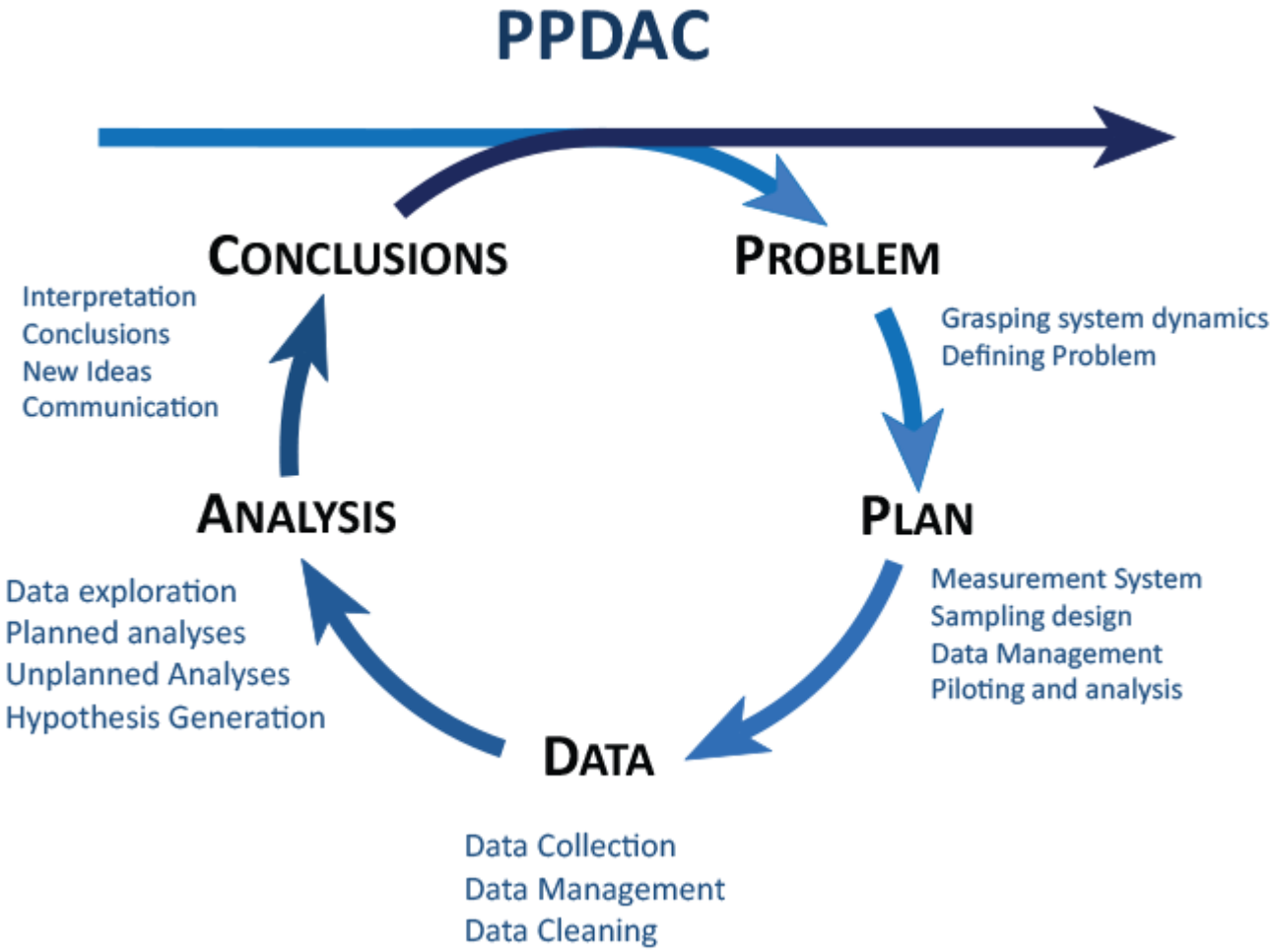
The place of data analysis in problem solving



This model shows the ***process of abstracting and solving a statistical problem to help solve a larger real problem***. A knowledge-based solution to the real problem requires better understanding of how some things work.

A data analyst often doesn't come into the picture until mid-way through the data step of PPDAC. To do a good job, an analyst needs to develop a good understanding of what has gone on before, particularly how the data was actually obtained. (Involving statisticians from the very start of the PPDAC process is much better as it leads to better-quality data that is better suited to answering an investigator's questions).

In the early stages, the problem is often poorly defined. People start with very vague ideas about what the problems are, what they need to understand and why. The **Problem** step is about trying to turn these vague feelings into much more precise goals, some very specific questions that should be able to be answered using data.

The **Plan** step is then about deciding what people/objects/entities to collect data on, what things we should "measure", and how we are going to do all of this.

The **Data** step is about obtaining the data, storing it and "whipping it into shape" (data cleaning). Data analysts are always involved with data cleaning, if only because we almost always discover problems with data during analysis.

The **Analysis** step and the **Conclusions** steps are about making sense of it all and then communicating what has been learned. There is always a back and forth involving doing analysis, tentatively forming conclusions and doing more analysis. The formation of conclusions typically involves the analyst and a subject-matter expert (e.g. someone who understands the business) who will "own" the conclusions.

Often, looking at a set of data will raise more questions than it answers. For this reason we may need to go around this cycle several times before we feel that we've learned what we needed to learn. But with real-world problems, there are always limits to the time and money that can be spent on data collection and analysis.

This article has been about purpose-collected data. We also analyse data that was collected for reasons unrelated to our problem. Such data is much less reliable (for reasons discussed in Week 5) but a whole lot cheaper!

(The PPDAC model was developed by R.J. Mackay and W. Oldford in the early 1990s. Our diagram is an elaboration of PPDAC.)

Common question(s)

How does PPDAC relate to PDCA, DMAIC and similar models used in management?

The PDCA (Plan, Do, Check, Act) cycle often used in management and its descendent DMAIC (Define, Measure, Analyze, Improve and Control) used to drive six-sigma projects are helpful for structured ways of thinking about addressing ***a need to act*** to solve a practical problem such as improving a manufacturing process or an administrative system.

PPDAC, on the other hand, is a structured way of thinking about addressing **a *need to know*** about something by collecting and analysing statistical data. It may need to be invoked many times in coming up with action plans. There are variants in many other fields addressing (sometimes subtly) different objectives.