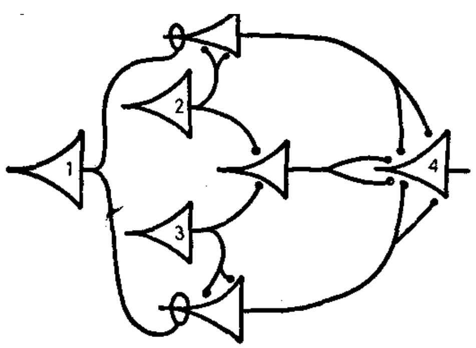
Machine learning and realism

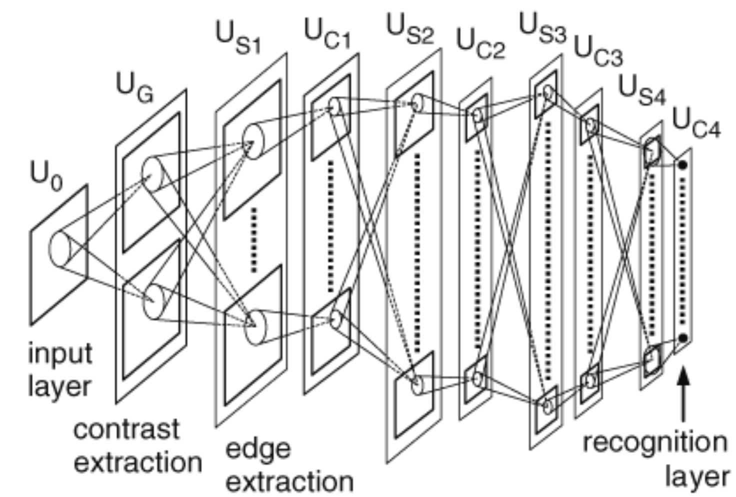


Mcculloch & Pitts (1943)

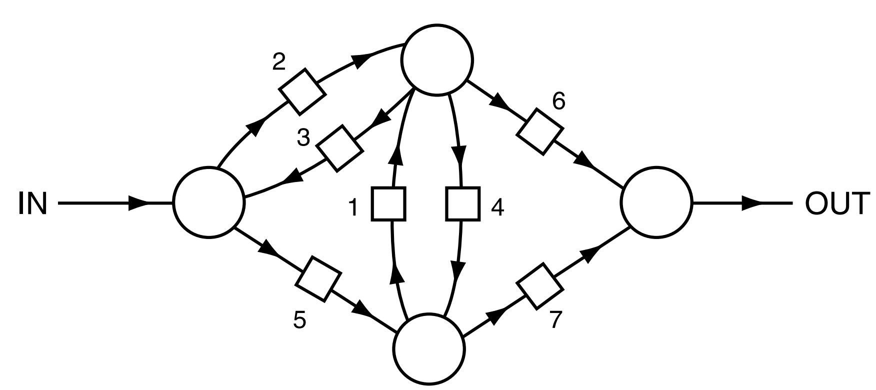
Ryan Reece (UCSC)

ryan.reece@cern.ch

with Nico Formanek (HLRS)



Fukushima (1980)



Turing (1948)

Outline

1. Reminders on the problem of induction and positivsm
2. Introduction to machine learning and neural nets
3. Examples of deep learning of in high-energy in physics
4. Clustering in feature space
5. Hypothesis tests and signficance
6. Natural kinds

The problem of induction

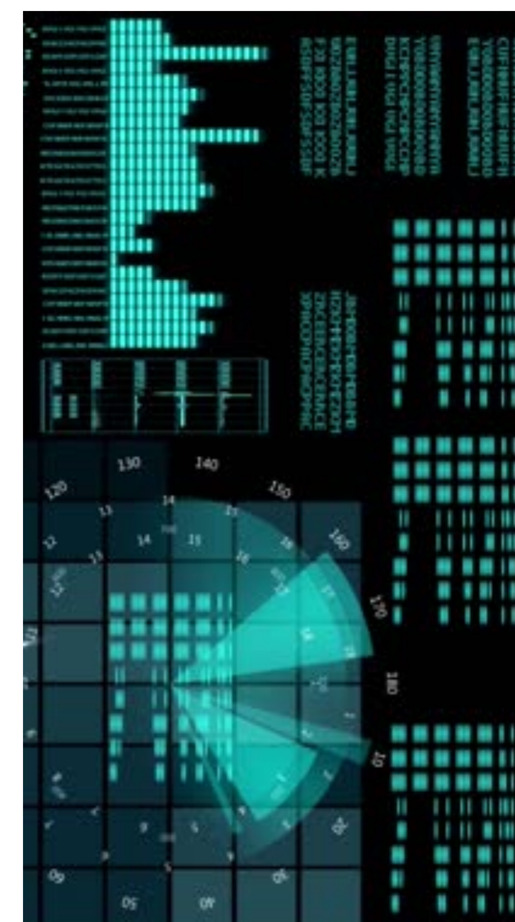
- We justify inferences with
  - ***deduction***: following by definition, logic, mathematics, "relations of ideas"
  - ***induction***: generalizing a universal based on limited data, drawing generalized conclusions from "matters of fact"
- Can we trust that "instances of which we have had no experience resemble those of which we have had experience"? (Hume, 1739)
- Induction is always susceptible possible "black swans".
- Russell's Thanksgiving turkey.


David Hume (1711-1776)

Sympathy for positivism

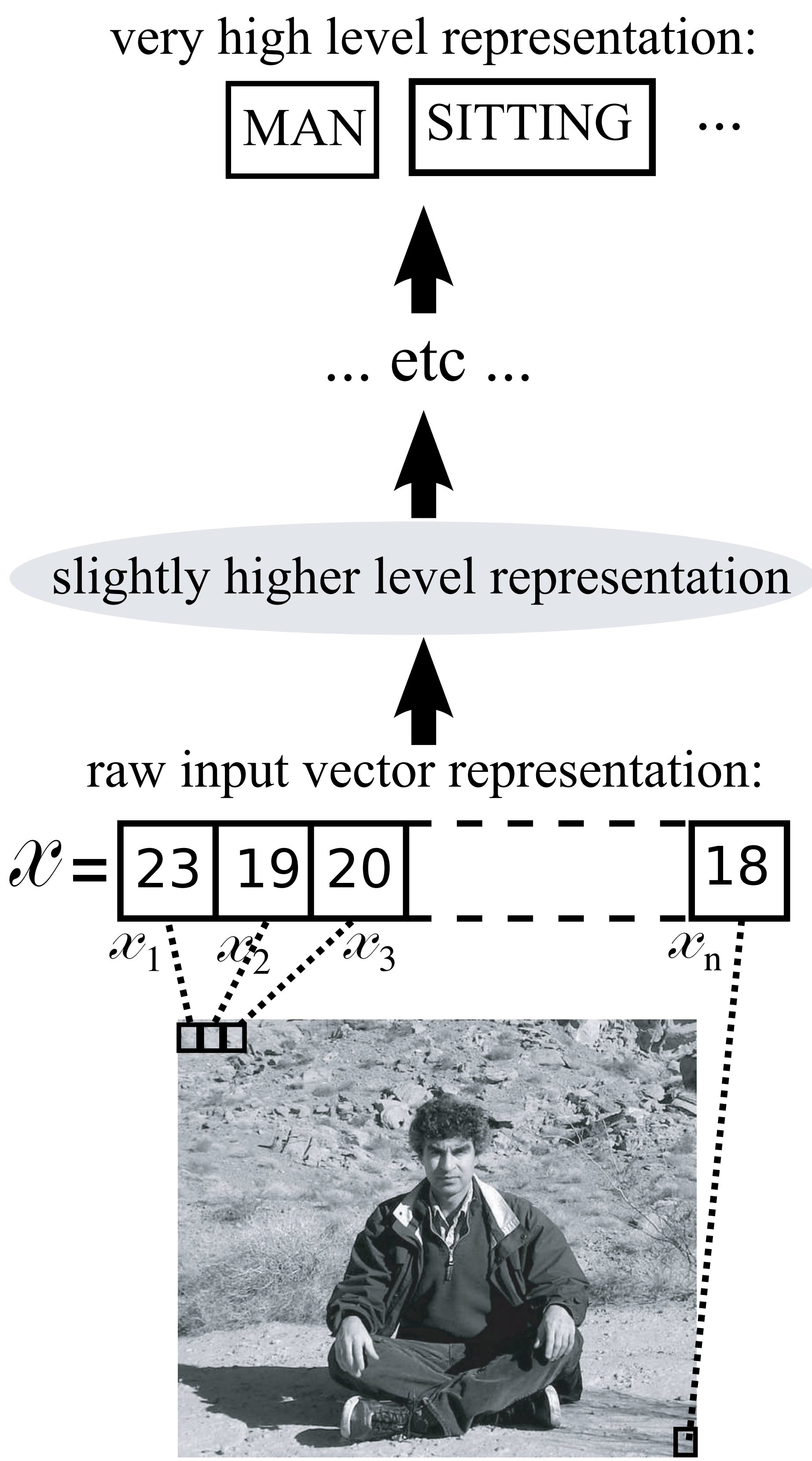
“Carnap never abandoned his belief that science and philosophy should be founded on a bedrock of logic… His unshakeable support of the primacy of both deductive and inductive logic made his position an increasingly isolated philosophy during the latter part of the Twentieth Century.

Meanwhile, technology has moved on… The agency of machines will steadily increase: think of robots, unmanned vehicles, industrial processes… so Carnap’s approach will become increasingly relevant, because highly sophisticated machine agents will certainly act on a basis of logic… Carnap’s faith in logic as the basis of one form of agency will have been vindicated.”



[MacFarlane, A. (2017). Rudolf Carnap (1891-1970). Philosophy Now, 118.]

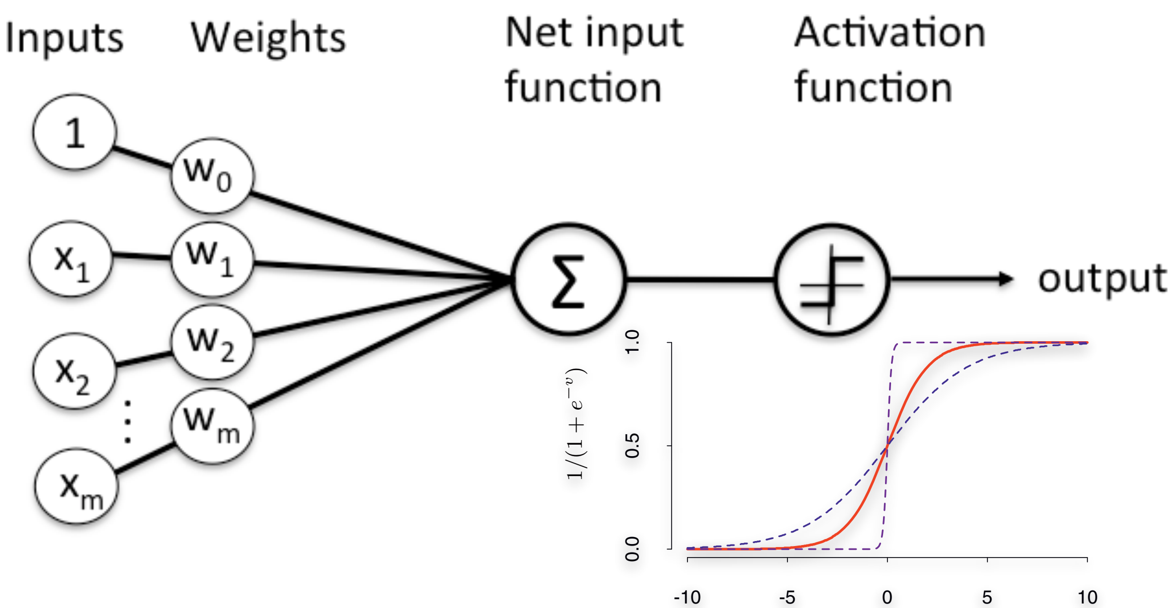
Machine learning

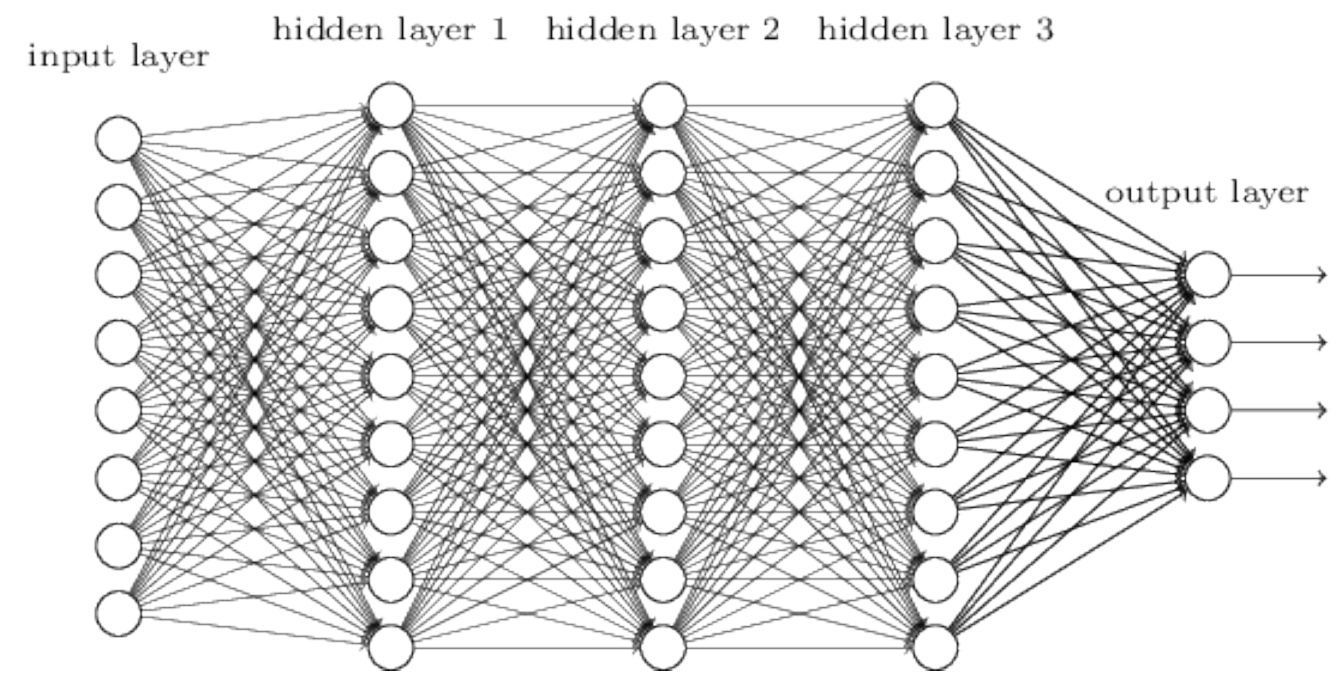


<u>The automation of</u>

- Pattern recognition
- Classification
- Data reduction
- Derivation of high-level representation
- Hypothesis testing

Neural Nets





$$y = f(x) = K\left(\sum_i w_i\, g_i(x)\right)$$

$$f(x) \overset{?}{\sim} w_i^g\, w_{ij}^h\, x_j$$

~ matrix multiplication except nonlinear activation

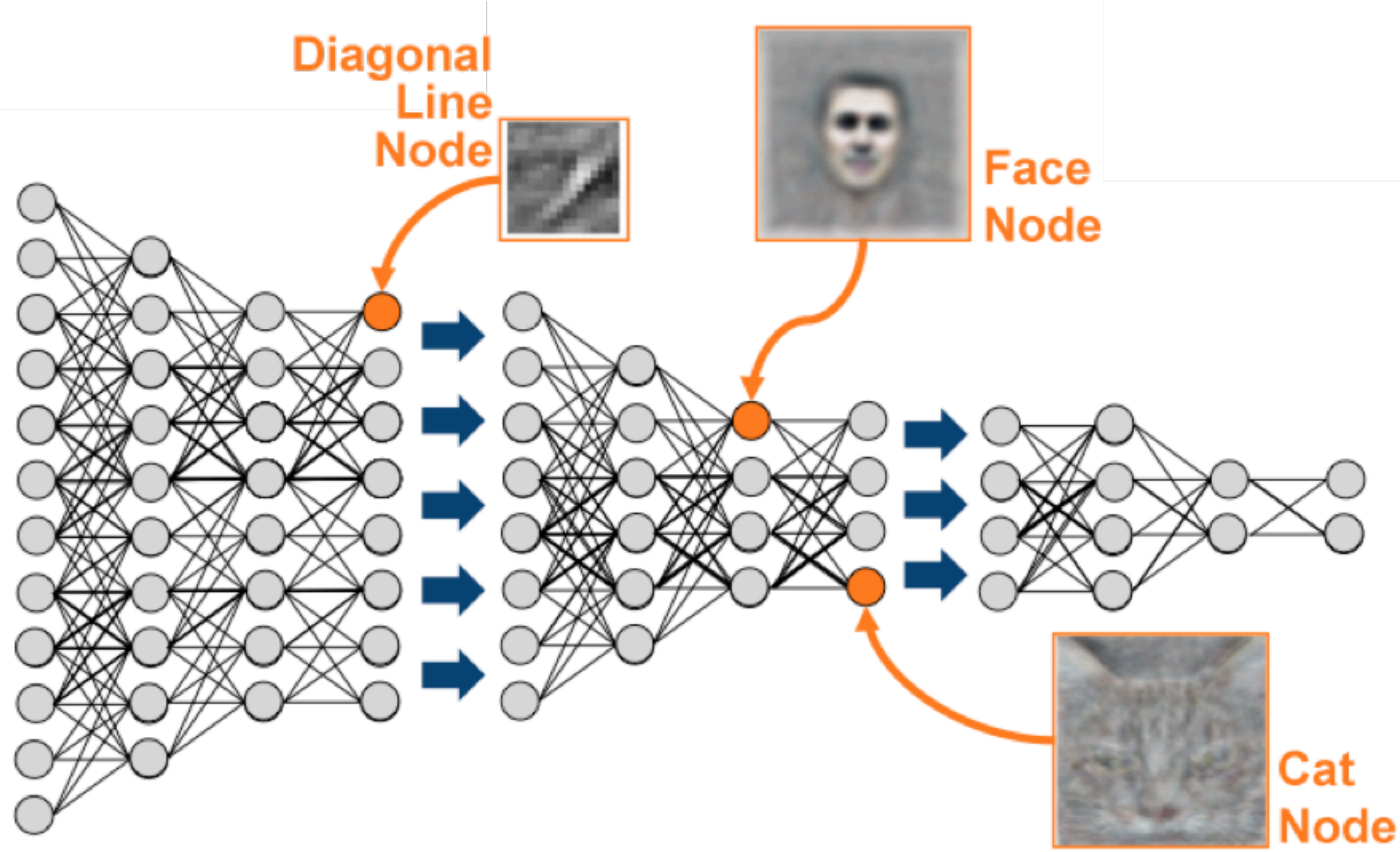
Neural nets have:

- input varaiables, x_i
- weights, w_{ij}
- activation function, $K_j(\cdot)$ (sigmoid, tanh, ...)
- output variables, y_j
- a *learning rule* to update the weights.
- a learning step is called an "*epoch*."
- Optimizing the weights is called "*training*."

- "*Deep*" networks have multiple hidden layers
- Can be used for *classification* or *regression*.
- Similar to other multivariate techniques, cutting on a classifier makes some acceptance blob in *x*-space.

Why go deep?

- Multiple layers allow for specialization and *feature extraction*.
- ***Now in "Deep Learning Renaissance"***



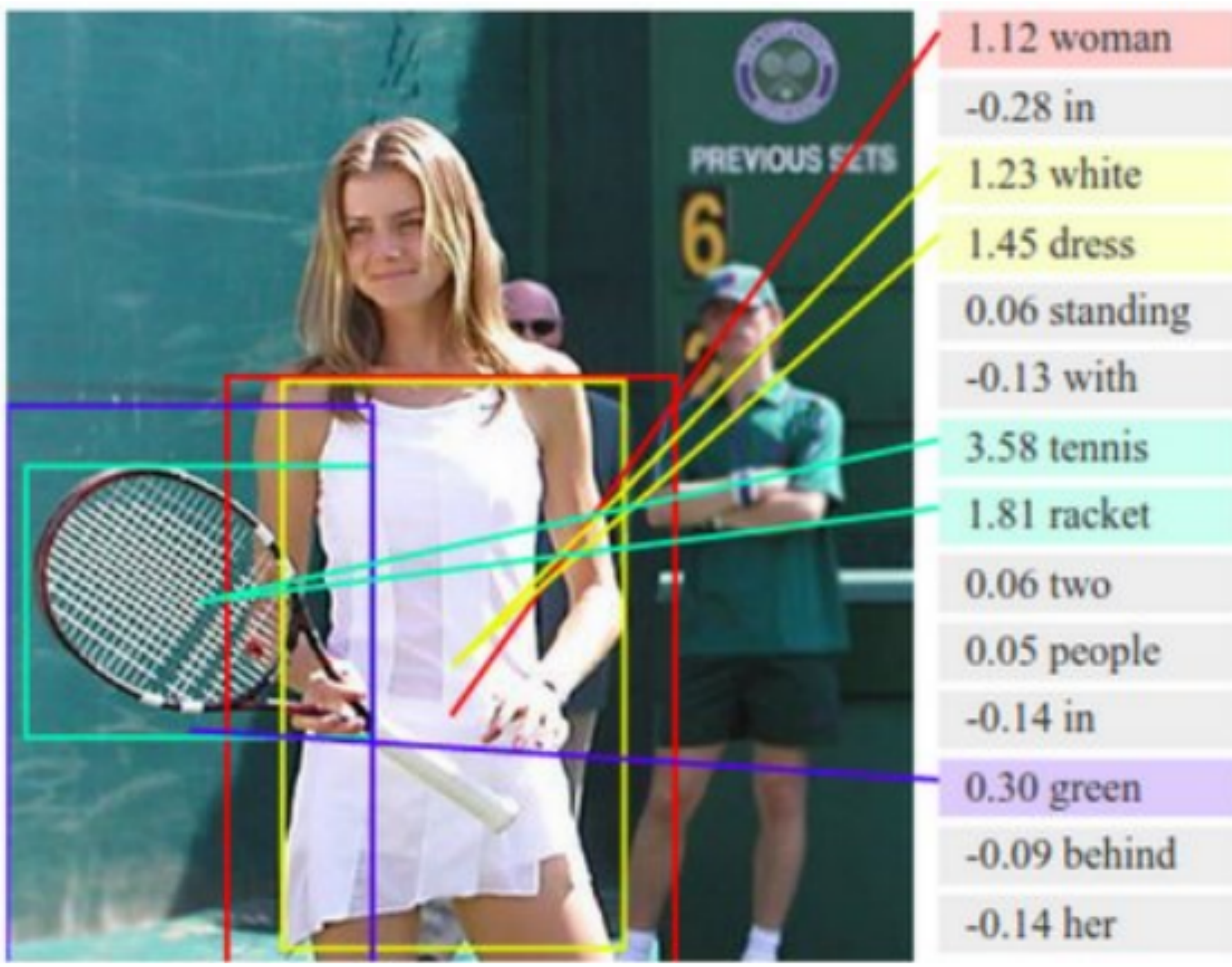
[Forbes/Google]

1. *<u>Better training</u>*: techniques and tools (e.g. SGD, layerwise training, smarter NN structures).
2. *<u>Better hardware</u>*: multicore, GPUs, bigger data centers, cloud computing, coming: neuromorphic computing.
3. *<u>More training</u>*: bigger datasets, search, the internet, open science.

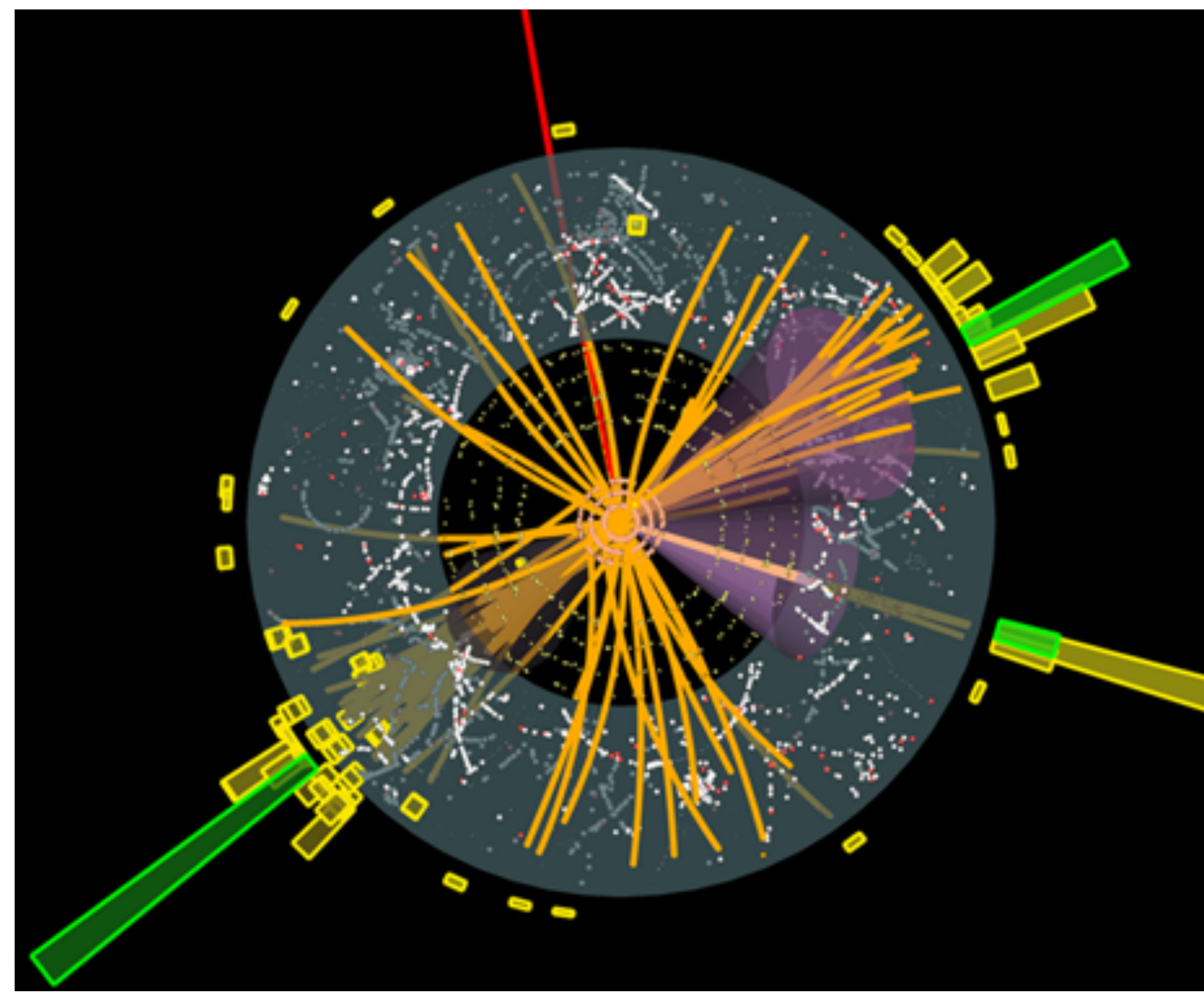
Learning to identify things

Is end-to-end learning from the raw data the future of particle physics reconstruction?

ImageNet competition example



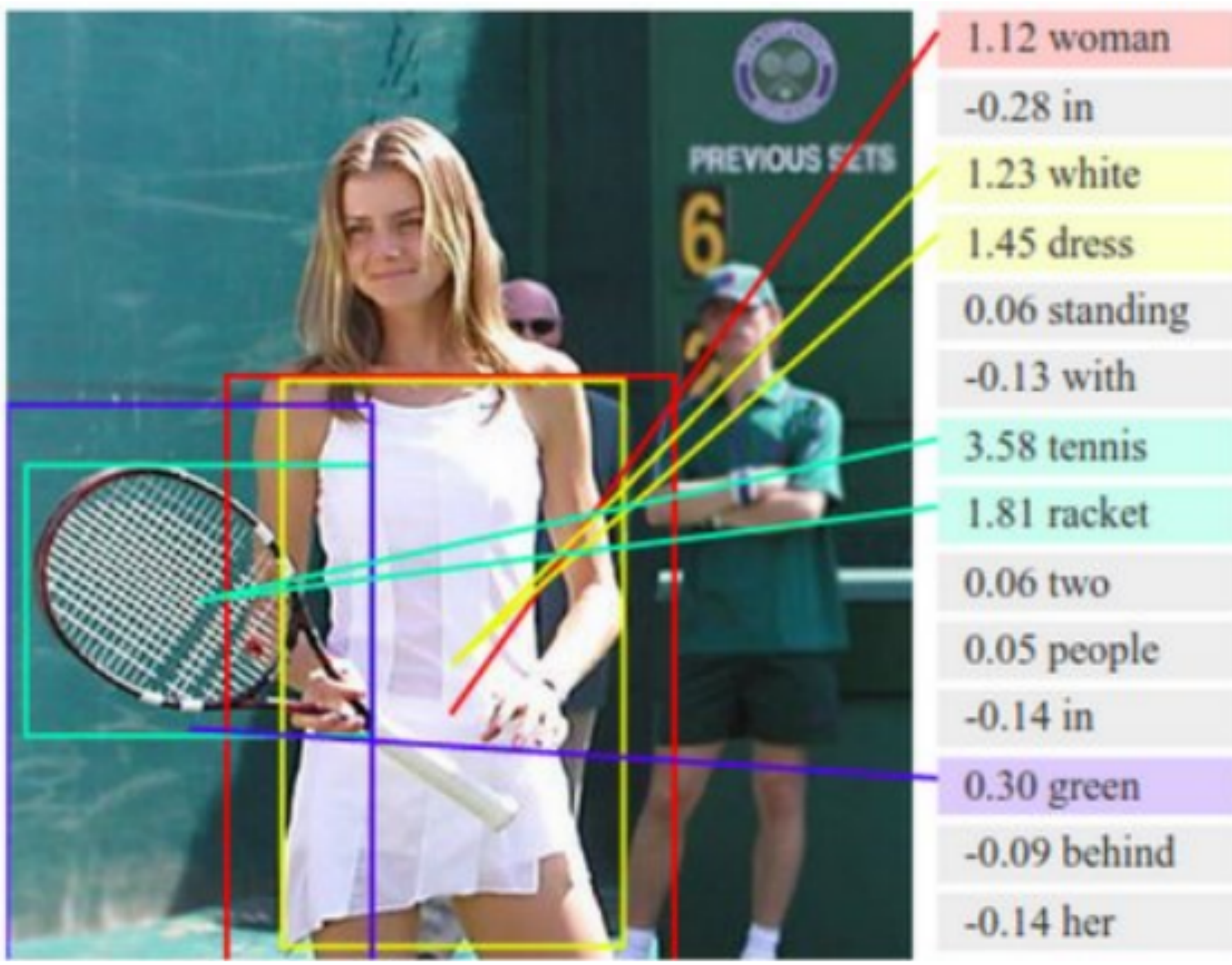
Future of ATLAS?



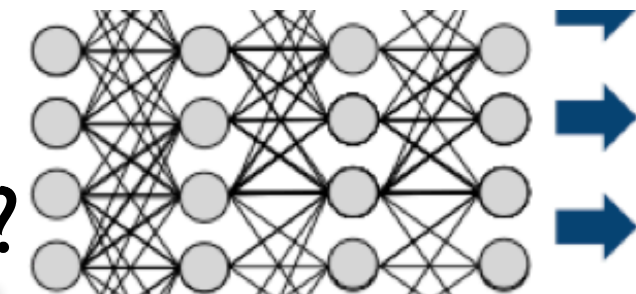
Learning to identify things

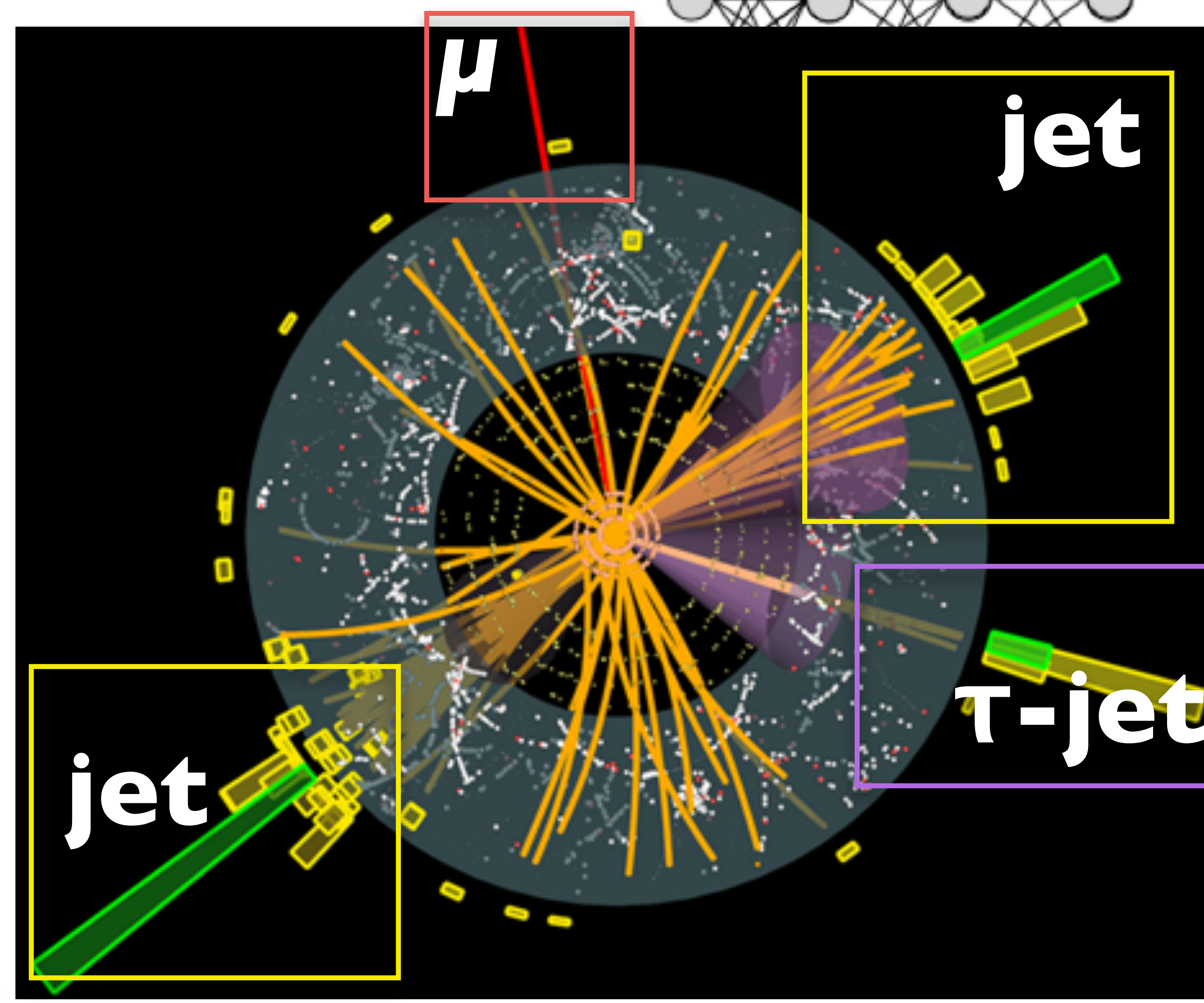
Is end-to-end learning from the raw data the future of particle physics reconstruction?

ImageNet competition example



Future of ATLAS?





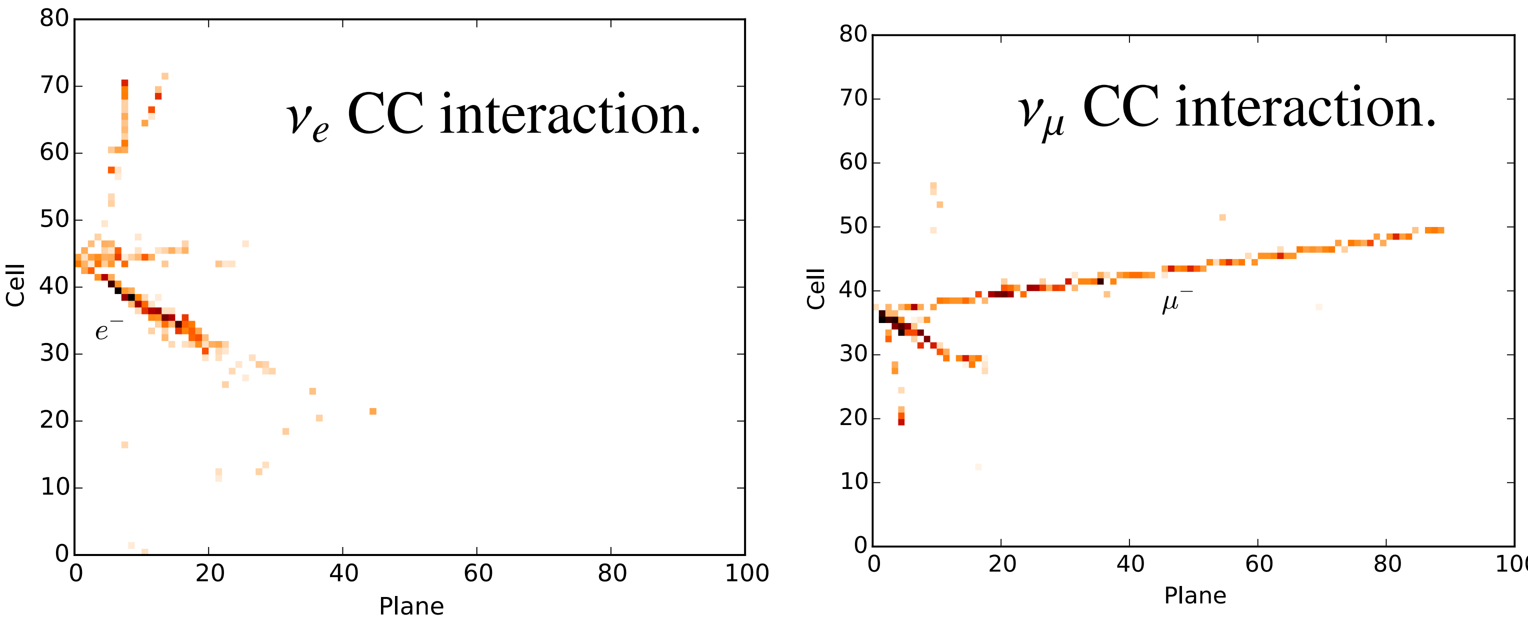
Deep Learning in HEP

- Deep learning does best with raw data and when there are unexploited features.
- raw channels → *tagging*
- basic kinematics → *features*

- Baldi *et al.* (2014). Searching for Exotic Particles in High-Energy Physics with Deep Learning. [1402.4735]
- 

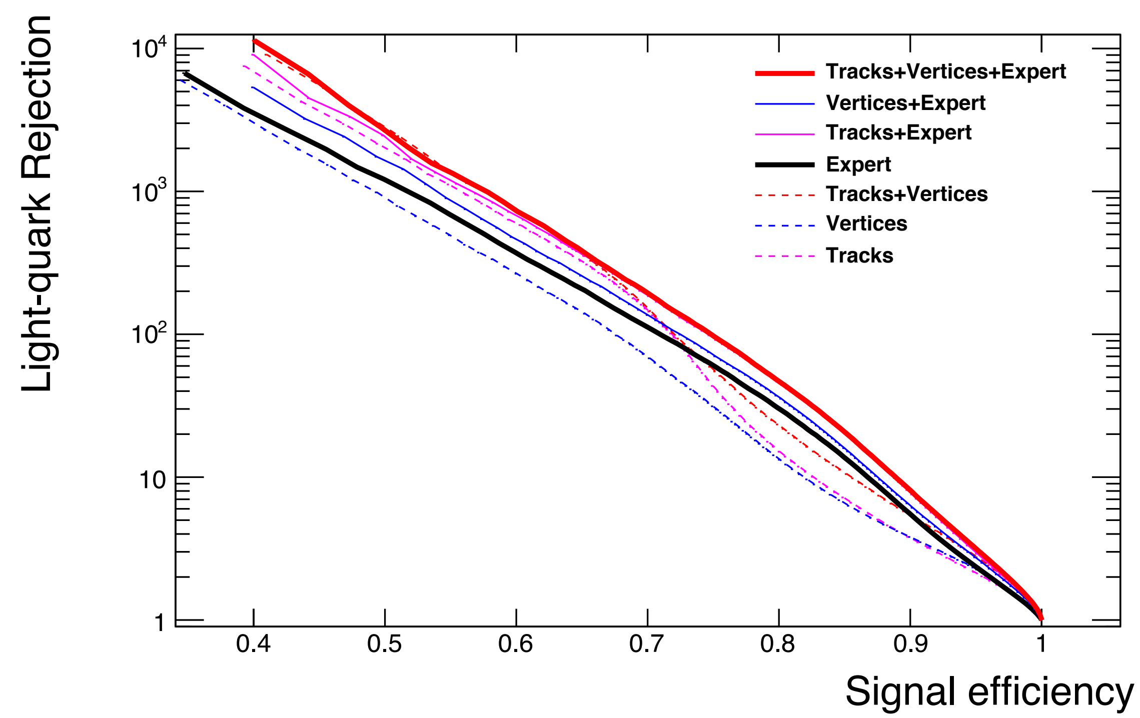
Baldi *et al.* (2015). Enhanced Higgs to $\tau^+\tau^-$ Search with Deep Learning. [1410.3469]

- Aurisano *et al.* (2016). A Convolutional Neural Network Neutrino Event Classifier. [1604.01444]



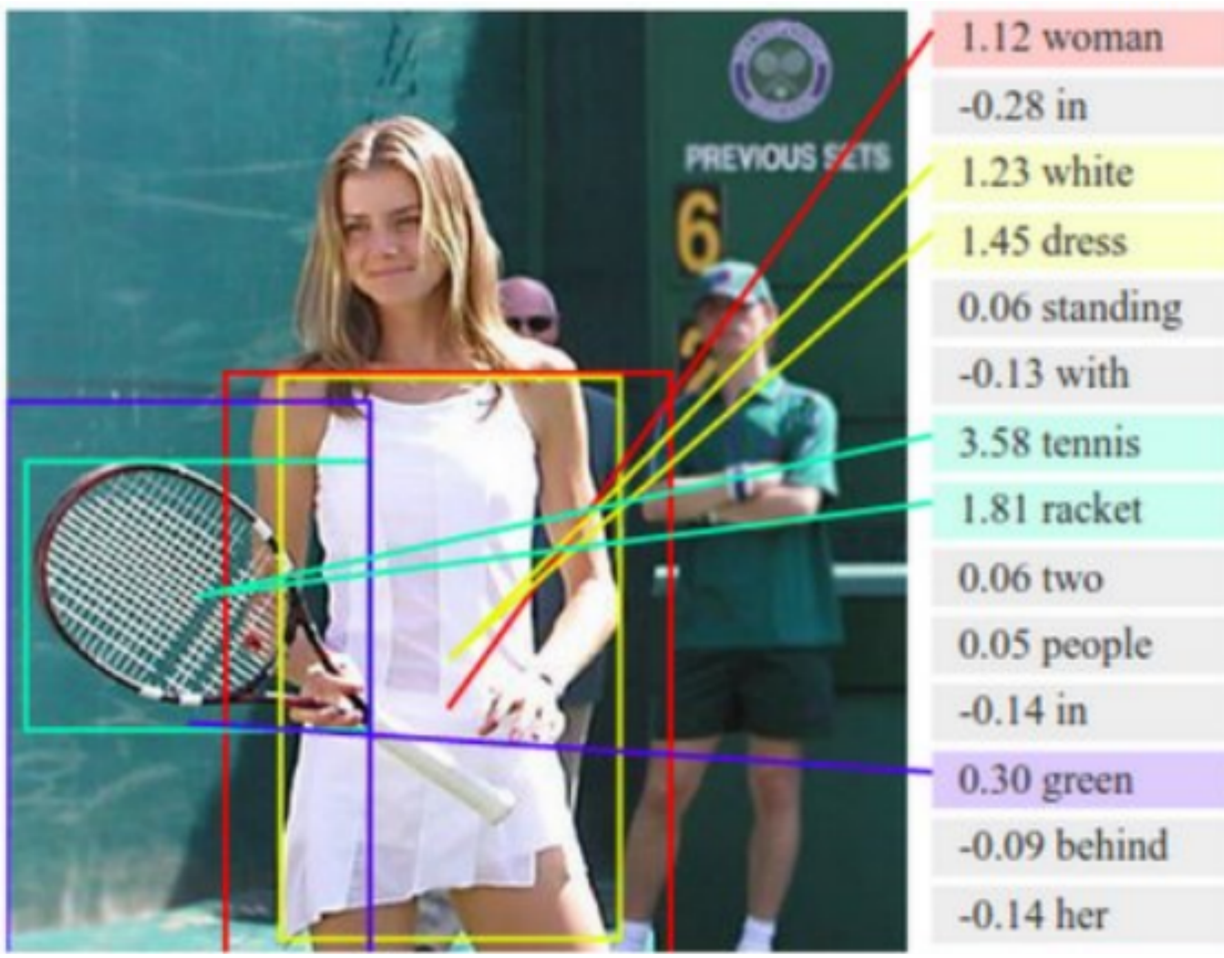
out performs NOvA's conventional reconstruction

- Guest et al. (2016). Jet Flavor Classification in HEP with DNNs. [1607.0863]

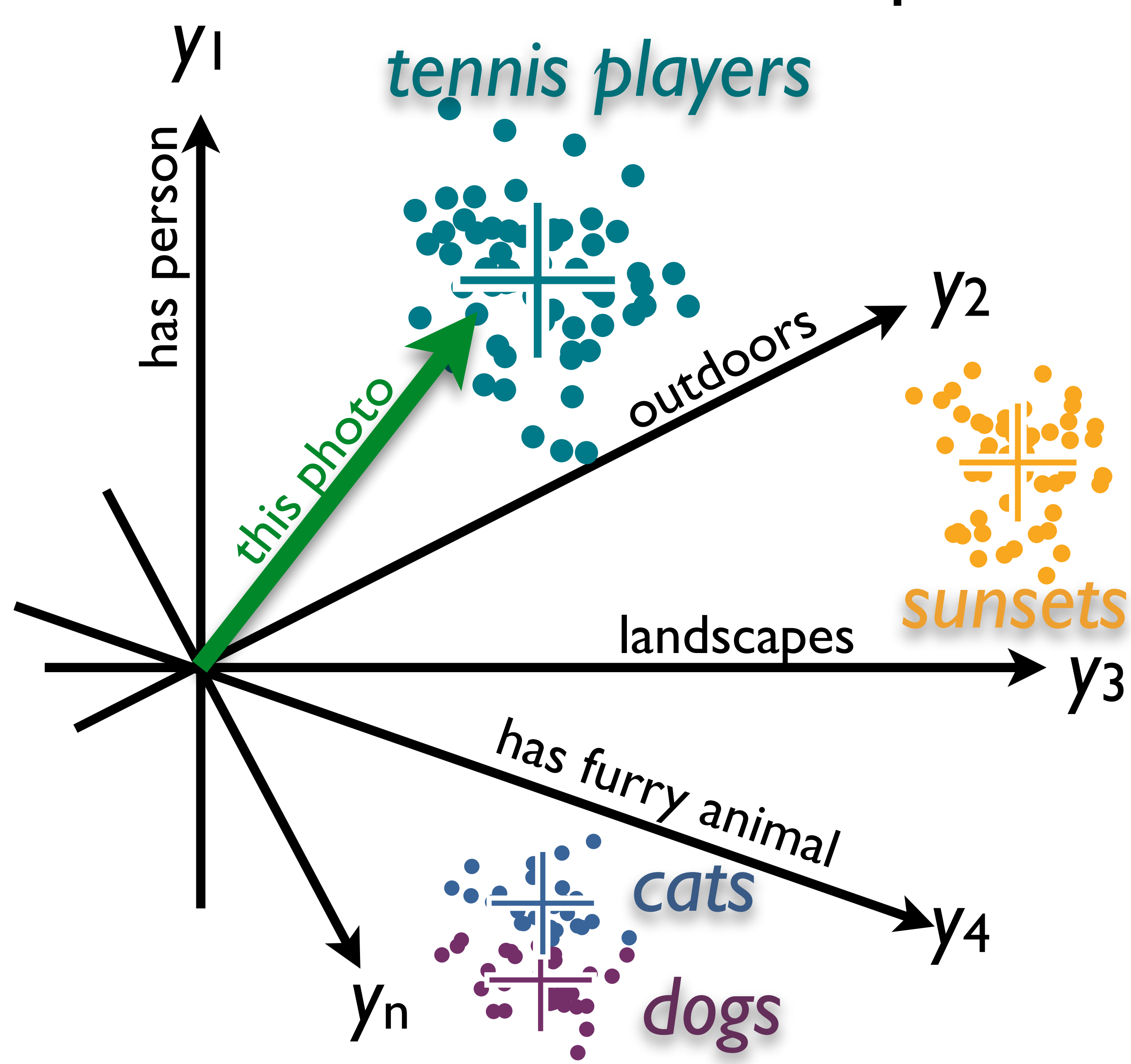


Unsupervised learning

Raw input features



clusters in feature space

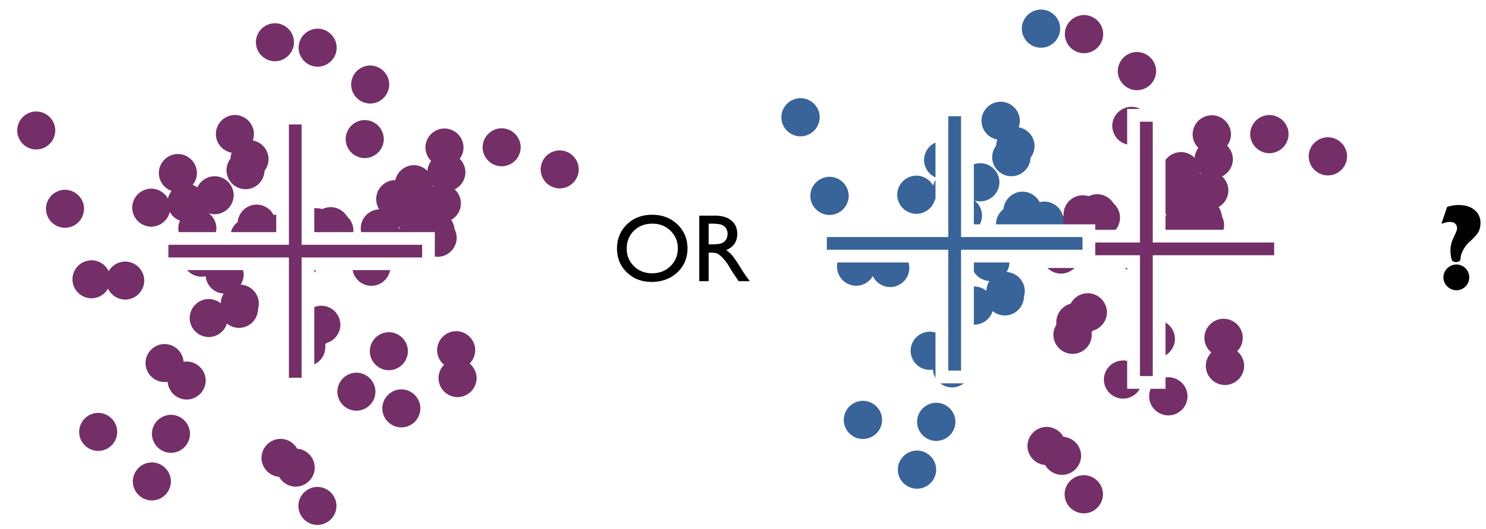


Vector Quantization (VQ)

"The typical vectors we use to represent concepts like images have about 4,000 dimensions," he says. "So, basically, it is a list of 4,000 numbers that characterises everything about an image." Vectors can describe an image, a piece of text or human interests. Reduced to a number, it's easy for computers to search and compare. If the interests of a person, represented by a vector, match the vector of an image, the person will likely enjoy the image. "***Basically, it reduces reasoning to geometry***," he says.

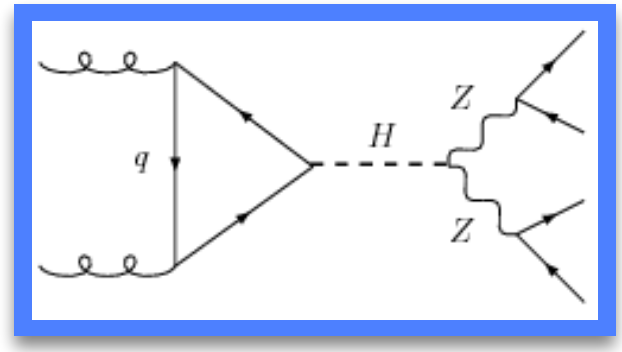
— Jeff Goodell of *Rolling Stone* quoting ML researcher Yann LeCunn at NYU and Facebook

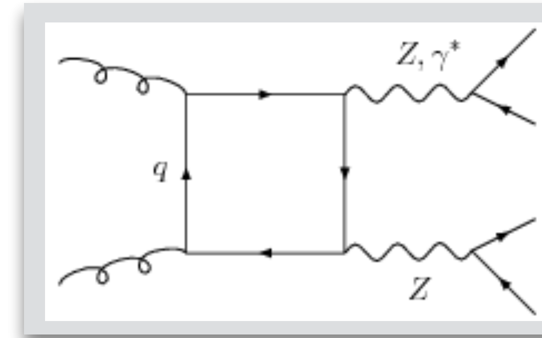
Cluster sensitivity



- Any experiment has empirical limits.
  - To discover structure within clusters (split them) may require better measurement precision, and/or better training samples, and/or exposure to new features (*y*-dimensions).
- Clustering is task, algorithm, or *model dependent* (in the case of maximum likelihood fitting).
  - Not everyone agrees with me: "It seems to me that a misguided desire for uniqueness" (Hennig, 2015)

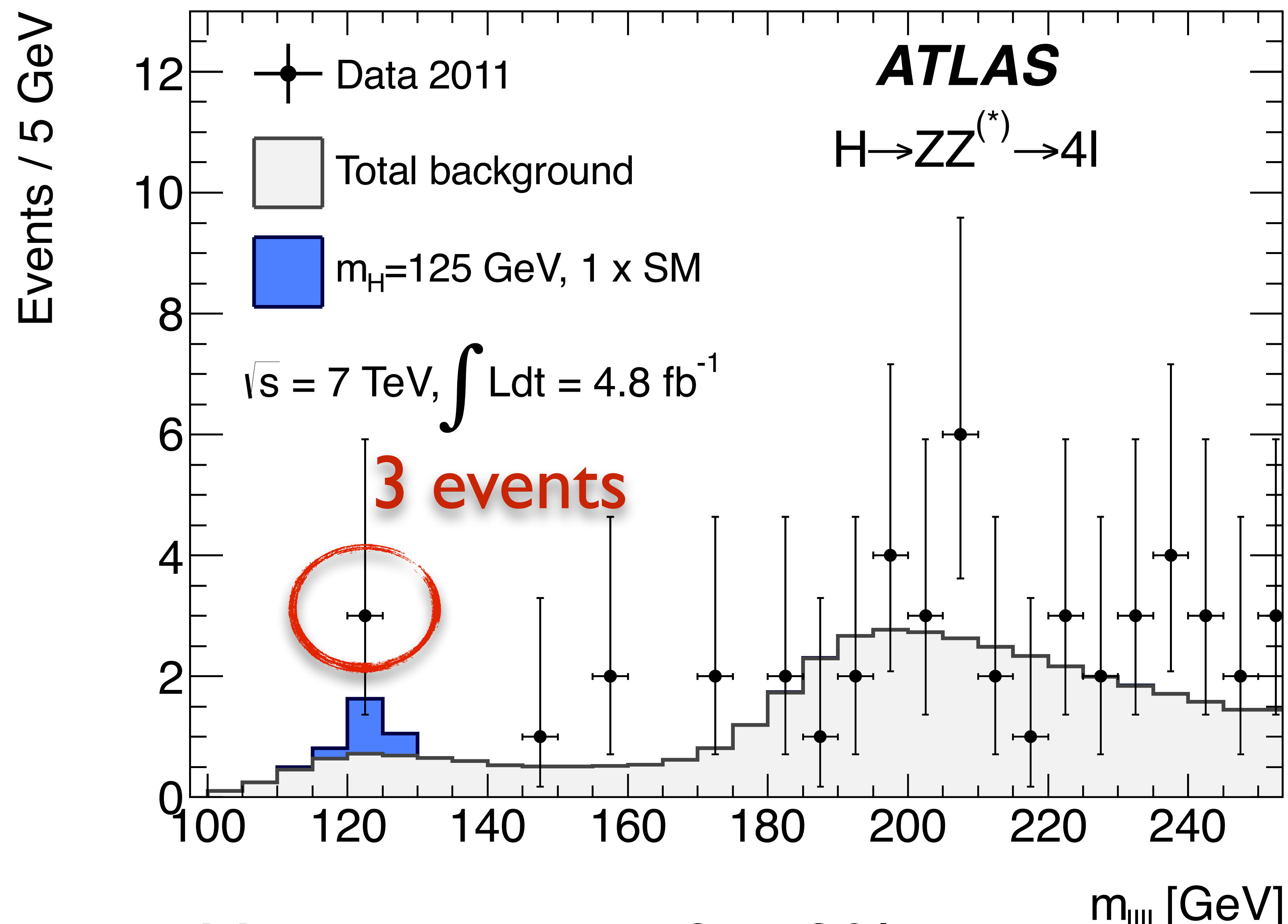
Is this significant?





Statistical and philosophical question:

- How can we be precise and rigorous about how confident we are that a model is wrong?
  - ***Hypothesis testing***



Has a local p_0 of $\approx$ 2%

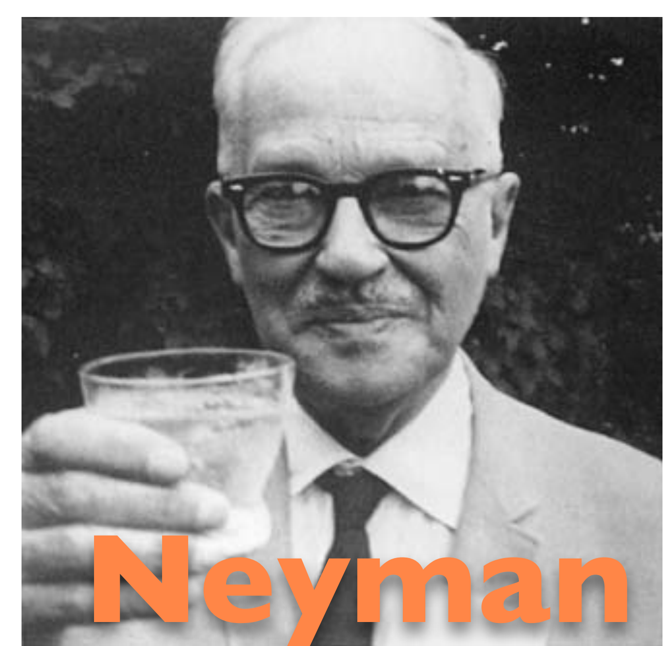
[arxiv:1207.0319]

Hypothesis testing

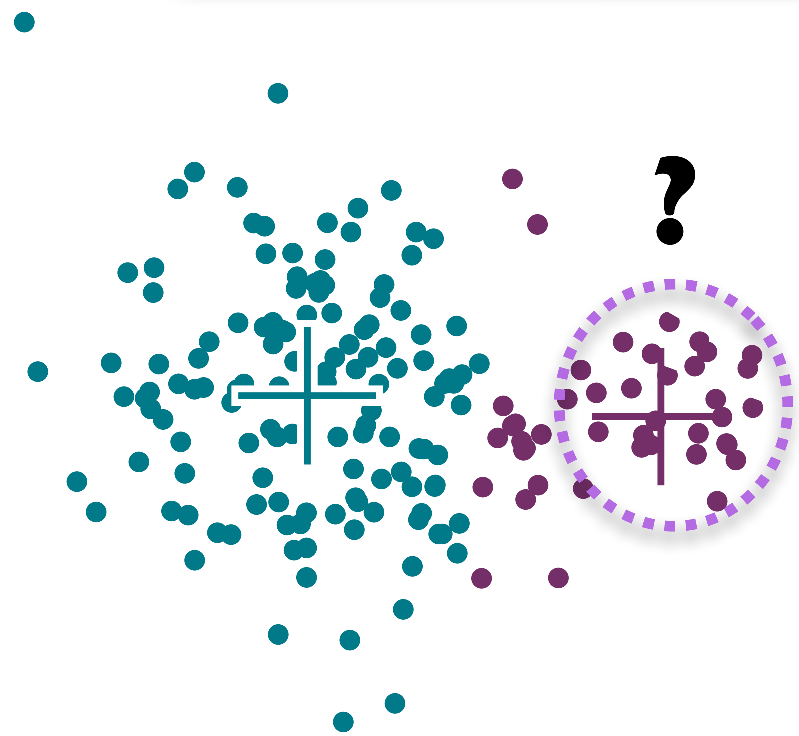
Table of error types		Null hypothesis (H_0) is	
		Valid/True	**Invalid/False**
Judgment of Null Hypothesis (H_0)	**Reject**	Type I error (False Positive, α)	Correct inference (True Positive, 1-β)
	Fail to reject	Correct inference (True Negative, 1-α)	Type II error (False Negative, β)
Type I = True H_0 but reject it (False Positive) **Type II = False H_0 but fail to reject it (False Negative)**			

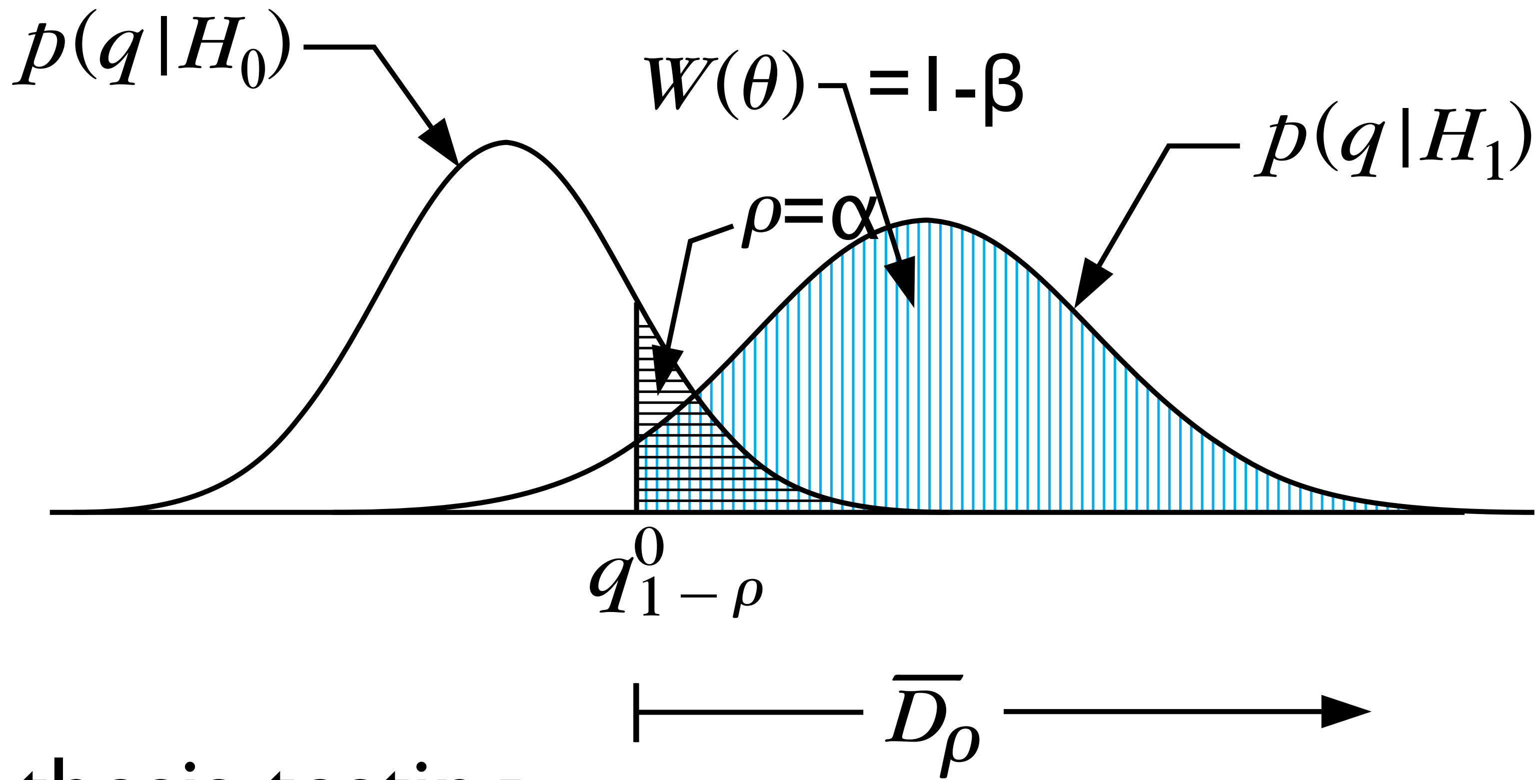
- Want to maximize power for a fixed false positive rate
- Particle physics has a tradition of claiming discovery at $5\sigma \Rightarrow p_0 = 2.9\times10^{-7}$ = 1 in 3.5 million
- Makes exclusions with p_0 = 5%, (95% CL "coverage").
- Neyman-Pearson lemma (1933): the most powerful test for fixed α is the likelihood ratio:

$$\frac{L(x|H_0)}{L(x|H_1)} > k_\alpha$$



Cluster discovery





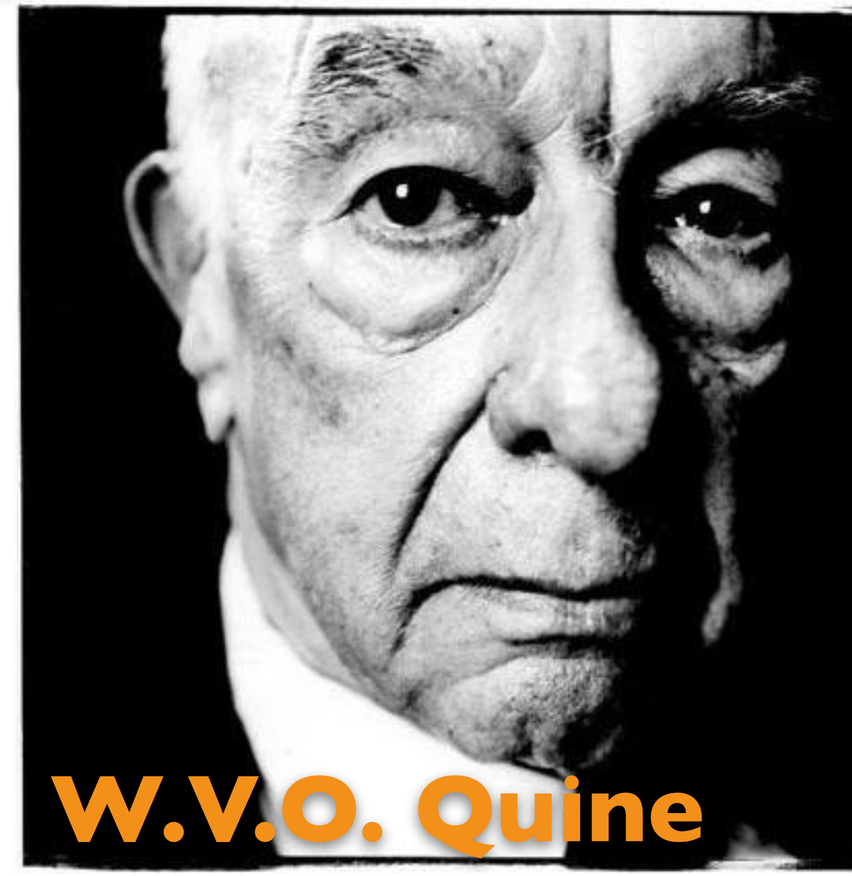
- Cluster validation via hypothesis testing
  ρ = α = p-value for H_0 (5% for 95% CL)

- Neyman-Pearson theory of confidence intervals
  $q \sim \; t_{\mathrm{NP}}(\boldsymbol{x}) = f(\boldsymbol{x}|H_1)\,/\,f(\boldsymbol{x}|H_0)$

- Can give frequentist confidence that:
  if the kind exists (H_1 is true), *then it fits the data better,*
  if H_0 *is true, then the observed data is rare at some confidence level.*

Natural kinds

- A natural kind is a natural (objective) grouping, as opposed to an artificial (constructed) one.
- "The human experience that the reality outside the observer's control seems to make certain distinctions between categories *inevitable*" (Hennig, 2015).
- They carve nature at its joints.
- E.g. atomic elements, 4 DNA bases, …
- Complex/evolving species are more problematic (the species problem).



Carbon ($_6$C)

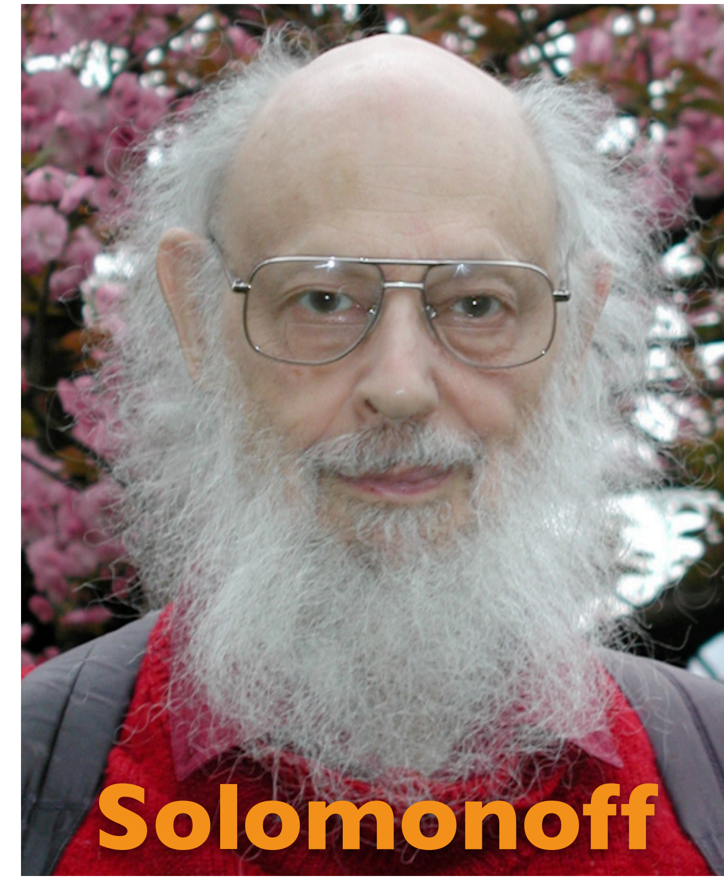
Gold ($_{79}$Au)

Coyote

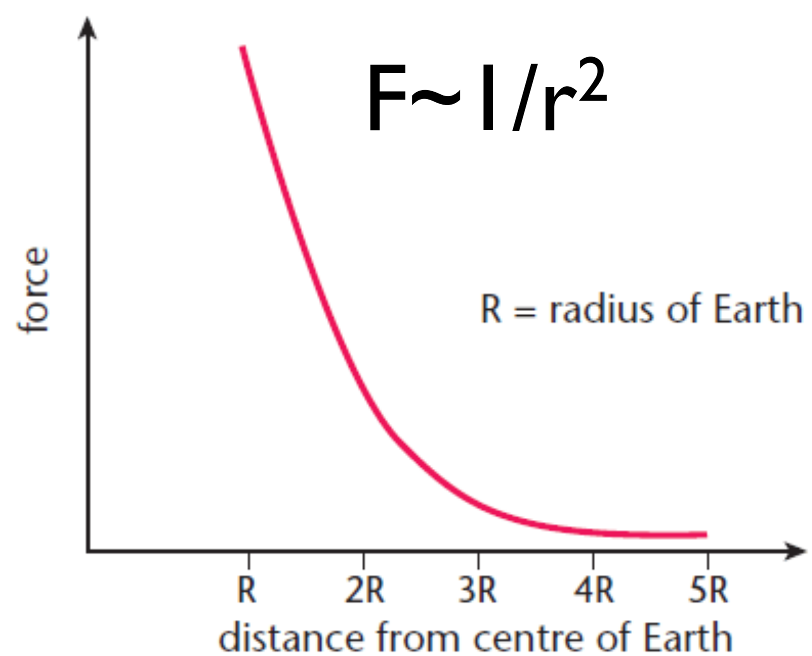
Coywolf

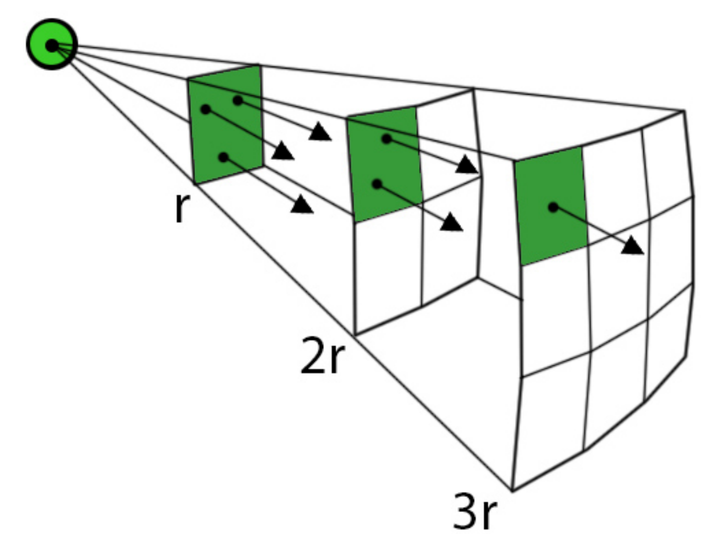
Convergence

- "In the physical sciences, the single "best" theory, is usually much better than the others, so selecting the single best law is not much different from ALP. In the complex sciences—such as sociology, psychology and geology—the tenth best theory may be not far behind the best, and ALP's weighing of all of them can be considerably different from choosing the single best one." [Solomonoff, R. J. (1996).]



- E.g. $1/r^2$ Newtonian gravity
- Effectiveness / abduction
- The "right" approximation





Summary

- Phrasing the scientific realism debate in machine learning terms can sharpen the disucssion.

Data → Features → Clustering → Natural Kinds **?**

- Antirealist might say:
  - Science finds empirically adequate patterns and regularities.
  - No need to think they are objectively real.
  - Machine learning will take this pattern finding out of human hands. "The End of Theory" [Anderson, C. (2008). *Wired.*]

- Realist retort:
  - Machine learning makes manifest that we can classify the world into kinds, arguably (nearly) independent of human convention.
  - ML is not the end of theory, in fact it is becoming one of our most powerful tools for discovering natural laws.

References

Alai, M. (2004). AI, Scientific Discovery and Realism. *Minds and Machines*, *14*, 21–42.

Aurisano, A. *et al.* (2016). A convolutional neural network neutrino event classifier. *Journal of Instrumentation*, *11*, P09001. https://arxiv.org/abs/1604.01444

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning*, *2*, 1–127.

Carnap, R. (1945a). On Inductive Logic. *Philosophy of Science*, *12*, 72–97.

———. (1945b). Two Concepts of Probability. *The Journal of Philosophy*, *5*, 513–532.

———. (1947a). On the Application of Inductive Logic. *Philosophy and Phenomenological Research*, *8*, 133–148.

———. (1947b). Probability as a Guide in Life. *The Journal of Philosophy*, *44*, 141–148.

———. (1950). Empiricism, Semantics, and Ontology. *Revue Internationale de Philosophie*, *4*, 20–40.

Cowan, G. (1998). *Statistical Data Analysis*. Oxford: Clarendon Press.

———. (2016). Statistics. In C. Patrignani *et al.* (Particle Data Group), *Chinese Physics C*, *40*, 100001. http://pdg.lbl.gov/2016/reviews/rpp2016-rev-statistics.pdf.

Cowan, G., Cranmer, K., Gross, E., & Vitells, O. (2011). Asymptotic formulae for likelihood-based tests of new physics. *European Physical Journal C*, *71*, 1544. https://arxiv.org/abs/1007.1727

Cranmer, K. (2015). Practical Statistics for the LHC. https://arxiv.org/abs/1503.07622

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459.

References

Guest, D. *et al.* (2016). Jet flavor classification in high-energy physics with deep neural networks. *Physical Review D*, *94*, 112002. https://arxiv.org/abs/1607.08633

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic.* Cambridge: Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters*, *64*, 53–62. https://arxiv.org/abs/1502.02555

Huang, C., Loy, C. C., & Tang, X. (2016). Unsupervised learning of discriminative attributes and visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5175–5184).

Hume, D. (2007). *An Enquiry Concerning Human Understanding.* (P. Millican, Ed.). Oxford: Oxford University Press. (Originally published in 1748).

James, F. (2006). *Statistical Methods in Experimental Particle Physics.* World Scientific.

Kendall, M. G. (1946). *The Advanced Theory of Statistics, Vol.II.* London: Charles Griffin & Company.

Korb, K. B. (2004). Introduction: Machine learning as philosophy of science. *Minds and Machines*, *14*, 433–440.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

MacFarlane, A. (2017). Rudolf Carnap (1891-1970). *Philosophy Now, 118.* https://philosophynow.org/issues/118/Rudolf_Carnap_1891-1970

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge.* Chicago: Chicago University Press.

References

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth.* Routledge.

———. (2016). The Realist Turn in the Philosophy of Science. http://philsci-archive.pitt.edu/12440/

Quine, W. V. O. (1969). Natural kinds. In *Ontological Relativity and Other Essays* (pp. 114–138). New York: Columbia University Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://arxiv.org/abs/1404.7828

Sellars, W. (1964). Induction as Vindication. *Philosophy of Science*, *31*, 197–231.

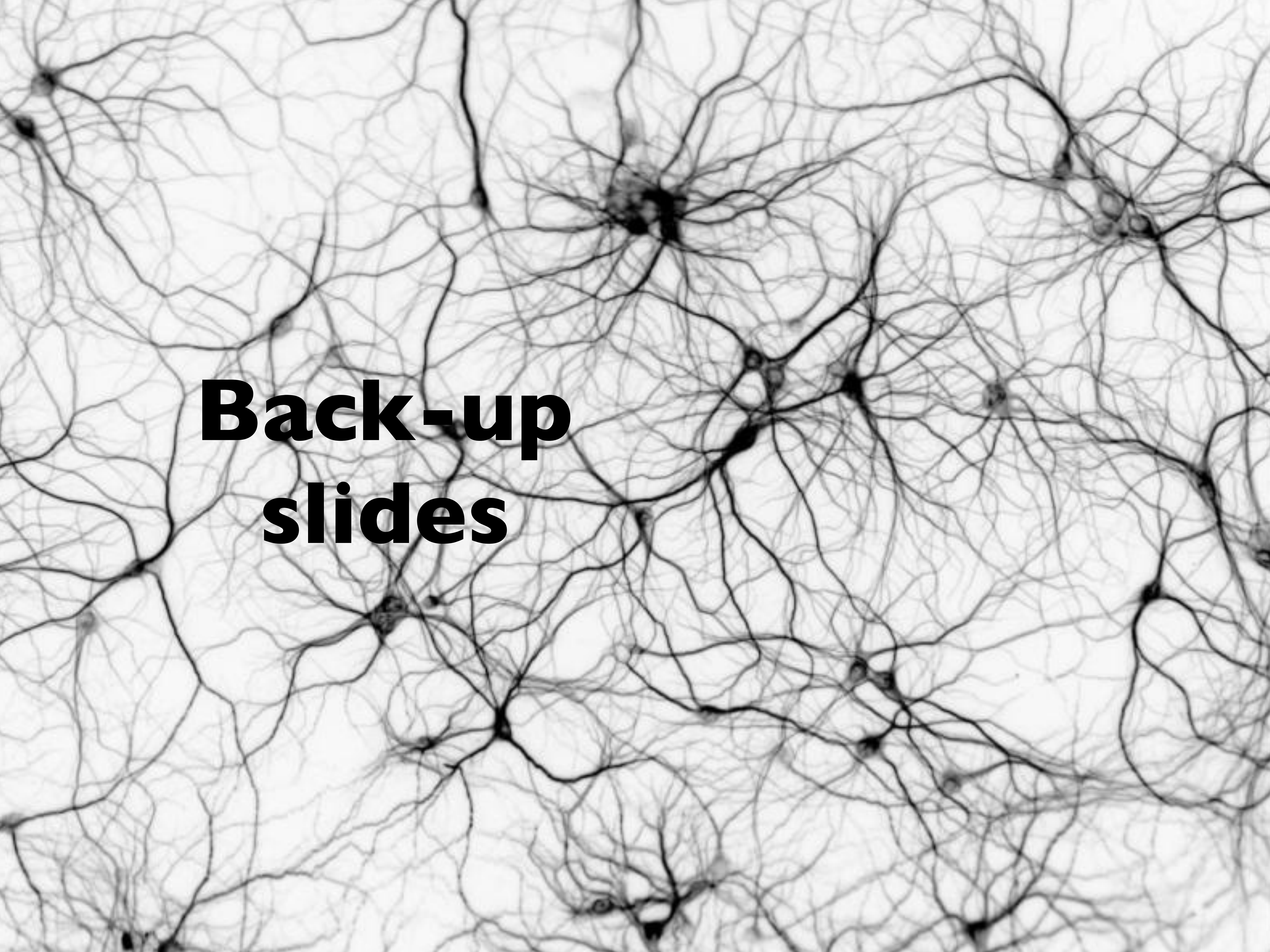
Sider, T. (2011). *Writing the Book of the World.* Oxford: Oxford University Press.

Solomonoff, R. J. (1996). Does Algorithmic Probability Solve the Problem of Induction? In *Information, Statistics and Induction in Science: Proceedings of the Conference, ISIS '96.* World Scientific. http://raysolomonoff.com/publications/isis96.pdf

———. (1997). The Discovery of Algorithmic Probability. *Journal of Computer and System Sciences*, *55*, 73–88. http://raysolomonoff.com/publications/barc97.pdf

Theodoridis, S. (2009). *Pattern Recognition.* London: Elsevier.

Back-up slides



Mayo's error statistics

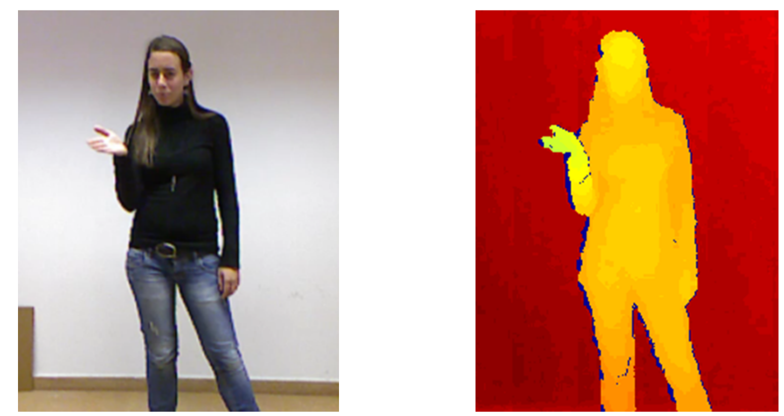
"The challenge, if one is not to abandon the view that science is characterized by rational methods of hypothesis appraisal, is either to develop more adequate models of inductive inference, or else to find some account of scientific rationality."

— Deborah Mayo (1996)
Error and the Growth of Experimental Knowledge

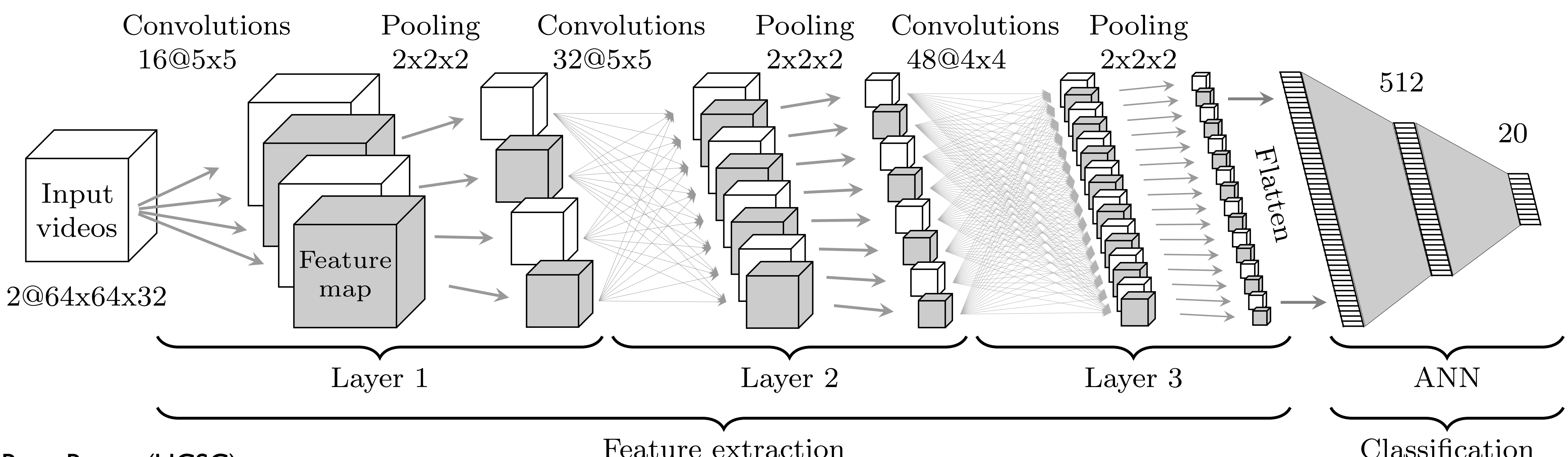
Examples of CNNs

- In 1990s, Yann LeCun pioneered Convolutional Neural Nets (CNN) and used them for Optical Character Recognition.
- Inspired by animal cortex.
- Now it is standard in image recognition and captioning, NLP, computer vision, etc.

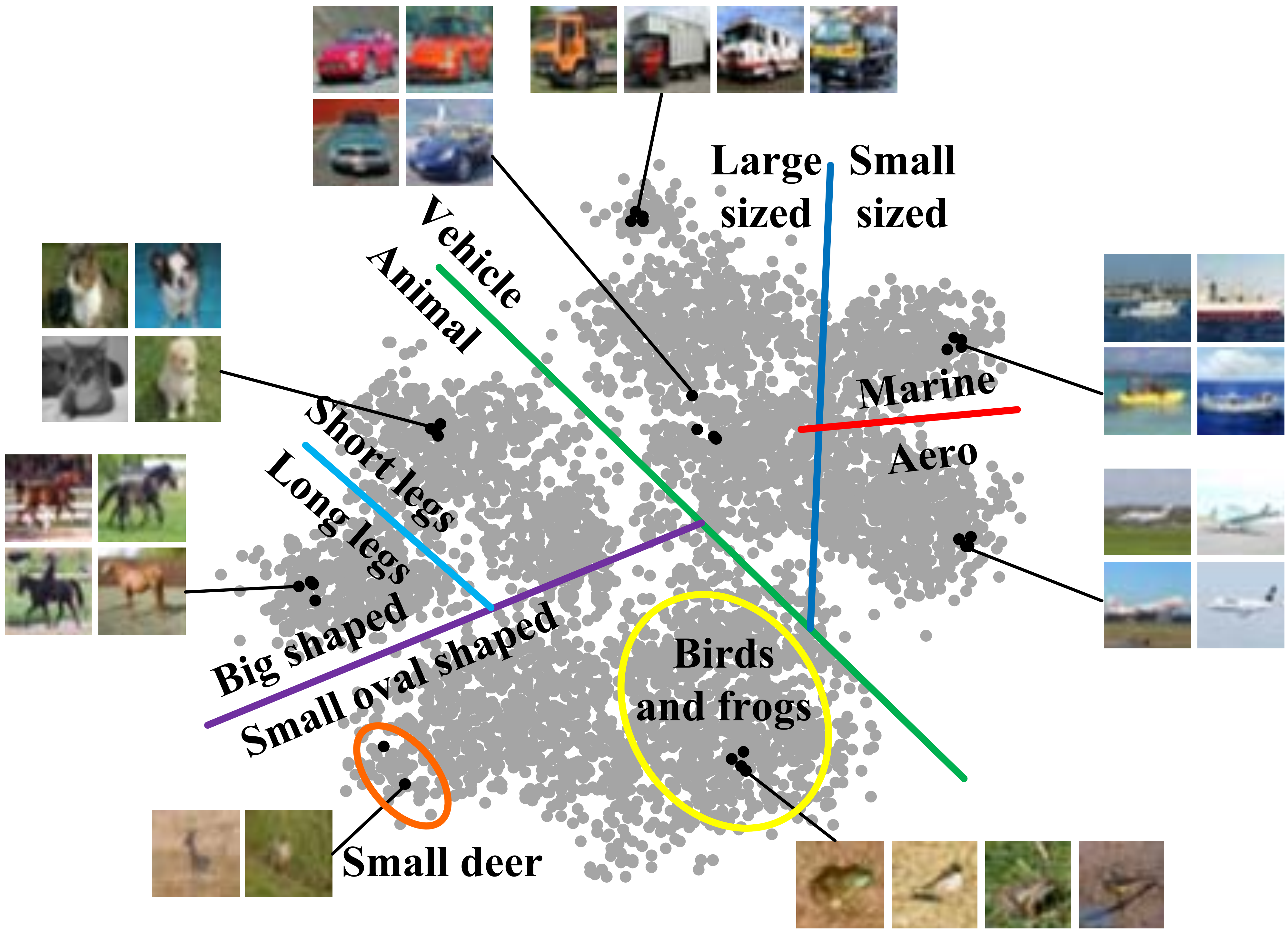
Pigou et al. (2014). Sign Language Recognition using Convolutional Neural Networks.



(a) RGB (b) Depth map



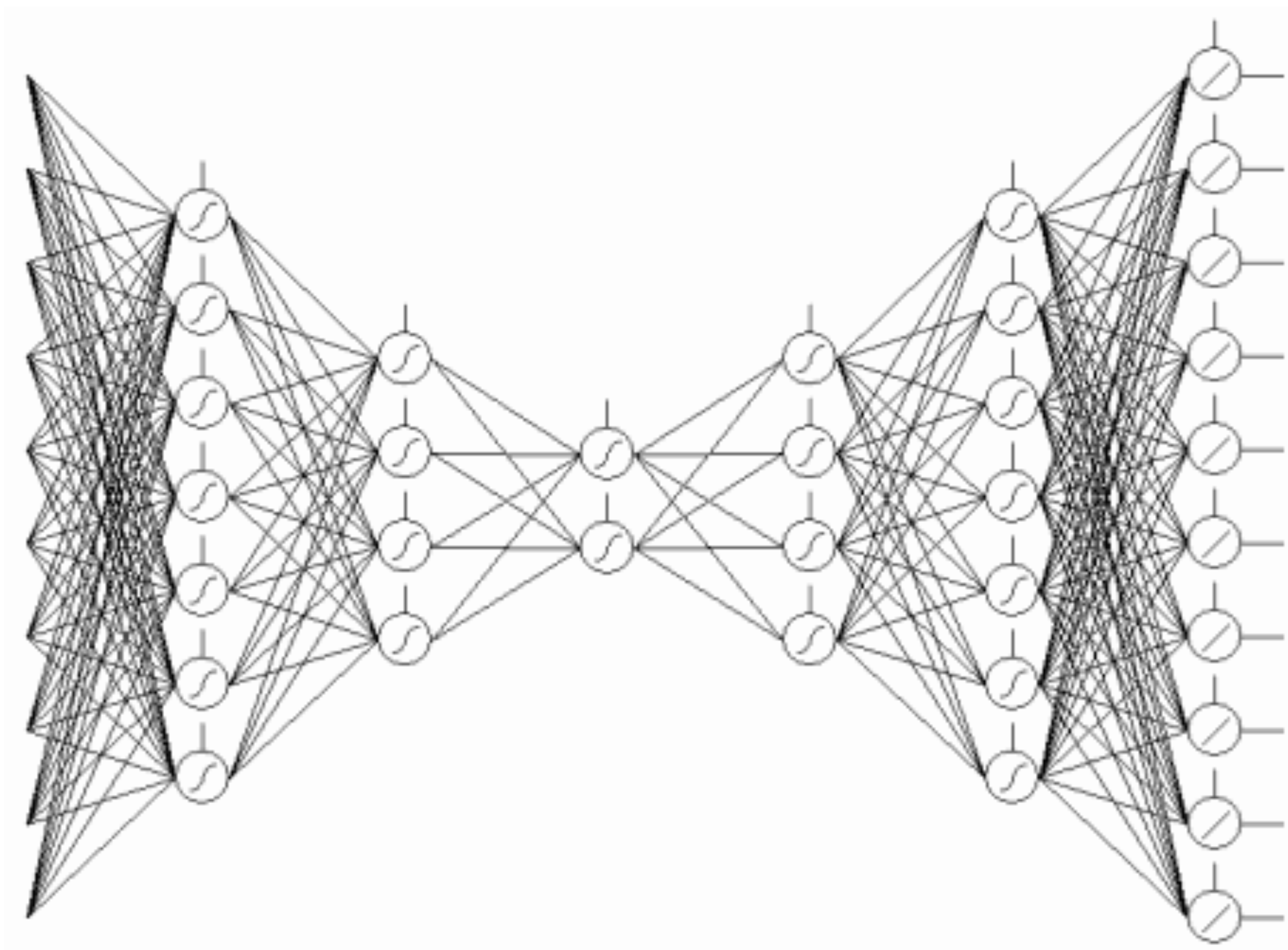
Unsupervised learning



[Huang, Loy, & Tang (2016)]

Autoencoder

x



x'

Another VQ example

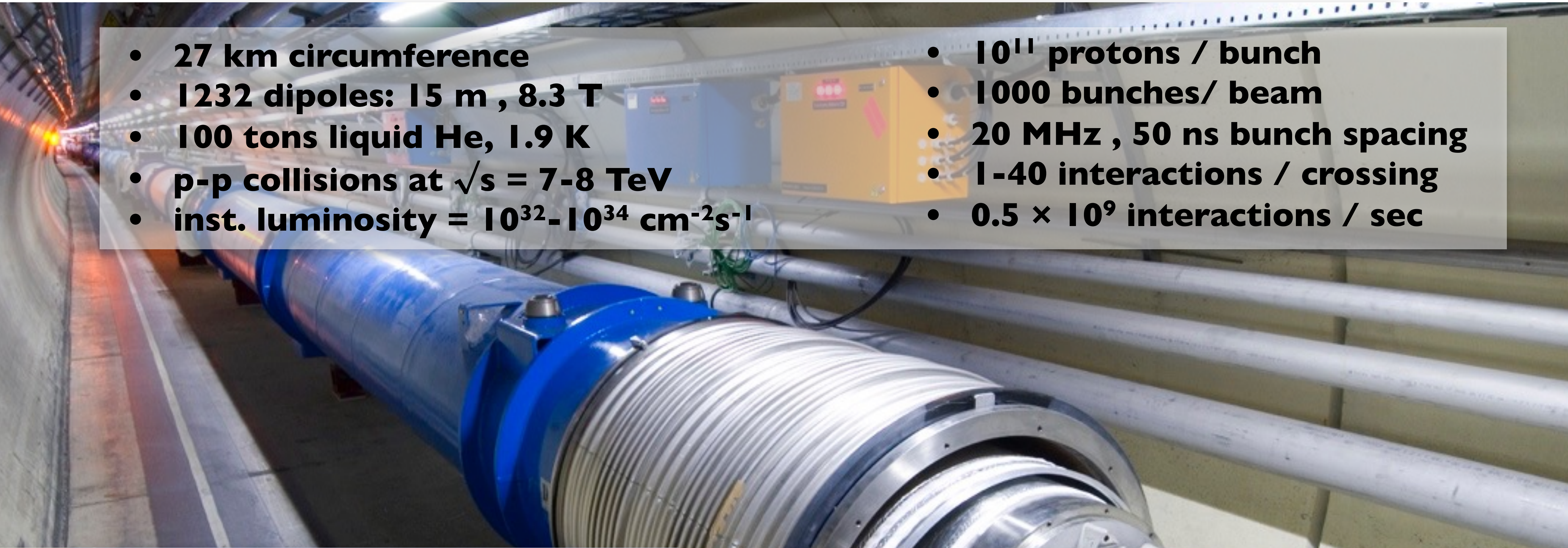
FIGURE 14.9. *Sir Ronald A. Fisher* (1890 − 1962) *was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a* 1024×1024 *grayscale image at* 8 *bits per pixel. The center image is the result of* 2×2 *block VQ, using* 200 *code vectors, with a compression rate of* 1.9 *bits/pixel. The right image uses only four code vectors, with a compression rate of* 0.50 *bits/pixel*

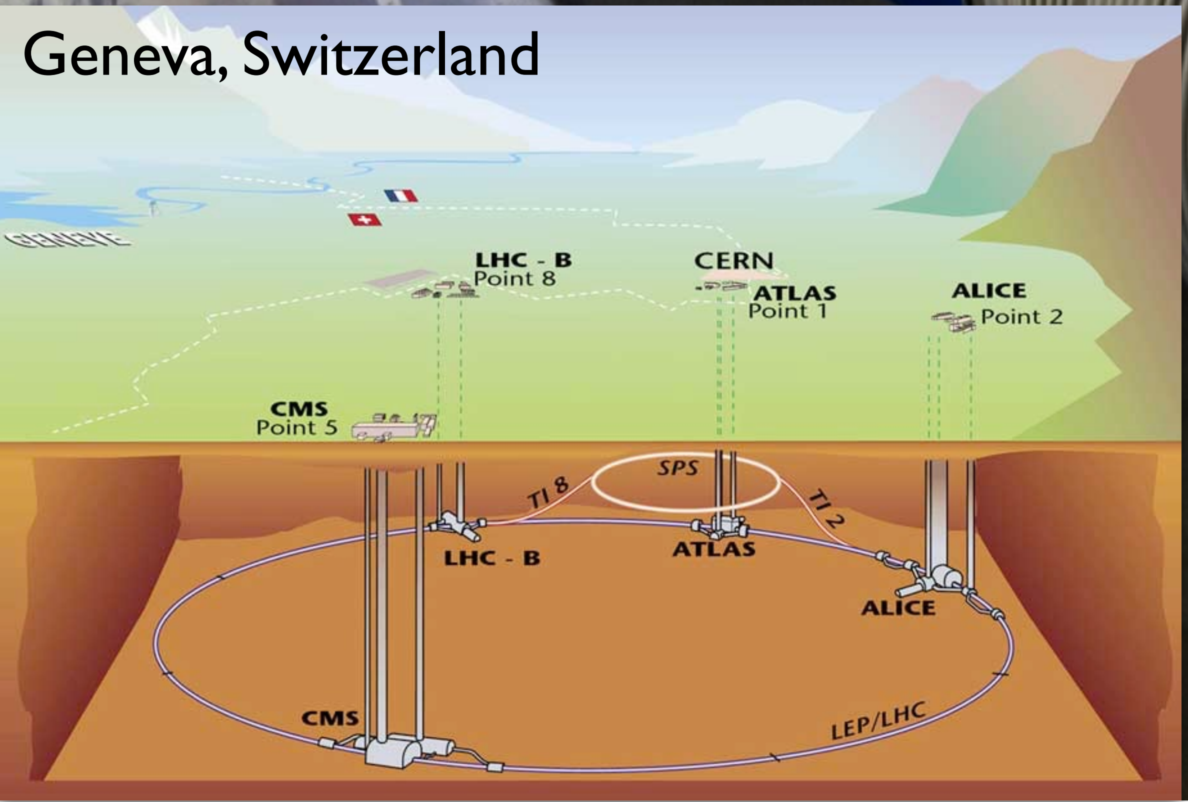
[Hastie, Tibshirani, & Friedman (2009). p. 514]

Large Hadron Collider

- 27 km circumference
- 1232 dipoles: 15 m , 8.3 T
- 100 tons liquid He, 1.9 K
- p-p collisions at $\sqrt{s}$ = 7-8 TeV
- inst. luminosity = 10^{32}-10^{34} $cm^{-2}s^{-1}$

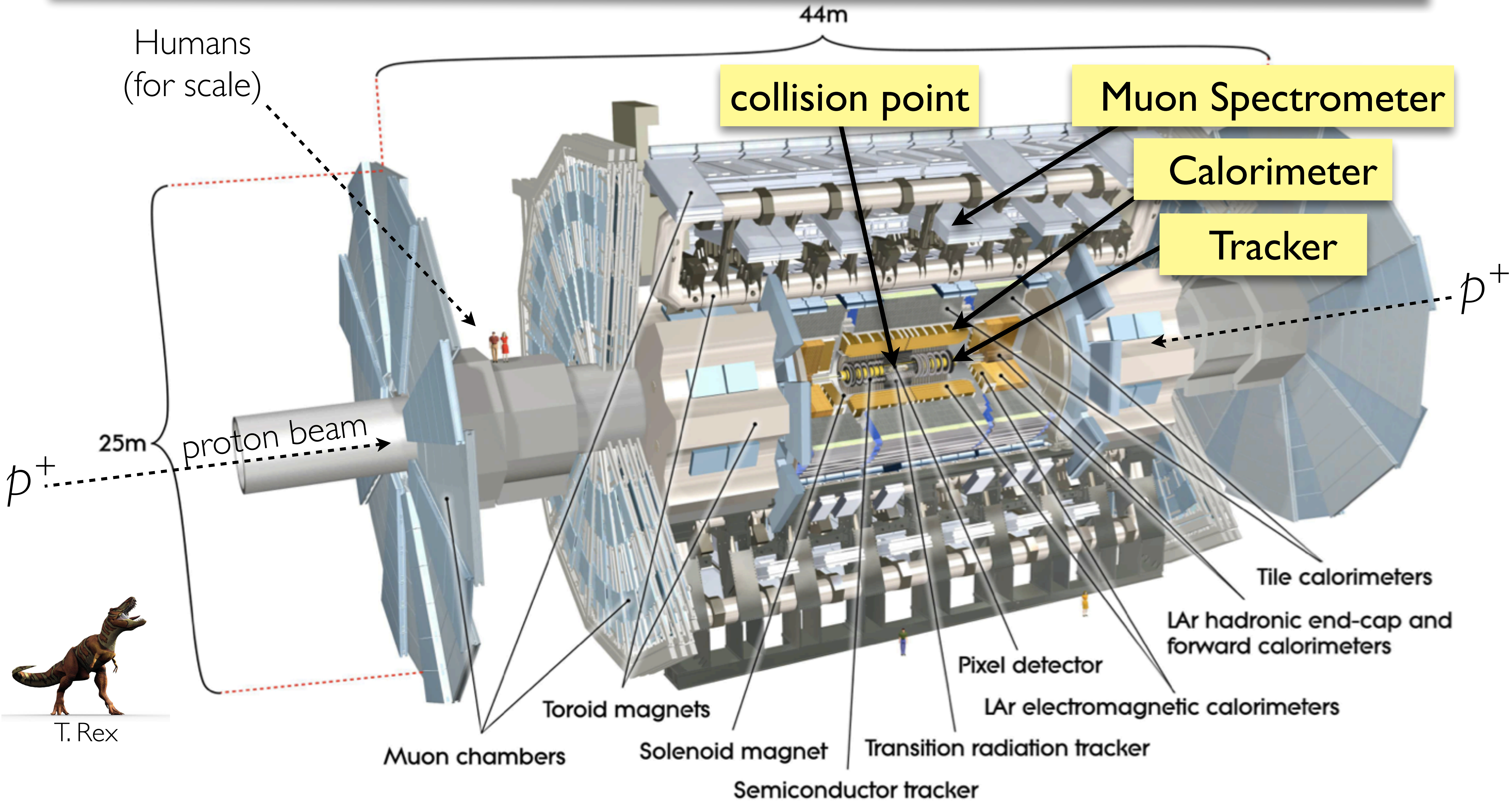
- 10^{11} protons / bunch
- 1000 bunches/ beam
- 20 MHz , 50 ns bunch spacing
- 1-40 interactions / crossing
- 0.5×10^{9} interactions / sec





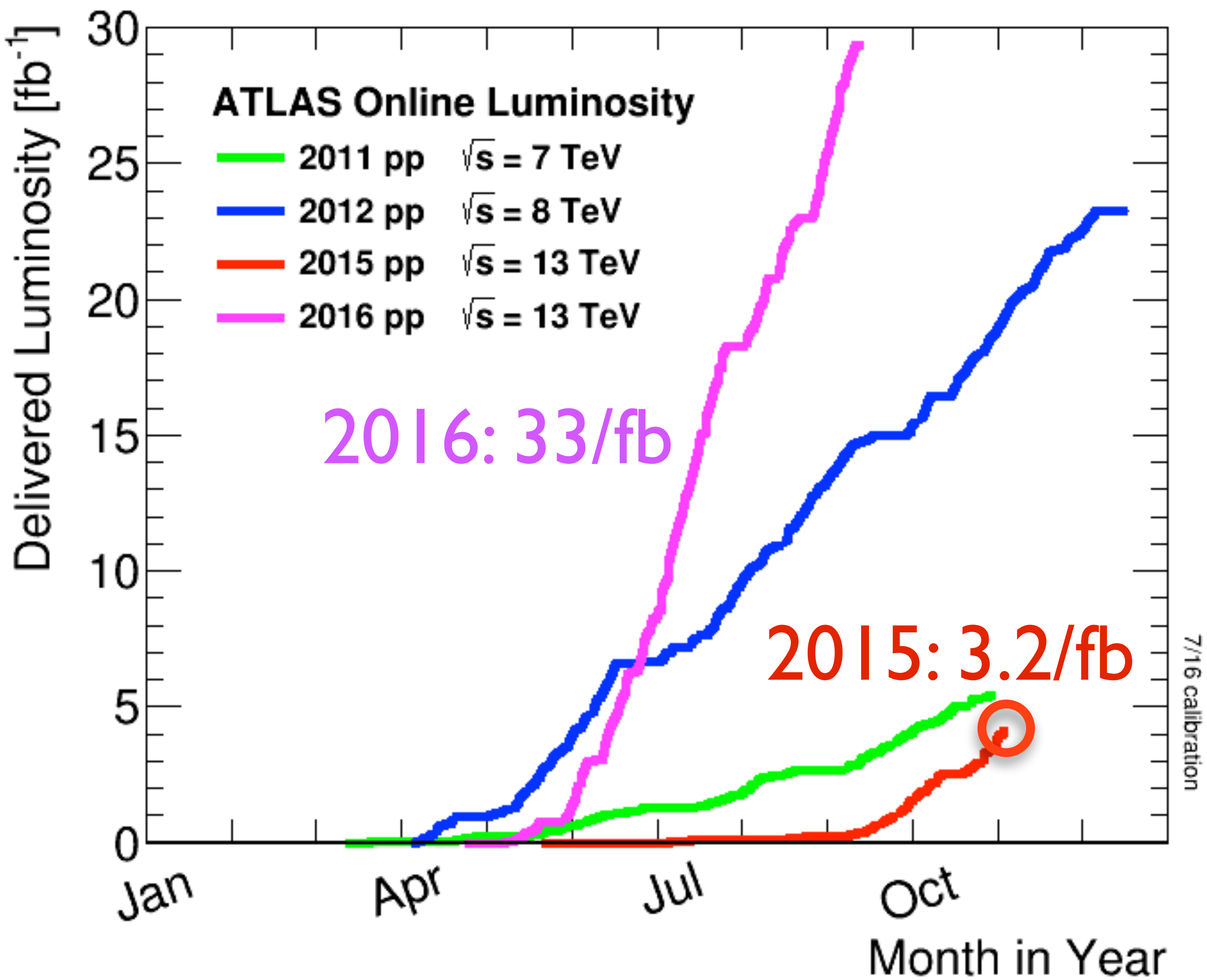
ATLAS Detector

ATLAS is a 7 story tall, 100 megapixel "camera", taking 3-D pictures of proton-proton collisions 40 million times per second, saving 10 million GB of data per year, using a world-wide computing grid with over 100,000 CPUs. The collaboration involves more than 3000 scientists and engineers.

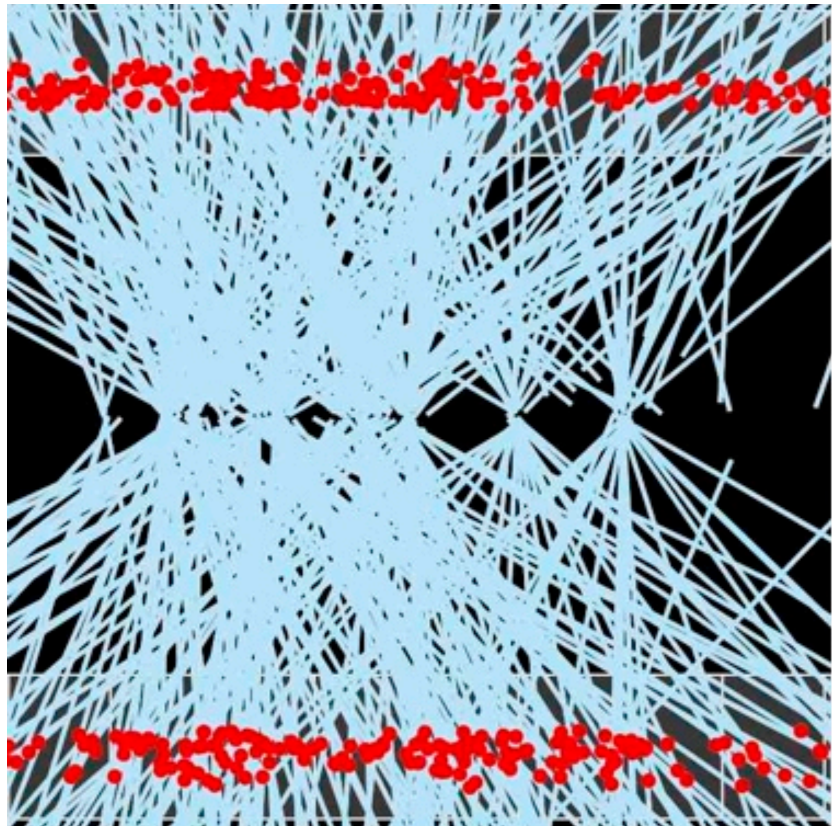


Datasets

The LHC has performed extremely well!!



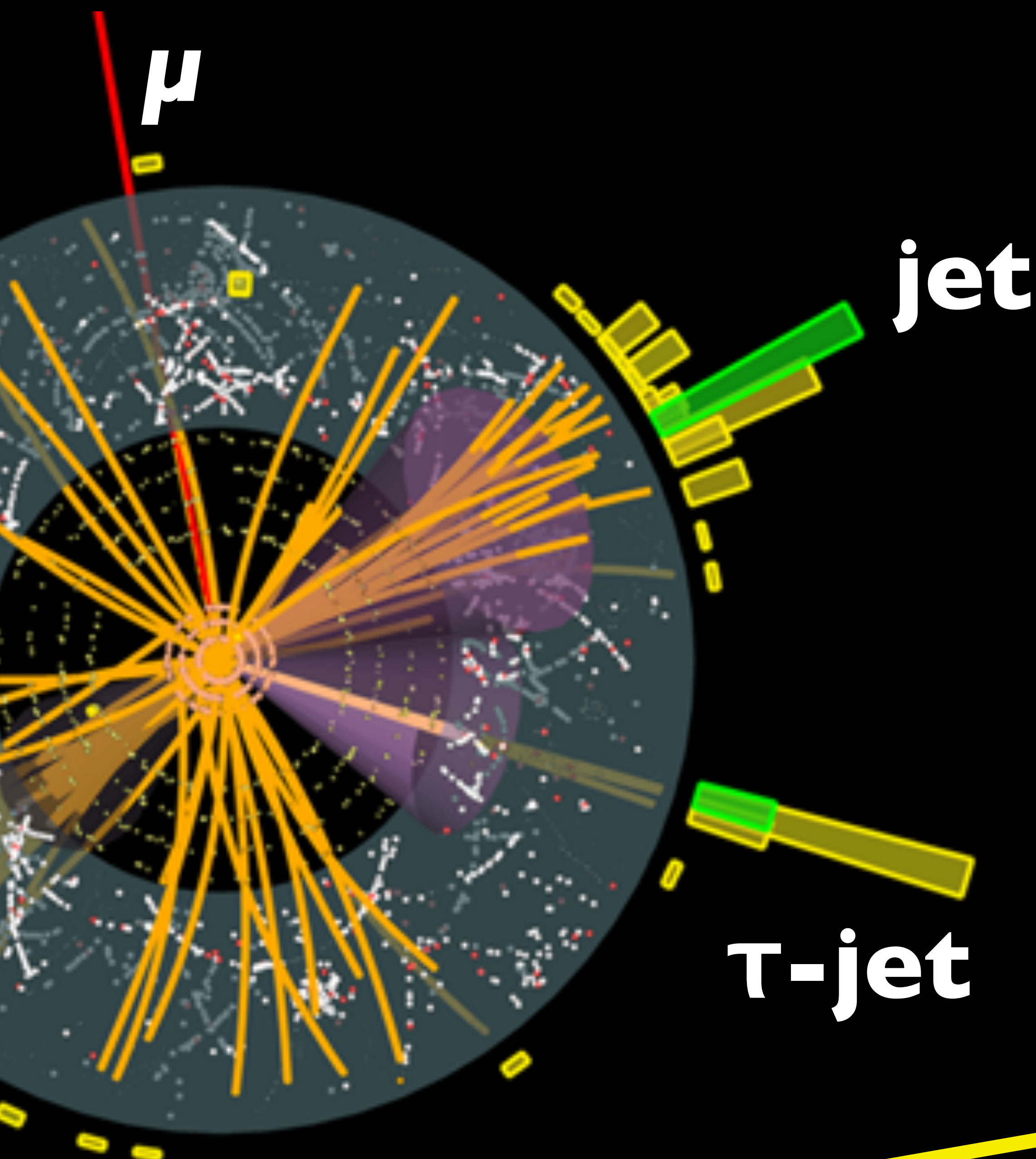
Recently broke inst. lumi. records > 10^{34} cm^{-2}s^{-1}



Typically 20-40 verticies per bunch crossing

Latest analyses combine collision data at $\sqrt{s}$=13TeV collected in the years 2015 and 2016, giving a total integrated lumi ≈ 36 fb^{-1}.

What do we reconstruct?



(main objects)

- muons
- electrons & photons
- jets of hadrons
- τ- and *b*-tagged jets
- missing energy

How do we search?

ATLAS Physics Groups

SM
W, *Z*, top,...

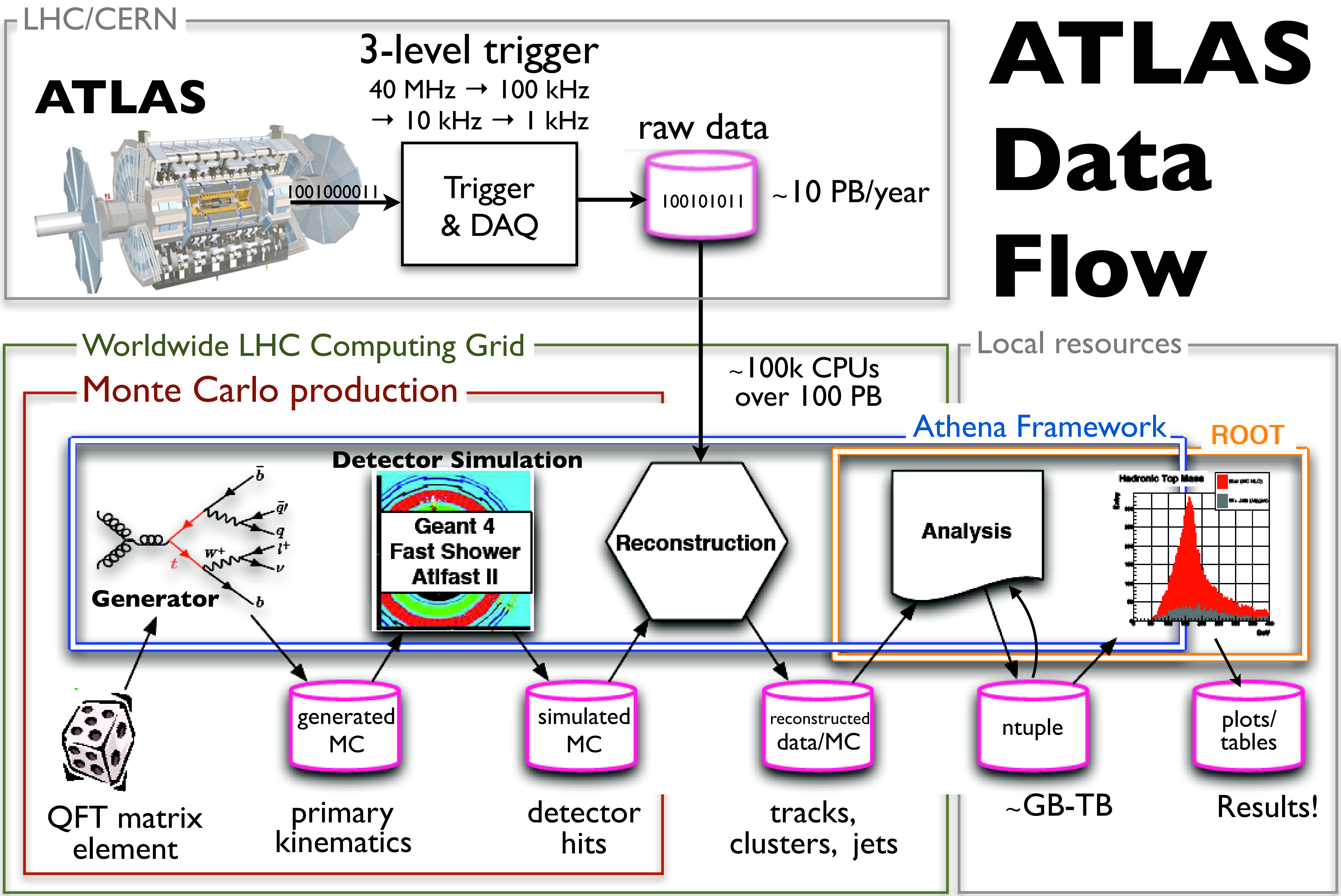
Higgs
$H \rightarrow \gamma\gamma$, *ZZ*, *WW*, ...

SUSY
l+jets, γ+jets, ...

Exotics
Z', *W'*, ...

Currently ATLAS has published 579+ papers

ATLAS Data Flow

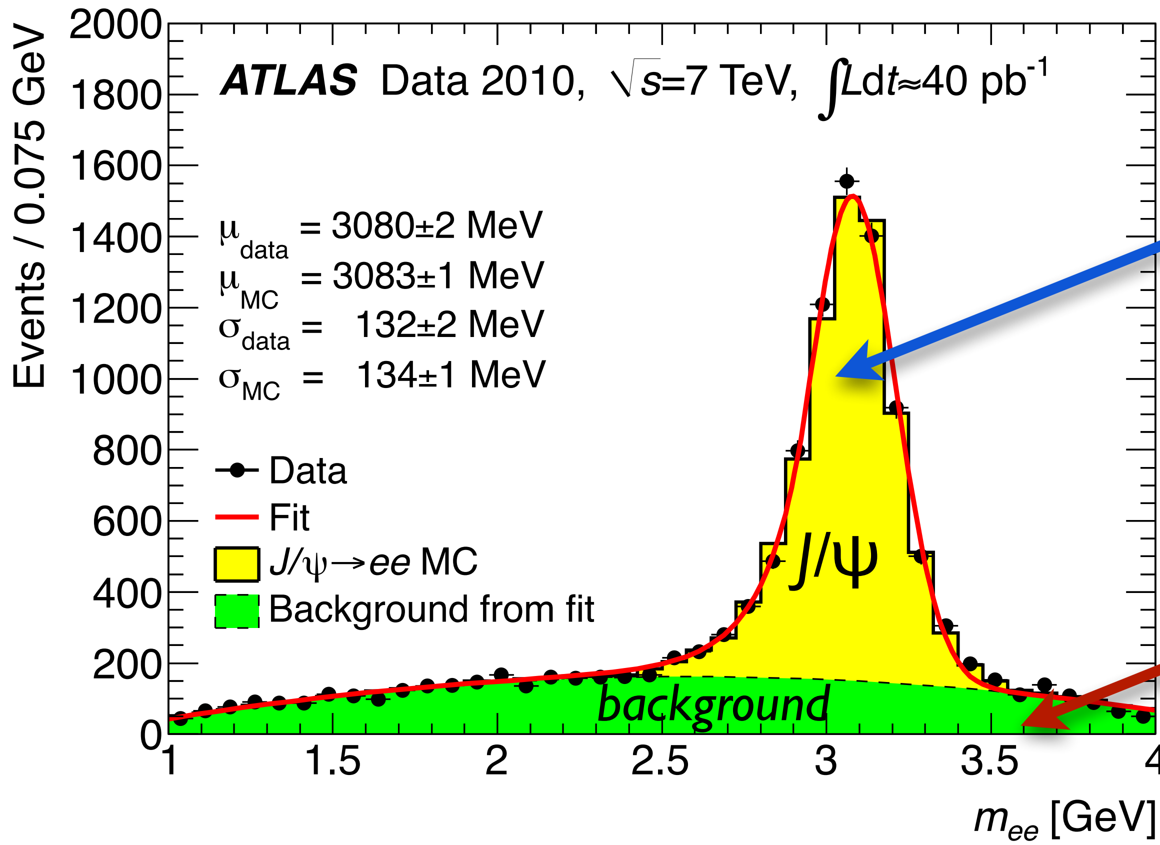


Building a model

N(expected) = *N(correct-ID)* + *N(fake)*

- ***Bottom-up***
- well-identified objects have scale factors from control regions
- estimated with detailed Monte Carlo simulation

- ***Top-down*, "data-driven"**
- various magic with data depending on the analysis and your creativity
- side-band fit
- fake-factor method



Bottom-up
Monte Carlo

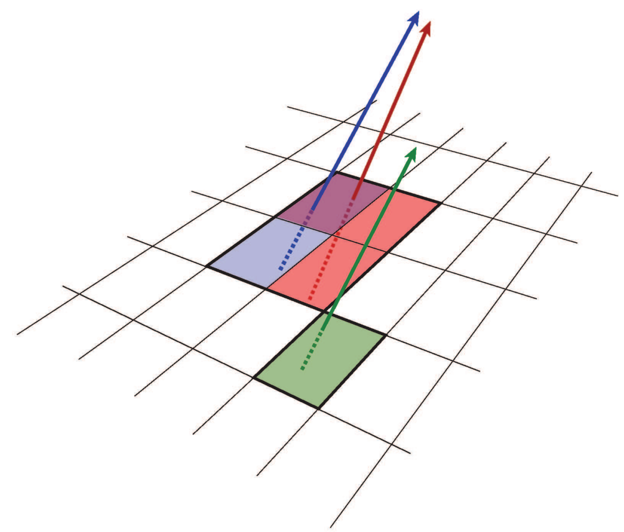
Data-driven
side-band fit

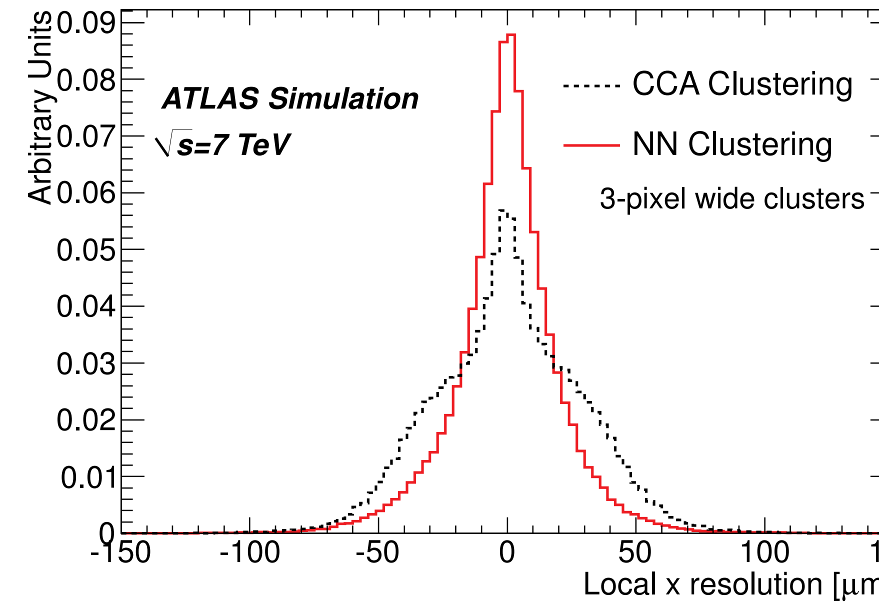
[arxiv:1110.3174]

NNs and BDTs in ATLAS

- Using NNs and other MVAs has been common in HEP for years, for pattern recognition, particle ID, event selection...
- In the past, always used *shallow* NNs.
- ATLAS uses NNs in many places, e.g. pixel clustering.
- Jet tagging for taus and b-quarks has used NNs in many iterations (also c, q/g).

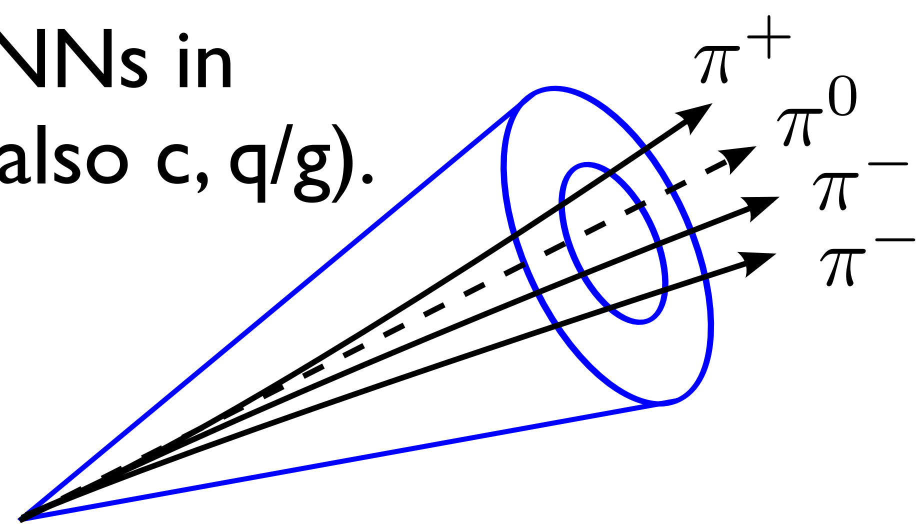
ATLAS pixel clustering with NNs

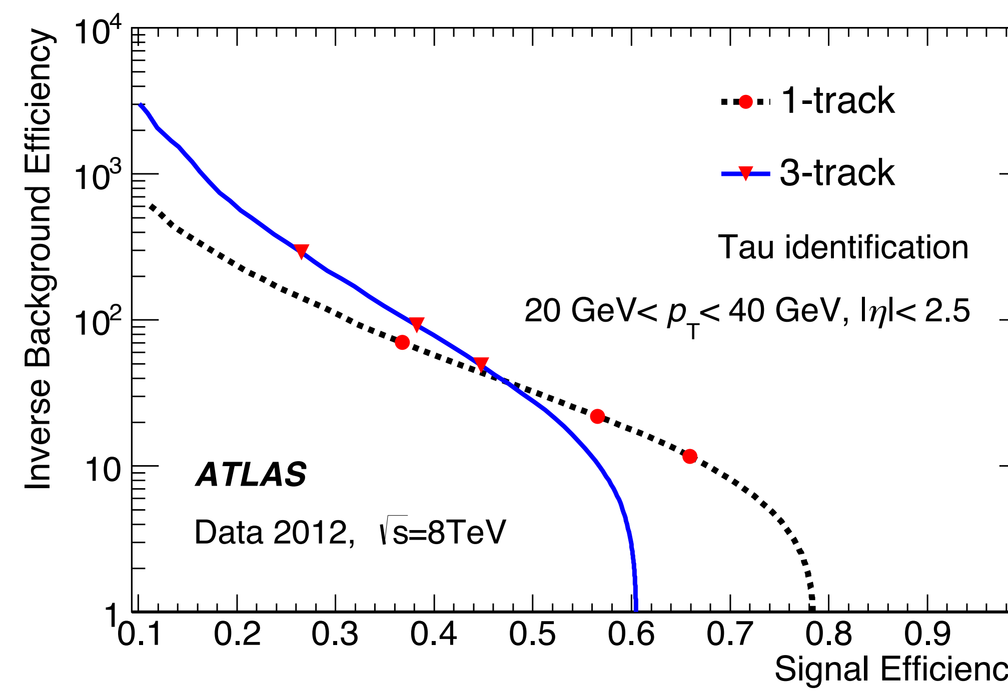




[1406.7690]

ATLAS tau identification with BDTs

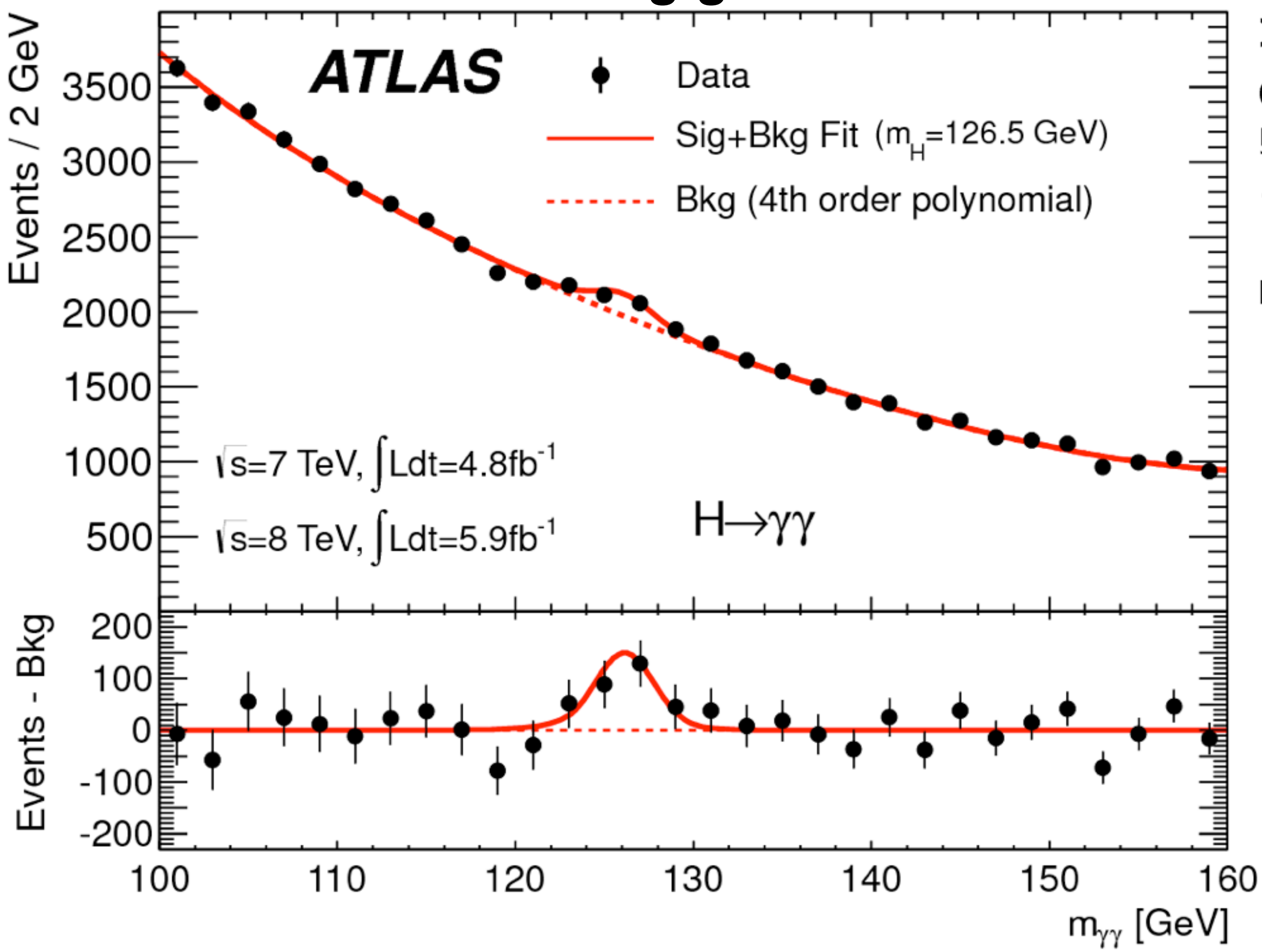




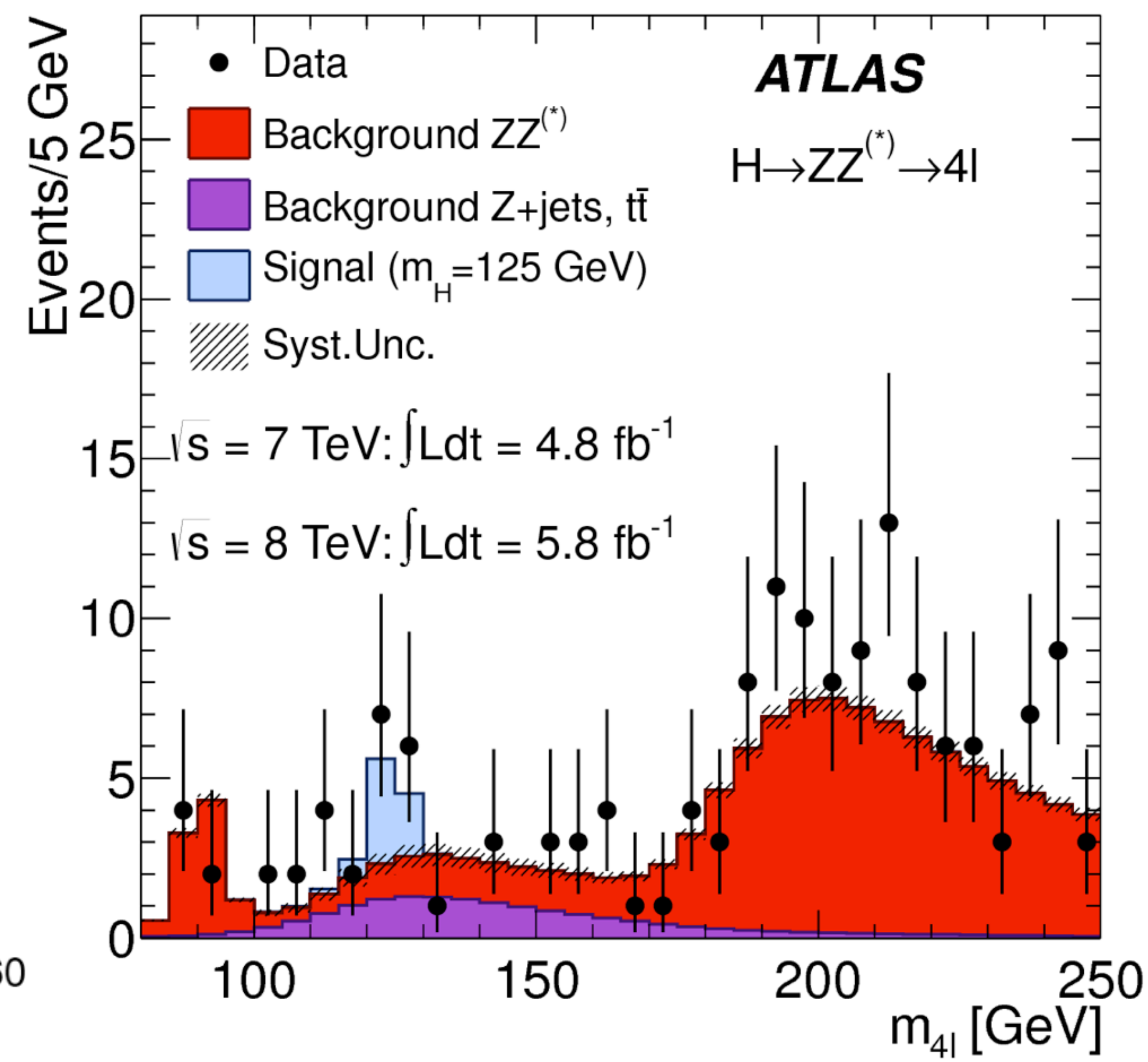
[1412.7086]

Higgs discovery

H→γγ

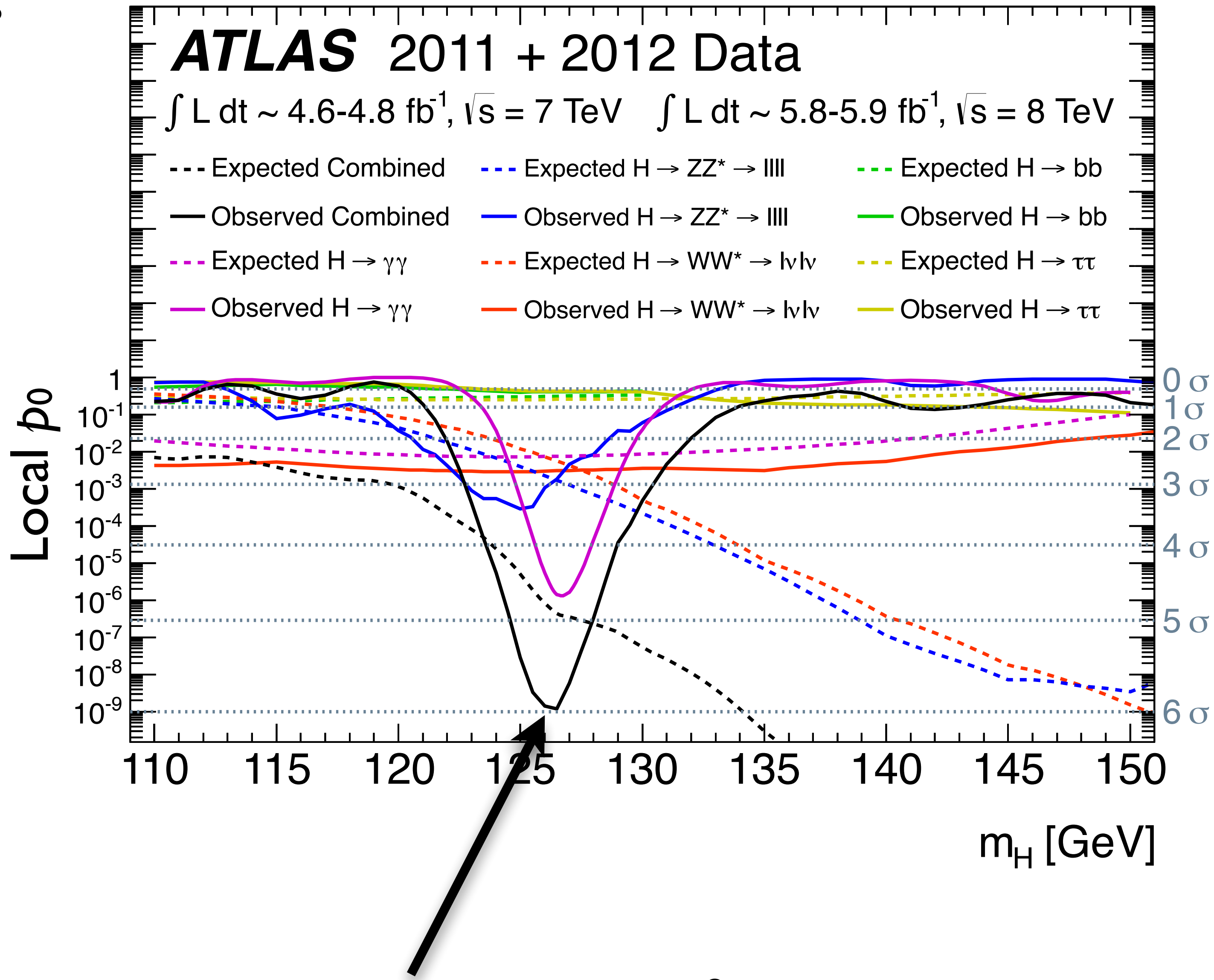


H→ZZ→4l

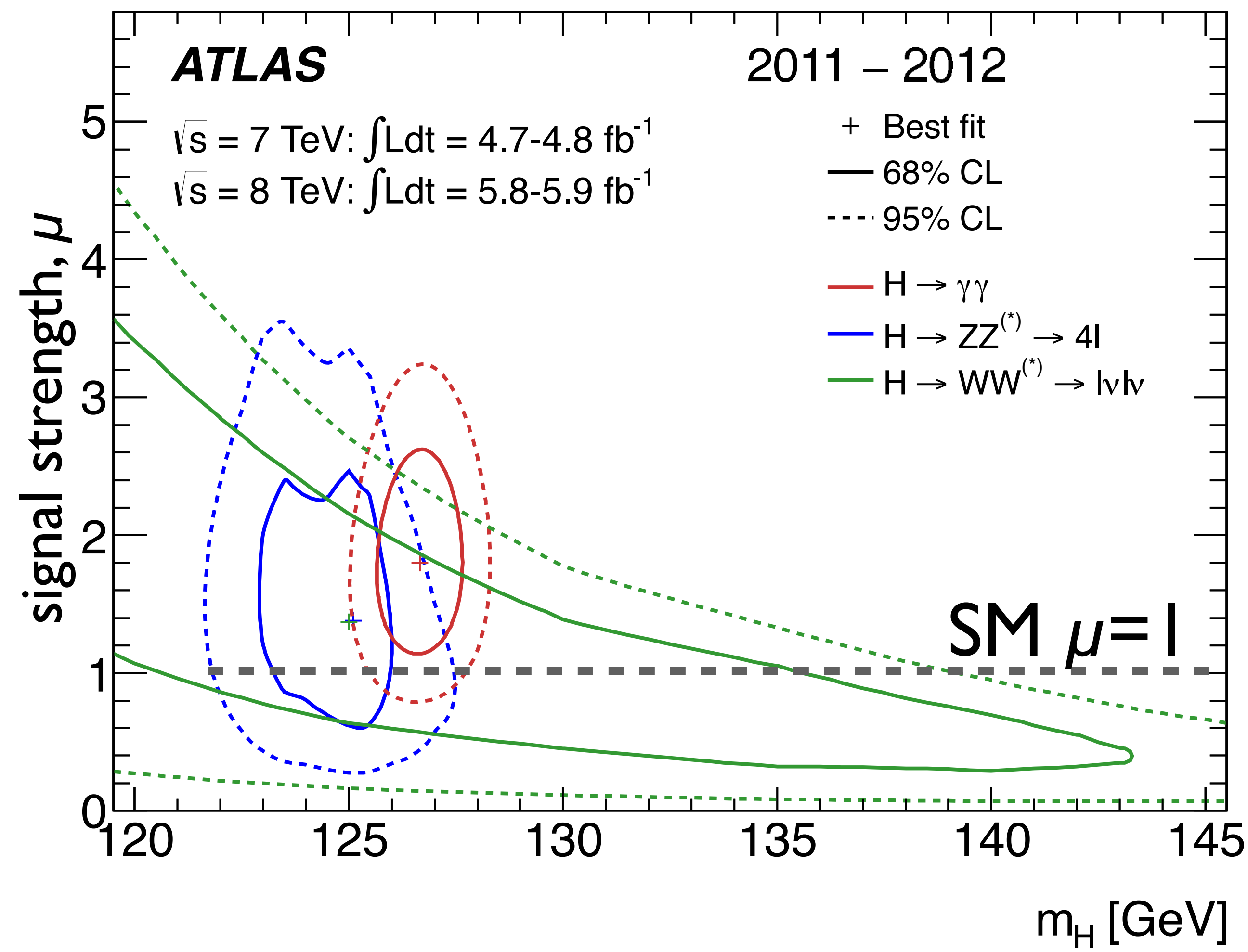


Higgs Confidence

Inconsistent with background only

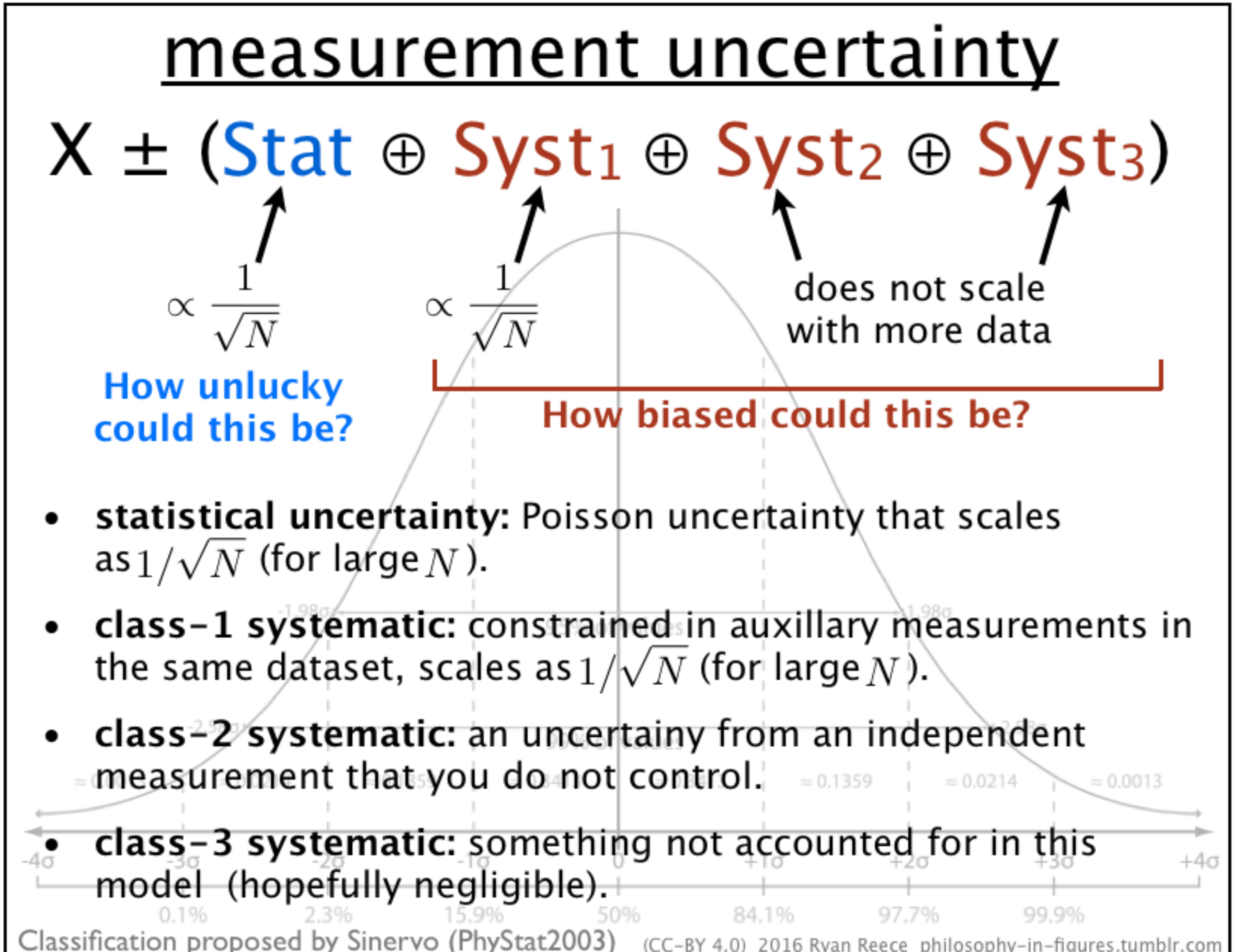


Consistent with SM Higgs



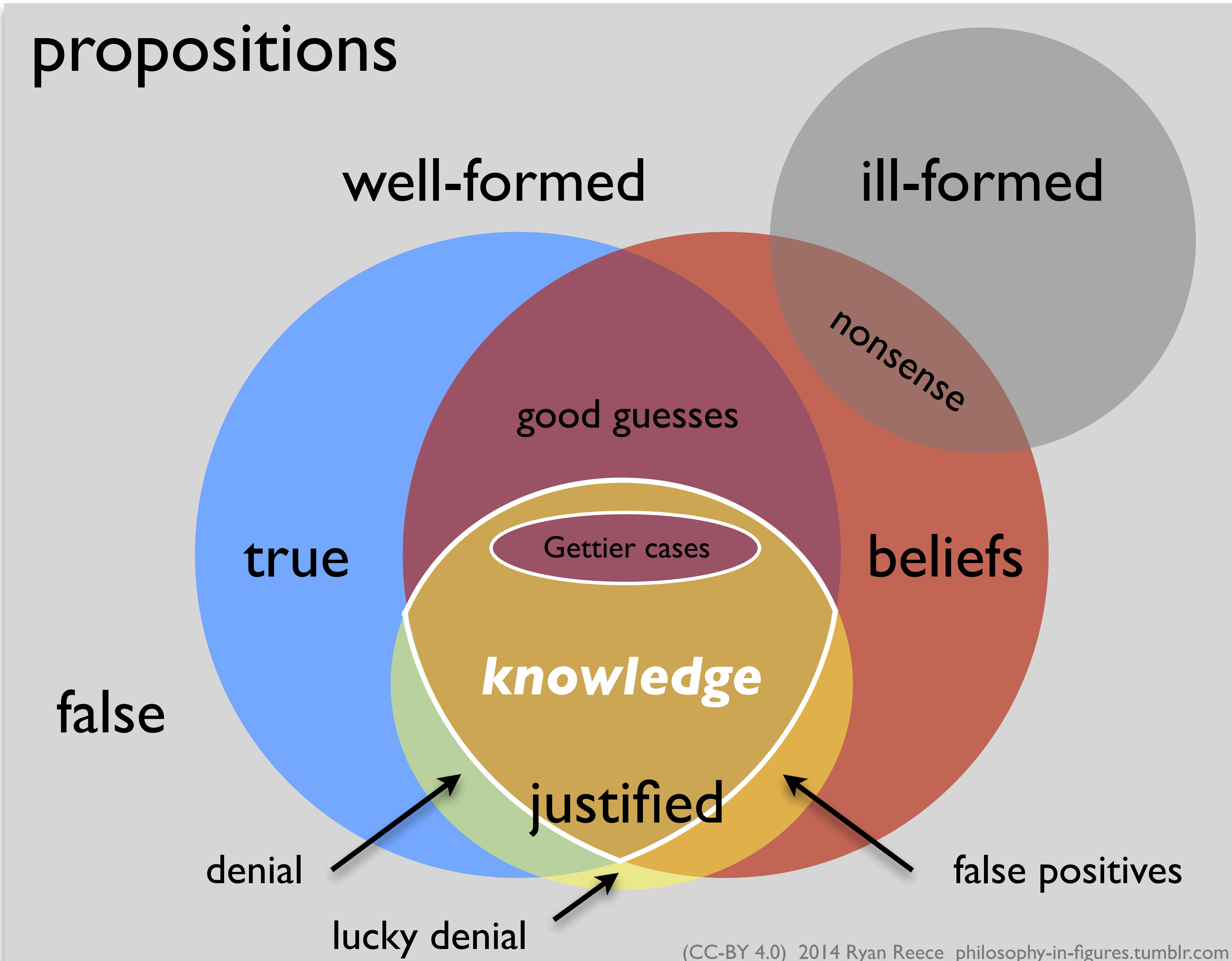
- Local $p_0 = 1.7 \times 10^{-9}$, corresponding to 5.9

Systematics



from: http://philosophy-in-figures.tumblr.com/

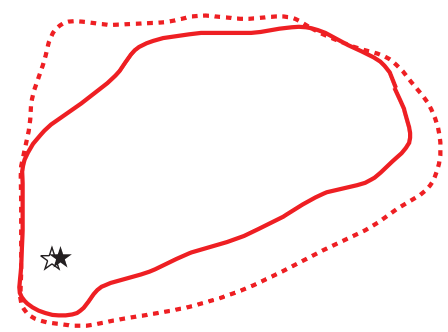
Knowledge = JTB-G



from: http://philosophy-in-figures.tumblr.com/

Confidence Intervals

- A ***frequentist confidence interval*** is constructed such that, given the model, if the experiment were repeated, each time creating an interval, 95% (or other CL) of the intervals would contain the true population parameter (*i.e.* the interval has ≈95% *coverage*).
  - They can be one-sided exclusions, *e.g.* $m(Z')$ > 2.0 TeV at 95% CL
  - Two-sided measurements, *e.g.* m_{H} = 125.1 ± 0.2 GeV at 68% CL
  - Contours in 2 or more parameters



- This ***is not the same*** as saying *"There is a 95% probability that the true parameter is in my interval"*. Any probability assigned to a parameter *strictly involves a Bayesian prior probability.*
- <u>Bayes' theorem:</u> P(Theory | Data) ∝ P(Data | Theory) P(Theory)

"likelihood" "prior"