

The American Soldier in World War II: Text Analytics Potential of Historical Survey Data

• • •

The AI'd Powers

Chandler Crescentini, Colin Lagator, Quinn Moore

Dr. Ed Gitre, Virginia Tech Department of History



CALL OF DUTY®
WWII

<https://www.gamespot.com/articles/the-surprising-way-call-of-duty-wwii-devs-captured/1100-6470488/>

<https://www.amazon.com/Saving-Private-Ryan-Tom-Hanks/dp/B008CBZX7L>

<https://images.app.goo.gl/z3kMTMoBtu91QMBFA>



<https://images.app.goo.gl/DYykD73QZC3xXbds6>



<https://images.app.goo.gl/xgfaUlUm7c47AMw69>

Survey S-106:
Postwar Job Plans of Officers
and Enlisted Men.

RESTRICTED

(60-62)

46. After the war, do you plan to go back and live in the same place — the same local area, say within 5 or 10 miles of where you lived before you entered the Army? (Check one)

- Yes, I plan to live there for a long time to come
- Yes, but only for a short time
- No
- Undecided

47. Up to now, how much had you thought about exactly what you will want to do after the war? (Check one)

- I had thought a good deal about it
- I had thought some about it
- I had hardly thought about it

48. (a) How much has serving in the Army changed the plans for your life that you had before you entered the Army? (Check one)

- It has delayed carrying out my plans but hasn't changed them much
- It has delayed carrying out some plans and has changed others
- It has changed most of my plans
- I really didn't have any very definite plans

(b) If some of your plans have been changed, what plans have been changed as a result of your Army service?

Army life has made me consider
settling down on a farm rather than
changing location every few years as
a teacher.

49. We would like you to use this and the following page to write anything else you have to say on any of the following subjects:

What you think you will do once you leave the Army.
What you would like to know about the post-war Army.
Whether you will want to get permanently established as soon as possible or will take a short-term job when you leave the Army.

I want to know in detail about educational help for veterans. I want to avail myself of such help and then settle down somewhere.
(continue writing on the next page)

Survey S-106:
Postwar Job Plans of Officers
and Enlisted Men.

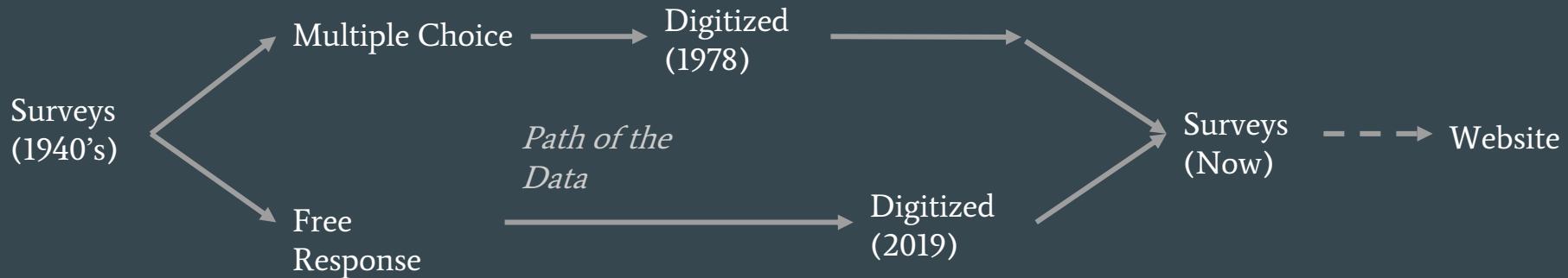
(b) If some of your plans have been changed, what plans have been changed as a result of your Army service?

Army life has made me consider
settling down on a farm rather than
changing location every few years as
a teacher

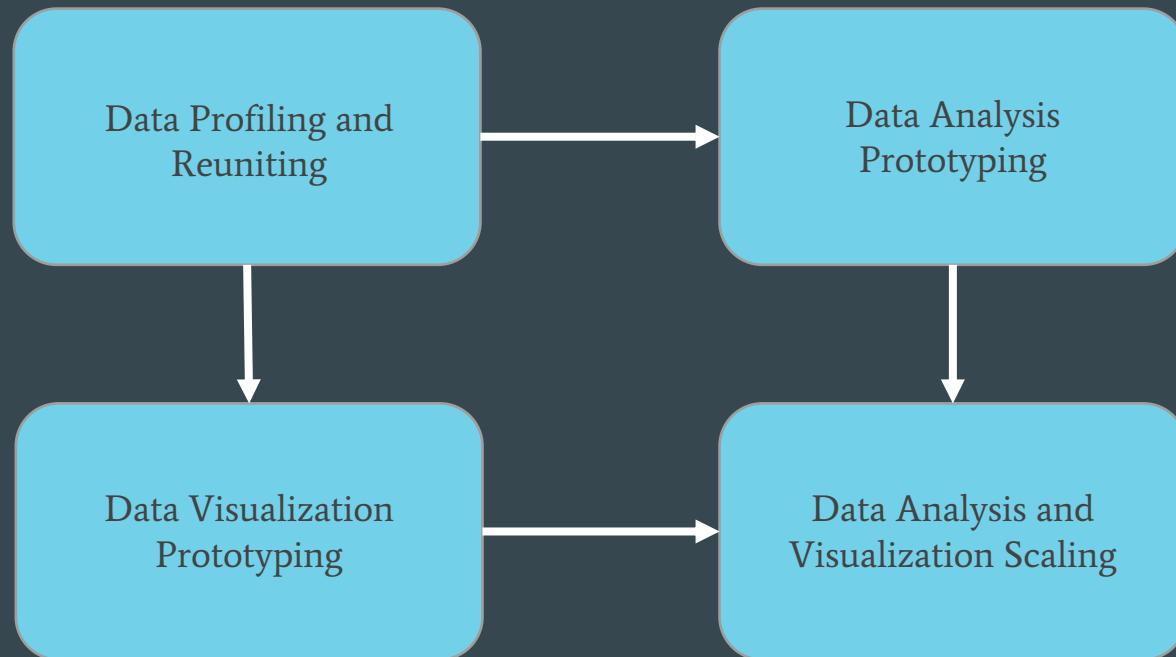
“Army life has made me consider settling down on a farm rather than changing
location every few years as a teacher”

Problem Statement

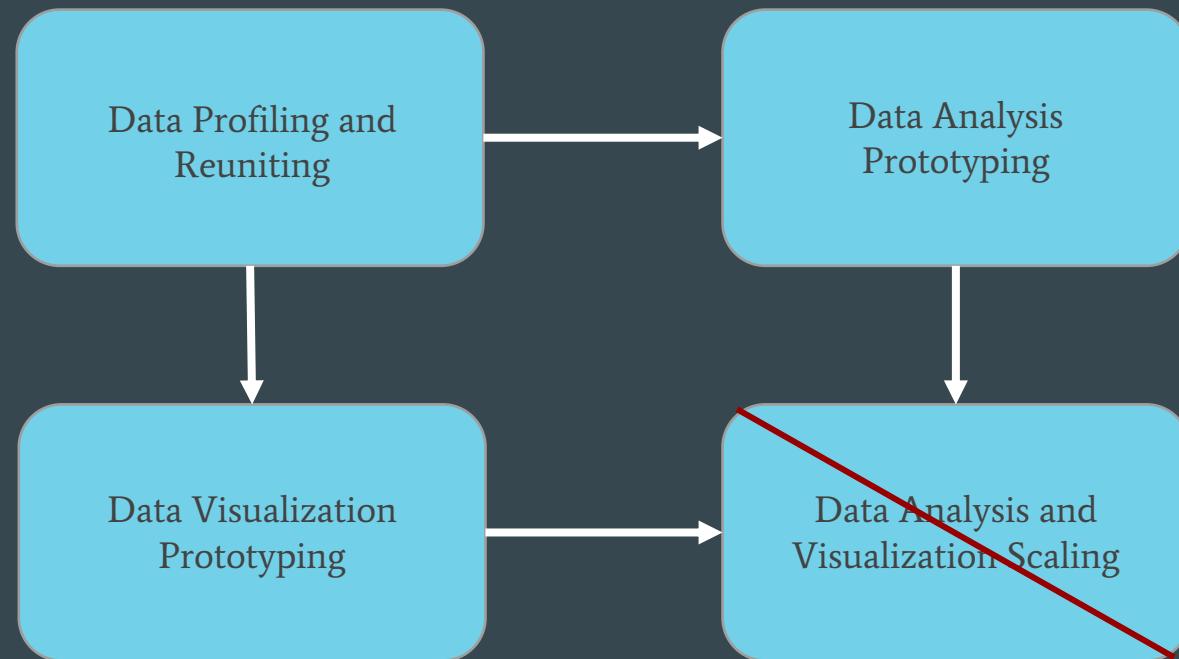
- Client's research goal is a public website
- Our role is to provide the tools
- Reunite multiple choice and free response sections
 - Structured and unstructured
- Demonstrate the potential of Natural Language Processing (NLP)



Project Components - Original



Project Components - Revised



Literature Review

- “Impact of Anonymity on Responses to Sensitive Questions”
 - Ong and Weiss (2000)
- “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”
 - Devlin, Chang, et al. (2018)
- “CUR Matrix Factorizations: Algorithms, Analysis, Applications”
 - Sorensen and Embree (2016)

Description of Solution Approach

- Reunite multiple choice and free response answers
 - Serial Numbers
- Natural Language Processing
 - Sentiment Analysis
 - N-grams
 - Bidirectional Encoder Representations from Transformers
 - CUR Decomposition

Data Reuniting

- Multiple choice and transcribed handwritten free response portions
 - Survey 106 - Post War Job Plans of Officers and Enlisted Men
- Match five digit serial numbers - not so simple
 - Needed to convert between the two forms

NUMBERING SCHEME FOR S-106 CPA BOTH OFFICERS AND
ENLISTED MEN

Cols. 1 & 2 = Branch of Service

<u>Branch</u>	<u>Classification for Both Officers & Enlisted Men</u>		<u>Col. 1</u>	<u>Col. 2</u>	<u>Total Cases</u>	<u>E. M.</u>	<u>Officers</u>
Infantry			1	1,2	646		97
Field Artillery			1	3	163		34
Coast Artillery			1	4	108		27
Anti-Aircraft			1	5	158		36
Engineers			2	1	312		68
Quartermaster			2	2	268		54
Ordnance			2	3	313		63
Signal Corps			2	4	281		65
Transportation Corps			2	5	47		-
Military Police			2	6	108		-
Chemical Warfare			2	7	69		-
Medical Department			2	8	304		83
Miscellaneous (mostly Hdqts. organizations)			2	9	326		30
Air Corps - Flying personnel			3	1,2,3	518		341
Air Corps - Ground personnel			3	5,6,7	1492		156

Col. 3 = Separate outfits within each Branch

Cols. 4 & 5 = Serial number within outfit.

TC 5113 1054

Branch	Roper Code	Orig. Code 1	Orig. Code 2
Infantry	01	1	1,2
Field Artillery	02	1	3
Coast Artillery	03	1	4
Anti-Aircraft	04	1	5
Engineers	05	2	1
Quartermaster	06	2	2
Ordnance	07	2	3
Signal Corps	08	2	4
Transportation Corps	09	2	5
Military Police	10	2	6
Chemical Warfare	11	2	7
Medical Department	12	2	8
Miscellaneous	13	2	9
Air Corps - Flying	14	3	1,2,3
Air Corps - Ground	15	3	5,6,7
No Code or No Data	00	0	0

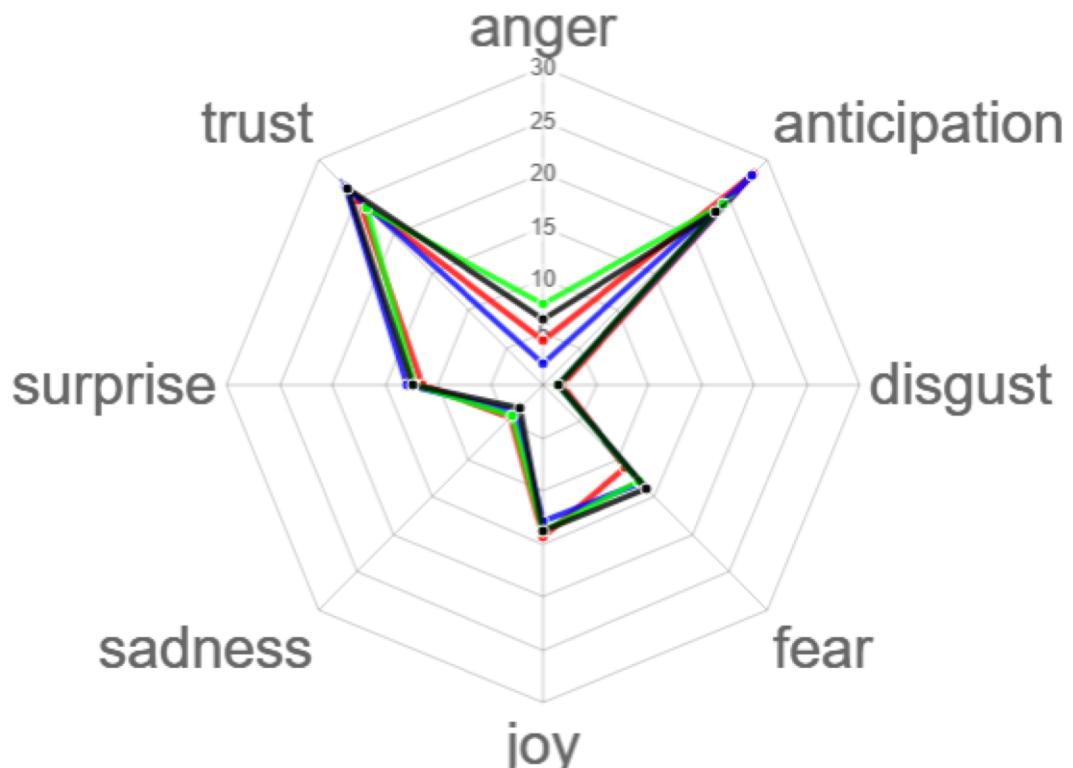
28463 → 12463

Sentiment Analysis

- Measure and categorize text data for attitudes and emotions.
- We wanted to do this to leverage the reunited survey.
 - Multiple choice and free response are paired
- We look at sentiment segmented by a multiple choice answer.
 - Group by age, pre-war job, branch, etc.

Sentiment Regarding Post-War Job Plans

Grouped by Combat Experience



"Have you been in combat or under enemy fire in this war?"

Yes, I have been in actual combat

Yes, I have been under enemy fire, but not actual combat.

No.

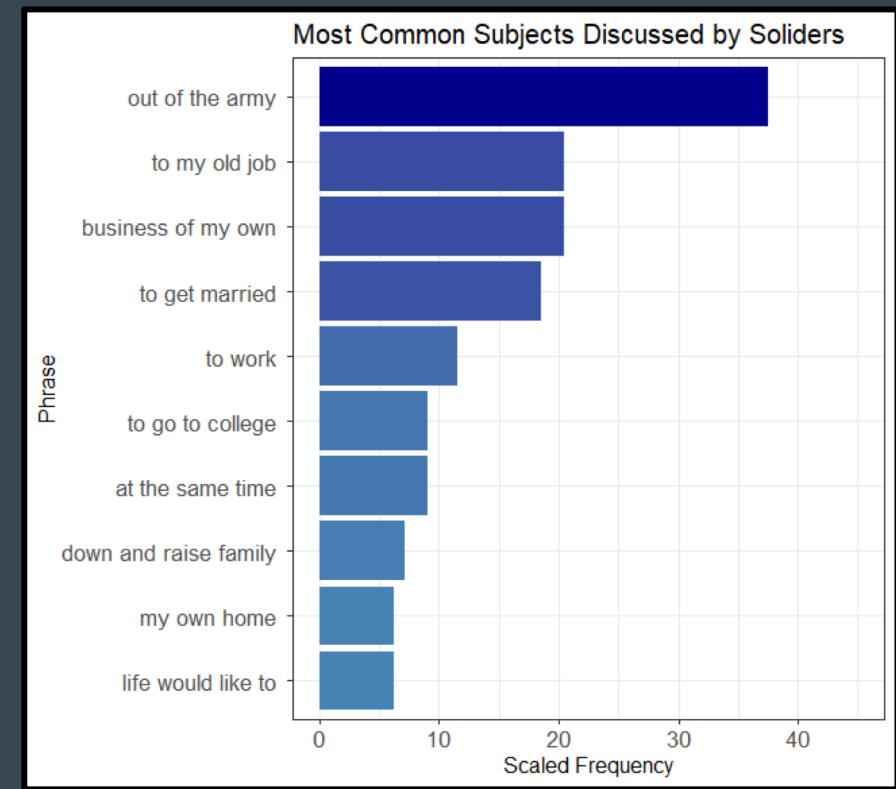
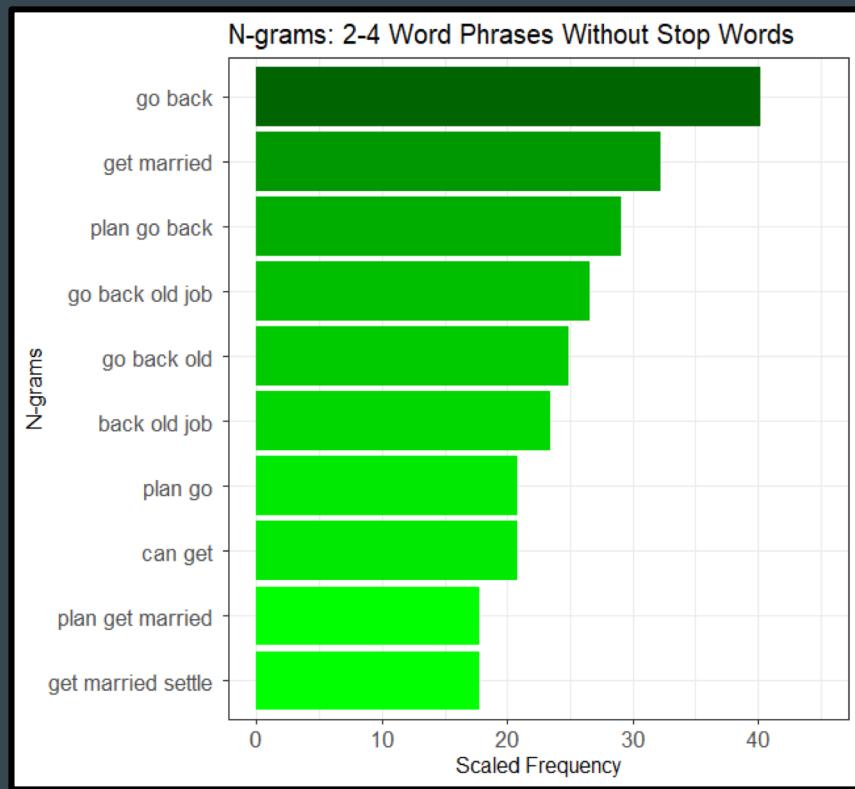
No answer.

N-grams

- Frequency of multiple adjacent words
 - “n” indicates number of words chained together
- Phrases commonly used by soldiers
- Scaled frequency of n-gram types

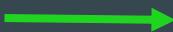
N	The American Soldier in WW2
1	<ul style="list-style-type: none">• The• American• Soldier• In• WW2
2	<ul style="list-style-type: none">• The American• American Soldier• Soldier in• in WW2
3	<ul style="list-style-type: none">• The American Soldier• American Soldier in• Soldier in WW2

N-grams



Bidirectional Encoder Representations from Transformers (BERT)

- Released by Google AI, October 2018
- Open source and pre-trained using the internet
- Handles context better than previous models
- Used to create vector embeddings of language

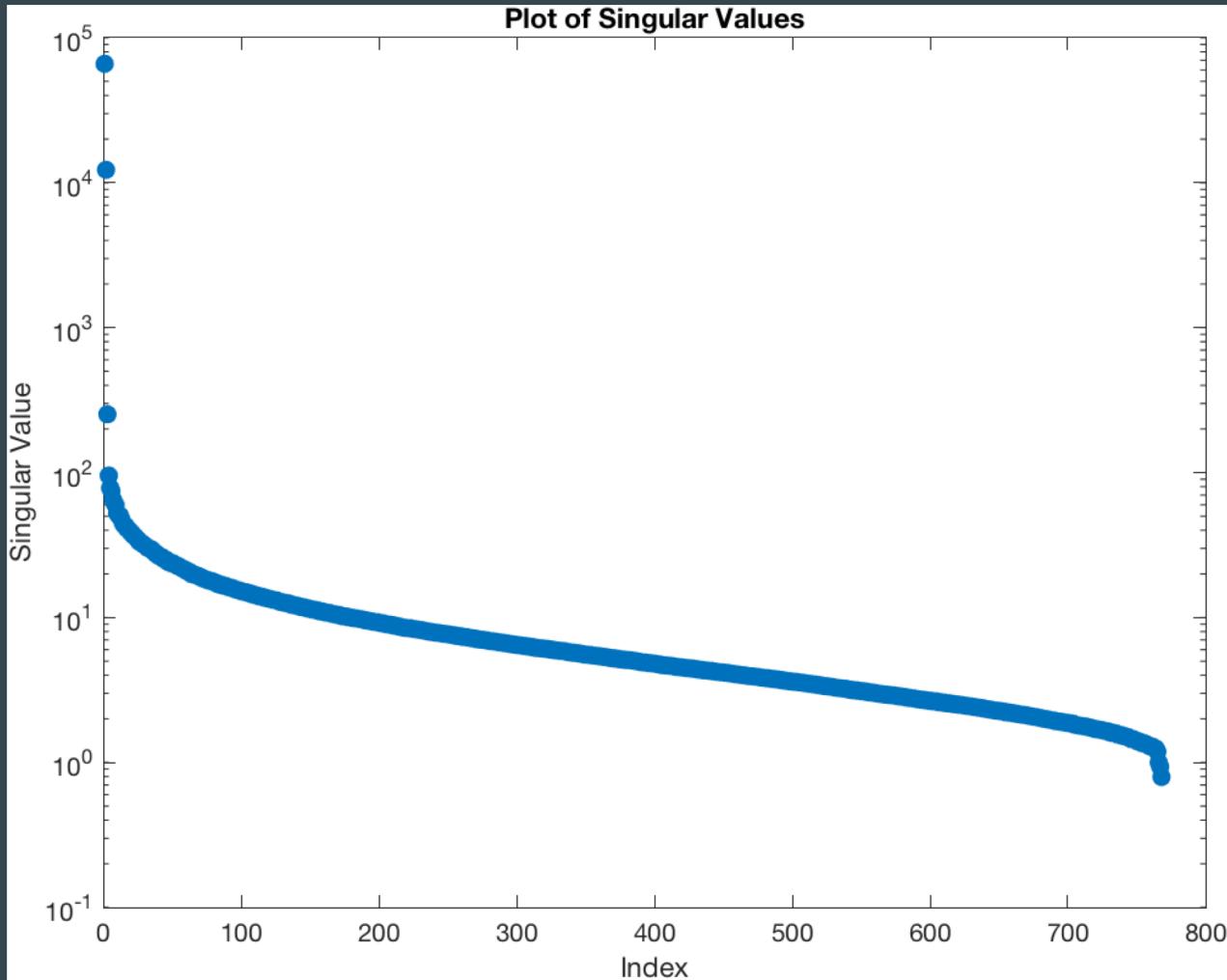
“Stick” 

-0.34	-0.84	0.20	-0.26	-0.12
-------	-------	------	-------	-------	------

CUR Decomposition

- Novel approach to text analytics
- Low-rank approximation similar to singular value decomposition (SVD)
- Used to show representative examples

$$\begin{array}{ccccc} \mathbf{A} & \approx & \mathbf{C} & \mathbf{U} & \mathbf{R} \\ m \times n & & m \times k & k \times k & k \times n \end{array}$$



CUR Decomposition Results

- My plans for after the war is to work for my factory, and in spare time do carpenter work.
- I am undecided
- All men who are qualified should have first chance with Air Lines.
- I wanna get home right now- this instant.
- Get married and build a home and raise a family.

Limitations

- Scale and variety of the data
- Surveys require different reuniting techniques
- ~62% of responses reunited with serial numbers
- Lexicon bias affects sentiment

Lessons Learned

- Interfacing with client and the work of others
- Iterating through analyses
- Looking deeper into results

Acknowledgements

- Dr. Mike Hughes
- Aaron Schroeder
- Gizem Korkmaz
- Dr. Ed Gitre
- Xavier Pleimling
- Dr. Mark Embree
- Brian Sanchez

<https://www.zooniverse.org/projects/tkotwim/the-american-soldier>

Conclusions and Future Work

- Data reuniting
- Segmenting of combat/non-combat troops
- Natural Language Processing
- CUR decomposition