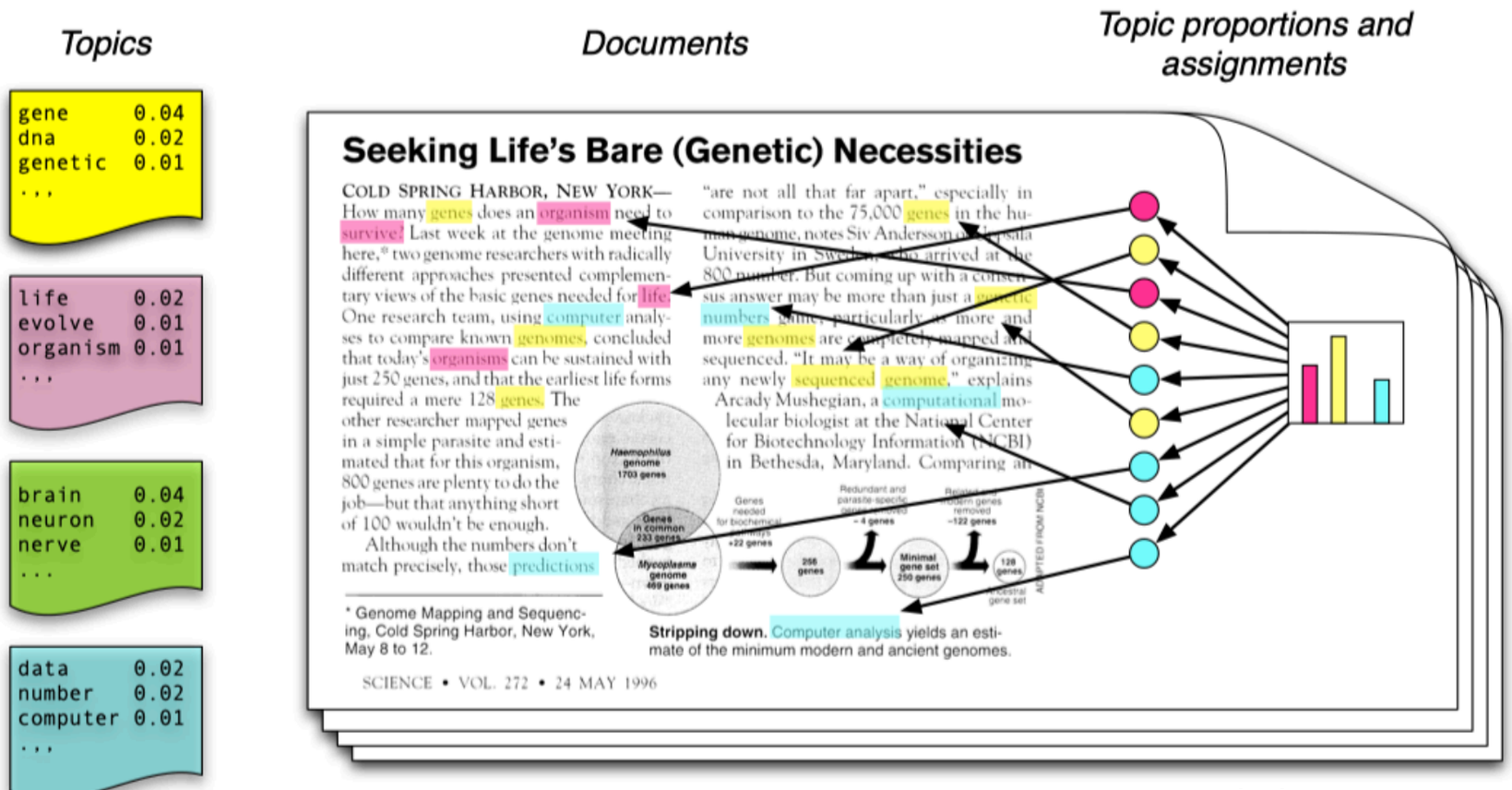
# Topic Models (I)

Runzhe Yang

2019/03/20

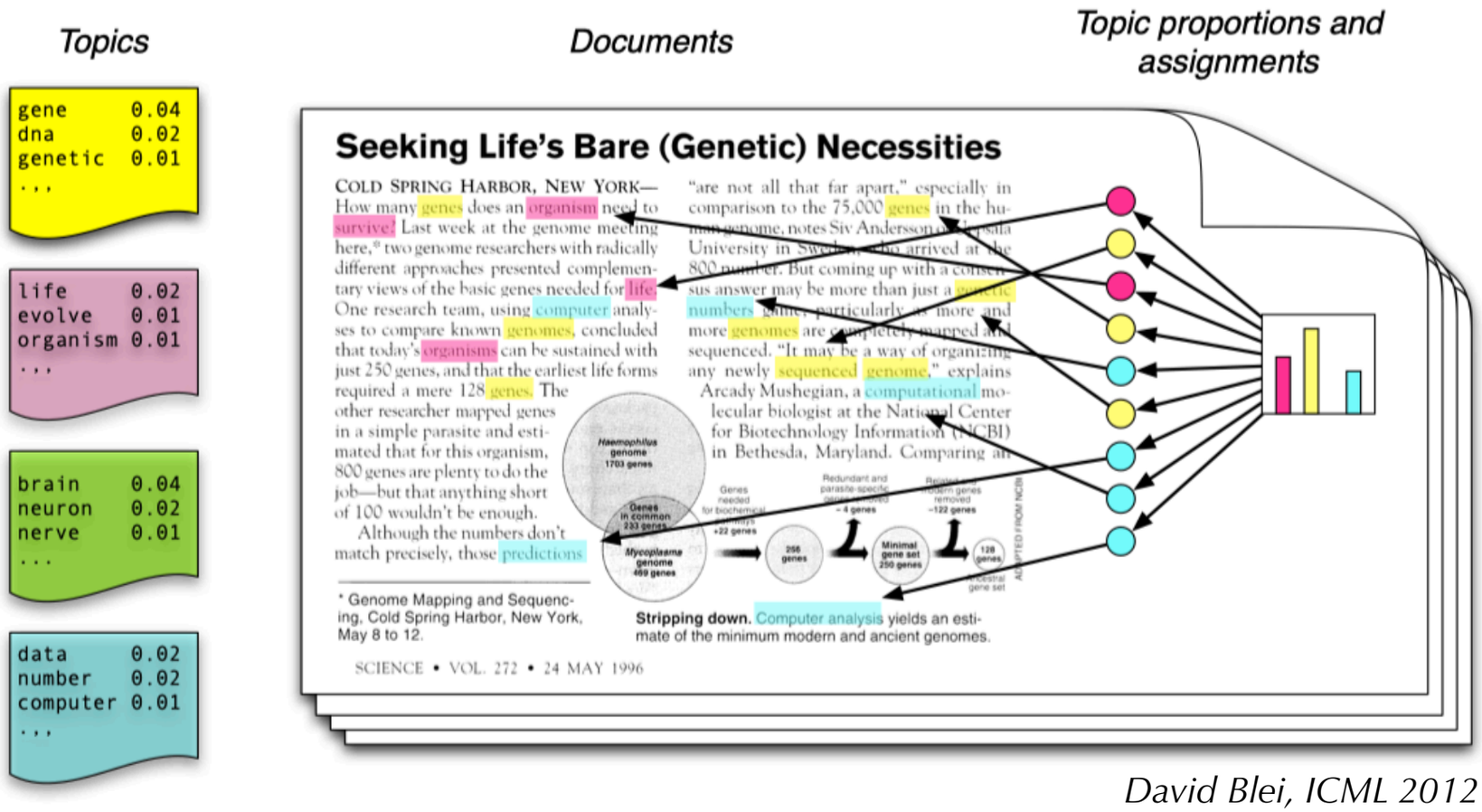
# What is “topic”?

A **hidden structure** that helps determine what **words** are likely to appear in a **document**.




*David Blei, ICML 2012*

# What is “topic”?

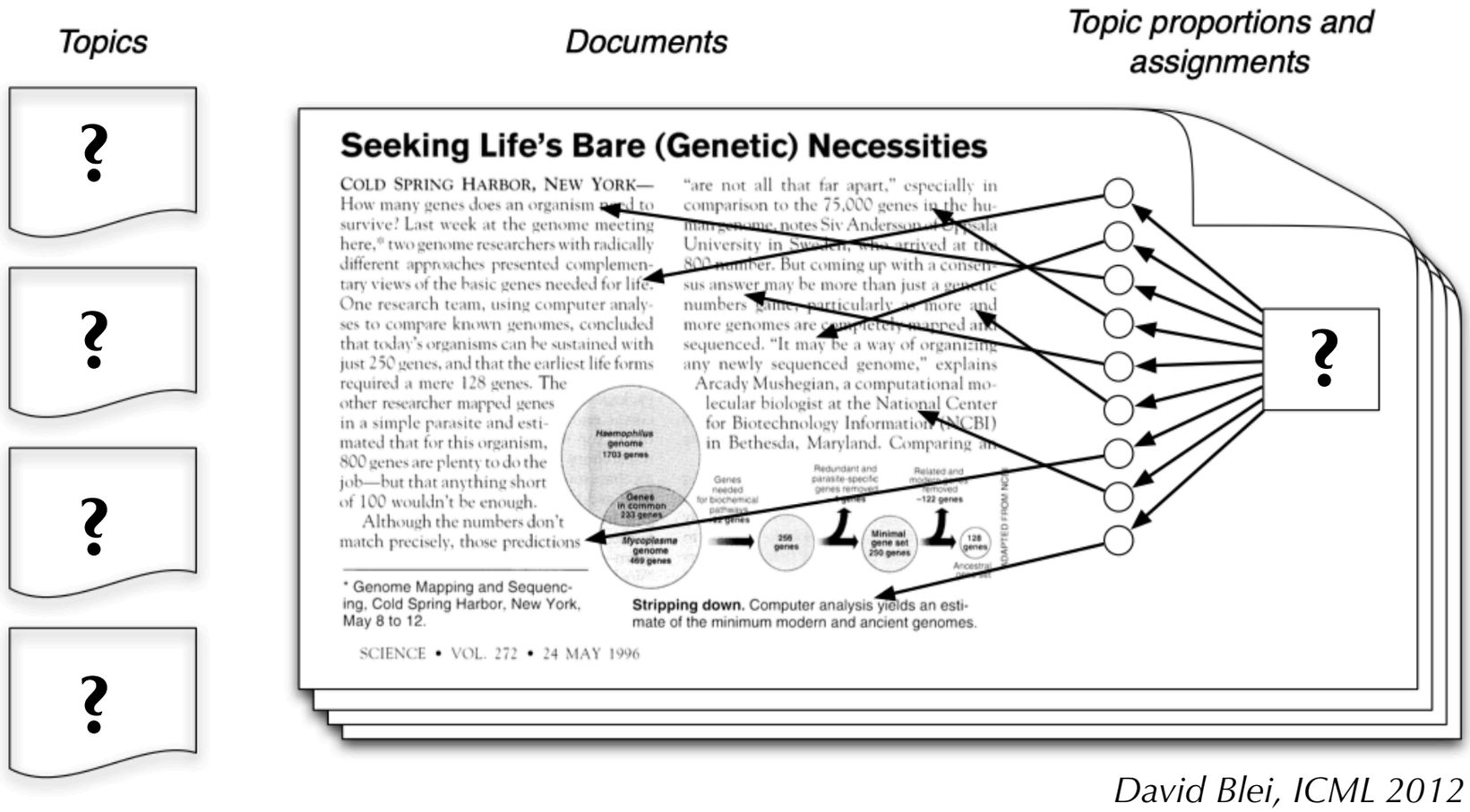



*David Blei, ICML 2012*

- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those **topics**

# What is “topic modeling”?

We only observe the **documents**.




*David Blei, ICML 2012*

The goal of ***topic modeling*** is
to infer the **hidden structures**. ——

1) what’re the topics of the corpus?
2) what’re the topics of this document?

Outline:

- **Classic Methods Overview**
  - Latent Semantic Indexing (LSI) - SVD
  - Probabilistic Latent Semantic Indexing (pLSI)
    - Equivalence to NMF & Autoencoder (?)
  - Latent Dirichlet Allocation (LDA)
    - Neural Topic Modeling (Gaussian prior)
  - **A comparison of SVD / NMF / LDA**

- **Community Detection & Topic Modeling**
  - ~~Stochastic Block Model (SBM)~~
  - ~~Hierarchical Stochastic Block Model (hSBM)~~

- **Beyond Probabilistic Generative Models**
  - Disynaptic Neural Network
  - Connection with Correlation Game
  - **Disynaptic Neural Topic Modelling?**

Outline:

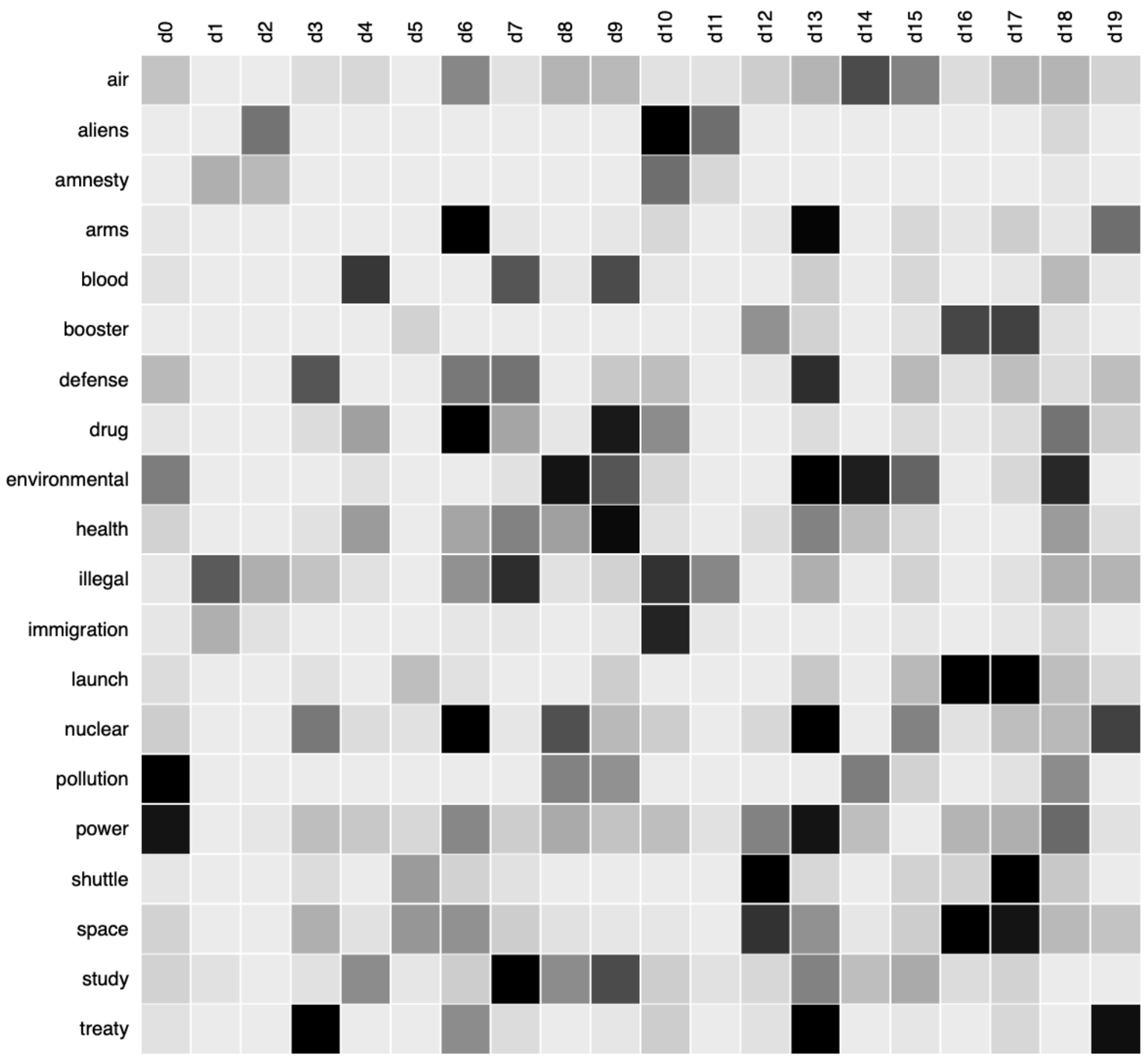
- **Classic Methods Overview**
  - Latent Semantic Indexing (LSI) - SVD
  - Probabilistic Latent Semantic Indexing (pLSI)
    - Equivalence to NMF & Autoencoder (?)
  - Latent Dirichlet Allocation (LDA)
    - Neural Topic Modeling (Gaussian prior)
  - **A comparison of SVD / NMF / LDA**

- **Community Detection & Topic Modeling**
  - Stochastic Block Model (SBM)
  - Hierarchical Stochastic Block Model (hSBM)

- **Beyond Probabilistic Generative Models**
  - Disynaptic Neural Network
  - Correlation Game
  - **Disynaptic Neural Topic Modelling?**

**Please stop and correct me whenever you think I am wrong!**

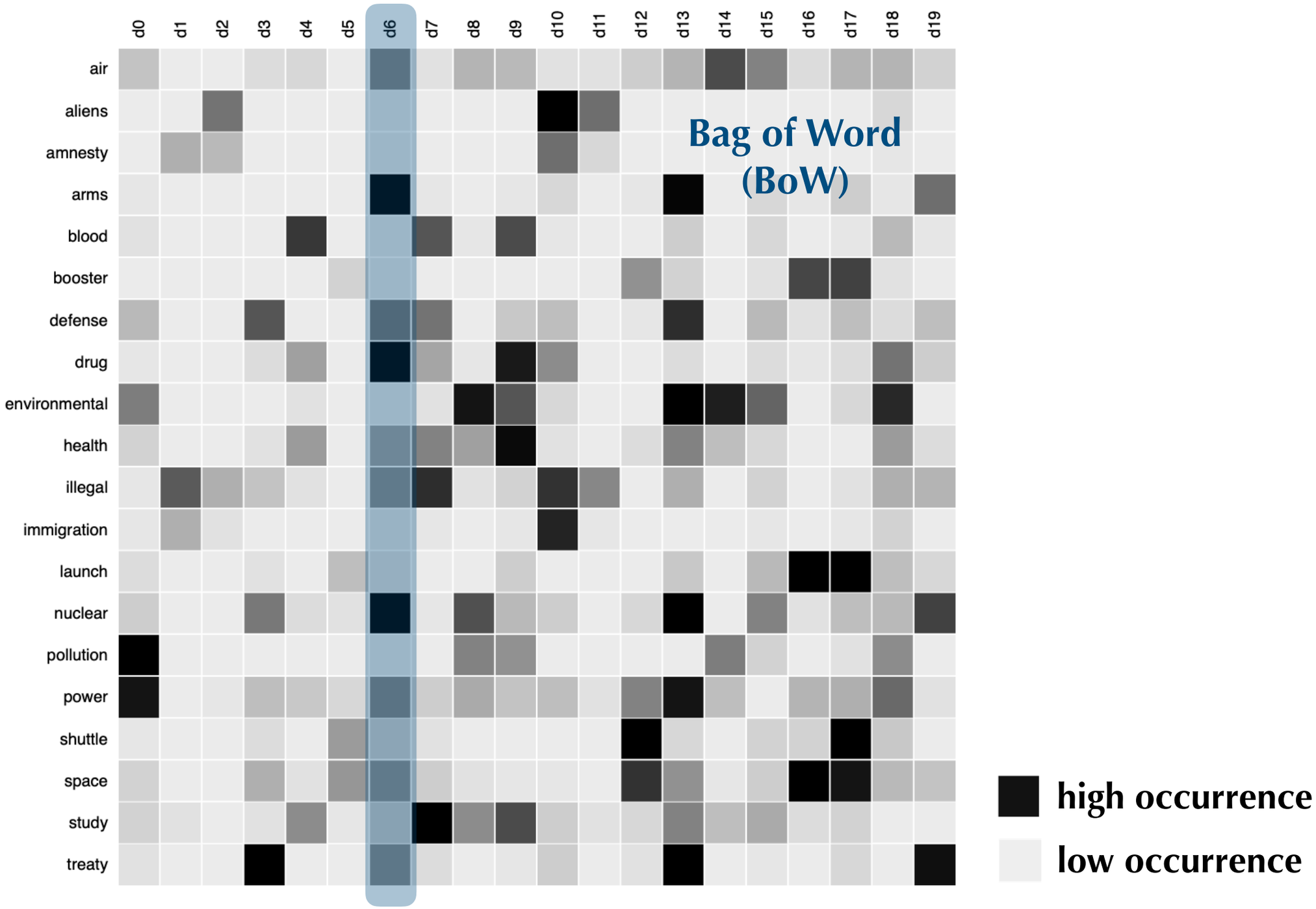
# Latent Semantic Indexing (LSI)




**Word x Document Matrix**

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

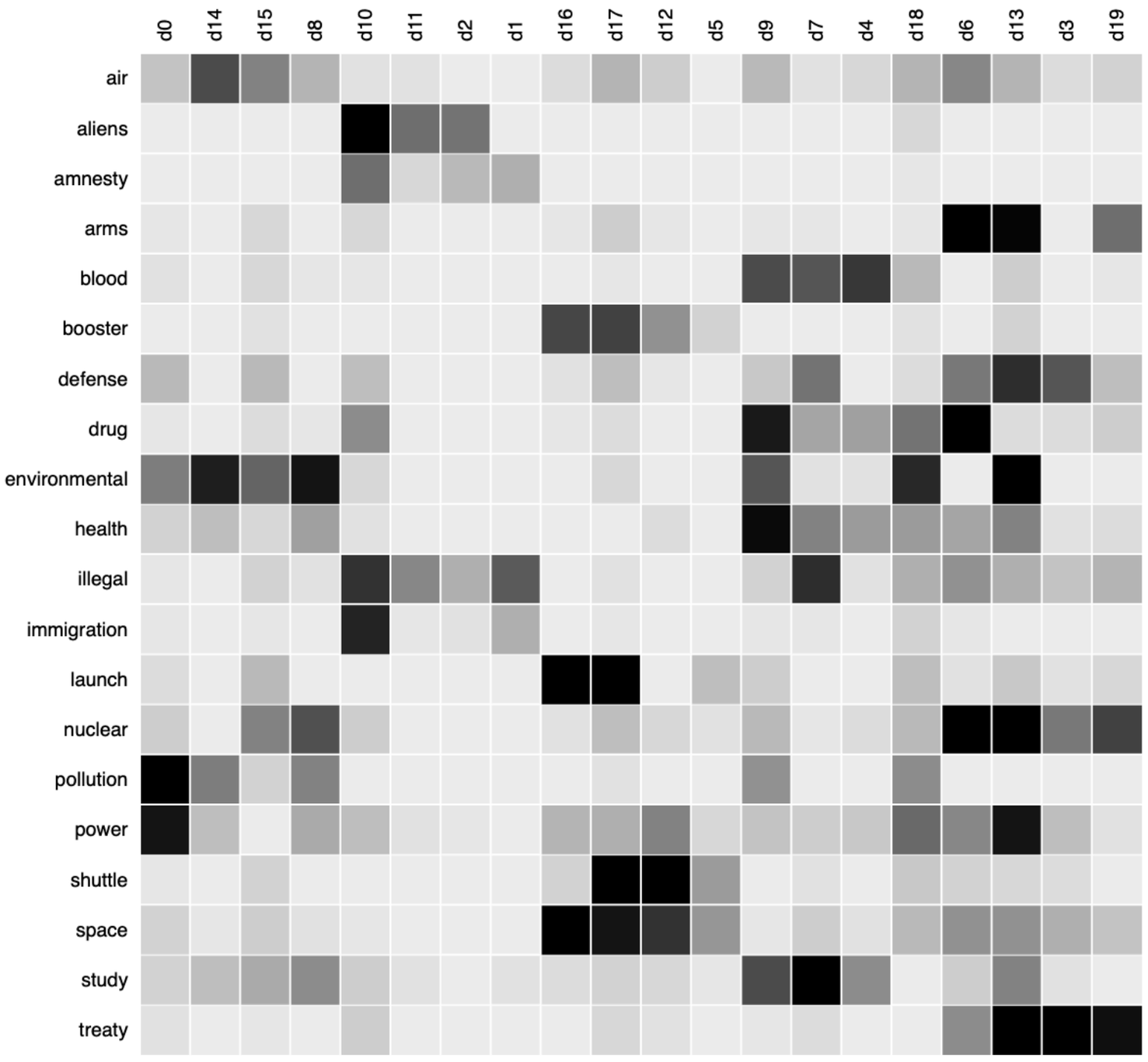
# Latent Semantic Indexing (LSI)




**Word x Document Matrix**

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

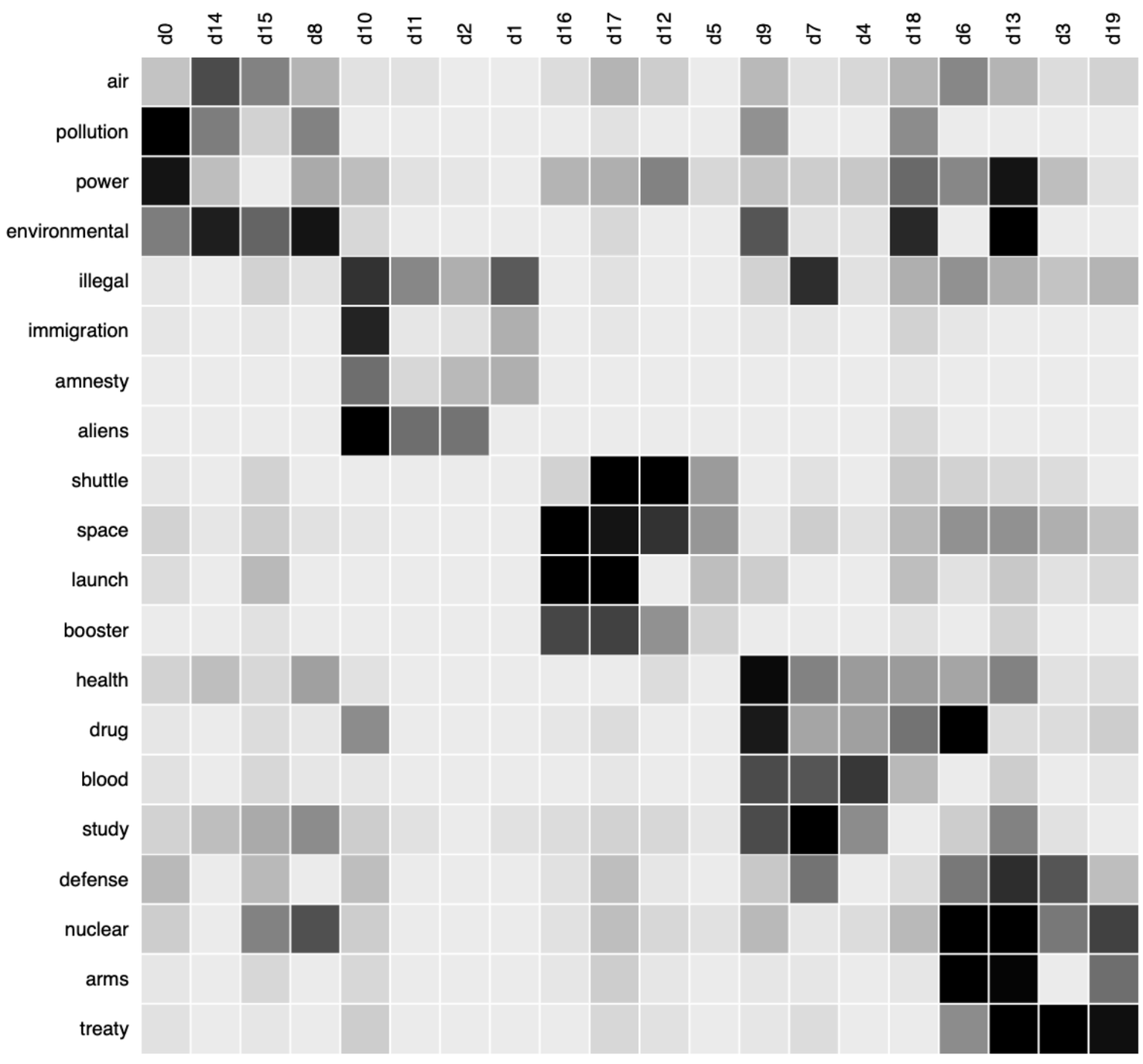
# Latent Semantic Indexing (LSI)




**Word x Document Matrix - Sort by columns (docs with similar words)**

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

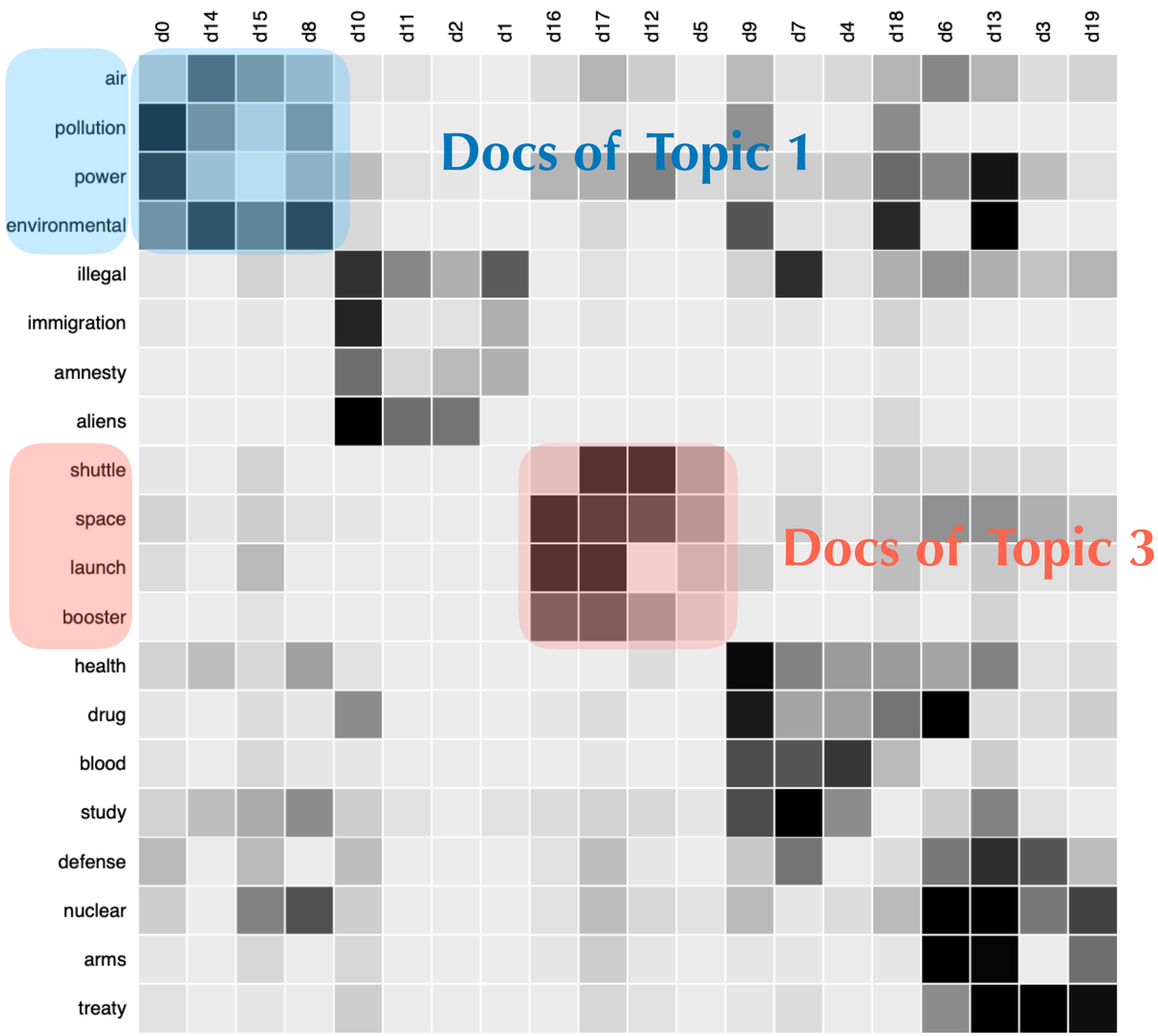
# Latent Semantic Indexing (LSI)




**Word x Document Matrix - Sort by rows (words occurring in similar docs)**

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)




**Topic 1**
**"environment"**

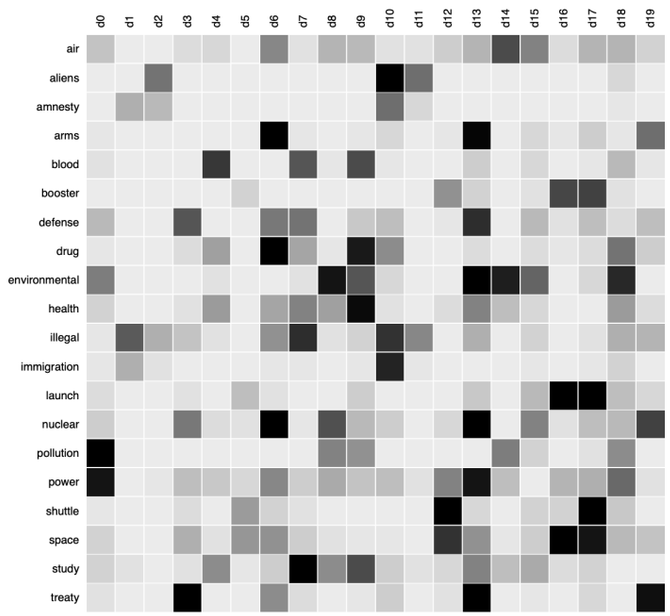
**Topic 3**
**"space travel"**

**Word x Document Matrix - Sort by rows (words occurring in similar docs)**

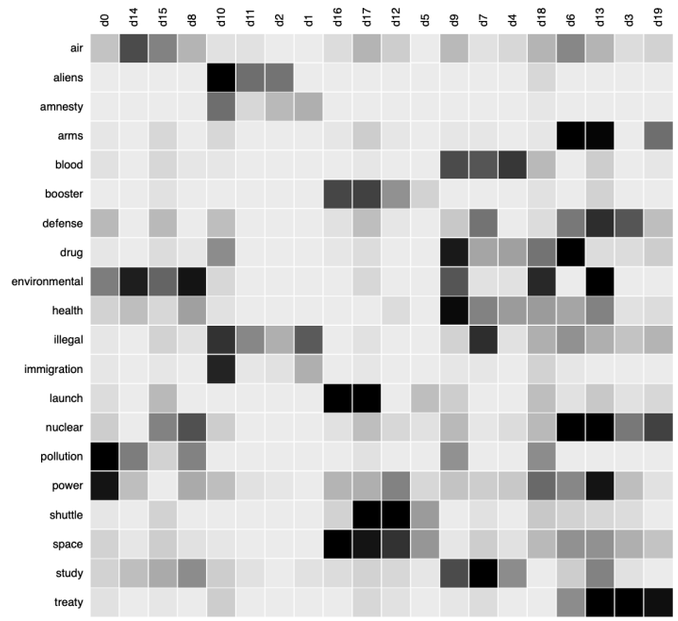
*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)

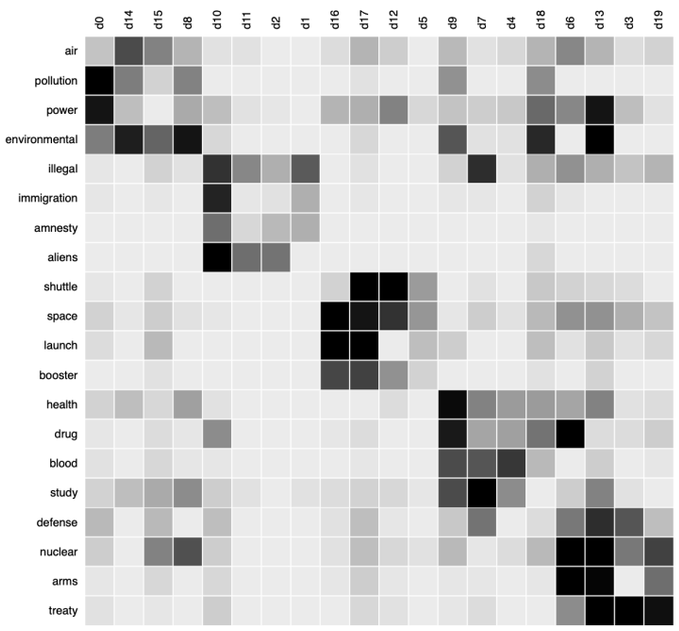
## How to sort / cluster?




**Word x Document Matrix**




**Sort by columns (docs with similar words)**

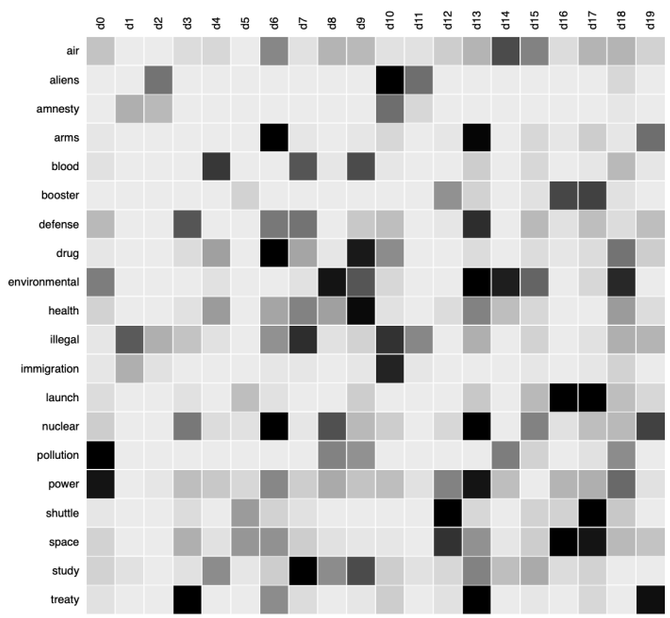



**Sort by rows (words in similar docs)**

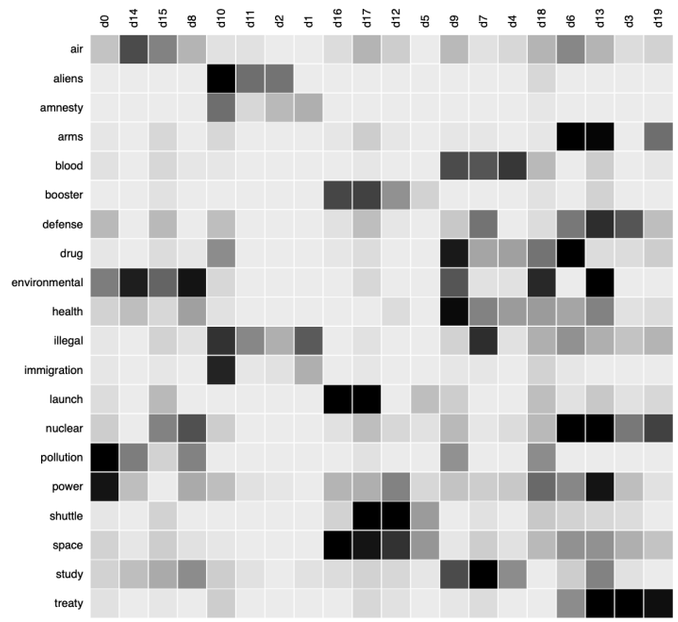
*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)

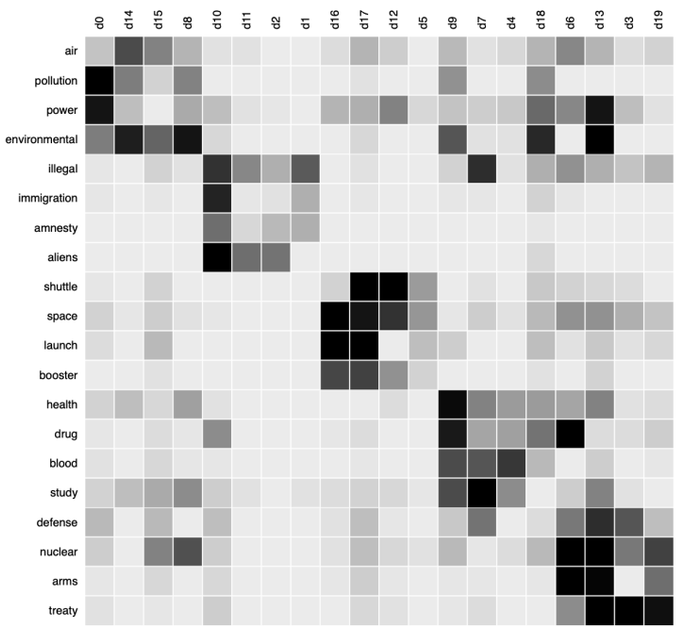
## How to sort / cluster?




**Word x Document Matrix**




**Sort by columns (docs with similar words)**




**Sort by rows (words in similar docs)**

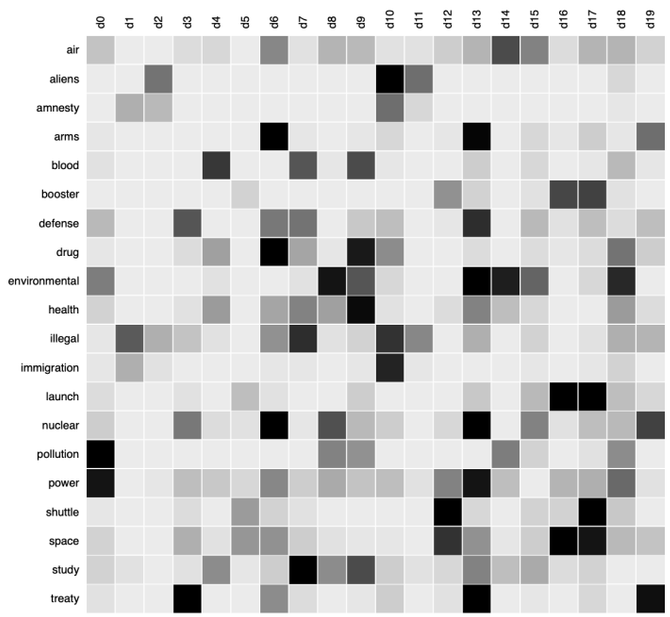
Directly use the cosine similarity of columns / rows?

*Expensive! Noisy! Overly Sparse!*

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)

## How to sort / cluster? - Use SVD




**Word x Document Matrix**

$X$

$U_k$

**Word x k**

$\Sigma_k$

**k x k**

$V_k^T$

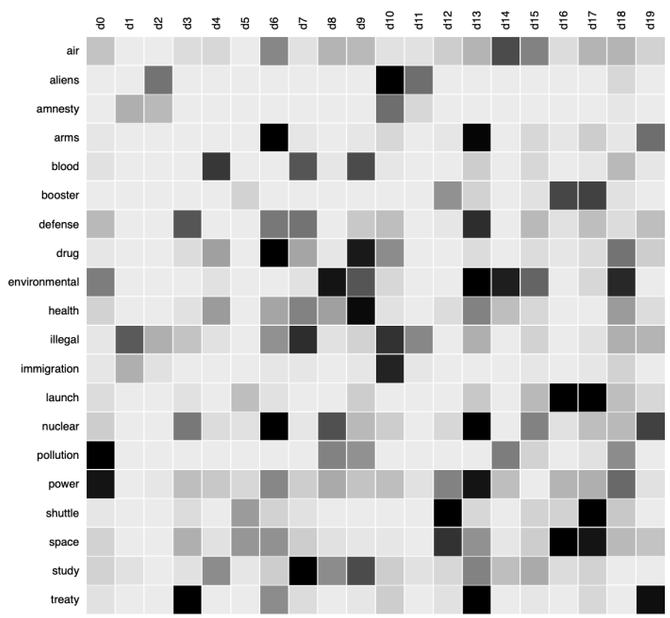
**k x Document**

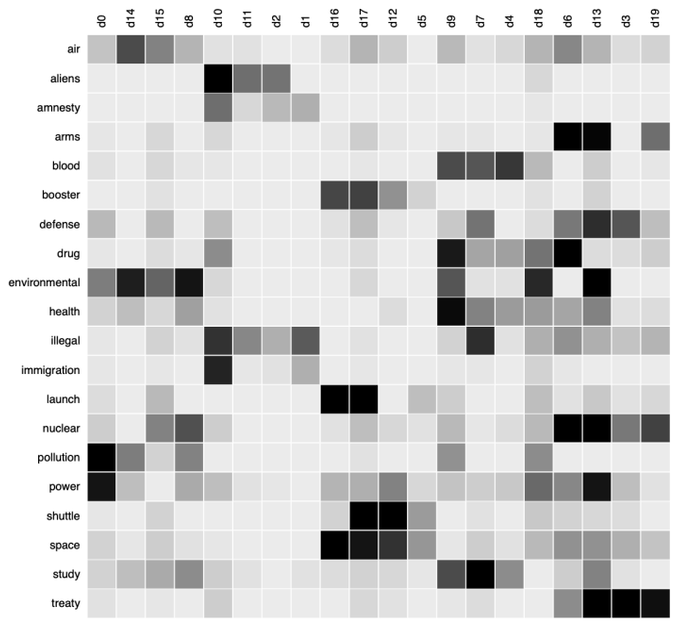
$$X \approx U_k \Sigma_k V_k^T$$

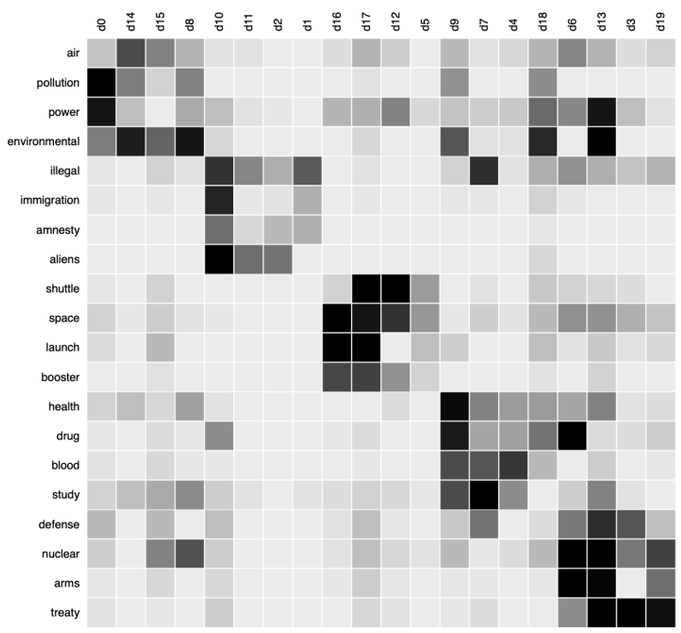
*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)

## How to sort / cluster? - Use SVD




**Word x Document Matrix**

$$X$$

$$X \approx U_k \Sigma_k V_k^T$$

**Sort by columns (docs with similar words)**

$X^T X$ **Document Correlation**

**sort**

$V_k^T$

**k x Document**

**Sort by rows (words in similar docs)**

$X X^T$ **Word Correlation**

**sort**

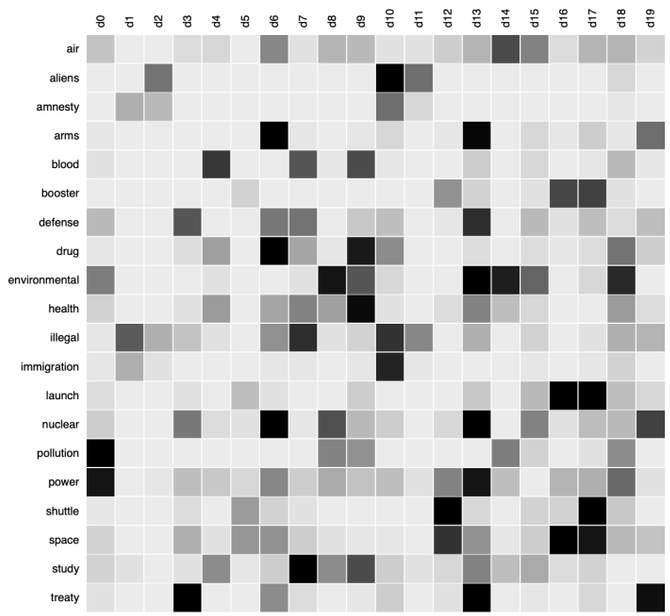
$U_k$

**Word x k**

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)

## How to sort / cluster? - Use SVD




**Word x Document Matrix**

$X$

$U_k$

**Word x k**

$\Sigma_k$

**k x k**

$V_k^T$

**k x Document**

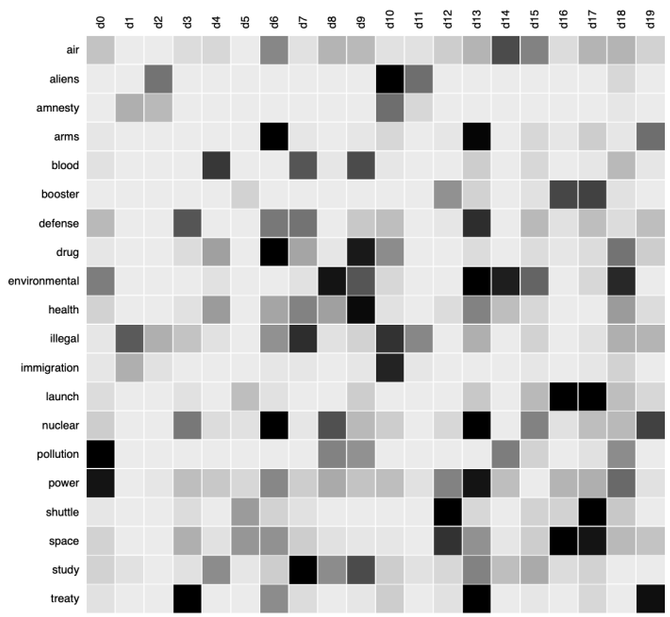
$$X \approx U_k \Sigma_k V_k^T$$

**Each column is a topic!**

*Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Latent Semantic Indexing (LSI)

## How to sort / cluster? - Use SVD




**Word x Document Matrix**

$X$

$U_k$

**Word x k**

**k x k**

$V_k^T$

**k x Document**

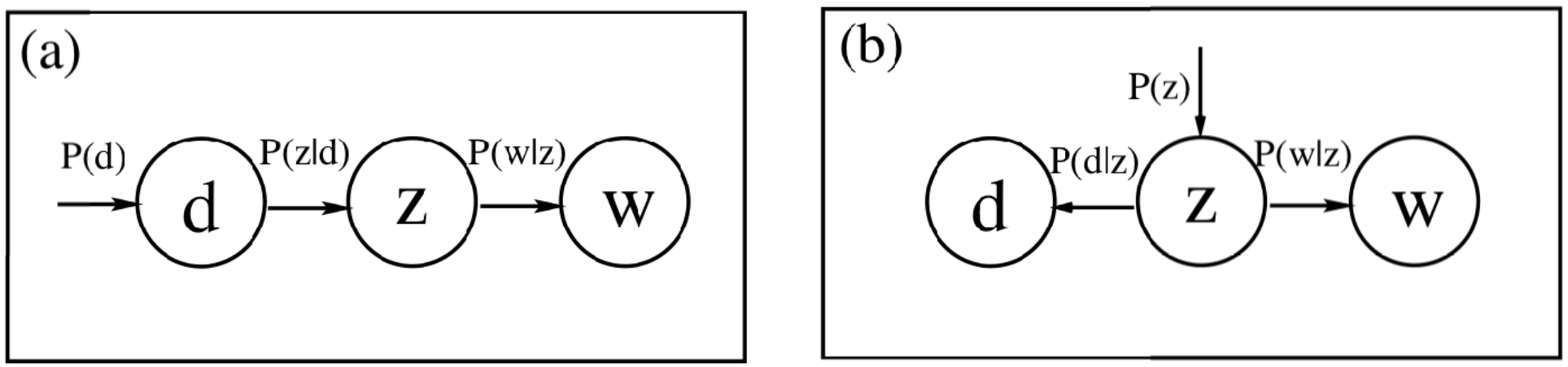
$$X \approx U_k \Sigma_k V_k^T$$

**Each column is a topic!**

**- elements of U and V can be negative, how to interpret?**

*S. Deerwester et al., Indexing by Latent Semantic Analysis, 1990*

# Probabilistic Latent Semantic Indexing (pLSI)



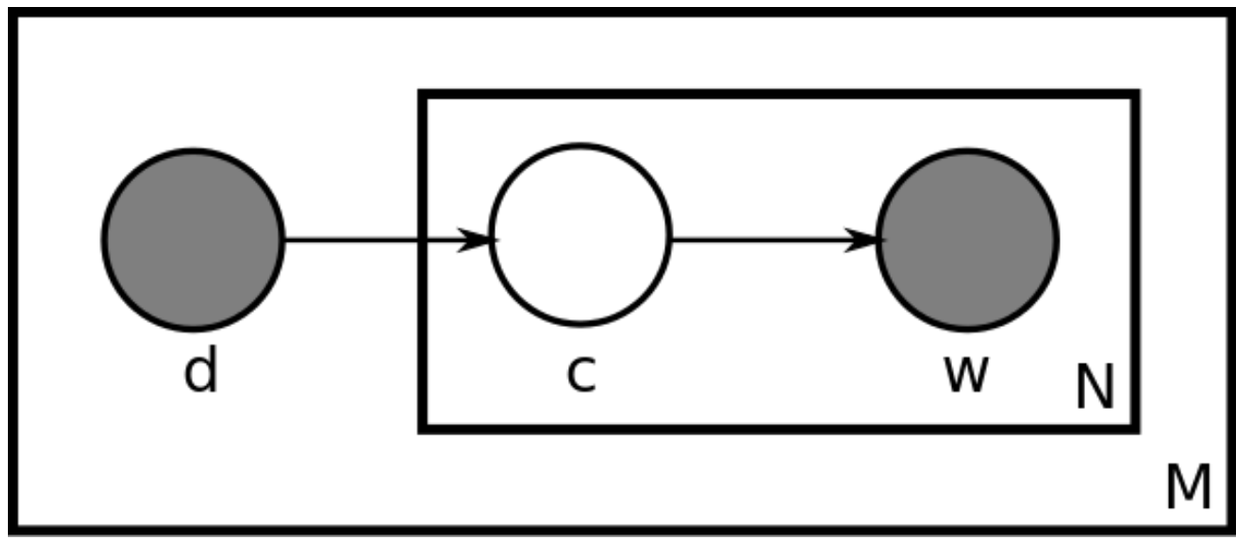

Asymmetric Symmetric

**two equivalent probabilistic views**

**d -** document
**c -** topic
**w -** word

*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

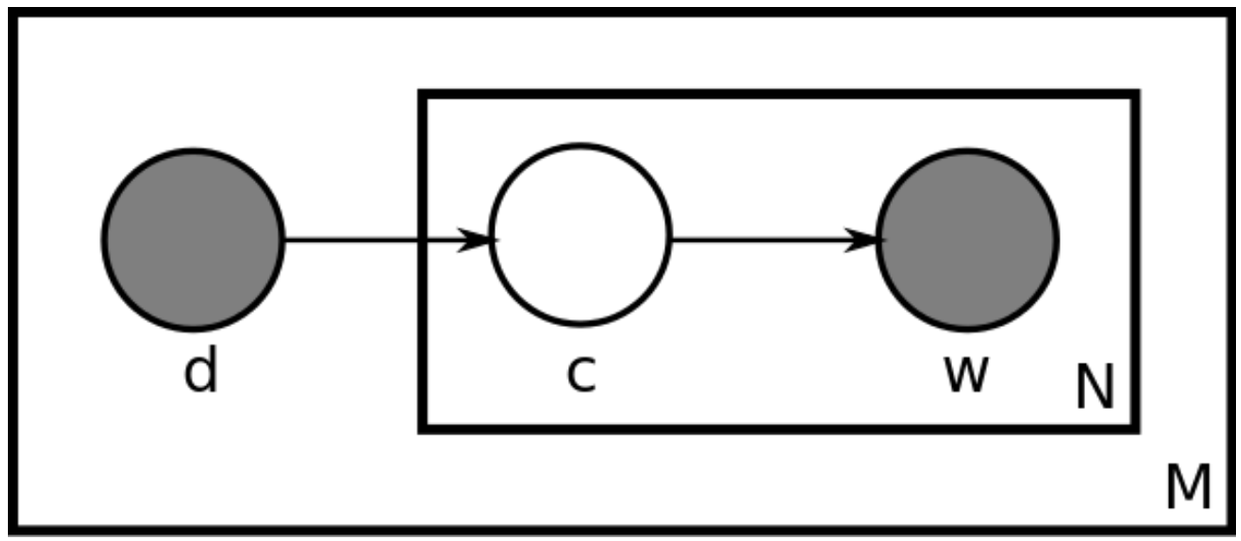
# Probabilistic Latent Semantic Indexing (pLSI)




$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

# Probabilistic Latent Semantic Indexing (pLSI)

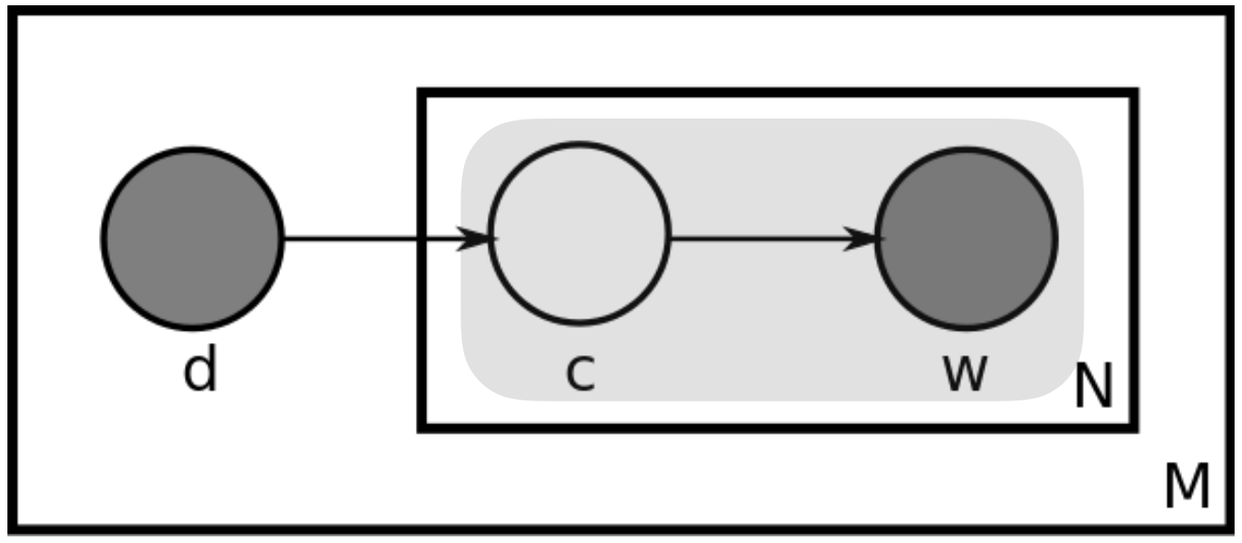



$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

Symmetric Asymmetric

*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

# Probabilistic Latent Semantic Indexing (pLSI)

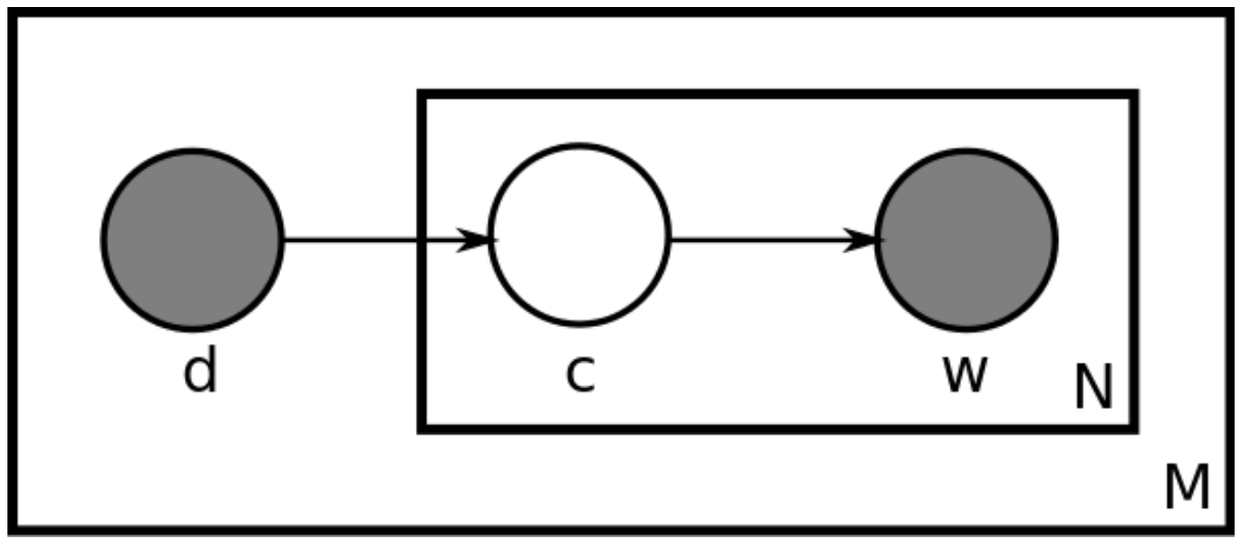



$$P(w,d) = \sum_c P(c)P(d|c)P(w|c) = P(d)\sum_c P(c|d)P(w|c)$$

Topics:


*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

# Probabilistic Latent Semantic Indexing (pLSI)




**Maximum likelihood estimate of the topic c**

E-Step: $$P(c|d,w) = \frac{P(d)P(c|d)P(w|c)}{P(d)\sum_{c' \in C} P(w|c')P(c'|d)} = \frac{P(c|d)P(w|c)}{\sum_{c' \in C} P(w|c')P(c'|d)}$$

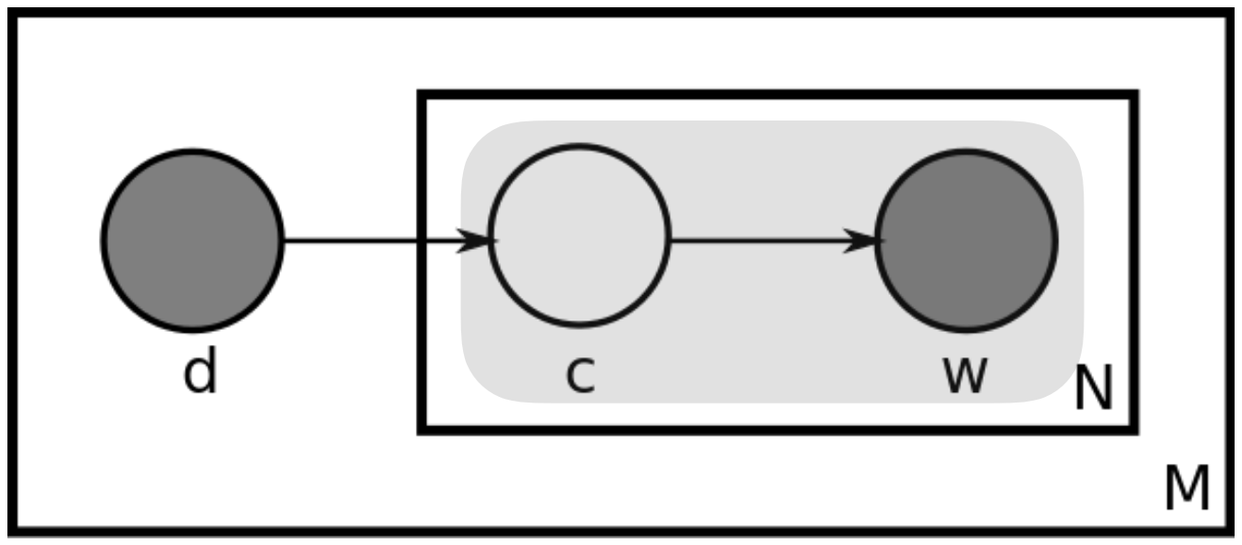
M-Step:

$$p(w|c) \propto \sum_{d \in \mathcal{D}} n(d,w)P(c|d,w) \qquad p(c) \propto \sum_{d \in \mathcal{D}} \sum_{w \in \mathcal{W}} n(d,w)P(c|d,w)$$

$$p(d|c) \propto \sum_{w \in \mathcal{W}} n(d,w)P(c|d,w) \qquad p(d|c) \propto p(d|c)p(c)$$

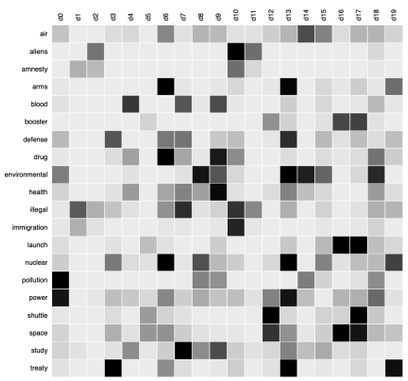
*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

# Probabilistic Latent Semantic Indexing (pLSI)




$$P(w|d) = \sum_c p(w|c)p(c|d)$$

$P(w|d)$




**Word x Document Frequency Matrix**

$\approx$

$P(w|c)$

**Word x k**

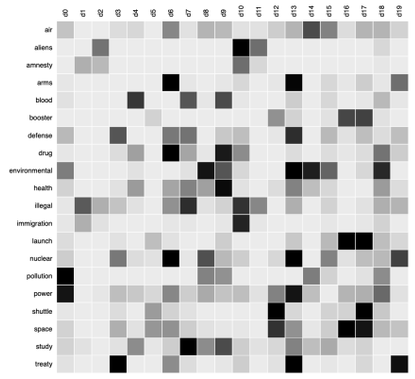
$P(c|d)$.

**k x Document**

*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

# Nonnegative Matrix Factorization (NMF)

$P(w|d)$




**Word x Document Frequency Matrix**

$\approx$

$P(w|c)$

**Word x k**

$P(c|d)$.

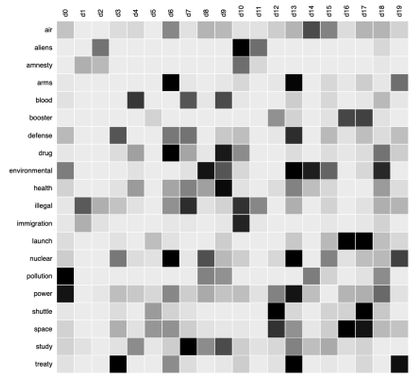
**k x Document**

$V$ $W \geq 0$ $H \geq 0$

*Lee & Seung, Learning the parts of objects by non-negative matrix factorization, Nature 1999*

# Nonnegative Matrix Factorization (NMF)

$P(w|d)$




**Word x Document Frequency Matrix**

$\approx$

$P(w|c)$

**Word x k**

$P(c|d)$.

**k x Document**

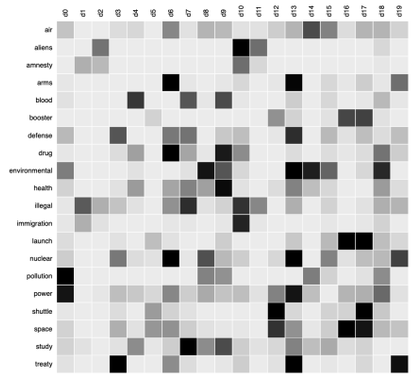
Given $V$, find $W \geq 0$ $H \geq 0$

to minimize $\|H - WH\|_F$ or $D(H||WH)$

*Lee & Seung, Learning the parts of objects by non-negative matrix factorization, Nature 1999*

# Nonnegative Matrix Factorization (NMF)

$P(w|d)$




**Word x Document Frequency Matrix**

$\approx$

$P(w|c)$

**Word x k**

$P(c|d)$.

**k x Document**

Given $V$, find $W \geq 0$ $H \geq 0$

to minimize $\|H - WH\|_F$ or $D(H||WH)$

**Similar to EM:**

$$W_{ia} \leftarrow W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}$$

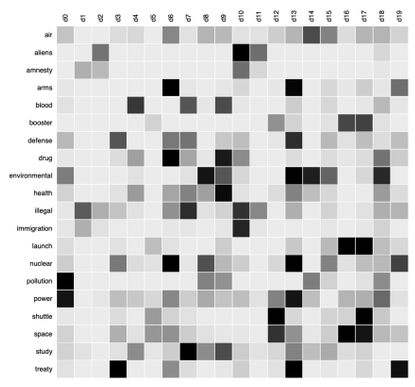
$$W_{ia} \leftarrow \frac{W_{ia}}{\sum_j W_{ja}}$$

$$H_{a\mu} \leftarrow H_{a\mu} \sum_i W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}$$

*Lee & Seung, Learning the parts of objects by non-negative matrix factorization, Nature 1999*

# Nonnegative Matrix Factorization (NMF)

$P(w|d)$




**Word x Document Frequency Matrix**

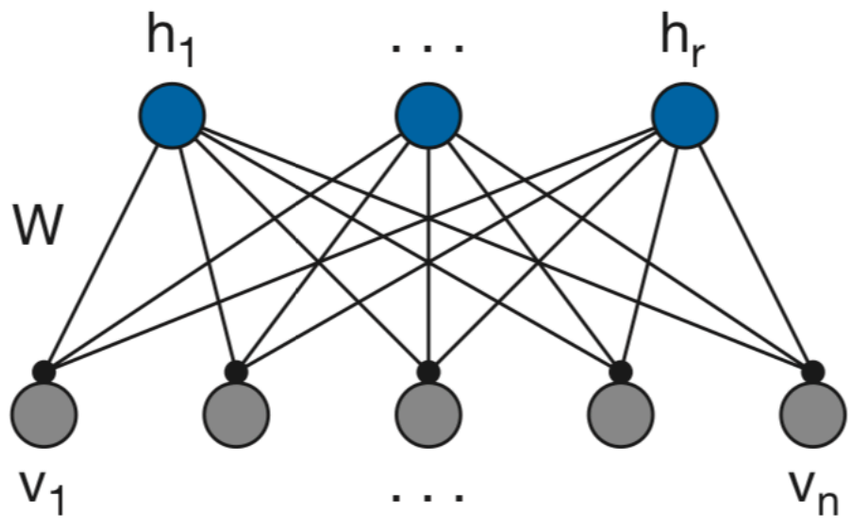
$\approx$

$P(w|c)$

**Word x k**

$P(c|d)$.

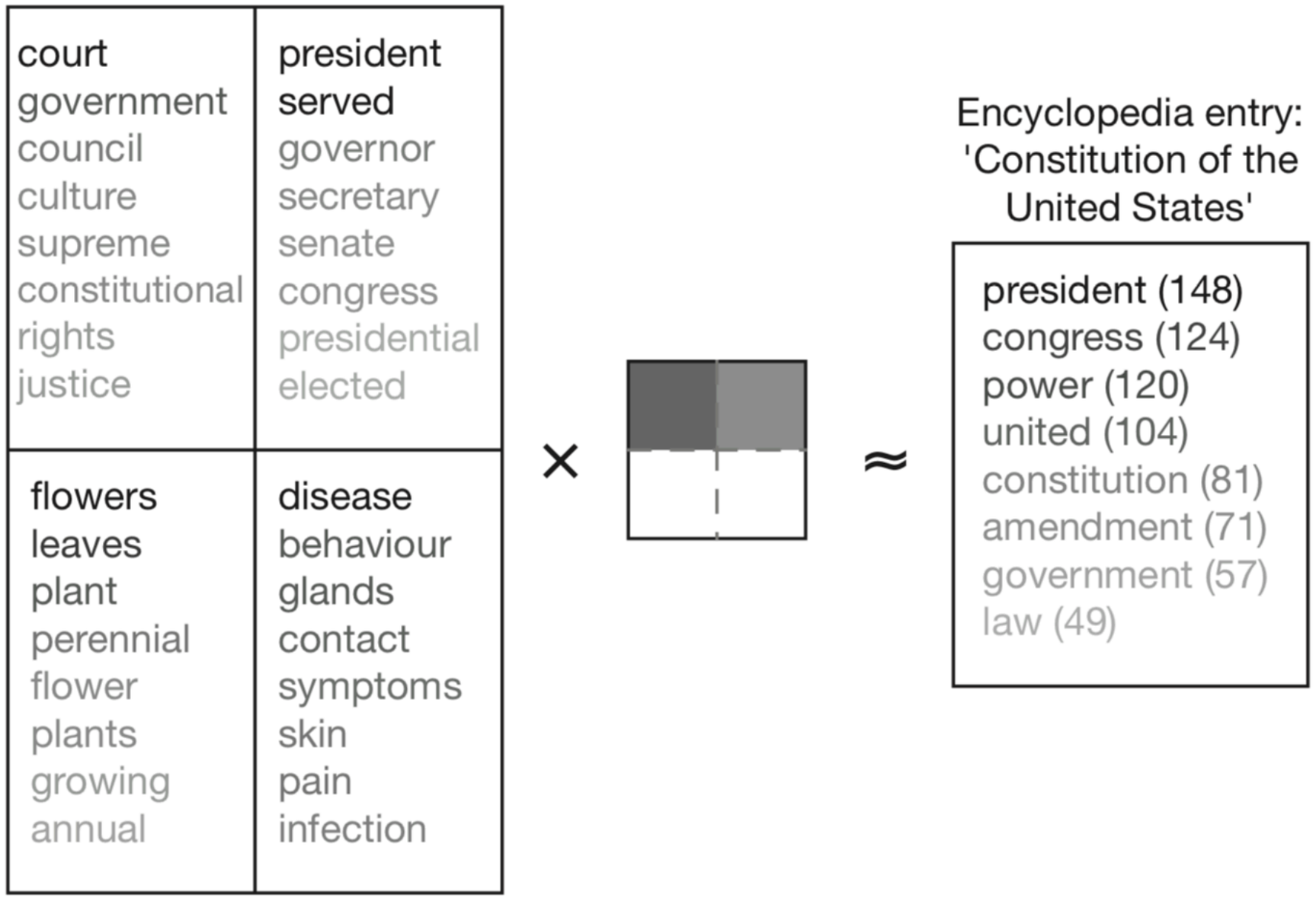
**k x Document**




**Probabilistic hidden variables model underlying non-negative matrix factorization.**
The bottom layer of nodes are generated from the hidden variables.
The influence of $h_a$ on $v_i$ is represented by a connection with strength $W_{ia}$.

*Lee & Seung, Learning the parts of objects by non-negative matrix factorization, Nature 1999*

# Nonnegative Matrix Factorization (NMF)




**Non-negative matrix factorization (NMF) discovers semantic features of 30,991 articles from the Grolier encyclopedia.**

*Lee & Seung, Learning the parts of objects by non-negative matrix factorization, Nature 1999*

# Nonnegative Matrix Factorization (NMF)

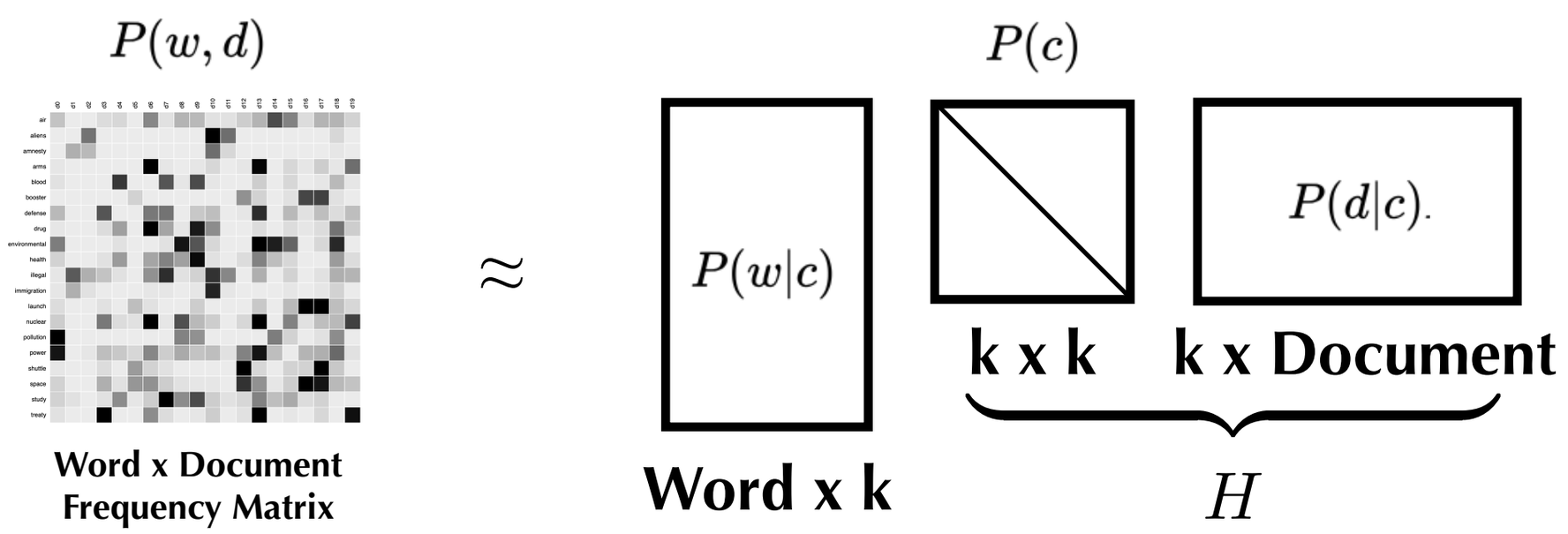
## Equivalence to pLSI — H is non-normalized?

**Proposition 1**. The objective function of PLSI is identical to the objective function of NMF, i.e.,

$$\max J_{\text{PLSI}} \quad \Longleftrightarrow \quad \min J_{\text{NMF}} \tag{6}$$

**Proposition 2**. Column normalized NMF of Eq.(1) is equivalent to the probability factorization of Eq.(4), i.e., $(CH^T)_{ij} = P(w_i, d_j)$.

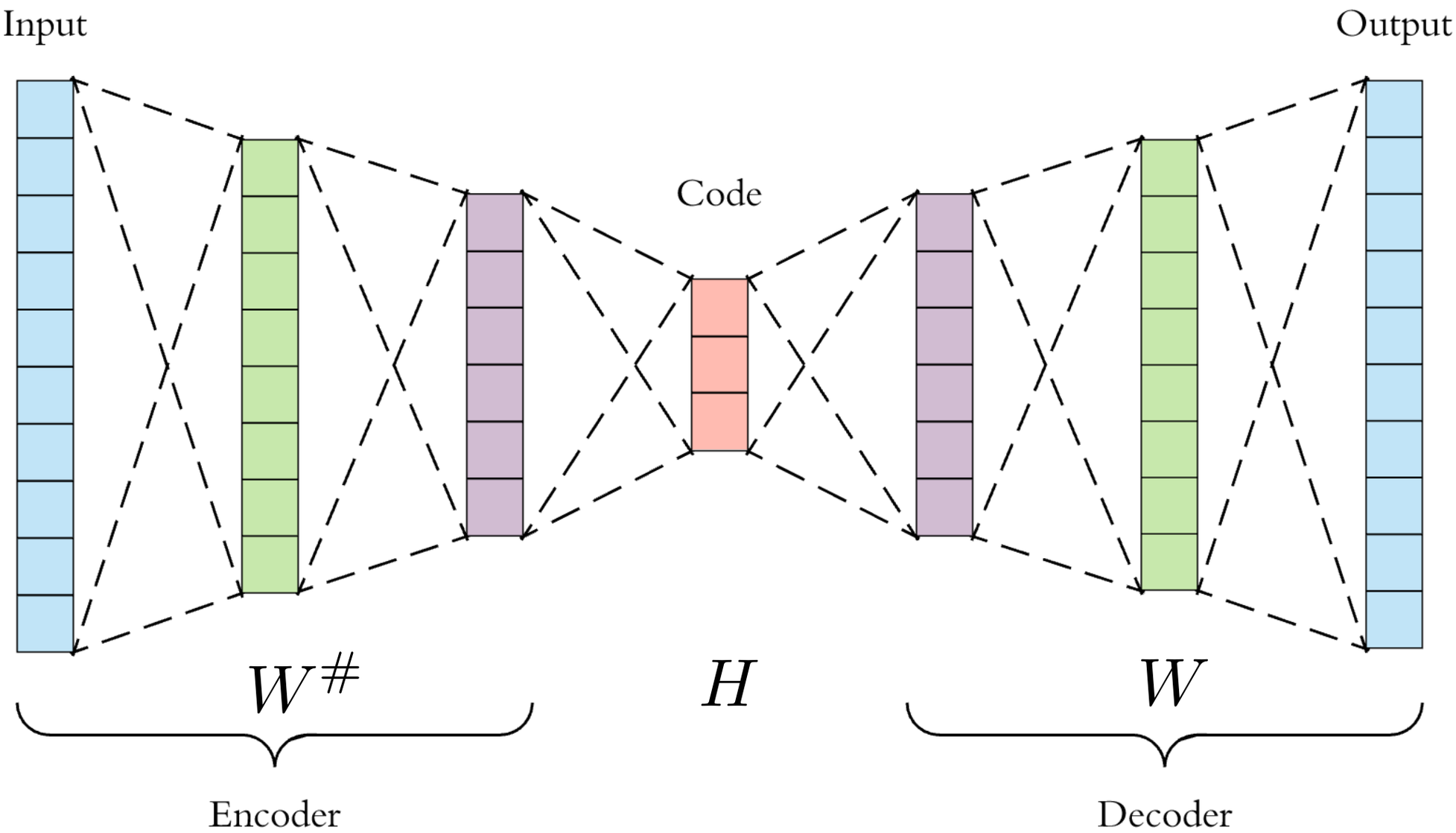
$$P(w,d) = \sum_c P(c)P(d|c)P(w|c)$$




*C. Ding et al, On the Equivalence Between Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing, AAAI 2006*

# Autoencoder (AE)

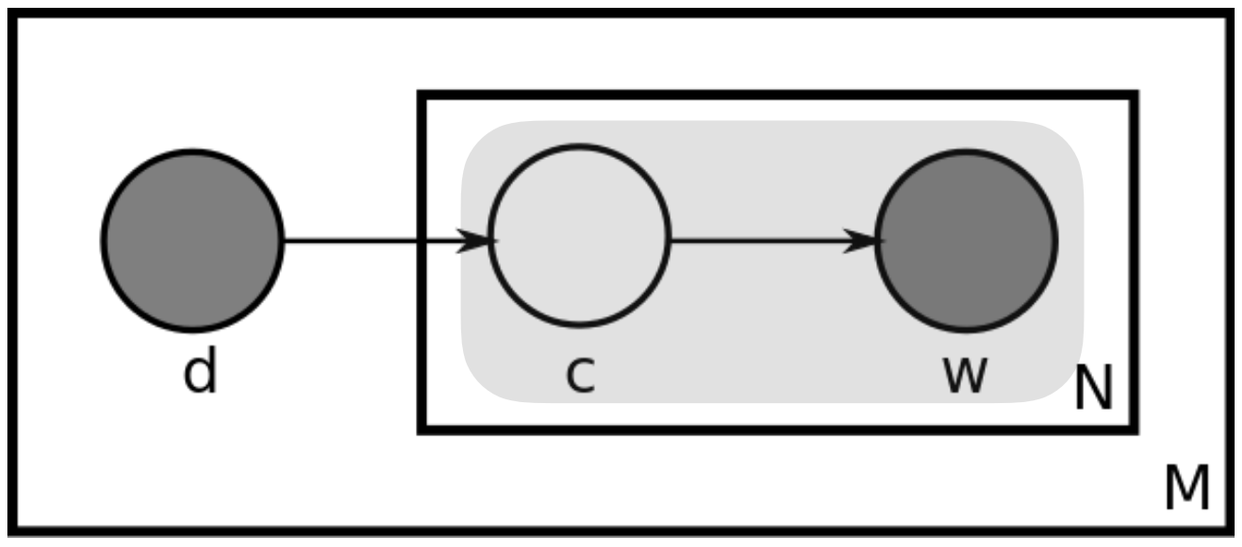
$$V \approx WH$$



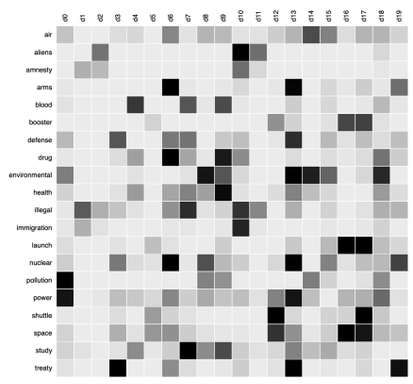

$$W^{\#}V = H \qquad WH \approx V$$

*Smaragdis & Venkataramani, A Neural Network Alternative to Non-Negative Audio Models, ICASSP 2017*

# Probabilistic Latent Semantic Indexing (pLSI)




$P(w,d)$




**Word x Document Frequency Matrix**

$\approx$

$P(w|c)$

**Word x k**

$P(c)$

**k x k**

$P(d|c).$

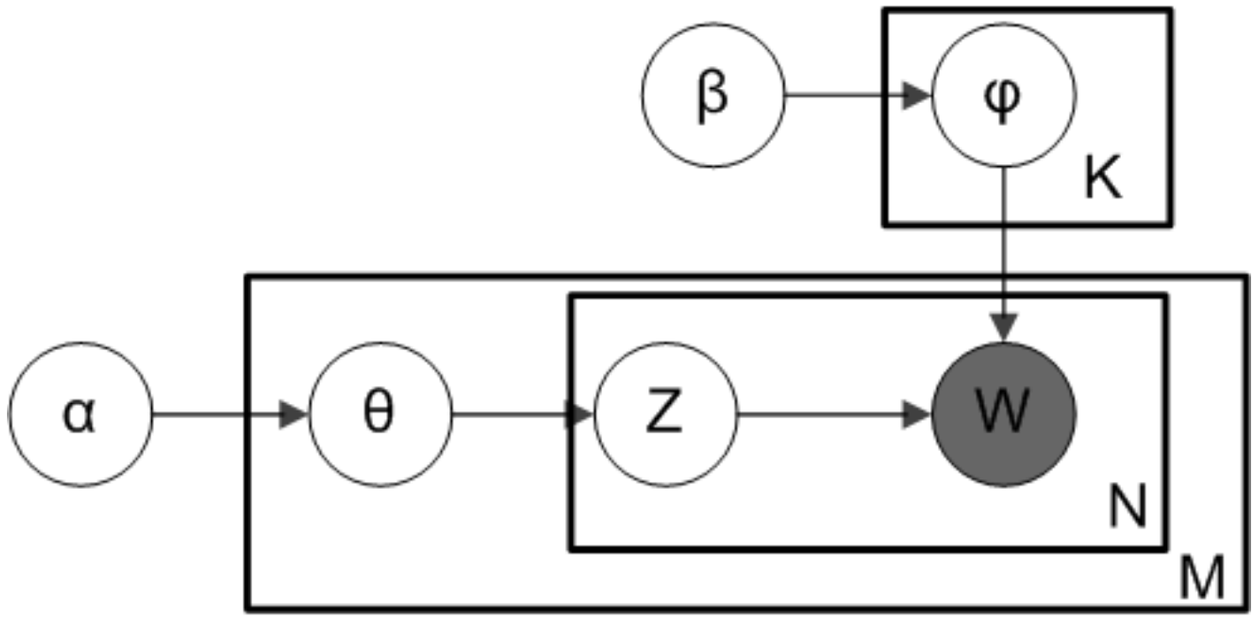
**k x Document**

$H$

**- MLE, no control on topic / word sparsity?**

*T. Hofmann, Probabilistic latent semantic indexing, SIGIR 1999*

# Latent Dirichlet Analysis (LDA)

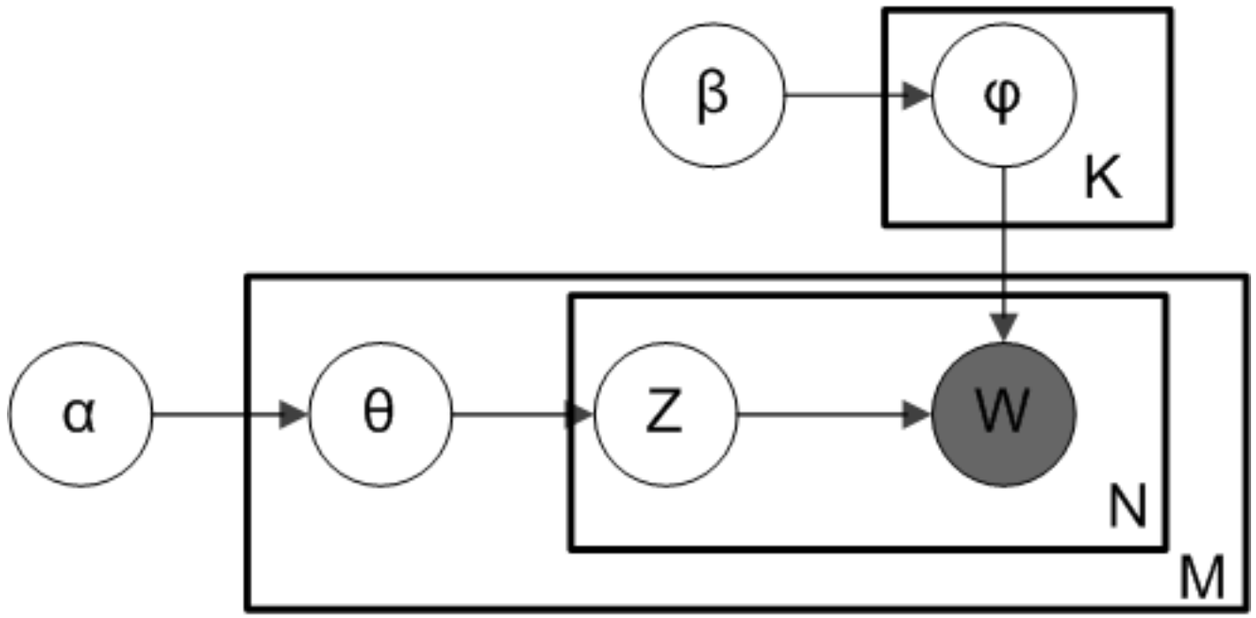



**z:** *topics,* **w:** *words*
**θ:** *topic distributions for documents*
**φ:** word distributions for topics

**α**: the parameter of the Dirichlet prior
on the per-document topic distributions
**ß:** *the parameter of the Dirichlet prior*
*on the per-topic word distribution*

*D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)*

# Latent Dirichlet Analysis (LDA)




## Generative process:

$$
\begin{aligned}
\boldsymbol{\varphi}_{k=1\ldots K} &\sim \text{Dirichlet}_V(\boldsymbol{\beta}) \\
\boldsymbol{\theta}_{d=1\ldots M} &\sim \text{Dirichlet}_K(\boldsymbol{\alpha}) \\
z_{d=1\ldots M, w=1\ldots N_d} &\sim \text{Categorical}_K(\boldsymbol{\theta}_d) \\
w_{d=1\ldots M, w=1\ldots N_d} &\sim \text{Categorical}_V(\boldsymbol{\varphi}_{z_{dw}})
\end{aligned}
$$

*D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)*

# Latent Dirichlet Analysis (LDA)

**α, ß control doc-topic, topic-word sparsity**

$$P(\boldsymbol{\theta}_j|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_{j,i}^{\alpha_i-1}$$

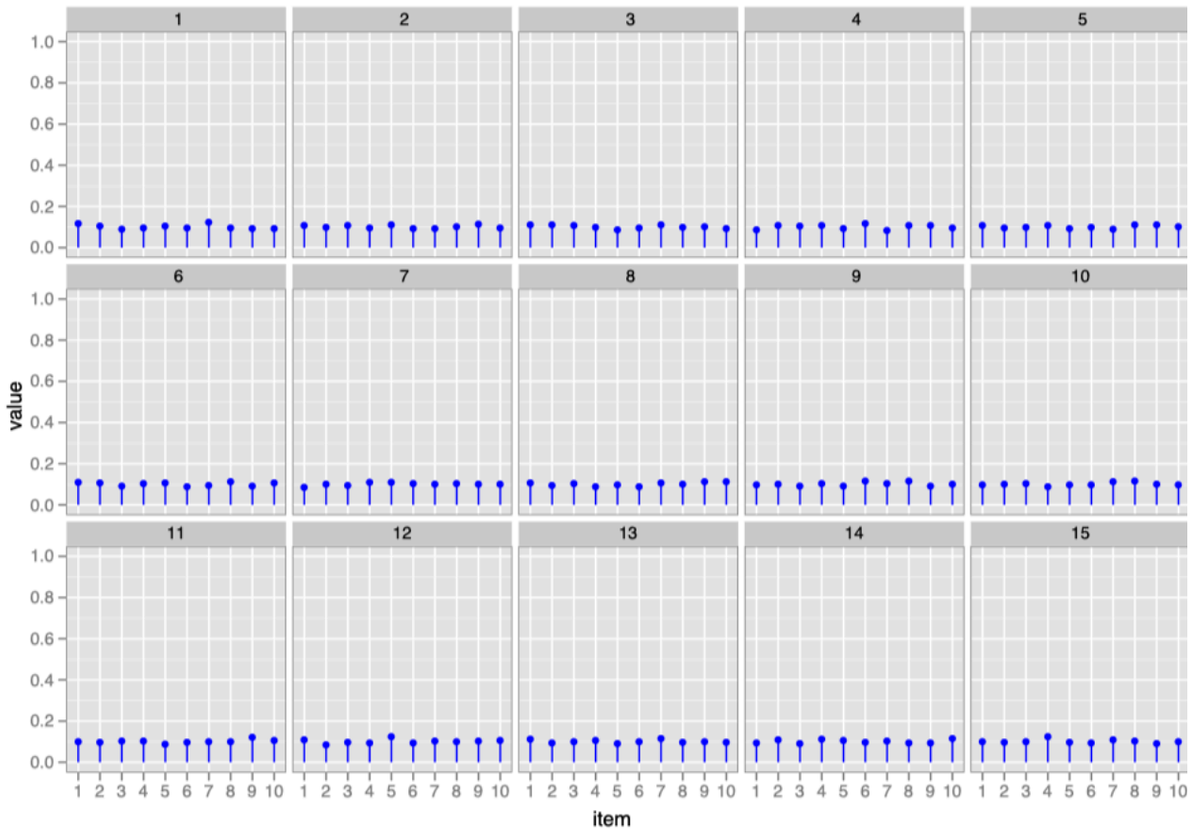
It is **conjugate** to the **multinomial**. Given a multinomial observation, the posterior distribution of θ is a **Dirichlet**.

*D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)*

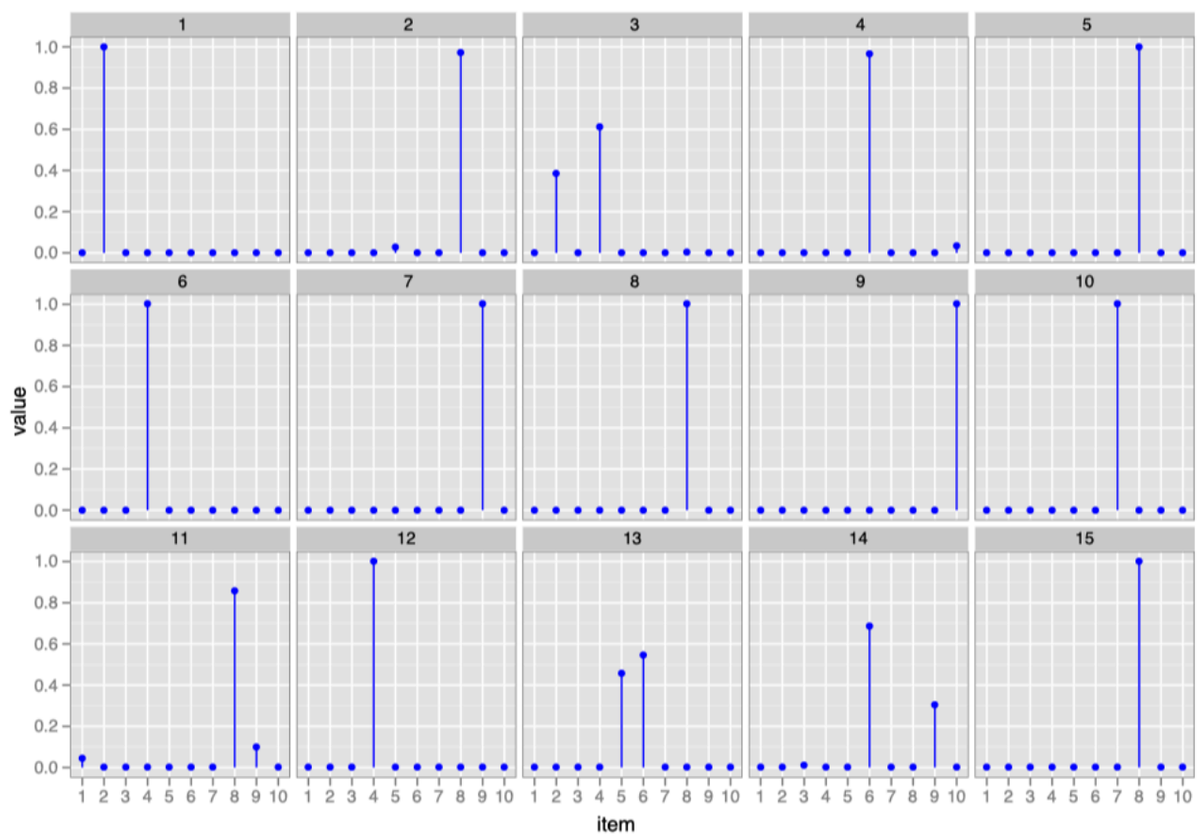
# Latent Dirichlet Analysis (LDA)

**α, ß controls doc-topic, topic-word sparsity**

$$P(\theta_j|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma(\alpha_i)}\prod_{i=1}^{K}\theta_{j,i}^{\alpha_i-1}$$



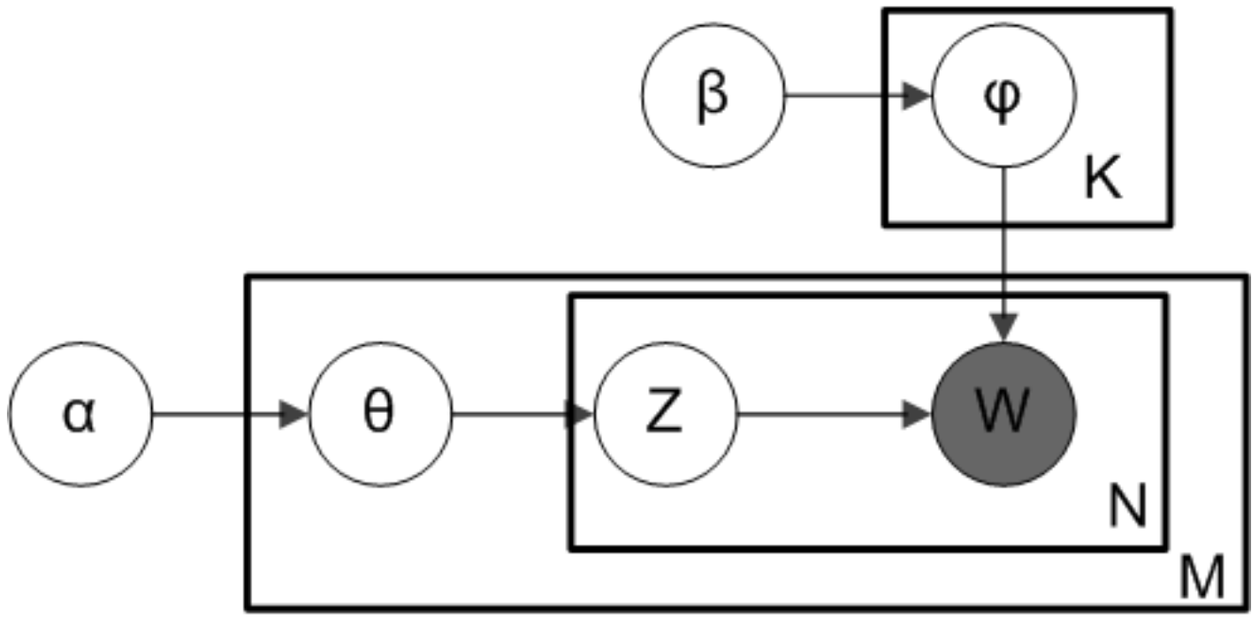

$\alpha = 100$




$\alpha = 0.01$

*D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)*

# Latent Dirichlet Analysis (LDA)



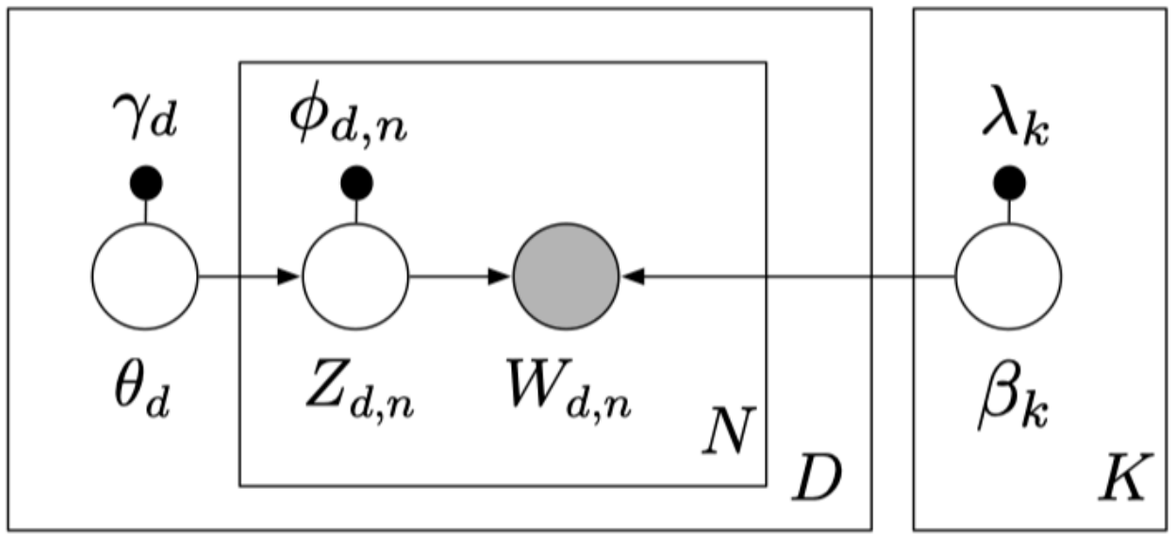

**Posterior inference for LDA:** $P(\varphi, \theta, z|w)$

$$\frac{P(\varphi, \theta, z, w)}{\int_{\varphi} \int_{\theta} \sum_{z} P(\varphi, \theta, z, w)}$$

Intractable

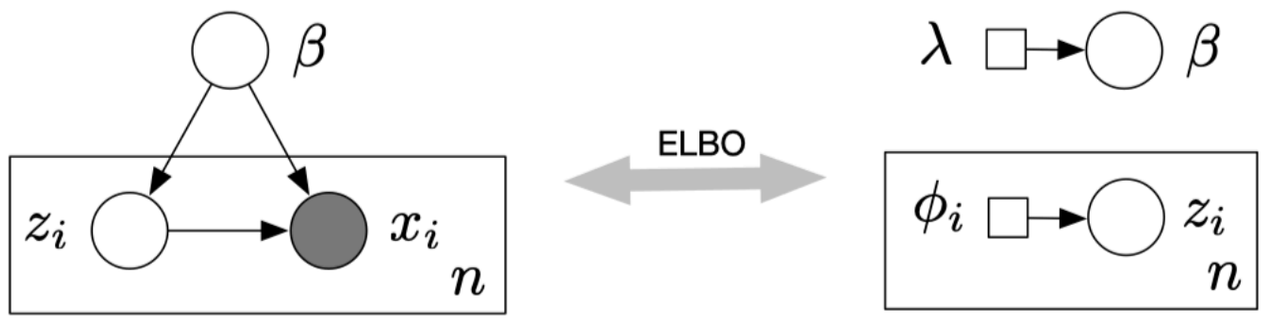
*D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)*

# Latent Dirichlet Analysis (LDA)



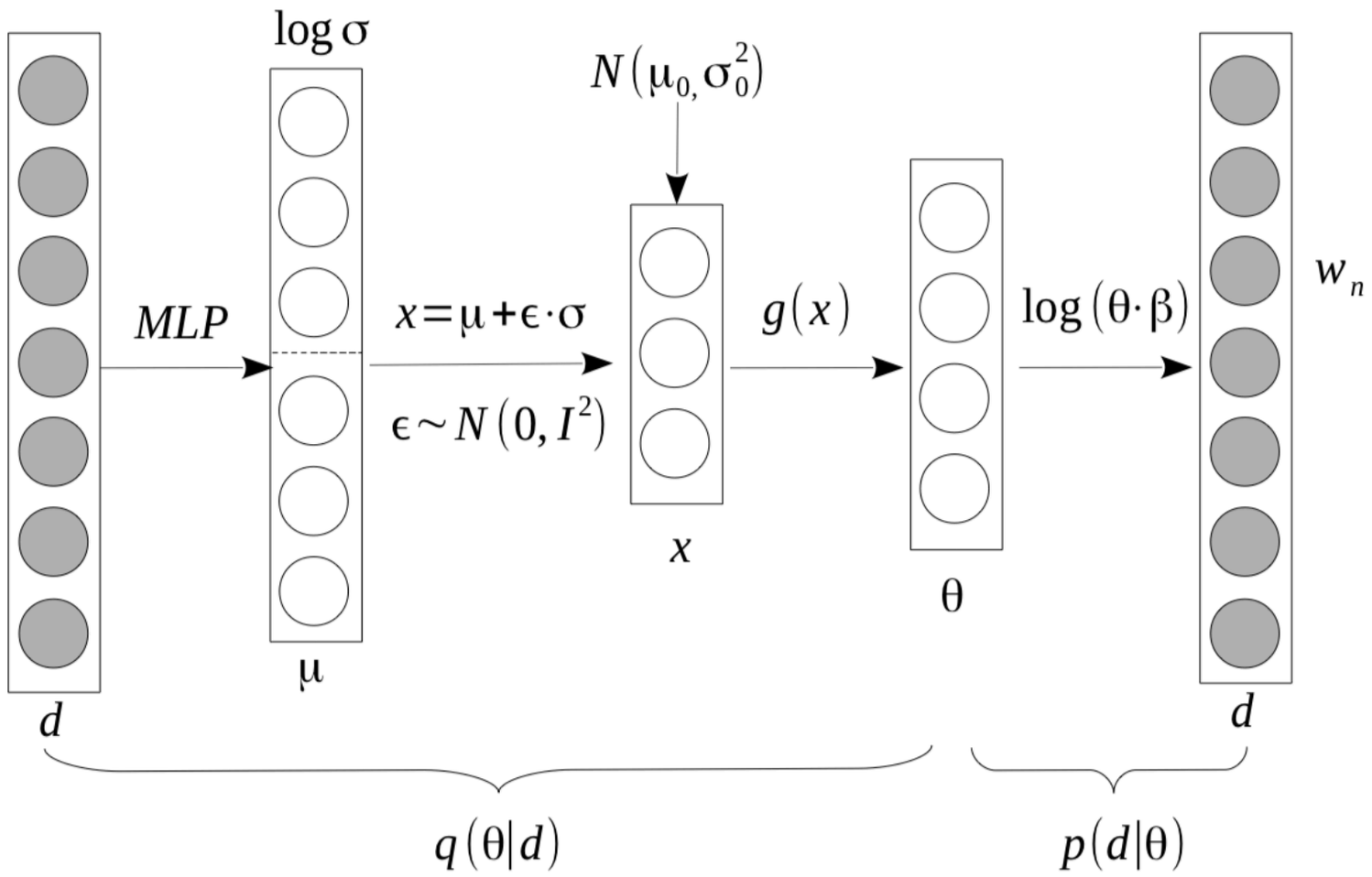

## Variational distribution $\Rightarrow$ Optimization problem

$$q(\beta,\theta,z)=\prod_{k=1}^{K} q(\beta_k|\lambda_k)\prod_{d=1}^{D} q(\theta_d|\gamma_d)\prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}) \quad \approx \quad P(\boldsymbol{\varphi},\boldsymbol{\theta},\boldsymbol{z}|\boldsymbol{w})$$




*D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet allocation. J. Mach. Learn. Res. (2003)*

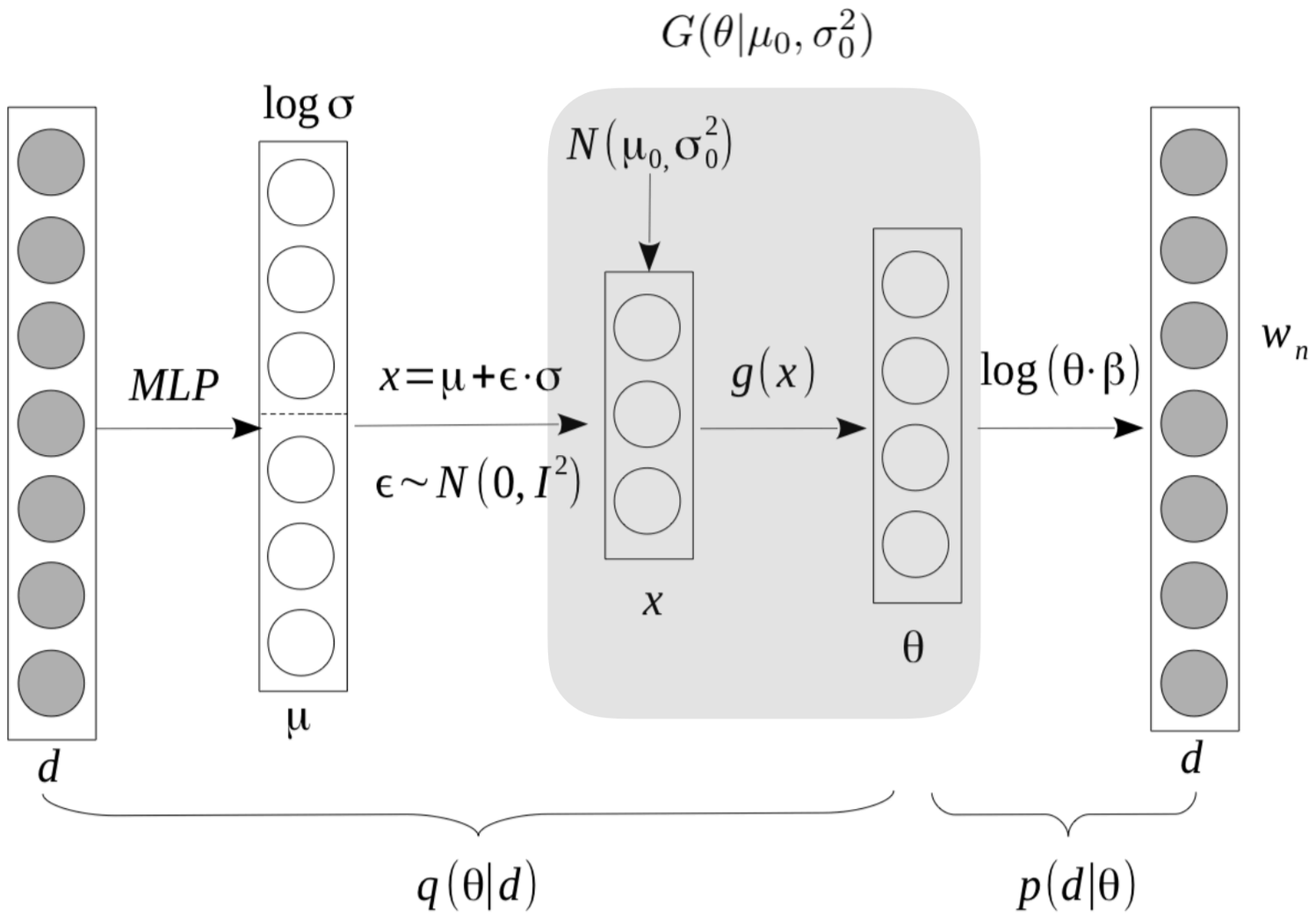
# Neural Topic Modeling (Gaussian prior)




**Network structure of the inference model q(θ | d), and of the generative model p(d | θ).**

*Y. Miao et al., Discovering Discrete Latent Topics with Neural Variational Inference. ICML 2017.*

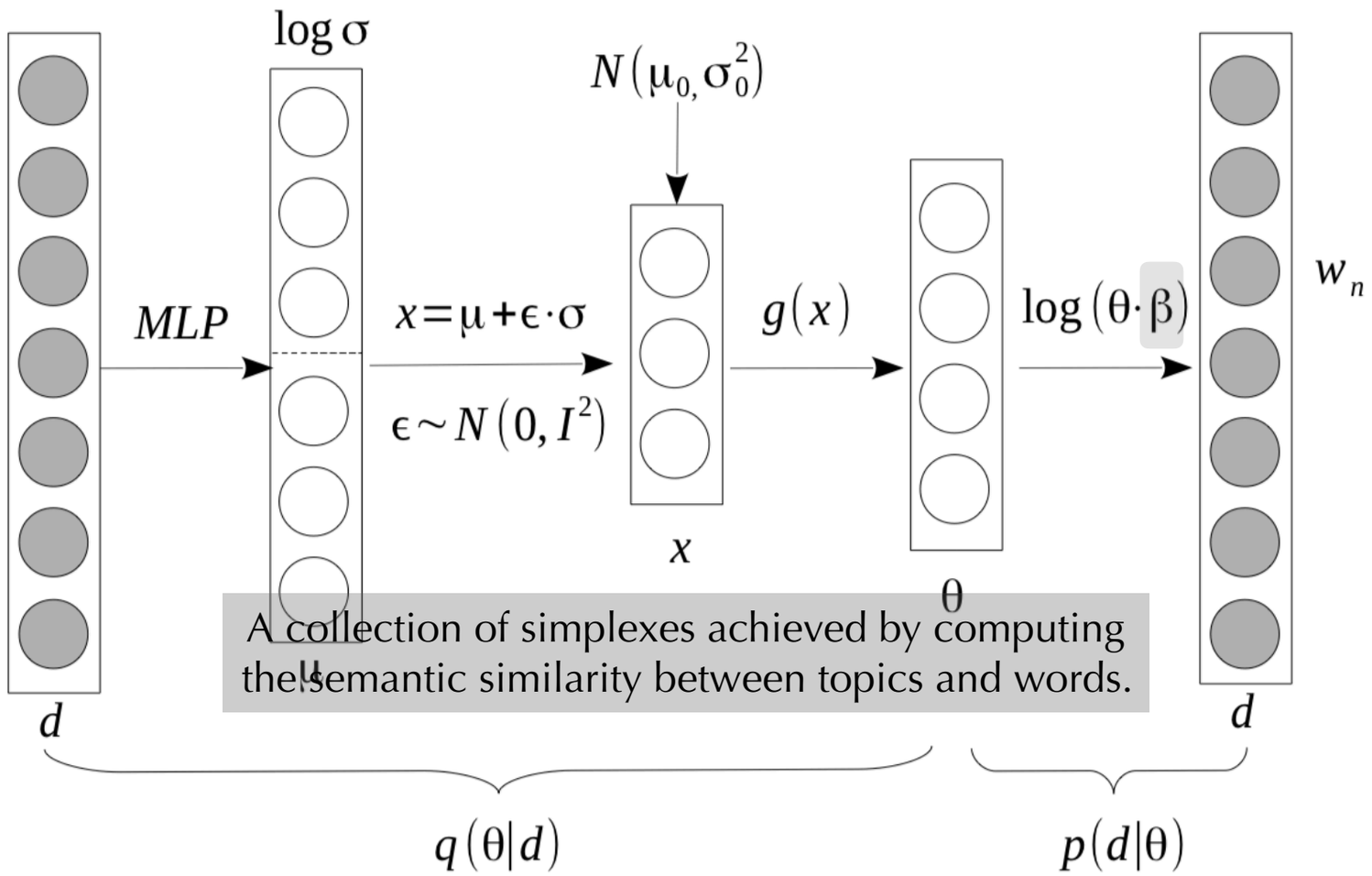
# Neural Topic Modeling (Gaussian prior)




**Network structure of the inference model q(θ | d), and of the generative model p(d | θ).**

*Y. Miao et al., Discovering Discrete Latent Topics with Neural Variational Inference. ICML 2017.*

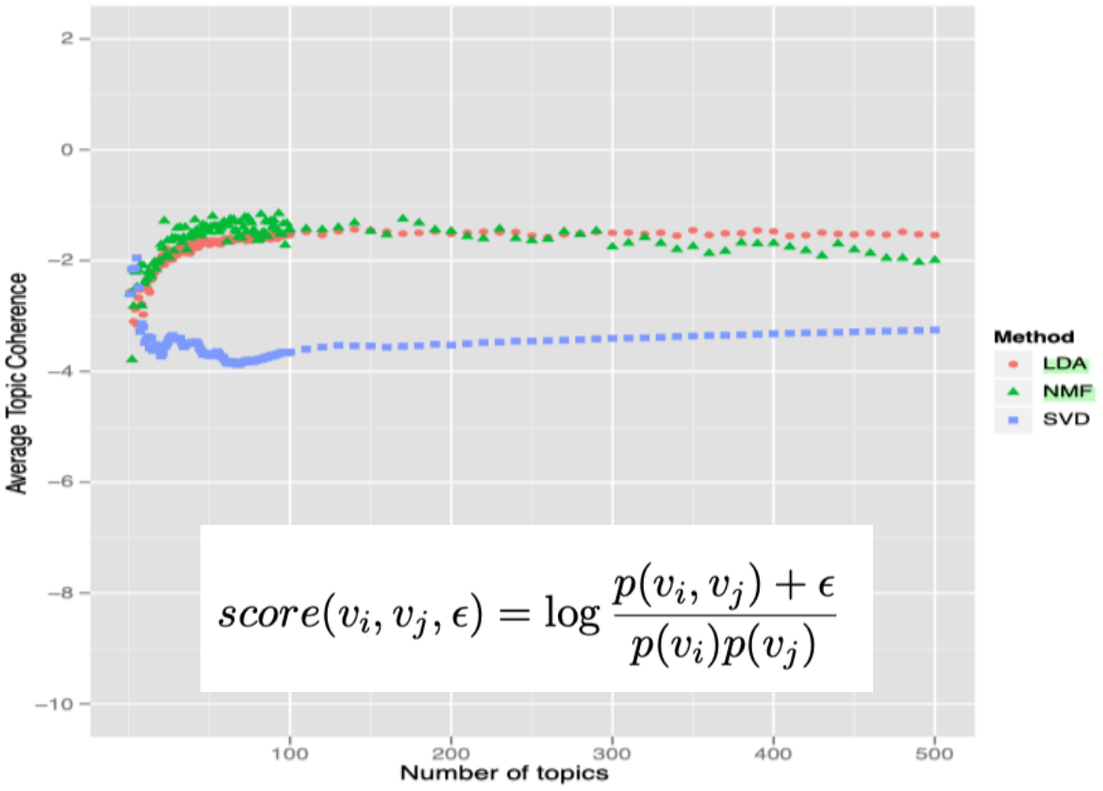
# Neural Topic Modeling (Gaussian prior)

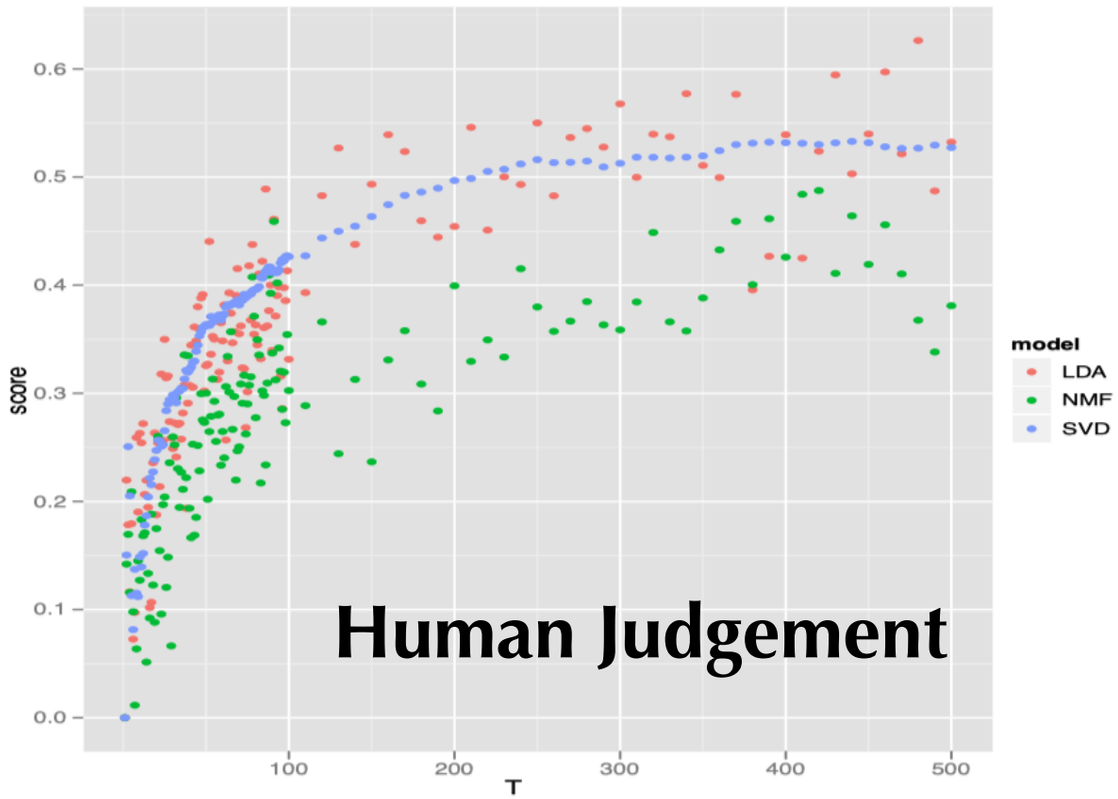



**Network structure of the inference model q(θ | d), and of the generative model p(d | θ).**

*Y. Miao et al., Discovering Discrete Latent Topics with Neural Variational Inference. ICML 2017.*

# Comparison - SVD, NMF, LDA



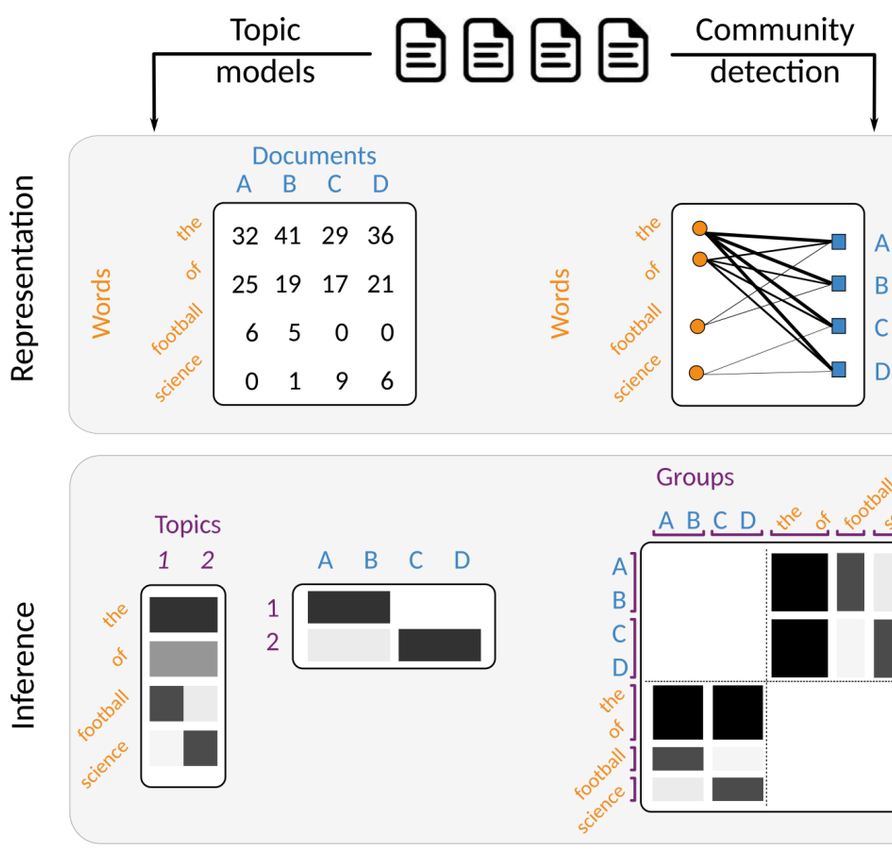

- **SVD fails to form individual topics that aggregate similar words**.
- Conversely, both NMF and LDA learn concise and coherent topics and achieved similar performance.
- **NMF learns more incoherent topics than LDA and SVD.**

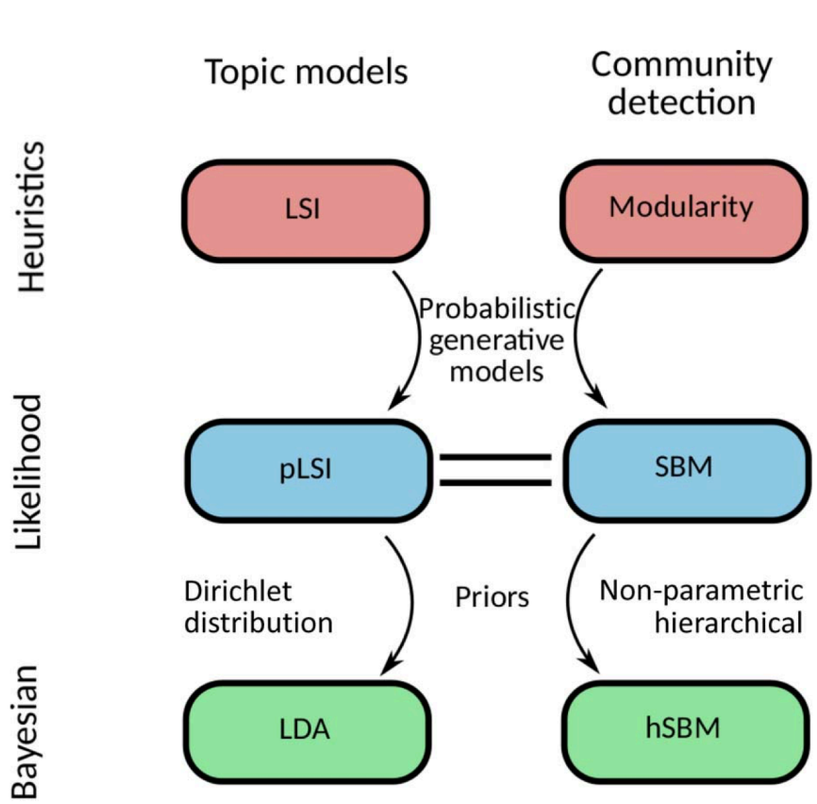
*K. Stevens et al., Exploring Topic Coherence over many models and many topics. EMNLP 2012.*

# Beyond Probabilistic Generative Models

**Main Drawbacks of LDA:**

- lacks an intrinsic methodology to choose the **number of topics** and contains a **large number of free parameters** that can cause overfitting.
- no **justification for the use of the Dirichlet prior** in the model formulation besides mathematical convenience.



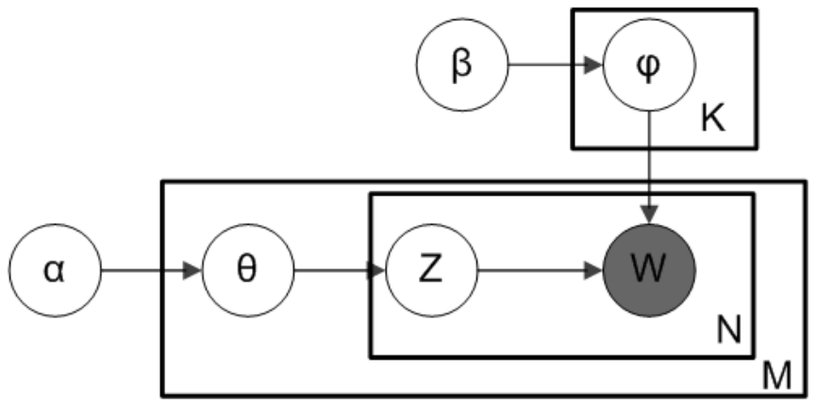



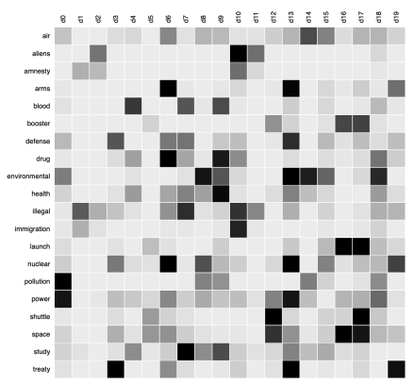

*M. Gerlach et al., A network approach to topic models. Science Advances 2018*

# Beyond Probabilistic Generative Models

**However, LDA is charming because**

- We have α, ß to control doc-topic, topic-word sparsity




**Word x Document Frequency Matrix**

≈

**Sparse**

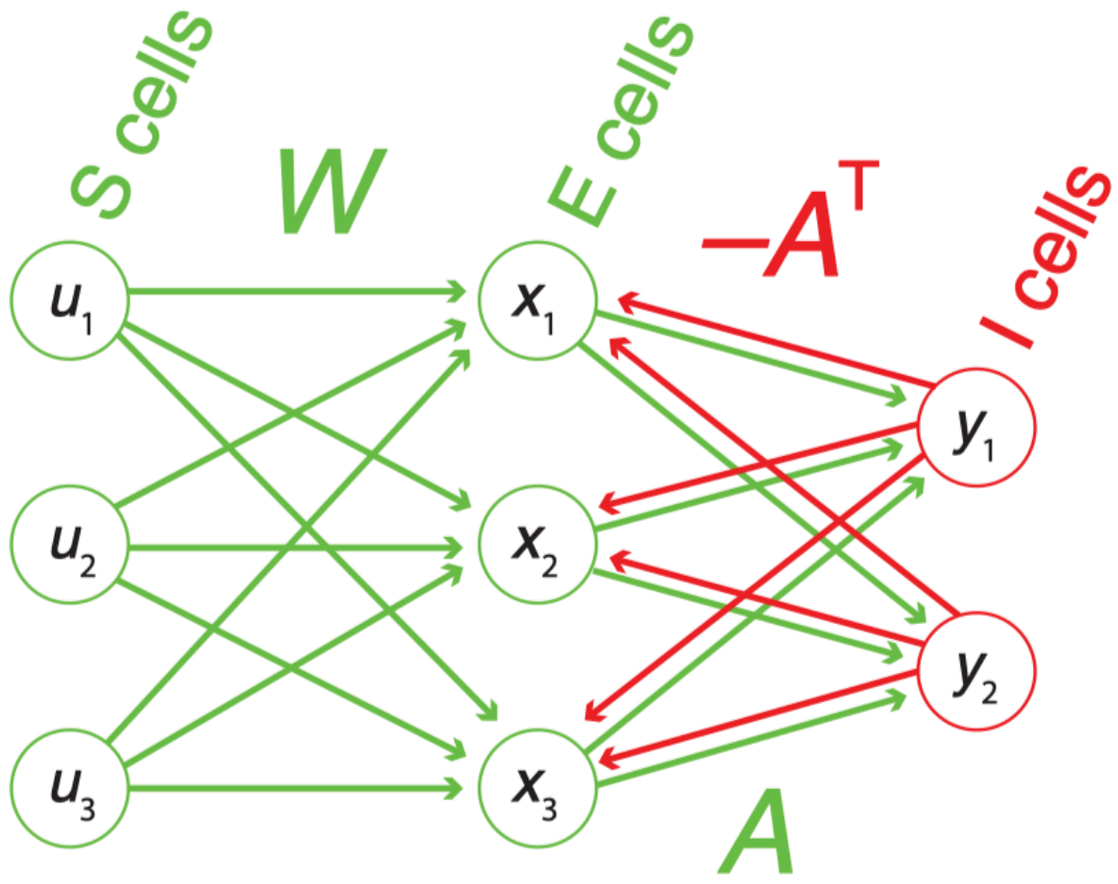
**Word x k**

**Sparse**

**k x Document**

Less sparse - more general topics
More spares - more representative topics

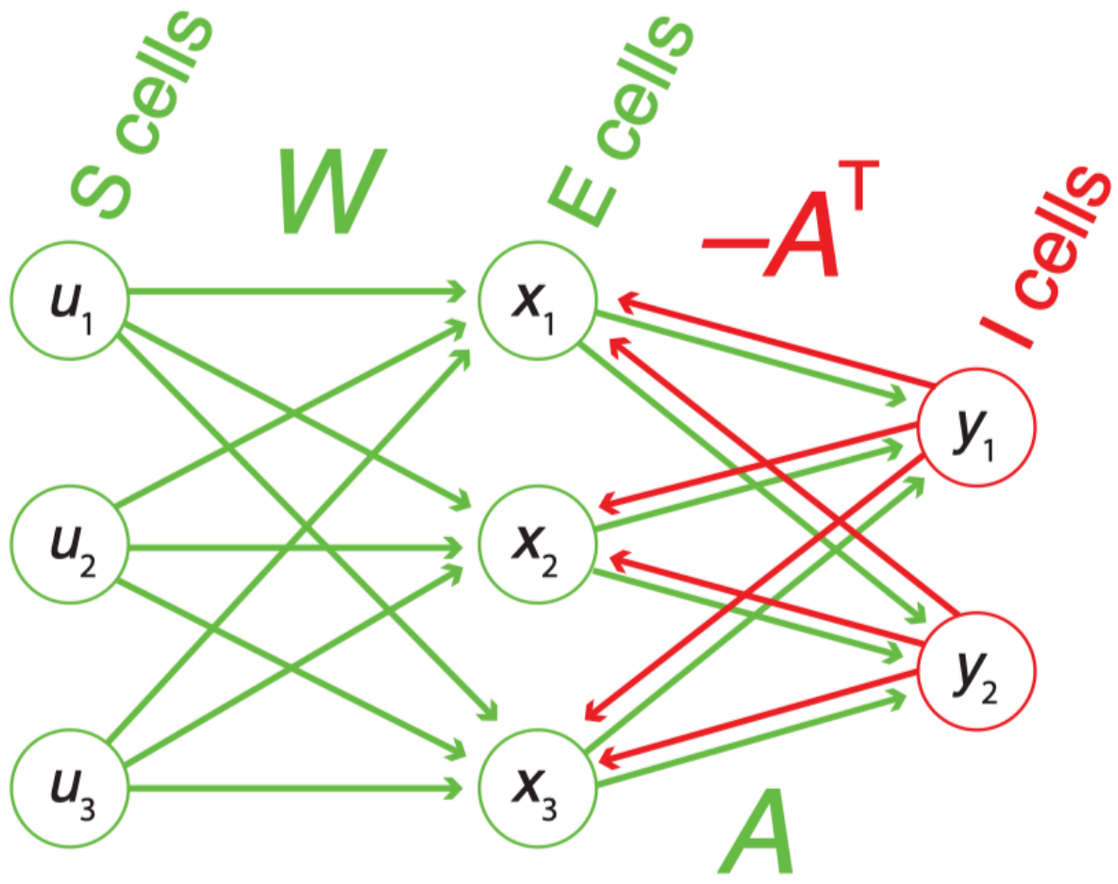
# Disynaptic Neural Networks




**Excitatory-inhibitory net respecting Dale's Law.**

- The S→E connection from $u_a$ to $x_i$ has strength $W_{ia}$ .
- The E neurons $x_i$ are reciprocally coupled with the I neurons $y_\alpha$ .
- The E → I connection from $x_i$ to $y_\alpha$ has strength $A_{\alpha i}$
- The I → E connection from $y_\alpha$ to $x_i$ has strength $-A_{\alpha i}$.

# Disynaptic Neural Networks

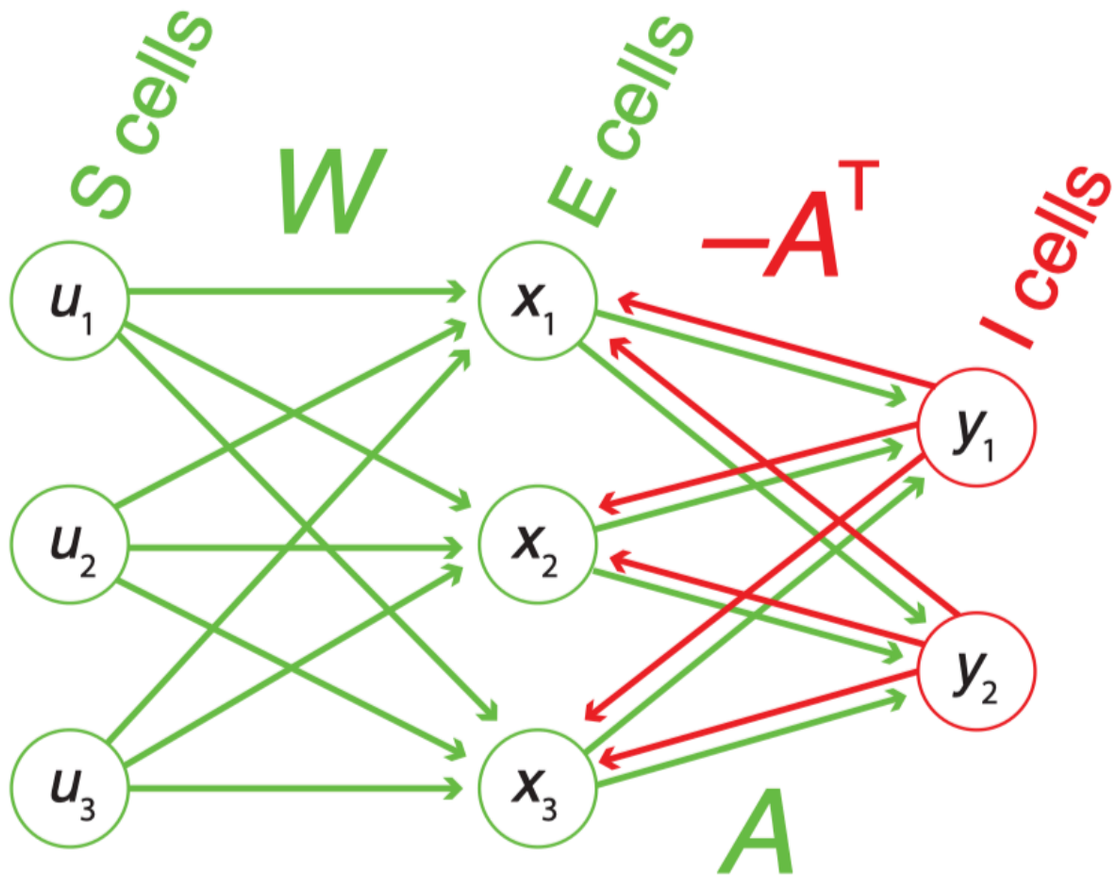



**Activity dynamics:**

$$x_i := \left[ (1 - dt)\, x_i + dt\, \lambda_i^{-1} \left( \sum_{a=1}^{n} W_{ia} u_a - \sum_{\alpha=1}^{r} y_\alpha A_{\alpha i} \right) \right]^+$$

$$y_\alpha = \sum_{i=1}^{m} A_{\alpha i} x_i$$

# Disynaptic Neural Networks



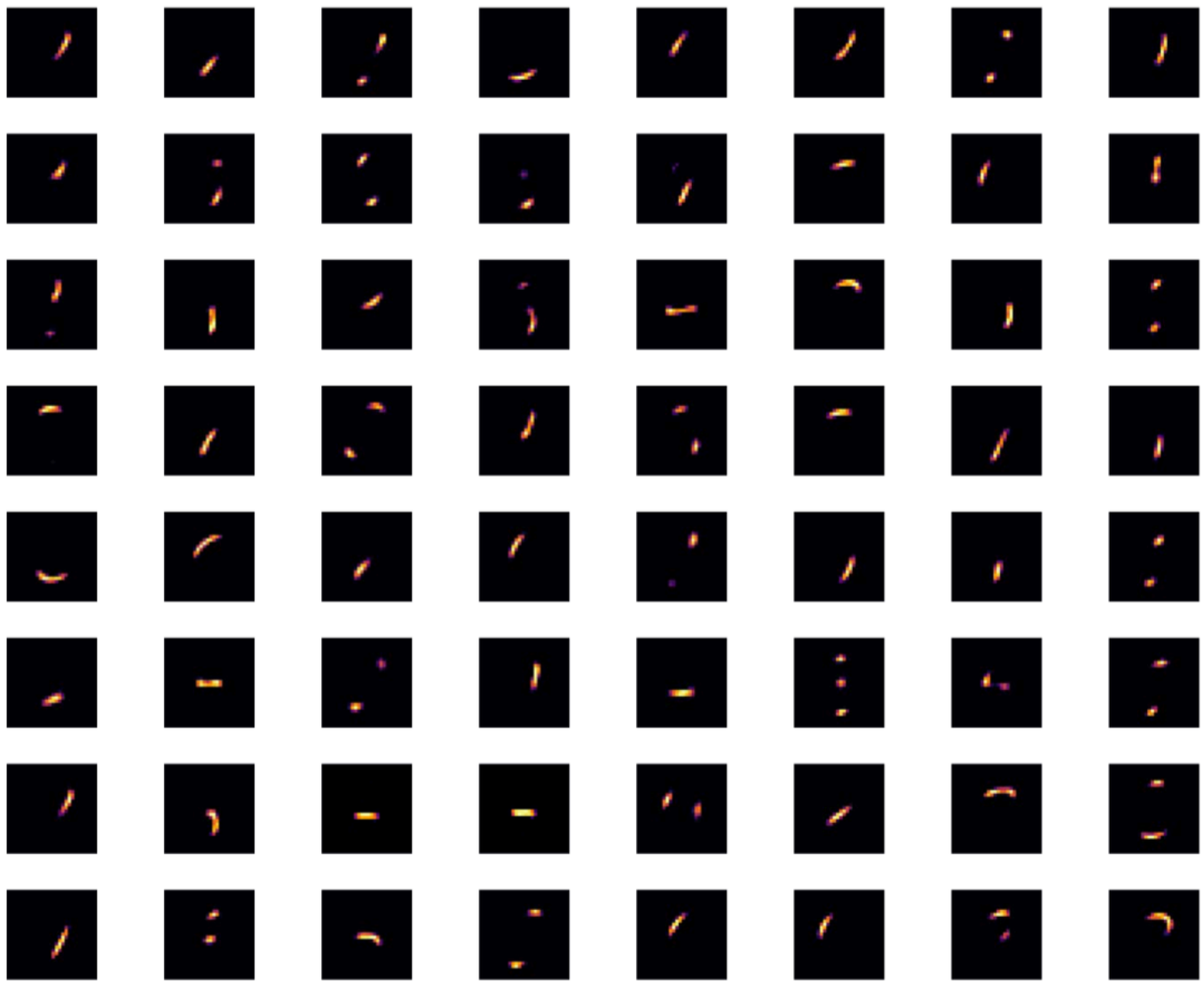

**Learning rules (Hebbian / Anti-hebbian learning):**

$$\Delta W_{ia} \propto x_i u_a - \gamma W_{ia} - \kappa \sum_b W_{ib}$$

$$\Delta A_{\alpha j} \propto y_\alpha x_j - \left(q^2 - p^2\right) A_{\alpha j} - p^2 \sum_i A_{\alpha i}$$

$$\Delta \lambda_i \propto x_i^2 - q^2$$

# Disynaptic Neural Networks

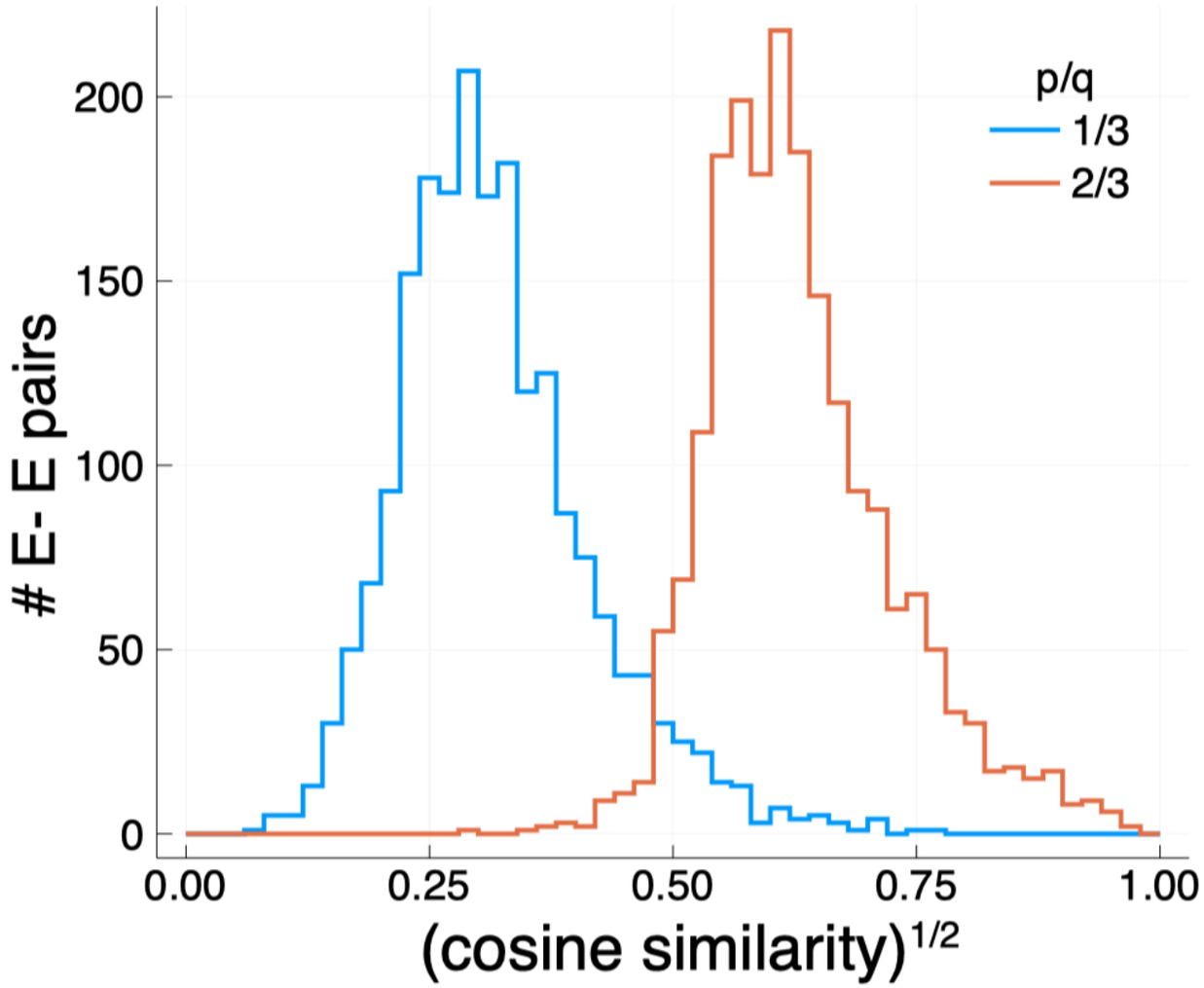


**For each E neuron, the convergent S → E connections constitute a sensory feature learned from the stimuli. (γ/κ = 5)**

# Disynaptic Neural Networks

**For each E neuron, the convergent S → E connections constitute a sensory feature learned from the stimuli. (γ/κ = 50)**

# Disynaptic Neural Networks




**The ratio p/q controls the degree of decorrelation.**

# Correlation Game

Define the goal of unsupervised learning
as the constrained optimization:

$$\max_{X \geq 0} \Phi^* \left( \frac{XU^\top}{T} \right) \text{ subject to copositivity of } D - \frac{XX^\top}{T}$$

# Correlation Game

Define the goal of unsupervised learning
as the constrained optimization:

$$\max_{X \geq 0} \Phi^* \left( \frac{XU^\top}{T} \right) \text{ subject to copositivity of } D - \frac{XX^\top}{T}$$

where $\Phi^*(C) = \max_{W \geq 0} \left\{ \sum_{ia} W_{ia} C_{ia} - \Phi(W) \right\}$, $D_{ij} = \begin{cases} q^2, & i = j, \\ p^2, & i \neq j \end{cases}$.

# Correlation Game

Define the goal of unsupervised learning
as the constrained optimization:

$$\max_{X \geq 0} \Phi^* \left( \frac{XU^\top}{T} \right) \text{ subject to copositivity of } D - \frac{XX^\top}{T}$$

where $\Phi^*(C) = \max_{W \geq 0} \left\{ \sum_{ia} W_{ia} C_{ia} - \Phi(W) \right\}$, $D_{ij} = \begin{cases} q^2, & i = j, \\ p^2, & i \neq j \end{cases}$.

convex conjugate of regularizer

$$\Phi(W) = \frac{\gamma}{2} \sum_{ia} W_{ia}^2 + \frac{\kappa}{2} \sum_i \left( \sum_a W_{ia} \right)^2$$

# Correlation Game

Define the goal of unsupervised learning
as the constrained optimization:

$$\max_{X \geq 0} \Phi^* \left( \frac{XU^\top}{T} \right) \text{ subject to copositivity of } D - \frac{XX^\top}{T}$$

where $\Phi^*(C) = \max_{W \geq 0} \left\{ \sum_{ia} W_{ia} C_{ia} - \Phi(W) \right\}$, $D_{ij} = \begin{cases} q^2, & i = j, \\ p^2, & i \neq j \end{cases}$.

$$\langle x_i^2 \rangle \leq D_{ii} \qquad \langle x_i x_j \rangle \leq D_{ij}$$

constraints on correlation

# Correlation Game

Define the goal of unsupervised learning
as the constrained optimization:

$$\max_{X \geq 0} \Phi^* \left( \frac{XU^\top}{T} \right) \text{ subject to copositivity of } D - \frac{XX^\top}{T}$$

$$\Downarrow$$

Introducing Lagrange multipliers $\mathbf{A}$ and $\Lambda$:

$$\min_{A,\Lambda \geq 0} \max_{X \geq 0} \left\{ \Phi^* \left( \frac{XU^\top}{T} \right) + \frac{1}{2} \mathrm{Tr} \left( D - \frac{XX^\top}{T} \right) \left( A^\top A + \Lambda \right) \right\}$$

# Correlation Game

- Original Problem(s): **(1)** $\equiv$ **(2)** $\equiv$ **(3)**

$$\underset{X\geq 0}{\max}\underset{L\geq 0}{\min}\underset{W\geq 0}{\max}\left\{\frac{1}{T}\operatorname{Tr} W^{\intercal}XU^{\intercal}-\Phi(W)+\frac{1}{2}\left(D-\frac{XX^{\intercal}}{T}\right)L\right\} \tag{1}$$

$$\underset{X\geq 0}{\max}\underset{W\geq 0}{\max}\underset{L\geq 0}{\min}\left\{\frac{1}{T}\operatorname{Tr} W^{\intercal}XU^{\intercal}-\Phi(W)+\frac{1}{2}\left(D-\frac{XX^{\intercal}}{T}\right)L\right\} \tag{2}$$

$$\underset{W\geq 0}{\max}\underset{X\geq 0}{\max}\underset{L\geq 0}{\min}\left\{\frac{1}{T}\operatorname{Tr} W^{\intercal}XU^{\intercal}-\Phi(W)+\frac{1}{2}\left(D-\frac{XX^{\intercal}}{T}\right)L\right\} \tag{3}$$

- Lagrangian Dual Problem(s): **(4)** $\equiv$ **(5)**

$$\underset{L\geq 0}{\min}\underset{X\geq 0}{\max}\underset{W\geq 0}{\max}\left\{\frac{1}{T}\operatorname{Tr} W^{\intercal}XU^{\intercal}-\Phi(W)+\frac{1}{2}\left(D-\frac{XX^{\intercal}}{T}\right)L\right\} \tag{4}$$

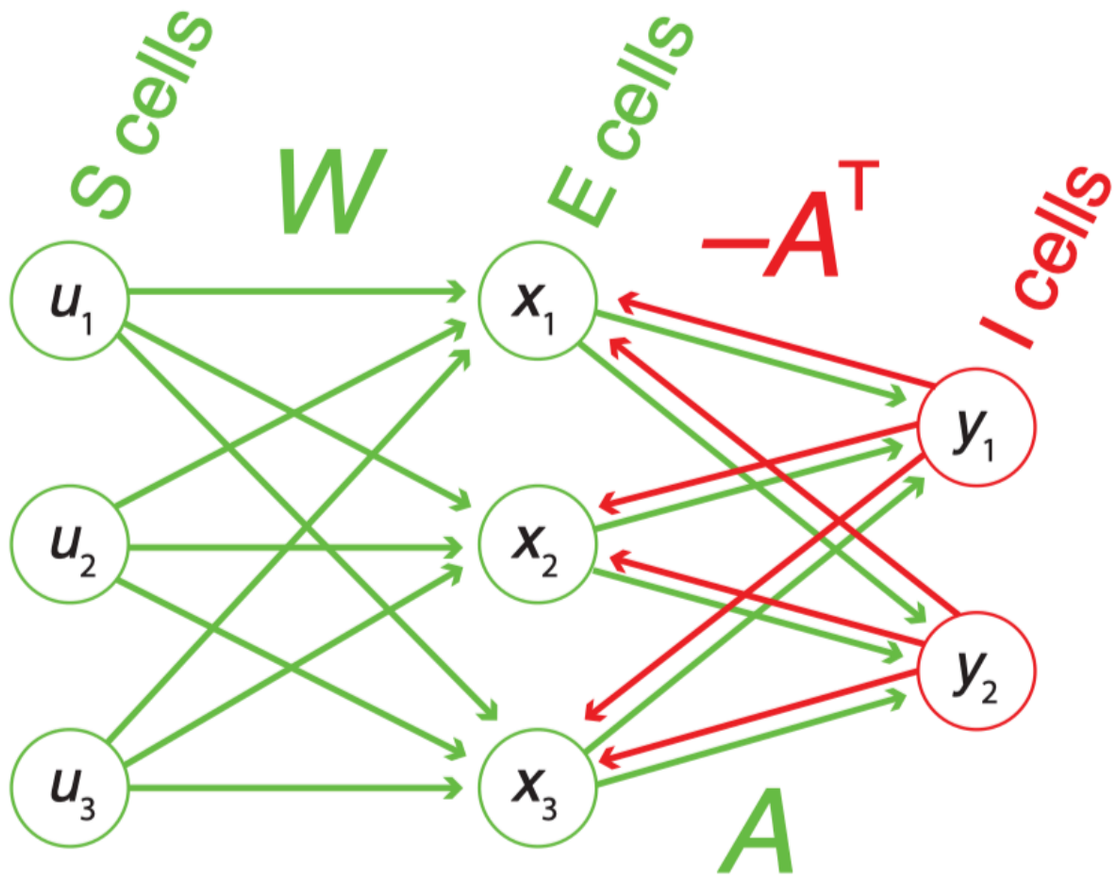
$$\underset{L\geq 0}{\min}\underset{W\geq 0}{\max}\underset{X\geq 0}{\max}\left\{\frac{1}{T}\operatorname{Tr} W^{\intercal}XU^{\intercal}-\Phi(W)+\frac{1}{2}\left(D-\frac{XX^{\intercal}}{T}\right)L\right\} \tag{5}$$

- Neural Network Algorithm:

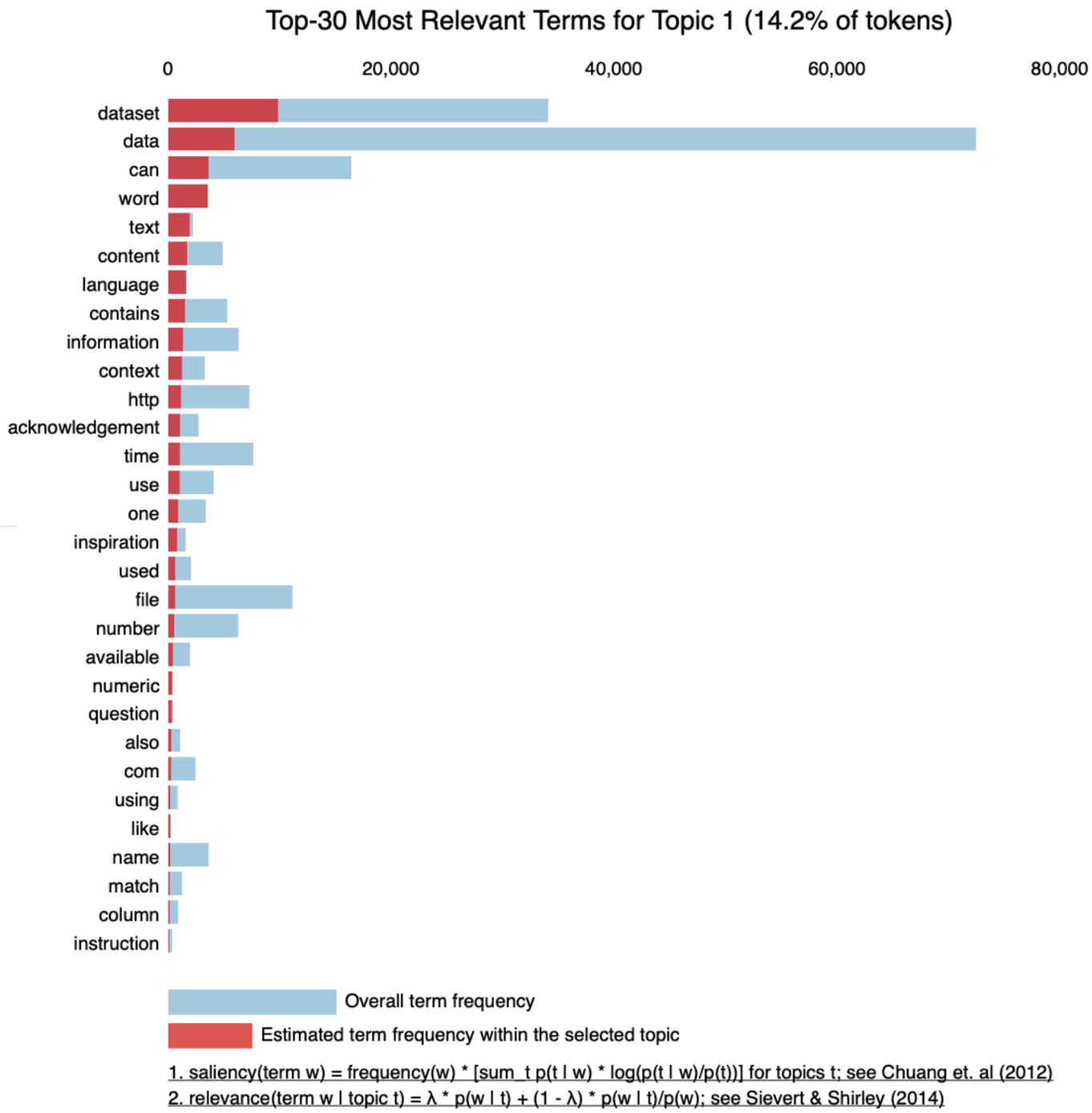
$$\underset{W\geq 0}{\max}\underset{L\geq 0}{\min}\underset{X\geq 0}{\max}\left\{\frac{1}{T}\operatorname{Tr} W^{\intercal}XU^{\intercal}-\Phi(W)+\frac{1}{2}\left(D-\frac{XX^{\intercal}}{T}\right)L\right\} \tag{6}$$

$$(1)\equiv(2)\equiv(3)\leq(6)\leq(4)\equiv(5)$$

# Disynaptic Neural Topic Modelling?

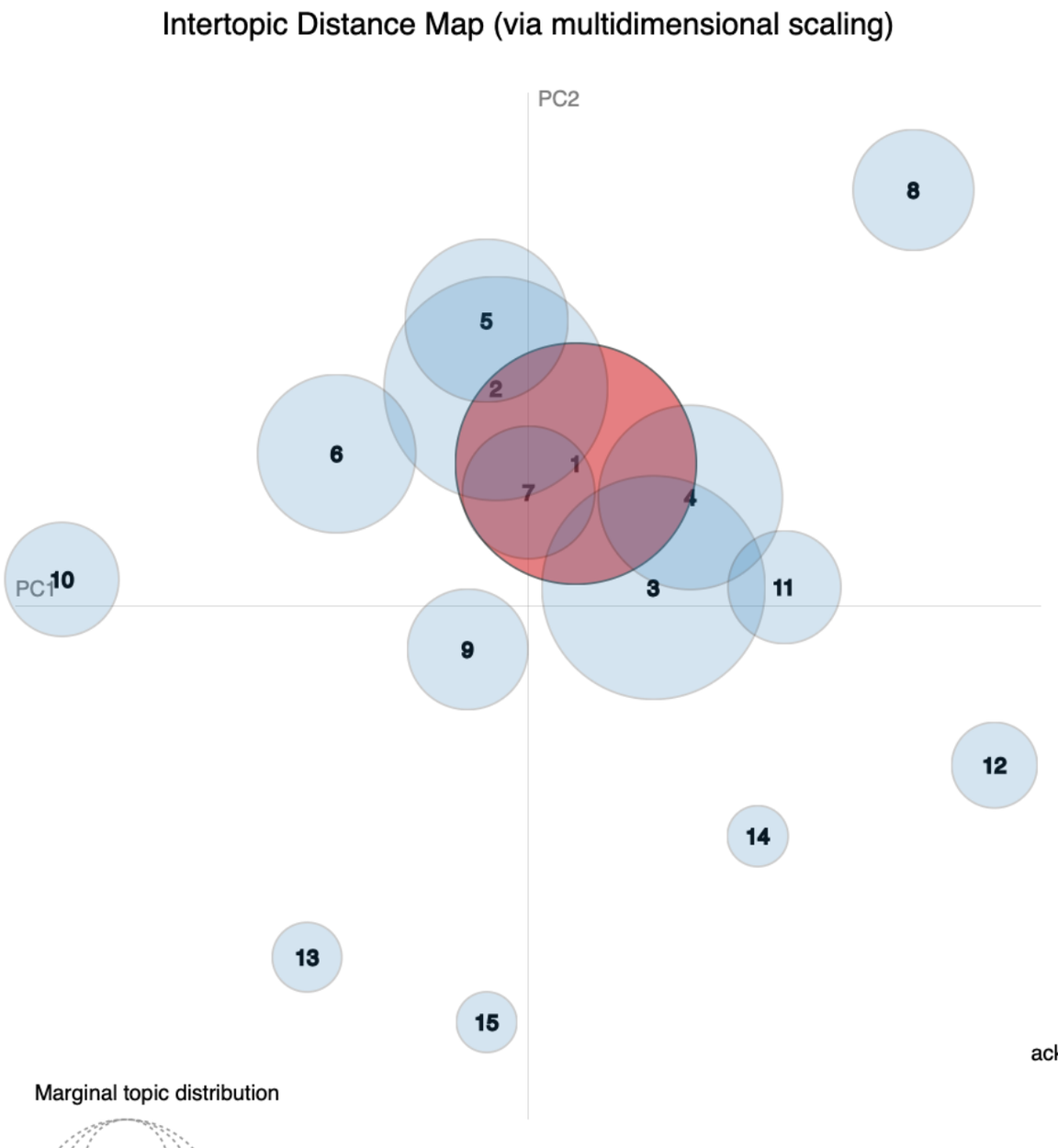


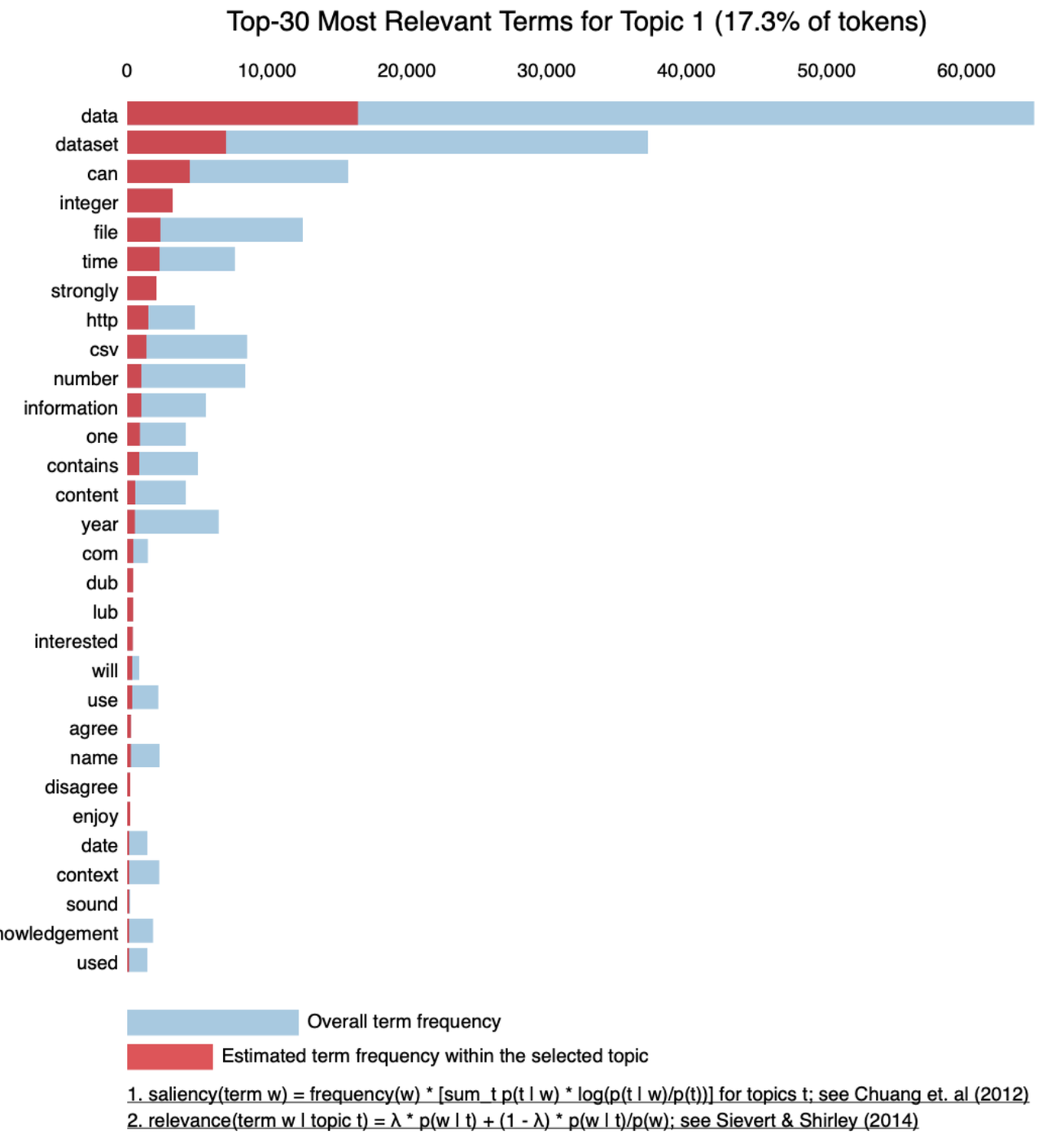
# Disynaptic Neural Topic Modelling?


**p/q = 1/3**

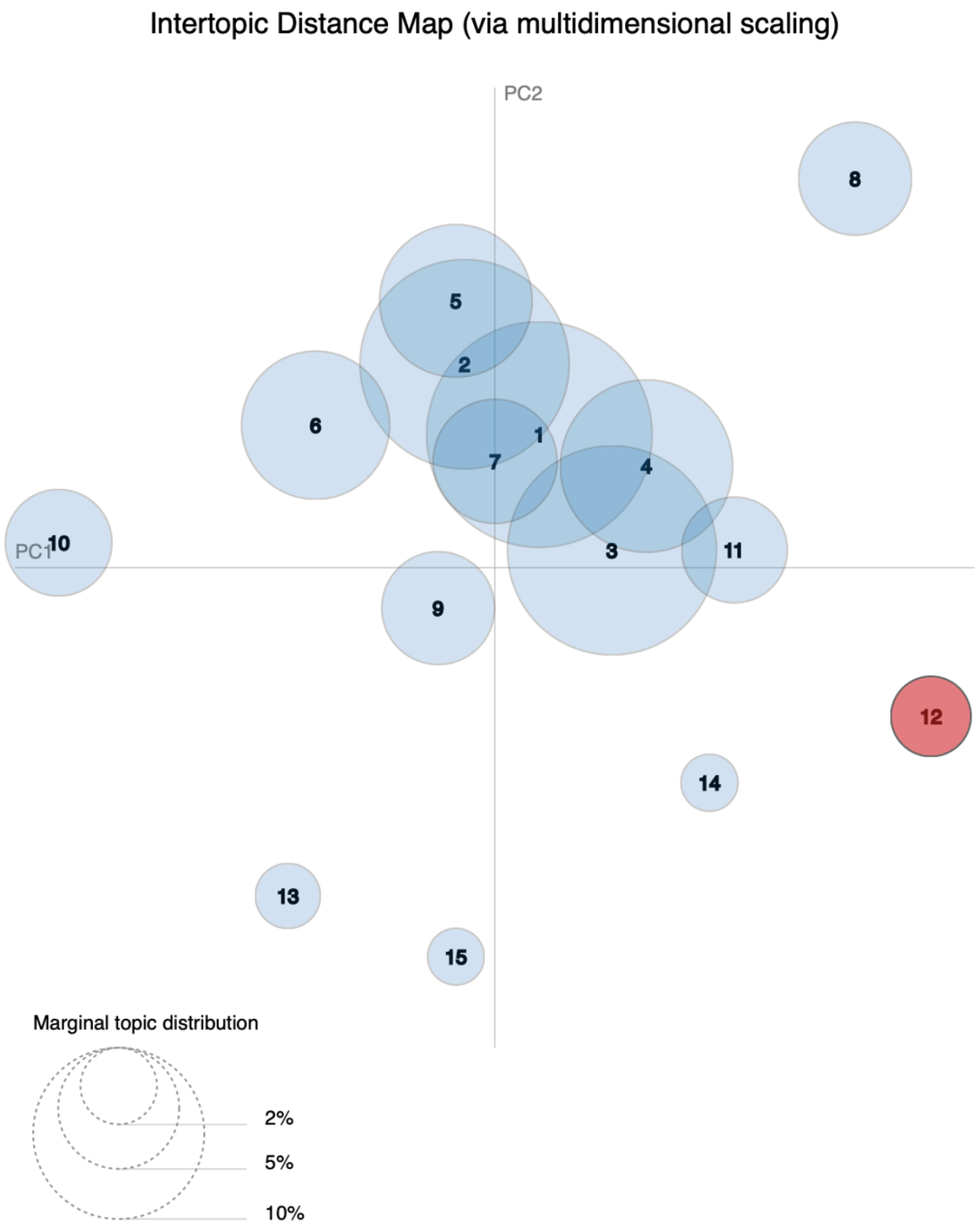
# Disynaptic Neural Topic Modelling?

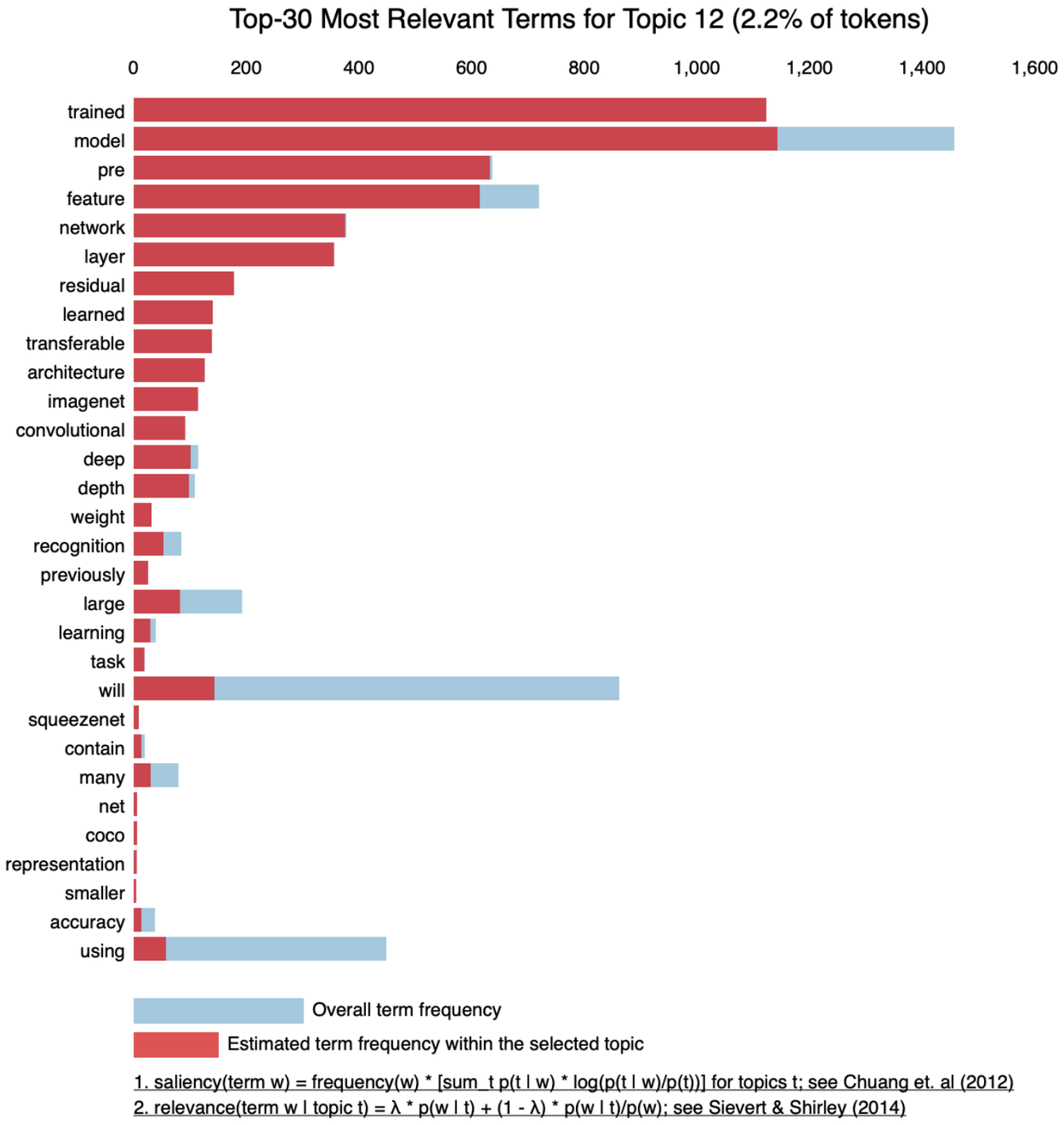



**p/q = 1/18**

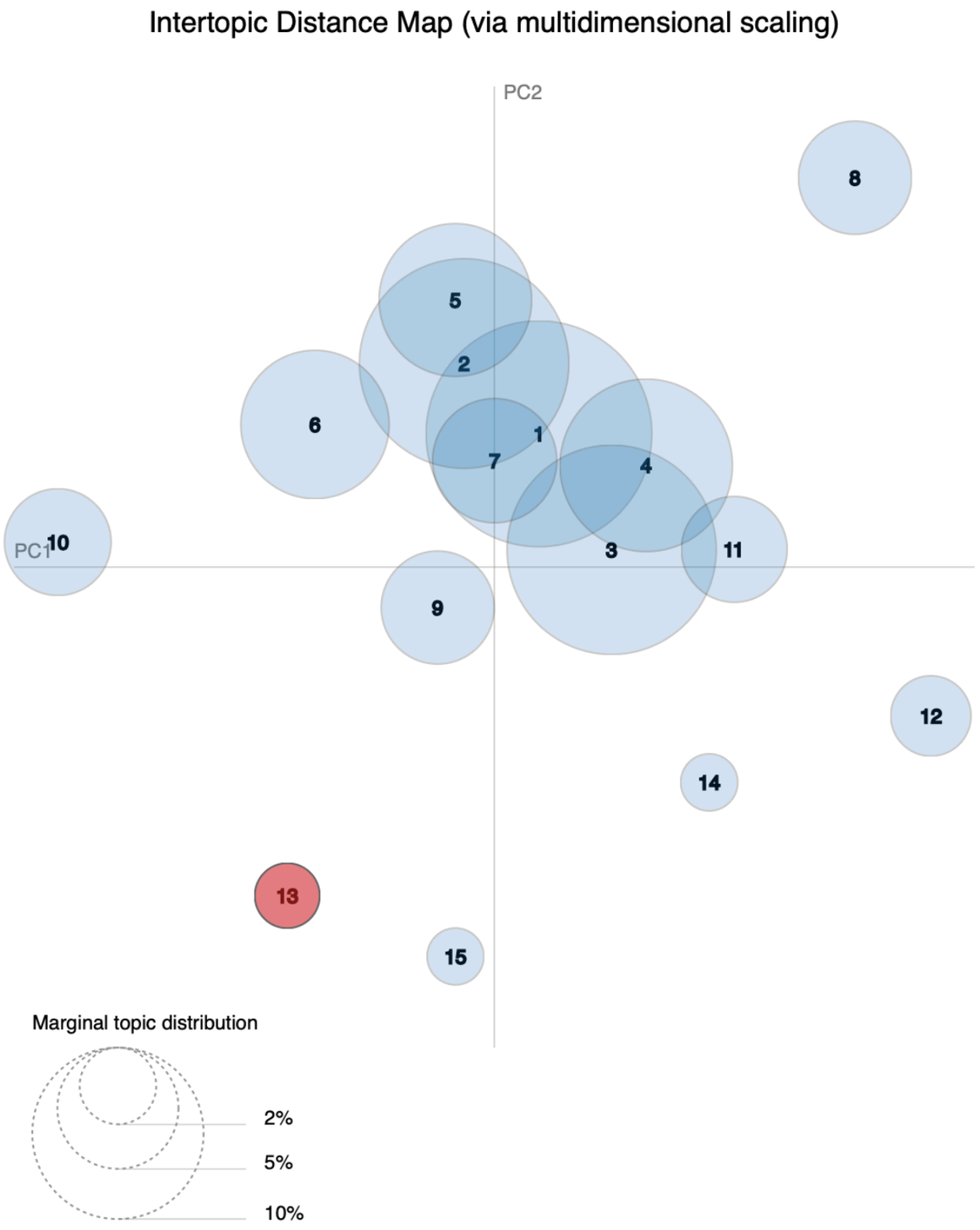
# Disynaptic Neural Topic Modelling?

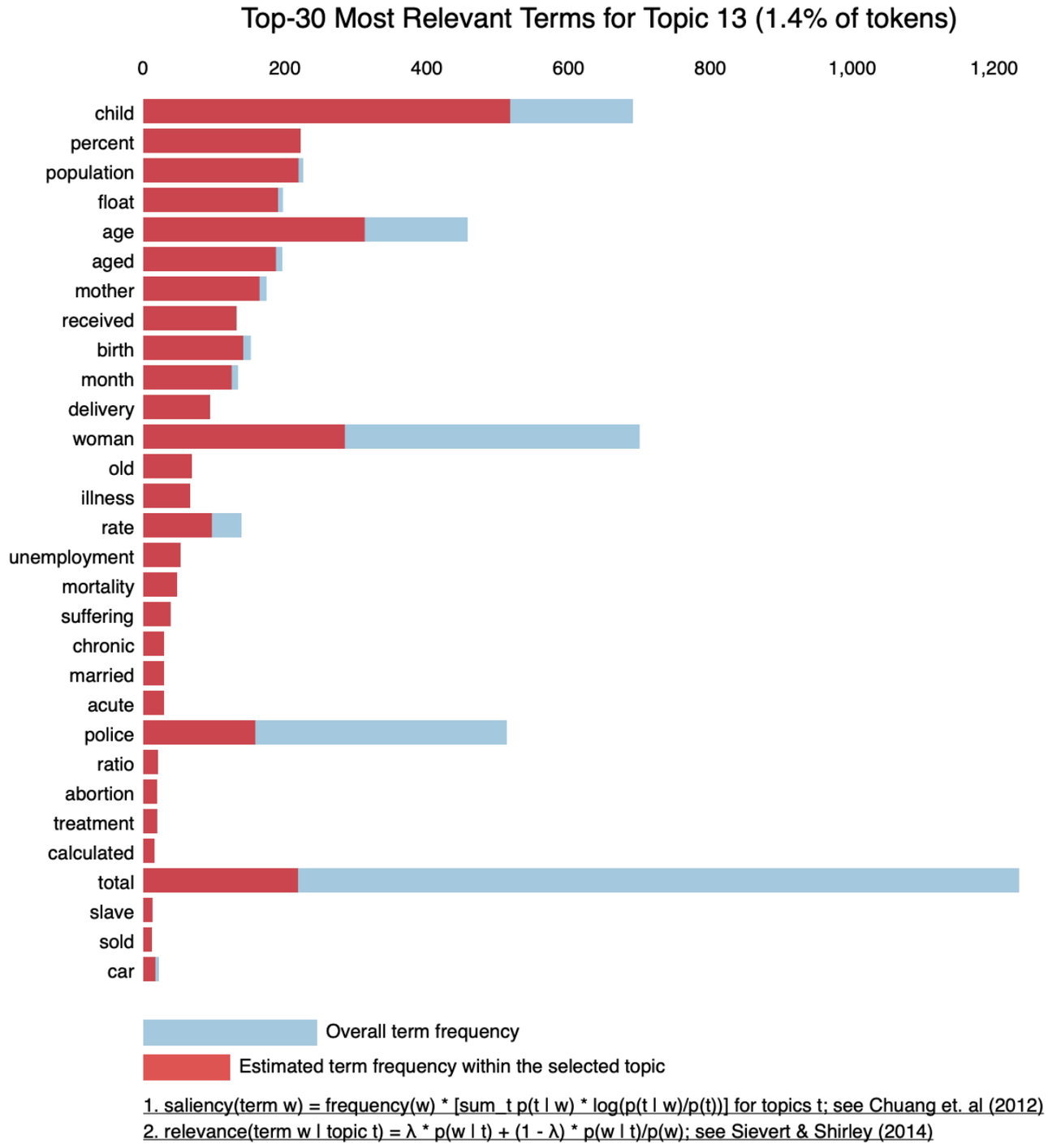



**p/q = 1/18, topic 12 “training neural nets”**

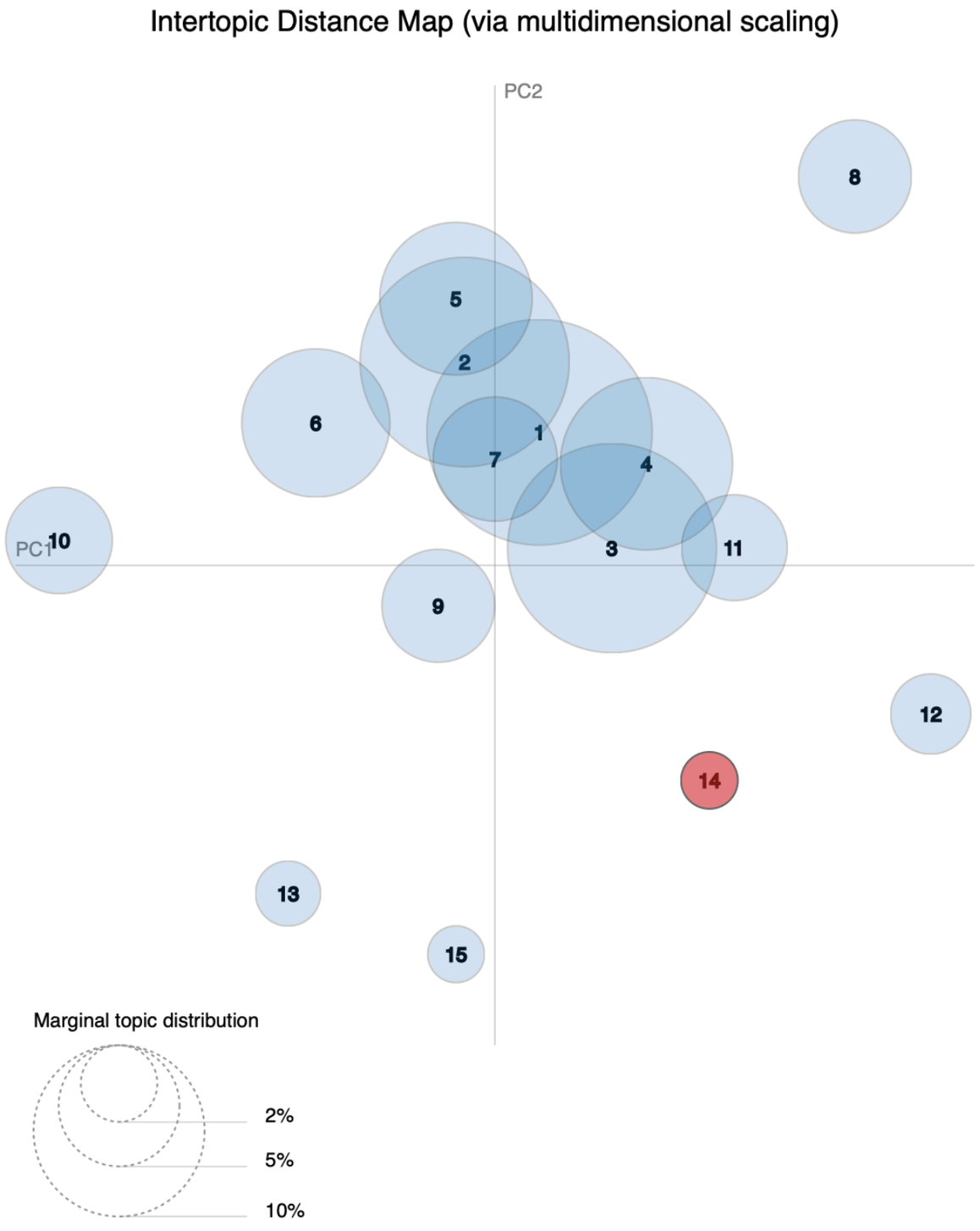
# Disynaptic Neural Topic Modelling?

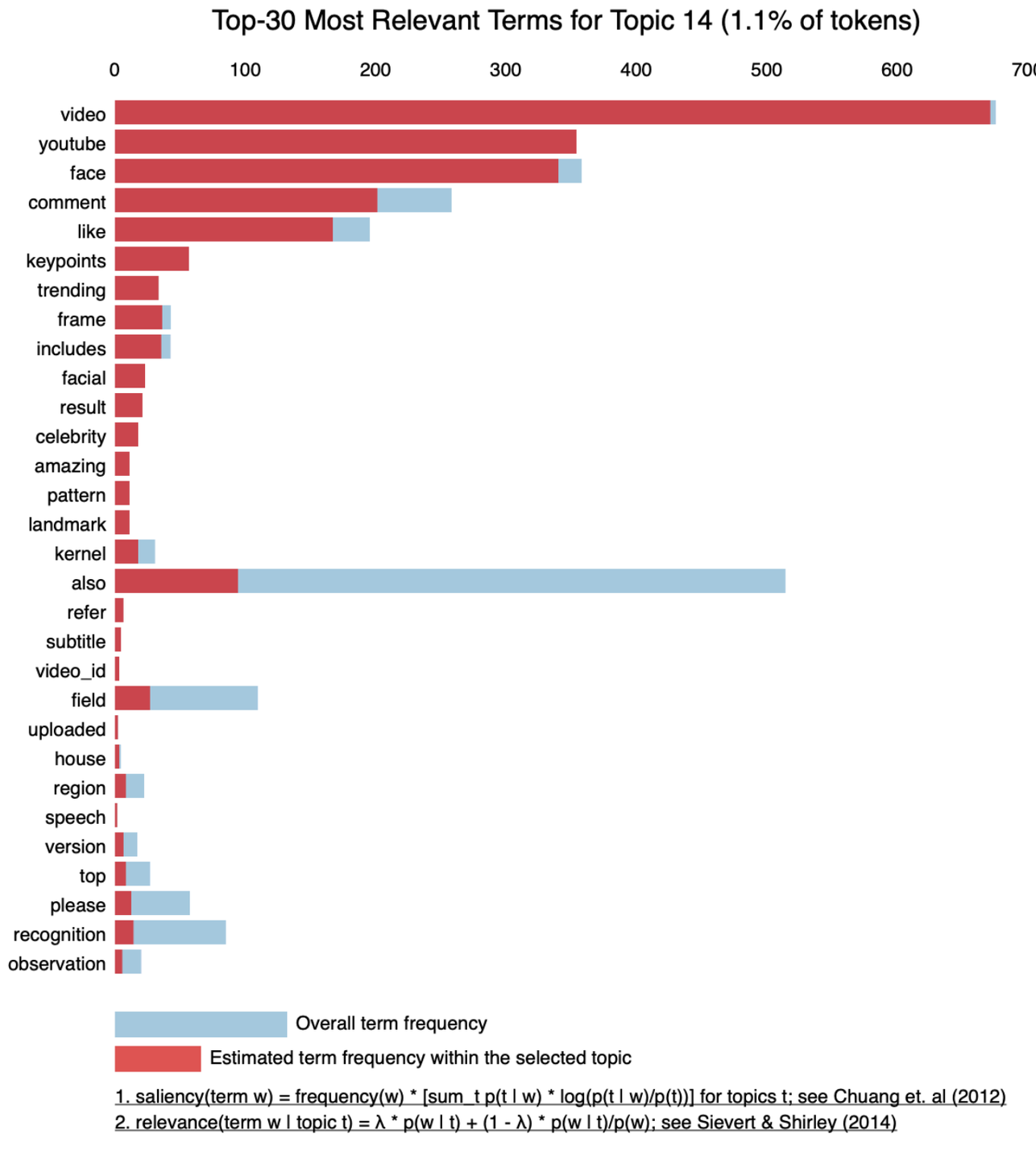



**p/q = 1/18, topic 13 “population”**

# Disynaptic Neural Topic Modelling?




**p/q = 1/18, topic 14 “video detection”**