**Bioinformatics Intern, The Center for Computational Biology at JHU** Baltimore, MD |Sep 2018 - Dec 2018

*Explored significant co-occurring or mutually exclusive mutated driver genes across cancer types*

* Extracted The Cancer Genome Atlas (TCGA) data using Bash
* Performed RNA-seq using R and shell scripting
* Applied fisher’s exact test, chi-square test and permutation test to identify ~4000 significant (p-value < 0.05) co-occurring or mutually exclusive mutated driver gene pairs
* Used pandas, numpy, matplotlib, scipy, scikit-learn, networkx in Python to visualize significant gene pairs network after Bonferroni correction
* Optimized ridge regression model using cross validation to predict the phenotype with 80.77% accuracy

你如果听了我之前关于data简历上的项目讲解的话，可以发现我一直在强调pipeline的问题，那么在做一个data相关的项目的时候，1. 先要告诉你的读者(可能是hr/recruiter/interviewer)，你为什么要做这个项目，也就是motivation，use case或者要达到的目的和项目功能。2. 说明data的获取来源，比如爬虫获得还是怎样，当然了，从raw data开始是工业界的人最希望看到的。如果实在data相对不是很complex，那么也可以先写一下如何进行的预处理(pre-processing)，包括怎么做的extraction，cleaning等等。3. 预处理之后可以想想怎样存取数据或者进行realtime的data互动，比如存到数据库中？那么这个过程包含数据库的设计。4. 开始使用数学模型来完成你的目的，比如用一些ml的常用算法，然后基于常用模型的优化，调参，比对各个模型的优势等等；5. 测试，是如何测试的最后结果？测试的结果怎么样？为什么你做的东西有价值？这个环节能量化就量化。能量化的结果最有说服力。6. 迭代整个pipeline，再不停优化。要说明优化办法。

数据科学简历修改指导：https://youtu.be/XuYUVlvzZPo