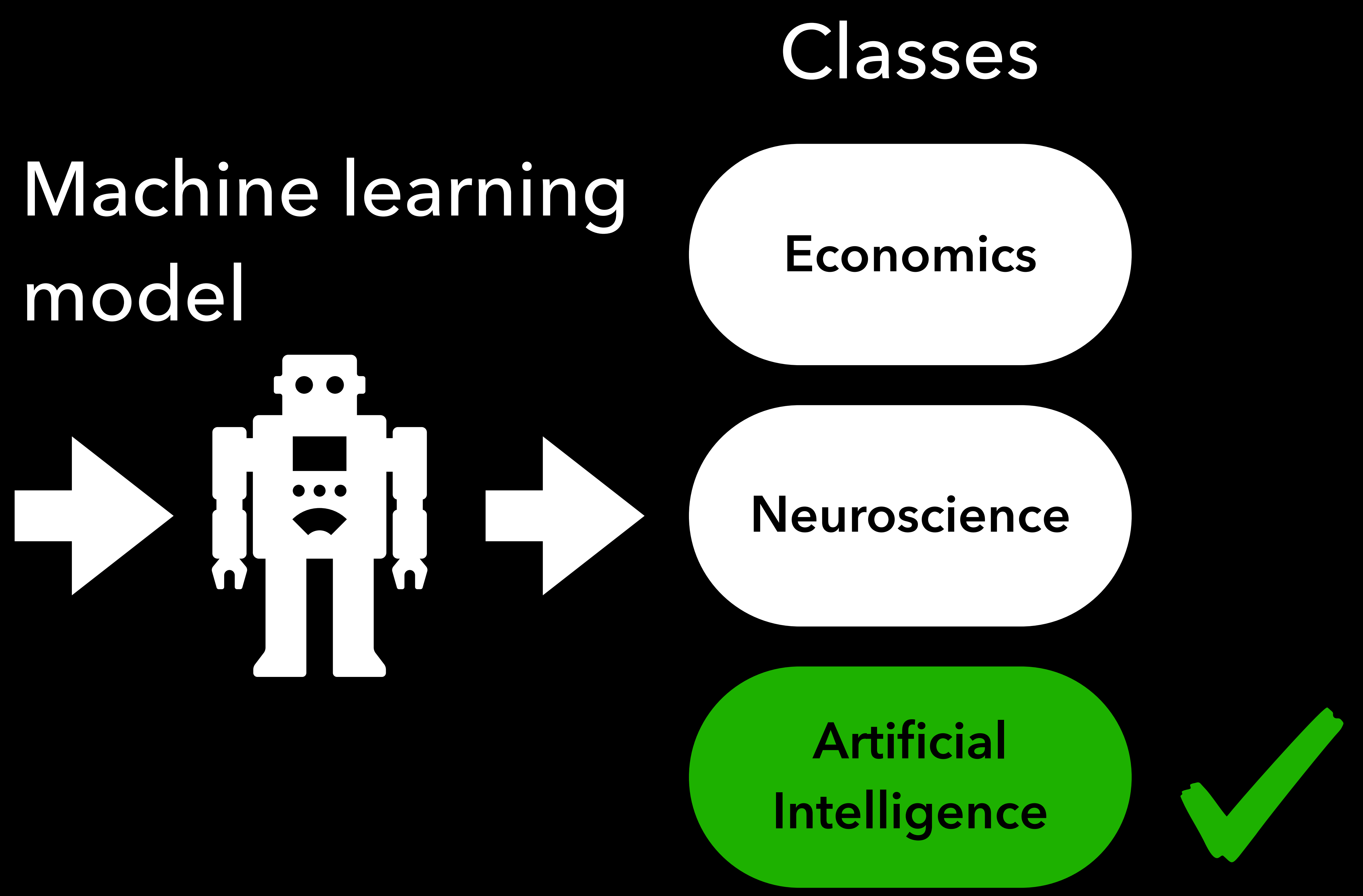
Three Aspects of Structure
Outline

- Induced structure
- **External structure**
- Internal/compositional structure

Motivational Example

Classifying research papers into topics

Paper title: **Deep learning**



Previous work: On research papers, *only title* just as good as *title* + *full-text*.

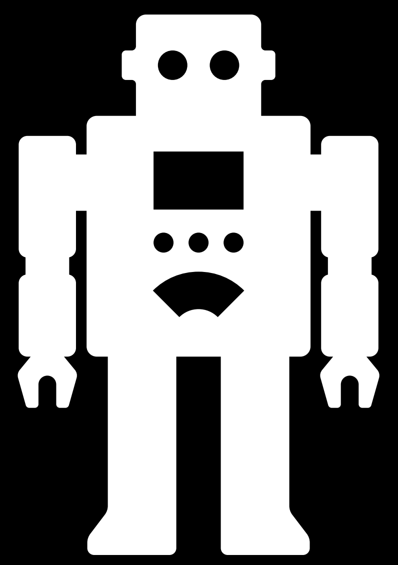
Galke et al. (K-CAP 2017); Mai, Galke, & Scherp (JCDL 2018)

Motivational Example
Some titles are hard to classify

Classes

Machine learning model

Paper title: *Optimal Brain Damage*



Economics

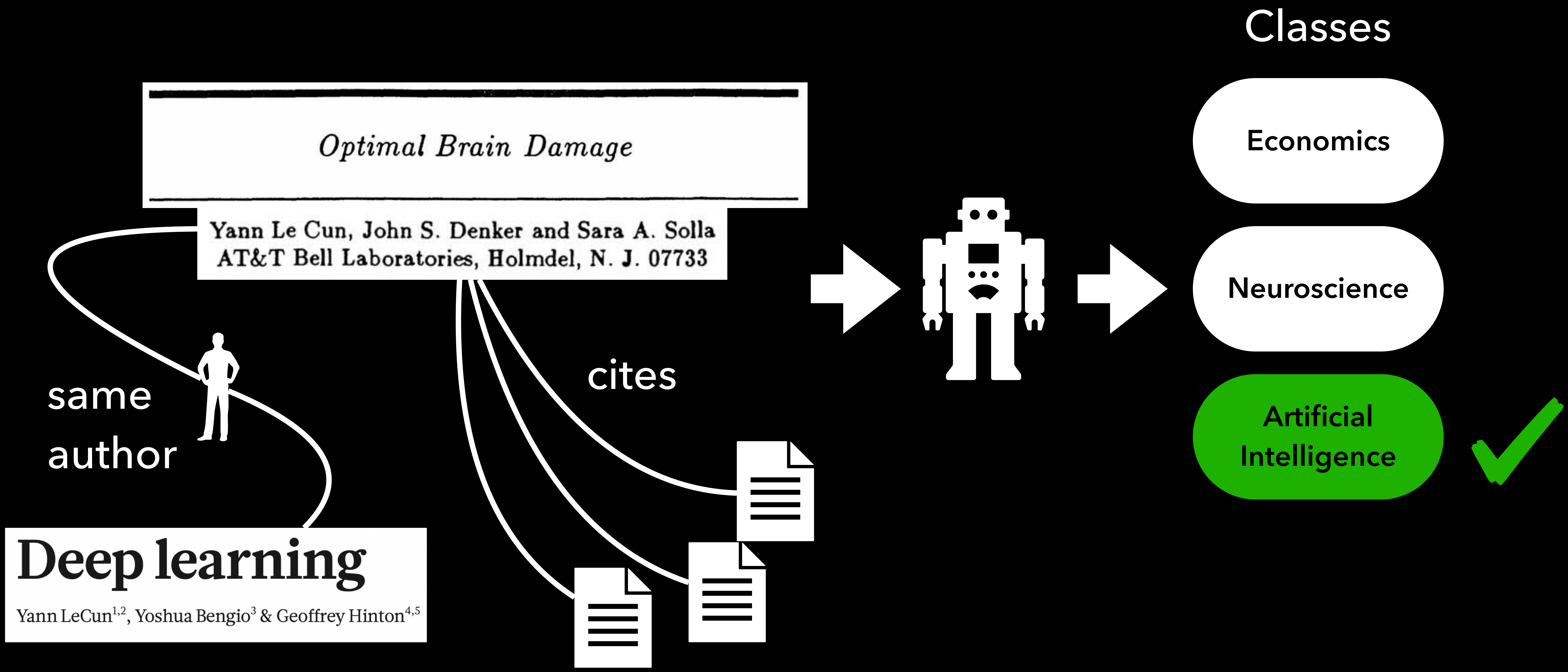
Neuroscience

Artificial Intelligence

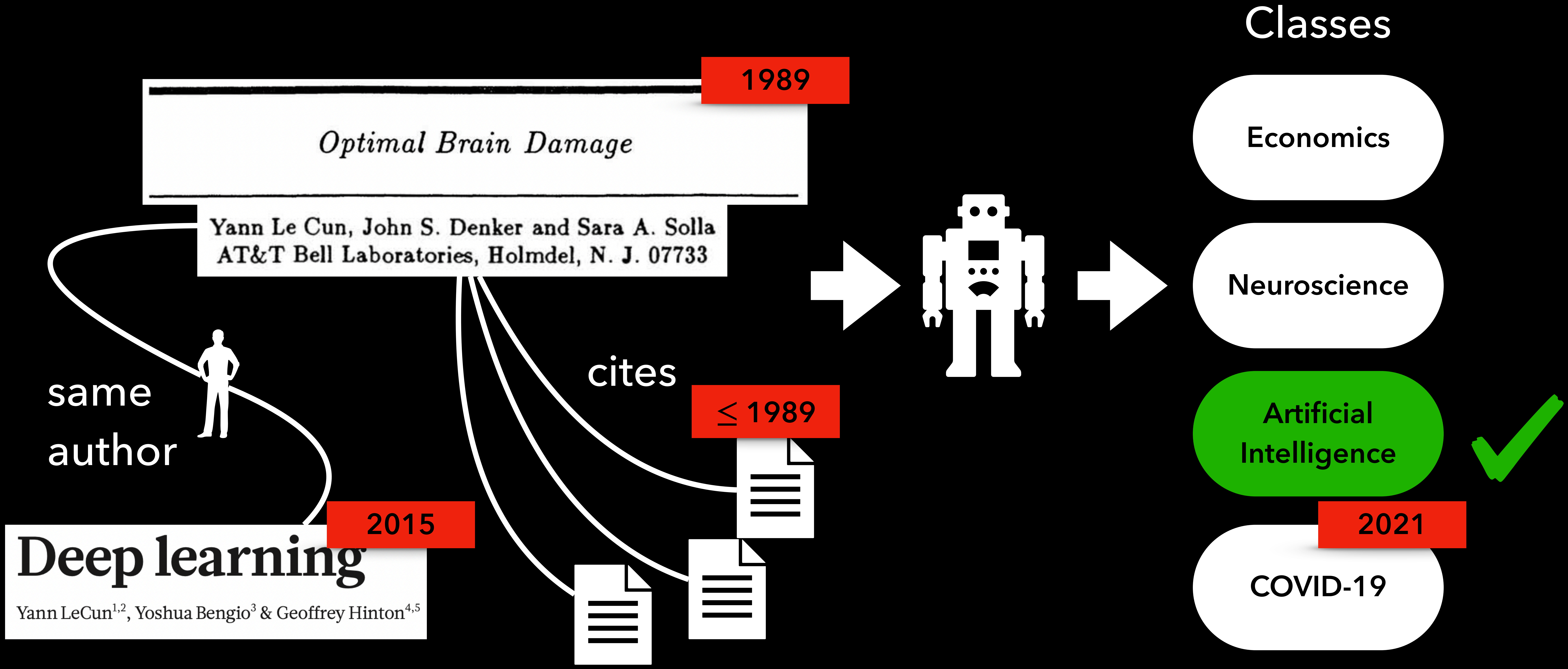
?

Motivational Example

Graph data to the rescue?



Motivational Example
But the world is dynamic



Even large language models fail without context

Paper title: *Optimal Brain Damage*

What is the topic of the research paper with the title "Optimal brain damage"?

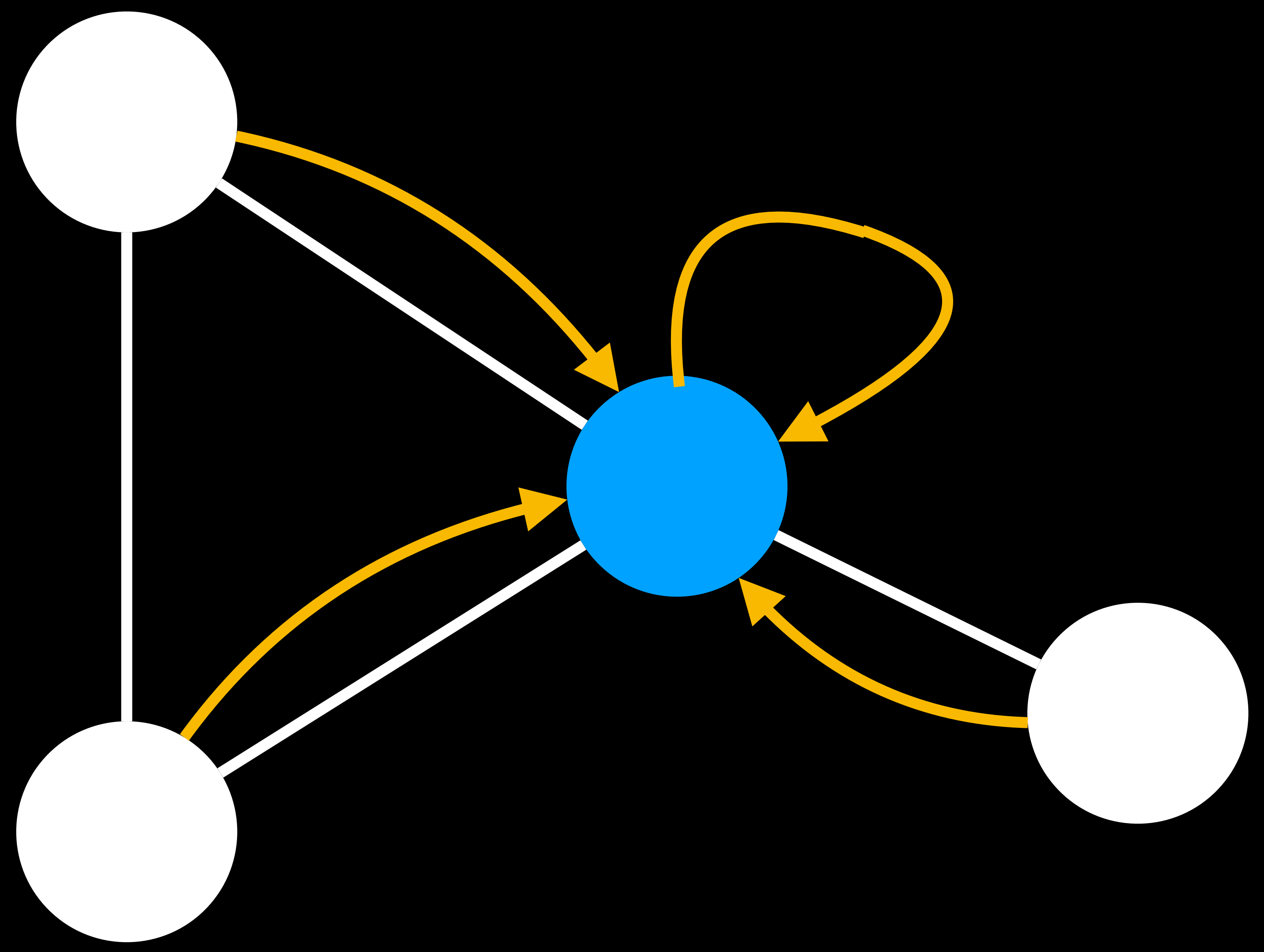
The topic of the research paper with the title "Optimal brain damage" is the study of how much damage to the brain is optimal for a person's health and well-being.

GPT-3/text-davinci-002

Wrong!

Graph Neural Networks

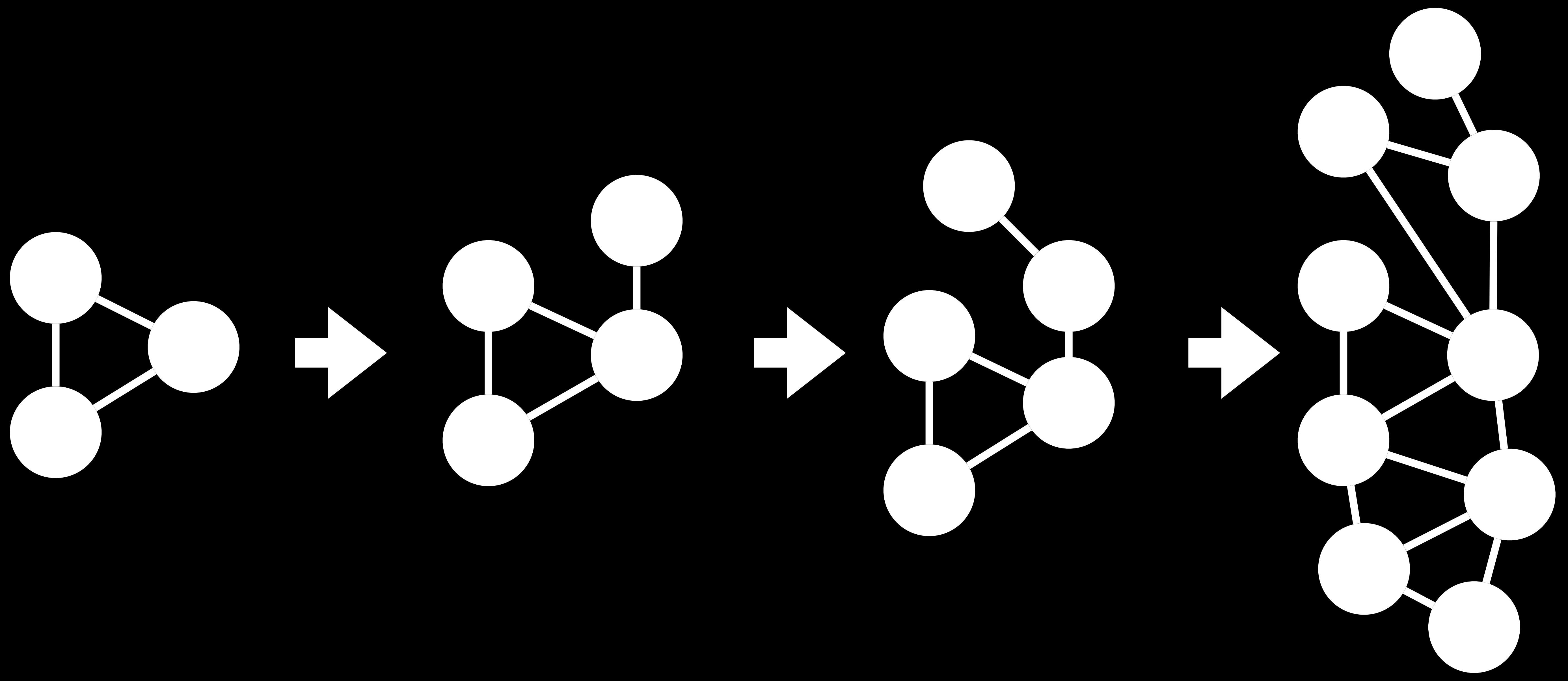
- Neighborhood aggregation
- Nonlinear transformation
- For each node simultaneously
- Stack multiple layers



$$\mathbf{h_i} = \sigma\left(\alpha_{ii}\mathbf{W}^{(\text{self})}\mathbf{x_i} + \sum_{j \in \mathcal{N}(i)} \alpha_{ij}\mathbf{W}^{(\text{neigh})}\mathbf{x_j}\right)$$

Evolving Graphs

- Real-world graphs evolve over time
  - Citation Graphs
  - Collaboration Graphs
  - Social Graphs
  - …

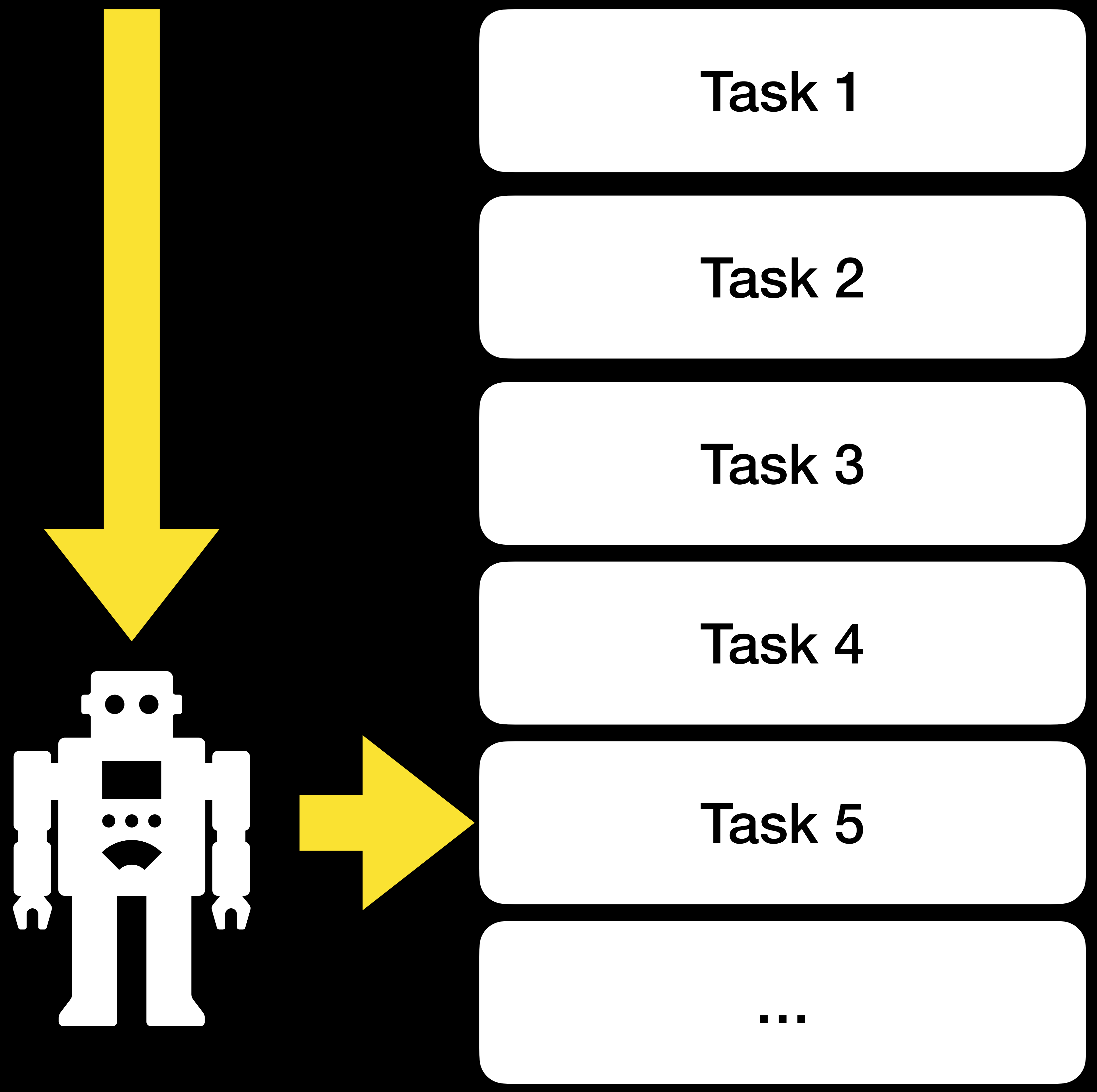


How can we adapt machine learning models to new graph data?

Galke et al. (IJCNN 2021)

Lifelong Learning

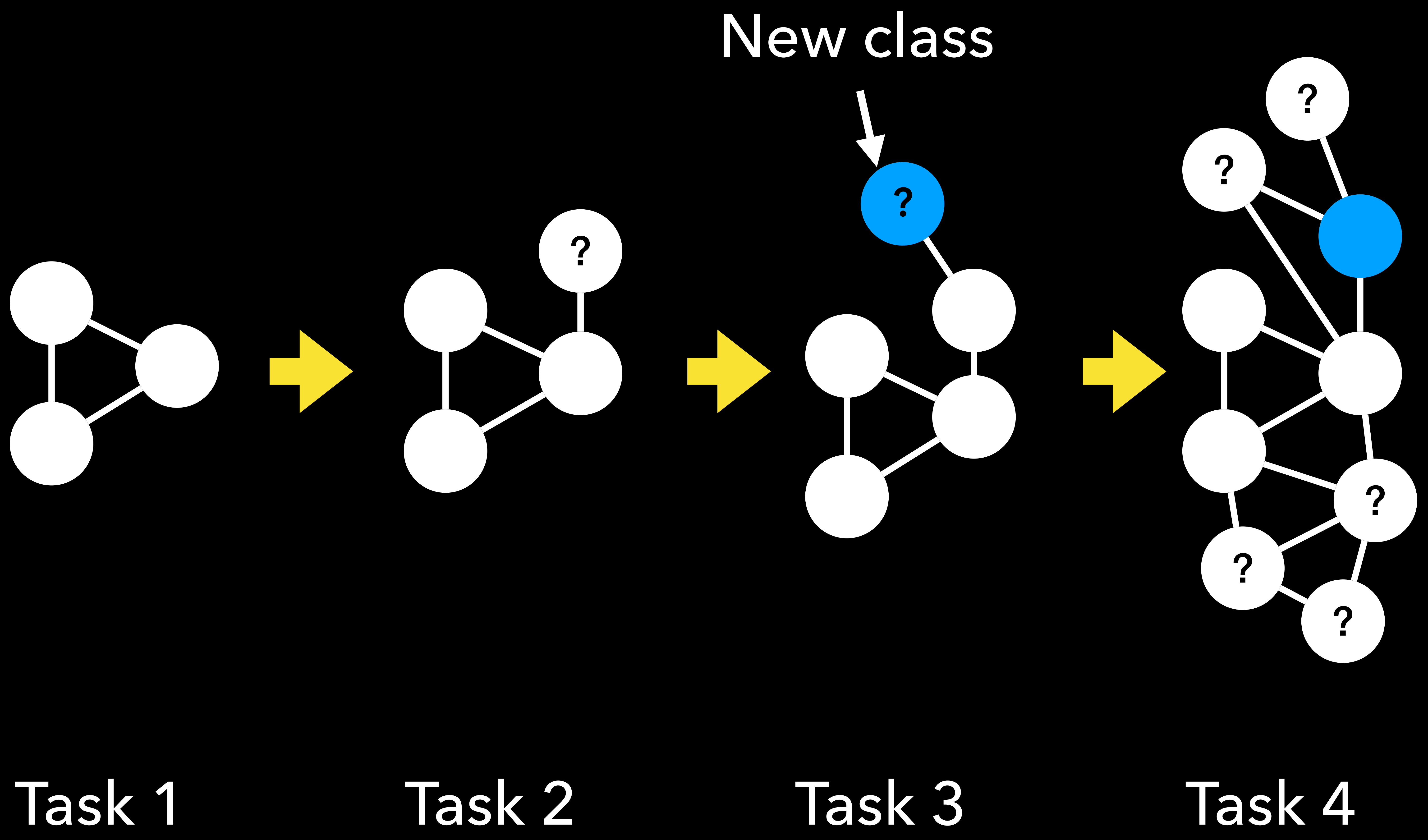
- Same model has to perform sequence of tasks
- Can make use of knowledge acquired in previous tasks



Thrun (1998)

Lifelong Learning on Evolving Graphs

- Train new model or adapt previous model?
- How much past data is needed?
- Can we detect when a new class appears?

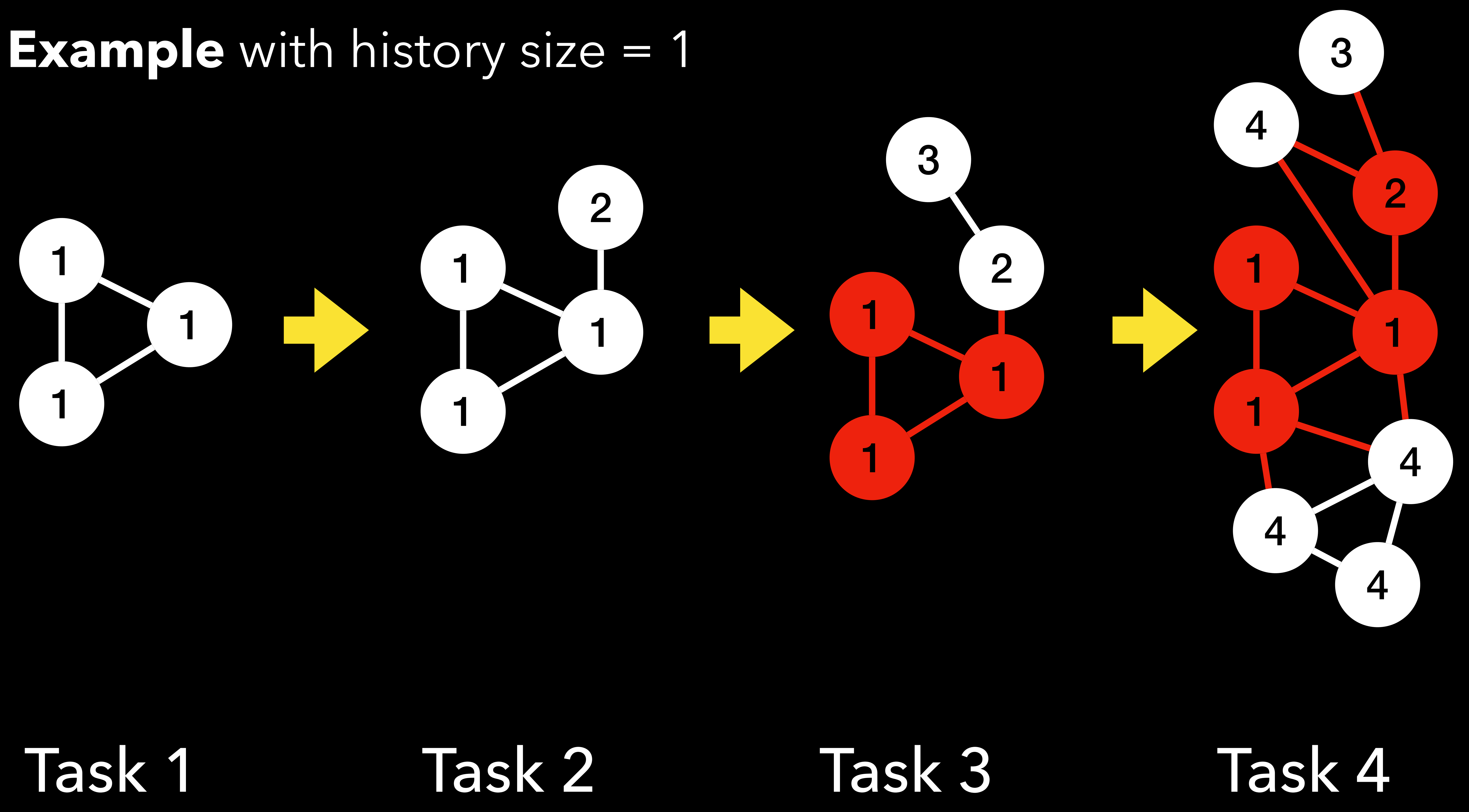


Galke et al. (IJCNN 2021)

Approach

- Incremental training with a sliding window (history size)
- Method to determine comparable history sizes from data
- Method for unseen class detection

Example with history size = 1



Galke et al. (IJCNN 2021)

New Datasets

for lifelong learning on graphs

Papers

Title

Publ. Year

Venue

		#nodes	#edges	#features	#tasks	#classes
Citation graphs	**DBLP-Easy**	45k	122k	2k	**12**	12 (**4 new**)
	DBLP-Hard	199k	644k	4k	**12**	73 (**23 new**)
Co-authorship graph	**PharmaBio**	68k	2.1M	5k	**18**	7

Galke et al. (IJCNN 2021)

Evaluating Multiple Tasks

- **Average Accuracy**
$$\frac{1}{T}\sum_{t\in 1,\ldots,T} \mathrm{acc}_t\left(f^{(t)}\right)$$

- **Forward Transfer**
$$\frac{1}{T-1}\sum_{t\in 2,\ldots,T} \mathrm{acc}_t\left(f^{(t)}_{\text{warm}}\right) - \mathrm{acc}_t\left(f^{(t)}_{\text{cold}}\right)$$

Reuse prev. model

Random reinit.

Incremental Training with Limited History

- Warm restarts (reuse prev. model) allows for smaller history sizes
- Medium history sizes good enough: 50% coverage (history size 3) → ~95% accuracy cmp. to full graph
- GNNs better than pure MLP → graph data helps

Accuracy / Forward Transfer on Task Sequence (DBLP-Hard)

DBLP-Hard	GraphSAGE	MLP
Hist. Size 1	**40.0 / +5.9**	38.3 / +7.4
Hist. Size 3	45.1 / +0.8	38.9 / +5.6
Hist. Size 6	46.7 / +0.2	38.3 / -0.7
Full Graph	47.1 / +0.3	36.7 / -1.1

Galke et al. (IJCNN 2021)

Unseen Class Detection

- Extension of Deep Open Classification (DOC, Shu et al., EMNLP 2017) to graphs and graph neural nets
  - Unsupervised, works by thresholding outputs
- Crucial to account for class imbalance (GDOC) by weighting the loss function

F1-Macro with Extra "Unseen Class" (DBLP-Hard)

DBLP-Hard	DOC (baseline)	GDOC (ours)
Hist. Size 1	1 %	**13 %**
Hist. Size 3	2 %	**15 %**
Hist. Size 6	5 %	**16 %**
Full Graph	8 %	**16 %**

Galke et al. (under review)

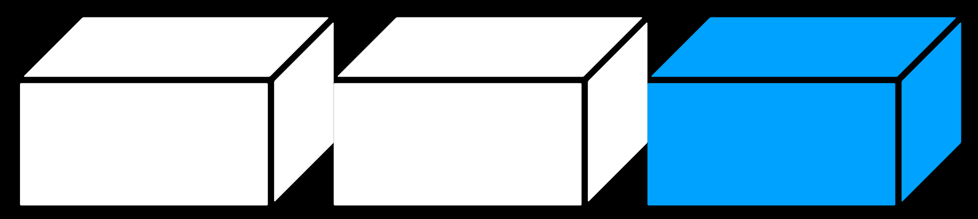
External Structure — Summary

- Graph neural networks can exploit external structure
- Evolution of graph data can be tackled with incremental training
- Parameter reuse is helpful for small history sizes
- Method to derive comparable history sizes across datasets
- In our datasets, small history sizes tend to be good enough

Code: GitHub.com/lgalke/lifelong-learning

Three Aspects of Structure
Outline

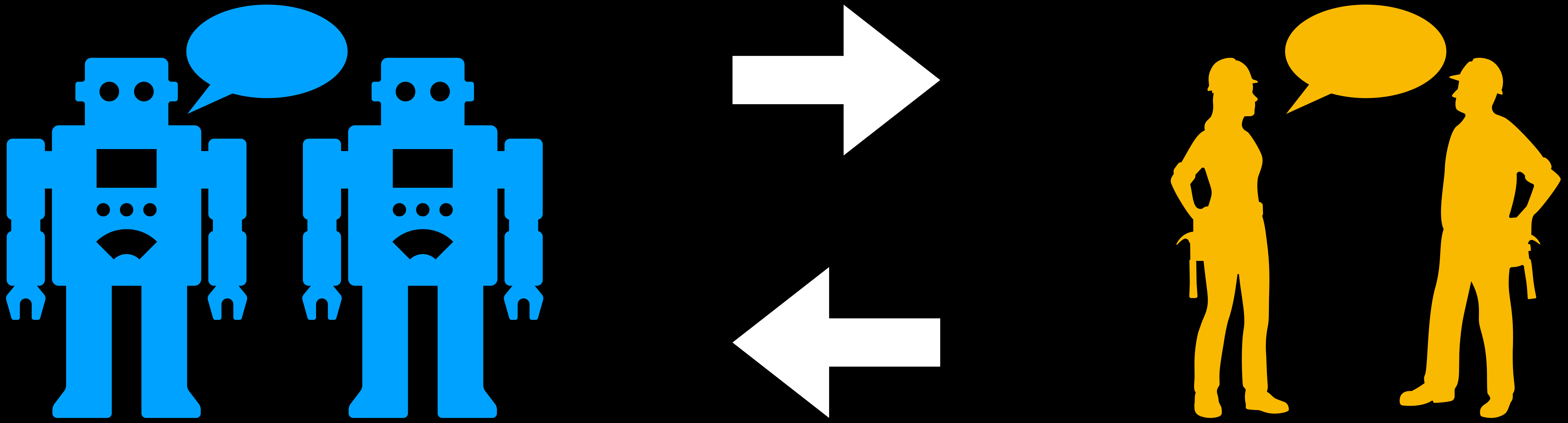
- Induced structure
- External structure
- **Internal/compositional structure (current project)**



Work in progress, feedback welcome!

High-level Aim

Learn more about human language?



Improved machine learning models?

Approach

Replicate lab experiments with neural network agents

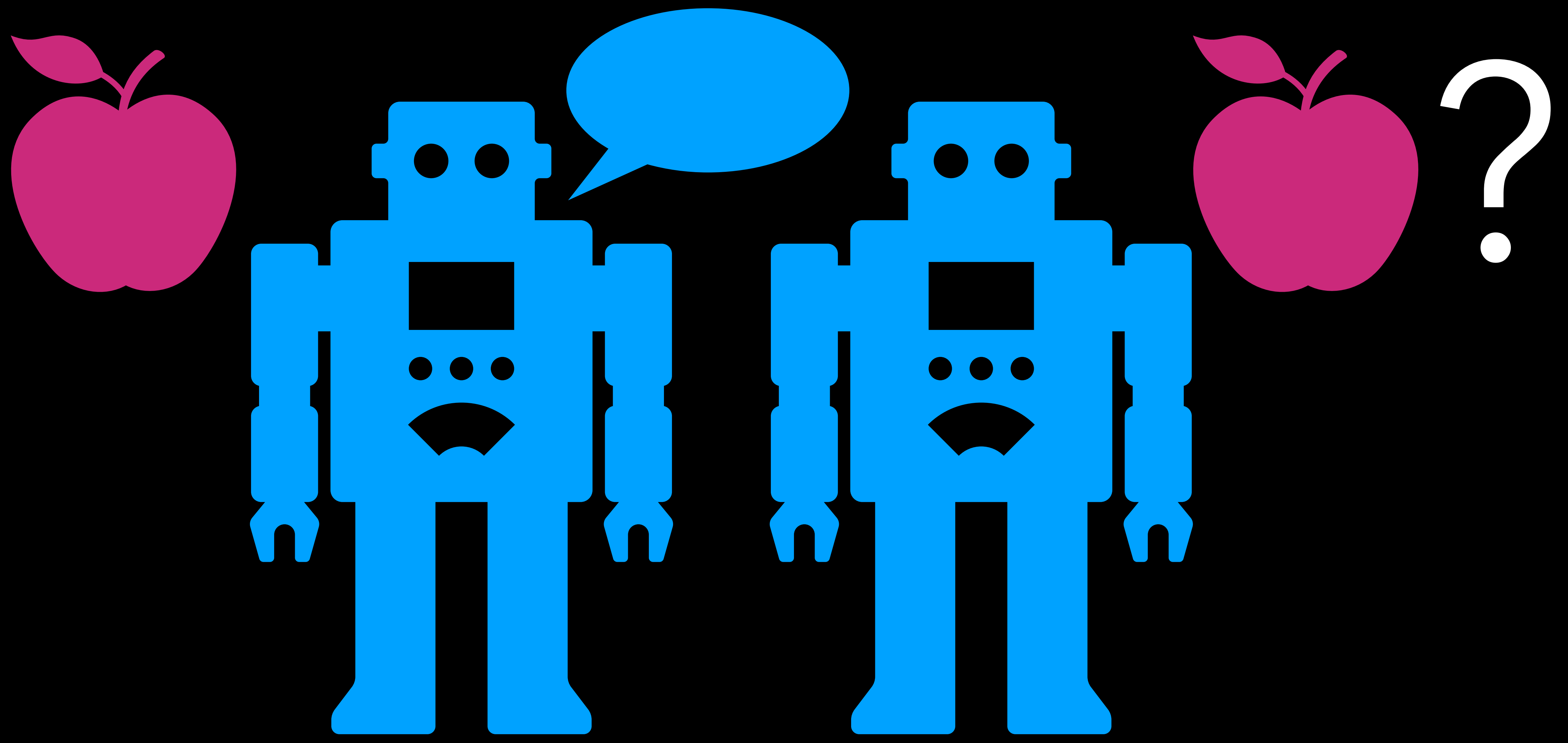
Larger communities create more systematic languages (Raviv et al., 2019)

More systematic languages are easier to learn (Raviv et al., 2021)

The Lewis Game

Same experimental playground

Larger communities create more systematic languages (Raviv et al., 2019)

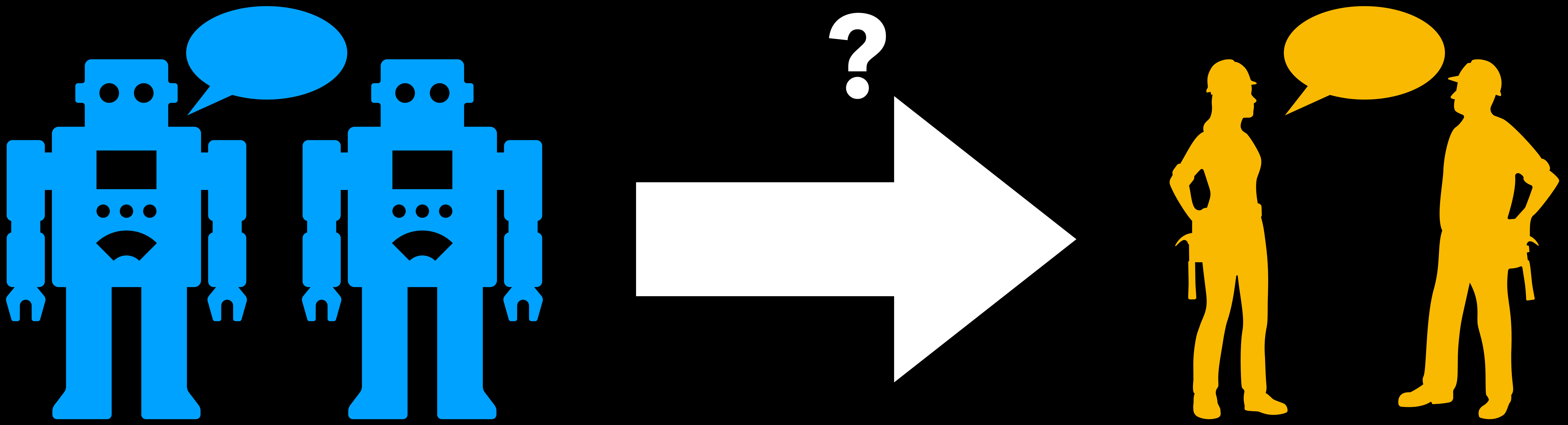


Emergent Communication

Experiments with Human Participants

Galke, Ram, & Raviv (2022)

What do machines tell us about humans?

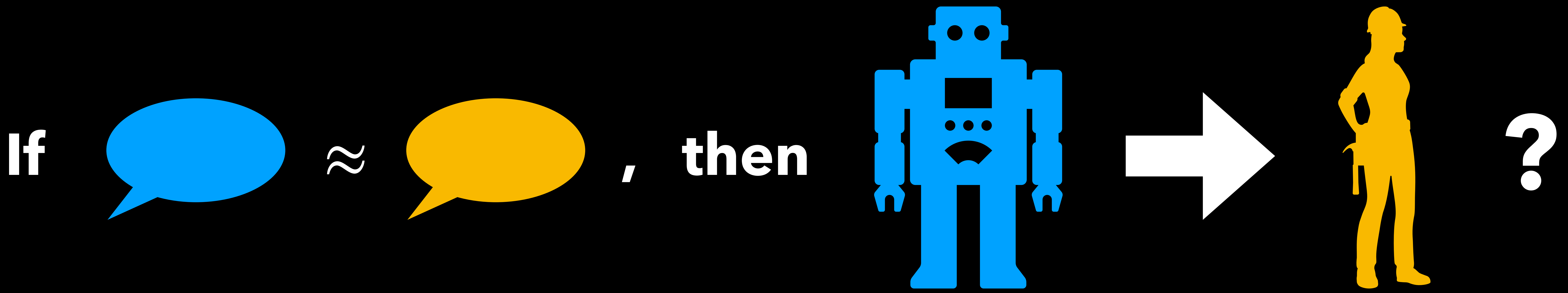


Emergent Communication

Human Language Evolution

Galke, Ram, & Raviv (2022)

What if machines were "Linguistically Plausible"?



Shared linguistic properties between machines and humans

Galke, Ram, & Raviv (2022)

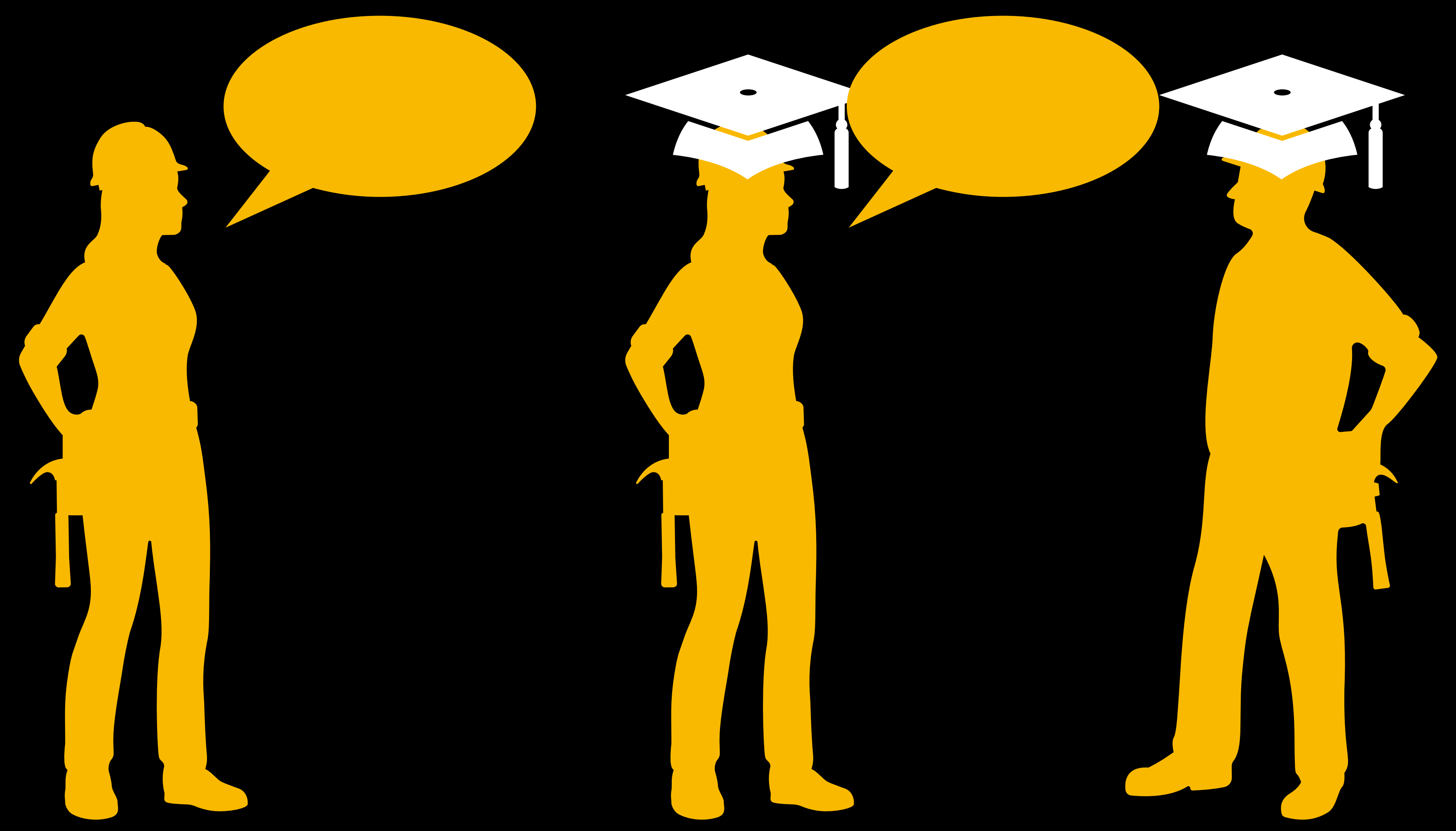
Larger groups → more structure

- **Humans**: Larger communities create more structured languages (Raviv et al., 2019)
- **Machines**: Hard to replicate, but population heterogeneity looks promising (Rita et al., 2022)

Galke, Ram, & Raviv (2022)

More structure → easier to learn

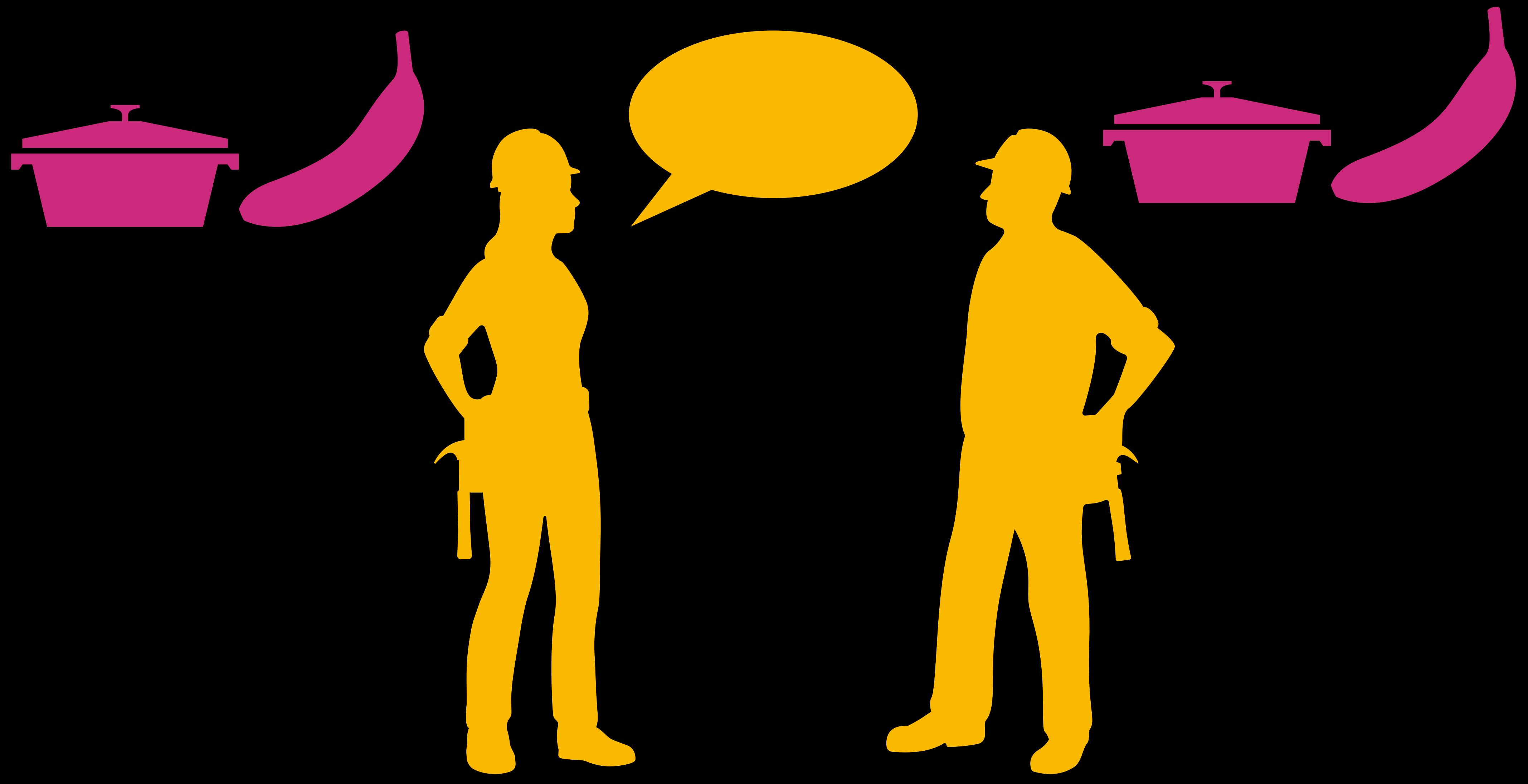
- **Humans/Machines**: Structure prevails in generational transmission, as simulated by listener reset (Li & Bowling, 2019)
- **Humans**: Structured languages are easier to learn (Raviv et al., 2021)



Galke, Ram, & Raviv (2022)

More structure → better generalization

- **Humans**: compose existing concepts to form new meanings
- **Machines**: can generalise without compositionality (Chaabouni et al., ACL 2020)



Galke, Ram, & Raviv (2022)

Linguistic phenomena in Humans & Machines

- Larger groups → more structure ✓ (partially)
- More structure → easier to learn ✓ (assumed)
- More structure → better generalization ✗ The big mystery

Galke, Ram, & Raviv (2022)

Trade-off between Simplicity and Expressivity
(Kirby et al., 2015)

But: Machines have virtually infinite memory → no need for simplicity

Galke, Ram, & Raviv (2022)

Linguistic phenomena in Humans & Machines

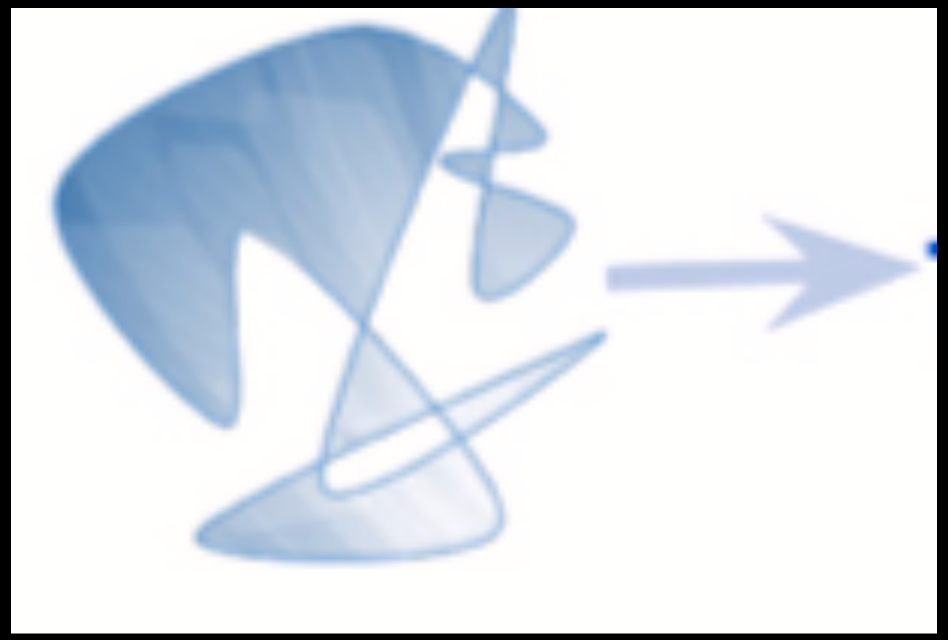
- Larger groups → more structure ✓ (partially)
- **More structure → easier to learn** ✓ (assumed)
- **More structure → better generalization** ✗ The big mystery

→ **Replicate** More systematic languages are easier to learn (Raviv et al., 2021) **with neural networks**

Do machines benefit from structure?

- 10 input languages with different degrees of structure
  - Created in lab experiments with humans (Raviv et al., 2019)
- **Training**: Exposure, guessing, production
- **Testing**: Memorization & generalisation
- Key difference to emergent communication:
  pure language learning without reinforcement learning

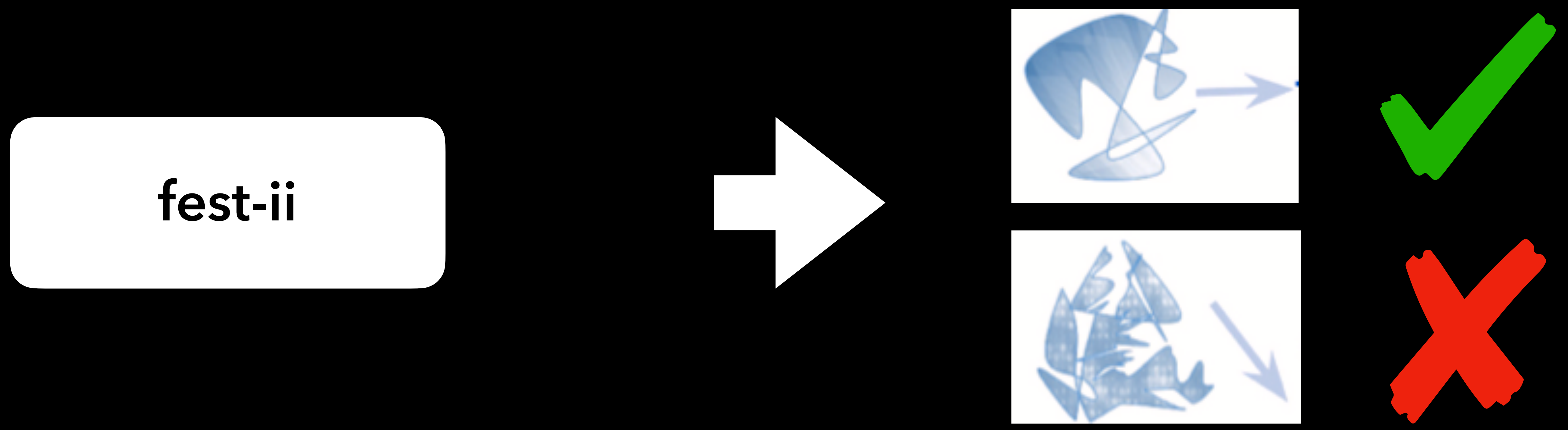
Production Block



fest-ii

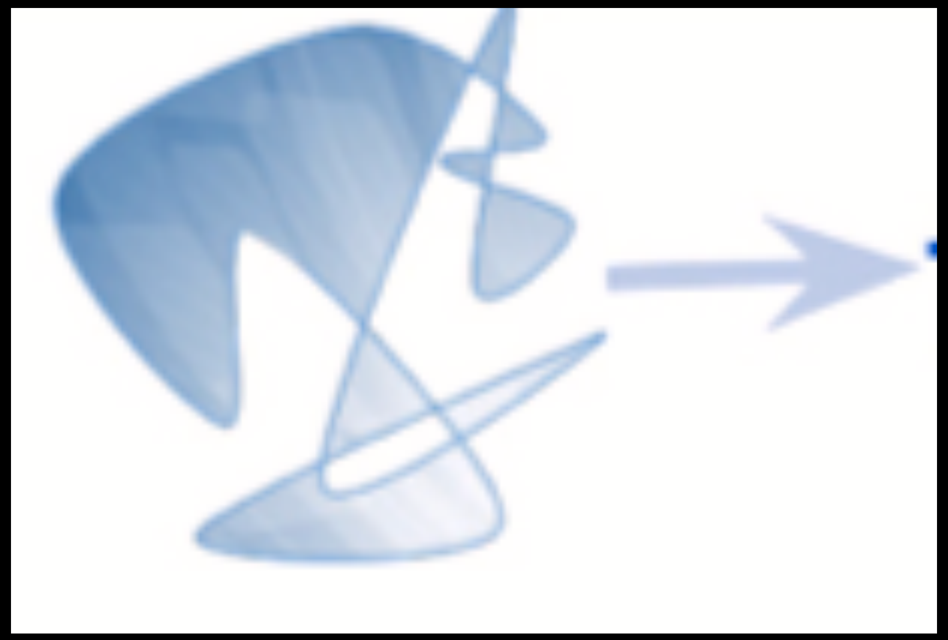
- **Input:** A scene
- **Task:** Produce a label to describe the scene
- **Output:** A label
- **Model:** Scene encoder and generative decoder

Guessing Block



- **Input:** a label, plus a list of candidate scenes
- **Task:** Find the right scene among distractors
- **Output:** Correct scene
- **Model:** Label and scene encoders + contrastive training objective

Exposure Block



fest-ii

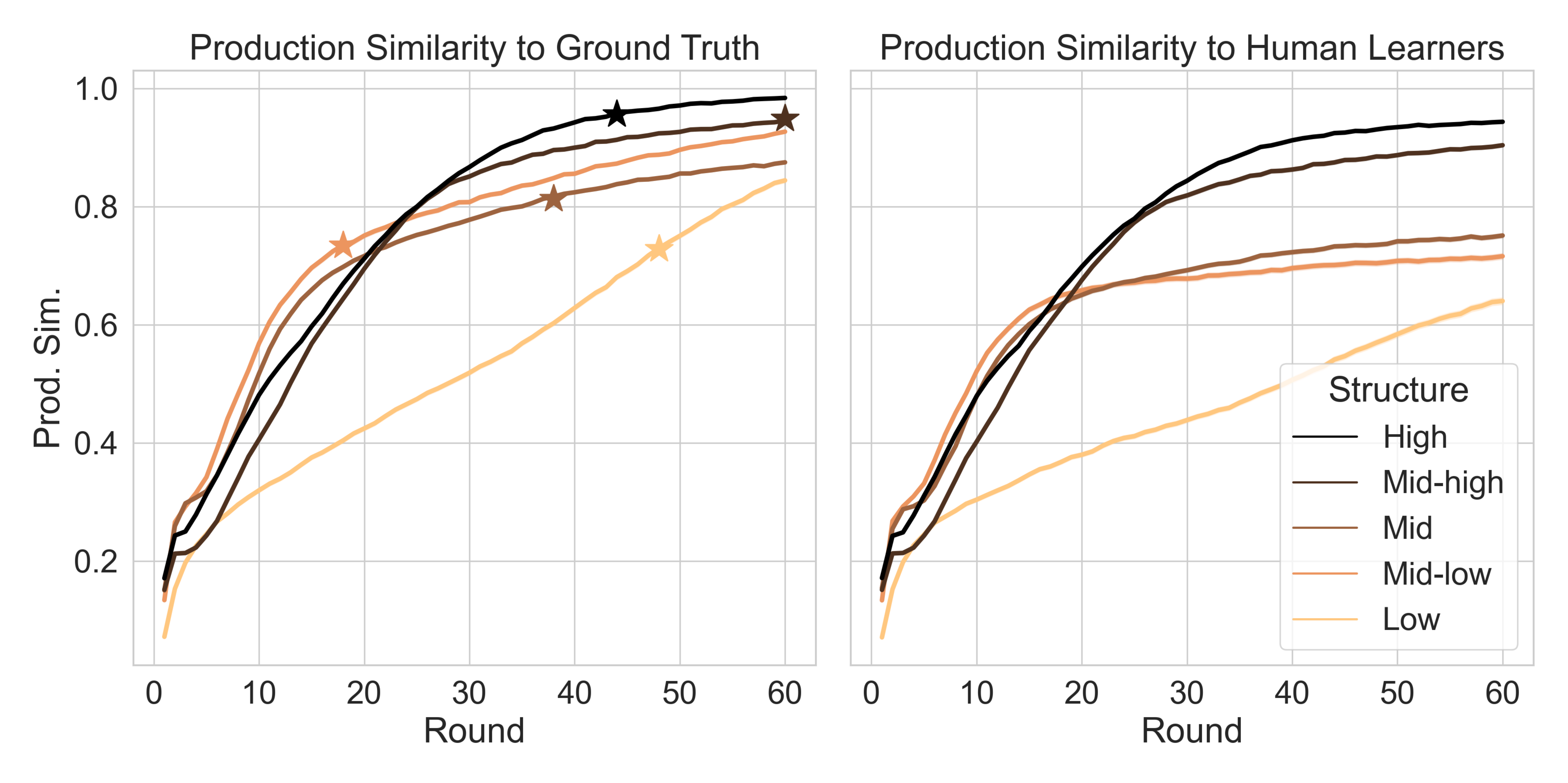
- **Input:** a label and the corresponding scene
- **Task:** Just look at the scenes
- **Output:** Nothing
- **Model:** Mix of generative and contrastive training

Model architecture

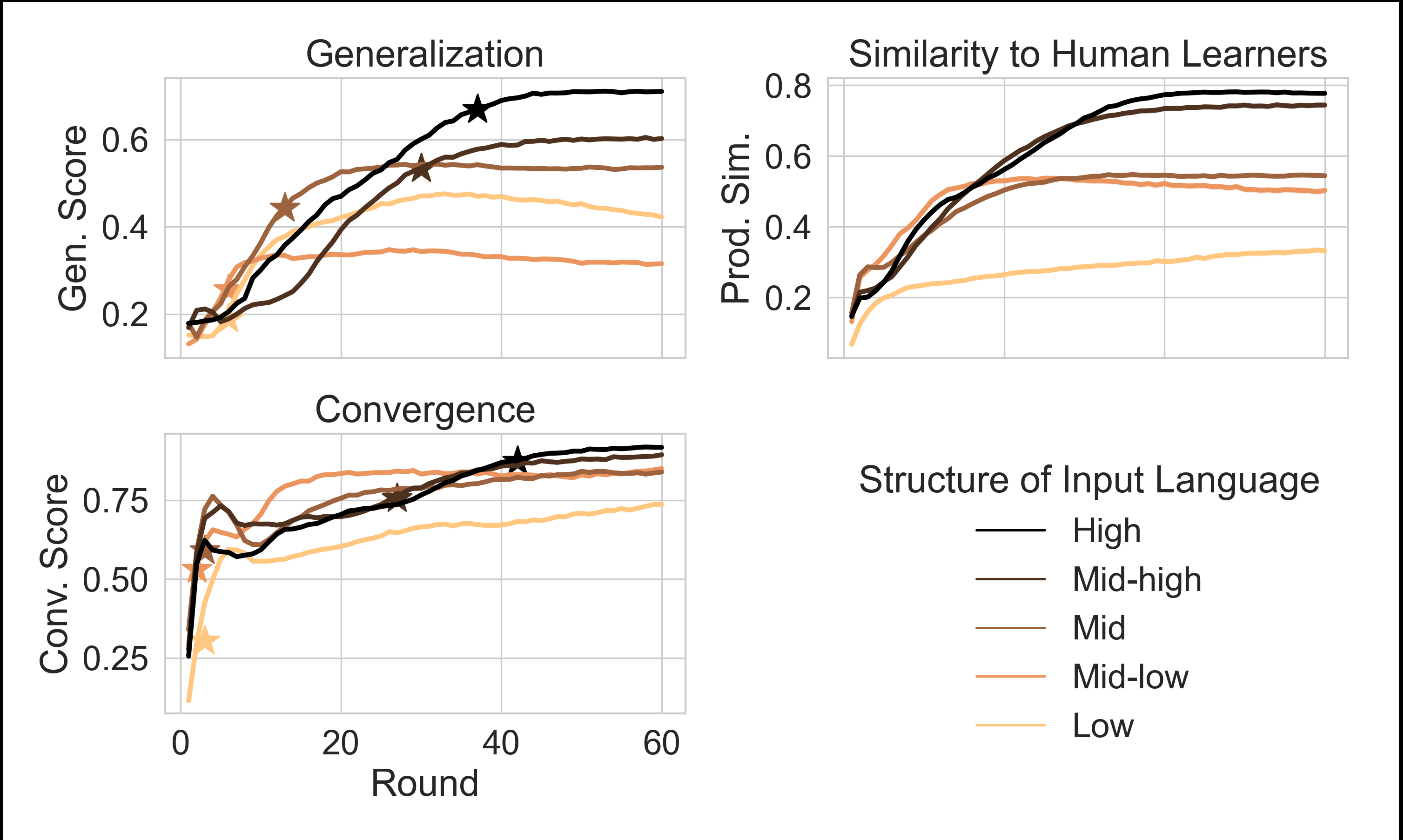
Metrics

- **Production Similarity**: average pairwise length-normalised edit distance
  - Prod. Sim. to ground truth of input languages
  - Prod. Sim. to human learners
- **Generalisation Score**: How systematic is the generalisation to new scenes compared to memorised labels for known scenes
- **Convergence Score**: To what extent do different agents come up with the same generalisations

Results of Memorization Test



Generalisation Test



Summary

- Induced structure in Text Classification
  - Pretrained transformers best, followed by Bag-of-words MLP
- External structure in evolving graphs
  - Graph neural nets helpful, replay buffer needed, parameter reuse helps
- Internal structure in language learning
  - More structure improves memorization and generalization

Thank you

Questions and feedback welcome!

- Acknowledgements:
  - Collaborators in current project: Limor Raviv and Yoav Ram
  - PhD advisor: Ansgar Scherp
- Feel free to follow me on Twitter for updates via @LukasGalke

References

Galke, L., Franke, B., Zielke, T., & Scherp, A. (2021). Lifelong Learning of Graph Neural Networks for Open-World Node Classification. *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533412

Galke, L., Mai, F., Schelten, A., Brunsch, D., & Scherp, A. (2017). Using Titles vs. Full-text as Source for Automated Semantic Document Annotation. *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017*, 20:1-20:4. https://doi.org/10.1145/3148011.3148039

Galke, L., Ram, Y., & Raviv, L. (2022, April 22). Emergent Communication for Understanding Human Language Evolution: What's Missing? *Emergent Communication Workshop*. EmeCom at ICLR. https://doi.org/10.48550/arXiv.2204.10590

Galke, L., & Scherp, A. (2022). Bag-of-Words vs. Graph vs. Sequence in Text Classification: Questioning the Necessity of Text-Graphs and the Surprising Strength of a Wide MLP. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4038–4051. https://doi.org/10.18653/v1/2022.acl-long.279

Mai, F., Galke, L., & Scherp, A. (2018). Using Deep Learning for Title-Based Semantic Subject Indexing to Reach Competitive Performance to Full-Text. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 169–178. https://doi.org/10.1145/3197026.3197039