Peter: I'm noticeable.
Corin: (*With a smile, a nice one.*) So you are, Peter. Would like some coffee? A bit long in the pot, I'm afraid.
Peter: Yes, please, I'm an addict.

Corin puts a cup and saucer on the table by Peter's chair and pours from his pot which has been simmering for awhile. He pours himself a cup and sits opposite Peter. They're not far apart.

Corin: So what favor can I do for you?
Peter: It's just a whim that literally hit me yesterday sitting on my couch in The Celt. I'd like to view *The van Heck Nativity* sometime if that could be arranged. My friend Mary Hex was sitting with me and mentioned you were in her stats class and after she ran off to that class, the connection connected in my brain.
Corin: Which class I skipped yesterday. Essay due for my Spanish lit course, don't you know. (*He indicates his Spanish books.*) Also, I'm lost in stats and, probabilistically speaking, will stay that way.

In any case, I think we can have you over to Red House to see the old thing just about any weekend, even this one. But why the special interest?
Peter: I'm interested in painting in a general way, but I read Peter Piper's book on *The Nativity* to learn about the artist, Ruprecht Paulus. In later life, he was primarily an urban planner, one of the designers of the Garden Cities of Brussels. I've been thinking of mathematically modeling the growth of cities and Paulus' name came up. I'd like to go to Brussels to see Paulus' urban development work, but I can't right now, but at least I could see his painting in Oxfordshire for a physical connection to the man.
Corin: Old Paulus, the master forger.
Peter: *The van Heck Nativity* is not a forgery, we both know it. It's a demonstration of the forger's art.
Corin: In fact, I've known "it" all my life. I once touched it all over just to be bad, but just gently.
Peter: I wouldn't do that. Just give it a good viewing, take some images if that is permitted.
Corin: That's fine absolutely. (*Then with some slyness.*) But I have favor to ask of you if it is convenient.
Peter: At your service, Corin.
Corin: Could you explain Bayesian analysis to me right now?
Peter: How much do you know?
Corin: Let's pretend I know nothing.
Peter now becomes Professor Courtney.
Peter: You understand it first as poetry. (*Slight pause.*) Do you have a pad of writing paper? (*Corin nods and picks it up to show.*) Why not pull up your chair beside me and let's give it a shot.

Corin pulls his chair to the right of Peter's. The writing pad bridges the arms of the chairs. Peter's green-ink pen writes "Universe." We follow on the pad the diagrams and equations as Peter explains them. The writing just appears more

quickly than the hand can write like on the YouTube videos. Occasionally, there are cuts to Peter and Corin as needed.

Peter: This is a Venn diagram. There are populations of events A and events B represented by partially overlapping ovals. Surrounding those ovals is a large oval containing events K, where K is a more general event. For key example, K may be all knowledge that we have. And surrounding the K oval is the big oval representing the universe of all events. The overlap region of A, B and K has a joint probablitiy of occurring $P(ABK)$, where here ABK is not a product, but means "A, B, and K in hand at the same time" or, precisely mathematically, the intersection of A, B, and K. We can expand $P(ABK)$ using the frequentist conception of probability: i.e., so many events out of so many trials. Behold:

$$P(ABK) = \frac{N_{ABK}}{N} = \frac{N_{ABK}}{N_{BK}} \frac{N_{BK}}{N_K} \frac{N_K}{N} = P(A|BK)P(B|K)P(K) \ ,$$

where the N's represent populations of events/trials taken to the limit of infinite size, the multiplied fractions follow from a licit why not, and $P(A|BK)$ etc. are conditional probabilities: i.e., probability of A given BK etc. The equation we have just derived is called the product rule in probability theory for obvious reasons. Note that probabilities determined empirically from actual populations of trials and events must always be less than exact since you cannot go to populations of infinite size. Nearly exact probabilities can be obtained from sufficiently large populations. Can exact probabilities ever be obtained? They can be obtained by assumption for heuristic reasons: i.e., for education, example, or advancing discovery. They can also be obtained from powerful theories you believe to be true. The example par excellence of the latter case is quantum mechanics—which we will not digress on.

There are experts who say that Bayes' theorem (which we will get to soon) and Bayesian analysis are more general than the frequentist conception in that probabilities can represent degrees of belief, not just frequencies. But that opinion invites the question, what does one mean by degrees of belief? I think it must mean "so many events out of so many trials" even if the events and trials are specified in a very vague way and any measured populations are less than infinite, often far less. In fact, in Bayesian analysis, they are often specified in a vague way and that is not a problem at all as we will discuss.

Now symmetry gives a second version of the product rule starting from the first one:

$$P(ABK) = P(B|AK)P(A|K)P(K) \ .$$

Equating the two versions and canceling the often only vaguely known $P(K)$ gives Bayes' theorem in symmetrical form

$$P(A|BK)P(B|K) = P(B|AK)P(A|K) \ .$$

More often, such as at Wikipedia, one sees the asymmetric forms of Bayes' theorem:

$$P(A|BK) = \frac{P(B|AK)P(A|K)}{P(B|K)} \qquad \text{and} \qquad P(B|AK) = \frac{P(A|BK)P(B|K)}{P(A|K)} \ .$$

The K is usually not seen in the derivation and the theorem. It's just taken as understood. But, in fact, we need the K to understand Bayesian analysis. So I made the K explicit.

Now Bayes' theorem is elementary both in being at the root of reality and directly intelligible to all sufficiently sophisticated intelligences, I think, without exception. It's a relationship among conditional probabilities. I like to think of Bayes' theorem as a Platonic ideal, one of many actually.

Now to be more specific, Bayes' theorem is general to all probability theory. Bayesian analysis is a special branch of probability theory rooted in Bayes' theorem, but having a lot more formalism of which we'll show only a little bit. By the by, the term Bayesian analysis is I think to be preferred to the term Bayesian inference which has a more restricted meaning which I will not explicate in the here and now.

But what is Bayesian analysis? First off, like evolution by natural selection, Bayesian analysis is a form of learning, but unlike evolution, Bayesian analysis requires intelligence in some sense. We humans and all biota to some degree have done what can be called for the nonce qualitative Bayesian analysis with fair-to-middling success since forever. In fact, in scientific matters, qualitative Bayesian analysis is the scientific method in action.

To explicate qualitative Bayesian analysis, say some trial is coming up for you with various possible outcomes which we often call events in the probability jargon as, in fact, we already have. All your past experience allows you to qualitatively estimate the probabilities of those events: certain, nearly certain, very likely, likely, 50-50, unlikely, very unlikely, impossible. Those probabilities are your prior probabilities or priors in Bayesian analysis jargon. After the trial, you have new knowledge and using that new knowledge you update your probabilities: the updated probabilities are your posterior probabilities or posteriors. The essence of the qualitative Bayesian analysis procedure is the estimating of priors and the updating to posteriors.

How do you calculate your probabilities for qualitative Bayesian analysis? Ultimately, I think, as I've already indicated by "so many events out of so many trials," but your calculation is usually very approximate math without any explicit numbers and usually there is no need to do better since the number of events and trials in your past is far short of what is needed to establish nearly exact probabilities which require, as aforesaid, large populations of events and trials and your events and trials are only approximately alike in any case which makes your whole calculation very vague. But you say now you have priors even for events that occur in trials that you have never had before. But there are all kinds of events and trials in life you have experienced including those you experience just vicariously in reading or in oral stories or in imagination. Out of all those events and trials, you are able to synthesize priors and, in fact, you almost always do and whether you want to or not. Sometimes synthesized priors lead to bad outcomes, but often they are useful.

Now what is Bayesian analysis without qualification? It quantifies the procedure of qualitative Bayesian analysis making use of Bayes' theorem and allows

you to find truth at least in the limit of ideal Bayesian analysis—with the exception of intractable cases which we will discuss eventually. We will, in fact, prove that ideal Bayesian analysis allows you to find truth (except for aforementioned intractable cases) which is the same as saying we will prove ideal Bayesian analysis which is the same as saying we will prove Bayesian analysis without qualification is generally useful.

However, we must note that Bayesian analysis is actually computationally very intensive in practical cases because such cases usually involve large data sets upon which elaborate calculations have to be done. Because of the computational requirements, only in recent decades with the advent of vast computer power has Bayesian analysis become generally useful. And it is very useful in many modern sciences: e.g., epidemiology, economics, ecology, particle physics, and cosmology all of which nowadays must analyze vast data sets. I hope to help make Bayesian analysis more useful in cliodynamics, the mathematical modeling of history, which is my own specialty of which I don't know very much yet.

But for what kind of learning is Bayesian analysis used? Well, many kinds in general. But most importantly, it is used to find the probability of the truth of hypotheses or theories. Note the terms hypothesis and theory can be used as synonyms, but, following usual understanding, a hypothesis is of limited application in understanding some aspect of reality and a theory is of much more general application in understanding some aspect of reality. Hereafter, I will usually just use the word theory meaning both hypothesis and theory for brevity.

But now you say isn't a theory just true or false setting aside the complication of partially true theories for simplicity in discussion. In an absolute sense, one theory has probability of truth 1 and all the others have probability of truth 0. But the absolute sense is not the to-your-knowledge sense. To give a simple example of the difference between the absolute probability of truth (which is 1 or 0) and the probability of being true to your knowledge, say I have coin under my left hand which in fact I do.

Peter's left hand is now seen flat on on the arm of his chair away from Corin.

The coin can be only one of heads or tails. One theory about it has absolute probability 1 and the other has absolute probability 0. But those probabilities are known only in the mind of God. The probabilities to your knowledge, Corin, are what?

Corin: 50-50.

Peter: Update your probabilities.

Peter takes away his hand.

Corin: Heads, 100 %.

Peter: (*Peter catches Corin's eye for a moment.*) Tails, look again.

Corin: (*Looking again.*) Devil.

Peter: Right the first time, heads.

Corin: (*Looking one more time.*) Double devil. What's the trick, Peter?

Peter: (*Pocketing the coin.*) Just power-of-suggestion trickery. The trickery hints at how hard Bayesian analysis can actually be—which we know from qualitative

Bayesian analysis which, as aforesaid, we've done forever. But forgetting the trickery, your knowledge was correctly updated from 50-50 to 100 % for heads.

You'll note that my example was really about a hypothesis, not a theory. You computed 50-50 for the two posssible hypotheses using the theory of probability of coin tosses which is, of course, the familiar binomial probability distribution. See, there's the Wikipedia article on the binomial probability distribution on my phone.

However, similar examples can be given for theories. To take one such example, say you have two plausible theories about time travel and they are the only two plausible ones you can think of and you think of them about equally often. In terms of the frequentist conception of probability, out of thinking trials there are only two plausible theory events and you estimate the two plausible theory events to have 50-50 probability for thinking trials. You have come up with many implausible theories and you assign them all zero probability. This is a rather vague (i.e., poorly specified) procedure and you could easily have assigned probabilities by some other vague procedure, but still following the frequentist conception of probability. In fact, practitioners of Bayesian analysis usually just assign equal initial probabilities (i.e., equal initial priors) to theories they think plausible and zero to all the others just as we've done for time travel. The assignment of equal probabilities to plausible theories is grandly called the principle of indifference. Actually, except probably in very well specified cases, there is no better way to assign initial probabilities to plausible theories. If you try to do better, you are effectively doing Bayesian analysis before you do Bayesian analysis rather than doing qualitative Bayesian analysis before you do Bayesian analysis. In fact, reasonably good initial probabilities are fine for practical Bayesian analysis and ideal Bayesian analysis can start with completely wrong ones and still work as we will prove.

There is major fact about the modern sciences in which Bayesian analysis is used which must be mentioned: their theories almost always have free parameters, and so are not fully-specified. Free parameters are almost always continuous variables that have to be set by the experimental/observational data itself. Thus, a theory with free parameters is, in one perspective, a continuous family of infinitely many theories, but that perspective is of no value, except that it must be understood. If your first interest is determining the true theory and not its free parameters, then Bayesian analysis makes use of marginalization. Marginalization is a whole lesson in itself which we will not do today.

Now I've promised a proof that ideal Bayesian analysis: i.e., a proof that it is a way to find truth (except for the aforementioned intractable cases). Note ideal Bayesian analysis is the ideal limit of Bayesian analysis without qualification. Both of forms of Bayesian analyis are, by the way, procedures. The proof of ideal Bayesian analysis is mathematical and has the certainty of a mathematical proof. Of course, we arn't going to be rigorous because we don't care about rigor—not today. We will do the proof right now and it's actually not so hard.

Say we have a set of theories T_i (which could be just hypotheses) about some aspect of reality, where i is an index that numbers the theories. For simplicity, we assume the set of theories T_i is finite. Actually, we can dispense with this assumption

(by using marginalization or otherwise), but having finite time at the moment, we won't. The theories are derived from initial knowledge (i.e., past data) which we now denote K_0, where the subscript 0 stands for step 0 of the procedure of ideal Bayesian analysis. We also have the initial probabilities $P(T_i|K_0)$ for all theories T_i being true given our initial knowledge K_0. In more elaborate Bayesian analysis jargon—which I won't use—the initial knowledge K_0 can be called the conditioning information or context. In the first step of a Bayesian analyis, we aquire new data and then we update our initial probabilities (i.e., our initial priors) to our initial posteriors using the newly acquired data and an updating formula which we will derive. The initial posteriors become the priors for the next step and so on until you find the true theory—except for the thrice aforementioned intractable cases.

Note the set of theories are those we consider plausible and we could set them have to equal initial probabilities by the principle of indifference. However, we do not need to do that and we will leave the $P(T_i|K_0)$ values general. Theories we consider implausible, we exclude from the set of theories which is the same as saying we assign them inital probabilities of zero. The probabilities $P(T_i|K_0)$ only need to have relative values: i.e., they do not have to sum to any particular number. However, it simplifies our proof of ideal Bayesian analysis to normalize them: i.e., to scale them to sum to 1: i.e., to scale them so that

$$\sum_i P(T_i|K_0) = 1 \ .$$

At face value, normalization means that theories not in the set absolutely definitely have probability zero. However, ideal Bayesian analysis does not take that normalization at face value. Some of excluded theories may actually have significant nonzero probabilities and if so, they might re-enter the ideal Bayesian analysis and we will show where in just a moment. What of the scaling of updated probabilities? The ideal Bayesian analysis automatically normalizes all updated probabilities as we will show.

I define the set of theories T_i as adequate in a Bayesian analysis sense if it contains the true theory. This is because the ideal Bayesian analysis will eventually find that true theory. What if the set does not contain the true theory? Then the ideal Bayesian analysis will eventually show that all theories in our set have zero probability which occurs when the mean likelihood (which we define in due course) goes to zero. In fact, whenever the mean likelihood goes to zero, the ideal Bayesian analysis requires us to introduce new theories (which, as we indicated just a moment ago, might include some theories we excluded initially) and eventually we will introduce the true theory though we will not know it is the true theory until the ideal Bayesian analysis is complete. Why are we guaranteed to introduce the true theory? Somewhere as we approach the limit of having all data relevant to the aspect of reality we are considering, we will be able to find the true theory to introduce (except for intractable cases). So even if the true theory is not in our original set of theories, ideal Bayesian analysis will find the true theory (except for intractable cases).

Now to be a bit more mathematical, we label each step of the ideal Bayesian analysis by index ℓ. We start with step $\ell = 0$ (i.e., no step done) and the first step is step $\ell = 1$. Now say we have completed to step $\ell - 1$. The next step is step ℓ. Before we do step ℓ, we have probabilities $P(T_i|K_{\ell-1})$, where our step $\ell - 1$ knowledge $K_{\ell-1} = K_0 D_1 D_2 D_3 \dots D_{\ell-1}$, where again the "product" is interpreted as the intersection and the D values are data acquisitions indexed by the step number. To start step ℓ, we acquire new data D_ℓ. The new data doesn't have to be experimental/observational data in an obvious sense. It could be other kinds of knowledge: for important example, a new theory about some aspect of reality that impacts our knowledge about the set of theories T_i about their different aspect of reality. For the ideal Bayesian analysis, we will assume that data D_ℓ has no uncertainty (i.e., it is exactly correct).

To obtain a preliminary version updating formula used to update the probabilities from priors $P(T_i|K_{\ell-1})$ to posteriors $P(T_i|K_\ell)$, we apply Bayes' theorem replacing A by T_i, B by data D_ℓ, and K by $K_{\ell-1}$. Said formula is

$$P(T_i|K_\ell) = P(T_i|D_\ell K_{\ell-1}) = \frac{P(D_\ell|T_i K_{\ell-1})P(T_i|K_{\ell-1})}{P(D_\ell|K_{\ell-1})} \ ,$$

where $K_\ell = D_\ell K_{\ell-1}$ (with the "product" interpreted as the intersection, of course) and $P(D_\ell|T_i K_{\ell-1})$ is called the likelihood—it is the probability of the data given the intersection $T_i K_{\ell-1}$ which includes knowledge about the experiment/observation used to obtain D_ℓ which we need to calculate said likelihood $P(D_\ell|T_i K_{\ell-1})$. Note likelihoods do not need to be normalized. All one needs is their relative sizes as we will show. However, one can normalize them if one wants too. What $P(D_\ell|K_{\ell-1})$ is we will discuss in a bit.

The likelihood $P(D_\ell|T_i K_{\ell-1})$ takes some more explication. For this explication, we will assume we have normalized the likelihoods so that summing them over i gives 1. Say the data D_ℓ is exactly true, the past knowledge $K_{\ell-1}$ includes all the particulars about experiment/observation that the theory T_i requires for a full prediction, and the theory T_i is an fully-specified deterministic theory. There are three cases. First case: if T_i is true, $P(D_\ell|T_i K_{\ell-1})$ must be 1. Second case: if T_i is false, $P(D_\ell|T_i K_{\ell-1})$ could be 1 if the theory fortuitously predicts D_ℓ—it can't predict all possible true events or it would be true, but it can predict them sometimes. In fact, the data D_ℓ is not decisive between the first and second cases: it does not tell us if the theory is true or false: it could be either. Third case: if $P(D_\ell|T_i K_{\ell-1}) = 0$, then the theory T_i is false and the data D_ℓ is decisive. But what if D_ℓ has experimental/observational uncertainty and/or $K_{\ell-1}$ does not include all the particulars about experiment/observation that the theory requires for a full prediction? There are ways of dealing with those cases, but those ways take us off into the details of non-ideal Bayesian analysis, and so we won't go into those cases. But what if T_i is a probabilistic theory? Then $P(D_\ell|T_i K_{\ell-1})$ could be anywhere in the range $[0, 1]$ whether the theory T_i is true or not. For example, say T_i is the theory of a perfect die: all it predicts is that the probability of throwing any of 6 numbers is 1/6. Thus, for the perfect die theory, $P(D_\ell|T_i K_{\ell-1}) = 1/6$ in all

cases: i.e., for all throws and whether the theory is true or not for the die you are actually using. Now probabilistic theories turn up all the time in the modern sciences mentioned above (e.g., epidemiology, etc.) and others in which Bayesian analysis is used. So normalized likelihoods that are neither 0 nor 1 turn up all the time in Bayesian analysis, whether ideal or not. But that in itself is not a problem for ideal Bayesian analysis, but you may have to go to infinite steps to reach the true theory.

A finicky point must be mentioned here. Some authorities assert that likelihoods are not actually probabilities. But they are. Likelihoods arise from Bayes' theorem which only has probabilities in it. Therefore, they are probabilities, and therefore can be meaningfully summed (without multiplying them by any factors, I mean) though one usually has no reason to do that. What likelihoods are not is probability densities (including in cases where they are set up for marginalization) and that is what those authorities mean I think: i.e., likelihoods are not probability densities. That they are not probability densities means likelihoods cannot be meaningfully integrated over just by themselves. In cases of marginalization, a likelihood is multiplied by a probability density and the product is a probablitiy density and that product is, in fact, what is meaningfully integrated over in marginalization. I will not explicate further here and now.

Now that we have dealt with the likelihood $P(D_\ell|T_iK_{\ell-1})$ in the preliminary version of the updating formula, what actually is the denominator $P(D_\ell|K_{\ell-1})$ in the preliminary version of the updating formula? It is the probability of getting the data D_ℓ given the knowledge $K_{\ell-1}$. However, in general, there might be many ways of specifying $P(D_\ell|K_{\ell-1})$ depending on the purpose you use it for or how you think of it. But for Bayesian analysis, there seems to be only one good specification: the mean likelihood $\langle P(D_\ell|T_iK_{\ell-1})\rangle$, where the angle brackets mean mean. Of course, we need a formula for the mean likelihood that leads to the ideal Bayesian analysis being proven true and that, as becomes clear below, happens to be

$$\begin{aligned}\langle P(D_\ell|T_iK_{\ell-1})\rangle &= \sum_i P(D_\ell|T_iK_{\ell-1})P(T_iK_{\ell-1}) \\ &= \sum_i P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1})P(K_{\ell-1}) \\ &= \sum_i P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1})\end{aligned}$$

where the likelihoods in the averaging process are weighted by their probabilities of turning up in the ideal Bayesian analysis we are doing (i.e., are weighted by $P(T_iK_{\ell-1})$: the probability of intersection $T_iK_{\ell-1}$ being true), where we have used the product rule, and where $P(K_{\ell-1})$ is suppressed in the third line since $P(K_{\ell-1}) = 1$ since we actually have $K_{\ell-1}$ in hand. Note if the likelihoods $P(D_\ell|T_iK_{\ell-1})$ are equal for all i and the probabilities $P(T_i|K_{\ell-1})$ are normalized, then $\langle P(D_\ell|T_iK_{\ell-1})\rangle = P(D_\ell|T_iK_{\ell-1})$, of course.

Now if our set of theories T_i were the only theories with nonzero probabilities, our theory probabilities were known to be exactly true just before step ℓ, and our

likelihoods were calculated exactly, then

$$P(D_\ell|K_{\ell-1}) = \langle P(D_\ell|T_iK_{\ell-1})\rangle$$

is exactly correct since $P(D_\ell|K_{\ell-1})$ can then only be expanded in the complete and exactly known terms $P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1})$. However, since the first two of those three conditions do not hold in general even in ideal Bayesian analysis, all we can say is that specifying $P(D_\ell|K_{\ell-1})$ by $\langle P(D_\ell|T_iK_{\ell-1})\rangle$ seems the only good choice for Bayesian analysis, ideal or otherwise, since we have the ingredients $P(D_\ell|T_iK_{\ell-1})$ and $P(T_i|K_{\ell-1})$ in hand. Moreover, if we chose our set of theories T_i very well (i.e., they were overwhelmingly the most probable theories by the time we have knowledge $K_{\ell-1}$), then it is very likely that $\langle P(D_\ell|T_iK_{\ell-1})\rangle$ by any specification well approximates $P(D_\ell|K_{\ell-1})$. However, no matter what our set of theories T_i, we will soon see that using the mean likelihood $\langle P(D_\ell|T_iK_{\ell-1})\rangle$ as a specification for $P(D_\ell|K_{\ell-1})$ allows the ideal Bayesian analysis.

Now in the preliminary version of the updating formula, we substitute for $P(D_\ell|K_{\ell-1})$ with mean likelihood $\langle P(D_\ell|T_iK_{\ell-1})\rangle$ and obtain the (final) updating formula for ideal Bayesian analysis

$$P(T_i|K_\ell) = \frac{P(D_\ell|T_iK_{\ell-1})}{\langle P(D_\ell|T_iK_{\ell-1})\rangle} P(T_i|K_{\ell-1}) \ .$$

The updating formula shows that the overall scalings of the priors $P(T_i|K_{\ell-1})$ and the likelihoods $P(D_\ell|T_iK_{\ell-1})$ have no effect since those scalings cancel out. Thus, there is no need to normalize either of the probabilities $P(T_i|K_{\ell-1})$ and the likelihoods $P(D_\ell|T_iK_{\ell-1})$. However, the posteriors are automatically normalized since the sum over all i on both sides of the updating formula gives

$$\sum_i P(T_i|K_\ell) = \frac{\sum_i P(D_\ell|T_iK_{\ell-1})P(T_i|K_{\ell-1})}{\langle P(D_\ell|T_iK_{\ell-1})\rangle} = \frac{\langle P(D_\ell|T_iK_{\ell-1})\rangle}{\langle P(D_\ell|T_iK_{\ell-1})\rangle} = 1 \ ,$$

unless all $P(D_\ell|T_iK_{\ell-1}) = 0$ in which case the posteriors are, in fact, undefined and all theories T_i are falsified. The upshot is that only the initial probabilities $P(T_i|K_0)$ are not automatically normalized by the ideal Bayesian analysis (unless all $P(D_\ell|T_iK_{\ell-1}) = 0$).

However, we did normalize the initial probabilities $P(T_i|K_0)$ for simplicity and that simplicity makes the following statement completely general to the ideal Bayesian analysis. Statement: the updating formula shows that the posterior $P(T_i|K_\ell)$ increases/decreases relative to the prior $P(T_i|K_{\ell-1})$ if likelihood $P(D_\ell|T_iK_{\ell-1})$ exceeds/subceeds the mean likelihood. In other words, we see explicitly mathematically how new data D_ℓ acts to update the priors $P(T_i|K_{\ell-1})$ to the posteriors $P(T_i|K_\ell)$. And since the posteriors are based on more data than the priors (i.e., the new data D_ℓ), the posteriors are more accurate probabilities than the priors usually. Why only usually? The new data D_ℓ, even it has no uncertainty (which we explicitly assumed), may be accidently unrepresentative of

the true theory, and so lead the ideal Bayesian analysis temporarily astray from the path to the true theory. But only temporarily for reasons we are just now going to give.

Say we keep doing updating steps in the ideal Bayesian analysis procedure (i.e., acquiring new data D_ℓ and updating from priors to posteriors using the updating formula) and our set of theories T_i is adequate (i.e., contains the true theory), are we guaranteed to arrive at truth? The truth being one theory T_i with probability 1 (i.e., $P(T_i|K_\ell) = 1$) and the others with probability zero (i.e., $P(T_i|K_\ell) = 0$). The answer is yes (with a pure math qualification given below) if we can keep doing steps until we reach $K_{\ell_{\max}}$ which is all knowledge relevant to the aspect of reality we are considering. However, $\ell_{\max}$ may have to go to infinity to have all relevant knowledge. Hypothetically, this could be the case for a theory that makes absolutely only probabilistic predictions. In fact, it's hard to think of such a theory and we won't try think of it now. In any case, we allow ideal Bayesian analysis to have infinite $\ell_{\max}$. But can we always reach $K_{\ell_{\max}}$ even with infinite $\ell_{\max}$. In pure math, there may be cases where we can't, but I will skip any discussion out sheer ignorance. What about cases in physical reality? My current thinking is that physical reality has enough features to interrogate about any aspect of reality that we can ideally reach $K_{\ell_{\max}}$ for any aspect of reality and reach it with finite $\ell_{\max}$ as long as our new data acquisitions do not become vanishingly small. Note the word "ideally" in the last statement. The features are there to be interrogated, but practically we may not be able to interrogate them. So in many practical cases, it is impossible to have all relevant knowledge (i.e., $K_{\ell_{\max}}$).

However, I'm not quite sure that even ideal Bayesian analysis can arrive at truth for all cases in physical reality. So I'm just going to call all cases (in pure mathematic and physical reality) where ideal Bayesian analysis can't reach truth by the already-introduced term intractable cases which admittedly take a lot more discussion to adequately define. And now I assert that we have proven that ideal Bayesian analysis starting with an adequate set of theories is guaranteed to arrive at truth, except for intractable cases which I won't keep mentioning all the time when they should be mentioned in order to not sound like a broken record.

What if our set of theories T_i is inadequate (i.e., does not contain the true theory)? At some step, all our theories fail to satisfy the new data D_ℓ: i.e., $P(D_\ell|T_iK_{\ell-1}) = 0$ for all i, and thus $\langle P(D_\ell|T_iK_{\ell-1})\rangle = 0$: i.e., the mean likelihood is zero. At this point, as mentioned earlier, you must introduce new theories (i.e., a new set of theories) which ideally you can do by deduction from your knowledge $K_\ell = D_\ell K_{\ell-1}$. Then you restart the ideal Bayesian analysis using the new set of theories. If your new set of theories turns out to be inadequate, you will eventually have to introduce another new set of theories and so on. The conclusion is that whether or not your initial set of theories is adequate, the ideal Bayesian analysis will find a set that is adequate and then the ideal Bayesian analysis will find the true theory at least after obtaining all relevant knowledge: i.e., $K_{\ell_{\max}}$. Except, to sound like a broken record, for intractable cases.

Actually, in doing ideal Bayesian analysis you are always free to introduce new

theories and discard old theories if your current knowledge $K_\ell = D_\ell K_{\ell-1}$ suggests the former are plausible and the latter are implausible. You do have to set new probabilites for the updated set of theories. Probably setting them all equal (i.e., using the principle of indifference) and, for simplicity, normalizing them is the best practice. Updating, your set of theories in this qualitative Bayesian analysis way is likely to speed the ideal Bayesian analysis to finding the true theory.

Of course, in many practical cases as aforementioned, it is impossible to have all relevant knowledge (i.e., $K_{\ell_{\max}}$), but one can often have enough relevant knowledge so that one theory is so probable that it can be accepted as true pending any future failure. But there is better perspective. Instead of saying "true pending any future failure," we say the theory is true despite any failures to come, but its realm of validity is just narrowed by those failures. Now you might say we are just lowering the bar, but that is an inadequate perspective for powerful theories about reality known to have vast realms of validity, that include the axiom that they hold in an ideal limit that can be closely approached in practice, and that have expanded our wisdom about reality. Examples of such theories are Newtonian physics, the germ theory of disease, and the theory of evolution by natural selection. Aside from absolute philosophical skepticism—which Bertrand Russell for one admitted is logically viable if perfectly useless—proven mathematical theories (including ideal Bayesian analysis itself—if you except intractable cases) are just true by logic and other powerful theories are true I maintain by the argument just given, *sans phrase*.

Howsoever, returning to the Bayesian analysis steps, ideal or otherwise, where truth is not yet in hand, we want to choose our data acquisitions D_ℓ judiciously. What does that mean? To go way back, Ernest Rutherford (1871–1937), the discoverer of the atomic nucleus in 1911, once said "If you need statistics, you did the wrong experiment." Of course, all aphorisms are both right and wrong including this one. But the valid kernel of Rutherford's aphorism is that if we can do the experiment that minimizes the need for statistics, good and if we can do the decisive experiment, better. But in many fields, such as those we mentioned above (epidemiology, etc.) and others, statistics, including most especially Bayesian analysis, can't be avoided. Rutherford flourished in a simpler age when there were more problems that didn't need much statistics—but they usually needed some, in fact. Now for Bayesian analysis, Rutherford's aphorism and "judiciously" mean the same thing: choose data acquisitions such that they decisively rule out theories as fast as possible and rule in more plausible theories as you go along, and so rule in the true/powerful theory pronto. Being judicious just saves a lot of kicking and screaming.

I should emphasize the ideal Bayesian analysis procedure that I've outlined is virtually impossible, except for trivial cases. In nontrivial cases, all kinds of simplifications and approximations are needed. A major one is that the likelihoods $P(D_\ell|T_i K_{\ell-1})$ and especially their marginalizable versions (which are the usual ones needed) are often very hard to compute even with supercomputers. This is often because data D_ℓ, the innocent looking data D_ℓ, is often in important cases petabytes and getting worse as science takes up harder and harder problems. In nontrivial

Bayesian analysis, there's a bit of an art to finding the optimum simplifications and/or approximations for any particular likelihood calculation. And there is even more than a bit of an art to marginalization. All too often for same data, people marginalize differently and come to very different conclusions.

Now a question for Corin. Given that the ideal Bayesian analysis procedure is virtually impossible except for trivial cases, why should one bother with it at all? Why not just describe non-ideal Bayesian analysis procedures?
Moment of thought.
Corin: The ideal Bayesian analysis procedure is the true limit of all the non-ideal ones. It proves that they should work to some degree and you can always improve how they work by getting closer to the true limit. You yourself implied as much about true/powerful theories just a moment ago. Now you cannot trust a theory which fails when you try applying it more exactly. In fact, such a theory is falsified. But to half take back what I just said, some theories apply exactly only in such ideal cases that they can only be approached closely in the imagination, but they are still very useful in understanding reality. I learnt all this in high-school physics and surprised I am to speak it *ex tempore*.
Peter: What is the relevance of ideal Bayesian analysis to the scientific method?
Corin: It is the ideal quantified scientific method, and thus is the proof of the scientific method. At least that is so to my satisfaction at this moment. But I'm still worried about those intractable cases.
Peter: A-plus, Corin.
Corin: Do you know anything about the Spanish novel from 1898 to 1975?
Peter: Let's pretend I don't.
Corin: Right.
Peter: How about lunch and a pint?
Corin: Do we have to discuss Bayesian analysis?
Peter: Only as applied to the football pools.
Corin: (*Thinking.*) I guess that's OK. (*Thinking.*) So going back to the beginning, could you come with me this Saturday to see *The van Heck Nativity* and Red House? You could stay the night. We'd come back Sunday morning, in my case, to drudgery.
Peter: Thank you very much. I'd be happy to do that.
Corin: The only family who's there is my sister Serena. She has autism, but she can talk. She's really quite social.
Peter: I'd be glad to talk to Serena. I'm a little experienced with interacting with people with autism. I've had students with that nature.
Corin: Of course, my older brother might make an appearance.
Peter: Isn't he the Member of Parliament for Rydal?
Corin: It's worse than that I'm afraid.
Peter: The stuff in the news?
Corin: Worse.
Peter: He's pro-Brexit?

Scene 3