**Data science vs statistics: some principal components of data science**

In this section we discuss different a…………….s to data analysis and how they relate to **broader trends** in data science. We call these principal components in the sense that they explain variation in the practice of data analysis. Each of the modes discussed below represents a spectrum between two methodologies, one more associated with data science and the other with classical statistics. These modes help explain why different communities seem to **talk past** each other and why some techniques (e.g. computation) have become more popular in recent times.

**Prediction vs. inference (do vs. understand)**

A computer scientist might **pejoratively** describe a linear or logistic regression as shallow and **quaint[[1]](#footnote-1)**. A statistician might be surprised at **the buzz** around deep learning and question why more principled, interpretable statistical models won’t **do the trick**. The point here is that these two imaginary academics are thinking about problems with different goals. The computer scientist is trying to build a system to ac…………sh a task; the statistician is typically trying to learn something about how the world works.

Prediction vs. inference is a spectrum. Many complicated problems have well defined subproblems which are closer to one end or the other end. The d……………..n we are trying to make here is maybe better described as engineering vs. science. Engineering is the business of creating a thing that does something. Science is the business of understanding how something works. Of course engineers use science and scientists use engineering. But if we focus on the end goal of a particular problem we can probably, reasonably c…………y that problem as either science or engineering.

Breiman *et al.* discusses many of the differences between predictive and inferential modelling. Predictive modelling often uses more so…………….d, computationally intensive models. This often comes with a loss in interpretability and general understanding about how the model works and what the data look like. Predictive modelling also places less e……………..s on theory/assumptions because there are fairly good, external **metrics** to tell the analyst how well they are doing.

Predictive modelling is one of the main **drivers** of artificial intelligence (AI). The fact that data can be used to help computers automate things is perhaps one of the most impactful innovations of recent decades. Early **attempts at** AI type applications i……………..d primarily rules based systems which did not use a lot of data; modern AI systems are typically based on deep learning and are extremely data hungry.

**Empirically vs. theoretically driven**

*Data science is exploratory data analysis gone mad*.—Neil Lawrence

Most quantitative fields of study do both theoretical and empirical work, e.g. theoretical vs. experimental physics. Within statistics, we might contrast **exploratory data analysis** (EDA) vs. c…………matory analysis, i.e. searching for hypotheses vs. attempting to c………..m a hypothesis (in this section theory refers to the scientific theory being tested).

It used to be that statistics and science were primarily theory driven. A scientist has a model of the world; they design and conduct an experiment to a……….s this model; then use hypothesis testing to confirm the results of the experiment. With changes ….. data availability and computing, the value of exploratory analysis, data mining, and using data to generate hypotheses has increased dramatically. EDA often prioritizes the ability to rapidly experiment, which means computation can dominate the analysis.

EDA can become problematic when the analyst puts too much f……….. in its results, i.e. when the analyst mistakes hypothesis generation for hypothesis confirmation. Every statistician can list problems with simply mining a data set for correlations (e.g. false discovery, sampling bias, etc). These problems do not mean the results are wrong, but rather they mean exploratory analysis provides much weaker e………….e for a hypothesis than confirmatory analysis. How much this matters depends on the context.

One problematic idea is that EDA can solve every problem. For example, in the controversially titled article, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” it was argued that EDA will replace the scientific method. We disagree. This article is an extreme example of the broader attitude that correlation, and **fancy models** applied to large data sets, can r…………. causal inference and the careful, time intensive scientific method. **The point** that EDA can c……………… to scientific applications **is well taken**, however, and this in fact is becoming more common. These applications likely **raise** many interesting and impactful research questions arising from the increased value in EDA.

**Problem first vs. hammer looking for a nail**

Some researchers take a hammer-looking-for-a-nail approach; the researcher has developed/studied a statistical procedure and then looks for problems where it might be applicable. Other researchers a…… to solve some particular problem from a domain. Note that the f………….r approach is strongly correlated with, but not equivalent to theoretical research.

Both research approaches are valid and productive, however the balance in academic statistics may have shifted too far to the former (hammer) approach. We include this section because data science is focused on problem solving and it is this problem solving which makes data analysis useful to other disciplines.

**The 80/20 rule**

The 80/20 rule of data analysis is:

One of the **under-appreciated** rules in statistical methods development is what I call the 80/20 rule (maybe could even by the 90/10 rule). The basic idea is that the first reasonable thing you can do to a set of data often is 80% of the way to the optimal solution. Everything after that is working on getting the last 20%.

A……….ing basic models to a data set often provides the most value (and/or solves the problem of interest much of the time). The 80/20 rule explains part of why a number of techniques have become more valuable than in the past and why the six areas of GDS[[2]](#footnote-2) emphasize previously undervalued areas. These include: data visualization, exploratory data analysis, data mining, programming, data storage/processing, computation with large datasets and communication.

**Source**: Iain Carmichael, J.S. Marron, “Data Science vs. Statistics: Two Cultures?”, ***Japanese Journal of Statistics and Data Science***, 21 April 2018

1. **quaint -** interesting or attractive with a slightly strange and old-fashioned quality [↑](#footnote-ref-1)
2. **GDS** – In this article we define data science as the union of 6 areas of **greater data science** which are borrowed from David Donoho’s article titled “50 Years of Data Science” (Donoho, 2017).

   1. Data gathering, preparation, and exploration

   2. Data representation and transformation

   3. Computing with data

   4. Data modeling

   5. Data visualization and presentation

   6. Science about data science [↑](#footnote-ref-2)