Science AMA Series: Dr. Anne Carpenter, Institute Scientist at the Broad Institute of MIT and Harvard providing background on the 2018 Data Science Bowl and questions on cell biology, microscopy, and computational biology. Ask Me Anything.

Anne$_{C}arpenter^{1}andr/ScienceAMAs^{1}$

[1]Affiliation not available

April 17, 2023

Abstract

Hi, I'm Dr. Anne Carpenter, I lead a computational research group at the Broad Institute of MIT and Harvard. My Ph.D. is in cell biology and my lab's expertise is in developing and applying algorithms and software for extracting information from biological images. My team's open-source CellProfiler software is used by thousands of biologists worldwide, and is accelerating the discovery of new medicines. We're passionate about developing tools to speed up research and discover cures to diseases. Nucleus detection is a very important part of this process because most of the human body's 30 trillion cells contain a nucleus full of DNA, the genetic code that programs each cell. Identifying nuclei allows researchers to identify each individual cell in a sample, and by measuring how cells react to various treatments, the researcher can understand the underlying biological processes at work. This year's Data Science Bowl is challenging teams to automate the process of identifying nuclei in images, to allow for more efficient drug testing (right now it takes ˜10 years for a new drug to come to market!) Check out my 5 minute video introduction to the challenge. My team (including yours truly!) hand-annotated more than 20,000 nuclei for the data challenge - we think it was worth it to solve this challenge. My lab's focus has been on deep learning algorithms and we'd love someone to beat our best efforts! Thanks for caring about the intersection of computer science and biology! You can catch me anytime on Twitter, and I'm here from 12-1PM to answer your questions about the challenge, my lab's work, and being a scientist. Ask me anything! Thank you all for joining me today! I'm done! There's still plenty of time to register and compete in the 2018 Data Science Bowl focused on algorithms to spot nuclei. The winning algorithms will be released to the community. Stay connected or join the competition by visiting DataScienceBowl.com. You can also learn more at NVIDIA's GPU Technology Conference: March 26 to 29th in San Jose, CA Booz Allen Hamilton will be hosting a Business Track focused on AI for Social Good as an Innovation Driver, Tuesday, March 27th. Thanks again for joining!

the WINNOWER

REDDIT

Science AMA Series: Dr. Anne Carpenter, Institute Scientist at the Broad Institute of MIT and Harvard providing background on the 2018 Data Science Bowl and questions on cell biology, microscopy, and computational biology. Ask Me Anything.

ANNE_CARPENTER R/SCIENCE

Hi, I'm Dr. Anne Carpenter, I lead a computational research group at the Broad Institute of MIT and Harvard. My Ph.D. is in cell biology and my lab's expertise is in developing and applying algorithms and software for extracting information from biological images.

My team's open-source CellProfiler software is used by thousands of biologists worldwide, and is accelerating the discovery of new medicines. We're passionate about developing tools to speed up research and discover cures to diseases.

Nucleus detection is a very important part of this process because most of the human body's 30 trillion cells contain a nucleus full of DNA, the genetic code that programs each cell. Identifying nuclei allows researchers to identify each individual cell in a sample, and by measuring how cells react to various treatments, the researcher can understand the underlying biological processes at work.

This year's Data Science Bowl is challenging teams to automate the process of identifying nuclei in images, to allow for more efficient drug testing (right now it takes ~10 years for a new drug to come to market!) Check out my 5 minute video introduction to the challenge.

My team (including yours truly!) hand-annotated more than 20,000 nuclei for the data challenge - we think it was worth it to solve this challenge. My lab's focus has been on deep learning algorithms and we'd love someone to beat our best efforts!

Thanks for caring about the intersection of computer science and biology! You can catch me anytime on Twitter, and I'm here from 12-1PM to answer your questions about the challenge, my lab's work, and being a scientist. Ask me anything!

Thank you all for joining me today! I'm done!

There's still plenty of time to register and compete in the 2018 Data Science Bowl focused on algorithms to spot nuclei. The winning algorithms will be released to the community. Stay connected or join the competition by visiting DataScienceBowl.com.

You can also learn more at NVIDIA's GPU Technology Conference: March 26 to 29th in San Jose, CA Booz Allen Hamilton will be hosting a Business Track focused on AI for Social Good as an Innovation Driver, Tuesday, March 27th.

Thanks again for joining!

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

DATE RECEIVED:
February 16, 2018

DOI:
10.15200/winn.151869.99087

ARCHIVED:

How well should one's Kaggle model perform in order to get hired into your research group without references, an education, or a background check?

symphonic_reeker

You can have Allen Goodman's job if you beat his score on the leaderboard :D

February 15, 2018

CITATION:
Anne_Carpenter , r/Science , Science AMA Series: Dr. Anne Carpenter, Institute Scientist at the Broad Institute of MIT and Harvard providing background on the 2018 Data Science Bowl and questions on cell biology, microscopy, and computational biology. Ask Me Anything., *The Winnower* 5:e151869.99087 , 2018 , DOI: 10.15200/winn.151869.99087

Hi! Not sure if my question is relevant but here goes. Can we target drugs specifically at cancer cells and if yes how close are we in developing a promising cure?

beathz

You've pinpointed the exact problem with almost all cancer drugs! They do not really specifically target cancer cells, only fast-growing cells in the body – this is why peoples' hair falls out and they vomit, because normal hair cells and stomach lining cells grow very quickly. The problem is that we cannot just "cure cancer" because there are, like, THOUSANDS of different types of cancer, even if you subdivide it by the part of the body, like lung cancer. So curing cancer is more like chipping away at a large iceberg, it's going to happen piece by piece, with some chunks bigger than others, but it will take a while.

Is CellProfiler using deep learning? If so, since when and how big of a difference did it make?

somewittyalias

Deep learning is not a routine part of using CellProfiler yet – there are a few plugins that can run particular trained models (for example, the MeasureFocus plug-in uses a convolutional neural network to classify whether an image is focused and the Wahlby lab has a plugin for image segmentation).

But we are hoping that changes by the end of the year! We've already invested a lot of resources in laying the groundwork for using deep learning to replace some of CellProfiler's existing algorithms (or developing new algorithms entirely).

If you're interested, you can look at Keras-ResNet, a library we use to build residual networks for feature extraction and Keras-RCNN, a library we use to solve object detection and image segmentation problems. We're always looking for new contributions to our software.

We've also been contributing to many of the foundational tools themselves (e.g. we've contributed to TensorFlow and Keras since the beginning) and a recent student from our lab is now at Facebook working on PyTorch (and he's been super helpful in making our work more accessible to that community).

Hi, thank you for this AMA. My question is: What do i need to study to became a good computational biologist? There's something mandatory to study?

kalpof

It definitely depends on your long-term goals… I think it is great to gain a deep knowledge in statistics and/or machine learning before diving into biology problems. If you instead begin by learning computational biology per se, then you are more likely to focus on practical tools in a given area and skip over the fundamentals. Which is fine so long as those tools are popular, or that problem is popular, but that doesn't give you a good foundation for a long career.

How widespread is automated image analysis in cell biology currently (percentage of manually tagged vs automated experiments)? How big of an impact can it have?

somewittyalias

Biologists are definitely getting savvier about quantifying images even if they have only a handful. I'd say the majority who make images now quantify them somehow. And experiments are getting bigger

and bigger as people use more automation for other steps of an experiment (pipetting, microscopy, etc.).

I would categorize users of my lab's software CellProfiler into two groups: there are scientists that are processing a handful of images and there are major labs processing millions of images using cloud computing. I find it incredible that both types are happy with the same software. Part of the reason for that is that the software is developed by a team of people that is actually using it intensively.

Why are images of nuclei so important to biology research? What are the biggest technical hurdles to segmenting nuclei?

bitfrosting

The answer is definitely not obvious! Some researchers actually study the nucleus but that isn't our main goal… identifying the nucleus is useful for almost all cell experiments because it gives us a single, clear landmark within the cell that is much easier to find than if we tried to find cell edges directly. I explain this more in the video for the Data Science Bowl: https://www.youtube.com/watch?v=Dbiq6l50zO8

The biggest hurdle in identifying nuclei is when cells are clumpy and close together. It's often hard for algorithms to tell whether two nuclei close together is a single lumpy nucleus or two separate nuclei. We suspect deep learning should be able to surpass classical algorithms at this.

Thank you for allocating the time to do this AMA. Going straight to the question, where do you think AI startups are needed most in the field of computational biology (or more specifically medical image analysis)? Do you think there is room for them to grow in terms of marketing and valuation? I am asking because having access to data is a high bar for small startups and without it, they cannot show their worth to investors to be able to afford data collection, a dilemma.

morteza_milani

Good afternoon and thanks for the question! There has been *huge* interest in digital pathology in the VC world so although there is a ton of opportunity for AI to make a difference for medical diagnosis… there will also be a ton of competition in this space. I would say what will differentiate the winners is careful design and collection of clinical data to design tools that work not just on test sets but in real life. Overfitting is a major danger here. I think everyone should read this article on Voodoo machine learning because these kinds of problems have fooled just about every field that's used machine learning! https://www.biorxiv.org/content/early/2016/06/19/059774