

The background of the slide is a complex network diagram. It consists of numerous nodes, represented by circles of various sizes and colors (blue, green, yellow, black, and white), interconnected by a dense web of thin lines. Some nodes are highlighted with larger, semi-transparent circles in blue, green, or yellow. The overall effect is a sense of a large, interconnected system, likely representing a social network or a data network.

What's in the mail box?

Filippo Medri
Metis 2018 San Francisco



❖ 3000 e-mails



- LSA + K-Means
- LDA

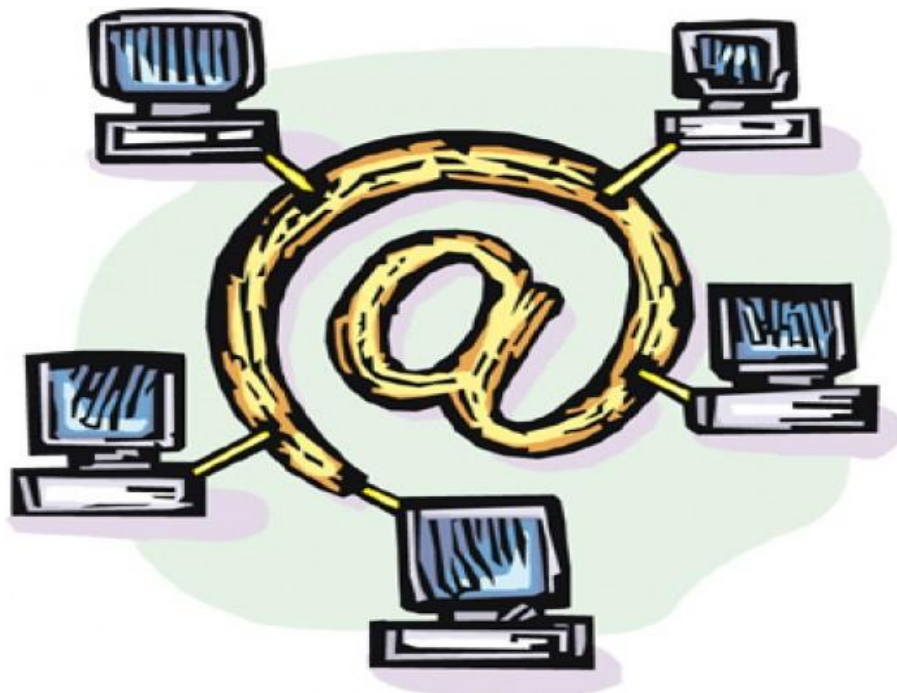
AGENDA

- ➔ Topic Discovery
- ➔ Automatic Categorization

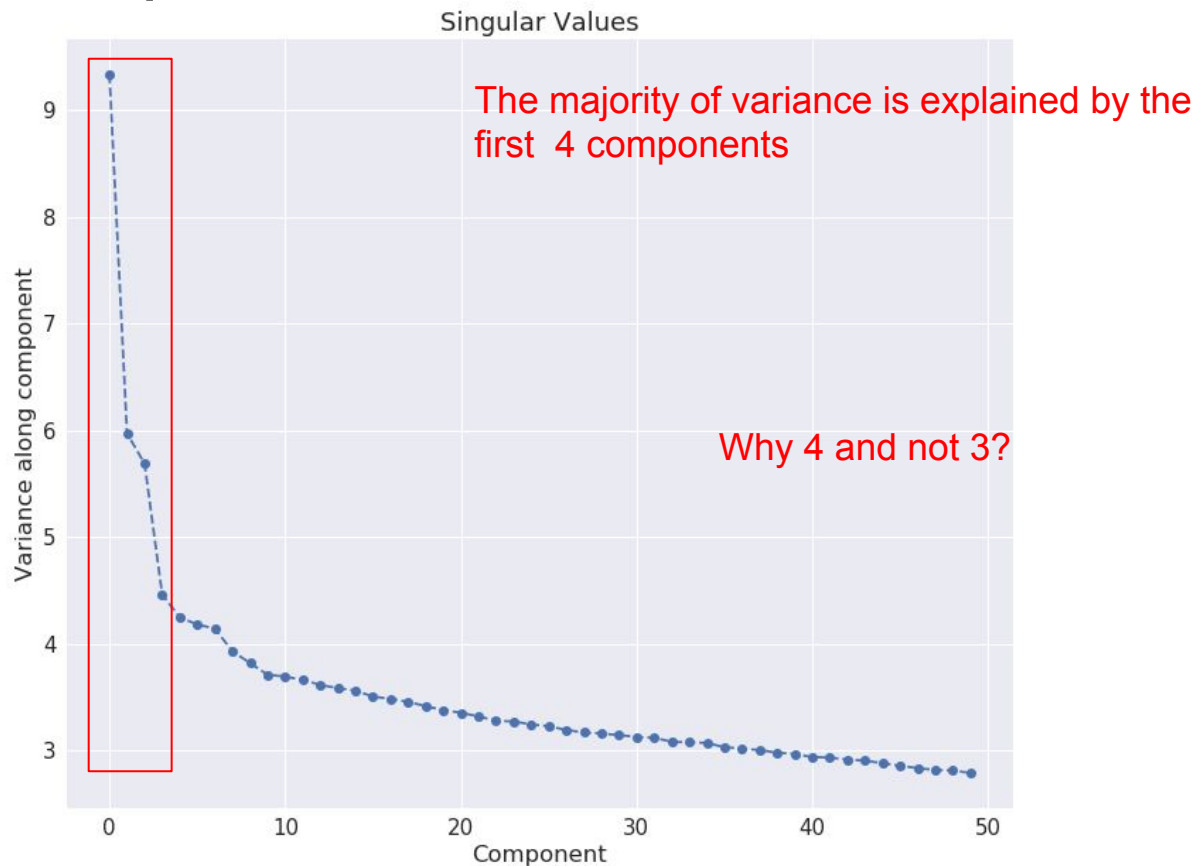


20Newsgroups Dataset

- ❖ 3 newsgroups :
 - comp.graphics
 - rec.autos
 - talk.misc.religion
- ❖ Removed **header**,
footers, **quotes**

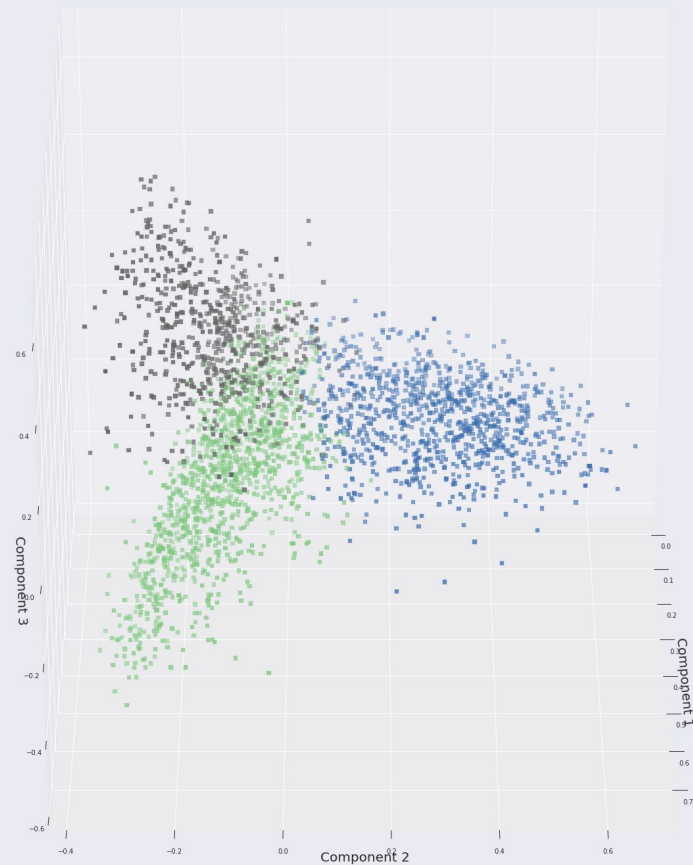


LSA - 50 components from 1000 features



3-Means

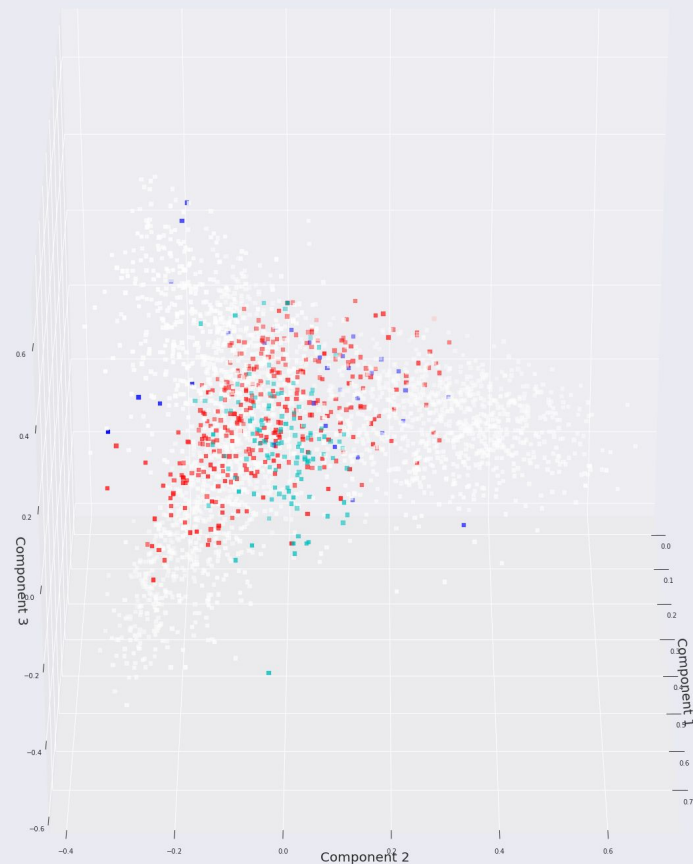
PEOPLE	thank	car
GOD	graphic	engine
THINK	image	new
CHRISTIAN	know	drive
SAY	program	good
KNOW	computer	speed
LIFE	looking	ford
MAKE	please	dealer
GOOD	file	problem
JESUS	use	price



Misclassification

Overall	17 %
<i>rec.autos</i>	35.3 %
comp.graphics	12.6 %
TALK.MISC.RELIGION	3.4 %

***WHITE: correctly
classified***



3-Means

PEOPLE	thank	car
GOD	graphic	engine
THINK	image	new
CHRISTIAN	know	drive
SAY	program	good
KNOW	computer	speed
LIFE	looking	ford
MAKE	please	dealer
GOOD	file	problem
JESUS	use	price

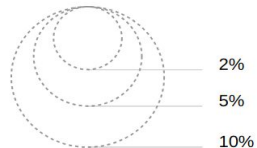


LDA - Unaspected topic

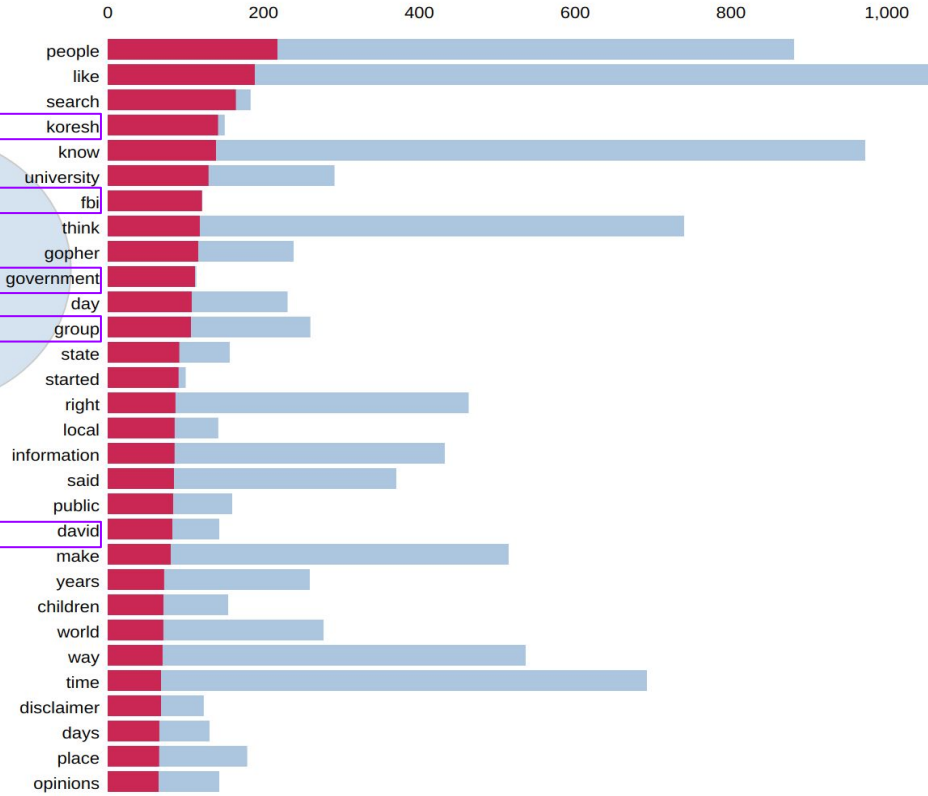
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 4 (12.8% of tokens)



Overall term frequency

Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))]] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

LDA - it is about David Koresh !!!

David Koresh - Wikipedia - Mozilla Firefox (Private Browsing)

David Koresh - Wikipedia

https://en.wikipedia.org/wiki/David_Koresh

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

David Koresh

From Wikipedia, the free encyclopedia

David Koresh (born **Vernon Wayne Howell**; August 17, 1959 – April 19, 1993) was the American leader of the [Branch Davidians](#) sect, believing himself to be its final [prophet](#).

Koresh came from a dysfunctional family background and was a member, and later a leader, of the [Shepherds Rod](#), a reform movement led by [Victor Houteff](#) that arose from within the [Seventh-day Adventist Church](#).


Koresh joined a spiritual [group](#) that was based at the Mount Carmel Center outside [Waco, Texas](#), where the [group](#) took the name "Branch Davidians". Here he competed for dominance with another leader named [George Roden](#), until Roden was jailed for murdering another rival.^[2]

The serving of arrest and search warrants by the U.S. [Bureau of Alcohol, Tobacco, and Firearms](#) (ATF) as part of an investigation into illegal possession of firearms and explosives provoked the historic [1993 raid](#) on the center.^[3] Four ATF agents and six Davidians were killed during the initial two-hour firefight, both sides claiming the other side fired first. The subsequent siege by the [FBI](#) of almost two months ended when the center was set on fire — Koresh and 79 others were found dead after the [conflagration](#).

Contents [\[hide\]](#)

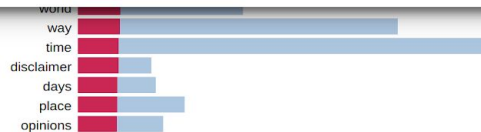
- Early life
- Ascent to leadership of the Branch Davidians

David Koresh



Koresh in 1987

Marginal topic distribution



Overall term frequency

Estimated term frequency within the selected topic

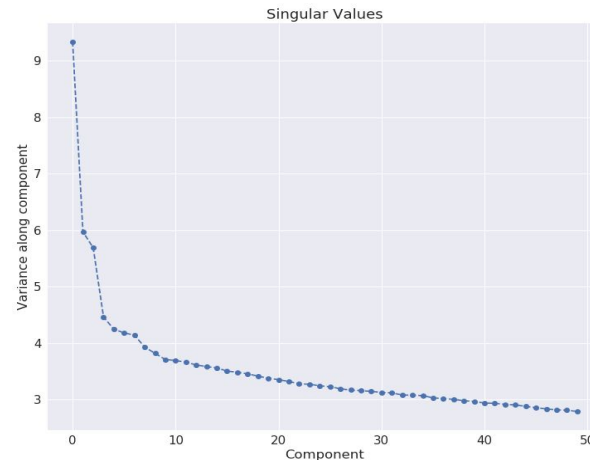
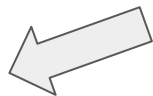
1. saliency($\text{term } w$) = $\text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)
2. relevance($\text{term } w$ | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

Conclusions

Automatic categorization



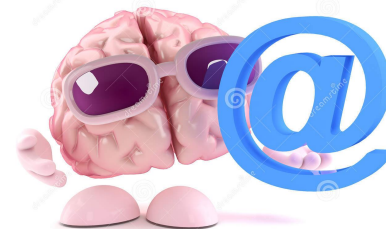
Component
distribution



Discovery of **new** topics !!!



Pre-training of
deep learning
models



A complex network diagram with numerous nodes and edges. Nodes are represented by circles of various sizes and colors (blue, green, yellow, black, white). Edges are thin lines connecting the nodes, forming a dense web. Some nodes are highlighted with larger circles or dashed outlines. The background is light blue with a subtle gradient.

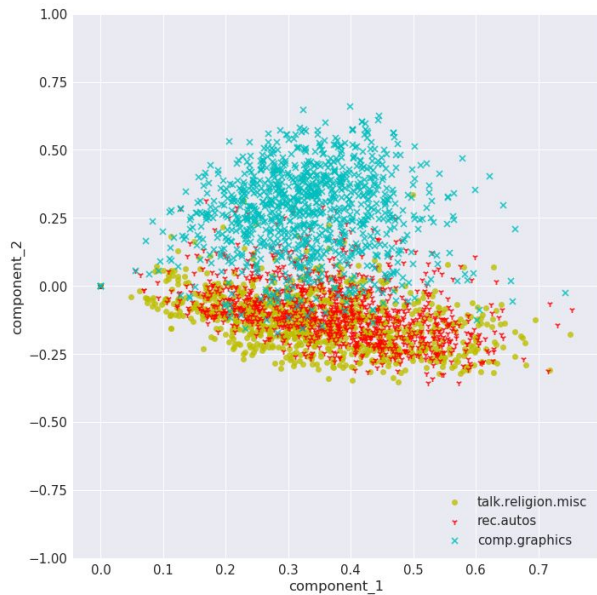
Thank You!!

Filippo Medri
Metis 2018 San Francisco

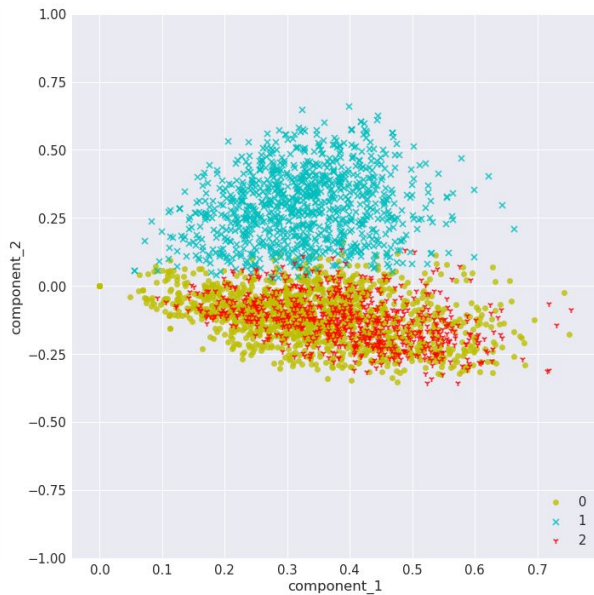
Discard Pile



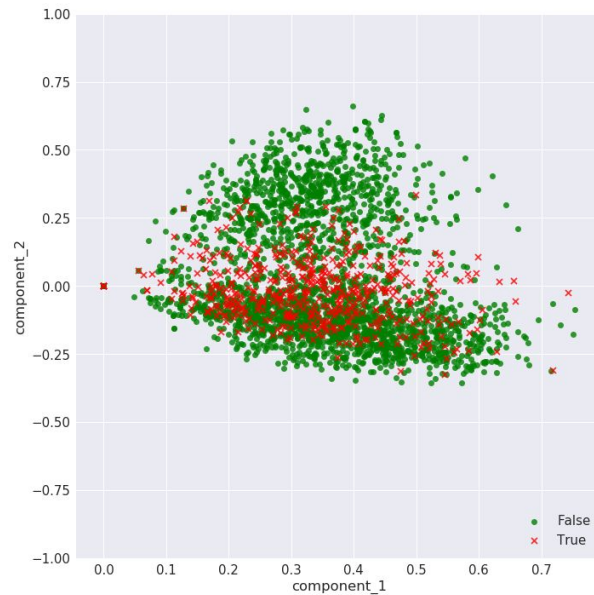
Clusters vs Category on first 2 principal components



Original Categories



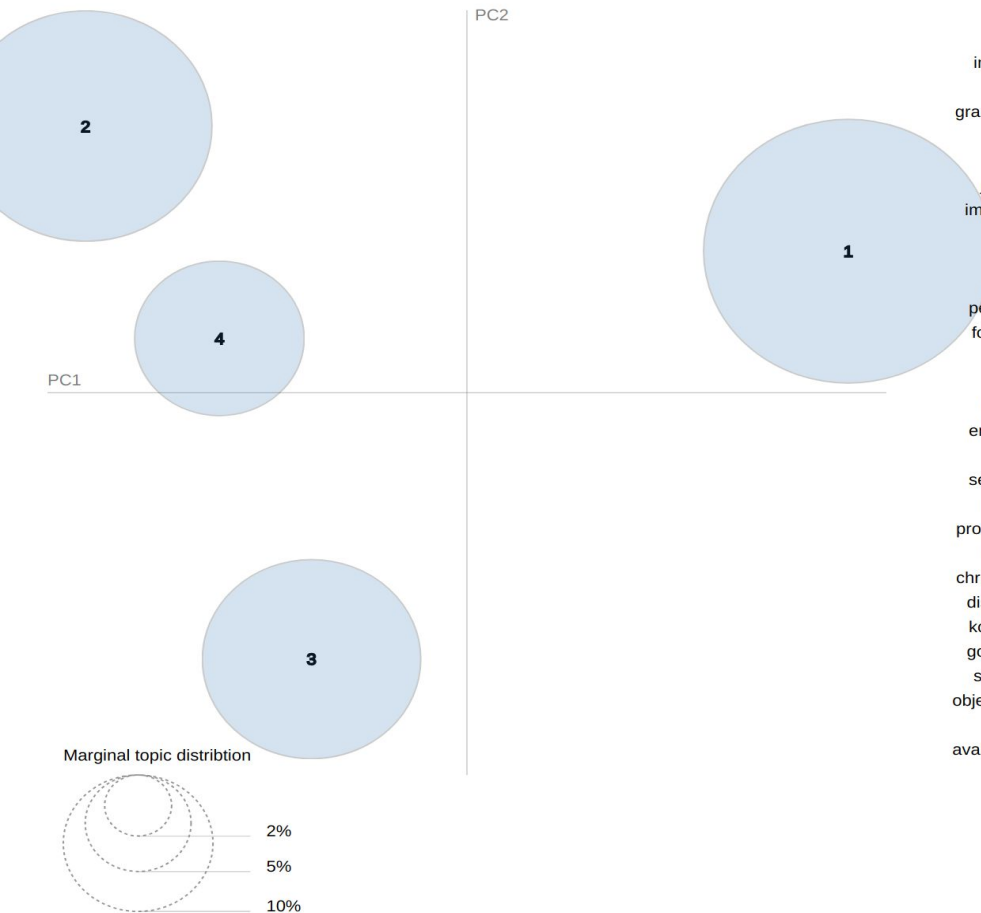
Clusters



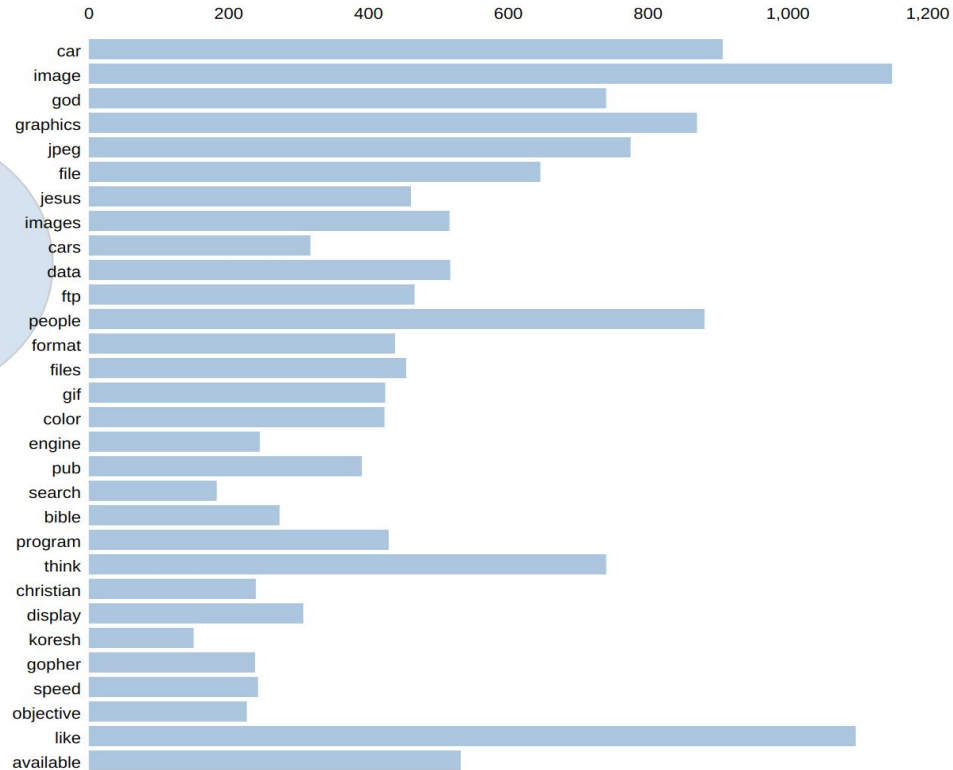
Misclassified

LDA - Overall

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Salient Terms¹



Overall term frequency

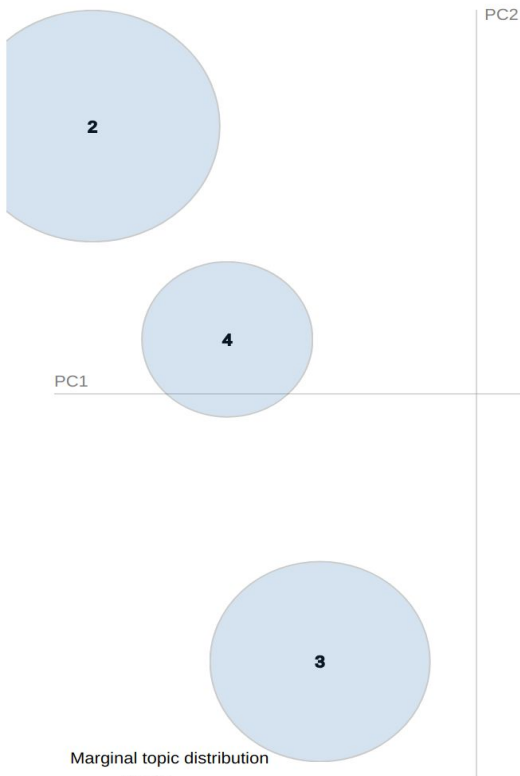
Estimated term frequency within the selected topic

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))]] for topics t; see Chuang et. al (2012)

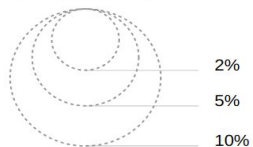
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

LDA - Graphics

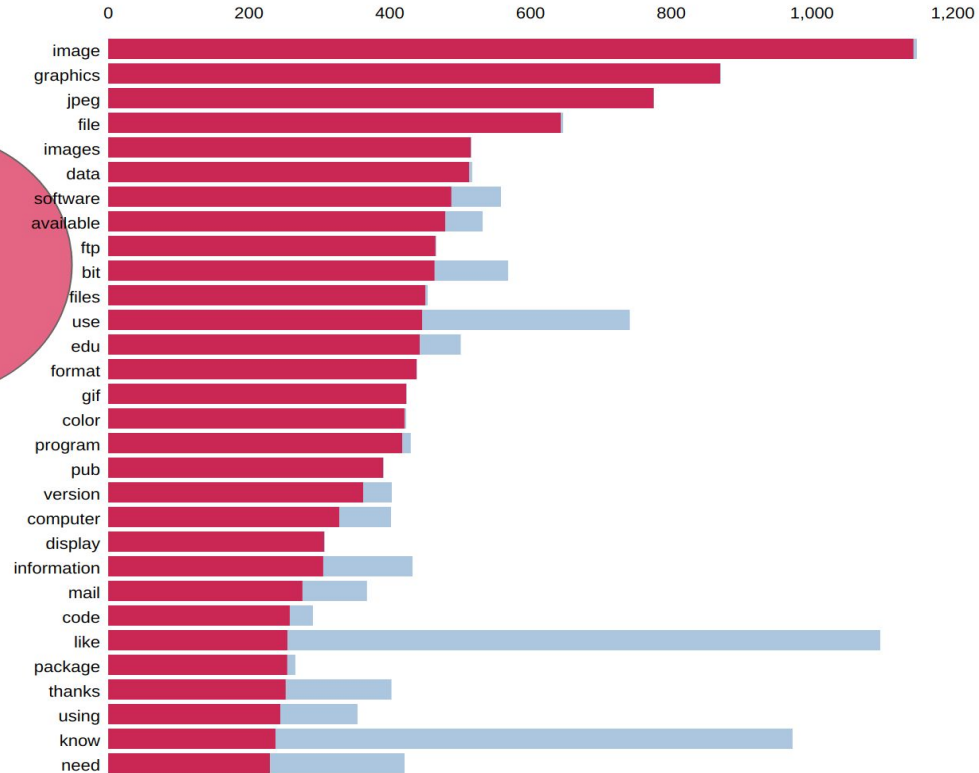
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 1 (37.4% of tokens)



Overall term frequency

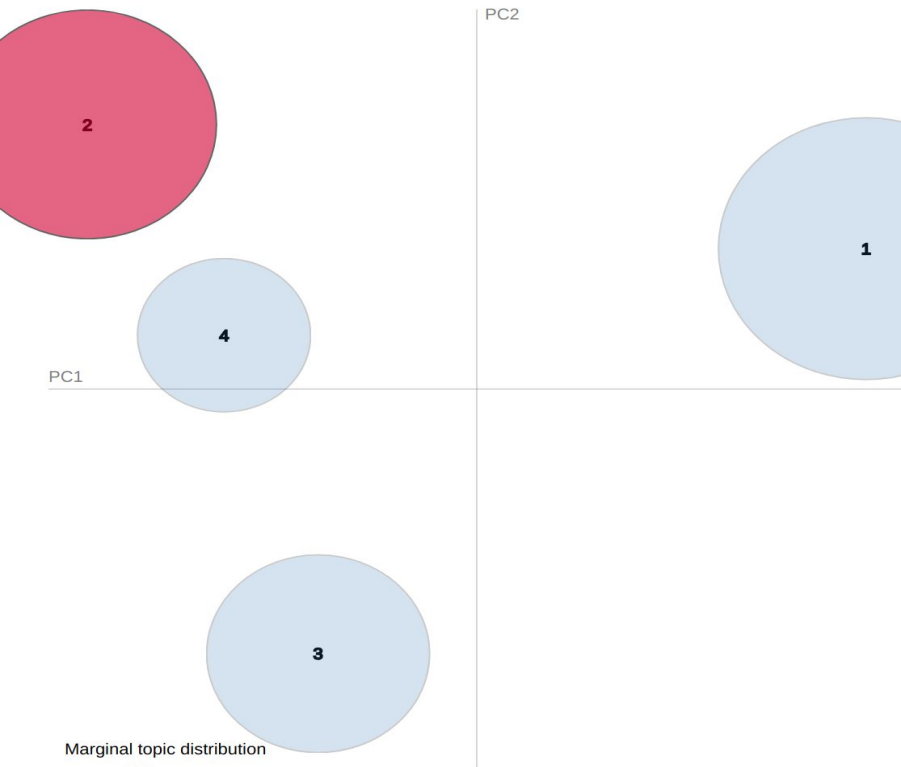
Estimated term frequency within the selected topic

1. $\text{saliency}(\text{term } w) = \text{frequency}(w) * [\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)

2. $\text{relevance}(\text{term } w | \text{topic } t) = \lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

LDA - Religion

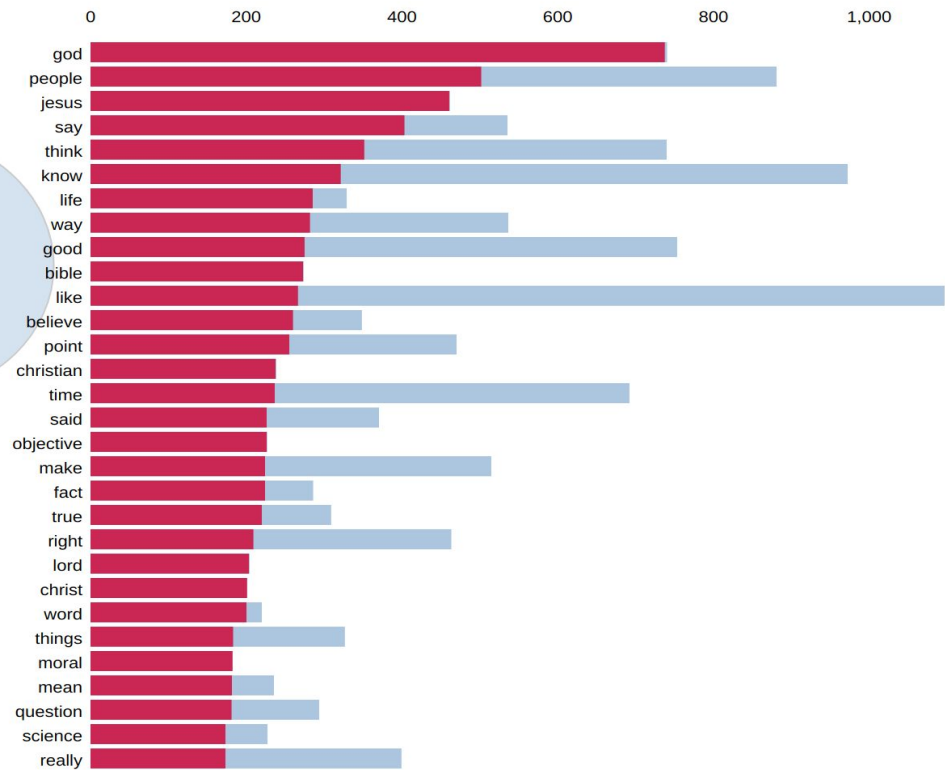
Intertopic Distance Map (via multidimensional scaling)



Marginal topic distribution



Top-30 Most Relevant Terms for Topic 2 (28.5% of tokens)



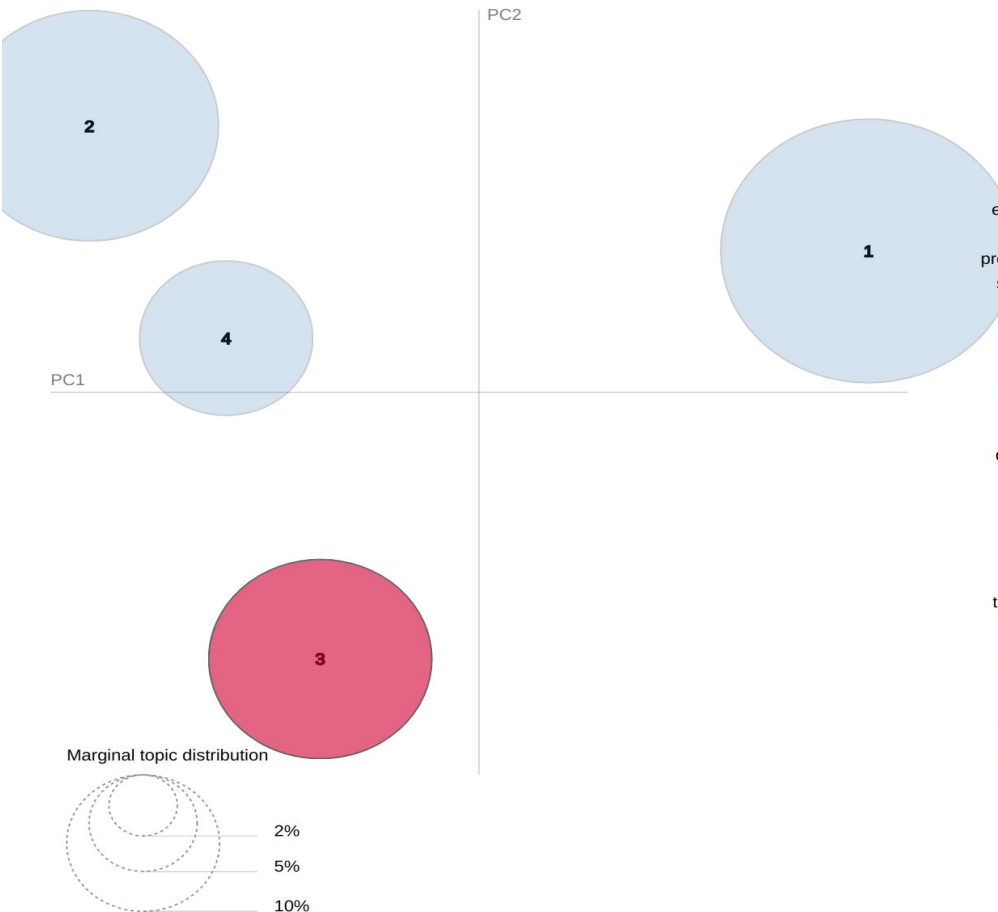
Overall term frequency

Estimated term frequency within the selected topic

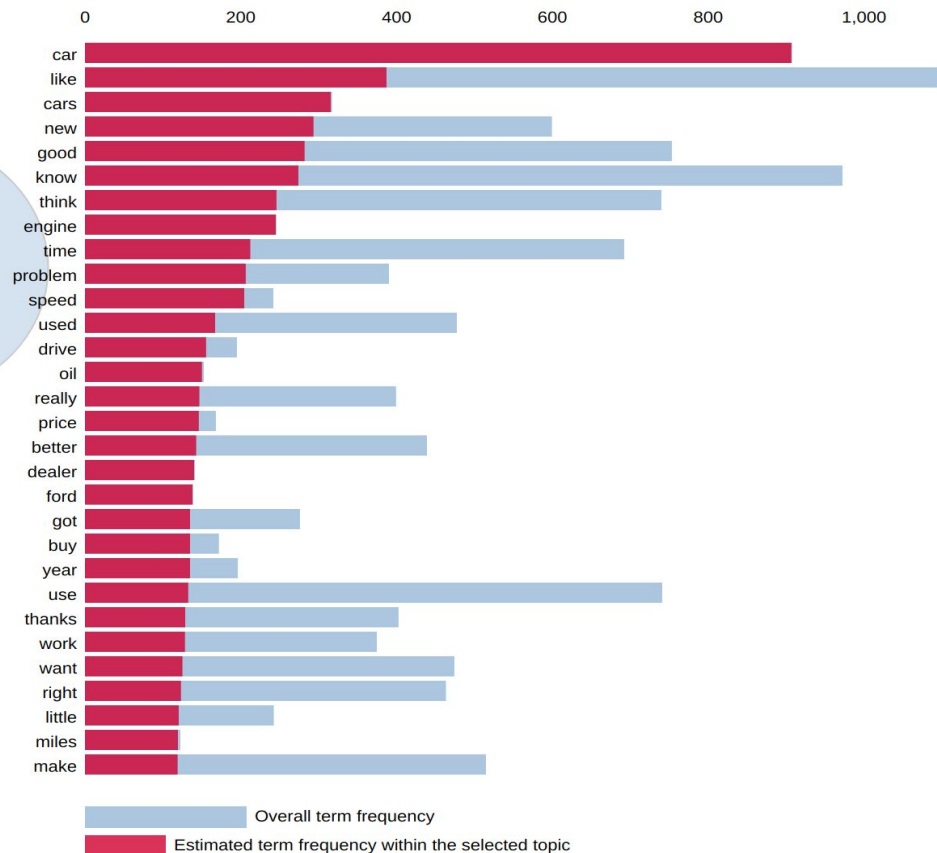
1. saliency(term w) = frequency(w) * $[\sum_t p(t | w) * \log(p(t | w)/p(t))]$ for topics t ; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t)/p(w)$; see Sievert & Shirley (2014)

LDA - CAR

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 3 (21.3% of tokens)



1. saliency(term w) = frequency(w) * $\left[\sum_t p(t | w) * \log(p(t | w) / p(t)) \right]$ for topics t ; see Chuang et. al (2012)
2. relevance(term w | topic t) = $\lambda * p(w | t) + (1 - \lambda) * p(w | t) / p(w)$; see Sievert & Shirley (2014)

*Can we extract the mail **topics**
from a mail box?*

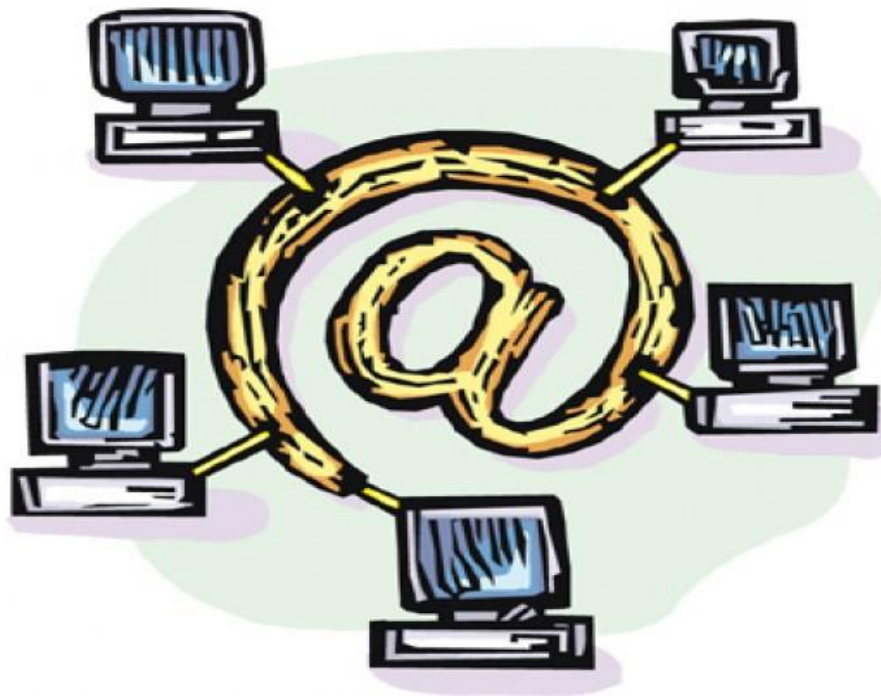


*Let's try with $LSA+K$ -means and
 LDA !*



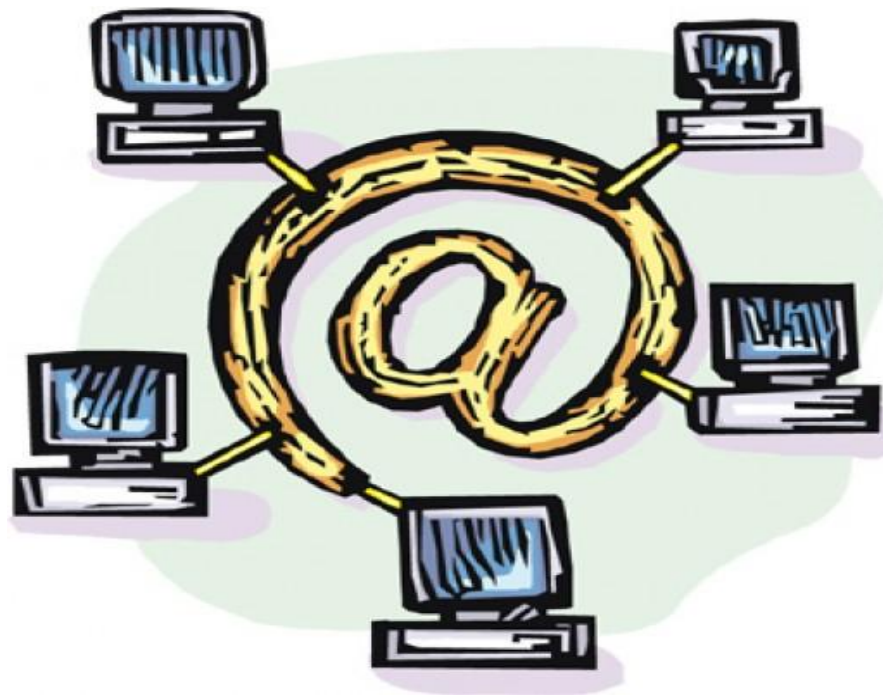
Can we extract the topics from a mail box?

- **Label each e-mail to one topic**
- Discover hidden topics
- 20NewsGroups Dataset
- LSA + K-Means
- Latent Dirichlet Analysis



Can we extract the topics from a mail box?

- **3000** e-mails from the 20Newsgroups dataset
- **3 newsgroups** :
 - comp.graphics
 - rec.cars
 - talk.misc.religion
- **2 modeling attempts:**
 - Dimensionality reduction through LSA + Clustering through K-Means
 - Latent Dirichlet Analysis
- **1.5 Goal(s) :**
 - **Automatic categorization of each e-mail to one of the 3 newsgroups main topic**
 - Find the presence of other topics



Example mail - Clean Up

Newsgroups: talk.religion.misc
From: bskendig@netcom.com (Brian Kendig)
Subject: Re: *** The list of Biblical contradictions
Message-ID: <bskendigC51CqB.K0r@netcom.com>
Organization: Starfleet Headquarters: San Francisco
References: <bskendigC50tnu.lno@netcom.com> <7912@blue.cis.pitt.edu>
Date: Tue, 6 Apr 1993 00:15:46 GMT
Lines: 30



Header

joslin@pogo.isp.pitt.edu (David Joslin) writes:
>
>I'm curious to know what purpose people think these lists serve.
>Lists like this seem to value quantity over quality, an "argument
>from article length." And the list you have here is of poorer
>quality than most.



Quotes

I agree, which is why I've asked for help with it.

The reason I'm working on this list is because I've recently had one too many Christians tell me "the Bible contains no contradictions whatsoever." They believe that it's true, and that it describes reality perfectly, and even predicts history before it happens.

Before I can carry on any sort of meaningful conversation with these people, I've got to SHOW them, with concrete evidence, that the Bible is not nearly as airtight as they thought. I hope to do that with this list.

Specifically: when I bring up the fact that Genesis contains two contradictory creation stories, I usually get blank stares or flat denials. I've never had a fundamentalist acknowledge that there are indeed two different accounts of creation.

--
//_ Brian Kendig Je ne suis fait comme aucun
/_/_ bskendig@netcom.com de ceux que j'ai vus; j'ose croire
//_ n'etre fait comme aucun de ceux qui existent.
/ The meaning of life Si je ne vaux pas mieux, au moins je suis autre.
/ is that it ends. -- Rousseau



Footer

Example mail - (Possibly) Relevant Tokens for Religion Topic

I agree, which is why I've asked for help with it.

The reason I'm working on this list is because I've recently had one too many **Christians** tell me "the **Bible** contains no contradictions whatsoever." They **believe** that it's true, and that it describes reality perfectly, and even predicts history before it happens.

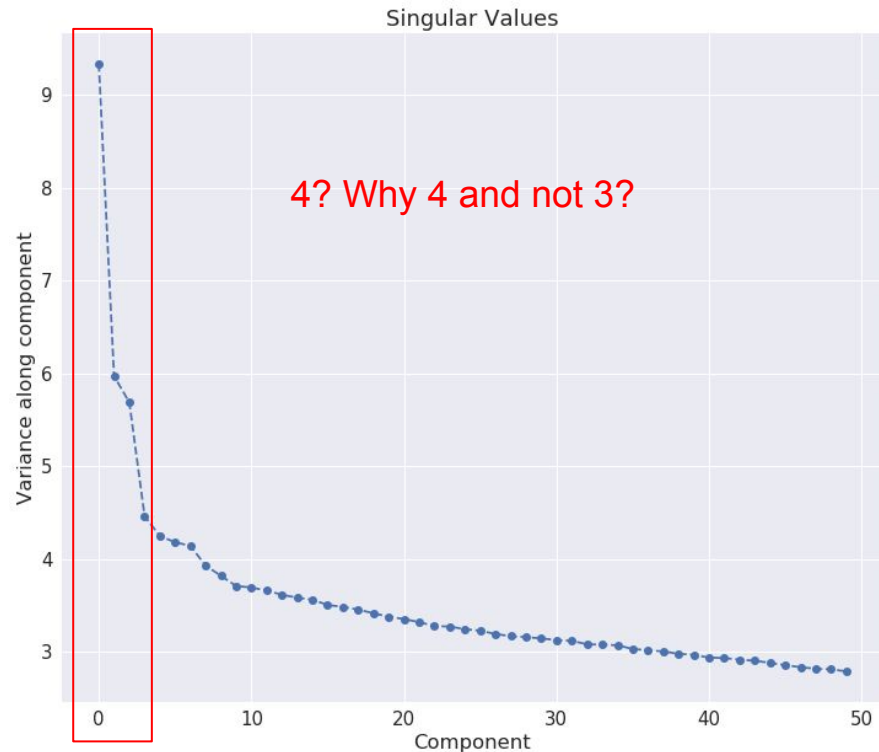
Before I can carry on any sort of meaningful conversation with these people, I've got to SHOW them, with concrete evidence, that the **Bible** is not nearly as airtight as they thought. I **hope** to do that with this list.

Specifically: when I bring up the fact that **Genesis** contains two contradictory **creation** stories, I usually get blank stares or flat denials. I've never had a **fundamentalist** acknowledge that there are indeed two different accounts of **creation**.

LSA

The majority of variance is explained by the first 4 components !!!

- 3000 Mail
- Each mail becomes a point in a 1000-dimensional space (Tfidf matrix)
- Reduction to 50-dimensional space(principal components) using truncated SVD
- Variance along component
- Singular values distribution imply high separation potential



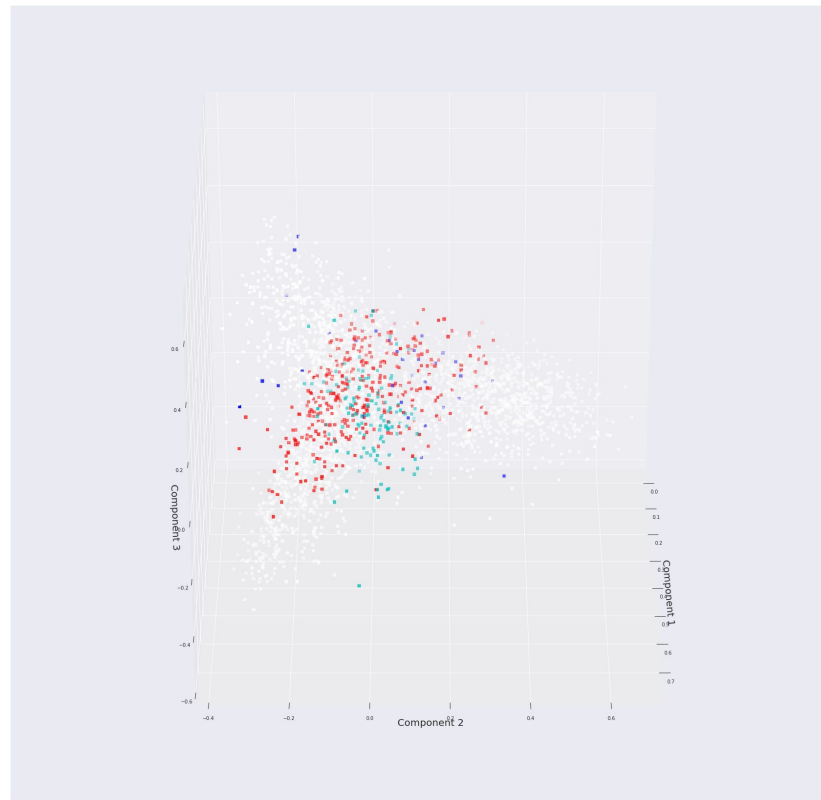
K-means Clusters

- Look for 3 clusters over 50 dimensional space (projected on the first 3 components)
- Most frequent terms per cluster:
 - Cluster 0: people god think like say know life make good jesus
 - Cluster 1: thanks graphics image know program files looking like file use
 - Cluster 2: car cars engine new like good speed ford dealer problem



K-means Clusters -results

- 513 Misclassification over 3000 e-mails:
 - Messages from rec.autos not in Cluster 2: **35.3 % (red)**
 - Messages from comp.graphics not in Cluster 1: **12.6% (cyan)**
 - Messages from talk.religion.misc not in Cluster 0: **3.4% (blue)**
 - Message in white correctly classified



Conclusions

- Dimensionality reduction through SVD improves the quality of clustering techniques for topic modeling
- Comparing the value of the components conveys informations about the topic distribution
- Extending the number of clusters beyond the number of expected categories by looking at the singular values can lead to the identification of new topics!
- **Future :**
 - Extend number of newsgroups
 - Train deep learning model on each clusters