Anatomy of a HEP-ex Analysis

CSU NUPAX CERN IRES

Week 11: Balancing Signal Efficiency and Background Rejection

Johan S Bonilla

Overview of an Analysis

- Select target signal (SM measurement, BSM search, etc)
  - Production mechanism and/or final state (aka channel)
  - Used as a benchmark for optimizing the analysis

- Identify Trigger
  - How would the most signal and least background appear?
  - Loosest selection of analysis

- Design Signal Region
  - What selections would best enhance signal?

- Estimate Background
  - Given signal region strategy, what is your background?
  - Can you trust simulations? If not, need to derive estimate from data

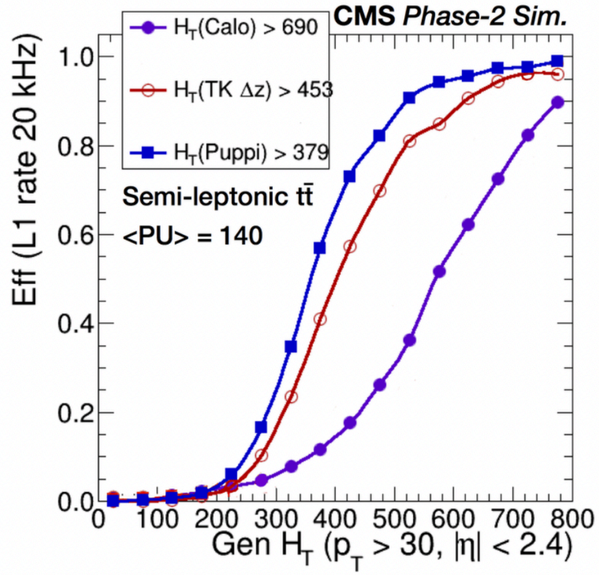
- Statistical Analysis
  - Multi-variate fit of expected signal and backgrounds in all regions

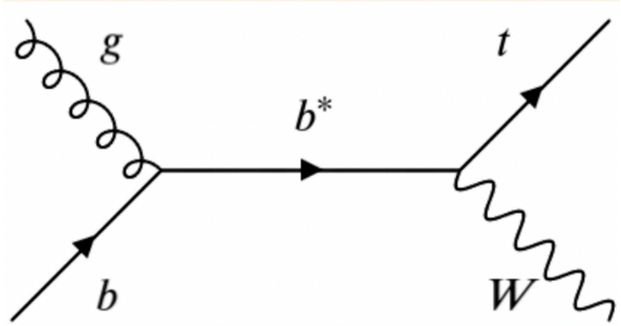
Preselections

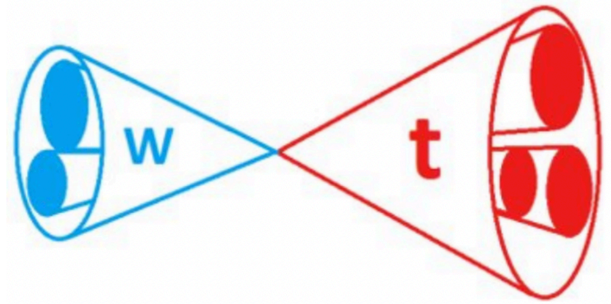
- Minimum selections to make sense of simulation+data
  - Selection of objects + thresholds
  - Trigger selection
  - Vetos on objects

- Contains all kinematic regions (SR/VR/CR)
  - Sometimes there are different SR/CR preselections
  - Helps diagnose general problems (weights, data/MC agreement, etc)

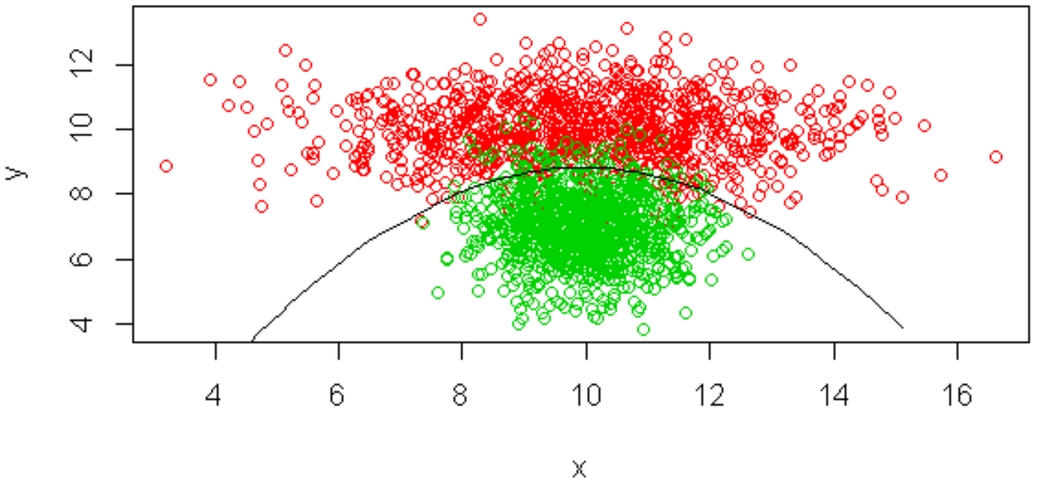
- Example: excited-b resonance
  - Think first, what finals state looks like
  - What if all-hadronic?
  - What if semi-leptonic?







What is a Signal Region

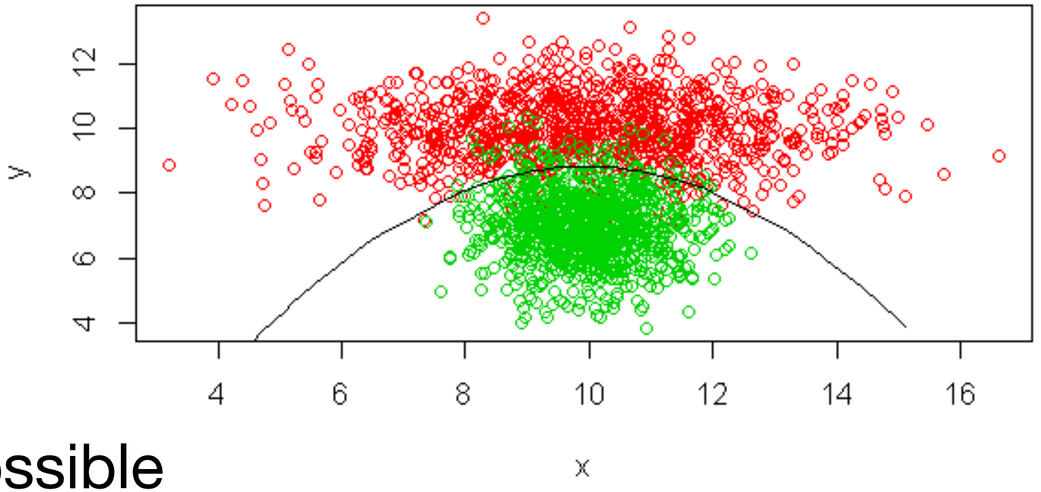


- A SR is a signal-rich region of phase-space
  - Signal efficiency should be as high as possible
  - Usually in a ‘corner’ of phase-space -> often modeling issues

- Defining a SR starts with understanding the signal kinematics
  - What physical processes characterize your signal?
  - Is there an observable that can discriminate signal from BG?

- Final SR depends on BG estimation
  - SR should have enough BG to be convincing (i.e. not too tight!)
  - SR should have a CR close enough to extrapolate results

What is a Control Region

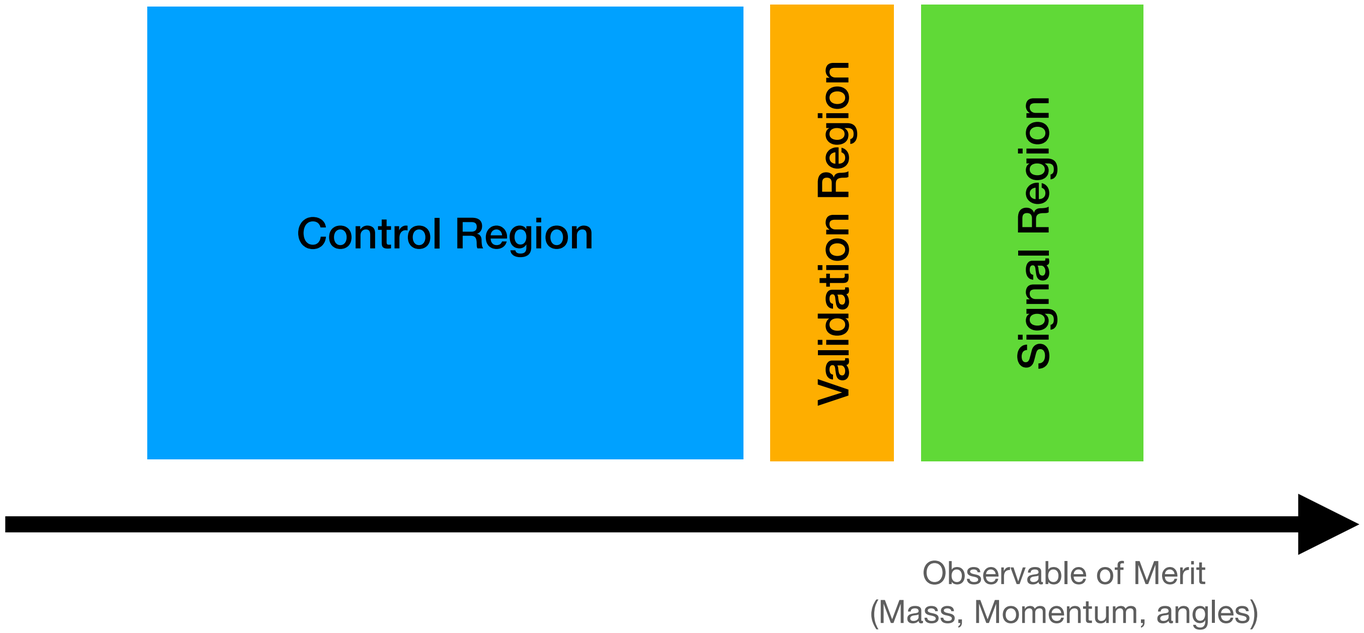


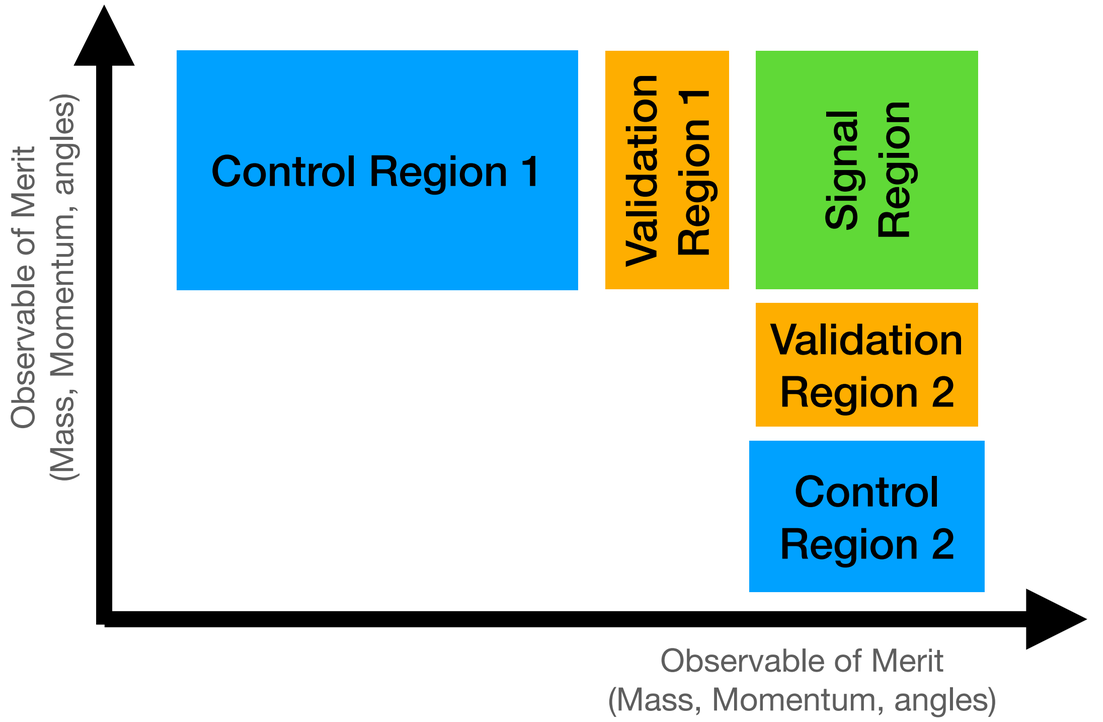
- A CR is a BG-rich region of phase-space
  - Signal efficiency should be as LOW as possible
  - Used to measure backgrounds CLOSE to SR -> to extrapolate CR to SR
  - Need to consider if SR is in a ‘corner’ of phase-space -> often modeling issues
  - Ideally one dedicated CR for each major BG

- Defining a CR starts with understanding the signal kinematics
  - What are the major sources of BG?
  - What cuts in SR can be flipped to measure BGs?
  - Can you trust all the simulations you have? If not -> data driven BG estimation

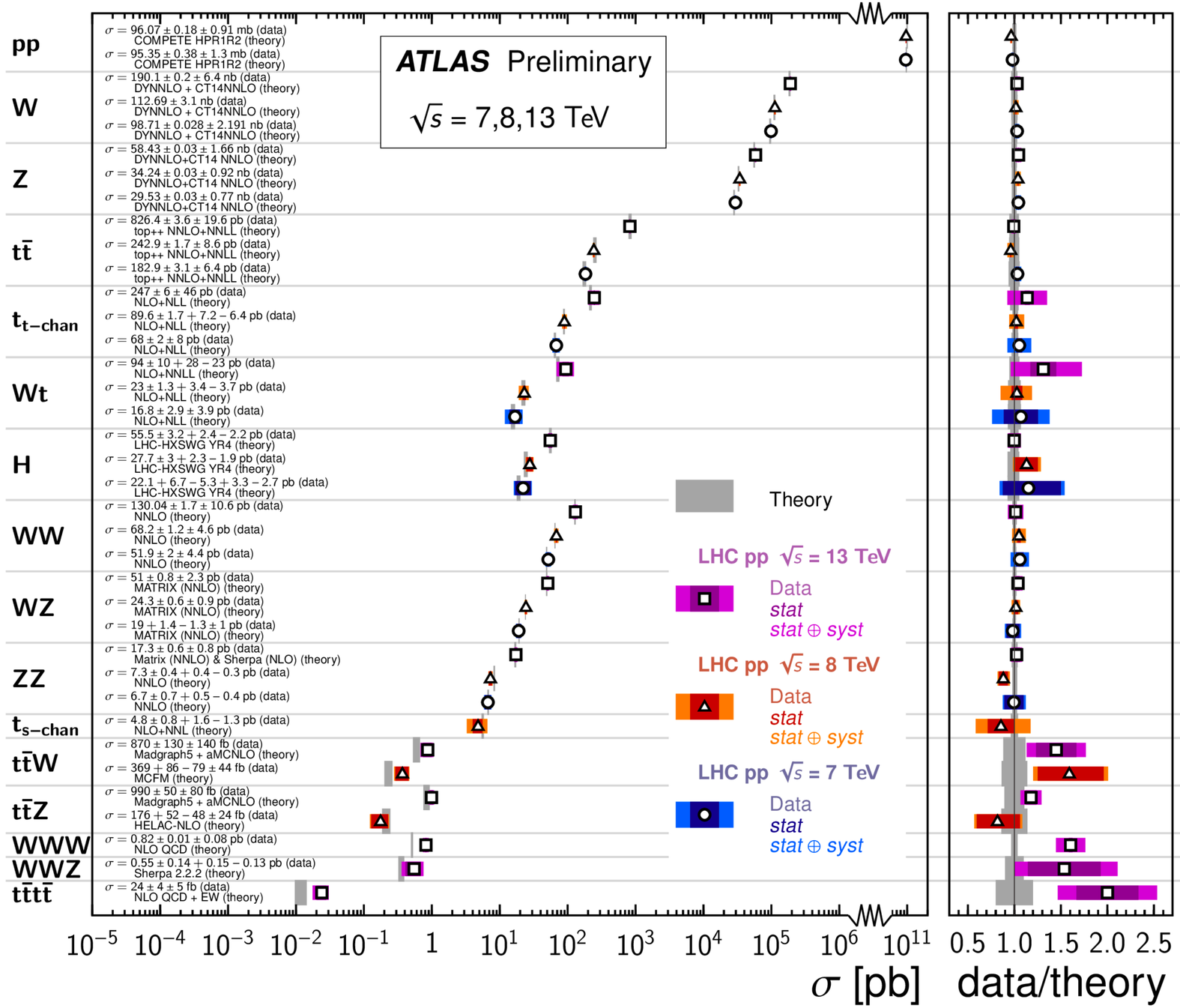
- Final SR depends on BG estimation
  - CR should have enough BG to be convincing (i.e. not too tight!)
  - Use a validation region (or VERY conservative unblinding) to verify CR->SR

Rough Sketch of SR/VR/CR





Standard Model Total Production Cross Section Measurements



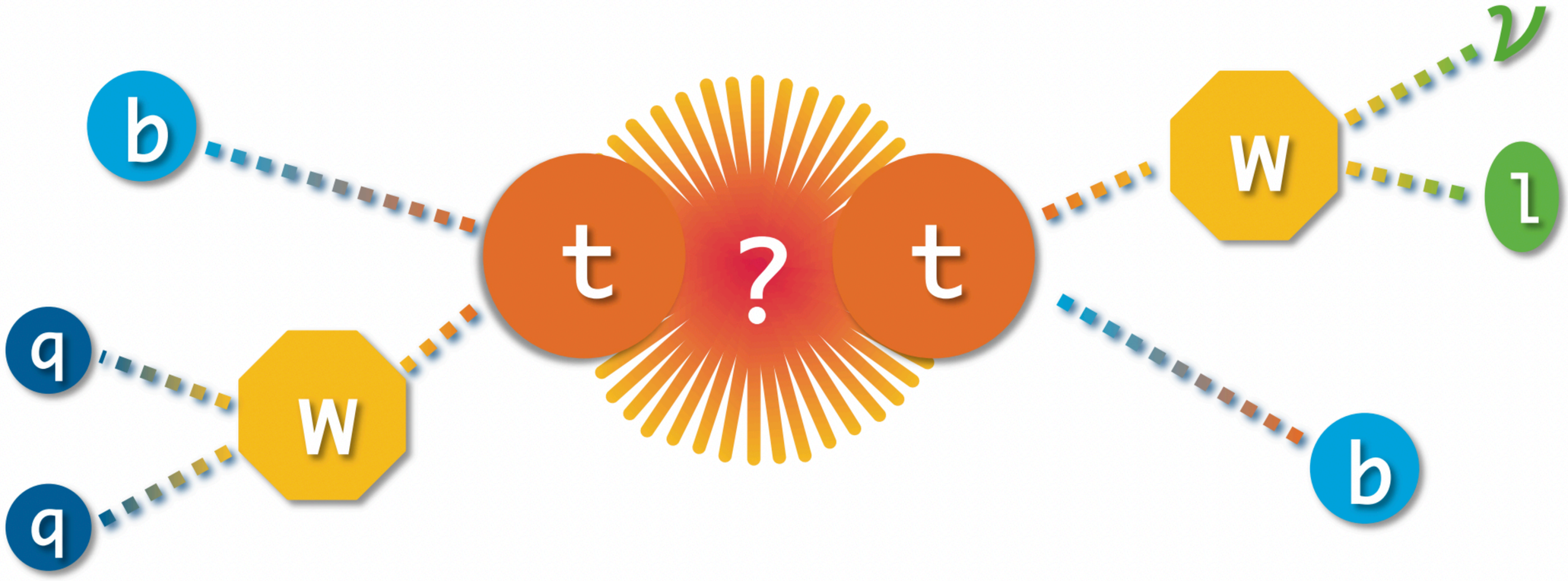
$\int \mathcal{L}\,dt$ [fb^{-1}]	Reference
50×10^{-8}	PLB 761 (2016) 158
8×10^{-8}	Nucl. Phys. B, 486-548 (2014)
0.081	PLB 759 (2016) 601
20.2	EPJC 79 (2019) 760
4.6	EPJC 77 (2017) 367
3.2	JHEP 02 (2017) 117
20.2	JHEP 02 (2017) 117
4.6	JHEP 02 (2017) 117
36.1	EPJC 80 (2020) 528
20.2	EPJC 74 (2014) 3109
4.6	EPJC 74 (2014) 3109
3.2	JHEP 04 (2017) 086
20.3	EPJC 77 (2017) 531
4.6	PRD 90, 112006 (2014)
3.2	JHEP 01 (2018) 63
20.3	JHEP 01, 064 (2016)
2.0	PLB 716, 142-159 (2012)
139	ATLAS-CONF-2022-002
20.3	EPJC 76 (2016) 6
4.5	EPJC 76 (2016) 6
36.1	EPJC 79 (2019) 884
20.3	PLB 763, 114 (2016)
4.6	Phys. Rev. D 87 (2013) 112001 arXiv:1408.5243
36.1	EPJC 79 (2019) 535
20.3	PRD 93, 092004 (2016)
4.6	EPJC 72 (2012) 2173
36.1	PRD 97 (2018) 032005
20.3	JHEP 01, 099 (2017)
4.6	JHEP 03, 128 (2013) PLB 735 (2014) 311
20.3	LB 756, 228-246 (2016)
36.1	PRD 99, 072009 (2019)
20.3	JHEP 11, 172 (2015)
139	Eur. Phys. J. C 81 (2021) 737
20.3	JHEP 11, 172 (2015)
139	arXiv:2201.13045
79.8	PLB 798 (2019) 134913
139	JHEP 11 (2021) 118

Status: February 2022

Let's Derive a Signal Region

Thinking Like a Physicist

- Ex: Search for new ttbar Resonances



Blinding Signal Regions

- How can an observer be biased in their measurements?
  - Can be tempted to adjust methods to yield stronger limits/results
  - Unexpected BG can obscure signal
  - Control Region of one analysis could be the Signal Region of another

- How to avoid biasing? Blinding!
  - Do not look at data in Signal Region until 100% ready
  - Use validation regions to confirm background estimation of control regions
  - ATLAS/CMS has policies of varying rigidity
  - CRs should be loose, always check signal contamination (< few%)

- ALWAYS ASK IF YOU CAN LOOK AT DATA!!!
  - Best to be safe, don’t be afraid to be the stickler!

End of April 5th Class

How to Measure Sensitivity

- Sensitivity: ability of an analysis to capture signal
  - Detector acceptance: can you see it with the machine
  - Signal efficiency: can you keep most of your signal
  - Background rejection: can you reduce rate of non-signal processes

- What affects sensitivity?
  - Detector hardware
  - Triggers
  - Taggers

$$\sigma = \frac{S}{\sqrt{B}}$$

Data Size Impact on Sensitivity

What is the impact of more data from LHC?

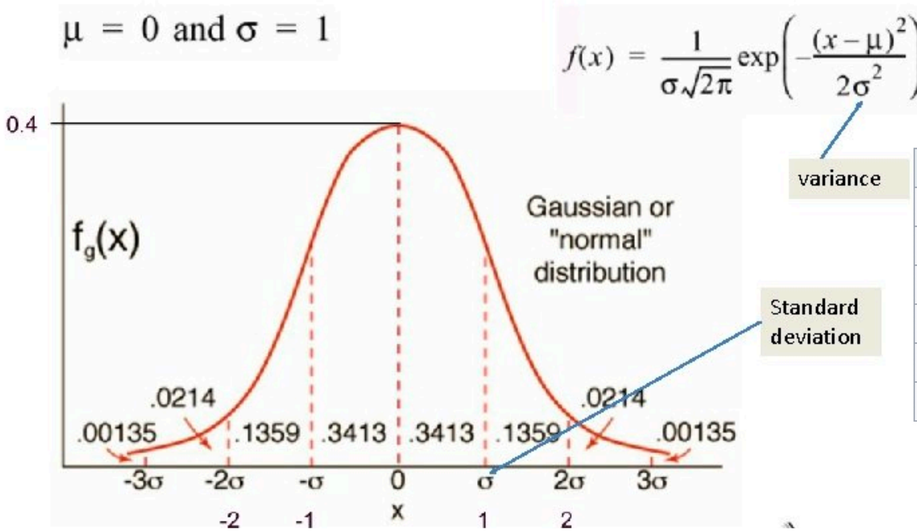
$$\sigma = \frac{S}{\sqrt{B}} \qquad \sigma = \frac{S}{\sqrt{S+B}} \qquad \sigma = \frac{S}{\sqrt{S+(B+\delta^2_{syst})}}$$

Sigma — One Word, Many Meanings

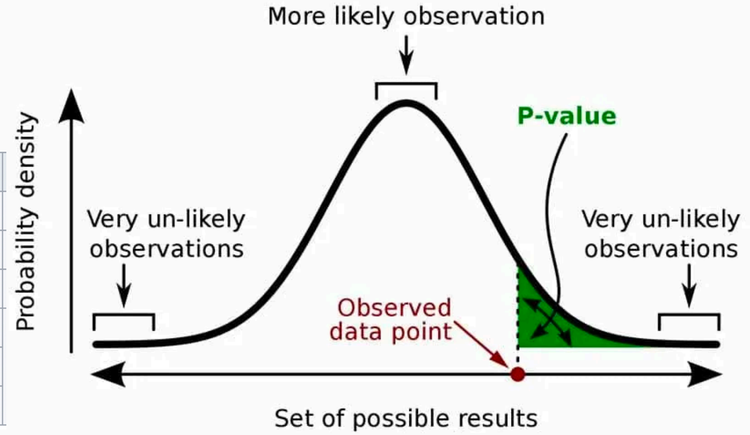
Introduction to measurement and exclusion limits

- Sigma: usually represents some width (distance) away from mean
  - Gaussian: 4, 5 sigma => ~1/16k, ~1/1.7M probability
  - Higgs discovery at 5 sigma (local)

- Confidence level: probability (p-value) that null hypotheses is true
  - Se no excess of events => set exclusion limits



n	$p = F(\mu + n\sigma) - F(\mu - n\sigma)$	i.e. $1 - p$	or 1 in p
1	0.682 689 492 137	0.317 310 507 863	3.151 487 187 53
2	0.954 499 736 104	0.045 500 263 896	21.977 894 5080
3	0.997 300 203 937	0.002 699 796 063	370.398 347 345
4	0.999 936 657 516	0.000 063 342 484	15 787.192 7673
5	0.999 999 426 697	0.000 000 573 303	1 744 277.893 62
6	0.999 999 998 027	0.000 000 001 973	506 797 345.897



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

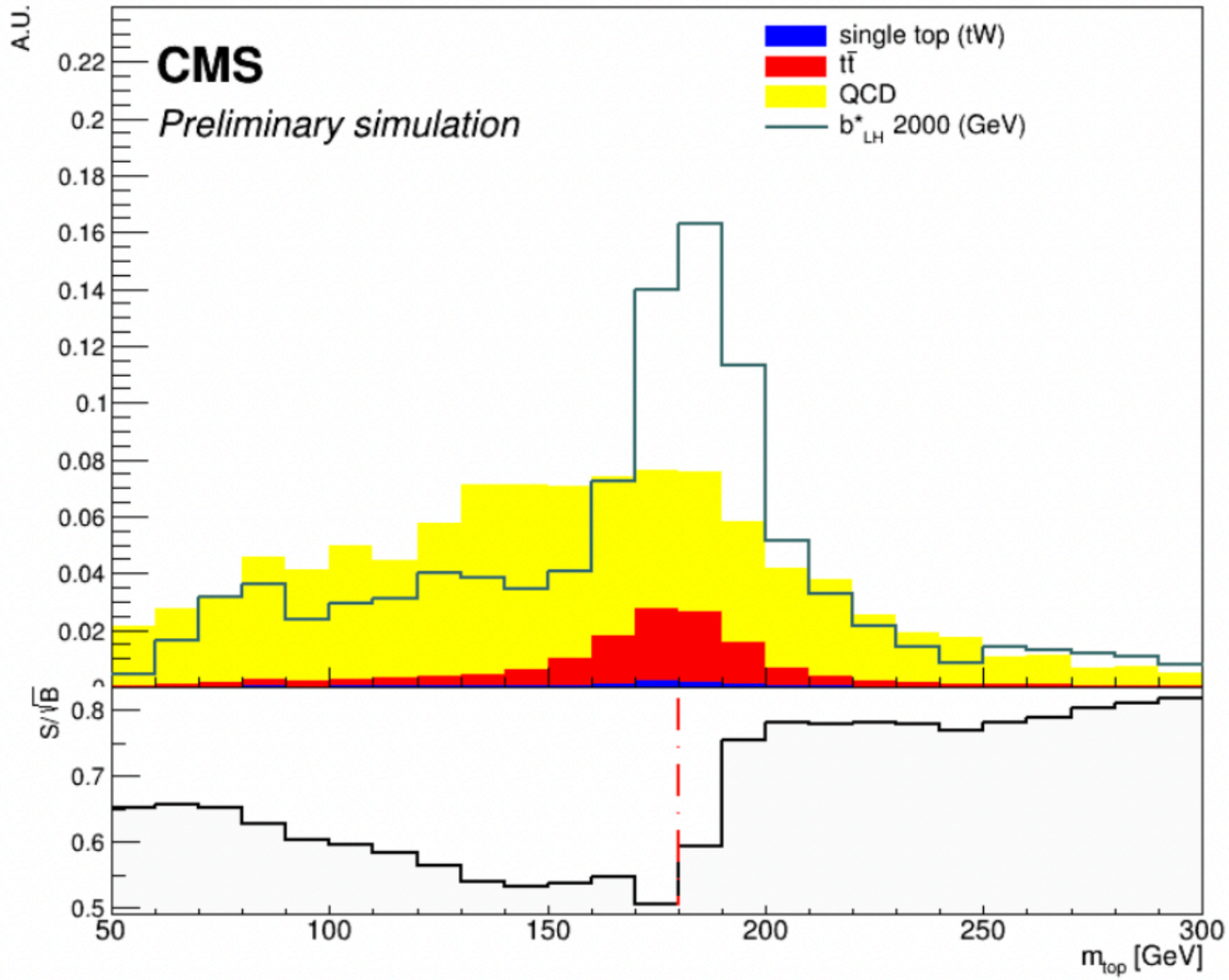
N-1 Plots

N = All selections
N-1 = All-but-one selection

Shows power of single observable
— Use S/sqrtB to estimate

Used in all analysis steps!

Worth to develop your own library



Homework

Due April 11th

- From your previous papers, find the SR/VR/CR used
  - Present them to the class in a few slides on Thursday
  - Also list preselection cuts and triggers used
- In RUCIO, find the signal and dominant BGs
  - Download a test file of each to your public space on lxplus
- Next Class: Plotting
  - Plots with no cuts
  - Plots with preselection cuts
  - N-1 plots with selections in paper
  - Speculate which cuts could be optimized