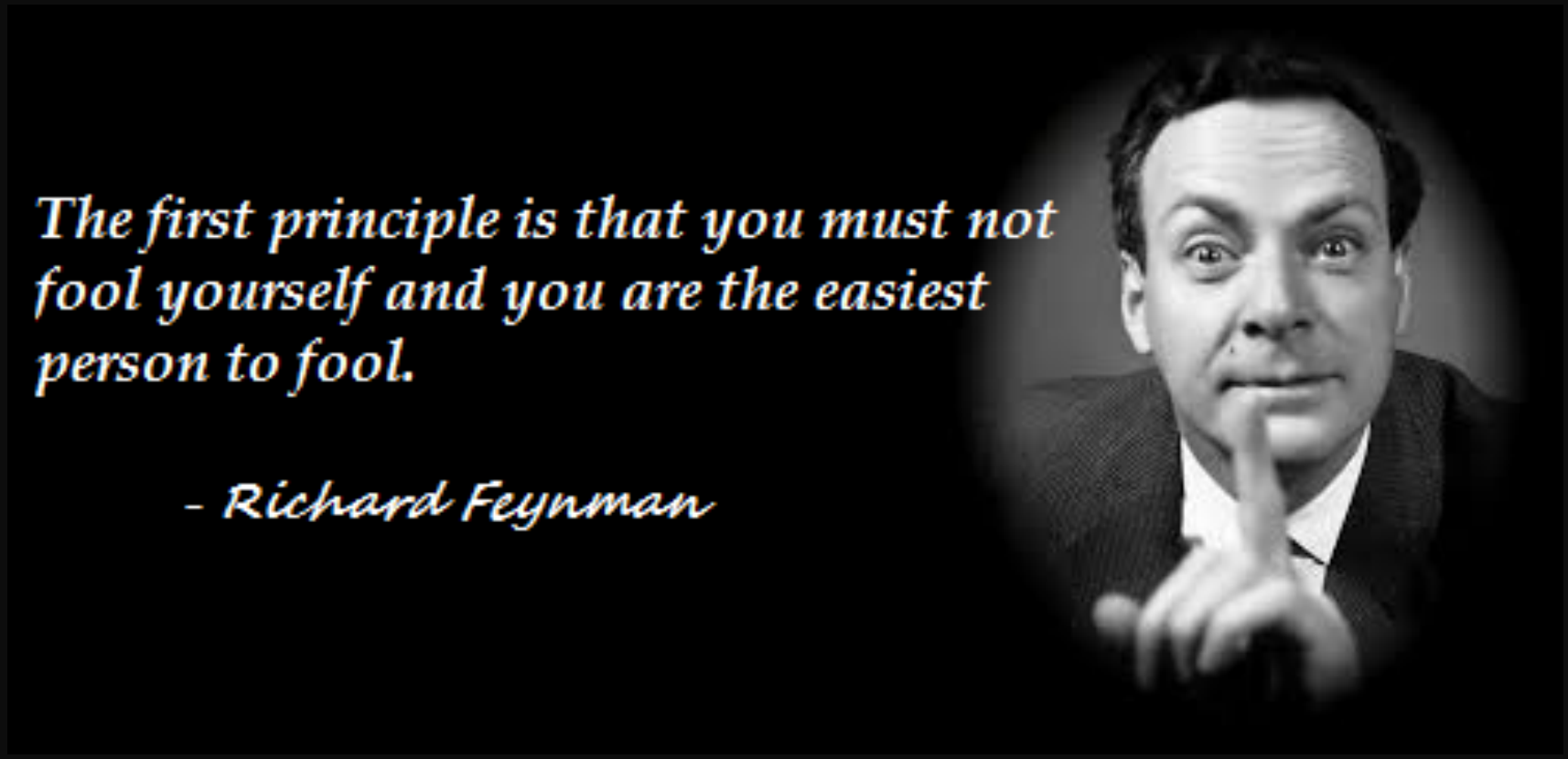
Bayesian Statistics

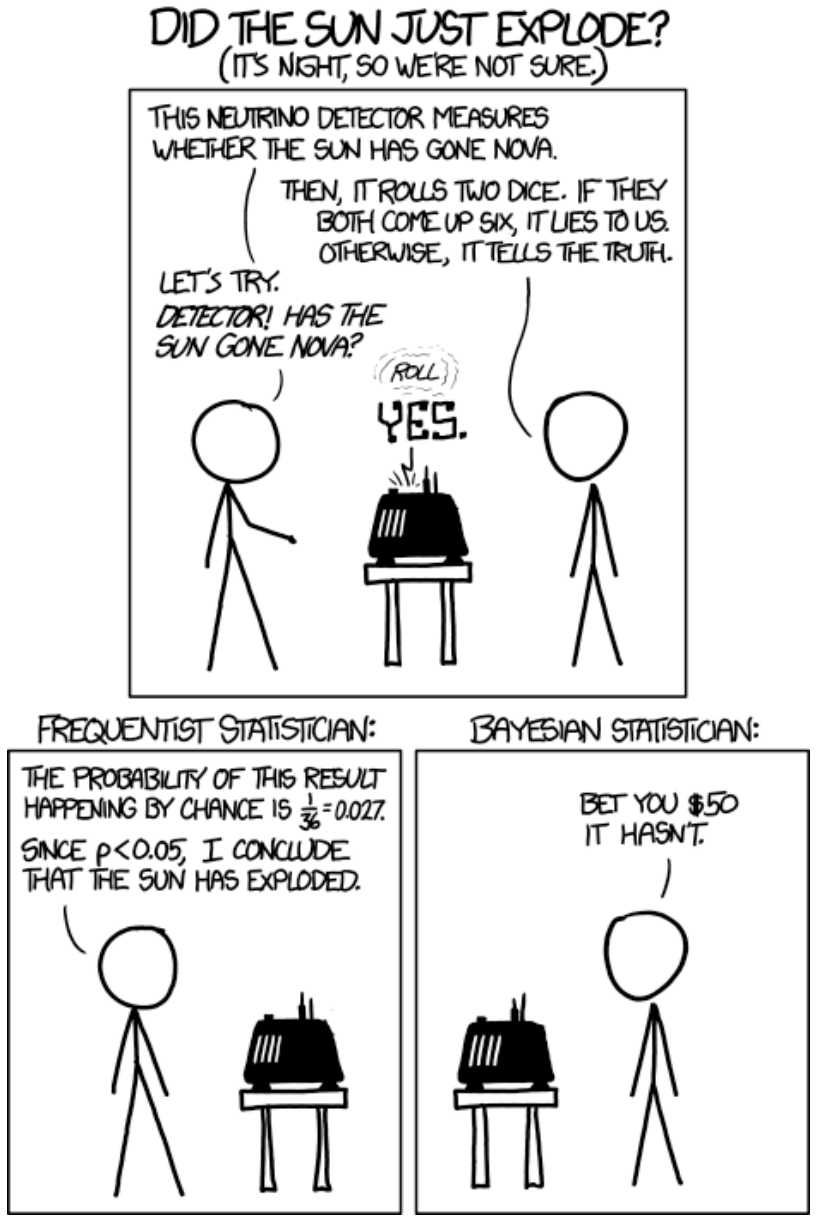


Bayesians and Frequentist don't agree on Planes and Elections

- Frequentist give probability of event
- Bayesians give confidence of an event happening [0,1]
- Are frequentist methods incorrect then?

No.

Frequentist methods are still useful or state-of-the-art in many areas. Tools such as least squares linear regression, LASSO regression, and expectation-maximization algorithms are all powerful and fast. Bayesian methods complement these techniques by solving problems that these approaches cannot, or by illuminating the underlying system with more flexible modeling.



History

- Bayes never published, Richard Price found it among his notes after Bayes' death in 1761, re-edited it, and published it. Unfortunately, virtually no one seems to have read the paper, and Bayes' method lay cold until the arrival of Laplace
- Pierre-Simon Laplace, came to believe that probability theory held the key, and he independently rediscovered Bayes' mechanism and published it in 1774.
- In 1781, Richard Price visited Paris, and word of Bayes' earlier discovery eventually reached Laplace. Laplace eventually refined it to the form we see today.
- Bayes method died again until the invention of machines, most famously being used by Alan Turing to crack the Enigma codes.
- Computational advances in the 90s launched the 'Bayesian revolution' in a long list of fields. This is only partly because Bayes' Theorem shows us the mathematically correct response to new evidence. It is also because Bayes' Theorem works.

Bayes Theorem

Likelihood
Probability of collecting this data when our hypothesis is true

Prior
The probability of the hypothesis being true before collecting data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

Bayes Theorem

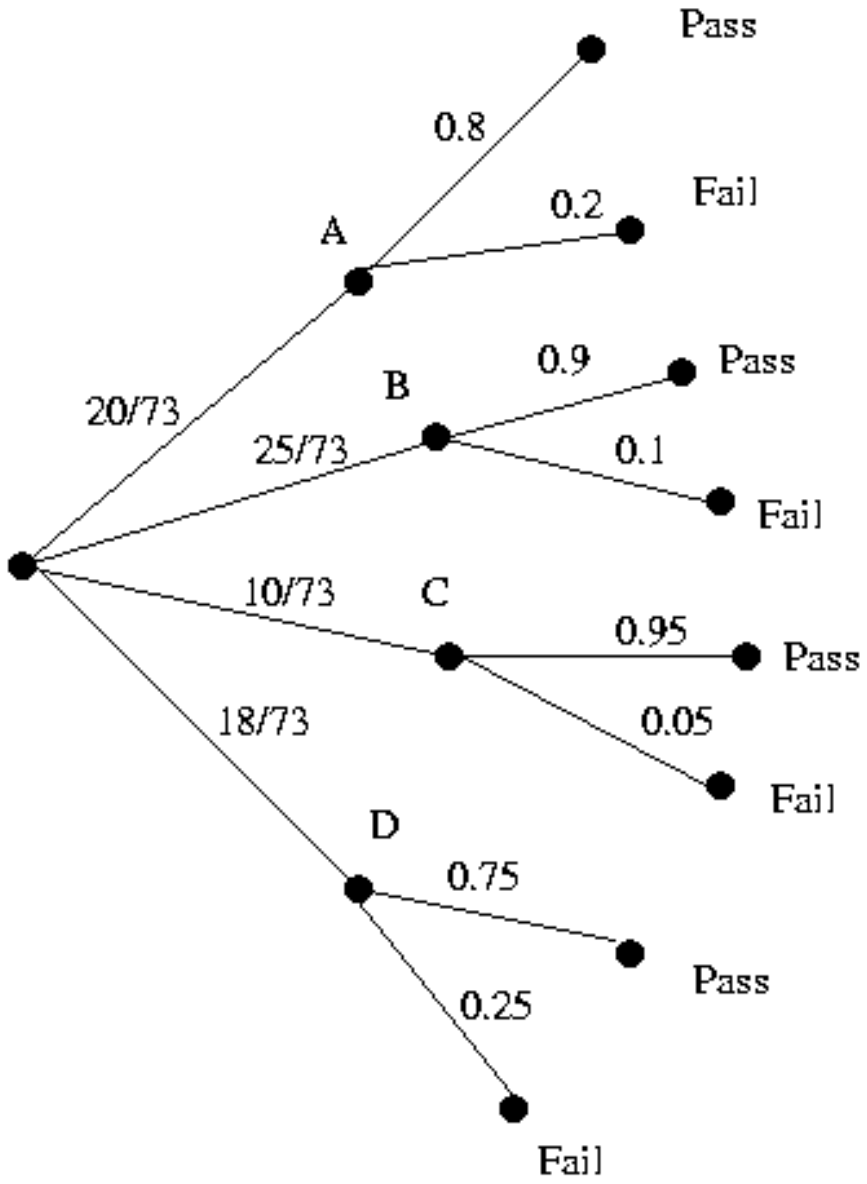
Likelihood
Probability of collecting this data when our hypothesis is true

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Prior
The probability of the hypothesis being true before collecting data

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?



1. Probability you will pass?

2. What is the probability that you will pass the class if you are in section D?

(read carefully here)

3. What is the probability that if you pass, you are in in section D?

Notice that the likelihood stays the same as the Frequentist and the Marginal is essentially a normalization factor.

Therefore the prior is the name of the punch line.

Bayes Theorem

Likelihood
Probability of collecting this data when our hypothesis is true

Prior
The probability of the hypothesis being true before collecting data

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

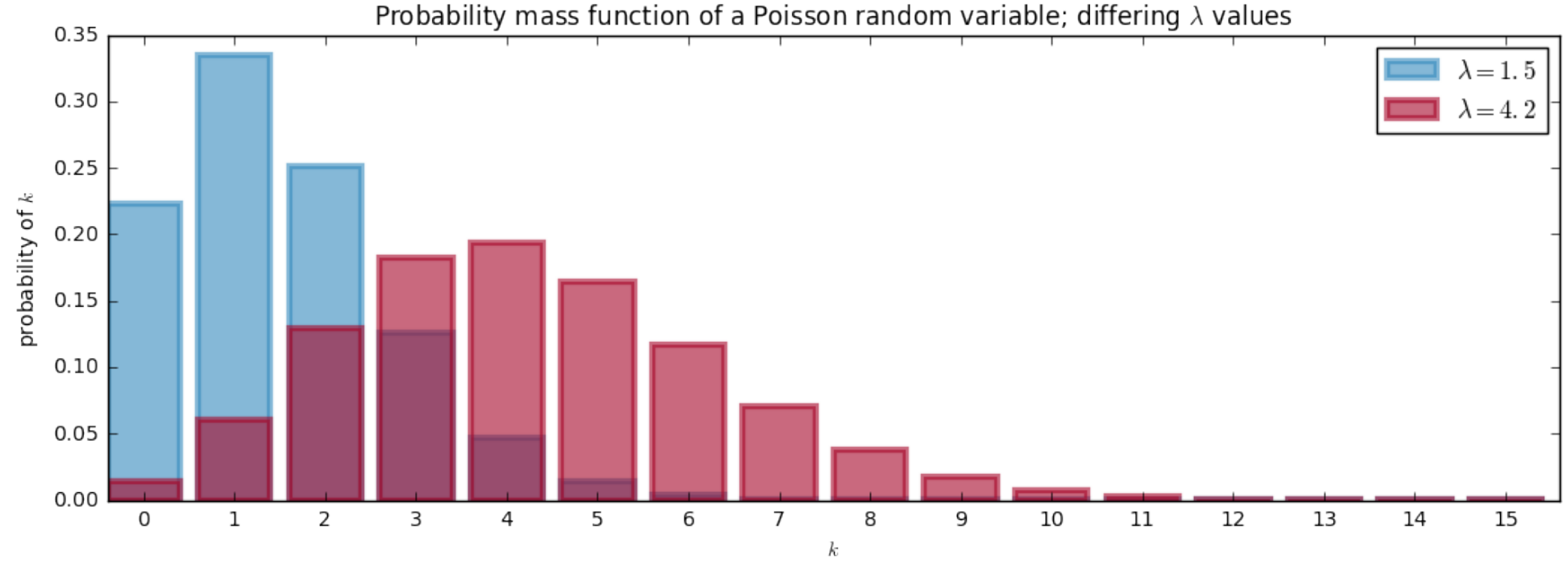
Probability Distributions

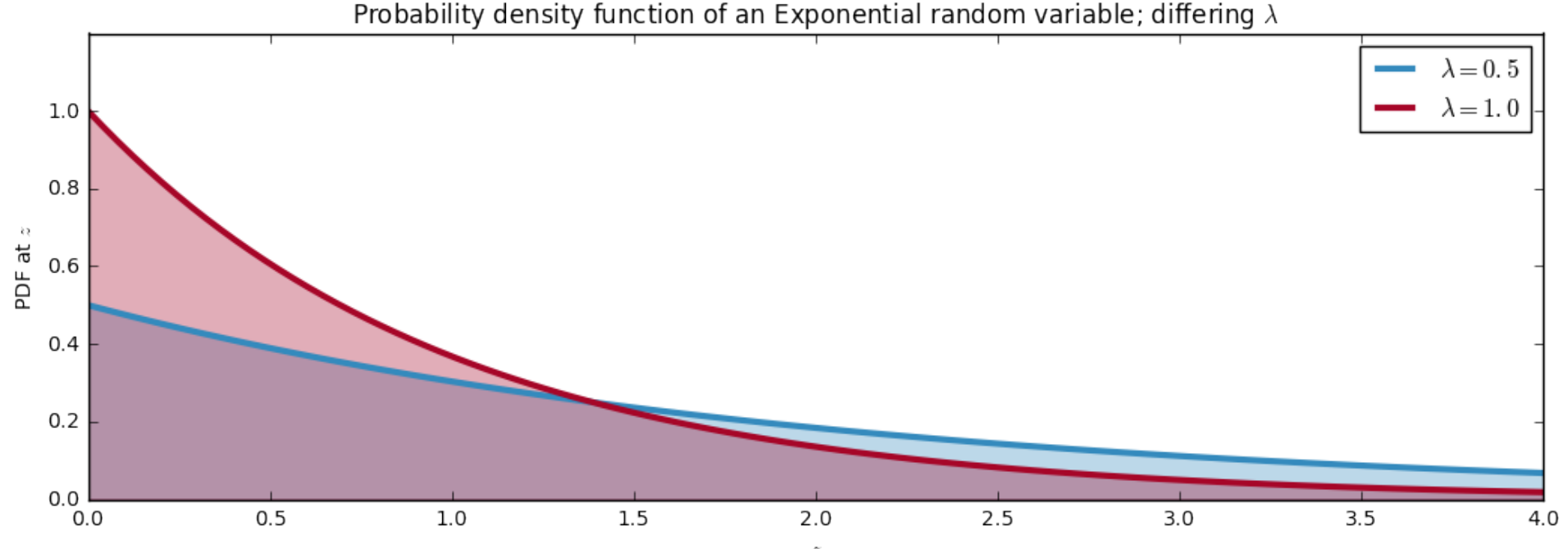
We can divide random variables into three classifications:

- Discrete: Discrete random variables may only assume values on a specified list. Things like populations, movie ratings, and number of votes are all discrete random variables. Discrete random variables become more clear when we contrast them with...

- Continuous: Continuous random variable can take on arbitrarily exact values. For example, temperature, speed, time, color are all modeled as continuous variables because you can progressively make the values more and more precise.

- Mixed: Mixed random variables assign probabilities to both discrete and continuous random variables, i.e. it is a combination of the above two categories.


Probability mass function of a Poisson random variable; differing λ values
probability of k
k
$\lambda = 1.5$
$\lambda = 4.2$


Probability density function of an Exponential random variable; differing λ
PDF at x
$\lambda = 0.5$
$\lambda = 1.0$

Bayes Theorem

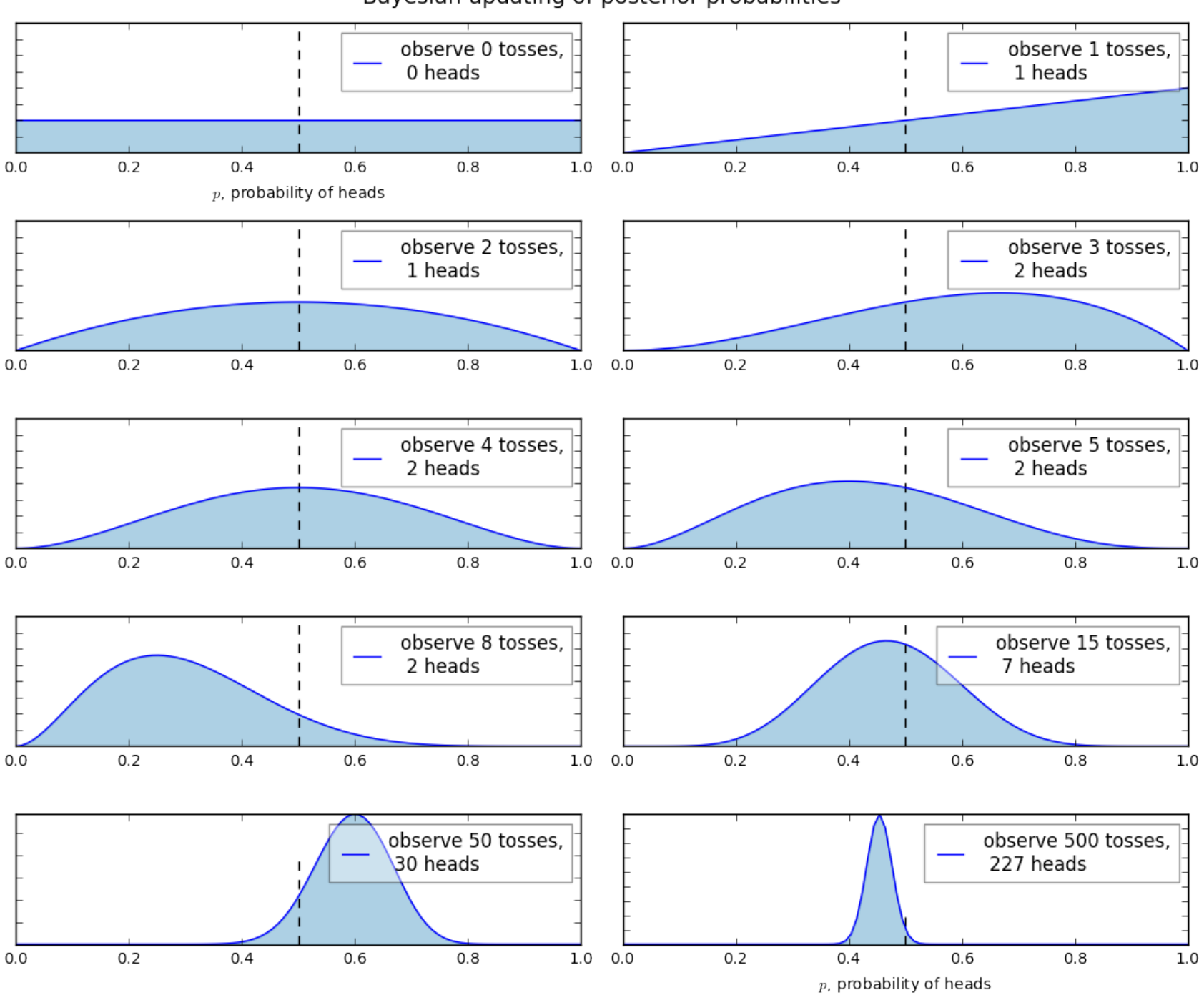
Likelihood
Probability of collecting this data when our hypothesis is true

$$P(H|D) = \frac{P(D|H)\,P(H)}{P(D)}$$

Prior
The probability of the hypothesis being true before collecting data

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

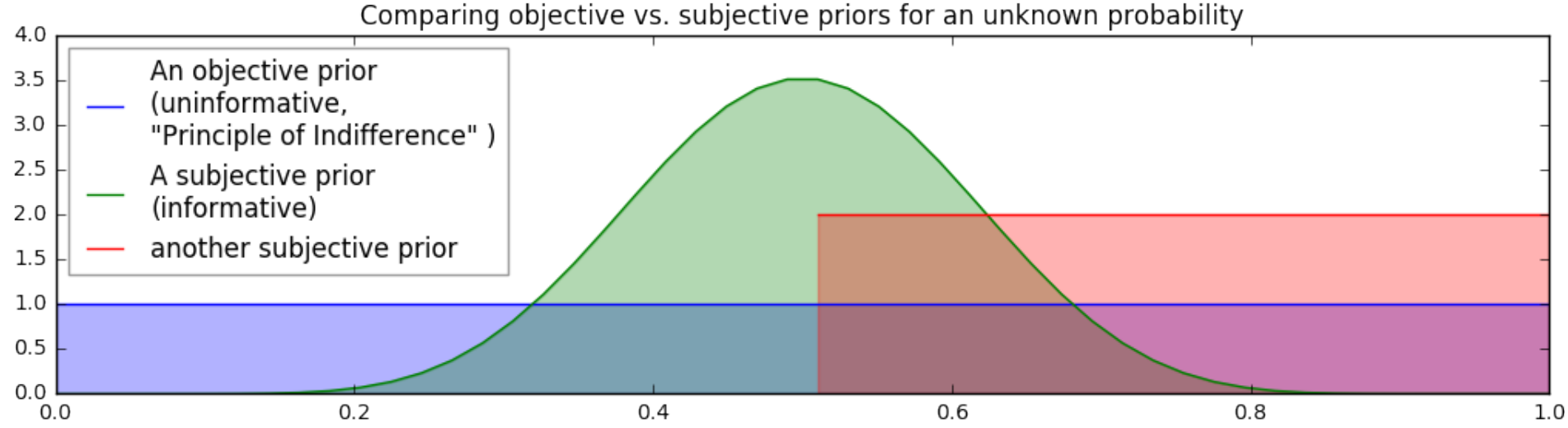

Bayesian updating of posterior probabilities
observe 0 tosses, 0 heads
observe 1 tosses, 1 heads
observe 2 tosses, 1 heads
observe 3 tosses, 2 heads
observe 4 tosses, 2 heads
observe 5 tosses, 2 heads
observe 8 tosses, 2 heads
observe 15 tosses, 7 heads
observe 50 tosses, 30 heads
observe 500 tosses, 227 heads
0.0
0.2
0.4
0.6
0.8
1.0
p, probability of heads

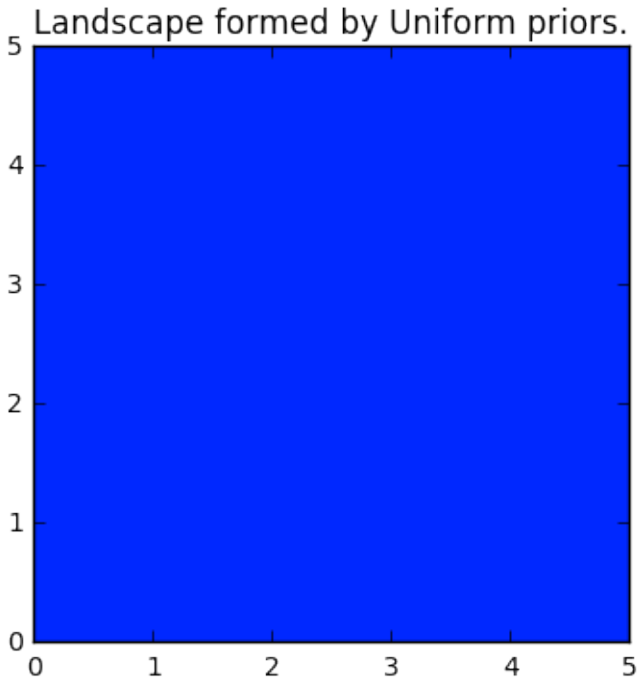
Subjective vs Objective priors

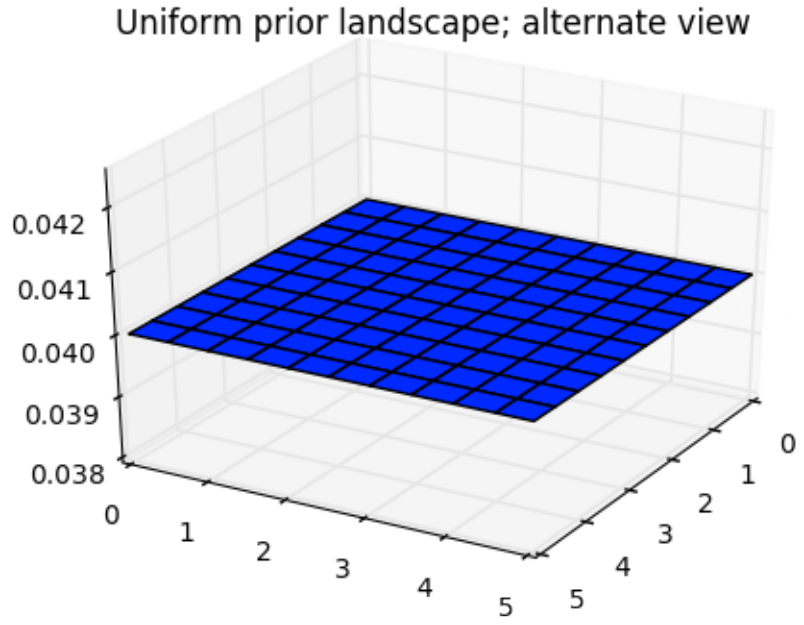
- objective priors, which aim to allow the data to influence the posterior the most, and subjective priors, which allow the practitioner to express his or her views into the prior.

Example is the coin flip, is a flat prior "The Principle of Indifference", which is a uniform distribution over the entire possible range of the unknown. Using a flat prior implies that we give each possible value an equal weighting.

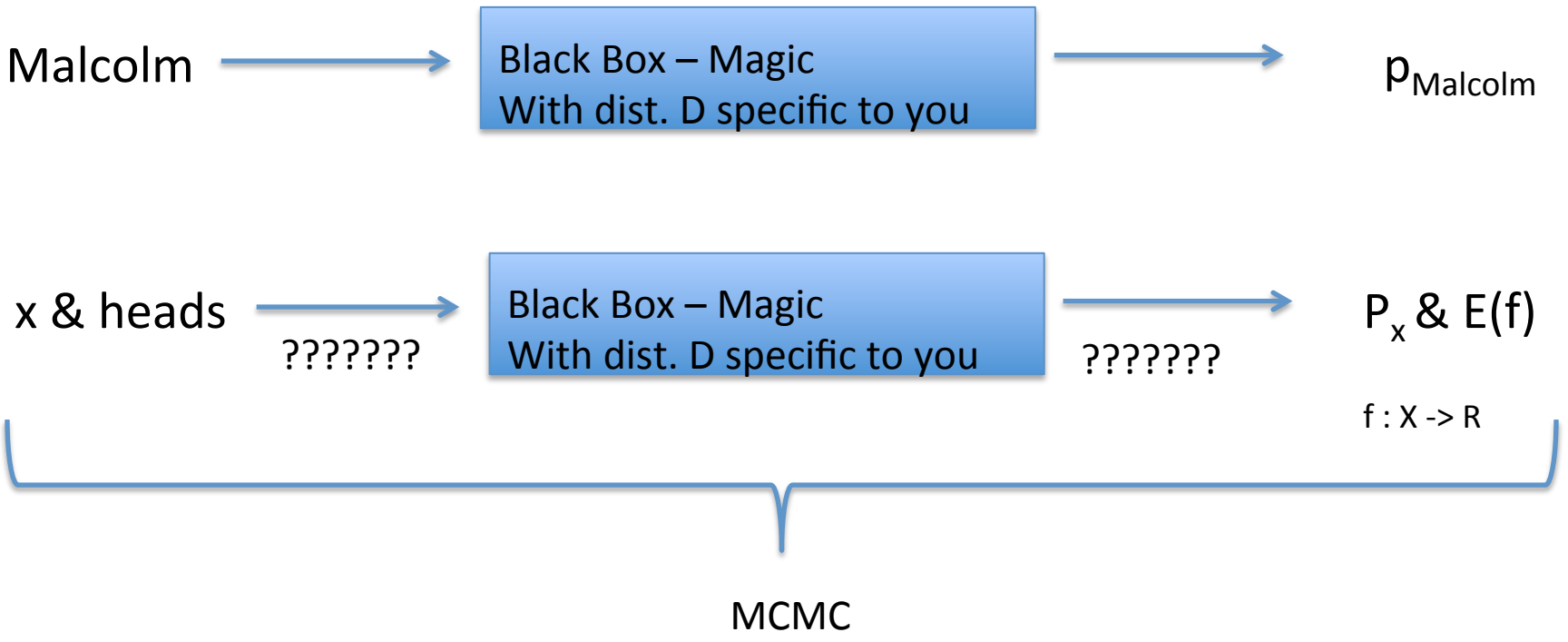
- Subjective Priors On the other hand, if we added more probability mass to certain areas of the prior, and less elsewhere, we are biasing our inference towards the unknowns existing in the former area. This is known as a subjective, or informative prior


Comparing objective vs. subjective priors for an unknown probability
An objective prior
(uninformative,
"Principle of Indifference")
A subjective prior
(informative)
another subjective prior
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


Landscape formed by Uniform priors.
5
4
3
2
1
0
0
1
2
3
4
5


Uniform prior landscape; alternate view
0.042
0.041
0.040
0.039
0.038
0
1
2
3
4
5
0
1
2
3
4
5

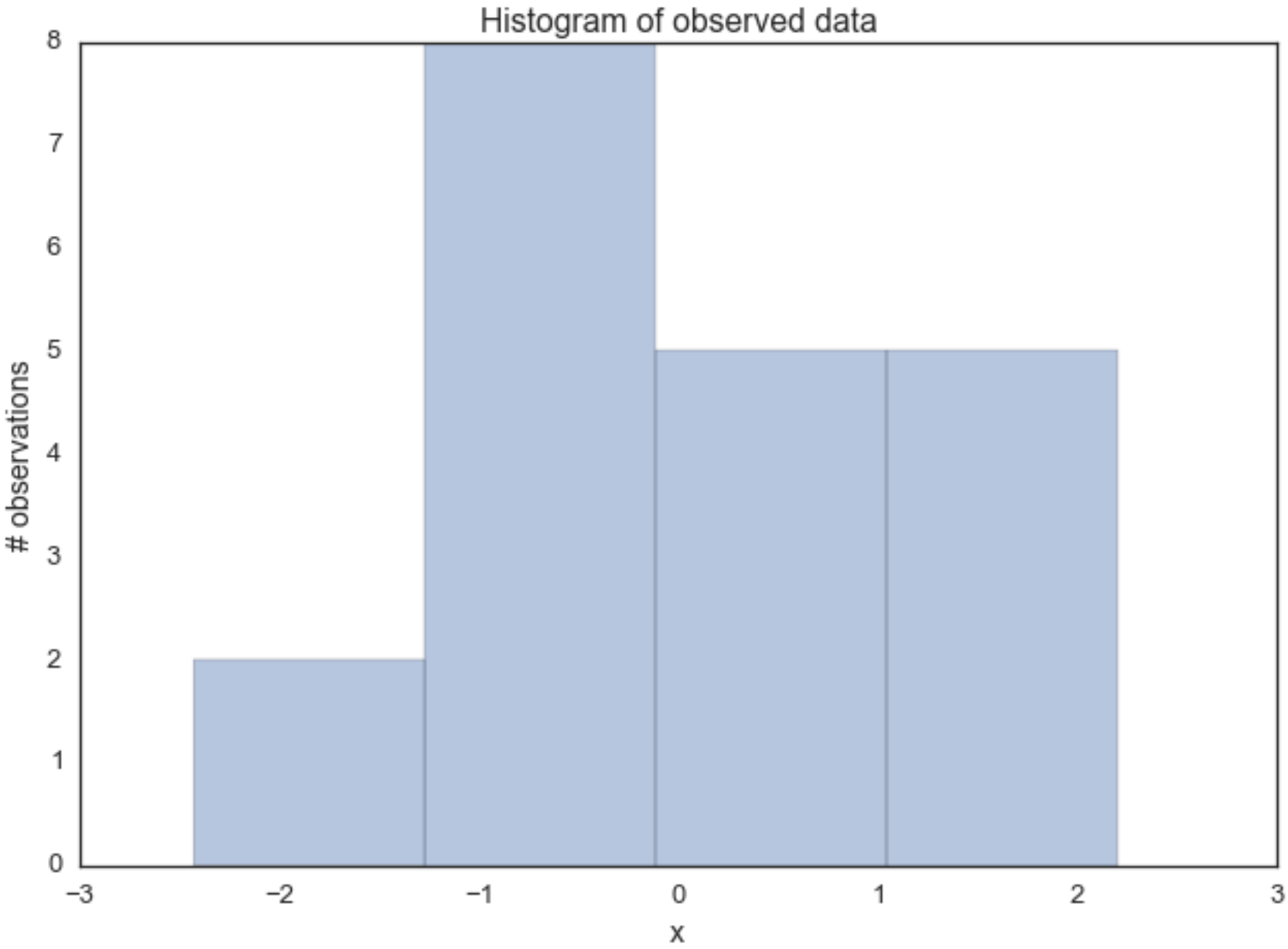
Problem Example

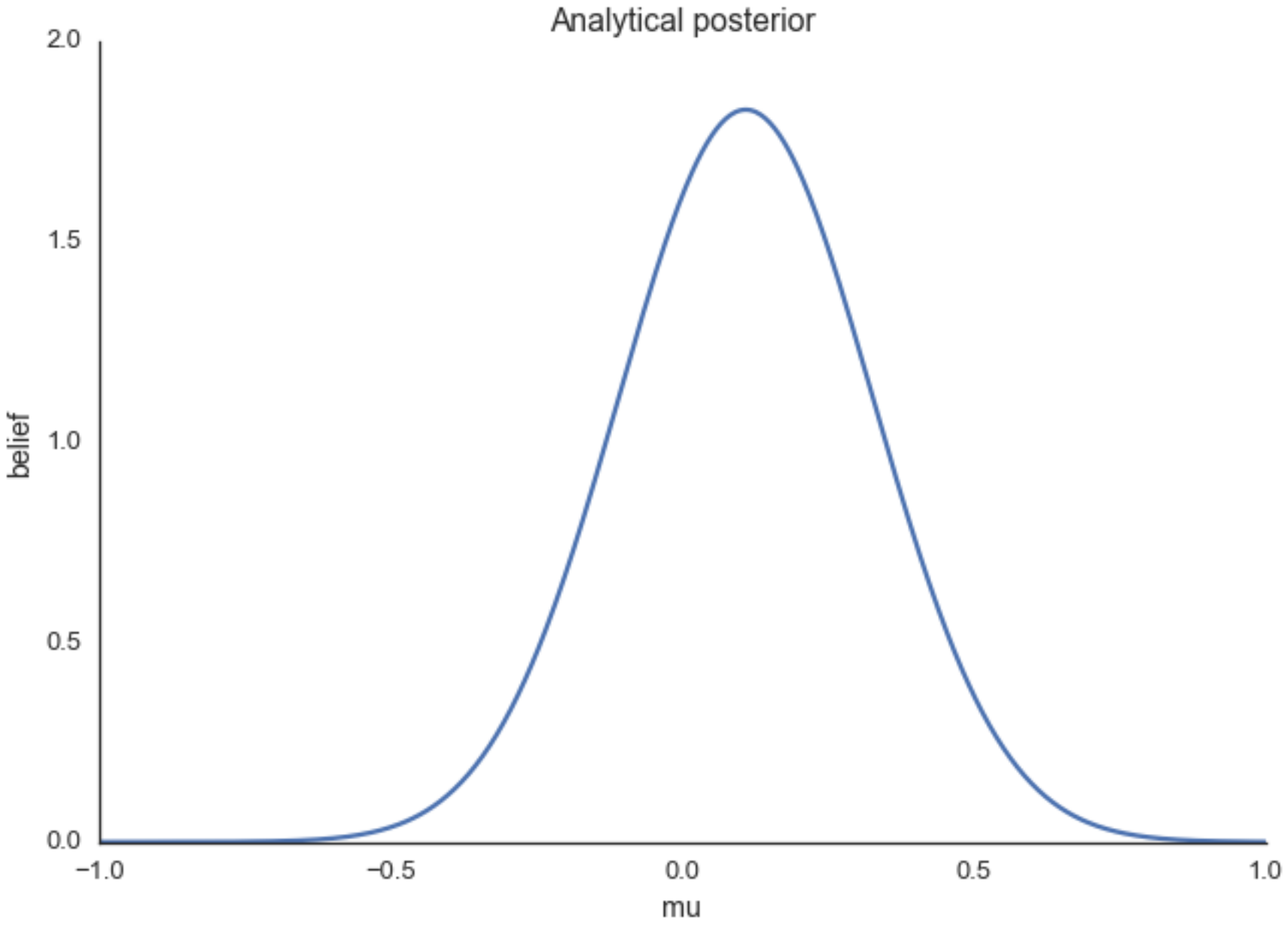


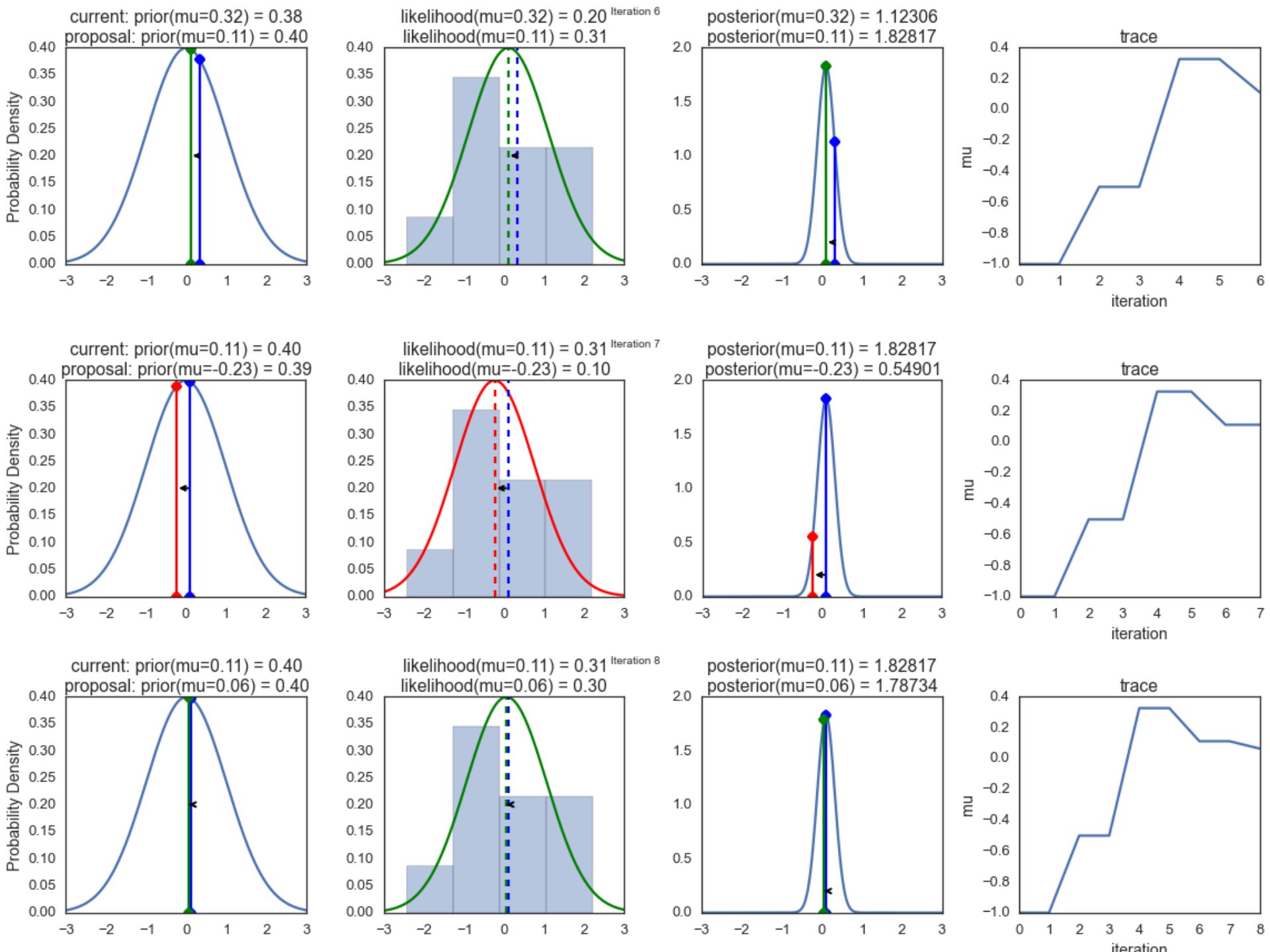
Putting These Methods into Practice

Markov Chain Monte Carlo (MCMC)

- Bayesian inference problem with N unknowns, we are implicitly creating an N dimensional space for the prior distributions to exist in. Associated with the space is an additional dimension, which we can describe as the surface, or curve, that sits on top of the space, that reflects the prior probability of a particular point.

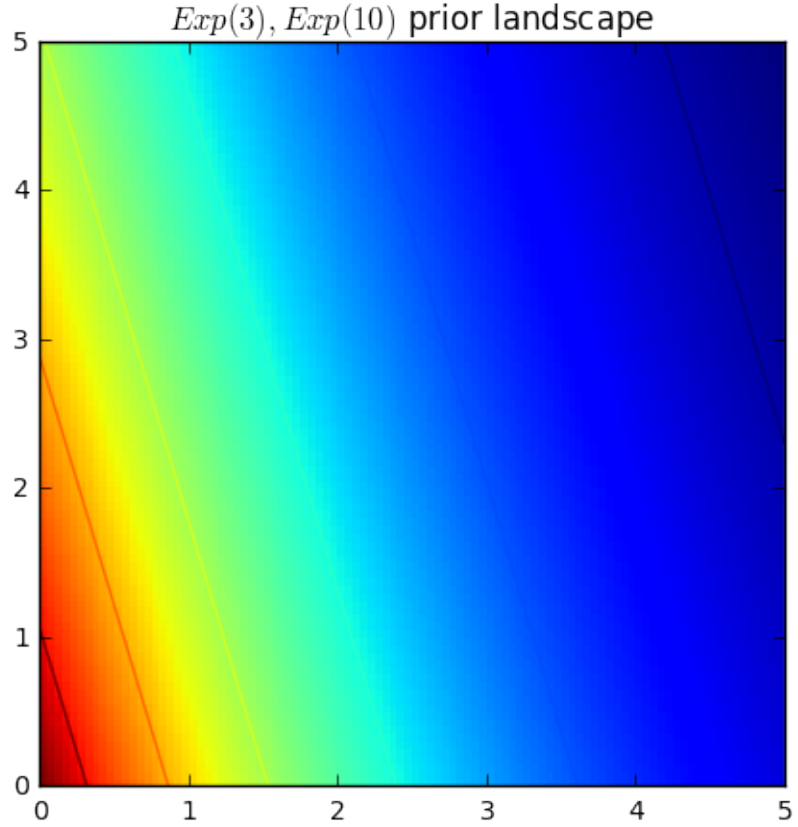

Histogram of observed data
observations
x
0
1
2
3
4
5
6
7
8
−3
−2
−1
0
1
2
3

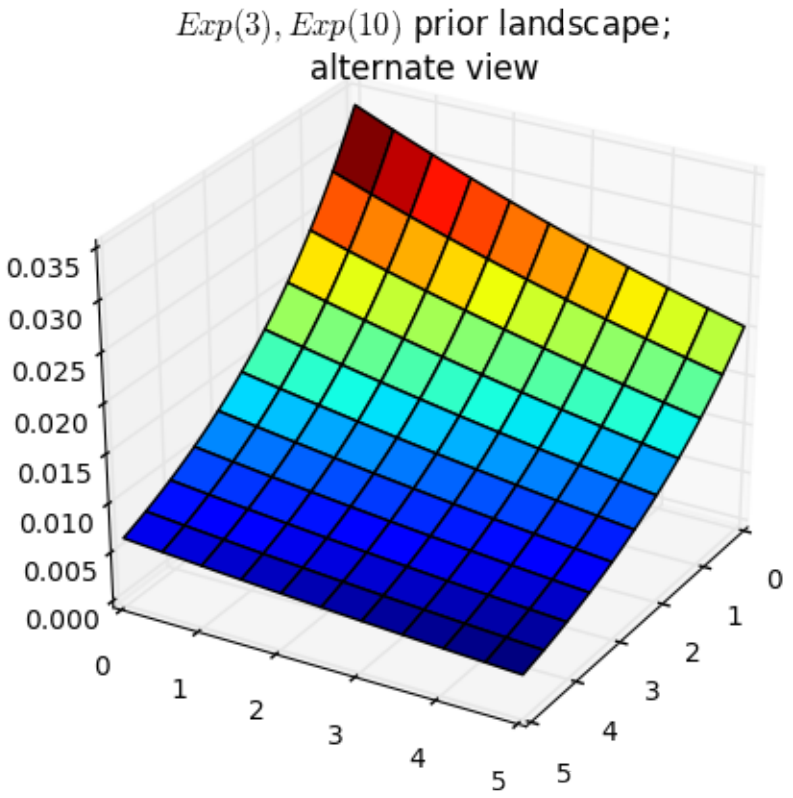

Analytical posterior
2.0
1.5
1.0
0.5
0.0
belief
−1.0
−0.5
0.0
0.5
1.0
mu


current: prior(mu=0.32) = 0.38
proposal: prior(mu=0.11) = 0.40
likelihood(mu=0.32) = 0.20
likelihood(mu=0.11) = 0.31
Iteration 6
posterior(mu=0.32) = 1.12306
posterior(mu=0.11) = 1.82817
trace
Probability Density
mu
iteration
current: prior(mu=0.11) = 0.40
proposal: prior(mu=-0.23) = 0.39
likelihood(mu=0.11) = 0.31
likelihood(mu=-0.23) = 0.10
Iteration 7
posterior(mu=0.11) = 1.82817
posterior(mu=-0.23) = 0.54901
trace
Probability Density
mu
iteration
current: prior(mu=0.11) = 0.40
proposal: prior(mu=0.06) = 0.40
likelihood(mu=0.11) = 0.31
likelihood(mu=0.06) = 0.30
Iteration 8
posterior(mu=0.11) = 1.82817
posterior(mu=0.06) = 1.78734
trace
Probability Density
mu
iteration

The Curse of Dimensionality

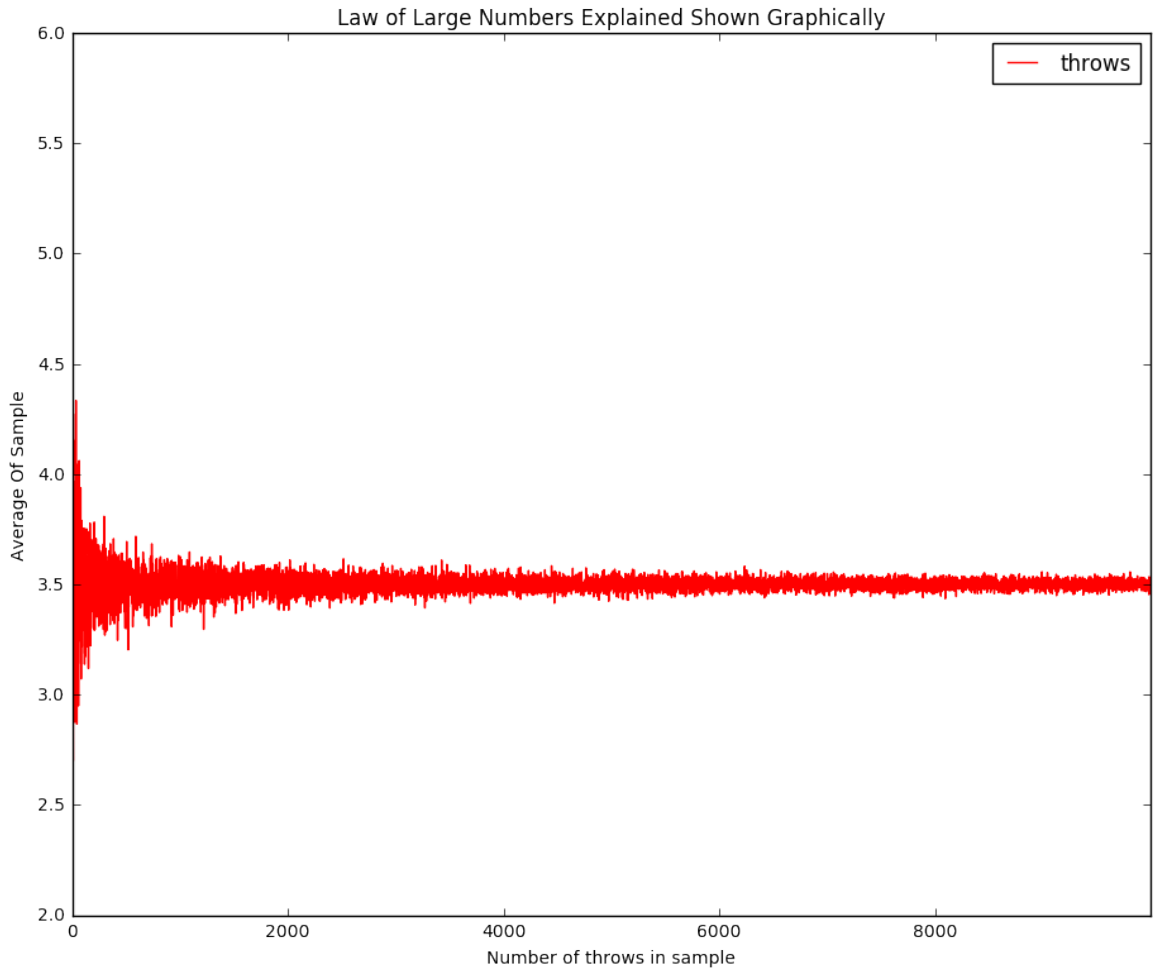
- We cannot naively search the space: any computer scientist will tell you that traversing N-dimensional space is exponentially difficult in N: the size of the space quickly blows-up as we increase N.





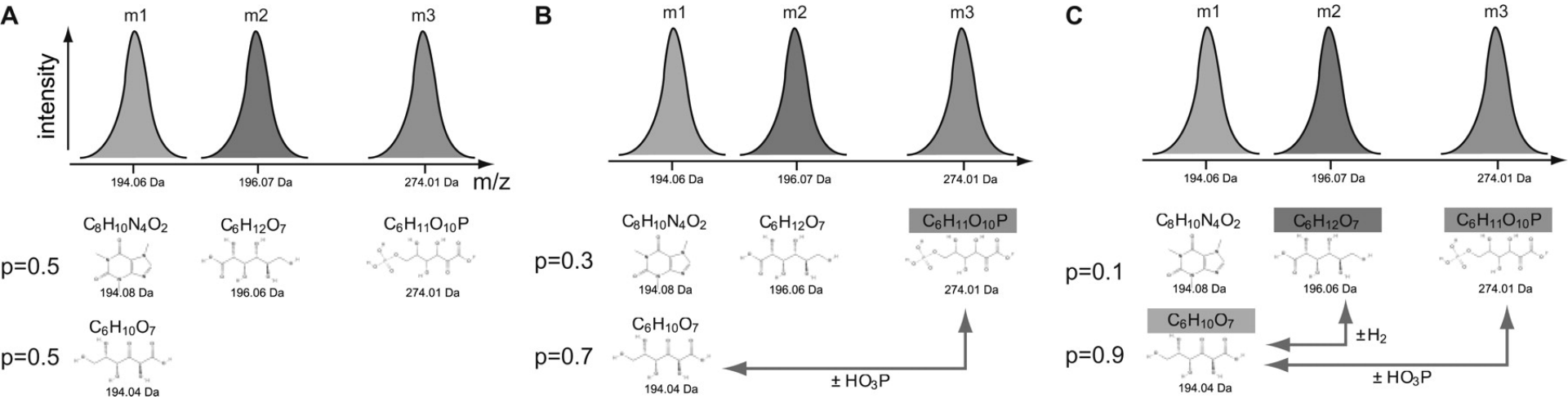
Building Intuition

- Law of Large Numbers



Example Putting It All Together

Probabilistic assignment of formulas to mass peaks in metabolomics experiments



Cartoon depicting the principle of our approach. Three peaks are observed in the mass spectrum. For mass m1 two empirical formulas are initially equally likely (P=0.5), as they both differ by 0.02 Da from the observed mass. Assignment of mass m3 as C6H11O10P provides support for the identification of m1 as C6H10O7 from which it differs by one phosphorylation reaction. Assignment of m2 further increases the confidence in the assignment of m1 to C6H10O7, with the posterior probability increasing to P=0.9.

Building Deeper Intuition, If we have time

Bayesian Point Estimate

- We need to distill our posterior distribution down to a single value (or vector in the multivariate case). If the value is chosen intelligently, we can avoid the flaw of frequentist methodologies that mask the uncertainty and provide a more informative result. The value chosen, if from a Bayesian posterior, is a Bayesian point estimate.

Building Deeper Intuition, If we have time

Loss Function

- Measuring how bad our current estimation is
- So far we have been under the unrealistic assumption that we know the true parameter. Of course if we knew the true parameter, bothering to guess an estimate is pointless.
- As we have a whole distribution of what the unknown parameter could be (the posterior), we should be more interested in computing the expected loss given an estimate. This expected loss is a better estimate of the true loss than comparing the given loss from only a single sample from the posterior. (frequentist method)

Thanks for listening and Questions?