Advanced Statistical Techniques in Particle Physics

Conference Summary

Bob Cousins, UCLA

22 March 2002

In this posted version, I have added a few slides (in this black font) containing a few of the things which I said verbally.

THANKS!

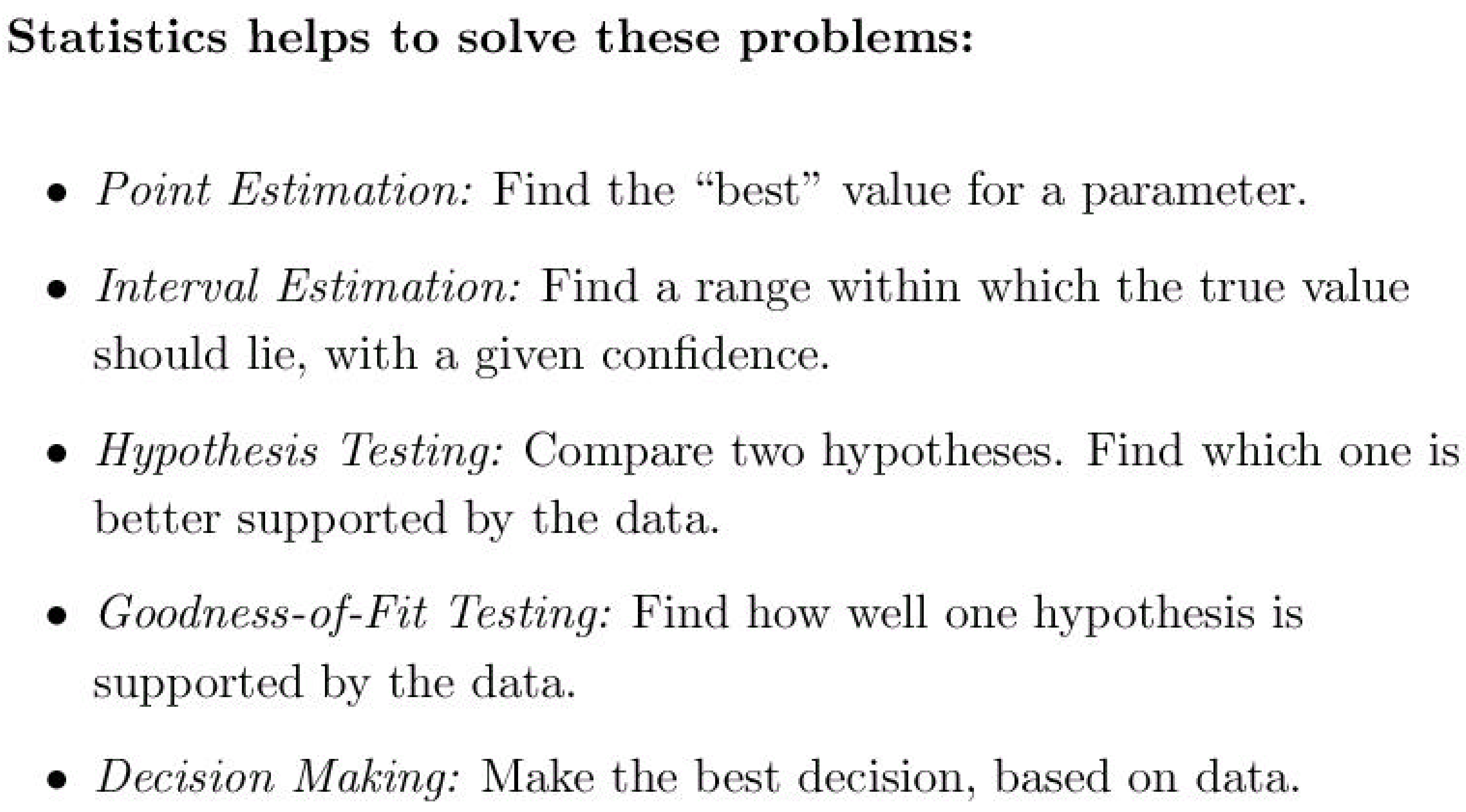
- On behalf of all of us, to the organizers.
- To all of you, for making this a stimulating week.
- To Michael Goldstein, for his help and good-natured tolerance.

So much to summarize: What will have a lasting impact?

- Well-founded methods from elsewhere in academe introduced into HEP and found to be practical and useful.
- HEP-specific adaptations of standard methods which are understood, practical, and useful.
- Methods in which “P” is defined.
- Lucid explanations of subtle issues.
- With higher threshold: completely new inventions by professional physicists/amateur statisticians.

Even with these criteria, there is not enough time to mention everything.
I apologize to those I left out.

Fred James

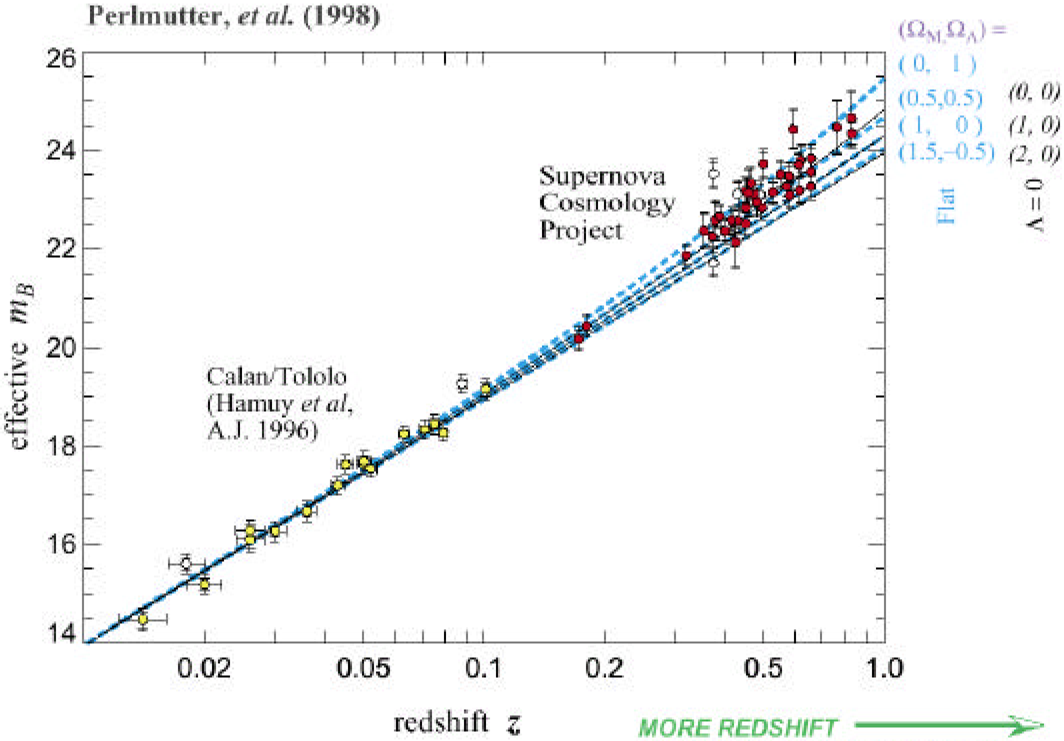


Statistics helps to solve these problems:

- *Point Estimation:* Find the "best" value for a parameter.
- *Interval Estimation:* Find a range within which the true value should lie, with a given confidence.
- *Hypothesis Testing:* Compare two hypotheses. Find which one is better supported by the data.
- *Goodness-of-Fit Testing:* Find how well one hypothesis is supported by the data.
- *Decision Making:* Make the best decision, based on data.

Important not to confuse these problems, e.g., interval estimation and goodness-of-fit testing.

FAINTER
(Farther)
(Further back in time)



MORE REDSHIFT
(More total expansion of universe since the supernova explosion)

In flat universe: $\Omega_M = 0.28$ [± 0.085 statistical] [± 0.05 systematic]

Prob. of fit to Λ = 0 universe: 1%

The “Other” PDF’s

- Illuminating talks by
  - Robert Thorne: Uncertainties in parton related quantities.
  - Mandy Cooper: The ZEUS NLO QCD fit to determine parton distributions and α_S.
  - Others in parallel: Blumlein…
- *Tough* business: uncertainties on *functions*!
- PDF’s matter: CDF compositeness, high-pT search reach.

(Thorne)

H1 $F_2^{e^+p}(x,Q^2)$ 1996-97 moderate Q^2 and 1996-97 high Q^2, and $F_2^{e^-p}(x,Q^2)$ 1998-99 high Q^2 small x.

ZEUS $F_2^{e^+p}(x,Q^2)$ 1996-97 small x wide range of Q^2.

NMC $F_2^{\mu p}(x,Q^2), F_2^{\mu d}(x,Q^2), (F_2^{\mu n}(x,Q^2)/F_2^{\mu p}(x,Q^2))$ medium x.

E665 $F_2^{\mu p}(x,Q^2), F_2^{\mu d}(x,Q^2)$ medium x.

BCDMS $F_2^{\mu p}(x,Q^2), F_2^{\mu d}(x,Q^2)$ large x.

SLAC $F_2^{\mu p}(x,Q^2), F_2^{\mu d}(x,Q^2)$ large x.

CCFR $F_2^{\nu(\bar{\nu})p}(x,Q^2), F_3^{\nu(\bar{\nu})p}(x,Q^2)$ large x , singlet, valence.

ZEUS $F_{2,c}^{e^+p}(x,Q^2)$ 1996-97 charm.

E605 $pN \rightarrow \mu\bar{\mu} + X$ large x sea.

E866 Drell-Yan asymmetry $\bar{u}, \bar{d}$ $\bar{d} - \bar{u}$.

CDF W-asymmetry u/d ratio at high x.

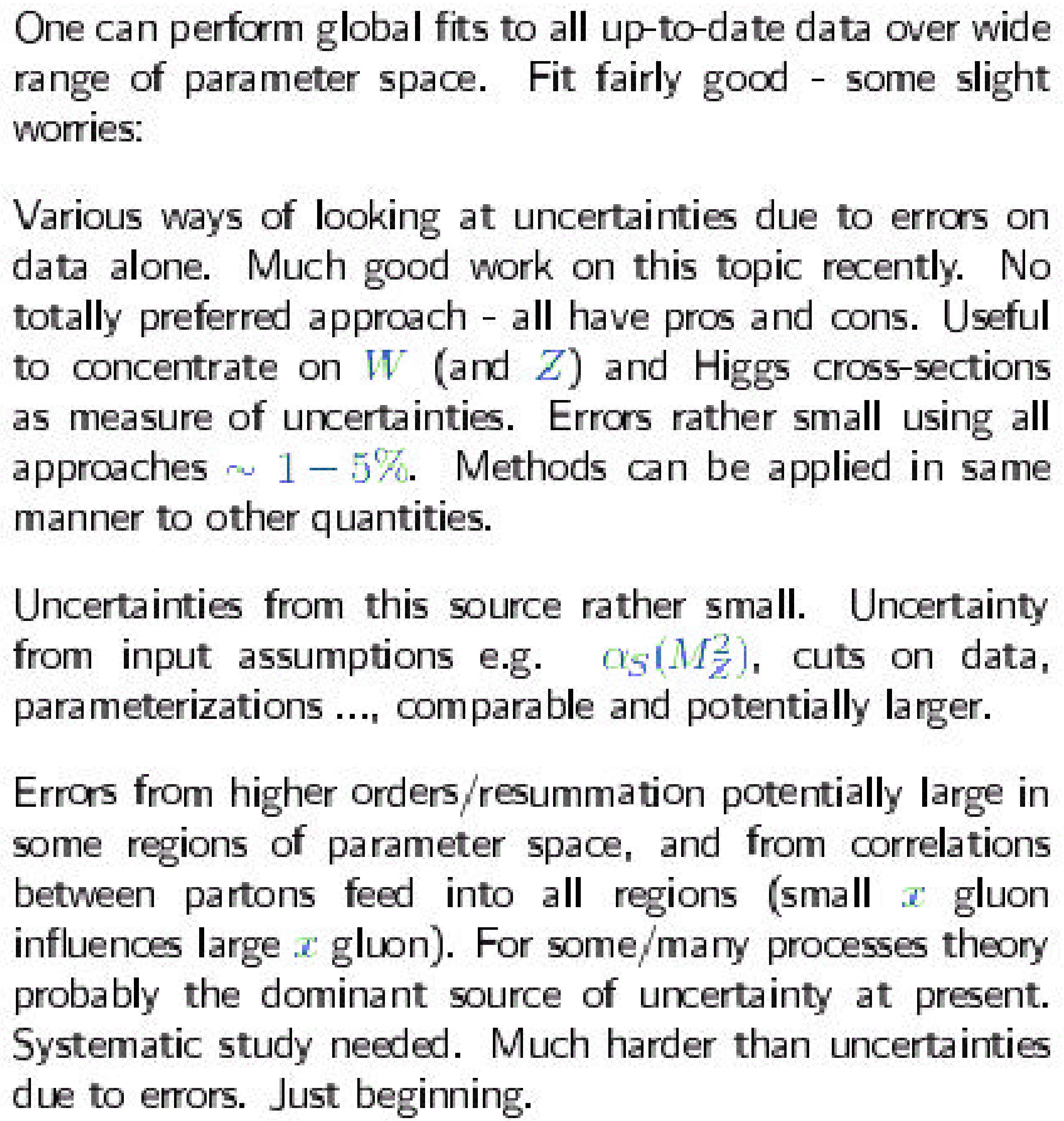
CDF Inclusive jet data high x gluon.

D0 Inclusive jet data high x gluon.

CCFR Dimuon data NuTev constrains strange sea.

Impressive progress!

(Thorne)



One can perform global fits to all up-to-date data over wide range of parameter space. Fit fairly good - some slight worries:

Various ways of looking at uncertainties due to errors on data alone. Much good work on this topic recently. No totally preferred approach - all have pros and cons. Useful to concentrate on W (and Z) and Higgs cross-sections as measure of uncertainties. Errors rather small using all approaches $\sim 1-5\%$. Methods can be applied in same manner to other quantities.

Uncertainties from this source rather small. Uncertainty from input assumptions e.g. $\alpha_S(M_Z^2)$, cuts on data, parameterizations ..., comparable and potentially larger.

Errors from higher orders/resummation potentially large in some regions of parameter space, and from correlations between partons feed into all regions (small x gluon influences large x gluon). For some/many processes theory probably the dominant source of uncertainty at present. Systematic study needed. Much harder than uncertainties due to errors. Just beginning.

1-5% on σ!

Higher order

Mandy Cooper: The ZEUS NLO QCD fit to determine parton distributions and α_S.

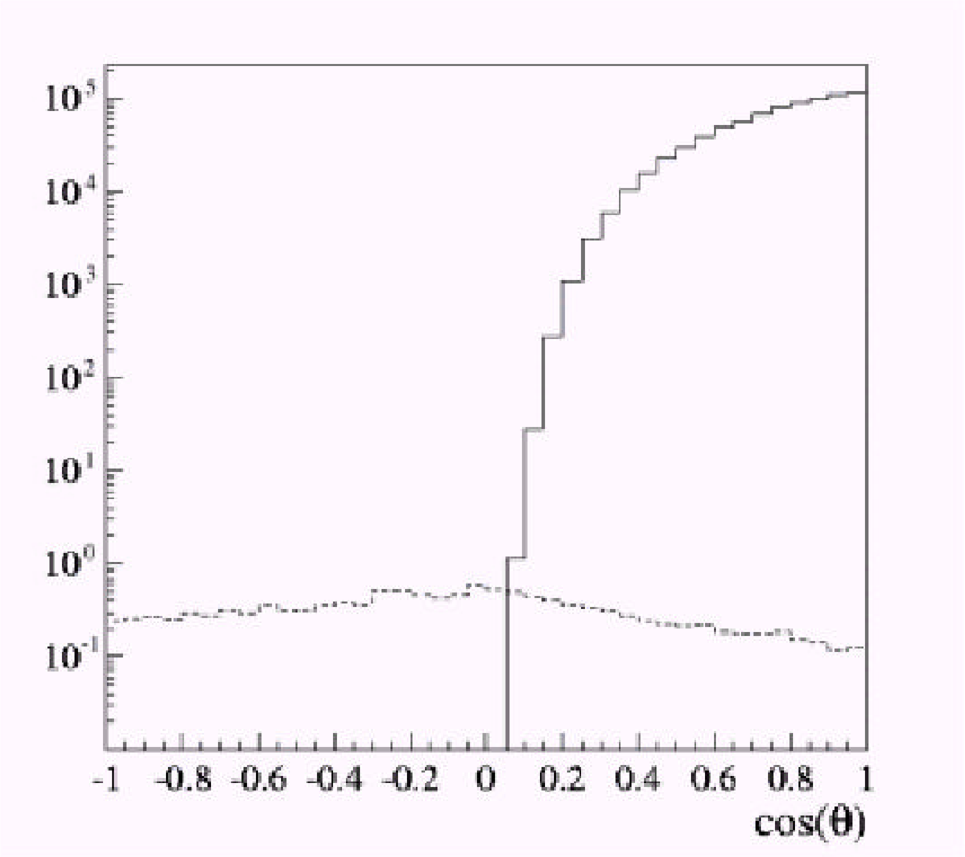
- Very stimulating talk, detailing many of the tough issues to be dealt with in a real-world example carried to completion.
- Along with Thorn, highlighted the issue of what $\Delta\chi^2$ to use for interval estimation.
- Some controversy over sqrt(2N); needs another look.
- Along with Thorn, the thorny problem of theoretical systematics: also needs another look?

Reports from the Trenches (I)

- Gary Hill and Tyce De Young: Application of Bayesian statistics to muon track reconstruction in Amanda
- Volker Blobel and Claus Kleinwort: A New method for the high-precision alignment of track detectors
- Nigel Smith and Dan Tovey: Dark Matter Searches
- R.K. Bock: Gamma/Hadron separation in atmospheric Cherenkov telescopes

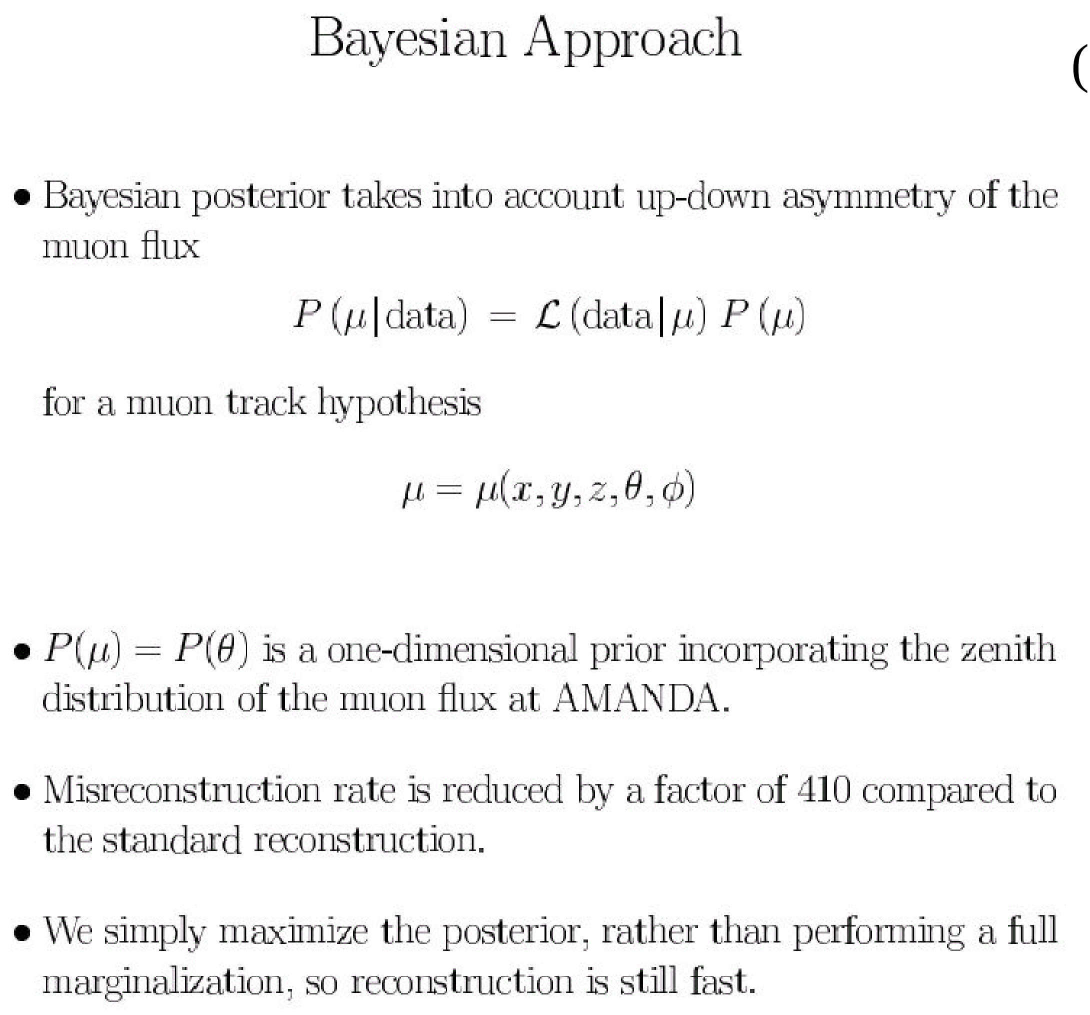
Identifying Muons from Neutrinos

- We wish to separate muons produced by neutrinos from muons produced in cosmic ray air showers



(Hill/DeYoung)

- We use the Earth as a filter, and observe the Northern Hemisphere



Bayesian Approach

(Hill/DeYoung)

- Bayesian posterior takes into account up-down asymmetry of the muon flux

$$P\left(\mu|\text{data}\right) \;=\; \mathcal{L}\left(\text{data}|\mu\right)\, P\left(\mu\right)$$

for a muon track hypothesis

$$\mu = \mu(x, y, z, \theta, \phi)$$

- $P(\mu) = P(\theta)$ is a one-dimensional prior incorporating the zenith distribution of the muon flux at AMANDA.
- Misreconstruction rate is reduced by a factor of 410 compared to the standard reconstruction.
- We simply maximize the posterior, rather than performing a full marginalization, so reconstruction is still fast.

But....

Very Interesting Technique!

- Let’s relate it to something we do: say particle ID in a detector:
  - In hot part of detector near beam: lots of background, we tighten particle-ID cuts
  - In lower-occupancy part of the detector away from beam, can loosen certain particle-ID cuts without letting in a lot of background
- Use our knowledge of position-dependent occupancy rates in Bayes’s Theorem to calculate the probability that a given particle in a given location is the species of interest.

Comments:

- If all input P’s are frequentist P’s, the output P(particle type | data) is a frequentist P.
- We can use this posterior frequentist P like any other observable for cuts, weights, etc. If we *independently* calibrate the signal efficiency/ background rejection of this use, there is nothing circular about using our knowledge of the input occupancies.
- If the input occupancy knowledge is imperfect it will not introduce a bias, but rather make the technique less powerful.

Bayes’s Theorem applies to any P satisfying the axioms of probability

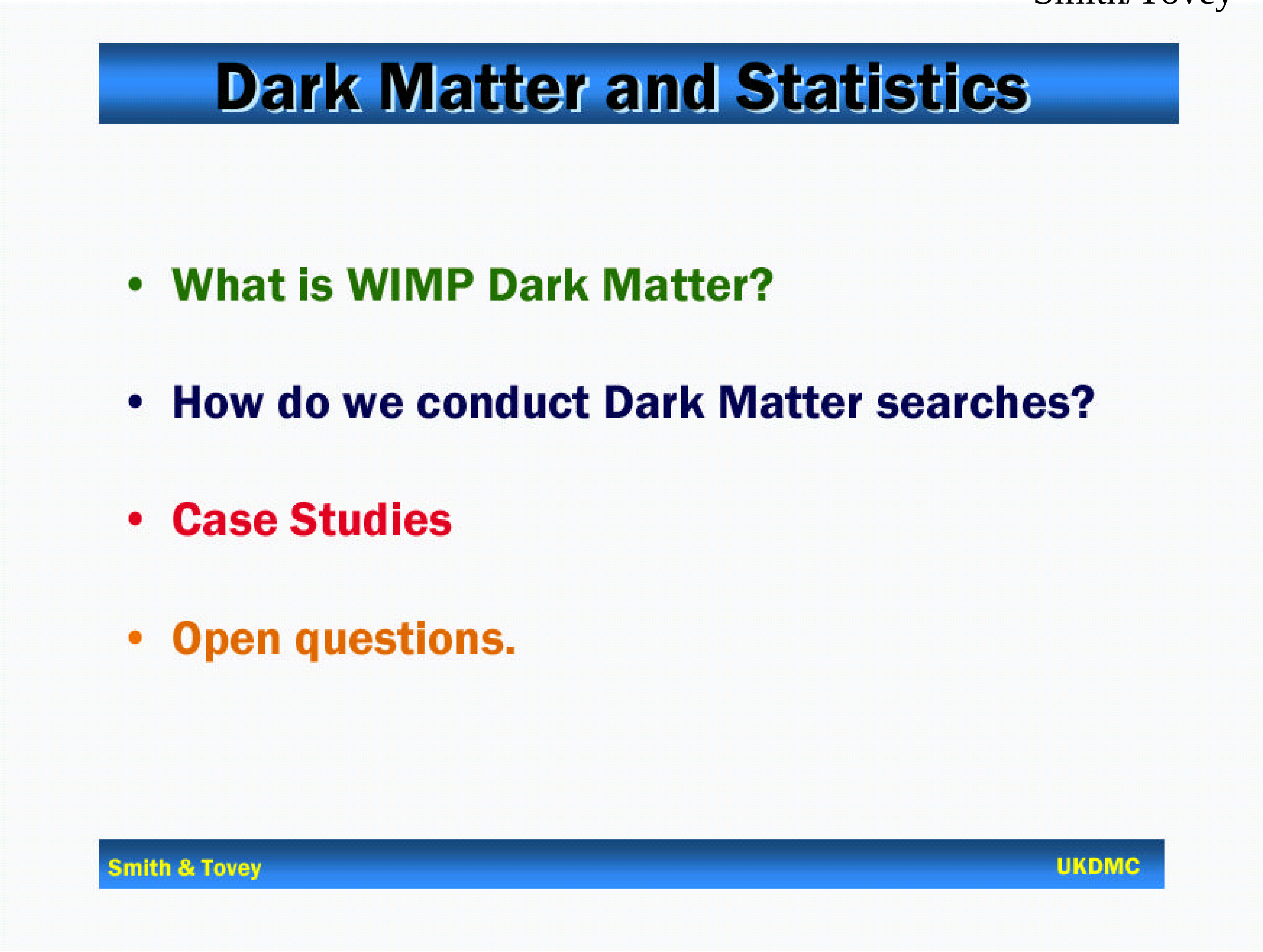
- Frequentist P: limiting frequency
  - Theorem not much use if the unknown is a constant of nature: P(unknown) = delta-function at unknown value
- Bayesian P: degree of belief
  - For constant of nature, P(unknown) can be combination of delta-function and continuous function, reflecting degree of belief
- Is the Amanda technique “Bayesian”?
  - Not if “Bayesian” implies “not frequentist”, as I think is common, even though frequency P is emulated in a certain application/limit of degree of belief.
- In any case, thanks for the instructive example!

Volker Blobel and Claus Kleinwort:

A New method for the high-precision alignment of track detectors

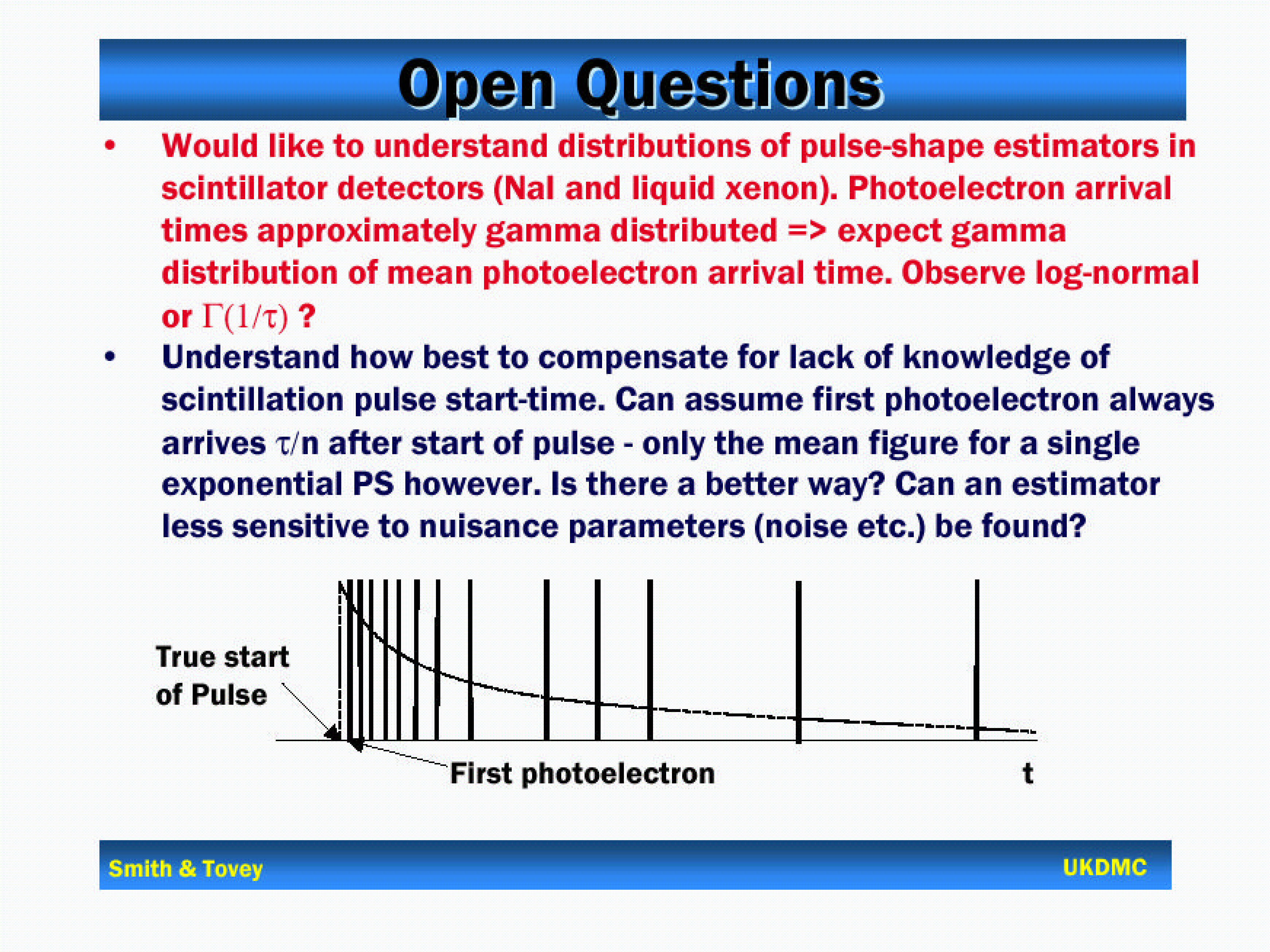
(transparency)

Just Beautiful!



Dark Matter and Statistics

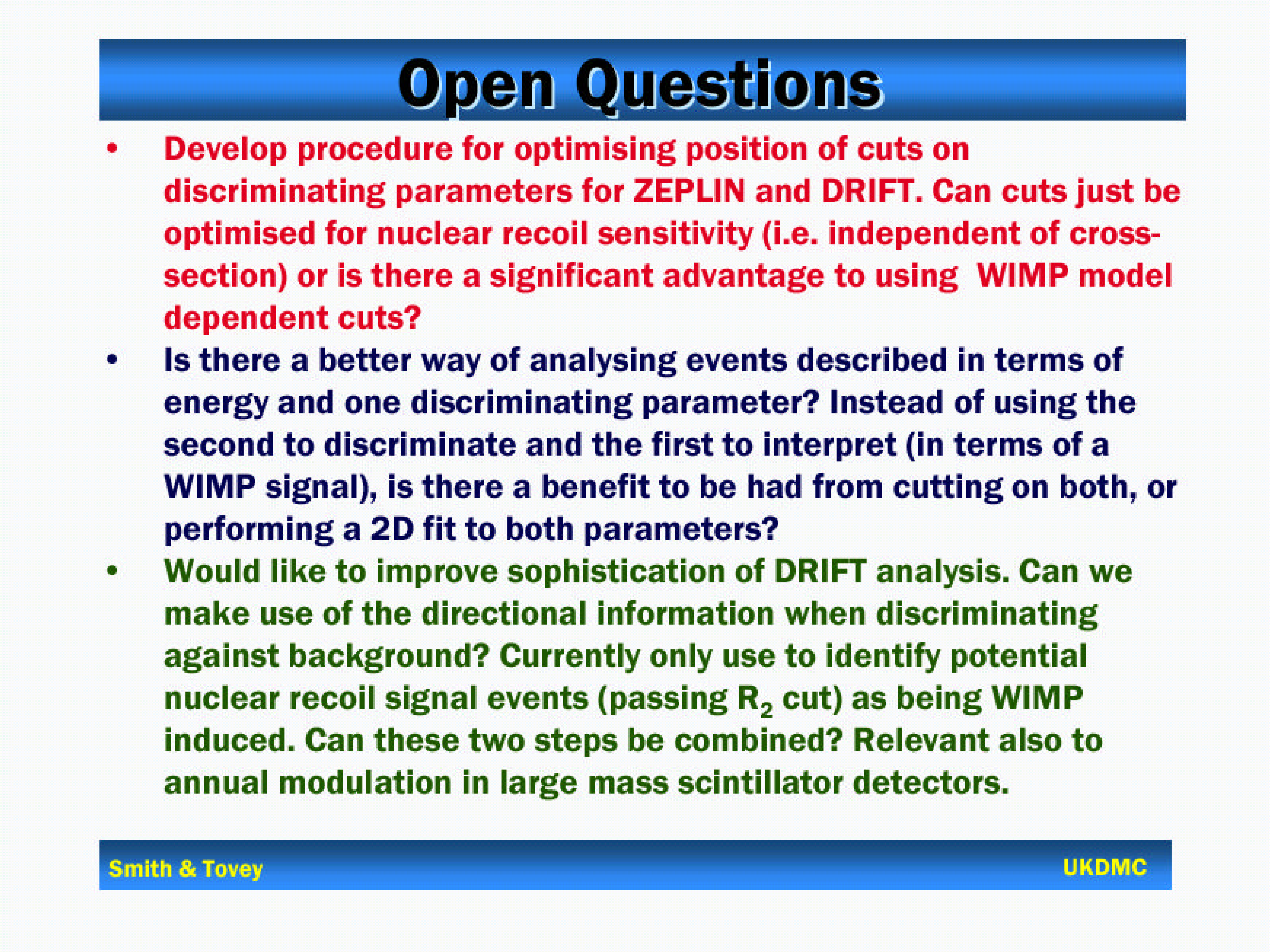
- What is WIMP Dark Matter?
- How do we conduct Dark Matter searches?
- Case Studies
- Open questions.



Open Questions

- Would like to understand distributions of pulse-shape estimators in scintillator detectors (NaI and liquid xenon). Photoelectron arrival times approximately gamma distributed => expect gamma distribution of mean photoelectron arrival time. Observe log-normal or $\Gamma(1/\tau)$?
- Understand how best to compensate for lack of knowledge of scintillation pulse start-time. Can assume first photoelectron always arrives τ/n after start of pulse - only the mean figure for a single exponential PS however. Is there a better way? Can an estimator less sensitive to nuisance parameters (noise etc.) be found?

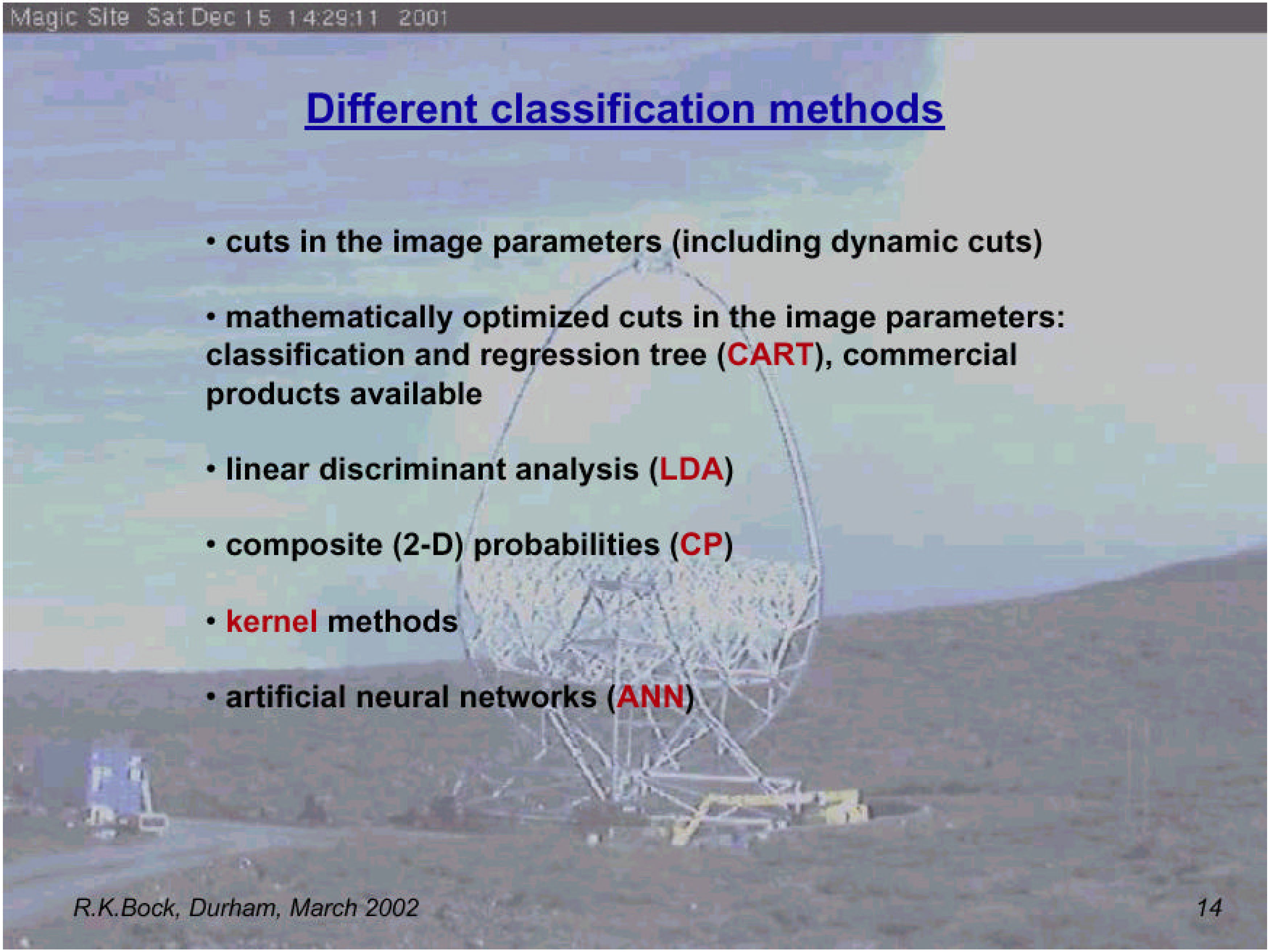
(Smith/Tovey)



Open Questions

- Develop procedure for optimising position of cuts on discriminating parameters for ZEPLIN and DRIFT. Can cuts just be optimised for nuclear recoil sensitivity (i.e. independent of cross-section) or is there a significant advantage to using WIMP model dependent cuts?
- Is there a better way of analysing events described in terms of energy and one discriminating parameter? Instead of using the second to discriminate and the first to interpret (in terms of a WIMP signal), is there a benefit to be had from cutting on both, or performing a 2D fit to both parameters?
- Would like to improve sophistication of DRIFT analysis. Can we make use of the directional information when discriminating against background? Currently only use to identify potential nuclear recoil signal events (passing R_2 cut) as being WIMP induced. Can these two steps be combined? Relevant also to annual modulation in large mass scintillator detectors.

This DM search is very important work, well-funded with first-rate collaboration and detectors. They came with some interesting statistics issues to discuss, and I am sorry that we could not give them the attention they deserve. I hope that some of the UK folks will take a look at these issues.



Different classification methods

- cuts in the image parameters (including dynamic cuts)

- mathematically optimized cuts in the image parameters: classification and regression tree (CART), commercial products available

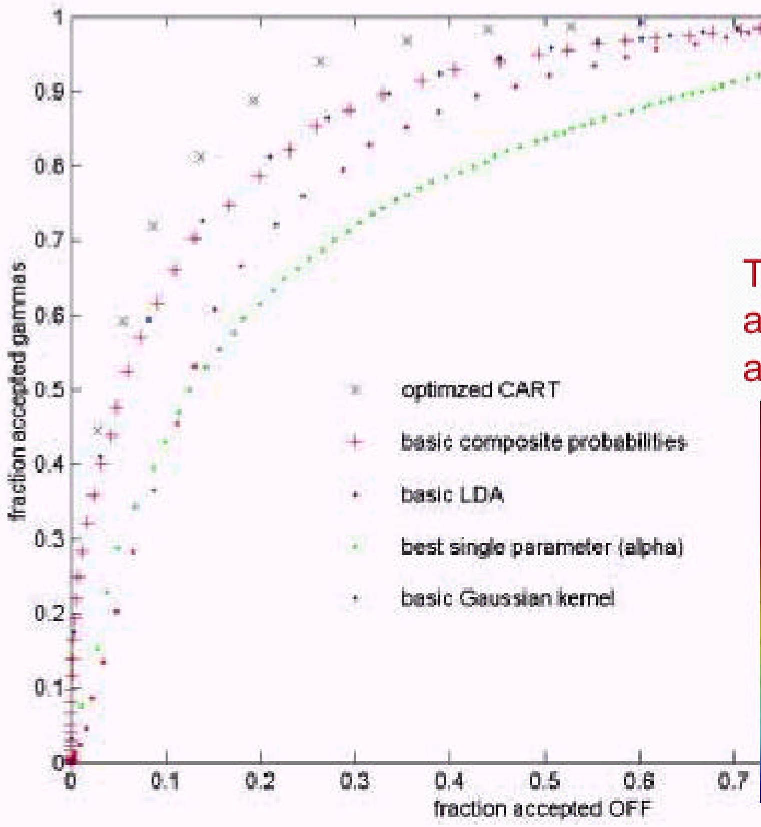
- linear discriminant analysis (LDA)

- composite (2-D) probabilities (CP)

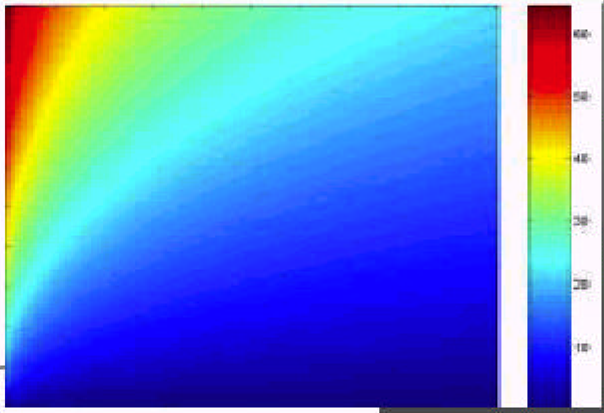
- kernel methods

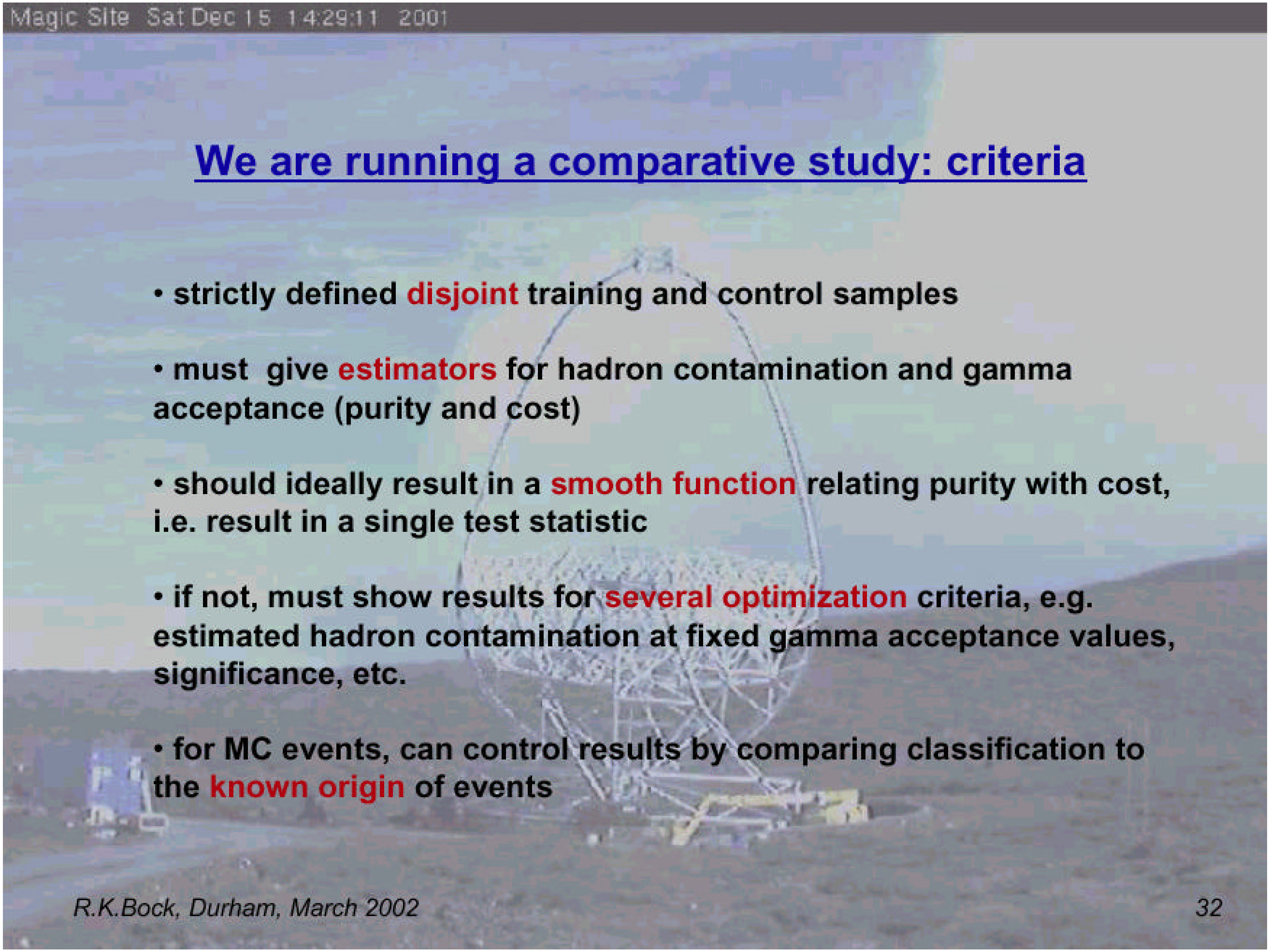
- artificial neural networks (ANN)

Different methods on the same data set



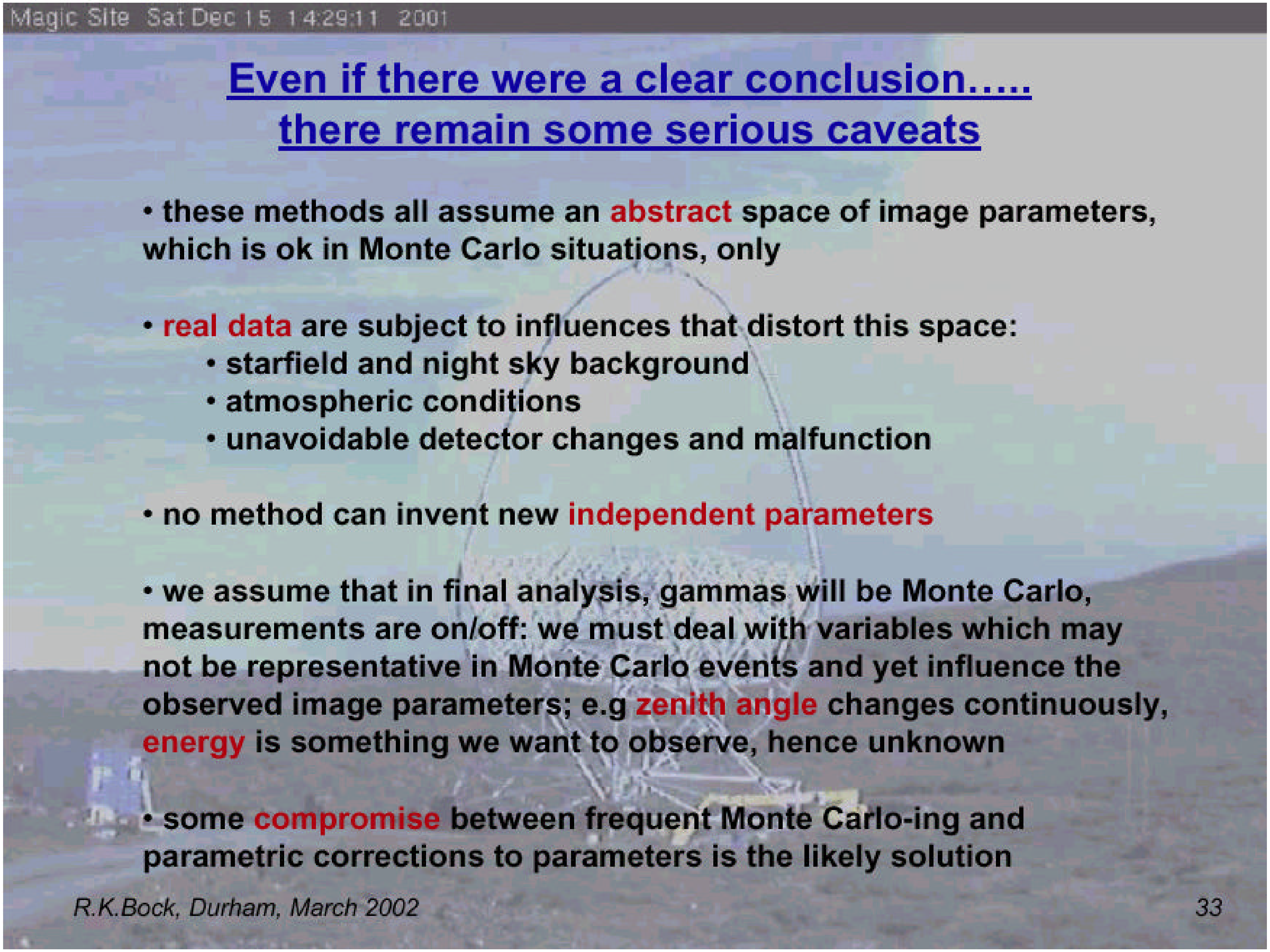
Typically, optimization parameters are fully defined by cost, purity, and sample size





We are running a comparative study: criteria

- strictly defined **disjoint** training and control samples
- must give **estimators** for hadron contamination and gamma acceptance (purity and cost)
- should ideally result in a **smooth function** relating purity with cost, i.e. result in a single test statistic
- if not, must show results for **several optimization** criteria, e.g. estimated hadron contamination at fixed gamma acceptance values, significance, etc.
- for MC events, can control results by comparing classification to the **known origin** of events



Magic Site Sat Dec 15 14:29:11 2001

Even if there were a clear conclusion…..
there remain some serious caveats

- these methods all assume an abstract space of image parameters, which is ok in Monte Carlo situations, only

- real data are subject to influences that distort this space:
  - starfield and night sky background
  - atmospheric conditions
  - unavoidable detector changes and malfunction

- no method can invent new independent parameters

- we assume that in final analysis, gammas will be Monte Carlo, measurements are on/off: we must deal with variables which may not be representative in Monte Carlo events and yet influence the observed image parameters; e.g zenith angle changes continuously, energy is something we want to observe, hence unknown

- some compromise between frequent Monte Carlo-ing and parametric corrections to parameters is the likely solution

R.K.Bock, Durham, March 2002 33

Lucid battle-tested studies like this should be required reading for us and our students!

Reports from the Trenches (II)

- Chris Parkes: Practicalities of combining analyses: W physics results at LEP
- Sergei Redin: Advanced Statistical Techniques in the muon g–2 experiment at BNL
- Bruce Yabsley: Statistical practice at the Belle experiment, and some questions
- Fabrizio Parodi et al: How to include the information coming from B_S^0 oscillations in CKM fits

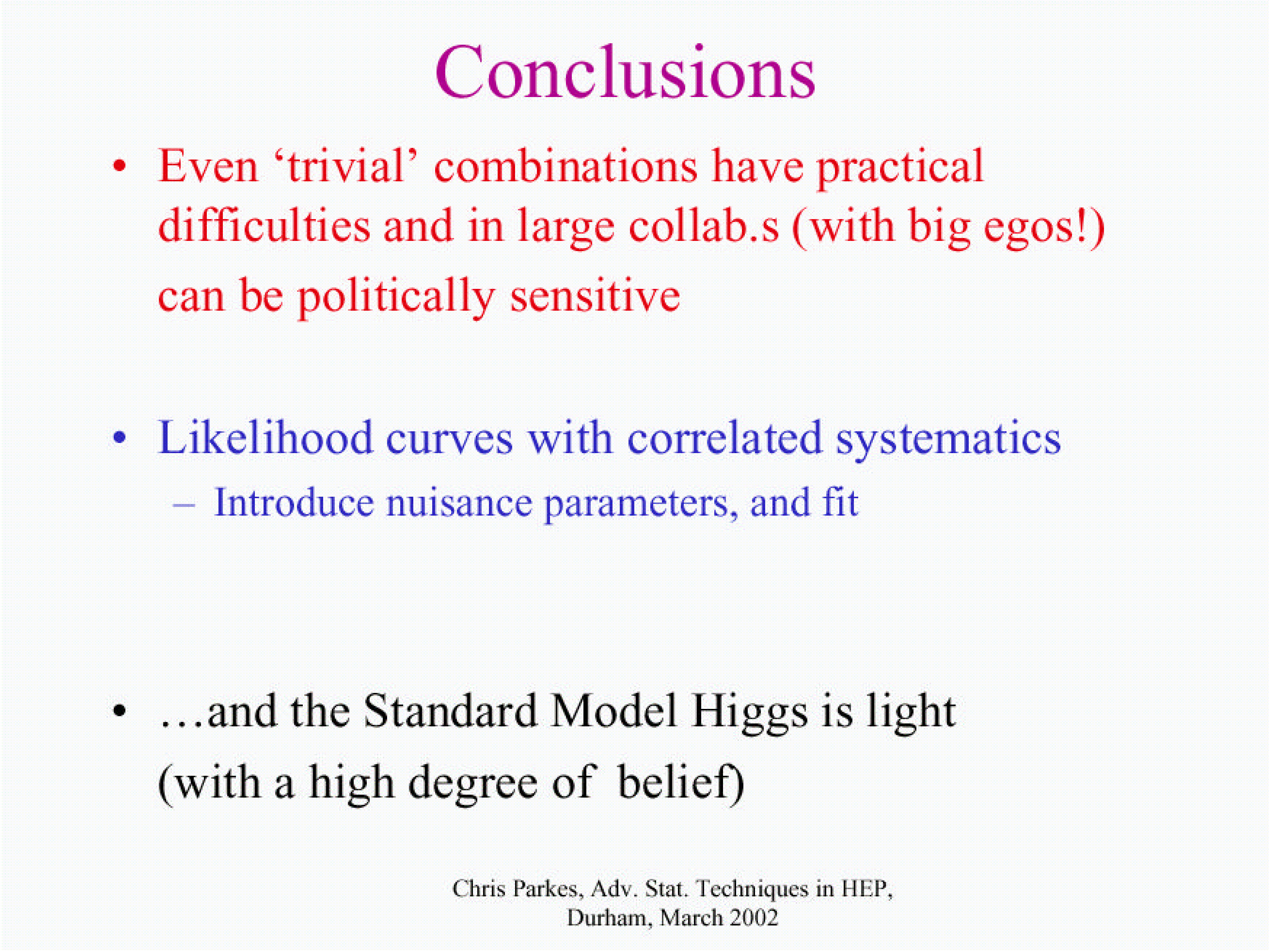
Practicalities of Combining Analyses:
W Physics Results at LEP

Chris
Parkes

Now the stuff you don’t normally see…

RC: An informative talk about both methodology and sociology!

An important reminder: pragmatic considerations (sometimes even irrational) can be as important as principles in order to get out a result.



Conclusions

- Even ‘trivial’ combinations have practical difficulties and in large collab.s (with big egos!) can be politically sensitive

- Likelihood curves with correlated systematics
  - Introduce nuisance parameters, and fit

- …and the Standard Model Higgs is light (with a high degree of belief)

- LEP experiments contained a sizable fraction of world HEP community, and reached very mature state of analysis.
  - We have much to learn from them, both theoretical and practical.

Sergei Redin: Advanced Statistical Techniques in the muon g–2 experiment at BNL

- ppm measurement!
- $G(t) = N_0\, e^{-t/\tau}\, [1 + A\cos(\omega_0 t+\phi)$
- Five parameters, ω_0 is the one of main interest for new physics
- Shows value of examining a problem analytically: can give insight that hard to get from M.C.

It’s a great reminder for our students that one can learn a lot by analytic calculation. We live at a time when a student’s first reaction may be to run a bunch of GEANT jobs, but it may take a lot of CPU and a lot of log paper to discover scaling laws which can be found by hand with some thought.