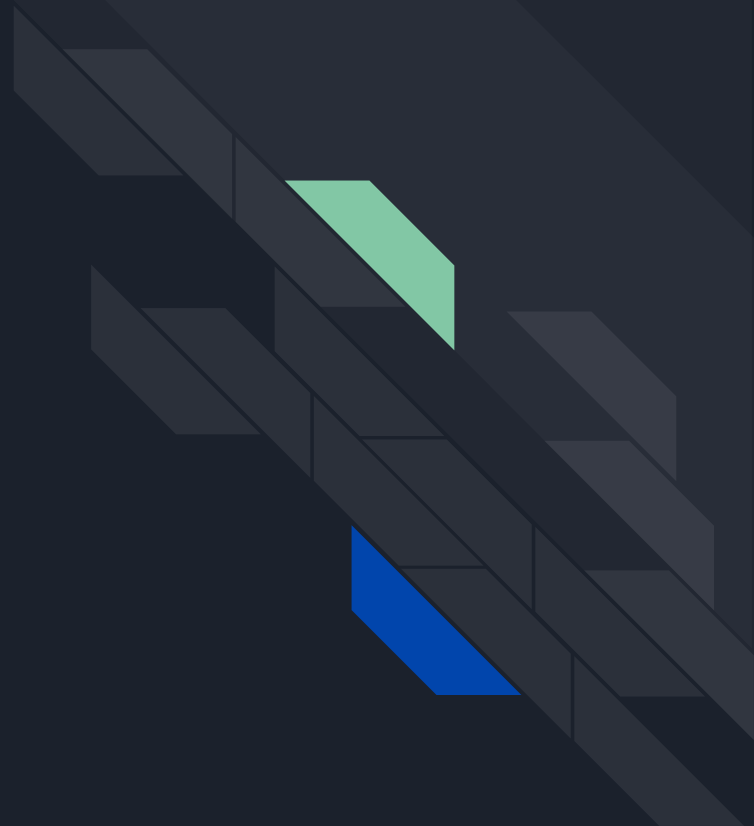


# Debiasing Word Embeddings

Abhishek Dalvi

Suhash Bollu





# Word Embeddings

- Word Embedding is a popular framework to represent text data as vectors of real numbers which has been used in many machine learning and NLP tasks.
- Word Embeddings are capable of capturing context of word in a document, semantic, syntactic similarity etc..



# Gender Bias in Word Embeddings

- Datasets generally tend to exhibit stereotypes and biases.
- Models trained using such datasets pickup and even amplify these biases.
- An example where analogy task exhibits female/male gender stereotypes to a disturbing extent.
  - *Eg.  $man - woman = computer\ programmer - homemaker$*



# Dataset and Bias Analysis

- The Word embeddings we use are GloVe 300-Dimensional Word Vectors trained on google news articles.
- We use a condensed vocabulary of the of the model having 26k words.
- Geometrically, the gender stereotype is captured by a direction in word embedding.
- Gender neutral words are linearly separable from gender definition words in word embedding.
- We have done gender bias analysis and mitigation on **occupation** only.



# Analogies to Women->Men

Words with same gender direction with thresholded distance

- Heroine ----> Hero
- Interior Designer ----> Architect
- Wig ----> Beard
- Nurse ----> Doctors
- Nanny ----> Chauffeur
- Gymnasts ----> Athletes
- Cocktails ----> Beers
- Vocalist ----> Drummers



# Analyzing gender bias in word embeddings

Cosine similarity between Word vector and Gender vector -> -ve(male biased)  
+ve(female biased)

Word	Similarity
Mathematician	-0.118
Maestro	-0.237
Sportsman	-0.1948
Librarian	0.266
Receptionist	0.273
Interior Designer	0.197



# Bias Mitigation

- Divide the 300 dimensional word embedding into 2(or more) subspace.
- The 1st subspace is the gender direction(1D).
- The 2nd subspace containing 299 dimensions is orthogonal to the gender direction.
- Neutralize gender independent word.(Doctor, Programmer)
- Equalize gender specific words.(Father~male, mother~female)



# Bias Mitigation

Doctor ●

● Nurse

Grandfather ●

● Grandmother

Beard ●

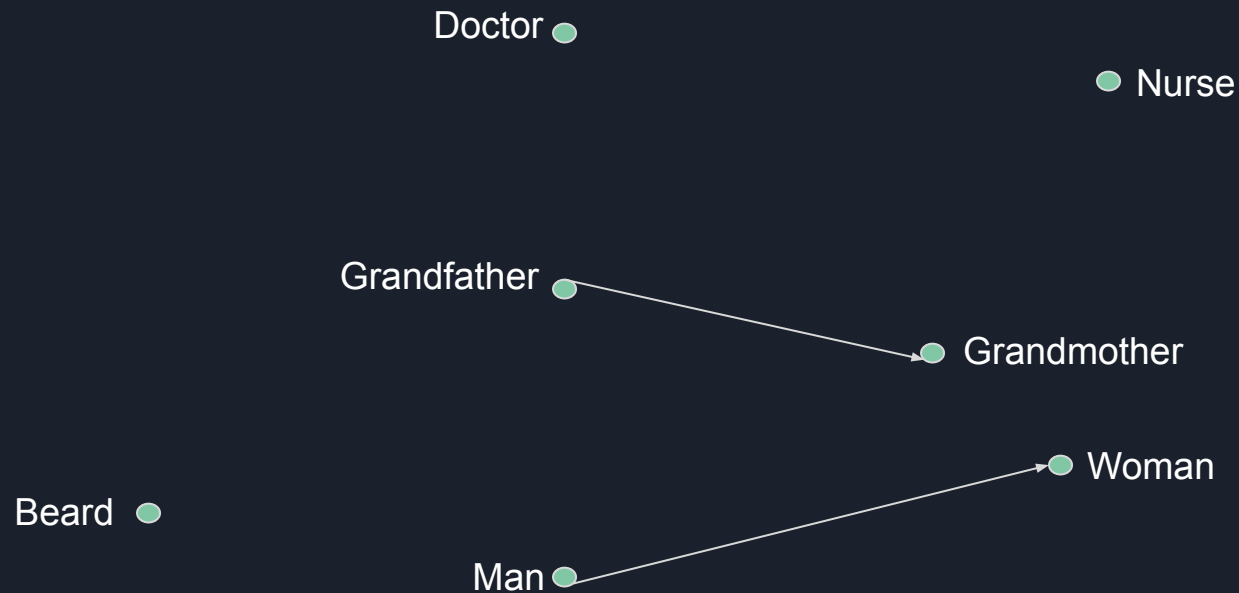
● Woman

Man ●



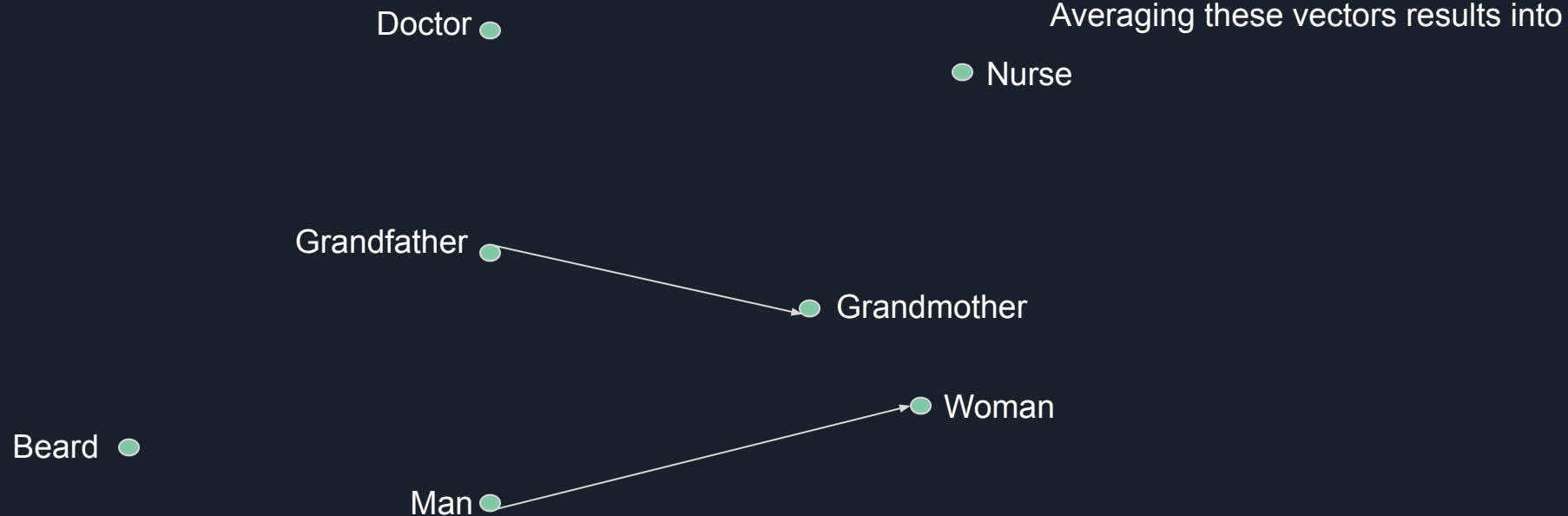


# Bias Mitigation

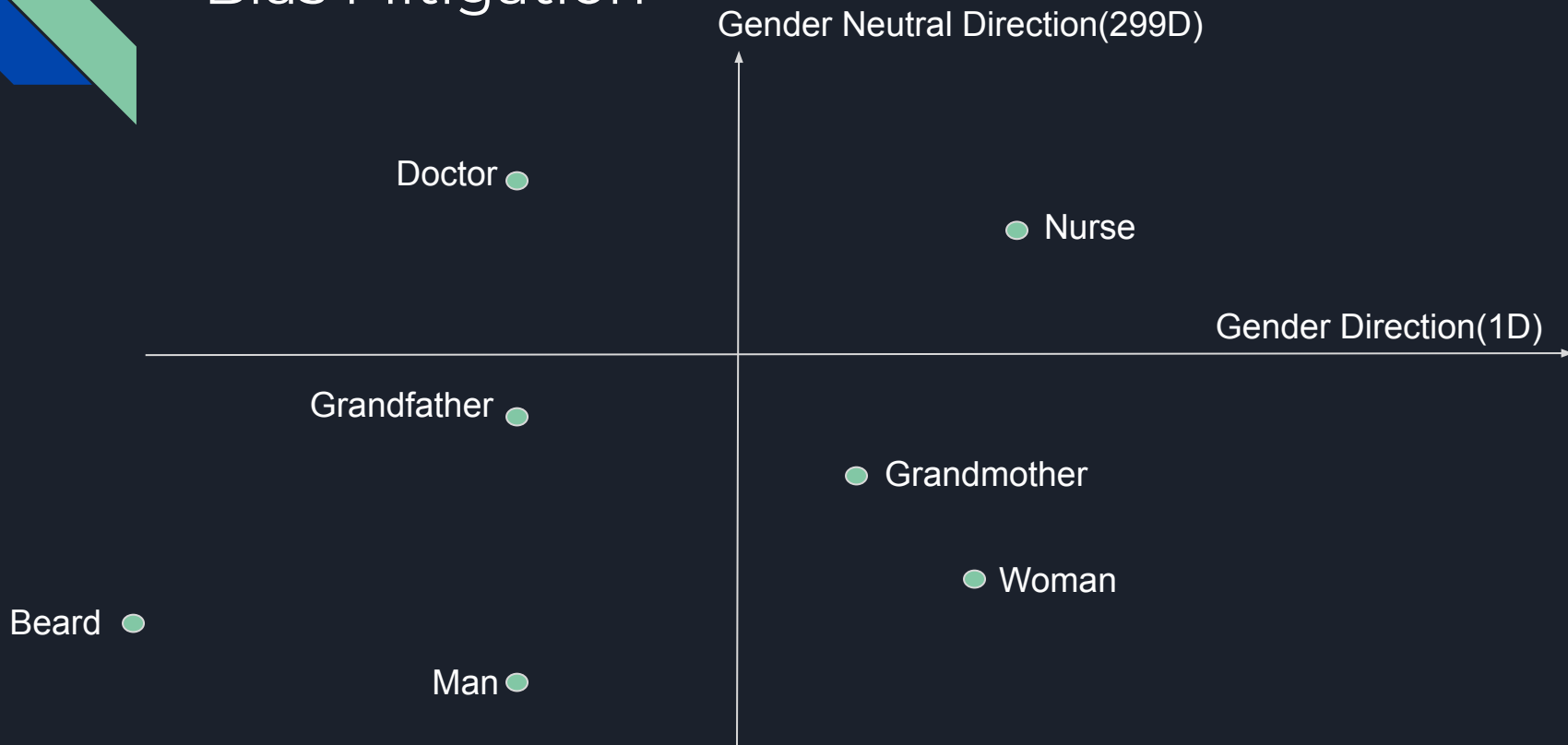




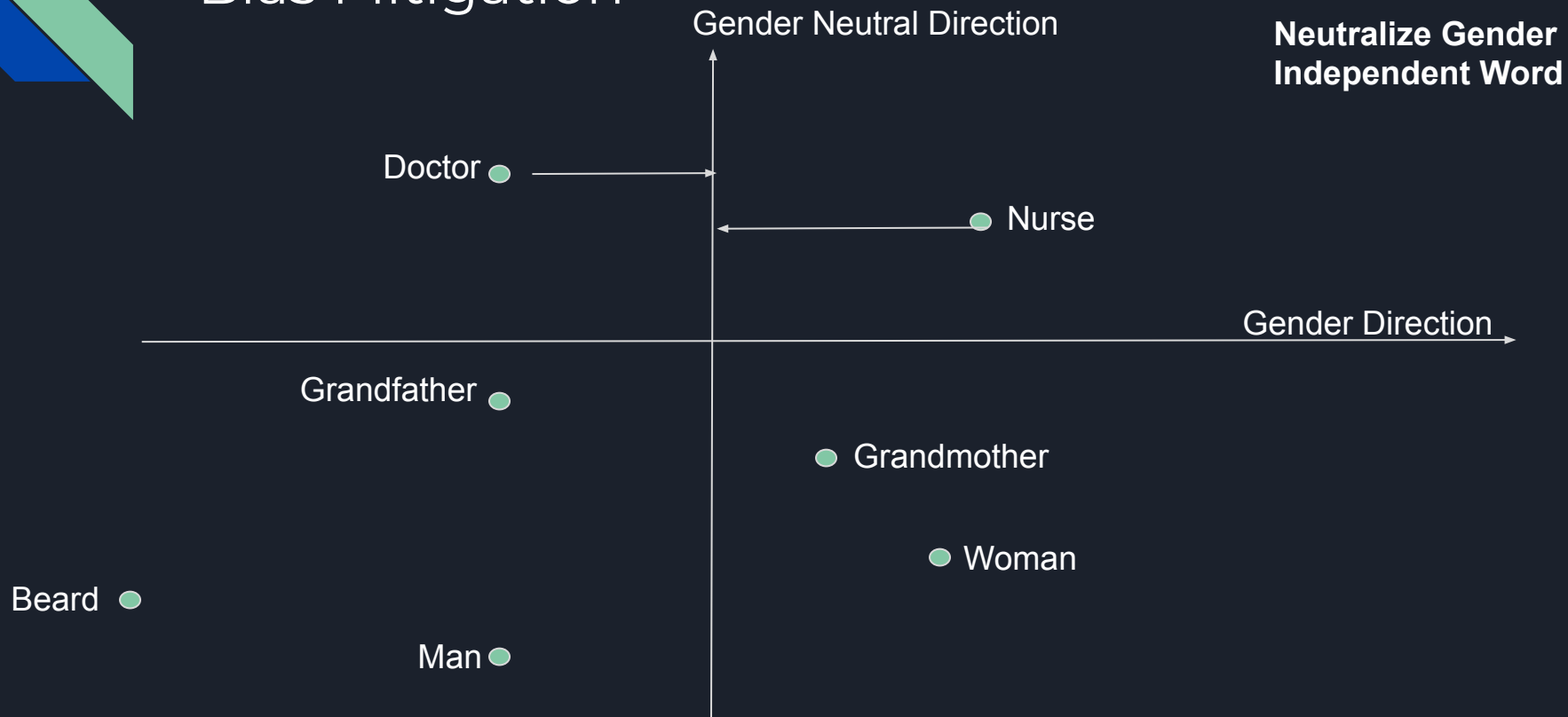
# Bias Mitigation



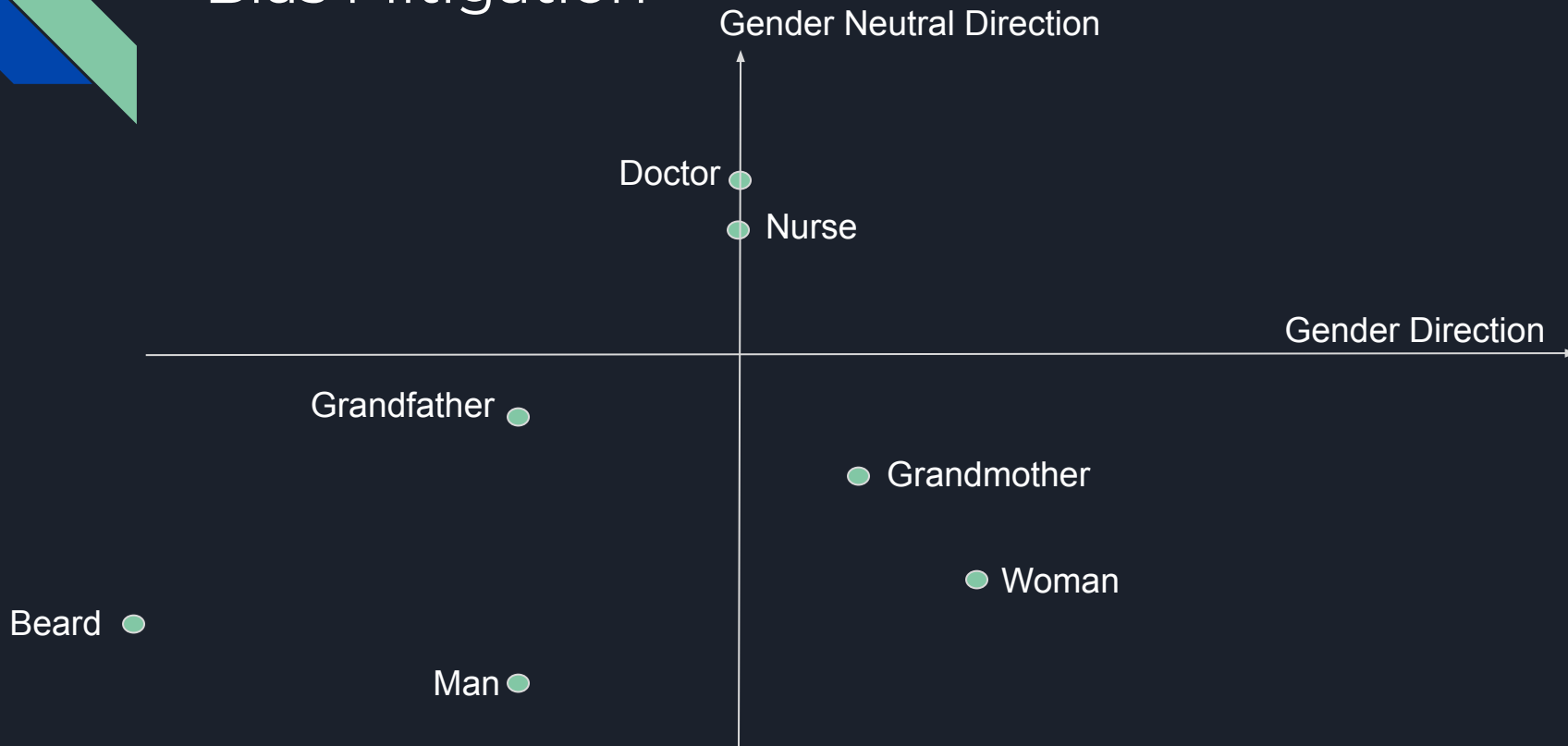
# Bias Mitigation



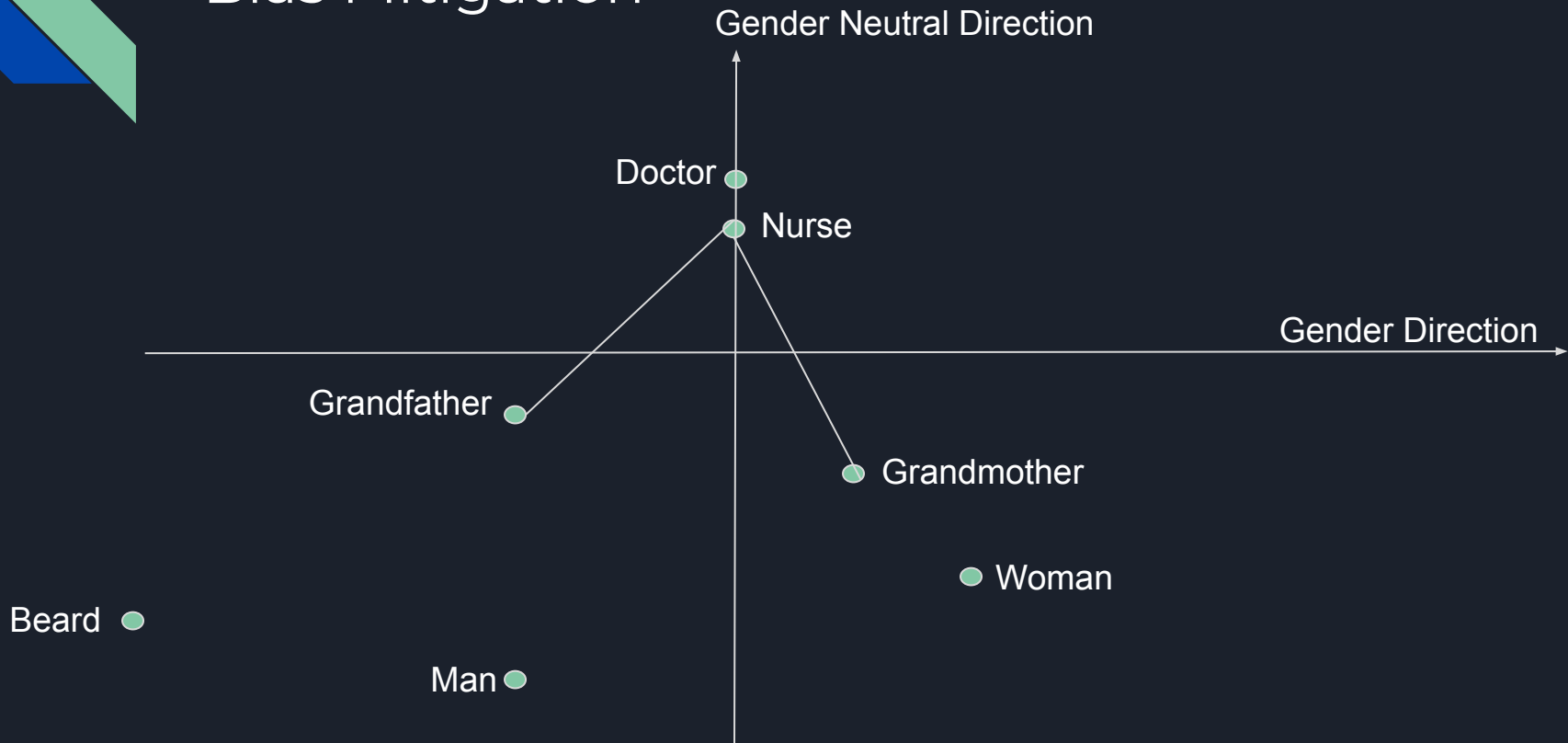
# Bias Mitigation



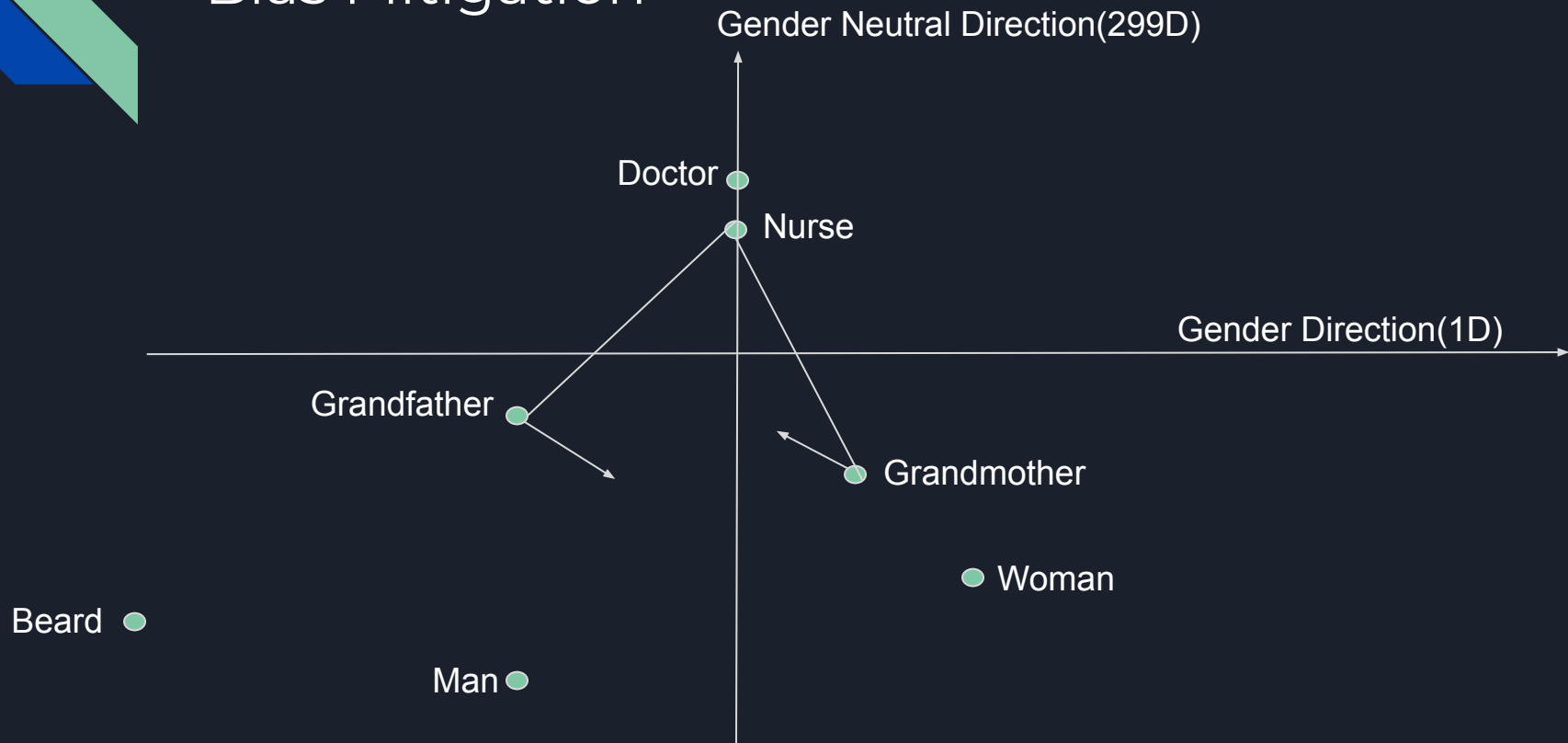
# Bias Mitigation



# Bias Mitigation



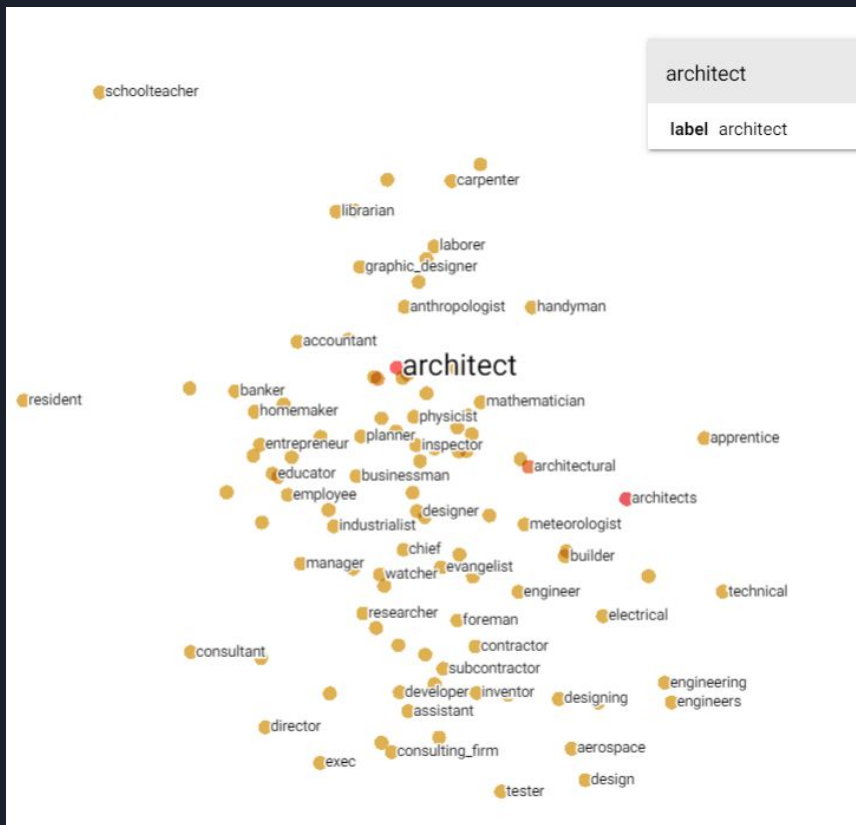
# Bias Mitigation



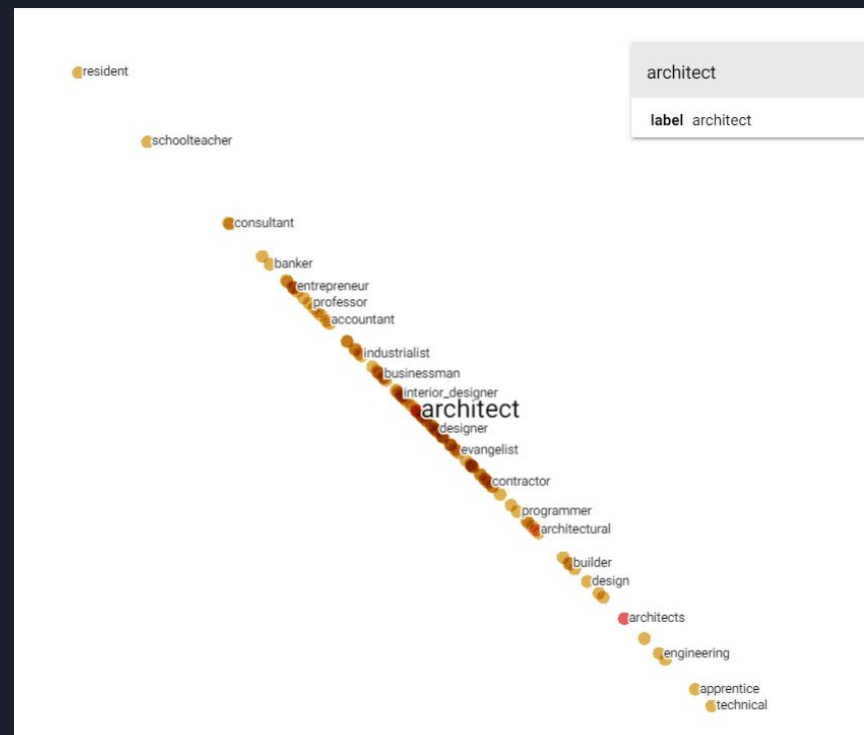
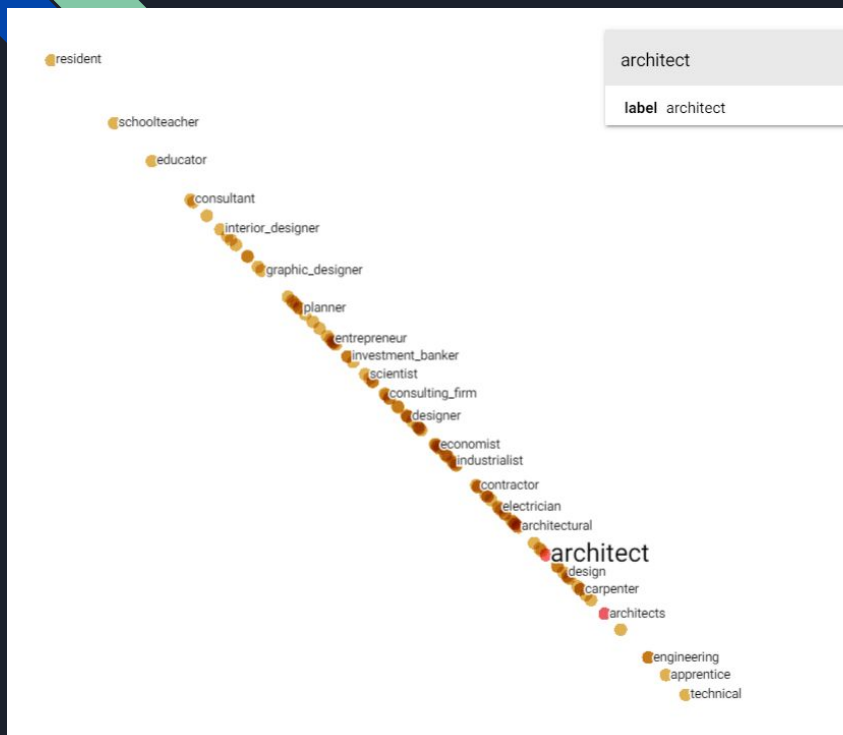



# Results





# Results(biased vs unbiased)





# Evaluating gender bias in word embeddings After Debiasing

Cosine similarity between Word vector and Gender vector -> -ve(male biased)  
+ve(female biased)

Word	Similarity(before debiasing)	Similarity(after debiasing)
Mathematician	-0.118	-0.071
Maestro	-0.237	-0.005
Sportsman	-0.1948	-0.003
Librarian	0.266	0.058
Receptionist	0.273	-0.060
Interior Designer	0.197	0.053



# Analogies to Women->Men after Debiasing

- Heroine ----> Hero
- Girls ----> Boys
- Wig ----> Beard
- Ladies ----> Gentlemen
- Hens----> Chicken
- Queens----> Kings
- Tamika----> Kareem



Thank you!