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

CHALLENGE

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 questions correctly will help 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. Only the latest submission will be considered. As a guide, this should take 3-4 hours but you are welcome to take more time if it would help. (*) denotes a required field.

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 two 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 might also give you a head start on solving the challenge questions. Consider following our blog(http://blog.thedataincubator.com/) and our twitter account(https://twitter.com/intent/user?screen_name=thedatainc) for more hints about becoming a data scientist!

If you are having browser troubles, we recommend using Chrome. If you have trouble downloading any files, we suggest using command line tools, rather than relying on a browser.

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.

Due to the volume of requests, we will only accept submissions via this form.

Q1: You have a chain with N links numbered 1 through N. Every minute, you draw a random link from a bag, and connect it to any other consecutively-numbered link that you drew before. For example, if you drew $4, 1, 5, 7, 3$, you would end up with three subchains: $1, (3, 4, 5), 7$. You keep on drawing until you have drawn all N links and connected them into a single chain of length N. Let M be the maximum number of subchains in this process.

What is the mean of the distribution of M if $N = 8$

0.123456789

What is the standard deviation of the distribution of M if $N = 8$

0.123456789

What is the mean of the distribution of M if $N = 16$

0.123456789

What is the standard deviation of the distribution of M if $N = 16$

0.123456789

What is the mean of the distribution of M if $N = 32$

0.123456789

What is the standard deviation of the distribution of M if $N = 32$

0.123456789

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 files Badges.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/Badges.xml.gz), Comments.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/Comments.xml.gz), PostLinks.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/PostLinks.xml.gz), Tags.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/Tags.xml.gz), PostHistory.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/PostHistory.xml.gz), Users.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/Users.xml.gz), Votes.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/Votes.xml.gz), Posts.xml.gz(http://thedataincubator-challenge.s3.amazonaws.com/HCHACdrgRBoVokdgdJNF/Posts.xml.gz) contain zipped XML files from statistics overflow(http://stats.stackexchange.com/), a smaller cousin to the more popular Stackoverflow(http://stackoverflow.com/). The schema is partially outlined here(http://meta.stackexchange.com/questions/2677/database-schema-documentation-for-the-public-data-dump-and-sede).

What fraction of posts contain the 5th most popular tag?

0.123456789

How much higher is the average answer's score than the average question's?

0.123456789

What is the Pearson's correlation between a user's reputation and total score from posts (for valid users)?

0.123456789

How many more upvotes does the average answer receive than the average question?

0.123456789

We are interested in what time of the day one should post to get a fast accepted response. Look at the median response time of the accepted answer as a function of the question post hour (from 0 to 23 inclusive). The response time is the length of time in between when the question was first posted and when the accepted answer was first posted in hours (as a decimal). What is the difference between the largest and smallest median response times across question post hours?

0.123456789

We would like to understand which actions lead to which other actions on stats overflow. For each valid user, create a chronological history of when the user took one of these three actions: posing questions, answering questions, or commenting. For each of these three possible actions, compute the unconditional probability of each action (three total) as well as the probability conditioned on the immediately preceding action (nine total). What is the largest quotient of the conditional probability of an action divided by its unconditioned probability?

0.123456789

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 (as opposed to academic projects that only researchers will care about). The project does not have to be completely novel. Here are some useful links about data sources on our blog (Post 1(http://blog.thedataincubator.com/2014/10/data-sources-for-cool-data-science-projects-part-1/) and Post 2(http://blog.thedataincubator.com/2014/10/data-sources-for-cool-data-science-projects-part-2/)).

Propose a project that uses a large, publically accessible dataset. Motivate the problem you are tackling, discuss the data source(s) you are using, and explain the the analysis you are performing. Do enough exploratory data analysis to convince one the project is viable and generate two interesting non-trivial plots supporting this. Explain the plots and give url links to those plots.

1. Problems of general interest are more interesting than academic research problems.
2. While their potential is important, projects are assessed primarily based on the success of analysis performed.
3. All things being equal, downloading a pre-formatted, pre-cleaned dataset intended for machine-learning is less impressive than scraping a webpage or pulling data from an API.
4. All things being equal, analysis of larger datasets is more impressive than analysis of smaller ones.
5. All things being equal, using other challenge question datasets demonstrates lack of creativity.

Propose a project.*

Link to public description of data source.*

http://blog.thedataincubator.com/2014/10/data-sources-for-cool-data-science-projects-part-1/

Link to 1st plot. Bonus points if you use a Heroku apps domain(https://www.heroku.com/) for your hosting.*

http://example.herokuapp.com/index.html

Link to 2nd plot. Bonus points if you use a Heroku apps domain(https://www.heroku.com/) for your hosting.*

http://example.herokuapp.com/index.html

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

1234

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

- [ ] 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 here(https://www.youtube.com/my_webcam) and upload it to Youtube (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 [e.g. https://www.youtube.com/embed/C9Dlgu2Lm6U]*

https://www.youtube.com/embed/C9Dlgu2Lm6U

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

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).*

9999

☐ 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

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