# Blind Analyses
# -or-
# “The Answer's Not in the Back of the Book”

Scott Oser

Colloquium at UBC

November 25, 2010

# Blind Analyses
# -or-
# “The Answer's Not in the Back of the Book”

Scott Oser

Sermon

~~Colloquium~~ at UBC

November 25, 2010

# Clever Hans

Ask Hans the horse to add any two numbers, and he tapped his hoof the correct number of time!

He could also tell time, work a calendar, and spell words.

His German spelling was much better than mine.

# Was Hans a fraud?

Seemingly not. Hans answered questions correctly even when his trainer was not in the room!

Psychologist Oskar Pfungst made a very important discovery: if no one in the room knew the correct answer to the question being asked of Hans, Hans didn't know the answer either!

Hans was apparently picking up on subtle cues given by the questioners! This is the “Clever Hans” effect.

Pfungst studied and identified a number of subtle non-verbal cues that the participants would give off. He then tried to purposely suppress these clues, doing his best not to tip the answers to Hans. He failed---the cues were involuntary, and for most people completely unconscious.

Hans was clever, but not in the way that people thought!

# Medical applications


Medicine has long recognized the importance of "blind analyses". Given that placebo effects do happen, and that patients interact with their doctors, just as Hans interacted with his questioners, the gold standard in medicine is the "double blind" study:

*Neither the physician or the patient should know whether the patient is receiving a real treatment or a placebo.*

These days you'd probably have trouble getting a non-blind drug study published!

But surely experiments on inanimate objects shouldn't have such worries. Should they?

# Gregor Mendel: scientific fraud?

Gregor Mendel is the father of genetics, having discovered the laws of genetic inheritance.

But his published data is very curious: data fits his model with $\chi^2$/dof = 41.6/84, which has P<7x10$^{-5}$

Possible explanations:
1) Did Mendel publish only his best data, throwing out results that disagreed with his model?
2) After he formulated his theory, did he just continue to take data until the agreement was excellent, only then deciding to stop?

# Do the physical sciences have to worry about this?

Mendel's plants presumably didn't know they were part of an experiment. Neither do electrons. Remember: medical trials are double blind because *both* the experimenter and the subject can be unconsciously influenced.

There's no reason to think that physicists or astronomers are any less human than anyone else. It's not hard at all to think of ways that even with the best of intentions you can inadvertently produce a biased result.

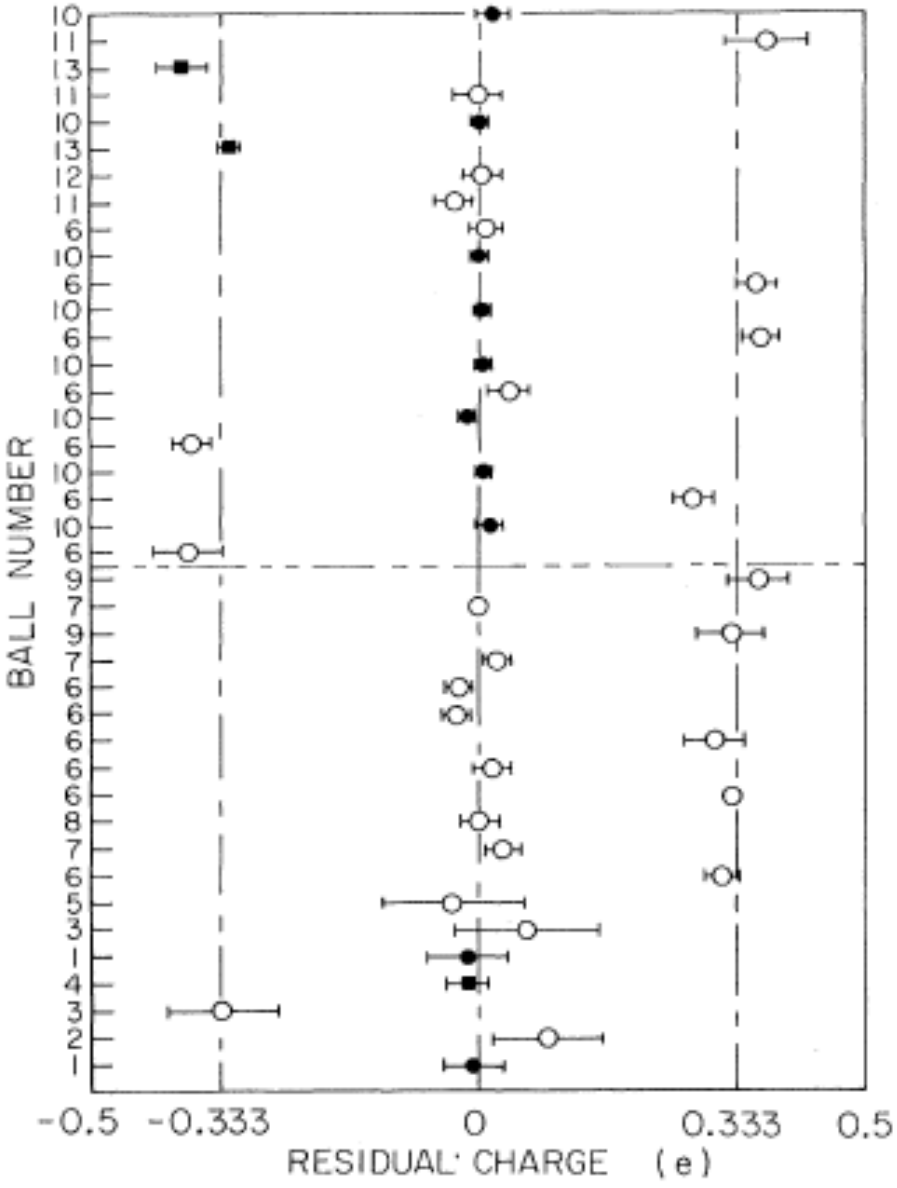
Are there any examples of this happening in the literature of the physical sciences?

# Fairbank's fractional charge measurement

In the early 1980's Fairbank *et al* did an experiment to look for fractional charges, with charges of ±1/3*e*. While quarks with fractional charges exist, they are supposed to be always confined inside bound states with integer charge.

To test this, superconducting niobium spheres were levitated in a magnetic field in a modern version of the Millikan oil-drop experiment, resulting in the data seen here:

Phys. Rev. Lett. 46, 967 - 970 (1981)



# Fairbank's fractional charge measurement

After this publication, Luis Alvarez suggested that the experimenters should redo the experiment, this time adding unknown random numbers to the charges, so that they did not know the true value of the charge until the analysis was complete. This was intended to prevent selection bias.

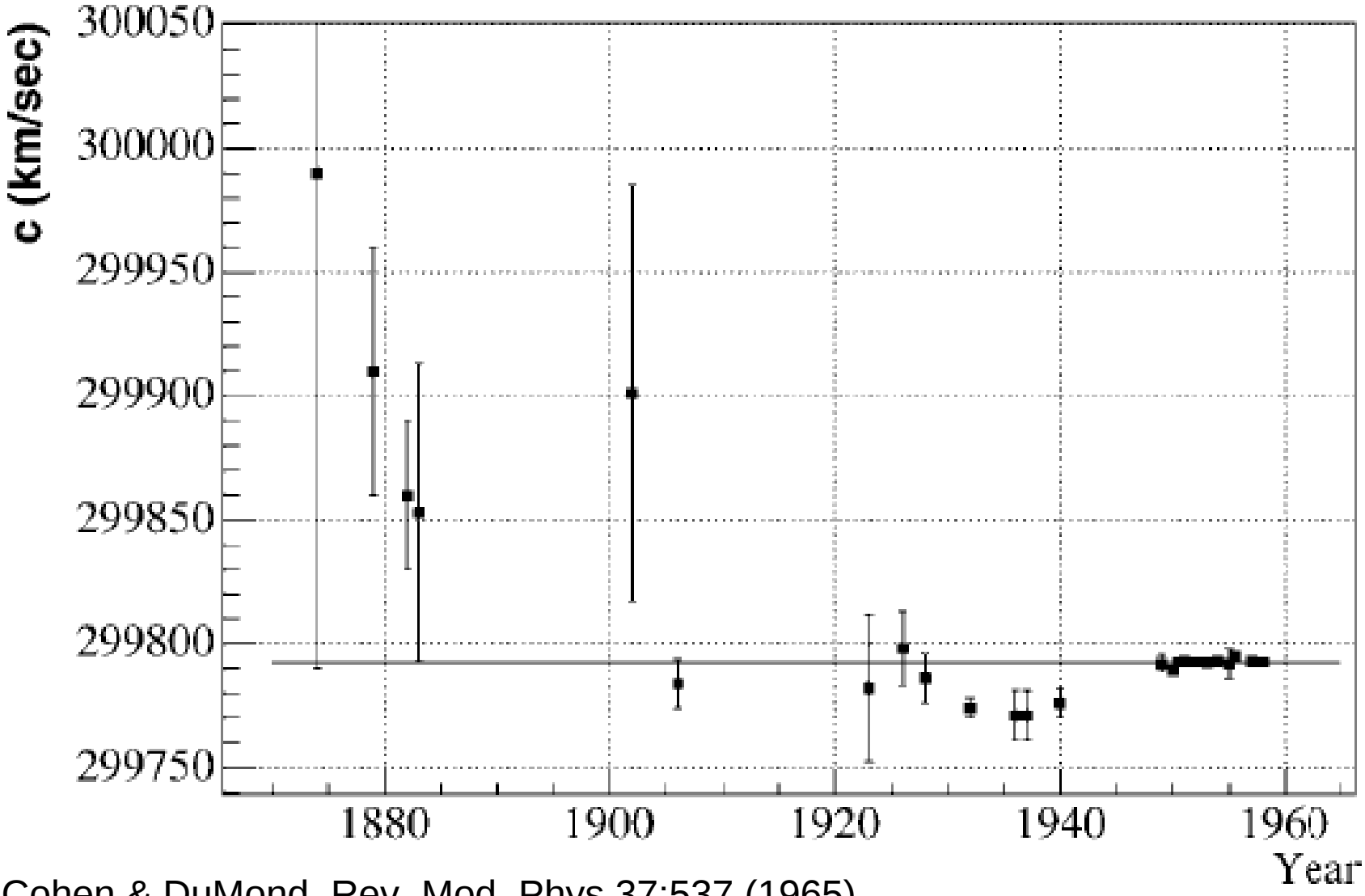
This procedure was tried on a new set of data. After the analysis was finished, the values of the random numbers were revealed and subtracted from the measured charges.

Results after unblinding showed no quantization at ±1/3*e*:

*Results from a blind analysis disagreed with the previous publications! To date, there is no credible evidence for free fractional charges.*

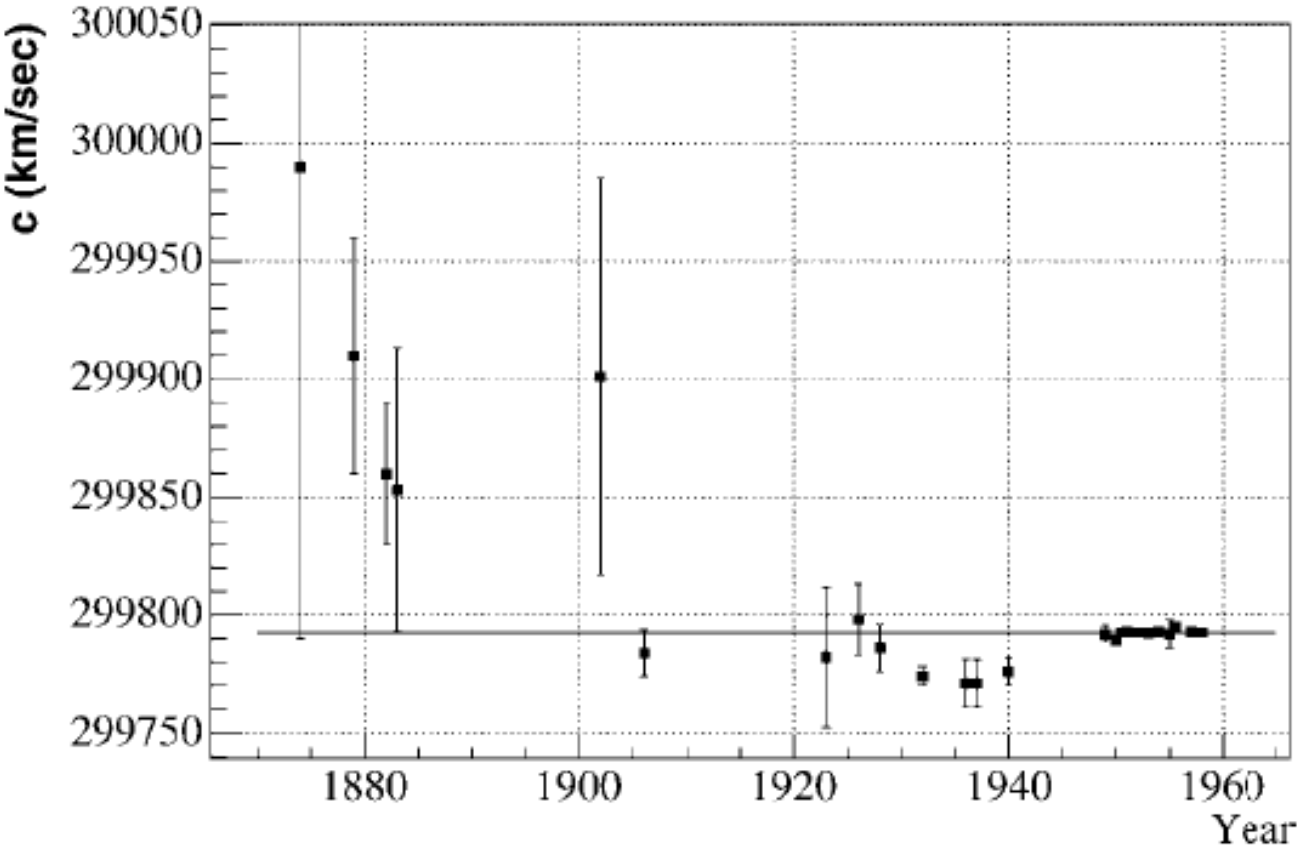
# Speed of light measurements

What if anything is wrong with this picture?




Cohen & DuMond, Rev. Mod. Phys 37:537 (1965)

# Speed of light measurements




Did you notice that:

1) Data points tend to agree with the previous data point better than they agree with the true value?

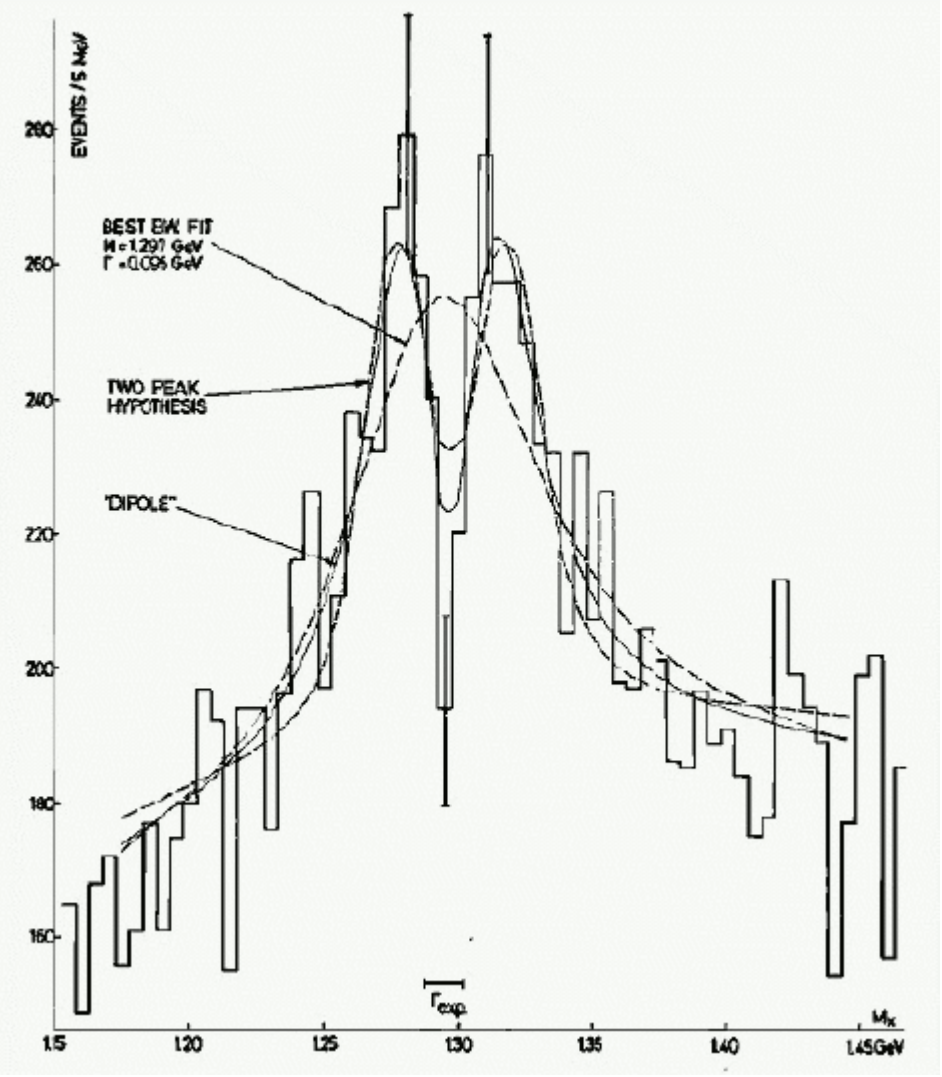
2) The existence of several data points (1930-1940) that are several standard deviations off the true value?

# The split $A_2$ peak

The $A_2$ is a meson produced in

$\pi^- + p \rightarrow p + M^+ M^-$

It was reported that the mass peak was split in two, with very high significance. Lots of theorists got very interested in explaining why.

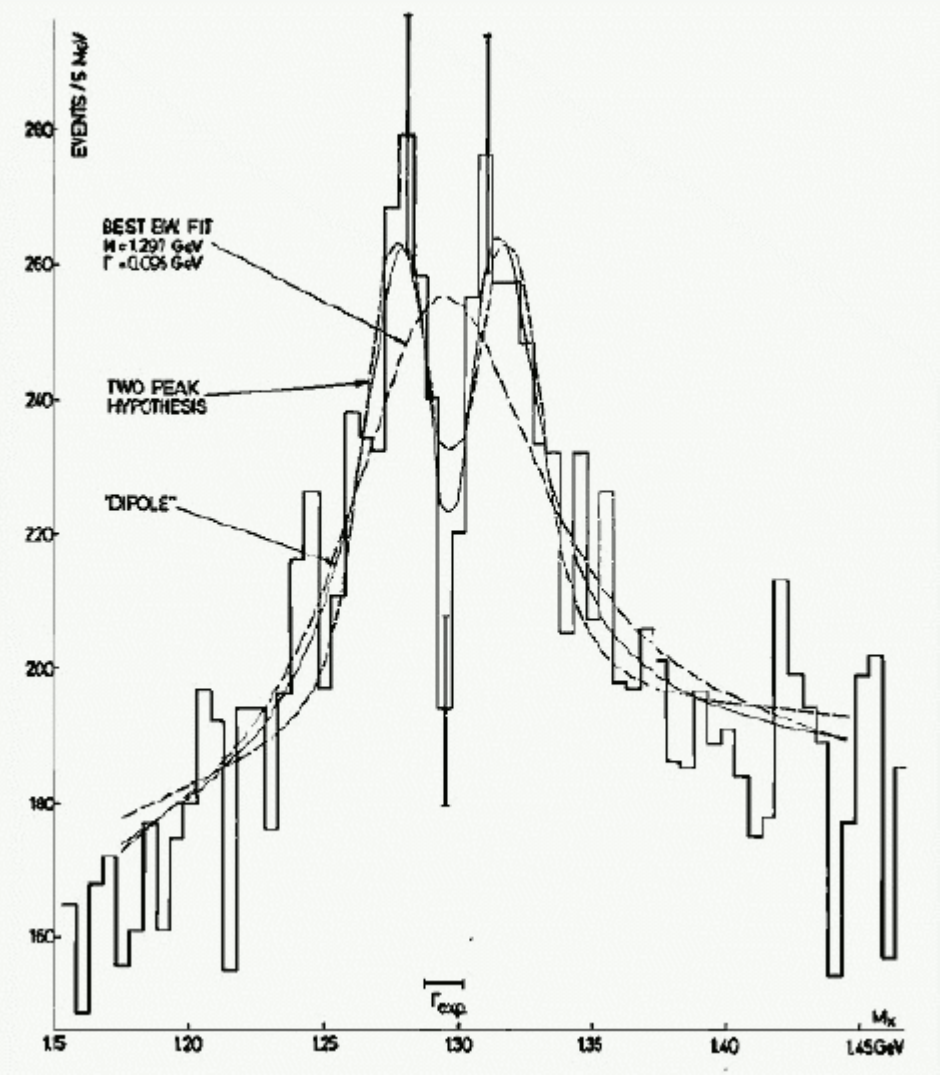



Kienzle et al., EMS 1968 proceedings

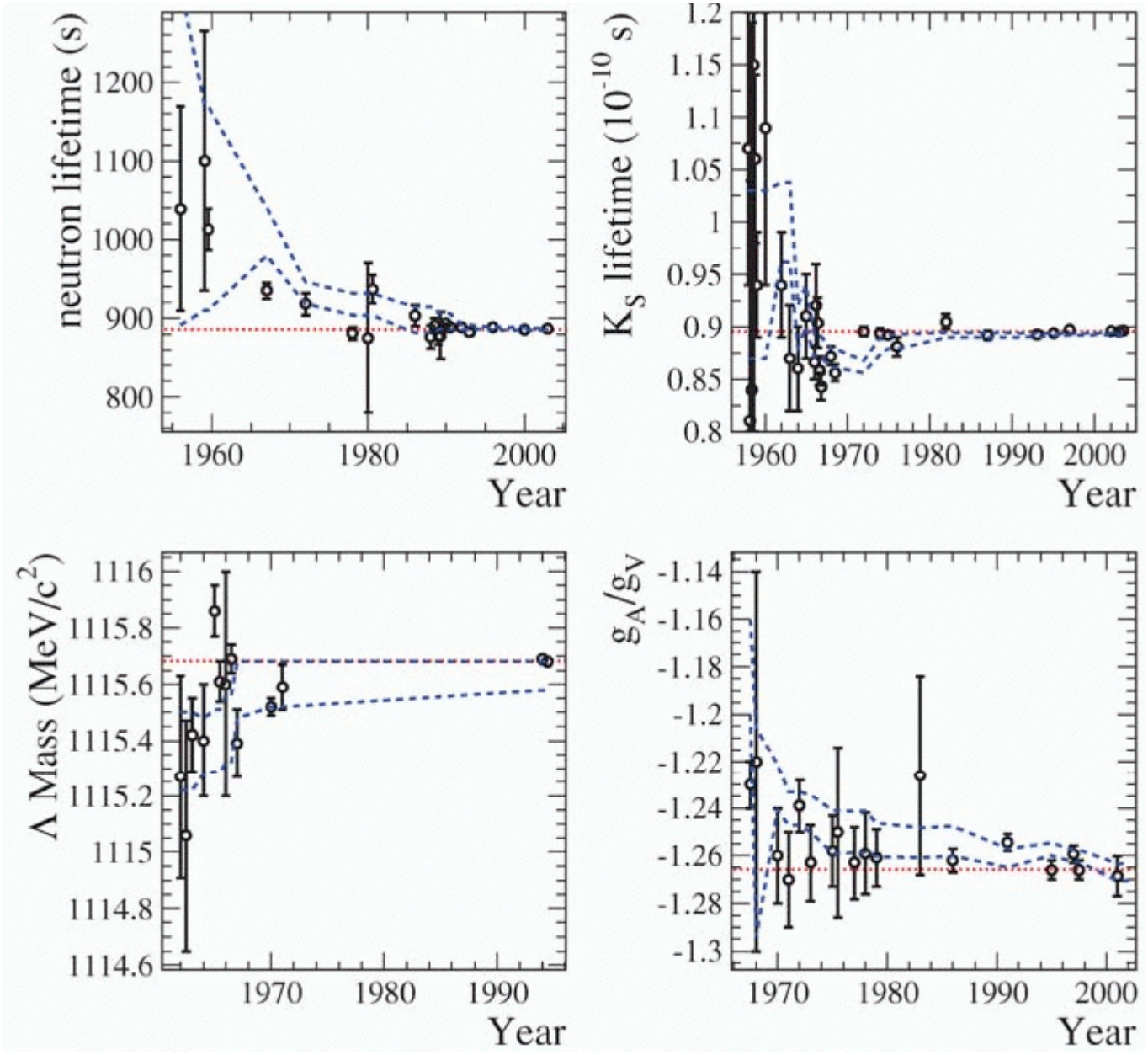
# The split $A_2$ peak

It turns out that one of the data cuts applied was to throw out whole runs in which the split was not seen.

Given the complexity of the detector, it was always possible to find some reason to throw out any run, and practically speaking the presence or absence of the split peak wound up being used as a selection criterion!



# PDG history plots: “follow the leader”




History plots for measurements of four fundamental particle properties

Hypothesis that each measurement is scattered with normal errors around the prior averages:
$\chi^2 = 131.2/83$ d.o.f.

Hypothesis that each is scattered around the eventual world average:
$\chi^2 = 249.7/82$ d.o.f.

# What is a blind analysis?

Blind analysis is a technique for avoiding biases in data analysis.

In a blind analysis, you analyze the data in such a way that you don't know what the final answer is going to be until the very last step, so you can't “tune” your result to get any particular answer.

In a blind analysis, you commit ahead of time to publish the result you get when you remove the blindness.

Blind analysis doesn't mean:

- you never look at the data
- you can't correct a mistake if you find one
- the analysis is necessarily correct---it's merely blind!
- conversely, a non-blind analysis doesn't necessarily give the wrong answer, but it does leave open the risk of bias.

# Dunnington's blind e/m analysis

In 1932 Frank Dunnington published one of the earliest blind analyses in physics. He was trying to measure the *e/m* ratio for the electron---there were several previous and widely discrepant values for this quantity.

He asked his machinist to build part of his apparatus at an unknown angle close to 340$^{o}$ and not to tell him the true value of the angle. This angle was one of the quantities in the formula for the *e/m* value.

Dunnington completed his analysis, got his "final" answer, and only then went back and measured the true value of the angle, putting it in place of the nominal 340$^{o}$. He then published the result.

*"It is also desirable to emphasize the importance of the human equation in accurate measurements such as these. It is easier than is generally realized to unconsciously work toward a certain value. One cannot, of course, alter or change natural phenomena (for example, the location of the current minimum in the present experiment), but one can, for instance, seek for those corrections and refinements which shift the results in the desired direction."*



*Phys. Rev. 43:404 (1932)*

# Mechanisms that produce bias

What are some of the ways you can bias a result?

# Mechanisms that produce bias: cut tuning

What are some of the ways you can bias a result?

1) Cut selection: normally you apply some selection criteria (“cuts”) to discard uninteresting data or events, in order to enhance your sensitivity to the signal. If you can directly observe the effect of these cuts on your final answer, you may be inclined to choose cut values that affect the answer in a subtle way.

This can be very easy to do, and it isn't always obvious that you've done it.

# Igor's Awesome Analysis

Igor is a graduate student in a big particle physics experiment. He wants to graduate this century.

His data set consists of ~2500 events. For each event, the detector measured the energy and the values of 10 quantities $X_1$ ... $X_{10}$ associated with the event. Each of these quantities can go from 0 to 1, and is distributed uniformly for normal events.

A theorist tells him that there ought to be a new particle between 4-6 GeV. Igor studies a bunch of Monte Carlo events for this model, and concludes that signal events are slightly more likely to have small values of $X_i$ than background. The Monte Carlo tells him that the optimal cut is $X_i$<~0.9. This in principle will remove 10% of the background but only a little signal. But his MC isn't good enough to tell him whether 0.88 or 0.92 is a better cut than 0.90.

# Igor's Automated Search

Since Monte Carlo can't exactly pin down what cut values Igor ought to use for each variable (0.88?  0.91?), he decides to optimize them on the data.

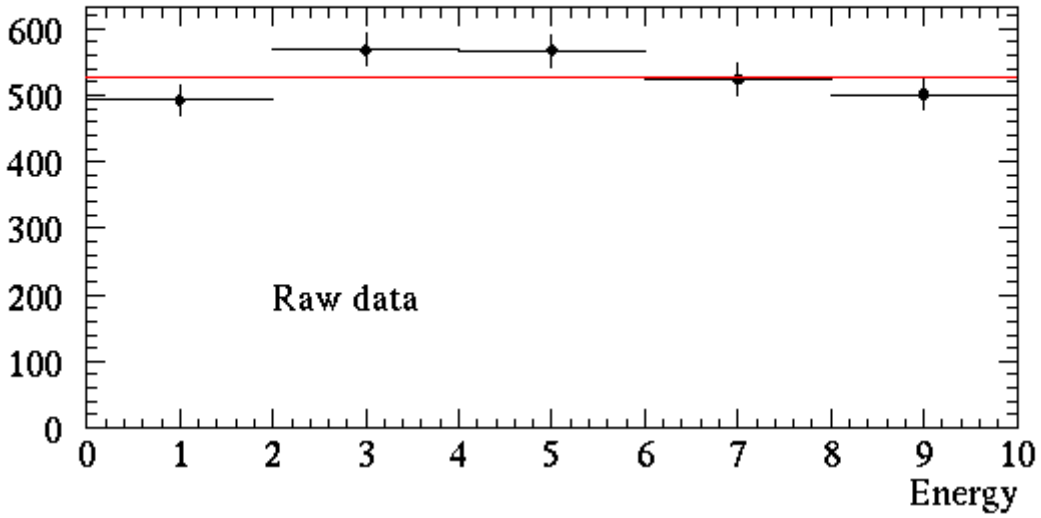
He applies each cut one-by-one, searching over all cut values between 0.88-0.92.  For each cut, he chooses a cut boundary that gives the largest signal significance between 4-6 GeV.  After all, he doesn't want to miss an important discovery.

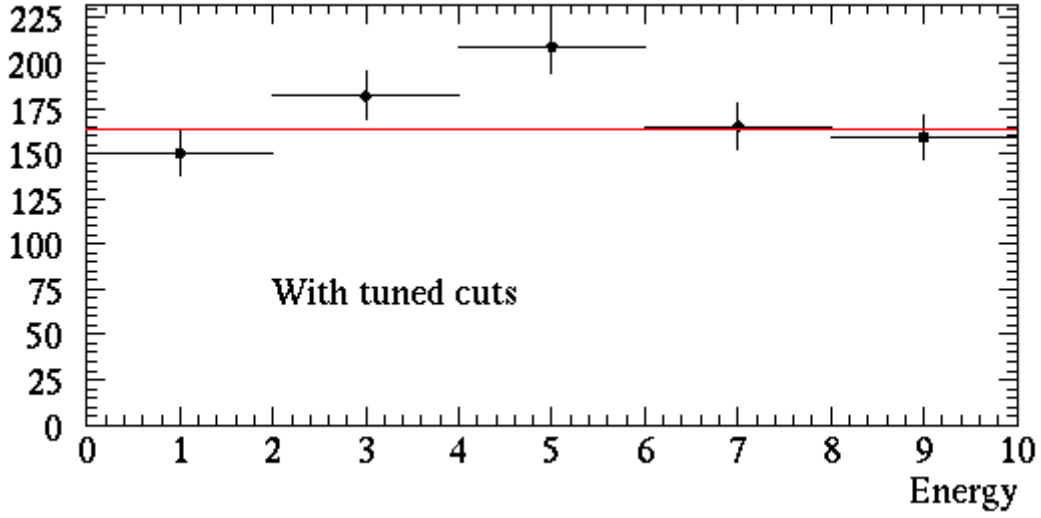
# Igor's Before And After Energy Spectra

It worked! In the raw data, there was no clear signal. But just as the Monte Carlo said, applying cuts on the $X_i$ enhanced the signal, eventually reaching >3σ with all cuts applied.

All cut values lie with 0.88-0.92, which is the range that Monte Carlo said was optimal. The cuts are scarcely tuned at all!

Igor says:
"Master will be pleased ..."



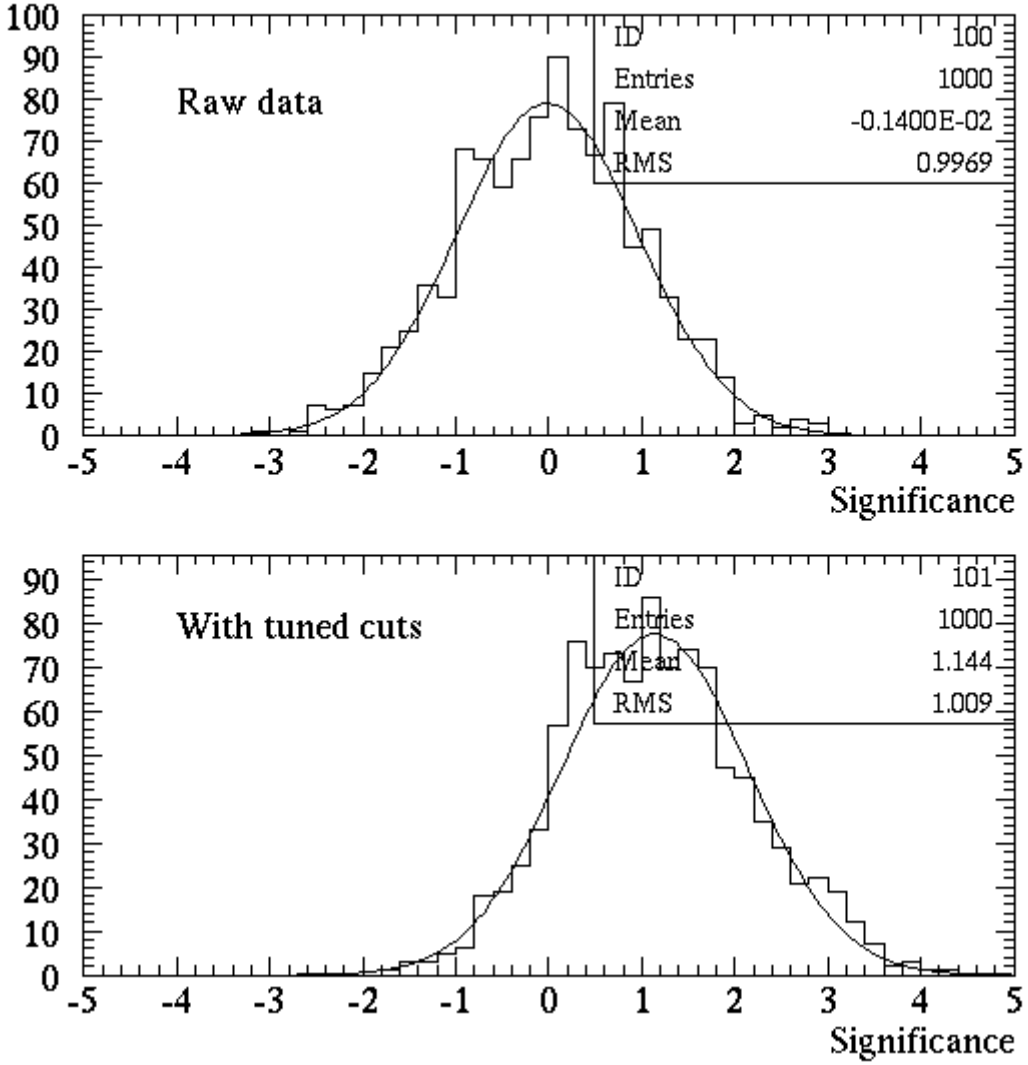
# Master was not pleased.

Master was not pleased. He used Monte Carlo to simulate the effect of Igor's procedure on 1000 fake data sets containing only background.

As expected, the raw data had an average signal significance of 0.

Igor's cut tuning on average produced a 1.1σ significance, and increased the chance of a 3σ result from 0.2% to 4.5%.

Tiny amounts of tuning can produce big effects!

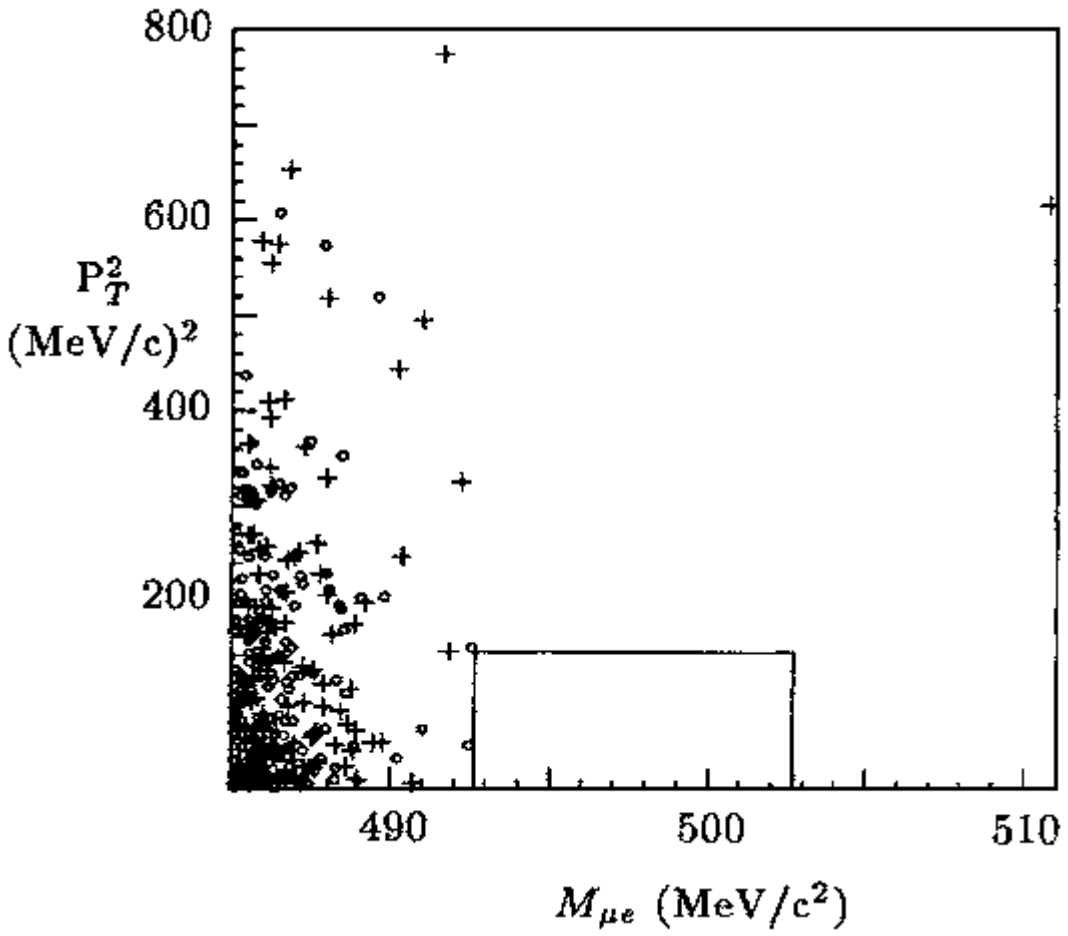


# Mechanisms that produce bias: rare events

Here's data from an experiment to look for $K \rightarrow \mu e$ decays. These decays are in principle forbidden by lepton flavour conservation.

If they do happen, they would produce an invariant mass around 498 MeV and a low value of $P_T^2$

We expect zero events. Given the data, where do you draw the signal box?

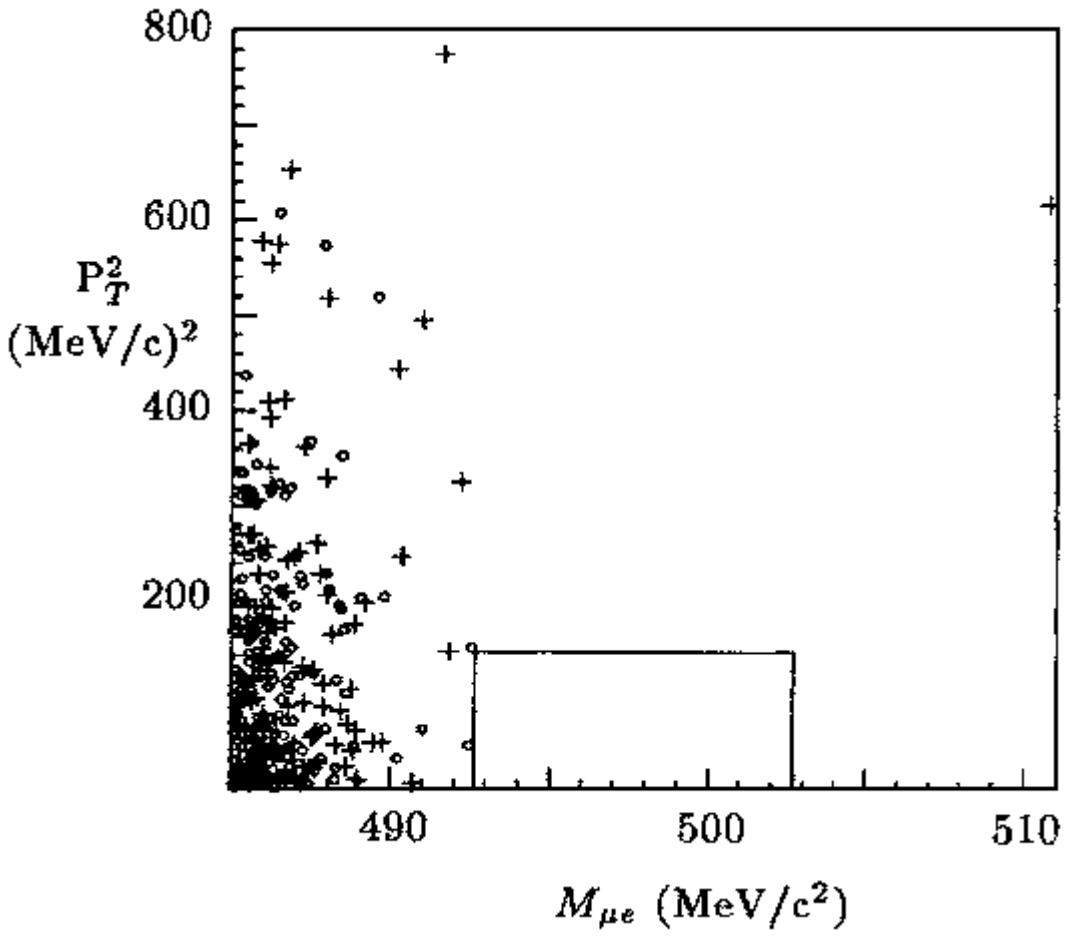



PRL 70:1049 (1993)

# Mechanisms that produce bias: rare events

The people who did this experiment were aware of two potential sources of bias:

1) drawing the box after having seen the data to avoid events
2) you find more events than expected in the box, and examine them one-by-one---maybe you decide that one track you thought was an electron looks atypical for an electron, so you decide to toss it.




They chose to do a “hidden box” analysis. They defined a box near the signal region that was “off limits”. No one could look at the events in that region.

# Mechanisms that produce bias: stop signs

*"If you don't like the weather, just wait a bit."*

When do you stop taking data? If you use the measured values to decide when to stop, you risk stopping as soon as the data fluctuates in the direction you expect it to. This will bias the results.

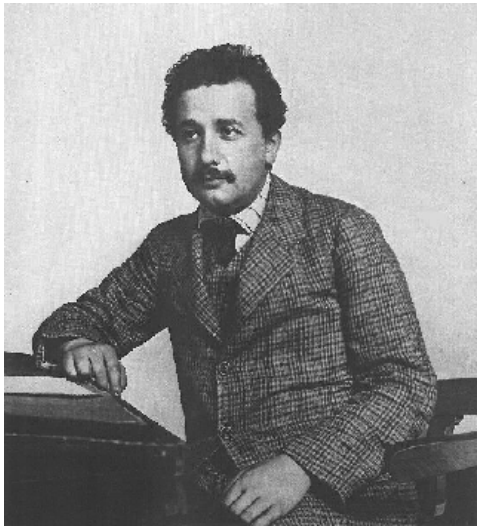

This applies to data analysis as well. Consider the following:

- you complete your measurement, and get a very strange result
- you spend a couple of days checking your analysis, and you find a code bug, which you fix
- after fixing the bug, your measurement agrees with your prediction

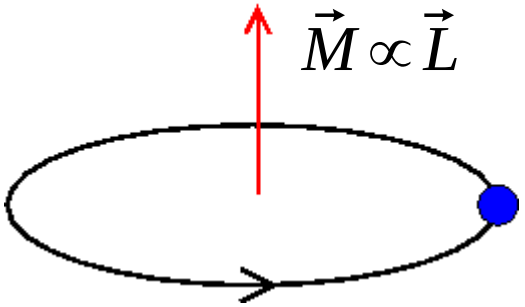
If your initial result had agreed with expectations, would you have ever found that code bug?

Knowing the final answer can never tell you whether your result is correct or not!

# Case Study: The Einstein-de Haas effect



Aspiring experimentalist




W.J. de Haas: luckiest lab assistant ever

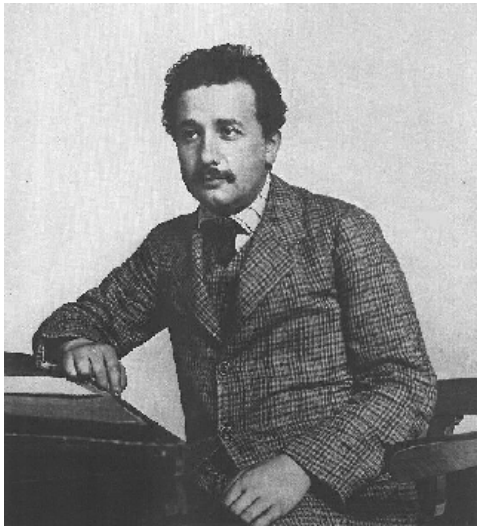
Is ferromagnetism caused by current loops from orbiting electrons?

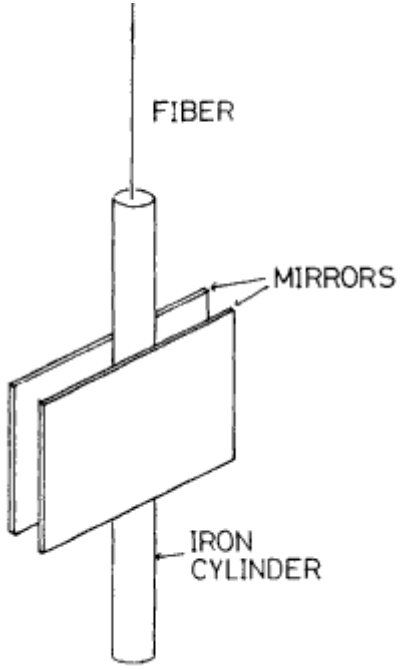
$$M = I(\pi r^2) = \left| \frac{e}{(2\pi r / v)} \right| (\pi r^2) = \frac{1}{2} evr$$

$$L = mvr$$

$$\frac{M}{L} = \frac{e}{2m}$$

# Case Study: The Einstein-de Haas effect




Basic idea: hit an iron cylinder with a magnetic field to magnetize it, look for any twisting due to acquired angular momentum.

Their measurement (1915): $\frac{M}{L} = \left(\frac{e}{2m}\right)(1.02 \pm 0.10)$

# Case Study: The Einstein-de Haas effect

$$\frac{M}{L} = \left| \frac{e}{2m} \right| (1.02 \pm 0.10)$$

Agrees with theory! Einstein moves on to other things (eg. discovering GR)

Other researchers pursue these measurements as a way to more precisely measure *e/m*

A problem: as experimenters try to reduce systematic error after another, all the measurements start converging on 2.0!

This is the correct value. Explanation of why g=2 had to wait for Dirac's equation.

What happened? This was a difficult experiment at the time, with lots of difficulties with stray magnetic fields, alignment, and magnetic saturation effects.

Einstein and de Haas knew of these things, but thought they had them under control once their result agreed with (wrong) theory.

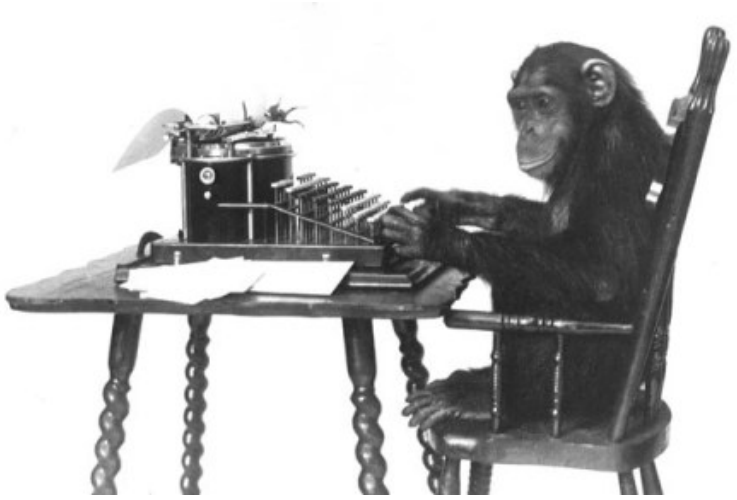
# What to do when you get a significant effect?

Suppose your colleague comes to you and says “I found this interesting 5σ effect in our data!” You check the data and see the same thing.  Should you call a press conference?

# What to do when you get a significant effect?

Suppose your colleague comes to you and says “I found this interesting 5σ effect in our data!” You check the data and see the same thing.  Should you call a press conference?

This depends not only on what your colleague has been up to, but also on how the data has been handled!



*A trillion monkeys typing on a trillion typewriters will, sooner or later, reproduce the works of William Shakespeare.*

*Don't be a monkey.*

# Trials factors

Did your colleague look at just one data distribution, or did she look at 1000?

Was she the only person analyzing the data, or have lots of people been mining the same data?

How many tunable parameters were twiddled (choice of which data sets to use, which cuts to apply, which data to throw out) before she got a significant result?

The underlying issue is called the “trials penalty”. If you keep looking for anomalies, sooner or later you're guaranteed to find them, even by accident.

Failure to account for trials penalties is one of the most common causes of bad but statistically significant results.

# Why trials factors are hard

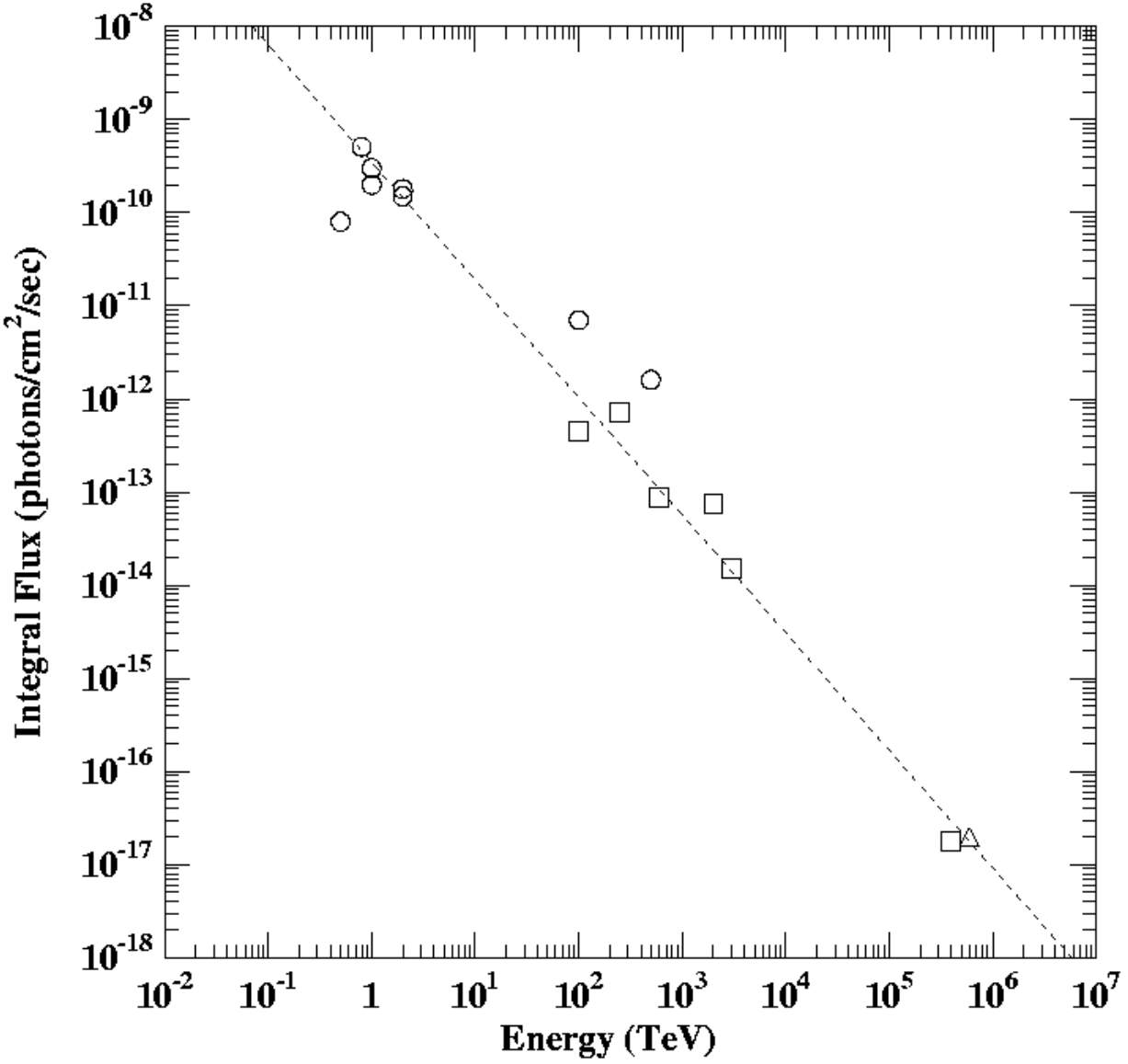
It can be really difficult to account for trials factors. For one thing, do you even know how many trials were involved?

Example: 200 medical researchers test 200 drugs. One researcher finds a statistically significant effect at the 99.9% C.L., and publishes. The other 199 find nothing, and publish nothing. You never hear of the existence of these other studies.

Chance of one drug giving a false signal: 0.1%.
Chance that at least one of 199 drugs will give a significant result at this level: 18%

Failing to publish null results is not only stupid (publish or perish, people!), but downright dangerous.

# An aside: gamma-ray astronomy history

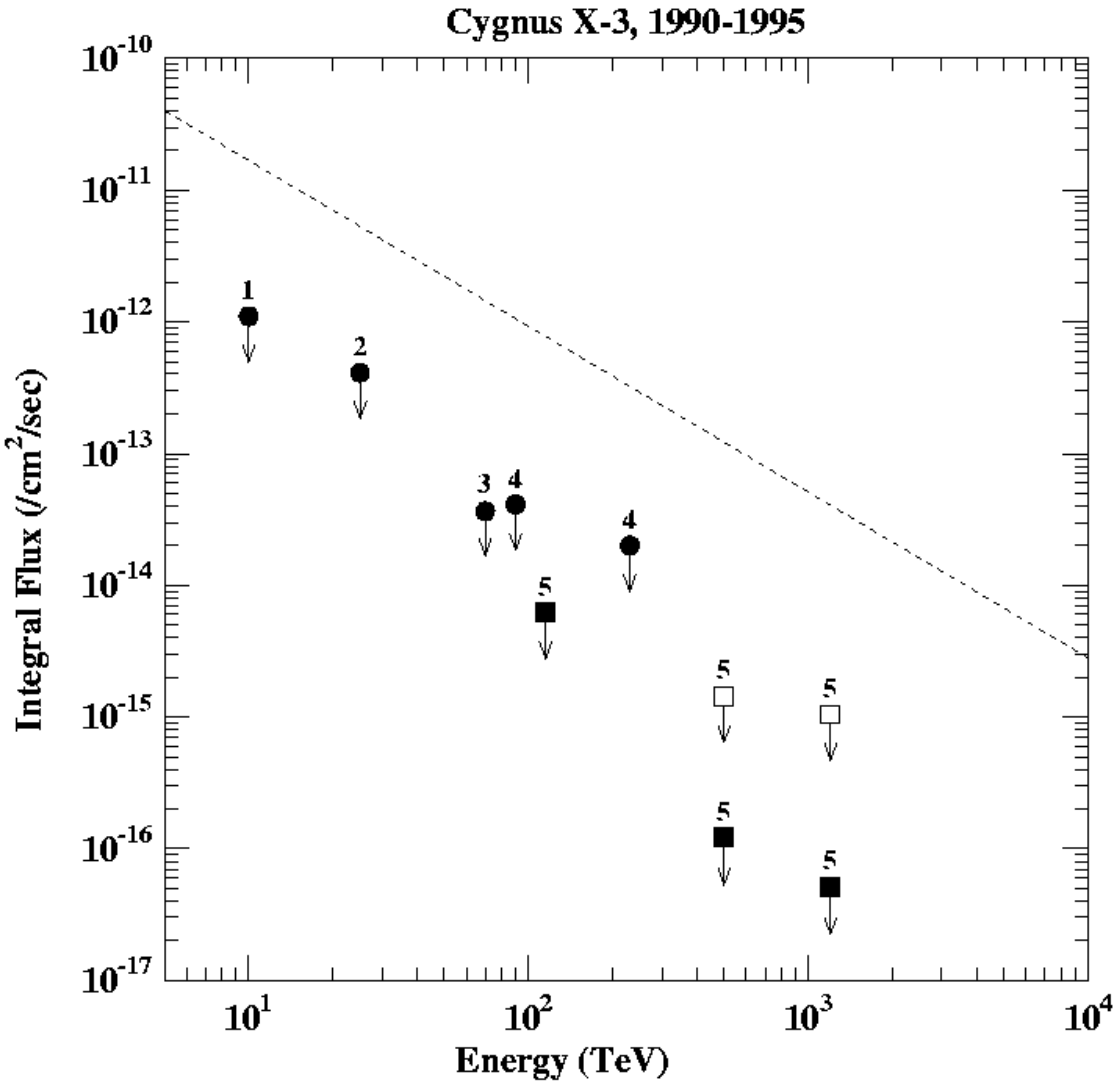



In the 1980's, many experiments operating at very different energy ranges detected high-energy gamma-rays from Cygnus X-3. Typical statistical significance was 3-4σ, and signals were hard to pull out---lot of data massaging.

But multiple independent measurements all claimed something, and the collective data was nicely fit by a consistent power law!

So much better detectors were built.

# Gamma-ray astronomy: the next generation




A. Borione et al, PRD 55:1714 (1997)

New detectors were orders of magnitude more sensitive. Using blind analyses they saw nothing!

It's possible, but highly conspiratorial, to imagine that Cygnus X-3 “turned off” just as the new experiments came on line.

A likelier interpretation of the earlier results is that they were a combination of statistical fluctuations and trial factors--- maybe people were so convinced that Cygnus was there that they kept manipulating their data until they “found something”.

Since sensitivity of experiments also follows a power law, this explains seemingly convincing energy spectrum.

# Moral

Science is littered with many examples of statistically significant, but wrong, results. Some advice:

- Be wary of data of marginal significance. Multiple measurements at 3σ are not worth a single measurement at 6σ.
- Think about possible biases in the analysis, especially if the analysis wasn't blind.
- Consider trials factors carefully, and quiz others about their own trials factors.
- Remember the following aphorism: “You get a 3σ result about half the time.”

# Hidden offset analyses

In many analyses, you might be able to hide the value of some parameter of the analysis. (Remember how Dunnington hid the machined angle in his apparatus for measuring the electron's *e/m* ratio?) Consider whether there is any element whose value you could hide.

Example: if you're doing the Cavendish experiment to measure G, have someone unconnected with the experiment measure the masses of the test masses for you and seal the result in an envelope. She then adds a random number to the measured mass and reports the "shifted" mass to you. You use that in your analysis. As a last step, you open the envelope, and replace the shifted mass with the correct mass, and calculate the final answer without changing anything else.

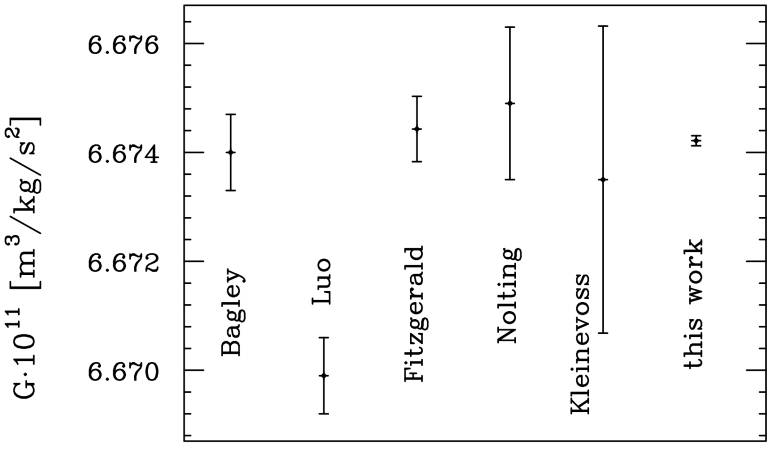



Figure from
PRL 85 (2000) 2869-2872

# Hidden fitting offsets

Do you fit your data?

One easy blindness scheme is to insert hidden offsets into your fitter. For example, one analysis of high-Z supernovae set up their fitter to fit for the cosmological parameters $\Omega_M$+X and $\Omega_\Lambda$+Y, instead of just fitting for $\Omega_M$ and $\Omega_\Lambda$[†].

Perhaps an even better approach would be to include a hidden ± sign in the fit, so you no longer can tell which direction the fit parameter changes if you vary something:

$$\Omega_{M,fitted} = \begin{pmatrix} +1 \\ -1 \end{pmatrix} \times \Omega_{M,true} + X$$

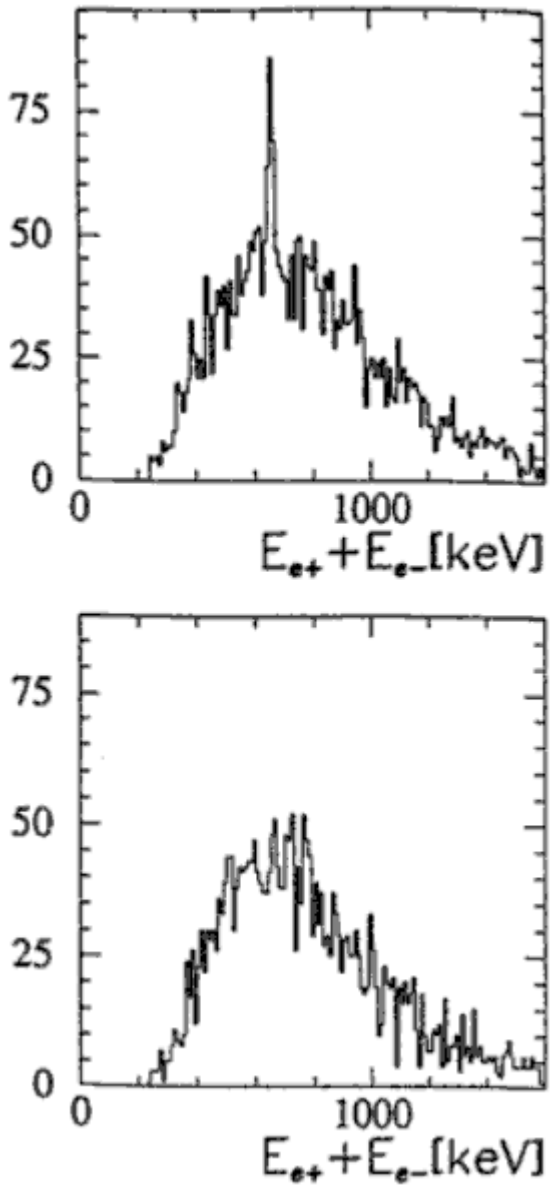
Here both the value of X and the choice of sign are hidden.

[†]Conley et al, Ap.J. 644 (2006) 1

# Data division

A very easy way to do a blind analysis is to use “data division”. Divide your data set into a small portion which is 100% unblind. You can do whatever you want to it. Once you're happy with the analysis on the unblind portion, you apply it to the rest of the data (the “blind portion”).

It's important to separately report the results from the unblind and blind data portions in your publication. Ideally you should base your final answer only on the blind data, since the unblind portion could be biased.




←This experiment thought it had discovered a new particle in heavy ion collisions ...

←Or they did until they tried data division on a new data set, and compared the open and blind results, getting this!

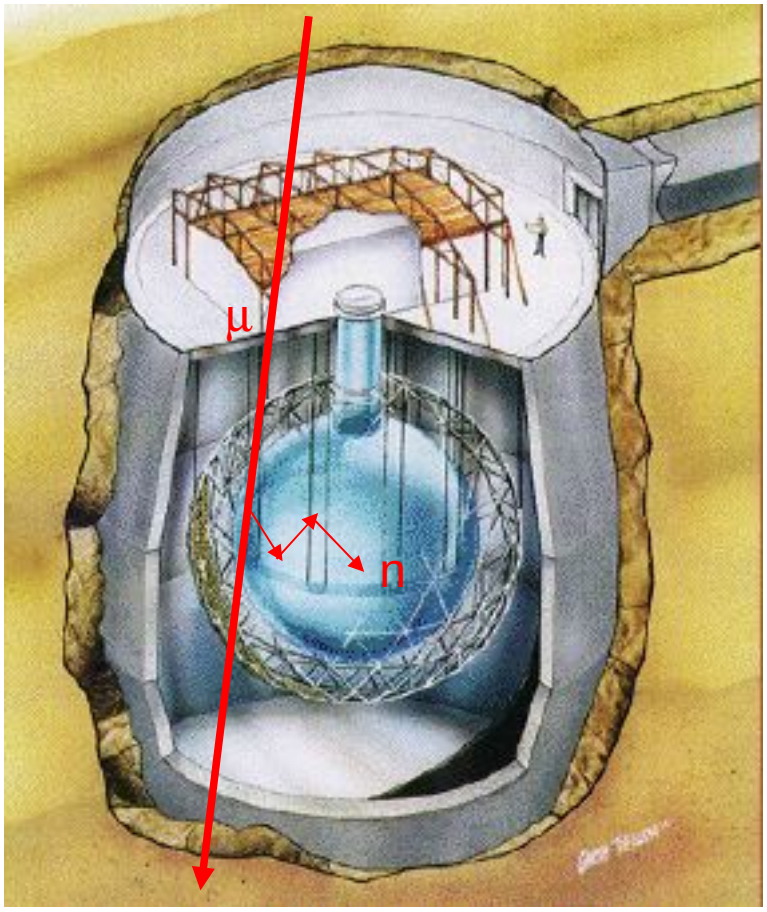
Ganz et al, Phys. Lett B 389:4-12 (1996)

# Adding fake events

Sometimes you can effectively add “fake” events to your data which will hide the answer. It's very important that the fake data be indistinguishable from your real data.

The SNO experiment counted neutrons knocked loose by solar neutrinos.

Muons also knock loose neutrons. Normally we cut these out. This gives an opportunity for blindness ...




Normal cut removes all events within X seconds after muon

t=0 t=X

Modified cut leaves an unknown fraction ? of these events in the data set

The blinded muon cut inserts an unknown number of extra neutrons into our data!

# Hiding events

In addition to adding events, you can hide events.

The simplest version is to add a cut that removes an unknown fraction of your signal events.

It's important that this be done in a way that makes it impossible to estimate the fraction by looking at the data. (For example, if your detector records calibration signals every 5 seconds, make sure these events aren't hidden, since their rate is known!)

SNO applied this on top of adding events to make the blindness two-directional (unknown number of background events added, and unknown number of real events hidden).

*Adding an unknown number of events, then removing a different unknown number, gave double blindness.*

# When blind analyses go wrong

If you do blind analyses, it's possible that something is going to go wrong:

1) You look in your signal box, see many more events than you expected, and realize that there is a background you forgot about.

2) You remove a hidden offset, and the final answer is physically absurd because you forgot to apply some correction.

3) A poorly constructed blind analysis could potentially bias the result in one direction. Be especially careful in cases where the analysts know the sign but not the magnitude of a shift in the data.

What do you do if you get egg on your face?

# Suck it up!

Blind analysis techniques do not prevent all mistakes.

Nor does it require you to publish a wrong result!

If you attempt a blind analysis and something goes wrong, you should:

- fully disclose your entire procedure, including what your “blind” answer was, and why you think it is in error. You will have more credibility admitting to a mistake and showing how you fixed it than hiding your dirty laundry and not doing a blind analysis at all.
- report your corrected answer, making clear that corrections were applied after blindness was removed, and so the final answer is not “blind”
- have a written procedure in place in advance describing what checks you will do after blindness is removed, and what actions you might take as a result of those checks

# Sociology of blind analyses

I have seen more fighting and acrimony about blind analyses than almost any topic in physics. Blind analyses, although gaining rapidly in popularity, are not yet universally applied or even supported. Some objections I've heard to blind analyses:

1) “Doing a blind analysis slows down the data analysis.”
Certainly this can happen, and partly by design. A careful blind analysis strategy can minimize delays, but a lot of the benefit of a blind analysis is that it purposefully forces analysts to slow down and check everything. If you're not supposed to “patch” mistakes after the fact, you spend more time making sure you get things right the first time, and will catch more mistakes. Here's a good strategy: list every possible check that you would do if you did an analysis and got an absurd or unexpected answer. Now carry out all of those same checks before you know the final answer.

# Sociology of blind analyses

2) “If the analysis is blind, how will I know if I got the right answer?”
This objection, once uttered aloud, often ends any debate about whether to do blind analyses!

3) “I'm going to make my full data set publicly available, so others can check my results. Therefore I don't need to do a blind analysis.”
I have heard this one from some CMBR people, for example. The objection doesn't hold---first of all, if you're being biased by a theoretical expectation, then the whole community can be subject to the same bias. Why can't everyone be biased? Furthermore, it sounds a lot like “I don't need to be very careful with my own analysis, because I can count on other people to find and correct my mistakes.” (!)

# Sociology of blind analyses

4) “Doing a blind analysis will limit my ability to perform sanity checks on the data, since I can't look at some of it.”
The extent to which this is true depends on how the blind analysis is implemented. The best blind analyses hide only the final answer and let you look at anything else you want in the data. Often a more careful design for blindness will help. Of course there may be a trade-off.

5) “Blindness isn't necessary, because the data is what it is. Psychological motives don't enter into it.”
History and common sense prove otherwise.

6) “By having multiple independent analyses, we get most of the benefits of blind analyses anyway.”
Independent analyses are good for finding errors. But what happens when the first analysis is wrong, and the second considers itself finished when it gets the same answer as the first?

# Conclusions

1. Scientists are human.  Everyone potentially has biases.

2. The history of physics and astronomy is littered with the wreckage of biased analyses: incorrect results, missed opportunities, and wasted effort.

3. “Blind analysis” is a general principle for avoiding bias in a result, and can be applied in almost any context.

4. We should be teaching our students, starting at the undergraduate level, about techniques for blind analysis.

5. My personal viewpoint: We should move towards a scientific culture in which blind analysis is the default, and non-blind analyses should have to justify why they should be considered for publication.

# Further reading

*Selectivity and Discord*, by Allan Franklin, University of Pittsburgh Press, 2002.

*How Experiments End*, by Peter Louis Galison, University of Chicago Press, 1987.

“Blind analysis”, P.F. Harrison, J. Phys G: Nucl. Part. Phys. 28  2679-2691, 2002.

“Benefits of Blind Analysis Techniques”, Joel G. Heinrich, University of Pennsylvania, CDF internal note CDF/MEMO/STATISTICS/PUBLIC/6576 http://www-cdf.fnal.gov/publications/cdf6576_blind.pdf

“Blind Analysis in Nuclear and Particle Physics”, Joshua R. Klein & Aaron Roodman, Ann. Rev. Nucl. and Part. Systems, Vol. 55, Issue 1, pp. 141-163, 2006.