Demystifying Data Science
Part 3 – Common Data Science Mistakes

Mistake 1: Not Defining the Problem (or Solving the Wrong Problem)

Time & resources can easily be wasted if the business & data science problems are not clearly defined. Remember that all analysis decisions are based on the problem that is trying to be solved!

Solution: Make sure there is collaboration between business leaders, domain experts, and data scientists to clearly define the problem and the desired outcome.

Mistake 2: Not Having the Right Data

Having the right data is essential – a model is only as good as the data that was used to train it. *"Garbage in, garbage out"*.

> *Not everything that can be counted counts, and not everything that counts can be counted.*
>
> - William Bruce Cameron

Solution: After defining your problem, see if you have the right data needed to solve the problem. Do you have adequate predictors? Do you have labelled responses? Is the data integrity high? How many data points do you have?

Mistake 3: Not Thinking About Deployment

At the start of the data science process, we should be thinking about how the results of the analysis will be used.

Solution:

Consider the following when thinking about deployment:

- **Modularity**: how will different parts of the data science process be organized?
- **Reproducibility**: will the results from the model be reproducible?
- **Scalability**: does the model need to scale to handle large datasets in a reasonable amount of time?
- **Extensibility**: how will the code be modified for future tasks?
- **Testing**: how will the code be tested as new versions are developed?

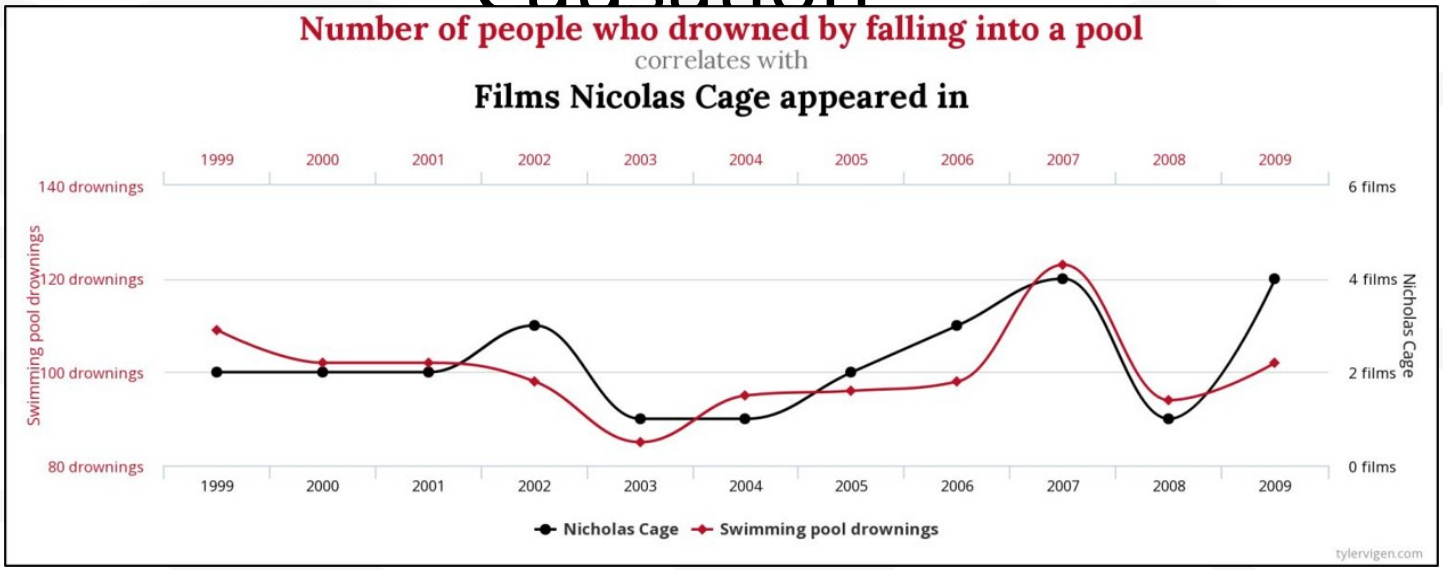
Mistake 4: Not Thinking Past Deployment

Once a model has been successfully deployed, the work is not finished. Models may need to be updated with new data, or they may need to be modified based on shifts in the economy, technology or customer behavior.

Solution: Think about the long term – how will you ensure that your model continues to be useful after it is deployed?

Mistake 5: Thinking Correlation = Causation

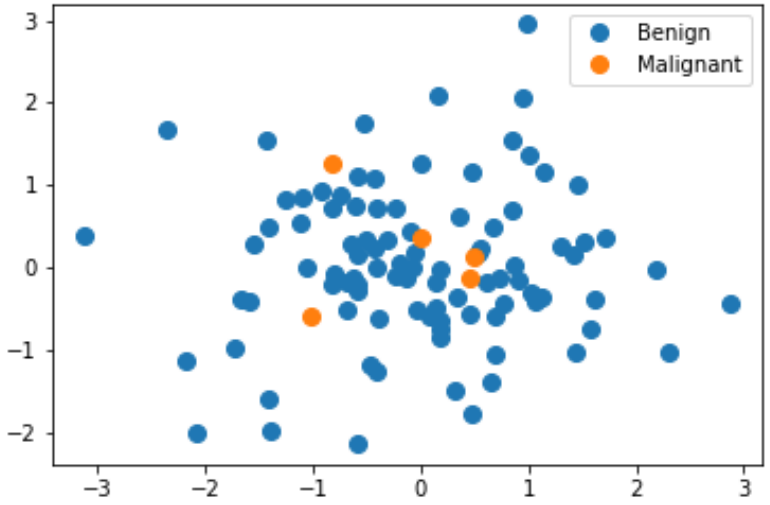
Correlation = 0.6686



Solution: Remember that correlation does not *necessarily* imply causation!

Mistake 6: Not Understanding Limitations of Methods

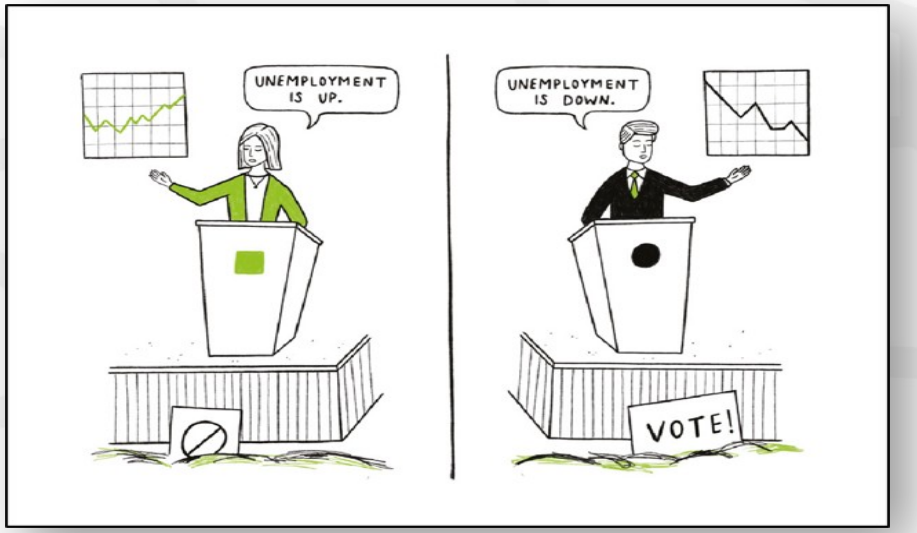
Our model has 90% accuracy in identifying benign tumors!



Solution: Ensure limitations of models are understood and communicated clearly to non-technical stakeholders. When in doubt, ask questions and do research.

Mistake 7: “Cherry Picking” Your Data

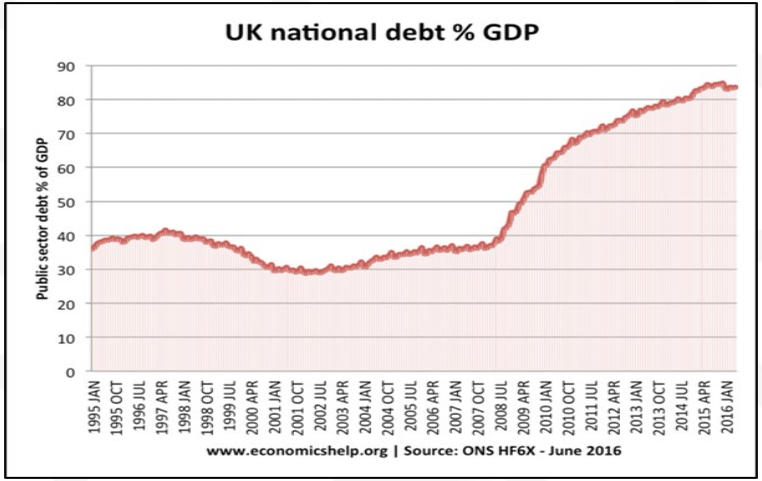
Data can often be used to support multiple and opposing agendas by selecting data or statistics that support that agenda only. This is commonly called “torturing” the data.

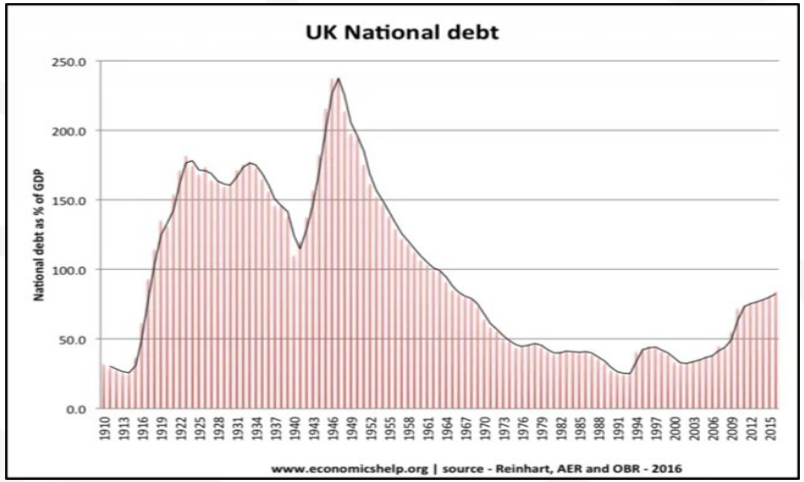


I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
- Sir Arthur Conan Doyle

Mistake 7: “Cherry Picking” Your Data

WARNING – You may be doing this if you find yourself striving to make your data agree with some pre-existing conclusion

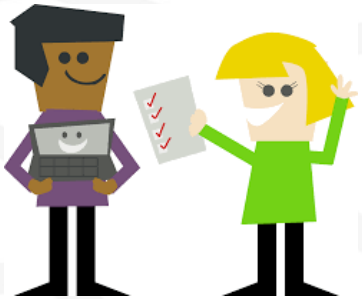




Solution: Do not remove data because it does not support your analysis. Make sure to show the whole picture.

Mistake 8: Not Having Peer Review

Peer review is a great way to ensure that work being produced is of high quality and high integrity. This is especially important for high consequence/high impact projects.

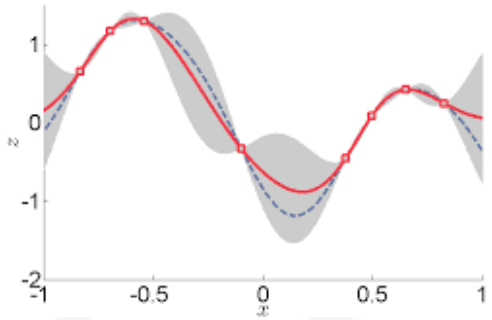


Solution: Encourage your team to set up a peer review process and to support collaboration across data scientists. If you're the only data scientist in your company, ask for peer review from others in the industry or submit your work to peer reviewed journals if possible.

Mistake 9: Ignoring Uncertainties

There is always uncertainty in data science and this is often overlooked by those who are new to the field. Sources of uncertainty may include:

- Measurement uncertainty in your data
- Model form uncertainty
- Sampling uncertainty



Solution: Be aware of uncertainties that could affect the results of your analysis. Look for ways to incorporate uncertainty in your results (e.g., producing a confidence interval on the classification accuracy for your machine learning model).