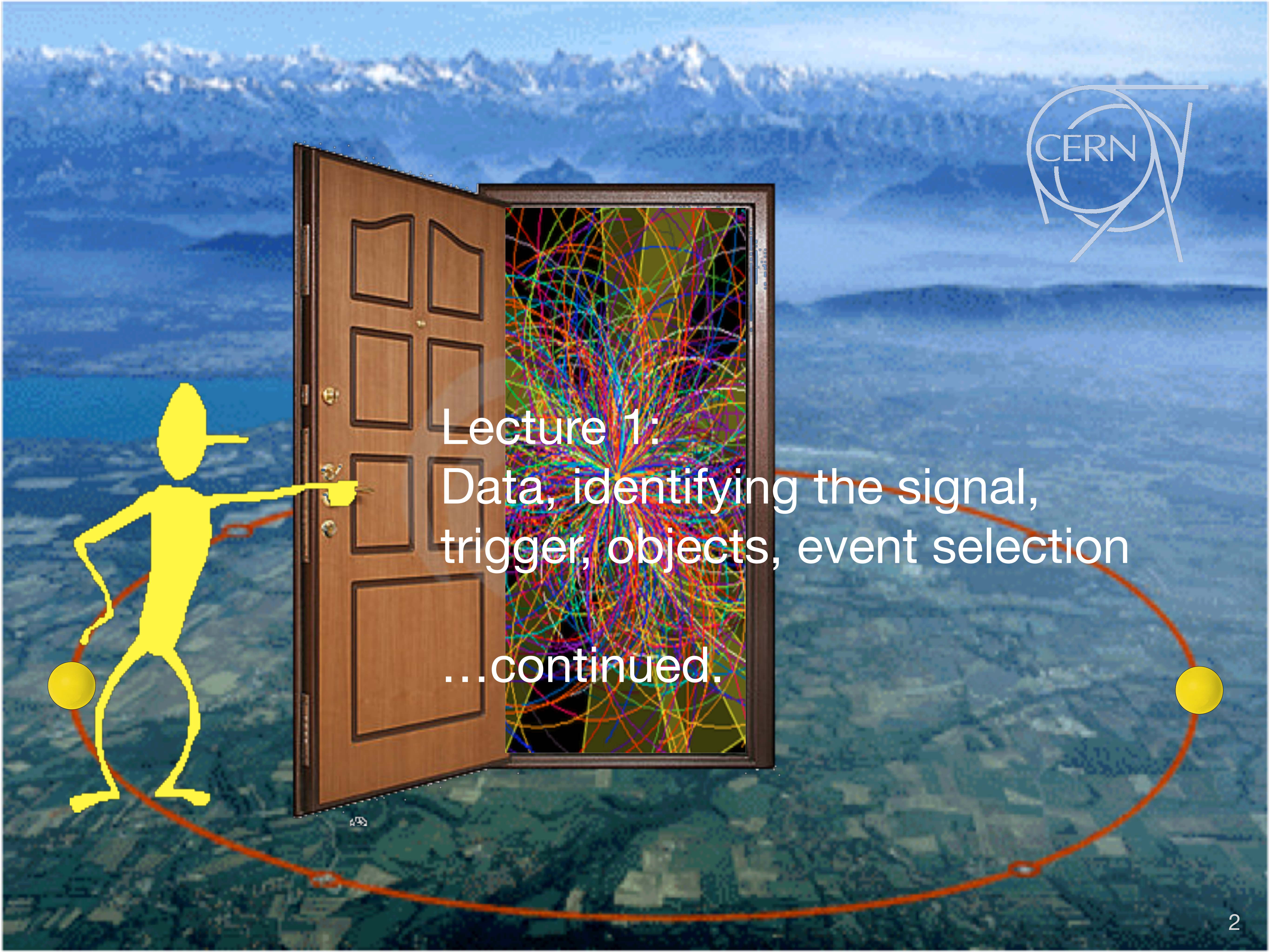
Experimental Methods and Physics at the LHC - II

Sezen Sekmen
Kyungpook National University / CMS

26th Vietnam School of Physics: Particles and Dark Matter
29 Nov - 11 Dec 2020, Quy Nhon & virtual

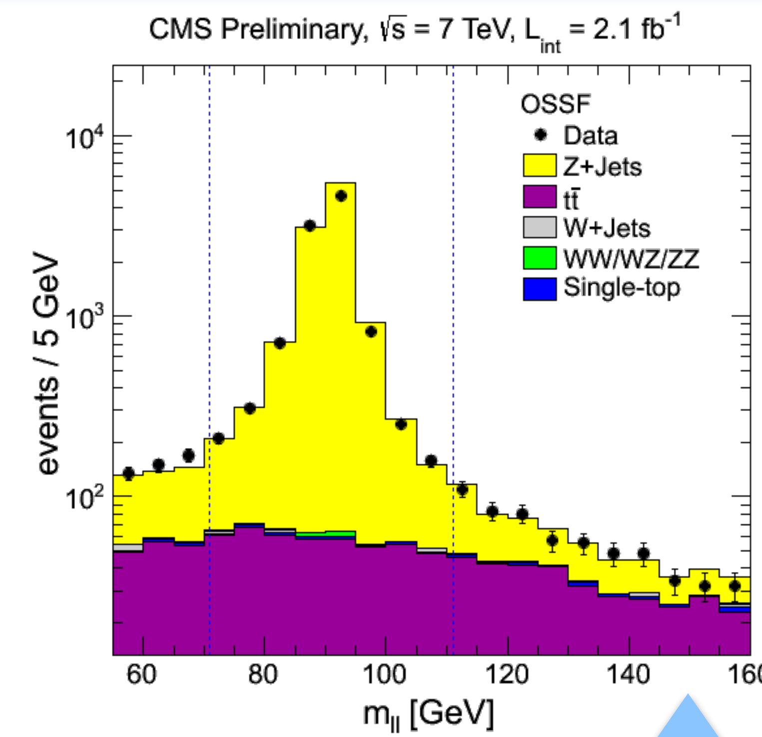

CERN
Lecture 1:
Data, identifying the signal,
trigger, objects, event selection
...continued.

Characterizing the signal

Good old invariant mass

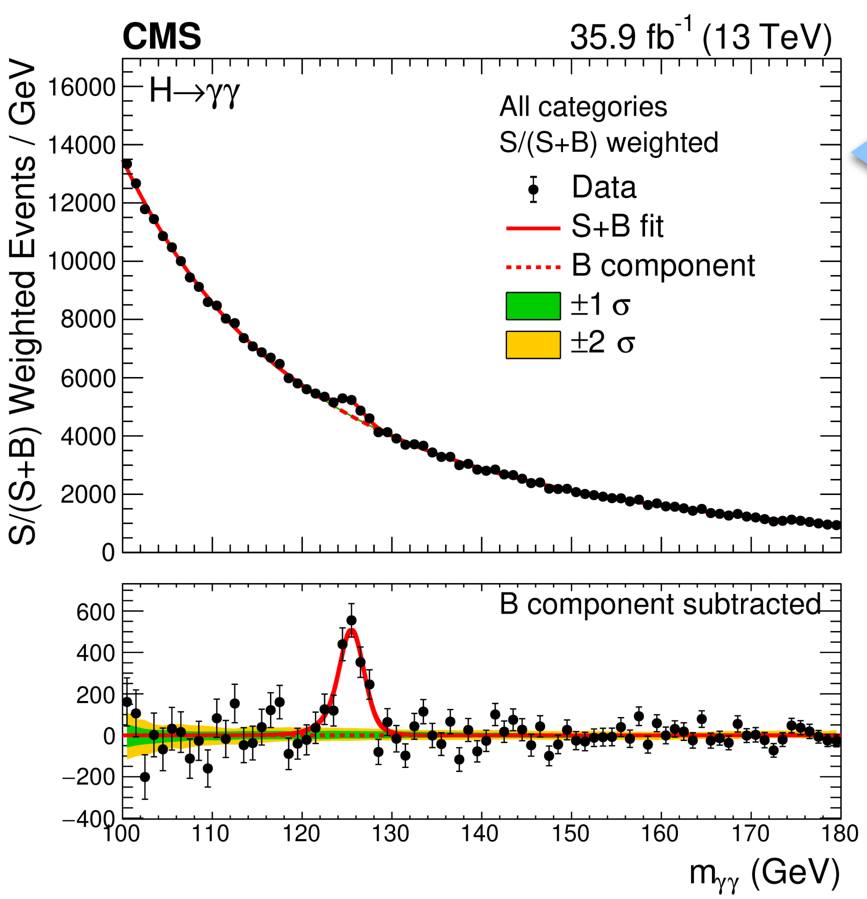
A mother particle decaying into I final state particles has the invariant mass:

$$m = \sqrt{\left(\sum_i E^i\right)^2 - \left(\sum_i \vec{p}^i\right)^2}$$



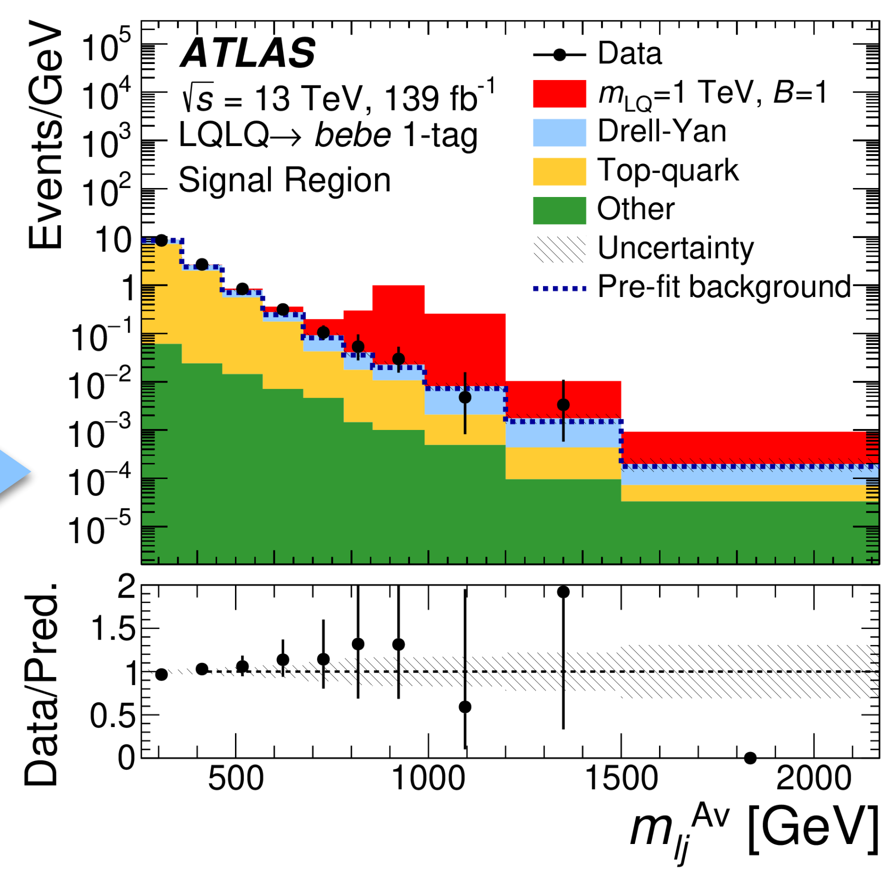
Inv. mass for a mother particle can be reconstructed if the 4-momenta of all its daughter particles are known. This happens when the decays products are visible.

Inv. mass is used when requiring the particles with a known mass (e.g. Zs) in selection, or when looking for new states.



Higgs observed at $m_{\gamma\gamma}$ = 125 GeV.

However no excess for leptoquarks in invariant mass distribution.



Inv. mass also helps new physics searches indirectly.

$m_{Dilepton}$ was used above in a search for SUSY with Z + jets + MET to find events with Zs.

Characterizing the signal
W transverse mass

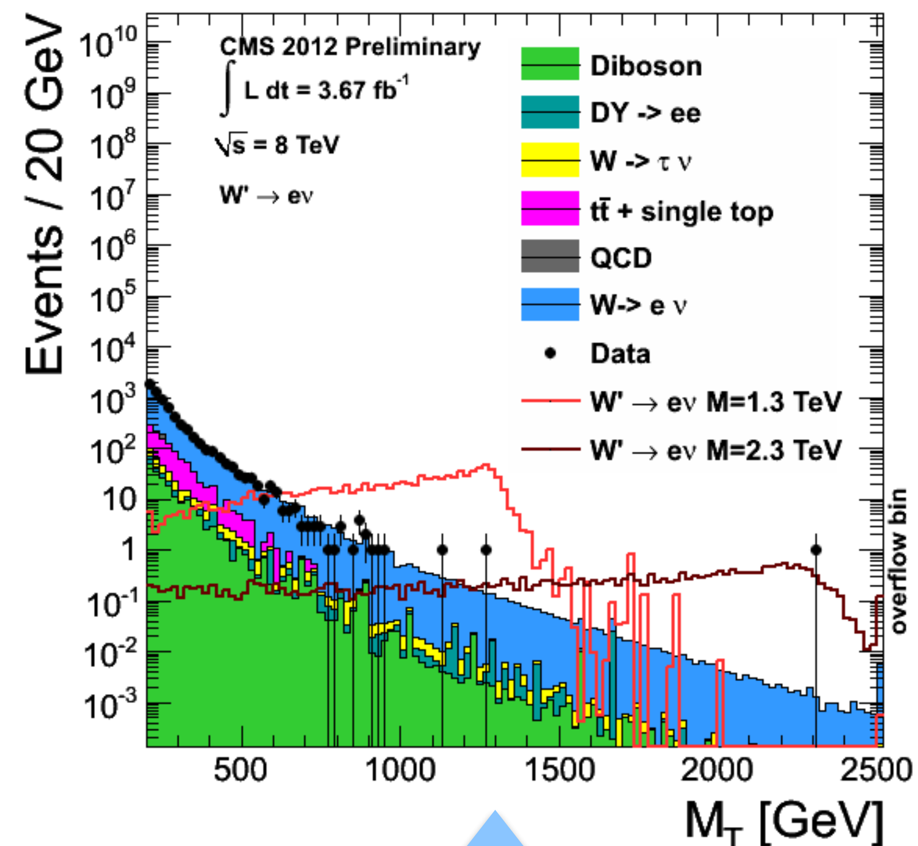
BUT…we do not always have access to full 4-momenta of the final state particles.

For example, in W→lν decays, invisible neutrinos escape the detector. If there is only one ν in the event, we can approximate ν transverse momentum p_T^{ν} by the MET. We define the transverse mass for W as:

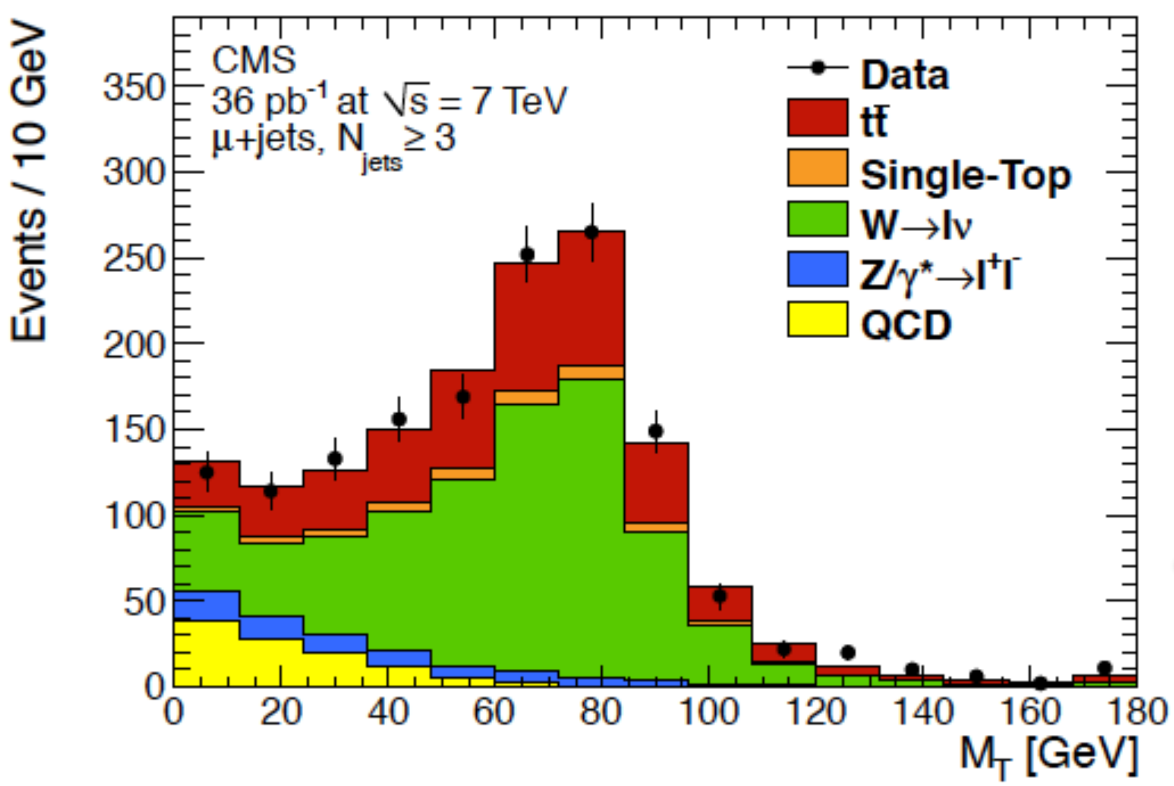
$$m_{T,W}^2 = m_\ell^2 + m_\nu^2 + 2(p_T^\ell p_T^\nu - \vec{p}_T^{\,\ell} \vec{p}_T^{\,\nu})$$

$$(m_\ell, m_\nu \sim 0 \rightarrow) \simeq 2 p_T^\ell p_T^\nu (1 - \cos\Delta\phi(\ell,\nu))$$

where $m_{T,W}^{max}$ gives m_W because $m_{T,W} < m_W$.



Used in new physics searches. M_T distribution for hypothetical W' particles where W'→eν.



W M_T is used extensively in top searches and searches for new physics with top-like particles as a discriminating variable in the event selection (Left: from a ttbar cross section measurement in leptons+jets channel).

Characterizing the signal

The "s"transverse mass

BUT…what if we have more than one invisible particle in the final state? Take the typical case

$$pp \to \tilde{q}_1 \tilde{q}_2 \to j_1 \tilde{\chi}_1 j_2 \tilde{\chi}_2$$

where ~χs are invisible. Two invisible particles make up MET. The stransverse mass

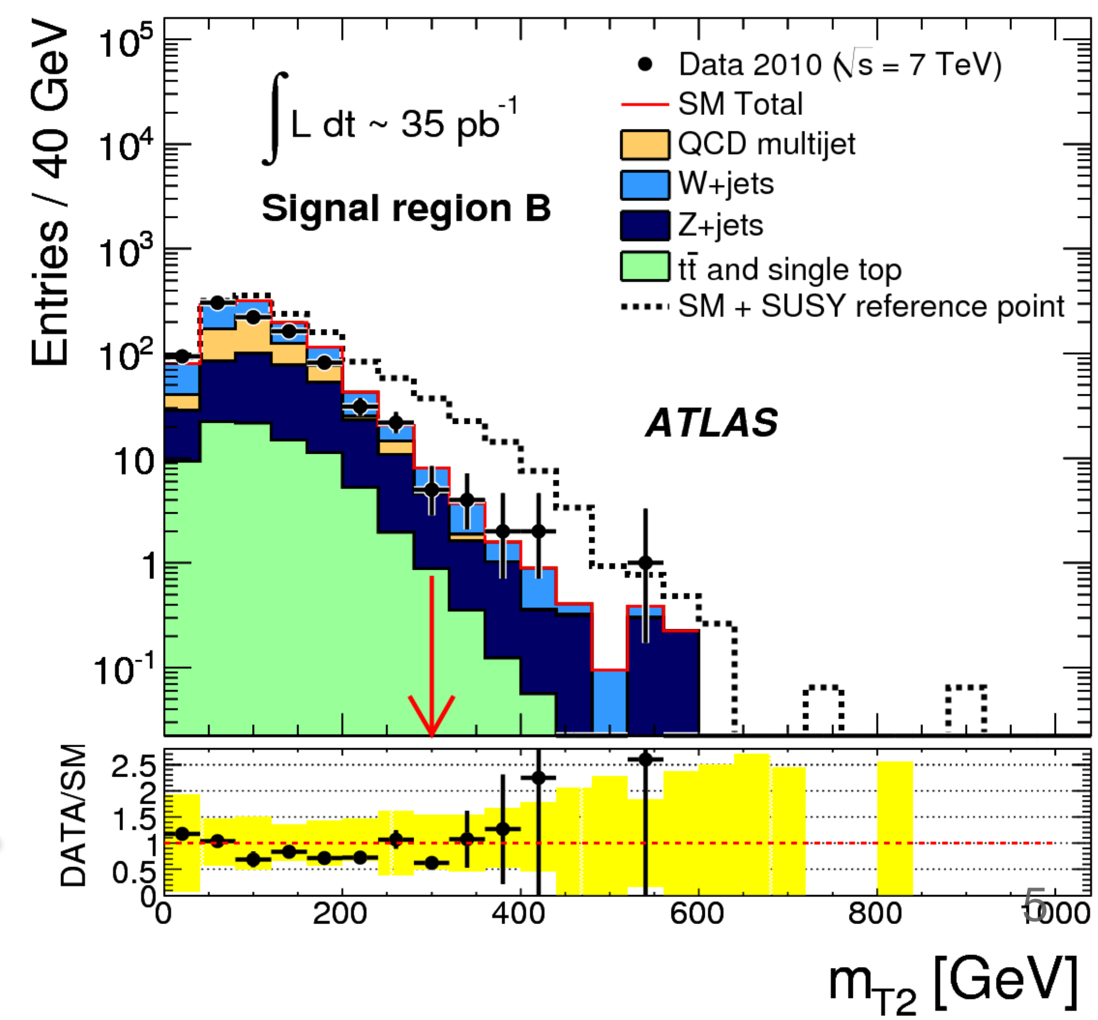
$$m_{T2}(m_{\tilde{\chi}}) = \min_{\vec{p}_T^{\tilde{\chi}1} + \vec{p}_T^{\tilde{\chi}2} = \vec{p}_T^{miss}} \left[\max \left(m_T(\vec{p}_T^{j1}, \vec{p}_T^{\tilde{\chi}1}), m_T(\vec{p}_T^{j2}, \vec{p}_T^{\tilde{\chi}2}) \right) \right] \leq m_{\tilde{q}}^2$$

suggests a way to decompose the MET into these particles.

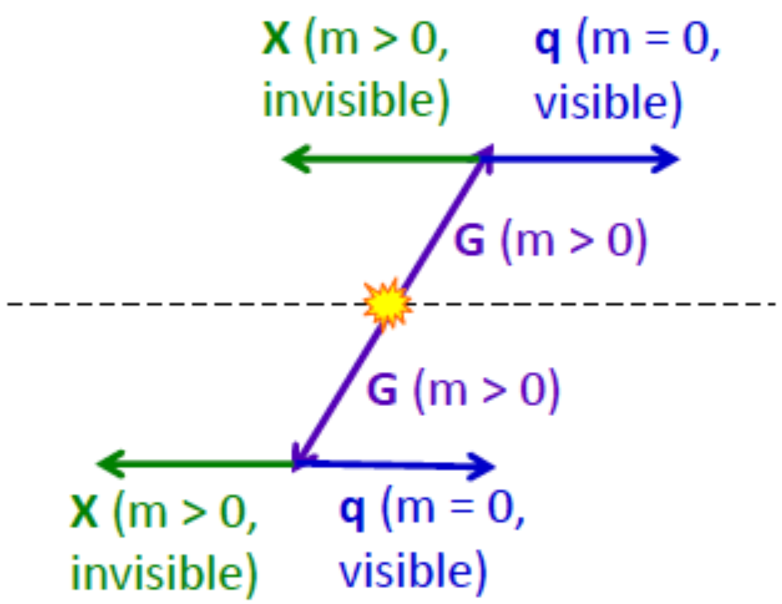
The minimization is over all possible partitions of the measured MET.

However, for massive ~χ, we need the ~χ mass for calculating m_{T2}. It is shown that for different input $m_{\sim\chi}$ values, maximum m_{T2} vs. $m_{\sim\chi}$ curve has a kink at the correct $m_{\sim\chi}$ value.

MT2 is used as a selection variable in SUSY searches in ATLAS and CMS



Razor kinematic variables



Suppose a signal with pair production of heavy particles G, each decaying to a massless visible particle χ and a massive invisible particle q.

In the G rest frame, the momentum of Q is a constant depending on heavy particle masses

$$|\vec{p}^{q}| = \frac{m_G^2 - m_\chi^2}{2m_G} = \frac{m_\Delta}{2}$$

Razor variables estimate the momentum of Q in the G rest frame using lab frame observables.

using longitudinal lab fr. observables:

$$M_R = \sqrt{\frac{(\vec{p}_z^{\,q_1} E^{q_2} - \vec{p}_z^{\,q_2} E^{q_1})^2}{(\vec{p}_z^{\,q_1} - \vec{p}_z^{\,q_2})^2 - (E^{q_1} - E^{q_2})^2}} \approx m_\Delta$$

For a signal with heavy G and χ, M_R distribution peaks at m_Δ. When there are no heavy particles M_R falls exponentially.

using transverse lab fr. observables:

$$M_T^R = \sqrt{\frac{E_T^{miss}}{2}(p_T^{q_1} + p_T^{q_2}) - \frac{1}{2}\vec{E}_T^{miss} \cdot (\vec{p}_T^{\,q_1} + \vec{p}_T^{\,q_2})} < m_\Delta$$

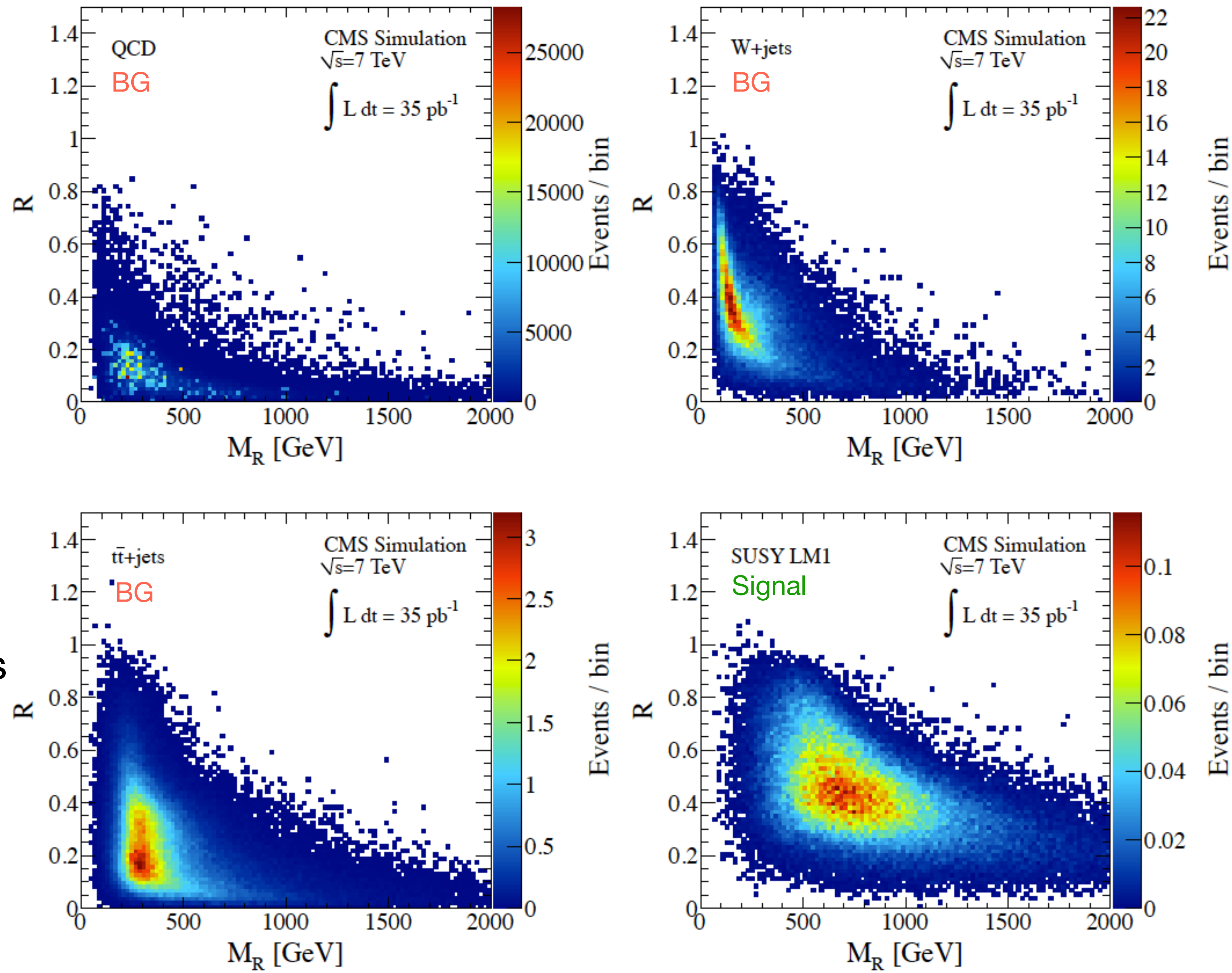
$$R \equiv M_T^R / M_R$$

M_T^R distribution has an endpoint at m_Δ.

Characterizing the signal
Razor variables

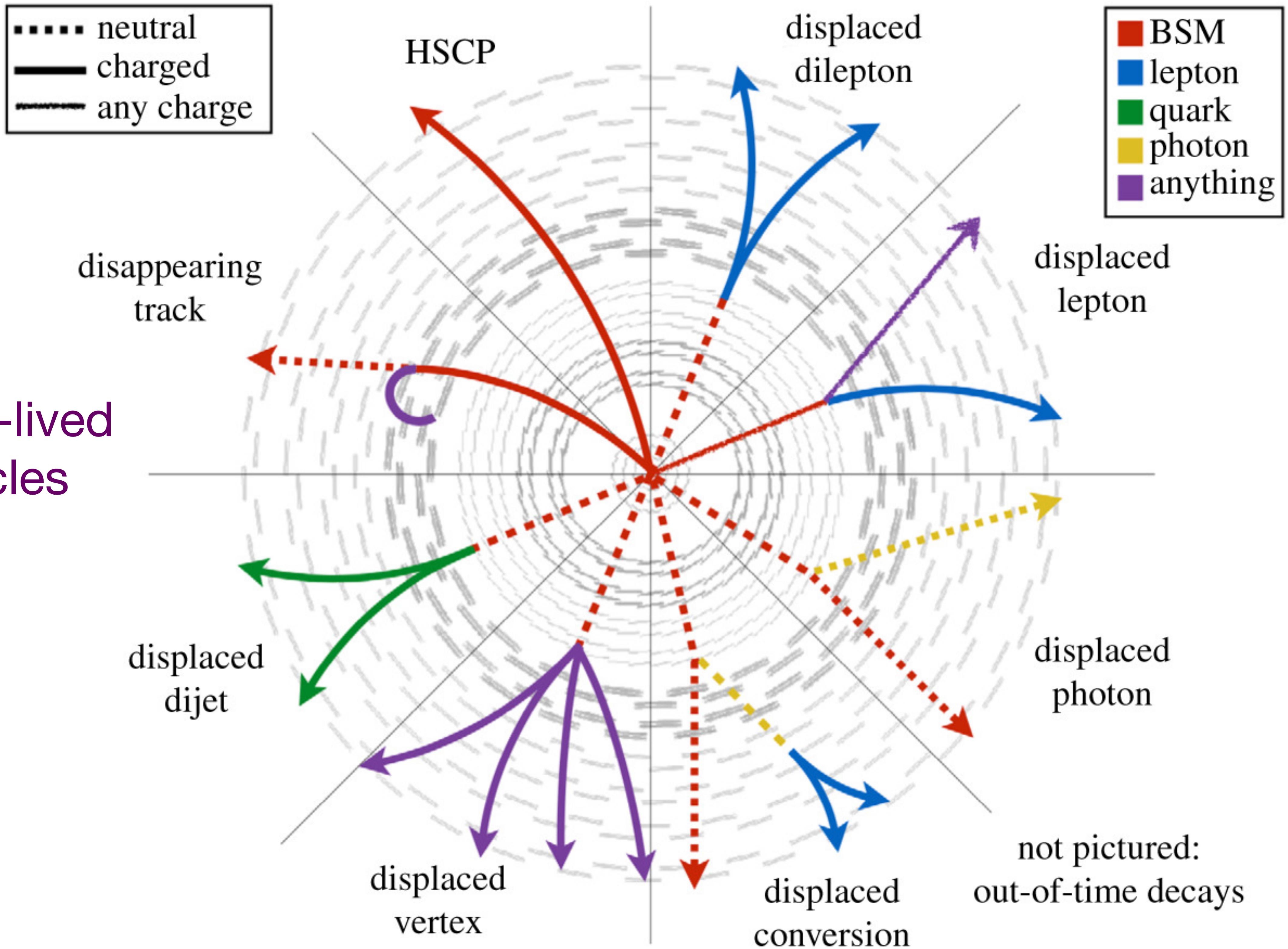
Most kinematic discriminators give an excesses in the tails (e.g. MET), but razor variables define a “bump”, hence they provide very good signal-BG discrimination.

For events with >2 visible objects, we partition these objects into 2 megajets, then compute M_R and R^2.



Long-lived particles

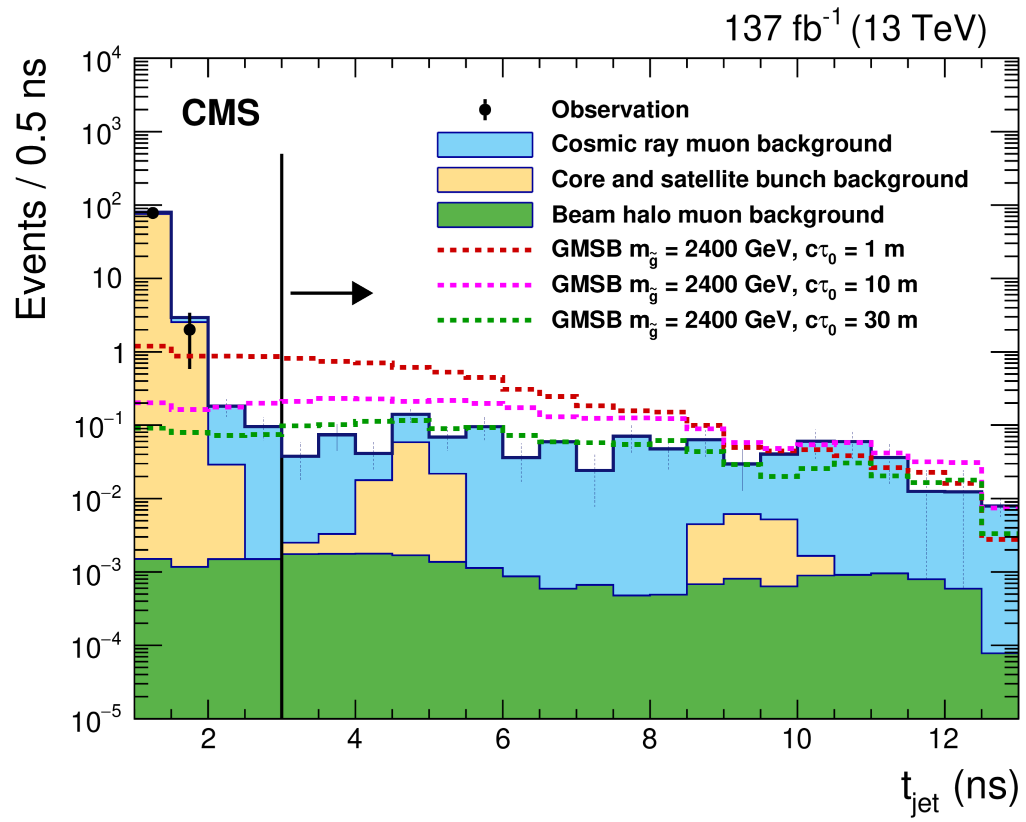
Long-lived particles

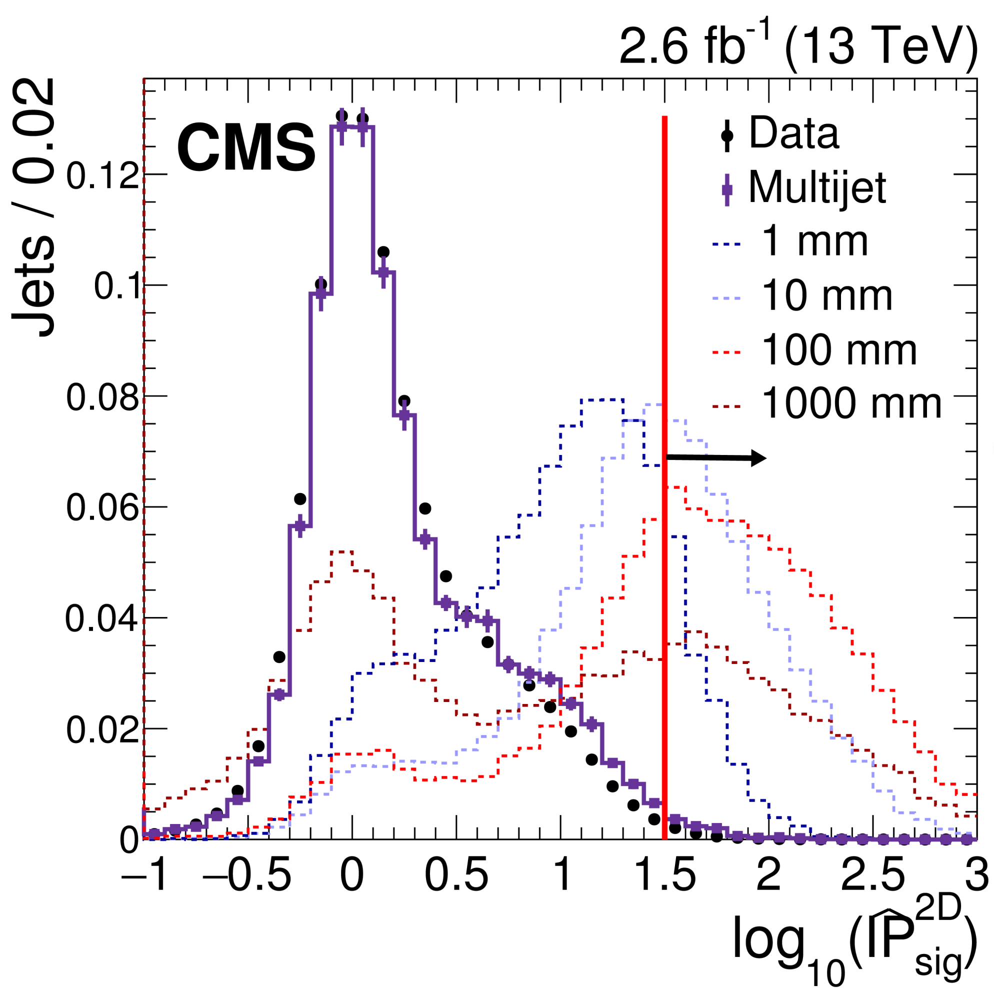


Characterizing the signal
Long-lived particles

Timing information of the object.

A long-lived BSM particle has bigger timing compared to SM particles.

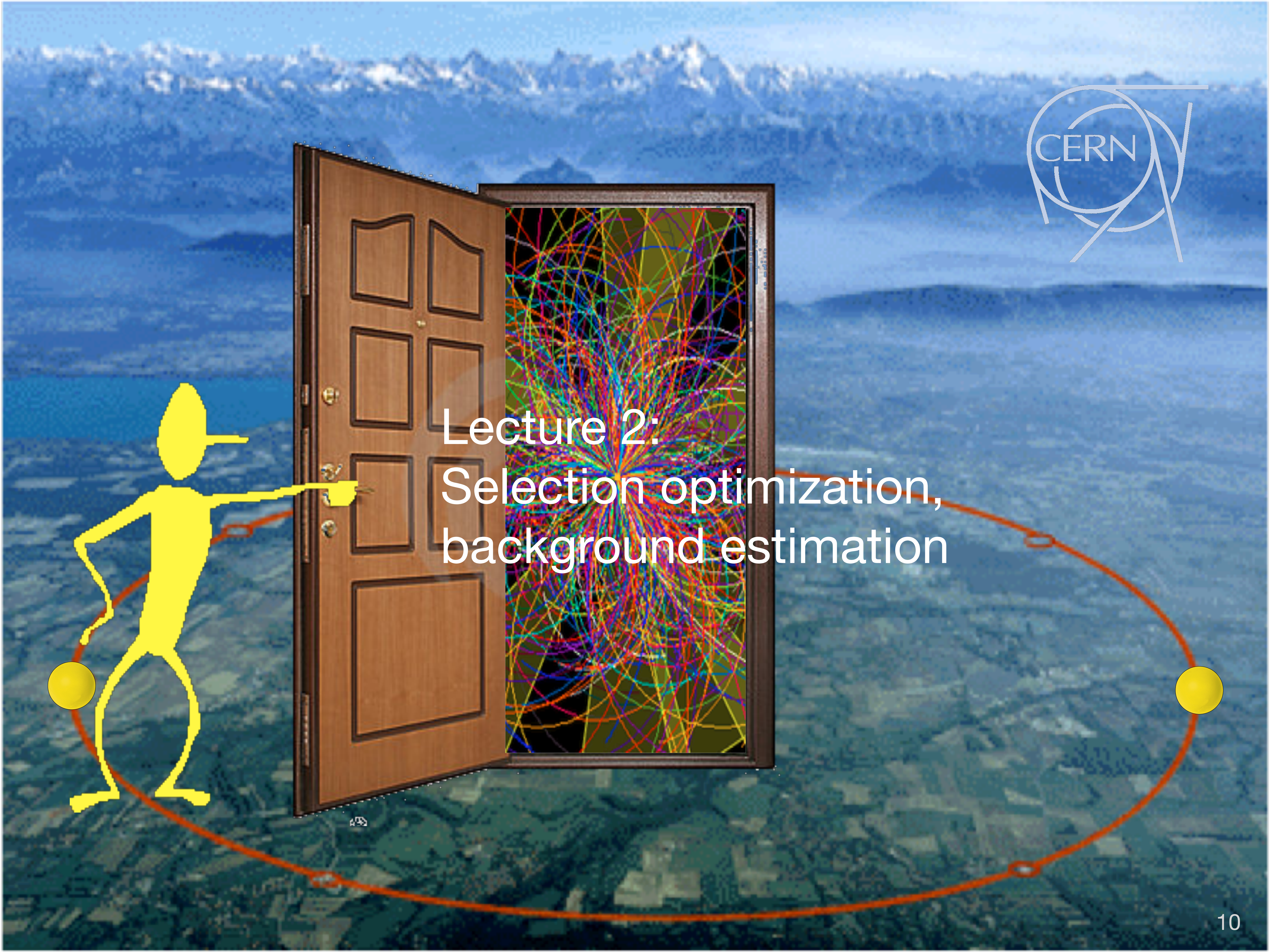




Displacement information of the object from the interaction point.

A long-lived particle can decay far away from the interaction point.

Lecture 2: Selection optimization, background estimation



Event selection and cutflows

- An event selection consists of a sequence of selections, i.e. cuts applied on event variables. Usually multiple event variables are used in an event selection.
  - e.g. to find a Z boson, first require 2 electrons or 2 muons, then make sure they have opposite electric charge, then calculate their invariant mass and require the value to be around the Z mass of 90 GeV, e.g. between 70 and 100 GeV.
- Cutflow: The sequence of cuts leading to a selection (and the number of events surviving them, or the selection efficiencies).
  - Selection efficiency: Number of events surviving the cuts over number of total events.
- Selection region / category: The phase space defined by a sequence of cuts.
- Signal region / category / search region: A selection region where signal can be observed with high significance.

Optimizing the selection

Event selection and cutflows

Example cutflow table from a CMS supersymmetry analysis:

Selection	$\mathrm{pp} \to \tilde{\mathrm{t}}\bar{\tilde{\mathrm{t}}}, \tilde{\mathrm{t}} \to \mathrm{t}\tilde{\chi}^0_1$ $m_{\tilde{\mathrm{t}}} = 950\,\mathrm{GeV}$ $m_{\tilde{\chi}^0_1} = 100\,\mathrm{GeV}$	$\mathrm{pp} \to \tilde{\mathrm{b}}\bar{\tilde{\mathrm{b}}}, \tilde{\mathrm{b}} \to \mathrm{b}\tilde{\chi}^0_1$ $m_{\tilde{\mathrm{b}}} = 1000\,\mathrm{GeV}$ $m_{\tilde{\chi}^0_1} = 100\,\mathrm{GeV}$	$\mathrm{pp} \to \tilde{\mathrm{q}}\bar{\tilde{\mathrm{q}}}, \tilde{\mathrm{q}} \to \mathrm{q}\tilde{\chi}^0_1$ $m_{\tilde{\mathrm{q}}} = 1400\,\mathrm{GeV}$ $m_{\tilde{\chi}^0_1} = 200\,\mathrm{GeV}$
$N_{\mathrm{jet}} \geq 2$	99.9 ± 0.2	98.8 ± 0.5	99.1 ± 0.5
$H_{\mathrm{T}} > 300\,\mathrm{GeV}$	98.7 ± 0.4	98.3 ± 0.5	98.9 ± 0.6
$H_{\mathrm{T}}^{\mathrm{miss}} > 300\,\mathrm{GeV}$	74.5 ± 1.2	79.6 ± 1.4	88.1 ± 1.4
$H_{\mathrm{T}}^{\mathrm{miss}} / H_{\mathrm{T}} \leq 1$	73.6 ± 1.3	78.2 ± 1.4	86.8 ± 1.5
$N_{\mathrm{muon}} = 0$	58.7 ± 1.4	77.9 ± 1.4	86.7 ± 1.5
$N^{(\mathrm{muon})}_{\mathrm{isolated\ tracks}} = 0$	58.2 ± 1.4	77.8 ± 1.4	86.7 ± 1.5
$N_{\mathrm{electron}} = 0$	47.2 ± 1.4	77.5 ± 1.5	86.4 ± 1.5
$N^{(\mathrm{electron})}_{\mathrm{isolated\ tracks}} = 0$	46.4 ± 1.4	77.2 ± 1.5	86.2 ± 1.5
$N^{(\mathrm{hadron})}_{\mathrm{isolated\ tracks}} = 0$	45.5 ± 1.4	76.8 ± 1.5	85.6 ± 1.5
$N_{\mathrm{photon}} = 0$	43.8 ± 1.4	75.2 ± 1.5	83.6 ± 1.6
$\Delta\phi_{H_{\mathrm{T}}^{\mathrm{miss}}, j_1} > 0.5$	43.6 ± 1.4	75.1 ± 1.5	83.5 ± 1.6
$\Delta\phi_{H_{\mathrm{T}}^{\mathrm{miss}}, j_2} > 0.5$	41.1 ± 1.4	70.6 ± 1.6	78.7 ± 1.7
$\Delta\phi_{H_{\mathrm{T}}^{\mathrm{miss}}, j_3} > 0.3$	39.8 ± 1.4	67.0 ± 1.6	74.4 ± 1.8
$\Delta\phi_{H_{\mathrm{T}}^{\mathrm{miss}}, j_4} > 0.3$	38.5 ± 1.4	64.5 ± 1.6	71.4 ± 1.9
Event quality filter	36.7 ± 1.4	61.4 ± 1.7	67.8 ± 1.9

Optimizing the selection

What is optimization?

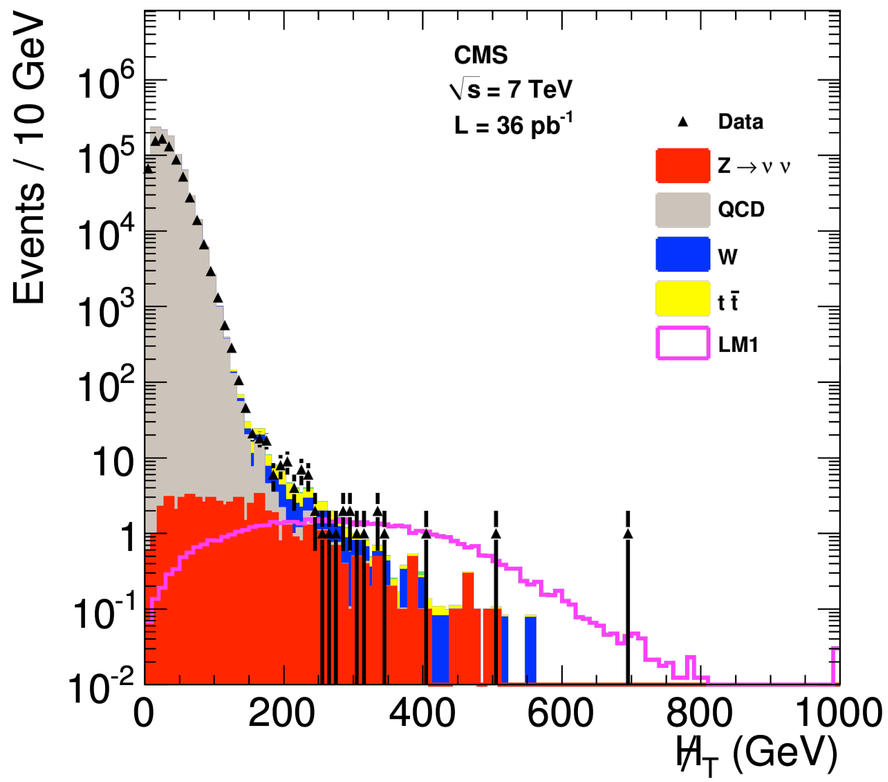
- Optimization of a selection involves finding the best selection (best cutflow) out of all possibilities which leads to the best sensitivity.
- Sensitivity: The capability of an analysis to observe a given physics process.
e.g. This analysis is sensitive to supersymmetric particles with mass 3 TeV.
  - Sensitivity implies high expected signal significance.
- Significance: A measure of the probability of rejecting the null hypothesis (i.e. background). (Formal definition is more complex, but we won't go into it here.).
  - A commonly used simple approximation:

$$s = \frac{N_{signal}}{\sqrt{N_{signal} + N_{background}}}$$

  - But more formal methods are used in real analyses.
- Optimization involves finding the selection that gives the best significance for a reasonable amount of data as well as results in the least amount of uncertainties.
- Optimization methods: "by eye", random grid search, machine-learning-based, etc.

Optimizing the selection

Rectangular cuts “by eye”



Missing hadronic transverse momentum:

$$\not{H}_T = H_T^{miss} = -\sum_i^{n\,jets} \vec{p}_T^{\,jet_i}$$

This one looks easy, doesn’t it?

Somewhere around 300 GeV?

The original CMS SUSY analysis used $H_T^{miss} > 250$ GeV

Hadronic transverse energy:

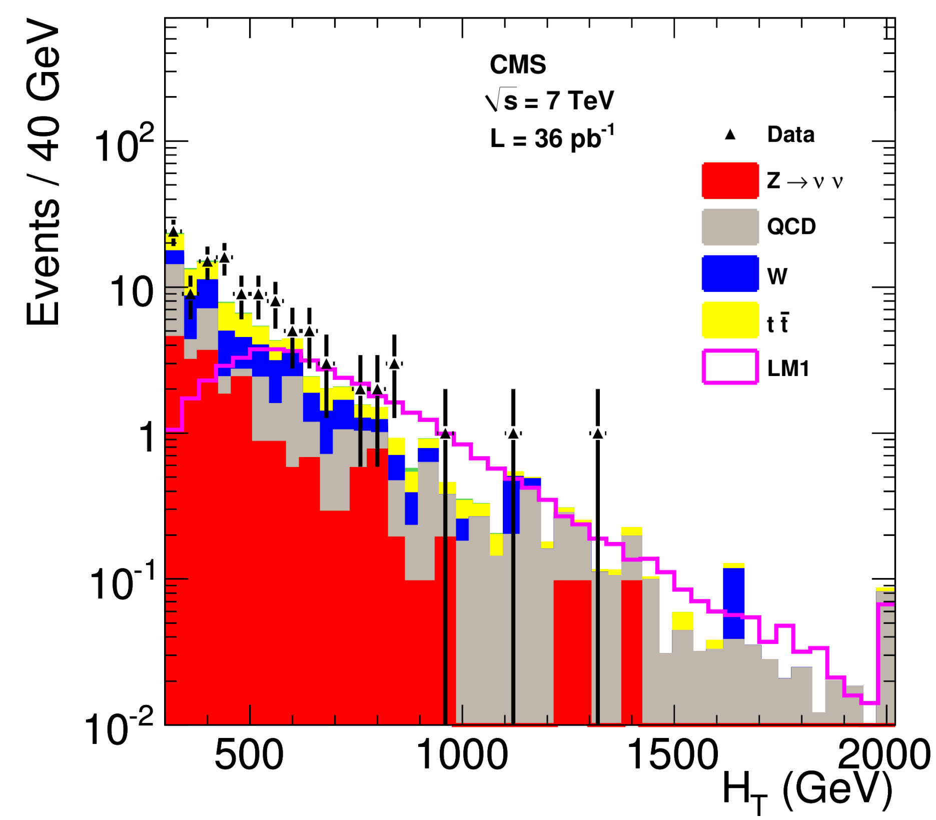
$$H_T = \sum_i^{n\,jets} p_T^{jet_i}$$

How about this one? Not so obvious…

The original analysis used $H_T > 500$.

Maybe we should try several random H_T values and find the H_T that gives the best significance?

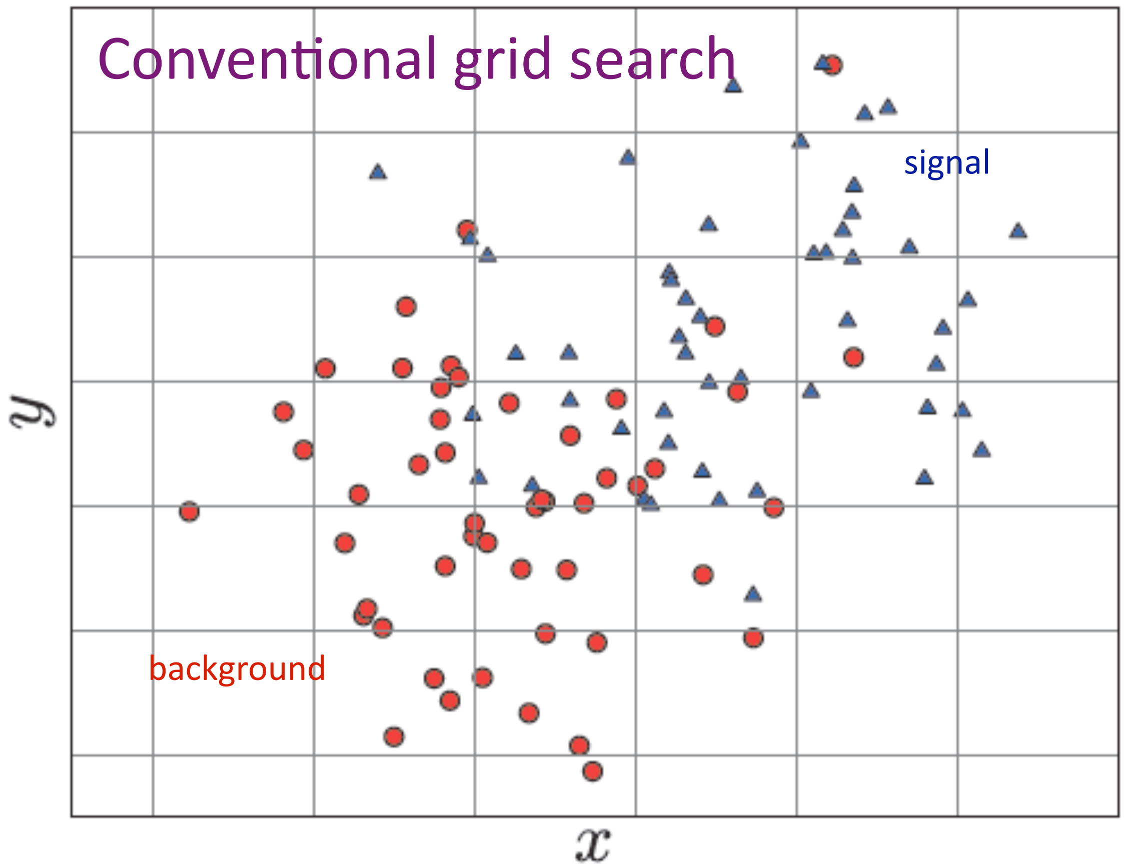
But what if we have many selection variables?

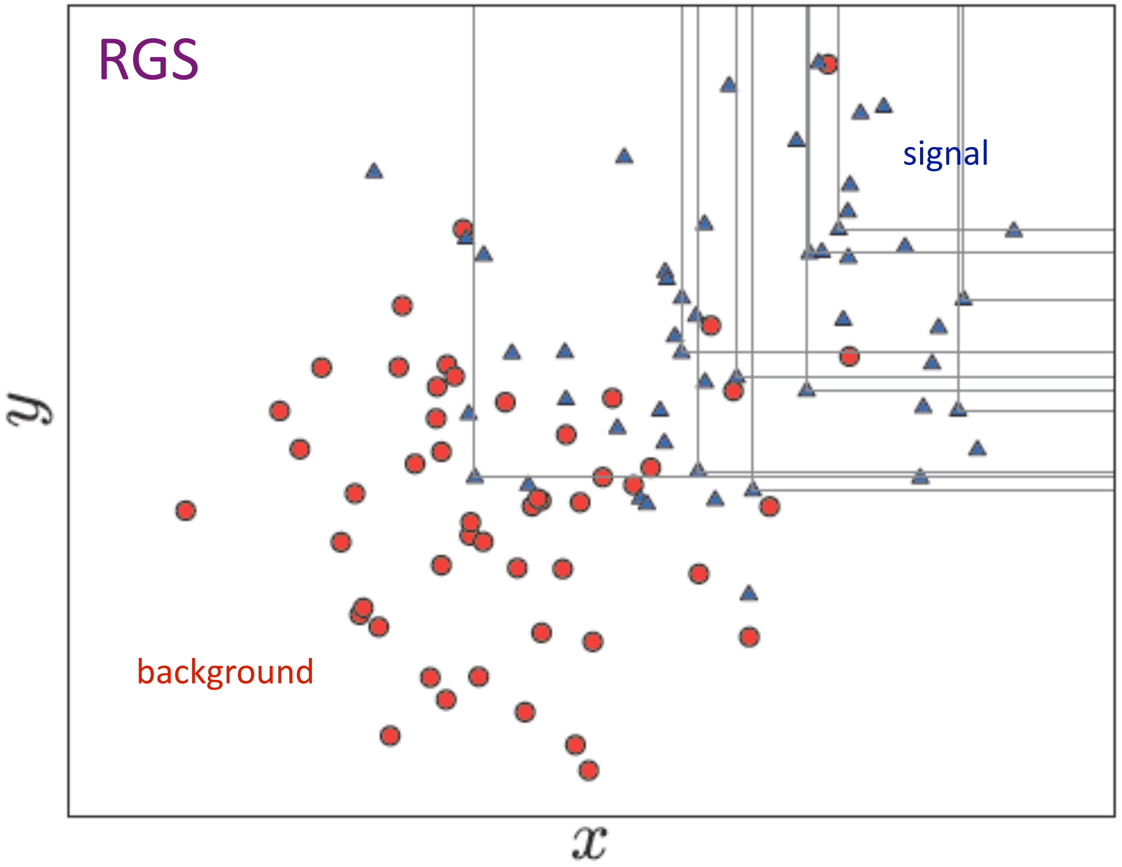


Optimizing the selection

Rectangular cuts by "Random grid search"

- Efficient sampling for rectangular cut optimization.
- We would like to find a selection that characterizes the signal final state.
- Most natural way to do this is to use the signal events themselves as candidate cut sets (i.e. use values for each cut variable in each signal event as cut candidates).
- Random Grid Search (RGS) tries every cut set, implements the selection, and finds the selection that is most optimal (e.g. maximizes significance, etc.).
- Easily generalized to all types of cuts (interval, box, staircase, etc.)
- Becomes very efficient for optimization over multiple parameters.



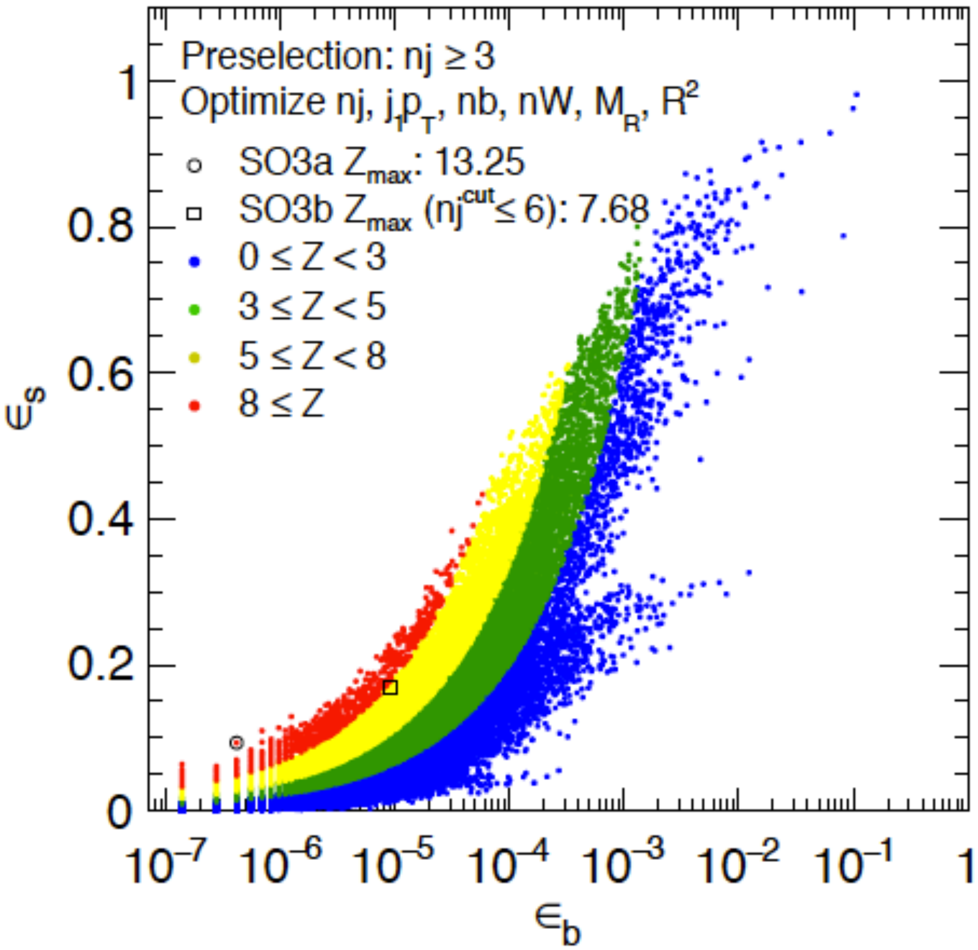


P. C. Bhat, H. B. Prosper, S. Sekmen, C. Stewart, arXiv:1706.09907

Optimizing the selection

Rectangular cuts by “Random grid search”

2-dimensional cuts optimization using RGS



Each colored point corresponds to one candidate selection.

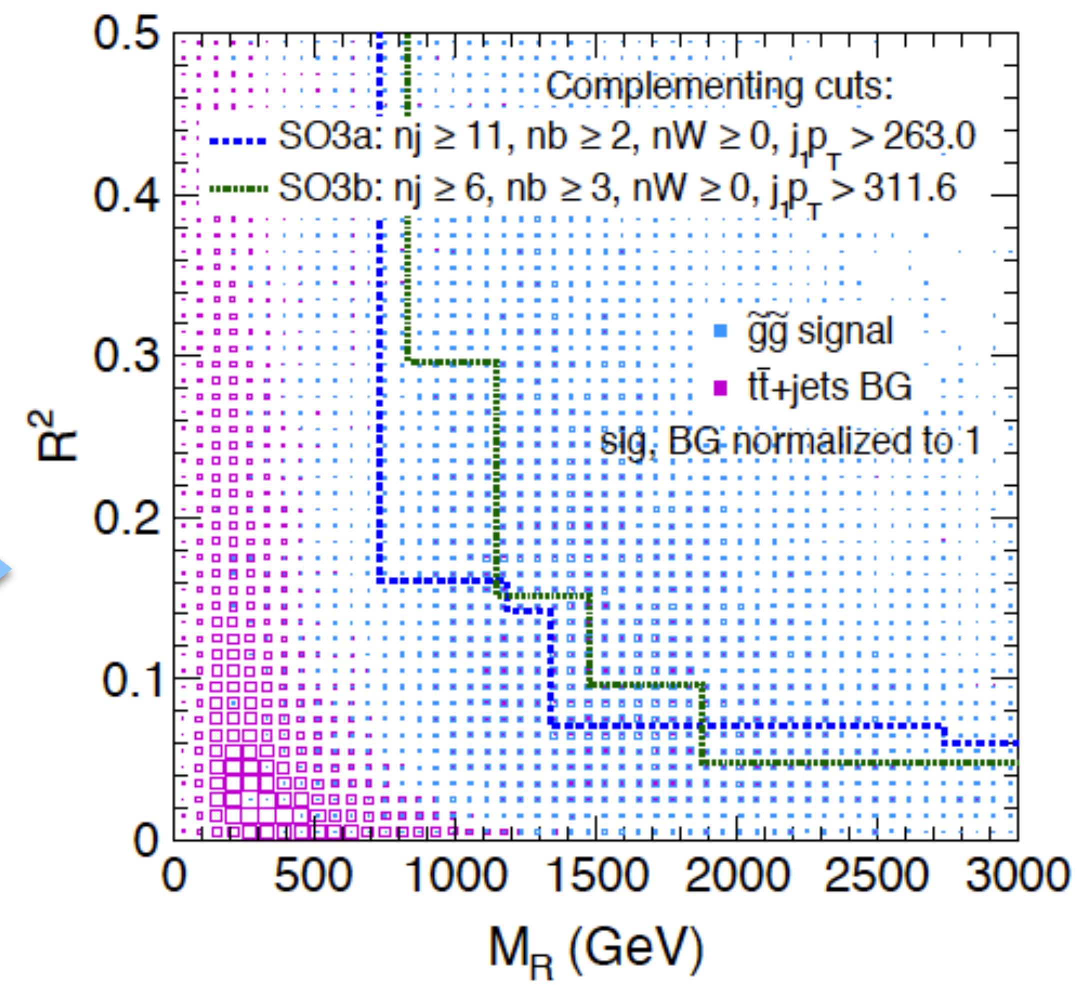
Different colors show significance values. Red is the best.

Optimal selections found to be SO3a, SO3b.

2-dimensional space of variables.

SO3a and SO3b selections are plotted in blue and green lines.

They are shown to effectively separate signal and background



P. C. Bhat, H. B. Prosper, S. Sekmen, C. Stewart, arXiv:1706.09907

Machine learning methods

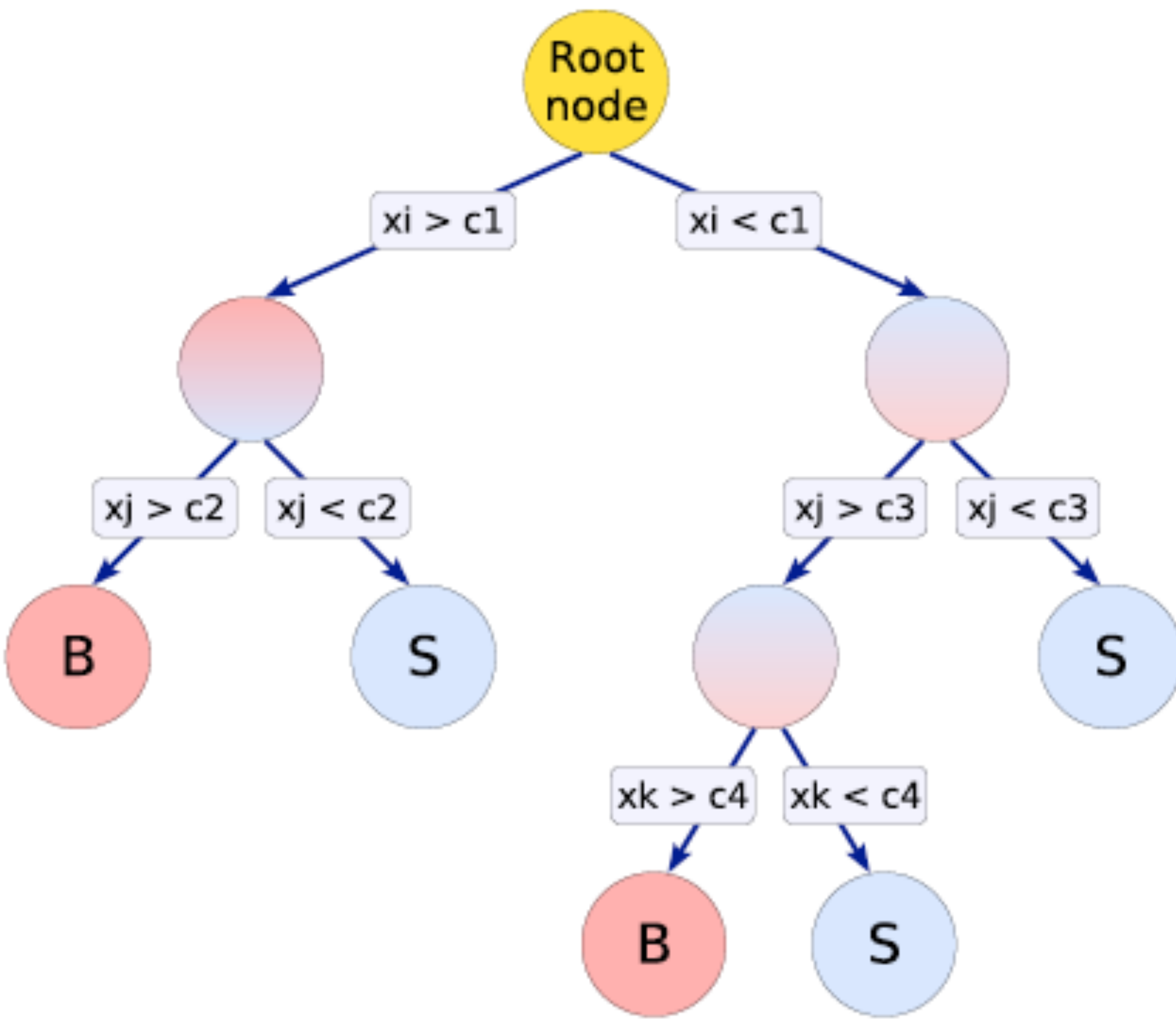
Machine learning methods are very useful in getting optimal event selections.

- Classification methods are used to categorize events into various groups, e.g. signal or background.
- They are also used in classifying objects, by obtaining the best identification criteria for objects. They are used for separating b-jets from light jets, separating boosted particles from non-boosted particles, separating long-lived particles from promptly decaying particles, etc.
- Traditional methods like boosted decision trees have been used for years, and greatly helped in discoveries like single top quark and the Higgs boson.
- Nowadays neural networks, even deep neural networks are becoming more mainstream, also due to wider availability of GPUs.
- ML methods are especially useful for cases with small signals buried under large backgrounds. When rectangular cuts do not yield sufficient sensitivity, they are applied to extract the utmost sensitivity.
- They are recently tested for generic searches for anomalies / signals in data.

Regression methods are also widely used, for robust measurements of quantities such as energy, mass, etc. …. BUT this is outside the scope of event selection.

ML: Decision trees

- A decision tree is a binary tree, a sequence of cuts paving the phase-space of the input variables.
- Repeated yes/no decisions on each selection variable is taken for an event until a stop criterion is fulfilled. Each node splits the data according to one attribute.
- For each variable, find the splitting value that gives the best separation.
- Trained with labeled data to maximize the the probability of assigning random events correctly as signal or background.

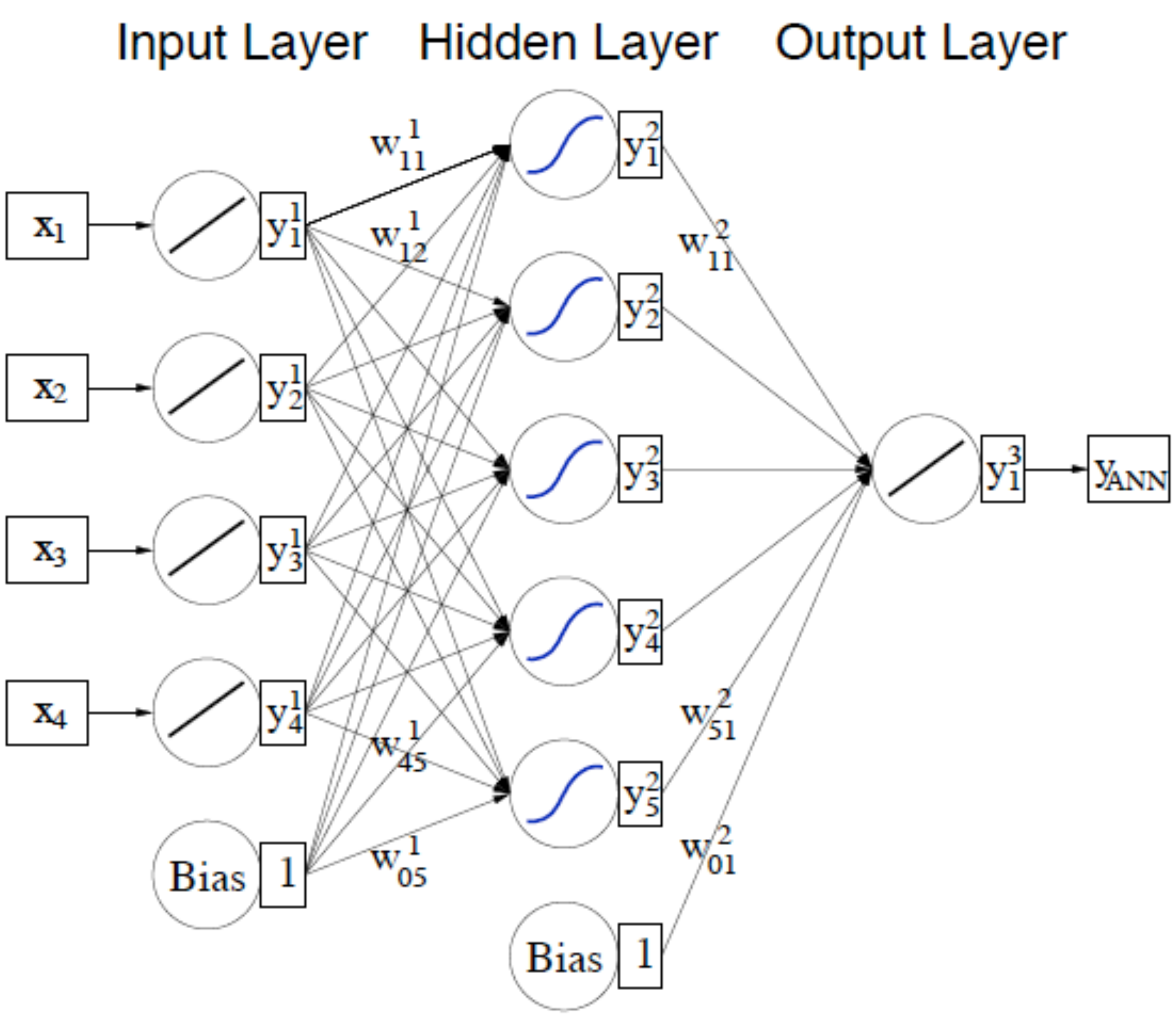


- Similar to rectangular cuts, but each selection depends on the previous one. Selection sequence effects the result.
- Boosting: Combine information from multiple trees.

Optimizing the selection

ML: Neural networks

- Inspired by the brain – neural networks are composed of “artificial neurons”.
- Approximates and outputs a discriminator which quantifies how signal-like the events are.

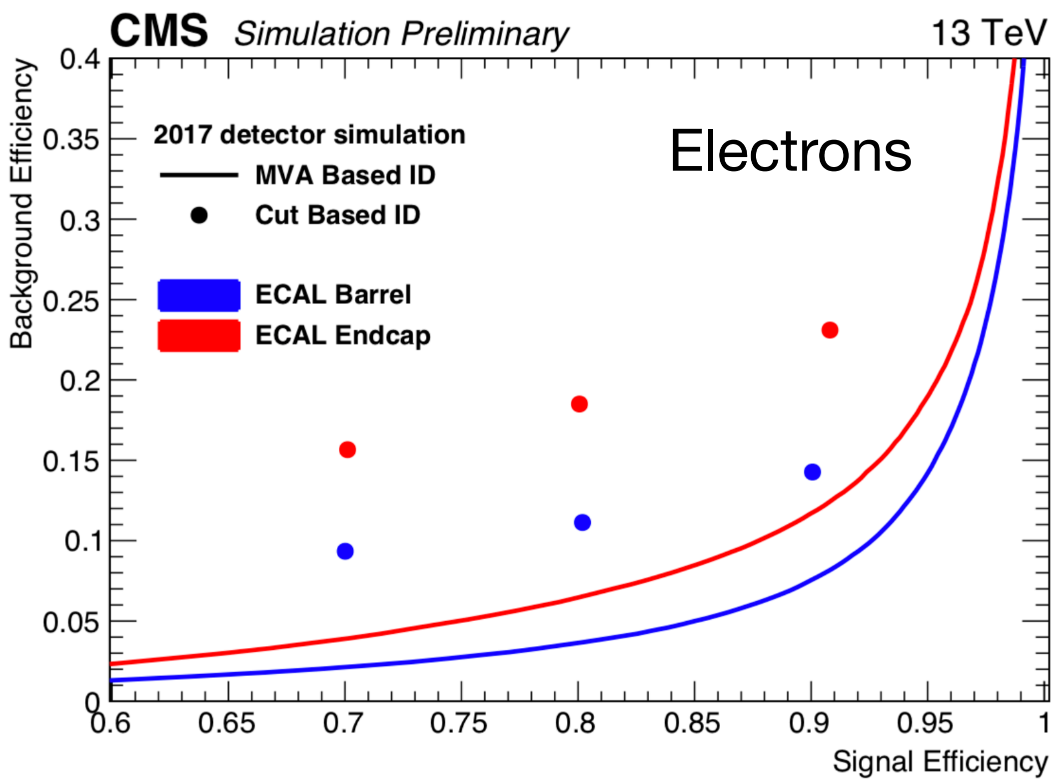


The recipe:

- Start from a set of input variables fed to the input layer
- For each neuron in the hidden layer, compute a weighted sum of the input variables.
- Transform the output with an activation function
- Repeat the operation for each neuron of the next hidden layer
- Output is a weighted sum (average) of the previous input layers.

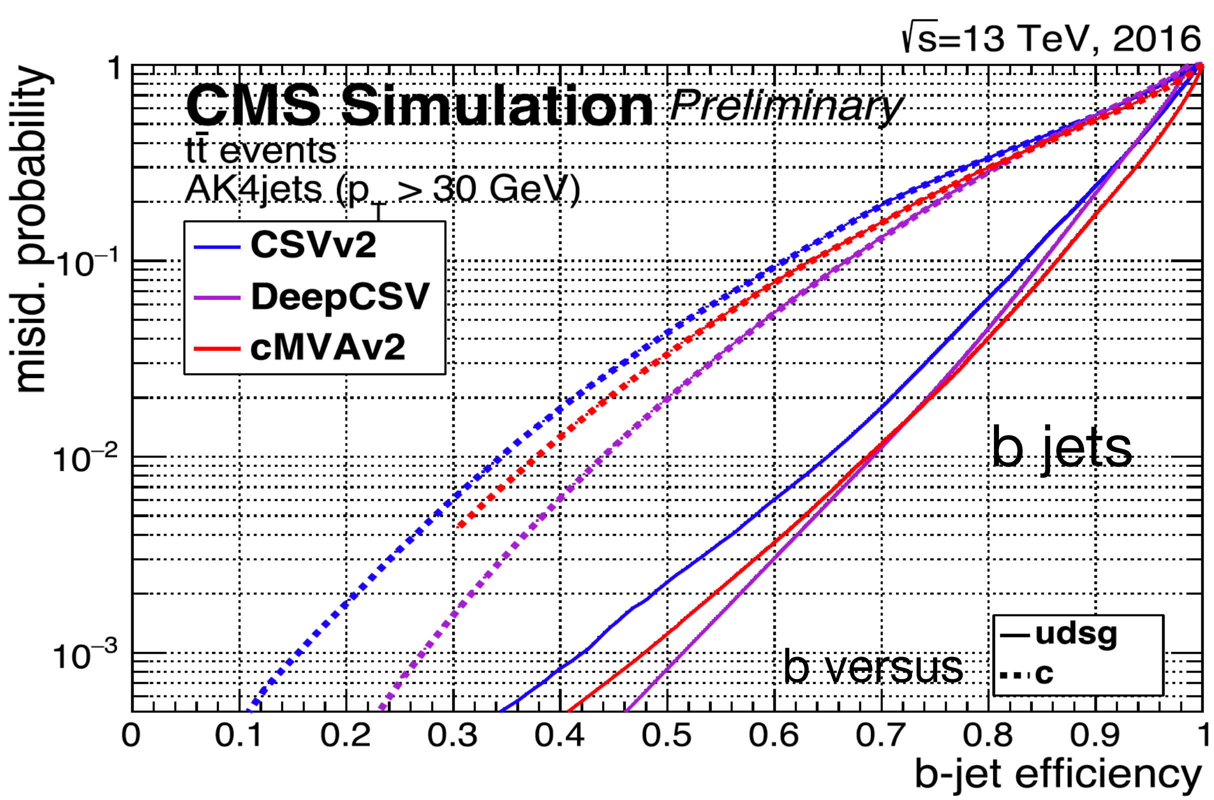
Optimizing the selection

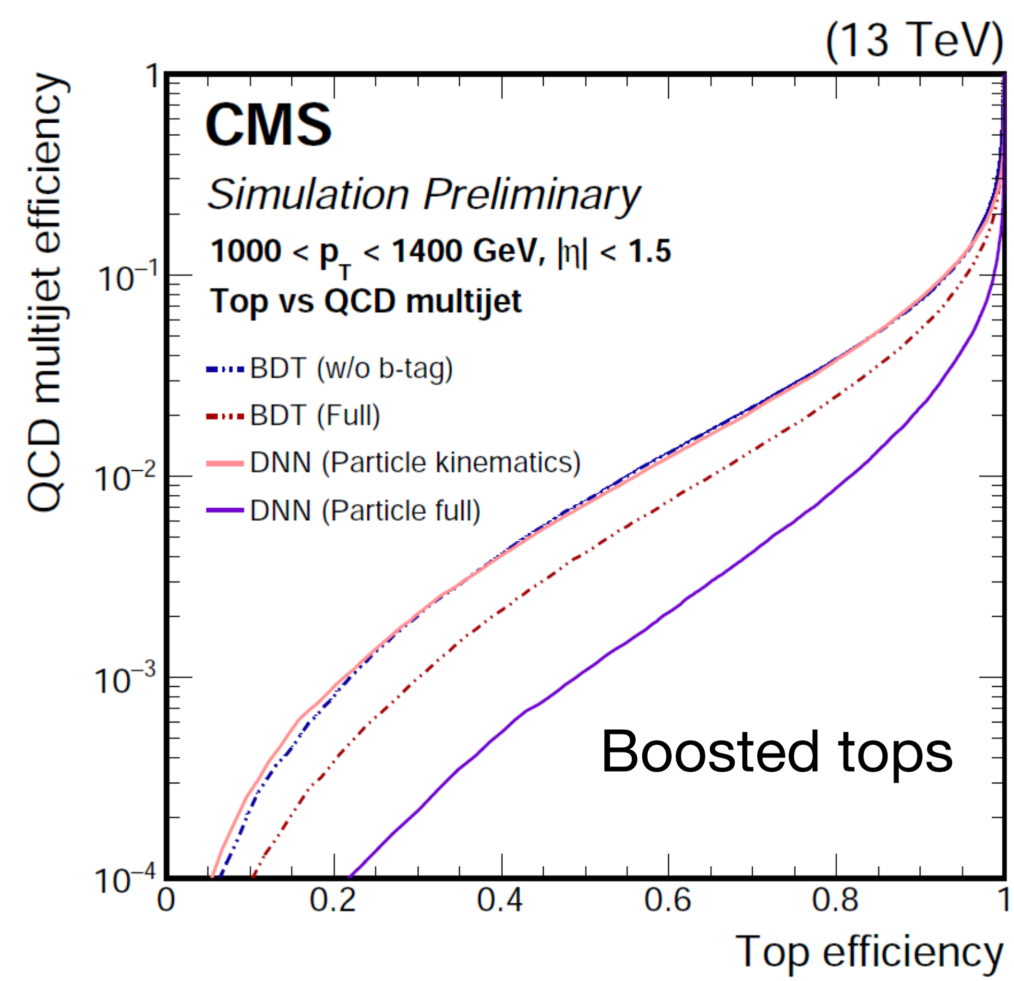
ML: Object classification



ML methods are used for object identification and classification. They provide better performance than cut based identification.

Better classified objects allow better signal characterization, optimization and measurement.



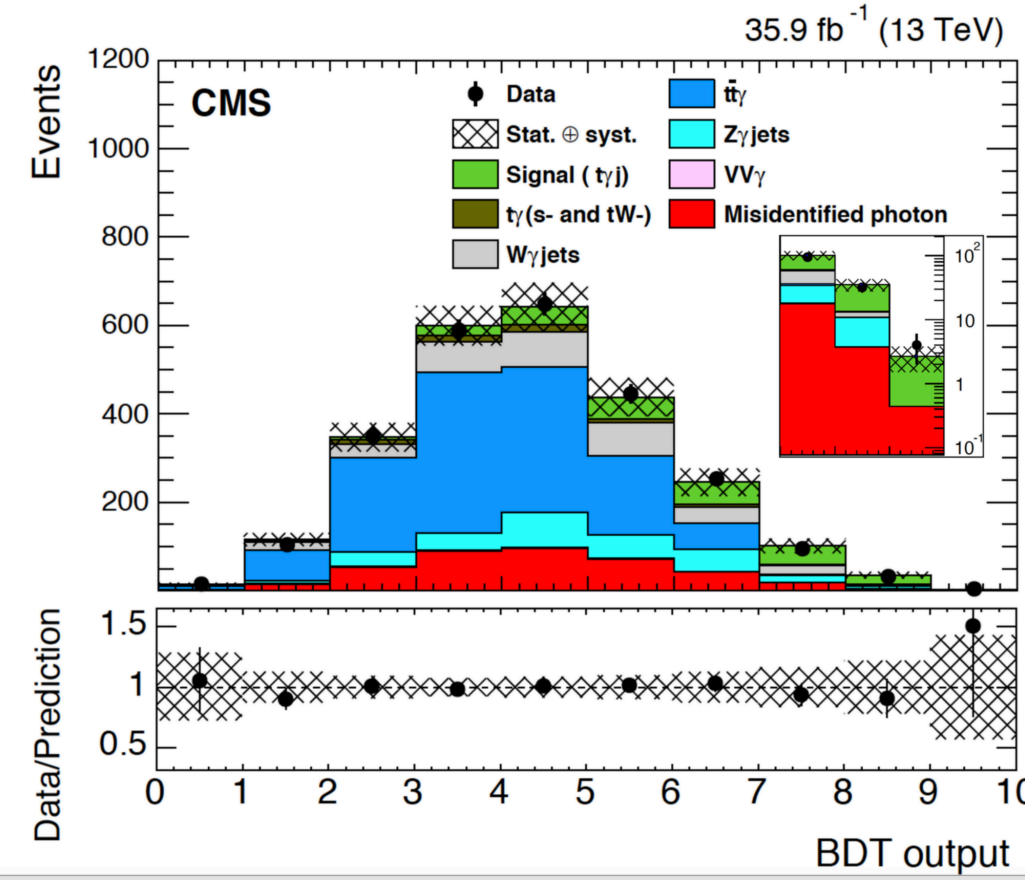


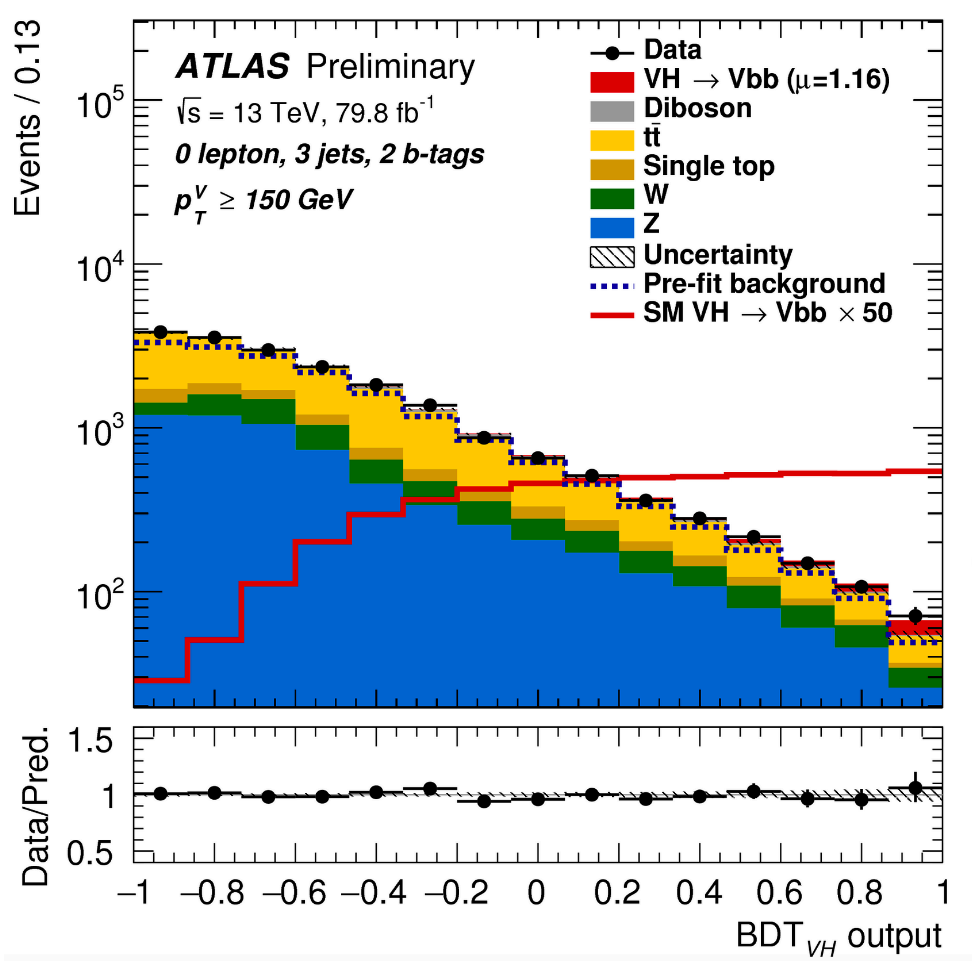
Optimizing the selection

ML: BDTs in searches

CMS used BDTs to observe and make measurements on the single top quark production.

Fun fact: Single top observation at Tevatron D0 (which used both BDTs and neural networks) was a historical analysis marking the recognition of these methods in HEP.





ATLAS observed Higgs in the decay channel of H —> bb in 2018 thanks to BDTs.

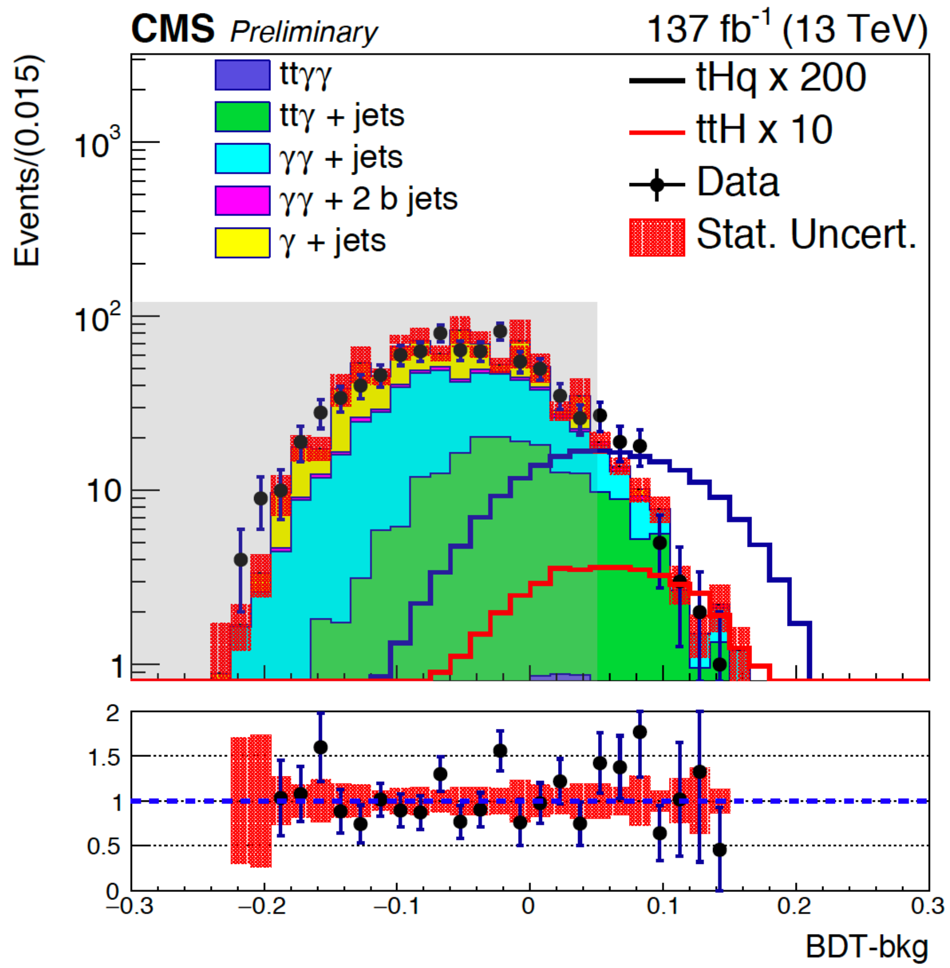
BDTs were trained for different selections and their results were combined.

Optimizing the selection

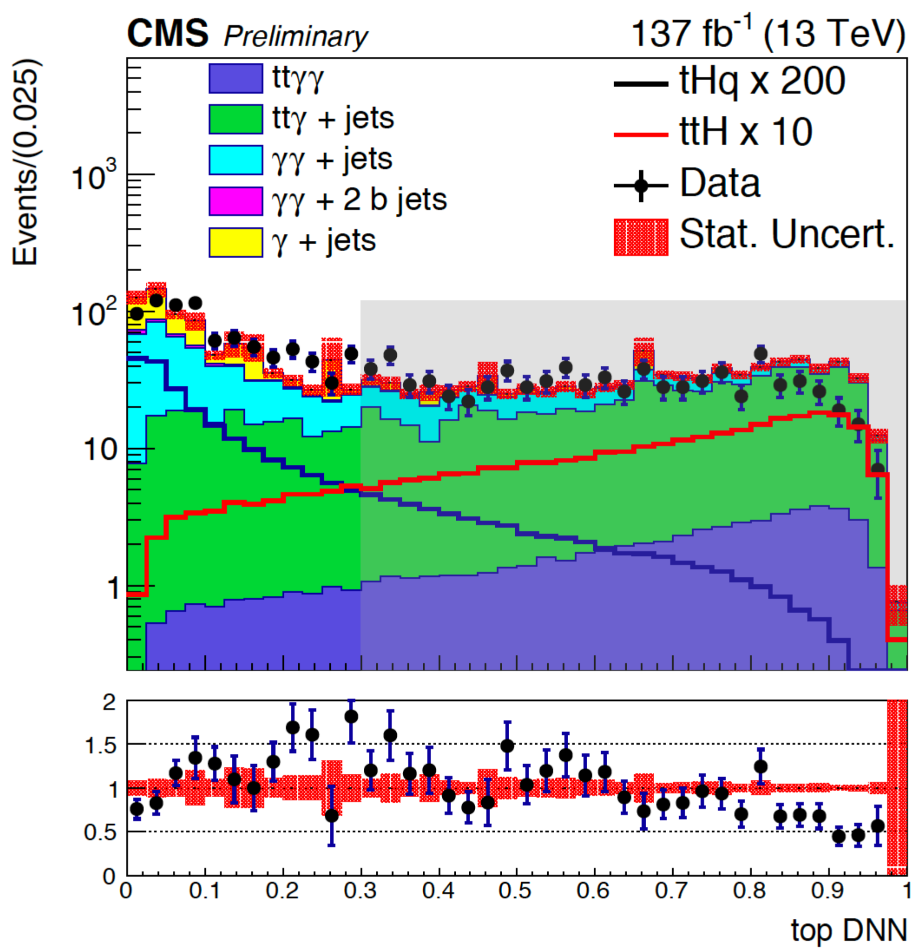
ML: DNNs in searches

DNN example: CMS analysis measuring Higgs properties in Higgs production in the ttH channel, with H —> γγ.

Higgs production in tHq channel has a similar final state to ttH, and also constitutes a background we must get rid of. Trained 2 DNNs for this purpose



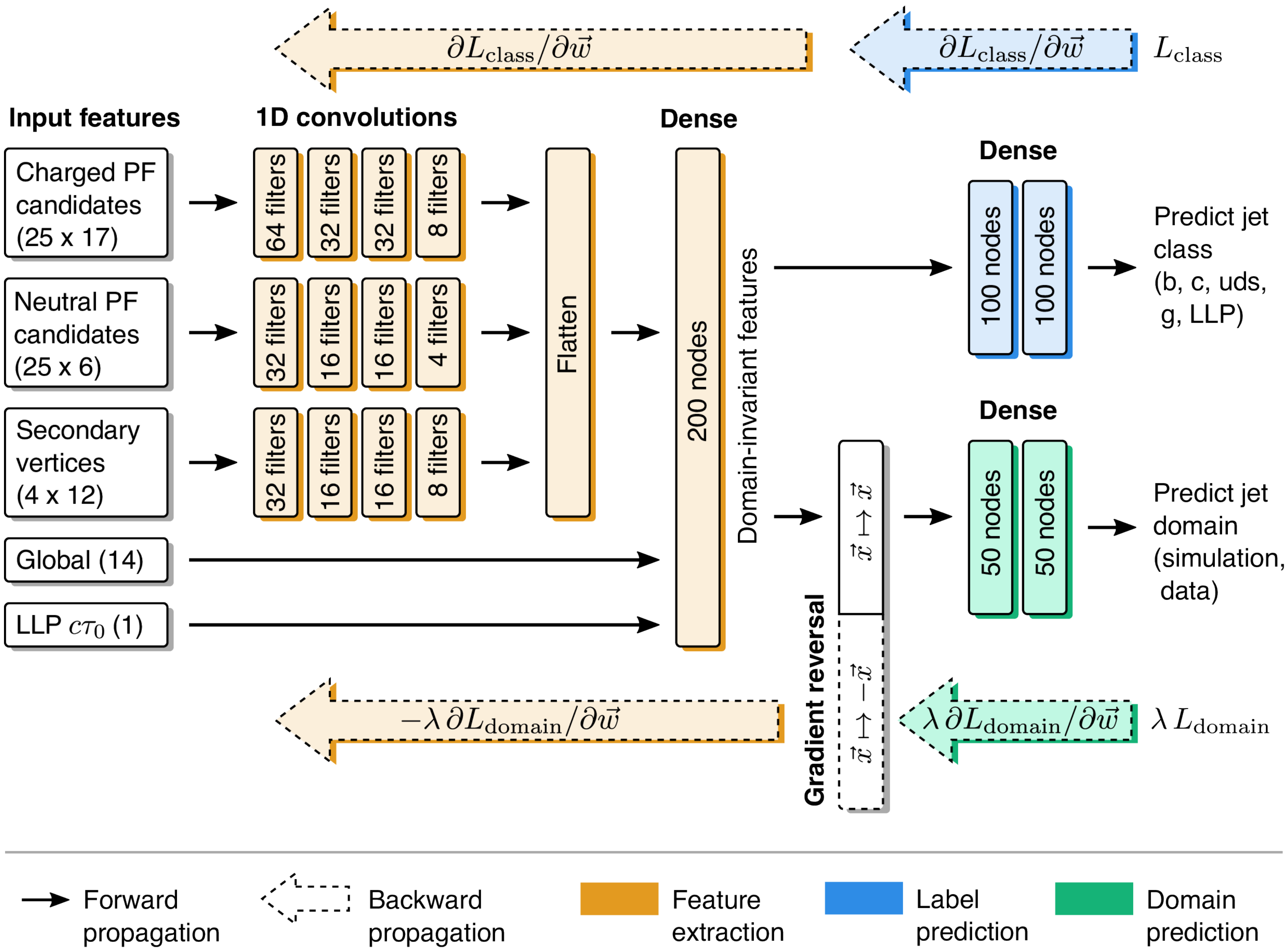
DNN to discriminate ttH + tHq from the rest of the backgrounds.



DNN to discriminate ttH from tHq and the rest of the backgrounds.

ML: DNNs in searches

CMS analysis searching for new long-lived particles decaying to jets uses a jet classification DNN. Here is how the architecture looks:



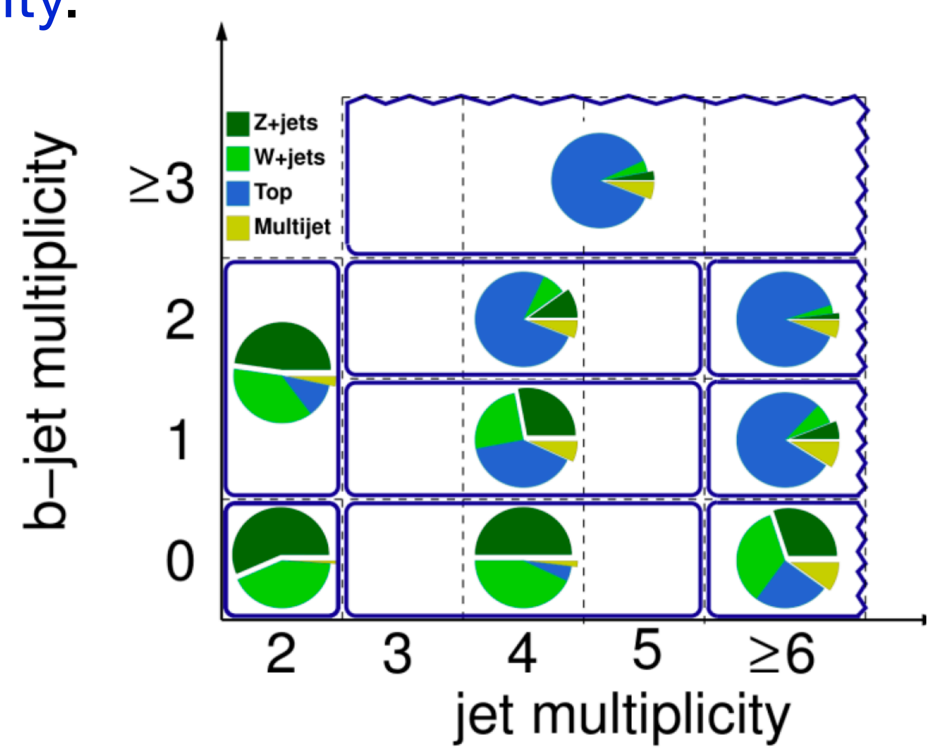
CMS EXO-19-011, arXiv:1912.12238

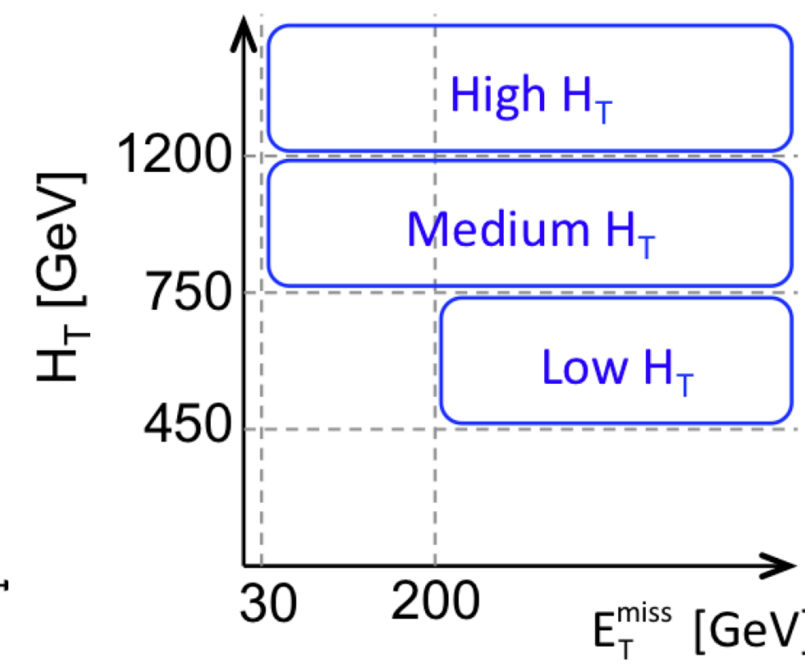
Optimizing the selection
Multiple regions

An analysis usually consists of multiple selection regions. Why?

- SM measurements: Design signal regions for
  - different production/decay channels (e.g. different Higgs production channels)
  - to focus on different kinematic properties (e.g. boosted top vs. non-boosted top)
- New physics searches: New physics models have (multiple) free parameters
  —> new particle properties like masses, branching ratios, etc. are variables
  —> design dedicated signal regions to cover different particles and all signatures with highest sensitivity.

Definition of signal regions from a CMS SUSY search looking for different SUSY particles: gluinos, squarks, stops, sbottoms.





Different multiplicities dedicated to different sparticles, or different decay model. Colors show BG composition

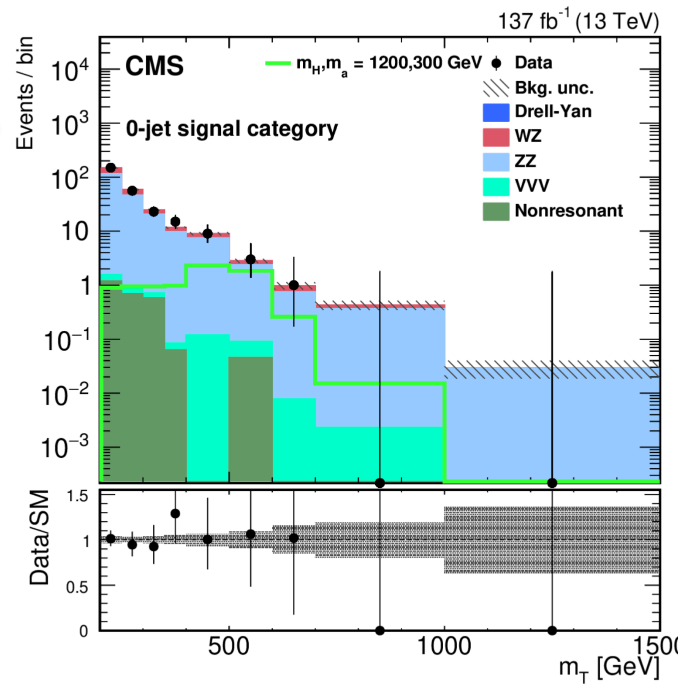
Different regions dedicated to different SUSY particle masses or decay kinematics.

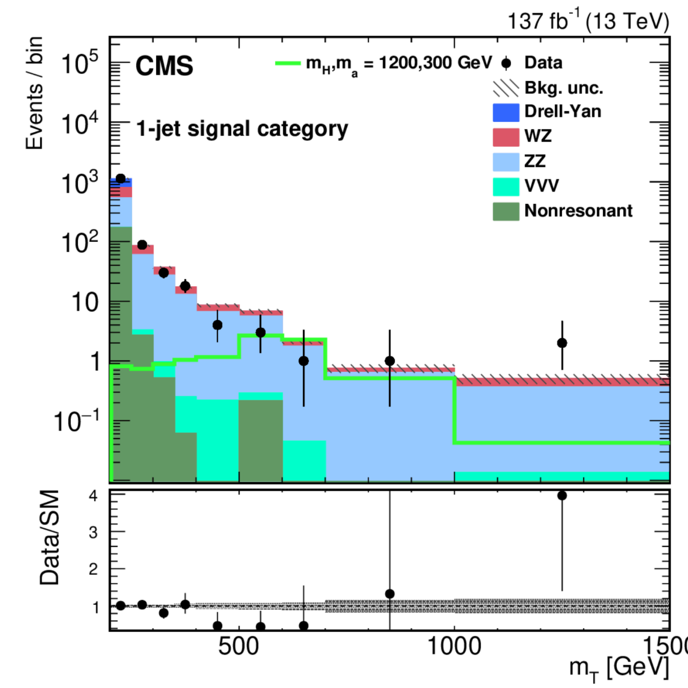
Optimizing the selection
Analysis variables, bins

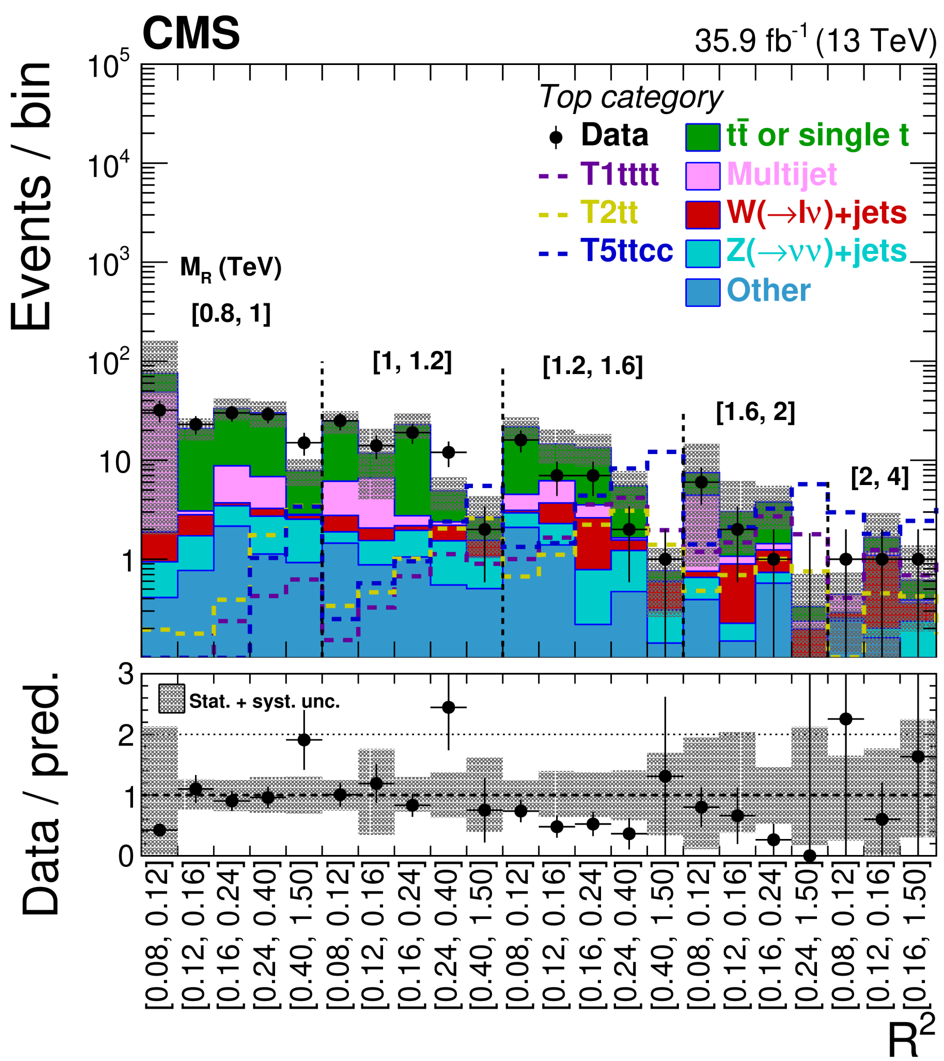
Once the signal region selections are done, we must decide which variable(s) we will use for testing the signal hypothesis with data.

Usually, these variables are divided into bins, i.e. discrete intervals.

m_T bins in two signal regions in a CMS dark matter search.







2-dimensional binning in razor variables M_R and R^2 in a CMS SUSY analysis.

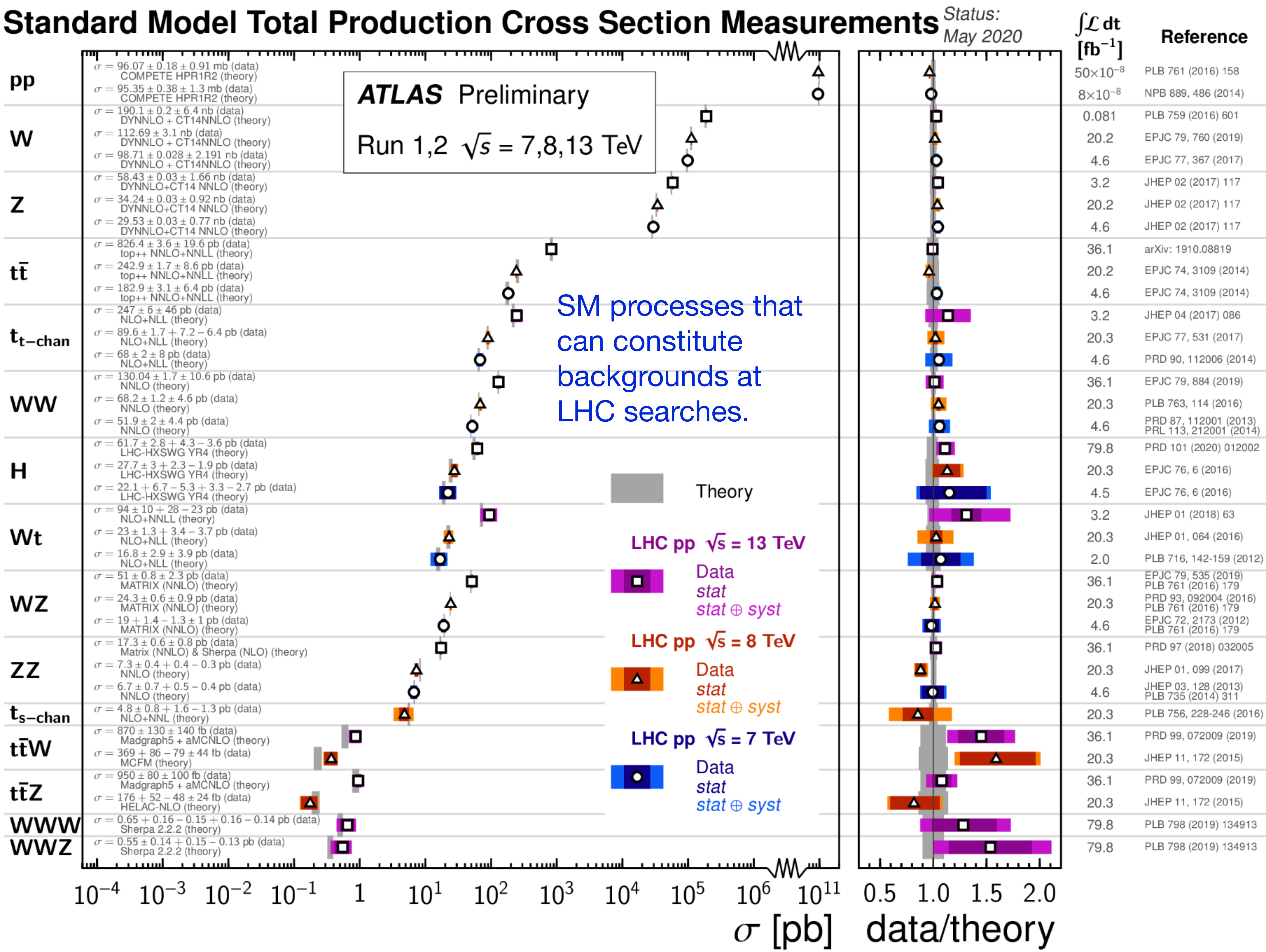
The analysis has 8 signal regions defined by jet, b-jet, lepton, W and top multiplicities.

The plot is shown for the boosted top signal region.

(5) BACKGROUUUNNNNDDSSSS!

Background estimation

SM backgrounds measured at LHC



SM processes that can constitute backgrounds at LHC searches.

Generic idea

We need to estimate the amount (and shape) of the irreducible backgrounds remaining in the signal region after signal selections.

This is a crucial part of analysis. Numerous methods exist and still being devised.

Use predictions from Monte Carlo simulation:

- Contains all our knowledge on theory and detector.
- We precisely know what physics MC events have.
- Long but persistent way from roughness to precision.

Use data-driven estimation methods:

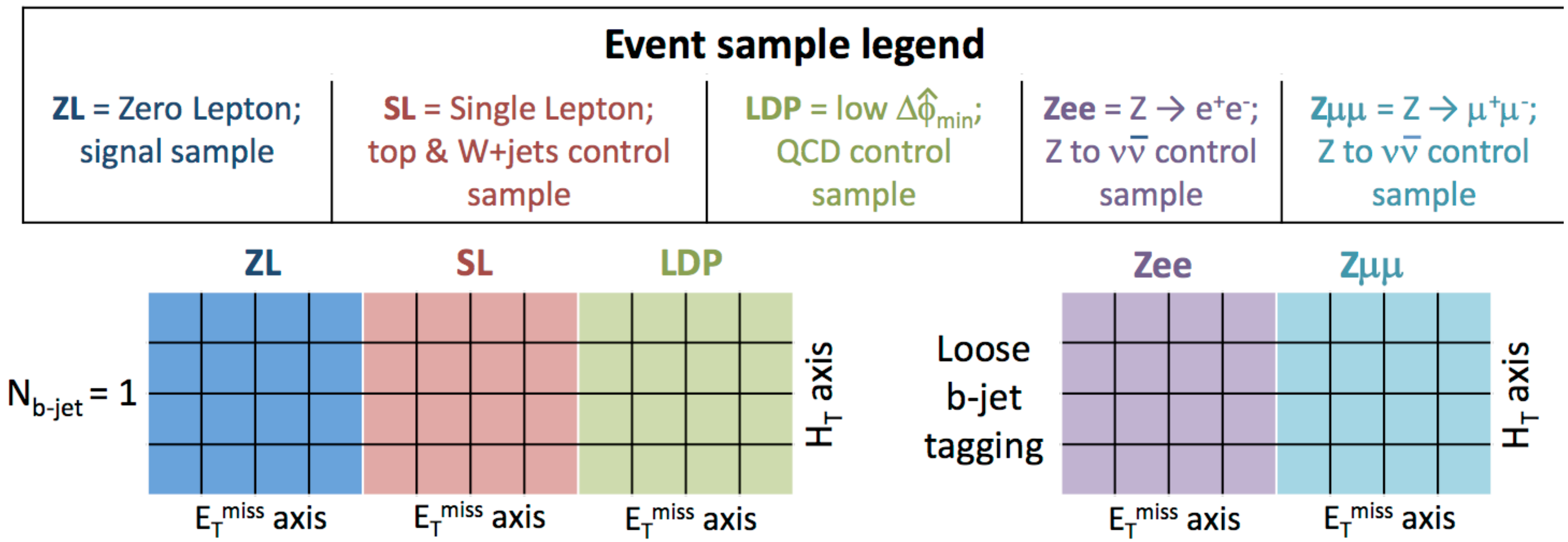
Common principle: use control regions

- Control region: A selection where background of interest is dominant while signal and other backgrounds are negligible.
- Must be disjoint from / not correlated with the signal region.
- Obtain information on BG from the control region and extrapolate it to the signal region.

Data and MC can work together:

- Data is used for fine-tuning MC.
- MC shapes of kinematic variables are used in data-driven methods.

Using control regions and MC ratios

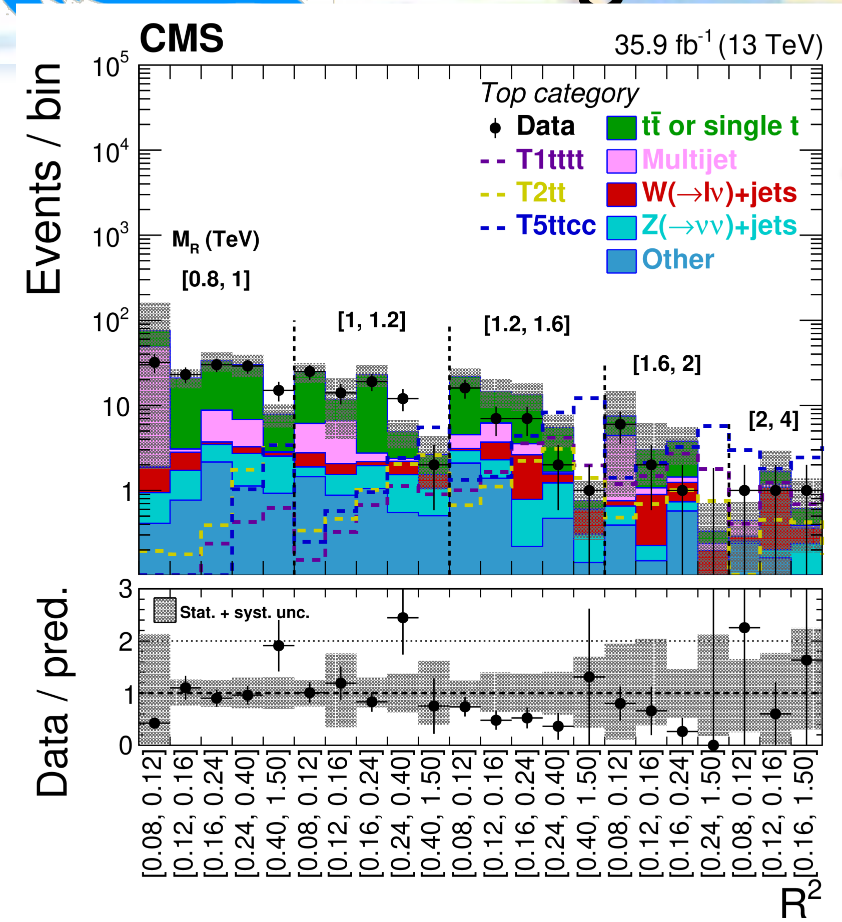


- Find control regions by reverting some of the signal region selection criteria.
- Find the amount of BG in every bin i in the control region.
- Then multiply this amount with BG expectation ratio between signal rand control regions obtained from MC:

$$N_{\mathrm{BG}}^{\mathrm{SR,\,i,\,estm}} = \frac{N_{\mathrm{BG}}^{\mathrm{SR,\,i,\,MC}}}{N_{\mathrm{BG}}^{\mathrm{CR,\,i,\,MC}}} \cdot N_{\mathrm{BG}}^{\mathrm{CR,\,i,\,data}}$$

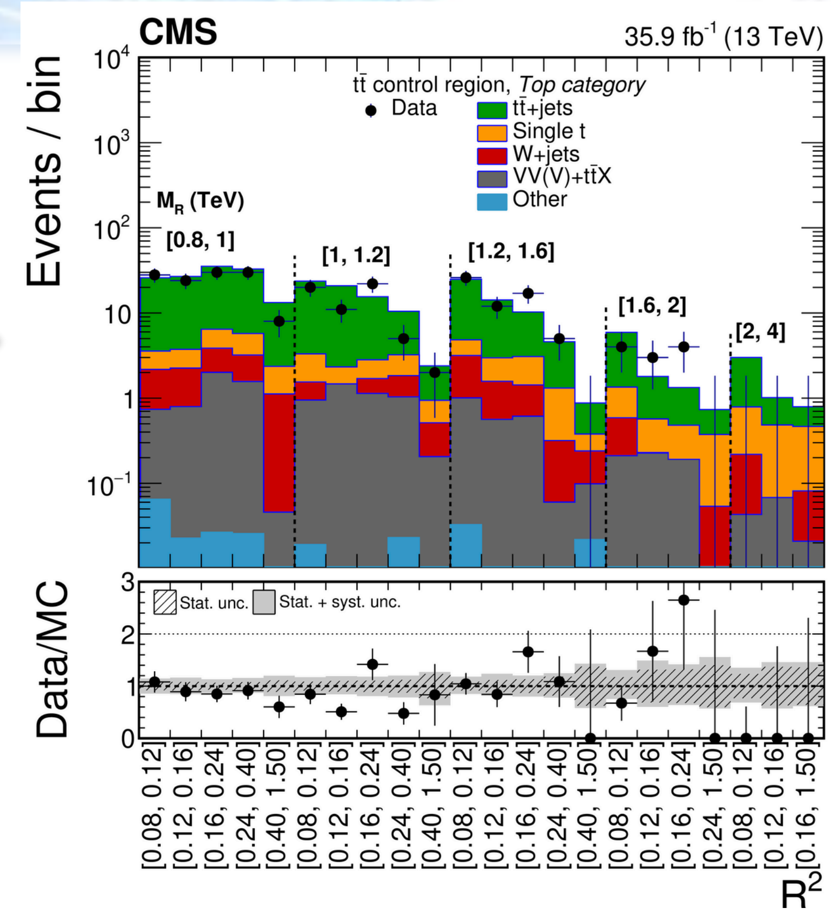
Background estimation

Using control regions and MC ratios

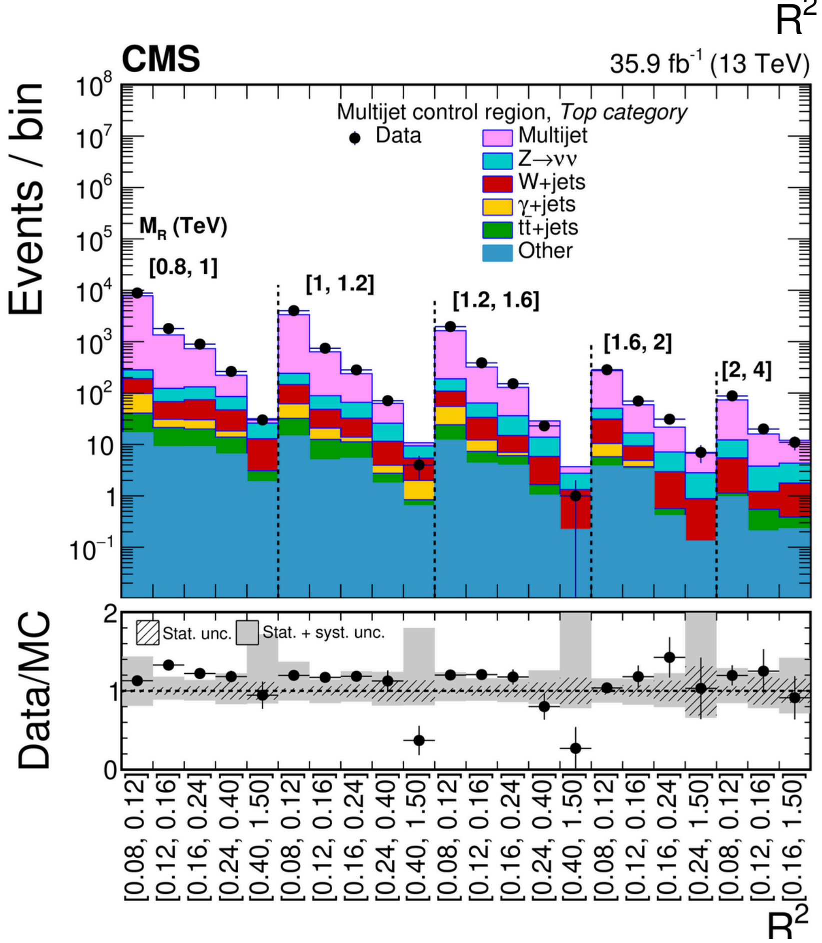


0 lepton signal region.

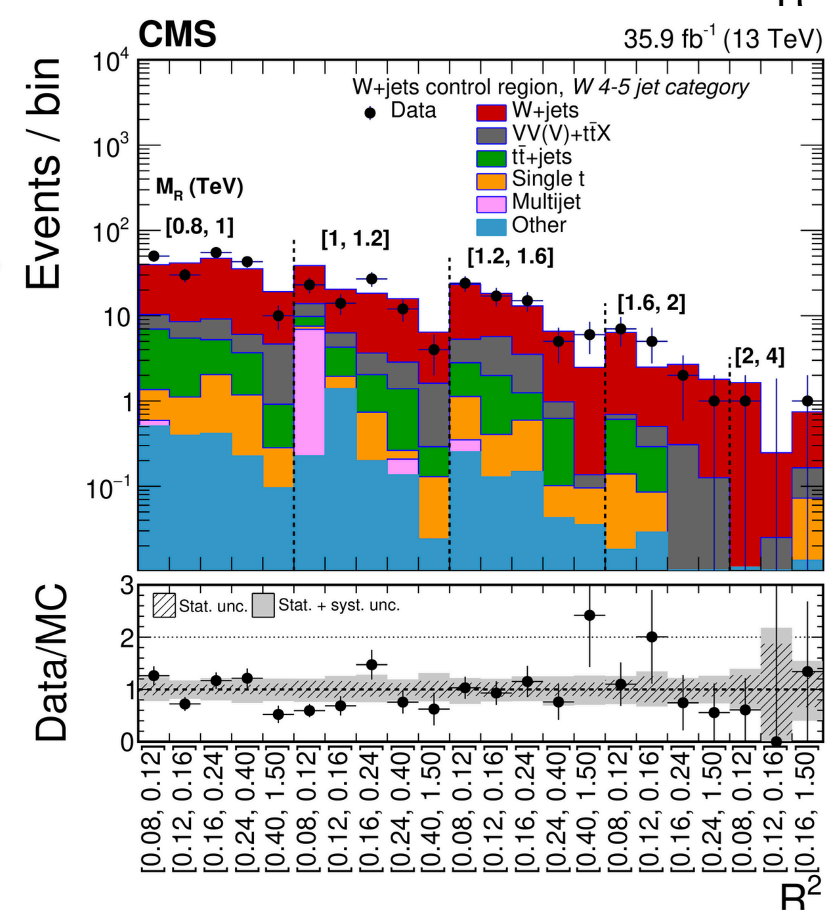
Single lepton + b jet control region for tt+jets



Signal and control regions in the CMS SUSY analysis with razor variables.



Single lepton + 0b jet control region for W+jets



Reverted $\Delta\phi_{min}$ QCD control region.

Replacing particles: Z →νν from Z →l^+l^-

Z(→νν)+jets is an irreducible BG for hadronic searches that use high MET. However there is no straightforward control region where Z(→νν)+jets is dominant.

But we can use the Z(→l^+l^-)+jets events to estimate the BG contribution from Z(→νν) +jets, since Z→νν and Z→l^+l^- events have the same kinematic characteristics.

- Select a l^+l^- events in a control region with l^+l^- invariant mass in the Z mass range (we assume this control region is signal-free).
- Count the leptons as MET, i.e.: add lepton momenta to MET and recalculate MET.
- Apply the MET cut and count the observed events.
- Z(→νν)+jets can be estimated as:

$$\mathrm{N}_{\mathrm{Z}\nu\nu}^{\mathrm{SR,\,i,\,estm}} = \frac{\mathrm{N}_{\mathrm{Z}\ell\ell}^{\mathrm{SR,\,i,\,MC}}}{\mathrm{N}_{\mathrm{Z}\ell\ell}^{\mathrm{CR,\,i,\,MC}}} \cdot \mathrm{N}_{\mathrm{Z}\ell\ell}^{\mathrm{CR,\,i,\,data}} \cdot \frac{\mathrm{BR}(\mathrm{Z}\to\nu\nu)}{\mathrm{BR}(\mathrm{Z}\to\ell\ell)}$$

ratio of branching ratios

- The estimate is corrected by the ratio of Z —> νν / Z —> ll BRs in order to have a correct estimate for the yield.

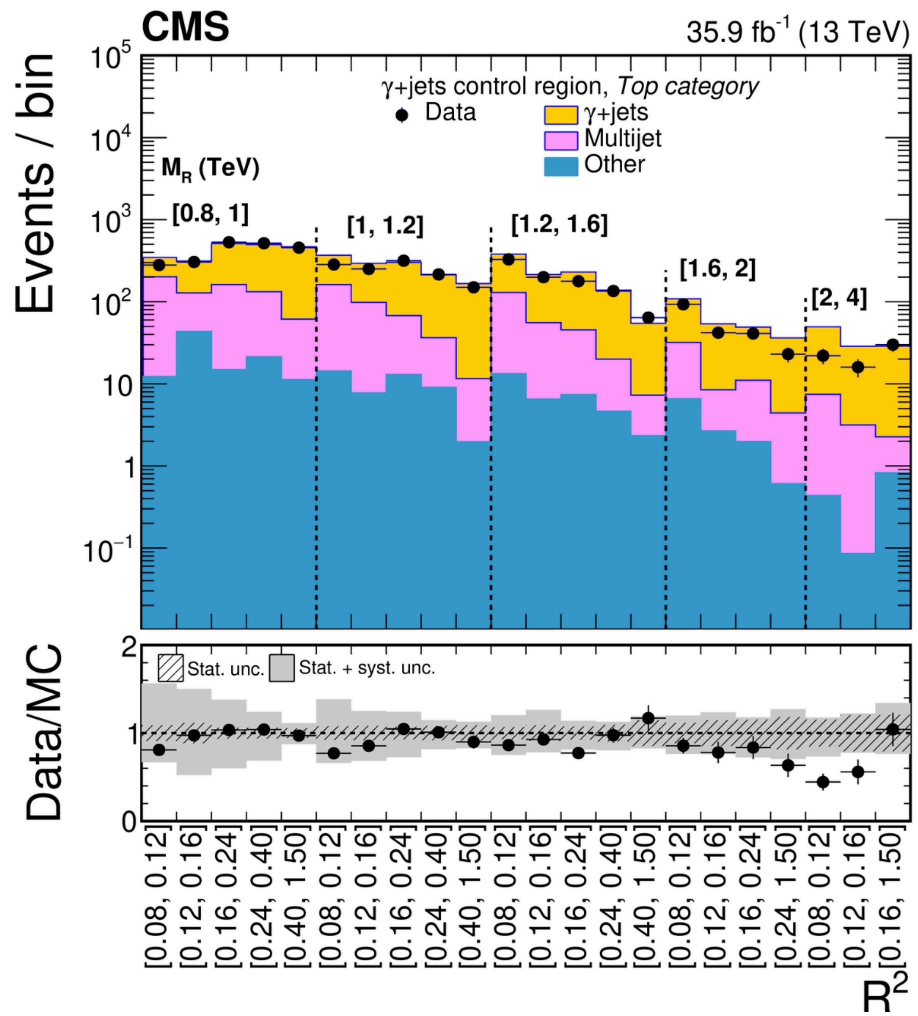
Replacing particles: Z →νν from Z →l+l-

Estimating Z(→νν)+jets from Z(→l+l-)+jets has one issue: Number of Z(→l+l-)+jets events in the control region is too low.

Another option is to use γ+jets events since

- γ+jets and Z+jets kinematics are reasonably similar.
- The BR ratio in the estimate formula is replaced by Z/γ cross section ratio.
- We also take into technical factors like photon purity, etc.

Single photon control region for γ+jets in the CMS SUSY analysis with razor variables.



Background estimation
Sideband method

Used in searches for resonances, where the BG has a smooth, well-described shape, and the signal peaks over the BG.

- Define a signal region, and find signal-free control regions, i.e. sideband regions around the signal region.
- Deduce the shape of the BG from the sidebands (polynomial, exponential, etc.?)
- Extrapolate the BG in sidebands to the signal region.
- Either count the extrapolated events under the signal peak – or -- fit the data distribution to BG shape + signal shape and extract the parameters of the BG function.

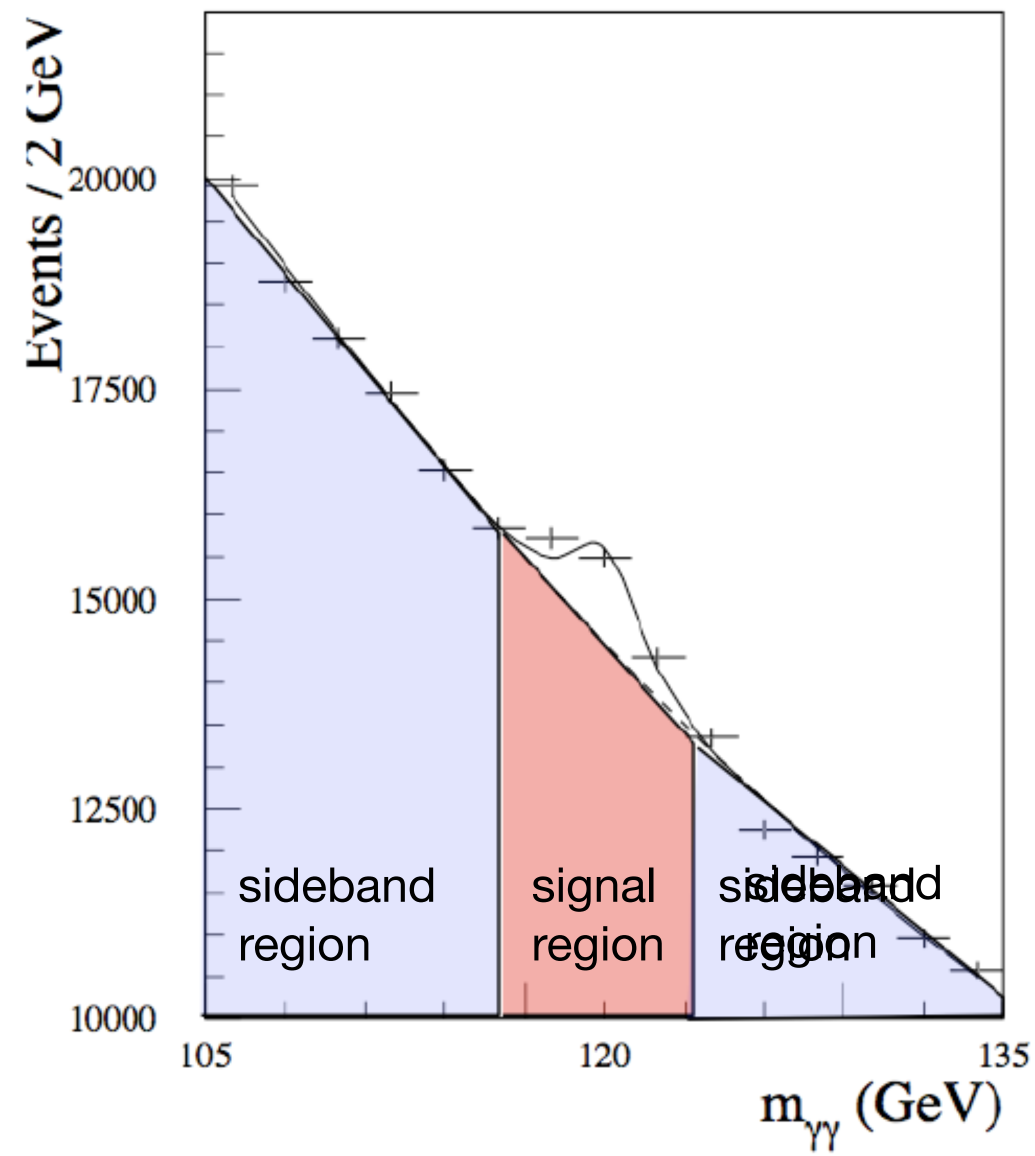
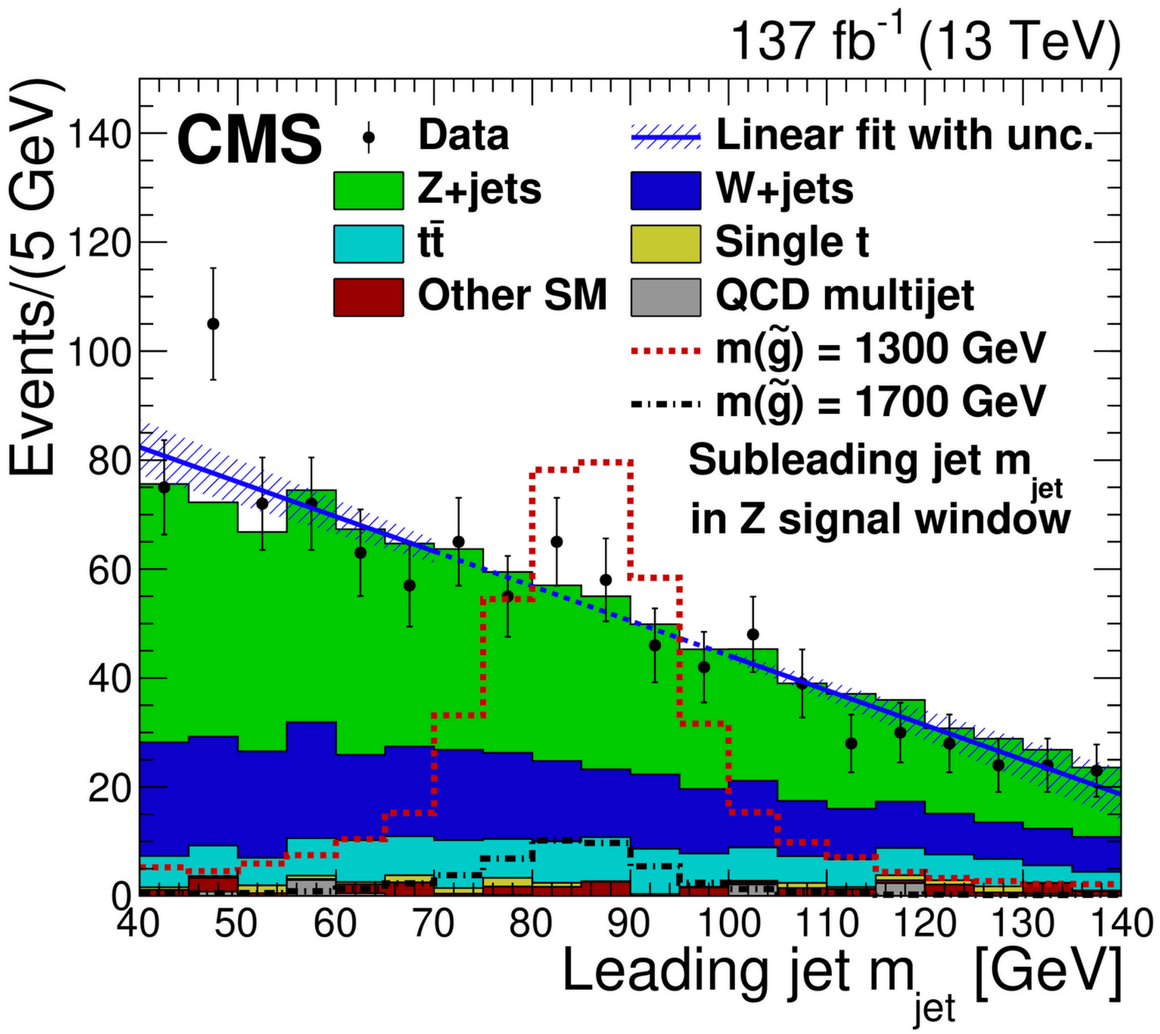


Figure from P. Govoni HCP2011 lectures

Sideband method

Used in searches for resonances, where the BG has a smooth, well-described shape, and the signal peaks over the BG.

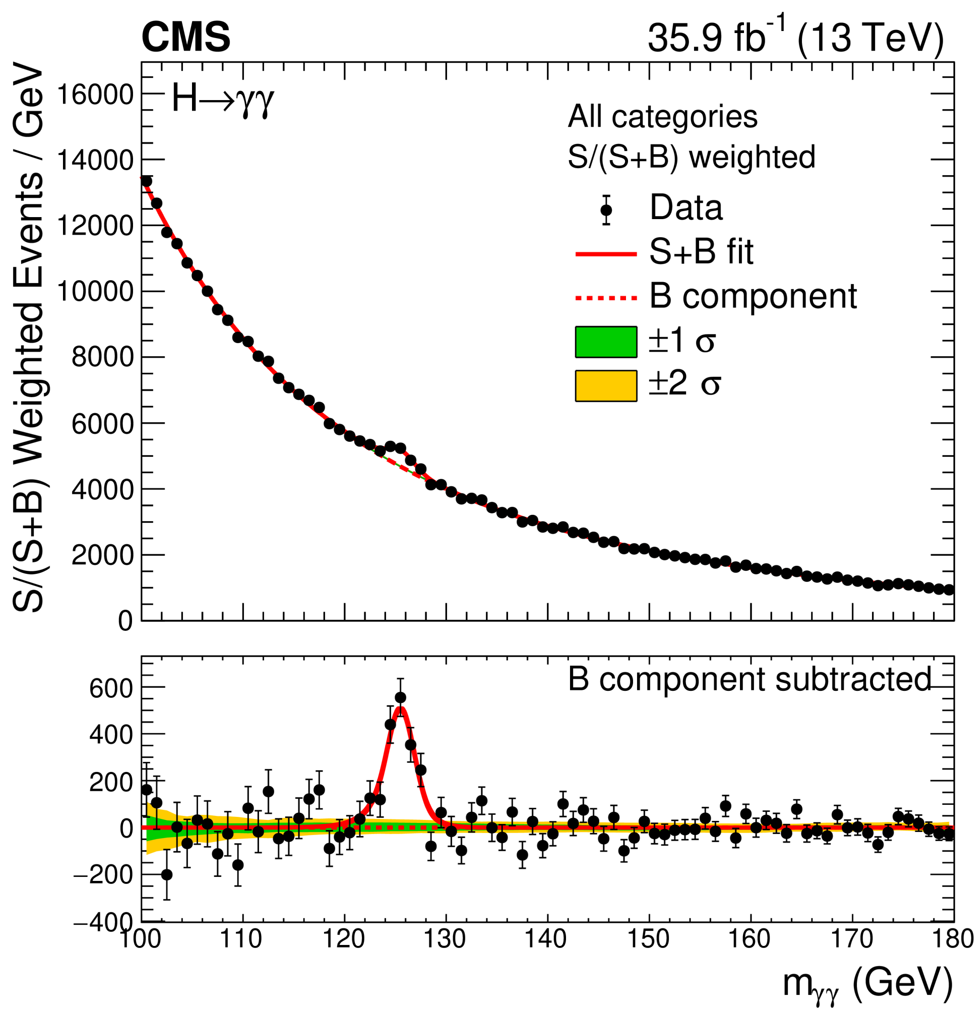
- Define a signal region, and find signal-free control regions, i.e. sideband regions around the signal region.
- Deduce the shape of the BG from the sidebands (polynomial, exponential, etc.?)
- Extrapolate the BG in sidebands to the signal region.
- Either count the extrapolated events under the signal peak – or -- fit the data distribution to BG shape + signal shape and extract the parameters of the BG function.

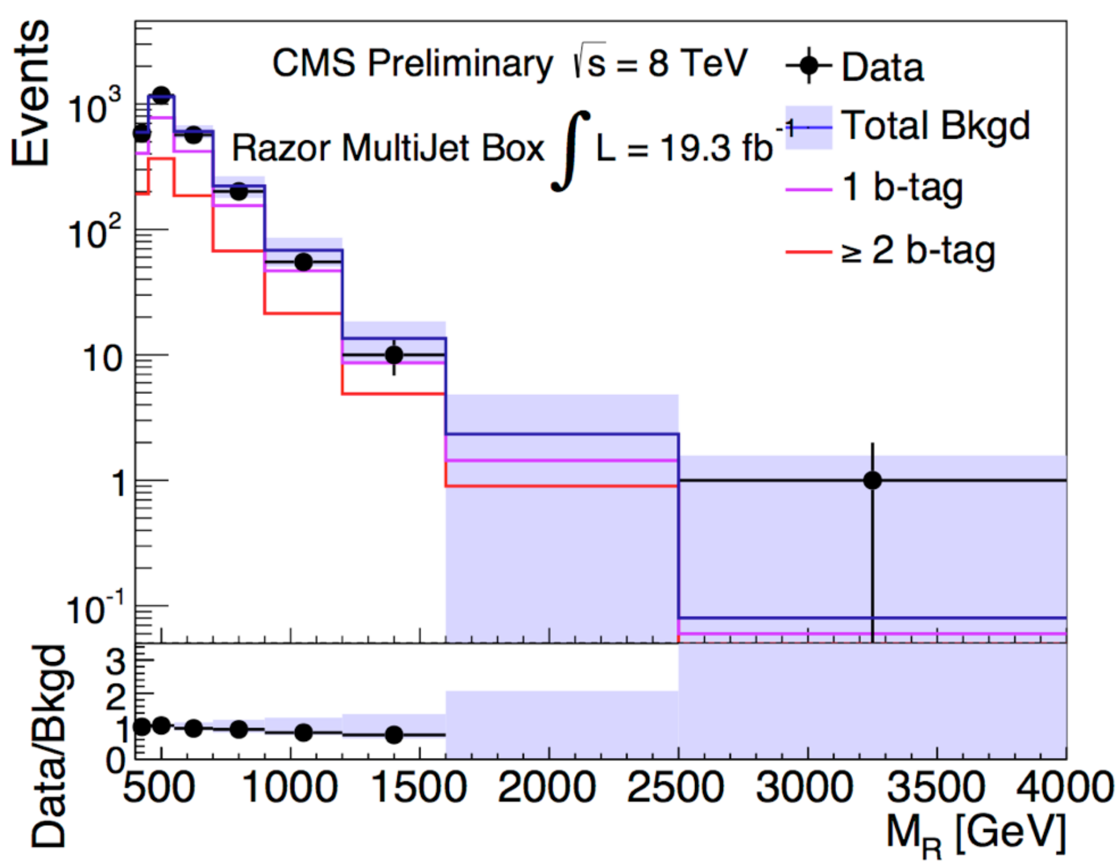


Fit to an analytical function

Sometimes the BG is well-described by an analytical function. If so:

- Find a control region dominated by the BG.
- Find an analytical function that describes the BG well.
- Fit the data to this analytical function in the control region and find the parameters of the analytical function.
- Extrapolate the fit to the signal region.





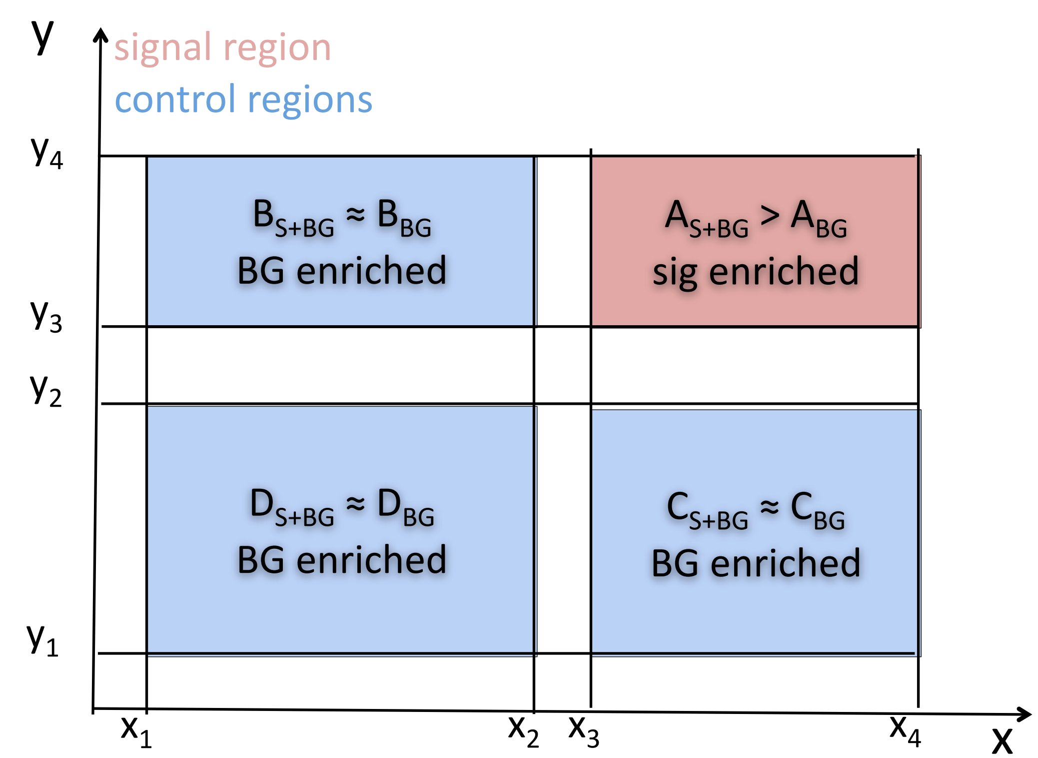
Matrix – or ABCD - method

When there exist two variables x and y for which the BG is uncorrelated, i.e. factorizable:

$$f^{BG}(x,y) = f^{BG}(x) \cdot f^{BG}(y)$$

- Apply all cuts except those on x and y on data
- Divide the x-y plane into 4-regions:
- When there is no signal, we have

$$\frac{N_A^{BG}}{N_B^{BG}} = \frac{N_C^{BG}}{N_D^{BG}}, \quad \frac{N_A^{BG}}{N_C^{BG}} = \frac{N_B^{BG}}{N_D^{BG}}$$



- In the presence of signal, A will be contaminated by the signal. But we can estimate the number of BG events in A from

$$N_A^{BG} = \frac{N_C^{BG} N_B^{BG}}{N_D^{BG}}$$

Note: Always beware the signal contamination in the control regions. Add it as a systematic.

Matrix – or ABCD - method

When there exist two variables x and y for which the BG is uncorrelated, i.e. factorizable:

$$f^{BG}(x,y) = f^{BG}(x) \cdot f^{BG}(y)$$

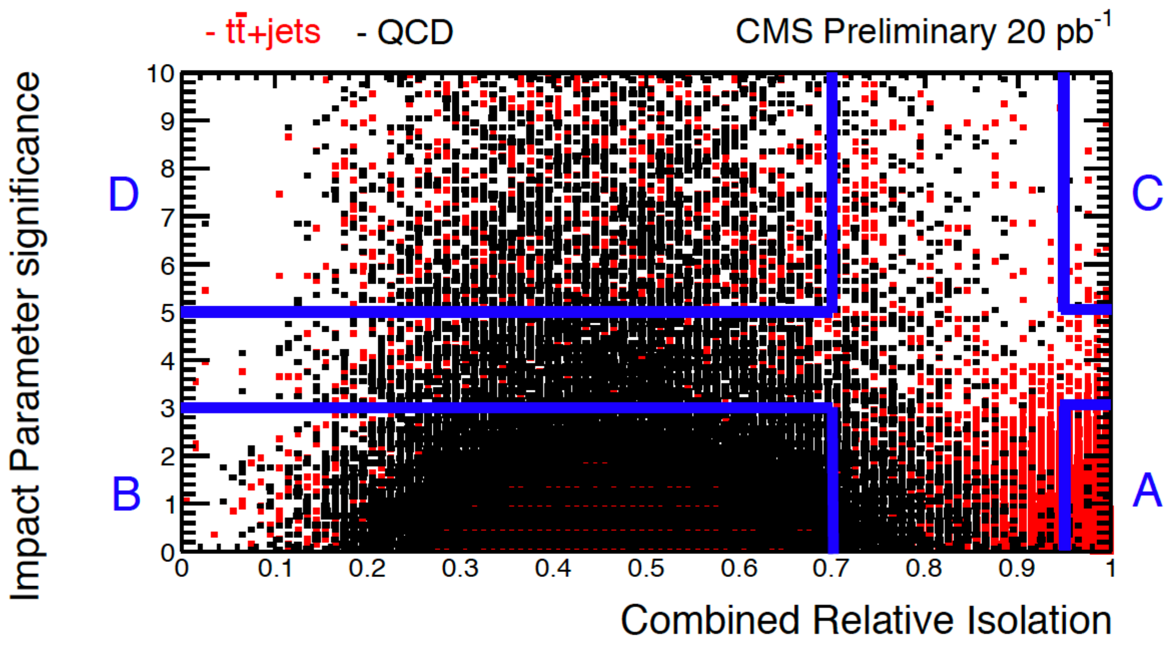
- Apply all cuts except those on x and y on data
- Divide the x-y plane into 4-regions:
- When there is no signal, we have

$$\frac{N_A^{BG}}{N_B^{BG}} = \frac{N_C^{BG}}{N_D^{BG}}, \quad \frac{N_A^{BG}}{N_C^{BG}} = \frac{N_B^{BG}}{N_D^{BG}}$$

- In the presence of signal, A will be contaminated by the signal. But we can estimate the number of BG events in A from

$$N_A^{BG} = \frac{N_C^{BG} N_B^{BG}}{N_D^{BG}}$$

CMS tt+jets cross section measurement in the muon+jets channel.



Note: Always beware the signal contamination in the control regions. Add it as a systematic.

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD in a signal region that has leptons. Real leptons come from the signal and fake leptons come from QCD (jets faking leptons). We define two event selections with loose and tight lepton ID criteria, which can be decomposed as:

$$
\begin{aligned}
N_{loose} &= N_{loose}^{real} + N_{loose}^{fake} \\
N_{tight} &= N_{tight}^{real} + N_{tight}^{fake} \\
\epsilon^k \equiv N_{tight}^k / N_{loose}^k \rightarrow &= \epsilon^{real} N_{loose}^{real} + \epsilon^{fake} N_{loose}^{fake}
\end{aligned}
$$

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD in a signal region that has leptons. Real leptons come from the signal and fake leptons come from QCD (jets faking leptons). We define two event selections with loose and tight lepton ID criteria, which can be decomposed as:

Get these counts from data

$$\boxed{N_{loose}} = N^{real}_{loose} + N^{fake}_{loose}$$

$$\boxed{N_{tight}} = N^{real}_{tight} + N^{fake}_{tight}$$

$$\epsilon^k \equiv N^k_{tight}/N^k_{loose} \rightarrow \quad = \epsilon^{real} N^{real}_{loose} + \epsilon^{fake} N^{fake}_{loose}$$

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD in a signal region that has leptons. Real leptons come from the signal and fake leptons come from QCD (jets faking leptons). We define two event selections with loose and tight lepton ID criteria, which can be decomposed as:

Get these counts from data

$$\boxed{N_{loose}} = N_{loose}^{real} + N_{loose}^{fake}$$

$$\boxed{N_{tight}} = N_{tight}^{real} + N_{tight}^{fake}$$

$$\epsilon^k \equiv N_{tight}^k / N_{loose}^k \rightarrow \quad = \boxed{\epsilon^{real}} N_{loose}^{real} + \boxed{\epsilon^{fake}} N_{loose}^{fake}$$

Find the ID efficiency, using e.g. tag and probe method.

Find this efficiency, i.e. the fake rate using e.g. a QCD control sample (e.g. low MET)

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD in a signal region that has leptons. Real leptons come from the signal and fake leptons come from QCD (jets faking leptons). We define two event selections with loose and tight lepton ID criteria, which can be decomposed as:

Get these counts from data

$$\boxed{N_{loose}} = \boxed{N^{real}_{loose}} + \boxed{N^{fake}_{loose}}$$

$$\boxed{N_{tight}} = N^{real}_{tight} + N^{fake}_{tight}$$

$$\epsilon^k \equiv N^k_{tight}/N^k_{loose} \rightarrow \quad = \boxed{\epsilon^{real}}\boxed{N^{real}_{loose}} + \boxed{\epsilon^{fake}}\boxed{N^{fake}_{loose}}$$

Solve the two equations simultaneously to get these numbers.

Find the ID efficiency, using e.g. tag and probe method.

Find this efficiency, i.e. the fake rate using e.g. a QCD control sample (e.g. low MET)

Fake rates method

The ratios of objects found by a tight identification over objects found by a loose identification is widely used as a BG estimation tool.

Suppose we would like to estimate QCD in a signal region that has leptons. Real leptons come from the signal and fake leptons come from QCD (jets faking leptons). We define two event selections with loose and tight lepton ID criteria, which can be decomposed as:

Solve the two equations simultaneously to get these numbers.

Get these counts from data

$$\boxed{N_{loose}} = \boxed{N^{real}_{loose}} + \boxed{N^{fake}_{loose}}$$

$$\boxed{N_{tight}} = N^{real}_{tight} + N^{fake}_{tight}$$

$$\epsilon^k \equiv N^k_{tight}/N^k_{loose} \rightarrow \quad = \boxed{\epsilon^{real}}\boxed{N^{real}_{loose}} + \boxed{\epsilon^{fake}}\boxed{N^{fake}_{loose}}$$

Find the ID efficiency, using e.g. tag and probe method.

Find this efficiency, i.e. the fake rate using e.g. a QCD control sample (e.g. low MET)

Finally obtain the number of BG events from $\epsilon^{fake} N^{fake}_{loose} = N^{fake}_{tight} = N_{BG}$

Tag-and probe method

- Tag and probe (TP) is a data-driven method used for measuring particle efficiencies. It is used for obtaining trigger, reconstruction, identification efficiencies. Mainly used for leptons.
- For TP, we need a mass resonance decaying to the object whose efficiency we want to measure (e.g. J/psi, upsilon, Z)
- We select two objects, a tag object and a probe object.
  - Tag object : Tight selection/ID criteria- we assume this is a real object.
  - Probe object: Very loose selection/ID criteria.
- We compute the diobject invariant mass of the tag object + probe object.
  - If the invariant mass is close to the resonance mass value, we assume that the probe object was a real object. Otherwise it should be a fake object.
- We take the real leptons inside the resonance mass window and apply on them the criteria of the selection, whose efficiency we want to measure
- Selection efficiency is computed as

$$\epsilon_{\text{selection}} = \mathrm{N}_{\text{selection}}^{\text{in mass window}} / \mathrm{N}_{\text{total}}^{\text{in mass window}}$$

Validating the estimates

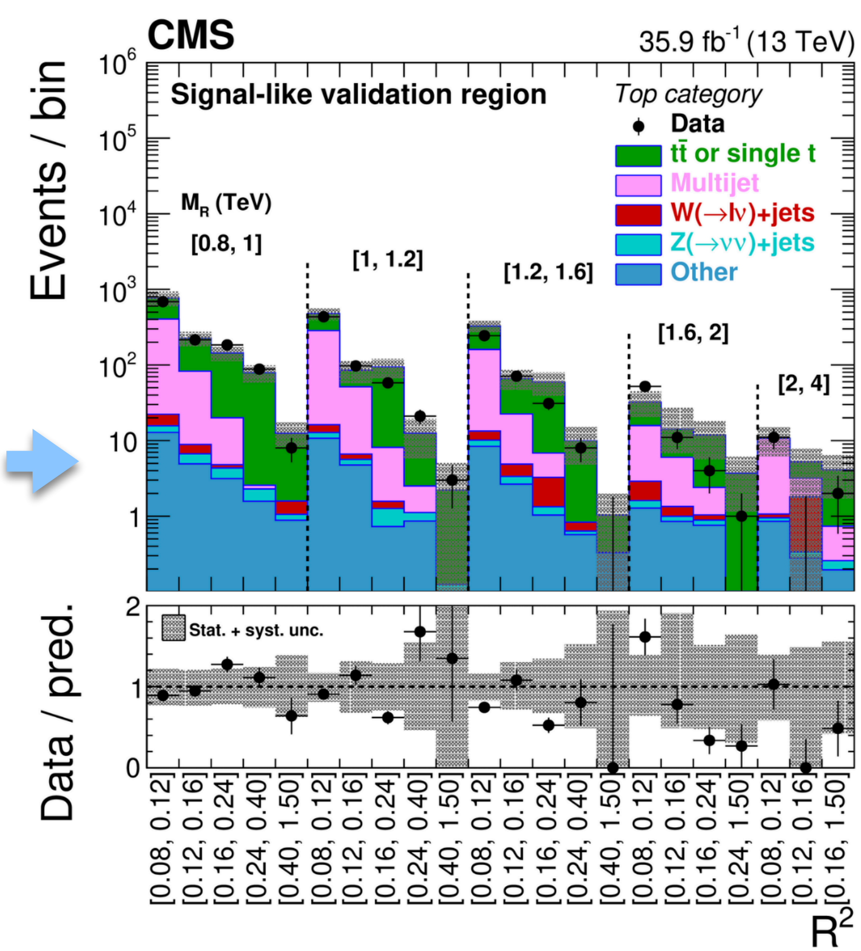
BG estimation methods must always be validated with closure tests or independent validation regions, or alternative methods.

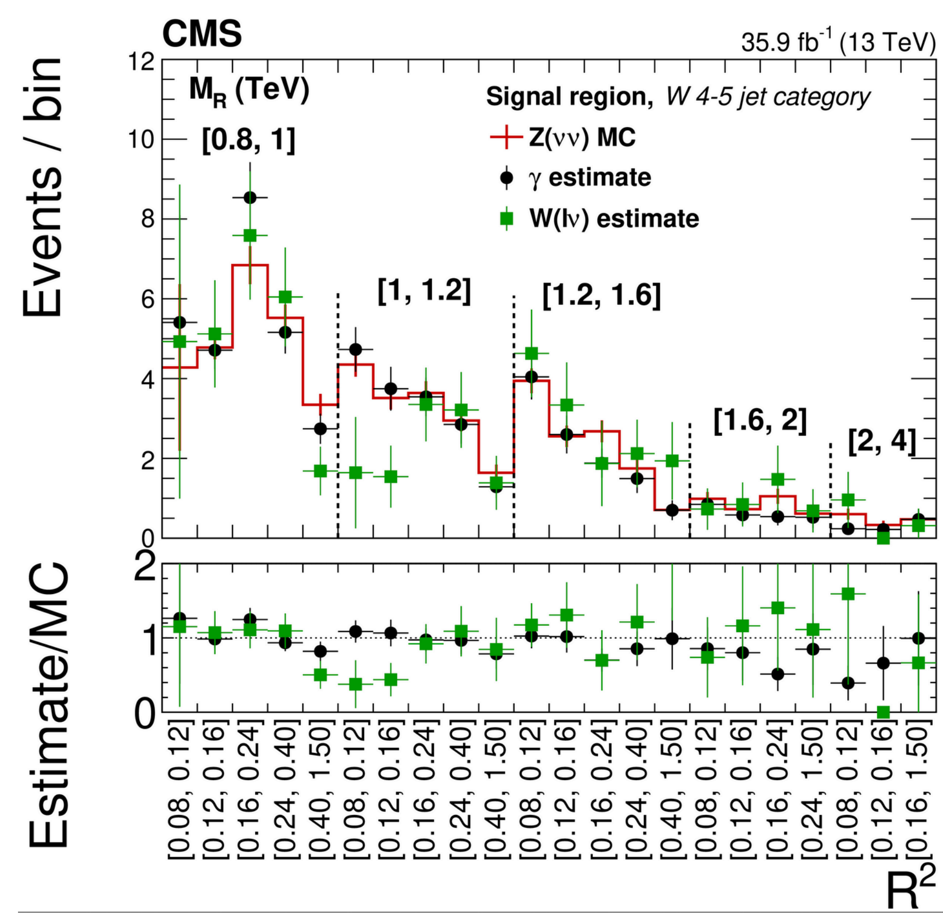
Closure tests : Validate the internal consistency of the method, e.g. validate the method using purely MC events.

Validation regions : Validate the method in independent dedicated regions. These can have a composition similar to the signal regions but be dominated by BG. Estimate should be equal to data.

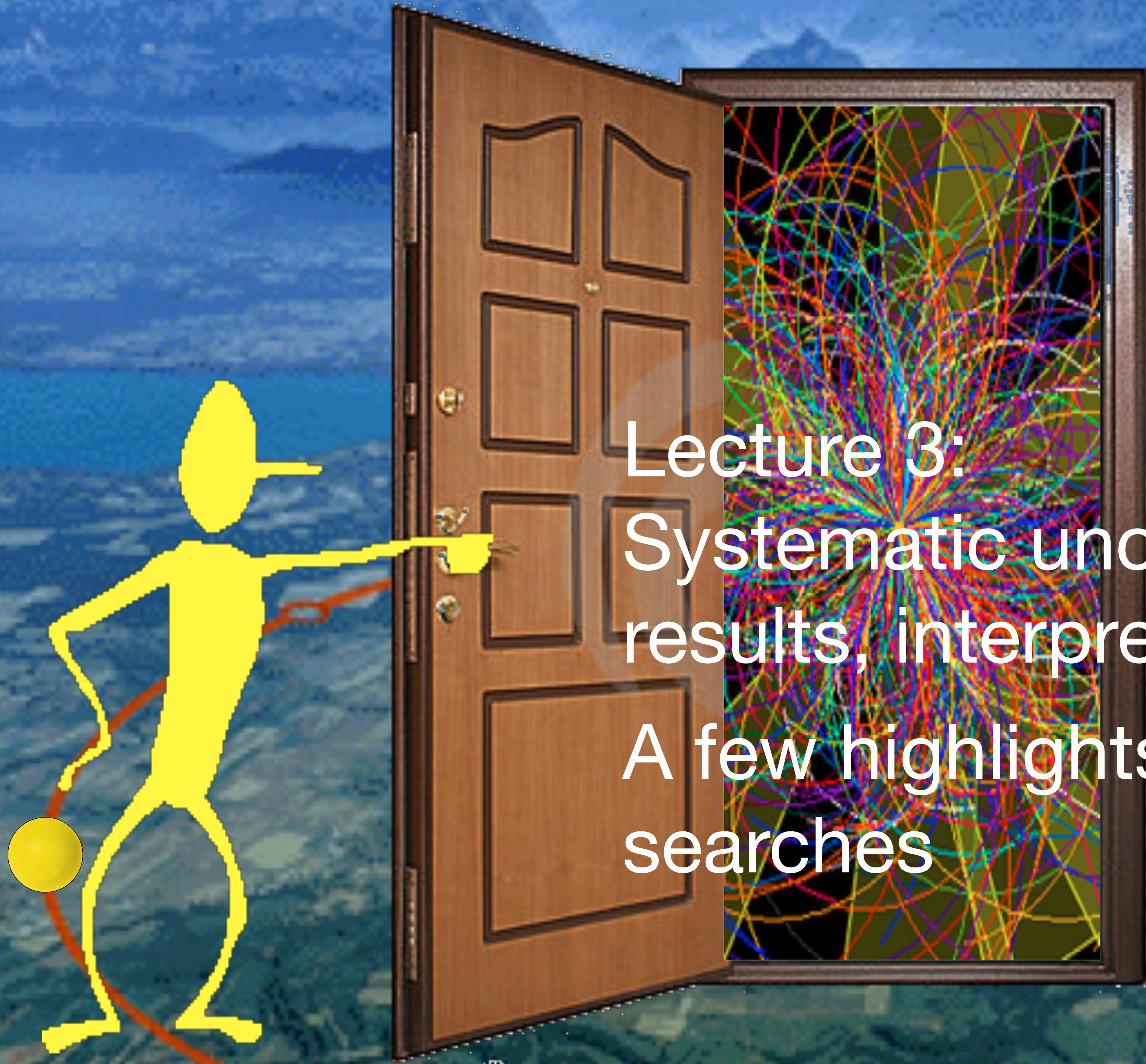
Alternative methods : Estimate the BG with multiple methods and compare the results.

Signal-like validation region from CMS SUSY razor.





Estimating Z(νν)+jets in 3 different ways from CMS SUSY razor.



Lecture 3:

Systematic uncertainties, results, interpretation

A few highlights from LHC searches