Python for Data Science 2

Lecture 4- Models at the Large Hadron Collider

Amir Farbin

What is a “Model”?

- The context of science, data science, machine learning, etc, the word model is often used…
  - e.g disease transmission model, plate model, Standard Model of particle physics, neutral network model, …
- In science, models often try to capture some phenomena in a mathematical or heuristic way.
  - There are likely assumptions in a model, which often are simplifications
- Why? to understand, define, quantify, visualize, simulate, predict, …
- Often the general knowledge of a discipline (e.g. particle physics), is captured in a model.
  - It therefore represents what is known.
- Scientific research then becomes the iterative process of
  - comparing the model predictions with experimental results
  - improve the model
- Key is to understand what experiments to best reveal the weaknesses of these models.

- To understand the most fundamental phenomena in nature, the Physicists application of the scientific method:

1. Build Models

- The **Standard Model of HEP** describes the building blocks of matter and their interactions.
    - SM requires the existence of a set of fundamental particles with very specific properties.
        - Tested and validated through experiments, including some of the most precise measurements ever.
    - Yet, has **failures**, e.g. no Dark Matter. Not compatible with Gravity.
    - And, has **inconsistencies**: e.g. both SM predicts the Higgs mass to be big and requires it to be small.
- Models are **expressed mathematically**
    - Ideally build on a minimal set of rules (principles/laws)
    - SM Mathematical Framework: Quantum Field Theory- Way beyond this class.
        - Renormalization: encapsulating unknown (effects at high energy) in measurable parameters.
        - SM is a Model with **19 parameters**. Masses of particles. Strengths of forces. etc…
            - None are fundamental like speed of light or Planck's constant.

2. **Build new model that tackle inconsistencies**.

3. **Create experiments (target weaknesses in the model)**

Physics Models

Classical:
Calculus (Infinitesimal)
Object (a particle) described by x, y, z, α,β,γ and their derivatives (ie momentum).

Atomic scales

Quantum Mechanics:
Probabilistic
A particle described by a complex wave-function P(x, y, z, α,β,γ + derv).

Relativity explains Electromagnetic Unification. EM fields inherently relativistic.

General Relativity:
Tensor Calculous
Gravity = Curvature of Space Time.
Gravity weak, so curvature ignorable on small scales.

Relativistic QM Requires Creation of Particles... leads to Fields

Classical Field Theory:
eg Electrodynamics
Interaction of particles with a dynamic field A(x, y, z).

Relativistic Particles

Quantum Field Theory:
eg: QED, Standard Model
Particles = Fields.
Gauge Symmetries = the classical dynamic Fields = Particles = Electroweak
Calculate Probability of interaction.

Building the SM

Field Theory + Quantum Mechanics + Relativity = Quantum Field Theory

Ingredients

1. Leptons + Quarks

FERMIONS matter constituents spin = 1/2, 3/2, 5/2, ...

Leptons spin = 1/2			
Flavor		Mass GeV/c^2	Electric charge
ν_e	electron neutrino	$<1\times10^{-8}$	0
e	electron	0.000511	–1
ν_μ	muon neutrino	<0.0002	0
μ	muon	0.106	–1
ν_τ	tau neutrino	<0.02	0
τ	tau	1.7771	–1

Quarks spin = 1/2			
Flavor		Approx. Mass GeV/c^2	Electric charge
u	up	0.003	2/3
d	down	0.006	–1/3
c	charm	1.3	2/3
s	strange	0.1	–1/3
t	top	175	2/3
b	bottom	4.3	–1/3

2. Three Local Gauge Symmetries

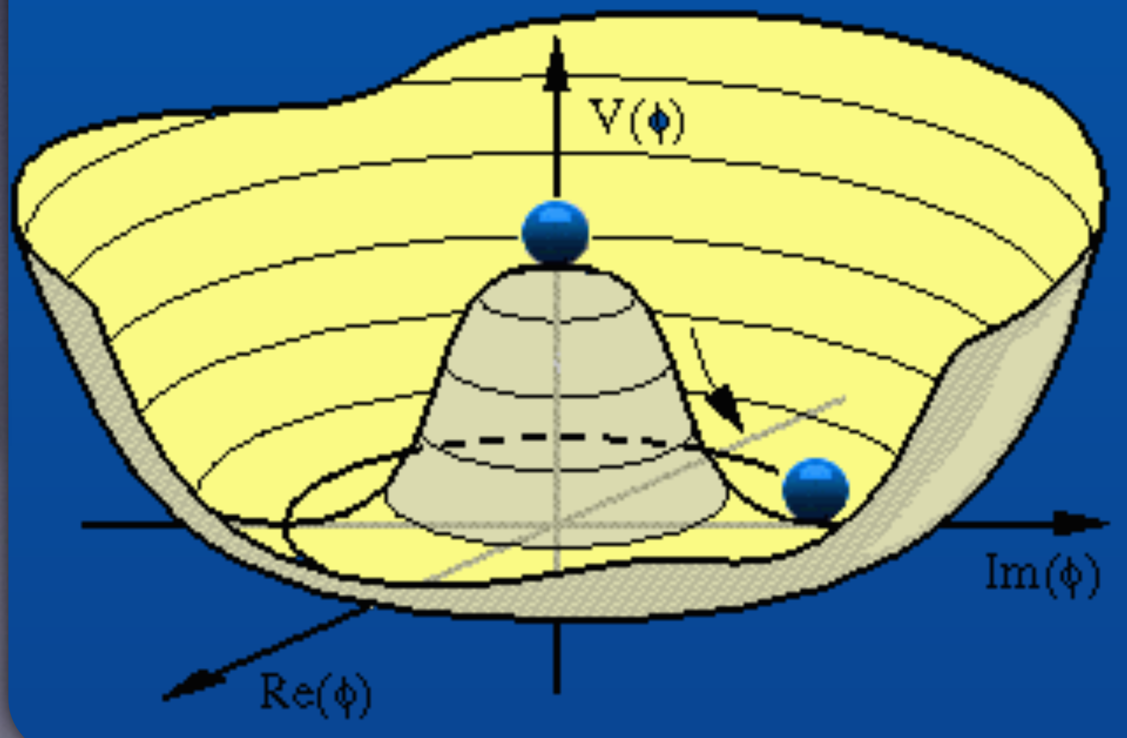
- Each Symmetry implies existence of a new set of boson (spin 1) fields
- These bosons "carry" the Forces

Unified Electroweak spin = 1		
Name	Mass GeV/c^2	Electric charge
γ photon	0	0
W$^-$	80.4	–1
W$^+$	80.4	+1
Z^0	91.187	0

Strong (color) spin = 1		
Name	Mass GeV/c^2	Electric charge
g gluon	0	0

3. Higgs Mechanism

- Every particle interacts with a scalar (spin 0) field
- This field has non-zero Vacuum Expectation Value
  - Breaks symmetry between E&M and Weak Forces
  - Gives masses to all particles w/o breaking Gauge Symmetries

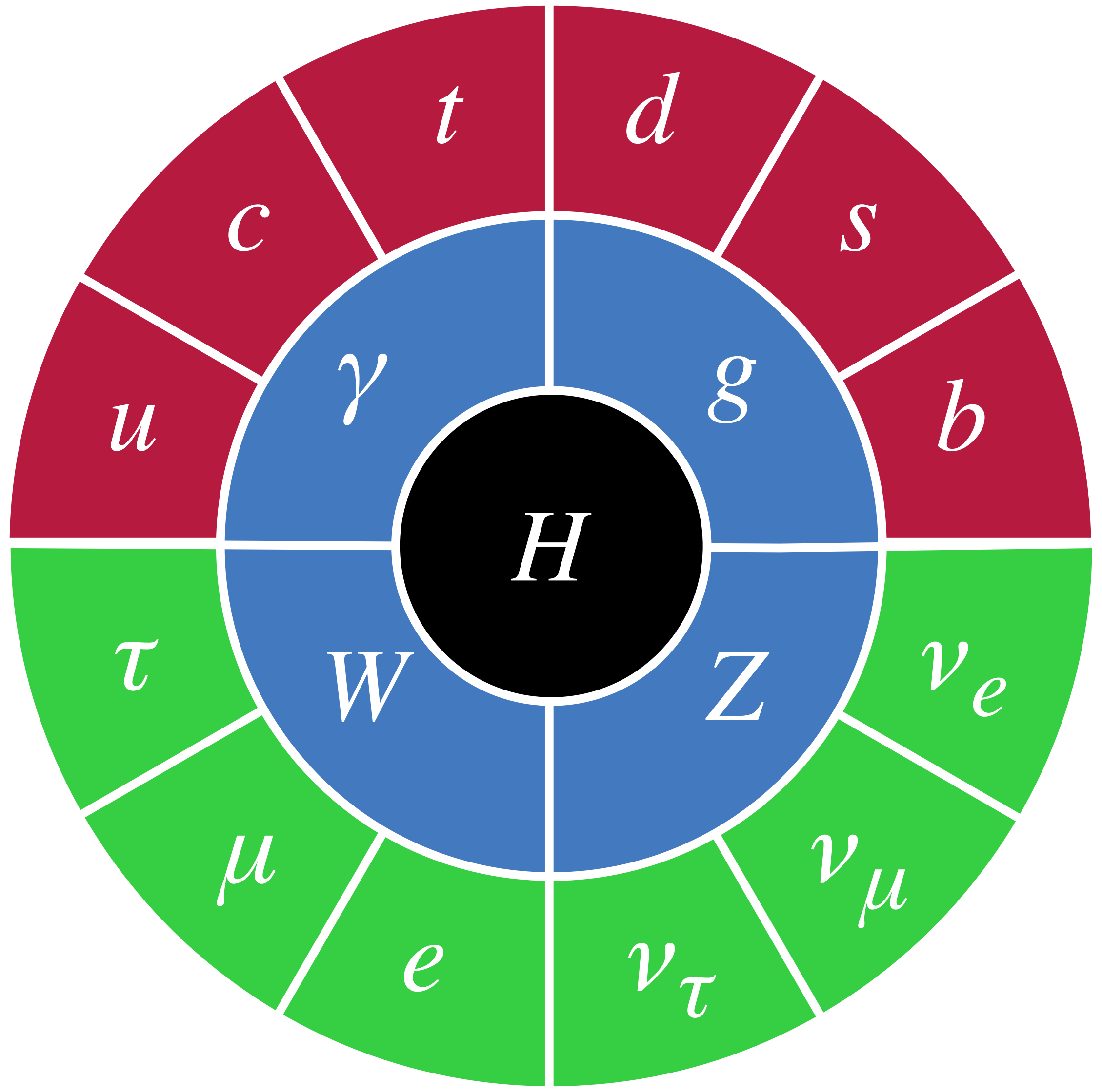


Standard Model

Agreement with all experimental results so far

PARTICLE PHYSICS: 19 PARAMETERS

$$
\begin{aligned}
\mathcal{L}_{SM} = {} & \underbrace{\frac{1}{4}\mathbf{W}_{\mu\nu}\cdot\mathbf{W}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a}_{\text{kinetic energies and self-interactions of the gauge bosons}} \\
& + \underbrace{\bar{L}\gamma^\mu\left(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu\right)L + \bar{R}\gamma^\mu\left(i\partial_\mu - \frac{1}{2}g'YB_\mu\right)R}_{\text{kinetic energies and electroweak interactions of fermions}} \\
& + \underbrace{\frac{1}{2}\left|\left(i\partial_\mu - \frac{1}{2}g\tau\cdot\mathbf{W}_\mu - \frac{1}{2}g'YB_\mu\right)\phi\right|^2 - V(\phi)}_{W^\pm, Z, \gamma, \text{and Higgs masses and couplings}} \\
& + \underbrace{g''(\bar{q}\gamma^\mu T_a q)G^a_\mu}_{\text{interactions between quarks and gluons}} + \underbrace{(G_1\bar{L}\phi R + G_2\bar{L}\phi_c R + h.c.)}_{\text{fermion masses and couplings to Higgs}}
\end{aligned}
$$



Symbol	Description	Value
m_e	Electron mass	511 keV
m_μ	Muon mass	105.7 MeV
m_τ	Tau mass	1.78 GeV
m_u	Up quark mass	1.9 MeV
m_d	Down quark mass	4.4 MeV
m_s	Strange quark mass	87 MeV
m_c	Charm quark mass	1.32 GeV
m_b	Bottom quark mass	4.24 GeV
m_t	Top quark mass	172.7 GeV
θ_{12}	CKM 12-mixing angle	13.1°
θ_{23}	CKM 23-mixing angle	2.4°
θ_{13}	CKM 13-mixing angle	0.2°
δ	CKM CP-violating Phase	0.995
g_1	U(1) gauge coupling	0.357
g_2	SU(2) gauge coupling	0.652
g_3	SU(3) gauge coupling	1.221
θ_{QCD}	QCD vacuum angle	~0
v	Higgs vacuum expectation value	246 GeV
m_H	Higgs mass	125 GeV

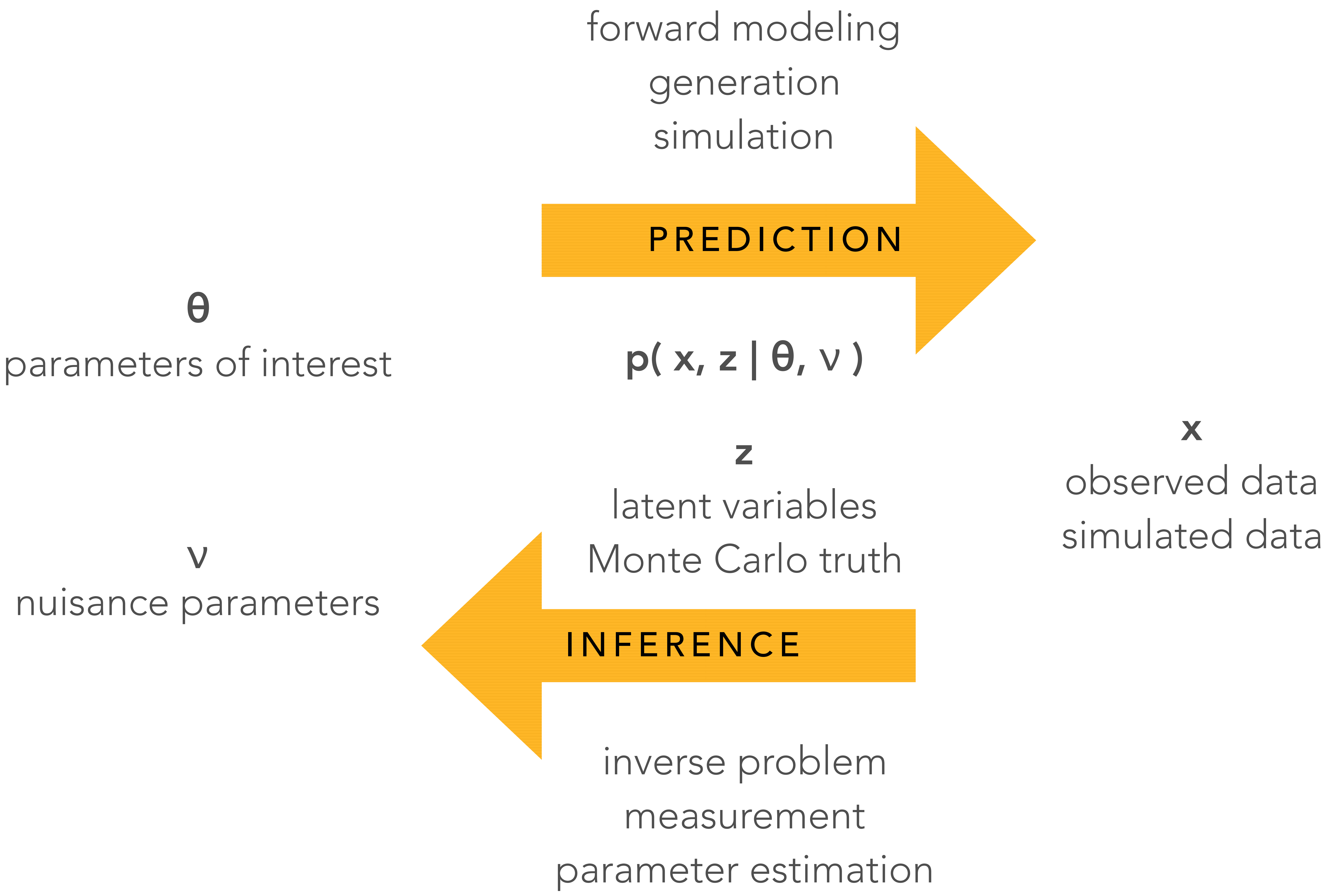
- Why Look Beyond the Standard Model?
  - Takes 19 parameters (eg Masses)... Why these values?
  - Still looking for the Higgs particle.
  - Gravity not included! Why gravity is so much weaker than everything else?
  - Misses a lot of the Universe: No Dark matter candidate. Can't explain Dark Energy.
  - Doesn't have enough asymmetry between matter/anti-matter to explain why we exist!
- At ~ 1 TeV of energy, some of the SM predictions don't make sense. So something ***new*** has to happen at 1 TeV.

Data Model

- Lets formalize what we mean by a ***dataset*** with a Probabilistic Model:
    - Assumption: Observed **Data is a mixture of *M* different processes**
    - Data set of *N* **data points**, $\{\{x_d\}_i\}$
        - each $\{x_d\}_i$ consisting of
            - *d* **observations** $\{x_d\}$
            - probability f_j of uniquely coming from one of *M* **classes**
                - each class has label c_j is indexed by j
                - dependent on parameters $\{\alpha_k\}_j$ (some **parameters of of interest**, some **nuisance parameters**)
        - Dependent on other parameters $\{\beta_l\}$
- $\Longrightarrow P(\{x\}|\theta) = P(\{x\}|\{f_j, c_j, \{\{\alpha_k\}_j\}, \{\beta_l\}) = \Sigma_j \, f_j \, P(\{x\}|c_j, \{\alpha_k\}_j, \{\beta_l\})$

What is it good for?

- If we know $P(X|\theta)$, what is it good for? (Statistical Inference)
  - ***Prediction***: Assume $\theta \Longrightarrow$ distribution of $\{x_d\}$.
  - ***Classification***: Observation $\{x_d\} \Longrightarrow$ most likely class c
  - ***Regression***: Dataset $\{\{x_d\}_i\} \Longrightarrow$ parameters of interest $\{\alpha_j^k\}$ or $\{\beta_l\}$
  - ***Hypothesis test***: Dataset $\{\{x_d\}_i\} \Longrightarrow$ is H_1 true (or H_0 null hypothesis)


forward modeling
generation
simulation
PREDICTION
θ
parameters of interest
p(x, z | θ, ν)
z
latent variables
Monte Carlo truth
x
observed data
simulated data
ν
nuisance parameters
INFERENCE
inverse problem
measurement
parameter estimation

Data Analysis

- Objectives:
  - ***Searches*** (hypothesis testing): Likelihood Ratio Test (Neyman-Pearson lemma)
  - ***Limits*** (confidence intervals): Also based on Likelihood
  - ***Measurements***: Maximum Likelihood Estimate

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

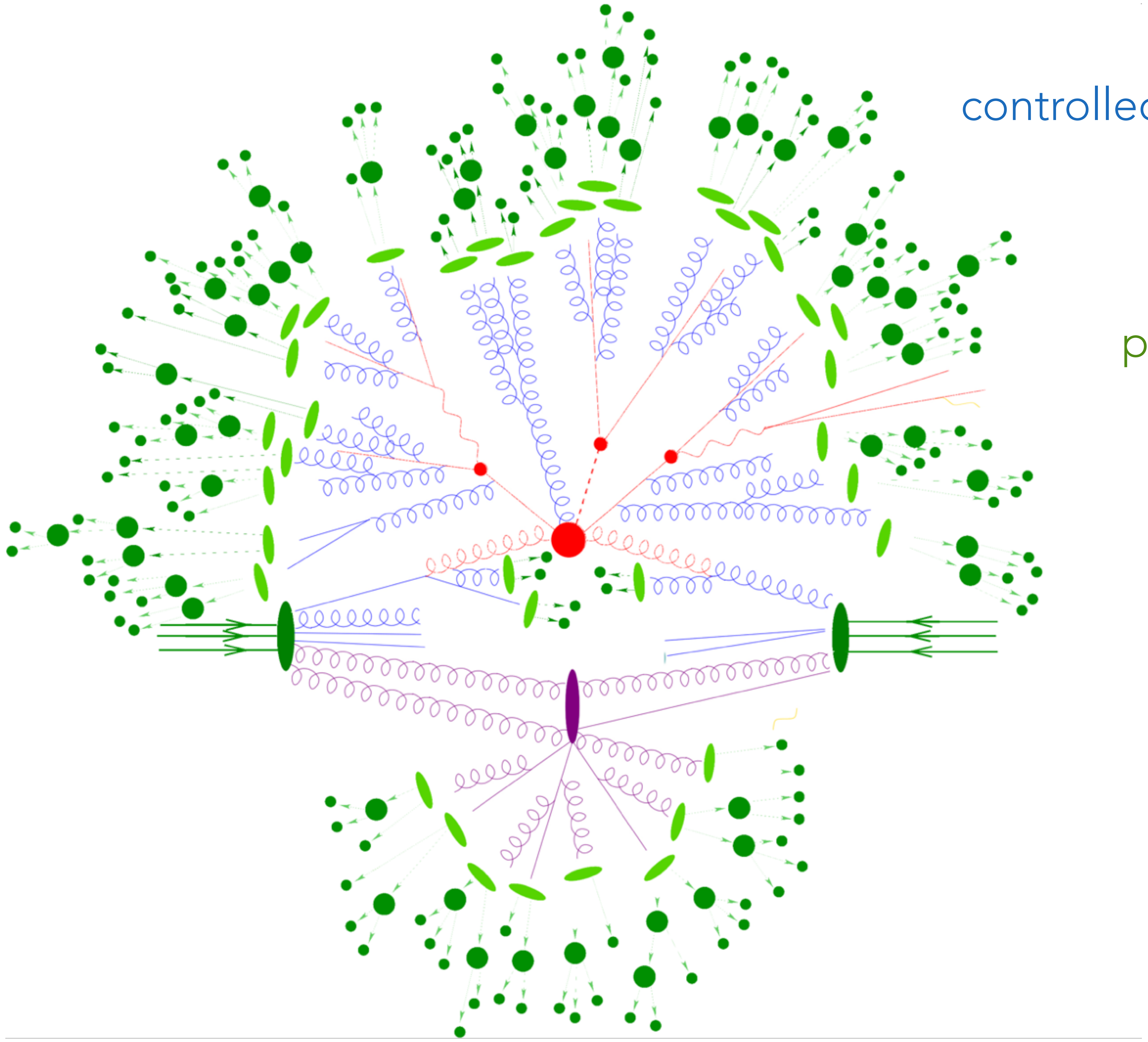
- ***Likelihood***

$$p(\{x\}|\theta) = \mathrm{Pois}(n|\nu(\theta)) \prod_{e=1}^{n} p(x_e|\theta)$$

  - *n* Independent Events (*e*) with Identically Distributed Observables ({*x*})
  - Significant part of Data Analysis is ***approximating the likelihood*** as best as we can.

Obtaining the Likelihood

- How do we obtain *P(X|θ)*?
    - In HEP, we have precise ***algorithmic simulations*** that generate $\{x_d\}$ given θ.
        - We estimate *P* by comparing observed x_d with simulated $\{x_d\}$.
    - We can build ***analytical first principle models***. *Matrix Element Method* is such a technique.
        - But it's technically difficult, computationally expensive, and only tractable with physics and detector simplifications.
    - We can build a statistical model *P* based on previous data.
        - We can using ML to learn *P* from simulation or data.



pencil & paper calculable from first principles

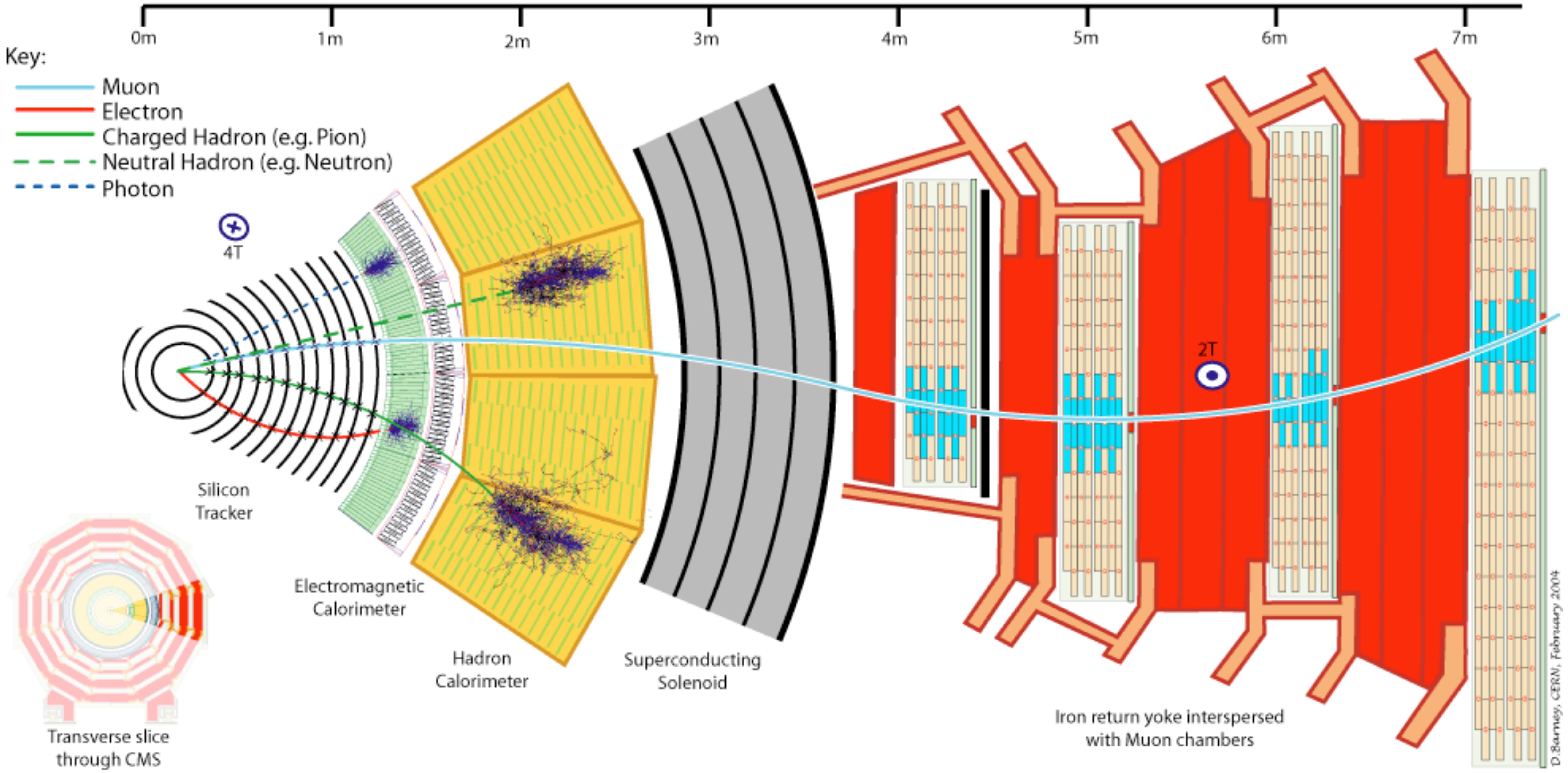
$p(z_1 \mid \theta)$

controlled approximation of first principles

$p(z_2 \mid z_1, \nu_1)$

phenomenological model

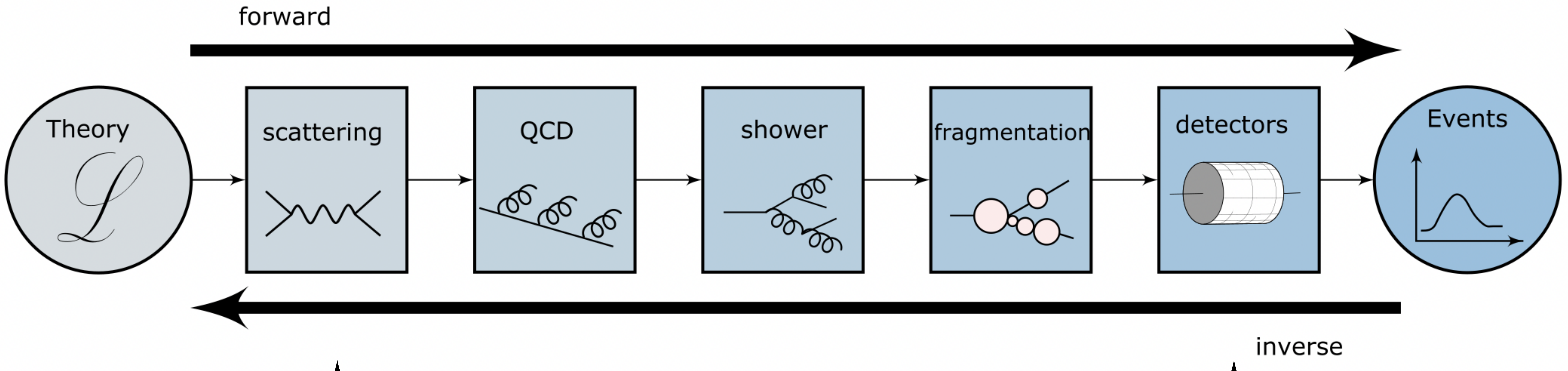
$p(z_3 \mid z_2, \nu_2)$



Detector Simulation $\mathbf{p(x \mid z3, \nu_3)}$:

- detailed engineering (eg. CAD)
- in situ measurements of temperature, magnetic field, alignment, calibration constants
- first-principles description of interaction of particles with matter
- measured interaction of particles with matter

Simulation for LHC



What we
want to know

What we
measure or simulate

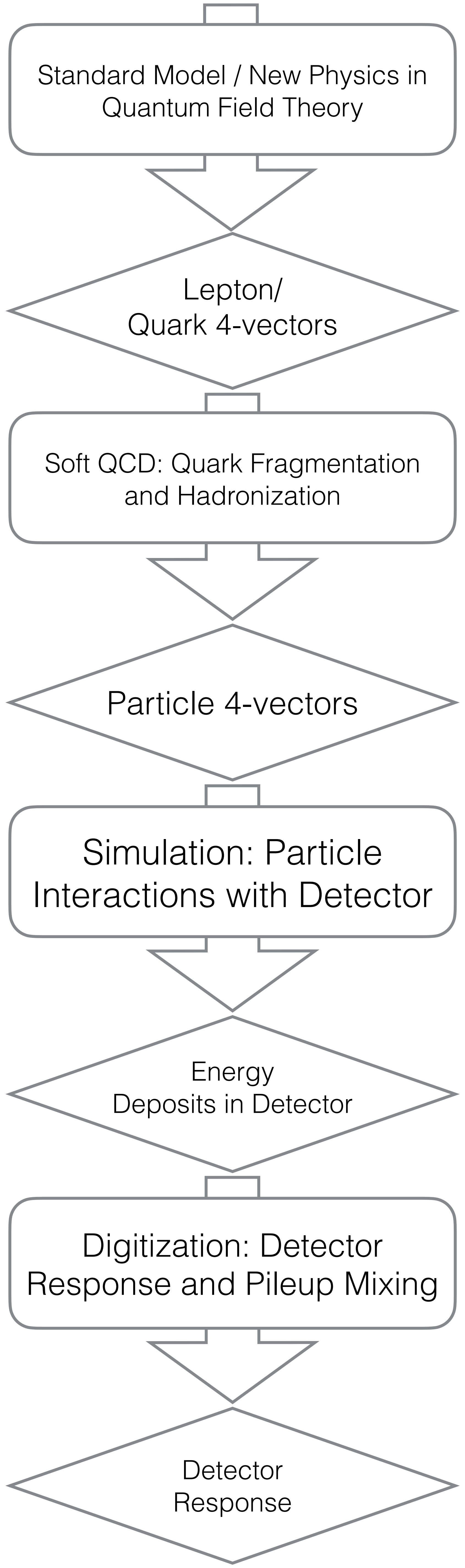
Algorithmic Simulation

- Physics is all about establishing a very precise "model" of the underlying phenomena… so in general ***we can model our data very well.***

- For example for LHC we do ***multi-step ab-initio simulations***:

  1. ***Generation***: Standard Model and New Physics are expressed in language of Quantum Field Theory.
     - ➡ Feynman Diagrams simplify perturbative prediction of HEP interactions among the most fundamental particles (leptons, quarks)
  2. ***Hadronization***: Quarks turn to jets of particles via Quantum Chromodynamics (QCD) at energies where theory is too strong to compute perturbatively.
     - ➡ Use semi-empirical models tuned to Data.
  3. ***Simulation***: Particles interact with the Detector via stochastic processes
     - ➡ Use detailed Monte Carlo integration over the "micro-physics"
  4. ***Digitization***: Ultimately the energy deposits lead to electronic signals in the O(100 Million) channels of the detector.
     - ➡ Model using test beam data and calibrations.

- Output is fed through ***same reconstruction as real data***.



Likelihood Approximations

- Need $P(\{x_d\}|\theta)$ of an observed event (*i*). The better we do, the more sensitive our measurements.
- Steps 2 (Hadronization) and 3 (Simulation) can only be done in the ***forward mode***…
  - ➡ ***cannot evaluate the likelihood.***
- So we simulate a lot of events and generally use histograms (a crude *Probability Density Estimator (PDE) technique)*
  - $\{x_d\}$ = {100M Detector Channels} or even { particle 4-vectors } are too high dimensional.
    - Curse of dimensionality… more on this later
  - Instead we derive $\{x_d\}$ = { small set of physics motivated observables } → ***Lose information.***
    - ***Isolate signal*** dominating regions of $\{x_d\}$ → ***Lose efficiency.***
    - Sometimes use ***ML-based classifiers*** to further reduce dimensionality and improve significance
    - ***Profile the likelihood*** in 1 or 2 (ideally uncorrelated) observables.