Data and the reproducibility of research results

What are the roles of 'data' and 'reproducibility' in the scientific method?

"Data is used to support scientific conclusions. In the context of a statistical analysis, data is used quantify the effect or presence of some real-world phenomenon on someoutcome of interest. Reproducibility is the concept that conclusions reached under some experimental method can be obtained again using the same methodology."

"I suppose that data is widely used to empirically test and prove a theory. Seeing a theory work on data of an actual application is very convincing when arguing about the usefulness of a new theoretical method. If the results cannot be reproduced, then all arguments based on that results become meaningless."

What are the federal requirements on sharing data?

"Based on the America COMPETES Act, the federal law requires civilian federal agencies to provide guidelines, policy, and procedures, to facilitate and optimize the open exchange of data and research. To me, this rule seems to be reasonable and beneficial for almost everyone."

To what extent do you think these rules are enforced?

"I am guessing they are often difficult to enforce especially since they seem to be relatively vague, with room for interpretation, rather than strict policies."

"I feel in some cases there are delays for data sharing and even in certain cases, data are not openly accessible. Although some confidential data need to be kept confidential but much of the collected data should be open after publication of the main paper tobe useable by other authors."

Advanced statistical methods often require sophisticated computational implementations. Should statistical researchers be expected to share their computer code on request?

A. *"Yes, going a step farther, I think that researchers should publish their code when they publish and article."*

B. *"Absolutely not. I believe that the methodology should be detailed in any report, but sharing computer code has potential to leave the innovative aspect of the experiment susceptible for theft without credit. I think the more the statistical researcher feels comfortable that the code is protected, the more likely it is reasonable for him or her to share the computer code."*

"Yes, I think statistical researchers have to share, because there are so many sophisticated computational implementations. Other people cannot reproduce your result ifyou dont share your code."
"Sharing codes would certainly help others to implement methods one provided in paper.Also, researchers normally would share their codes only if they are really confidentabout their methods. Whats more, such kindness and transparency would definitelyhelp build ones reputation."
"Yes. Sharing the code can help distribute proposed methodology quickly and broadly."
"It is perhaps appropriate to share partial code or the framework of the code whenrequested."

What is the difference between data and a statistical model for the data? For example, comment on the assertion "Let $y_1, \ldots, y_n$ be independent identically distributed data."

"Data refer to directly observed quantities or features, while a model is a hypothesized process by which the data were generated, or pattern that the data follow. So in this example the y_i values are data, but the statement that they are i.i.d. is a statistical model."

"We must be careful not to confuse data with the abstractions we use to analyze them." (William James, 1842–1910)

The remaining questions consider the following hypothetical case study:
Ben is a Statistics PhD student who has written computer code for a simulation study to test a new statistical theory and methodology which he is developing. He plans to put the results in his thesis and to publish them in a journal paper. The results of the simulations are usually consistent with his theoretical analysis. However, sometimes the code crashes, particularly when investigating more extreme values of the parameter space. Ben has checked and rechecked the code very carefully, and cannot find any error. He decides that there must be some weird numerical effect, perhaps to do with occasional extremely large or small numbers. Ben decides to report the results only in the region of the parameter space where the code never crashed.

Is Ben's course of action a reasonable balance between the necessity to make progress on his thesis and his desire to report correct results?

A. *"No. Ben should solve this problem or write this problem in his paper."*

B. *"I think that Bens course of action is quite reasonable as he is not fabricating any kind of study."*

"[. . .] Ben should speak to his advisor and try to determine what may be going on."

"Ben seems to prefer wanting to make progress for his thesis more than a correct result report. But I think it will be fine if he points out exactly the parameterspace where his code runs well. It can still be a contribution to the scientific field."

What are the 'data' in this example? What is 'reproducibility' in this context?

"Data here are all the simulation results (consisitant ones and also inconsistant ones). Reproducibility here means that if other researchers try to implement Bens approachin the same senerio, then they should get similar results/performances with results reported by Ben."

Ben asks your opinion on how to proceed. What is your advice?

"Ben should fix the bug!"

"It is ok that the code is crashing in some regions. Try to figure out why, is this an issue with numerical storage in computers, seek advice from a computer science person tosee if there is some way to fix it. If not, Ben can surely publish his paper if the model is fine, his theoretical results are perfect and his simulations back up his theory most of the case and maybe just point out that it fails in this region and my code works for this type of datasets."

"The minimum Ben should do is to state the region of the parameter space where the reported results come from and to report that other region yield unexpected results."

"Obtain input from others who are very experienced with coding to see if there are any errors in the code. If none are found, consult an expert on numerical stability issuesand try to find evidence that indeed the code crashes due to numerical instabilities."

"I would encourage Ben to ask his advisor; however, my personal opinion is that he ought to pin down the numerical effect before proceeding."