Algorithmic and Data Transparency in NYC Agencies: Tools and Strategies

Julia Stoyanovich
Drexel University & Princeton CITP

data RESPONSIBLY

Outline

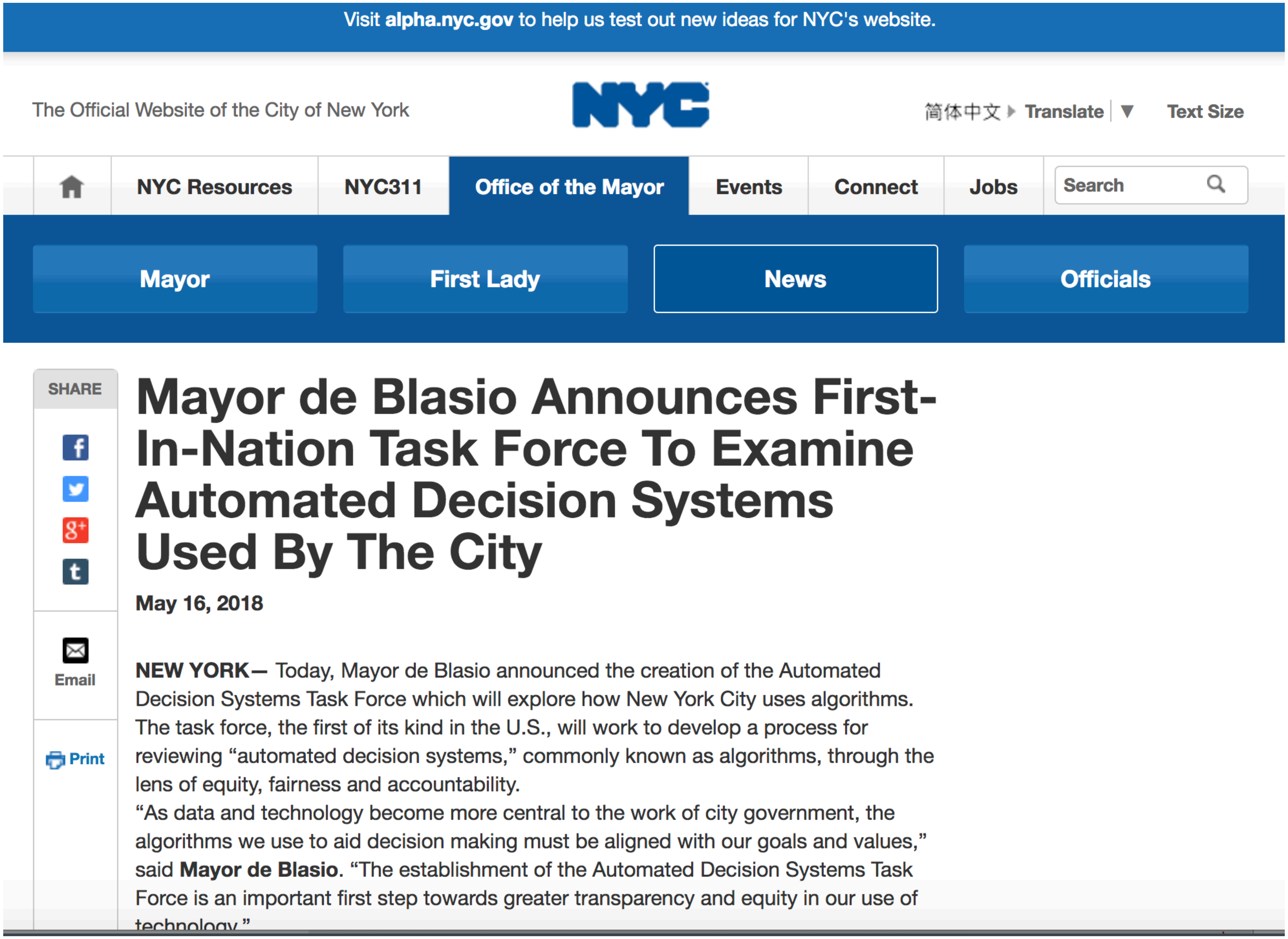
- Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies
- comments on the Law
- strategies for success

Summary of Int. No. 1696-A

Form an automated decision systems (**ADS**) task force that surveys current use of algorithms and data in City agencies and develops procedures for:

- requesting and receiving an explanation of an algorithmic decision affecting an individual (3(b))
- interrogating ADS for bias and discrimination against members of legally-protected groups (3(c) and 3(d))
- allowing the public to assess how ADS function and are used (3(e)), and archiving ADS together with the data they use (3(f))

The ADS Task Force


Visit alpha.nyc.gov to help us test out new ideas for NYC's website.
The Official Website of the City of New York
NYC
简体中文 Translate ▼ Text Size
NYC Resources
NYC311
Office of the Mayor
Events
Connect
Jobs
Search
Mayor
First Lady
News
Officials
SHARE
Email
Print

Mayor de Blasio Announces First-In-Nation Task Force To Examine Automated Decision Systems Used By The City

May 16, 2018

NEW YORK— Today, Mayor de Blasio announced the creation of the Automated Decision Systems Task Force which will explore how New York City uses algorithms. The task force, the first of its kind in the U.S., will work to develop a process for reviewing “automated decision systems,” commonly known as algorithms, through the lens of equity, fairness and accountability.

“As data and technology become more central to the work of city government, the algorithms we use to aid decision making must be aligned with our goals and values,” said **Mayor de Blasio**. “The establishment of the Automated Decision Systems Task Force is an important first step towards greater transparency and equity in our use of technology.”

Point 1

algorithmic transparency is not synonymous with releasing the source code

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

syntactic vs. semantic transparency

the interplay between code and data

Point 2

algorithmic transparency requires data transparency

data is used in training, validation, deployment

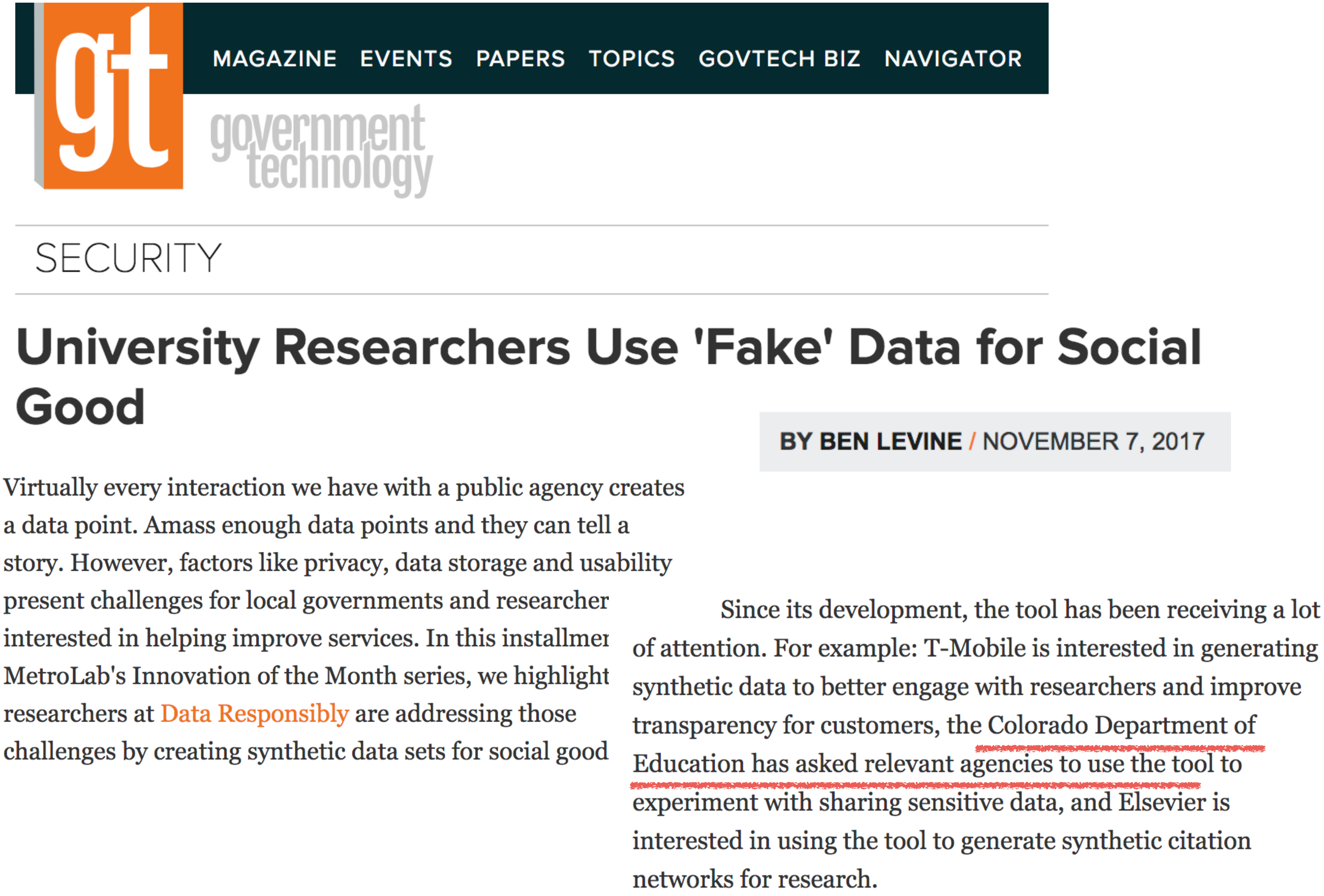
validity, accuracy, applicability can only be understood in the data context

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; dataset composition, statistical properties, sources of bias; validation methodologies



SECURITY

University Researchers Use 'Fake' Data for Social Good

BY BEN LEVINE / NOVEMBER 7, 2017

Virtually every interaction we have with a public agency creates a data point. Amass enough data points and they can tell a story. However, factors like privacy, data storage and usability present challenges for local governments and researcher interested in helping improve services. In this installmen MetroLab's Innovation of the Month series, we highlight researchers at Data Responsibly are addressing those challenges by creating synthetic data sets for social good

Since its development, the tool has been receiving a lot of attention. For example: T-Mobile is interested in generating synthetic data to better engage with researchers and improve transparency for customers, the Colorado Department of Education has asked relevant agencies to use the tool to experiment with sharing sensitive data, and Elsevier is interested in using the tool to generate synthetic citation networks for research.

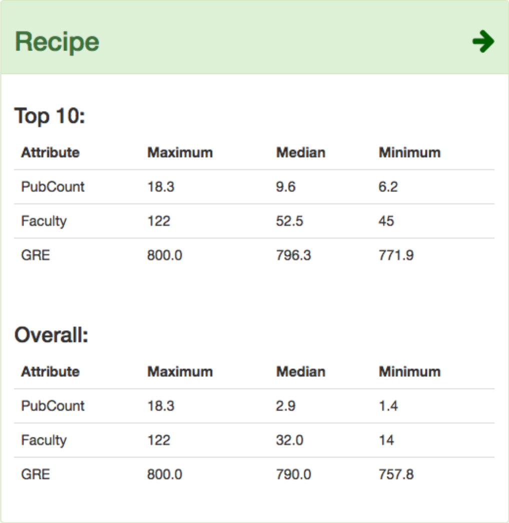
http://www.govtech.com/security/University-Researchers-Use-Fake-Data-for-Social-Good.html

Point 4

actionable transparency requires interpretability

explain assumptions and effects, not details of operation

engage the public - technical and non-technical



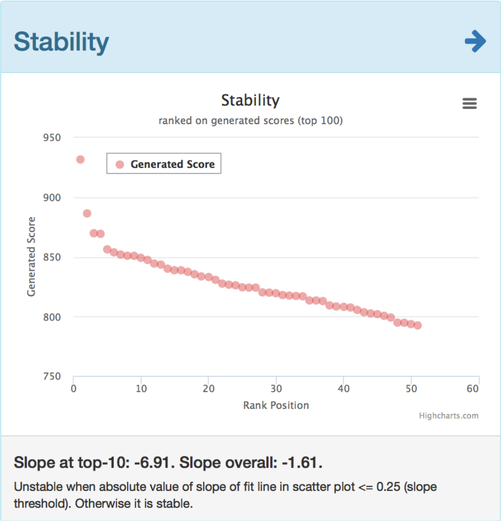
Recipe

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:

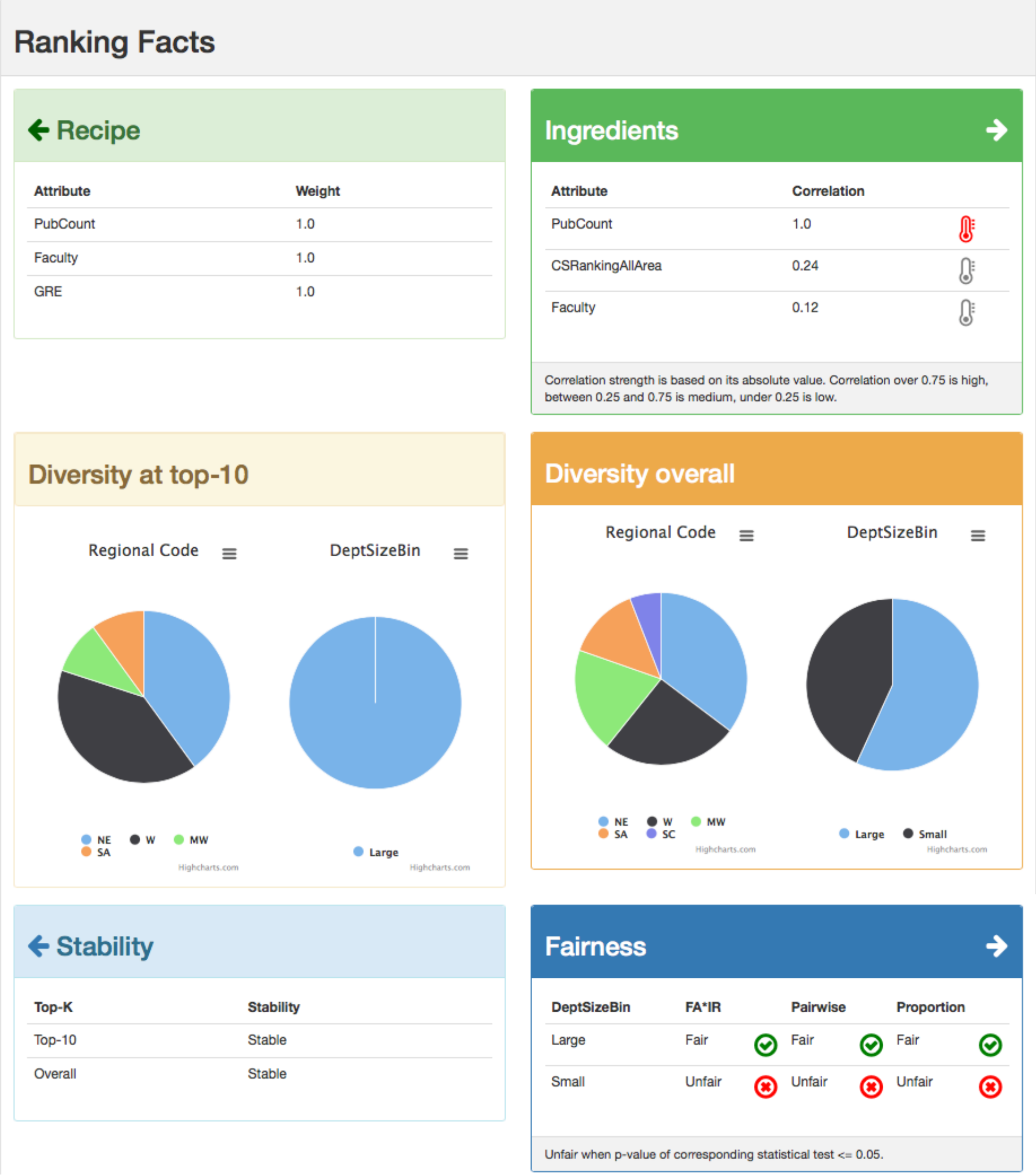
Attribute	Maximum	Median	Minimum
PubCount	18.3	2.9	1.4
Faculty	122	32.0	14
GRE	800.0	790.0	757.8



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.



Ranking Facts

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.

Diversity at top-10

Diversity overall

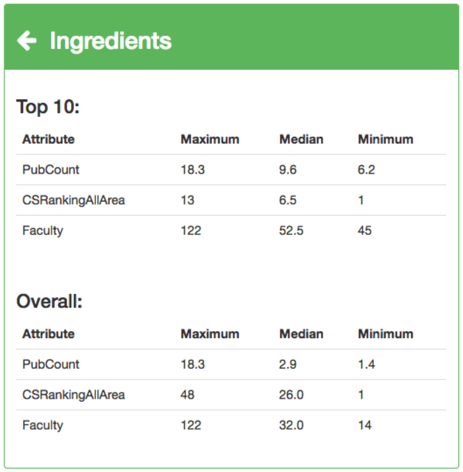
Stability

Top-K	Stability
Top-10	Stable
Overall	Stable

Fairness

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

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



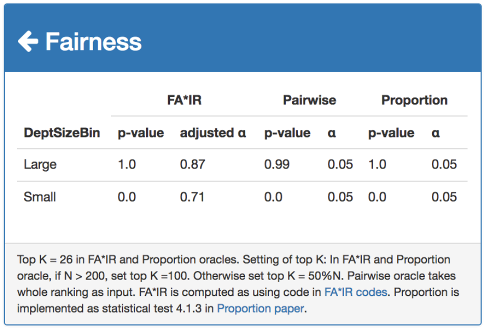
Ingredients

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



Fairness

	FA*IR		Pairwise		Proportion	
DeptSizeBin	p-value	adjusted α	p-value	α	p-value	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/

data RESPONSIBLY

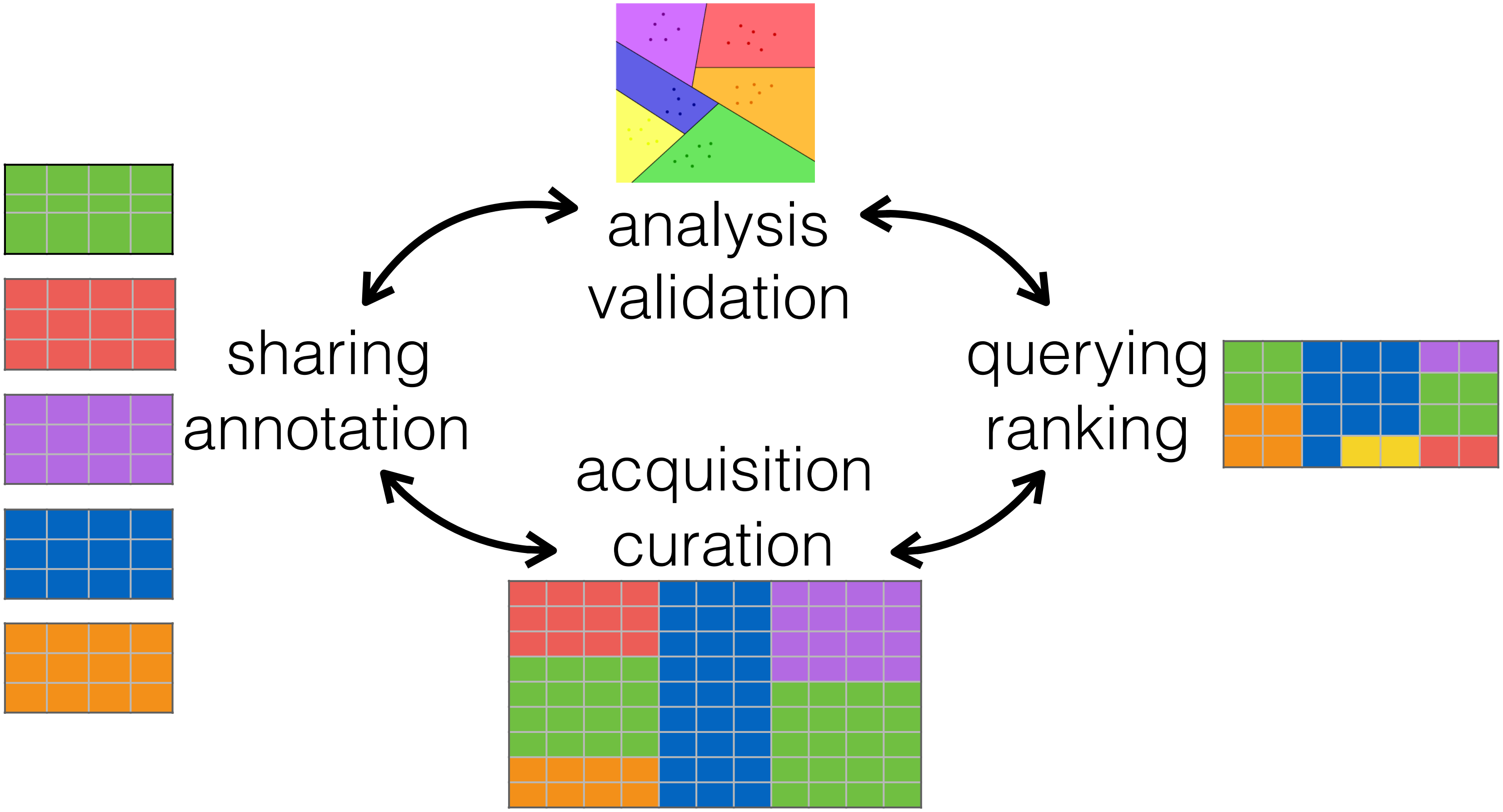
Point 5

transparency 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

The data science lifecycle



analysis
validation

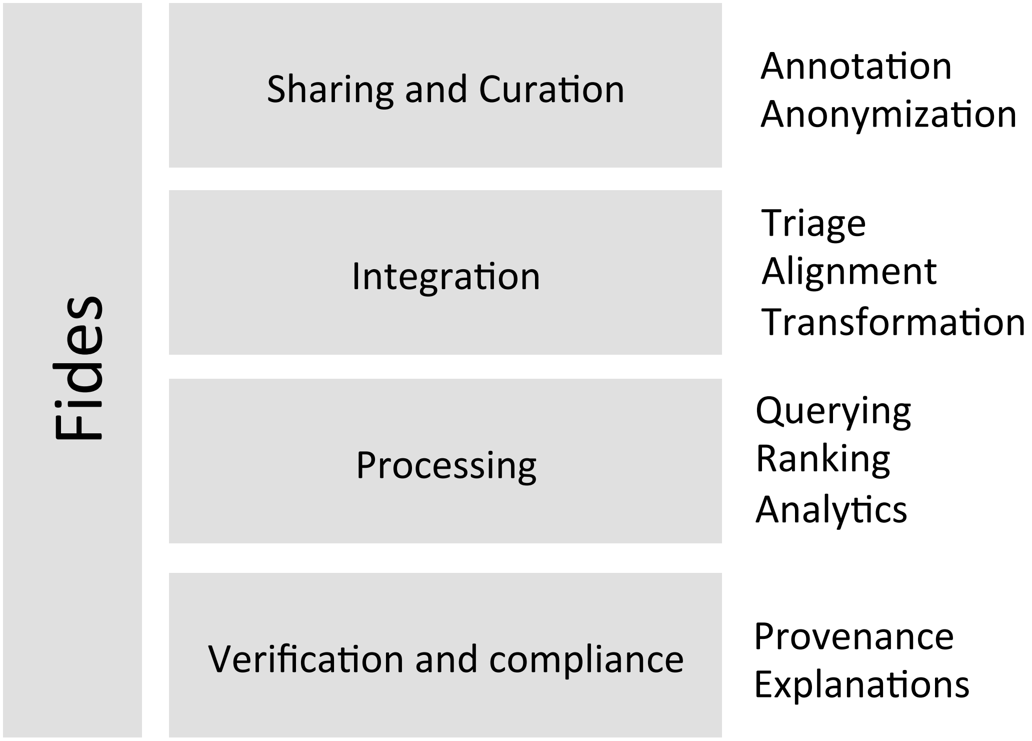
sharing
annotation

querying
ranking

acquisition
curation

responsible data science requires a holistic view of the data lifecycle

Responsibility by design



Fides

Sharing and Curation: Annotation, Anonymization

Integration: Triage, Alignment, Transformation

Processing: Querying, Ranking, Analytics

Verification and compliance: Provenance, Explanations

Systems support for responsible data science

Responsibility by design, managed at all stages of the lifecycle of data-intensive applications

responsible data science requires a holistic view of the data lifecycle

Stoyanovich, Howe, Abiteboul, Miklau, Sahuguet, Weikum - *SSDBM 2017*

Point 6

transparency is a challenge and an opportunity

lots of ongoing research, but not a solved problem

will require time and resources to get right - we need all hands on deck

the GDPR is drawing tremendous technological investment in the EU, the NYC algorithmic transparency law should be our opportunity

Strategies

build on NYC Open Data Law

leverage public engagement

leverage the research community

learn from others