Student Name: Han Weiyu
Matriculation Number: G2301835L

Story-telling with Graphics and Visualizations (BS6203)
Assignment 1

---

1. **Task 1**
   Go to Pinterest, and review some data science or machine learning or programming infographics (e.g. https://www.pinterest.com/kourouklides/data-science/?autologin=true or https://www.pinterest.com/gohwils/data-science/)

   Find one (or a few) that you really like (for presentation and content)

   Try to list down why you enjoyed it.

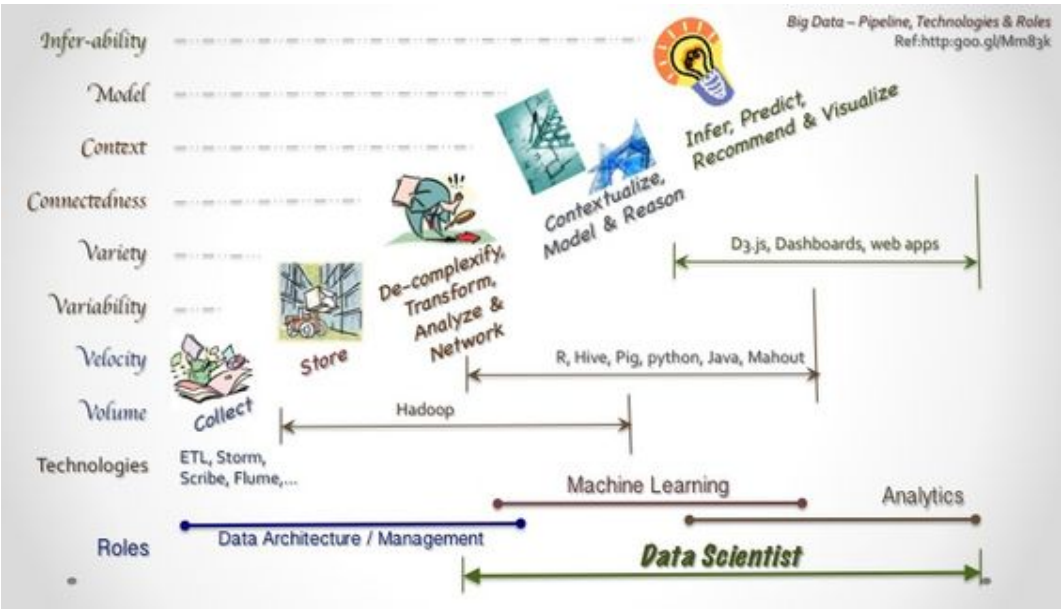

Figure 1: The favorite diagram

The key advantages of the graphic can be listed as follows: simple but much-information, using different color to show the difference, tightly coupled logic.

First, the graphic is simple on the first sight and it only contains two axes and some points. However, it conveys much information to the viewers. Specifically, the X axis shows both the roles and technologies. Besides, it also conveys the encompassing relationship between them. For example, the length of the green, red, brown and blue lines is a good illustration of data scientists often make effort on machine learning and analytics, rather than data architecture and management. The graphic, though simple, but visually conveys these messages to us.

Second, the graphic uses different colors to convey different messages. Within the same discussing class, different color are uses to help viewers to identify the difference of contents, which is often

helpful to comprehension.

Third, the logic connection are tightly coupled in the graphic. The five points in the graphic correspond well to the x and y axes, signalling the discipline to which the application belongs, the tools used and the biggest challenges.

To conclude, the graphic illustrates the enough information but simple enough, which is helpful to clear the logic and comprehesion. And that is the reason why I like the graph most.

2. **Task 2**

The UPR (Unfolded Protein Response) is a cellular stress response that is activated by an accumulation of unfolded or misfolded proteins in the lumen of the endoplasmic reticulum. In this scenario, the UPR has three aims: initially to restore normal function of the cell by halting protein translation, degrading misfolded proteins, and activating the signalling pathways that lead to increasing the production of molecular chaperones involved in protein folding. If these objectives are not achieved within a certain time span or the disruption is prolonged, the UPR aims towards apoptosis. There are 3 pathways that feed into a mechanism known as the UPR. This requires turning on any of 3 potential paths via IRE1, PERK and ATF6. Let us also assume that when turning on these 3 paths, all downstream targets are also all turned on, and there is no suppression.

A student performed a series of knock outs. And for each knock out, measured which genes are still inducible to create the UPR. The results are shown in the Venn diagram here.

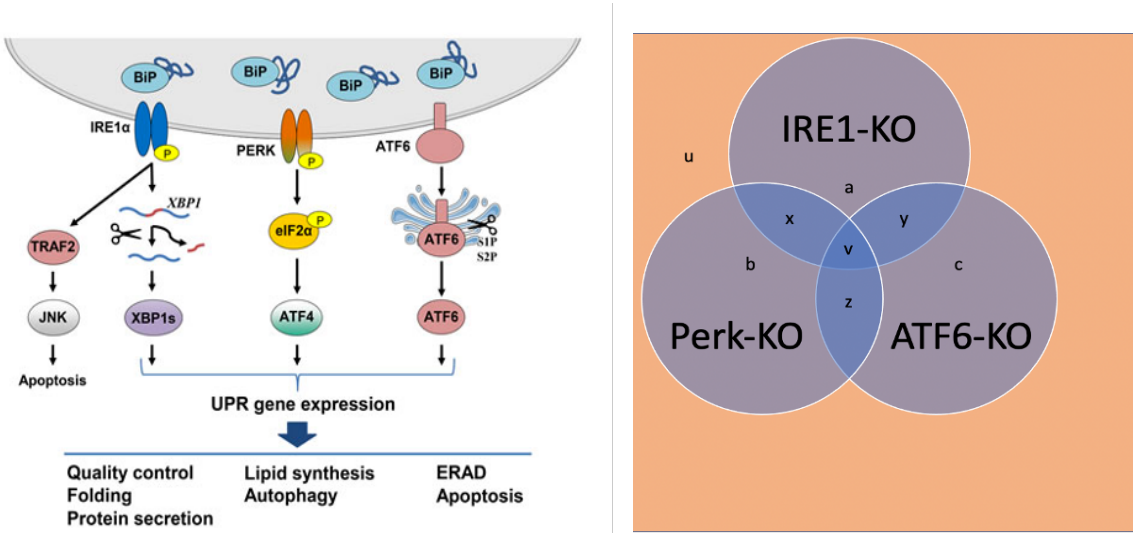

Figure 2: Genes involved and Venn diagram

Some assumptions are made to describe the result better. "a", "b" and "c" does not contain any overlapping shading. "u" is the outside part of the three circle. "x" is the overlapping of IRE1-KO

and Perk-KO. And so on for "y" and "z". "v" represents all the three circles' overlapping area.

(a) **"Does "a" show the set of genes downstream of IRE1? If not, what is it?"**

No, "a" shows the genes that still works when knocking out IRE1. In other words, "a" is not the downstream of IRE1 but the downstream of Perk and ATF6. (If there is not any more paths exist in this mechanism.)

(b) **"What is the expected value of "v" if only 3 paths exist?"**

The expected value of "v" is **0** if no more paths exist. "v" represents the genes still inducible to create the UPR when knocking down IRE1, Perk and ATF6. There will be no genes still inducible to create the UPR since all the genes are downstreams of the IRE1, Perk and ATF6, if they are the only 3 paths exist.

(c) **"What should you expect the value of v to be if more than 3 paths exist?"**

The expected value of "v" is **no less than 0**. There are two situations if different assumptions are made. First assumption is all the other paths are related with the 3 paths mentioned. The knocking down of the 3 path will affect the other paths and finally the expected value of "v" is then **0**. Another assumption is there is at least one path that is independent with the 3 paths. In this situation, the expected value of "v" will be **more than 0**.

(d) **"Do you expect "b" and "c" to be 0?"**

Actually, the "b" and "c" cannot be 0 at the same time if there are no more paths and all 3 paths contribute to the mechanism. "b" represents all the downstream genes that only works when IRE1-KO and ATF6 both exist. "b" and "c" are both 0 means "z" is 0 as well. In other words, IRE1 does no contribution to this process, which contradicts the assupmtion that all 3 paths contribute to the mechanism.

(e) **"Is "x" the common set of genes shared between IRE1 and Perk?"**

"x" is the overlapping part of knocking down IRE1 set and knocking down Perk set. In other words, the genes are not included in both IRE1 and Perk. And it is more likely to be the downstream gene of ATF6.

(f) **"What is the expected value of "u"? Is it possible to even get "u"?"**

The expected value of "u" relies on the definition of the whole set and the information we know about this set. If assuming that the whole set is the genes observed, the "u" means the genes is not related with the UPR mechanism. Whether the student knock down either of the 3 paths or not, the "u" genes are not inducible to create the UPR. It is possible to get "u" if the whole set could be observed.

(g) **"Are there situations where regions a b and c are non-empty?"**

Yes, "a" represents the genes that only works when Perk and ATF both exist. "b" and "c" are as so on. "a", "b" and "c" are non-empty indicates all 3 paths are not definitely independent with each other. Some genes can work only if two of them exists.

(h) **"Does v correspond to triple knock out?"**

No, "v" are those genes which are still inducible to create the UPR even after all the 3 paths are knocked out. In other words, if "v" is not 0, it is certain to be other path can create the UPR.