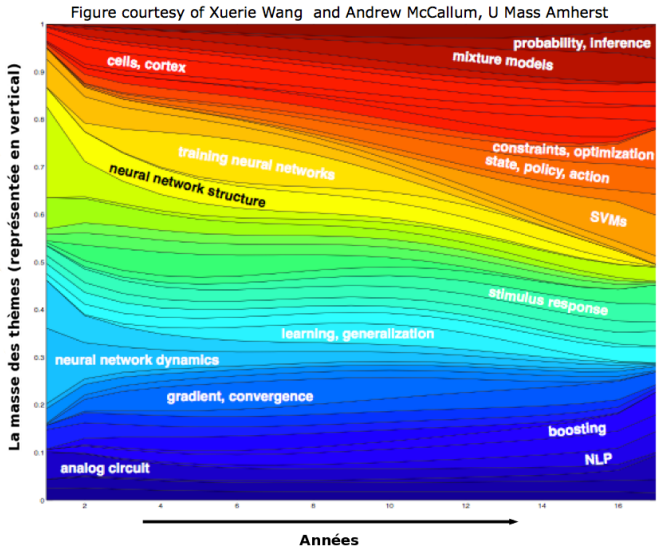
Predictive topic modeling : Complex Networks approach using dynamics of author's communities
Application on health corpora (QALY)

Karim Sayadi
karim.sayadi@etu.ephe.fr

(a) Annotate according to the topics

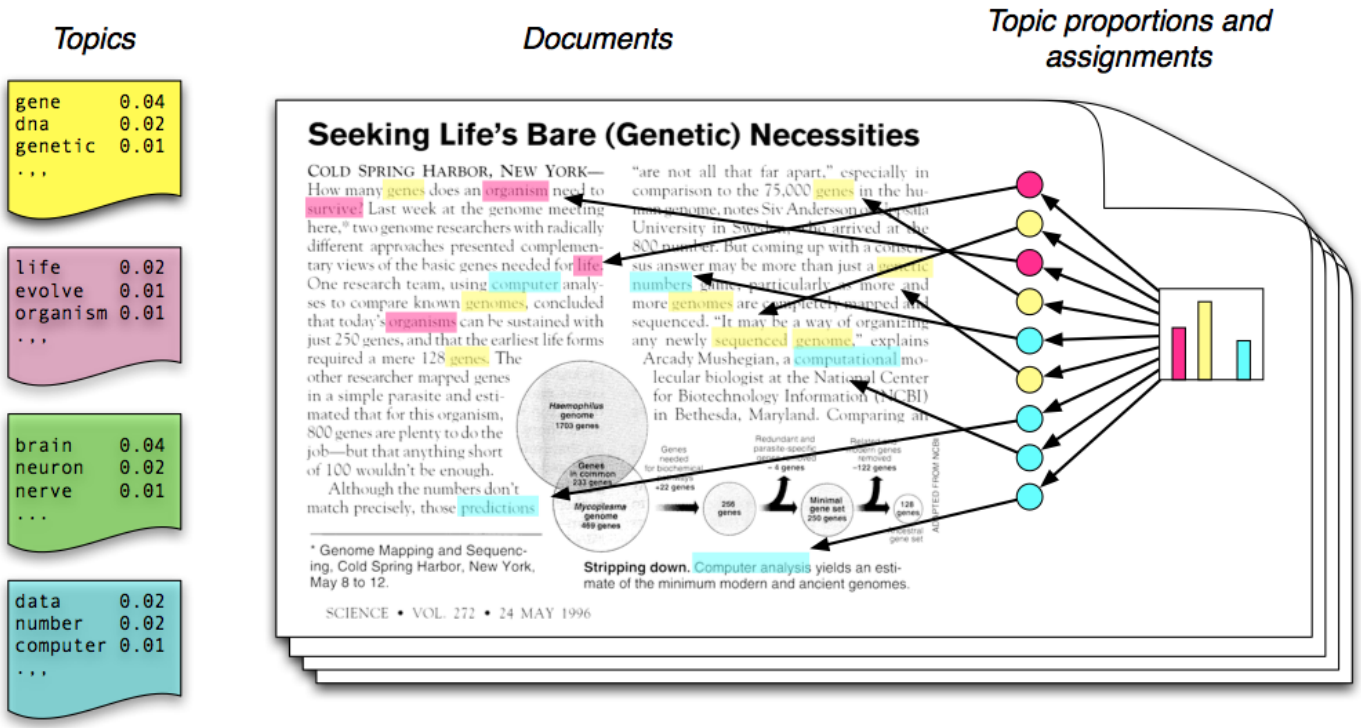


(b) Topics over time

Plan

1. Probabilistic Graphical Models for topic modeling
   - The main problem
   - Latent Dirichlet Allocation method
2. LDA implementation and parameters estimation
   - Gibbs Sampling algorithm
3. Contribution and application
   - The Pre-topological Semantic Analysis algorithm
   - Classification and visualization of the corpora
4. Conclusion and futur directions

Outline

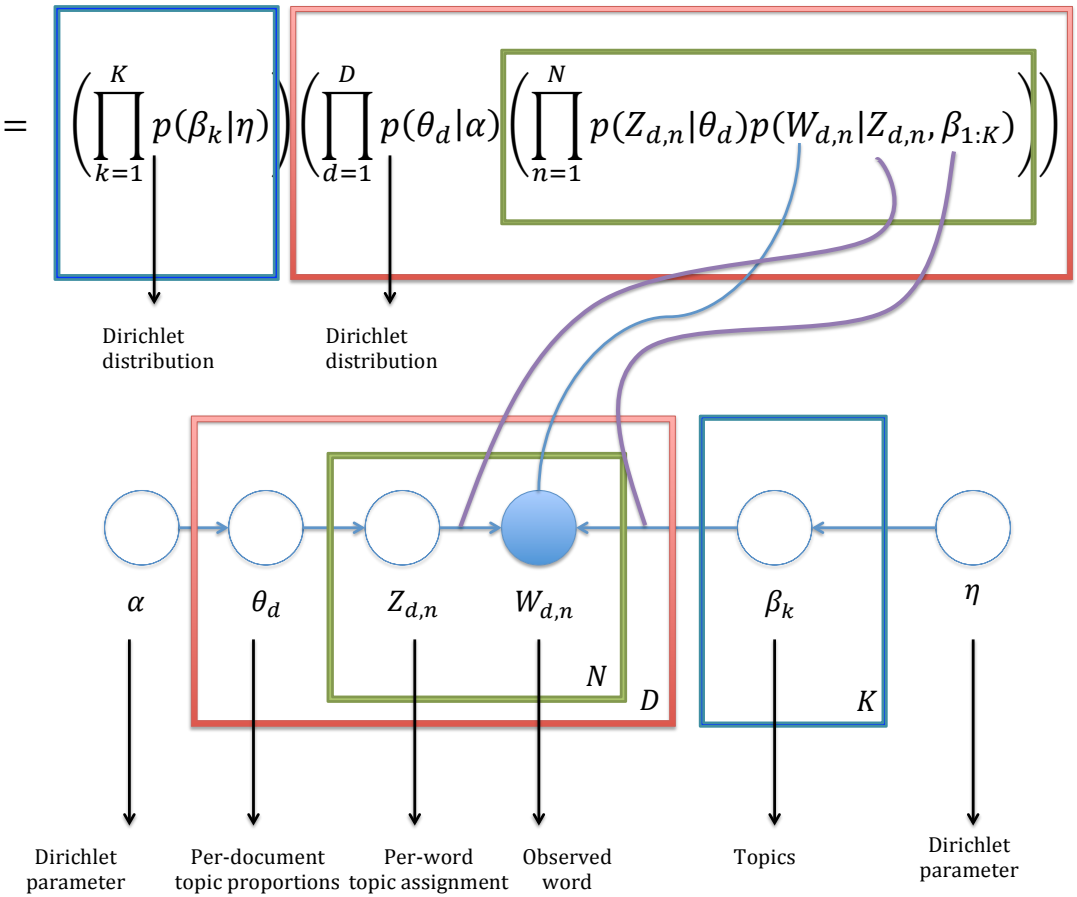


- A topic ? $\rightarrow$ probabilistic distribution over words
- A document ? $\rightarrow$ probabilistic distribution over topics
- ☞ Probabilistic Model
- Method : Latent Dirichlet Allocation [1]

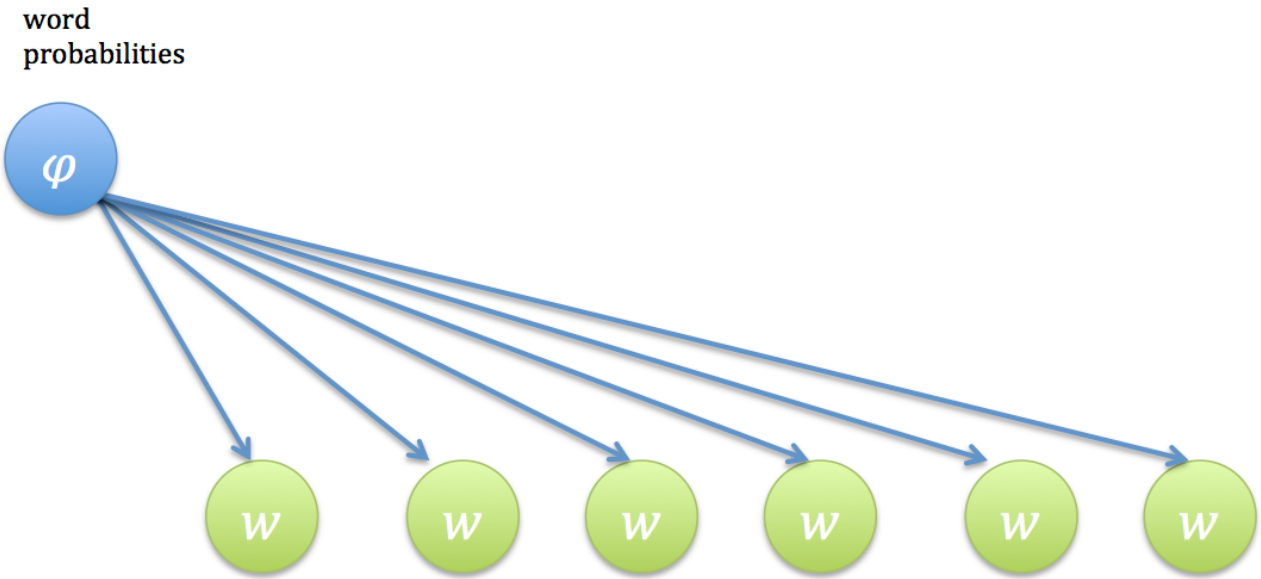
1. D. M. BLEI, M. I. Jordan A.Y. NG et J. LAFFERTY. "Latent dirichlet allocation". In : *Journal of Machine Learning Research* 993-1022.3 (2003)

The generative process

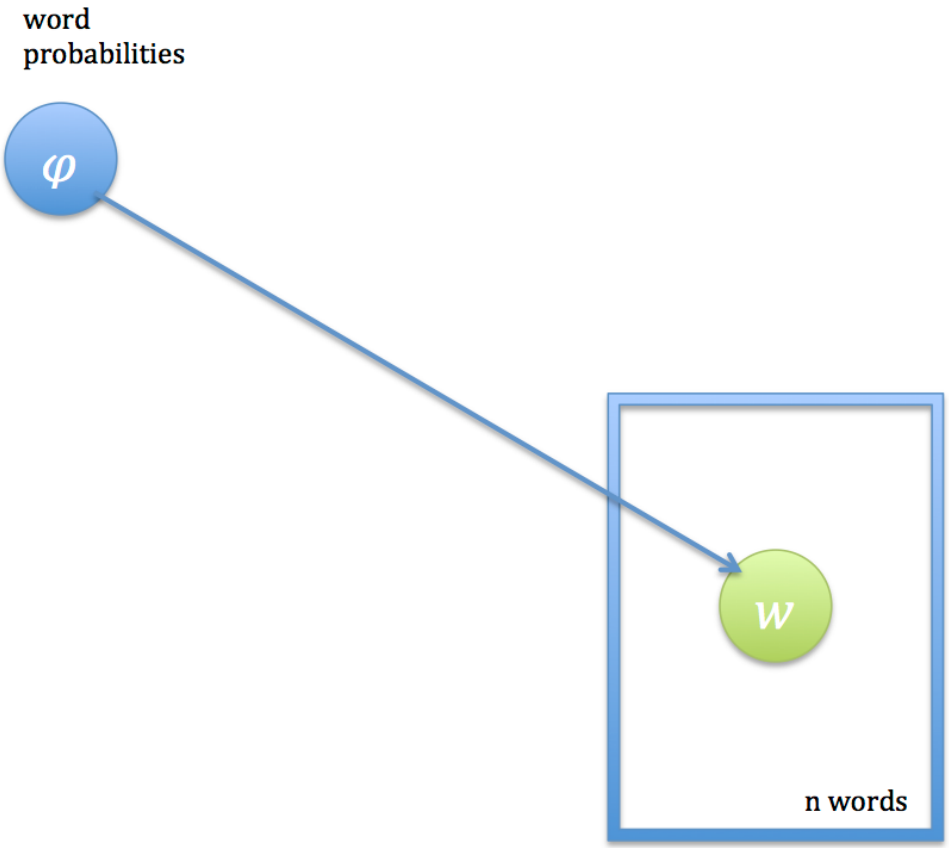
$$p(\beta_k, \eta, \theta_d, \alpha, Z_{d,n}, W_{d,n}) = \left(\prod_{k=1}^{K} p(\beta_k|\eta)\right)\left(\prod_{d=1}^{D} p(\theta_d|\alpha)\left(\prod_{n=1}^{N} p(Z_{d,n}|\theta_d)p(W_{d,n}|Z_{d,n}, \beta_{1:K})\right)\right)$$



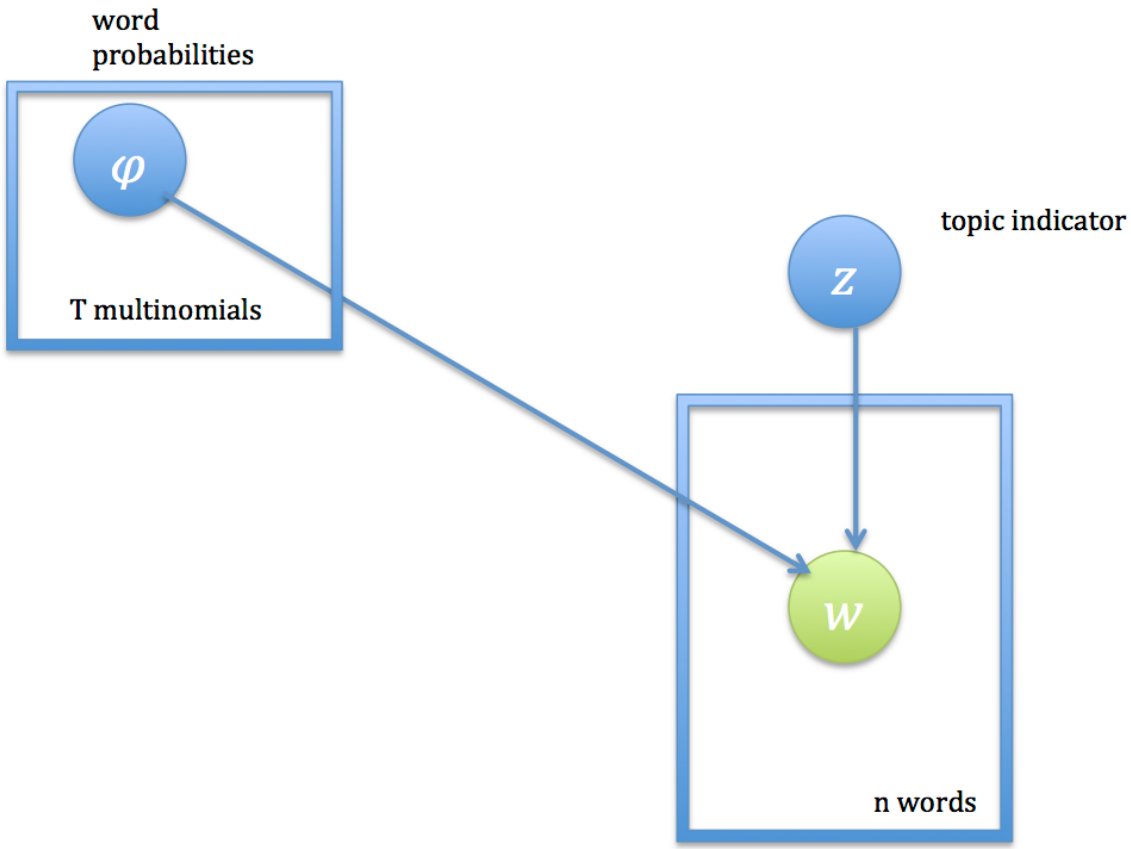
Mixture Models : One Topic per Document



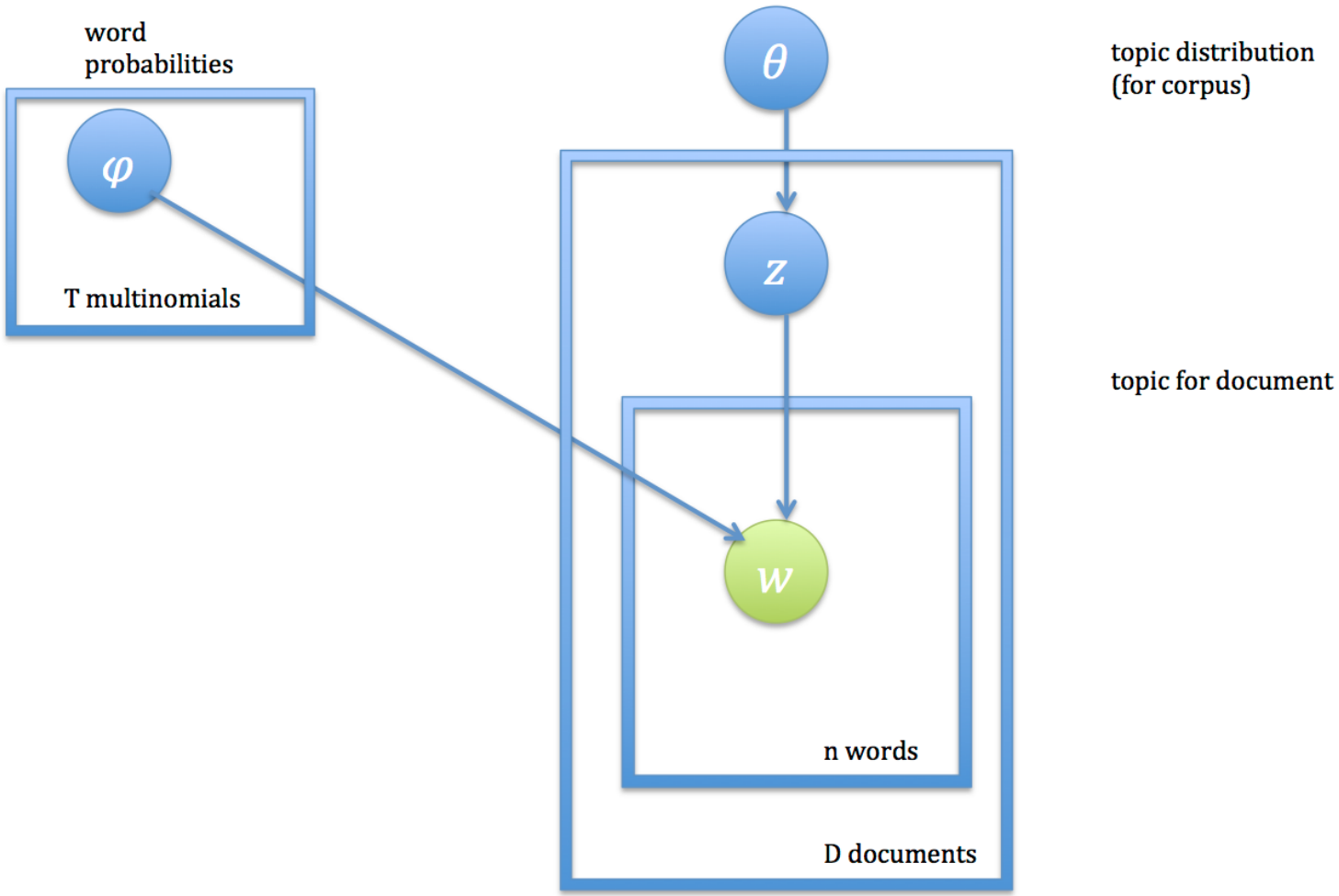
Mixture Models : One Topic per Document



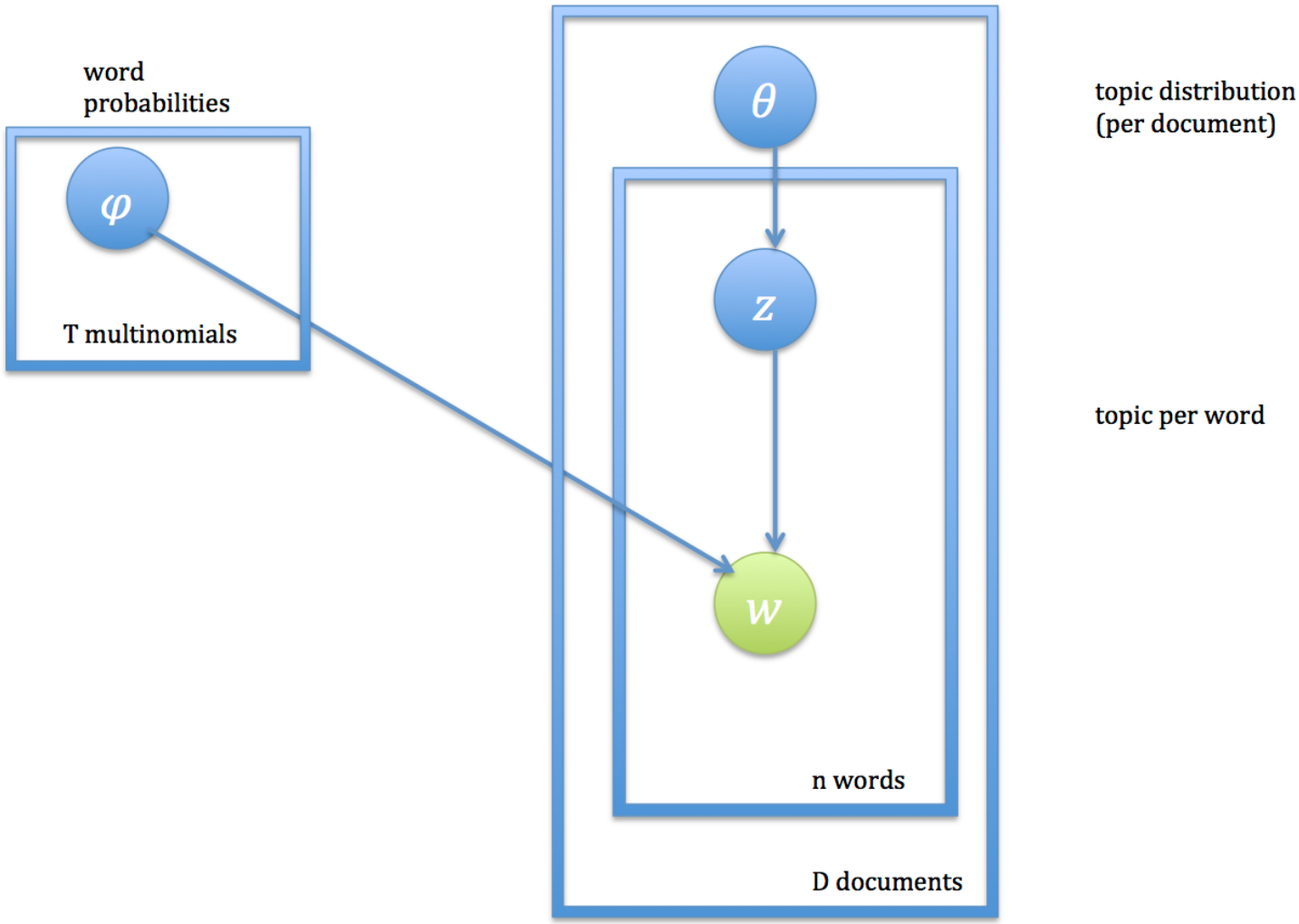
Mixture Models : One Topic per Document



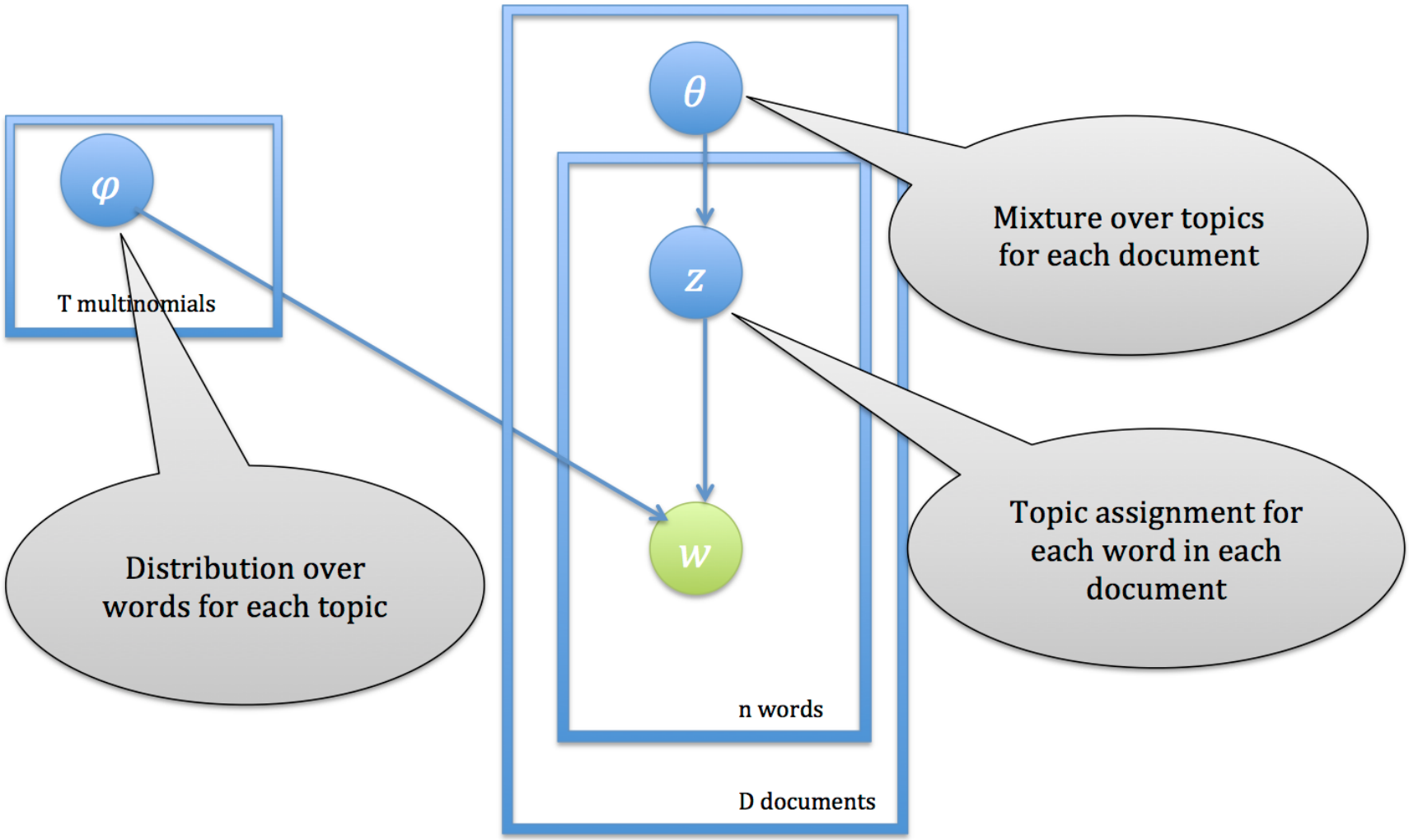
Mixture Models : One Topic per Document



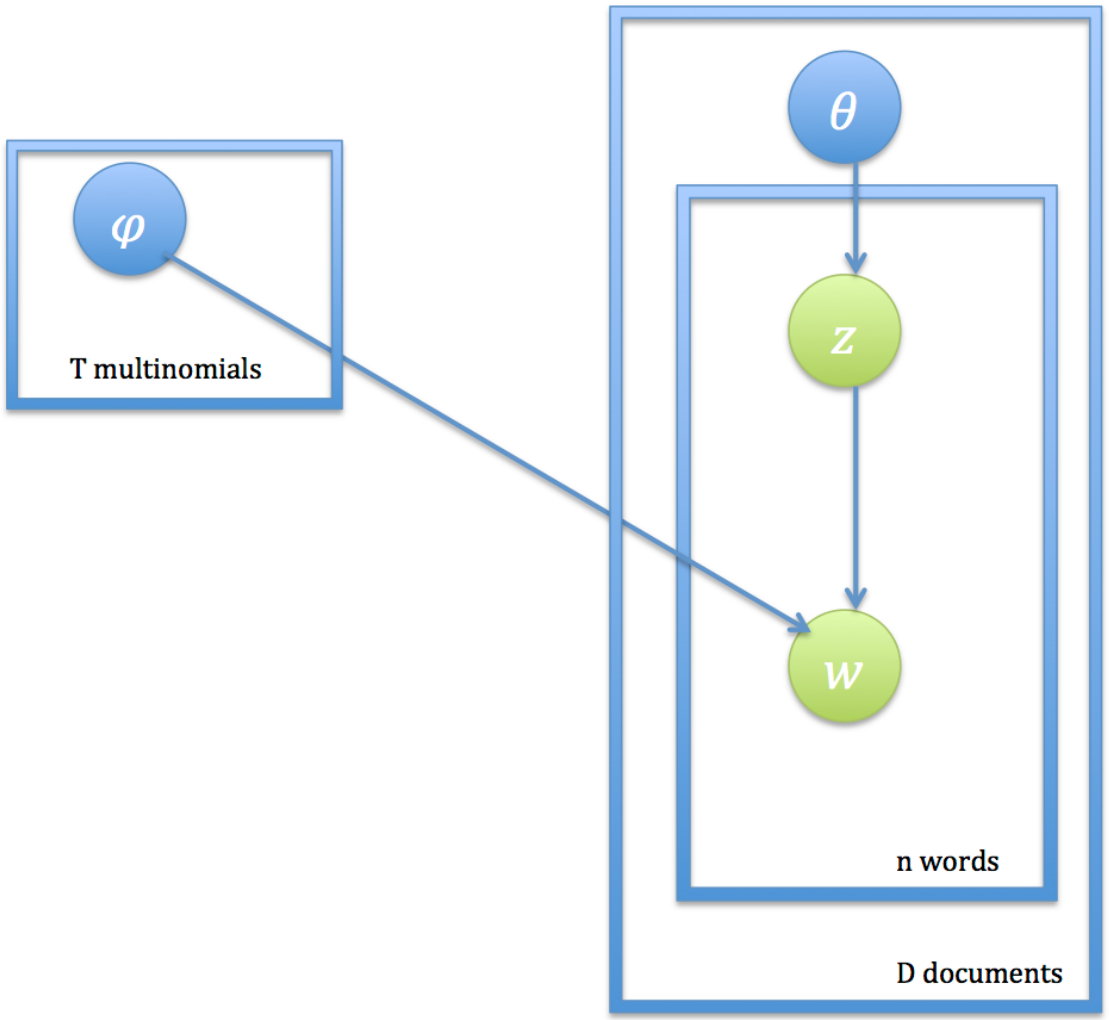
Topic Model Documents = Mixtures of Topics



What do we need to learn ?



What do we need to learn ?



Gibbs Sampling :
Sample $P(z \mid data, priors)$.
Marginalizing over θ and φ

Outline

Learning algorithm for LDA

Input

- N documents, each as "*bag of words*"
- Number of topics T

Learning via Collapsed Gibbs Sampling

- Sample z's, marginalize over θ and ϕ
- Given z's, easy to estimate θ and ϕ

Output

- ϕ : Topic-word probability distributions for each topic
- θ : Document-topic probability distributions
- z's : Assignment of each word in each doc to a topic

The diagram of LDA's implementation components

Outline

PSA algorithm : LDA and a pre-topological operator

- Red points represent the author's communities writing on a particular topic.
- Yellow points represent the authors that are becoming a part of the communities.

Informal description of the algorithm PSA

Input

- A corpora E with a set of documents.
- T lists of the authors of different scientific articles.

Detection of the author's communities and the construction of the network

- Topic and author's list extraction for the corpora E.
- Link the different documents depending on the topic that they express and the communities where they are produced.

Output

- Z sets of topics.
- Scale free network modeling the interaction between the communities depending on the topics that they write in.

Most likely words from top topics

topic 1	topic 2	topic 3	topic 4	topic 5
fracture 0.0881	copd 0.0165	estimation 0.1211	care 0.3200	vaccin 0.1444
women 0.0462	asthma 0.0165	data 0.0516	health 0.1893	cost 0.0603
year 0.0390	pulmonary 0.0113	time 0.0376	expenditure 0.0190	effect 0.0337
osteoporosis 0.0269	respiratory 0.0105	study 0.0273	resource 0.0164	hpv 0.0226
treatment 0.0265	omalizumab 0.0093	sample 0.0223	resources 0.0164	influenza 0.0178
hips 0.0250	lung 0.0073	statistics 0.0164	insure 0.0084	prevent 0.0163
estrogen 0.0221	quality 0.0069	regression 0.0109	spend 0.0066	age 0.0132
bone 0.0144	allergies 0.0065	observation 0.0097	justify 0.0044	observation 0.0097
denosumab 0.0122	allergy 0.0065	bayesian 0.0069	outcome 0.0019	immun 0.0129
postmenopaus 0.0103	epidemiology 0.0048	covariance 0.0069	evaluation 0.0019	rotavirus 0.0107
raloxifen 0.007	persist 0.0048	error 0.0066	payer 0.0014	dose 0.0091

TABLE: Results of running the implementation of LDA with 999 iteration of the Gibbs sampling algorithm. Value of the hyper-parameters are $\alpha = 0.5$ and $\beta = 0.1$

Visualization of the scale free Documents and authors communities network.

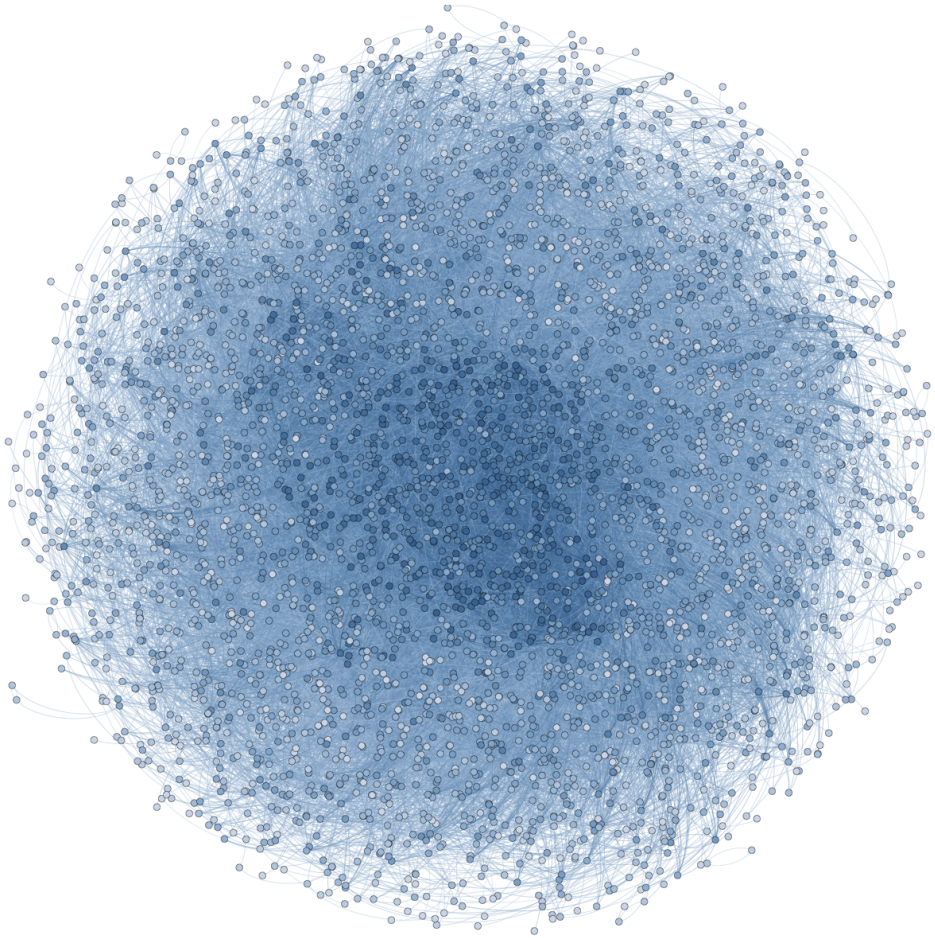


FIGURE: Scale free network constructed from the topics and the authors in QALY corpora. Communities of authors who write on the same subject are in dark blue. The nodes represent documents, while the connexion between them are established flowing the communities of authors and the topics that they write on.

Outline

Conclusion

- The extracted topics capture meaningful structure in the data
  $\rightarrow$ Consistent with the key words provided by the authors.
- Our objectif : Improve the algorithm efficiency to deal with the dynamics corpora that grows over the time.
  - ☞ Parallel solution for a distributed computation

Thanks for your attention