

---

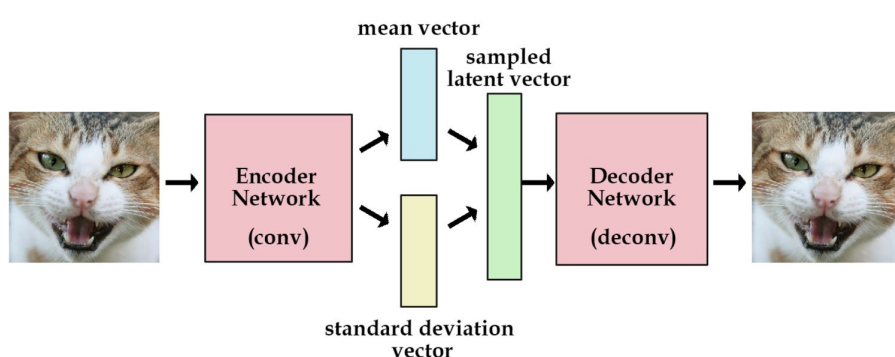
J  r  mie KALFON, UCL, Computational Neuroscience

## Academics

I am a Biomedical Engineer in Computer Science & Electronics (*resume*), thanks to my sundry training, I have gained a knowledge about many different topics (*courses\_list*). However, French engineering schools do not focus on academic tasks and it is -as I have explained it in my (*motivation letter*)- why I have chosen a second master at the University of Kent in Computational Intelligence. I want to pursue my research career in this direction, focusing on data science and learning techniques. More specifically, what has been capturing my attention, this last couple of years, is artificial neural networks and their forever more diversification, from new RNNs architectures (*Graves et al.*) to GANs (*Goodfellow et al.*) and VAEs. It has never ceased to retain my attention in a number of ways.

First, through their steadily strengthening links to theoretical neuroscience (*Hassabis et al.*) which is a subject I always liked getting to know about. It is shown by my involvement in CaImAn<sup>1</sup> (*Giovannucci et al. -in review-*) and participation to HBP workshops.

Second, because of their vast range of applications, from computer graphics to medical screenings to neuroscience itself, and many more. Such applications that are increasingly popular in domains where vast amounts of data are produced. Their recent utilisation (*Alipanahi et al.*) on genomics data will no doubt take on a tremendous importance in the years to come.



Simple architecture of a VAE property of Kevin Frans

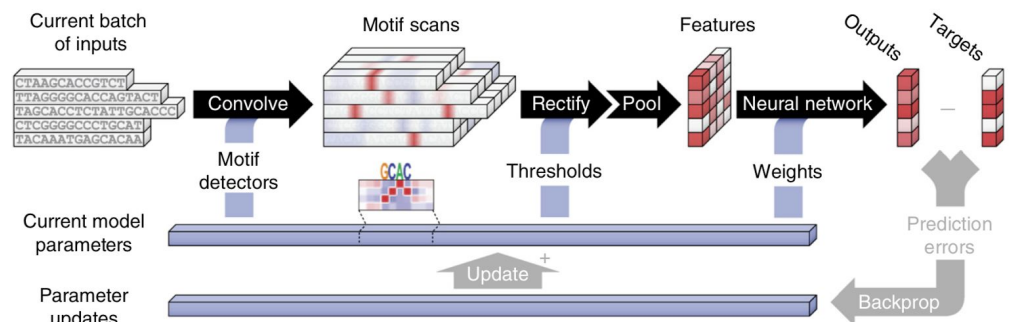
Similarly to many cases where relevant output is sparse and acquiring it demands a lot of effort, learning and treatment of data has to be as unsupervised -Principal Component Analysis, Independent Component Analysis (*Hyv  rinen & Morioka NIPS16*) and Variational AutoEncoders- as possible. Focusing on finding patterns.

1. It uses gaussian kernels and a constrained matrix factorization (using Autoregressive processes, Constrained deconvolution with sparsity control, noise estimation and overlaps management) to divide the input image into a spatial location and a temporal activity of the neurons from a deconvolution of their calcium dynamics.

## Objectives

My master thesis has recently been defined as a study, in collaboration with Tobias Von Der Haar and Dominique Chu about the analysis of RNA codons using unsupervised learning techniques to try to decipher the problem of “codon usage bias” which shows that, triplets encoding the same amino acid are not present in equal proportion in the genome.

This task which should result into a joint publication seems to represent a great introduction to my research objective for the next years, which will consist in applying unsupervised methods to genomic data such as to extract information from this inherently opaque source. Allowing to bridge the gap between genotype (information contained in our cells) and phenotype (the result of the expression of the genotype).



*Architecture of a neural network to extract information from codons. It is however done in a supervised fashion. published by Babak Alipanahi Nature biotechnology All rights reserved.*

I am well aware that such an objective can be the one of a decade. However, It is believable that the integration of data science in life sciences and systems biology is only in its infancy and that the open source availability of data (NAR's list), the rapid increase of its production through personalised medicine, NextGenerationSequencing as well as its direct application into domains that receive incredible amount of resources makes it a great direction to take for data science and machine learning.

## Plan

- A first part of this project is to learn more about unsupervised learning methods -going to UCL Gatsby seems to be an efficient way of reaching it-
  - Then it is essential to have a detailed categorized map of the features and potential of various feature extracting methods.
  - As for codon usage bias, a lot of unanswered questions can be at least partially solved or better asked using such methods. It is thus necessary to create a thoughtful map of such questions, their possible answers and the type of data required to answer them.
  - In addition to a comprehensive knowledge of the open databases and tools readily available on the web (QBF) it is then trivial to select the problems and produce an architecture at least able to compute some relevant information.

- A second part of the plan is to refine the model.
  - There are many parallels made between gene feature extraction models and language processing methods (*DanQ*, *gene2vec*).
  - However, genes are influenced also by protein bindings, the environment and more. They can change entire meanings, roles and influences of groups of genes.
  - The machine learning community starts appropriating itself the realm of the genome. Nonetheless, ideas are mostly reused from other concepts. Yet the genome is a system whose purpose -in addition to encoding information- is to do basic forms of computation (*Adleman*). Moreover the search space can be highly constrained since the results are dictated by a known set of chemical relations, thus driving the need for different approaches.
  
- Lastly it is sure that an incremental approach will help having a worthy project rapidly in time and to adapt new decisions and opportunities to change.
  - I have been inspired by some great machine learning open source projects such as CalmAn of course, but also Scikit-learn and Pytorch.

### Addition :

1. It goes without saying that such a task needs to emphasize on interdisciplinary collaboration and exchange of knowledge, as well as on producing a project which is sufficiently usable (UI, API, documentation) and versatile (simplicity, scalability, modularity) as to be reused by other groups.
  
  2. While trying to be as precise as possible, this first research plan tries to demonstrate an ability to plan, to understand some valuable concepts and a knowledge of the discipline. A Ph.D. is also a process of refinement of the plan while gaining a deeper knowledge of the subject. I really wish for Gatsby to be the place for me to explore, refine, and accomplish this plan.
-