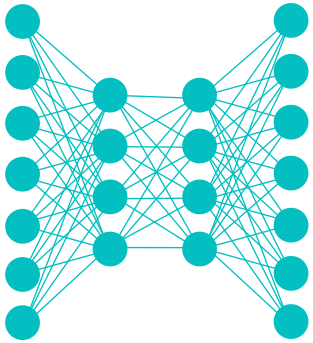
Lecture Notes for

Deep Learning

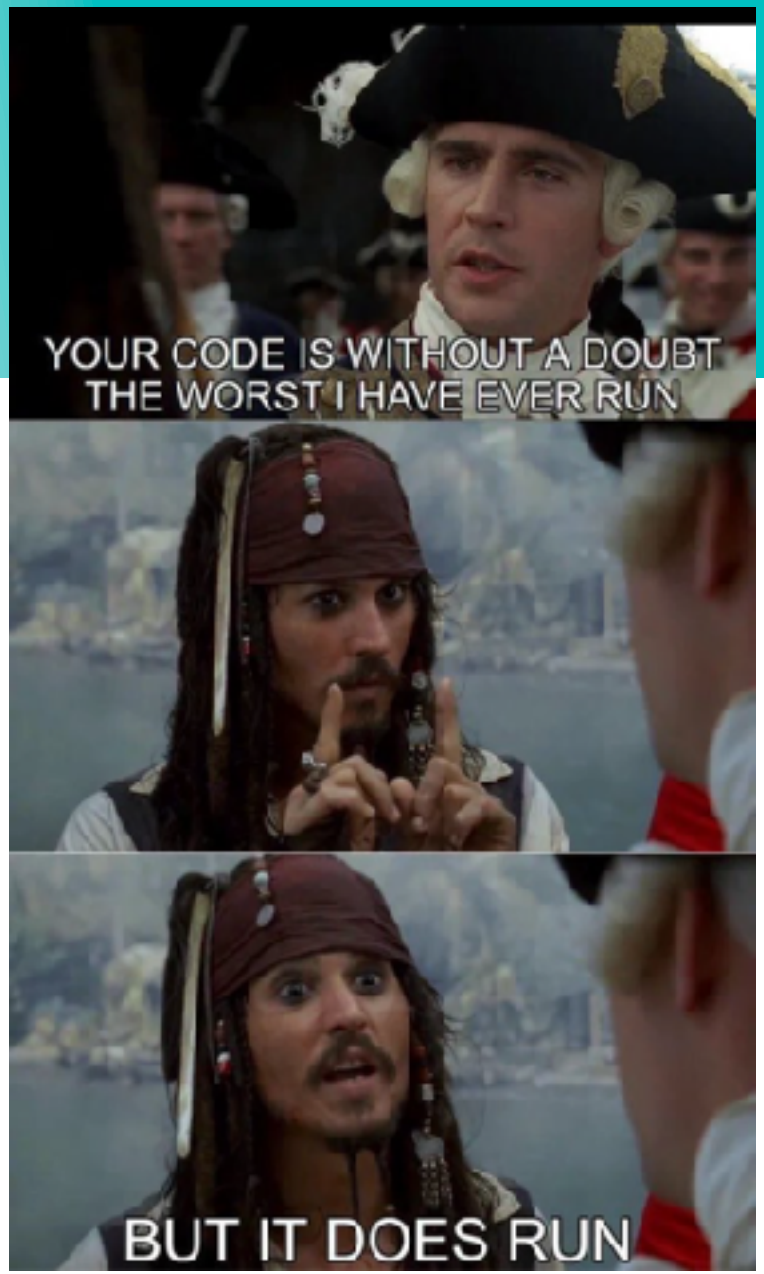
De-biasing Strategies



Logistics and Agenda

- Logistics
  - Lab due soon!
  - Office hours this week: Wednesday
  - Exams: Dates now included
- Last Time:
  - Case Studies
  - MITRA Paper
  - NLP Embeddings
- Agenda
  - Implicit and Explicit de-biasing
  - Town Hall Lab One

De-biasing Strategies

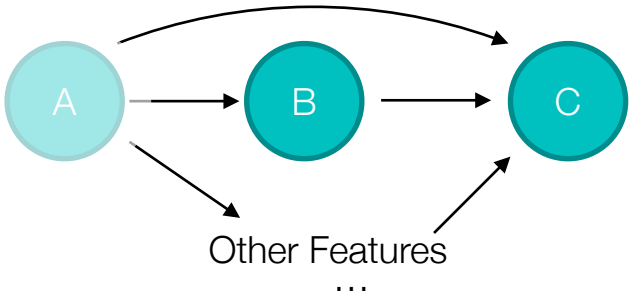


De-biasing Strategies

- **Explicit**:
  - Incorporate fairness in loss function, $\mathscr{L}_{bias}(\cdot)$
  - Incorporate other constraints during training that limit bias
  - Identify features or model parts that exacerbate bias
    - Prune out or decrease influence of offending parts of model, while tracking performance trade off (if any)
- **Implicit**:
  - Incorporate additional knowledge sources that *might* mitigate bias
    - More structured relationships could help disambiguate
  - Open question: *Do LLMs have a unique ability to identify bias form training data, and mitigate it?*

Measuring Reliability and Fairness

- Identify potential bias, groups defined by attribute "A"
- **Fairness through unawareness**, no knowledge of A:
  $f(\mathcal{X})_{\setminus A} \rightarrow \mathcal{Y}$ (omission of data)



- **Individual Fairness**, similar individuals are classified similarly: $d(i,j) < \epsilon \rightarrow f(\mathcal{X}^{(i)}, A^{(i)}) \approx f(\mathcal{X}^{(j)}, A^{(j)})$
  - where d is a measure of if i,j individuals similarity

Defining which individuals should be close can be difficult or expensive to collect…

Measuring Reliability and Fairness

- **Demographic Parity**: $f(\mathcal{X} \mid A = 0) \approx f(\mathcal{X} \mid A = 1)$
  - Attribute should never influence outcome…
- **Equal Opportunity**: Positive class not influenced by A

$$f(\mathcal{X} \mid A = 0, Y = 1) \approx f(\mathcal{X} \mid A = 1, Y = 1)$$

Can be good in many situations, but tend to decrease performance when some groupings should influence outcomes

- **Counterfactual fairness**: $\lfloor f(X_a) \rfloor = \lfloor f(X_{a'}) \rfloor$ for a given set of groups, a and a' (need group definition)
- **EDDI**: $$\frac{1}{|A|} \sum_{a \in A} \frac{ER_a - ER_A}{\max(ER_A, \quad 1 - ER_A)}$$
- **Minimum Difference**: Minority class confidences distribution should match majority (distribution matching definition requires careful attention)

Kusner et al. "Counterfactual Fairness" in Proceedings of Neurips 2017

Fairness as Loss Functions

- Identify: measure differences in reliability for identified groups, measure **statistical difference** and **impact**
- Develop examples of interest with counterfactual fairness,
  - original example: features with X_a where A=a
  - counterfactual: features with $X_{a'}$ where A=a` *and outcome should not change, expert judged*
- **Counterfactual loss**:
  - $\mathscr{L}_{cf} = \|f(X_a) - f(X_{a'})\|^2$ **or other measure of closeness**
  - $\mathscr{L}_{tot} = \mathscr{L}_{bce} + \lambda \cdot \mathscr{L}_{cf}$
- **Min Diff**, define two groups, a,b that should be similar:
  $\mathscr{L}_{md} = \mu(f(X_a)) - \mu(f(X_b))$
- **EDDI Loss:** Minimize EDDI over batches

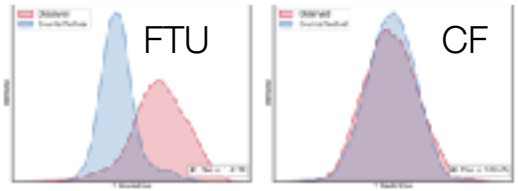
Only defined for validation set, run after full epoch

A result on common datasets

Base and Unaware | **Counter Factual Training or EO**

Synthetic Loan Data

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	FLAP$_1$(O)	FLAP$_2$(O)	FLAP$_1$(M)	FLAP$_2$(M)	OB$_1$	OB$_2$
↑ ACC	0.6618	0.6481	0.6224	0.6237	0.6224	0.6237	0.6224	0.6237	0.6224	**0.6406**	0.6279
↑ AUC	0.9457	0.8986	0.5867	**0.6682**	0.5714	0.5868	0.5837	0.5875	0.5863	0.5704	0.5856
↓ CF-metrics	0.6291	0.3906	0.0031	0.0355	0.0034	0.0016	0.0032	**0.0002**	**0.0002**	0.0011	0.0026
↓ CF Bound	0.8690	0.9464	0.1836	0.1071	0.0918	0.0937	0.1847	0.0690	**0.0670**	0.0830	0.2340
↓ EO Fairness	0.5469	0	0.0156	**0**	0.0336	0.0321	0.0156	0.0301	0.0180	**0**	**0**
↓ AA Fairness	0.6235	0.4559	**5.6e-18**	0.0370	**1.1e-18**	**3.3e-18**	**6.7e-18**	0.0012	0.0038	**4.6e-17**	**4.3e-17**



Two groups identified and their distributions, KL measure difference. Lower diff is better.

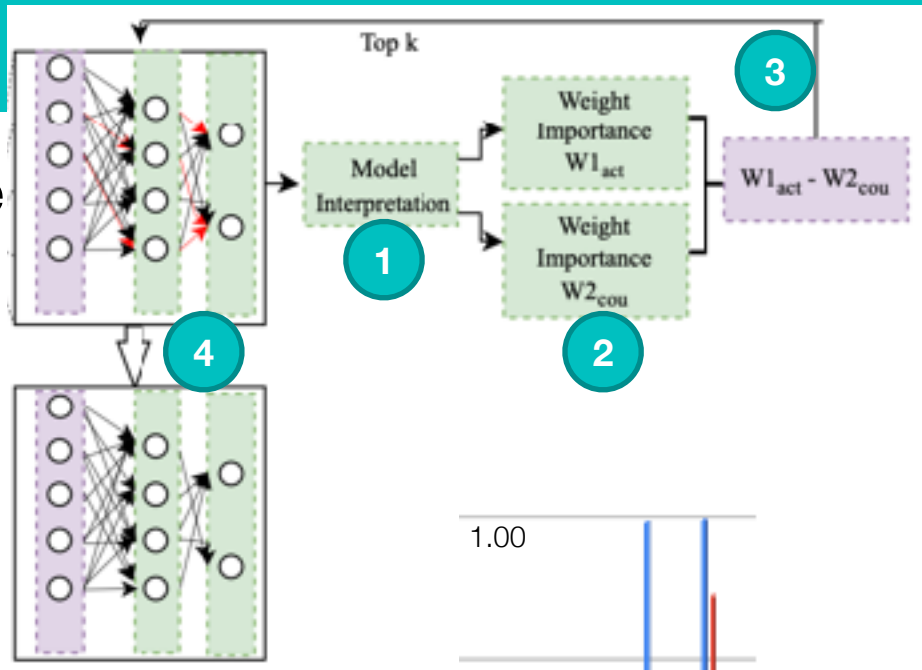
COMPAS Data: who will reoffend in next two years

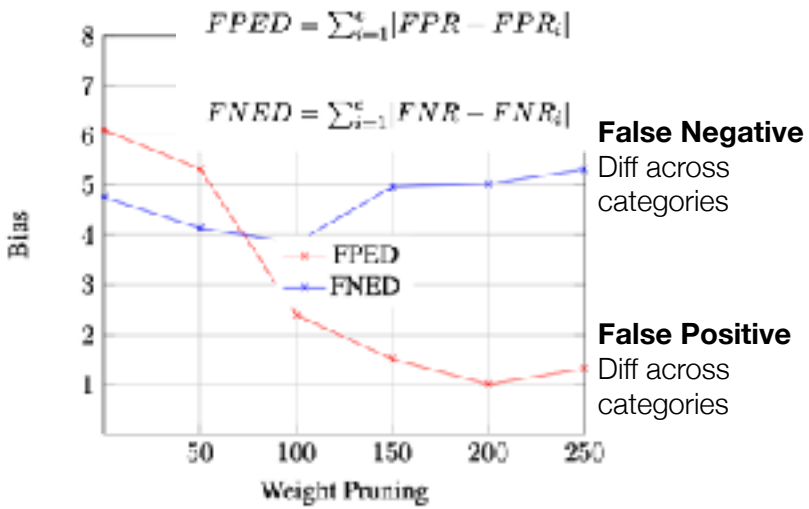
Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	FLAP$_1$(O)	FLAP$_2$(O)	FLAP$_1$(M)	FLAP$_2$(M)	OB$_1$	OB$_2$
↑ ACC	0.5744	0.5726	0.5598	**0.5710**	0.5609	0.5605	0.5599	0.5607	0.5607	0.5666	0.5674
↑ AUC	0.7206	0.7225	0.6928	0.7225	0.6927	0.6927	0.6928	0.7015	0.7019	**0.6764**	**0.6744**
↓ CF-metric	0.2274	0.1406	0.0054	0.1377	0.0060	0.0058	0.0054	**0.0026**	0.0027	0.0060	0.0065
↓ EO Fairness	0.1046	0	0.1374	**0**	0.1405	1.7e-06	3.3e-06	6.7e-07	1.2e-06	**0**	**0**
↓ AA Fairness	0.2258	0.1460	**0**	0.1424	**0**	2.9e-07	5.6e-07	8.2e-07	3.0e-07	1.6e-16	1.1e-16

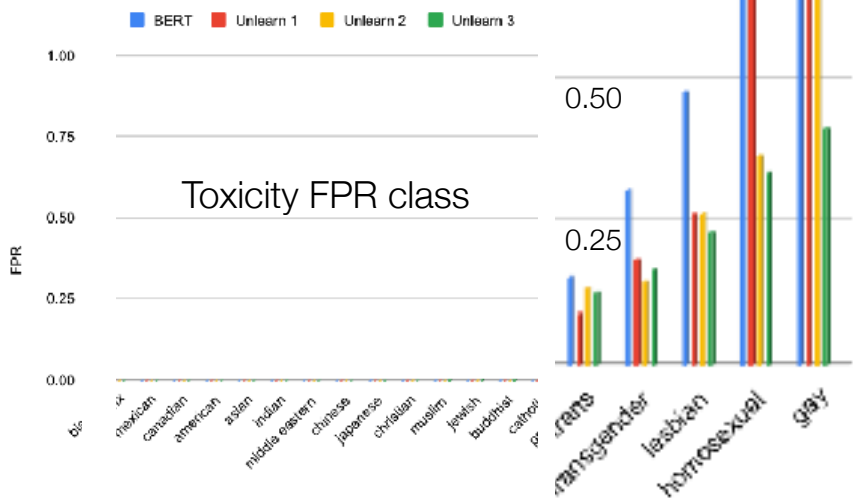
https://arxiv.org/pdf/2403.17852v1 Chen and Zhu, Counterfactual Fairness through Transforming Data Orthogonal to Bias, 2024

BiasWipe

- Run model explainability with feature sensitivity or importances 1
- Identify counter factual groups and measure bias 2
- Identify weights that are most influential on CF examples (k-top) 3
- Prune weights (make zero) 4





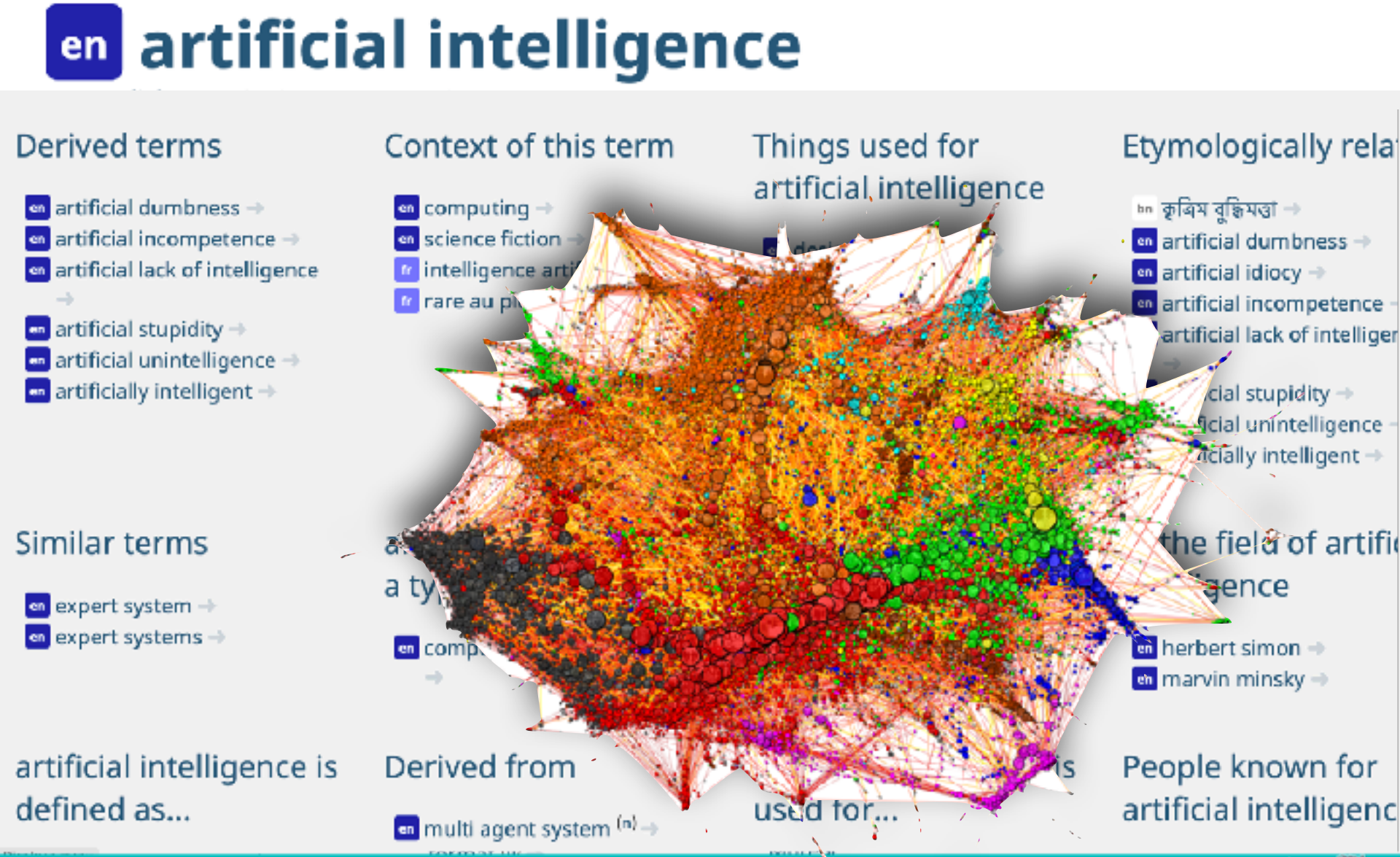


Mamta et al., BiasWipe: Mitigating Unintended Bias in Text Classifier through Model Interpretability, EMNLP 2024 https://aclanthology.org/2024.emnlp-main.1172.pdf

Implicit De-biasing

- Adding new knowledge can eliminate unintended bias
- ***Can it?***

ConceptNet, a Knowledge Graph



ConceptNet Numberbatch

- Create with a Knowledge Graph (from multiple sources with relations like *UsedFor*, *PartOf*, *etc*.)
- Based KG edges, perturb existing embeddings (like GloVe) to optimize:

$$\Psi(Q) = \sum_{i=1}^{n} \left[\alpha_i \|q_i - \hat{q}_i\|^2 + \sum_{(i,j) \in E} \beta_{ij} \|q_i - q_j\|^2 \right]$$

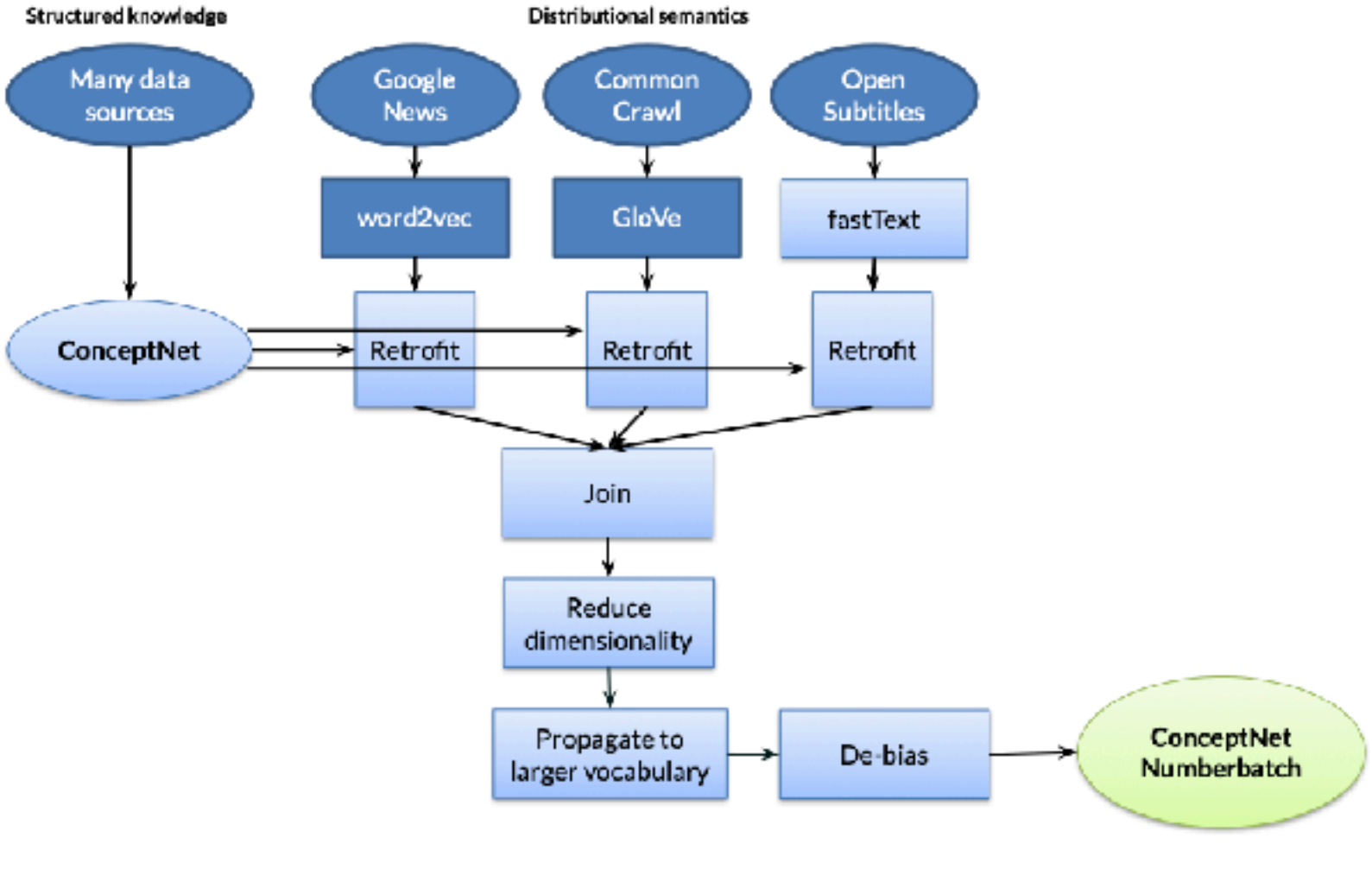
new embed — old embed — neighbors from KG

(keep similar to original) — (make similar according to other knowledge)

- Easy to optimize the objective by averaging neighbors in the ConceptNet KG
- Multiple embeddings achieved by merging through “retrofitting” which projects onto a shared matrix space (with SVD)

ConceptNet 5.5: An Open Multilingual Graph of General Knowledge, Speer et al., 2017

Building ConceptNet Numberbatch



LUMINOSO

Aside: Transparency in Research

ConceptNet is all you need

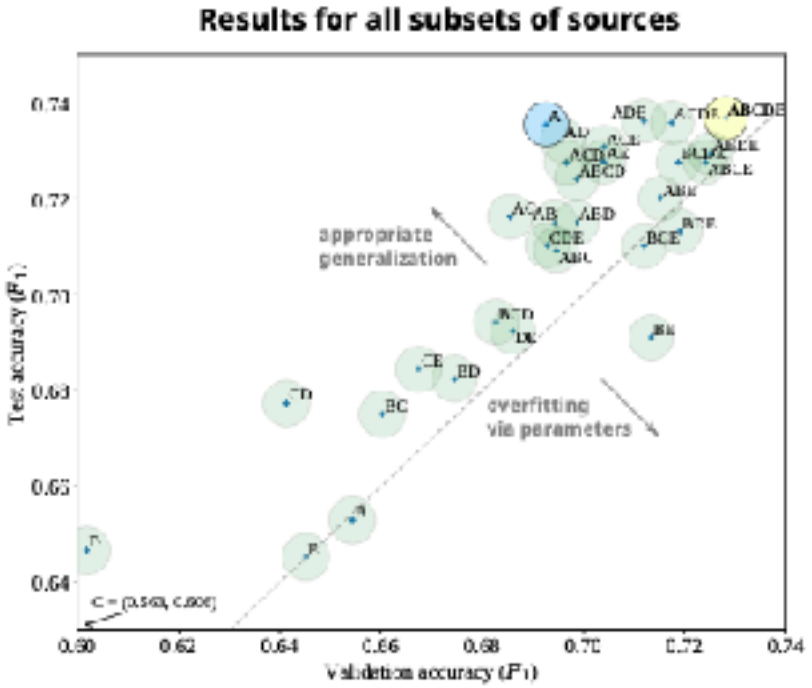
Our full classifier used the linear combination of 5 types of input features shown above. This point is labeled **ABCDE** on the graph to the right. The other points are ablated versions of the classifier, trained on subsets of the five sources.

We found that the single feature of ConceptNet similarity (**A**) performed just as well on the test data as the full classifier, despite its lower validation accuracy.

This one-feature classifier could be more simply described as a heuristic over cosine similarities of ConceptNet embeddings:

$$\text{sim}(term_1, art) - \text{sim}(term_2, art) > 0.0961$$

It seems that the test data contained distinctions that can already be found by comparing ConceptNet embeddings, and that more complex features may have simply provided an opportunity to overfit to the validation set by parameter selection.



This graph shows the validation and test accuracy of classifiers trained on subsets of the five sources of features. Ellipses indicate standard error of the mean, assuming that the data is sampled from a larger set.

https://blog.conceptnet.io/2018/06/naacl2018-poster.pdf

Aside: ConceptNet Numberbatch

As a kid, I used to hold marble racing tournaments in my room, rolling marbles simultaneously down plastic towers of tracks and funnels. I went so far as to set up a bracket of 64 marbles to find the fastest marble. I kind of thought that running marble tournaments was peculiar to me and my childhood, but now I've found out that marble racing videos on YouTube are a big thing! Some of them even have overlays as if they're major sporting events.

In the end, there's nothing special about the fastest marble compared to most other marbles. It's just lucky. If one ran the tournament again, the marble champion might lose in the first round. But the one thing you could conclude about the fastest marble is that it was no *worse* than the other marbles. A bad marble (say, a misshapen one, or a plastic bead) would never luck out enough to win.

In our paper, we tested 30 alternate versions of the classifier, including the one that was roughly equivalent to this very simple system. We were impressed by the fact that it performed as well as our real entry. And this could be because of the inherent power of ConceptNet Numberbatch, or it could be because it's the lucky marble.

-Robyn Speer
http://blog.conceptnet.io

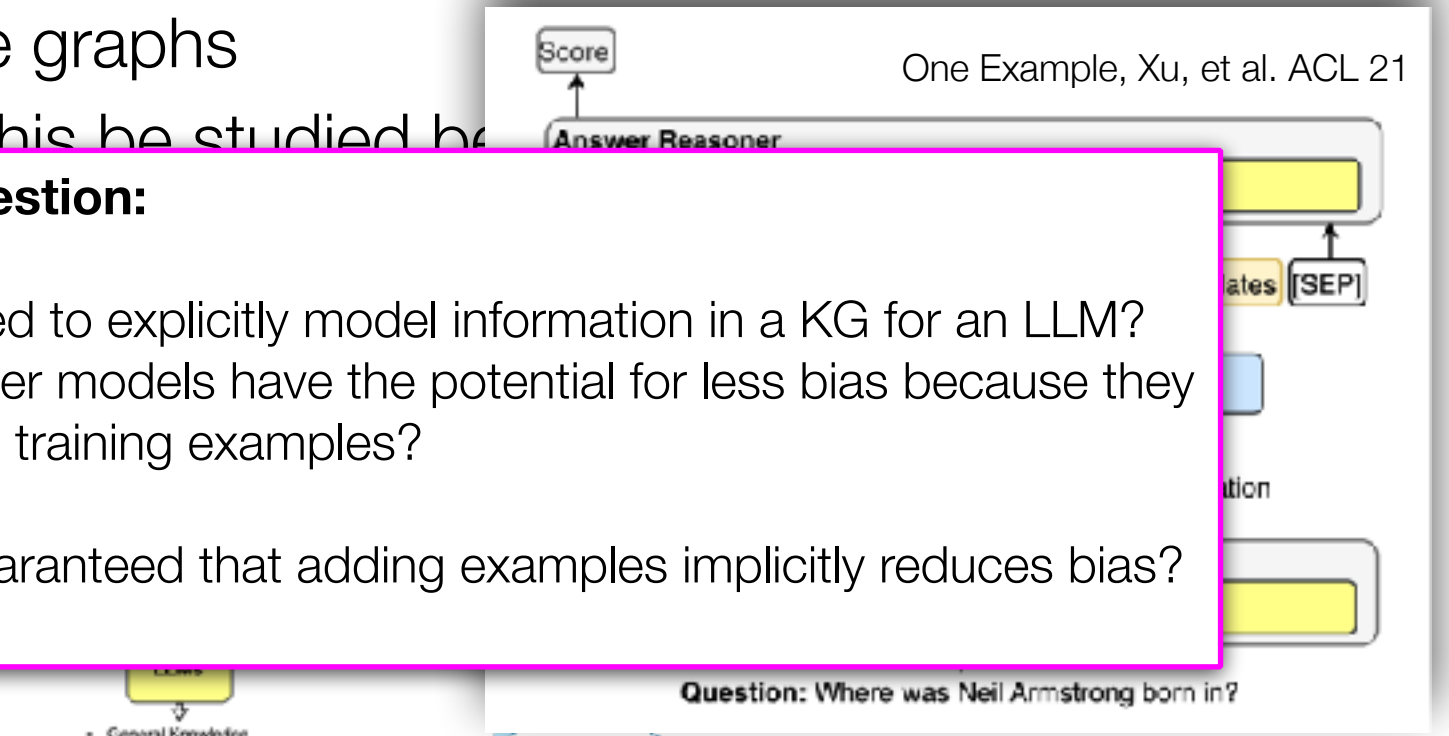
Beyond Word Embeddings

- ConceptNet Numberbatch is designed to help reduce bias in word embeddings analogy through incorporating knowledge graphs
- How can this be studied be

Open Question:

Do we need to explicitly model information in a KG for an LLM? Or, do larger models have the potential for less bias because they have more training examples?

Are we guaranteed that adding examples implicitly reduces bias?



Lecture Notes for

Deep Learning

De-biasing strategies

Next Time:
Transfer Learning

Reading: Chollet Article