Universidade Federal do Espírito Santo
Centro Tecnológico
Programa de Pós-Graduação em Informática

Impacto da Adequação à Lei Geral de Proteção de Dados Pessoais na Metrificação da Qualidade de Dados

Leandro Furlam Turi

Giovanni Comarela

Thu Oct 25 20:26:07 BST 2012
Steve Song:

Open Knowledge Foundation

> Dados sobre grupos marginalizados dizem que os ciganos na Europa ou tribos específicas na África podem ser potencialmente uma ferramenta de opressão.
>
> Trata-se de informação pessoal? Ou são dados que só são significativos em conjunto?

Conglomerado Santander: Política de Privacidade

Conjunto de dados: Nome; Nome Social; Data de nascimento; Sobrenome; CPF; CNPJ; Carteira de identidade; Carteira Nacional de Habilitação (CNH); Carteira de Trabalho e Previdência Social (CTPS); Cédula de Identidade de Estrangeiro (CIE); Registro Nacional de Estrangeiros (RNE); Protocolo de solicitação da CIE Protocolo do pedido de refúgio de que trata o art. 21 da Lei nº 9.474 de 22 de julho de 1997; Passaporte Guia de acolhimento de que trata o § 3 do artº 101 da Lei nº 8.069, de 13 de julho de 1990 (Estatuto da criança e do Adolescente); Idade; Nacionalidade; E-mail; Naturalidade; Nome da mãe; Nome do pai; Endereço Residencial; Endereço Comercial; Estado Civil; Sexo; Telefones residenciais; Telefones comerciais; Telefones celulares; Condição Pessoal (espólio, interdito, deficiente, etc...); Renda; Patrimônio; IMEI do celular; Origem racial; Geolocalização; Foto; Filmagens; Biometria; Áudio/voz; Pessoa Politicamente exposta; Título de eleitor; Documentos profissionais (CREA, OAB e etc.); PIS/NIS - Programa de integração social; Profissão; Formação Acadêmica; IP; Cookies; Dados Transacionais de Contas; Dados Transacionais de Cartão; Dados Transacionais de Operação de Crédito.

Conglomerado Santander: Política de Privacidade

Conjunto de dados: Nome; Nome Social; Data de nascimento; Sobrenome; CPF; CNPJ; Carteira de identidade; Carteira Nacional de Habilitação (CNH); Carteira de Trabalho e Previdência Social (CTPS); Cédula de Identidade de Estrangeiro (CIE); Registro Nacional de Estrangeiros (RNE); Protocolo de solicitação da CIE Protocolo do pedido de refúgio de que trata o art. 21 da Lei nº 9.474 de 22 de julho de 1997; Passaporte Guia de acolhimento de que trata o § 3 do artº 101 da Lei nº 8.069, de 13 de julho de 1990 (Estatuto da criança e do Adolescente); Idade; Nacionalidade; E-mail; Naturalidade; Nome da mãe; Nome do pai; Endereço Residencial; Endereço Comercial; Estado Civil; Sexo; Telefones residenciais; Telefones comerciais; Telefones celulares; Condição Pessoal (espólio, interdito, deficiente, etc...); Renda; Patrimônio; IMEI do celular; Origem racial; Geolocalização; Foto; Filmagens; Biometria; Áudio/voz; Pessoa Politicamente exposta; Título de eleitor; Documentos profissionais (CREA, OAB e etc.); PIS/NIS - Programa de integração social; Profissão; Formação Acadêmica; IP; Cookies; Dados Transacionais de Contas; Dados Transacionais de Cartão; Dados Transacionais de Operação de Crédito.

Uma das dificuldades fundamentais é que as informações extraídas podem ser tendenciosas, ruidosas, desatualizadas, incorretas, enganosas e, portanto, não confiáveis.

BERTI-EQUILLE; BORGE-HOLTHOEFER, 2015

OBJETIVOS

Analisar a viabilidade e a adaptabilidade de processos de mensuração de qualidade de dados através de diferentes formas de adequação à LGPD

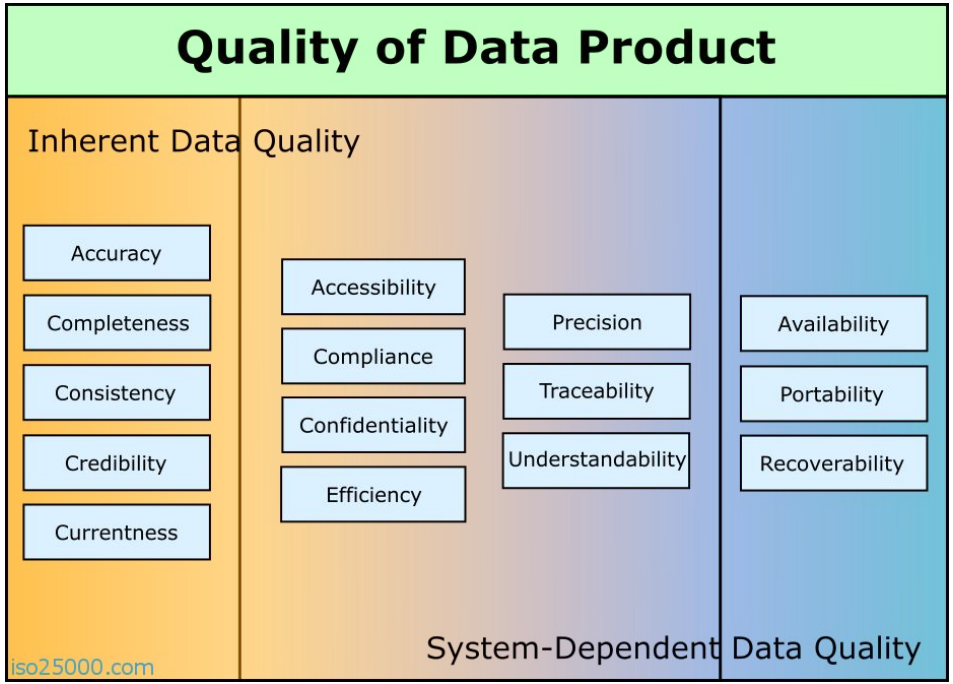
1. categorizar abordagens e requisitos gerais atuais para cada métrica de qualidade de dados, demonstrando estruturas e desafios;
2. resumir ferramentas, métodos e tecnologias existentes em qualidade de dados;
3. implementar e automatizar tais técnicas em conjunto com algoritmos de suporte para o processo de qualidade de dados e automatização do processo;
4. comparar e avaliar os resultados obtidos com cada técnica de anonimização em relação à base original.

Como a LGPD influencia na mensuração da qualidade de uma base de dados? Isto reflete-se em projetos de ciência de dados?

Qualidade de dados

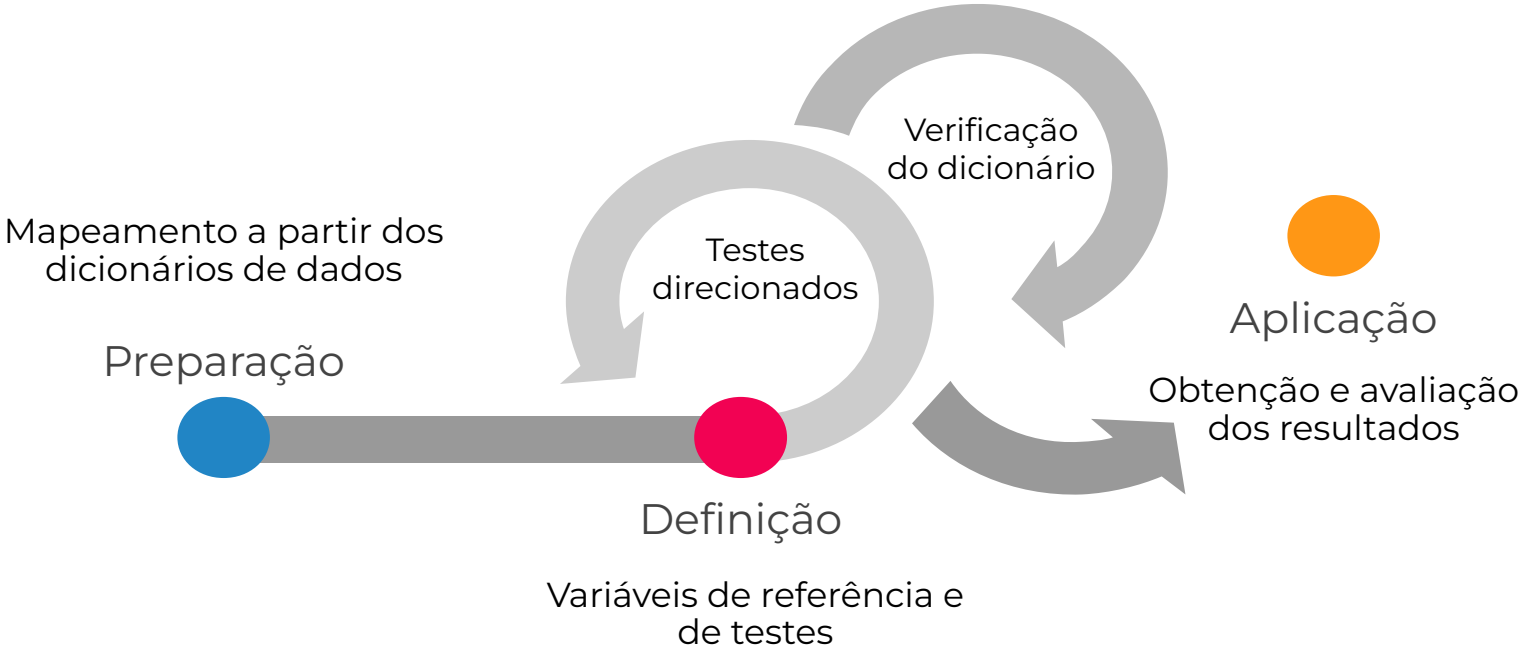
- Pré-requisito para se obter bons resultados no processo de análise de dados;
- Conceito multidimensional, pois diferentes aspectos devem ser considerados;
- Diretamente relacionada à confiabilidade dos dados de entrada.

Características definidas pela ISO/IEC 25012

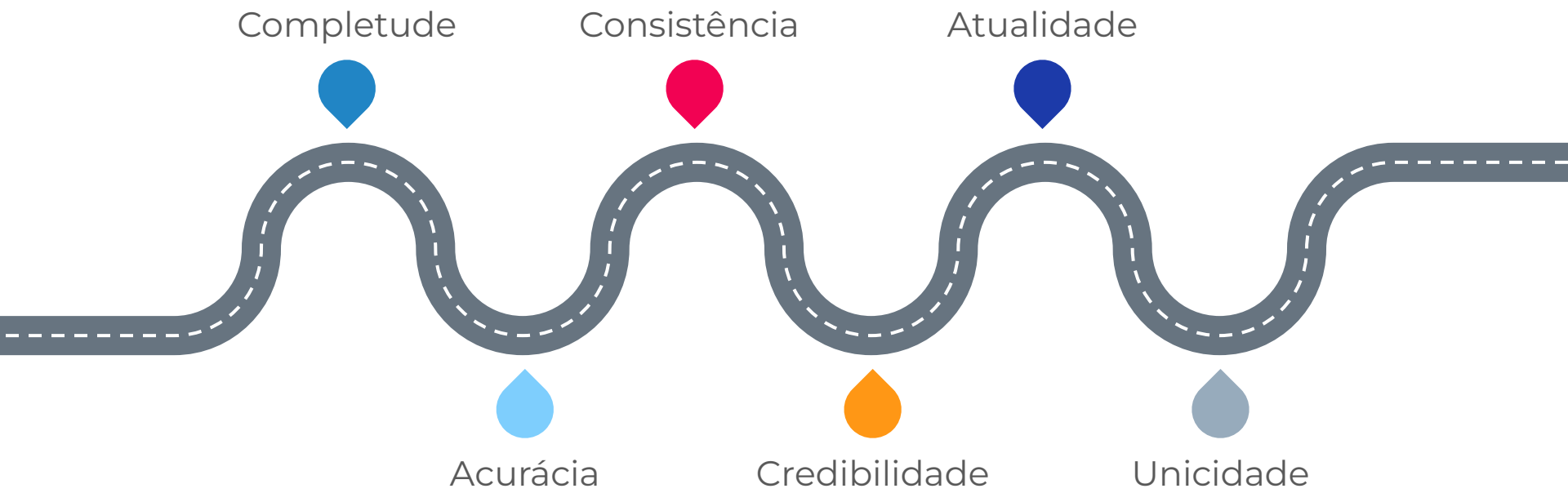


https://iso25000.com/

Metrificação da qualidade de dados



Métricas de qualidade



LEI Nº 13.709, DE 14 DE AGOSTO DE 2018

Lei Geral de Proteção de Dados Pessoais (LGPD)

Art. 2º A disciplina da proteção de dados pessoais tem como fundamentos:

I - o respeito à privacidade;

II - a autodeterminação informativa;

III - a liberdade de expressão, de informação, de comunicação e de opinião;

IV - a inviolabilidade da intimidade, da honra e da imagem;

V - o desenvolvimento econômico e tecnológico e a inovação;

VI - a livre iniciativa, a livre concorrência e a defesa do consumidor; e

VII - os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais.

LEI Nº 13.709, DE 14 DE AGOSTO DE 2018

Lei Geral de Proteção de Dados Pessoais (LGPD)

Art. 5º Para os fins desta Lei, considera-se:

I - dado pessoal: informação relacionada a pessoa natural identificada ou identificável;

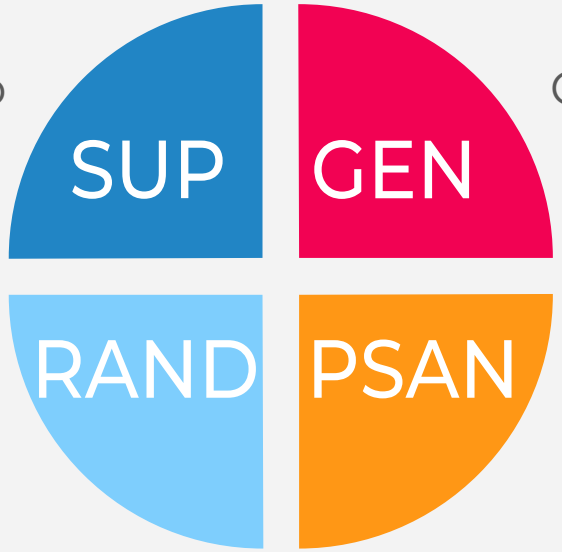
II - dado pessoal sensível: dado pessoal sobre origem racial ou étnica, convicção religiosa, opinião política, filiação a sindicato ou a organização de caráter religioso, filosófico ou político, dado referente à saúde ou à vida sexual, dado genético ou biométrico, quando vinculado a uma pessoa natural;

III - dado anonimizado: dado relativo a titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião de seu tratamento;

Técnicas de anonimização

Supressão

Generalização

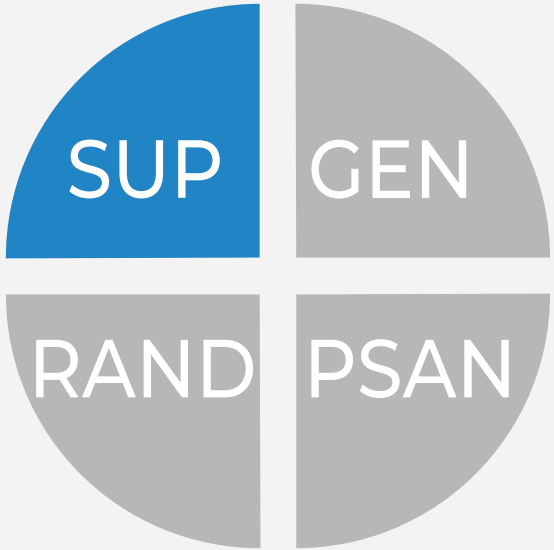


Randomização

Pseudoanonimização

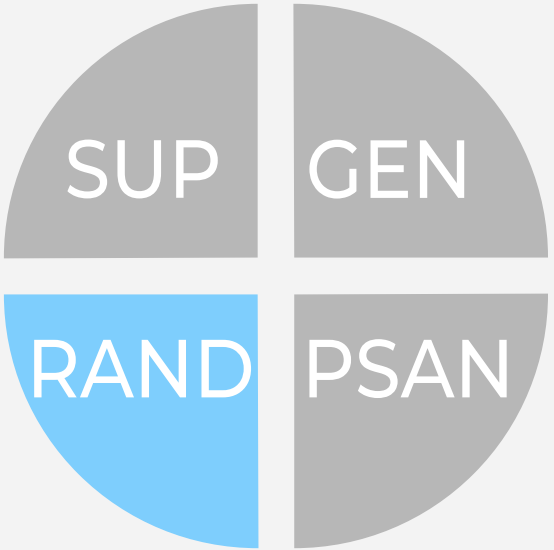
Supressão

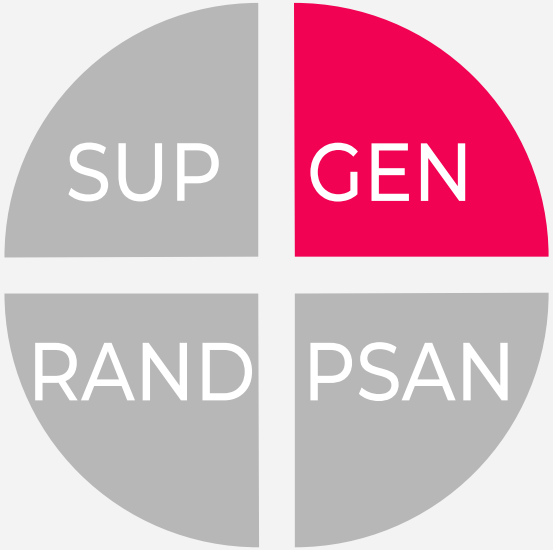
nome	*cpf*
Bruno Santos	~~123.456.789-10~~
Maria Silva	~~234.567.891-01~~



Randomização

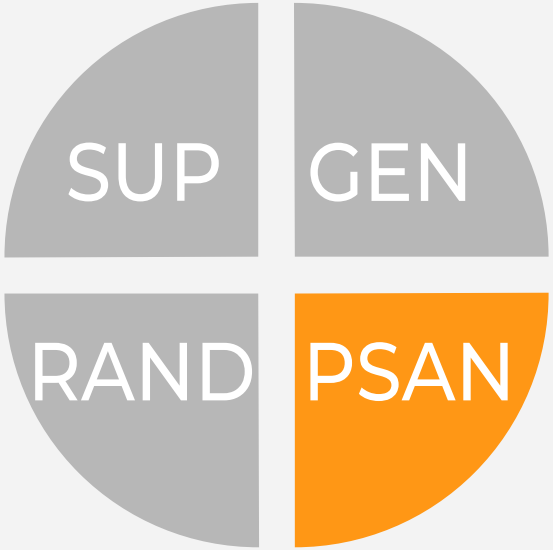
nome	*cpf*
Bruno Santos	123.456.789-10
Maria Silva	234.567.891-01





Generalização

nome	*idade*
Bruno Santos	< 25
Maria Silva	> 60



Pseudoanonimização

nome	*cpf*
Bruno Santos	506cfdfe5e9ed27...
Maria Silva	5647d122bc00c4...

SHA224

METODOLOGIA

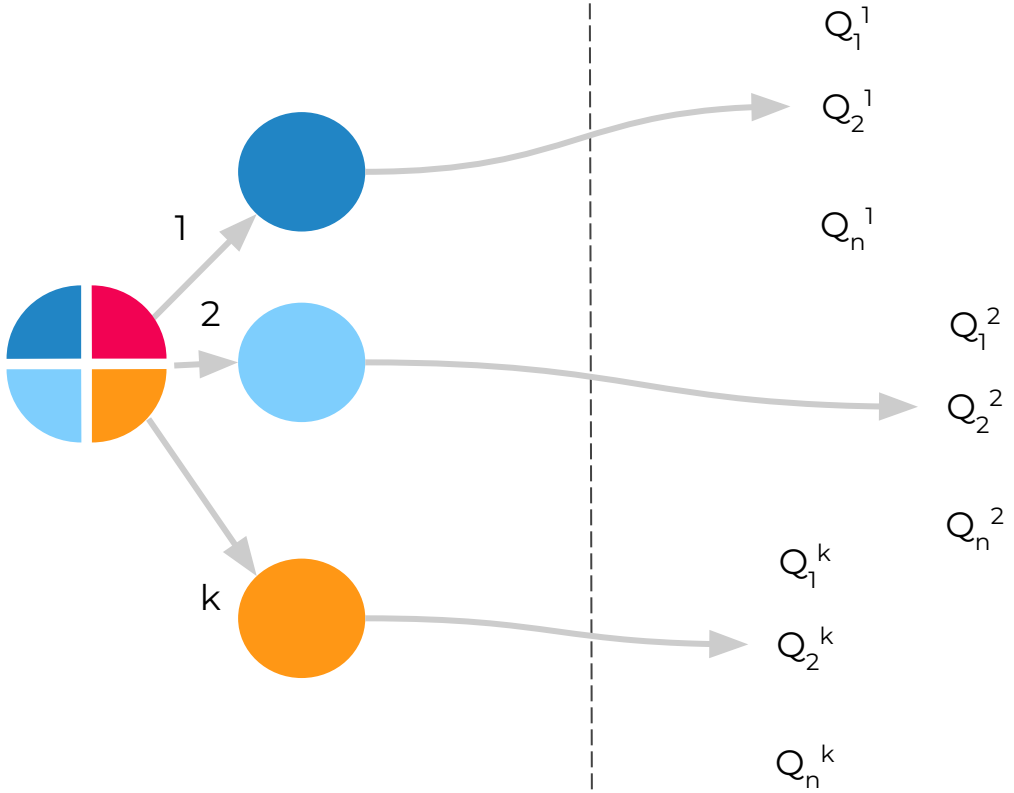
Qualidade desejável

Q_1

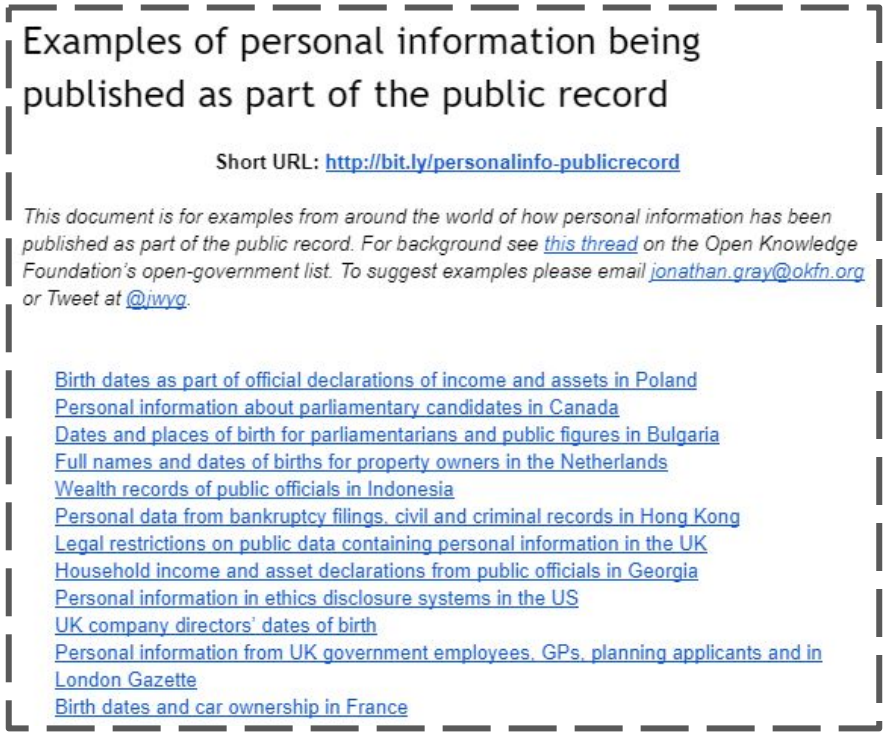
Q_2

Q_n

LGPD



Bases de dados



Examples of personal information being published as part of the public record

Short URL: http://bit.ly/personalinfo-publicrecord

This document is for examples from around the world of how personal information has been published as part of the public record. For background see this thread on the Open Knowledge Foundation's open-government list. To suggest examples please email jonathan.gray@okfn.org or Tweet at @jwyg.

Birth dates as part of official declarations of income and assets in Poland
Personal information about parliamentary candidates in Canada
Dates and places of birth for parliamentarians and public figures in Bulgaria
Full names and dates of births for property owners in the Netherlands
Wealth records of public officials in Indonesia
Personal data from bankruptcy filings, civil and criminal records in Hong Kong
Legal restrictions on public data containing personal information in the UK
Household income and asset declarations from public officials in Georgia
Personal information in ethics disclosure systems in the US
UK company directors' dates of birth
Personal information from UK government employees, GPs, planning applicants and in London Gazette
Birth dates and car ownership in France

http://bit.ly/personalinfo-publicrecord

athleteEvents	dados básicos sobre atletas e resultados de medalhas de Atenas 1896 a Rio 2016
Canada	informações pessoais sobre candidatos parlamentares no Canadá
eleicoes	dados sobre os candidatos brasileiros das eleições de 1996 a 2020, no estado do Espírito Santo
EuropeanSoccer	atributos de jogadores do futebol profissional Europeu
Poland	declarações de rendimentos e bens na Polônia
Sinasc	Sistema de Informação sobre nascidos vivos no Brasil, no estado do Espírito Santo

athleteEvents

Dados básicos sobre atletas e resultados de medalhas de Atenas 1896 a Rio 2016

id	name	sex	age	height	...	sport	event	medal	region	notes
1	A Dijiang	M	24.0	180.0	...	Basketball	Basketball Men's Basketball	None	China	None
2	A Lamusi	M	23.0	170.0	...	Judo	Judo Men's Extra-Lightweight	None	China	None
602	Abdoureheman	M	22.0	182.0	...	Boxing	Boxing Men's Middleweight	None	China	None
1463	Ai Linuer	M	25.0	160.0	...	Wrestling	Wrestling Men's Lightweight, Greco-Roman	None	China	None
1464	Ai Yanhan	F	14.0	168.0	...	Swimming	Swimming Women's 200 metres Freestyle	None	China	None
...	...	..	...	...	...	...	...	...	...	...
120575	Mamorallo Tjoka	F	23.0	150.0	...	Athletics	Athletics Women's Marathon	None	Lesotho	None
120575	Mamorallo Tjoka	F	27.0	150.0	...	Athletics	Athletics Women's Marathon	None	Lesotho	None
122166	M'apotlaki Ts'elho	F	15.0	NaN	...	Athletics	Athletics Women's 4 x 100 metres Relay	None	Lesotho	None
122215	Lefa Tsapi	M	23.0	170.0	...	Boxing	Boxing Men's Welterweight	None	Lesotho	None
122299	Mosolesa Tsie	M	20.0	175.0	...	Boxing	Boxing Men's Welterweight	None	Lesotho	None

fonte: Kaggle
[270767 linhas x 17 colunas]

Canada

Informações pessoais sobre candidatos parlamentares no Canadá

	date_of_birth	place_of_birth	...	language_preference	cause_of_death
0	02/01/1922	Sherbrooke, Quebec, Canada	...	None	None
1	10/09/1966	Aiha, Beqaa, Lebanon	...	English, French, Arabic	None
2	1821-03-12	St. Andrews, Lower Canada	...	None	cancer
3	1822-01-09	York, Upper Canada	...	None	None
4	1850-05-03	Lennoxville, Canada East	...	None	None
5	03/07/1908	Newcastle, New Brunswick, Canada	...	None	None
6	None	None	...	None	None
7	1853-08-13	Cornwall Township, Canada West	...	None	None
8	None	None	...	None	None
9	27/08/1915	Carleton, Quebec, Canada	...	None	None
10	None	None	...	None	None
11	None	None	...	None	None

fonte: Parlamento canadense
[50 linhas x 8 colunas]

eleicoes

Dados sobre os candidatos brasileiros das eleições de 1996 a 2020, no estado do Espírito Santo

	ano	cpf	data_nascimento	...	sigla_unidade_federativa	sigla_unidade_federativa_nascimento	titulo_eleitoral
2378	1996	1.000000e+00	1954-10-21	...	ES	None	1
2379	1996	1.000000e+00	1962-04-22	...	ES	None	1
2380	1996	1.000000e+00	1962-11-19	...	ES	None	1
2381	1996	1.000000e+00	1952-10-28	...	ES	None	1
2382	1996	1.000000e+00	1957-02-15	...	ES	None	1
...	...	...	...	...	...	...	...
2385595	2020	1.328582e+10	1988-10-19	...	ES	ES	30474201457
2385596	2020	7.577511e+09	1975-10-12	...	ES	ES	16241531430
2385597	2020	7.253578e+10	1962-03-04	...	ES	ES	18435691465
2385598	2020	7.919510e+09	1979-09-25	...	ES	ES	23395951481
2385599	2020	3.912826e+10	1949-07-27	...	ES	ES	7531441430

fonte: Tribunal Superior Eleitoral
[59507 linhas x 25 colunas]

EuropeanSoccer

Atributos de jogadores do futebol profissional Europeu

	player_name	birthday	height	weight	...	gk_handling	gk_kicking	gk_positioning	gk_reflexes
0	Aaron Appindangoye	1992-02-29 00:00:00	182.88	187	...	11.0	10.0	8.0	8.0
1	Aaron Appindangoye	1992-02-29 00:00:00	182.88	187	...	11.0	10.0	8.0	8.0
2	Aaron Appindangoye	1992-02-29 00:00:00	182.88	187	...	11.0	10.0	8.0	8.0
3	Aaron Appindangoye	1992-02-29 00:00:00	182.88	187	...	10.0	9.0	7.0	7.0
4	Aaron Appindangoye	1992-02-29 00:00:00	182.88	187	...	10.0	9.0	7.0	7.0
...	...	...	...	...	...	...	...	...	...
183973	Zvjezdan Misimovic	1982-06-05 00:00:00	180.34	176	...	20.0	84.0	20.0	20.0
183974	Zvjezdan Misimovic	1982-06-05 00:00:00	180.34	176	...	20.0	73.0	20.0	20.0
183975	Zvjezdan Misimovic	1982-06-05 00:00:00	180.34	176	...	20.0	73.0	20.0	20.0
183976	Zvjezdan Misimovic	1982-06-05 00:00:00	180.34	176	...	20.0	73.0	20.0	20.0
183977	Zvjezdan Misimovic	1982-06-05 00:00:00	180.34	176	...	9.0	78.0	7.0	15.0

fonte: Kaggle
[183978 linhas x 43 colunas]

Poland

Declarações de rendimentos e bens na Polônia

	date_and_place_of_birth	occupation	...	district	number_of_votes
0	1972-05-16, Kraków	nauczyciel akademicki	...	None	None
1	1959-01-04, Krzeszowice	poseł	...	13 Kraków	29 686 (4,57%)
2	1974-05-30, Dąbrowa Górnicza	samorządowiec	...	32 Katowice	12 148 (3,62%)
3	1961-06-26, Elbląg	dziennikarz	...	25 Gdańsk	41 795 (7,90%)
4	1949-01-19, Duszniki Wlkp.	rolnik	...	38 Piła	14 438 (4,14%)
..	...	...	...	...	...
455	1969-08-05, Świebodzice	radca prawny	...	2 Wałbrzych	10 688 (3,78%)
456	1973-01-13, Białystok	prawnik	...	24 Białystok	12 141 (2,33%)
457	1970-08-23, Chełm	politolog	...	20 Waszawa	8 665 (8665%)
458	1954-10-01, Plimkaim	menedżer	...	1 Legnica	7 694 (1,78%)
459	1983-06-11, Warszawa	prawnik	...	19 Warszawa	18 864 (1,37%)

fonte: Parlamento polonês
[460 linhas x 10 colunas]

Sinasc

Sistema de Informação sobre nascidos vivos no Brasil, no estado do Espírito Santo

	loc_nasc	cod_mun_nasc	idade_mae	est_civ_mae	esc_mae	...	tp_func_resp	td_doc_resp	dt_declarac	par_idade	kotelchuck
0	1.0	1600303.0	18.0	2.0	2.0	...	NaN	NaN	NaN	NaN	NaN
1	1.0	1721208.0	21.0	2.0	3.0	...	NaN	NaN	NaN	NaN	NaN
2	1.0	2112209.0	19.0	2.0	3.0	...	NaN	NaN	NaN	NaN	NaN
3	1.0	2101202.0	26.0	2.0	4.0	...	NaN	NaN	NaN	NaN	NaN
4	1.0	2111300.0	32.0	1.0	2.0	...	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...
1135347	1.0	330455.0	30.0	2.0	5.0	...	4.0	5.0	19102020.0	0.0	5.0
1135348	1.0	351830.0	23.0	1.0	4.0	...	2.0	3.0	16092020.0	1.0	2.0
1135349	1.0	352690.0	26.0	2.0	5.0	...	5.0	4.0	17082020.0	1.0	5.0
1135350	1.0	410690.0	32.0	1.0	1.0	...	2.0	3.0	30092020.0	1.0	5.0
1135351	1.0	510795.0	34.0	2.0	5.0	...	2.0	3.0	20082020.0	1.0	2.0

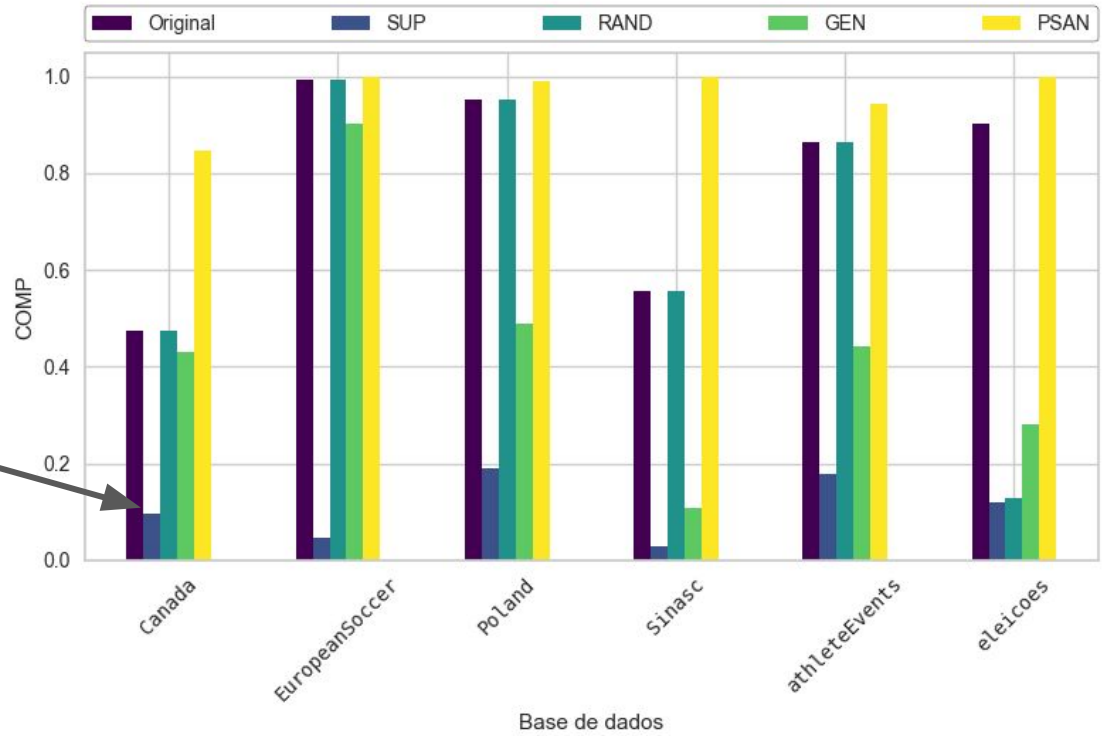
fonte: DataSUS

[1135352 linhas x 70 colunas]

RESULTADOS

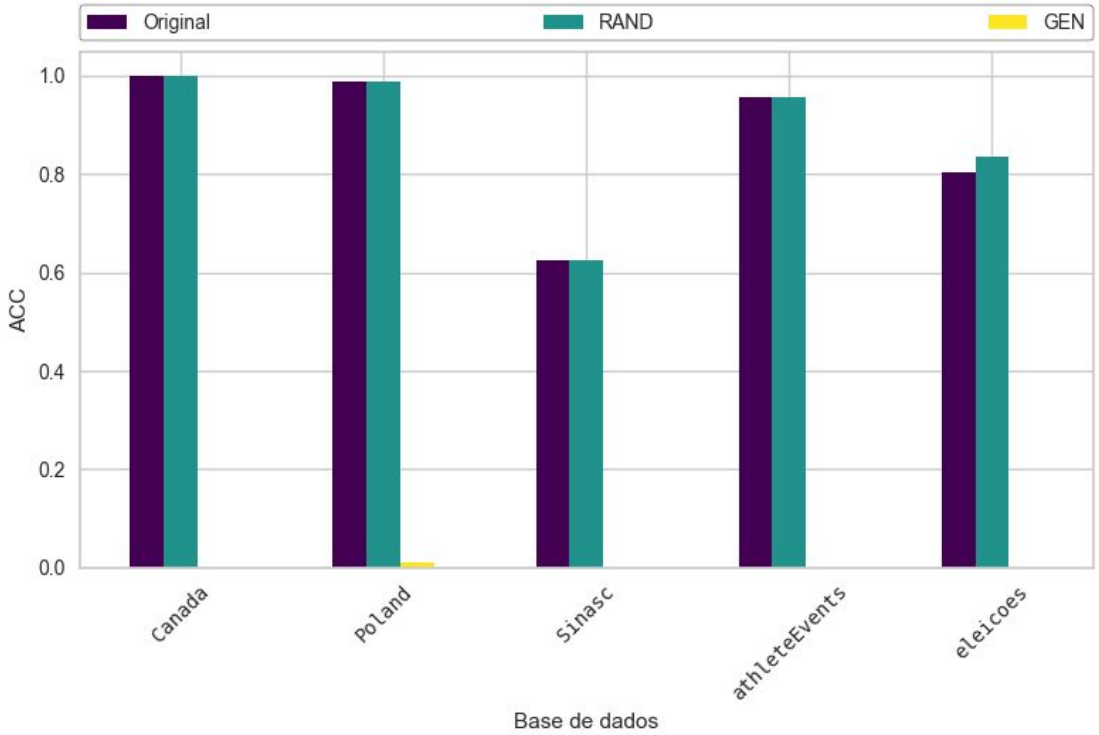
Queda de para a técnica de supressão, seguida pela queda para generalização

muitas variáveis anuladas



Completude

Para supressão e pseudoanonimização sequer foi possível metrificar



Acurácia

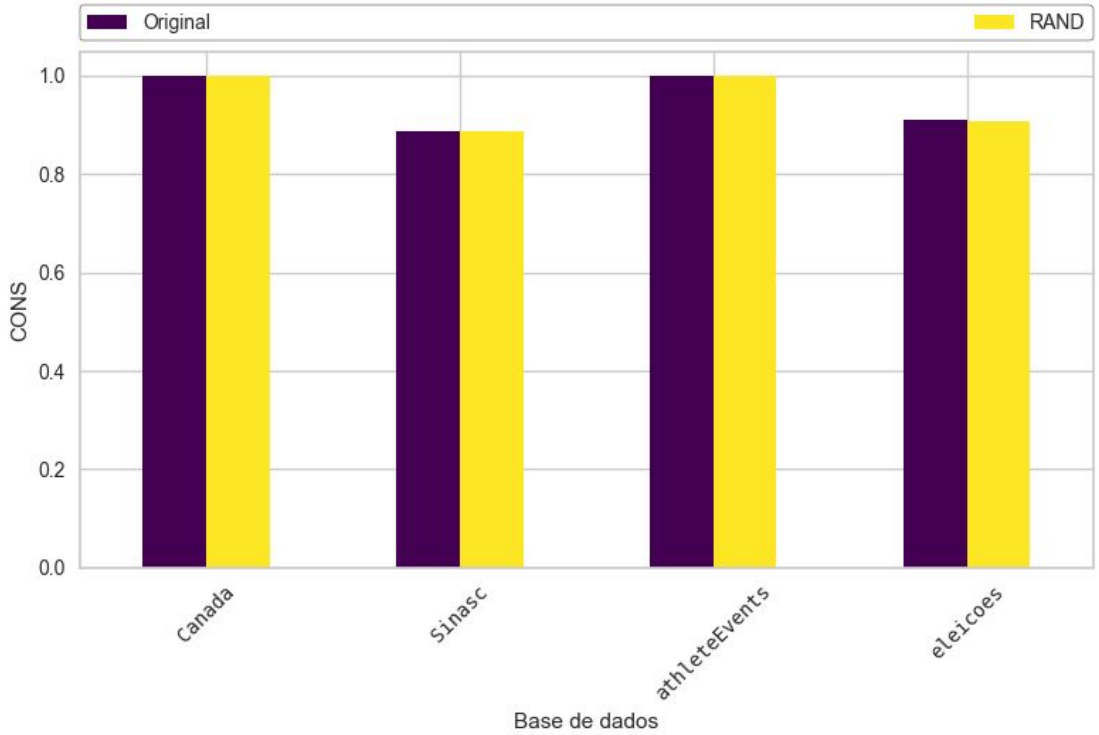
Tão grande é a alteração realizada que nem sequer é possível realizar alguma verificação

dt_nasc
06-05-1998
26-07-2022

Pseudoanonimização

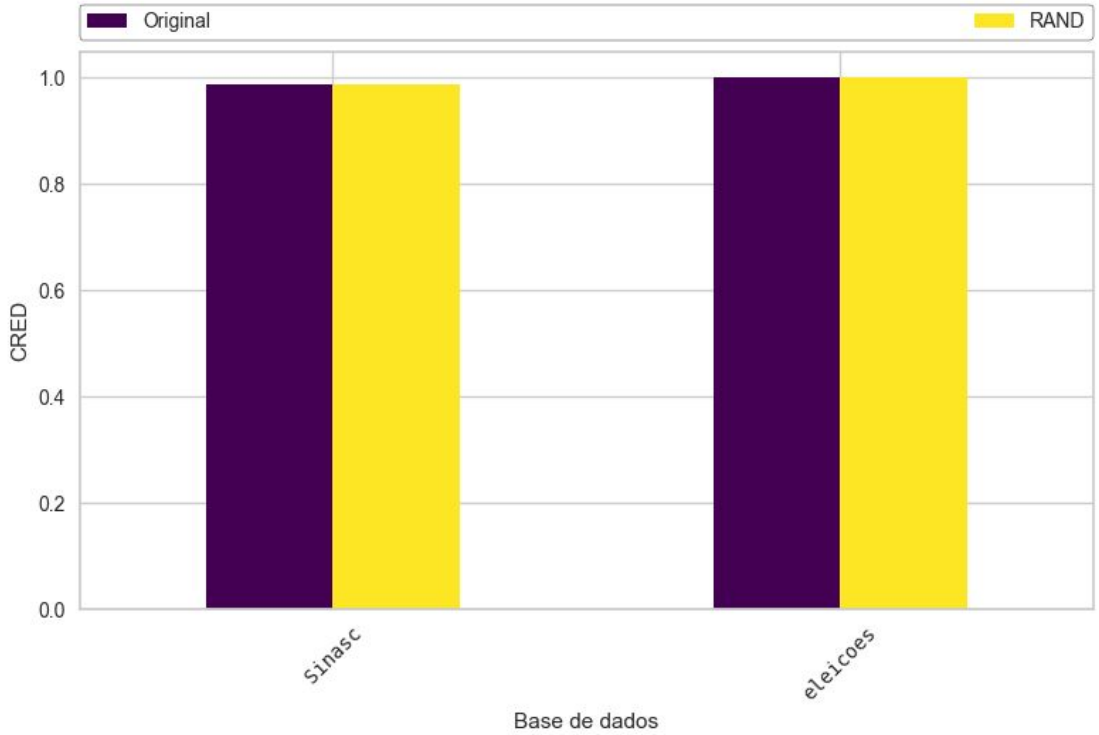
dt_nasc
7ec256f8508137fb7f8c...
287e22c2693d132e7d9...

Para supressão e pseudoanonimização sequer foi possível metrificar



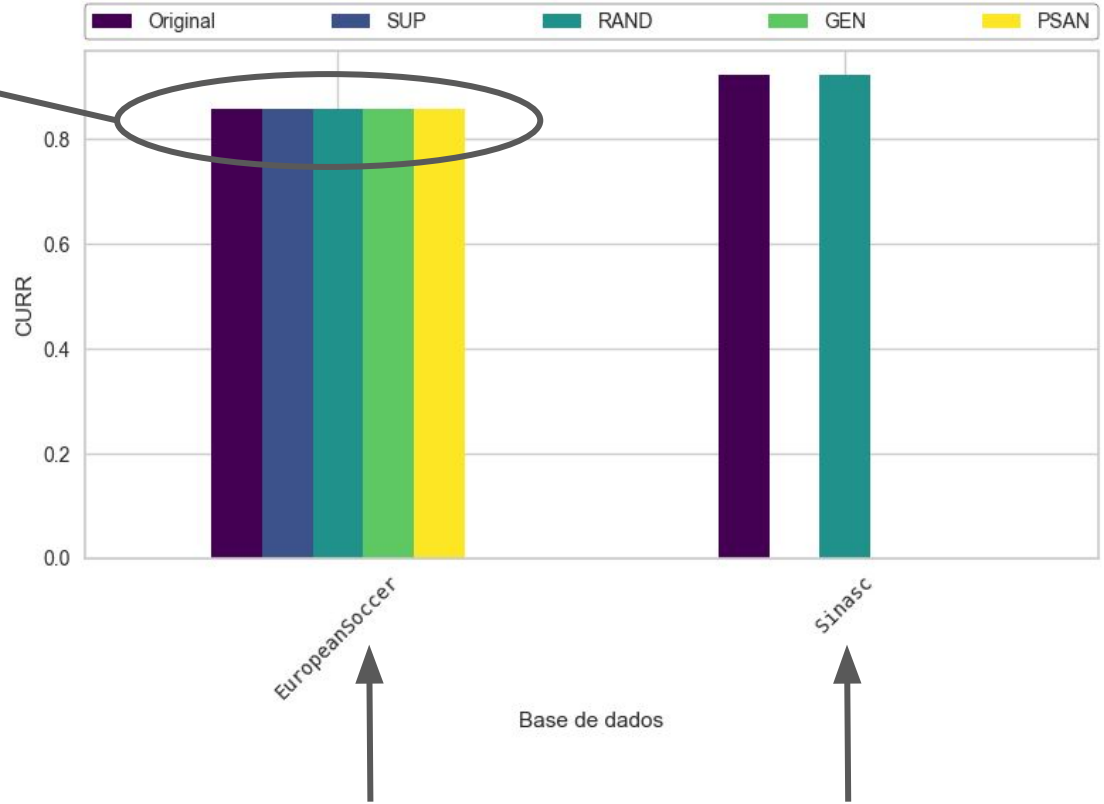
Consistência

Para supressão e pseudoanonimização sequer foi possível metrificar



Credibilidade

Não compreendem dados sensíveis

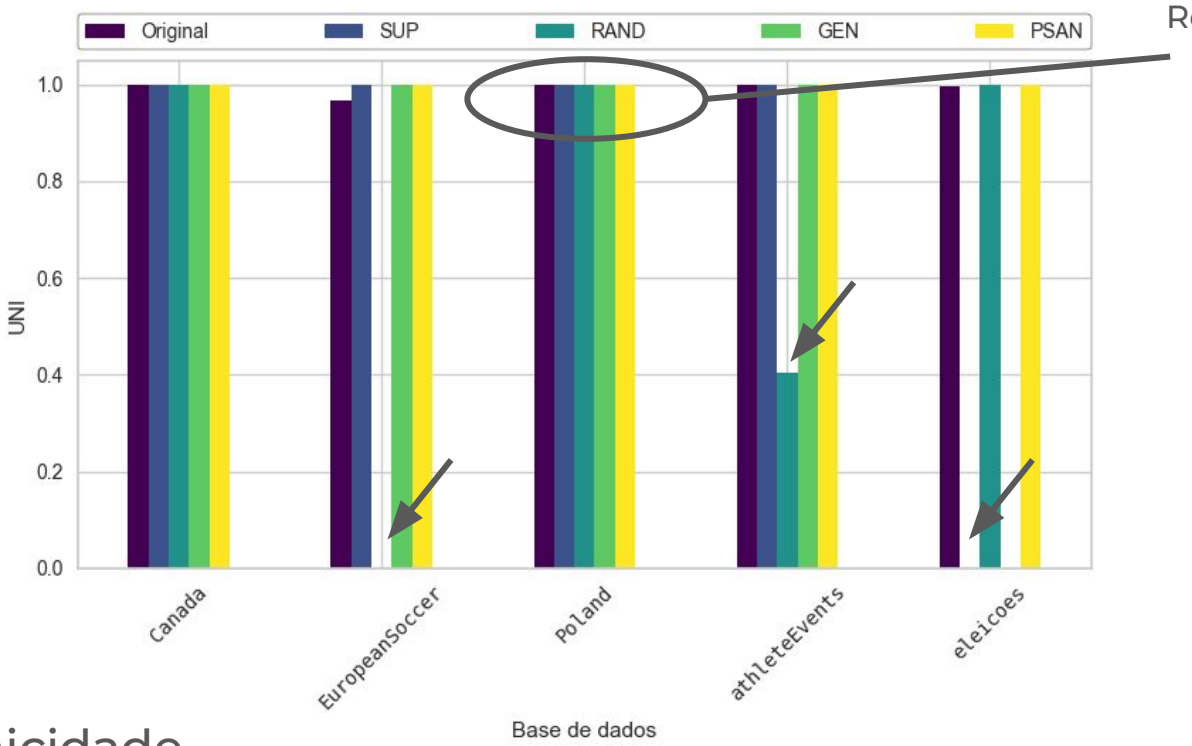


Atualidade

informações sem lacunas entre anos

informação submetida em no máximo um ano

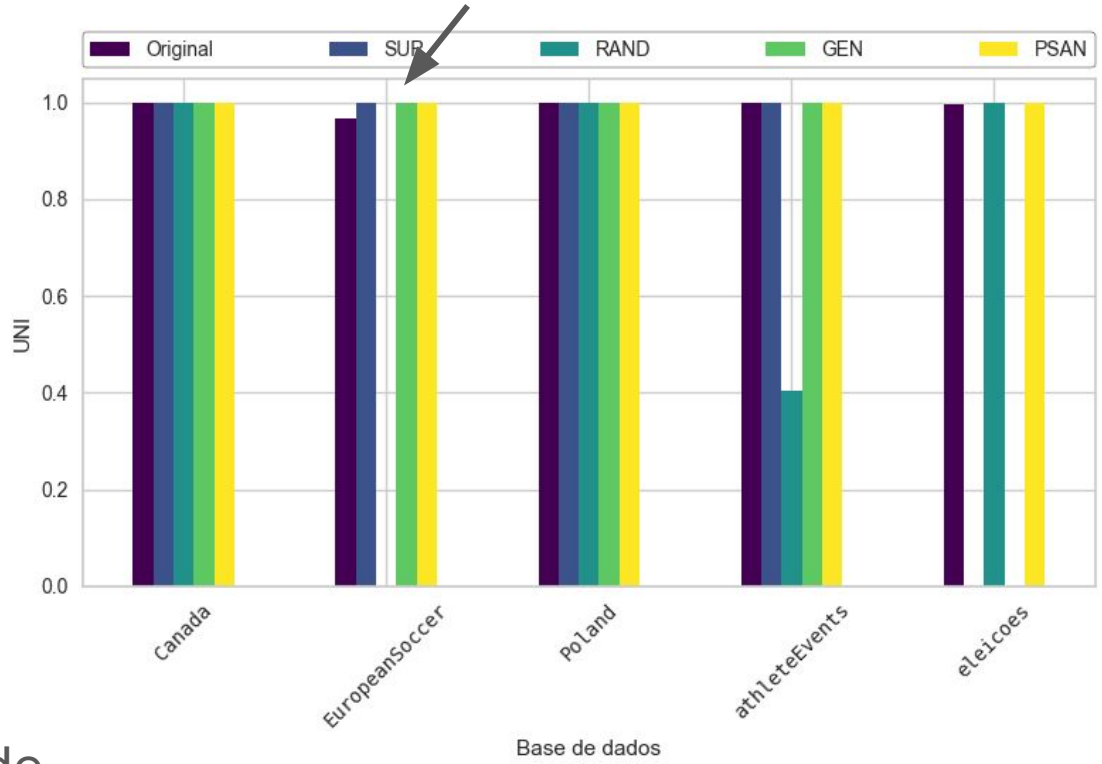
Randomização é a mais favorável a apresentar erros, ao embaralhar a ordem das informações



Registros não se repetem

Unicidade

Supressão, generalização e pseudoanonimização tornam as variáveis mais amplas, e menos susceptíveis a erros



Unicidade

As variáveis tornam-se mais amplas, e menos susceptíveis a falhas

cod_mun_nasc
3205309
3204351

Generalização

cod_mun_nasc
32
32

- **Supressão**: a ação de anular variáveis torna-as incompletas;
- **Generalização**: a aplicação da técnica de forma completa iguala-se à supressão;
- **Randomização**: altera-se a pertinência da informação;
- **Pseudoanonimização**: deve-se observar se a sintaxe alterada não altera a semântica intrínseca.

CONCLUSÕES

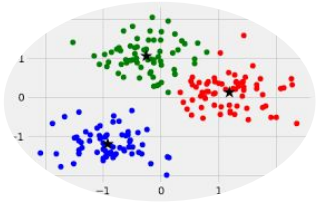
Conclusões

Os ensaios realizados demonstraram que, em maior ou menor grau, todas as maneiras de corresponder-se à legislação mostraram-se passíveis a falhas, sejam elas em relação à disposição da informação, ao preenchimento, sintáticas, semânticas e outras, fazendo com que a adequação deva ser conformada ao fim no qual o projeto de ciência de dados se dará.

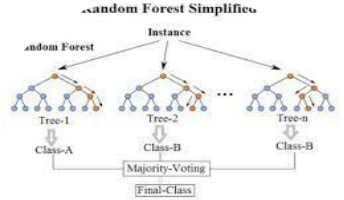
Trabalhos futuros

Variáveis com boa qualidade

Aplicação em modelos



KMeans



Random Forest

Obrigado!