Data bias measurement: a geometrical approach through frames

Andrea Trenta
UNINFO UNI CT 533
Technical Committee Artificial Intelligence
Italy
andrea.trenta@dataqualitylab.it

***Abstract*— In previous papers [8], [9] we discussed ISO/IEC 25000 application when new quality measures are defined. In continuity with papers above we show, through the definition of new data quality measures for bias, how to handle additional and new measures in a SQuaRE perspective. The method proposed is intended applicable in general.**
In the present paper:
- data bias is identified as a quality issue
- some notions about frames theory are recalled and
- two quality measures for data bias are proposed and
- one of them is proposed as ISO/IEC 25024 conforming measure

Keywords: data quality, measures, eigenvalue, bias, fairness, ISO, ISO/IEC 25024, frame, metric, AI, ML, PCA

I. INTRODUCTION

A well-known problem in ML is the bias-variance dilemma: to find the optimal complexity of the model that minimize output errors while giving independency from changes of training dataset (i.e. balancing underfitting and overfitting) [19]:

$$E_D = B^2 + V =$$

$$= E_D^2[y(x;D) - h(x)] + E_D[\{y(x;D) - E_D\,[y(x;D)]\}^2]$$

where:

E_D is the expected squared output error

B is the bias

V is the variance

x is the input vector

$h(x)$ is the regression function characterizing the model

y(x;D) is the prediction function of x over the dataset D

So, the expected squared error E_D is due to the squared error generated by the regression function adopted (bias) and also due to the behavior of the prediction function around its average for the dataset D (variance), in other words the sensitivity to the variation of dataset.

The bias-variance decomposition is of low practical value, because it requires to know all the datasets the machine will handle, whereas in practice we have only a single observed dataset and we need to predict\train the behavior of the machine at its best. Moreover, the U-shaped error function of the bias-variance optimization doesn't hold for deep neural networks [29].

For those reasons, in the following we don't refer to the bias-variance dilemma, but simply refer to data bias as statistical features of a dataset in a ML context[i]. Such statistical features can be measured by several indexes (e.g. Gini, Shannon, see [16], [17], [18]) and the new ones that we are going to introduce in this paper.

Moreover, this paper tries to recall a wider issue: how to address the manifold of measures that are continuously discovered, including, but not only, AI measures: in our view they can be all addressed under the ISO/IEC 25000 umbrella [8][9].

In the following we introduce an application and some considerations taken from frame theory and close to the Principal Component Analysis [19], [7]. Intuitively, PCA finds the (hyper)ellipsoid that best fit the dataset, by centering dataset in the origin, and the axis of the (hyper)ellipsoid are the eigenvectors of the covariance matrix. In a similar view, we reshape the (hyper)ellipsoid into an (hyper)sphere and translate the dataset over its surface to assess its spread with the help of frame theory.

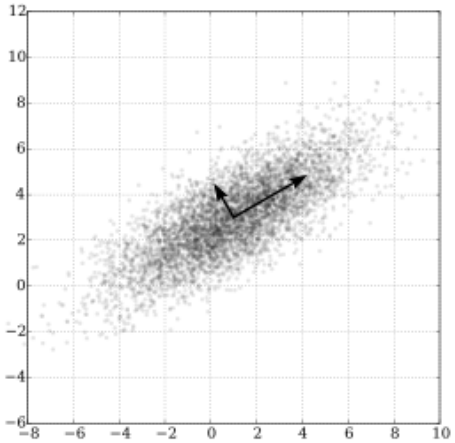


Figure 1 PCA of a multivariate gaussian distribution (source: Nicoguaro - wikimedia)

In our application, we consider each sample-point as the edge of a vector with the other edge in the origin and measure the overall span of such vectors.

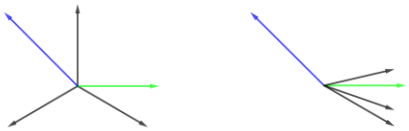
Firstly, we recall the definition of frame and frame bounds with an example taken from [10].

Definition 1: For a Hilbert space H_m of dimension m and with inner product $<\cdot,\cdot>_{H_m}$, a finite or countable collection of vectors $\varphi_{i\,(i\in I)} \subset H_m$ is said to be a frame of H_m if there exist constants $0< c \leq C$ such that

$$c \parallel \varphi \parallel^2 \leq \sum_{i\in I} |< \varphi_i , \varphi >|^2 \leq C \parallel \varphi \parallel^2 \qquad (1)$$

for all $\varphi \in H_m$

A frame is said to be tight if $c = C$



(A) (B)

Figure 2 Examples of frame spread

The φ_i vectors are the black ones in (A) and (B) and are frames of R^2. Blue and green vectors are instances of φ.

With reference to figure 2 above, the black vectors are our frame and it is easy to realize that in (A) they are spreader in the space than in (B).

It is also understandable at a glance that in (B) the green vector maximizes, as it forms narrow angles, the sum of dot product of each black vectors with the green one, leading to find C; and the blue vector minimizes, as it forms wide angles, the sum of dot product of each black vectors with the blue one, so leading to find c.

In the same way, it is easy to check that in (A) the sum of the dot products of the green vector with the black ones, has the same value of the sum[1] of the dot product of the blue vector with the black ones, leading to $c = C$, so the frame in (A) is tight.

Fortunately, it is possible to calculate tightness, that is the difference between C and c, not by φ trial and error but by the covariance matrix generated with mutual dot product among vectors, as C and c are respectively its maximum and the minimum eigenvalue.

II. DATASET AND THE FRAME MODEL

The first proposed bias measure is based on the following theorems and definitions [14]:

-when the frame bounds c and C are equal, a frame is said to be tight.

Defined $\Phi = \begin{pmatrix} \varphi_1 \\ \varphi_2 \\ \cdot\cdot \\ \varphi_N \end{pmatrix}$ the matrix NxM that collects vectors φ_i of the frame, then:

-the upper and lower frame bounds (see C and c above) of a frame are given by the largest and smallest eigen values of the frame operator $S = \Phi\,\Phi^T$ respectively.

-the non-zero eigen values of the frame operator S are the same of the non-zero eigen values of the Gram matrix $G = \Phi^T\Phi$.

-the rank of G is M. The M eigenvalues of G are positive.

In this proposal we:

(a) handle a numeric dataset as it was a frame: for a set of N tuples over a set of M attributes, then in the frame view the number M of attributes is the space dimension and the number N of tuples is the number of vectors of the frame;

(b) then, we measure data bias in the same way we measure frame tightness, and in particular:

(b.1) measure difference between upper and lower bounds of the frame

(b.2) measure Frame Potential

From those assumptions follows that a non-biased dataset is found when the corresponding frame is tight.

The measures b.1 and b.2 can be considered equivalent for the purpose of evaluating bias of dataset; in this case even only one of them can be adopted, and the choose between the two can be driven by the computational effort required. In this paper we explore mainly the measure b.1.

Measure b.1

The basis of our analysis is the calculation of lower and upper frame bounds with the following steps:

1. Collect a numeric data table with N tuple and M attributes and define it as a set of N row vectors $\{ \Phi_1, \Phi_2, \ldots \Phi_N \}$;

2. Build the matrix (NxM) $\Phi = \begin{pmatrix} \Phi 1 \\ \Phi 2 \\ \cdot\cdot \\ \Phi N \end{pmatrix}$

then

3. Compute the Gramian matrix (MxM) $G = \Phi^T\Phi$

4. Compute M eigenvalues λ_i (i=1,..M) of G

5. Sort the (non-zero) eigenvalues of G in descending order

6. Find the upper eigenvalue λ_{max} and the lower eigenvalue λ_{min}

[1] We mean the $\parallel \varphi \parallel^2$ normalized squared sum according (1)

7. Compute the difference D = λ_{max} - λ_{min}

8. Assess the value of D considering that D = 0 means a tight frame.

Measure b.2

The second proposed bias measure is based on the following theorems and definitions.

The frame potential FP is defined as [9]:

FP = $\sum_{i,j\epsilon I} |\langle \varphi_i, \varphi_j \rangle|^2$

where φ_i are the frame vectors.

In this measure, the step 4 above and further ones are replaced by the following

4. Compute the FP from matrix G = $\Phi^T\Phi$

5. Assess the value of FP considering that a minimum value of FP, is reached when the frame is tight.

For computing step 4, consider that $\langle \varphi_i, \varphi_j \rangle$ with $i, j\epsilon I$ are the diagonal and upper -right elements (or lower – left as G is symmetric) of the Gramian matrix; FP in other words is the sum of the squared upper (or lower) elements of the Gramian, including diagonal ones; this measure may be easier to compute than the previous one.

A first example for measure b.1 and b.2

Consider a dataset with attributes "Age" and "Income"; domains are 6 age groups [20-30), [30-40), [40-50), [50-60), [60-70), [70-80) and 7 income categories [10-20K€), [20-30k€), [30-40k€), [40-50k€), [50-60k€), [60-70k€), [70-80k€); here three samples of (Age, Income) are collected:

Age Income

$$\Phi = \begin{pmatrix} 1 & 4 \\ 3 & 1 \\ 6 & 7 \end{pmatrix} = \begin{pmatrix} t \\ u \\ v \end{pmatrix} \qquad G = \begin{pmatrix} 46 & 49 \\ 49 & 66 \end{pmatrix}$$

From G we calculate measures b.1:

D = λ_{max} - λ_{min} = 106 – 6 = 100

and measure b.2:

FP = $46^2+66^2+49^2$= 8873

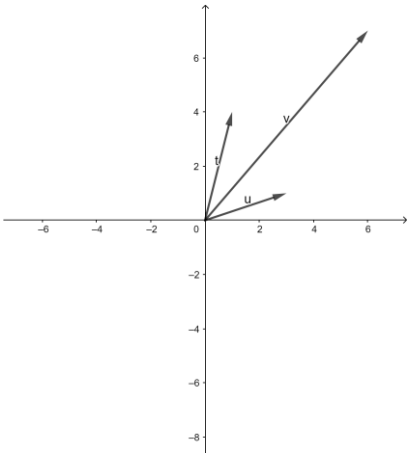


Figure 3 Frame vectors t,u,v are the rows of matrix Φ

The measure b.1 is responsive to tuples order (e.g. swapping tuples in general leads to different measure values) and so it gives a measure of tightness of ordered tuples, where tightness is defined according (1). As we want a measure not responsive to the tuples order, in the following we explain how to solve this issue.

III. A COMPARISON WITH PCA

To explain visually the approach, we compare PCA (figure 1) with our method:

(i) in PCA, if we find equal G eigenvalues $\lambda_1 = \lambda_2 \ldots = \lambda_M$, we can conclude that there isn't any dominant component and the volume fitting data is an (hyper)sphere;

(ii) similarly in our method, if we firstly project the data over an (hyper)sphere surface, and if we then find equal G' eigenvalues $\lambda'_1 = \lambda'_2 \ldots = \lambda'_M$, we can conclude that the projected data are evenly spread over the (hyper)sphere surface because they are the edges of an equiangular tight frame[2].

Possibly, to gain more information about data bias, both the approaches (i) and (ii) can be adopted. In this paper we consider only the approach (ii).

IV. APPLICATION REMARKS

According the method (ii), before applying steps 1-8, we apply the following 0.a, 0.b, 0.c steps:

0.a discretize vectors coordinates domains
0.b vertex mean translation so that the barycenter (average of the translated vertices) is zero.
0.c normalize vectors module to unit

As an example, we apply the measure b.1 over the dataset of covid-19 vaccinated people in Italy (https://github.com/italia/covid19-opendata-vaccini).

The dataset contains the number of COVID-19 vaccine injections grouped by 9 age range [16-20), [20-30), [30-

[2]as for equiangular tight frames holds G=N\M*I, where I is the identity matrix [25].

40), [40-50), [50-60), [60-70), [70-80), [80-90), [90-further).

As elder people were firstly vaccinated (generally 2 injections required for vaccination), the histogram of injections people shows the desired polarization in the higher age groups.

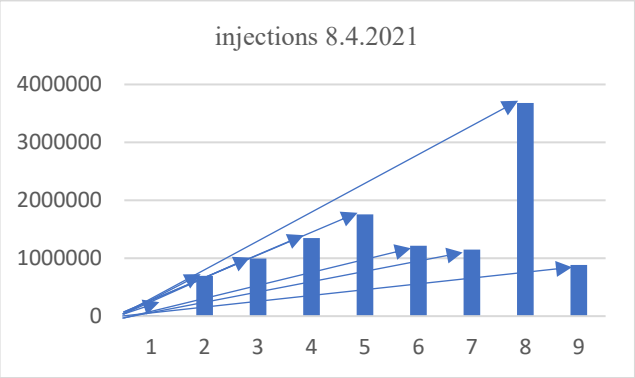


Figure 4 Injections per group age at 8.4.2021

Starting from elder people, vaccination was progressively extended to mid-age people, so about one and half month later it is found a different shaped histogram in figure 5.

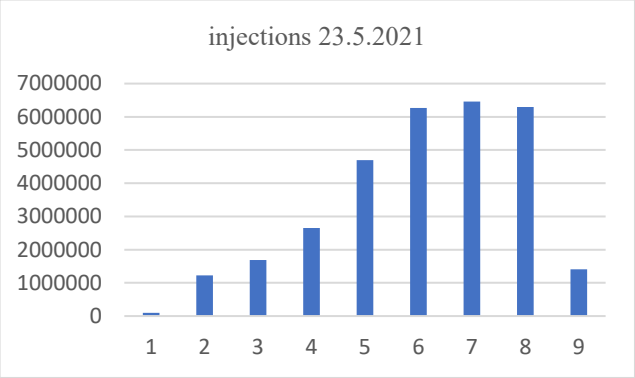


Figure 5 Injections per group age at 23.5.2021

As we said before, measure b.1 is not understandable if directly applied to the original dataset: as shown in figure 4, it leads to a sort of evaluation of the shape of the histogram, so we instead apply step 0.a dividing the domain of #vaccine_injections in 9 intervals and then apply step 0.b and 0.c. After normalization, we process $\Phi_{norm8.4.2021}$ and $\Phi_{norm23.5.2021}$ and we have respectively the results (figure 6):

$\lambda_{norm8.4.2021\_1}$=2,92,

$\lambda_{norm8.4.2021\_2}$=6,10 $D_{norm8.4.2021}$=3,18

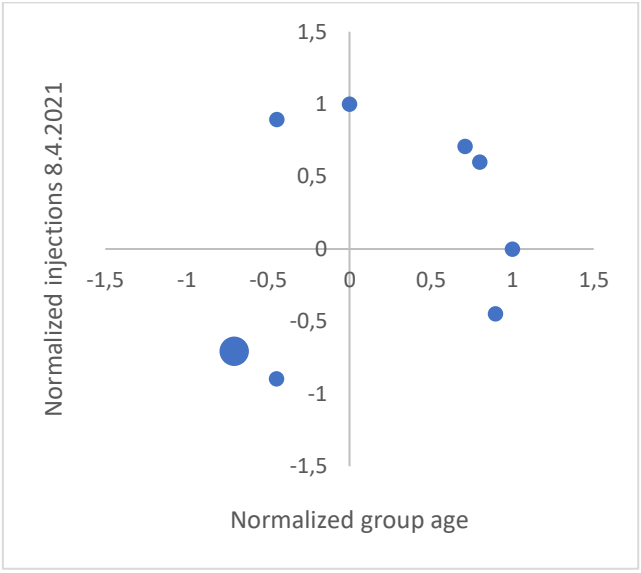


Figure 6 Injections (3-steps normalization) at 8.4.2021

and (figure 7):

$\lambda_{norm23.5.2021\_1}$=2,2

$\lambda_{norm23.5.2021\_2}$=6,84 $D_{norm23.5.2021}$=4,64

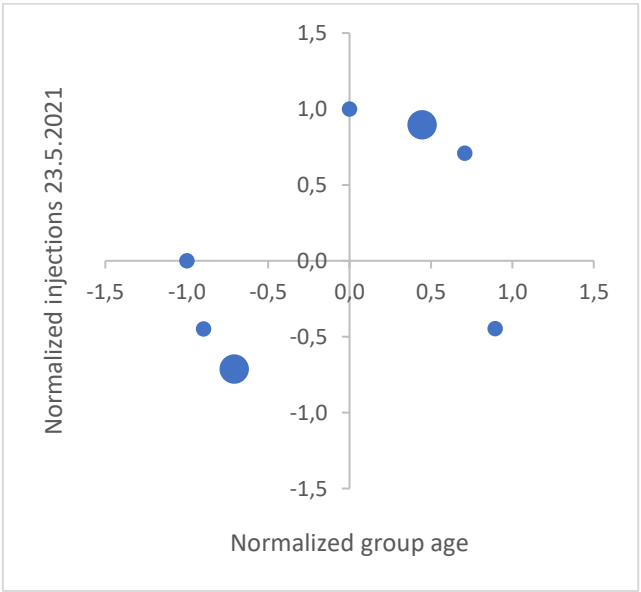


Figure 7 Injections (3-steps normalization) at 23.5.2021

Note: the bigger circles stand for overlapping points

Note that, applying the normalization steps to a generic dataset, we have the maximum D_{norm}=8 for a uniform distribution, so the 23.5.2021 value it is closer to a uniform distribution than the 8.4.2021 value and that is the behavior expected[3], because the aim was to have, later in time, high values of vaccinated people in all group ages (i.e. less

[3] Note that is also fulfilled [24] the normalized frames condition N= $\sum_{i=1}^{M} \lambda_i$

"tight"). Visually, in figure 7 some points got closer than in figure 6 and two more collapsed; and this is what we expected, as moving towards a uniform distribution means that even more points get closer or collapse.

It is interesting transpose, for the purpose of assessing bias, some results from frame theory, for example:

-care should be taken in choosing M and N, because some couples (M, N) don't correspond to Equiangular Tight Frames[4] [25] and\or don't correspond to "highly symmetric frames"[5] [26].

From a bias point of view, this could mean that for some (M, N) couples it's easier to build non-biased data.

Moreover, the minimum value for Frame Potential FP for unit-norm tight frames, [9] is:

$FP = N$ if $M \leq N$

$FP = M^2\backslash N$ if $M \geq N$

and this helps to assess the data that have an FP close to the minimum, as it means they are "as orthogonal" to each other as possible.

V. PROPOSAL

To sum up, with this proposal we address the issue of finding a data quality measurement function (i.e. metric) through geometrical calculation.

Its application is envisaged for, but not for only, evaluation of sampling bias across multiple attributes, as for example the protected ones [12]: a well-known issue in modern societies are the inequalities and with the measure above we can overall assess the bias of a population dataset over multiple attributes like "income", "ethnicity", "group age", instead of assessing bias against single or couples of attributes.

At the same time, we highlight the need to handle the manifold of measures that are discovered by the community of researchers with the approach explained in [8]: the new measure b.1 "tightness" can be defined in terms of a new measure conforming to [6] and\or to [28].
The measure b.1 is documented in SQuaRE format in Table 1. For the time being, we make the assumption that "tightness" is relevant to completeness characteristic, further refinements about characteristic relevance will depend on the work progress in [28]. For the scope of this paper, Table 1 shows a simplified measure documentation; a comprehensive way of measure documentation is described in [3].

ID	Com-I-4-IT-10
Name	Data values completeness
Description	Tightness of normalized dataset
Measurement function	$X = A-B = \lambda_{max} - \lambda_{min}$ $\lambda_{max}, \lambda_{min}$ are max, min eigenvalue of $G = \Phi_{norm}{}^T \Phi_{norm}$ matrix Φ_{norm} is built from dataset normalized according steps 0.a, 0.b, 0.c
DLC	All Data Life Cycle
Target Entity	Dataset with N tuples and M attributes
Property	Data value
NOTE X=0 means "tight" according the definition of frame theory	
NOTE ID includes additional part IT-10 [3]	

Table 1 Completeness measure - Tightness

VI. FURTHER STUDIES

Whereas data bias measurement starting from a given dataset appears relatively easy, designing a dataset (frame) starting from a level of bias (tightness), is not so simple [13]; this result, as far as possibly others from frame theory, can be taken into consideration when looking for an optimal training dataset for Machine Learning.

Dataset spread measurement appears useful in conjunction with classification[6] and it holds also for not pre-trained machine like SVM (Support Vector Machines).

Due to the use of ML models in many fields (see Perceptron in mechanical statistic [31]), we cannot exclude other fields of application for this early study.

The metric is applicable with some tricks [7] also to images and it will be analyzed in a future paper.

VII. CONCLUSION

In this early study the measures b.1 and b.2 appear suitable to measure data sample bias [12], that in turns is mainly related to accuracy and completeness data quality model characteristic [6], [28]. They can be considered in SC7 WG6 and appear relevant to SC42 WG2 and WG3 work in progress on A.I. [27], [28].

The manifold of metrics available for industry and research[7], including the one introduced in this paper, can be addressed in the ISO/IEC 25000 perspective: applying the process described in [8], all the measures, including b.1 and b.2 presented in this paper, can be defined as ISO/IEC 25000 conforming measures.

[4] From [25]: ∄ RETF (19, 76) and ∄ RETF (20, 96); notation means "there not exists a real equiangular tight frame with parameters (M, N)"

[5] E.g.: there are no "highly symmetric" tight frames of five vectors in C^3, but there are tight frames of five vectors like vertexes of a trigonal bipyramid

[6] In general classification is easier when data are not spread.

[7] See an example of the manifold of benchmarks in https://paperswithcode.com/sota and [30].

VIII. ACKNOWLEDGEMENTS

The author would like to thank Antonio Vetrò, who encouraged this work, Alessandro Simonetta, who suggested some enhancements, Domenico Natale, for his foundational contributions in the field of data quality, and Roberto Li Voti, who believed in this project.

REFERENCES

[1] ISO/IEC 25010:2011 Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models

[2] ISO/IEC 25012:2008 Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Data quality model

[3] ISO/IEC 25020:2019, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality measurement framework.

[4] ISO/IEC 25022:2016, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of quality in use.

[5] ISO/IEC 25023:2016, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of system and software product quality.

[6] ISO/IEC 25024:2015, Systems and Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Measurement of data quality

[7] A. Trenta: ISO/IEC 25000 quality measures for A.I.: a geometrical approach Proceedings APSEC IWESQ 2020 (http://ceur-ws.org/Vol-2800/, ISSN 1613-0073)

[8] D. Natale, A. Trenta: Examples of practical use of ISO/IEC 25000 Proceedings APSEC IWESQ 2019 (http://ceur-ws.org/Vol-2545/, ISSN 1613-0073)

[9] Jelena Kovacevic, Amina Chebira: An Introduction to Frames. Found. Trends Signal Process. 2(1): 1-94 (2008)

[10] Mishal Assif P. K., Mohammed Rayyan Sheriff, Debasish Chatterjee: Measure of quality of finite-dimensional linear systems: A frame-theoretic view. CoRR abs/1902.04548 (2019)

[11] John J Benedetto and Joseph D Kolesar. Geometric properties of Grassmannian frames for R^2 and R^3 EURASIP Journal on Advances in Signal Processing, 2006(1): 049850, 2006.

[12] ISO/IEC 24027 draft Information technology - Artificial Intelligence (AI) – Bias in AI systems and AI-aided decision making

[13] https://www3.math.tu-berlin.de/numerik/www.fusionframe.org/index_application.html

[14] Shailesh Kumar - Results on equiangular tight frames https://www.slideshare.net/

[15] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems (pp. 6626–6637)

[16] Mecati, M., Cannavò, F. E., Vetrò, A., & Torchiano, M. (2020, August). Identifying Risks in Datasets for Automated Decision–Making. In International Conference on Electronic Government (pp. 332-344). Springer, Cham. https://link.springer.com/chapter/10.1007%2F978-3-030-57599-1_25

[17] Beretta E., Vetrò A., Lepri B., De Martin J. C. D. (2021) Detecting discriminatory risk through data annotation based on Bayesian inferences. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 794-804). https://doi.org/10.1145/3442188.3445940

[18] Beretta E., Vetrò A., Lepri B., De Martin J. C. (2018) Ethical and socially-aware data labels. In Annual International Symposium on Information Management and Big Data (pp. 320-327). Springer, Cham. https://link.springer.com/chapter/10.1007%2F978-3-030-11680-4_30

[19] Bishop C., (2006) Pattern recognition and machine learning, chapter 7 Sparse Kernel Machines ISBN-13: 978-0387-31073-2

[20] Kailash Ahirwar (2019) Generative Adversarial Networks Projects ISBN-13: 978-1789136678

[21] Borji (2018) Pros and Cons of GAN Evaluation Measures https://arxiv.org/pdf/1802.03446.pdf

[22] ISO/IEC 22989 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (2021 draft)

[23] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero, Samira Shabanian, Sina Honari (2021) Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics https://openreview.net/forum?id=OTnqQUEwPKu

[24] Ole Christensen (2016) An Introduction to Frames and Riesz Bases 2nd edition ISBN: 978-3-319-25613-9

[25] Matthew Fickus, Dustin G. Mixon (2016) Tables of the existence of equiangular tight frames https://arxiv.org/abs/1504.00253

[26] Helen Broome, Shayne Waldron (2013) On the construction of highly symmetric tight frames and complex polytopes http://dx.doi.org/10.1016/j.laa.2013.10.003

[27] ISO/IEC 25059 (draft) Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality Model for AI-based systems

[28] ISO/IEC 5259-2 (draft) Artificial intelligence — Data quality for analytics and ML — Part 2: Data quality measures

[29] Double Trouble in Double Descent: Bias and Variance(s) in the Lazy Regime (2020) Stephane d'Ascoli, Maria Refinetti, Giulio Biroli, Florent Krzakala https://arxiv.org/pdf/2003.01054.pdf

[30] ISO/IEC DIS 23053 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

[31] The simplest model of jamming (2015) Silvio Franz, Giorgio Parisi https://arxiv.org/pdf/1501.03397.pdf

[i] Some definitions:
Data bias: "data properties that if unaddressed lead to AI systems that perform better or worse for different groups"
Bias: "systematic difference in treatment of certain objects, people, or groups in comparison to others" and "Treatment is any kind of action, including perception, observation, representation, prediction, or decision. [12, 22]
Data quality checking "process in which data is examined for completeness, bias and other factors which affect its usefulness for an AI system" [22]