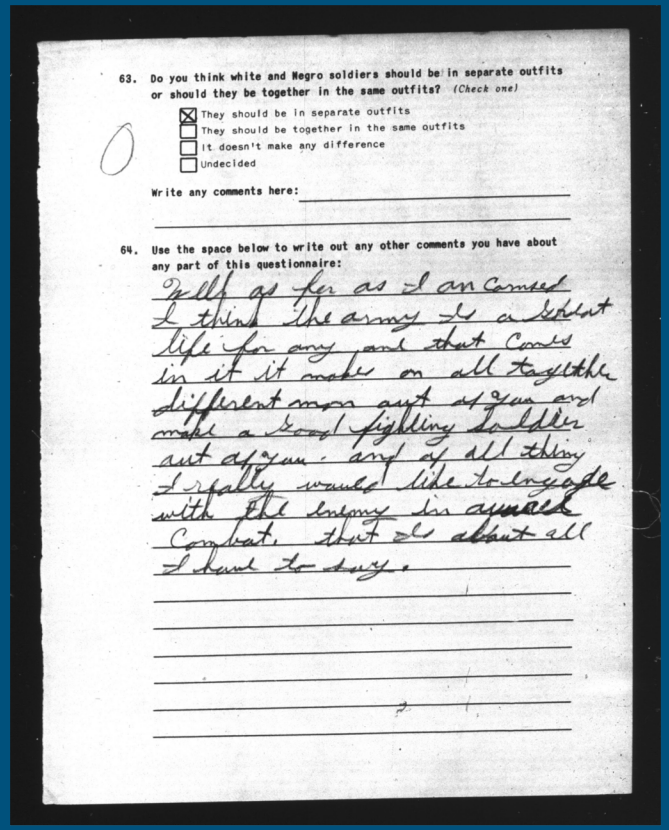
BERT

The AI'd Powers
Chandler Crescentini, Colin Lagator, Quinn Moore

Why BERT?

- **65,000** free response survey answers
  - Transcribed by the public
- Too much to analyze by hand
- Computers require numbers as inputs
- How do we convert the words to numbers?



63. Do you think white and Negro soldiers should be in separate outfits or should they be together in the same outfits? *(Check one)*

- [x] They should be in separate outfits
- [ ] They should be together in the same outfits
- [ ] It doesn't make any difference
- [ ] Undecided

Write any comments here:

64. Use the space below to write out any other comments you have about any part of this questionnaire:

Well as far as I am concerned I think the army is a decent life for any one that comes in it it make on all tryithe different man out of you and make a good fighting soldier out of you. and of all thing I really would like to engage with the enemy in a attack Combat. that is about all I have to say.

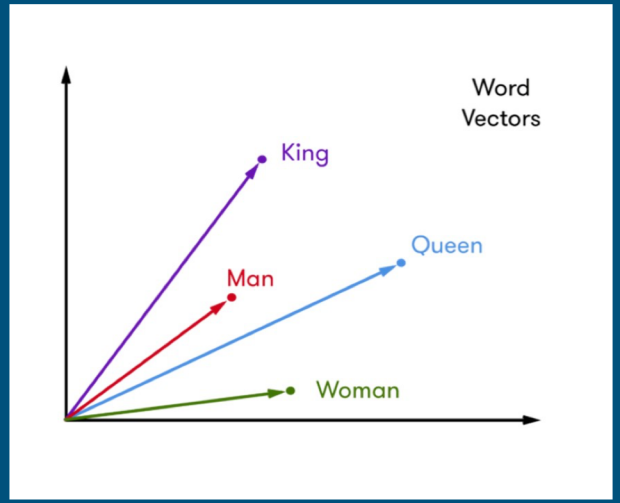
Example free response from Survey-32W Attitudes Towards Black Soldiers

Vector Embeddings of Language

- Representing language as numbers
- **“Stick”**

-0.34	-0.84	0.20	-0.26	-0.12	….

  - First five rounded values of the vector embedding for “stick”



https://www.sentiance.com/2018/01/29/unlabeled-visits/

- “King” + “Man” = “Queen” + “Woman”
- **Problem**: meaning changes depending on context!

Language Embedding Techniques

- **ELMo (2017)**
  - Birth of contextualized vector embeddings
- **ULM-FiT (2018)**
  - Introduced ability to fine-tune to specific tasks
- **OpenAI Transformer (2019)**
  - Model pre-trained using diverse unlabeled data

Bidirectional Encoder Representations from Transformers (BERT)

- Huge leap forward for Natural Language Processing (NLP)
- Open source and pre-trained using the internet
- Released by Google AI, October 2018
- Very large model

http://s3.amazonaws.com/images.seroundtable.com/google-bert-algorithm-update-1572000459.jpg

BERT Training

- **Masked Language Model**
  - Bidirectional conditioning with masks

Input: The man went to the $[MASK]_1$. He bought a $[MASK]_2$ of milk .
Labels: $[MASK]_1$ = store; $[MASK]_2$ = gallon

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

BERT Training

- **Two-sentence tasks**
  - Given two sentences A and B, is B the sentence that follows A?

Sentence A = The man went to the store.
Sentence B = He bought a gallon of milk.
Label = IsNextSentence

Sentence A = The man went to the store.
Sentence B = Penguins are flightless.
Label = NotNextSentence

https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

Applications

- Sentence classification
  - **Input**: email message
  - **Output**: spam/not spam
- Sentiment analysis
  - **Input**: movie/product review
  - **Output**: positive/negative
- Fact checking
  - **Input**: sentence
  - **Output**: claim/not claim

BERT (# of tokens)	SciBERT (# of tokens)	BioBERT (# of tokens)
English Wikipedia: 2.5B BooksCorpus: 0.8B	Biomedical paper: 2.5B Computer Science paper: 0.6B	English Wikipedia: 2.5B BooksCorpus: 0.8B PubMed Abstracts: 4.5B PMC full text: 13.5B

https://towardsdatascience.com/how-to-apply-bert-in-scientific-domain-2d9db0480bd9