

# The Humility Project

## Word Counts, Clustering, and Search Engine

Tyler Perini, Undergraduate

Dr. Amy Langville, Mathematics

Dr. Jen Wright, Psychology

Dr. Thomas Nadelhoffer, Philosophy

- 1) Motivation**
- 2) Data Collection**
- 3) Comparative Analysis**
- 4) Topic Analysis**
- 5) Predictive Analysis**
- 6) Conclusions**

# Motivation

# Templeton Grant for Humility Project: Research Objectives

- (1) Develop and validate a new psychometric tool—the Humility Scale (HS)—that measures humility along several dimensions
  - (2) Explore Humility from a developmental perspective
- (3) Explore the interaction between humility and conformity or susceptibility to peer pressure
- (4) Analyze a set of legendary and living moral exemplars and appropriate comparison groups

# Templeton Grant for Humility Project: Research Objectives

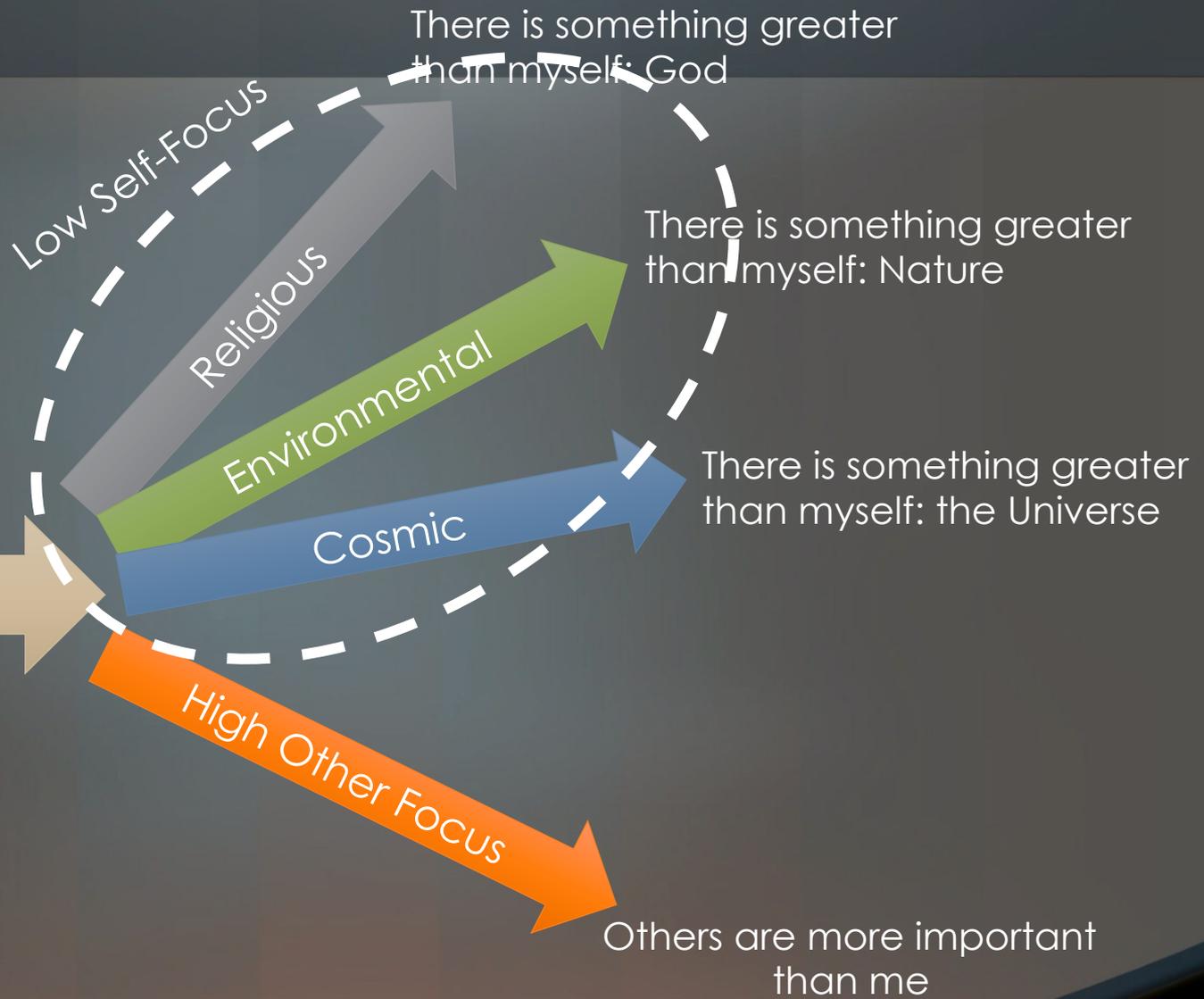
- (1) Develop and validate a new psychometric tool—the Humility Scale (HS)—that measures humility along several dimensions
  - (2) Explore Humility from a developmental perspective
- (3) Explore the interaction between humility and conformity or susceptibility to peer pressure
- (4) Analyze a set of legendary and living moral exemplars and appropriate comparison groups
  - Develop an empirical tool that assesses the humility of people by way of analyzing existing speeches

# Research Objective #4

- (1) Find statistically significant patterns in the language usage of individuals that correlate to his or her own humility score
- (2) Refine pattern identification for use in accurately predicting humility scores from text sources
- (3) Implement methods that use an open-vocabulary approach for more natural analysis of language usage
- (4) Visually display findings in a clean and efficient diagram that is both appealing and informative

# Data

Humility



# Humility Scale

Cosmic:

- I often find myself pondering my smallness in the face of the vastness of the universe.
- I often think about the fragility of existence.

# Humility Scale

## Cosmic:

- I often find myself pondering my smallness in the face of the vastness of the universe.
- I often think about the fragility of existence.

## Environmental:

- Humans have to learn to share the Earth with other species.
- We should always try to be in harmony with Mother Nature.

## Religious:

- Ultimately, there is a Supreme Being who gets all of the credit and glory for our individual accomplishments.
- I accept my total dependence upon the grace of God.

## Other Focus:

- My friends would say I focus more on others than I do myself.
- I care about the welfare others, at times more than my own welfare.

# Data Collection

Using Mturk, we ask volunteer surveyors the following questions which target their beliefs to each facet of humility:

- 1) When you reflect on your life, broadly speaking, how would you best describe your relationship with the surrounding universe/cosmos -- and your beliefs/attitudes about that relationship?

# Data Collection

Using Mturk, we ask volunteer surveyors the following questions which target their beliefs to each facet of humility:

- 1) When you reflect on your life, broadly speaking, how would you best describe your relationship with the surrounding universe/cosmos -- and your beliefs/attitudes about that relationship?
- 2) ... God or a creator/supreme being/higher power...?
- 3) ... earth/environment/"mother nature"...?
- 4) ... fellow human beings...?

In addition, we have them complete the humility scale.

# Small Data vs. Big Data

- Aim to have the best and cleanest training set possible.
- We weren't sure that those who scored highly on the humility scale necessarily *wrote* humbly, or vice versa.

# Small Data vs. Big Data

- Aim to have the best and cleanest training set possible.
- We weren't sure that those who scored highly on the humility scale necessarily *wrote* humbly, or vice versa.
- By hand, we went through hundreds of volunteer responses and categorized them as Humble, Not Humble, or Neutral.

<b>Cosmic</b>	<b>Environ.</b>	<b>Religious</b>	<b>Other Focus</b>
93%	96%	100%	78%

# Small Data vs. Big Data

- Aim to have the best and cleanest training set possible.
- We weren't sure that those who scored highly on the humility scale necessarily *wrote* humbly, or vice versa.
- By hand, we went through hundreds of volunteer responses and categorized them as Humble, Not Humble, or Neutral.

<b>Cosmic</b>	<b>Environ.</b>	<b>Religious</b>	<b>Other Focus</b>
93%	96%	100%	78%

Even with high reliability percentages, we noticed that within a single response, there are humble and not humble sentences. This is how we decided on a *sentence-by-sentence* analysis.

# Coding for Humility

## Humble

- Everybody is made of “star stuff”, which really makes you feel like a part of the universe.

## Not Humble

- I had the realization that everything exists for my benefit.

# Coding for Humility

## Humble

- Everybody is made of “star stuff”, which really makes you feel like a part of the universe.
- I feel that as a member of the human race it is my duty to protect the earth and do what I can to be nice to “her”.

## Not Humble

- I had the realization that everything exists for my benefit.
- I believe that I have dominion over the earth, I have the right to use it how I want.

# Coding for Humility

## Humble

- Everybody is made of “star stuff”, which really makes you feel like a part of the universe.
- I feel that as a member of the human race it is my duty to protect the earth and do what I can to be nice to “her”.
- But I believe with every fiber of my being, that there is a God, who is truly merciful and just.

## Not Humble

- I had the realization that everything exists for my benefit.
- I believe that I have dominion over the earth, I have the right to use it how I want.
- I feel that the idea of a God to be somewhat naive.

# Coding for Humility

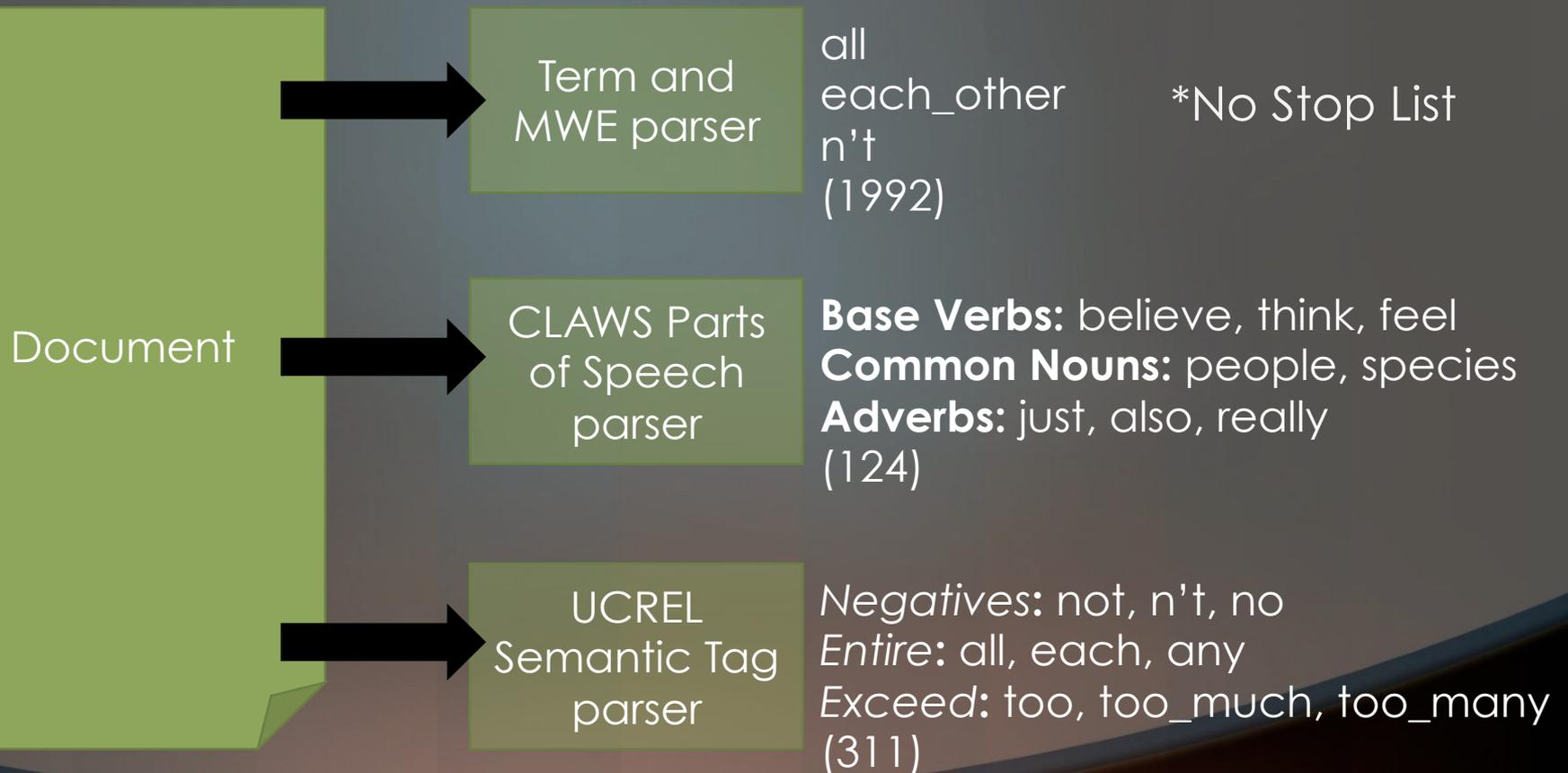
## Humble

- Everybody is made of “star stuff”, which really makes you feel like a part of the universe.
- I feel that as a member of the human race it is my duty to protect the earth and do what I can to be nice to “her”.
- But I believe with every fiber of my being, that there is a God, who is truly merciful and just.
- I believe in a helping hand and a loving heart!

## Not Humble

- I had the realization that everything exists for my benefit.
- I believe that I have dominion over the earth, I have the right to use it how I want.
- I feel that the idea of a God to be somewhat naive.
- I don't trust too many people, not even the people I get close to.

# Text Analysis Tool: WMatrix



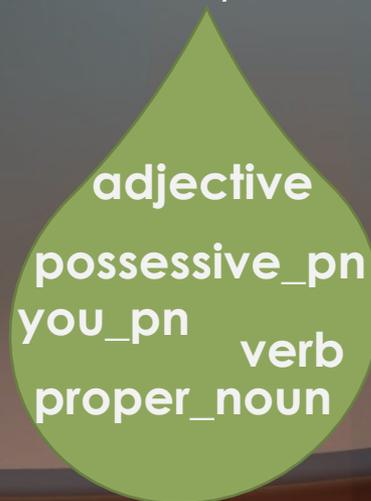
# Text Parsing: Bags of Words

When we parse sentences for this information,  
we lose its underlying *structure*

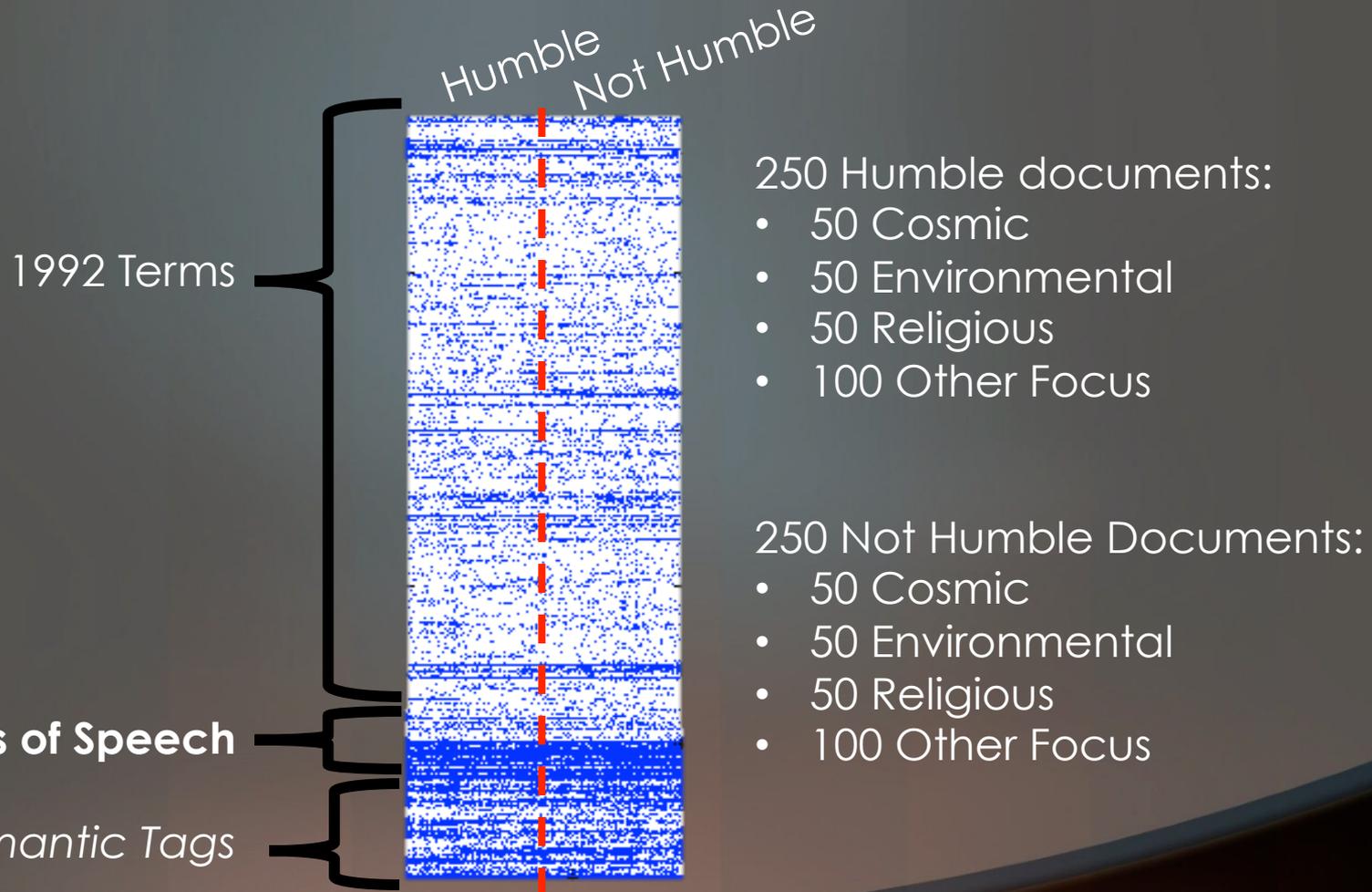
Language is thus represented as unsorted bags of terms, parts of speech, and semantic categories, which is relatively minimalistic.

.....You should *only* love your God.....

.....You should love your *only* God.....



# Feature-by-Document Matrix



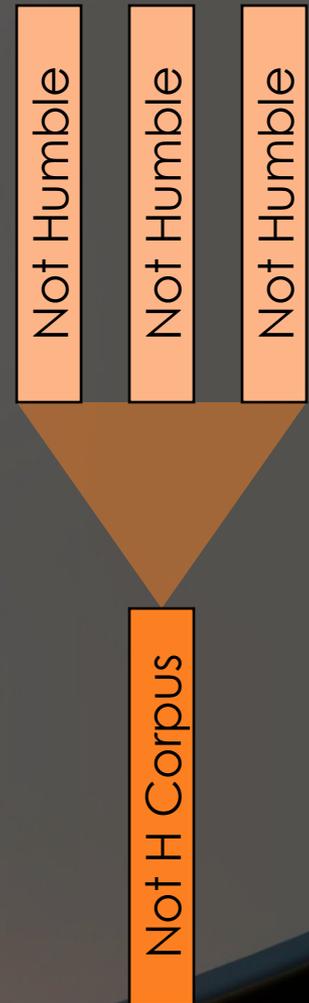
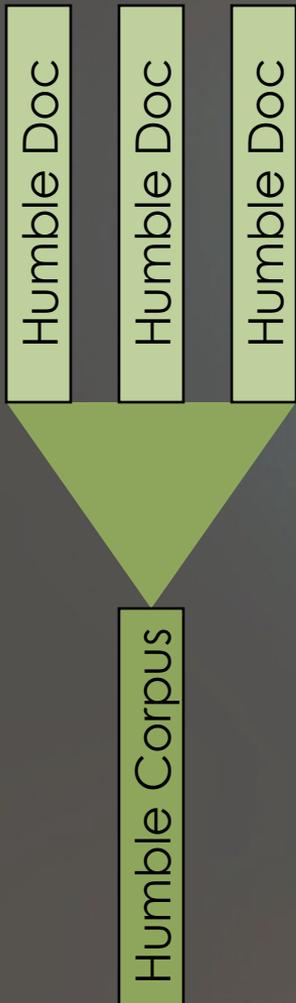
# Comparative Analysis

# Forming Corpora

Every document is a column vector of its feature composition.

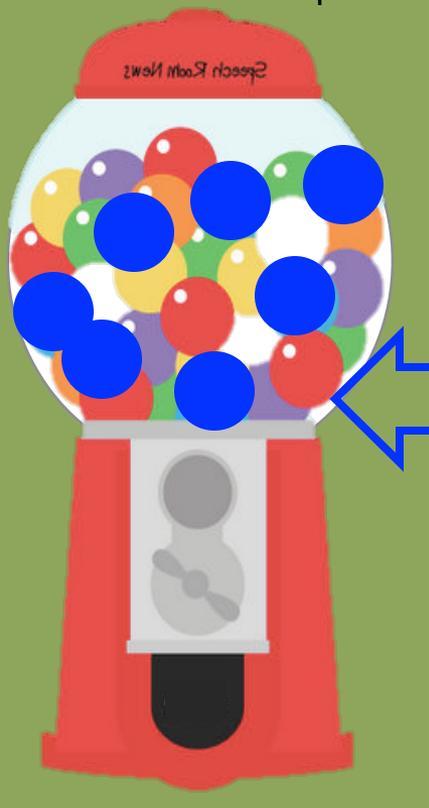
We use the vector sum of all Humble Documents to form the vector for the **Humble Corpus**.

We form the **Not Humble Corpus** in the same way.



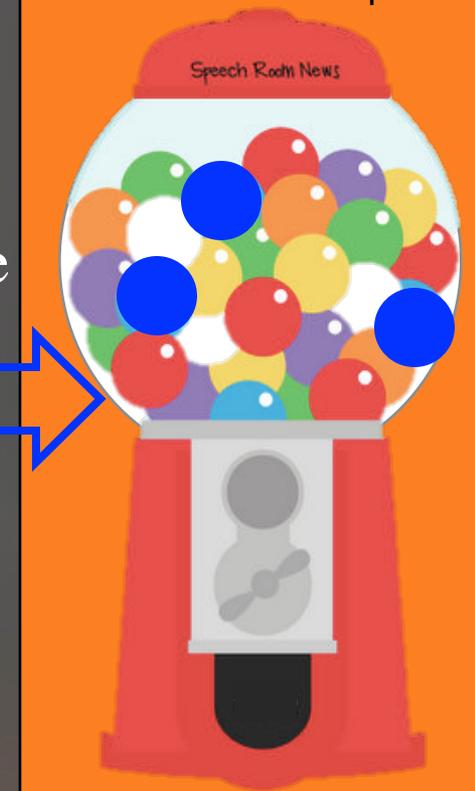
# Log Likelihood

Humble Corpus



A statistical calculation that gives the significance of a term's overuse or underuse between two documents.

Not-Hum Corpus



# Log Likelihood

It parallels the Chi-Squared Test, however LL is preferred for instances when there are less than 5 observed outcomes, such as in word counting.

The equation is based on the observed outcomes (O) wrt expected outcomes (E).

$$LL = 2 * \left( O_1 \ln \left( \frac{O_1}{E_1} \right) + O_2 \ln \left( \frac{O_2}{E_2} \right) \right)$$

# Log Likelihood

- We assume an even distribution to determine the Expected Frequencies.

$$E_1 = \frac{N_1(O_1 + O_2)}{N_1 + N_2}$$

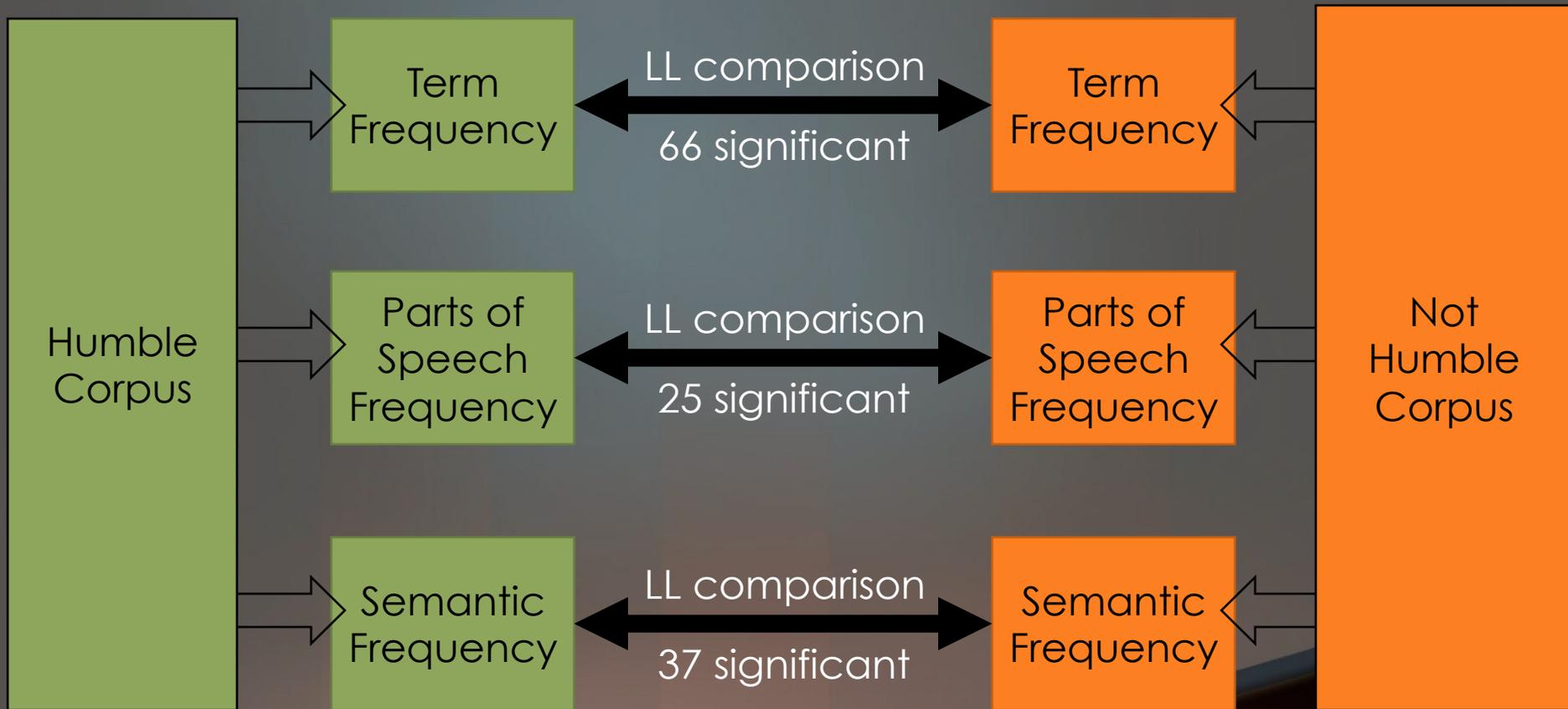
# Log Likelihood

- We assume an even distribution to determine the Expected Frequencies.

$$E_1 = \frac{N_1(O_1 + O_2)}{N_1 + N_2}$$

- The LL calculation incorporates
  - The size of each corpus,
  - How many times that feature was used, and
  - If a features was used more (or less) than was expected.
- Then we use a Chi-Squared table to determine the significance of the LL score.

# Comparing Features



# Humble vs. Not Humble

Inclusive vs. Exclusive

all  
we  
and  
us  
together  
everything  
our  
human\_beings

people  
or  
they  
them  
themselves  
my\_own  
some  
generally

# Humble vs. Not Humble

Inclusive vs. Exclusive

all  
we  
and  
us  
together  
everything  
our  
human\_beings

people  
or  
they  
them  
themselves  
my\_own  
some  
generally

Humble writers prefer to include themselves when generalizing:

'we', 'us', 'our' ('human\_beings' is an exception)

Not Humble writers tend to exclude themselves:

'people', 'they', 'them', 'themselves'

# Humble vs. Not Humble

## Inclusive vs. Exclusive

all  
we  
and  
us  
together  
everything  
our  
human\_beings

people  
or  
they  
them  
themselves  
my\_own  
some  
generally

Humble writers tend to use all-inclusive words:

'all', 'together', 'everything'

Not Humble writers use words that imply exclusions:

'my\_own', 'some', 'generally'

# Humble vs. Not Humble

## Inclusive vs. Exclusive

all  
we  
and  
us  
together  
everything  
our  
human\_beings

people  
or  
they  
them  
themselves  
my\_own  
some  
generally

Humble writers tend to use 'and' which appends to inclusive lists:

"I feel a connection to the stars , and to the planets... "

Not Humble writers tend to use 'or' which appends to exclusive lists:

"I am not 'part of the universe or part of the stars'. "

# Humble vs. Not Humble

Idealistic vs. Cynical

respect  
treat  
love  
believe  
positive  
helpful

hard  
bad  
*unethical*: evil, ashamed  
*foolish*: stupid, naïve  
*bad*: bad, crappy  
*sad*: suffering, tragedy

Humble writers are often optimistic on their outlooks on people/life:

“Strangers can often be very helpful and caring.”

Not Humble writers tend to point out the problems with people/life:

“There is too much evil and too much irrationality...”

# Humble vs. Not Humble

## Emphasizing Equality

each\_other  
each  
fellow  
same / different

## Qualifying/Analytic

think  
really  
if  
too\_many  
seem  
**adverbs:** just, also, really  
*diminishers:* simply,  
somewhat

Humble writers use language that break boundaries and hierarchies:

“... we should help each other and love each other...”

Not Humble writers tend to infuse their own judgmental views:

“Earth is simply something to be used for its resources.”

# Humble vs. Not Humble

## Obligation to Duty

try  
should  
help  
challenges

**helping:** protect, help  
**obligation:** should, need

## Negative Separation

n't  
not  
care  
any  
about  
anything  
kind  
was

Humble writers acknowledge obligations they owe to concepts:

“as part of the human race it is my duty to protect the earth.”

Not Humble writers tend to dismiss other belief systems:

“I do not, and can not, have any kind of relationship with it.”

# Humble vs. Not Humble

## Function Words

take  
can  
yet  
that  
of

do  
about  
does

# Topic Analysis

# NMF Decomposition

- NMF is a non-unique decomposition that approximates two matrices  $W$  and  $H$  such that  $WH \approx A$ .
- Very similar to SVD decomposition, except all values are positive.
- $W$  Matrix groups features into descriptive topics.
- $H$  Matrix shows the topic composition of original documents.

$A_{\text{term-by-doc}}$

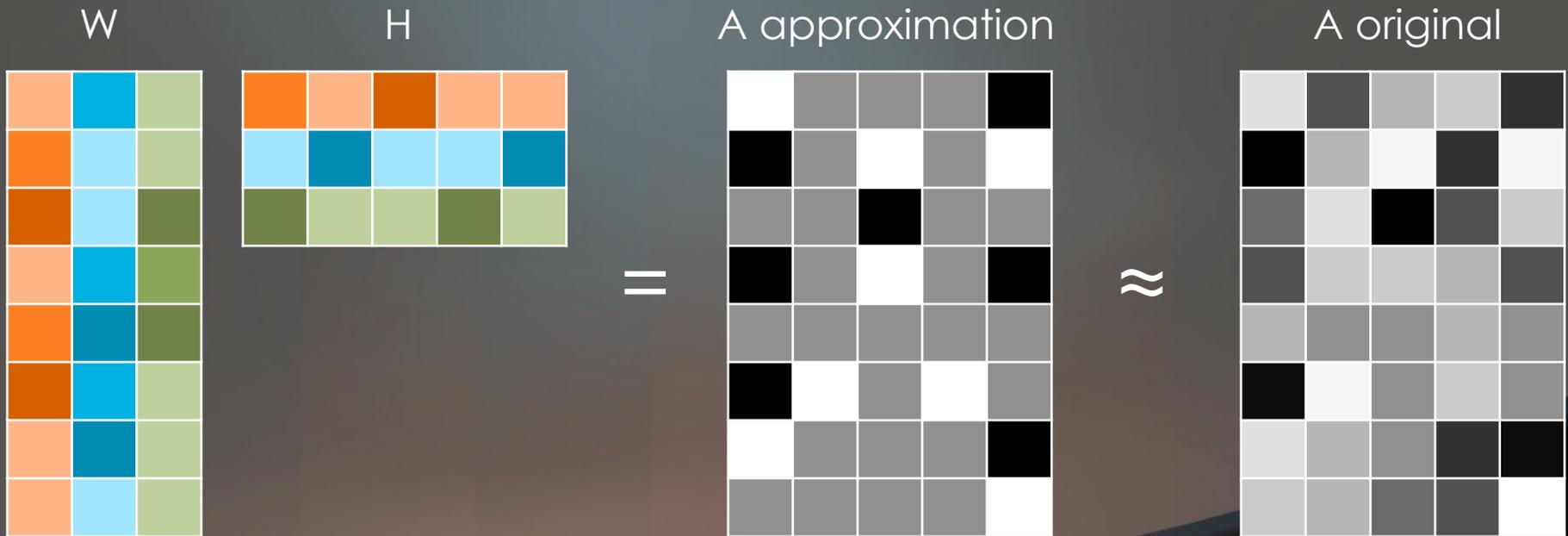
=

$W_{\text{Term-by-topic}}$

$H_{\text{topic-by-doc}}$

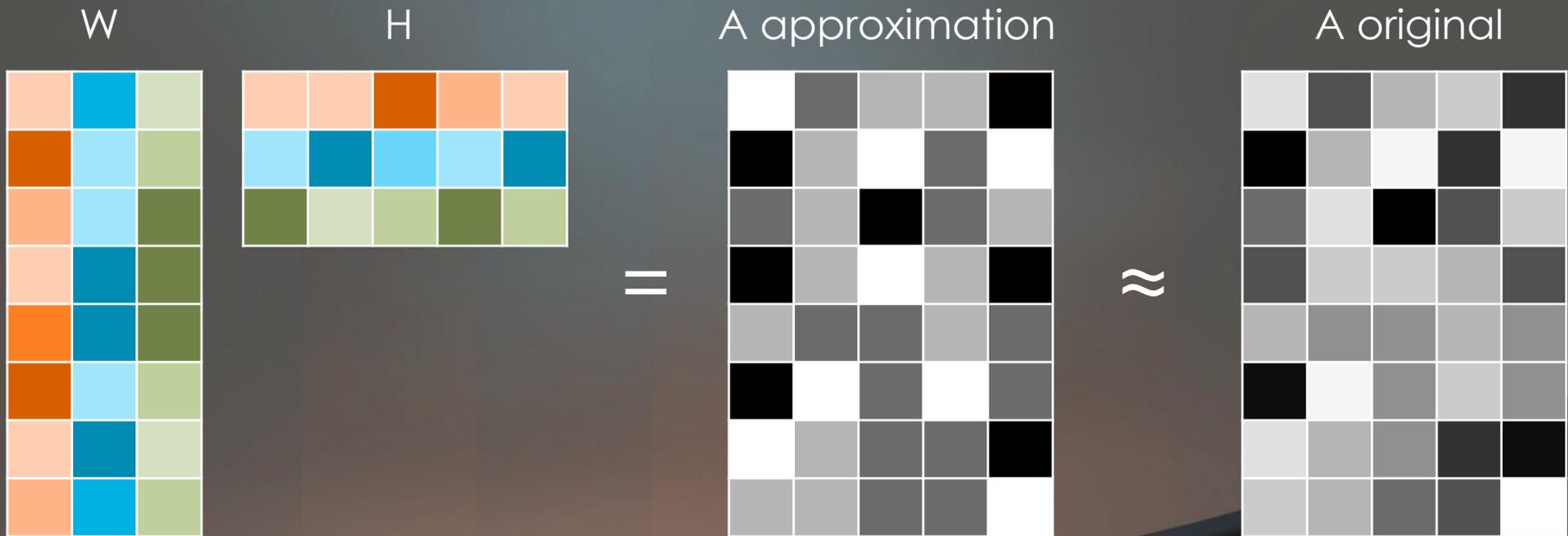
# NMF Decomposition

- W and H are initiated with random values



# NMF Decomposition

- $W$  and  $H$  are initiated with random values
- Then they are adjusted so that  $WH \rightarrow A$



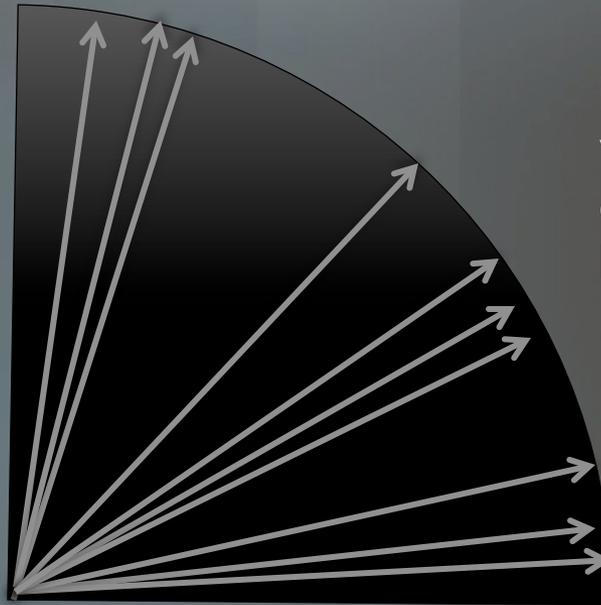
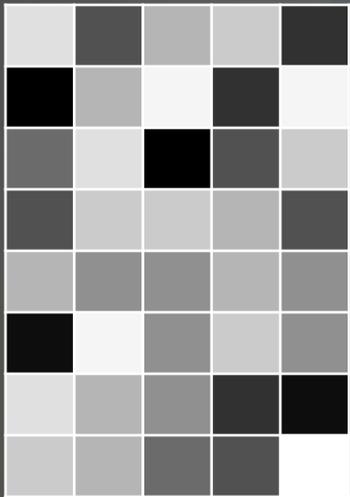
# NMF Decomposition

- $W$  and  $H$  are initiated with random values
- Then they are adjusted so that  $WH \rightarrow A$
- This continues until  $A - WH \approx 0$



# Geometry of NMF Decomposition

A original

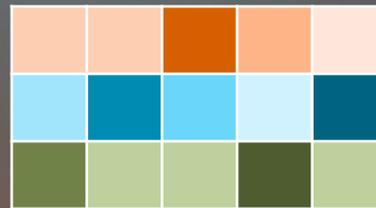


Documents in A are vectors that exist in one quadrant of hyperspace

W

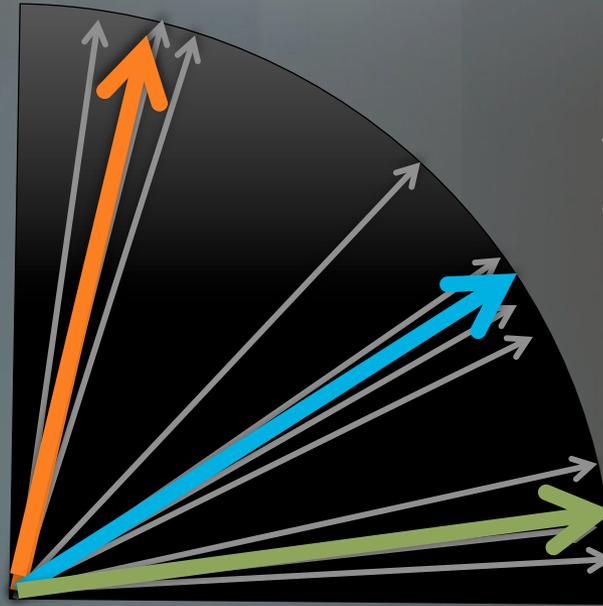
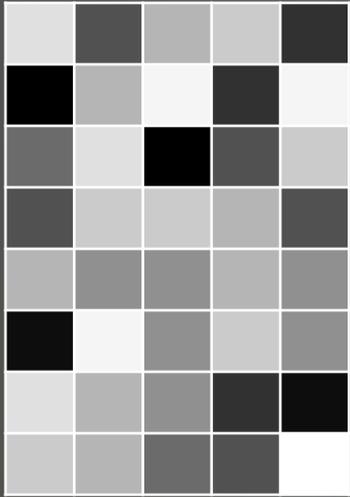


H



# Geometry of NMF Decomposition

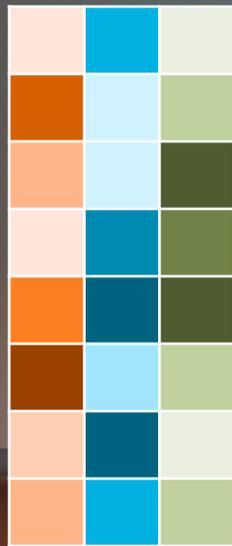
A original



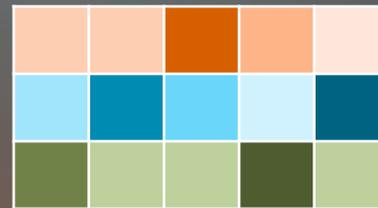
Documents in A are vectors that exist in one quadrant of hyperspace

Vector of W are topic vectors that strongly describe the data set

W

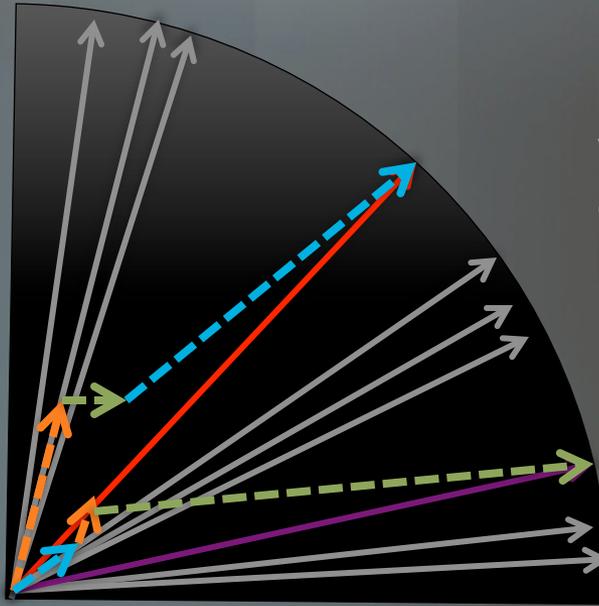
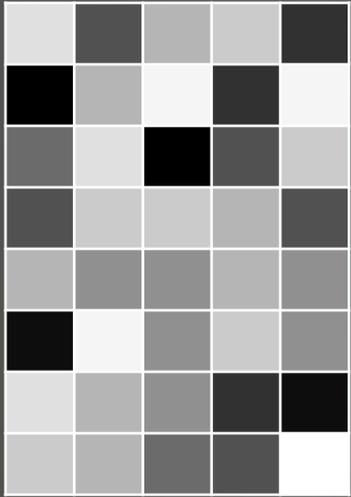


H



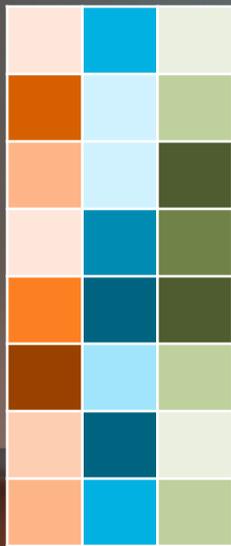
# Geometry of NMF Decomposition

A original



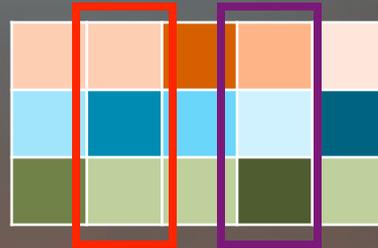
Documents in A are vectors that exist in one quadrant of hyperspace

W



Vector of W are topic vectors that strongly describe the data set

H



Columns of H give the topic decomposition of each document

# Weighting Schemes

- 3 Weights normally used to standardize the data:  
**Local Weight, Global Weight, Document Norm.**
- Experimentation with ‘Standard’ Global Weight resulted in unintended partitions but in strong predictive accuracies.
- Developed a new ‘Comparative’ Global Weight to train NMF on the desired partition, but it did not improve predictive accuracy

# Standard Global Weight

*Inverse Document Frequency*

The standard global weight emphasizes terms unique to a few documents

Maximum Weight →



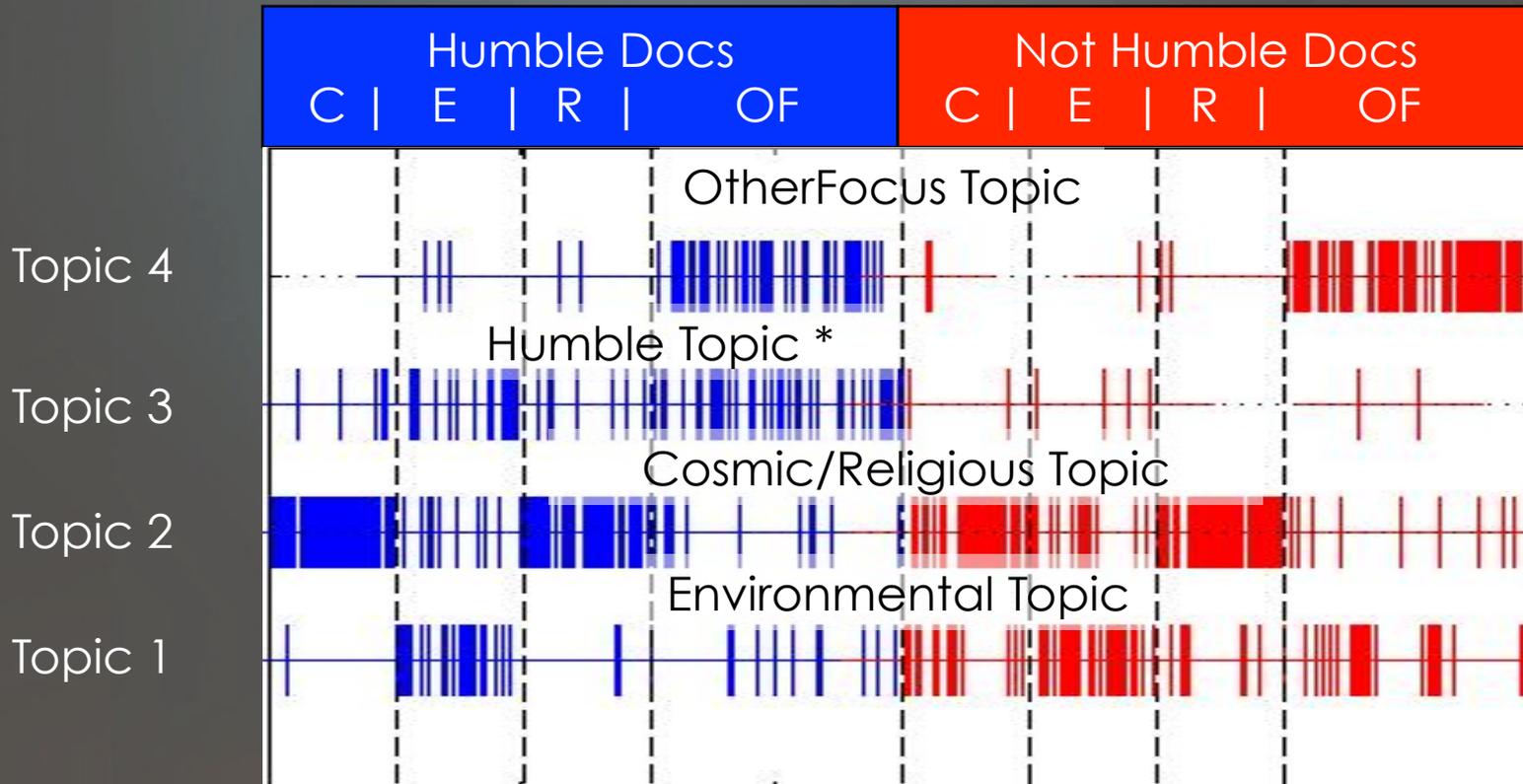
Medium Weight →



Minimum Weight →



# NMF with Standard Weight

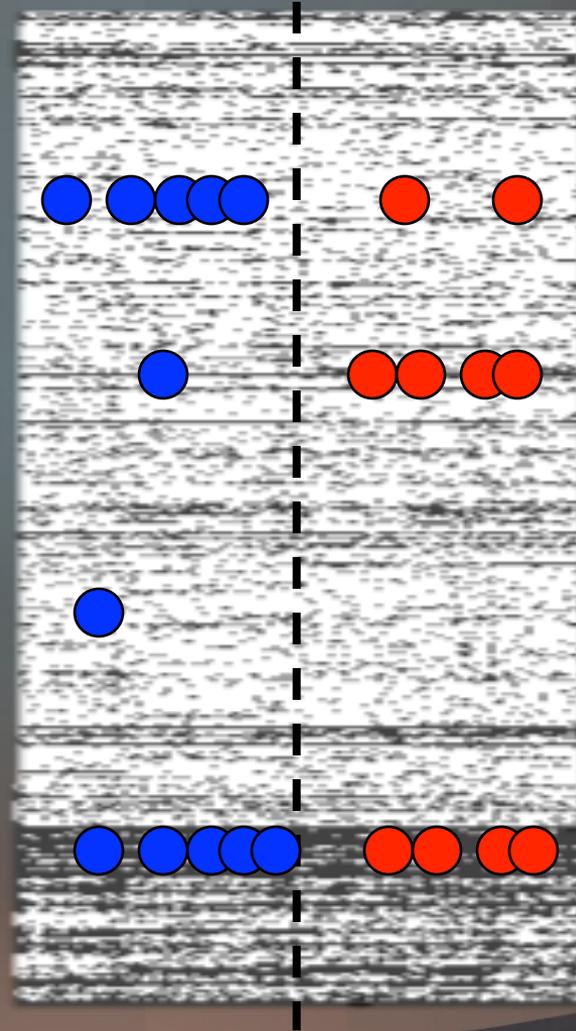


With the standard weighting, NMF naturally partitions by the divisions we already know exist.

# Comparative Global Weight

The new comparative weight emphasizes terms determined to be overused across the predetermined division

Terms get weighted heavily where overused...



...while that term gets weighted minimally where underused

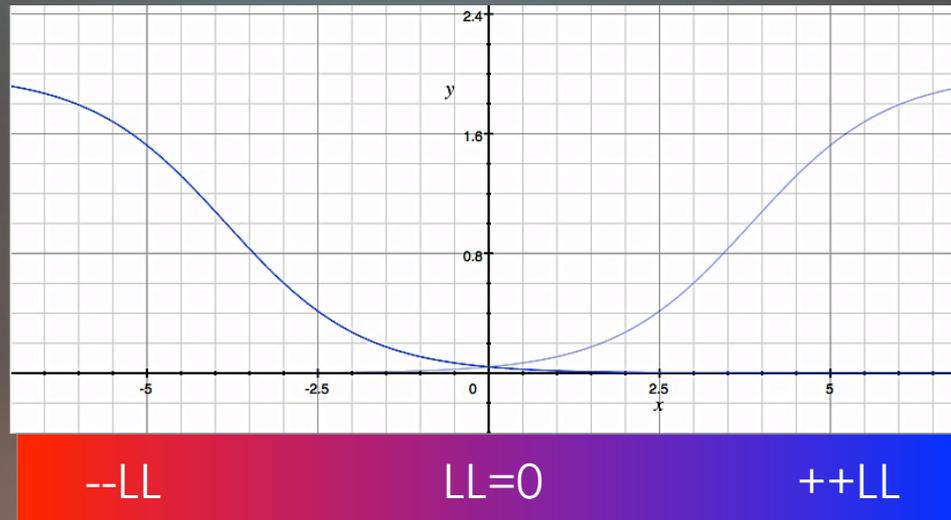
Terms either unique to a few documents --or-- shared amongst all document should also get minimal weight

# Comparative Global Weight

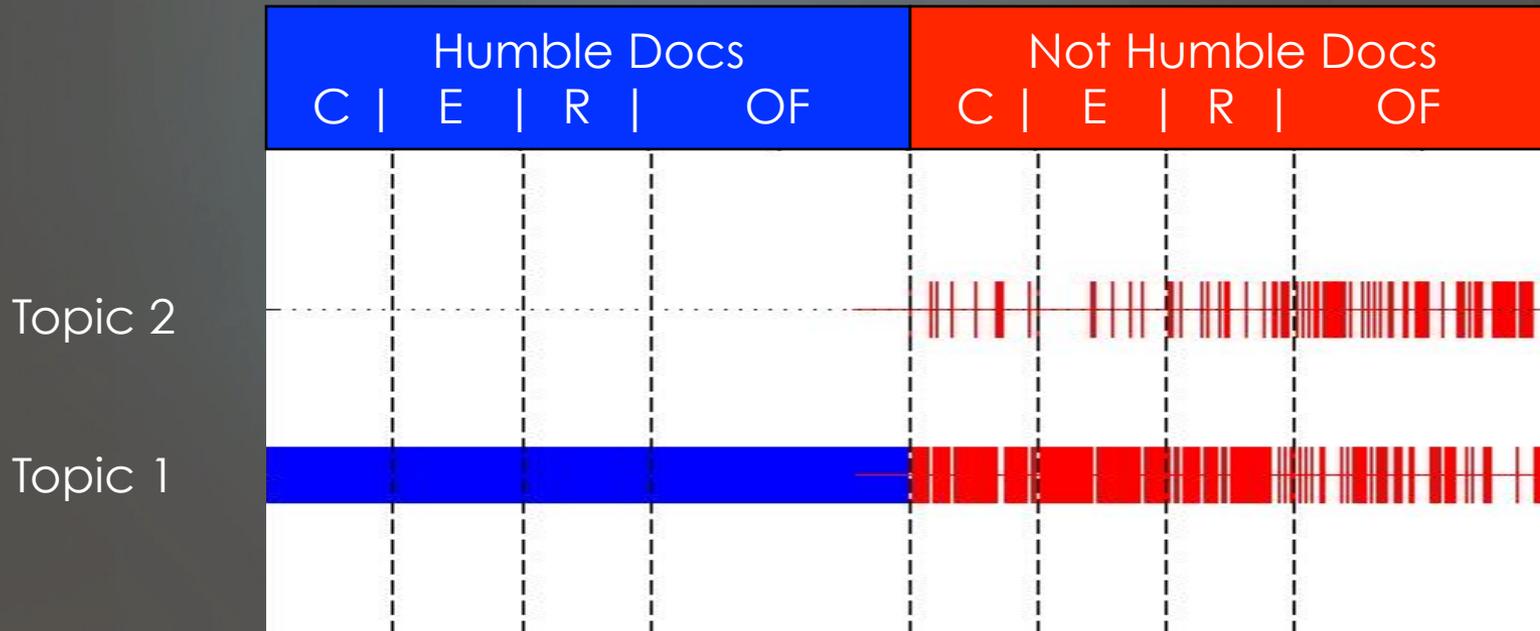
Weight is dependent upon LL score from before

$$f_X(x) = \frac{2}{1 + e^{x+3.84}}$$

$$f_H(x) = \frac{2}{1 + e^{-x+3.84}}$$



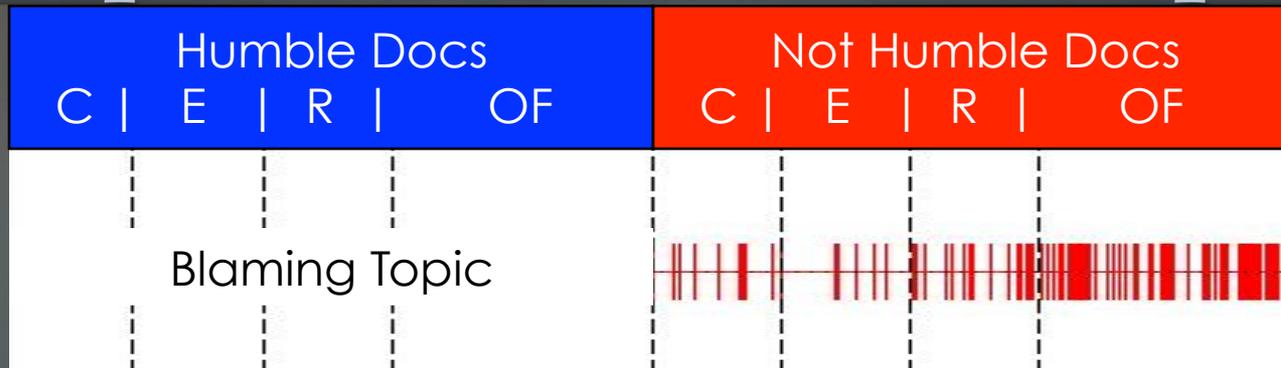
# 2 Topic Analysis



First pulls apart the Not Humble Other Focus Documents.

# Example Documents for Topic 2

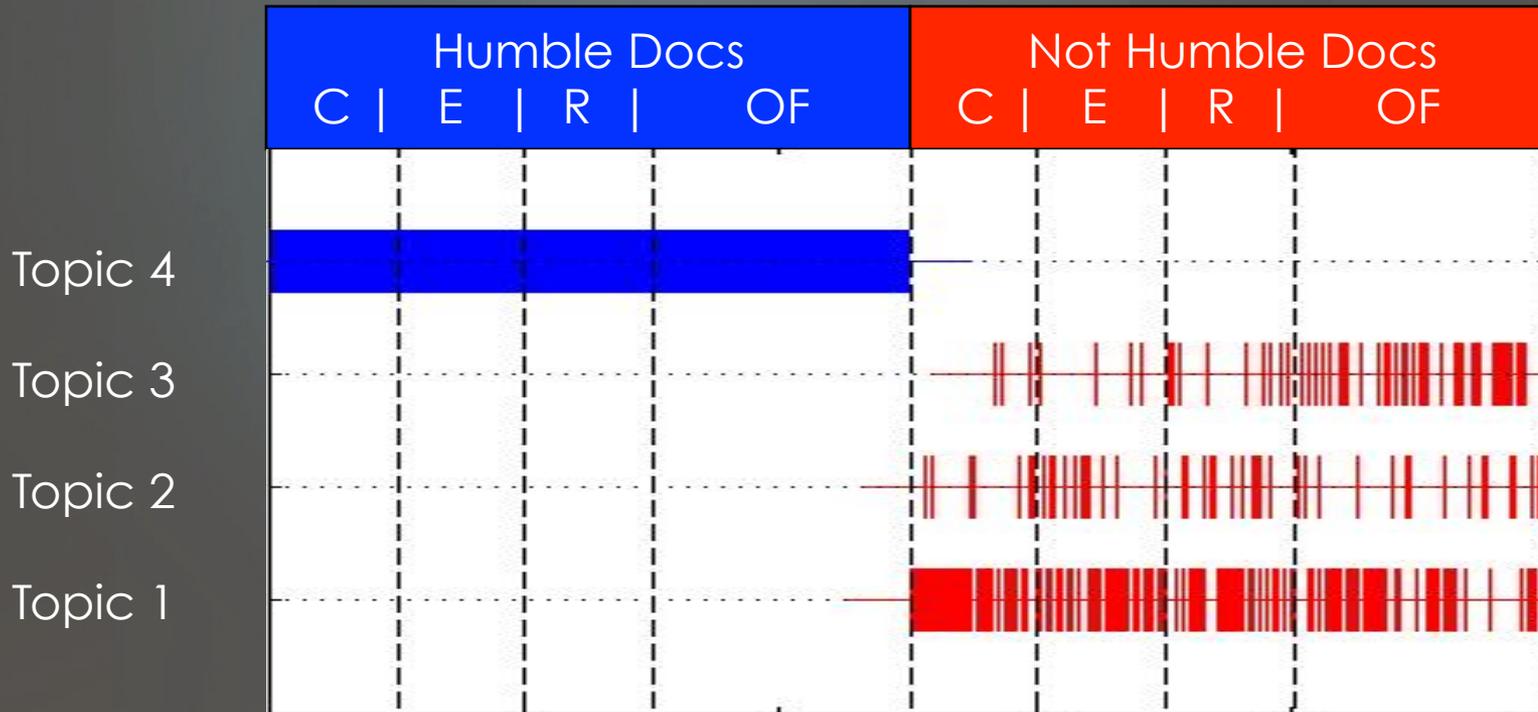
Topic 2



- Many people are untrustworthy; many are lazy; some are even evil.
- While there are good people, there are many bad people.
- People in general can be pretty disappointing.
- People can hardly be trusted.

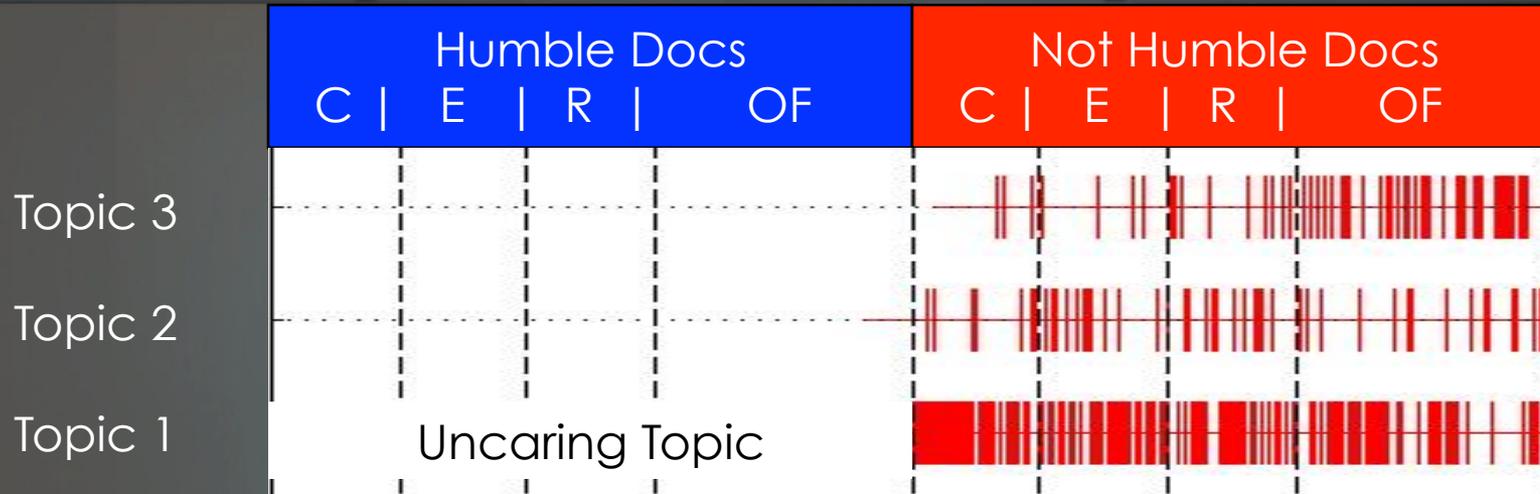


# 4 Topic Analysis



Breaks down Not Humble documents into three topics before it can divide Humble documents into two distinct topics.

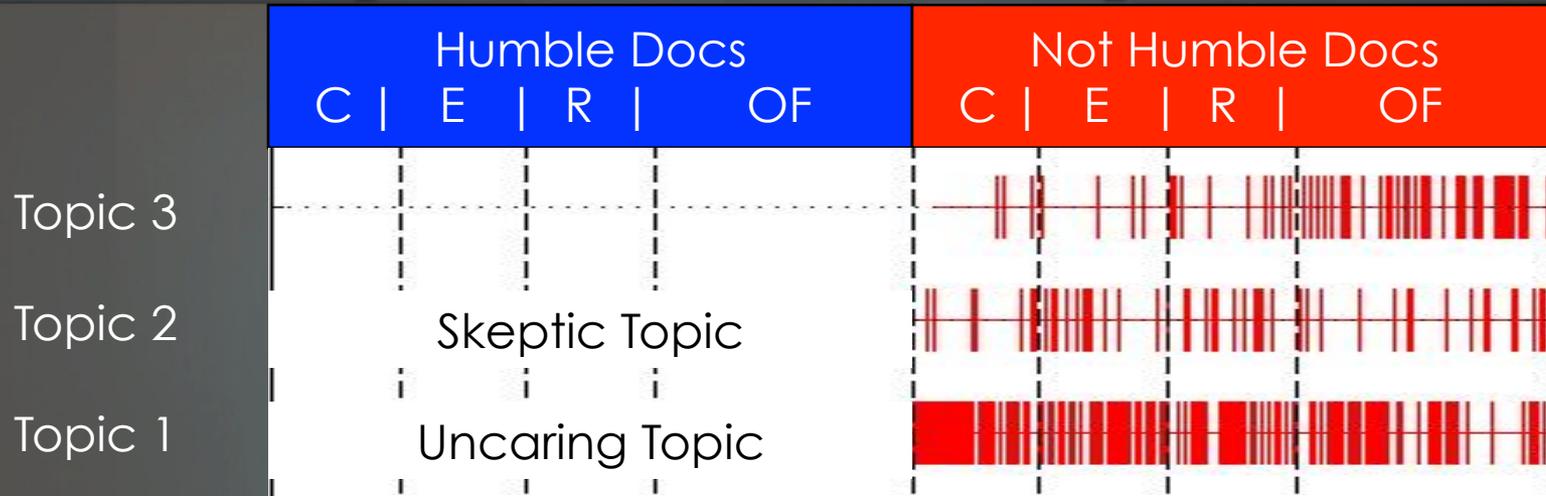
# Example Documents for Topics 1, 2, and 3



## Topic 1

- My family is big into nature. I'm not. I don't really care about it much.
- I don't see how my life ties into the universe.
- I don't care about the environment or mother nature.
- Sadly, I don't care one way or another.

# Example Documents for Topics 1, 2, and 3



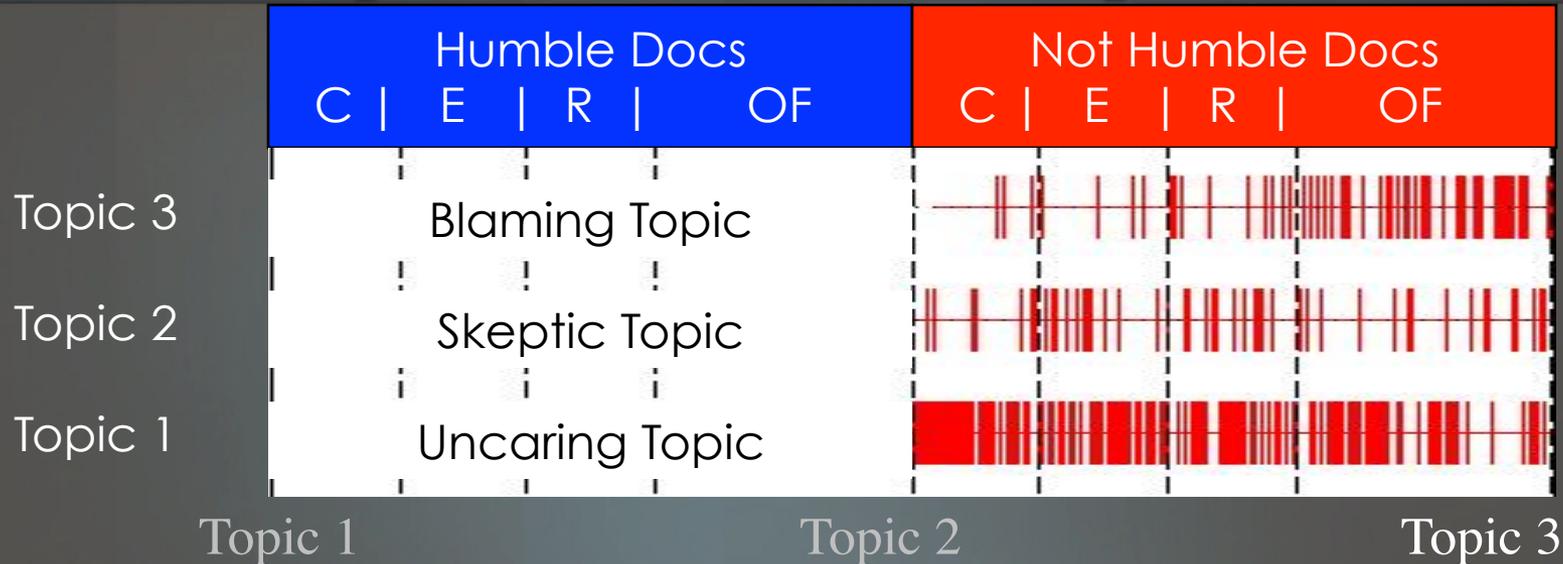
Topic 1

Topic 2

- My family is big into nature. I'm not. I don't really care about it much.
- I don't see how my life ties into the universe.
- I don't care about the environment or mother nature.
- Sadly, I don't care one way or another.

- God just seems like a crappy story.
- I guess somewhere along the way the idea of God just seemed to get very ridiculous to me.
- Based on interactions and stories for just a few bad incidents, it is hard to trust others.

# Example Documents for Topics 1, 2, and 3



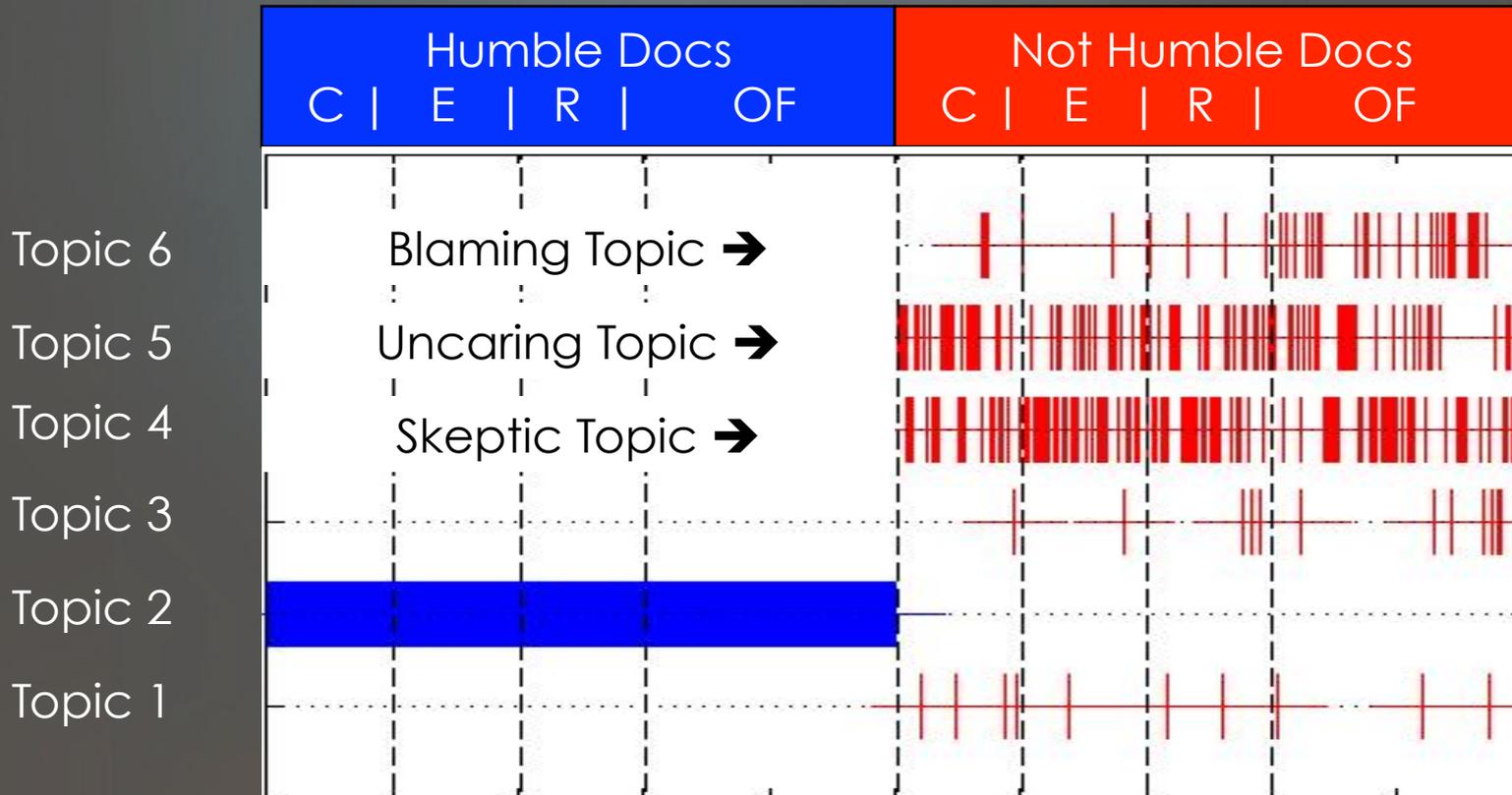
- My family is big into nature. I'm not. I don't really care about it much.
- I don't see how my life ties into the universe.
- I don't care about the environment or mother nature.
- Sadly, I don't care one way or another.

- God just seems like a crappy story.
- I guess somewhere along the way the idea of God just seemed to get very ridiculous to me.
- Based on interactions and stories for just a few bad incidents, it is hard to trust others.

- Many people are untrustworthy; many are lazy; some are even evil.
- While there are good people, there are many bad people.
- People in general can be pretty disappointing.
- People can hardly be trusted.

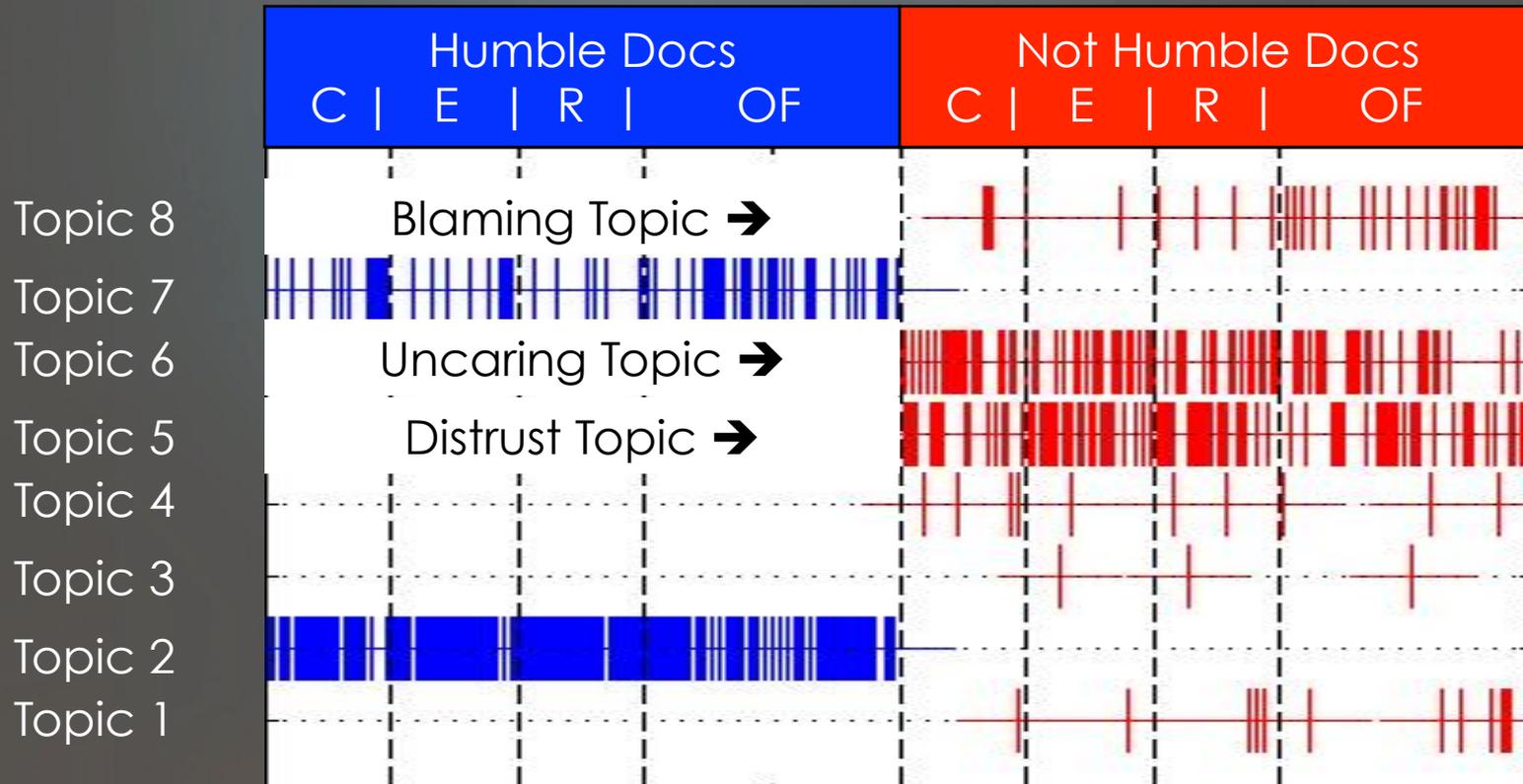


# 6 Topic Analysis



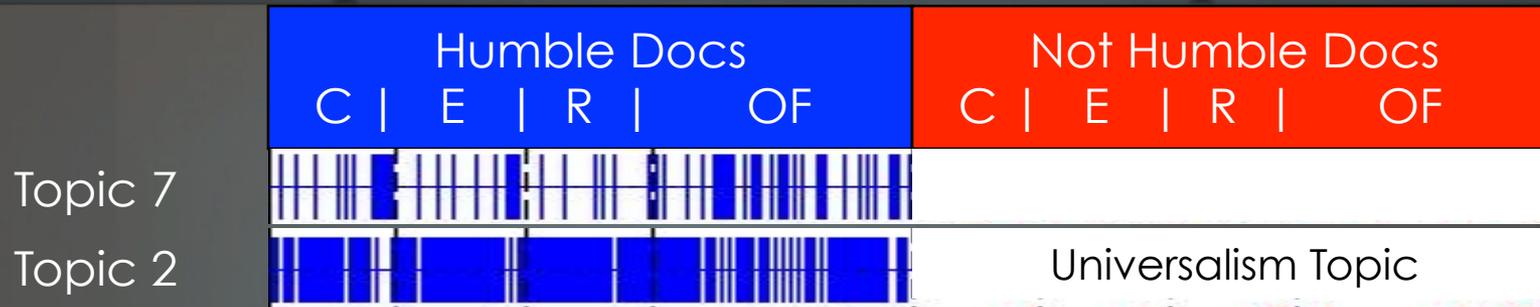
Further breaks down Not Humble documents into 5 topics before it can divide Humble documents into 2 distinct topics.

# 8 Topic Analysis



Only when we break the data set into 8 topics do we finally see Humble documents divided into 2 distinct topics.

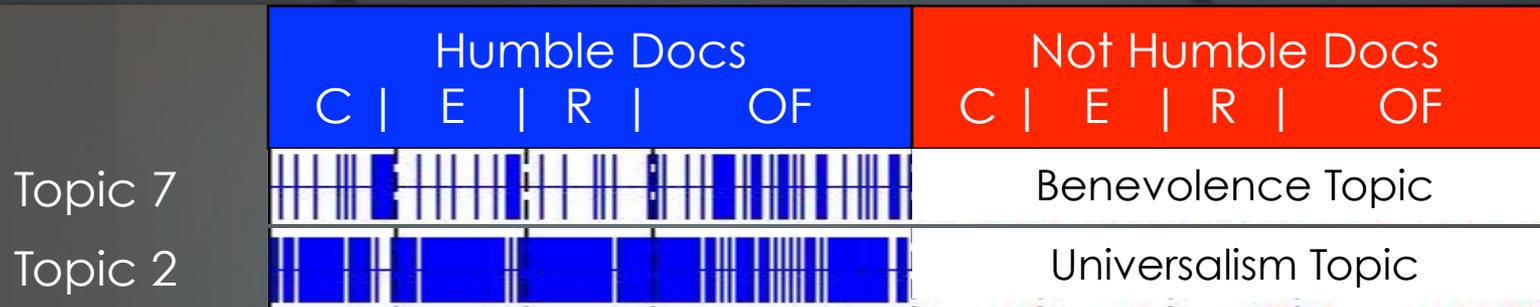
# Example Documents for Topics 2 and 7



## Topic 2

- I believe that the majority of people are innately good, and I am a trusting person.
- I feel like the majority of people I have met are good and want to be friendly with other humans.
- I acknowledge that everyone has an equal status and equal rights to be who they are.
- I believe most people are good and some just battle their demons better than others.

# Example Documents for Topics 2 and 7



## Topic 2

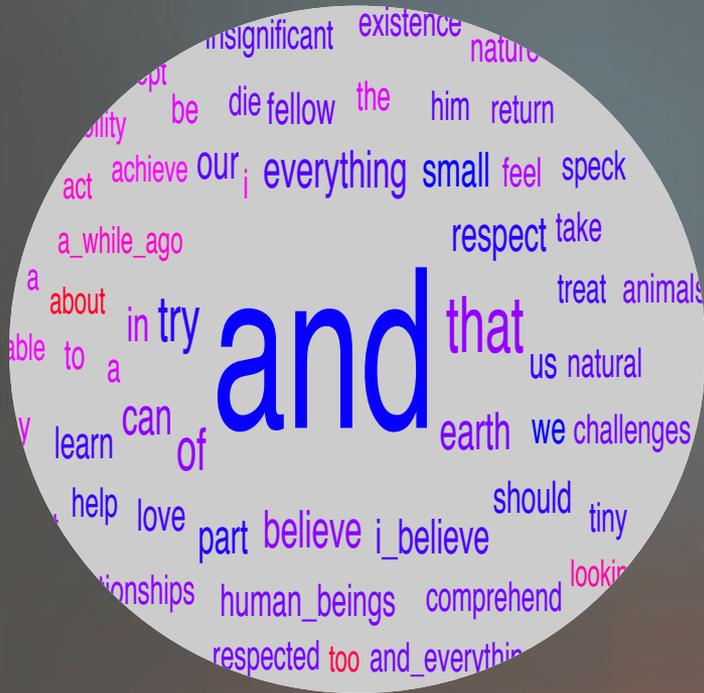
- I believe that the majority of people are innately good, and I am a trusting person.
- I feel like the majority of people I have met are good and want to be friendly with other humans.
- I acknowledge that everyone has an equal status and equal rights to be who they are.
- I believe most people are good and some just battle their demons better than others.

## Topic 7

- I have a genuine concern for the well being of all mankind.
- I do feel like we all have responsibilities for each other.
- I love how we are all different.
- Relationships with other people are all we have to judge our life by.
- From this relationship comes the incentive for all my other relationships.

# Characteristic Features of Universalism and Benevolence

Universalism



Benevolence



Humble

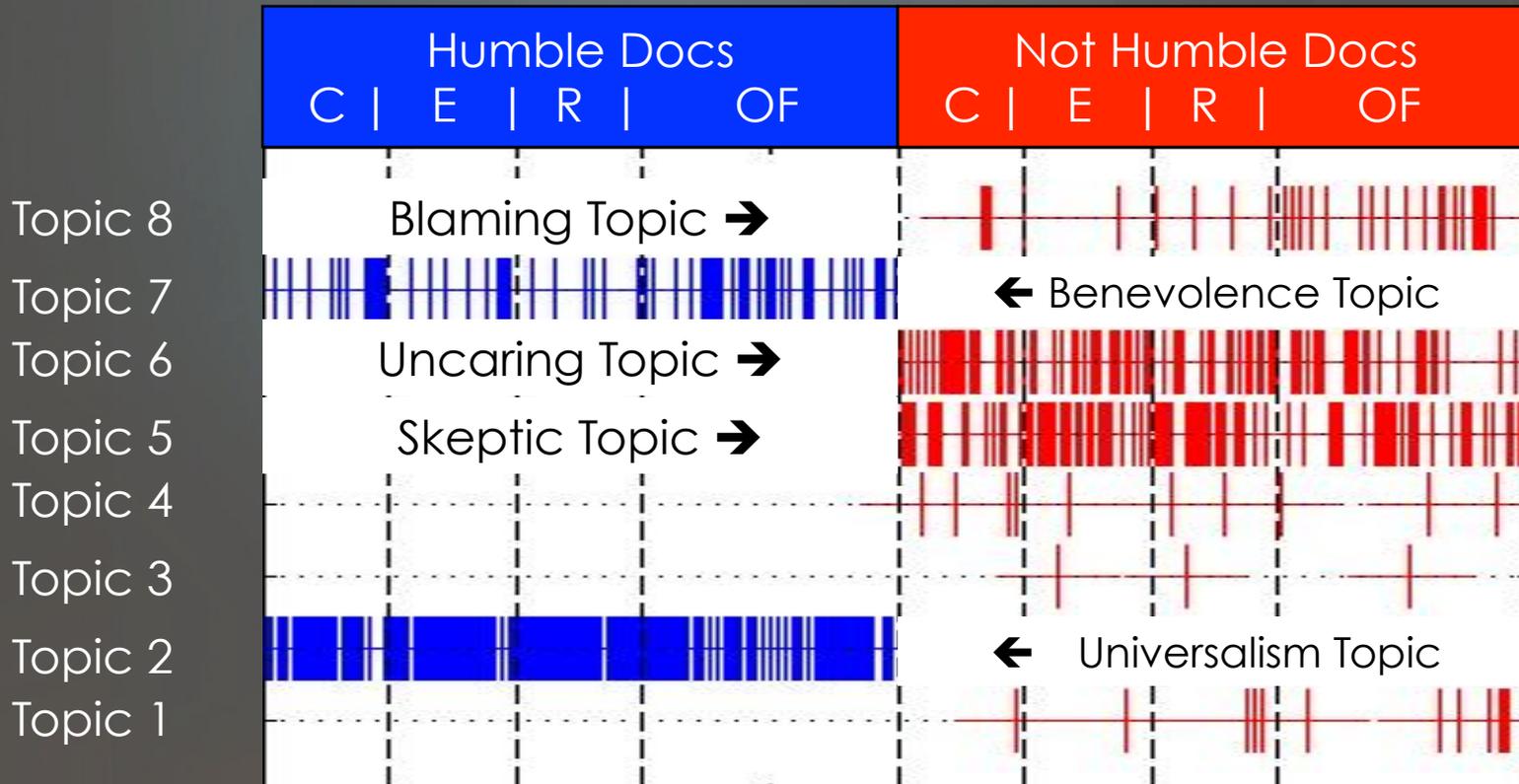
95% Significance

Neutral

95% Significance

Not Humble

# 8 Topic Analysis



# Predictive Analysis

# Cross-Validation & Predictive Power

- Randomly remove 25 Humble and 25 Not Humble documents from training set
- Classify ‘query documents’ by their single nearest neighbor via cosine similarity
- Repeat 1,000 times for overall accuracies:

k=8, ‘standard’ global weight

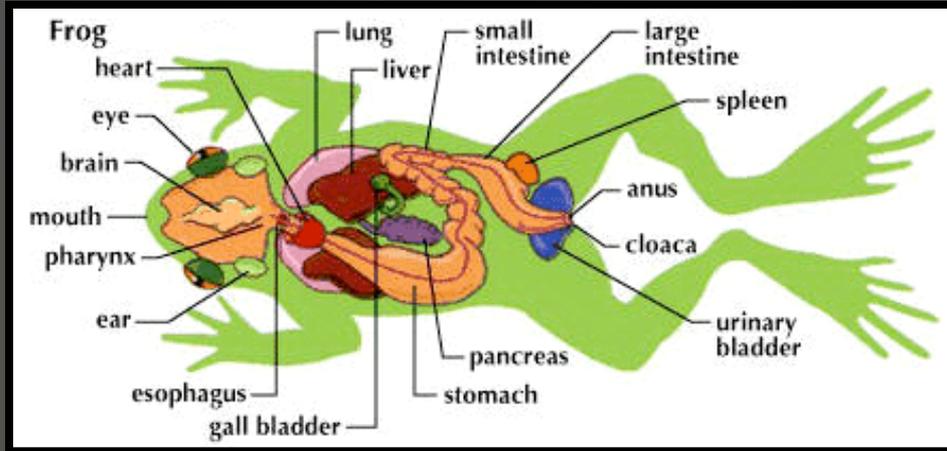
	Humble	Non Humble
Humble Docs classified as...	66.6%	33.4%
Non Humble Docs classified as...	32.7%	67.3%

# Conclusions

# Conclusions about Methodology

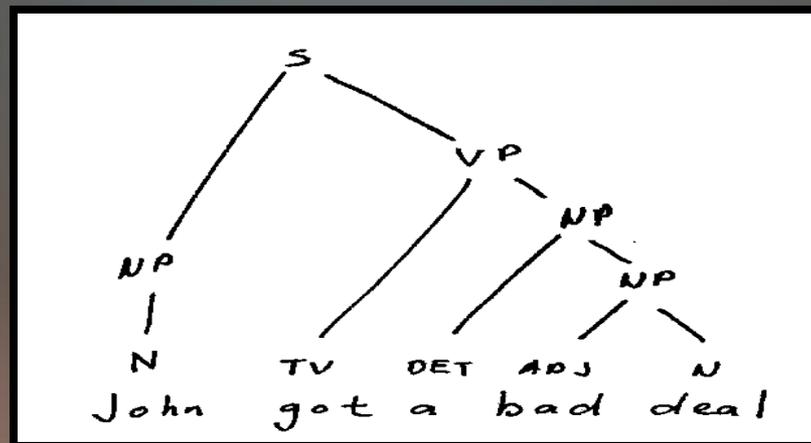
- Analysis is only at the word level, where sentences are represented as unstructured ‘bags of words’
- Through two independent analyses, we found complementary and insightful clusters of words in their responses
- A new weighting mechanism was developed to emphasize partitioning on desired characteristic features
- Standard weighting scheme proved to be more effective in predictions than new weighting scheme

# Language >>> Sum of its Parts



Eyes	Heart	Bladder	Liver
2	1	1	1
Cloaca	Spleen	Lungs	Brain
1	1	2	1

Just like you don't know the anatomy of a frog by counting its organs, we cannot make conclusions about Language just by word counts.



# Linguistics vs. Word Counts

**L**anguage the universal potential of humans' incredible ability to communicate

phonology    syntax    pragmatics

morphology    semantics    cross-linguistic



the way people use Language to communicate

the words English-speakers use when primed to respond to specific questions

# Conclusions about Humility

Since we asked participants directly about their relationships to these concepts, we can make strong inferences about their belief systems (but not about Language)

Humble	Not Humble
<p>These belief systems tend to be much simpler &amp; inclusive with fewer exceptions &amp; qualifications</p> <p>They most often value benevolence or universalism</p> <p>When talking about their beliefs, these people</p> <ul style="list-style-type: none"><li>▪ include themselves and everybody else in generalizations,</li><li>▪ value both the larger context and the ‘little guys’, and</li><li>▪ break boundaries and hierarchies.</li></ul>	<p>There are more ways to be <i>not humble</i> than to be <i>humble</i></p> <p>They most often dismiss others’ belief systems instead of explicitly sharing their own</p> <p>When talking about their beliefs, these people</p> <ul style="list-style-type: none"><li>▪ frequently blame others,</li><li>▪ display skepticism rather than trust, and</li><li>▪ disconnect through the simple <i>lack of caring</i>.</li></ul>

# Future Research

# Questions?

Thank you so much for your time  
and this opportunity!!



# Log Likelihood

- Log Likelihood is a statistical calculation that gives the significance of a term's overuse (or underuse) between two documents.
- It parallels the Chi-Squared Test, however LL is preferred for instances when there are less than 5 observed outcomes, such as in word counting.
- The equation is based on the observed outcomes (O) and expected outcomes (E).

$$LL = 2 * \left( O_1 \ln \left( \frac{O_1}{E_1} \right) + O_2 \ln \left( \frac{O_2}{E_2} \right) \right)$$