Future directions of Fairness-aware Data Mining

Recommendation, Causality, and Theoretical Aspects

Toshihiro Kamishima[*1] and **Kazuto Fukuchi**[*2]
joint work with Shotaro Akaho[*1], Hideki Asoh[*1], and Jun Sakuma[*2,3]

[*1]*National Institute of Advanced Industrial Science and Technology (AIST), Japan*
[*2]*University of Tsukuba, and* [*3]*JST CREST*

Workshop on Fairness, Accountability, and Transparency in Machine Learning
In conjunction with the ICML 2015 @ Lille, France, Jul. 11, 2015

I'm Toshihiro Kamishima, and he is Kazuto Fukuchi.
This is joint work with Shotaro Akaho, Hideki Asoh, and Jun Sakuma.
Part of this slide is available at the Slideshare. Please check a Twitter timeline with #icml2015 hash tag.

Outline

Foods for discussion about new directions of fairness in DM / ML

New Applications of Fairness-Aware Data Mining

- Applications of FADM techniques, other than anti-discrimination, especially in a recommendation context

New Directions of Fairness

- Relations of existing formal fairness with causal inference and information theory
- Introducing an idea of a fair division problem and avoiding unfair treatments

Generalization Bound in terms of Fairness

- Theoretical aspects of fairness not on training data, but on test data

✤ We use the term “fairness-aware” instead of “discrimination-aware,” because the word “discrimination” means classification in a ML context, and this technique applicable to tasks other than avoiding discriminative decisions

In this talk, we try to provide foods for discussion about new directions of fairness in a data mining or a machine learning context.
We show these three major topics:
First, we explore applications of FADM techniques, other than anti-discrimination, especially in a recommendation context.
Second, after reviewing relations of existing formal fairness with causal inference and information theory, we present new directions of fairness: Introducing an idea of a fair division problem and avoiding unfair treatments.
Finally, Kazuto Fukuchi will talk about the learning theory, generalization bound in terms of fairness.

PART I

Applications of Fairness-Aware Data Mining

Let's start Part 1: Applications of fairness-aware data mining.

Fairness-Aware Data Mining

Fairness-aware Data mining (FADM)
data analysis taking into account potential issues of fairness

[Romei+ 2014]

Two major tasks of FADM

- **Unfairness Detection:** Finding unfair treatments in database
- **Unfairness Prevention:** Building a model to provide fair outcomes

Unfairness Prevention

- S **: sensitive feature:** representing information that is wanted not to influence outcomes
- **Other factors:** Y: target variable, $\mathbf{X}$: non-sensitive feature

⬇

Learning a statistical model from potentially unfair data sets so that the sensitive feature does not influence the model's outcomes

We begin with what is fairness-aware data mining.
FADM is data analysis taking into account potential issues of fairness.
Here, we focus on a unfairness prevention task.
A sensitive feature represents information that is wanted not to influence outcomes, such as socially sensitive information.
The goal of this unfairness prevention task is to learn a statistical model from potentially unfair data sets so that the sensitive feature does not influence the model's outcomes.

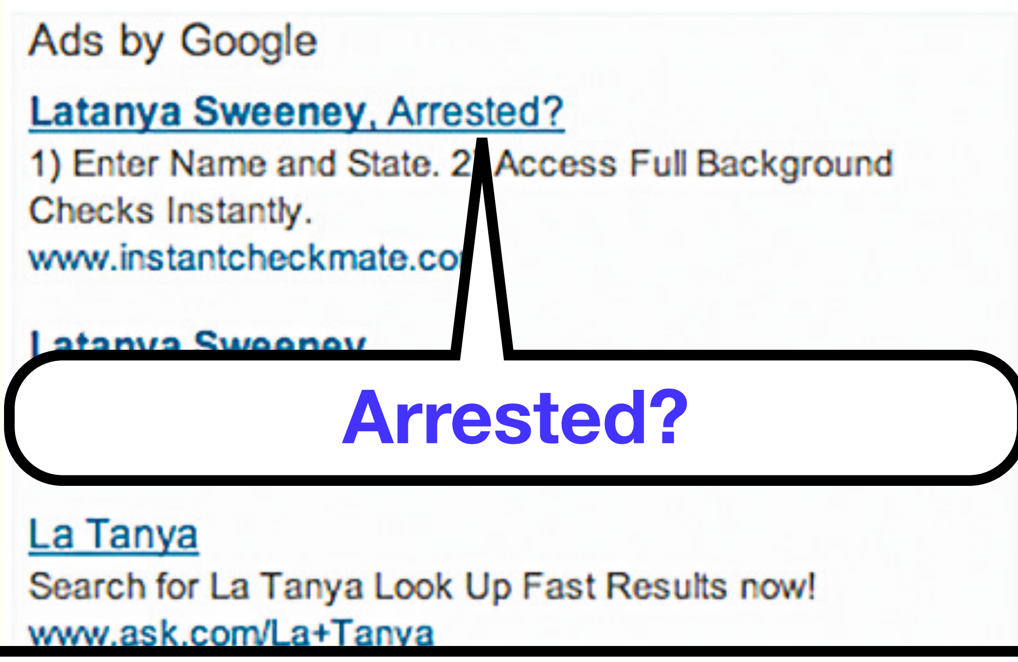
Anti-Discrimination

[Sweeney 13]

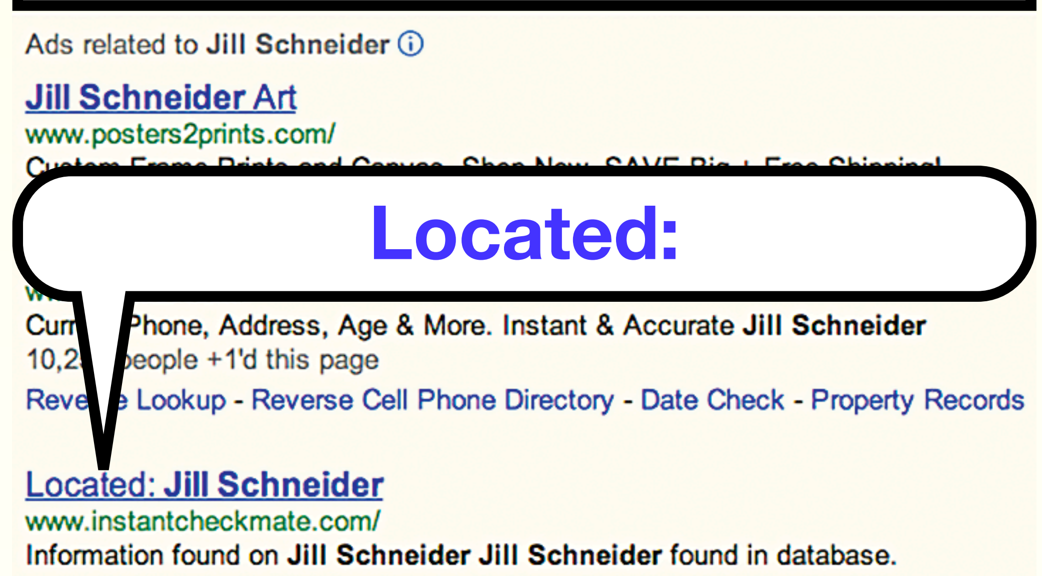
obtaining socially and legally anti-discriminative outcomes

Advertisements indicating arrest records were more frequently displayed for names that are more popular among individuals of African descent than those of European descent

African descent names



European descent names



sensitive feature = users' socially sensitive demographic information

↓

anti-discriminative outcomes

Unfairness prevention methods have been mainly applied to obtain socially and legally anti-discriminative outcomes.
This is Sweeney's well-known case.
We consider that unfairness prevention methods are useful for other types of applications; so, we will show these potential applications.

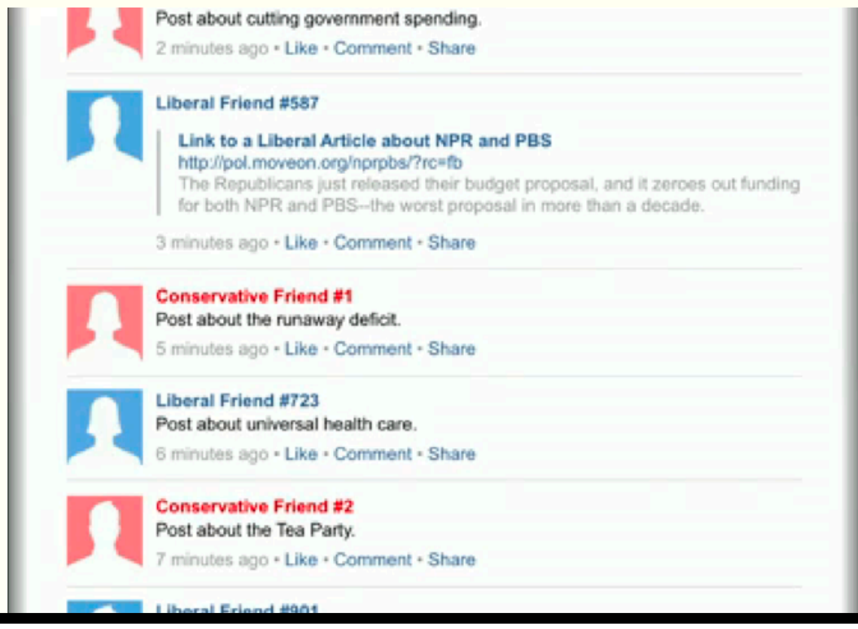
Unbiased Information

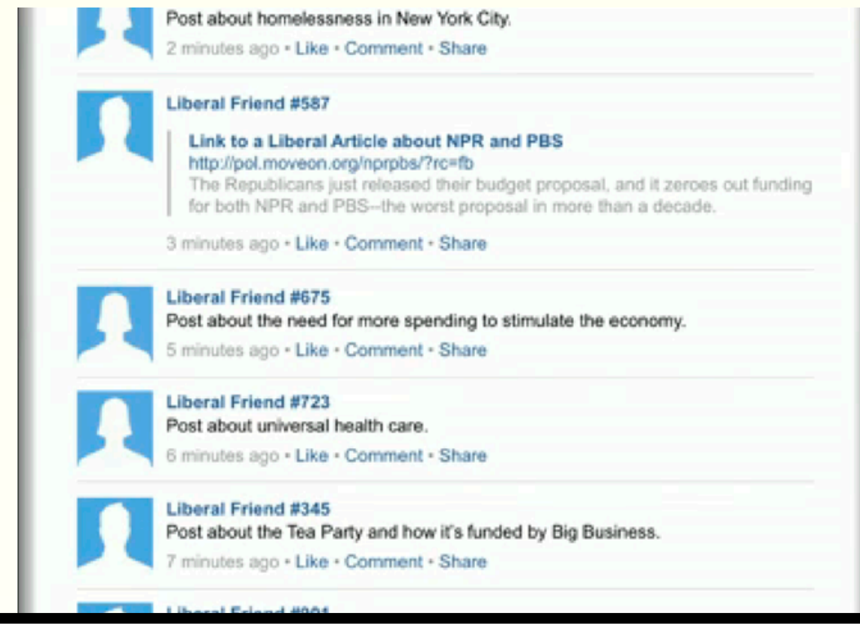
[Pariser 2011, TED Talk by Eli Pariser, http://www.filterbubble.com, Kamishima+ 13]

avoiding biased information that doesn't meet a user's intention

Filter Bubble: a concern that personalization technologies narrow and bias the topics of information provided to people

To fit for Pariser's preference, conservative people are eliminated from his friend recommendation list in a social networking service





sensitive feature = political conviction of a friend candidate

unbiased information in terms of candidates' political conviction

The first application is avoiding biased information that doesn't meet a user's intention.
Pariser show an example of a friend recommendation list.
To fit for his preference, conservative people are eliminated form his friend recommendation list, while this fact is not noticed to him.
In this case, a political conviction of a friend candidate is specified as a sensitive feature.
Then, a recommender will be able to provide unbiased information in terms of candidates' political conviction.

Fair Trading

[Kamishima+ 12, Kamishima+ 13]

equal treatment of content providers

Online retail store

- The site owner directly sells items
- The site is rented to tenants, and the tenants also sells items

In the recommendation on the retail store, the items sold by the site owner are constantly ranked higher than those sold by tenants

⬇

Tenants will complain about this unfair treatment

sensitive feature = a content provider of a candidate item

⬇

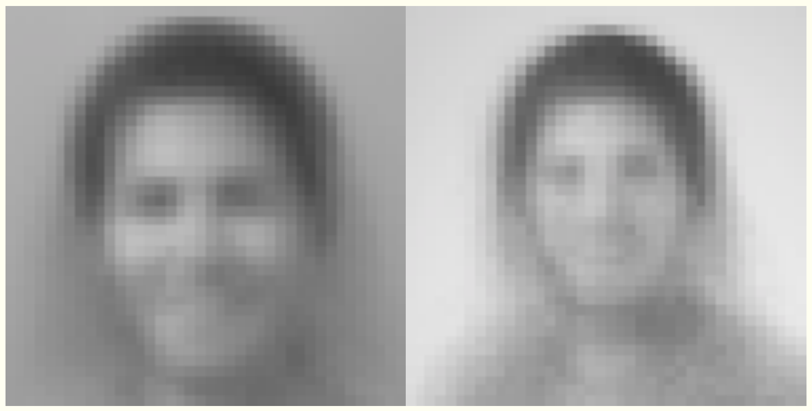
site owner and its tenants are equally treated in recommendation

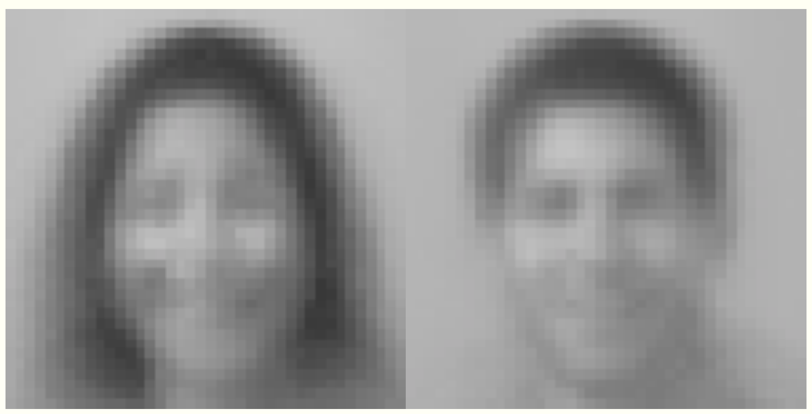
The second application is to encourage fair trading by equal treatment of content providers.
Consider an online retail store.
The site owner directly sells items. Additionally, the site is rented to tenants, and the tenants also sells items.
In the recommendation on the retail store, if the items sold by the site owner are constantly ranked higher than those sold by tenants, then tenants will complain about this unfair treatment.
In this case, a content provider of a candidate item is specified as a sensitive feature.
Then, site owner and its tenants can be equally treated in recommendation by using FADM techniques.

Ignoring Uninteresting Information

[Gondek+ 04]

non-redundant clustering : find clusters that are as independent from a given uninteresting partition as possible





clustering facial images

- A simple clustering method finds two clusters: one contains only faces, and the other contains faces with shoulders
- A data analyst considers this clustering is useless and uninteresting
- By ignoring this uninteresting information, more meaningful female- and male-like clusters could be obtained

sensitive feature = uninteresting information

⬇

ignore the influence of uninteresting information

The third application is ignoring uninteresting information.
This is an example of clustering facial images:
A simple clustering method finds two clusters: one contains only faces, and the other contains faces with shoulders.
A data analyst considers this clustering is useless and uninteresting.
By ignoring this uninteresting information, more useful male and female clusters could be obtained.
In this case, uninteresting information is specified as a sensitive feature.
Ignoring the influence of uninteresting information is helpful for meaningful outcomes.

Part I: Summary

A belief introduction of FADM and a unfairness prevention task

- Learning a statistical model from potentially unfair data sets so that the sensitive feature does not influence the model's outcomes

FADM techniques are widely applicable

- There are many FADM applications other than anti-discrimination, such as providing unbiased information, fair trading, and ignoring uninteresting information

Part 1: summary.
After a belief introduction of FADM and a unfairness prevention task, we explore FADM applications other than anti-discrimination: providing unbiased information, fair trading, and ignoring uninteresting information.

PART II

New Directions of Fairness

Let’s move on to Part 2: new directions of fairness.

PART II: Outline

Discussion about formal definitions and treatments of fairness in data mining and machine learning contexts

Related Topics of a Current Formal Fairness

- connection between formal fairness and causal inference
- interpretation in view of information theory

New Definitions of Formal Fairness

- Why statistical independence can be used as fairness
- Introducing an idea of a fair division problem

New Treatments of Formal Fairness

- methods for avoiding unfair treatments instead of enhancing fairness

In this part, we will discuss formal definitions and treatments of fairness in data mining or machine learning contexts.
We will provide three topics:
First, we review connection between formal fairness and causal inference, and then show their interpretation in view of information theory.
Second, we discuss new direction of formal fairness Introducing an idea of a fair division problem.
Third, we examine methods for avoiding unfair treatments instead of enhancing fairness.

Causality

[Žliobaitė+ 11, Calders+ 13]

Unfairness Prevention task
optimization of accuracy under causality constraints

An example of university admission in [Žliobaitė+ 11]

sensitive feature: S
gender
male / female

$\not\longrightarrow$

target variable: Y
acceptance
accept / not accept

Fair determination: the gender does not influence the acceptance

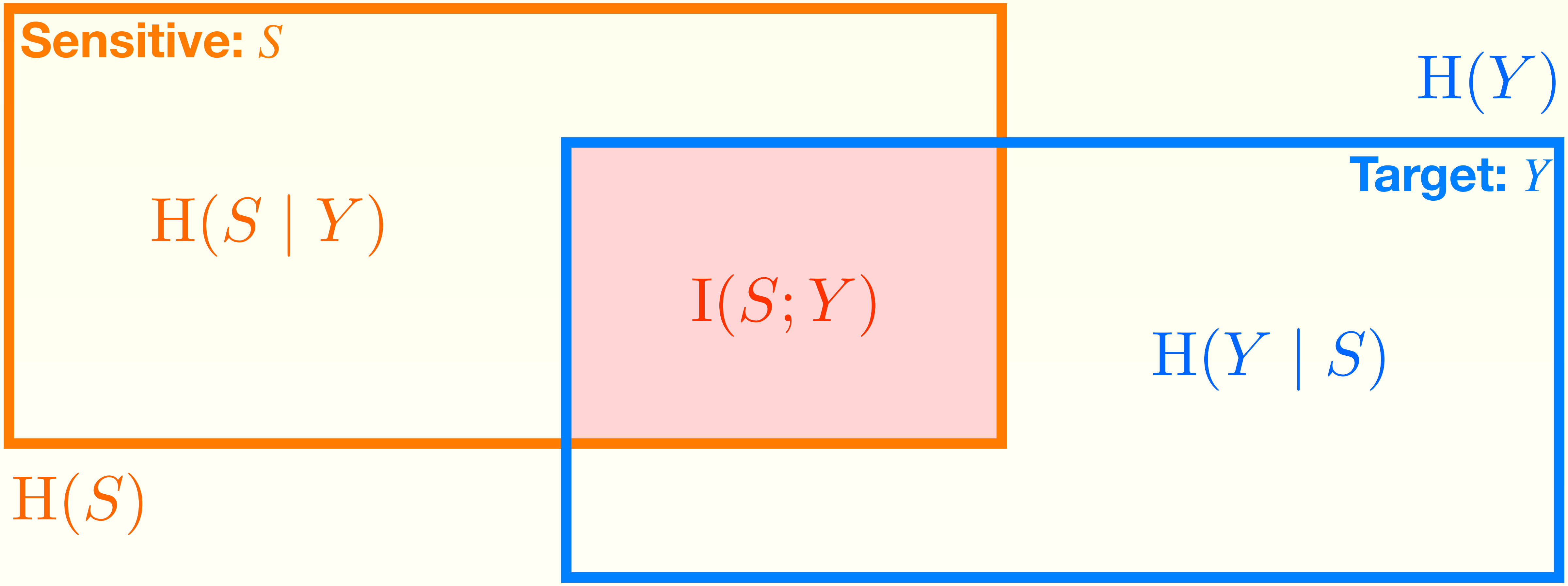
⬇

statistical independence: $Y \perp\!\!\!\perp S$

A unfairness prevention task can be stated as an optimization problem of accuracy under causality constraints.
We therefore explain connection between formal fairness and causal inference using an example of university admission in Žliobaitė's paper.
If the gender does not influence the acceptance, the determination is considered as fair.
Formally, this condition corresponds to statistical dependence between a sensitive feature and a target variable.

Information Theoretic Interpretation

Information theoretical view of a fairness condition



statistical independence between S and Y implies
zero mutual information: $\mathrm{I}(S; Y) = 0$

⬇

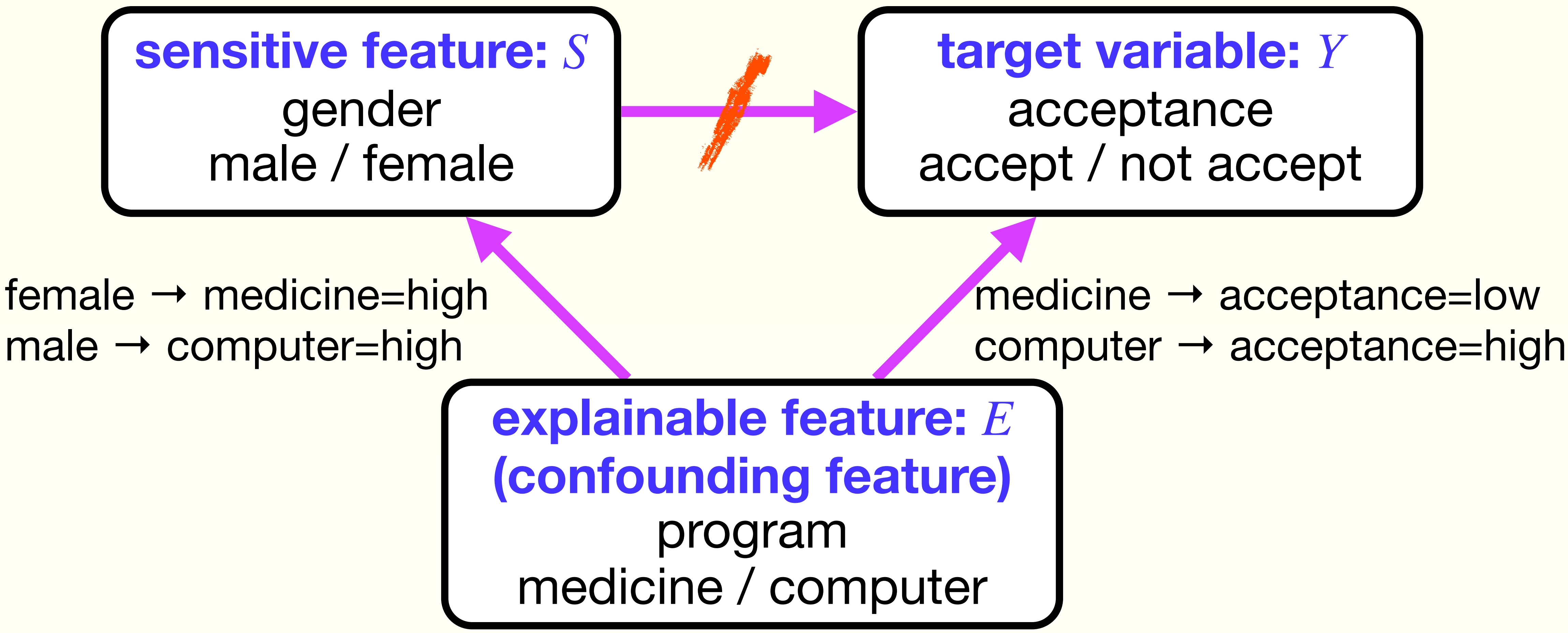
the degree of influence S to Y can be measured by $\mathrm{I}(S; Y)$

This Venn diagram shows information-theoretical view of a fairness condition.
Because statistical independence between S and Y implies zero mutual information, the degree of influence S to Y can be measured by the area of this part; I(S; Y).
Among the total uncertainty about Y, this portion is influenced from S.
To prevent unfairness, we have to reduce this area.

Causality with Explainable Features

[Žliobaitė+ 11, Calders+ 13]

An example of fair determination
even if S and Y are not independent

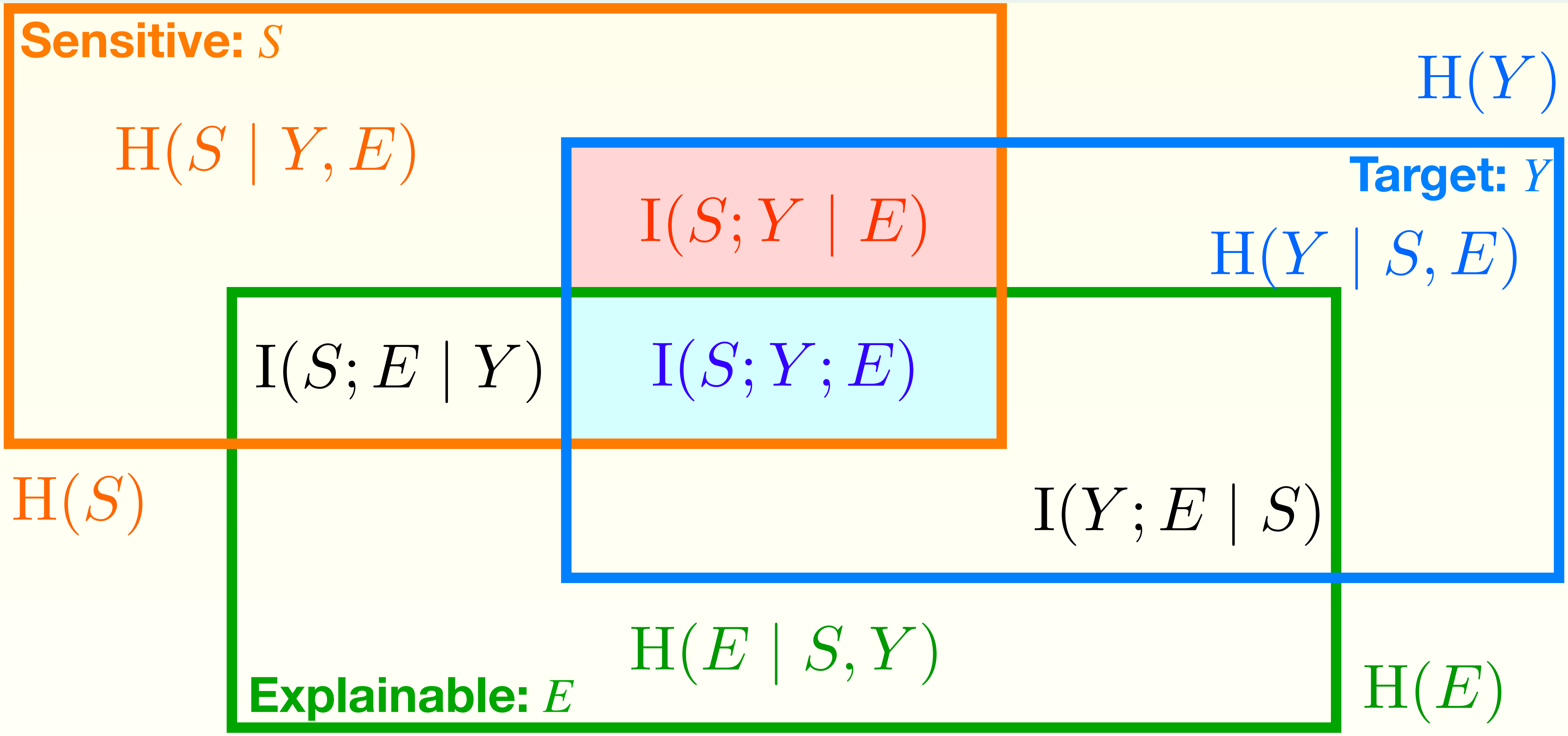


Removing the pure influence of S to Y, excluding the effect of E

conditional statistical independence: $Y \perp\!\!\!\perp S \mid E$

This is an example of fair determination even if S and Y are not independent.
The acceptance ratio of females is lower. However, the determination is still regarded as fair, if this is because females more frequently applied harder programs.
Such a factor that influences both S and Y is called explainable feature in a FADM context and is called confounding feature in Rubin’s causal inference context.
To remove pure influence of S to Y, excluding the effect of E, the conditional statistical independence between Y and S given E have to be satisfied.

Information Theoretic Interpretation



the degree of conditional independence between Y and S given E

conditional mutual information: $\mathrm{I}(S; Y \mid E)$

We can exploit additional information $\mathrm{I}(S; Y; E)$ to obtain outcomes

Again, this Venn diagram shows information-theoretical view of fairness condition.
Because the degree of the independence between S and Y given E, can be measured by conditional mutual information between S and Y given E; I(S; Y|E).
To remove the pure influence of S to Y, we have to reduce this area.
This means that we can exploit additional information I(S; Y; E) to obtain outcomes by adopting explainable features.

Why outcomes are assumed as being fair?

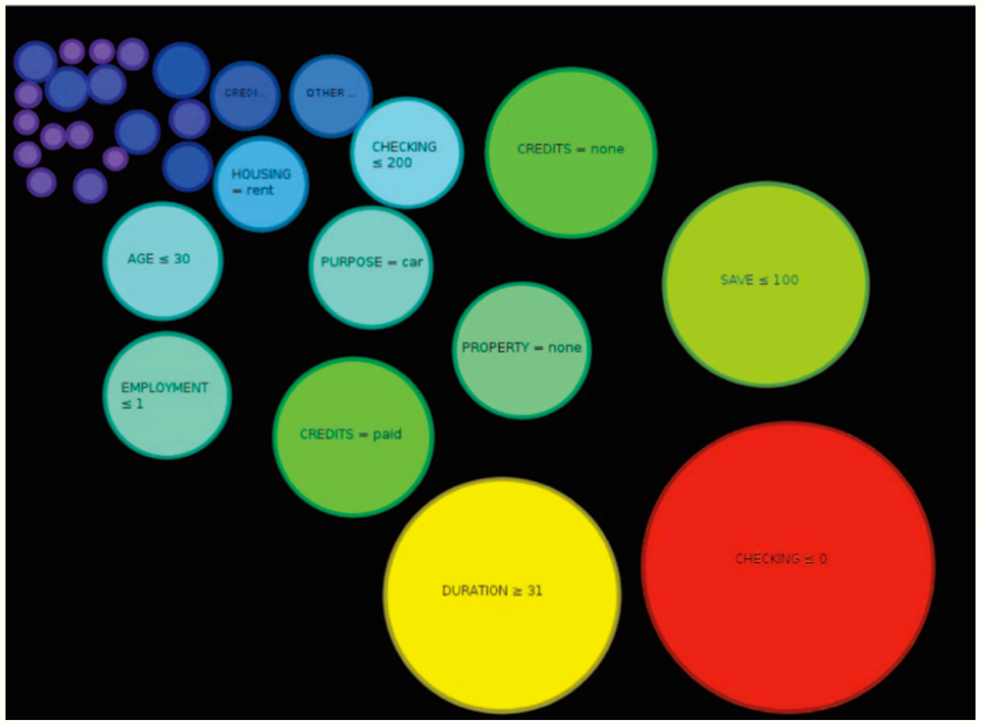
Why outcomes are assumed as being fair,
if a sensitive feature does not influence the outcomes?

⬇

All parties agree with the use of this criterion,
may be because this is objective and reasonable

Is there any way for making an agreement?

- In this view, [Brendt+ 12]'s approach is regarded as a way of making agreements in a wisdom-of-crowds way.
- The size and color of circles indicate the size of samples and the risk of discrimination, respectively

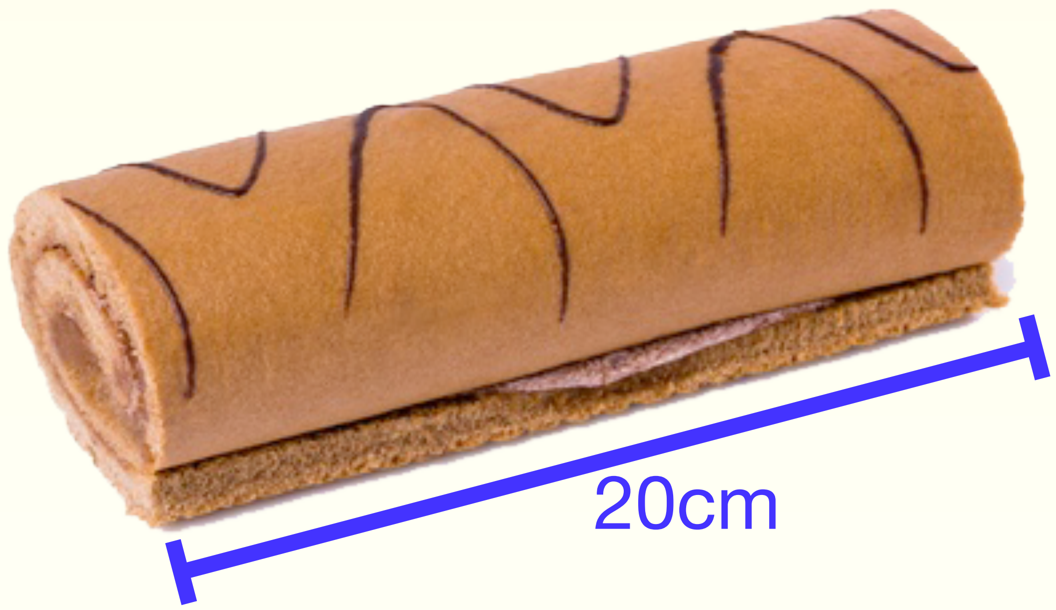


To further examine new directions, we introduce a fair division problem

We have shown connection of fairness with causality and its information theoretical view.
Here, we go back to a fundamental question.
Why outcomes are assumed as fair, if a sensitive feature does not influence the outcomes?
It would be because all parties would agree with the use of this criterion.
Is there any way for making agreements?
In this view, Brendt's approach is regarded as a way of making agreements in a wisdom-of-crowds way.
To further examine new directions, we introduce a fair division problem.

Fair Division

Alice and Bob want to divide this swiss-roll **FAIRLY**



Total length of this swiss-roll is 20cm

Alice and Bob get half each based on **agreed common measure**

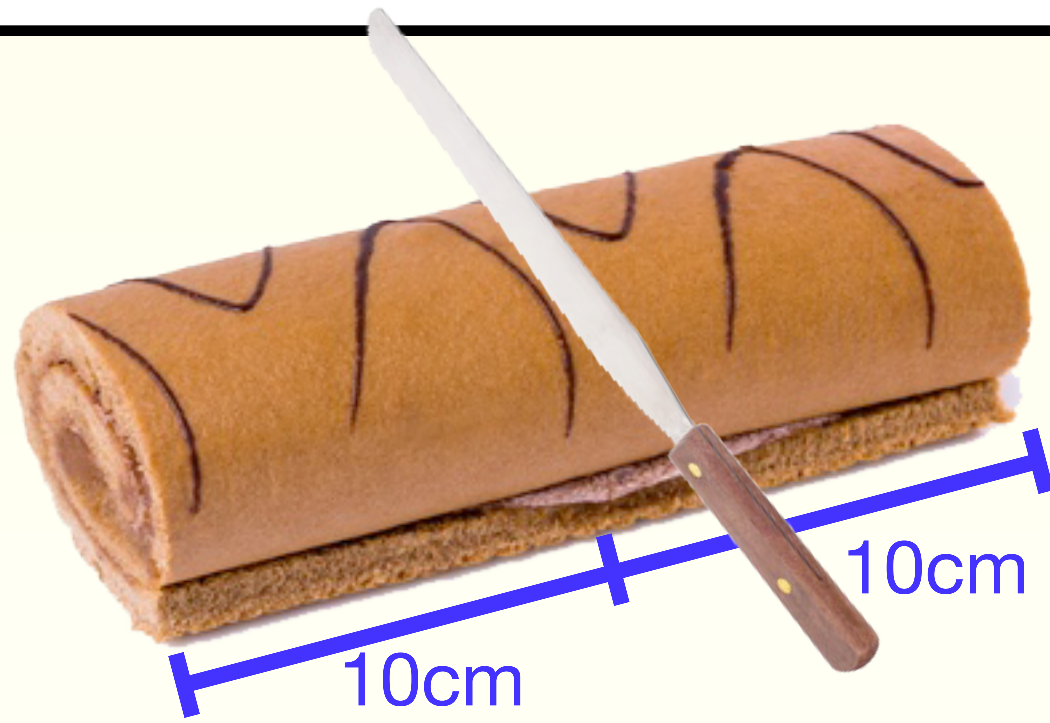
⬇

This approach is adopted in current FADM techniques

Some people in this room may say “this is a manifold,” but this is a swiss-roll cake.
Alice and Bob want to divide this swiss-roll fairly.
A simple method is like this: [PUSH] Total length of this swiss-roll is 20cm. [PUSH] Then, divide the swiss-roll into 10cm each.
This procedure regarded as fair, because Alice and Bob get half each based on agreed common measure.
This approach is adopted in current FADM techniques.

Fair Division

Alice and Bob want to divide this swiss-roll **FAIRLY**



divide the swiss-roll into 10cm each

Alice and Bob get half each based on **agreed common measure**

⬇

This approach is adopted in current FADM techniques

Some people in this room may say “this is a manifold,” but this is a swiss-roll cake.
Alice and Bob want to divide this swiss-roll fairly.
A simple method is like this: [PUSH] Total length of this swiss-roll is 20cm. [PUSH] Then, divide the swiss-roll into 10cm each.
This procedure regarded as fair, because Alice and Bob get half each based on agreed common measure.
This approach is adopted in current FADM techniques.

Fair Division

Unfortunately, Alice and Bob don’t have a scale

Alice cut the swiss-roll exactly in halves based on her own feeling

envy-free division: Alice and Bob get a equal or larger piece based on **their own measure**

Unfortunately, Alice and Bob don’t have a scale. Don’t worry, they don’t need to fight.
[PUSH] First, Alice cut the swiss-roll exactly in halves based on her own feeling.
[PUSH] Then, Bob pick a larger piece based on his own feeling.
Alice believes that the sizes of two pieces are the same, and Bob believes that he get larger one.
This condition is called by envy-free division.

Fair Division

Unfortunately, Alice and Bob don't have a scale

Bob pick a larger piece based on his own feeling

envy-free division: Alice and Bob get a equal or larger piece based on **their own measure**

Unfortunately, Alice and Bob don't have a scale. Don't worry, they don't need to fight.
[PUSH] First, Alice cut the swiss-roll exactly in halves based on her own feeling.
[PUSH] Then, Bob pick a larger piece based on his own feeling.
Alice believes that the sizes of two pieces are the same, and Bob believes that he get larger one.
This condition is called by envy-free division.

Fair Division

- There are n parties
- Every party i has one's own measure $m_i(P_j)$ for each piece P_j

Fairness in a fair division context

- **Envy-Free Division:** Every party gets a equal or larger piece than other parties' pieces based on one's own measure

$$m_i(P_i) \geq m_i(P_j),\ \forall i, j$$

- **Proportional Division:** Every party gets an equal or larger piece than $1/n$ based on one's own measure; Envy-free division is proportional division

$$m_i(P_i) \geq 1/n,\ \forall i$$

- **Exact Division:** Every party gets a equal-sized piece

$$m_i(P_j) = 1/n,\ \forall i, j$$

More formally, there are n parties, and every party has one's own measure for each piece of a cake.
The condition of envy-free division is stated that every party gets a equal or larger piece than other parties' pieces based on one's own measure.
For a fair division problem, the other types of fairness conditions have been discussed, such as proportional division or exact division.

Envy-Free in a FADM Context

Current FADM techniques adopt common agreed measure

⬇

Can we develop FADM techniques using an envy-free approach?

This technique can be applicable without agreements on fairness criterion

FADM under envy-free fairness

Maximize the utility of analysis, such as prediction accuracy, under the envy-free fairness constraints

A Naïve method for Classification

- Among n candidates k ones can be classified as positive
- Among all ${}_{n}C_{k}$ classifications, **enumerate those satisfying envy-free conditions** based on parties' own utility measures
  - ex. Fair classifiers with different sets of explainable features
- Pick the classification whose **accuracy is maximum**

Open Problem: Can we develop a more efficient algorithm?

Can we develop FADM techniques using an envy-free approach?
This technique is highly attractive, because it can applicable without agreements on fairness measures among parties.
If admissions are considered as pieces of cake, they are divided by concerned groups.
This is one possible formulation of FADM under envy-free fairness; Maximize the utility of analysis, such as prediction accuracy, under the envy-free fairness constraints.
A naïve method is an exhaustive search, like this; but, it is practically infeasible.
We leave as an open problem: Can we develop a more efficient algorithm?

Fairness Guardian

Current fairness prevention methods are designed so as **to be fair**

Example: Logistic Regression + Prejudice Remover [Kamishima+ 12]

The objective function is composed of classification loss and fairness constraint terms

$$\boxed{-\sum_{\mathcal{D}} \ln \Pr[Y \mid \mathbf{X}, S; \Theta] + \frac{\lambda}{2} \|\Theta\|_2^2} + \eta \boxed{\mathrm{I}(Y; S)}$$

Fairness Guardian Approach

Unfairness is prevented by enhancing fairness of outcomes

We have discussed new directions of fairness criteria. We then examine how to treat fairness.
Current fairness prevention methods are designed so as to be fair.
This is an example of our logistic regression with a prejudice remover regularizer.
The objective function is composed of classification loss and fairness constraint terms.
Here, we call this approach by a fairness guardian.
Unfairness is prevented by enhancing fairness of outcomes.

Fair Is Not Unfair?

A reverse treatment of fairness:
not to be unfair

One possible formulation of a unfair classifier
Outcomes are determined ONLY by a sensitive feature

$$\Pr[Y \mid S;\ \mathbf{\Psi}^*]$$

Ex. Your paper is rejected, just because you are not handsome

Penalty term to maximize the KL divergence between
a pre-trained unfair classifier and a target classifier

$$D_{\mathrm{KL}}[\Pr[Y \mid S;\ \mathbf{\Psi}^*] \| \Pr[Y \mid X, S; \mathbf{\Theta}]]$$

We here explore a reverse treatment of fairness: not to be unfair.
For this purpose, we built unfair classifier.
This would be one possible formulation of a unfair classifier
Outcomes are determined ONLY by a sensitive feature.
For example, your paper is rejected, just because you are not handsome.
To avoid this unfair classifier, we tested penalty term to maximize the KL divergence between a pre-trained unfair classifier and a target classifier.

Unfairness Hater

Unfairness Hater Approach
Unfairness is prevented by avoiding unfair outcomes

This approach was almost useless for obtaining fair outcomes, but…

- **Better Optimization**
  The fairness-enhanced objective function tends to be non-convex; thus, adding a unfairness hater may help for avoiding local minima
- **Avoiding Unfair Situation**
  There would be unfair situations that should be avoided;
  Ex. Humans' photos were mistakenly labeled as gorilla in auto-tagging [Barr 2015]

There would be many choices between to be fair and not to be unfair that should be examined

We call this approach by a unfairness hater.
Unfortunately, this approach was almost useless for obtaining fair outcomes.
However, this approach may be useful for these situations:
First, better optimization: The fairness-enhanced objective function tends to be non-convex; thus, adding a unfairness hater may help for avoiding local minima.
Second, There would be unfair situations that should be avoided; For example, humans' photos were mistakenly labeled as gorilla in auto-tagging.
There would be many choices between to be fair and not to be unfair that should be examined.

Part II: Summary

Relation of fairness with causal inference and information theory

- We review a current formal definition of fairness by relating it with Rubin's causal inference; and, its interpretation based on information theory

New Directions of formal fairness without agreements

- We showed the possibility of formal fairness that does not presume a common criterion agreed between concerned parties

New Directions of treatment of fairness by avoiding unfairness

- We discussed that FADM techniques for avoiding unfairness, instead of enhancing fairness.

Part 2: summary.
In this part, we showed three topics:
First, we review a current formal definition of fairness by relating it with Rubin's causal inference; and, its interpretation based on information theory.
Second, We showed the possibility of formal fairness that does not presume a common criterion agreed between concerned parties.
Third, We discussed that FADM techniques for avoiding unfairness, instead of enhancing fairness.

PART III
Generalization Bound in terms of Fairness

Part 3: Generalization bound in terms of fairness.

Part III: Introduction

There are many technical problems to solve in a FADM literature, because tools for excluding specific information has not been developed actively.

- **Types of Sensitive Features**
  Non-binary sensitive feature
- **Analysis Techniques**
  Analysis methods other than classification or regression
- **Optimization**
  Constraint terms make objective functions non-convex
- **Fairness measure**
  Interpretable to humans and having convenient properties
- **Learning Theory**
  Generalization ability in terms of fairness

There are many technical problems to solve in a FADM literature, because tools for excluding specific information has not been developed actively.
Among these problems, Kazuto Fukuchi will talk about generalization ability in terms of fairness

Kazuto Fukuchi's Talk

Conclusion

Applications of Fairness-Aware Data Mining

- Applications other than anti-discrimination: providing unbiased information, fair trading, and excluding unwanted information

New Directions of Fairness

- Relation of fairness with causal inference and information theory
- Formal fairness introducing an idea of a fair division problem
- Avoiding unfair treatment, instead of enhancing fairness

Generalization bound in terms of fairness

- Generalization bound in terms of fairness based on f-divergence

Additional Information and codes

http://www.kamishima.net/fadm

Acknowledgments: This work is supported by MEXT/JSPS KAKENHI Grant Number 24500194, 24680015, 25540094, 25540094, and 15K00327

This is a summary of our talk.
That’s all I have to say. Thank you for your attention.

Bibliography I

A. Barr.
Google mistakenly tags black people as ‘gorillas,’ showing limits of algorithms.
The Wall Street Journal, 2015.
⟨http://on.wsj.com/1CaCNlb⟩.

B. Berendt and S. Preibusch.
Exploring discrimination: A user-centric evaluation of discrimination-aware data mining.
In *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pages 344–351, 2012.

T. Calders, A. Karim, F. Kamiran, W. Ali, and X. Zhang.
Controlling attribute effect in linear regression.
In *Proc. of the 13th IEEE Int'l Conf. on Data Mining*, pages 71–80, 2013.

T. Calders and S. Verwer.
Three naive Bayes approaches for discrimination-free classification.
Data Mining and Knowledge Discovery, 21:277–292, 2010.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel.
Fairness through awareness.
In *Proc. of the 3rd Innovations in Theoretical Computer Science Conf.*, pages 214–226, 2012.

Bibliography II

- K. Fukuchi and J. Sakuma.
  Fairness-aware learning with restriction of universal dependency using f-divergences.
  arXiv:1104.3913 [cs.CC], 2015.

- D. Gondek and T. Hofmann.
  Non-redundant data clustering.
  In *Proc. of the 4th IEEE Int'l Conf. on Data Mining*, pages 75–82, 2004.

- T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma.
  Considerations on fairness-aware data mining.
  In *Proc. of the IEEE Int'l Workshop on Discrimination and Privacy-Aware Data Mining*, pages 378–385, 2012.

- T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma.
  Enhancement of the neutrality in recommendation.
  In *Proc. of the 2nd Workshop on Human Decision Making in Recommender Systems*, pages 8–14, 2012.

- T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma.
  Fairness-aware classifier with prejudice remover regularizer.
  In *Proc. of the ECML PKDD 2012, Part II*, pages 35–50, 2012.
  [LNCS 7524].

Bibliography III

T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma.
Efficiency improvement of neutrality-enhanced recommendation.
In *Proc. of the 3rd Workshop on Human Decision Making in Recommender Systems*, pages 1–8, 2013.

E. Pariser.
The filter bubble.
⟨http://www.thefilterbubble.com/⟩.

E. Pariser.
The Filter Bubble: What The Internet Is Hiding From You.
Viking, 2011.

A. Romei and S. Ruggieri.
A multidisciplinary survey on discrimination analysis.
The Knowledge Engineering Review, 29(5):582–638, 2014.

L. Sweeney.
Discrimination in online ad delivery.
Communications of the ACM, 56(5):44–54, 2013.

I. Žliobaitė, F. Kamiran, and T. Calders.
Handling conditional discrimination.
In *Proc. of the 11th IEEE Int'l Conf. on Data Mining*, 2011.

Bibliography IV

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork.
Learning fair representations.
In *Proc. of the 30th Int'l Conf. on Machine Learning*, 2013.