E-SCIENCE

Artigo: Tracking and Analyzing the Evolution of Provenance from Scripts

Aluno: Fernando Tamberlini Alves

Data: 26/04/2022

Resumo da Apresentação

01 Breves Conceitos de Proveniência

02 Artigo Científico

Tracking and Analyzing the Evolution of Provenence from Scripts
Pimentel, J.F., Freire, J, Braganholo, V., Murta, L.
IPAW - 2016

03 Uma breve análise da Proveniência do Estudo do Imperial College sobre a Previsão do Impacto do COVID-19 no Brasil

Conceitos de Proveniência

Formas de Proveniência

Prospectiva
Especificação dos passos do experimento

Retrospectiva
Dados capturados durante a execução de cada ensaio (trial) do experimento

Formas de Abordagem das Ferramentas de Proveniência

Ferramentas de gerência de configuração
Exemplo: CVS, SVN, GIT

Ferramentas de proveniência baseadas em scripts
Exemplo: noWorkflow, YesWorkflow, StarFlow, RDataTracker

Ferramentas de proveniência baseadas em workflow
Exemplo: Vistrails, Kepler

Artigo Científico

Título: Tracking and Analyzing the Evolution of Provenence from Scripts
Tradução Livre: Rastreamento e análise da evolução da proveniência a partir de scripts

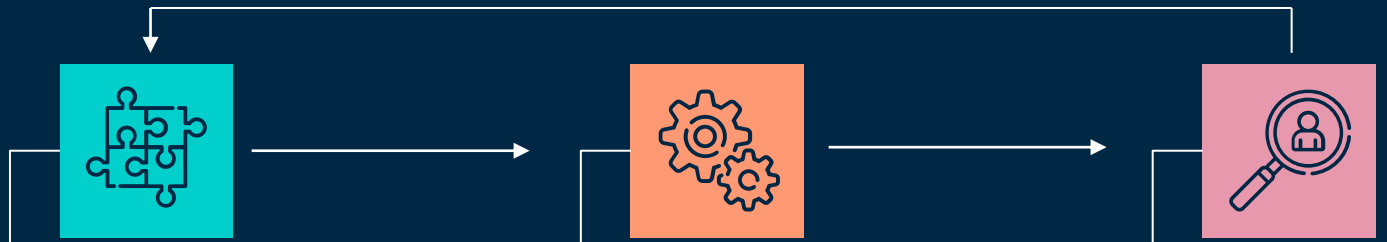
Autores: Pimentel, J.F, , Braganholo, V., Murta, L. – Universidade Federal Fluminense
Freire, J – New York University

Local de Publicação: International Provenance and Annotation Workshop – IPAW de 2016 (ipaw.info)

Seções do Artigo:

1. Introdução – Apresenta o ciclo de vida dos experimentos, a importância do rastreamento da evolução dos experimentos e as questões a serem respondidas a partir do rastreamento;
2. Trabalhos Relacionados – Apresentas as 3(três) categorias de abordagens das ferramentas de proveniência analisando a forma de rastreamento da evolução da proveniência;
3. Evolução da Proveniência baseada em scripts – Descreve a proposta do trabalho apresentando um modelo de versionamento e a forma de comparação dos ensaios de um experimento;
4. Implementação no noWorkFlow –Apresenta a implementação da proposta no noWorkflow
5. Avaliação – Descreve como a implementação da proposta no noWorkflow responde as questões mencionadas na seção de introdução
6. Conclusão – Conclui que a proposta é capaz de registrar toda a proveniência de cada ensaio e apresenta os trabalhos futuros

Ciclo de Vida dos Experimentos (script-based)



01

Hipótese

Definição da hipótese e codificação dos scripts

02

Execução

Execução dos scripts sobre a massa de dados específica

03

Análise dos Resultados

Análise dos resultados a partir de visualizações para confirmar a hipótese

Questões

Por que é necessário rastrear a proveniência ao longo dos ensaios executados durante o ciclo de vida de um experimento?

Q1 – Se um cientista executou um experimento duas vezes, mas trocou algums dos procedimentos no 2° ensaio, qual é a diferença?

Q2 – Comparando múltiplas execuções de acordo com os seus respectivos parâmetros, qual é a diferença no comportamento do experimento?

Q3 – Como as diferentes entradas se relacionam com as diferentes saídas?

Q4 – Rastreando o histórico dos ensaios, que partes da execução do experimento falha frequentemente?

Q5 – Que ensaio está relacionado com um determinado ensaio?

Q6 – Um ensaio foi derivado de qual outro ensaio?

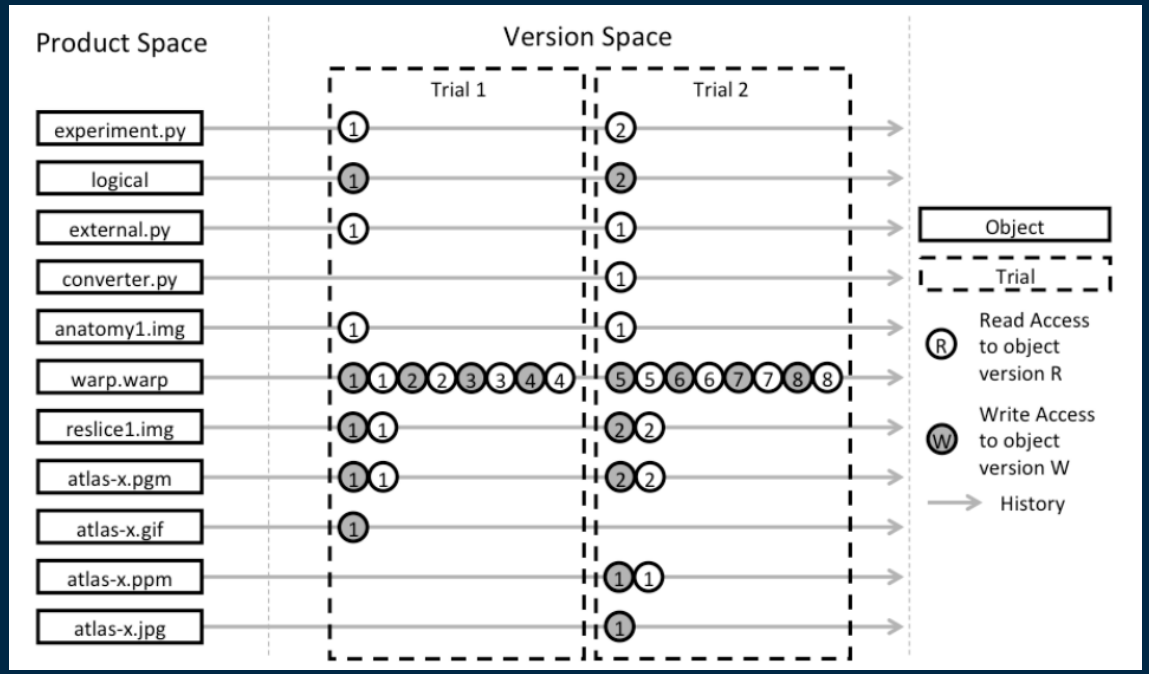
Q7 – Quais ensaio estão disponíveis e quais as suas respectivas durações?

Q8 – Quantos ensaios estão associados com um determinado Código?

Q9 – Quantos ensaio apresentaram falhas?

Adaptadas do 1° Desafio de Proveniência e ProveBench workshpos

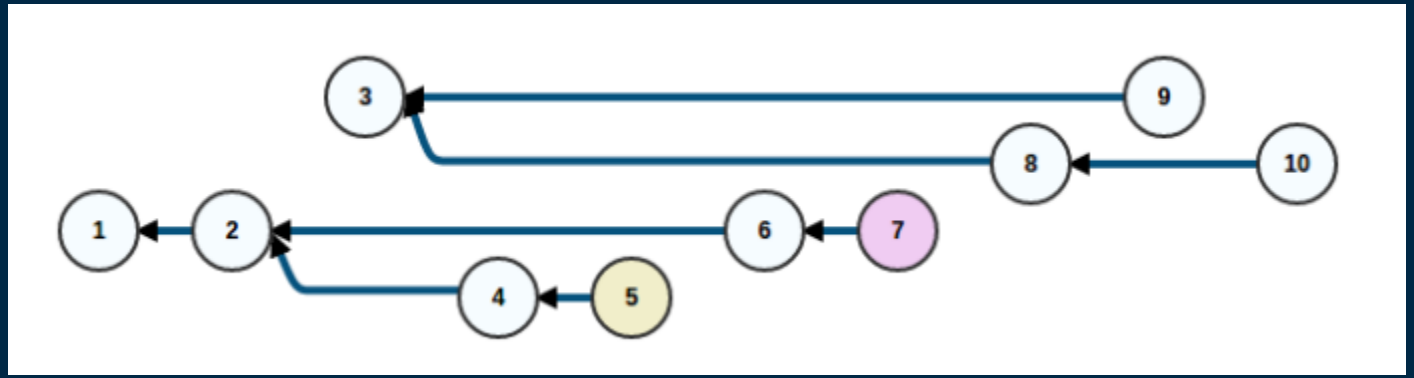
Modelo de Versionamento – Exemplo



- Área do Produto (Product Space) é como o produto é estruturado sendo que no caso o produto é o experimento. É onde contém os scripts, dados etc;

- Área do Versionamento (Version Space) é como o produto é versionado. No caso existem dois níves de versionamento: versão do ensaio e versão dos arquivos objetos do ensaio;

Histórico de Evolução do Experimento



Grafo que representa o histórico de evolução do experimento

- Os nós são as versões dos ensaios (trials);
- O nó 5 é um exemplo de ensaio (trial) de backup (não foi executado);
- O nó 7 é um exemplo de ensaio (trial) não finalizado;
- O nó 3 é um exemplo de ensaio (trial) que não teve origem nos ensaios anteriores;
- Poderá haver ramificações (branches) a partir de um ensaio (nós 2 e 3)

Representação do Modelo de Versionamento

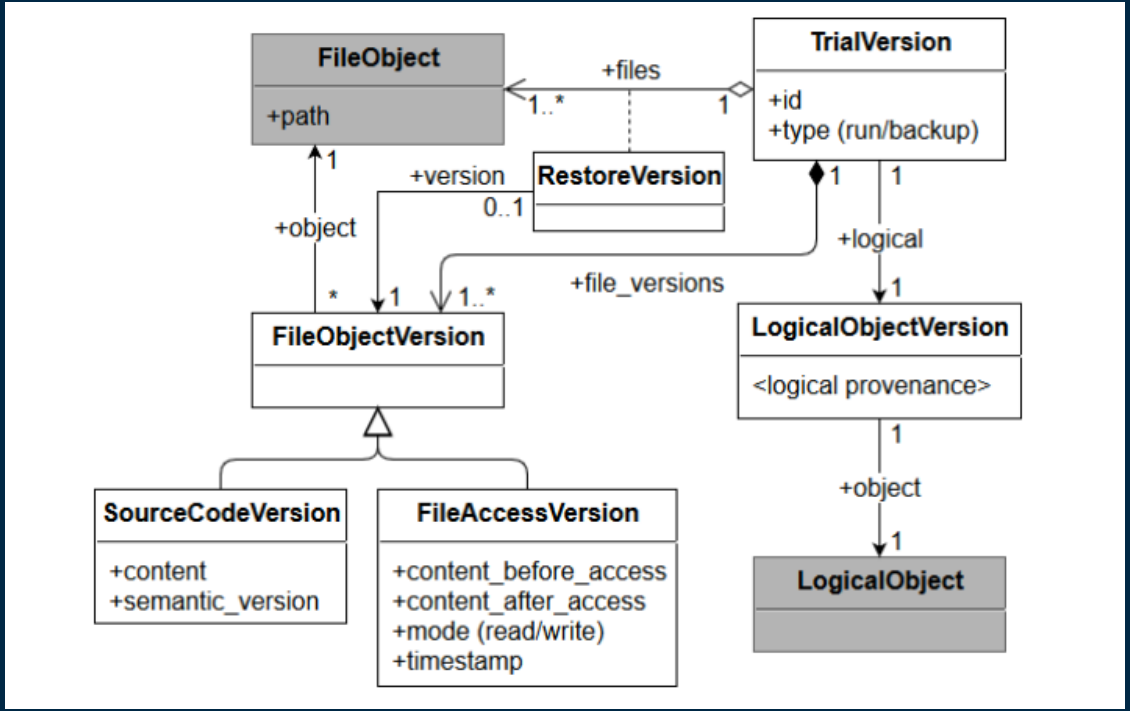


Diagrama que representa o modelo de versionamento proposto

- FileObject armazena a proveniência prospective;
- LogicalObject representa a proveniência retrospective;
- Aponta-se a Cardinalidade entre as entidades;
- Um ensaio possui apenas um LogicalObject;
- Um ensaio possui de 1 a N FileObject

Implementação no noWorkflow

Como comparar os ensaios de um experimento a partir do rastreamento do histórico no noWorkflow?

Comparando as versões entre os arquivos de cada ensaio e seus respectivos atributos.

Reportandos os arquivos incluídos e excluídos em cada ensaio.

Comparando os grafos de ativação de cada ensaio.

`> now run <script>` -> A partir da execução do script é criado um ID para o ensaio; coletado os dados de proveniência e armazendo em 2(dois) bancos de dados (uma base para armazenar os *file object* e outro para armazenar o *logical object* e os metadados dos *file object).* O noWorkflow armazena o script principal, módulos importados e acessos no *file object provenence* e a data de início e fim do ensaio, commandos, situação de execução, dados do ambiente, chamadas de função, valores retornados, variáveis, dados de dependência de várias são armazenadas no *logical provenence;*

`> now restore <trial_id>` -> restaura um versão de um ensaio;

`> now history` -> visualiza o histórico de evolução do experimento;

`> now diff <trial_id1> <trial_id2>` -> compara dois ensaios;

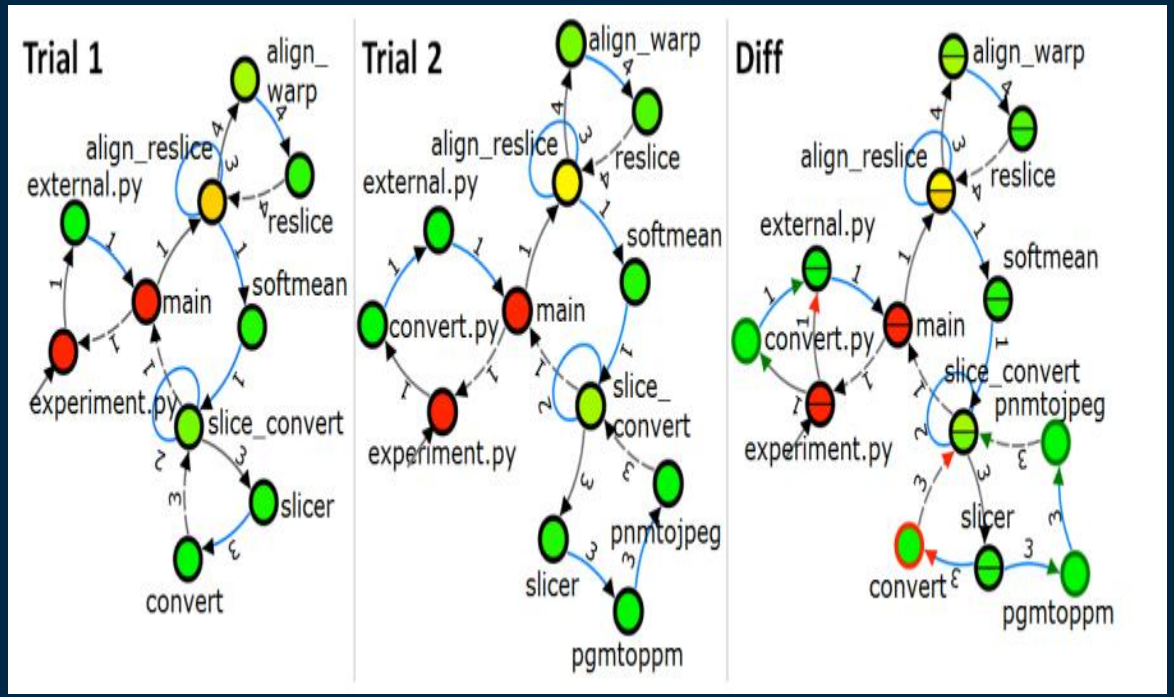
Exemplo de comparação entre ensaios

```
$ now diff 1 2 -f --brief
[now] trial diff:
  Start changed from 2016-02-11 04:49:09.008354
                  to 2016-02-11 04:49:09.898675
  Finish changed from 2016-02-11 04:49:09.536409
                   to 2016-02-11 04:49:10.276422
  Duration text changed from 0:00:00.528055 to 0:00:00.377747
  Code hash changed from cd1be11a2308ab217327a7d361138cb7f6c25106
                     to 2f637ec102961a7677e3f629ab88612d8875f04f
  Parent id changed from <None> to 1

[now] Brief file access diff
[Additions]            | [Removals]              | [Changes]
(rb) atlax-x.ppm       | (w) atlax-x.gif (new)   |
(w) atlax-x.jpg (new)  | (w) atlax-x.pgm (new)   |
(w) atlax-x.pgm        | (w) reslice1.img (new)  |
(w) atlax-x.ppm (new)  | (wb) warp.warp (new)    |
(w) reslice1.hdr       | ...                     |
(wb) warp.warp         |                         |
...                    |                         |
```

- Exemplo mostra as diferenças da proveniência lógica entre os ensaios 1 e 2;
- A verificação de igualdade dos arquivos é feita através de hashcode;
- Comparação entre os versões de código poderá ser feita através de ferramentas externas de comparação;
- O processo de comparação de chamadas de função é mais complexa. Ela transforma os ensaios para formato em grafo, depois em lísta de nós e aplica o algoritmo LCS sobre as listas. Esse processo é feita através do commando `now vis`

Comparação entre ensaios – chamadas



- Algoritmo LCS – Maior Subsequência Comum
- Cada nó representa um chamada e suas cores representam sua duração (Vermelho mais lento e verde mais rápido)
- Setas pretas são início de chamadas, setas azuis chamadas subsequentes e setas tracejadas represtam retorno;
- Nós com borda preta estão em ambos ensaios, nós e setas com borda vermelha só estão no 1° ensaio e nós e setas verdes só estão no 2° ensaio;

Avaliação da Proposta de Trabalho

A implementação na ferramenta noWorkflow foi capaz de responder as questões apontadas?

Q1 – Se um cientista executou um experimento duas vezes, mas trocou algums dos procedimentos no 2° ensaio, qual é a diferença? Q2 – Comparando múltiplas execuções de acordo com os seus respectivos parâmetros, qual é a diferença no comportamento do experimento? R: A visualização da comparação entre as chamadas de função entre dois ensaios é capaz de respodenre a Q1 e Q2

Q3 – Como as diferentes entradas se relacionam com as diferentes saídas? R: O comando now diff –f é capaz de comparer os acessos aos arquivos de dados.

Q4 – Rastreando o histórico dos ensaios, que partes da execução do experimento falha frequentemente? R: Um commando de consulta SQL é capaz de saber os resultados dos ensaios.

Q5 – Que ensaio está relacionado com um determinado ensaio? Q6 – Um ensaio foi derivado de qual outro ensaio? Q7 – Quais ensaio estão disponíveis e quais as suas respectivas durações? Q8 – Quantos ensaios estão associados com um determinado Código? Q9 – Quantos ensaio apresentaram falhas? R: A ferramenta dispõe de uma visualização da evolução de histórico de proveniência capaz de responder estas questões.

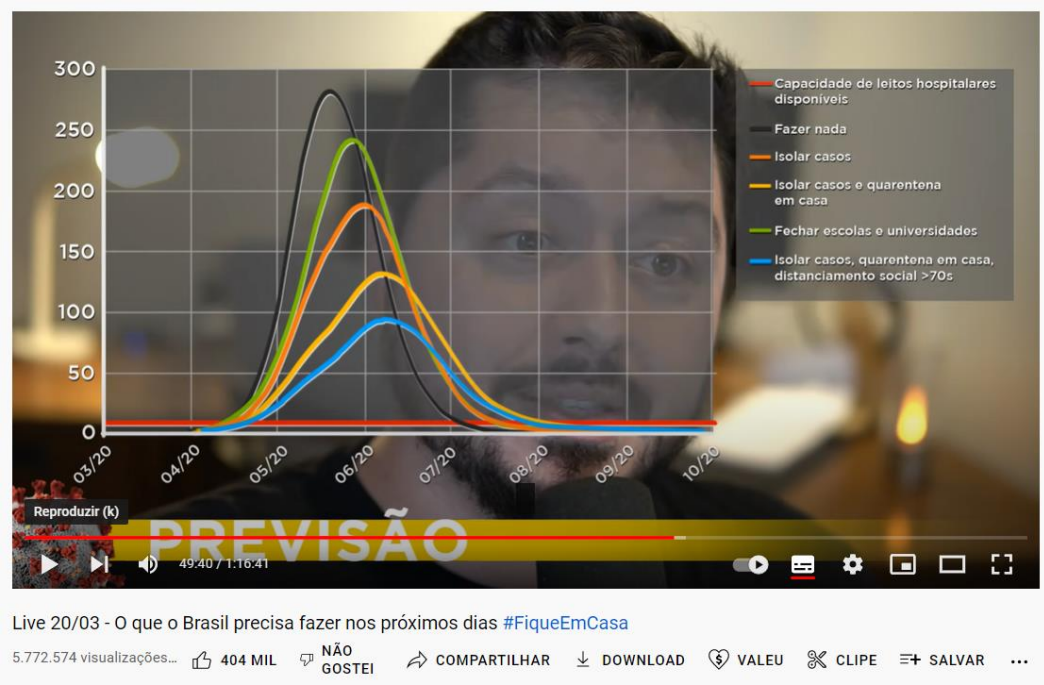
Conclusão e trabalhos futuros

A proposta apresentada se mostrou capaz de coletar todo histórico de proveniência de uma experimento e ela está implementada na ferramenta noWorkflow

O uso de ferramenta pode impactar o tempo de execução do experimento e em razão de armazenar informações de fino grão poderá comprometer a escalabilidade em termo de tempo de execução e armazenamento

Pretende-se explorar alternativas de detecção de mudança de arquivo e melhorar o algoritmo de comparação de chamadas de função

Proveniência no Estudo da Imperial College



Previsão de Mortes no Brasil
Estudo da Imperial College

- Atila Iamarino – D. Sc. Biologia e Youtuber com canal com 1,56 milhão de inscritos
- Vídeo disponibilizado em 2020
- Mais de 5 milhões visualizações
- Faz uma previsão de 1 milhão de mortes até Ago/2020 no Brasil
- Projeção no Brasil a partir do histórico de outros países.

Proveniência no Estudo da Imperial College

Previsão de Mortes no Brasil – Estudo da Imperial College

Estudo Geral

The global impact of COVID-19 has been profound, and the public health threat it represents is the most serious seen in a respiratory virus since the 1918 H1N1 influenza pandemic. Here we present the results of epidemiological modelling which has informed policymaking in the UK and other countries in recent weeks. In the absence of a COVID-19 vaccine, we assess the potential role of a number of public health measures – so-called non-pharmaceutical interventions (NPIs) – aimed at reducing contact rates in the population and thereby reducing transmission of the virus.

Estudo Específico para o Brasil

O Brasil é atualmente o epicentro da COVID-19 na América Latina. Neste relatório descrevemos a epidemia brasileira usando três medidas epidemiológicas: número de infecções, número de óbitos e número de reprodução. Nossa modelagem requer um número mínimo de óbitos para que tendências possam ser estimadas, portanto limitamos nossa análise aos 16 estados com mais de cinquenta óbitos reportados. A distribuição de óbitos entre os estados é altamente heterogênea, com cinco estados --- São Paulo, Rio de Janeiro, Ceará, Pernambuco e Amazonas --- contabilizando 81% dos óbitos reportados até o momento (...) Embora em escala nacional a epidemia brasileira ainda seja relativamente inicial, nossos resultados sugerem que mais ações são necessárias para limitar a disseminação e prevenir sobrecarga do sistema de saúde.

Proveniência no Estudo da Imperial College

Outros estudos de previsão – COVID

A pandemia da COVID-19 no Brasil: a série de projeções do Institute for Health Metrics and Evaluation e a evolução observada, maio a agosto de 2020

Em termos de acurácia, dados disponíveis sugerem que erros nas estimativas de óbitos por COVID-19 no Brasil, realizadas por grupos internacionais, são semelhantes; exceto as do ICL, que apresentaram superestimação bem maior.

Proveniência no Estudo da Imperial College

Críticas ao Código do Modelo de Ferguson

- Sue Denin (pseudônimo)– ex-engenheira do Google – Blog Lockdown Sceptics
  - Code Review of Fergunson's Model
  - Código disponibilizado no github
  - Código em C++ em um único arquivo com 15 mil linhas
  - Saídas não determinísticas, sem testes e equações não documentadas
  - Comentários sobre a revisão

Proveniência no Estudo da Imperial College

Conclusão – Críticas ao Código do Modelo de Ferguson

All papers based on this code should be retracted immediately. Imperial's modelling efforts should be reset with a new team that isn't under Professor Ferguson, and which has a commitment to replicable results with published code from day one.

On a personal level, I'd go further and suggest that all academic epidemiology be defunded. This sort of work is best done by the insurance sector. Insurers employ modellers and data scientists, but also employ managers whose job is to decide whether a model is accurate enough for real world usage and professional software engineers to ensure model software is properly tested, understandable and so on. Academic efforts don't have these people, and the results speak for themselves.

Proveniência no Estudo da Imperial College

A inexistência da proveniência no estudo da Imperial College impediu a reprodutibilidade do modelo proposto para assim ser corroborado por outros cientistas ?