Assessing AI Models for Release

Lessons from GRAIMatter and SACRO

The AI-SDC toolkit

Jim Smith, Professor of Interactive AI, UWE

james.smith@uwe.ac.uk

Contents

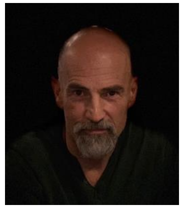
- What were GRAIMatter and SACRO
- What are the risks of AI?
- GRAIMAtter/SACRO recommendations & findings
- The AI-SDC toolkit
- Getting involved in the community

DAREUK HDRUK Health Data Research UK ADRUK UK Research and Innovation SACRO

GRAIMatter project

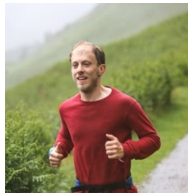
Prof Emily Jefferson (PI): Director of HIC TRE

Prof Felix Ritchie: 5 Safes and Disclosure Control



Prof Jim Smith: AI Models

Dr Christian Cole: Senior Lecturer



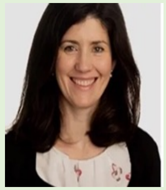
Dr Simon Rogers: Principal Engineer - AI Models

Dr James Liley: Assistant Professor in biostatistics

Law and Ethics

Prof Angela Daly: Regulation and governance of digital technologies, data protection, AI ethics

Dr Francesco Tava: Applied ethics, privacy and trust



Maeve Malone: Lecturer in Intellectual Property law and Healthcare Law and Ethics

Dr Xaroula Kerasidou: Researcher

Dr Smarti Reel

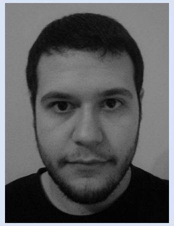
Dr Esma Mansouri-Benssassi

Alba Crespi Boixader

Dr Richard Preen

Andrew McCarthy

Professor Josep Domingo-Ferrer



Dr Alberto Blanco Justicia

International experts

Antony Chuter

Jillian Beggs

PPIE Co-leads

Guidelines and Resources for AI Model Access from TrusTEd Research environments: DARE Sprint exemplar

SACRO partners? (alphabetically)

Universities

- Aberdeen
- Dundee
- Durham
- Edinburgh
- Oxford
- UWE

Public Data Bodies

- Health Data Research UK
- NHS Scotland
- Public Health Scotland
- Research Data Scotland

TREs

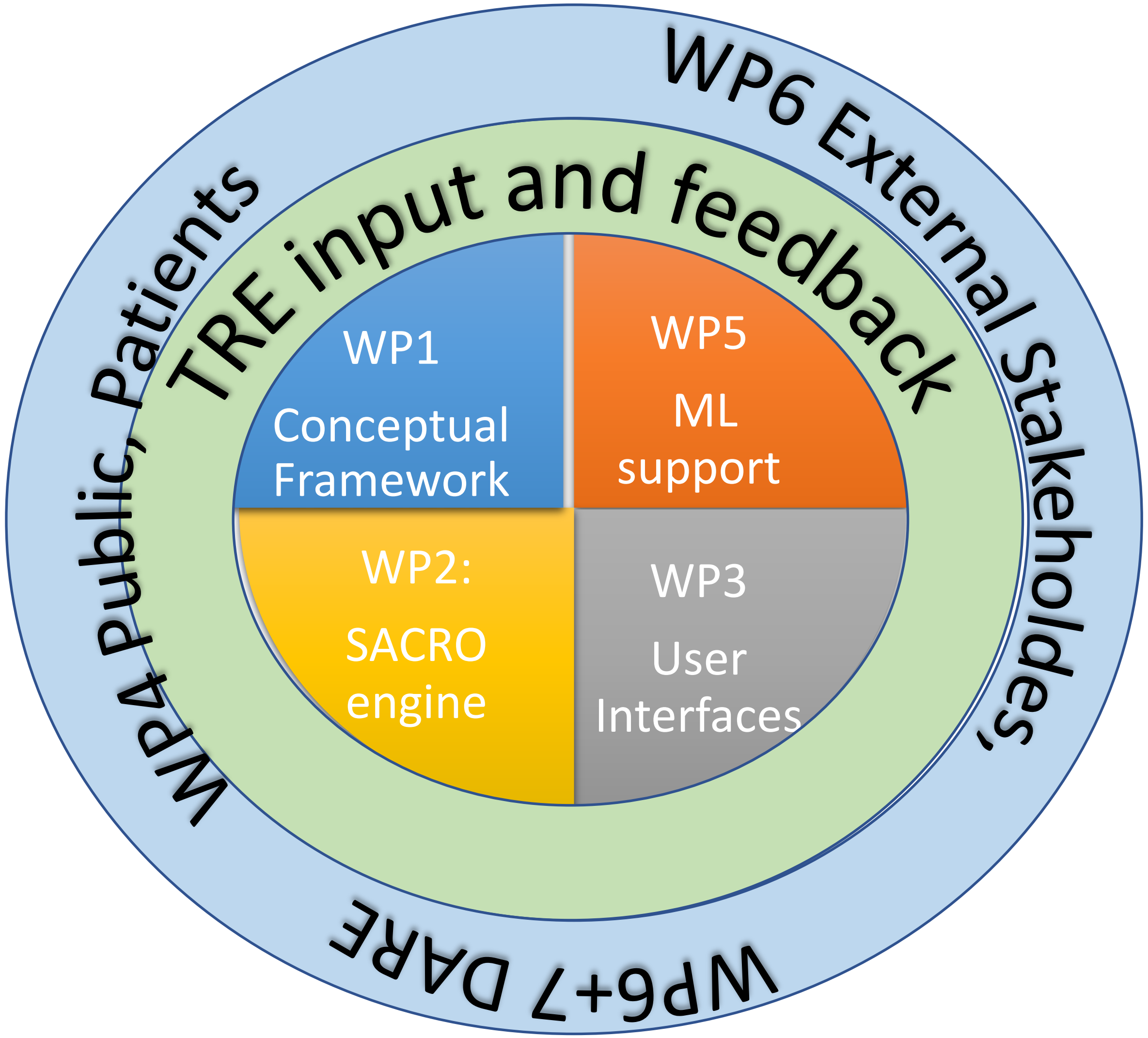
- DASH (Aberdeen/Grampian)
- DataLoch (Edinburgh)
- HIC (Dundee)
- eDRIS (Public Health Scot)
- OpenSafely (Oxford)

External collaborators / steering group:

UK: ONS, NHS-Digital, and SAIL Databank

Global: Eurostat, Bundesbank ,SDC-GESIS, ICPSR (US)

SACRO: DARE Driver project
Semi Automated Checking of Research Outputs



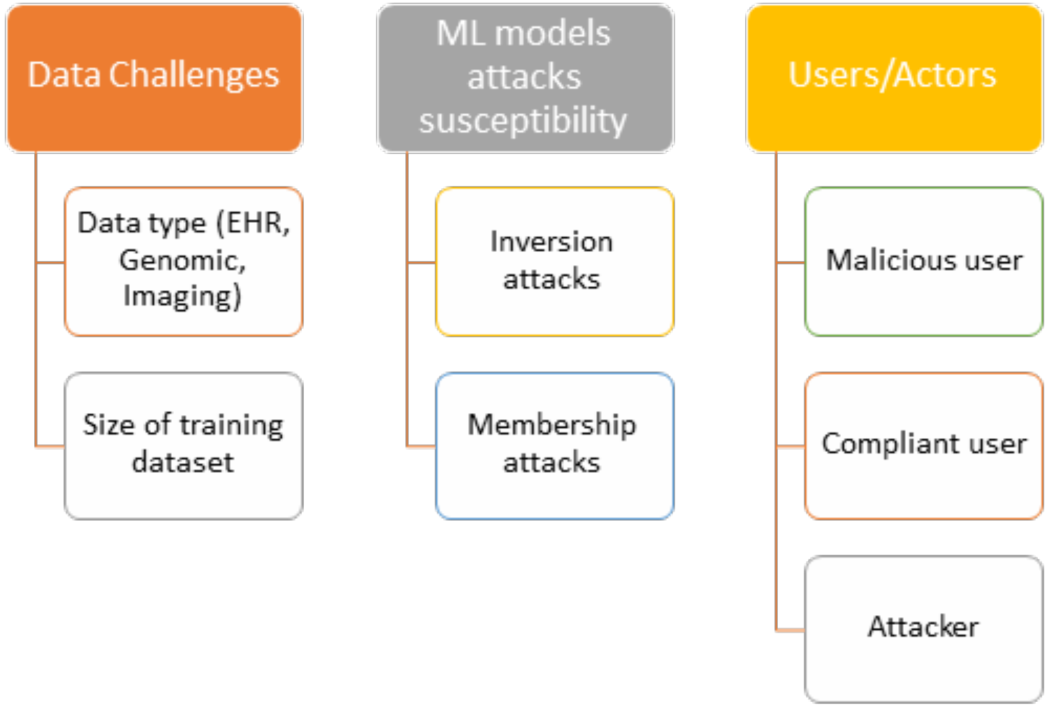
AI-relevant outputs:

- Consensus statement about use of automation
  - Inevitable when outputs are ML
  - HDR UK, ONS, ...
  - Practice & training related expectations
- Refinements to AI-SDC toolkit
  - More attacks
  - Closer links to ‘trad-SDC’ theory
  - Support for ‘user journeys’

Informed by a number of case studies providing advice and support to our partner TREs

(Additional) Risks from AI

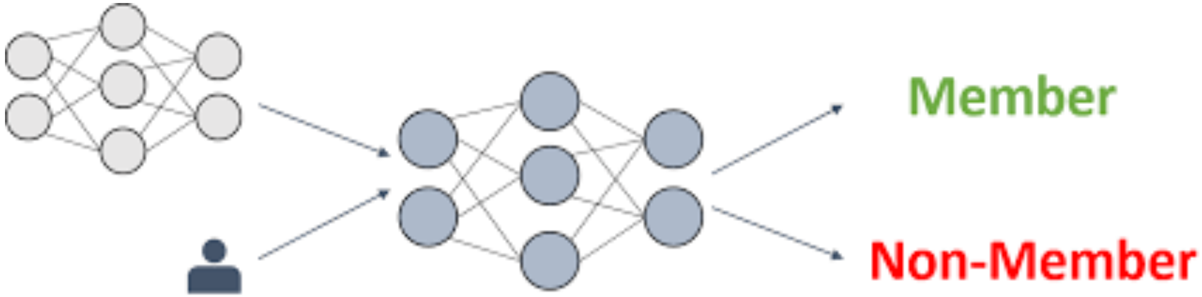
- Malicious user: hide row level identifiable data within exported data arrays
- Non-malicious user: unknowingly train AI model which incorporates training data directly
- Trained models **always** remember aspects of training data; exports can be susceptible to malicious attacks



Disclosure control of machine learning models from trusted research environments (TRE): New challenges and opportunities, Heliyon, Volume 9, Issue 4, 2023, e15143, ISSN 2405-8440, https://doi.org/10.1016/j.heliyon.2023.e15143

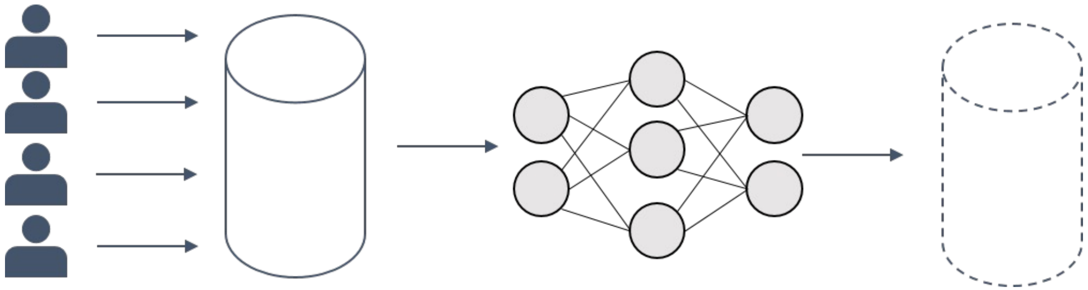
Background: Attacks

- Membership Inference attack: Was?
- Attribute Inference Attack : What?
- Model Inversion attack: Who?



Age	smoker		Diabetes	...
56	y	...	N	...
34	N	...	?	...

DOI:10.1145/2810103.2813677



https://franziska-boenisch.de/posts/2020/12/model-inversion/

Membership Inference Attack Simulation Framework

Target dataset
- Health records
- Medical images
- Synthetic data

Trained model
- Random Forest
- Decision Tree
- SVM
- XGBoost
- etc.

Membership Inference Attack
- Worst Case scenario
- Adversary attack scenarios
- LIRA attack

Attack risk
- AUC
- Attacker advantage
- False Alarm Rate
- New metrics

The gap between ML-Privacy and SDC theory

ML-Privacy Research typically comes from a ‘big-tech’ perspective so

- Asks Different Questions:
  - SDC: If I know X is in the sample what else can I infer from this output?
  - Membership Inference: Can I predict if X was in the sample?
  - Attribute Inference: Are my guesses about X better if X in training set?
- Uses Different Metrics:
  - SDC: Risk to Most Vulnerable Person
  - ML Privacy research: Mean risk to all people
  - Differential Privacy: risk averaged over {people} x {guesses}
- Takes Different Approaches:
  - SDC: Concept of *reasonable* risk based on theory /statistical arguments
  - ML Privacy: empirical results: lots of methodological problems[1,2,3,4]

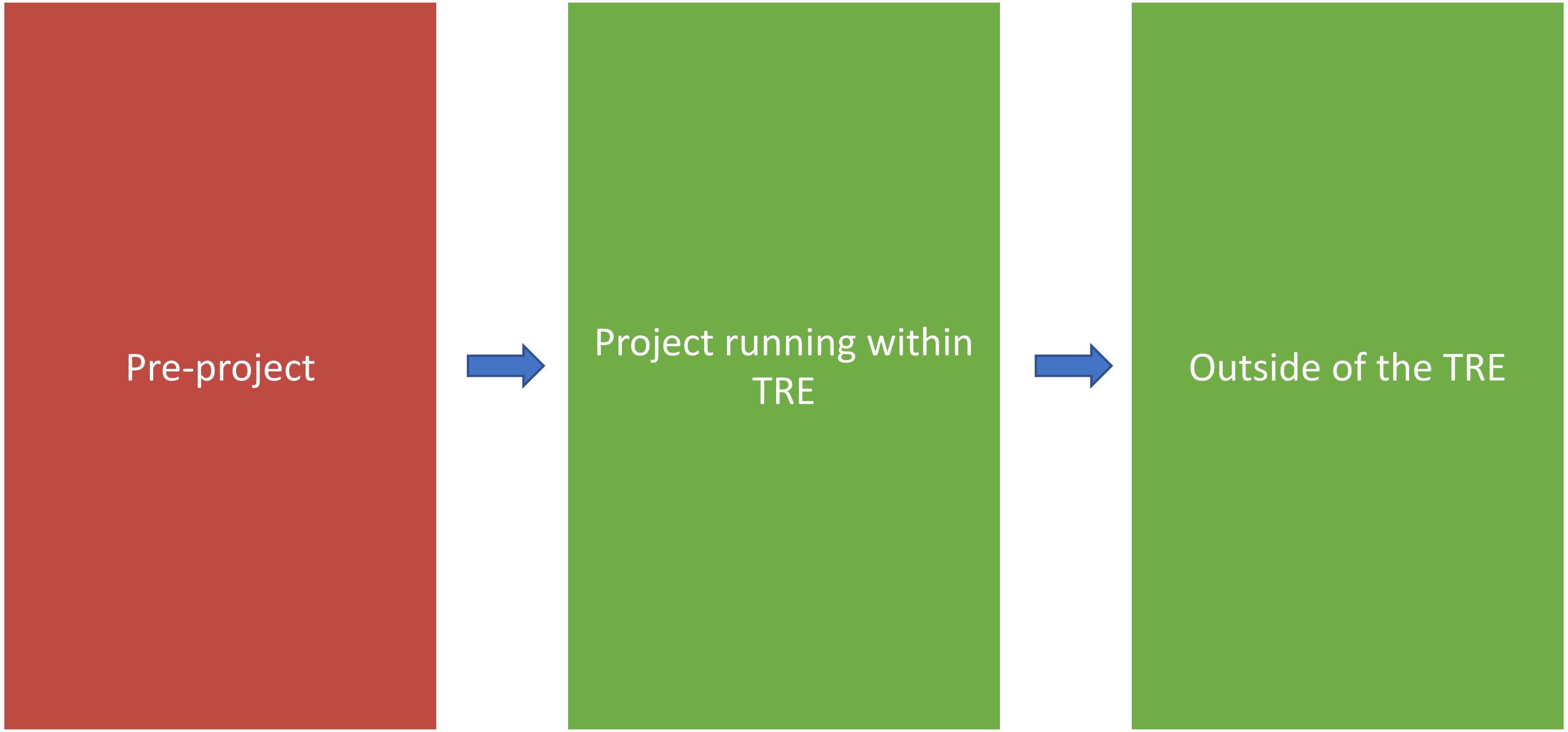
Principle-Based OSDC: It is easy to to say no, when does “not no” mean “ok”?

GRAIMATTER key recommendations

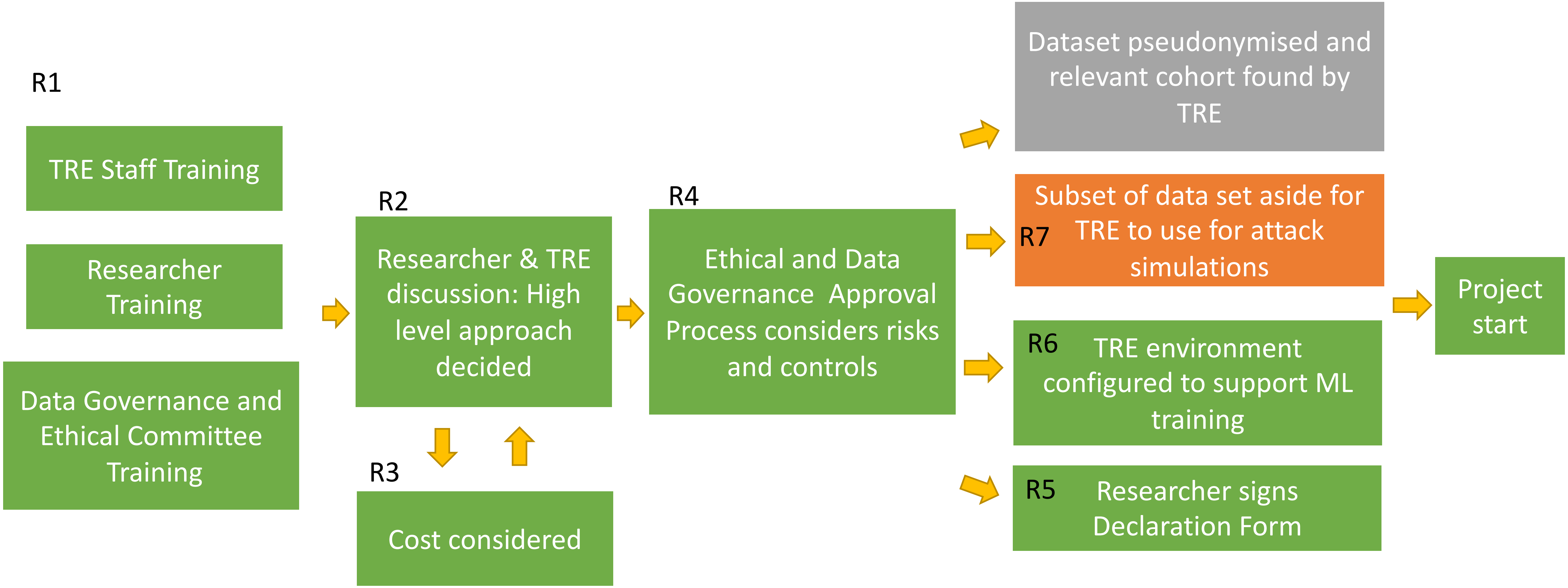
1. Discussions about SDC need to begin during project inception
    1. So a coherent case can be made to approval boards (PBPP etc.)
    2. Because it may rely on some data being set aside for risk assessment
    3. Deployment scenario: Model Disclosure Controls vs Model Query Controls
    4. Type of model proposed
    5. Preprocessing vs deep learning?
2. (Amended) Legal agreements may be needed.

Jefferson, E. *et al* (2022)
'GRAIMATTER Green Paper: Recommendations for disclosure control of trained Machine Learning (ML) models from Trusted Research Environments (TREs)'.
Zenodo. doi: 10.5281/zenodo.7089491.

Categorised the 13 recommendations based on the stage in the project life cycle



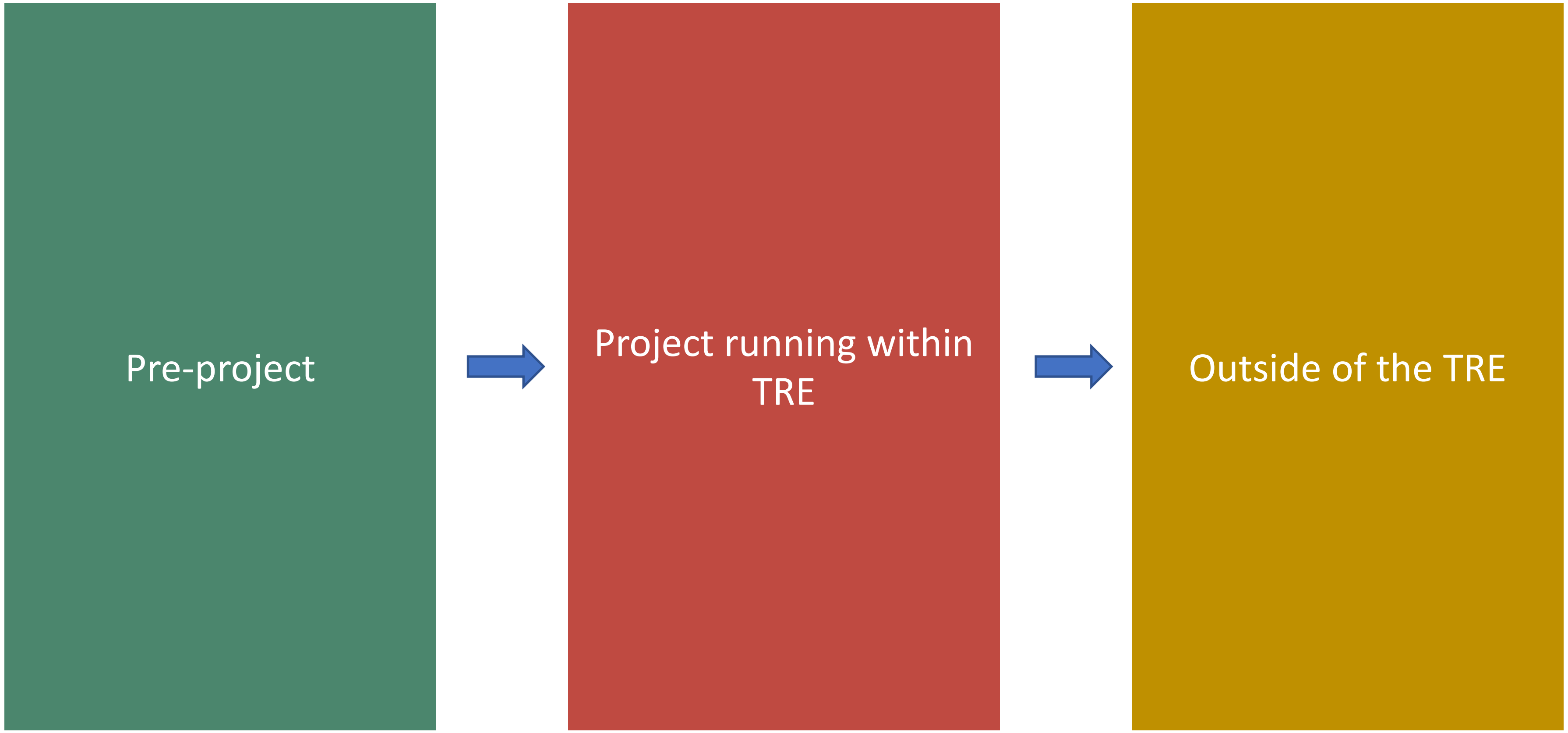
Processes: Pre-Project



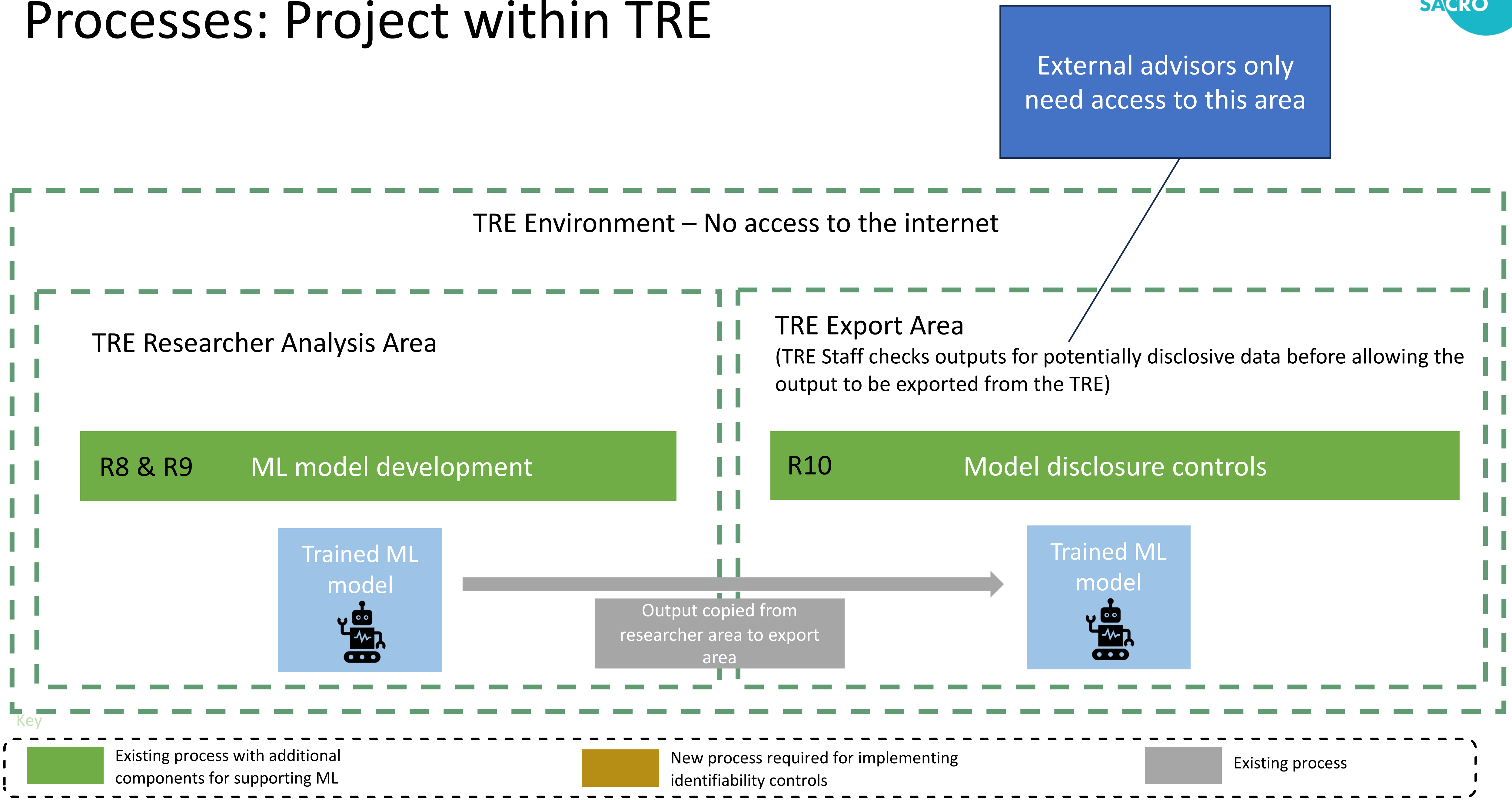
Key

Existing process with additional components for supporting ML	New process required for implementing identifiability controls	Existing process

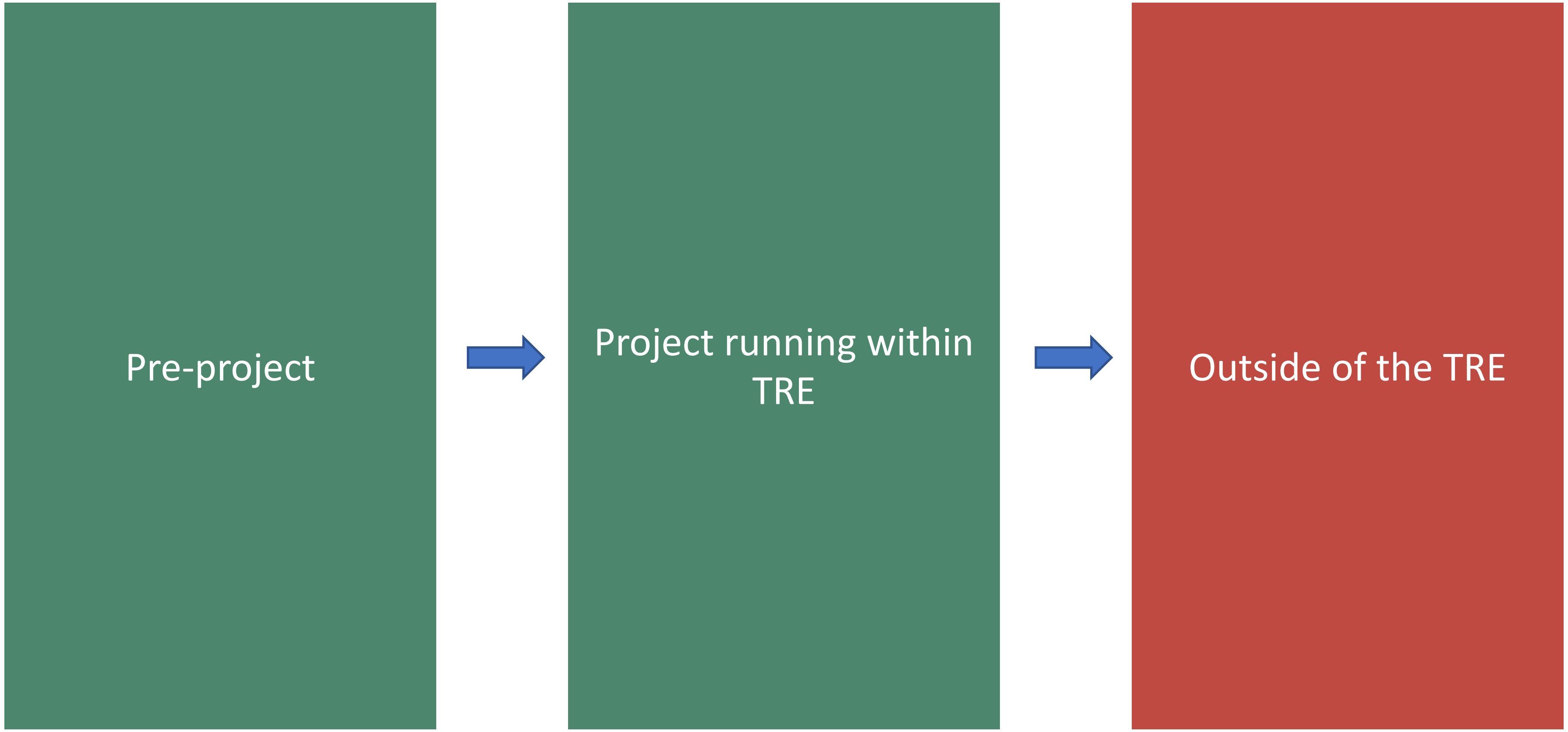
Explaining each recommendation based on the stage in the project life cycle



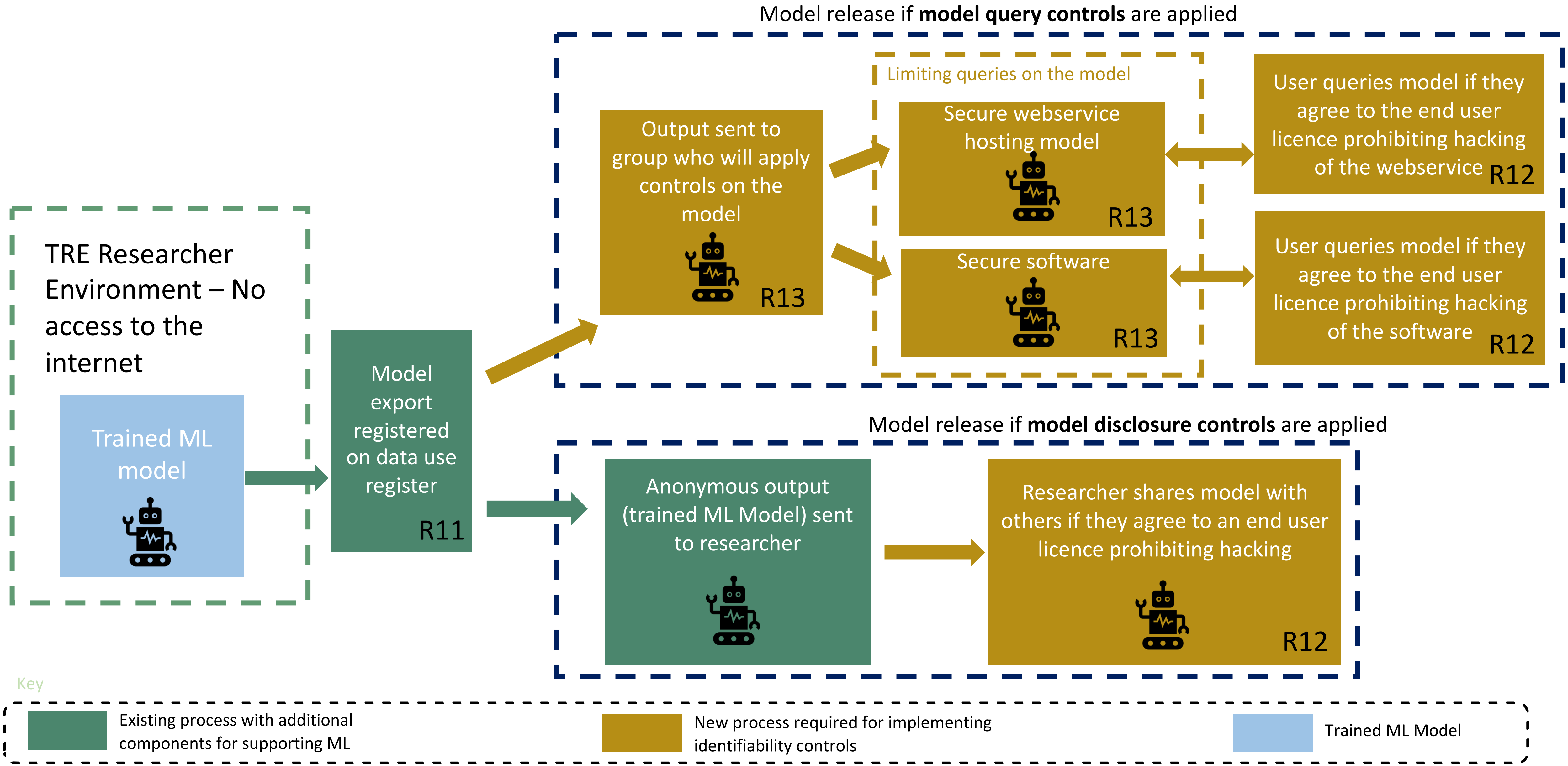
Processes: Project within TRE



Explaining each recommendation based on the stage in the project life cycle



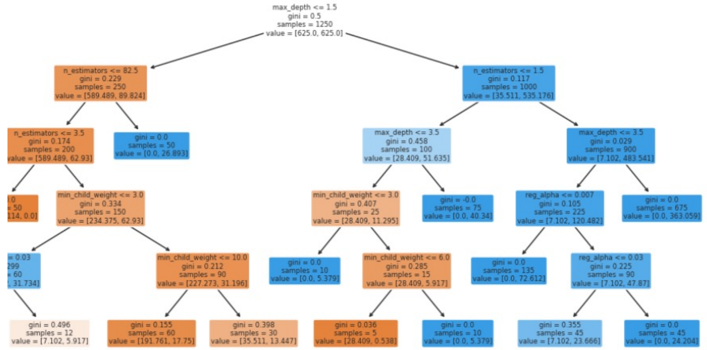
Processes: Model Release



DAREUK HDRUK Health Data Research UK ADRUK UK Research and Innovation SACRO

Predicting 'Unnecessary Risk':xgboost example

Train to predict



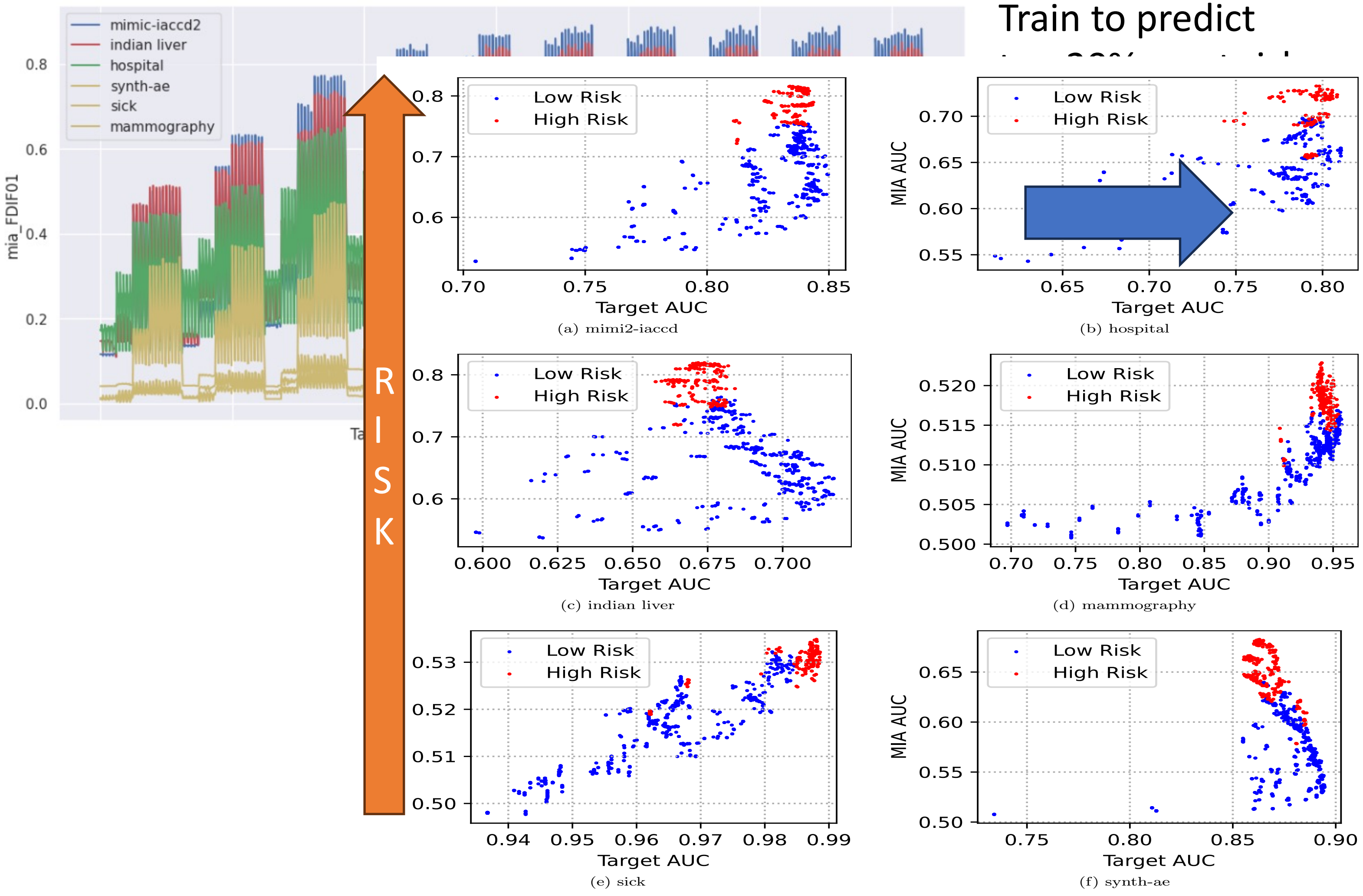
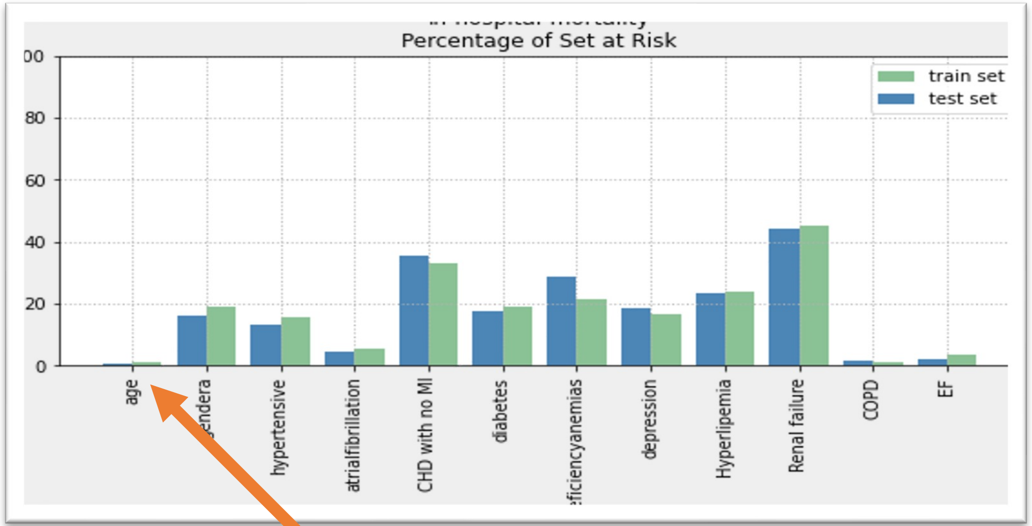


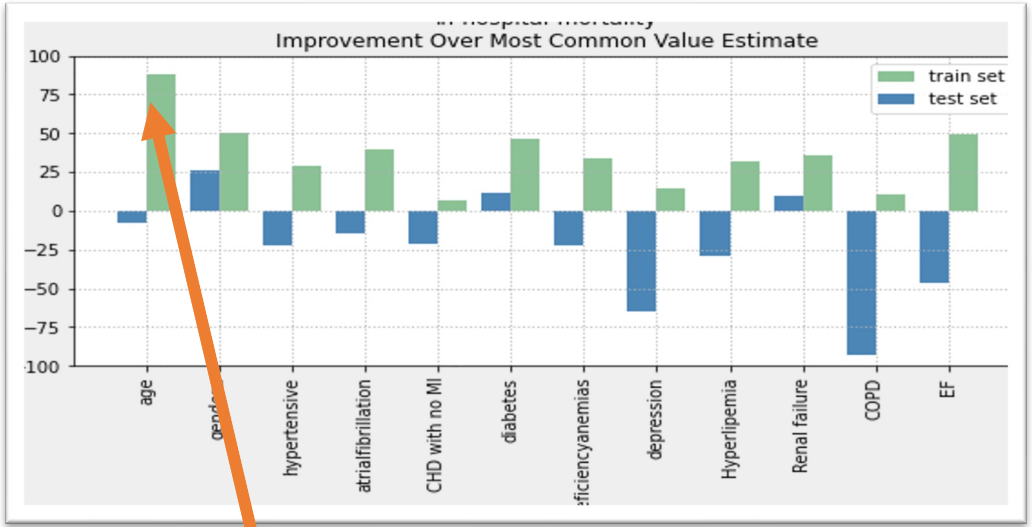
Figure 3: XGBoost Classifier Target Model

QUALITY

GRAIMAtter findings:
Attribute Inference Attacks work(sometimes)

Target is 'naïve' random forest trained on hospital mortality
Contrast results for records used to train target model (green) or not (blue)





Not all records are vulnerable to inference
- Sometimes attack says 'don't know'

Inference accuracy
- Up to 100% on training set data
- Worse than baseline for test data

Graimatter: proof of concept ‘SafeModel’ wrappers

Python wrappers around common algorithms

- Set parameters to “safe” values when model is created.
- Chooses Differentially Private version of algorithm if available

Researcher uses them just like the version they are used to

- But then calls request_release()
  - Checks for common user errors
  - Produces report for TRE output checkers

- SafeDecisionTree()
- SafeRandomForest()
- SafeSVC()
- SafeKeras()

GRAIMatter created prototypes to explore the concept and develop guidelines for how wrappers can / should work.

AI-SDCtoolkit: Attacks and metrics available

Structural Attacks

- K-anonymity
- **Degrees of freedom**
- Class Disclosure Risk (2 variants)
- Unnecessarily Risky hyper-parameter combination

User-behaviour Attacks

(built in to SafeModel wrappers)

- Unsafe Hyper-parameters?
- Failure to use DP optimizer?
- Manual changes to model or hyper-params?
- Optimizer object included in DNN?

Attribute Inference Attacks

- Single most likely value (categorical)
- Prediction within +/- upper/lower acceptance bounds of actual (continuous)
- Report increase in vulnerability for train vs test

Membership Inference Attacks

- Likelihood Ratio
- Worst-Case MIA
  - Based on probabilities

MIA Metrics

- Advantage
- Generalisation error of target
- TPR,FPR ... & derived stats (AUC, pAUC)
- TPR@low FPR[1]
- **PDIF/FDIF:** focussed on extremes of attack confidence

SACRO focus:
Making the AI-SDC 'attacks' easier to use

USER JOURNEYS

'Best Case'

- Instance of safemodel*X* classifier
- preprocessing code
- training and test set

LOTS of risks we can:

- assess
- Possibly rule out

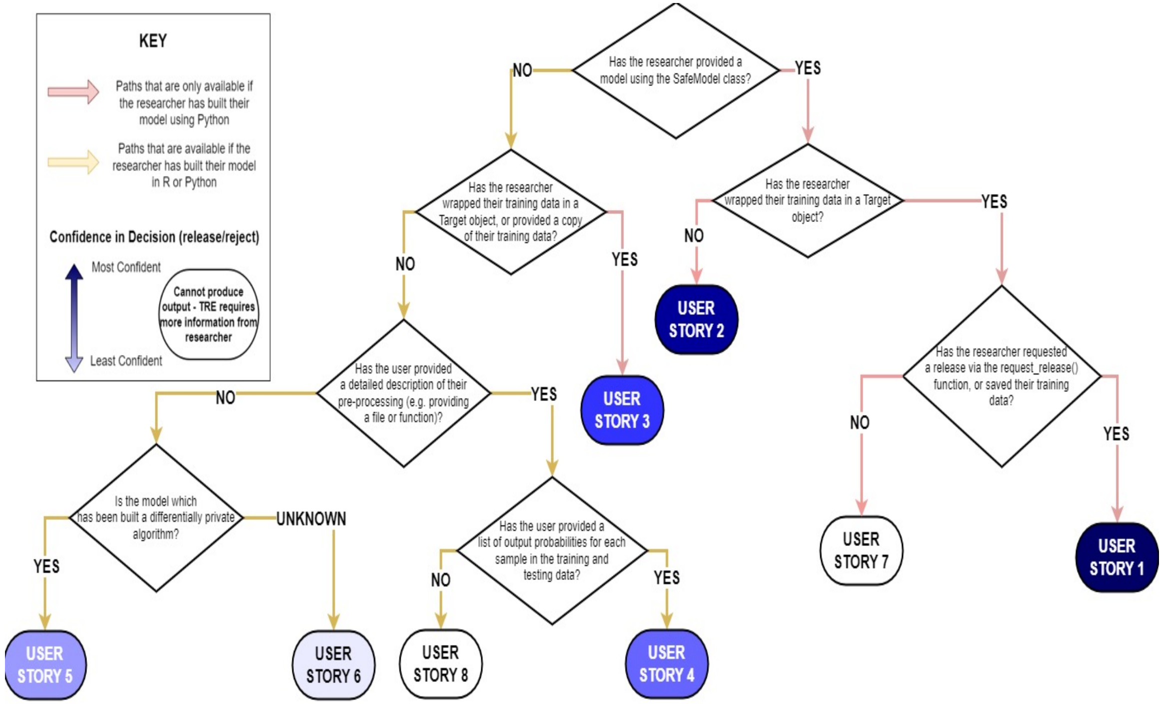
'Worst Case'

- Model created from some non-standard library
- No way to replicate preprocessing
- No details of training/test split

Very hard to

- **run tests**
- **recommend release**

AI-SDC: User Stories



Easy to configure python scripts for most case cases

- Run as many tests as they can given info available
- Gather ***lots*** of metrics about
  - Target model performance
  - Attack model performance
- Produce a summary report

So it’s all sorted then …?

No,

but we’re getting there

Gaining a better understanding of:

- Causes of vulnerability
- How to describe risk?
- Role of PET technologies
- What is ‘sufficient preprocessing’?

Join the aisdc community?

- SDC-Reboot@jiscmail.ac.uk DARE funded Community of Interest
  - Covers all things 'automated checking'
  - So necessarily covers all things relating to assessing AI models
  - ML focussed workshop 7th February:
- https://github.com/AI-SDC/AI-SDC
  - All the ai-sdc tools and 'user stories' scripts
  - Suggest improvements
  - Contribute code -(pytorch anyone?)

Thanks for listening

References

1. Carlini, Nicholas, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 'Membership Inference Attacks From First Principles'. arXiv, 12 April 2022. http://arxiv.org/abs/2112.03570.
2. Hintersdorf, Dominik, Lukas Struppek, and Kristian Kersting. 'To Trust or Not To Trust Prediction Scores for Membership Inference Attacks'. arXiv, 24 January 2023. http://arxiv.org/abs/2111.09076.
3. Rezaei, Shahbaz, and Xin Liu. 'On the Difficulty of Membership Inference Attacks'. arXiv, 22 March 2021. http://arxiv.org/abs/2005.13702.
4. ———. 'On the Discredibility of Membership Inference Attacks'. arXiv, 28 April 2023. http://arxiv.org/abs/2212.02701.