Nutritional Labels for Data and Models *

Julia Stoyanovich
New York University
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
University of Washington
Seattle, WA, USA
billhowe@uw.edu

Abstract

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often used outside of the original context for which they were intended. In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a nutritional label, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Nutritional labels are derived automatically or semi-automatically as part of the complex process that gave rise to the data or model they describe, embodying the paradigm of interpretability-by-design. In this paper we further motivate nutritional labels, describe our instantiation of this paradigm for algorithmic rankers, and give a vision for developing nutritional labels that are appropriate for different contexts and stakeholders.

1 Introduction

An essential ingredient of successful machine-assisted decision-making, particularly in high-stakes decisions, is interpretability — allowing humans to understand, trust and, if necessary, contest, the computational process and its outcomes. These decision-making processes are typically complex: carried out in multiple steps, employing models with many hidden assumptions, and relying on datasets that are often repurposed — used outside of the original context for which they were intended.[1] In response, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we propose to develop interpretability and transparency tools based on the concept of a *nutritional label*, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes. Short of setting up a chemistry lab, the consumer would otherwise

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

[*] This work was supported in part by NSF Grants No. 1926250, 1916647, and 1740996.

[1] See Section 1.4 of Salganik's "Bit by Bit" [24] for a discussion of data repurposing in the Digital Age, which he aptly describes as ''mixing readymades with custommades."

have no access to this information. Similarly, consumers of data products cannot be expected to reproduce the computational procedures just to understand fitness for their use. Nutritional labels, in contrast, are designed to support specific decisions by the consumer rather than completeness of information. A number of proposals for hand-designed nutritional labels for data, methods, or both have been suggested in the literature[9, 12, 17]; we advocate deriving such labels automatically or semi-automatically as a side effect of the computational process itself, embodying the paradigm of *interpretability-by-design*.

Interpretability means different things to different stakeholders, including individuals being affected by decisions, individuals making decisions with the help of machines, policy makers, regulators, auditors, vendors, data scientists who develop and deploy the systems, and members of the general public. Designers of nutritional labels must therefore consider *what* they are explaining, *to whom*, and *for what purpose*. In the remainder of this section, we will briefly describe two regulatory frameworks that mandate interpretability of data collection and processing to members of the general public, auditors, and regulators, where nutritional labels offer a compelling solution (Section 1.1). We then discuss interpretability requirements in data sharing, particularly when data is altered to protect privacy or mitigate bias (Section 1.2).

1.1 Regulatory Requirements for Interpretability

The European Union recently enacted a sweeping regulatory framework known as the General Data Protection Regulation, or the GDPR [30]. The regulation was adopted in April 2016, and became enforceable about two years later, on May 25, 2018. The GDPR aims to protect the rights and freedoms of natural persons with regard to how their personal data is processed, moved, and exchanged (Article 1). The GDPR is broad in scope, and applies to "the processing of personal data wholly or partly by automated means" (Article 2), both in the private sector and in the public sector. Personal data is broadly construed, and refers to any information relating to an identified or identifiable natural person, called the *data subject* (Article 4).

According to Article 4, lawful processing of data is predicated on the data subject's *informed consent*, stating whether their personal data can be used, and for what purpose (Articles 6, 7). Further, data subjects have *the right to be informed* about the collection and use of their data. [2] Providing insight to data subjects about the collection and use of their data requires technical methods that support interpretability.

Regulatory frameworks that mandate interpretability are also starting to emerge in the US. New York City was the first US municipality to pass a law (Local Law 49 of 2018) [32], requiring that a task force be put in place to survey the current use of "automated decision systems" (ADS) in city agencies. ADS are defined as "computerized implementations of algorithms, including those derived from machine learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions." The task force is developing recommendations for enacting algorithmic transparency by the agencies, and will propose procedures for: (i) requesting and receiving an explanation of an algorithmic decision affecting an individual (Section 3 (b) of Local Law 49); (ii) interrogating ADS for bias and discrimination against members of legally protected groups, and addressing instances in which a person is harmed based on membership in such groups (Sections 3 (c) and (d)); (iii) and assessing how ADS function and are used, and archiving the systems together with the data they use (Sections 3 (e) and (f)).

Other government entities in the US are following suit. Vermont is convening an Artificial Intelligence Task Force to "... make recommendations on the responsible growth of Vermont's emerging technology markets, the use of artificial intelligence in State government, and State regulation of the artificial intelligence field." [33]. Idaho's legislature has passed a law that eliminates trade secret protections for algorithmic systems used in criminal justice [31]. In early April 2019, Senators Booker and Wyden introduced the Algorithmic Accountability Act of 2019 to the US Congress [6]. The Act, if passed, would use "automated decision systems impact assessment" to address and remedy harms caused by algorithmic systems to federally protected classes of people. The act

[2] `https://gdpr-info.eu/issues/right-to-be-informed/`

empowers the Federal Trade Commission to issue regulations requiring larger companies to conduct impact assessments of their algorithmic systems.

The use of nutritional labels in response to these and similar regulatory requirements can benefit a variety of stakeholders. The designer of a data-driven algorithmic method may use them to validate assumptions, check legal compliance, and tune parameters. Government agencies may exchange labels to coordinate service delivery, for example when working to address the opioid epidemic, where at least three sectors must coordinate: health care, criminal justice, and emergency housing, implying a global optimization problem to assign resources to patients effectively, fairly and transparently. The general public may review labels to hold agencies accountable to their commitment to equitable resource distribution.

1.2 Interpretability with Semi-synthetic Data

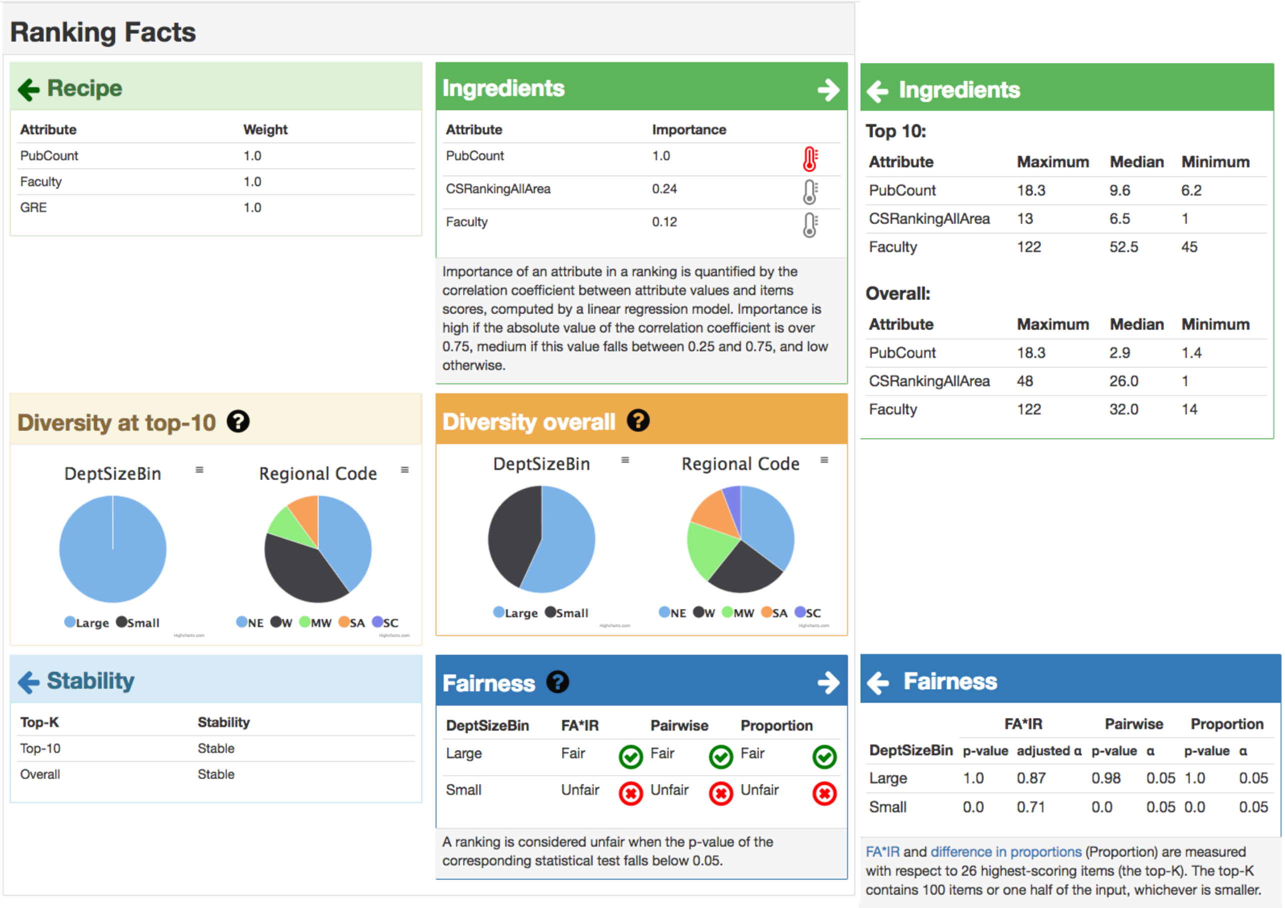
A central issue in machine-assisted decision-making is its reliance on historical data, which often embeds results of historical discrimination, also known as *structural bias*. As we have seen time and time again, models trained on data will appear to work well, but will silently and dangerously reinforce discrimination [1, 7, 13]. Worse yet, these models will legitimize the bias — "the computer said so." Nutritional labels for data and models are designed specifically to mitigate the harms implied by these scenarios, in contrast to the more general concept of "data about data."

Good datasets drive research: they inform new methods, focus attention on important problems, promote a culture of reproducibility, and facilitate communication across discipline boundaries. But research-ready datasets are scarce due to the high potential for misuse. Researchers, analysts, and practitioners therefore too often find themselves compelled to use the data they have on hand rather than the data they would (or should) like to use. For example, aggregate usage patterns of ride hailing services may overestimate demand in early-adopter (i.e., wealthy) neighborhoods, creating a feedback loop that reduces service in poorer neighborhoods, which in turn reduces usage. In this example, and in many others, there is a need to alter the input dataset to achieve specific properties in the output, while preserving all other relevant properties. We refer to such altered datasets as *semi-synthetic*.

Recent examples of methods that produce semi-synthetic data include database repair for causal fairness [25], database augmentation for coverage enhancement [4], and privacy-preserving and bias-correcting data release [21, 23]. A semi-synthetic datasets may be altered in different ways. Noise may be added to it to protect privacy, or statistical bias may be removed or deliberately introduced. When a dataset of this kind is released, its composition and the process by which it was derived must be made interpretable to a data scientist, helping determine fitness for use. For example, datasets repaired for racial bias are unsuitable for studying discrimination mitigation methods, while datasets with bias deliberately introduced are less appropriate for research unrelated to fairness. This gives another compelling use case for nutritional labels.

2 Nutritional Labels for Algorithmic Rankers

To make our discussion more concrete, we now describe Ranking Facts, a system that automatically derives nutritional labels for rankings [36]. Algorithmic decisions often result in scoring and ranking individuals — to determine credit worthiness, desirability for college admissions and employment, and compatibility as dating partners. Algorithmic rankers take a collection of items as input and produce a ranking – a sorted list of items – as output. The simplest kind of a ranker is a score-based ranker, which computes a score for each item independently, and then sorts the items on their scores. While automatic and seemingly objective, rankers can discriminate against individuals and protected groups [5], and exhibit low diversity at top ranks [27]. Furthermore, ranked results are often unstable — small changes in the input or in the ranking methodology may lead to drastic changes in the output, making the result uninformative and easy to manipulate [11]. Similar concerns apply in cases where



Ranking Facts

Recipe

Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Ingredients

Attribute	Importance
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Importance of an attribute in a ranking is quantified by the correlation coefficient between attribute values and items scores, computed by a linear regression model. Importance is high if the absolute value of the correlation coefficient is over 0.75, medium if this value falls between 0.25 and 0.75, and low otherwise.

Ingredients

Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
CSRankingAllArea	13	6.5	1
Faculty	122	52.5	45

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
CSRankingAllArea	48	26.0	1
Faculty	122	32.0	14

Diversity at top-10

Diversity overall

Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair ✓	Fair ✓	Fair ✓
Small	Unfair ⊗	Unfair ⊗	Unfair ⊗

A ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05.

Fairness

	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	α
Large	1.0	0.87	0.98	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

FA*IR and difference in proportions (Proportion) are measured with respect to 26 highest-scoring items (the top-K). The top-K contains 100 items or one half of the input, whichever is smaller.

Figure 1: Ranking Facts for the CS departments dataset. The Ingredients widget (green) has been expanded to show the details of the attributes that strongly influence the ranking. The Fairness widget (blue) has been expanded to show the computation that produced the fair/unfair labels.

items other than individuals are ranked, including colleges, academic departments, and products.

In a recent work, we developed Ranking Facts, a nutritional label for rankings [36]. Ranking Facts is available as a Web-based tool[3], and its code is available in the open source [4]. Figure 1 presents Ranking Facts that explains a ranking of Computer Science departments. The data in this example was obtained from CS Rankings[5], augmented with attributes from the NRC dataset [6]. Ranking Facts is made up of a collection of visual widgets, each with an overview and a detailed view. Each widget addresses an essential aspect of transparency and interpretability, and is based on our recent technical work on fairness [3, 35], diversity [8, 27, 28, 34], and stability [2] in algorithmic rankers. We now describe each widget in some detail.

2.1 Recipe and Ingredients

These two widgets help to explain the ranking methodology. The Recipe widget succinctly describes the ranking algorithm. For example, for a linear scoring formula, each attribute would be listed together with its weight. The

[3] http://demo.dataresponsibly.com/rankingfacts/

[4] https://github.com/DataResponsibly/RankingFacts

[5] https://github.com/emeryberger/CSRankings

[6] http://www.nap.edu/rdp/

Ingredients widget lists attributes most material to the ranked outcome, in order of importance. For example, for a linear model, this list could present the attributes with the highest learned weights. Put another way, the explicit intentions of the designer of the scoring function about which attributes matter, and to what extent, are stated in the Recipe, while Ingredients may show attributes that are actually associated with high rank. Such associations can be derived with linear models or with other methods, such as rank-aware similarity in our prior work [27]. The detailed Recipe and Ingredients widgets list statistics of the attributes in the Recipe and in the Ingredients: minimum, maximum and median values at the top-10 and over-all.

2.2 Stability

The Stability widget explains whether the ranking methodology is robust on this particular dataset. An unstable ranking is one where slight changes to the data (e.g., due to uncertainty and noise), or to the methodology (e.g., by slightly adjusting the weights in a score-based ranker) could lead to a significant change in the output. This widget reports a stability score, as a single number that indicates the extent of the change required for the ranking to change. As with the widgets above, there is a detailed Stability widget to complement the overview widget.

An example is shown in Figure 2, where the stability of the ranking is quantified as the slope of the line that is fit to the score distribution, at the top-10 and over-all. A score distribution is unstable if scores of items in adjacent ranks are close to each other, and so a very small change in scores will lead to a change in the ranking. In this example the score distribution is considered unstable if the slope is 0.25 or lower. Alternatively, stability can be computed with respect to each scoring attribute, or it can be assessed using a model of uncertainty in the data. In these cases, stability quantifies the extent to which a ranked list will change as a result of small changes to the underlying data. A complementary notion of stability quantifies the magnitude of change as a result of small changes to the ranking model. We explored this notion in our recent work, briefly discussed below.

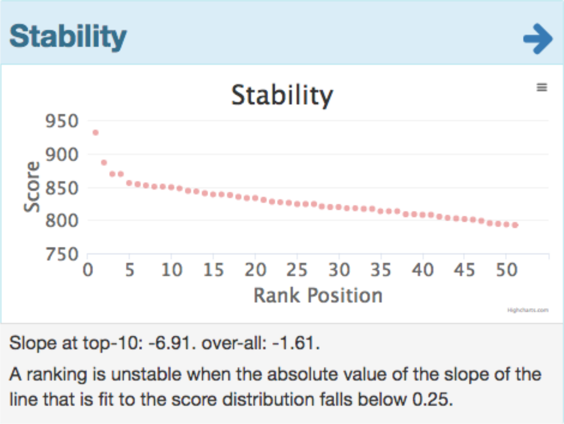
In [2] we develped methods for quantifying the stability of a score-based ranker with respect to a given dataset. Specifically, we considered rankers that specify non-negative weights, one for each item attribute, and compute the score as a weighted sum of attribute values. We focused on a notion of stability that quantifies whether the output ranking will change due to a small change in the attribute weights. This notion of stability is natural for consumers of a ranked list (i.e., those who use the ranking to prioritize items and make decisions), who should be able to assess the magnitude of the *region in the weight space* that produces the observed ranking. If this region is large, then the same ranked order would be obtained for many choices of weights, and the ranking is stable. But if this region is small, then we know that only a few weight choices can produce the observed ranking. This may suggest that the ranking was engineered or “cherry-picked” by the producer to obtain a specific outcome.

2.3 Fairness

The Fairness widget quantifies whether the ranked output exhibits statistical parity (one interpretation of fairness) with respect to one or more sensitive attributes, such as gender or race of individuals [35]. We denote one or several values of the sensitive attribute as a protected feature. For example, for the sensitive attribute gender, the assignment gender=F is a protected feature.

A variety of fairness measures have been proposed in the literature [38], with a primary focus on classification or risk assessment tasks. One typical fairness measure for classification compares the proportion of members of a protected group (e.g., female gender or minority race) who receive a positive outcome to their proportion in the overall population. For example, if the dataset contains an equal number of men and women, then among the individuals invited for a job interview, one half should be women. A measure of this kind can be adapted to rankings by quantifying the proportion of members of a protected group in some selected set of size k (treating the top-k as a set).

In [35], we were the first to propose a family of *fairness measures specifically for rankings*. Our measures are based on a generative process for rankings that meet a particular fairness criterion (fairness probability f) and



Slope at top-10: -6.91. over-all: -1.61.

A ranking is unstable when the absolute value of the slope of the line that is fit to the score distribution falls below 0.25.

Figure 2: Stability: detailed widget.

are drawn from a dataset with a given proportion of members of a binary protected group (p). This method was subsequently used in FA*IR [37] to quantify fairness in every prefix of a top-k list. We also developed a pairwise measure that directly models the probability that a member of a protected group is preferred to a member of the non-protected group.

Let us now return to the Fairness widget in Figure 1. We select a binary version of the department size attribute DeptSizeBin from the CS departments dataset as the sensitive attribute, and treat the value and "small" as the protected feature. The summary view of the Fairness widget in our example presents the output of three fairness measures: FA*IR [37], proportion [38], and our own pairwise measure. All these measures are statistical tests, and whether a result is fair is determined by the computed p-value. The detailed Fairness widget provides additional information about the tests and explains the process.

2.4 Diversity

Fairness is related to diversity: ensuring that different kinds of objects are represented in the output of an algorithmic process [8]. Diversity has been considered in search and recommender systems, but in a narrow context, and was rarely applied to profiles of individuals. The Diversity widget shows diversity with respect to a set of demographic categories of individuals, or a set of categorical attributes of other kinds of items [8]. The widget displays the proportion of each category in the top-10 ranked list and over-all, and, like other widgets, is updated as the user selects different ranking methods or sets different weights. In our example in Figure 1, we quantify diversity with respect to department size and to the regional code of the university. By comparing the pie charts for top-10 and over-all, we observe that only large departments are present in the top-10.

This simple diversity measure that is currently included in Ranking Facts can be augmented by, or replaced with, other measures, including, for example, those we developed in our recent work [28, 34].

3 Learning Labels

The creation of nutritional labels is often cast as a design problem rather than a computational problem [9, 12]. Standard labels with broad applicability can amortize the cost of design, but the diversity of datasets, methods, and desirable properties for nutritional labels suggest a learning approach to help develop labels for a variety of situations. Since opaque automation is what motivated the need for labels in the first place, automating their creation may seem like a step backwards. But there are several benefits:

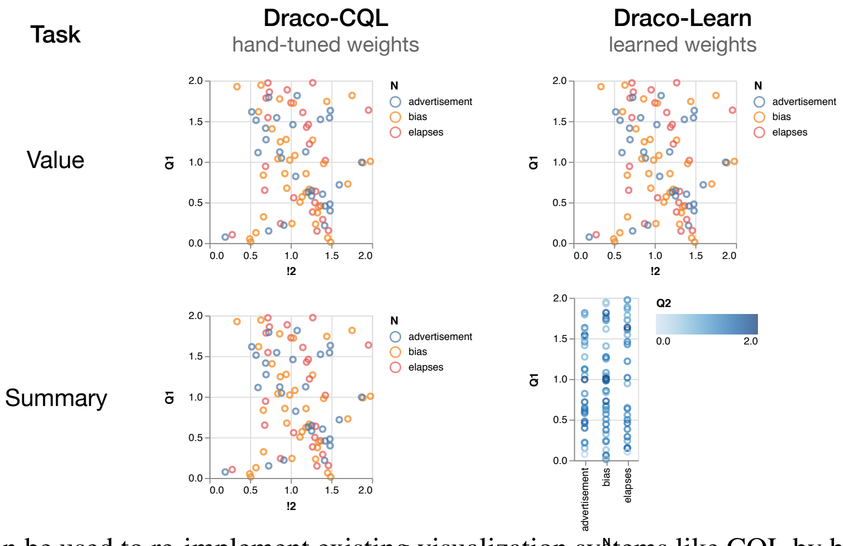


Figure 3: Draco can be used to re-implement existing visualization systems like CQL by hand-tuning weights (left) or be used to learn weights automatically from preference data (right). The visualizations selected can vary significantly, affording customization for specific applications. A similar approach can be used when generating nutritional labels for data and models.

- Coverage: *some* information provided in (nearly) *all* cases is preferable to *all* information provided in *some* cases, as there are many models and datasets being deployed.
- Correctness: Hand-designed labels imply human metadata attachment, but curation of metadata is essentially an unsolved problem. Computable labels reduce reliance on human curation efforts.
- Retroactivity: Some information can only be manually collected at the time of data collection (e.g., demographics of authors in a speech corpus to control for nationality bias). This opportunity is lost for existing datasets. However, inferring relevant properties based on the content of the data may be "better than nothing."

We now consider two approaches to the problem of learning labels, one based on the visualization recommendation literature, and one based on bin-packing optimization.

3.1 Learning as Visualization Recommendation

Moritz et al. proposed Draco [19], a formal model that represents visualizations as sets of logical facts, and represents design guidelines as a collection of hard and soft constraints over these facts, following an earlier proposal for the VizDeck system [14]. Draco enumerates the visualizations that do not violate the hard constraints and finds the most preferred visualizations according to the weights of the soft constraints. Formalized visualization descriptions are derived from the Vega-Lite grammar [26] extended with rules to encode expressiveness criteria [16], preference rules validated in perception experiments, and general visualization design best practices. Hard constraints *must* be satisfied (e.g., shape encodings cannot express quantitative values), whereas soft constraints express a preference (e.g., temporal values should use the x-axis by default). The weights associated with soft constraints can be learned from preference or utility data, when available (see example in Figure 3).

Draco implements the constraints using Answer Set Programming (ASP) semantics, and casts the problem of finding appropriate encodings as finding optimal answer sets [10]. Draco has been extended to optimize for constraints over multiple visualizations [22], and adapted for use in specialized domains.

Using Draco (or similar formalizations), the specialized constraints governing the construction of nutritional labels can be developed in the general framework of ASP, while borrowing the foundational constraints capturing general visualization design principles. This approach can help reduce the cost of developing hundreds of application-specific labels by encoding common patterns, such as including descriptive statistics in all labels, or only showing fairness visualizations when bias is detected.

3.2 Learning as Optimization

Sun et al. proposed MithraLabel [29], focusing on generating task-specific labels for datasets to determine fitness for specific tasks. Considering the dataset as a collection of items over a set of attributes, each widget provides specific information (such as functional dependencies) about the whole dataset or some selected part of it. For example, if a data scientist is considering the use of a number-of-prior-arrests attribute to predict likelihood of recidivism, she should know that the number of prior arrests is highly correlated with the likelihood of re-offending, but it introduces bias as the number of prior arrests is higher for African Americans than for other races due to policing practices and segregation effects in poor neighborhoods. Widgets that might appear in the nutritional label for prior arrests include the count of missing values, correlation with the predicted attribute or a protected attribute, and the distribution of values.

4 Properties of a nutritional label

The database and cyberinfrastructure communities have been studying systems and standards for metadata, provenance, and transparency for decades [20, 18]. We are now seeing renewed interest in these topics due to the proliferation of data science applications that use data opportunistically. Several recent projects explore these concepts for data and algorithmic transparency, including the Dataset Nutrition Label project [12], Datasheets for Datasets [9], and Model Cards [17]. All these method rely on manually constructed annotations. In contrast, our goal is to *generate labels automatically or semi-automatically*.

To differentiate a nutritional label from more general forms of metadata, we articulate several properties:

- **Comprehensible**: The label is not a complete (and therefore overwhelming) history of every processing step applied to produce the result. This approach has its place and has been extensively studied in the literature on scientific workflows, but is unsuitable for the applications we target. The information on a nutritional label must be short, simple, and clear.

- **Consultative**: Nutritional labels should provide actionable information, rather than just descriptive metadata. For example, universities may invest in research to improve their ranking, or consumers may cancel unused credit card accounts to improve their credit score.

- **Comparable**: Nutritional labels enable comparisons between related products, implying a standard. The IEEE is developing a series of ethics standards, known as the IEEE P70xx series, as part of its Global Initiative on Ethics of Autonomous and Intelligent Systems.[7] These standards include “IEEE P7001: Transparency of Autonomous Systems” and “P7003: Algorithmic Bias Considerations” [15]. The work on nutritional labels is synergistic with these efforts.

- **Concrete**: The label must contain more than just general statements about the source of the data; such statements do not provide sufficient information to make technical decisions on whether or not to use the data.

Data and models are chained together into complex automated pipelines — computational systems “consume” datasets at least as often as people do, and therefore also require nutritional labels! We articulate additional properties in this context:

[7] https://ethicsinaction.ieee.org/

- **Computable**: Although primarily intended for human consumption, nutritional labels should be machine-readable to enable specific applications: data discovery, integration, automated warnings of potential misuse.
- **Composable**: Datasets are frequently integrated to construct training data; the nutritional labels must be similarly integratable. In some situations, the composed label is simple to construct: the union of sources. In other cases, the biases may interact in complex ways: a group may be sufficiently represented in each source dataset, but underrepresented in their join.
- **Concomitant**: The label should be carried with the dataset; systems should be designed to propagate labels through processing steps, modifying the label as appropriate, and implementing the paradigm of transparency by design.

5 Conclusions

In this paper we discussed work on transparency and interpretability for data and models based on the concept of a nutritional label. We presented Ranking Facts, a system that automatically derives nutritional labels for rankings, and outlined directions for ongoing research that casts the creation of nutritional labels as a computational problem, rather than as purely a design problem.

We advocate interpretability tools for a variety of datasets and models, for a broad class of application domains, and to accommodate the needs of a variety of stakeholders. These tools must be informed by an understanding of how humans perceive algorithms and the decisions they inform, including issues of trust and agency to challenge or accept an algorithm-informed decision. These tools aim to reduce bias and errors in deployed models by preventing the use of an inappropriate dataset or model at design time. Although the extent of data misuse is difficult to measure directly, we can design experiments to show how well nutritional labels inform usage decisions, and design the tools accordingly. More broadly, we see the review of human-curated and machine-computed metadata as a critical step for interpretability in data science, which can lead to lasting progress in the use of machine-assisted decision-making in society.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: Risk assessments in criminal sentencing. *ProPublica*, May 2016.

[2] Abolfazl Asudeh, H. V. Jagadish, Gerome Miklau, and Julia Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.

[3] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019.*, pages 1259–1276, 2019.

[4] Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 554–565, 2019.

[5] Danielle K. Citron and Frank A. Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89, 2014.

[6] Cory Booker, Ron Wyden, Yvette Clarke. Algorithmic Accountability Act. `https://www.wyden.senate.gov/imo/media/doc/Algorithmic\%20Accountability\%20Act\%20of\%202019\%20Bill\%20Text.pdf`, 2019. [Online; accessed 3-May-2019].

[7] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, October 2018.

[8] Marina Drosou, HV Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. Diversity in Big Data: A review. *Big Data*, 5(2), 2017.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[10] Martin Gebser. *Proof theory and algorithms for answer set programming*. PhD thesis, University of Potsdam, 2011.

[11] Malcolm Gladwell. The order of things. *The New Yorker*, February 14, 2011.

[12] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677, 2018.

[13] David Ingold and Spencer Soper. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, April 2016.

[14] Alicia Key, Bill Howe, Daniel Perry, and Cecilia R. Aragon. Vizdeck: self-organizing dashboards for visual analytics. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 681–684, 2012.

[15] Ansgar R. Koene, Liz Dowthwaite, and Suchana Seth. IEEE p7003^{TM} standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 38–41, 2018.

[16] Jock D. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.*, 5(2):110–141, 1986.

[17] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229, 2019.

[18] Luc Moreau, Bertram Ludäscher, Ilkay Altintas, Roger S. Barga, Shawn Bowers, Steven P. Callahan, George Chin Jr., Ben Clifford, Shirley Cohen, Sarah Cohen Boulakia, Susan B. Davidson, Ewa Deelman, Luciano A. Digiampietri, Ian T. Foster, Juliana Freire, James Frew, Joe Futrelle, Tara Gibson, Yolanda Gil, Carole A. Goble, Jennifer Golbeck, Paul T. Groth, David A. Holland, Sheng Jiang, Jihie Kim, David Koop, Ales Krenek, Timothy M. McPhillips, Gaurang Mehta, Simon Miles, Dominic Metzger, Steve Munroe, Jim Myers, Beth Plale, Norbert Podhorszki, Varun Ratnakar, Emanuele Santos, Carlos Eduardo Scheidegger, Karen Schuchardt, Margo I. Seltzer, Yogesh L. Simmhan, Cláudio T. Silva, Peter Slaughter, Eric G. Stephan, Robert Stevens, Daniele Turi, Huy T. Vo, Michael Wilde, Jun Zhao, and Yong Zhao. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, 2008.

[19] Dominik Moritz, Chenglong Wang, Gregory Nelson, Halden Lin, Adam M. Smith, Bill Howe, and Jeffrey Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2019.

[20] Open provenance. `https://openprovenance.org`. [Online; accessed 14-August-2019].

[21] Haoyue Ping, Julia Stoyanovich, and Bill Howe. Datasynthesizer: Privacy-preserving synthetic datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 42:1–42:5, 2017.

[22] Zening Qu and Jessica Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Trans. Vis. Comput. Graph.*, 24(1):468–477, 2018.

[23] Luke Rodriguez, Babak Salimi, Haoyue Ping, Julia Stoyanovich, and Bill Howe. MobilityMirror: Bias-adjusted transportation datasets. In *Big Social Data and Urban Computing - First Workshop, BiDU@VLDB 2018, Rio de Janeiro, Brazil, August 31, 2018, Revised Selected Papers*, pages 18–39, 2018.

[24] Matthew J. Salganik. *Bit By Bit: Social Research in the Digital Age*. Princeton University Press, 2019.

[25] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019.*, pages 793–810, 2019.

[26] Arvind Satyanarayan, Dominik Moritz, Kanit Wongsuphasawat, and Jeffrey Heer. Vega-lite: A grammar of interactive graphics. *IEEE Trans. Vis. Comput. Graph.*, 23(1):341–350, 2017.

[27] Julia Stoyanovich, Sihem Amer-Yahia, and Tova Milo. Making interval-based clustering rank-aware. In *EDBT 2011, 14th International Conference on Extending Database Technology, Uppsala, Sweden, March 21-24, 2011, Proceedings*, pages 437–448, 2011.

[28] Julia Stoyanovich, Ke Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In *Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018.*, pages 241–252, 2018.

[29] Chenkai Sun, Abolfazl Asudeh, H. V. Jagadish, Bill Howe, and Julia Stoyanovich. MithraLabel: Flexible dataset nutritional labels for responsible data science. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

[30] The European Union. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). `https://gdpr-info.eu/`, 2016. [Online; accessed 15-August-2019].

[31] The Idaho house of Representatives. House Bill No. 118. `https://legislature.vermont.gov/bill/status/2018/H.378`, 2019. [Online; accessed 15-August-2019].

[32] The New York City Council. Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies. `https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2-9C42-461253F9C6D0`, 2017. [Online; accessed on 15-August-2019].

[33] Vermont General Assembly. An act relating to the creation of the Artificial Intelligence Task Force. `https://legislature.idaho.gov/wp-content/uploads/sessioninfo/2019/legislation/H0118A2.pdf`, 2018. [Online; accessed 15-August-2019].

[34] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6035–6042, 2019.

[35] Ke Yang and Julia Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 22:1–22:6, 2017.

[36] Ke Yang, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, H. V. Jagadish, and Gerome Miklau. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1773–1776, 2018.

[37] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo A. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1569–1578, 2017.

[38] Indre Zliobaite. Measuring discrimination in algorithmic decision making. *Data Min. Knowl. Discov.*, 31(4):1060–1089, 2017.