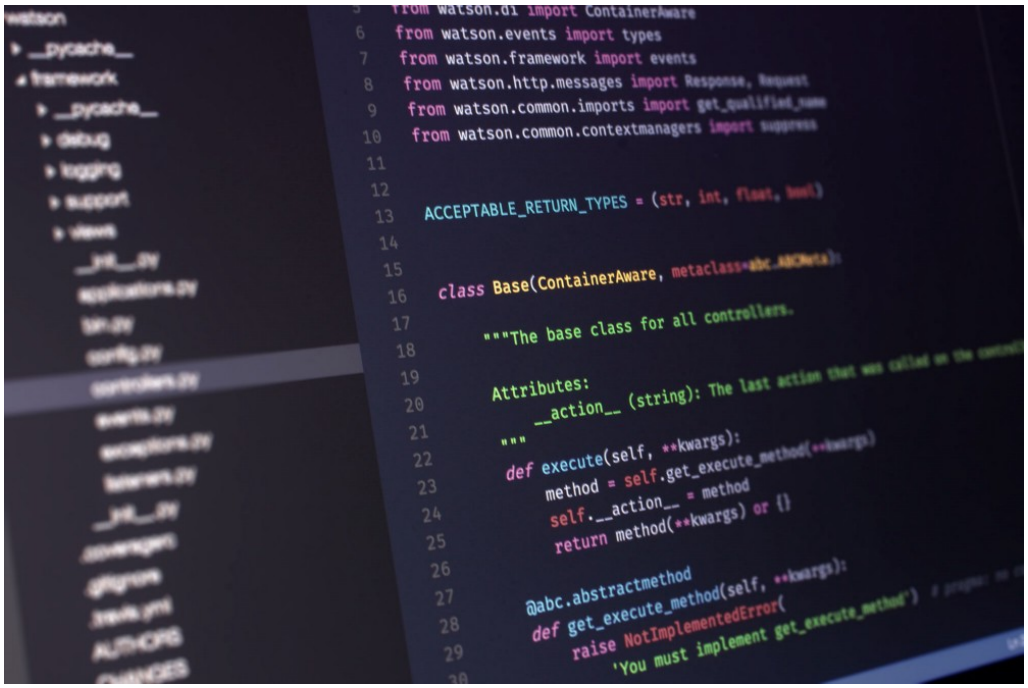
Lidando com Dados na Pesquisa Computacional Reprodutível



Este artigo faz parte da série **Por que devo tornar minha Pesquisa Computacional Reprodutível?** Se você ainda não leu, dê um pulinho por lá, caso já saiba do que estamos falando vá em frente e aproveite a leitura!

Os dados em qualquer pesquisa, são um item especialmente delicados, já que nem sempre os dados utilizados podem ser compartilhados publicamente. Dados médicos, genéticos entre outros, têm este tipo de restrição. Para os dados que podem ser compartilhados as ferramentas de controle de versão podem ser utilizadas como os conhecidos git e svn. Também existem alguns repositórios online específicos que podem ajudar nesta tarefa como **Github**ou **Bitbucket.**

Dicas

1. Se possível, disponibilize os dados. Estes podem ser de dois tipos: Dados Brutos ou Dados Normalizados, deixe explícito qual tipo de dado está sendo disponibilizado. O dado bruto é aquele sem nenhum tipo de filtro e alteração, os Normalizados são aqueles que passaram por alguma adaptação ou alteração por conta do autor.
2. Caso utilize dados normalizados inclua os dados brutos e scripts que chegaram aquele resultado, juntamente com as sementes do gerador de números randômicos, caso seja utilizada. Os dados e scripts podem ser disponibilizados num ambiente de controle de versão.
3. Deixe o computador fazer o trabalho difícil: Caso precise realizar alguma tarefa repetitiva como numa base de dados de imagem cortá-las no mesmo tamanho específico, automatize esta tarefa. Isso garante dados mais consistentes e evita dados com erros ou fora do padrão especificado.

4. Informe as decisões do projeto: Alguns dados foram removidos da amostra utilizada? Além de disponibilizar os scripts que alteram a amostra, indique o motivo de utilizar aquele tipo de normalização. Através dessa informação outros pesquisadores podem ter visões diferentes do seu trabalho e até gerarem novas propostas.
5. Quando possível, ofereça a informação da procedência dos dados. Afinal de contas, não queremos fazer pesquisa com dados não confiáveis. Informar a origem dos seus dados dá mais transparência e confiança em seu projeto.