The Bayesian Workflow

> **Let us cheer up a little bit on the Bayesian side of things to its frequentist counterpart...**

On one hand, Bayesian thinking has a broader bandwidth regarding how a Data Scientist can choose how to model the data coming from any population or system of interest. For instance, as in Lecture 4 - Markov Chain Monte Carlo, Stan, and Complex Bayesian Models, we can model the mean daily counts of bike-share rides according to how the weather is behaving in terms of thresholds of temperatures which will end up with interesting **random latent variables** in the forms of "*elbows*" and "*slopes*". This would be translated into specific mathematical functions **shaping our determined Poisson likelihood**, allowing us to include further features as a **systematic component** making a Poisson regression catered to our specific statistical needs. Furthermore, we combine this likelihood with our prior knowledge (possibly coming from experts in the field) on our latent variables, which makes our model a comprehensive tool combining this knowledge with our gathered evidence contained in the likelihood (i.e., **the Bayes' rule in action**).

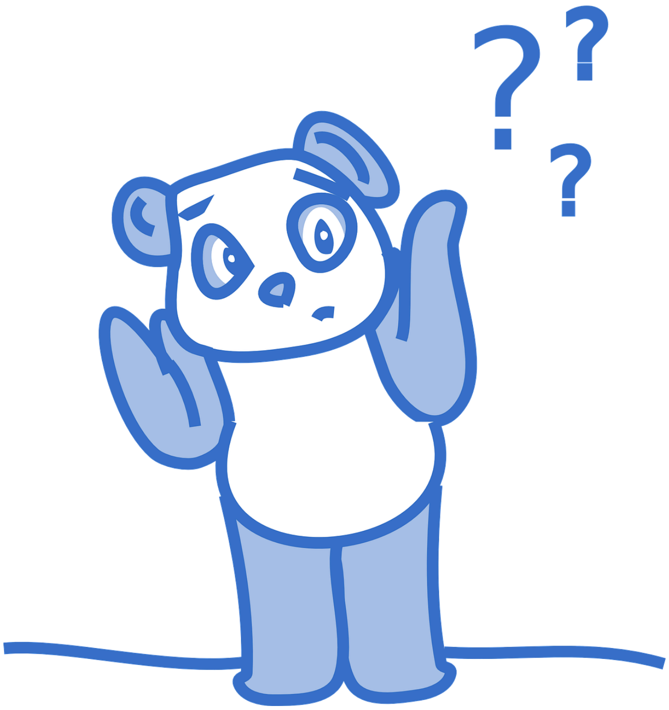


Fig. 1 The Questioning Panda comes back.

On the other hand, on the frequentist side, we might encounter a "traditional" Poisson regression involving a **systematic component** with the so-called **fixed** regression parameters (i.e., $\beta_0, \beta_1, \ldots, \beta_k$ in the presence of k regressors or features) which would involve a linear combination along with the observed values of the regressors per row in the **training dataset**. This frequentist generalized linear model (GLM) might look like a standard "*statistical recipe*" to make inference or predict a count-type response (i.e., the Y) in the presence of a set of features (i.e., the Xs). Note that this way of modelling data in the form of frequentist recipes has a certain appeal for mainstream Data Science practice, given its ease in explaining all statistical findings to our stakeholders, which is understandable to a certain extent. On top of that, tools like the **DSCI 562's regression workflow** (a.k.a. **the Data Science workflow**) make this frequentist modelling handy to execute so we can come with a **coherent final storytelling** in line with the **initial main statistical inquiries** of our stakeholders.

Having said all this, given the broader bandwidth we have in Bayesian data modelling, we might wonder:

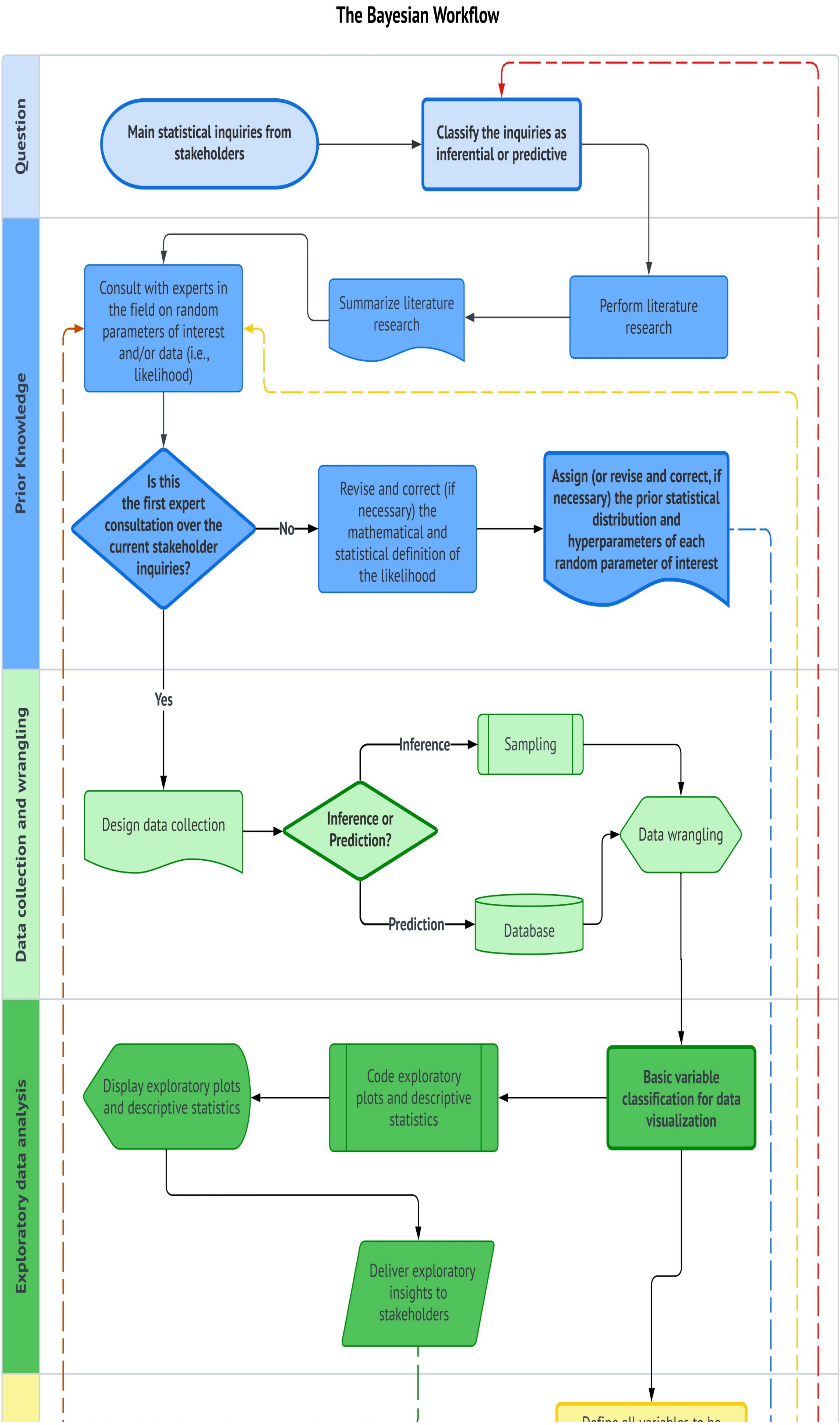
> **Is it also possible to have a general Bayesian workflow to approach any data modelling problem in general?**

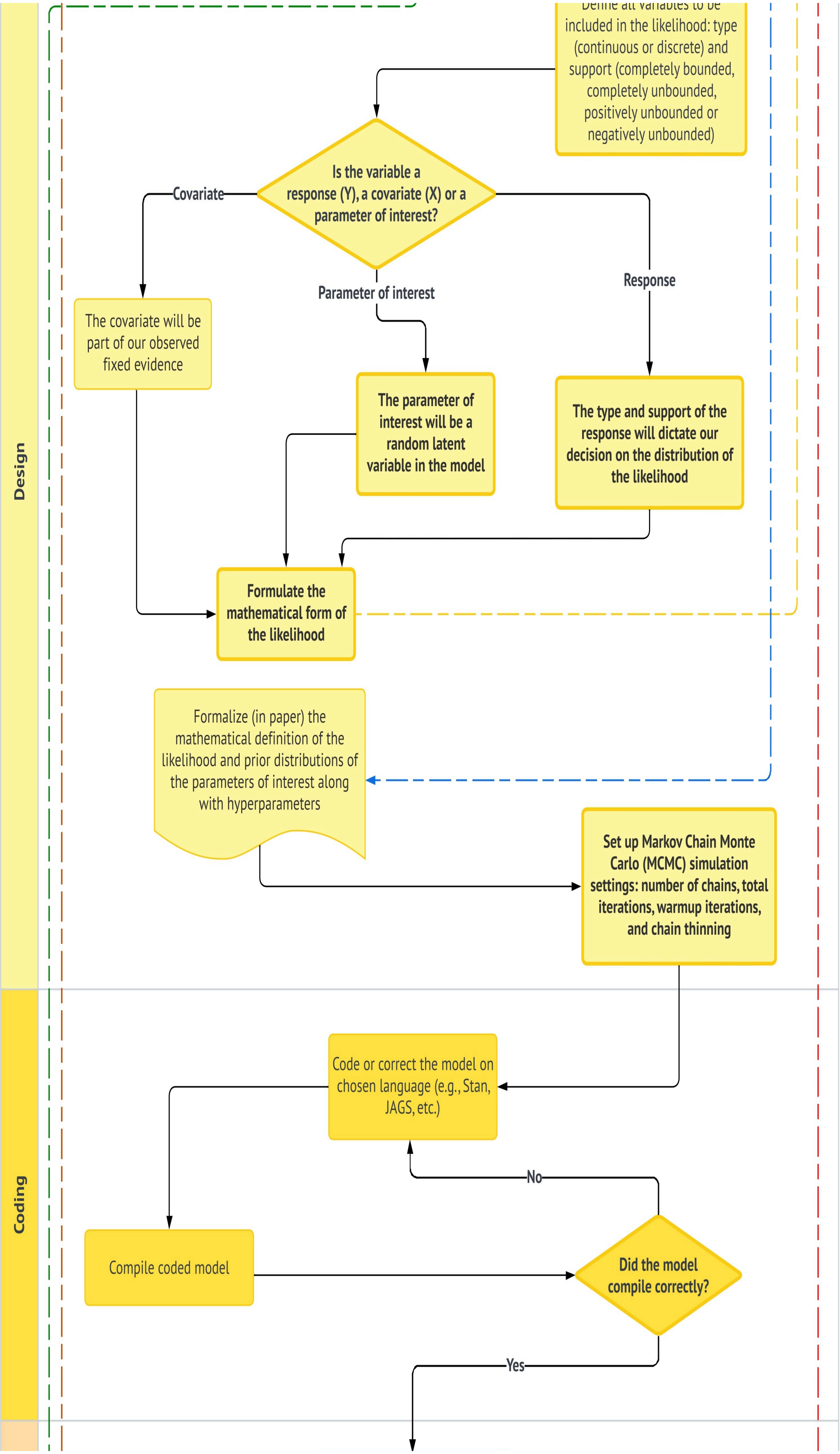
The answer to the above question is "*yes*" to a certain extent. Accordingly, besides incorporating the application of the Bayes' rule involving prior knowledge and observed evidence in this workflow, we need to define our **simulation settings** in regards to our chosen **Markov Chain Monte Carlo (MCMC)** approach to approximate the posterior distributions of our **random parameters of interests** (i.e., our **random latent variables**) combined with **model diagnostics** (as in Lecture 8 - More Hierarchical Modelling and MCMC Diagnostics). Finally, by also incorporating the **main statistical inquiries** and **final storytelling** with **proper variable classification** and **initial exploratory data analysis (EDA)**, this Bayesian workflow can be depicted as in Fig. 2. Note this workflow is briefly implied at the beginning of Lecture 3 - Bayesian Statistics in Action: The Beta-Binomial Model and Lecture 4 - Markov Chain Monte Carlo, Stan, and Complex Bayesian Models.

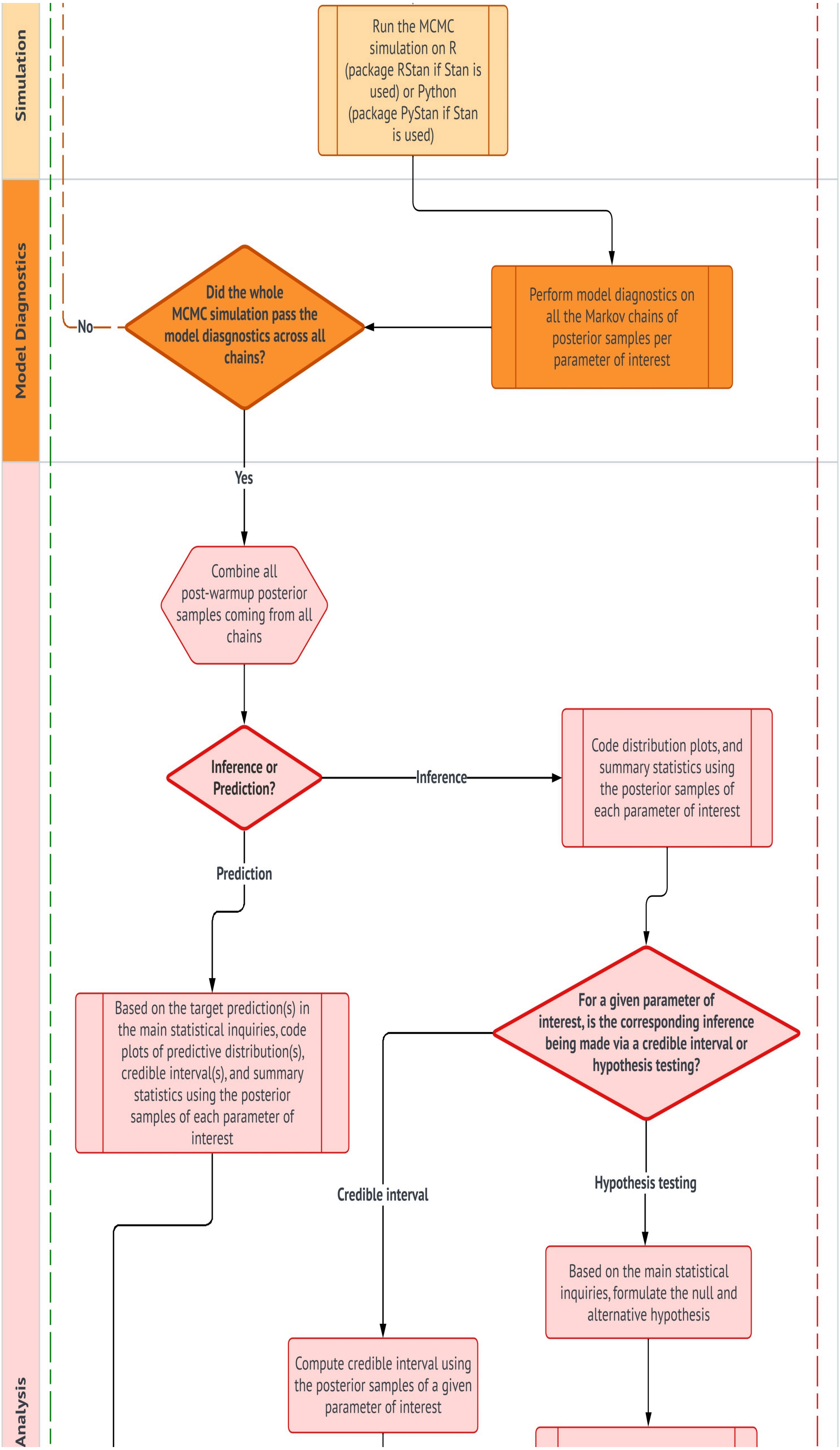
Therefore, let us illustrate each one of the stages (i.e., horizontal lanes in the workflow):

1. **Question.** This stage refers to the main statistical inquiries of our stakeholders (**inferential and/or predictive**).

2. **Prior knowledge.** This stage involves consulting experts in our problem regarding the random parameters of interest, their prior distributions, hyperparameters, and our likelihood.
3. **Data collection and wrangling.** This is merely how data for our likelihood will be collected (according to the nature of the inquiries) along with its corresponding wrangling.
4. **Exploratory data analysis.** This stage is related to basic variable classification, which is used to perform data visualization so we can deliver descriptive and exploratory insights to our stakeholders.
5. **Design.** One of the most important stages in Bayesian data modelling. It includes a comprehensive design of the likelihood and prior setup (according to our defined random parameters of interest). Note this stage is iterative with the one related to **prior knowledge**. Furthermore, this stage also outlines the simulation settings for MCMC. To check how MCMC performs the simulation, in general, you can check Insights of Markov Chain Monte Carlo via the Gamma-Poisson Model.
6. **Coding.** Once the model is designed, we must code and compile it.
7. **Simulation.** The coded and compiled model will be used to perform the corresponding MCMC simulation in `R` or `Python`.
8. **Model diagnostics.** All our Markov chains of posterior samples per parameter of interest will be properly assessed to proceed to our **posterior analysis** (check Lecture 8 - More Hierarchical Modelling and MCMC Diagnostics).
9. **Posterior analysis.** This stage includes all posterior summaries and conclusions, depending on the nature of the main statistical inquires.
10. **Storytelling.** The conclusions of our data findings catered to our stakeholders. Note this stage might be iterative concerning the very first stage of the whole workflow.


The Bayesian Workflow
Question
Main statistical inquiries from stakeholders
Classify the inquiries as inferential or predictive
Prior Knowledge
Consult with experts in the field on random parameters of interest and/or data (i.e., likelihood)
Summarize literature research
Perform literature research
Is this the first expert consultation over the current stakeholder inquiries?
No
Revise and correct (if necessary) the mathematical and statistical definition of the likelihood
Assign (or revise and correct, if necessary) the prior statistical distribution and hyperparameters of each random parameter of interest
Yes
Data collection and wrangling
Inference
Sampling
Design data collection
Inference or Prediction?
Data wrangling
Prediction
Database
Exploratory data analysis
Display exploratory plots and descriptive statistics
Code exploratory plots and descriptive statistics
Basic variable classification for data visualization
Deliver exploratory insights to stakeholders
Define all variables to be


Define all variables to be included in the likelihood: type (continuous or discrete) and support (completely bounded, completely unbounded, positively unbounded or negatively unbounded)
Is the variable a response (Y), a covariate (X) or a parameter of interest?
Covariate
Parameter of interest
Response
The covariate will be part of our observed fixed evidence
The parameter of interest will be a random latent variable in the model
The type and support of the response will dictate our decision on the distribution of the likelihood
Formulate the mathematical form of the likelihood
Formalize (in paper) the mathematical definition of the likelihood and prior distributions of the parameters of interest along with hyperparameters
Set up Markov Chain Monte Carlo (MCMC) simulation settings: number of chains, total iterations, warmup iterations, and chain thinning
Design
Coding
Code or correct the model on chosen language (e.g., Stan, JAGS, etc.)
No
Compile coded model
Did the model compile correctly?
Yes


Simulation
Run the MCMC simulation on R (package RStan if Stan is used) or Python (package PyStan if Stan is used)
Model Diagnostics
Perform model diagnostics on all the Markov chains of posterior samples per parameter of interest
Did the whole MCMC simulation pass the model diasgnostics across all chains?
No
Yes
Analysis
Combine all post-warmup posterior samples coming from all chains
Inference or Prediction?
Inference
Code distribution plots, and summary statistics using the posterior samples of each parameter of interest
Prediction
Based on the target prediction(s) in the main statistical inquiries, code plots of predictive distribution(s), credible interval(s), and summary statistics using the posterior samples of each parameter of interest
For a given parameter of interest, is the corresponding inference being made via a credible interval or hypothesis testing?
Credible interval
Hypothesis testing
Based on the main statistical inquiries, formulate the null and alternative hypothesis
Compute credible interval using the posterior samples of a given parameter of interest

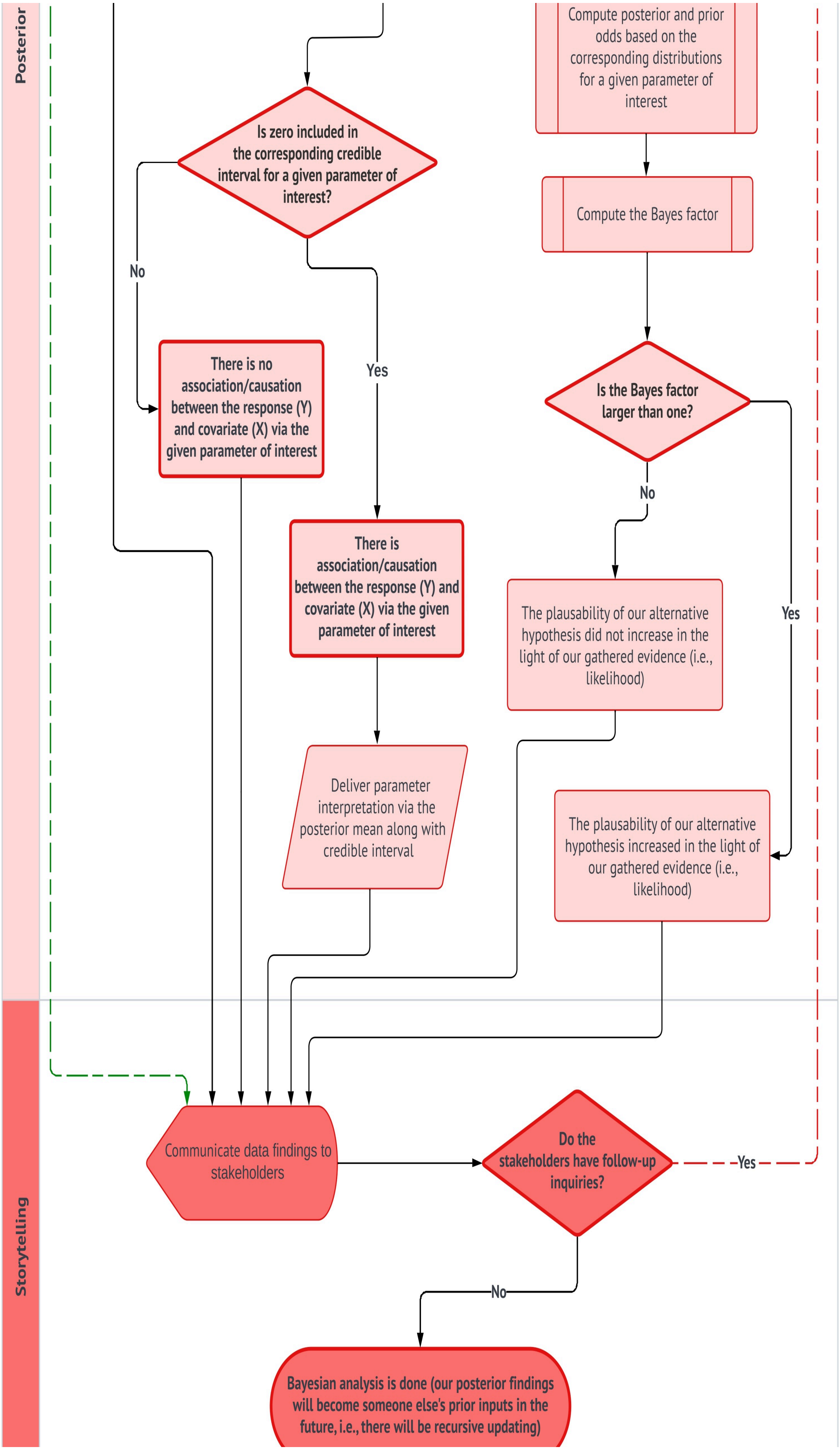

Posterior
Compute posterior and prior odds based on the corresponding distributions for a given parameter of interest
Is zero included in the corresponding credible interval for a given parameter of interest?
Compute the Bayes factor
No
Yes
There is no association/causation between the response (Y) and covariate (X) via the given parameter of interest
Is the Bayes factor larger than one?
No
There is association/causation between the response (Y) and covariate (X) via the given parameter of interest
The plausability of our alternative hypothesis did not increase in the light of our gathered evidence (i.e., likelihood)
Yes
Deliver parameter interpretation via the posterior mean along with credible interval
The plausability of our alternative hypothesis increased in the light of our gathered evidence (i.e., likelihood)
Storytelling
Communicate data findings to stakeholders
Do the stakeholders have follow-up inquiries?
Yes
No
Bayesian analysis is done (our posterior findings will become someone else's prior inputs in the future, i.e., there will be recursive updating)

Fig. 2 General Bayesian workflow for **inferential** and **predictive** inquiries. Components in bold letters indicate crucial steps in the workflow, whereas dashed coloured arrows depict jumps across more than one stage.