> DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY
>
> — HARVARD BUSINESS REVIEW

CHALLENGE

Warning: We suggest you use Chrome(https://www.google.com/chrome/browser/desktop/index.html) as your browser (possibly using Incognito Mode) if you experience any errors.

Please answer as many questions as you can. We do not expect you to answer all the questions (they are mostly optional) but answering more optional questions correctly will help you and answering them incorrectly will not hurt you. **Please give all numerical answers to 10 digits of precision. Partial credit will be given to answers that agree to less than 10 digits.** You can resubmit your answers on this form as often as you would like. We keep track of your latest submissions and they are populated here. (*) denotes a required field. The basic ground rules and suggestions are:

- **Answer the questions yourself without asking others for assistance**. This is a test of your ability to answer realistic questions. You will be asked questions of similar difficulty during the phone interview so cheating will not help.
- **Do not share the questions or your answers with anyone.** This includes posting your questions or solutions on services like quora, stackoverflow, or github. Doing so gives others an unfair advantage and may also disqualify you from this or future fellowships.
- **Submit early.** We highly recommend aiming to submit the answers well ahead of the deadline as "unforeseeable" technical difficulties have prohibited highly-qualified last-minute applicants from submitting and continuing previously.
- **Submit often.** You can submit your challenge solutions as often as you would like. Only the last submitted challenge is kept so we recommend you submit your answers as you complete them.

A few helpful hints:

1. **Want to get a head start on being a data scientist?** We want all semifinalists to get as much out of the challenge questions as possible. So we've written three(http://blog.thedataincubator.com/2015/09/painlessly-deploying-data-apps-with-bokeh-flask-and-heroku/) blog(http://blog.thedataincubator.com/2015/01/processing-data-like-a-professional-data-scientist/) posts(http://blog.thedataincubator.com/2015/01/a-cs-degree-for-data-science-part-i-efficient-numerical-computation/) that might get you thinking about mathematics and computation differently. They will also give you a head start on solving the challenge questions. For additional hints on the challenge, follow us on Twitter(http://twitter.com/intent/user?screen_name=thedatainc), LinkedIn(https://www.linkedin.com/company/the-data-incubator), and Facebook(https://www.facebook.com/dataincubator/).
2. **Having browser troubles?** We recommend using Chrome(https://www.google.com/chrome/browser/desktop/index.html) (possibly using Incognito Mode).
3. **Having trouble downloading any files?** We suggest using command-line tools, rather than relying on a browser.
4. **Want to avoid being a statistic?** Every application cycle, a number of applicants wait until the last minute to submit, only to discover "unforeseeable" last-minute glitches that prevent submission. We suggest not waiting until the deadline to submit.
5. **Found something ambiguous?** We realize some questions are ambiguous. Most real-world questions are. This is a test of whether you can prioritize important effects and combine real-world knowledge with theory.
6. Due to the volume of requests, we will only accept submissions via this form.

Q1: A knight in standard international chess is sitting on a board as follows

0	1	2	3
4	5	6	7
8	9	10	11
12	13	14	15

The knight starts on square "0" and makes T jumps to other squares according to the allowable moves in Chess (so that at each space, it has between two to four valid moves). The knight chooses amongst the allowable moves at each jump uniformly at random and keeps track of the running sum S of keys on which it lands. See below for specific questions and answers.

After $T = 16$ moves, what is the mean of the quantity S modulo 13?

1.234567890

What is the standard deviation?

0.9876543210

After $T = 512$ moves, what is the mean of the quantity S modulo 311?

1.234567890

What is the standard deviation?

0.9876543210

After $T = 16$ moves, what is the probability that the sum is divisible by 5, given that it is divisible by 13?

0.9876543210

After $T = 512$ moves, what is the probability that the sum is divisible by 7, given that it is divisible by 43?

0.9876543210

Please provide the script used to generate this result (max 10000 characters).

In what language is the script written?

- ◯ C/C++
- ◯ Fortran
- ◯ IDL
- ◯ Java
- ◯ MATLAB
- ◯ Perl
- ◯ Python
- ◯ R
- ◯ Stata
- ◯ SQL
- ◯ VBA
- ◯ Other

Q2:

The UK has a dataset(https://data.gov.uk/dataset/road-accidents-safety-data) on vehicle accidents. Please dowload the "2014 All STATS19 data (accident, casualties and vehicle tables) for 2005 to 2014." Information on the variables can be found at the bottom of the page under additional links. In addition, the form which is used to record data by police officers can be found here(http://docs.adrn.ac.uk/888043/mrdoc/pdf/888043_stats19-road-accident-injury-statistics-report-form.pdf).

What fraction of accidents occur in urban areas? Report the answer in decimal form.

0.6426569109

When is the most dangerous time to drive? Find the hour of the day that has the highest occurance of fatal accidents, normalized by the total number of accidents that occured in that hour. For your answer, submit the corresponding frequency of fatal accidents to all accidents in that hour. Note: round accident times down. For example, if an accident occured at 23:55 it occured in hour 23.

0.9876543210

There appears to be a linear trend in the number of accidents that occur each year. What is that trend? Return the slope in units of increased number of accidents per year.

1.234567890

Do accidents in high-speed-limit areas have more casualties? Compute the Pearson correlation coefficient between the speed limit and the ratio of the number of casualties to accidents for each speed limit. Bin the data by speed limit.

0.9876543210

How many times more likely are you to be in an accident where you skid, jackknife, or overturn (as opposed to an accident where you don't) when it's raining or snowing compared to nice weather with no high winds? Ignore accidents where the weather is unknown or missing.

1.234567890

How many times more likely are accidents involving male car drivers to be fatal compared to accidents involving female car drivers? The answer should be the ratio of fatality rates of males to females. Ignore all accidents where the driver wasn't driving a car.

1.234567890

We can use the accident locations to estimate the areas of the police districts. Represent each as an ellipse with semi-axes given by a single standard deviation of the longitude and latitude. What is the area, in square kilometers, of the largest district measured in this manner?

1234.567890

How fast do the number of car accidents drop off with age? Only consider car drivers who are legally allowed to drive in the UK (17 years or older). Find the rate at which the number of accidents exponentially decays with age. Age is measured in years. Assume that the number of accidents is exponentially distributed with age for driver's over the age of 17.

0.9876543210

Please provide the script used to generate this result (max 10000 characters).

In what language is the script written?

◯ C/C++	◯ Fortran	◯ IDL	◯ Java
◯ MATLAB	◯ Perl	◯ Python	◯ R
◯ Stata	◯ SQL	◯ VBA	◯ Other

Q3: This question is required.

Propose a project to do while at The Data Incubator. We want to know about your ability to think at a high level. Try to think of projects that users or businesses will care about that are also relatively unanalyzed. Here are some useful links about data sources on our blog(http://blog.thedataincubator.com/tag/data-sources/) as well as the archive of data sources on Data is Plural(http://tinyletter.com/data-is-plural/archive). You can see some final projects of previous Fellows on our YouTube Page(https://www.youtube.com/playlist?list=PLOE4k9MRzZanWmZ7MBrJFi7ZekYmVqEIV).

Propose a project that uses a large, publicly accessible dataset. Explain your motivation for tackling this problem, discuss the data source(s) you are using, and explain the analysis you are performing. At a minimum, you will need to do enough exploratory data analysis to convince someone that the project is viable and generate two interesting non-trivial plots supporting this. *The most impressive applicants have even finished a "rough draft" of their projects and have derived non-obvious meaningful conclusions from their data.* Explain the plots and give url links to them. For guidance on how to choose a project, check out this blog post(http://blog.thedataincubator.com/2017/01/how-employers-judge-data-science-projects/).

Propose a project.*

Link to project:
http://nbviewer.jupyter.org/github/udothemath1984/000017_challenge/blob/master/challenge_proposal_v0.ipynb

Link to public description of data source.*

https://www.census.gov/data/developers/data-sets/business-dynamics.html

Link to 1st plot. You are highly encouraged to use Heroku apps domain(https://www.heroku.com/) for an app or Github(https://www.github.com/) to display a notebook.*

http://nbviewer.jupyter.org/github/udothemath1984/000017_challenge/blob/master/103017_challenge_code.ipyr

Link to 2nd plot. You are highly encouraged to use Heroku apps domain(https://www.heroku.com/) for an app or Github(https://www.github.com/) to display a notebook.*

http://nbviewer.jupyter.org/github/udothemath1984/000017_challenge/blob/master/103017_challenge_code.ipyr

How much data did you analyze (in MB)?*

500

How did you obtain your dataset? (Please check all that apply.)

- [x] I downloaded a dataset available online.
- [ ] I used a provided API.
- [ ] I scraped data from a webpage.
- [ ] Other (please explain).

We want to know your communication style. Record a video of yourself giving a high-level proposal of your project to a non-technical person. The video should be no longer than 1 minute and should be at a higher level than the previous explanation.

Record a video of yourself and upload it to YouTube(https://support.google.com/youtube/answer/57407) (and not another video hosting service). Be sure to make the video unlisted (but not private!) so people without the link cannot find it on Google (go here(https://www.youtube.com/my_videos), click "Edit" on your video, select unlisted from the privacy dropdown menu(static/images/youtube-unlisted.png), and save your changes). You can use either your webcam or a smartphone.

Once complete, please provide the *embed* URL of the video. To find this URL (**NOT** the entire iframe tag), on the video's normal watch page, you can click Share → Embed(/static/images/embed.png), and take the link from inside the 'src' attribute of the tag. It looks something like this: https://www.youtube.com/embed/y9tX5whl2U

Please provide the EMBED URL to your video*

https://www.youtube.com/embed/MIuez8fBov4

Note: youtube videos take some time to process after uploading, and your video won't validate until processing is complete. Please allow 10 to 15 minutes for this to take place.

Please provide the script used to generate this result (max 10000 characters).*

Link:
http://nbviewer.jupyter.org/github/udothemath1984/000017_challenge/blob/master/103017_challenge_code.ipynb

In what language is the script written?

- ○ C/C++
- ○ Fortran
- ○ IDL
- ○ Java
- ○ MATLAB
- ○ Perl
- ◉ Python
- ○ R
- ○ Stata
- ○ SQL
- ○ VBA
- ○ Other

For future challenge questions, how many hours did it take you to complete this challenge? This will not be considered in your application (please just enter a number).*

12

☑ By submitting this form, you certify that your answers are the result of your own work and not copied from another individual or source. *

SUBMIT SAVE

You can save your work and return to this page at any point. Once you have filled out the required fields, your challenge submission will be considered 'complete'.

Saved! We have saved a copy of your submission. You can come back before the challenge is due to modify answers. If you have submitted a fully valid challenge at this point, your status has been updated to reflect this. ×

> WITH LOADS OF DATA YOU WILL FIND RELATIONSHIPS THAT AREN'T REAL. BIG DATA ISN'T ABOUT BITS, IT'S ABOUT TALENT.

— FORBES MAGAZINE