Big Science, Big Data and Statistics

Harrison B. Prosper
Florida State University

DESY, Hamburg, Germany

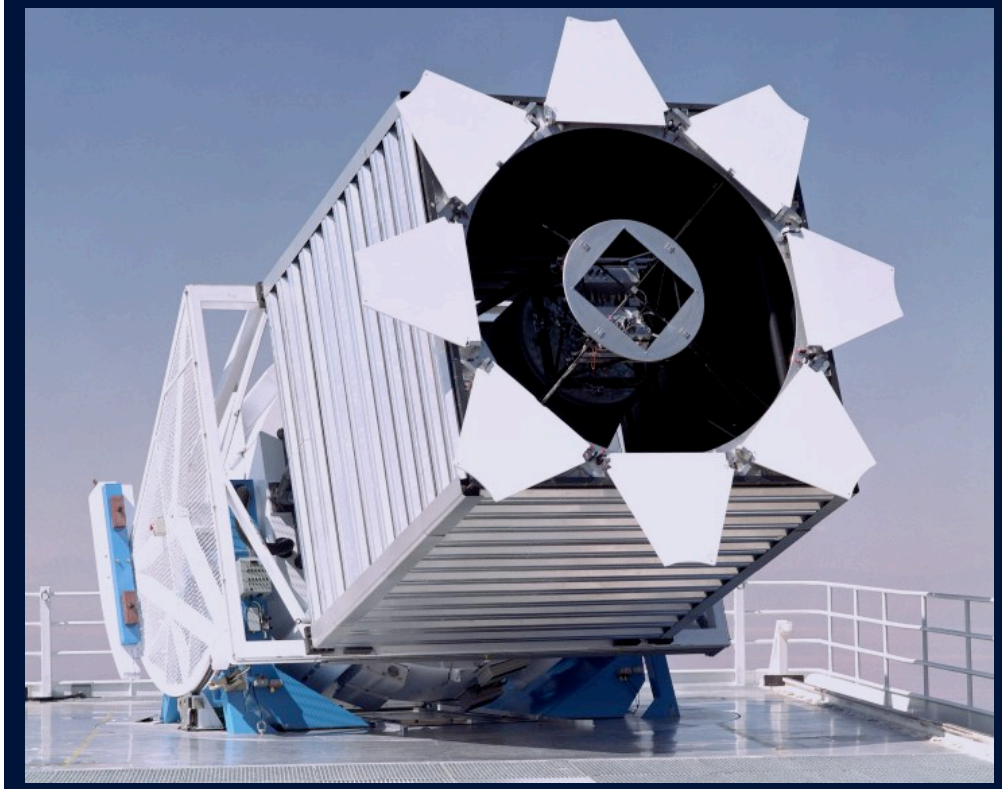
April 4, 2014

Outline

- Big Science
- Big Data
- Statistics *in* Particle Physics

BIG SCIENCE

"WHEN the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy." *The Economist*, Feb 25th 2010

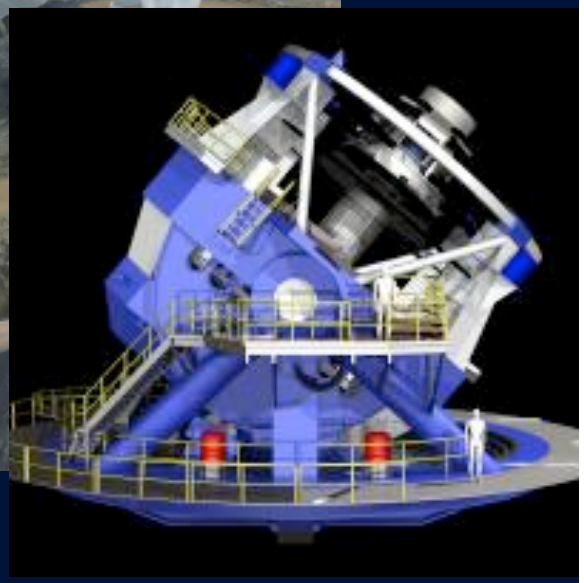


SDSS III

http://www.sdss3.org/press/dr9.php

The Large Synoptic Survey Telescope

A 3200 Megapix camera that will make a 10-year movie of half the sky in multiple wavelength bands, starting ~2022

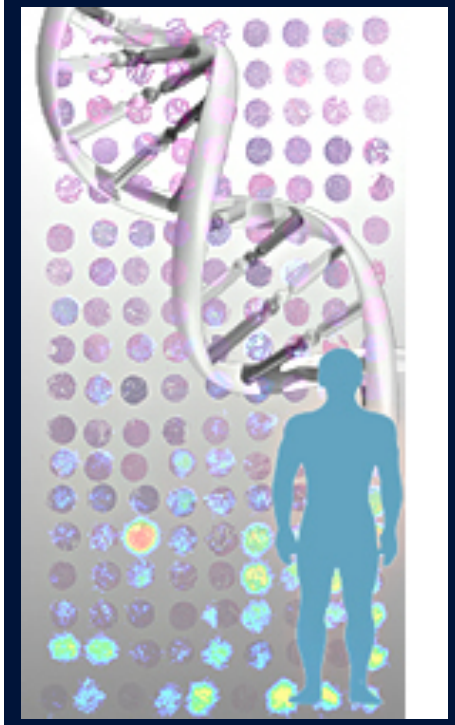


Hanford, Washington

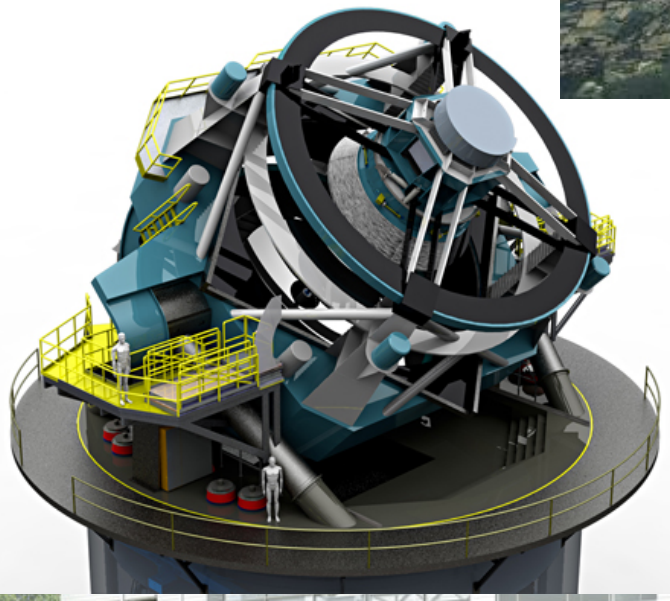
Livingston, Louisiana

AdvLIGO
Virgo

LHC



US
National
Human
Genome
Project



LSST

XFEL

BIG DATA

Data Data Everywhere

"As of 2012, about 2.5 EB (exabytes) of data are created each day, and that number is doubling every 40 months or so."

Harvard Business Review, October 2012

1 EB = 1 000 000 000 GB

Data Data Everywhere

1 byte a single character

1 kilobyte a short story

1 megabyte a small novel

1 gigabyte a movie (TV resolution)

1 terrabyte printed paper from 50,000 trees

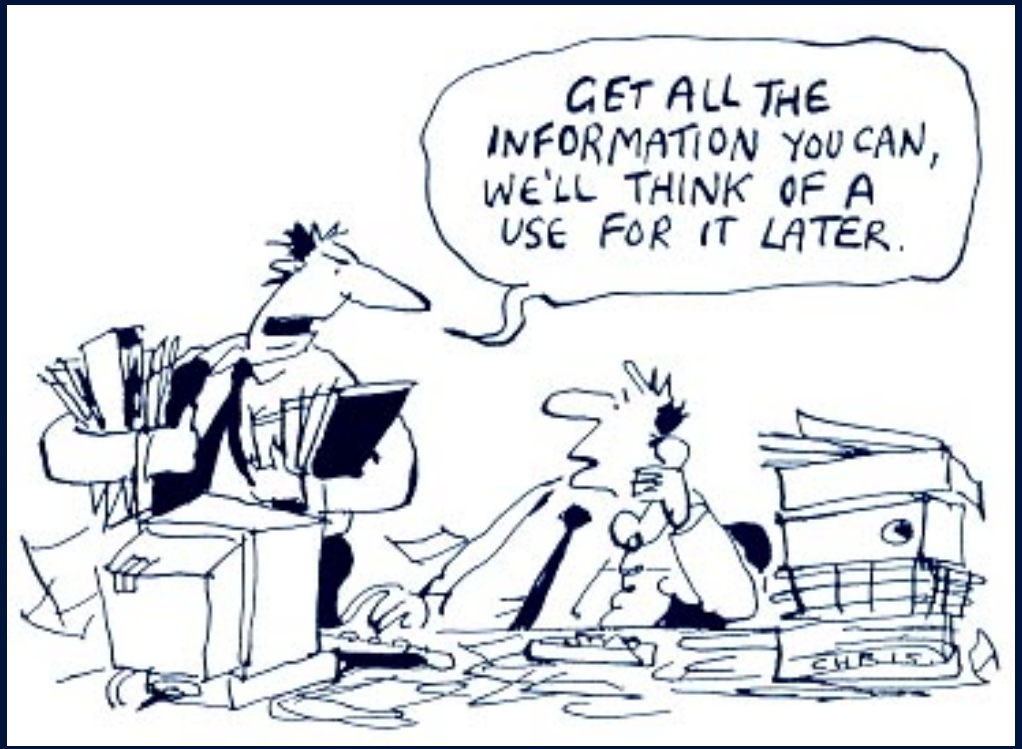
1 petabyte 5 years of EOS data

1 exabyte all words ever spoken

http://highscalability.com/blog/2012/9/11/how-big-is-a-petabyte-exabyte-zettabyte-or-a-yottabyte.html

Data Data Everywhere

Project	Expected Data Size	Period
SDSS	100 TB	2000 – 2015
LSST	100 000 TB	2022 – 2032
LHC	15 000 000 TB	2010 – 2035



“Imagination is everything.”

Albert Einstein

Open Data is the Future! Why?

- Governments will insist on it!

Open Data is the Future! Why?

Office of Science and Technology Policy

About OSTP | Pressroom | OSTP Blog | Divisions | Initiatives | R&D Budgets | Resc

Expanding Public Access to the Results of Federally Funded Research

Subsc

Posted by Michael Stebbins on February 22, 2013 at 12:04 PM EDT

E-Mail Tweet Share +

The Obama Administration is committed to the proposition that citizens deserve easy access to the results of scientific research their tax dollars have paid for. That's why, in a policy memorandum released today, OSTP Director John Holdren has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally

Open Data is the Future! Why?

as of midnight April 3rd 2014

In the age group 15 – 65 years, there are ~**5.0 billion brains**, many of whom are extremely smart. Compare this with the **~3,000** permanent brains at CERN and DESY, plus the **~10,000** visiting brains!

Open Data is the Future! Why?

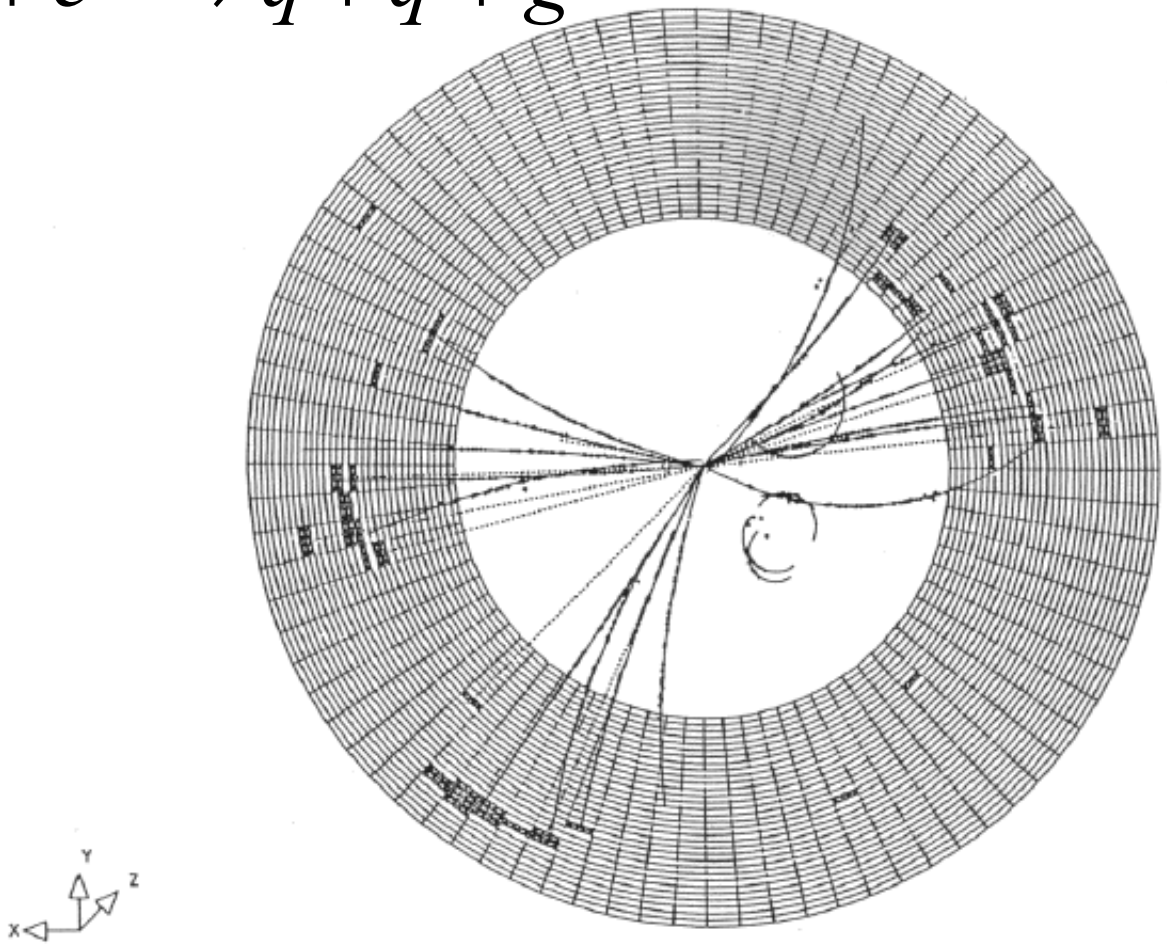
- Reproducibility. If it is science, ideally, it is reproducible.
- Data acquired today may yield new science tomorrow.

Hamburg, 10 December 1979

Courtesy Prof. Robin Marshall

DESY, JADE – 1979

$$e^+ + e^- \rightarrow q + \bar{q} + g$$



*** SUMS (GEV) *** PTOT 35.768 PTRANS 29.964 PLONG 15.788 CHARGE -2
TOTAL CLUSTER ENERGY 15.169 PHOTON ENERGY 4.893 NR OF PHOTONS 11

DESY, JADE – 2009

Publications based on resurrected JADE data (1997-2009)

7) Study of moments of event shapes and a determination of alpha(S) using e+ e- annihilation data from Jade.
Christoph Pahl (Munich, Max Planck Inst. & Munich, Tech. U.) , Siegfried Bethke, Stefan Kluth, Jochen Schieck, the JADE collaboration (Munich, Max Planck Inst.) . MPP-2008-135, May 8, 2009. 14pp. Eur.Phys.J.C60:181-196,2009, e-Print: arXiv:0810.2933 [hep-ex]

6) Determination of the Strong Coupling alpha(S) from hadronic Event Shapes and NNLO QCD predictions using JADE Data.
By JADE Collaboration (S. Bethke et al.). MPP-2008-131, Oct 2008. 9pp., Submitted to Eur.Phys.J.C. e-Print: arXiv:0810.1389 [hep-ex]

5) Measurement of the strong coupling alpha(s) from the four-jet rate in e+ e- annihilation using JADE data.
By JADE Collaboration (J. Schieck et al.). MPP-2006-161, 2006. 11pp. Eur.Phys.J.C48:3-13,2006, Erratum-ibid.C50:769,2007.
e-Print: arXiv:0707.0392 [hep-ex]

4) Measurement of the longitudinal and transverse cross-sections in e+ e- annihilation at s**(1/2) = 35-GeV - 44-GeV.
By JADE Collaboration (M. Blumenstengel et al.). MPI-PHE-2001-11, Jun 2001. 12pp., Phys.Lett.B517:37-46,2001.
e-Print: hep-ex/0106066

3) QCD analyses and determinations of alpha(s) in e+ e- annihilation at energies between 35-GeV and 189-GeV.
By JADE collaboration and OPAL Collaboration (P. Pfeifenschneider et al.). CERN-EP-99-175, Dec 1999. 49pp.
Eur.Phys.J.C17:19-51,2000. e-Print: hep-ex/0001055

2) C parameter and jet broadening at PETRA energies.
By JADE Collaboration (O. Biebel et al.). PITHA-98-21A, Mar 1999. 14pp. Phys.Lett.B459:326-334,1999., e-Print: hep-ex/9903009

1) A Study of event shapes and determinations of alpha-s using data of e+ e- annihilations at s**(1/2) = 22-GeV to 44-GeV.
By JADE Collaboration (P.A. Movilla Fernandez et al.). PITHA-97-27, Aug 1997. 36pp. , Eur.Phys.J.C1:461-478,1998.
e-Print: hep-ex/9708034

==========

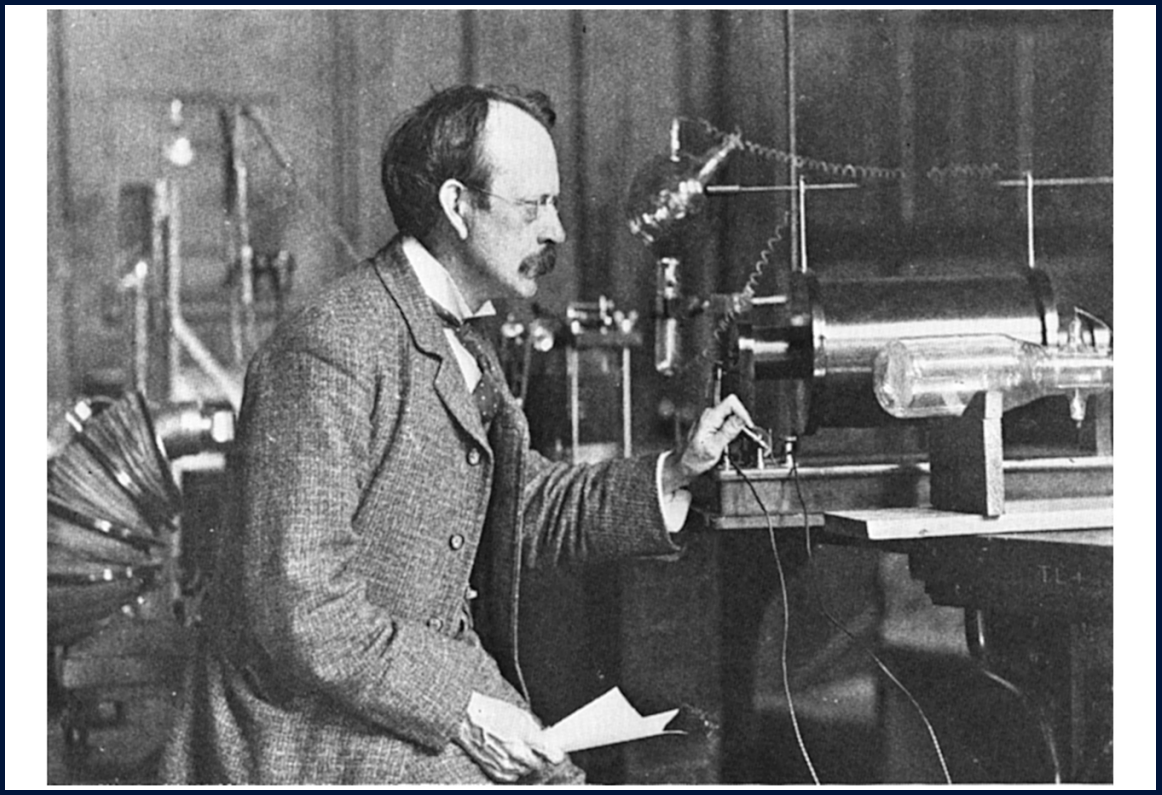
3) Tests of analytical hadronisation models using event shape moments in {\epem} annihilation.
C. Pahl, S. Bethke, O. Biebel, S. Kluth, J. Schieck . MPP-2009-38, Apr 2009. 17pp. e-Print: arXiv:0904.0786 [hep-ex]

2) Tests of power corrections for event shapes in e+ e- annihilation. P.A. Movilla Fernandez, S. Bethke, O. Biebel, S. Kluth (Munich, Max Planck Inst.) . MPI-PH-2001-005, May 2001. 27pp., Eur.Phys.J.C22:1-15,2001. e-Print: hep-ex/0105059

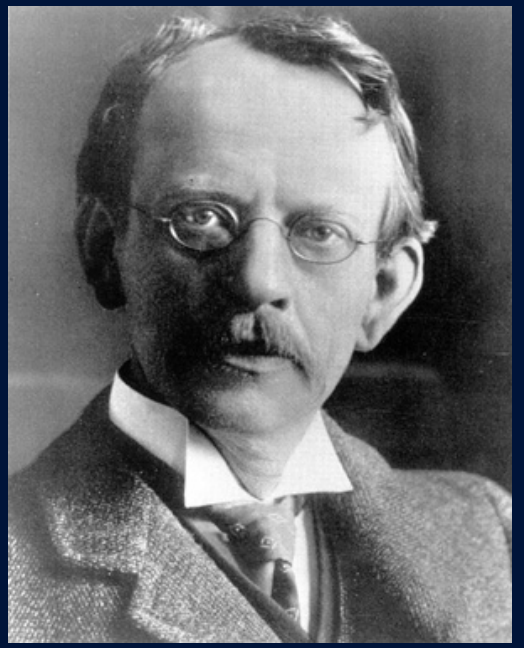
1) A Measurement of the QCD color factors using event shape distributions at s**(1/2) = 14-GeV to 189-GeV.
S. Kluth, P.A. Movilla Fernandez, S. Bethke, C. Pahl, P. Pfeifenschneider (Munich, Max Planck Inst.) . MPI-PHE-2000-19, Dec 2000. 25pp.
Eur.Phys.J.C21:199-210,2001. e-Print: hep-ex/0012044

STATISTICS *IN PARTICLE PHYSICS*

Particle Physics, Then…



The discovery of the electron, 1897
J.J. Thomson



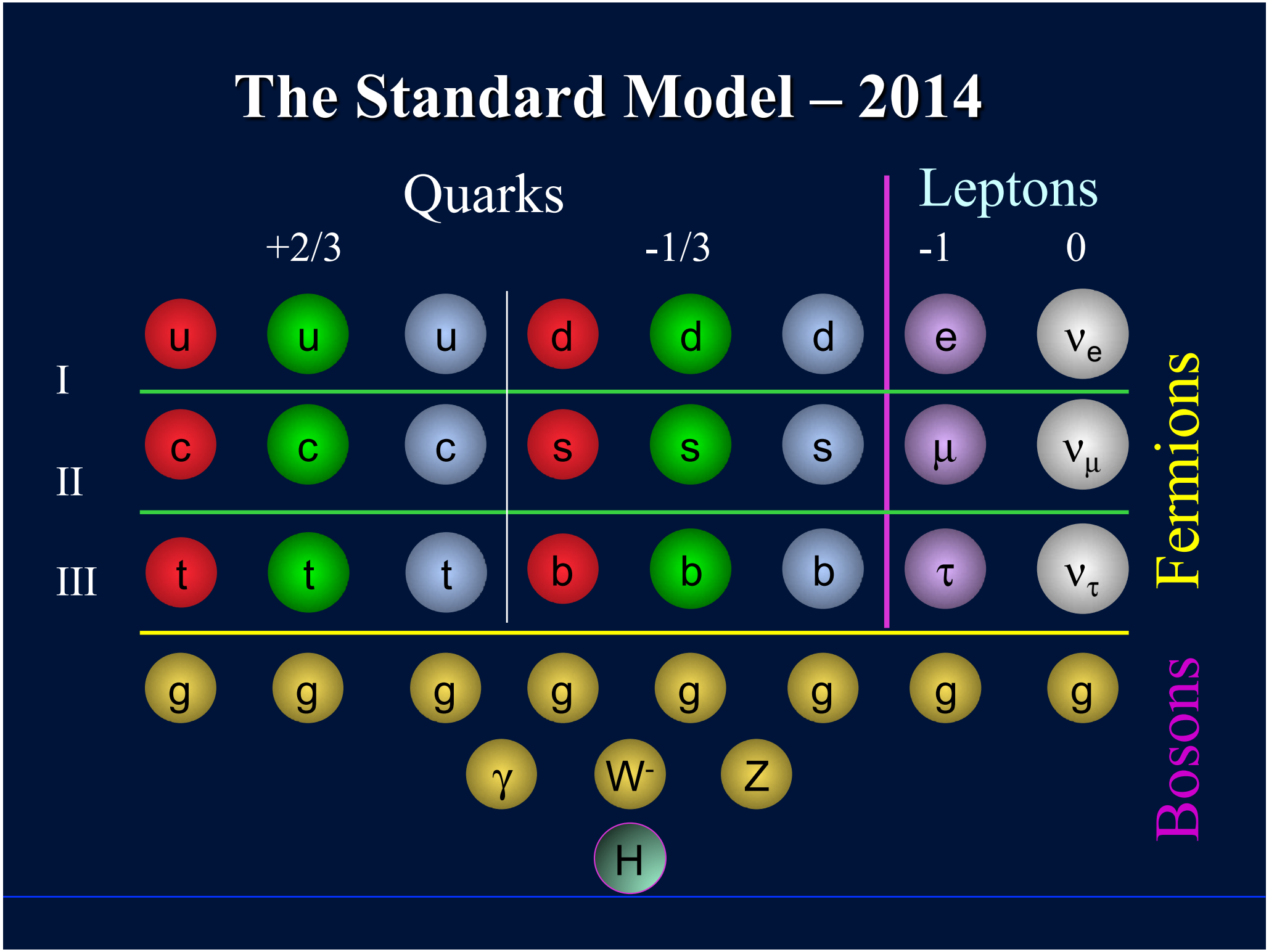
The Large Hadron Collider

…and Now!

One Ring to rule them all, One Ring to find them,
One Ring to bring them all, And in the darkness
reveal them


CMS Exclusion Potential
Not-yet-excluded region: ~[115-130] GeV
The five decay modes discussed today have comparable sensitivities for exclusion.
Most analyses used in this combination have been re-optimized. In order to avoid the possibility of an unintended bias, all selection criteria in the analyses of the 2011 and 2012 data were fixed before looking at the result in the signal region.
CMS Preliminary
95% CL limit on σ/σSM
Higgs boson mass (GeV)

The Standard Model – 2014



A Very Short List of Questions

- What determines the values of particle masses?
- Why is the Higgs boson so light?
- Why is the <Higgs field> = 246 GeV in the vacuum?
- What is the origin of the observed pattern of particles?
- What is the origin of the observed particles symmetries?
- What is the origin of particle quantum numbers?

: : :

Technicolor

Supersymmetry ***

Compositeness ***

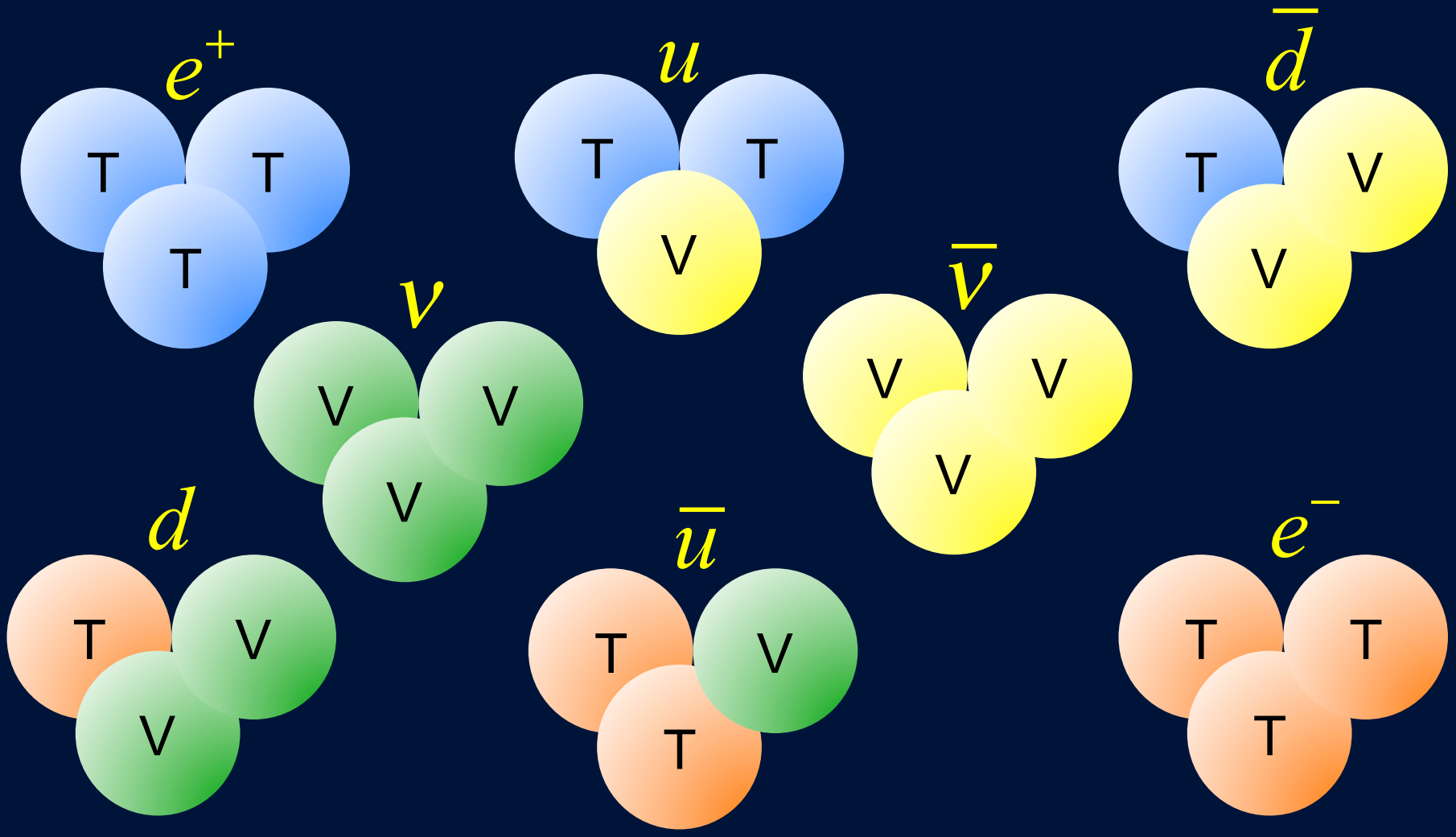
Extra Dimensions

Strings

Brane Worlds

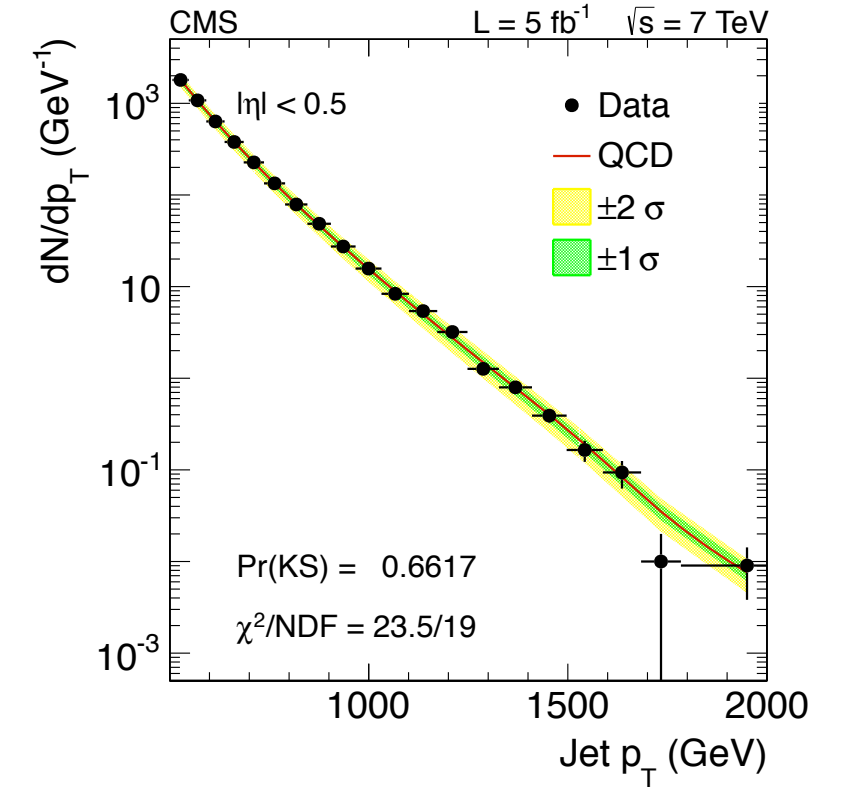
Multiverse

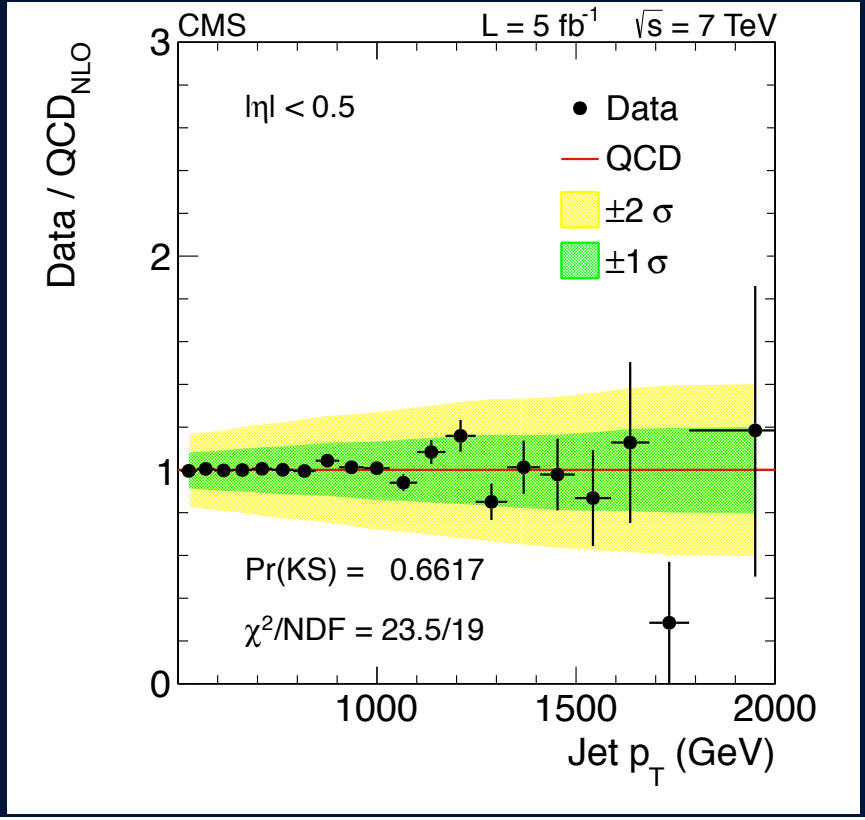
The Generation Model



B. Robson, "The Generation Model and the Origin of Mass", Int. J. Mod. Phys. E18 (2009)

The Search for Compositeness





PHYSICAL REVIEW D 87, 052017 (2013)

Search for contact interactions using the inclusive jet p_T spectrum in pp collisions at $\sqrt{s} = 7$ TeV

S. Chatrchyan *et al.**
(CMS Collaboration)

(Received 21 January 2013; published 26 March 2013)

The Decade Ahead

"the most pressing question at the LHC will be to figure out whether there is *any* evidence for physics beyond the standard model, and then most broadly what theoretical framework best describes the new physics"

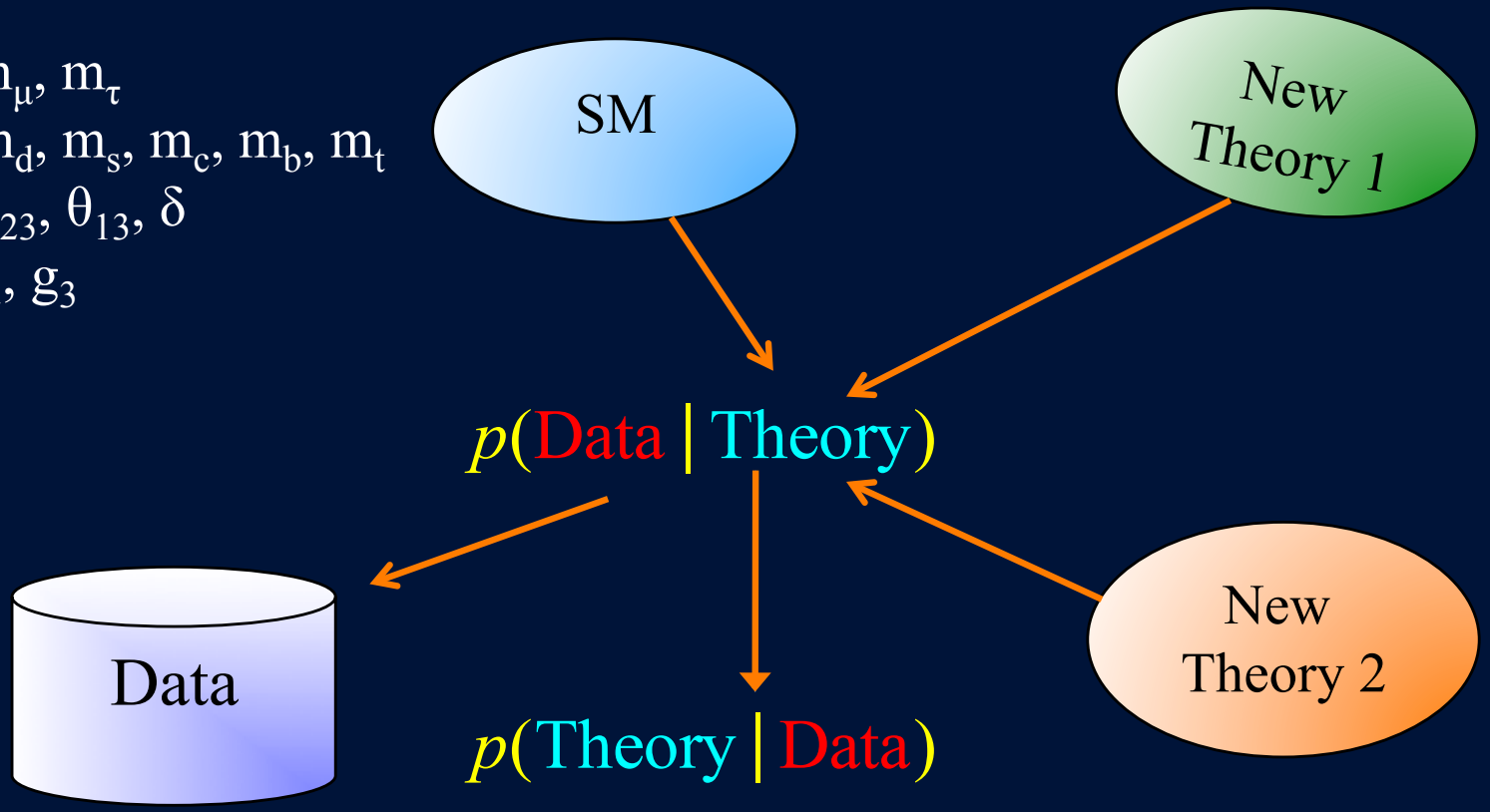
Supersymmetry and the LHC Inverse Problem,

N. Arkani-Hamed, G.L.Kane, J. Thaler, L.Wang, JHEP **0608**, 070 (2006)

The Decade Ahead

All interesting theories are ***multi-parameter*** models

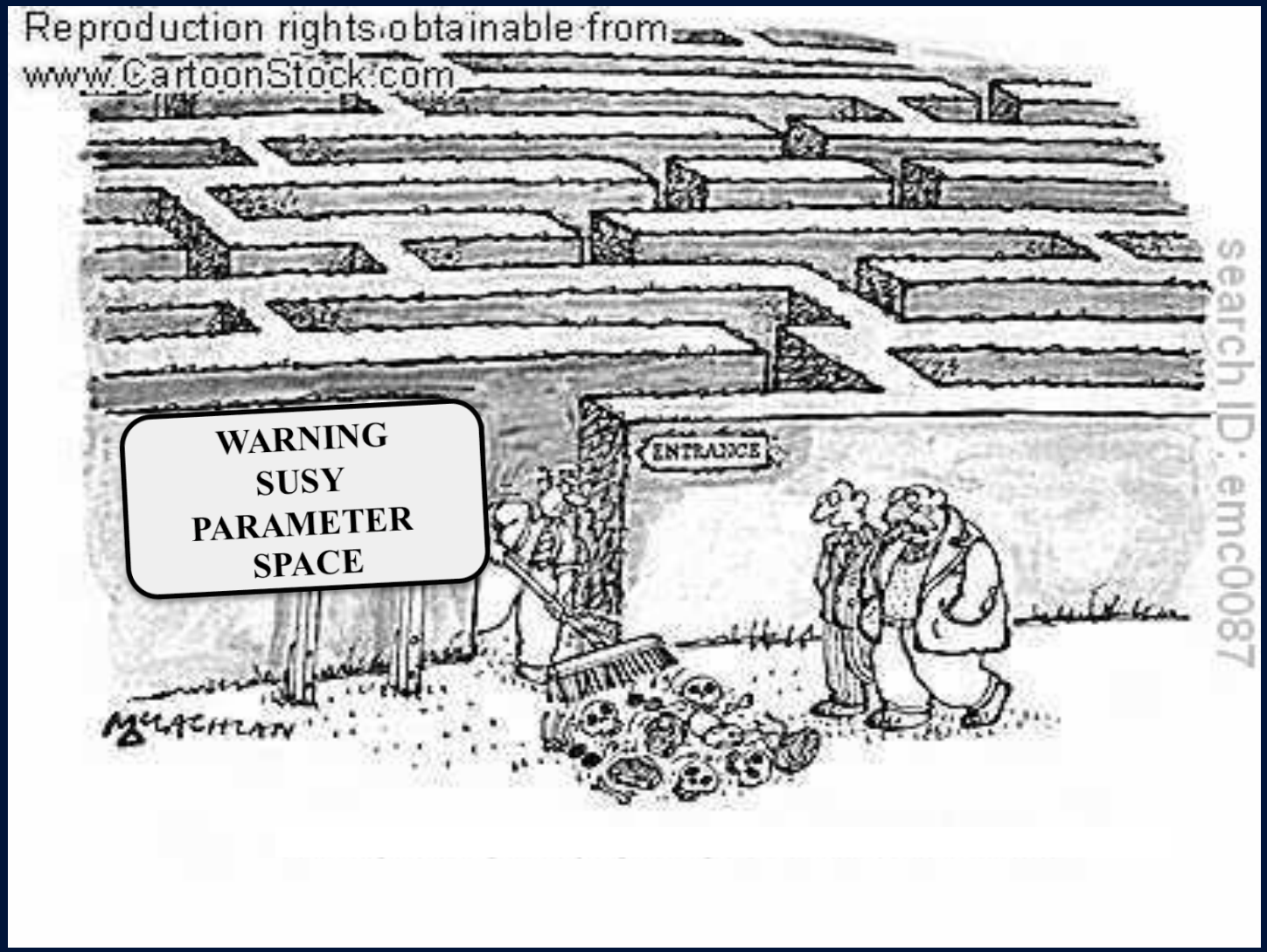
m_e, m_μ, m_τ
m_u, m_d, m_s, m_c, m_b, m_t
θ_{12}, θ_{23}, θ_{13}, δ
g_1, g_2, g_3
θ_{QCD}
μ, λ



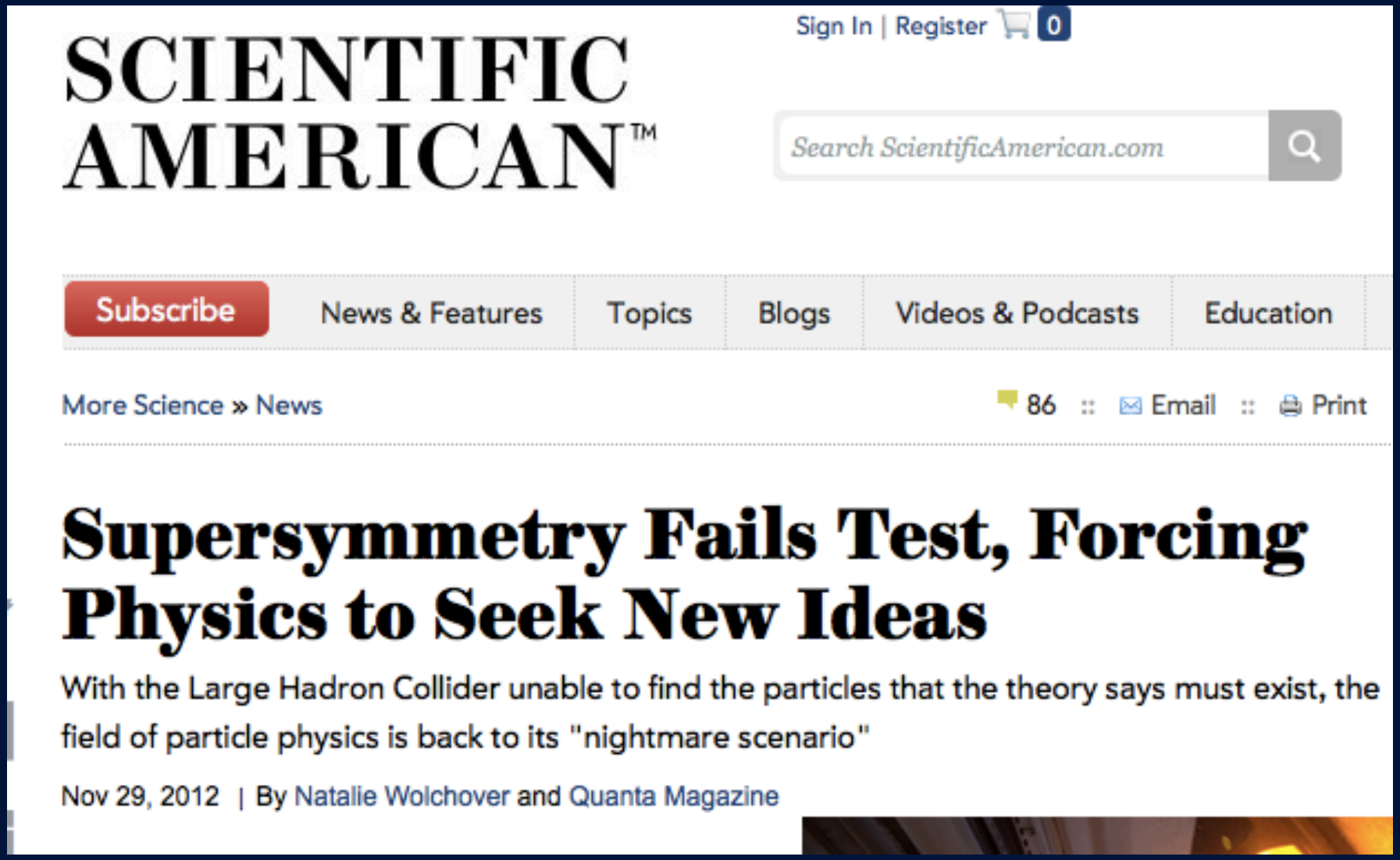
Basic *statistical* questions:

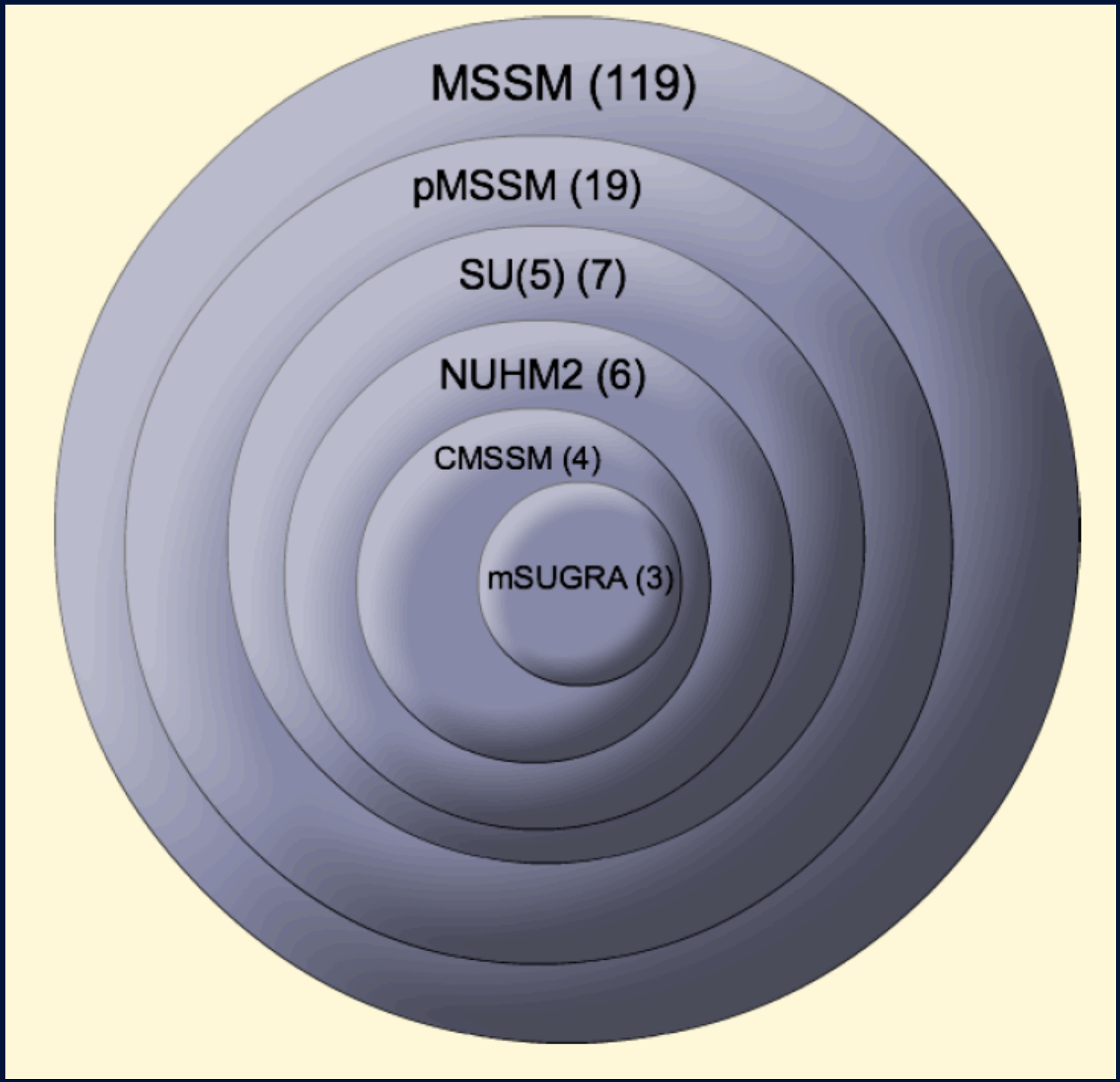
1. Which theories are preferred, given the data?
2. And which parameter sub-spaces within these theories?

The Minimal Supersymmetric SM



Alas Poor SUSY! I Knew Her…




MSSM (119)
pMSSM (19)
SU(5) (7)
NUHM2 (6)
CMSSM (4)
mSUGRA (3)

Statistical Challenges

- How should one incorporate previous information about a theory?
- How should one design analyses to test the predictions of a multi-parameter theory?
- How can one find the parameterization of a theory, such as the pMSSM, that best captures what can be learned about the theory, experimentally?
- How should one compare one theory versus another?

Statistical Challenges

The most fruitful way *to think* about an *inference* problem is Bayesian:

p(Data | Theory, Experimental)

to

p(Theory | Data)

The most convincing way *to validate* an inference procedure is frequentist.

Statistical Challenges

posterior density = prior × likelihood / prior

$$p(\text{Theory}\,|\,\text{Data}) = \pi(\text{Theory})\frac{\int p(\text{Data})\,|\,\text{Theory, Exp.})\;\pi(\text{Exp.})\,d\text{Exp.}}{p(\text{Data})}$$

Computational Challenges:

- Calculating the likelihood when one has ~billion events
- Calculating π(Exp.) (experimental uncertainties)
- Calculating π(Theory) (theoretical uncertainties)
- Calculating the integral

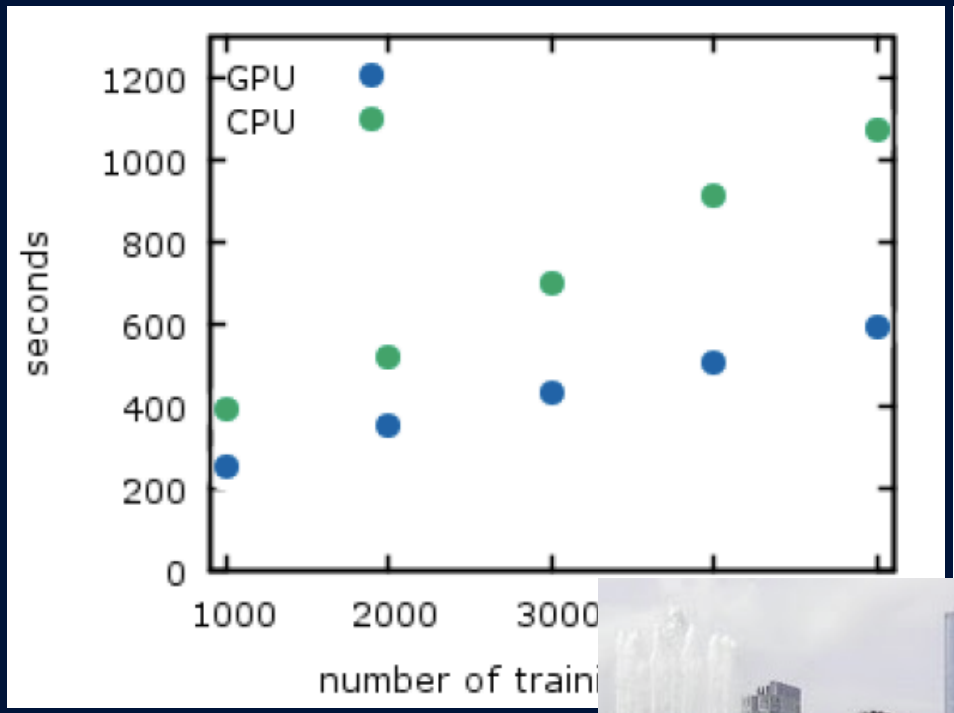
Statistical Challenges

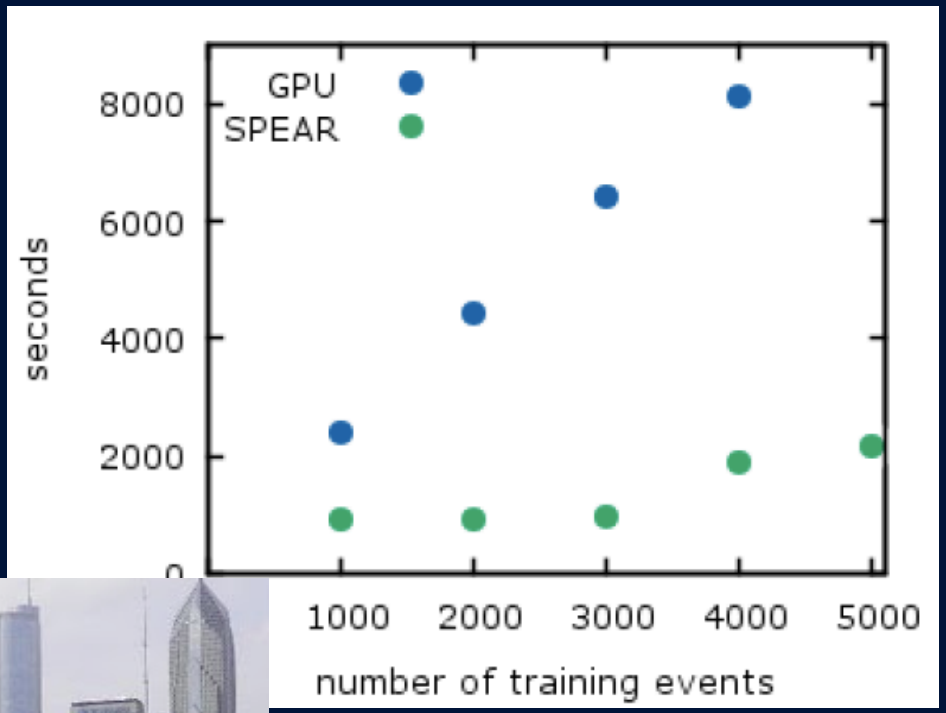
Statistical Challenges

Massively parallel computation using Graphics Processing Units

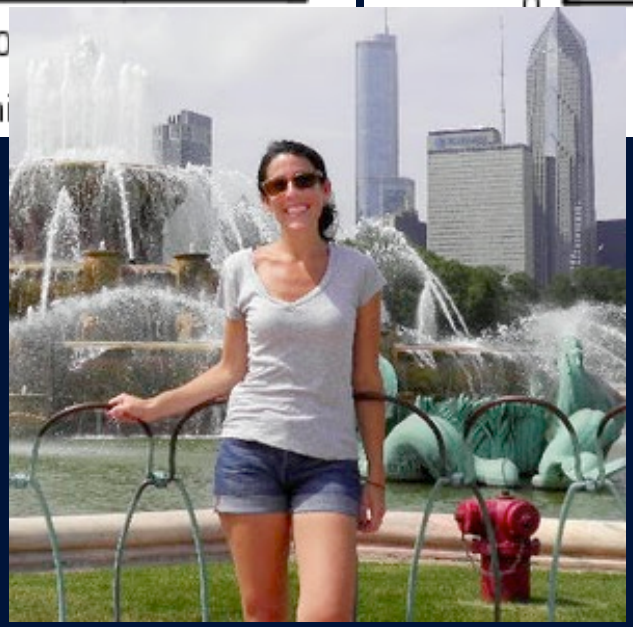
Image courtesy of NVIDIA Corp.

Statistical Challenges





Bayesian Neural Networks training on GPUs



Michelle Perry
PhD, April 2014

The future of statistics:

Think Bayesian!

Act frequentist!

Solve computationally!

"Prediction is very difficult, especially about the future" Niels Bohr The END