

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

O gênero cabe em um nome? : cuidados na inferência automática em bases científicas

Jesús P. Mena-Chalco

jmenac@gmail.com

À primeira vista, inferir gênero a partir de um nome parece uma tarefa fácil. Muitas pessoas, ao lerem nomes como Maria, João, Ana ou José, imaginam rapidamente uma associação provável. Sistemas computacionais fazem algo parecido, recebem um nome, consultam uma base de referência e devolvem uma classificação, geralmente “feminino” ou “masculino”, acompanhada de uma probabilidade. Parece direto, mas não é.

Entre o nome escrito e uma classificação automática existem várias decisões metodológicas:



Jesús P. Mena-Chalco

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1-9, jul. 2026. ISSN: 2965-3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

(i) qual base de referência usar, (ii) como tratar acentos, (iii) como lidar com nomes compostos, (iv) quais grafias considerar equivalentes, (v) qual limiar de confiança adotar e, principalmente, (vi) quando reconhecer que não há evidência suficiente. Em outras palavras, o desafio não é apenas classificar nomes. O desafio está em classificar com responsabilidade.

A ciência brasileira, vista a partir da Plataforma Lattes, é formada por milhões de trajetórias e também por milhões de nomes. Em uma análise recente com 9.549.171 CVs coletados em fevereiro de 2026, identificamos 283.669 primeiros nomes distintos. Desses, 185.797 aparecem uma única vez. Ao mesmo tempo, os 100 primeiros nomes mais frequentes concentram 44,6% dos registros. O primeiro nome Maria, sozinho, aparece 278.868 vezes.

Esses números revelam dois mundos convivendo na mesma base. De um lado, há um núcleo de nomes muito frequentes, para os quais a inferência automática costuma funcionar melhor. De outro, há uma longa cauda de nomes pouco frequentes, raros, compostos, estrangeiros, ambíguos ou com grafias variantes. Cada um desses nomes pode aparecer poucas vezes, mas, juntos, eles representam uma parte importante da diversidade nominal da comunidade científica brasileira.



Nomes, frequências e cauda longa / Imagem: Divulga-CI

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1-9, jul. 2026. ISSN: 2965-3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

Mas, por que isso importa?

Porque estudos sobre desigualdades de gênero na ciência precisam, muitas vezes, analisar grandes bases de dados. Queremos saber, por exemplo, como mulheres e homens aparecem em diferentes áreas do conhecimento, em redes de colaboração, em posições de autoria, em bolsas, orientações ou trajetórias acadêmicas. Mas, nem sempre a informação sobre gênero está disponível publicamente. Na Plataforma Lattes, por exemplo, o CV público não informa esse atributo. Por isso, muitos estudos recorrem a métodos indiretos, usando primeiros nomes como pista estatística.

Essa estratégia pode ser útil, especialmente em análises agregadas. No entanto, ela exige cuidado. Uma base nacional, pública e importante como a base de nomes do IBGE pode ser

uma referência valiosa para o contexto brasileiro. Serviços externos, como ferramentas internacionais de inferência por nome, também podem ajudar em comparações. Mas nenhuma dessas fontes deve ser tratada como resposta definitiva. Elas são referências estatísticas, não verdades individuais.

Nomes são objetos culturais, históricos e linguísticos. Eles mudam conforme região, geração, família, migração, religião, modismos, preferências pessoais e circulação global. Por isso, a inferência automática simplifica um fenômeno muito mais complexo do que parece.

Um exemplo simples está nos nomes compostos. Se um sistema considera apenas a primeira palavra, pode perder informação relevante. Em nomes como Maria Clara, Ana Paula, João Pedro ou Luiz Felipe, o segundo elemento pode reforçar

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1-9, jul. 2026. ISSN: 2965-3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

a interpretação do primeiro. Em outros casos, pode alterar a granularidade da análise. Maria, por exemplo, é um primeiro nome extremamente frequente, mas combinações como Maria Clara, Maria Eduarda, Maria José ou Maria do Socorro carregam informações distintas que podem ser perdidas quando o método observa apenas o primeiro nome.

A dificuldade também aparece nas grafias variantes. Isabella e Izabella, Rafaela e Rafaella, Cristian, Crhistian e Crystian, Stephanie e Stephany podem parecer formas próximas para uma pessoa familiarizada com esses nomes. Para um método baseado apenas em correspondência exata, porém, elas podem ser tratadas como nomes completamente diferentes. Isso pode reduzir a cobertura do método ou gerar classificações frágeis.

Um exemplo ajuda a entender o problema. Pense em um nome como Roniberto. Para quem conhece a pessoa, ou para quem está familiarizado com certos padrões de nomes no Brasil, pode parecer relativamente fácil imaginar uma associação provável. Para um sistema automático, no entanto, a pergunta é outra: esse nome aparece na base de referência? Com quantas ocorrências? A distribuição observada é suficientemente concentrada em uma categoria? Há nomes parecidos que poderiam ajudar na aproximação? Ou a melhor resposta seria reconhecer que não há evidência suficiente?

Esse tipo de situação é comum na cauda longa dos nomes. Nomes como Roniberto, Josenildo, Edivânia, Rosivânia, Advane ou Marileide podem carregar pistas culturais para leitores humanos, mas podem ser

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1–9, jul. 2026. ISSN: 2965–3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

pouco frequentes, ausentes ou instáveis em bases de referência. Nesses casos, um método responsável não deve forçar uma resposta apenas para aumentar a cobertura. ***Ele precisa preservar o direito à incerteza: quando a evidência é baixa, a resposta mais adequada pode ser “não sei com segurança”.***

Há ainda nomes ambíguos ou dependentes do contexto cultural, como Ariel, Dominique, Sasha, Andrea ou Alex. Em alguns países, um nome pode ser mais associado a uma categoria; em outros, pode ter distribuição diferente. Mesmo dentro de um mesmo país, a interpretação pode variar conforme geração, região ou grupo cultural. Por isso, a inferência automática não deve ser entendida como identificação individual, mas como aproximação probabilística para análises agregadas.



Exemplos de nomes com ambiguidade e baixa representação / Imagem: Divulga-CI

Essa distinção é fundamental. Inferir gênero por nome não é o mesmo que conhecer a identidade de gênero de uma pessoa. O nome não determina identidade. Além disso, categorias binárias são insuficientes para representar a diversidade das experiências individuais. Por esse motivo, a inferência por nome deve ser usada apenas em estudos

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1–9, jul. 2026. ISSN: 2965–3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

agregados, com transparência, incerteza explícita e limites bem definidos. Ela não deve ser usada para classificar indivíduos em decisões administrativas, avaliativas ou institucionais.

Em bases grandes, pequenos erros podem se acumular. Se os nomes classificados incorretamente estiverem distribuídos de forma desigual entre regiões, áreas do conhecimento, origens linguísticas ou grupos sociais, os resultados finais podem produzir distorções. Um método aparentemente eficiente no total pode funcionar muito bem para nomes frequentes e muito mal para nomes raros. Por isso, não basta medir quantos currículos foram classificados. É necessário perguntar: quais nomes ficaram sem classificação? Em quais grupos de frequência o método errou mais? Onde a incerteza se concentrou?

INFERÊNCIA DE GÊNERO

A inferência automática pode apontar tendências, mas não certezas.



ASSOCIAÇÃO PROVÁVEL

O nome se associa mais fortemente a um gênero.



AMBÍGUO

O nome é compartilhado por mais de um gênero ou apresenta padrão misto.



EVIDÊNCIA INSUFICIENTE

Dados escassos ou inexistentes. Não é possível inferir com segurança.



NÃO SEI COM SEGURANÇA

Trate resultados com cautela.

Cuidados na inferência automática de gênero em nomes / Imagem: Divulga-CI

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1–9, jul. 2026. ISSN: 2965–3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

A organização da informação científica precisa lidar cada vez mais com esse tipo de problema. Bases de dados são fundamentais para compreender a ciência, mas elas não falam sozinhas. Entre o dado bruto e o indicador final existem escolhas: como normalizar nomes, como tratar acentos, como lidar com nomes compostos, qual base usar como referência, qual limiar de confiança adotar, quando classificar e quando se abster. Essas decisões parecem técnicas, mas têm consequências sociais.

Estudar os nomes da Plataforma Lattes, portanto, não é apenas uma curiosidade. É uma forma de melhorar a qualidade dos sistemas de informação científica e de produzir indicadores mais confiáveis sobre a ciência brasileira. Também é uma oportunidade para mostrar que métodos

computacionais devem ser transparentes, avaliáveis e sensíveis às limitações dos dados.

O problema não é usar dados.

O problema é usá-los como se não tivessem incertezas.

Em estudos sobre gênero na ciência, essa cautela é ainda mais necessária. A inferência automática pode ser uma ferramenta útil, mas deve ser tratada como aproximação probabilística, não como verdade individual. Em muitos casos, a resposta mais científica e mais ética não é classificar a qualquer custo. É reconhecer, com responsabilidade: aqui, eu não sei com segurança.

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1–9, jul. 2026. ISSN: 2965–3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

Nota do autor

* Neste texto, a expressão “inferência de gênero” é usada em sentido metodológico e agregado. Ela se refere à estimativa probabilística feita a partir de primeiros nomes, e não à identificação da identidade de gênero de pessoas específicas.

** A “cauda longa” ocorre quando poucos elementos aparecem muitas vezes, enquanto muitos elementos aparecem poucas vezes. No caso dos nomes, isso significa que poucos primeiros nomes concentram grande parte dos registros, enquanto milhares de nomes aparecem raramente.

Referências

LARIVIÈRE, Vincent; NI, Chaoqun; GINGRAS, Yves; CRONIN, Blaise; SUGIMOTO, Cassidy R.

Bibliometrics: global gender disparities in science. **Nature**, London, v. 504, n. 7479, p. 211–213, 12 dez. 2013. Disponível em:

<https://doi.org/10.1038/504211a>.

Acesso em: 4 jul. 2026.

LOCKHART, Jeffrey W.; KING, Molly M.; MUNSCH, Christin L.

Name-based demographic inference and the unequal distribution of misrecognition.

Nature Human Behaviour, London, v. 7, n. 7, p. 1084–1095, jul. 2023. Disponível em:

<https://doi.org/10.1038/s41562-023-01587-9>. Acesso em: 4 jul. 2026.

SANTAMARÍA, Lucía; MIHALJEVIĆ, Helena. Comparison and benchmark of name-to-gender

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1–9, jul. 2026. ISSN: 2965–3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.

DIVULGA-CI

Revista de Divulgação Científica em Ciência da Informação

PERSPECTIVA

inference services. **PeerJ**

Computer Science, London, v. 4,

e156, 2018. Disponível em:

<https://doi.org/10.7717/peerj-cs.156>

. Acesso em: 4 jul. 2026.

TEODORO, Olívia Rico;
MENA-CHALCO, Jesús P. A
distribuição de nomes de
pesquisadores na Plataforma
Lattes: frequência e unicidade em
larga escala. In: ENCONTRO
BRASILEIRO DE BIBLIOMETRIA E
CIENTOMETRIA, 10., 2026, Curitiba.
Trabalho aceito para
apresentação na modalidade
Pecha Kucha. Curitiba:
Universidade Federal do Paraná,
2026.

Redação: Jesús P. Mena-Chalco

Fotografia: Jesús P. Mena-Chalco

Diagramação: Naiara Raíssa da
Silva Passos

Sobre o autor

Jesús P. Mena-Chalco

Professor do curso de
Bacharelado em Ciência e
Tecnologia e do Programa de
Pós-Graduação em Ciência da
Computação junto à
Universidade Federal do ABC.
Pesquisador associado da Rede
Nacional de Ciência para a
Educação.

Doutor e mestre em Ciência da
Computação pela Universidade
de São Paulo. Mestre em Ciência
da Computação pela
Universidade de São Paulo.
Graduado em Engenharia de
Sistemas pela Universidade
Nacional de San Agustín.
Pós-doutor em Ciência da
Computação pela Universidade
de São Paulo.

Divulga-CI, Porto Velho, v. 4, n. 7, p. 1–9, jul. 2026. ISSN: 2965–3436



Todo o conteúdo da Revista Divulga-CI está sob uma licença Creative Commons Atribuição 4.0 Internacional. Ao serem publicados por esta Revista, os conteúdos são de Acesso Aberto e de livre uso para compartilhar e adaptar, sendo necessário dar o crédito apropriado, prover um link para a licença e indicar se mudanças foram feitas.