

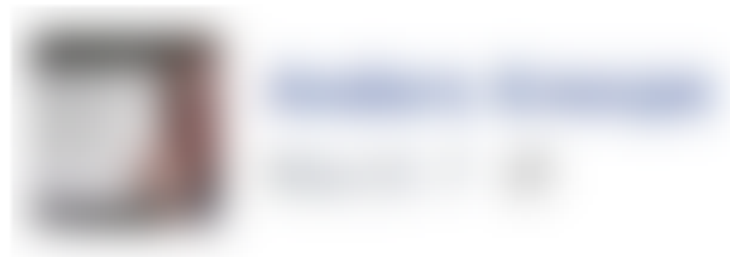
Generative Modeling: The Cure for the Common Systematic



Michael Betancourt
@betanalpha
Centre for Research
in Statistical Methodology,
University of Warwick

PhyStat-U
The Kavli Institute for the Physics
and Mathematics of the Universe,
Kashiwa, Japan
May 31, 2016

“Systematic uncertainties” are ubiquitous in real analyses but, frustratingly, absent from most statistical pedagogy.



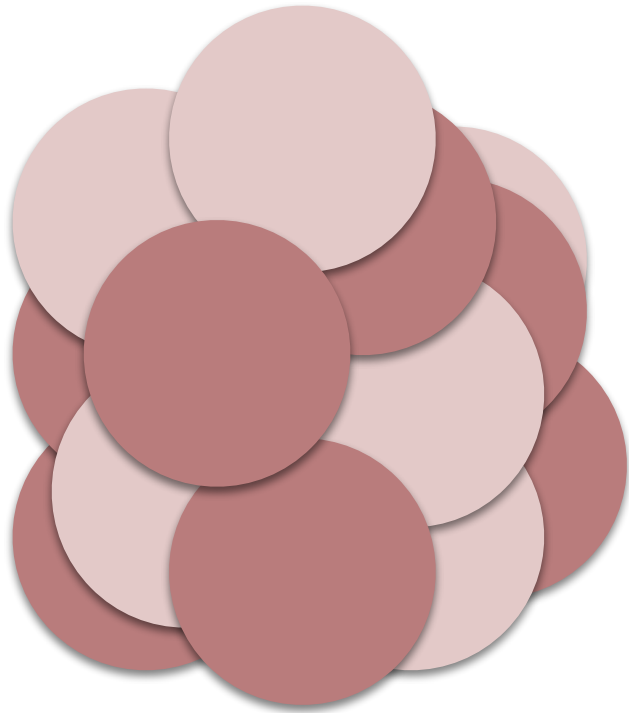
There is no right way to calculate systematic uncertainties; there are only wrong ways.

and 2 others like this.

Share

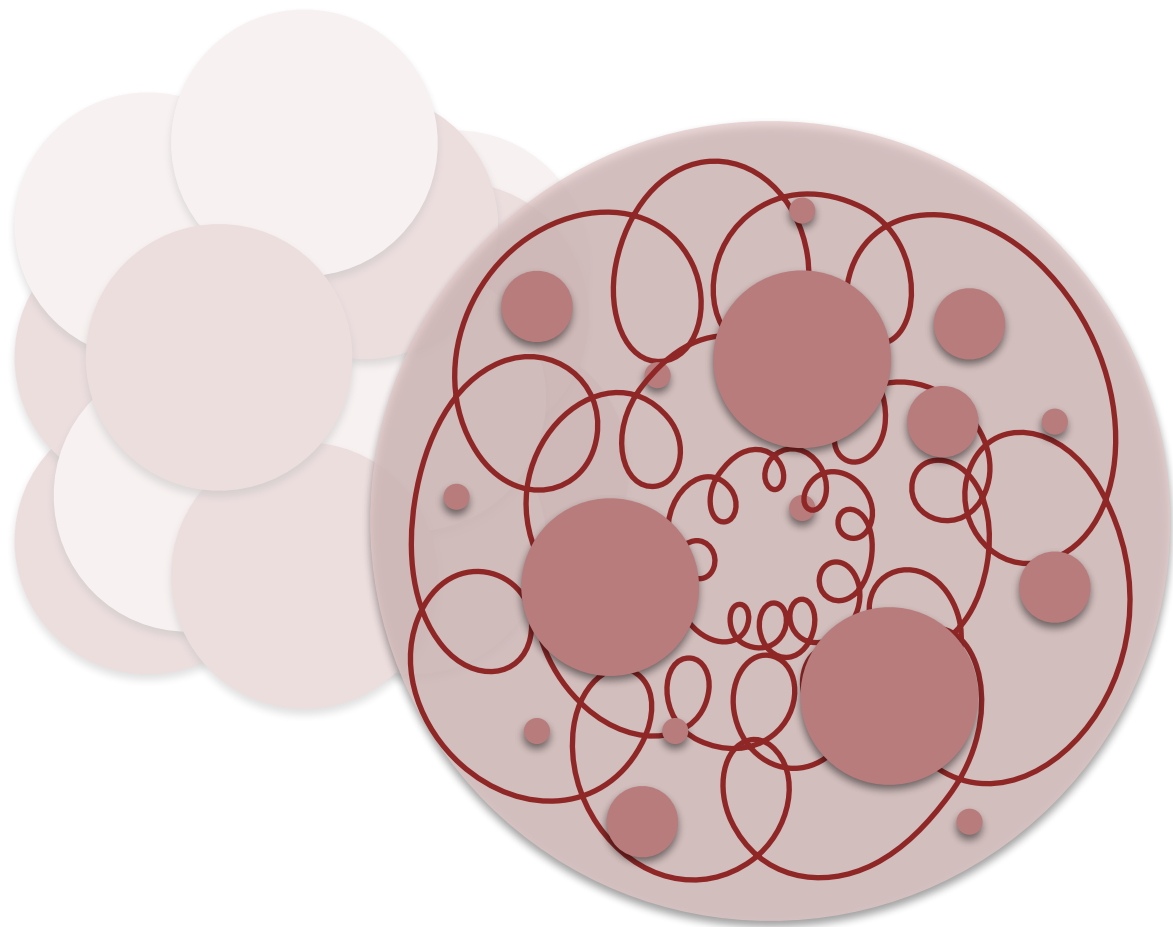
6 3

The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.



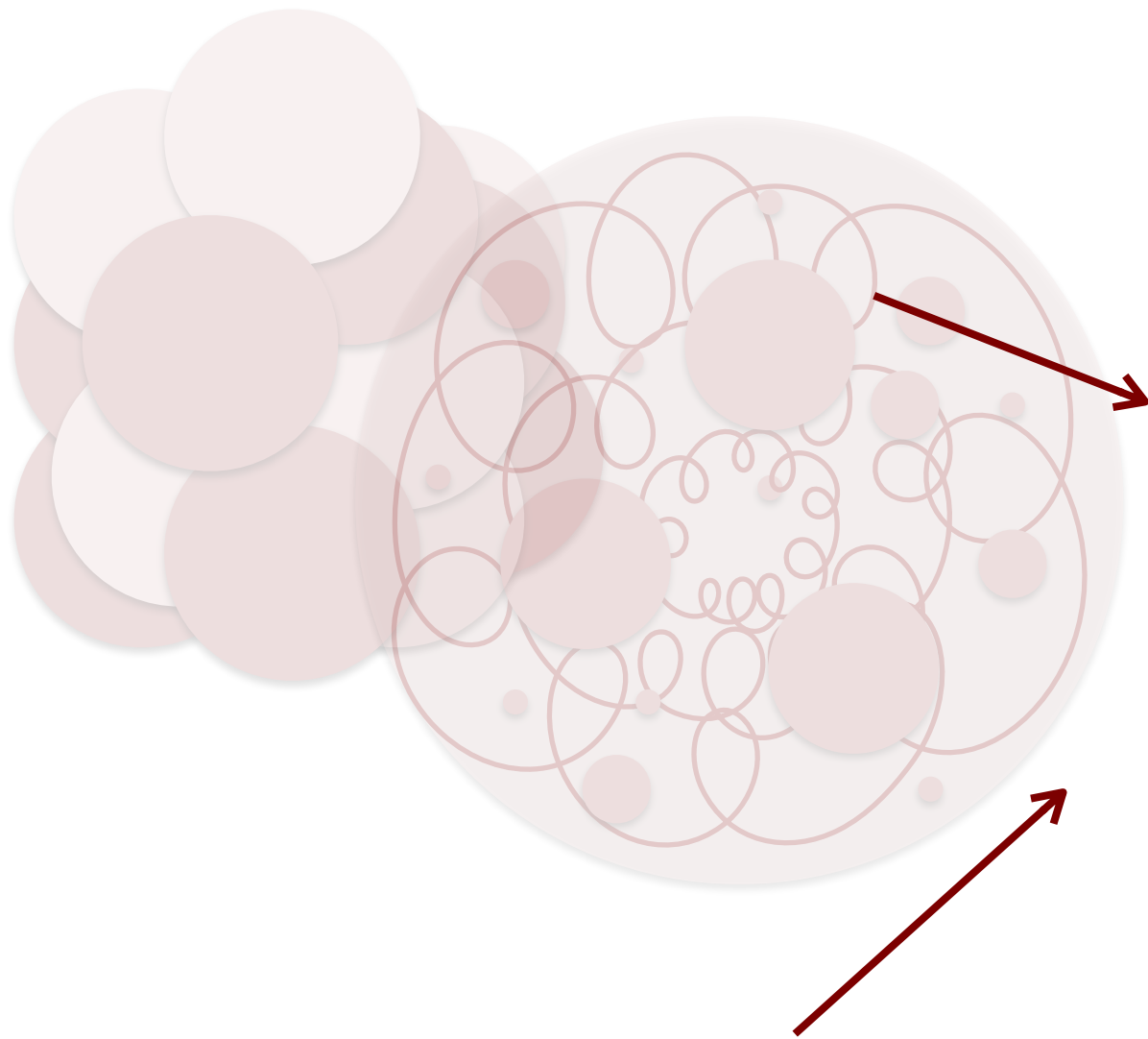
$$\theta_{\text{th}} \rightarrow FS \rightarrow \mathcal{D}$$

The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.



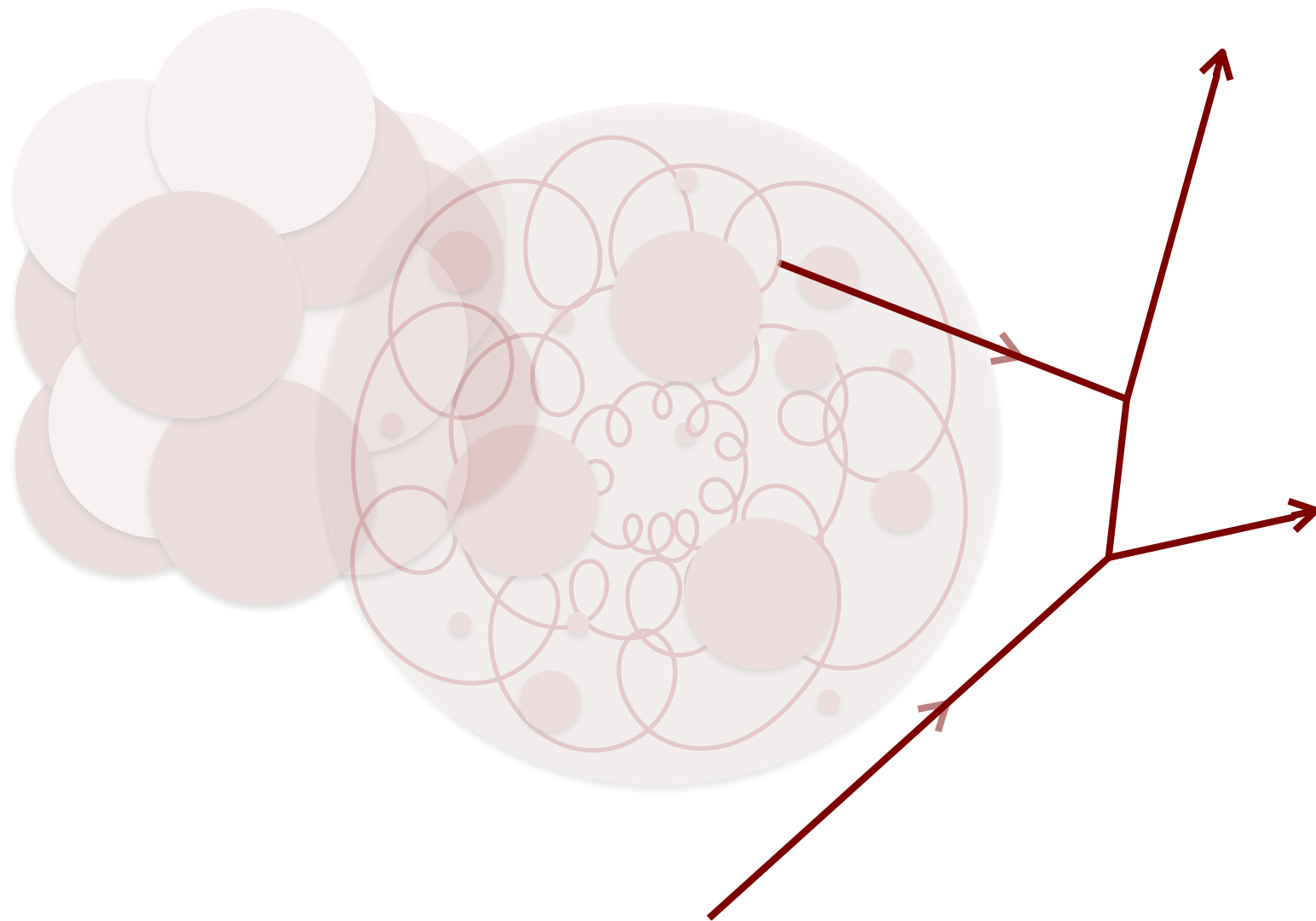
$$\theta_{\text{th}} \rightarrow FS \rightarrow \mathcal{D}$$

The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.



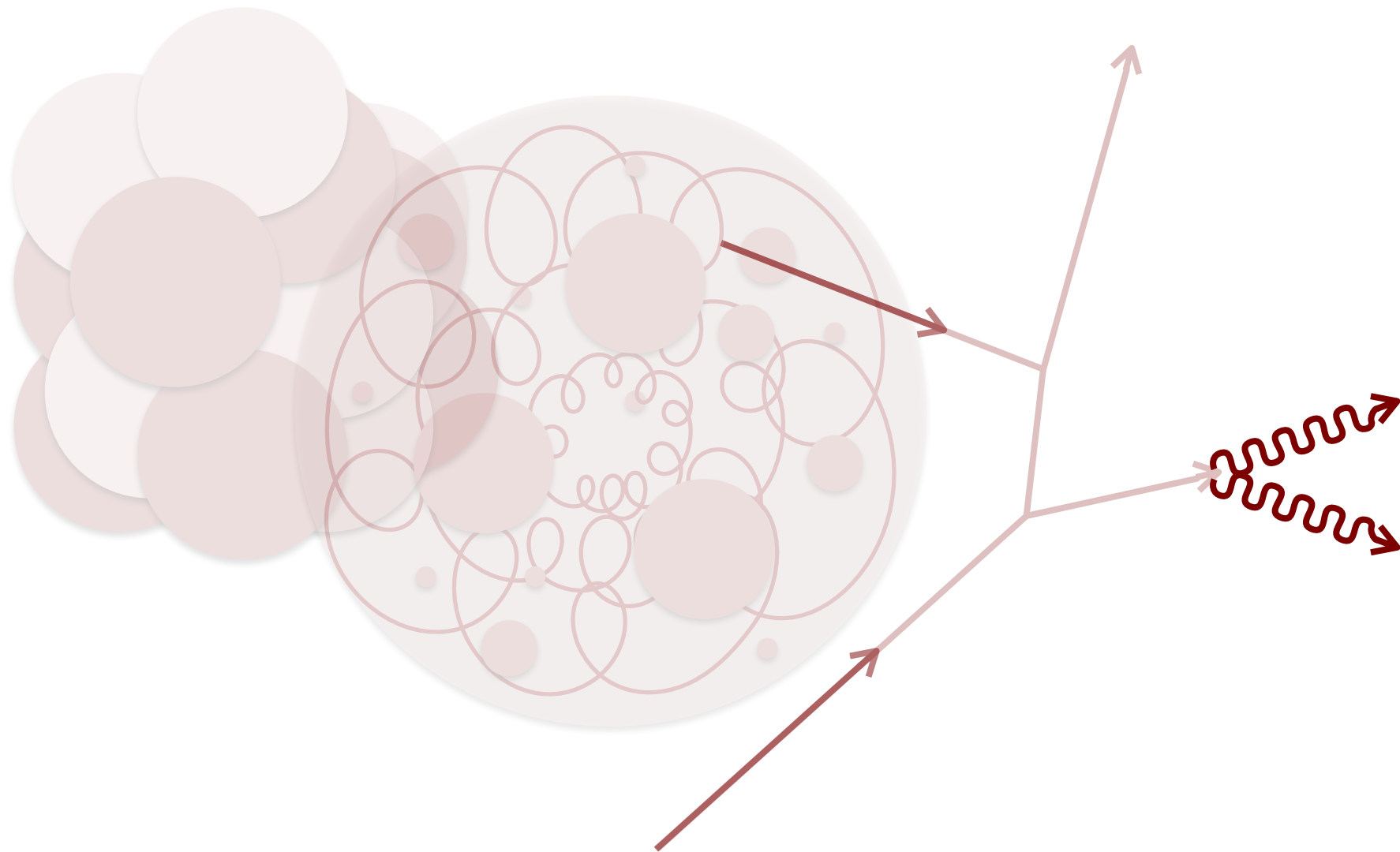
$$\theta_{\text{th}} \rightarrow FS \rightarrow \mathcal{D}$$

The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.



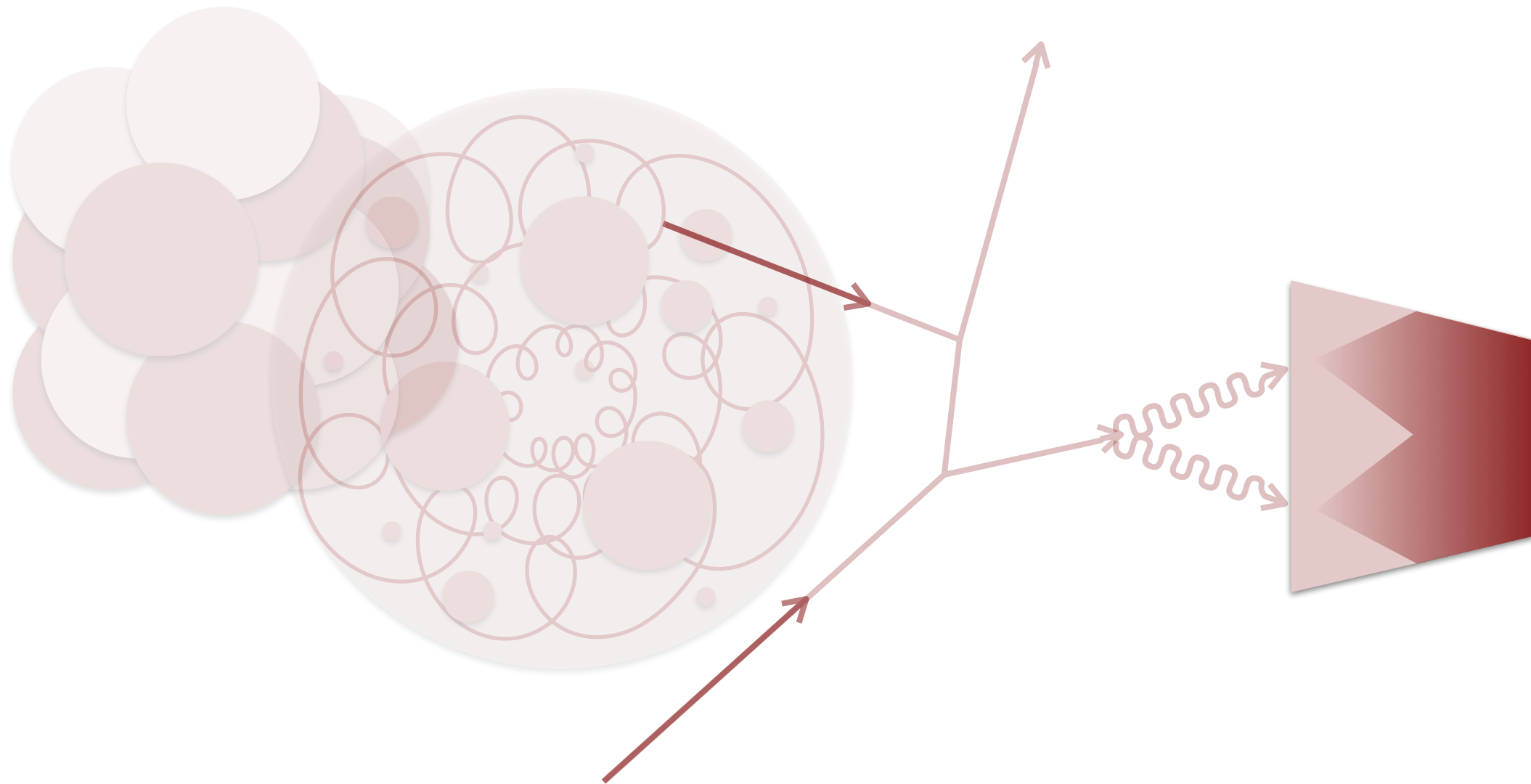
$$\theta_{\text{th}} \rightarrow FS \rightarrow \mathcal{D}$$

The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.



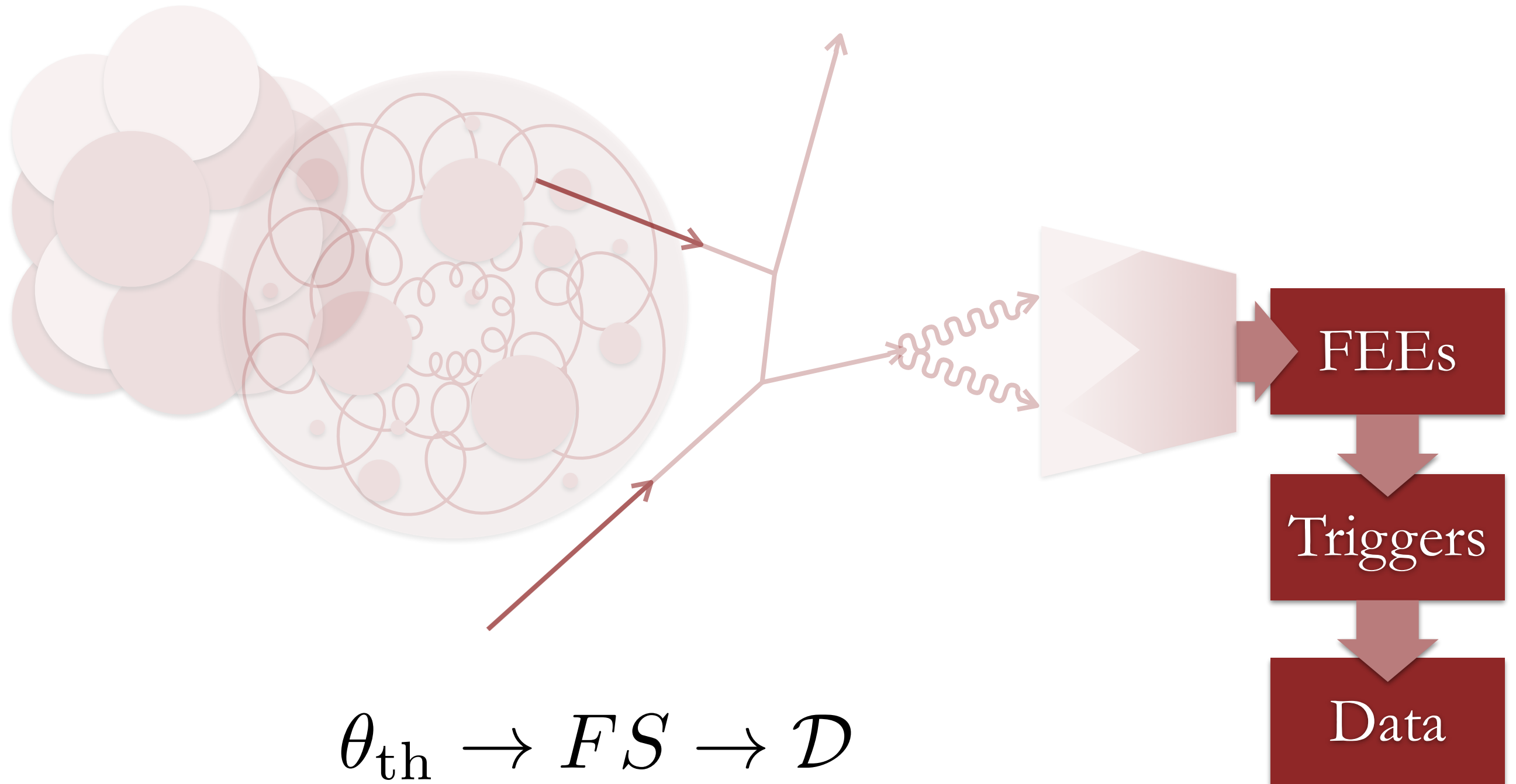
$$\theta_{\text{th}} \rightarrow FS \rightarrow \mathcal{D}$$

The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.

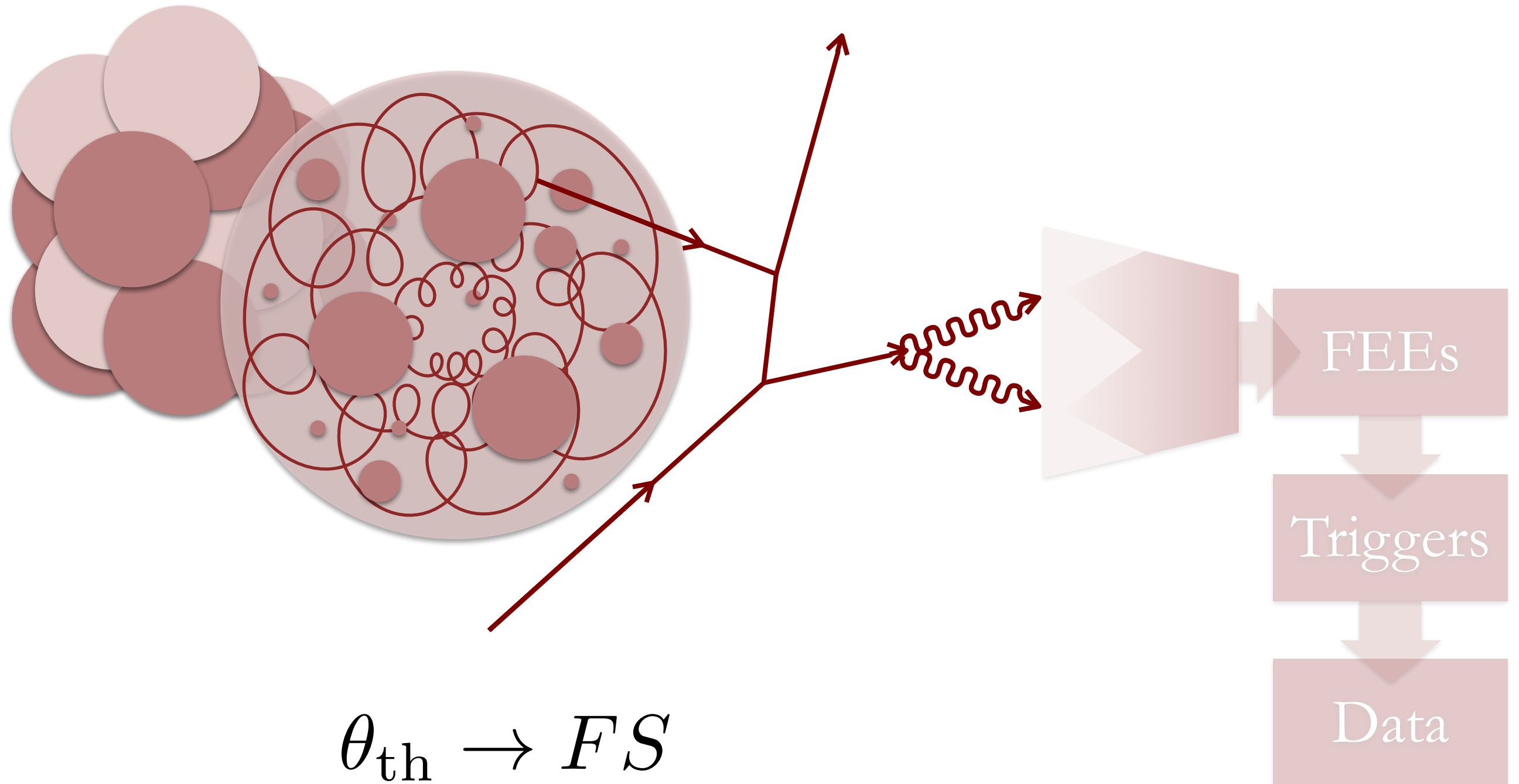


$$\theta_{\text{th}} \rightarrow FS \rightarrow \mathcal{D}$$

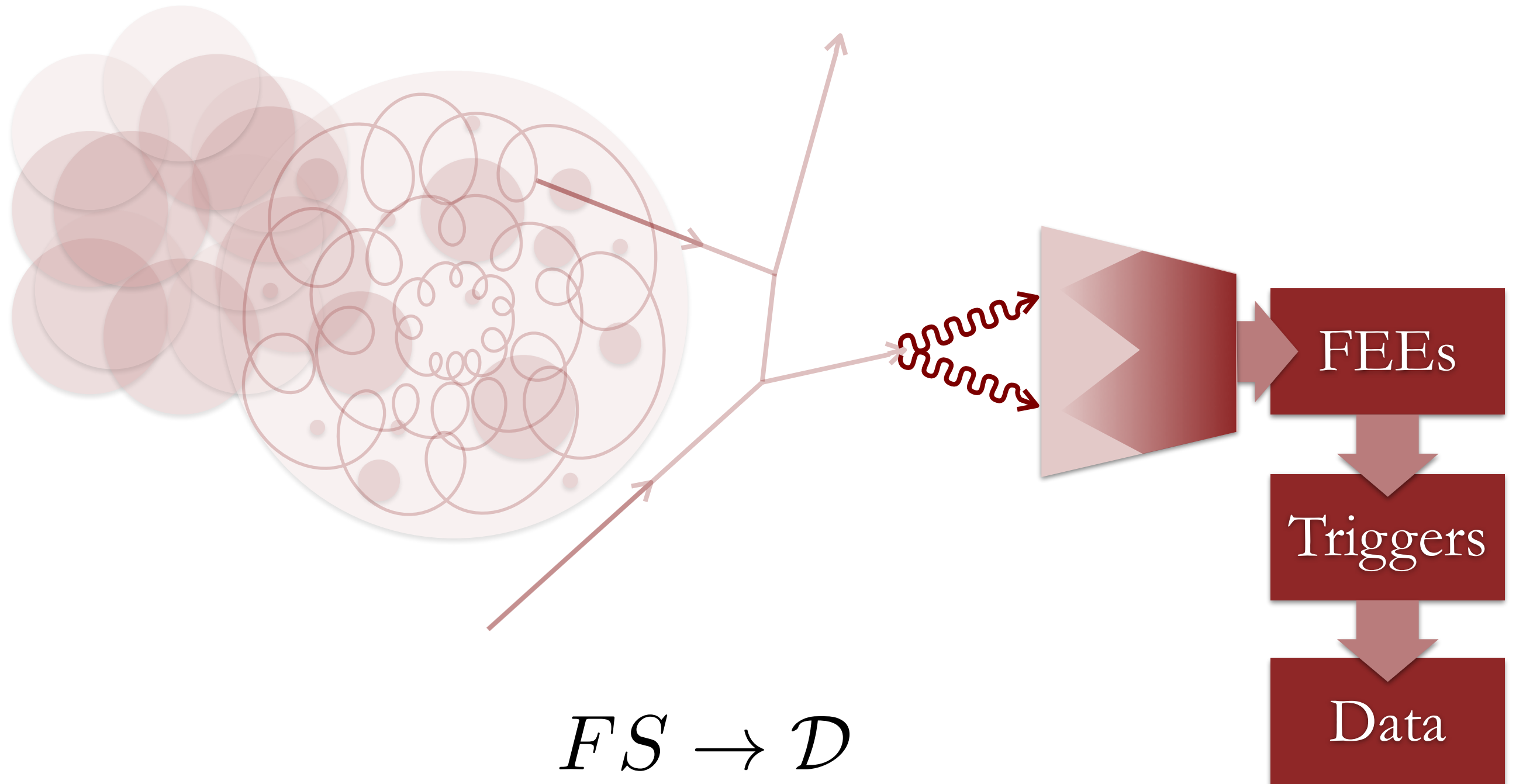
The sophistication of physics models relative to the simple models presented in statistics classes doesn't help.



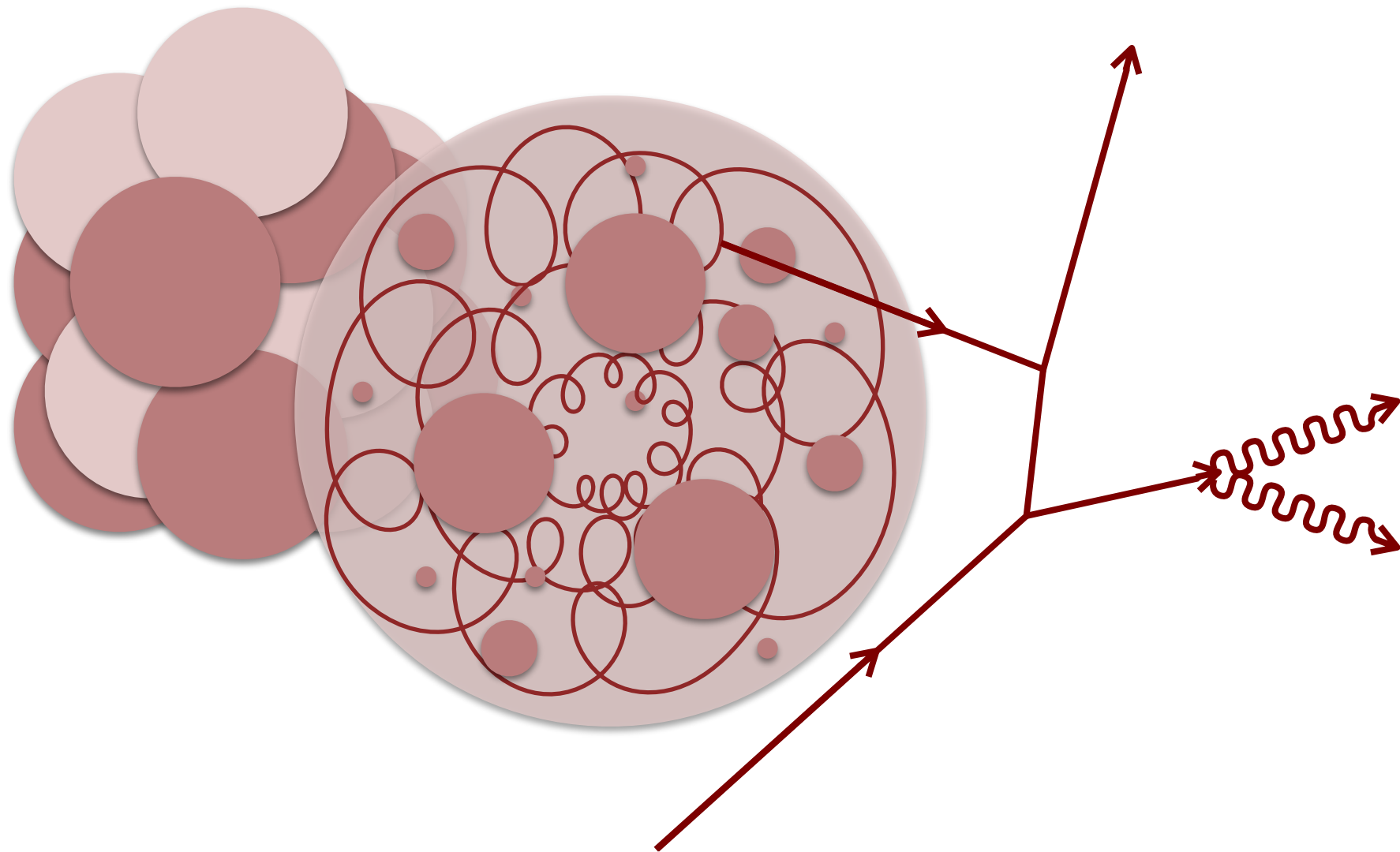
To simplify analyses, these models are decomposed into “theoretical” models and “experimental” models.



To simplify analyses, these models are decomposed into “theoretical” models and “experimental” models.

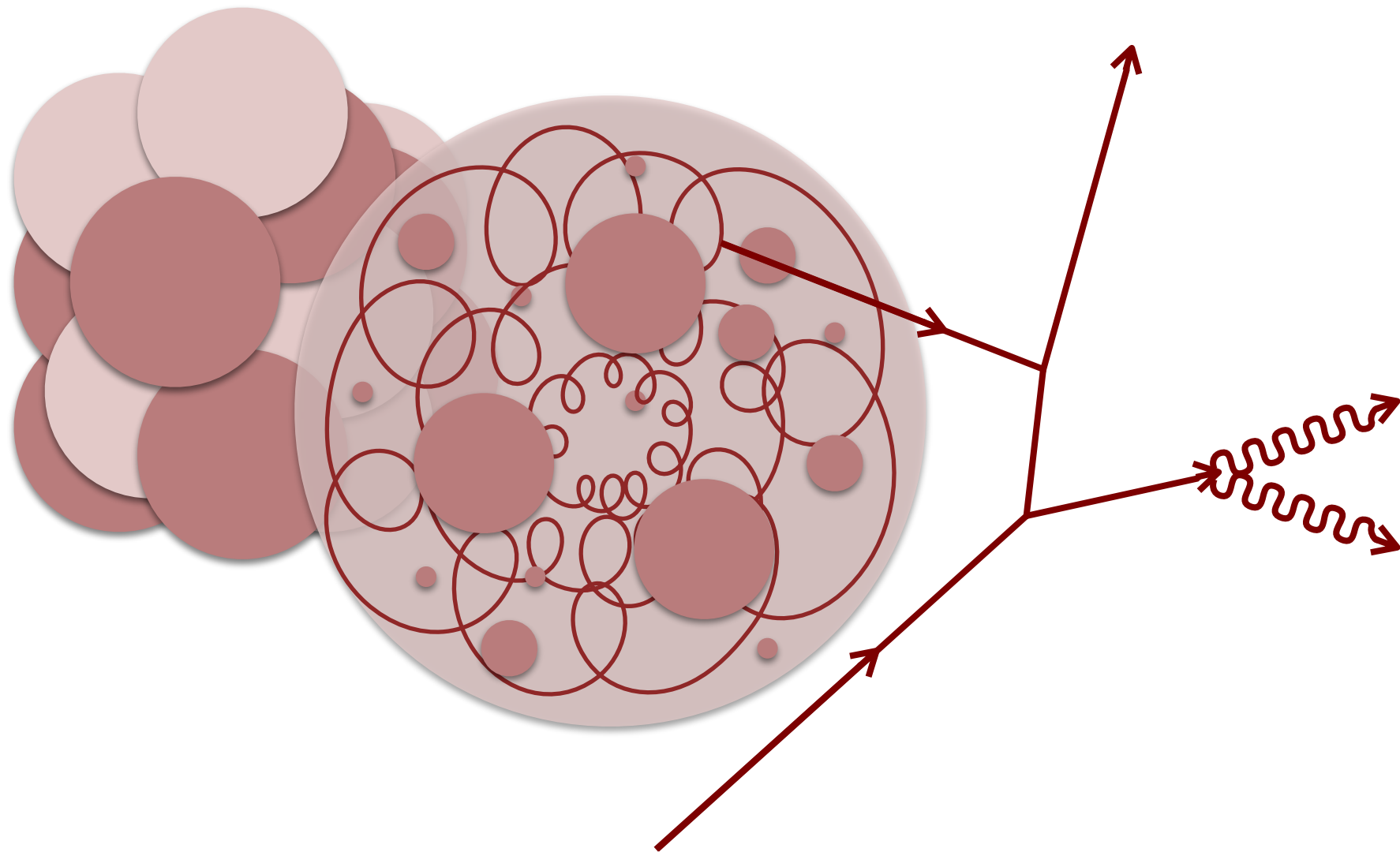


Theoretical models focus on the generation of the final state from the underlying physics.



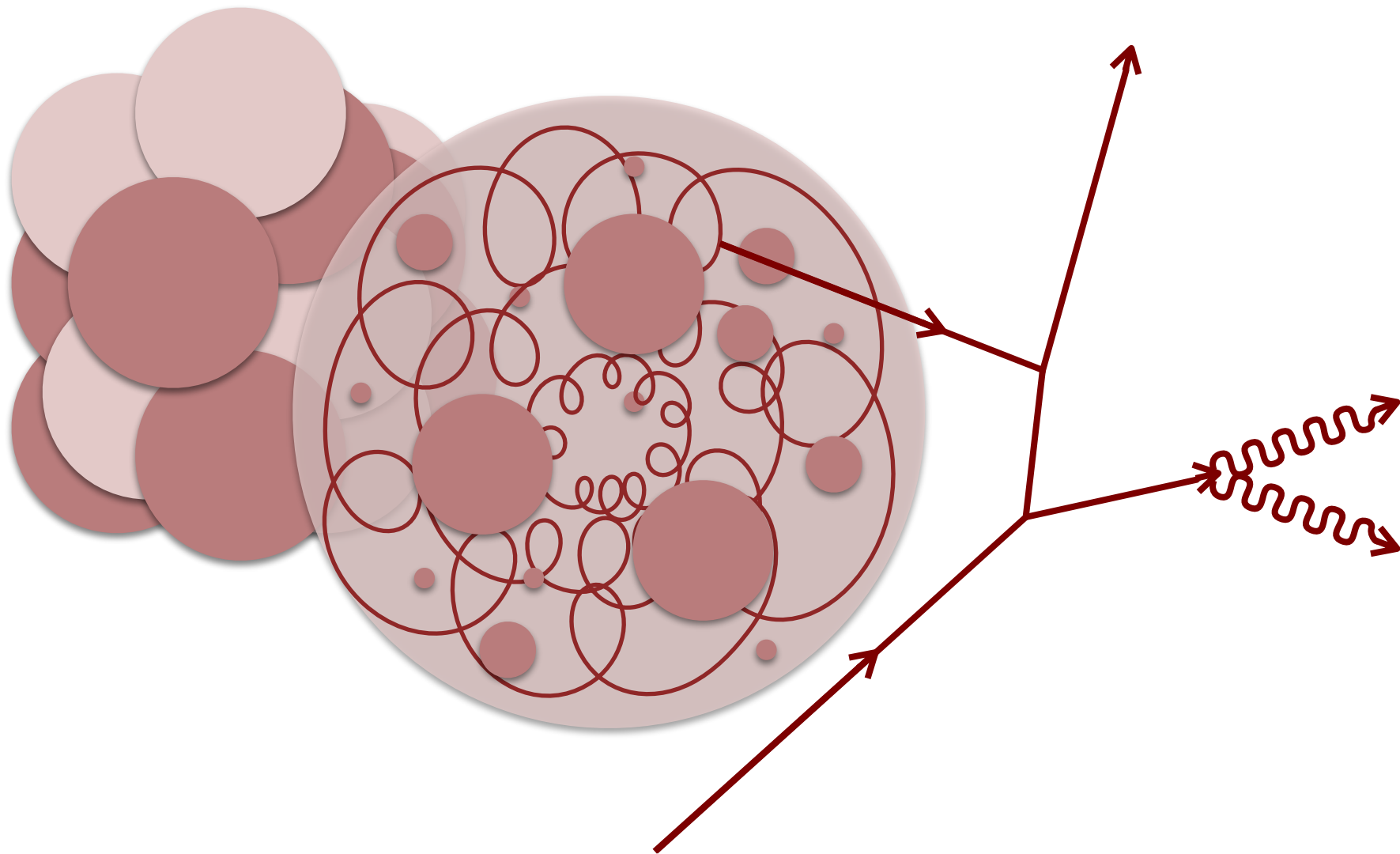
$$FS(\theta_{\text{th}})$$

But there are many nuisance parameters to consider, and quantifying the systematic uncertainties can be awkward.



$$FS(\theta_{\text{th}}, \tilde{\theta}_{\text{th}})$$

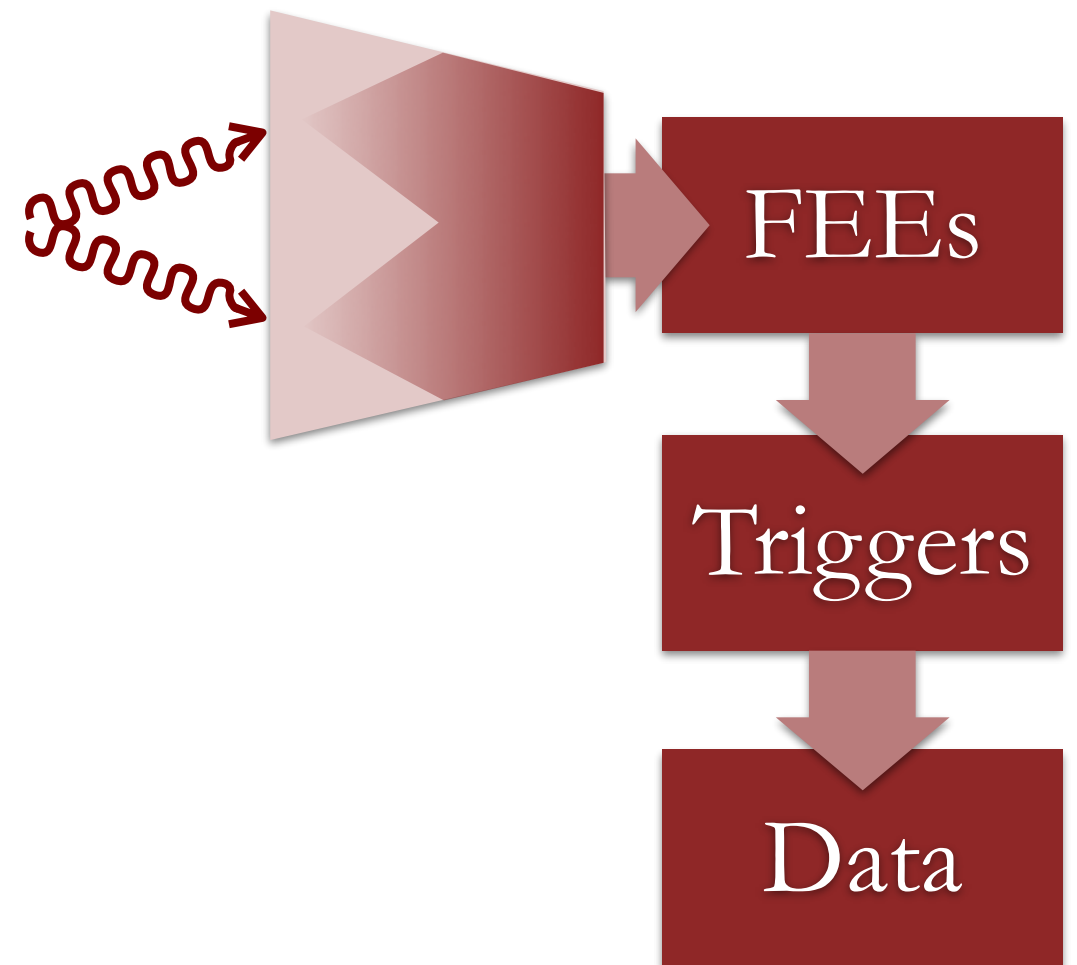
But there are many nuisance parameters to consider, and quantifying the systematic uncertainties can be awkward.



$$FS(\theta_{\text{th}}, \tilde{\theta}_{\text{th}}) \approx FS(\theta_{\text{th}}) \pm \delta FS_{\text{th}}$$

Experimental models try to deconvolve the measured data to recover the final state.

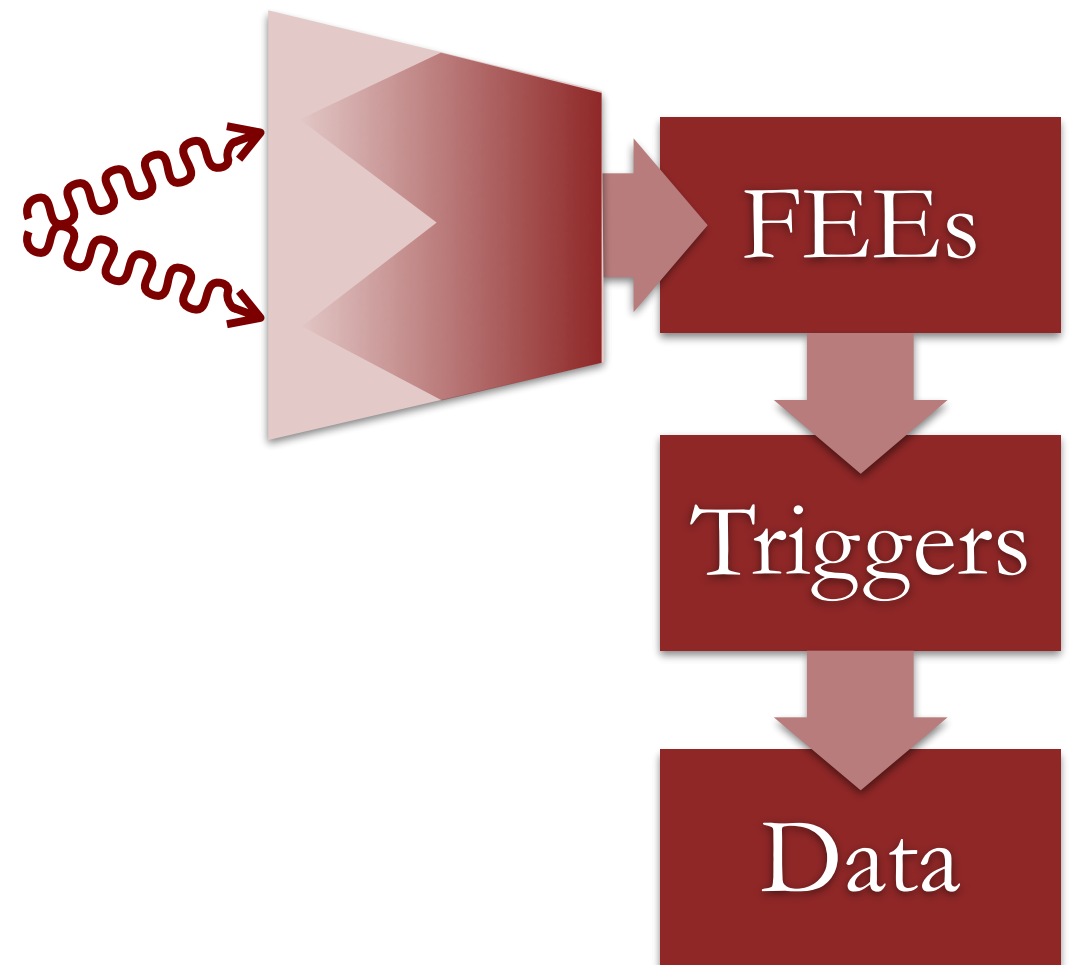
$$\mathcal{D}(FS, \theta_{\text{exp}}) \rightarrow FS(\mathcal{D}, \theta_{\text{exp}})$$



Experimental models try to deconvolve the measured data to recover the final state.

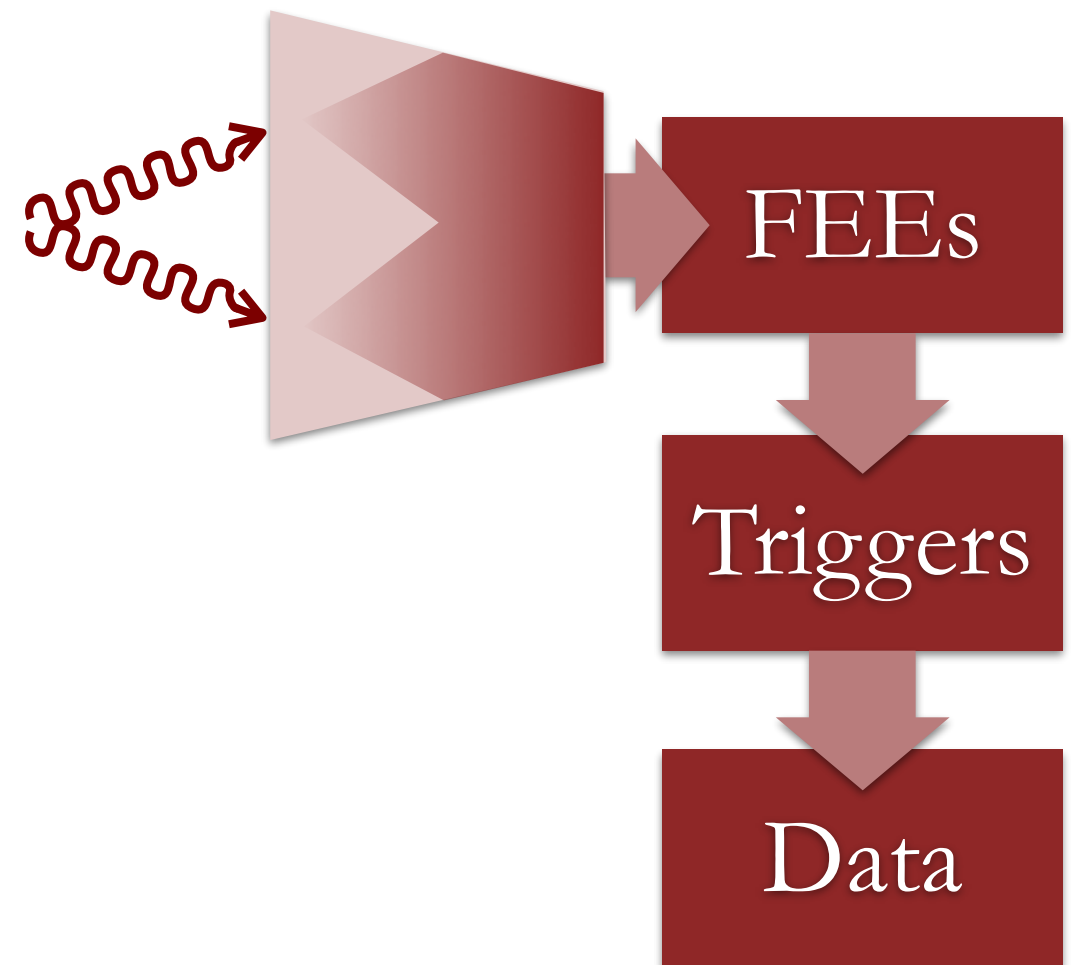
$$FS(\mathcal{D}, \theta_{\text{exp}}) \approx FS(\mathcal{D}) \pm \delta FS_{\text{stat}}$$

$$\mathcal{D}(FS, \theta_{\text{exp}}) \rightarrow FS(\mathcal{D}, \theta_{\text{exp}})$$



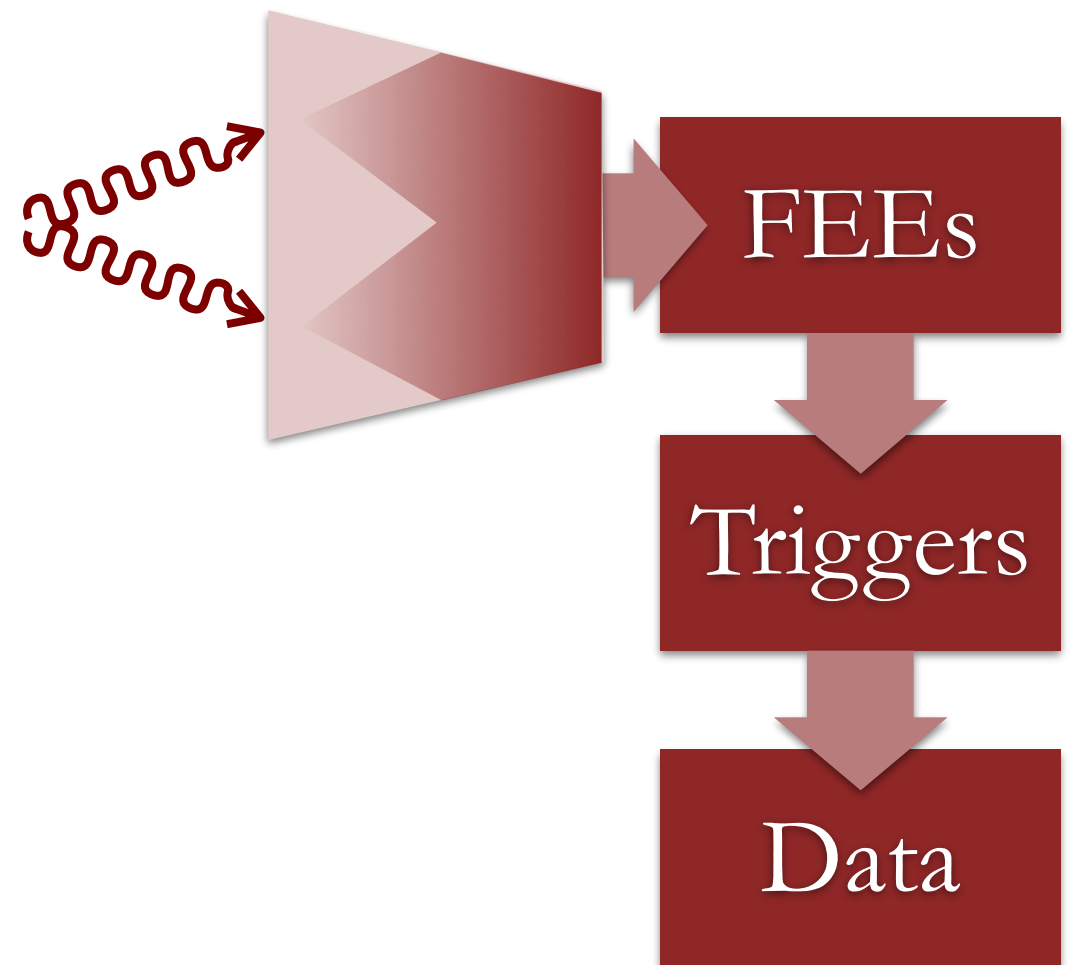
This process also has nuisance parameters, and we have to consider *both* statistical and systematic uncertainties.

$$FS(\mathcal{D}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$$

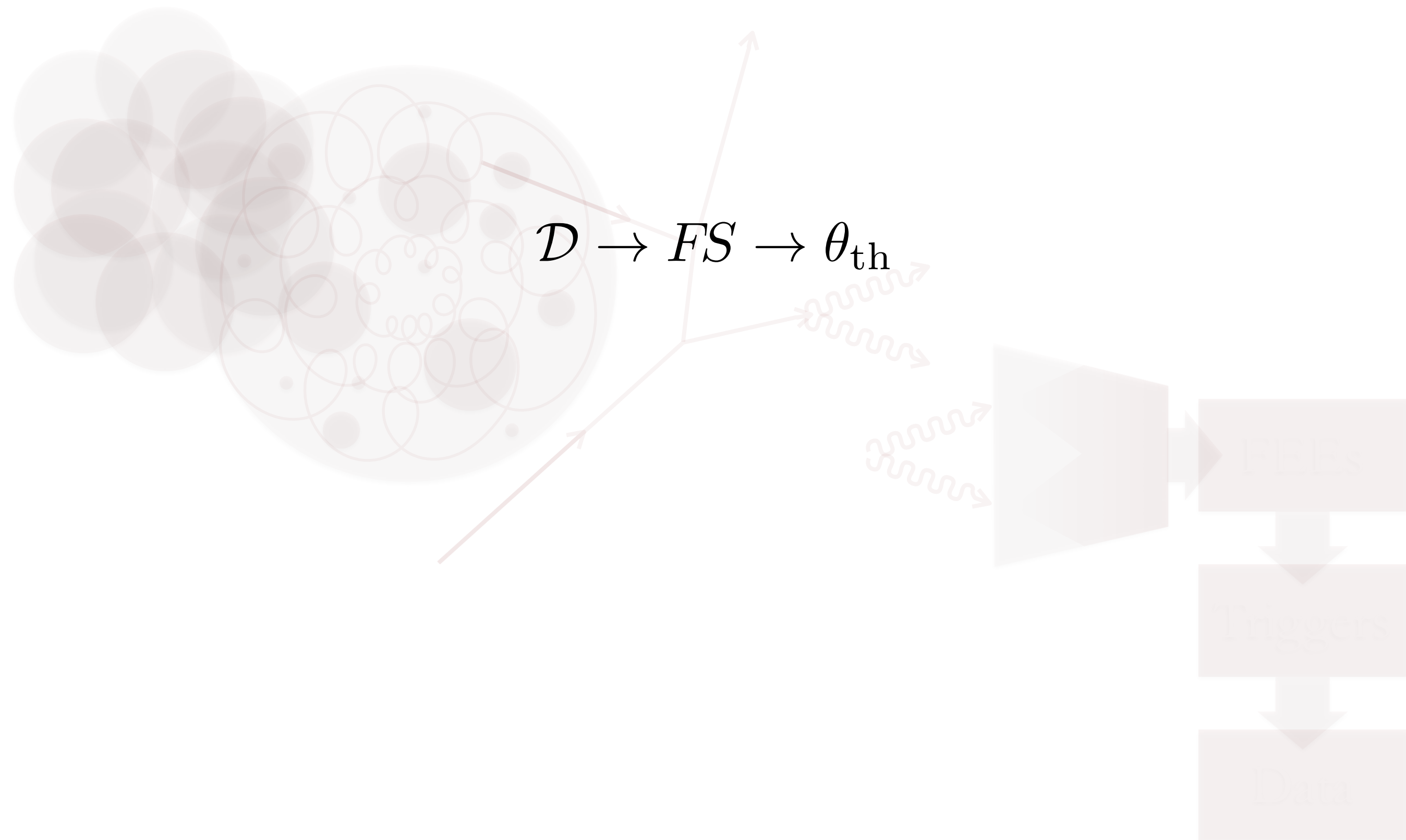


This process also has nuisance parameters, and we have to consider *both* statistical and systematic uncertainties.

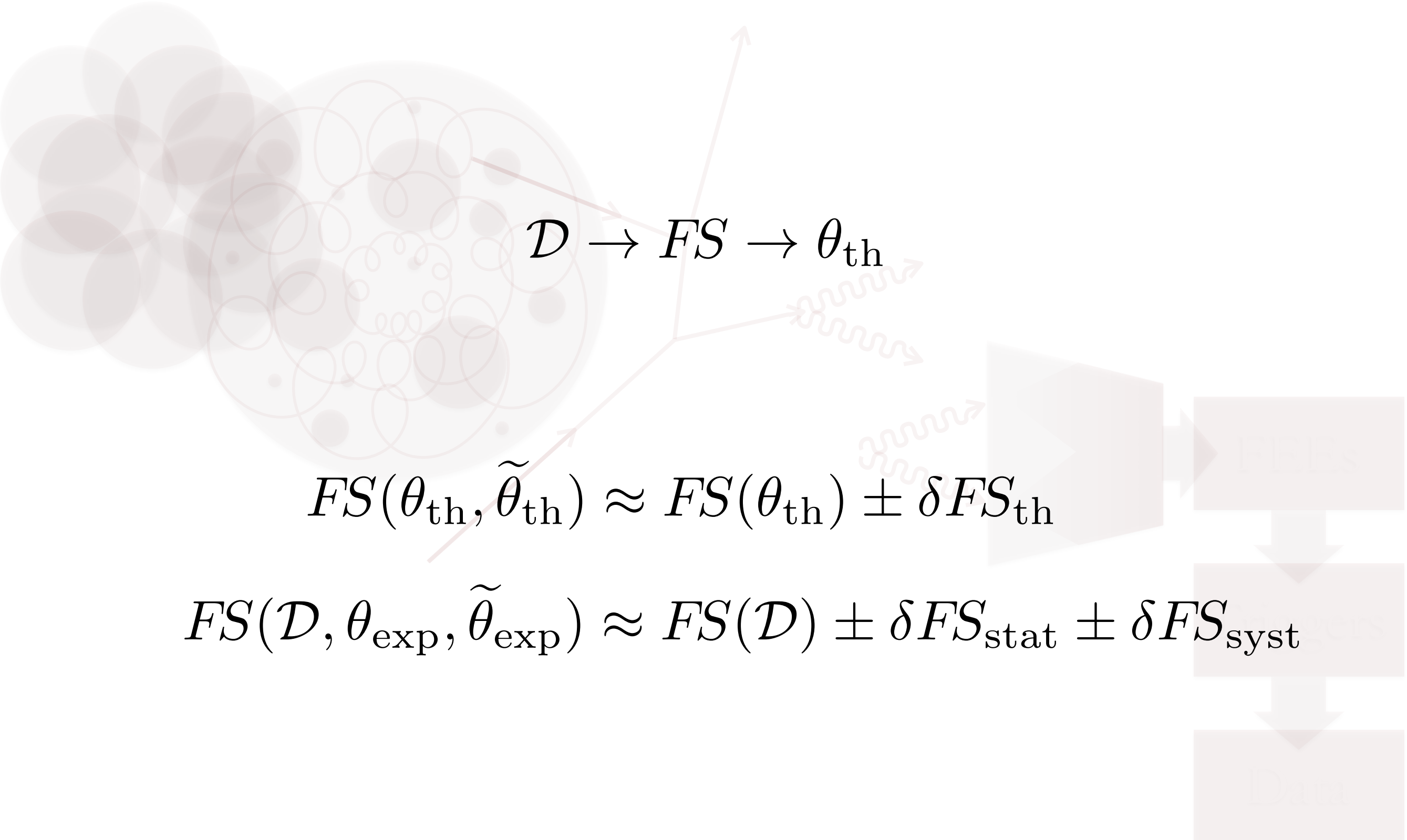
$$FS(\mathcal{D}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}}) \approx FS(\mathcal{D}) \pm \delta FS_{\text{stat}} \pm \delta FS_{\text{syst}}$$



What we really want to know is how the data informs the underlying physics, so how do we combine these results?



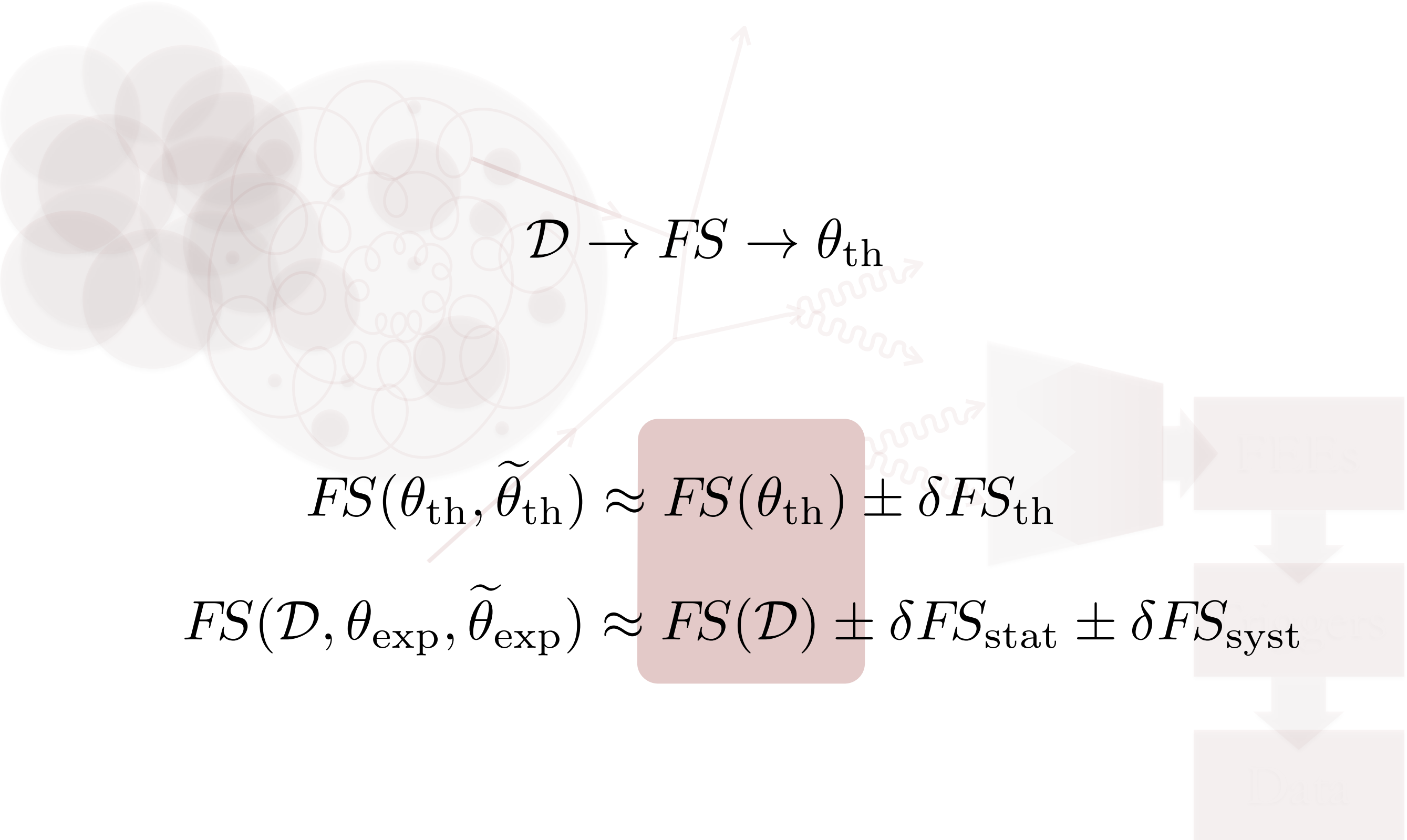
What we really want to know is how the data informs the underlying physics, so how do we combine these results?


$$\mathcal{D} \rightarrow FS \rightarrow \theta_{\text{th}}$$

$$FS(\theta_{\text{th}}, \tilde{\theta}_{\text{th}}) \approx FS(\theta_{\text{th}}) \pm \delta FS_{\text{th}}$$

$$FS(\mathcal{D}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}}) \approx FS(\mathcal{D}) \pm \delta FS_{\text{stat}} \pm \delta FS_{\text{syst}}$$

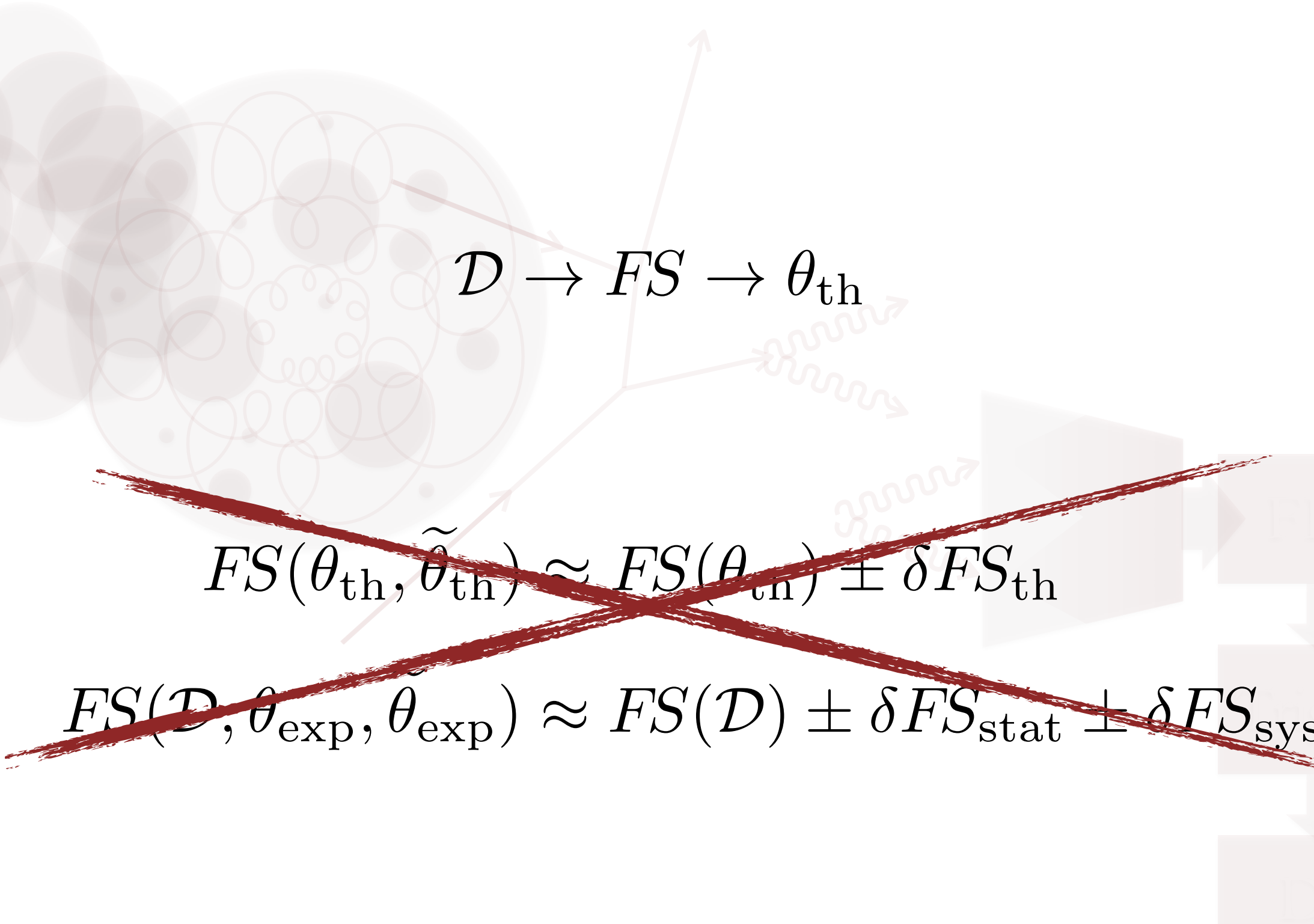
What we really want to know is how the data informs the underlying physics, so how do we combine these results?


$$\mathcal{D} \rightarrow FS \rightarrow \theta_{\text{th}}$$

$$FS(\theta_{\text{th}}, \tilde{\theta}_{\text{th}}) \approx FS(\theta_{\text{th}}) \pm \delta FS_{\text{th}}$$

$$FS(\mathcal{D}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}}) \approx FS(\mathcal{D}) \pm \delta FS_{\text{stat}} \pm \delta FS_{\text{syst}}$$

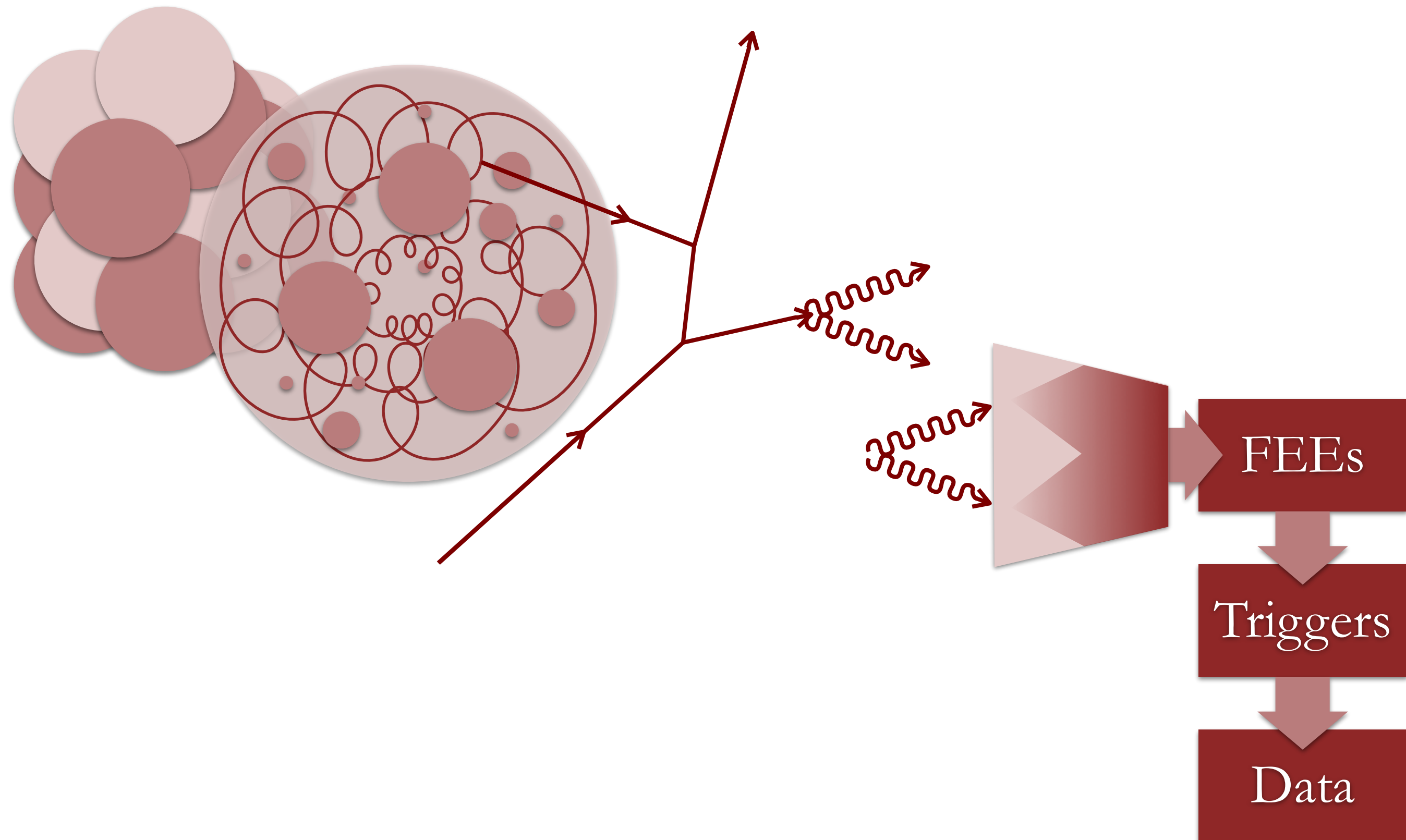
There is no single answer because this is not a valid statistical procedure! It's not supposed to make sense.


$$\mathcal{D} \rightarrow FS \rightarrow \theta_{\text{th}}$$

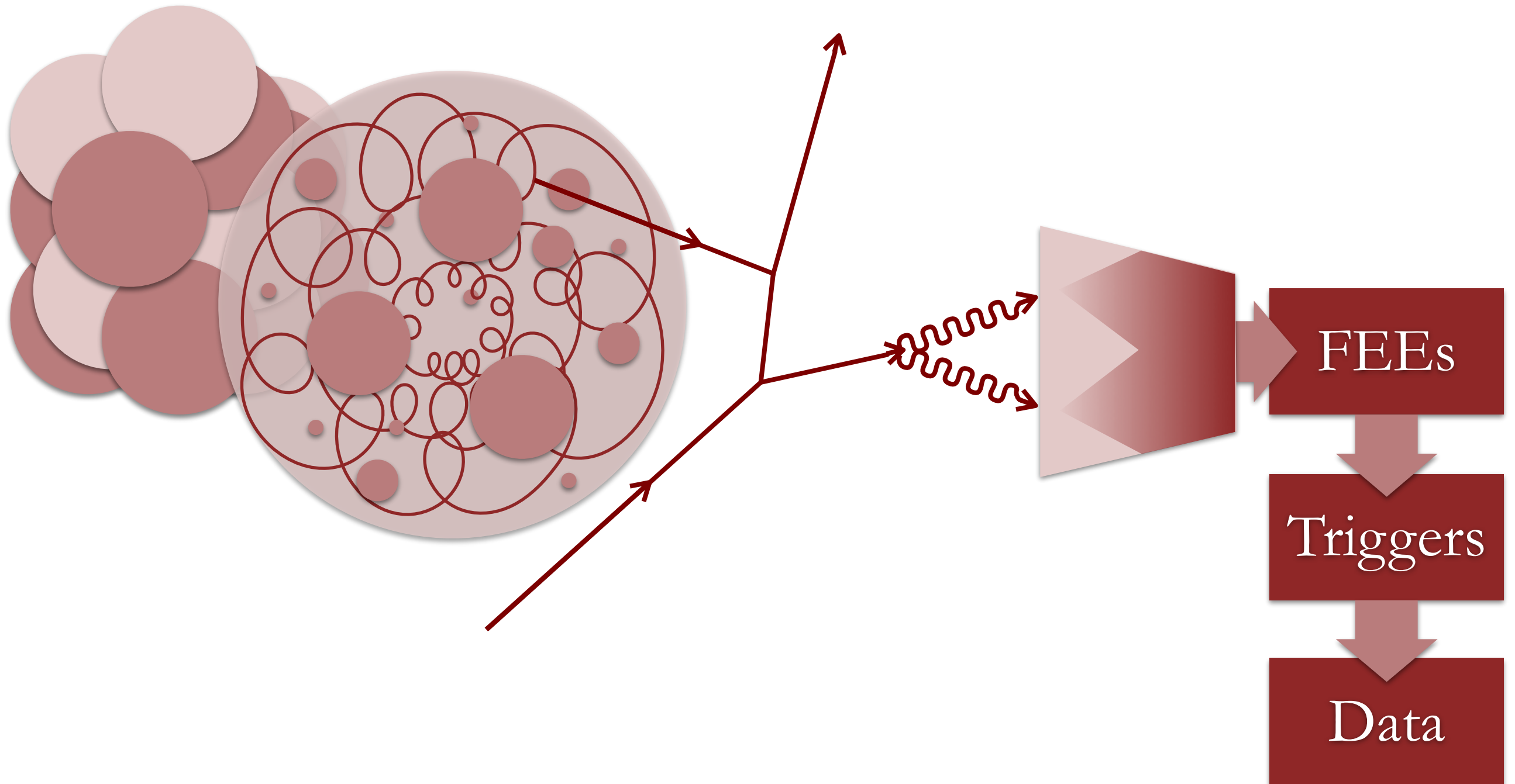
~~$$FS(\theta_{\text{th}}, \tilde{\theta}_{\text{th}}) \approx FS(\theta_{\text{th}}) \pm \delta FS_{\text{th}}$$~~

~~$$FS(\mathcal{D}, \theta_{\text{exp}}, \theta_{\text{exp}}) \approx FS(\mathcal{D}) \pm \delta FS_{\text{stat}} \pm \delta FS_{\text{syst}}$$~~

Proper statistical models do not discriminate between phases of the data generating process.



Proper statistical models do not discriminate between phases of the data generating process.



Nor do they make any distinction between relevant parameters and nuisance parameters.

θ_{th}

$\tilde{\theta}_{\text{exp}}$

$\tilde{\theta}_{\text{th}}$

θ_{exp}

Nor do they make any distinction between relevant parameters and nuisance parameters.

$$(\theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$$

All valid statistical procedures are based on a single joint model of the data generating process, or likelihood.

$$\pi(\mathcal{D} \mid \theta)$$

All valid statistical procedures are based on a single joint model of the data generating process, or likelihood.

$$\pi(\mathcal{D} \mid \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$$

The likelihood self-consistently incorporates the sensitivity of all of these parameters *automatically*.

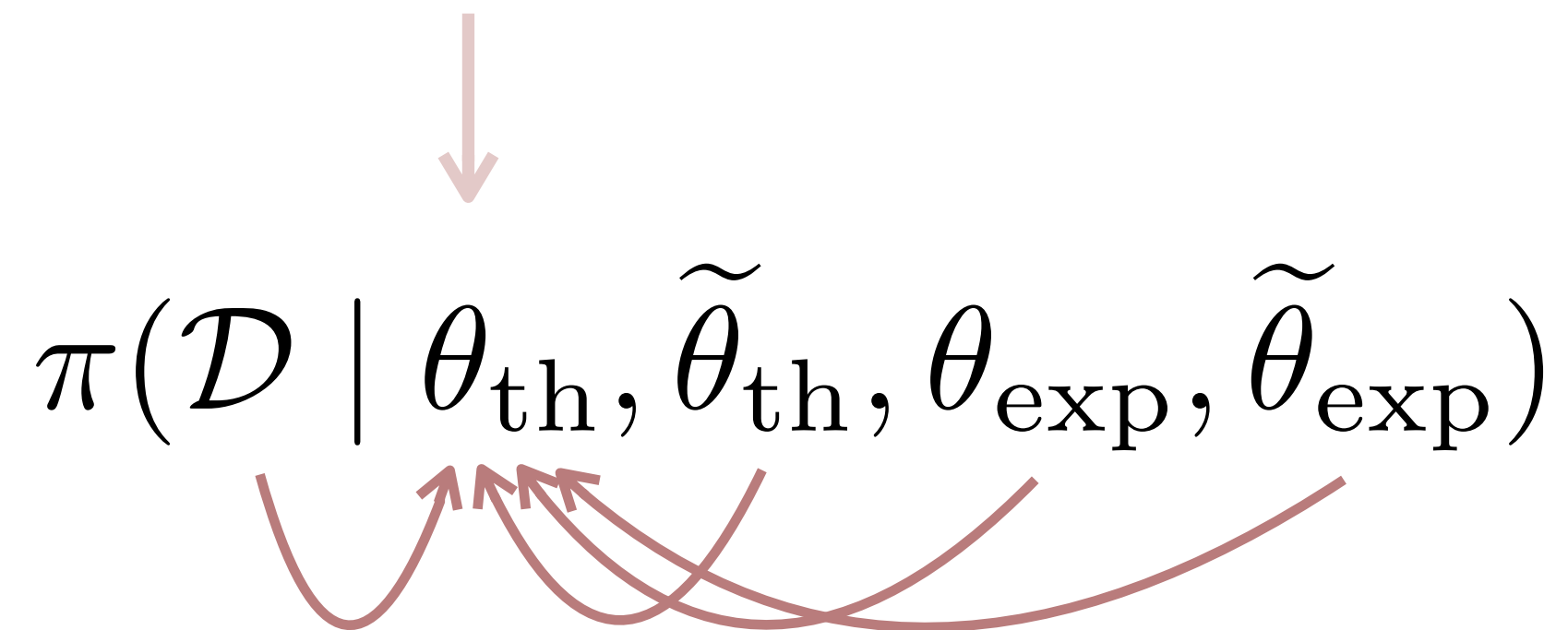


$$\pi(\mathcal{D} \mid \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$$

The likelihood self-consistently incorporates the sensitivity of all of these parameters *automatically*.


$$\pi(\mathcal{D} \mid \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$$

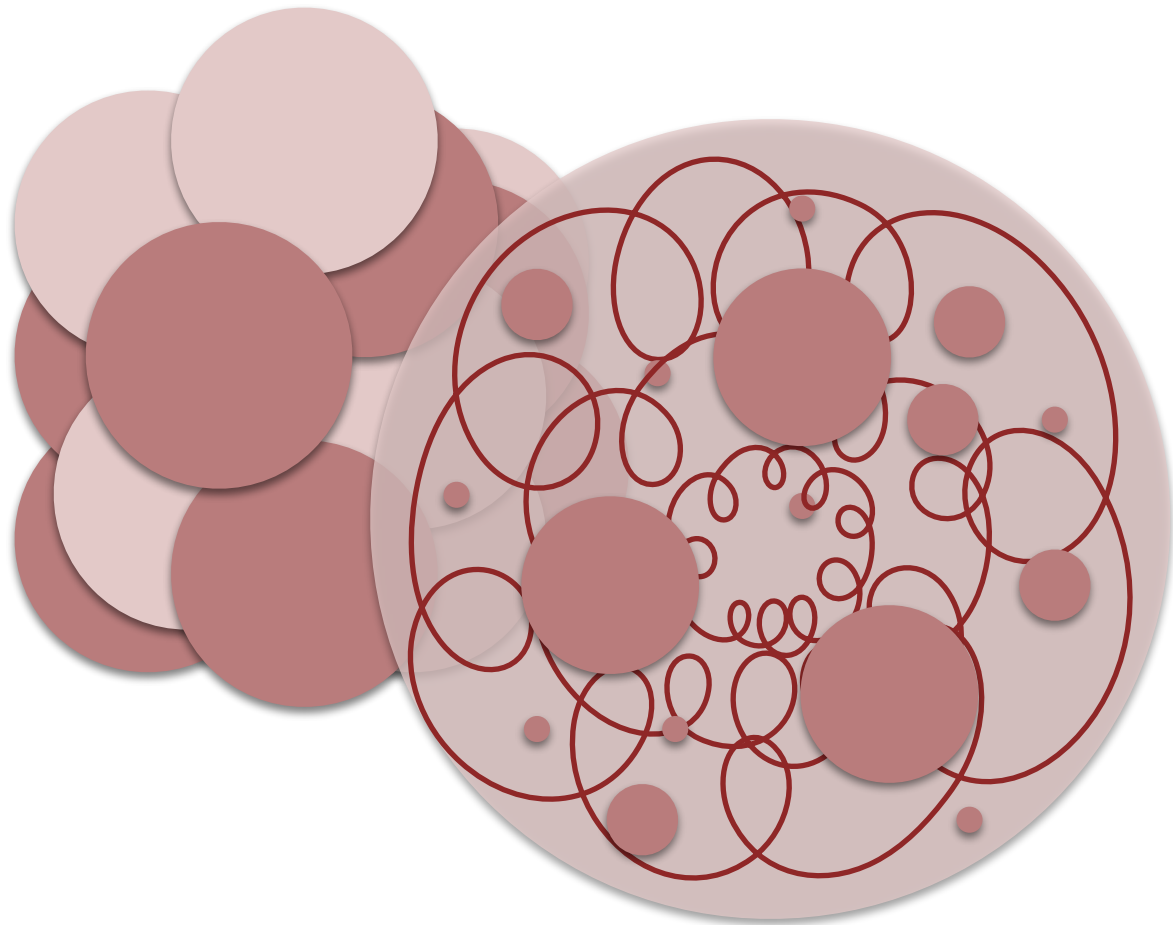
The likelihood self-consistently incorporates the sensitivity of all of these parameters *automatically*.



A diagram illustrating the likelihood function $\pi(\mathcal{D} \mid \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$. A light red arrow points down from the text above to the expression. Below the expression, four curved red arrows point upwards towards the parameters: one from θ_{th} , one from $\tilde{\theta}_{\text{th}}$, one from θ_{exp} , and one from $\tilde{\theta}_{\text{exp}}$. These arrows represent the sensitivity of the likelihood to each parameter.

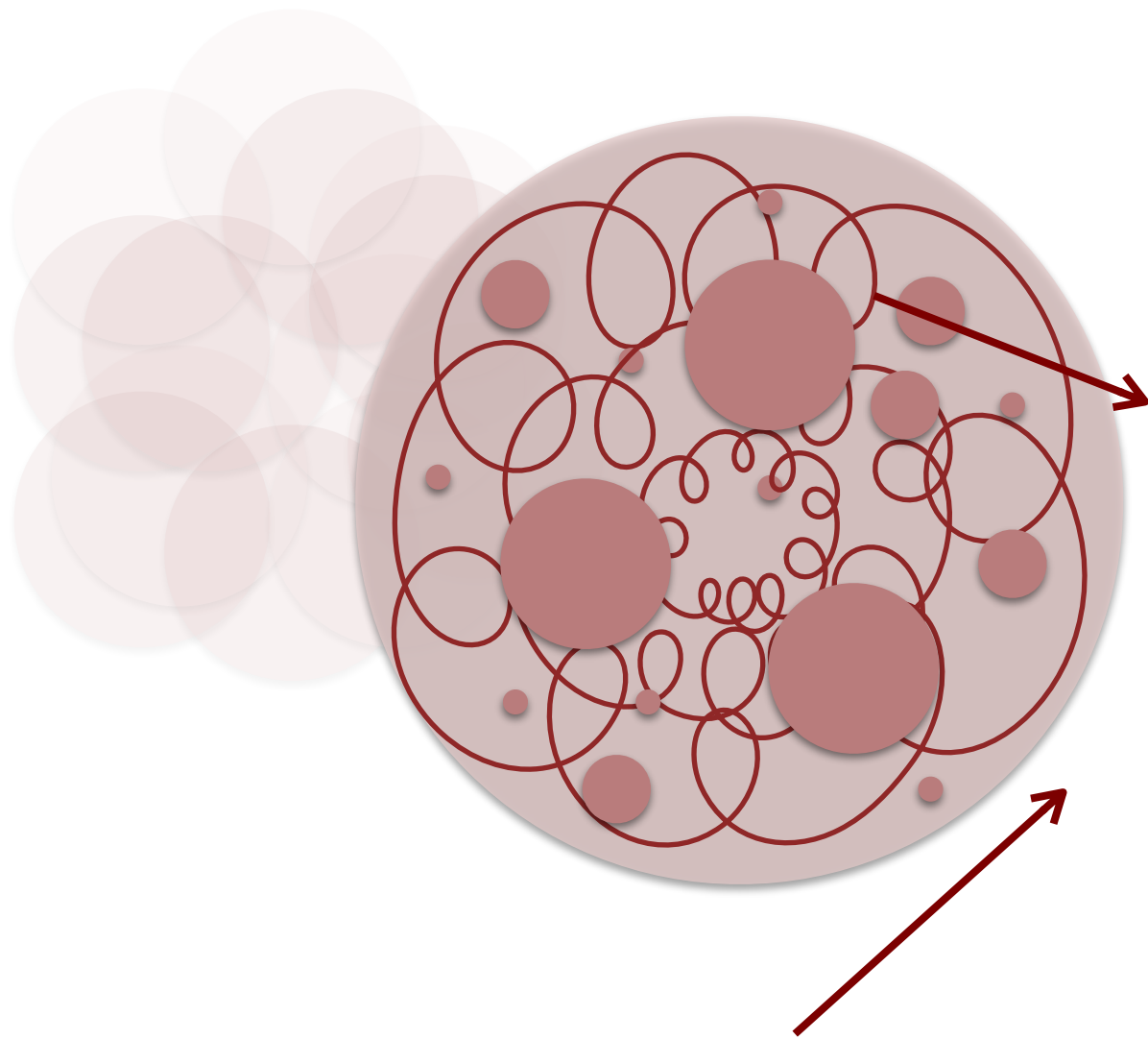
$$\pi(\mathcal{D} \mid \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}})$$

Moreover, the likelihood is naturally built *generatively*
with conditional probability distributions.



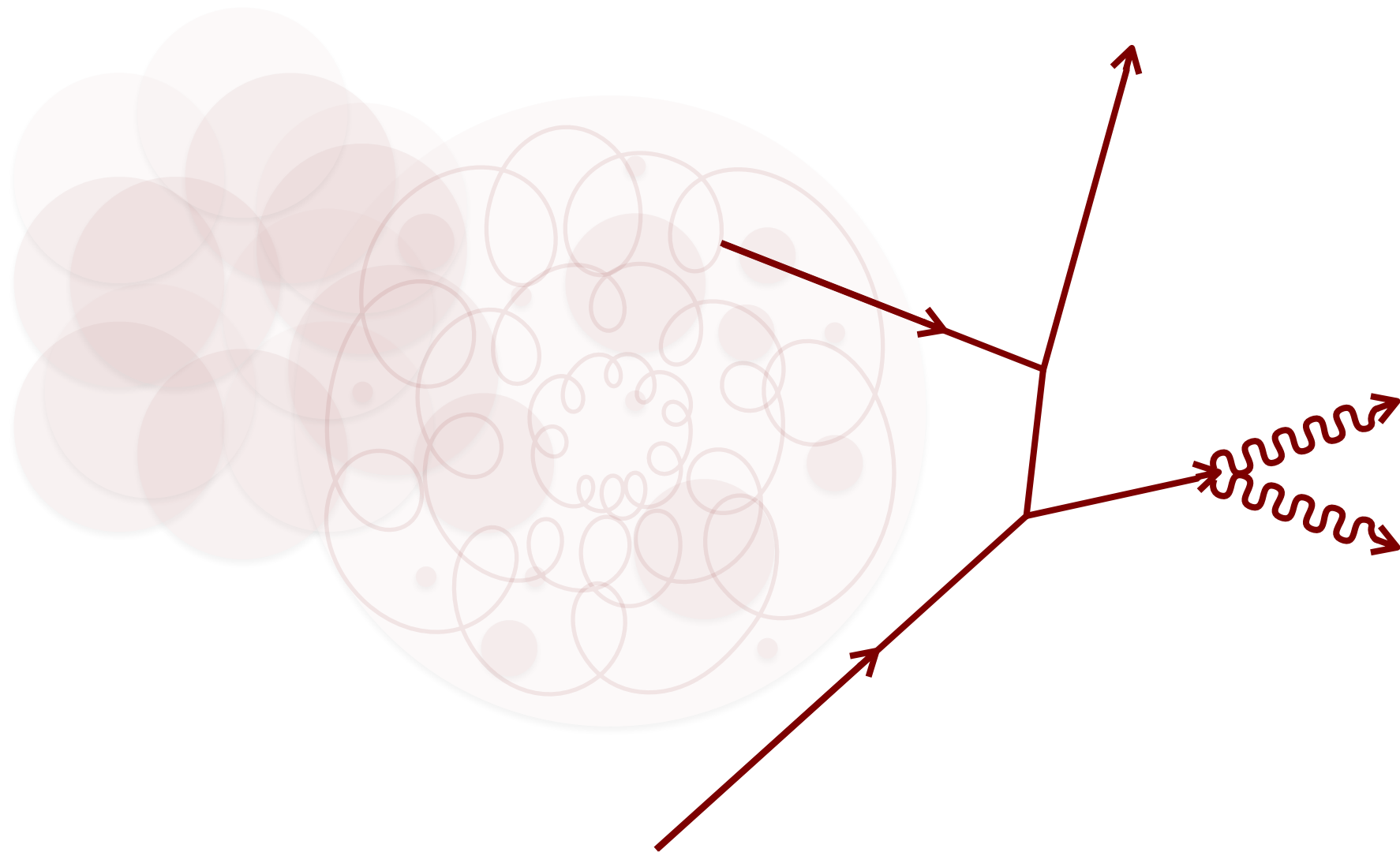
$$\pi(\text{nucleon} \mid \text{nucleus}, \tilde{\theta}_{\text{th}})$$

Moreover, the likelihood is naturally built *generatively* with conditional probability distributions.



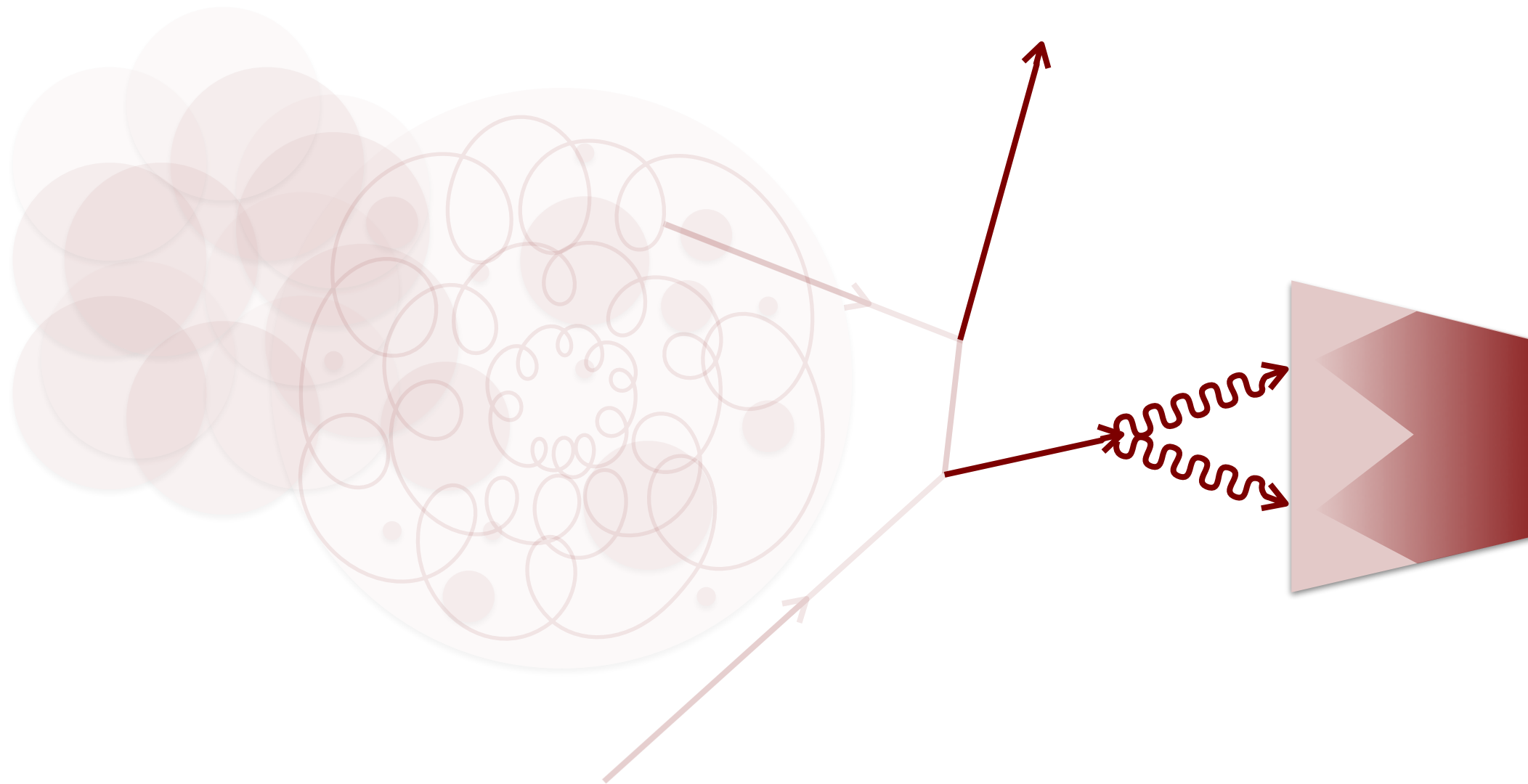
$$\pi(IS \mid \text{nucleon}, \tilde{\theta}_{\text{th}})$$

Moreover, the likelihood is naturally built *generatively* with conditional probability distributions.



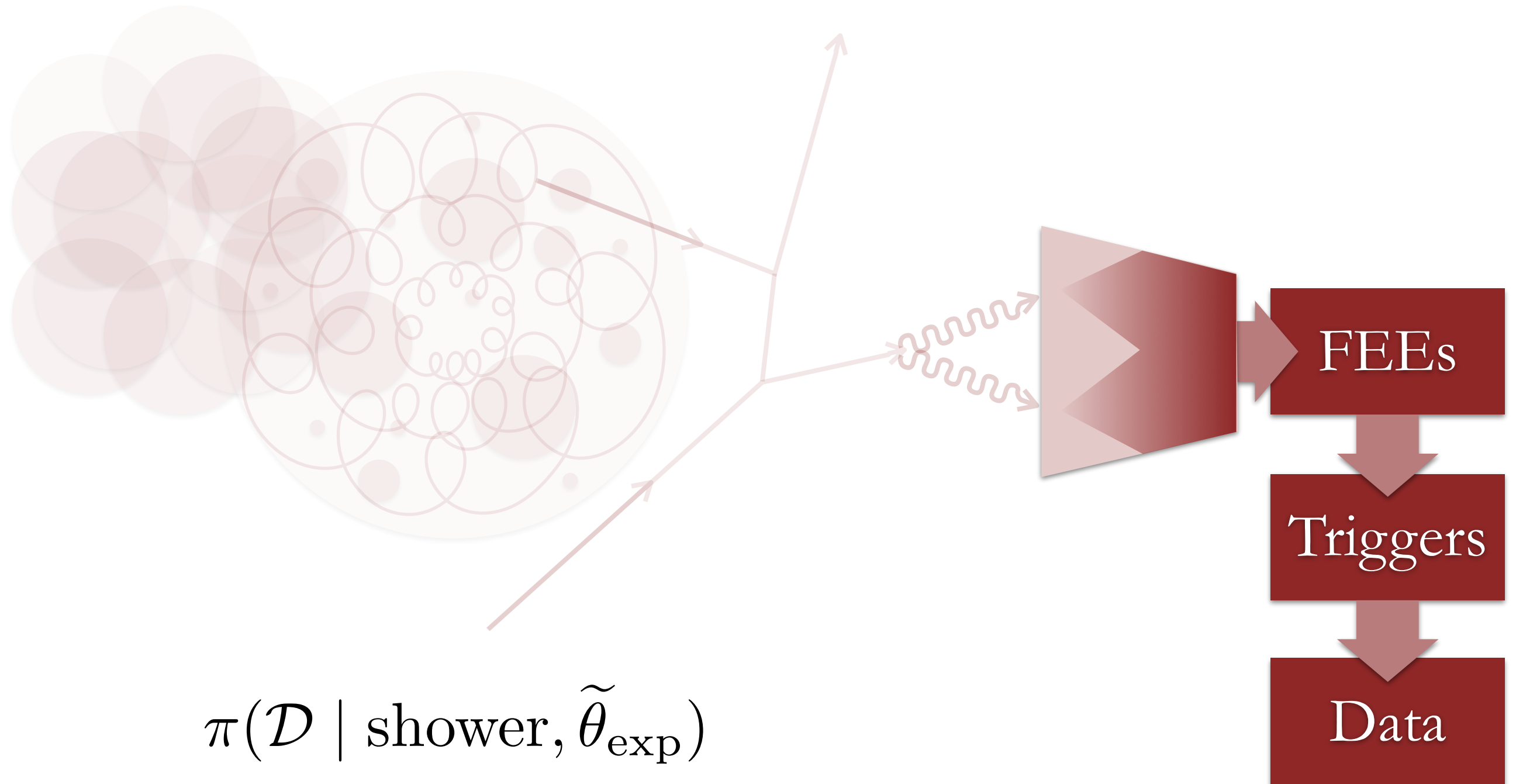
$$\pi(FS \mid IS, \theta_{\text{th}})$$

Moreover, the likelihood is naturally built *generatively* with conditional probability distributions.

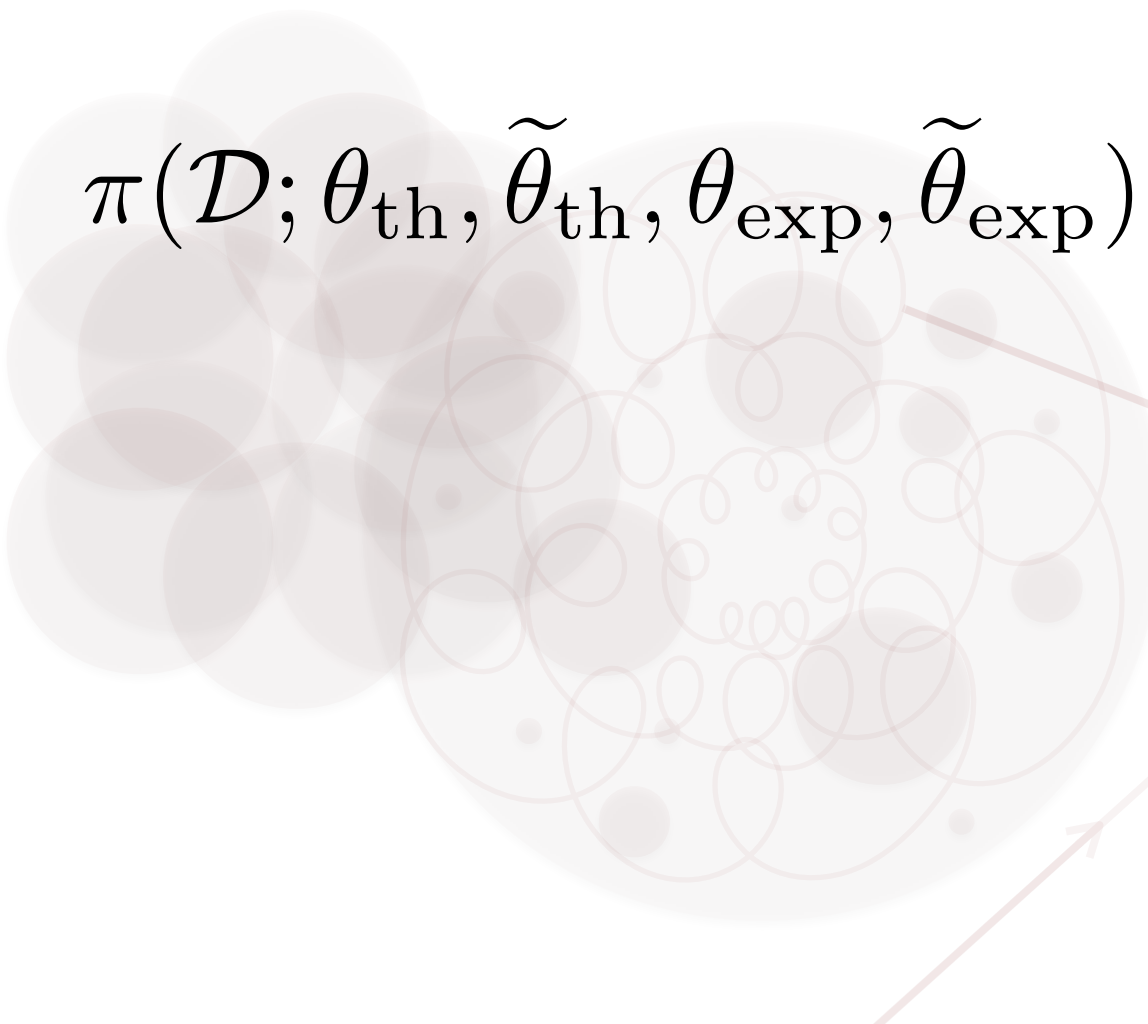
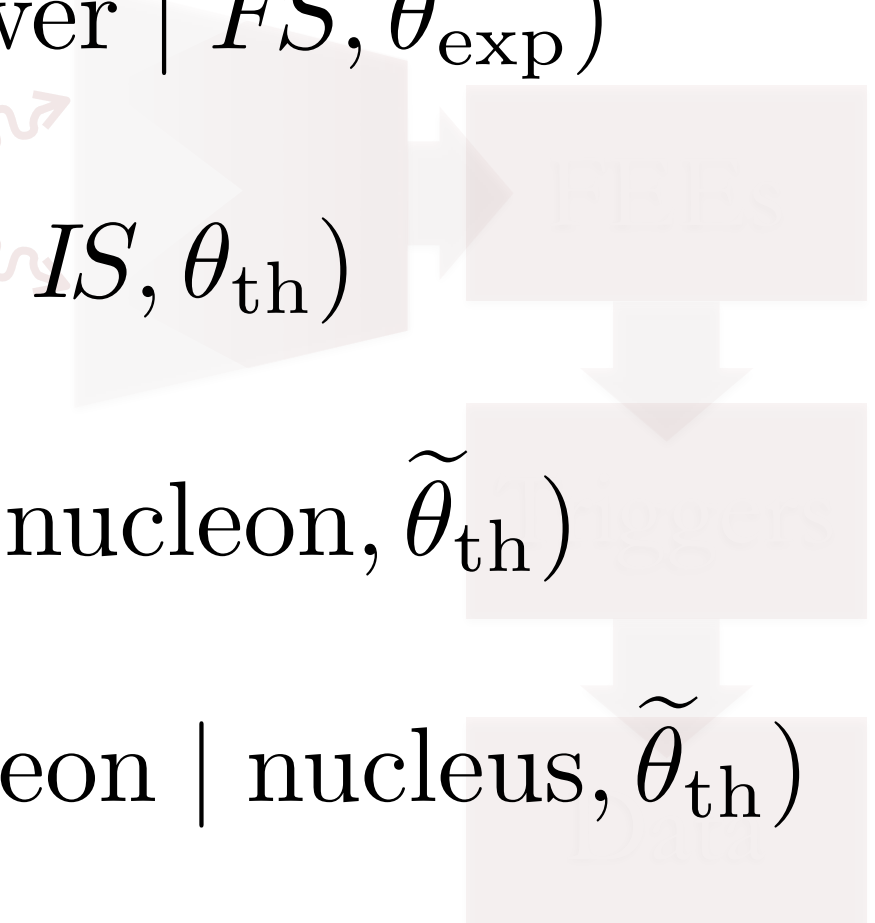


$$\pi(\text{shower} \mid FS, \theta_{\text{exp}})$$

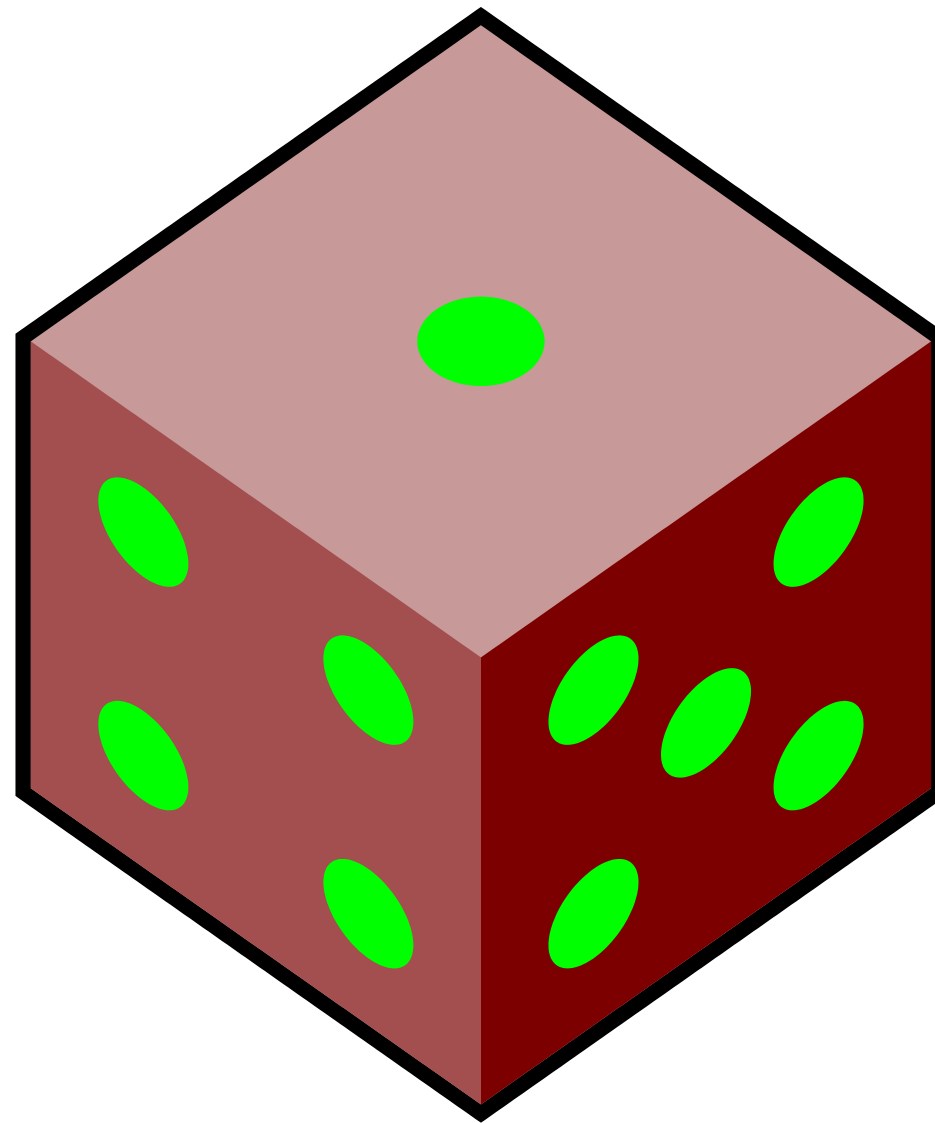
Moreover, the likelihood is naturally built *generatively* with conditional probability distributions.



Moreover, the likelihood is naturally built *generatively* with conditional probability distributions.


$$\begin{aligned}\pi(\mathcal{D}; \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}}) = & \pi(\mathcal{D} \mid \text{shower}, \tilde{\theta}_{\text{exp}}) \\ & \times \pi(\text{shower} \mid FS, \theta_{\text{exp}}) \\ & \times \pi(FS \mid IS, \theta_{\text{th}}) \\ & \times \pi(IS \mid \text{nucleon}, \tilde{\theta}_{\text{th}}) \\ & \times \pi(\text{nucleon} \mid \text{nucleus}, \tilde{\theta}_{\text{th}})\end{aligned}$$


The next step depends on the questions we want to answer and the assumptions we are willing to make.



Frequentist methods use the likelihood to identify parameters compatible with the data *on average*.

$$\hat{\theta}(\mathcal{D})$$

Frequentist methods use the likelihood to identify parameters compatible with the data *on average*.

$$\hat{\theta}(\mathcal{D})$$

$$\mathcal{L}(\hat{\theta}(\mathcal{D}), \theta)$$

Frequentist methods use the likelihood to identify parameters compatible with the data *on average*.

$$\hat{\theta}(\mathcal{D})$$

$$\mathcal{L}(\theta) = \int \mathcal{L}(\hat{\theta}(\mathcal{D}), \theta) \pi(\mathcal{D} \mid \theta) \mathrm{d}\mathcal{D}$$

In general, nuisance parameters have to enter into the estimator design.

$$\hat{\theta}(\mathcal{D})$$

$$\mathcal{L}(\theta, \tilde{\theta}) = \int \mathcal{L}(\hat{\theta}(\mathcal{D}), \theta) \pi(\mathcal{D} \mid \theta, \tilde{\theta}) d\mathcal{D}$$

In general, nuisance parameters have to enter into the estimator design.

$$\hat{\theta}(\mathcal{D})$$

$$\mathcal{L}(\theta, \tilde{\theta}) = \int \mathcal{L}(\hat{\theta}(\mathcal{D}), \theta) \pi(\mathcal{D} \mid \theta, \tilde{\theta}) d\mathcal{D}$$

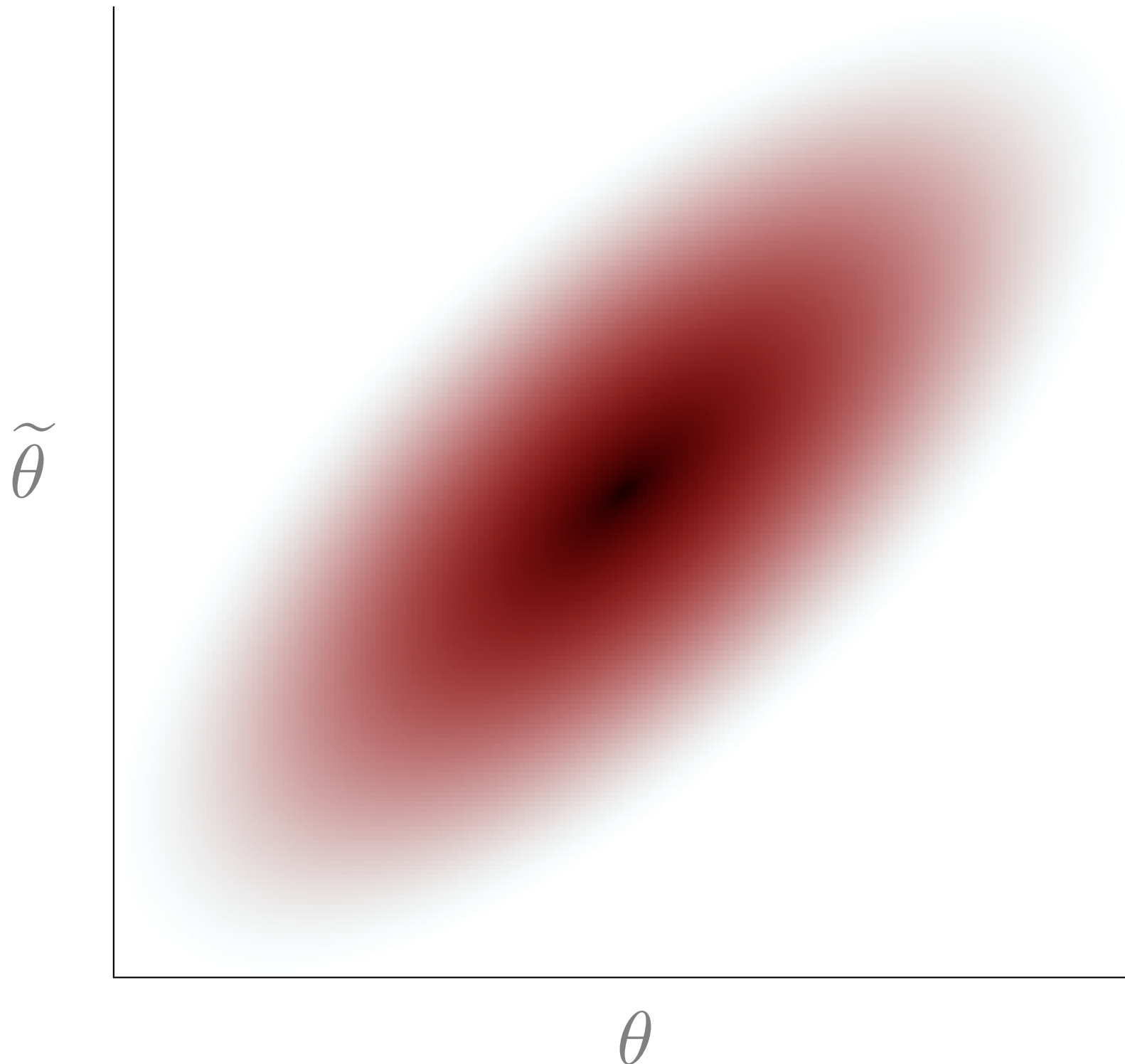
$$\mathcal{L}(\theta)_{\max} = \max_{\tilde{\theta}} \mathcal{L}(\theta, \tilde{\theta})$$

In *some* applications, relevant parameters can be isolated with techniques such as profiling.

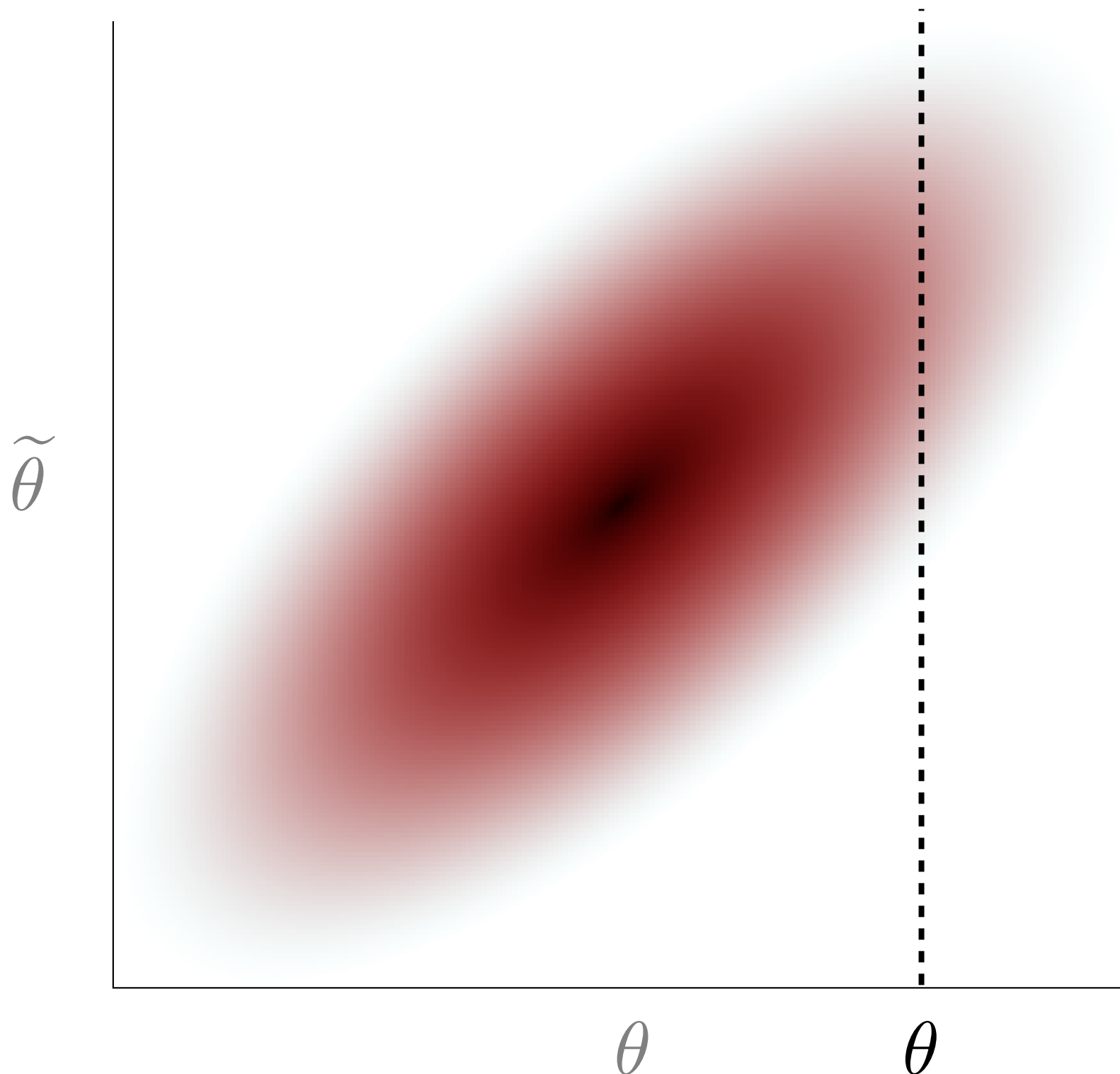
$$\hat{\tilde{\theta}}(\theta) = \operatorname{argmax}_{\tilde{\theta}} \pi(\mathcal{D} \mid \theta, \tilde{\theta})$$

$$\hat{\pi}(\mathcal{D} \mid \theta) = \pi(\mathcal{D} \mid \theta, \hat{\tilde{\theta}}(\theta))$$

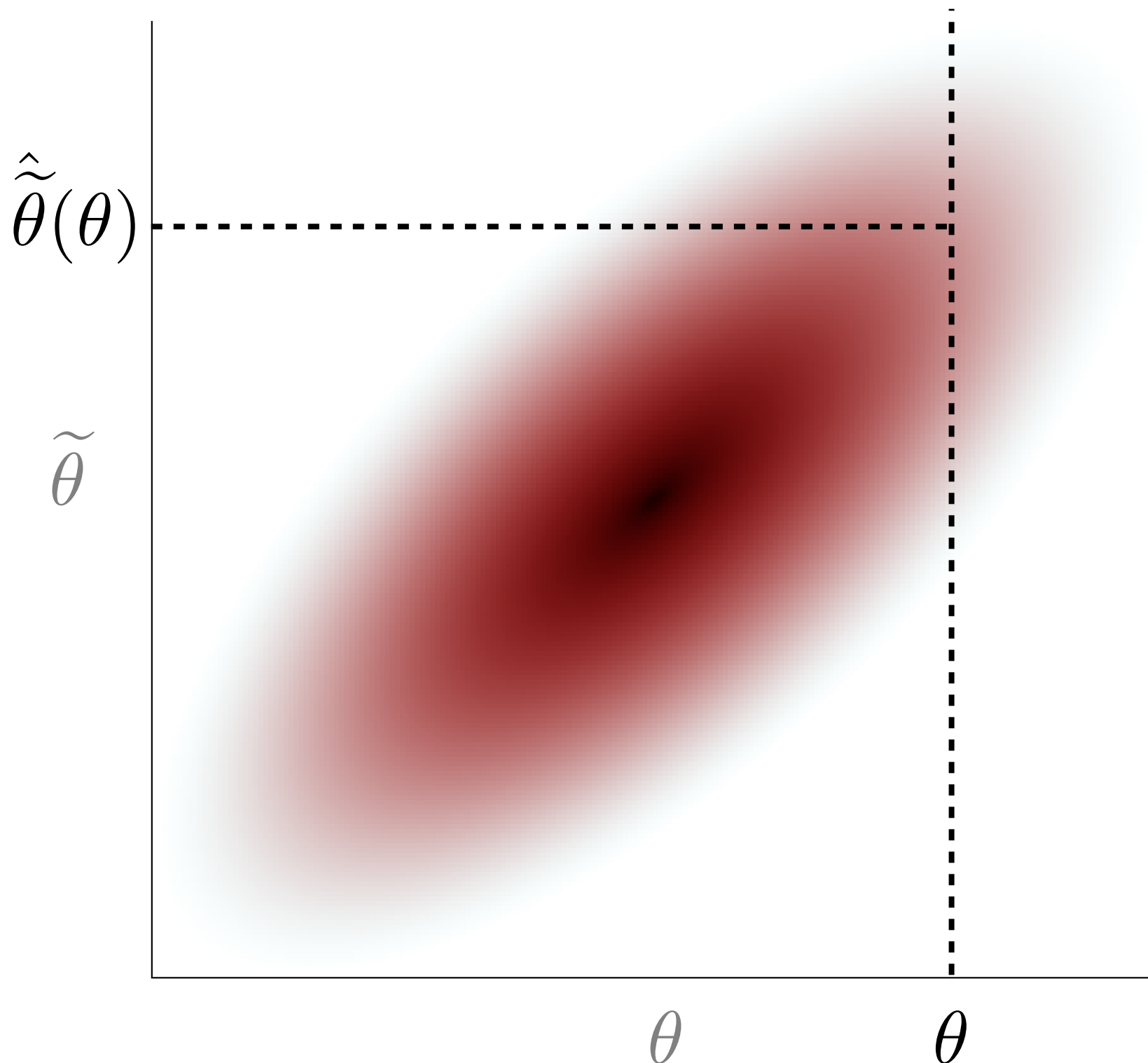
In *some* applications, relevant parameters can be isolated with techniques such as profiling.



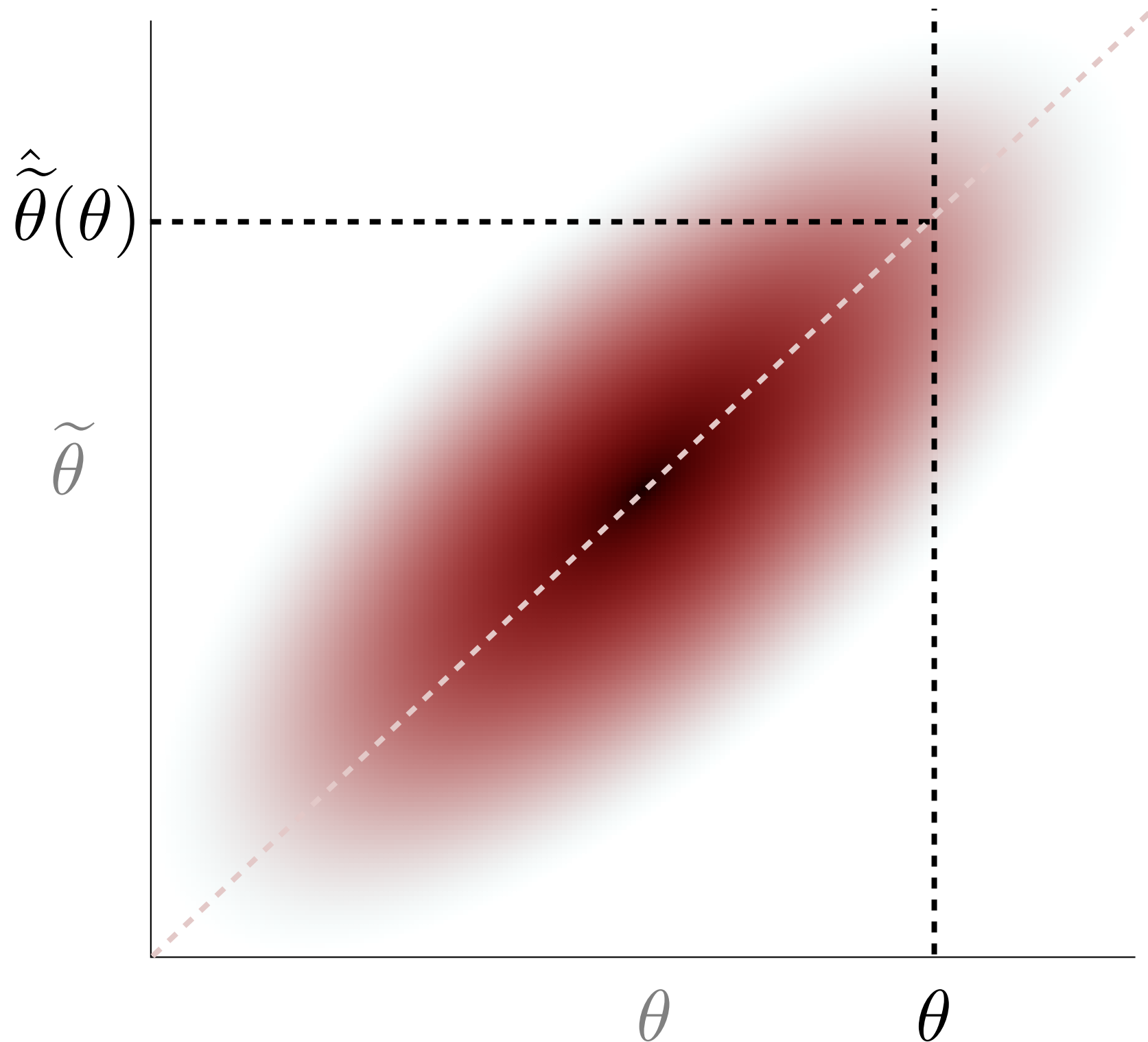
In *some* applications, relevant parameters can be isolated with techniques such as profiling.



In *some* applications, relevant parameters can be isolated with techniques such as profiling.



In *some* applications, relevant parameters can be isolated with techniques such as profiling.



Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi(\mathcal{D} \mid \theta)$$

Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi(\mathcal{D} \mid \theta)$$

Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi(\mathcal{D} \mid \theta)$$

Adding a prior distribution “inverts” the likelihood, identifying the parameters consistent with the data.

$$\pi(\mathcal{D} \mid \theta)$$

Adding a prior distribution “inverts” the likelihood, identifying the parameters consistent with the data.

$$\pi(\mathcal{D} \mid \theta) \pi(\theta)$$

Adding a prior distribution “inverts” the likelihood, identifying the parameters consistent with the data.

$$\pi(\theta \mid \mathcal{D}) \propto \pi(\mathcal{D} \mid \theta) \pi(\theta)$$

Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.

$$\pi(\theta \mid \mathcal{D})$$

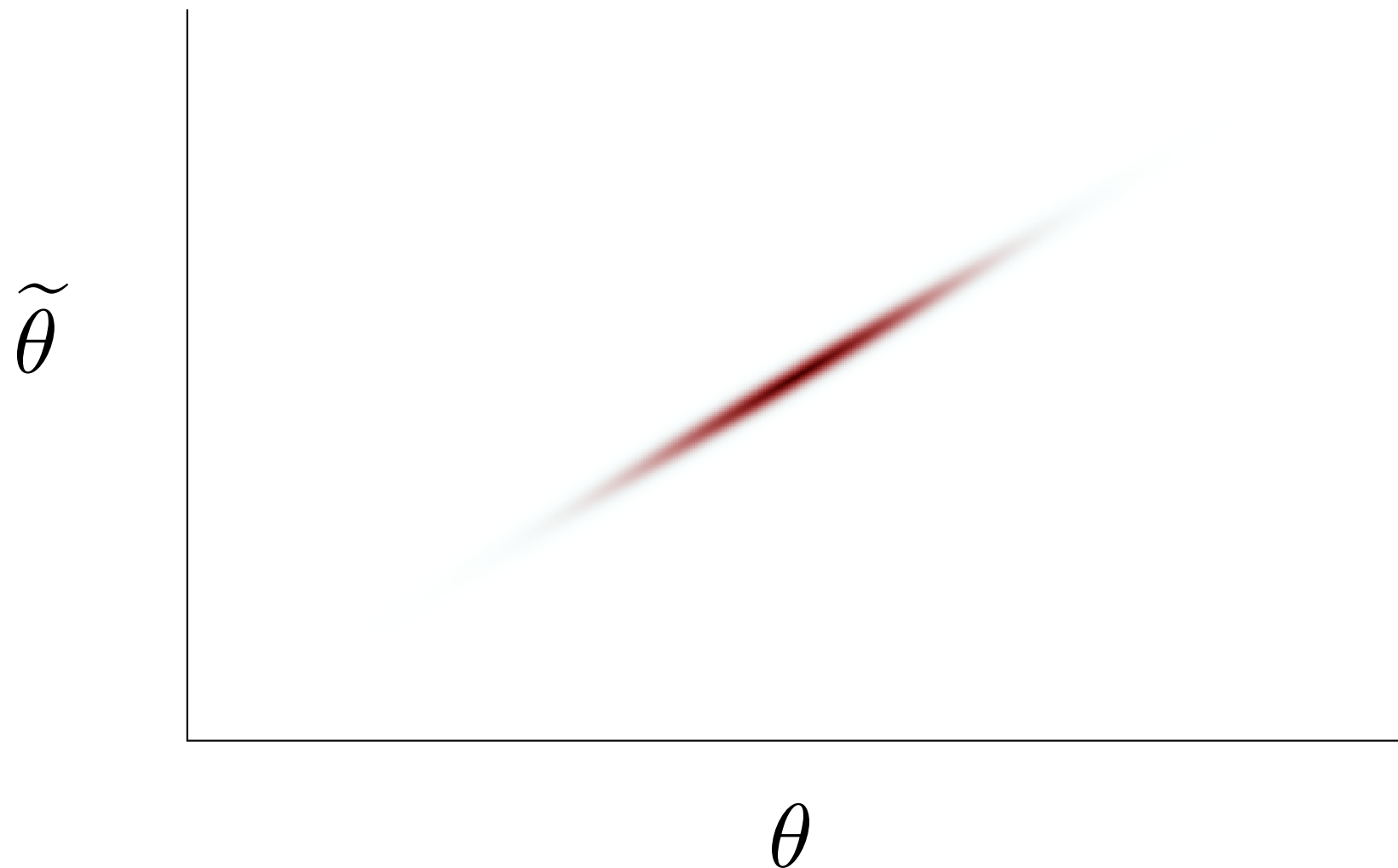
Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.

$$\pi(\theta, \tilde{\theta} \mid \mathcal{D})$$

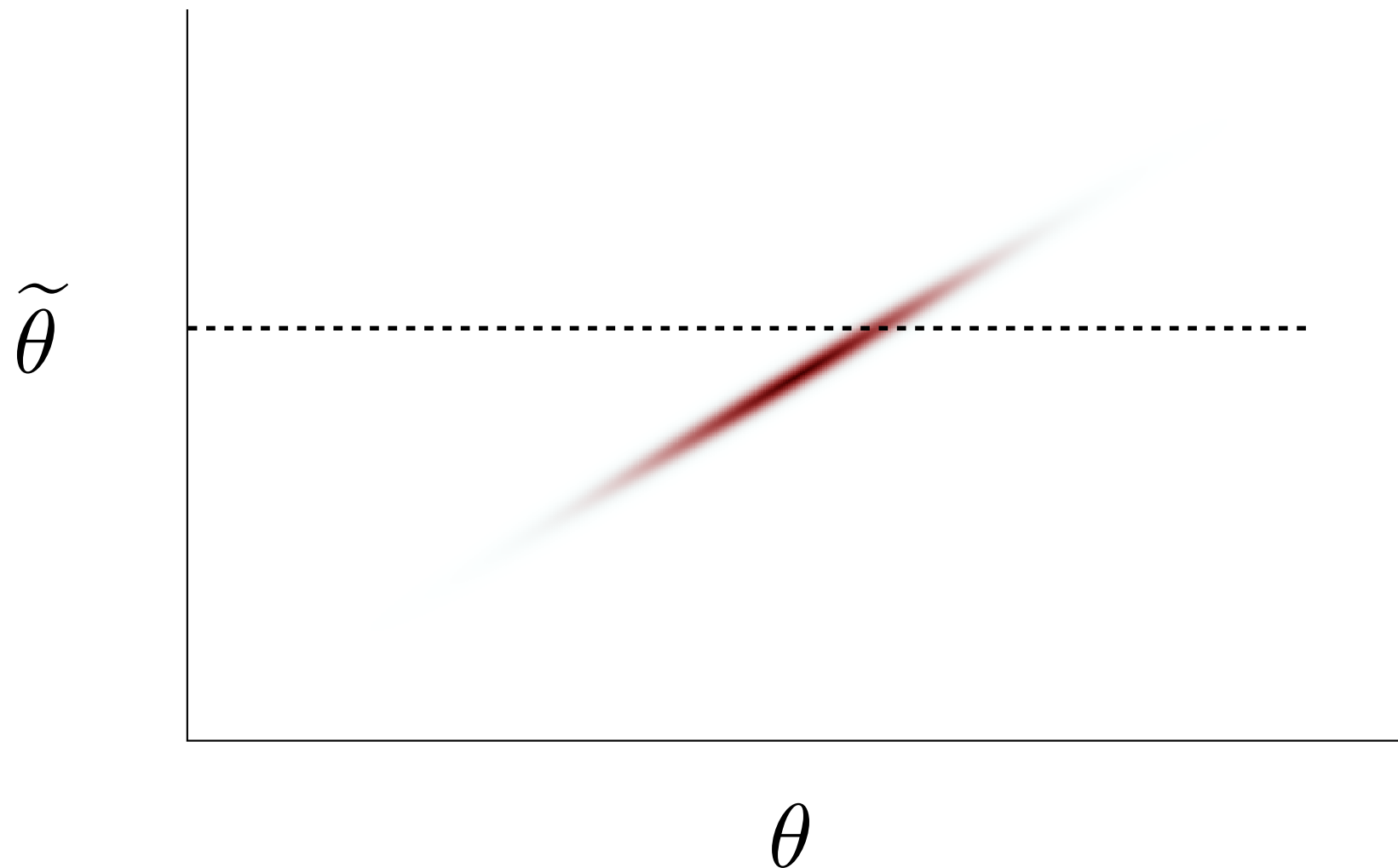
Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.

$$\pi(\theta \mid \mathcal{D}) = \int \pi(\theta, \tilde{\theta} \mid \mathcal{D}) d\tilde{\theta}$$

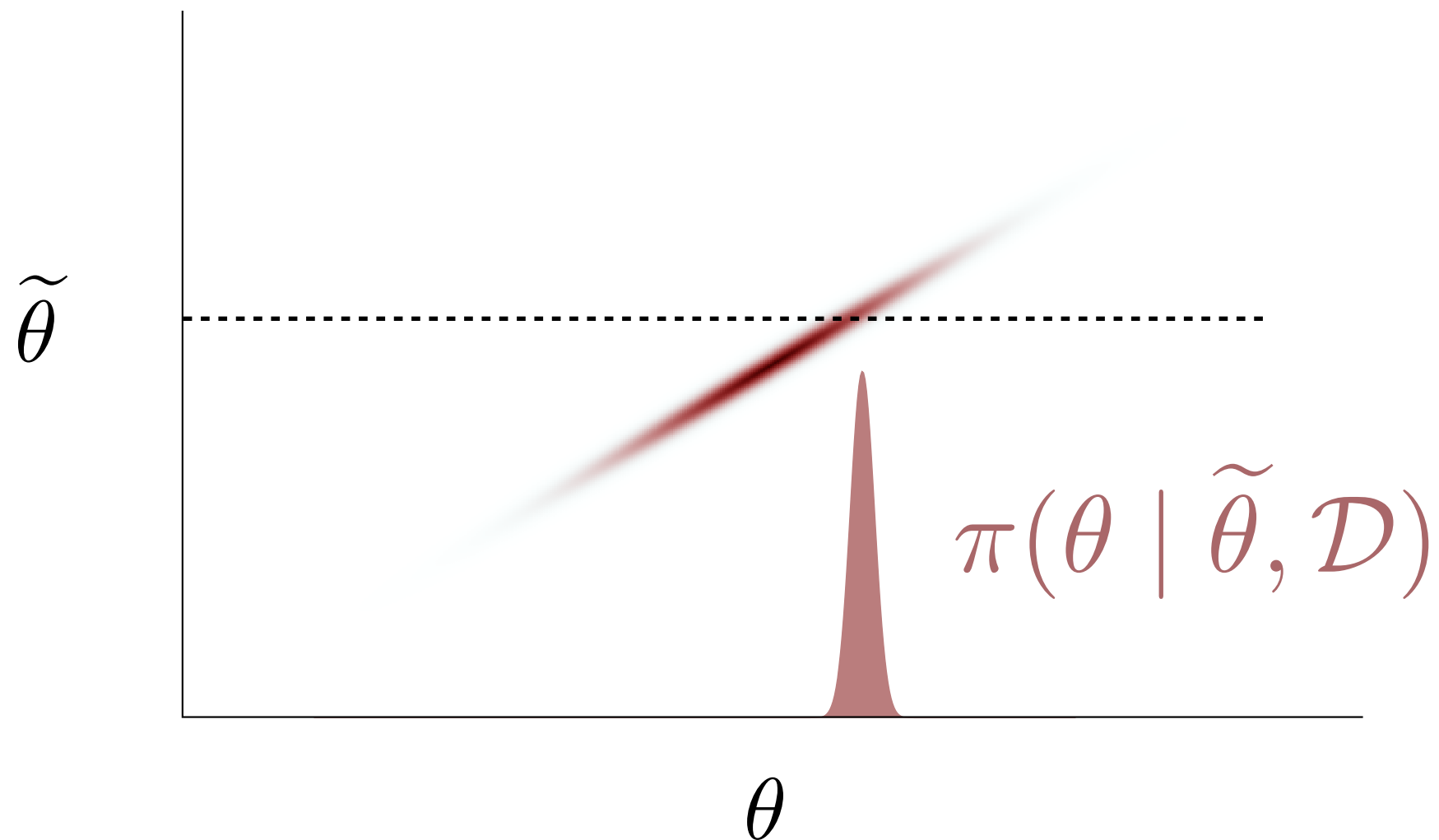
Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.



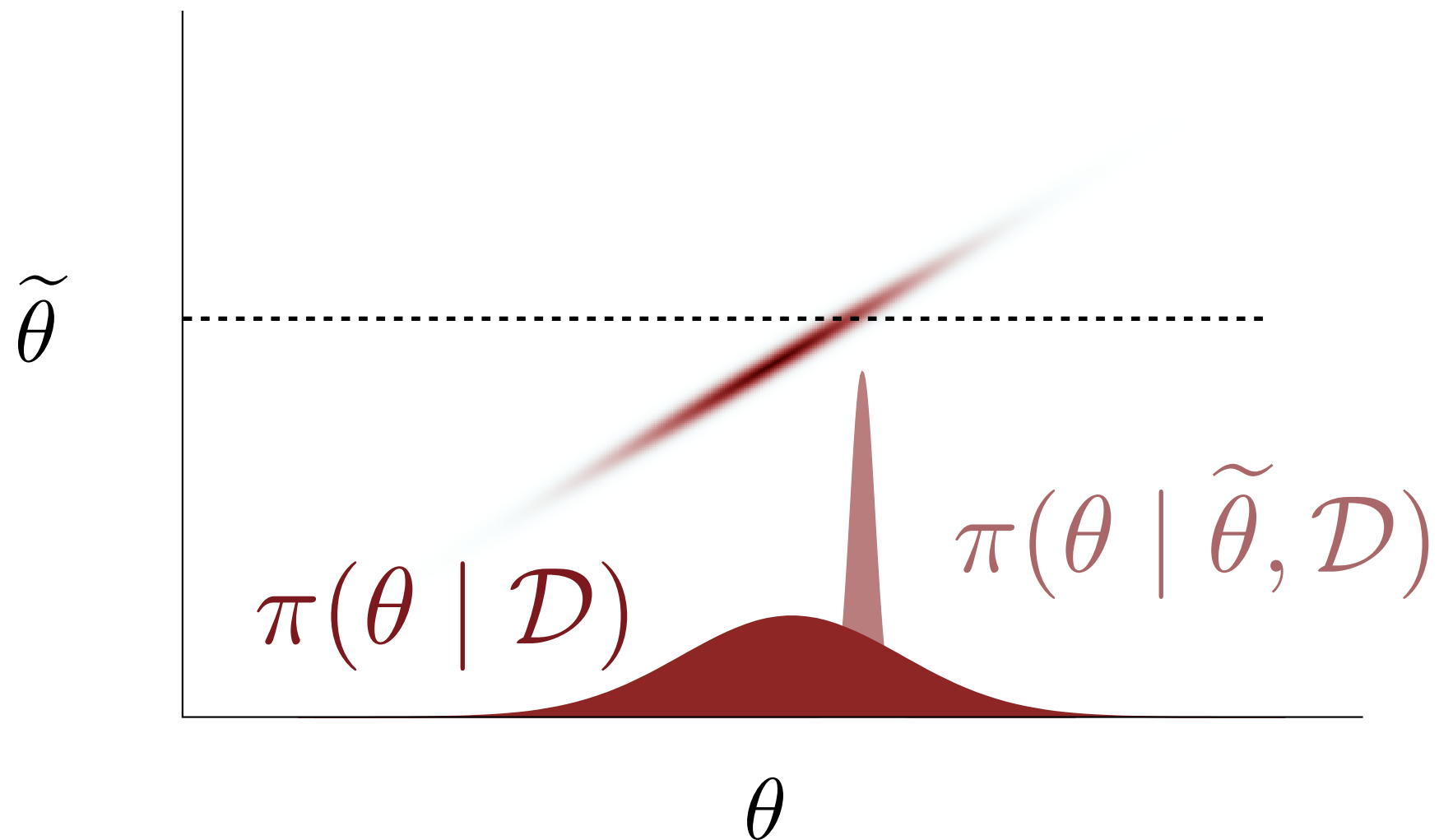
Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.



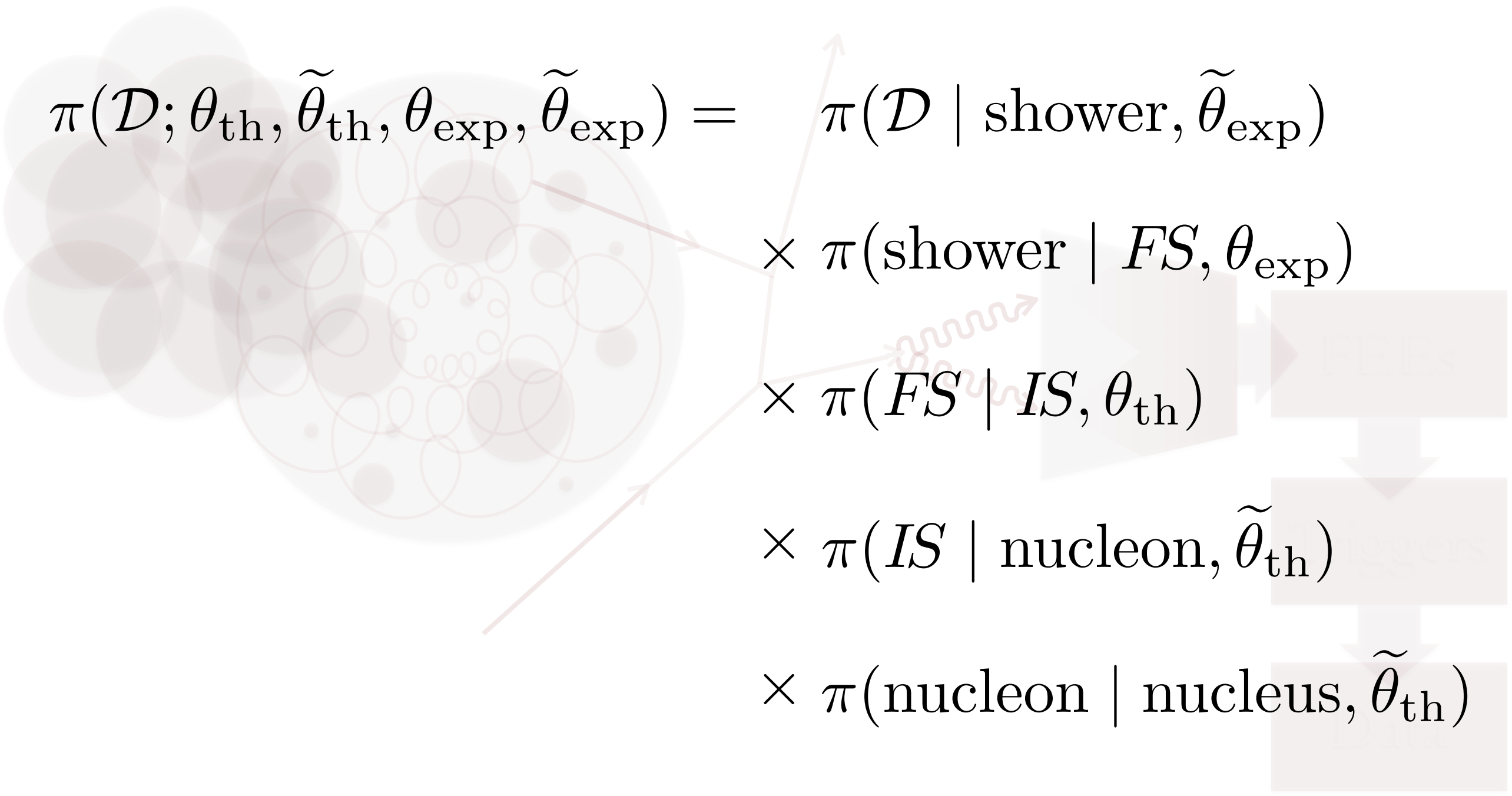
Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.



Relevant parameters are isolated by marginalizing nuisance parameters out of the posterior.

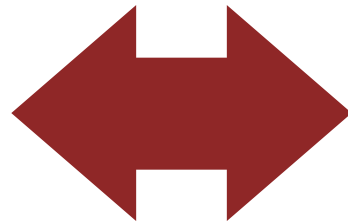


Generative modeling is hard, but any steps towards a completely generative model will improve analyses.


$$\begin{aligned}\pi(\mathcal{D}; \theta_{\text{th}}, \tilde{\theta}_{\text{th}}, \theta_{\text{exp}}, \tilde{\theta}_{\text{exp}}) = & \pi(\mathcal{D} \mid \text{shower}, \tilde{\theta}_{\text{exp}}) \\ & \times \pi(\text{shower} \mid FS, \theta_{\text{exp}}) \\ & \times \pi(FS \mid IS, \theta_{\text{th}}) \\ & \times \pi(IS \mid \text{nucleon}, \tilde{\theta}_{\text{th}}) \\ & \times \pi(\text{nucleon} \mid \text{nucleus}, \tilde{\theta}_{\text{th}})\end{aligned}$$

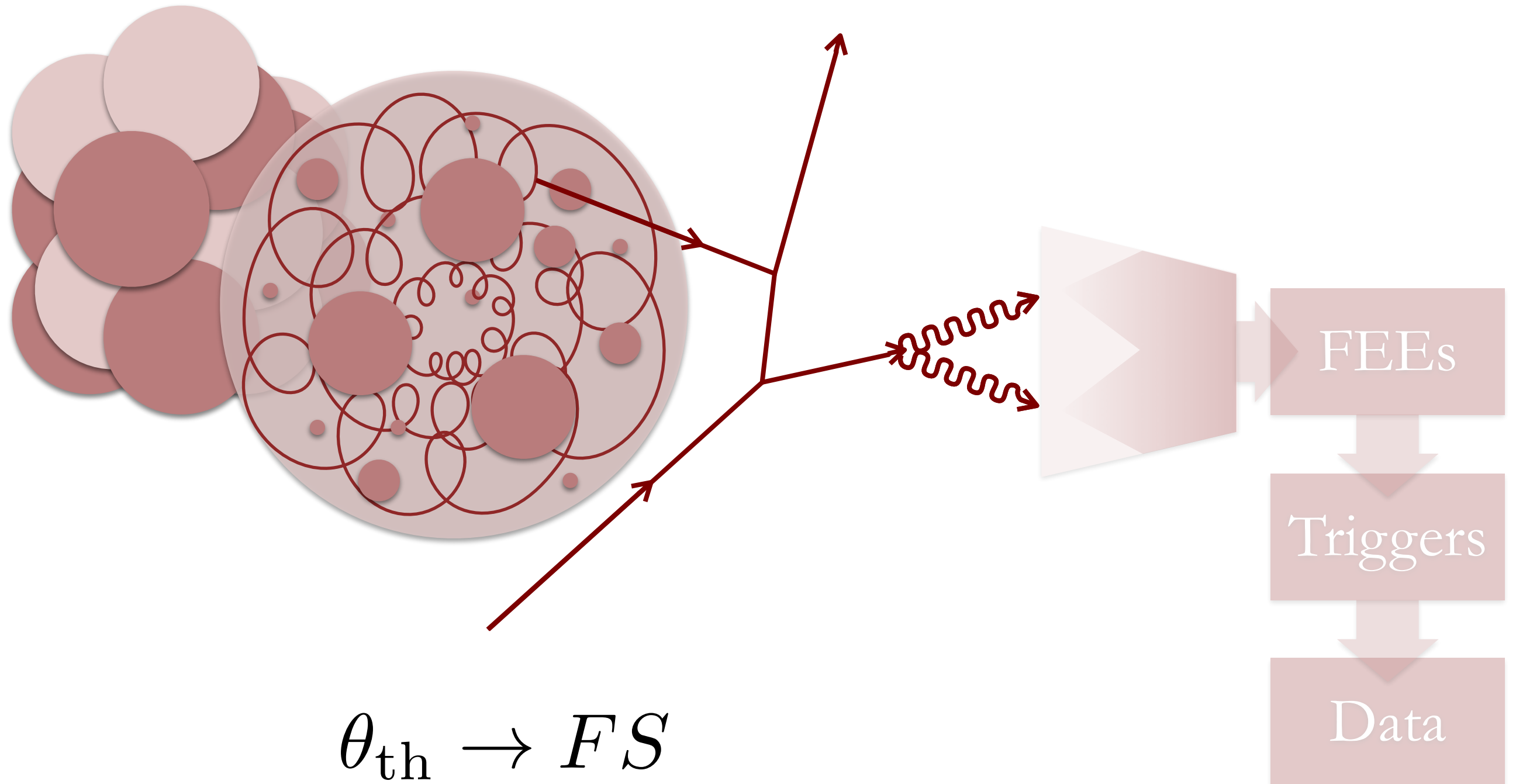
New tools like probabilistic programming facilitate the specification of these model components.

```
data {  
  int<lower=1> N;  
  real x[N];  
}  
transformed data {  
  vector[N] mu;  
  cov_matrix[N] Sigma;  
  for (i in 1:N)  
    mu[i] <- 0;  
  for (i in 1:N)  
    for (j in 1:N)  
      Sigma[i,j] <- exp(-pow(x[i] - x[j],2))  
        + if_else(i==j, 0.1, 0.0);  
}  
parameters {  
  vector[N] y;  
}  
model {  
  y ~ multi_normal(mu,Sigma);  
}
```

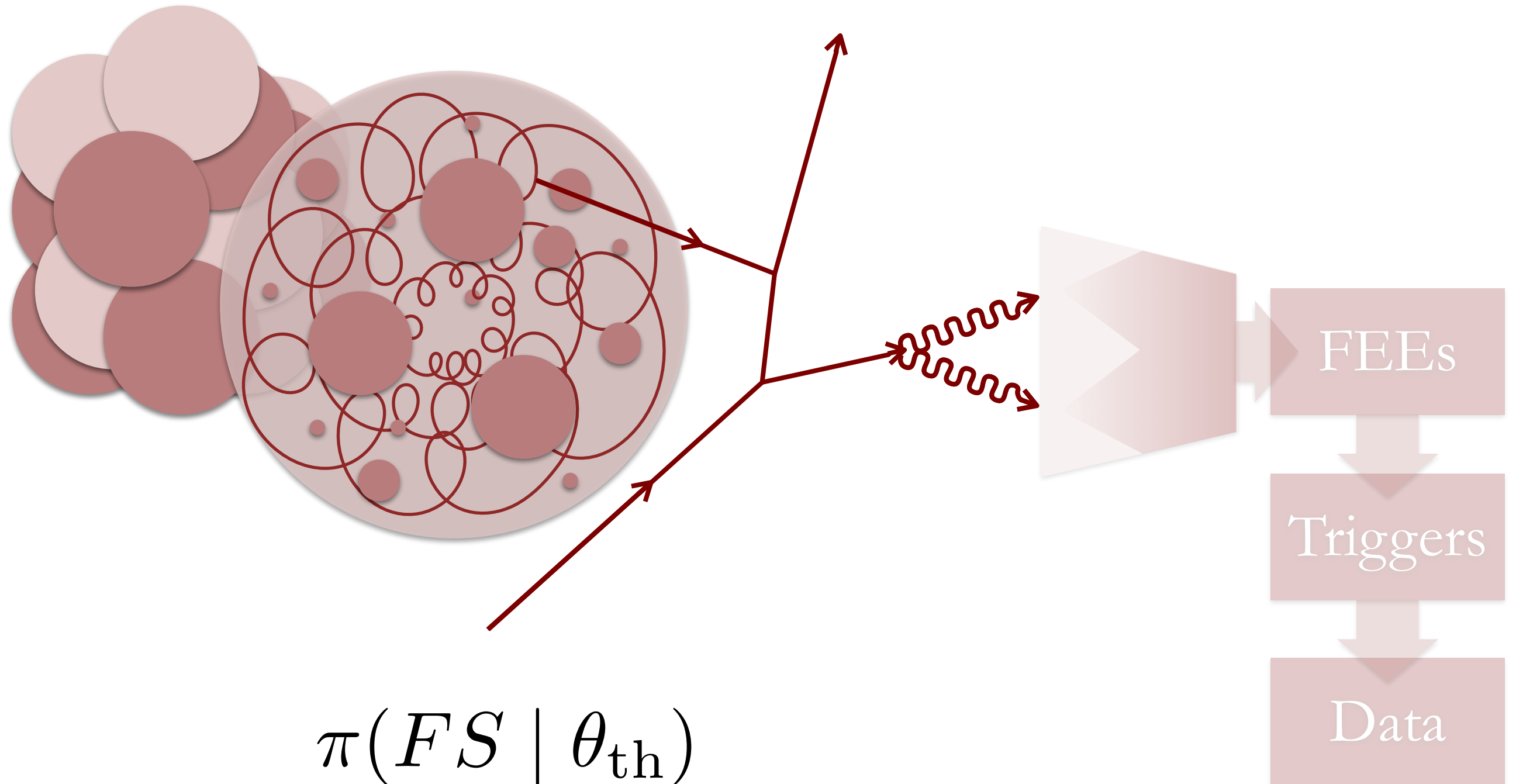


$$\pi(\mathcal{D} \mid \theta)$$

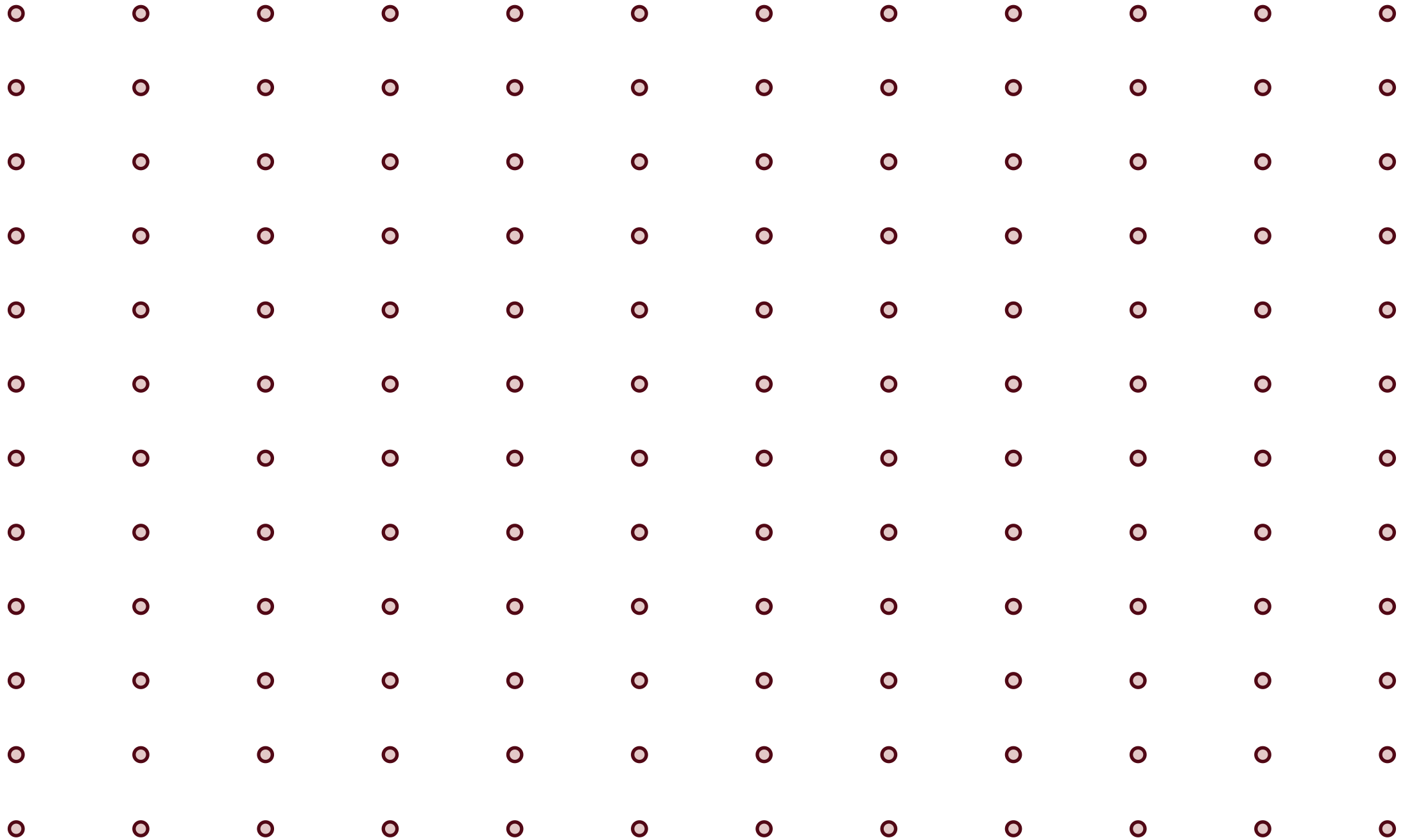
Finally, event generators are really just stochastic representations of intermediate probability distributions.



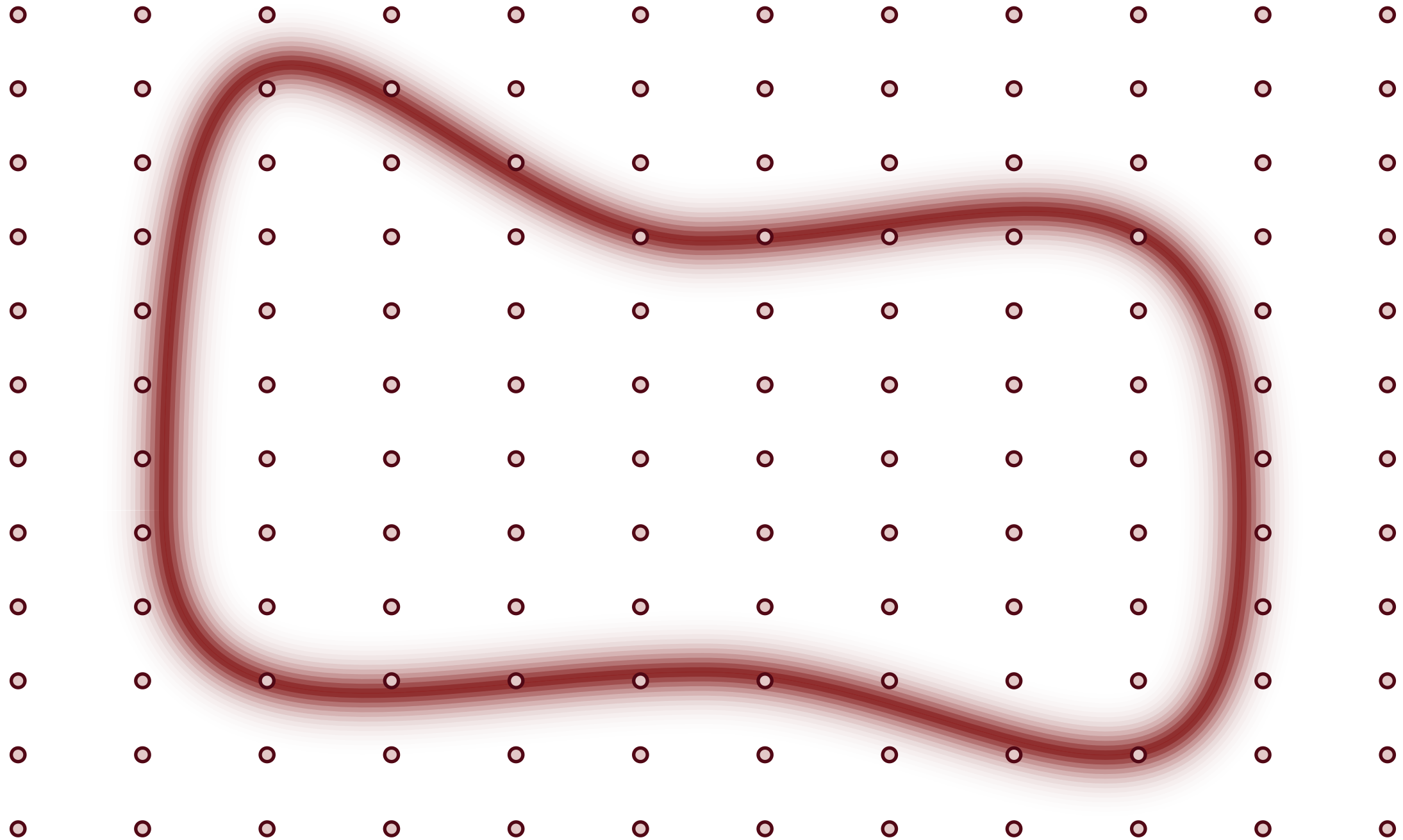
Finally, event generators are really just stochastic representations of intermediate probability distributions.



Proper statistical methods, however, can admit much more efficient use of computational resources.



Proper statistical methods, however, can admit much more efficient use of computational resources.



Proper statistical methods, however, can admit much more efficient use of computational resources.

