Methods in Open Science

Introduction

Image from NASA: https://commons.wikimedia.org/wiki/File:Earth_Western_Hemisphere_transparent_background.png#filelinks

Welcome to Methods in Open Science. The primary goal of this course is to explore methods that scientists and analysts use to process, explore, and analyze data; document, present, and communicate results; and generate useful knowledge from raw data using reproducible methods.

This course has several broad components. The more conceptual material will predominantly be presented using web-based slides and modules. Annotations and text will be provided throughout in the Notes section to the right of the slide. I also include some embedded videos, some interactive web content, links to web resources, references to publications, and book recommendations.

I have also provided materials associated with learning to code and analyze data in both the Python and R languages and computational environments. This material is intentionally redundant: you can complete the assignments in the language of your choosing or attempt them in both environments. In the world of data science, Python and R are both used, and many practitioners are bilingual. So, I have included examples in both languages. This material is presented as webpages with code examples and explanations, videos, examples, and assignments.

This course is meant for those who have no prior experience working with data analytics or coding. So, it is fine if these concepts are completely new to you. I

have tried to focus on key concepts and hands-on applications. Once you have completed this course, you may want to take a deeper dive into some of the broad topics discussed.

Let's get started with some introductions and an overview of data science and reproducible science.

Course Goals

- Explain the characteristics of data and the methods we use to extract useful information from them
- Implement data cleaning, manipulation, and summarization to prepare raw data as input to additional analyses
- Code at an intermediate level in the Python or R language/computational environment
- Construct effective graphs and data visualizations
- Execute and interpret statistical tests and assess the appropriateness of tests and input data for exploring a specific hypothesis
- Critique modeling methods for making predictions and assess model output
- Perform data science experiments to address specific questions using appropriate techniques and data

2

This slide lays out the goals of this course. Please take some time to read through these goals. If this seems overwhelming, don't worry. We will work though this material using a variety methods, and I'll provide links to additional resources for those that want to continue learning.

Thanks

AmericaView and United State Geological Survey

WV View is supported by AmericaView and the U.S. Geological Survey under Grant/Cooperative Agreement No. G18AP00077. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the opinions or policies of the U.S. Geological Survey. Mention of trade names or commercial products does not constitute their endorsement by the U.S. Geological Survey.

National Science Foundation

This work was also funded by the National Science Foundation (NSF) (Federal Award ID No. 2046059: "CAREER: Mapping Anthropocene Geomorphology with Deep Learning, Big Data Spatial Analytics, and LiDAR").

3

This course was funded by AmericaView, which obtains funding from the United States Geological Survey. AmericaView supports geospatial and remote sensing research, education, and outreach. This larger organization funds state views, such as West Virginia View. I am currently the principal investigator for West Virginia View. Over the past several years, we have generated a series of courses, which we host on our West Virginia View website, with funding from AmericaView. So, this course would not have been possible without this support.

This material was also funded by the National Science Foundation (NSF). Thanks to the NSF for making this course possible.

Contact Info

Aaron.Maxwell@mail.wvu.edu

GitHub

https://github.com/maxwell-geospatial

YouTube

https://www.youtube.com/channel/UCpSooZG5FtMyKwGnHaGnvrA/videos

http://www.wvview.org/index.html

4

Here are links to my GitHub, YouTube, and West Virginia View webpages. I have also included my email. Feel free to reach out if you have any questions or comments, and please check out the other instructional materials that we provide through these mediums.

Example Data: Auto MPG

- https://archive.ics.uci.edu/ml/datasets/Auto+MPG
- n = 398 (total number of records)
- Attributes
  - mpg: continuous
  - cylinders: multi-valued discrete
  - displacement: continuous
  - horsepower: continuous
  - weight: continuous
  - acceleration: continuous
  - model year: multi-valued discrete
  - origin: multi-valued discrete
  - car name: string (unique for each instance)

5

Throughout the lecture modules, we will make use of two datasets. The Auto MPG dataset is made available by the University of California Irvine Machine Learning Repository. A total of 398 records are provided that include different pieces of information for specific car models including fuel efficiency, number of cylinders, engine displacement, horsepower, weight, acceleration, model year, and car name. This is a famous dataset that is used in a lot of machine learning examples.

Example Data: Seeds

- https://archive.ics.uci.edu/ml/datasets/seeds
- n = 210
- Attributes
  - area
  - perimeter
  - Compactness (C = 4*pi*A/P^2)
  - Length of kernel
  - Width of kernel
  - Asymmetry coefficient
  - Length of kernel groove

The examined group comprised kernels belonging to three different varieties of wheat: Kama, Rosa, and Canadian (70 elements each).

6

The second dataset is also provided by the UCI Machine Learning repository. It consists of 210 measurements of three different types of wheat kernels: Kama, Rosa, and Canadian. There are 70 random samples of each type and a variety of physical measurements of the kernels have been included, such as area, perimeter, compactness, length of kernel, and width of kernel.

Book Recommendation

Baumer, B.S., Kaplan, D.T. and Horton, N.J., 2017. *Texts in Statistical Science: Modern Data Science with R.* Chapman and Hall/CRC.

https://www.taylorfrancis.com/books/mono/10.1201/9781315113760/texts-statistical-science-benjamin-baumer-daniel-kaplan-nicholas-horton

https://mdsr-book.github.io/mdsr2e/

7

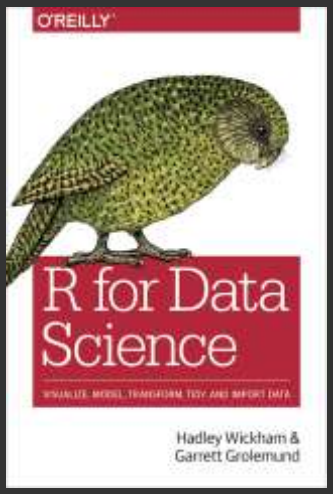
There is no required text for this class. However, if you would like to use a textbook, I would recommend *Modern Data Science in R* by Baumer, Kaplan, and Horton. Note that you can buy a physical copy of the book or access it free at the second link on the slide. This is a great text!

Again, no textbook is required to work through this material. However, I will provide book recommendations throughout the course to help you further your learning or investigate topics in greater detail after completing the course.

Book Recommendation

Wickham, H. and Grolemund, G., 2016. *R for data science: import, tidy, transform, visualize, and model data.* " O'Reilly Media, Inc.".

https://hadley.nz/
https://r4ds.had.co.nz/



8

This is another great text to consider.

Data Science

I will now introduce the field of data science and why it is worth learning about.

What is Data Science?

Science that explores how to extract actionable knowledge from raw data.

Draws from:

- Mathematics
- Statistics
- Computer Science
- Machine Learning
- Deep Learning
- Computer Vision
- Database Management
- High Performance Computing
- Discipline-Specific Knowledge

10

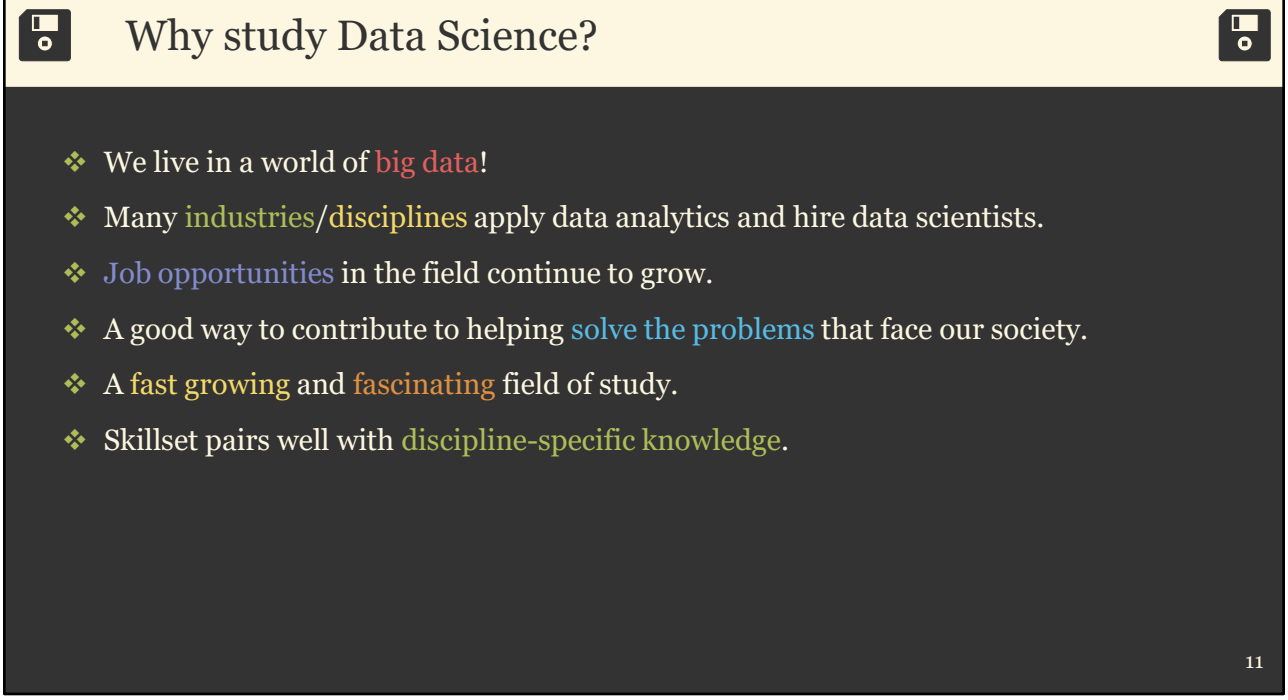
Although data science and data analytics have been undertaken for years in private, non-profit, government, research-and-development, and academic settings, the concept of data science as a distinct field is fairly new. Generally, data science explores how to extract actionable knowledge from raw data.

Data science developed from and continues to draw from a variety of fields including mathematics, statistics, and computer science. Many tasks in data science relate to making predictions, so machine learning, deep learning, and computer vision are also key components of the field. We will discuss these concepts later in the course, and you will develop and assess your own models in some of the provided assignments.

Since data scientists deal with data, database technologies are also important. We need to be able to efficiently store and query a variety of data types. Modern datasets can be very large, so high-performance computing is important since not all analyses can be performed using laptop or desktop computers. Data science has been made possible by “big data” and continues to adapt to allow us to deal with and make use of large datasets.

Data science is used in a variety of disciplines, so discipline-specific knowledge is very important. It is difficult to analyze data and extract useful information from them if you don’t understand their meaning. A knowledge of a specific

discipline combined with data science skills can be a road to developing marketable skillsets and an impactful career.



There are many topics to study and many options as to how to spend our time and energy, so why devote time and mental energy into learning data science? This slide highlights why you should consider pursuing a study of this specific topic.

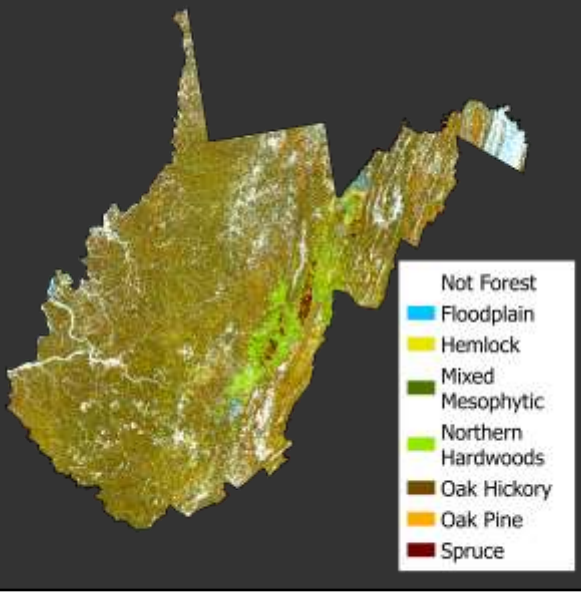
We live in a world of big data. Large volumes of raw data are collected constantly, and individuals that have the skills to analyze them are in demand. Although we can't predict the future, it is unlikely that this will change anytime soon. In fact, a wide variety of industries and disciplines hire data scientists or use data science methods. This includes business, health care, the physical sciences, the social sciences, and even arts and entertainment. As a result, there are currently a lot of job opportunities for data scientists.

Further, data scientists can develop the skills required to help address a variety of problems facing our society, such as global climate change, food security, health care quality and access, and community resiliency. This is a fast growing and fascinating field of study. So, if you like to continually learn and develop new skills or augmenting existing skillsets, data science may be for you.

As mentioned above, discipline-specific knowledge is also key. For example, a public health professional could benefit from a knowledge of both epidemiology and data science. You could be a forester that does some data

science or a data scientist that works with forestry data. In short, if you decide to study data science in college it could pair well with another degree, either as a dual degree or as a minor.

My interest in Data Science



- I am a geospatial scientist
- I work with geographic information science (GISc) and remote sensing
- I work with Earth data

12

I developed an interest in data science from within the Earth Sciences. I primarily work with geospatial, or map, data, and data science has become a key component of our field. From the perspective of data science, geospatial data and methods are also key. I have no formal training in data science. Instead, I developed these skills during and after completing my graduate work, as I found that they were necessary to do the type of work I wanted to do and answer the types of questions I wanted to answer.

For example, the map on this page was created by my graduate student Faith Hartley and myself. Specifically, it was created using satellite data and machine learning methods. Machine learning has in many ways revolutionized our ability to extract information from geospatial and remotely sensed data.

Although learning a new skillset can be frustrating and daunting, I have found learning data science to be very rewarding, and it has allowed me to do the type of work I want to do and answer the types of questions I want to answer. In short, it was a good investment of time and energy for me. If you are committed to learning data science, I think you will find that it will be rewarding and worth the struggle. This course was generated to make it easier for you to get started in this field.

What disciplines use Data Science?

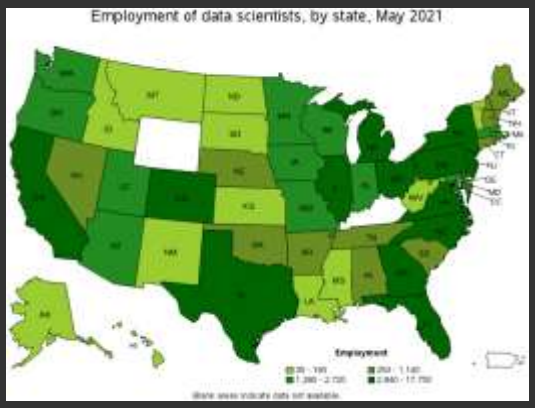
- Engineering
- Computer Science
- Public Health/Epidemiology
- Biology/Ecology
- Medical Imaging
- Physics/Astronomy
- Geology/Earth Science
- Resource Management
- Forestry/Fire Science
- Geography/Geospatial Science
- Business/Economics
- Archeology
- Sociology/Anthropology
- Agriculture
- Arts and Entertainment
- And More!

13

As mentioned above, data science draws from and is used in a variety of disciplines. This slide provides some examples. So, opportunities abound!

Why is Data Science growing so quickly?

- Large volumes of data
- The Internet
- Technology and sensors
- Advances in computing
- Advances in machine learning and deep learning
- Filling a crucial need



https://www.bls.gov/oes/current/oes152051.htm

14

The image on this slide was taken from the Bureau of Labor Statistics and highlights the growth of data science as a discipline and career field.

So, why has data science grown so quickly and also so recently? Part of this can be attributed to the availability of data, such as those generated by the internet and sensors, along with advancements in computing, machine learning, and deep learning. We now have the data and tools required to answer complex questions with data. Further, there is a crucial need to extract useful and actionable information from raw data to inform our decision making and help solve societal problems.

With any changes in technology, science, and/or available methods, questions of best practices and ethics are raised. We will discuss this throughout this course. Broadly, it is important that we represent our data and findings in ways that are not biased or misleading, intentional or otherwise. Again, this is a point we will return to throughout this course. It is also important that data analyses and scientific research is reproducible and rigorous.

Some common misconceptions about Data Science

I can't learn data science because:

The math is too difficult.

I won't be able to learn how to code.

I don't have the required resources.

15

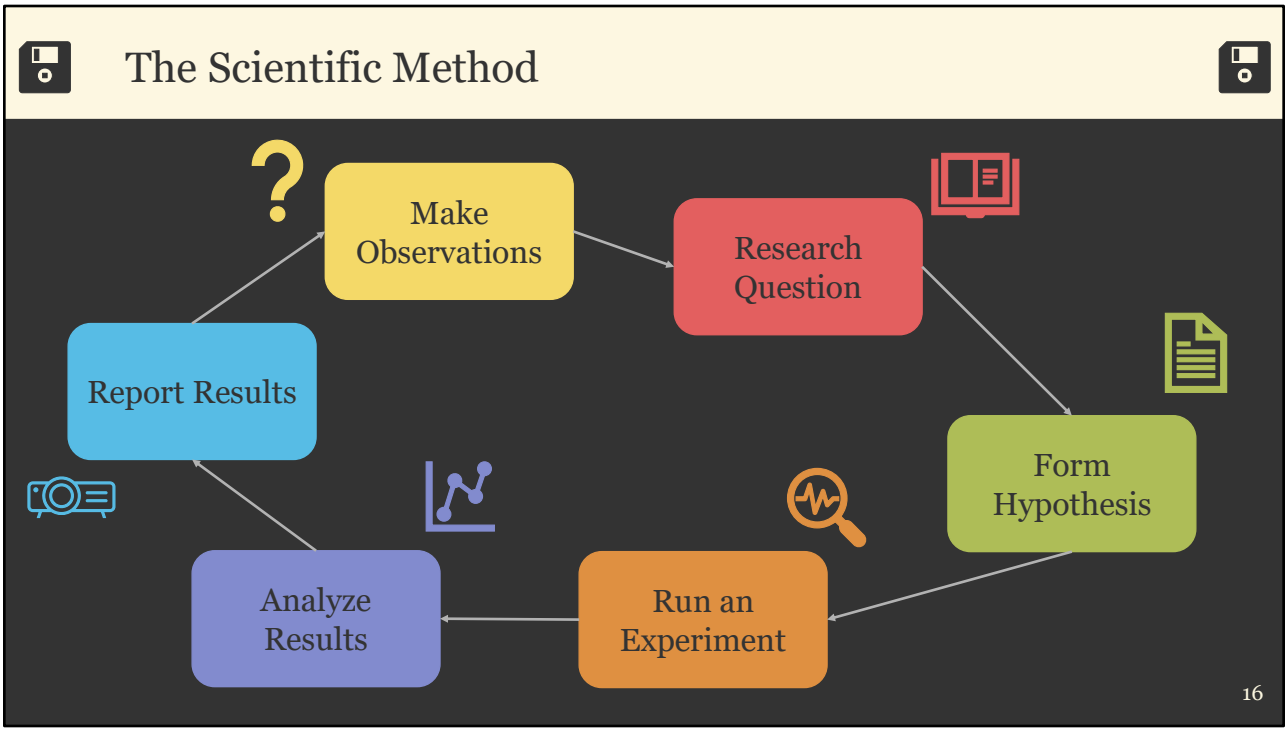
The beginning of any journey can feel daunting. However, I want to assure you that you will be able to develop the skills necessary to learn data science. Although some of the algorithms and techniques used can be mathematically complicated, a conceptual, practical understanding can go a long way. That is the focus of this course.

Learning how to code is not as difficult as you might think. In fact, I think you will be surprised at how easy you pick up this skill. I have found that it is easiest to learn a skill when you are applying it to real questions or real problems that interest you. So, in this course you will learn to code by applying these skills to real problems.

A variety of free and/or cheap resources are available to learn data science and coding methods. This entire course is free and openly available thanks to AmericaView and the NSF. There are other websites, books, and videos that are also free. I will point out other resources throughout this course. Note that the two texts mentioned above are freely available online.

There are some resources that may not be free but are affordable. For example, some great data science textbooks are very reasonably priced. Again, I will make book recommendations throughout this course.

To work through this course, you will not need to buy any software. You will need access to a computer, but it does not need to be a powerful computer. In fact, the code examples and assignments can be executed in a web browser using Google Colab.



The field of data science makes use of the scientific method to answer questions.

The scientific method begins with observation and the formulation of a question. Background research may indicate that this question has already been answered and is not in need of additional investigation. However, if there appears to be a knowledge gap, the researcher can form a hypothesis and conduct an experiment to test it. The results of the experiment can then be analyzed, and the results can be presented or reported.

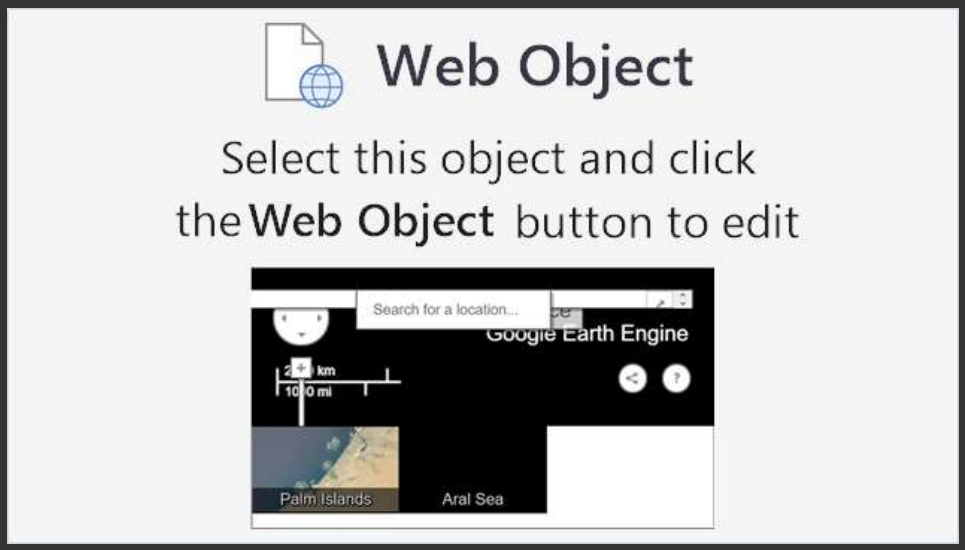
The scientific method may seem lofty, but we use this means of answering questions and understanding the world on a day-to-day basis. For example, you may observe that your car is making a strange noise and question why this is happening. You could then do some background research to determine what could be the cause. If the background research does not pinpoint the problem or only helps to narrow down the problem to a list of potential causes, you can then form a hypothesis and test it. Your hypothesis might be that the noise is caused by a bad wheel bearing. Your experiment could consist of observing whether or not the noise occurs all the time or just when the vehicle is moving. You could determine if it changes pitch as your speed varies. Lastly, you could lift the car and see if the tire seem loose. If your experiments indicate that it is indeed the wheel bearing you could then conclude that your hypothesis is likely

true and proceed to replace it or take it to the shop.

Scientific investigations often lead to more investigations. If your hypothesis is not supported by your experiment and subsequent analysis, then you may form a new hypothesis and design a new experiment to test it. If your experiment supports your hypothesis then new or more specific questions may be raised that can lead to more research, hypothesis formulation, and experimentation.

In short, the scientific method offers a rigorous and systematic means to explore our world. Since experimentation yields data that needs to be analyzed, scientists can benefit from developing skills in data science and data scientists can benefit from thinking in terms of the scientific method.

Example 1: Water Resources

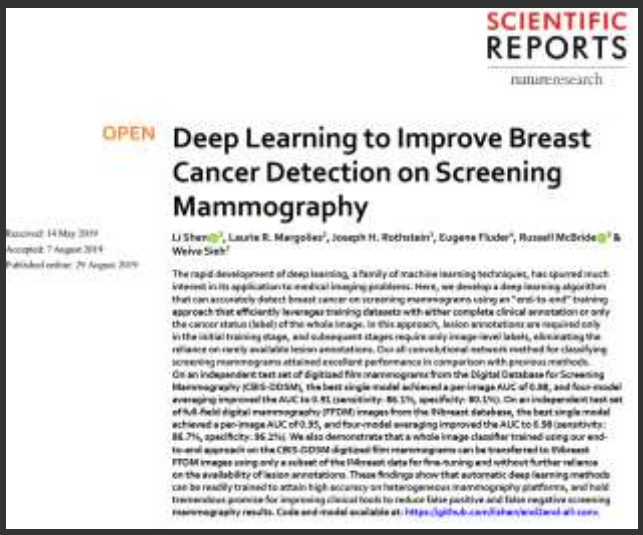


https://global-surface-water.appspot.com

17

In the last few slides in this section, I wanted to highlight some practical uses of data science. The website linked on this page represents a global mapping of the extent of water and changes in the extent of water. This was undertaken using publicly available satellite data and a computing platform made available by Google called Google Earth Engine. With these resources, the researchers were able to generate a global product in a reasonable amount of time that can even be updated as new images are made available. This is a truly amazing product that only recently has been possible to create.

Example 2: Cancer Diagnosis



SCIENTIFIC REPORTS

nature research

OPEN

Deep Learning to Improve Breast Cancer Detection on Screening Mammography

Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride & Weiva Sieh

The rapid development of deep learning, a family of machine learning techniques, has spurred much interest in its application to medical imaging problems. Here, we develop a deep learning algorithm that can accurately detect breast cancer on screening mammograms using an "end-to-end" training approach that efficiently leverages training datasets with either complete clinical annotation or only the cancer status (label) of the whole image. In this approach, lesion annotations are required only in the initial training stage, and subsequent stages require only image-level labels, eliminating the reliance on rarely available lesion annotations. Our all convolutional network method for classifying screening mammograms attained excellent performance in comparison with previous methods. On an independent test set of digitized film mammograms from the Digital Database for Screening Mammography (CBIS-DDSM), the best single model achieved a per-image AUC of 0.88, and four-model averaging improved the AUC to 0.91 (sensitivity: 86.1%, specificity: 80.1%). On an independent test set of full-field digital mammography (FFDM) images from the INbreast database, the best single model achieved a per-image AUC of 0.95, and four-model averaging improved the AUC to 0.98 (sensitivity: 86.7%, specificity: 96.1%). We also demonstrate that a whole image classifier trained using our end-to-end approach on the CBIS-DDSM digitized film mammograms can be transferred to INbreast FFDM images using only a subset of the INbreast data for fine-tuning and without further reliance on the availability of lesion annotations. These findings show that automatic deep learning methods can be readily trained to attain high accuracy on heterogeneous mammography platforms, and hold tremendous promise for improving clinical tools to reduce false positive and false negative screening mammography results. Code and model available at: https://github.com/lishen/end2end-all-conv

Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R. and Sieh, W., 2019. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, *9*(1), pp.1-12.

18

Data science is also of great importance in the field of medical diagnostics and imaging. This paper discusses using deep learning methods to improve and augment the screening of breast cancer. Algorithms are trained and models are developed to interpret mammograms and augment the interpretation undertaken by doctors and radiologists. This highlights a key theme in data science and computer vision: the goal is generally not to replace humans but to assist and augment humans to improve the accuracy, effectiveness, and efficiency of key tasks.

Summary of Key Points

- Data Science explores how to extract actionable knowledge from raw data.
- It is grounded in the scientific method.
- It draws from a wide variety of disciplines.
- Skills in data science are often paired with content knowledge.

19

All sections in this course end with a summary of key points. Make sure to read through these slides.

This is the end of this lecture module.

Please return to the West Virginia View Webpage for additional content.

Thanks! Hope you found this useful.