Learn Bayes, MCMC and Stan 2017!

Andrew Gelman
Jonah Sol Gabry
Michael Betancourt @betanalpha

Columbia University

New York, NY
August 23 - August 25, 2017

Wednesday

- Foundations of Bayesian Inference
- Bayesian Computation and Markov Chain Monte Carlo
- Bayesian Modeling with Stan

Wednesday

- Foundations of Bayesian Inference
- Bayesian Computation and Markov Chain Monte Carlo
- Bayesian Modeling with Stan

Thursday

- Regression Modeling with Stan

Wednesday

- Foundations of Bayesian Inference
- Bayesian Computation and Markov Chain Monte Carlo
- Bayesian Modeling with Stan

Thursday

- Regression Modeling with Stan

Friday

- Hierarchical Modeling with Stan

Foundations of Bayesian Inference

Attempts to learn from measurements are complicated by the natural variability of measurements.

$\mathcal{D}$

Attempts to learn from measurements are complicated by the natural variability of measurements.

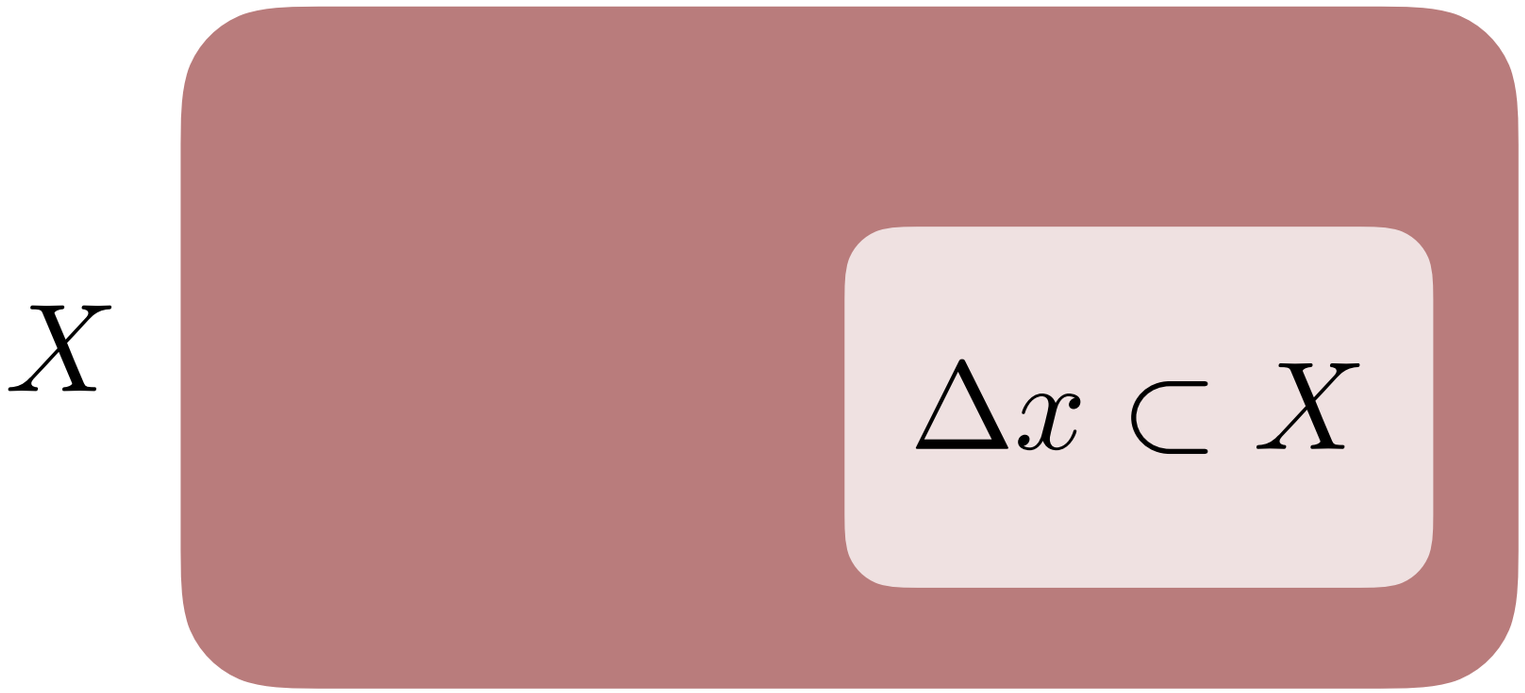
$$\pi(\mathcal{D})$$

Quick aside: what exactly is a probability distribution?

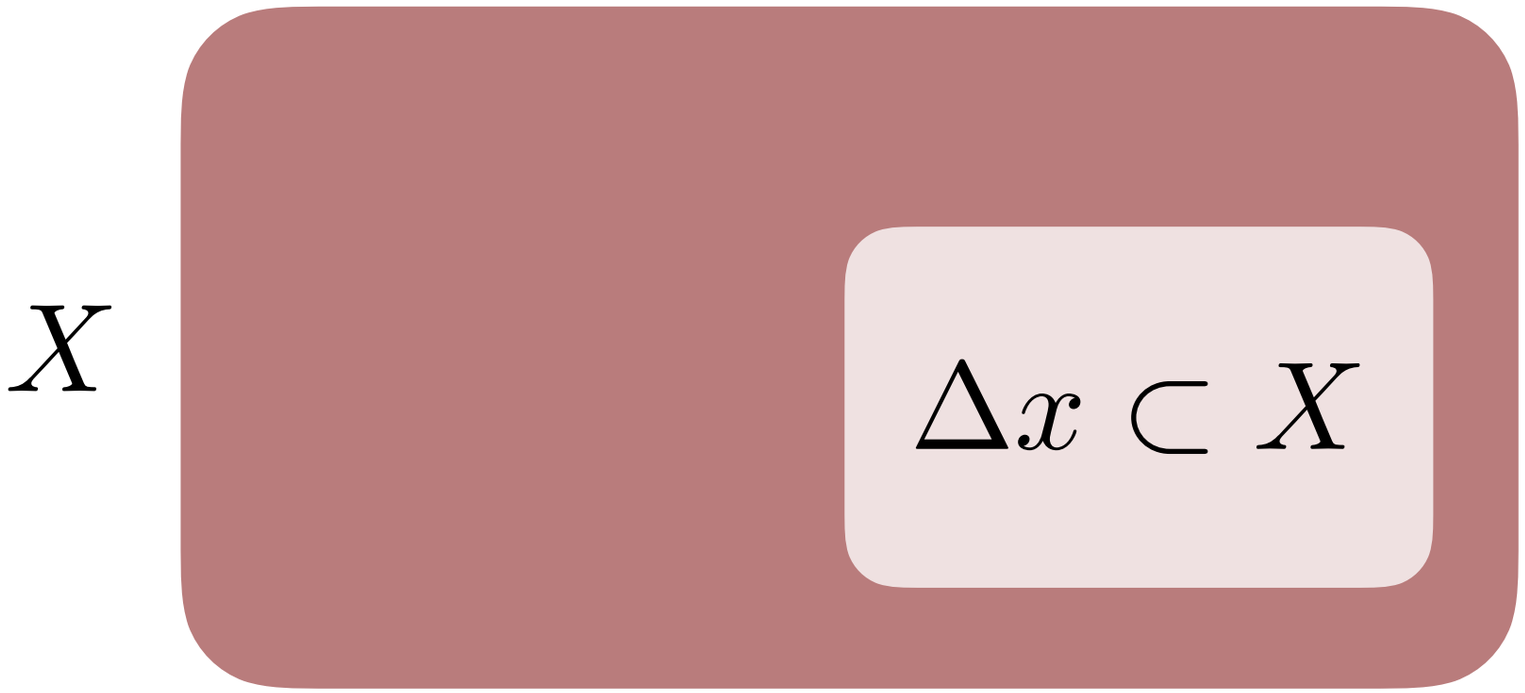
$\pi(\mathcal{D})$?

A probability distribution assigns probability to neighborhoods in the measurement space.

A probability distribution assigns probability to neighborhoods in the measurement space.



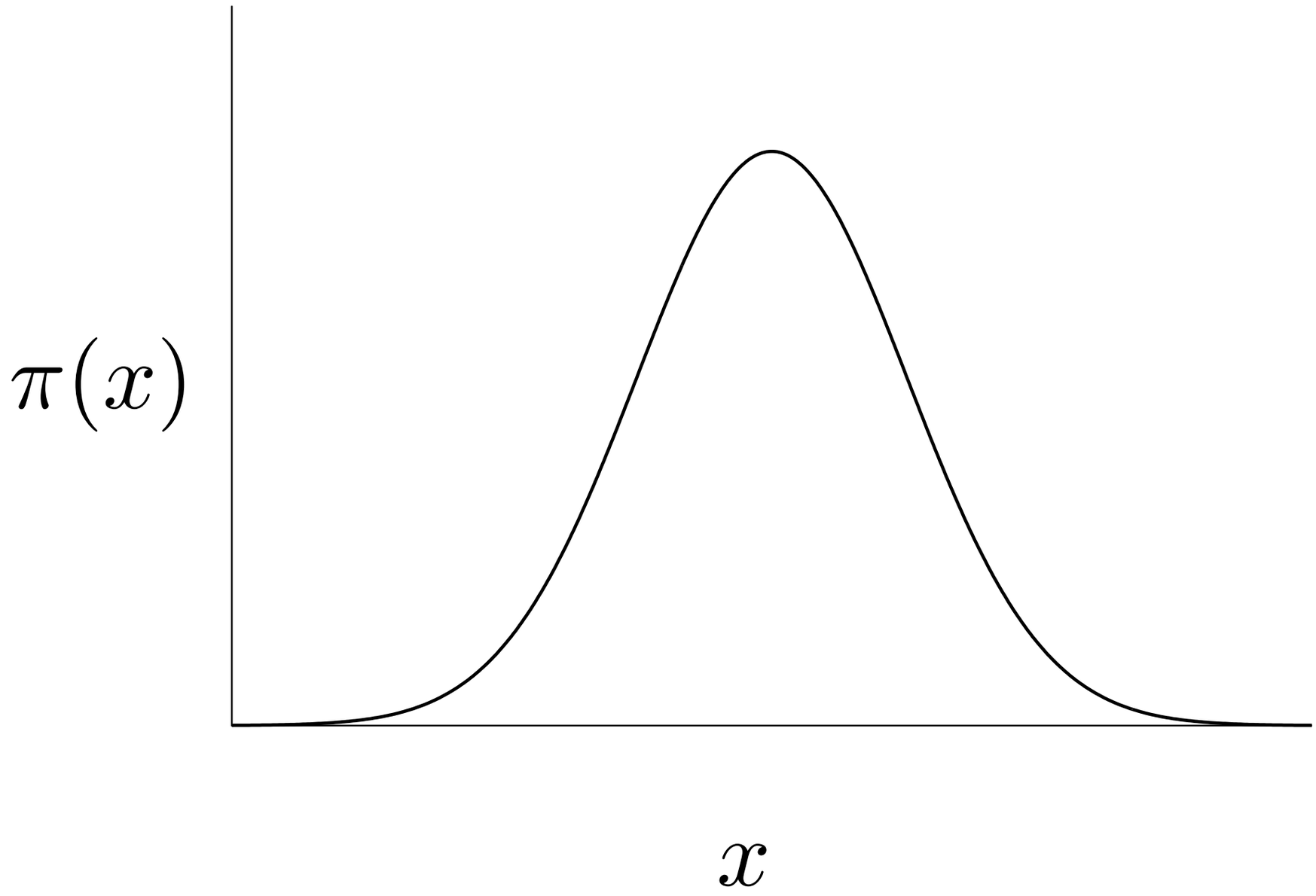
A probability distribution assigns probability to neighborhoods in the measurement space.



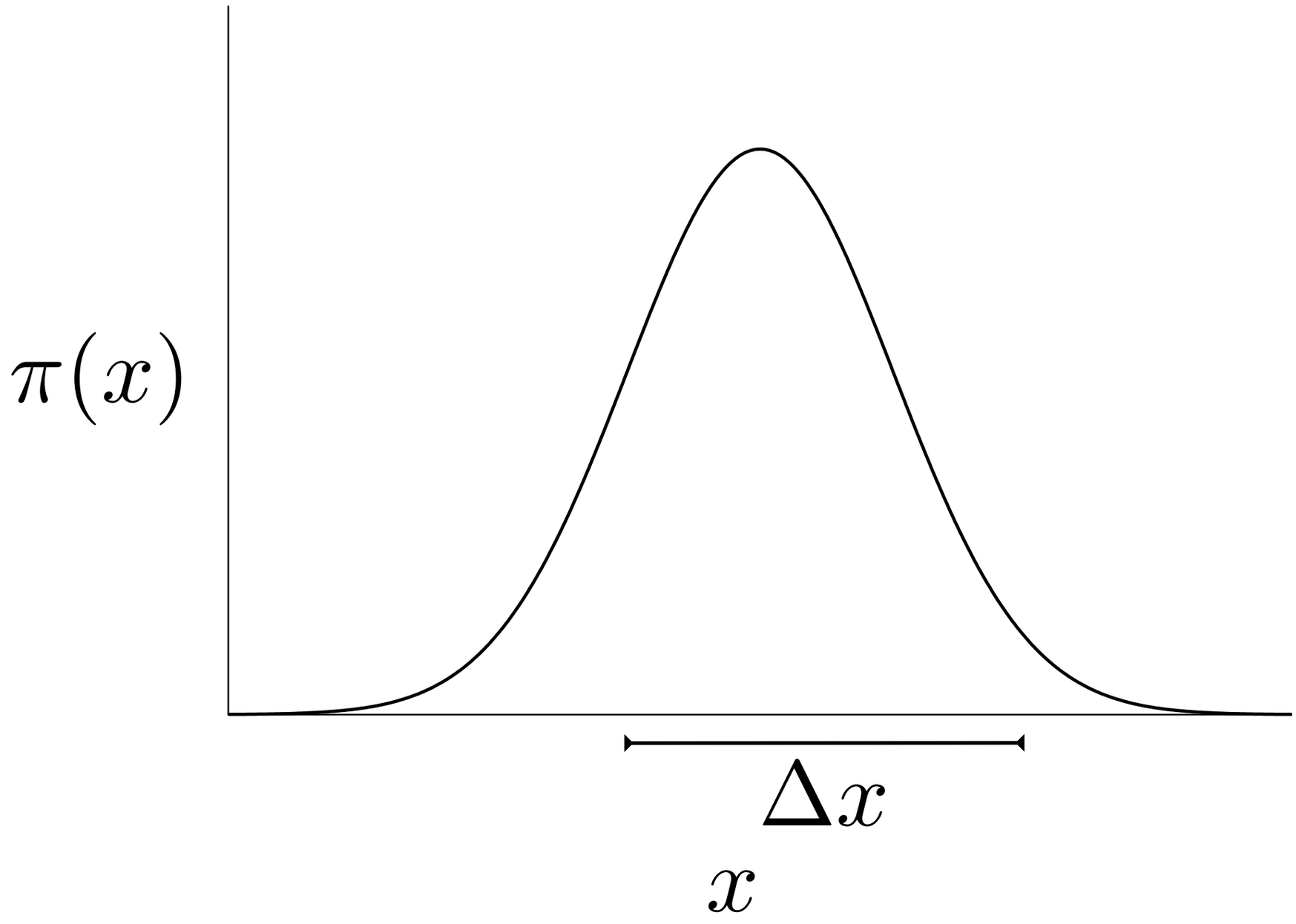
$$\pi[\Delta x] \in [0, 1]$$

$$\pi[X] = 1$$

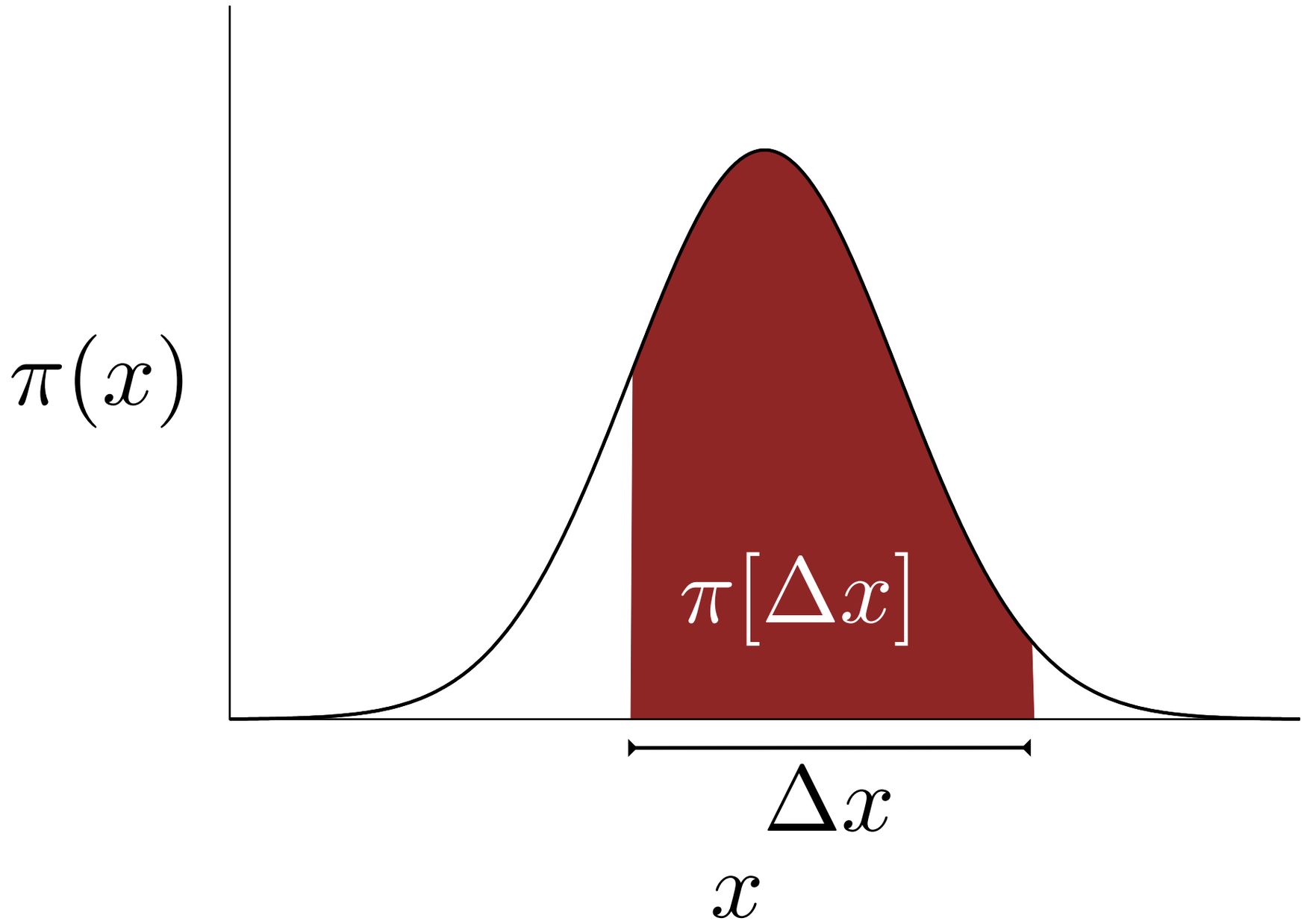
Typically we represent probability distributions with probability density functions which we can *integrate*.



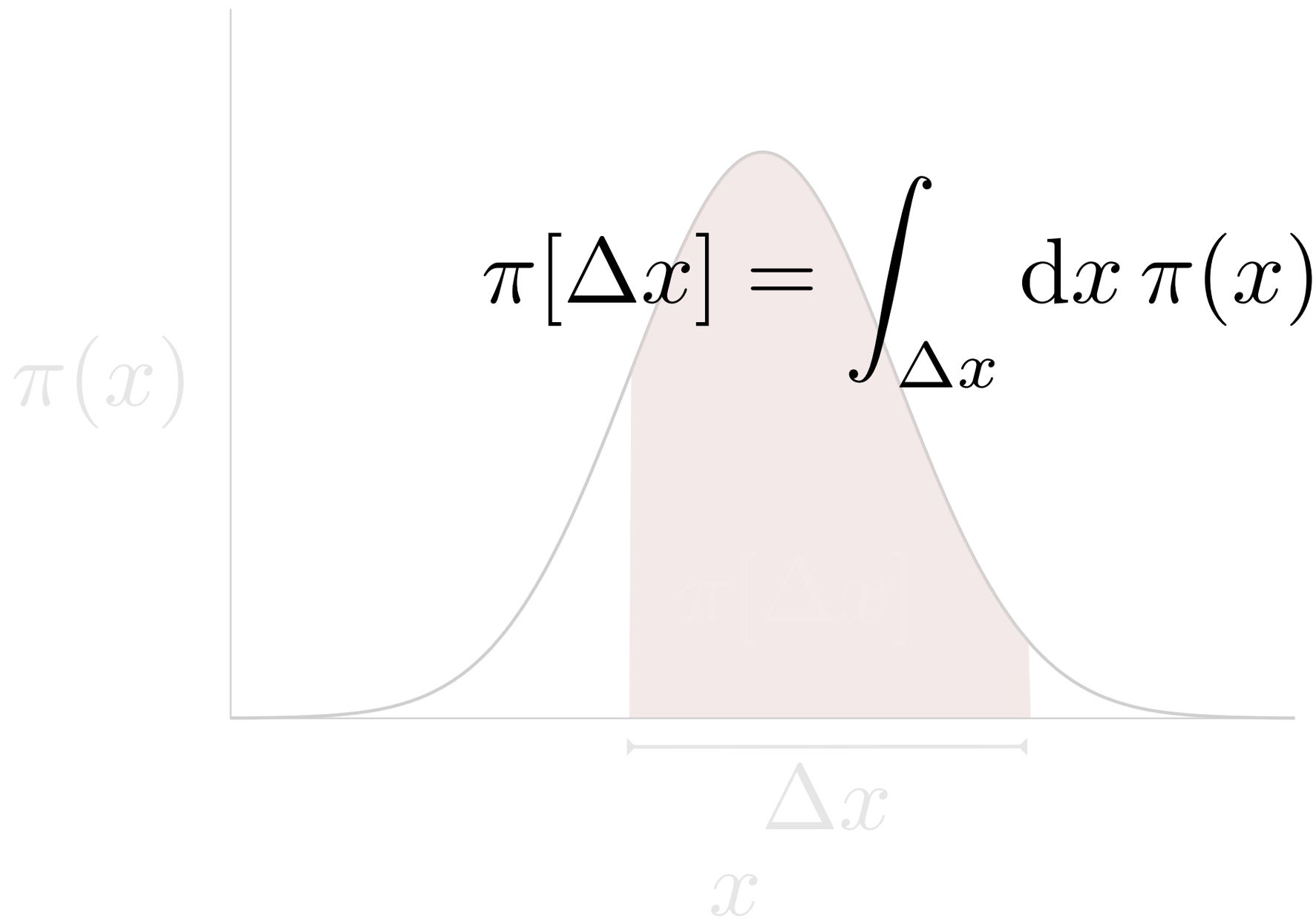
Typically we represent probability distributions with probability density functions which we can *integrate*.



Typically we represent probability distributions with probability density functions which we can *integrate*.



Typically we represent probability distributions with probability density functions which we can *integrate*.



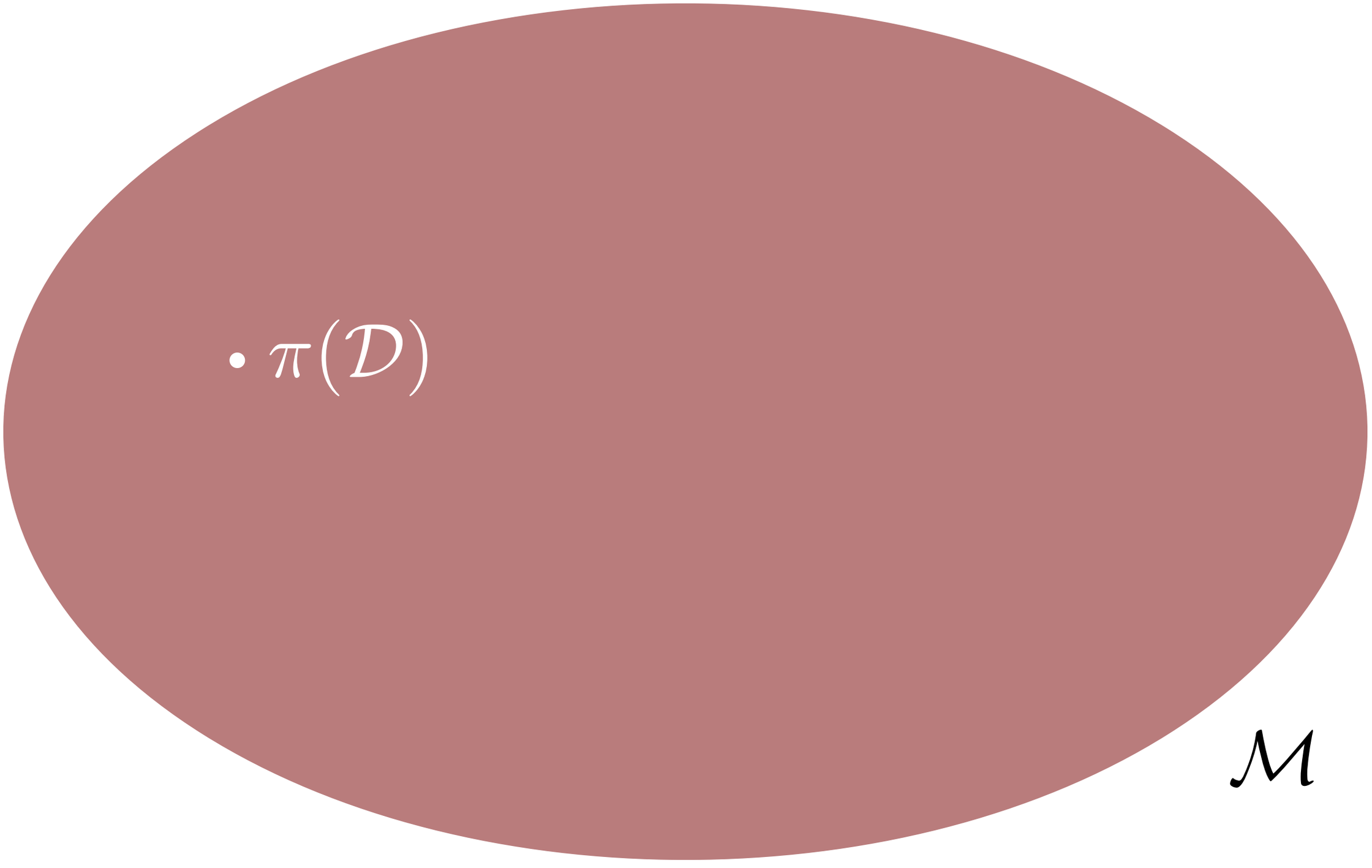
We can also use these probabilities to weight function outputs and define *expectation values*.

$$\mathbb{E}[f] = \int_X \mathrm{d}x\, \pi(x) f(x)$$

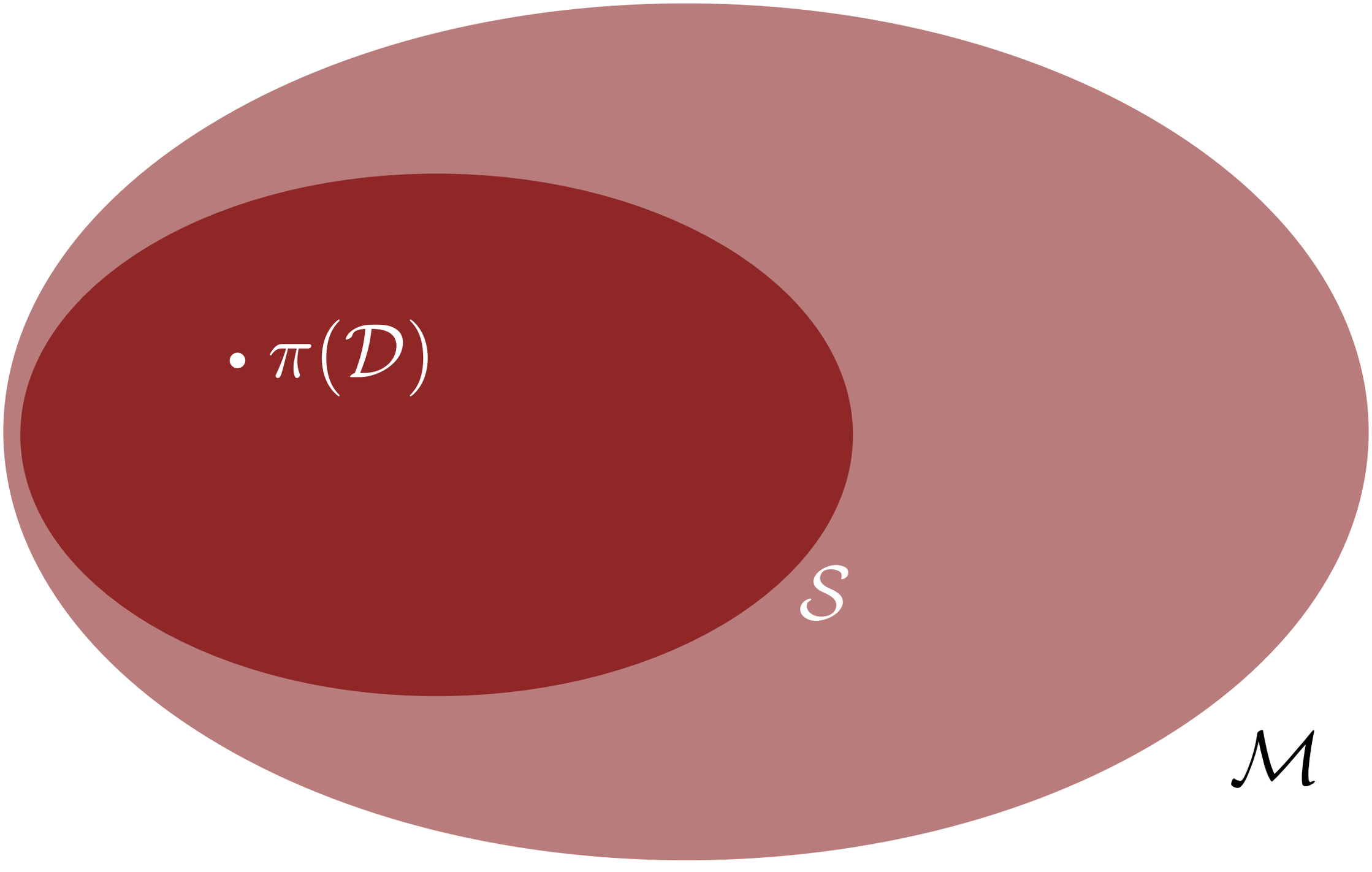
Let's return to our assumed latent data generating process,
a probability distribution over the measurement space.

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

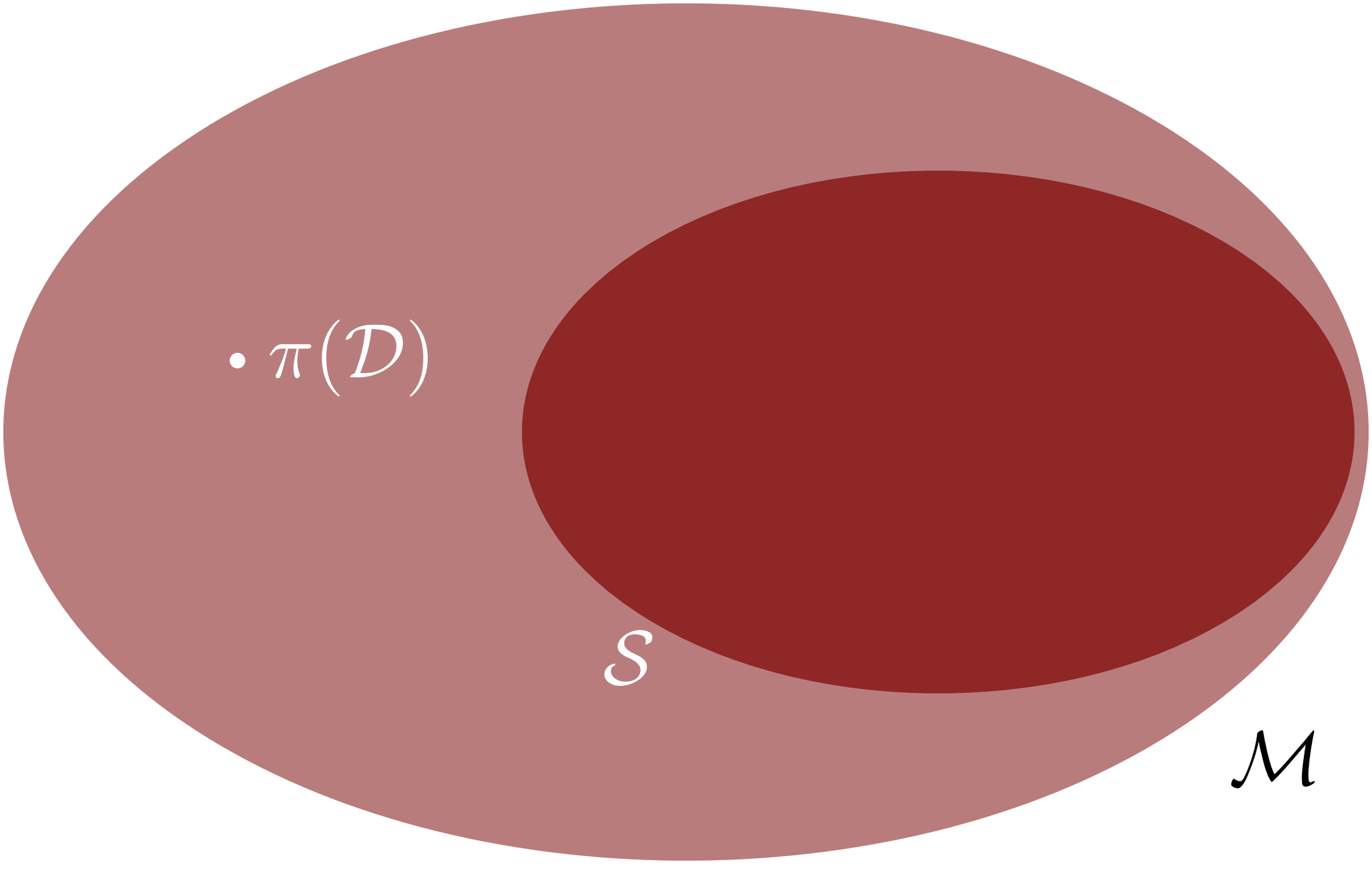
The "true" data generation process must lie in the space of all possible data generating processes, $\mathcal{M}$.



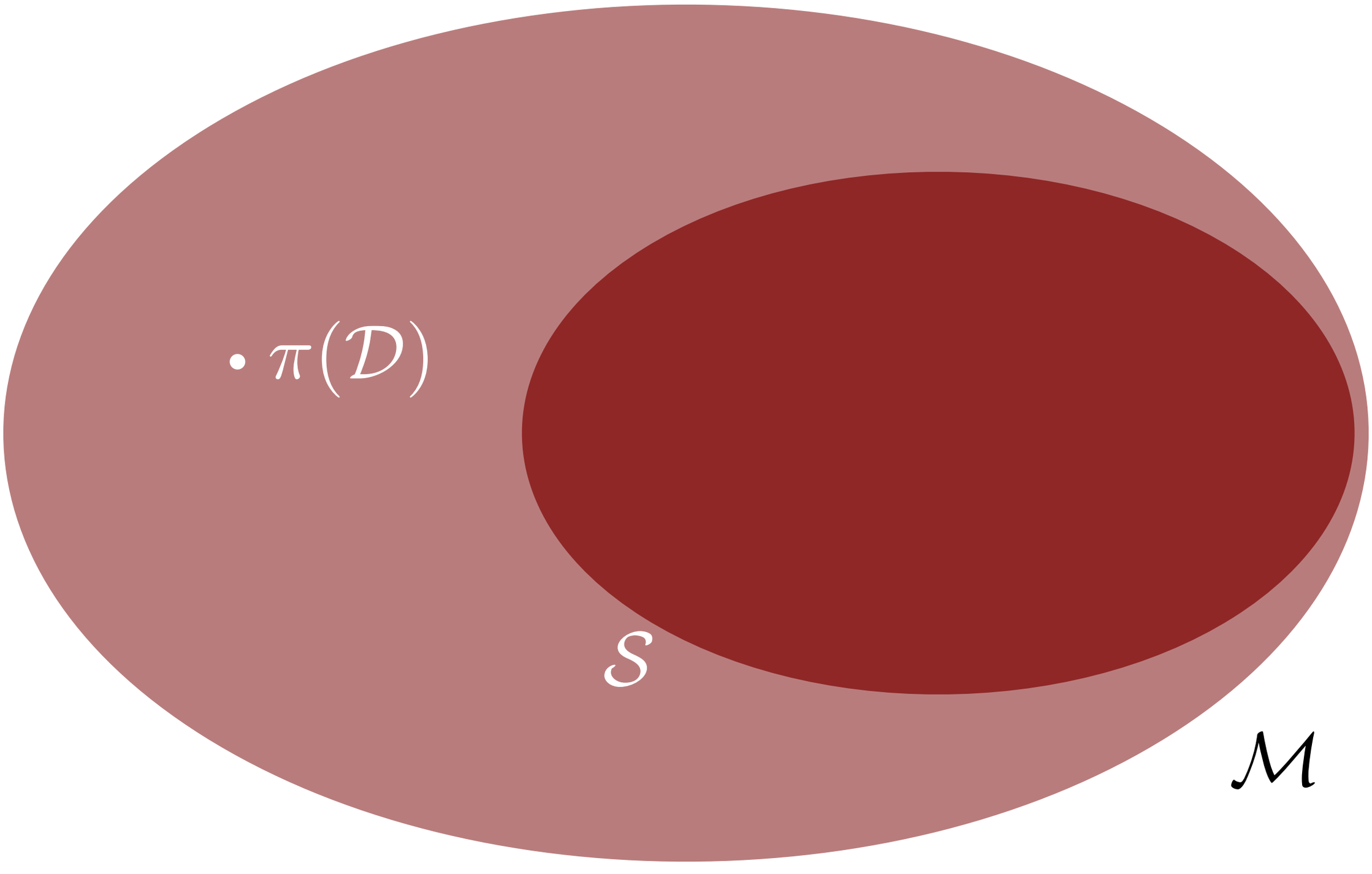
But in practice we have to consider only a small selection of processes $\mathcal{S} \subset \mathcal{M}$, sometimes called the *small world*.



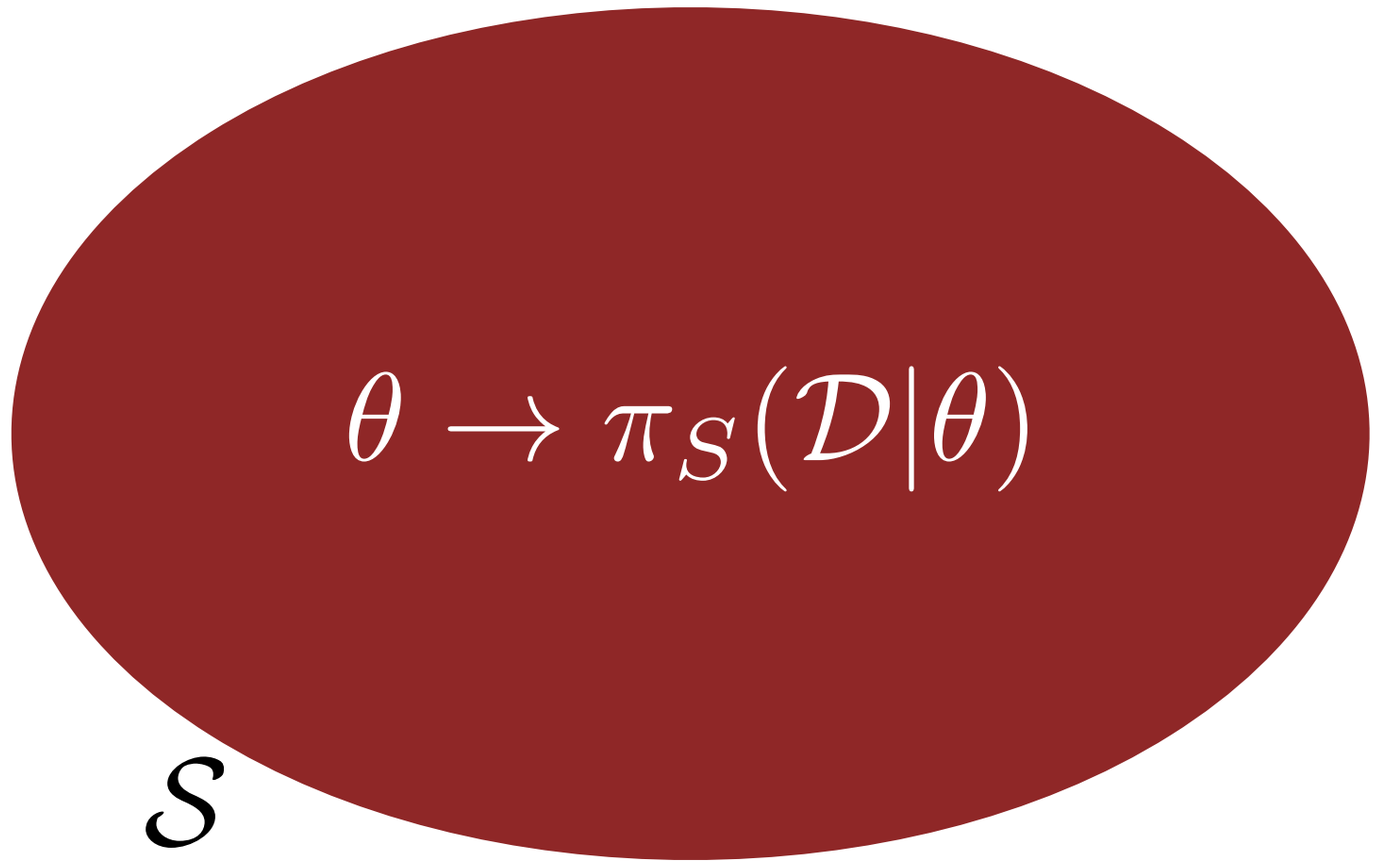
The true process, if it exists, may not be an element of the small world, but our inferences may still be meaningful.



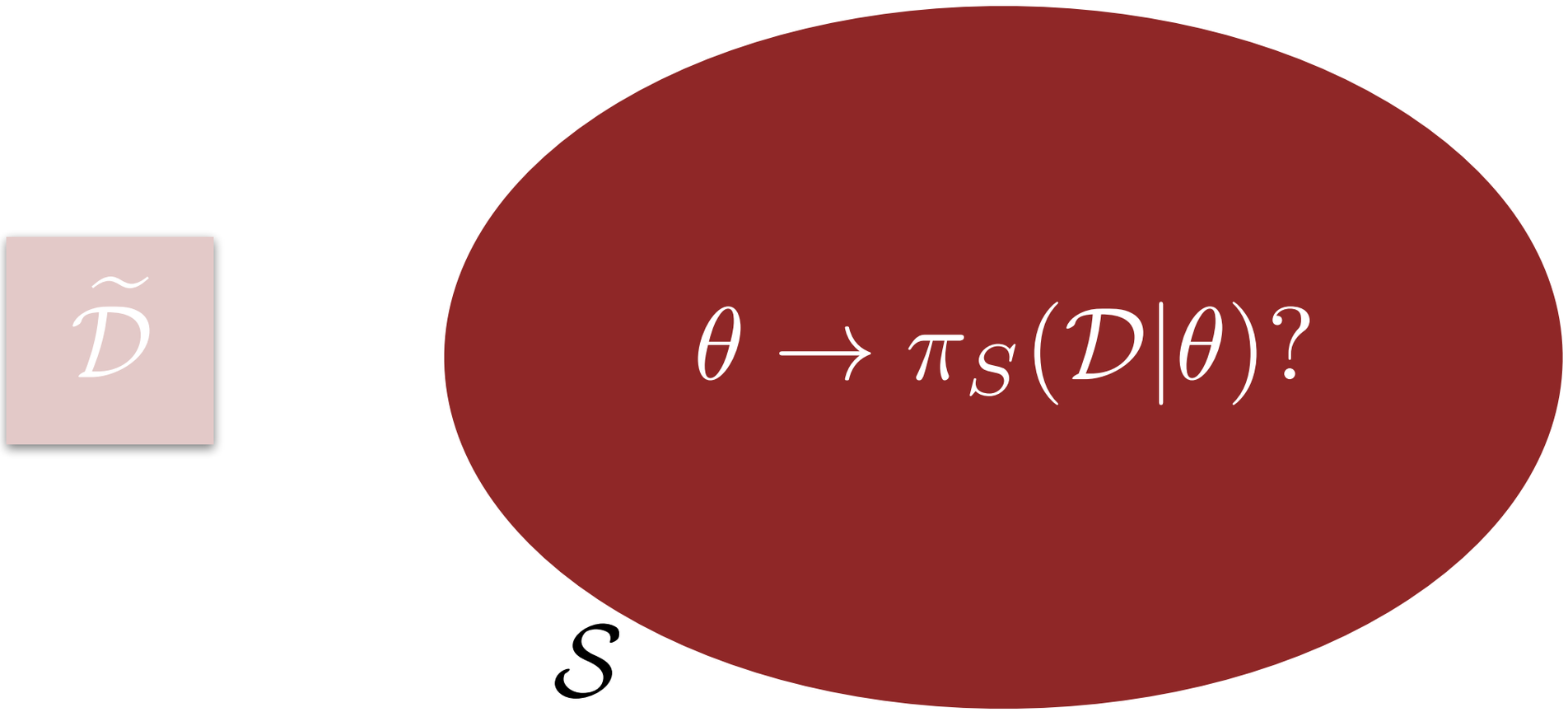
"All models are wrong but some are useful".
-George Box



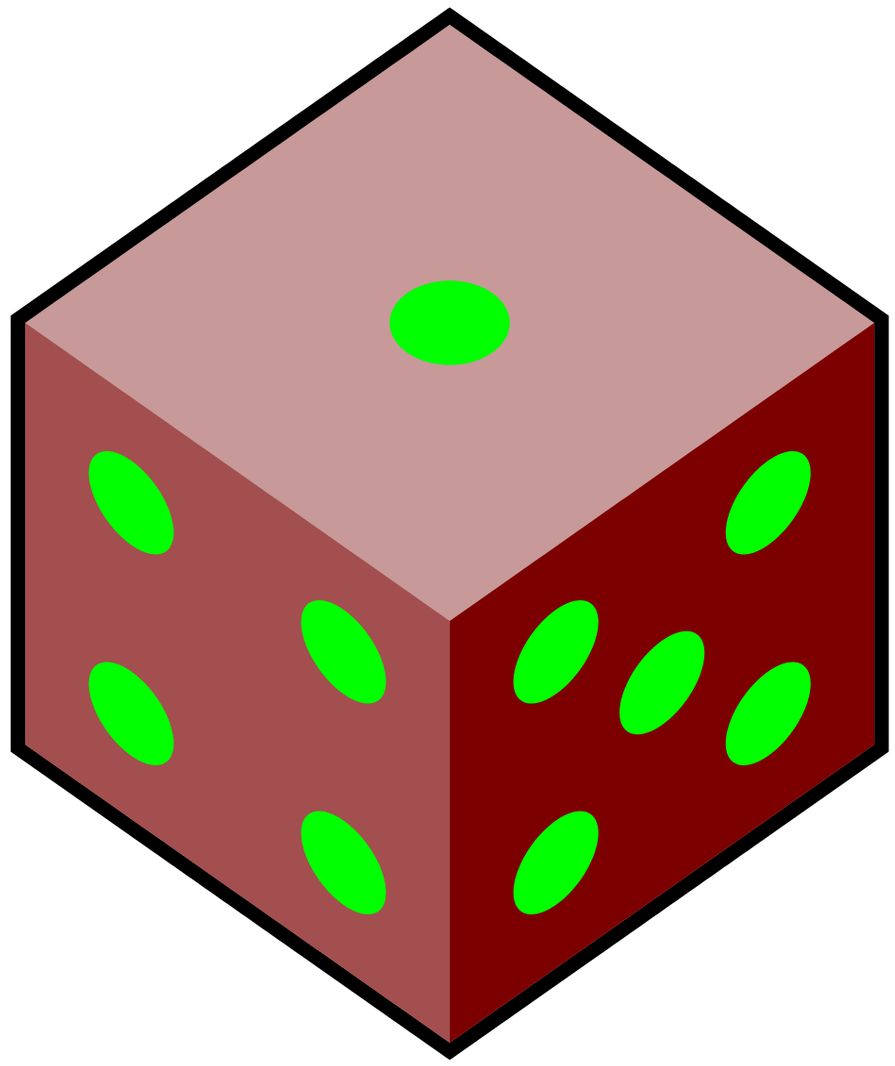
Any inferential model is then a choice of small world,
or a *likelihood* of distributions over measurements.



Inference is the identification of those points in the small world *consistent* with a given measurement.



How we define consistency, however, depends exactly on how we define probability itself.



In *frequentist statistics*, probability is defined in terms of frequencies and so can be applied to only the data.

$$\pi_S(\mathcal{D}|\theta)$$

In *frequentist statistics*, probability is defined in terms of frequencies and so can be applied to only the data.

$$\pi_S(\mathcal{D}|\theta)$$

In *frequentist statistics*, probability is defined in terms of frequencies and so can be applied to only the data.

$$\pi_S(\mathcal{D}|\theta)$$

Frequentist methods compute expectations with respect to the data to identify estimators that work well *on average*.

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

Frequentist methods compute expectations with respect to the data to identify estimators that work well *on average*.

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

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

Frequentist methods compute expectations with respect to the data to identify estimators that work well *on average*.

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

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

Bayesian inference, however, treats probability as a general measure of uncertainty.

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

$$\pi_S(\mathcal{D}|\theta)$$

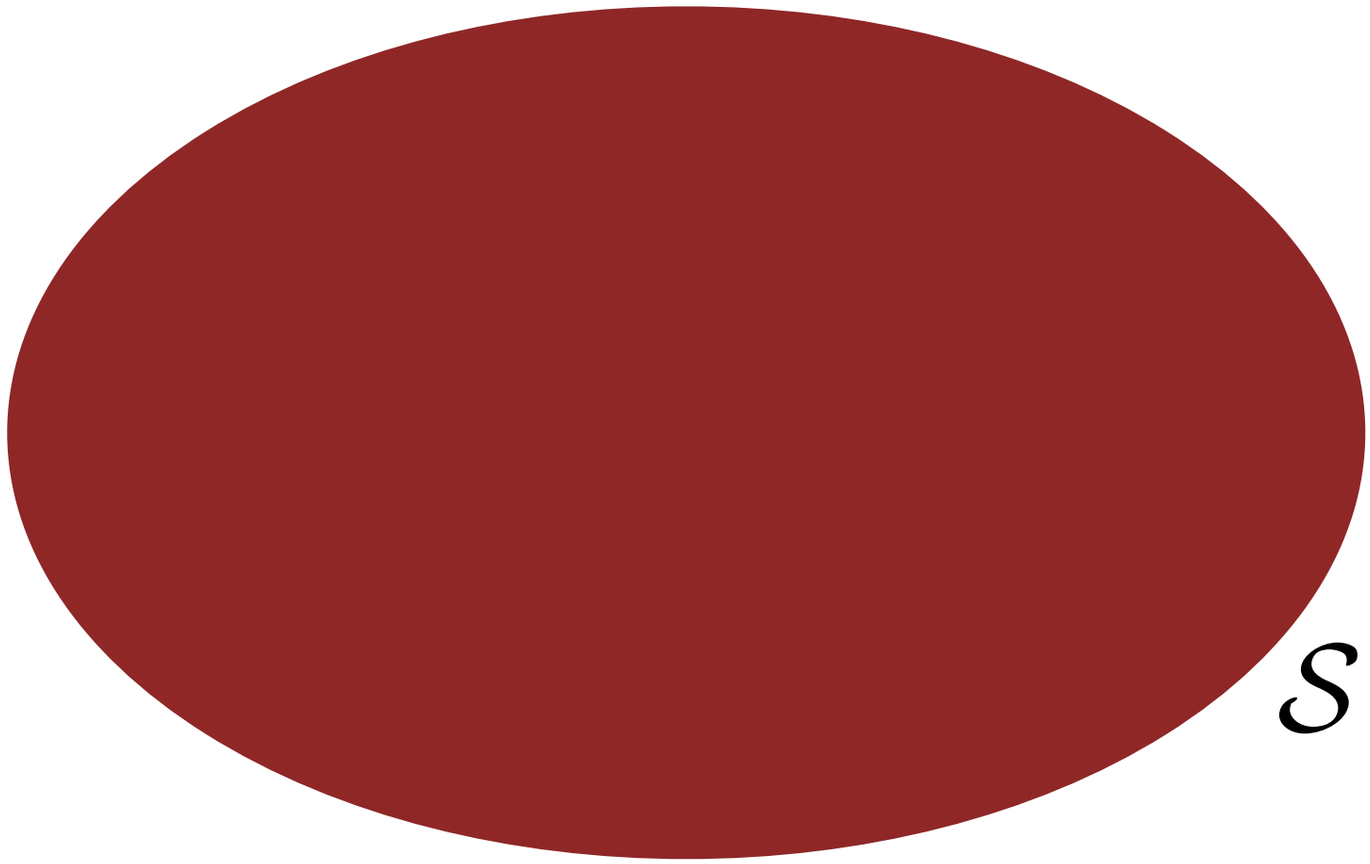
Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi_S(\mathcal{D}|\theta)$$

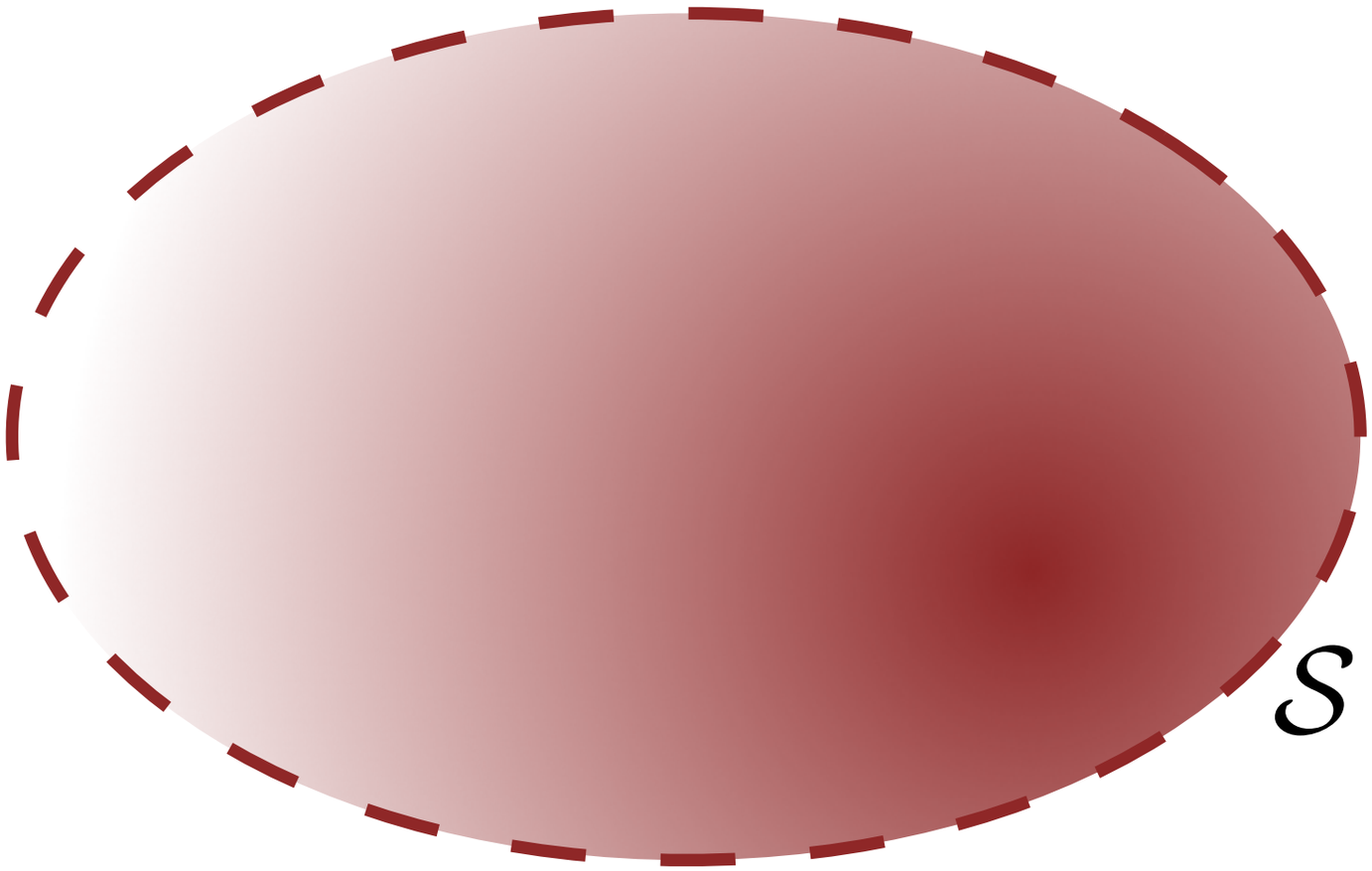
Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi_S(\mathcal{D}|\boldsymbol{\theta})$$

In this more general perspective we quantify consistency using a probability distribution over the small world.



In this more general perspective we quantify consistency using a probability distribution over the small world.



Uncertainty within the model is just
an application of *Bayes' Theorem*.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

The *prior* incorporates any prior knowledge about the model space before the data are measured.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

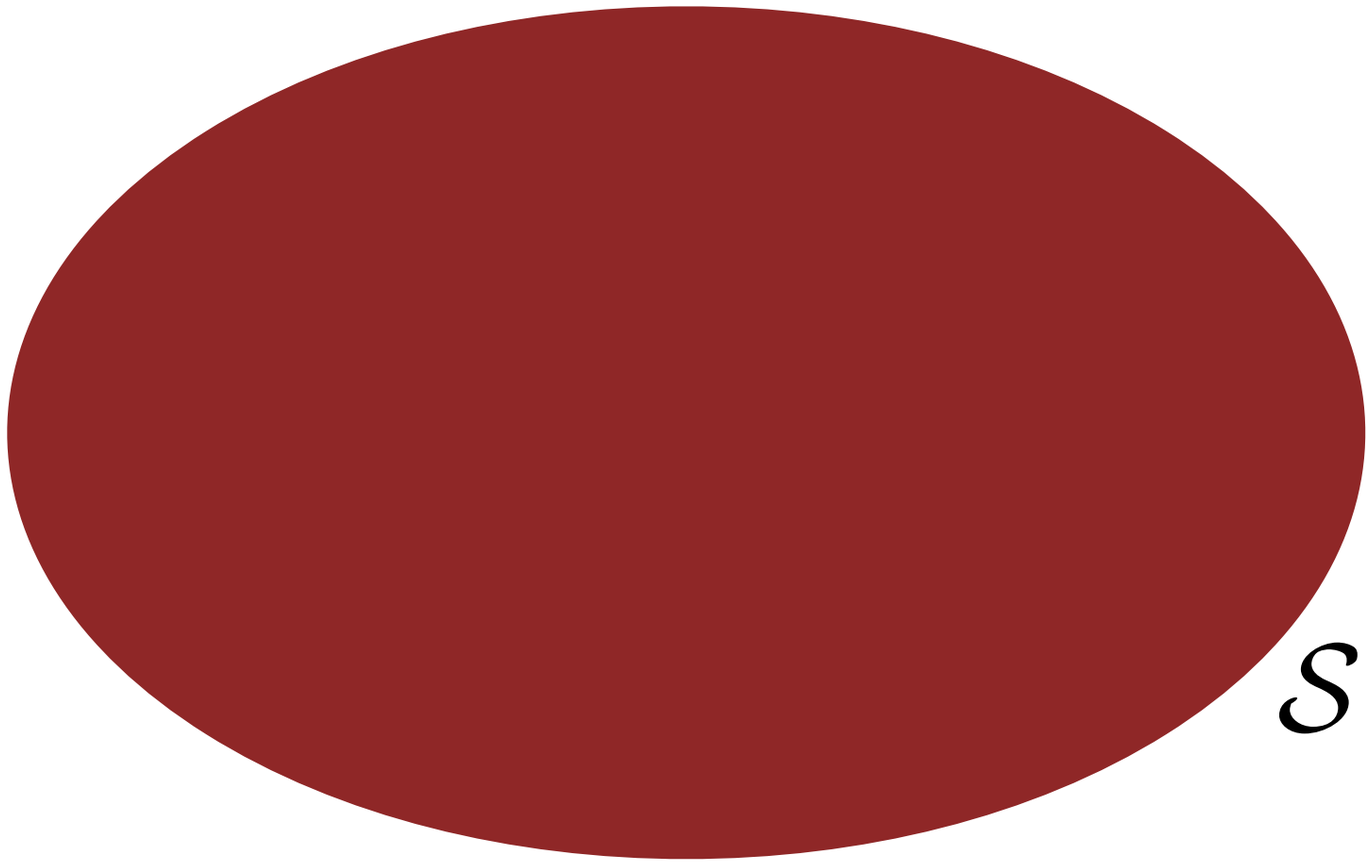
The *likelihood* is similar to the frequentist approach:
a generative model of the data.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

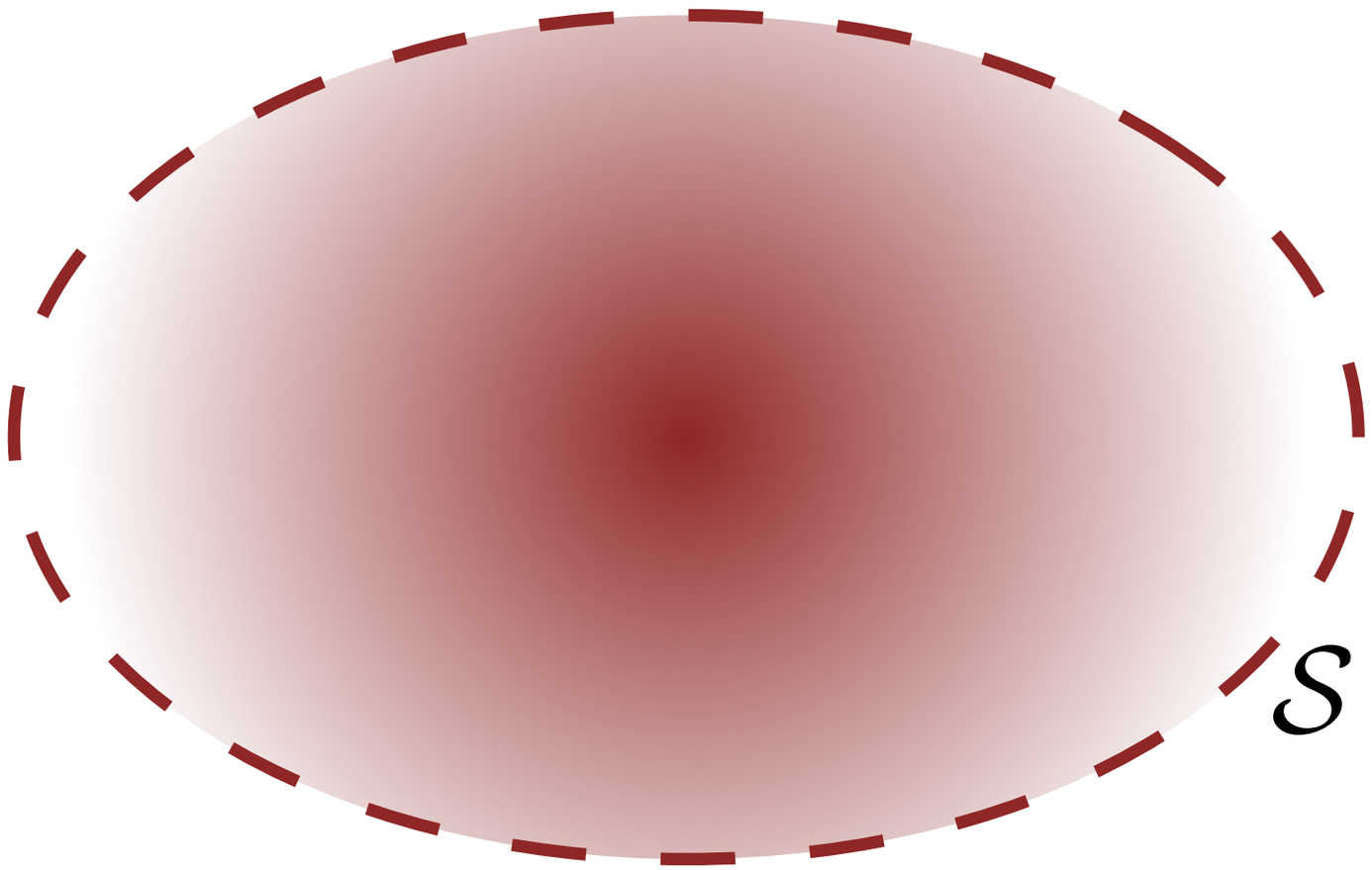
The *posterior* incorporates both the information into the prior and the data into a final uncertainty in the model.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

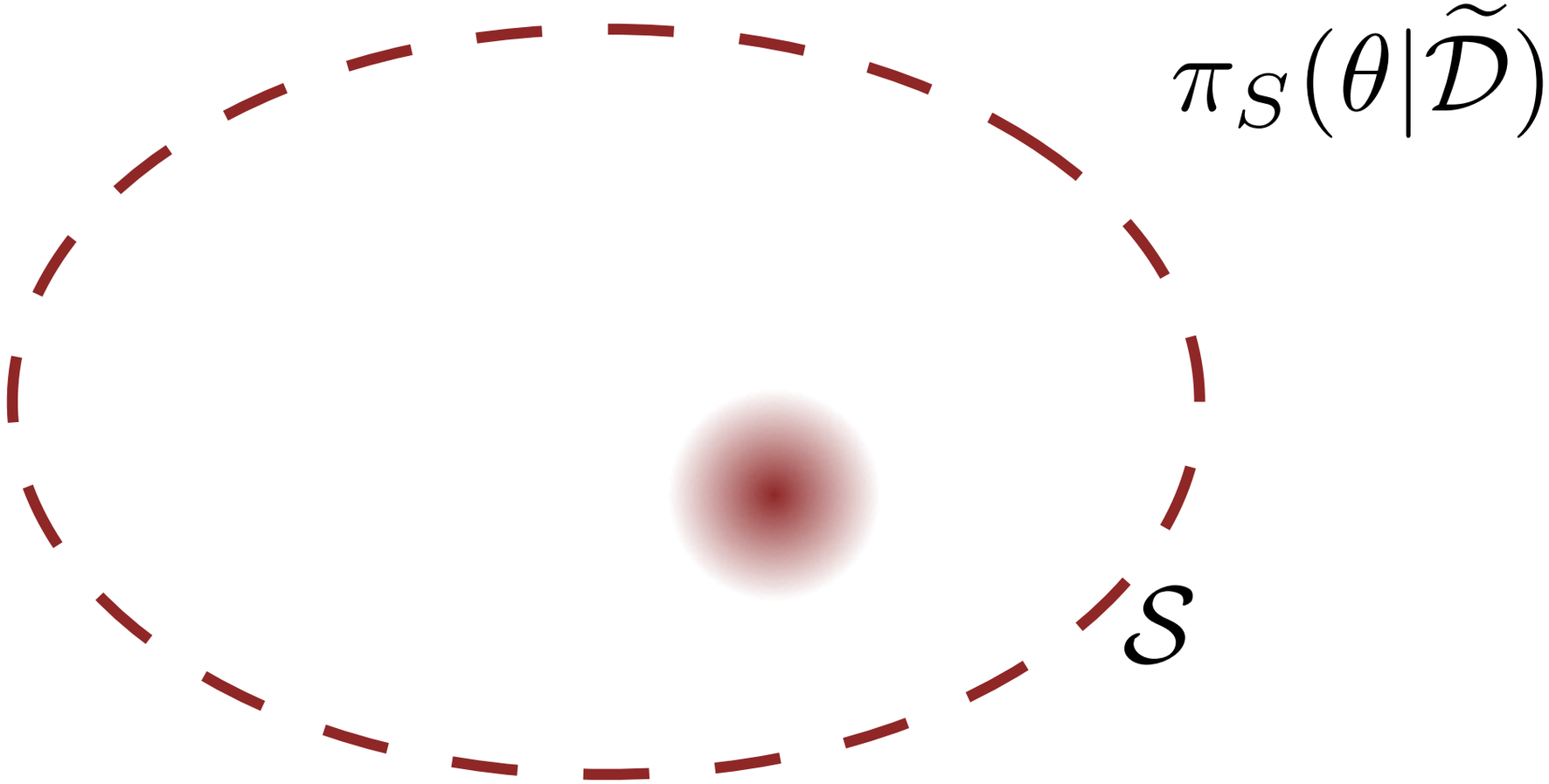
Conditioning on the measurement reduces our uncertainty amongst the data generation processes.



Conditioning on the measurement reduces our uncertainty amongst the data generation processes.



Conditioning on the measurement reduces our uncertainty amongst the data generation processes.



From a Bayesian perspective, all inferential questions are answered by expectations.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta)$$

Expectations include means and variances for posterior summaries and expected utility for decision making.

$$\mu = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta$$

Expectations include means and variances for posterior summaries and expected utility for decision making.

$$\mu = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta$$

$$\sigma^2 = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta^2 - \mu^2$$

Expectations include means and variances for posterior summaries and expected utility for decision making.

$$\mu = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta$$

$$\sigma^2 = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta^2 - \mu^2$$

$$U(A) = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, U(A, \theta)$$

Expectations also allow us to incorporate systematic effects via *marginalization*.

$$\pi_S(\theta_1, \theta_2, \ldots, \theta_n | \mathcal{D})$$

Expectations also allow us to incorporate systematic effects via *marginalization*.

$$\pi_S(\theta_1, \theta_2, \ldots, \theta_n | \mathcal{D})$$

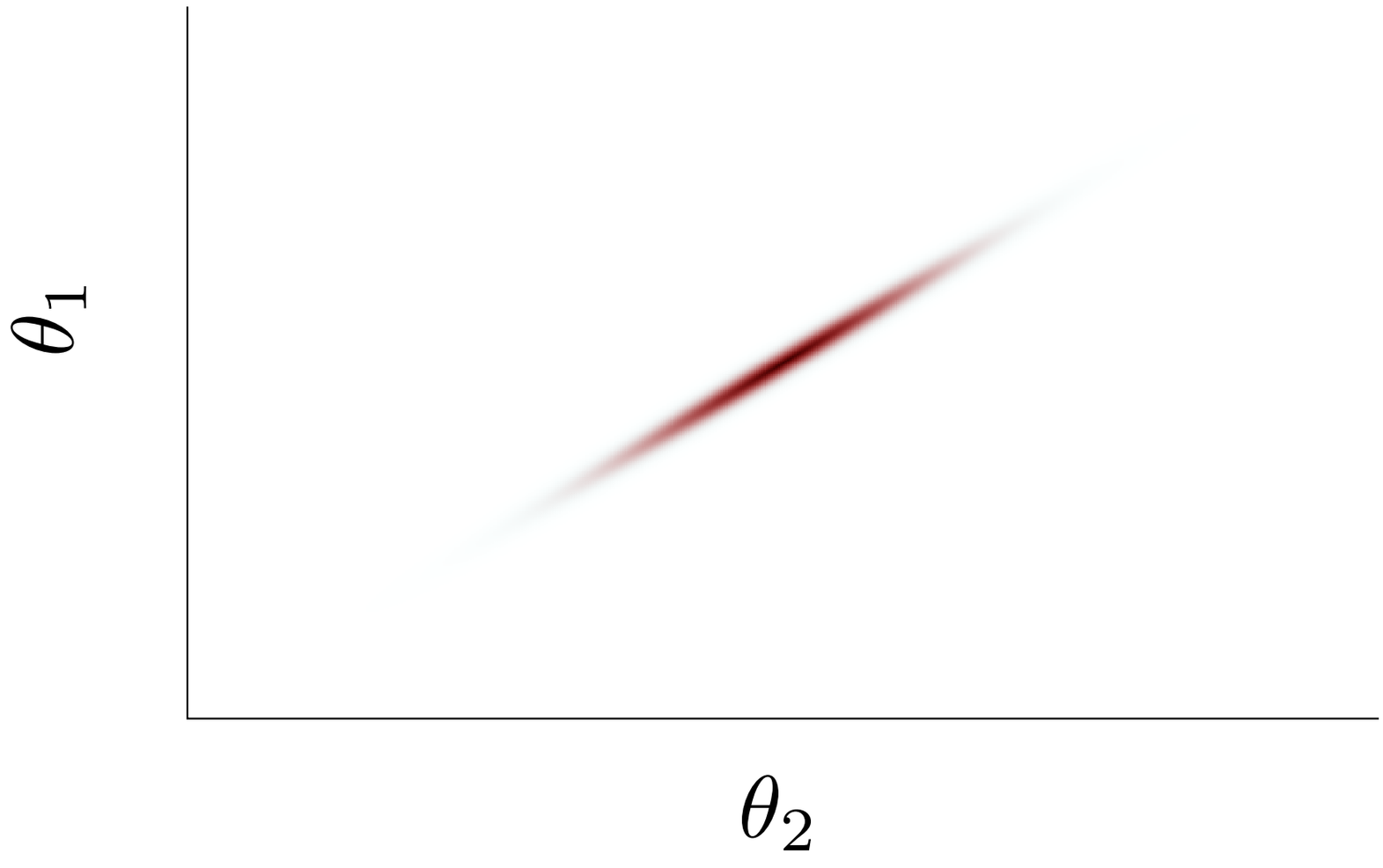
Expectations also allow us to incorporate systematic effects via *marginalization*.

$$\pi_S(\theta_1, \theta_2, \ldots, \theta_n | \mathcal{D})$$

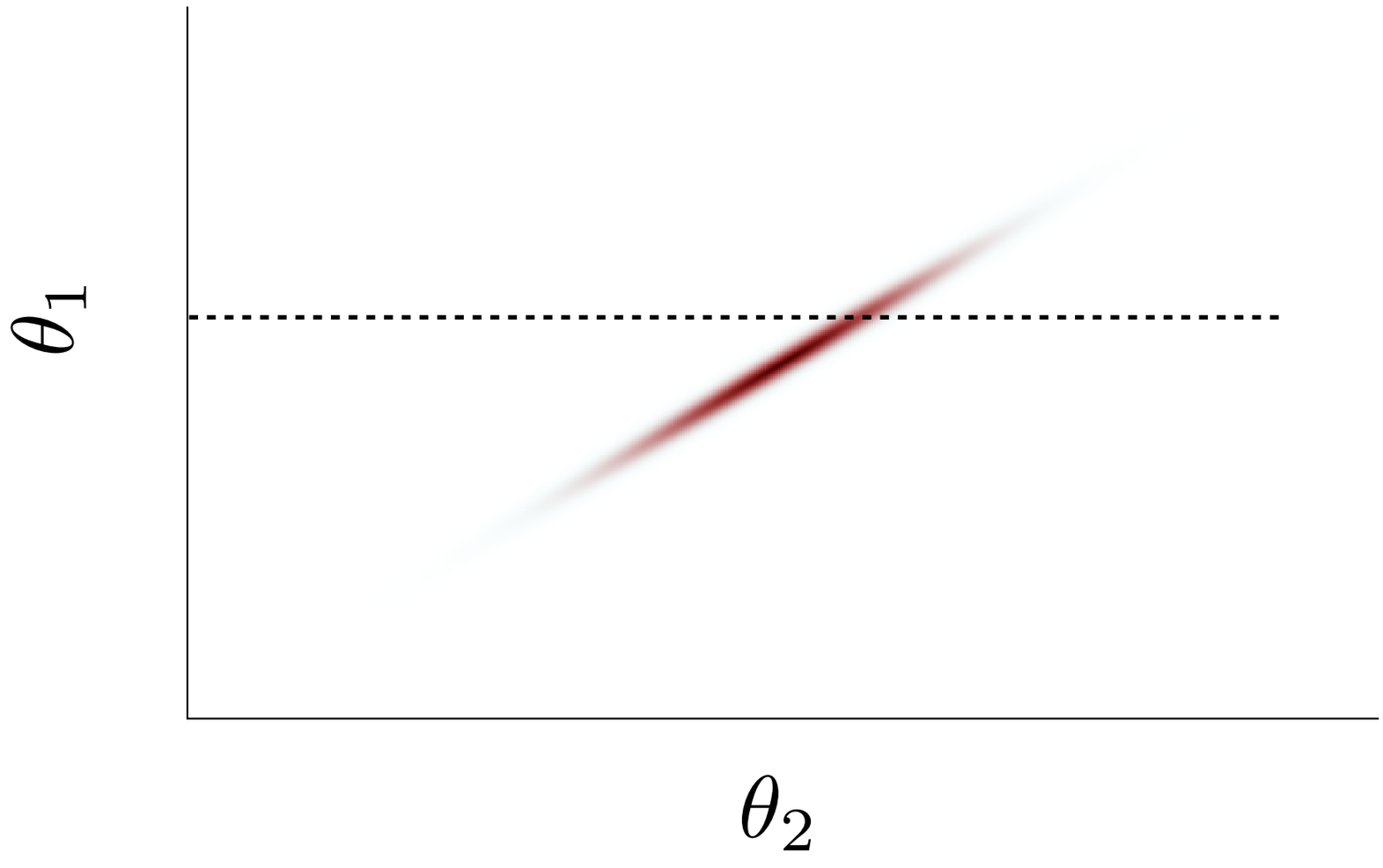
Expectations also allow us to incorporate systematic effects via *marginalization*.

$$\pi_S(\theta_2, \ldots, \theta_n|\widetilde{\mathcal{D}}) = \int \mathrm{d}\theta_1 \, \pi_S(\theta_1, \theta_2, \ldots, \theta_n|\widetilde{\mathcal{D}})$$

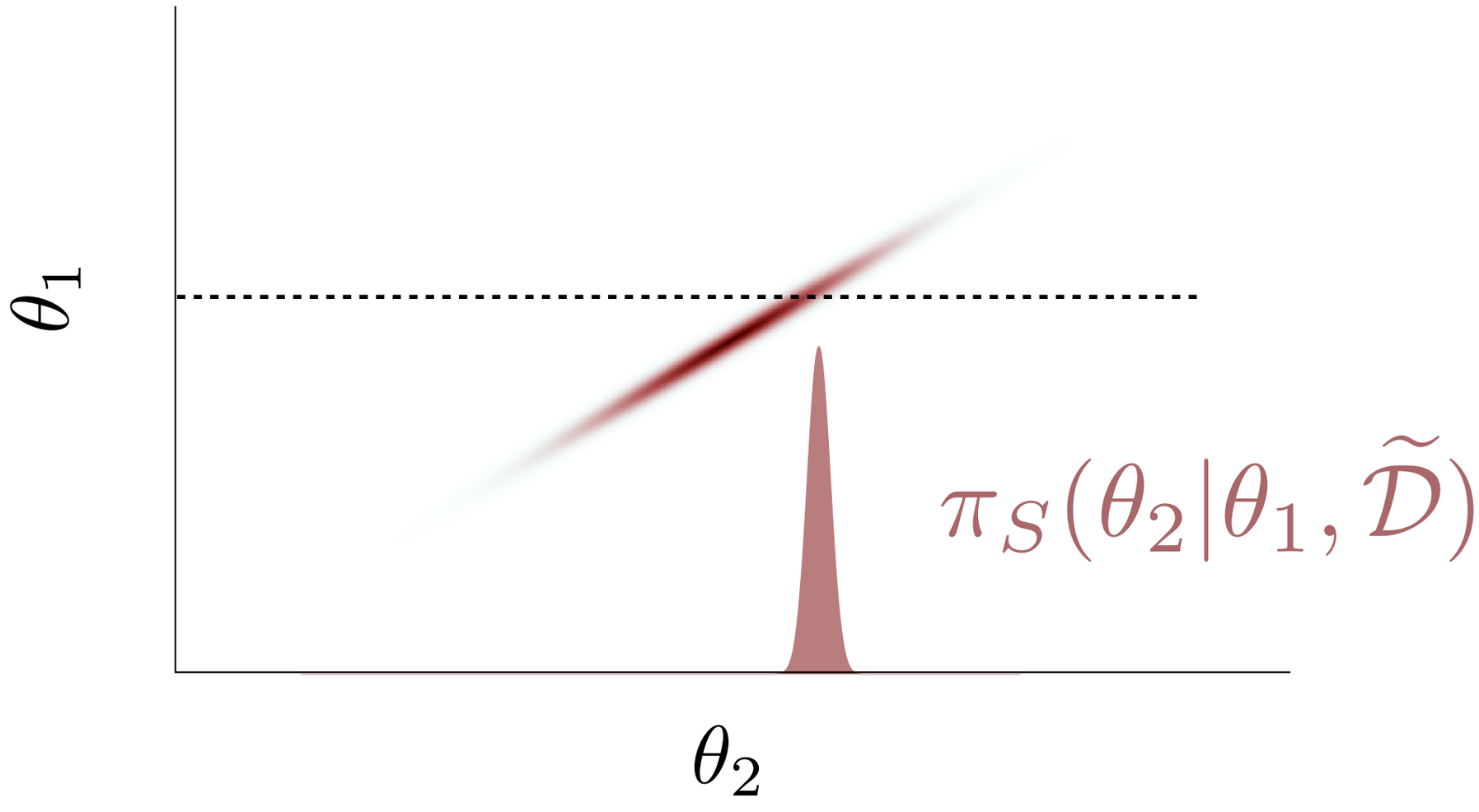
Expectations also allow us to incorporate systematic effects via *marginalization*.



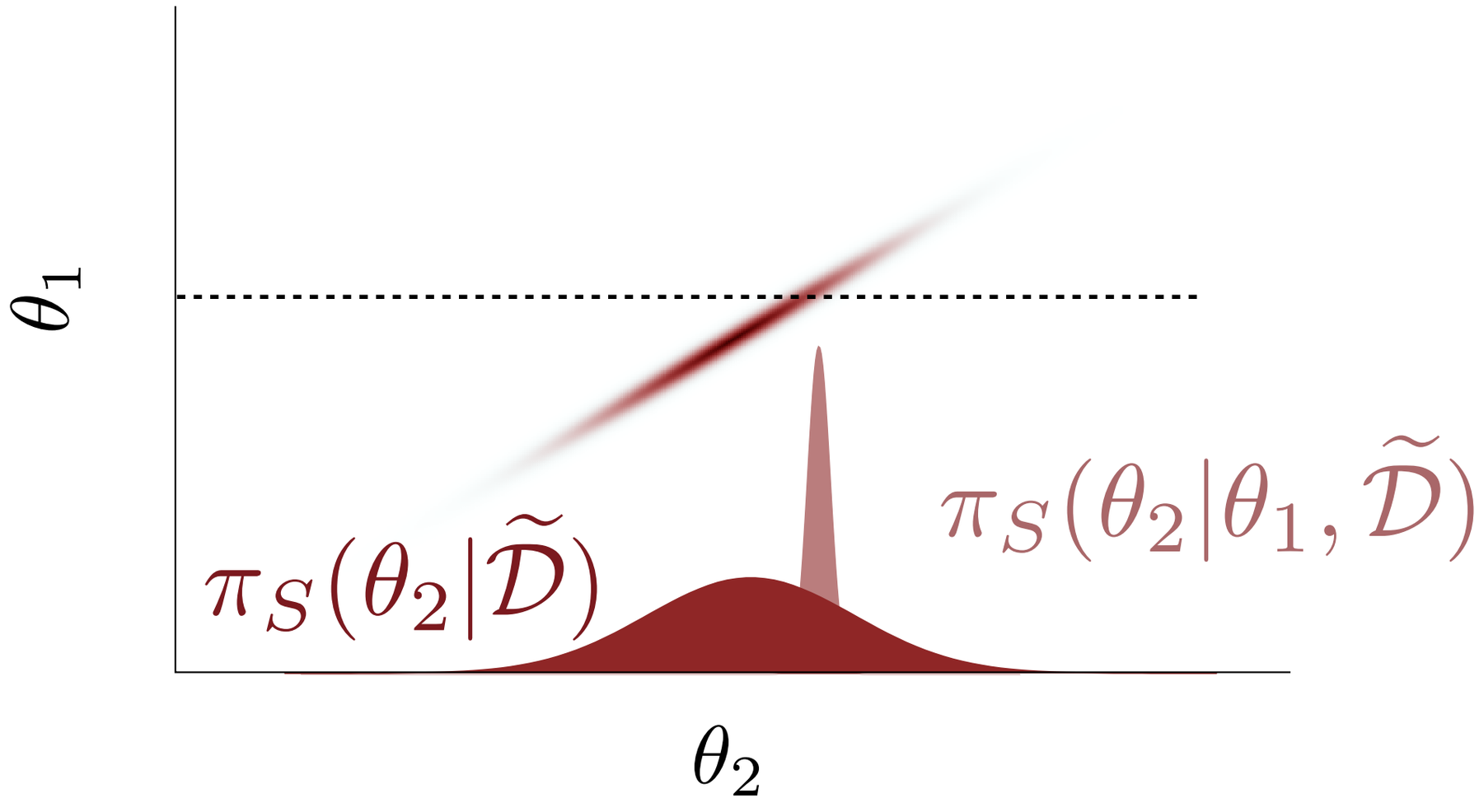
Expectations also allow us to incorporate systematic effects via *marginalization*.



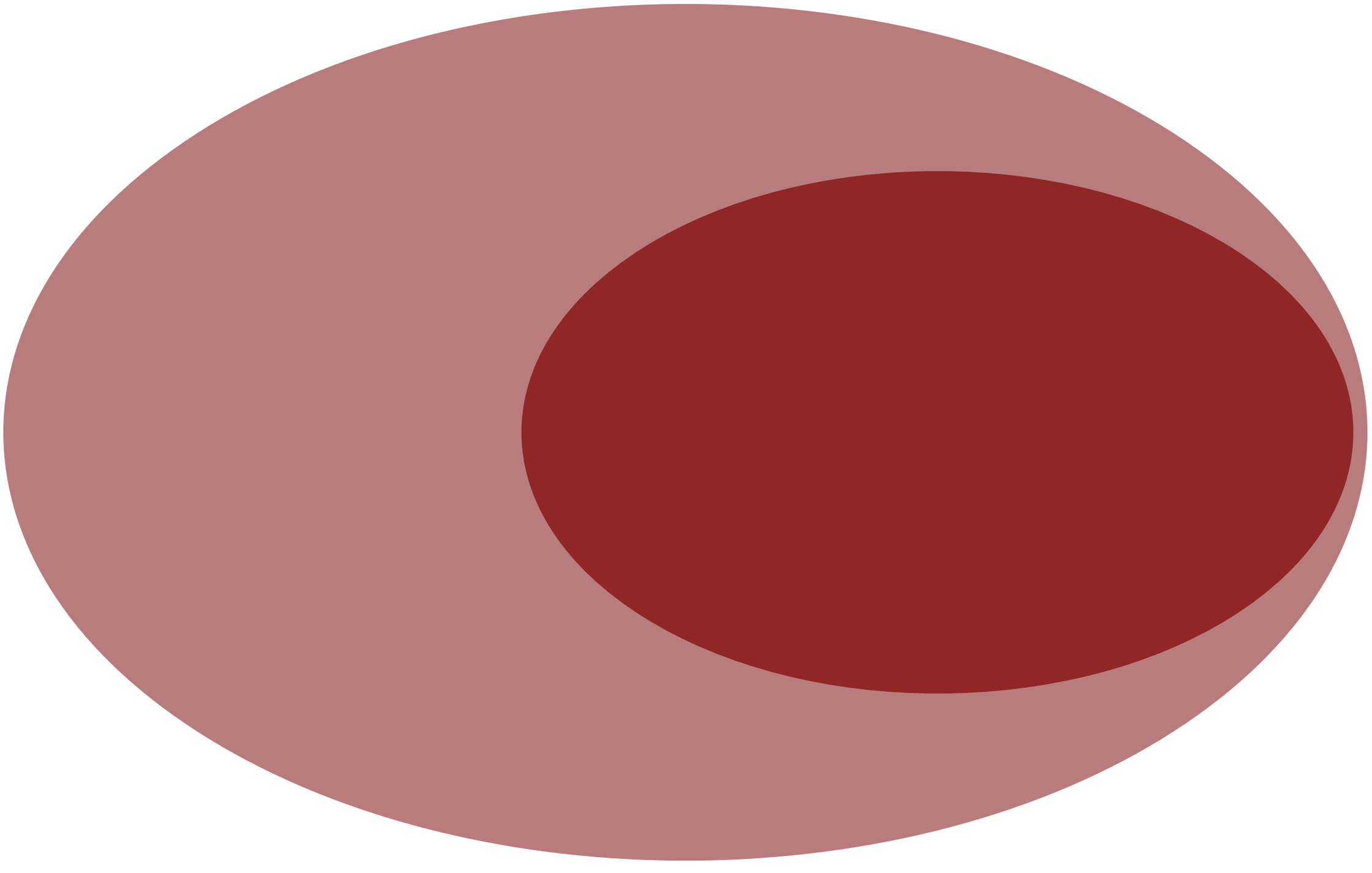
Expectations also allow us to incorporate systematic effects via *marginalization*.



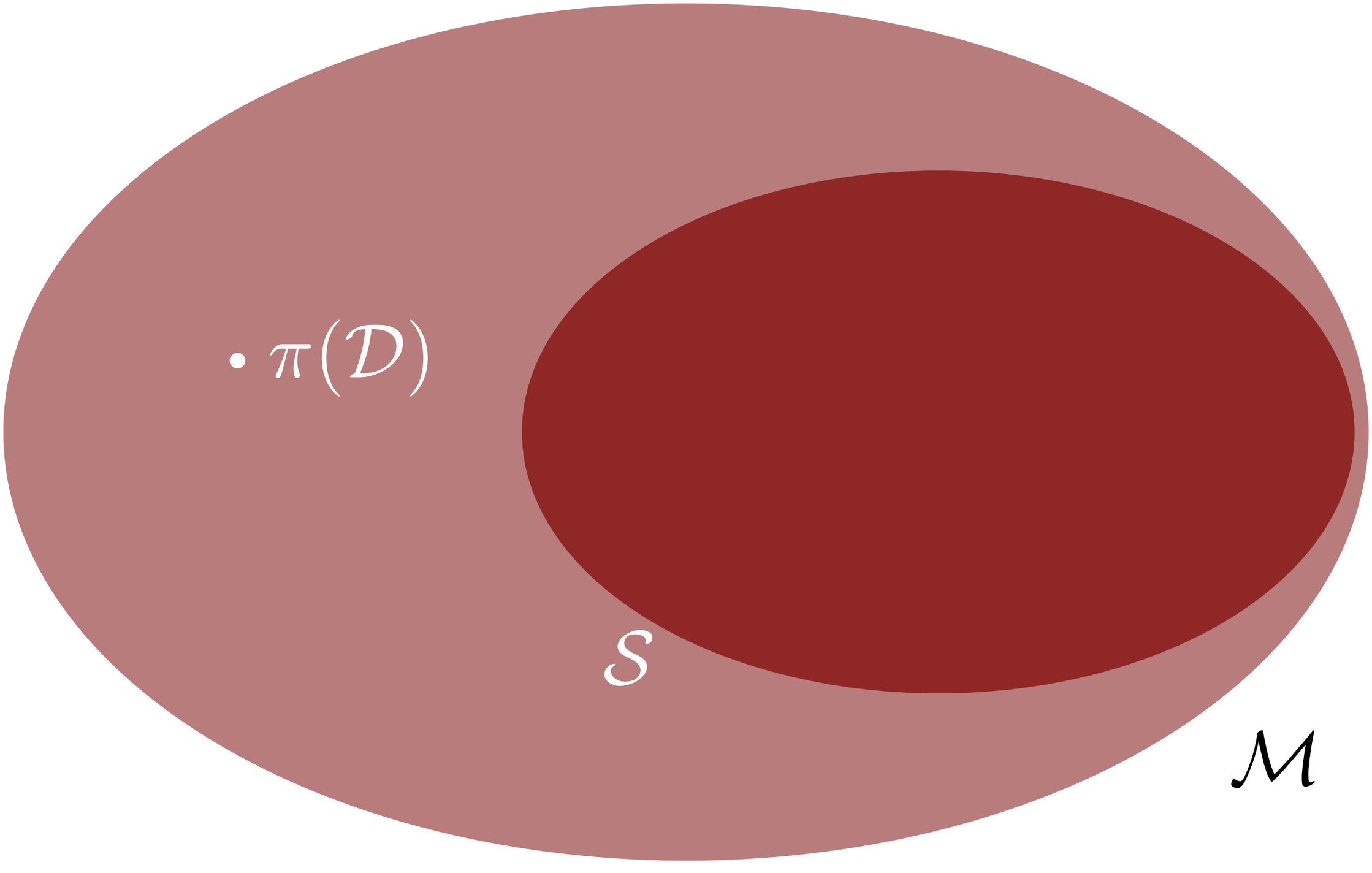
Expectations also allow us to incorporate systematic effects via *marginalization*.



Predictive Model Comparison



The only way to test the assumptions themselves
is by considering *predictive performance*.



The average data generation process over the entire model is given by the *posterior predictive distribution*.

$$\pi_S(\mathcal{D}|\theta)$$

The average data generation process over the entire model is given by the *posterior predictive distribution*.

$$\pi_S(\mathcal{D}|\widetilde{\mathcal{D}}) = \int \mathrm{d}\theta\, \pi_S(\mathcal{D}|\theta)\pi_S(\theta|\widetilde{\mathcal{D}})$$

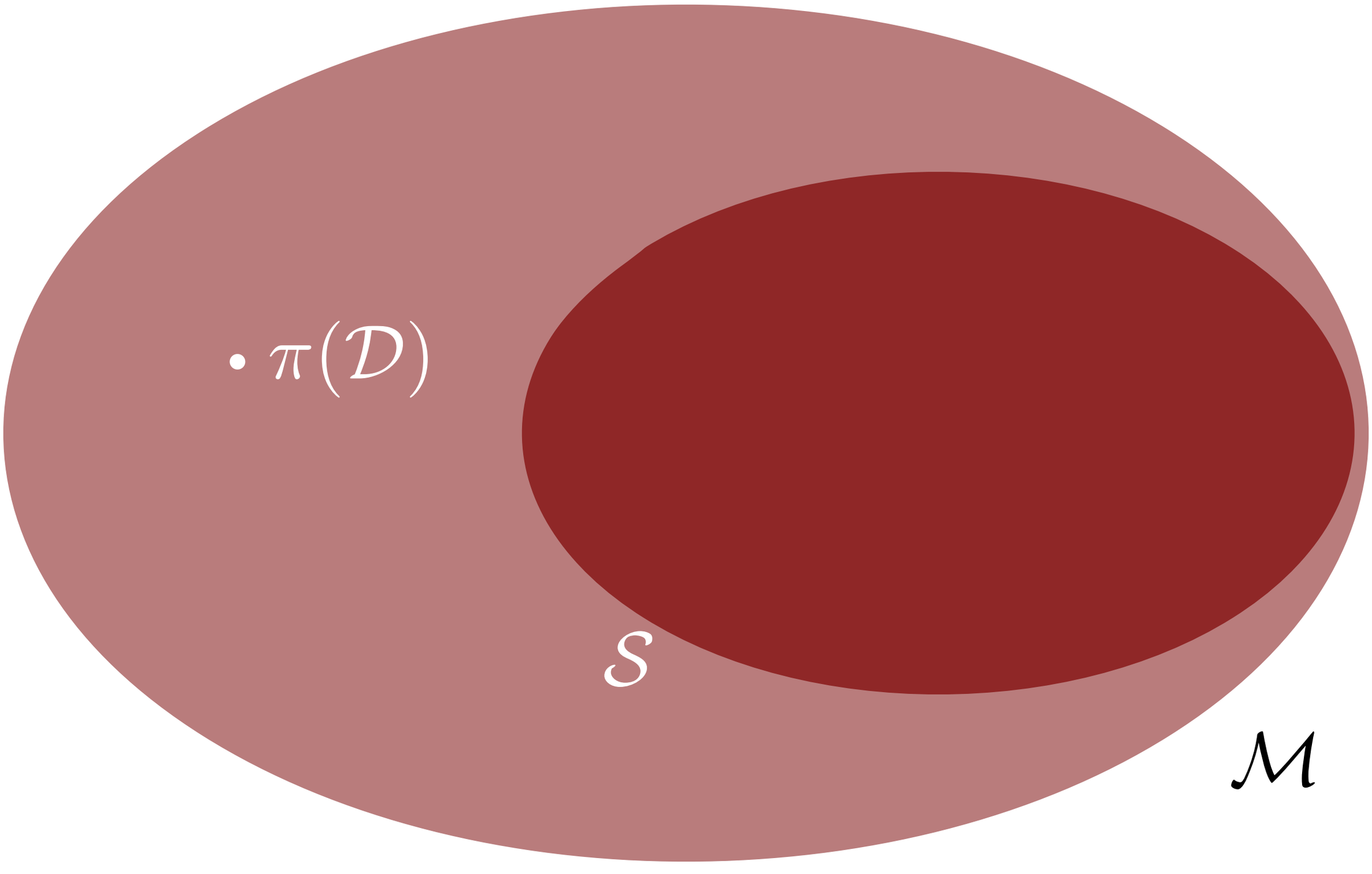
Comparing the posterior predictive distribution to the true data generation process tests our assumptions.

$$\pi_S(\mathcal{D}|\mathcal{D}) \quad \overset{?}{\approx} \quad \pi(\mathcal{D})$$

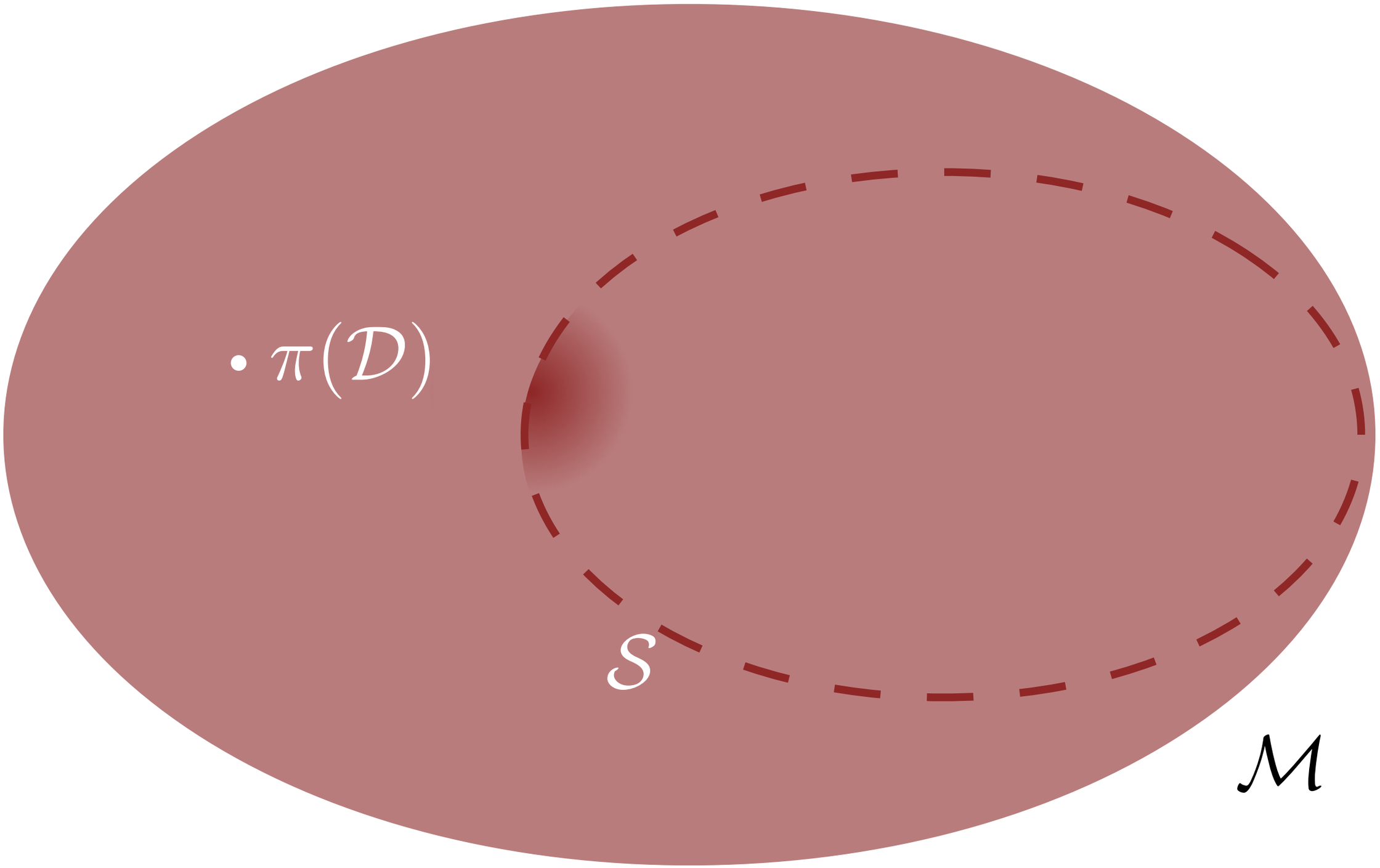
Comparing the posterior predictive distribution to the true data generation process tests our assumptions.

$$\pi_S(\mathcal{D}|\mathcal{D}) \quad ? \quad \mathcal{D}$$

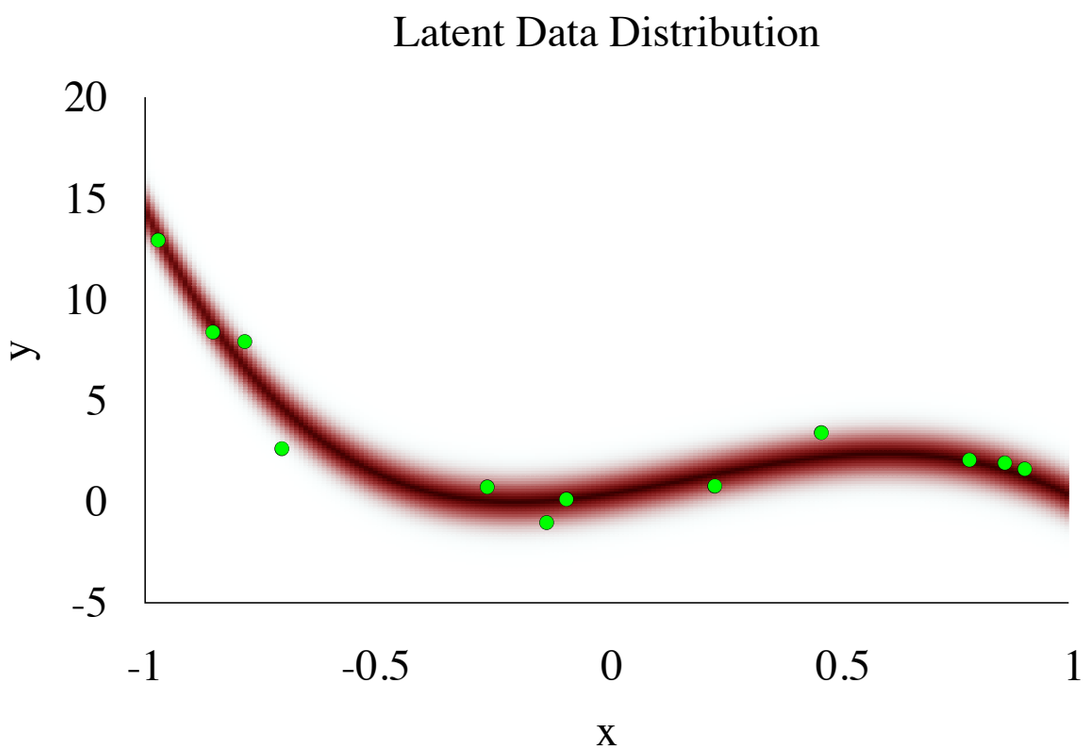
When the small world does not contain the latent data generating process our models will, in general, misfit.

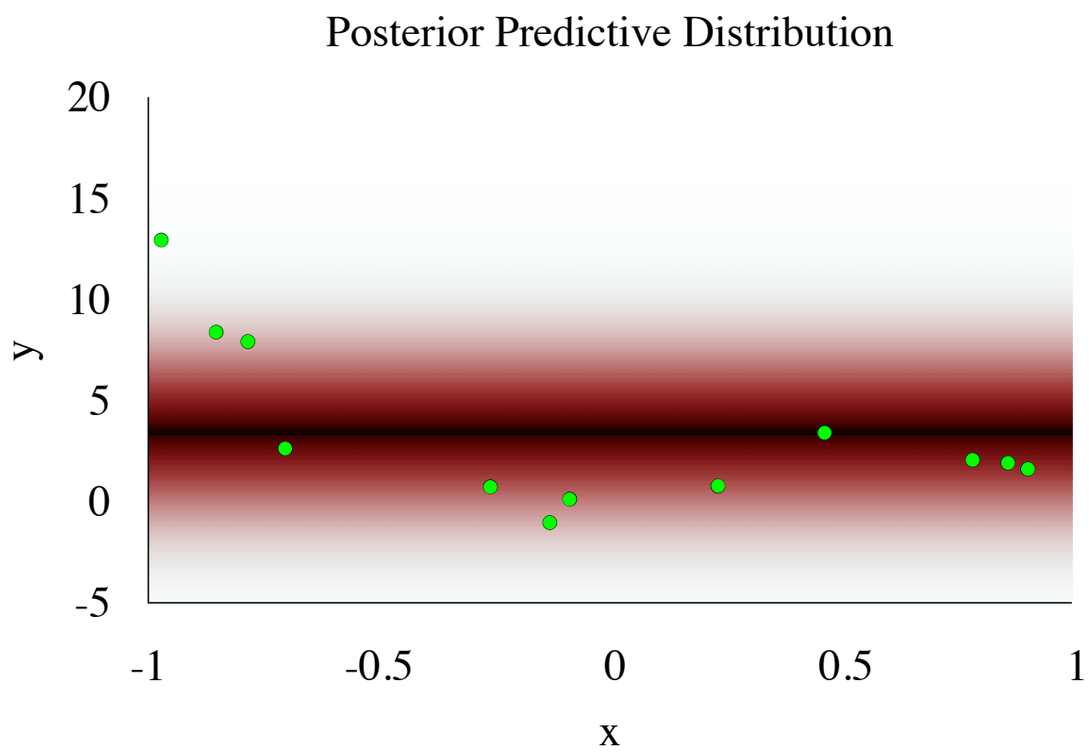


When the small world does not contain the latent data generating process our models will, in general, misfit.

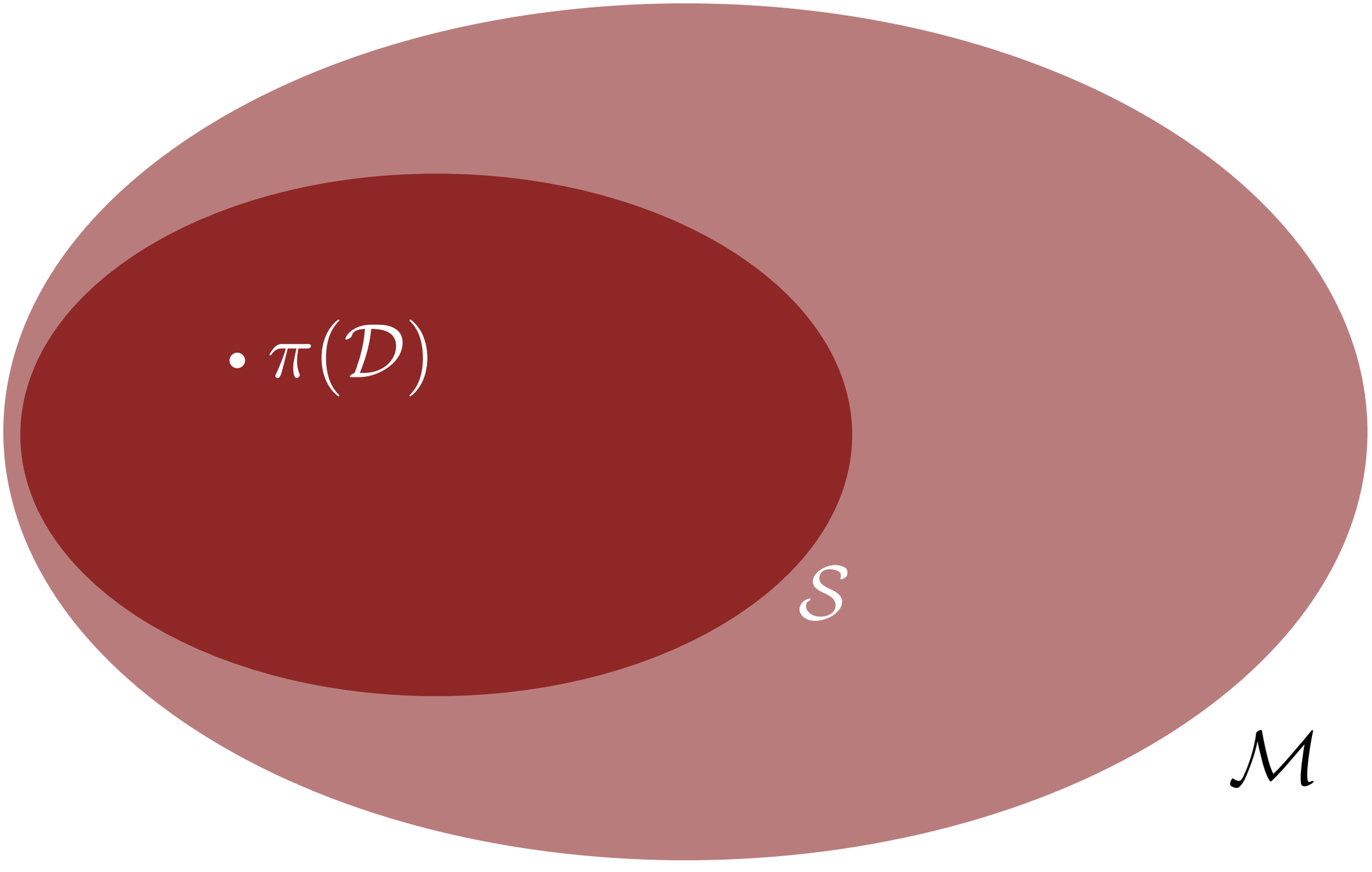


Fortunately, misfit results in tension between predictive distributions and measurements.

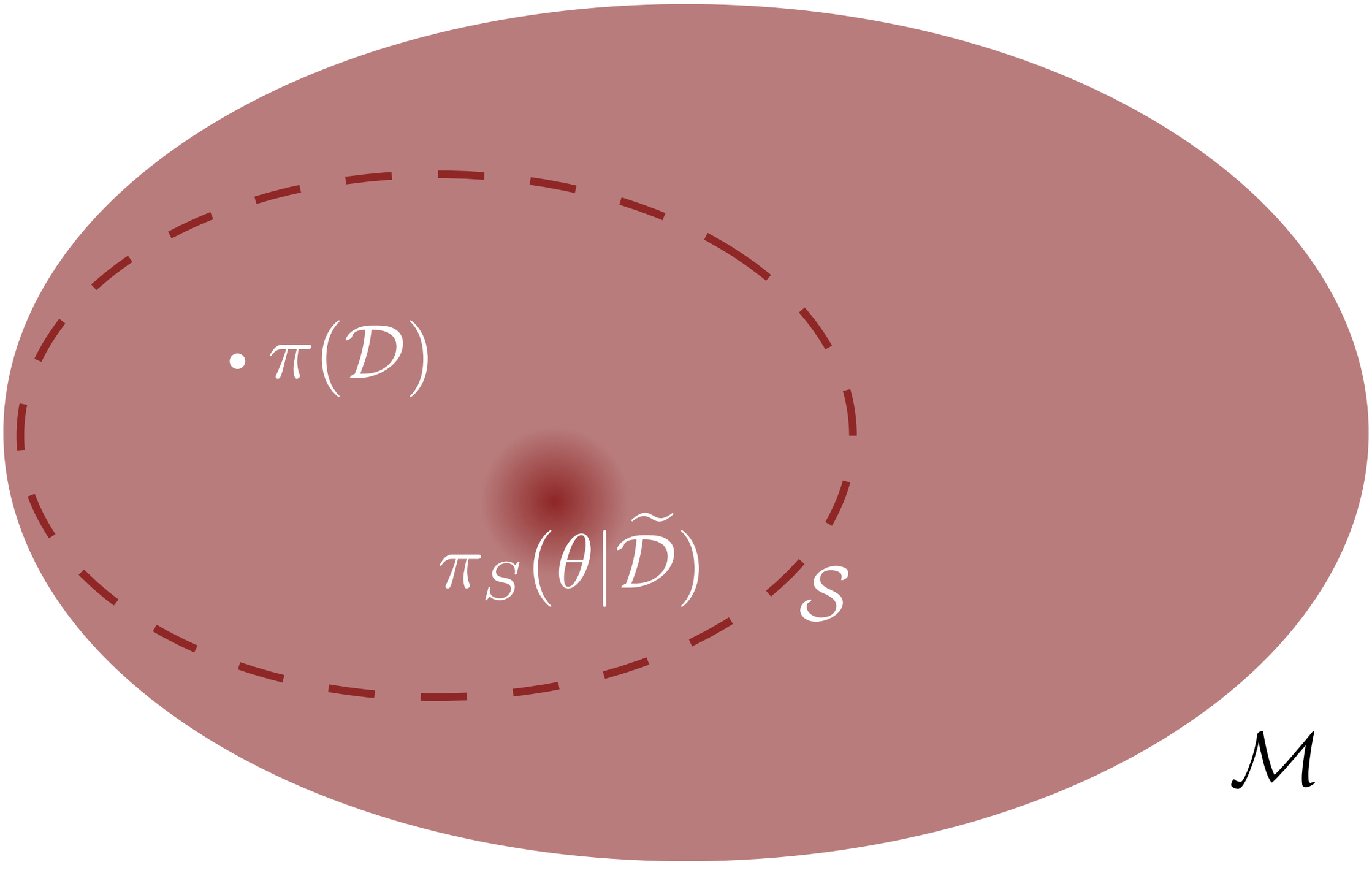




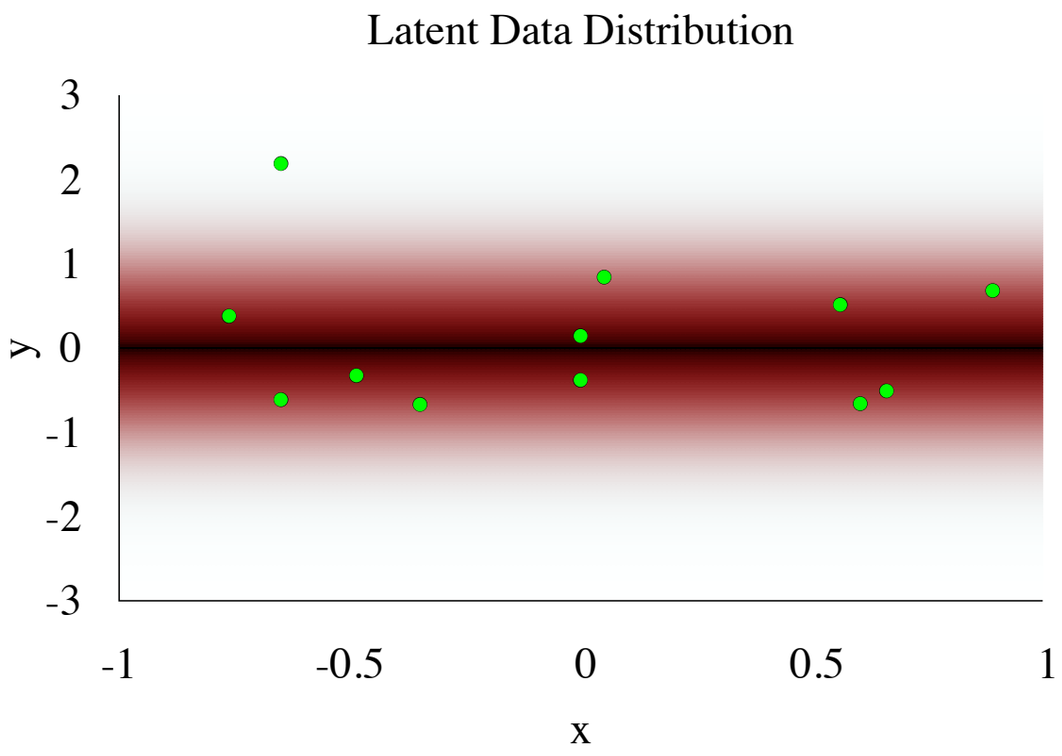
Even if the small world does contain the latent data generating process, however, our models can still overfit.

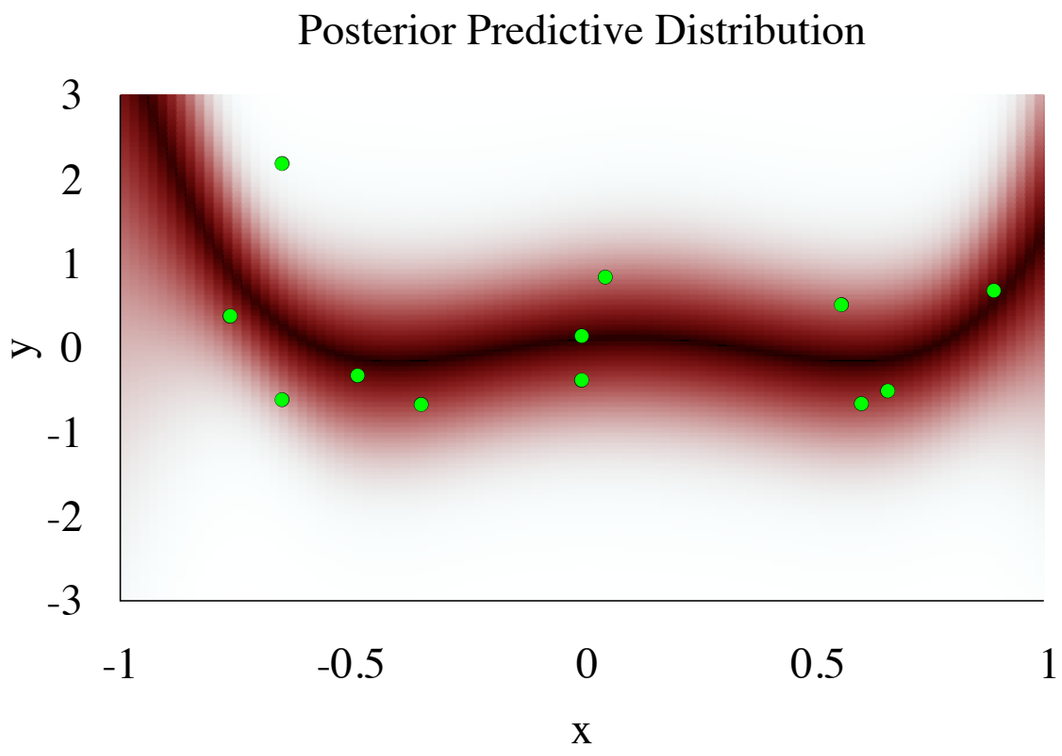


Even if the small world does contain the latent data generating process, however, our models can still overfit.



As with misfit, overfitting manifests as tension between predictive distributions and measurements.





Posterior predictive checks visually compare the predictive distribution to the measurement.

$$\theta \sim \pi_S(\theta|\widetilde{\mathcal{D}})$$

Posterior predictive checks visually compare the predictive distribution to the measurement.

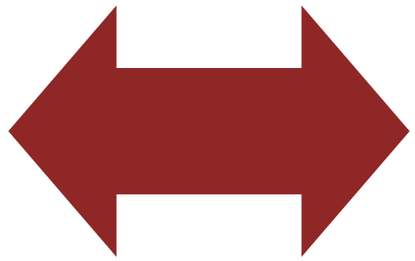
$$\theta \sim \pi_S(\theta|\widetilde{\mathcal{D}})$$

$$\mathcal{D} \sim \pi_S(\mathcal{D}|\theta)$$

Posterior predictive checks visually compare the predictive distribution to the measurement.

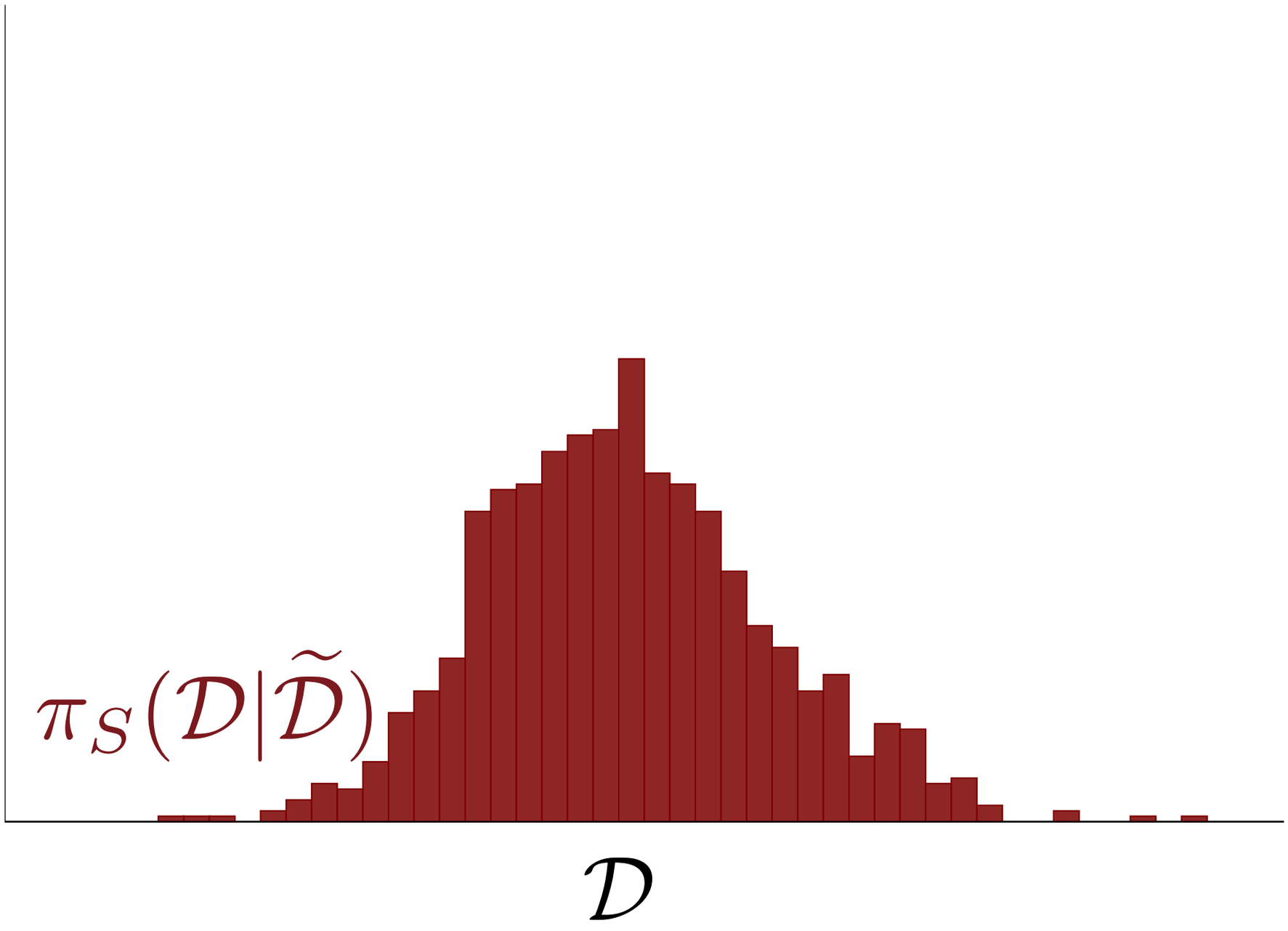
$$\theta \sim \pi_S(\theta|\mathcal{D})$$

$$\mathcal{D} \sim \pi_S(\mathcal{D}|\theta)$$

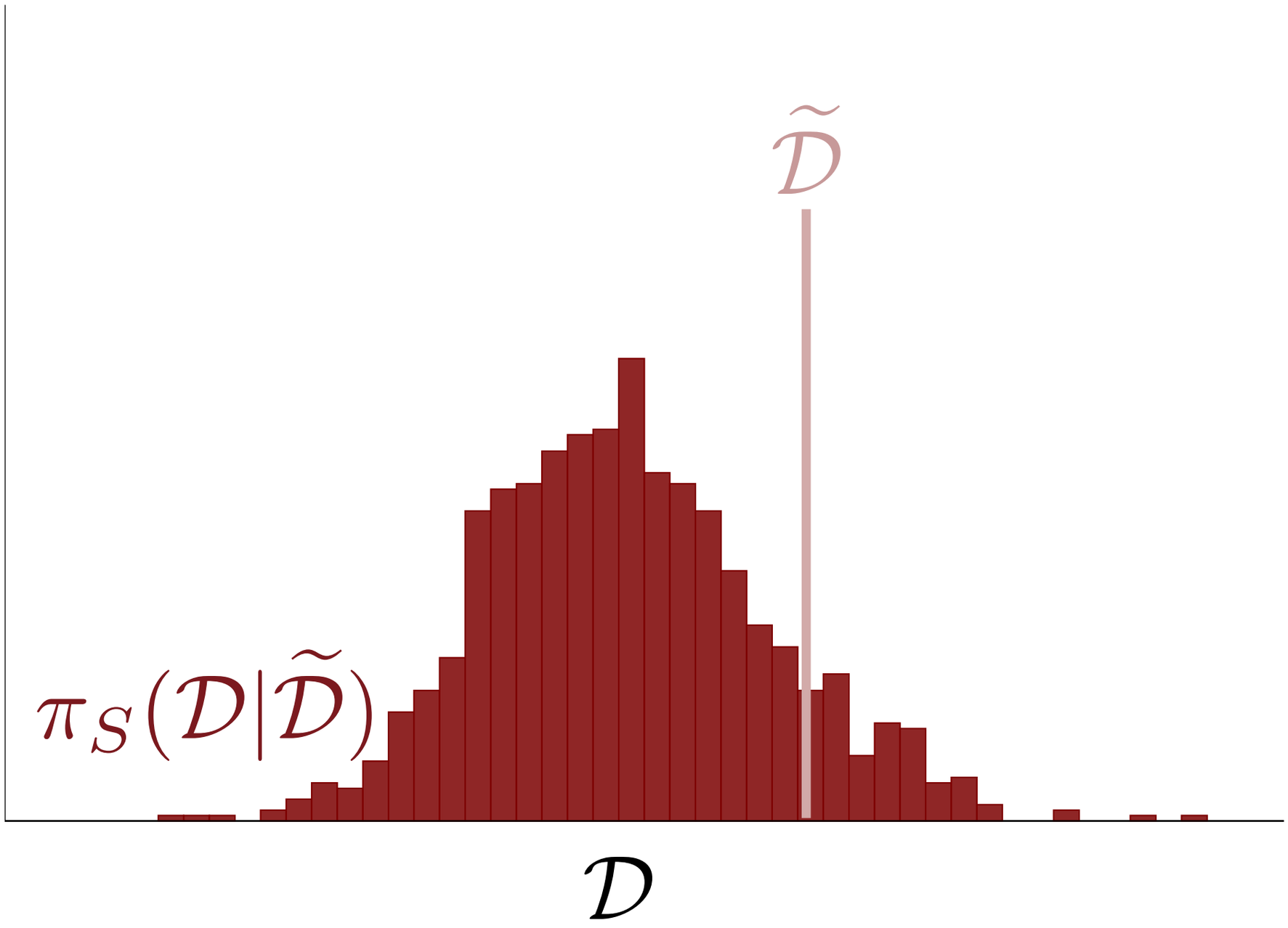


$$\mathcal{D} \sim \pi_S(\mathcal{D}|\widetilde{\mathcal{D}})$$

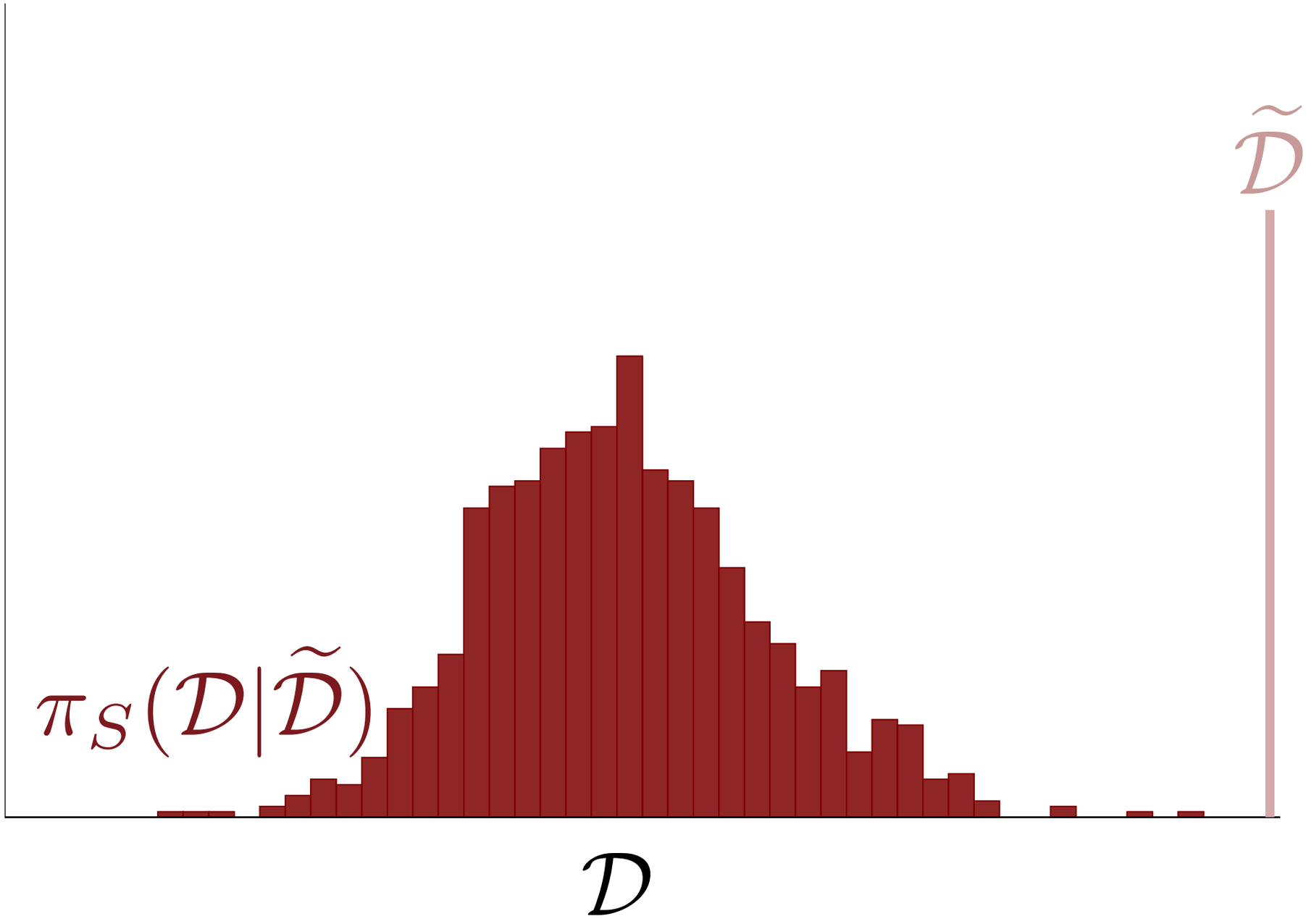
Firstly we can check to see how consistent the measurement is with the inferred predictive distribution.



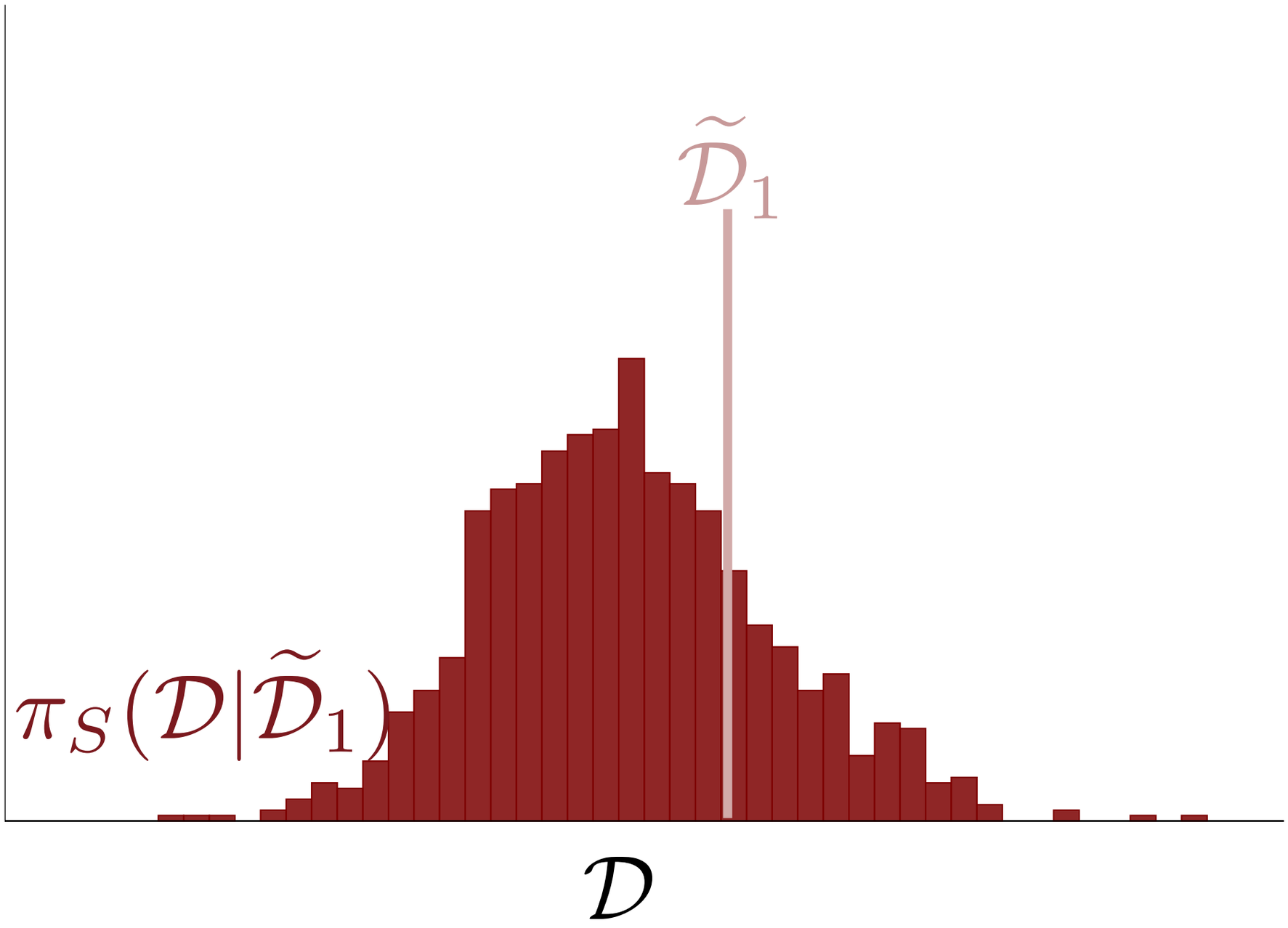
Firstly we can check to see how consistent the measurement is with the inferred predictive distribution.



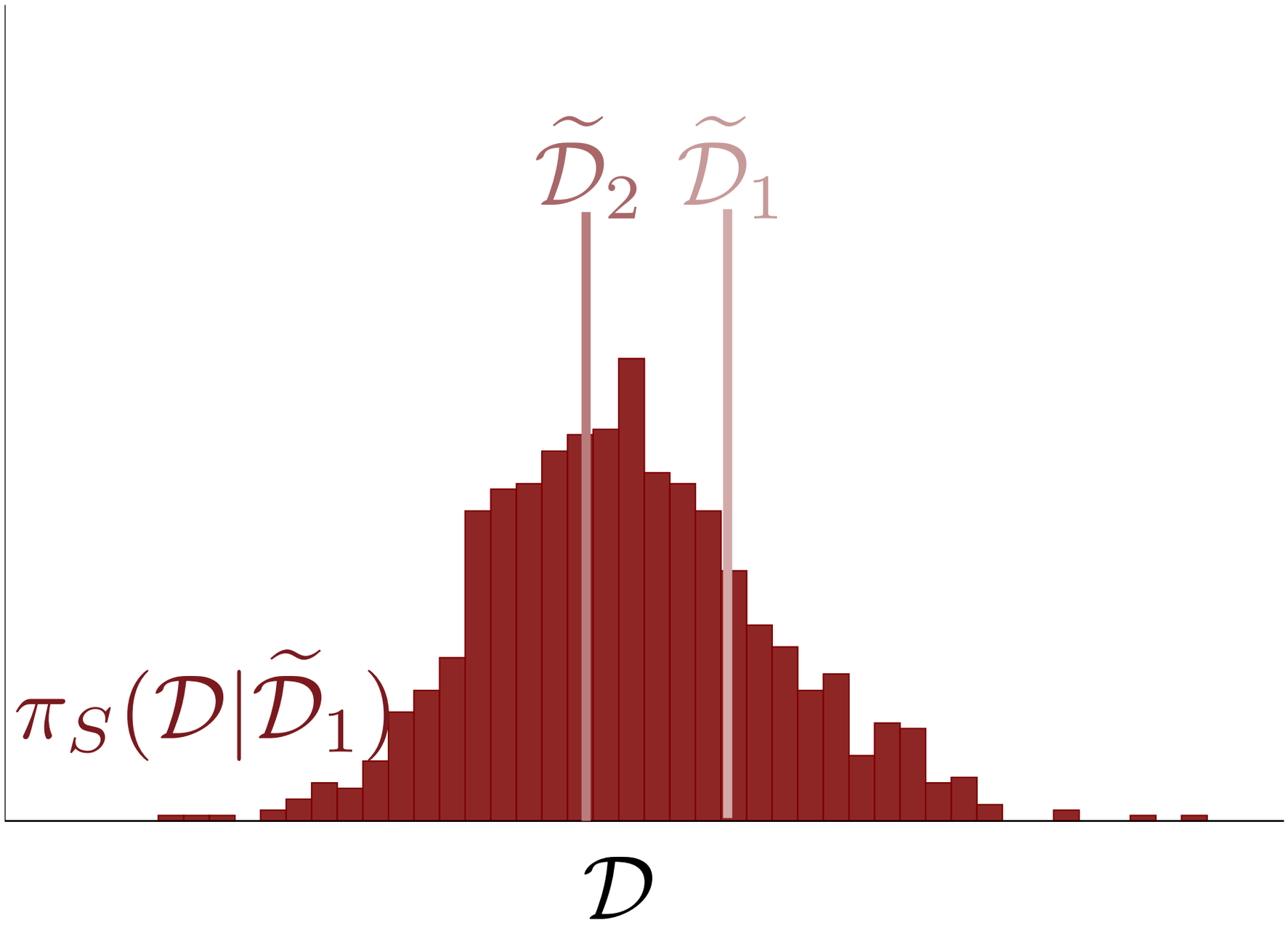
Firstly we can check to see how consistent the measurement is with the inferred predictive distribution.



Similarly, we can check for overfitting by comparing held-out or partitioned measurements.



Similarly, we can check for overfitting by comparing held-out or partitioned measurements.



Similarly, we can check for overfitting by comparing held-out or partitioned measurements.