Student Instructions

Galaxy Zoo & The Power of the Crowd

NOTE: Before beginning, make a copy of this document and rename it YourName_GalaxyZoo101. That is the version you should record your responses on.

I. Background Reading

Activity Overview:

In this activity, you will learn about the power of citizen science and crowdsourcing through classifying a small subset of galaxies based on their shape. You will be able to compare your galaxy classifications to those made by your classmates, Zooniverse volunteers, and a computer programmed to classify galaxies.

What is Zooniverse?

Zooniverse is the largest online platform for people-powered research (a.k.a citizen science). Nearly 2 million people around the world are Zooniverse volunteers, and they help actual research teams make discoveries by parsing through sets of image or text-based data. Anyone can be a Zoonivese volunteer, all you need is an email address! No specialized background, training, or expertise is required. What are you waiting for? Create your own Zooniverse account!

What is Galaxy Zoo?

Galaxy Zoo (http://www.galaxyzoo.org) is an online citizen science project that has engaged over 150,000 volunteers since its founding in 2007. The Galaxy Zoo Project was founded after data started coming back from the Sloan Digital Sky Survey (SDSS). The SDSS is the first and largest digital map of the Universe, and it has provided astronomers with images of over one million galaxies! Galaxy Zoo volunteers are asked to look at, and answer questions about images of galaxies to bin them into separate categories. When a volunteer begins classifying in Galaxy Zoo, they are led to an image of a specific galaxy. This project uses a decision tree method to classify the data, meaning each new question is dependent on how the volunteer answer the first question. This method allows for the exploration of different features of the specific galaxy-type selected. The question sequence is structured in such a way to mimic the thought process that actual astronomers use when they look at galaxy data. To date, findings from the Galaxy Zoo dataset has led to upwards of 60 published papers on topics like galaxy evolution, galaxy color and its relation to stellar evolution, and galaxy mergers.

A famous example of an extremely rare object found through Galaxy Zoo is “Hanny’s Voorwerp.” Hanny’s Voorwerp is a rare object called a quasar ionization echo, and was discovered by a 25-

year old Dutch teacher, Hanny van Arkel while she was classifying galaxies using Galaxy Zoo. Take a look at this video that describes the project in more detail.

Why Crowdsourcing?

A graduate student researcher was tasked with classifying the shape of more than 900,000 galaxies. It was estimated working 24 hours a day, 7 days a week would take an individual 3-5 years to classify all of the galaxies in the sample ***once.*** When the researchers asked for help from the public to classify the data using Galaxy Zoo, more than **100,000 volunteers** made over **40 million galaxy classifications** in just over **6 months!**

If you want to learn a little more about the benefits of crowdsourcing, read page 3 (the summary page) of this PDF.

Is my answer even correct? Why am I qualified to do this?

These are questions we get asked all the time by students getting started with citizen science. The short answer is, you **are** qualified, and the power of crowdsourcing takes a lot of the pressure off with regards to whether your classification is “correct” or not. For Galaxy Zoo in particular, **at least 40 volunteers classify each galaxy**. That way, if a few people get it wrong, it doesn’t matter. The scientists go with the majority. The science team considers a classification to be correct if more than 70% of volunteers agree (or converge) on the same answer -- the consensus result. **Past research into the accuracy of crowdsourcing compared to experts shows that the volunteer consensus result agrees with experts 97% of the time** (Swanson et al. 2015). Long story short, you can do this, and if you’re not sure if your answer is correct the power of the crowd will help you out.

Why doesn’t the research team just get a computer to classify all of these galaxies?

A major effort in citizen science right now is integrating computers alongside human classifications. For example, Galaxy Zoo researchers are using volunteer classifications to train a computer to accurately classify galaxies. The goal is for a computer to be able to efficiently, accurately, and reliably classify galaxies. In this activity you'll see examples where the computer still struggles at making a correct classification. So more work remains to be done to fully train the computer. But even once the computer is fully reliable, it will be essential for humans to remain in-the-loop. For example, it is an inherently human talent to flag the rare, the unusual, and/or unexpected (like Hanny's Voorwerp mentioned above).

II. Galaxy Zoo 101 Activity

Before Getting Started:

1. Please read the Background Reading section above
2. Have your Zooniverse account set up
3. Join the classroom for your class via the link your instructor sent you

Instructions

1. Visit this link and make a copy of the spreadsheet. You will use this to track your own galaxy classifications for later comparison with the class
2. Visit the Zooniverse website and sign-in with your account
3. Once you're signed in, visit the project page link your instructor provided. This will take you to the galaxy classification site. When you arrive, **be sure you see a green banner that states** "You are classifying as a student of your classroom" **at the top of the screen**. If you do not, verify that you are logged in and try the link again.
4. Classify each galaxy image as either:
    a. Spiral
    b. Elliptical
    c. Merger
    d. Star or Artifact
5. If you get stuck at any point you can click on 'NEED SOME HELP WITH THIS TASK' right below your answer choices, or click the 'FIELD GUIDE' tab on the right-hand side of the screen to see some examples of each galaxy type
6. Make note of your galaxy classification in your personal spreadsheet from Step 1 before moving on. The galaxy ID number is noted on the top left corner of each image.
7. Once you've typed your classification choice into your spreadsheet (next to the corresponding galaxy ID number), you can move onto the next galaxy by clicking on the **'Done'** button.
8. Work through all of the galaxies in the dataset. There are 22 total. Then wait for your instructor to initiate the next part of the activity.

III. Analyzing Your Results

In this part of the activity, you'll analyze the results of the galaxy classifications you just completed. You'll compare your classifications to those made by your classmates, Zooniverse volunteers, and the computer. Your instructor will provide the results of all classifications you and your classmates have completed.

Instructions

1. Open the spreadsheet from Part II where you recorded your own classification results for each galaxy
2. Open the results spreadsheet shared with you by your instructor, and make a copy into your own google drive
3. Look at the columns in the results spreadsheet, each row corresponds to a single galaxy from the dataset
    a. Column A: Galaxy ID number

b. Column B: Total number of classifications (from your class)
c. Columns C-F: Total number of votes by you and your classmates for;
    i. Column C: Spiral
    ii. Column D: Elliptical
    iii. Column E: Merger
    iv. Column F: Star/Artifact
d. Column G: The formal name of the galaxy in the Sloan Digital Sky Survey
e. Column H: Links to a webpage for each galaxy where you can view each galaxy image again
f. Columns I-L: Original Galaxy Zoo volunteers' votes for each galaxy type;
    i. Column I: Spiral
    ii. Column J: Elliptical
    iii. Column K: Merger
    iv. Column L: Star/Artifact
g. Column M: Whether the computer classified the galaxy correctly or incorrectly

Questions

1. Find a galaxy in the dataset where you ***agreed*** with the classification of your class overall.

NOTE: You can check your classifications by matching the Galaxy ID number in your spreadsheet to the Galaxy ID numbers in the bigger spreadsheet your teacher sent out.

a. What type of galaxy is it?

b. Did the Galaxy Zoo volunteers agree with the class?

c. Briefly describe what made this galaxy easy to classify for you.

2. Find a galaxy in the dataset where you ***disagreed*** with the classification of the majority of your class.

a. What type of galaxy did you classify it as?

b. What did the majority of students classify it as?

c. Do your neighbors agree or disagree with you?

d. Did the original Galaxy Zoo volunteers agree or disagree with you?

e. Find a neighboring student who classified the galaxy differently from you. Summarize how you and your neighbor disagree. Be specific. What were the features in the image that informed your decision? How about their decision? Ultimately, why was their classification different from yours?

f. After talking to each other today, did you want to change your vote? Did you get your neighbor to change their vote? If so, why?

3. For how many of the galaxies in the dataset did ***your class*** converge on one answer. Remember, converge in this case means that at least 70% of the class chose the same answer. Was this majority of the galaxies in the dataset? What types of galaxies were easier to classify for your class? Which were harder? Why?

4. For how many of the galaxies in the dataset did the ***Zooniverse volunteers*** converge on one answer? Remember, converge in this case means that at least 70% of the volunteers chose the same answer. Was this majority of the galaxies in the dataset?

In Questions 5 - 6 we'll look at some of the outliers:

5. Galaxy ID #s 11, 19, and 20 are galaxies for which the volunteers could not converge on an answer. Look at the images linked in Column H for these galaxies. What characteristics about these galaxies might make them harder to classify?

6. Now let's look at the computer classifications (Column M). Currently the computer program is only able to place the galaxies into two categories: 'elliptical' or 'not elliptical'. 'Not elliptical' includes spiral, merger, and star/artifact. If the computer could not decide on a classification, its choice is labeled 'Unable to Determine'.

a. How many of the galaxies were classified as ‘Unable to Determine’ by the computer?

b. Look at the images of the ‘Unable to Determine’ galaxies using the links in Column H. What are some possible characteristics that made these galaxies difficult for the computer to classify?

c. What types of galaxies does the computer classify well?

NOTE: There will always be images that are difficult to classify and require further detailed analysis by the research team, but this is typical of the scientific process! The majority of galaxies can be reliably classified. For the subset of images that cannot be reliably classified, Galaxy Zoo researchers include those galaxies in the uncertainties and error bars in their analyses. Over the past decade, Zooniverse projects have established a strong track record of the reliability of crowdsourced data, including through over 200 peer-reviewed publications. Recent efforts integrating computers alongside human classifications are providing huge gains in efficiency, but also reinforcing that computers are only as good as the training they receive and the need to keep humans in-the-loop to discover the rare, unusual, and/or unexpected.

7. Now that you’ve learned about crowdsourcing;

a. What do you think are the potential advantages and disadvantages of having a single astronomer classify thousands of galaxies?

b. What are some advantages and disadvantages of having the public classify the galaxies for the astronomer?

8. Can you think of any other experiences in your life where you participate in a form of crowdsourcing, or where you use crowdsourced data to inform a decision?

Now that you've learned about how easily you can contribute to real research through crowdsourcing, check out the over 100 projects on the Zooniverse platform across the disciplines, from astronomy to zoology, climate science to cancer research, arts and the humanities.