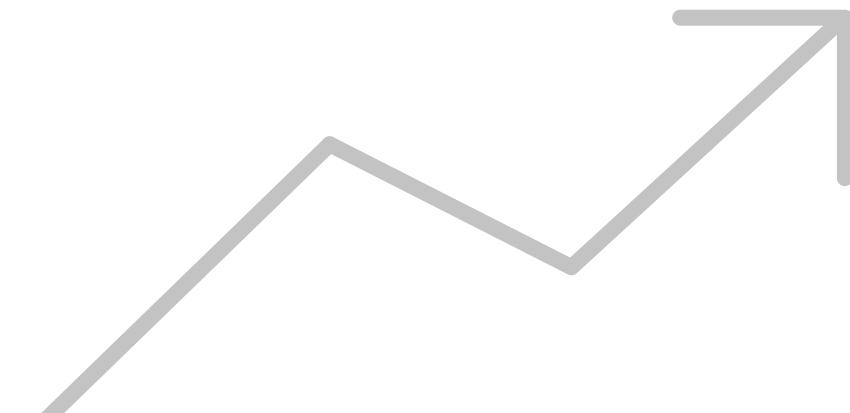


# Quality Dimensions of Machine Learning in Official Statistics

Florian Dumpert, Younes Saidani

Federal Statistical Office of Germany

Joint work with: Christian Borgs, Alexander Brand, Andreas Nickl, Alexandra Rittmann,  
Johannes Rohde, Christian Salwiczek, Nina Storfinger, Selina Straub



# Official Statistics

## It is not about ...

60 3. Uniformly Most Powerful Tests

yields the largest number of miles per hour. Analogously in the present problem the most valuable points  $x$  are those with the highest value of

$$r(x) = \frac{P_1(x)}{P_0(x)}$$

The points are therefore rated according to the value of this ratio and selected for  $S$  in this order, as many as one can afford under restriction (3.6). Formally this means that  $S$  is the set of all points  $x$  for which  $r(x) > c$ , where  $c$  is determined by the condition

$$P_0\{X \in S\} = \sum_{x: r(x) > c} P_0(x) = \alpha.$$

Here a difficulty is seen to arise. It may happen that when a certain point is included, the value  $\alpha$  has not yet been reached but that it would be exceeded if the point were also included. The exact value  $\alpha$  can then either not be achieved by at all, or it can be attained only by breaking the preference order established by  $r(x)$ . The resulting optimization problem has no explicit solution. (Algorithms for obtaining the maximizing set  $S$  are given by the theory of linear programming.) The difficulty can be avoided, however, by a modification which does not require violation of the  $r$ -order and which does lead to a simple explicit solution, namely by permitting randomization.<sup>2</sup> This makes it possible to split the next point, including only a portion of it, and thereby to obtain the exact value  $\alpha$  without breaking the order of preference that has been established for inclusion of the various sample points. These considerations are formalized in the following theorem, the fundamental lemma of Neyman and Pearson.

Theorem 3.2.1 Let  $P_0$  and  $P_1$  be distributions possessing densities  $p_0$  and  $p_1$  respectively with respect to a measure  $\mu$ .<sup>3</sup>

(i) Evidence. For testing  $H : p_0$  against the alternative  $K : p_1$ , there exists a test  $\phi$  and a constant  $k$  such that

$$E_0\phi(X) = \alpha \quad (3.7)$$

and

$$\phi(x) = \begin{cases} 1 & \text{when } p_1(x) > kp_0(x), \\ 0 & \text{when } p_1(x) < kp_0(x). \end{cases} \quad (3.8)$$

(ii) Sufficient condition for a most powerful test. If a test satisfies (3.7) and (3.8) for some  $k$ , then it is most powerful for testing  $p_0$  against  $p_1$  at level  $\alpha$ .

(iii) Necessary condition for a most powerful test. If  $\phi$  is most powerful at level  $\alpha$  for testing  $p_0$  against  $p_1$ , then for some  $k$  it satisfies (3.8) a.e.  $\mu$ . It also satisfies (3.7) unless there exists a test of size  $< \alpha$  and with power 1.

PROOF. For  $\alpha = 0$  and  $\alpha = 1$  the theorem is easily seen to be true provided the value  $k = +\infty$  is admitted in (3.8) and  $0 \cdot \infty$  is interpreted as 0. Throughout the proof we shall therefore assume  $0 < \alpha < 1$ .

<sup>2</sup>In practice, typically neither the breaking of the  $r$ -order nor randomization is considered acceptable. The common solution, instead, is to adopt a value of  $\alpha$  that can be attained exactly and therefore does not present this problem.

<sup>3</sup>There is no loss of generality in this assumption, since one can take  $\mu = P_0 + P_1$ .

(i): Let  $\alpha(c) = P_0\{p_1(X) > kp_0(X)\}$ . Since the probability is computed under  $P_0$ , the inequality need be considered only for the set where  $p_0(x) > 0$ , so that  $\alpha(c)$  is the probability that the random variable  $p_1(X)/p_0(X)$  exceeds  $c$ . Thus  $\alpha(c)$  is a cumulative distribution function, and  $\alpha(c)$  is nonincreasing and continuous on the right,  $\alpha(-\infty) - \alpha(c) = P_0\{p_1(X)/p_0(X) = c\}$ ,  $\alpha(-\infty) = 1$ , and  $\alpha(\infty) = 0$ . Given any  $0 < \alpha < 1$ , let  $c_0$  be such that  $\alpha(c_0) \leq \alpha \leq \alpha(c_0 - 0)$ , and consider the test  $\phi$  defined by

$$\phi(x) = \begin{cases} 1 & \text{when } p_1(x) > c_0 p_0(x), \\ \frac{\alpha - \alpha(c_0)}{\alpha(c_0 - 0) - \alpha(c_0)} & \text{when } p_1(x) = c_0 p_0(x), \\ 0 & \text{when } p_1(x) < c_0 p_0(x). \end{cases}$$

Here the middle expression is meaningful unless  $\alpha(c_0) = \alpha(c_0 - 0)$ ; since then  $E_0\{\phi(X) = c_0 p_0(X)\} = 0$ ,  $\phi$  is defined a.e. The size of  $\phi$  is

$$E_0\phi(X) = P_0\left\{\frac{p_1(X)}{p_0(X)} > c_0\right\} + \frac{\alpha - \alpha(c_0)}{\alpha(c_0 - 0) - \alpha(c_0)} P_0\left\{\frac{p_1(X)}{p_0(X)} = c_0\right\} = \alpha,$$

so that  $c_0$  can be taken as the  $k$  of the theorem.

(ii): Suppose that  $\phi$  is a test satisfying (3.7) and (3.8) and that  $\phi^*$  is any other test with  $E_0\phi^*(X) \leq \alpha$ . Denote by  $S^+$  and  $S^-$  the sets in the sample space where  $\phi(x) - \phi^*(x) > 0$  and  $< 0$  respectively. If  $x$  is in  $S^+$ ,  $\phi(x)$  must be  $> 0$  and  $\phi^*(x) = \phi^*(p_1 - kp_0) < 0$ . In the same way  $p_1(x) > kp_0(x)$  for all  $x$  in  $S^+$ , and hence  $p_1(x) > kp_0(x)$ . The difference in power between  $\phi$  and  $\phi^*$  therefore satisfies

$$\int (\phi - \phi^*)(p_1 - kp_0) d\mu = \int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0) d\mu \geq 0.$$

The difference in power between  $\phi$  and  $\phi^*$  is

$$\int (\phi - \phi^*)p_1 d\mu \geq k \int (\phi - \phi^*)p_0 d\mu \geq 0,$$

as was to be proved.

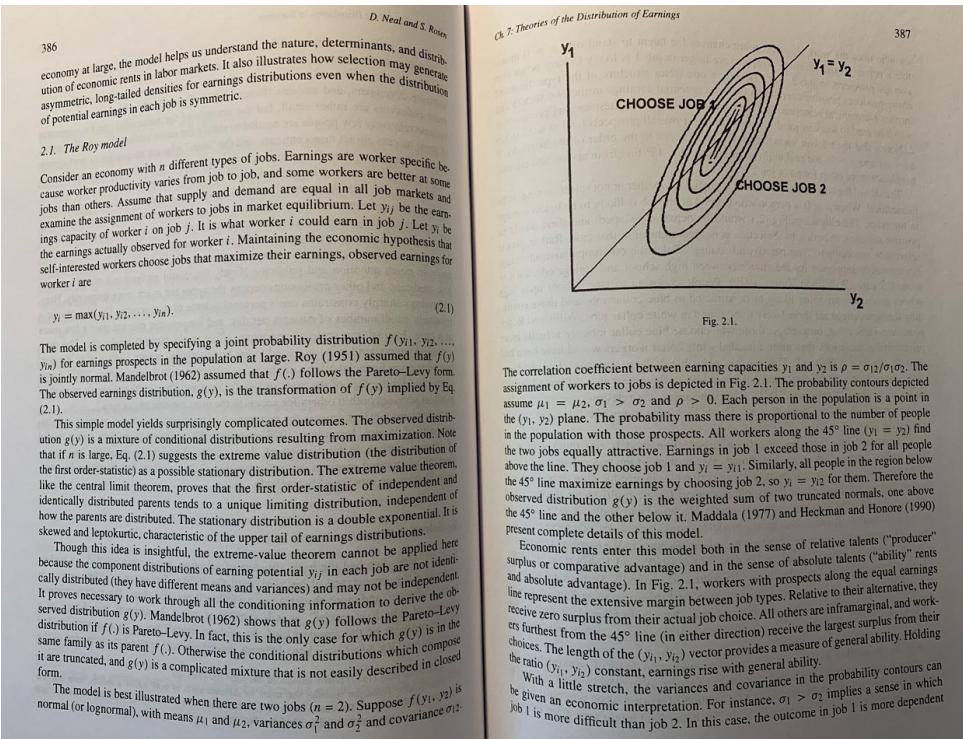
(iii): Let  $\phi^*$  be most powerful at level  $\alpha$  for testing  $p_0$  against  $p_1$ , and let  $\phi$  satisfy (3.7) and (3.8). Let  $S$  be the intersection of the set  $S^+ \cup S^-$ , on which  $\phi$  and  $\phi^*$  differ, with the set  $\{x : p_1(x) \neq kp_0(x)\}$ , and suppose that  $\mu(S) > 0$ . Since  $(\phi - \phi^*)(p_1 - kp_0)$  is positive on  $S$ , it follows from Problem 2.4 that

$$\int_{S^+ \cup S^-} (\phi - \phi^*)(p_1 - kp_0) d\mu = \int_S (\phi - \phi^*)(p_1 - kp_0) d\mu > 0$$

and therefore  $\mu(S) = 0$ , as was to be proved.

If  $\phi^*$  were of size  $< \alpha$  and power  $< 1$ , it would be possible to include in the rejection region additional points or portions of points and thereby to increase the power until the power is 1 or the size is  $\alpha$ . Thus either  $E_0\phi^*(X) = \infty$  or  $E_1\phi^*(X) = 1$ . ■

The proof of part (iii) shows that the most powerful test is uniquely determined by (3.7) and (3.8) except on the set on which  $p_1(x) = kp_0(x)$ . On this set,  $\phi$  can be defined arbitrarily provided the resulting test has size  $\alpha$ . Actually, we have shown that it is always to define  $\phi$  to be constant over this boundary set. In the trivial case that there exists a test of power 1, the constant  $k$  of (3.8) is 0, and one will accept  $H$  for all points for which  $p_1(x) = kp_0(x)$  even though the test may then have size  $< \alpha$ .



Lehmann EL, Romano JP (2005) Testing Statistical Hypotheses, 3rd edition, Springer  
Atkinson AB, Bourguignon F (2000) Handbook of Income Distribution Vol. 1, North Holland

# Official Statistics

## It is about ...

Die Statistik für Bundeszwecke (Bundesstatistik) hat im föderativ gegliederten Gesamtsystem der amtlichen Statistik die Aufgabe, **laufend Daten über Massenerscheinungen zu erheben, zu sammeln, aufzubereiten, darzustellen und zu analysieren**. Für sie gelten die Grundsätze der Neutralität, Objektivität und fachlichen Unabhängigkeit. Sie gewinnt die Daten unter Verwendung wissenschaftlicher Erkenntnisse und unter Einsatz der jeweils sachgerechten Methoden und Informationstechniken. Durch die Ergebnisse der Bundesstatistik werden gesellschaftliche, wirtschaftliche und ökologische Zusammenhänge für Bund, Länder einschließlich Gemeinden und Gemeindeverbände, Gesellschaft, Wirtschaft, Wissenschaft und Forschung aufgeschlüsselt. **Die Bundesstatistik ist Voraussetzung für eine am Sozialstaatsprinzip ausgerichtete Politik.** [...]

In the federally structured overall system of official statistics, statistics for federal purposes (federal statistics) have the task of **continuously collecting, collating, processing, presenting and analysing data on mass phenomena**. It is governed by the principles of neutrality, objectivity and professional independence. It collects data using scientific knowledge and appropriate methods and information technology. The results of federal statistics provide a breakdown of social, economic and ecological relationships for the Federation, the Länder including municipalities and municipal associations, society, the economy, science and research. **Federal statistics are a prerequisite for a policy oriented towards the welfare state principle.** [...]

§ 1 BStatG ([https://www.gesetze-im-internet.de/bstatg\\_1987/\\_1.html](https://www.gesetze-im-internet.de/bstatg_1987/_1.html)) and an unauthorised translation by the author of the slides

# Official Statistics

## How we work: The Generic Statistical Business Process Model (GSBPM)

| Specify needs                        | Design                                     | Build   | Collect                            | Process                            | Analyse                           | Disseminate                                  | Evaluate                     |
|--------------------------------------|--|---|------------------------------------|------------------------------------|-----------------------------------|--|------------------------------|
| 1.1 Identify needs                   | 2.1 Design outputs                         | 3.1 Reuse or build collection instruments             | 4.1 Create frame and select sample | 5.1 Integrate data                 | 6.1 Prepare draft outputs         | 7.1 Update output systems                    | 8.1 Gather evaluation inputs |
| 1.2 Consult and confirm needs        | 2.2 Design variable descriptions           | 3.2 Reuse or build processing and analysis components | 4.2 Set up collection              | 5.2 Classify and code              | 6.2 Validate outputs              | 7.2 Produce dissemination products           | 8.2 Conduct evaluation       |
| 1.3 Establish output objectives      | 2.3 Design collection                      | 3.3 Reuse or build dissemination components           | 4.3 Run collection                 | 5.3 Review and validate            | 6.3 Interpret and explain outputs | 7.3 Manage release of dissemination products | 8.3 Agree an action plan     |
| 1.4 Identify concepts                | 2.4 Design frame and sample                | 3.4 Configure workflows                               | 4.4 Finalise collection            | 5.4 Edit and impute                | 6.4 Apply disclosure control      | 7.4 Promote dissemination products           |                              |
| 1.5 Check data availability          | 2.5 Design processing and analysis         | 3.5 Test production systems                           |                                    | 5.5 Derive new variables and units | 6.5 Finalise outputs              | 7.5 Manage user support                      |                              |
| 1.6 Prepare and submit business case | 2.6 Design production systems and workflow | 3.6 Test statistical business process                 |                                    | 5.6 Calculate weights              |                                   |  |                              |
|                                      |  | 3.7 Finalise production systems                       |                                    | 5.7 Calculate aggregates           |                                   |  |                              |
|                                      |  |   |                                    | 5.8 Finalise data files            |                                   |  |                              |

<https://statswiki.unece.org/display/GSBPM>



**„Bad quality reduces trust very, very fast.“**

*Walter Radermacher, 2022*

# Starting Points

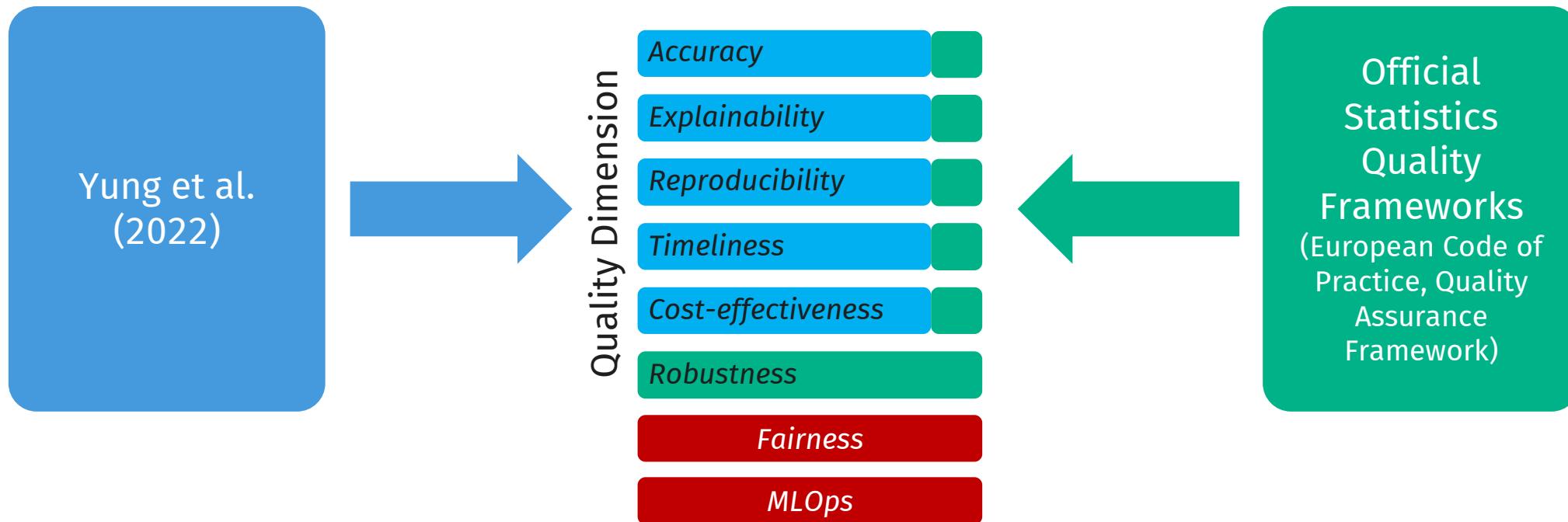
- **Quality Framework for Statistical Algorithms<sup>1)</sup> with its dimensions**
  - Accuracy
  - Explainability
  - Reproducibility
  - Timeliness
  - Cost-effectiveness
- **Quality frameworks in official statistics<sup>2)</sup>**



1) Yung W et al (2022) A quality framework for statistical algorithms. Statistical Journal of the IAOS, 38(1), 291–308

2) <https://www.destatis.de/DE/Methoden/Qualitaet/qualitaetshandbuch.html>, <https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework>, <https://unstats.un.org/unsd/unsystem/documents/UNSQAF-2018.pdf>

# Quality Dimensions



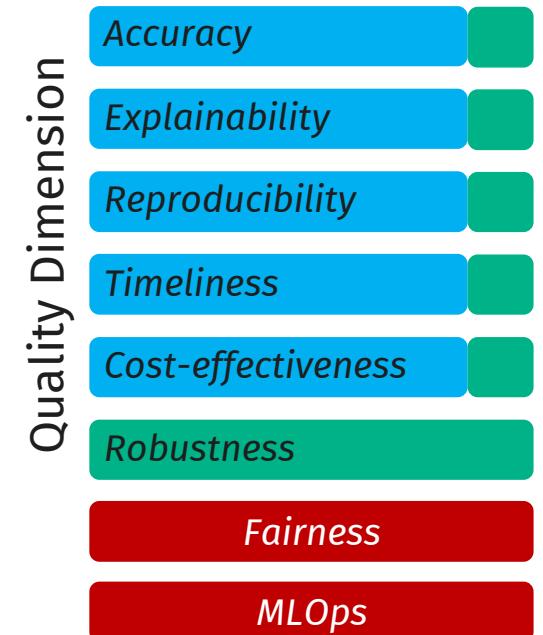
Yung W et al (2022) A quality framework for statistical algorithms. Statistical Journal of the IAOS, 38(1), 291–308

<https://ec.europa.eu/eurostat/web/quality/european-quality-standards/quality-assurance-framework>

Saidani Y et al (2023) Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik. AStA Wirtschafts- und Sozialstatistisches Archiv, 17(3-4), 253–303

# Our Three Main Contributions

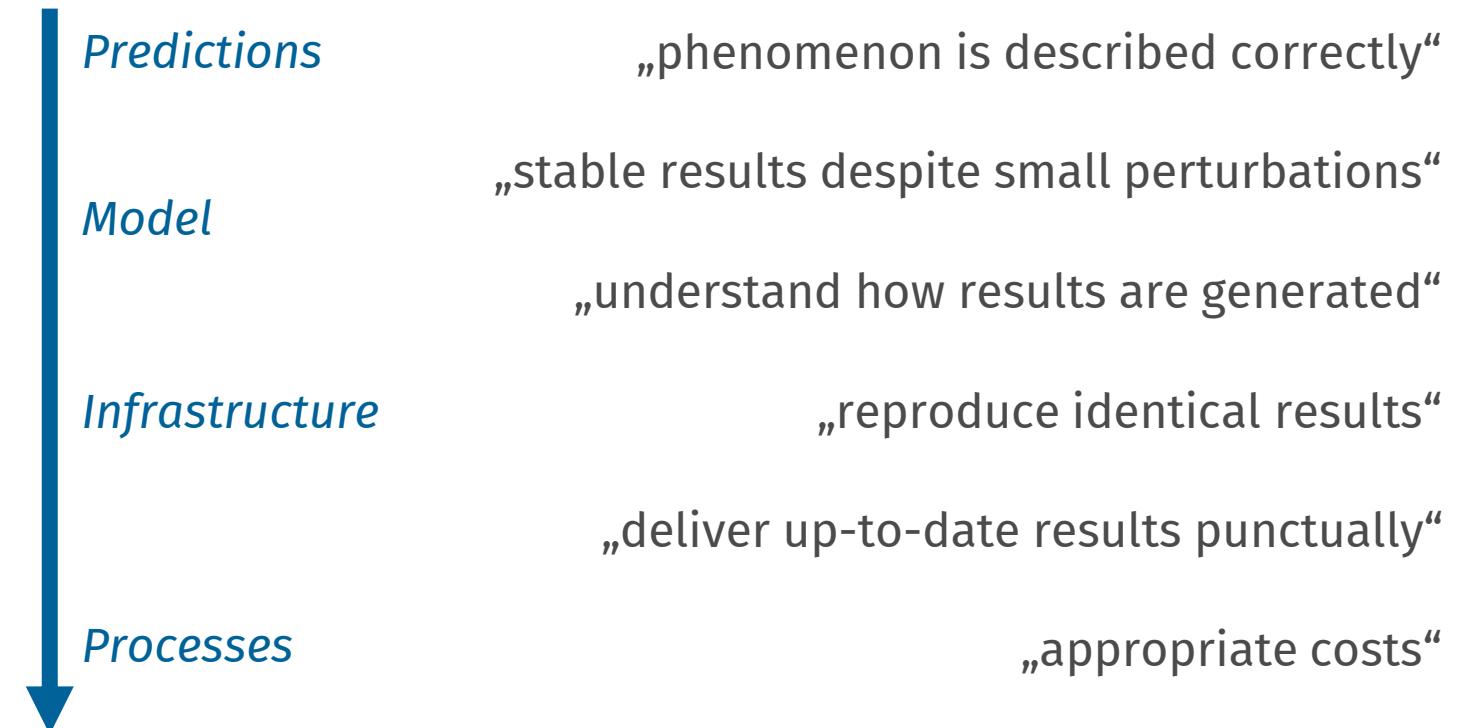
- » **Robustness** is proposed as a stand-alone quality dimension
- » Machine learning operations (**MLOps**) and **fairness** are discussed as two cross-cutting issues
- » Suggestions are made how quality assurance can be conducted **in practice** for each quality dimension



Saidani Y et al (2023) Qualitätsdimensionen maschinellen Lernens in der amtlichen Statistik. AStA Wirtschafts- und Sozialstatistisches Archiv, 17(3-4), 253–303

# Proposed Quality Dimensions

- » Accuracy
- » Robustness
- » Explainability
- » Reproducibility
- » Timeliness & Punctuality
- » Cost-effectiveness



# Our Approach

- » What is needed to ensure compatibility of ML applications used in official statistics production with existing official statistics quality standards?
  1. **Quality dimensions:** What does it mean for ML to have „high quality“?
  2. **Quality guidelines:** How to implement quality along the above dimensions during development?
  3. **Quality indicators:** How to evaluate quality in development & production?
  4. **Quality documentation:** How to communicate quality of ML in an appropriate, standardised way?
- » Work-in-progress: 1. completed, 2. in progress, 3. & 4. pending

# Quality Guidelines – Some Examples

| Dimension                | Guideline   |
|--------------------------|---|
| Accuracy                 | The quality measures relevant for the subject matter experts were determined, i.e. it was specified which quality measures were to be considered. |
| Robustness               | Target variables that are to be the subject of robustness were defined.   |
| Explainability           | If the other quality dimensions were (approximately) the same, the more explainable model was used.   |
| Reproducibility          | It was ensured that the data is identical regardless of the time of access, i.e. archived unchanged at a known location.                          |
| Timeliness & Punctuality | Sufficient time was planned for the conceptualisation, selection and testing of possible ML solutions.  |
| Cost-effectiveness       | As part of a proof of concept, the existing database was analysed and the feasibility of the project assessed.                                    |

# Fairness in Official Statistics

- » Aim/Definition: to **avoid treating certain (sub-)groups unjustifiably differently** in a relevant way by or as a result of statistical procedures (like ML)
- » In the context of official statistics, **effects are usually indirect**, e.g., through political decisions based on the published data
- » **Example:** statistical aggregates are systematically over- or underestimated for certain sub-groups (e.g., economic sectors, types of households, regions, ...)
- » **Connections** to accuracy (imbalanced data) and explainability

# By the Way

- » International conference
- » Around 150 participants
- » 19 countries
- » 5 keynotes, 36 talks
- » [www.destatis.de/ml-conference](http://www.destatis.de/ml-conference)



The screenshot shows the homepage of the conference website. At the top, there is a banner with a blue-to-red gradient featuring a blurred photo of people. The text on the banner reads "Conference on Foundations and Advances of Machine Learning in Official Statistics" and "3<sup>rd</sup> to 5<sup>th</sup> April, 2024". Below the banner, the navigation menu includes "Homepage", "About us", "Events", and "Machine Learning in Official Statistics". On the left side, there is a sidebar with links to "Program", "Scientific Committee", "Venue", and "Call for Abstracts". The main content area has a large logo for "ML in Official Statistics Wiesbaden 2024" with a blue circle containing the letters "ML" and the text "in Official Statistics" below it. To the right of the logo, there is a note: "Please find below the preliminary program of the conference on Foundations and Advances of ML in Official Statistics. Please note that this program may be subject to change." Below this note is a link "TO THE PROGRAM".

# Thank you! Questions?

**Florian Dumpert**

Federal Statistical Office of Germany

[florian.dumpert@destatis.de](mailto:florian.dumpert@destatis.de)

+49 611 75-3887

[www.destatis.de](http://www.destatis.de)

