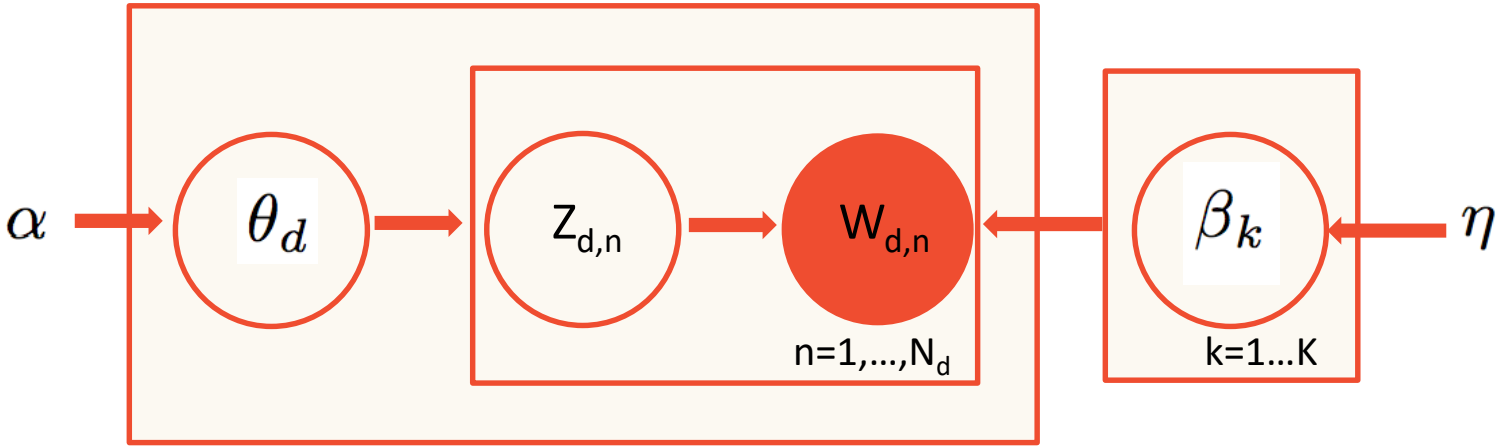
Latent Dirichlet Allocation (LDA)

A Bayesian Unsupervised Learning Model

$$
\begin{aligned}
Z_{d,n}|\theta_d &\sim Multinomial(\theta_d) \\
W_{d,n} \mid Z_{d,n},\ \beta &\sim Multinomial(\beta_{Z_{d,n}}) \\
\beta_k|\eta &\sim Dirichlet(\eta) \\
\theta_d|\alpha &\sim Dirichlet(\alpha)
\end{aligned}
$$



Agenda Items

- LDA – introduction & some notations (graphical models)
- LDA – theory and intuition
- LDA – application
- LDA – extensions
- References

LDA – Introduction & some notations (graphical models)

What is Latent Dirichlet Allocation (LDA)?

- ***Latent Dirichlet Allocation (LDA)*** is a topic model, which was first proposed by Blei et al. in 2003.
- ***Topic models*** are "[probabilistic] latent variable models of documents that exploit the correlations among the words and latent semantic themes" (Blei and Lafferty, 2007)
- The name "topics" signifies the hidden, to be estimated, variable relations (=distributions) that link words in a vocabulary and their occurrence in documents.
- *Simple intuition*: A document is seen as a mixture of topics.

So, what is LDA?

- **It's a way of automatically discovering topics that documents contain.**
- **LDA represents documents as mixtures of topics that spit out words with certain probabilities**.

LDA – First example

Suppose you have the following set of sentences (or documents):

1. I like to drink coffee.
2. Colombian coffee beans are better than Ethiopian.
3. Machine learning algorithms and reporting are data scientist's common tasks.
4. A member of DS team created a report for the daily consumption of coffee in Profusion.

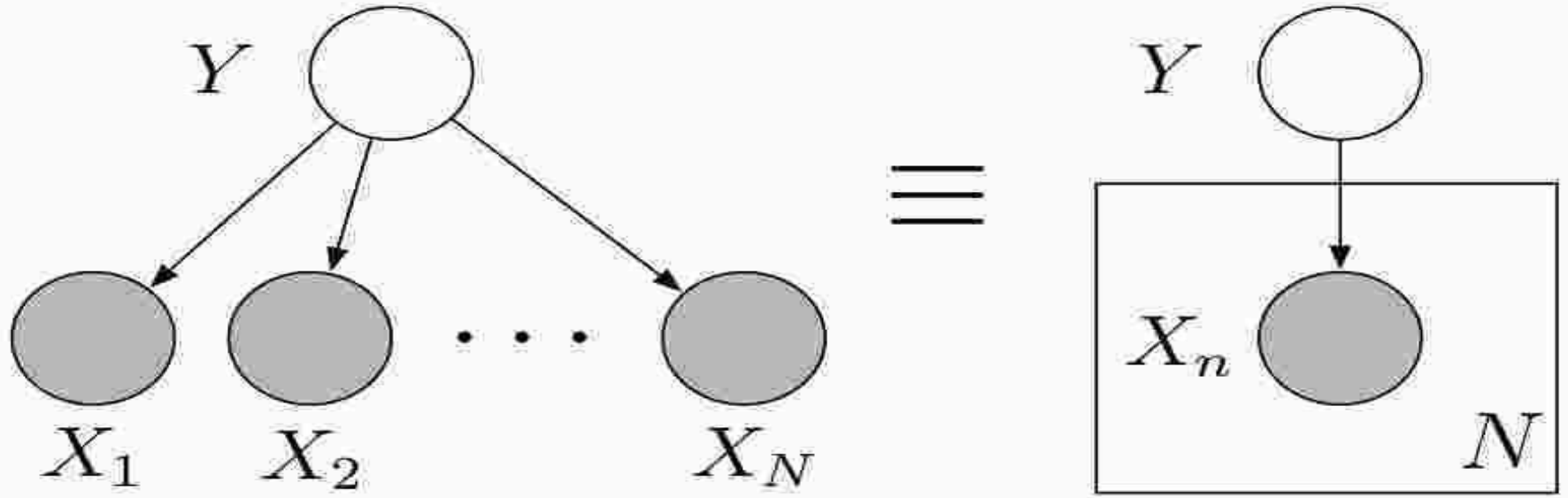
For example, given these sentences and asked for 2 topics, LDA might produce something like:

- **Topic A**: 50% coffee, 15% bean, 10% Colombian, 10% Ethiopian, … (at which point, you could interpret topic A to be about **coffee lovers**)
- **Topic B**: 20% machine, 20% learning, 20% reporting, 15% data, … (at which point, you could interpret topic B to be about **data science**)

- **Sentences 1 and 2**: 100% Topic A
- **Sentences 3**: 100% Topic B
- **Sentence 4**: 60% Topic A, 40% Topic B

Graphical Model

Introduction



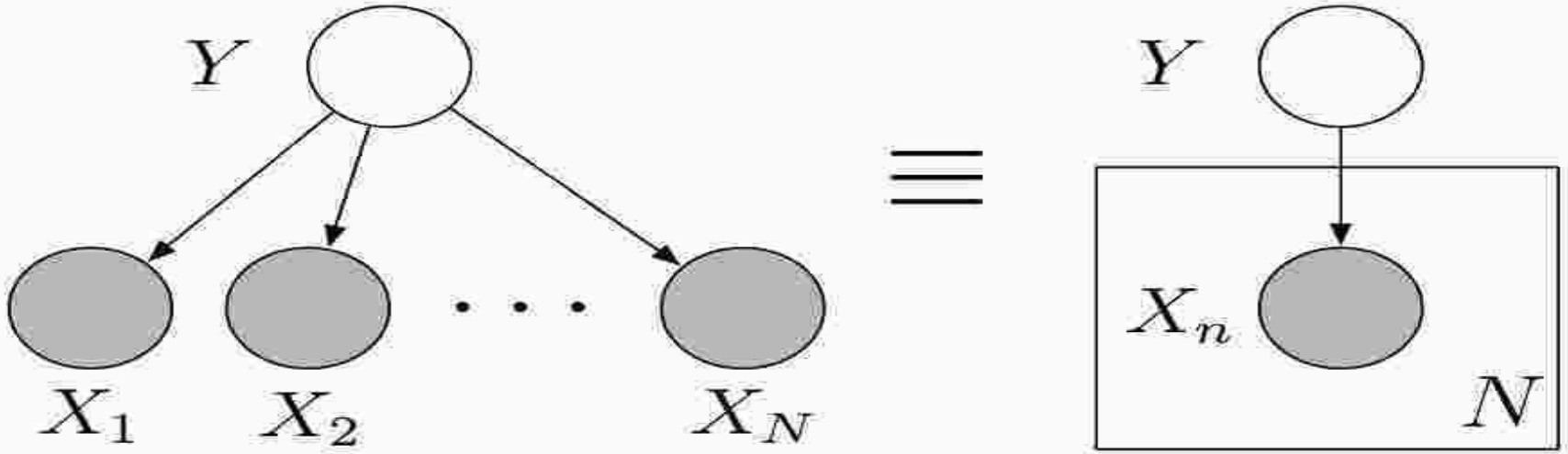
Nodes are random variables

Edges denote possible dependence

Observed variables are shaded

Plates denote replicated structure

Graphical models



Structure of the graph defines the pattern of conditional dependence between the ensemble of random variables

E.g., this graph corresponds to

$$p(y, x_1, \ldots, x_N) = p(y) \prod_{n=1}^{N} p(x_n \mid y)$$

LDA – theory and intuition

LDA – From scratch

How are words in a document generated?

One possibility:

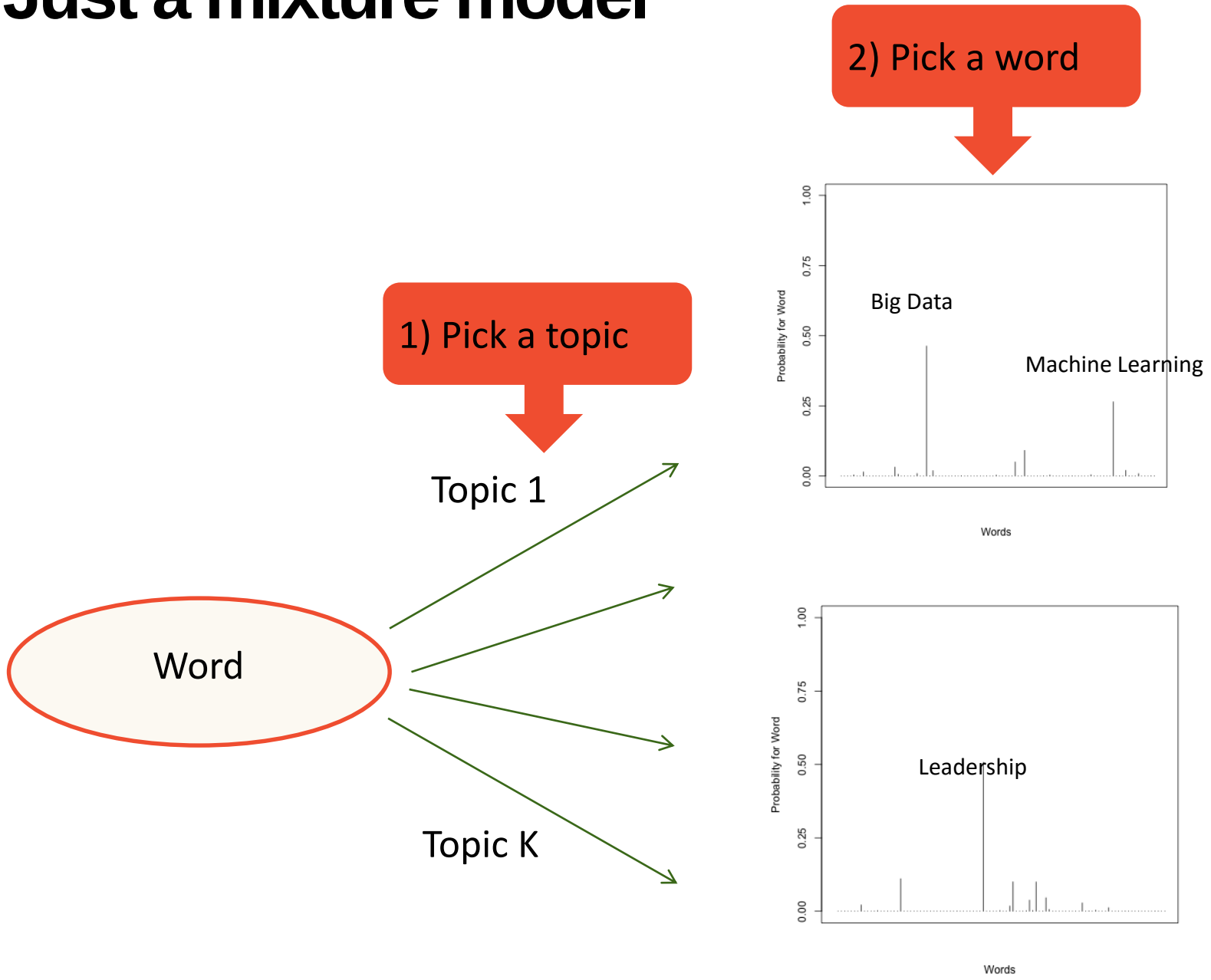
Each word comes from different topics (bag of words: ignore order)

$$P(word) = \sum_{k=1}^{K} P(Topic_k) P(word|Topic_k)$$

Mixture Weight for Topic k

Multinomial Distribution over ALL words based on topic k

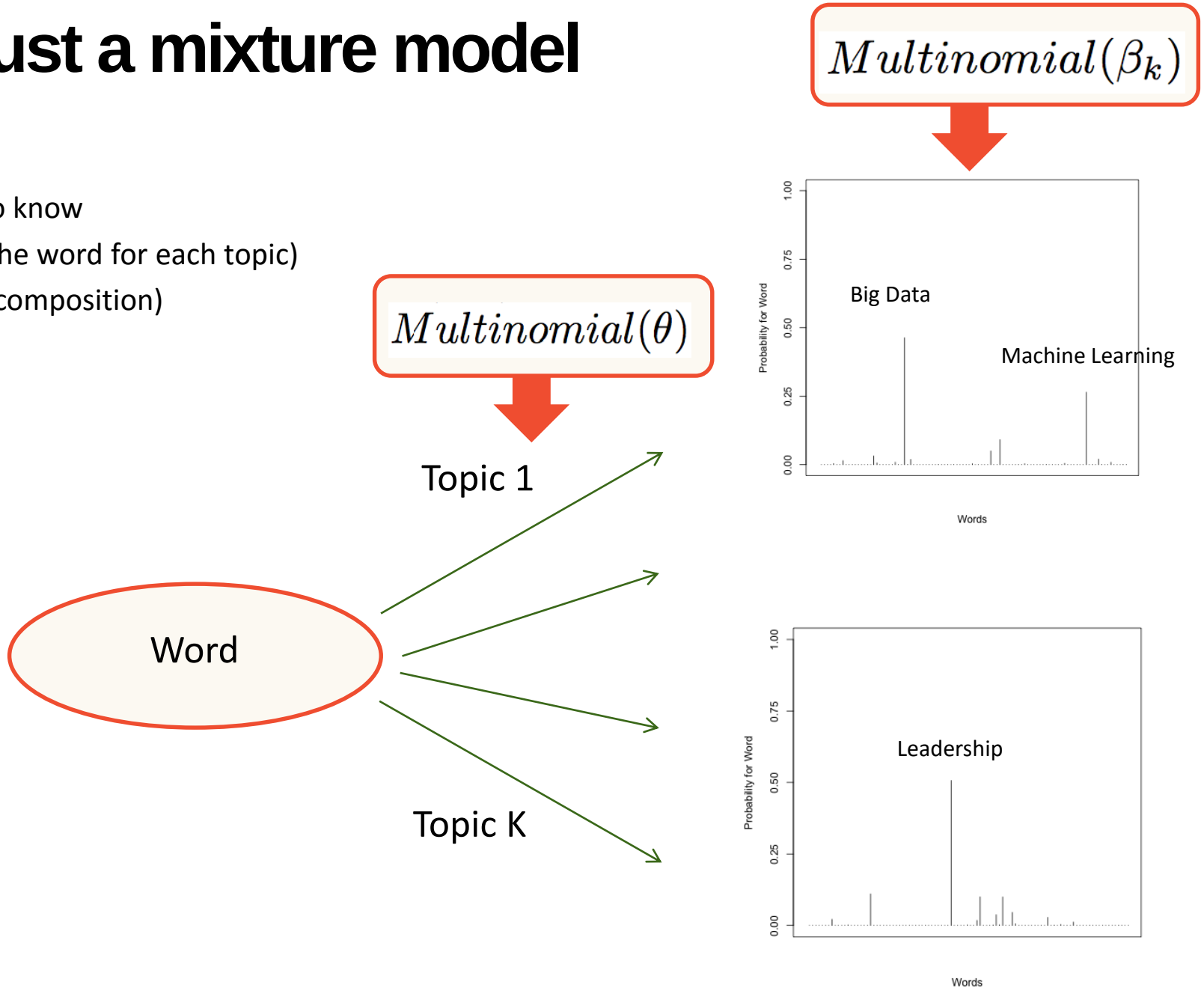
LDA - Just a mixture model



LDA - Just a mixture model

So we really want to know

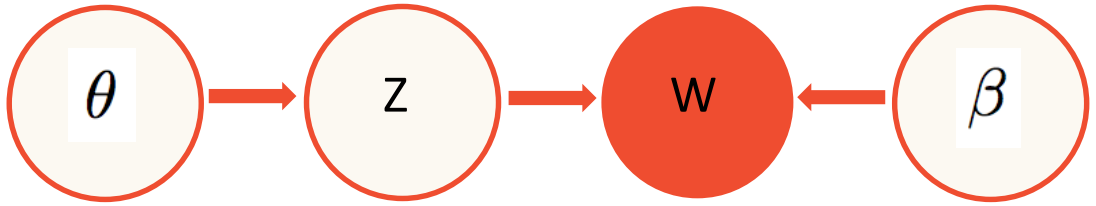
1) Z (cluster for the word for each topic)
2) θ (document composition)
3) β (key words)



Graphical representation

1 word from 1 document generative process

$$
\begin{aligned}
Z \mid \theta &\sim Multinomial(\theta) \\
W \mid Z, \beta &\sim Multinomial(\beta_Z)
\end{aligned}
$$



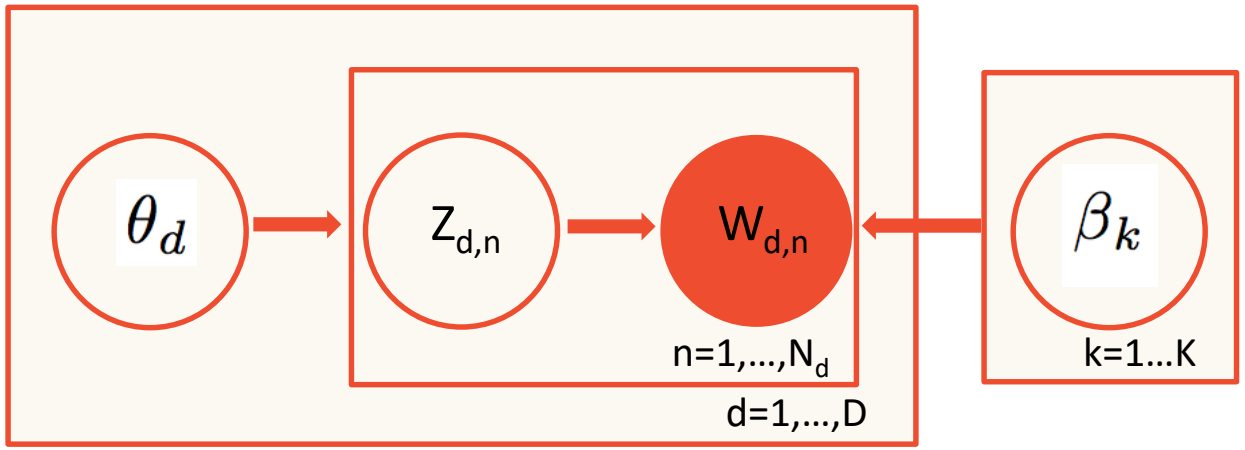
More general

K: number of topics

N_d: number of words

D: number of documents

$$Z_{d,n}|\theta_d \sim Multinomial(\theta_d)$$
$$W_{d,n} \mid Z_{d,n}, \beta \sim Multinomial(\beta_{Z_{d,n}})$$



Thinking Bayesian

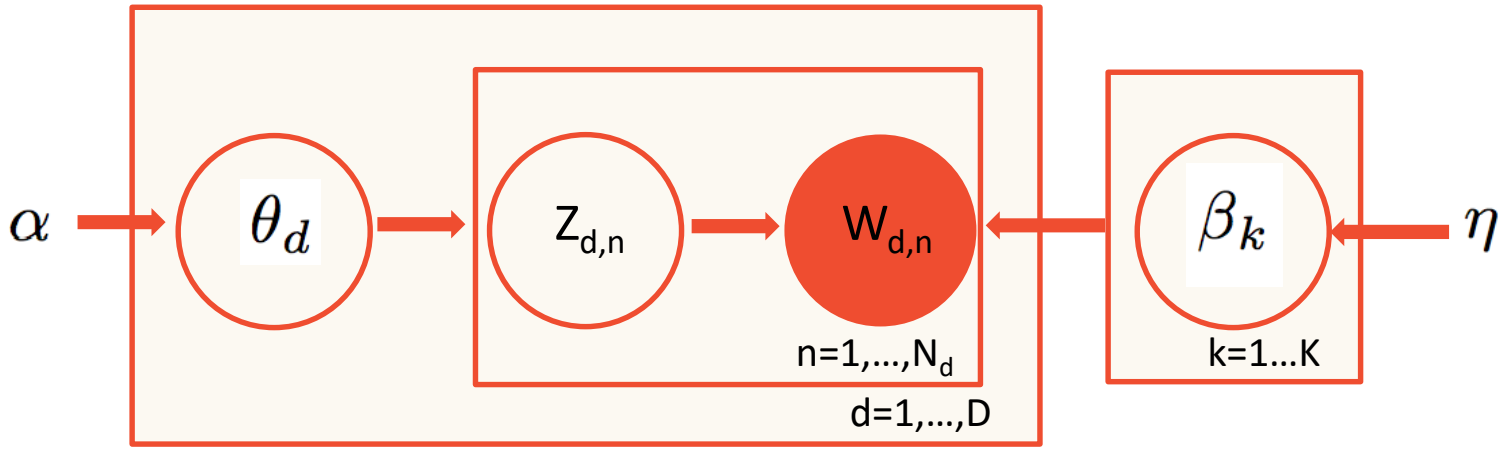
Bayesian: But what about the distribution for θ_d and β_k ??

K: number of topics

N_d: number of words

D: number of documents

$$
\begin{aligned}
Z_{d,n}|\theta_d &\sim Multinomial(\theta_d) \\
W_{d,n} \mid Z_{d,n},\ \beta &\sim Multinomial(\beta_{Z_{d,n}}) \\
\beta_k|\eta &\sim Dirichlet(\eta) \\
\theta_d|\alpha &\sim Dirichlet(\alpha)
\end{aligned}
$$



LDA as hierarchical Bayesian model

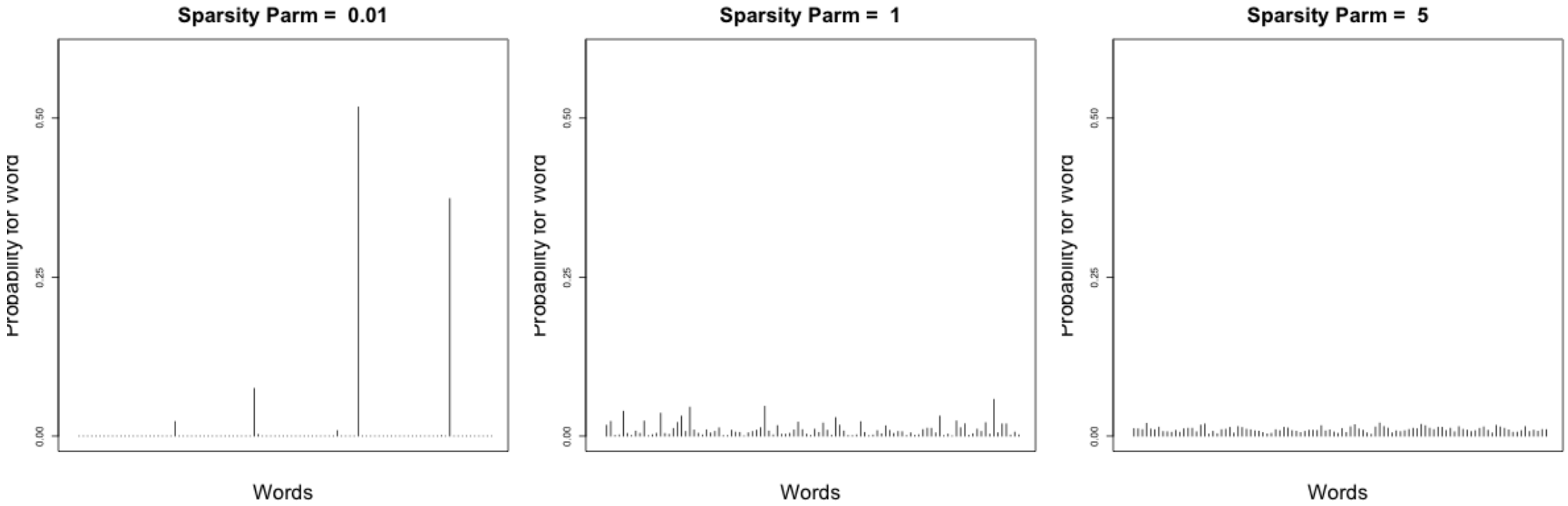
$$
\begin{aligned}
Z_{d,n}|\theta_d &\sim Multinomial(\theta_d) \\
W_{d,n} \mid Z_{d,n},\ \beta &\sim Multinomial(\beta_{Z_{d,n}}) \\
\beta_k|\eta &\sim Dirichlet(\eta) \\
\theta_d|\alpha &\sim Dirichlet(\alpha)
\end{aligned}
$$

α and η control the "sparsity" of the weights for the multinomial.

Implications: a priori we assume

- Topics have few key words
- Documents only have a small subset of topics

Dirichlet Distribution with Different Sparsity Parameters



Calculate the posterior

How do we fit this model?

$$
\begin{aligned}
Z_{d,n}|\theta_d &\sim Multinomial(\theta_d) \\
W_{d,n} \mid Z_{d,n},\ \beta &\sim Multinomial(\beta_{Z_{d,n}}) \\
\beta_k|\eta &\sim Dirichlet(\eta) \\
\theta_d|\alpha &\sim Dirichlet(\alpha)
\end{aligned}
$$

Want the posterior:

$$P\left(\theta_{1,\ldots D}, \beta_{1,\ldots,K}, Z_{1,\ldots,D;1,\ldots N} \mid W_{1,\ldots,D;1,\ldots N}\right)$$

Worst part of Bayesian Analysis…..personally speaking

Computing the posterior

$$P\left(\theta_{1,\ldots D}, \beta_{1,\ldots,K}, Z_{1,\ldots,D;1,\ldots N} \mid W_{1,\ldots,D;1,\ldots N}\right)$$

- Mean field variational methods (Blei et al., 2001,2003)
- Expectation propagation (Minka and Lafferty, 2002)
- Collapsed Gibbs sampling (Griffiths and Steyvers, 2002)
- Collapsed variational inference (The et al., 2006)

Application

LDA in the R Environment

R packages for LDA

1. **tm**
2. **topicmodels**
3. **lda**

1. Introductions and examples to **topicmodels** are available from:

- the package reference documentation which can be viewed via the R help system, or as a single PDF file (topicmodels.pdf at http://cran.r-project.org/package=topicmodels), or
- the authors' paper in the Journal of Statistical Software (Grun and Hornik, 2011).

For data input the topicmodels package depends on the package tm (text mining)

2. The current functionality of **tm** is documented in detail in the package's online help and in its reference manual (tm.pdf at http://cran.r-project.org/package=tm). tm is handling text mining tasks. It enables the user to import data from various formats and offers standard tools for pre-processing and further transformation and filtering of texts and text collections. (removal of punctuation, numbers, whitespace and stop words, conversion to lower case, stemming and so on.)

3. **lda** package implements latent Dirichlet allocation (Lda) and related models. This includes (but is not limited to) sLda [(supervised Lda, see Blei and McAuliffe (2007))], [related topic models (Rtm, see Chang and Blei (2009))], and the mixed-membership stochastic blockmodel [(Mmsb, see Airoldi et al. (2008))].

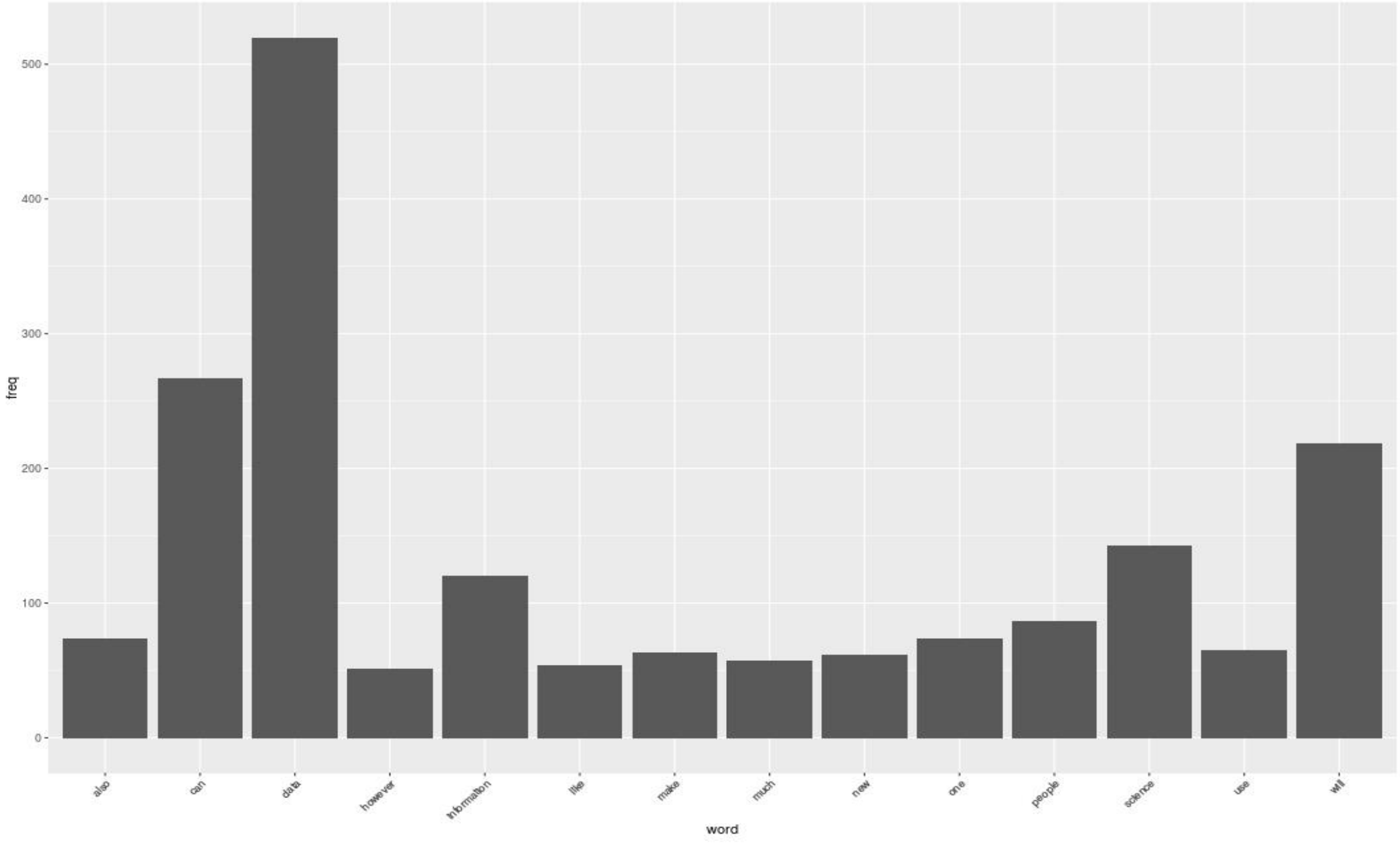
Application - Profusion blogs

59 blogs (15 June 2014 – 31 March 2016)

Title	Timestamp
Out-smarting the ad fraudsters	Posted on March 31, 2016
The smart future of marketing	Posted on March 24, 2016
My quantified self	Posted on March 17, 2016
Building better construction through data science	Posted on March 10, 2016
Does data science exist?	Posted on March 3, 2016
Intelligent infrastructure in a smart city	Posted on February 25, 2016
Wearables for Charities and NGOs	Posted on February 18, 2016
Wearables taking flight	Posted on February 11, 2016
Swapping Safe Harbour for a Shield	Posted on February 4, 2016
Smart local authorities â€“ Data, Austerity, Resources and Devolution	Posted on January 28, 2016
It pays to be different in Data Science	Posted on January 21, 2016
How data will shape finance in 2016	Posted on January 14, 2016
Lighting up with the IoT	Posted on January 7, 2016
How much is a Christmas advert worth?	Posted on December 31, 2015
Santa is a data scientist	Posted on December 24, 2015
Bayesian statistics and â€¦ the sunrise!	Posted on December 17, 2015
Data science can prove PRâ€™s worth	Posted on December 10, 2015
Smarter testing in email marketing	Posted on December 3, 2015
Healthy use of data in occupational health	Posted on November 26, 2015
Uncovering your customersâ€™ true desires	Posted on November 20, 2015
Life without internet	Posted on November 18, 2015
A statistics walk-through	Posted on November 12, 2015
Data science can make banking more personal	Posted on November 5, 2015
Thor, wearables and construction	Posted on October 29, 2015
Smart city thinking today	Posted on October 22, 2015
What can you actually do with wearable data?	Posted on October 15, 2015
Keeping your data personal in the â€˜smartâ€™ era	Posted on October 8, 2015
Seeing the forest and the trees	Posted on October 1, 2015
What data science means to me	Posted on September 24, 2015
KTPs to plug the skills gap	Posted on September 17, 2015

Title	Timestamp
Embracing data science in Dubai	Posted on September 10, 2015
Wearables herald a brave new world for marketers	Posted on September 3, 2015
An emotional response to a technical device â€“ my wearables experience!	Posted on August 27, 2015
Making data science work for you	Posted on August 20, 2015
Smart Cities mean smarter marketing	Posted on August 13, 2015
Does the nothing box exist?	Posted on August 6, 2015
What I think about data protection â€“ a respectful rebuttal to Hasan Akyol	Posted on July 10, 2015
Wearable wobbles	Posted on June 30, 2015
Should individuals be solely responsible for protecting their privacy and security online?	Posted on June 18, 2015
Whatâ€™s in store for wearables data and marketing?	Posted on June 7, 2015
Here comes the data science bitâ€¦	Posted on April 23, 2015
You have to WEAR them	Posted on March 27, 2015
5 questions everybody should ask about their email marketing campaign	Posted on March 23, 2015
How big data can improve your commute	Posted on February 27, 2015
Everything you need to know about clustering and how it can help your business	Posted on January 20, 2015
Profusion has Landed: a digital dÃ©butante in Dubai	Posted on October 27, 2014
Testing: will you be star baker?	Posted on October 10, 2014
Make it personal â€“ personalisation in the digital age	Posted on September 10, 2014
Lily pads: Markov chains and the customer journey	Posted on September 9, 2014
The survivorâ€™s guide to digital marketing internships	Posted on September 4, 2014
Seeing is believing	Posted on August 27, 2014
The new gmail unsubscribe button â€“ a chance for marketers to prove themselves	Posted on August 26, 2014
A handy guide to buzzwords in digital marketing â€“ part 1	Posted on August 5, 2014
Not just a pretty face? Amazon and the art of secret signals	Posted on July 14, 2014
How we use trees* to predict customer behaviour: random forests in action (*not real ones)	Posted on July 8, 2014
Donâ€™t fall at the last hurdle â€“ the essential rules of copy	Posted on June 26, 2014
Are there such things as digital buying signals?	Posted on June 18, 2014
Pioneering principles still apply in the world of personalisation	Posted on June 16, 2014
A stock take â€“ where have we been and where are we going?	Posted on June 15, 2014

What words do they commonly use?



Relationships Between Terms

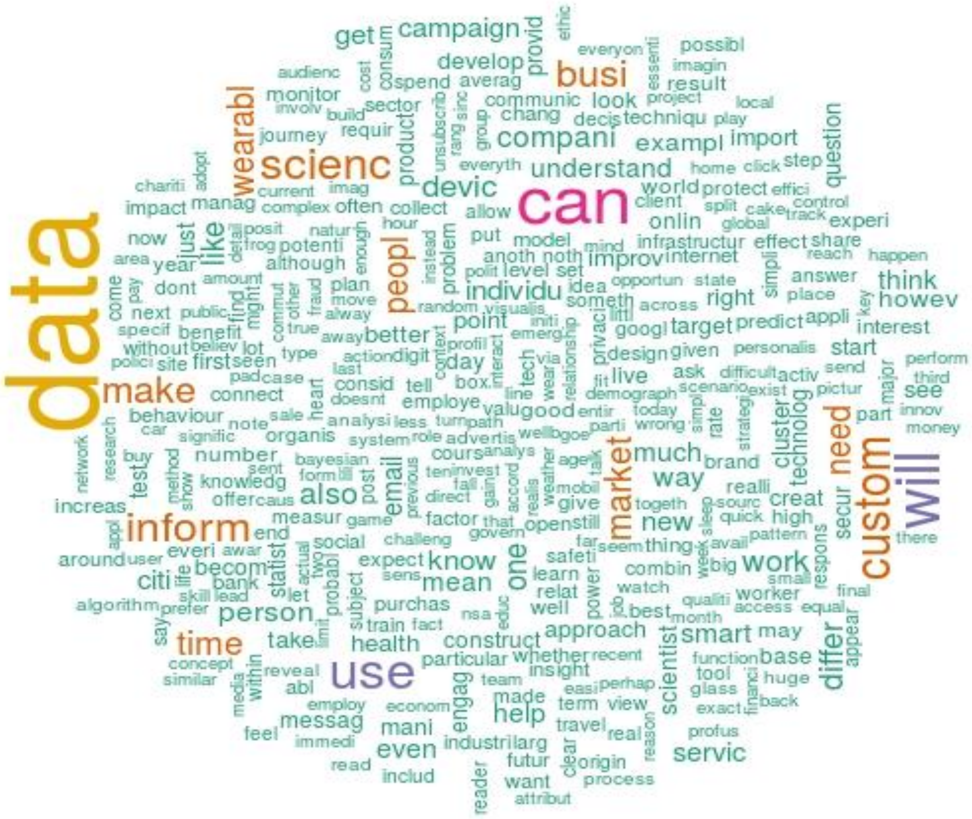
- **Data**
  - ✓ science: 0.68
  - ✓ consequently: 0.64
  - ✓ team: 0.64
  - ✓ business:0.61
  - ✓ scientists:0.61

- **Science**
  - ✓ data: 0.68
  - ✓ business:0.64
  - ✓ capability: 0.64
  - ✓ encouraging: 0.64
  - ✓ promises: 0.64

- **will**
  - ✓ benefits: 0.61
  - ✓ technology: 0.61

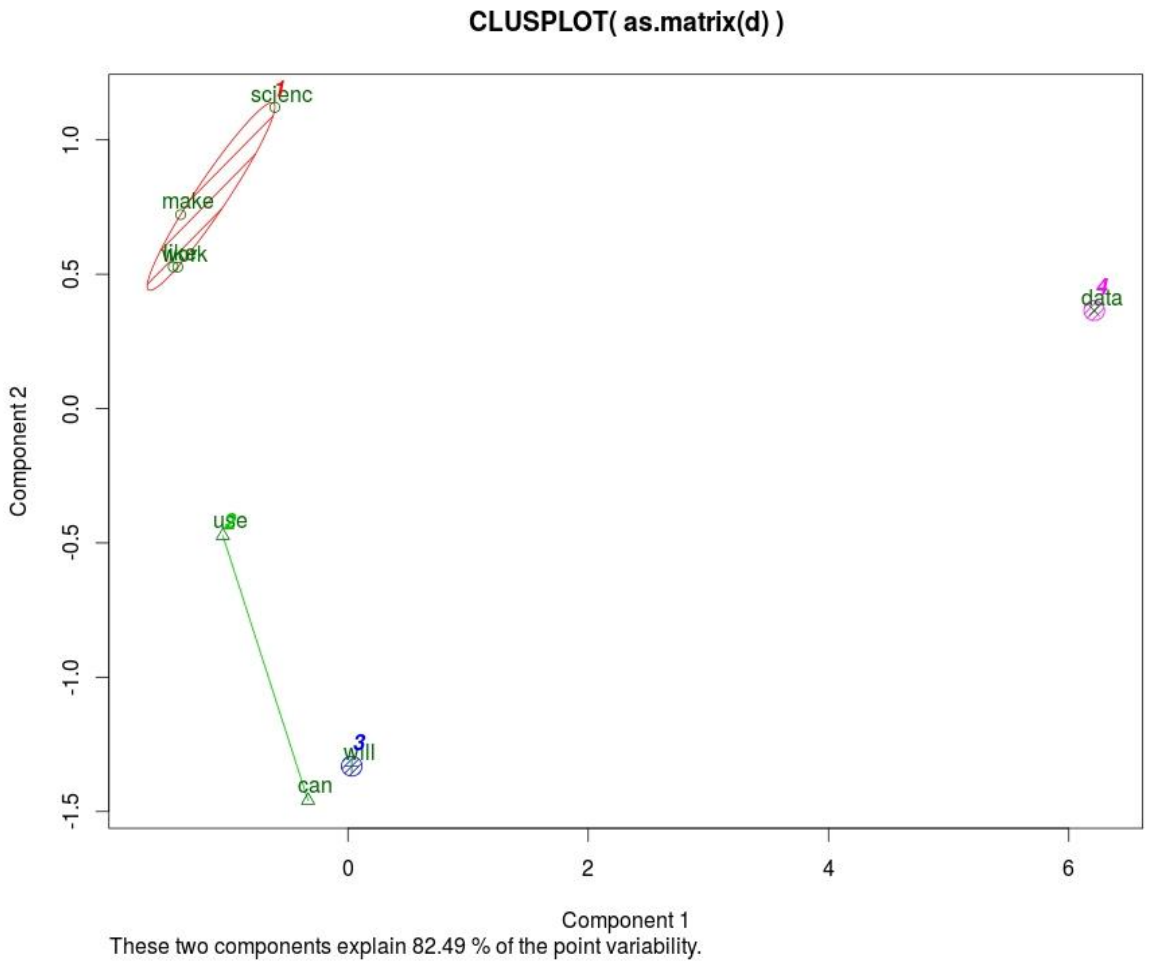
- **people**
  - ✓ perform: 0.72
  - ✓ highest: 0.71
  - ✓ technique: 0.69
  - ✓ rarely: 0.68
  - ✓ sale:0.66
  - ✓ engaged: 0.65
  - ✓ encourage: 0.64
  - ✓ purchasing: 0.62

Word Cloud

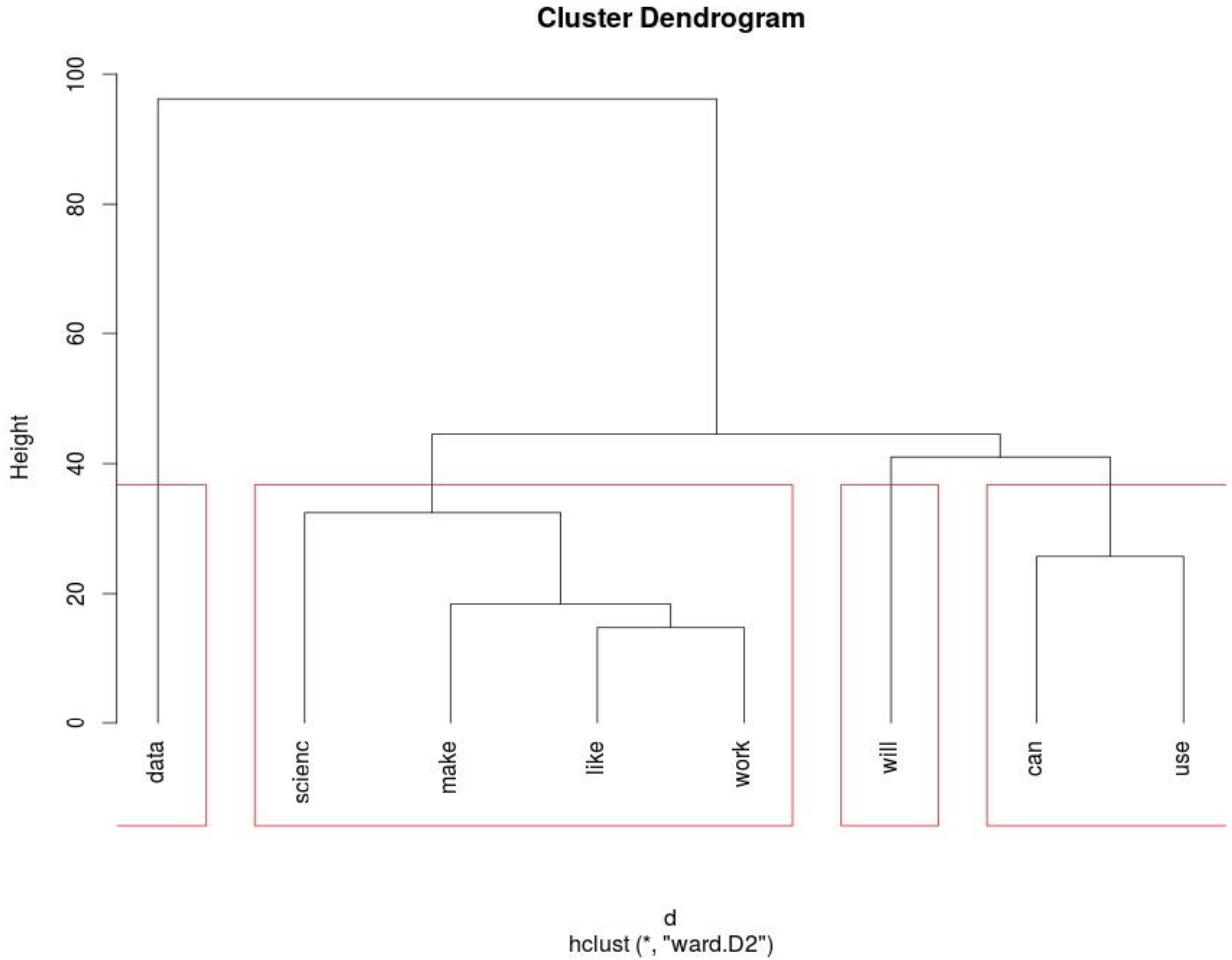

data
can
will
custom
use
scienc
inform
make
time
busi
market
peopl
need
wearabl
campaign
compani
understand
individu
differ
work
know
mean
person
help
servic

Clustering

- K-means clustering



- Hierarchal Clustering

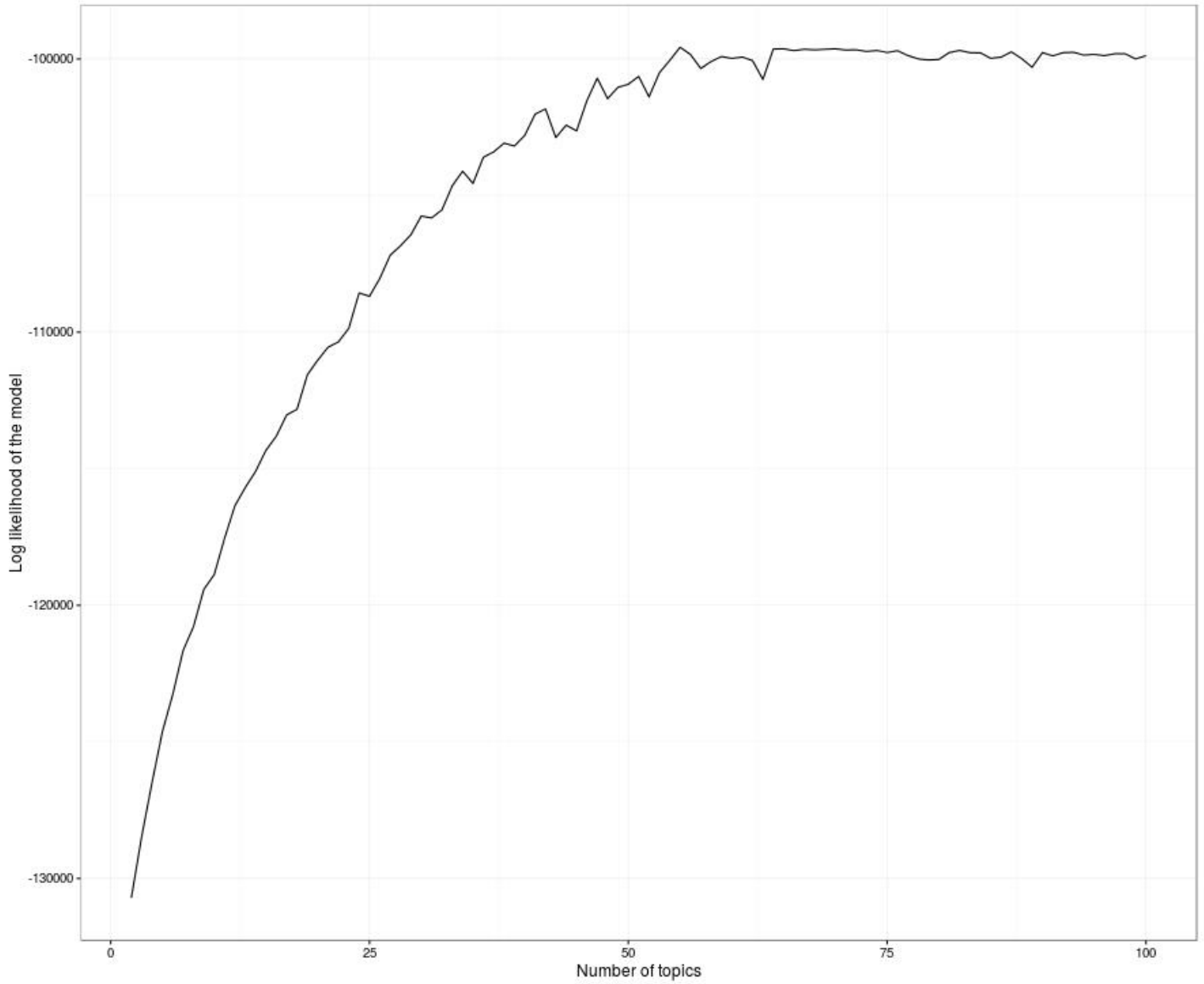


Topic modelling

Optimal number of topics

Rule of thumb:

- The correct number is the number you want, just like with k-means.
- Try out different values of k, select the one has largest likelihood.



LDA

- Custom: 0.0193
- Data: 0.0160
- Can: 0.0137
- Email: 0.0106
- Use: 0.0096

DS & email

- Data: 0.0263
- Wearable: 0.0146
- Can: 0.0128
- Inform: 0.0122
- Will: 0.0102

DS & wearables

- Data: 0.0358
- Can: 0.0182
- Science: 0.0164
- Business: 0.0147
- Use: 0.0108

DS & business

- Data: 0.0336
- Will: 0.0177
- Can: 0.0111
- Smart: 0.0093
- Construct: 0.0085

DS & smart cities

LDA extensions

There are numerous extensions to the standard LDA model:

- hierarchical Dirichlet processes (Hdp, Teh et al., 2006b)
- dynamic topic models (Dtm, Blei and Lafferty, 2006)
- correlated topic models (Ctm, Blei and Lafferty, 2007)
- etc.

References

- D. M. Blei, A. Ng, and M. I. Jordan. Latent Dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.
- D. M. Blei. Topic models. Video lecture, 2009. URL http://videolectures.net/mlss09uk_blei_tm/.
- D. M. Blei. Introduction to probabilistic topic models, 2011. URL http://www.cs.princeton.edu/~blei/papers/Blei2011.pdf.
- D. M. Blei and J. D. Lafferty. A correlated topic model of Science. Annals of Applied Statistics, 1(1): 17–35, 2007.
- H.-A. Chang and C.-H. Yu. Latent Dirichlet allocation. Seminar paper, 2007. URL http://cassava.fudan.edu.cn/seminars/ppt/lecture-lda.pdf.
- Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101:1566–1581, 2006b
- Y. Wang, P. Sabzmeydani, and G. Mori. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. Lecture Notes in Computer Science, 4814:240–254, 2007

Questions?

Thanks !!!!