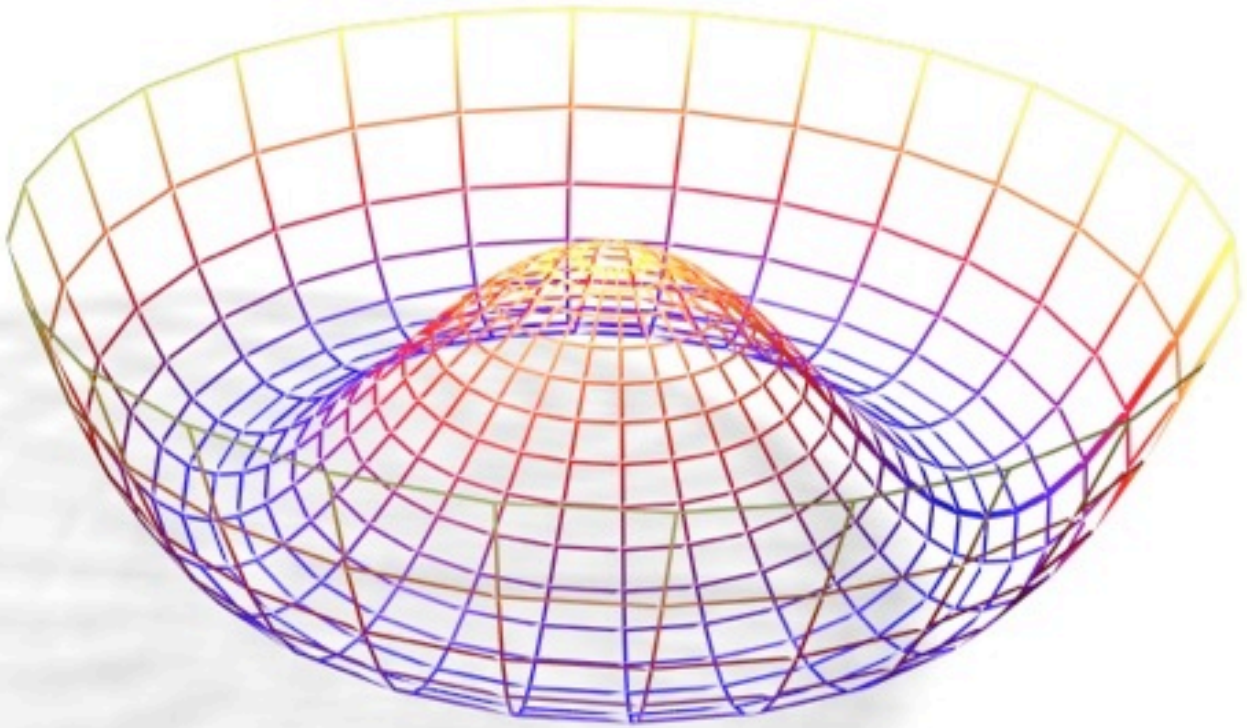
DASPOS Data and Software Preservation for Open Science

Center for Cosmology and Particle Physics

RECAST: Analysis Archival



Kyle Cranmer,
New York University

Together with Itay Yavin and support from the Perimeter Institute

Introduction

Focus of data preservation and open access has been on “the data”

- collision data: particle momenta, energies, etc.
- meta data: (documentation?)
- necessary software to use data

This is a good goal, but

- very difficult
  - several full time people for years arriving at ATLAS data model)
- even if data model is solved, using the data properly and obtaining a robust scientific result is hard
  - I went back to 10 year old ALEPH data and was afraid I was forgetting something

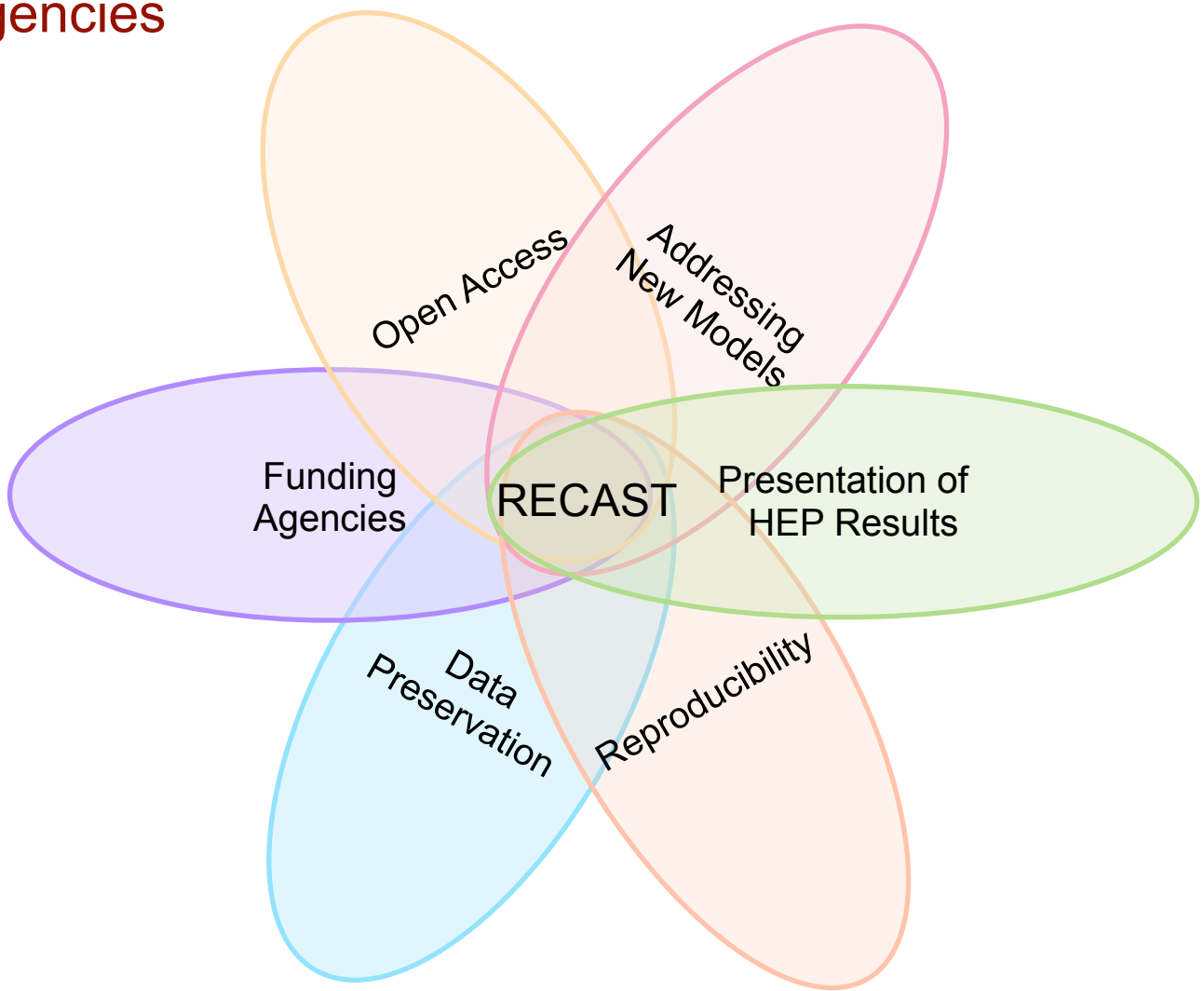
Relatively little discussion on preserving the analyses (>100 / year)

- not preserved, not reproducible, and not accessible in detail
- analyses are robust and can be extended to new scientific questions
  - In short there is demand for archiving the analyses, it’s an easier problem technically, and I argue that it has more scientific impact than “the data”

The bigger picture

There are several overlapping discussions where RECAST is relevant:

- **Physics**: presentation of LHC results, addressing new models
- **Policy**: data preservation, open access, reproducibility, requests from funding agencies



Blueprint of Data Preservation in High-Energy Physics

(or rather "Status report")

arXiv:1205.4667

DPHEP Study Group for Data Preservation and Long Term Analysis in High Energy Physics

DPHEP BLUEPRINT DRAFT 0.95 23/02/2012

5. DPHEP Common Projects

Among the projects and needs of the various experiments, certain areas have emerged as opportunities for common efforts to build preservation infrastructure within the community. Common infrastructures for data preservation can be jointly developed with the coordination and support of DPHEP, pooling the interests and resources of collaborations and other interested parties to eliminate duplication of effort.

In particular, technologies for data preservation, such as automated validation frameworks, are areas of potential collaboration. The development of a data storage solution for long term preservation, given that day-to-day systems are unlikely to provide the level of data integrity security required may also be investigated. In addition, the RECAST project offers an alternative to maintaining the full data set.¶

5.3 The RECAST Framework

While there are many technical obstacles to preparing and using archived HEP data, one of the most challenging is the propagation of institutional wisdom needed to use the data properly. Even if the community is able to overcome the technical challenges in preserving HEP data, it will continue to be difficult to extract meaningful scientific results. The RECAST framework provides a complementary approach.

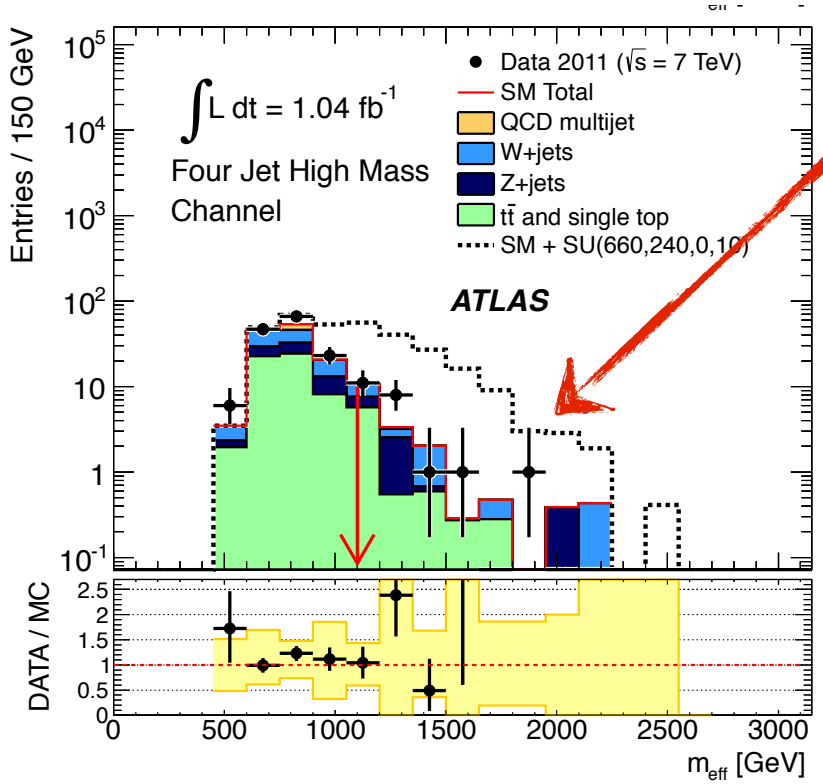
Analysis re-interpretation: "recasting"

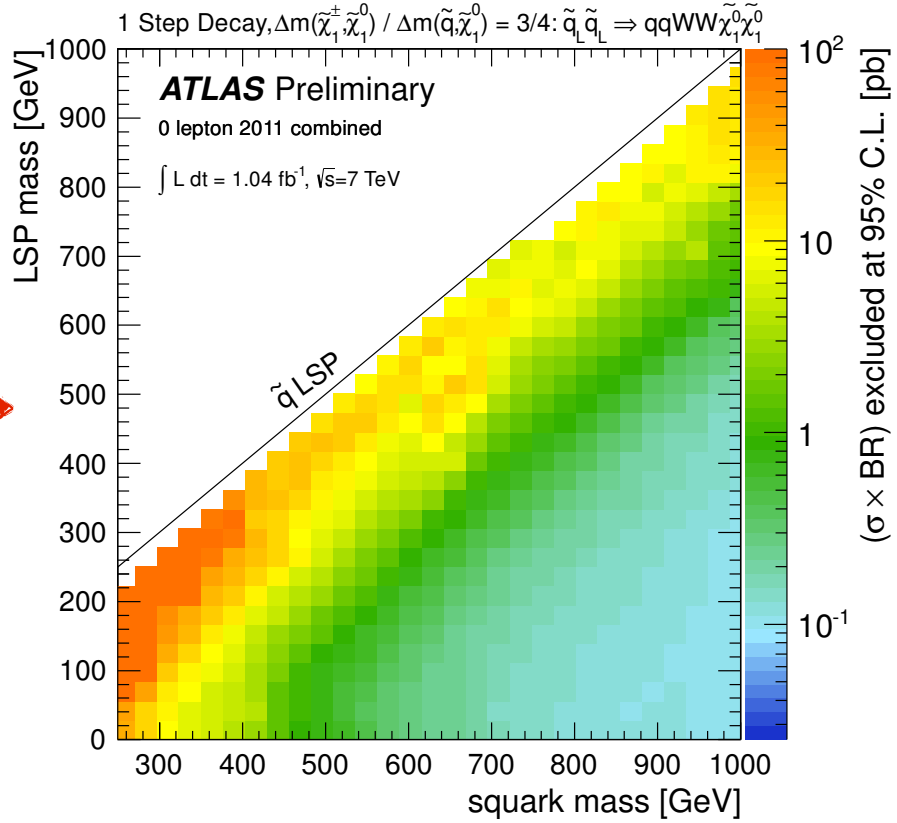
Information | Discussion (0) | Files

ATLAS Note

Report number	ATLAS-CONF-2011-155
Title	**Search for supersymmetry with jets and missing transverse momentum:** Additional model interpretations
Corporate Author(s)	The ATLAS collaboration
Imprint	14 Nov 2011. - mult. p.
In:	Hadron Collider Physics Symposium 2011,
Subject category	Detectors and Experimental Techniques
Accelerator/Facility, Experiment	CERN LHC ; ATLAS
Free keywords	supersymmetry ; SUSY ; simplified model ; extra dimensions

Additional model interpretations require a new signal histogram, which is obtained internally through full analysis chain.



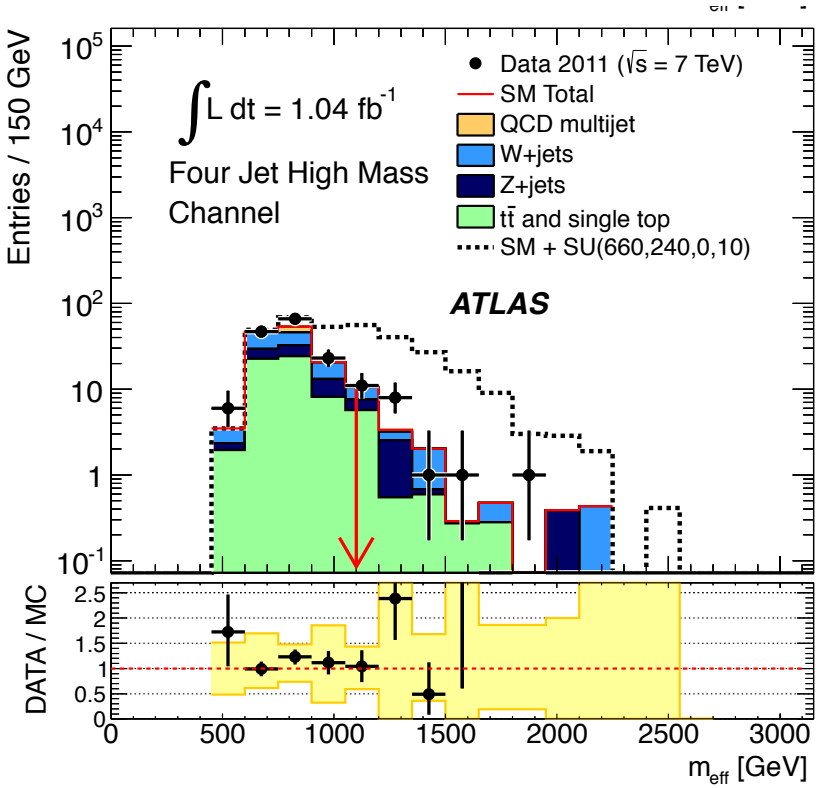


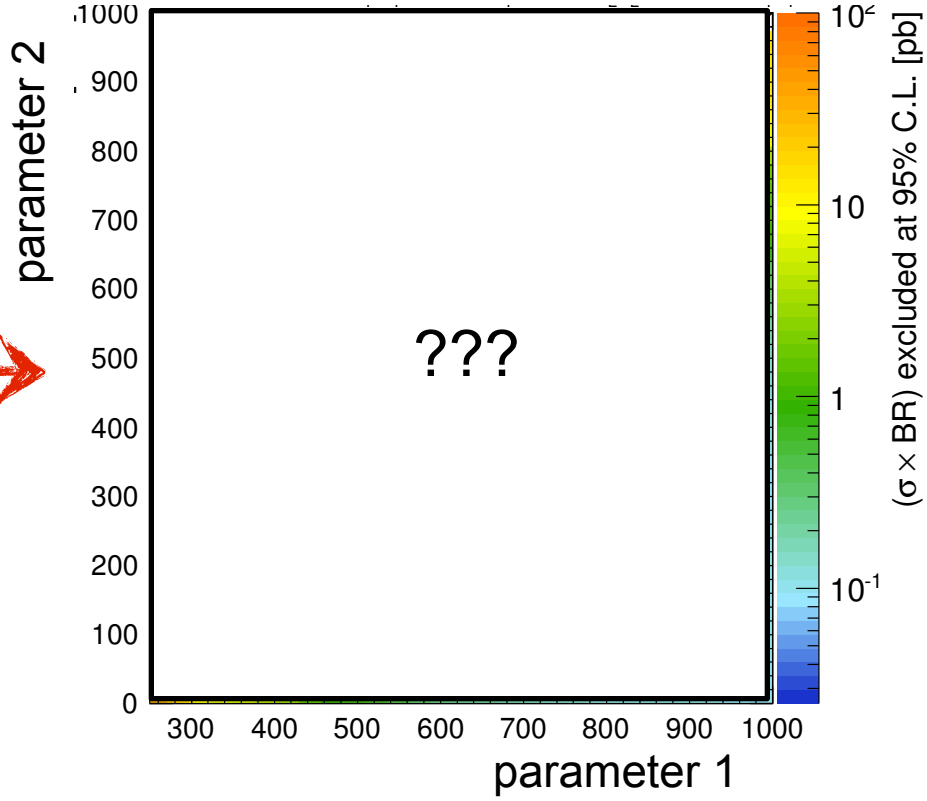
Status of analysis reinterpretation

SUSY analyses are doing a pretty good job of recasting the results of their searches into simplified models and other SUSY scenarios.

But there are many other analyses that do not provide any re-intepretation and many (most) models are not being addressed.

RECAST is a framework for recasting that aims to collect, standardize, and facilitate the processing of recast requests from the community.



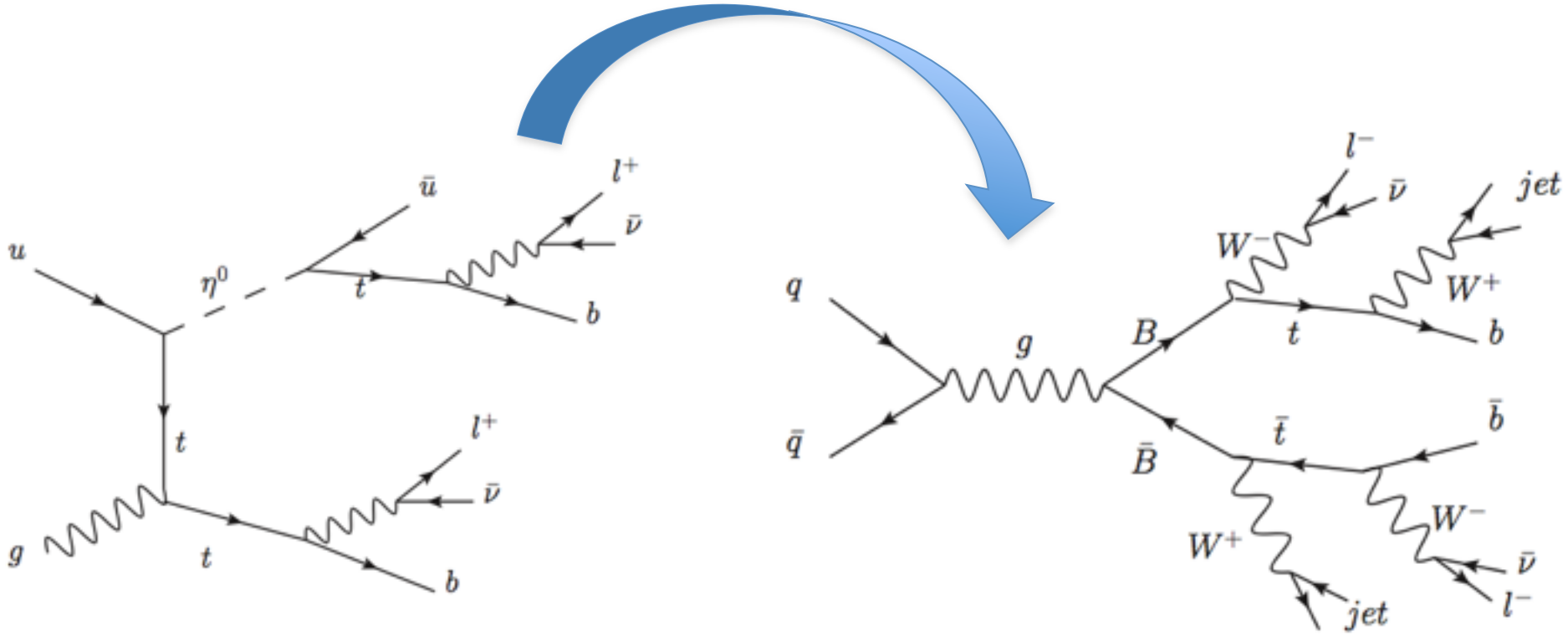


CDF 4th Generation Search

Center for Cosmology and Particle Physics

D. Whiteson for CDF recasted a previous search for maximal flavor violating scalars into a search for 4th generation b-quarks.

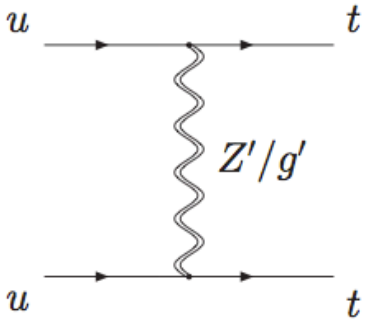
Both scenarios lead to $\ell^{\pm}\ell^{\pm}bj\not{E}_T$

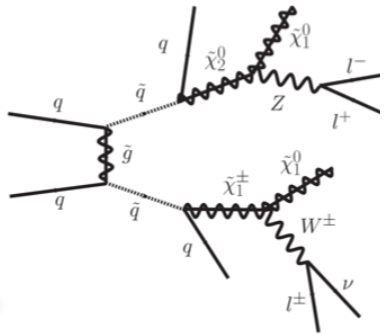


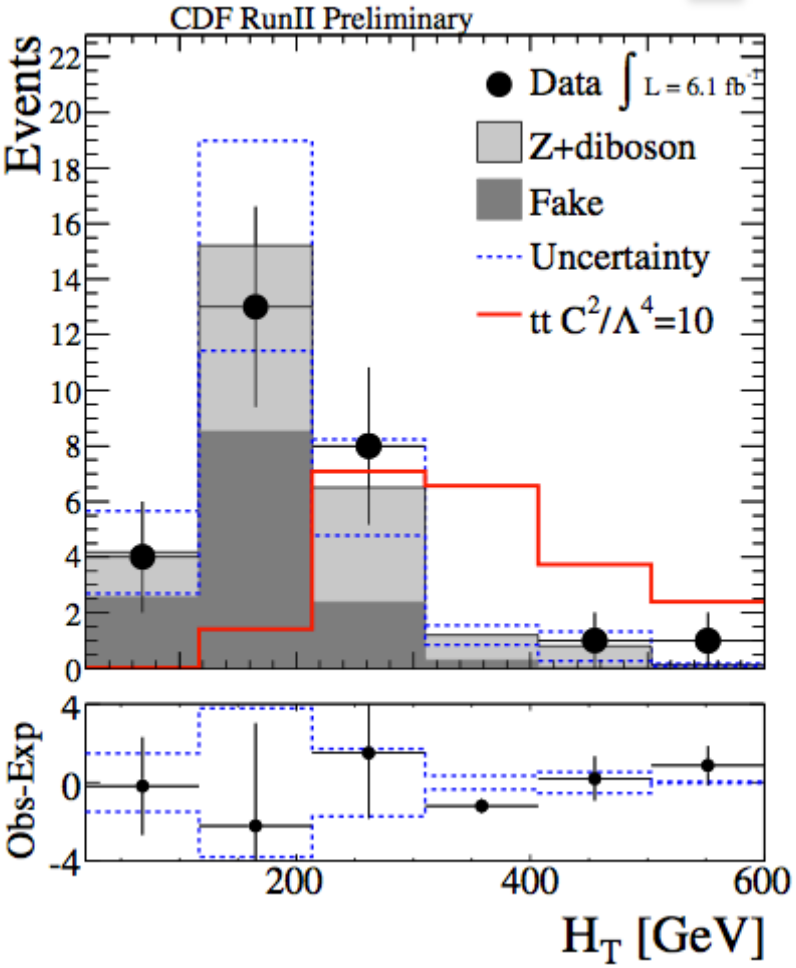
Phys. Rev. Lett. 102
hep-ex/0809.4903

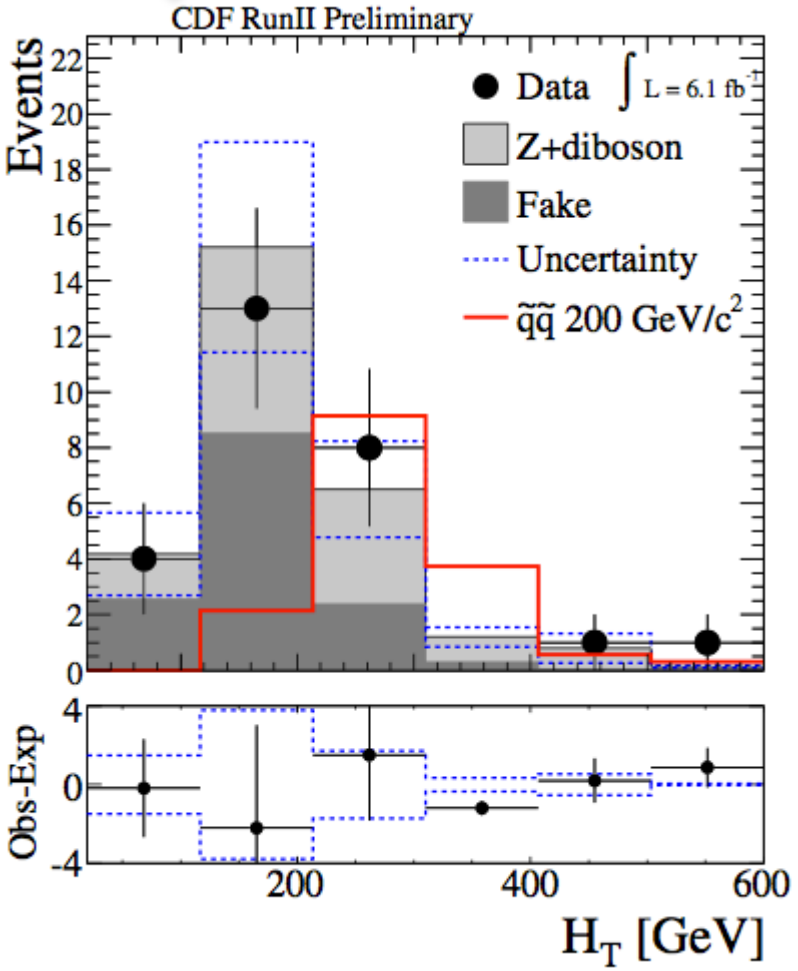
Phys. Rev. Lett. 104
hep-ex/0912.1057

Recent CDF same-sign suite









Take home points for RECAST

- Does not require access to or reprocessing of the data
- Does not involve design of new event selection criteria
- Does not require additional estimates of background rates or systematic uncertainties

- Extends the impact of existing experimental searches
- Targets physics scenarios of interest to the community
- Provides accurate interpretation of model-independent and signature-based searches in the context of a specific model
- Facilitates the consideration of new models even after the analysis is done
- Allows collaborations to control the approval of new results
- Complements data archival efforts

Discussed regularly in LHC Data Preservation meetings.
Part of upcoming DPHEP report to ICFA.

DPHEP: Which preservation model?

The RECAST approach requires Level 4 infrastructure, but does not require archiving the data.

- DPHEP table is useful, but should not say the preservation models are inclusive

Preservation Model	**Use case**
1. Provide additional documentation	Publication-related information search
2. Preserve the data in a simplified format	Outreach, simple training analyses
3. Preserve the analysis level software and data format	Full scientific analysis based on existing reconstruction
4. Preserve the reconstruction and simulation software and basic level data	Full potential of the experimental data

Table 1: Various preservation models, listed in order of increasing complexity. Subsequent models are inclusive. For example, preservation model 4 also includes steps and use cases described in models 1,2 and 3.

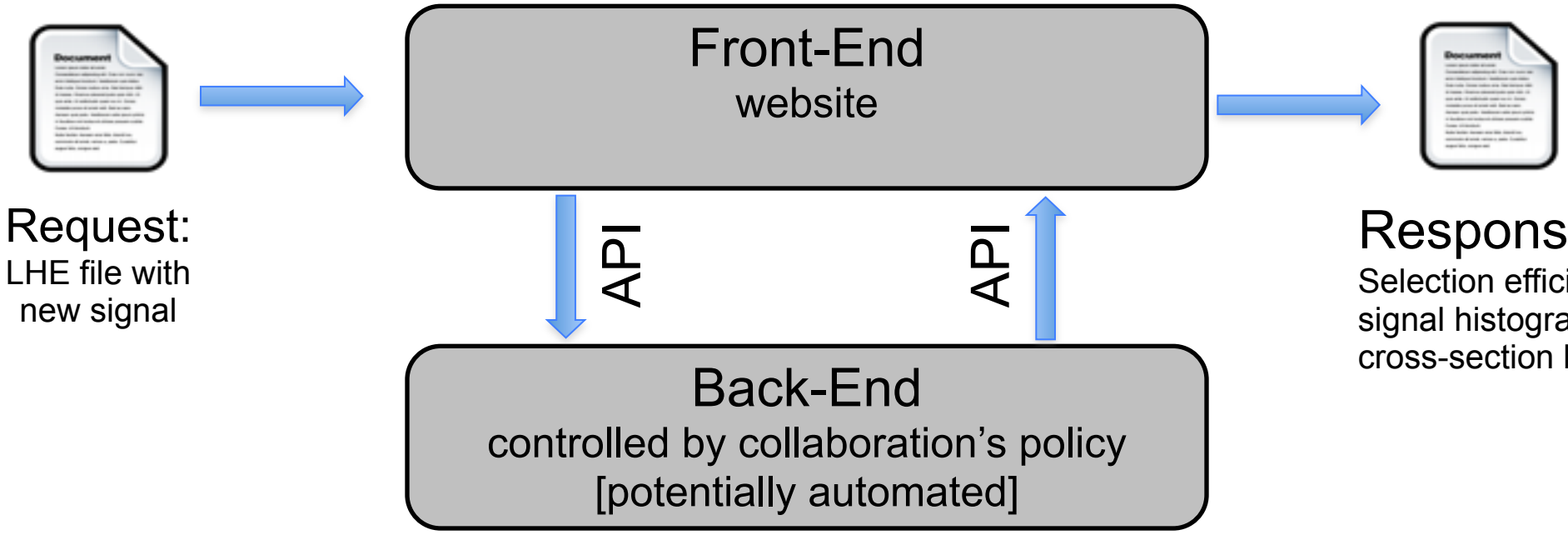
K.C., Itay Yavin [hep-ex/1010.2506], JHEP.

RECAST is a framework for recasting that aims to collect, standardize, and facilitate the processing of recast requests from the community.

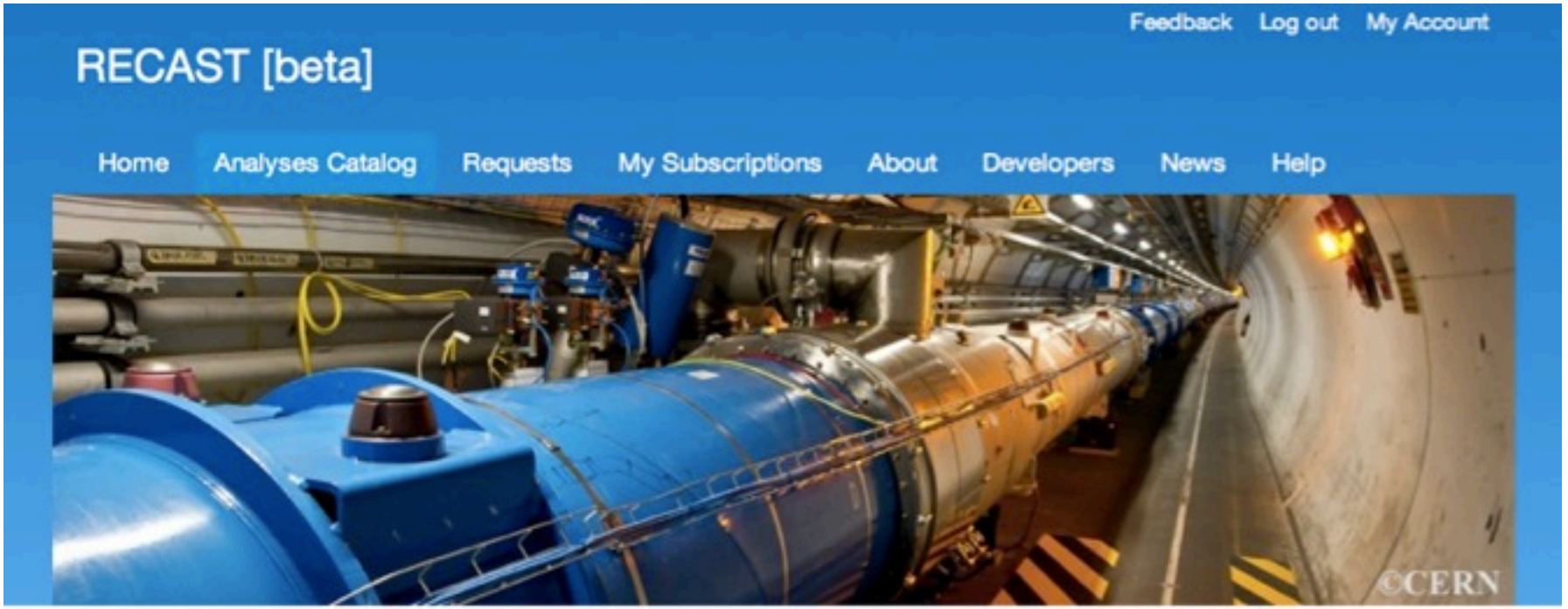
- cuts don't change: re-use background estimates and observation from original analysis
- what is needed is to archive the analysis cuts & provide a pipeline for new signal
- data is kept private, still goes through necessary approval process as determined by collaboration, original paper receives citation & recognition (doi's tracked by INSPIRE)

RECAST front-end is a website that collects and organizes the Requests and Responses

- standardizes request & response format, API allows for process to be automated,
- back-end implementation is up to collaboration



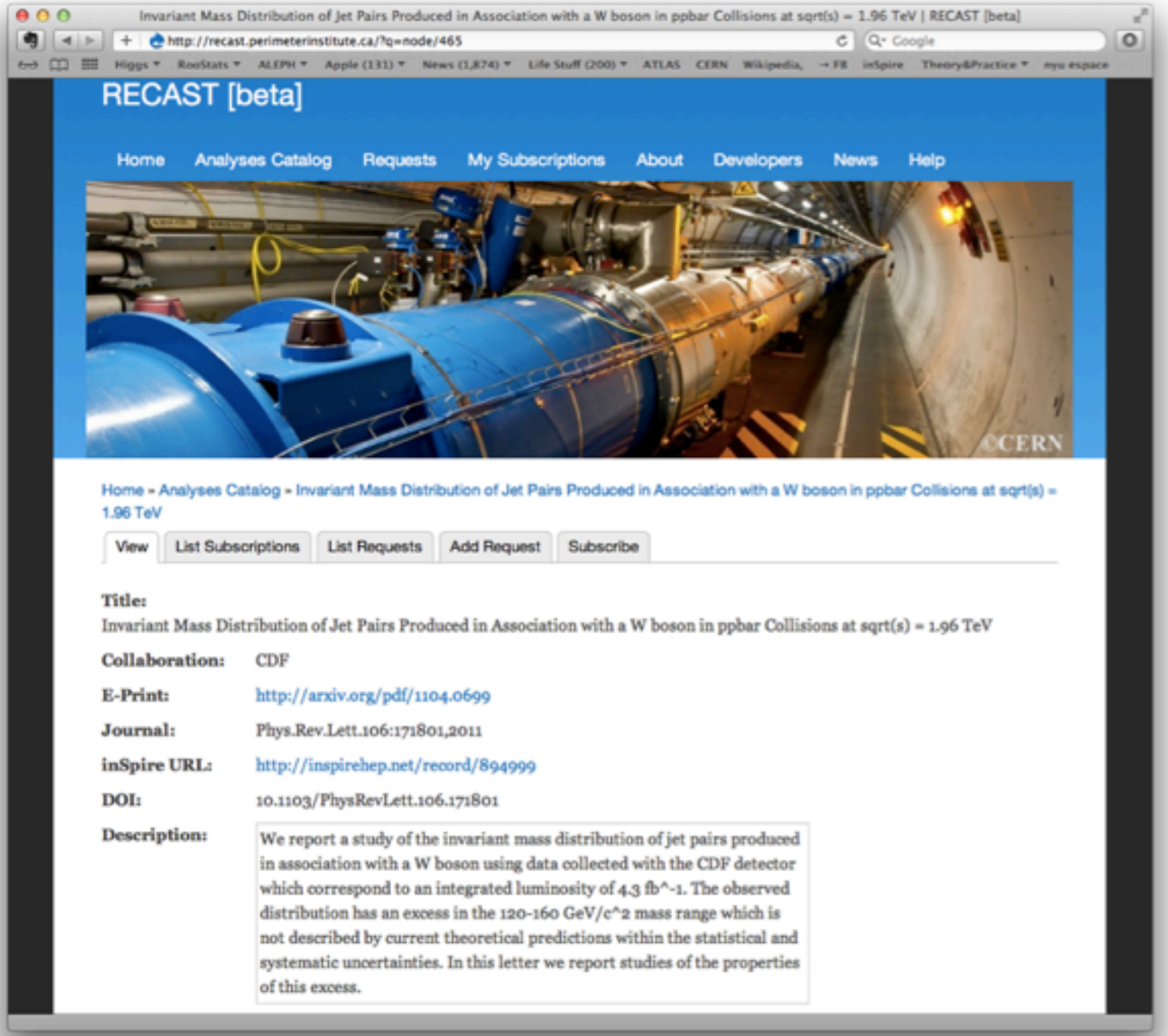
http://recast.perimeterinstitute.ca



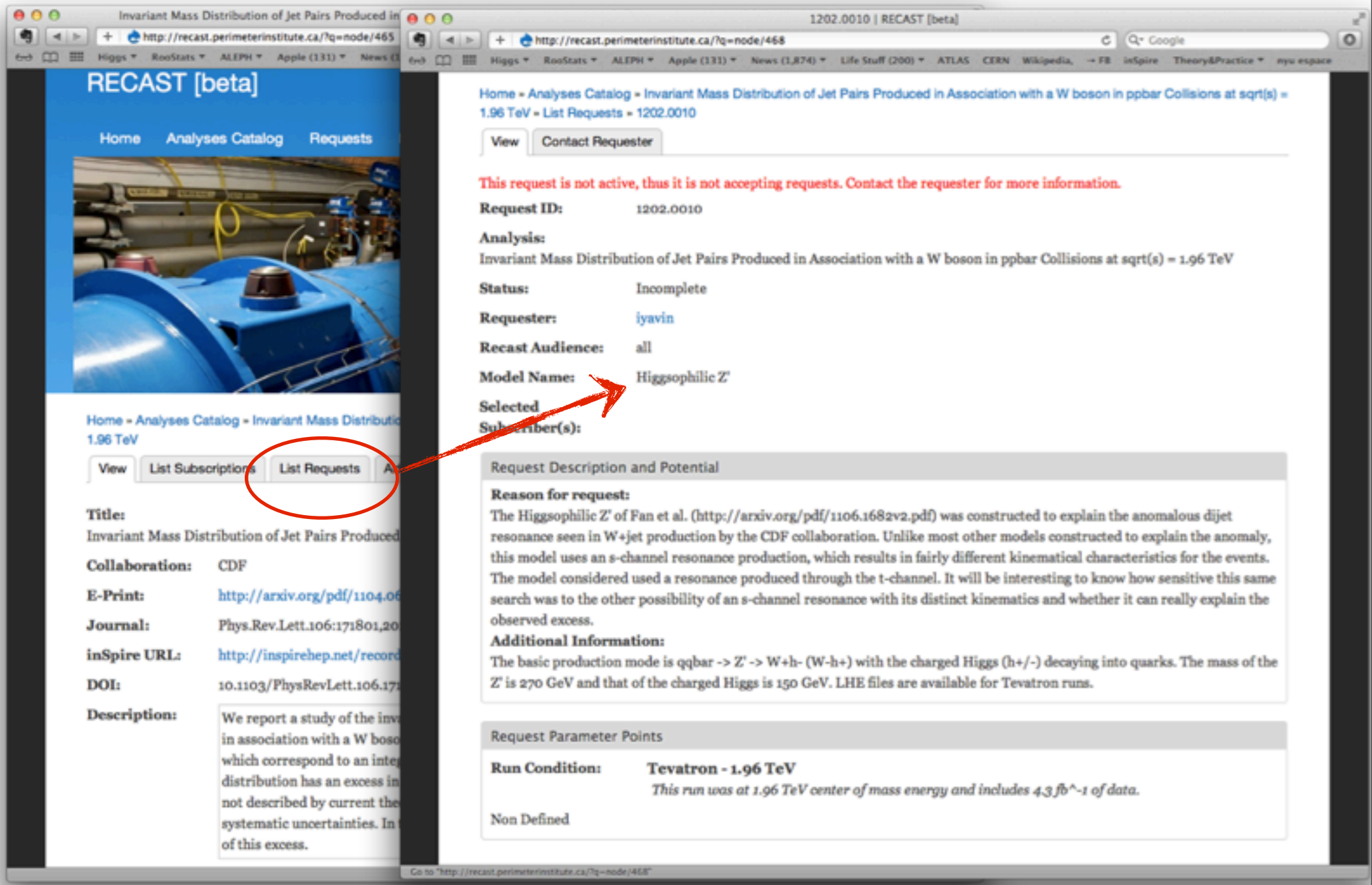
Analysis Catalog | New Analysis

Analysis	Collaboration	Subscribers	Requests	Active	Completed	Actions
Search for Diphoton Events with Large Missing Transverse Momentum in 1 fb^-1 of 7 TeV Proton-Proton Collision Data with the ATLAS Detector	Atlas	1	1	0	1	view | edit | un-subscribe | add request
Search for New Phenomena in Events with Three or more Charged Leptons	Atlas	0	0	0	0	view | edit | subscribe | add request
Multileptonic SUSY searches	CMS	2	1	1	0	view | un-subscribe | add request
Search for Anomalous Production of Multilepton Events and R-Parity-Violating Supersymmetry in √s = 7 TeV	CMS	0	0	0	0	view | subscribe | add request
Search for a Vector-like Quark with Charge 2/3 in t + Z Events from pp Collisions at sqrt(s) = 7 TeV	CMS	0	0	0	0	view | subscribe | add request

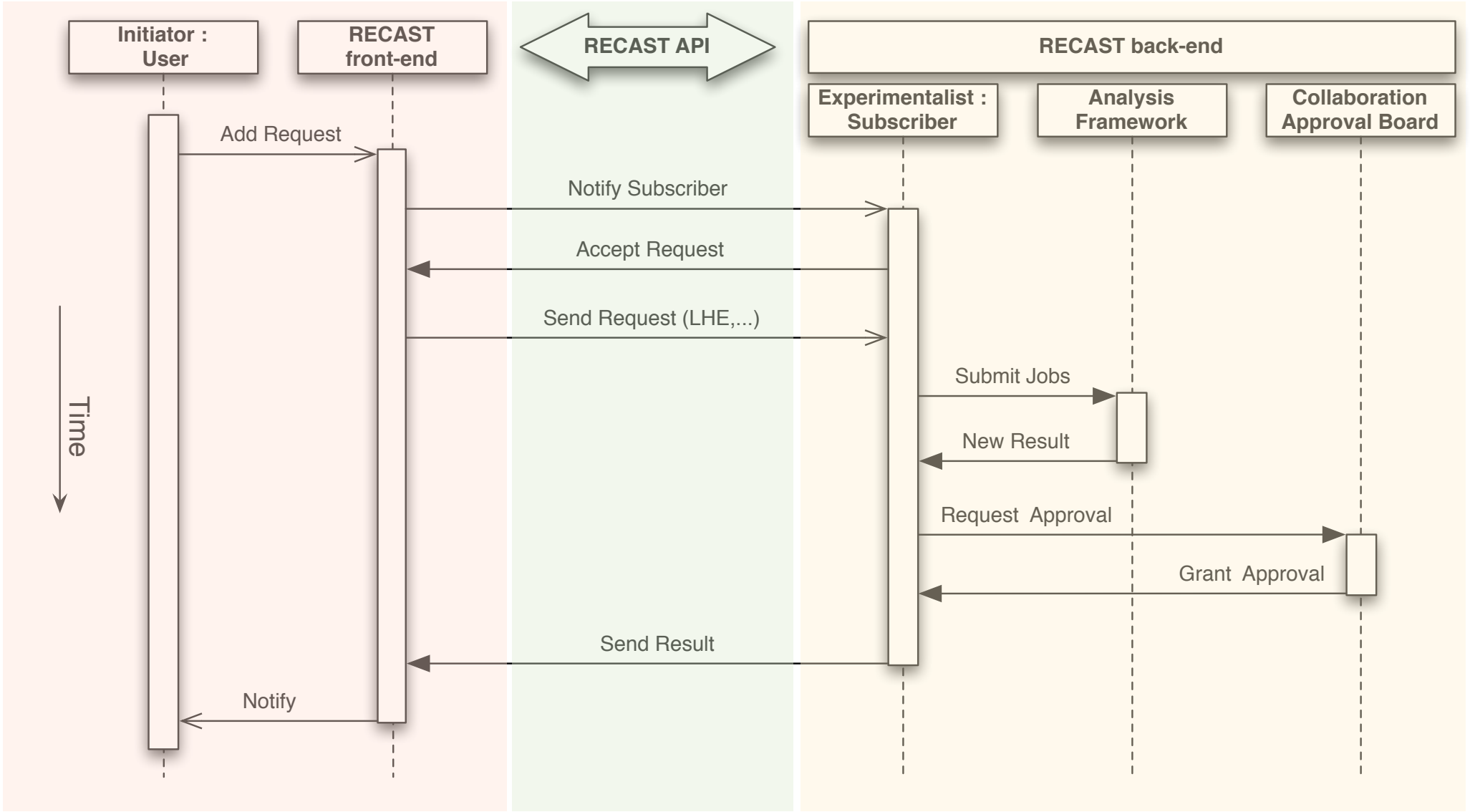
A CDF Example



A CDF Example



High Level Design



K.C., Itay Yavin [hep-ex/1010.2506], JHEP.

A beta versions of the front-end website and API have been developed thanks to support from the Perimeter Institute.

Example API from command line

Documentation: http://recast.perimeterinstitute.ca/sites/default/files/PIRESTfulWebService_v1.1.pdf

```
cranmer@172-26-28-145:~$ curl http://recast.perimeterinstitute.ca/api/recast-analysis
-
   uuid: c3afcb6c-ce2f-9104-ed00-6a5cfb02a401
   title: >
      Invariant Mass Distribution of Jet Pairs Produced in
      Association with a W boson in ppbar Collisions at sqrt(s) =
      1.96 TeV
   number_of_requests: 1
   collaboration: CDF
```

Supports XML, JSON, and YAML

```
cranmer@172-26-28-145:~$ curl http://recast.perimeterinstitute.ca/api/recast-request/71003820-bfec-aae4-25fd-76c713993f50
---
-
    uuid: 71003820-bfec-aae4-25fd-76c713993f50
    title: 1202.0010
    requestor: iyavin
    status: 0
    analysis-uuid: c3afcb6c-ce2f-9104-ed00-6a5cfb02a401
    audience:
    subscribers:
    predefined-model: Higgsophilic Z'
    new-model-information:
    reason-for-request: >
        The Higgsophilic Z' of Fan et al.
        (http://arxiv.org/pdf/1106.1682v2.pdf) was constructed to
        explain the anomalous dijet resonance seen in W+jet
        production by the CDF collaboration. Unlike most other
        models constructed to explain the anomaly, this model uses
        an s-channel resonance production, which results in fairly
        different kinematical characteristics for the events. The
        model considered used a resonance produced through the
        t-channel. It will be interesting to know how sensitive this
        same search was to the other possibility of an s-channel
        resonance with its distinct kinematics and whether it can
        really explain the observed excess.
    additional-information: "The basic production mode is qqbar -> Z' -> W+h- (W-h+) with the charged Higgs (h+/-) decaying into
quarks. The mass of the Z' is 270 GeV and that of the charged Higgs is 150 GeV. LHE files are available for Tevatron runs."
    model-type:
    parameter-points: [ ]
```

What is needed to put analysis into RECAST?

Reminder:

- input to the chain is a standard (Les Houches) input file of signal events
- output is the signal expectation (eg. histogram) after selection cuts, ideally propagated through to a cross-section limit on events of this type.

One must run the standard chain:

- pythia -> simulation -> reconstruction (-> ntuplization)
  - collaboration could choose to use fast simulation here
  - Perimeter has offered a dedicated cluster to support simulation needed for recasting
- hard part: selection code needs to be preserved
  - Avoid over-engineering, a tarball with a makefile that can process input.root and produce the output signal histogram can suffice.
  - must capture analysis code shortly after approval of result
- hard part: collaboration approval
  - ideally a stream-lined re-approval process since the selection, background estimate, and data have not changed

In principle, first steps can be automated, but start with a by-hand effort

Making this happen

The investment by the Perimeter Institute to produce the beta version of the web site and API was a major step (completed Dec. 2011)

- theorists are encouraged to start submitting requests

The most important issue for CDF right now is to capture the event selection code for the final analyses before they are subject to 'code rot'

- book-keeping information about which ntuples are used so that one can create ntuples of the same format for a new signal
- event selection code
- framework for processing requests can be come later

Computing resources are a concern:

- remember, collaborations choose which requests to respond to
- the Perimeter Institute has offered access to 100 CPU cores that can be used by the collaborations for RECAST (details to be worked out, but think of it as an external tier-3 dedicated to RECAST)

Usage patterns to be determined:

- maybe the most important usage of RECAST will be to provide a few validated 'anchor points' for larger parameters scans based on fast simulation

Interaction with INSPIRE & HEPData

INSPIRE: Advanced discussions regarding the interaction of RECAST and INSPIRE.

- New responses get DOI number from CERN, are aggregated with original collaboration publication. New responses are citable objects, and increment citation of original paper
- New requests (from theorists) get their own DOI number, so theorist can put out paper pointing to their request independently of a collaboration response

HEPData: Signal histograms from the new model could go into HEPData.

Les Houches: Meeting on Feb. 13 discussing presentation of LHC results

Searches for New Physics: Les Houches Recommendations for the Presentation of LHC Results

We also note at this point that the RECAST [26] project would allow one to obtain the signal contribution to the likelihood for an arbitrary theoretical model, thus allowing one to build a higher-level framework for analysis re-interpretation.

Backup

Examples & Interest

HEP 6 records found

1. **Electroweakinos Hiding in Higgs Searches.**
Mariangela Lisanti, Neal Weiner. arXiv:1112.4834. Dec 2011. 7 pp.
e-Print: **arXiv:1112.4834 [hep-ph]**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Abstract and Postscript and PDF from arXiv.org
Detailed record

2. **Natural SUSY Endures.**
Michele Papucci, Joshua T. Ruderman (LBL, Berkeley & UC, Berkeley), Andreas Weiler (DESY & CERN). DESY-11-193
e-Print: **arXiv:1110.6926 [hep-ph]**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Abstract and Postscript and PDF from arXiv.org
Detailed record - Cited by 23 records

3. **The Status of GMSB After 1/fb at the LHC.**
Yevgeny Kats, Patrick Meade, Matthew Reece, David Shih. RUNHETC-2011-19, YITP-SB-11-36. Oct 2011. 43 pp.
e-Print: **arXiv:1110.6444 [hep-ph]**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Abstract and Postscript and PDF from arXiv.org
Detailed record - Cited by 18 records

4. **Simplified Models for LHC New Physics Searches.**
LHC New Physics Working Group Collaboration (Daniele Alves (SLAC) *et al.*). arXiv:1105.2838. May 2011. 40 pp.
e-Print: **arXiv:1105.2838 [hep-ph]**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Abstract and Postscript and PDF from arXiv.org
Detailed record - Cited by 38 records

5. **LHC Charge Asymmetry as Constraint on Models for the Tevatron Top Anomaly.**
Nathaniel Craig (Princeton, Inst. Advanced Study & Rutgers U., Piscataway), Can Kilic, Matthew J. Strassler (Rutgers U.
Published in **Phys.Rev. D84 (2011) 035012**
e-Print: **arXiv:1103.2127 [hep-ph]**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Abstract and Postscript and PDF from arXiv.org; Journal Server - Phys.Rev.
Detailed record - Cited by 25 records

6. **Discovering a Higgs boson decaying to four jets in supersymmetric cascade decays.**
Brando Bellazzini, Csaba Csaki (Cornell U., CIHEP), Jay Hubisz, Jing Shao (Syracuse U.). Dec 2010. 10 pp.
Published in **Phys.Rev. D83 (2011) 095018**
e-Print: **arXiv:1012.1316 [hep-ph]**
References | BibTeX | LaTeX(US) | LaTeX(EU) | Harvmac | EndNote
Abstract and Postscript and PDF from arXiv.org; Journal Server - Phys.Rev.
Detailed record - Cited by 10 records

Latest Requests

Request	Analysis	Model
1111.0012	Search for Supersymmetry in pp Collisions at 7 TeV in Events with Jets and Missing Transverse Energy	Littlest Higgs with T-Parity

These are examples from theorists requesting recasts before the website was running.

Request examples

Latest Requests

Request	Analysis	Model	Status
1111.0012	Search for Supersymmetry in pp Collisions at 7 TeV in Events with Jets and Missing Transverse Energy	Littlest Higgs with T-Parity	Completed
1202.0008	Search for New Phenomena in ttbar Events With Large Missing Transverse Momentum in Proton-Proton Collisions at sqrt(s) = 7 TeV with the ATLAS Detector	Light stops	Incomplete
1202.0004	Search for the Standard Model Higgs boson in the H->WW->llnunu decay mode using 1.7 fb-1 of data collected with the ATLAS detector at sqrt(s)=7 TeV	Light Electroweakinos	Active
1202.0008	Constraints on Anomalous μ+μ+ and Like-sign Top Quark Pair Production	B-L Spontaneous R-parity Violation	Incomplete
1202.0006	Multileptonic SUSY searches	B-L Spontaneous R-parity Violation	Incomplete
1202.0005	Search for New Phenomena in Events with Three or more Charged Leptons	B-L Spontaneous R-parity Violation	Incomplete
1112.0002	Search for Diphoton Events with Large Missing Transverse Momentum in 1 fb^-1 of 7 TeV Proton-Proton Collision Data with the ATLAS Detector	Demo Model	Completed
1202.0007	Search for New Phenomena in Events with Four Charged Leptons	B-L Spontaneous R-parity Violation	Incomplete
1202.0010	Invariant Mass Distribution of Jet Pairs Produced in Association with a W boson in ppbar Collisions at sqrt(s) = 1.96 TeV	Higgsophilic Z'	Incomplete
1202.0009	Bounds on an anomalous dijet resonance in W+jets production in ppbar collisions at sqrt{s} =1.96 TeV	Higgsophilic Z'	Incomplete

Since announcing the RECAST website more widely in the last week, the first requests have started to come in.

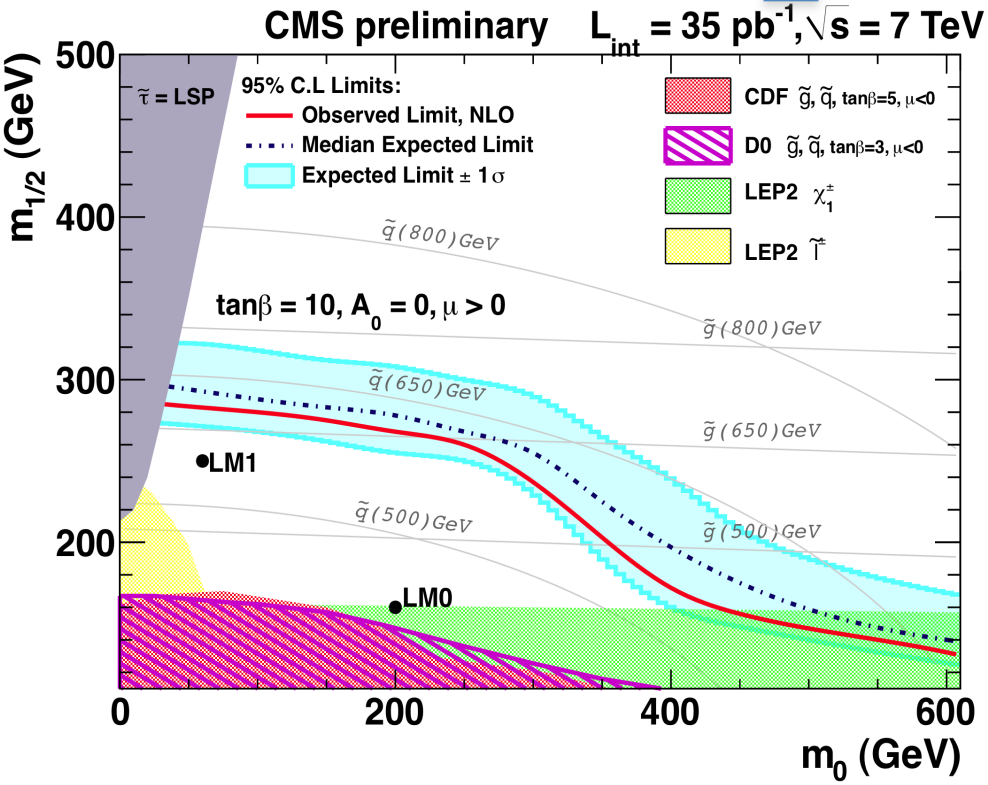
- here one for CDF Wjj analysis

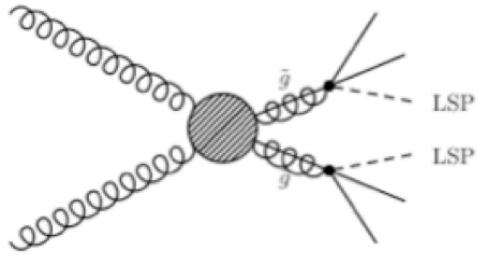
Examples of Recasting

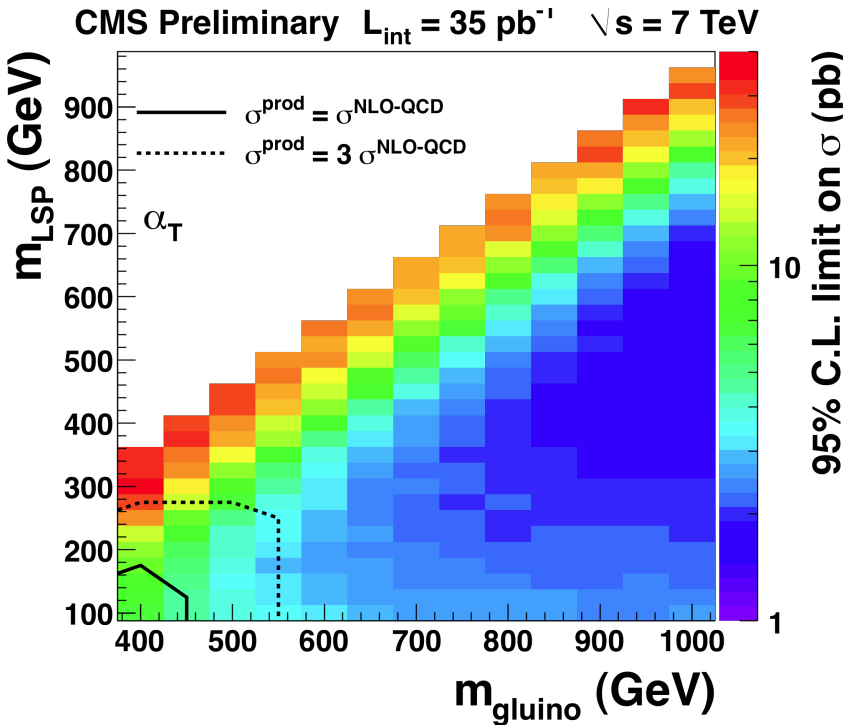
Recasting simplified models

Current practice by SUSY searches to present alternative interpretations is an example of the recasting technique

this requires estimate of signal efficiency for alternative model

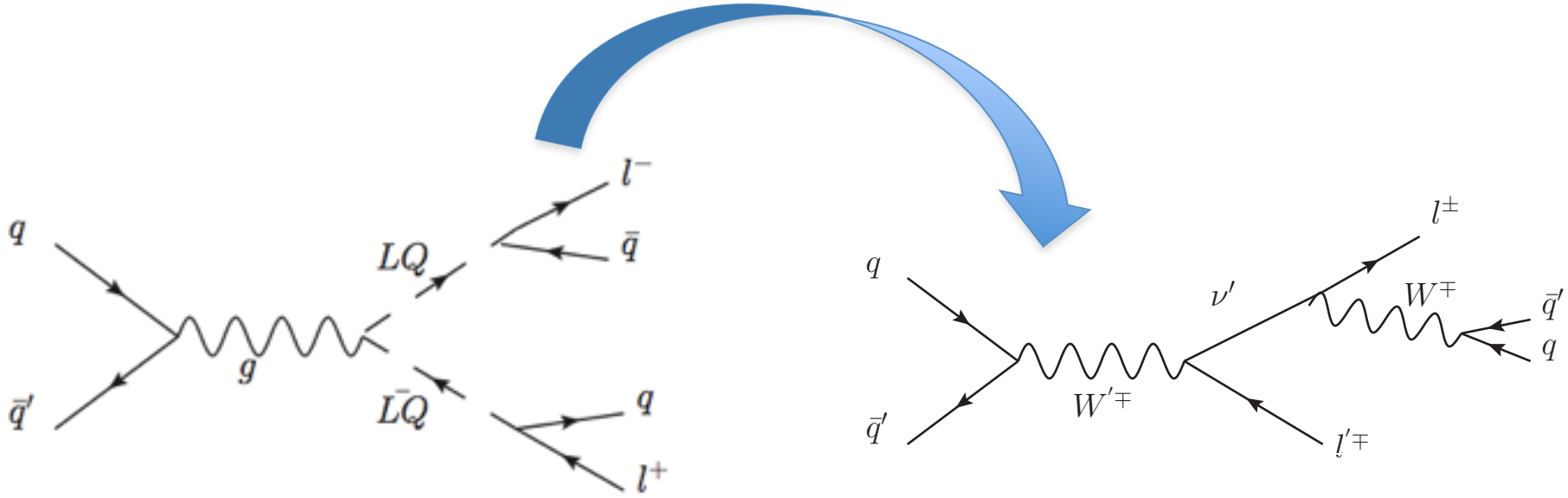






W' hunt from Leptoquark search

M. Schmaltz and C. Spethmann suggested a recast of a leptoquark search that was done by DØ to place bounds on W' particles expected in Little Higgs theories,

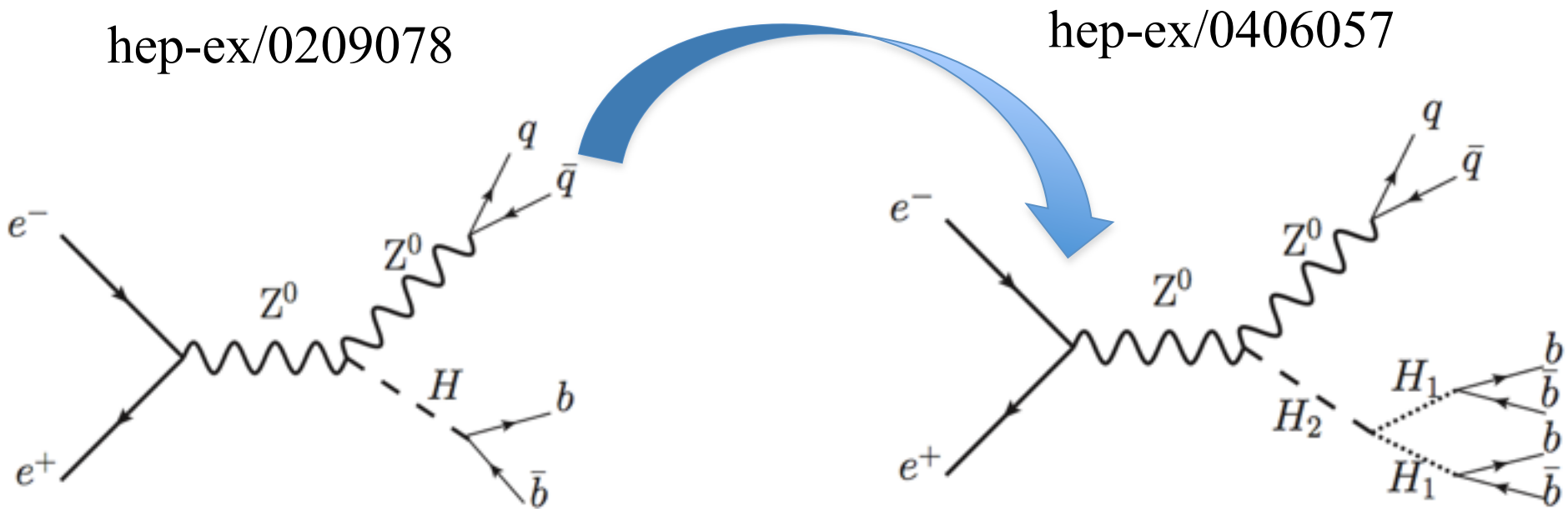


Phys.Lett.B636
hep-ex/0601047

????.????

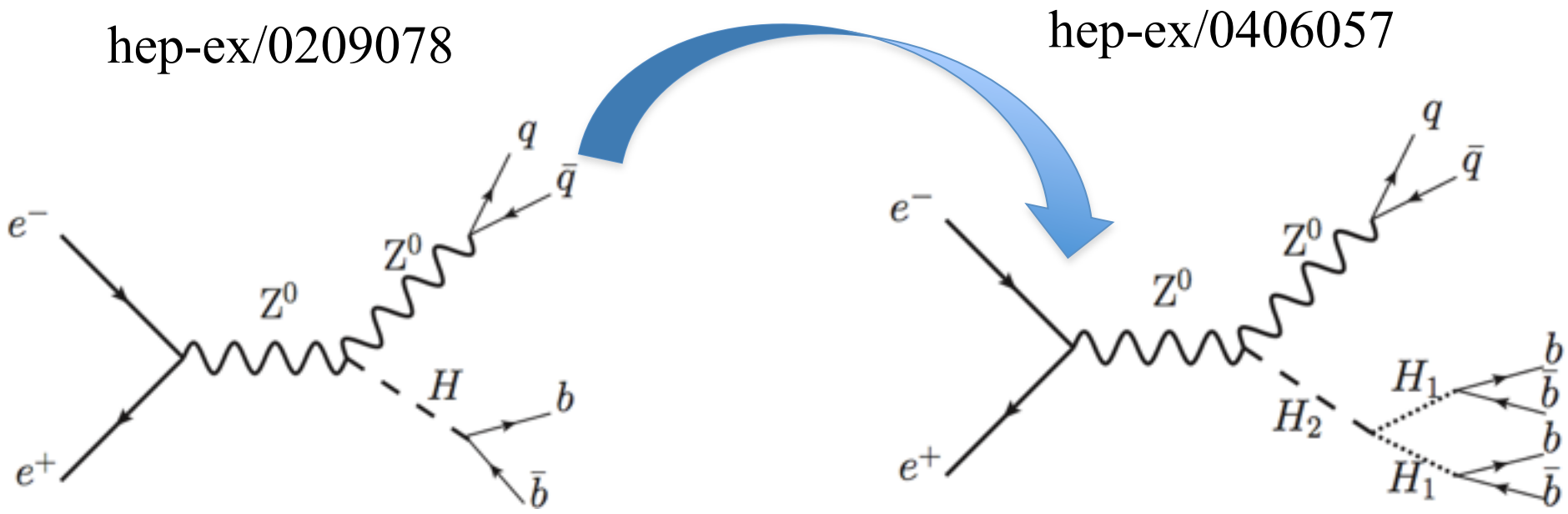
OPAL Higgs Searches

In hep-ex/0406057 OPAL recasted a previous search for Standard Model Higgs to place constraints on MSSM Higgs scenarios



OPAL Higgs Searches

In hep-ex/0406057 OPAL recasted a previous search for Standard Model Higgs to place constraints on MSSM Higgs scenarios

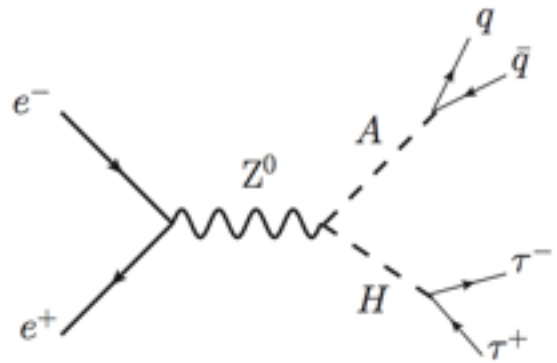


Efficient recasting

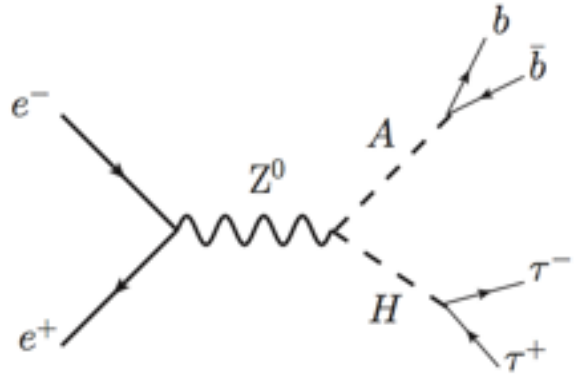
$m_{\mathcal{H}_2}$ (GeV)	$m_{\mathcal{H}_1}$ (GeV)	Efficiency for the process $\mathcal{H}_2Z \to b\bar{b}b\bar{b}q\bar{q}$ at $\sqrt{s}$				
		192 GeV	196 GeV	200 GeV	202 GeV	206 GeV
100.	12.	0.689	0.684	0.717	0.733	0.693
100.	20.	0.651	0.639	0.653	0.659	0.586
100.	30.	0.460	0.461	0.461	0.470	0.480
100.	40.	0.270	0.260	0.283	0.315	0.323
100.	48.	0.328	0.325	0.361	0.392	0.400

DELPHI Higgs Searches

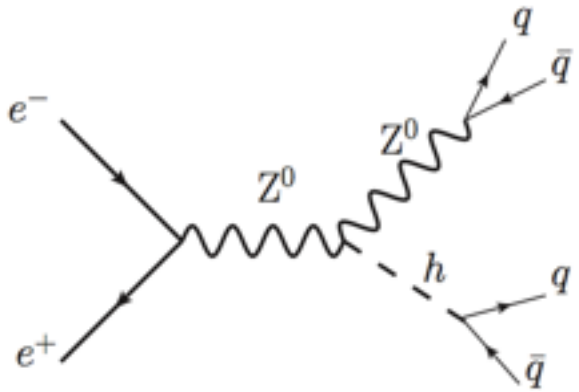
Similar recasting of previous SM Higgs searches was done at DELPHI



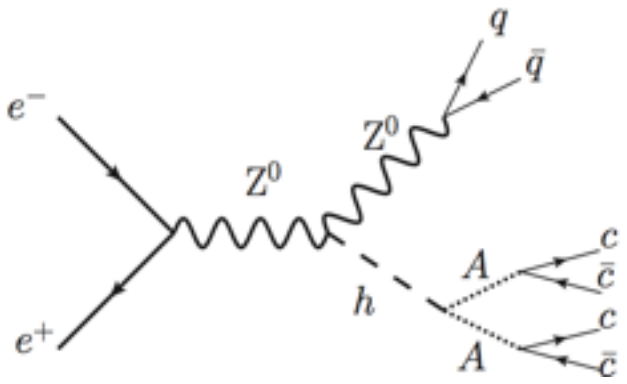
DELPHI Col., Eur. Phys. J. C38 (2004)



DELPHI Col., Eur.Phys.J. C54 (2008)



DELPHI Col., Eur. Phys. J. C23 (2002)

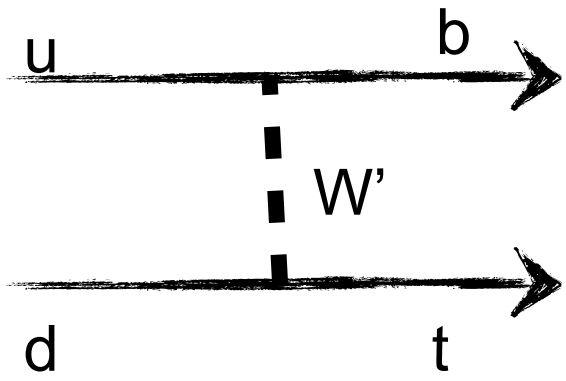


DELPHI Col., Eur.Phys.J. C54 (2008)

Recasting single-top

Several models for A_{FB}^{tt} have a t-channel Z' or W' with unusual flavor couplings

- often implies new single-top production modes
- existing single-top analyses with 35 pb^{-1} are already sensitive to these models



Signal region for these models