A Meta-Analysis of LHC Results

Sevim Açıksöz[1]
Erkcan Özcan[1,2]

1. **Boğaziçi Üniversitesi**, Dept. of Physics, Bebek, İstanbul
2. **Feza Gürsey Center for Physics & Mathematics**, Boğaziçi University, Üsküdar, İstanbul

1. Motivation
2. What is an Exclusion Plot?
3. Vectorial Analysis
4. Bitmap Analysis
5. Problems
6. Preliminary Results

Motivation

IN FOCUS NEWS

REPRODUCIBILITY

Statisticians issue warning on *P* values

Statement aims to halt missteps in the quest for certainty.

BY MONYA BAKER

Misuse of the *P* value — a common test for judging the strength of scientific evidence — is contributing to the number of research findings that cannot be reproduced, the American Statistical Association (ASA) warned on 8 March. The group has taken the unusual step of issuing principles to guide use of the *P* value, which it says cannot determine whether a hypothesis is true or whether results are important.

This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter, says executive director Ron Wasserstein. The society's members had become increasingly concerned that the *P* value was being misapplied, in ways that cast doubt on statistics generally, he adds.

In its statement, the ASA advises researchers to avoid drawing scientific conclusions or making policy decisions purely on the basis of *P* values (R. L. Wasserstein and N. A. Lazar *Am. Stat.* http://doi.org/bc6d; 2016). Researchers should describe not only the data analyses that produce statistically significant results, the society says, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust.

Véronique Kiermer, executive editor of the Public Library of Science journals, says that the ASA's statement lends weight and visibility to longstanding concerns over undue reliance on the *P* value. "It is also very important in that it shows statisticians, as a profession, engaging with the problems in the literature outside of their field," she adds.

P values are commonly used to test (and dismiss) a 'null hypothesis', which generally states that there is no difference between two groups, or that there is no correlation between a pair of characteristics. The smaller the *P* value, the less likely an observed set of values would occur by chance — assuming that the null hypothesis is true. A *P* value of 0.05 or less is generally taken to mean that a finding is statistically significant and warrants publication. But that is not necessarily true, the ASA statement notes.

A *P* value of 0.05 does not mean that there is a 95% chance that a given hypothesis is correct. Instead, it signifies that if the null hypothesis is true, and all other assumptions made are valid, there is a 5% chance of obtaining a result at least as extreme as the one observed. And a *P* value cannot indicate the importance of a finding; for instance, a drug can have a statistically significant effect on patients' blood glucose levels without having a therapeutic effect.

Giovanni Parmigiani, a biostatistician at the Dana Farber Cancer Institute in Boston, Massachusetts, says that misunderstandings about what information a *P* value provides often crop up in textbooks and practice manuals. A course correction is long overdue, he adds. "Surely if this happened twenty years ago, biomedical research could be in a better place now."

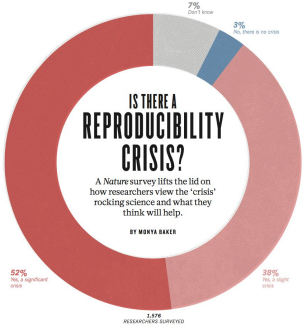
FRUSTRATION ABOUNDS

Criticism of the *P* value is nothing new. In 2011, researchers trying to raise awareness about false positives gamed an analysis to reach a statistically significant finding: that listening to music by the Beatles makes undergraduates younger (J. P. Simmons *et al. Psychol. Sci.* **22**, 1359–1366; 2011). More controversially, in 2015, a set of documentary filmmakers published conclusions from a purposely shoddy clinical trial — supported by a robust *P* value — to show that eating chocolate helps people to lose weight. (The article has since been retracted.)

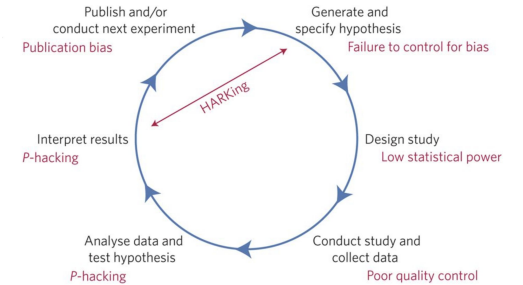
But Simine Vazire, a psychologist at the University of California, Davis, and editor of the journal *Social Psychological and Personality Science*, thinks that the ASA statement could help to convince authors to disclose all of the statistical analyses that they run. "To the extent that people might be sceptical, it helps to have statisticians saying, 'No, you can't interpret *P* values without this information,'" she says.

More drastic steps, such as a ban on publishing *P* values in articles instituted by at least one journal, could be counter-productive, says Andrew Vickers, a biostatistician at Memorial Sloan Kettering Cancer Center in New York City. He compares attempts to bar the use of *P* values to addressing the risk of automobile accidents by warning people not to drive — a message that many in the target audience would probably ignore. Instead, Vickers says that researchers should be instructed to "treat statistics as a science, and not a recipe".

But a better understanding of the *P* value will not take away the human impulse to use statistics to create an impossible level of certainty, warns Andrew Gelman, a statistician at Columbia University in New York City. "People want something that they can't really get," he says. "They want certainty." ■



From: A manifesto for reproducible science



An idealized version of the hypothetico-deductive model of the scientific method is shown. Various potential threats to this model exist (indicated in red), including lack of replication[5], hypothesizing after the results are known (HARKing)[7], poor study design, low statistical power[2], analytical flexibility[51], *P*-hacking[4], publication bias[3] and lack of data sharing[6]. Together these will serve to undermine the robustness of published research, and may also impact on the ability of science to self-correct.

Reproducibility in science has become a serious issue in the recent years. Even research published in top journals has become suspect.

Part of the problem is p-hacking, intentional or otherwise.

How is the field of experimental HEP doing? Are we able to estimate our uncertainties fairly? Can we be a good example to the wider scientific world?

nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0399-z

Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015

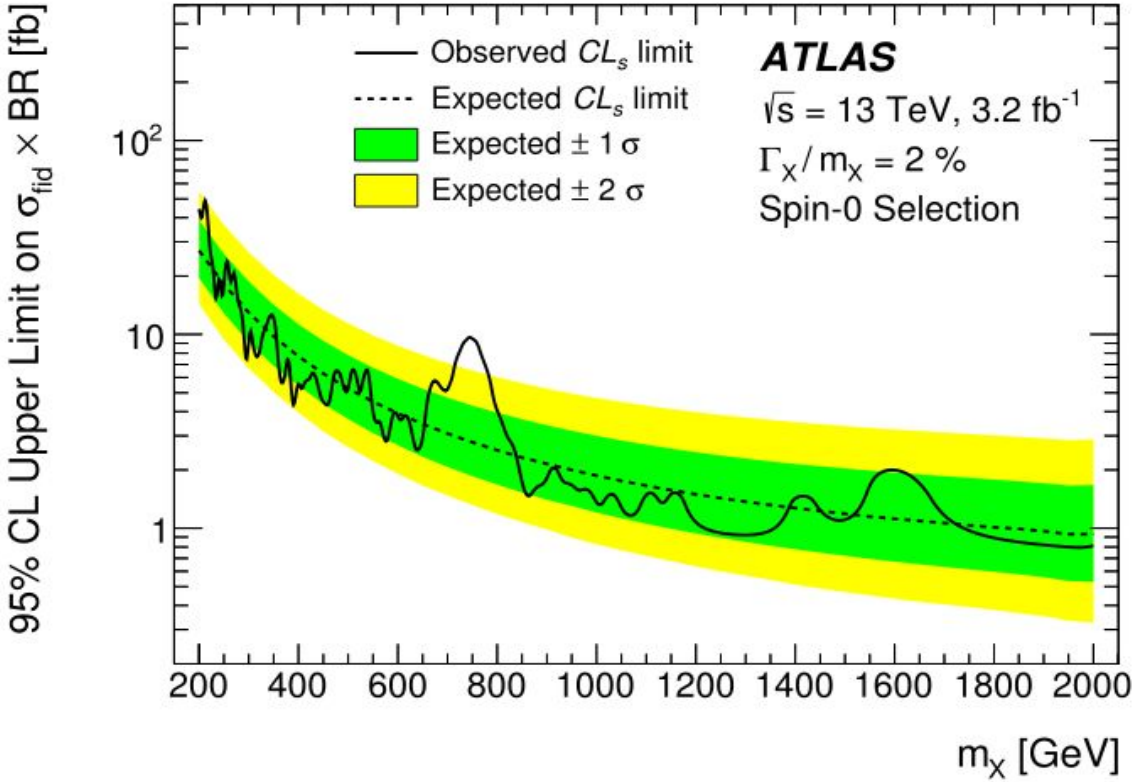
Colin F. Camerer[1,16], Anna Dreber[2,16], Felix Holzmeister[3,16], Teck-Hua Ho[4,16], Jürgen Huber[3,16], Magnus Johannesson[2,16], Michael Kirchler[3,5,16], Gideon Nave[6,16], Brian A. Nosek[7,8,16*], Thomas Pfeiffer[9,16], Adam Altmejd[2], Nick Buttrick[7,8], Taizan Chan[10], Yiling Chen[11], Eskil Forsell[12], Anup Gampa[7,8], Emma Heikensten[2], Lily Hummer[8], Taisuke Imai[13], Siri Isaksson[2], Dylan Manfredi[6], Julia Rose[3], Eric-Jan Wagenmakers[14] and Hang Wu[15]

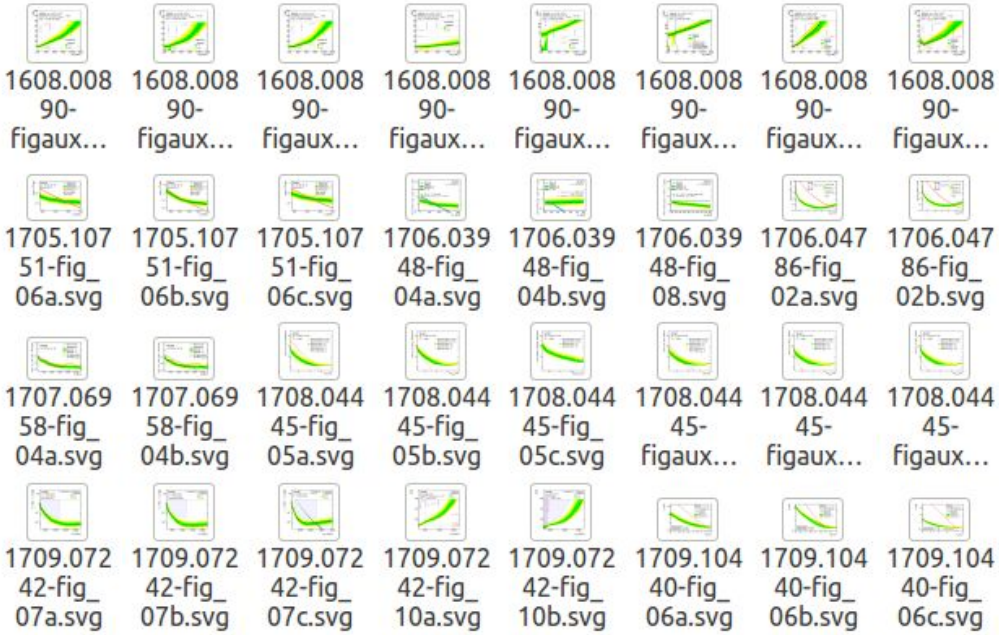
Being able to replicate scientific findings is crucial for scientific progress[1–15]. We replicate 21 systematically selected experimental studies in the social sciences published in *Nature* and *Science* between 2010 and 2015[16–36]. The replications follow analysis plans reviewed by the original authors and pre-registered prior to the replications. The replications are high powered, with sample sizes on average about five times higher than in the original studies. We find a significant effect in the same direction as the original study for 13 (62%) studies, and the effect size of the replications is on average about 50% of the original effect size. Replicability varies between 12 (57%) and 14 (67%) studies for complementary replicability indicators. Consistent with these results, the estimated true-positive rate is 67% in a Bayesian analysis. The relative effect size of true positives is estimated to be 71%, suggesting that both false positives and inflated effect sizes of true positives contribute to imperfect reproducibility. Furthermore, we find that peer beliefs of replicability are strongly related to replicability, suggesting that the research community could predict which results would replicate and that failures to replicate were not the result of chance alone.

a significant effect in the same direction as the original studies for 61% of replications[15]. Both the RPP and the EERP had high statistical power to detect the effect sizes observed in the original studies. However, the effect sizes of published studies may be inflated even for true-positive findings owing to publication or reporting biases[40–42]. As a consequence, if replications were well powered to detect effect sizes smaller than those observed in the original studies, replication rates might be higher than those estimated in the RPP and the EERP.

We provide evidence about the replicability of experimental studies in the social sciences published in the two most prestigious general science journals, *Nature* and *Science* (the Social Sciences Replication Project (SSRP)). Articles published in these journals are considered exciting, innovative and important. We include all experimental studies published between 2010 and 2015 that (1) test for an experimental treatment effect between or within subjects, (2) test at least one clear hypothesis with a statistically significant finding, and (3) were performed on students or other accessible subject pools. Twenty-one studies were identified to meet these criteria. We used the following three criteria in descending order to determine which treatment effect to replicate

Exclusion Plots





ATLAS

394 Plots from 63 articles (2016-2021)

CMS

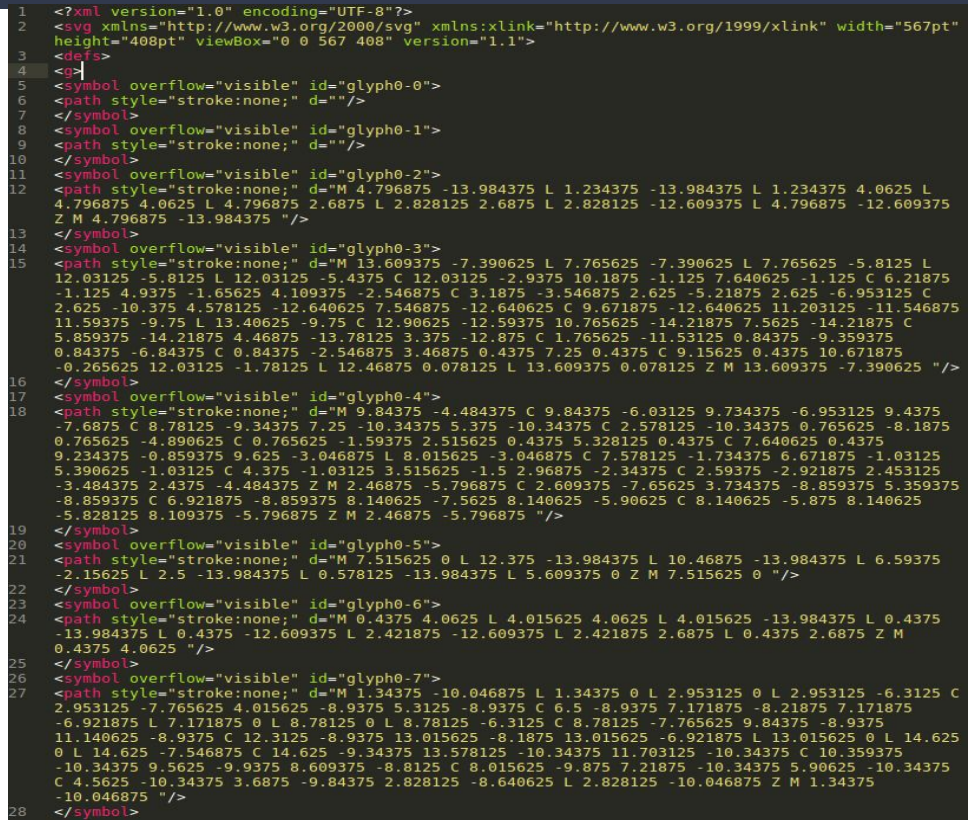
441 Plots from 86 articles (2016-2021)

Mostly Exotics Papers

Vector Graphics

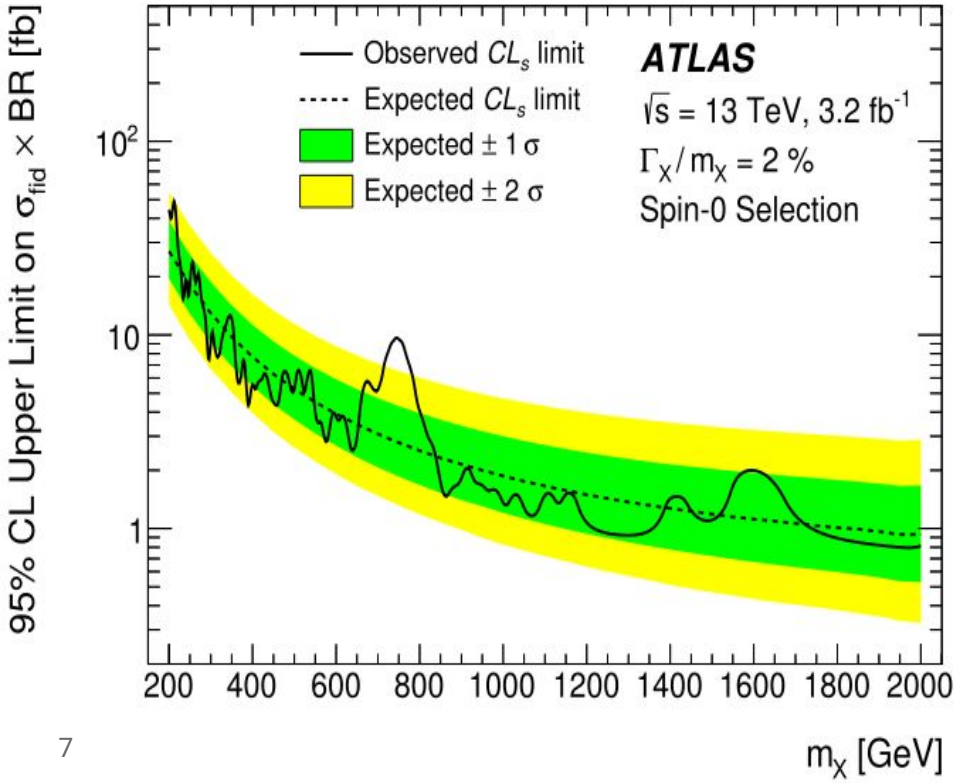
- EPS and PDF files are converted to SVG format by using Linux TEX packages eps2pdf and pdf2svg
- SVG files are composed of paths and their attributes.
- Python library svgpathtools to parse svg files.
- First step to extract needed lines is filtering objects' attributes with respect to RGB decimal color codes.

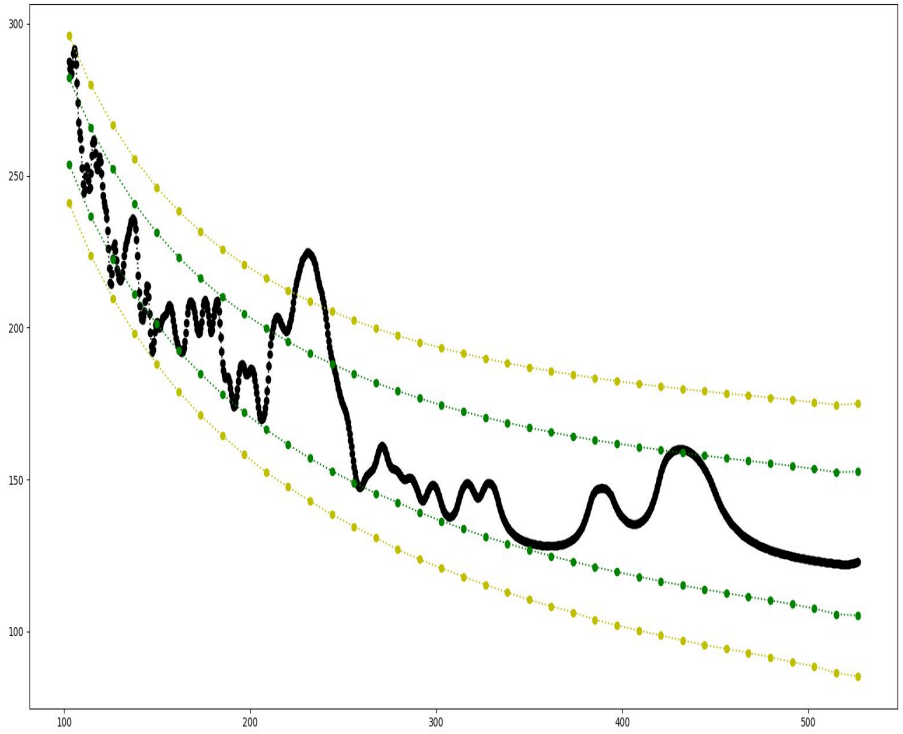
  Green : rgb(0%,100%,0%)
  Yellow : rgb(100%,100%,0%)
  Black : rgb(0%,0%,0%)



```
<?xml version="1.0" encoding="UTF-8"?>
<svg xmlns="http://www.w3.org/2000/svg" xmlns:xlink="http://www.w3.org/1999/xlink" width="567pt" height="408pt" viewBox="0 0 567 408" version="1.1">
<defs>
<g>
<symbol overflow="visible" id="glyph0-0">
<path style="stroke:none;" d=""/>
</symbol>
<symbol overflow="visible" id="glyph0-1">
<path style="stroke:none;" d=""/>
</symbol>
<symbol overflow="visible" id="glyph0-2">
<path style="stroke:none;" d="M 4.796875 -13.984375 L 1.234375 -13.984375 L 1.234375 4.0625 L 4.796875 4.0625 L 4.796875 2.6875 L 2.828125 2.6875 L 2.828125 -12.609375 L 4.796875 -12.609375 Z M 4.796875 -13.984375 "/>
</symbol>
<symbol overflow="visible" id="glyph0-3">
<path style="stroke:none;" d="M 13.609375 -7.390625 L 7.765625 -7.390625 L 7.765625 -5.8125 L 12.03125 -5.8125 L 12.03125 -5.4375 C 12.03125 -2.9375 10.1875 -1.125 7.640625 -1.125 C 6.21875 -1.125 4.9375 -1.65625 4.109375 -2.546875 C 3.1875 -3.546875 2.625 -5.21875 2.625 -6.953125 C 2.625 -10.375 4.578125 -12.640625 7.546875 -12.640625 C 9.671875 -12.640625 11.203125 -11.546875 11.59375 -9.75 L 13.40625 -9.75 C 12.90625 -12.59375 10.765625 -14.21875 7.5625 -14.21875 C 5.859375 -14.21875 4.46875 -13.78125 3.375 -12.875 C 1.765625 -11.53125 0.84375 -9.359375 0.84375 -6.84375 C 0.84375 -2.546875 3.46875 0.4375 7.25 0.4375 C 9.15625 0.4375 10.671875 -0.265625 12.03125 -1.78125 L 12.46875 0.078125 L 13.609375 0.078125 Z M 13.609375 -7.390625 "/>
</symbol>
<symbol overflow="visible" id="glyph0-4">
<path style="stroke:none;" d="M 9.84375 -4.484375 C 9.84375 -6.03125 9.734375 -6.953125 9.4375 -7.6875 C 8.78125 -9.34375 7.25 -10.34375 5.375 -10.34375 C 2.578125 -10.34375 0.765625 -8.1875 0.765625 -4.890625 C 0.765625 -1.59375 2.515625 0.4375 5.328125 0.4375 C 7.640625 0.4375 9.234375 -0.859375 9.625 -3.046875 L 8.015625 -3.046875 C 7.578125 -1.734375 6.671875 -1.03125 5.390625 -1.03125 C 4.375 -1.03125 3.515625 -1.5 2.96875 -2.34375 C 2.59375 -2.921875 2.453125 -3.484375 2.4375 -4.484375 Z M 2.46875 -5.796875 C 2.609375 -7.65625 3.734375 -8.859375 5.359375 -8.859375 C 6.921875 -8.859375 8.140625 -7.5625 8.140625 -5.90625 C 8.140625 -5.875 8.140625 -5.828125 8.109375 -5.796875 Z M 2.46875 -5.796875 "/>
</symbol>
<symbol overflow="visible" id="glyph0-5">
<path style="stroke:none;" d="M 7.515625 0 L 12.375 -13.984375 L 10.46875 -13.984375 L 6.59375 -2.15625 L 2.5 -13.984375 L 0.578125 -13.984375 L 5.609375 0 Z M 7.515625 0 "/>
</symbol>
<symbol overflow="visible" id="glyph0-6">
<path style="stroke:none;" d="M 0.4375 4.0625 L 4.015625 4.0625 L 4.015625 -13.984375 L 0.4375 -13.984375 L 0.4375 -12.609375 L 2.421875 -12.609375 L 2.421875 2.6875 L 0.4375 2.6875 Z M 0.4375 4.0625 "/>
</symbol>
<symbol overflow="visible" id="glyph0-7">
<path style="stroke:none;" d="M 1.34375 -10.046875 L 1.34375 0 L 2.953125 0 L 2.953125 -6.3125 C 2.953125 -7.765625 4.015625 -8.9375 5.3125 -8.9375 C 6.5 -8.9375 7.171875 -8.21875 7.171875 -6.921875 L 7.171875 0 L 8.78125 0 L 8.78125 -6.3125 C 8.78125 -7.765625 9.84375 -8.9375 11.140625 -8.9375 C 12.3125 -8.9375 13.015625 -8.1875 13.015625 -6.921875 L 13.015625 0 L 14.625 0 L 14.625 -7.546875 C 14.625 -9.34375 13.578125 -10.34375 11.703125 -10.34375 C 10.359375 -10.34375 9.5625 -9.9375 8.609375 -8.8125 C 8.015625 -9.875 7.21875 -10.34375 5.90625 -10.34375 C 4.5625 -10.34375 3.6875 -9.84375 2.828125 -8.640625 L 2.828125 -10.046875 Z M 1.34375 -10.046875 "/>
</symbol>
```

Skeleton Plot



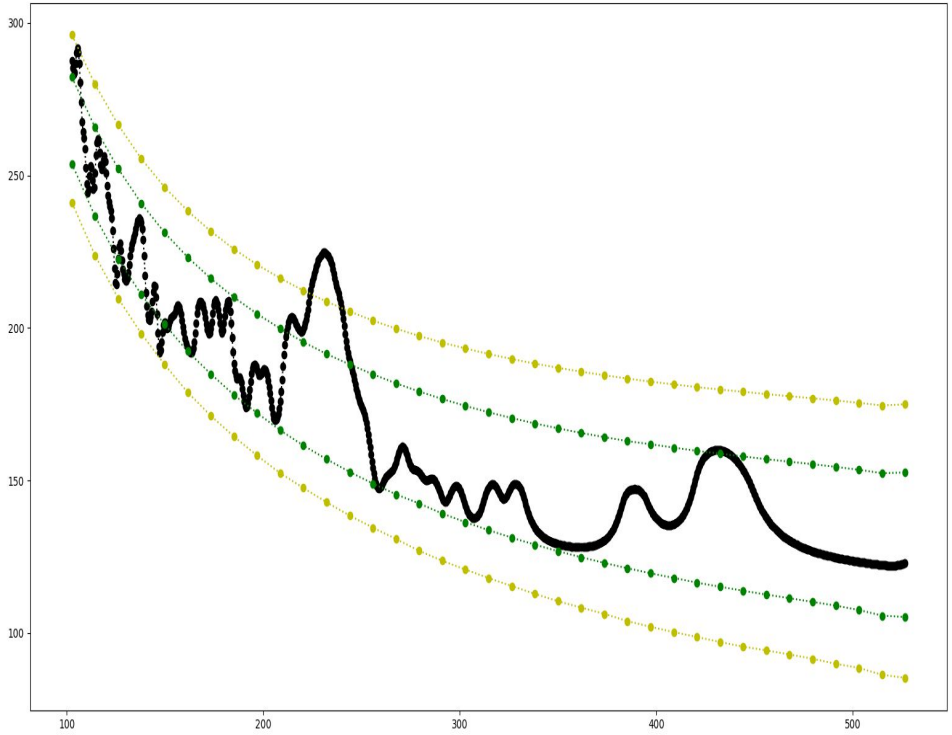


Algorithm

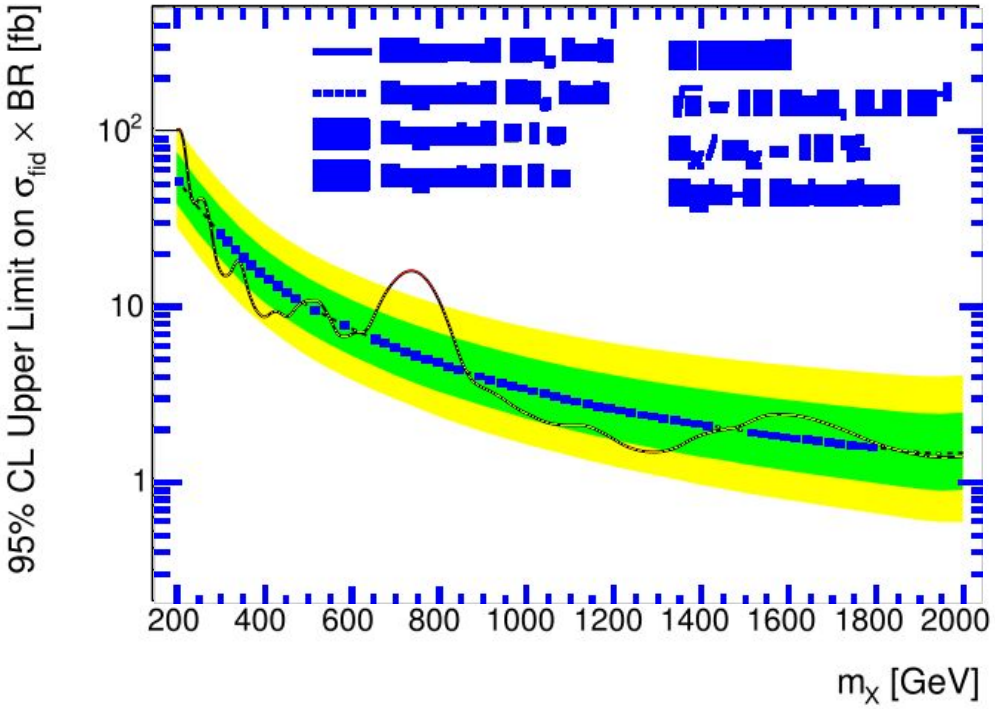
For point in blackline:

If green and yellow line includes x coordinate of the point in the black line:

Compare this point's y coordinate with yellow and green lines' matched points to decide the point is in which sigma band.

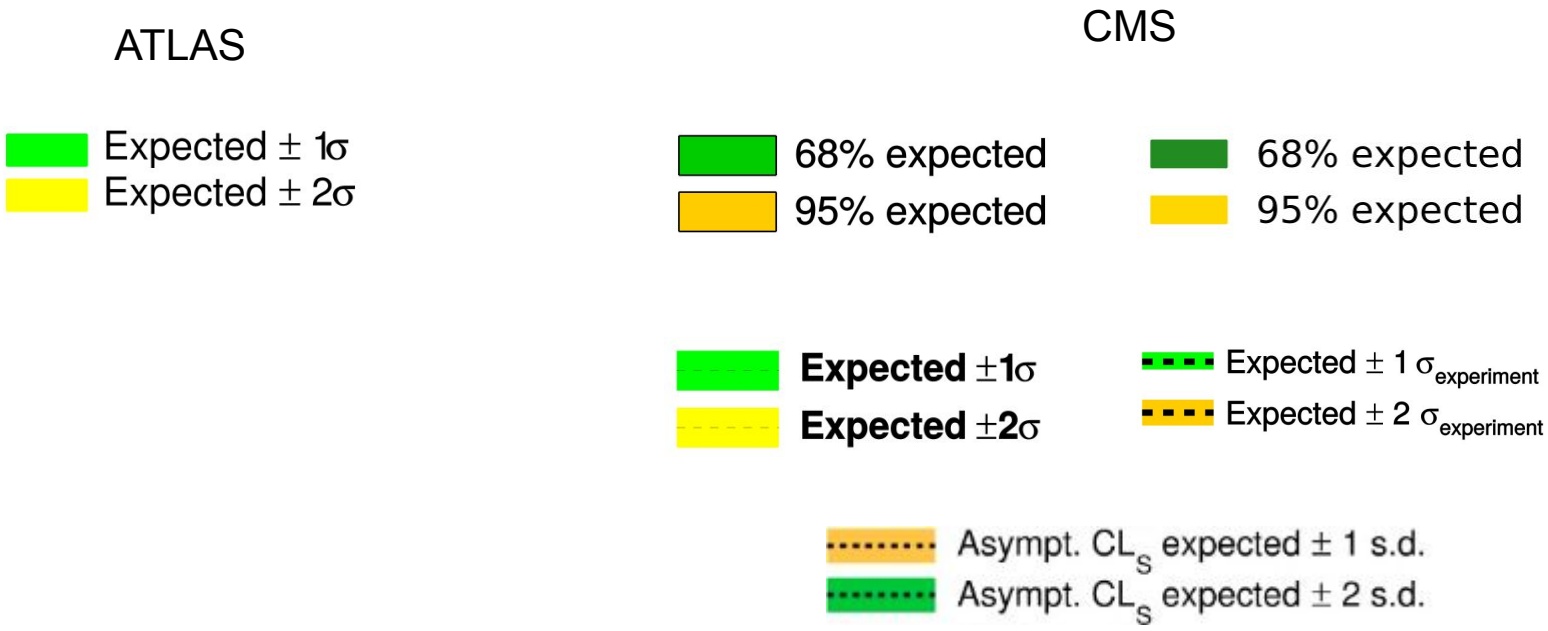


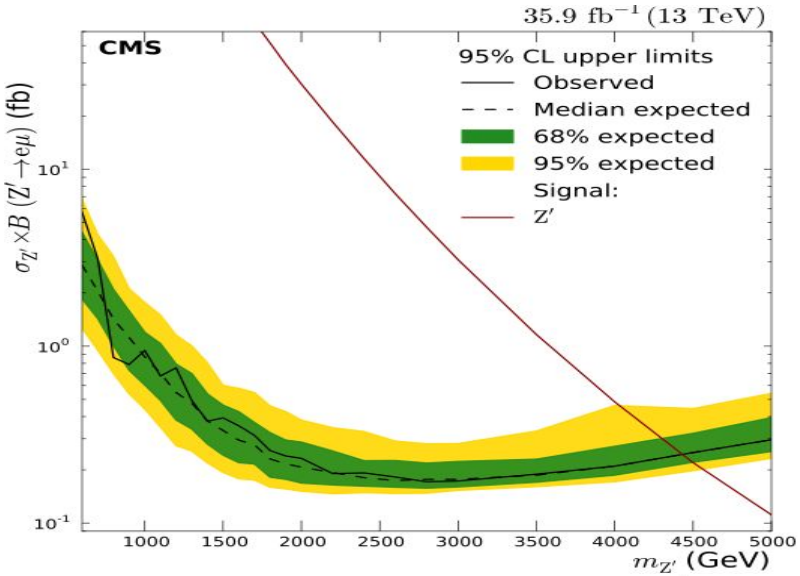
Analysis with Bitmap Images

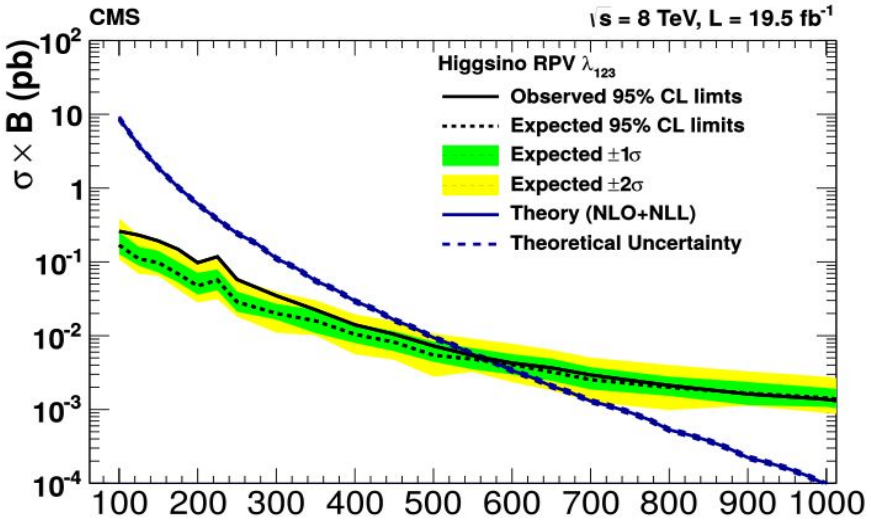


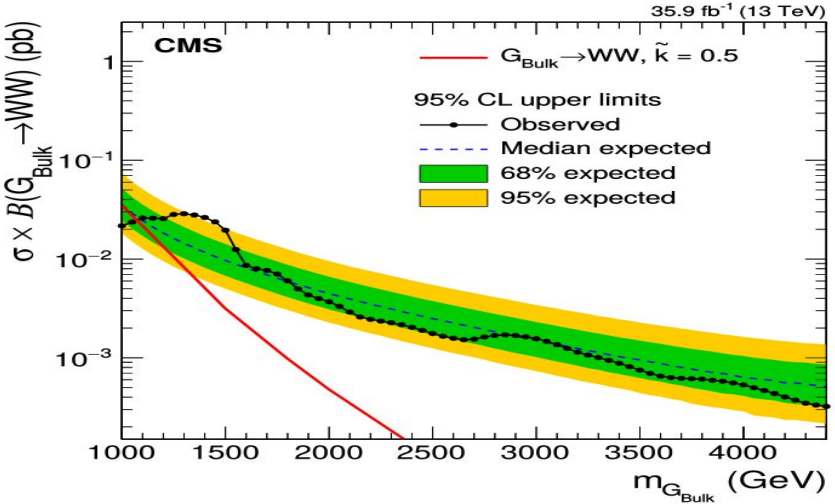
1. Determine the frame of the image & erase it.
2. A rectangle search algorithm handles the green/yellow legend boxes.
3. Ticks on the frame, letters and numbers are identified using OpenCV and blue (picked as a neutral color) boxes are drawn over them.
4. Loop along the x-axis. For each column of pixels:
   a. Scan from top to bottom to identify 1D clusters of black points. Discard cases with multiple or zero black clusters.
   b. Similarly identify the beginning & end points of yellow, green & white areas.
   c. Determine the position of the black cluster with respect to the identified color bands.

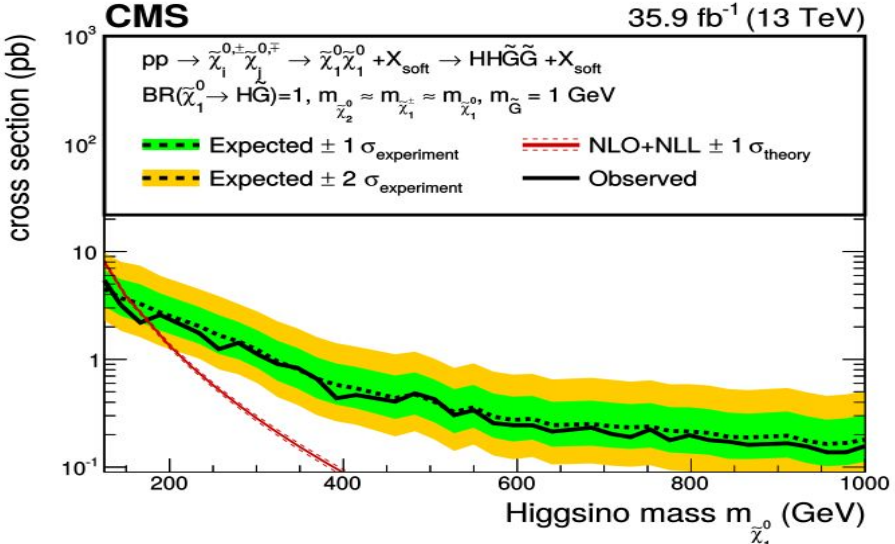
Non-Standardized Color Codes & Marker Styles

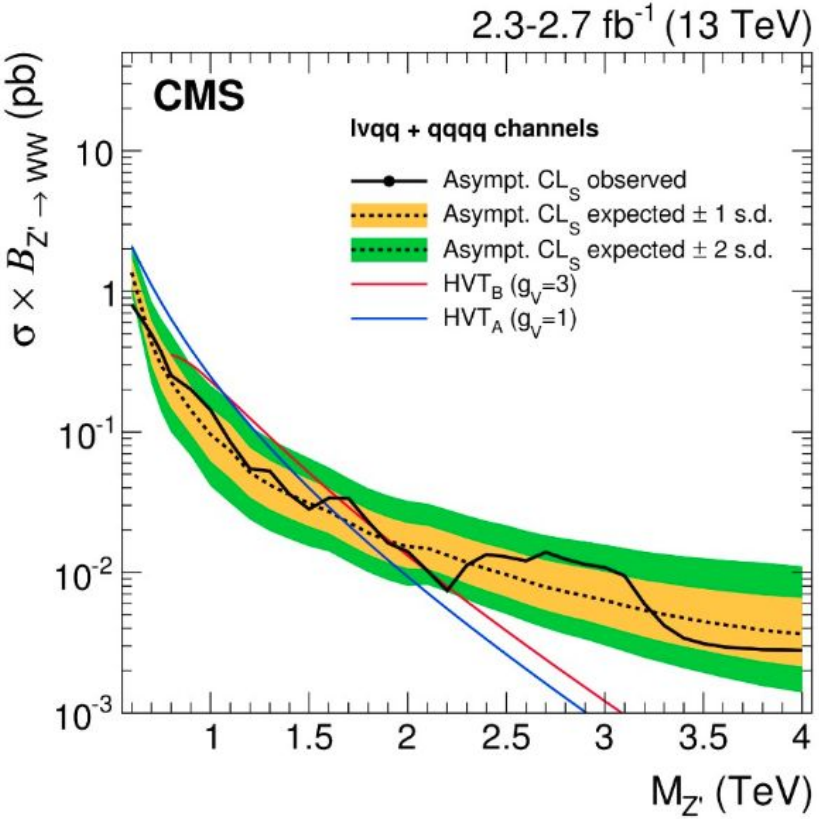



35.9 fb^{-1} (13 TeV)
CMS
95% CL upper limits
Observed
Median expected
68% expected
95% expected
Signal:
Z′
$\sigma_{Z'} \times B\,(Z' \to e\mu)$ (fb)
$m_{Z'}$ (GeV)


CMS
$\sqrt{s}$ = 8 TeV, L = 19.5 fb^{-1}
Higgsino RPV λ_{123}
Observed 95% CL limts
Expected 95% CL limits
Expected ±1σ
Expected ±2σ
Theory (NLO+NLL)
Theoretical Uncertainty
σ × B (pb)


35.9 fb^{-1} (13 TeV)
CMS
$G_{Bulk} \to WW$, $\tilde{k}$ = 0.5
95% CL upper limits
Observed
Median expected
68% expected
95% expected
$\sigma \times B(G_{Bulk} \to WW)$ (pb)
$m_{G_{Bulk}}$ (GeV)


CMS
35.9 fb^{-1} (13 TeV)
$pp \to \tilde{\chi}_i^{0,\pm} \tilde{\chi}_j^{0,\mp} \to \tilde{\chi}_1^0 \tilde{\chi}_1^0 + X_{soft} \to HH\tilde{G}\tilde{G} + X_{soft}$
$BR(\tilde{\chi}_1^0 \to H\tilde{G})$=1, $m_{\tilde{\chi}_2^0} \approx m_{\tilde{\chi}_1^\pm} \approx m_{\tilde{\chi}_1^0}$, $m_{\tilde{G}}$ = 1 GeV
Expected ± 1 $\sigma_{experiment}$
Expected ± 2 $\sigma_{experiment}$
NLO+NLL ± 1 σ_{theory}
Observed
cross section (pb)
Higgsino mass $m_{\tilde{\chi}_1^0}$ (GeV)



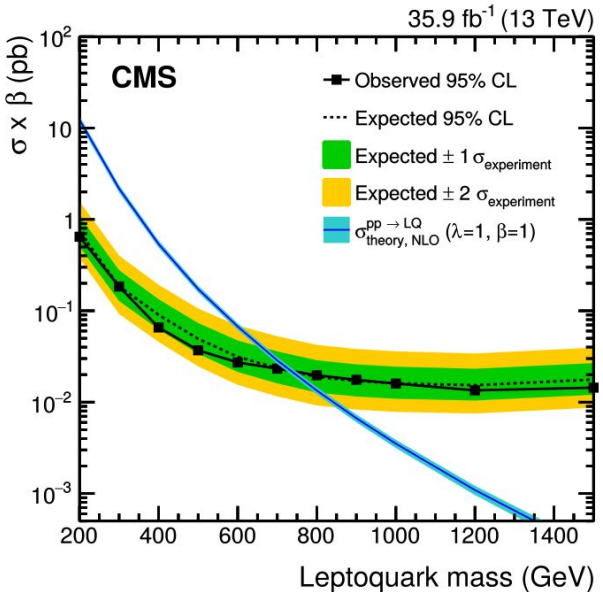
Inverse Colors for Sigma Bands

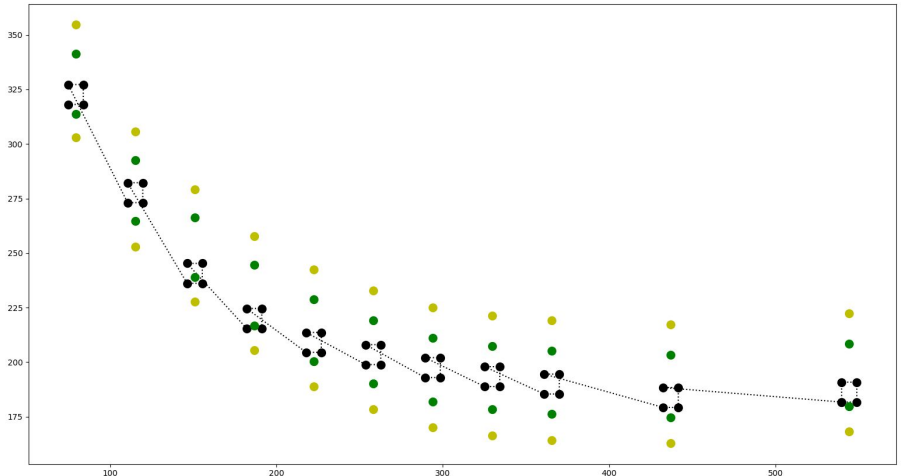
- Parsing problems because of inconsistent SVG attributes' keywords
- Marker used plots on the observed line
- Vector graphic formats are not available

Fraction of plots not considered fit for analysis:

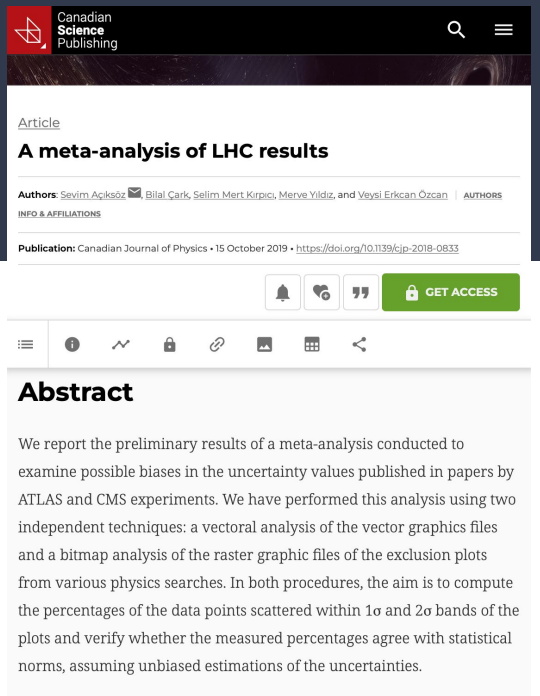
ATLAS : 43%

CMS : 43%





First Round Results



ATLAS - Vector

4157.33 points from 136 plots
68.98% ± 0.72
96.03% ± 0.30

CMS - Vector

2206.75 points from 98 plots
74.04% ± 0.93
95.23% ± 0.45

Two Points per Plot

331 points from 136 plots
69.98% ± 2.56
92.54% ± 1.40

Two Points per Plot

247 points from 98 plots
76.92% ± 2.68
93.93% ± 1.52

ATLAS - Bitmap

58449 pixel columns from 139 Plots
72.6%
94.1%

Canadian Journal of Physics • https://doi.org/10.1139/cjp-2018-0833

Adding Papers from 2019-2021

ATLAS
5425.45 points from 226 plots
70.73% ± 0.63
96.13% ± 0.27

CMS
3206.63 points from 253 plots
72.09% ± 0.79
94.38% ± 0.41

Two Points per Plot
266.76 points from 226 plots
70.12% ± 2.88
93.78% ± 1.52

Two Points per Plot
231.13 points from 253 plots
73.18% ± 2.91
94.31% ± 1.58

What's next?

- The overall results from both ATLAS and CMS experiments are quite encouraging.

- Analysis year by year
- Network mapping over cross referenced papers to eliminate possible correlations across papers.
  - Already performed for the past set of papers, to be updated.
- Correlations between consecutive points
  - Quantify (ex. Spearman coefficient)
  - Model their possible effects.

- We invite our colleagues to use the HEPData Repository so that analyses like this will be easily performed by future scientists.

ATLAS	A: 4 F: 23 P: 1401 81.51% 96.22% 3.78%	A: 9 F: 46 P: 1603 68.18% 97.57% 2.43%	A: 8 F: 69 P: 3608 66.69% 95.07% 4.93%	A: 21 F: 138 P: 6612 70.19% 95.92% 4.08%
CMS	A: 6 F: 35 P: 1318 72.23% 94.92% 5.08%	A: 14 F: 35 P: 675 71.11% 98.37% 1.63%	A: 10 F: 28 P: 933 85.21% 98.50% 1.50%	A: 30 F: 98 P: 2926 76.11% 96.86% 3.14%