UNIVERSITY OF CAMBRIDGE

A tale of two crises: COVID-19 and ML reproducibility

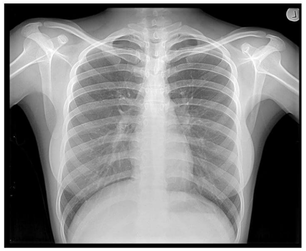
Trustworthy AI for Medical and Health Research Workshop

Michael Roberts
Department of Applied Mathematics and Theoretical Physics

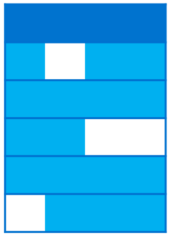
email: mr808@cam.ac.uk

A pandemic of reproducibility issues

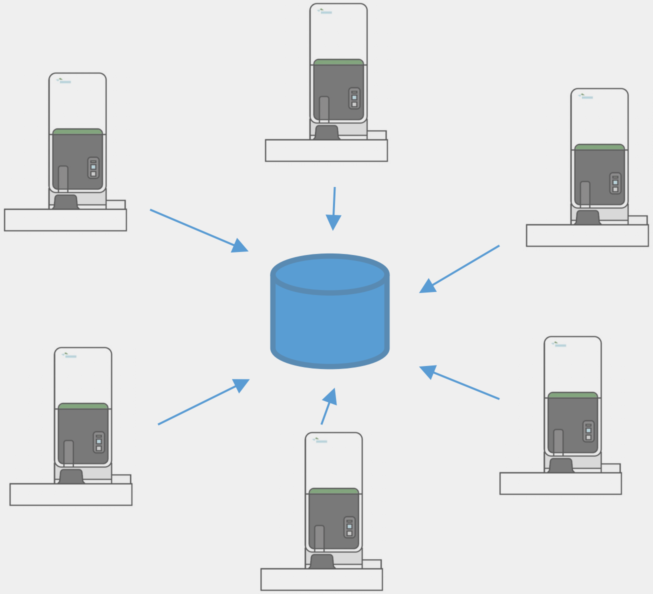
ML for COVID-19



ML for Incomplete Data



ML at scale



ML and code

python PyTorch

jupyter TensorFlow

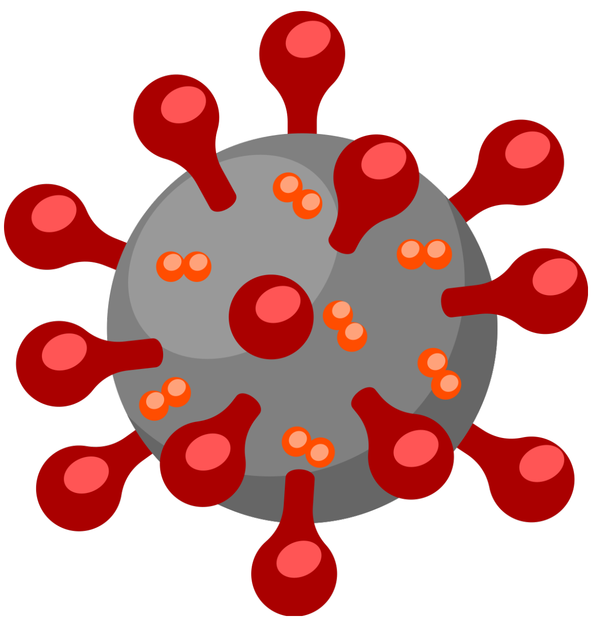
UNIVERSITY OF CAMBRIDGE

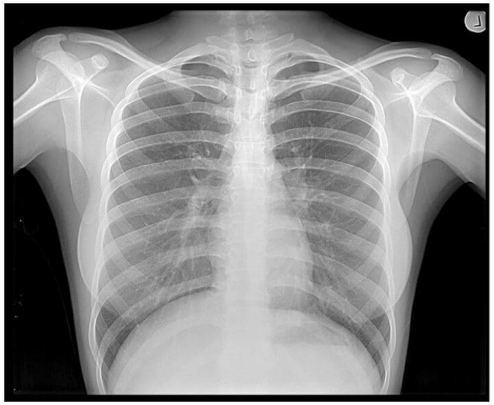
email: mr808@cam.ac.uk

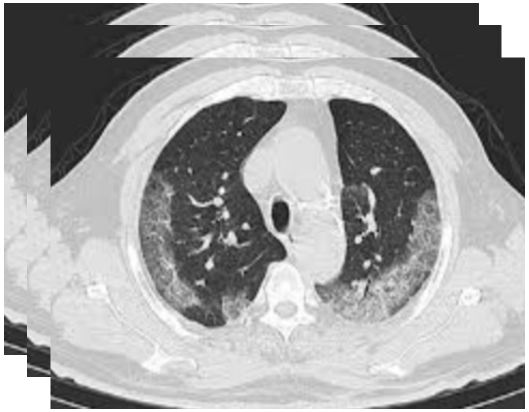
Issues with …

Joint with Derek Driggs, Matthew Thorpe, Julian Gilbey, Angelica I. Aviles-Rivero, Cathal McCague, James Rudd, Evis Sala, Carola-Bibiane Schönlieb and many AIX-COVNET members

ML for COVID-19 imaging







Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans.
Nature Machine Intelligence 3, 199–217 (2021).
https://doi.org/10.1038/s42256-021-00307-0

email: mr808@cam.ac.uk

COVID-19: A perfect test case for image based ML

- Never before have we had access to data:
  - at such scale for a single disease
  - all collected around the same time period
  - collected on the same machines
  - for a disease with a short infection time

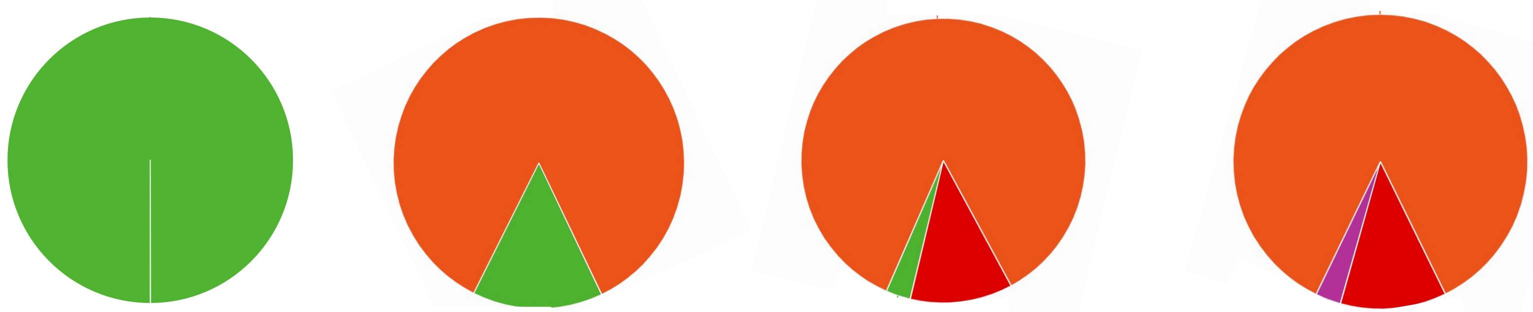
So why did image based ML fail to contribute significantly to the COVID-19 pandemic?

Roberts, M., Driggs, D., Thorpe, M. et al. *Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans.* Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Systematic Review

- **Eligibility:** Any papers using ML and CXR/CT imaging for COVID-19 diagnosis or prognosis.

Screened	**Eligible for Review**	**After Quality Review**	**Of Safe Clinical Utility**
2,212	320	62	0



UNIVERSITY OF CAMBRIDGE

Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Basic pitfalls: Frankenstein datasets

- Know where your data comes from!

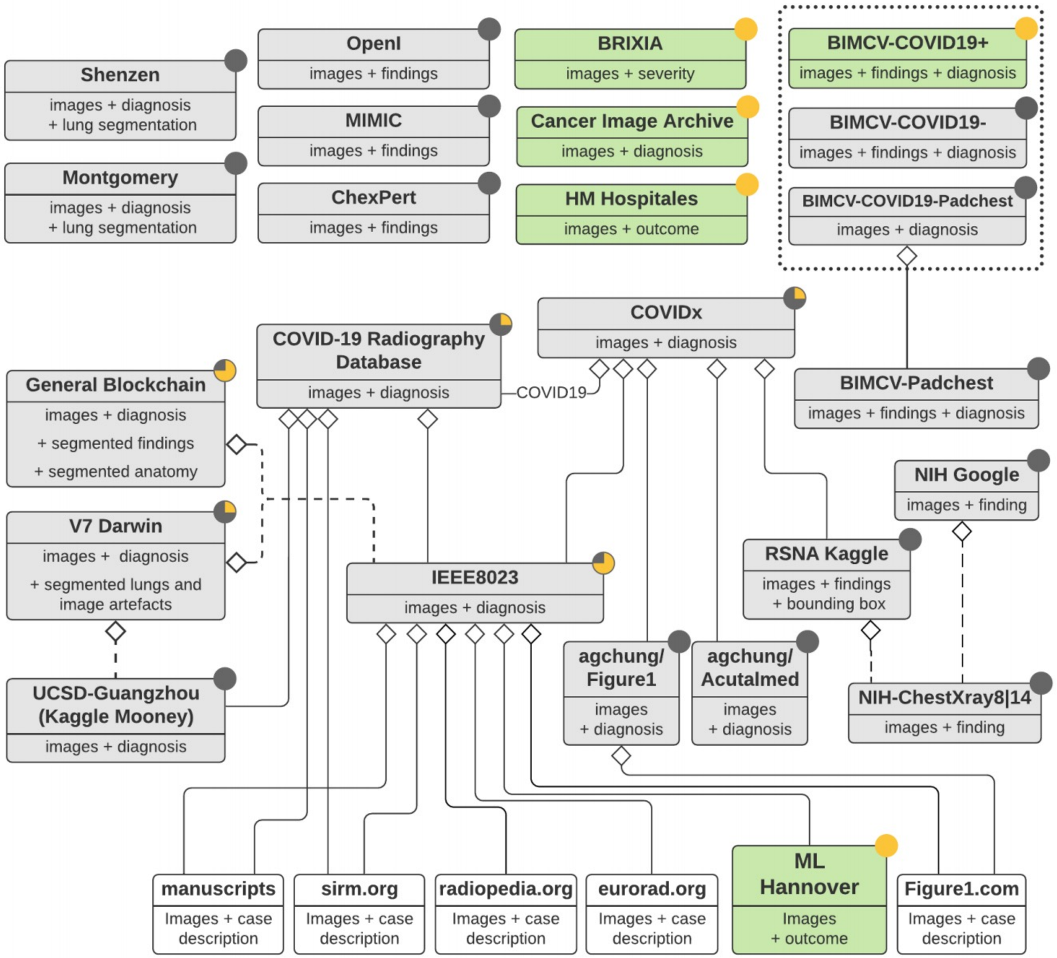


Figure from Cruz et al. (2021)

Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Basic pitfalls: Frankenstein datasets

- Know where your data comes from!

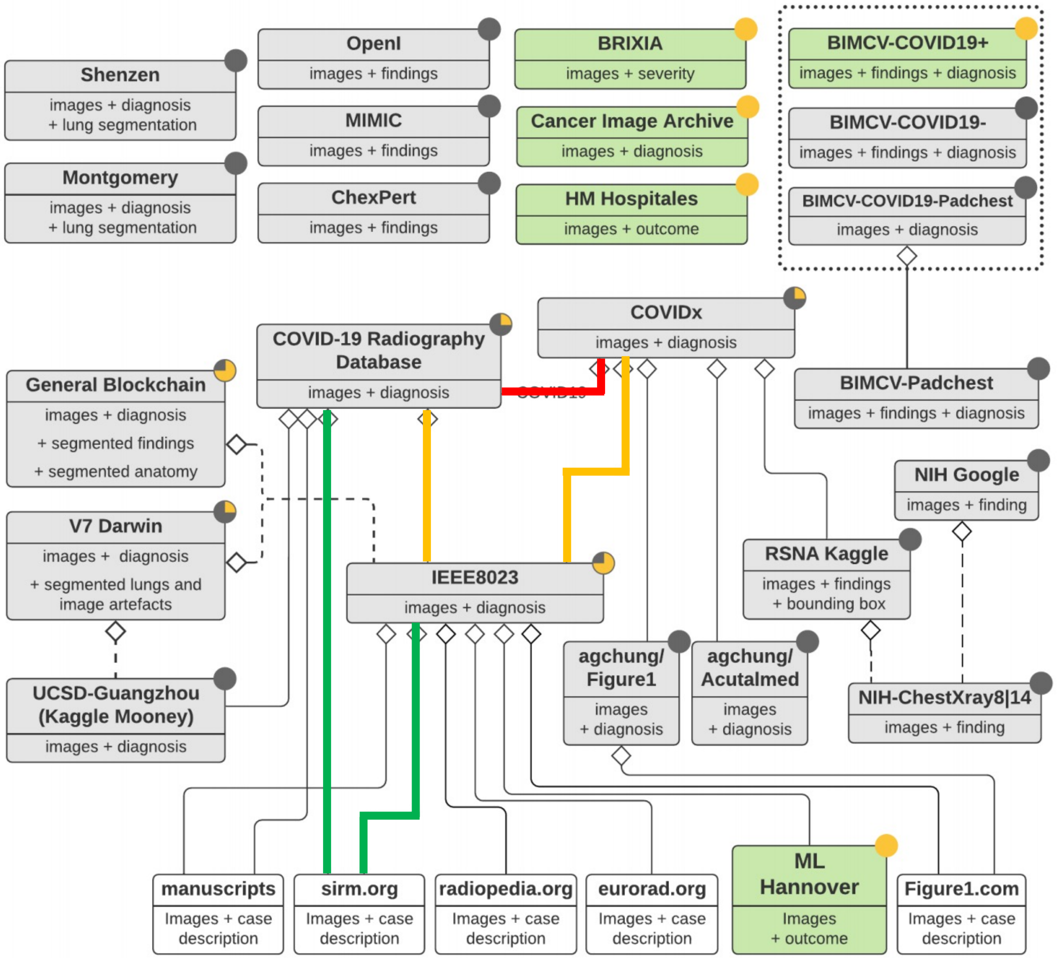
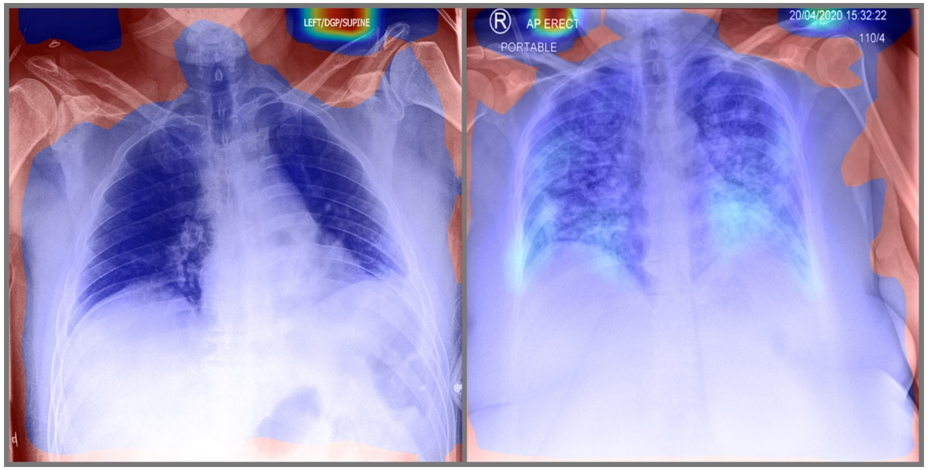


Figure from Cruz et al. (2021)

Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Basic pitfalls: biases in images

- Know where your data comes from!
- Appreciate the biases in your data.

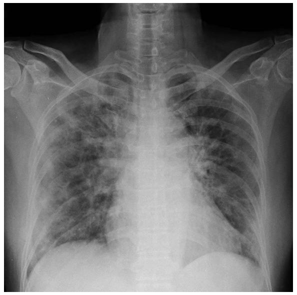


UNIVERSITY OF CAMBRIDGE

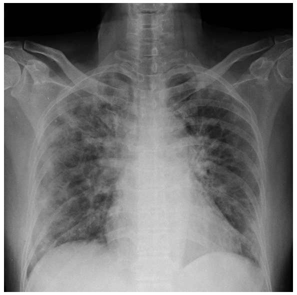
Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

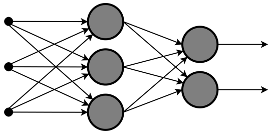
Basic pitfalls: biases in labels

- Know where your data comes from!
- Appreciate the biases in your data.
- Ground truth assigned based on images.



Positive
Negative



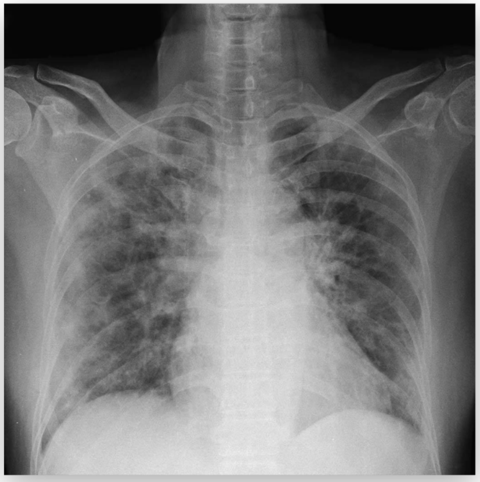


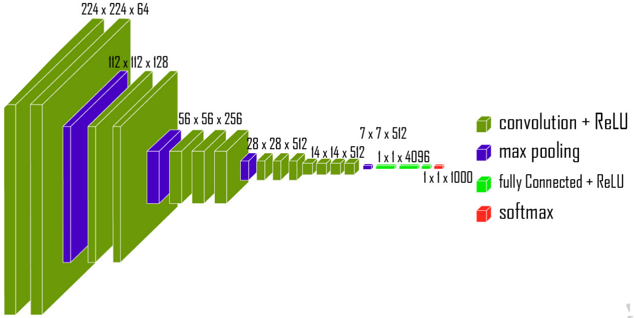
Positive
Negative

Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Basic pitfalls: biases in models

- Know where your data comes from!
- Appreciate the biases in your data.
- Ground truth assigned based on images.
- Resolution driven by pretrained networks



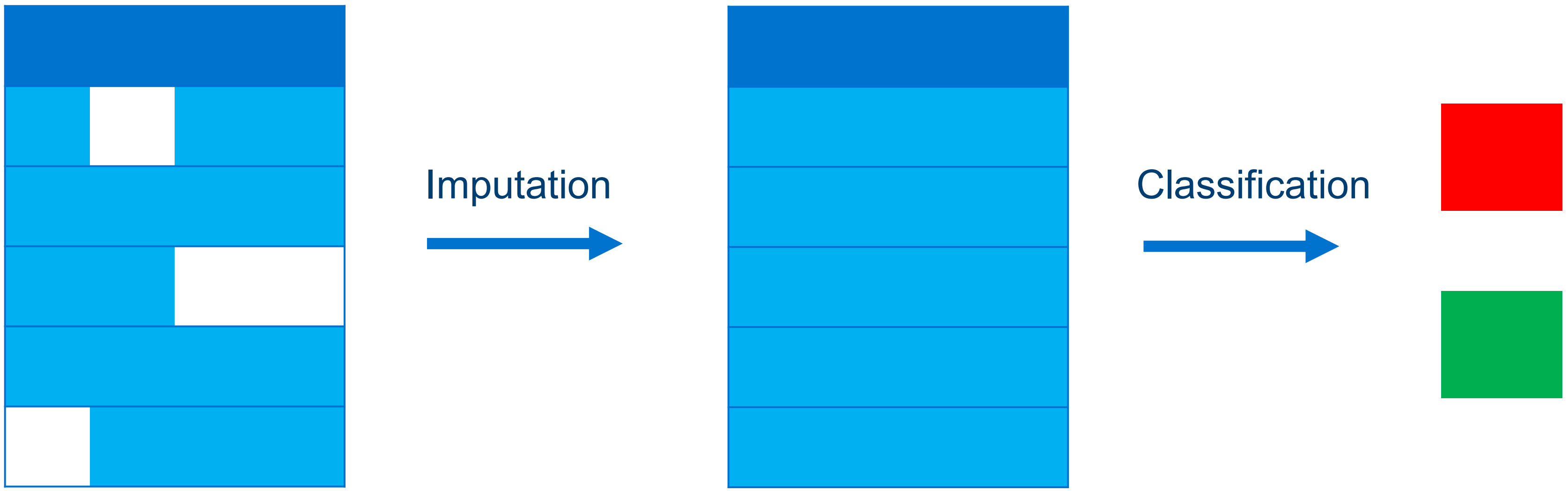


Roberts, M., Driggs, D., Thorpe, M. et al.
Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence 3, 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Issues with …

Joint with Tolou Shadbahr, Julian Gilbey, Jan Stanczuk, Philip Teare, Sören Dittmer, John Aston, Carola-Bibiane Schönlieb and many AIX-COVNET members

ML for Incomplete Data



Shadbahr, T.*, **Roberts, M.***, Stanczuk, J.*, Gilbey, J.*, Teare, P.*, Dittmer, S., ... & Schönlieb, C.-B.
Classification of datasets with imputed missing values: does imputation quality matter? (2022). Under Review.
arXiv preprint arXiv:2206.08478.

email: mr808@cam.ac.uk

Why is imputation important?

- QRISK is a model for predicting your risk of a cardiovascular disease.

[1] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study *BMJ* 2007; 335 :136 doi:10.1136/bmj.39261.471806.55

[2] https://www.bmj.com/rapid-response/2011/11/01/multiple-imputation-needs-be-used-care-and-reported-detail

Why is imputation important?

- QRISK is a model for predicting your risk of a cardiovascular disease.
- Initially, the published paper [1] found no link between cholesterol and outcome.

[1] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study *BMJ* 2007; 335 :136 doi:10.1136/bmj.39261.471806.55

[2] https://www.bmj.com/rapid-response/2011/11/01/multiple-imputation-needs-be-used-care-and-reported-detail

Why is imputation important?

- QRISK is a model for predicting your risk of a cardiovascular disease.
- Initially, the published paper [1] found no link between cholesterol and outcome.
- Other researchers [2] found that when only the complete data was considered, the link to cholesterol was found.

[1] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study *BMJ* 2007; 335 :136 doi:10.1136/bmj.39261.471806.55

[2] https://www.bmj.com/rapid-response/2011/11/01/multiple-imputation-needs-be-used-care-and-reported-detail

Why is imputation important?

- QRISK is a model for predicting your risk of a cardiovascular disease.
- Initially, the published paper [1] found no link between cholesterol and outcome.
- Other researchers [2] found that when only the complete data was considered, the link to cholesterol was found.
- After improving the imputation method, the link to cholesterol was recovered.

[1] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study *BMJ* 2007; 335 :136 doi:10.1136/bmj.39261.471806.55

[2] https://www.bmj.com/rapid-response/2011/11/01/multiple-imputation-needs-be-used-care-and-reported-detail

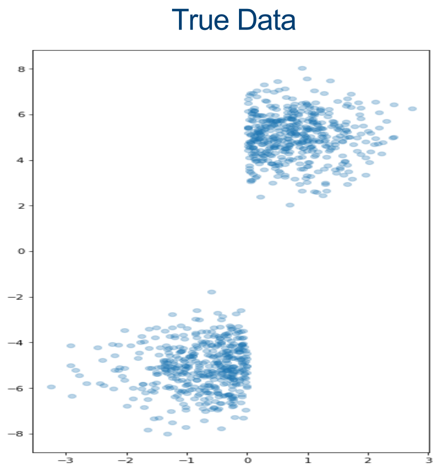
Why is imputation important?

- QRISK is a model for predicting your risk of a cardiovascular disease.
- Initially, the published paper [1] found no link between cholesterol and outcome.
- Other researchers [2] found that when only the complete data was considered, the link to cholesterol was found.
- After improving the imputation method, the link to cholesterol was recovered.
- An improved algorithm is now a standard used in the UK NHS.

[1] Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P et al. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study *BMJ* 2007; 335 :136 doi:10.1136/bmj.39261.471806.55

[2] https://www.bmj.com/rapid-response/2011/11/01/multiple-imputation-needs-be-used-care-and-reported-detail

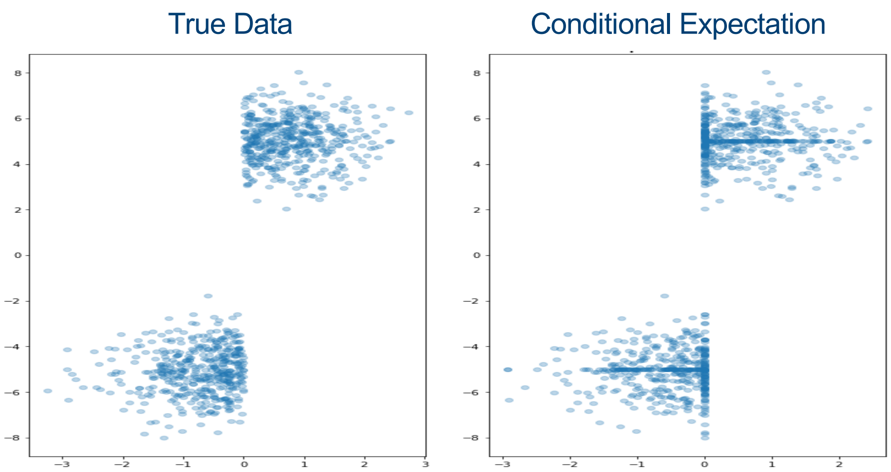
How should we measure quality?



Mean / root mean square error is a common metric for measuring imputation quality.

Shadbahr, T.*, **Roberts, M**.*, Stanczuk, J.*, Gilbey, J.*, Teare, P.*, Dittmer, S., ... & Schönlieb, C.-B. Classification of datasets with imputed missing values: does imputation quality matter? (2022). Under Review. *arXiv preprint arXiv:2206.08478*.

How should we measure quality?



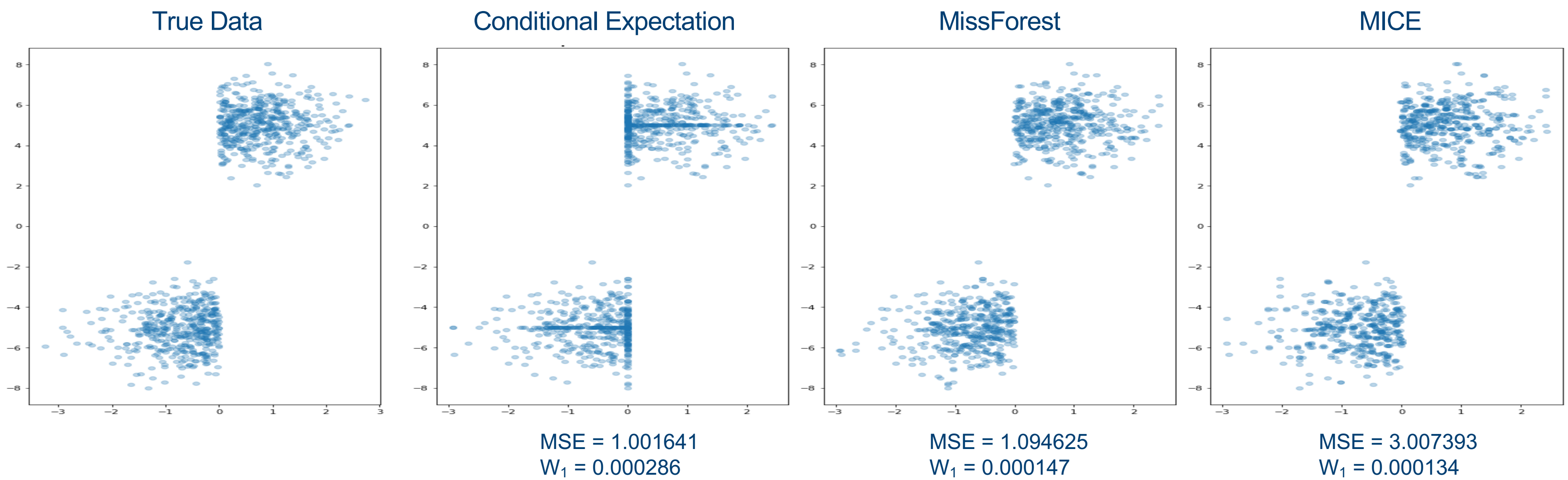
MSE = 1.001641
W_1 = 0.000286

Mean / root mean square error is a common metric for measuring imputation quality.

W_1 = Wasserstein distance between imputed and real value distributions

Shadbahr, T.*, **Roberts, M.***, Stanczuk, J.*, Gilbey, J.*, Teare, P.*, Dittmer, S., ... & Schönlieb, C.-B. Classification of datasets with imputed missing values: does imputation quality matter? (2022). Under Review. *arXiv preprint arXiv:2206.08478.*

How should we measure quality?



Mean / root mean square error is a common metric for measuring imputation quality.

W_1 = Wasserstein distance between imputed and real value distributions

Shadbahr, T.*, **Roberts, M.***, Stanczuk, J.*, Gilbey, J.*, Teare, P.*, Dittmer, S., ... & Schönlieb, C.-B. Classification of datasets with imputed missing values: does imputation quality matter? (2022). Under Review. *arXiv preprint arXiv:2206.08478.*

What are the issues?

- We find that many new methods fail both to:
  - recreate data distributions
  - give stable imputations
- This compromises model interpretability
- Missingness may be informative for the models

Shadbahr, T.*, **Roberts, M.***, Stanczuk, J.*, Gilbey, J.*, Teare, P.*, Dittmer, S., ... & Schönlieb, C.-B. Classification of datasets with imputed missing values: does imputation quality matter? (2022). Under Review. *arXiv preprint arXiv:2206.08478.*

Joint with Sören Dittmer, James Rudd, John Aston, Carola-Bibiane Schönlieb

Issues with …

ML and code

python™

TensorFlow

PyTorch

jupyter

Dittmer, S.*, **Roberts, M.***,., ... & Schönlieb, C.-B.
Navigating the challenges in creating complex data systems: a development philosophy (2022). Under Review.
arXiv:2210.13191.

email: mr808@cam.ac.uk

Marauding as software engineers

10 years ago, this was a software engineer….

email: mr808@cam.ac.uk

Marauding as software engineers

10 years ago, this was a software engineer….

Now it is also a:

- Data scientist
- Mathematician
- Statistician
- Machine learning engineer

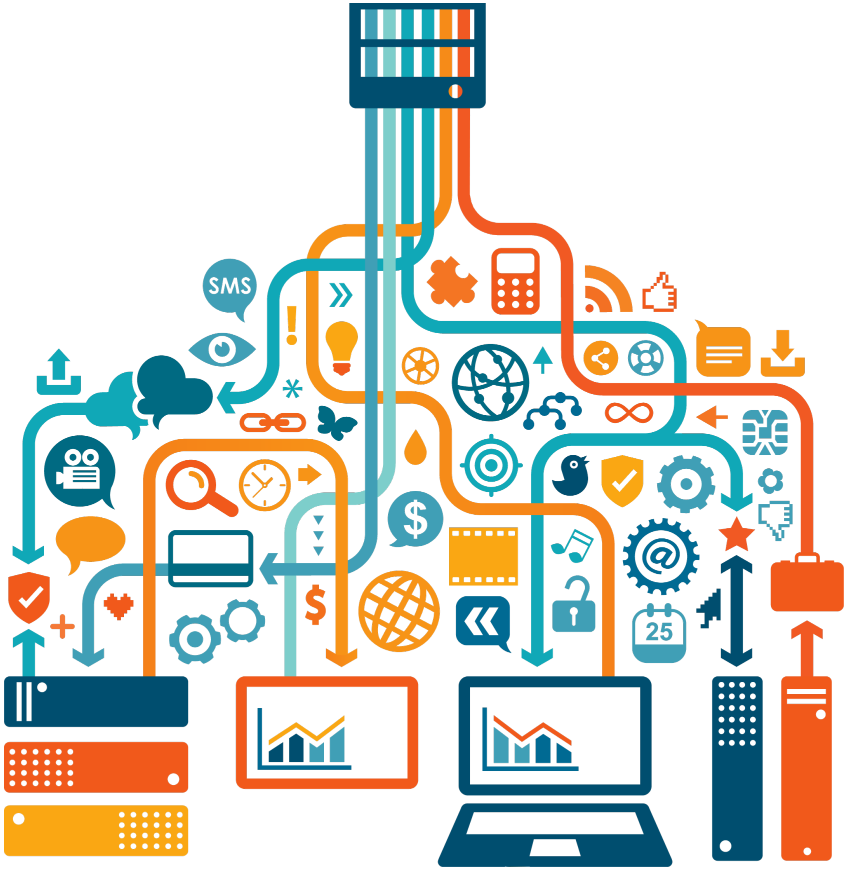
email: mr808@cam.ac.uk

Issues with the new world

An explosion of data and sources

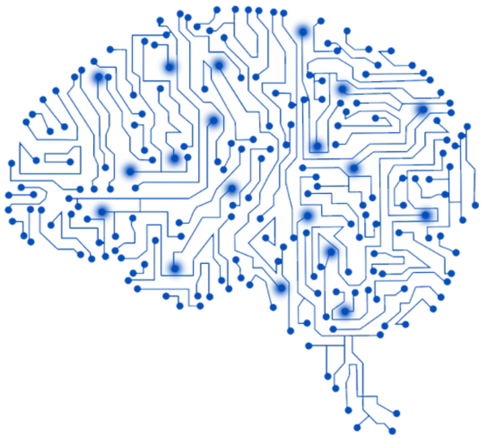
We must often structure and preprocess this data ourselves

Without deep domain knowledge, we have lost control of the biases in the data



email: mr808@cam.ac.uk

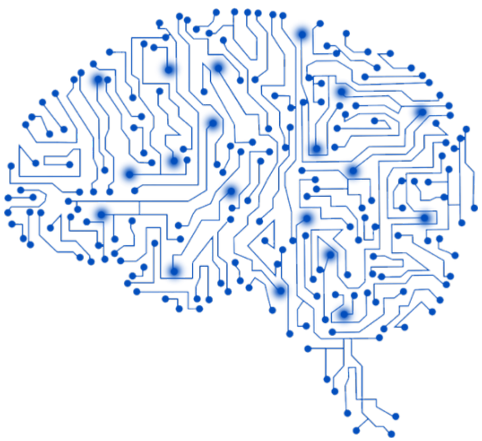
Issues with the new world



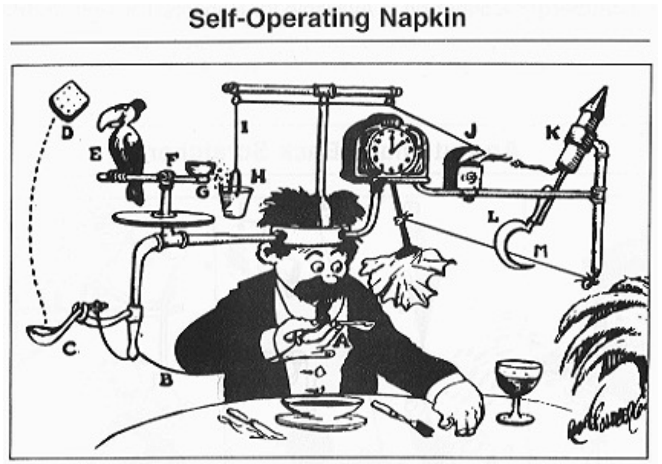
Idealistic illustration of machine learning

email: mr808@cam.ac.uk

Issues with the new world



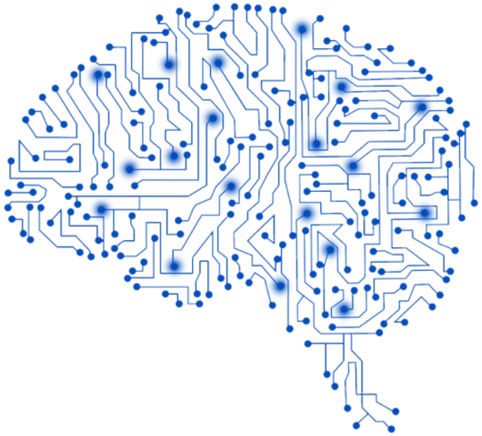
Idealistic illustration of machine learning



The reality …

email: mr808@cam.ac.uk

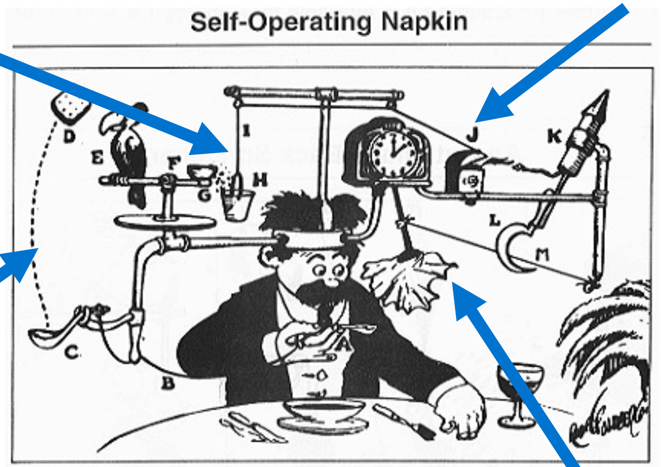
Issues with the new world



Idealistic illustration of machine learning

Data pre-processing

Model training



Data loading

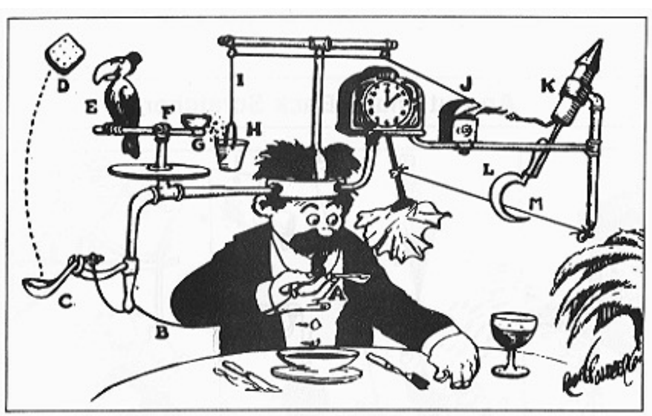
The reality …

Evaluation

UNIVERSITY OF CAMBRIDGE

email: mr808@cam.ac.uk

Issues with the new world



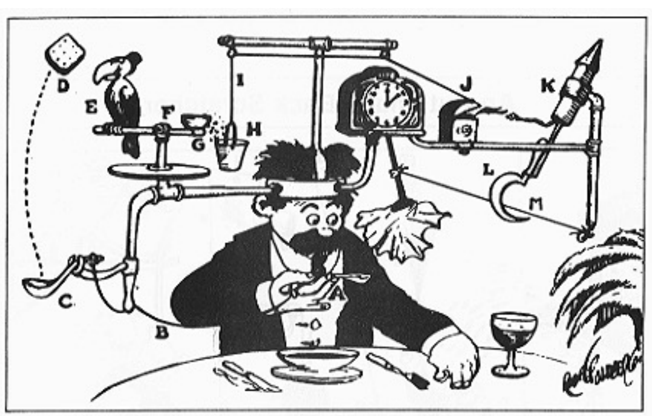
Typical workflow for an ML project

- Code is built as a monolith
- It is tinkered with until it stops giving errors
- Train using all the data and then analyse the results

?

email: mr808@cam.ac.uk

Issues with the new world



Typical workflow for an ML project

- Code is built as a monolith
- It is tinkered with until it stops giving errors
- Train using all the data and then analyse the results

?

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…
- Create a study plan in advance addressing imbalance, power, etc

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…
- Create a study plan in advance addressing imbalance, power, etc
- The literature needs purifying.

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…
- Create a study plan in advance addressing imbalance, power, etc
- The literature needs purifying.
- We all need to use checklists, journals need to enforce their use

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…
- Create a study plan in advance addressing imbalance, power, etc
- The literature needs purifying.
- We all need to use checklists, journals need to enforce their use
- Complex datasets + complex systems = a nightmare for reproducibility

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…
- Create a study plan in advance addressing imbalance, power, etc
- The literature needs purifying.
- We all need to use checklists, journals need to enforce their use
- Complex datasets + complex systems = a nightmare for reproducibility
- Practitioners need to rethink coding methodologies, e.g. AGILE

email: mr808@cam.ac.uk

General recommendations / thoughts

- We need to rethink the incentive structure for machine learning based research.
- Clinical trials are slow, and in phases, for a reason…
- Create a study plan in advance addressing imbalance, power, etc
- The literature needs purifying.
- We all need to use checklists, journals need to enforce their use
- Complex datasets + complex systems = a nightmare for reproducibility
- Practitioners need to rethink coding methodologies, e.g. AGILE
- If you don't enforce your assumptions in the code, someone will break it

email: mr808@cam.ac.uk

Thank you

email: mr808@cam.ac.uk