Multiview Clustering via Canonical Correlation Analysis

Karen Livescu, Sham Kakade, Karthik Sridharan

Toyota Technological Institute at Chicago

Kamalika Chaudhuri

University of California at San Diego

Motivation

Clustering is a stand-alone task or pre-processing step in

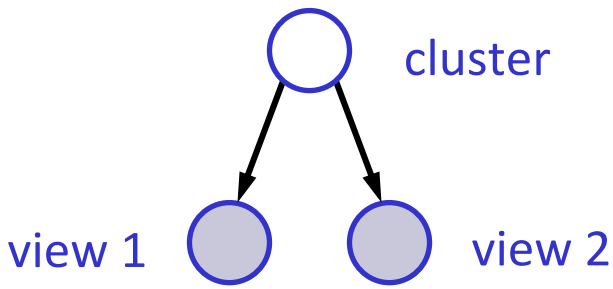
- Data analysis
- Vector quantization (for compression, speech recognition, ...)
- Density estimation
- Information retrieval
- Semi-supervised learning for classification/regression

Clustering in high dimensions is “hard”, so data may be projected to lower-dimensional space via PCA or random projection

- Does not differentiate between noise and signal dimensions
- PCA behavior depends on the coordinate system
- Theoretical guarantees depend on stringent separation requirements

Motivation (2)

- Can take advantage of multiple views to find meaningful dimensions (e.g. audio + video, text and link structure, images + captions, ...)
- Given a data set of paired vectors $\{(x^{(1)}_1, x^{(2)}_1), \ldots, (x^{(1)}_n, x^{(2)}_n)\}$, can think of each $(x^{(1)}_i, x^{(2)}_i)$ as a sample from the same cluster (class), plus (high-dimensional) additive noise
- We think of the two views as *independent given the hidden class*
- If noise is independent in the two views (e.g. audio noise vs. video lighting), then the correlated dimensions are related to the hidden class



Algorithm

CCA finds directions $w^{(1)}_i$ and $w^{(2)}_i$ that maximize the correlations between the projections of $X^{(1)} = \{x^{(1)}_i\}$ and $X^{(2)} = \{x^{(2)}_i\}$

- The first pair of directions is $\arg \max_{w^{(1)}_1, w^{(2)}_1} \text{corr}\left(w^{(1)}_1 x^{(1)}, w^{(2)}_1 x^{(2)}\right)$
- Subsequent direction vectors maximize the correlation subject to orthogonality with the previous vectors

Algorithm Given paired vectors $\{(x^{(1)}_1, x^{(2)}_1), \ldots, (x^{(1)}_n, x^{(2)}_n)\}$, $X^{(1)} = \{x^{(1)}_i\}$, $X^{(2)} = \{x^{(2)}_i\}$:

1. Find the top k CCA directions $W^{(1)} = \{w^{(1)}_{i:k}\}$, $W^{(2)} = \{w^{(2)}_{i:k}\}$
2. Project samples: $X^{(1)}_p = W^{(1)} X^{(1)}$, $X^{(2)}_p = W^{(2)} X^{(2)}$
3. Cluster $X^{(1)}_p$ or $X^{(2)}_p$, e.g. via k-means

Theoretical results [Chaudhuri and Kakade 2008]

Assumptions:

1. *Uncorrelated views* conditioned on the class

2. *Nondegeneracy*: $\mathrm{cov}(X^{(1)}, X^{(2)})$ has rank equal to the number of clusters k

Theoretical results [Chaudhuri and Kakade 2008]

- **Theorem** Suppose the source distribution is a mixture of k Gaussians and Assumptions 1 and 2 hold. If in *at least one view* $v \in \{1,2\}$,

$$||\mu^{(l)}_v - \mu^{(l)}_v|| > C\, \sigma^* k^{1/4} (\log (kn/\delta))^{1/2}$$

where σ^* is the maximum standard deviation *in the subspace containing the means*, then with probability $1-\delta$ the algorithm correctly classifies all examples in view v given a data set of size $O^*(d/\lambda_{min}^2)$

- **Theorem** If the source distribution is a mixture of log-concave distributions, then the separation requirement is

$$||\mu^{(l)}_v - \mu^{(l)}_v|| > C\, \sigma^* k^{1/2} \log (kn/\delta))$$

- In the case of PCA, the separation requirement depends on σ^*_d, the maximum deviation along *any* dimension

Experiment 1: Audio-visual clustering of speakers

Data: VidTIMIT

- 41 speakers, speaking 10 utterances each
- Audio + face video recorded in studio environment with no significant lighting/pose variation
- 25 image frames per second

Audio features: Spectra computed over 40ms frames (1501 dimensions)

Image features: Pixels of face region (2394 dimensions)

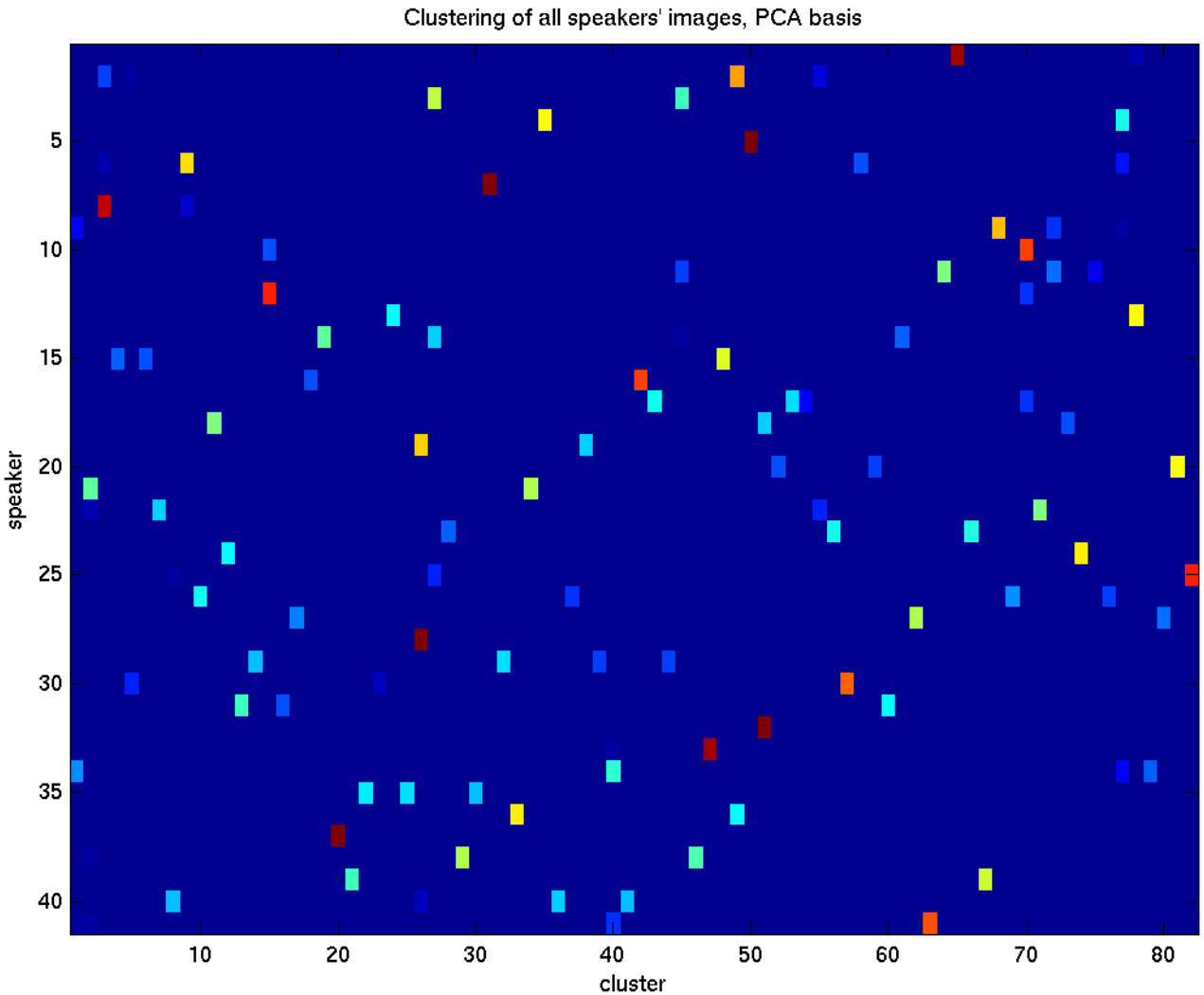
CCA/PCA to 45 dimensions

K-means clustering into 82 clusters (2 per speaker)

Note:

- Many clusterings possible (speakers, phonemes, clothing, ...)
- Here we consider the “target” clusters to be speakers
- Images expected to cluster well
- Audio not expected to cluster well

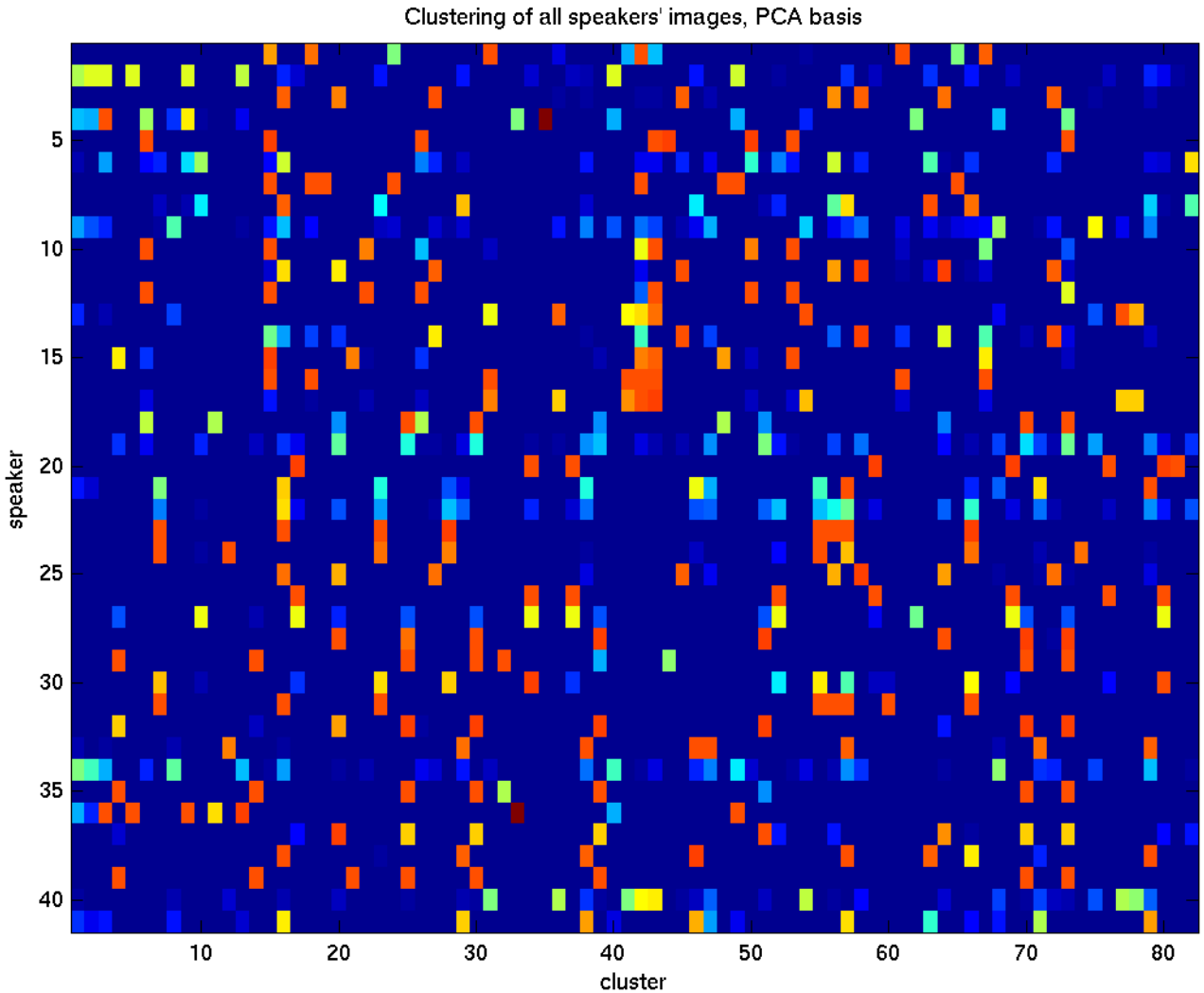
Experiment 1 selected results: Image clustering



	PCA	CCA
H(sp|clus)	0.27	0.35
perplexity $2^{H(sp\|clus)}$	1.21	1.27

Exp't 1 selected results: Image clustering with occlusions

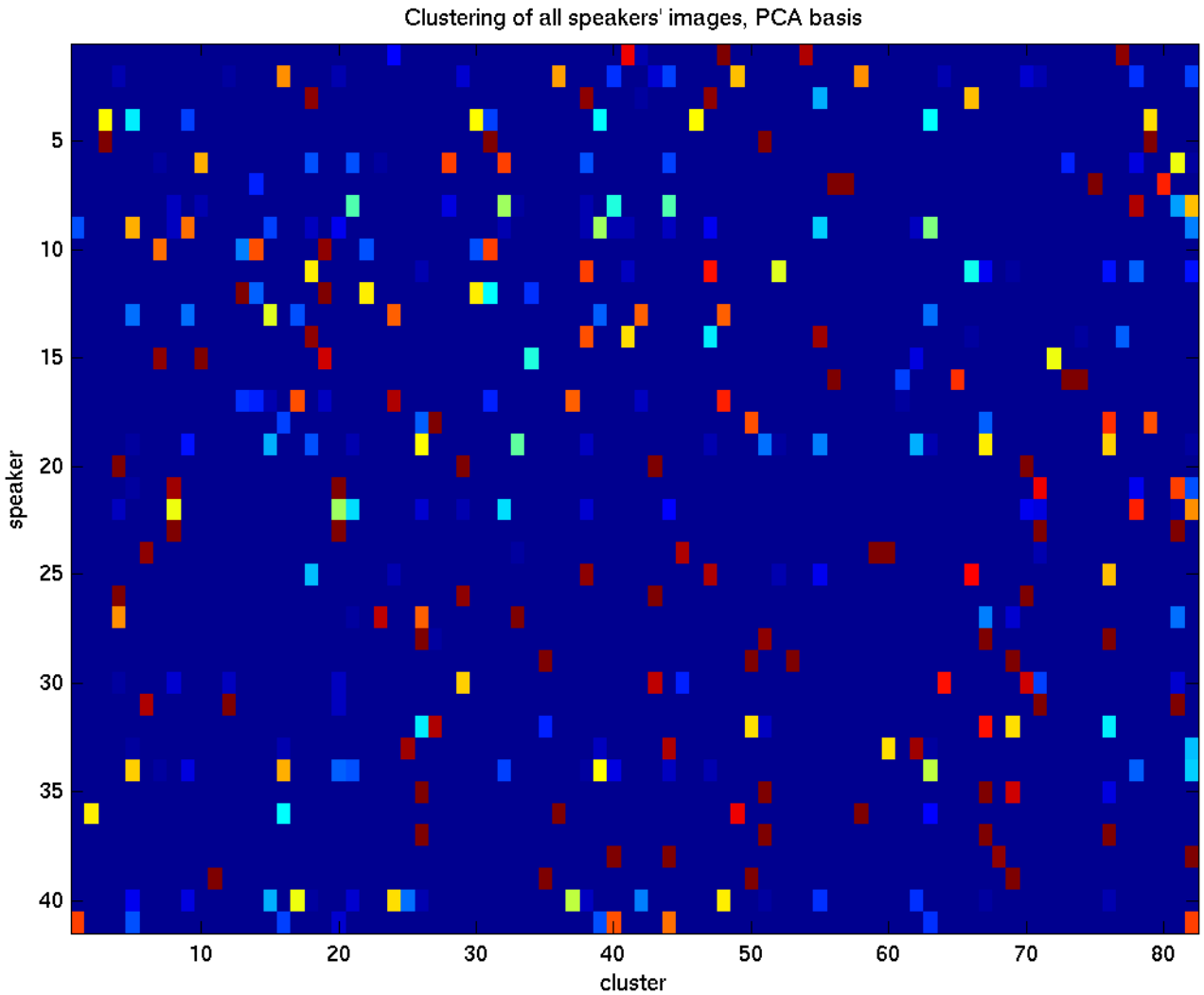
PCA results are greatly degraded; CCA results are not



	PCA	CCA
H(sp|clus)	2.72	0.33
perplexity $2^{H(sp\|clus)}$	6.59	1.26

Exp't 1 selected results: Image clustering with translations

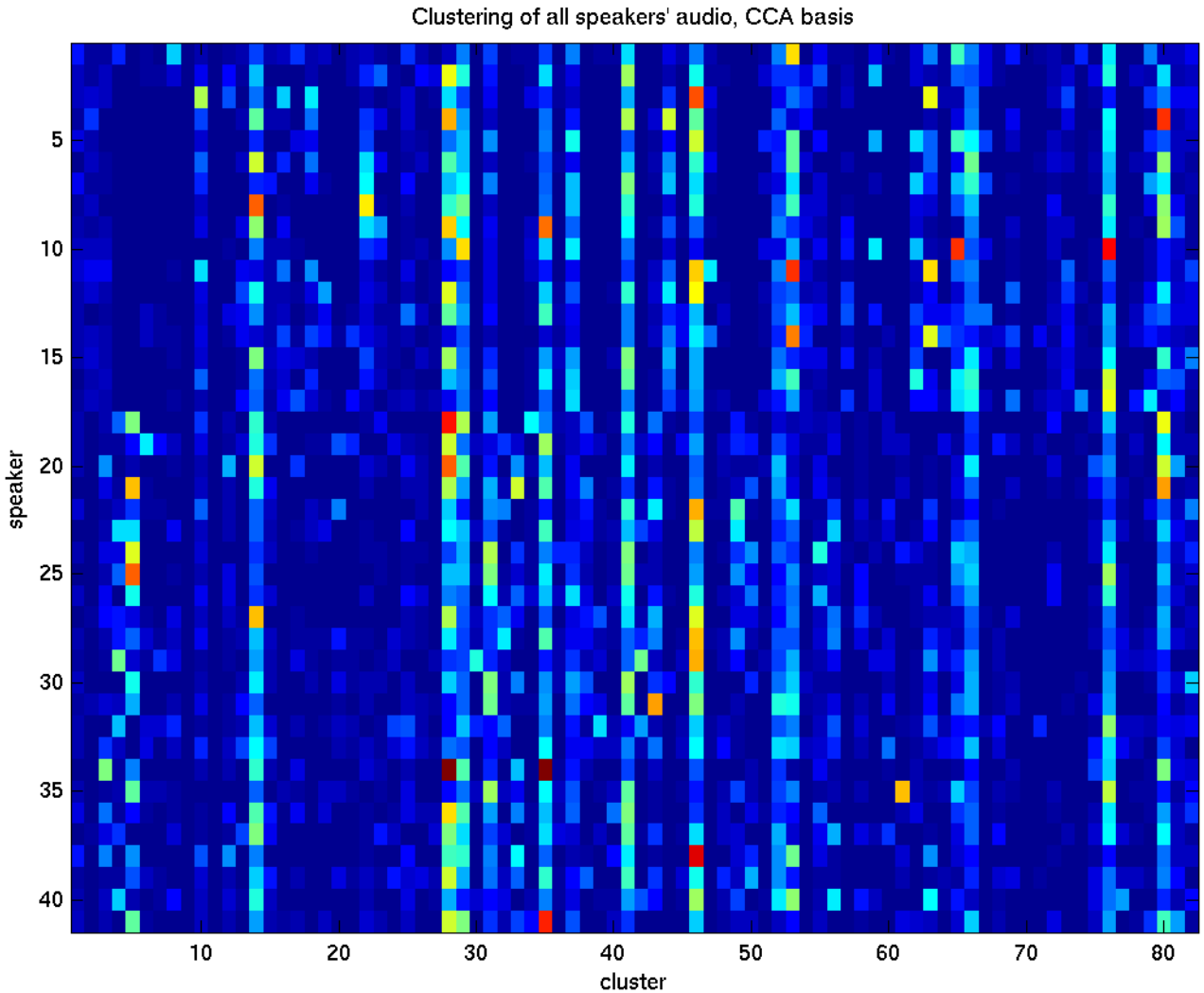
PCA results are greatly degraded; CCA results are not



	PCA	CCA
H(sp|clus)	1.77	0.83
perplexity $2^{H(sp\|clus)}$	3.41	1.77

Exp't 1 selected results: Audio clustering

Consistent .2-.3 bit reduction in conditional entropy from PCA to CCA across experiments; currently investigating additional audio features



	PCA	CCA
H(sp|clus)	4.98	4.76
perplexity $2^{H(sp\|clus)}$	31.6	27.1

Experiment 2: Clustering Wikipedia articles

Data: 128,000 Wikipedia pages

Text features: Count of each word in the article (~8 million dimensions)

Link features: Number of links from/to each article (~12 million dimensions)

Initial dimensionality reduction to 1000 dimensions using random projections

Projection via CCA/PCA to 20 dimensions

Used a hierarchical clustering procedure:

1. Perform CCA/PCA and find N clusters using k-means
2. For each cluster in (1) larger than a threshold, perform CCA/PCA on data in this cluster, and cluster into N smaller clusters

Experiment 2: Example CCA-space clusters

Creationism; War; Afrocentrism; Wahhabism; Harvey Milk; Saunders Lewis; Jean-Jacques Rousseau; William Joice; Opera; Idi Amin; John Donne; History of the telescope; Asceticism; Afterlife; Information; Satire; Notary public; Romantic nationalism; Thomas Aquinas; List of Estonians; List of fantasy authors; List of Germans; Vatican City; Torah; Old Catholic Church; Apostles' Creed; Mahmoud Abbas; Boris Yeltsin; Ulysses S. Grant; William Jennings Bryan; Pipe organ; George IV of the UK; George V of the UK; Emperor Taizong of Tang; David I of Scotland	Hanoi Hilton; USS Hornet; Battle of Peleliu; House (astrology); Hua Mulan; 1972 in sports; Moselle; Duke of York; Spam (Monty Python); History of Vanuatu; Liliopsida; 1908 Summer Olympics; Lucy Liu; Celia Cruz; National Cartoonists Society; Balearic Islands; SCSI; Asexuality; Copyright infringement of software; Lamb of God; Calendar date; Lapland War; Totalitarianism; History of Sweden; European Youth Parliament; Roman mythology; Kingdom of Israel; List of railway companies; Maundy money	Florence Nightingale; Texaco; Poltergeist; Isadora Duncan; John Allan Muhammad; 1935 in music; 1934 in music; 1939 in music; 1921 in music; Star Trek; 1776 (musical); The Godfather Part II; Dial M for Murder; Jimmy Durante; 2006 Commonwealth Games; 1968 in music; 1977 in music; Comet Hale-Bopp; High Speech Rail; Steve Wozniak; James Carville; David Copperfield; Lynyrd Skynyrd; Poison (band); LL Cool J; RZA; Stand; Bob Dole; Sergei Prokofiev; Peter Hain; Ravana; Michael Porillo; George Pickett	Roswell, SD; LaFayette, KY; Old Ripley, IL; Hainesville, IL; Belleville, WI; Bethany, IL; South Point, Ohio; Ashwaubenon, WI; North Miami, FL; Davenport, IA; Spartanburg, SC; Hurricane Lili; Scranton, PA; Myopia; Influenza; Motion picture rating system; FEMA; Vineland, NJ; Fair Lawn, NJ; Medford, NJ; Palisades Park, NJ; South River, NJ; Gillespie County, TX; Hidalgo County, NM; Finney County, KS; Onslow County; NC; Brady Township, PA; Argentine Township, MI; Augusta Gharter Township, MI	Natural number; Linear subspace; Cauchy distribution; Wrench; List of matrices; Hydrogenation; Alpha helix; Heat pump; Soil; Bicarbonate; Campfire; Thermal depolymerization; Cardiac arrest; Neuron; Transformer; Hepatocellular carcinoma; Yellow fever; Isopropyl alcohol; Gorilla; Make (software); Ostinato; Control flow; PHP; HTML; Objective-C; Vienna Development Method; Comparison of Java C++; Dot product; Elliptic integral; Catalan number; Expected value; Coenzyme; Neutron; History of geodesy; Orbital period

Experiment 2: Example PCA-space clusters

Hugh the Great; Eddy Duchin; Spoilt Bastard; Irish presidential election, 1990; Hadley, NY; Medicine Park, OK; Council Grove, KS; Perryville, KY; Albert of Sweden; Ferdinand I of Romania; Luigi Boccherini; Daniel Bernoulli; Market research; Gram; Metastability; Corticosteroid; Misdemeanor; 19 (number); Kent Brockman; Bourne shell; So I Married an Axe Murderer; Peyo; Lipschitz continuity; Equations of Motion; Conjugate transpose; Square number; Datamax UV-1; Tribes of Galway; Brethren; Iwi	Meeme, WI; West Point, WI; Wesley, ME; Charlotte, VT; Glen Ridge, FL; Geraldine, MA; Taylorsville, MI; Ovid, Colorado; Conway, ND; Grano, ND; Onamia, MN; Geneva, IA; Margaret Avison; Bodomi; Uncle Scrooge Adventure; Demographics of Saint Helena; William Prout; Union Hill, IL; Waldo, OH; East Nassau, NY; Rochester, OH; Mill Hall, PA; Attu Station, AL; Roxbury, NH; Lone Rock, WI; Baca County, CO; Comanche County, OK; Polk County, GA	Lisp (programming language); Species; Alkane; Hindu calendar; Sexual harassment; Symmetry; Book of Job; Marginalism; Creativity; Rumi; Assam; Classical liberalism; NATO; Republic of Macedonia; Action 14; Novel; Lombards; Wars of the Roses; History of Bolivia; Thabo Mbeki; W. H. Auden; Apollo; Gustav Mahler; James Joyce; New Mexico; Islamic art; Nevada; Metro Manila; Mount Athos; Avignon; Bangalore; Kolkata; Jabba the Hutt; Andrea Dworkin;	Cary Grant; Hormel; Wilson Pickett; Microsoft Windows; Louis Kahn; Sugar Ray Leonard; UNIVAC; Nat King Cole; Supergrass; Whoopi Goldberg; Baby boomer; Magi; Received Pronunciation; Common Pheasant; Samuel Mudd; Charlie Brown; Super 8 mm film; Loki; Equus (play); Second Battle of Fort Fisher; Romanos I; IBM Personal System/2; Action figure; Albertus Magnus; Roland Freisler; Giacomo Puccini; Frederick William II of Prussia; Wiki; Borzoi; Dynamic Host Configuration Protocol; Yinglish	List of abbreviations in the CIA world fact book; List of rural districts of Germany; Bilge; Orthoptera; Bill Fitch; Muirne; Ministry; 1770 in literature; John Bell Williams; Brooke Burke; Yeardley Smith; Anastasios II (emperor); Mary River National Park; List of diseases (S); Lake Township, MN; List of Acer species; Yucca; List of oboists; List of male film actors; Sparidae; 58 BC; Xanthine; 19; Pyrrhic; 100 gigametres; 177 BC; Businessperson; GNU compiler for JAVA; Strawberry Field; NJ Route 63; Cucurbitales; Rita Johnston

Conclusions

- In the right conditions, CCA outperforms PCA as a dimensionality reduction technique before k-means clustering

- Ongoing work: Application to semi-supervised learning for speaker and speech recognition, analysis of hierarchical clustering