Mixed Membership Word Embeddings for Computational Social Science

James Foulds
University of Maryland, Baltimore County

Overview

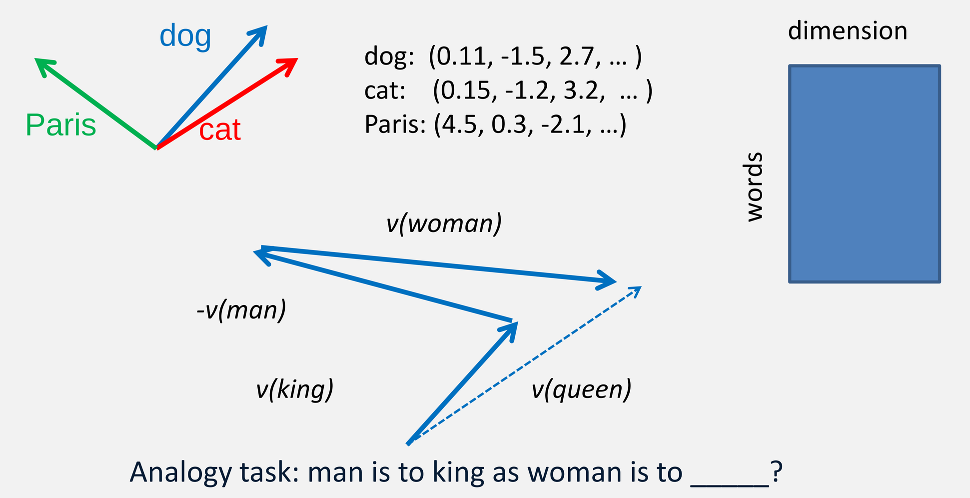
- **Word embeddings** find similarities between words, leading to **improved performance** for many NLP tasks:
  - translation, part-of-speech tagging, chunking, NER, …
- Allow NLP "from scratch," **without feature engineering**
- Typically trained in **big data** setting
- Have not yet been widely adopted for **computational social science** research due to the following limitations:
  - Target corpus of interest is often **not big data**
  - It is important for the model to be **interpretable**
- I propose a method for training **interpretable word embeddings** without **big data**, for **computational social science**, leveraging insights from **topic models**

Contributions

- **Interpretable, statistically efficient embedding model**
  - *Key insight:* **Mixed membership representation** for parameter sharing while retaining model flexibility
- **Efficient training algorithm,** using recent advances from both topic models and word embeddings:
  - **Metropolis-Hastings-Walker** (Li et al., 2014)
  - **Noise-contrastive estimation** (Gutmann and Hyvarinen, 2010, 2012) Proposed training algorithm is (amortized) sublinear time in the vocabulary size and number of topics
- Extensive quantitative experimental results
- Computational social science case studies
- **Practical recommendations** and insights based on these results, especially on the use of *generic big data embeddings*, which is a very common practice in NLP

Background: Word Embeddings

- Represent **dictionary words** with **vectors**. Similar words have similar vectors.



Skip-gram model (Mikolov et al., 2013)

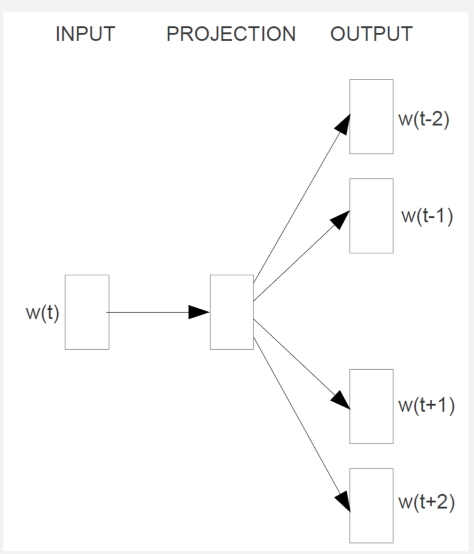


Figure due to Mikolov et al. (2013)

A log-bilinear classifier for the context of a given word

$$p(w_j|w_i) \propto \exp(v'^{\intercal}_{w_j} v_{w_i} + b_j)$$

v_w : "input" vectors
v'_w : "output" vectors
b_j : bias term

- Simple model, scales easily to large data sets
  - Can beat deep neural network models

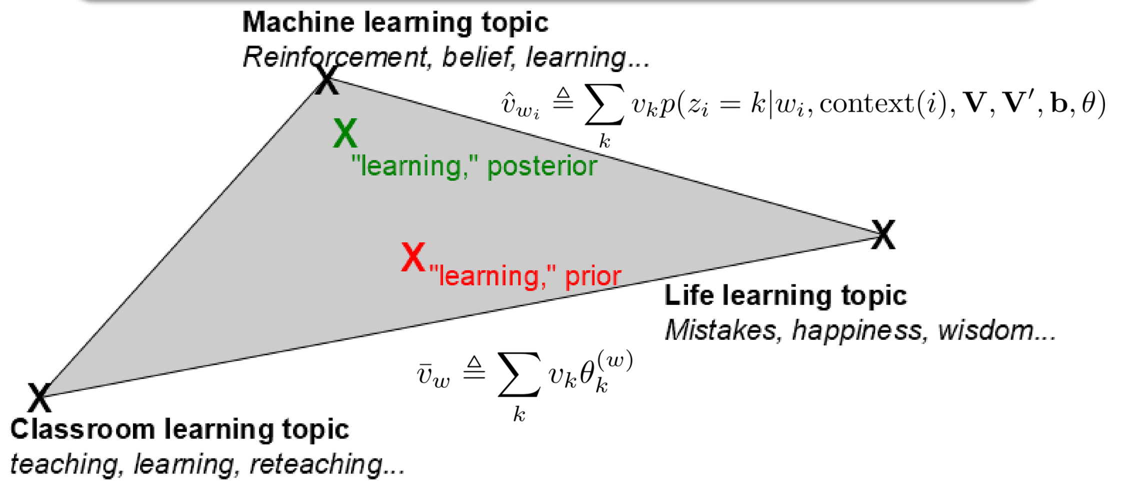
Connections to Topic Models, and Mixed Membership Extension to the Skip-Gram

- Skip-gram corresponds to a **supervised naïve Bayes topic model**, up to its parameterization via embeddings
- I propose **topic model** and **mixed membership** variants
- **Mixed membership** models provide parameter sharing
- Can use **fewer vectors than words** for small data, while retaining substantial representational power

	Skip-gram	Skip-gram topic model
Naive Bayes (Mikolov et al., 2013)	For each word in the corpus w_i For each word $w_c \in context(i)$ Draw w_c via $p(w_c\|w_i) \propto exp(v'^{\intercal}_{w_c} v_{w_i} + b_{w_c})$	For each word in the corpus w_i For each word $w_c \in context(i)$ Draw w_c via $p(w_c\|w_i) = \text{Discrete}(\phi^{(w_i)})$
Mixed membership (Proposed method)	For each word in the corpus w_i Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$ For each word $w_c \in context(i)$ Draw w_c via $p(w_c\|z_i) \propto exp(v'^{\intercal}_{w_c} v_{z_i} + b_{w_c})$	For each word in the corpus w_i Draw a topic $z_i \sim \text{Discrete}(\theta^{(w_i)})$ For each word $w_c \in context(i)$ Draw w_c via $p(w_c\|z_i) = \text{Discrete}(\phi^{(z_i)})$

reparameterize → (Skip-gram to Skip-gram topic model)
mixed membership extension ↓
← reparameterize (mixed membership topic model to proposed method)

Mixed Membership Word Embeddings



$$\hat{v}_{w_i} \triangleq \sum_k v_k p(z_i = k | w_i, \text{context}(i), \mathbf{V}, \mathbf{V}', \mathbf{b}, \theta)$$

$$\bar{v}_w \triangleq \sum_k v_k \theta_k^{(w)}$$

Context: "*We used an SVM when* ***learning*** *to predict the class labels.*"

Word embeddings are convex combinations of topic embeddings

- Words have **mixed membership distributions** over topics $\theta^{(w)}$
- **Topics** have embeddings v_k, words don't. Resolves **polysemy**
- **Fewer vectors than words**: **statistical efficiency** on small data
- Word embeddings recovered as prior mean $\bar{v}_w$ or posterior mean vectors $\hat{v}_{w_i}$**-convex combinations of topic embeddings**
- **Interpretable**: topics can be interpreted via top words lists, word embeddings are defined in terms of topic embeddings
- **Context can be leveraged** to improve embeddings $\bar{v}_w$ to $\hat{v}_{w_i}$ via the posterior distribution over topics for a word token w_i

Inference for MM Skip-Gram Topic Model

- Bayesian inference via collapsed Gibbs sampling

$$p(z_i = k|\cdot) \propto \left(n_k^{(w_i)\neg i} + \alpha_k\right) \prod_{c=1}^{|\text{context}(i)|} \frac{n_{w_c^{(i)}}^{(k)\neg i} + \beta_{w_c^{(i)}} + n_{w_c^{(i)}}^{(i,c)}}{n^{(k)\neg i} + \sum_{w'} \beta_{w'} + c - 1}$$

- Scale to many topics: **Metropolis-Hastings-Walker**
- Alias table data structure, amortized O(1) sampling
- "Mixture of experts" proposal, alias tables for words

$$c \sim \text{Uniform}(|\text{context}(w_i)|)\ ,\ q_{w_c}(k) \propto \frac{n_{w_c}^{(k)} + \beta_{w_c}}{n^{(k)} + \sum_{w'} \beta_{w'}}$$

- Simulated annealing to escape early local maxima

Amortized Sublinear Time Training for MM Skip-Gram

- Online EM impractical - O(KV) updates
- Key insight: MMSG topic model **equivalent** to word embedding model (*up to the Dirichlet prior*)
  - **Pre-solve E-step** via topic model CGS MHW algorithm
  - Apply **noise-contrastive estimation** to solve M-step

Experimental Results

Top Words in Topics

Model	Input word = "Bayesian" Top words in topic for input word. Top 3 topics for word shown for mixed membership models.
SGTM	model networks learning neural bayesian data models approach network framework
SG	belief learning framework models methods markov function bayesian based inference
MMSGTM	neural bayesian networks mackay computation framework practical learning weigend backpropagation model models bayesian prior data parameters likelihood priors structure graphical monte carlo chain markov sampling mcmc method methods model bayesian
MMSG	neural networks weigend bayesian data mackay learning computation practical probability model data models priors algorithm bayesian likelihood set parameters carlo monte mcmc chain reversible sampling model posterior

SG = skip-gram, TM = topic model, MM = mixed membership.

Predicting Held-Out Words

Dataset	Frequency baseline	Google +context	SG	SG +context	MMSG prior	MMSG posterior	SGTM	SGTM +context	MMSGTM prior	MMSGTM posterior
NIPS	0.029	0.027	0.038	0.031	0.037	**0.062**	**0.055**	**0.064**	**0.046**	**0.074**
SOTU	0.021	0.021	0.025	0.023	0.022	**0.034**	**0.036**	**0.046**	**0.032**	**0.045**
Shakespeare	0.015	**0.032**	0.020	0.010	0.015	0.019	**0.025**	**0.043**	0.020	**0.025**
Du Bois	0.028	0.033	0.045	0.037	0.041	**0.053**	**0.052**	**0.081**	**0.050**	**0.066**

Bold indicates statistically significant improvement versus SG

Prediction task:
- Predict context words, given an input word.
- Treat as ranking problem, report mean reciprocal rank (MRR) metric
- Training on **target corpus beats generic big-data** vectors (except for Shakespeare dataset)

- Using the **full context helps** (posterior over topic or summing vectors)
- **Mixed-membership models** (w/ posterior) **beat naïve Bayes models**
- **Topic models beat embedding models**

Downstream Tasks: Classification and Regression

Dataset	#Classes	#Topics	Tf-idf	Google	MMSG	SG	MMSGTM	SG+MMSG	SG+MMSG+Google
20 Newsgroups	20	200	83.33	52.50	55.58	59.50	64.08	66.55	72.53
Reuters-150	150	500	73.04	53.65	65.26	69.53	66.97	70.63	71.20
Ohsumed	23	500	43.07	20.56	31.82	37.57	32.41	39.53	40.27
SOTU (RMSE)	Regression	500	19.57	8.64	12.73	10.57	21.88	9.94	8.15

- Document categorization (classification accuracy, larger is better), and predicting the year of SOTU addresses (RMSE, smaller is better).
- **Target corpus beats generic big-data vectors** (except for SOTU, which is very small)
- **Skip-gram beats MMSG** for classification/regression – loss of granularity
- But, **concatenating the different vectors improves performance** over individual embeddings
  - MMSG, SG, generic Google vectors learn **complementary information**

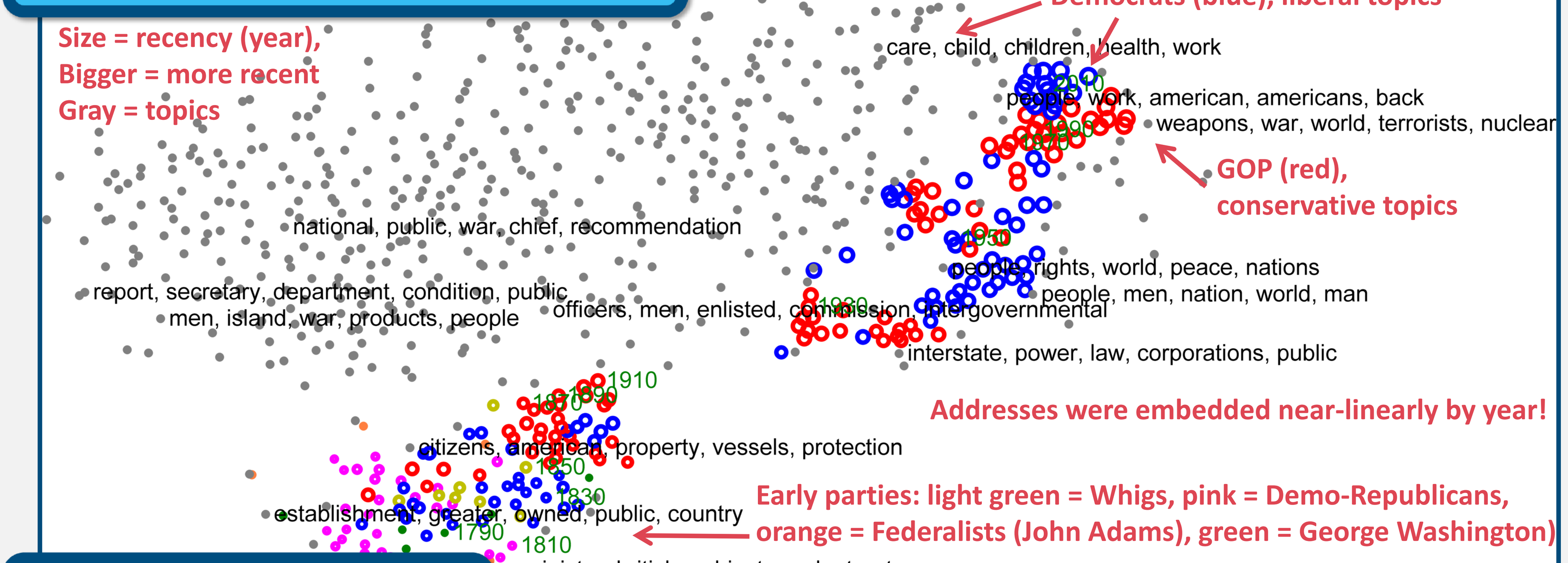
Vector Composition in Topic Space

Nearest topic after composition of mean vectors for words

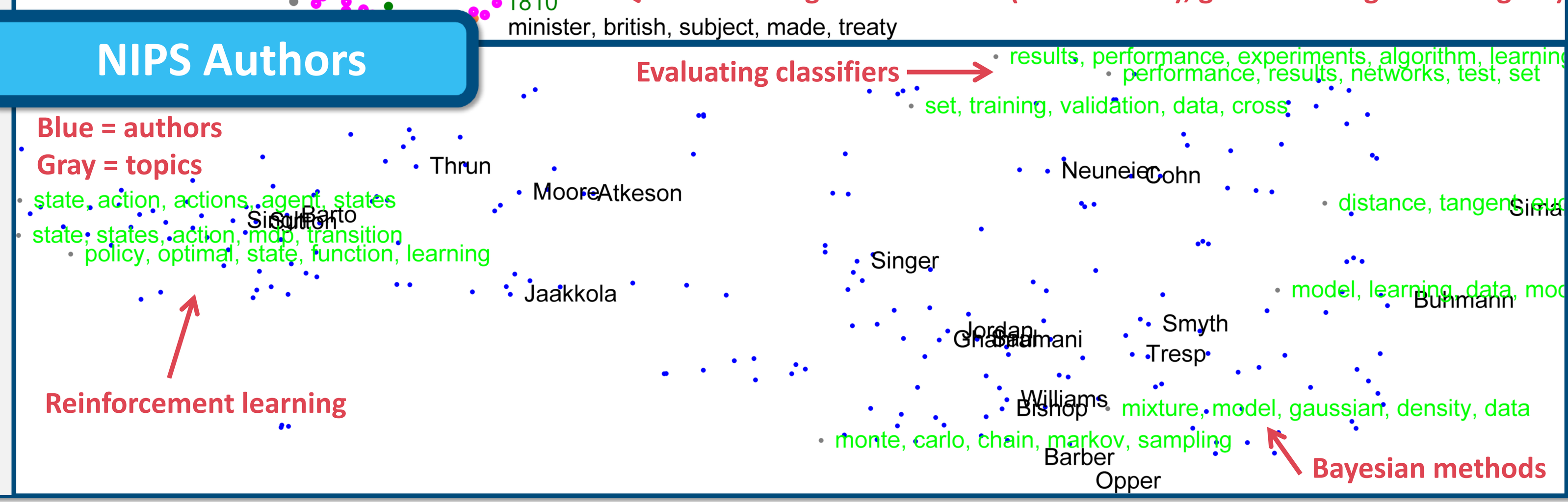
object + recognition	objects visual object recognition model
character + recognition	recognition segmentation character
speech + recognition	speech recognition hmm system hybrid
computer + vision	computer vision ieee image pattern
computer + science	university science colorado department
bias + variance	error training set data performance
covariance + variance	gaussian distribution model matrix

Data Visualization: Document, Topic, and Author Embeddings on State of the Union Addresses and NIPS Articles (t-SNE Projections)

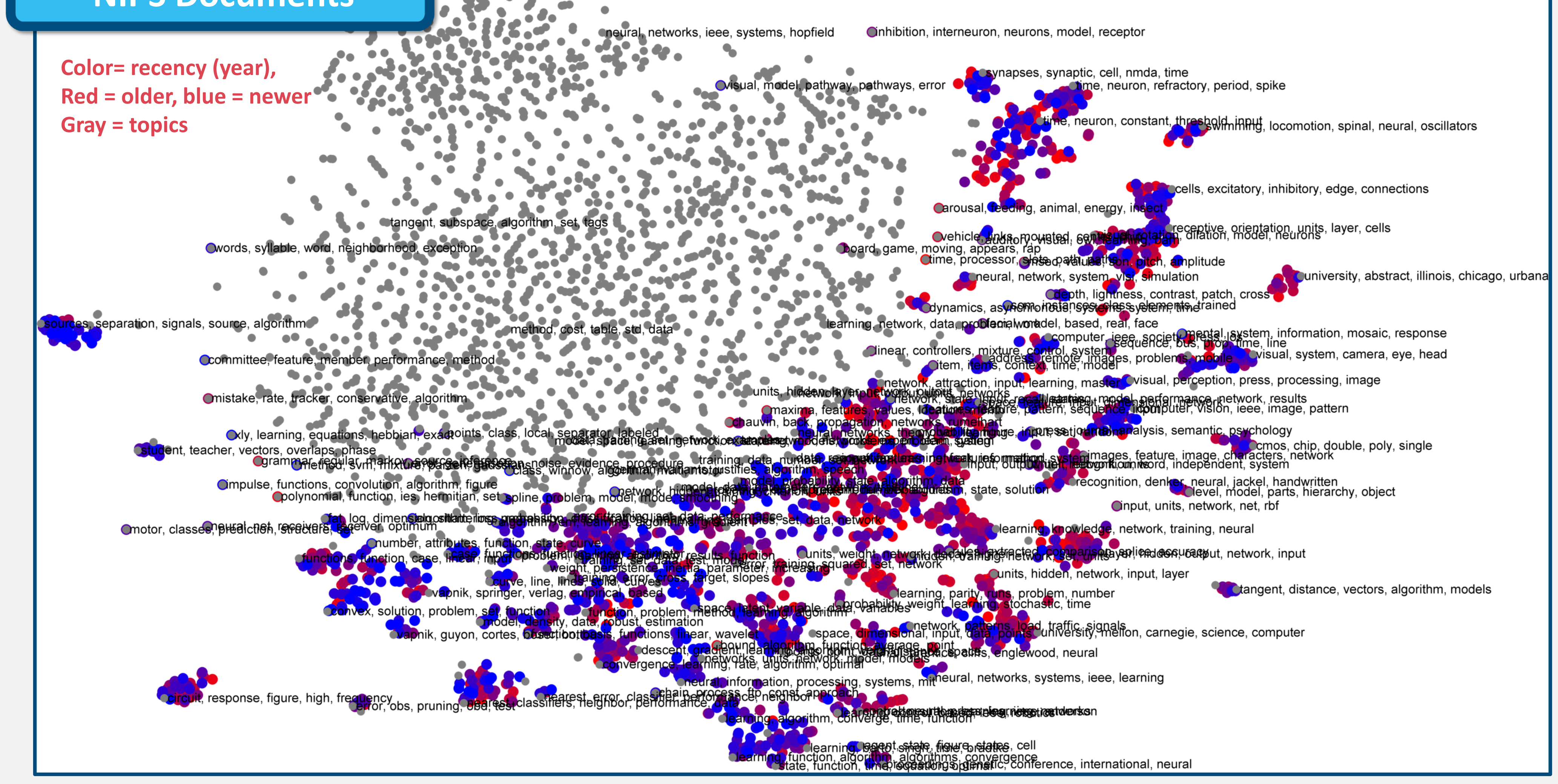
State of the Union Addresses



NIPS Authors



NIPS Documents



jfoulds@umbc.edu