CS 594: 2- From Data to Action
How things can go wrong

Abolfazl Asudeh
Fall 2020
9/1/2020

- References:
  - Solon Barocas, Moritz Hardt, Arvind Narayanan. "Fairness in Machine Learning, Limitations and Opportunities" (book) – link available in the course page
  - (Tutorial) A. Asudeh, H. V. Jagadish. Fairly Evaluating and Scoring Items in a Data Set. PVLDB, 13(12): 3445-3448, 2020

Side note: (happening now) VLDB'20 is a FREE Gem...

- Keynote: Responsible Data Management. Julia Stoyanovich
- Tutorial: Fairly Evaluating and Scoring Items in a Data set. Asudeh and Jagadish
- Several papers:
  - Fair Task Assignment in Spatial Crowdsourcing
  - Rank Aggregation Algorithms for Fair Consensus
  - ...

The loop of Data-driven Decision Making

State of the world

[FairML book]

Scoring, Ranking, and Classification

Happens everywhere,

Scoring is a common way to perform evaluations such as ranking, classification, or selection.

Scoring can have other uses, e.g. directly uesd. Similarly, ranking can be done without scoring, e.g. by pairwise comparison.

In the following, we consider scoring for ranking and classification, performed each on their own and performed jointy.

Notations

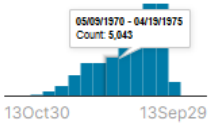
Data set

- Rows: Items/Object/tuples…
- Columns: attributes/features…
- Training data: iid sample from the underlying data distribution

Case_ID	Agency_T…	LastName	FirstName
51950	PRETRIAL	Fisher	Kevin
51950	PRETRIAL	Fisher	Kevin
51950	PRETRIAL	Fisher	Kevin
51956	PRETRIAL	KENDALL	KEVIN
51956	PRETRIAL	KENDALL	KEVIN
51956	PRETRIAL	KENDALL	KEVIN

Scoring Attributes/ Input Features

- Vector $X = \{X_1, \ldots, X_m\}$
- Observations
- Used for Evaluation

DateOfBirth	ScaleSet	AssessmentRea...	Language	LegalStatus
	Risk and Prescreen 96% All Scales 4%	1 unique value	English 100% Spanish 0%	Pretrial 62% Post Sentence 30% Other (4932) 8%
12/05/92	Risk and Prescreen	Intake	English	Pretrial
12/05/92	Risk and Prescreen	Intake	English	Pretrial
12/05/92	Risk and Prescreen	Intake	English	Pretrial
09/16/84	Risk and Prescreen	Intake	English	Pretrial
09/16/84	Risk and Prescreen	Intake	English	Pretrial
09/16/84	Risk and Prescreen	Intake	English	Pretrial

Target Value / True Label

- Y: (usually) a non-ordinal, categorical attribute
- Ground-truth value of evaluation (class labels)
- Unseen
- Evaluation (prediction) outcome

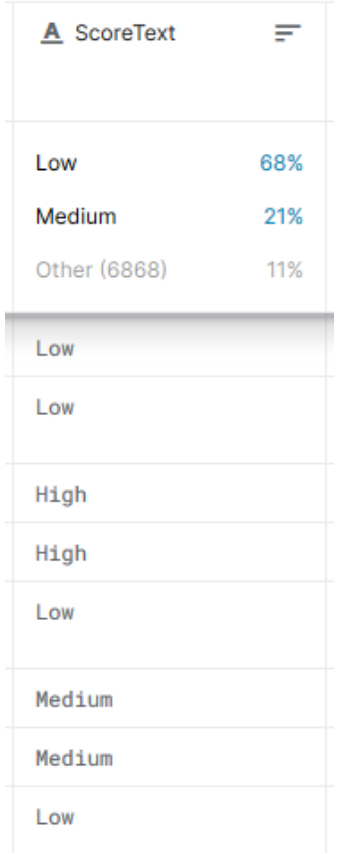
Sensitive Attribute(s)

- S: sensitive attribute(s) such as race and gender that identify demographic groups such as male, black, etc
- Protected Group: Minority groups
  - e.g.: Female, black, ...
- Protected attribute? → alternative for sensitive attribute

Sex_Code_Text		Ethnic_Code_Text	
Male	78%	African-American	44%
Female	22%	Caucasian	36%
		Other (12042)	20%
Male		Caucasian	
Male		Caucasian	
Male		Caucasian	
Male		Caucasian	
Male		Caucasian	

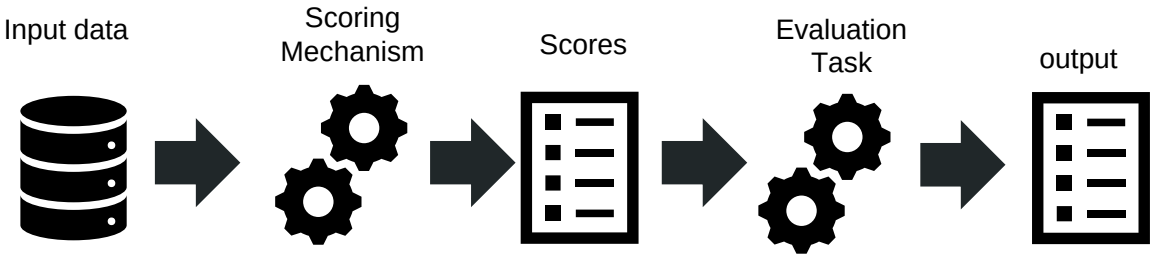
Evaluation

- **Scoring:** f_θ; *e.g.* $f_\theta(X) = \theta^\top X$
  - f : *Mechanism*
  - θ : *Parameters*
- **Evaluation based on scoring:**
  - $\mathrm{h}(f_\theta(X))$
  - We simplify the notation as f_θ or even h_θ



How does evaluation work?

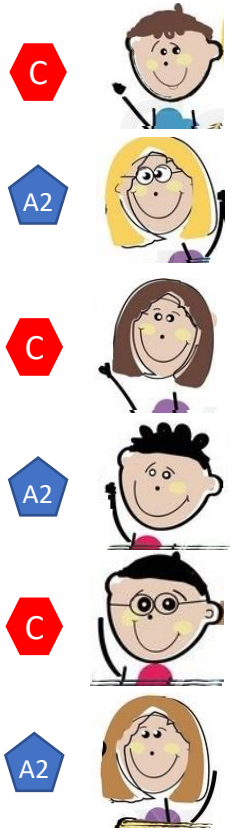
Score-based Evaluation



Toy Example

A2 Ann Arbor

D Chicago



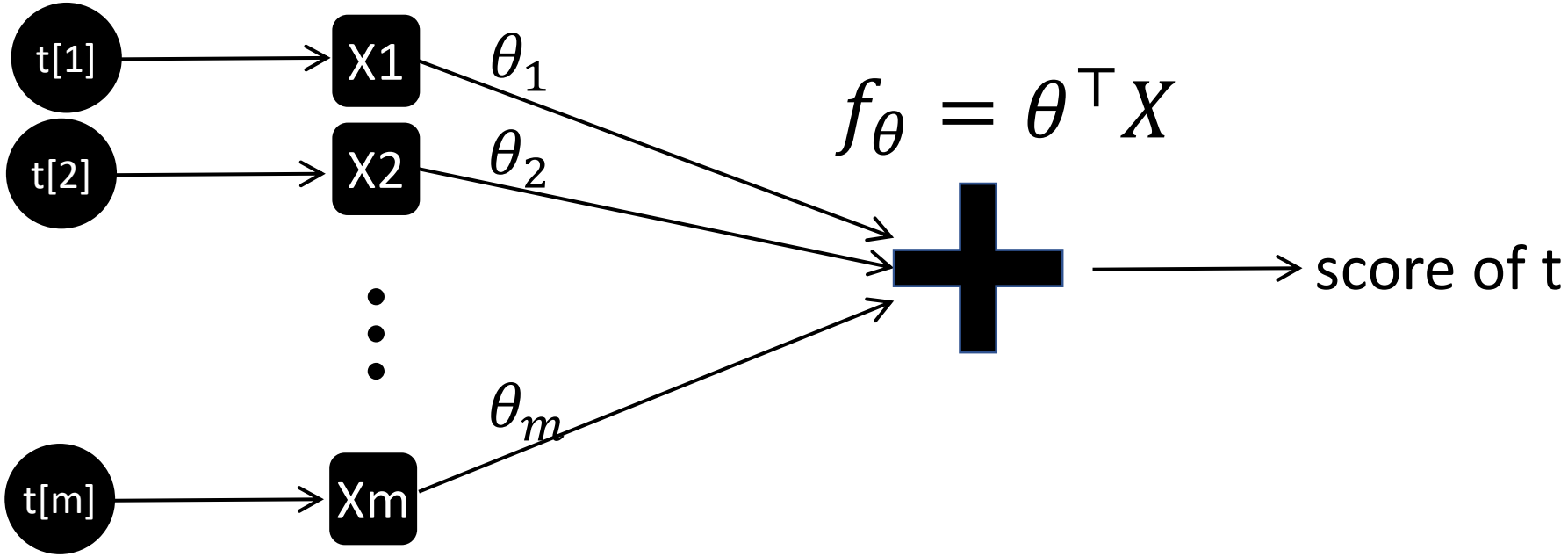
Suppose you own a real estate agency with two branches in Ann Arbor and Chicago.

You want to give bonus to

(1) Top-3 agents

(2) Successful agents

Scoring Mechanism: Linear Scoring



Toy Example

Sale -- Normalized
Customer Satisfaction -- Normalized

0.63	0.71
0.72	0.65
0.58	0.78
0.7	0.68
0.53	0.82
0.61	0.79

$$\sum \theta_i x_i$$

$$\theta_1 \square + \theta_2 \square$$

$1\square + 1\square$

1.34
1.37
1.36
1.38
1.35
1.4

Converting non-linear to linear scoring

- Add non-linear terms as new attributes.
  - Example: $\mathrm{f} = 3X_1^2 + 5X_2^2 + X_1 + 2X_2$
  - Set $X_1' = X_1, X_2' = X_2, X_3' = X_1^2, X_4' = X_2^2$ as the scoring attributes
  - → $\mathrm{f} = 3X_3' + 5X_4' + X'_1 + 2X'_2$

- Use Log function to convert multiplication/exponential functions to linear
  - Example: $\mathrm{f} = 2^{X_1}.X_2^5$
  - Set $X_1' = X_1, X_2' = \log X_2$ as the scoring attributes
  - → $f' = \log f = (\log 2)\, X_1' + 5X_2'$

Ranking based on scoring

- Sort the scores to get the ranking
- (Select the top-k)

Toy Example

Sale -- Normalized

Customer Satisfaction -- Normalized

$$\sum \theta_i x_i$$

$$\theta_1 \square + \theta_2 \square$$

Classification based on scoring

- Use score thresholds to specify decision boundaries
- For binary classification, for example, the scores above the threshold are classified as +1 (or accept) and the ones below it as -1 (or reject).

Toy Example

Sale -- Normalized
Customer Satisfaction -- Normalized






0.63	0.71	1.34	0
0.72	0.65	1.37	1
0.58	0.78	1.36	1
0.7	0.68	1.38	1
0.53	0.82	1.35	0
0.61	0.79	1.4	0

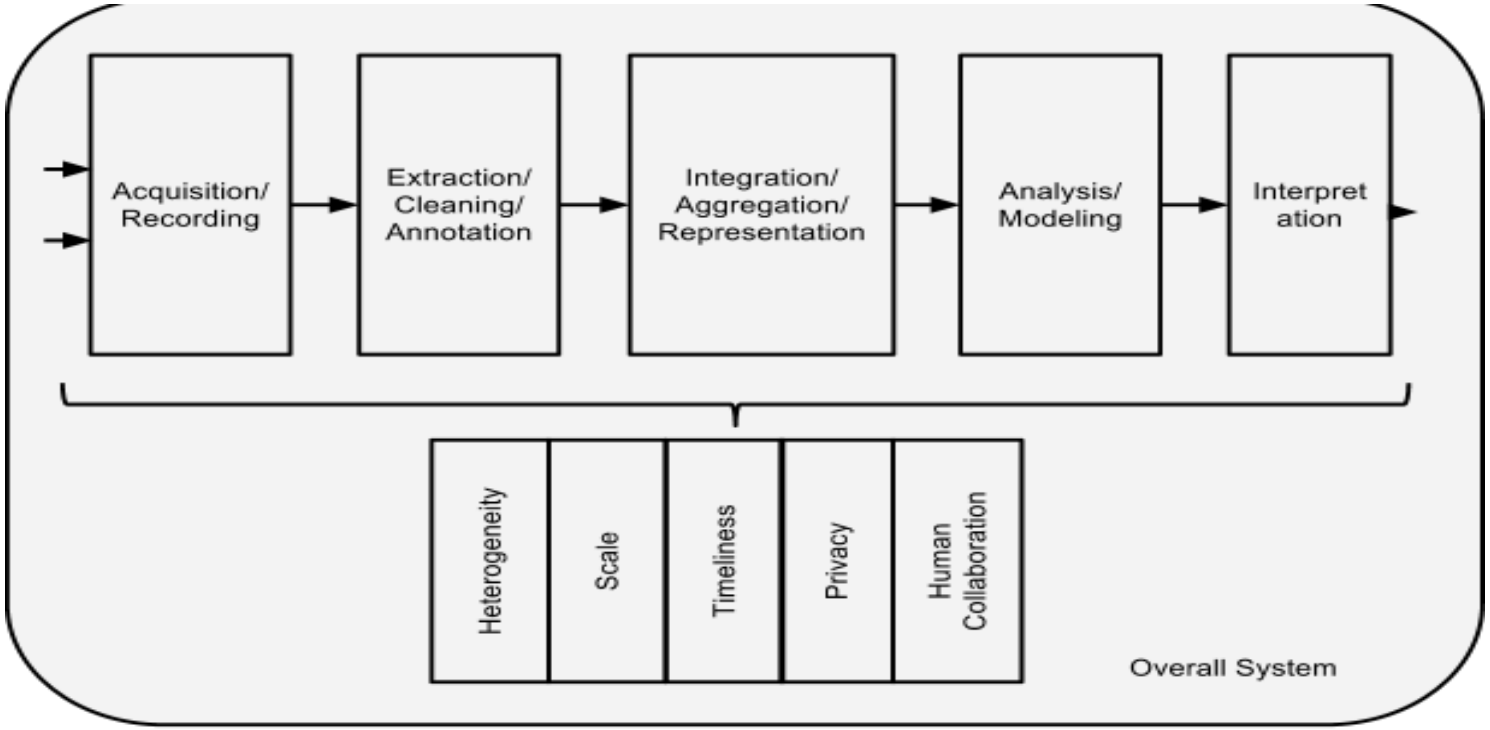
1 + 1

th=1.357

Measurements

Attribute Definition: Features

- Who collected the data? → Error, Bias
- Bias in the Data Preparation process



The Big Data Pipeline (CACM 2014)

Attribute Definition: How to Discretize

- Major accident?
  - Subjective
  - Non-standard forms

Attribute Definition: Sensitive Attributes

- New York Times:
  - “Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago”
  - From 1980 to 2015, the percentages of black, Hispanic students and others have grown.
- Shift in the definition of sensitive attributes:
  - multiracial category was only recently introduced in 2008
  - Many students who might have checked the “white” or “black” box checked the “multiracial” box instead.

Attribute Definition: Target Variable

- Successful Employee
- High-risk for cancer

Proxy attributes

- Example 1: criminal risk assessment

Target variable: who went on to commit a crime

It is hard (or not possible) to check who committed crime.

  - → we use arrests as a proxy

- Example 2: College Admission

Target Variable: Who is expected to be successful

  - → We use GPA as proxy

Stereotypes

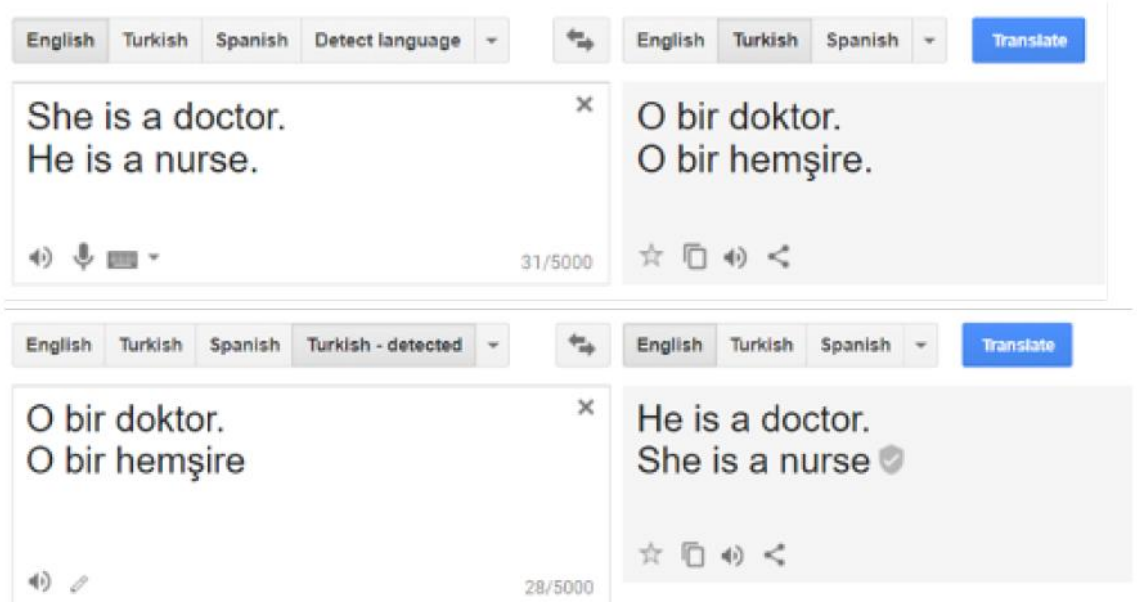
- Some patterns in the training data (smoking is associated with cancer) represent knowledge

- Others (girls like pink and boys like blue) represent stereotypes that we might wish to avoid learning

- Learning algorithms have no general way to distinguish between these two types of patterns, because they are the result of social norms and moral judgments.
    - → models will extract stereotypes that might be harmful

From Data To Representatives (Embeddings)

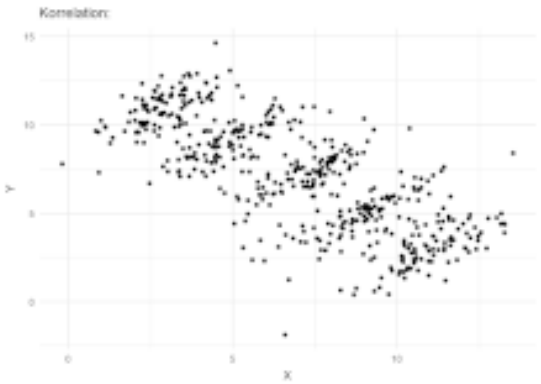
[20] Translating from English to Turkish, then back to English injects gender stereotypes.**



Pitfalls of action, Feedback loops

- Automated Decision Systems (ADS) impact the world
- Issue of Drift
- Correlation v.s. Causation

Associational Fairness can be misleading



- Simpson's Paradox
  - e.g.: UC Berkeley's 1973 Gender Bias case

	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	**44%**	4321	35%

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**

* Image and data are taken from Wikipedia

How to measure the impact of actions?