COMP6237 Data Mining

Finding Independent Features

Jonathon Hare
jsh2@ecs.soton.ac.uk

Content based on material from slides on NMF from Derek Greene at UCD (http://derekgreene.com/slides/nmf_insight_workshop.pdf)
and
David Blei's MLSS slides on LDA (http://www.cs.columbia.edu/~blei/talks/Blei_MLSS_2012.pdf)

Introduction

- Topic Models
- Non-negative Matrix Factorisation
- Brief introduction to Probabilistic approaches

Problem statement

- When we looked at LSA, we saw that it created *concepts* that were linear mixtures of words
    - But, the weightings were unconstrained, and could be negative
        - Very difficult to interpret or give semantic meaning to the topic
- Would be really nice if we could determine *thematic* **topics** for a corpus of documents

Topic Modelling

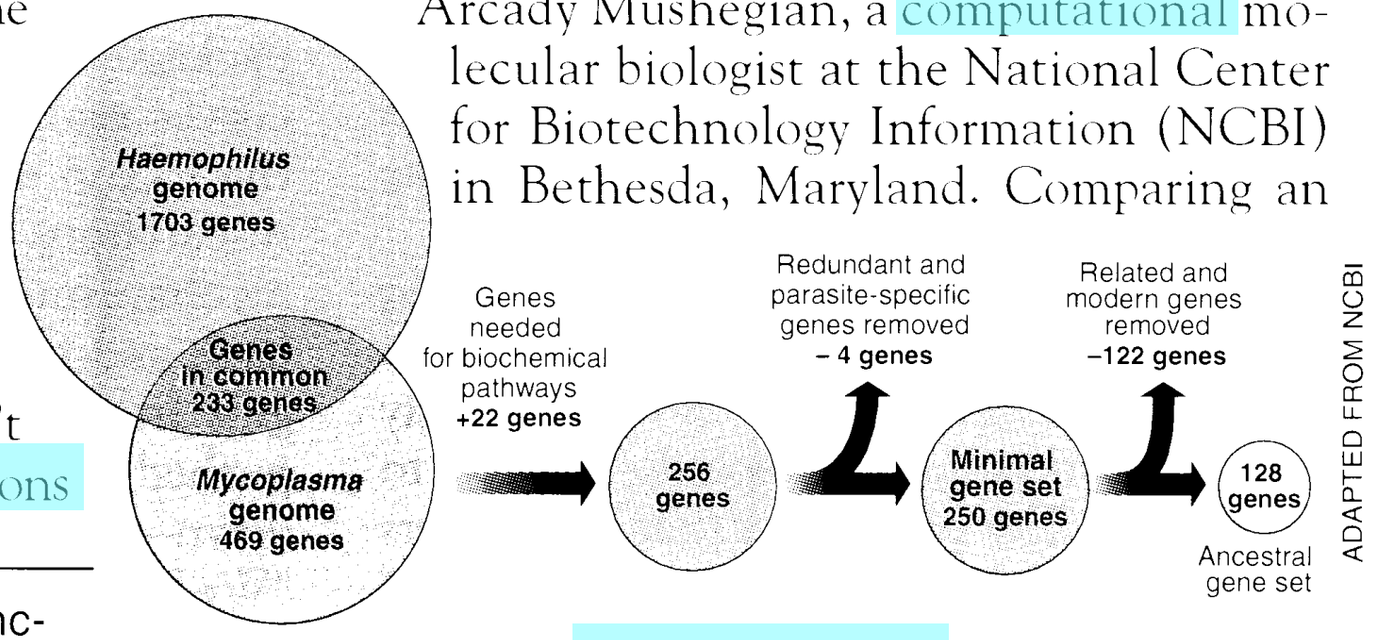
- Topic models uncover the **hidden thematic structure** in document collections.

- These algorithms help us develop new ways to
  - search
  - browse
  - summarise

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Simple intuition: documents exhibit multiple topics

Key topic modelling techniques

- There are many!
  - Probabilistic Latent Semantic Analysis (PLSA)
  - Latent Dirichlet Allocation (LDA)
  - Pachinko allocation (PAM)
  - Non-Negative Matrix Factorisation ([N]NMF)

Key topic modelling techniques

- There are many!
  - Probabilistic Latent Semantic Analysis (pLSA)
  - Latent Dirichlet Allocation (LDA)
  - Pachinko allocation (PAM)

  Probabilistic Models
- Non-Negative Matrix Factorisation ([N]NMF)

Relationship to clustering

- Clearly topic modelling is related to clustering
  - trying to group documents into similar sets
- Topic models in a sense perform *soft clustering*
  - A document can belong to a weighted **mixture** of topics

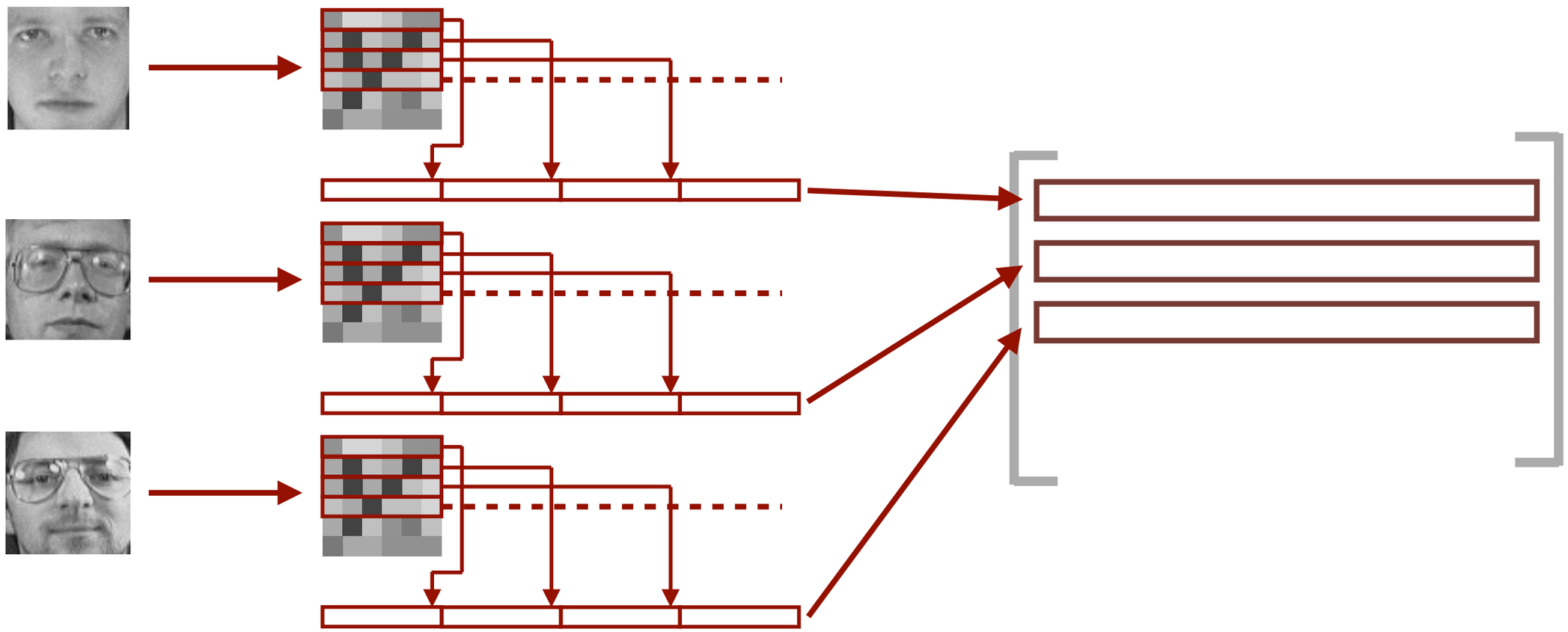
Non-negative Matrix Factorisation

Non-negative Matrix Factorisation

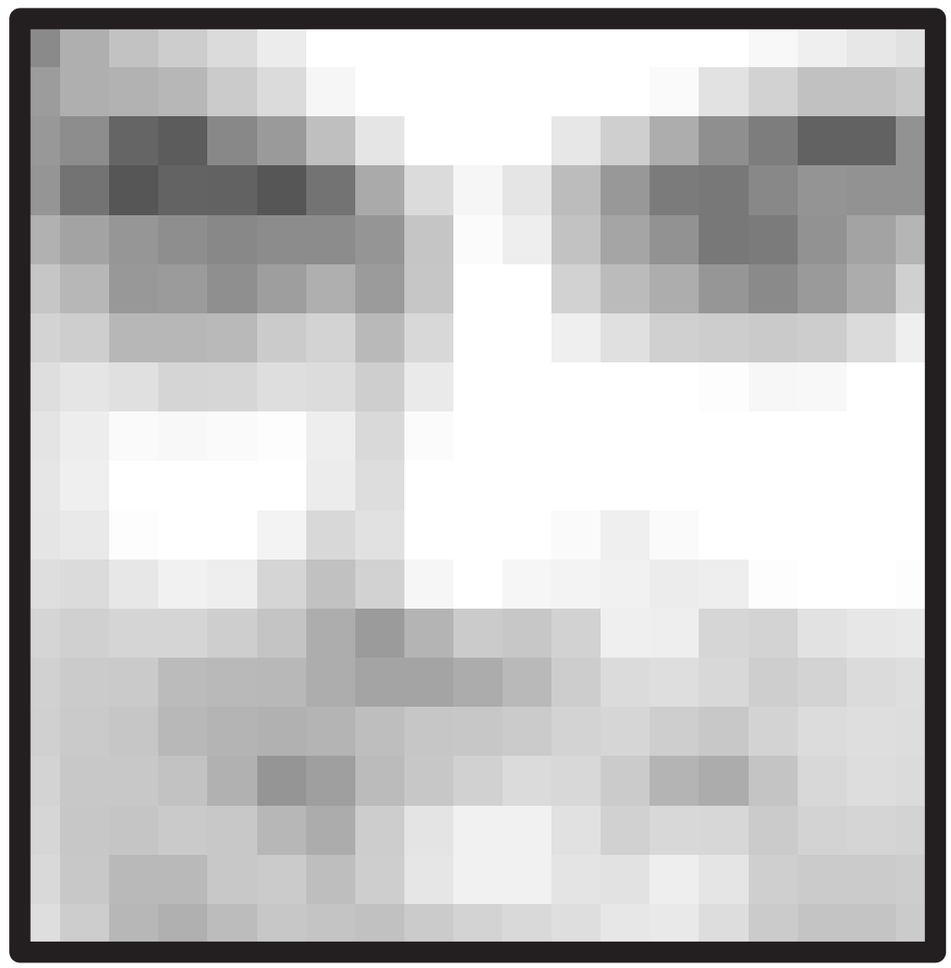
- **NMF**: an unsupervised family of algorithms that simultaneously perform dimension reduction and clustering.
    - Sometimes referred to as NNMF
    - Also known as positive matrix factorisation (PMF) and non-negative matrix approximation (NNMA).
- No strong statistical justification or grounding.
- But has been successfully applied in a range of areas:
    - Bioinformatics (e.g. clustering gene expression networks).
    - Image processing (e.g. face detection).
    - Audio processing (e.g. source separation).
    - **Text analysis**

NMF Overview

- NMF produces a "**parts-based**" decomposition of the latent relationships in a data matrix
  - Like SVD/PCA can reduce the dimensionality
- Best explained visually in comparison to PCA of a data matrix formed from images:



Use low-rank basis from PCA and NNMF to reconstruct this face:

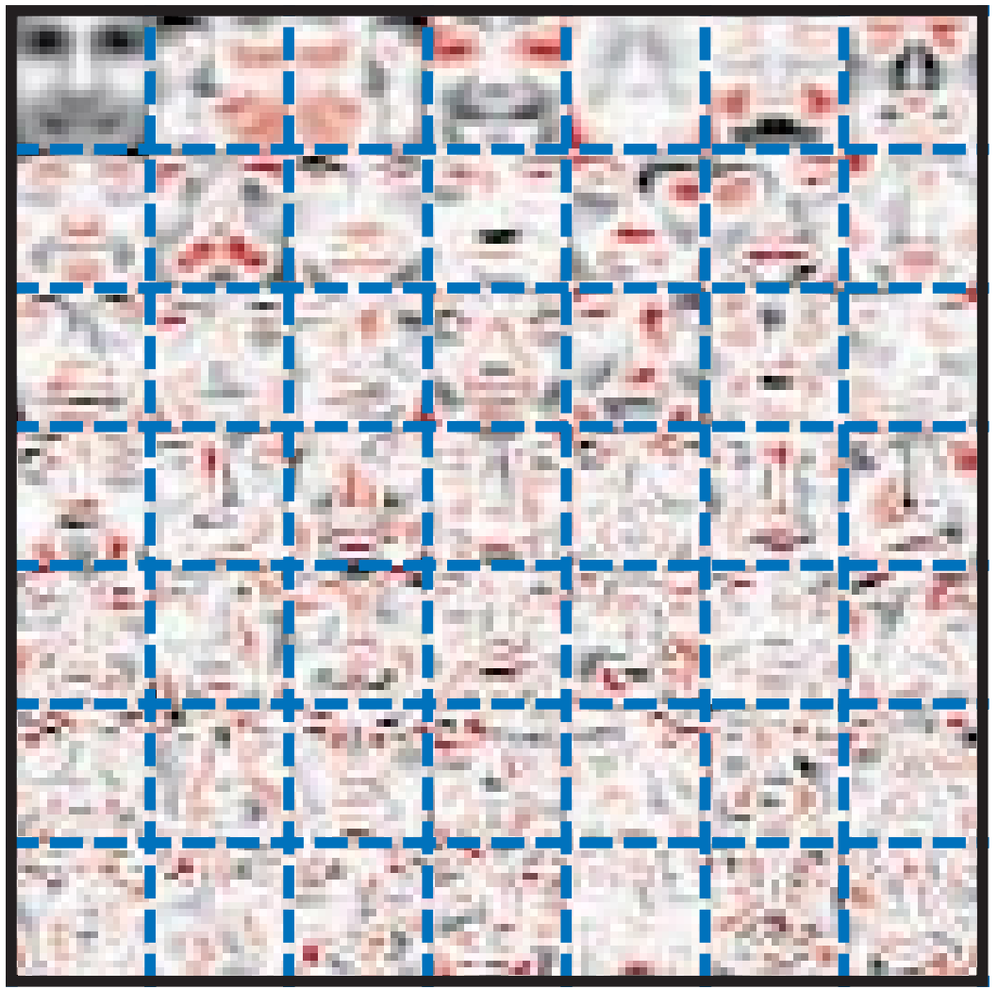


Daniel D. Lee and H. Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". *Nature* **401** (6755): 788–791

Red signifies -ve

Stronger colour indicates larger values

PCA



×

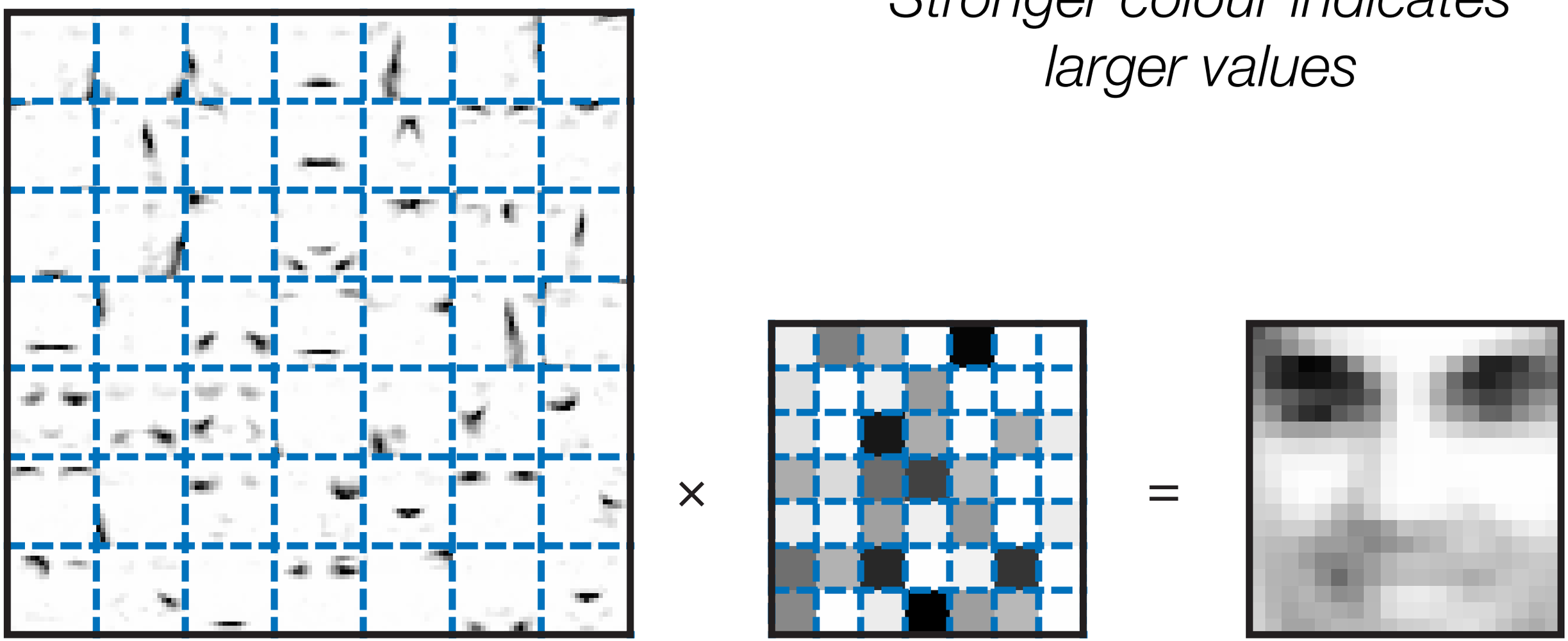
=

Daniel D. Lee and H. Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". *Nature* **401** (6755): 788–791

Red signifies -ve

Stronger colour indicates larger values

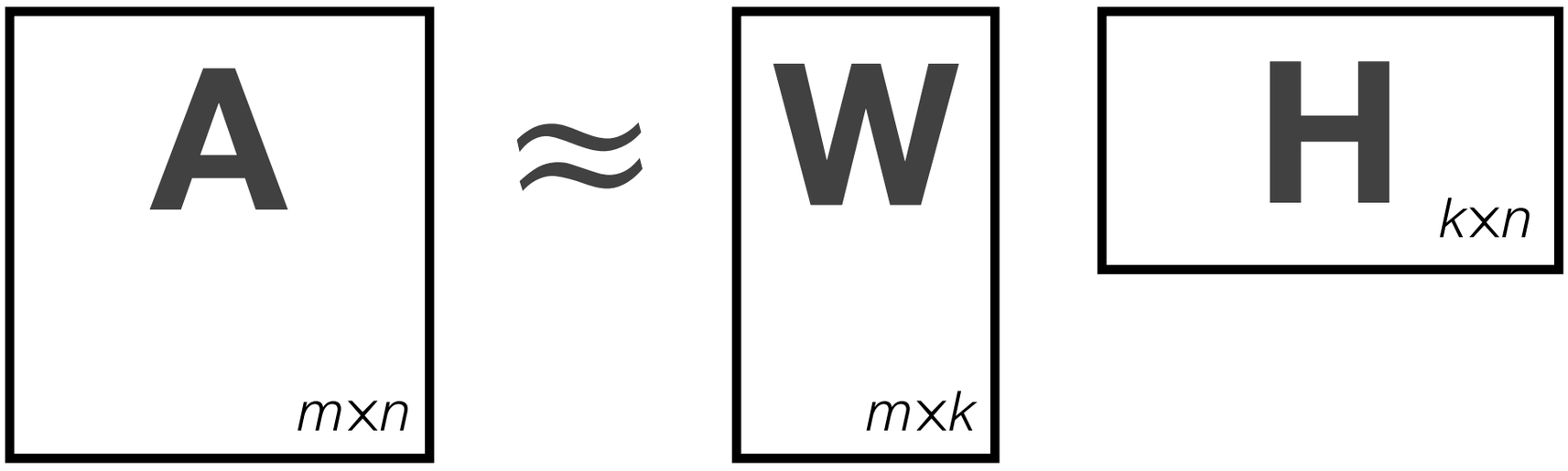
NMF



Daniel D. Lee and H. Sebastian Seung (1999). "Learning the parts of objects by non-negative matrix factorization". *Nature* **401** (6755): 788–791

NMF Overview

- Given a non-negative matrix A, find k-dimension approximation in terms of non-negative factors W and H:



W≥0, H≥0

Data Matrix
(rows=features, cols=items)

Basis Matrix
(rows=features)

Coefficient Matrix
(cols=items)

- Approximate each item (i.e. column of **A**) by a linear combination of k reduced dimensions or “basis vectors” in **W**.

- Each basis vector can be interpreted as a cluster. The memberships of items in these clusters encoded by **H**.

NMF Algorithm Components

- **Input**: Non-negative data matrix (**A**), number of basis vectors (*k*), initial values for factors **W** and **H** (e.g. random matrices).

- **Objective Function**: Some measure of reconstruction error between **A** and the approximation **WH**.

Euclidean Distance (Lee & Seung, 1999)

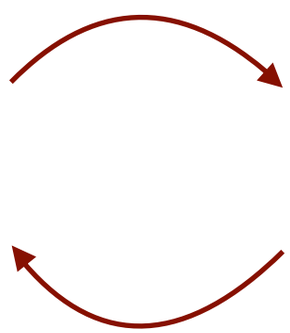
$$\frac{1}{2}||\mathbf{A}-\mathbf{WH}||_{\mathrm{F}}^2=\sum_{i=1}^{n}\sum_{j=1}^{m}(A_{ij}-(WH)_{ij})^2$$

- Optimisation Process: Local EM-style optimisation to refine **W** and **H** in order to minimise the objective function.

- Common approach is to iterate between two multiplicative update rules until convergence:

1. Update **H**

$$H_{cj}\leftarrow H_{cj}\frac{(W\mathbf{A})_{cj}}{(W\mathbf{WH})_{cj}}$$



2. Update **W**

$$W_{ic}\leftarrow W_{ic}\frac{(\mathbf{A}H)_{ic}}{(\mathbf{WH}H)_{ic}}$$

NMF Variants

- **Different objective functions:**
  - KL divergence; Bregman divergences…
- **More efficient optimisation:**
  - Alternating least squares with projected gradient method for sub-problems.
- **Constraints:**
  - ***Enforcing sparseness in outputs***.
  - Incorporation of background information (Semi-NMF).
- **Different inputs:**
  - Symmetric matrices - e.g. document-document cosine similarity matrix.

Topic Modelling with NMF

- Basic methodology:

    1. Construct vector space model for documents (after stop-word filtering), resulting in a term-document matrix **A**.

    2. Apply TF-IDF term weight normalisation to **A**.

    3. Normalize TF-IDF vectors to unit length.

    4. Initialise factors (randomly or using NNDSVD(**A**)).

    5. Compute NMF of A.

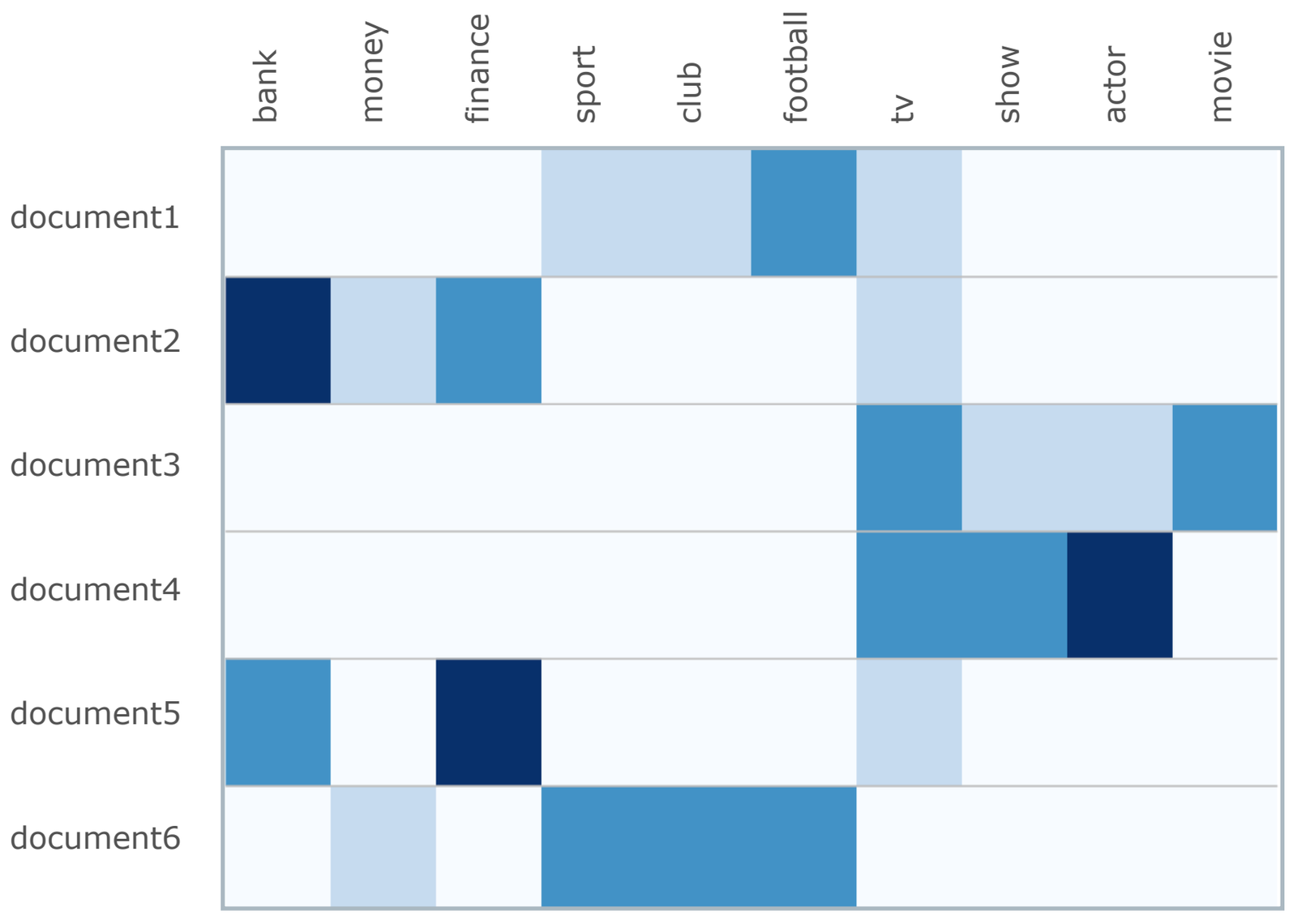
- Interpreting NMF output:

    - **Basis vectors**: the **topics (clusters)** in the data.

    - **Coefficient matrix**: the membership weights for documents relative to each topic (cluster).

NMF Topic Modelling: Simple example

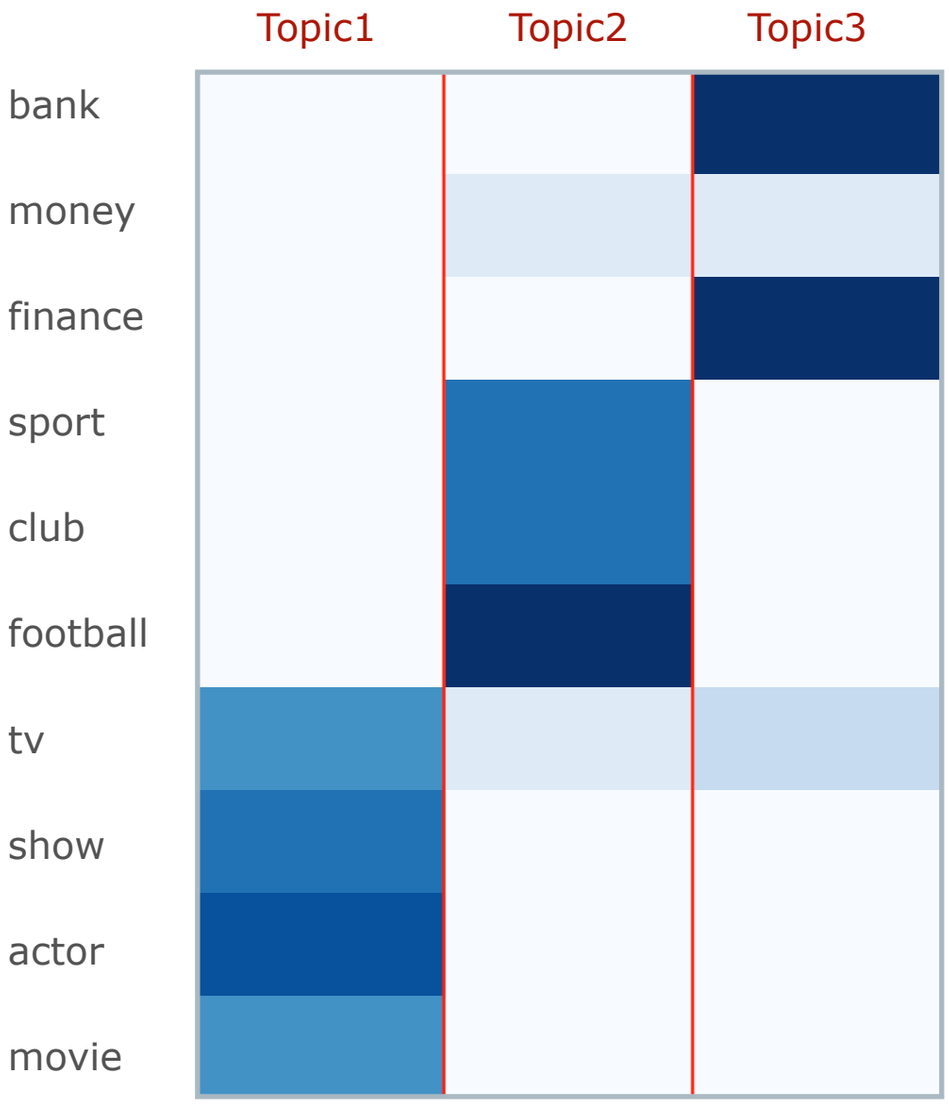
Document-Term Matrix **A**
(6 rows x 10 columns)



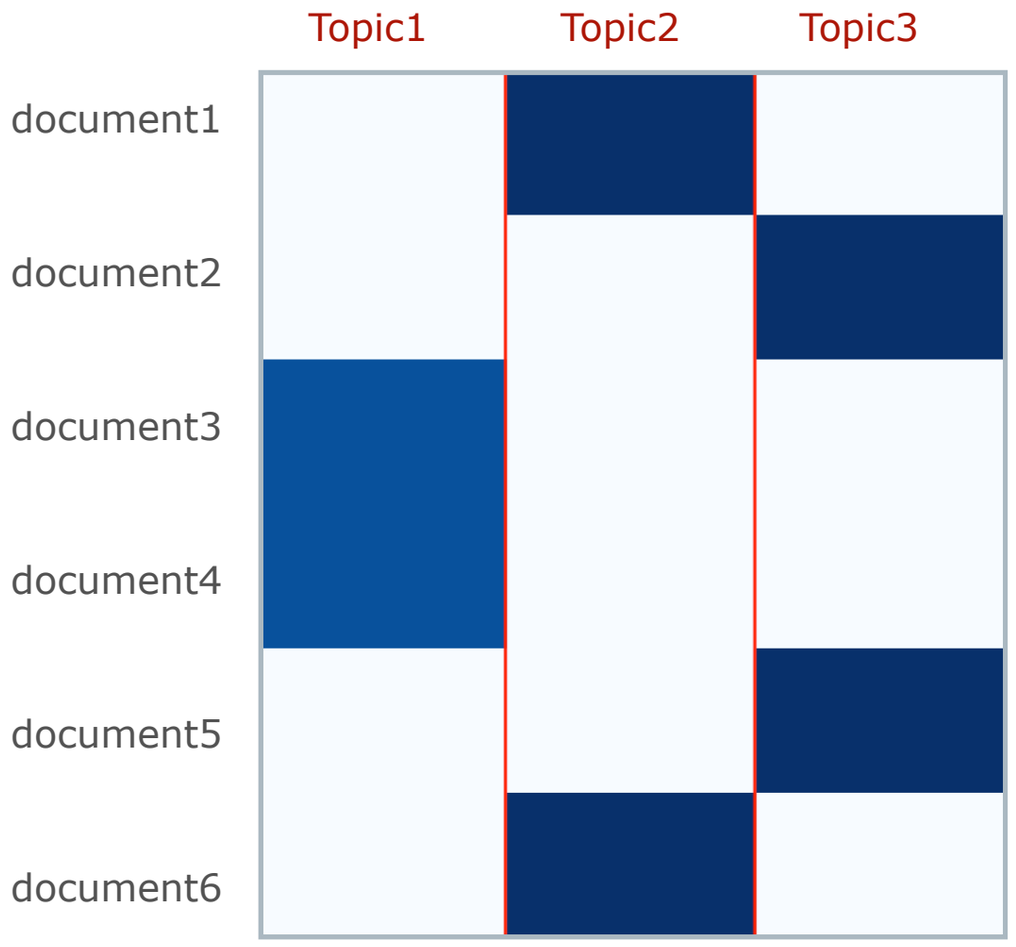
- Apply TF-IDF and unit length normalisation to rows of **A**.
- Run Euclidean NMF on normalised **A** (k=3, random initialisation).

NMF Topic Modelling: Simple example

Basis vectors **W**: topics (clusters)



Coefficients **H**: memberships for documents



Challenge: Selecting K

- The selection of number of topics k is often performed manually.
  - **No definitive model selection strategy.**
- Various alternatives comparing different models:
  - Compare reconstruction errors for different parameters.
    - *Natural bias towards larger value of k.*
  - Build a “consensus matrix” from multiple runs for each *k*, assess presence of block structure.
  - Examine the *stability* (i.e. agreement between results) from multiple randomly initialised runs for each value of *k*.

Challenge: initialisation

- Standard random initialisation of NMF factors can lead to **instability**
    - i.e. significantly different results for different runs on the same data matrix.
- **NNDSVD**: Nonnegative Double Singular Value Decomposition
    - Provides a deterministic initialisation with no random element.
    - Chooses initial factors based on positive components of the first k dimensions of SVD of data matrix **A**.
    - Often leads to significant decrease in number of NMF iterations required before convergence.

Experiment: BBC News Articles

- Collection of 2,225 BBC news articles from 2004-2005 with 5 manually annotated topics
- Apply NMF with k=5 to 2,225 x 9,125 matrix.
- Extract topic "descriptions" based on *top-ranked* terms in basis vectors.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
growth	mobile	england	film	labour
economy	phone	game	best	election
year	music	win	awards	blair
bank	technology	wales	award	brown
sales	people	cup	actor	party
economic	digital	ireland	oscar	government
oil	users	team	festival	howard
market	broadband	play	films	minister
prices	net	match	actress	tax
china	software	rugby	won	chancellor

Experiment: Irish Economy Dataset

- Collection of 21k news articles from 2009-2010 relating to the economy (Irish Times, …).
- Extracted all named entities from articles (person, org, location), and constructed 21,496 x 3,014 article-entity matrix.
- Apply NMF (k=8) and examine topics on basis of top-ranked entities

Topic 1	Topic 2	Topic 3	Topic 4
nama	european_union	allied_irish_bank	hse
brian_lenihan	europe	bank_of_ireland	dublin
green_party	greece	anglo_irish_bank	mary_harney
ntma	lisbon_treaty	dublin	department_of_health
anglo_irish_bank	ecb	irish_life_permanent	brendan_drumm

Topic 5	Topic 6	Topic 7	Topic 8
usa	aer_lingus	uk	brian_cowen
asia	ryanair	dublin	fine_gael
new_york	dublin	northern_ireland	fianna_fail
federal_reserve	daa	bank_of_england	green_party
china	christoph_mueller	london	brian_lenihan

Experiment: IMDB Dataset

- Documents constructed from IMDB Keywords for set of 21k movies.
- Applied NMF (k=10) to 20,923 x 5,528 movie-keyword matrix.
- Topic “descriptions” based on top ranked keywords in basis vectors appear to reveal genres and genre cross-overs.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
cowboy	bmovie	martialarts	police	superhero
shootout	atgunpoint	combat	detective	basedoncomic
cowboyhat	bwestern	hero	murder	superheroine
cowboyboots	stockfootage	actionhero	investigation	dccomics
horse	gangmember	brawl	policedetective	secretidentity
revolver	duplicity	fistfight	detectiveseries	amazon
sixshotter	gangleader	disarming	murderer	culttv
outlaw	deception	warrior	policeofficer	actionheroine
rifle	sheriff	kungfu	policeman	twowordtitle
winchester	povertyrow	onemanarmy	crime	bracelet

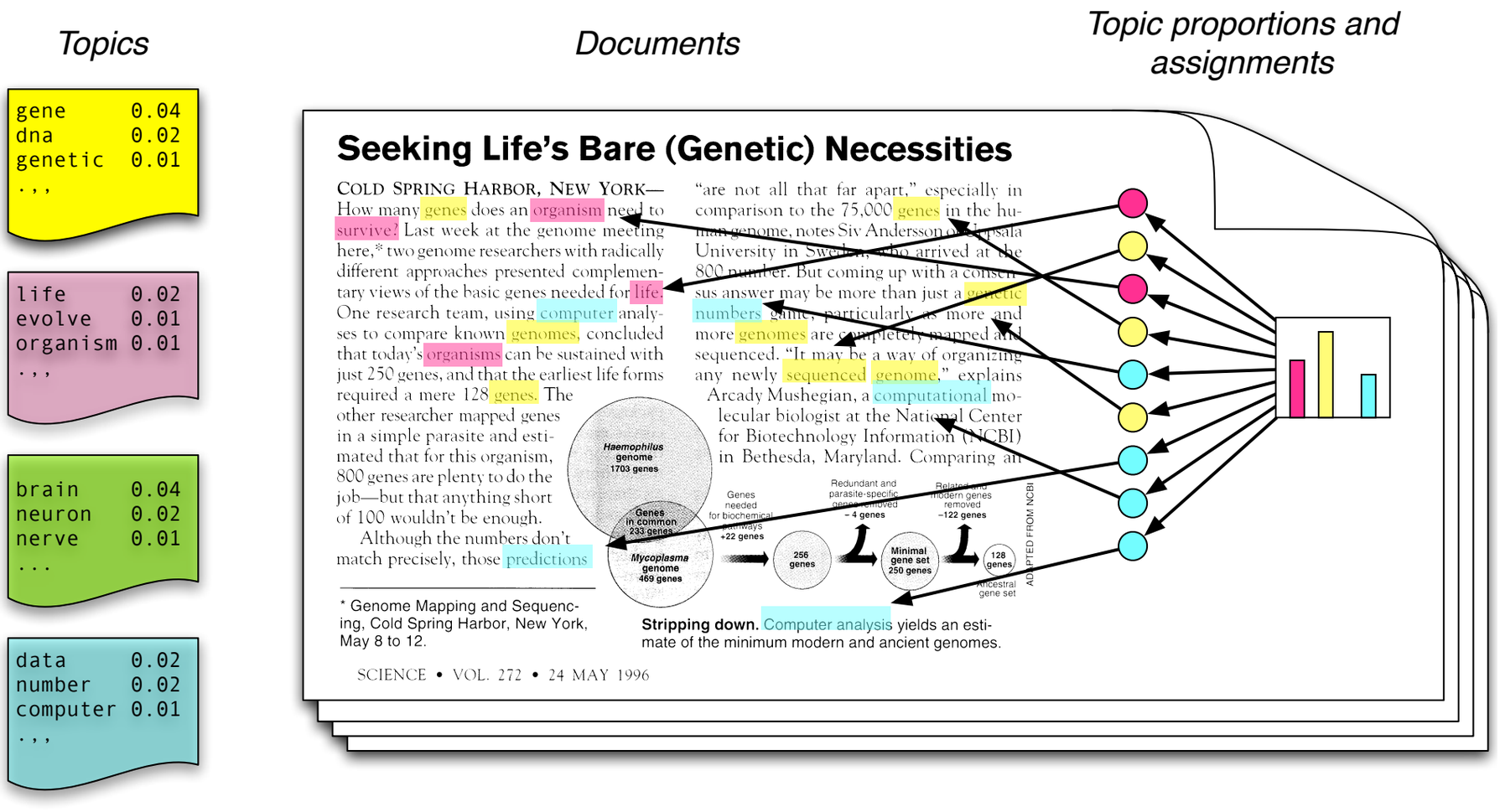
Experiment: IMDB Dataset

- Documents constructed from IMDB Keywords for set of 21k movies.
- Applied NMF (k=10) to 20,923 x 5,528 movie-keyword matrix.
- Topic “descriptions” based on top ranked keywords in basis vectors appear to reveal genres and genre cross-overs.

Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
worldwartwo	monster	love	newyorkcity	shotinthechest
soldier	alien	friend	manhattan	shottodeath
battle	cultfilm	kiss	nightclub	shotinthehead
army	supernatural	adultery	marriageproposal	punchedintheface
1940s	scientist	infidelity	jealousy	corpse
nazi	surpriseending	restaurant	engagement	shotintheback
military	demon	extramaritalaffair	party	shotgun
combat	occult	photograph	hotel	shotintheforehead
warviolence	possession	tears	deception	shotintheleg
explosion	slasher	pregnancy	romanticrivalry	shootout

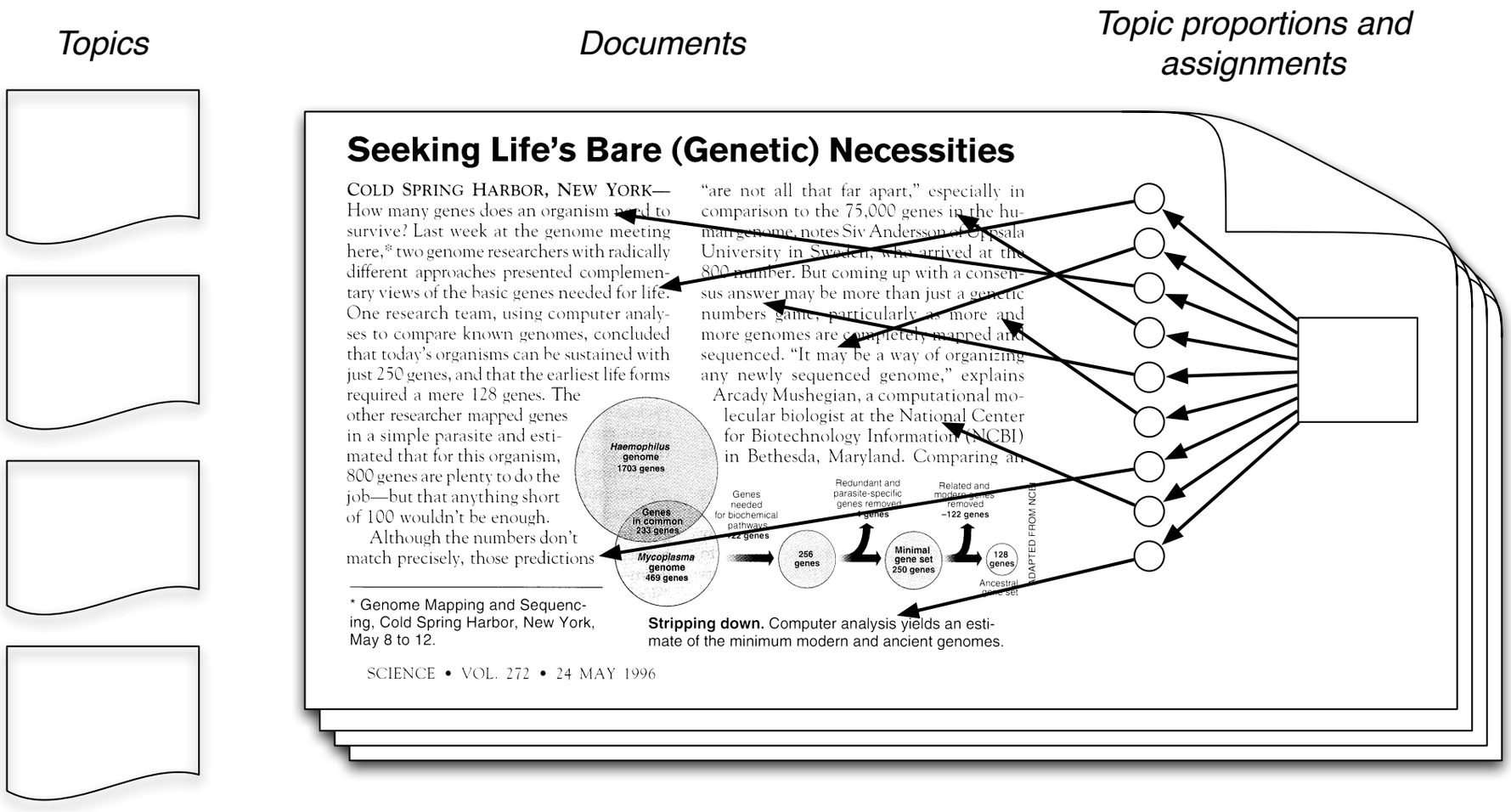
A brief overview of the key probabilistic models

Typical probabilistic model



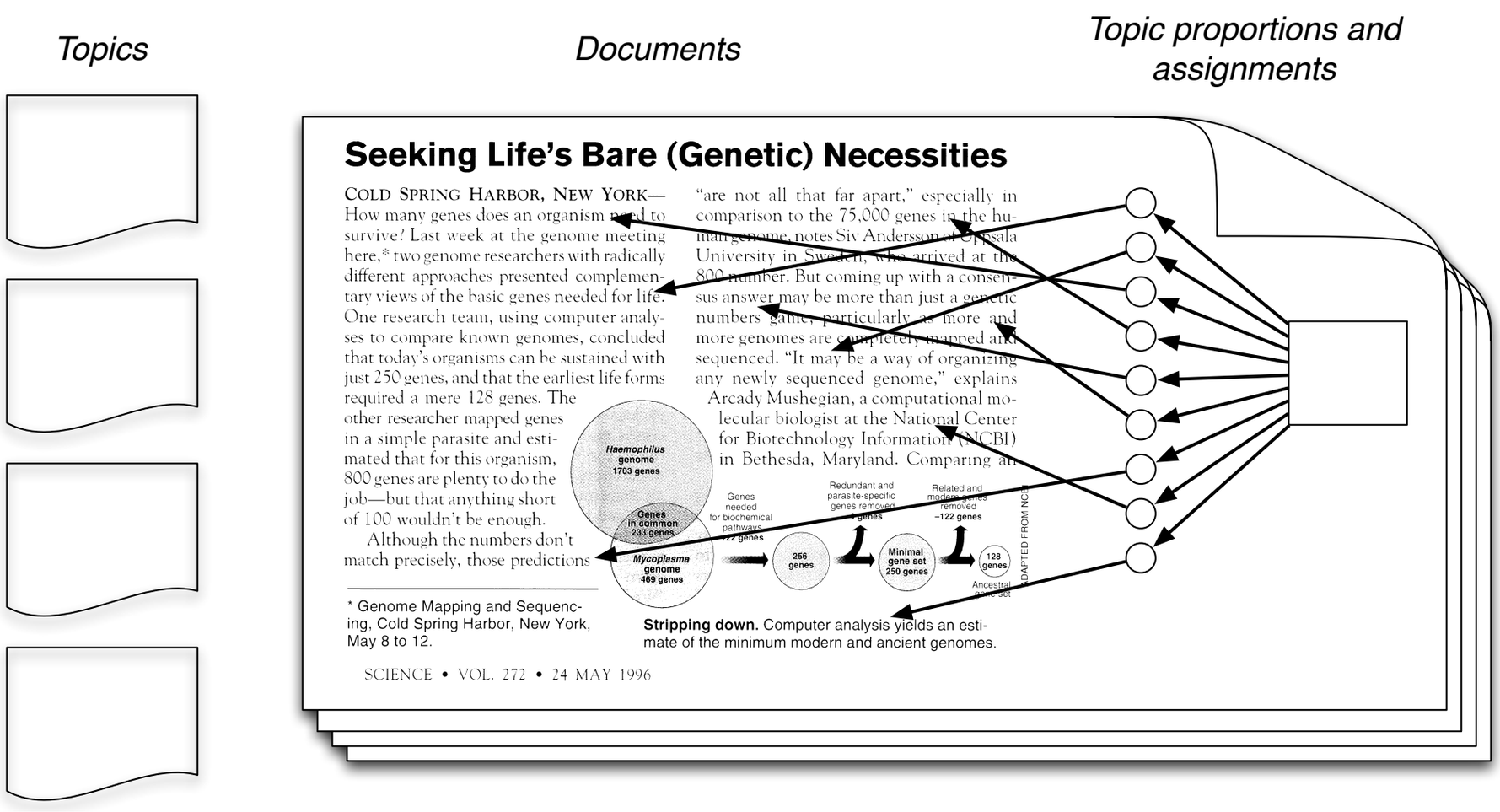
- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

Typical probabilistic model



- In reality, we only observe the documents
- The other structures are **hidden variables**

Typical probabilistic model



- Our goal is to **infer** the hidden variables
  - i.e. compute their distribution conditioned on the documents

$$p(\text{topics, proportions, assignments} \mid \text{documents})$$

Probabilistic Latent Semantic Analysis

- Given a corpus, observations produced in the form of pairs of words and documents (*w,d*)

- Each observation is associated with an unobserved latent class variable, *c*

- PLSA model assumes that the probability of a co-occurrence *P(w,d)* is a mixture of conditionally independent *multinomial distributions*:

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

Latent Dirichlet Allocation

- ***Popular* Bayesian extension to PLSA**
  - Add a Dirichlet prior on the per-document topic distribution
  - Makes the model **fully generative** (i.e. we can sample new documents)
- Parameters must be learned using **Bayesian inference** (e.g. variational Bayes/Gibbs Sampling/etc)
- In practice: better than PLSA for small datasets; with lots of data tends to perform similarly

Experiment: Science Articles

- **Data**: The OCR'ed collection of Science from 1990–2000
  - 17K documents, 11M words, 20K unique terms (stop words and rare words removed)
- **Model**: 100-topic LDA model using variational inference.

Summary

- Topic modelling as an important part of data mining of unstructured data
    - Key idea is that documents/items belong to or are made up of a number of topics
        - Typically a small subset of the overall set of topics
- All the models we've looked at have 1 key parameter that must be manually tuned: the number of topics