Quantifying and Reducing Gender Bias in Russian Word Embeddings

Team 2
Anna Koval
Ekaterina Kovalenko
Anastasiia Ryzhova

NLA Project
Skoltech 2018

Word embeddings and gender bias

Gender specific words:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

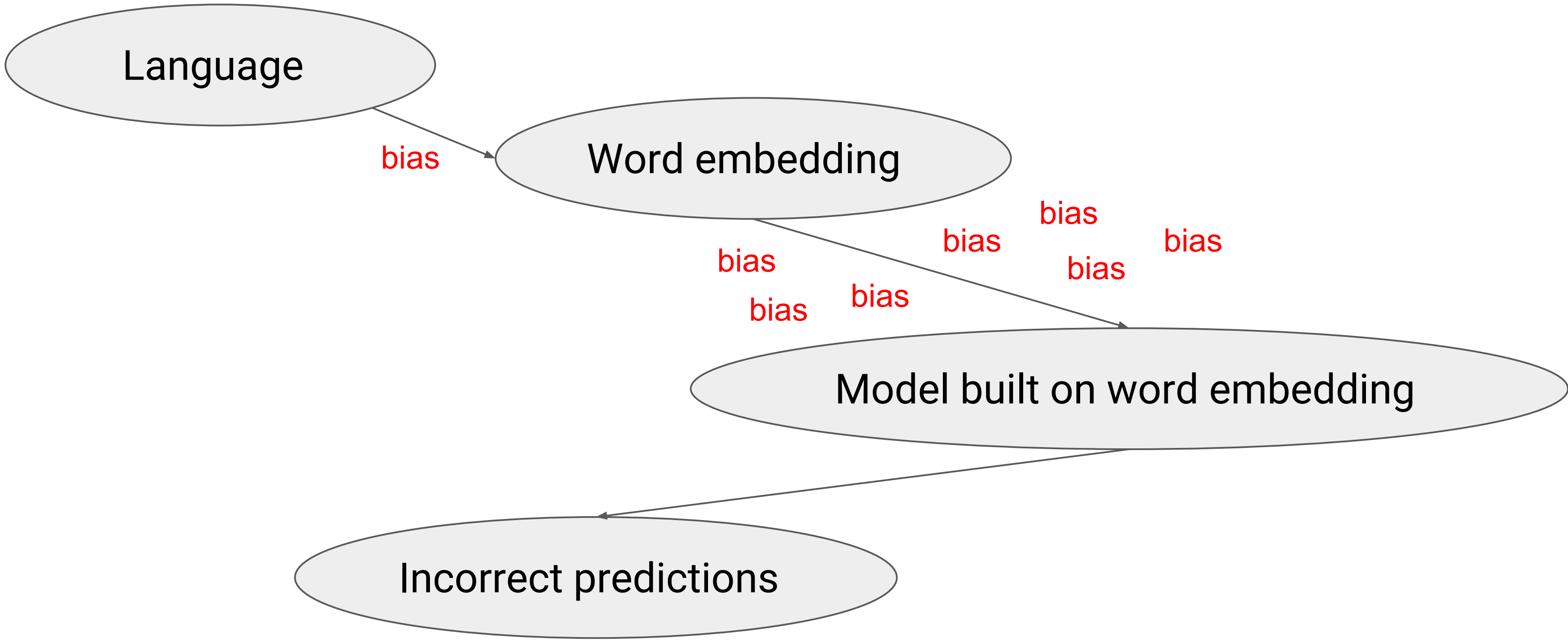
Gender neutral words with bias:

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

Problem

- **Gender bias** in a word embedding tends to be **amplified** in a model trained on this word embedding
- Training ML model on unbalanced sample is a well-known problem
- Web search example

Problem


Language
bias
Word embedding
bias
bias
bias
bias
bias
bias
bias
Model built on word embedding
Incorrect predictions

Existing solution

- Data: English *Word2vec* word embeddings
- Detect gender bias in names of occupations (*receptionist, computer programmer*)
- Reduce the bias by solving an optimization problem:

$$\min_{X \preceq 0} \|AXA^T - AA^T\|_F^2 + \lambda \|PXb^T\|_F^2 \qquad X = TT^T$$

- SVD on *A* gives

$$\min_{X \preceq 0} \|\Sigma V^T (X - I) V \Sigma\|_F^2 + \lambda \|PXb^T\|_F^2$$

T.Bolukbasi, K.-W. Chang, J.Zou, V.Saligrama, A.Kalai. 2016. [*Quantifying and Reducing Stereotypes in Word Embeddings*]

Our workflow

RusVectōrēs

Form matrix of
word embeddings *E*

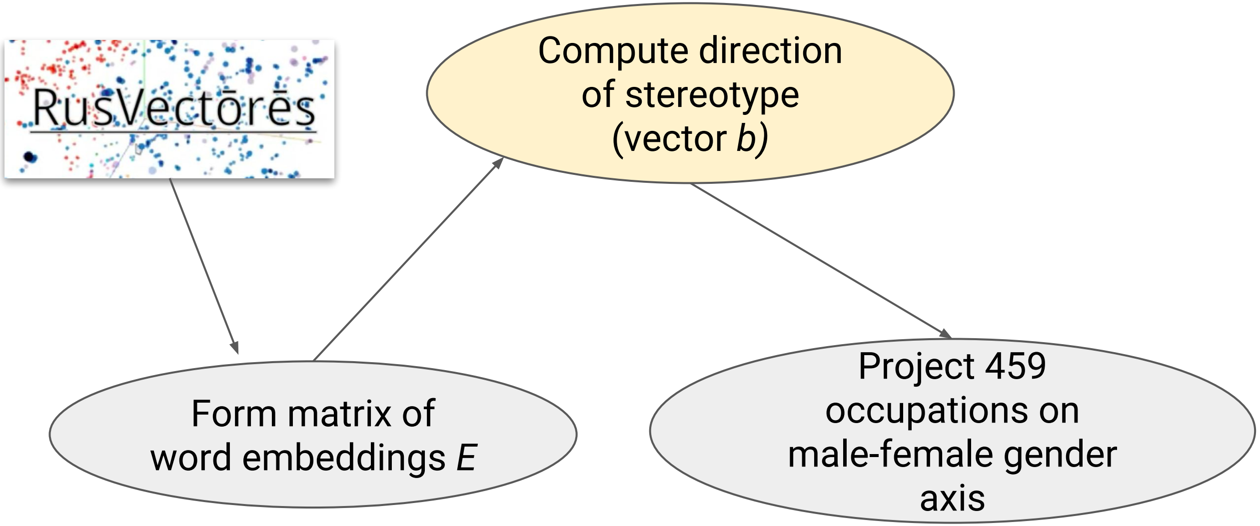
Our workflow

RusVectōrēs

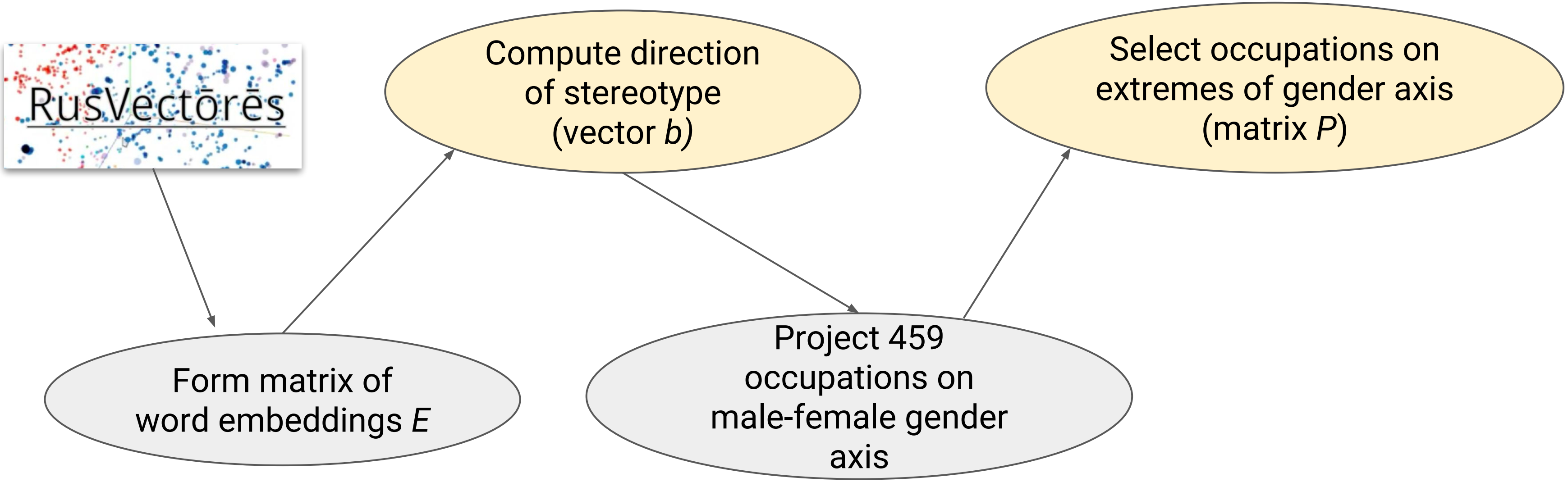
Compute direction
of stereotype
(vector *b*)

Form matrix of
word embeddings *E*

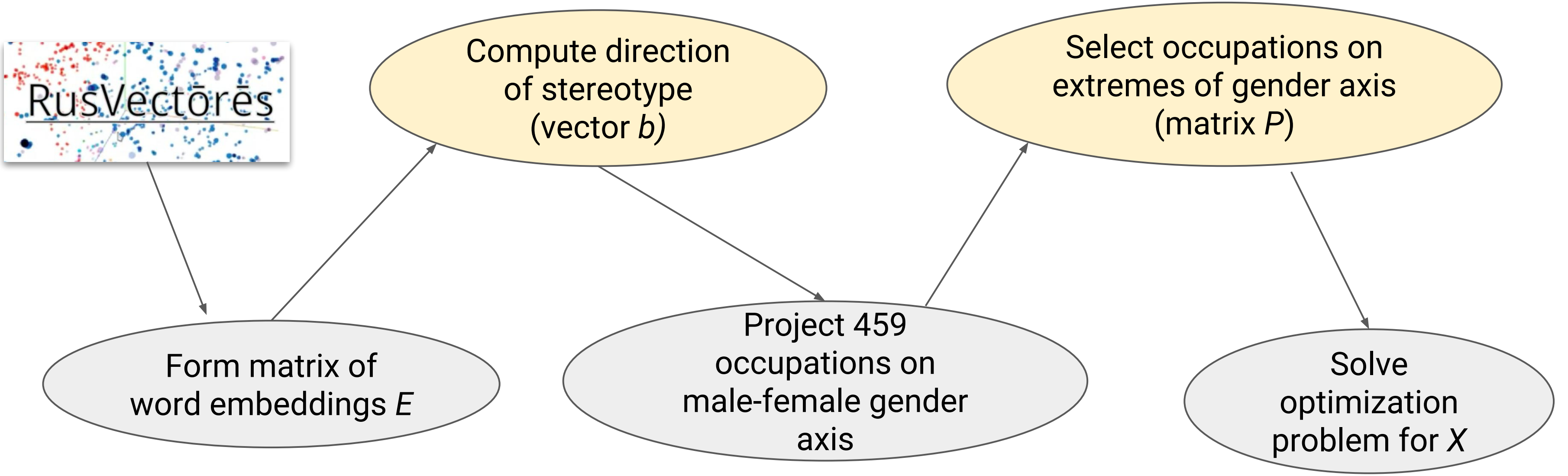
Our workflow



Our workflow



Our workflow



$$\min_{X \le 0} \|AXA^T - AA^T\|_F^2 + \lambda \|PXb^T\|_F^2$$

$$A = E - P$$

Our workflow

RusVectōrēs

Form matrix of word embeddings *E*

Compute direction of stereotype (vector *b*)

Project 459 occupations on male-female gender axis

Select occupations on extremes of gender axis (matrix *P*)

Solve optimization problem for *X*

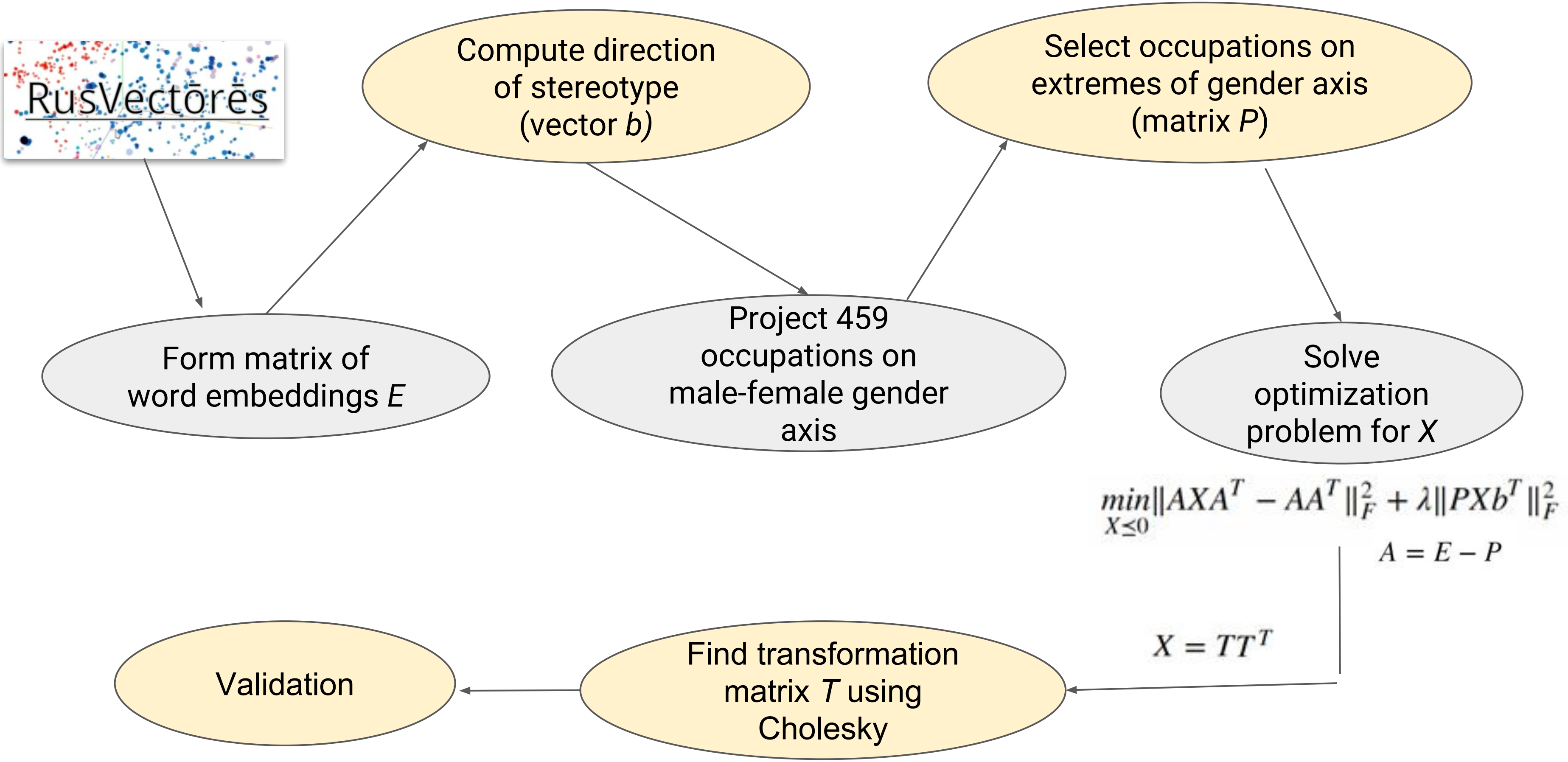
$$\min_{X \le 0} \|AXA^T - AA^T\|_F^2 + \lambda \|PXb^T\|_F^2$$

$$A = E - P$$

$$X = TT^T$$

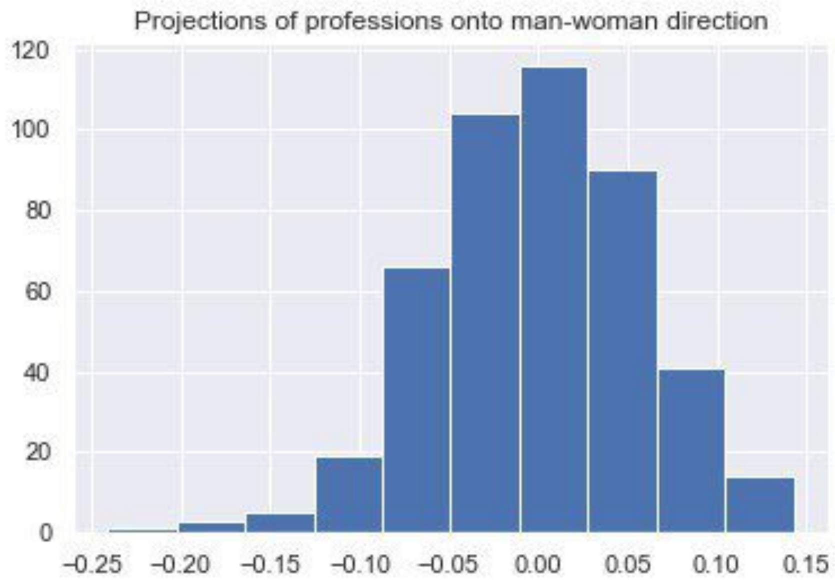
Find transformation matrix *T* using Cholesky

Our workflow



Results: bias detection

Frequencies



female *male*

Projections on gender axis

Extreme *man* occupations (1 stdev)

менеджер
шеф-повар
губернатор
директор
президент
тренер
программист

Extreme *woman* occupations (1.5 stdev)

учитель
повар
библиотекарь
корректор
художник
врач

Results: bias reduction

Variance of projections before and after transformation

decreased by **1.5** for words with bias

	Projection before transformation	**Projection after transformation**
губернатор	0.09	0.04
менеджер	0.10	0.06
шеф-повар	0.09	0.04
повар	-0.10	-0.07
библиотекарь	-0.09	-0.05

Discussion

Results:

1. Bias detected but not very extreme compared to research on w2v Google news
2. Debiasing transformation worked to a limited extent only

Conclusions for future:

1. Apply algorithm on word embeddings trained on **news** (we expect more bias)
2. Take a larger matrix P (more occupations)
3. Tune in *lambda* parameter in the optimization task

Methodology can be applied to other languages and other types of bias (e.g. racial)

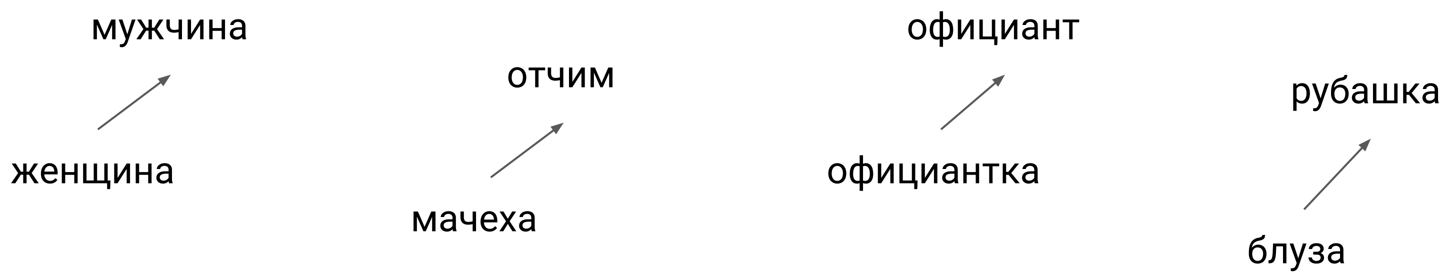
Thank you for your attention!

Our data

- Word embedding trained on the Russian National Corpus
- Corpus size: 250 million words
- Vocabulary size: 195 071
- Vector size: 300

Image source

Computing vector *b* with PCA



- Select 45 pairs of words that reflect gender opposites
- Compute 45 differences between pairs of vectors and stack them into a matrix
- Do PCA to find a principal vector - direction of stereotype

PCA returned singular values with very slow decay → we used $v_{\text{мужчина}} - v_{\text{женщина}}$

Optimization problem

$$\min_{X \succeq 0} \|AXA^T - AA^T\|_F^2 + \lambda \|PXB^T\|_F^2$$

$$X = TT^T$$

- Direction of stereotype *b*
- Matrix of biased words *P*
- Matrix of background words *A*
- Transformation matrix *T*

Validation

- Compute projections of biased words on the gender axis
- Compute the variance of these projections before and after transformation
- The variance should decrease and get close to zero

Contribution of team members

Top 3 for each where 1 is largest contribution:

Anna Koval

1. Presentation and report
2. Linguistics-related decisions
3. Programming (data preprocessing, optimization)

Ekaterina Kovalenko

1. Programming (data preprocessing, svd, optimization)
2. Math
3. Presentation and report

Anastasiia Ryzhova

1. Programming (data preprocessing, projections, optimization)
2. Math
3. Presentation and report