Reprodutibilidade em Ciência dos Dados

Ivan Marin
Vivo Data Labs
ivan.smarin@telefonica.com

O que é

Reprodutibilidade

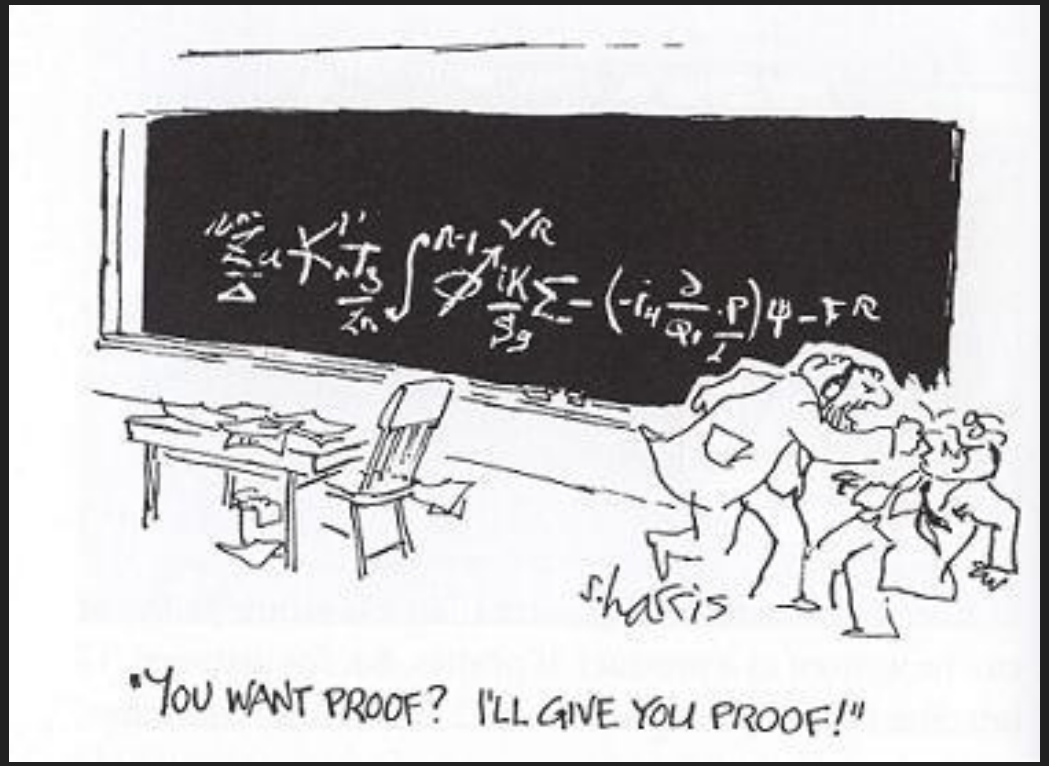
é a capacidade de um experimento ou estudo ser ***replicado***

pelo ***mesmo*** pesquisador ou por ***outro grupo***

de forma ***independente*** e ***completa***

O que ***não*** é

Sidney Harris



O que ***não*** é

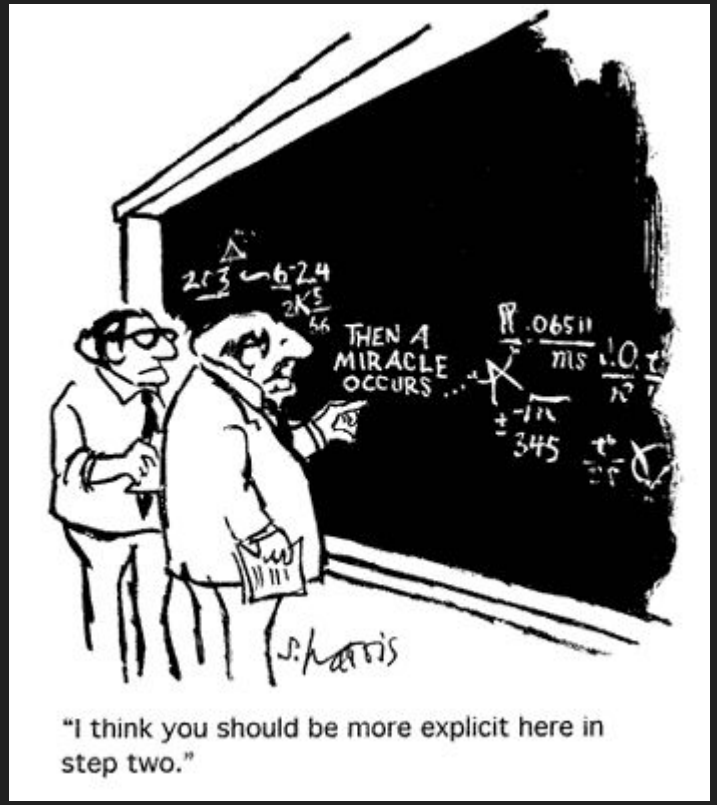
- Uma ferramenta (ou um conjunto de ferramentas)
- Um padrão fixo de código ou documentação
- Útil somente no meio acadêmico
- Garantia de entrega de resultados
- Método de planejamento de projeto
- Versionamento

Para quê serve?

- Garantir que os resultados gerados em um projeto sejam replicáveis
- Não perder o conhecimento gerado durante o andamento da pesquisa
- Poder analisar o processo se houver dúvidas ou erros
- Fundamentar as conclusões
- Permitir a expansão, modificação ou continuação do projeto
- Permitir a comparação dos resultados ou metodologia entre projetos
- Analisar o histórico e as etapas

Para quê serve?

Sidney Harris



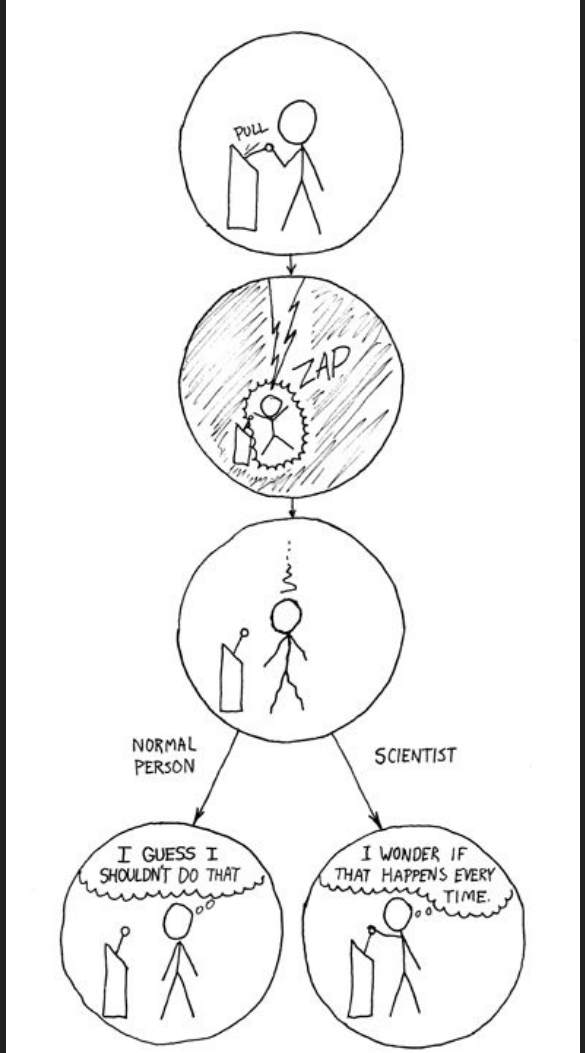
"I think you should be more explicit here in step two."

E a quem pode ajudar?

- Pesquisadores
- Cientistas de Dados
- Engenheiros de Software
- Gerentes de Projeto
- Programadores
- Equipe de Suporte
- Editores e revisores de periódicos

XKCD obrigatória

https://xkcd.com/242/



Como fazer

Para que um experimento seja reprodutível, ele deve contemplar

- A documentação das hipóteses envolvidas
- O código fonte utilizado para todas as etapas (da ingestão até a visualização e relatórios)
- Os dados utilizados para gerar os resultados
  - Ou pelo menos uma amostra significativa para testar os métodos
- Versionamento dos componentes
- Os resultados do experimento

Como fazer

Código fonte:

- Código, dados e documentação versionados
- As bibliotecas computacionais devem ser isoladas do sistema
- Todo o código deve ser portável para outras arquiteturas

Como fazer

Dados:

- Os dados devem ser identificados por data e origem
- Dados brutos nunca devem ser modificados
- Dados transformados podem ser armazenados
- Estatísticas e análises intermediárias podem ser armazenados

Como fazer

Resultados e documentação:

- Todos os resultados devem ser autogerados, sem intervenção
- A documentação deve ser autogerada
- Os resultados podem estar prontos a serem compartilhados em sua forma final
- Um histórico da evolução dos resultados e documentação pode estar disponível

Hora do show: Projeto de Ciência de Dados

- Projeto de ciência dos dados para análise de base de eleitores
- Python (claro!)

Estrutura:

- `data`
- `src`
- `doc`
- `analysis`
- `results`

Projeto de Ciência de Dados

Ferramentas:

- `virtualenv`
- `pip`
- `cookiecutter`
- `jupyter notebook`
- `sklearn`
- `tensorflow`

Projeto de Ciência de Dados

Ferramentas externas:

- github.com/deeplearningsp
- Google Cloud Computing
- IBM Bluemix

Projeto de Ciência de Dados

Etapas:

1. Criar o ambiente de desenvolvimento com diretórios, `virtualenv` e `pip`
2. Popular definições do problema em `doc` e em `results`
3. Estruturar o código para ingestão de dados
4. Fazer a ingestão dos dados brutos em `data`
5. E que comecem as análises
6. Parte iterativa: a cada etapa, fazer um ciclo entre
    a. análise
    b. visualização
    c. relatório/documentação
7. Gerar código para resultados finais
8. Gerar resultados finais e apresentação em `results`

Here be dragons

- Os resultados são mais importantes que a estrutura
- O código em um projeto de Ciência de Dados ***não*** é o objetivo final
- A estrutura deve ser adaptada ao projeto a ser executado (e não vice-versa)

"*A foolish consistency is the hobgoblin of little minds*"

(Ralph Waldo Emerson)

Perguntas?

Obrigado!

(Algumas) referências

http://mfactor.sdf.org/data-science-workflow-with-reproducible-research.html

http://drivendata.github.io/cookiecutter-data-science/

https://en.wikipedia.org/wiki/Reproducibility

https://en.wikipedia.org/wiki/Reproducibility_Project

https://en.wikipedia.org/wiki/Replication_crisis

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/

Útil?

“In any moment of decision, the best thing you can do is the right thing, the next best thing is the wrong thing, and the worst thing you can do is nothing.”

Theodore Roosevelt