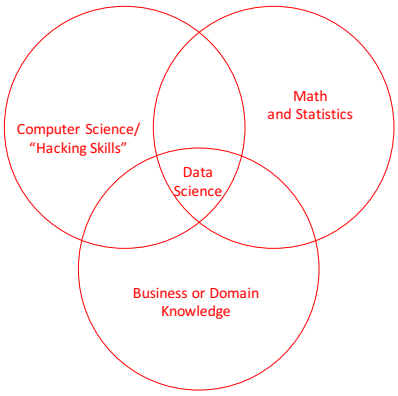
DNSC 6279 – Quiz 1

1. **(4 pts.)** Draw the Data Science Venn diagram:




Order around Venn diagram does not matter, but data science must be in the middle.

(Include at least the three major labels and the location of Data Science in the diagram.)

2. **(1 pt.)** Give an example of, or describe, semi-structured data.

Semi-structured data has a known or dependable structure or patterns, but is not stored in organized tables of rows and columns. Examples of semi-structured data are XML, JSON, and computer logs. (Other reasonable examples are acceptable.)

3. **(1 pt.)** Give an example of, or describe, unstructured data.

Unstructured data is not stored in tables of organized rows and columns. Examples of unstructured data might include raw text, images, videos, and/or sound recordings. (Other reasonable examples are acceptable.)

4. **(1 pt.)** What is the basic difference between supervised learning and unsupervised learning?

Supervised learning takes advantage of the presence of a labeled, known target/y/dependent variable. Unsupervised learning attempts to learn from data without the benefit of a target/y/dependent variable.

5. **(1 pt.)** Give an example of, or describe, deploying a predictive model.

Deploying a model means moving all the logic of the model necessary to make predictions on new data, including data preparation steps, from a development environment, such as a laptop or a workstation, into an operational computer system where it is used to make decisions automatically for an organization. An example of a deployed model is a model used to authorize credit card transactions. (Other reasonable examples are acceptable.)

6. **(1 pt.)** Give an example of a legitimate business decision process in which the use of analytics would be inappropriate or superfluous?

Businesses should not use analytics to make obvious decisions. If a business decision is obvious, the business should just act, not waste time and resources on a data mining project.

Businesses should not use analytics to make decisions in situations where there is little or no past data available.

(Other reasonable examples are acceptable.)

7. **(1 pt.)** Is "big data" always better than having a smaller data set for training a model? Give one reason to support your answer.

No, bigger data is not always better.

Sampling is a scientifically justifiable and well-understood mathematical operation designed explicitly for making big data easier to use without losing important information.

Sometimes sampling or resampling (bagging, boosting etc.) can even lead to better models.

Big data can contain a lot of redundant, repeated information.

Big data might contain very little pertinent information about a specific topic of interest.

Often a designed experiment than collects a smaller sample of relevant data contains better information than a massive set of "data exhaust" – data collected as a by-product of some other organizational process.

(Other reasonable explanations are acceptable.)