Neural Word Embeddings: Adding the Human into the Loop

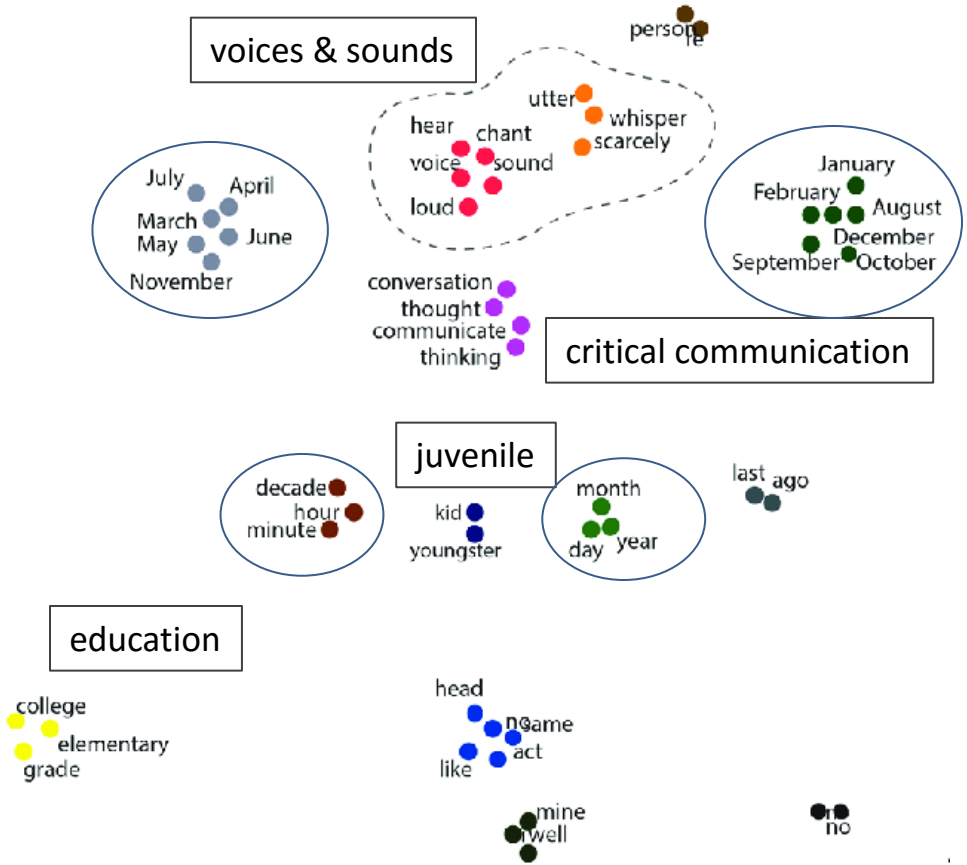
Klaus Mueller, PhD

Visual Analytics and Imaging (VAI) Lab
Computer Science Department
Stony Brook University

What's a Word Embedding?

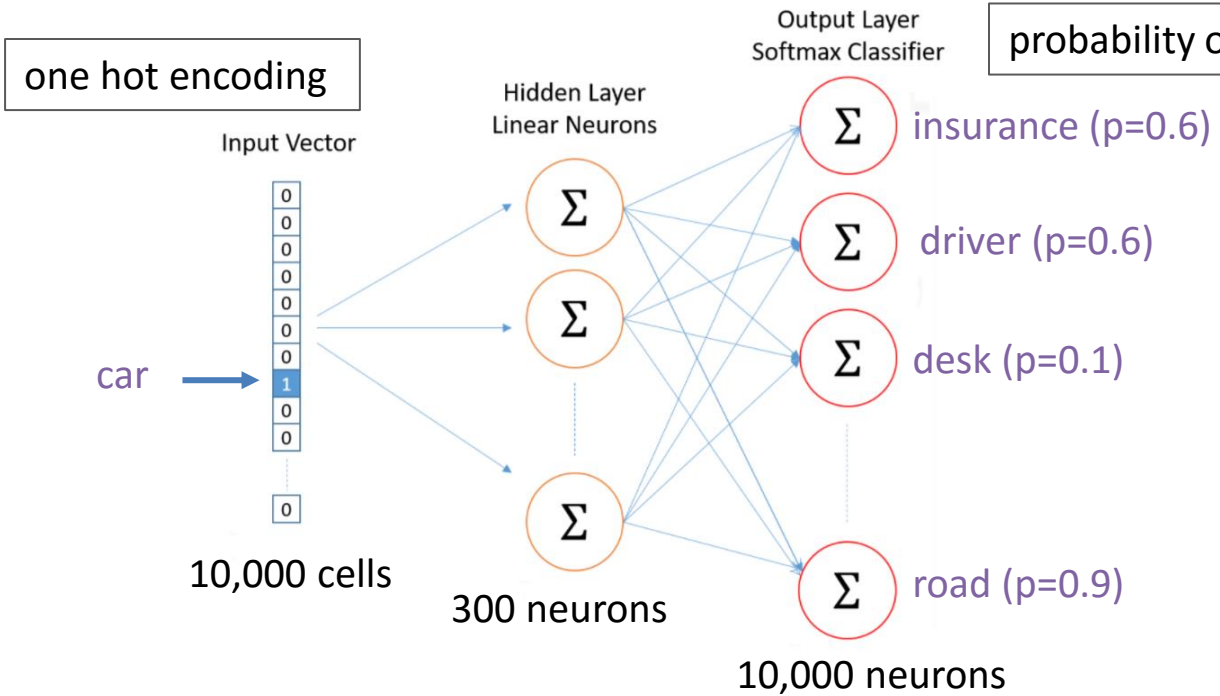
A mapping where words associated with similar concepts locate into close neighborhoods

- Layout is determined by the mapping's training process
- Isn't always perfect
- It pays to visually confirm (and edit) the layout
- Add the human in the loop



What's a *Neural* Word Embedding?

Use a shallow neural network to determine the embedding

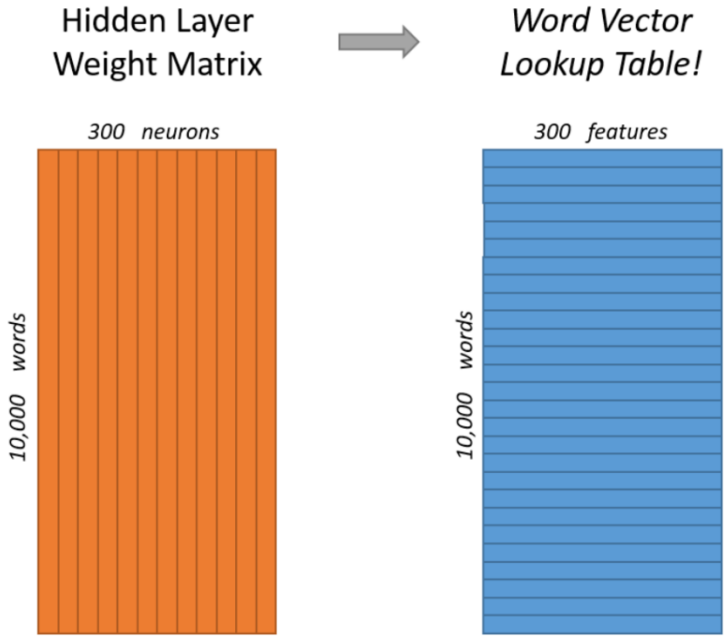


text corpus (massive)

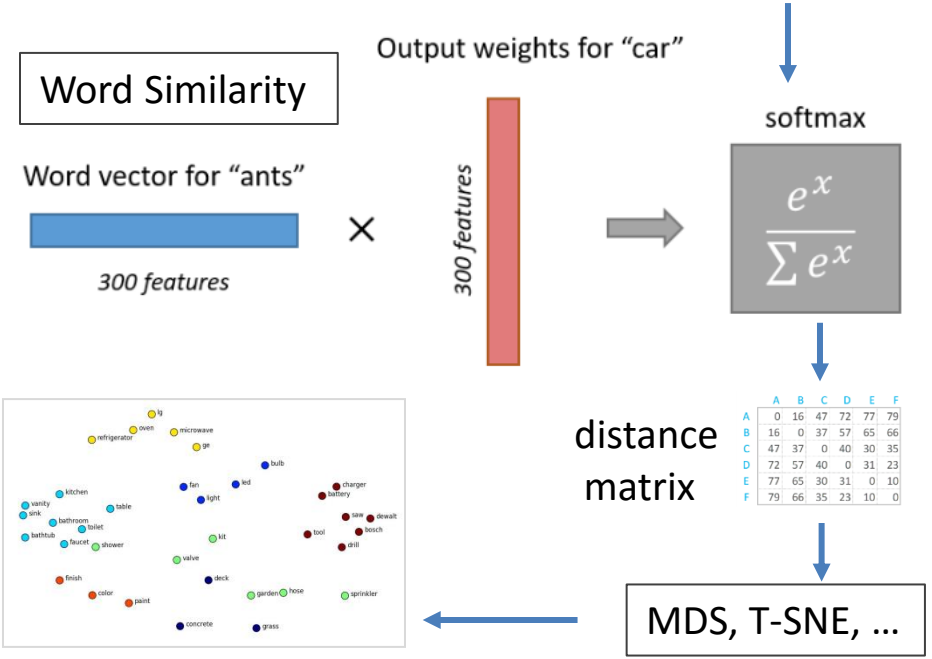
…
..[the car cruises the road]..
..[a driver steers his car]..
..[a road full of cars]..
..[do you have car insurance]
..[a desk in a room]..
…

Constructing a Word Embedding

What we will keep are the hidden layer weights



Probability that if you randomly pick a word nearby "ants", that it is "car"



Implementations

Word2Vec, Glove

- Train them with your own text corpus
- Inherent problem is “conflation of word sense”
- There is only one mapping per word

More advanced algorithms like BERT, etc.

- Operate on a wider, sentence-level context
- Can produce multiple mappings for a given word based on its semantics
- We have used Word2Vec in our work due to its simplicity

Applications

Next I will discuss a few applications of Word Embeddings

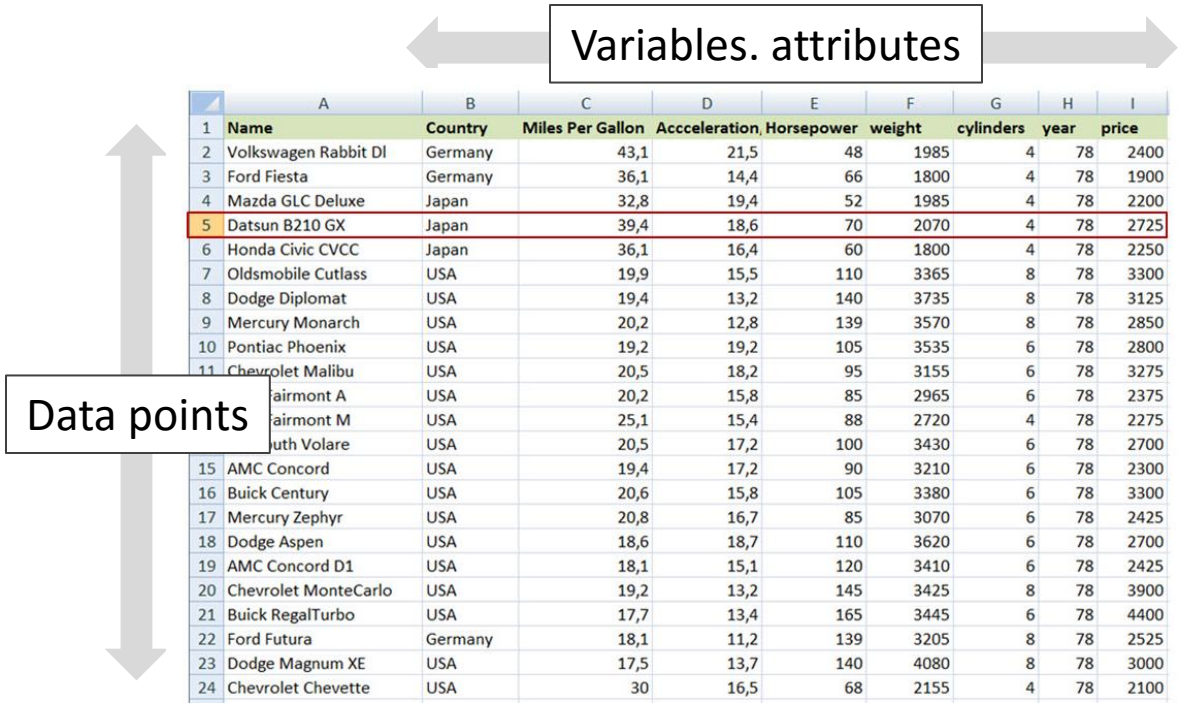
- Semantic Subspace Clustering for High-Dimensional Data
- Construction of Semantic/Data Driven Attribute Hierarchies
- Detecting Word Bias in Word Embeddings

All were done in the VAI Lab

- Salman Mahmood (PhD, now at Google)
- Bhavya Ghai, Md Naimul Hoque (PhD 2B)

High-Dimensional Data

Any dataset with many rows (data points) and many columns (variables, attributes)



	A	B	C	D	E	F	G	H	I
1	Name	Country	Miles Per Gallon	Accceleration	Horsepower	weight	cylinders	year	price
2	Volkswagen Rabbit Dl	Germany	43,1	21,5	48	1985	4	78	2400
3	Ford Fiesta	Germany	36,1	14,4	66	1800	4	78	1900
4	Mazda GLC Deluxe	Japan	32,8	19,4	52	1985	4	78	2200
5	Datsun B210 GX	Japan	39,4	18,6	70	2070	4	78	2725
6	Honda Civic CVCC	Japan	36,1	16,4	60	1800	4	78	2250
7	Oldsmobile Cutlass	USA	19,9	15,5	110	3365	8	78	3300
8	Dodge Diplomat	USA	19,4	13,2	140	3735	8	78	3125
9	Mercury Monarch	USA	20,2	12,8	139	3570	8	78	2850
10	Pontiac Phoenix	USA	19,2	19,2	105	3535	6	78	2800
11	Chevrolet Malibu	USA	20,5	18,2	95	3155	6	78	3275
12	airmont A	USA	20,2	15,8	85	2965	6	78	2375
13	airmont M	USA	25,1	15,4	88	2720	4	78	2275
14	uth Volare	USA	20,5	17,2	100	3430	6	78	2700
15	AMC Concord	USA	19,4	17,2	90	3210	6	78	2300
16	Buick Century	USA	20,6	15,8	105	3380	6	78	3300
17	Mercury Zephyr	USA	20,8	16,7	85	3070	6	78	2425
18	Dodge Aspen	USA	18,6	18,7	110	3620	6	78	2700
19	AMC Concord D1	USA	18,1	15,1	120	3410	6	78	2425
20	Chevrolet MonteCarlo	USA	19,2	13,2	145	3425	8	78	3900
21	Buick RegalTurbo	USA	17,7	13,4	165	3445	6	78	4400
22	Ford Futura	Germany	18,1	11,2	139	3205	8	78	2525
23	Dodge Magnum XE	USA	17,5	13,7	140	4080	8	78	3000
24	Chevrolet Chevette	USA	30	16,5	68	2155	4	78	2100

Projection into 2D

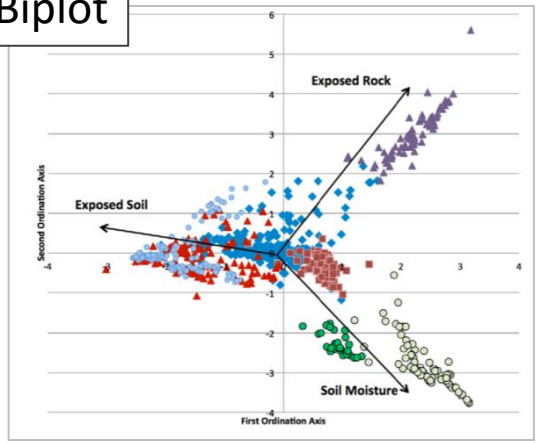
Required for visualization

- Principal Component Analysis (PCA)
- Multidimensional Scaling (MDS)
- t-SNE, UMAP, etc.

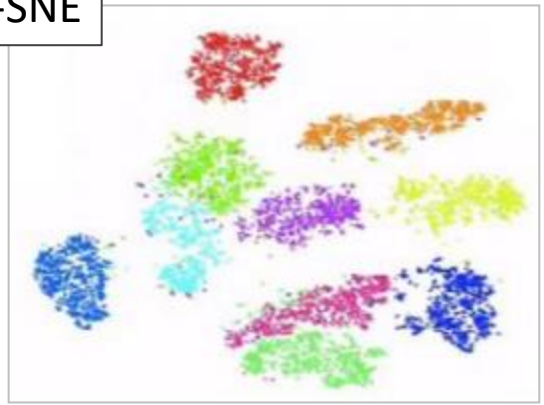
The essential problem with all of these is:

- Involving too many dimensions can hide subtle patterns in the data
- We would want to isolate these structures prior to projection

Biplot



t-SNE



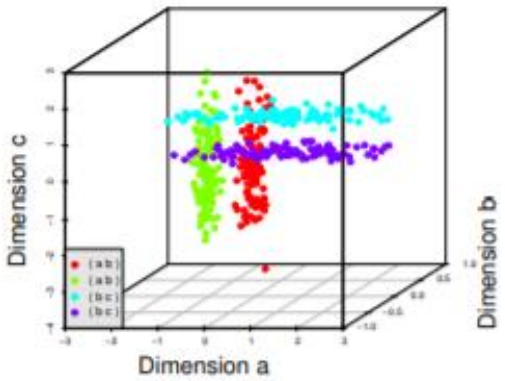
Dimension Reduction in Subspaces

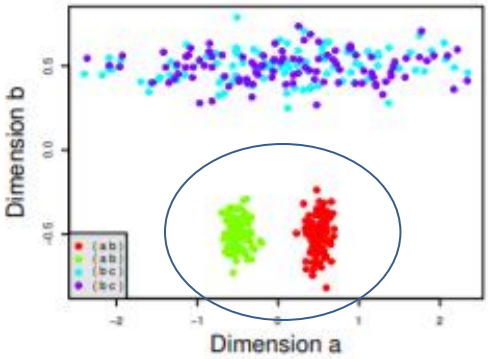
Think of a high-D dataset as an eco-system of subspace clusters

- Each subspace cluster can have a different relevant attribute set
- Subspace clustering can significantly reduce data dimensionality

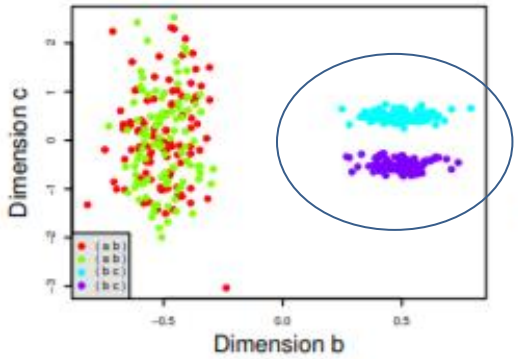
What's a Subspace?

A subspace is a projection of data into a reduced dimension set that isolates a particular cluster

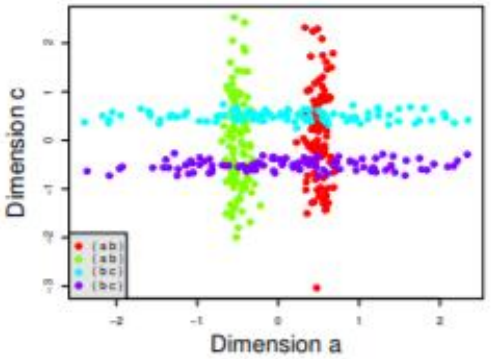




(a) Dims a & b



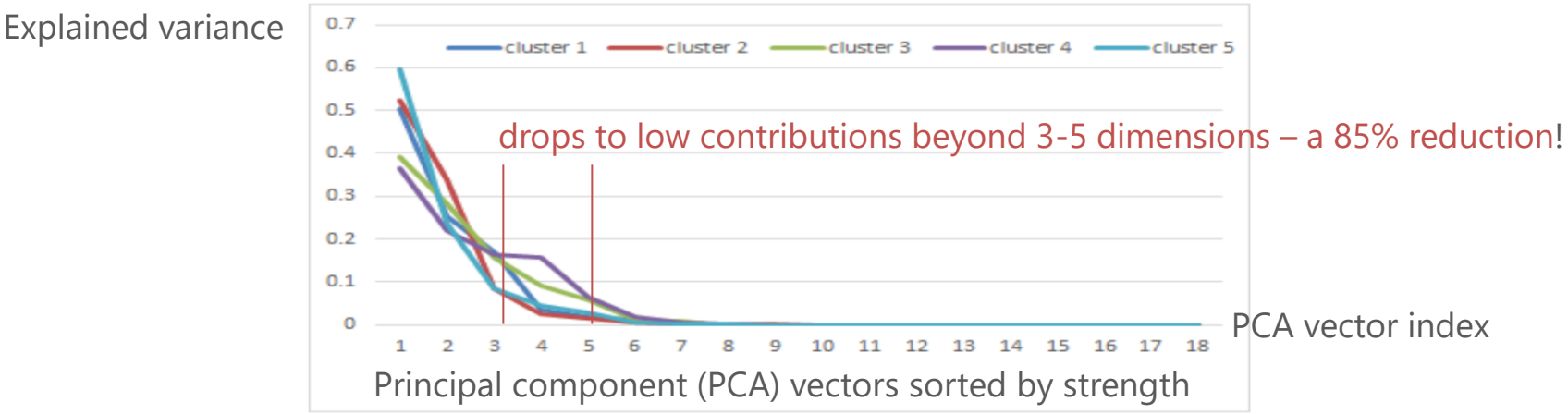
(b) Dims b & c



(c) Dims a & c

Dimension Reduction in Subspaces

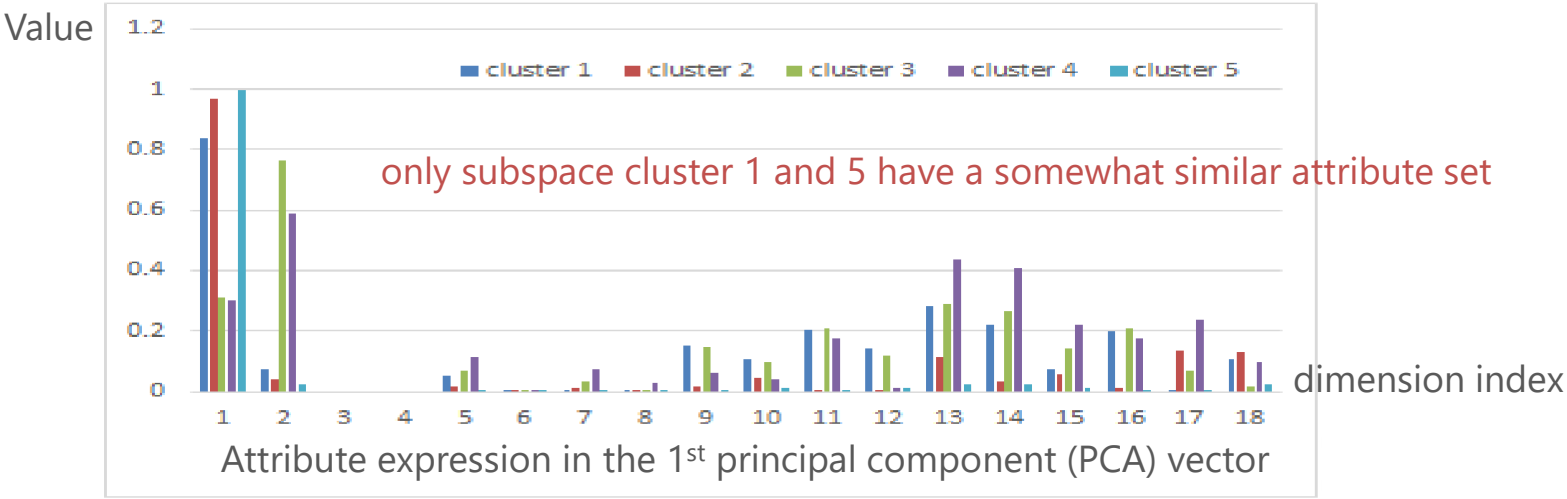
Example: Image segmentation dataset with 19 dimension



B. Wang. K. Mueller, "Does 3D really make sense for visual cluster analysis? Yes!" International Workshop on 3DVis: Does 3D Really Make Sense for Data Visualization? (held jointly with VIS 2014), Paris, France, November 2014.

Dimension Reduction in Subspaces

Example: Image segmentation dataset with 19 dimension



B. Wang. K. Mueller, "Does 3D really make sense for visual cluster analysis? Yes!" International Workshop on 3DVis: Does 3D Really Make Sense for Data Visualization? (held jointly with VIS 2014), Paris, France, November 2014.

Subspace Clustering is Non-Trivial

Subspace clustering is not as easy as it sounds

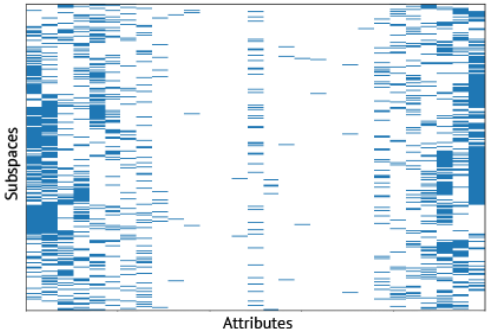
- some algorithms are CLIQUE, SURFING, …

SURFING

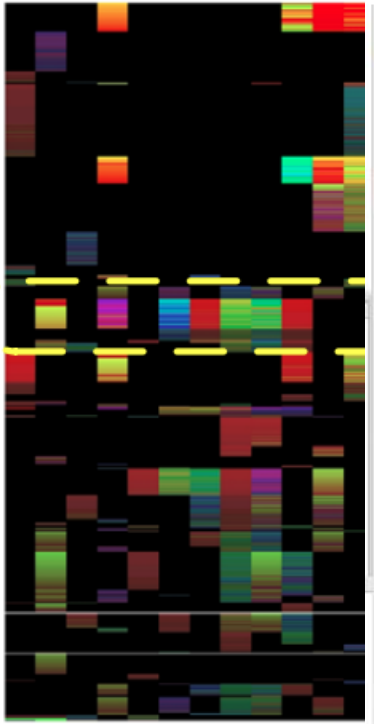
- runs bottom-up using heuristic search with a quality metric
- generates new subspaces from subspaces already known to be interesting

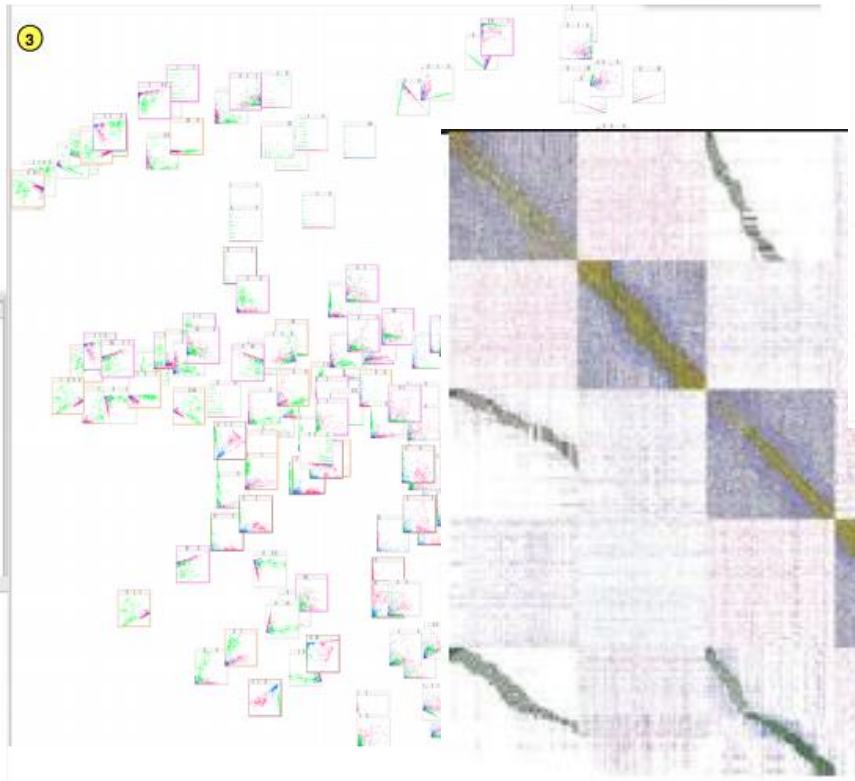
Example

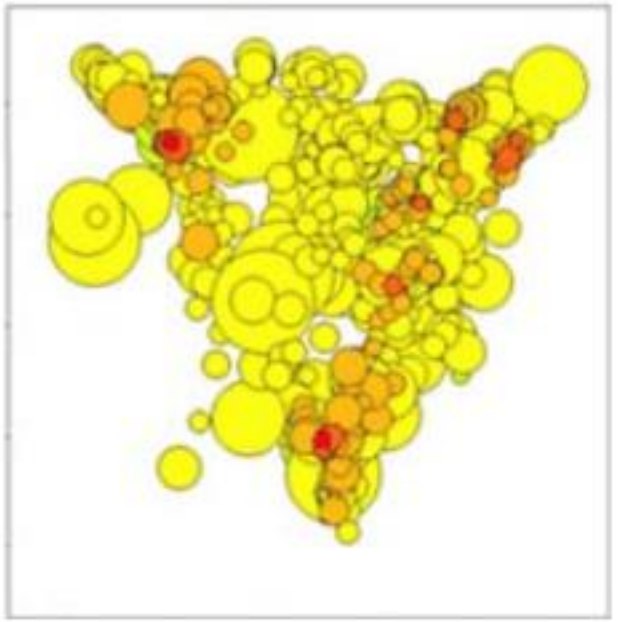
- Filipino Family Income & Expenditure dataset
- generates a whopping 2,011 subspaces!
- ≥3 dimensions shown
- ordered with Jaccard distance

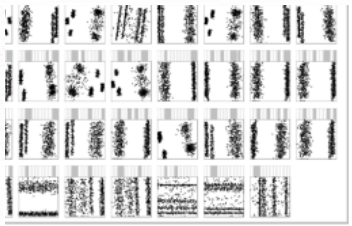


Subspace Explosion

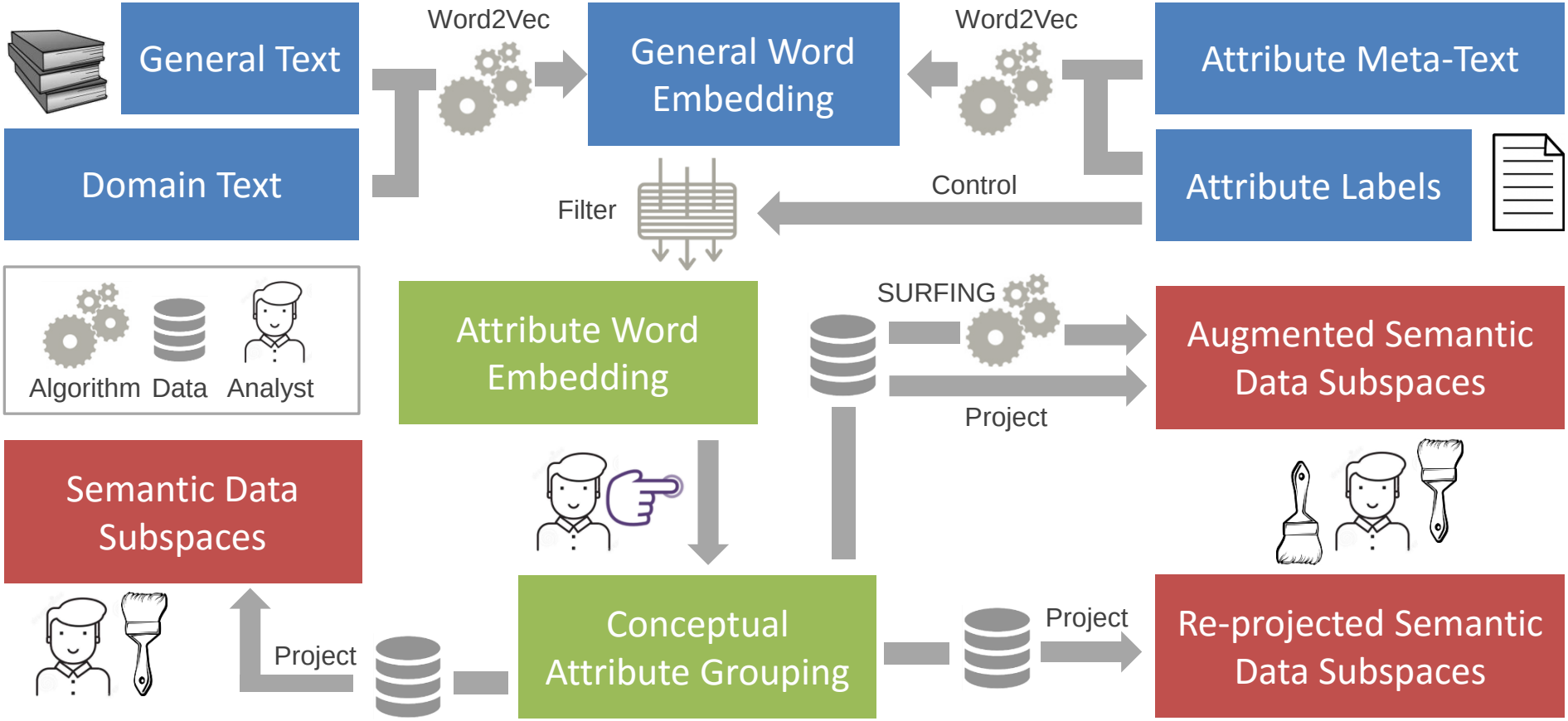








Our Solution:
Semantic Subspace Clustering


General Text
Word2Vec
General Word Embedding
Word2Vec
Attribute Meta-Text
Domain Text
Filter
Control
Attribute Labels
Algorithm
Data
Analyst
Attribute Word Embedding
SURFING
Augmented Semantic Data Subspaces
Project
Semantic Data Subspaces
Project
Conceptual Attribute Grouping
Project
Re-projected Semantic Data Subspaces

Filipino Income & Expenditures Dataset

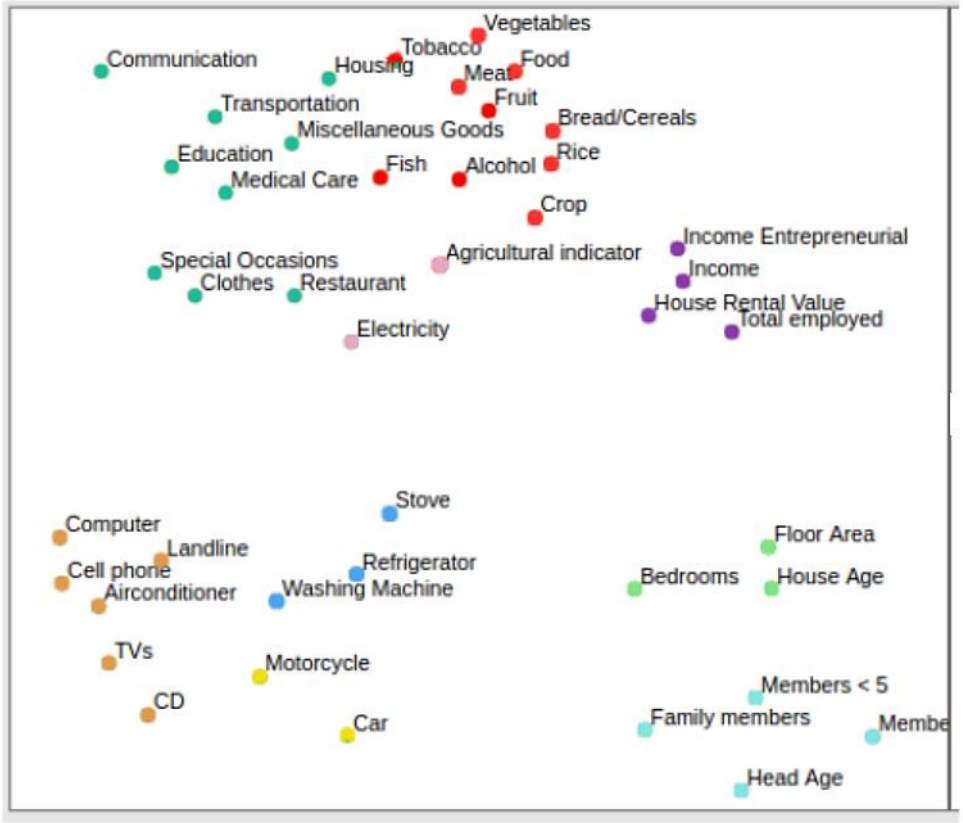
Semantic clusters derived from the Attribute Word Embedding

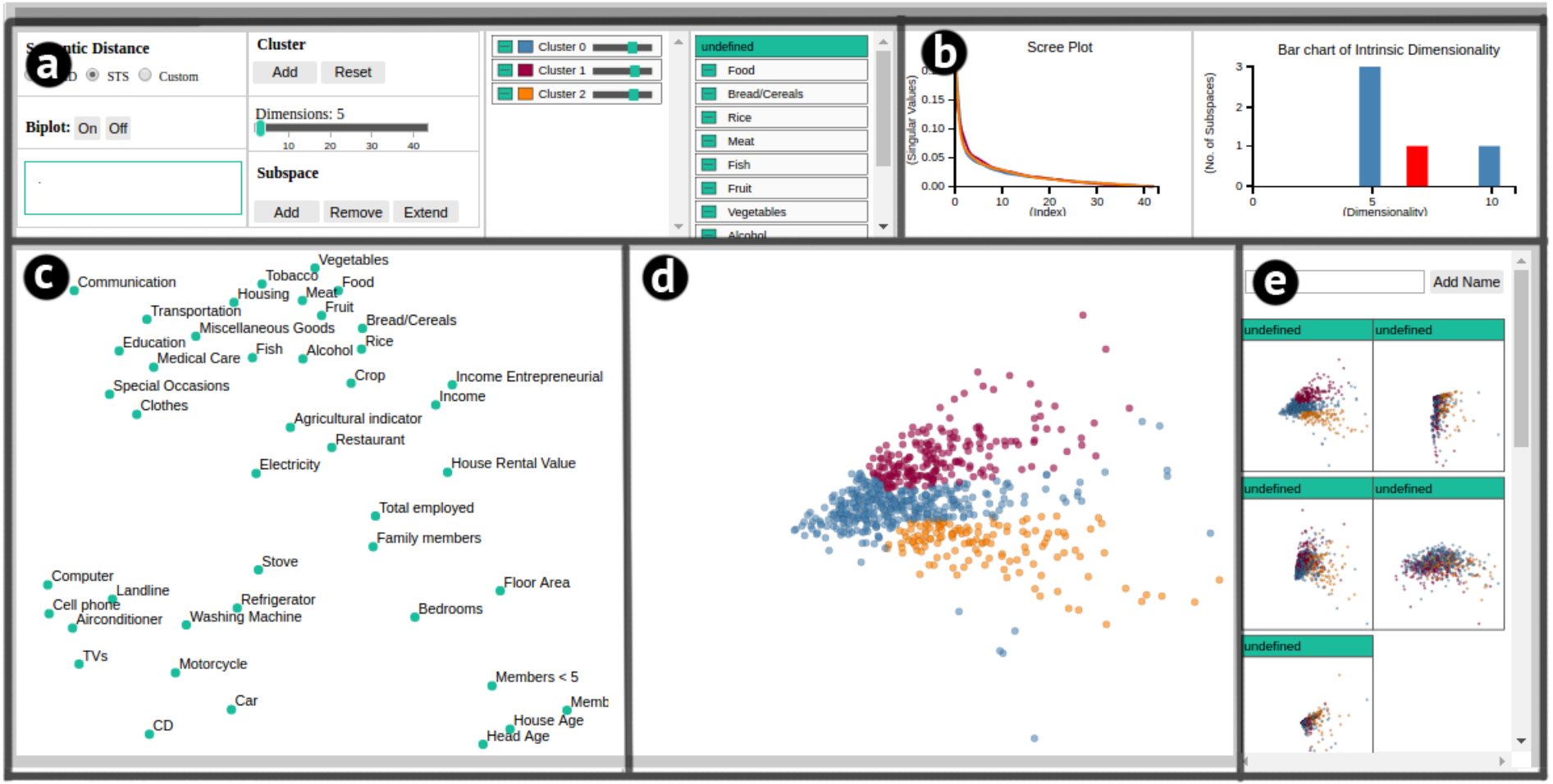
Expenditures
- Food
- Utilities

Ownership
- Electronics
- Transport
- Appliances

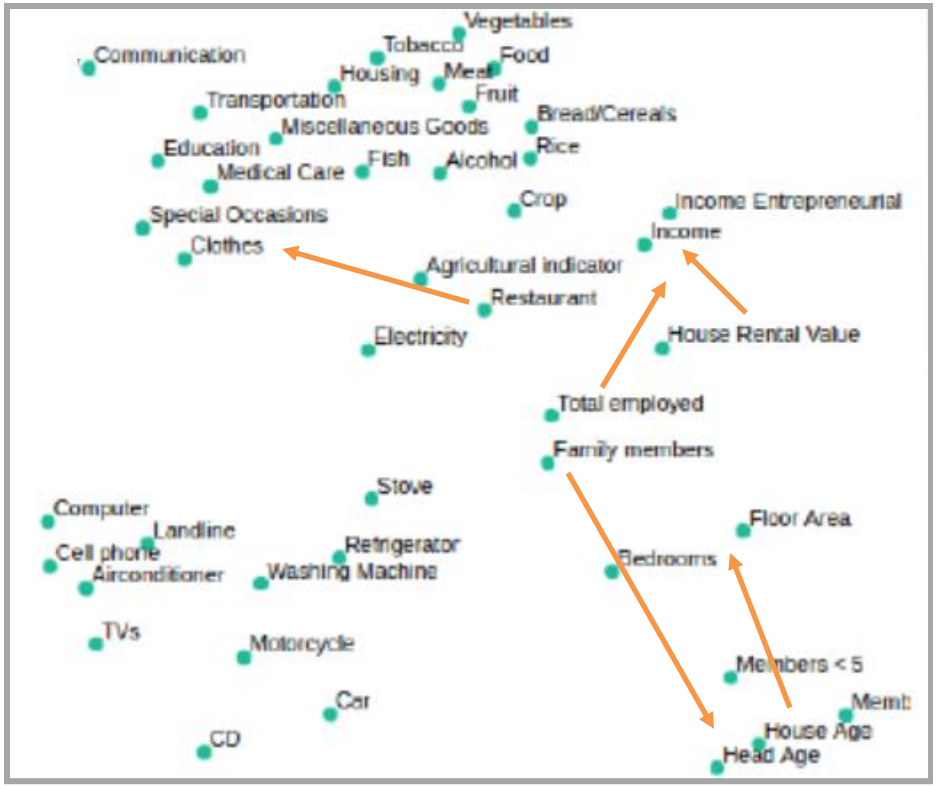
Housing

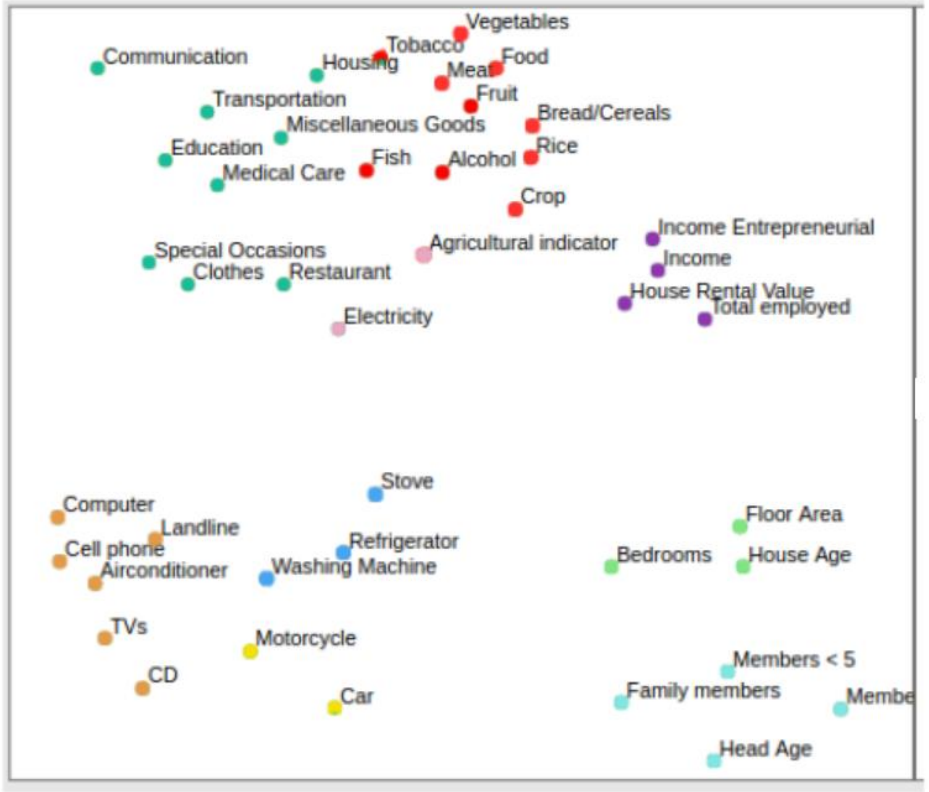
Income




a
Semantic Distance
D
STS
Custom
Cluster
Add
Reset
Biplot:
On
Off
Dimensions: 5
10
20
30
40
Subspace
Add
Remove
Extend
Cluster 0
Cluster 1
Cluster 2
undefined
Food
Bread/Cereals
Rice
Meat
Fish
Fruit
Vegetables
Alcohol
b
Scree Plot
(Singular Values)
0.15
0.10
0.05
0.00
0
10
20
30
40
(Index)
Bar chart of Intrinsic Dimensionality
(No. of Subspaces)
3
2
1
0
0
5
10
(Dimensionality)
c
Vegetables
Communication
Tobacco
Food
Housing
Meat
Transportation
Fruit
Miscellaneous Goods
Bread/Cereals
Education
Rice
Fish
Alcohol
Medical Care
Crop
Income Entrepreneurial
Special Occasions
Income
Clothes
Agricultural indicator
Restaurant
Electricity
House Rental Value
Total employed
Family members
Stove
Computer
Floor Area
Landline
Refrigerator
Cell phone
Bedrooms
Airconditioner
Washing Machine
TVs
Motorcycle
Members < 5
Car
CD
House Age
Head Age
d
e
Add Name
undefined
undefined
undefined
undefined
undefined

Semantic Space Interactions





Semantic Subspace Exploration (1)

Biplot reveals two dimension clusters

- Crop, Bread, Rice (basic staples)
- Other (more luxurious) foods

Distinguish these for further visual analysis

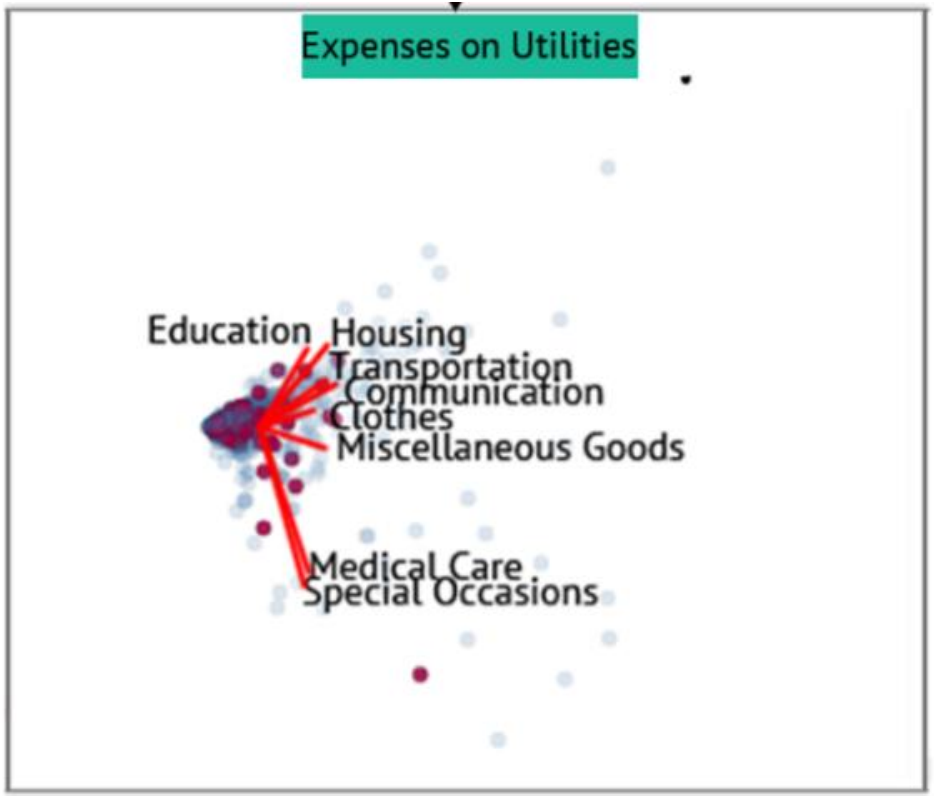
- Color these “Basic Staples” household points in magenta

Semantic Subspace Exploration (2)

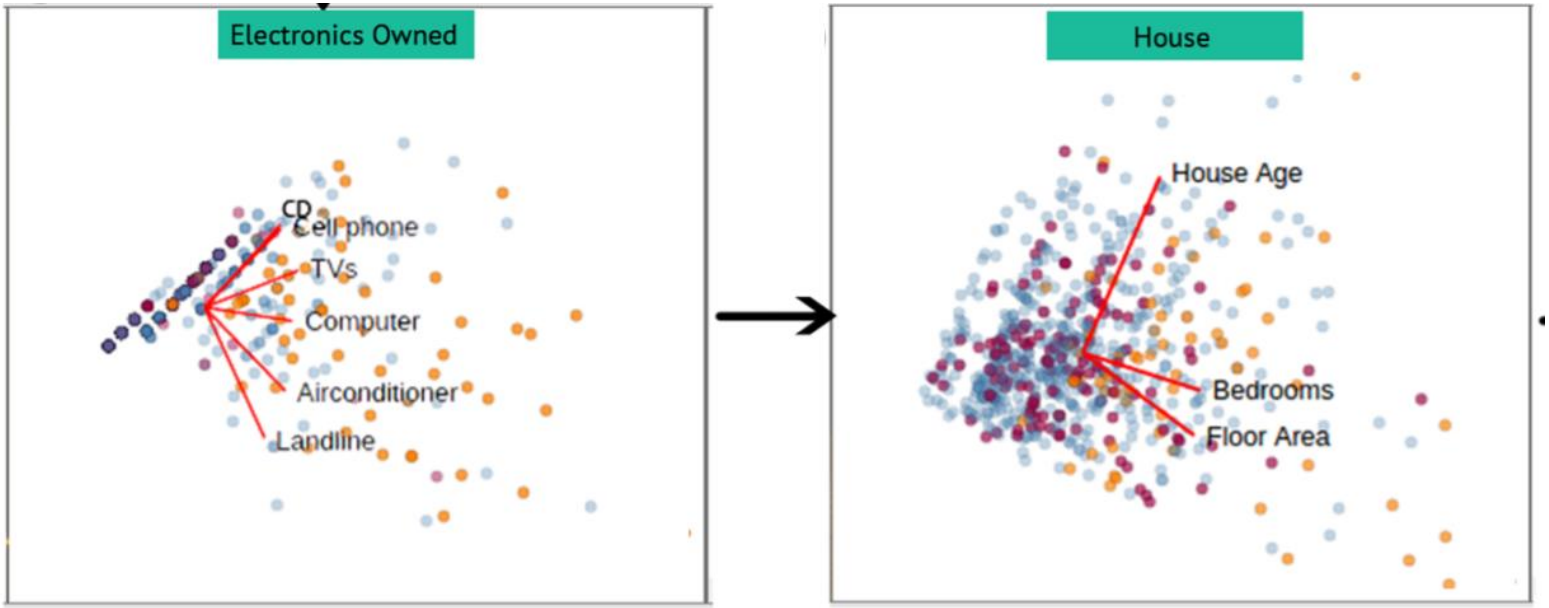
Make blue clusters less opaque to focus on the magenta “Basic Staples” households

- “Basic staples” households do not spend much on utilities
- Medical Care is different from other utilities, similar to Special Occasions

Color “high spenders” in yellow



Semantic Subspace Exploration (3)



"High spender" households own more electronics and live in larger houses

"Basic staples" households own few electronics (left) and smaller houses (right)

There is no distinction in terms of house age (right)

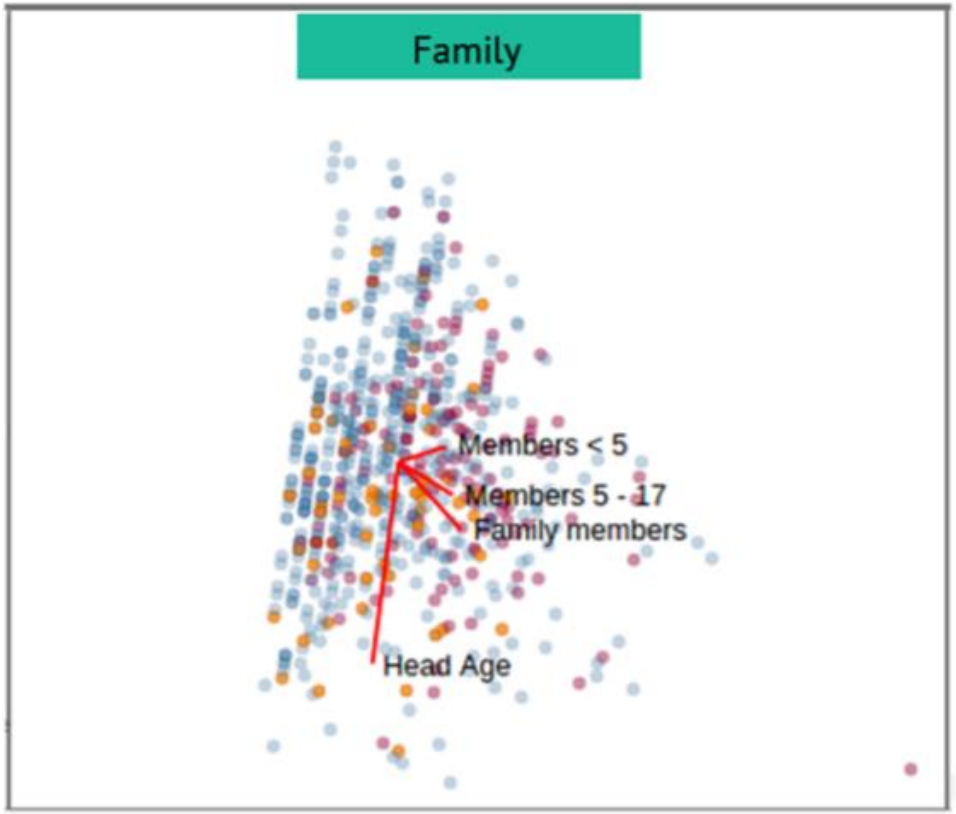
Semantic Subspace Exploration (4)

"Basic staples" households are more evenly distributed

"High spenders" seems to be older families

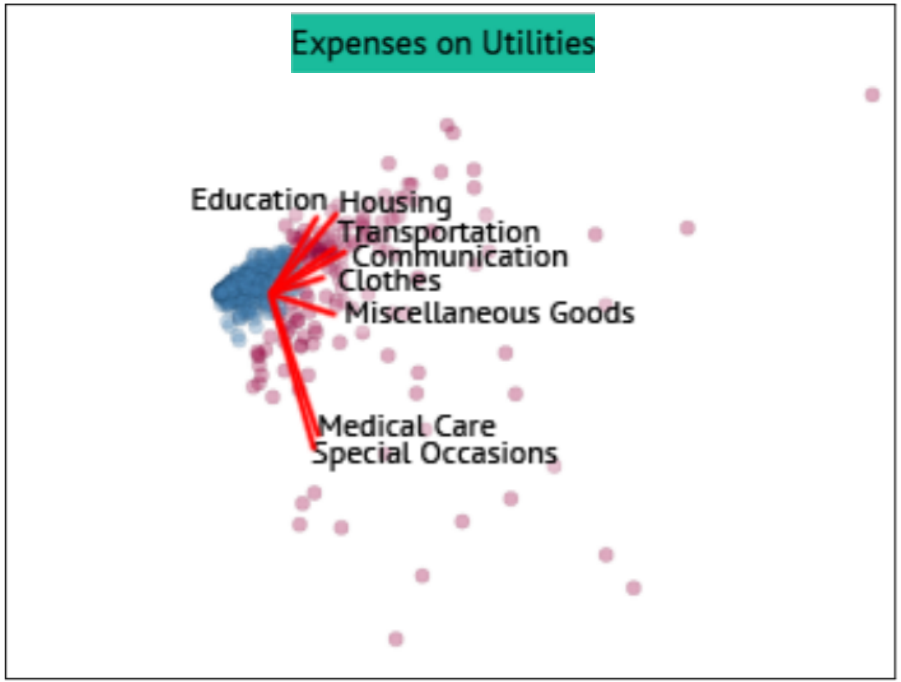
The overall finding is thus:

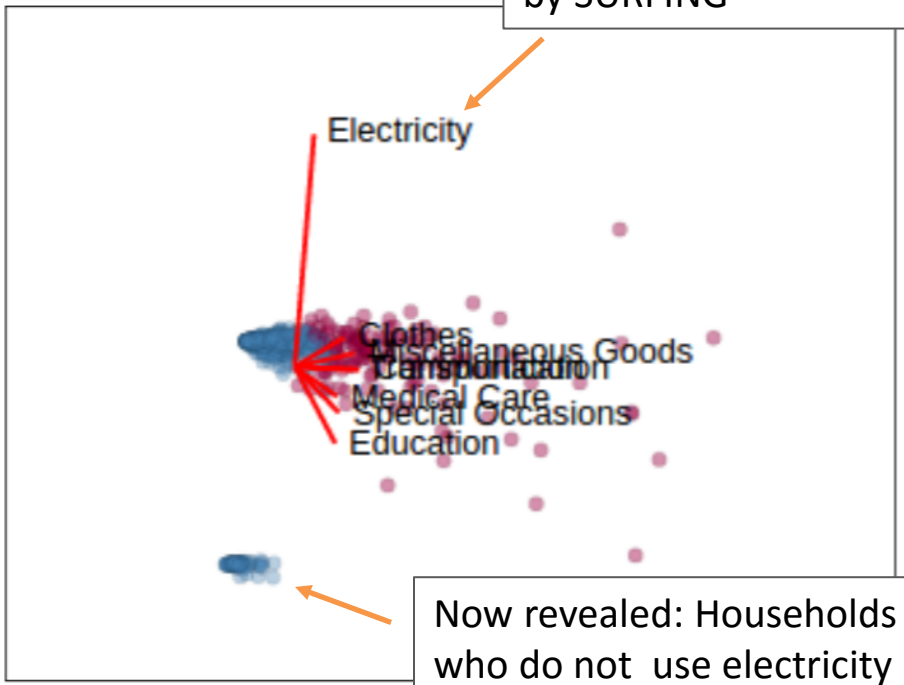
- In Filipino families Rice, Bread/Cereal and Crops make up a major portion of the food consumption in households with less economic resources.



Extend a Semantic Subspace

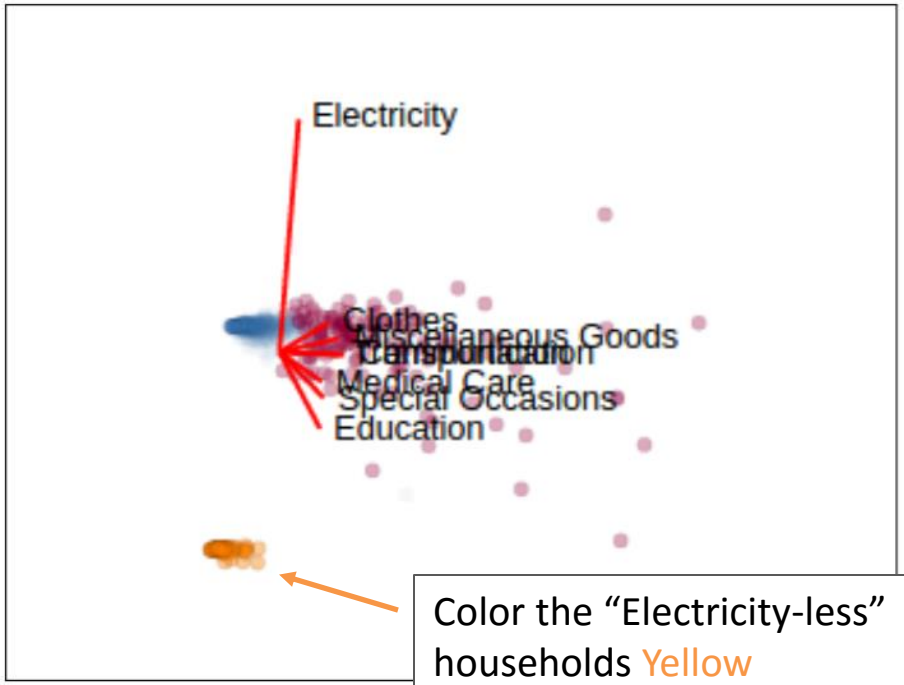
Using SURFING to observe exterior influences





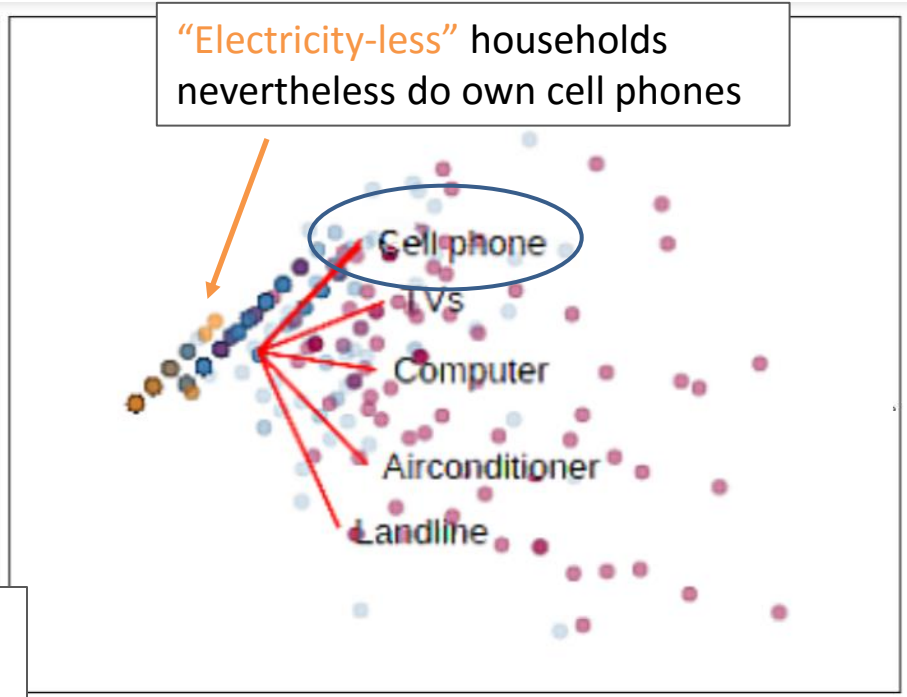
Explore It Further

Electronics use electricity – are there differences?

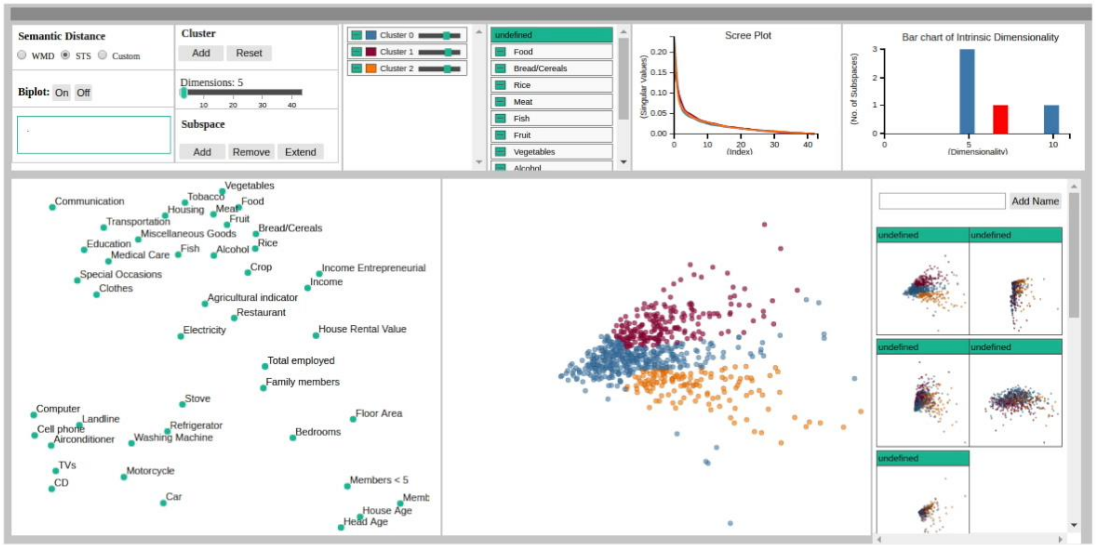


Color the “Electricity-less” households Yellow

“Electricity-less” households nevertheless do own cell phones



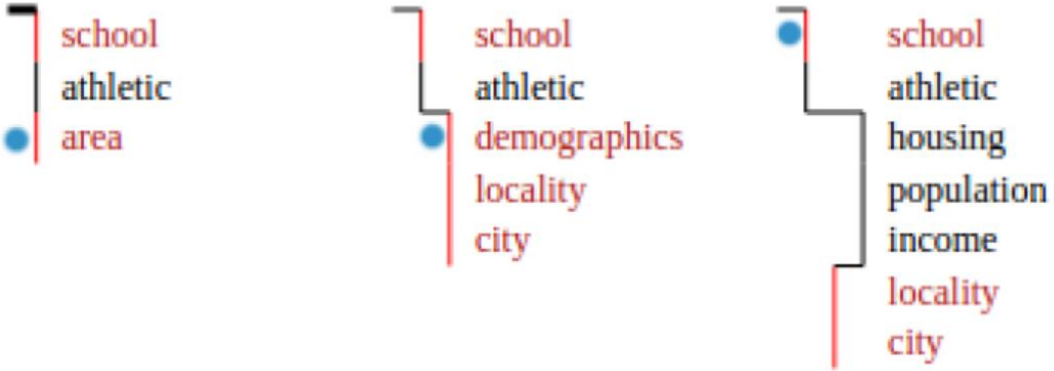
Semantic Subspace Learning



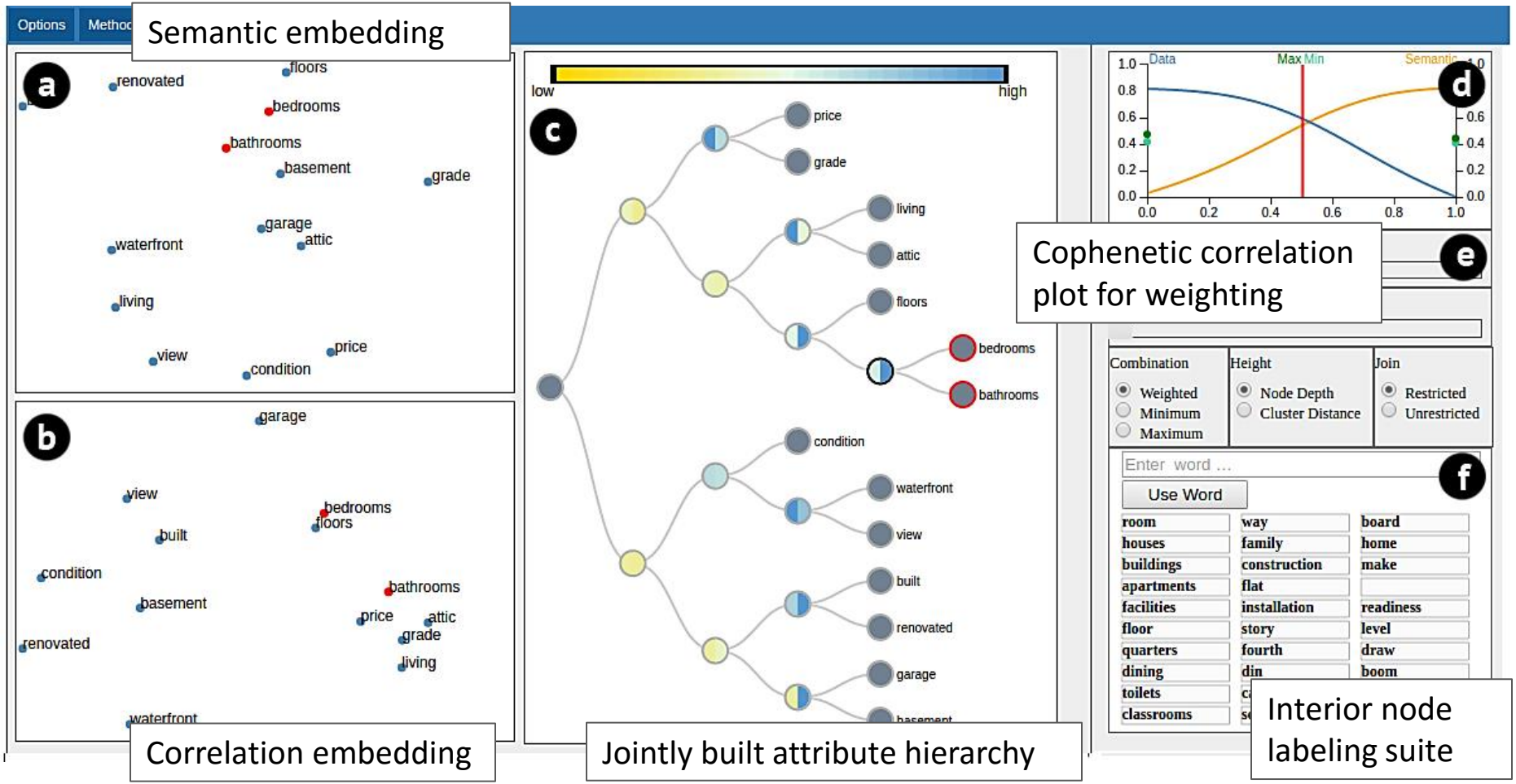
Related Work: Taxonomizer

A semantic tool to manage large attribute spaces

- Semantically explore different aspects of the attribute space

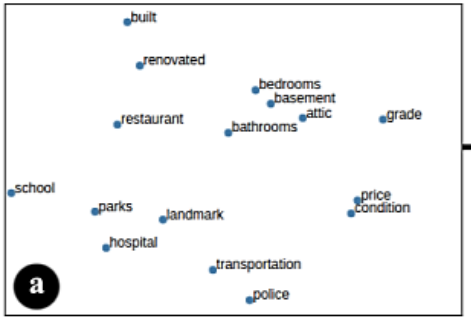


- Could now explore a selected front by a series of scatterplots, PCA, or low-D embedding


Options
Semantic embedding
a
renovated
floors
bedrooms
bathrooms
basement
grade
garage
attic
waterfront
living
view
price
condition
b
garage
view
bedrooms
floors
built
condition
bathrooms
basement
price
attic
grade
renovated
living
waterfront
Correlation embedding
c
low
high
price
grade
living
attic
floors
bedrooms
bathrooms
condition
waterfront
view
built
renovated
garage
Jointly built attribute hierarchy
d
Data
Max Min
Semantic
Cophenetic correlation plot for weighting
e
Combination
Weighted
Minimum
Maximum
Height
Node Depth
Cluster Distance
Join
Restricted
Unrestricted
f
Enter word ...
Use Word
room
way
board
houses
family
home
buildings
construction
make
apartments
flat
facilities
installation
readiness
floor
story
level
quarters
fourth
draw
dining
din
boom
toilets
classrooms
Interior node labeling suite

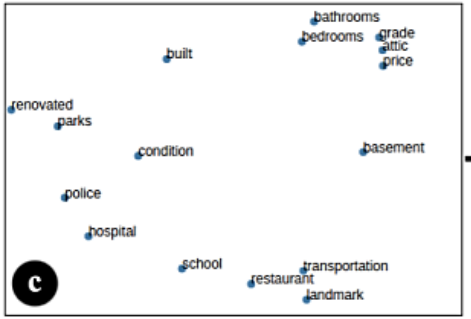
Cosine-distance based NMDS layout

Semantic Space

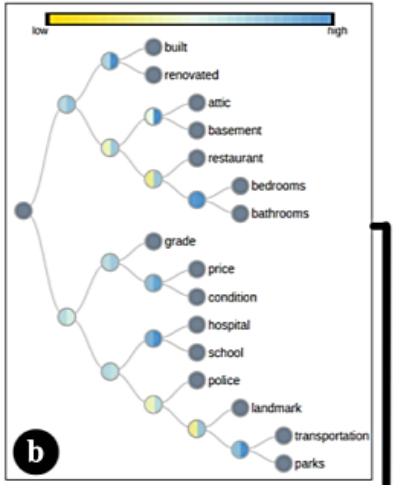


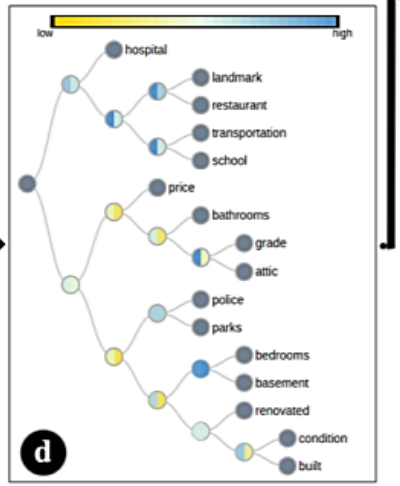
Correlation-distance based NMDS layout

Data Space



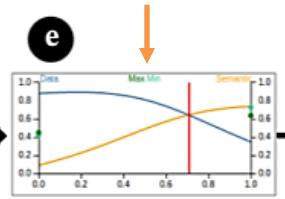
Dendrograms





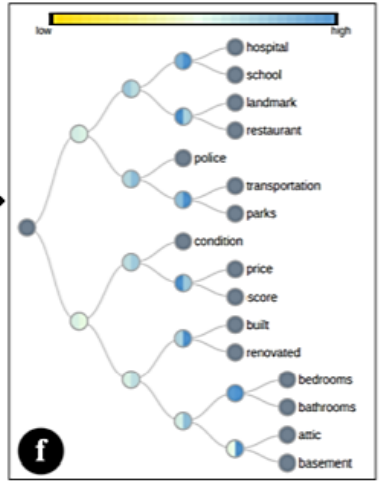
Joint Hierarchy Construction

User sets the semantic/data preference

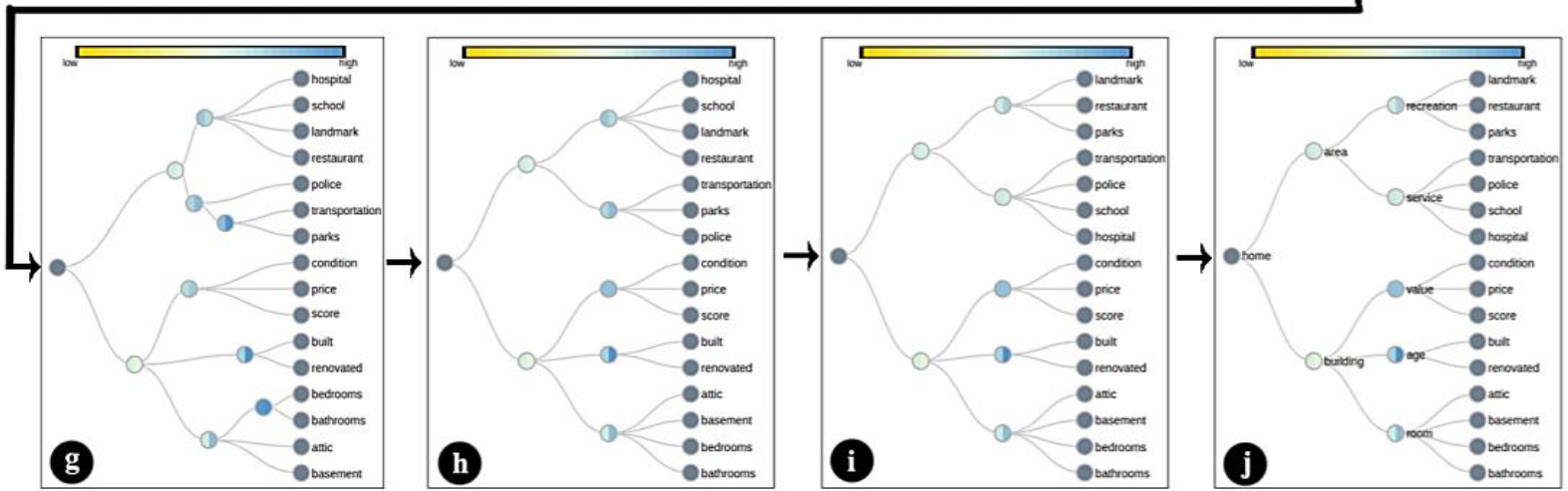


Cophenetic Distance compares the weighted hierarchy with the 2 layouts

Joint dendrogram



User-Driven Refinement and Labeling

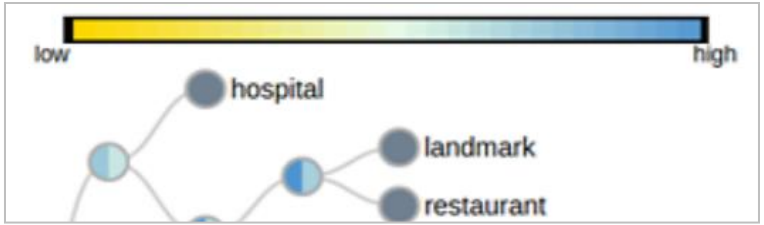


The user can change the hierarchy via drag and drop operations

- The node color encodes the quality loss

The system aids in inner node labeling

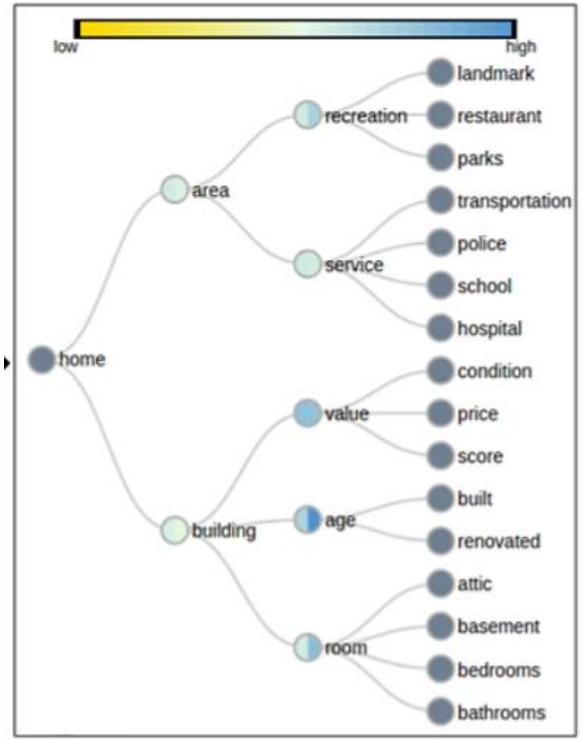
- Parent chosen by contextual inclusion
- Uses the degree of entailment measure



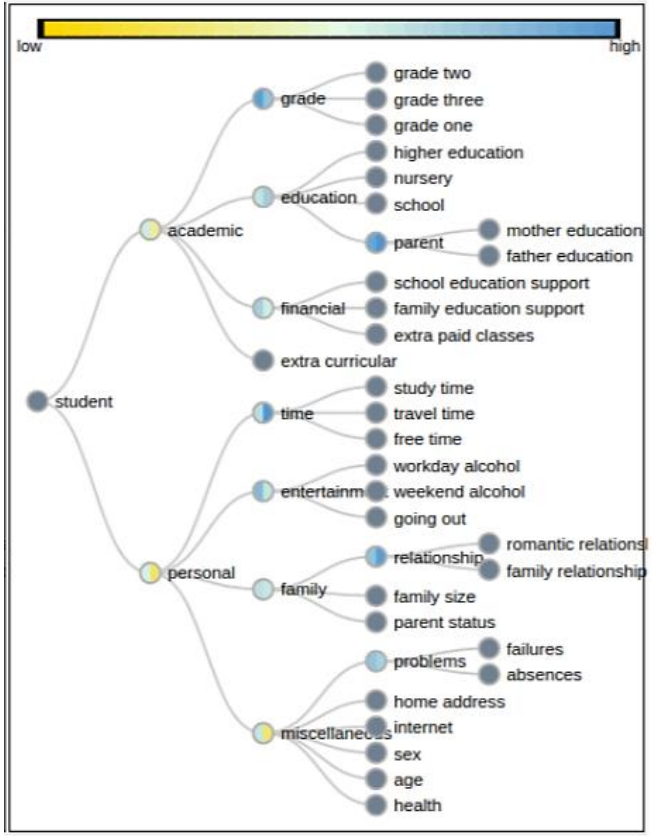
Node color = group quality (Dunn Index)

- Left: quality in the data space
- Right: quality in the semantic space

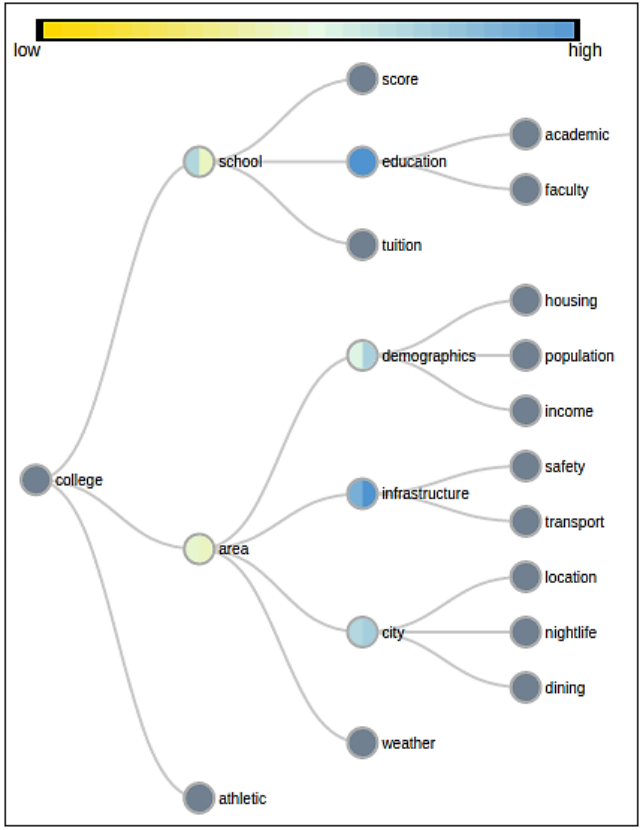
Results



Kings County Housing DS



Student Performance Dataset



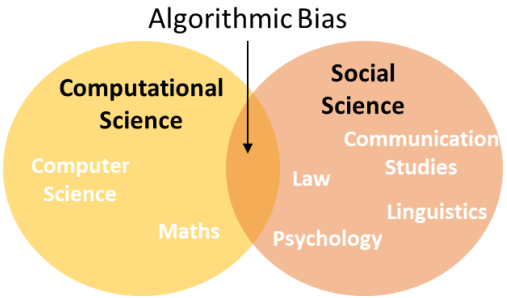
University dataset

More Detail

S. Mahmood, K. Mueller, "Taxonomizer: Interactive Construction of Fully Labeled Hierarchical Groupings from Attributes of Multivariate Data," *IEEE Trans. on Visualization and Computer Graphics*, 26 (9): 2875-2890, 2019.

Next up:
Detecting Intersectional Bias in Word Embeddings

Algorithmic Bias



Algorithmic bias is

- When Algorithms exhibit preference for or prejudice against certain sections of society based on their identity
- Generally emanates from biased training data

Which sub-domains of AI are affected?

- ALL
- Minorities & underrepresented groups are worst hit

Sources of Bias

If the training data or the code developer is biased, the Algorithm will be biased

Books

Teacher

Learn

Human

Training data

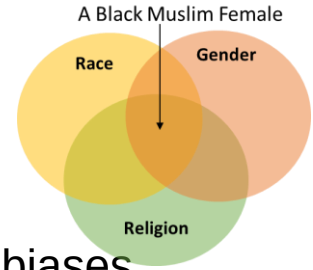
CGPA	GRE_Verbal	TOEFL	International	Admitted
3.5	168	117	No	✓
3.7	165	119	No	✓
3.4	167	118	No	✓
3.8	155	106	Yes	✗
3.9	160	108	Yes	✗
3.7	157	110	Yes	✗

Developers

Learn

Algorithm

Bias in Word Embeddings



- Studies have shown that word embeddings can learn and exhibit social biases based on race, gender, ethnicity, etc. that are encoded in the training dataset.

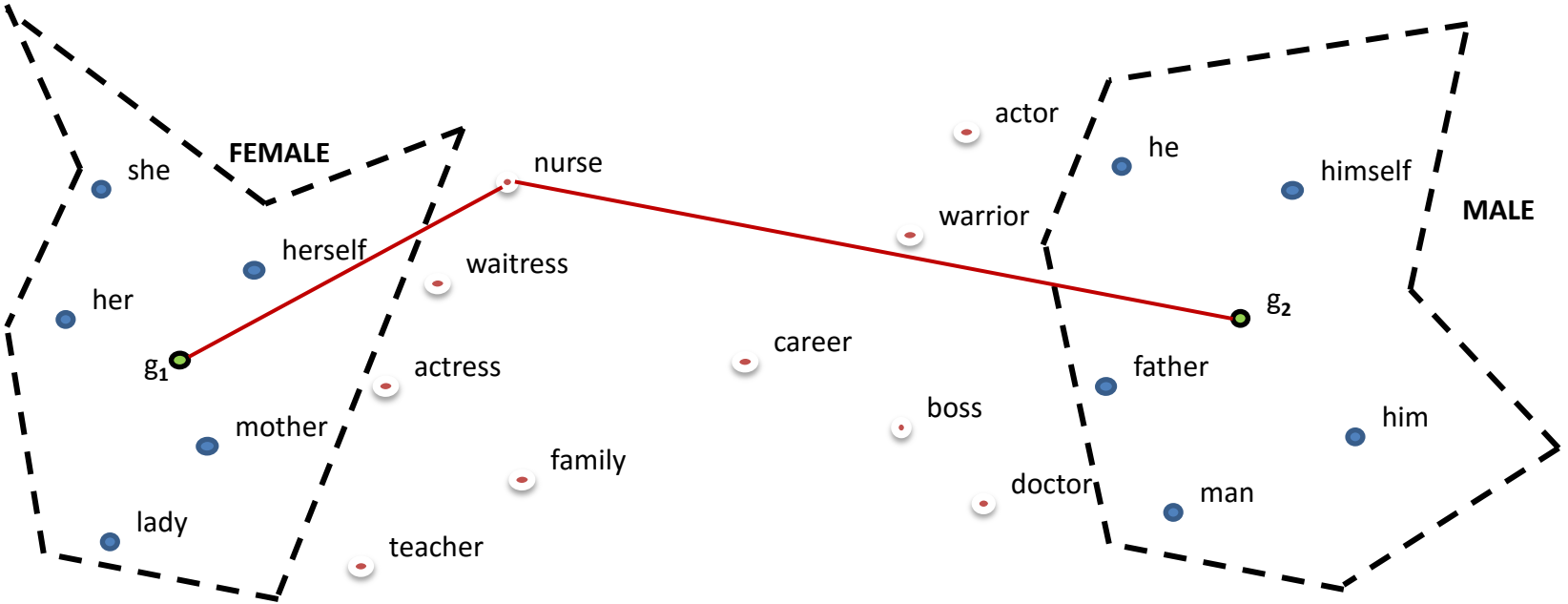
- Biases in word embeddings manifested as stereotypes or undesirable associations between words.
  - For example, Males are disproportionately linked with career and math while females are linked with family and arts.

- Existing literatures primarily focuses on biases based on Gender (93%), Race (54%) and Age/Religion (10%).

- In the real world, an individual might face compound discrimination due to their affiliation with multiple social categorizations based on factors like gender, race, religion, etc.
  - For example, Black Muslim Females.

Bias_score(word) = distance(word, g1)

Bias Quantification

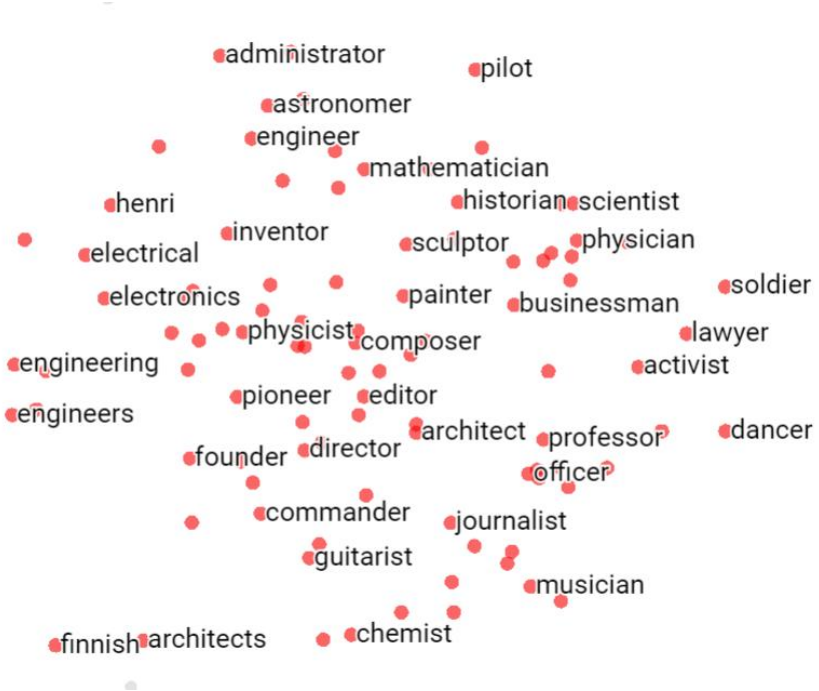


Ideally, neutral words should be equidistant from either cluster

Google’s Embedding Projector

Lacks some critical features for an effective bias discovery tool

- Primarily meant to visualize any embedding space say words, images, etc.
- Limited to only two kinds of bias
- Only a single word can be used to define a group
    - For example, ‘he’ to represent Males
- Doesn’t support any bias quantification algorithm



Design Goals

Word Associations

Given a word say ‘priest’, our tool should help identify the different subgroups the word is associated to, along with the degree of association say Priest -> Males, Christianity, Old, etc.

Bias Exploration

Our tool should support quick and intuitive exploration of words associated with a single subgroup, say *Males,* or an intersectional group, say *Black Muslim Females*.

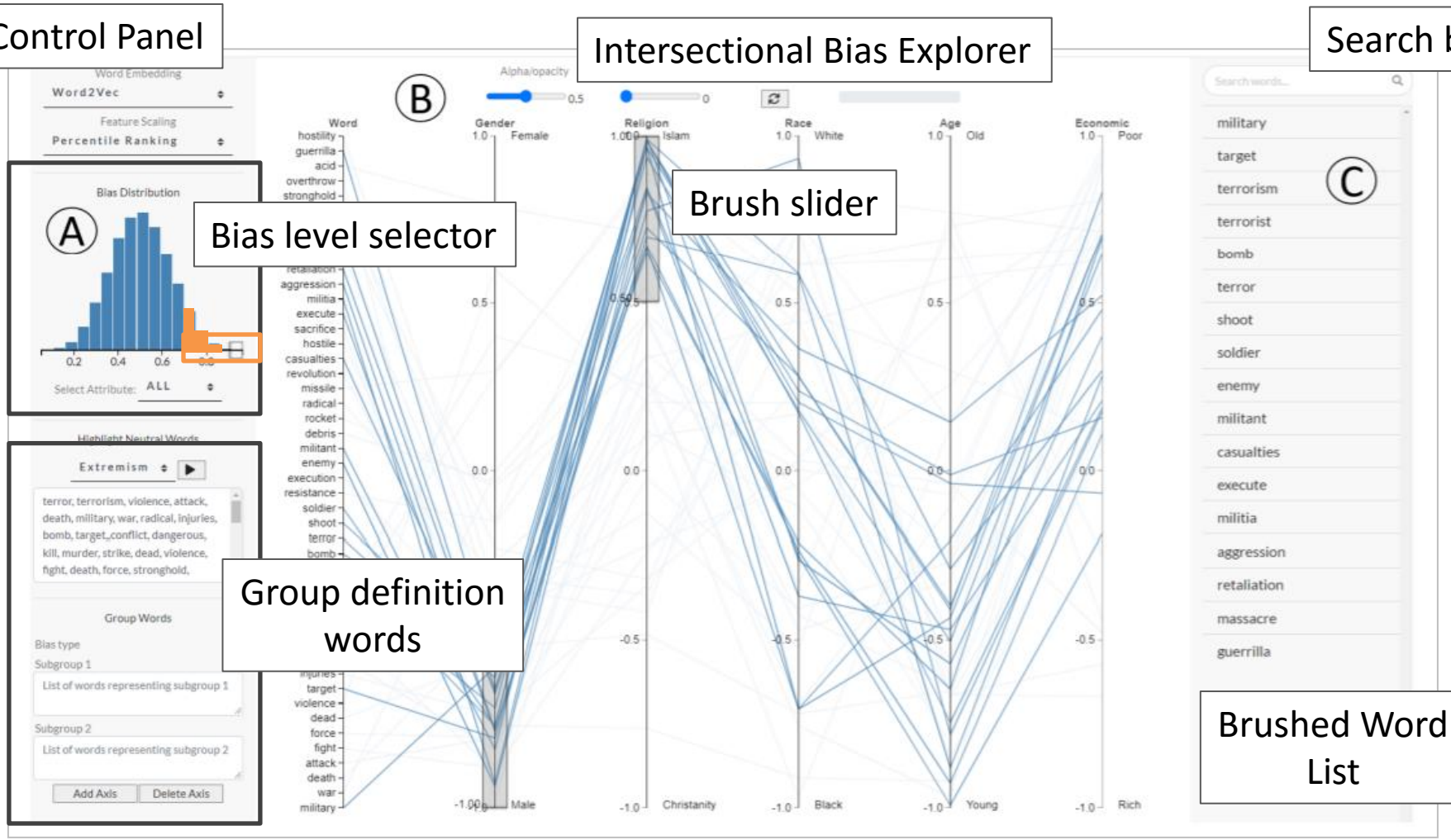
Bias Types

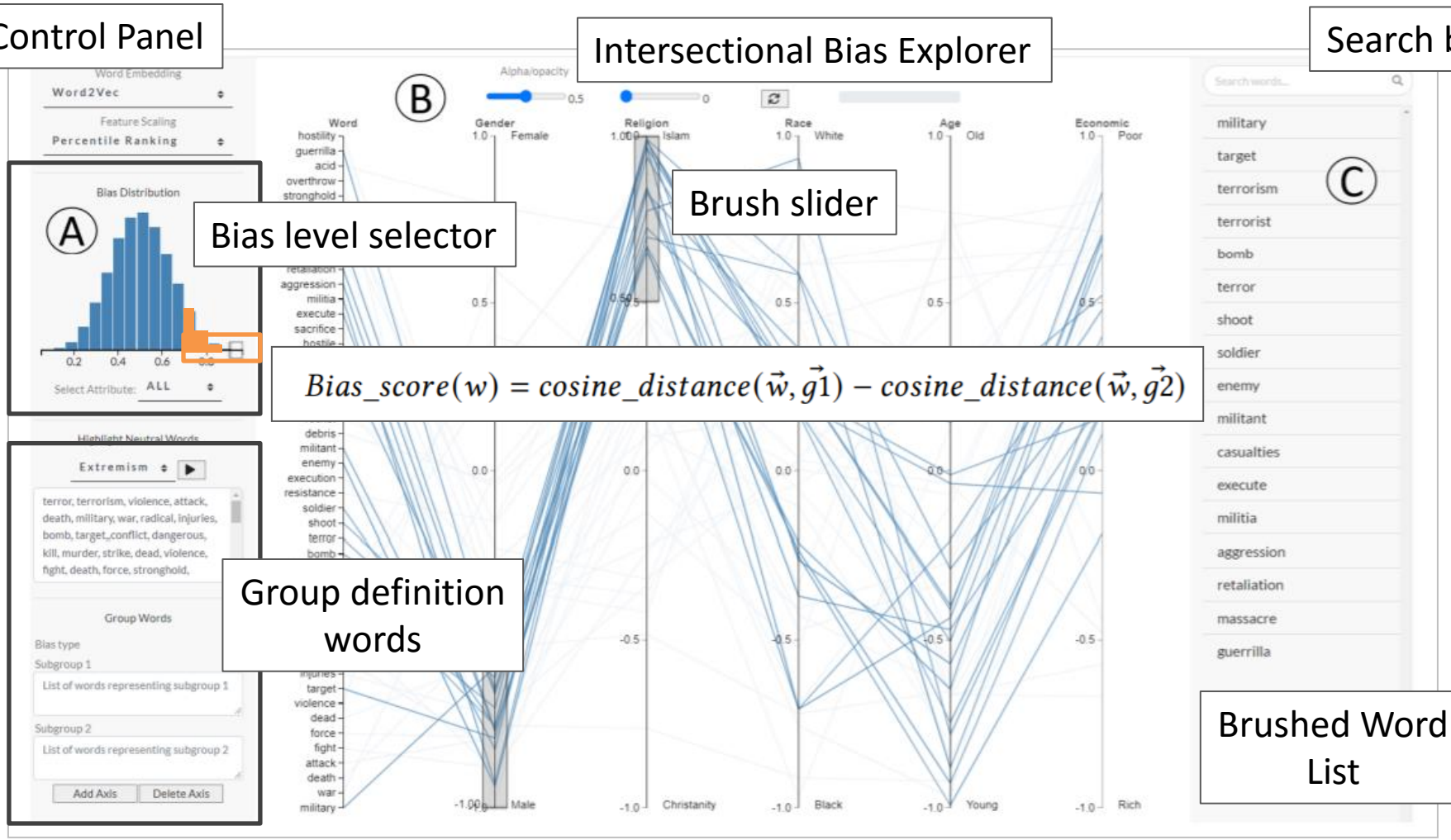
Our tool should support the exploration of well-known biases, but also under-reported biases based on physical appearance, political leanings, etc. or any user-defined bias type.

Data Volume

Our tool should be designed to deal with a large volume of data at both the back and the front end to ensure a smooth user experience.

Our Tool…

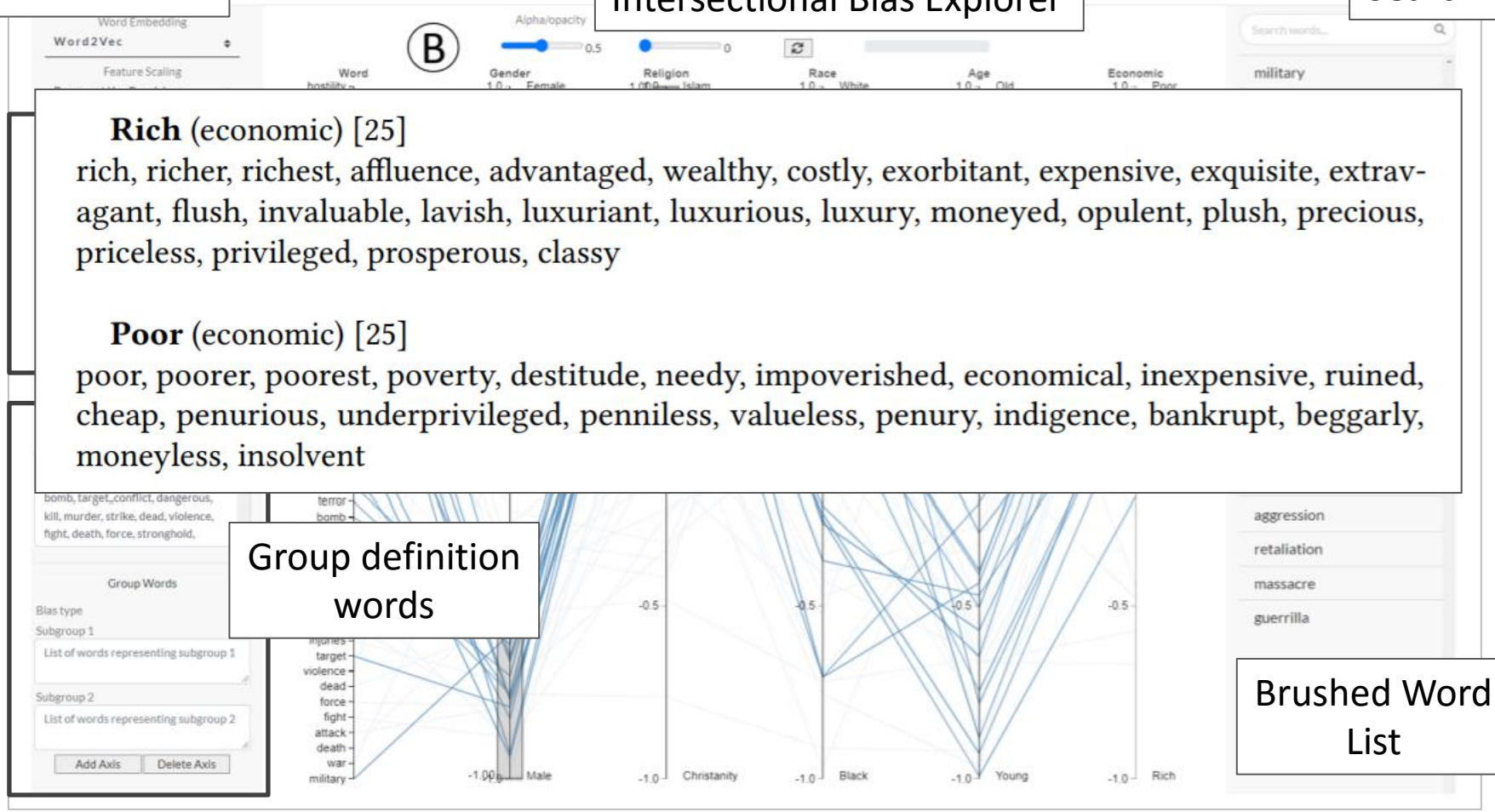

Control Panel
Intersectional Bias Explorer
Search box
Word Embedding
Word2Vec
Feature Scaling
Percentile Ranking
Bias Distribution
A
Bias level selector
Select Attribute:
ALL
Highlight Neutral Words
Extremism
terror, terrorism, violence, attack, death, military, war, radical, injuries, bomb, target,,conflict, dangerous, kill, murder, strike, dead, violence, fight, death, force, stronghold,
Group Words
Bias type
Subgroup 1
List of words representing subgroup 1
Subgroup 2
List of words representing subgroup 2
Add Axis
Delete Axis
Group definition words
B
Alpha/opacity
Brush slider
Word
Gender
Religion
Race
Age
Economic
Female
Male
Islam
Christianity
White
Black
Old
Young
Poor
Rich
hostility
guerrilla
acid
overthrow
stronghold
aggression
militia
execute
sacrifice
hostile
casualties
revolution
missile
radical
rocket
debris
militant
enemy
execution
resistance
soldier
shoot
terror
bomb
target
violence
dead
force
fight
attack
death
war
military
C
military
target
terrorism
terrorist
bomb
terror
shoot
soldier
enemy
militant
casualties
execute
militia
aggression
retaliation
massacre
guerrilla
Brushed Word List


Control Panel
Intersectional Bias Explorer
Search box
Word Embedding
Word2Vec
Feature Scaling
Percentile Ranking
Bias Distribution
A
Bias level selector
Select Attribute: ALL
Highlight Neutral Words
Extremism
terror, terrorism, violence, attack, death, military, war, radical, injuries, bomb, target,,conflict, dangerous, kill, murder, strike, dead, violence, fight, death, force, stronghold,
Group Words
Bias type
Subgroup 1
List of words representing subgroup 1
Subgroup 2
List of words representing subgroup 2
Add Axis
Delete Axis
Group definition words
B
Alpha/opacity
Brush slider
Word
Gender
Religion
Race
Age
Economic
Female
Male
Islam
Christianity
White
Black
Old
Young
Poor
Rich
$Bias\_score(w) = cosine\_distance(\vec{w}, \vec{g1}) - cosine\_distance(\vec{w}, \vec{g2})$
C
military
target
terrorism
terrorist
bomb
terror
shoot
soldier
enemy
militant
casualties
execute
militia
aggression
retaliation
massacre
guerrilla
Brushed Word List

Control Panel

Intersectional Bias Explorer

Search box



Rich (economic) [25]
rich, richer, richest, affluence, advantaged, wealthy, costly, exorbitant, expensive, exquisite, extravagant, flush, invaluable, lavish, luxuriant, luxurious, luxury, moneyed, opulent, plush, precious, priceless, privileged, prosperous, classy

Poor (economic) [25]
poor, poorer, poorest, poverty, destitude, needy, impoverished, economical, inexpensive, ruined, cheap, penurious, underprivileged, penniless, valueless, penury, indigence, bankrupt, beggarly, moneyless, insolvent

Group definition words

Brushed Word List

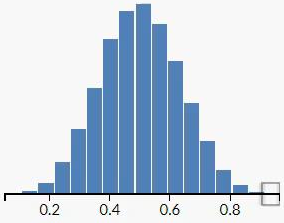
Word Embedding

Word2Vec

Feature Scaling

Percentile Ranking

Bias Distribution



Select Attribute: ALL

Highlight Neutral Words

Profession

teacher, author, mechanic, broker, baker, surveyor, laborer,,surgeon, gardener, painter, dentist, janitor, athlete, manager, conductor, carpenter, housekeeper, secretary,

Group Words

Bias type

Subgroup 1

List of words representing subgroup 1

Subgroup 2

List of words representing subgroup 2

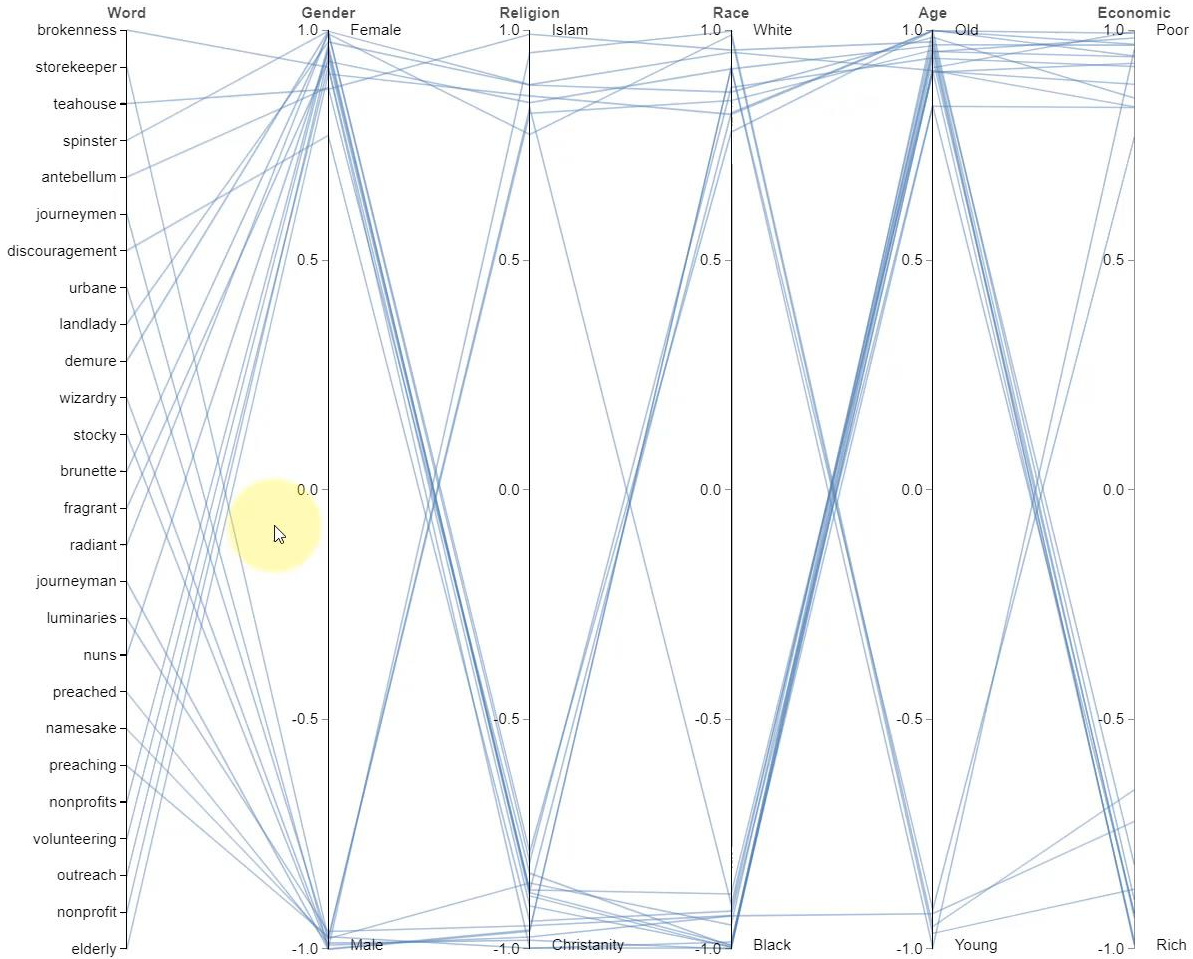
Add Axis | Delete Axis

Alpha/opacity 0.5

Curve Smoothness 0

Reset Brush

Selected Words: 26/50K



Search words...

Our Findings

Intersectional Groups	Associated Words
Poor - Young - Black	disaster, struggle, tackle, chaos, woes, hunger, uprising, desperation, insecurity, rampage, roadblocks
Rich - Old - White	formal, attractive, appealing, desirable, castle, desserts, seaside, golfing, cordial, bungalow, fanciful, warmly, salty
Black - Muslim - Male	gun, assassination, bullets, bribes, thugs, looted, dictators, electrocuted, cowards, agitating, storekeeper, looter, bleeping
Young - Christian - Male	career, dominant, brilliant, lone, terrific, heroes, superb, epic, monster, prowess, heavyweights, excelled, superstars
Poor - Female	ostracism, brokenness, mortgages, eviction, brothels, witchcraft, traumatized, discrimination, sterilizations
White - Christian - Female	romantic, nuns, virgin, republicans, peachy, platonic, convent, radiant, unspoiled, unpersuasive, soppy, drippy, soapy

Potential Use Cases

Auditing Tool

- Researchers and data scientists could audit a word embedding model for different kinds of biases before deploying it for any downstream application.

Educational Tool

- Use in classrooms to teach about bias in AI.

Promoting Accessibility

- Algorithmic bias is an interdisciplinary problem with relevance in law, linguists, sociology, digital humanities, etc.

More Detail

B. Ghai, Md. N. Hoque, K. Mueller, “WordBias: An Interactive Visual Tool for Discovering Intersectional Biases Encoded in Word Embeddings” *ACM CHI (Late Breaking Work)*, May 8-13, 2021.

Some Ongoing Work in That Area

Extend support for more languages

- Chinese, Spanish, French, etc.

Visualize & debias for multiple sensitive groups simultaneously

- within an interactive visual interface

Augmented writing for groups with special needs

- e.g. people who stutter (our FLUENT tool, ACM SIGACCESS 21)

General algorithmic bias identification & mitigation in general

- paper to come, watch this space

Thanks

For more information. visit:
https://www.cs.stonybrook.edu/~mueller/

My co-authors/PhD students: Bhavya Ghai, Md Naimul Hoque, Salman Mahmood

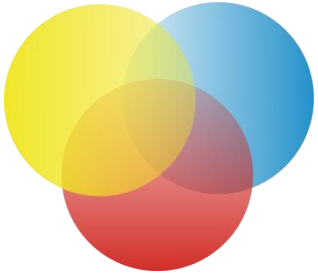
Funding agencies: NSF, NIH, DOD, DOE

Transparency

Interactive visual interfaces boost transparency

Accountability

Human in-charge can be held accountable



Multidisciplinary

Human experts infuse domain knowledge into the system

Trust

Human in the loop brings more trust into the system