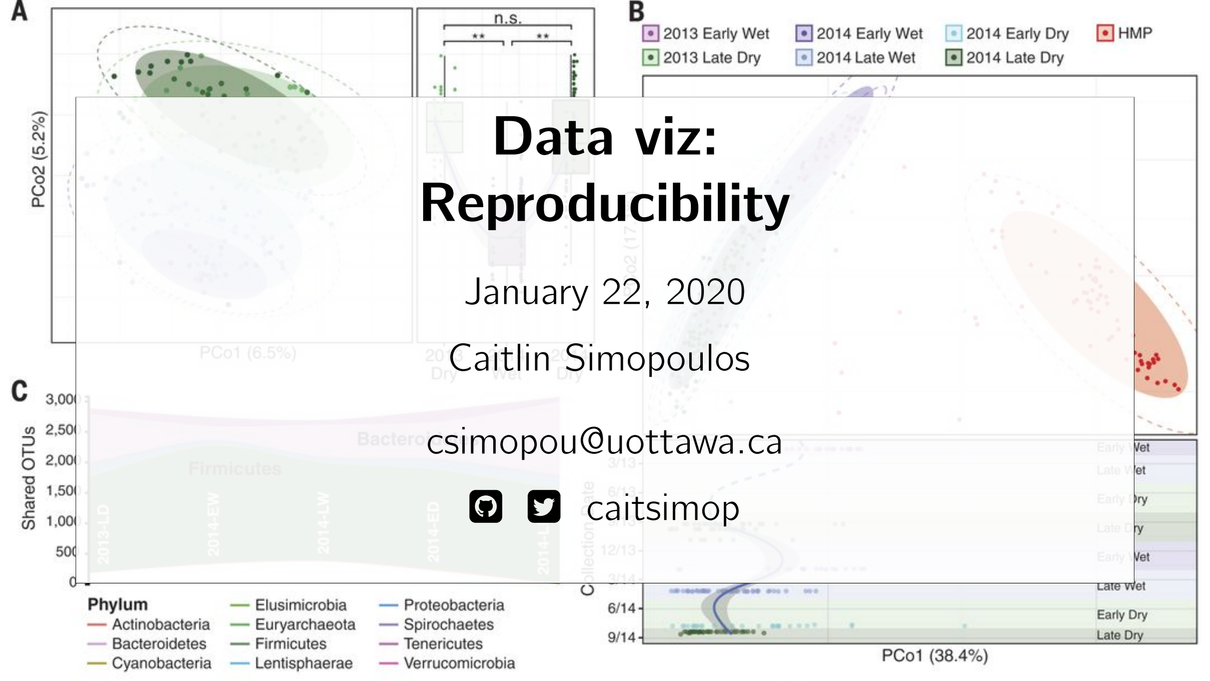


Data viz: Reproducibility

January 22, 2020

Caitlin Simopoulos

csimopou@uottawa.ca

caitsimop

Overview

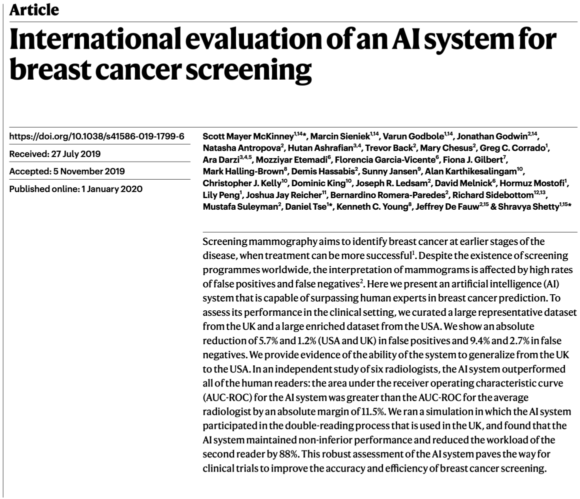
1. What is reproducible data science?
2. Why is reproducibility essential when analyzing/visualizing data?
3. What is required to reproduce figures?
4. An example on how to be reproducible using R.

What is reproducible data science?

It is **raw data** and a set of instructions that explain **exactly** how you:

- Completed your statistical analyses (from raw data)
- Produced your figures and tables

Is this helpful?



Article

International evaluation of an AI system for breast cancer screening

https://doi.org/10.1038/s41586-019-1799-6

Received: 27 July 2019

Accepted: 5 November 2019

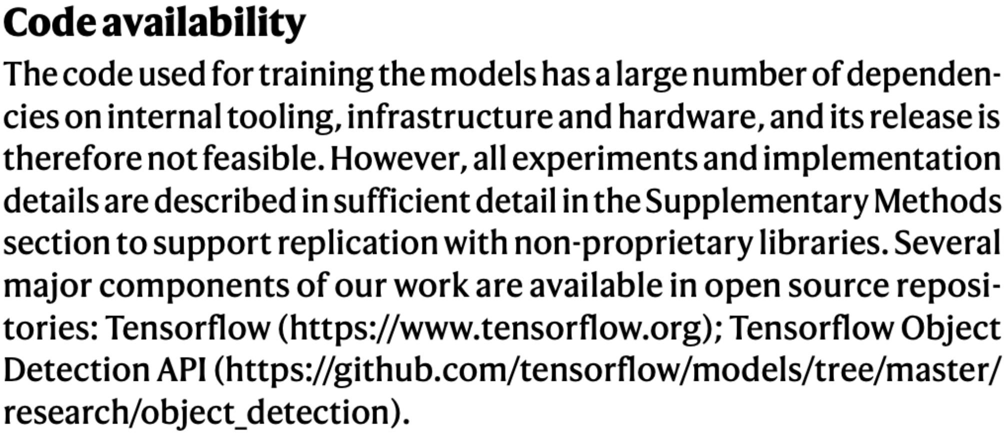
Published online: 1 January 2020

Scott Mayer McKinney[1,14]*, Marcin Sieniek[1,14], Varun Godbole[1,14], Jonathan Godwin[2,14], Natasha Antropova[2], Hutan Ashrafian[3,4], Trevor Back[2], Mary Chesus[2], Greg C. Corrado[1], Ara Darzi[3,4,5], Mozziyar Etemadi[6], Florencia Garcia-Vicente[6], Fiona J. Gilbert[7], Mark Halling-Brown[8], Demis Hassabis[2], Sunny Jansen[9], Alan Karthikesalingam[10], Christopher J. Kelly[10], Dominic King[10], Joseph R. Ledsam[2], David Melnick[6], Hormuz Mostofi[1], Lily Peng[1], Joshua Jay Reicher[11], Bernardino Romera-Paredes[2], Richard Sidebottom[12,13], Mustafa Suleyman[2], Daniel Tse[1]*, Kenneth C. Young[8], Jeffrey De Fauw[2,15] & Shravya Shetty[1,15]*

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To assess its performance in the clinical setting, we curated a large representative dataset from the UK and a large enriched dataset from the USA. We show an absolute reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives. We provide evidence of the ability of the system to generalize from the UK to the USA. In an independent study of six radiologists, the AI system outperformed all of the human readers: the area under the receiver operating characteristic curve (AUC-ROC) for the AI system was greater than the AUC-ROC for the average radiologist by an absolute margin of 11.5%. We ran a simulation in which the AI system participated in the double-reading process that is used in the UK, and found that the AI system maintained non-inferior performance and reduced the workload of the second reader by 88%. This robust assessment of the AI system paves the way for clinical trials to improve the accuracy and efficiency of breast cancer screening.

Nature. 2020. https://doi.org/10.1038/s41586-019-1799-6

Is this helpful?



Code availability

The code used for training the models has a large number of dependencies on internal tooling, infrastructure and hardware, and its release is therefore not feasible. However, all experiments and implementation details are described in sufficient detail in the Supplementary Methods section to support replication with non-proprietary libraries. Several major components of our work are available in open source repositories: Tensorflow (https://www.tensorflow.org); Tensorflow Object Detection API (https://github.com/tensorflow/models/tree/master/research/object_detection).

Nature. 2020. https://doi.org/10.1038/s41586-019-1799-6

Why are we even talking about reproducibility?

1. It saves **you** time in the long run
   - Small OR big changes to figures are less painful
   - Can create automated pipeline for new data!
2. Journals often require code to be published so that the entire pipeline can be reproduced
   - **Shiny apps** ensure code will run and be used by researchers who may be uncomfortable with programming
3. It's just good science

What is required for reproducible analysis/figs?

1. Raw data
2. Code used to **normalize** and **analyze** data
3. Output data file that can be directly visualized (**optional**)
4. File that discloses software/package versions; download date of any files used (*e.g.* databases)
5. A "README" file that explains how to use your code
    - can include previous point
6. An **organized** repository to store these required files

What is required for reproducible analysis/figs?

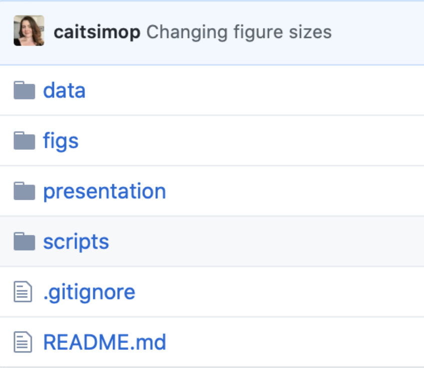
1. Raw data
2. Code used to **normalize** and **analyze** data
3. Output data file that can be directly visualized (**optional**)
4. File that discloses software/package versions; download date of any files used (*e.g.* databases)
5. A "README" file that explains how to use your code
    - can include previous point
6. An **organized** repository to store these required files

Everything needed to get someone from raw data to your figures.

An example:

https://github.com/caitsimop/reproducible-figs

Notice the directory/folder structure



caitsimop Changing figure sizes

data

figs

presentation

scripts

.gitignore

README.md

- Organized
- Obvious
- Includes ALL files needed

Follow along with me in R (if you want), or just watch me on the screen