UWE Bristol | University of the West of England

University of Dundee

Statistical disclosure control for machine learning models

Data Research Access and Governance Network

Susan Krueger

Esma Mansouri-Benssassi

Felix Ritchie

Jim Smith

Additional input: Christian Cole, Emily Jefferson, Simon Rogers, Amy Tilbrook & others

November 2021

Let's celebrate success…

- Trusted research environments (safe havens, secure access etc)
  - the great success story of 21st Century data sharing
- Machine learning
  - Now feasible, accessible and effective
- ML within a TRE?
  - Teach models on very sensitive data
  - Familiar, controlled environment => very safe

The problem

- No SDC guidelines for ML models

⇒Is it the same type of problem as existing models?

⇒If so, can we adapt existing rules?

⇒If not…what?

Machine learning

- Repeated (but unrepeatable) analysis of the same dataset
- Multiple analyses/analytical methods combined
  - Multiple batches/subsets of the data
- Aim: to allow accurate predictions/classifications on new data

Machine learning attacks

- Most common attacks:
  - Model inversion: recreating the data from parameters eg
    - create noisy images
    - minimise prediction error given weights (inverting the learning process)
    - Re-optimise input sample, and repeat
  - Membership: was X in the training data?
    - Models perform better on the data they're trained on
    - Outliers/boundaries can be informative

- Attack scenarios:
  - Black box – access to predictions only
  - White box – access to parameters/architecture

Initial approach: Same sort of problem

- ML models: generate parameters to represent reality
- Statistical models: generate parameters to represent reality

⇒Treat as regressions ('safe statistics')

Problem 1 overfitting

	Regression model	Machine learning model
Type of data	Data table	Images
N	10,000	100,000
K	100	10,000
Estimation method	Deterministic	Non-deterministic
Interest	Parameters	Prediction

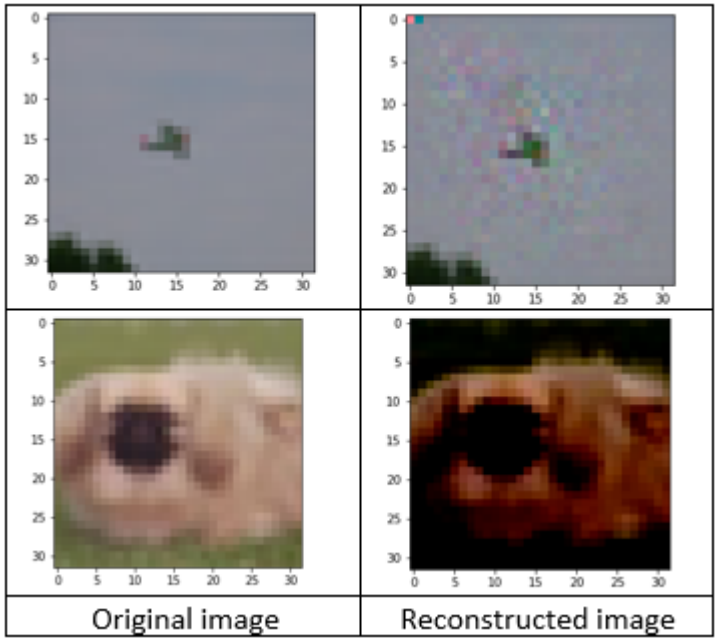
- Technically, doesn't change anything
  - in practice…
  - But no longer human-checkable

Problem 2: what is disclosure?

- Regression: coefficients are of interest
- ML: designed to produce predictions of the whole data item
  - What is the boundary between good prediction and re-identification?

Model inversion attack

16% accurately reconstructed
(9% with DP noise addition)



Membership attack

56% attack accuracy

Problem 3: intruder motivations

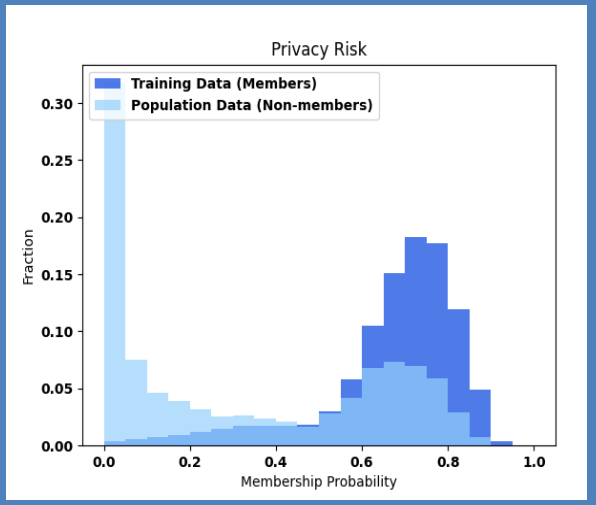
- Regression analyses in RDCs – no incentive for researcher to falsify results
  - Can see and remember interesting data points
    - In theory, incentive in RJSs but risk of capture higher
- What if you want to extract an image?
  - No way to remember it
  - Lots of potential to hide it?

Ways forward

- Can we quantify risk?

Privacy risk for a Blackbox attack

Privacy risk for a Whitebox attack



Calculations using ML Privacy Meter (Murakonda et al, 2020)

Ways forward

- Can we quantify risk?
  - yes, to some extent
- Can we identify statistical solutions?
  - ~~No idea~~ Probably
- Can we identify non-statistical solutions?
  - ~~No idea~~ Probably

- Big problems/uncertainties:
  - Variety of models
  - Motivation of attackers

Next steps

- Greatly expanded team
  - Dundee: Emily Jefferson, Christian Cole
  - Edinburgh: Amy Tilbrook
  - NHS Scotland: Simon Rogers
  - Plus large group of interested parties across academia, govt and health service
    - Seeking funding to develop formal project

Next steps – defining scope

Disclosure Control on trained models – Concept

21.09.21

Intent

To determine risk measures and a set of effective controls for disclosure risk of trained machine learning models from Trusted Research Environments.

Scope

Models trained on data provided and intended to remain in the Safe Haven.

Objectives

Research and define risks, evaluate these against likelihood (given other controls) and measure actual risk. Work these up to practical controls.

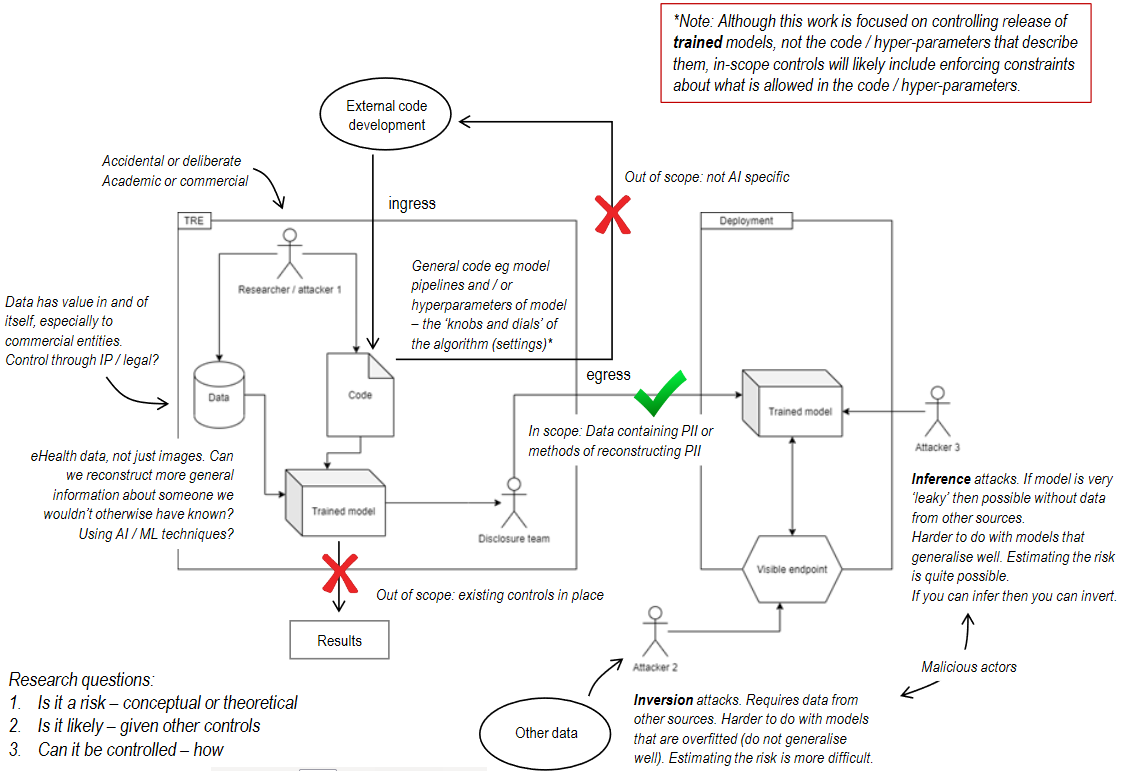
Deliverables

Method of assessing disclosure risk associated with trained models.

Actionable insights / recommendations with evidence to support.

Milestones / Resources

See Plan on next slide.



Next steps - ambition

- December 2021
  - Context: ML baseline risk
    - Benssassi-Manzouri et al 2021 https://arxiv.org/ftp/arxiv/papers/2111/2111.05628.pdf
  - Context: ML operations in TREs
    - Ritchie et al 2021 [link when ready]
- May 2021
  - Methods paper
  - Provisional practice guidance
- July 2021
  - Example use case

Questions?

- University of Dundee:
  - Susan Krueger
  - Esma Mansouri-Benssassi
- University of the West of England
  - Felix Ritchie
  - Jim Smith