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

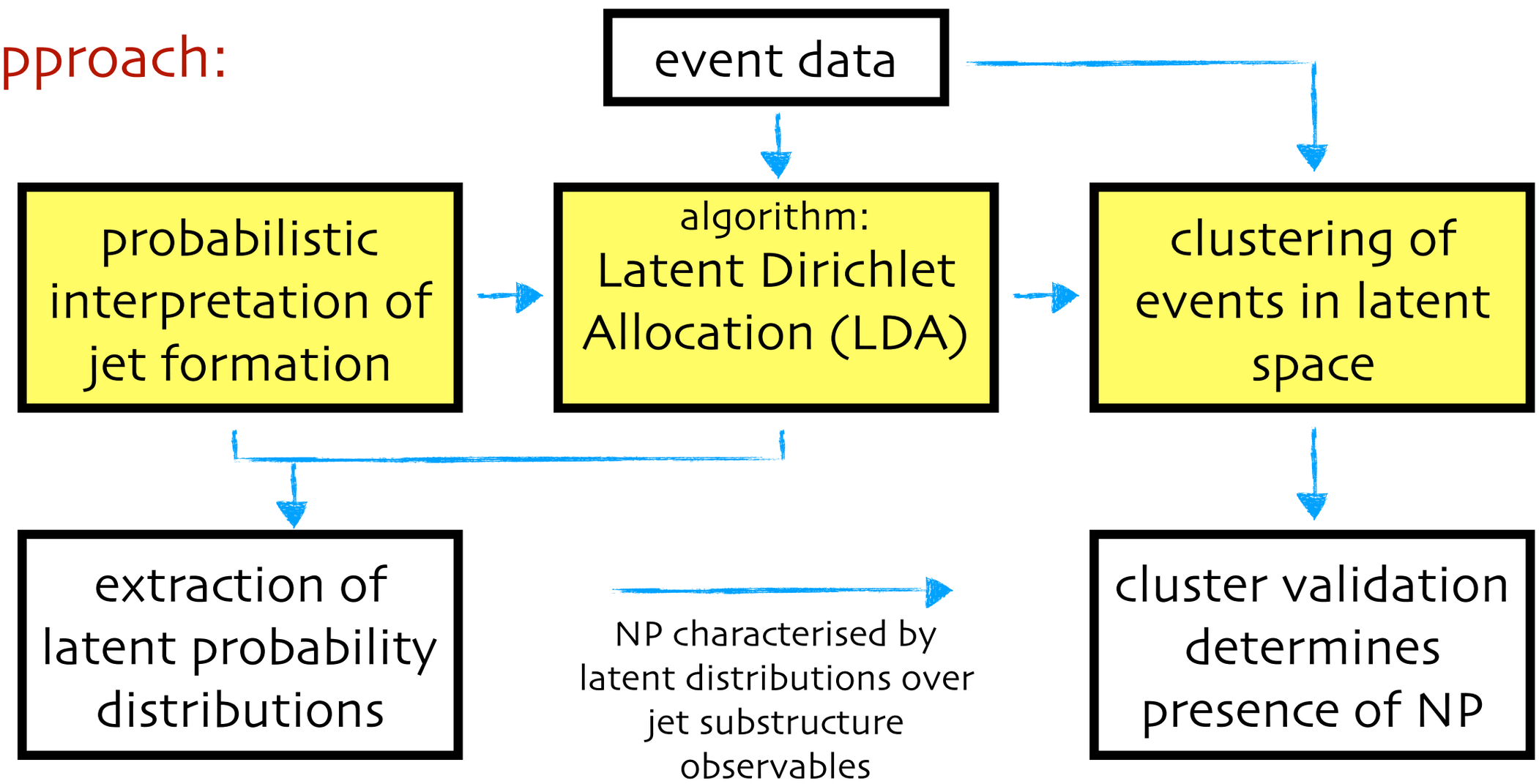
Planck2019, Granada, 4th June

Institut
"Jožef Stefan"
Ljubljana, Slovenija

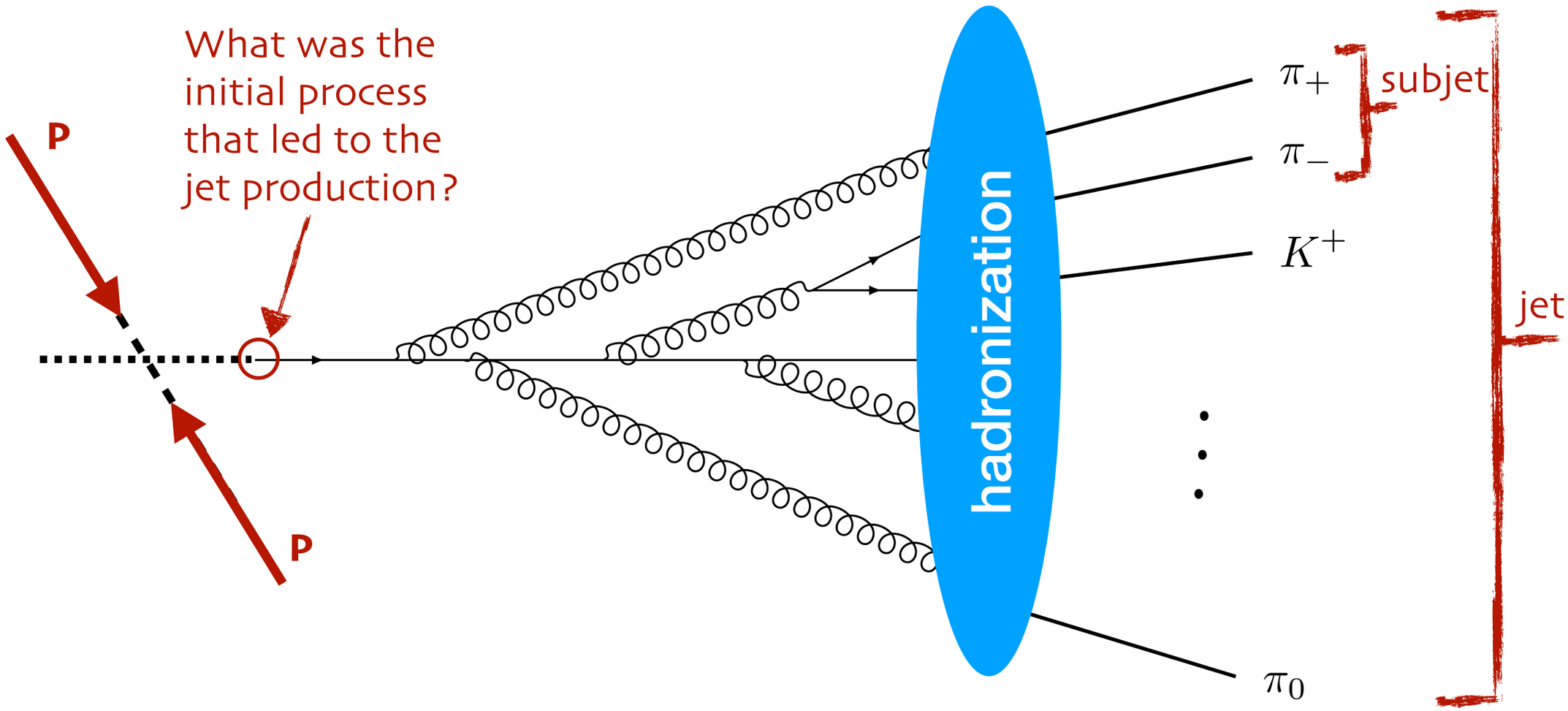
Overview

Goal: completely unsupervised new physics searches in jets
+ characterisation of new physics

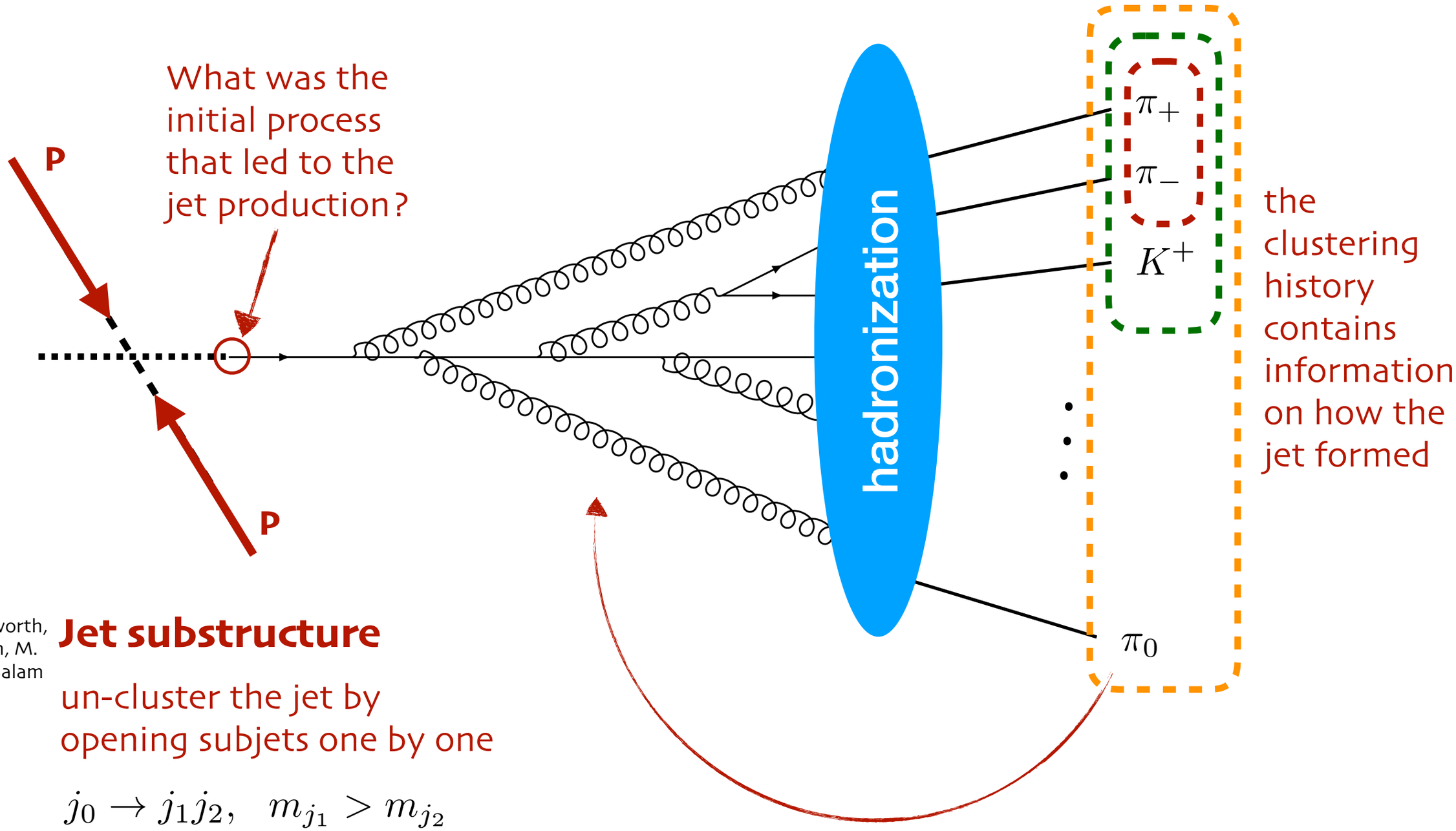
Approach:



Jets and substructure



Jets and substructure



J. M. Butterworth, A. R. Davison, M. Rubin, G. P. Salam (2008)

Jet substructure

un-cluster the jet by opening subjets one by one

$$j_0 \to j_1 j_2, \quad m_{j_1} > m_{j_2}$$

Substructure observables

Un-clustering a jet gives you a set of observables at each splitting:

$$o_{j_0} = \left\{ m_{j_0} \,,\ \frac{m_{j_1}}{m_{j_0}} \,,\ \frac{m_{j_2}}{m_{j_1}} \,,\ \frac{\min(p_{T,1}^2, p_{T,2}^2)}{m_{j_0}^2} \Delta R_{1,2}^2 \right\}$$

binning → "feature"

subjet mass

mass drop

probabilistic interpretation of jet formation:

- an event sample is a distribution over the features
- different underlaying processes are represented by different distributions
- jets are formed by sampling from these distributions

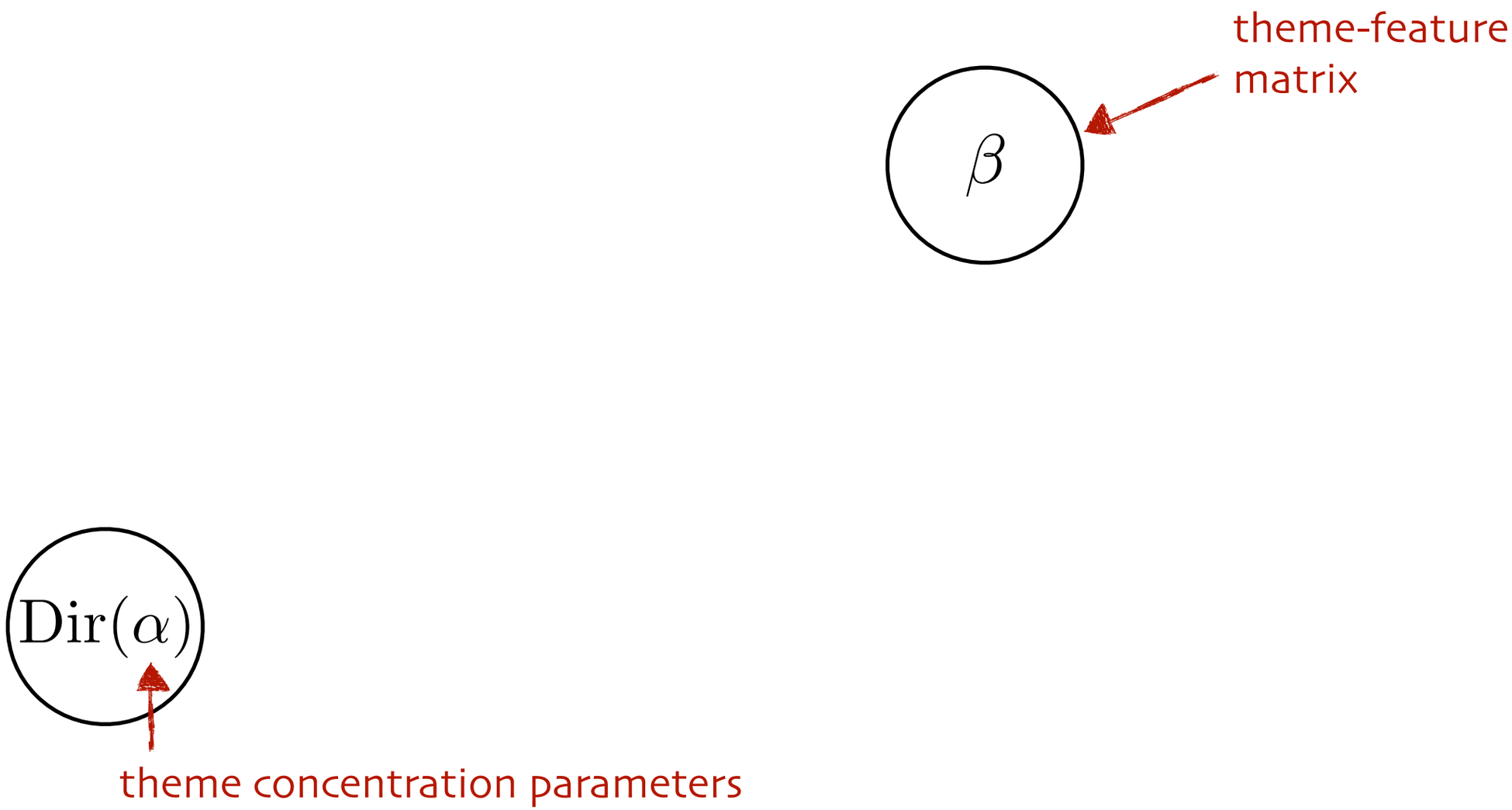
e.g. a top jet would be generated by sampling predominantly from one distribution, and a QCD jet predominantly from another distribution

- these distributions are referred to as "topics" or "themes"

Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

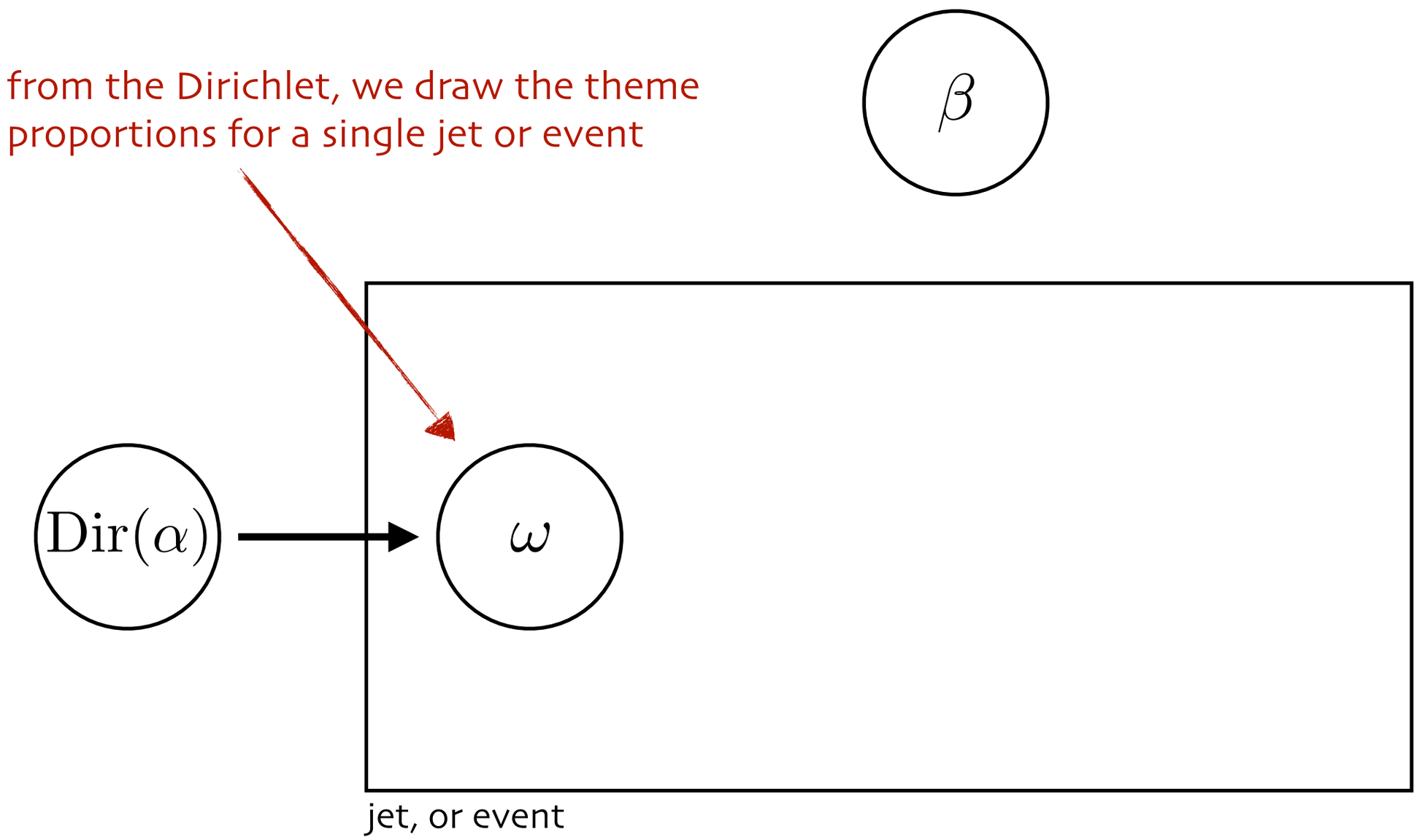
The LDA process for generating jets or events



Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events



Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

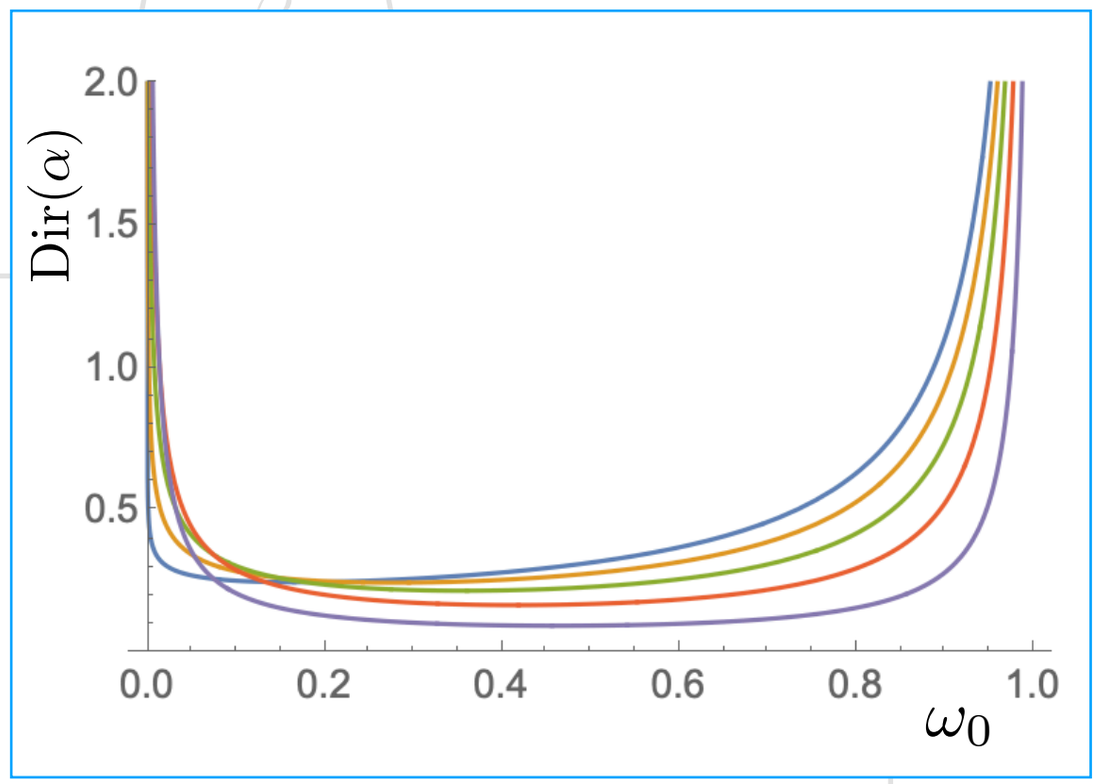
The LDA process for generating jets or events

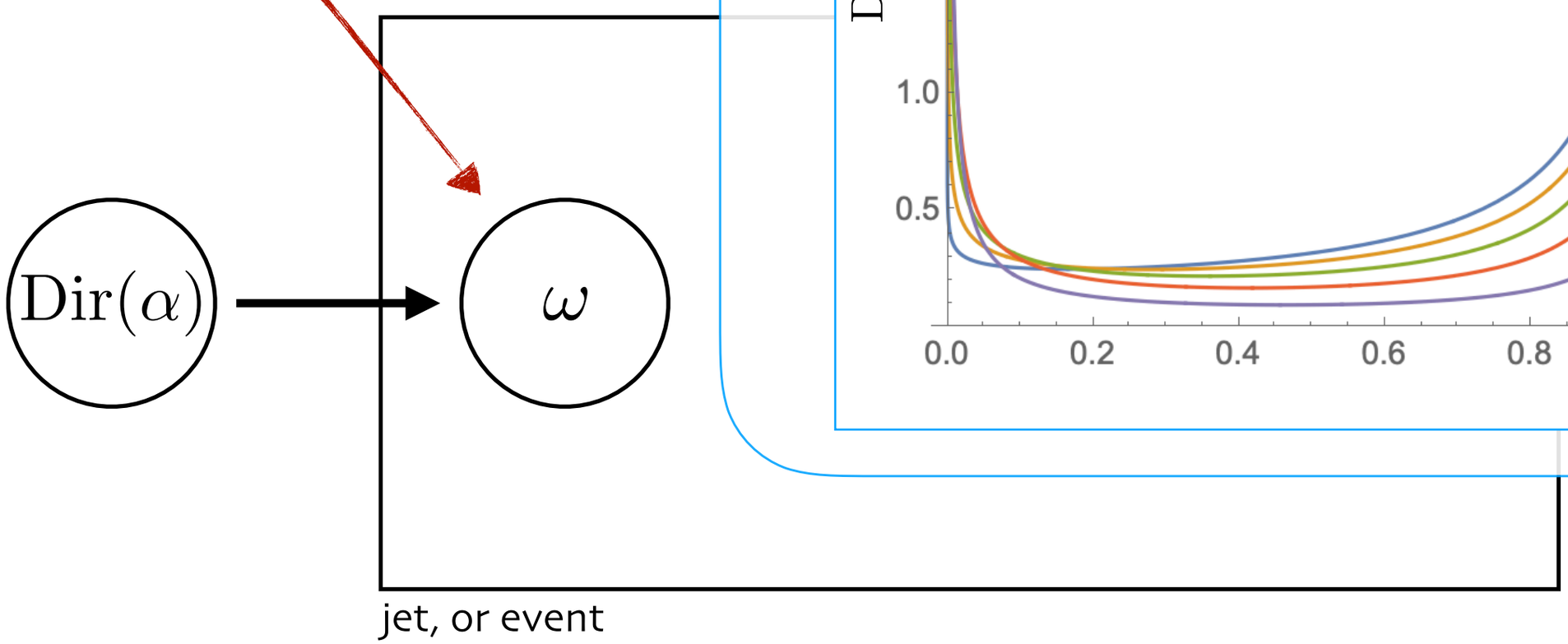
2 themes: one signal, one background

$$\Rightarrow \alpha = [\alpha_0, \alpha_1],\ \omega = [\omega_0, 1-\omega_0]$$

$$S/B \simeq \alpha_1/\alpha_0$$

from the Dirichlet, we draw the theme proportions for a single jet or event

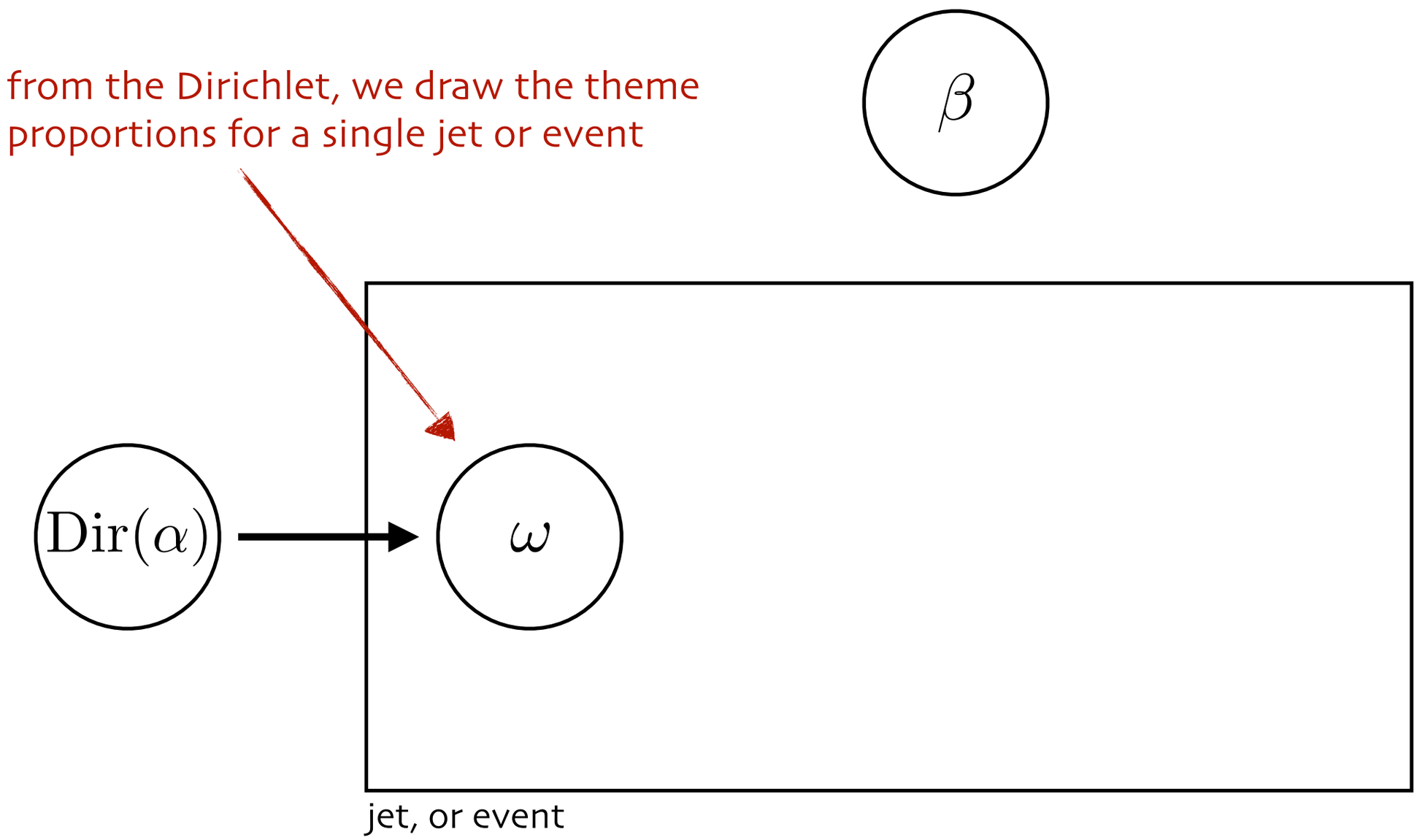




Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events

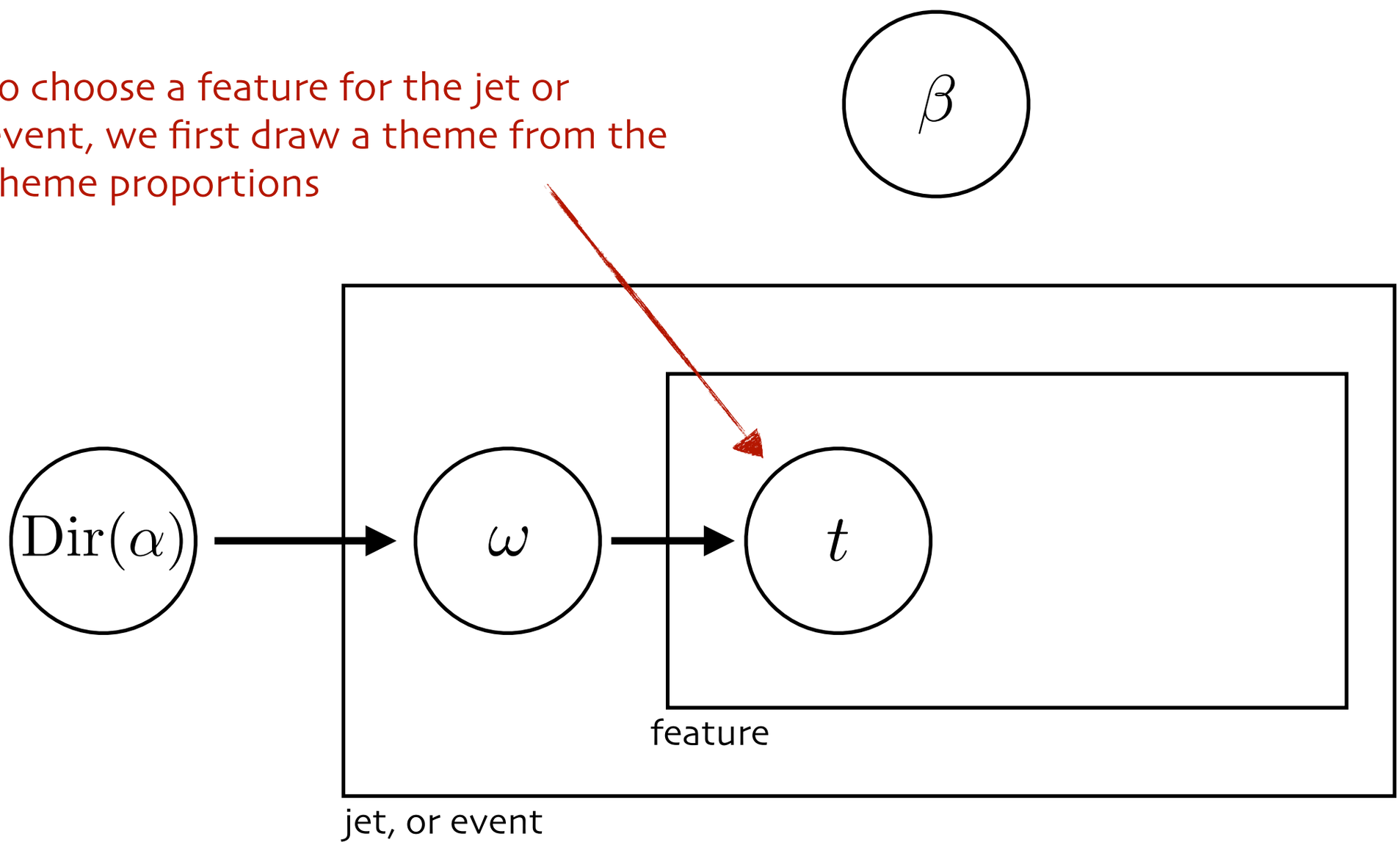


Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events

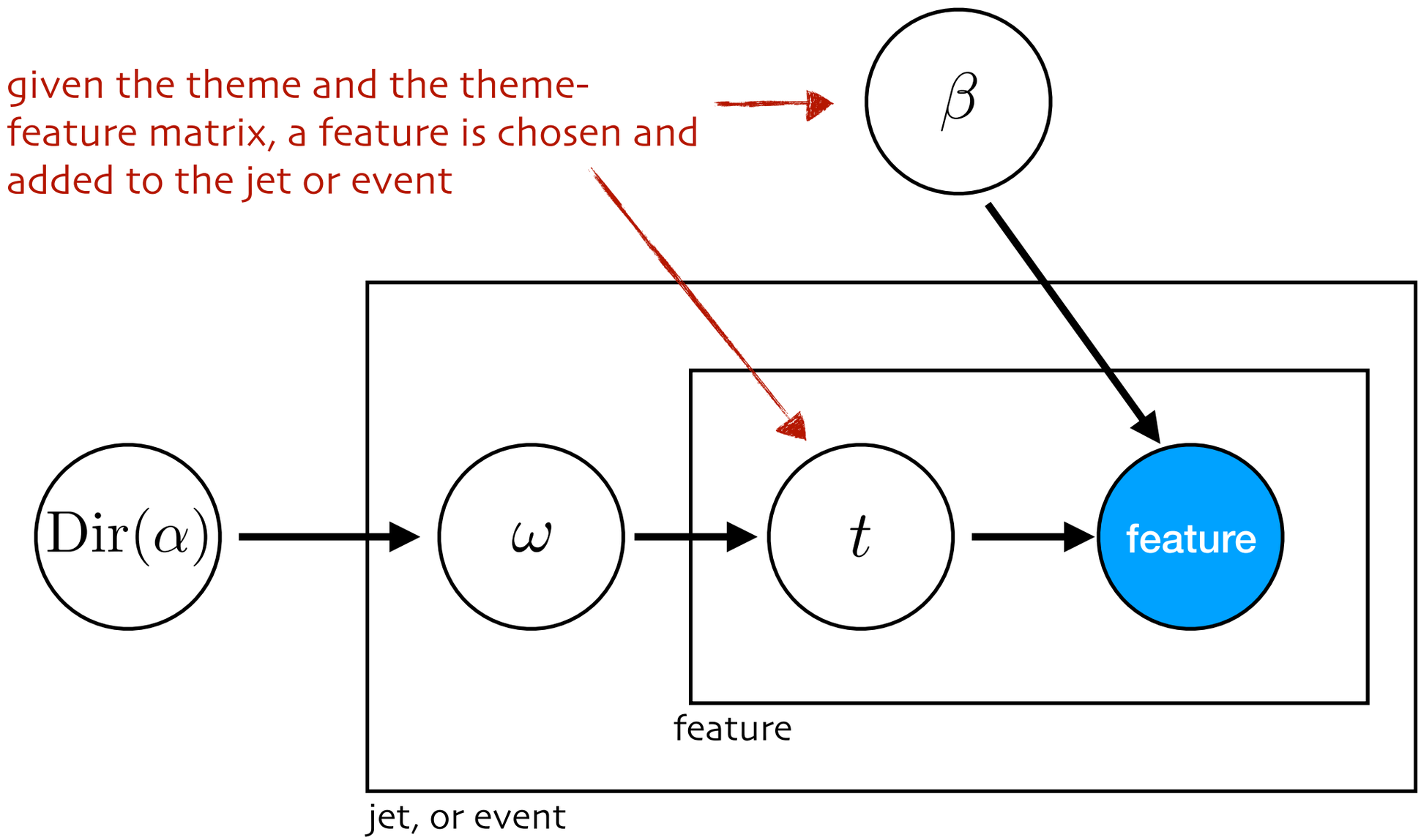
to choose a feature for the jet or event, we first draw a theme from the theme proportions



Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events



Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events

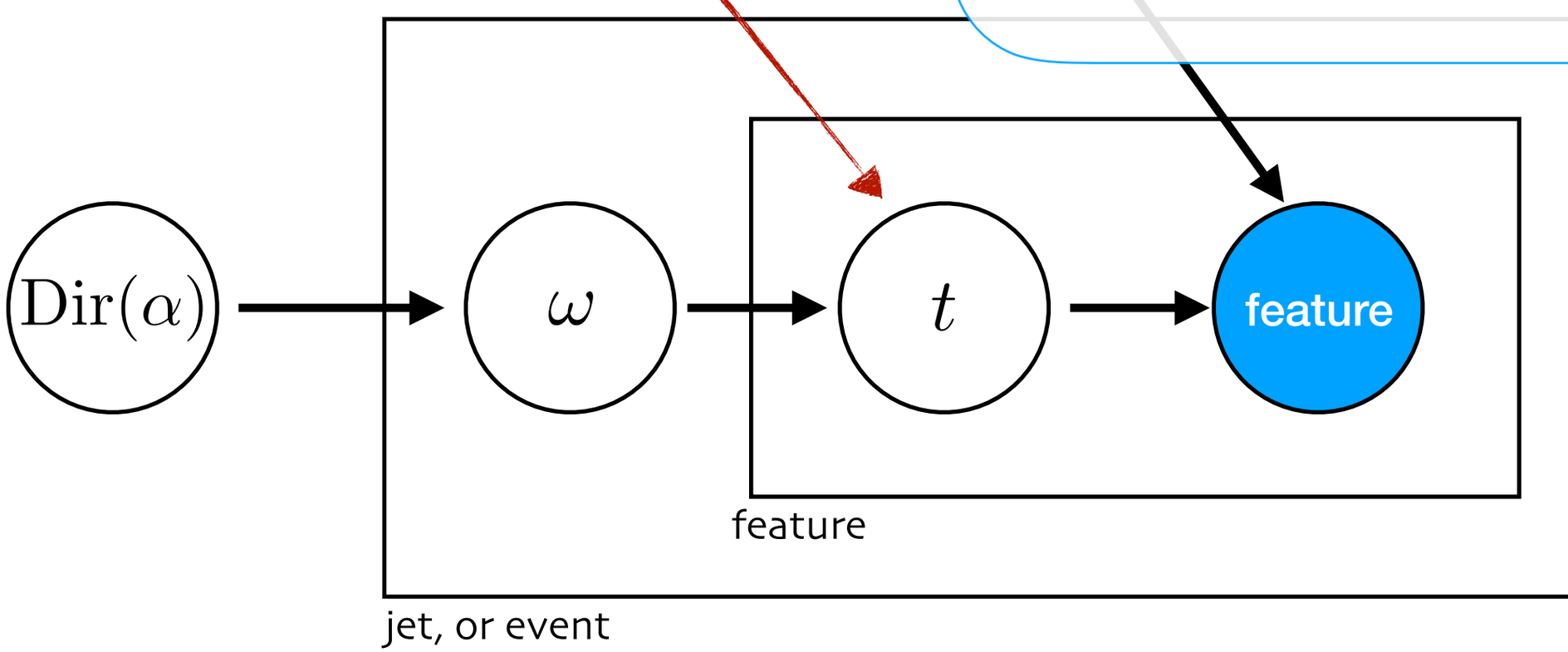
2 themes: one signal, one background

$$\Rightarrow \alpha = [\alpha_0, \alpha_1],\ \omega = [\omega_0, 1-\omega_0]$$

$$\beta = \begin{pmatrix} P_B(o_{j_0}) \\ P_S(o_{j_0}) \end{pmatrix}$$

distributions over feature space

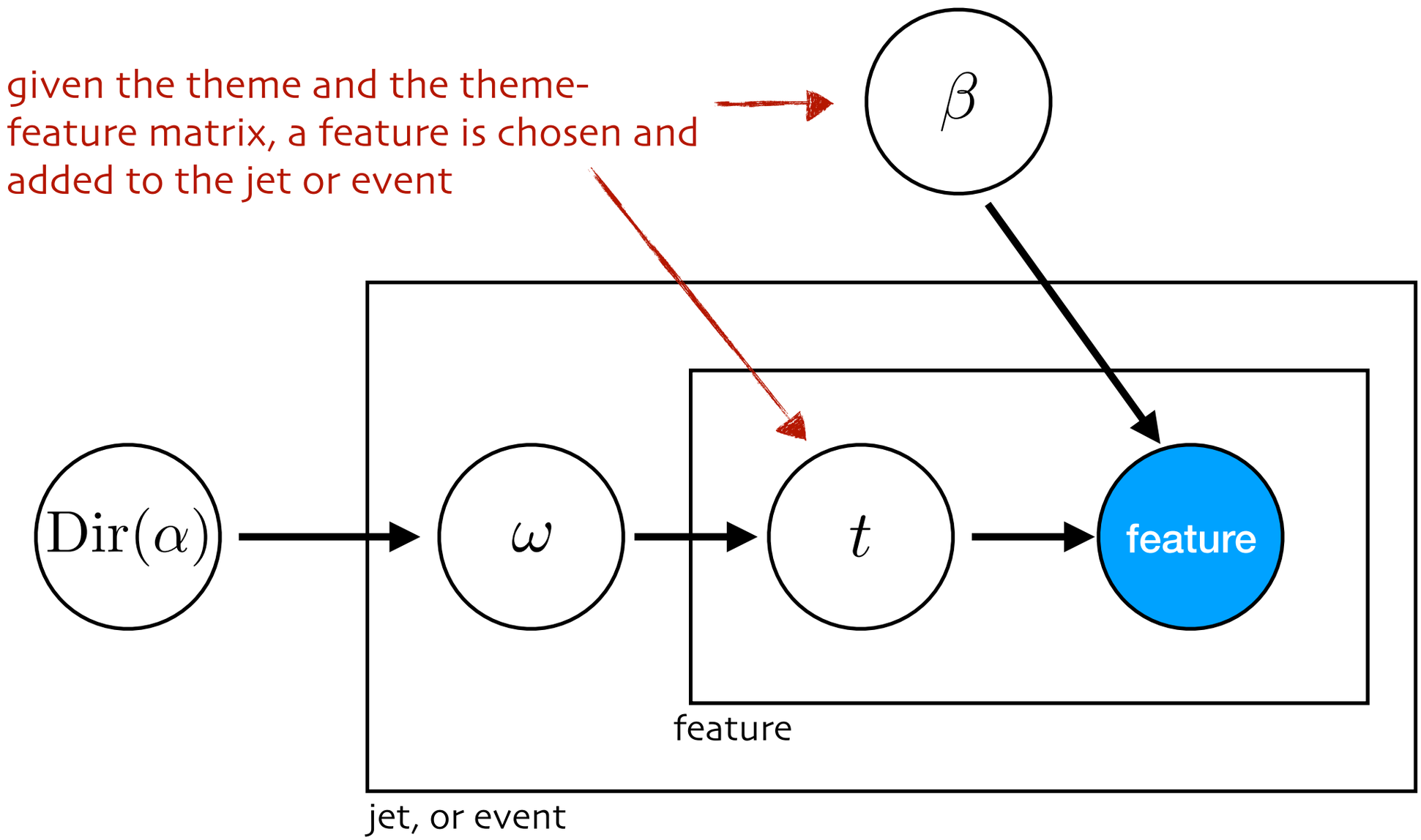
given the theme and the theme-feature matrix, a feature is chosen and added to the jet or event



Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events

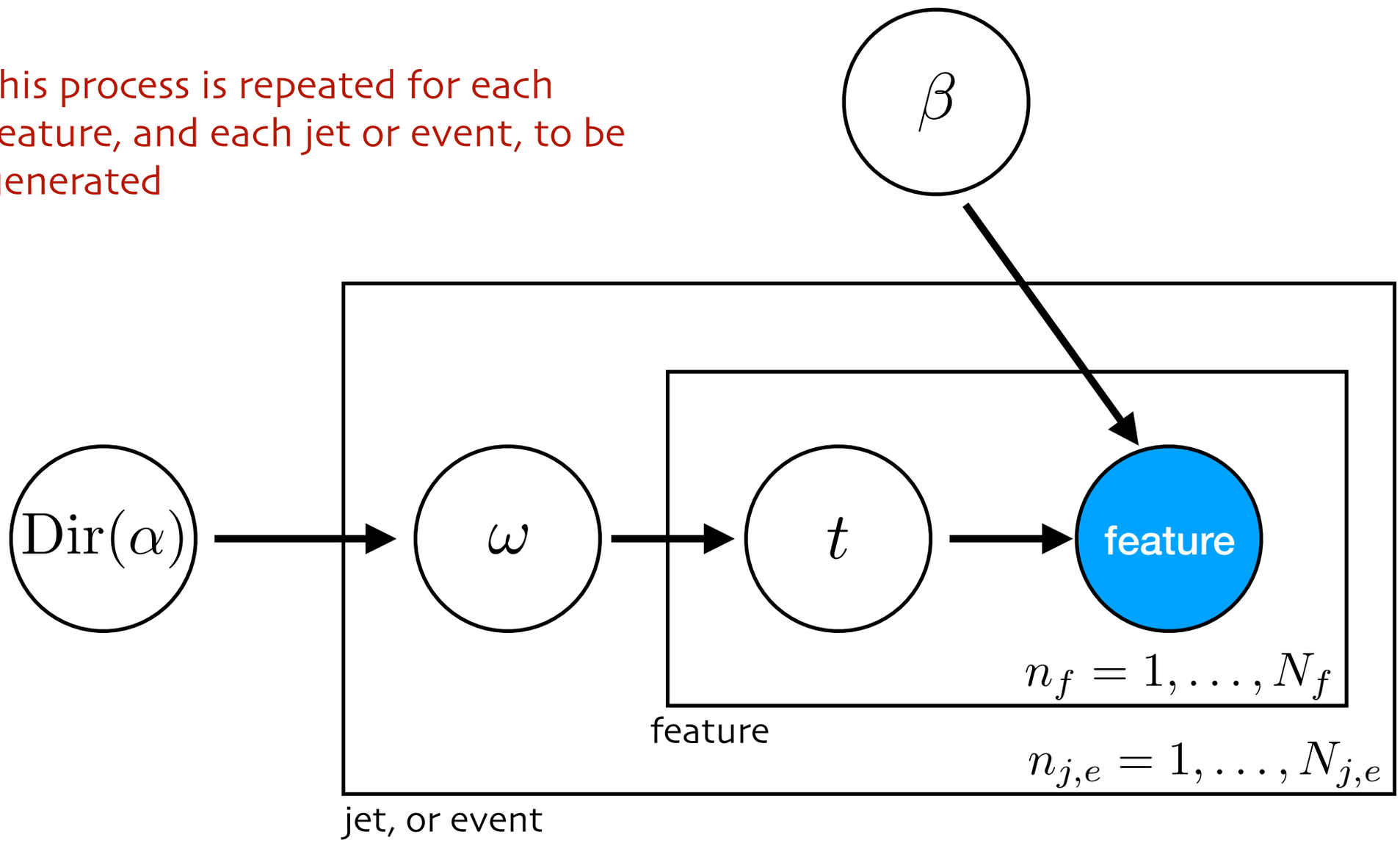


Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

The LDA process for generating jets or events

this process is repeated for each feature, and each jet or event, to be generated



Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

$$p(\text{jet}|\alpha,\beta) = \int_{\omega} p(\omega|\alpha) \overbrace{\prod_{f\in\text{jet}}}^{\text{probability of choosing all features in jet}} \left(\sum_{t} p(t|\omega) p(f|t,\beta) \right)$$

latent parameters

probability of choosing some theme proportions

probability of choosing some feature

marginalise over possible chosen theme proportions

marginalise over possible chosen themes

Latent Dirichlet Allocation

D. M. Blei, A. Y. Ng, M. I. Jordan, J. Lafferty (2003)

$$p(\text{jet}|\alpha,\beta) = \int_{\omega} p(\omega|\alpha) \overbrace{\prod_{f\in\text{jet}}}^{\text{probability of choosing all features in jet}} \left(\sum_{t} p(t|\omega) p(f|t,\beta) \right)$$

latent parameters

probability of choosing some theme proportions

probability of choosing some feature

marginalise over possible chosen theme proportions

marginalise over possible chosen themes

Infer latent parameters: $$p(\alpha,\beta|\text{jets}) = \frac{p(\text{jets}|\alpha,\beta)p(\alpha,\beta)}{\int d\alpha d\beta \; p(\text{jets}|\alpha,\beta)}$$

Bayes theorem

numerical techniques required

Variational Bayes technique
GENSIM software

M. D. Hoffman, D. M. Blei, F. Bach (2010)
R. Rehurek, P. Sojka (2010)

co-occurrence of features!

LDA new physics tagging

W′ search: $pp \to W' \to \phi W \to WWW,\ m_{W'} = 3\ \mathrm{TeV},\ m_\phi = 400\ \mathrm{GeV}$

K. Agashe, J. H. Collins, P. Du, S. Hong, D. Kim, R. K. Mishra (2018)

W′ event "features" → $m_{j_0} \sim m_\phi,\ m_W \quad \& \quad \dfrac{m_{j_1}}{m_{j_0}} \sim \dfrac{m_W}{m_\phi}$

QCD events ~ featureless

Setup:

1 - mixed unlabelled samples: $S/B = \ 0.011,\ 0.0058$

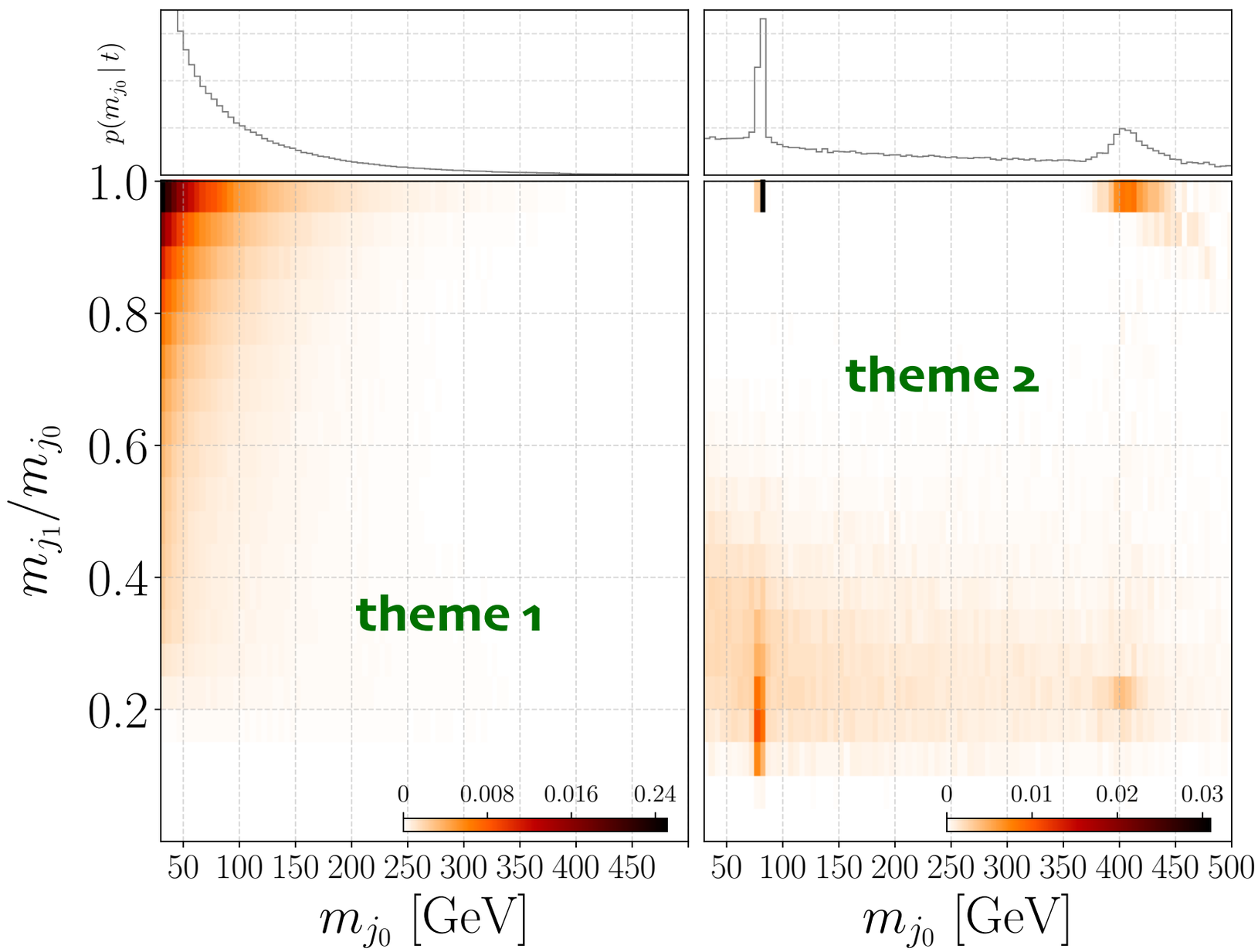
$m_{jj} \in [2750, 3250]\ \mathrm{GeV}$

3 - train with two themes, extract latent distributions

4 - infer latent content of jets/events: $\mathrm{event} = [\omega_0, 1 - \omega_0]$ ← classification

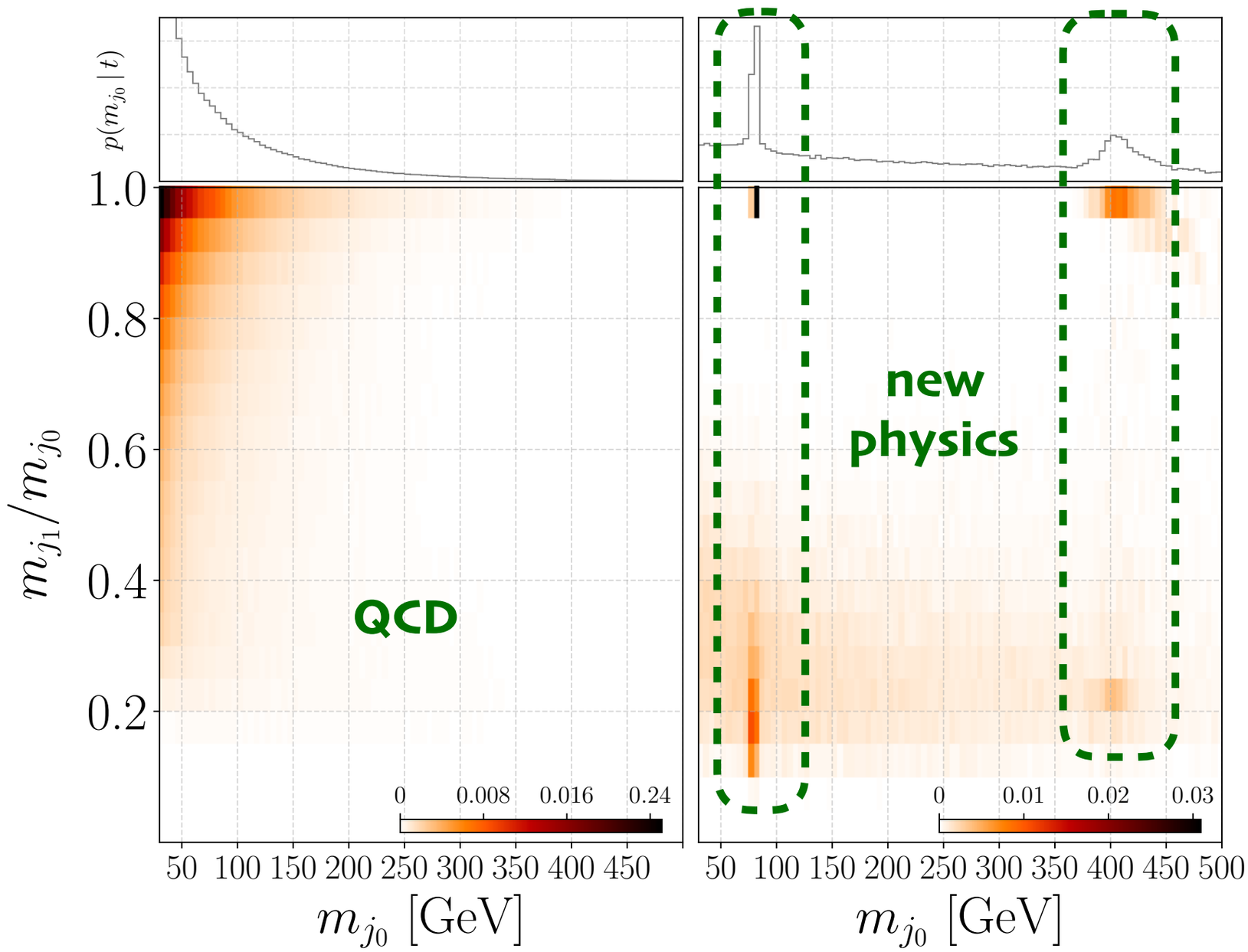
LDA new physics tagging

Latent themes uncovered by the algorithm for $S/B = 0.011$, $\alpha = [0.989, 0.011]$



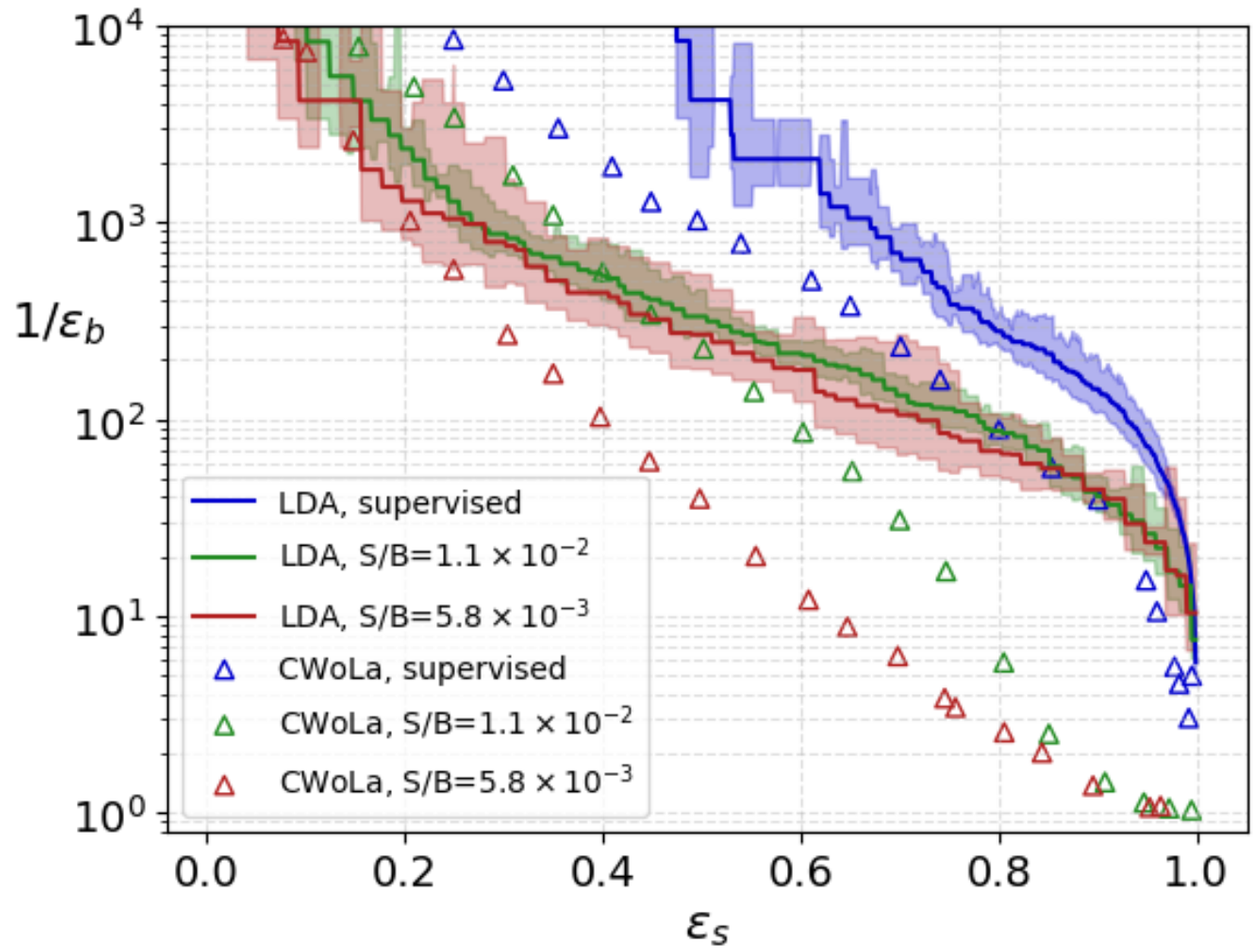
LDA new physics tagging

Latent themes uncovered by the algorithm for $S/B = 0.011$, $\alpha = [0.989, 0.011]$



LDA new physics tagging

Measure performance with ROC curves:



results compared to CWoLa tagger

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

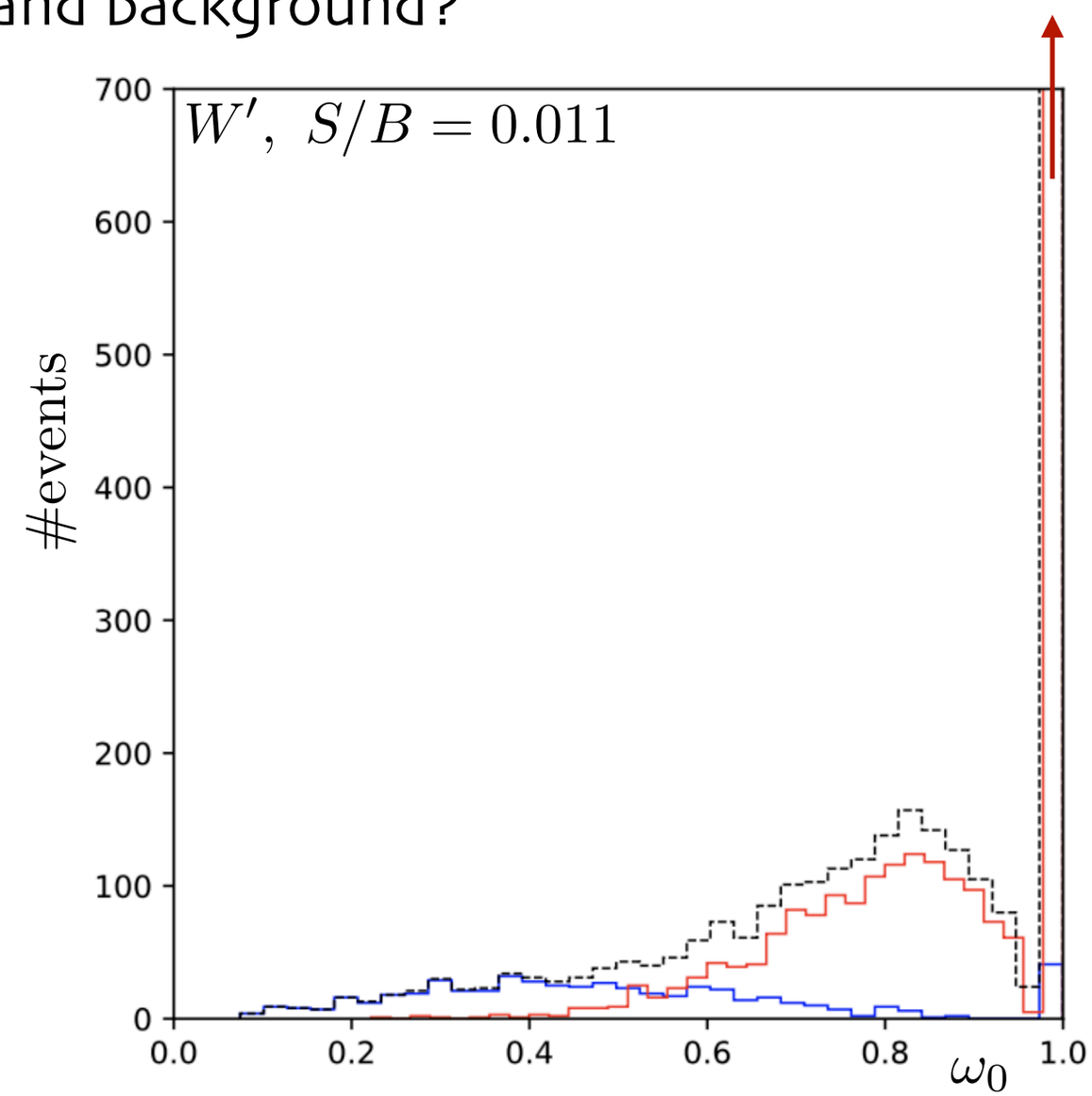
results have been k-folded, k=10, to estimate robustness

Clustering in latent space

Unsupervised → no truth labels → $\text{event} = [\omega_0, 1 - \omega_0]$

How do we know where to split signal and background?

How do we even know there's a signal?

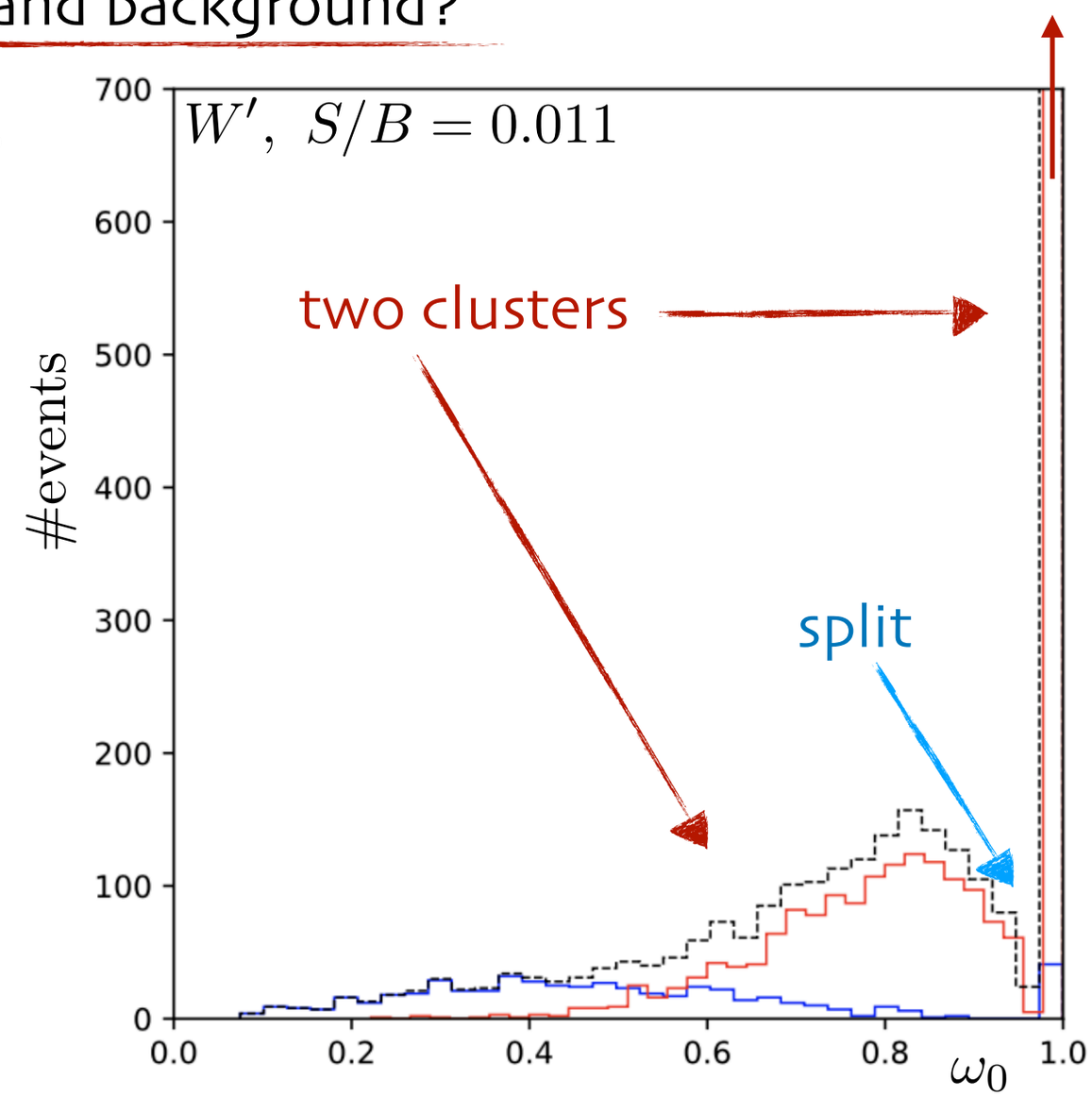


Clustering in latent space

Unsupervised → no truth labels → $\text{event} = [\omega_0, 1 - \omega_0]$

How do we know where to split signal and background?

How do we even know there's a signal?



clustering algorithm determines split

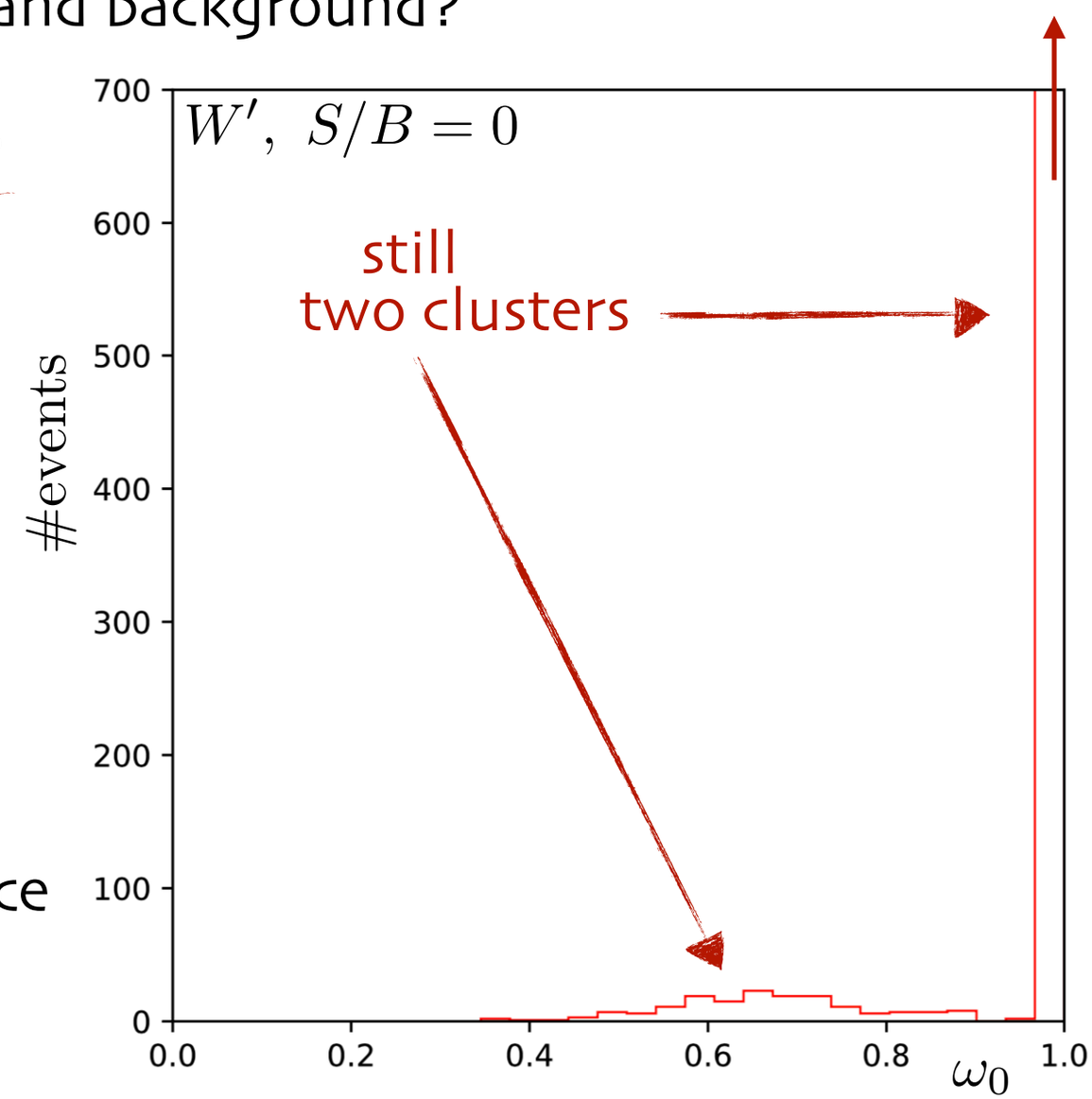
In this case:
signal is cleaned from $S/B = 0.011$
to ~ 0.4

Clustering in latent space

Unsupervised → no truth labels → $\text{event} = [\omega_0, 1 - \omega_0]$

How do we know where to split signal and background?

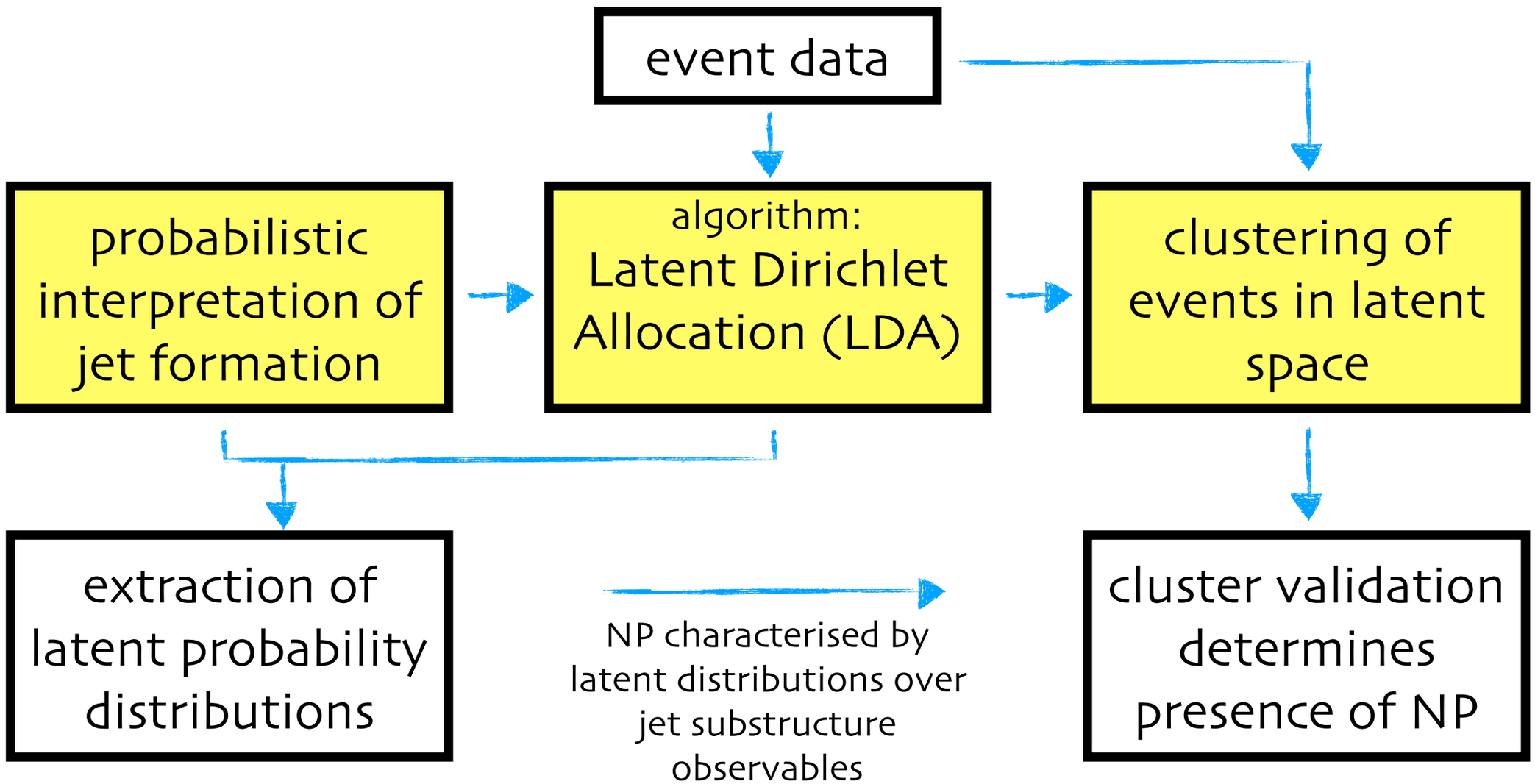
How do we even know there's a signal?



We can:

- use cluster validation techniques
  - internal measures
- use more than 2 themes
  - 3 themes: cluster in 2D latent space
- HDP: #themes not fixed

Summary and next steps



So far: only a proof-of-concept

Next steps: implement a NP search, extend to other signals and test cases
(also: LHC Olympics + Dark Machines challenges)

additional slides

LDA top tagging

top jet "features" → $m_{j_0} \sim m_t, m_W \quad \& \quad \frac{m_{j_1}}{m_{j_0}} \sim \frac{m_W}{m_t}$

QCD jets ~ featureless

Our LDA method has four modes: unsupervised or supervised jet or event classification

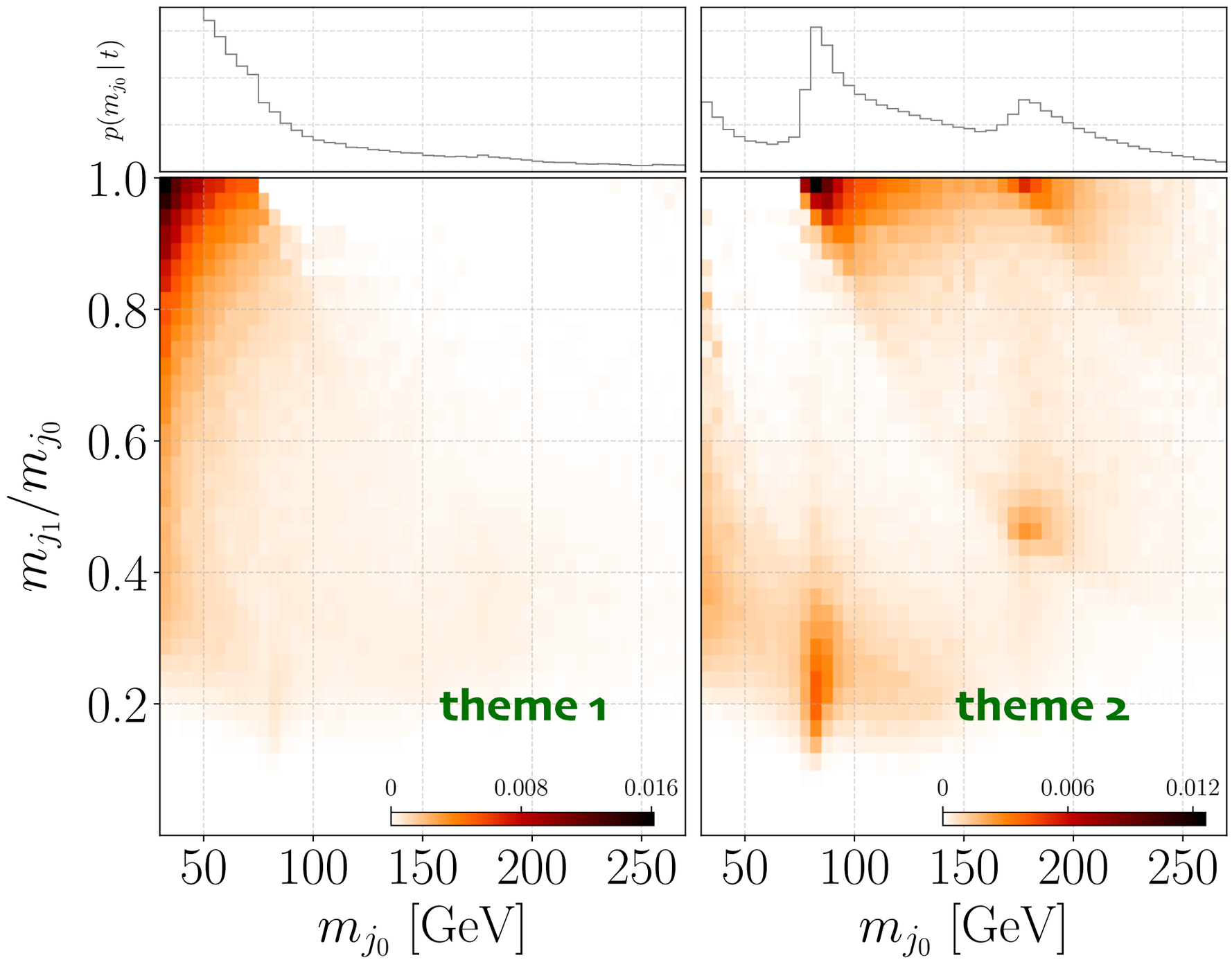
1 - mixed unlabelled samples: $S/B = 1,\ 1/9,\ 1/99$

3 - train with two themes, extract latent distributions

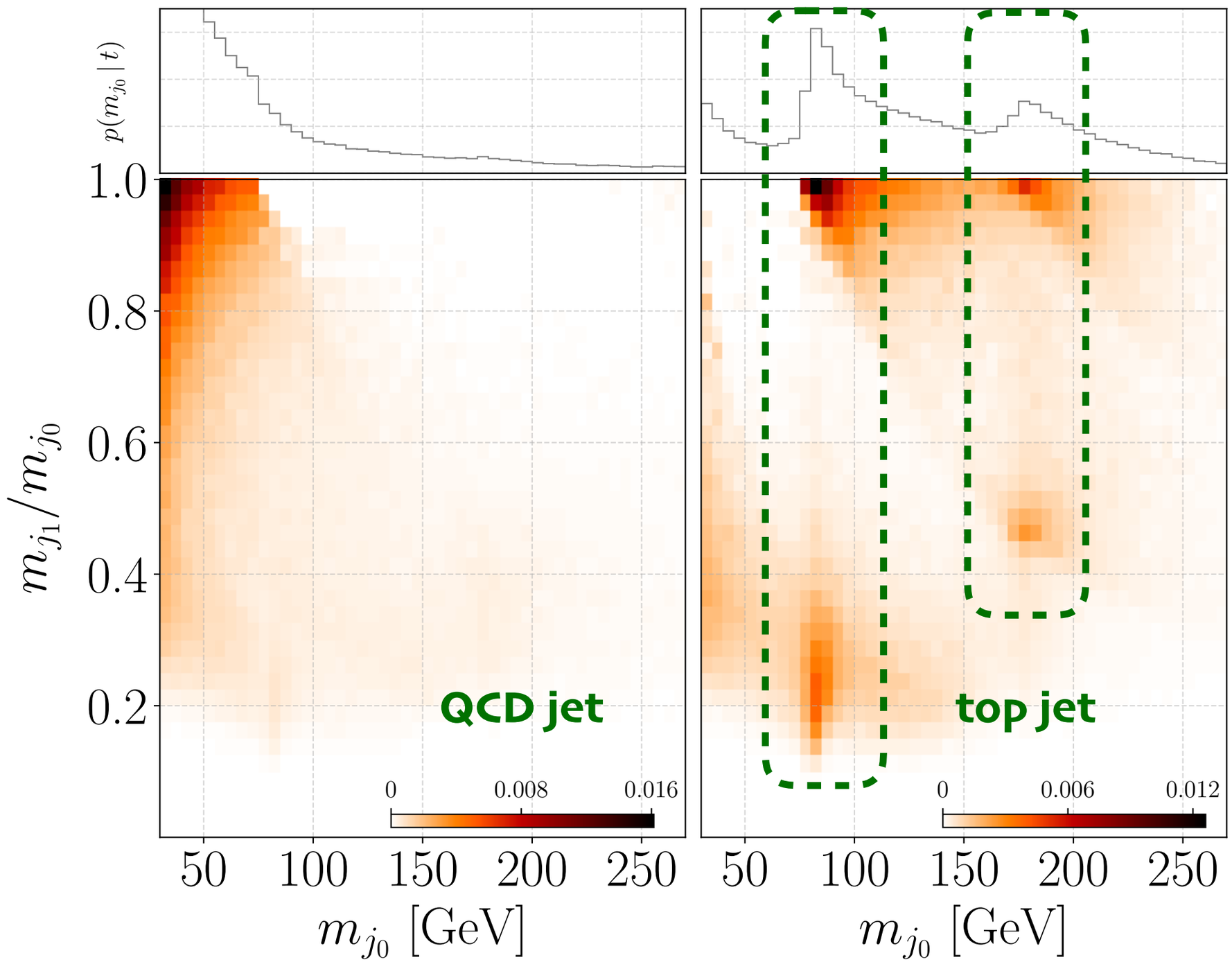
4 - infer latent content of jets/events:

$$\text{event} = [\omega_0, 1 - \omega_0]$$ ← classification

LDA top tagging

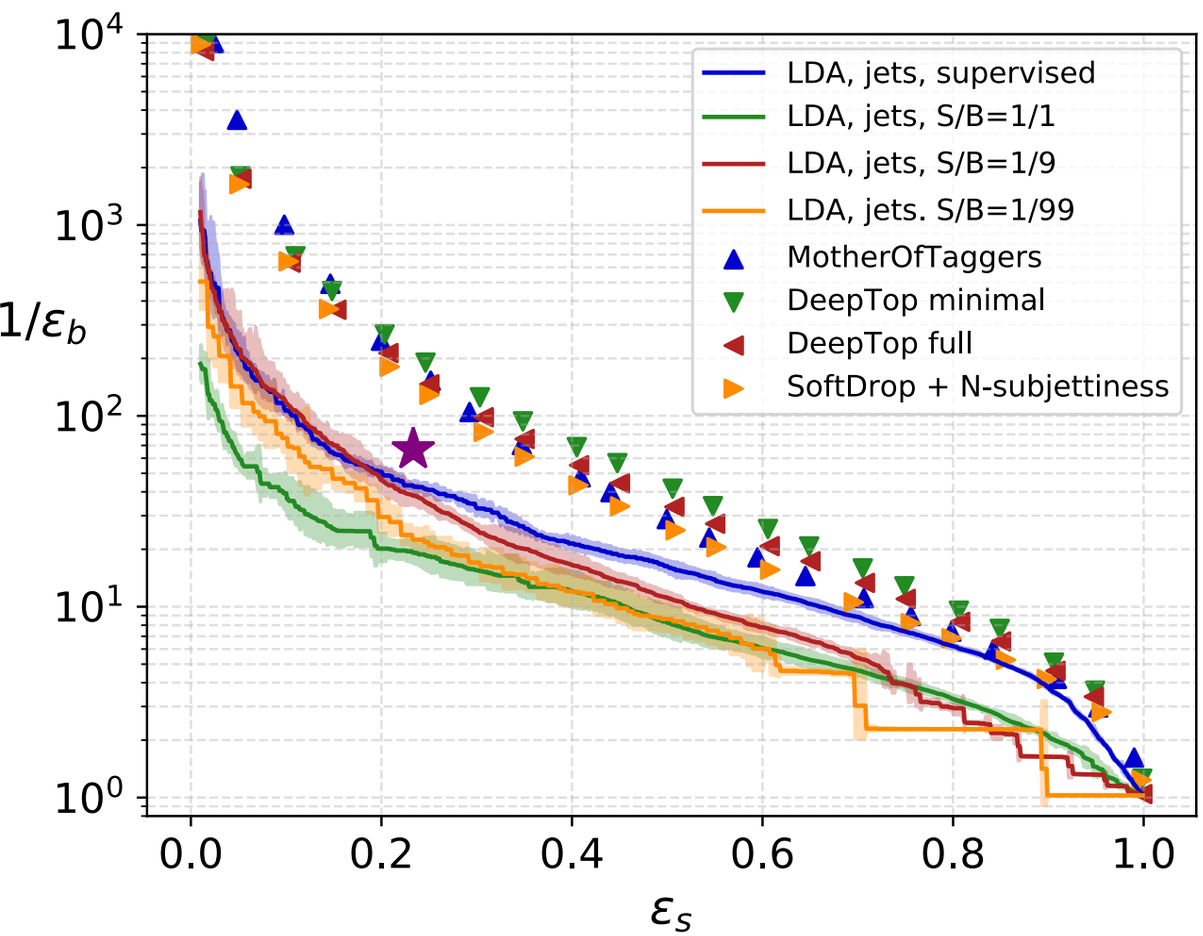


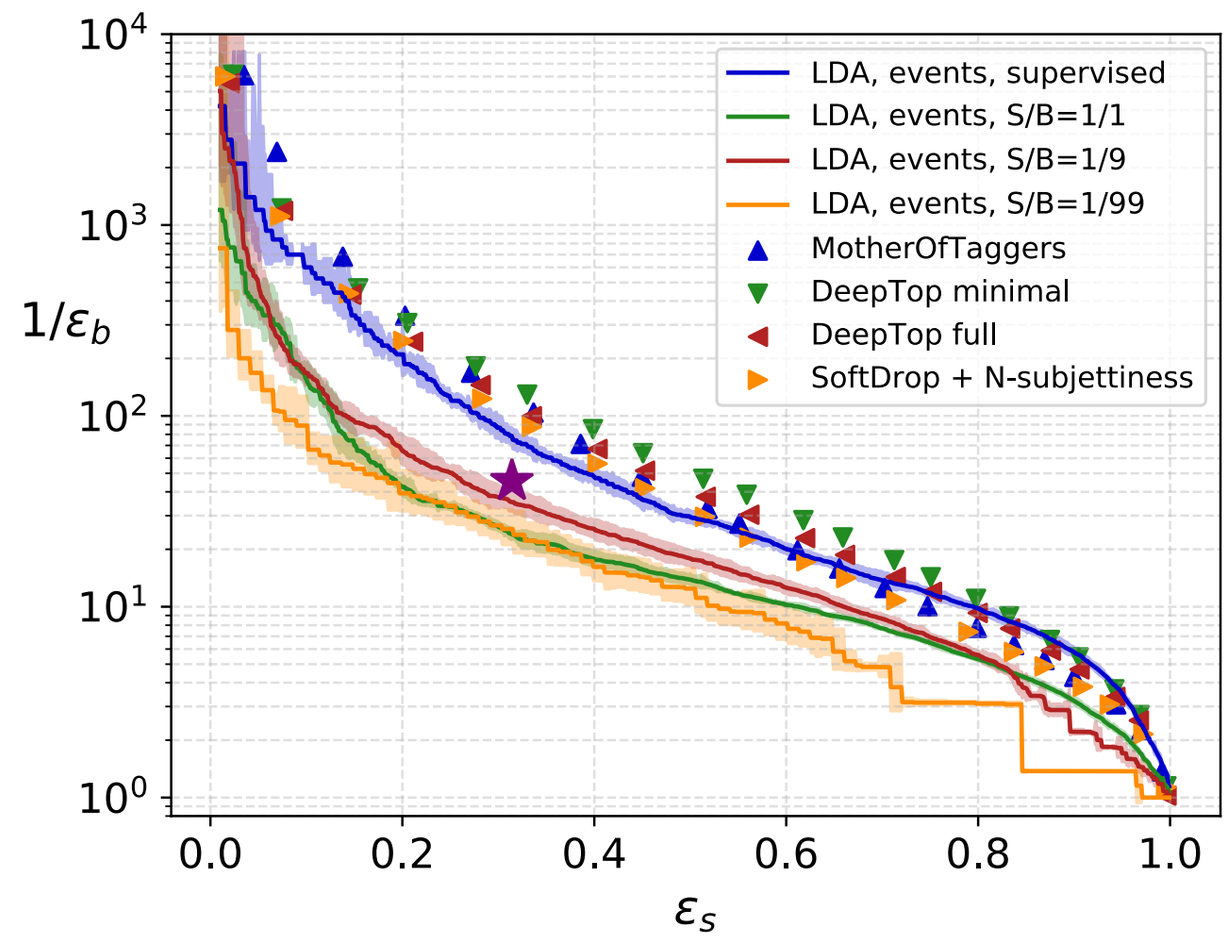
LDA top tagging



LDA top tagging

Measure performance with ROC curves:





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

results compared to JH top tagger (purple star) and DeepTop

results have been k-folded, k=10, to estimate robustness

Beta distribution

2 themes: one signal, one background

$$\Rightarrow \alpha = [\alpha_0, \alpha_1],\ \omega = [\omega_0, 1-\omega_0]$$

$$\begin{aligned}\mathrm{Dir}(\omega_0;\alpha) =& \mathrm{B}(\omega_0;\alpha_0,\alpha_1)\\ =& \frac{\Gamma(\alpha_0+\alpha_1)}{\Gamma(\alpha_0)\Gamma(\alpha_1)}\omega_0^{\alpha_0-1}(1-\omega_0)^{\alpha_1-1}\end{aligned}$$

$$\int_0^1 d\omega_0\ \mathrm{B}(\omega_0;\alpha_0,\alpha_1)\,(\omega_0 t_0 + (1-\omega_0)t_1)$$

$$= \frac{\alpha_0 t_0 + \alpha_1 t_1}{\alpha_0+\alpha_1}$$

$$\longrightarrow \quad S/B \simeq \alpha_1/\alpha_0$$