Uncovering latent jet substructure

Barry M. Dillon
Jozef Stefan Institute, Ljubljana, Slovenia

Based on: **hep-ph/1904.04200**
BMD, Darius A. Faroughy, Jernej F. Kamenik
+ Manuel Szewc

BOOST2019, Boston, 23rd July 2019

Institut
"Jožef Stefan"
Ljubljana, Slovenija

Overview

Goal: - unsupervised new physics searches in jets
- train on mixed, unlabelled, imbalanced samples
- sensitivity to small S/B (~1%)
- extract descriptions of signal and background
- interpret 'what the machine has learned'
- no control/side-band regions

Approach: - inspired by Natural Language Processing (NLP)
- generative 'topic' model for jet formation
- extract topics from jet substructure
- use topics for classification
- … all based on Latent Dirichlet Allocation

Related work:
On the topic of jets - Metodiev, Thaler
Operational definition of quark and gluon jets - Komiske, Metodiev, Thaler
CWoLa - Metodiev, Nachman, Thaler
LLP - Dery, Nachman, Rubbo, Schwartzman
CWoLa bump hunt - Collins, Howe, Nachman
(binary)JUNIPR - Andreassen, Feige, Frye, Schwartz

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

LDA is used for text classification, it is based on a generative model for text documents.

Vocabulary:
list of all words in all documents

Topics:
probability distribution over this vocabulary

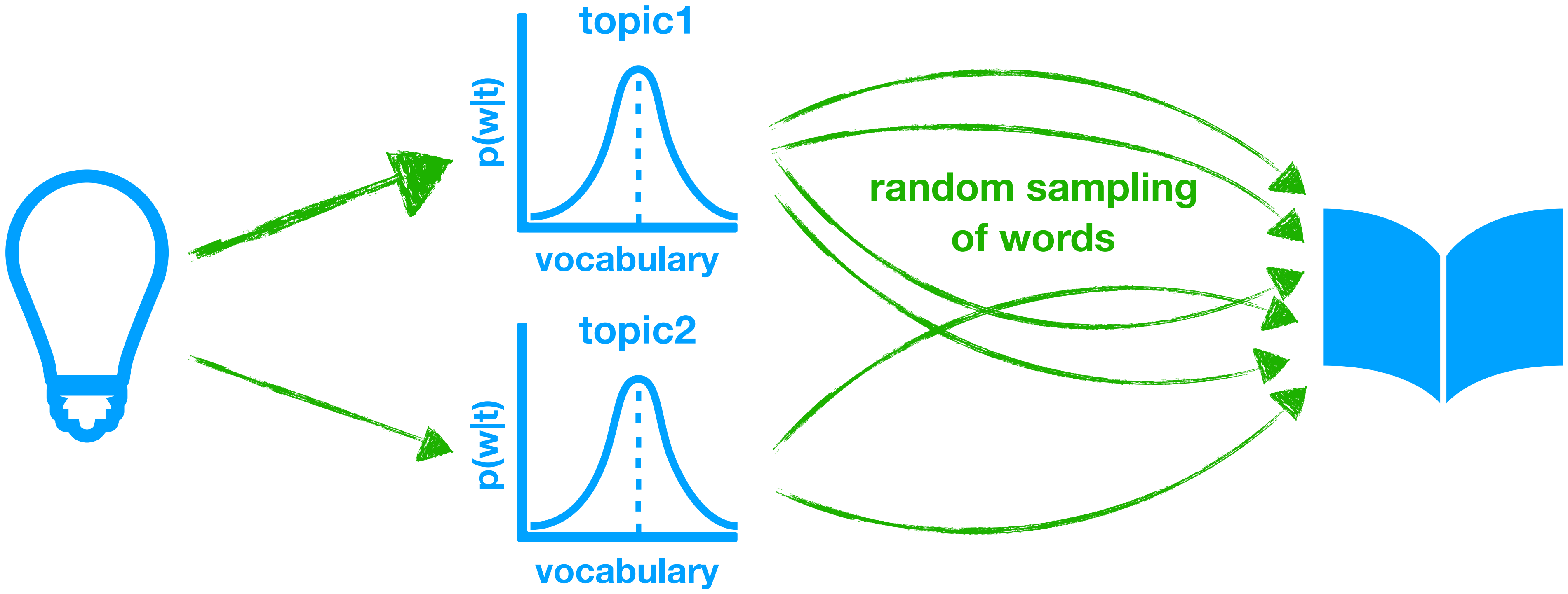
Document:
an un-ordered list of words, sampled from the topics

Latent structure:
The topic distributions, and the prior on topic proportions

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

LDA is used for text classification, it is based on a generative model for text documents.

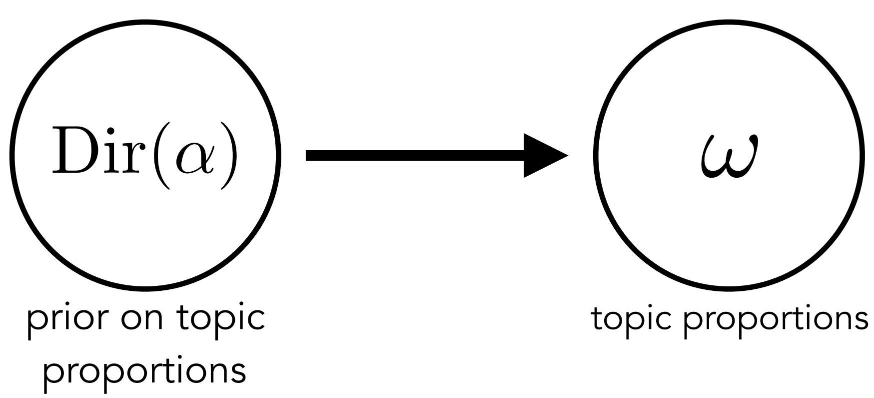


A document is a mixture of topics - no hard assignment
Bag-of-words approximation - no semantic structure is modelled

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

A generative model for a document with K topics

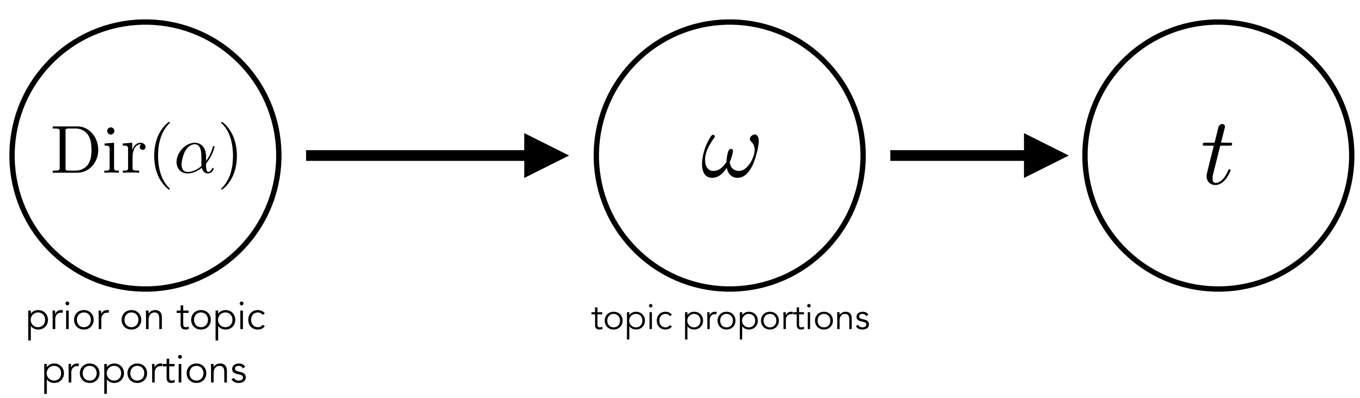


Step 1: sample proportions for each topic, a K-dimensional multinomial ω

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

A generative model for a document with K topics

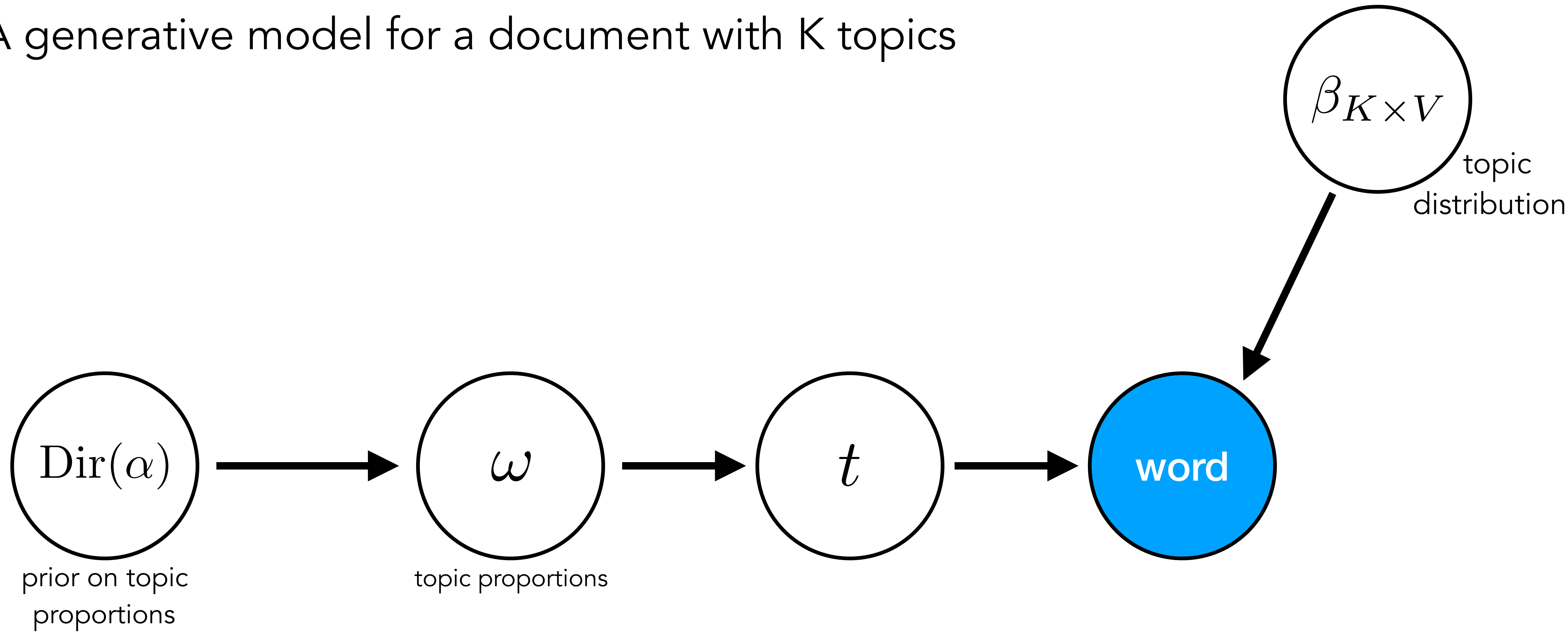


Step 2: sample a single topic from the multinomial

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

A generative model for a document with K topics

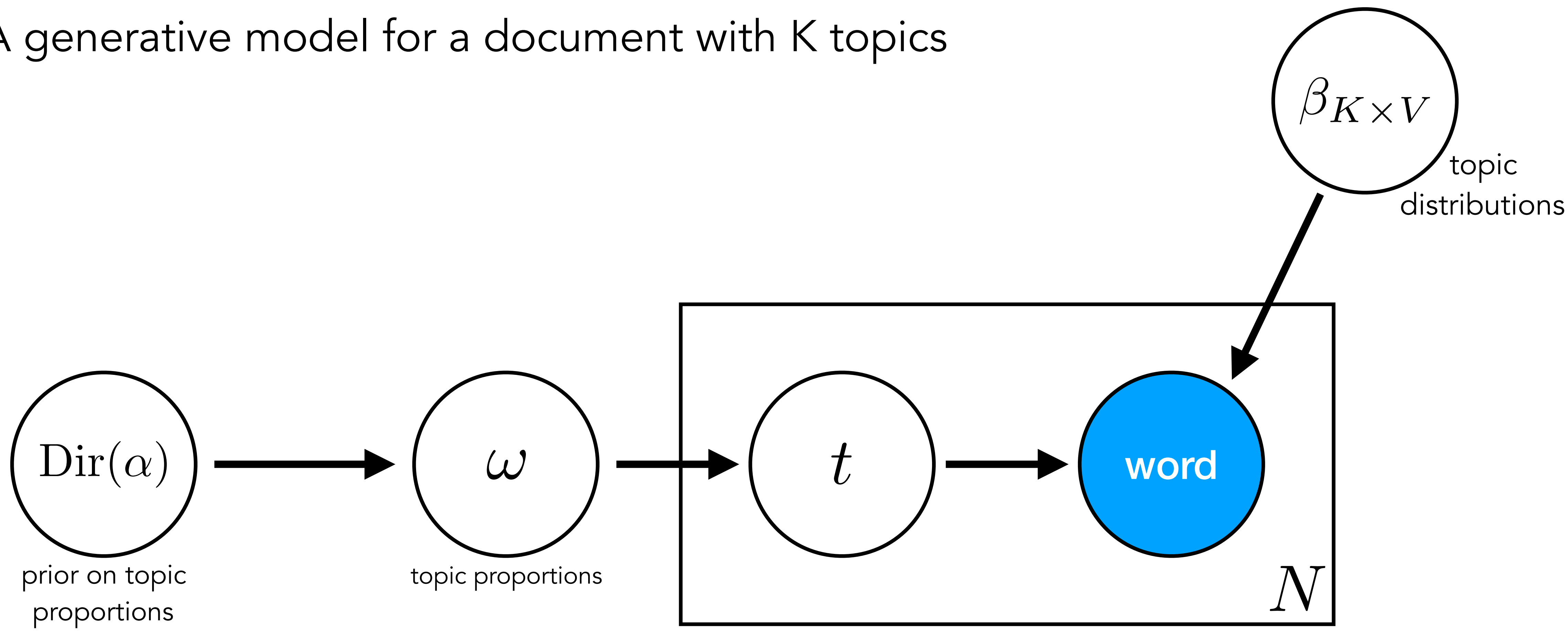


Step 3: sample a word from the appropriate topic distribution

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

A generative model for a document with K topics

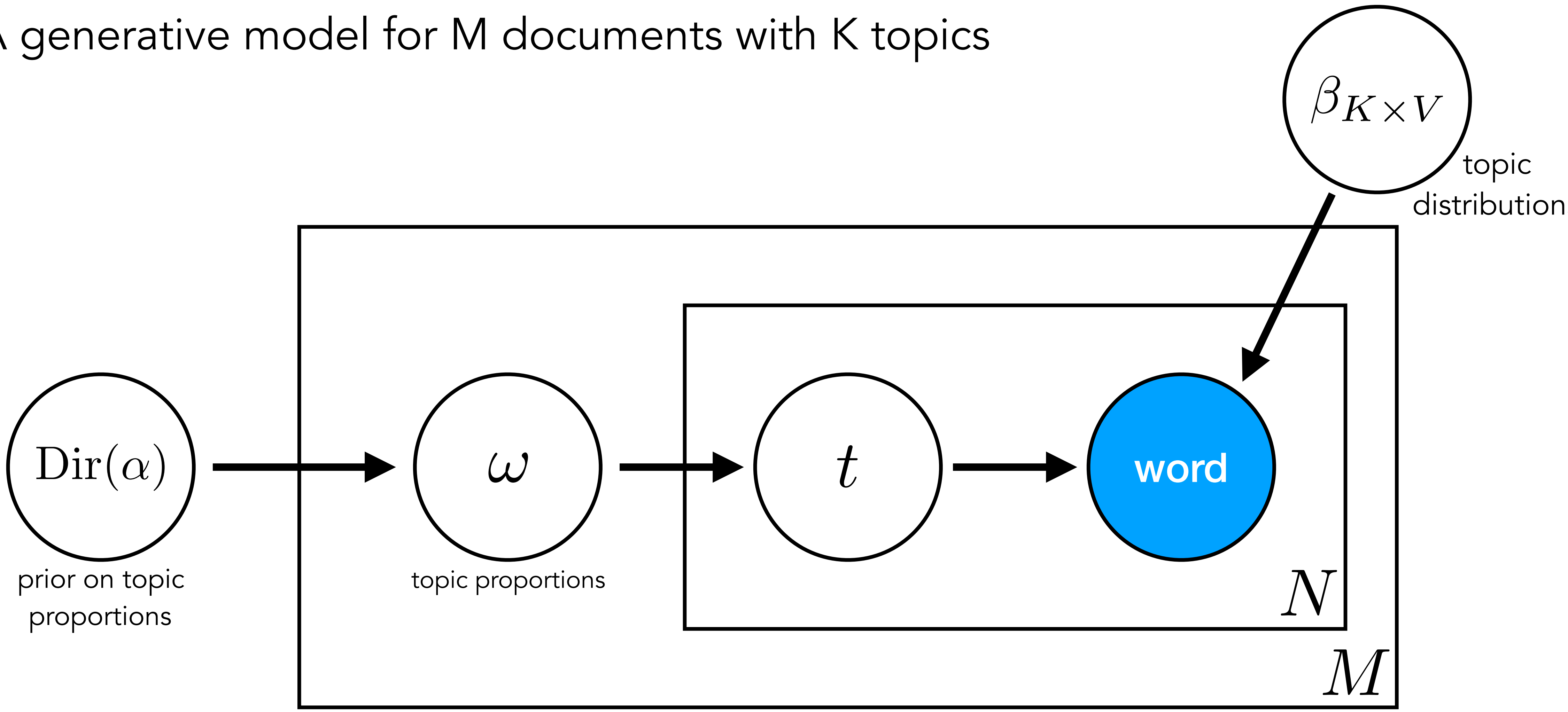


- repeat this for each of the N words in the document

Latent Dirichlet Allocation

Blei, Ng, Jordan
Journal of Machine Learning Research
(2003)

A generative model for M documents with K topics



- repeat again for each of the M documents you want to generate

Latent Dirichlet Allocation

The jet physics edition

LDA is used for text classification, it is based on a generative model for text documents.

~~**Vocabulary:**~~ **substructure features**
list of all words in all documents

~~**Topics:**~~ **underlaying physics**
probability distribution over this vocabulary

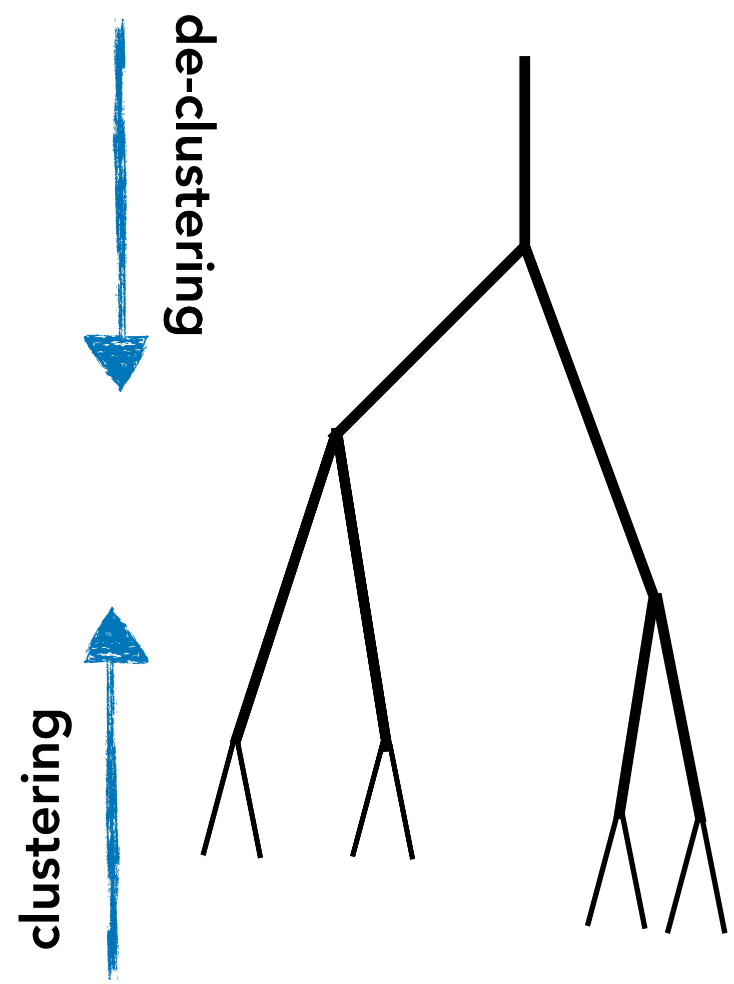
~~**Document:**~~ **a jet or an event**
an un-ordered list of words, sampled from the topics

~~**Latent structure:**~~ **physics underlaying the jets**
The topic distributions, and the prior on topic proportions

Substructure features

We use clustering history of a jet as a proxy for how its substructure was formed

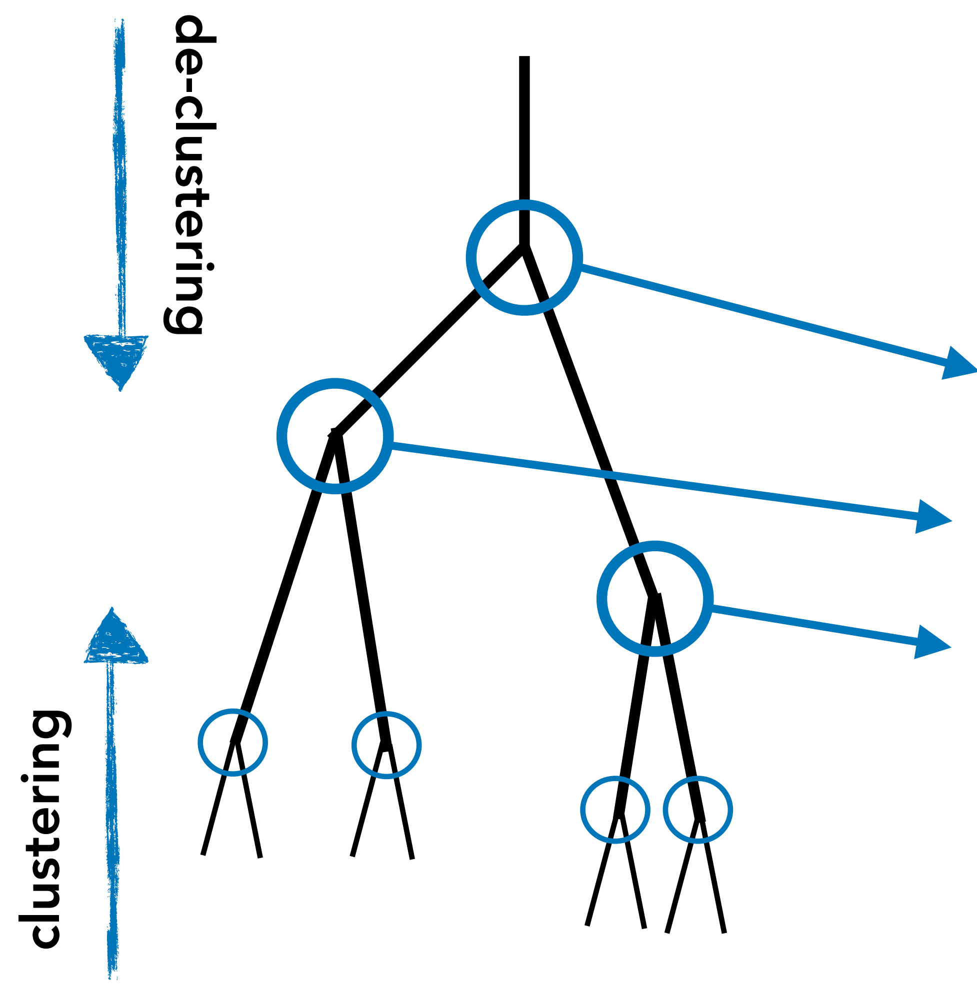
The ‘features’ should encapsulate this tree-structure



Substructure features

We use clustering history of a jet as a proxy for how its substructure was formed

The ‘features’ should encapsulate this tree-structure



$$\text{feature1} = \left[m_{\text{child}}, \frac{m_{\text{parent1}}}{m_{\text{child}}}, \frac{m_{\text{parent2}}}{m_{\text{parent1}}}, \Delta R_{\text{parents}} \right]$$

$$\text{feature2}$$

$$\ldots$$

$$\boxed{\text{jet} = \{\text{feature1}, \text{feature2}, \text{feature3}, \ldots\}}$$

Extracting the topics

The goal is, given a sample of data, to extract the underlaying topics.

The generative model gives us

$$p(\text{jet}|\alpha, \beta) = \int_\omega p(\omega|\alpha) \prod_{f \in \text{jet}} \left(\sum_t p(t|\omega) p(f|t, \beta) \right)$$

Given a sample of jets, we want to find the set of topics that maximise the log likelihood that the jets were sampled

$$\beta^{\text{MLE}}_{K \times V} = \underset{\beta}{\text{argmax}} \;\; \log \left(\prod_{i=1}^{M} P(\text{jet}_i|\alpha, \beta) \right)$$

- we used the gensim software (it implements a variational EM algorithm)

Radim Rehurek, Petr Sojka, 2010

Topics for new physics

W' example: $pp \to W' \to \phi W \to WWW$ $\quad S/B = 0.011$

J. H. Collins, K. Howe, B. Nachman (2019)

$m_{W'} = 3\ \text{TeV},\ m_\phi = 400\ \text{GeV}$ $\quad m_{jj} \in 2730 - 3190\ \text{GeV}$

Training:

- ~80,000 events
- 2 topics
- mixed, unlabelled sample of QCD and W' events, S/B=0.011
- Dirichlet parameters: $\alpha_1 = 0.99, \quad \alpha_2 = 0.01$

What we want:

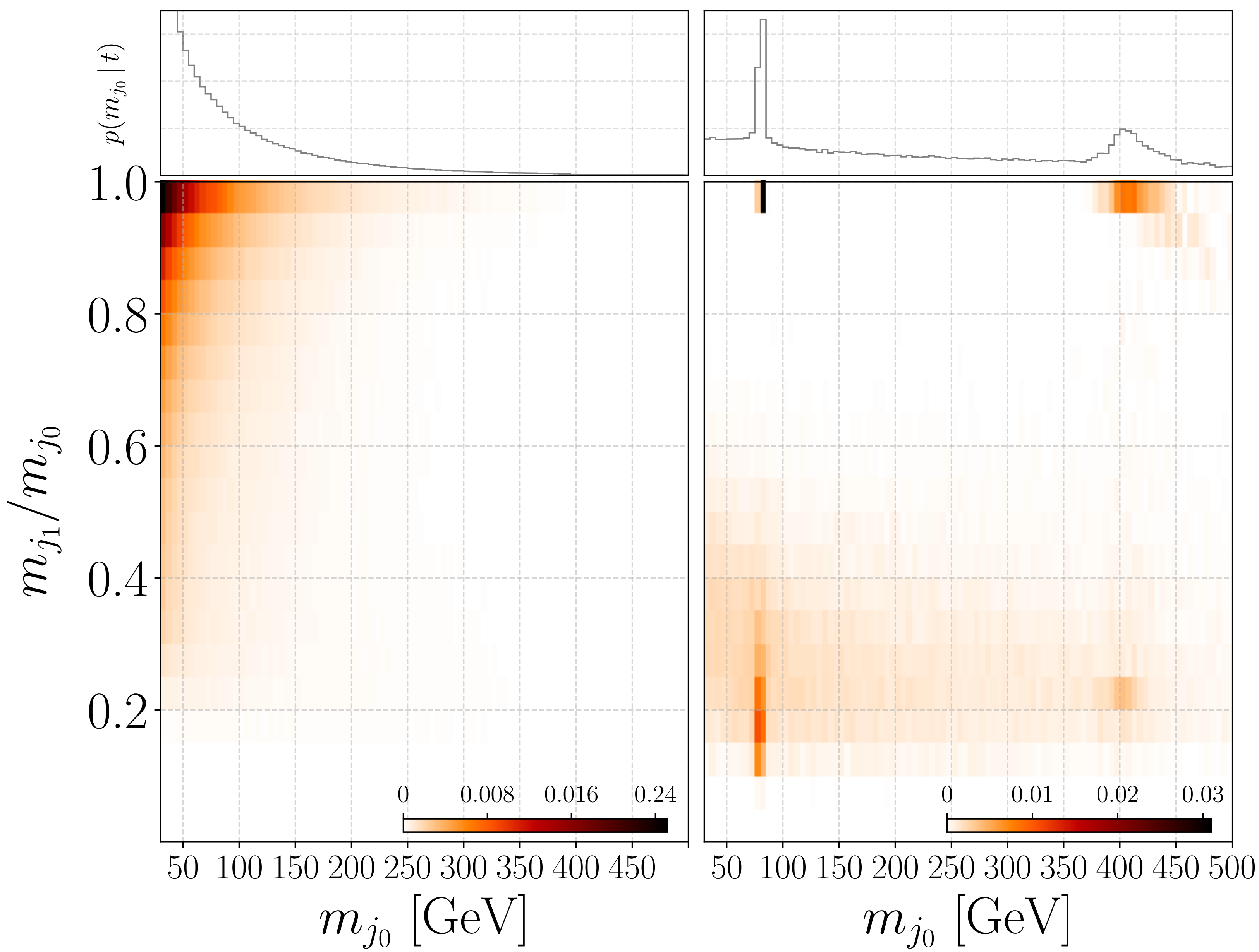
- extraction of 2 topics related to signal and background
- use topics for classification/clustering

Topics for new physics

W′ example: $pp \to W' \to \phi W \to WWW$ $S/B = 0.011$

J. H. Collins, K. Howe, B. Nachman (2019)

$m_{W'} = 3\ \text{TeV},\ m_\phi = 400\ \text{GeV}$ $m_{jj} \in 2730 - 3190\ \text{GeV}$

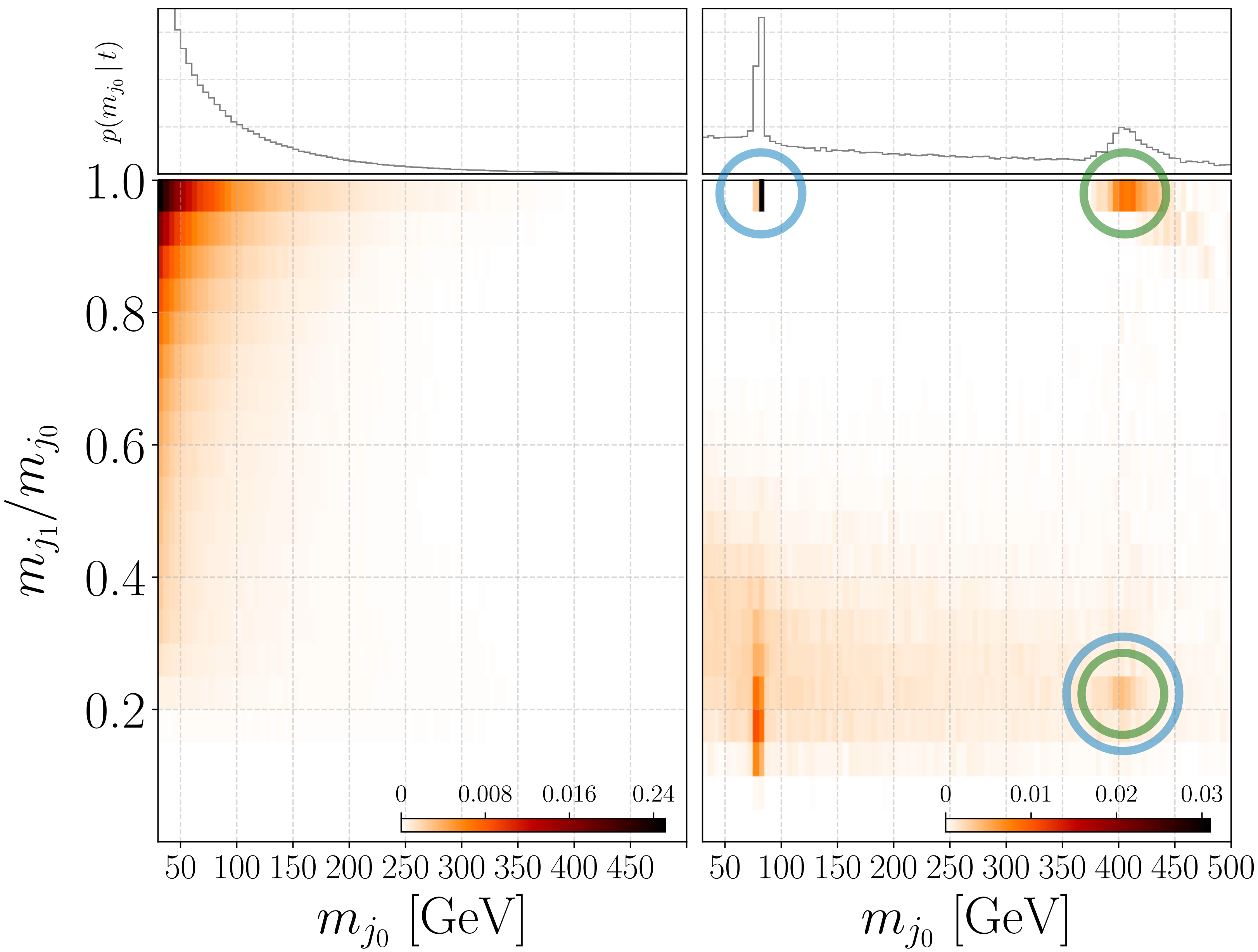


Topics for new physics

W' example: $pp \to W' \to \phi W \to WWW$ $\quad S/B = 0.011$

J. H. Collins, K. Howe, B. Nachman (2019)

$m_{W'} = 3\ \text{TeV},\ m_\phi = 400\ \text{GeV} \quad m_{jj} \in 2730 - 3190\ \text{GeV}$



Classification of new physics

W′ example: $pp \to W' \to \phi W \to WWW$ $S/B = 0.011$

J. H. Collins, K. Howe, B. Nachman (2019)

$m_{W'} = 3\ \text{TeV},\ m_\phi = 400\ \text{GeV}$ $m_{jj} \in 2730 - 3190\ \text{GeV}$

Two methods for classification:

1 - classify using topic proportions inferred in each event

maximise the likelihood for each event to extract ω_1

2 - classify using a likelihood ratio built from the extracted topics

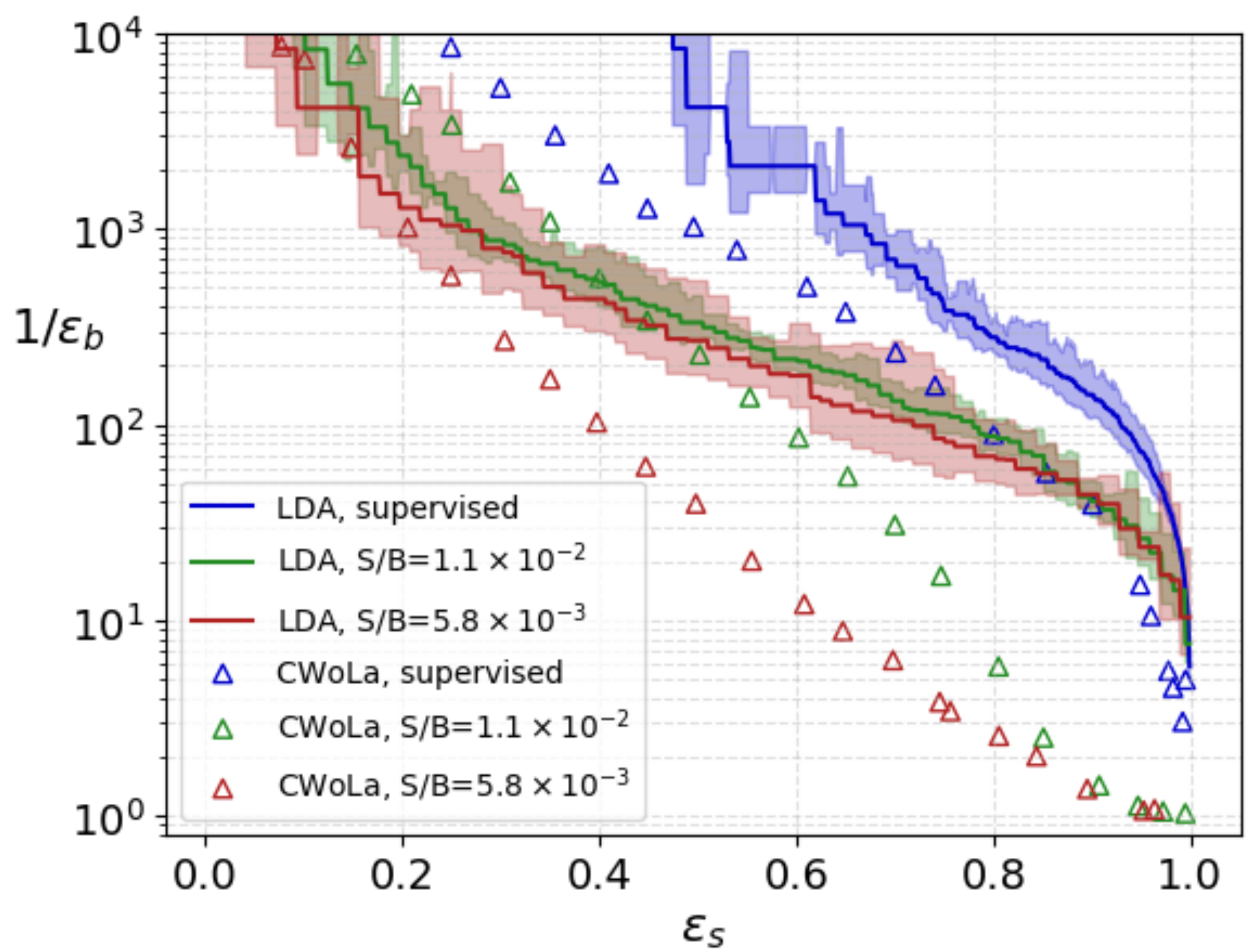
$$\frac{p(\text{event}|\phi_2)}{p(\text{event}|\phi_1)} \leq \text{threshold}$$

Classification of new physics

W′ example: $pp \to W' \to \phi W \to WWW$ $S/B = 0.011$

J. H. Collins, K. Howe, B. Nachman (2019)

$m_{W'} = 3\ \mathrm{TeV},\ m_\phi = 400\ \mathrm{GeV}$ $m_{jj} \in 2730 - 3190\ \mathrm{GeV}$



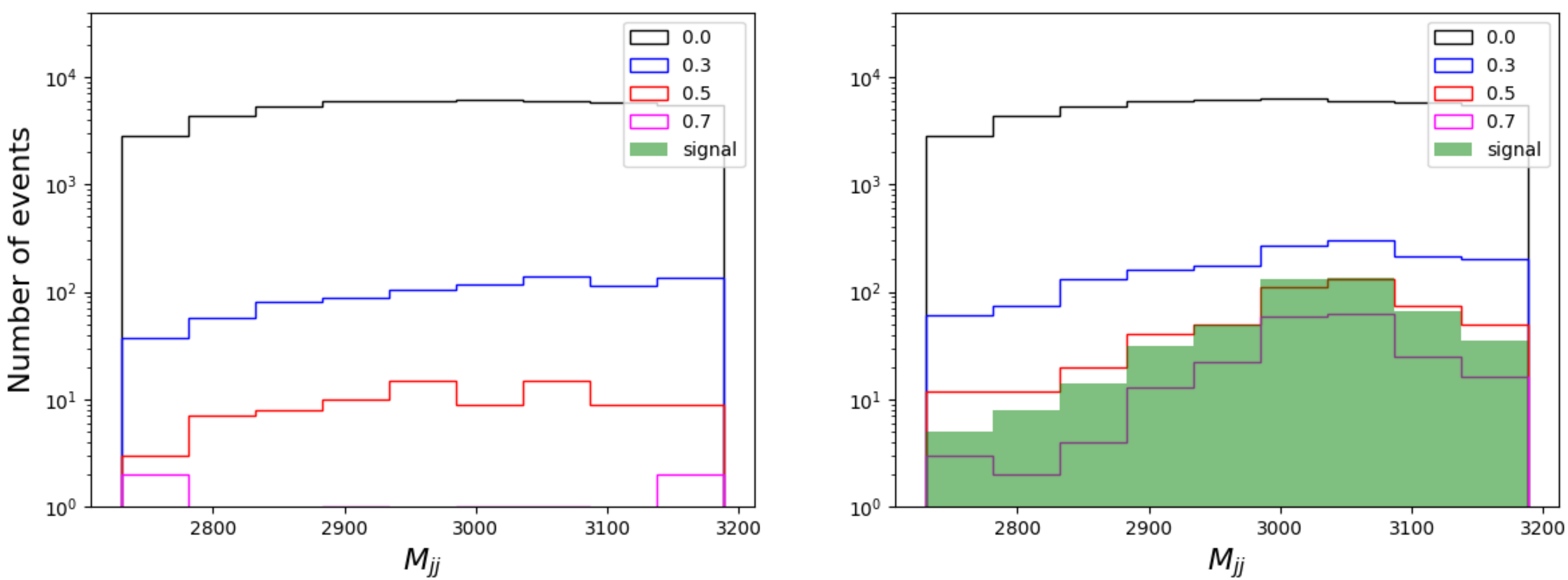
errors estimated using k-folding, with k=10

Uncovering the new physics

W′ example: $pp \to W' \to \phi W \to WWW$ $S/B = 0.011$

J. H. Collins, K. Howe, B. Nachman (2019)

$m_{W'} = 3\ \text{TeV},\ m_\phi = 400\ \text{GeV}$ $m_{jj} \in 2730 - 3190\ \text{GeV}$



LDA with no signal vs LDA with signal

Summary

Latent Dirichlet Allocation allows us to attack several interesting problems:

- ✓ train on mixed, unlabelled, imbalanced samples
- ✓ sensitivity to small S/B (~1%)
- ✓ extract descriptions of signal and background
- ✓ interpret 'what the machine has learned'
- ✓ no control/side-band regions

So far, this is a proof-of-concept.

Next steps:

- build a specialised topic model for jets/events
- explore classifiers with more than two topics
- hierarchical topic models: number of topics is learned from data
- can this be optimised with deep learning?
- better understand the relationship between the topics and the physics
- more applications: SM, more BSM, pile-up mitigation, …

additional slides

The Dirichlet prior

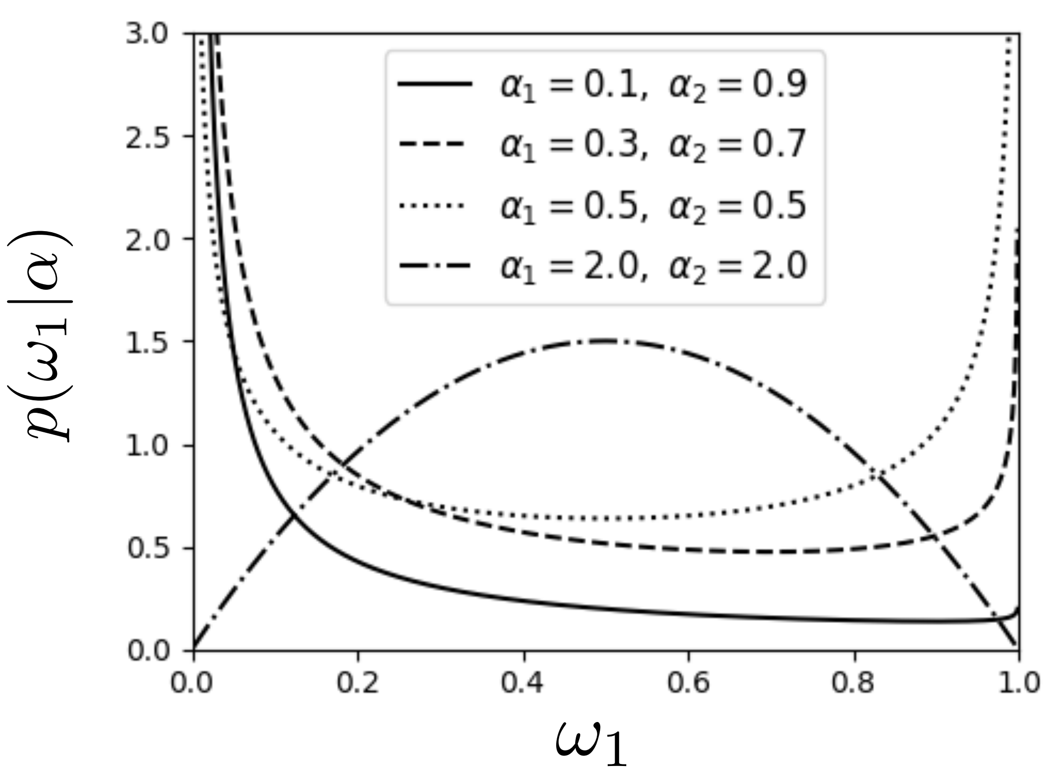
This determines the prevalence of the topics both in the whole sample, and per jet/event

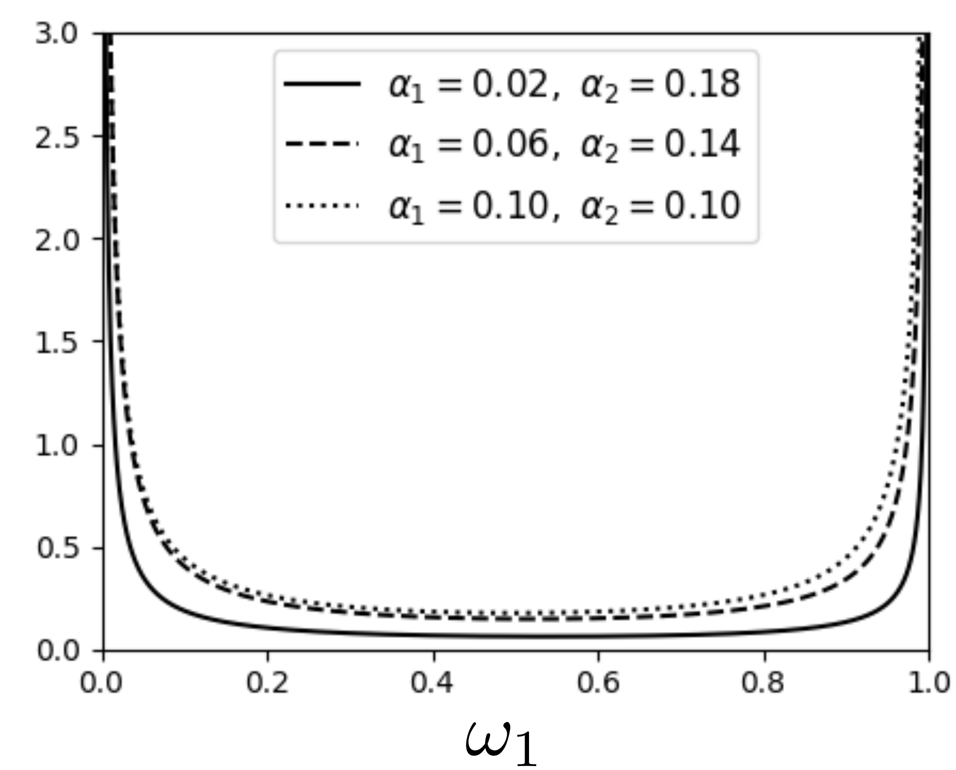
- crucial for sensitivity to small S/B

Basic scenario: **two topics** - one signal, one background

Dirichlet distribution → beta distribution

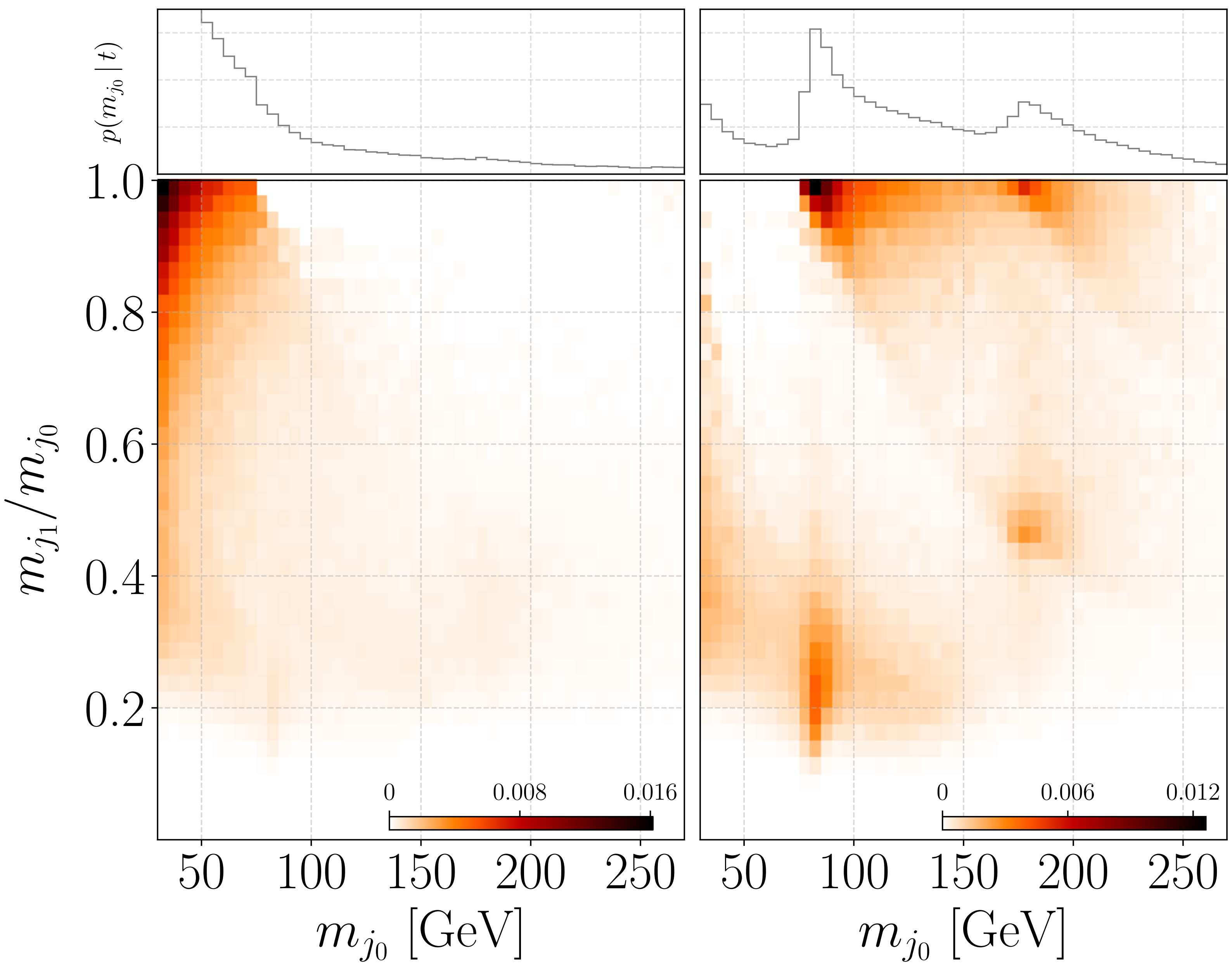
$$\mathrm{B}(\omega_1;\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}\omega_1^{\alpha_1-1}(1-\omega_1)^{\alpha_2-1} \qquad \frac{\alpha_2}{\alpha_1} = \frac{\omega_2}{\omega_1}\bigg|_{\mathrm{sample}}$$





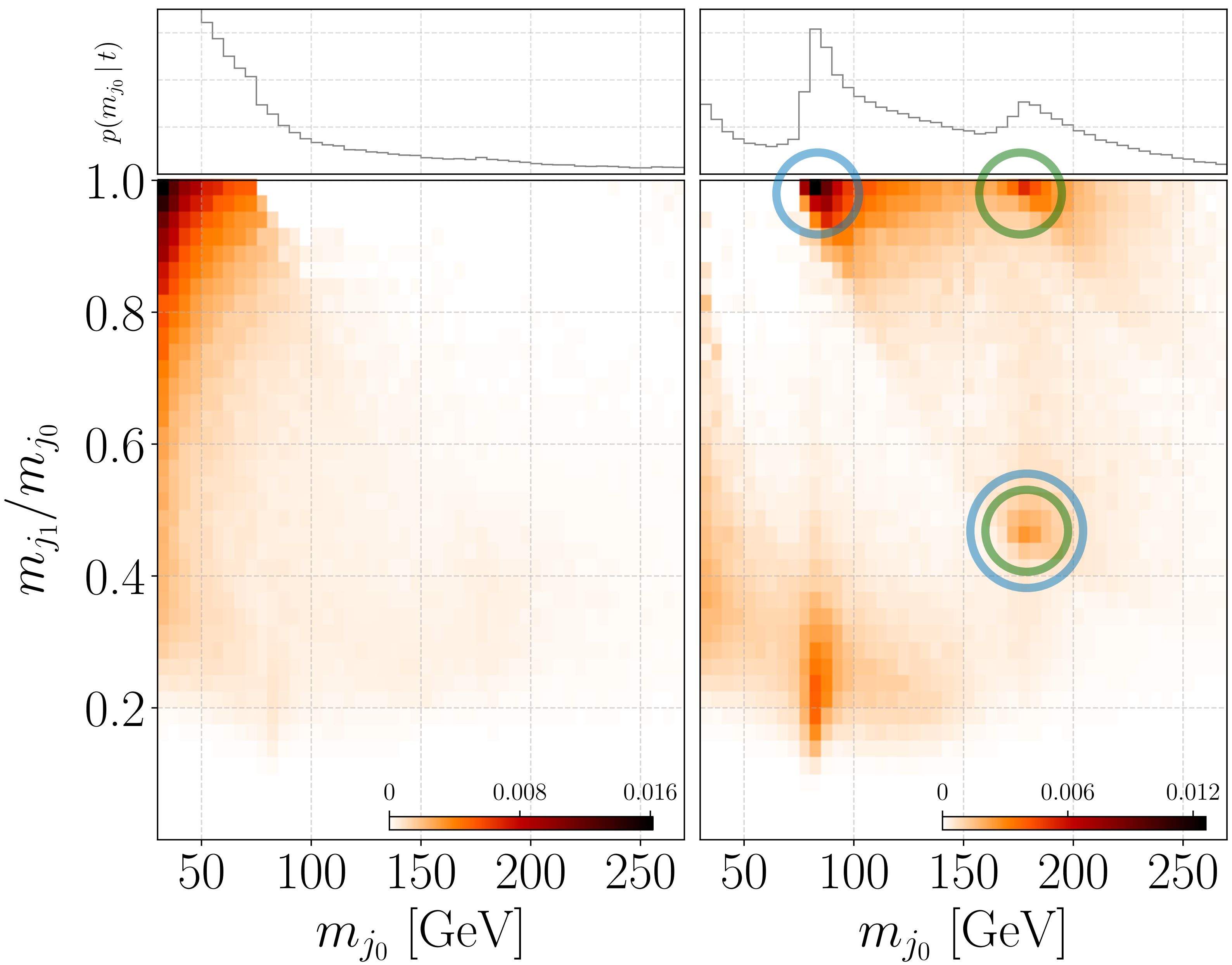
Topics for tops

$$pp \to t\bar{t} \to W^+W^-b\bar{b}, \quad S/B = 1, \quad \alpha_1 = \alpha_2 = 0.5$$



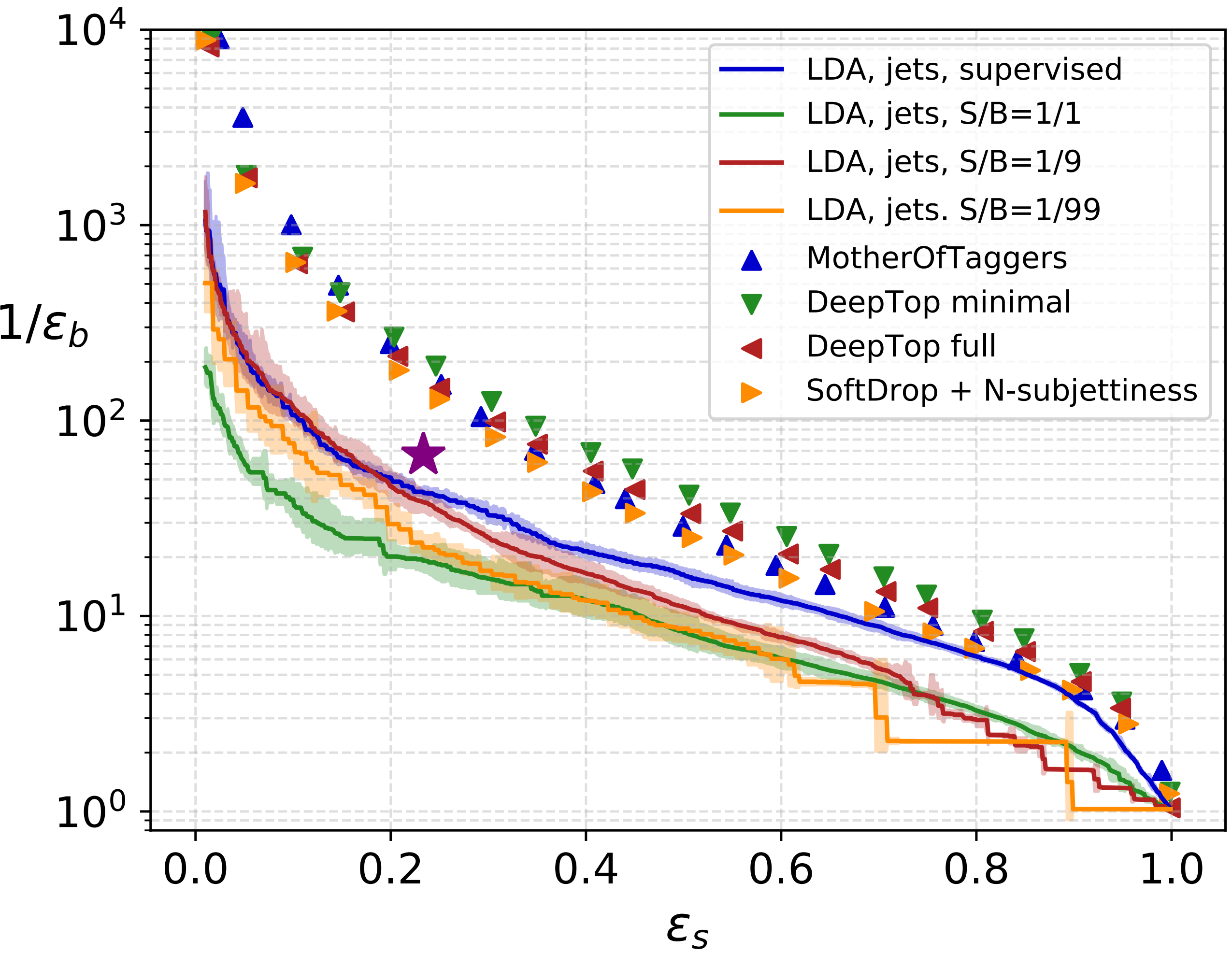
Topics for tops

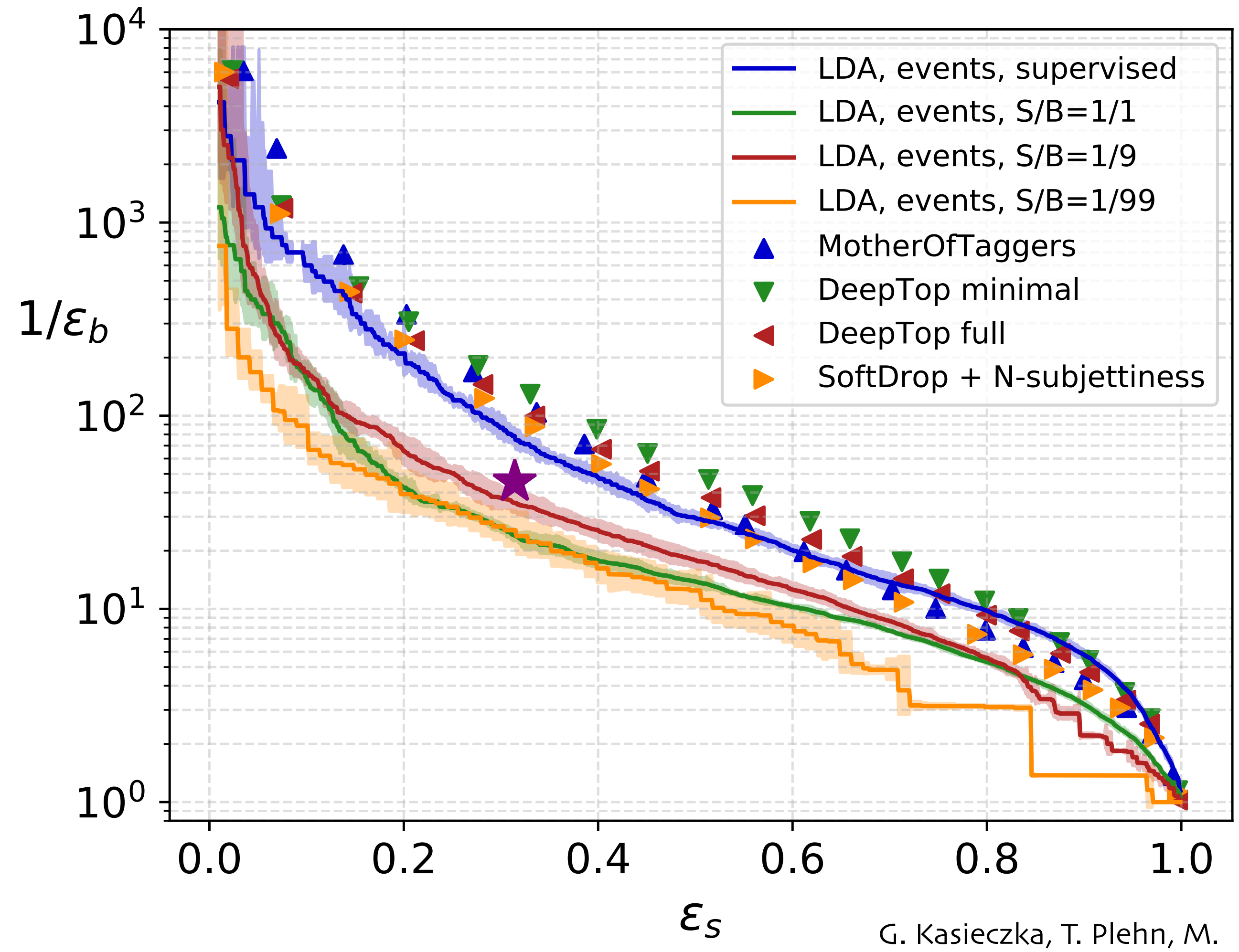
$$pp \to t\bar{t} \to W^+W^- b\bar{b}, \quad S/B = 1, \quad \alpha_1 = \alpha_2 = 0.5$$



Classification of tops

$$pp \to t\bar{t} \to W^+W^-b\bar{b}, \quad S/B = 1, \quad \alpha_1 = \alpha_2 = 0.5$$





G. Kasieczka, T. Plehn, M. Russell, T. Schell (2017)

errors estimated using k-folding, with k=10