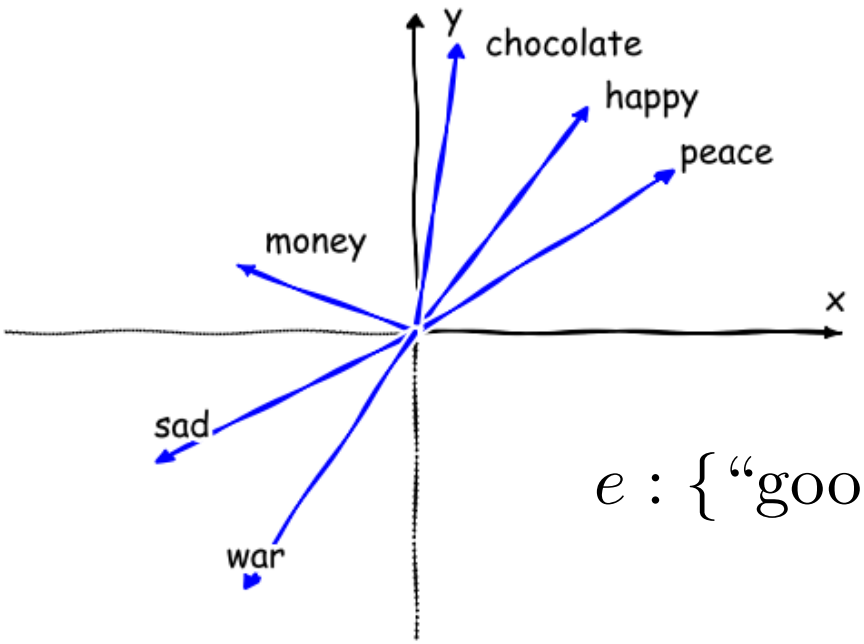
Analytical Methods for Interpretable Ultradense Word Embeddings

PHILIPP DUFTER, HINRICH SCHÜTZE
CIS – LMU MUNICH

EMNLP 2019 – HONG KONG

Word Embeddings are Ubiquitous



$$e : \{\text{“good“}, \text{“peace“}, \dots\} \rightarrow \mathbb{R}^d$$

$$E \in \mathbb{R}^{n \times d}$$

Dimensions are not Interpretable

$$e(\text{“money”}) = \begin{pmatrix} 0.5 \\ -0.8 \\ 0.2 \\ 1.5 \end{pmatrix}$$

Interpretability = first dimension correlated with e.g., sentiment

Interpretability is desirable:

- Retain relevant information
- Discard irrelevant information
- Supports goal of interpretable networks

Two Possibilities to Obtain Interpretable Embedding Spaces

i) Introduce new embedding algorithms
ii) Transform existing embedding spaces

Our Focus

$$E' = EQ$$

$E \in \mathbb{R}^{n \times d}$ original space
$Q \in \mathbb{R}^{d \times d}$ orthogonal
$E' \in \mathbb{R}^{n \times d}$ interpretable space

Distance-preserving

\+ Maintains downstream performance
\- Assumes interpretable dimensions exist in the original space

Simplification: Identify Interpretable Dimension

$$e' = Eq$$

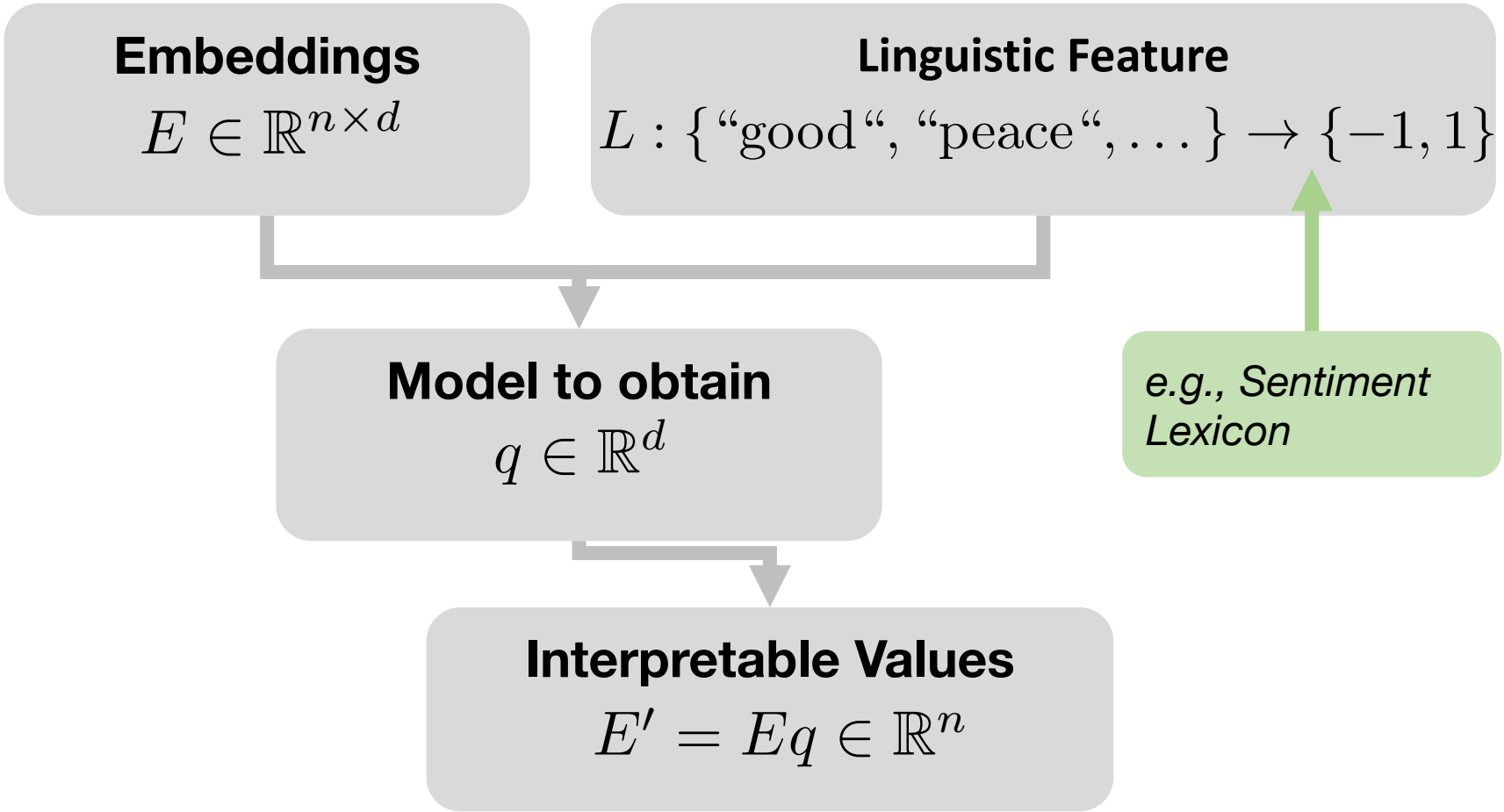
$E \in \mathbb{R}^{n \times d}$ original space

$q \in \mathbb{R}^{d \times 1}$ L2-normed

$e' \in \mathbb{R}^{n \times 1}$ interpretable dimension

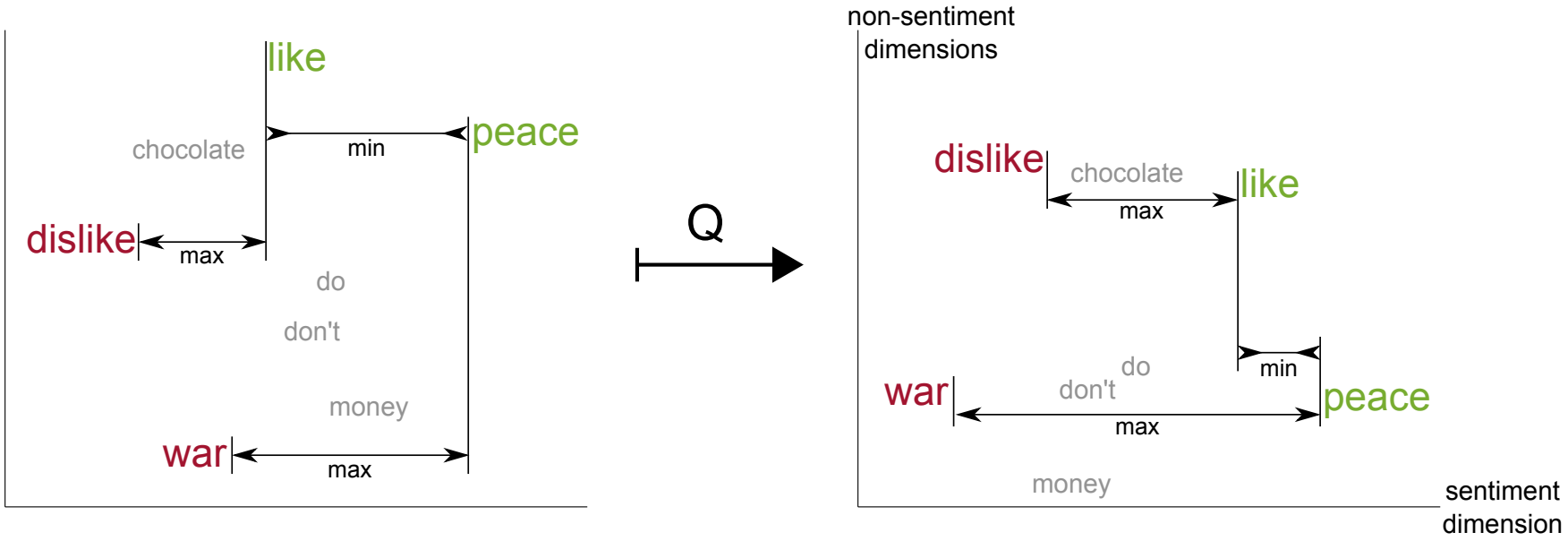
Most results generalize easily to multiple dimensions

Pretrained Embeddings and Linguistic Feature Required



Intuition Behind Existing Densifier Model

(Rothe et al. 2016)



Densifier is a Nonlinear Optimization Problem

(Rothe et al. 2016)

$$\max_{q} \sum_{\text{different label}} \|q^\intercal(e_v - e_w)\| - \sum_{\text{same label}} \|q^\intercal(e_v - e_w)\|$$

Solve with gradient descent
(potential re-orthogonalization for
multi-dimensional case)

Modified Objective: DensRay

$$\max_{q} \sum_{\text{different label}} \|q^\intercal (e_v - e_w)\|_2^2 - \sum_{\text{same label}} \|q^\intercal (e_v - e_w)\|_2^2$$

Simple closed form solution available

DensRay Objective can be Reduced to **Ray**leigh-Quotient

$$\max_q \sum_{\text{different label}} \|q^\intercal (e_v - e_w)\|_2^2 - \sum_{\text{same label}} \|q^\intercal (e_v - e_w)\|_2^2$$

$$= \max_q q^\intercal \Big(\sum_{\text{different label}} (e_v - e_w)(e_v - e_w)^\intercal - \sum_{\text{same label}} (e_v - e_w)(e_v - e_w)^\intercal \Big) q$$

$$=: \max_q q^\intercal A q$$

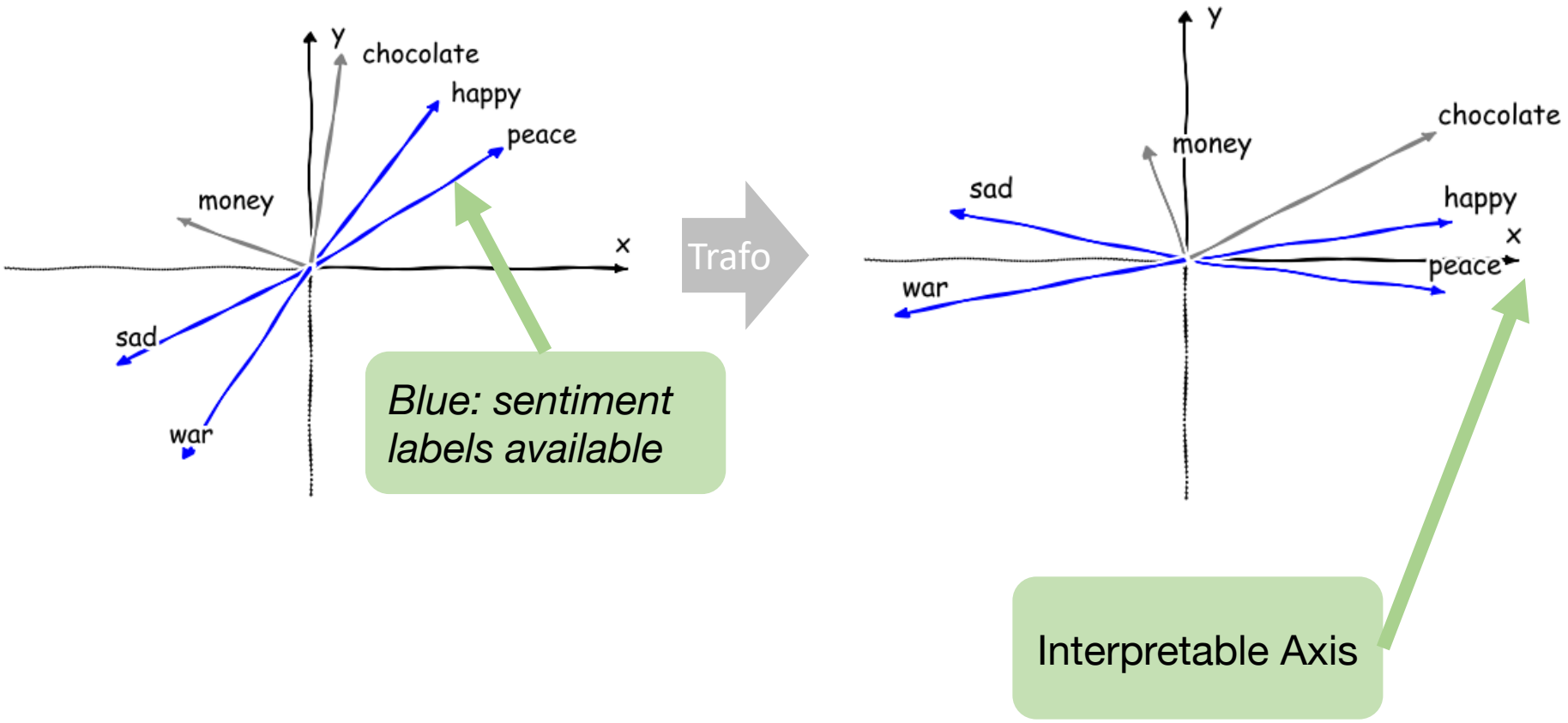
Solution: eigendecomposition of A
For multiple dimensions: always orthogonal Q.

Consider 3 Applications for DensRay

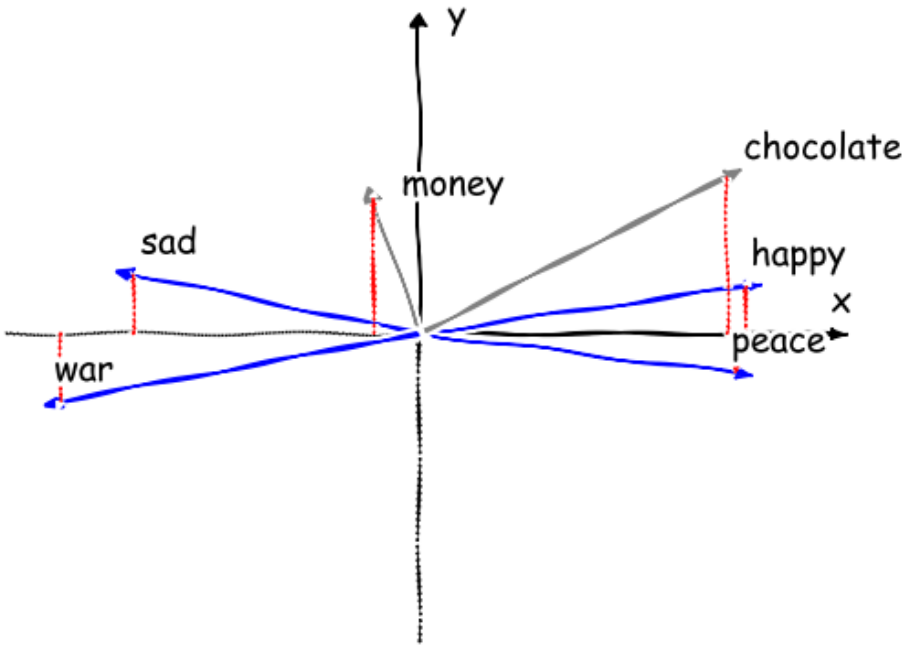
#BenderRule: **English**

1. (Sentiment) Lexicon Induction
2. Removing Gender Information
3. Set-Based Word Analogy

Experiment 1: Sentiment Lexicon Induction



Use Scores along Interpretable Axis as Sentiment Scores



Method:

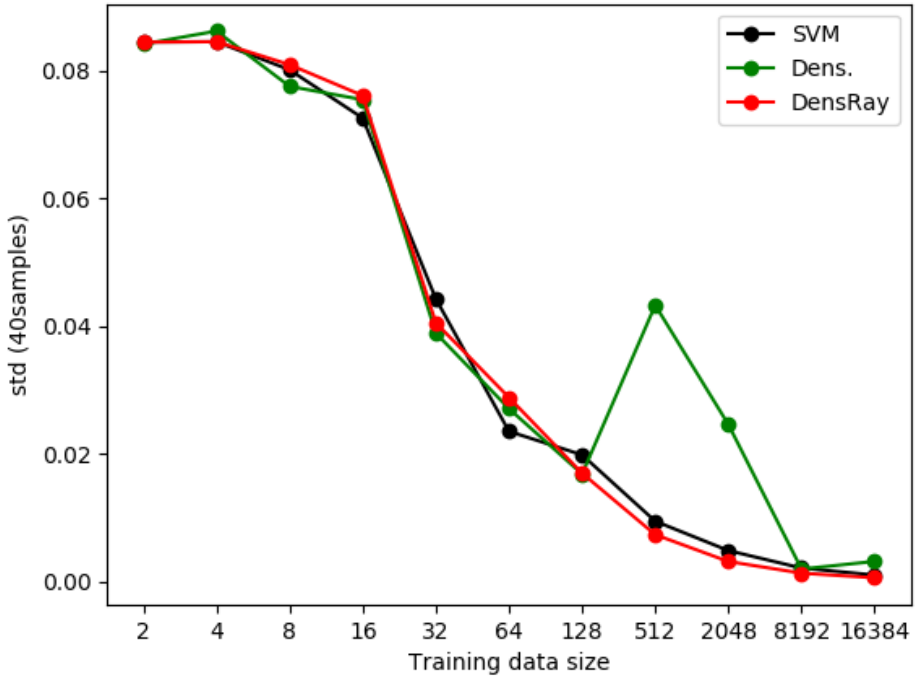
1. Use values along interpretable axis as predicted scores
2. Compute Kendall's Tau Rank Correlation with gold test lexicon

DensRay performs the same as Densifier

Macro Mean across 8 lexica	**Densifier**	**DensRay**	**SVM**
Kendall‘s Tau	0.580	0.581	**0.618**

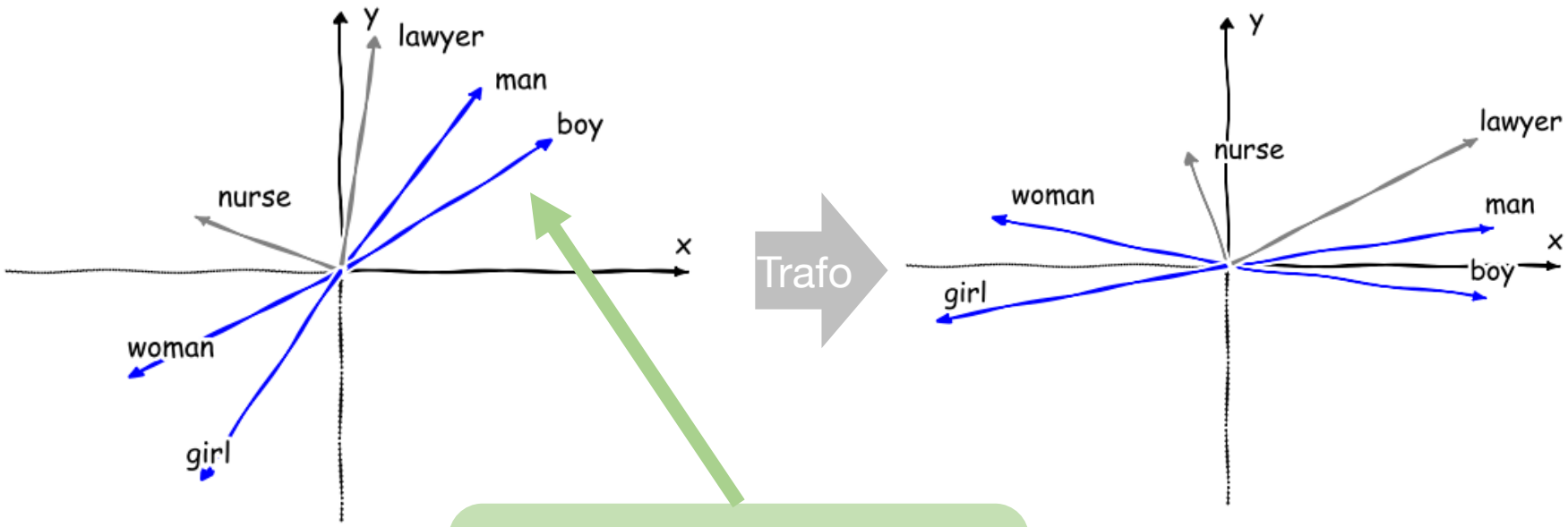
- Densifier and DensRay perform the same
- SVM Baseline is strongest

DensRay is more robust



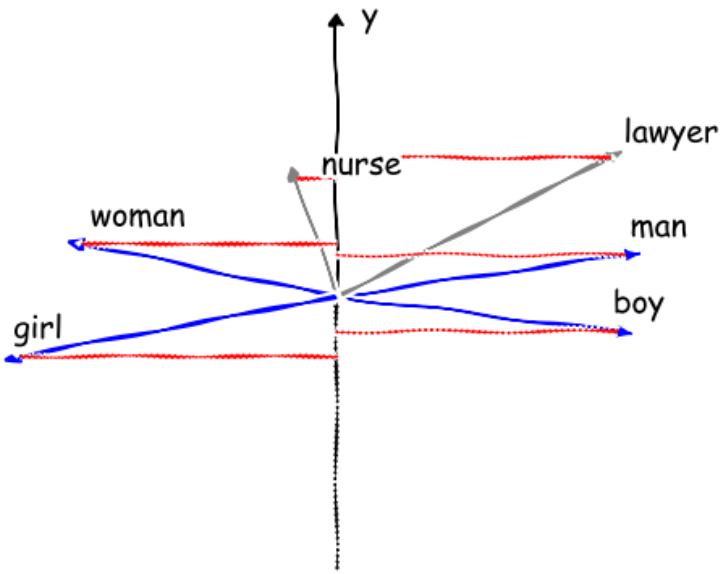
- In some cases Densifier fails to converge
- DensRay more robust than Densifier

Experiment 2:
Removing Gender Information



Use word-analogy pairs for "male-female" as lexicon

Remove the Gender Dimension



Method:

1. Simple approach: remove the gender dimension
2. Evaluate qualitatively

Examining Bias with Respect to Occupations Names

(Drozd et al. 2016)

	Original Space		Modified Space	
Cosine Similarity	**„Woman“**	**„Man“**	**„Woman“**	**„Man“**
„Actress“	0.46	0.23	0.38	0.31
„Nurse“	0.33	0.12	0.25	0.18
„Gangster“	0.20	0.34	0.28	0.32

Examining Bias with Respect to Occupations Names

(Drozd et al. 2016)

	Original Space			Modified Space		
Cosine Similarity	„Woman“		„Man“	„Woman“		„Man“
„Actress“	0.46	**0.23**	0.23	0.38	**0.07**	0.31
„Nurse“	0.33	**0.21**	0.12	0.25	**0.07**	0.18
„Gangster“	0.20	**0.14**	0.34	0.28	**0.04**	0.32

Experiment 3: Set-Based Word Analogy

(Drozd et al. 2016)

$$\text{man} - \text{woman} \simeq \text{king} - \text{queen}$$

$$a - a' \simeq b - b'$$

$$\hat{a} = \arg\max_{v \in V} \ \text{score}(v) \ \cdot \ \text{similiarity}(a, v)$$

Use e.g., Logistic Regression (Drozd et al. 2016) or score along interpretable axis

DensRay performs similar to Logistic Regression

macro mean across categories	DensRay	SVM	Logistic Regr.
FastText + BATS	0.60	0.60	**0.61**
GoogleNews + BATS	**0.48**	0.46	0.45
FastText + Google Analogy	**0.91**	0.90	0.89
GoogleNews + Google Analogy	**0.88**	0.85	0.87

Performance very similar.
DensRay beneficial for „difficult“ analogies.

Summary

- Interpretability through rotation **never harms and is useful**.
- **Analytical solutions** like SVMs or DensRay yield best performance.
- DensRay is **analytic**, **simple and works reliably** across 3 considered tasks.

- Source Code: **https://github.com/pdufter/densray**

Acknowledgements and References

Thanks to **Zentrum Digitalisierung.Bayern**, **European Research Council** (# 740516), and the anonymous **reviewers**.

Drozd, A., Gladkova, A. and Matsuoka, S., 2016, December. Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 3519-3530).

Rothe, S., Ebert, S. and Schütze, H., 2016, June. Ultradense Word Embeddings by Orthogonal Transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 767-777).

Questions