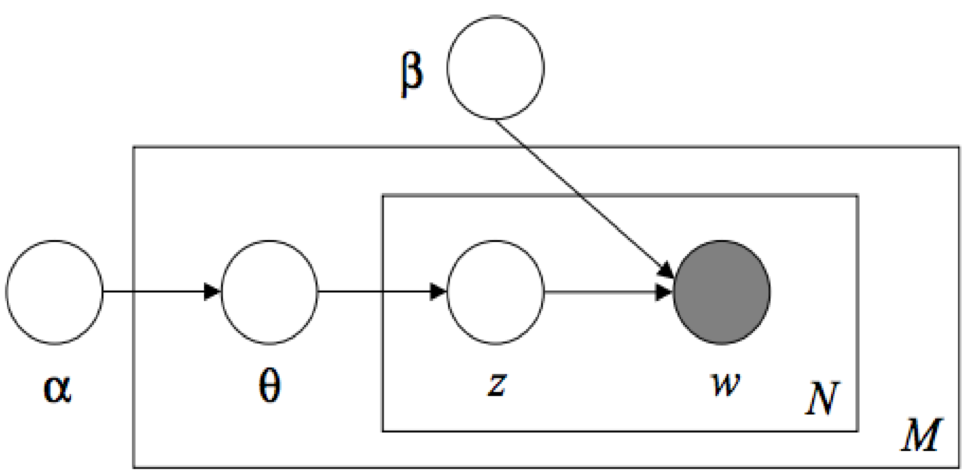
Karlsruhe Institute of Technology

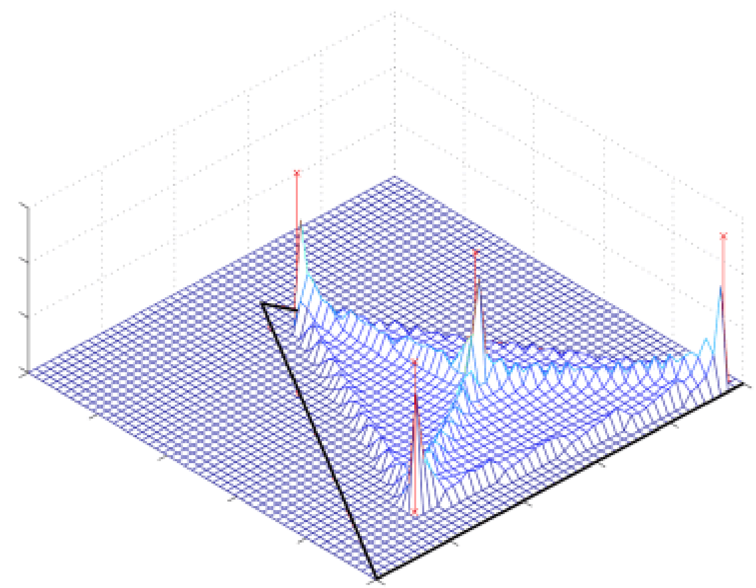
Topic Recognition
Latent Dirichlet Allocation

Algorithmic Methods in the Humanities · June 23, 2016

Florian Becker

INSTITUTE OF THEORETICAL INFORMATICS · ALGORITHMICS GROUP





More and more text

http://www.passion-estampes.com/npe/newsletter-francois-schuiten.html

More and more text

http://www.passion-estampes.com/npe/newsletter-francois-schuiten.html

Mass production of text:

- $>$ 4000 peer-reviewed papers / day
- nearly 3 million blog posts / day
- 500 million tweets / day

(Probabilistic) Topic Models - Intro

- Automatically extract **topics** from documents
- Organizing and searching of large collections of text

(Probabilistic) Topic Models - Intro

Karlsruhe Institute of Technology

- Automatically extract **topics** from documents
- Organizing and searching of large collections of text

Algorithm: Corpus $\rightarrow$ Topics

Input	corpus, int K (number of topics)
Output	K topics

(Probabilistic) Topic Models - Intro

- Automatically extract **topics** from documents
- Organizing and searching of large collections of text

Algorithm: Corpus $\rightarrow$ Topics

Input corpus, int K (number of topics)
Output K topics

Distribution over words

(Probabilistic) Topic Models - Intro

- Automatically extract **topics** from documents
- Organizing and searching of large collections of text

Algorithm: Corpus $\rightarrow$ Topics

Input corpus, int K (number of topics)
Output K topics

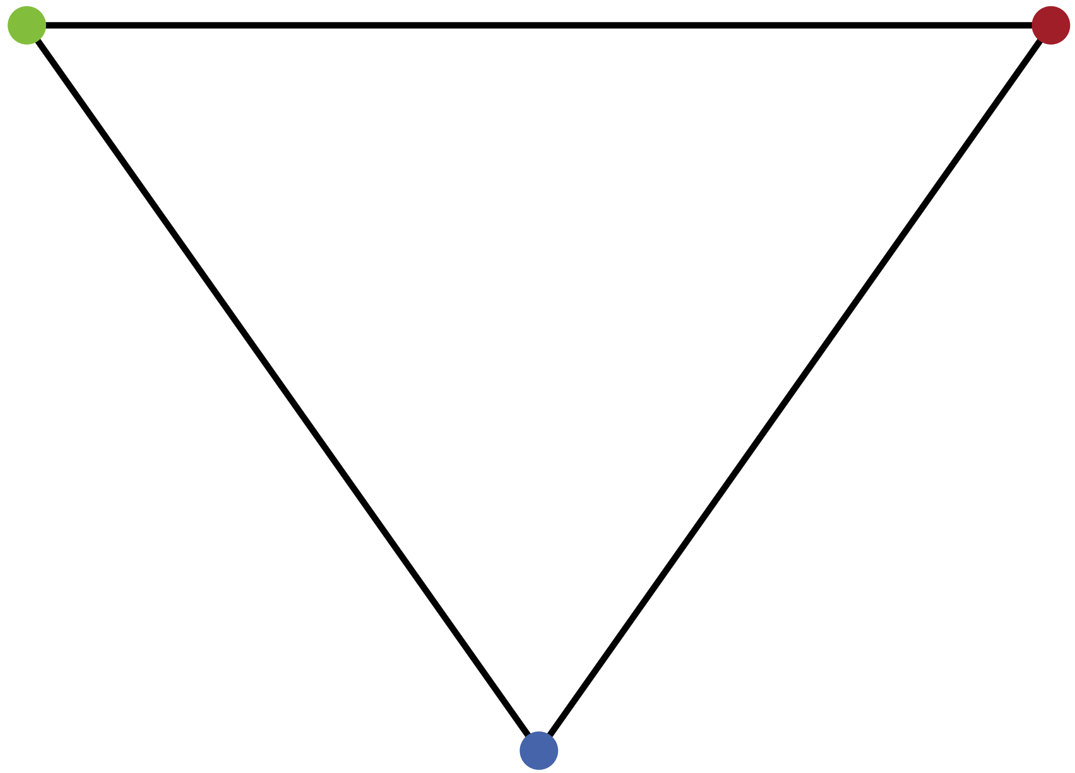
Distribution over words

0.15*algorithm
0.1*complexity
0.05*program
0.05*turing
...
...

Institute of Theoretical Informatics
Algorithmics Group

Topic Simplex

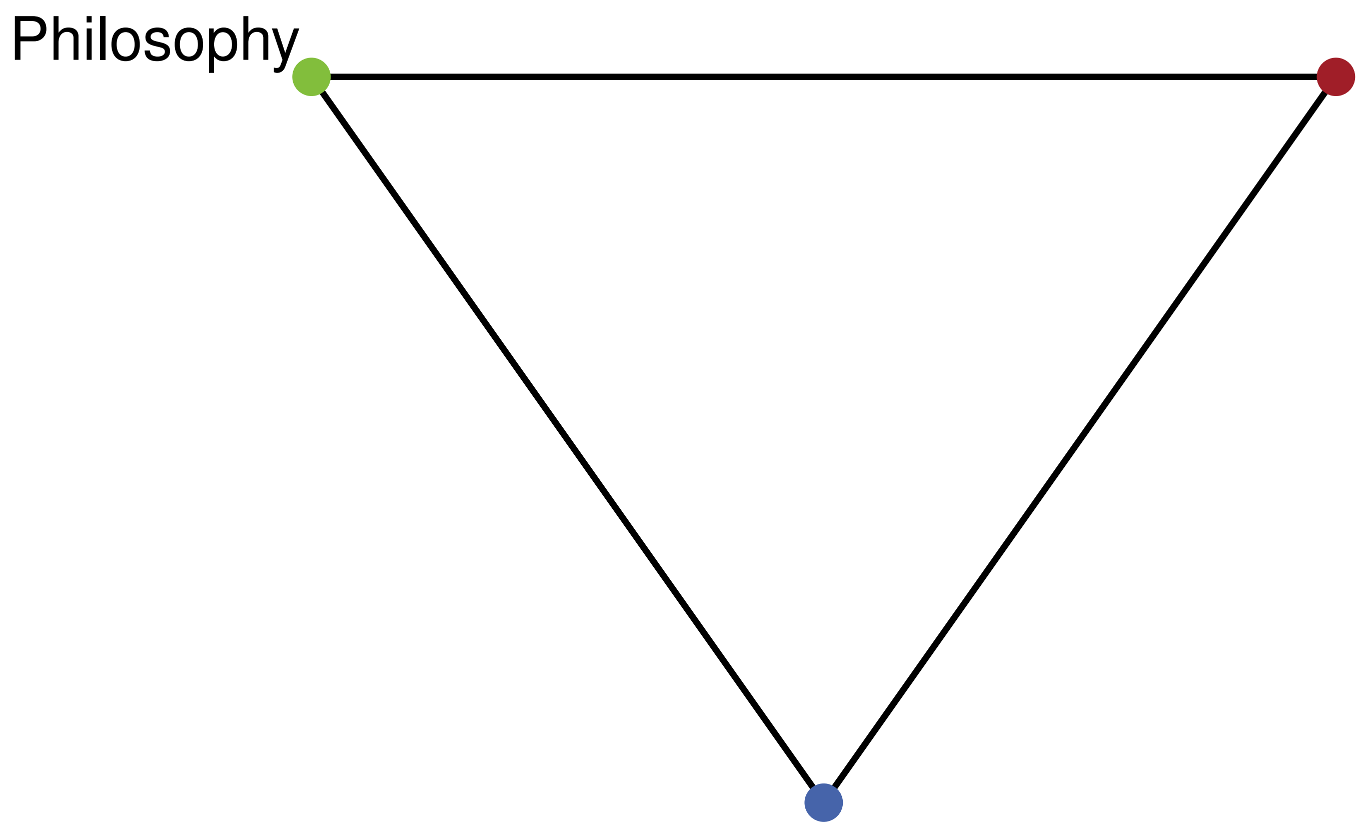
A document is a distribution over topics



Institute of Theoretical Informatics
Algorithmics Group

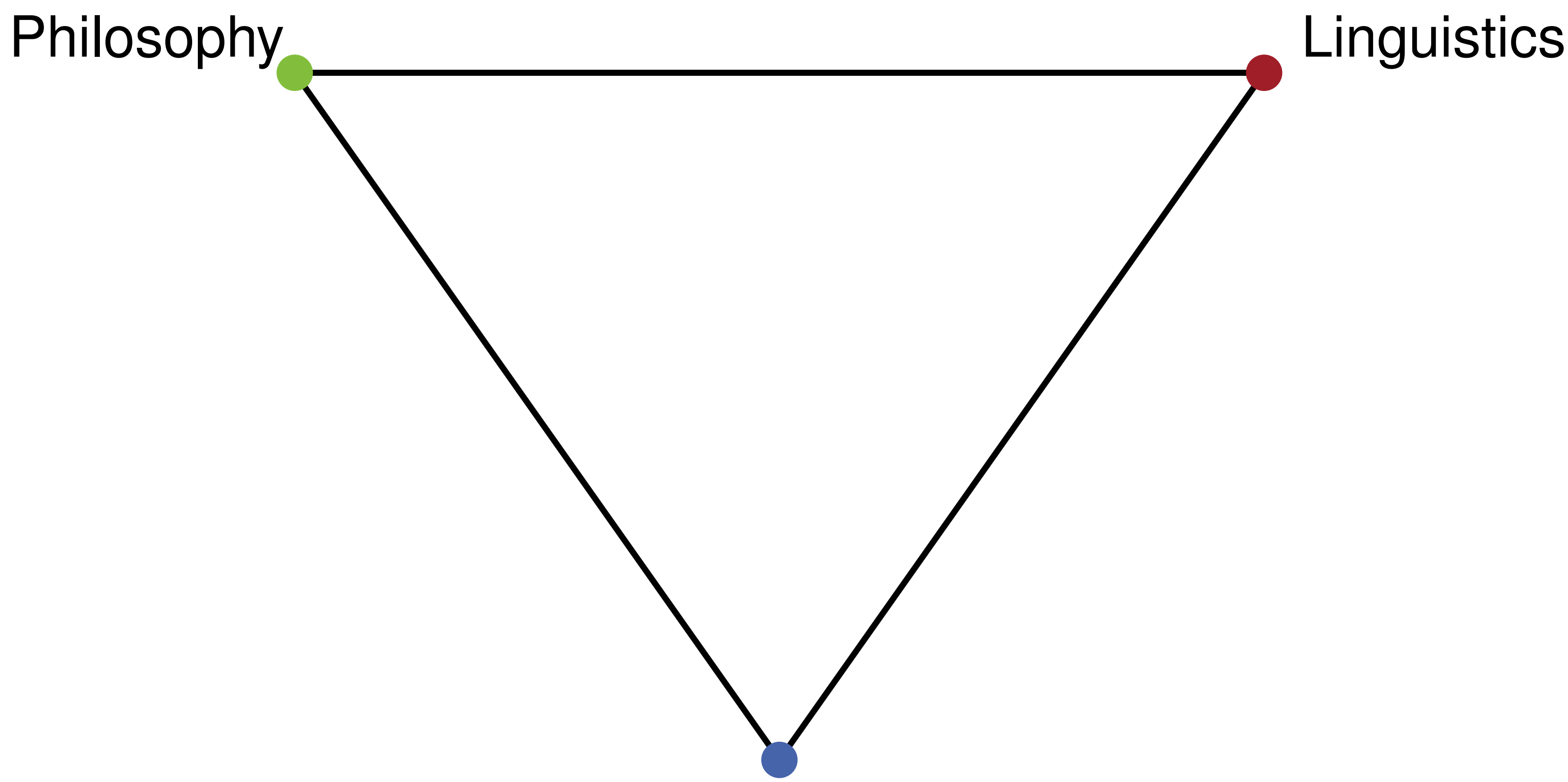
Topic Simplex

A document is a distribution over topics



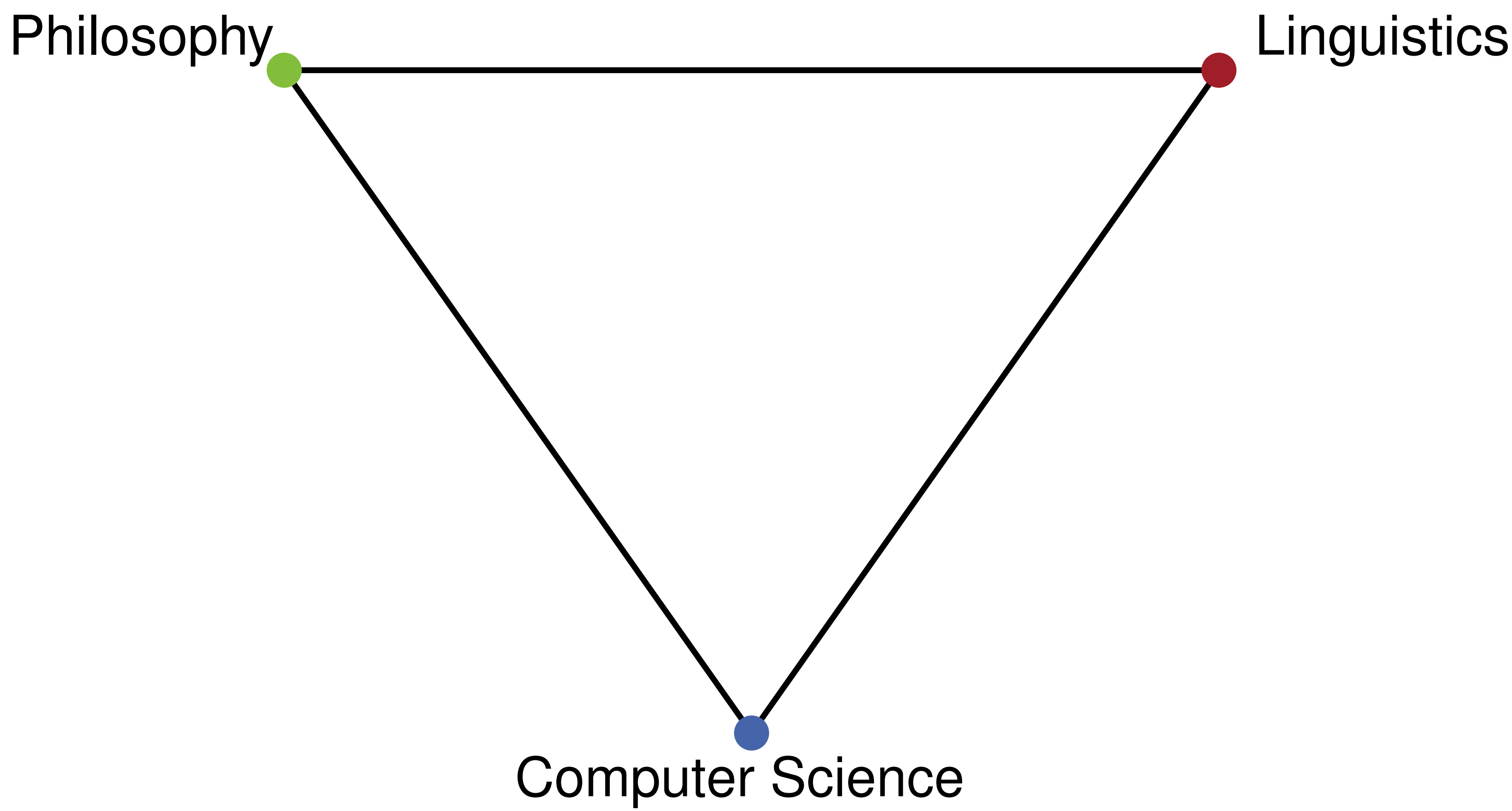
Topic Simplex

A document is a distribution over topics



Topic Simplex

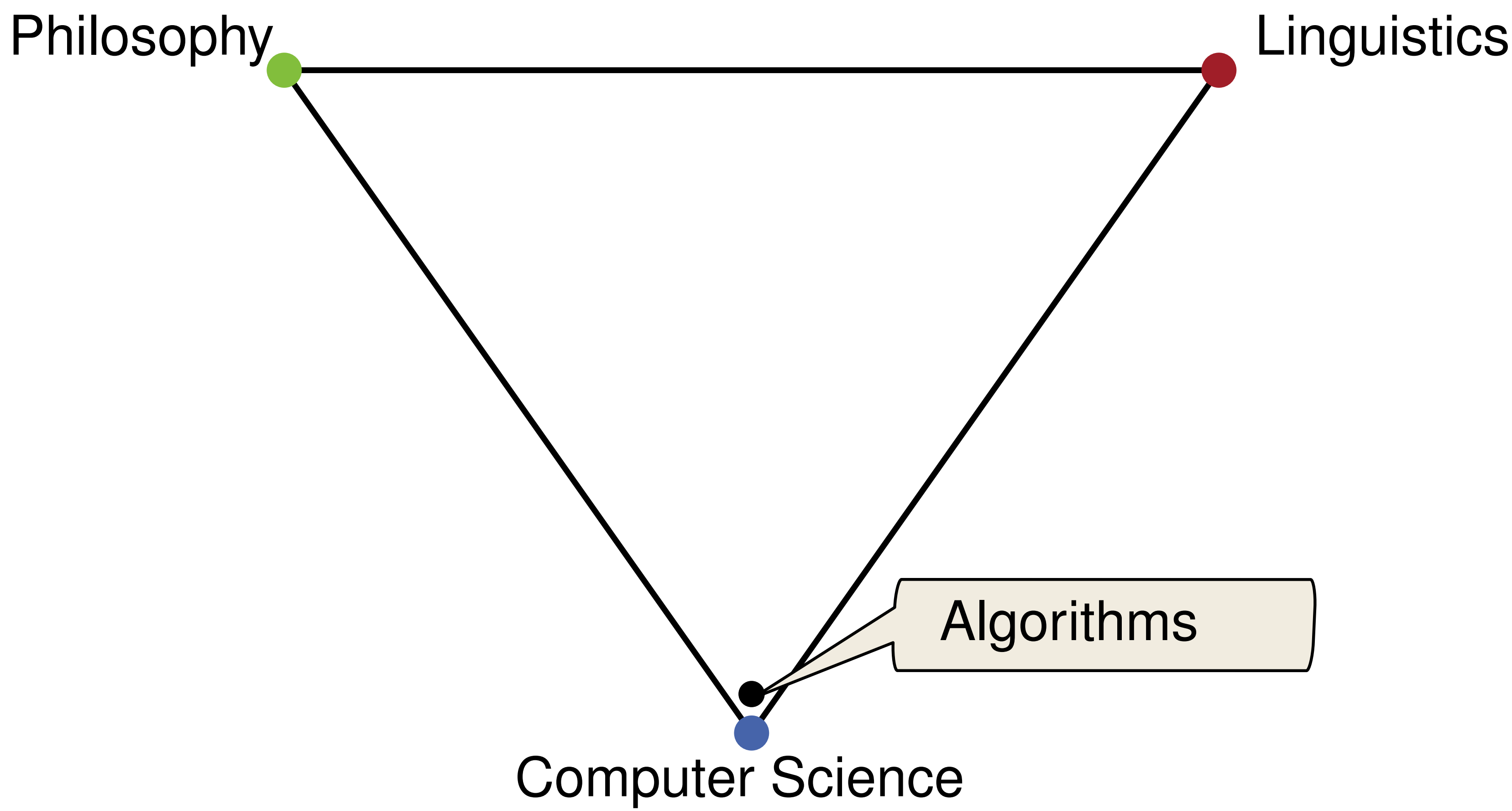
A document is a distribution over topics



Institute of Theoretical Informatics
Algorithmics Group

Topic Simplex

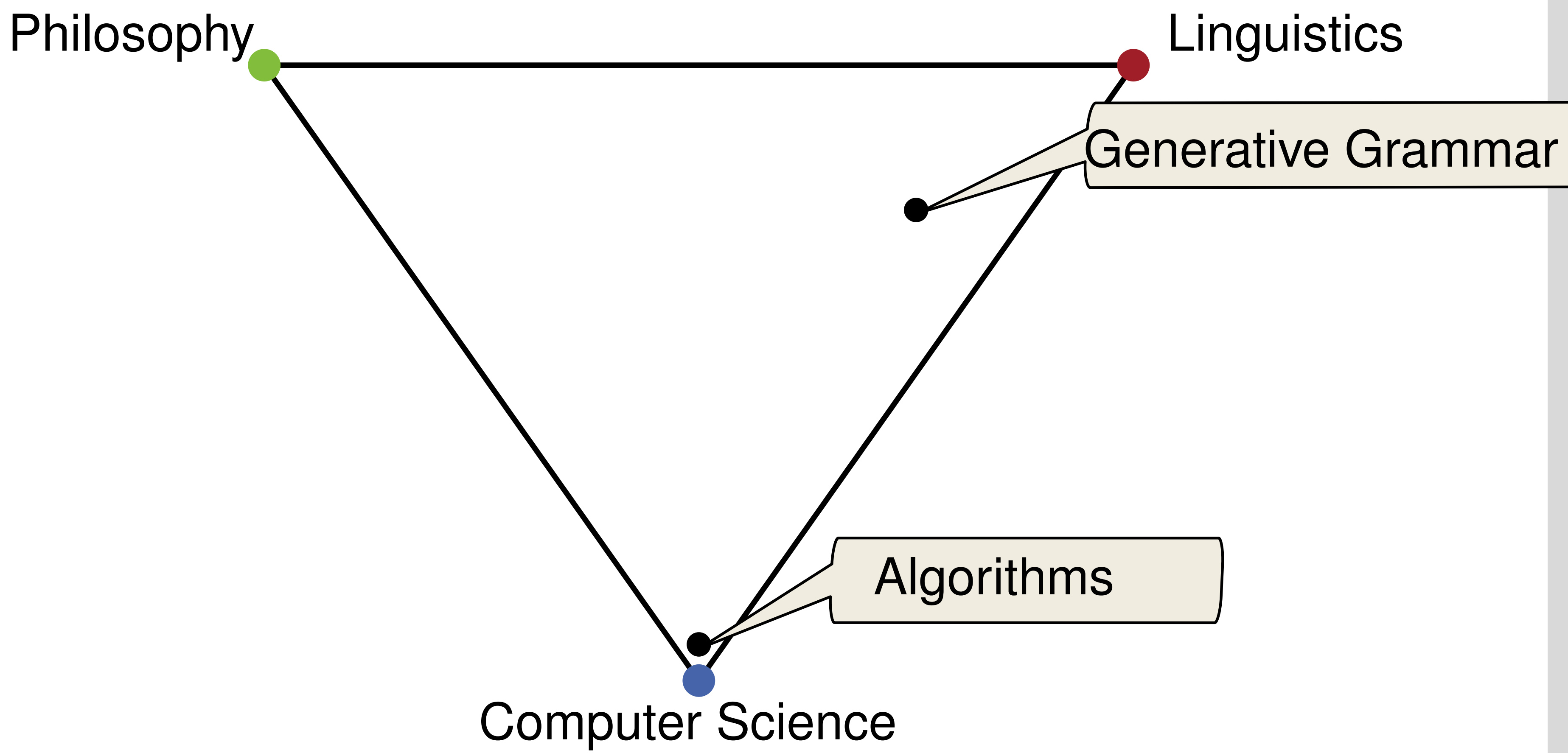
A document is a distribution over topics



Institute of Theoretical Informatics
Algorithmics Group

Topic Simplex

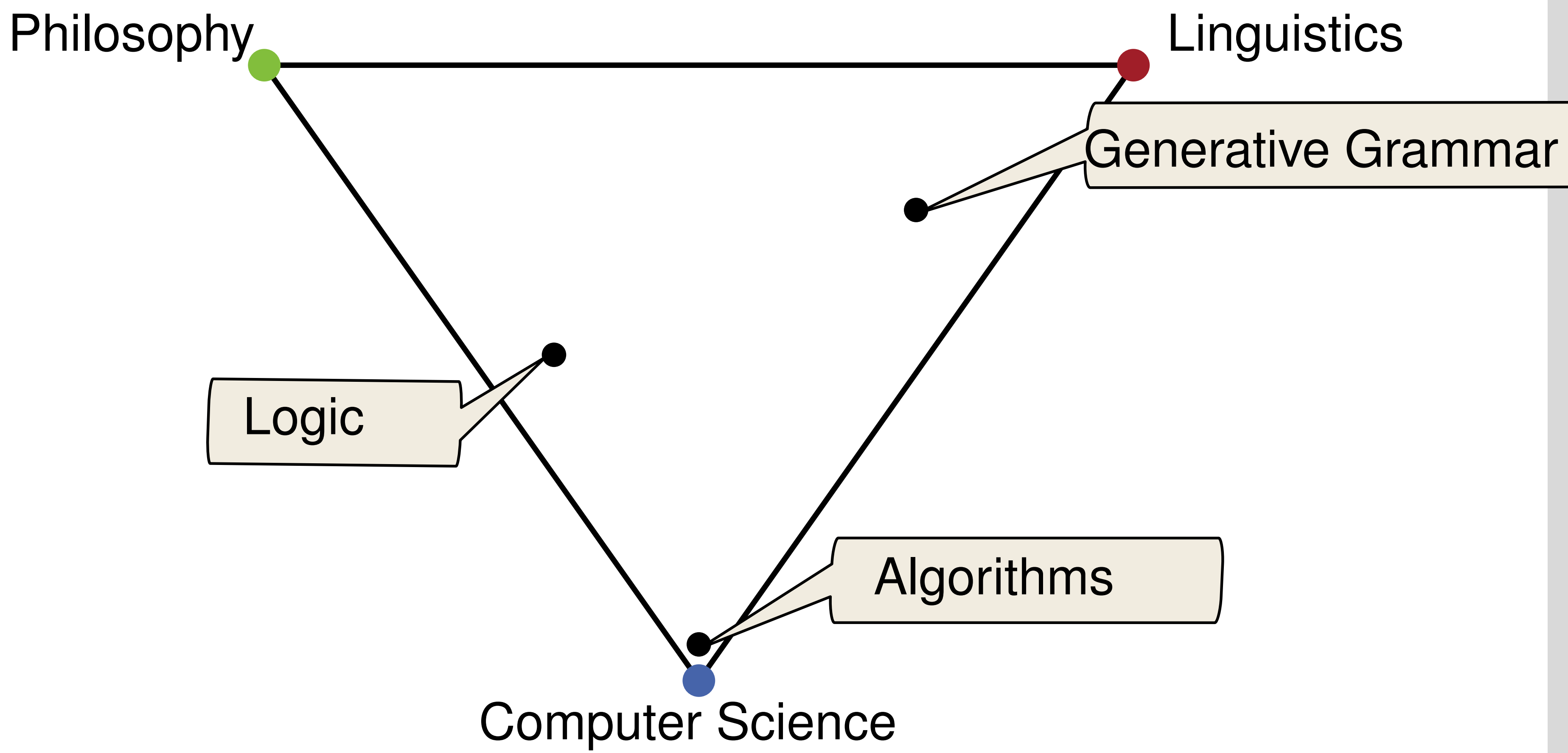
A document is a distribution over topics



Institute of Theoretical Informatics
Algorithmics Group

Topic Simplex

A document is a distribution over topics



Institute of Theoretical Informatics
Algorithmics Group

Outline

Karlsruhe Institute of Technology

Latent Dirichlet Allocation: What does it do?

Outline

Karlsruhe Institute of Technology

Latent Dirichlet Allocation: What does it do?

- Assumptions
- Generative Process
- Dirichlet Distribution

Institute of Theoretical Informatics
Algorithmics Group

Outline

Karlsruhe Institute of Technology

Latent Dirichlet Allocation: What does it do?

- Assumptions
- Generative Process
- Dirichlet Distribution

Demo

Institute of Theoretical Informatics
Algorithmics Group

Outline

Latent Dirichlet Allocation: What does it do?

- Assumptions
- Generative Process
- Dirichlet Distribution

Demo

Latent Dirichlet Allocation: How does it do it?

Institute of Theoretical Informatics
Algorithmics Group

Outline

Latent Dirichlet Allocation: What does it do?

- Assumptions
- Generative Process
- Dirichlet Distribution

Demo

Latent Dirichlet Allocation: How does it do it?

- Inference
- Gibbs Sampling

Institute of Theoretical Informatics
Algorithmics Group

Outline

Latent Dirichlet Allocation: What does it do?

- Assumptions
- Generative Process
- Dirichlet Distribution

Demo

Latent Dirichlet Allocation: How does it do it?

- Inference
- Gibbs Sampling

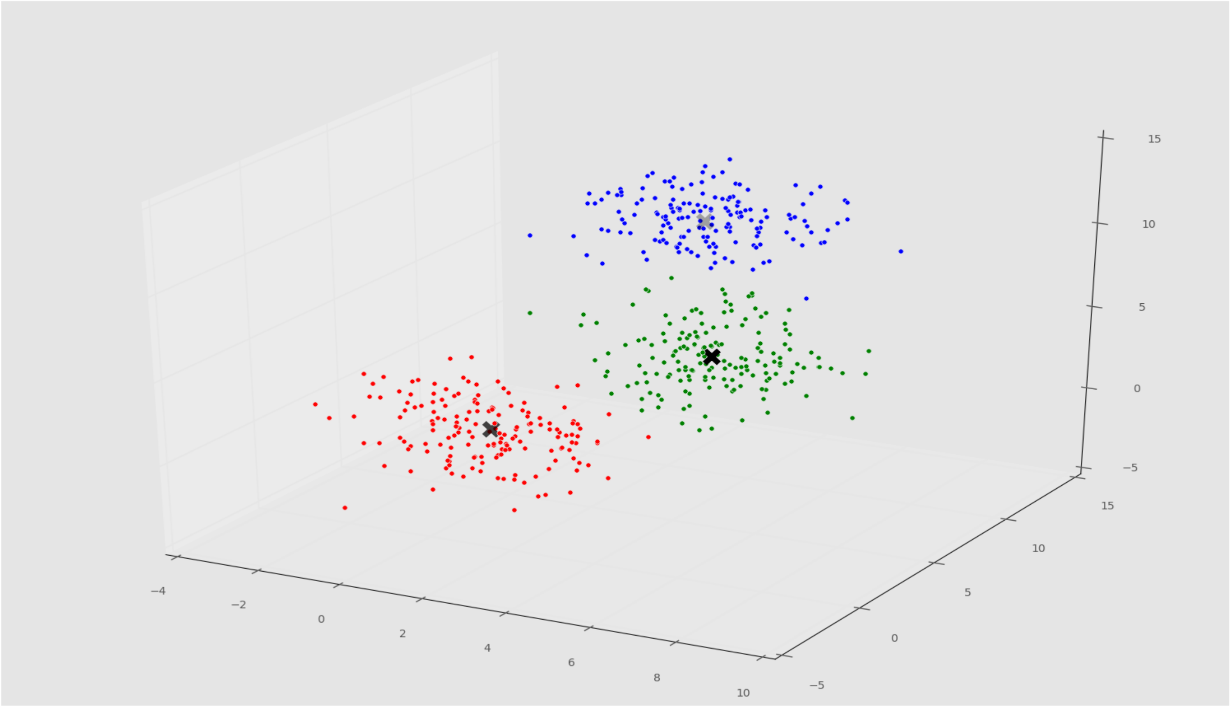
Conclusion

Institute of Theoretical Informatics
Algorithmics Group

Latent Dirichlet Allocation

Latent Dirichlet Allocation

- Unsupervised Learning Model
- Finding clusters of similar texts
- Generative Model



Latent Dirichlet Allocation

Karlsruhe Institute of Technology

Assumptions

- A document is represented as a bag of words
- A document is about multiple topics
- A topic is a distribution over words
- Order of documents in corpus does not matter
- Every document is generated by a **generative process**

Generative Process

Algorithm: Generative Process

1. Choose $\theta_i \sim \text{Dir}(\alpha)$,
2. Choose $\varphi_k \sim \text{Dir}(\beta)$
3. For each of the word positions i, j
   (a) Choose a topic $z_{i,j} \sim \text{Multinomial}(\theta_i)$.
   (b) Choose a word $w_{i,j} \sim \text{Multinomial}(\varphi_{z_{i,j}})$

$j \in \{1, \ldots, N_i\}$, and $i \in \{1, \ldots, D\}$

N_i - Number of words in document i
D - Number of documents

Institute of Theoretical Informatics
Algorithmics Group

Dirichlet Distribution

Binomial Distribution $\rightarrow$ Multinomial Distribution $\rightarrow$ Dirichlet Distribution

Institute of Theoretical Informatics
Algorithmics Group

Dirichlet Distribution

Binomial Distribution $\rightarrow$ Multinomial Distribution $\rightarrow$ Dirichlet Distribution

Binomial Distribution (PMF)

$$f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Institute of Theoretical Informatics
Algorithmics Group

Dirichlet Distribution

Binomial Distribution $\rightarrow$ Multinomial Distribution $\rightarrow$ Dirichlet Distribution

Binomial Distribution (PMF)

$$f(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

success/failure experiments
Example: fair coin, 6 tosses

Probability of 5 heads?

$$\Pr(5 \text{ heads}) = f(5) = \Pr(X = 5) = \binom{6}{5} 0.5^5 (1 - 0.5)^{6-5} \approx 0.09375$$

Institute of Theoretical Informatics
Algorithmics Group

Dirichlet Distribution

Multinomial Distribution (PMF)

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Dirichlet Distribution

Multinomial Distribution (PMF)

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

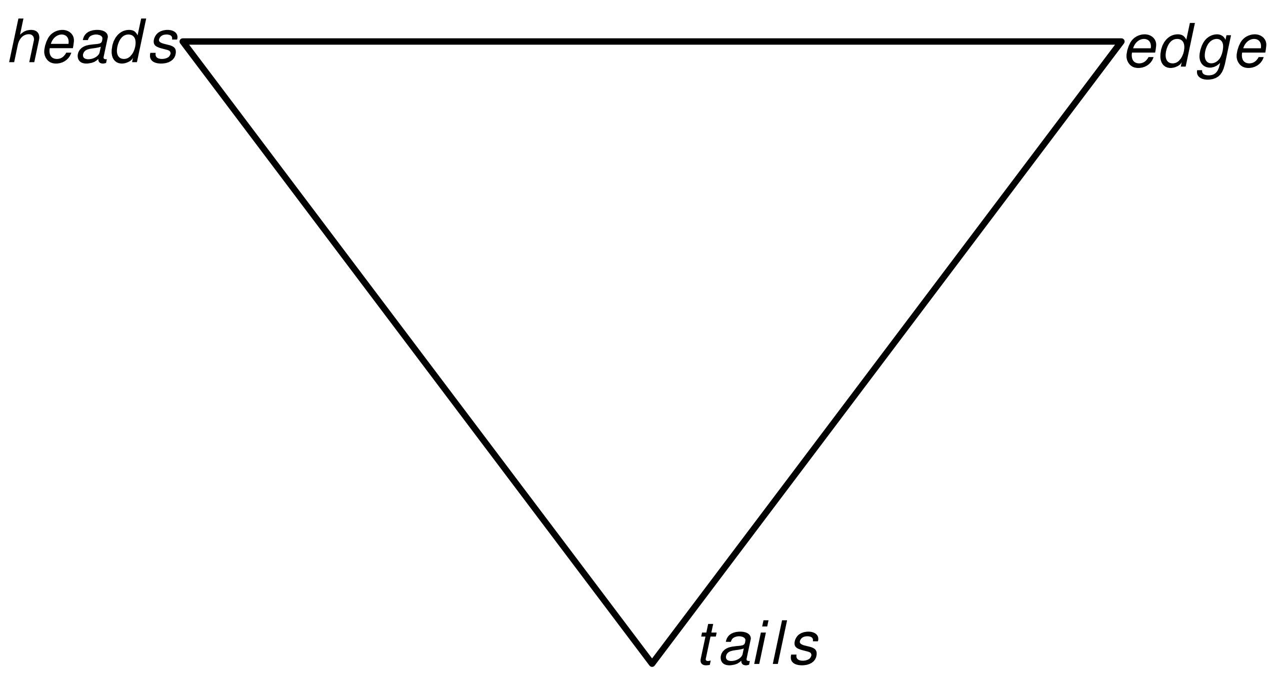
Some event with 3 outcomes: $X = [x_1, x_2, x_3]$

Dirichlet Distribution

Multinomial Distribution (PMF)

$$f(x_1, \dots, x_k; n, p_1, \dots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Some event with 3 outcomes: $X = [x_1, x_2, x_3]$



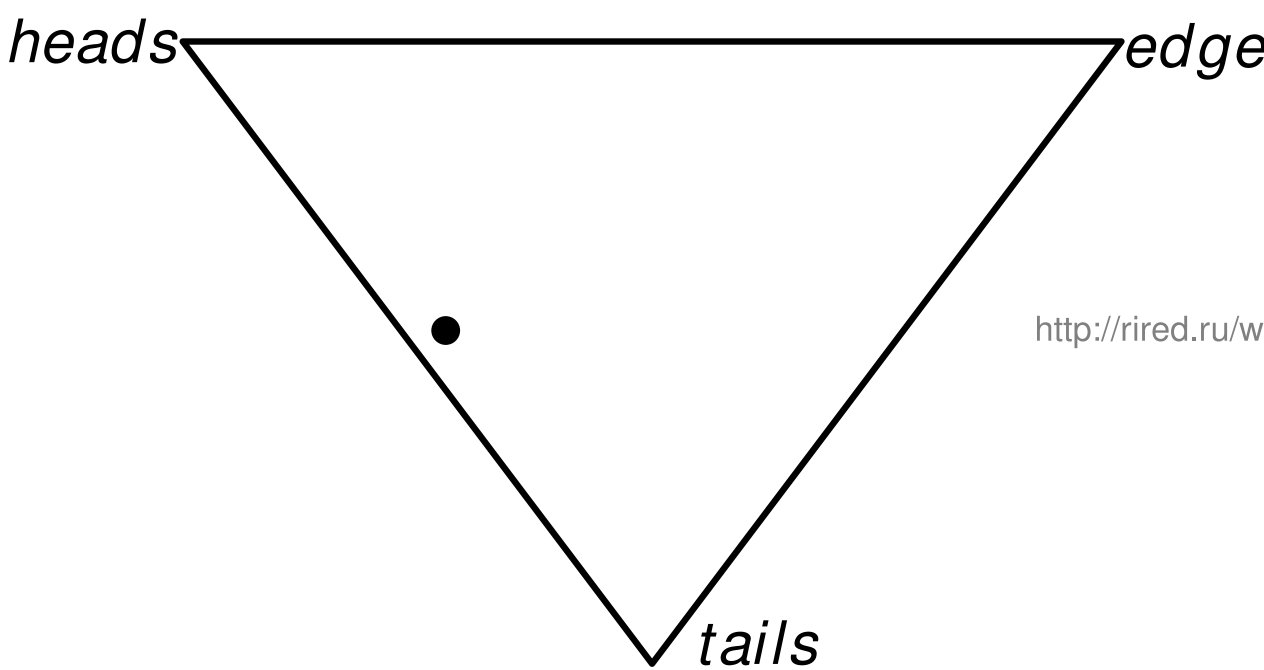
Institute of Theoretical Informatics
Algorithmics Group

Dirichlet Distribution

Multinomial Distribution (PMF)

$$f(x_1, \dots, x_k; n, p_1, \dots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Some event with 3 outcomes: $X = [x_1, x_2, x_3]$



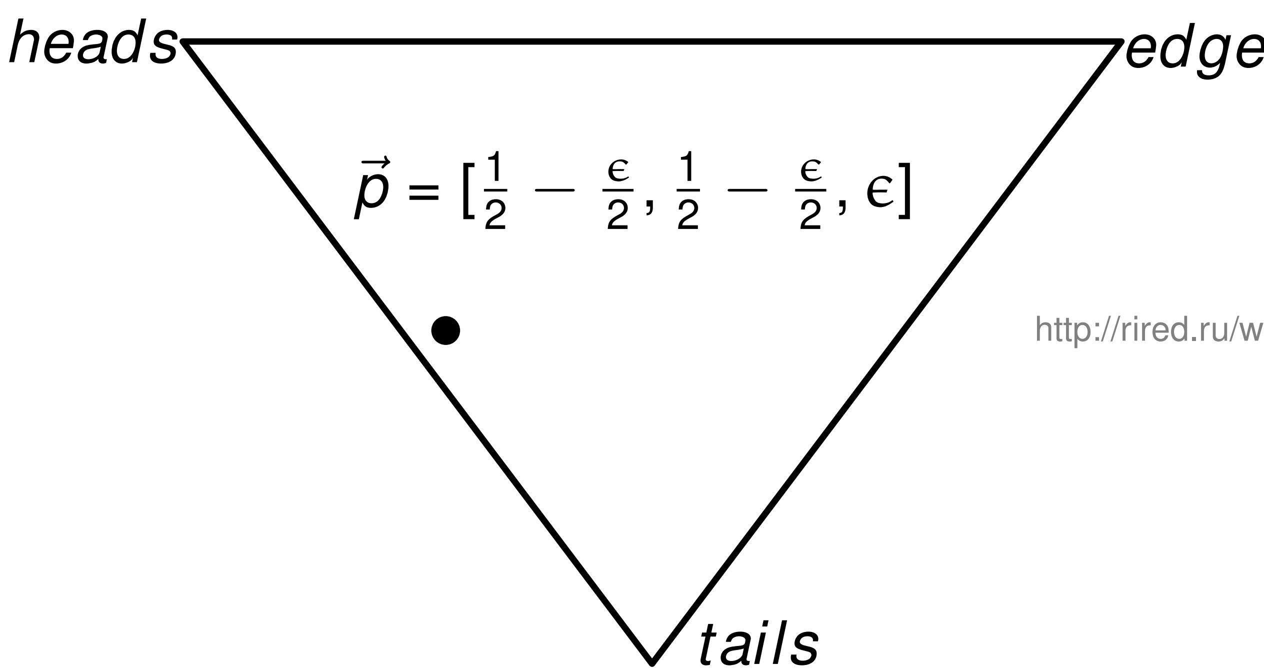
http://rired.ru/wp-content/uploads/2013/03/85142

Dirichlet Distribution

Multinomial Distribution (PMF)

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Some event with 3 outcomes: $X = [x_1, x_2, x_3]$



http://rired.ru/wp-content/uploads/2013/03/85142

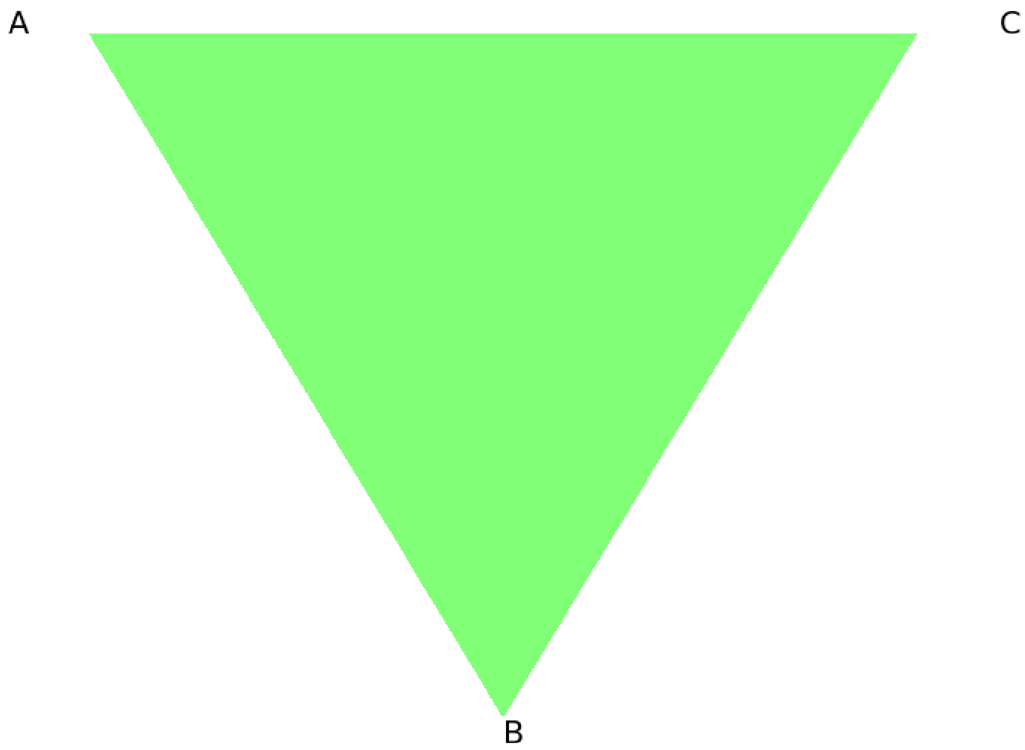
Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

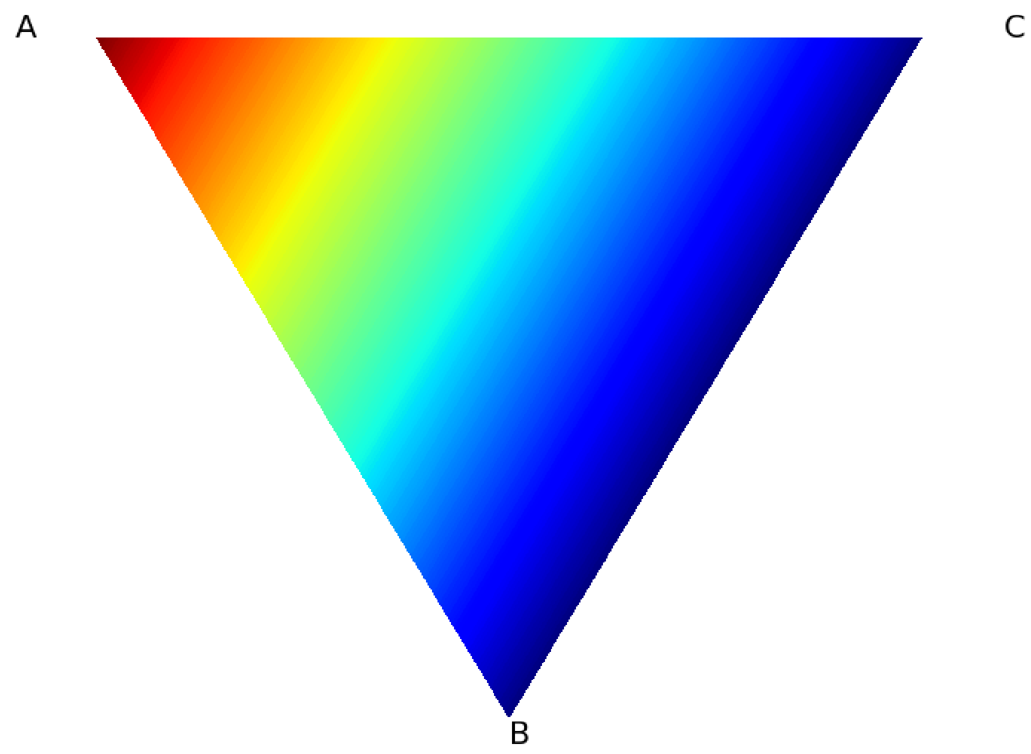
$\alpha = [1, 1, 1]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

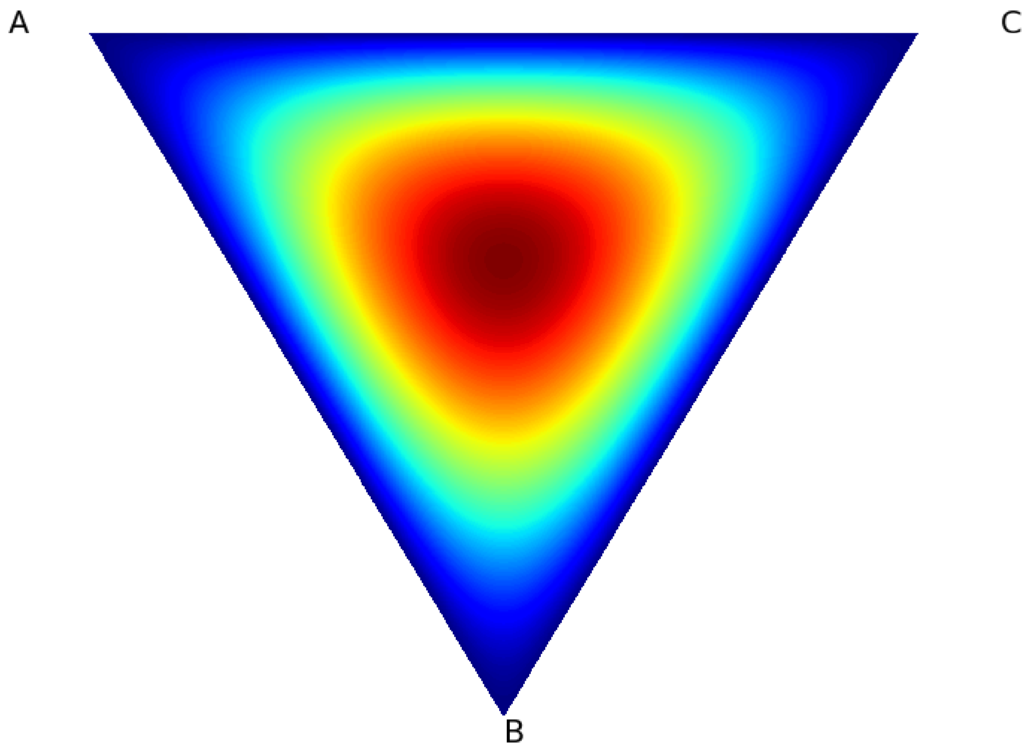
$\alpha = [2, 1, 1]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

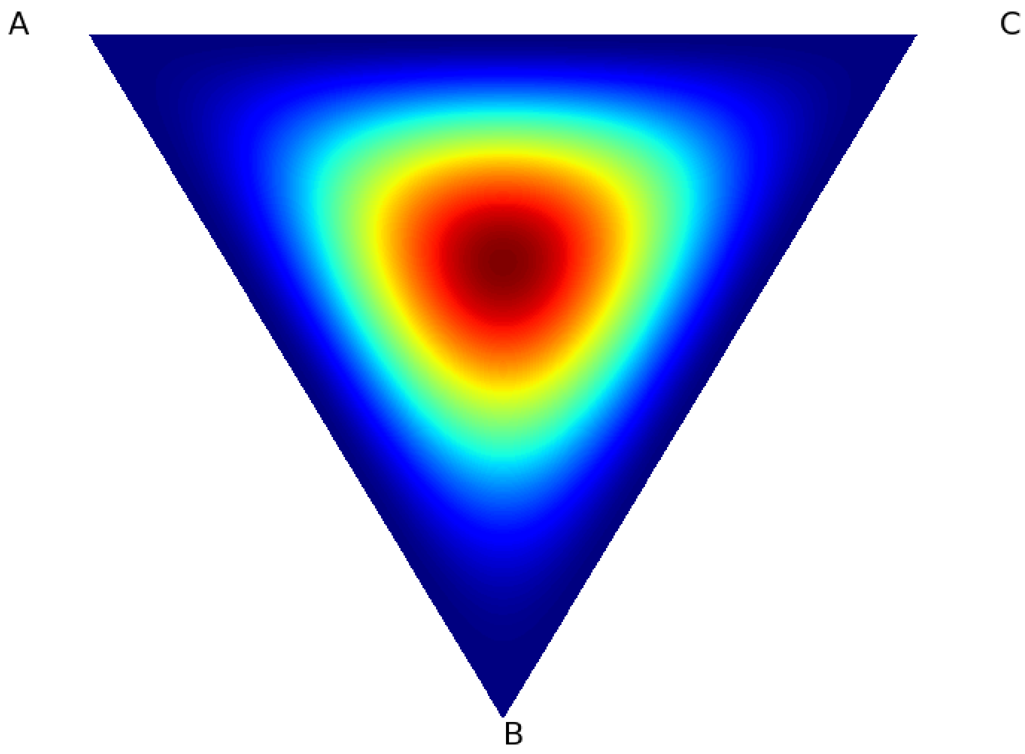
$\alpha = [2, 2, 2]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

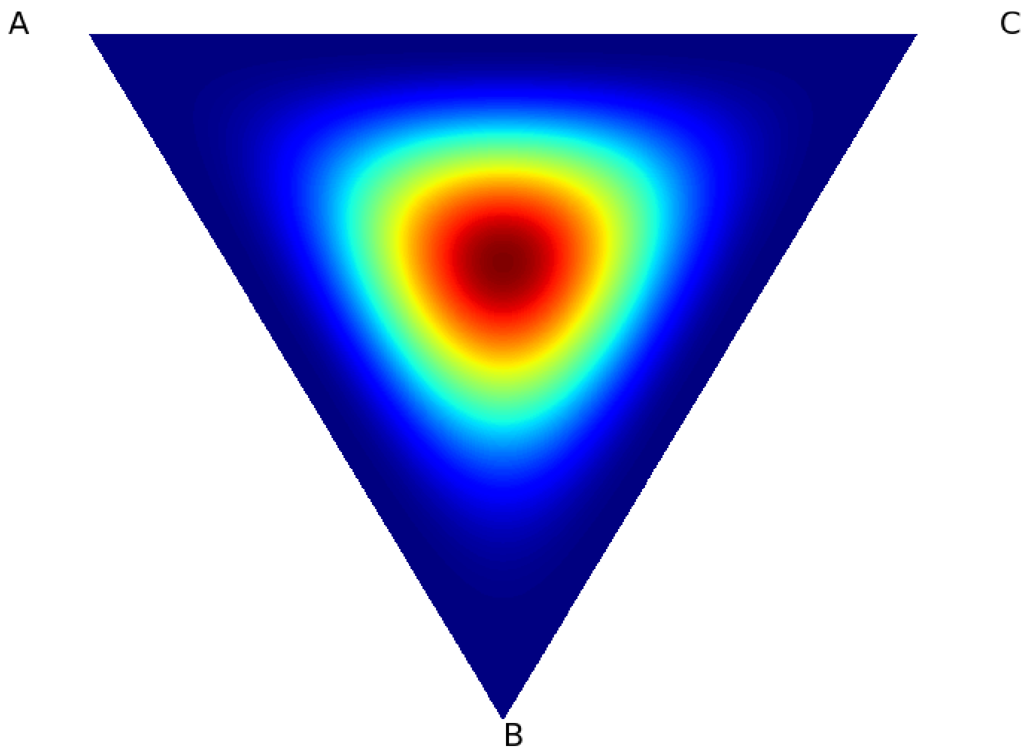
$\alpha = [3, 3, 3]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

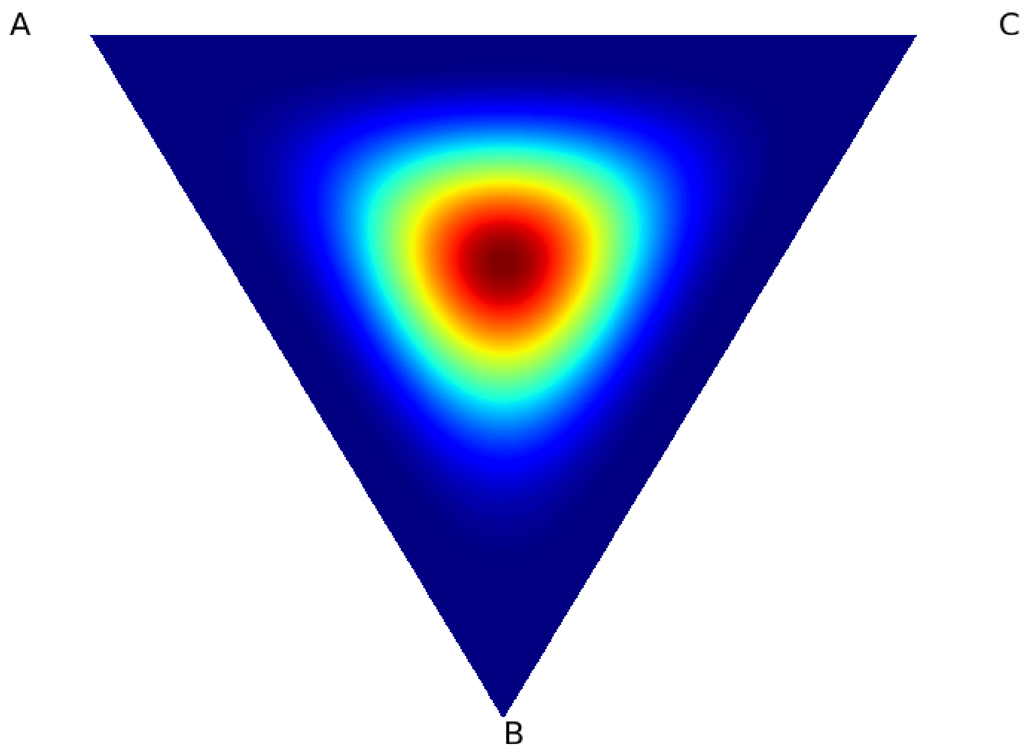
$\alpha = [4, 4, 4]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

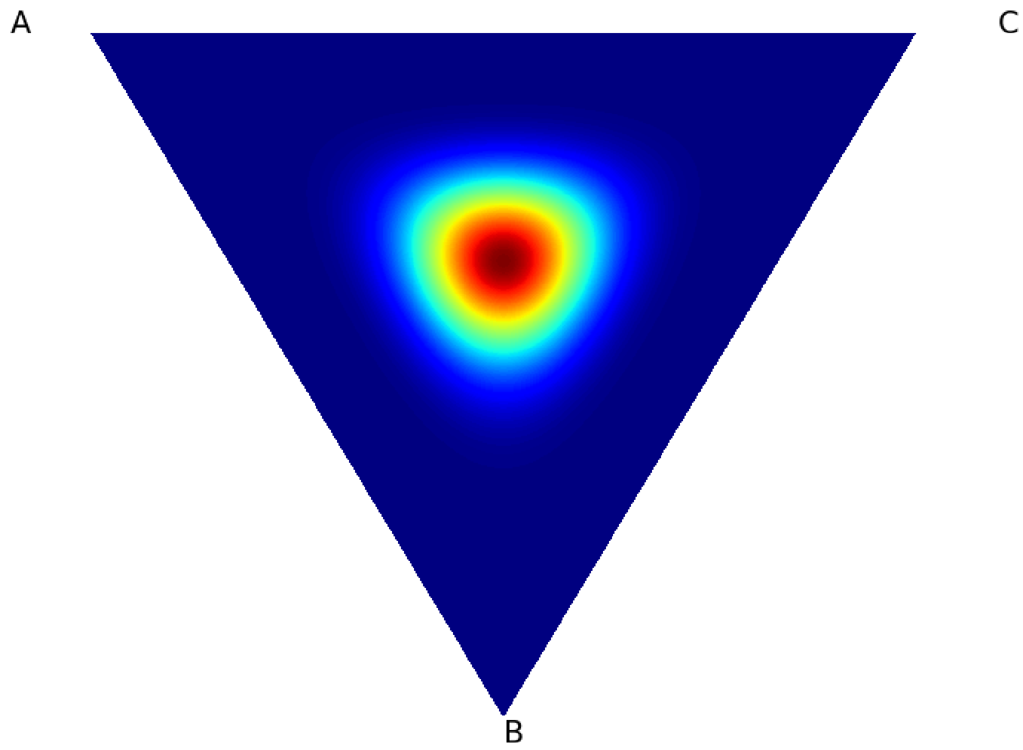
$\alpha = [5, 5, 5]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

$\alpha = [10, 10, 10]$



Dirichlet Distribution

The Dirichlet distribution $Dir(\alpha)$ is a distribution over the space of multinomial distributions.

$\alpha = [0.9, 0.9, 0.9]$

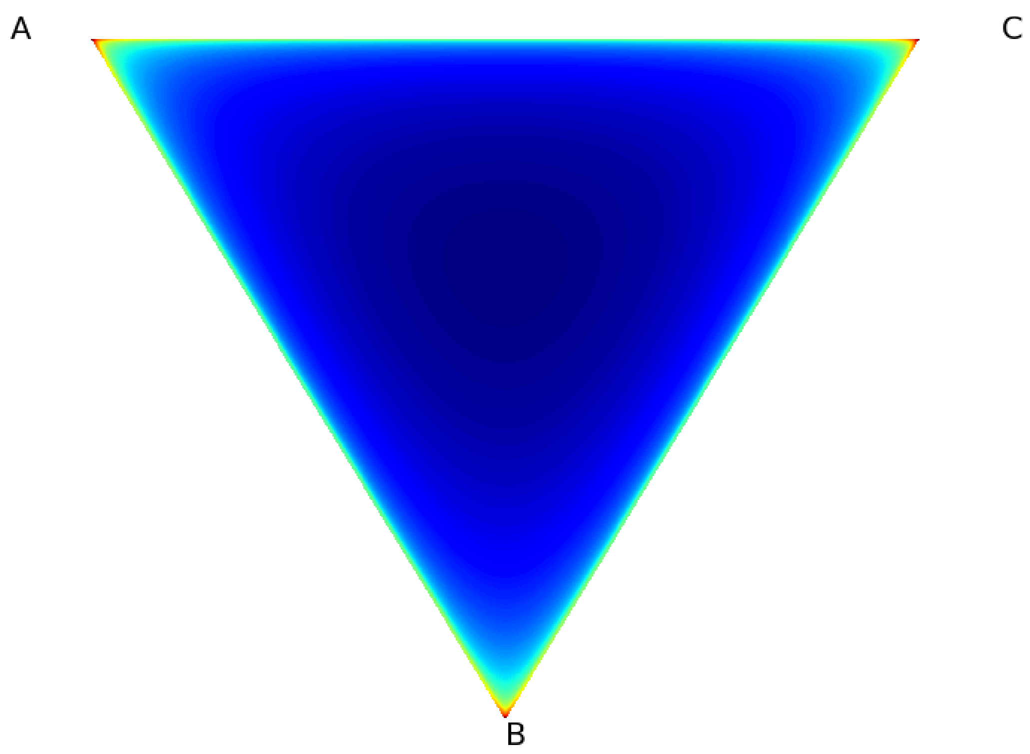
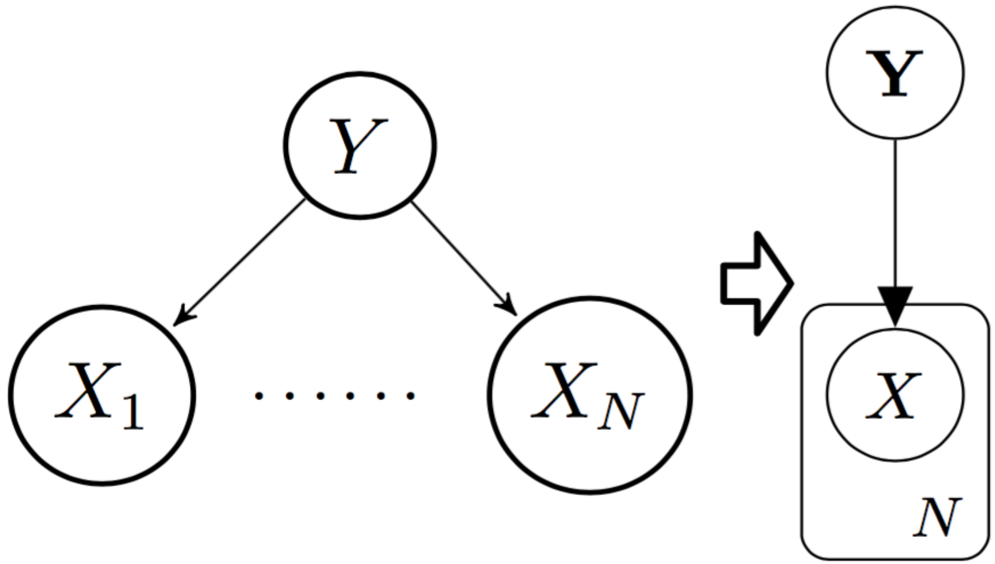


Plate Notation

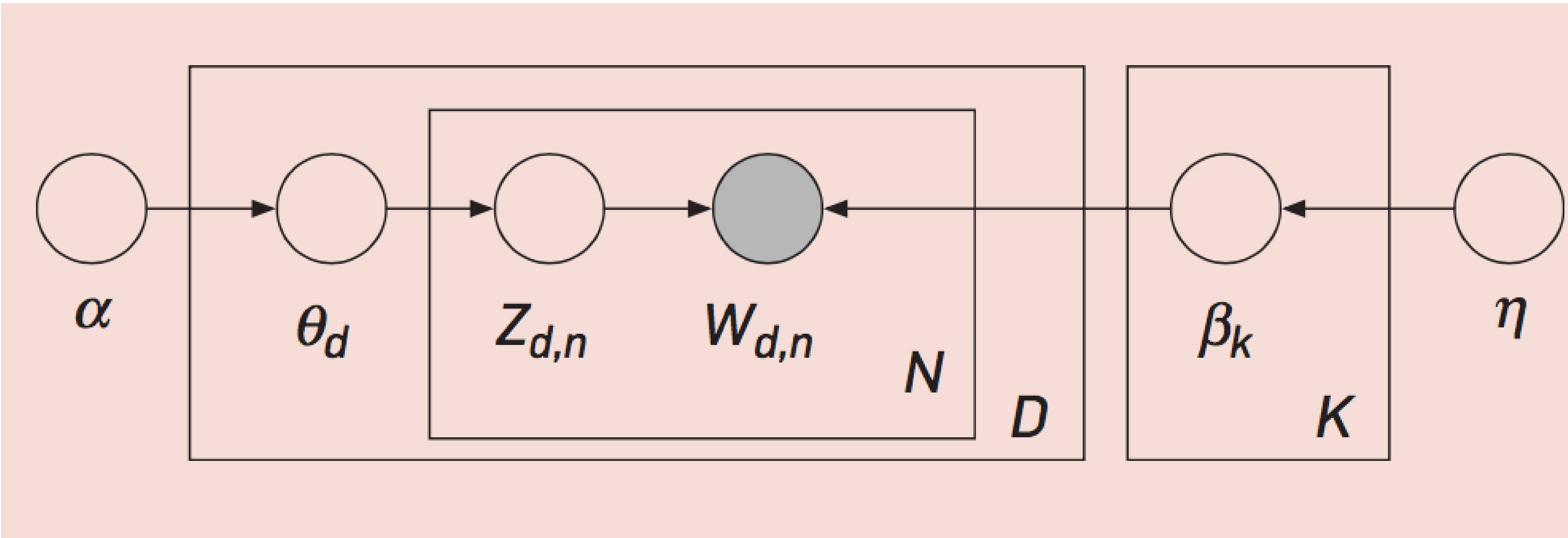
Karlsruhe Institute of Technology



- Vertex $\equiv$ random variable
- Edge $\equiv$ dependence

Institute of Theoretical Informatics
Algorithmics Group

Plate Notation: LDA



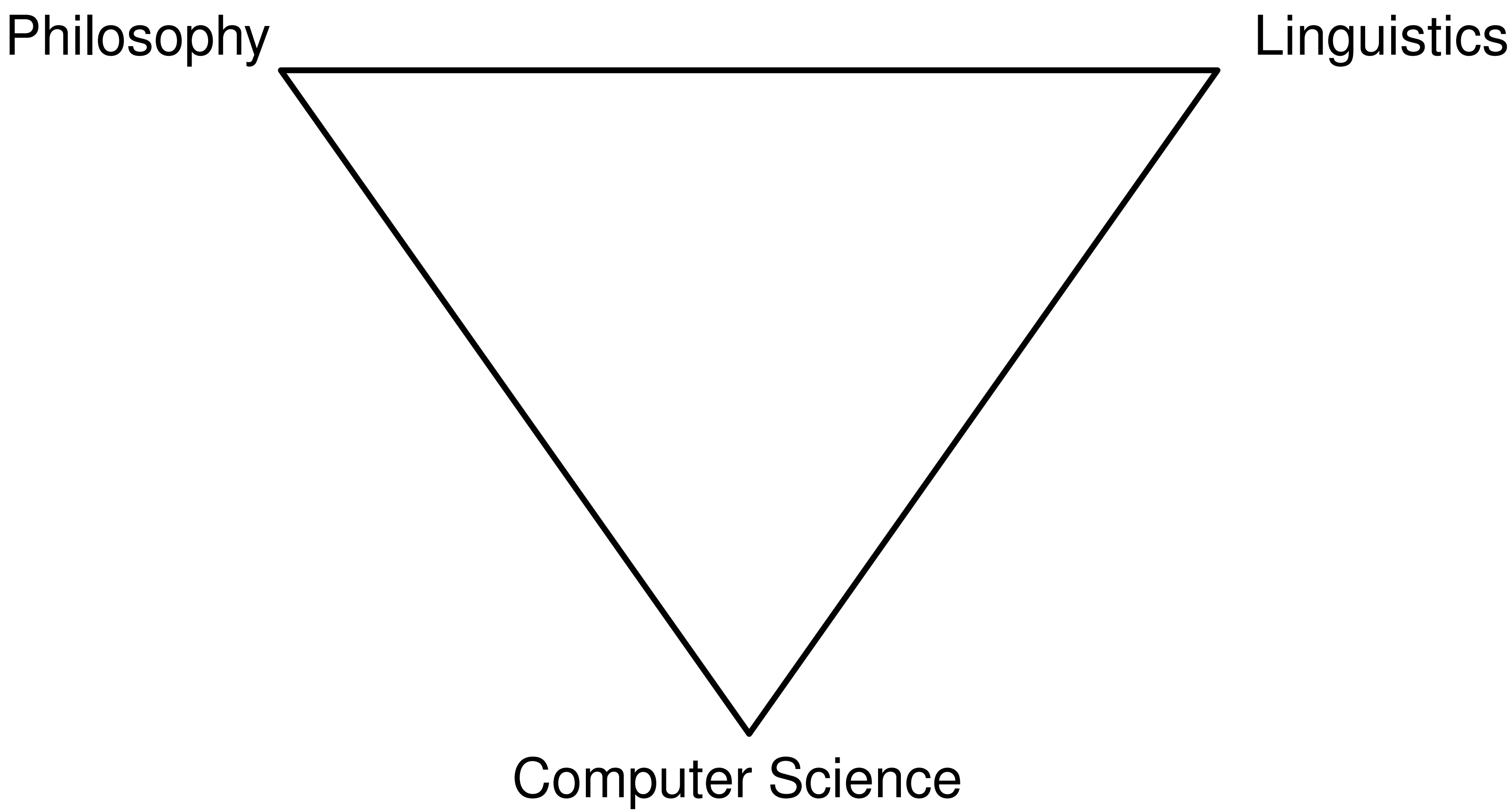
α - Dirichlet parameterization

β_k - topics (dist. over words)

θ_d - topic proportions for d^{th} document

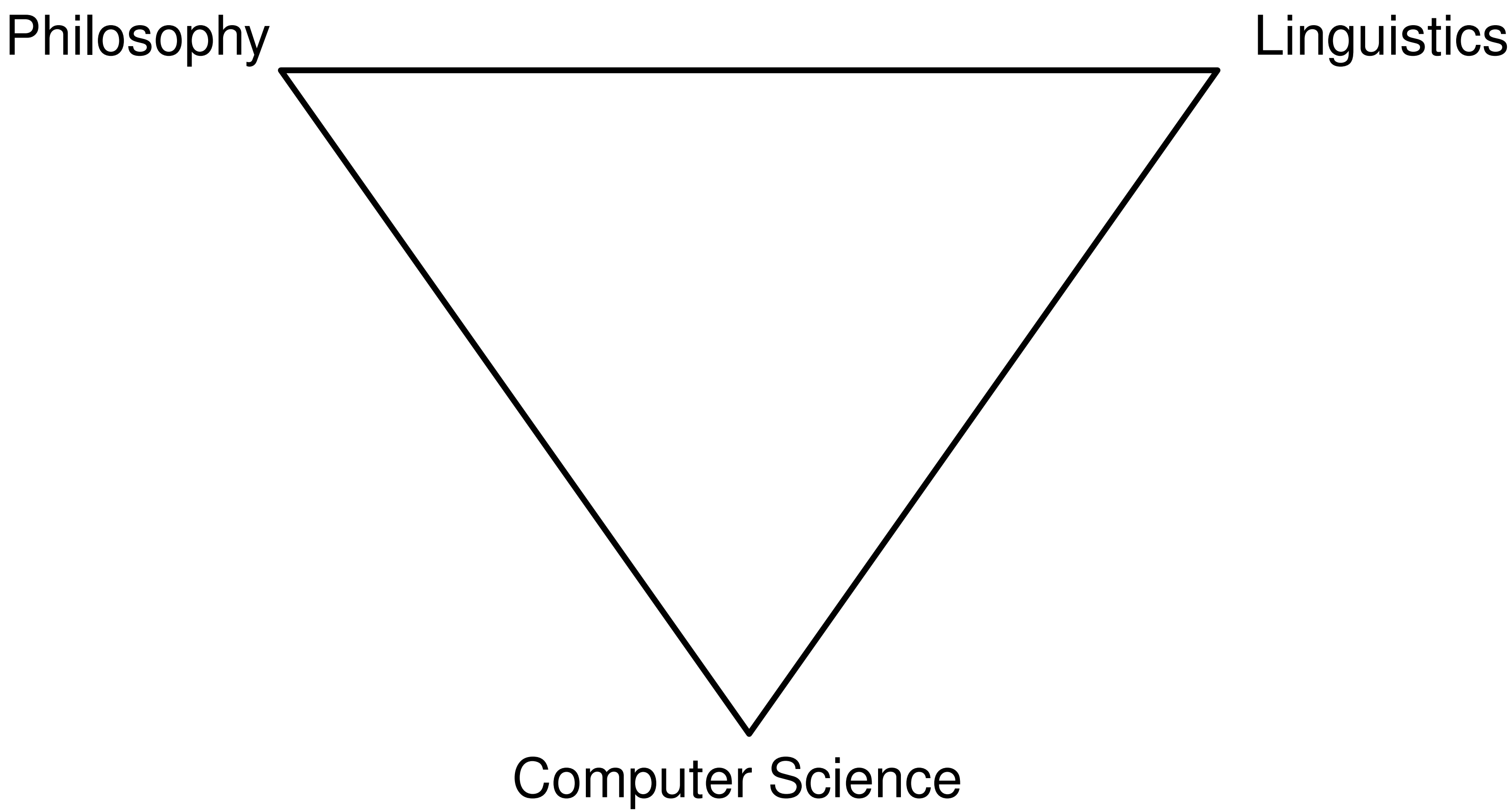
$z_{d,n}$ - topic assignment for n^{th} word in d^{th} document

LDA: Demo



LDA: Demo

Depending on how the corpus changes. . .



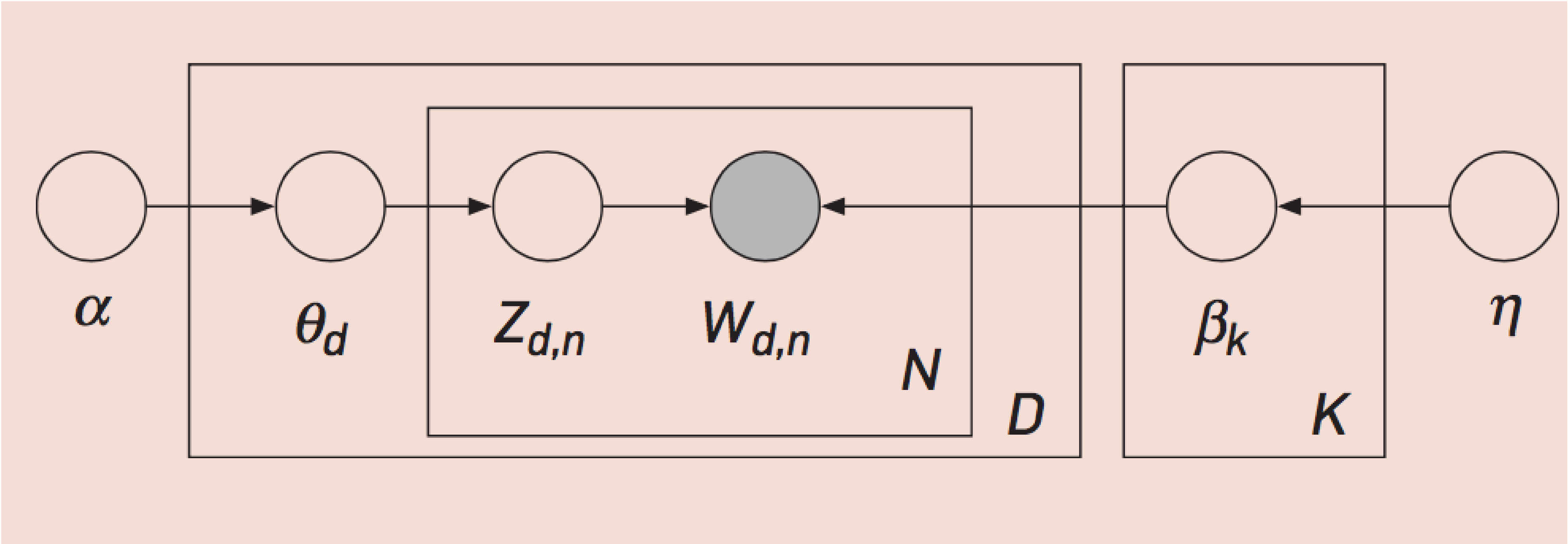
Institute of Theoretical Informatics
Algorithmics Group

LDA and Inference

- Goal: Automatically discover topics from a collection of documents
- Only documents themselves are *observed*
- topics, per-document topic distributions, and the per-document per-word topic assignments is *hidden*

LDA and Inference

- Goal: Automatically discover topics from a collection of documents
- Only documents themselves are *observed*
- topics, per-document topic distributions, and the per-document per-word topic assignments is *hidden*



Inference

Karlsruhe Institute of Technology

How to infer latent variables?

Inference

How to infer latent variables?

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

β_k - topics (dist. over words)
θ_d - topic proportions for d^{th} document
$z_{d,n}$ - topic assignment for n^{th} word in d^{th} document
w_d - observed words

Institute of Theoretical Informatics
Algorithmics Group

Inference

How to infer latent variables?

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

marginal probability

β_k - topics (dist. over words)
θ_d - topic proportions for d^{th} document
$z_{d,n}$ - topic assignment for n^{th} word in d^{th} document
w_d - observed words

Institute of Theoretical Informatics
Algorithmics Group

Inference

Problem: Marginal probability is intractable to compute

Institute of Theoretical Informatics
Algorithmics Group

Inference

Problem: Marginal probability is intractable to compute

Could only be computed theoretically:
Sum the joint distribution over every possible instance of the hidden topic structure.

Institute of Theoretical Informatics
Algorithmics Group

Inference

Karlsruhe Institute of Technology

Problem: Marginal probability is intractable to compute

Could only be computed theoretically:
Sum the joint distribution over every possible instance of the hidden topic structure.

In other words: Sum over all possible ways of assigning each observed word of the collection to one of the topics.

Inference

Problem: Marginal probability is intractable to compute

Could only be computed theoretically:
Sum the joint distribution over every possible instance of the hidden topic structure.

In other words: Sum over all possible ways of assigning each observed word of the collection to one of the topics.

$\Rightarrow$ Approximation !

Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling

- used for Bayesian inference
- randomized algorithm
- Markov Chain Monte Carlo Algorithm
- Method to find (good) topics

Gibbs Sampling - Example

Text mining algorithms can be used to find structure in text corpora like Plato's *dialogues*

Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling - Example

Text mining algorithms can be used to find structure in text corpora like Plato's *dialogues*

-	-	-	-	-	-	-
text	mining	algorithms	structure	corpora	Aristotle	dialogues

Gibbs Sampling - Example

Karlsruhe Institute of Technology

Text mining algorithms can be used to find structure in text corpora like Plato's *dialogues*

1. Randomly assign words to topics

1	3	2	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

Gibbs Sampling - Example

Text mining algorithms can be used to find structure in text corpora like Plato's *dialogues*

1. Randomly assign words to topics

1	3	2	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

2. Do this for all documents in corpus

Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling - Example

1	3	2	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

Gibbs Sampling - Example

1	3	2	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

	1	**2**	**3**
text	65	54	59
mining	21	4	12
algorithms	100	74	122
structure	20	12	14
corpora	5	2	12
Plato	35	33	42
dialogues	24	27	31

Gibbs Sampling - Example

1	3	2	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

	1	**2**	**3**
text	65	54	59
mining	21	4	12
algorithms	100	74	122
structure	20	12	14
corpora	5	2	12
Plato	35	33	42
dialogues	24	27	31

Counts from **all** documents

Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling - Example

sample word *algorithm*

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

	1	**2**	**3**
text	65	54	59
mining	21	4	12
algorithms	100	74	122
structure	20	12	14
corpora	5	2	12
Plato	35	33	42
dialogues	24	27	31

Counts from **all** documents

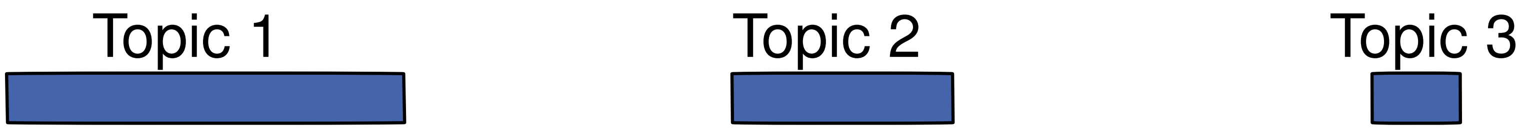
Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

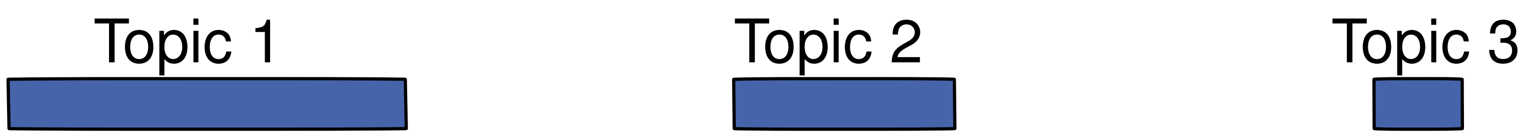
3. Topic distribution in this document



Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

3. Topic distribution in this document



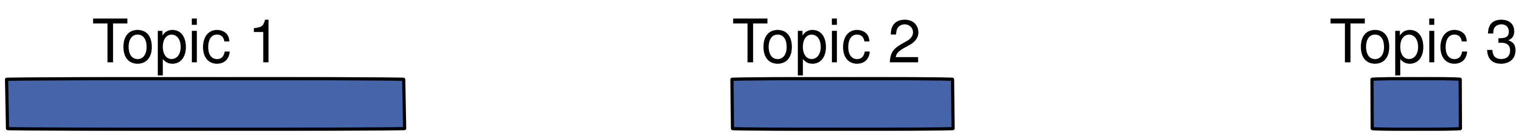
4. Word distribution over topics

Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

3. Topic distribution in this document



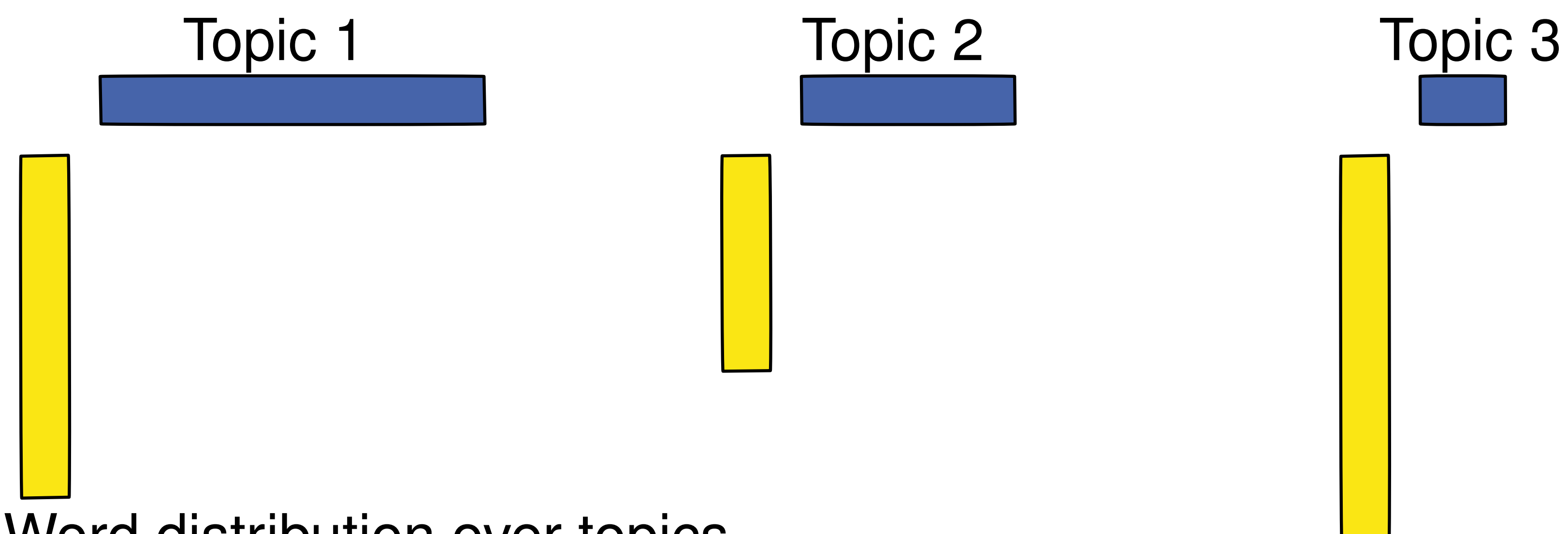
4. Word distribution over topics

	1	**2**	**3**
text	65	54	59
mining	21	4	12
algorithms	100	74	122
structure	20	12	14
corpora	5	2	12
Plato	35	33	42
dialogues	24	27	31

Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

3. Topic distribution in this document



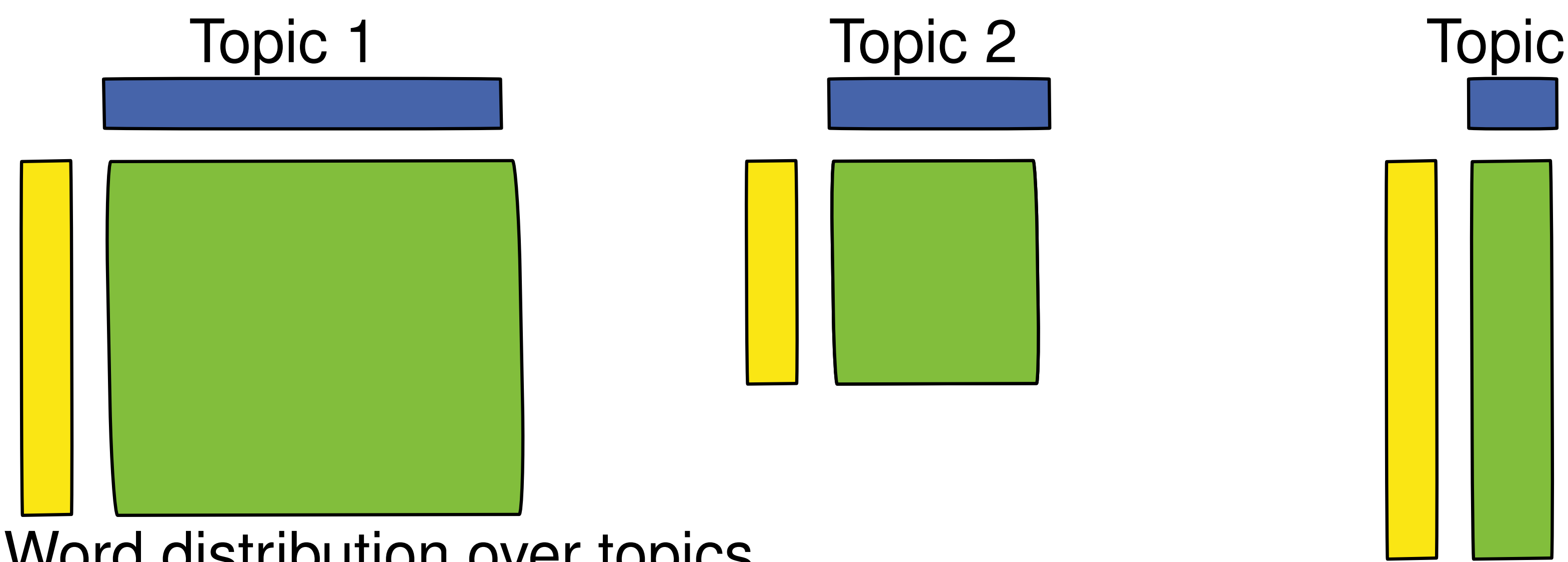
4. Word distribution over topics

Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

3. Topic distribution in this document

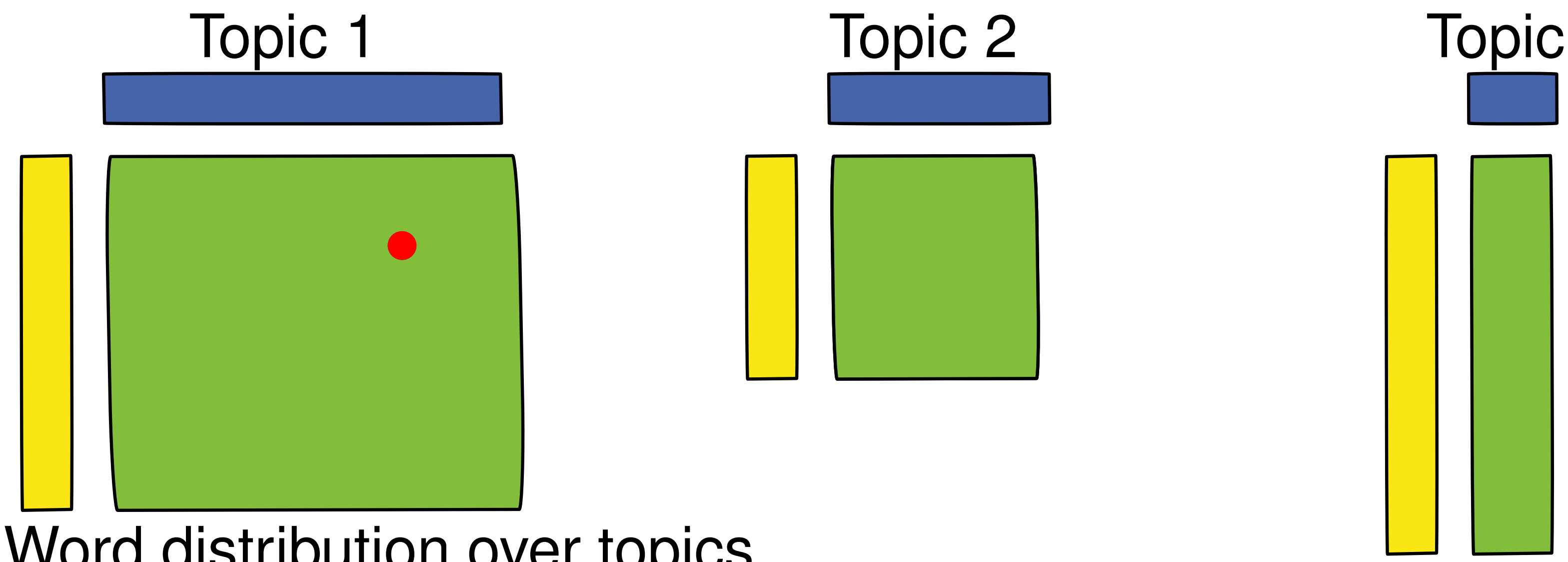


4. Word distribution over topics

Gibbs Sampling - Example

1	3	???	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

3. Topic distribution in this document



4. Word distribution over topics
5. Sample according to green area

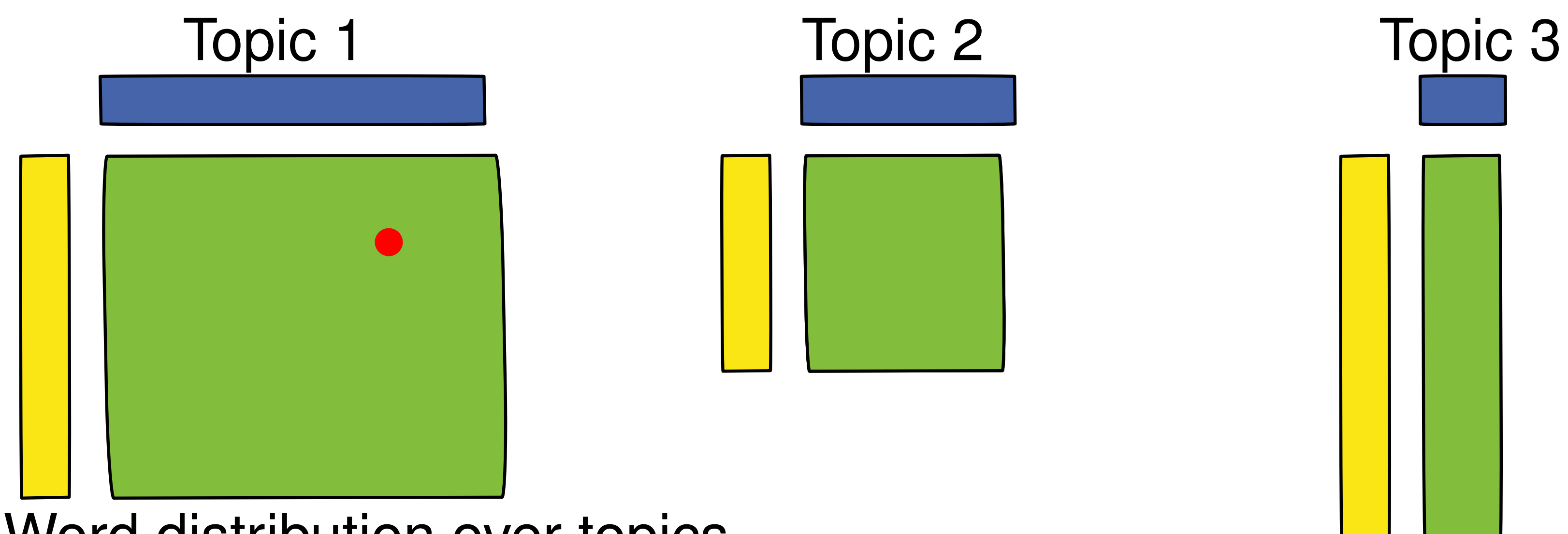
Institute of Theoretical Informatics
Algorithmics Group

Gibbs Sampling - Example

reassign to Topic 1

1	3	**1**	1	2	1	2
text	mining	algorithms	structure	corpora	Plato	dialogues

3. Topic distribution in this document



4. Word distribution over topics
5. Sample according to green area

Institute of Theoretical Informatics
Algorithmics Group

Conclusion - Take home message

Wrap up

Conclusion - Take home message

Wrap up

- Topic models find the hidden topical patterns that pervade a unstructured collection of text
  - Generative process as a model of how texts are composed
  - Words are **allocated** according a Dirichlet distribution over topics

Conclusion - Take home message

Karlsruhe Institute of Technology

Wrap up

- Topic models find the hidden topical patterns that pervade a unstructured collection of text
  - Generative process as a model of how texts are composed
  - Words are **allocated** according a Dirichlet distribution over topics
- Inference
  - Gibbs sampling can be used for approximating the hidden variables

Resources

- http://www.cs.columbia.edu/ blei/
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. ”Latent dirichlet allocation.” the Journal of machine Learning research 3 (2003): 993-1022. APA
- Porteous, Ian, et al. ”Fast collapsed gibbs sampling for latent dirichlet allocation.” Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.