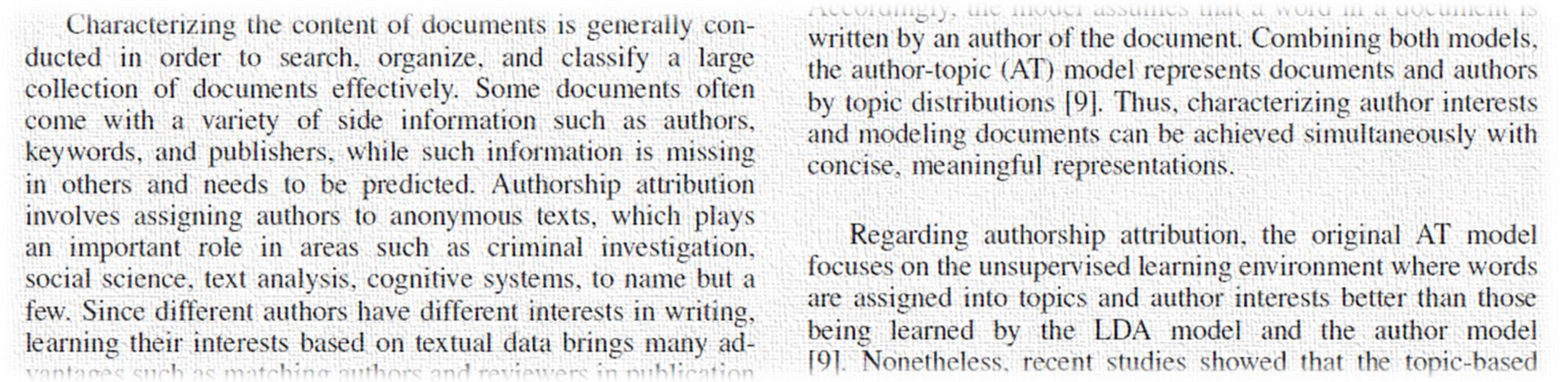

Characterizing the content of documents is generally conducted in order to search, organize, and classify a large collection of documents effectively. Some documents often come with a variety of side information such as authors, keywords, and publishers, while such information is missing in others and needs to be predicted. Authorship attribution involves assigning authors to anonymous texts, which plays an important role in areas such as criminal investigation, social science, text analysis, cognitive systems, to name but a few. Since different authors have different interests in writing, learning their interests based on textual data brings many ad-

written by an author of the document. Combining both models, the author-topic (AT) model represents documents and authors by topic distributions [9]. Thus, characterizing author interests and modeling documents can be achieved simultaneously with concise, meaningful representations.

Regarding authorship attribution, the original AT model focuses on the unsupervised learning environment where words are assigned into topics and author interests better than those being learned by the LDA model and the author model [9]. Nonetheless, recent studies showed that the topic-based

Who Wrote This? Textual Modeling with Authorship Attribution in Big Data

---

Naruemon Pratanwanich (**Ploy**)

Dr. Pietro Lio'

14 Dec 2014

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

2nd International Workshop on High Dimensional Data Mining (HDM'14)

Dec 14-17, 2014 ICDM SHENZHEN, CHINA

Overview

- **Goals**:
  - To predict authors of a given document
  - To discover new knowledge *i.e. author contribution, author interests,* and the underlying *topics*
- **Our belief**:
  - A document and its authors have the overlapping sets of the underlying topics.
- **Latent variables**:
  - Per-document author distribution
  - Per-author topic distribution
  - Per-topic word distribution

Overview

- **Goals**:
  - To predict authors of a given document
  - To discover new knowledge *i.e. author contribution, author interests,* and the underlying *topics*
- **Our belief**:
  - A document and its authors have the overlapping sets of the underlying topics.
- **Latent variables**:
  - Per-document author distribution
  - Per-author topic distribution
  - Per-topic word distribution

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Overview

- **Goals**:
  - To predict authors of a given document
  - To discover new knowledge *i.e. author contribution*, *author interests*, and the underlying *topics*
- **Our belief**:
  - A document and its authors have the overlapping sets of the underlying topics.
- **Latent variables**:
  - Per-document author distribution
  - Per-author topic distribution
  - Per-topic word distribution

Outline

- **Introduction**
  - Probabilistic generative models (LDA and AT models)
- **Our model**
  - Supervised Author-Topic (SAT) model
- **Results**
  - Model fitness
  - Information discovery
  - Model performance for supervised learning
- **Conclusion**
  - Applications

INTRODUCTION

Probabilistic Generative Models

- Latent Dirichlet Allocation (LDA)
- Author-Topic (AT) model

Latent Dirichlet Allocation (LDA)

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Given MRI images,
support vector machine
is a technique used to
classify diseased cells
and healthy cells.

Given X-ray images,
neural network
is an algorithm used to
identify abnormal bones
and normal bones.

Email: np394@cam.ac.uk

Latent Dirichlet Allocation (LDA)

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

Latent Dirichlet Allocation (LDA)

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

A document

For each position:
  choose a **topic**
  choose a word in the topic

Latent Dirichlet Allocation (LDA)

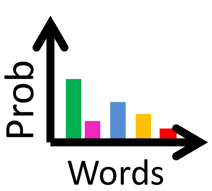
- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

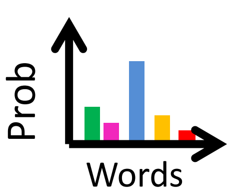
Given MRI images,
support vector machine
is a technique used to
classify diseased cells
and healthy cells.

Given X-ray images,
neural network
is an algorithm used to
identify abnormal bones
and normal bones.

A document = 

For each position:
choose a **topic**
choose a word in the topic

Topic ■ 



Topic ■

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Author-Topic (AT) Model

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Author: Ploy and Lio'

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

Author: Ploy and James

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

Author-Topic (AT) Model

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Author: Ploy and Lio'

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

Author: Ploy and James

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

A document = {author$_i$}

For each position:
choose an **author**
choose a **topic**
choose a word in the topic

Author-Topic (AT) Model

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Author: Ploy and Lio’

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

Author: Ploy and James

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

A document = {author$_i$}

For each position:
choose an **author**
choose a **topic**
choose a word in the topic

Author-Topic (AT) Model

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Author: Ploy and Lio'

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

Author: Ploy and James

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

A document = {author$_i$}

For each position:
  choose an **author**
  choose a **topic**
  choose a word in the topic

Email: np394@cam.ac.uk

Author-Topic (AT) Model

- Purpose in text mining: organise, search, etc.
- Generative processes of writing a document

Author: Ploy and Lio’

Given MRI images, support vector machine is a technique used to classify diseased cells and healthy cells.

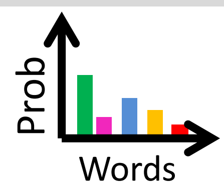
Author: Ploy and James

Given X-ray images, neural network is an algorithm used to identify abnormal bones and normal bones.

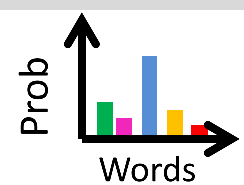
A document = $\{\text{author}_i\}$

For each position:
 choose an **author**
 choose a **topic**
 choose a word in the topic

Topic ■



Topic ■



Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

OUR MODEL

Supervised Author-Topic (SAT) Model

Inference method

Author-Topic (AT) Model

- Equal author contributions

Given MRI images,
support vector machine
is a technique used to
classify diseased cells
and healthy cells.

Given X-ray images,
neural network
is an algorithm used to
identify abnormal bones
and normal bones.

A document = {author$_i$}

For each position:
choose an **author**
choose a **topic**
choose a word in the topic

Author-Topic (AT) Model

- Equal author contributions

Given MRI images,
support vector machine
is a technique used to
classify diseased cells
and healthy cells.

Given X-ray images,
neural network
is an algorithm used to
identify abnormal bones
and normal bones.

A document = {author$_i$}

For each position:
choose an **author** **Uniform**
choose a **topic**
choose a word in the topic

Supervised Author-Topic (SAT) Model

- Unequal author contributions

Given MRI images,
support vector machine
is a technique used to
classify diseased cells
and healthy cells.

Given X-ray images,
neural network
is an algorithm used to
identify abnormal bones
and normal bones.

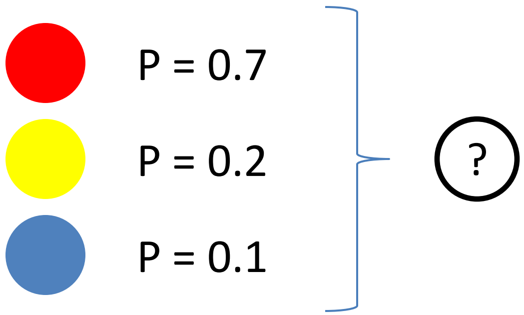
A document = {(author$_i$,p$_i$)}

For each position:
choose an **author** ~ **p**
choose a **topic**
choose a word in the topic

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Distributions

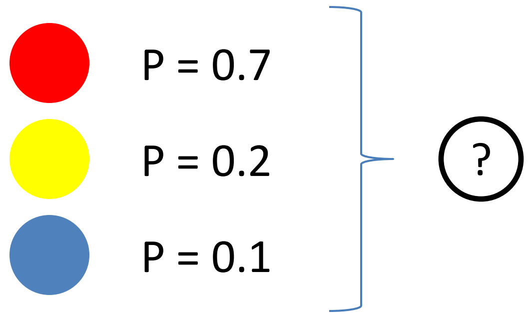
- Multinomial distribution is ..



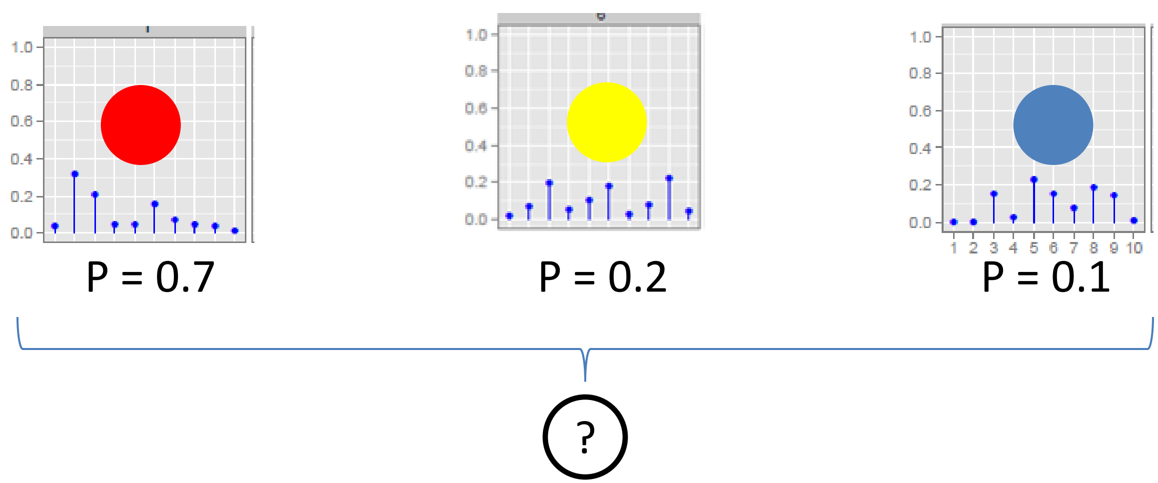
Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Distributions

- Multinomial distribution is ..

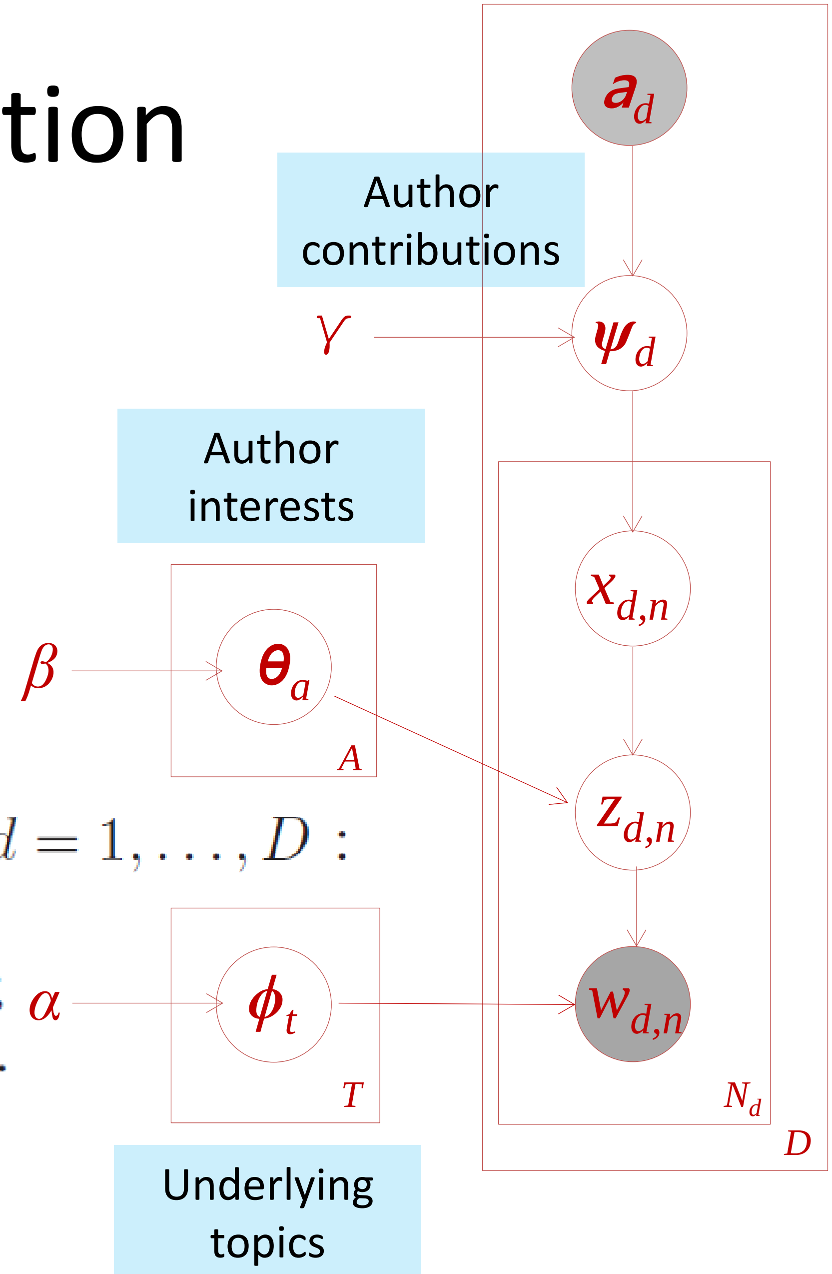


- Dirichlet is a distribution of distribution***s***



Model Description

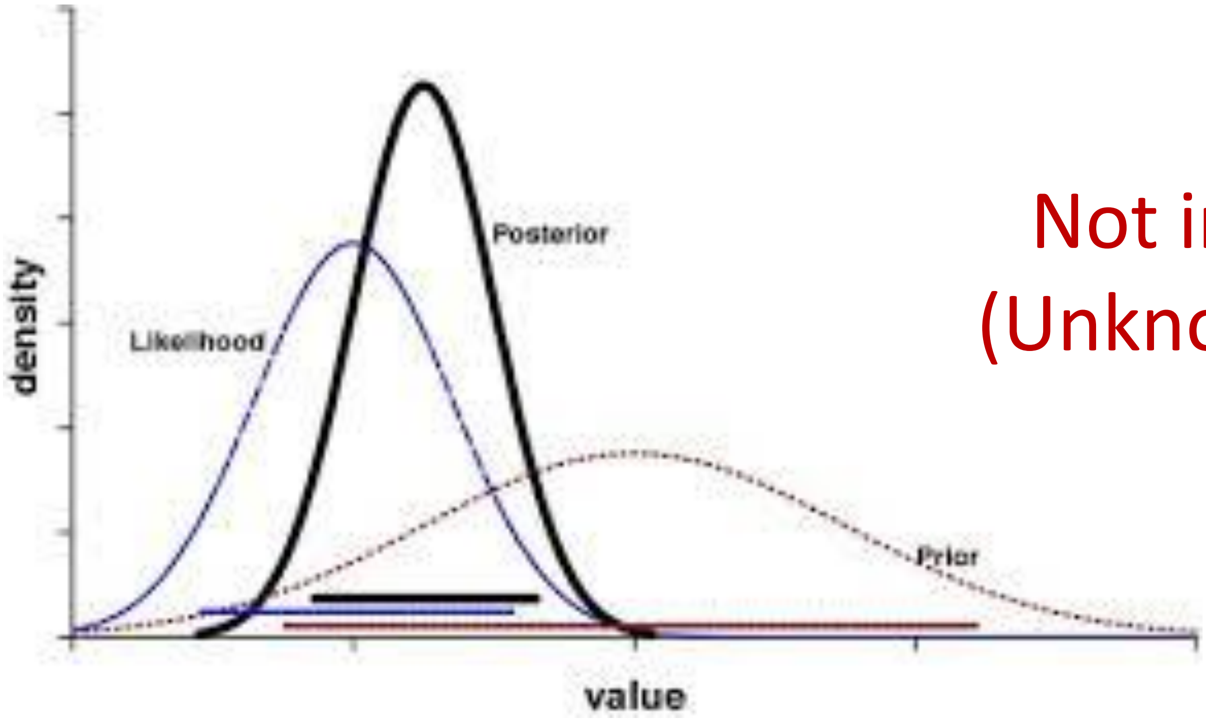
1) For each topic $t = 1, \ldots, T$:
   a) $\phi_t \sim \text{Dirichlet}(\alpha)$;
2) For each author $a = 1, \ldots, A$:
   a) $\boldsymbol{\theta}_a \sim \text{Dirichlet}(\beta)$;
3) For each document $d = 1, \ldots, D$:
   a) $\psi_d \sim \text{Dirichlet}(\gamma \mathbf{a}_d)$;
4) For each word $n = 1, \ldots, N_d$ and $d = 1, \ldots, D$:
   a) $x_{d,n} \sim \text{Multinomial}(\psi_d)$;
   b) $z_{d,n} \sim \text{Multinomial}(\boldsymbol{\theta}_{x_{d,n}})$;
   c) $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$.



Bayesian Inference

- Bayes' theorem ($\mathcal{M}$ is the set of latent variables)

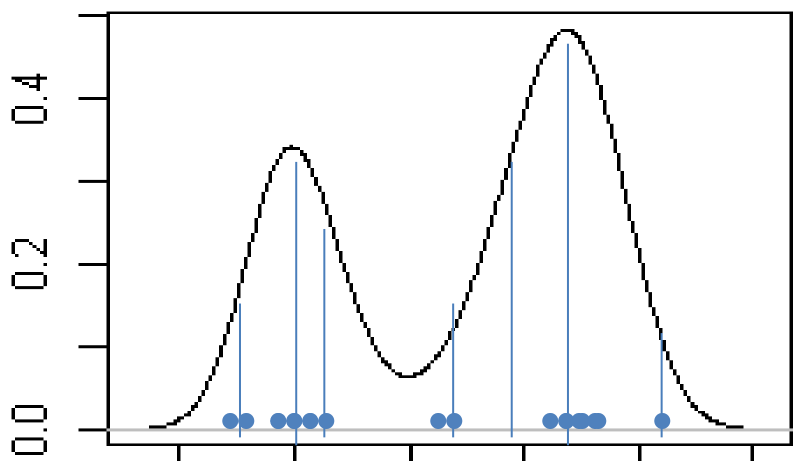
$$P(\mathcal{M}|data) = \frac{P(data|\mathcal{M}) \times P(\mathcal{M})}{P(data)}$$

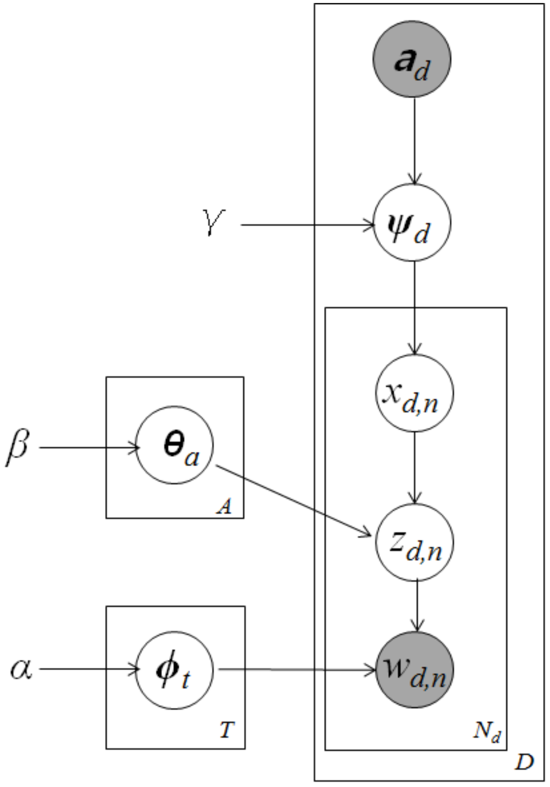


Not in a closed form
(Unknown distribution)

Approximate Inference

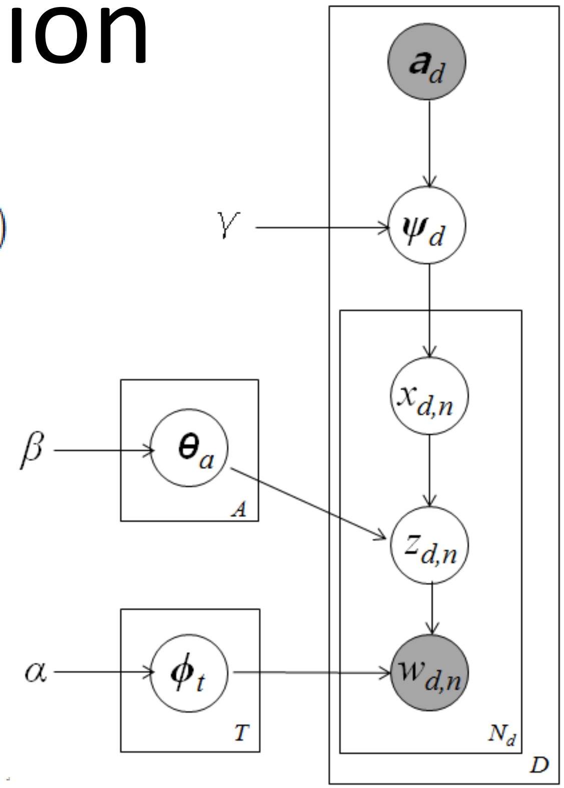
- Sampling method - Gibbs sampling
  - Iteratively random one variable, given others fixed
  - Infinite time -> True distribution





Posterior Distribution

$$P(x_{d,n} = a, z_{d,n} = t | w_{d,n} = w, \boldsymbol{w}_{\backslash(d,n)}, \boldsymbol{x}_{\backslash(d,n)}, \boldsymbol{z}_{\backslash(d,n)}, \alpha, \beta, \gamma, \mathbf{a}_d)$$
$$\propto \frac{C^{TV}_{t,v,\backslash(d,n)} + \alpha}{\sum_{v'} C^{TV}_{t,v',\backslash(d,n)} + V\alpha} \times \frac{C^{AT}_{a,t,\backslash(d,n)} + \beta}{\sum_{t'} C^{AT}_{a,t',\backslash(d,n)} + T\beta} \times \frac{C^{DA}_{d,a,\backslash(d,n)} + \gamma^{a_{d,a}}}{\sum_{a'} (C^{DA}_{d,a',\backslash(d,n)} + \gamma^{a_{d,a'}})}$$



Intuitively, the probability of assigning a word w to a topic t written by an author a depends on three probabilities

- how likely the word w belongs to the topic t
- how likely the topic t is written by the author a
- how likely the author a contributes to the document d

RESULTS

Convergence Analysis & Model Fitness

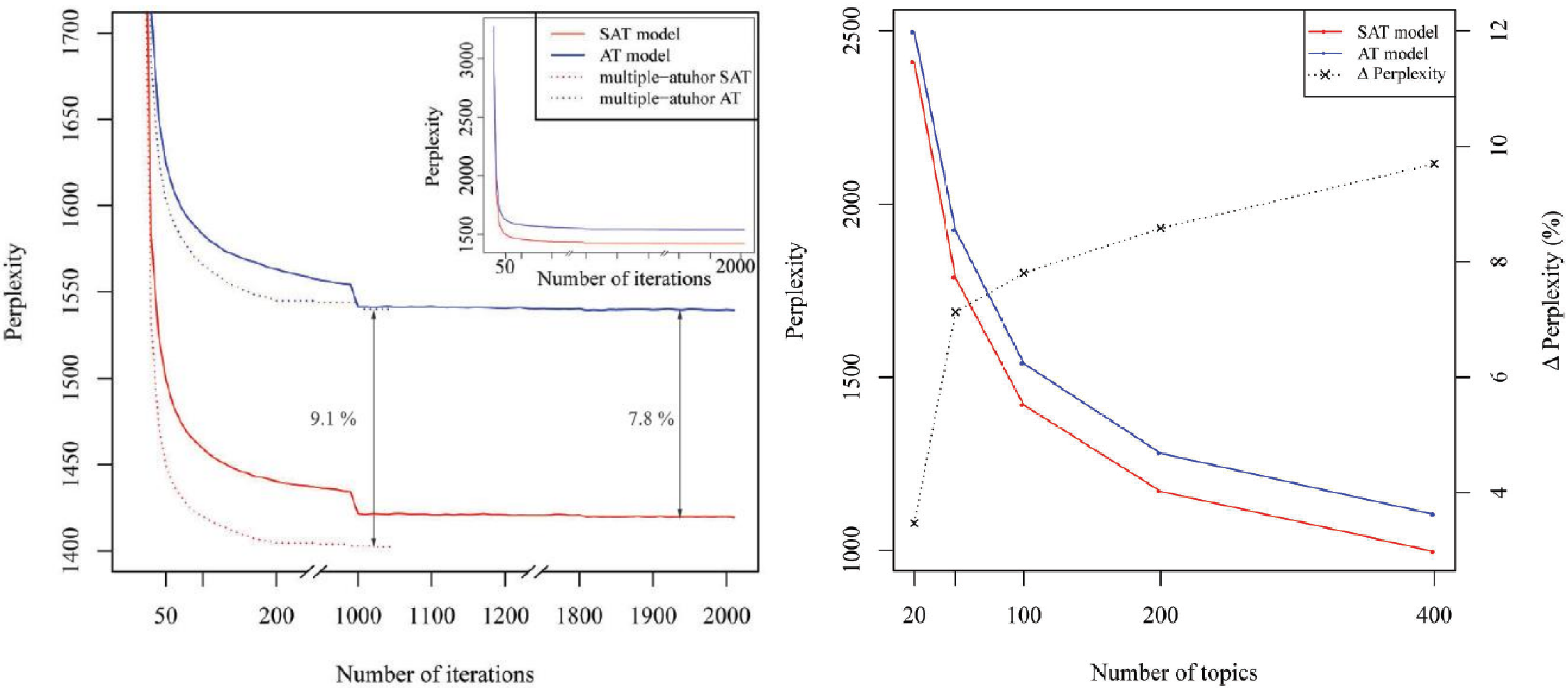
Information Discovery

Supervised Learning

Email: np394@cam.ac.uk

Convergence Analysis

$$perplexity = \exp\left(-\frac{\sum_{d=1}^{D} \log p(w_d|\alpha, \beta, \gamma, \mathbf{a}_d)}{\sum_{d=1}^{D} N_d}\right)$$



(a) (b)

Author Interests

Author: Cole R		Author: Agin P		Author: Ghahramani Z		Author: MacKay D		Author: Bishop C	
Topics	Probability	Topics	Probability	Topics	Probability	Topics	Probability	Topics	Probability
64	0.895	58	0.842	22	0.551	53	0.335	53	0.379
11	0.070	78	0.152	63	0.180	78	0.20	78	0.334
				52	0.098	22	0.183	22	0.189
				78	0.071	10	0.093	86	0.067
				87	0.036			63	0.027

Topics

Topic: 10 - Kernel learning		Topic: 11 - Classification		Topic: 22-Maximum likelihood		Topic : 48 - Face recognition	
Words	Probability	Words	Probability	Words	Probability	Words	Probability
rbf	0.048	classification	0.048	model	0.040	face	0.046
experts	0.028	classifier	0.045	data	0.025	images	0.024
basis	0.024	class	0.039	models	0.021	faces	0.022
expert	0.021	training	0.031	probability	0.019	recognition	0.020
gating	0.019	classifiers	0.026	likelihood	0.015	facial	0.017
network	0.018	classes	0.016	mixture	0.014	image	0.017
radial	0.017	feature	0.015	distribution	0.013	human	0.009
networks	0.017	pattern	0.014	parameters	0.013	based	0.009
mixture	0.016	decision	0.012	em	0.012	view	0.008
gaussian	0.013	nearest	0.011	density	0.011	system	0.008
Topic: 52 - Gradient algorithm		**Topic: 53 - Bayesian/Monte Carlo**		**Topic: 58 - Protein structure**		**Topic: 63 - Network**	
Words	Probability	Words	Probability	Words	Probability	Words	Probability
function	0.024	bayesian	0.028	protein	0.023	network	0.050
algorithm	0.019	gaussian	0.025	chain	0.021	units	0.031
learning	0.017	prior	0.022	region	0.016	input	0.030
gradient	0.013	posterior	0.019	structure	0.015	learning	0.025
vector	0.012	distribution	0.016	mouse	0.014	output	0.023
convergence	0.010	evidence	0.014	proteins	0.014	training	0.023
problem	0.009	monte	0.012	human	0.013	hidden	0.023
linear	0.009	carlo	0.012	sequences	0.012	networks	0.022
case	0.008	mackay	0.009	prediction	0.010	unit	0.019
algorithms	0.008	noise	0.009	sequence	0.010	layer	0.019
Topic: 64 - Speech recognition		**Topic: 78**		**Topic: 86 - PCA**		**Topic: 87 - Statistical mechanics**	
Words	Probability	Words	Probability	Words	Probability	Words	Probability
speech	0.039	data	0.022	matrix	0.038	energy	0.036
recognition	0.029	set	0.019	pca	0.028	boltzmann	0.028
word	0.023	number	0.013	linear	0.026	temperature	0.020
system	0.018	figure	0.012	principal	0.025	annealing	0.017
training	0.015	results	0.012	analysis	0.017	units	0.013
hmm	0.013	model	0.012	component	0.017	state	0.012
speaker	0.012	neural	0.011	components	0.015	field	0.011
context	0.011	learning	0.010	covariance	0.013	machine	0.011
network	0.009	function	0.009	eigenvectors	0.012	probability	0.008
neural	0.008	training	0.009	subspace	0.011	signature	0.008

Author Interests

Author: Cole R		Author: Agin P		Author: Ghahramani Z		Author: MacKay D		Author: Bishop C	
Topics	Probability	Topics	Probability	Topics	Probability	Topics	Probability	Topics	Probability
64	0.895	58	0.842	22	0.551	53	0.335	53	0.379
11	0.070	78	0.152	63	0.180	78	0.20	78	0.334
				52	0.098	22	0.183	22	0.189
				78	0.071	10	0.093	86	0.067
				87	0.036			63	0.027

Topics

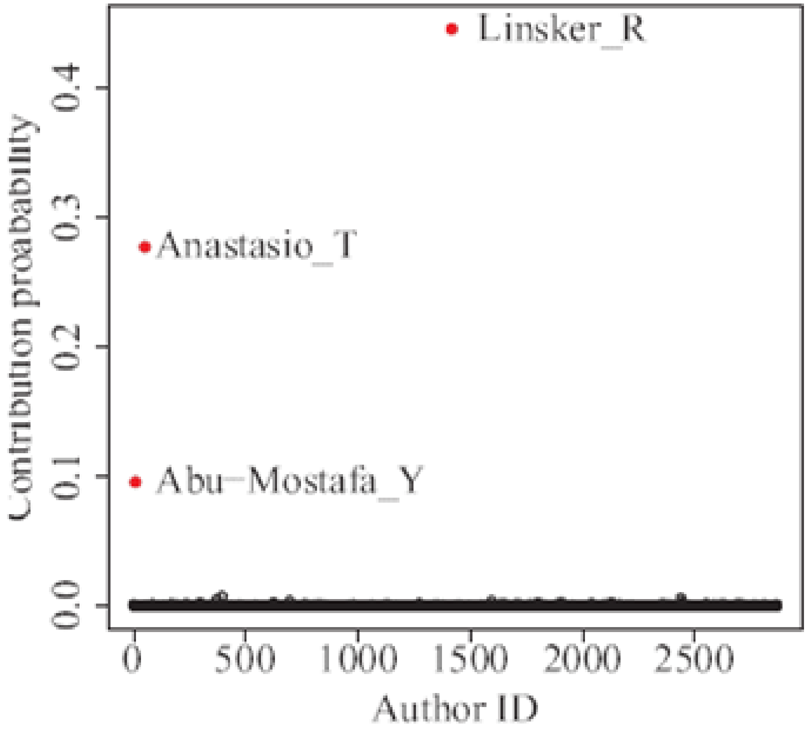
Topic: 10 - Kernel learning		Topic: 11 - Classification		Topic: 22-Maximum likelihood		Topic : 48 - Face recognition	
Words	Probability	Words	Probability	Words	Probability	Words	Probability
rbf	0.048	classification	0.048	model	0.040	face	0.046
experts	0.028	classifier	0.045	data	0.025	images	0.024
basis	0.024	class	0.039	models	0.021	faces	0.022
expert	0.021	training	0.031	probability	0.019	recognition	0.020
gating	0.019	classifiers	0.026	likelihood	0.015	facial	0.017
network	0.018	classes	0.016	mixture	0.014	image	0.017
radial	0.017	feature	0.015	distribution	0.013	human	0.009
networks	0.017	pattern	0.014	parameters	0.013	based	0.009
mixture	0.016	decision	0.012	em	0.012	view	0.008
gaussian	0.013	nearest	0.011	density	0.011	system	0.008

Topic: 52 - Gradient algorithm		Topic: 53 - Bayesian/Monte Carlo		Topic: 58 - Protein structure		Topic: 63 - Network	
Words	Probability	Words	Probability	Words	Probability	Words	Probability
function	0.024	bayesian	0.028	protein	0.023	network	0.050
algorithm	0.019	gaussian	0.025	chain	0.021	units	0.031
learning	0.017	prior	0.022	region	0.016	input	0.030
gradient	0.013	posterior	0.019	structure	0.015	learning	0.025
vector	0.012	distribution	0.016	mouse	0.014	output	0.023
convergence	0.010	evidence	0.014	proteins	0.014	training	0.023
problem	0.009	monte	0.012	human	0.013	hidden	0.023
linear	0.009	carlo	0.012	sequences	0.012	networks	0.022
case	0.008	mackay	0.009	prediction	0.010	unit	0.019
algorithms	0.008	noise	0.009	sequence	0.010	layer	0.019

Topic: 64 - Speech recognition		Topic: 78		Topic: 86 - PCA		Topic: 87 - Statistical mechanics	
Words	Probability	Words	Probability	Words	Probability	Words	Probability
speech	0.039	data	0.022	matrix	0.038	energy	0.036
recognition	0.029	set	0.019	pca	0.028	boltzmann	0.028
word	0.023	number	0.013	linear	0.026	temperature	0.020
system	0.018	figure	0.012	principal	0.025	annealing	0.017
training	0.015	results	0.012	analysis	0.017	units	0.013
hmm	0.013	model	0.012	component	0.017	state	0.012
speaker	0.012	neural	0.011	components	0.015	field	0.011
context	0.011	learning	0.010	covariance	0.013	machine	0.011
network	0.009	function	0.009	eigenvectors	0.012	probability	0.008
neural	0.008	training	0.009	subspace	0.011	signature	0.008

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Author Contributions

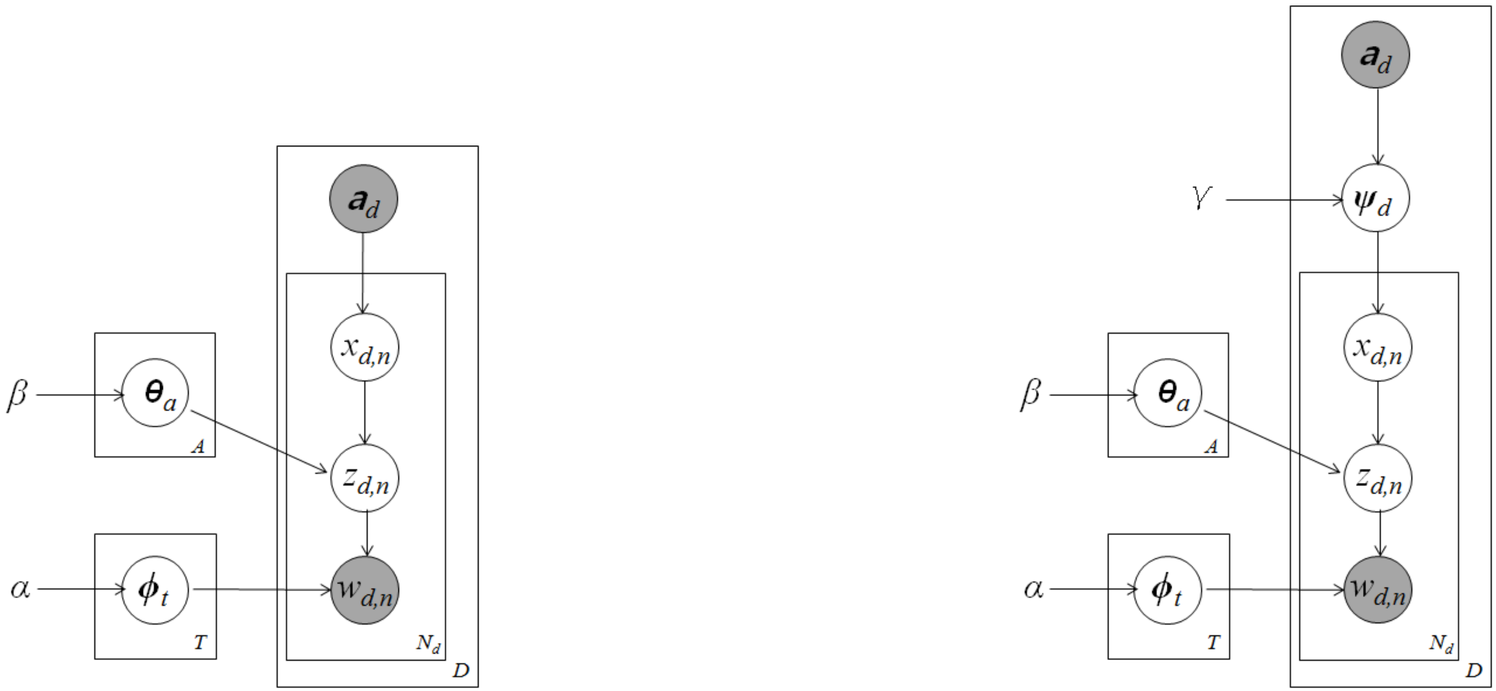


Test document

- 1 document from Abu-Mostafa Y (0.17%)
- 2 documents from Anastasio T (0.33%)
- 3 documents from Linsker R (0.50%)

Who Wrote This ?

To discover the meaningful structures underlying documents, probabilistic generative models which employ the abstract definition of topics as a fundamental concept to generate words have gained popularity for document analysis as unsupervised learning techniques. In topic-based generative models, a document is described by a particular topic proportion, where a topic is defined as a distribution over words. After latent Dirichlet allocation (LDA), a mixed-membership topic model, was introduced, many studies have proposed a great number of model variations. The primary goal of such extensions is to incorporate side information or meta-data together with words in the texts for better characterization of



(a) AT model

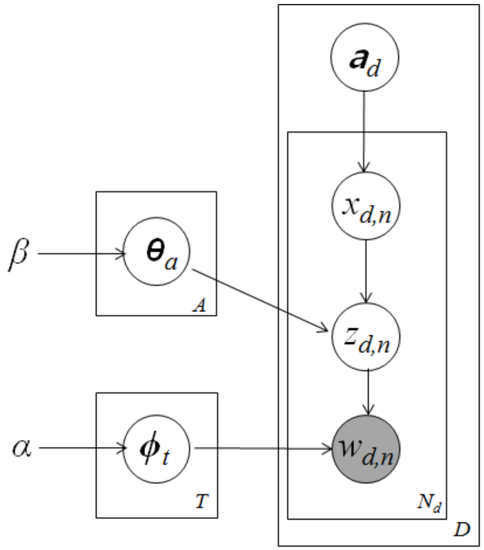
(b) SAT model

Naruemon Pratanwanich (Ploy)

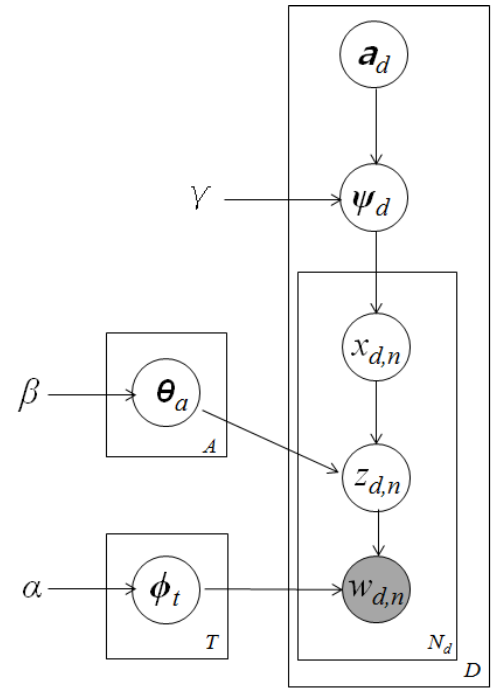
Email: np394@cam.ac.uk

Who Wrote This ?

To discover the meaningful structures underlying documents, probabilistic generative models which employ the abstract definition of topics as a fundamental concept to generate words have gained popularity for document analysis as unsupervised learning techniques. In topic-based generative models, a document is described by a particular topic proportion, where a topic is defined as a distribution over words. After latent Dirichlet allocation (LDA), a mixed-membership topic model, was introduced, many studies have proposed a great number of model variations. The primary goal of such extensions is to incorporate side information or meta-data together with words in the texts for better characterization of



(a) AT model



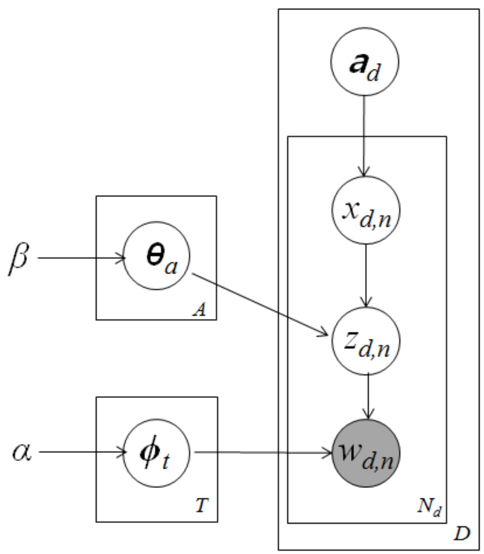
(b) SAT model

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

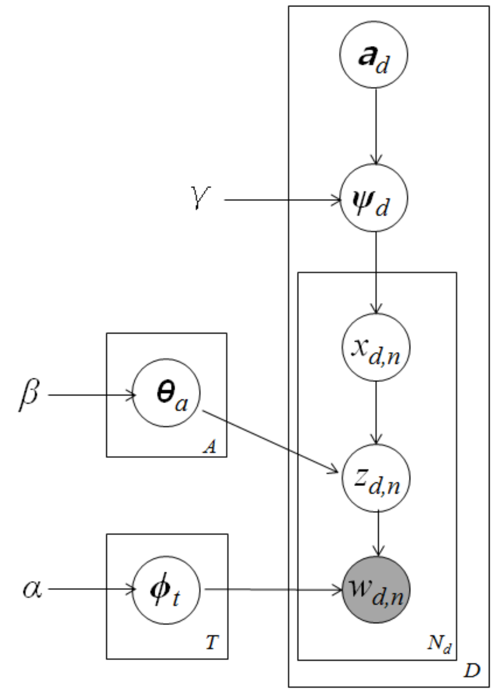
Who Wrote This ?

To discover the meaningful structures underlying documents, probabilistic generative models which employ the abstract definition of topics as a fundamental concept to generate words have gained popularity for document analysis as unsupervised learning techniques. In topic-based generative models, a document is described by a particular topic proportion, where a topic is defined as a distribution over words. After latent Dirichlet allocation (LDA), a mixed-membership topic model, was introduced, many studies have proposed a great number of model variations. The primary goal of such extensions is to incorporate side information or meta-data together with words in the texts for better characterization of

$$\tilde{\psi}_d = P(a \in \mathbf{a}_d | \tilde{\mathbf{w}}_d, \mathcal{M}) = \frac{\sum_{n=1}^{N_d} \delta(\tilde{x}_{d,n} = a)}{N_d}$$



(a) AT model

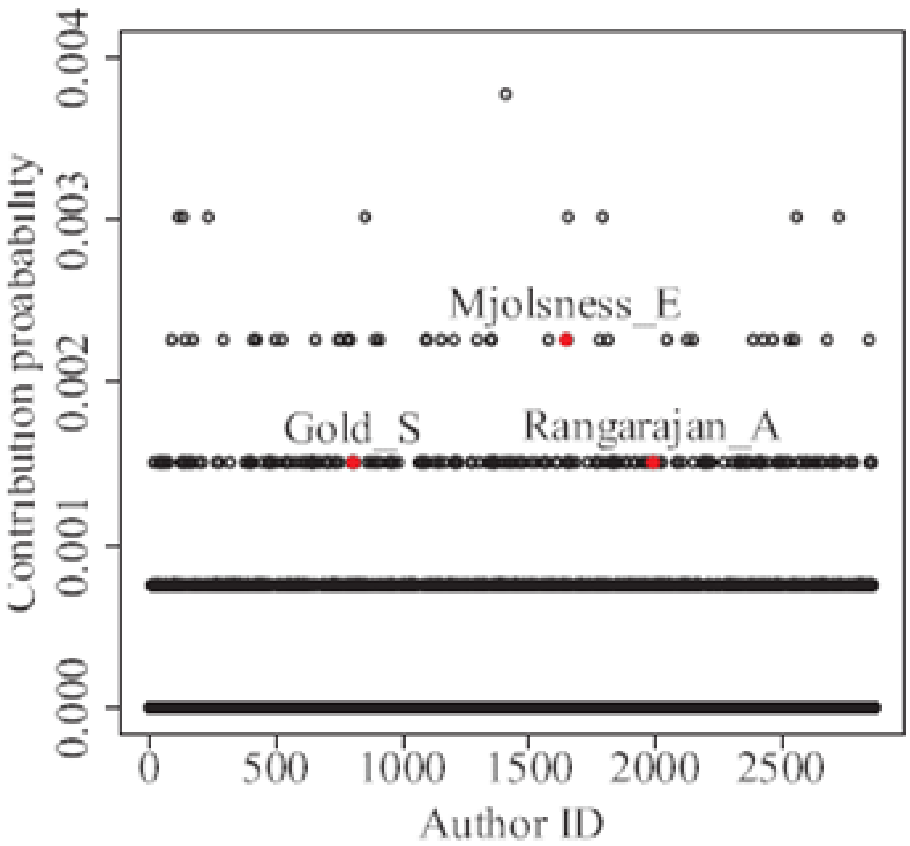


(b) SAT model

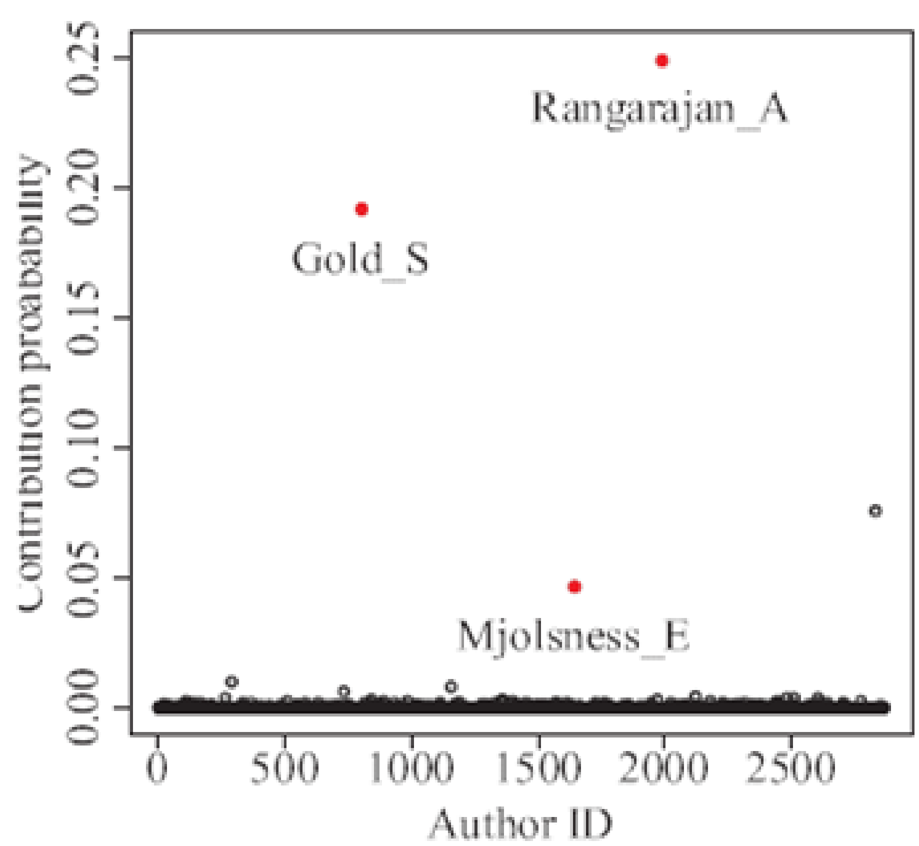
Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Example of Predictive Distribution

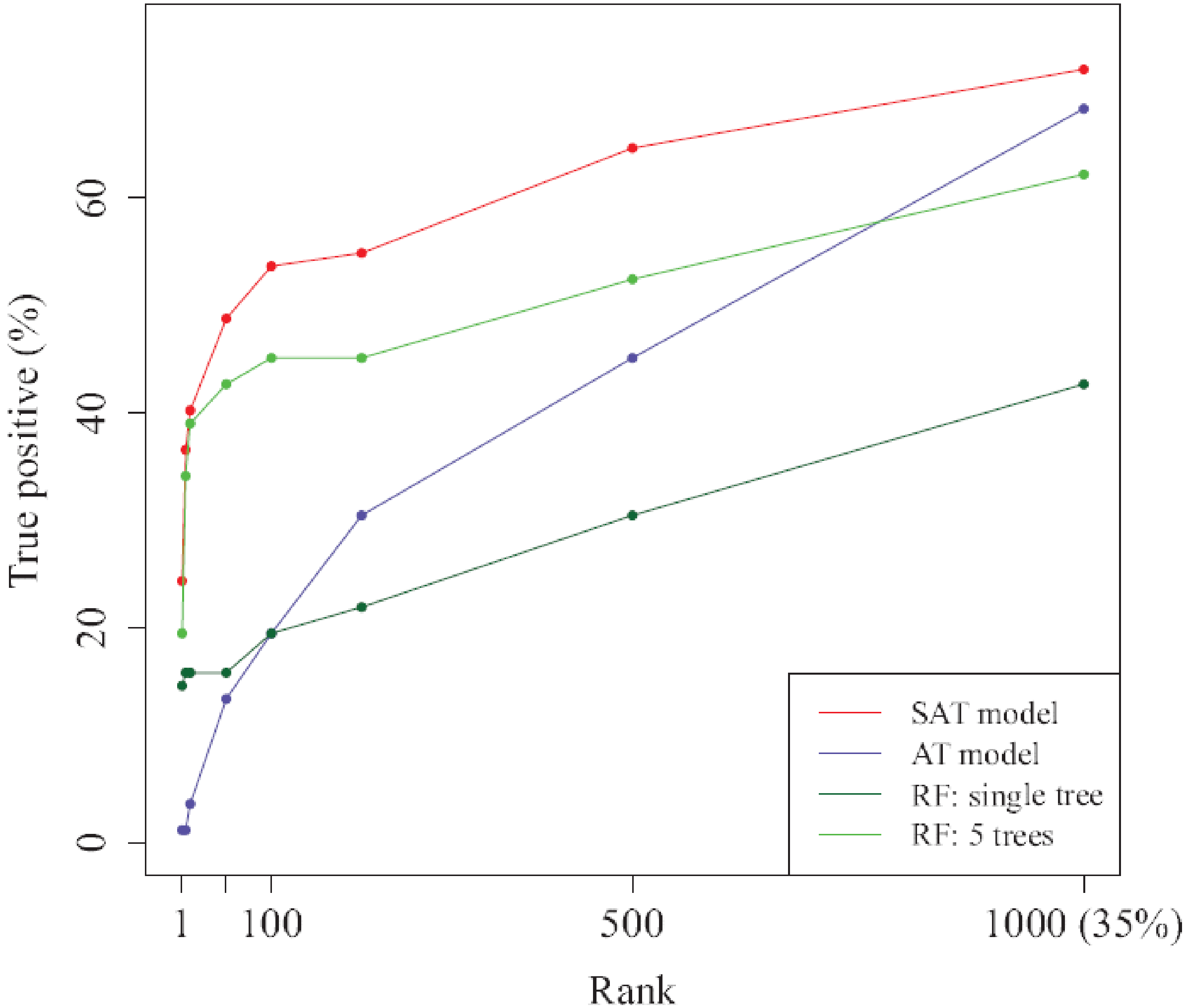
AT model



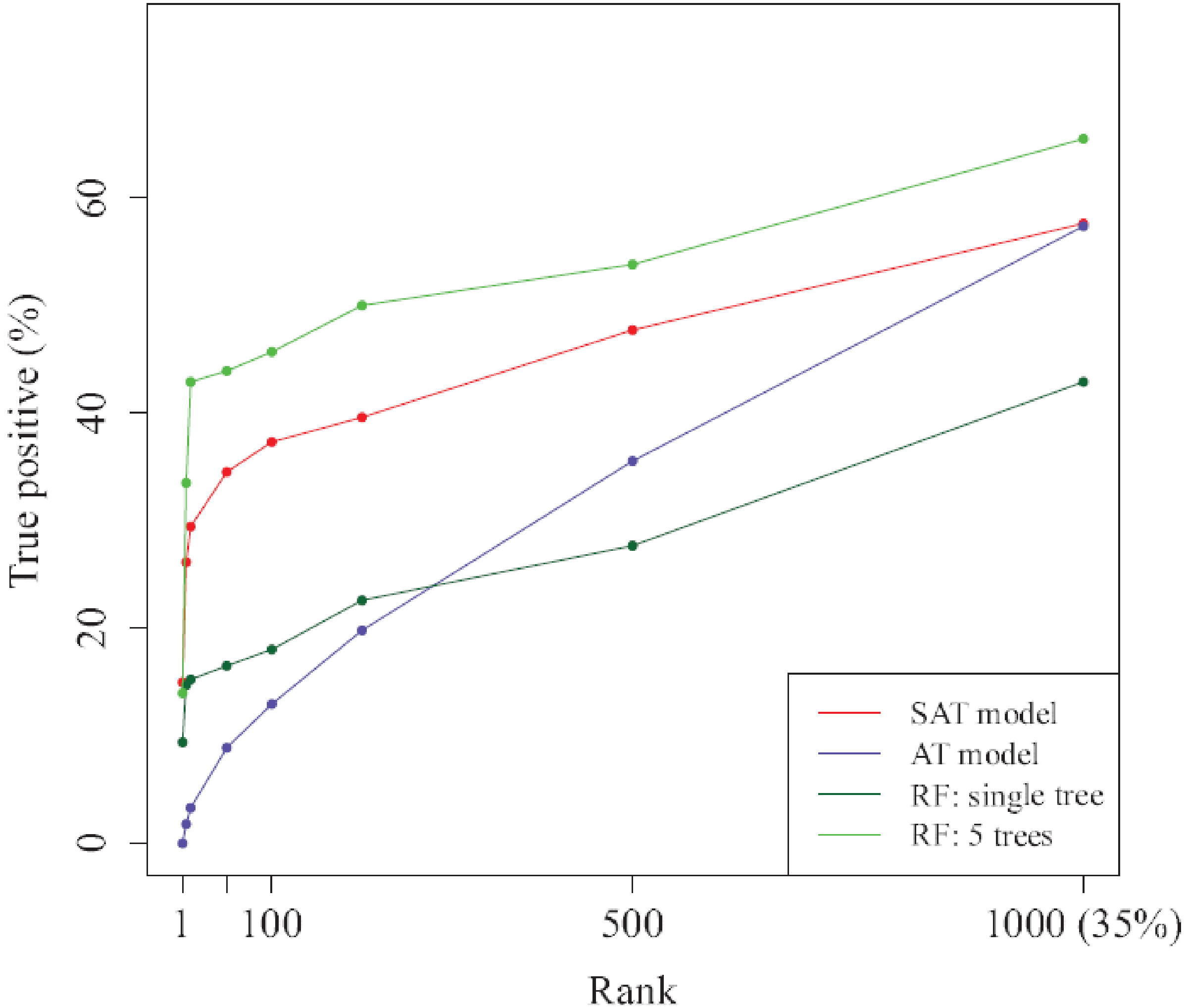
SAT model



TFs vs Ranks for Prediction on **Single-Author** Documents



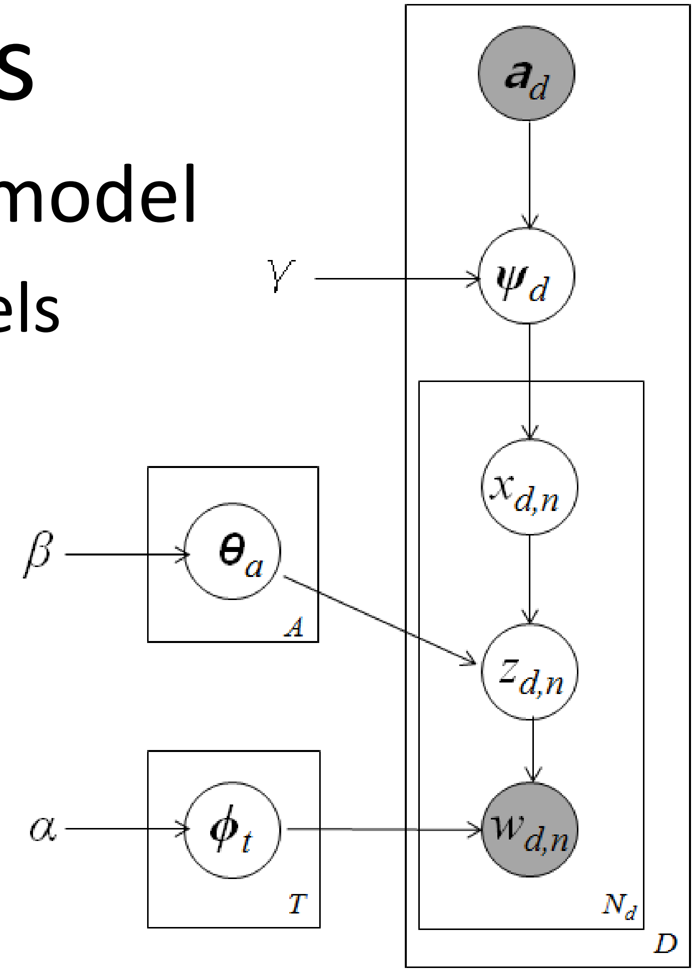
TFs vs Ranks for Prediction on **Multiple-Author** Documents



Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Conclusions

- Supervised Author-Topic (SAT) model
  - Probabilistic *latent* variable models
  - Information discovery
    - Topic-word distributions
    - Author (topic-based) interests
    - Author contributions
  - Authorship attribution



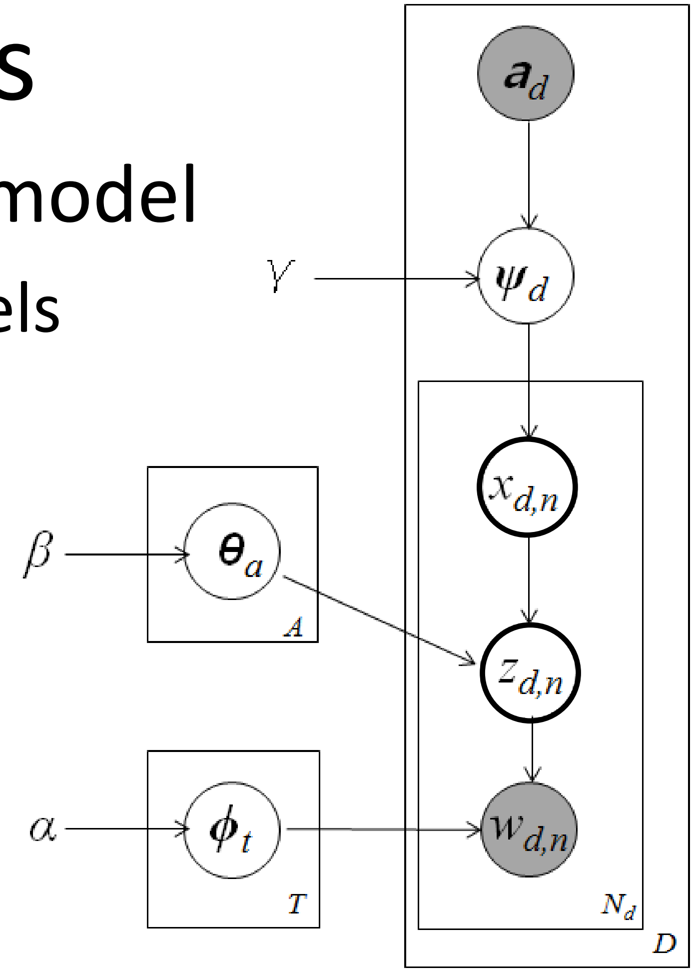
Abstract—By representing large corpora with concise and neaningful elements, topic-based generative models aim to reluce the dimension and understand the content of documents. Those techniques originally analyze on words in the documents, out their extensions currently accommodate meta-data such as uthorship information, which has been proved useful for textual nodeling. The importance of learning authorship is to extract uthor interests and assign authors to anonymous texts. Author-Topic (AT) model, an unsupervised learning technique, successfully exploits authorship information to model both documents und author interests using topic representations. However, the

is increasing, data in the hig and prediction authorship prediction have largely relied on discriminative modeling techniques that depend crucially on a variety of features such as word functions, word length distributions, and word contents [1]. The main drawback is that they generate a "black box" that makes it hard to understand why they give high performance on prediction.

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Conclusions

- Supervised Author-Topic (SAT) model
  - Probabilistic *latent* variable models
  - Information discovery
    - Topic-word distributions
    - Author (topic-based) interests
    - Author contributions
  - Authorship attribution



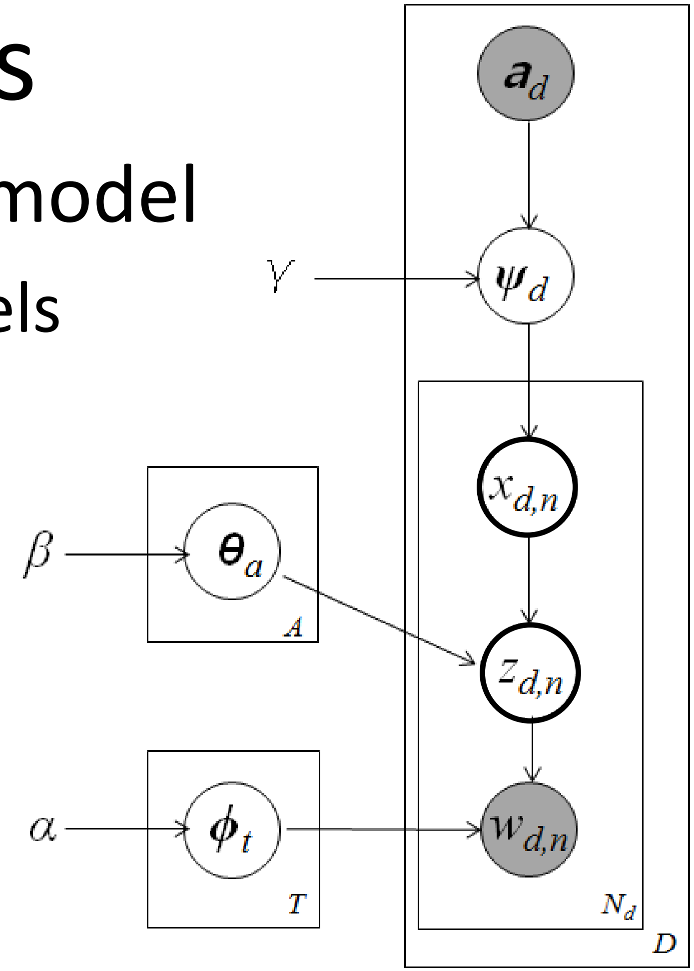
Abstract—By representing large corpora with concise and neaningful elements, topic-based generative models aim to reluce the dimension and understand the content of documents. Those techniques originally analyze on words in the documents, out their extensions currently accommodate meta-data such as uthorship information, which has been proved useful for textual nodeling. The importance of learning authorship is to extract uthor interests and assign authors to anonymous texts. Author-Topic (AT) model, an unsupervised learning technique, successfully exploits authorship information to model both documents and author interests using topic representations. However, the

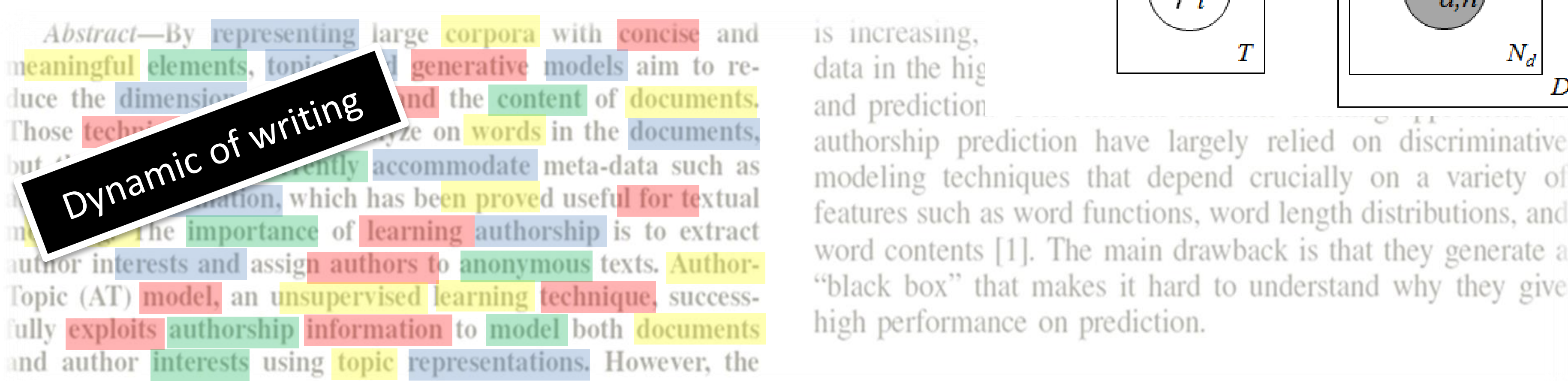
is increasing, data in the hig and prediction authorship prediction have largely relied on discriminative modeling techniques that depend crucially on a variety of features such as word functions, word length distributions, and word contents [1]. The main drawback is that they generate a "black box" that makes it hard to understand why they give high performance on prediction.

Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Conclusions

- Supervised Author-Topic (SAT) model
  - Probabilistic *latent* variable models
  - Information discovery
    - Topic-word distributions
    - Author (topic-based) interests
    - Author contributions
  - Authorship attribution

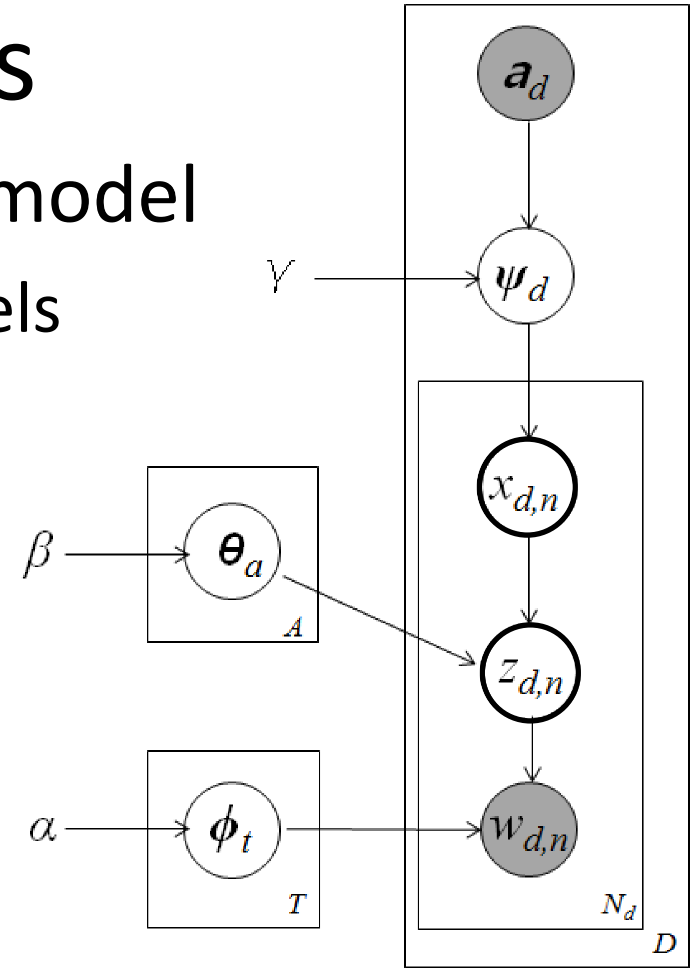


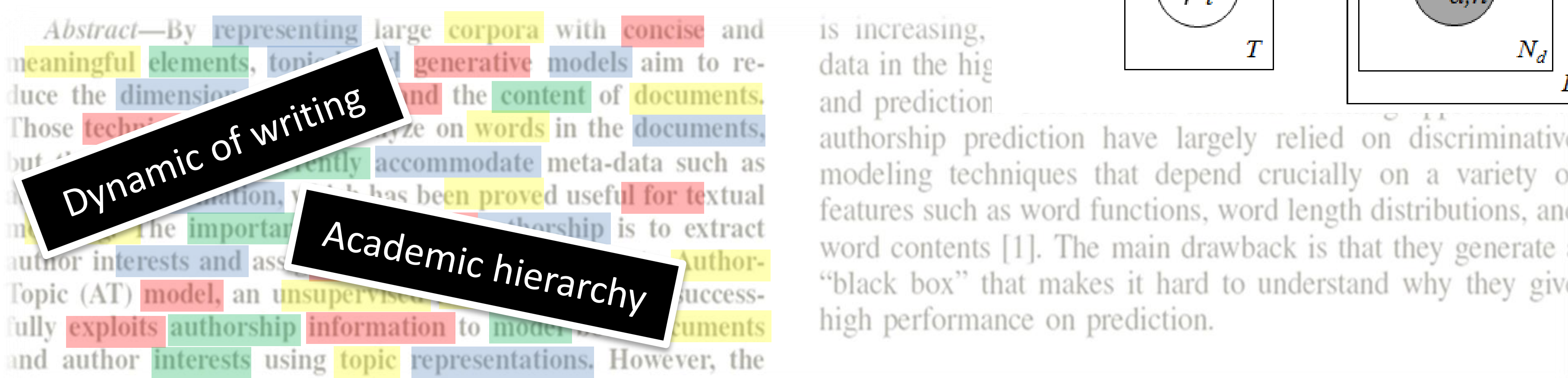


Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Conclusions

- Supervised Author-Topic (SAT) model
  - Probabilistic *latent* variable models
  - Information discovery
    - Topic-word distributions
    - Author (topic-based) interests
    - Author contributions
  - Authorship attribution

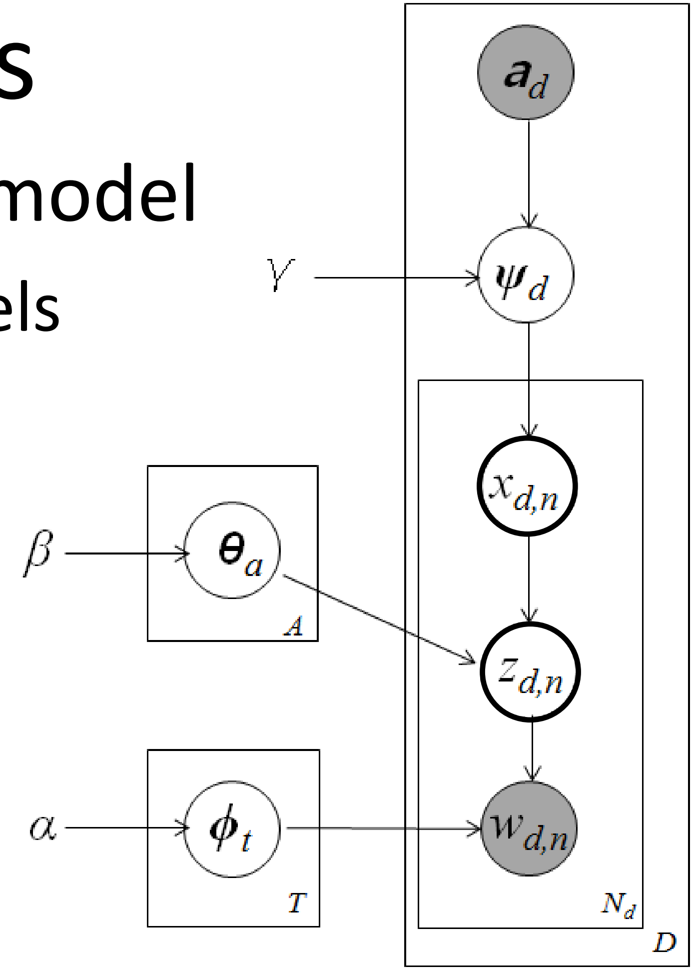


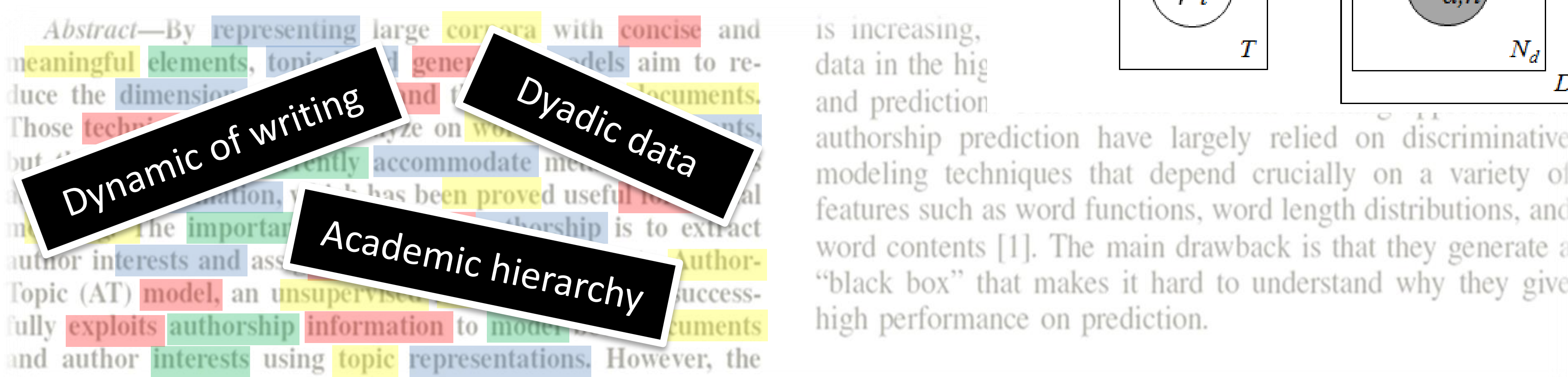


Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Conclusions

- Supervised Author-Topic (SAT) model
  - Probabilistic *latent* variable models
  - Information discovery
    - Topic-word distributions
    - Author (topic-based) interests
    - Author contributions
  - Authorship attribution

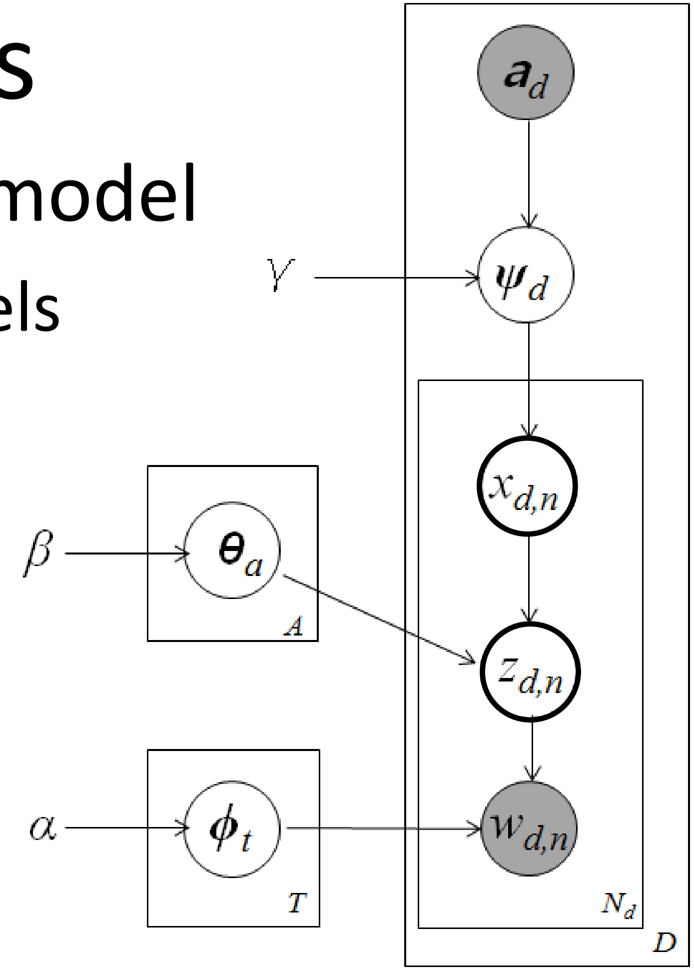


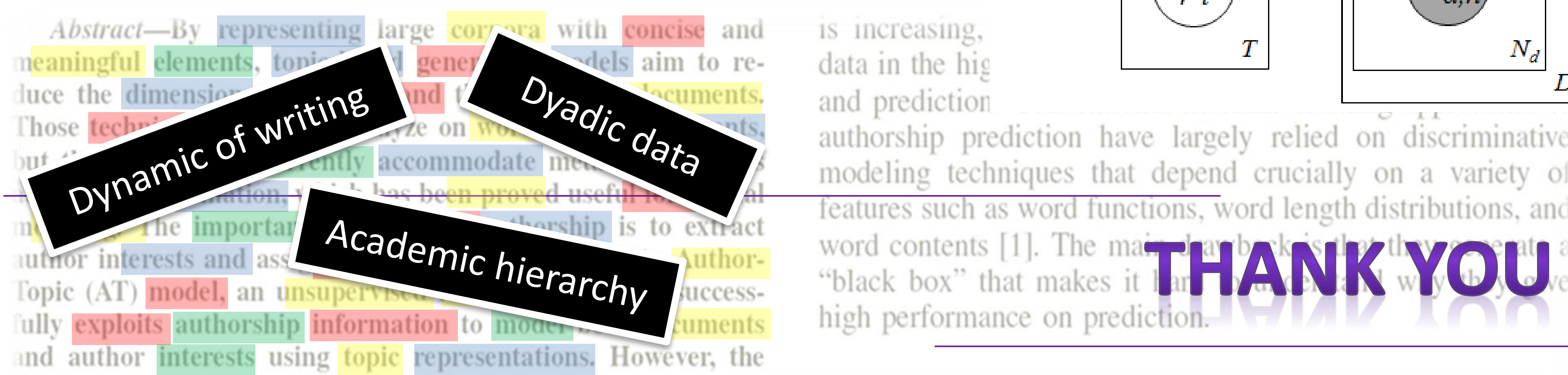


Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk

Conclusions

- Supervised Author-Topic (SAT) model
  - Probabilistic *latent* variable models
  - Information discovery
    - Topic-word distributions
    - Author (topic-based) interests
    - Author contributions
  - Authorship attribution





Naruemon Pratanwanich (Ploy) Email: np394@cam.ac.uk