

Unsupervised event classification using probabilistic modelling

Phys.Rev. D100 (2019) no.5, 056002 [arxiv:1904.04200]

Barry M. Dillon

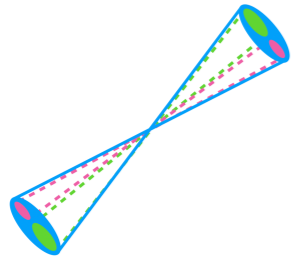
in collaboration with:
D. Faroughy, J. Kamenik, & M. Swezc

Particle Physics in the Computing Frontier, 11/12/2019

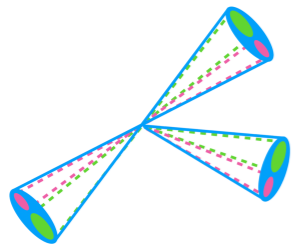
Outline of the talk



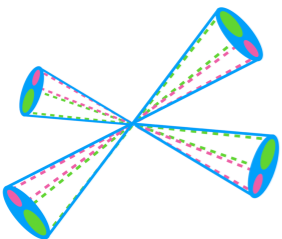
Jets in the Standard Model and beyond



Uncovering latent jet substructure



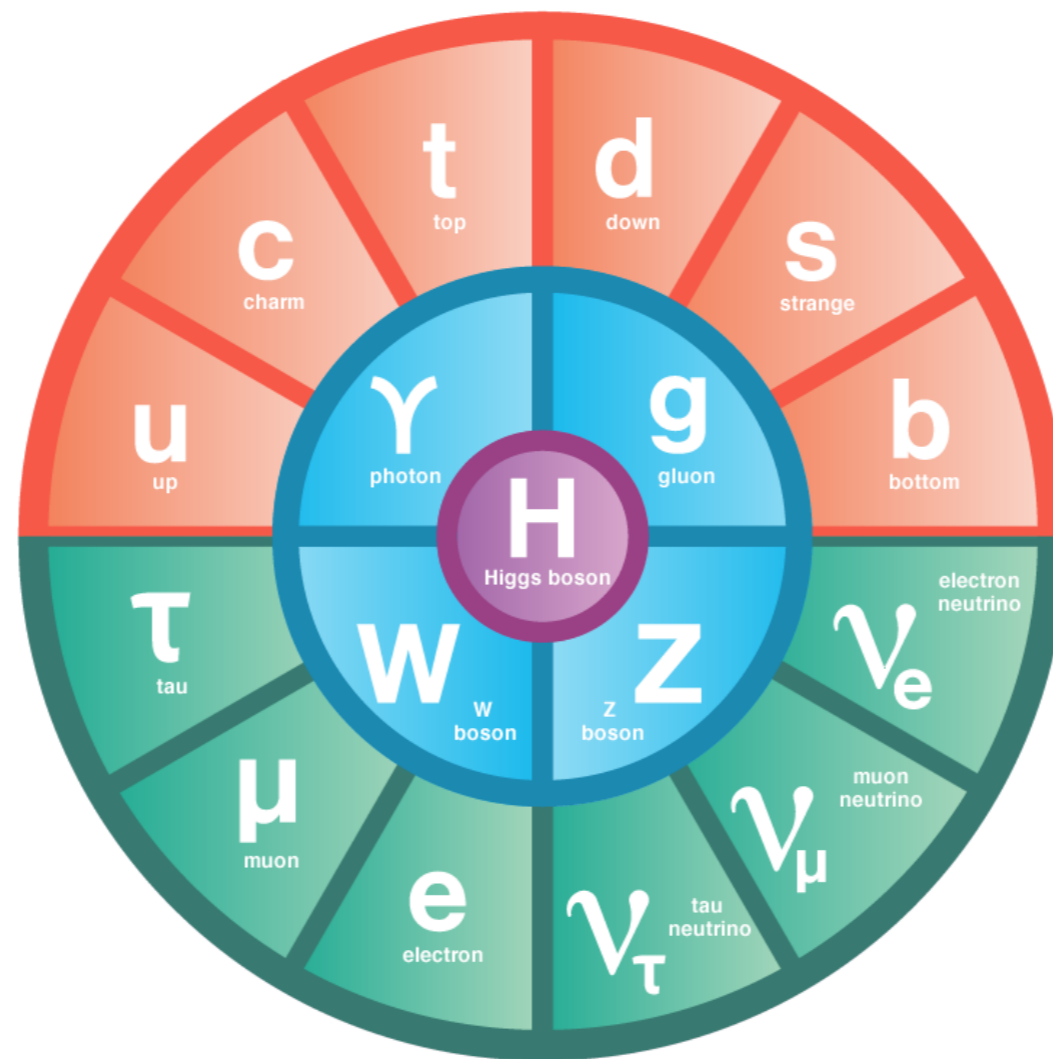
Data-driven/unsupervised top-tagging



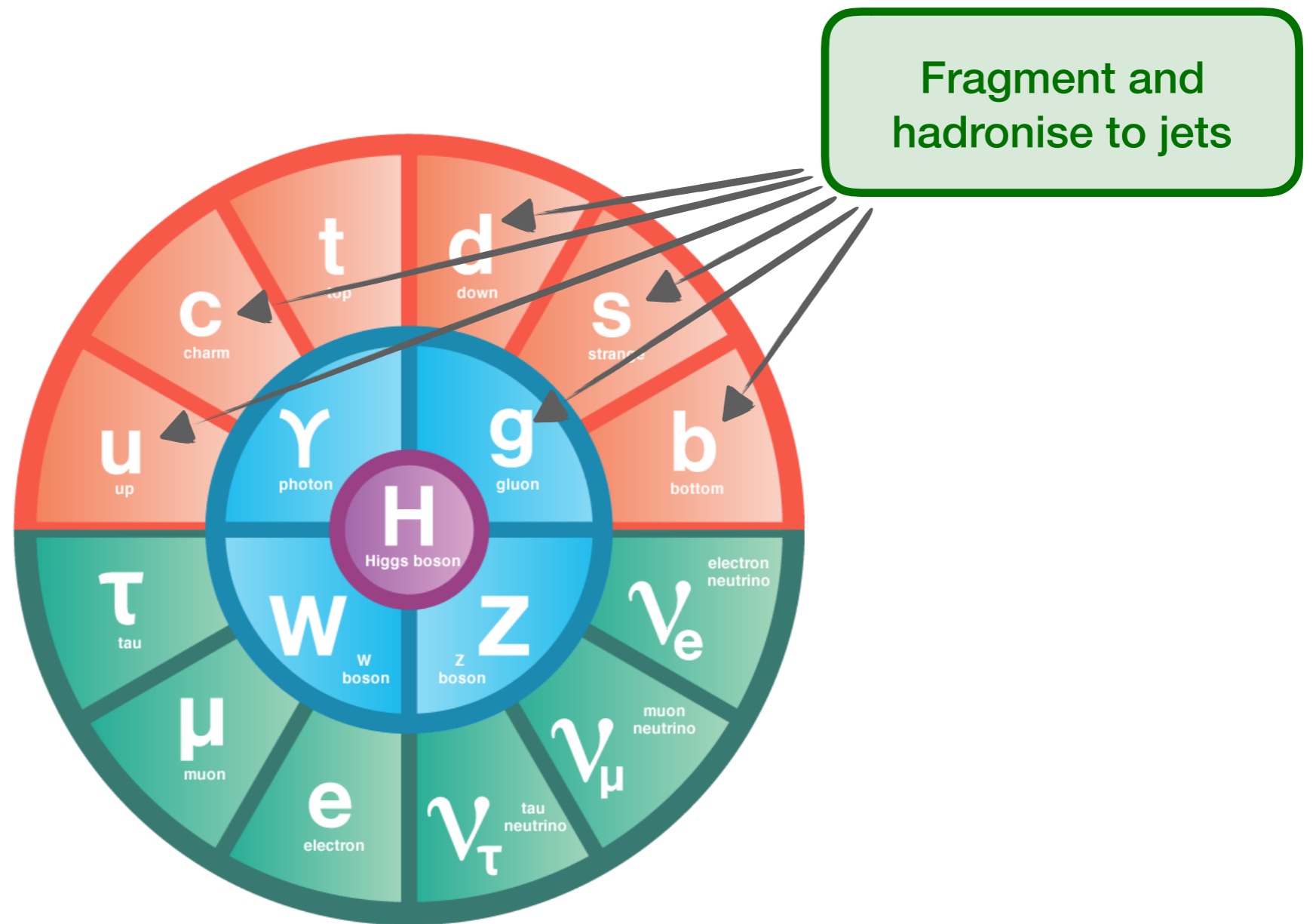
Unsupervised search for new physics

Jets in the Standard Model

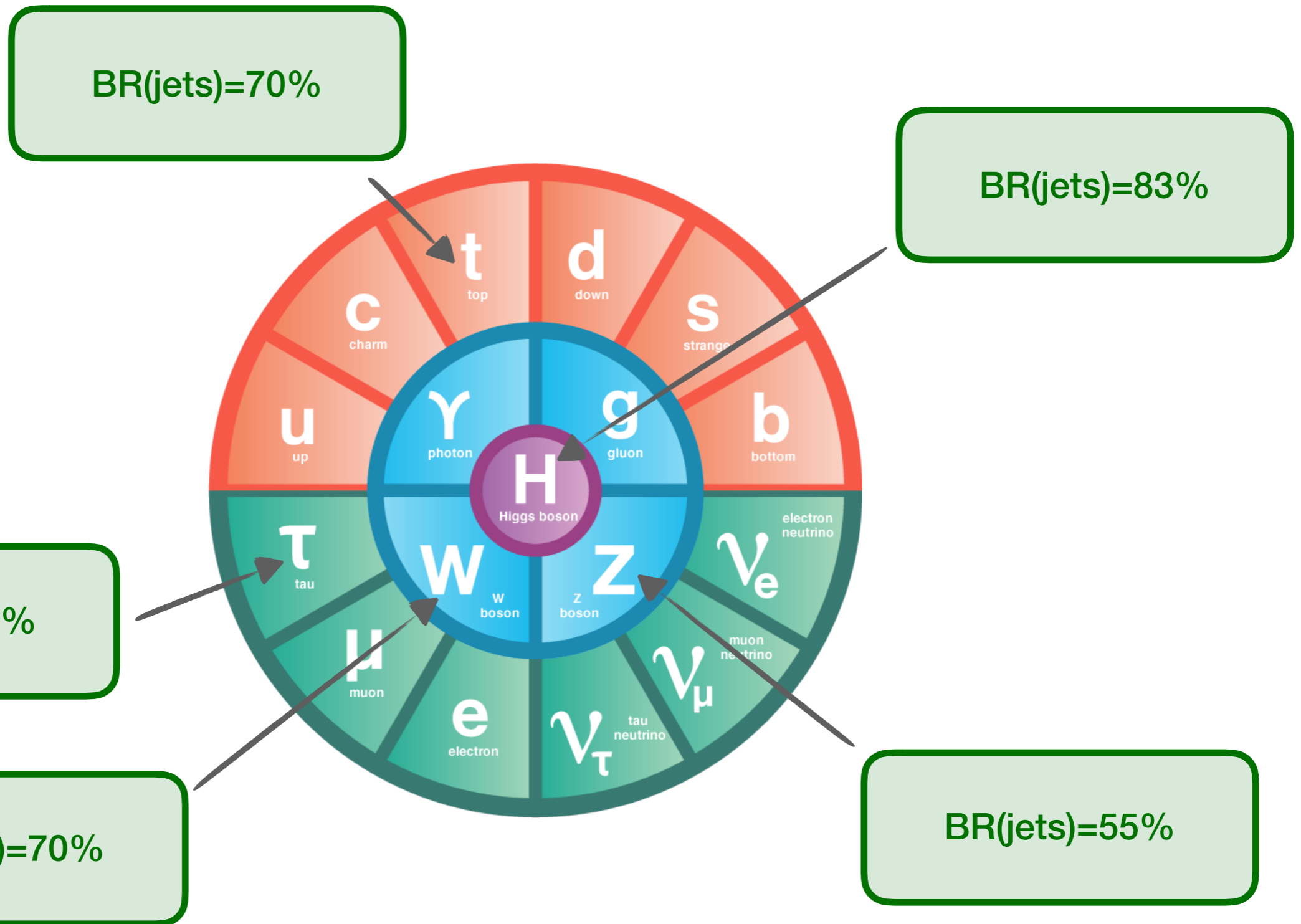
Jets in the Standard Model



Jets in the Standard Model

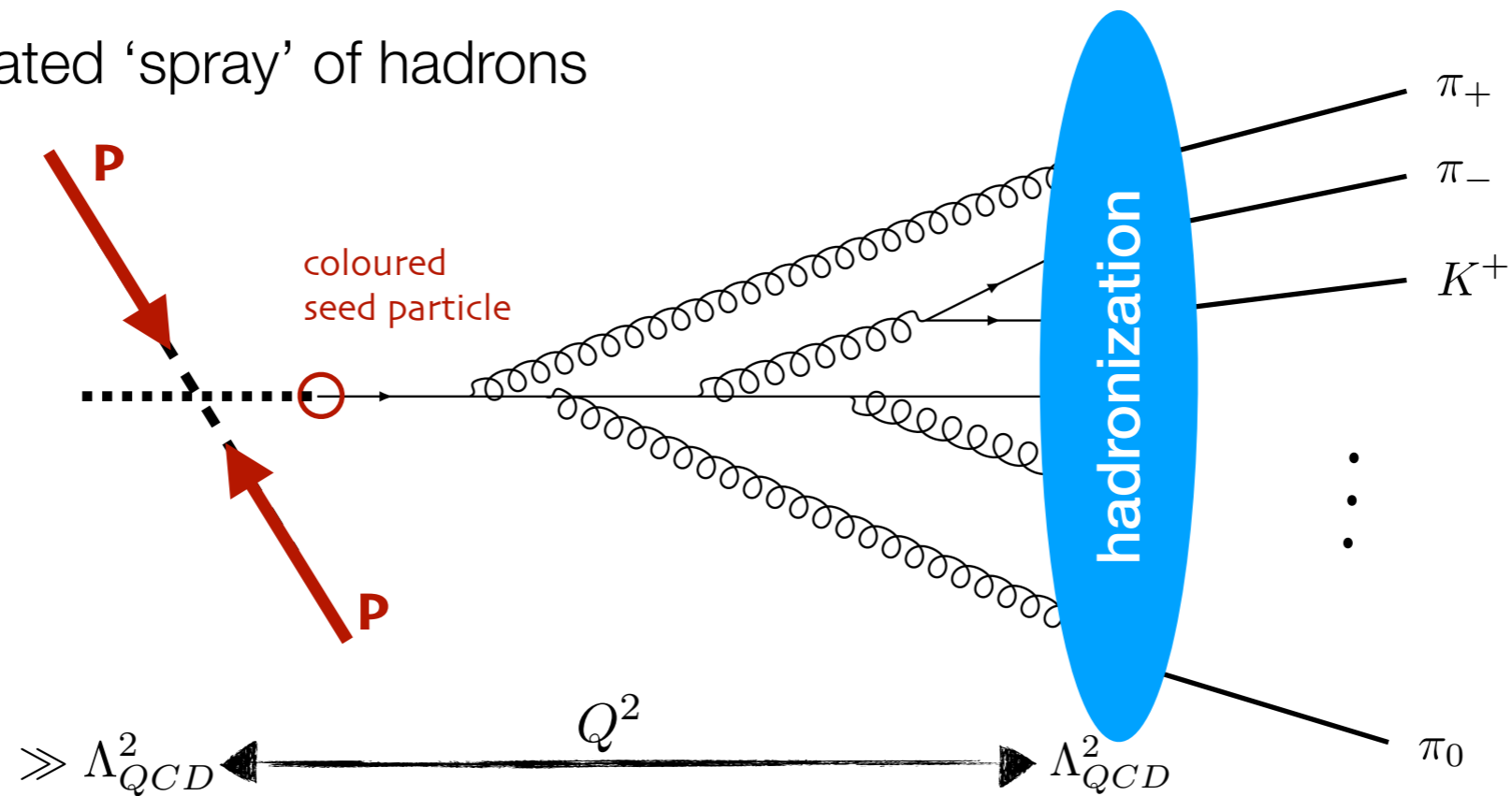


Jets in the Standard Model



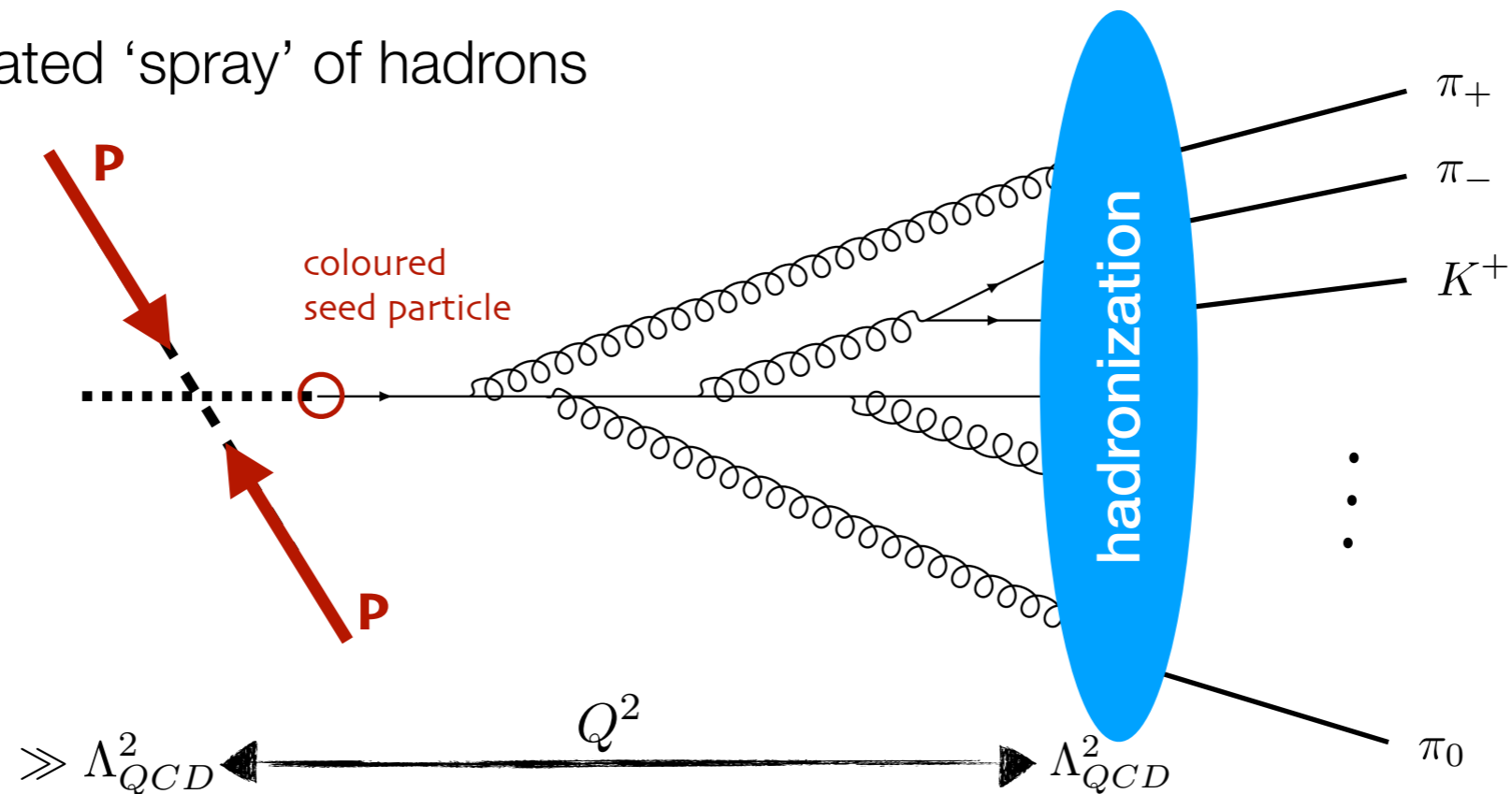
Jet basics

Jet: a collimated 'spray' of hadrons



Jet basics

Jet: a collimated 'spray' of hadrons



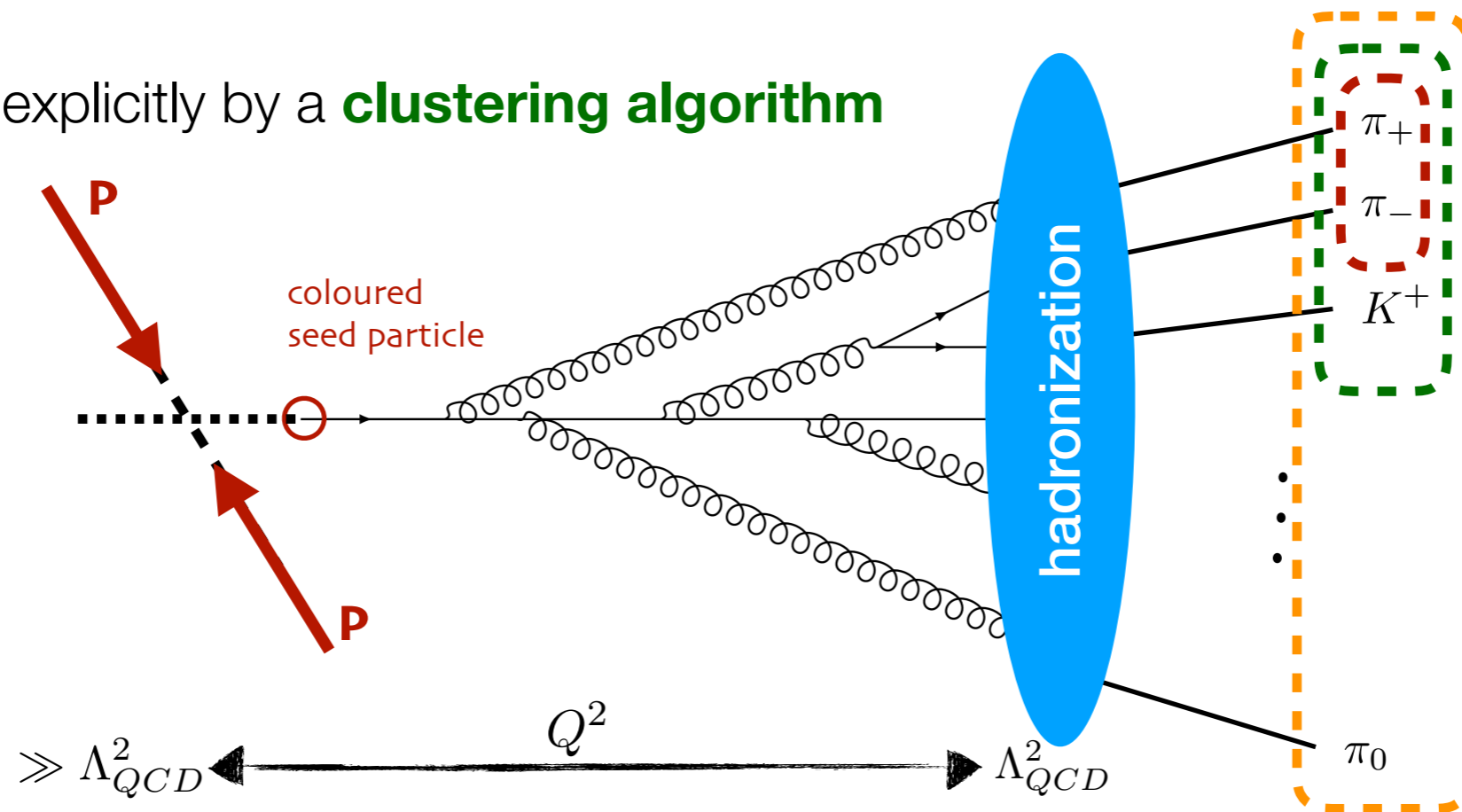
QCD factorises. At large energy scales:

$$P_q(z) = C_F \frac{1 + (1-z)^2}{z} \quad P_g(z) = C_A \left(2 \frac{1-z}{z} + z(1-z) + \frac{n_f T_R}{C_A} (z^2 + (1-z)^2) \right)$$

$$q_p \rightarrow q(1-z)p + g_{zp} \quad g_p \rightarrow g(1-z)p + g_{zp} \quad g_p \rightarrow q(1-z)p + q_{zp}$$

Jet basics

Jet: defined explicitly by a **clustering algorithm**



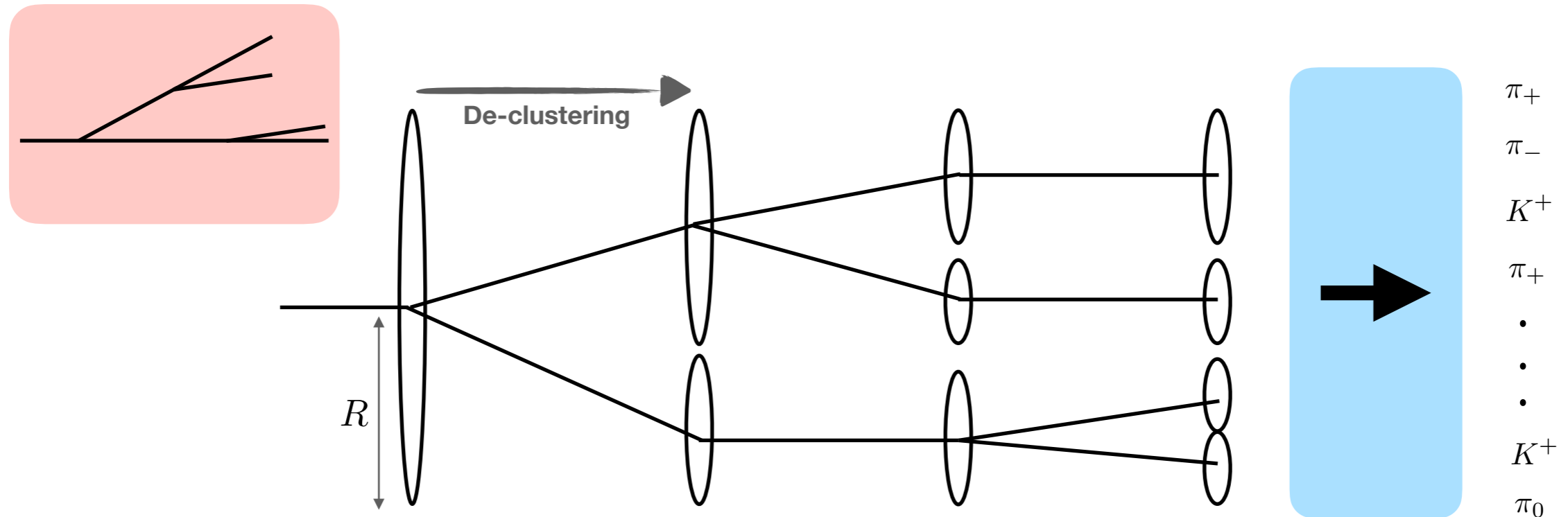
Procedure:

- 1 - calculate pairwise distance between every pair of objects
- 2 - merge the closest two objects
- 3 - repeat until no two particles are closer than a distance **R**

$$d_{ij} = \min(p_{Ti}^r, p_{Tj}^r) \frac{R_{ij}}{R}$$

Jet basics

Jet: defined explicitly by a **clustering algorithm**

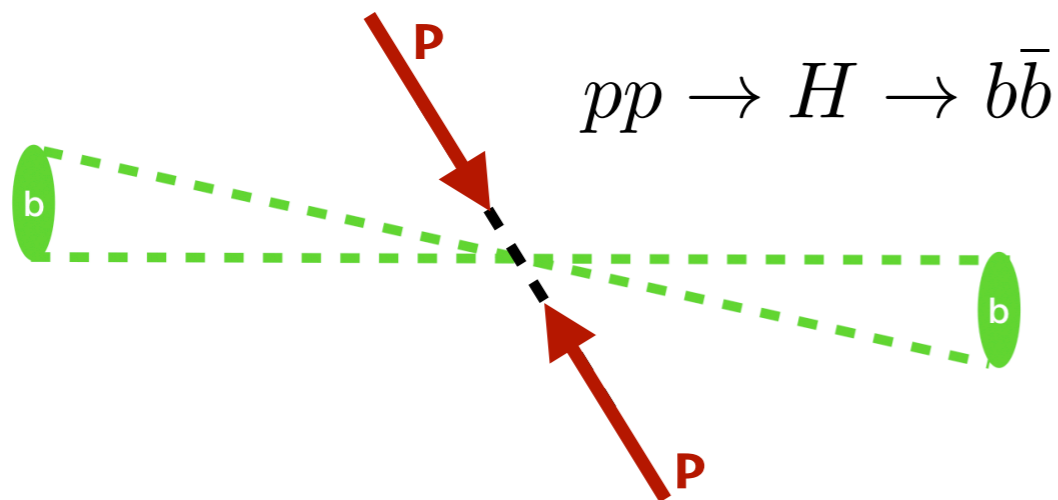


Important:

The **clustering history** of a jet can be viewed as a classical 'proxy' to how the jet formed as a series of hard splittings.

Boosted jet substructure

Boosted: large transverse momentum

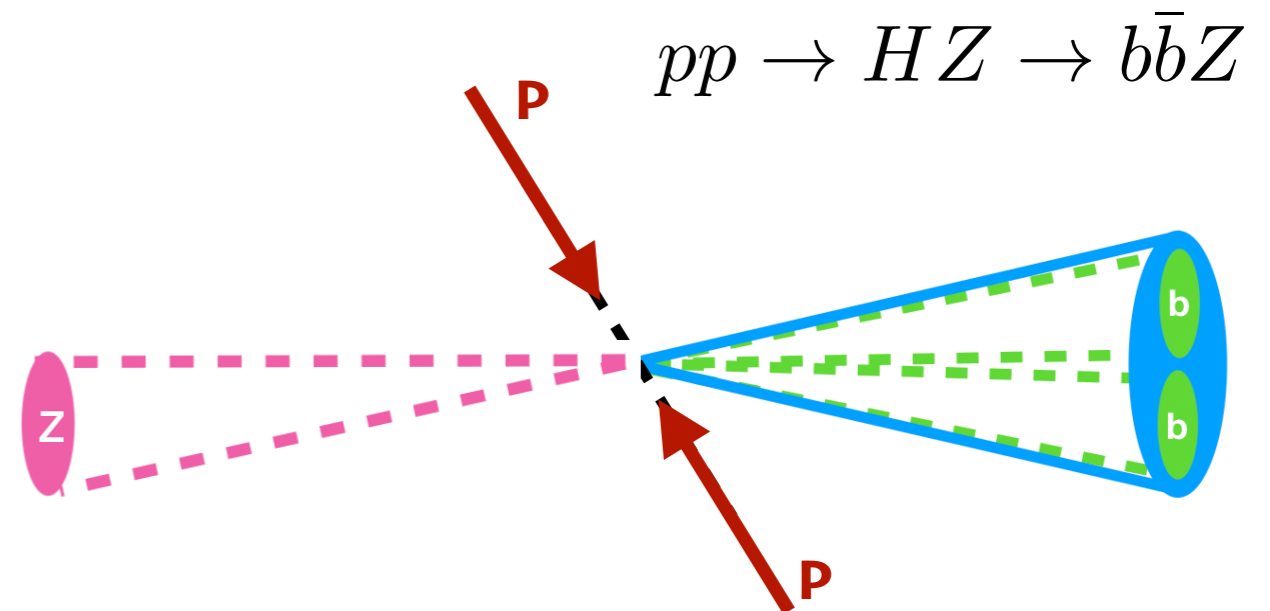


In HZ at high transverse momentum the backgrounds are massively reduced.

A recoiling Higgs can have a boost

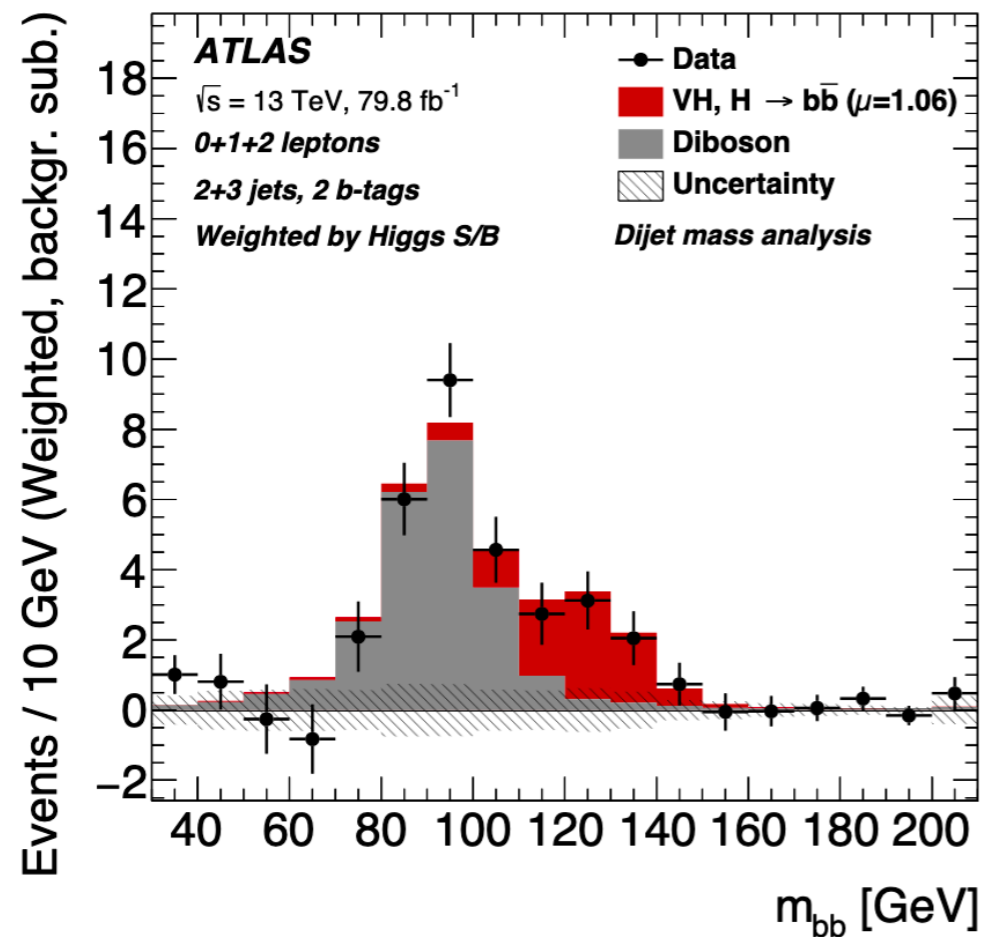
Angular separation of bottoms: $R_{b\bar{b}} \sim \frac{m_h}{p_T}$

Jet substructure techniques used to tag this process. (Butterworth et al 2008)



Boosted jet substructure

Boosted: large transverse momentum

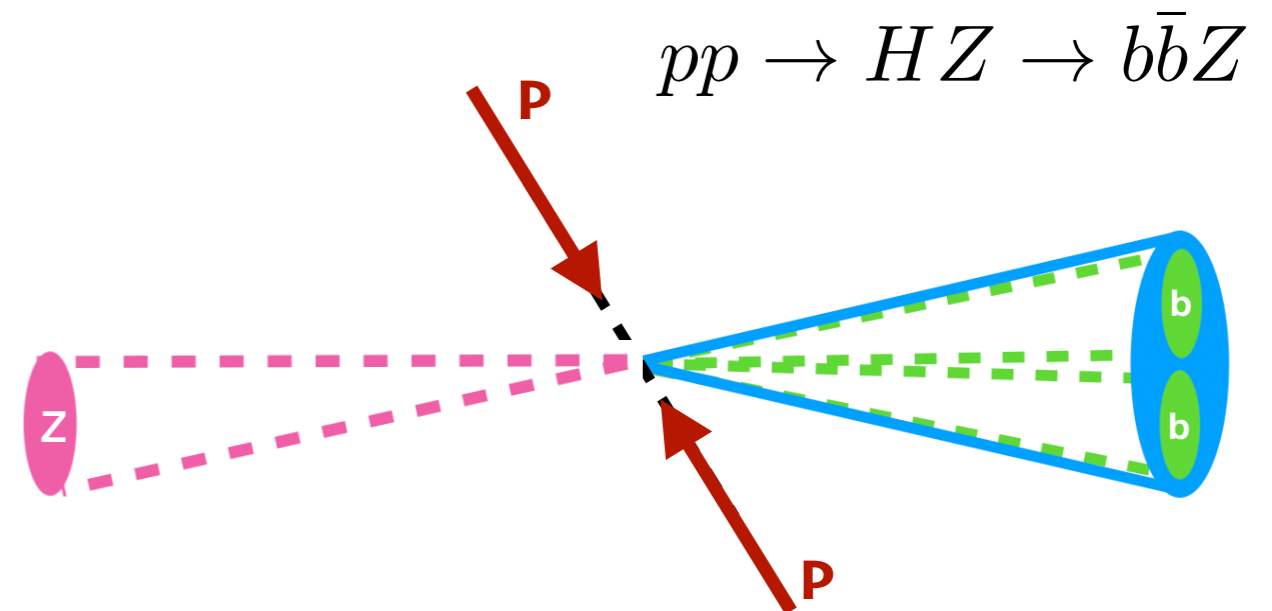


Observed in 2018 by the ATLAS and CMS collaborations! (Phys. Lett. B 786 (2018) 59)

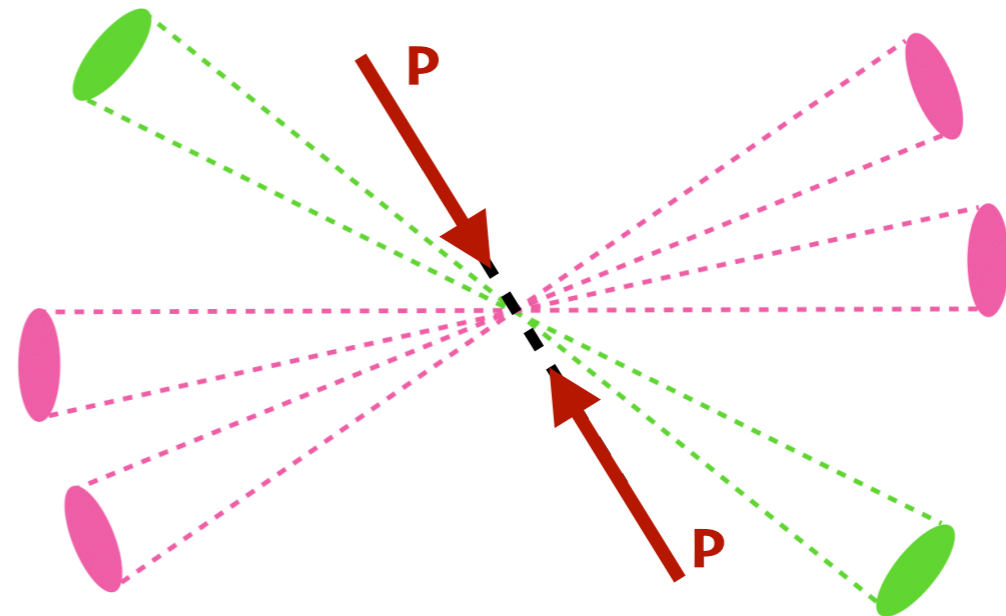
A recoiling Higgs can have a boost

Angular separation of bottoms: $R_{b\bar{b}} \sim \frac{m_h}{p_T}$

Jet substructure techniques used to tag this process. (Butterworth et al 2008)



Top-tagging

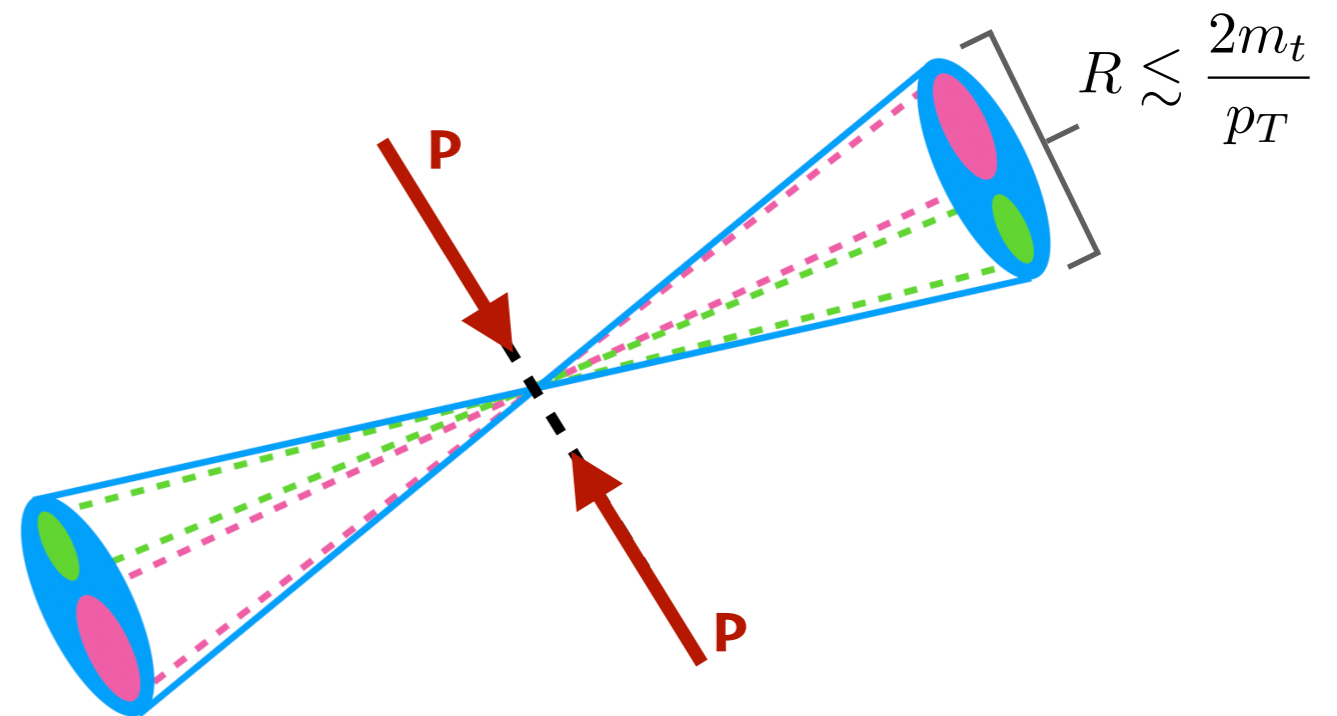


$$pp \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

(Observed first at Tevatron 1995)

At low transverse momentum all decay products can be resolved.

Top-tagging

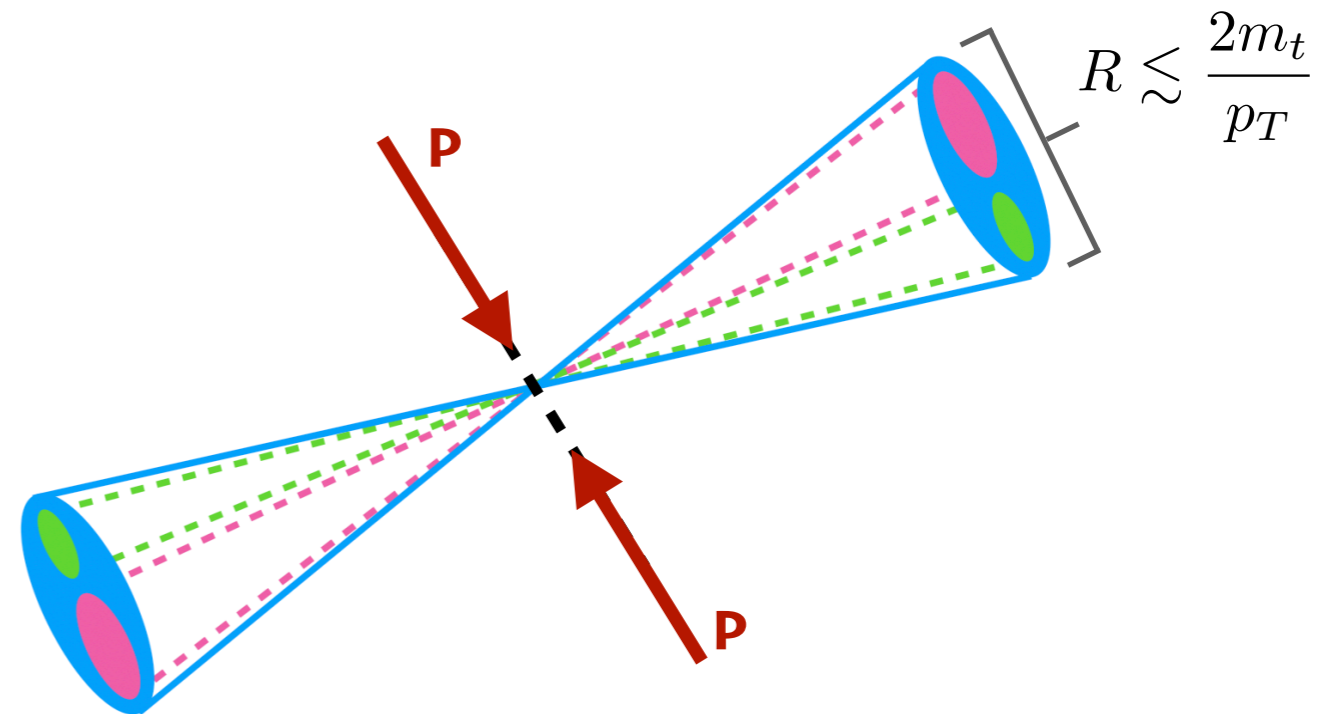
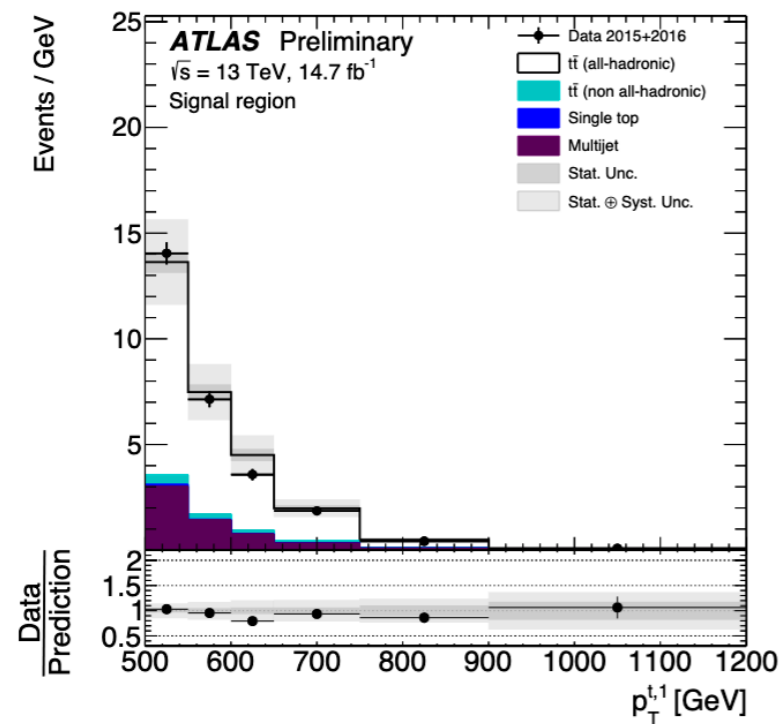
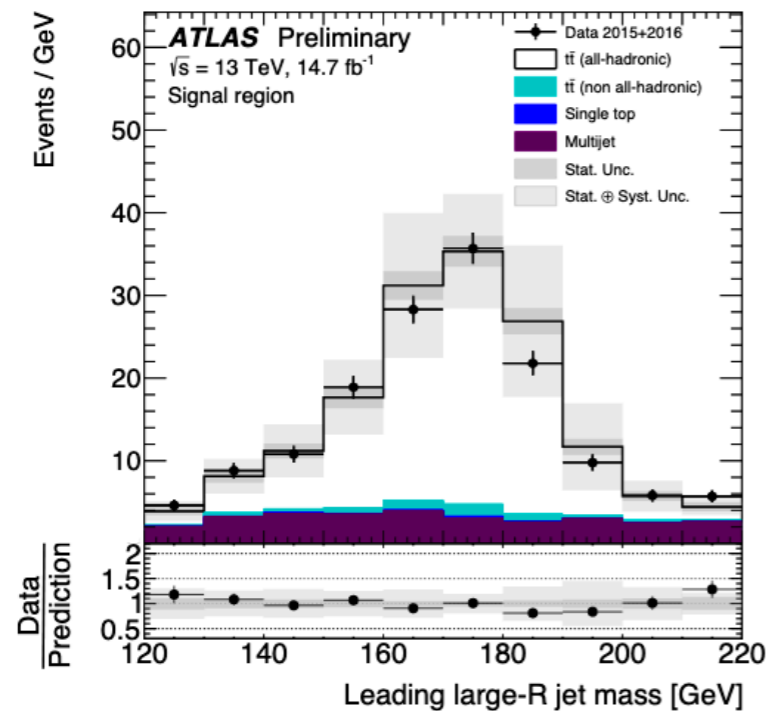


$$pp \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

← (ATLAS, PRD 2018)

Boosted kinematics force all decay products into a single jet.
... lower statistics at large boosts.

Top-tagging



$$pp \rightarrow t\bar{t} \rightarrow (W^+ \bar{b})(W^- b)$$

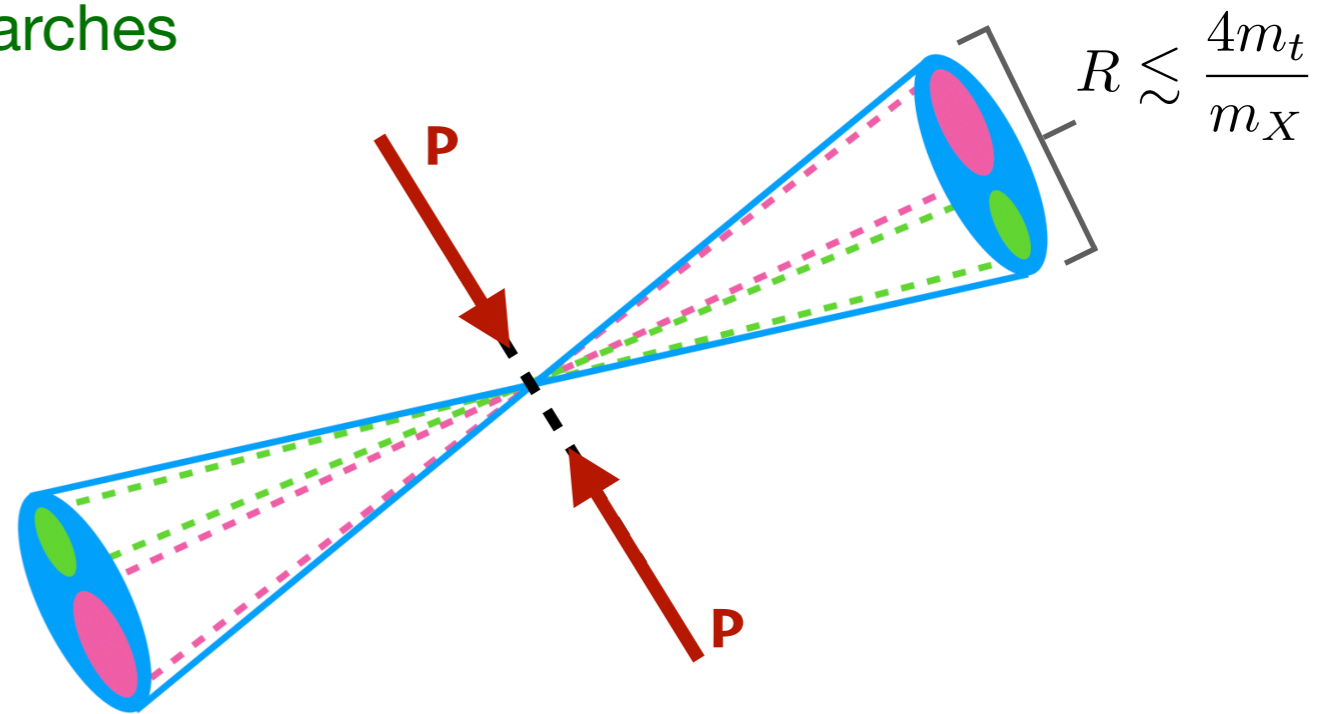
← (ATLAS, PRD 2018)

Boosted kinematics force all decay products into a single jet.
 ... lower statistics at large boosts.

Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

Johns-Hopkins top-tagger

(Kaplan et al, 2008)

- de-cluster the jet, analyse sub-clusters
- keep hard splittings:

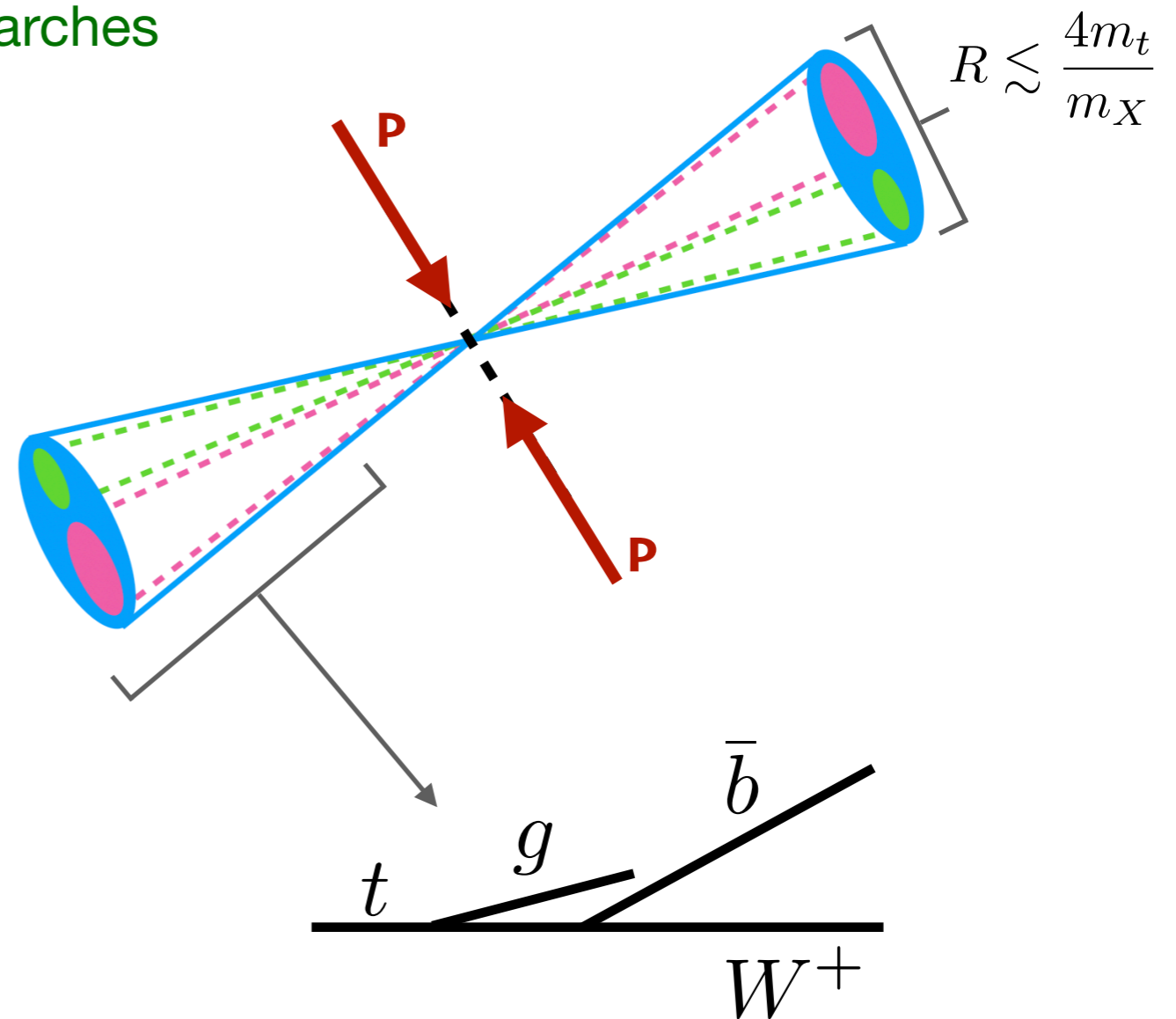
$$\min \frac{p_{T,j_i}}{p_T^{\text{hard}}} > 0.1, \quad R_{j_1,j_2} > 0.19$$

- tag as top if there are sub-clusters with:

$$m_{jjj} = m_t \pm 30 \text{ GeV}$$

$$m_{jj} = m_W \pm 15 \text{ GeV}$$

$$\cos\theta_h < 0.7$$



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

N-subjettiness

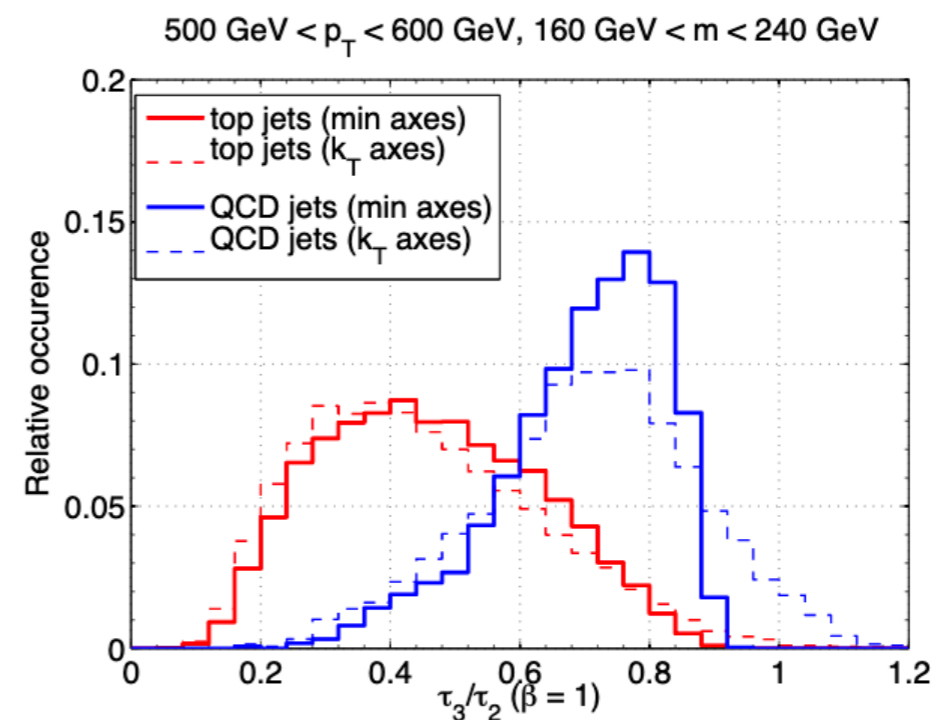
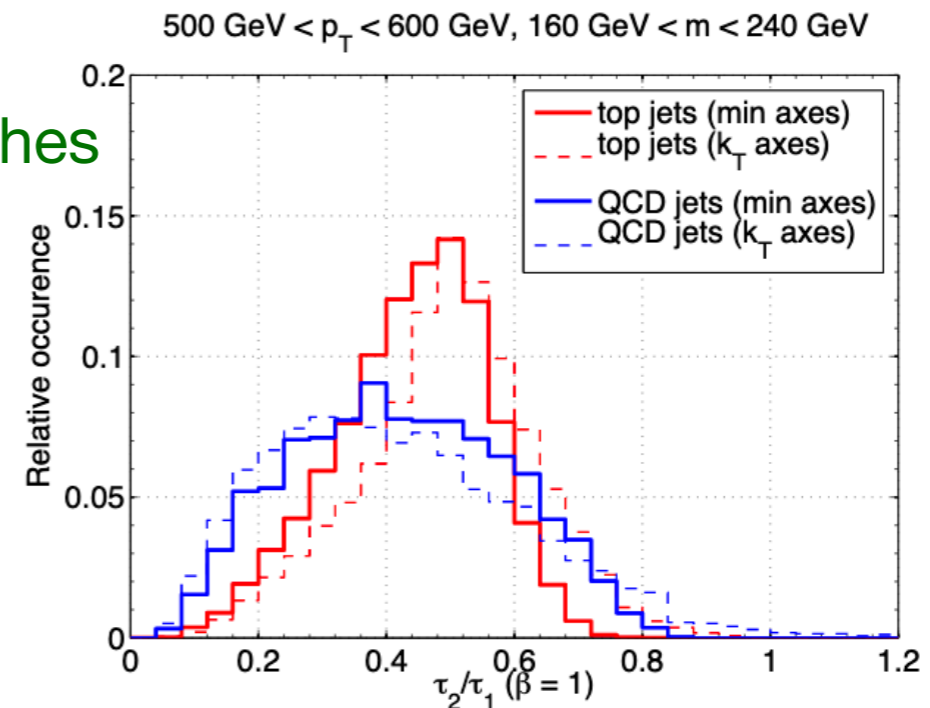
(Thaler et al, 2011)

$$\tau_N = \frac{\sum_{\alpha} p_{T,\alpha} \min_{k=1,\dots,N} R_{k,\alpha}}{\sum_{\alpha} p_{T,\alpha} R_0}$$

Measures distances of particles from N sub-axis, directions of candidate sub-jets.

Sums over particles in the jet.

Ratios indicate ‘pronginess’ of jets.



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

Supervised machine learning

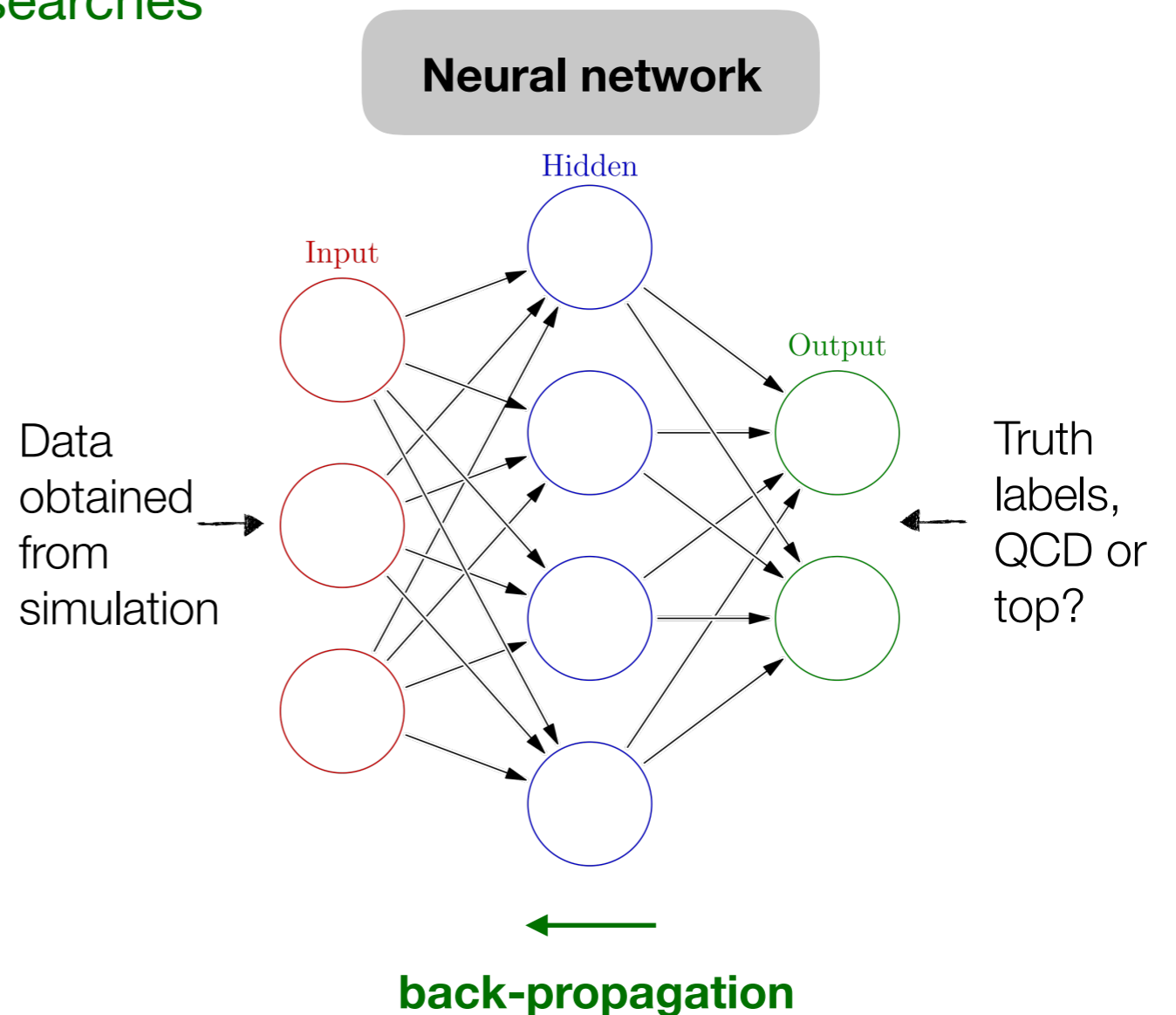
(Machine learning landscape of top-taggers, 2019)

Neural networks trained on Monte-Carlo events to classify between QCD jets and top jets.

Taggers evaluated with ROC curves.

$$\text{ROC}(x) = \epsilon_B^{-1}(\epsilon_S(x))$$

$$\text{AUC}(x) = \int_0^1 dx (1 - \epsilon_B(\epsilon_S(x)))$$



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

Supervised machine learning

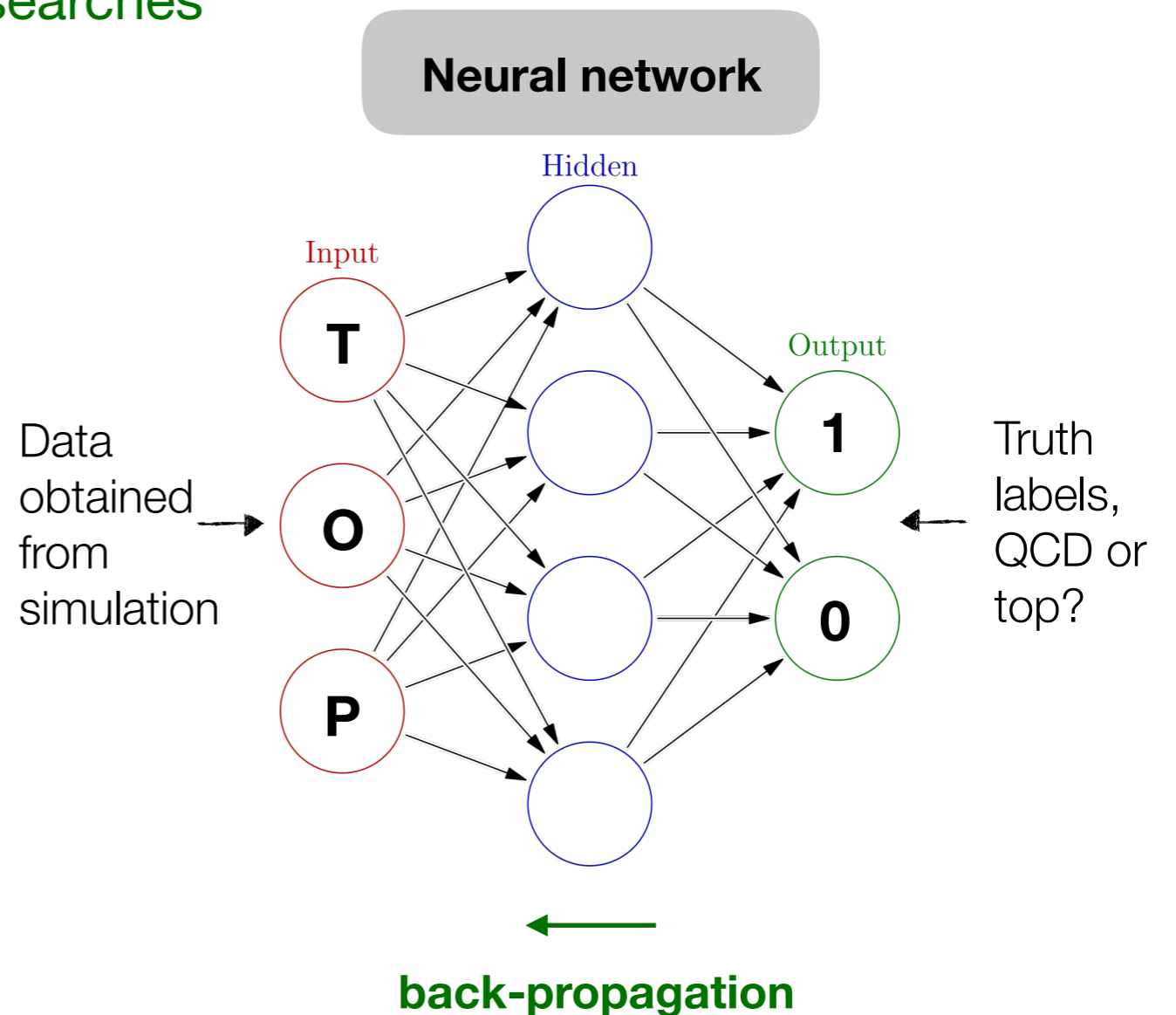
(Machine learning landscape of top-taggers, 2019)

Neural networks trained on Monte-Carlo events to classify between QCD jets and top jets.

Taggers evaluated with ROC curves.

$$\text{ROC}(x) = \epsilon_B^{-1}(\epsilon_S(x))$$

$$\text{AUC}(x) = \int_0^1 dx (1 - \epsilon_B(\epsilon_S(x)))$$



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

Supervised machine learning

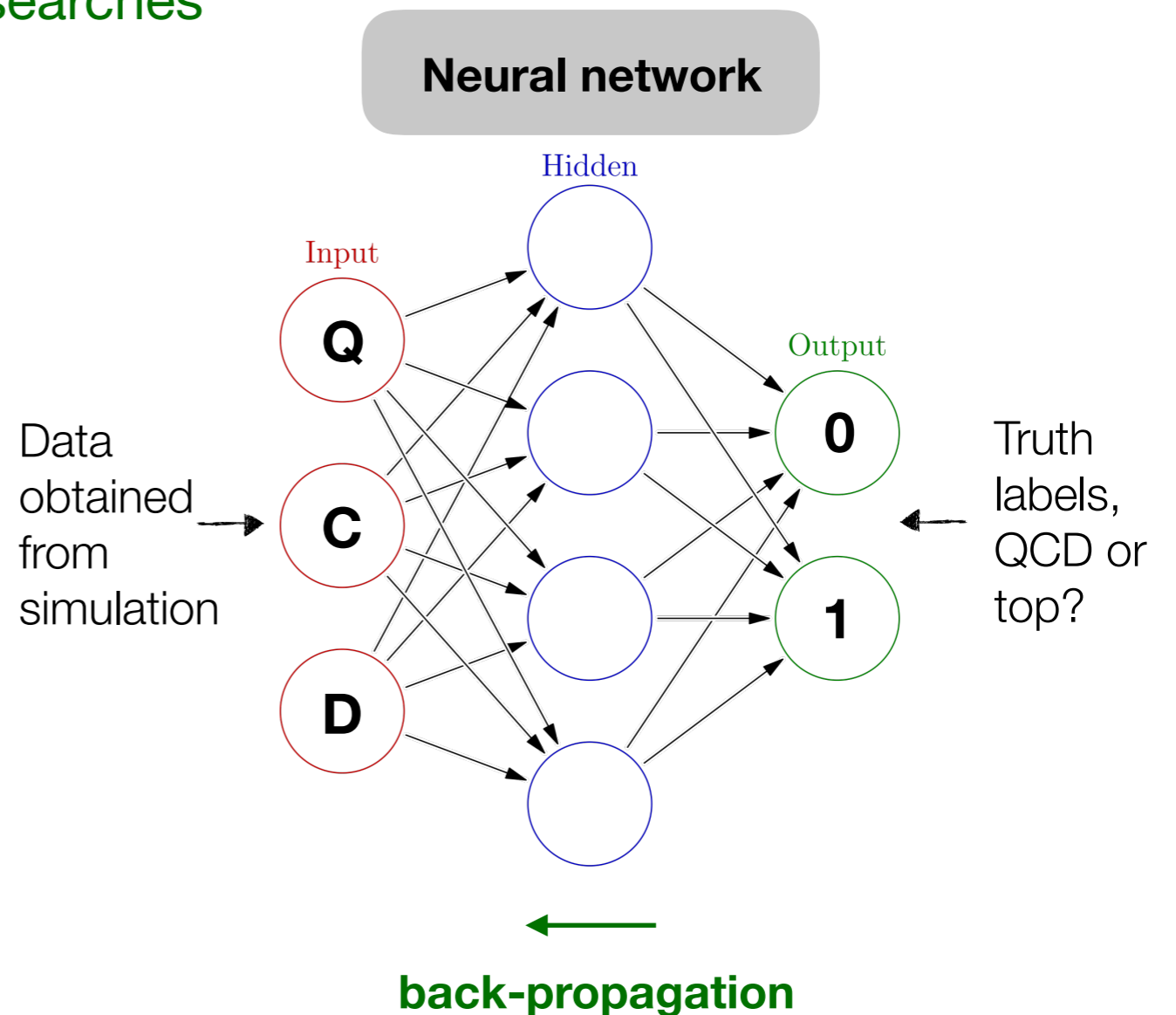
(Machine learning landscape of top-taggers, 2019)

Neural networks trained on Monte-Carlo events to classify between QCD jets and top jets.

Taggers evaluated with ROC curves.

$$\text{ROC}(x) = \epsilon_B^{-1}(\epsilon_S(x))$$

$$\text{AUC}(x) = \int_0^1 dx (1 - \epsilon_B(\epsilon_S(x)))$$



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

Supervised machine learning

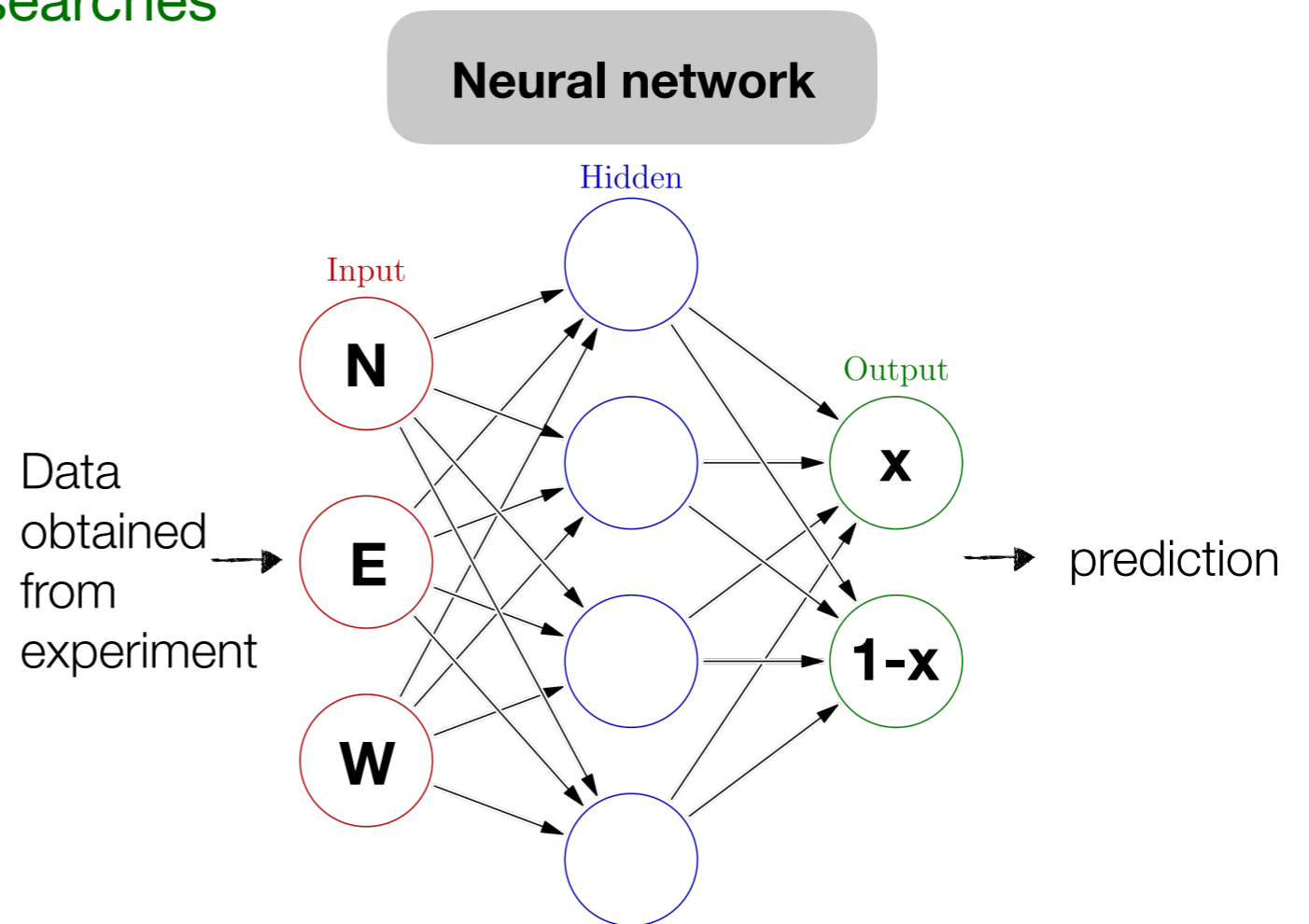
(Machine learning landscape of top-taggers, 2019)

Neural networks trained on Monte-Carlo events to classify between QCD jets and top jets.

Taggers evaluated with ROC curves.

$$\text{ROC}(x) = \epsilon_B^{-1}(\epsilon_S(x))$$

$$\text{AUC}(x) = \int_0^1 dx (1 - \epsilon_B(\epsilon_S(x)))$$



Top-tagging

Pair-production of top quarks in NP searches

$$pp \rightarrow X \rightarrow t\bar{t} \rightarrow (W^+\bar{b})(W^-b)$$

Supervised machine learning

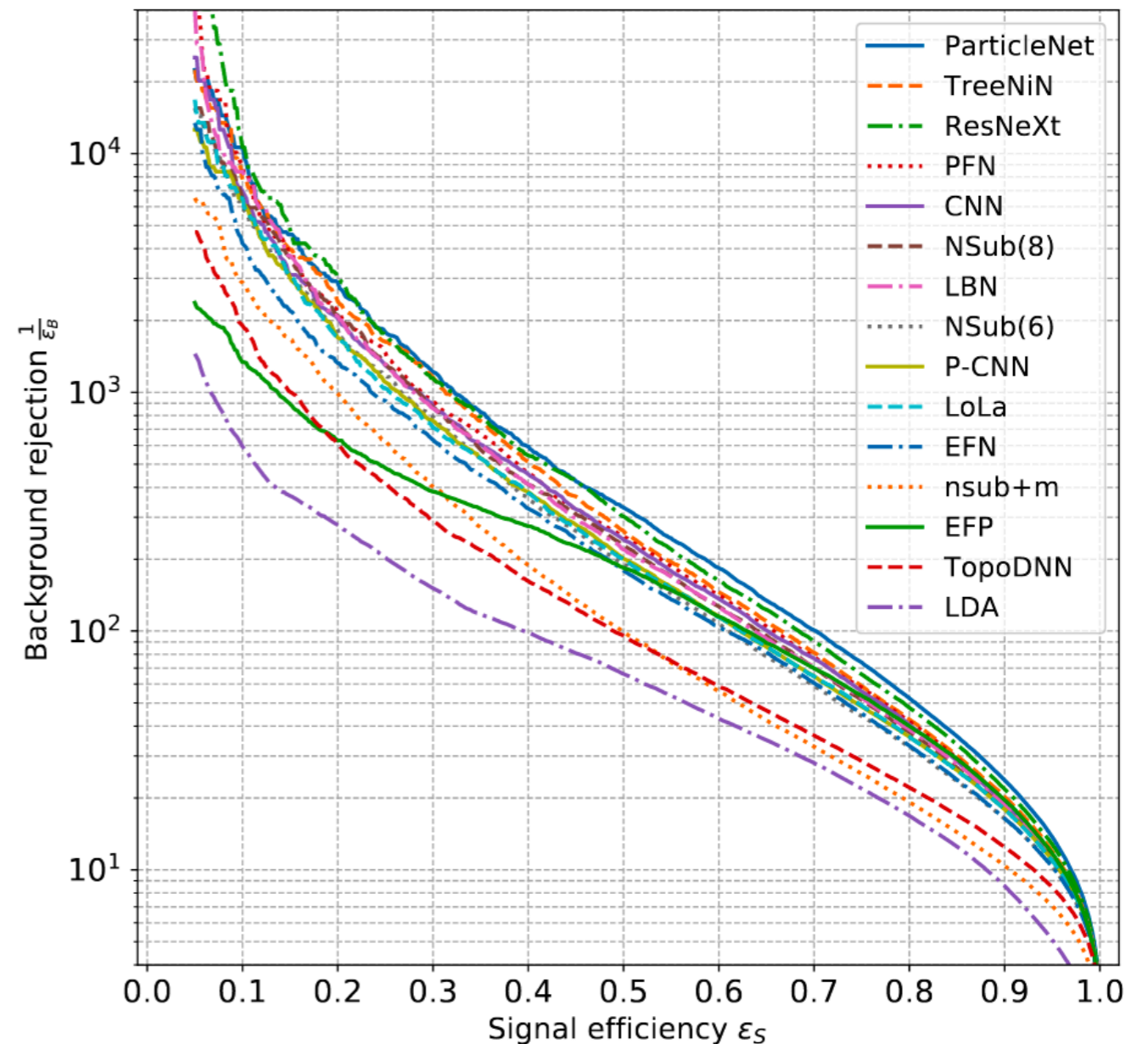
(Machine learning landscape of top-taggers, 2019)

Neural networks trained on Monte-Carlo events to classify between QCD jets and top jets.

Taggers evaluated with ROC curves.

$$\text{ROC}(x) = \epsilon_B^{-1}(\epsilon_S(x))$$

$$\text{AUC}(x) = \int_0^1 dx (1 - \epsilon_B(\epsilon_S(x)))$$



Aim of this work

Aim of the work

Tagging \longrightarrow samples 'enriched' with signal jets

They do so with knowledge of top quark decay channels and kinematics.

(also: b-tagging, tau-tagging, W/Z tagging quark vs. gluon tagging
With: BDTs, neural networks, jet images, energy flow polynomials, de-mix, ...)

Aim of the work

Tagging \longrightarrow samples 'enriched' with signal jets

They do so with knowledge of top quark decay channels and kinematics.

(also: b-tagging, tau-tagging, W/Z tagging quark vs. gluon tagging
With: BDTs, neural networks, jet images, energy flow polynomials, de-mix, ...)

Our aim is to do this without knowing anything about the top quark.

We need two things:

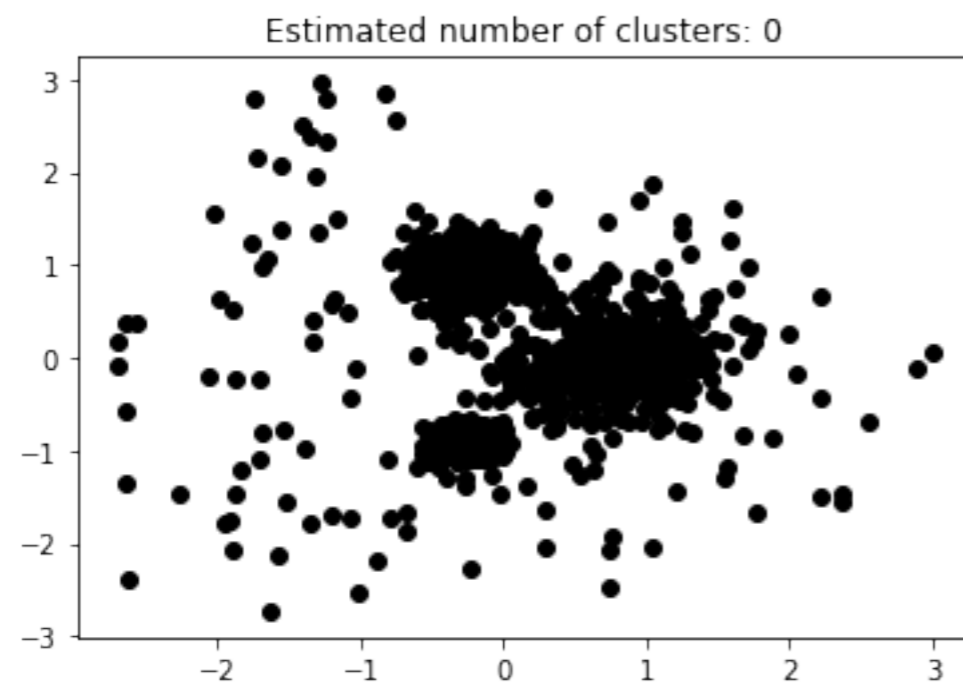
- 1 - knowledge of Quantum Field Theory
- 2 - machine learning tools

**Unsupervised
classification of jets**

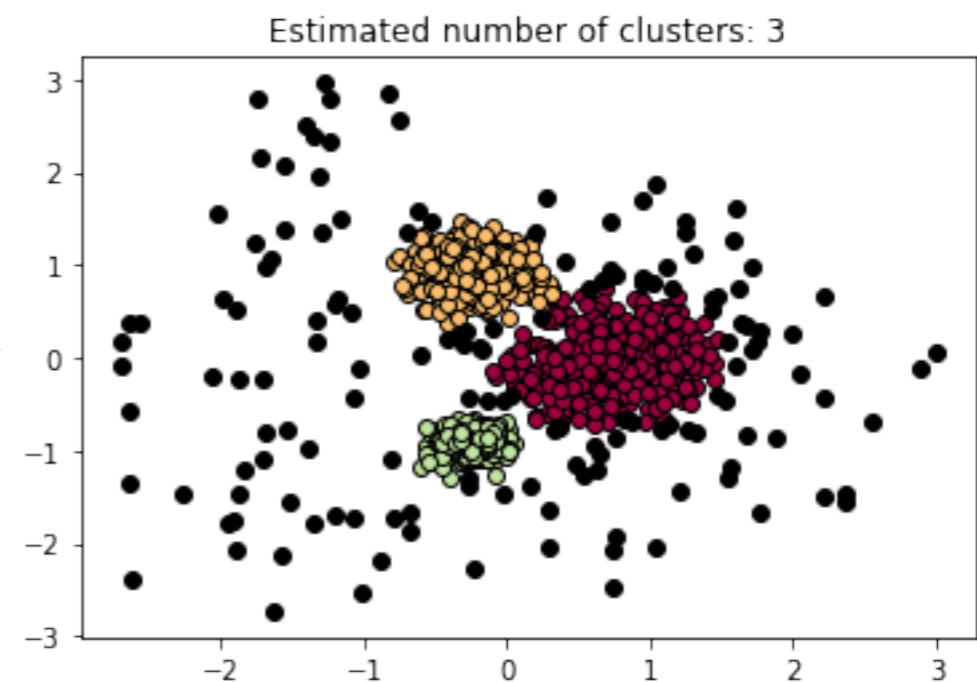
Unsupervised machine learning

Unsupervised learning: an algorithm that helps find previously unknown patterns in a data set **without pre-existing labels**.

Simplest example: clustering algorithms.



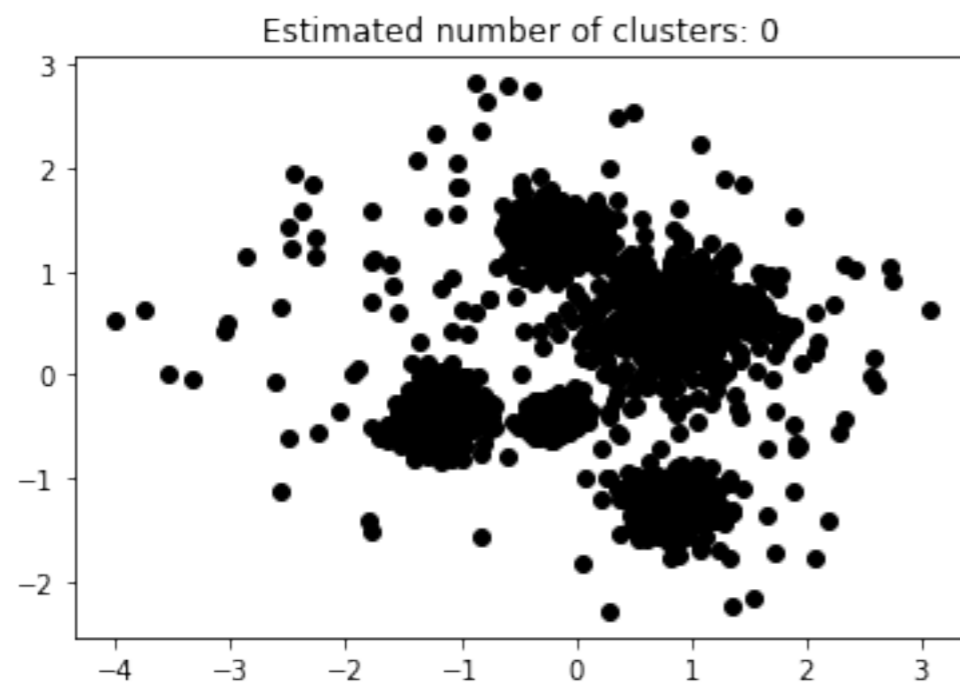
DBSCAN



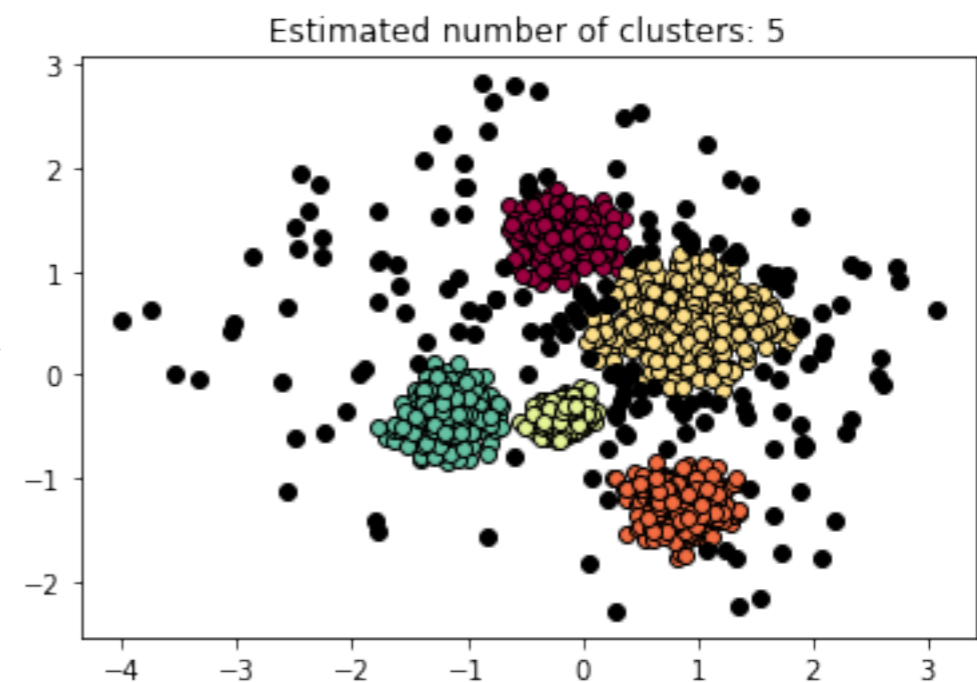
Unsupervised machine learning

Unsupervised learning: an algorithm that helps find previously unknown patterns in a data set **without pre-existing labels**.

Simplest example: **clustering algorithms**.



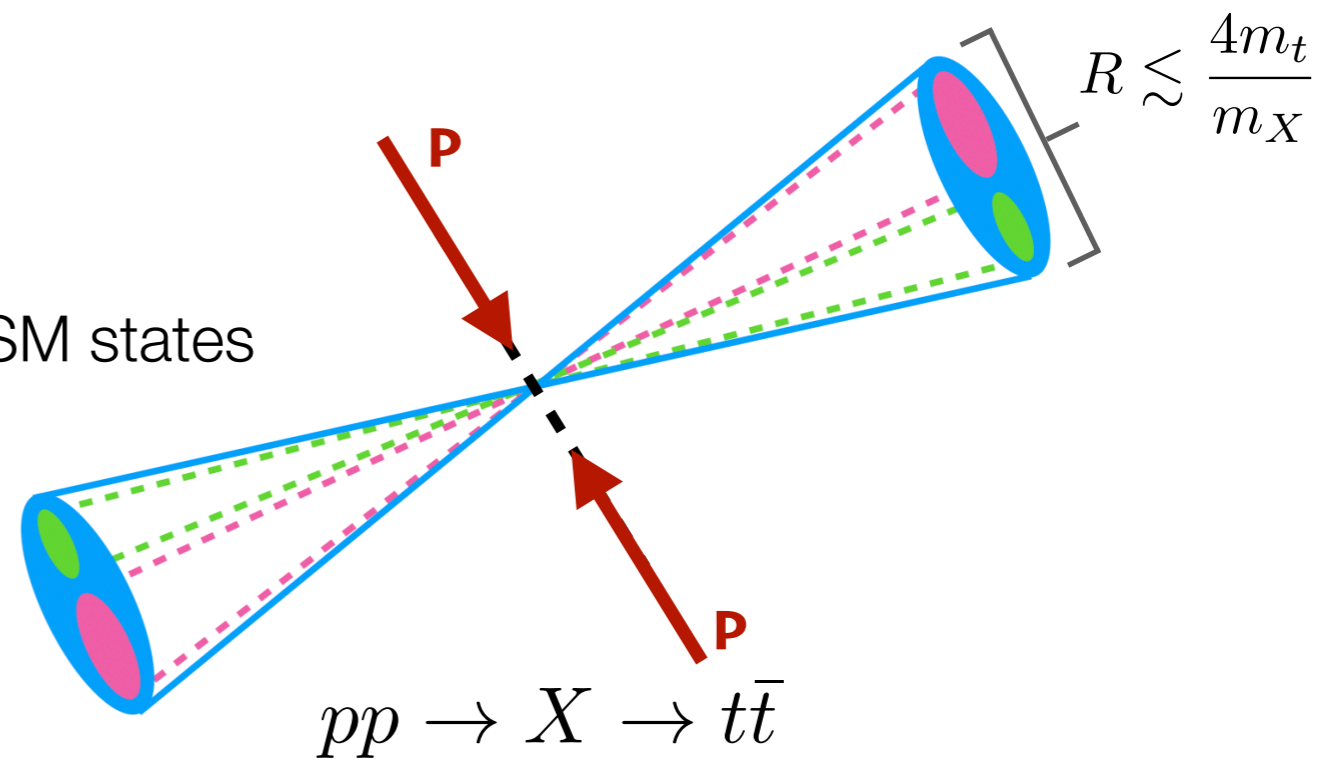
DBSCAN
→



Why unsupervised?

Supervised algorithms are useful when:

- measuring SM processes
- searching for NP decaying directly to SM states



Not so useful when the NP gives rise to a non-trivial jet substructure.

In this case we need unsupervised algorithms.

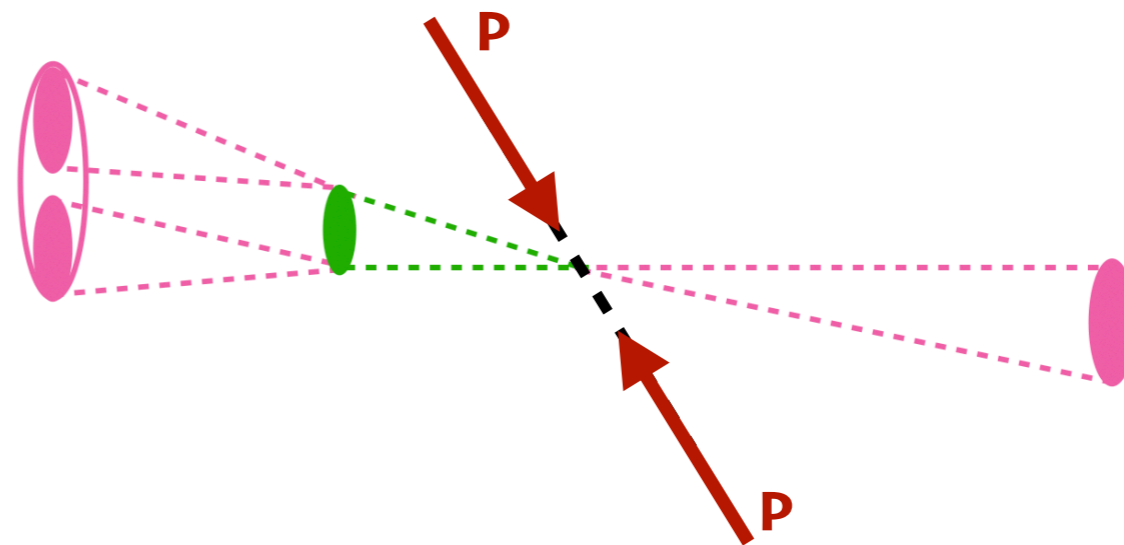
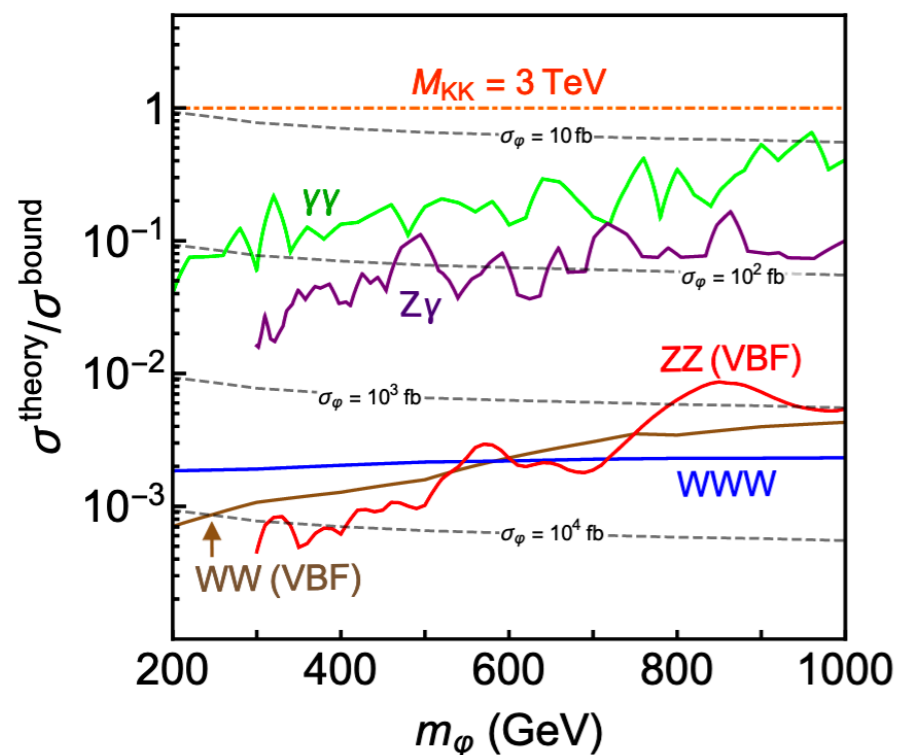
Why unsupervised?

For example: **'stealth bosons'** (Aguilar-Saavedra 2017, Agashe et al. 2018)

$$pp \rightarrow W' \rightarrow \phi W \rightarrow WWW$$

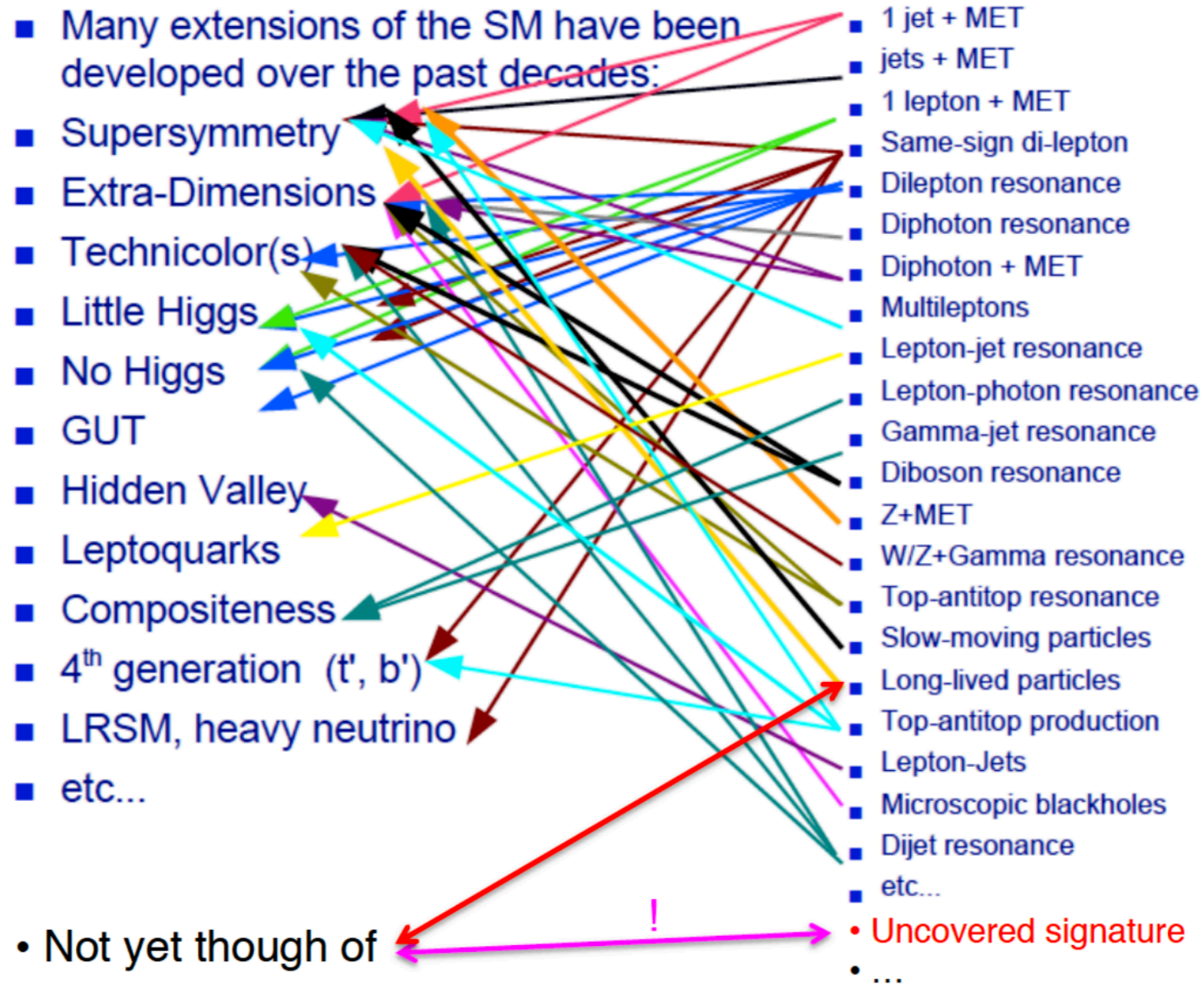
$$m_{W'} = 3 \text{ TeV}$$

$$m_\phi = 400 \text{ GeV}$$



- Scenarios arise in Randall-Sundrum type models.
- W' is a Kaluza-Klein mode, the scalar the radion.
- **Signal easily found with a dedicated tagger.**

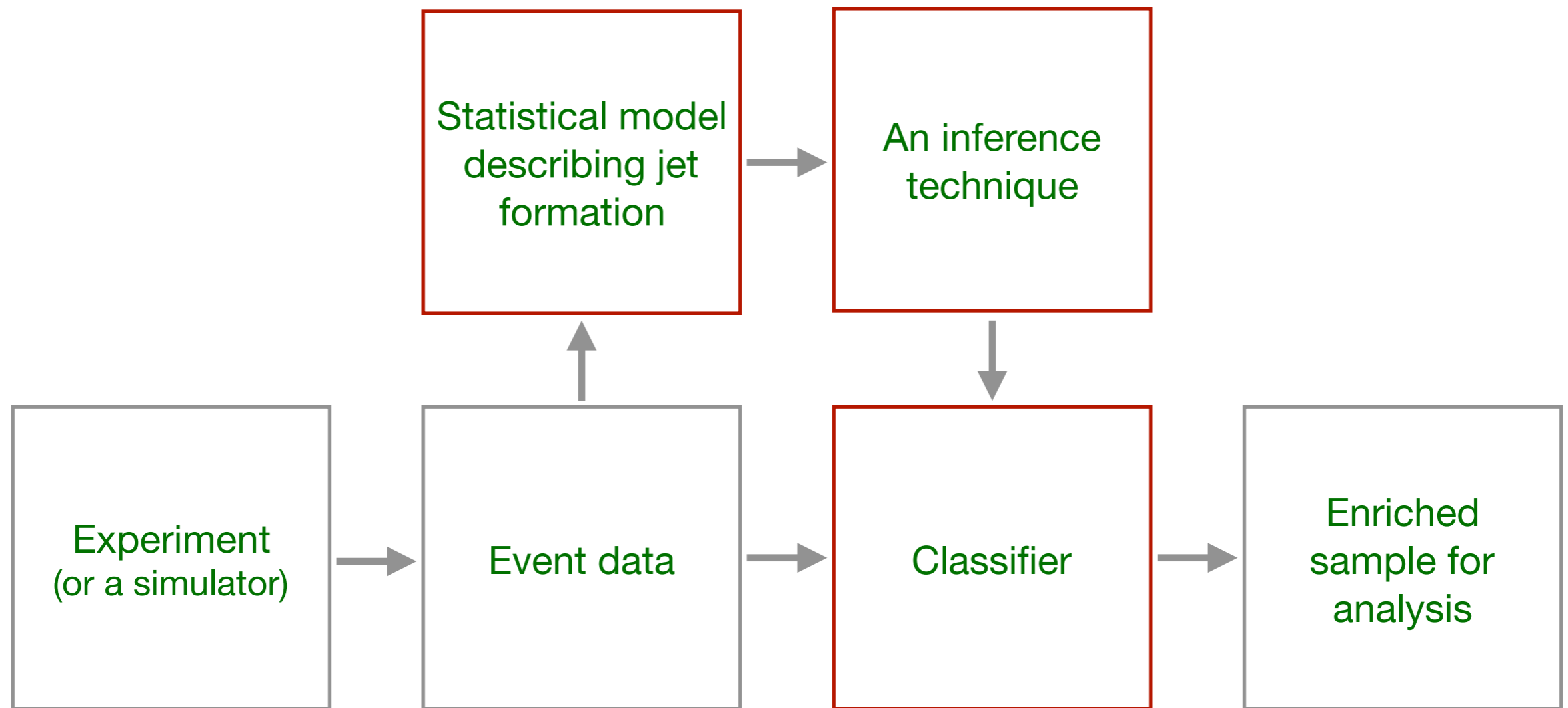
Why unsupervised?



(Juste, ATLAS talk, 2017)

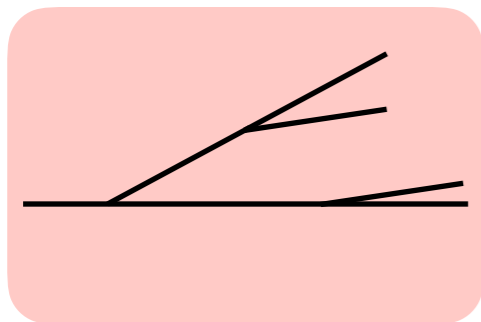
Uncovering latent jet substructure

The basic idea



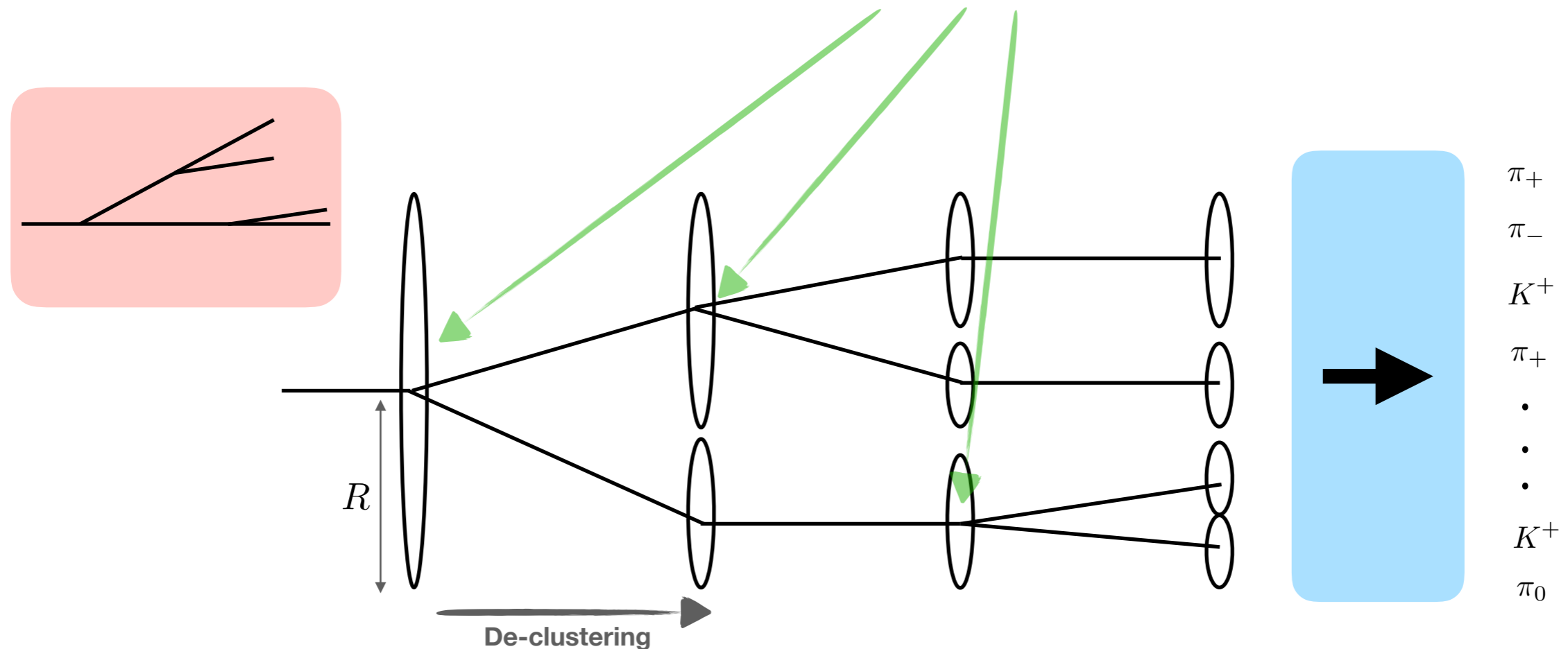
Modelling a jet

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$



Modelling a jet

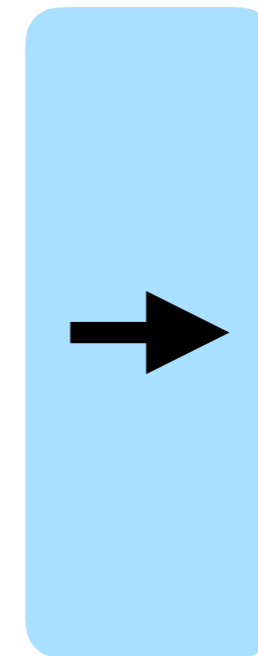
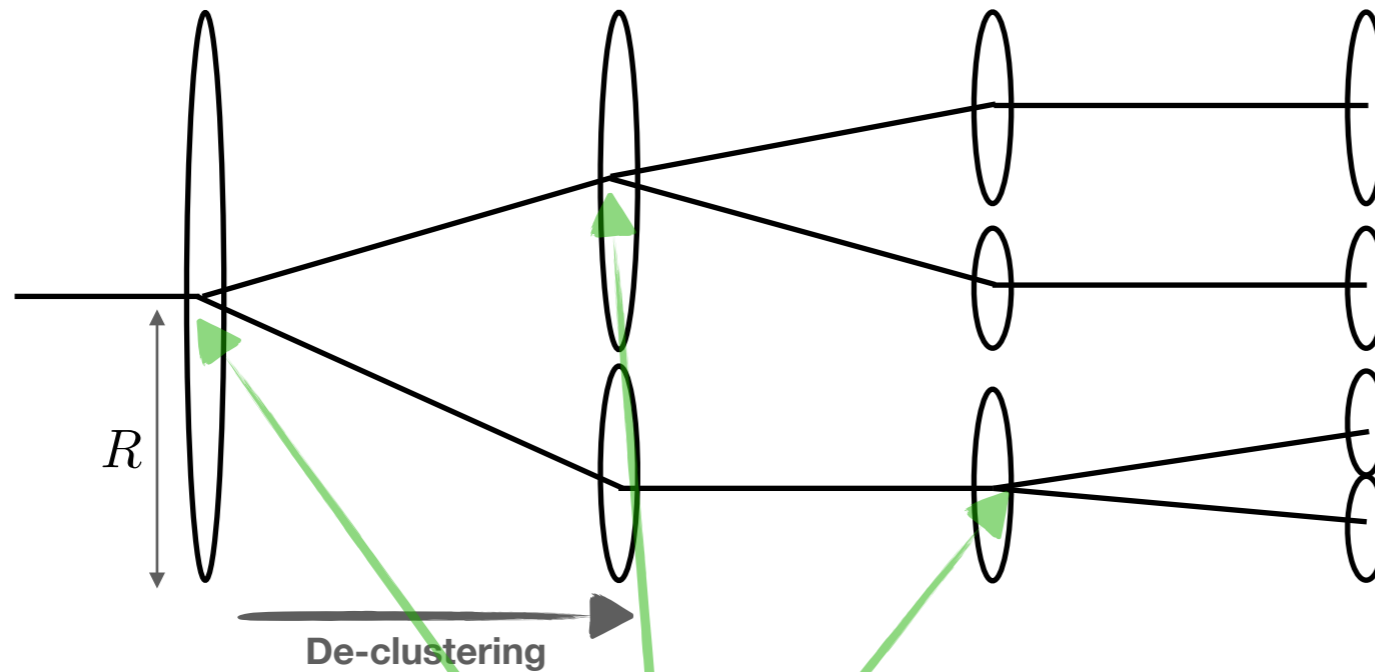
An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$



How do we represent the splittings?

Modelling a jet

$$e_j = \{f_1, f_2, \dots, f_n\}$$



π_+
 π_-
 K^+
 π_+
 \cdot
 \cdot
 \cdot
 K^+
 π_0

Mass representation

Each splitting represents: $j_0 \rightarrow j_1, j_2 \quad m_{j_1} > m_{j_2}$

Intuitive to define splittings as: $f_i = [m_{j_0}, \frac{m_{j_1}}{m_{j_0}}]$
 [subjct mass, mass drop]

Observables are binned.
 No concept of distance
 between splittings.

The Lund plane

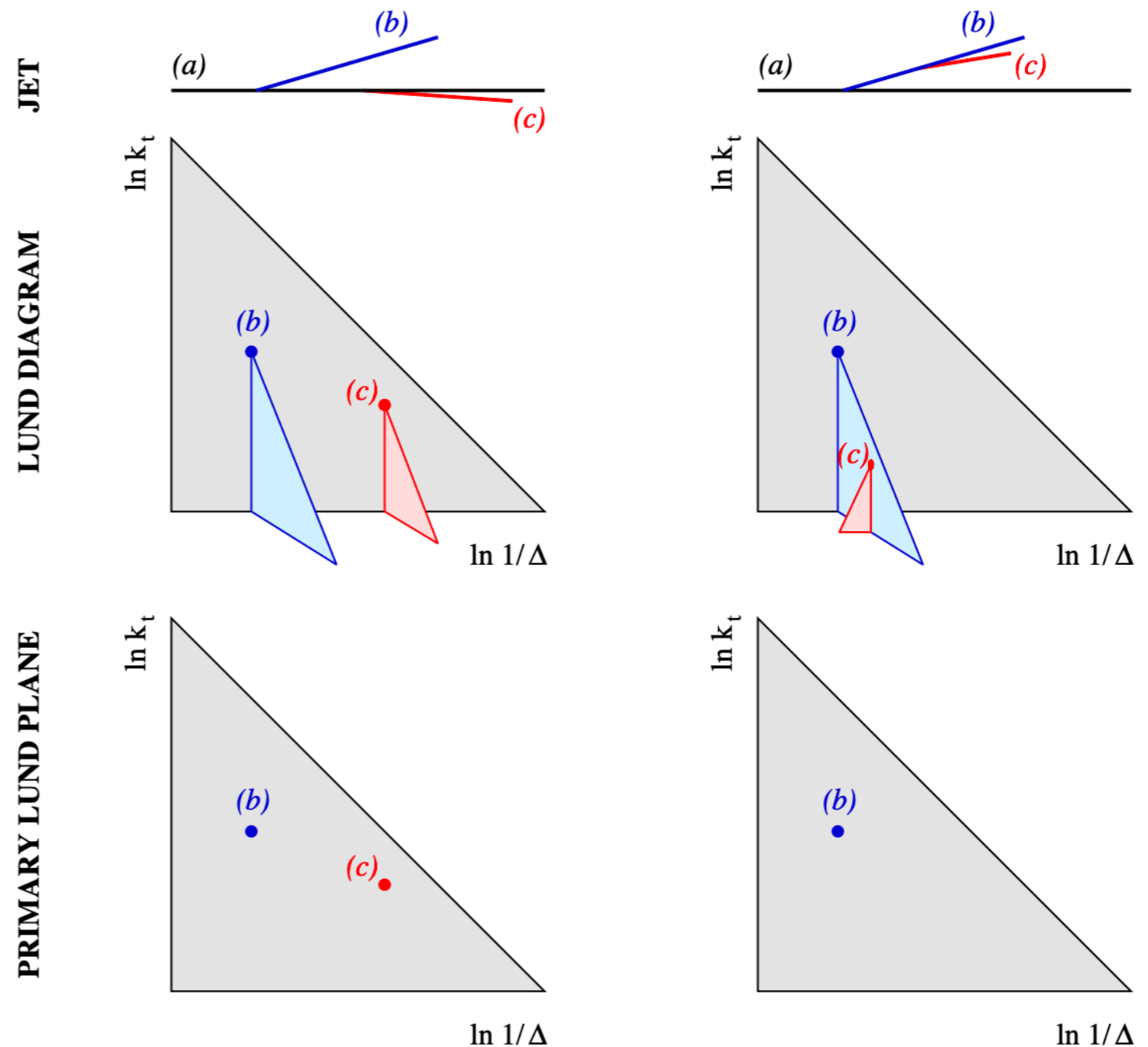
Different kinematic regions are naturally separated into soft, collinear, and large angle splittings.

$$\Delta = R_{12}$$

$$k_T = p_{T,2} R_{12}$$

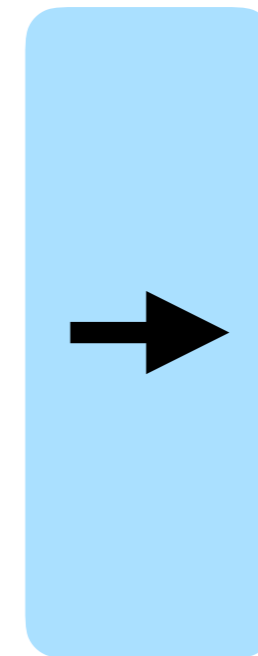
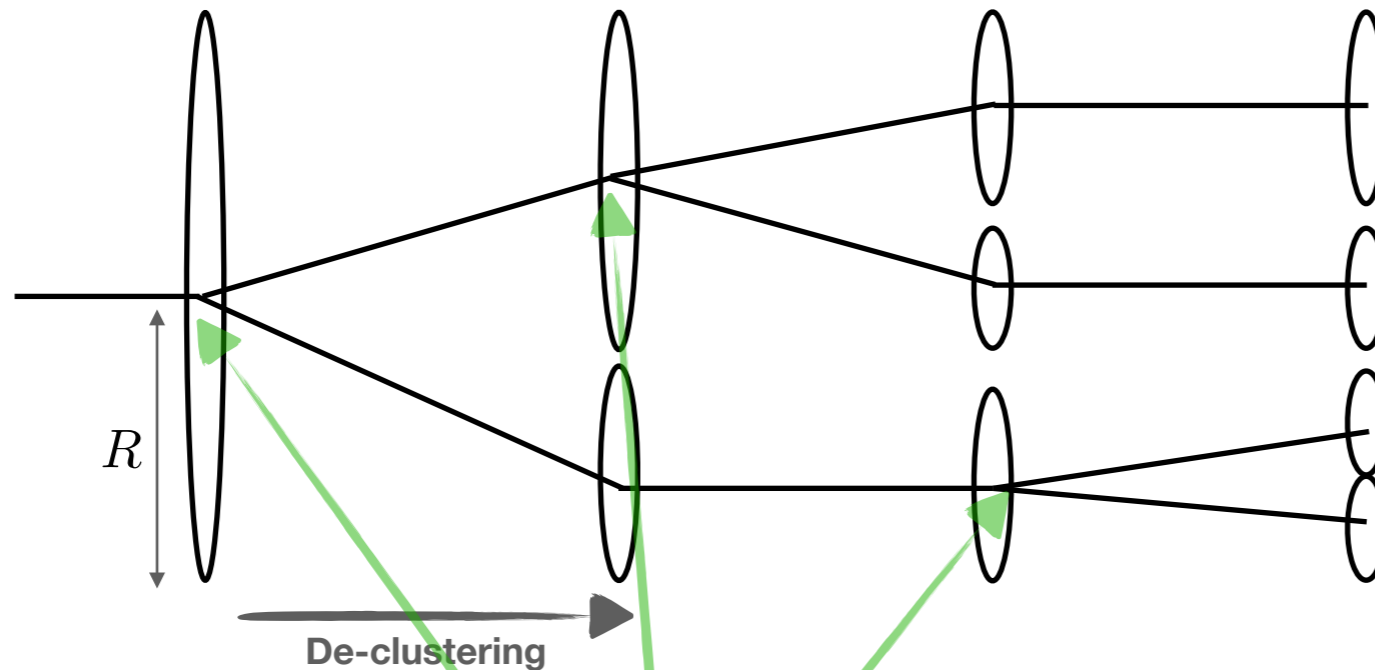
(Dreyer et al, JHEP 2018)

Describes the **radiation pattern** inside the jet.



Modelling a jet

$$e_j = \{f_1, f_2, \dots, f_n\}$$



π_+
 π_-
 K^+
 π_+
 \cdot
 \cdot
 \cdot
 K^+
 π_0

Lund representation

Each splitting represents: $j_0 \rightarrow j_1, j_2 \quad m_{j_1} > m_{j_2}$

Lund definition of splittings as: $f_i = [R_{12}, k_T]$

[angular splitting, kT splitting]

Observables are binned.
 No concept of distance
 between splittings.

The simplest model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

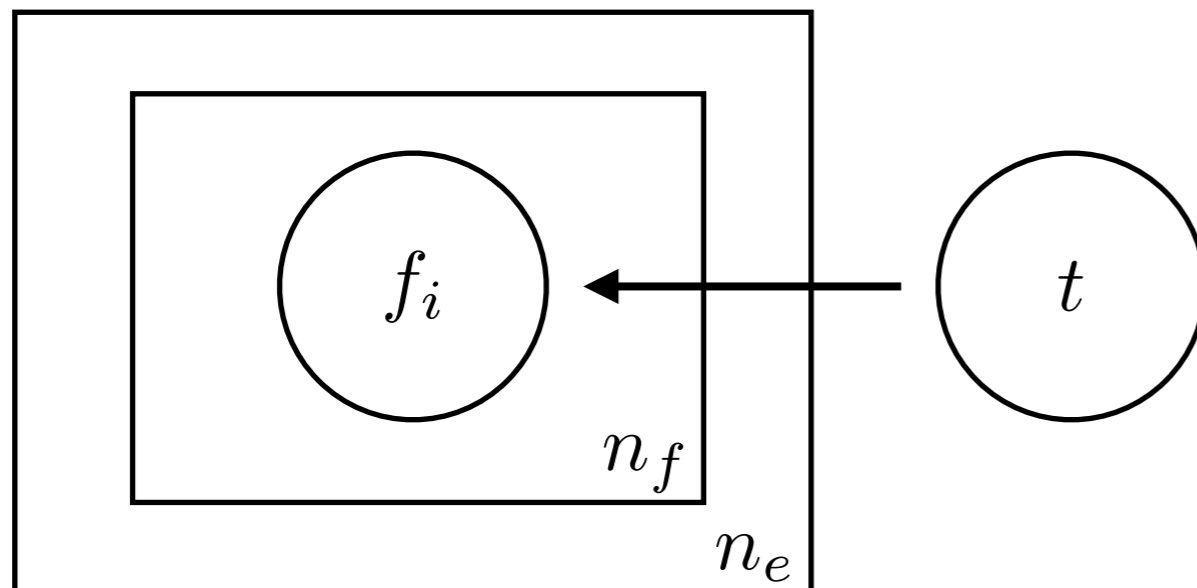
De Finetti's theorem: $P(e) = P(f_1, \dots, f_{n_f}) = \prod_{i=1}^{n_f} P(f_i|t)$

The simplest model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

De Finetti's theorem: $P(e) = P(f_1, \dots, f_{n_f}) = \prod_{i=1}^{n_f} P(f_i|t)$

Generative process:



Draw f_i randomly from $P(f|t)$

Repeat n_f times for the jet

Repeat for each of the n_e jets

Everything generated from a single distribution/process

A mixture model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by one of K processes.

$$P(e|\theta, t_z) = \sum_{z=1}^K \theta(z) \prod_{i=1}^{n_f} P(f_i|t_z)$$

A mixture model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by one of K processes.

$$P(e|\theta, t_z) = \sum_{z=1}^K \theta(z) \prod_{i=1}^{n_f} P(f_i|t_z)$$

Generative process:

Draw z randomly from $\theta(z)$

Draw f_i randomly from the appropriate $P(f_i|t_z)$

Repeat n_f times for the jet

Repeat for each of the n_e jets

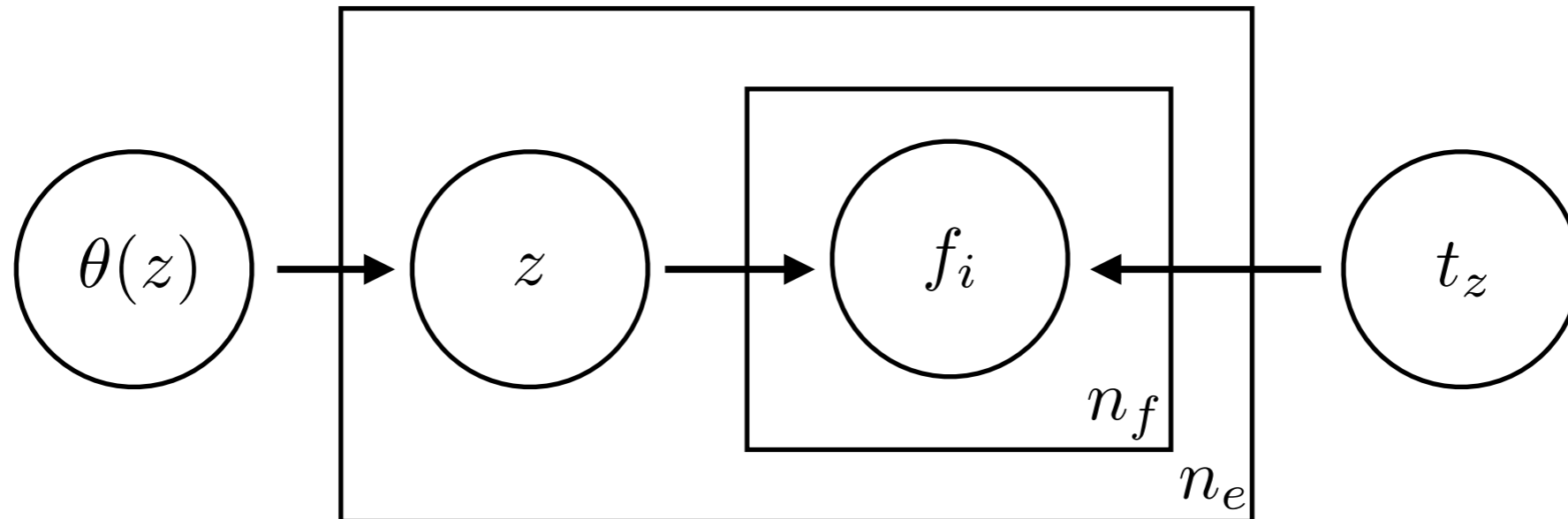
A mixture model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by one of K processes.

$$P(e|\theta, t_z) = \sum_{z=1}^K \theta(z) \prod_{i=1}^{n_f} P(f_i|t_z)$$

Generative process:



A mixed-membership model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by **a mixture of K processes**.

This is the most natural model for jet physics.

Each top jet is composed of:

- a series of hard splittings related to the underlying decay (e.g. $t \rightarrow Wb$)
- a series of soft QCD splittings independent of the hard process.

A mixed-membership model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by **a mixture of K processes**.

This is the most natural model for jet physics.

Generative process:

Draw $\theta(z)$ randomly from the prior $\pi(\theta)$

Draw z randomly from $\theta(z)$

Draw f_i randomly from the appropriate $P(f_i|t_z)$

Repeat n_f times for the jet

Repeat for each of the n_e jets

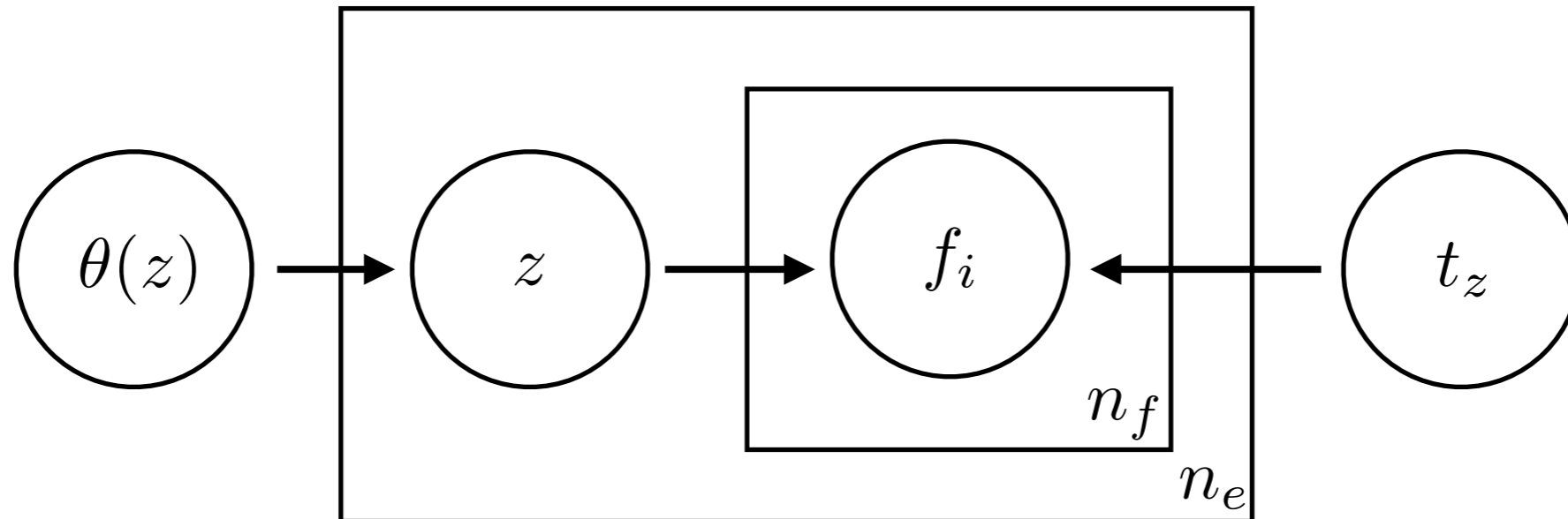
A mixed-membership model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by **a mixture of K processes**.

This is the most natural model for jet physics.

Generative process:



A mixed-membership model

An event/jet = a list of splittings, i.e. $e_j = \{f_1, f_2, \dots, f_n\}$

Let's assume each jet is generated by **a mixture of K processes**.

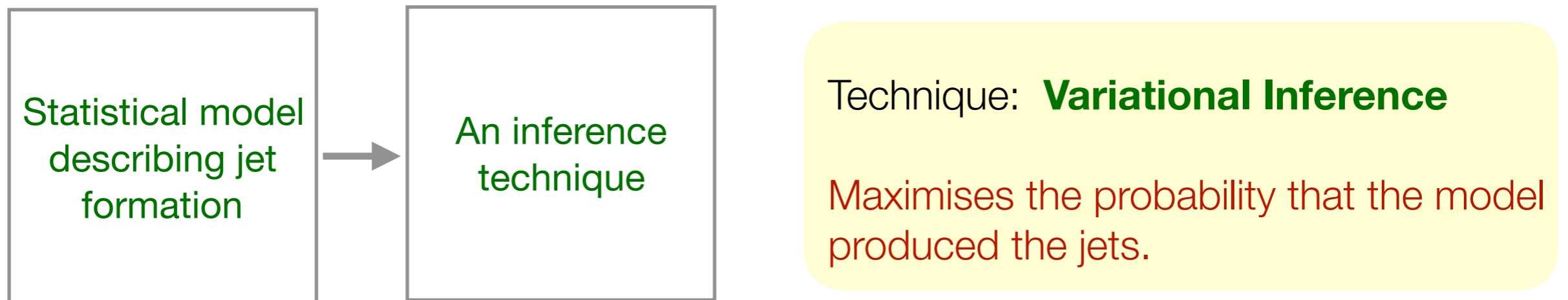
This is the most natural model for jet physics.

The probability of a single jet to be generated from the model is given by:

$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$

This model is called **Latent Dirichlet Allocation (LDA)**.

Inference



$$P(\mathcal{D}|\beta) \quad P(\beta|\mathcal{D}) = \frac{P(\mathcal{D}|\beta)P(\beta)}{P(\mathcal{D})}$$

Uses: **co-occurring** observables within the event.

Software: **gensim** (Radim Řehůřek and Petr Sojka, 2010)

Gibbs sampling can also be used, but is slower and does not typically yield superior results.

Inference

Variational inference is similar to the simpler Expectation-Maximisation (EM) algorithm:

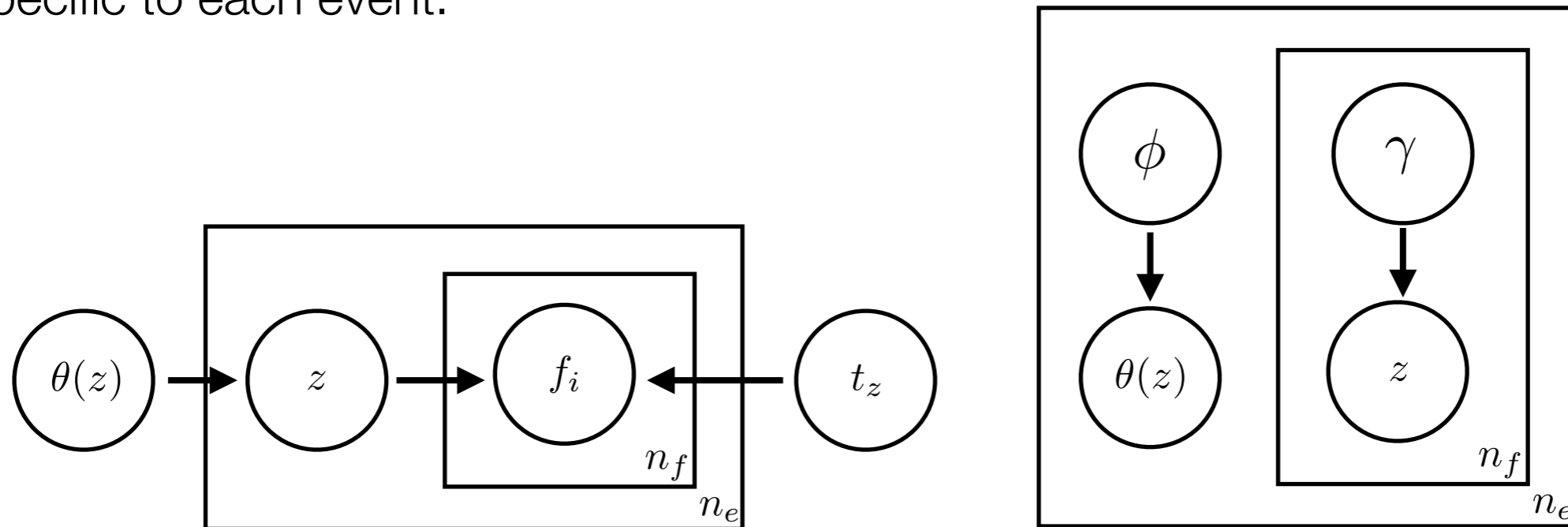
1. Randomly initialise model parameters
2. Use the **model parameters** to estimate the **latent parameters**
3. Use the **latent parameters** to estimate the **model parameters**
4. Calculate the likelihood
5. Repeat steps 2-4 until the likelihood is within some threshold

This works well when the model is relatively simple, for example in a Gaussian mixture model.

It does not work for the mixed-membership models discussed previously. In this case we need approximate inference methods such as variational inference.

Inference

Variational inference relies on using a simplified approximation to LDA (a proxy), specific to each event.



$$P(e, \theta, \vec{z} | \vec{\alpha}, t_z) = \pi(\theta | \vec{\alpha}) \prod_{i=1}^{n_f} z_i P(f_i | t_z) \quad \Bigg| \quad q(\theta, \vec{z} | \phi, \gamma) = q(\theta | \gamma) \prod_{i=1}^{n_f} q(z_i | \phi_i)$$

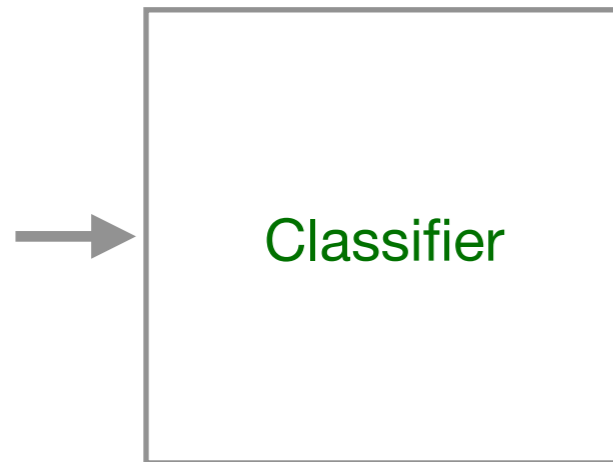
Inference

Variational inference relies on using a simplified approximation to LDA (a proxy), specific to each event.

1. Randomly initialise the latent distributions
2. **For each event:**
Fit $q(\theta, \vec{z}|\phi, \gamma)$ to the function $P(e, \theta, \vec{z}|\vec{\alpha}, t_z)$
You have a different γ and ϕ for each event
3. **Use the proxy model for each event, to fit and update the latent distributions**
4. Calculate the likelihood
5. Repeat steps 2-4 until the likelihood is within some threshold

The procedure is explained in good detail in the appendices of the original paper:
'Latent Dirichlet Allocation': Blei, Ng, and Jordan, 2003

Classification



Unsupervised: we classify on the data we infer on

Similar to a clustering procedure like DBSCAN or K-means.

Use LDA to infer the proportions of the different processes in each event:

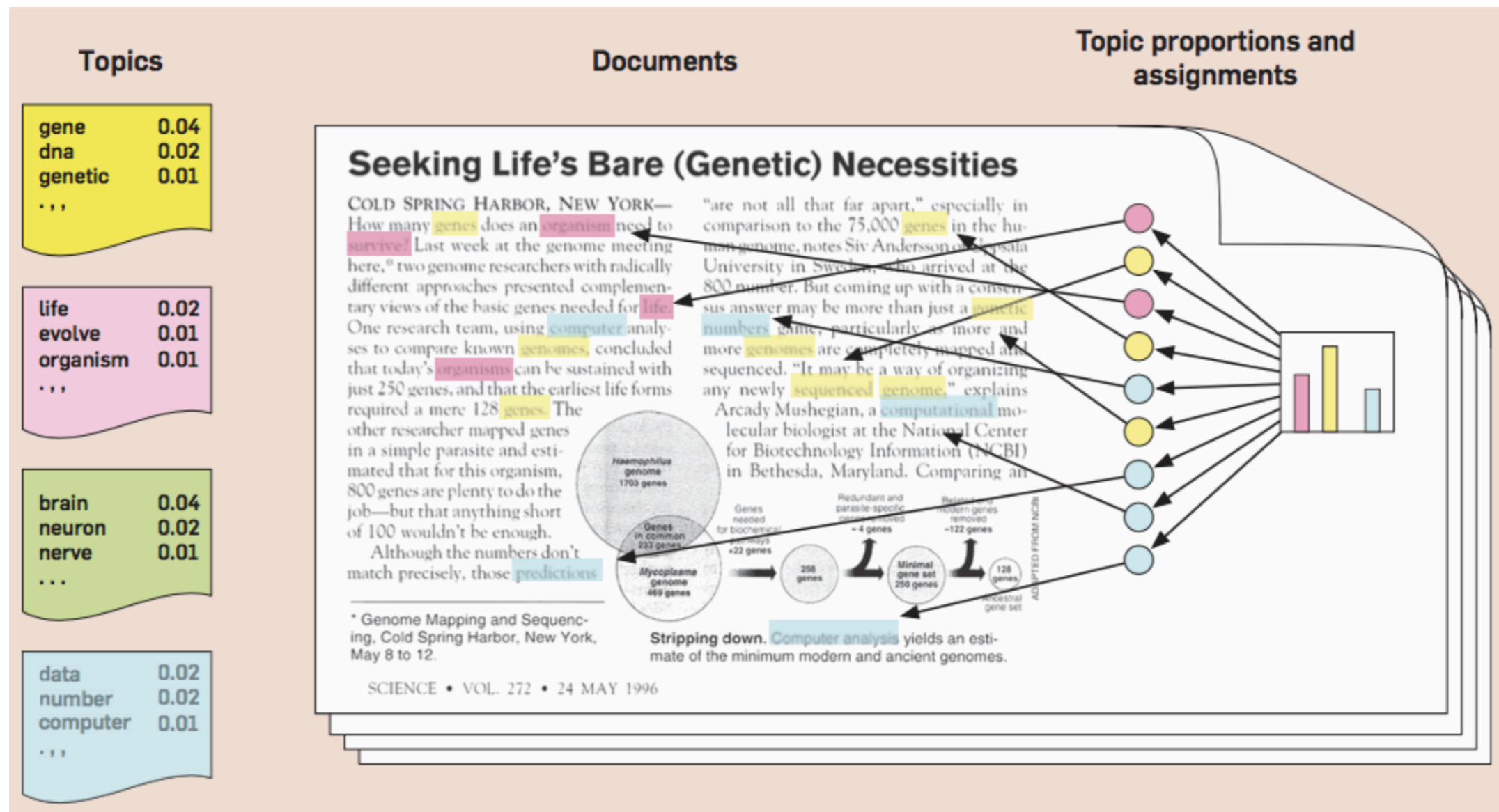
$$\hat{\theta}(z) = \operatorname{argmax}_{\theta} P(e, \theta(z) | \vec{\alpha}, t_z)$$

Construct alternative test-statistic, such as the likelihood ratio:

$$L(e) = L(f_1, \dots, f_{n_f}) = \prod_{i=1}^{n_f} \frac{P(f_i | t_S)}{P(f_i | t_B)}$$

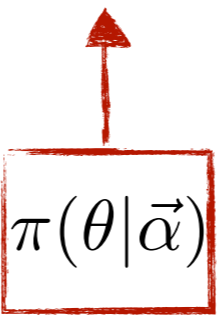
Topic modelling

LDA can be used to ‘discover’ **topics** within text documents.
A topic is simply a latent probability distribution over the vocabulary.



Prioritising rare events

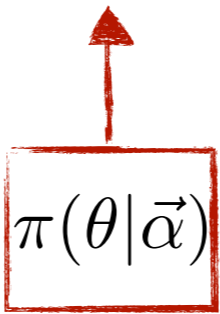
The Dirichlet prior

$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$


One signal and one background: $K=2 \Rightarrow \vec{\alpha} = [\alpha_0, \alpha_1]$
Dirichlet = the Beta function.

Useful re-definition: $\rho = \alpha_1 / \alpha_0 \quad \Sigma = \alpha_0 + \alpha_1$

The Dirichlet prior

$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$


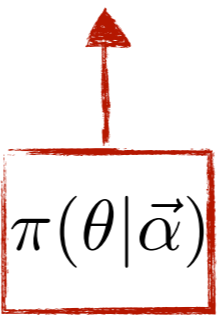
One signal and one background: $K=2 \Rightarrow \vec{\alpha} = [\alpha_0, \alpha_1]$
Dirichlet = the Beta function.

Useful re-definition: $\rho = \alpha_1 / \alpha_0 \quad \Sigma = \alpha_0 + \alpha_1$

$$\int_0^1 d\theta \pi(\theta|\alpha_0, \alpha_1) [\theta P(f_i|t_B) + (1 - \theta) P(f_i|t_S)]$$
$$= \frac{1}{1 + \rho} (P(f_i|t_B) + \rho P(f_i|t_S))$$

ρ represents the ratio
of signal to background
processes occurring in
the sample

The Dirichlet prior

$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$


One signal and one background: $K=2 \Rightarrow \vec{\alpha} = [\alpha_0, \alpha_1]$
Dirichlet = the Beta function.

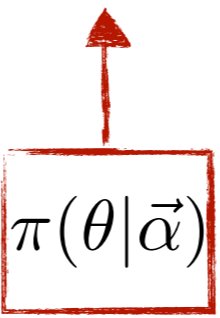
Useful re-definition: $\rho = \alpha_1 / \alpha_0 \quad \Sigma = \alpha_0 + \alpha_1$

Σ controls the distribution of signal and background processes in each event

$\Sigma \ll 1 \rightarrow$ events mostly composed of a single process

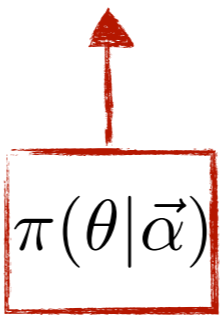
$\Sigma \gg 1 \rightarrow$ events composed of a large mixture of processes

The Dirichlet prior

$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$


How to fix the hyper-parameters? $\rho = \alpha_1/\alpha_0$ $\Sigma = \alpha_0 + \alpha_1$

The Dirichlet prior

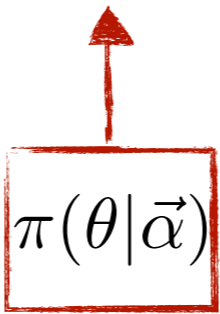
$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$


How to fix the hyper-parameters? $\rho = \alpha_1 / \alpha_0$ $\Sigma = \alpha_0 + \alpha_1$

Choose those that maximise the probability for the data to be generated.

With unsupervised learning, we cannot use truth labels to choose the best model

The Dirichlet prior

$$P(e|\vec{\alpha}, t_z) = \int d\theta \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^K \theta(z) P(f_i|t_z)$$


How to fix the hyper-parameters? $\rho = \alpha_1 / \alpha_0$ $\Sigma = \alpha_0 + \alpha_1$

Choose those that maximise the probability for the data to be generated.

With unsupervised learning, we cannot use truth labels to choose the best model

In practice, maximise: perplexity = $e^{-\frac{\log(P(\text{events}|\rho, \Sigma))}{N}}$

Good model := hyper-parameters that minimise perplexity.

Data-driven/unsupervised top-tagging

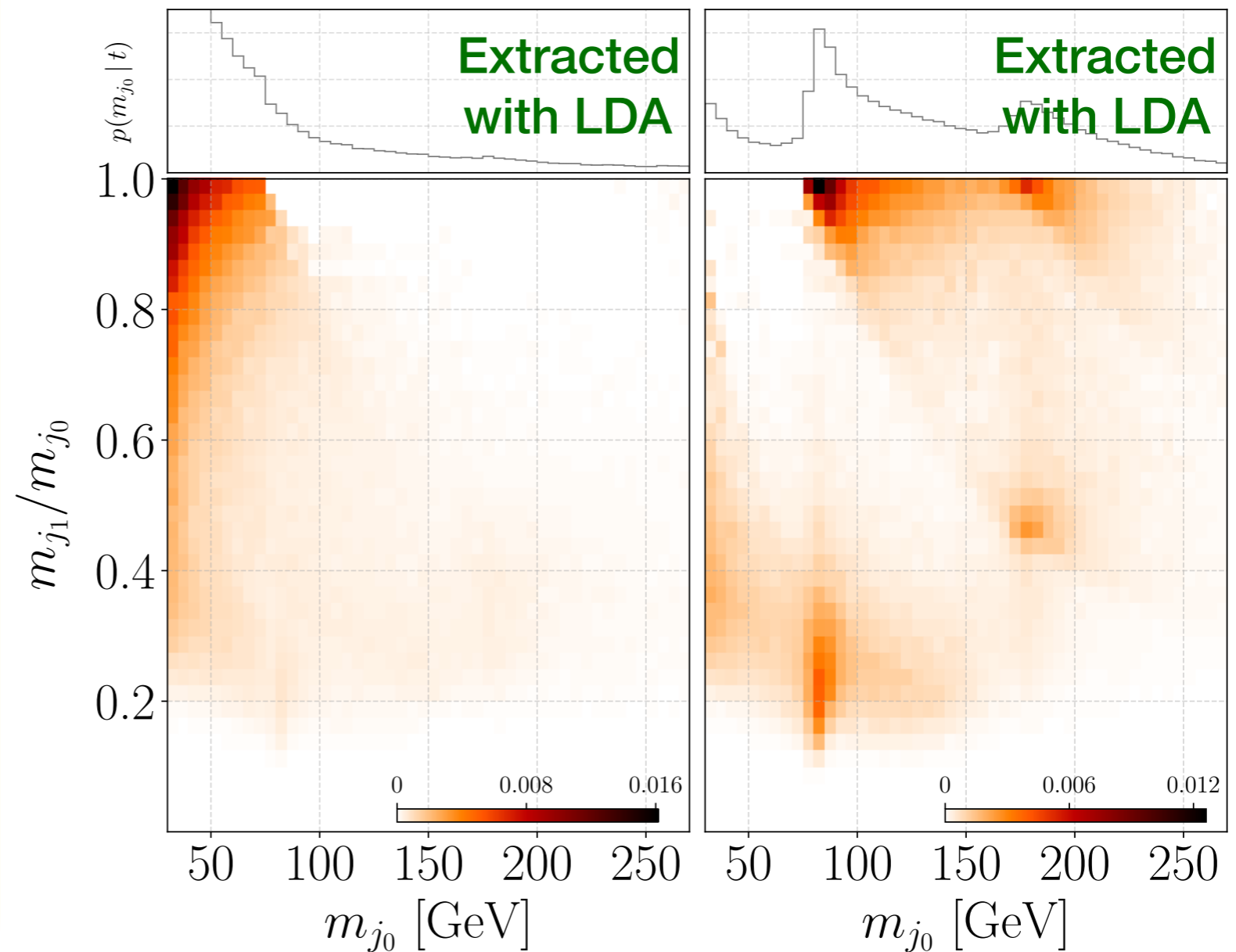
Finding ttbar with LDA

1. **Data:** mixed
unlabelled samples of
QCD and ttbar di-jets,
 $\sim 80,000$ events,
 $p_T \in [350, 450] \text{ GeV}$
 $S/B = 1, 1/9, 1/99$

2. **Model:** LDA with K=2
(Signal and Background)

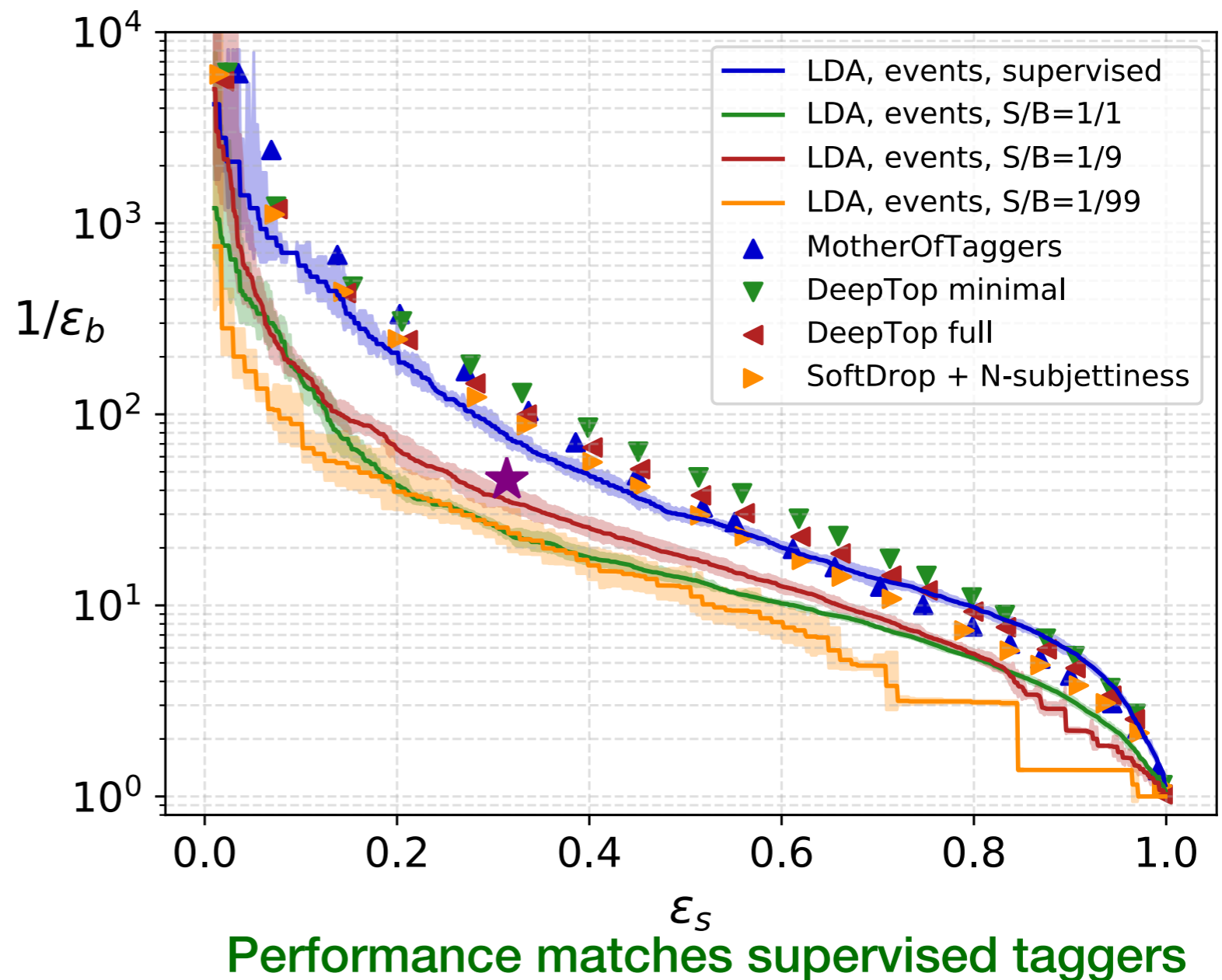
3. **Subjet masses and
mass drop describe
the top decay chain!**

$$t \rightarrow W^+ \bar{b}$$



Finding $t\bar{t}$ with LDA

1. **Data:** mixed unlabelled samples of QCD and $t\bar{t}$ di-jets, $\sim 80,000$ events, $p_T \in [350, 450]\text{GeV}$
 $S/B = 1, 1/9, 1/99$
2. **Model:** LDA with $K=2$ (Signal and Background)
3. **Purple star** is the Johns-Hopkins tagger.



Unsupervised search for new physics

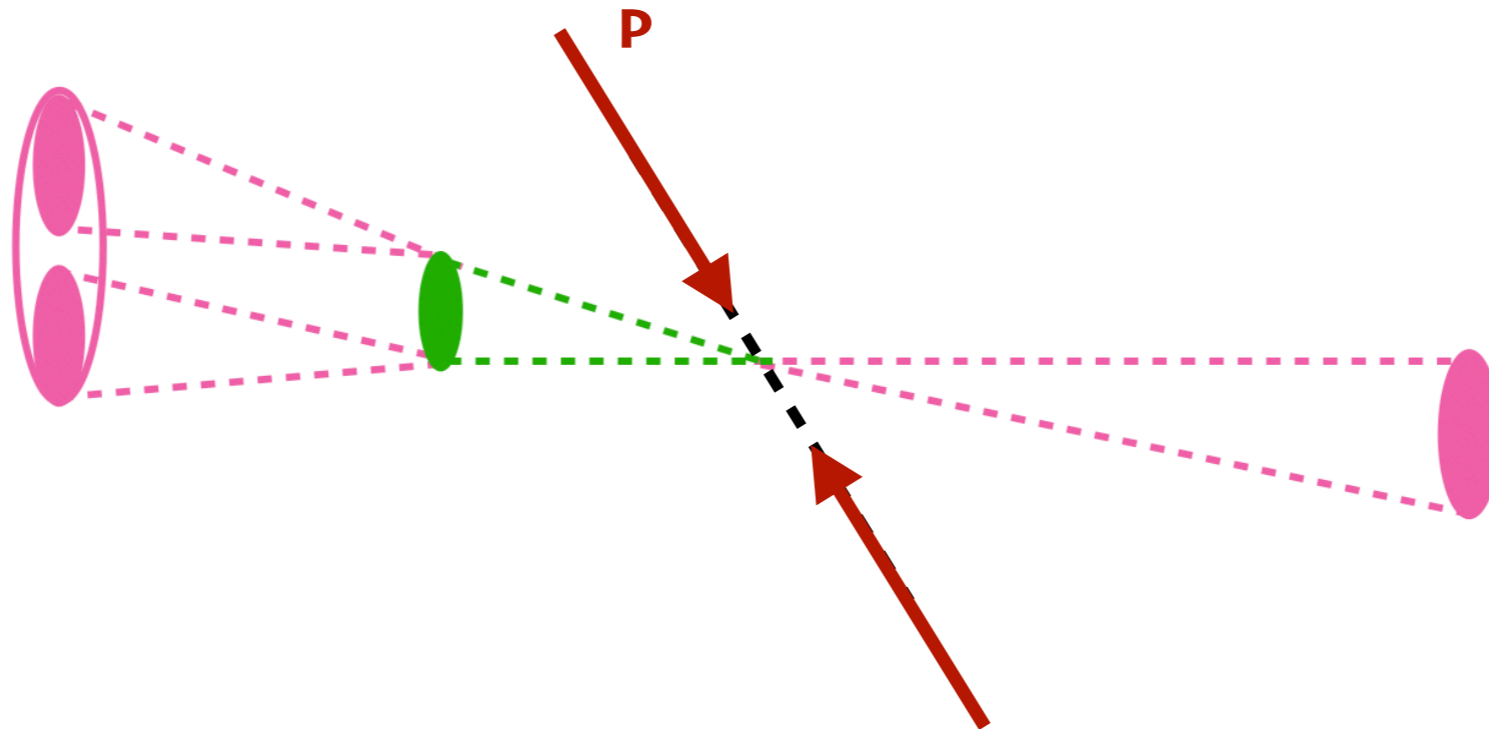
BSM jets

The RS ‘**stealth boson**’ case: $pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$

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

Background: QCD di-jets

Data: **mixed unlabelled sample**, $S/B = 1\%$



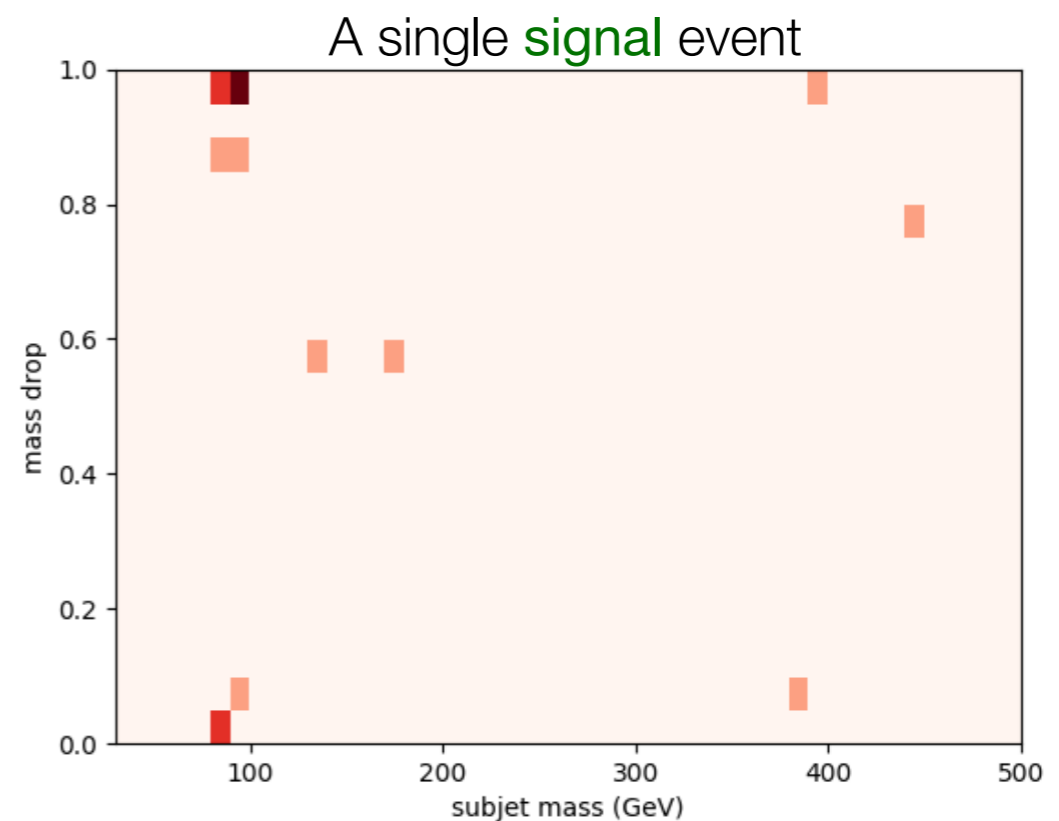
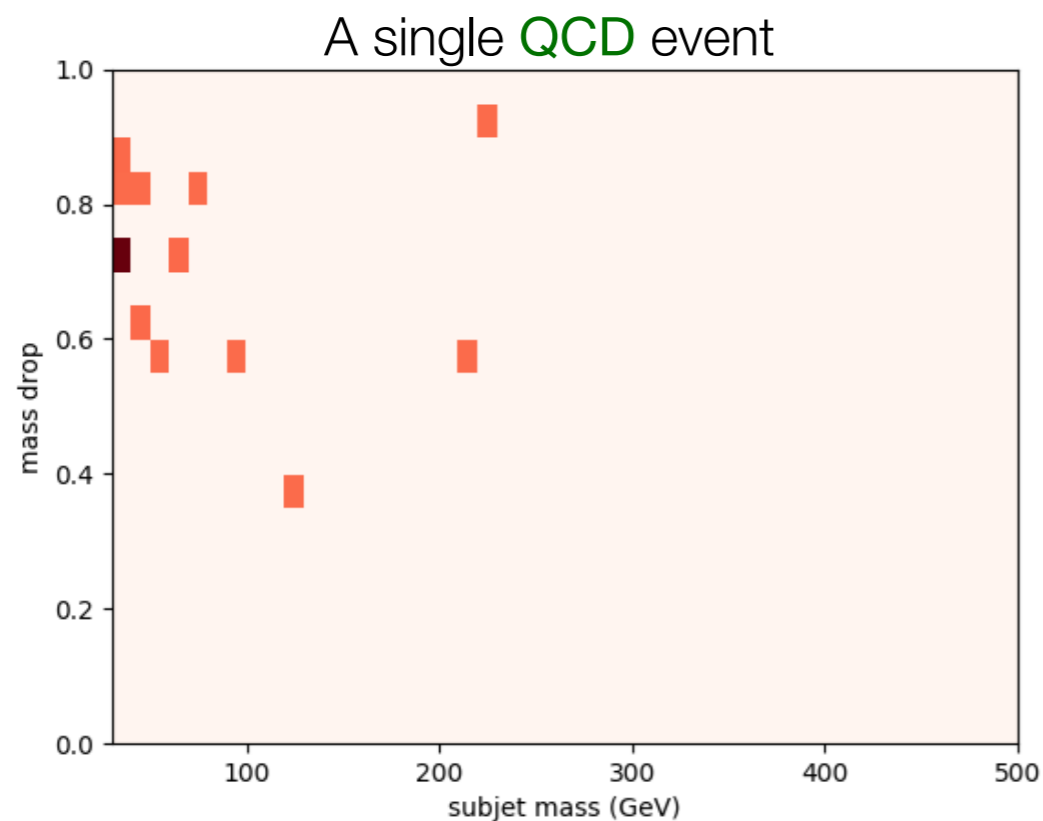
BSM jets

The RS ‘stealth boson’ case: $pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$

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

Background: QCD di-jets

Data: mixed unlabelled sample, $S/B = 1\%$



BSM jets

The RS ‘stealth boson’ case:

$$pp \rightarrow W' \rightarrow \phi W \rightarrow WWW$$

$$m_{W'} = 3 \text{ TeV}$$

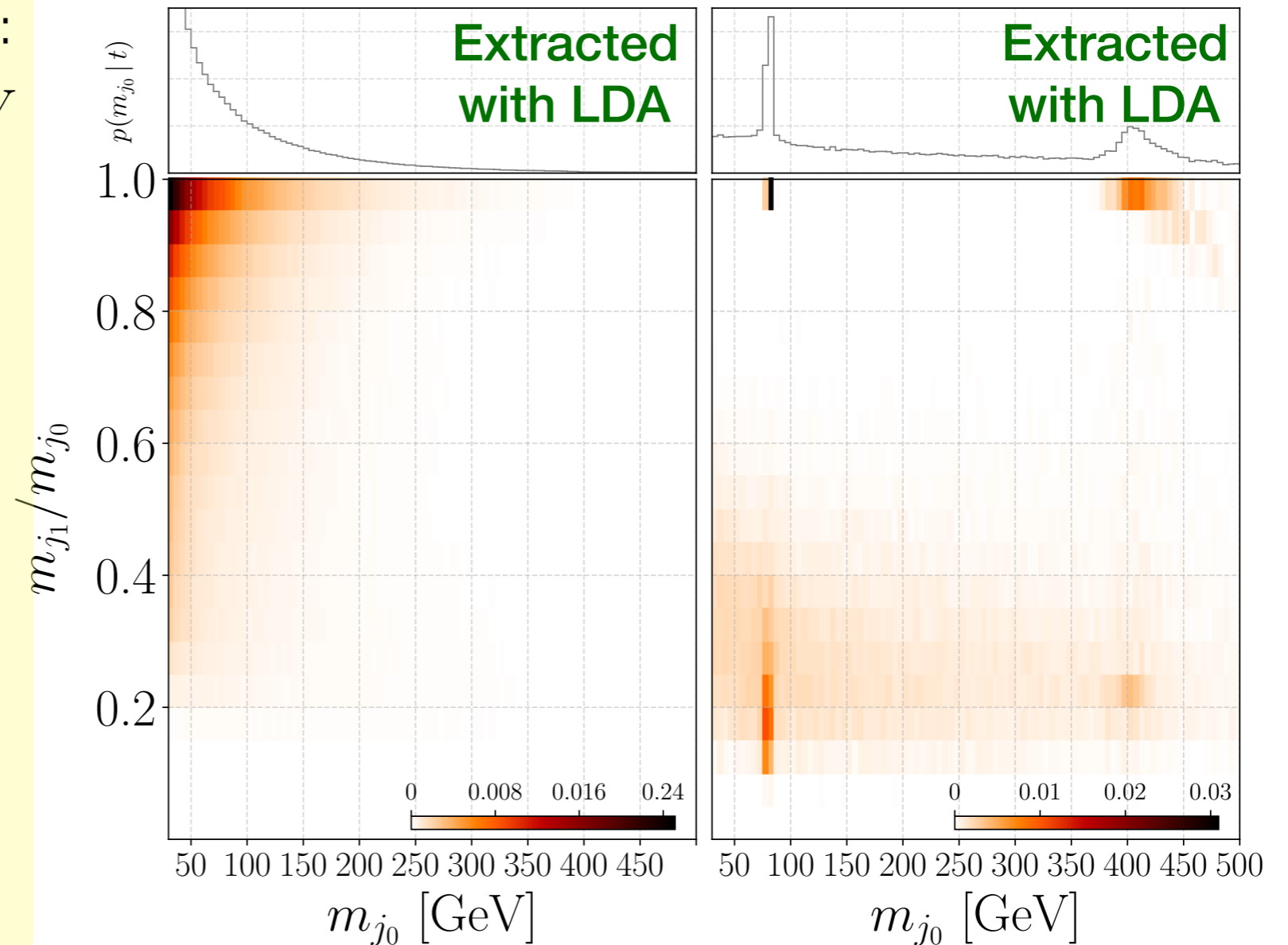
$$m_\phi = 400 \text{ GeV}$$

$$m_{jj} \in [2730, 3190] \text{ GeV}$$

Background: QCD di-jets

Data: mixed unlabelled sample, $S/B = 1\%$

Very clear signal distribution!



BSM jets

The RS ‘stealth boson’ case:

$$pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$$

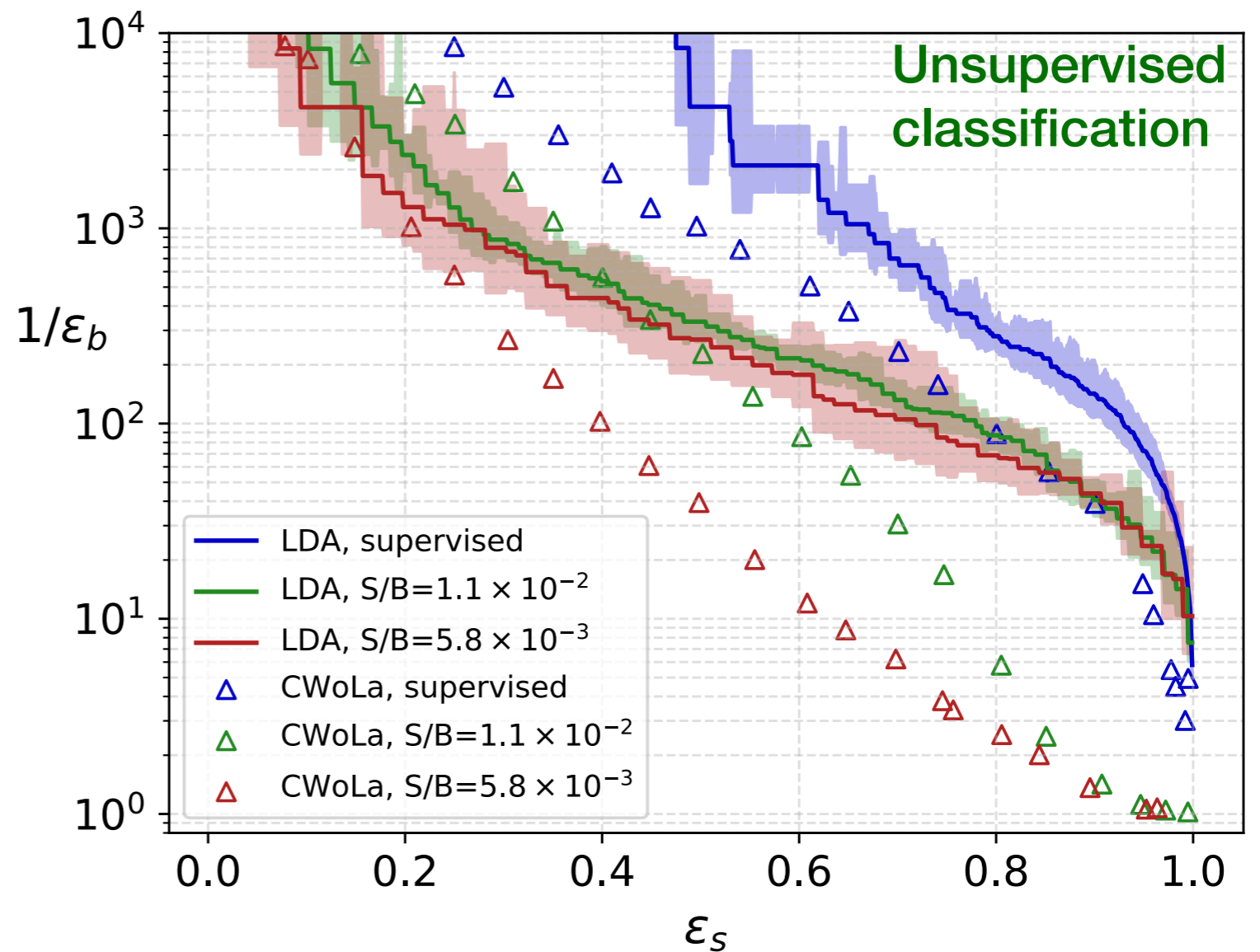
$$m_{W'} = 3 \text{ TeV}$$

$$m_\phi = 400 \text{ GeV}$$

$$m_{jj} \in [2730, 3190] \text{ GeV}$$

Background: QCD di-jets

Data: mixed unlabelled
sample, $S/B = 1\%$



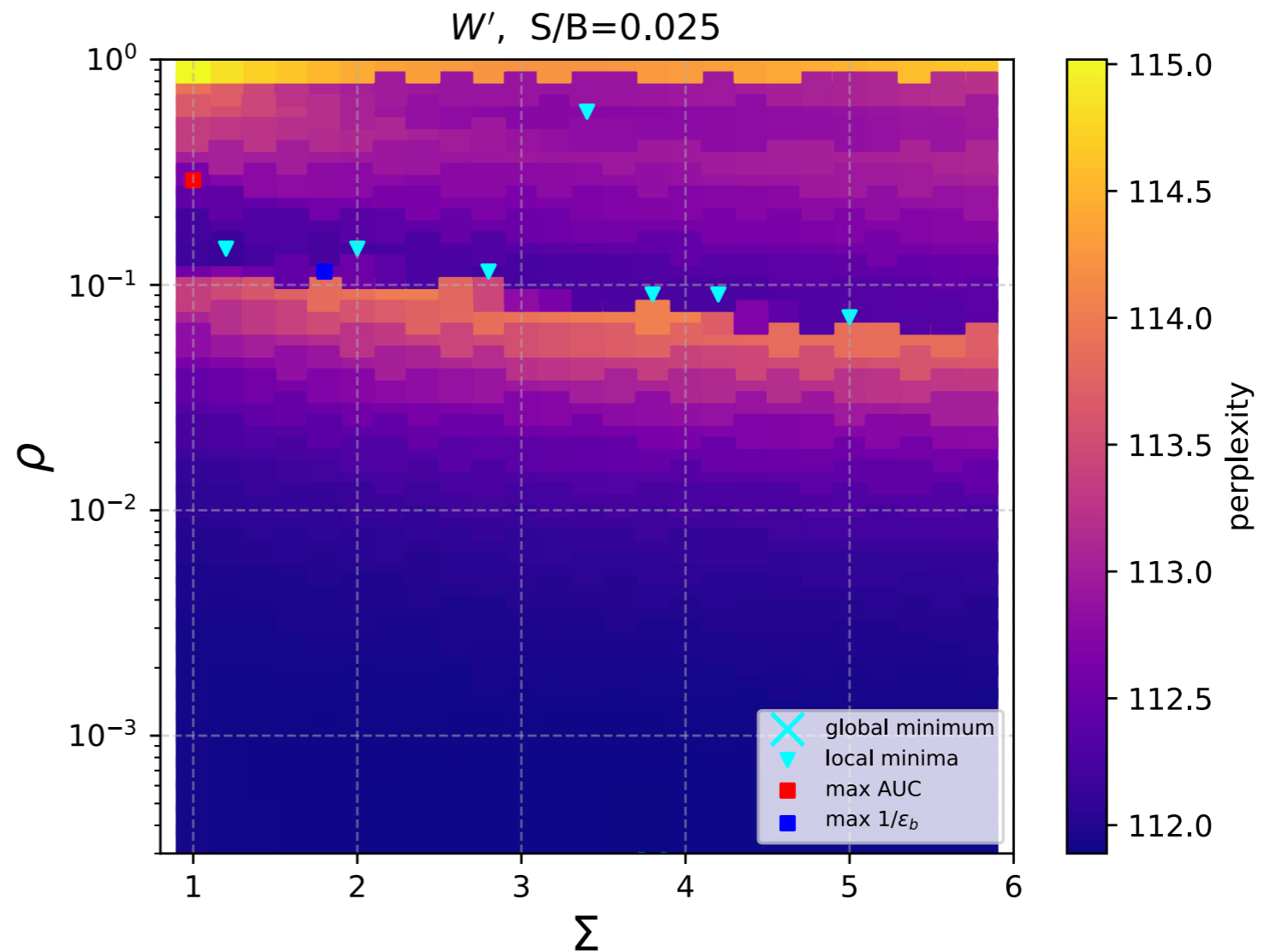
Comparison to CWoLa (Collins et al, 2019)

Fixing the hyper-parameters

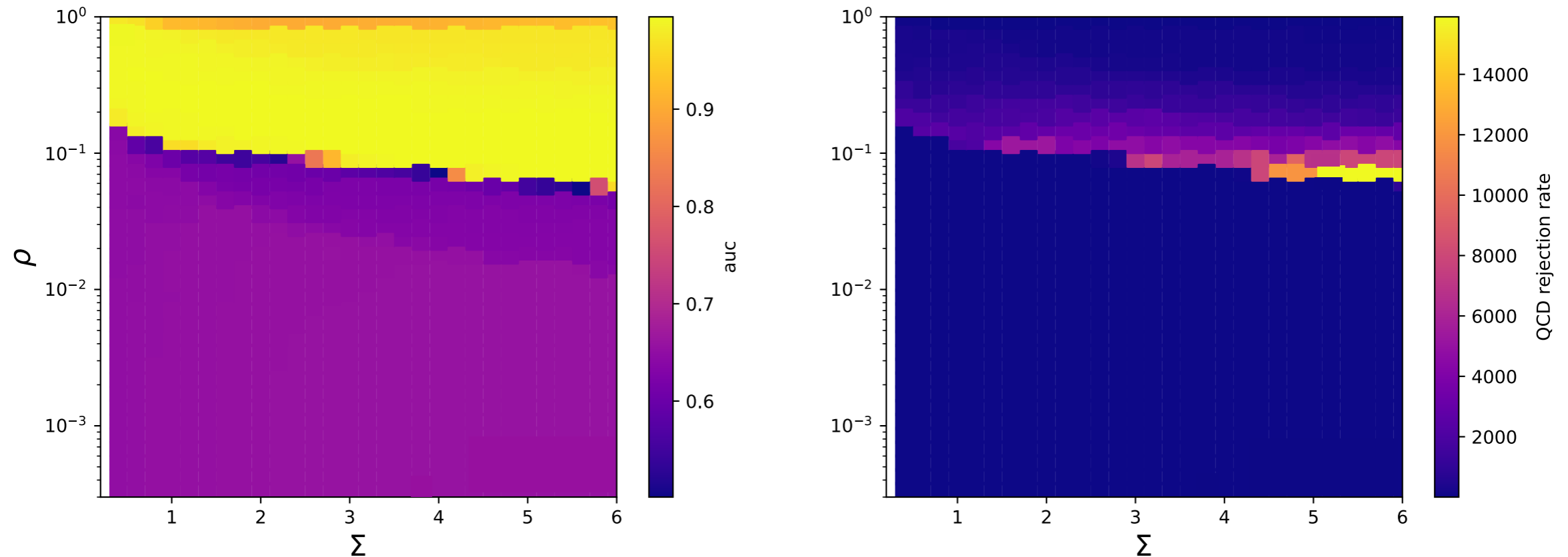
We scan over the hyper-parameters:

lots of local minima,
close to models with best
AUC and best rejection
rate at fixed mis-tag.

global minimum
at vanishing ρ ,
but this is a trivial
solution.



Fixing the hyper-parameters



High-performance regions match those of (local) minimum perplexity.

The prior/hyper-parameters allow for a focusing on small S/B.

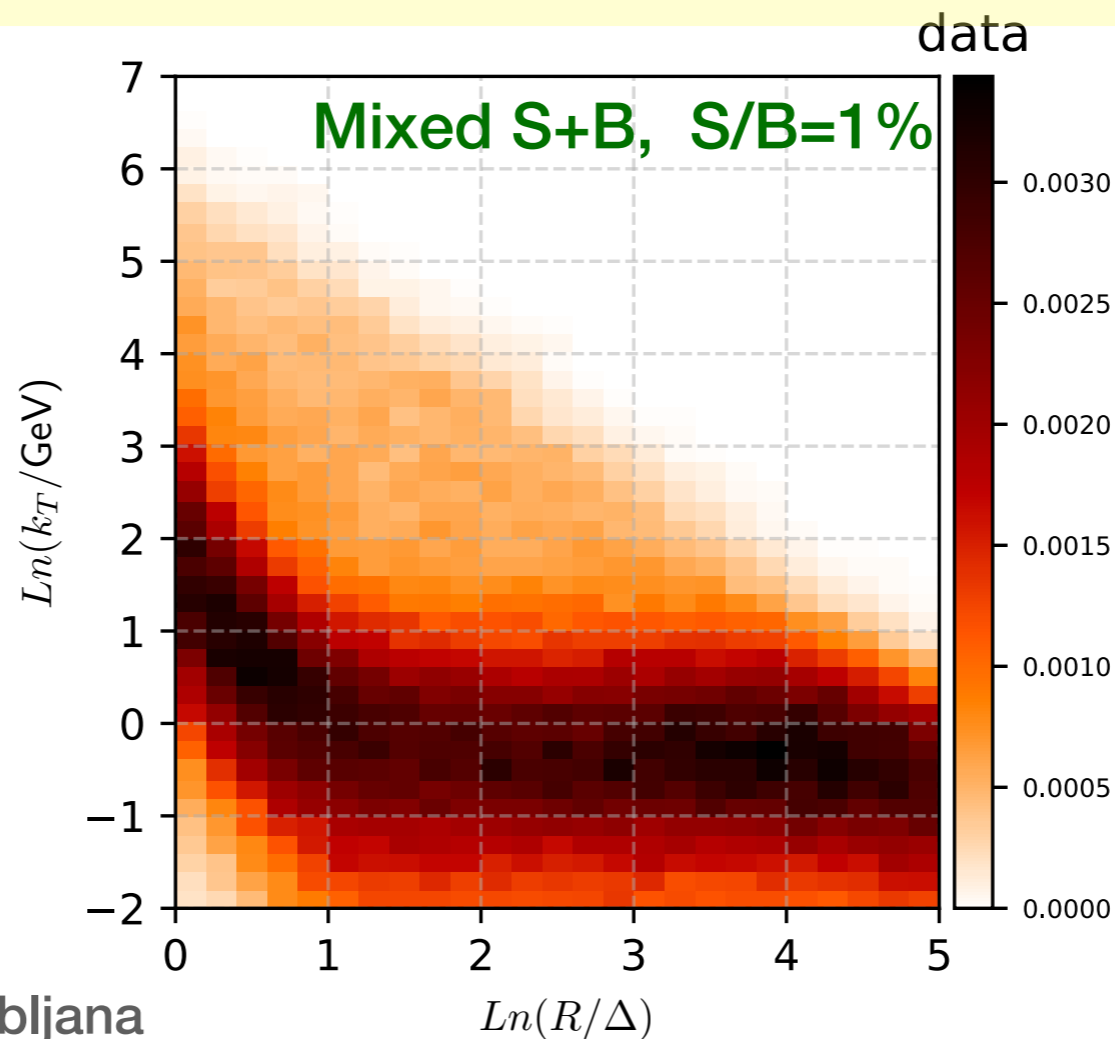
BSM jets in the Lund plane

The RS ‘stealth boson’ case: $pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$

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

Background: QCD di-jets

Data: mixed unlabelled sample, $S/B = 1\%$



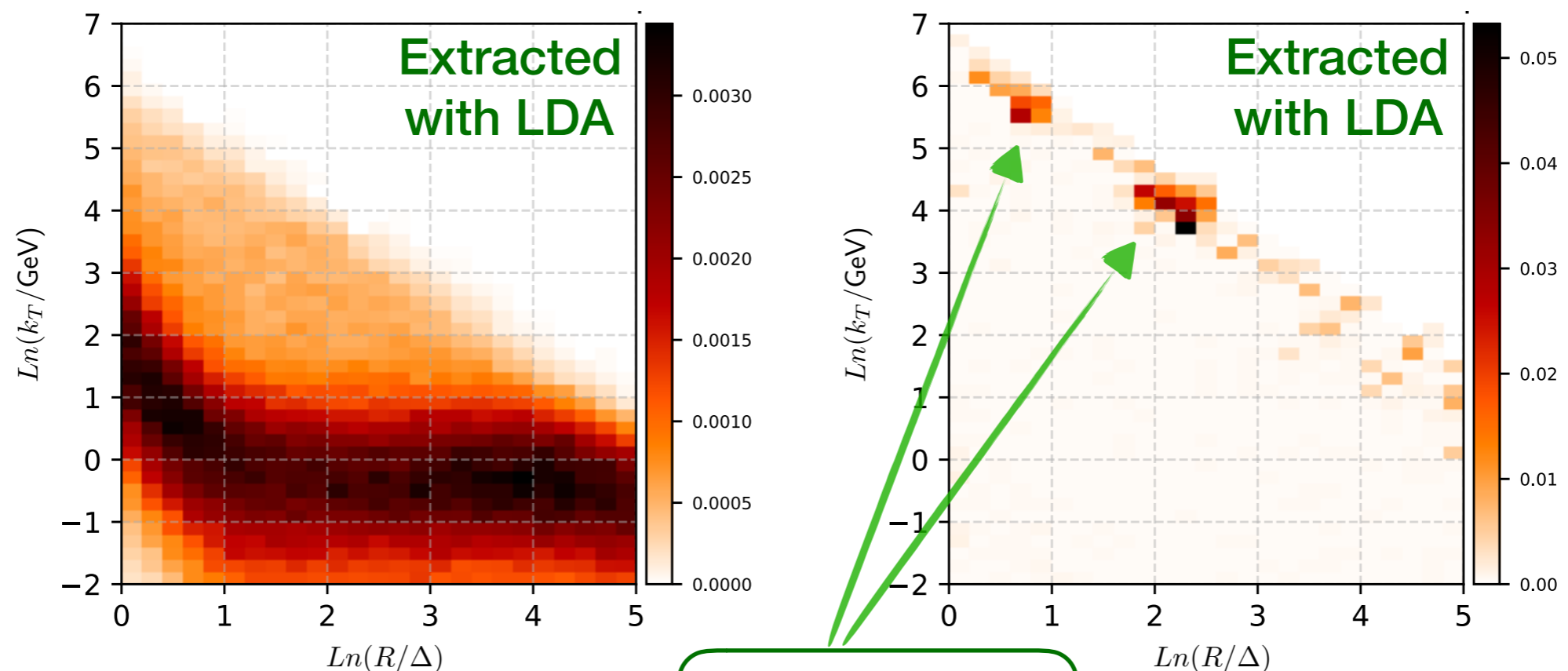
BSM jets in the Lund plane

The RS ‘stealth boson’ case: $pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$

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

Background: QCD di-jets

Data: mixed unlabelled sample, $S/B = 1\%$



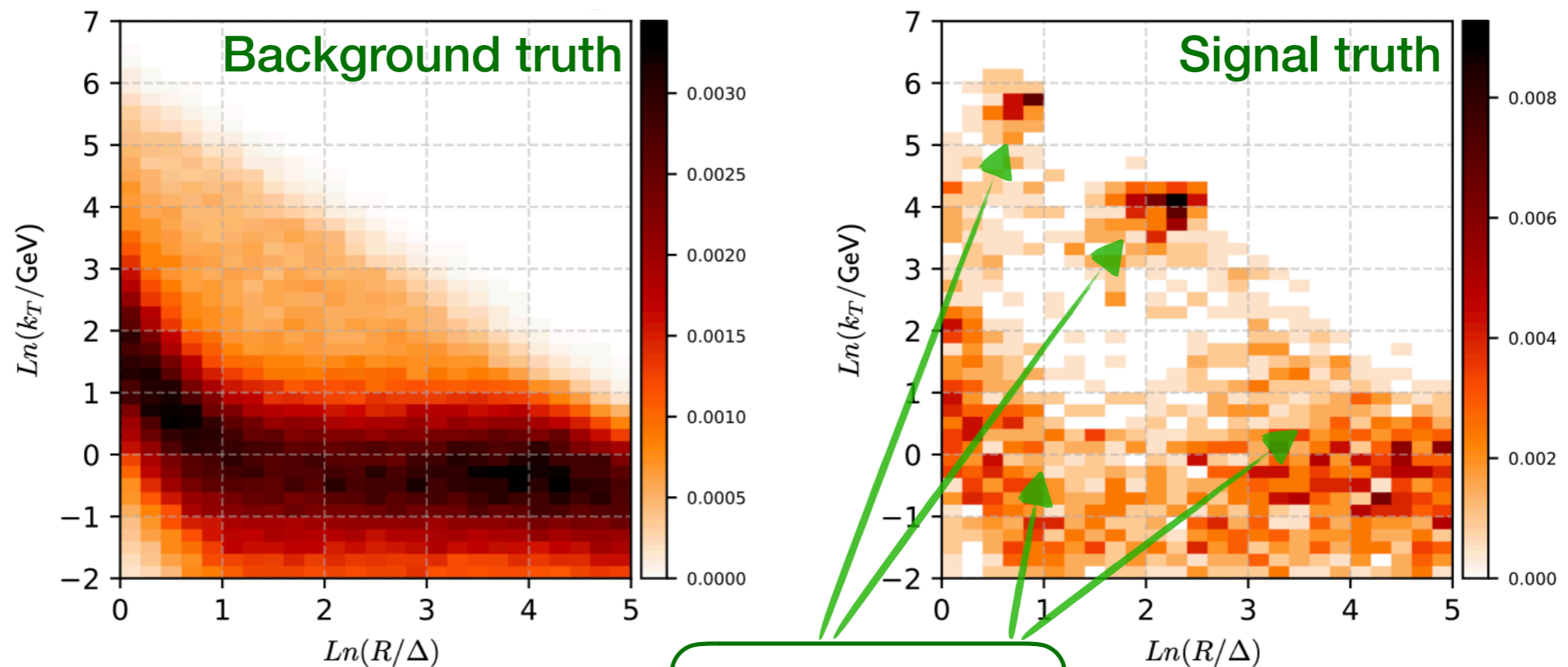
BSM jets in the Lund plane

The RS ‘stealth boson’ case: $pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$

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

Background: QCD di-jets

Data: mixed unlabelled sample, $S/B = 1\%$



LDA separates soft and hard physics!

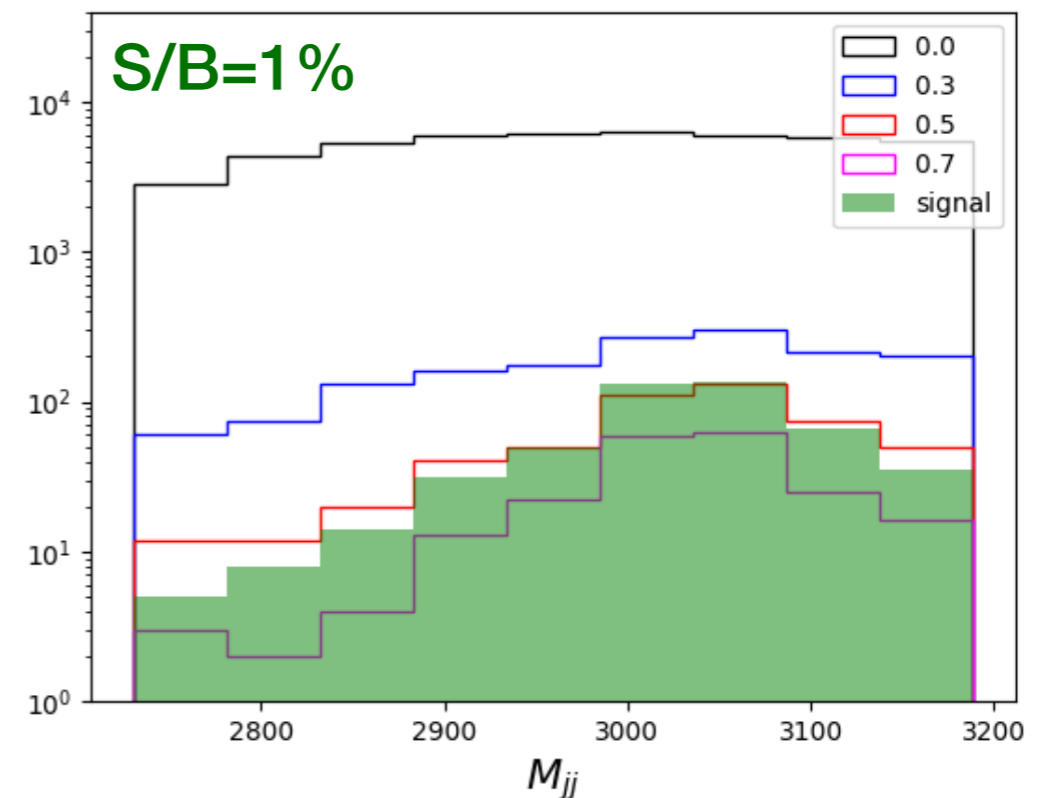
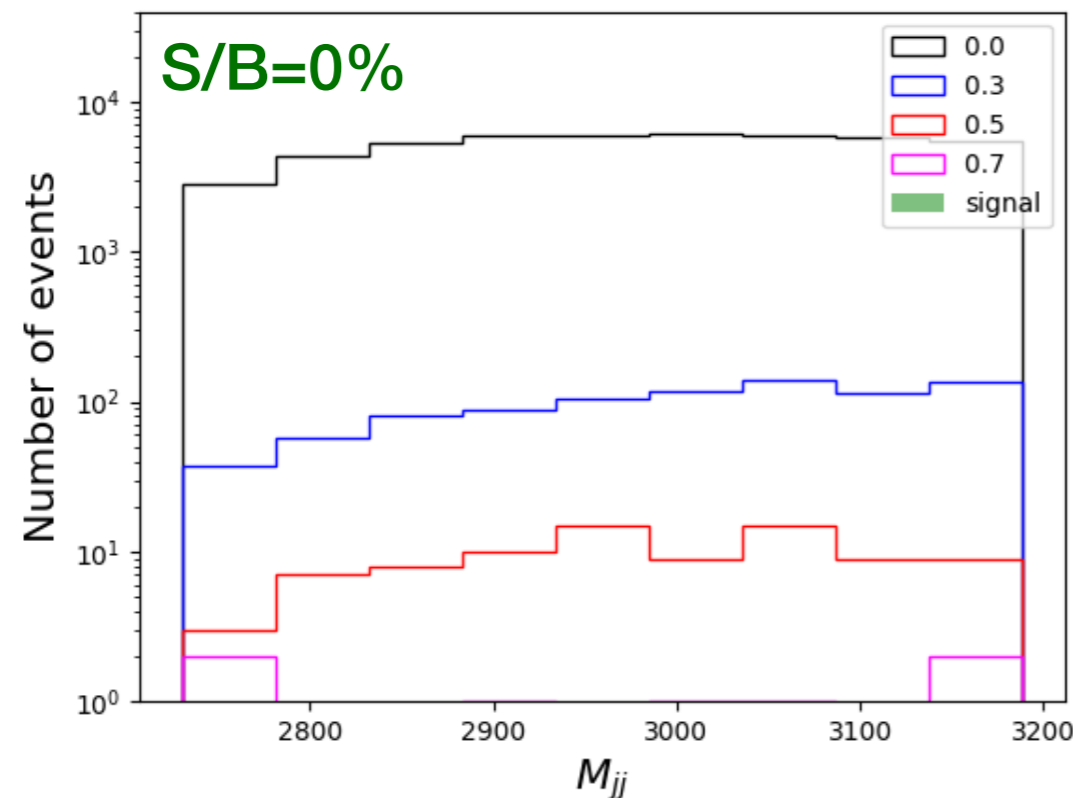
Unsupervised **bump hunting** with LDA

The RS ‘**stealth boson**’ case: $pp \rightarrow W' \rightarrow \phi W \rightarrow W W W$

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

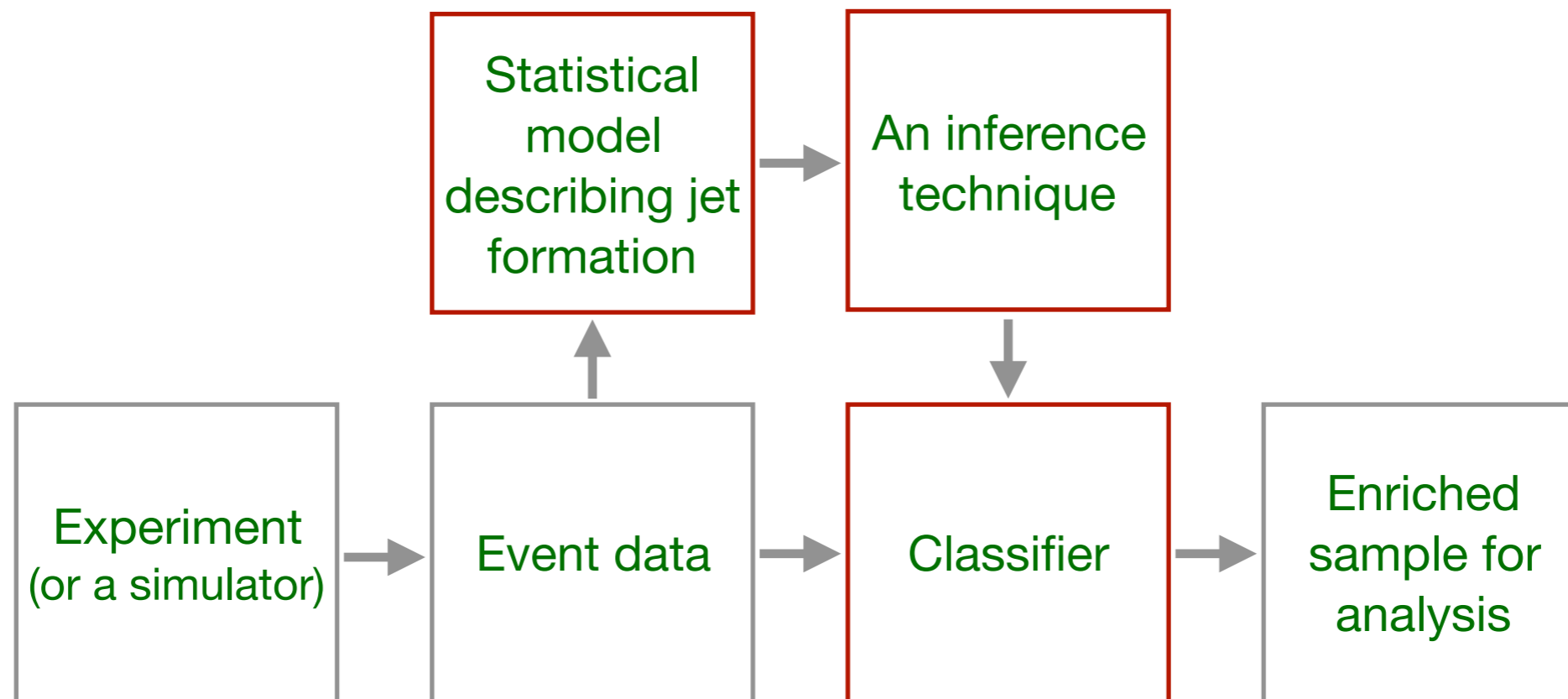
Background: QCD di-jets

Data: **mixed unlabelled sample**, $S/B = 1\%$



Conclusions

Conclusions



LDA can extract rare signals from data, with no a-priori knowledge of the signal

Also: separation of perturbative and non-perturbative physics, bump hunt tool, ...

Next steps? Go beyond multi-jets, include pile-up, full bump hunt analysis, ...

Additional slides

Discussion

Future directions

- In what ways could the LDA algorithm/model be upgraded?
(Aspect extraction, semi-supervised learning, ...)
- How can these methods be implemented in a realistic search strategy?
- Are the advantages gained with LDA as compared to anomaly detectors important?
- Can these methods be used to search for general event topologies?
How would the data be represented?
What modifications to the LDA algorithm would be beneficial?