# Introduction (preview)

# *raw structure - add more details and references*

## Introduction of machine learning (ML)

**Brief introduction of ML.** ML have already transferred the traditional data mining like linear regression or covariance analysis into a new area. The application of ML is introduced in many industries including web search, speech recognition, natural language processing (NLP), image classification (computer vision) and web search advertising (product recommendations).

**Why ML perform better than traditional statistical methods in certain (which) areas?**

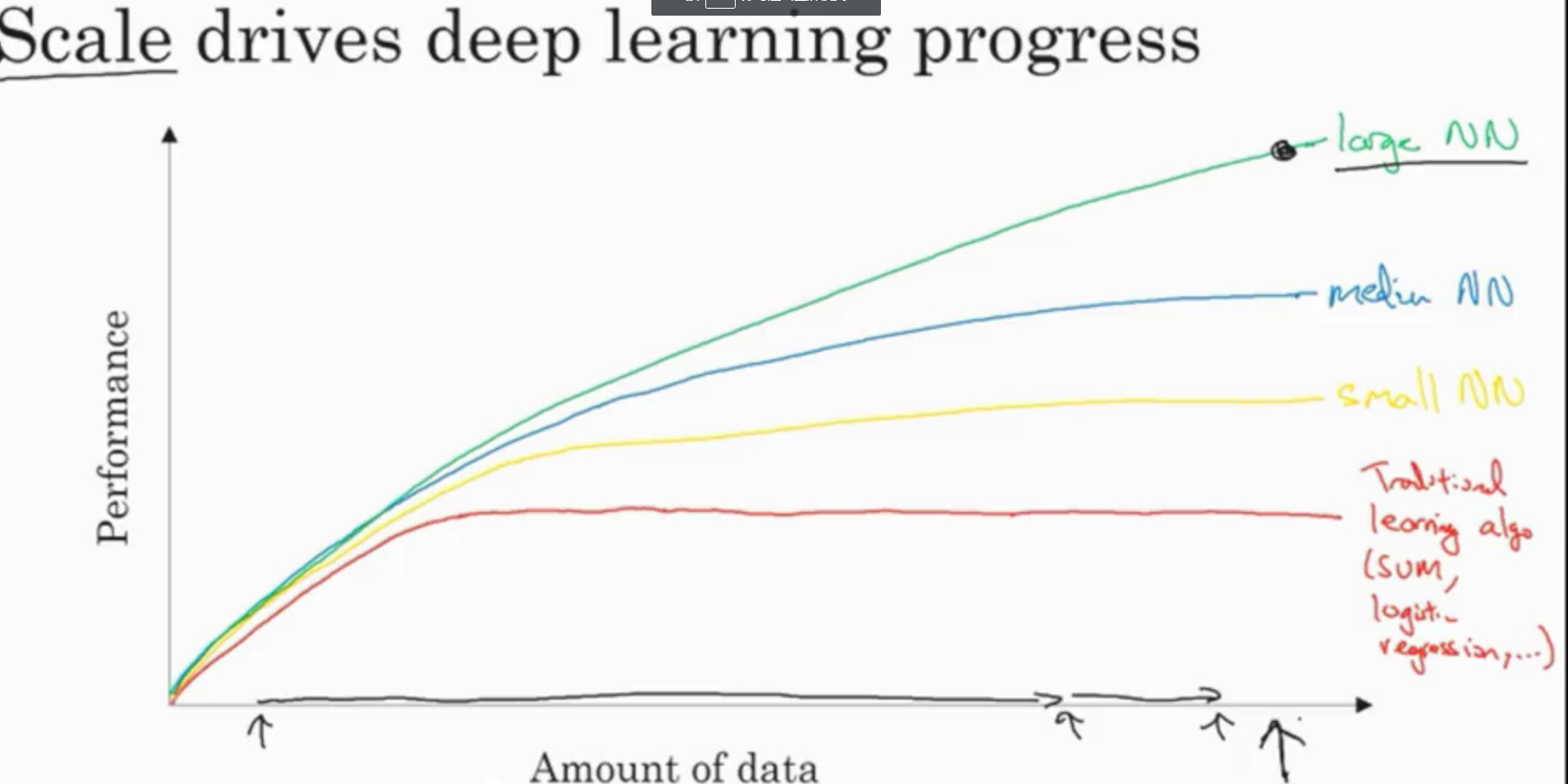
**Few types of ML. Supervised and unsupervised ML. NN as supervised ML.**

**In the field of Biology,** advanced methods have been developed in mining clinical data, sorting health information. Especially for image recognition, computer are programmed to help identifying experimental data. For instance, scientists used ML to recognize the condition of experimental subjects (e.g. flies or cells). In genetics, however, the large amount of sequencing data enable ML to predict disease factors either by inputting sequences or probe-set.

## Introduction of deep learning (DP) and neural network (NN)

**Brief introduction of DP.** Deep learning is the rapid developing ML algorithm in the past decades. How when? NN is one of the key algorithm in DP.

**As the development of computational technology, the size of data is growing rapidly.** Now we have more device to store and accumulate mega-data and we have the equivalent computational ability (improvement of the performance of CPU and GPU). Consequently we can have the output of the model much faster and then we can optimize our NN model in term of shorter time nowadays. On the other hand, a traditional learning algorithm like support vector machine or logistic regression may have a lower accuracy if the input data is large.



**What is NN?** NN can then identify the pattern in data by generating many “neurons” in the models and neurons represent the complex connection of the features behind the hidden layers (each layer have a set of neurons representing an predictors of the weight of each feature in one iteration). Cost function, forward propogation and backward propogation and gradient descent and layers. The process is similar to recognizing faces by distinguishing different noses, eyes or faces identities in samples first as human brain does. NN are remarkably good at figuring out functions that accurately map from input to output.

**Why NN solve curve regression instead of linear regression?** Unlike simple linear correlation input and output prediction, the connection of input and output in NN is more complicated and it is usually not linear (activation function: ReLU, leaky ReLU, tanh or sigmoid or)? For data with many features, NN can analyze the data by attaching the features (e.g. size, solubility and thermostability of chemicals) into a vector and stack them into a matrix.

**Types of DP.**

## DP in synthetic Biology

Most importantly and fundamentally, introduction. Types: Convolutional networks or convolutional neural networks(CNN) are often applied to images. WHY? recurrent neural networks (RNN) and long short term memory (LSTM) are normal sequential models. WHY?

Here we would like to design, build and fit a new structure of NN for pipeline in synthetic biological. Not just tuning hyper-parameters, regularizing data and train a normal model, a new strategy and structure of the model need to be optimized.

Introduction of the existed DP in synthetic Biology.

## DP in synthetic Biology

Hyperparameter tuning, regularization, how to diagnose key variants and preprocessing data

Implement momentum armrest prop and the ad authorization algorithm.

Switch your data into train, development or dev also called holdout cross-validation sets and test sets, has changed in the era of deep learning.

**RDKIT**

**SMILES and ECFP4**

Most importantly and fundamentally, introduction. Types: Convolutional networks or convolutional neural networks(CNN) are often applied to images. WHY? recurrent neural networks (RNN) and long short term memory (LSTM) are normal sequential models. WHY?

CNN in SMILES(Duvenaud et al., 2015).

常用的二维描述符采用基于指纹方法生成的拓扑描述符，如 MDL Public Keys、SciTegic ECFP(extended connectivity fingerprints)等。相似度的计算方法也有很多。最常见的是 T a n i m o t o 系数，其计算公式为：S T = C/(A+B-C)，A 和 B 分别为化合物 A 和 B 中所定义特

征结构的数目，C 为两个化合物中共有的特征结构的数目。图 1 中的范例展示了多种分子描述符的生成方法[8]。

Here we would like to design, build and fit a new structure of NN for pipeline in synthetic biological. Not just tuning hyper-parameters, regularizing data and train a normal model, a new strategy and structure of the model need to be optimized.

Introduction of the existed DP in synthetic Biology.

### Methods

**Data Source**

Our pipeline takes as input the SMILES [30] string encoding of each molecule, which is then converted into a graph using RDKit [20]. We also used RDKit to produce the extended circular fingerprints used in the baseline. Hydrogen atoms were treated implicitly.

Visualization: Structure Diagram Generation algorithms.

In our convolutional networks, the initial atom and bond features were chosen to be similar to those used by ECFP: Initial atom features concatenated a one-hot encoding of the atom’s element, its degree, the number of attached hydrogen atoms, and the implicit valence, and an aromaticity indicator.

The bond features were a concatenation of whether the bond type was single, double, triple,

or aromatic, whether the bond was conjugated, and whether the bond was part of a ring.**Data preprocessing**

Cross-validation – 100% training set and 30% testing set. Will this lead to the over-fitting?

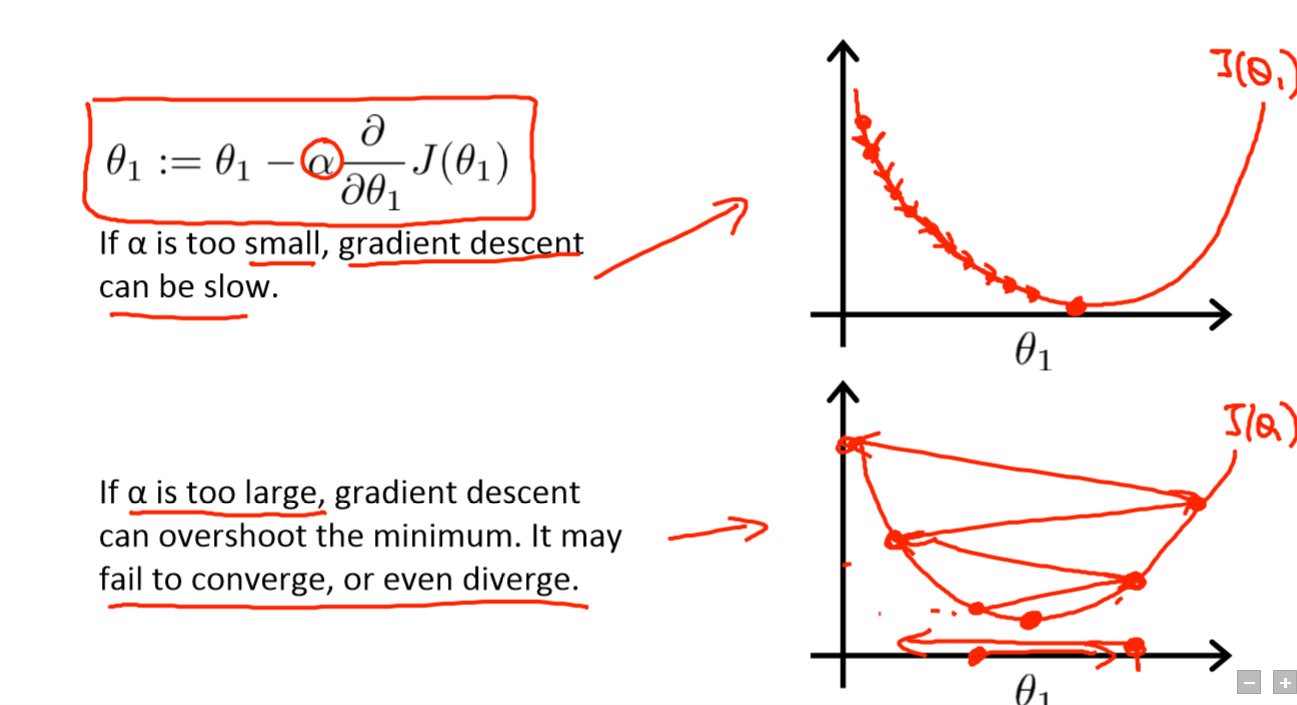
**Hyper-parameters**

W0, b0 and learning rate alpha (number of iteration), hidden layer l, choice of activation function and number of neurons.

Empirical process. As the development of the computer infrastructure, the hyper-parameters need to be tested as CPU or GPU updated and so on by maybe by cross-validation.

Cost function – learning rate plot. Learning rate alpha accelerate our learning and let the cost function converge on a stable value.

The learning rate is one of the most important hyper-parameters to tune for training deep neural networks. It decide how far to move the weights in the direction of the gradient for a mini-batch. If the learning rate is low it will slow down the descent process. On the contrary, if the learning rate is high, then training may not converge or even diverge.

Leslie N. Smith describes a powerful technique to select a range of learning rates for a neural network in section 3.3 of the 2015 paper “Cyclical Learning Rates for Training Neural Networks” . We need to select a point on the graph with the fastest decrease in the loss.

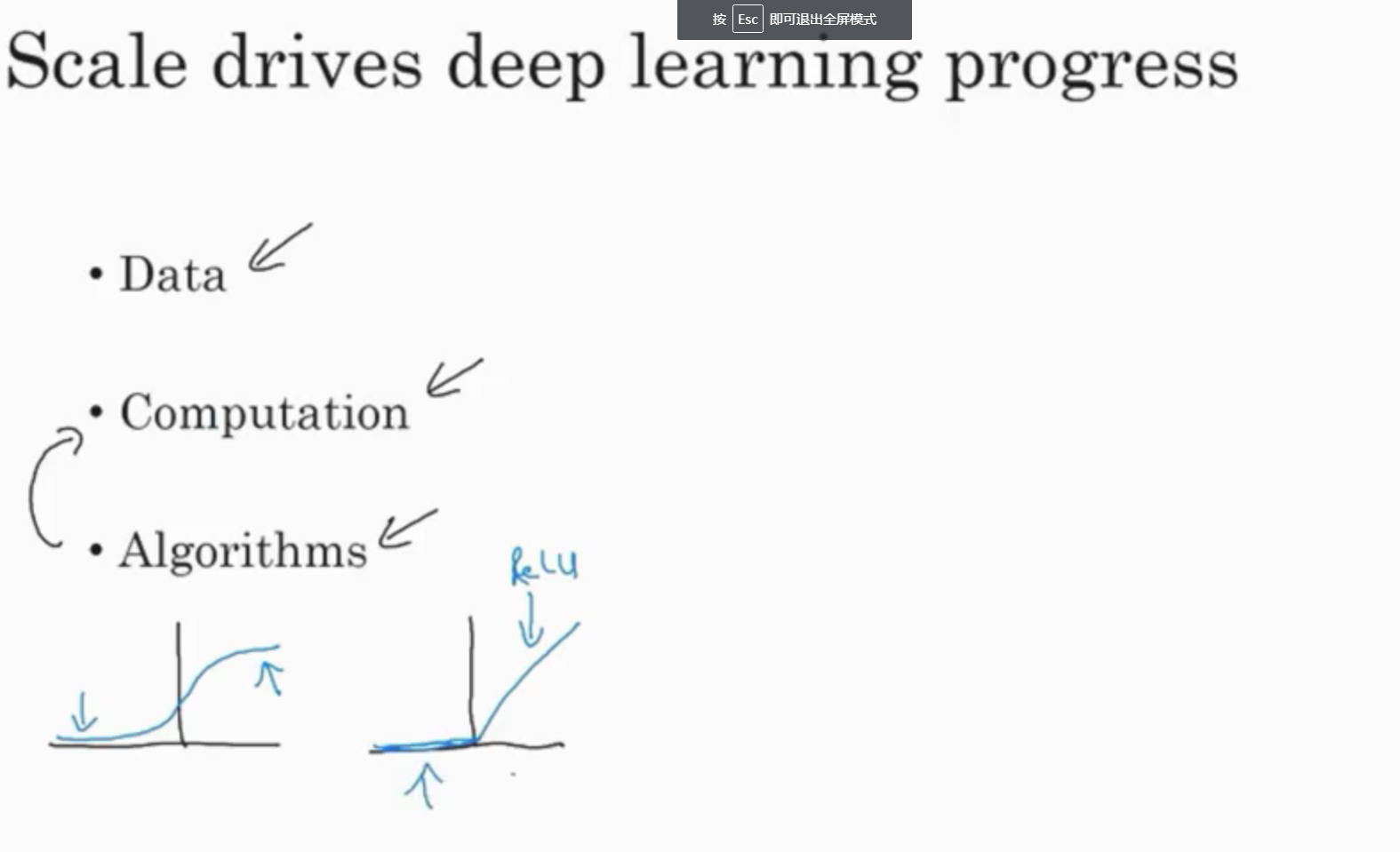
Jeremy Howard and his team at USF Data Institute developed fast.ai, a deep learning library that is a high-level abstraction on top of PyTorch. It’s an easy to use and yet powerful toolset for training state of the art deep learning models. Jeremy uses the library in the latest version of the Deep Learning course (fast.ai).

The library provides an implementation of the learning rate finder. You need just two lines of code to plot the loss over learning rates for your model:

**Activation function**

WHY ReLU and sigmoid?. ReLU rectify the output by taking a zero if the input is lower than 0.

The derivative of sigmoid function becomes relatively small if the independent variable goes up (in this case, the hypothesis function). In a NN models with many layers, using sigmoid as an activation function will cause a very slow rate in training the model. Is the input features need to be centerlized (larger than 0 or not). If so we need a tanh. So we switching to ReLU as our activation function in the hidden layers to make gradient descent works faster (the principle of Bayesian). In the output layer we used sigmoid as activation function simply because sigmoid restrict the output range from 0 and 1, and it fit the output as a binary classification (labeled the data as 0 and 1 indicating significant and non-significant in prediction).



### Conclusion and innovation

Scientists are still working on developing new algorithm to make DP work even better. There are more and more digital data in synthetic biology throwing in the data mining subject. The hardware is developing really fast allowing us to build even lager neural network. In the future, there will concretely be new algorithm of larger and more complicated neural network in processing data in synthetic biology.

The DP model now takes us a lot of time to test the best hyper-parameters. As the improment of the algorithm of DP, we are looking forward to automatically-tuned DP algorithm.

Duvenaud, D.K.*, et al.* Convolutional networks on graphs for learning molecular fingerprints. In, *Advances in neural information processing systems*. 2015. p. 2224-2232.