Unifying Data Science Summary Cheatsheet

A cheatsheet designed to summarize the key concepts and considerations one should bring to the evaluation of different types of data science projects when either reading / commenting on the work of others, or when trying to evaluate one's own work.

Normative v. Positive Questions

Data can only help us answer *positive* questions (questions about how the world *is*, or is likely to be in the future). Data can never fully answer questions about how the world *should be*. It can characterize trade-offs or likely consequences of different actions, but the *desirability* of different outcomes always depend on your value system! So always differentiate between positive statements based on data and normative statements that reflect your value system.

Descriptive Questions

Definition: A project aiming to describe how the world *is*, but not making strong claims about *why* the world looks the way it does (i.e. not making a causal claim).

Purpose: Theory generation / prioritization. Causal inference is great when you have a specific cause you want to test; descriptive analyses help give you a sense of what causes you might want to investigate in detail.

Key Concept: All descriptive analyses require **summarizing** data to extract patterns and make it understandable. But it is incumbent on the researcher to ensure they are doing so faithfully.

Key Questions:

- Does the project provide evidence that the summary statistics provided accurately characterize all relevant patterns in the data (plots; not just means but also medians, skewness; etc.)?
- If dimensionality reduction methods are used, are diagnostics and rationales provided (clustering, PCA indices)?
- Can you imagine a structure to the data that would generate the summary statistics presented but not fit the pattern the researcher describes? If so, what additional summary statistics would rule out those patterns?

Causal Questions

Definition: A project aiming to explain *why* we see the patterns we see by identifying the causal factors giving rise to the patterns we observe in the world.

Purpose: Testing a *specific* causal relationship. Usually motivated by prior descriptive analyses.

Key Concepts for Internal Validity: Correlations imply causation when our treatment and control groups have the same **Potential Outcomes**, meaning that outcomes for both groups would be the same both:

- if they were both treated ($E(Y_{T=1}|D=1) = E(Y_{T=1}|D=0)$), and
- if neither were treated ($E(Y_{T=0}|D=1) = E(Y_{T=0}|D=0)$)

But because potential outcomes are fundamentally unobservable, nothing *in the data* can ever tell us for sure if this condition is satisfied, so we need to think critically about:

1. how the different groups ended up being either treated or not,
2. whether that process may have resulted in the groups being different in a way that is also likely to affect the outcome we care about, and
3. if so, whether our research design can adequately control for those differences.

Also, with causal inference methods we're almost always calculating *Average Treatment Effects*, so remember to think about what variation in treatment effects that averaging may be hiding!

Key Concepts for External Validity: For all causal designs, you should ask: given the design of the experiment and the population of subjects studied, do you think the results of this study can be generalized, and if so, in which contexts would that be appropriate (and in which context would that *not* be appropriate)?

Randomized Experiments / A-B Testing

Key Questions (Internal Validity):

- Was randomization successful (i.e. do treatment and control groups look the same)?
- Did treating the treatment group impact the control group in any way (i.e. was SUTVA violated)?

Regression

Key Questions (Internal Validity):

- What determined whether people were treated or not? Do you think that process is likely to have resulted in the treatment group being different from the control group in ways that are likely to also affect your outcome of interest?
- Do you think the controls included in the regression adequately address the differences in potential outcomes described above?
- Are results sensitive to small changes in our controls (either what we include, or in specification of control), which suggest there are big differences in potential outcomes, and which suggests that results are highly dependent on precisely controlling for confounders (which there's a good chance we we'll never quite be able to do)?

Matching

Key Questions (Internal Validity):

- What determined whether people were treated or not? Do you think that process is likely to have resulted in the treatment group being different from the control group in ways that are likely to also affect your outcome of interest?
- Do you think the controls included in the regression adequately address the differences in potential outcomes described above?
- Is there enough common support that we think we're getting good estimates (i.e. could we get good matches)?
- Remember: matching is the same as regression, just without having to make linear functional form assumptions!

Pre-Post

Key Questions (Internal Validity):

- Do we think that our units had the same potential outcomes in the pre and post periods (i.e. were other things changing between the pre and post periods that would have generated different outcomes absent treatment)?
- If so, do we think our controls adequately address those differences in potential outcomes described above?

Differences-in-Differences

Key Questions (Internal Validity):

- Do our units have the same potential outcomes in terms of their *trends*? In other words:
  - Do the units have parallel trends before treatment, and
  - would they continue to have the same parallel trends after treatment if no treatment had occurred?
- If not, do we think those conditions are met after the inclusion of additional controls?

Prediction

Definition: A project with the goal of making *extrapolations* beyond the data used to build the model.

Key Concepts: As with causal inference, the performance of a model on data we don't use for model fitting and testing is fundamentally unobservable, so nothing *in the data* can ever tell us for sure if our model will extrapolate well.

Prediction from Causal Inference

- Do we think the original study had strong **internal validity** (did it properly estimate a causal effect for the population studied), and
- do we think it is generalizable (has **external validity**) to the context where we want to apply it's findings.

Prediction from Machine Learning

- Machine learning can be sensitive to *context-dependent proxies* – things that are correlated with causal factors in the context we study, but which may not be correlated in other places. So always ask: do we think we're applying our model in a context that is subject to the same dynamics as the place where we got our training/testing data?
- It's almost impossible to know if this is happening with a black box algorithm, so always test to see if you can get similar performance from an interpretable model before jumping to a black box algorithm.
- If your target outcome has bias, your algorithm will have an affirmative incentive to be biased too, so be careful about target selection!