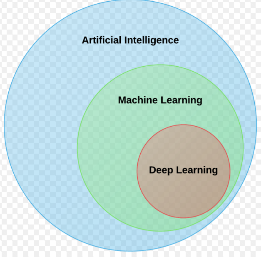
**The Marriage of Machine Learning and Clinical Research: A Proposal**

**Abstract/ Instruction:**

The definitions we need to get the discussion started:

Artificial Intelligence (AI) is kind of intelligence learned by machines. Machine Learning is a technique within AI to give computers the ability to learn without explicit programming and handcraft rules. Deep Learning is a special technique of machine learning which utilizes neutral networks providing the most powerful machine learning technique in real life applications.



Machine learning, especially deep learning, is very successful in fields like computer vision. In the broader medical field, a very first successful application is within medical image diagnosing, which is very close to general image classification in computer vision. 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.

1. What’s the problem?

To support the need for data transparency and to ensure patient privacy, all clinical data required for public use should be anonymized.

1. What’s the current approach to this problem?

Develop and write code for handcrafted rules. For example, we might shuffle the values of the variable “sex”

1. Can we translate it to machine learning problem?

A deep learning technique called Generative Adversarial Nets(GAN) is widely used for image replication. For example, with real dog images as inputs, GAN can learn though the images and then create images which look like dogs. Here is the idea to apply it to the de-identification problem:

We have original dataset (name it A) to be anonymized. At very beginning, we use GAN to generate fake dataset B, the B is sent to A for checking. If B doesn’t look like A, the model adjusts its parameters and creates a new B which will sent to A for check again and the process goes on, until B does look like A. So, the final B could serve as the final de-identified data for public use.

1. Do we have data?

Yes in this field, there are lots of data!