# Data Scientist: Python Test

A common task in data science is to pull data via a RESTful API and parse the output accordingly.

For this test, we ask you to write a Python script to access and process json content from MyGene.info for a given list of genes. Following this you will be asked to pull in associated publication information via the Entrez E-utilities API.

Before starting the test, familiarise yourself with the MyGene.info REST API here:

http://mygene.info/v3/api

You may also wish to refer to the documentation here: <http://docs.mygene.info/en/latest/doc/query_service.html#query-syntax>

and here:

<https://dataguide.nlm.nih.gov/eutilities/utilities.html>

Please send your solution as a zip file, including a readme with details of any dependencies.

# Task 1:

* 1. From the MyGeneInfo API, use the “Gene query service" GET method to return details on the following GENE symbols, filtered for species, “human": CDK2, FGFR1, SLC6A4
  2. From the returned json, parse out the “name", “symbol" and “entrezgene" values and print to screen

# Task 2:

2.1) Using the appropriate identifier from the above result, send a query to the MyGeneInfo “Gene annotation services" method for each gene

2.2) From the resulting json, collate up to 5 generif descriptions per gene

2.3) Write the results to an Excel spreadsheet with columns: gene\_symbol, gene\_name, entrez\_id, generifs

# Task 3:

Use the Pubmed IDs associated with the above generif content to extract additional bibliographic information.

e.g.

<https://dataguide.nlm.nih.gov/eutilities/utilities.html#esummary>

Hint: from Bio import Entrez