“ DATA SCIENTIST: THE
OF THE 21ST CENT

— HARVARD BUSINESS RE

CHALLENGE

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

Please answer as many questions as you can. We do not expect you to mostly optional) but answering more questions correctly will help you. **Pl answers to 10 digits of precision. Partial credit will be given to ansv digits.** You can resubmit your answers on this form as often as you wou will be considered. (*) denotes a required field. A few helpful hints:

1. **Want to get a head start on being a data scientist?** We want all of the challenge questions as possible. So we've written three(http://blog.thedataincubator.com/2015/09/painlessly-dep flask-and-heroku/) blog(http://blog.thedataincubator.com/2015 professional-data-scientist/) posts(http://blog.thedataincubator data-science-part-i-efficient-numerical-computation/) that might mathematics and computation differently. They will also give you a challenge questions. For additional hints on the challenge, follow u Twitter(http://twitter.com/intent/user?screen_name=thedatainc LinkedIn(https://www.linkedin.com/company/the-data-incubatc

Facebook(https://www.facebook.com/dataincubator/).

2. **Having browser troubles?** We recommend using Chrome(https://www.google.com/chrome/browser/desktop/inc Incognito Mode).
3. **Having trouble downloading any files?** We suggest using comm on a browser.
4. **Want to avoid being a statistic?** Every application cycle, a numb minute to submit, only to discover "unforseable" last-minute glitche suggest not waiting until the deadline to submit.
5. **Found something ambiguous?** We realize some questions are a questions are. This is a test of whether you can prioritize important knowledge with theory.
6. Due to the volume of requests, we will only accept submissions via

Q1: You see a stream of T numbers, each ranged 1 through 10 (inclusiv two registers. The first is in a 'max' register which holds the largest N nu is in the 'last' register which holds the last N numbers encountered. Let M in the 'max' register and L be the product of the numbers in the 'last' reg the stream.

Consider the case where $N = 2$ and $T = 8$.

What is the mean of $M - L$?

47.71416969

What is the stan

24.80027723

Now consider $N = 4$ and $T = 32$. Consider the difference $M - L$.

What is the mean of $M - L$?

7892.507078

What is the stan

1592.360155

What is the conditional probability that $M - L \geq a$, given $M - L \leq b$.

... when $N = 2$, $T = 8$, $a = 32$, and $b = 64$?

0.6173959116

... when $N = 4$, $b = 4096$?

0.9373277046

Please provide the script used to generate this result (max 10000 cl

```
#!/usr/bin/perl

use POSIX;
```

```
my $start = time; # To measure execution time

print "\n\nProcessing for T=8 and N=2    Please wait!\n\n";
```

In what language is the script written?

○ C/C++	○ Fortran	○ IDL
○ Matlab	◉ Perl	○ Python
○ Stata	○ SQL	○ VBA

Q2:

With the rise of computer-aided police dispatch systems, many municipa
calls for service. These may include both calls from the public and from
provide this data to the public. New Orleans is one of these, with all of th
2011 available on their Open Data(https://data.nola.gov/) website.

For each of the questions below, use the New Orleans Calls for Service
2015. The data can be found in New Orleans' Open Data portal. Each ye
2011(https://data.nola.gov/api/views/28ec-c8d6/rows.csv?accessType
2012(https://data.nola.gov/api/views/rv3g-ypg7/rows.csv?accessType
2013(https://data.nola.gov/api/views/5fn8-vtui/rows.csv?accessType=
2014(https://data.nola.gov/api/views/jsyu-nz5r/rows.csv?accessType=
2015(https://data.nola.gov/api/views/w68y-xmk6/rows.csv?accessTyp
description of the data can be found here(https://data.nola.gov/Public-
for-Service-2012/rv3g-ypg7/about), for example.

What fraction of calls are of the most common type?

0.2451365000

**Some calls resu
dispatched to th
arrival time. Wh
(dispatch to arri
only valid (i.e. n**

270.0000000

Work out the average (mean) response time in each district. What is the difference between the average response times of the districts with the longest and shortest times?

191.3115567

**We can define s
that occur more
over the whole
the conditional
given a district
of that event typ
which have mor
some events ha
and are reporte
should be ignor**

10.47005835

Find the call type that displayed the largest percentage decrease in volume between 2011 and 2015. What is the fraction of the 2011 volume that this decrease represents? The answer should be between 0 and 1.

0.9463035594

**The disposition
taken to addres
how the disposi
hour of the reco
disposition who
disposition vari
What is its varia
minimum fractio**

0.1796894078

We can use the call 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?

25.31510320

**The calls are as
of calls will rece
priorities. To un
the most variati
call whose most
smallest fractio
is that smallest**

0.0001247423

Please provide the script used to generate this result (max 10000 cl

```
#!/usr/bin/python
__author__ = 'Gungor Ozer'

import math,numpy
import csv
from csv import reader
```

In what language is the script written?

○ C/C++	○ Fortran	○ IDL
○ Matlab	○ Perl	◉ Python
○ Stata	○ SQL	○ VBA

Q3: This question is required.

Propose a project to do while at The Data Incubator. We want to know a level. Try to think of projects that users or businesses will care about (as that only researchers will care about). The project does not have to be c useful links about data sources on our blog (Post 1(http://blog.thedatai sources-for-cool-data-science-projects-part-1/) and Post 2(http://blog.thedataincubator.com/2014/10/data-sources-for-cool-d

Propose a project that uses a large, publically accessible dataset. Motiva discuss the data source(s) you are using, and explain the the analysis yo exploratory data analysis to convince one the project is viable and gener plots supporting this. Explain the plots and give url links to those plots.

1. High-impact problems of general interest are more interesting than you solve the problem, will anyone care? Identifying interesting pro leaving the academy.
2. While their potential is important, projects are assessed primarily b performed. We are looking for data scientists who are able to devli
3. Downloading a pre-formatted, pre-cleaned dataset intended for ma Kaggle datasets) is less impressive than pulling data from an API c realworld data does not come neatly pre-packaged.
4. All things being equal, using other challenge question datasets der We're looking for creative, original thinkers.
5. All things being equal, analysis of larger datasets is more impressiv
6. All things being equal, people who demonstrate the ability to use g Heroku for hosting(https://www.heroku.com/) will be viewed mo following this git tutorial(https://try.github.io/) or these Heroku tutorials(https://devcenter.heroku.com/start) in your favorite lan

Propose a project.*

New York Taxi and Limousine Commission release the cab ride data monthly, wl times, locations, base fare, and tip. I believe there is a lot to learn from this data t pleasant for both drivers and passengers. It would also be of interest to many tra in general.

Link to public description of data source.*

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

Link to 1st plot. You are highly encouraged to use a Heroku apps domain(https://www.heroku.com/) for your hosting.*

https://gungor.cartodb.com

Link to 2nd plot. You are highly encouraged to use a Heroku apps domain(https://www.heroku.com/) for your hosting.*

http://i282.photobucket.com/albums/kk277/dersimden/cab_rides_Jan_2016_sum

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

1985965

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 g
your project to a non-technical person. The video should be no longer th
higher level than the previous explanation.

Record a video of yourself here(https://www.youtube.com/my_webcar
not another video hosting service). Be sure to make the video unlisted (b
the link cannot find it on Google (go here(https://www.youtube.com/m
video, select unlisted from the privacy dropdown menu(static/images
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
on the video's normal watch page, you can click Share → Embed(/static
the link from inside the 'src' attribute of the tag. It looks something like th
https://www.youtube.com/embed/y9tX5whl2U

Please provide the EMBED URL to your video*

https://www.youtube.com/watch?v=1qo3AJdpgv0

Please provide the script used to generate this result (max 10000 cl

```
#!/usr/bin/python
__author__ = 'Gungor Ozer'

import numpy
import math
import csv
```

In what language is the script written?

- ○ C/C++
- ○ Fortran
- ○ IDL
- ○ Matlab
- ○ Perl
- ◉ Python
- ○ Stata
- ○ SQL
- ○ VBA

For future challenge questions, how many hours did it take you to c will not be considered in your application (please just enter a numb

14

☐ By submitting this form, you certify that your answers are the result of from another individual or source. *

SUBMIT

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

— FORBES MAGAZINE