We hear lots of requests on porting R to Clinical Computing Platform. Before R really touches the ground, I’d like to share some thoughts on why R is a not a good choice, considering there is a strong alternative, Python.

In this paper, I will compare the strength and weakness of R and Python, on various tasks like data manipulation, statistical analysis, system integration. Although R is well known in our industry and has some good things, I argue that Python is a better fit:

1. Python is a better designed programming language;

2. For various programming tasks, Python offers more consistent and unified packages stack, while in R, the packages are scattered;

3. Python is better at system integration due to its reputation as glue language;

4. Considering its elegant syntax and unified ecosystem, Python is easier to learn for SAS programmers.

Jiangtang('JT') Hu is a life sciences consultant at d-Wise of Raleigh, North Carolina. Before d-Wise, JT worked for Sanofi and SAS.

JT got a master degree on software engineering, and has wide interests on SAS, Python, R and machine learning. JT blogs at <http://jiangtanghu.com/blog>.

Machine learning, especially deep learning, is very successful in fields like computer vision. In the broader medical field, a very first successful application is medical image diagnosing, which is very close to general image classification. But on the clinical research side, the power of machine learning has not yet been realized.

Within clinical research, we handcraft lots of rules for various tasks. Considering the large number of manual rules and associated code we develop, it’s the perfect situation to think about the possibility of using machine learning techniques for the tasks without explicit programming. So, how can we get started?

As an example, I will demonstrate on how to use machine learning to perform clinical data de-identification using a deep learning technique called Generative Adversarial Nets(GAN). This technique can generate data that looks like the original data and the generated data could serve as de-identified data for public uses.