Unsupervised Dimensionality Reduction

ECE57000: Artificial Intelligence, Fall 2019

David I. Inouye

Announcements

- Must submit term paper PDF to TWO assignments on Blackboard
  - Peer assessment
  - Instructor assessment
- Course project
  - No deadline extension
  - Paper should be **5.5-6** pages long (excluding references)
  - Strict limit on 6 pages, can include appendix
  - I'd suggest finish term paper with minimum viable product, then improve if time available
  - Looking for deeper understanding via writing and implementation (rather than results themselves)

Announcements: (Tentative) Review format

- 1. Please summarize the key idea in each published paper that this term paper reports on in one sentence.
  - If the paper does not have clear headings for the 3 selected papers (e.g., the paper has a single "Related Works" section), please just choose 3 papers that are cited and discussed in the related works section. Some term papers may discuss more than 3 papers.
- 2. Please summarize the implementation that this term paper reports. State what the implementation takes as input (in one sentence). State what the implementation produces as output (in one sentence). Describe the algorithm in English, mathematical notation, or pseudocode.
- 3. Please summarize the experiments/evaluations and results in one sentence.
- 4. What didn't you understand in this term paper (one sentence)?
- 5. How can the author improve this term paper (one sentence)?
- Please "reverse" rank (where 5 is best) this selection in comparison to the other papers you are reviewing, i.e., 5 = Best paper, ..., 1= Worst paper.

Announcements: A few principles for reviews (Credit: Prof. Jeffrey Siskind)

1. It is imperative to be polite in reviews.
2. The primary purpose of the review is not to criticize the author or their work; it is to help them improve their work.
3. The most helpful things in reviews are suggestions about how to improve the paper.
4. Telling the author what you understood and what you didn't also helps the author improve the paper.

Why dimensionality reduction?
Lower computation costs

- Suppose original dimension is large like $\mathrm{d} = 10000$
(e.g., images, DNA sequencing, or text)

- If we reduce to $k = 100$ dimensions, the training algorithm can be sped up by $100\times$

Why dimensionality reduction? Visualization

- Allows 2D scatterplot visualizations even of high-dimensional data (2D projection of digits)

https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

Why dimensionality reduction?
Noise reduction via reconstruction

Demo of PCA via sklearn

- Random projections vs PCA projections
- Visualizations of
  - Minimum reconstruction error
  - Maximum variance
  - Explained variance based on k
- Code examples
  - Digits
  - Eigenfaces

Relation to clustering: One-hot vectors vs continuous vectors

- k-means clustering can be seen as reducing the dimensionality to k latent categories
  - Each category can be represented by a one-hot vector of length k
    e.g., if $k = 3$, $z_i \in \{[1,0,0], [0,1,0], [0,0,1]\}, \forall i$
  - Every instance can only "belong" to one category
- In dimensionality reduction techniques, the latent vectors can have non-zeros for all k latent dimensions
  - e.g., if $k = 3$, $z_i \in \mathbb{R}^3, \forall i$

Relation to clustering: K-means objective can be reformulated as seeking the best approximation to X with low rank constraint ($k < d$)

- Original k-means objective

$$\min_{\substack{\mathcal{C}_1,\ldots,\mathcal{C}_k \\ \mu_1,\ldots,\mu_k}} \sum_{j=1}^{k} \sum_{x \in \mathcal{C}_j} \|x - \mu_j\|_2^2$$

- Equivalent to the following objective

$$\min_{Z,M} \|X - ZM\|_F^2$$

$$\text{where } Z \in \{0,1\}^{n\times k}, \sum_j z_{ij} = 1, \forall i$$

$$\text{and } M \in \mathbb{R}^{k\times d}$$

- What if we relax the constraint on Z?

Derivation of equivalence between two objectives for k-means

- $y_i \in \{1, \dots, k\}$ is the cluster label for each instance
- z_i is the corresponding one hot vector to y_i
- $M = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}$ is the matrix of mean vectors
- $\sum_{j=1}^{k} \sum_{x \in \mathcal{C}_j} \left\| x - \mu_j \right\|_2^2 = \sum_{i=1}^{n} \left\| x_i - \mu_{y_i} \right\|_2^2 = \sum_{i=1}^{n} \| x_i - z_i M \|_2^2 = \sum_{i=1}^{n} \sum_{s=1}^{d} \left(x_{is} - z_i^T m_s \right)^2 = \left(\sqrt{\sum_{i=1}^{n} \sum_{s=1}^{d} \left(x_{is} - z_i^T m_s \right)^2} \right)^2 = \| X - ZM \|_F^2$

Principal Component Analysis (PCA) can be seen as minimizing the reconstruction error of the data using only $k \leq d$ components

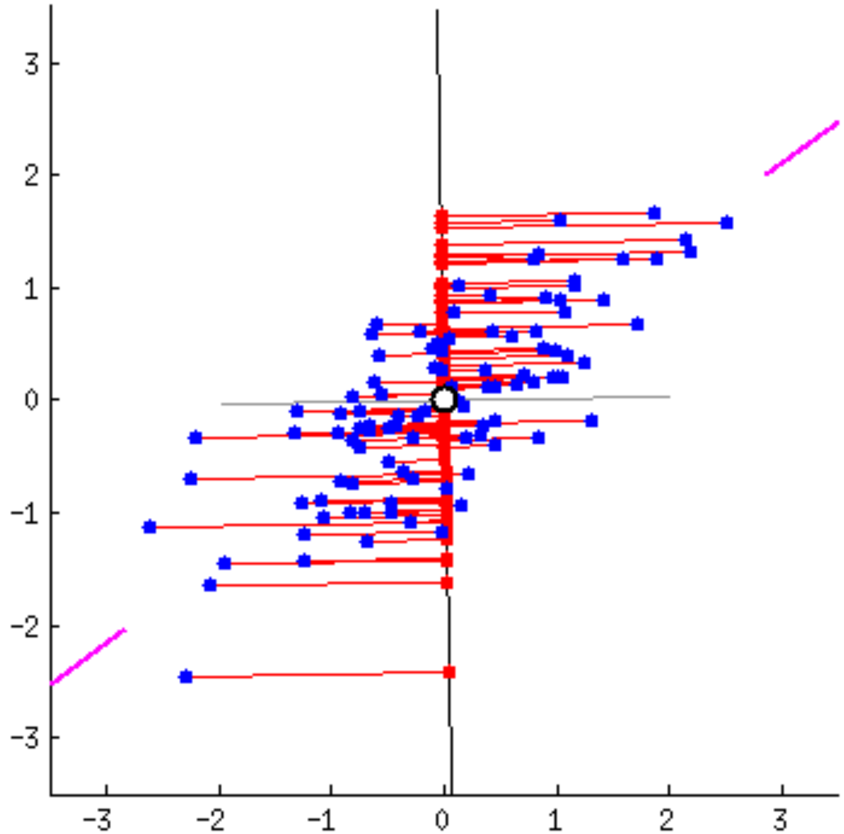
- (compare errors on board - cluster vs. PCA)
- Similar to clustering except Z is unconstrained and W^T has orthogonal rows

$$\min_{\mathrm{Z,W}} \|X_c - ZW^T\|_F^2$$

- where

  $\mathrm{X_c} = \mathrm{X} - \mu_x$ (centered)
  $Z \in \mathbb{R}^{n \times k}$ (latent representation)
  $W^T \in \mathbb{R}^{k \times d}$ (principal components)
  $w_s^T w_t = 0, w_s^T w_s = \|w_s\|_2 = 1, \forall s, t$
  (orthogonal constraint)

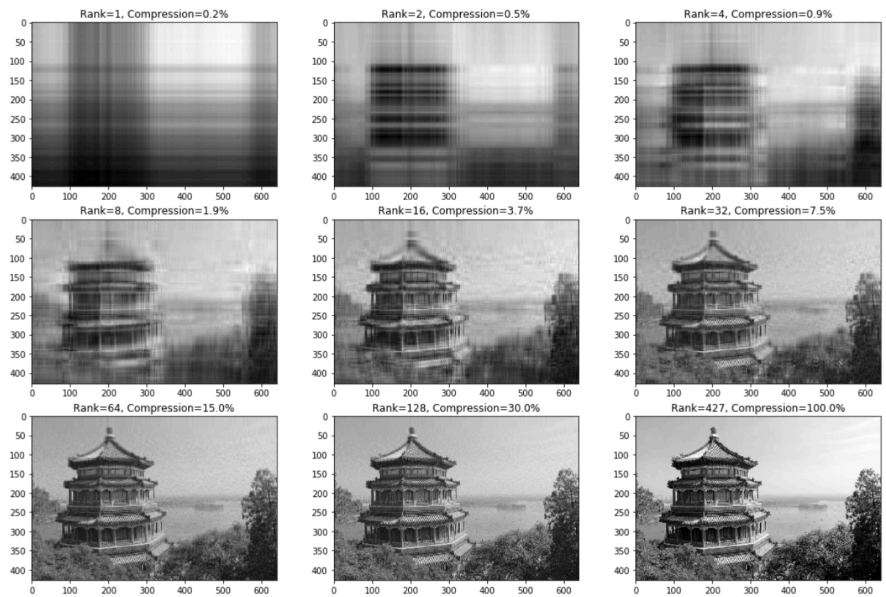
Minimum reconstruction error (red bars)



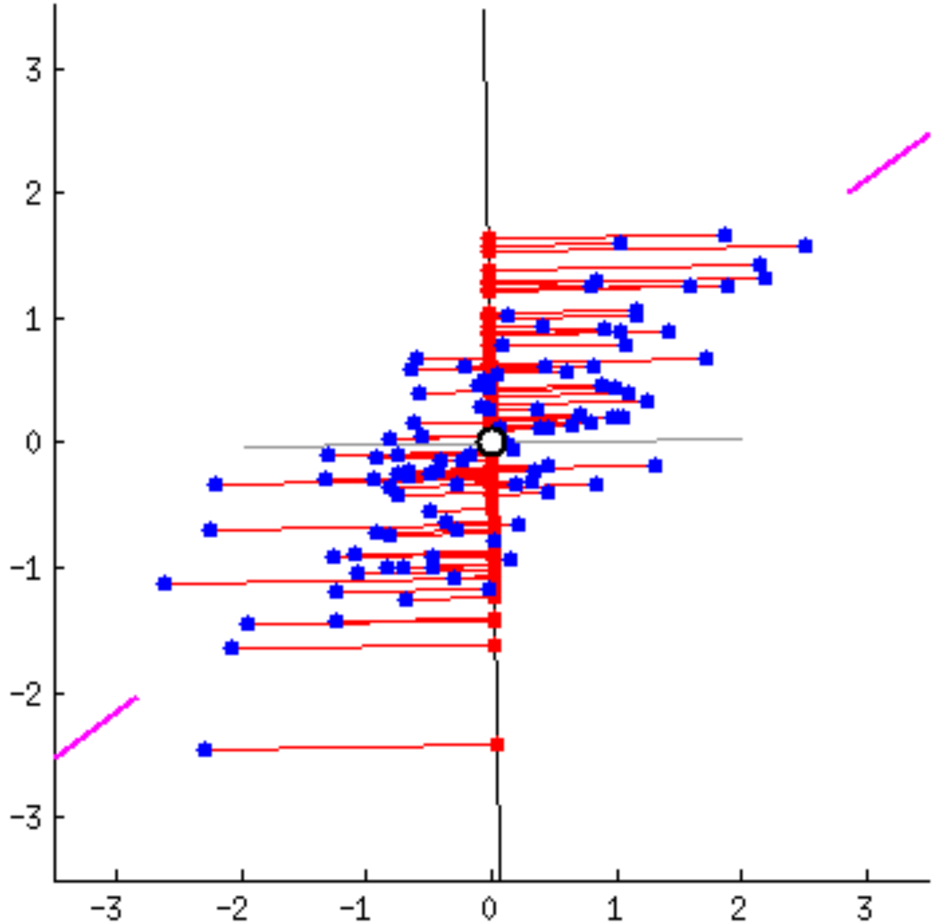
https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

The solution can be computed as the top k right singular vectors via SVD (lowest reconstruction error)

- If $X_c = USV^T$, then the solution to the previous problem is simply $W^T = V_{1:k}^T$
  - i.e. the first k singular vectors
- Remember if $k = d$, then perfect reconstruction



Minimum reconstruction error (red bars) = Maximum latent variance (spread of red points)



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

Minimizing reconstruction error is equivalent to maximizing variance of latent projection

- (compare interpretations on board)
- Consider one-dimensional projection, i.e. $k = 1$
- Let $z = w^T x$, where $\|w\|_2 = 1$
- What is the empirical variance?
  - For simplicity, we assume X has a mean of 0.

$$\widehat{var}[z] = \hat{\mathbb{E}}[(z - \mu_z)^2] = \hat{\mathbb{E}}[z^2]$$
$$= \hat{\mathbb{E}}[(w^T x)(w^T x)] = \hat{\mathbb{E}}[w^T (x x^T) w]$$
$$= w^T \hat{\mathbb{E}}[x x^T] w = w^T \hat{\Sigma}_x w$$

- Thus we have the following:

$$\max_w w^T \hat{\Sigma}_x w \,, \qquad s.t. \|w\|_2 = 1$$

The solution is the eigenvector with the largest eigenvalue of $\hat{\Sigma}_x$
For general k, the solution is the top k eigenvectors

- Suppose $\hat{\Sigma}_x = Q\Lambda \mathrm{Q}^{\mathrm{T}}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ (because $\hat{\Sigma}_x$ is positive semi-definite)
- Then, $w^* = q_1 = \arg\max_w w^T \hat{\Sigma}_x w$
- The more general case

$$W^* = Q_{1:k} = \arg \max_{W \in \mathbb{R}^{d \times k}} \sum_{j=1}^{k} w_j^T \hat{\Sigma}_x w_j$$

$$\text{where } w_s^T w_t = 0, w_s^T w_s = \|w_s\|_2 = 1, \forall s, t$$

Both approaches get the first k right singular vectors of X_c

- Minimize reconstruction error
  - Singular value decomposition (SVD) of $X_c = USV^T$
  - Solution: $W^T = V_{1:k}^T$

- Maximize variance of latent projection
  - Eigendecomposition of covariance
    $\widehat{\mathbb{E}}[xx^T] = X_c^T X_c = (USV^T)^T(USV^T)$
    $= (VSU^T)(USV^T) = VS(U^T U)SV^T = VS^2V^T$
    $= Q\Lambda Q^T$
  - Solution: $W^T = Q_{1:k}^T = V_{1:k}^T$

Non-negative matrix factorization (NMF) provides more of a part-based representation than PCA

- Objective NMF

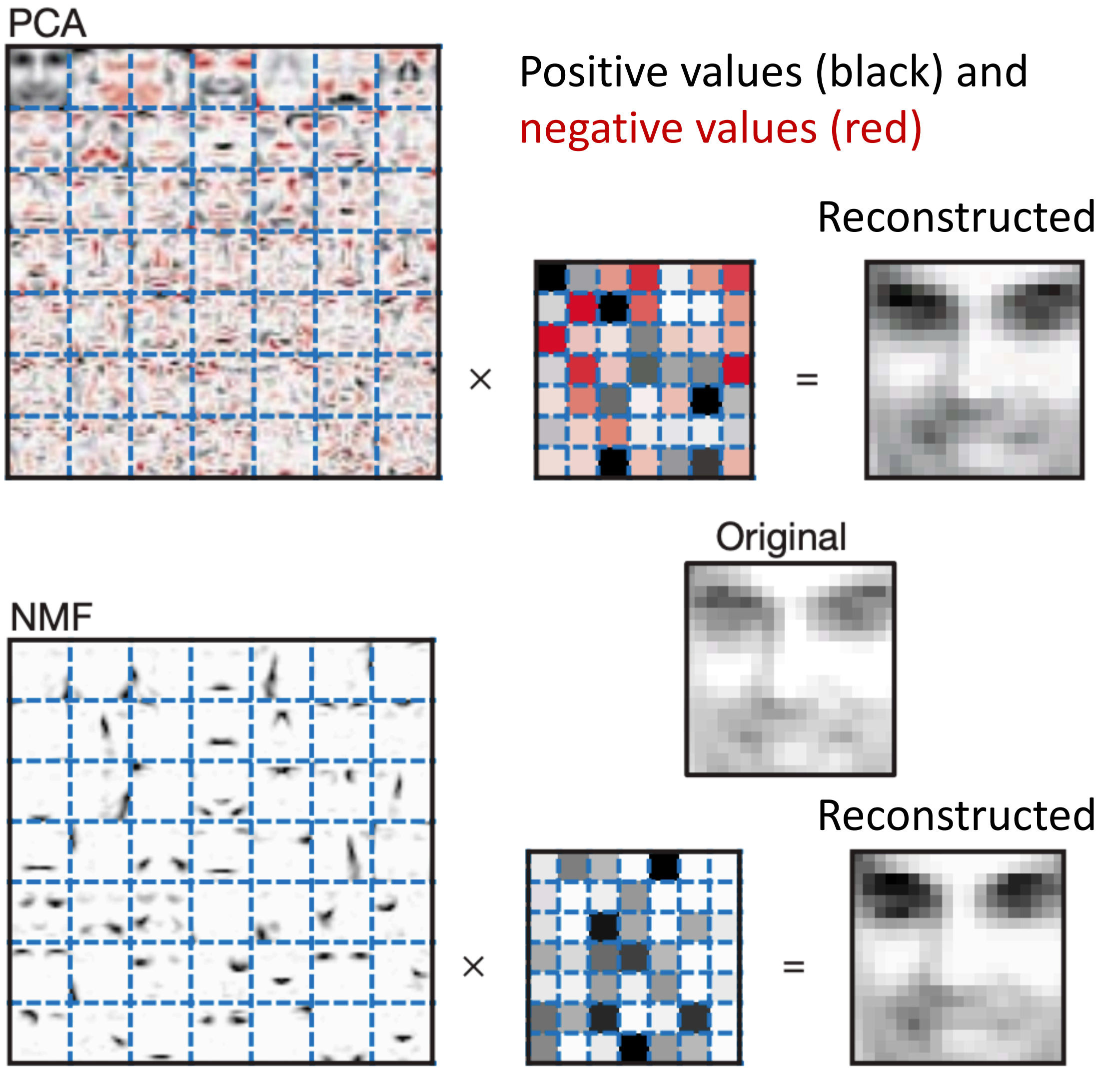
$$\min_{Z,W} \|X - ZW^T\|_F^2$$

where

$X \in \mathbb{R}_{+}^{n\times d}$

$Z \in \mathbb{R}_{+}^{n\times k}$

$W^T \in \mathbb{R}_{+}^{k\times d}$



Lee, Daniel D and Seung, H Sebastian (1999). "Learning the parts of objects by non-negative matrix factorization" (PDF). *Nature*. **401** (6755): 788–791. http://www.columbia.edu/~jwp2128/Teaching/E4903/papers/nmf_nature.pdf

NMF on document-word count matrix can be seen to identify underlying topics/factors

- Suppose we have a collection of n documents and there are d unique words
- Let each dimension correspond to the count of that word in the document
- Example:
  - "Intelligent applications creates intelligent business processes"
  - "Bots are intelligent applications"
  - "I do business intelligence"
- Non-negative document-word matrix

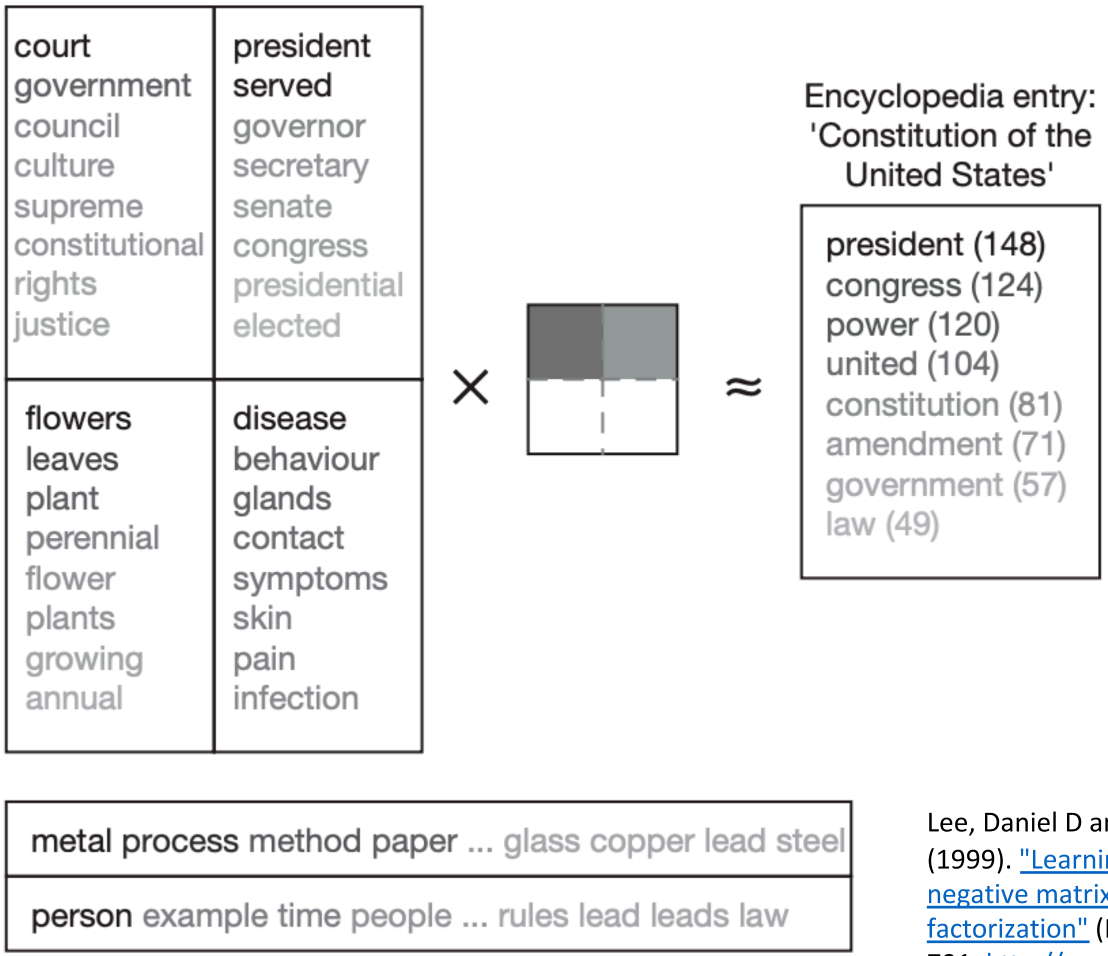
	intelligent	applications	creates	business	processes	bots	are	i	do	intelligence
Doc 1	2	1	1	1	1	0	0	0	0	0
Doc 2	1	1	0	0	0	1	1	0	0	0
Doc 3	0	0	0	1	0	0	0	1	1	1

https://www.darrinbishop.com/blog/2017/10/text-analytics-document-term-matrix/

NMF on encyclopedia articles reveals underlying topics in each document

4 selected components or "topics"

Other irrelevant topics



Lee, Daniel D and Seung, H Sebastian (1999). ["Learning the parts of objects by non-negative matrix factorization"](http://www.columbia.edu/~jwp2128/Teaching/E4903/papers/nmf_nature.pdf) (PDF). *Nature*. **401** (6755): 788–791. http://www.columbia.edu/~jwp2128/Teaching/E4903/papers/nmf_nature.pdf

<u>Probabilistic Latent Semantic Analysis (PLSA)</u> can be seen as non-negative matrix factorization with KL divergence loss (instead of squared error)

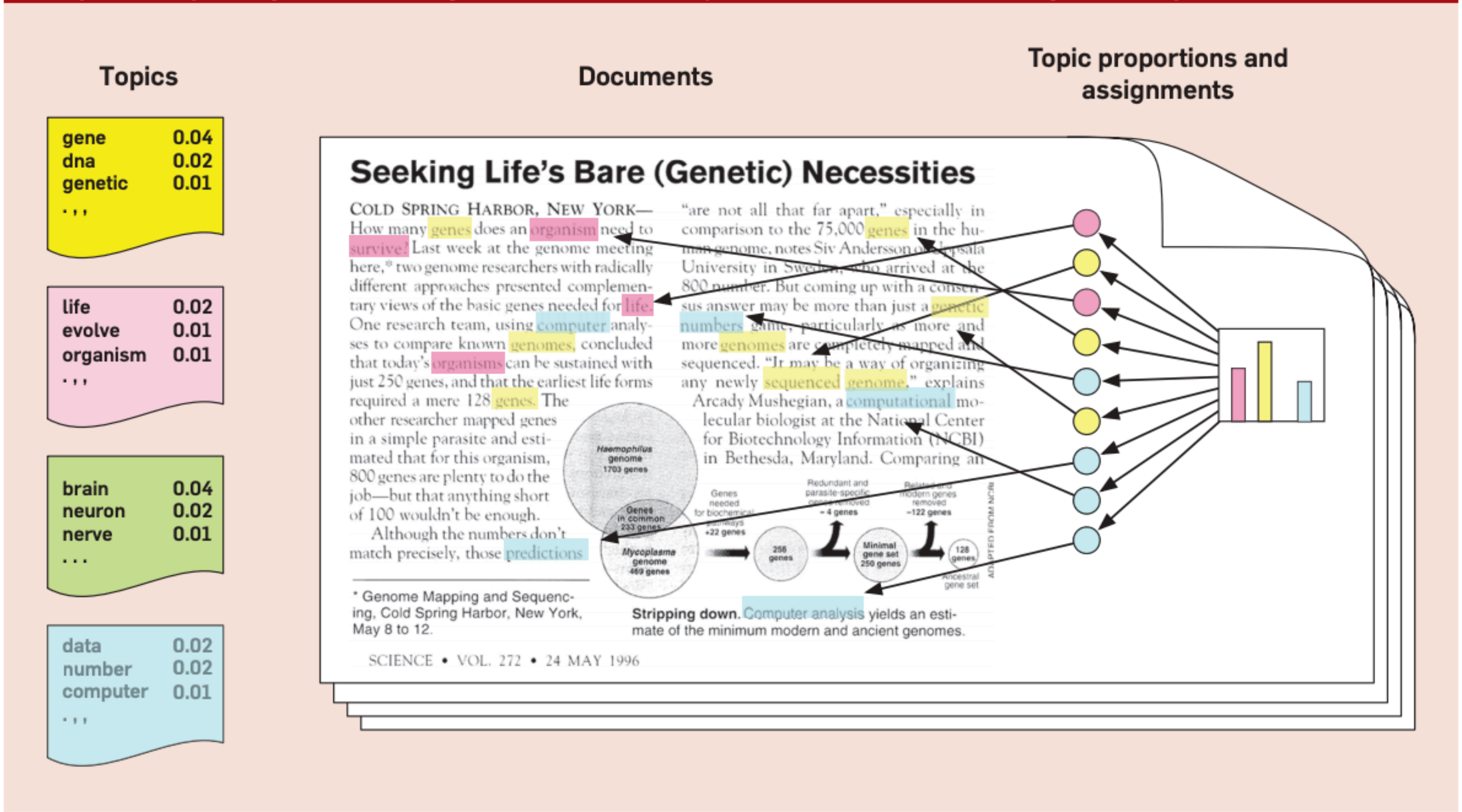
government	president	banks	pct	union	marks	gold	billion
tax	chairman	debt	january	air	currency	steel	dlrs
budget	executive	brazil	february	workers	dollar	plant	year
cut	chief	new	rise	strike	german	mining	surplus
spending	officer	loans	rose	airlines	bundesbank	copper	deficit
cuts	vice	dlrs	1986	aircraft	central	tons	foreign
deficit	company	bankers	december	port	mark	silver	current
taxes	named	bank	year	boeing	west	metal	trade
reform	board	payments	fell	employees	dollars	production	account
billion	director	billion	prices	airline	dealers	ounces	reserves
trading	american	trade	oil	vs	areas	food	house
exchange	general	japan	crude	cts	weather	drug	reagan
futures	motors	japanese	energy	net	area	study	president
stock	chrysler	ec	petroleum	loss	normal	aids	administration
options	gm	states	prices	mln	good	product	congress
index	car	united	bpd	shr	crop	treatment	white
contracts	ford	officials	barrels	qtr	damage	company	secretary
market	test	community	barrel	revs	caused	environmental	told
london	cars	european	exploration	profit	affected	products	volcker
exchanges	motor	imports	price	note	people	approval	reagans

Thomas Hofmann, *Learning the Similarity of Documents : an information-geometric approach to document retrieval and categorization*, Advances in Neural Information Processing Systems 12, pp-914-920, MIT Press, 2000

Equivalence formalized in:
Gaussier, E., & Goutte, C. (2005, August). Relation between PLSA and NMF and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 601-602). ACM.

The more well-known variant of <u>topic modeling</u> is called <u>Latent Dirichlet Allocation (LDA)</u>

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



David M. Blei, Probabilistic Topic Models, *Communications of the ACM*, 2012.