



DataLoch

ML/AI Model Development in the DataLoch Process

July 2023

Machine Learning (ML) is “a subset of Artificial Intelligence, that automatically learns patterns from datasets. It can be used to help humans better understand complex data, or make predictions based upon new, unseen data”.

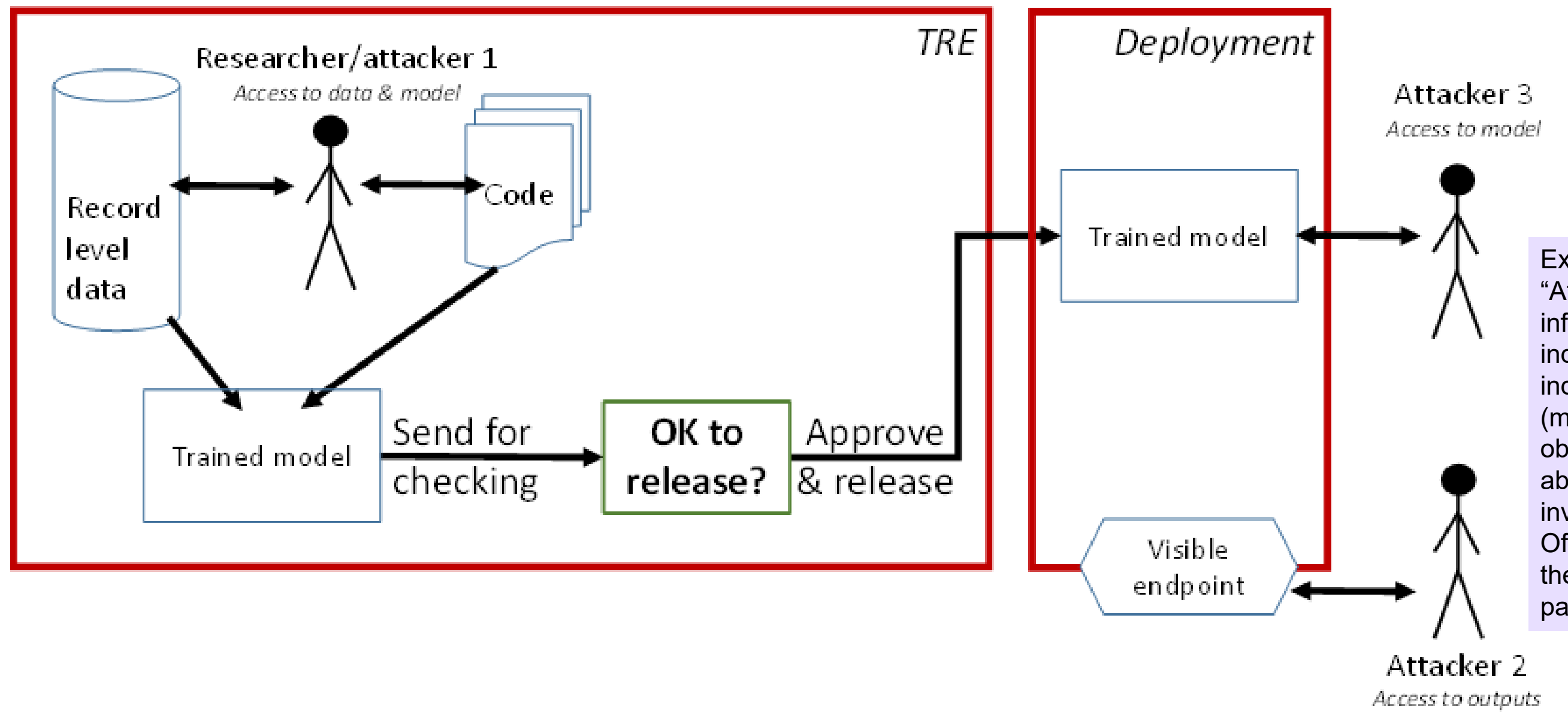
Unlike traditional statistical research models, where the method is specified by the researcher, ML models are provided with an approach to learning and goals and left to work out the method.

The models repeatedly interrogate the data, often in multiple stages and possibly with multiple learning approaches. The resulting model (the reason the ML process ends up with a model configured in a particular way) may not be understandable or fully explainable even by the model designer.

DataLoch have had several approaches to support the development of AI/ML for different uses – usually as part of a pathway to utilising the model on live NHS data to support care. DataLoch mostly supports initial model development/testing if models CAN be developed on data.

AI/ML Model Disclosure Risks

Researcher 1 Risks: Trusted Researchers Doesn't know about risks/issues. Accidentally designs model to be disclosive/include disclosive information



External malicious actor risks: "Attack" model to infer information about one or more individuals – either that an individual is within the dataset (membership inference) or obtain further information about an individual (model inversion/attribute inference). Often done by reconstructing the model using the parameters

Figure 4 Summary of TRE output scenarios and attack possibilities

AI/ML Model Disclosure Challenges – reconstructing data (image example)

Image used to train ML model
(part of training dataset)

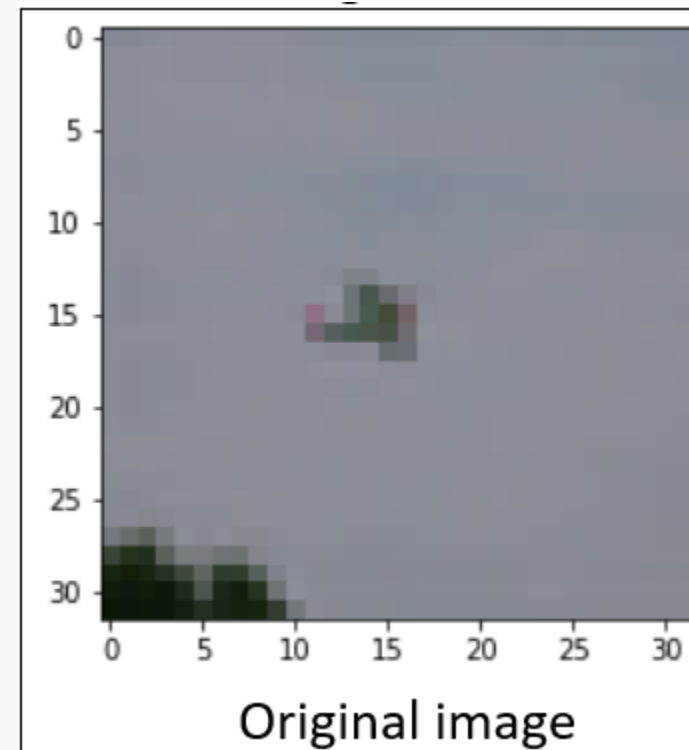
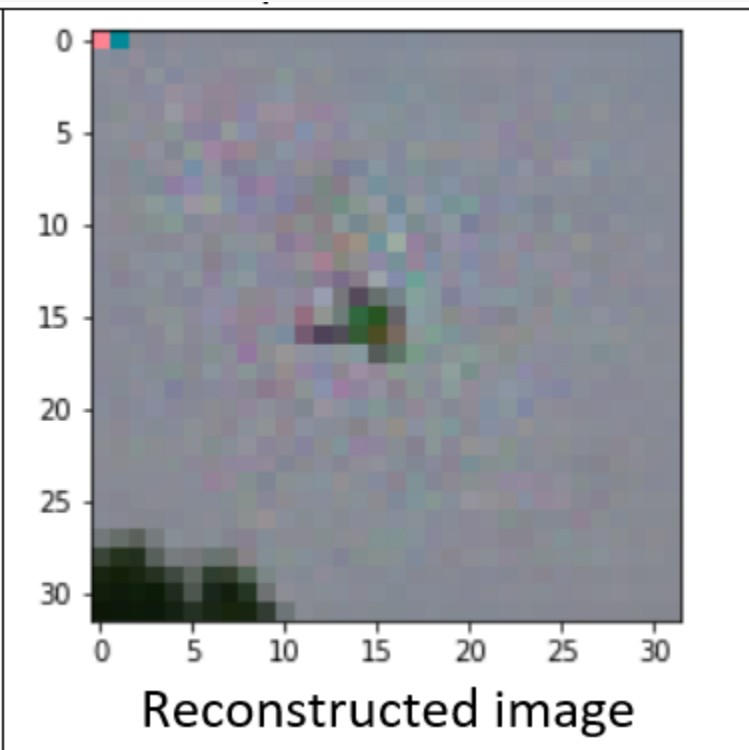
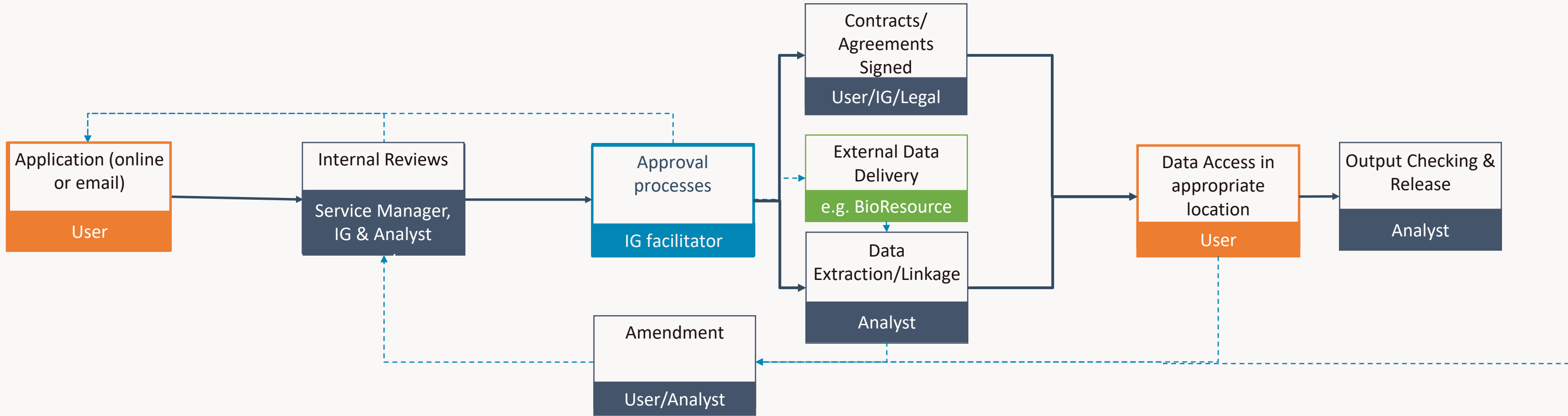


Image reconstructed using model
outputs by ethical hacker



1 Theoretical – what is disclosure? Is the picture on the right “identifiable” as the picture on the left?

Project Delivery Process - current



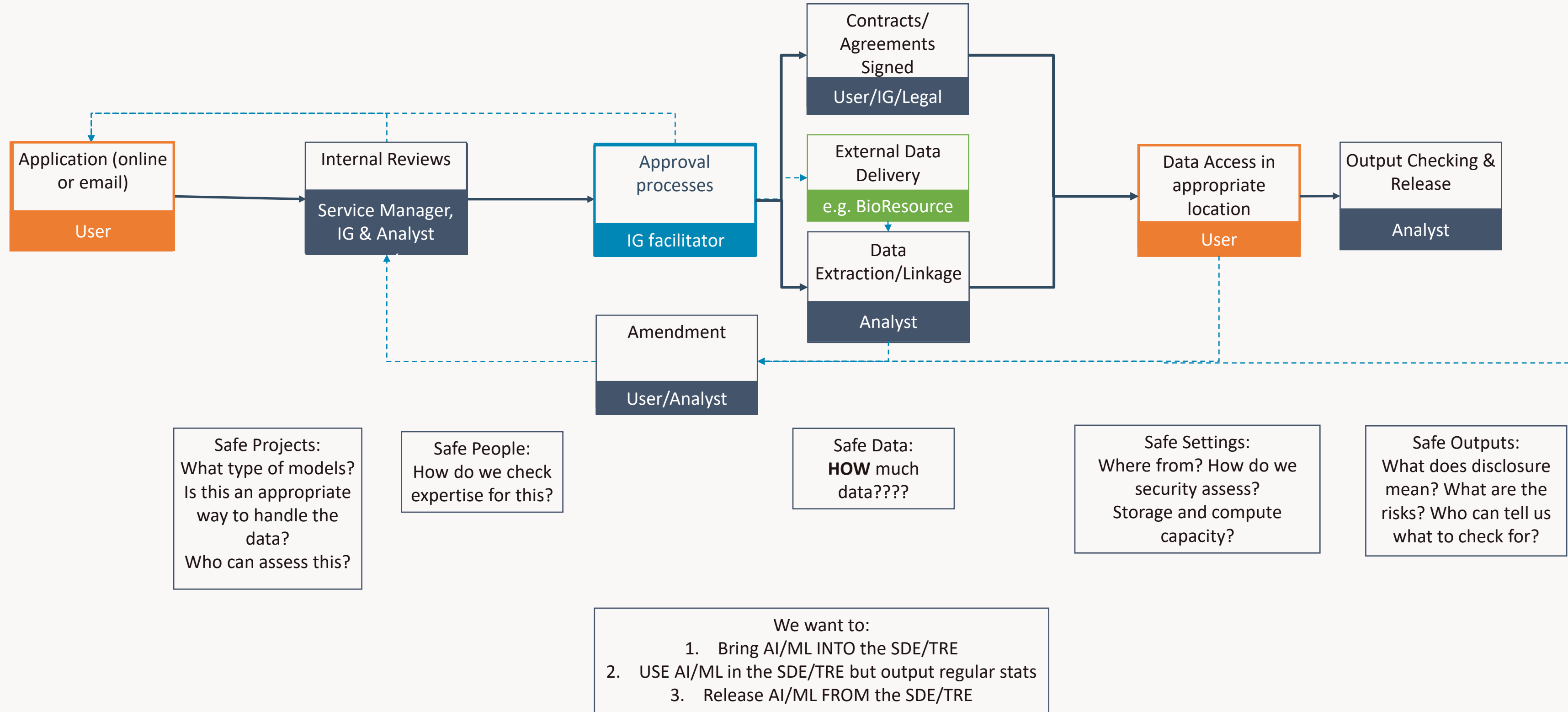
Researcher requests - We want to e.g.:

- Create a dashboard of X features about an individual
- Understand and predict risks of developing Y disease
- Model how to identify Z condition for future work/clinical trials
 - Validate models from elsewhere on local data

Translation in IG terms: We want to:

1. Bring AI/ML INTO the SDE/TRE
2. USE AI/ML in the SDE/TRE but output regular stats
3. Release AI/ML FROM the SDE/TRE

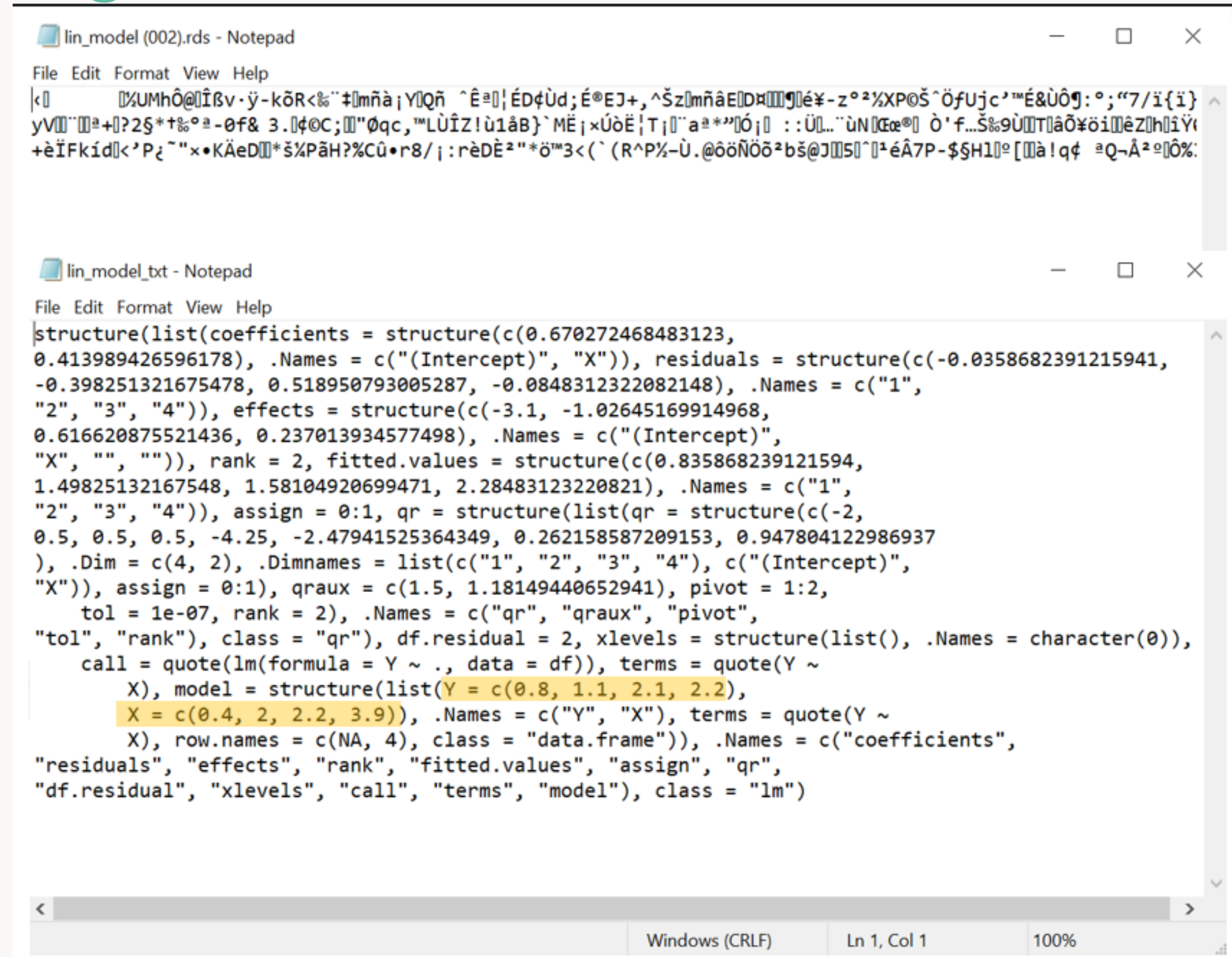
Project Delivery Process - current



AI/ML Model Disclosure Challenges

Model as binary (machine readable) file

Same model as human readable file – some models encode data within – without seeing the final model in this form, we wouldn't know.



```
lin_model (002).rds - Notepad
File Edit Format View Help
|<|      |%UMhÔ@|Î|Û·ÿ-kõR<%~#|mñà|Y|Qñ ^Ê|ÉDçÙd;É®EJ+,^Šz|mñâE|D#|é¥-z°²%XP@Š^ÖfUjc'™É&ÙÔ¶:°;“7/i{i}
yV|+|?2§*+|°-øf& 3. |ç|C;|q|c,™LÛÍZ!ù1âB} `MË|×ÙòÈ|T|`a°**|Ó| : :Ü...`ùn|æ| ò'f...Š%9Û|T|â|õ|ø|i|èz|h|i|Y|
+è|Fkíd|<'P;~"×•KÄeD|*š|PãH?%Cû•r8/|:rèDÈ²"*ö°³<(`(R^P%-Ù.@ôöÑÖö²bš@J|5|^°¹éÁ7P-$§H1|°[|à!qç ³Q-Å²°|Ô%:

lin_model_txt - Notepad
File Edit Format View Help
|structure(list(coefficients = structure(c(0.670272468483123,
0.413989426596178), .Names = c("(Intercept)", "X")), residuals = structure(c(-0.0358682391215941,
-0.398251321675478, 0.518950793005287, -0.0848312322082148), .Names = c("1",
"2", "3", "4")), effects = structure(c(-3.1, -1.02645169914968,
0.616620875521436, 0.237013934577498), .Names = c("(Intercept)",
"X", "", "")), rank = 2, fitted.values = structure(c(0.835868239121594,
1.49825132167548, 1.58104920699471, 2.28483123220821), .Names = c("1",
"2", "3", "4")), assign = 0:1, qr = structure(list(qr = structure(c(-2,
0.5, 0.5, 0.5, -4.25, -2.47941525364349, 0.262158587209153, 0.947804122986937
), .Dim = c(4, 2), .Dimnames = list(c("1", "2", "3", "4"), c("(Intercept)",
"X")), assign = 0:1), qraux = c(1.5, 1.18149440652941), pivot = 1:2,
  tol = 1e-07, rank = 2), .Names = c("qr", "qraux", "pivot",
"tol", "rank"), class = "qr"), df.residual = 2, xlevels = structure(list(), .Names = character(0)),
  call = quote(lm(formula = Y ~ ., data = df)), terms = quote(Y ~
  X), model = structure(list(Y = c(0.8, 1.1, 2.1, 2.2),
  X = c(0.4, 2, 2.2, 3.9)), .Names = c("Y", "X"), terms = quote(Y ~
  X), row.names = c(NA, 4), class = "data.frame"), .Names = c("coefficients",
"residuals", "effects", "rank", "fitted.values", "assign", "qr",
"df.residual", "xlevels", "call", "terms", "model"), class = "lm")
```

2 Practical – how and what can we check?

Application form updated to include sufficient ML/AI detail for Risk Assessment discussions

Governance/Ethics Reviews

- Training of Reviewers
- Consider independent review
- Understand and use risk assessment
- Data Use Register (Extended) updated

Contracts/Agreements

All projects

- Updated Org Framework Agreement: Controls after release
- Updated/Extended User Agreements/Conditions of Use

Some projects

- Project Specific DL/R confidentiality agreements (re IP)
- Project Specific - End User licence for model
- Bespoke project DSAs (if models contain personal data)

Application (online or email)
User

Internal Reviews/ App Refinement
User, SM, IG, Analyst

Approval processes
IG facilitator

Contracts/Agreements

- IG/User/Legal
- External Data Delivery
e.g. BioResource
- Data Extraction/Linkage
Analyst

Recommended additions for ML

Data Access in appropriate location
User

Output Checking & Release
Analyst

Release to public domain
User

Release to another SDE/Host in DL?
Analyst

Pre-Application/General Recommendations.

- Find ongoing AI/ML expertise to support
- Training of TRE Staff in basics of ML and risks
- Understand likely tools/infrastructure required and agree with EPCC
- Agree process and standard controls with Data Controllers
- Refine costing model

Out of DL control

- 10.2.10: Safe wrapper principles developed.
- 12.2.2 More funding provided to DL for this
- 11.2.5: Researchers complete training on AI/ML - including risks & controls

Amendment
User/Analyst

Risk Assessment

Documented by researcher:

- Benefits of model release
- Types of model
- Location of model release (TRE/not)
- Model life expectancy
- Use controls when model released
- Data/pipeline archiving requirements
- Tools required to develop models
- Legal implications

Documented by DL and agreed with researcher:

- Types of controls/process DL will implement for this project (standard/bespoke)
- Overall risks based on above (using matrix from GRAIMatter)
- Updated costs (based on ML tech and disclosure requirements)

10.2.6 Analyst keeps aside % of relevant data for risk assessment

Researcher Analysis/Model Creation

- Agreed Tools /Infrastructure provided to train ML models
- Researchers use agreed controls e.g.
 - Safe wrappers
 - Differentially private methods
 - Aggregate data training (instance based models)
- Researchers provide data dictionary for any model to be released (and base models)

Output Checking

- DL runs tests on model AND base models to check risk – using [checklist](#)
 - Run model against % set aside data
 - Check model performs as expected
 - Check Size
 - Check code
 - Check file type
 - Run attack simulations
- Extra checks/tests on:
 - Federated learning models
 - Models released outside any SDE (external party)
- Outsource technical checks (if required)
- Revisit Risk Assessment – is release model safe?
- Save snapshot of software/pipeline for archiving
- Data Use Register updated

DataLoch
 Application form updated to include sufficient ML/AI detail for Risk Assessment discussions

Governance/Ethics Reviews

- Training of Reviewers
- Consider independent review

Framework Agreement: Controls after release

Application (online or email)
 User

Internal Reviews/ App Refinement
 User, SM, IG, Analyst

Approval processes
 IG facilitator

Contracts/Agreements
 IG/User/Legal
 External Data Delivery
 e.g. BioResource
 Data Extraction/Linkage
 Analyst

In place so far

Data Access in appropriate location
 User

Output Checking & Release
 Analyst

Release to public domain
 User

Release to another SDE/Host in DL?
 Analyst

Pre-Application/General Recommendations.
 Find ongoing AI/ML expertise to support

Amendment
 User/Analyst

Risk Assessment

Documented by researcher:

- Benefits of model release
- Types of model
- Location of model release (TRE/not)
- Model life expectancy
- Use controls when model released
- Data/pipeline archiving requirements
- Tools required to develop models
- Legal implications

Researcher Analysis/Model Creation

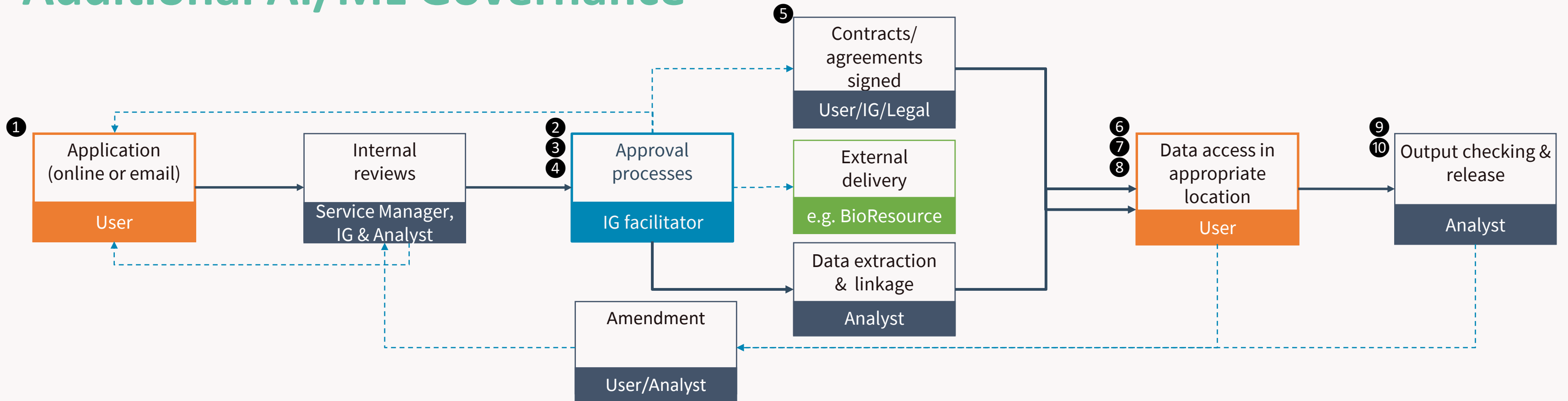
- Agreed Tools /Infrastructure provided to train ML models
- Researchers use agreed controls e.g.
 - Safe wrappers
 - Differentially private methods
 - Aggregate data training (instance based models)
- Researchers provide data dictionary for any model to be released (and base models)

Output Checking

- Extra checks/tests on:
 - Federated learning models
 - Models released outside any SDE (external party)
- Outsource technical checks (if required)
- Revisit Risk Assessment – is release model safe?

Out of DL control
 10.2.10: Safe wrapper principles developed.

Additional AI/ML Governance



1 Application Form updated to include ML/AI details

2 Approvals include some AI/ML expertise
 3 Additional ethical considerations included in decision, i.e. data bias and onward use of model
 4 In development - Upskill reviewers and DataLoch staff

5 Updated Organisational Framework Agreement to include controls after release

6 Some approved AI/ML tools within AAW
 7 Testing - Researcher controls such as safe wrappers
 8 Researcher provides data dictionary for models

9 Output request includes model characteristics and researcher declaration
 10 Outsource security tests (e.g. attack simulations, code reviews)

DataLoch AI/ML risk assessment
(to be discussed with researcher during application/governance processes)

Version 0.1

1. What type of models will you be building? (delete as appropriate)

A. Instance based - e.g. KNN, SVM, Gaussian?

B. Deep learning/expecting high numbers of parameters

C.

RISK ASSESSMENT

A living document created during application build, that supports discussions at approval and is used to evaluate model risk at output release

- Model benefits
- Model type
- Data archiving requirements
- Life expectancy
- Controls when live
- Tools/software required to develop the model
- Legal implications
- Output security test results

Specific Health Data Issues

- **Is it research? (governance differences)**
- **Classed as medical devices**
- **Implementation gap after development**

Questions for SDAP

- **Any experience in any of the scenarios? What have you done?**
- **What expertise is out there to assess some of the 5 Safe questions?**
- **Sharing forms/materials?**
- **Any researcher engagement done?**
- **Training recommendations?**
- [Get in touch: Amy.tilbrook@ed.ac.uk](mailto:Amy.tilbrook@ed.ac.uk)



DataLoch

Amy.tilbrook@ed.ac.uk

www.dataloch.co.uk