Responsible Data Science

Interpretability & Legal Frameworks

May 2 & 4, 2022

Prof. George Wood

Center for Data Science
New York University

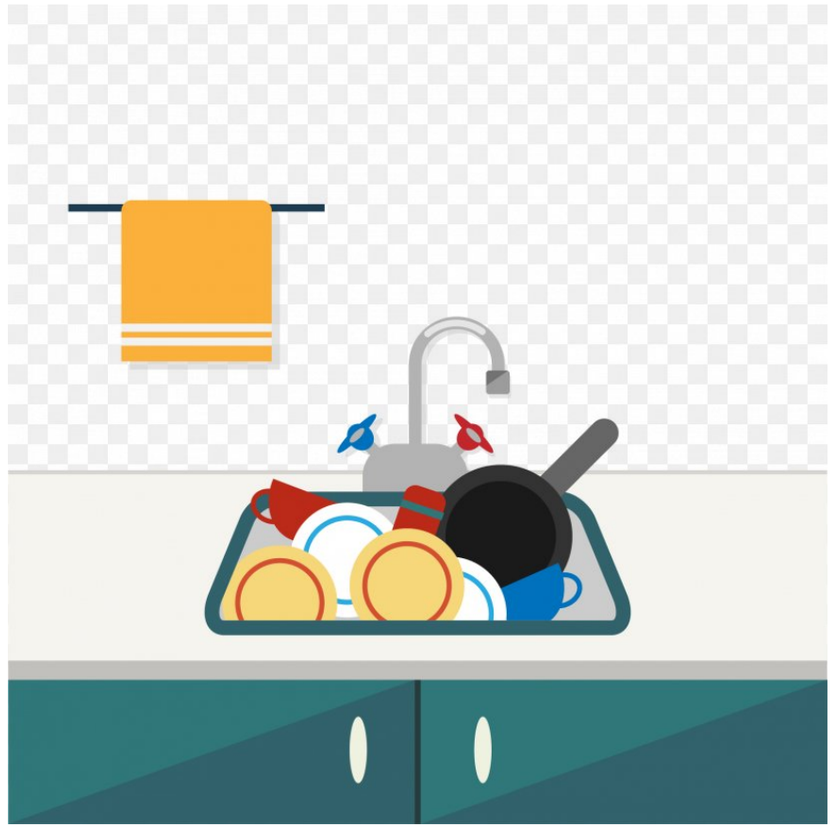
NYU | Center for Data Science

r/ai

What is interpretability?

- **Explaining black-box models**
- **Online ad targeting**
- **Interpretability**

A kitchen sink? Or a foundational concept for responsible data science?



(Image source)

Algorithmic rankers

https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

Input: database of items (individuals, colleges, cars, …)

Score-based ranker: computes the score of each item using a **known formula**, often a monotone aggregation function, then sorts items on score

Output: permutation of the items, complete or top-k

Do we have transparency?

	$\mathcal{D}$		f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.72	0.65	1.37
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35
t_6	0.61	0.79	1.4

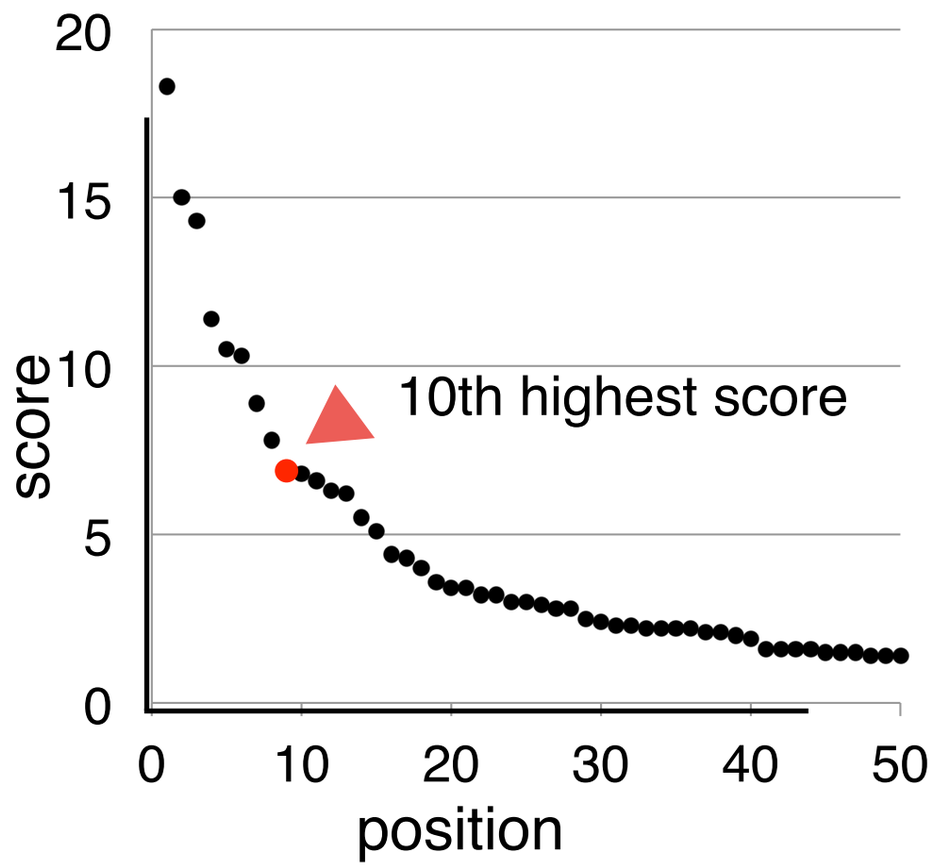
We have syntactic transparency, but lack interpretability!

Opacity in algorithmic rankers

https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

Reason 1: The scoring formula alone does not indicate the relative rank of an item.

Scores are absolute, rankings are relative. Is 5 a good score? What about 10? 15?



Opacity in algorithmic rankers

https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

Reason 2: A ranking may be unstable if there are tied or nearly-tied items.

Rank	Institution	Average Count	Faculty
1	▶ Carnegie Mellon University	18.4	123
2	▶ Massachusetts Institute of Technology	15.6	64
3	▶ Stanford University	14.8	56
4	▶ University of California - Berkeley	11.5	50
5	▶ University of Illinois at Urbana-Champaign	10.6	56
6	▶ University of Washington	10.3	50
7	▶ Georgia Institute of Technology	8.9	81
8	▶ University of California - San Diego	8	51
9	▶ Cornell University	7	45
10	▶ University of Michigan	6.8	63
11	▶ University of Texas - Austin	6.6	43
12	▶ University of Massachusetts - Amherst	6.4	47

Opacity in algorithmic rankers

https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

Reason 3: A ranking methodology may be unstable: small changes in weights can trigger significant re-shuffling.

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

1. Porsche Cayman 193

2. Chevrolet Corvette 186

3. Lotus Evora 182

1. Chevrolet Corvette 205

2. Lotus Evora 195

3. Porsche Cayman 195

1. Lotus Evora 205

2. Porsche Cayman 198

3. Chevrolet Corvette 192

https://www.newyorker.com/magazine/2011/02/14/the-order-of-things

Opacity in algorithmic rankers

https://freedom-to-tinker.com/2016/08/05/revealing-algorithmic-rankers/

Reason 4: The **weight of an attribute** in the scoring formula **does not determine its impact** on the outcome.

Rank	Name	Avg Count	Faculty	Pubs	GRE
1	CMU	18.3	122	2	791
2	MIT	15	64	3	772
3	Stanford	14.3	55	5	800
4	UC Berkeley	11.4	50	3	789
5	UIUC	10.5	55	3	772
6	UW	10.3	50	2	796
		• • • •			
39	U Chicago	2	28	2	779
40	UC Irvine	1.9	28	2	787
41	BU	1.6	15	2	783
41	U Colorado Boulder	1.6	32	1	761
41	UNC Chapel Hill	1.6	22	2	794
41	Dartmouth	1.6	18	2	794

Given a score function:

$$0.2 * faculty +$$

$$0.3 * avg\ cnt +$$

$$0.5 * gre$$

Rankings are not benign!

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

Rankings are not benign. They enshrine very particular ideologies, and, at a time when American higher education is facing a crisis of accessibility and affordability, we have adopted **a de-facto standard of college quality** that is uninterested in both of those factors. And why? Because a group of magazine analysts in an office building in Washington, D.C., decided twenty years ago to **value selectivity over efficacy**, to **use proxies** that scarcely relate to what they're meant to be proxies for, and to **pretend that they can compare** a large, diverse, low-cost land-grant university in rural Pennsylvania with a small, expensive, private Jewish university on two campuses in Manhattan.

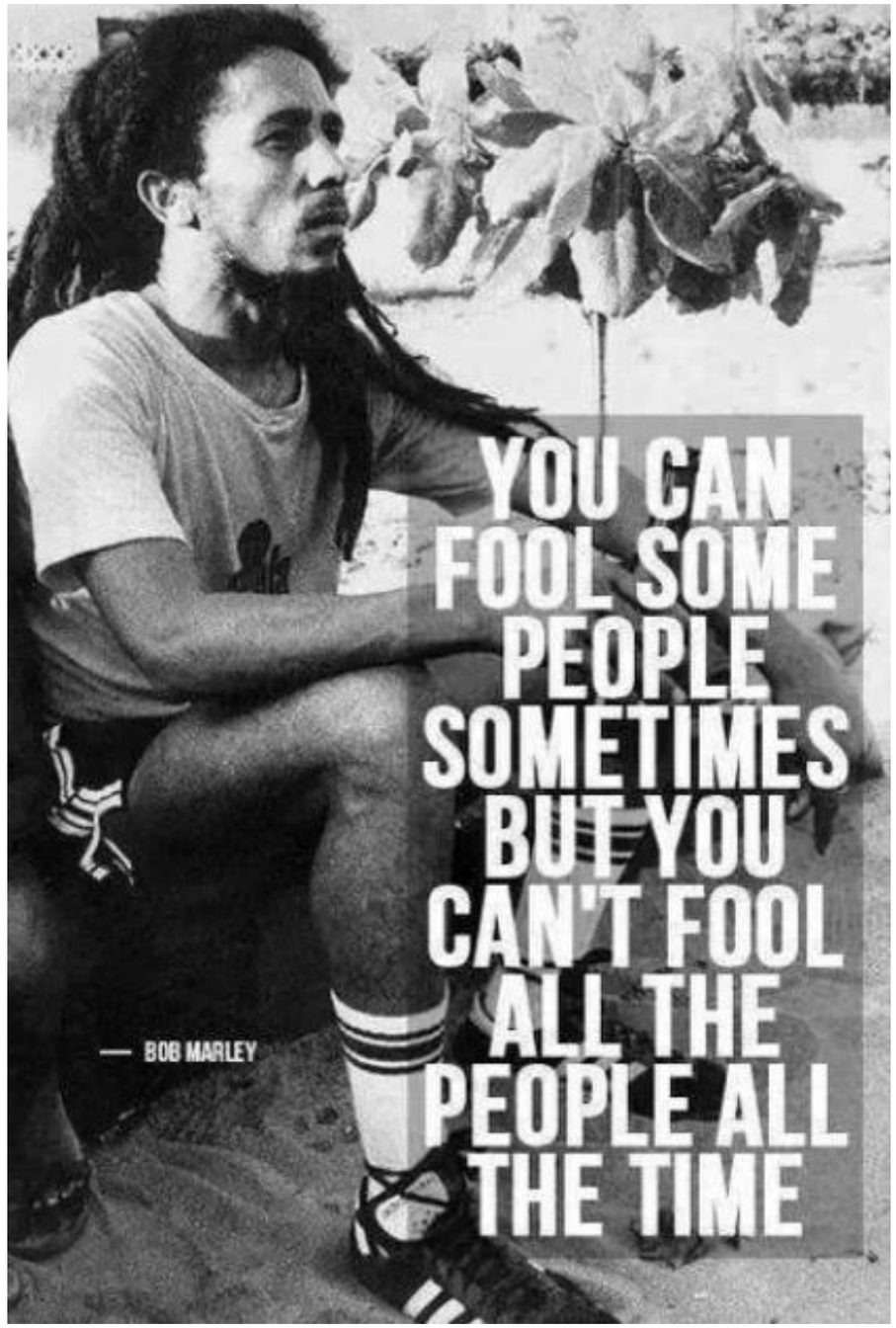
Interpretability in the service of trust!

Gladwell makes the point that rankings are claiming objectivity, yet are comparing apples and oranges.

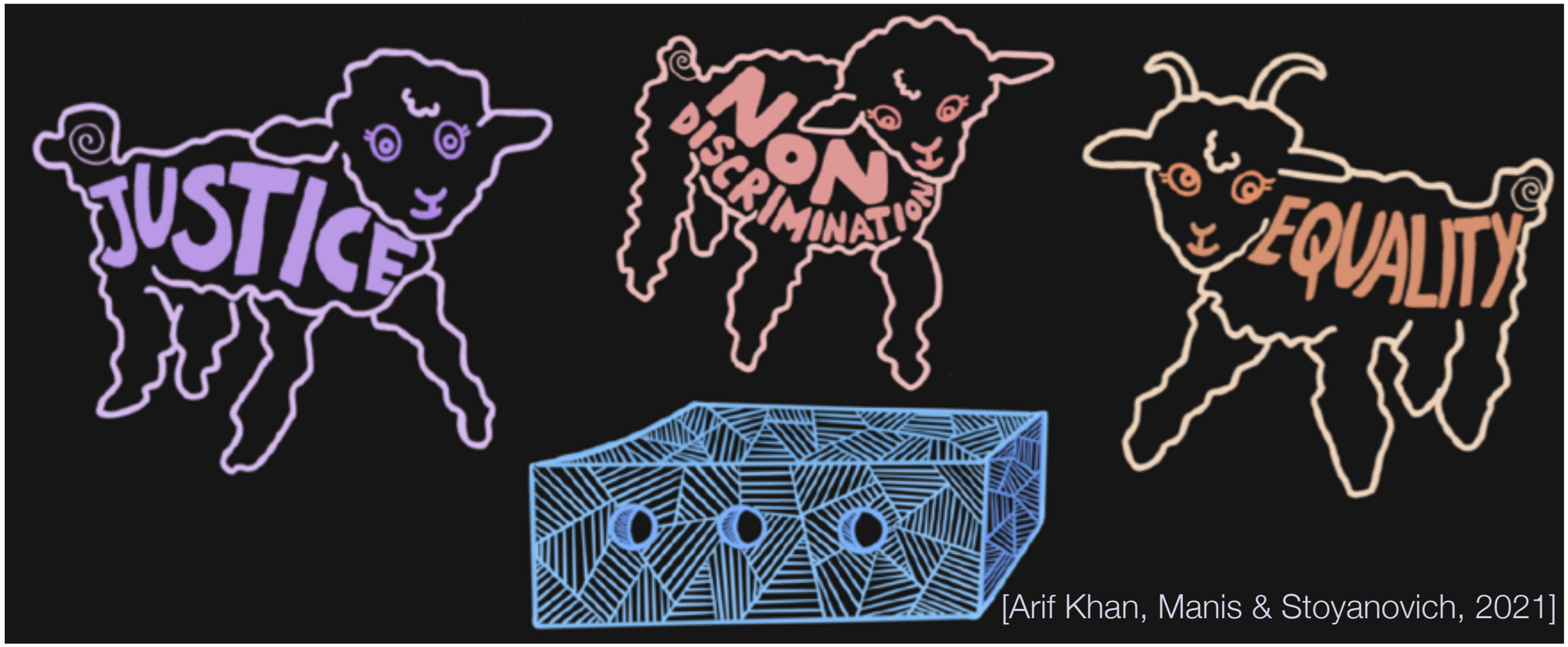
In that sense, **a score-based ranker is a quintessential "black box" of data science**, and perhaps the simplest possible such black box.

AI is a red herring, privacy / IP / gaming arguments are overused. The truly difficult issues are that:

1) using math to pretend that we are correct when making intrinsically subjective decisions reinforcing the balance of power in society
2) that math / objectivity is used as a substitute for trust, but **trust must run deeper than math**!
3) need to find a kind of an interpretability that will enable trust!



The fairness you asked for is inside this box



[Arif Khan, Manis & Stoyanovich, 2021]

data protection:
the GDPR

r/ai

General Data Protection Regulation
GDPR

Welcome to gdpr-info.eu. Here you can find the official PDF of the Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018 as a neatly arranged website. All Articles of the GDPR are linked with suitable recitals. The European Data Protection Regulation is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe. If you find the page useful, feel free to support us by sharing the project.

Quick Access

adopted in April 2016

enforced since May 25, 2018

https://gdpr-info.eu/

r/ai

GDPR: scope and definitions

Article 2: Material Scope

- This Regulation applies to the processing of personal data wholly or partly by automated means and to the processing other than by automated means of personal data which form part of a filing system or are intended to form part of a filing system.

Article 4: Definitions

- ‘**personal data**’ means any information relating to an identified or identifiable natural person (‘**data subject**’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
- ‘**processing**’ means **any operation** or set of operations which is performed on personal data or on sets of personal data, **whether or not by automated means**, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction;

GDPR: scope and definitions

Article 4: Definitions

- ‘**controller**’ means the natural or legal person, public authority, agency or other body which, alone or jointly with others, **determines the purposes and means of the processing** of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law;
- ‘**processor**’ means a natural or legal person, public authority, agency or other body which **processes personal data on behalf of the controller**;
- ‘**consent**’ of the data subject means any **freely given, specific, informed and unambiguous** indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, **signifies agreement to the processing of personal data** relating to him or her;

Art. 7 GDPR

Conditions for consent

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.

2. [1] If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. [2] Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.

Art. 7 GDPR

Conditions for consent

3. [1] The data subject shall have the right to withdraw his or her consent at any time. [2] The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. [3] Prior to giving consent, the data subject shall be informed thereof. [4] It shall be as easy to withdraw as to give consent.

4. When assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.

Chapter 3

Rights of the data subject

Chapter 3

Rights of the data subject

Section 3	–	**Rectification and erasure**
Article 16	–	Right to rectification
Article 17	–	Right to erasure (‘right to be forgotten’)
Article 18	–	Right to restriction of processing
Article 19	–	Notification obligation regarding rectification or erasure of personal data or restriction of processing
Article 20	–	Right to data portability

Removing personal data

The right to be forgotten (Article 17)

- Similar laws exist in other jurisdictions, e.g., Argentina (since 2006)
- Resulted in many dereferencing requests to search engines
- Often seen as controversial: **reasons?**
- May conflict with other legal requirements, or with technical requirements

Also, just technically challenging:

- have to re-engineer the data management stack, **what are the issues**?
- what about models?

[S. Abiteboul and J. Stoyanovich, 2019]

r/ai

Chapter 3

Rights of the data subject

Section 3	–	**Rectification and erasure**
Article 16	–	Right to rectification
Article 17	–	Right to erasure (‘right to be forgotten’)
Article 18	–	Right to restriction of processing
Article 19	–	Notification obligation regarding rectification or erasure of personal data or restriction of processing
Article 20	–	Right to data portability

Moving personal data

The right to data portability (Article 20)

- Aims to prevent vendor lock-in
- What are some technical difficulties?
  - Suppose you want to move your photos from Service A to Service B?
  - What about moving your social interactions from Service A to Service B?
- Can we look at this from the point of view of **inter-operability** rather than moving data?

[S. Abiteboul and J. Stoyanovich, 2019]

r/ai

Moving personal data

Data Transfer Project

Download White Paper

About Community Documentation Updates FAQ

About us

The Data Transfer Project was launched in 2018 to create an open-source, service-to-service data portability platform so that all individuals across the web could easily move their data between online service providers whenever they want.

The contributors to the Data Transfer Project believe portability and interoperability are central to innovation. Making it easier for individuals to choose among services facilitates competition, empowers individuals to try new services and enables them to choose the offering that best suits their needs.

Current contributors include:

What is the Data Transfer Project

Data Transfer Project (DTP) is a collaboration of organizations committed to building a common framework with open-source code that can connect any two online service providers, enabling a seamless, direct, user initiated portability of data between the two platforms.

Learn More

https://datatransferproject.dev/

r/ai

Chapter 3

Rights of the data subject

Section 4	**–**	**Right to object and automated individual decision-making**
Article 21	–	Right to object
Article 22	–	Automated individual decision-making, including profiling

Recital 58

The principle of transparency*

[1] The principle of transparency requires that any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language and, additionally, where appropriate, visualisation be used. [2] Such information could be provided in electronic form, for example, when addressed to the public, through a website. [3] This is of particular relevance in situations where the proliferation of actors and the technological complexity of practice make it difficult for the data subject to know and understand whether, by whom and for what purpose personal data relating to him or her are being collected, such as in the case of online advertising. [4] Given that children merit specific protection, any information and communication, where processing is addressed to a child, should be in such a clear and plain language that the child can easily understand.

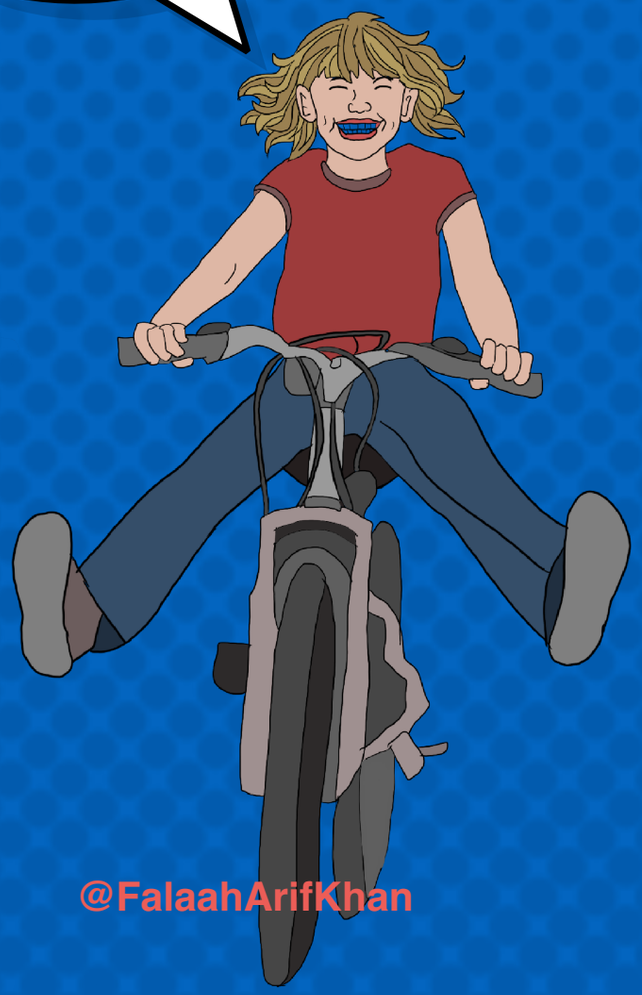
from data to impacts:
algorithmic impact
statements

r/ai

Regulating ADS?

Precautionary

Nah! I'm fine!



The Anti-Elon
@antiElon

Regulation rocks!

2.3K 9.2K 126K

Risk-based

Setting the stage: “Big Data Policing”

“Despite its growing popularity, predictive policing is in its relative infancy and is still mostly hype. Current prediction is akin to early weather forecasting, and, like Big Data approaches in other sectors, mixed evidence exists about its effectiveness.

Cities such as Los Angeles, Atlanta, Santa Cruz, and Seattle have enlisted the predictive policing software company PredPol to predict where property crimes will occur. Santa Cruz reportedly “saw burglaries drop by 11% and robberies by 27% in the first year of using [PredPol’s] software.” Similarly, Chicago’s Strategic Subject List—or “heat list”—of people most likely to be involved in a shooting had, as of mid-2016, predicted more than 70% of the people shot in the city, according to the police.

But two rigorous academic evaluations of predictive policing experiments, one in Chicago and another in Shreveport, have shown no benefit over traditional policing. **A great deal more study is required to measure both predictive policing’s benefits and its downsides**. “

what are the potential benefits?

what are the potential downsides?

[Andrew Selbst, 2017]

r/ai

How to regulate “Big Data Policing”

“While policing is just one of many aspects of society being upended by machine learning, and potentially exacerbating disparate impact in a hidden way as a result, it is a particularly useful case study because of how little our legal system is set up to regulate it.”

The Fourth Amendment: *The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.*

“[…] the Fourth Amendment’s reasonable suspicion requirement is inherently a “small data doctrine,” rendering it impotent in even its primary uses when it comes to data mining.”

new legal strategies are needed

[Andrew Selbst, 2017]

r/ai

How to regulate “Big Data Policing”

“ Regarding predictive policing specifically, society lacks basic knowledge and transparency about both the technology’s efficacy and its effects on vulnerable populations. Thus, this Article proposes a regulatory solution designed to fill this knowledge gap—to make the police do their homework and show it to the public before buying or building these technologies.”

Main contribution: Algorithmic Impact Statements (AISs)

“Impact statements are designed to **force consideration of the problem at an early stage**, and to document the process so that the public can learn what is at stake, **perhaps as a precursor to further regulation**. The primary problem is that no one, including the police using the technology, yet knows what the results of its use actually are.”

[Andrew Selbst, 2017]

r/ai

Algorithmic Impact Statements (AISs)

- Modeled on the Environmental Impact Statements (EISs) of the 1969 National Environmental Policy Act (NEPA)
- GDPR requires “data protection impact assessments (DPIAs) whenever data processing “is likely to result in a high risk to the rights and freedoms of natural persons”
- Privacy impact statements (PIAs) are used to assess the risks of using personally identifiable information by IT systems

The gist:

- Explore and evaluate all reasonable alternatives
- Include the alternative of “No Action”
- Include appropriate mitigation measures
- Provide opportunities for public comment

[Andrew Selbst, 2017]

r/ai

Canadian ADS directive

r/ai

Government of Canada | Gouvernement du Canada

Home → How government works → Policies, directives, standards and guidelines

Directive on Automated Decision-Making

The Government of Canada is increasingly looking to utilize artificial intelligence to make, or assist in making, administrative decisions to improve service delivery. The Government is committed to doing so in a manner that is compatible with core administrative law principles such as transparency, accountability, legality, and procedural fairness. Understanding that this technology is changing rapidly, this Directive will continue to evolve to ensure that it remains relevant.

Date modified: 2019-02-05

- Took effect on April 1, 2019, compliance by April 1, 2020
- Applies to any ADS developed or procured after April 1, 2020
- **Reviewed automatically every 6 months**

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Definitions

Appendix A: Definitions

- **Administrative Decision** Any decision that is made by an authorized official of an institution as identified in section 9 of this Directive pursuant to powers conferred by an Act of Parliament or an order made pursuant to a prerogative of the Crown that affects legal rights, privileges or interests.
- **Algorithmic Impact Assessment** A framework to help institutions better understand and reduce the risks associated with Automated Decision Systems and to provide the appropriate governance, oversight and reporting/audit requirements that best match the type of application being designed.
- **Automated Decision System** Includes any technology that either assists or replaces the judgement of human decision-makers. These systems draw from fields like statistics, linguistics, and computer science, and use techniques such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets.

Objectives

Section 4: Objectives and Expected Results

- **4.1** The objective of this Directive is to ensure that Automated Decision Systems are deployed in a manner that **reduces risks** to Canadians and federal institutions, and **leads to more efficient, accurate, consistent, and interpretable decisions** made pursuant to Canadian law.
- **4.2** The expected results of this Directive are as follows:
  - Decisions made by federal government departments are data-driven, responsible, and complies with procedural fairness and due process requirements.
  - Impacts of algorithms on administrative decisions are assessed and negative outcomes are reduced, when encountered.
  - Data and information on the use of Automated Decision Systems in federal institutions are made available to the public, when appropriate.

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Requirements

Section 6.1: Algorithmic Impact Assessment (excerpt)

- **6.1.1 Completing** an Algorithmic Impact Assessment **prior to the production** of any Automated Decision System.
- **6.1.2** …
- **6.1.3 Updating** the Algorithmic Impact Assessment when system functionality or the scope of the Automated Decision System changes.
- **6.1.4 Releasing the final results of Algorithmic Impact Assessments** in an accessible format via Government of Canada websites and any other services designated by the Treasury Board of Canada Secretariat pursuant to the Directive on Open Government.

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Requirements

Section 6.2: Transparency

- providing notice before decisions
- providing explanations after decisions
- access to components
- release of source code, unless it’s classified Secret, Top Secret or Protected C

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

Impact Assessment Levels

Decisions classified w.r.t. impact on:

- the rights of individuals or communities,
- the health or well-being of individuals or communities,
- the economic interests of individuals, entities, or communities,
- the ongoing sustainability of an ecosystem.

Level I: no impact: impacts are reversible and brief

Level II: moderate: impacts are likely reversible and short-term

Level III: high: impacts are difficult to reversible and ongoing

Level IV: very high: impacts are irreversible and perpetual

higher impact levels lead to more stringent requirements

https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

r/ai

so what's algorithmic transparency?

r/ai

Point 1

algorithmic transparency is not synonymous with releasing the source code

publishing source code helps, but it is sometimes unnecessary and often insufficient

Point 2

algorithmic transparency requires data transparency

data is used in training, validation, deployment

validity, accuracy, applicability can only be understood in the data context

data transparency is necessary for all ADS, not only for ML-based systems

Point 3

data transparency is not synonymous with making all data public

release data whenever possible;

also release:

data selection, collection and pre-processing methodologies; data provenance and quality information; known sources of bias; privacy-preserving statistical summaries of the data

r/ai

Point 3

data transparency is not synonymous with making all data public

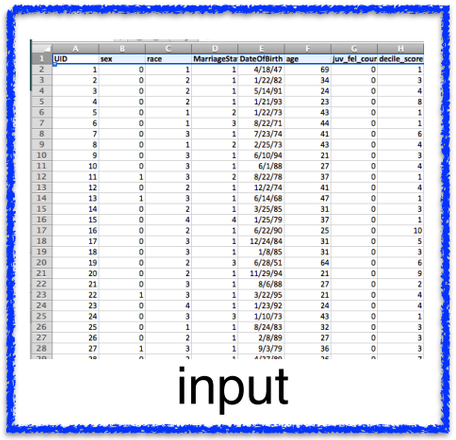
release data whenever possible;

also release:

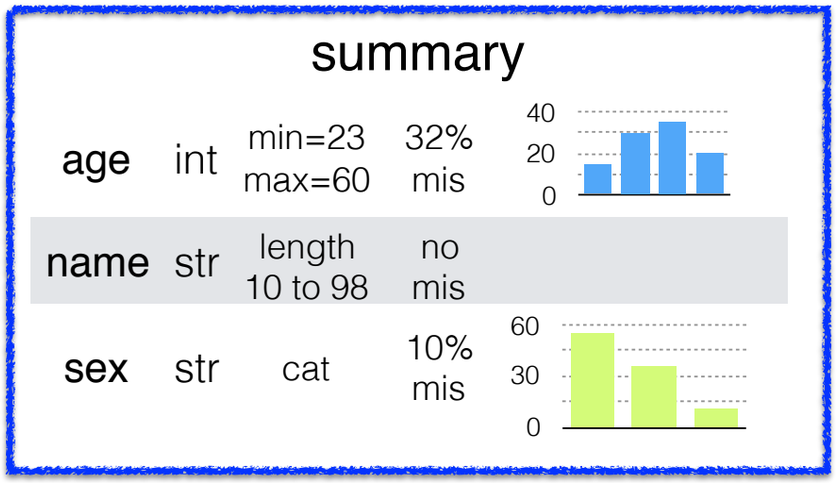
data selection, collection and pre-processing methodologies; data provenance and quality information; known sources of bias; privacy-preserving statistical summaries of the data

r/ai

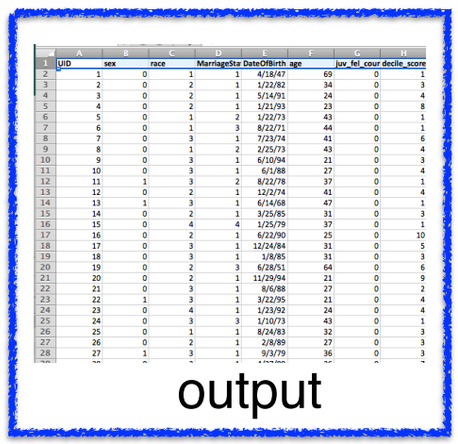
Data Synthesizer



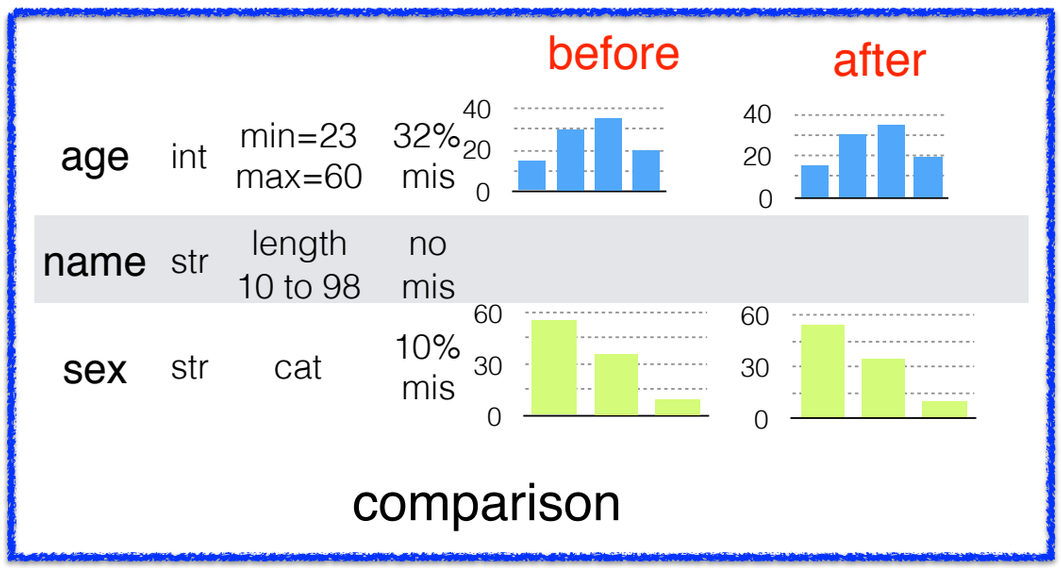
Data Describer



Data Generator



Model Inspector



[Ping, Stoyanovich, Howe *2017*] http://demo.dataresponsibly.com/synthesizer/

Point 4

actionable transparency requires
interpretability

explain assumptions and effects, not details of operation

engage the public - technical and non-technical

“Nutritional labels” for data and models

Ranking Facts

Recipe

Top 10:

Attribute	Maximum	Median	Minimum
PubCount	18.3	9.6	6.2
Faculty	122	52.5	45
GRE	800.0	796.3	771.9

Overall:

Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
Faculty	122	32.0	14
GRE	800.0	790.0	757.8

Recipe

Attribute	Weight
PubCount	1.0
Faculty	1.0
GRE	1.0

Ingredients

Attribute	Correlation
PubCount	1.0
CSRankingAllArea	0.24
Faculty	0.12

Correlation strength is based on its absolute value. Correlation over 0.75 is high, between 0.25 and 0.75 is medium, under 0.25 is low.

Ingredients

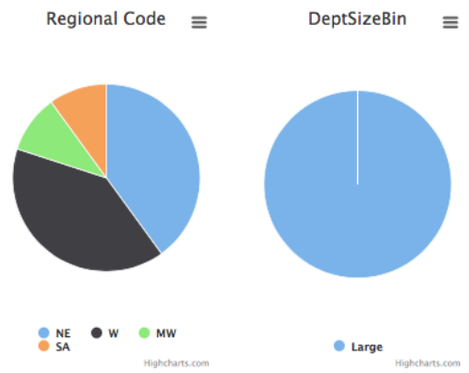
Top 10:

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

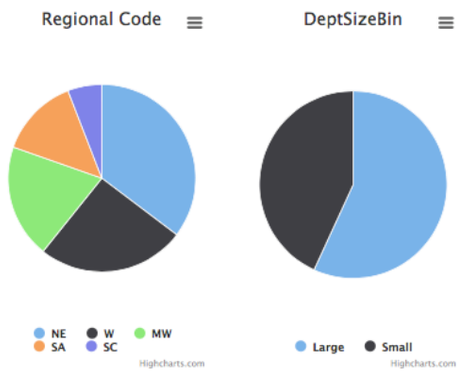
Overall:

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

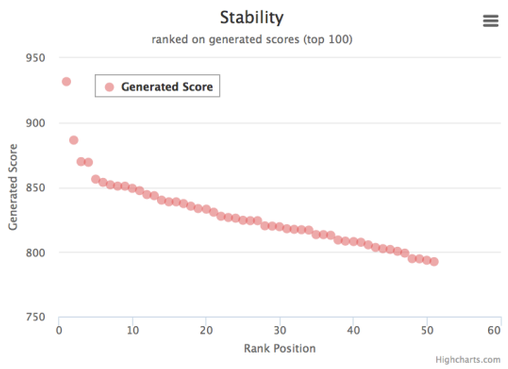
Diversity at top-10



Diversity overall



Stability



Slope at top-10: -6.91. Slope overall: -1.61.

Unstable when absolute value of slope of fit line in scatter plot <= 0.25 (slope threshold). Otherwise it is stable.

Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

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

Unfair when p-value of corresponding statistical test <= 0.05.

Fairness

DeptSizeBin	FA*IR p-value	FA*IR adjusted α	Pairwise p-value	Pairwise α	Proportion p-value	Proportion α
Large	1.0	0.87	0.99	0.05	1.0	0.05
Small	0.0	0.71	0.0	0.05	0.0	0.05

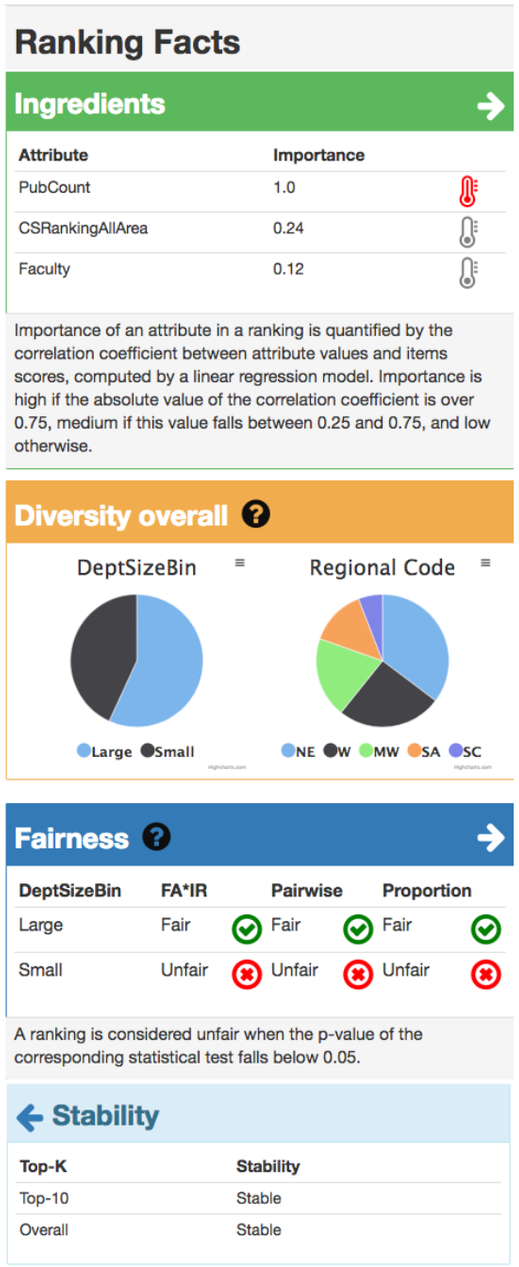
Top K = 26 in FA*IR and Proportion oracles. Setting of top K: In FA*IR and Proportion oracle, if N > 200, set top K =100. Otherwise set top K = 50%N. Pairwise oracle takes whole ranking as input. FA*IR is computed as using code in FA*IR codes. Proportion is implemented as statistical test 4.1.3 in Proportion paper.

http://demo.dataresponsibly.com/rankingfacts/nutrition_facts/

[K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, HV Jagadish, G. Miklau; 2018]

r/ai

Properties of a nutritional label



comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

concrete: helps determine a dataset's fitness for use for a given task

computable: produced as a "by-product" of computation - interpretability-by-design

[Stoyanovich and Howe, 2019]

r/ai

Point 5

transparency / interpretability by design, not as an afterthought

provision for transparency and interpretability at every stage of the data lifecycle

useful internally during development, for communication and coordination between agencies, and for accountability to the public

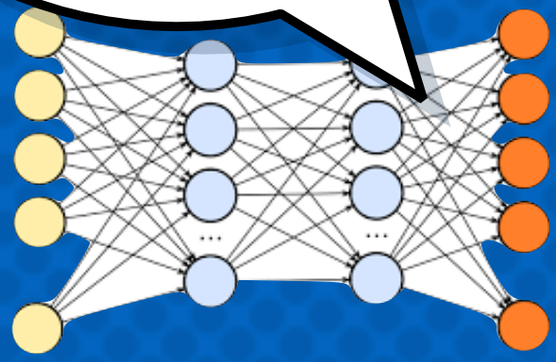
r/ai

Frog's eye view

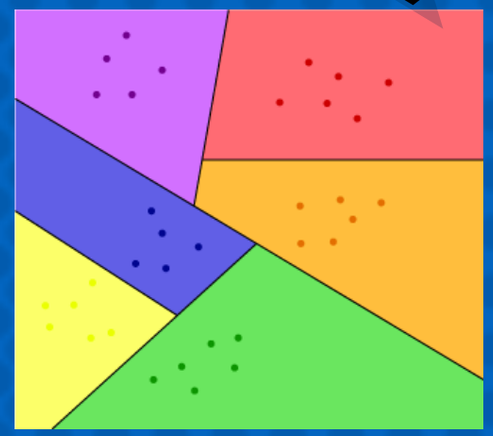
where did the data come from?

	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageSta	DateOfBirth	age	juv_fel_cour	decile_score
2	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3

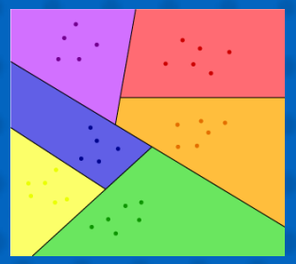
what happens inside the box?



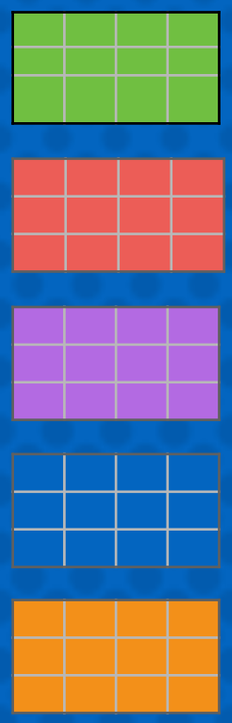
how are results used?



Data lifecycle of an ADS

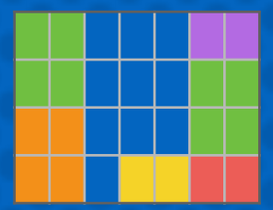


analysis
validation

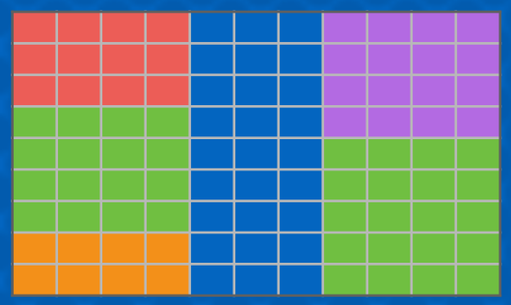


sharing
annotation

querying
ranking



acquisition
curation



interpretability in the eye of the stakeholder

r/ai

What are we explaining?

process (same for everyone? why is this the process?) vs. outcome

procedural justice aims to ensure that algorithms are perceived as fair and legitimate

data transparency is unique to algorithm-assisted decision-making, relates to the justification dimension of interpretability

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

To whom are we explaining and why?

accounting for the needs of different stakeholders

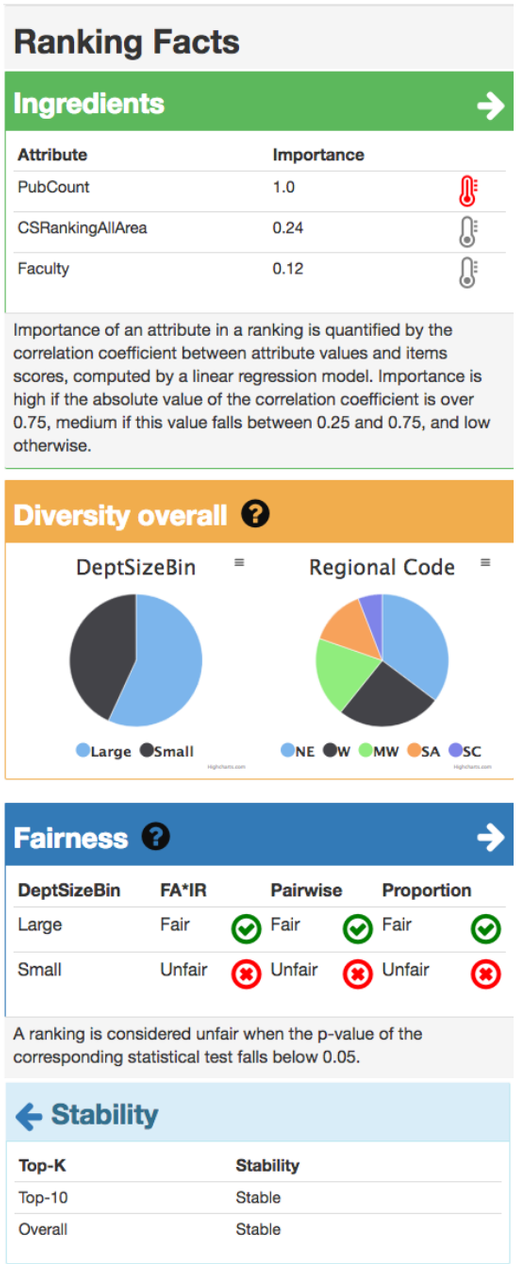
social identity - people trust their in-group members more

moral cognition - is a decision or outcome morally right or wrong?

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

How do we know that we explained well?



nutritional labels! :)

… but do they work?

[J. Stoyanovich, J. Van Bavel, T. West, 2020]

r/ai

regulating automated hiring systems

r/ai

Regulating hiring ADS: Int 1894-2020

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

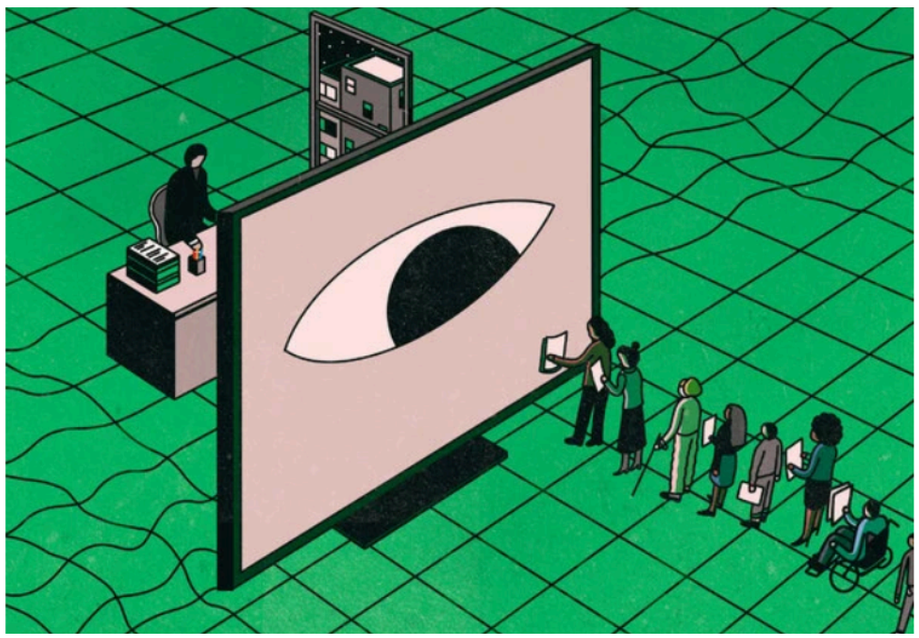
This bill would **regulate the use of automated employment decision tools**, which, for the purposes of this bill, encompass certain systems that use algorithmic methodologies to filter candidates for hire or to make decisions regarding any other term, condition or privilege of employment. This bill would prohibit the sale of such tools if they were not the **subject of an audit for bias** in the past year prior to sale, were not sold with a yearly bias audit service at no additional cost, and were not accompanied by a notice that the tool is subject to the provisions of this bill. This bill would also require any person who uses automated employment assessment tools for hiring and other employment purposes to **disclose to candidates, within 30 days, when such tools were used** to assess their candidacy for employment, and the **job qualifications or characteristics** for which the tool was used to screen. Violations of the provisions of the bill would incur a penalty.

Hiring ADS regulation

The New York Times March 17, 2021

We Need Laws to Take On Racism and Sexism in Hiring Technology

Artificial intelligence used to evaluate job candidates must not become a tool that exacerbates discrimination.



By Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich

Ms. Givens is the chief executive of the Center for Democracy & Technology. Ms. Schellman and Dr. Stoyanovich are professors at New York University focusing on artificial intelligence.

https://www.nytimes.com/2021/03/17/opinion/ai-employment-bias-nyc.html

The measure must require companies to **publicly disclose what they find when they audit their tech for bias**. Despite pressure to limit its scope, the City Council must ensure that the bill would address discrimination in all forms — on the basis of not only race or gender but also disability, sexual orientation and other protected characteristics.

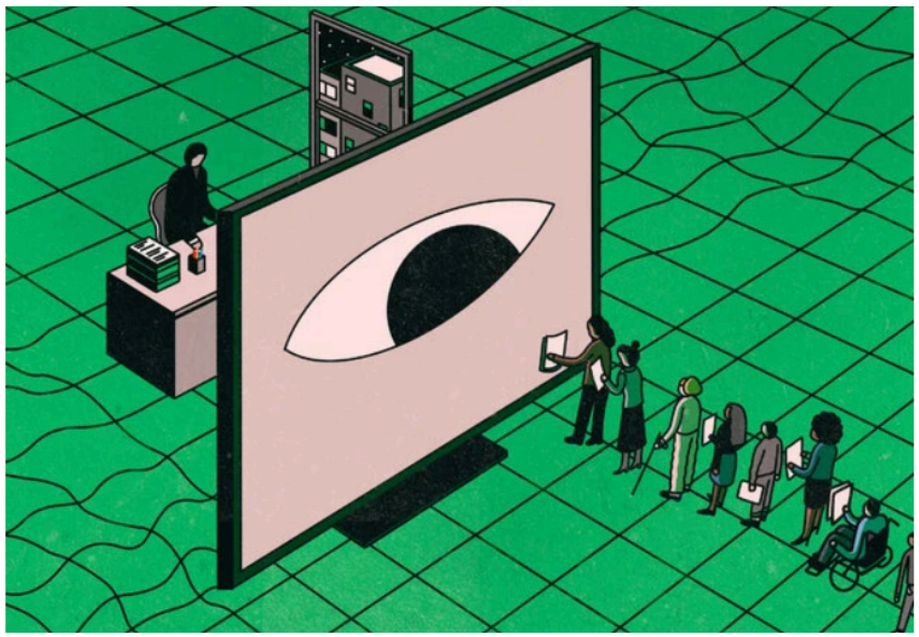
These audits should consider the circumstances of **people who are multiply marginalized** — for example, Black women, who may be discriminated against because they are both Black and women. Bias audits conducted by companies typically don't do this.

Hiring ADS regulation

The New York Times March 17, 2021

We Need Laws to Take On Racism and Sexism in Hiring Technology

Artificial intelligence used to evaluate job candidates must not become a tool that exacerbates discrimination.



By Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich

Ms. Givens is the chief executive of the Center for Democracy & Technology. Ms. Schellman and Dr. Stoyanovich are professors at New York University focusing on artificial intelligence.

https://www.nytimes.com/2021/03/17/opinion/ai-employment-bias-nyc.html

The bill should […] require validity testing, to **ensure that the tools actually measure what they claim to**, and it must make certain **that they measure characteristics that are relevant for the job**. Such testing would interrogate whether, for example, candidates' efforts to blow up a balloon in an online game really indicate their appetite for risk in the real world — and whether risk-taking is necessary for the job.

… [T]he City Council must require vendors to tell candidates how they will be screened by an automated tool **before** the screening, so candidates know what to expect. People who are blind, for example, may not suspect that their video interview could score poorly if they fail to make eye contact with the camera. If they know what is being tested, they can engage with the employer to seek a fairer test.

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected

Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich

Updated Sept. 22, 2021 11:00 am ET

Artificial-intelligence tools are seeing ever broader use in hiring. But this practice is also hotly criticized because we rarely understand how these tools select candidates, and whether the candidates they select are, in fact, better qualified than those who are rejected.

To help answer these crucial questions, **we should give job seekers more information about the hiring process and the decisions**. The solution I propose is a twist on something we see every day: **nutritional labels**. Specifically, job candidates would see simple, standardized labels that show the factors that go into the AI's decision.

https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

r/ai

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected

Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich

Updated Sept. 22, 2021 11:00 am ET

ACCOUNTANT

Acme Partners

Qualifications:	BS in accounting, GPA >3.0, Knowledge of financial and accounting systems and applications
Personal data to be analyzed:	An AI program could be used to review and analyze the applicant's personal data online, including LinkedIn profile, social media accounts and credit score.
Additional assessment:	AI-assisted personality scoring

ALERT: Applicants for this position DO NOT have the option to selectively decline use of AI analysis for any of their personal data or to review and challenge the results of such analysis.

https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313

r/ai

New York City Local Law 144 of 2021

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

December 11, 2021

This bill would require that a **bias audit** be conducted on an automated employment decision tool prior to the use of said tool. The bill would also require that candidates or employees that reside in the city **be notified about the use of such tools** in the assessment or evaluation for hire or promotion, as well as, **be notified about the job qualifications and characteristics that will be used** by the automated employment decision tool. Violations of the provisions of the bill would be subject to a civil penalty.