



Avaliação de validade

Proposta de experimento - Medição e
experimentação

INTRODUÇÃO

- Objetivo
 - Este laboratório investiga a atividade de code review em repositórios populares do GitHub, analisando fatores que influenciam no merge de Pull Requests e no número de revisões realizadas.
- Amostra
 - Foram coletados dados de 3.616 PRs válidos dos 200 repositórios mais populares do GitHub, aplicando filtros para garantir que passaram por processo de review humano.



AMEAÇA À VALIDADE DE CONCLUSÃO

BAIXO PODER ESTATÍSTICO

O PROBLEMA PRINCIPAL:

- Amostra limitada: 12 desenvolvedores, ~100 PRs
- Risco de Erro Tipo II (falso negativo)
- Estudo pode não ter sensibilidade suficiente

O CONSEQUÊNCIA DIRETA:

- Concluir que "IA não tem efeito" quando na verdade tem
- Resultados estatisticamente não-significativos, mesmo com efeitos práticos
- Decisões baseadas em evidências fracas

RAZÃO ESPECÍFICA DO ESTUDO:

- Restrições do contexto organizacional real
- Dificuldade de recrutar mais participantes
- Trade-off entre robustez estatística e viabilidade

AMEAÇA À VALIDADE INTERNA

VAZAMENTO DE TRATAMENTOS (CARRYOVER)

NATUREZA DA AMEAÇA:

- Design within-subjects (todos usam ambas condições)
- Aprendizado/hábitos de uma condição contaminam a outra
- Dificuldade em isolar efeito "puro" da IA

CENÁRIOS CONCRETOS:

- Dev que começa com IA: pode aplicar padrões aprendidos depois
- Dev frustrado com IA: pode desengajar nas revisões tradicionais
- Curva de aprendizado afeta medidas de eficiência

IMPACTO NAS HIPÓTESES:

- Superestimar benefícios da IA (se aprendizado ajudar)
- Subestimar benefícios (se frustração prejudicar)
- Dados misturados = causalidade incerta

AMEAÇA À VALIDADE DE CONSTRUCTO

VIÉS DE DESEJABILIDADE SOCIAL

ORIGEM DO VIÉS:

- Participantes sabem que avaliam "nova tecnologia"
- Pressão organizacional para aceitar inovações
- Medo de parecer "resistente à mudança"

ÁREAS MAIS AFETADAS:

- Questionários de percepção (Likert 1-5)
- Avaliações de utilidade e confiança
- Feedback qualitativo em entrevistas

RISCO PARA AS CONCLUSÕES:

- Métricas subjetivas artificialmente inflacionadas
- Discrepância entre "o que dizem" e "o que fazem"
- Percepção não reflete realidade de uso

AMEAÇA À VALIDADE EXTERNA

CONTEXTO ESPECÍFICO DEMASIADO

LIMITAÇÕES DO CONTEXTO:

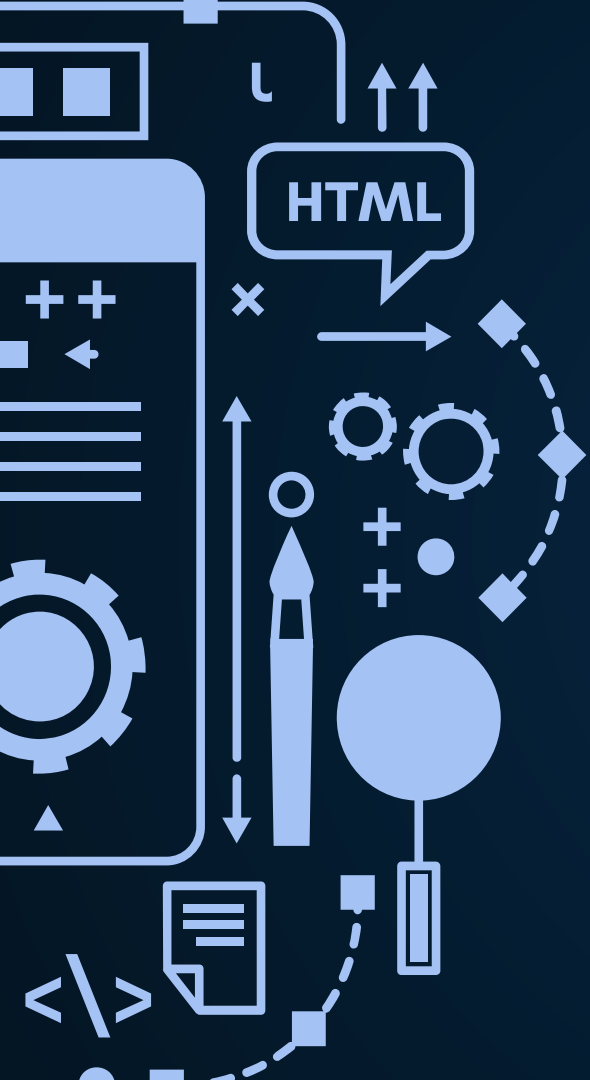
- Apenas UMA empresa/organização
- Stack específica: TypeScript, React, Node.js
- Cultura e processos únicos da empresa

PERFIS RESTRITOS:

- Desenvolvedores com experiências similares
- Mesmo nível de maturidade em PRs
- Mesmas ferramentas e pipelines

BARREIRAS À GENERALIZAÇÃO:

- Outras stacks (Java, Python, C#) podem ter resultados diferentes
- Startups vs. grandes empresas: dinâmicas diferentes
- Outras ferramentas de IA: desempenho variável



Obrigado!