

Towards responsible data management

Julia Stoyanovich
Drexel University, USA



data RESPONSIBLY

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 Naughton

The New York Times

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

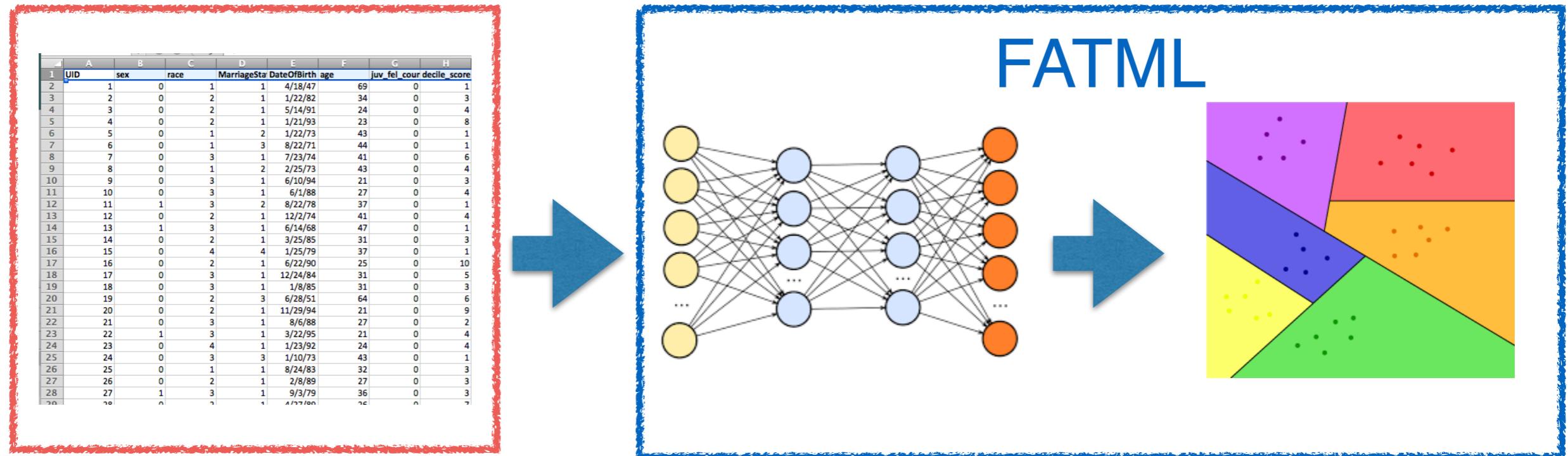
Last year, the total number of sheltered and unsheltered homeless people in the city was 75,323, which included 1,706 people between ages 18 and 24. The actual number of young people is significantly higher, according to the service providers, who said the census mostly captured young people who received 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.

Ending urban homelessness

- A variety of **services**: rapid rehousing, transitional housing, emergency shelter, permanent supportive housing
- A variety of **support mechanisms**: substance abuse treatment, mental health treatment, protection for victims of domestic violence
- Challenges
 - **recommend** pathways through the system
 - **evaluate** effectiveness of interventions
 - **measure** performance of the coordinated system of homeless assistance



Piece of cake!



done?

goal: responsible data science

Data, Responsibly

fairness



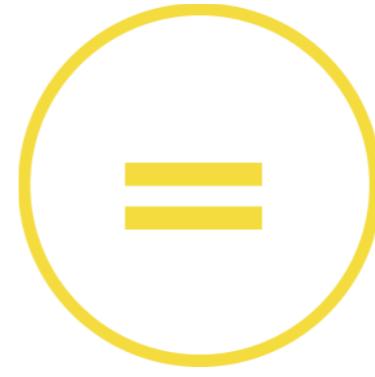
diversity



data protection

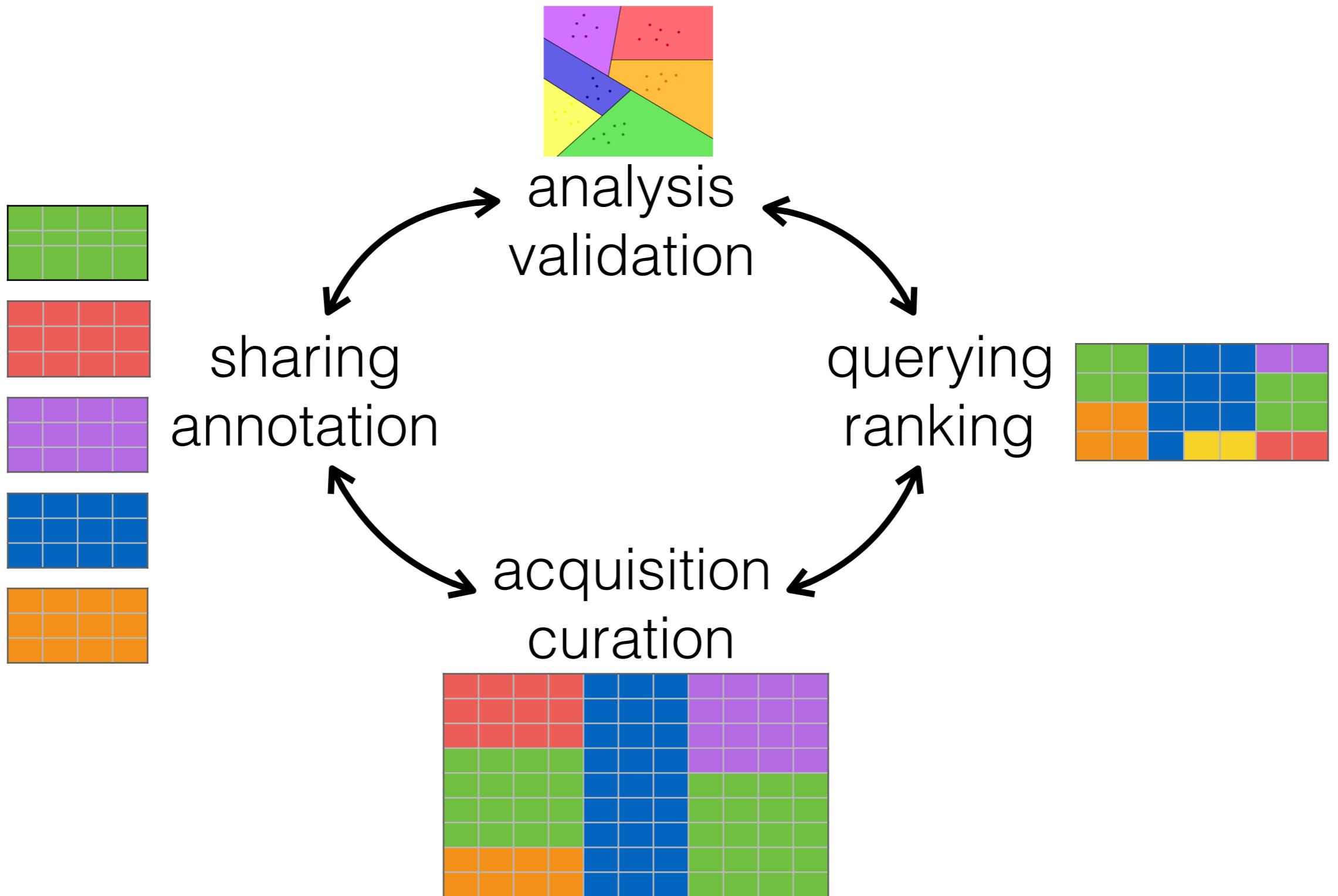


transparency



neutrality

Data science lifecycle



How did we get the data?

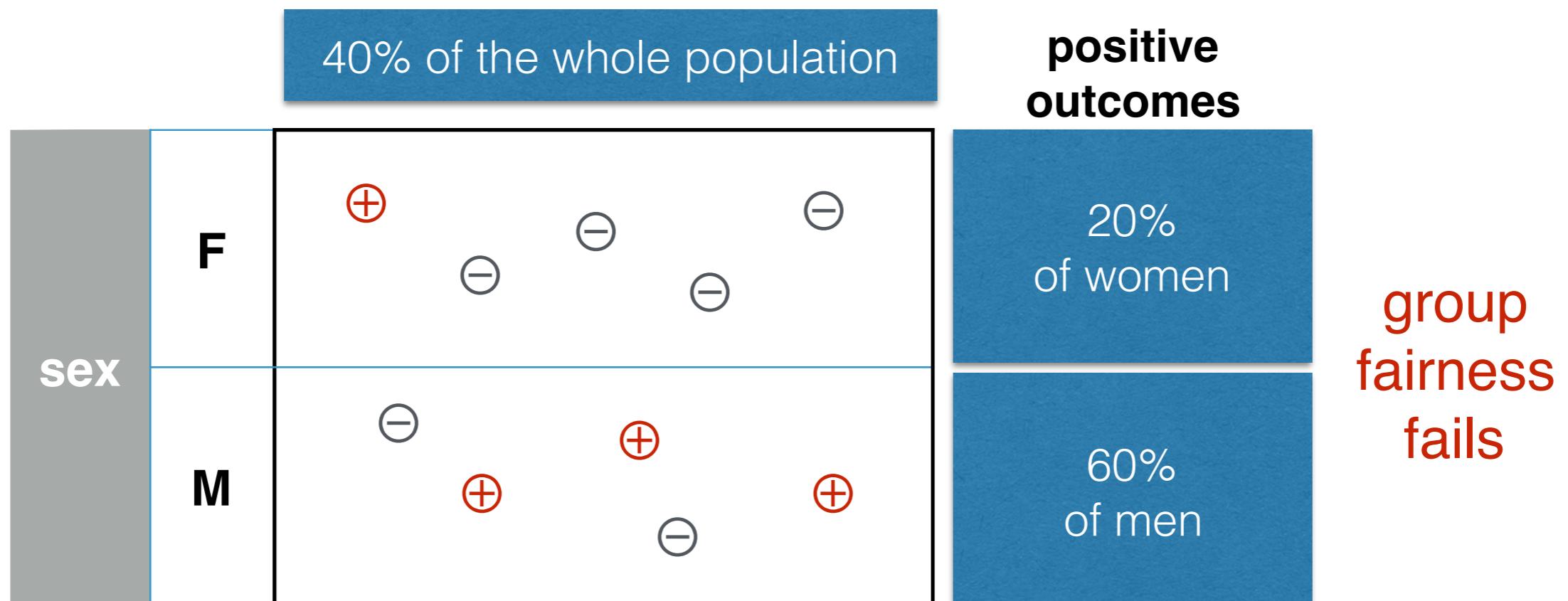
- A multitude of datasets gathered from local communities (HMIS), data is **weakly structured**: inconsistencies, missing values, hidden and apparent bias
- Some data was **anonymized**, other data was **not shared** in fear of violating regulations or the trust of participants
- Shared data was **triaged, aligned, integrated** (ETL + SQL)
- Integrated data was then **filtered** (SQL) and **prioritized** (sorted/ranked), and only then passed as input to the learning module

responsible data science starts with responsible data collection, sharing, integration, querying, ranking - with **responsible data management!**

An important property: fairness

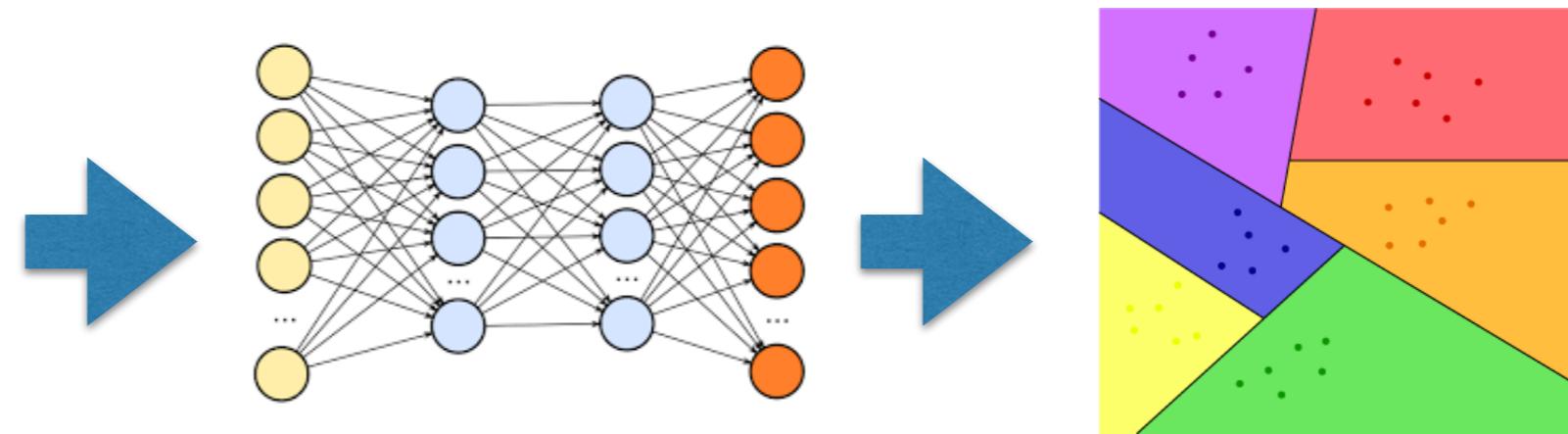
Group fairness (aka **statistical parity**)

demographics of the individuals receiving any outcome are the same as demographics of the underlying population



Let's revisit the analytics step

1	A	B	C	D	E	F	G	H
2	UID	sex	race	MarriageStat	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
29	29	0	3	1	8/27/80	26	0	7



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

fix the model!

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

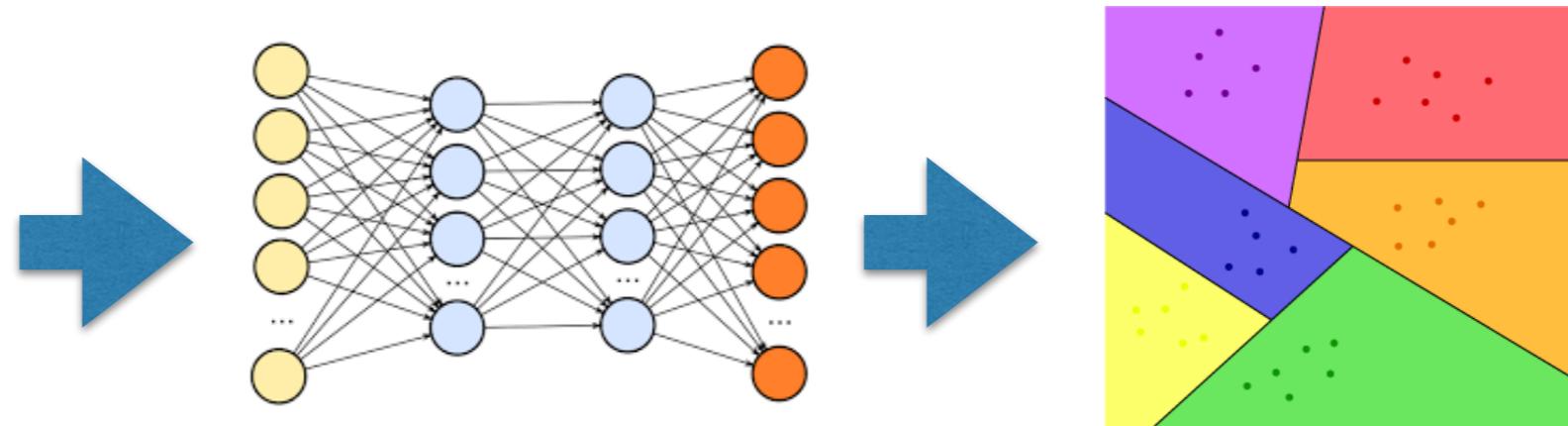
select * from R
where status = 'unsheltered'
and length > 2 month
and length > 1 month

10% female

40% female

Let's revisit the analytics step

1	A	B	C	D	E	F	G	H
1	UID	sex	race	MarriageStat	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
29	29	0	3	1	8/27/80	24	0	7



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

of course, but maybe...

fix the model!

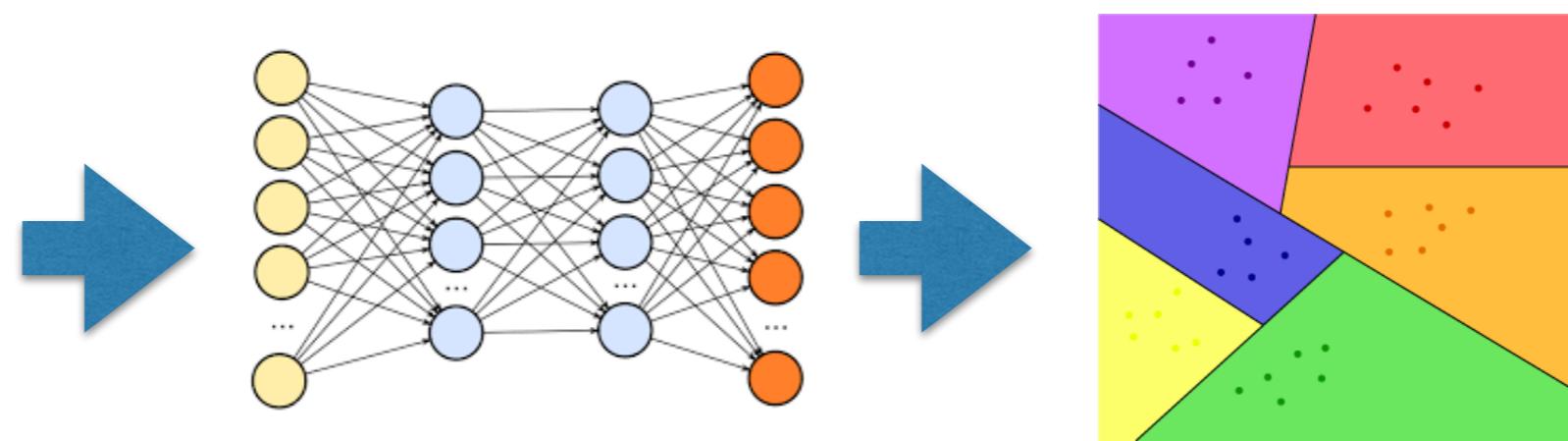
mental health info was missing for this population

go back to the triage step, look for additional datasets



Let's revisit the analytics step

1	A	B	C	D	E	F	G	H
2	UID	sex	race	MarriageStat	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
29	29	0	3	1	8/27/80	24	0	7



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

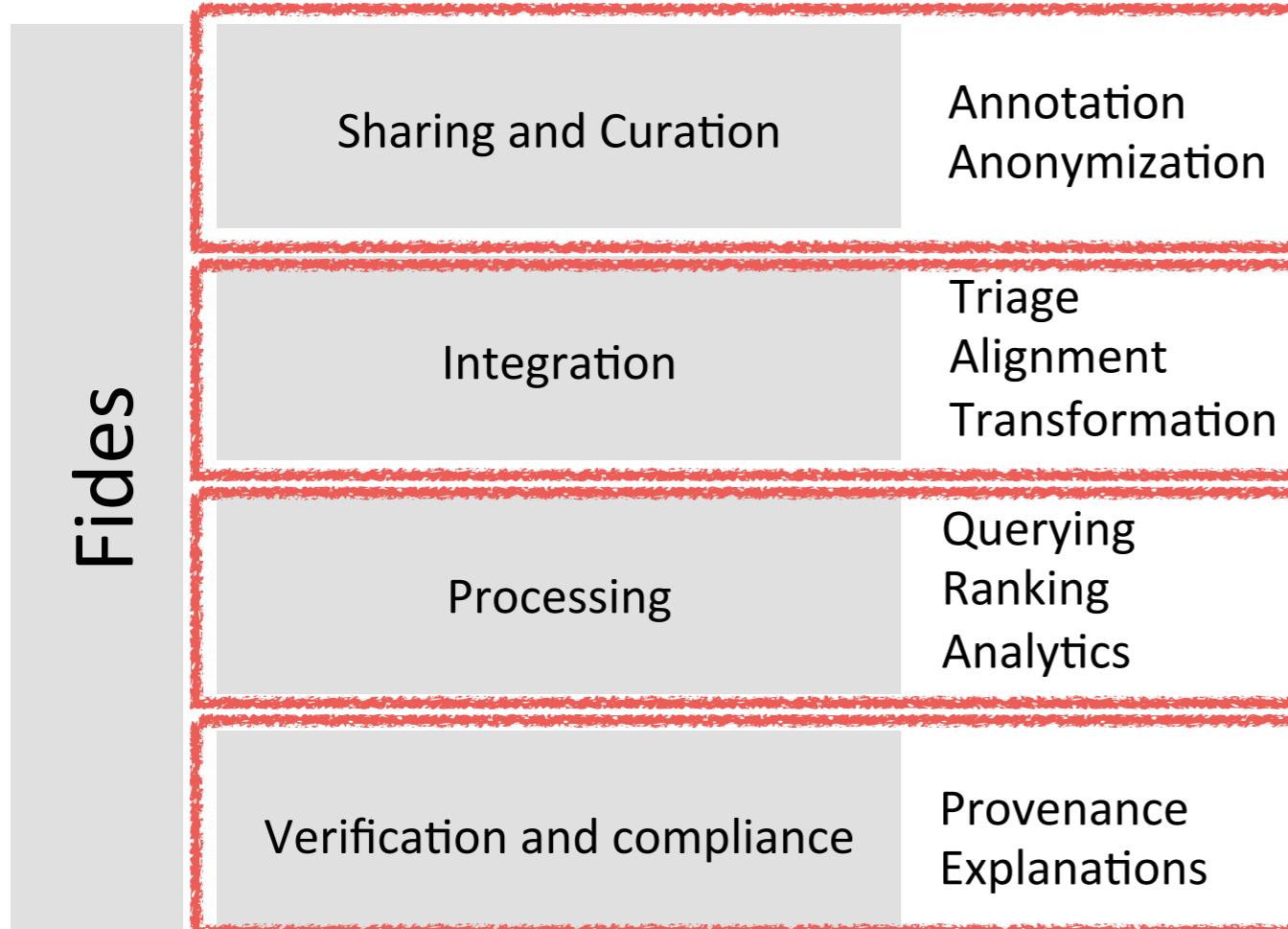
unlikely to help!

fix the model??

minors data was not shared

revisit the data sharing step, help data providers share their data while adhering to regulation and trust of the participants

Systems support for responsible data science



Fides: A responsible data science platform.

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

Application: DS for social good / urban homelessness

key point: holistic view of the lifecycle, information about both data and process, allow us to do much more!



Report from Dagstuhl Seminar 16291

Data, Responsibly

Edited by

Serge Abiteboul¹, Gerome Miklau², Julia Stoyanovich³, and
Gerhard Weikum⁴

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)

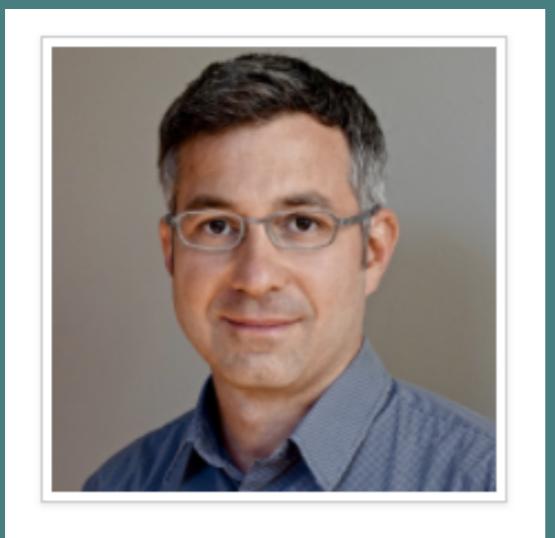
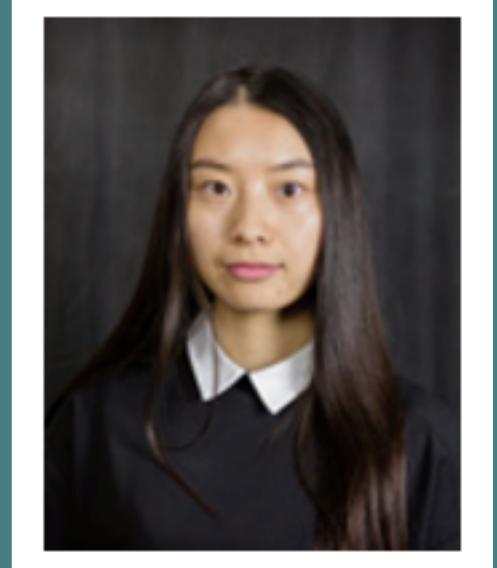
Edited by

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.

<https://arxiv.org/pdf/1701.09007.pdf>



Thank you!



data *RESPONSIBLY*

