Follow the data!

Algorithms and systems for responsible data science

Julia Stoyanovich
Drexel University & Princeton CITP

data RESPONSIBLY

NYC Algorithmic Transparency Law

1/11/2018

Int. No. 1696-A: A Local Law in relation to automated decision systems used by agencies

THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

Sign In

LEGISLATIVE RESEARCH CENTER

Council Home | Legislation | Calendar | City Council | Committees

RSS | Alerts

Details | Reports

File #:	Int 1696-2017 Version: A	Name:	Automated decision systems used by agencies.
Type:	Introduction	Status:	Enacted
		Committee:	Committee on Technology
On agenda:	8/24/2017		
Enactment date:	1/11/2018	Law number:	2018/049
Title:	A Local Law in relation to automated decision systems used by agencies		
Sponsors:	James Vacca, Helen K. Rosenthal, Corey D. Johnson, Rafael Salamanca, Jr., Vincent J. Gentile, Robert E. Cornegy, Jr., Jumaane D. Williams, Ben Kallos, Carlos Menchaca		
Council Member Sponsors:	9		
Summary:	This bill would require the creation of a task force that provides recommendations on how information on agency automated decision systems may be shared with the public and how agencies may address instances where people are harmed by agency automated decision systems.		
Indexes:	Oversight		
Attachments:	1. Summary of Int. No. 1696-A, 2. Summary of Int. No. 1696, 3. Int. No. 1696, 4. August 24, 2017 - Stated Meeting Agenda with Links to Files, 5. Committee Report 10/16/17, 6. Hearing Testimony 10/16/17, 7. Hearing Transcript 10/16/17, 8. Proposed Int. No. 1696-A - 12/12/17, 9. Committee Report 12/7/17, 10. Hearing Transcript 12/7/17, 11. December 11, 2017 - Stated Meeting Agenda with Links to Files, 12. Hearing Transcript - Stated Meeting 12-11-17, 13. Int. No. 1696-A (FINAL), 14. Fiscal Impact Statement, 15. Legislative Documents - Letter to the Mayor, 16. Local Law 49, 17. Minutes of the Stated Meeting - December 11, 2017		

data RESPONSIBLY

NYC Algorithmic Transparency Law

10/16/2017

THE NEW YORKER

By Julia Powles December 20, 2017

ELEMENTS

NEW YORK CITY'S BOLD, FLAWED ATTEMPT TO MAKE ALGORITHMS ACCOUNTABLE

Automated systems guide the allocation of everything from firehouses to food stamps. So why don't we know more about them?

Photograph by Mario Tama / Getty

data RESPONSIBLY

The original draft

8/16/2017

Int. No. 1696

By Council Member Vacca

A Local Law to amend the administrative code of the city of New York, in relation to automated processing of data for the purposes of targeting services, penalties, or policing to persons

Be it enacted by the Council as follows:

1 Section 1. Section 23-502 of the administrative code of the city of New York is amended
2 to add a new subdivision g to read as follows:
3 g. Each agency that uses, for the purposes of targeting services to persons, imposing
4 penalties upon persons or policing, an algorithm or any other method of automated processing
5 system of data shall:
6 1. Publish on such agency's website, the source code of such system; and
7 2. Permit a user to (i) submit data into such system for self-testing and (ii) receive the
8 results of having such data processed by such system.
9 § 2. This local law takes effect 120 days after it becomes law.

MAJ
LS# 10948
8/16/17 2:13 PM

this is **NOT** what was adopted

data RESPONSIBLY

Summary of Int. No. 1696-A

1/11/2018

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))

we've come a long way from the original draft!

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

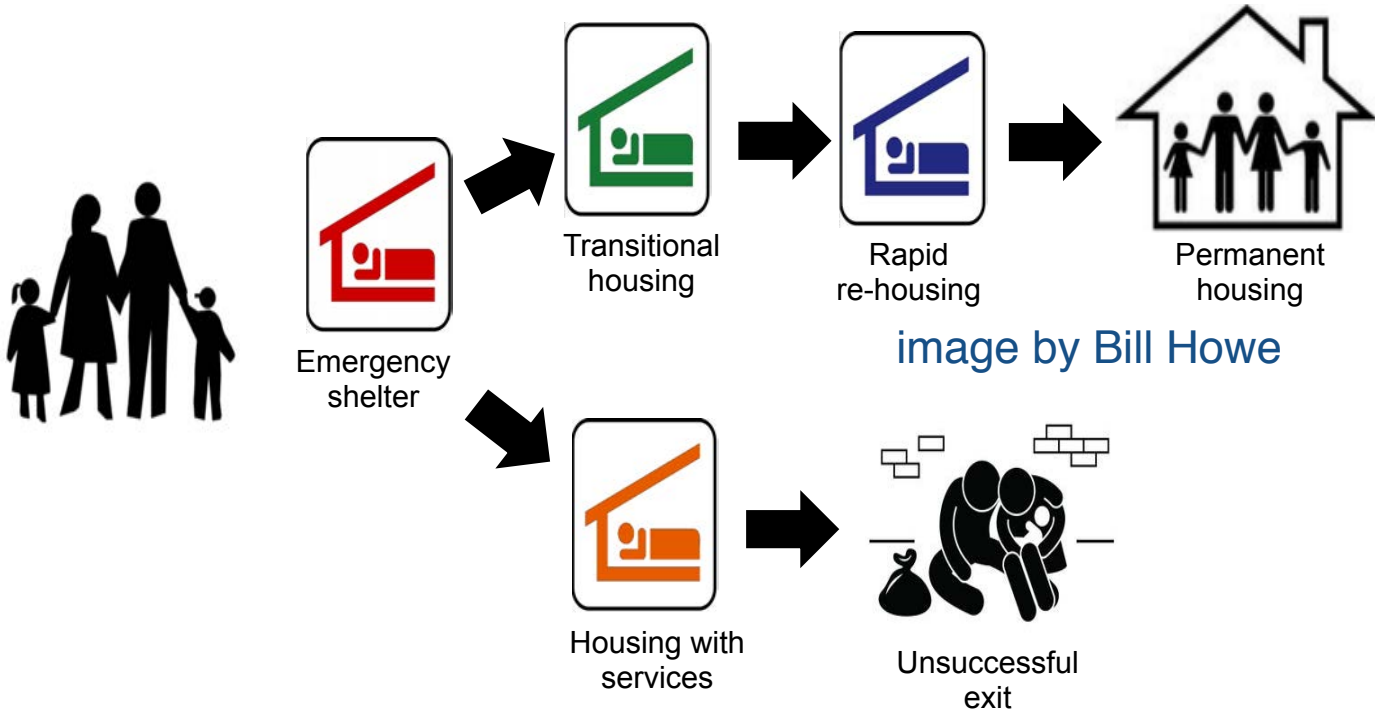
May 16, 2018

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."

data RESPONSIBLY

ADS example: urban homelessness



- **Allocate** interventions: services and support mechanisms
- **Recommend** pathways through the system
- **Evaluate** effectiveness of interventions, pathways, over-all system

Mayor de Blasio Scrambles to Curb Homelessness After Years of Not Keeping Pace

By J. DAVID GOODMAN and NIKITA STEWART JAN. 13, 2017

Volunteers during the homeless census in February 2015. In a decision made by Mayor Bill de Blasio, New York City stopped opening shelters for much of that year. Stephanie Keith for The New York Times

The New York Times

https://www.nytimes.com/2017/01/13/nyregion/mayor-de-blasio-scrambles-to-curb-homelessness-after-years-of-not-keeping-pace.html

Ms. Glen emphasized that the construction of new housing takes several years, a long-term solution whose effect on homelessness could not yet be evaluated.

Homeless Young People of New York, Overlooked and Underserved

By NIKITA STEWART FEB. 5, 2016

Abdul, 23, at Safe Horizon in Harlem, has been homeless since 2010. Jake Naughto

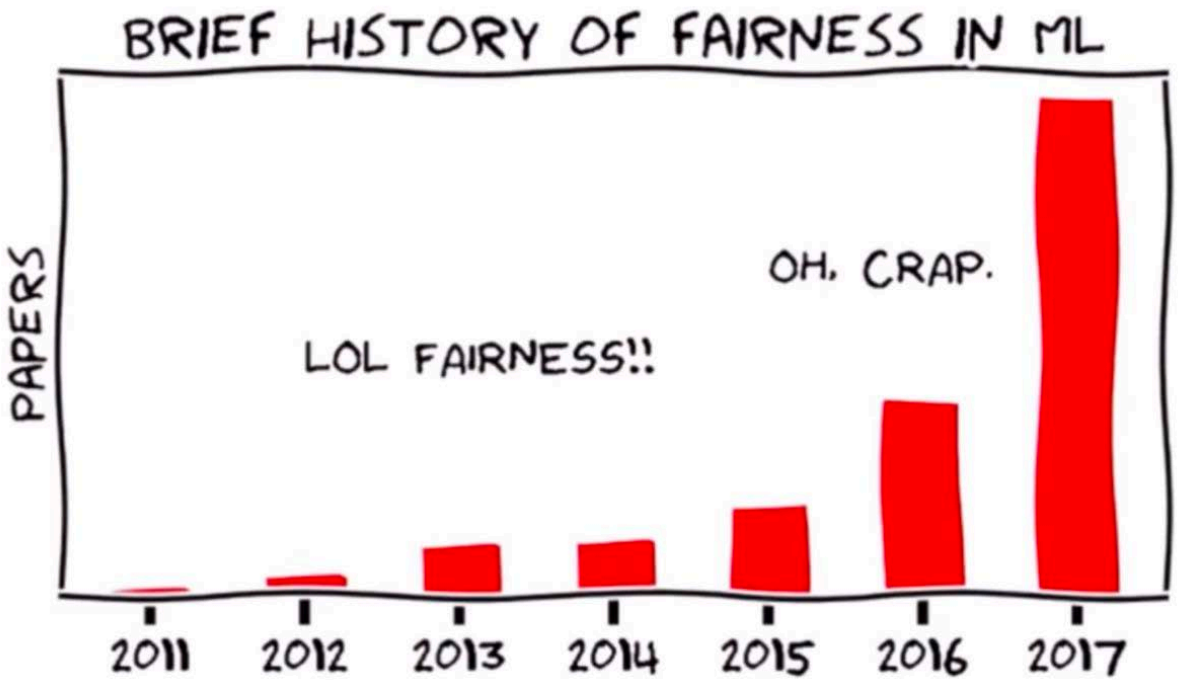
The New York Times

https://www.nytimes.com/2016/02/06/nyregion/young-and-homeless-in-new-york-overlooked-and-underserved.html

Last year, the total number of sheltered and unsheltered homeless people in the city was 75,323, which included 1,706 people between ages 18 and 24. The actual number of young people is significantly higher, according to the service providers, who said the census mostly captured young people who received social services. The census takers were not allowed to enter private businesses, including many of the late-night spots where young people often create an ad hoc shelter by pretending to be customers.

Responsible data science

- Be **transparent** and **accountable**
- Achieve **equitable** resource distribution
- Be cognizant of the **rights** and **preferences** of individuals

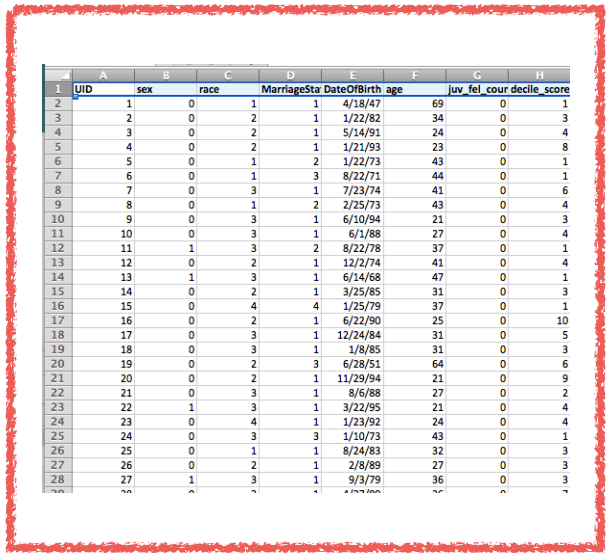


by Moritz Hardt

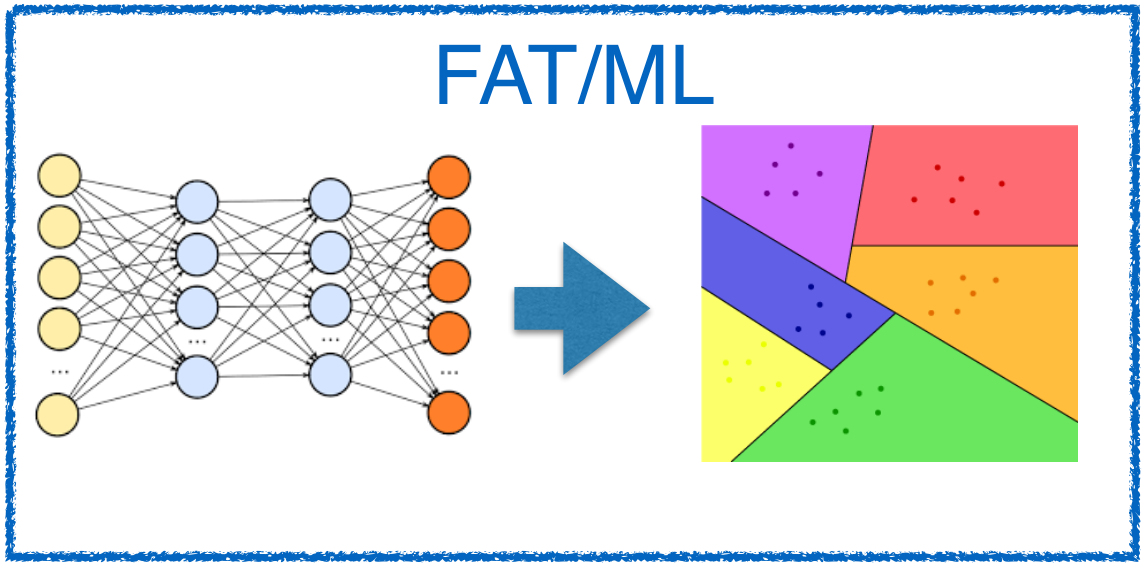
data RESPONSIBLY

Responsible data science

- Be **transparent** and **accountable**
- Achieve **equitable** resource distribution
- Be cognizant of the **rights** and **preferences** of individuals



	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



done?

but where does the data come from?

data RESPONSIBLY

Responsible data science

- Be **transparent** and **accountable**
- Achieve **equitable** resource distribution
- Be cognizant of the **rights** and **preferences** of individuals

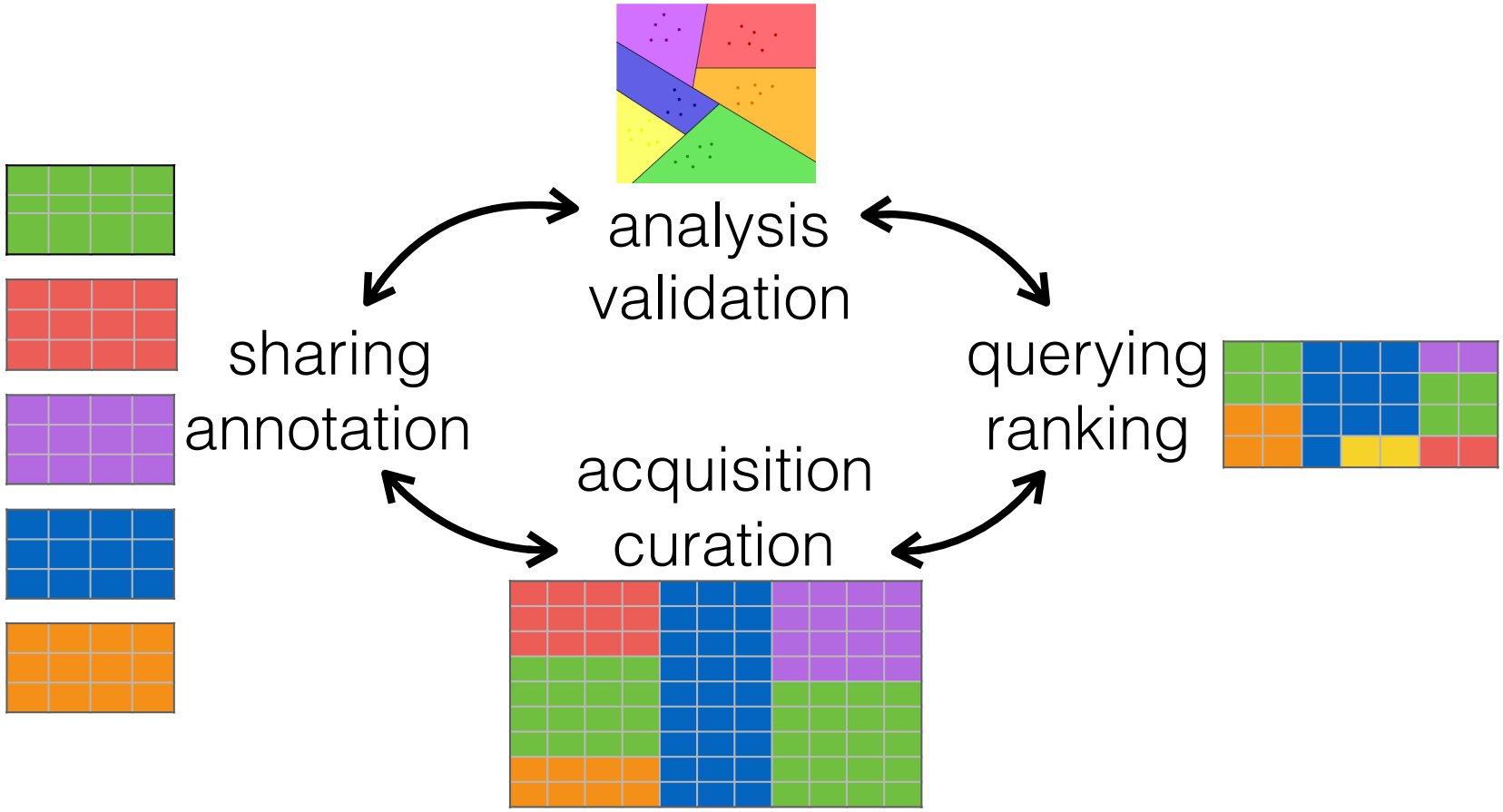
fairness

diversity

transparency

data protection

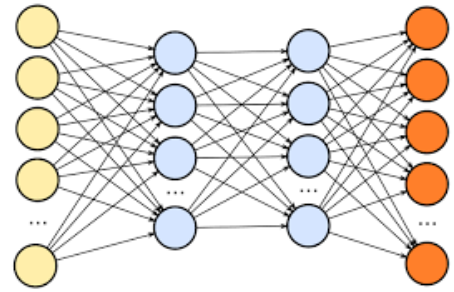
The data science lifecycle

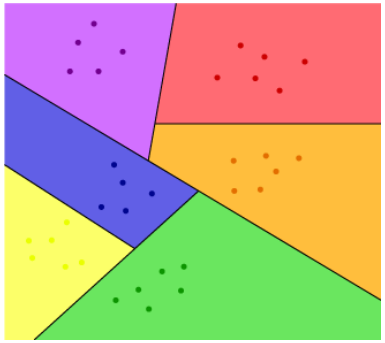


responsible data science requires a holistic view of the data lifecycle

Revisiting the analytics step

	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





finding: women are underrepresented in some outcome groups (group fairness)

fix the model!

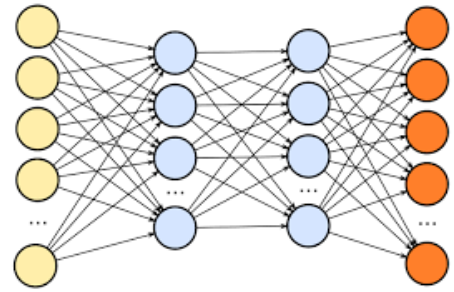
of course, but maybe… the input was generated with:

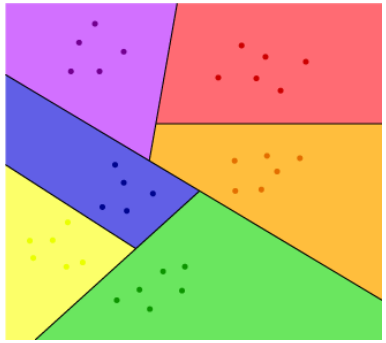
select * from R
where status = ‘unsheltered’
and length > 2 month

10% female

Revisiting the analytics step

	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





finding: women are underrepresented in some outcome groups (group fairness)

fix the model!

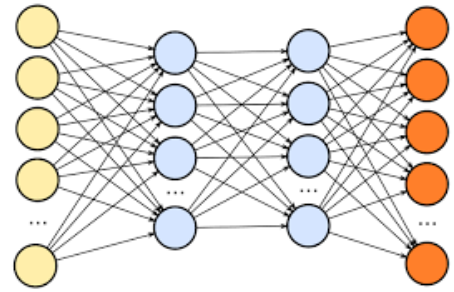
of course, but maybe… the input was generated with:

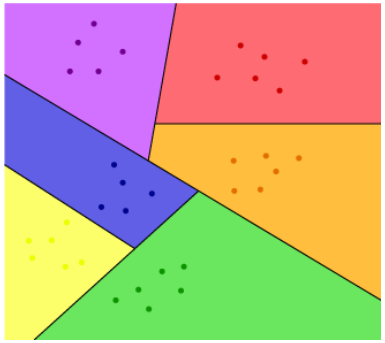
select * from R
where status = ‘unsheltered’
and length > 1 month

40% female

Revisiting the analytics step

	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





finding: young people are recommended pathways of lower effectiveness (high error rate)

fix the model!

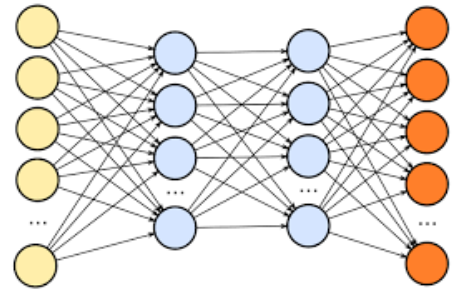
of course, but maybe…

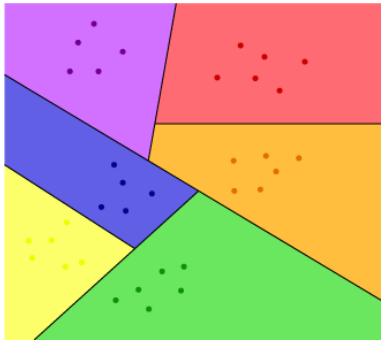
mental health info was missing for this population

go back to the data acquisition step, look for additional datasets

Revisiting the analytics step

	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





finding: minors are underrepresented in the input, compared to their actual proportion in the population (insufficient data)

unlikely to help!

fix the model??

minors data was not shared

go back to the data sharing step, help data providers share their data while adhering to laws and upholding the trust of the participants

Fides: responsibility by design

[BIGDATA] Foundations of responsible data management 09/2017-

Fides: 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

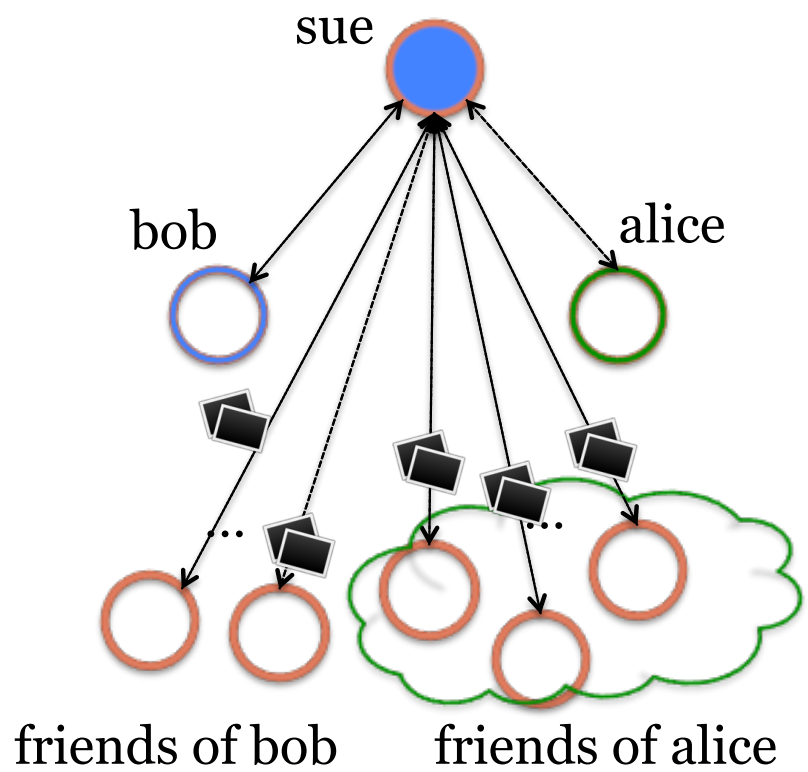
Responsibility by design, managed at all stages of the lifecycle of data-intensive applications

Applications: data science for social good

responsible data science requires a holistic view of the data lifecycle

Collaborative access control

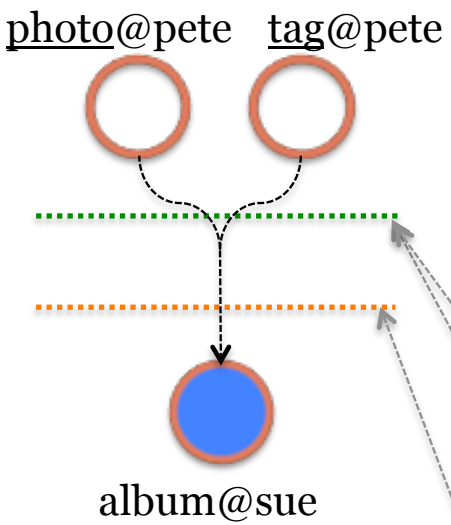
- Data owner specifies **access control** annotations on the **base relations**
- The system **automatically propagates** these annotations from base relations to views
- Based on **fine-grained provenance techniques** - because we know the data and the process!
- The environment: distributed datalog with delegation
- Implemented in a **system**, demonstrates that the overhead of access control is modest!



joint with Moffitt [Drexel], Abiteboul [INRIA], Miklau [UMass] - [SIGMOD 2015]

Collaborative access control

```
[at sue] album@sue($ph, pete) :- photo@pete($ph),
                                 tag@pete($ph, alice), tag@pete($ph, bob)
```



photo@pete(fname)
wildparty
awww

tag@pete(pic, name)	
wildparty	alice
wildparty	bob
wildparty	pete
wildparty	sue
awww	pete

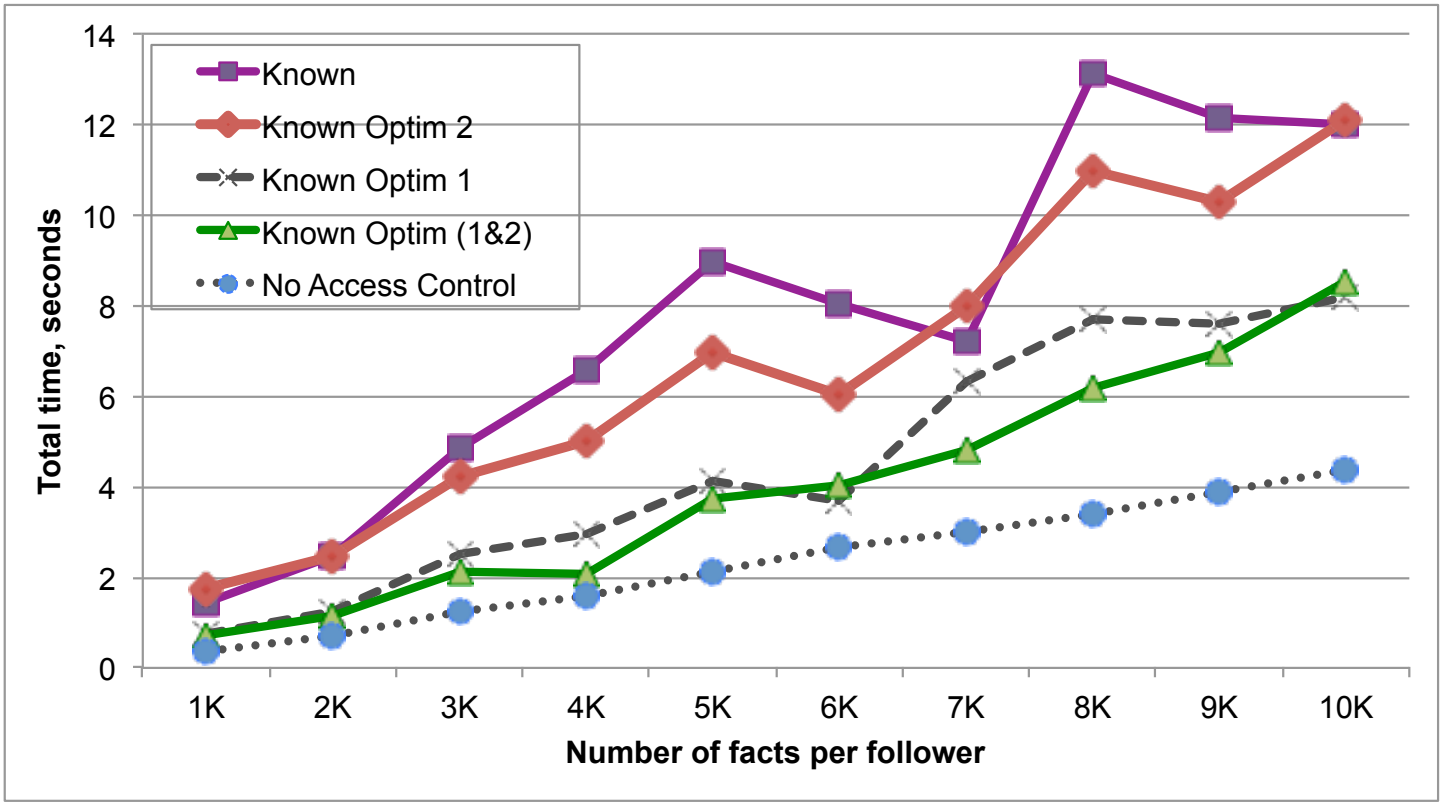
acl@pete(rel, pset, priv)		
photo	{alice, bob, pete, sue}	READ
tag	Ω	READ

acl@sue(rel, pset, priv)		
album	{sue}	WRITE

album+@sue(pic, source,pset,priv)			
wildparty	pete	{alice, bob, pete, sue}	READ

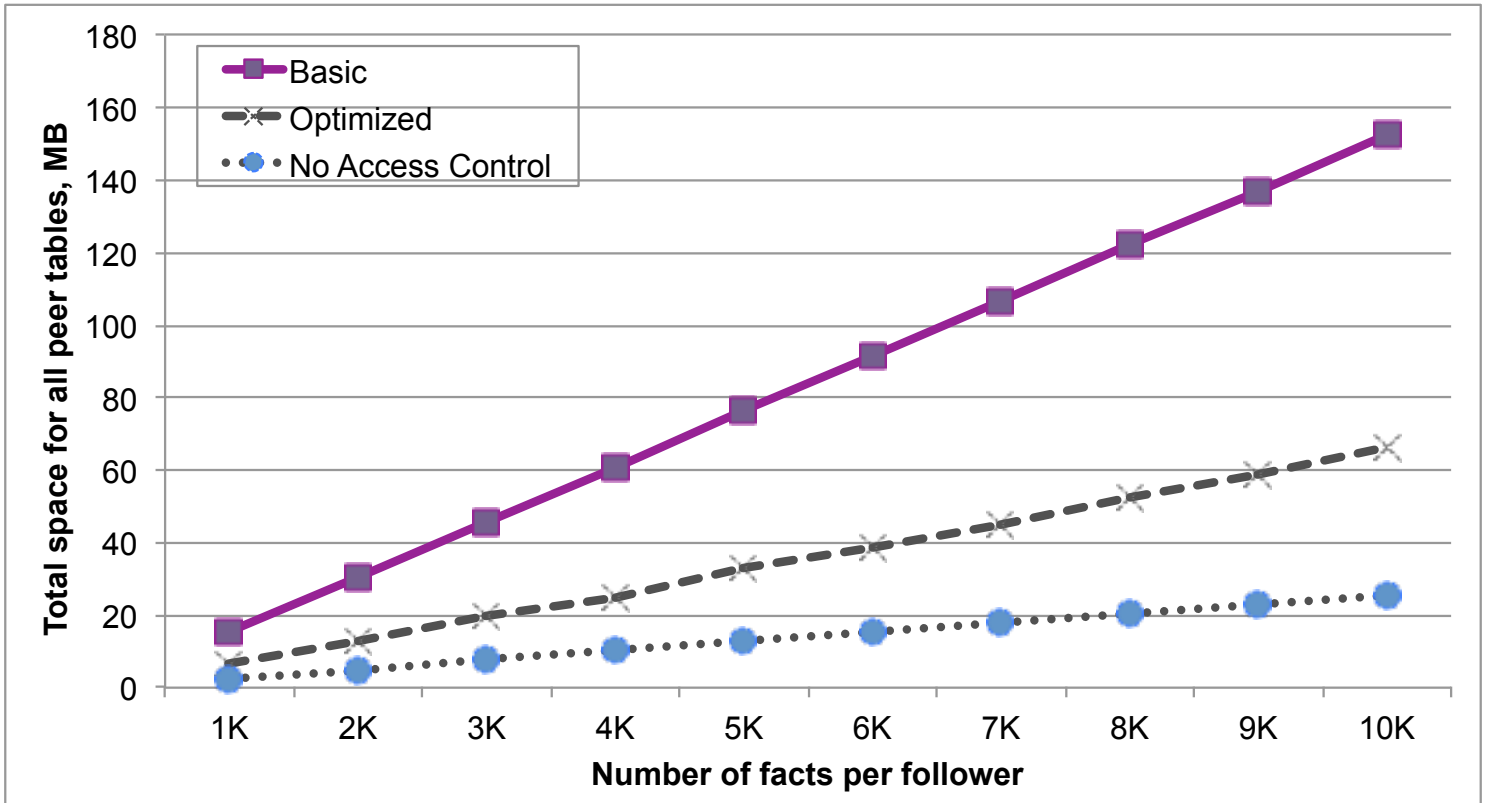
joint with Moffitt [Drexel], Abiteboul [INRIA], Miklau [UMass] - [SIGMOD 2015]

A taste of experimental results: time



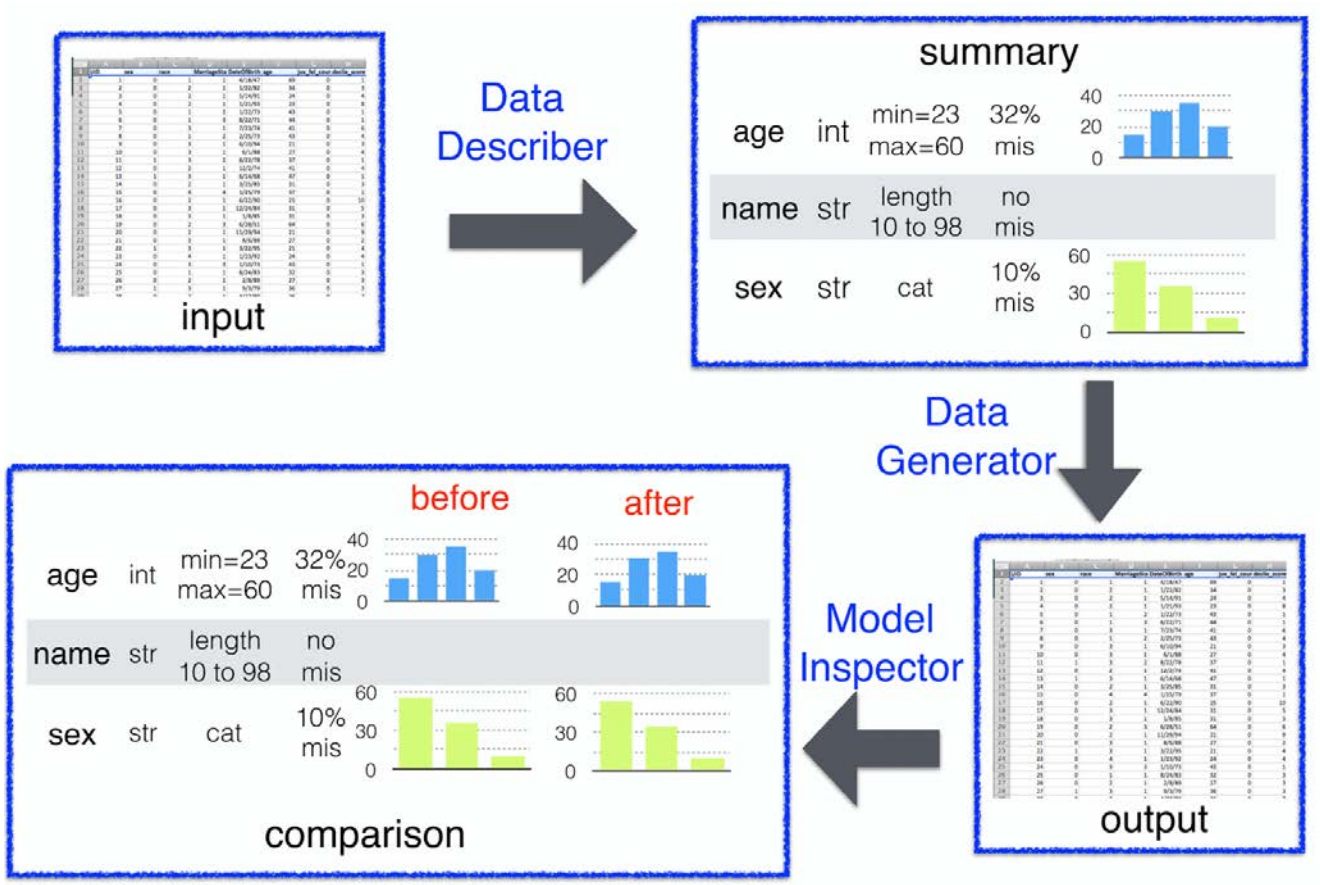
joint with Moffitt [Drexel], Abiteboul [INRIA], Miklau [UMass] - [SIGMOD 2015]

A taste of experimental results: space



joint with Moffitt [Drexel], Abiteboul [INRIA], Miklau [UMass] - [SIGMOD 2015]

DataSynthesizer: usable differential privacy



http://demo.dataresponsibly.com/synthesizer/

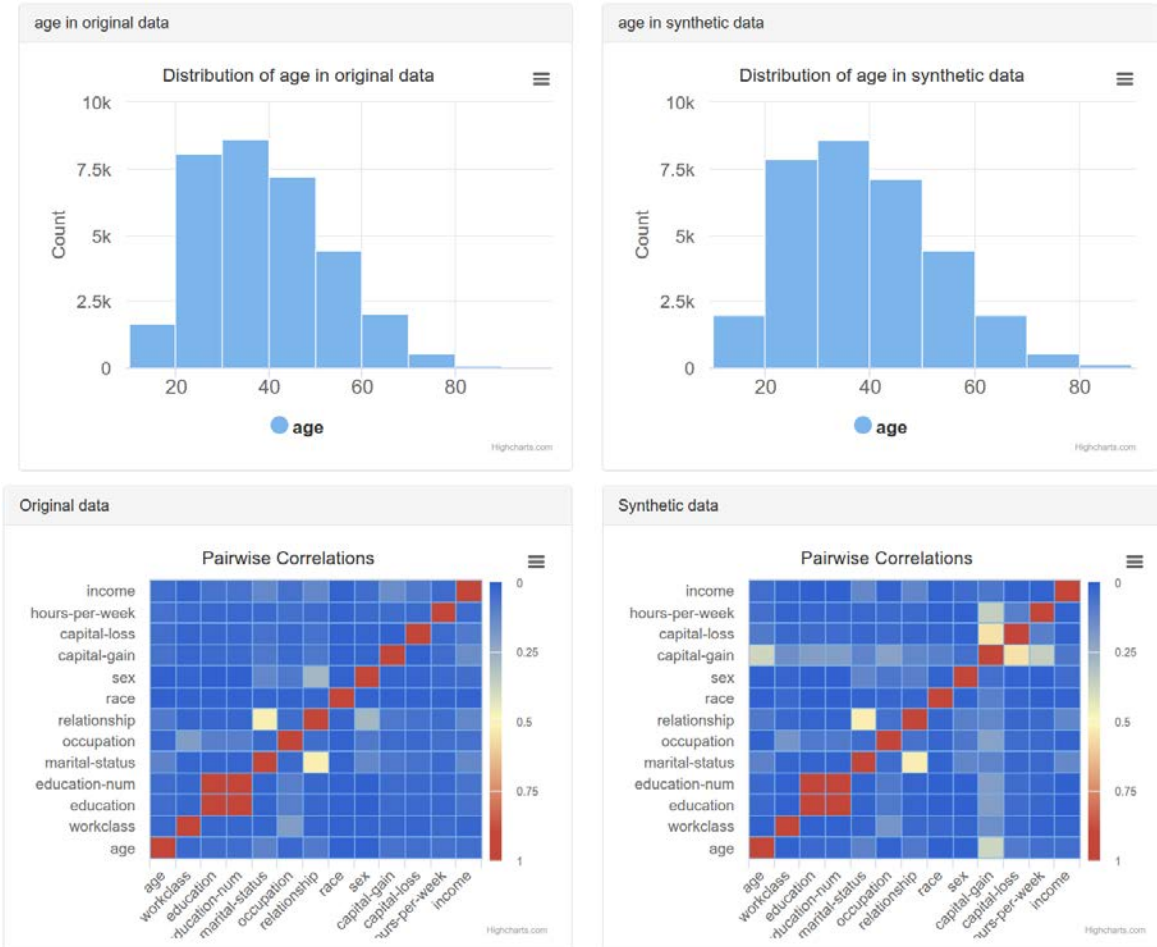
joint with Ping [Drexel] and Howe [UW] - [SSDBM 2017, D4GX 2017]

DataSynthesizer

- Easy to use: a CSV file as input, no schema description
- Generates and releases synthetic datasets that are
  - privacy-preserving - **differentially private**
  - statistically similar to real data
- There modes of operation
  - random type-consistent values
  - independent attributes - based on noisy histograms
  - correlated attributes - privately learn a Bayesian Network
- Interesting **translational research** challenges: usability / important standard assumptions of DP work don't hold in practice

joint with Ping [Drexel] and Howe [UW] - [SSDBM 2017, D4GX 2017]

But does it work?



http://demo.dataresponsibly.com/synthesizer/

joint with Ping [Drexel] and Howe [UW] - [SSDBM 2017, D4GX 2017]

government technology

MetroLab “Innovation of the Month”

SECURITY

University Researchers Use 'Fake' Data for Social Good

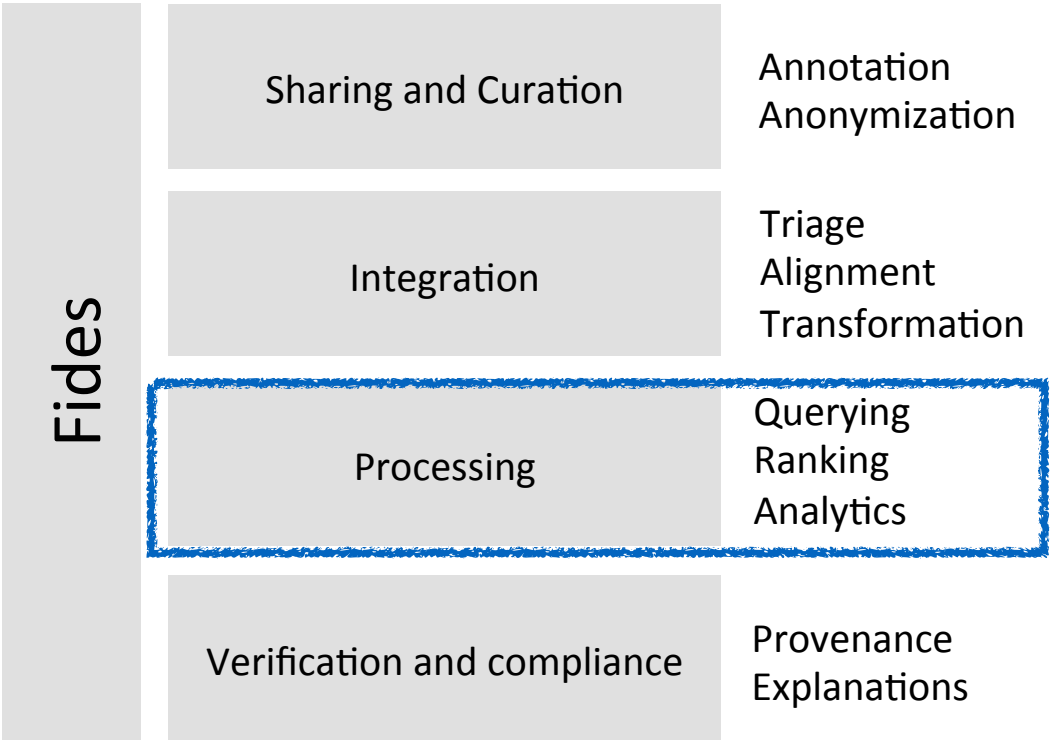
BY BEN LEVINE / NOVEMBER 7, 2017

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 installmer 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

Fides: a responsible data science platform



Systems support for responsible data science

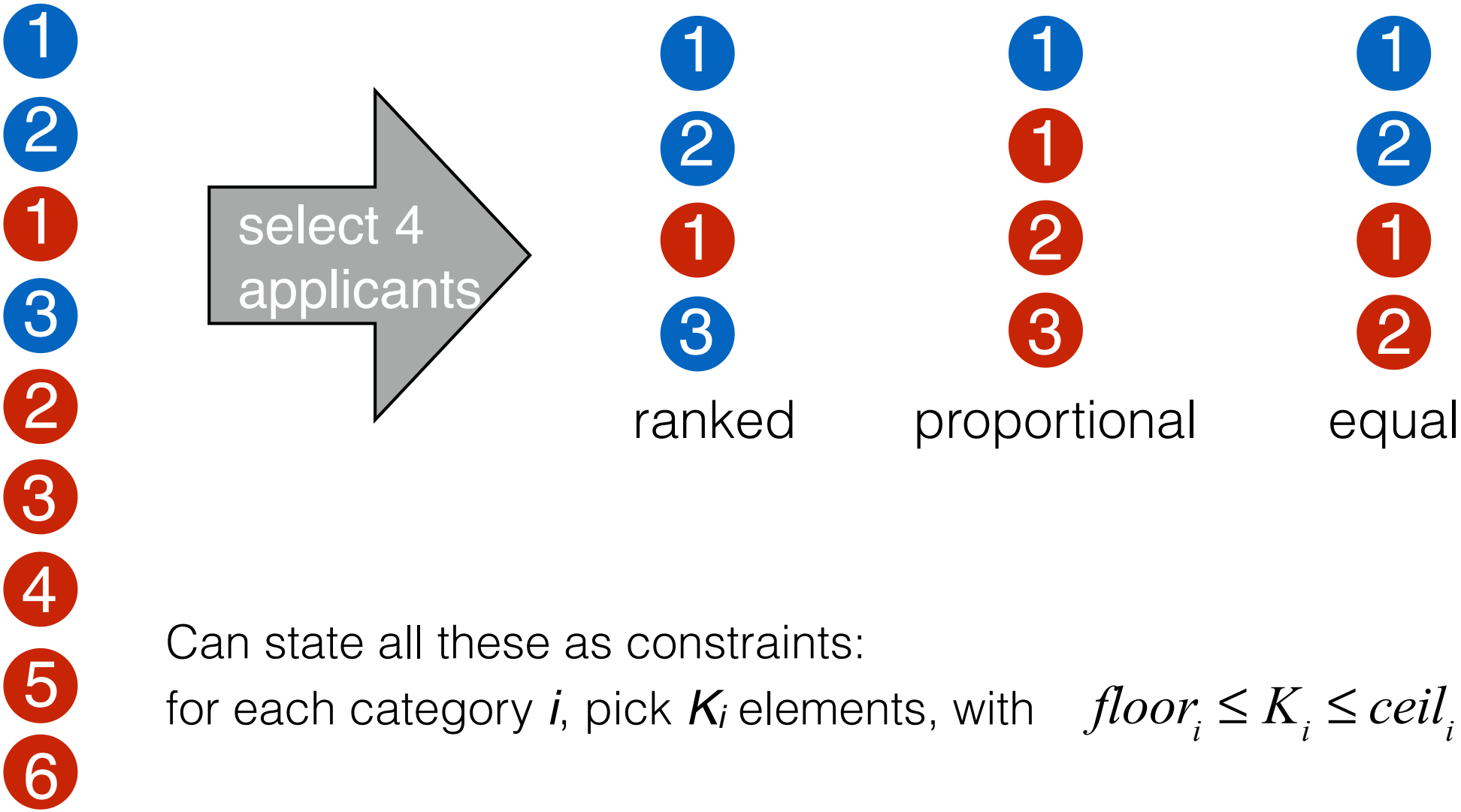
Responsibility by design, managed at all stages of the lifecycle of data-intensive applications

Applications: data science for social good

[BIGDATA] Foundations of responsible data management, 09/2017-

data RESPONSIBLY

Job applicant selection

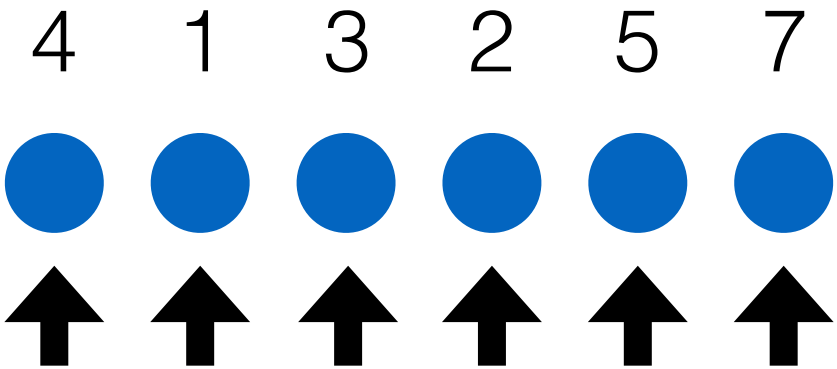


Can state all these as constraints:

for each category *i*, pick K_i elements, with $floor_i \leq K_i \leq ceil_i$

Hiring a job candidate

Goal: Hire a candidate with a high score



Candidates arrive one-by-one

A candidate's score is revealed when the candidate arrives

Decision to accept or reject a candidate made on the spot

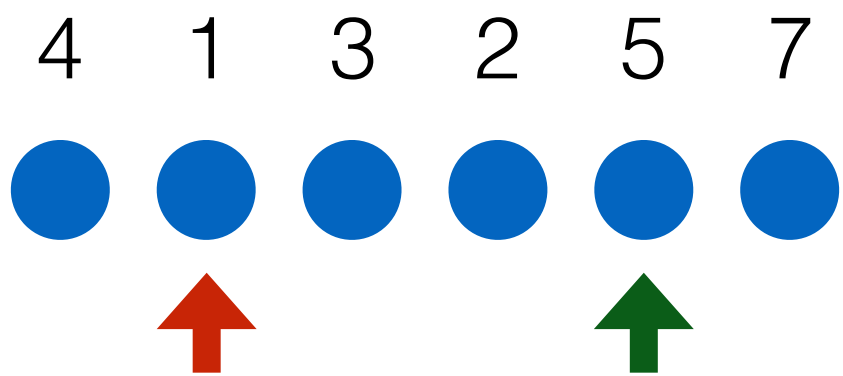
The Secretary Problem

Goal: Design an algorithm for picking **one** element of a **randomly ordered** sequence, to maximize the probability of picking the **maximum element** of the entire sequence.

$N = 6$

$S = \left\lfloor N/e \right\rfloor = 2$

$T = 4$



Competitive ratio

$1/e$

the best possible!

Consider, and reject, the first S candidates

Record T, the best seen score among the first S candidates

Accept the next candidate with score better than T

K-choice Secretary

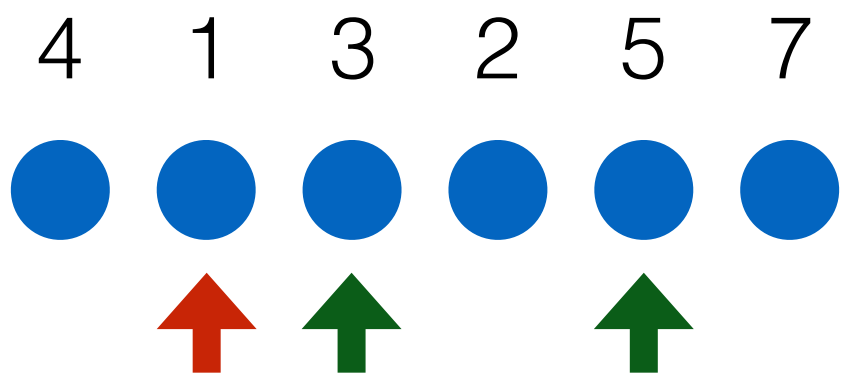
[Babaioff et al., 2007]

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

$N = 6 \quad K = 2$

$S = \left\lfloor N/e \right\rfloor = 2$

$T = \{1, 4\}$



Competitive ratio

$1/e$

far from optimal

Consider, and reject, the first *S* candidates

Record *K* best scores among the first *S* candidates, call this *T*

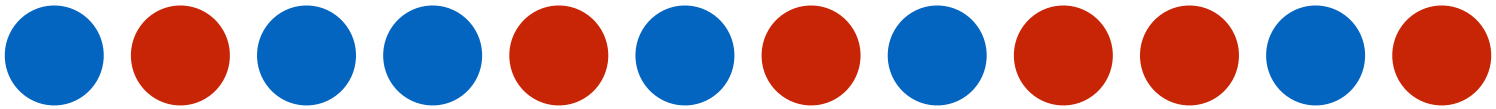
Whenever a candidate arrives whose score is higher than the minimum in *T*, accept the candidate and delete the minimum from *T*

Diverse K-choice Secretary

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

For each **category** *i*, pick K_i elements, with $floor_i \leq K_i \leq ceil_i$

6 4 1 8 2 3 1 2 9 5 7 5



$$N_{red} = N_{blue} = 6$$

$$K = 3$$

$$1 \leq K_{red}, K_{blue} \leq 2$$

Accept *floor* items for each category from per-category streams

$$slack = K - (floor_{red} + floor_{blue})$$

Accept the remaining *slack* items irrespective of category membership, but subject to *ceil*

joint with Yang [Drexel] and Jagadish [UMich] - [EDBT 2018]

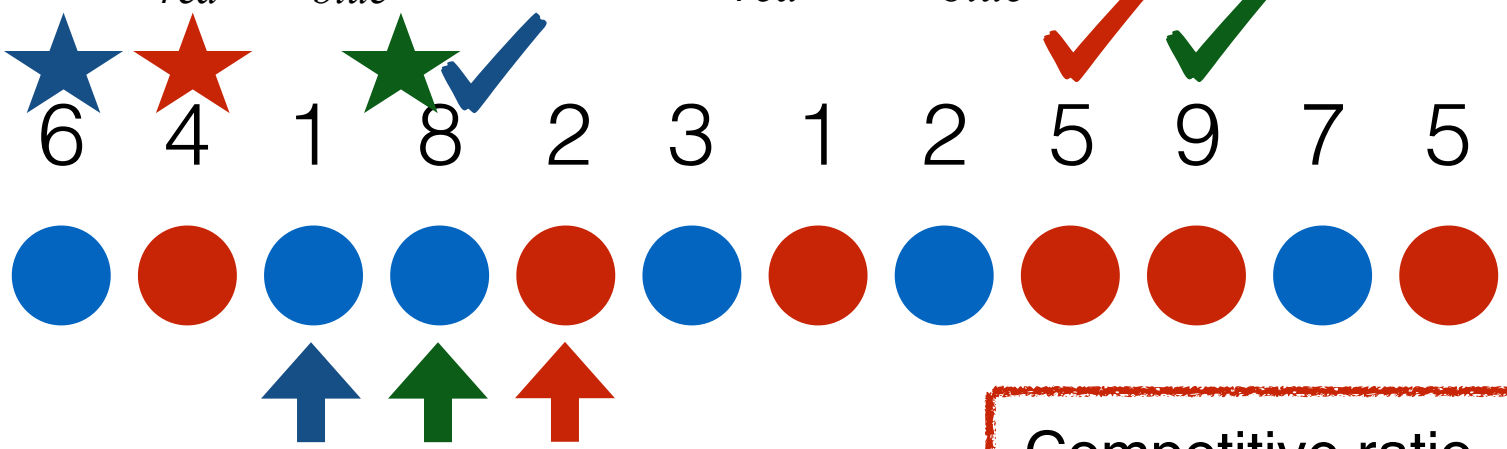
Diverse K-choice Secretary

$N_{red} = N_{blue} = 6$

$K = 3 \quad 1 \le K_{red}, K_{blue} \le 2$

$slack = 1$

$S_{red} = S_{blue} = 2 \quad S = 4$



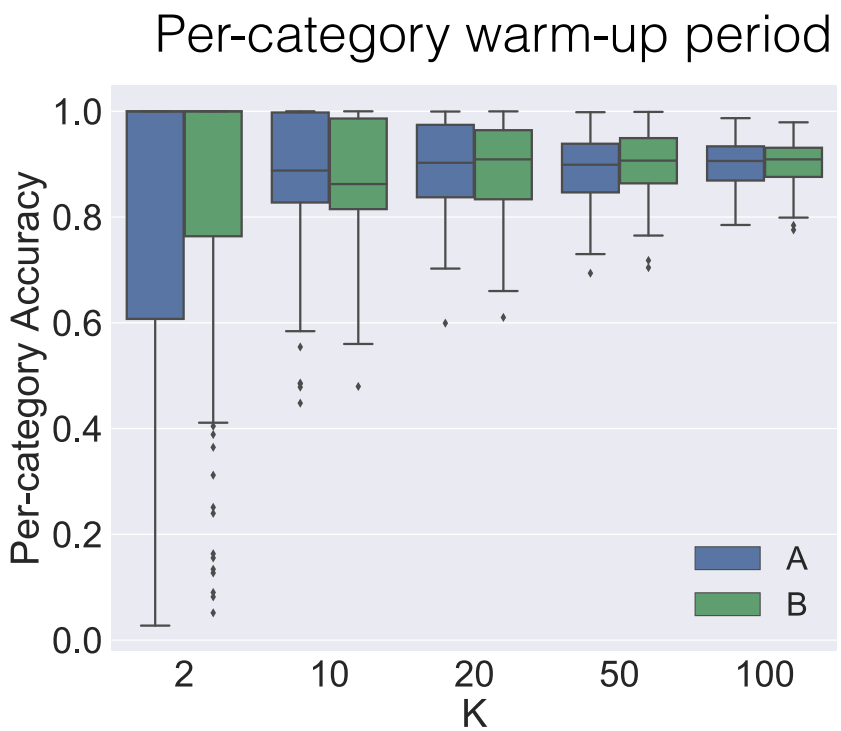
Competitive ratio

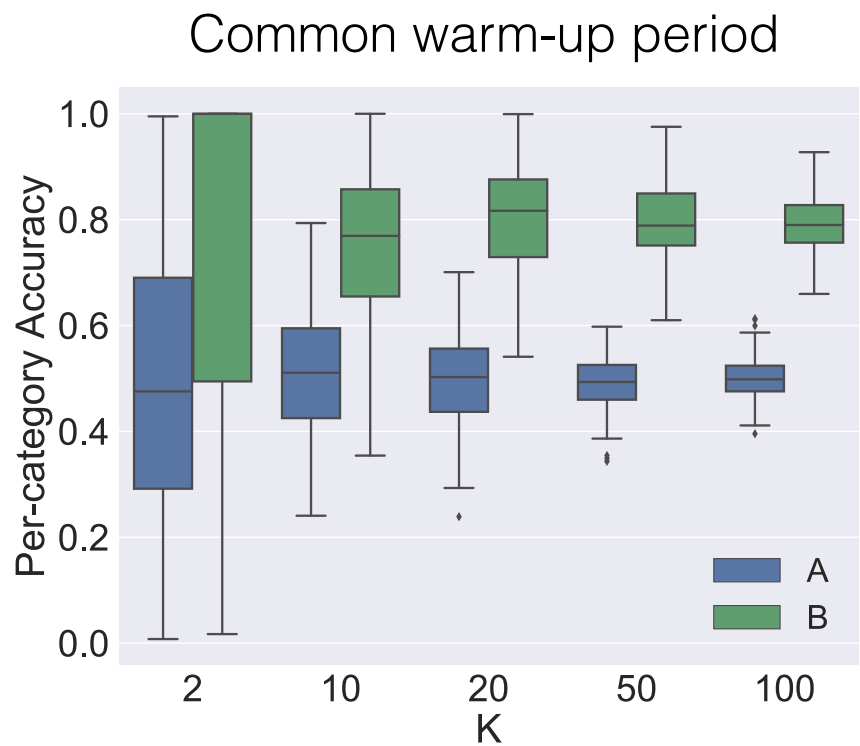
$1/e$

far from optimal

joint with Yang [Drexel] and Jagadish [UMich] - [EDBT 2018]

Per-category warm-up is crucial





synthetic data with categories A and B, score depends on category, lower for A

diversity by design

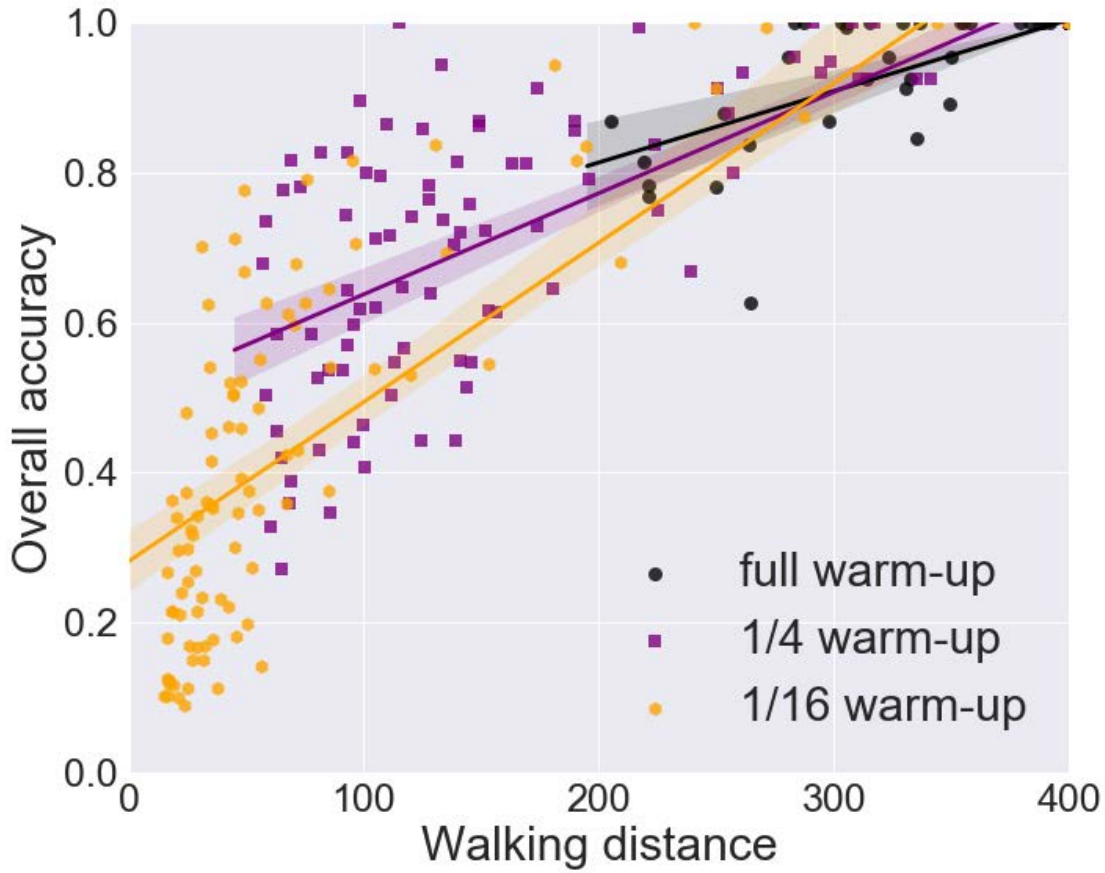
joint with Yang [Drexel] and Jagadish [UMich] - [EDBT 2018]

Diversity is achievable

Forbes US Richest: N=400, K=4 (27 female, 373 male)

diversity on gender: select 2 per gender

Warm-up can be shorter



Forbes US Richest: N=400, K=4 (27 female, 373 male)

deferred list variant, diversity on gender: select 2 per gender

Lack of diversity: harms and approaches

The New York Times

Artificial Intelligence's White Guy Problem

By KATE CRAWFORD JUNE 25, 2016

Like all technologies before it, artificial intelligence will reflect the values of its creators. So **inclusivity matters** — from who designs it to who sits on the company boards and which ethical perspectives are included.

Otherwise, **we risk constructing machine intelligence that mirrors a narrow and privileged vision of society**, with its old, familiar biases and stereotypes.

REVIEW

Diversity in Big Data: A Review

Marina Drosou[1], H.V. Jagadish[2], Evaggelia Pitoura[1], and Julia Stoyanovich[3,*]

Big Data
Volume 5 Number 2, 2017

DOI: 10.1089/big.2016.0054

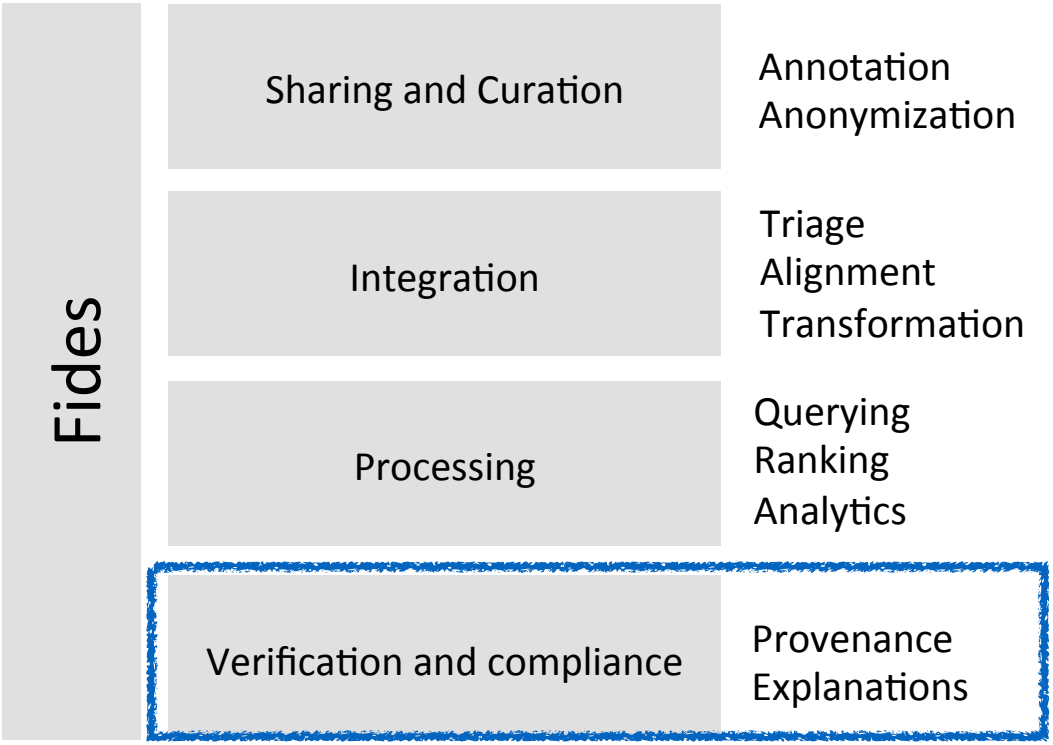
Abstract
Big data technology offers unprecedented opportunities to society as a whole and also to its individual members. At the same time, this technology poses significant risks to those it overlooks. In this article, we give an overview of recent technical work on diversity, particularly in selection tasks, discuss connections between diversity and fairness, and identify promising directions for future work that will position diversity as an important component of a data-responsible society. We argue that diversity should come to the forefront of our discourse, for reasons that are both ethical—to mitigate the risks of exclusion—and utilitarian, to enable more powerful, accurate, and engaging data analysis and use.

Keywords: data; diversity; empirical studies; models and algorithms; responsibly

+ Fairness in ranked outputs, joint with Yang [Drexel] [FATML 2016] [SSDBM 2017]

data RESPONSIBLY

Fides: a responsible data science platform



Systems support for responsible data science

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

Applications: data science for social good

[BIGDATA] Foundations of responsible data management, 09/2017-

data RESPONSIBLY

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

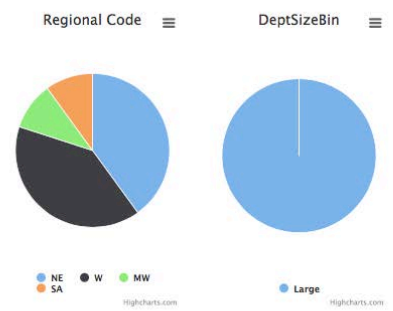
Top 10:

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

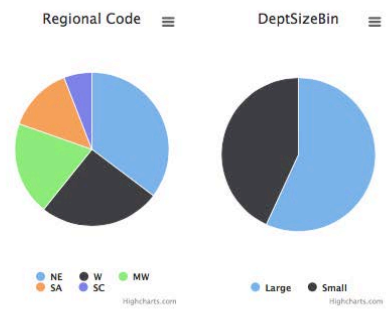
Overall:

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

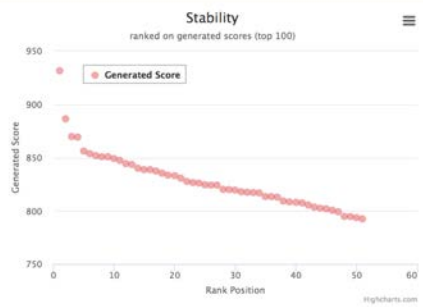
Diversity at top-10



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

	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/

joint with Yang [Drexel], Howe [UW], Jagadish & Asudeh [UMich], Miklau [UMass] - [SIGMOD 2018]

How do we make an impact?

- **An emerging community of research and practice:**
  - FAT*: Conference on Fairness, Accountability and Transparency
- **Getting the existing technical communities on board:**
  - SIGMOD 2018 session, VLDB 2018 debate, EDBT 2016 tutorial, …
- **Policy:**
  - NYC algorithmic transparency law
  - ACM Code of Ethics, CPEDS
- **"Translation":**
  - Let's build tools! Data Synthesizer, Ranking Facts, ….
  - PhillyOpenData

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

Report from Dagstuhl Seminar 16291

Data, Responsibly

Edited by

Serge Abiteboul[1], Gerome Miklau[2], Julia Stoyanovich[3], and Gerhard Weikum[4]

1 ENS – Cachan, FR, serge.abiteboul@inria.fr
2 University of Massachusetts – Amherst, US, miklau@cs.umass.edu
3 Drexel University – Philadelphia, US, stoyanovich@drexel.edu
4 MPI für Informatik – Saarbrücken, DE, weikum@mpi-inf.mpg.de

The goals of the seminar were to assess the state of data analysis in terms of fairness, transparency and diversity, identify new research challenges, and derive an agenda for computer science research and education efforts in responsible data analysis and use.

An important goal of the seminar was to **identify opportunities for high-impact contributions to this important emergent area specifically from the data management community**.

http://drops.dagstuhl.de/opus/volltexte/2016/6764/pdf/dagrep_v006_i007_p042_s16291.pdf

Research Directions for Principles of Data Management (Dagstuhl Perspectives Workshop 16151)

Edited by

Serge Abiteboul, Marcelo Arenas, Pablo Barceló, Meghyn Bienvenu, Diego Calvanese, Claire David, Richard Hull, Eyke Hüllermeier, Benny Kimelfeld, Leonid Libkin, Wim Martens, Tova Milo, Filip Murlak, Frank Neven, Magdalena Ortiz, Thomas Schwentick, Julia Stoyanovich, Jianwen Su, Dan Suciu, Victor Vianu, and Ke Yi

1 Introduction

In April 2016, a community of researchers working in the area of Principles of Data Management (PDM) joined in a workshop at the Dagstuhl Castle in Germany. The workshop was organized jointly by the Executive Committee of the ACM Symposium on Principles of Database Systems (PODS) and the Council of the International Conference on Database Theory (ICDT). The mission of this workshop was to identify and explore some of the most important research directions that have high relevance to society and to Computer Science today, and where the PDM community has the potential to make significant contributions. This report describes the family of research directions that the workshop focused on from three perspectives: potential practical relevance, results already obtained, and research questions that appear surmountable in the short and medium term. This report organizes the identified research challenges for PDM around seven core themes, namely *Managing Data at Scale*, *Multi-model Data*, *Uncertain Information*, *Knowledge-enriched Data*, *Data Management and Machine Learning*, *Process and Data*, and *Ethics and Data Management*. Since new challenges in PDM arise all the time, we note that this list of themes is not intended to be exclusive.

Dagstuhl Manifestos 7(1): 1-29 (2018)

INTERNATIONAL | POLITIQUE | SOCIÉTÉ | ÉCO | CULTURE | IDÉES | PLANÈTE | SPORT | SCIENCES

Sciences

SCIENCES | Vidéos | Archéologie | Affaire de logique | Astronomie | Biologie | Cerveau | Géophysic

ÉDITION ABONNÉS

Plaidoyer pour une analyse « responsable » des données

Face aux risques d'atteinte à la vie privée, les chercheurs en informatique Serge Abiteboul et Julia Stoyanovich plaident pour une collecte et une analyse des données impartiales, transparentes et accessibles à tous.

LE MONDE SCIENCE ET TECHNO | 12.10.2015 à 20h47 • Mis à jour le 19.10.2015 à 16h16

Association For Computing Machinery acm ACM SIGMOD Blog

Serge Abiteboul and Julia Stoyanovich

NOVEMBER 20, 2015

DATA, RESPONSIBLY

Big Data

(This blog post is an extended version of an October 12, 2015 Le Monde op-ed article (in French))

Our society is increasingly relying on algorithms in all aspects of its operation. We trust algorithms not only *to help carry out routine tasks*, such as accounting and automatic manufacturing, but also *to make decisions on our behalf*. The sorts of decisions with which we now casually entrust algorithms range from unsettling (killer drones), to tedious (automatic trading), or deeply personal (online dating). Computer technology has tremendous power, and with that power comes immense responsibility. Nowhere is the need to control the power and to judiciously use technology more apparent than in massive data analysis, known as big data.

Responsible data science

- Be **transparent** and **accountable**
- Achieve **equitable** resource distribution
- Be cognizant of the **rights** and **preferences** of individuals

fairness

diversity

transparency

data protection

DB+COMSOC: databases meet computational social choice

[NSF III + BSF] DBCOMSOC, 2018-

United States – Israel
Binational Science Foundation

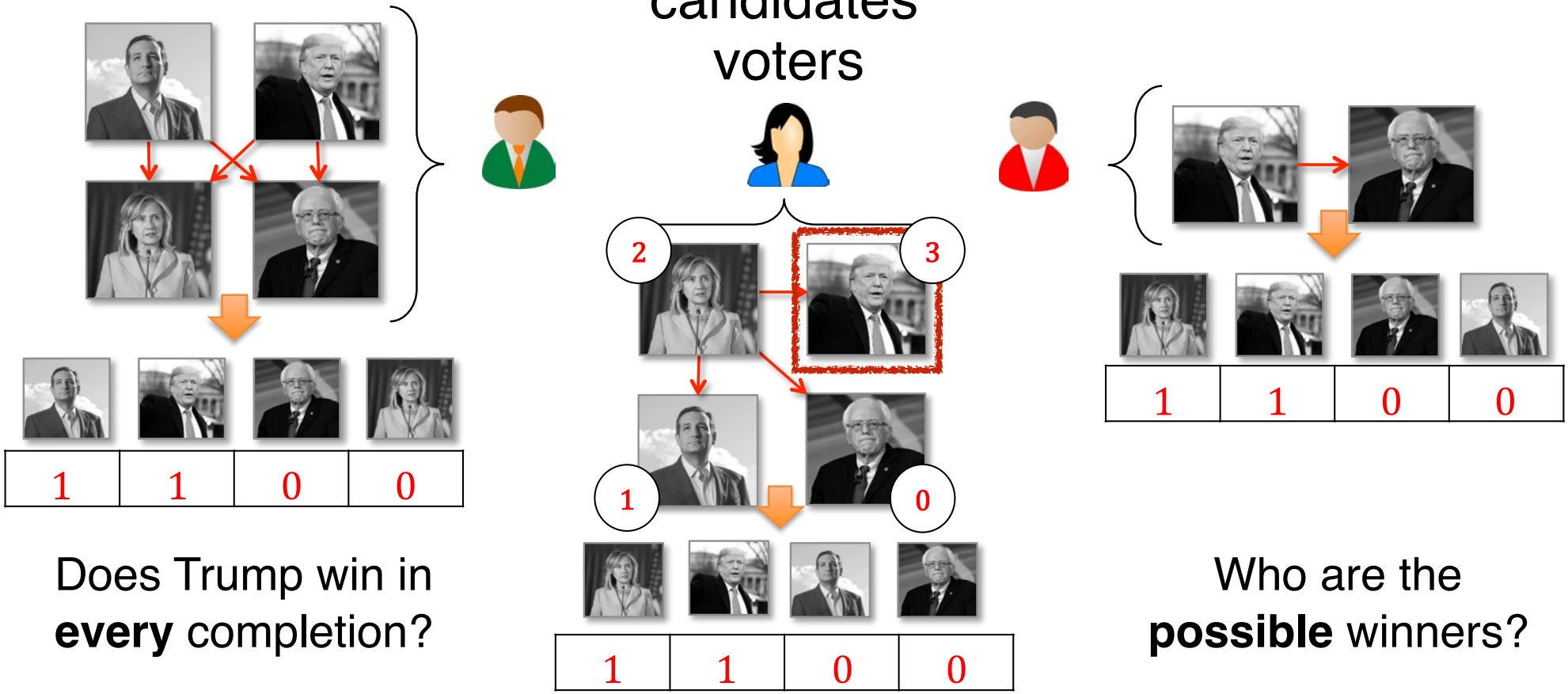
data RESPONSIBLY

Elections and winners

TEASER!

scoring rules:
plurality, veto,
2-approval…

candidates
voters



Does Trump win in **every** completion?

Who are the **possible** winners?

joint with Kimelfeld [Technion] and Kolaitis [UC Santa Cruz] [IJCAI 2018]

data RESPONSIBLY

Context makes a difference!

TEASER!




scoring rules:
plurality, veto,
2-approval…

candidates
voters



Is **every** winner pro-choice?

Is it **possible** that the first spouse will be US-born?

Candidates

cand	party	spouse born	pro-choice
Clinton	D	USA	yes
Trump	R	Slovenia	no
Rubio	R	USA	no
Sander	D	USA	yes

Does Trump win in **every** completion?

Who are the **possible** winners?

joint with Kimelfeld [Technion] and Kolaitis [UC Santa Cruz] [IJCAI 2018]

Thank you!