Modeling text data

word representations

You shall know a word by the company it keeps

-JR Firth (1957)

government debt problems turning into banking crises as has happened in

saying that Europe needs unified banking regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

word2vec

"the cat sat on floor"

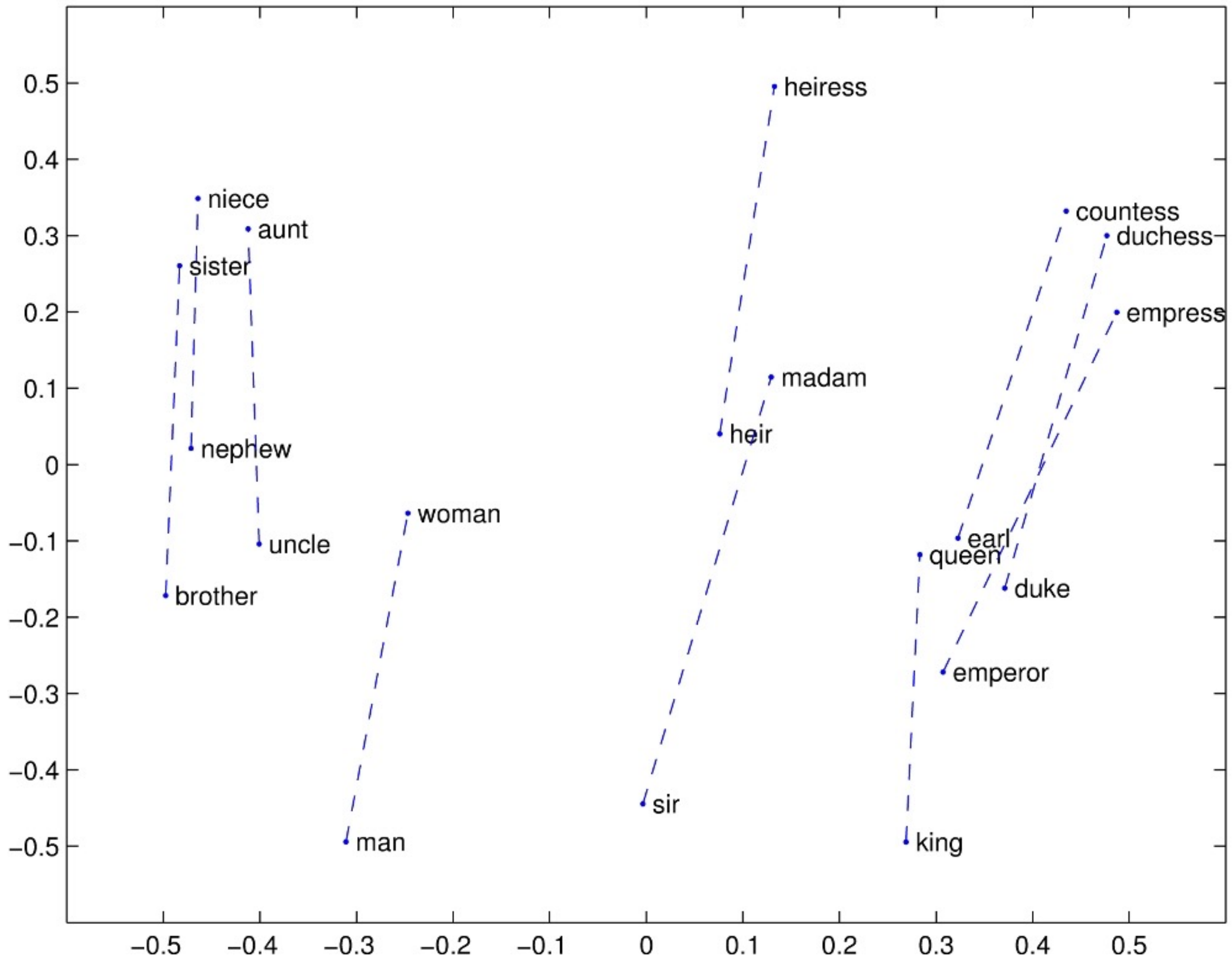
positive sample

cat flew

negative sample

$$P(positive) = \sigma(v_c \cdot v_w)$$

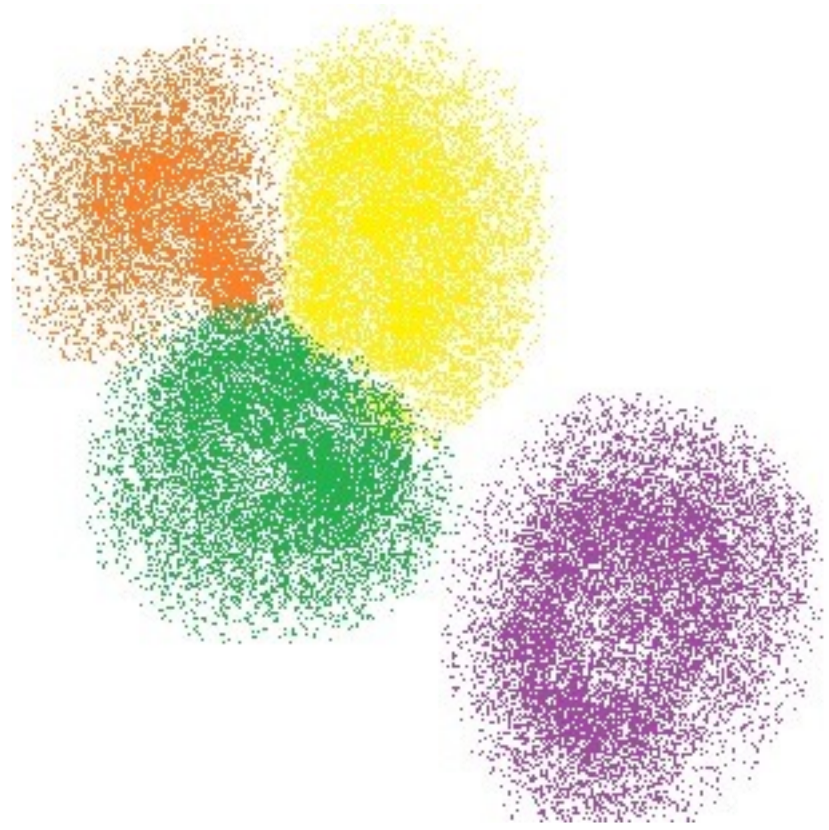
analogies



Berlin-Germany+France=Paris

clustering

- want to group data, may not necessarily have labels

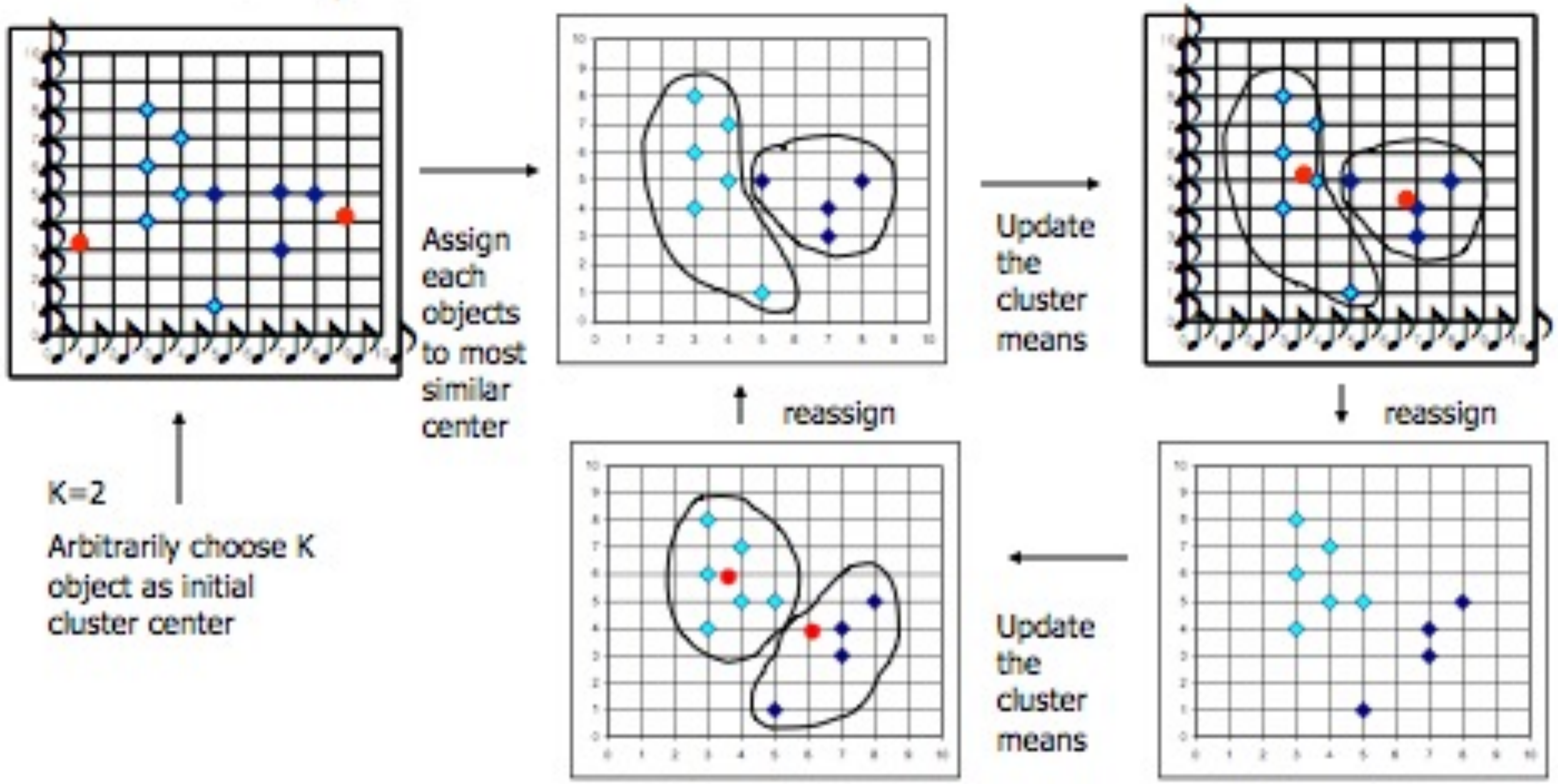


k-means clustering

- start with initial guess for number of clusters K and group membership of each data point
- iteratively come up with better and better clusters
- at each iteration compute the center of each cluster using the mean of each group
- re-assign membership so that each point joins the group of the closest center

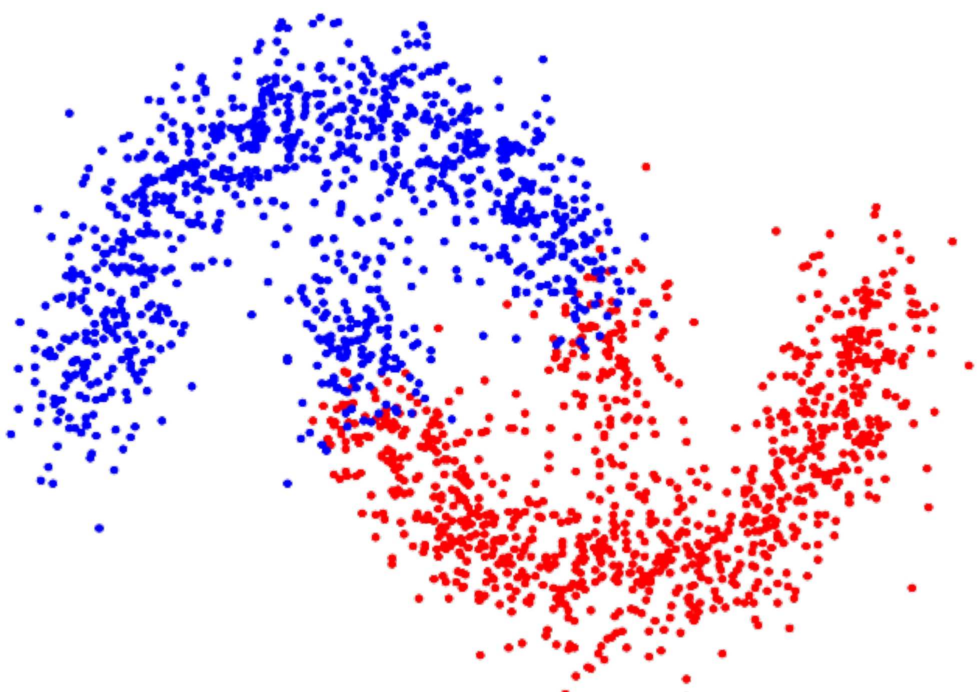
The *K-Means* Clustering Method

- Example

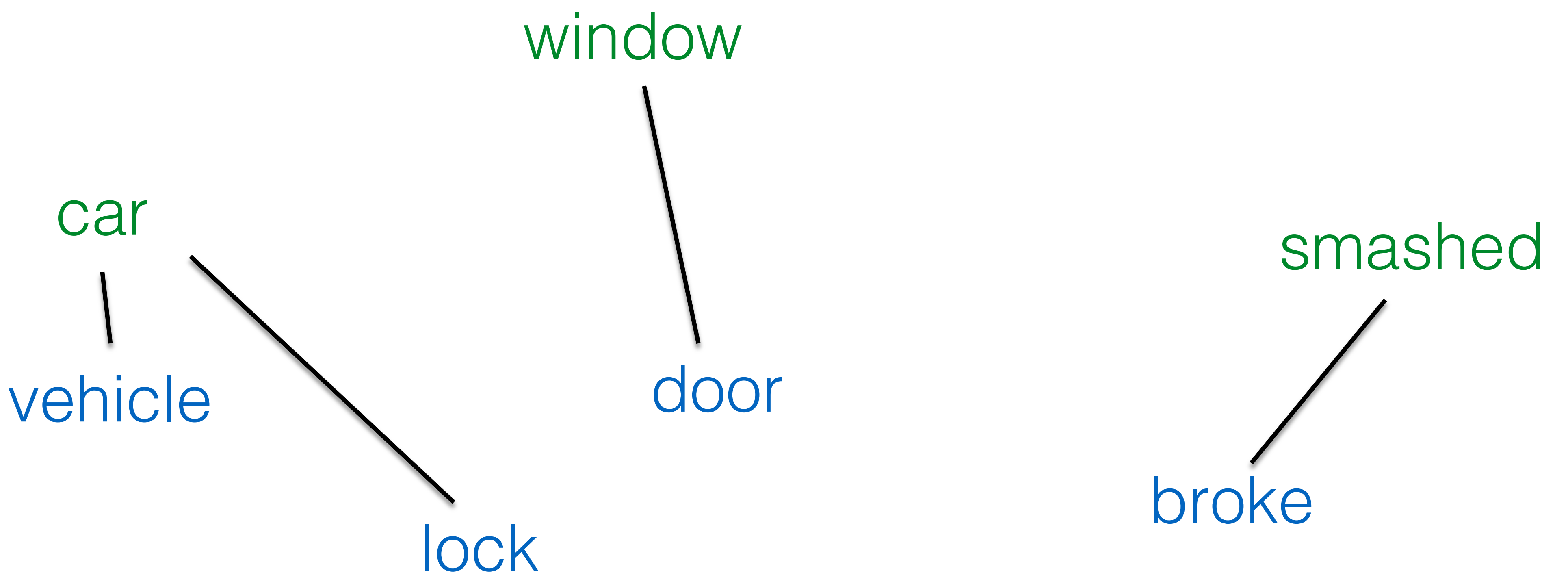


strengths/weaknesses

- simple to compute, method is easy to explain
- need to specify K, which may be unknown
- only finds convex clusters



Word movers distance



Word movers distance+clustering

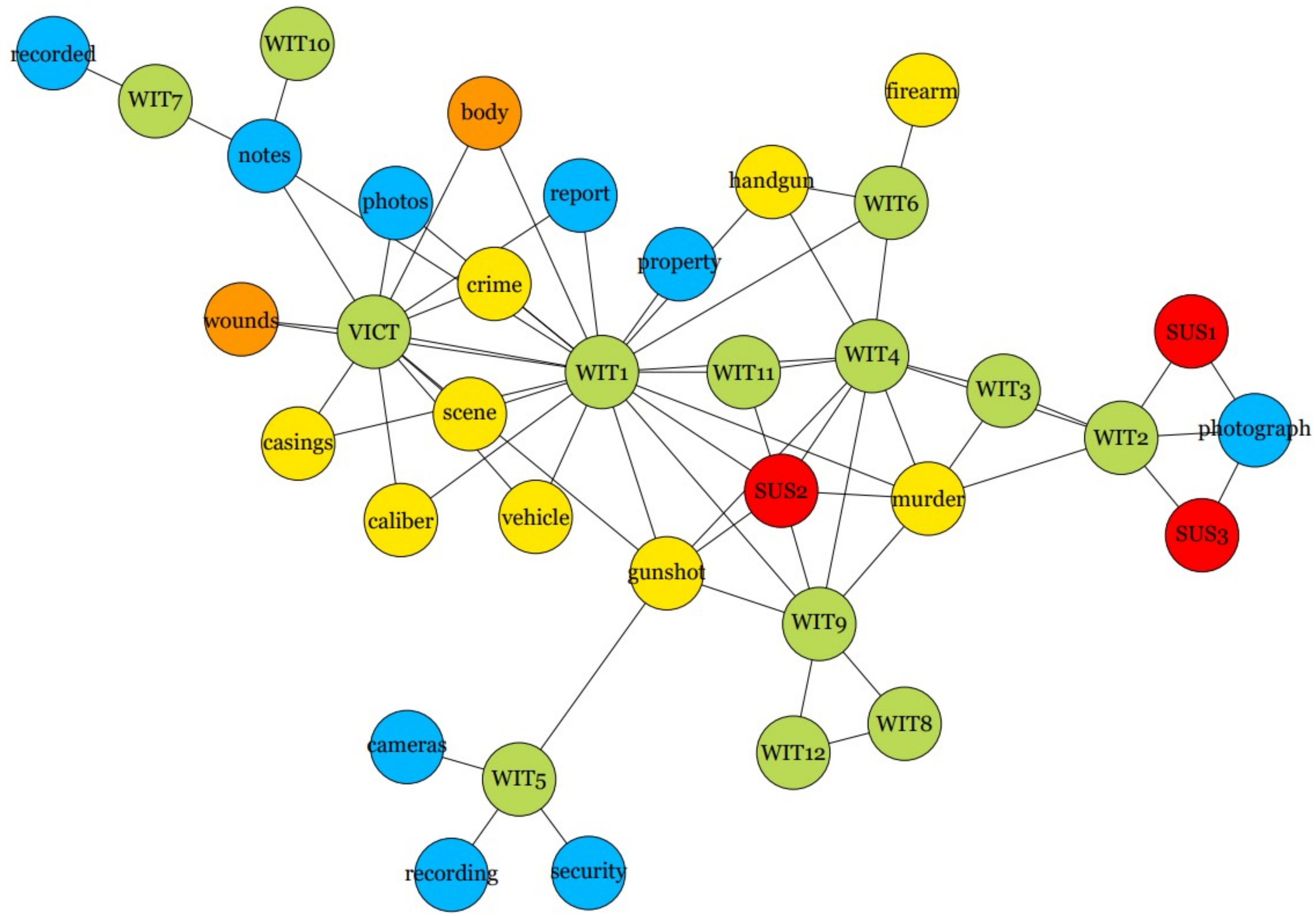
POST

I came home to opened, empty packages on my porch. I'm at [redacted]. Anyone else get hit today? Somehow it's even nastier that they left the empty boxes instead of just stealing them

TOPIC

door	said
front	location
house	package
came	someone
get	told

Text mining homicide investigation chronologies



Pandey et al, ICDMW, 2020

Keyword expansion

- Example: find all songs about driving a vehicle
- Start with a list of keywords (car, truck, drive, driving)
- Then find songs with those keywords and look at other words in song that have similar word vectors, expand list
- Repeat several times

Named entity recognition

Text mining homicide investigation chronologies

TABLE II: Initial List for Identifying Types of Evidence in Text

Evidence Type	Keywords
Documentary Evidence	tapes, recording, surveillance, photo, video, camera, photograph
Physical Evidence	weapon, gun, knife, gunshot, caliber, casing
Forensic Evidence	dna, blood, fingerprint, autopsy

TABLE III: Evidence List after applying Keyword Expansion

Evidence Type	Keywords
Documentary Evidence	tapes, recording, surveillance, photo, video, camera, photograph, print, letter, security, camera, printout, record, recording, report, notes, document, monitor, footage, warrant, property, picture, chronology, log
Physical Evidence	weapon, gun, knife, gunshot, caliber, casing, handgun, firearm, item, shooting, bullet, murder, crime, scene, crimescene, shot, kill, stab, revolver, fire, discovery, criminal, kick, vehicle, veh
Forensic Evidence	wound, body, polygraph, exam, examination, test, hair, impression

Pandey et al, ICDMW, 2020

Text mining homicide investigation chronologies

DET1 and DET2 arrived at crime scene, located at ADDRESS . Victim in streetcovered with sheet . Victim identified at scene by his Sister WIT as VICT , GENDER/ETHNICITY AGE . Victim had multiple gunshot wounds to his chest , back and possibly to BODYPART . I/O 's conducted crime scene investigation See IR Report and Notes .Recoveredevidence, two .45 caliber casings . Coroner 's Investigator DET3 took charge of thevictim's body and assigned Coroner 's Case No . XXXX . DET1 took possession oftwo cell phones in victim 's pockets and searched victim 's MODEL MAKE , parked on westcurb on ADDRESS . Provided victim 's vehicle keys to WIT . SID PhotographerNAME XXXX took photos that were directed by DET1 , C # XXXX .

Pandey et al, ICDMW, 2020

Text mining homicide investigation chronologies

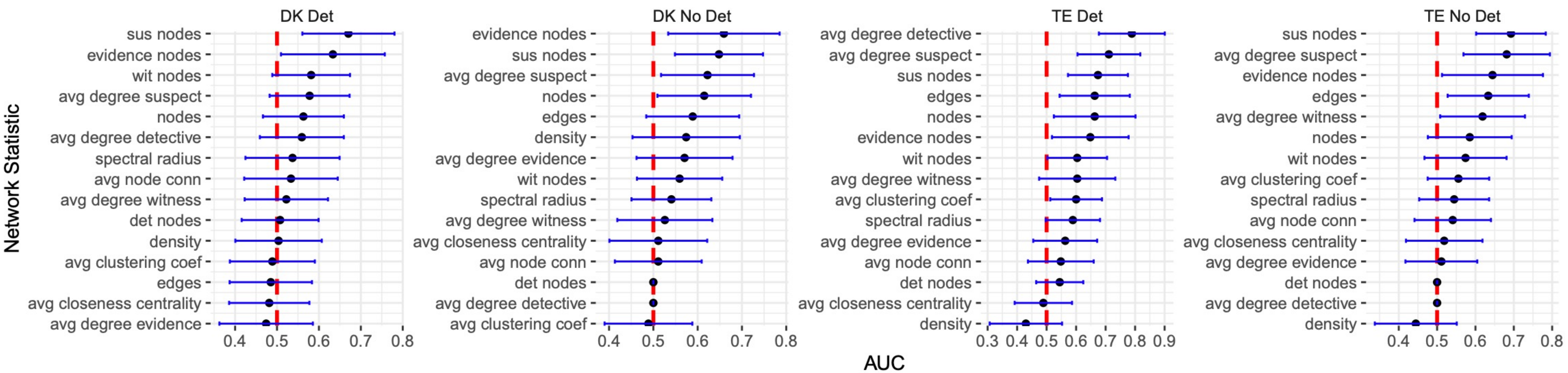
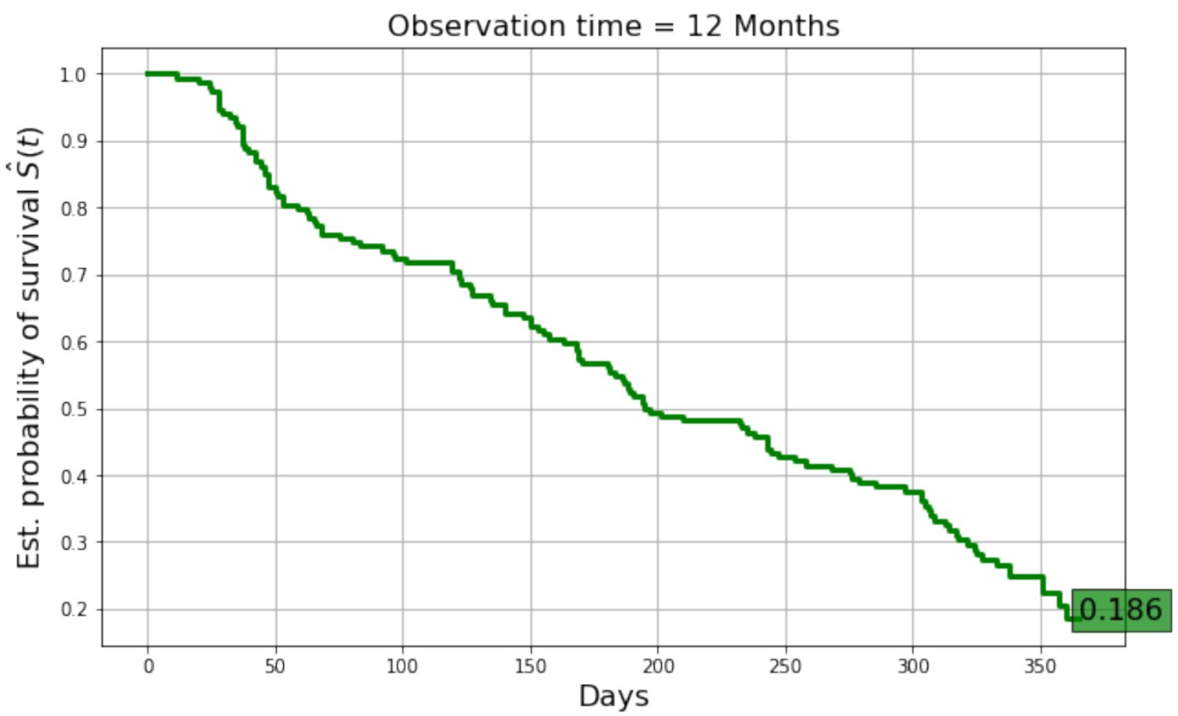


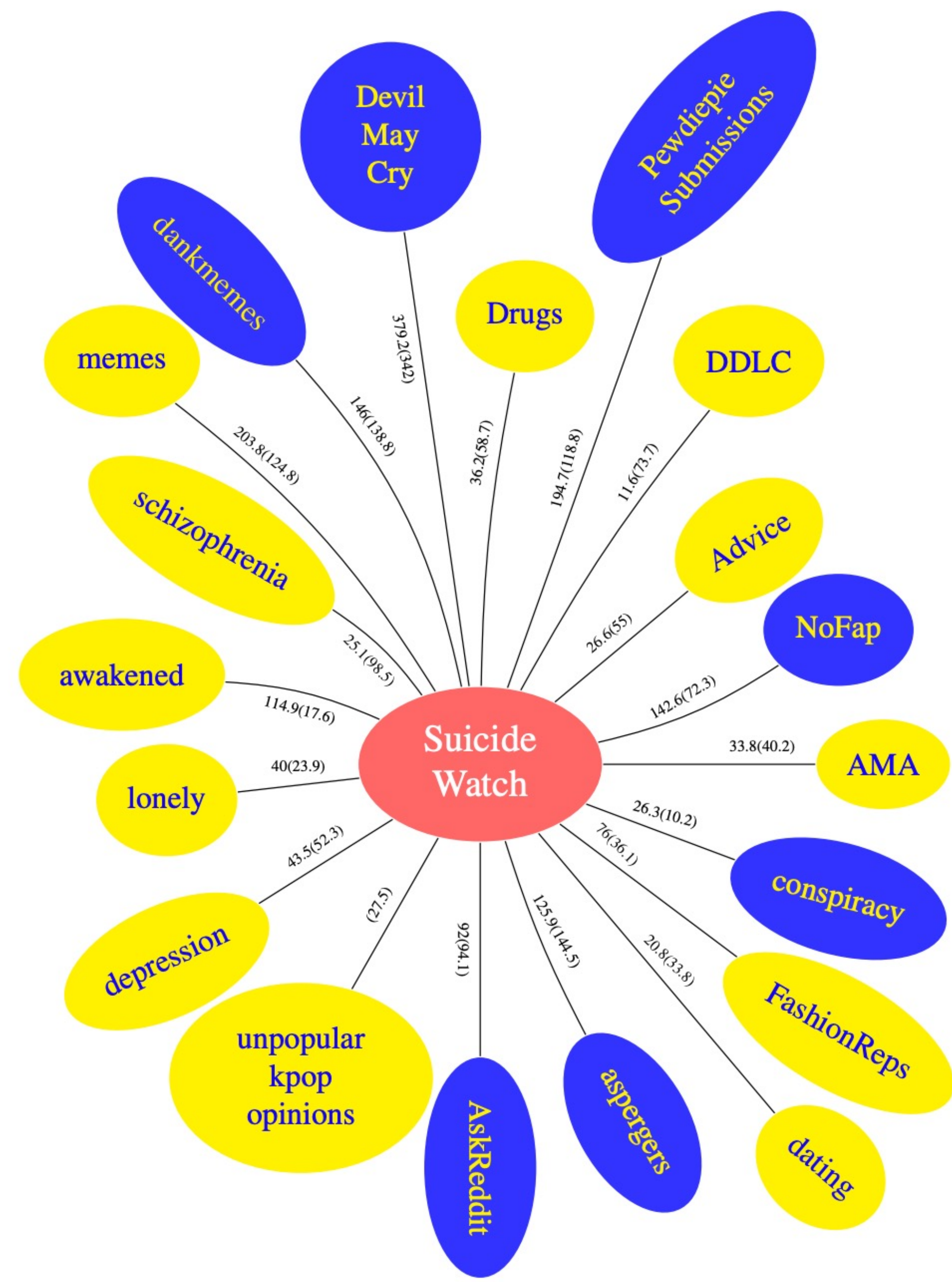
Fig. 7: AUC and standard error for each network statistic in week 1 of the investigation.

Pandey et al, ICDMW, 2020

Modeling subreddit transitions on Reddit



Kaplan-Meier curve for transition from casual drug use sub-reddits to r/RedditorsInRecovery



Modeling subreddit transitions on Reddit

Cox Hazard Function

$$\lambda_i(t) = \lambda_0(t)\exp\{\boldsymbol{x}^\top \boldsymbol{\beta}^{(i)}\}$$

Table 3: Cox Model Results Summary. **Train/test split of 1,775 (1665 censored) and 592 (352 censored) users, respectively. C-Index shown for models using different feature sets. The model using drug utterances, keywords, and LIWC features performed best on training set using 5-fold cross validation and gave a test-set C-Index of 0.820. Test set data consisted of 45 observed and 592 censored examples.**

Model	C-Index
Doc2Vec	0.790
Doc2Vec + drugs + keywords + LIWC	0.788
Drugs + keywords + LIWC	**0.820**
Test Set Performance	***0.820***

Lu, Sridhar, Pandey, Hasan, Mohler, KDD 2019

Modeling subreddit transitions on Reddit

Table 4: Top 10 Explanatory Covariates

Drug Name	C-Index
Heroin	0.748
Buprenorphine	0.702
LSD	0.687
Psilocybin	0.628
Oxycodone	0.623
Marijuana	0.621
Ecstasy	0.614
Fentanyl	0.610
Oxymorphone	0.608
Amphetamine	0.597

LIWC feature	C-Index
Leisure	0.668
Period	0.646
Time	0.646
Ingest	0.645
Informal	0.642
Netspeak	0.633
Focuspresent	0.630
Relativ	0.627
Nonflu	0.612
Money	0.610

Table 5: One-Year Survival Probability by Top Drug Mention

Drug Name	**Surv. Prob.**	**Drug Name**	**Surv. Prob.**
Ecstasy	0.987	fentanyl	0.820
LSD	0.981	cocaine	0.774
benzodiazepines	0.877	oxycodone	0.767
marijuana	0.872	Heroin	0.502
methamphetamine	0.824	Buprenorphine	0.498

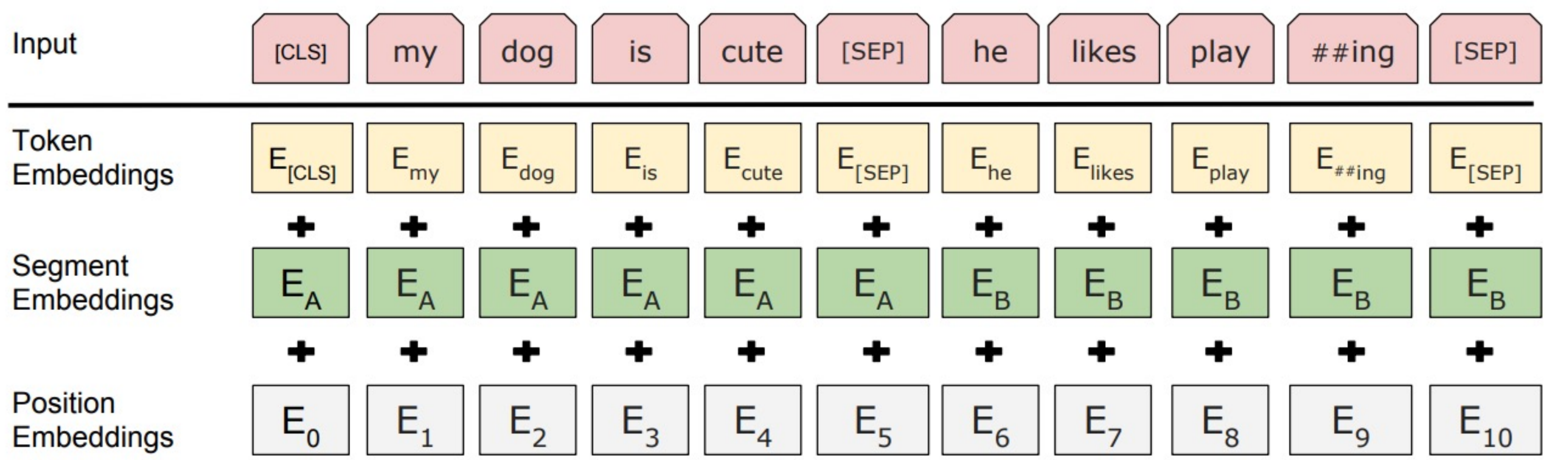
Lu, Sridhar, Pandey, Hasan, Mohler, KDD 2019

Text analysis libraries

- NLTK (tagging, named entity recognition, diagramming)
- GENSIM (topic modeling, word representations)
- SPACY (deep learning based tagging NER, word representations)

Bidirectional encoder representations from transformers (BERT)

- Masked token pre-training: mask 15% of input words and then predict them using output embedding
- Masked sentence pre-training: input pairs of sentences and then predict if they are next sentence pairs or not
- Architecture:
  - L=12 transformer blocks (layers)
  - H=768 embedding size
  - A =12 self attention heads
  - 110M parameters (340M larger version)



Question/answering with a fine-tuned BERT (C. Khanna)

Text: New York (CNN) -- More than 80 Michael Jackson collectibles -- including the late pop star's famous rhinestone-studded glove from a 1983 performance -- were auctioned off Saturday, reaping a total $2 million. Profits from the auction at the Hard Rock Cafe in New York's Times Square crushed pre-sale expectations of only $120,000 in sales. The highly prized memorabilia, which included items spanning the many stages of Jackson's career, came from more than 30 fans, associates and family members, who contacted Julien's Auctions to sell their gifts and mementos of the singer. Jackson's flashy glove was the big-ticket item of the night, fetching $420,000 from a buyer in Hong Kong, China. Jackson wore the glove at a 1983 performance during \"Motown 25,\" an NBC special where he debuted his revolutionary moonwalk. Fellow Motown star Walter \"Clyde\" Orange of the Commodores, who also performed in the special 26 years ago, said he asked for Jackson's autograph at the time, but Jackson gave him the glove instead. "The legacy that [Jackson] left behind is bigger than life for me,\" Orange said. \"I hope that through that glove people can see what he was trying to say in his music and what he said in his music.\" Orange said he plans to give a portion of the proceeds to charity. Hoffman Ma, who bought the glove on behalf of Ponte 16 Resort in Macau, paid a 25 percent buyer's premium, which was tacked onto all final sales over $50,000. Winners of items less than $50,000 paid a 20 percent premium

Q: Where was the Auction held?

A: Hard rock cafe in new york ' s times square