Capítulo 16

Data Wrangling: o que é? para que serve?

Ricardo Ceneviva
Março de 2025

Introdução

A ciência de dados vem transformando as ciências sociais ao oferecer novas formas de coletar, analisar e interpretar dados em larga escala. Perguntas clássicas sobre comportamento humano e organização social podem agora ser exploradas com auxílio de grandes bases de dados digitais – de registros de telefonia móvel a interações em redes sociais online – e de métodos computacionais avançados (LAZER et al., 2021). Contudo, tirar proveito dessas novas fontes exige um pipeline estruturado que vá desde a formulação da pergunta de pesquisa até a consideração de questões de governança e ética dos dados. Este relatório apresenta os fundamentos teórico-metodológicos para uma aula intitulada "Ciência de Dados aplicada às ciências sociais: pipeline, programação e automação". O objetivo central é mostrar como seguir um fluxo de trabalho coeso – pergunta → dados → métodos → métricas → governança – pode auxiliar pesquisadores sociais a aproveitar ferramentas de ciência de dados sem perder de vista rigor metodológico e princípios éticos (SALGANIK, 2019).

Como motivação, tomamos dois estudos canônicos: (1) Blumenstock, Cadamuro e On (2015), que usam metadados de telefonia móvel para predizer pobreza em Ruanda, ilustrando a abordagem preditiva; e (2) Bond et al. (2012), um experimento com 61 milhões de usuários do Facebook para estimular votação, exemplificando inferência causal em plataformas digitais. Esses casos demonstram, respectivamente, o potencial de modelos preditivos para fins sociais e de experimentos randomizados em larga escala para estimar efeitos causais. Em ambos, veremos como a pergunta de pesquisa orienta as escolhas de dados e métodos, quais métricas avaliam o sucesso e quais riscos operacionais e de governança emergem em cada etapa.

A estrutura deste relatório segue o pipeline proposto. Primeiro, discutimos o conceito de pipeline de ciência de dados e o fluxo de pesquisa social digital segundo Salganik (2019), detalhando etapas de coleta programática de dados (incluindo scraping responsável) e de aplicação de surveys digitais. Em seguida, abordamos práticas de preparação de dados, engenharia de atributos, automação de análises e reprodutibilidade. Depois, contrastamos objetivos e critérios de predição vs. inferência causal, mostrando que cada abordagem requer métricas e validações próprias (SHMUELI, 2010). Na sequência, analisamos os

dois estudos empíricos mencionados, mapeando em cada um a relação entre pergunta, método, métricas e questões de governança. Por fim, discutimos aspectos de governança, ética e conformidade legal – com destaque para a LGPD (Lei Geral de Proteção de Dados) – que devem permear todo o processo de pesquisa.

Ao longo do texto, integramos referências essenciais que situam essas práticas no estado da arte. Isso inclui obras fundamentais em modelagem estatística e aprendizado de máquina (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2021), discussões sobre diferenças entre paradigmas estatísticos e computacionais (BREIMAN, 2001; SHMUELI, 2010), novos tipos de dados como textos (GRIMMER; ROBERTS; STEWART, 2022; GENTZKOW; KELLY; TADDY, 2019), e diretrizes de reprodutibilidade e boas práticas de programação científica (PENG, 2011; WILSON et al., 2017; MARWICK; BOETTIGER; MULLEN, 2018). Espera-se que, com essa base, os alunos de ciências sociais compreendam não apenas o que é feito em ciência de dados, mas como e por que fazê-lo de maneira responsável e alinhada às perguntas substantivas de pesquisa.

Ciência de Dados e Pesquisa Social Digital

Um pipeline de ciência de dados é a sequência organizada de etapas pela qual uma investigação orientada por dados progride, desde a concepção da pergunta até a obtenção de resultados e sua avaliação sob critérios técnicos e éticos. Em linhas gerais, podemos delinear as etapas como: (a) formular a pergunta ou hipótese de pesquisa, vinculada a um problema social substantivo; (b) identificar e coletar os dados apropriados para respondê-la, seja por fontes digitais existentes (observação passiva) ou por instrumentos de pesquisa ativos (surveys, experimentos); (c) aplicar métodos analíticos ou computacionais adequados – que podem incluir algoritmos de aprendizado de máquina, modelos estatísticos inferenciais, ou desenhos experimentais, conforme o caso; (d) avaliar os resultados por meio de métricas de desempenho ou testes de significância, garantindo que as conclusões sejam válidas; e (e) implementar práticas de governança de dados, que envolvem tanto a gestão responsável (privacidade, segurança, conformidade legal) quanto a documentação e comunicação transparente dos processos e achados.

Importante notar que esse pipeline raramente é linear ou estanque: muitas vezes o processo é iterativo, ajustando métodos ou coletando dados adicionais conforme se aprende com etapas intermediárias. Ademais, princípios de ética e governança devem estar presentes transversalmente – por exemplo, verificando-se requisitos legais já na coleta de dados e considerando-se implicações sociais ao interpretar métricas.

A estrutura proposta por Salganik (2019) para pesquisa social na era digital fornece um arcabouço conceitual alinhado com esse pipeline. Salganik sugere que pesquisadores podem observar comportamentos através de dados digitais existentes (como registros de telefonia ou redes sociais), perguntar diretamente a pessoas através de surveys e enquetes digitais, experimentar mediante testes controlados online, e colaborar com o público ou outros pesquisadores em projetos de ciência cidadã ou crowdsourcing. Cada um desses modos fornece dados de natureza distinta – observacionais, declarativos, experimentais,

colaborativos – que enriquecem a pesquisa social quando integrados ao fluxo de trabalho. Por exemplo, no caso de Blumenstock et al. (2015), combinaram-se observação (metadados telefônicos) com perguntas (survey socioeconômico) para criar e validar um modelo preditivo. Já o estudo de Bond et al. (2012) se apoia na experimentação em massa via plataforma digital. Assim, o pipeline deve ser suficientemente flexível para abarcar diferentes estratégias de coleta e análise, mantendo, porém, uma coesão entre pergunta, dados, método, métrica e governança. Em todos os casos, a documentação rigorosa e a consideração de riscos (como viés de seleção em dados observacionais ou questões de consentimento em experimentos online) precisam acompanhar o processo.

Coleta Programática de Dados e Scraping Responsável

O passo de coleta de dados em contextos digitais muitas vezes envolve a extração automática de informações de websites, plataformas online ou bancos de dados remotos. Para isso, é crucial adotar práticas de scraping responsáveis. Em primeiro lugar, deve-se priorizar APIs oficiais quando disponíveis, pois elas geralmente fornecem acesso estruturado aos dados de forma autorizada. Quando o acesso é obtido por raspagem direta de sites (web scraping), é imperativo respeitar as instruções contidas no arquivo `robots.txt` do domínio (que indica quais áreas podem ou não ser vasculhadas) e os termos de serviço da plataforma (que podem restringir coleta automatizada). Por exemplo, se um pesquisador pretende coletar dados de perfis em rede social, deve verificar se a plataforma permite tal uso para fins acadêmicos e em que condições.

Além disso, práticas de minimização e registro são recomendadas. Coletar apenas os dados necessários para a pergunta de pesquisa está alinhado ao princípio da necessidade (conforme discutido na LGPD, ver seção de Governança) e reduz riscos em caso de vazamentos. Dados pessoais sensíveis ou identificadores diretos devem ser anonimizados o mais cedo possível no processo – idealmente no próprio momento da coleta. Deve-se também respeitar limites de taxa (*rate limits*) nas requisições para não sobrecarregar servidores e evitar bloqueios; ferramentas de scraping normalmente permitem inserir pausas entre acessos. Todo o procedimento de coleta deve ser cuidadosamente documentado: registram-se os parâmetros usados (por exemplo, palavras-chave de busca, intervalo de datas), a data e hora da coleta e o código ou script utilizado. Esse registro permite tanto a reprodutibilidade (PENG, 2011) quanto eventuais auditorias ou atualizações posteriores dos dados coletados. Em suma, a coleta programática responsável concilia a eficiência da automação – essencial para lidar com grandes volumes de dados na ciência social computacional – com o respeito a normas éticas e legais de obtenção de dados.

Surveys Digitais e Amostragem

Paralelamente aos dados obtidos de forma passiva, muitas pesquisas exigem dados coletados ativamente por meio de surveys digitais. Com a migração de questionários para plataformas online, é comum que as amostras obtidas sejam não probabilísticas – por exemplo, respondentes recrutados via redes sociais ou painéis voluntários na internet. Essa falta

de aleatoriedade estrita na seleção pode introduzir vieses de representatividade: certos grupos populacionais podem ficar sub ou super-representados entre os respondentes (por motivos de acesso à internet, engajamento online, etc.). Para mitigar esses problemas, os pesquisadores empregam técnicas de pós-estratificação e ponderação. Essencialmente, ajustam-se pesos para os respondentes de modo que a distribuição amostral se aproxime de marginais conhecidas da população (por exemplo, calibrando para que a composição por faixa etária, gênero e região na amostra reflita dados censitários). Métodos como *raking* ou utilização de cotas pré-definidas são comumente aplicados com esse fim.

Outra prática recomendada é a validação externa: sempre que possível, comparar as estimativas do survey digital com fontes confiáveis ou dados oficiais. Por exemplo, se uma enquete online medir intenção de voto, seus resultados podem ser contrastados com pesquisas eleitorais tradicionais ou com o resultado real da eleição, a fim de avaliar o viés e calibrar o modelo preditivo. Além disso, documentar detalhadamente o processo de amostragem e coleta – incluindo o texto das perguntas, o período de aplicação e o modo de recrutamento dos participantes – é fundamental para transparência. Essa documentação permite que outros entendam o contexto dos dados e a adequação da amostra ao objetivo da pesquisa (WICKHAM; ÇETINKAYA-RUNDEL; GROLEMUND, 2023). No caso do estudo de Blumenstock et al. (2015), por exemplo, embora os autores tenham conduzido um survey telefônico com amostragem aleatória estratificada de usuários de celular, é sabido que os mais pobres sem acesso a telefone não poderiam ser incluídos – uma limitação de cobertura que precisa ser reconhecida na análise. Em suma, surveys digitais ampliam o alcance e rapidez da coleta de dados em ciências sociais, mas exigem cuidados metodológicos para assegurar que as inferências sejam válidas e bem suportadas pelos dados coletados.

Preparação de Dados e Engenharia de Atributos

Uma vez obtidos os dados brutos, inicia-se a etapa de preparação e engenharia de atributos. A preparação de dados – frequentemente chamada de *data wrangling* – envolve limpar e organizar os dados de forma que estejam prontos para análise. Isso inclui tratar valores ausentes ou inconsistentes, eliminar duplicatas, corrigir erros de digitação ou formatação e integrar diferentes fontes de dados quando necessário (por exemplo, combinando dados de survey com registros digitais através de um identificador comum). É recomendável realizar checagens de integridade: verificar se os totais batem após junções, se unidades de medida estão consistentes e se não há outliers extremos indicando possíveis erros de coleta. Ferramentas como `pandas` no Python (McKINNEY, 2022) ou o pacote `tidyverse` no R (WICKHAM, 2014) são amplamente utilizadas para essas tarefas, pois facilitam transformações e inspeções rápidas nos dados.

Parte essencial dessa fase é montar um dicionário de dados, documentando cada variável (coluna) e seu significado, unidades, categorias possíveis etc. Isso garante que tanto o pesquisador quanto outros repliquem ou compreendam a análise no futuro, evitando ambiguidades sobre o que cada atributo representa.

Em seguida, entra a engenharia de atributos: a criação ou seleção de variáveis derivadas que melhor representem os padrões relevantes no conjunto de dados (KUHN; JOHNSON, 2019). Dados originais muitas vezes precisam ser transformados para extrair informações úteis – por exemplo, no estudo de Blumenstock et al. (2015), a partir dos registros brutos de ligações telefônicas foram construídas métricas como número de contatos distintos, frequência de uso noturno do telefone, volume de gastos com recarga, entre outras. Essas variáveis derivadas capturam aspectos do comportamento do indivíduo que podem ter correlação com sua condição socioeconômica. Técnicas de redução de dimensionalidade ou seleção de variáveis (como análise de componentes principais ou métodos *lasso*) também podem ser aplicadas para diminuir ruído e evitar a sobrecarga de atributos irrelevantes. Conforme Harris et al. (2020) destacam, bibliotecas de computação científica como `NumPy` permitem manipulação eficiente de *arrays* de dados, o que é fundamental para calcular inúmeras transformações nos atributos de forma vetorizada e reprodutível.

Ao final da preparação, os dados devem estar organizados em um formato analítico – frequentemente *tidy data*, isto é, cada linha representando uma unidade de observação e cada coluna, uma variável pertinente (WICKHAM, 2014). Todo código de limpeza e engenharia deve ser salvo e versionado (por exemplo, com controle de versão via Git), de modo que haja rastreabilidade completa das modificações realizadas no conjunto de dados. Essa disciplina de documentação e versionamento é parte integrante das boas práticas recomendadas em computação científica (WILSON et al., 2017), assegurando que do dado bruto ao dado final analisado exista um histórico claro e reproduzível.

Automação, Documentação e Reprodutibilidade

Por fim, um pipeline robusto deve incluir a automação e documentação de todas as etapas analíticas, visando máxima reprodutibilidade. Em projetos de ciência de dados, é prática recomendada usar um *task runner* (como `Makefile`, `Snakemake` ou scripts bem estruturados) para automatizar a execução sequencial das tarefas: por exemplo, da coleta à limpeza, da geração de modelos à produção de relatórios. Com isso, caso novos dados sejam coletados ou ajustes sejam necessários, todo o processo pode ser reexecutado de forma consistente, minimizando erros manuais (MARWICK; BOETTIGER; MULLEN, 2018). Durante a execução das rotinas, é útil implementar mecanismos de *logging* – registros automáticos de eventos – para guardar informações como parâmetros utilizados, versões de pacotes de software, semente aleatória empregada em algoritmos estocásticos, duração de treinamentos de modelo etc. Tais registros facilitam detectar a causa de resultados inesperados e garantir que outro pesquisador possa reproduzir o ambiente analítico (PENG, 2011).

A geração de relatórios automáticos de métricas é outra prática valiosa. Ferramentas como Jupyter Notebooks (no Python) ou RMarkdown (no R) podem compilar, ao final da análise, tabelas e visualizações resumindo o desempenho do modelo preditivo (p. ex., erros de predição, valores de AUC) ou os resultados de um experimento (p. ex., diferenças de média entre grupos, intervalos de confiança). Esses relatórios padronizados permitem

acompanhar a evolução do projeto e rapidamente comunicar resultados a interessados. No caso de modelos de *machine learning*, vêm ganhando espaço os chamados *Model Cards*, documentos breves que descrevem de forma acessível o propósito do modelo, os dados usados em seu treinamento, suas métricas de desempenho global e por subgrupos, além de limitações e recomendações de uso. A elaboração de *Model Cards* e a realização de auditorias estratificadas – isto é, verificar como o modelo performa para diferentes segmentos da população (como mulheres vs. homens, diferentes faixas etárias, regiões distintas) – aumentam a transparência e ajudam a identificar possíveis vieses ou desigualdades no desempenho (ATHEY; IMBENS, 2019).

Por fim, especialmente quando modelos preditivos são implantados em sistemas contínuos, é crucial implementar monitoramento de *drift*. Isso significa acompanhar se a distribuição dos dados de entrada ou as características da população-alvo mudam ao longo do tempo, o que poderia degradar a acurácia ou validade do modelo. Caso seja detectado um desvio substancial (por exemplo, um modelo treinado com dados de 2019 pode perder acurácia ao ser aplicado em 2025 se padrões de comportamento mudaram), protocolos de re-treinamento ou recalibração devem ser acionados. Esses cuidados de automação, documentação e monitoramento alinham-se às diretrizes de qualidade e confiabilidade em inteligência artificial, como o *AI Risk Management Framework* do NIST, e garantem que a pesquisa computacional em ciências sociais mantenha traços auditáveis e refutáveis, tal qual se espera de qualquer ciência madura.

Predição vs. Inferência Causal

Uma distinção fundamental a ser compreendida pelos pesquisadores é a diferença entre um exercício de predição e um de inferência causal (SHMUELI, 2010). Embora ambos façam parte do *toolkit* da ciência de dados e utilizem muitas vezes técnicas semelhantes, seus objetivos e critérios de sucesso diferem substancialmente. Leo Breiman (2001) caracterizou isso como duas culturas na modelagem: uma voltada a explicar os dados através de modelos estatísticos tradicionais (por exemplo, regressões paramétricas, que buscam interpretar coeficientes) e outra focada em prever com alta acurácia usando algoritmos flexíveis de *machine learning*, frequentemente tratados como "caixas-pretas". Na prática, essas abordagens são complementares, mas é crucial escolher o caminho de acordo com a pergunta de pesquisa.

Em problemas de predição, a meta é estimar com precisão o valor de uma variável de interesse em novos casos ou no futuro. Por exemplo, predizer a probabilidade de um certo indivíduo estar em situação de pobreza a partir de seus registros de celular (caso de Blumenstock et al., 2015) ou prever quais eleitores têm maior probabilidade de votar dado seu histórico de comportamento. O ênfase está em desempenho fora da amostra: o modelo é bom em generalizar para dados não vistos? Métricas quantitativas são usadas para avaliar isso. Se a variável alvo for contínua, medidas de erro como RMSE (raiz do erro quadrático médio) e MAE (erro absoluto médio) resumem o quão distantes, em média, as predições ficam dos valores verdadeiros. No caso de uma classificação binária

(por exemplo, predizer se alguém está acima ou abaixo da linha da pobreza), avalia-se a capacidade de discriminação através da AUC-ROC (área sob a curva ROC), que indica a probabilidade de o modelo dar pontuação mais alta a um caso positivo do que a um negativo. Também se verifica a calibração das probabilidades preditas, por exemplo com o *Brier score*, que mede o erro quadrático médio das probabilidades previstas vs. ocorrências observadas. Um bom modelo preditivo deve tanto distinguir bem os casos (alta AUC) quanto atribuir probabilidades bem ajustadas à realidade (baixo Brier). Para evitar sobreajuste (*overfitting*), é praxe utilizar validação cruzada ou manter um conjunto de teste segregado, avaliando o modelo em dados que não foram usados no treinamento (JAMES et al., 2021). Em suma, o sucesso na predição é medido pela acurácia e robustez das previsões em novos dados – e não necessariamente pela interpretabilidade do modelo ou significância estatística de parâmetros internos.

Já nos estudos de efeito causal, a pergunta típica é "qual o impacto de X sobre Y?", assumindo um cenário em que se busca isolar a relação causal entre uma variável de intervenção (X) e um desfecho (Y). Aqui, a preocupação central é com validade interna e identificação correta do efeito – isto é, garantir que a variação observada em Y seja devida a X e não a fatores de confusão. Os métodos empregados vão desde experimentos aleatórios controlados até modelos estatísticos com suposições de identificabilidade (ANGRIST; PISCHKE, 2009; GELMAN; HILL, 2007). A avaliação de resultados foca em estimativas pontuais do efeito e sua incerteza: por exemplo, calcular o ATE (*Average Treatment Effect*) – o efeito médio do tratamento X na população – e o ATT (*Average Treatment effect on the Treated*) – efeito médio entre aqueles que receberam X. Esses valores geralmente são expressos em unidades do desfecho ou em pontos percentuais de diferença entre grupos. No estudo de Bond et al. (2012), por exemplo, o tratamento (mensagem social no Facebook) aumentou em aproximadamente 0,39 pontos percentuais a probabilidade de votar dos usuários diretamente expostos, em comparação a um grupo de controle que não recebeu mensagem. Além da estimativa central, importam os intervalos de confiança e testes de significância: se o intervalo de confiança do efeito inclui zero (ou seja, possibilidade de nenhum efeito real), concluímos não haver evidência estatística de impacto; se é estreito e distante de zero, reforça-se a confiança no efeito detectado. Contudo, mais do que "significância", muitas vezes interessa a magnitude substantiva do efeito – se 0,39 p.p. é um impacto pequeno ou relevante em termos eleitorais, por exemplo, deve ser interpretado à luz do contexto.

Outro aspecto peculiar à inferência causal é considerar efeitos indiretos e análises de sensibilidade. Em certas situações, os efeitos de X em Y podem propagar-se por redes ou mecanismos complexos, não se limitando a quem sofreu a intervenção diretamente. Bond et al. (2012) mostraram que o estímulo ao voto gerou um efeito indireto através da rede social: amigos de quem recebeu a mensagem também passaram a votar ligeiramente mais, amplificando o impacto total. Para captar esses efeitos de rede, são necessários desenhos experimentais ou modelos adicionais que estimem a influência de indivíduos tratados sobre não tratados em seu entorno. Ademais, quando não se tem um experimento perfeito, análises de sensibilidade testam quão robustos os resultados são frente a violações de

premissas (por exemplo, quanta variação não observada seria necessária para anular o efeito estimado).

Em síntese, predição e inferência causal demandam mentalidades diferentes quanto à validação. O preditivo preocupa-se em generalizar predições acuradas, usando dados do passado para acertar o futuro próximo, enquanto o causal preocupa-se em isolar relações de causa e efeito, frequentemente sacrificando alguma acurácia preditiva em prol de modelos interpretáveis e identificáveis. Não há hierarquia absoluta entre as abordagens – a escolha depende da pergunta: se queremos antecipar fenômenos ou identificar mecanismos e avaliar políticas. Vale notar que, nas ciências sociais, combinações são possíveis: um pesquisador pode primeiro construir um modelo preditivo para identificar fatores relevantes e depois desenhar um experimento ou análise causal focando nesses fatores. Ferramentas de *machine learning* podem auxiliar na descoberta de heterogeneidades de efeitos (ATHEY; IMBENS, 2019) ou no controle de muitas variáveis de confusão em modelos causais. Em todos os casos, manter claro o objetivo final – explicar ou prever – é essencial para adotar os critérios de validação corretos e interpretar os resultados adequadamente.

Estudos Canônicos: Aplicações de Predição e Causalidade

Predizendo Pobreza a partir de Metadados Telefônicos (Blumenstock, Cadamuro & On, 2015)

O estudo de Blumenstock, Cadamuro e On (2015) tornou-se um marco por demonstrar como big data pode auxiliar no mapeamento de indicadores sociais tradicionalmente escassos. A pergunta central era: é possível estimar a riqueza ou pobreza de indivíduos – e por extensão de regiões – usando seus padrões de uso de telefone celular? Em países em desenvolvimento, onde censos e pesquisas domiciliares são infrequentes, essa abordagem promete obter estimativas atualizadas de pobreza a baixo custo, complementando ou, em parte, substituindo métodos tradicionais.

Para investigar isso, Blumenstock e colegas obtiveram um enorme conjunto de dados de telefonia móvel em Ruanda: registros anonimizados de chamadas e mensagens (CDRs, *call detail records*) de ~1,5 milhão de usuários de uma operadora ao longo de um ano. Esses dados registram, por exemplo, quantas chamadas cada pessoa fez, para quantos contatos distintos, durações, uso de crédito, mobilidade aproximada entre torres, etc. Como não havia um "rótulo" direto de riqueza nas chamadas, os autores conduziram um survey por telefone com uma amostra estratificada de 856 usuários, perguntando sobre bens duráveis, condições de moradia e outras variáveis socioeconômicas. Com essas respostas, construíram um índice de riqueza para cada respondente (essencialmente via análise de componentes principais sobre bens possuídos, aproximando o conceito de riqueza relativa, similar ao índice DHS). Importante: obteve-se consentimento informado dos participantes para vincular suas respostas aos dados de suas ligações (BLUMENS-

TOCK; CADAMURO; ON, 2015, p. 1073). Assim, formou-se um conjunto de dados mesclando atributos telefônicos com nível de riqueza conhecido para esses 856 indivíduos – que pôde então ser usado como amostra de treinamento.

No pipeline deste estudo, a etapa de preparação e engenharia de atributos foi crítica. Dos logs brutos de chamadas, foram extraídas centenas de variáveis descritivas do comportamento de cada usuário. Por exemplo, contaram o número total de chamadas, distinguiram chamadas recebidas de realizadas, o número de contatos únicos, a regularidade de uso ao longo do dia e da semana, volume de gastos com recarga de créditos, e assim por diante. A maioria desses atributos tinha distribuição altamente assimétrica (alguns usuários quase não usavam o telefone, outros usavam intensivamente), então transformações como logaritmos foram aplicadas para atenuar escala. Em seguida, os autores aplicaram técnicas de seleção de atributos para evitar sobreajuste, retendo um subconjunto informativo de variáveis. Com os dados prontos, ajustou-se um modelo preditivo – que não é descrito em minúcias no artigo, mas pode ser interpretado como um algoritmo de *machine learning* (possivelmente regressão regularizada ou árvores de decisão) capaz de estimar o índice de riqueza de um indivíduo a partir de seus atributos telefônicos.

Os resultados foram avaliados por validação cruzada: ao predizer a riqueza dos respondentes não usados no treinamento, obteve-se uma correlação de aproximadamente $r = 0{,}68$ entre o valor predito e o real, indicando boa acurácia preditiva (BLUMENSTOCK; CADAMURO; ON, 2015, p. 1073). Em termos simples, o modelo conseguiu distinguir relativamente bem quem eram os mais pobres e os mais ricos na amostra apenas olhando seus metadados de ligações. Além disso, para validar utilidade prática, os autores aplicaram o modelo treinado a milhões de usuários (todos aqueles na base de CDR) gerando predições de riqueza para a população em geral. Agregando essas predições por região geográfica, conseguiram produzir mapas de distribuição de riqueza em alta resolução, os quais mostraram padrões semelhantes aos captados por pesquisas oficiais em Ruanda – mas atualizados e obtidos com muito menor custo. Esse feito evidenciou o potencial de generalização do modelo: mesmo sem dados censitários recentes, as predições via celular capturaram desigualdades regionais plausíveis.

Apesar do êxito, os autores e a comunidade reconhecem diversos riscos e limitações associados. Primeiro, há o problema da cobertura: quem não usa telefone celular (ou usa de forma compartilhada) permanece invisível ao modelo, possivelmente excluindo exatamente os indivíduos mais pobres e marginalizados – o que poderia levar a subestimar a pobreza extrema se interpretado indevidamente (LAZER et al., 2021, discutem o perigo de gaps em dados digitais). Segundo, questões de privacidade são centrais: embora os dados utilizados fossem anonimizados, o próprio uso de metadados pessoais para inferir informação sensível (riqueza) levanta preocupações éticas. Em contextos reais, aplicar tal modelo requer aderência a princípios legais como finalidade e consentimento – no estudo, o uso dos CDR foi autorizado pela operadora e os participantes do survey consentiram, mas num *deployment* governamental em larga escala seria necessário profundo escrutínio ético e provavelmente consentimento coletivo ou bases legais bem estabelecidas. Terceiro, existe o risco de viés algorítmico: o modelo aprendeu padrões específicos de Ruanda em 2009, e

sua validade em outros países ou tempos depende de quão similares forem os padrões de uso telefônico. Alterações tecnológicas (por exemplo, adoção massiva de smartphones) ou culturais podem reduzir a generalização do modelo.

Por fim, Blumenstock et al. (2015) enfatizam a necessidade de documentação cuidadosa do modelo – equivalente a um "model card" – incluindo suas métricas de erro e limitações, antes de usar seus resultados para guiar políticas públicas. Em suma, esse estudo ilustra o pipeline preditivo completo em uma aplicação social: partiu-se de uma pergunta clara (estimar pobreza), coletaram-se dados digitais e de survey, engenharam-se variáveis, construiu-se e validou-se um modelo, e discutiram-se abertamente os desafios de levar a técnica à prática de forma responsável.

Experimento de Mobilização Social em Rede (Bond et al., 2012)

Como contraponto focado em inferência causal, o estudo conduzido por Bond e colaboradores (2012) exemplifica o poder – e os dilemas – de se realizar experimentos em grande escala através de plataformas digitais. A pesquisa indagou: mensagens de incentivo ao voto exibidas no Facebook podem aumentar efetivamente a participação eleitoral? E esse impacto se propaga através da rede social dos usuários?. Tradicionalmente, estudos de mobilização política envolviam experimentos de campo com dezenas de milhares de pessoas no máximo. Aqui, aproveitando o Facebook como infraestrutura, foi possível engajar 61 milhões de usuários adultos dos EUA em um experimento natural durante o dia da eleição legislativa de 2010 – um volume sem precedentes à época.

O desenho experimental foi simples e elegante. No dia da eleição, usuários foram aleatoriamente designados a um de três grupos: um grupo de tratamento "social" (aproximadamente 60 milhões de pessoas) que viu no topo de seu feed uma mensagem lembrando-os de votar, acompanhada de um botão "Eu Votei" e, crucialmente, de imagens de até seis amigos do Facebook que já haviam clicado no botão; um segundo grupo de tratamento informativo (cerca de 600 mil pessoas) que recebeu essencialmente a mesma mensagem de incentivo ao voto e o botão, mas sem mostrar os amigos – ou seja, sem o componente de pressão social; e um grupo de controle (outros ˜600 mil usuários) que não recebeu nenhuma mensagem especial na plataforma naquele dia. Dessa forma, pôde-se isolar o efeito da exposição à mensagem (comparando tratado vs. controle) e, adicionalmente, o efeito específico da inserção de sinais sociais de amigos (comparando grupo social vs. grupo informativo). Vale notar que os participantes não sabiam explicitamente que estavam em um experimento; a intervenção veio embutida na experiência usual da plataforma, um ponto que posteriormente gerou discussões éticas sobre consentimento e transparência em experimentos online.

A coleta de dados de resultado integrou tanto comportamentos online quanto dados externos. Do próprio Facebook registrou-se quem clicou no botão "Eu Votei" e quem clicou no link para localizar o local de votação (ações interpretadas como medidas de auto-expressão política e busca de informação, respectivamente). Contudo, para medir o efeito final de interesse – o voto real – os pesquisadores realizaram um *record linkage* pós-

eleição: cruzaram a identidade de ~6,3 milhões de usuários (que puderam ser associados com alta confiança) com registros públicos de quem efetivamente votou nas eleições. Isso permitiu verificar, para uma fração grande da amostra, se a pessoa votou de fato (validado oficialmente), e não apenas se declarou ter votado no Facebook.

Os resultados revelaram efeitos estatisticamente significativos embora modestos em magnitude. Comparando o grupo tratamento social com o controle, encontrou-se um aumento absoluto de cerca de 0,39 pontos percentuais na taxa de comparecimento às urnas entre aqueles que receberam a mensagem social (BOND et al., 2012, p. 295). Pode parecer um incremento pequeno – menos de meio ponto percentual – mas num universo de 60 milhões de pessoas, isso se traduz em centenas de milhares de votantes adicionais. Os autores enfatizam que em eleições acirradas até mesmo variações dessa ordem podem alterar resultados (eles citam o exemplo da eleição presidencial de 2000 decidida por poucos votos na Flórida). Mais interessante ainda, ao dissecar o impacto indireto, constatou-se que a maior parte do efeito total veio da influência entre amigos: usuários que viram amigos próximos interagindo com a mensagem acabaram também mais propensos a votar, mesmo que eles próprios não recebessem nenhuma mensagem (o design permitiu estimar esse *spillover* examinando amigos de indivíduos nos diferentes grupos). Ou seja, houve um efeito de rede amplificador: a mensagem social não só motivou diretamente alguns usuários a votar, mas gerou uma cascata de estímulo via laços pessoais, aumentando o comparecimento também entre amigos destes. Essa conclusão – de que "fortes laços" entre amigos próximos foram vetores fundamentais da difusão – conecta-se a teorias sociológicas sobre influência interpessoal e demonstra o valor de medir efeitos além do indivíduo focal.

Do ponto de vista do *pipeline* metodológico, esse trabalho destaca a validação causal rigorosa: a randomização garantiu grupos comparáveis, e foram calculados efeitos com testes de significância (a diferença de 0,39 p.p. era significativa a $p = 0{,}02$, ou seja, muito improvável de ser aleatória) e intervalos de confiança estreitos. A análise também explorou heterogeneidades: por exemplo, verificou-se que o efeito direto da mensagem social (versus mensagem informativa) em clicar "Eu Votei" foi de +2,08% – indicando que ver amigos engajados aumentou a propensão de expressar o voto – enquanto o efeito em buscar local de votação foi positivo porém menor (+0,26%). Assim, diferentes métricas de comportamento apresentaram sensibilidades distintas ao tratamento.

Embora inovador e esclarecedor, o estudo de Bond et al. (2012) suscitou debates sobre ética e governança de experimentos em plataformas. Pelo lado da ética, críticos apontaram que os usuários não deram consentimento explícito para participar de um experimento político; a intervenção foi sutil e aderente à finalidade (incentivar o voto pode ser visto como de interesse público), mas, ainda assim, envolve manipulação deliberada do conteúdo mostrado. Questões sobre transparência e possível necessidade de anuência informada foram levantadas, embora os autores tenham agido dentro dos termos de uso do Facebook. Esse caso prenunciou discussões mais amplas que emergiram em 2014 com o experimento do "contágio emocional" no Facebook, e hoje contribui para demandas de maior supervisão e diretrizes claras para pesquisas em larga escala com usuários de

plataformas (por exemplo, exigência de revisão ética independente mesmo para estudos conduzidos internamente por empresas).

Quanto à replicabilidade, há um desafio evidente: a escala e singularidade do contexto (uma rede social privada com dezenas de milhões de participantes) fazem com que poucos pesquisadores fora das grandes empresas consigam realizar intervenção similar. Isso levanta a questão da governança sobre dados e experimentação: resultados como os de Bond et al. têm implicações sociais importantes (como entender meios eficazes de aumentar participação cívica), porém dependemos de colaborações público-privadas ou de regulamentações que permitam acesso acadêmico a plataformas para serem reproduzidos ou estendidos. Em resposta a isso, tem-se discutido exigências de auditorias independentes em plataformas online e incentivos para que dados agregados de experimentos sejam compartilhados para escrutínio científico.

Em resumo, o estudo de Bond e colegas demonstra como o pipeline "pergunta → dados → método → métricas → governança" se manifesta em um experimento social digital massivo. A pergunta (pode uma intervenção online mudar comportamento político?) guiou um método experimental com coleta de dados tanto online quanto offline, medido por métricas causais (diferenças percentuais, efeitos diretos e em rede) e avaliado quanto à sua significância. Ao final, impôs reflexões profundas sobre governança ética de pesquisas, mostrando que não basta obter resultados – é preciso avaliar como eles foram obtidos e sob quais responsabilidades sociais e legais.

Governança, Ética e Conformidade

Qualquer aplicação de ciência de dados nas ciências sociais deve, do início ao fim, estar ancorada em princípios de ética e conformidade legal. No contexto brasileiro, a Lei Geral de Proteção de Dados (LGPD, Lei n.º 13.709/2018) estabelece diretrizes claras para o tratamento de dados pessoais, que se aplicam mesmo em pesquisas acadêmicas. Quatro princípios centrais merecem destaque: finalidade, adequação, necessidade e transparência/direitos do titular. Finalidade significa que os dados coletados devem ter um propósito específico e legítimo, informado ao titular – não se deve usar dados para objetivos incompatíveis com a razão original da coleta. Adequação refere-se a garantir que o tratamento dos dados seja compatível com o contexto e expectativas do titular (por exemplo, se um indivíduo forneceu dados para um estudo de saúde, não seria adequado reutilizá-los para pesquisa de marketing sem consentimento adicional). Necessidade implica a coleta mínima: deve-se limitar os dados ao estritamente necessário para a pergunta de pesquisa, evitando excessos (em linha com a ideia de minimização mencionada no pipeline técnico). Já transparência e direitos do titular englobam a obrigação de informar claramente os participantes sobre o uso de seus dados e garantir direitos como acesso, retificação e eventual exclusão. Em pesquisa, isso se traduz em fornecer termos de consentimento esclarecedores e vias de contato caso o participante queira mais informações ou retirar seus dados (SALGANIK, 2019, cap. 6, discute extensivamente ética em pesquisa digital).

No caso de dados pessoais sensíveis ou identificáveis, a LGPD exige uma base legal

para o tratamento. Nas atividades de ensino e pesquisa científica, as bases legais mais comuns são o consentimento ou o chamado legítimo interesse, além de previsões específicas para pesquisa. O consentimento, quando viável, deve ser livre, informado e explícito – por exemplo, respondentes de um survey online assinalando que concordam em participar e ter suas respostas utilizadas no estudo. Este foi o caminho adotado por Blumenstock et al. (2015) ao obterem consentimento dos entrevistados para combinar seus dados. Entretanto, nem sempre é possível obter consentimento de milhões de usuários cujos dados foram obtidos em agregados (como CDRs anonimizados ou posts públicos em redes). Nesses casos, o pesquisador pode recorrer ao interesse legítimo, que permite o tratamento de dados sem consentimento desde que atendidos certos critérios: avaliação de impacto que demonstre que o uso dos dados tem baixo risco aos direitos individuais, benefício público ou científico claro, e salvaguardas como anonimização. A LGPD também contém uma exceção para pesquisa (art. 7º, IV, e art. 11, II, c): dados pessoais podem ser tratados para fins de estudo por órgão de pesquisa, garantida, sempre que possível, a anonimização. Essa cláusula exige que os pesquisadores implementem salvaguardas rígidas – por exemplo, usar dados anonimizados ou pseudonimizados, proteger a confidencialidade com rigor, e não usar os dados para fins comerciais ou decisões que afetem os indivíduos participantes. Em suma, mesmo sob base legal, é imperativo adotar uma postura de respeito ao indivíduo: coletar dados de usuários de plataformas ou cidadãos requer considerar o ponto de vista deles (eles concordariam com esse uso? há risco de dano ou discriminação?).

Outro pilar da governança de projetos de ciência de dados é a documentação e auditoria dos processos e dos modelos gerados. A documentação envolve registrar todas as etapas metodológicas (como já destacado, via dicionários de dados, código versionado, etc.) e também produzir relatórios acessíveis que resumam resultados e implicações éticas. Por exemplo, ao treinar um modelo preditivo, é recomendável elaborar um documento de "cartão do modelo" (*Model Card*) descrevendo em linguagem clara o objetivo do modelo, quais dados o alimentaram, que desempenho ele obteve em métricas relevantes e – crucialmente – quaisquer vieses identificados. Essa última parte conecta-se à auditoria: avaliar o desempenho do modelo de forma estratificada. Isso significa verificar, por exemplo, se um modelo de predição de pobreza funciona com a mesma acurácia para diferentes grupos étnicos, regiões ou gêneros. Caso se observe disparidades (digamos, o modelo funciona pior para um determinado grupo minoritário), isso precisa ser reportado e investigado. Auditorias independentes podem ser realizadas, onde terceiros examinariam o algoritmo e seus resultados para garantir conformidade com critérios de não discriminação e equidade. Em experimentos, algo semelhante ocorre: deve-se reportar se a eficácia do tratamento variou substancialmente entre subgrupos (o que pode indicar que a intervenção só beneficiou alguns, por exemplo).

Por fim, iniciativas de marcos normativos em IA vêm reforçar a importância de gestão de risco e transparência. O *framework* de gerenciamento de risco em IA do NIST (NIST AI RMF), por exemplo, recomenda que desenvolvedores e pesquisadores identifiquem possíveis impactos nocivos de seus sistemas, implementem medidas de mitigação e monitorem continuamente o desempenho e consequências. Embora pensado para sistemas

de inteligência artificial em implantação, esses princípios se aplicam igualmente a projetos acadêmicos: é preciso avaliar antecipadamente riscos como vazamento de dados pessoais, uso indevido de predições (p. ex., um modelo de Blumenstock et al. poderia ser mal utilizado para negar crédito a pessoas pobres, o que seria eticamente problemático), ou manipulação de comportamentos sem consentimento (como no experimento de Bond et al., 2012, que exige ponderar o impacto na autonomia dos participantes). Pesquisadores devem estar preparados para prestar contas (*accountability*) de suas escolhas – desde justificar por que certo dado era necessário até mostrar como os resultados foram validados e comunicados aos participantes ou público.

Em resumo, a camada de governança não é acessória, mas parte integral do pipeline de ciência de dados. Integridade científica e responsabilidade social andam juntas: garantir conformidade com a LGPD e princípios éticos não apenas protege os sujeitos da pesquisa, mas também confere maior credibilidade e aceitação aos resultados. Na prática, incorporar governança significa planejar a pesquisa já prevendo mecanismos de proteção de dados, envolver comitês de ética institucionais quando aplicável, ser transparente nas publicações sobre limitações e possíveis vieses, e manter uma postura pró-ativa de revisão e melhoria contínua das práticas à luz do feedback de pares e da sociedade.

Considerações Finais

Ao longo deste relatório, mapeamos como a ciência de dados pode ser aplicada de forma rigorosa e ética às ciências sociais, estruturando o trabalho em um pipeline bem definido que liga perguntas substantivas a dados, métodos, métricas e considerações de governança. Ficou evidente que a pergunta de pesquisa deve ditar as ferramentas empregadas: se o objetivo é preditivo, valorizamos acurácia e generalização; se é explicativo/causal, priorizamos identificação e interpretação. Essa distinção conceitual – reforçada por autores como Breiman (2001) e Shmueli (2010) – nos ajuda a evitar confusões entre "correlação e causalidade" ou entre busca de precisão versus compreensão de mecanismos.

Os estudos de caso analisados, Blumenstock et al. (2015) e Bond et al. (2012), ilustram na prática esses dois polos e também mostram as convergências possíveis. Ambos usam grandes volumes de dados e técnicas computacionais avançadas; ambos requerem validação cuidadosa (seja por validação-cruzada, seja por aleatorização); e ambos enfrentam questões de viés e ética que transcendem a análise numérica pura. Essa interseção é onde reside o potencial mais promissor da ciência de dados social: a capacidade de combinar abordagens. Por exemplo, métodos de *machine learning* podem auxiliar em inferências causais (ATHEY; IMBENS, 2019), identificando padrões ou variáveis relevantes que depois são testadas em desenhos experimentais. Inversamente, conceitos de causalidade podem enriquecer análises preditivas – por exemplo, evitando-se interpretar erroneamente modelos preditivos como se fossem explicativos, ou incorporando variáveis de forma consciente para refletir mecanismos teóricos.

Também destacamos que a competência técnica (programação em Python/R, uso de bibliotecas como NumPy, pandas, scikit-learn, ou *frameworks* do *tidyverse*) deve vir acom-

panhada de uma cultura de reprodutibilidade e colaboração. Ferramentas de controle de versão, documentação clara e adoção de práticas como as sugeridas por Wilson et al. (2017) e Marwick, Boettiger e Mullen (2018) não são extras, mas fundamentos para que o trabalho em ciência de dados seja confiável e passível de escrutínio público. Em áreas socialmente sensíveis, isso é ainda mais crucial: compartilhar dados (quando possível legalmente) e código contribui para acumulatividade do conhecimento e para a confiança nos achados.

Por fim, reforça-se a importância de formação interdisciplinar para os cientistas sociais do século XXI. As leituras integradas – dos manuais clássicos de modelagem estatística (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; JAMES et al., 2021) aos novos horizontes de "texto como dado" (GRIMMER; ROBERTS; STEWART, 2022) e mensuração digital de comportamentos (LAZER et al., 2021) – indicam que há um vasto repertório disponível. O desafio e a oportunidade estão em saber escolher a ferramenta certa para cada pergunta, e fazer as perguntas certas em face das novas ferramentas. A ética e a governança, por sua vez, asseguram que inovar metodologicamente não signifique comprometer valores fundamentais ou a confiança do público.

Em resumo, a ciência de dados aplicada às ciências sociais, quando orientada por um pipeline claro e consciente, permite aprofundar insights sobre a vida social em escala e detalhe antes impossíveis. Ao mesmo tempo, impõe aos pesquisadores a responsabilidade de elevar seus padrões de transparência, rigor e respeito aos indivíduos por trás dos dados. Com o equilíbrio adequado entre técnica e reflexão crítica – exatamente o foco desta aula – os estudantes estarão aptos a aproveitar o melhor dos dois mundos: o poder preditivo das modernas técnicas computacionais e a solidez inferencial e ética da tradição científica social.

Referências Bibliográficas

ANGRIST, Joshua D.; PISCHKE, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009.

ATHEY, Susan; IMBENS, Guido W. Machine learning methods that economists should know about. *Annual Review of Economics*, v. 11, p. 685–725, 2019.

BLUMENSTOCK, Joshua E.; CADAMURO, Gabriel; ON, Robert. Predicting poverty and wealth from mobile phone metadata. *Science*, v. 350, n. 6264, p. 1073–1076, 2015.

BOND, Robert M. *et al.* A 61-million-person experiment in social influence and political mobilization. *Nature*, v. 489, p. 295–298, 2012.

BREIMAN, Leo. Statistical modeling: the two cultures. *Statistical Science*, v. 16, n. 3, p. 199–231, 2001.

GELMAN, Andrew; HILL, Jennifer. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press, 2007.

GENTZKOW, Matthew; KELLY, Bryan; TADDY, Matt. Text as Data. *Journal of Economic Literature*, v. 57, n. 3, p. 535–574, 2019.

GRIMMER, Justin; ROBERTS, Margaret E.; STEWART, Brandon M. *Text as Data: A*

New Framework for Machine Learning and the Social Sciences. Princeton: Princeton University Press, 2022.

HARRIS, Charles R. *et al.* Array programming with NumPy. *Nature*, v. 585, p. 357–362, 2020.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. New York: Springer, 2009.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. *An Introduction to Statistical Learning: With Applications in R*. 2. ed. New York: Springer, 2021.

KUHN, Max; JOHNSON, Kjell. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton: CRC Press, 2019.

LAZER, David M. J. *et al.* Meaningful measures of human society in the twenty-first century. *Nature*, v. 595, p. 189–196, 2021.

MARWICK, Ben; BOETTIGER, Carl; MULLEN, Lincoln. Packaging data analytical work reproducibly using R (and friends). *The American Statistician*, v. 72, n. 1, p. 80–88, 2018.

McKINNEY, Wes. *Python for Data Analysis: Data Wrangling with pandas, NumPy, and Jupyter*. 3. ed. Sebastopol: O'Reilly Media, 2022.

PEDREGOSA, Fabian *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PENG, Roger D. Reproducible research in computational science. *Science*, v. 334, n. 6060, p. 1226–1227, 2011.

SALGANIK, Matthew J. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press, 2019.

SHMUELI, Galit. To explain or to predict? *Statistical Science*, v. 25, n. 3, p. 289–310, 2010.

WICKHAM, Hadley. Tidy data. *Journal of Statistical Software*, v. 59, n. 10, p. 1–23, 2014.

WICKHAM, Hadley; ÇETINKAYA-RUNDEL, Mine; GROLEMUND, Garrett. *R for Data Science*. 2. ed. Sebastopol: O'Reilly Media, 2023.

WILSON, Greg *et al.* Good enough practices in scientific computing. *PLOS Computational Biology*, v. 13, n. 6, e1005510, 2017.

Bibliografia Anotada

SALGANIK, Matthew J. *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press, 2019. Salganik apresenta um panorama abrangente de como a era digital transformou a pesquisa social, articulando quatro métodos-chave (observar, perguntar, experimentar e colaborar) e enfatizando a ética em cada um. Sua tese central é que aproveitar dados e tecnologias novas exige combinar criatividade metodológica com princípios sólidos, tornando esta obra fundamental para estruturar a aula e integrar técnica

com reflexão ética.

BLUMENSTOCK, Joshua E.; CADAMURO, Gabriel; ON, Robert. Predicting poverty and wealth from mobile phone metadata. *Science*, v. 350, n. 6264, p. 1073–1076, 2015. Este artigo demonstra o potencial de usar dados massivos (metadados telefônicos) e aprendizado de máquina para estimar pobreza em países em desenvolvimento. Mostra, com um estudo de caso em Ruanda, como modelos preditivos podem reproduzir estatísticas socioeconômicas com rapidez e baixo custo, servindo de fio condutor na aula para discutir pipeline preditivo, métricas de validação e questões de privacidade e cobertura de dados.

BOND, Robert M. *et al.* A 61-million-person experiment in social influence and political mobilization. *Nature*, v. 489, p. 295–298, 2012. Trabalho pioneiro que relata um experimento randomizado via Facebook com 61 milhões de participantes, investigando se mensagens sociais online aumentam a participação eleitoral. A principal contribuição é evidenciar efeitos causais diretos e em rede no comportamento político. Na aula, ilustra a aplicação de metodologia experimental em larga escala, destacando desenho, mensuração de efeito (pontos percentuais) e reflexões sobre ética e governança em pesquisas digitais.

BREIMAN, Leo. Statistical modeling: the two cultures. *Statistical Science*, v. 16, n. 3, p. 199–231, 2001. Breiman argumenta que estatísticos tradicionalmente focados em modelos paramétricos ("data modeling") deveriam abraçar a cultura "algorítmica" das técnicas preditivas, que frequentemente obtêm maior acurácia em problemas complexos. Essa provocativa distinção entre duas culturas ajuda a classe a entender diferenças de perspectiva entre inferência clássica e ciência de dados moderna, estimulando uma visão mais aberta ao uso de métodos de *machine learning* em ciências sociais.

SHMUELI, Galit. To explain or to predict? *Statistical Science*, v. 25, n. 3, p. 289–310, 2010. Shmueli discute rigorosamente a distinção entre modelos explicativos (focados em testar teorias e identificar relações causais) e modelos preditivos (focados em acurácia de previsão). A autora argumenta que conflitar os dois objetivos leva a erros metodológicos. Esta leitura é relevante para a aula pois fundamenta teoricamente a seção de predição vs. causalidade, clarificando aos alunos por que precisão preditiva e estimativa imparcial de efeitos exigem abordagens e métricas diferentes.

ANGRIST, Joshua D.; PISCHKE, Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009. Nesta obra voltada a praticantes, Angrist e Pischke desmistificam técnicas econométricas para identificação causal, enfatizando métodos como experimentos naturais, variáveis instrumentais e regressão descontínua. A tese central é que econometria aplicada pode ser simples e robusta ("mostly harmless") se focar em desenhos bem identificados. Para a aula, serve como referência clássica sobre inferência causal, reforçando conceitos como ATE/ATT e a importância de desenho de pesquisa para validade.

ATHEY, Susan; IMBENS, Guido W. Machine learning methods that economists should know about. *Annual Review of Economics*, v. 11, p. 685–725, 2019. Artigo que faz ponte entre economia e ciência de dados, apresentando aos economistas técnicas de *machine learning* (árvores, florestas, redes neurais, etc.) e discutindo como elas podem auxiliar tanto

previsões quanto inferências causais (por exemplo, na descoberta de heterogeneidade de tratamentos). A relevância para a aula está em mostrar a convergência atual dos campos: fornece exemplos de uso combinado de ML e econometria, e alerta para cuidados como *overfitting* e validade fora da amostra, alinhando-se aos tópicos de predição e auditoria de modelos.

LAZER, David M. J. *et al.* Meaningful measures of human society in the twenty-first century. *Nature*, v. 595, p. 189–196, 2021. Lazer e colaboradores discutem como as novas fontes de dados digitais (redes sociais, celulares, rastros online) oferecem oportunidades inéditas para medir fenômenos sociais, mas também apresentam desafios de validade, representatividade e governança. O artigo defende indicadores sociais combinando dados tradicionais e big data. Sua inclusão na aula contextualiza por que iniciativas como a de Blumenstock são importantes, ao mesmo tempo alertando sobre "pontos cegos" e a necessidade de métodos híbridos e responsabilidade no uso de dados massivos.

PENG, Roger D. Reproducible research in computational science. *Science*, v. 334, n. 6060, p. 1226–1227, 2011. Peng argumenta que, em pesquisas computacionais, a publicação de resultados deve vir acompanhada do código e dados utilizados, para que outros possam reproduzir exatamente os achados – elevando o padrão de credibilidade da ciência. Ele destaca casos em que a irreprodutibilidade gerou problemas. Para a aula, este texto reforça a motivação das práticas reprodutíveis discutidas (controle de versão, compartilhamento de código, documentação completa), mostrando que não se trata apenas de capricho, mas de um imperativo científico.

WILSON, Greg *et al.* Good enough practices in scientific computing. *PLOS Computational Biology*, v. 13, n. 6, e1005510, 2017. Este artigo propõe um conjunto de "boas práticas mínimas" que qualquer pessoa trabalhando com dados e código deveria adotar – como organizar arquivos de forma clara, usar controle de versão, automatizar tarefas rotineiras e registrar *metadata*. A premissa é que pequenas melhorias sistemáticas evitam erros e aumentam a eficiência do trabalho científico. Na aula, serve como guia prático subjacente às recomendações técnicas do pipeline (especialmente nas seções de preparação, automação e documentação), mostrando que seguir padrões simples pode alavancar a qualidade e colaboração em projetos de ciência de dados.

KUHN, Max; JOHNSON, Kjell. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton: CRC Press, 2019. Kuhn e Johnson oferecem um tratado aprofundado sobre como criar, transformar e selecionar variáveis para melhorar modelos preditivos. Eles enfatizam técnicas para extrair sinal relevante dos dados e evitar armadilhas como sobreajuste ou *leakage*. A relevância desta obra para a aula está no subsídio à etapa de engenharia de atributos: fundamenta o porquê de se investir tempo na preparação dos dados e como abordagens sistemáticas de *feature engineering* podem definir o sucesso de um projeto de predição.

GRIMMER, Justin; ROBERTS, Margaret E.; STEWART, Brandon M. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press, 2022. Os autores apresentam um panorama e uma metodologia integrada para análise de texto nas ciências sociais, combinando conceitos de estatística e apren-

dizado de máquina aplicados a dados textuais (discursos, redes sociais, documentos). A obra defende que texto, antes tratado de forma qualitativa, pode ser transformado em dados quantificáveis para inferir fenômenos sociais em larga escala. Para os propósitos da aula, este livro exemplifica a expansão do pipeline de ciência de dados para além de números tradicionais – mostrando aos alunos que técnicas similares às discutidas (como preparação de dados, modelagem, validação) também se aplicam no domínio de texto, que é cada vez mais relevante em pesquisas sociais.