Bayesian Networks

Week 10

April 1, 2024

Spring 24 | CUSP-GX 7033 – Machine Learning for Cities | Dr. Anton Rozhkov

Today’s Outline

Last class

- Definitions of fairness and bias
- The need for fairness in algorithms: motivation and examples
- Fairness-aware algorithms.
  - Preventing disparate impact: a case study in criminal justice

Today's Outline

Last class

- Definitions of fairness and bias
- The need for fairness in algorithms: motivation and examples
- Fairness-aware algorithms.
  - Preventing disparate impact: a case study in criminal justice
  - **Group fairness: tweaking ML algorithms to prevent discrimination**
  - **Calibration: Detecting and fixing systematic biases in risk prediction**
- **Bayesian Networks – introduction**
- **Lab: Bayesian Networks**

Algorithmic Fairness
(cont)

Websites Vary Prices, Deals Based on Users' Information

By JENNIFER VALENTINO-DEVRIES, JEREMY SINGER-VINE and ASHKAN SOLTANI
December 24, 2012

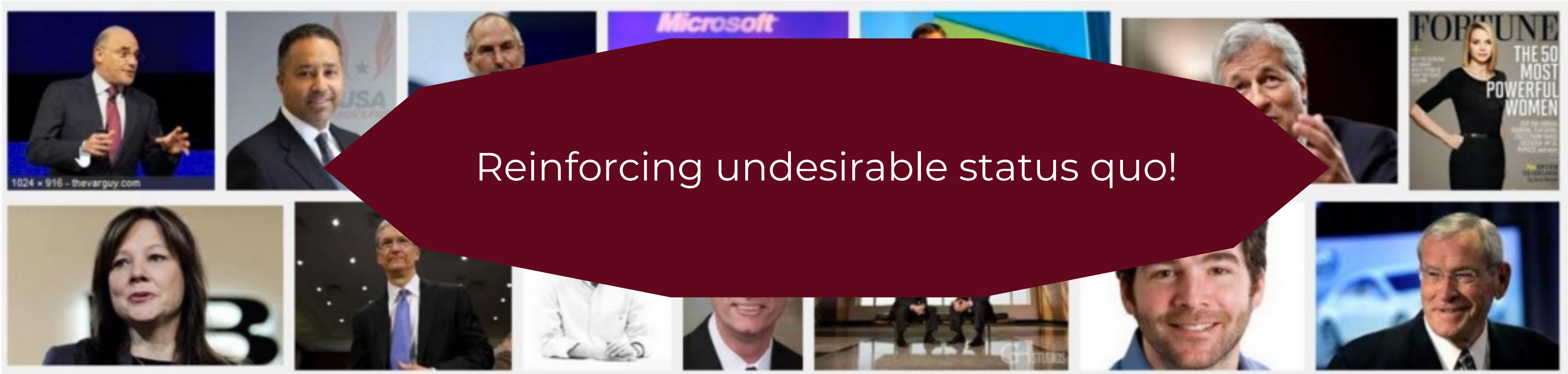
It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was $15.79, while the price on Trude Frizzell's screen, just a few miles away, was $14.29.

A key difference: where Staples seemed to think they were located.

Unforeseen consequences!

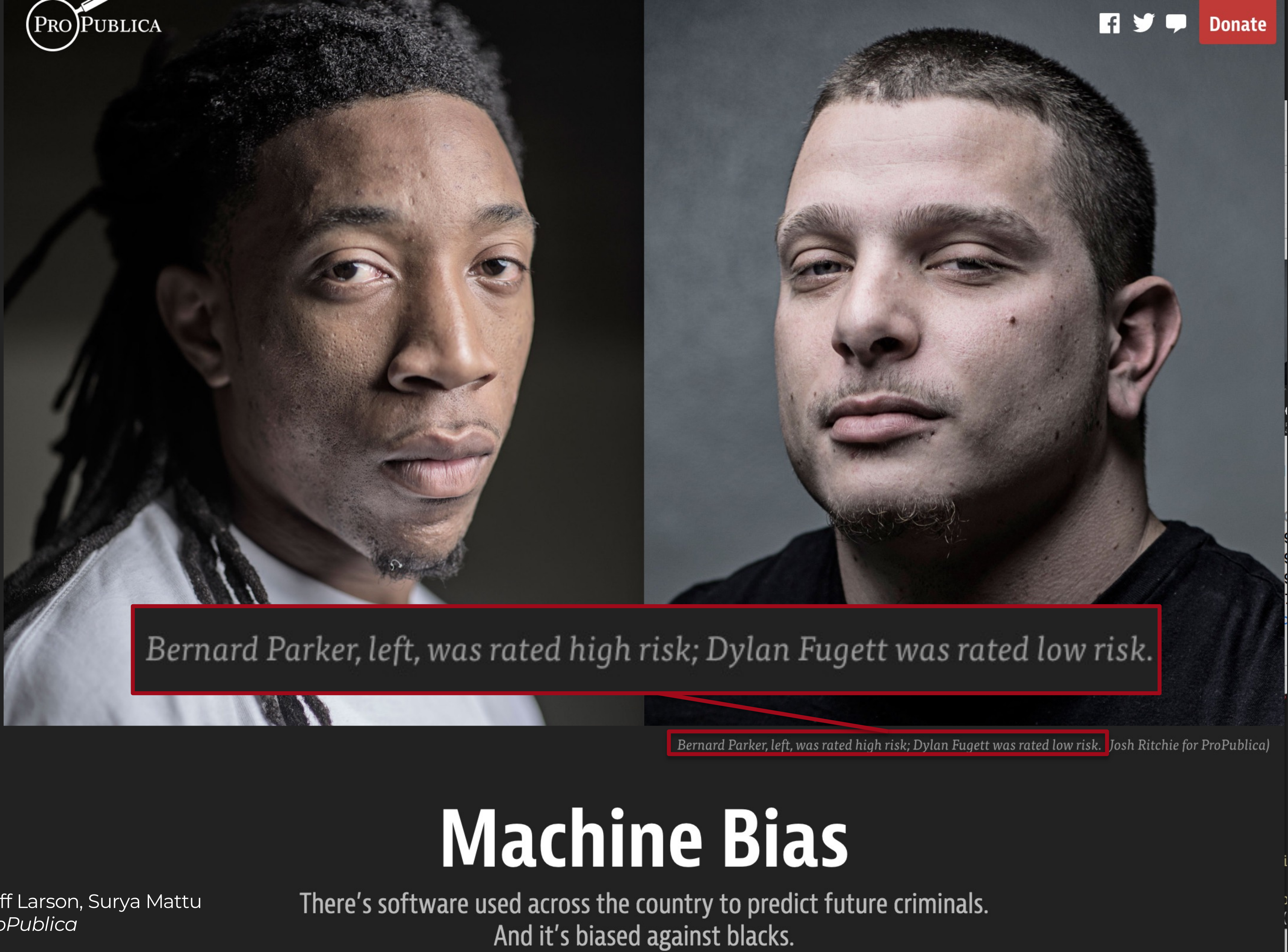
http://www.wsj.com/articles/SB10001424127887323777204578189391813881534

<u>What happened</u>: lower store density in poor & ethnic minority neighborhoods → higher prices → racially disparate impact.



Reinforcing undesirable status quo!

IMAGE: PERCENTAGE OF WOMEN IN TOP 100 GOOGLE IMAGE SEARCH RESULTS FOR CEO IS: 11 PERCENT. PERCENTAGE OF US CEOS WHO ARE WOMEN IS: 27 PERCENT. view more ›



Source: Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, *ProPublica*

GROUP FAIRNESS: TWEAKING ML ALGORITHMS TO PREVENT DISCRIMINATION

NYU

Statistical Parity

or “group fairness”: an entirely different notion of fairness

$$\mathbb{P}(\ \text{hired} \mid \text{man}\ ) = \mathbb{P}(\ \text{hired} \mid \text{woman})$$

- Reasonable fairness criterion in settings such as **employment**
  - *Not reasonable in risk assessment*

- Lots of recent work on constructing models that satisfy statistical parity

- **Caution:**
  - Statistical parity shouldn’t be an end in itself
  - E.g., Could just hire top 10% of men and a random 10% of women
  - *Self-fulfilling prophecy* of discrimination (Dwork et. al.)

NYU

Discrimination

“Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category. Justified distinctions are exceptions explicitly admitted by law, such as imposing a minimum age for voting in elections, or that are proven (sometimes in court) as being objective and legitimate, such as requiring a man for a male character in a film. Some groups, traditionally subject to discrimination, are explicitly listed as ‘protected groups’ by national and international human rights laws.”

from “A multidisciplinary survey on discrimination analysis”

How Algorithms Can Discriminate

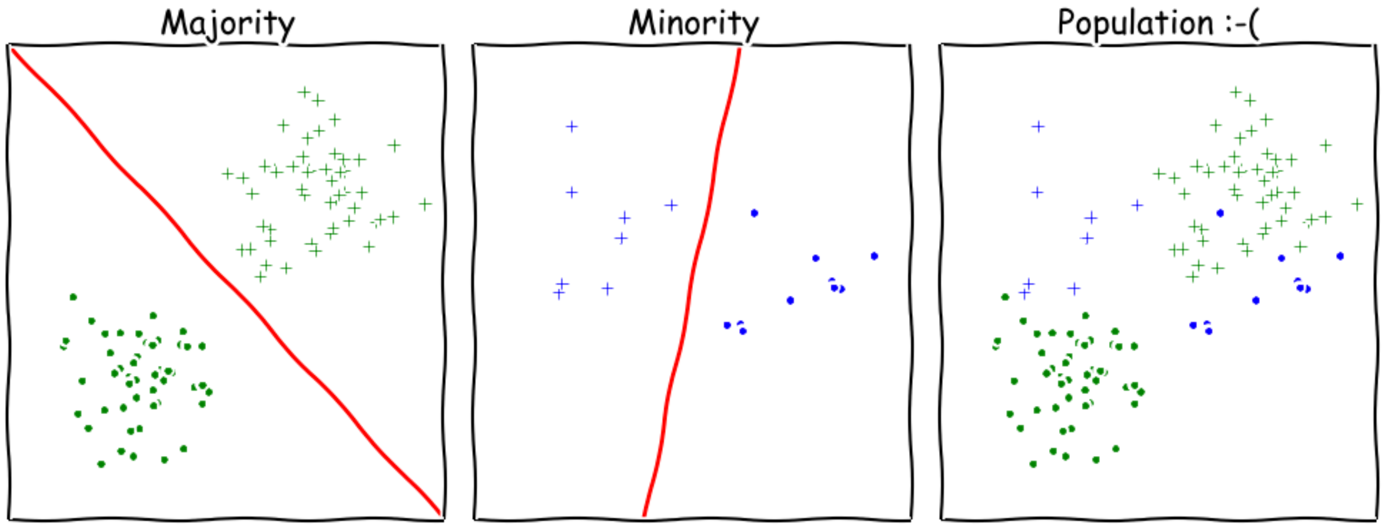
- Problem specification (definition of target variable, features, etc.)

e.g. electing to hire on the basis of predicted tenure can be more likely to have a disparate impact on certain protected classes than hiring decisions that turn on some estimate of worker productivity

How Algorithms Can Discriminate

- Problems with training data:

– the data generating process itself was inherently discriminatory
– biased/imbalanced data samples



How Algorithms Can Discriminate

- Via proxies

“when the criteria that are genuinely relevant in making rational and well-informed decisions also happen to serve as reliable proxies for class membership. In other words, the very same criteria that correctly sort individuals according to their predicted likelihood of excelling at a job may also sort individuals according to class membership”

How Algorithms Can Discriminate

via proxies

hidden attribute

Customer no.	Gender	Age	Hp	Driving style	Risk
#1	Male	30 years	High	Aggressive	+
#2	Male	35 years	Low	Aggressive	-
#3	Female	24 years	Med.	Calm	-
#4	Female	18 years	Med.	Aggressive	+
#5	Male	65 years	High	Calm	-
#6	Male	54 years	Low	Aggressive	+
#7	Female	21 years	Low	Calm	-
#8	Female	29 years	Med.	Calm	-

NYU

from “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures,” Calders & Žliobaitė

Simple Fixes that Don’t Work

removing the sensitive attribute

Customer no.	~~Ethnicity~~	Work exp.	Postal code	Loan decision
#1	~~European~~	12 years	1212	+
#2	~~Asian~~	2 years	1010	-
#3	~~European~~	5 years	1221	+
#4	~~Asian~~	10 years	1011	-
#5	~~European~~	10 years	1200	+
#6	~~Asian~~	5 years	1001	-
#7	~~European~~	12 years	1212	+
#8	~~Asian~~	2 years	1010	-

from “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures,” Calders & Žliobaitė

Simple Fixes that Don't Work

Building separate models for each value of the sensitive attribute

Applicant no.	Gender	Test score	Level	Acceptance
#1	Male	82	A	+
#2	Female	85	A	+
#3	Male	75	B	+
#4	Female	75	B	-
#5	Male	65	A	-
#6	Female	62	A	-
#7	Male	91	B	+
#8	Female	81	B	+

from “Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures,” Calders & Žliobaitė

Adverse Affect and the 80% Rule

Adverse effect refers to a total employment process which results in a significantly higher percentage of a protected group in the candidate population being rejected for employment, placement, or promotion. The difference between the rejection rates for a protected group and the remaining group must be statistically significant at the .05 level. In addition, if the acceptance rate of the protected group is greater than or equal to 80% of the acceptance rate of the remaining group, then adverse effect is said to be not present by definition (Section 7.1).

from Biddle

Adverse Affect and the 80% Rule

Definition 1.1 (Disparate Impact ("80% rule")). *Given data set* $D = (X, Y, C)$, *with* protected attribute X *(e.g., race, sex, religion, etc.), remaining attributes* Y, *and binary class to be predicted* C *(e.g., "will hire"), we will say that* D *has* disparate impact *if*

$$\frac{\Pr(C = YES|X = 0)}{\Pr(C = YES|X = 1)} \leq \tau = 0.8$$

for positive outcome class YES *and majority protected attribute* 1 *where* $\Pr(C = c|X = x)$ *denotes the conditional probability (evaluated over* D*) that the class outcome is* $c \in C$ *given protected attribute* $x \in X$.[1]

from "Certifying and removing disparate impact",

Adverse Affect and the 80% Rule

consider a classifier defined by a **decision boundary;**

to satisfy the 80% rule our classifier must satisfy

$$\frac{P(d_\theta(\mathbf{x}) > 0 | X = 0)}{P(d_\theta(\mathbf{x}) > 0 | X = 1)} \geq 0.80$$

Unfortunately, this isn't a very well-behaved objective function
(as a function of the parameters theta)

NYU

One Quantitative Learning Approach to Fairness: Decision Boundary Covariance

decision boundary

$$
\begin{aligned}
\mathrm{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})]\bar{d}_{\boldsymbol{\theta}}(\mathbf{x}) \\
&= \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] \\
&\approx \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i),
\end{aligned}
$$

sensitive class

"Learning Fair Classifiers." Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, Krishna P. Gummadi

http://arxiv.org/pdf/1507.05259v3.pdf

Decision Boundary Covariance

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i - \bar{\mathbf{z}})\,\boldsymbol{\theta}^T\mathbf{x}_i$$

this will serve as our measure of unfairness

making this small does NOT guarantee that the 80% rule will be satisfied; but as we will see, in practice a small value of the covariance will typically lead to a balanced ratio of

$$\frac{P(d_\theta(\mathbf{x}) > 0 | X = 0)}{P(d_\theta(\mathbf{x}) > 0 | X = 1)}$$

Maximizing Accuracy Under Fairness Constraints

$$p(y_i = 1|\mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + e^{\boldsymbol{\theta}^T \mathbf{x}_i}}$$

our constrained optimization problem
with fairness constraints:

$$\begin{array}{ll} \text{minimize} & -\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \leq \mathbf{c}, \\ & \frac{1}{N}\sum_{i=1}^{N} (\mathbf{z}_i - \bar{\mathbf{z}})\, \boldsymbol{\theta}^T \mathbf{x}_i \geq -\mathbf{c}. \end{array}$$

logistic regression case

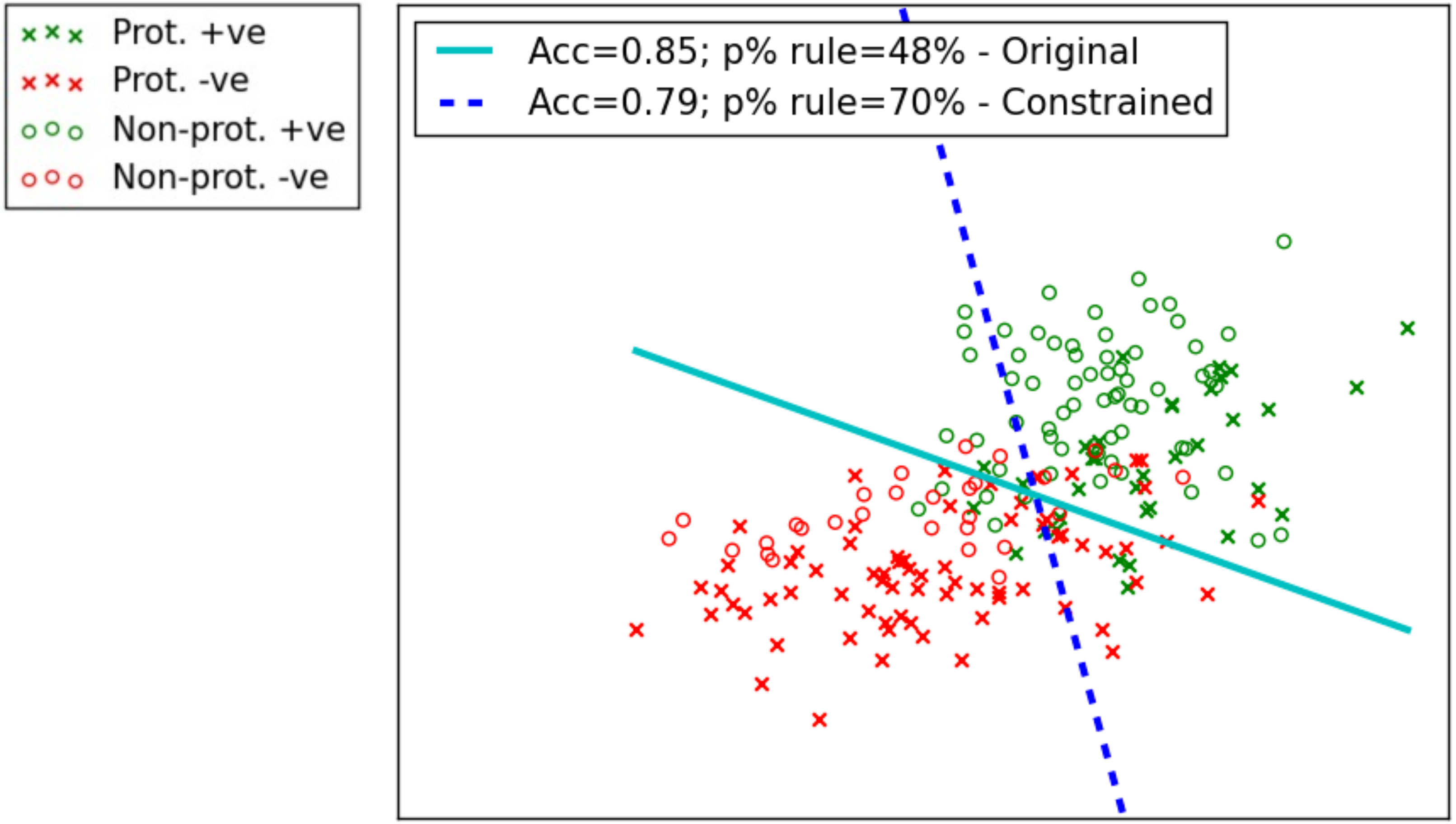
NYU

Maximizing Accuracy Under Fairness Constraints

$$
\begin{array}{ll}
\text{minimize} & \|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i \\
\text{subject to} & y_i\boldsymbol{\theta}^T\mathbf{x}_i \geq 1 - \xi_i, \forall i \in \{1,\ldots,n\} \\
& \xi_i \geq 0, \forall i \in \{1,\ldots,n\}, \\
& \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \bar{\mathbf{z}}\right)\boldsymbol{\theta}^T\mathbf{x}_i \leq \mathbf{c}, \\
& \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{z}_i - \bar{\mathbf{z}}\right)\boldsymbol{\theta}^T\mathbf{x}_i \geq -\mathbf{c},
\end{array}
$$

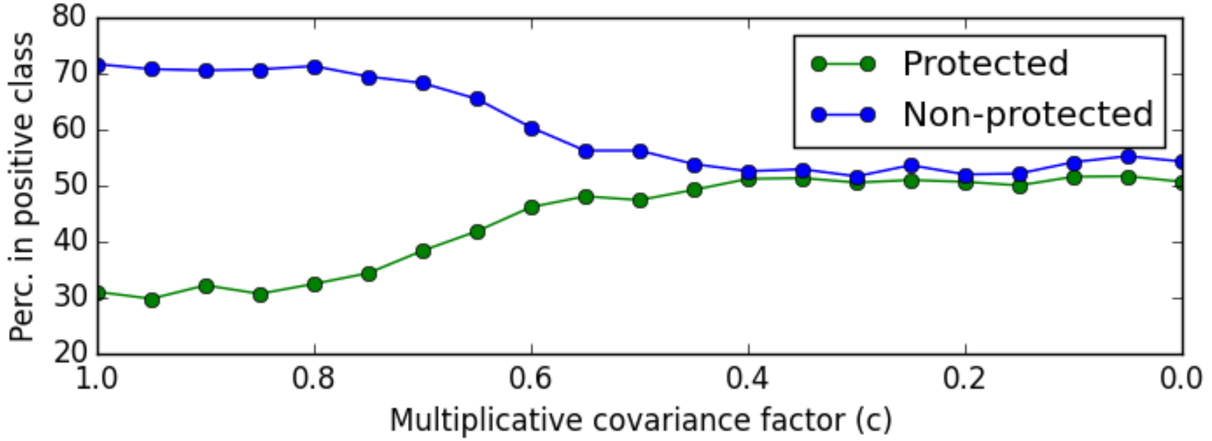
linear SVM case

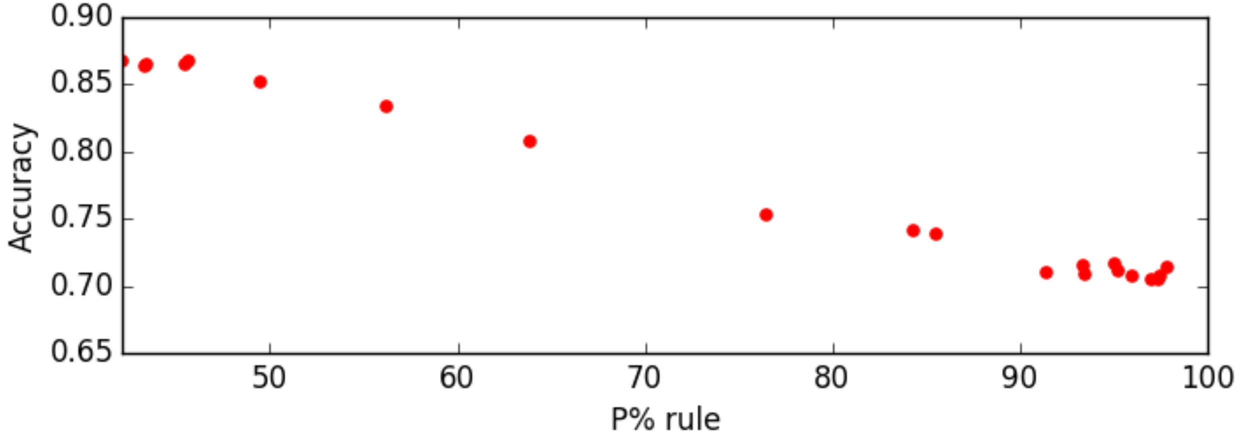
Maximizing Accuracy Under Fairness Constraints



https://github.com/mbilalzafar/fair-classification

Maximizing Accuracy Under Fairness Constraints





NYU

CALIBRATION: DETECTING AND FIXING SYSTEMATIC BIASES IN RISK PREDICTION

NYU

Another Perspective

Whether our goal is to achieve group fairness or reduce disparate impacts, our first step should be to predict risk as accurately as possible.

In particular, they wish to **detect** and **correct** any systematic **biases** in risk prediction that a classifier may have (i.e., over-predicting or under-predicting risk for a specific attribute or combination of attributes).

Thus, a new **subset scan** method was developed to identify subgroups where classifier predictions are significantly biased (Zhang & Neill, 2016).

Assume a dataset with inputs x_i, binary labels $y_i \in \{0,1\}$, and the classifier's risk predictions $\hat{p}_i = \Pr(y_i = 1)$.

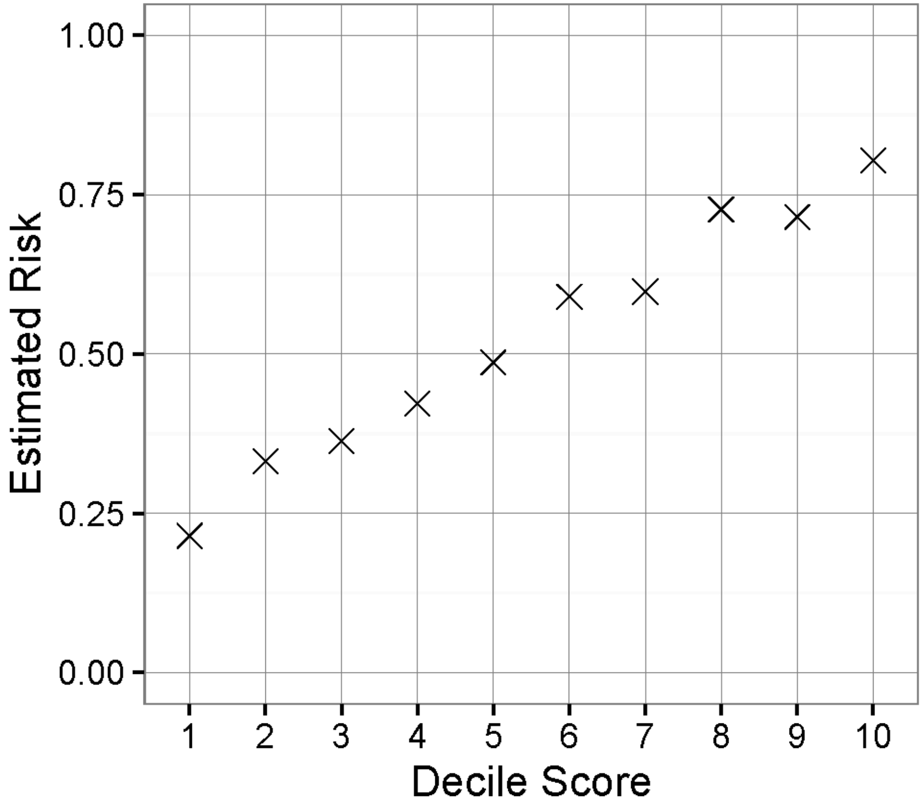
Search space: subspaces defined by a subset of values for each attribute (e.g., "white and Asian males under 25")

Score function: a log-likelihood ratio statistic. H_0: $\hat{p}_i$ correctly calibrated; $H_1(S)$: constant multiplicative increase or decrease in odds of $y_i = 1$ for subspace S.

NYU

Source: *the approach* used by Dr. Daniel Neill

Results of Bias Scan on COMPAS

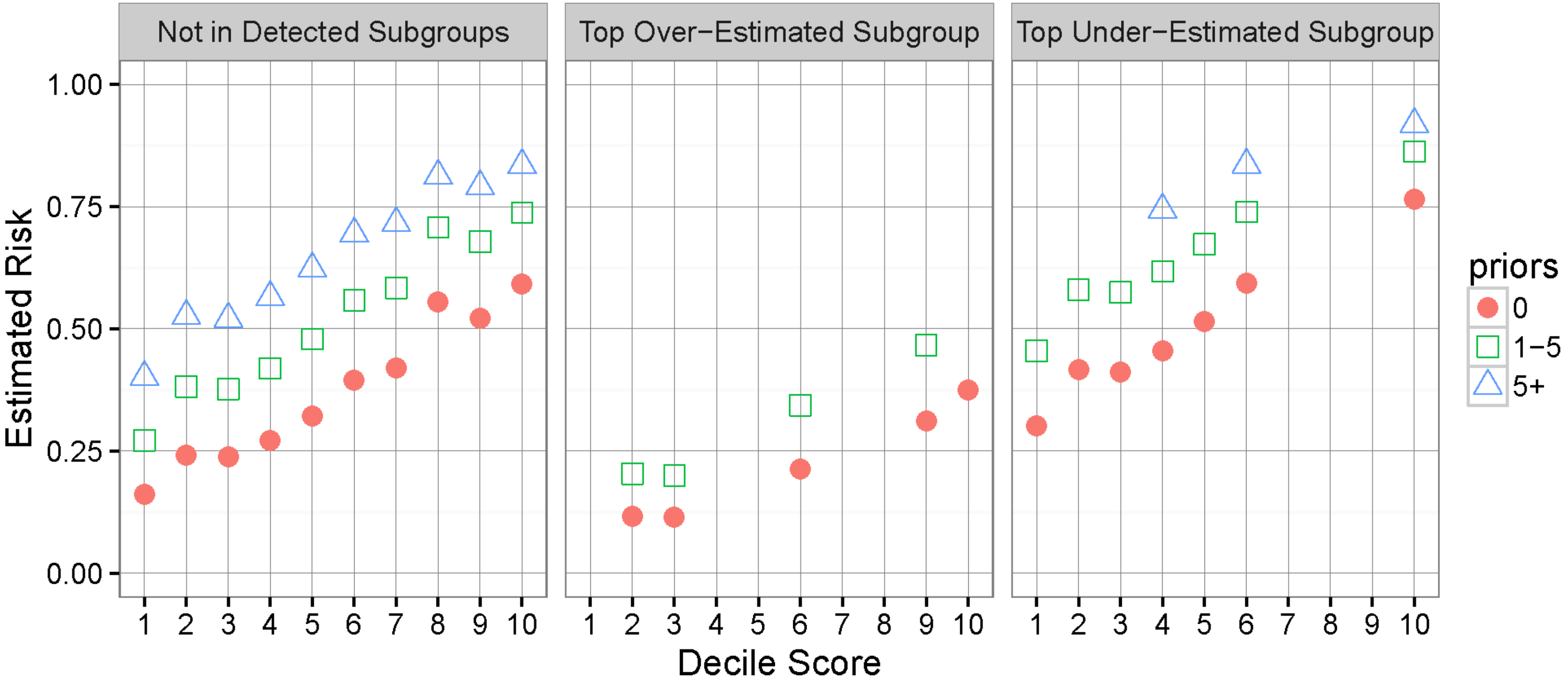


Start with maximum likelihood risk estimates for each COMPAS decile score.

Detection result 1: COMPAS underestimates the importance of prior offenses, overestimating risk for 0 priors, and underestimating risk for 5 or more priors.

Detection result 2: Even controlling for prior offenses, COMPAS still underestimates risk for males under 25, and overestimates risk for females who committed misdemeanors.

Results of Bias Scan on COMPAS

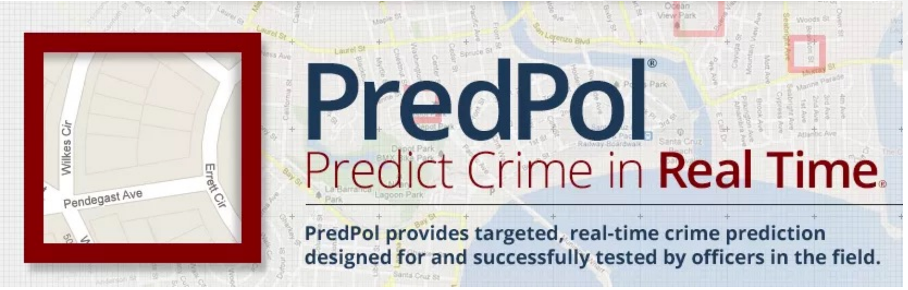


After controlling for prior offenses and membership in the two detected subgroups, there are no significant systematic biases in prediction.

NYU

Thorny question: given individual risk predictions, what to do with them?

The Bigger Picture



Chronicle Of Social Change

Chronicle Webpage

www

California Bets on Big Data to Predict Child Abuse

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016

THE WHITE HOUSE
WASHINGTON

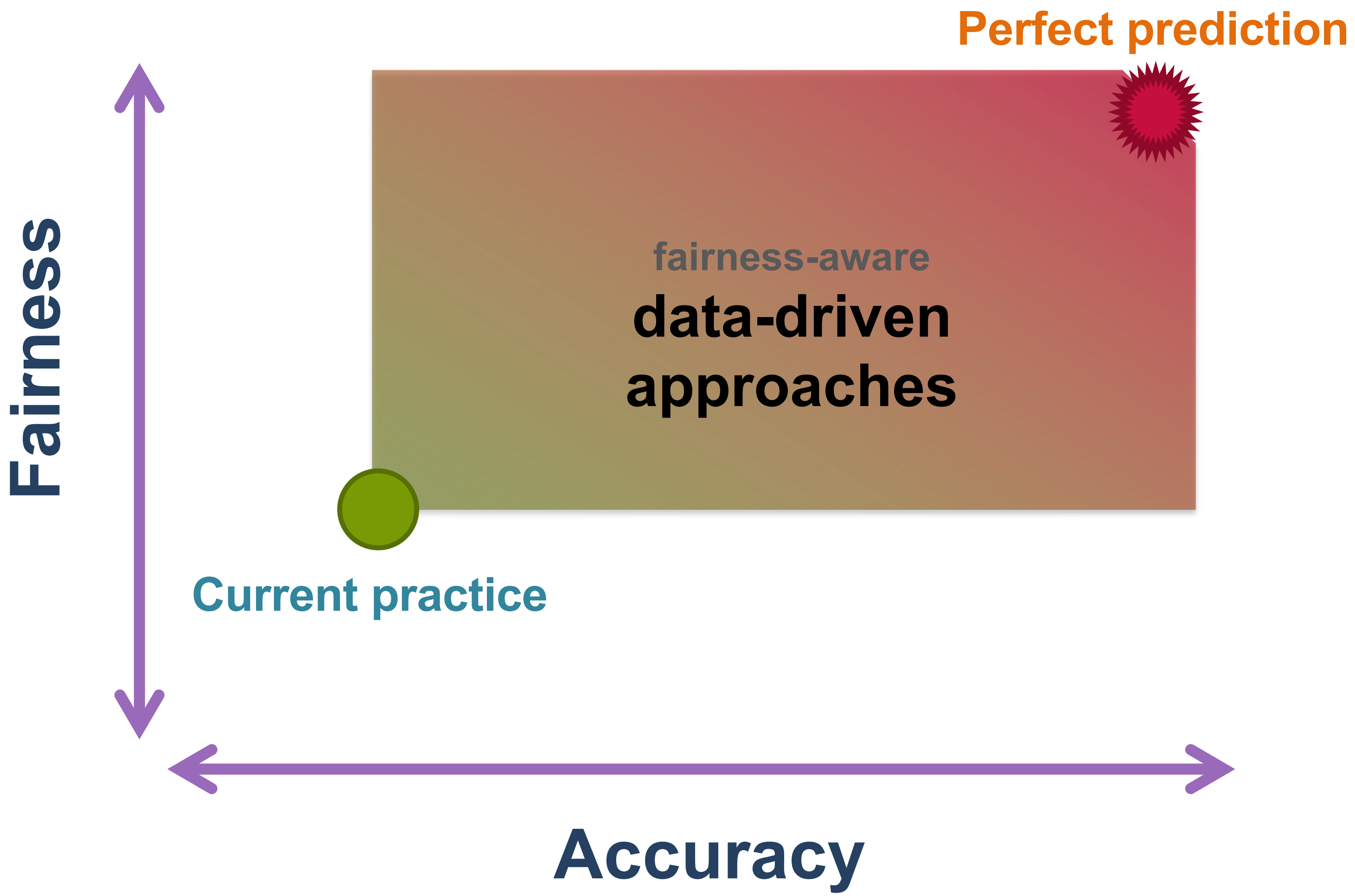
CADE METZ BUSINESS 07.11.16 7:00 AM

ARTIFICIAL INTELLIGENCE IS SETTING UP THE INTERNET FOR A HUGE CLASH WITH EUROPE

GETTY IMAGES

NYU

Should We Adopt Data-Driven Approaches?



References

1. Lecture slides:
“Data-driven discrimination and fairness-aware classification” (M. Jankowiak)
“Bias and discrimination in data-driven decision making” (A. Chouldechova)
“Identifying significant predictive bias in classifiers” (Z. Zhang and D.B. Neill)
2. Resources for fairness, accountability, and transparency in ML: https://www.fatml.org/resources/relevant-scholarship
3. A. Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data, 5(2): 153-163, 2017.
4. Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. https://arxiv.org/pdf/1611.08292.pdf . In NIPS Workshop on Interpretable Machine Learning, 2016.
5. A Romei & S Ruggieri. A multidisciplinary survey on discrimination analysis. http://www.di.unipi.it/~ruggieri/Papers/ker.pdf
6. S. Barocas and A.D. Selbst. Big Data’s Disparate Impact. In 104 California Law Review 671, 2016.
7. Žliobaitė, I. Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4): 1060-1089, 2017. http://www.zliobaite.com/publications
8. M. Bilal Zafar, et al. Learning Fair Classifiers. Tech. report, 2016. http://arxiv.org/pdf/1507.05259v3.pdf
9. S. Feldman et al.. Certifying and removing disparate impact. In Proc. KDD 2015, http://sorelle.friedler.net/papers/kdd_disparate_impact.pdf

Bayesian Network and Causality

Some Motivation for Bayes Nets

Urban systems consist of many interconnected sub-systems. We wish to understand the complex dependencies between these systems, both predictive and causal. We use it to **model** “typical” system behavior, to **detect** anomalies, trends and patterns, and to inform possible **interventions**.

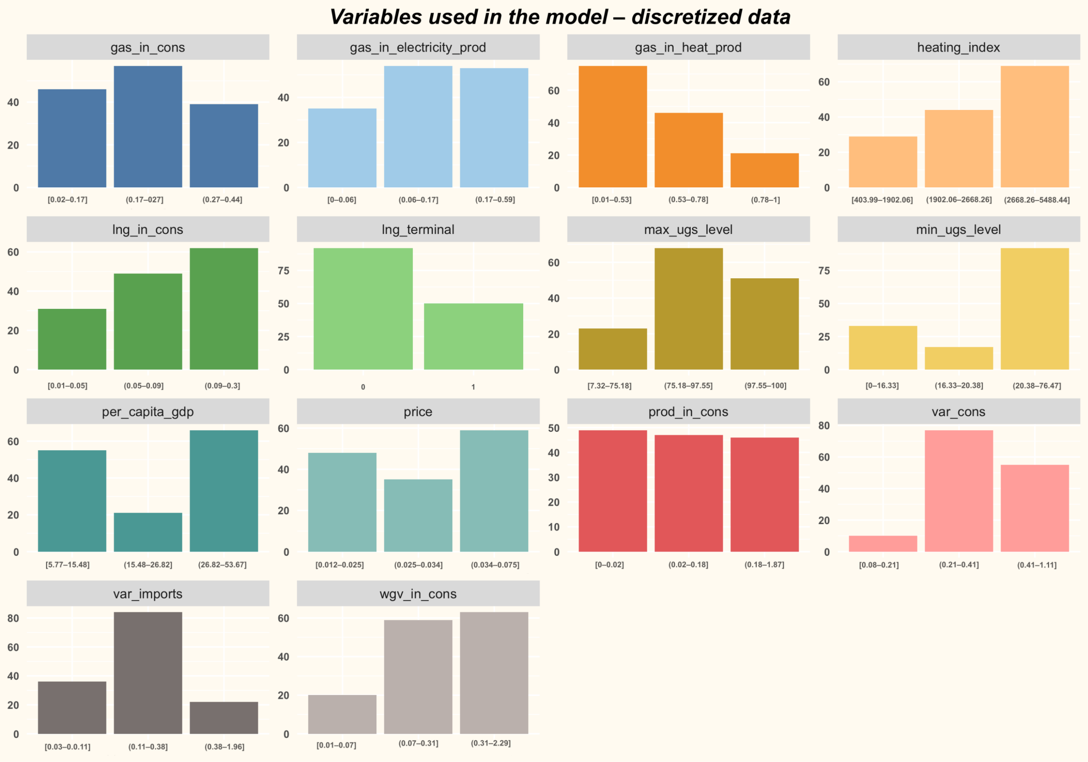
In this course, we will use Bayesian networks both to model probabilistic dependencies between many variables and to infer the causal relationships between them.

Why Bayesian Networks?

- An easily interpretable graphical representation of the relationships between a set of variables.
- Bayes Nets can be specified manually or learned automatically from data and enable computationally efficient probabilistic inferences.
- Many practical and successful applications in medicine, manufacturing, and failure diagnosis:

Diagnosis: infer Pr (problem type | symptoms)
Prediction: infer probability distributions for values that are expensive or impossible to measure.
Anomaly Detection: detect observations that are very unlikely (i.e. have low probabilities given the model).
Active Learning: choose the most informative diagnostic test to perform given these observations.

Example: Underground Gas Storage



Source: *Kosowski, P.; Kosowska, K.; Nawalaniec, W. Application of Bayesian Networks in Modeling of Underground Gas Storage Energy Security. Energies 2022, 15, 5185*

Introduction to Bayesian Networks

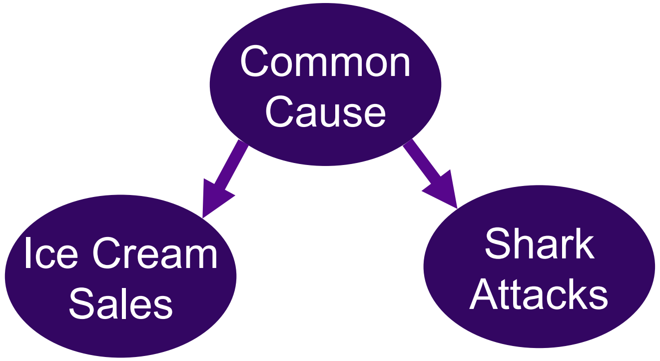
Recent Dow-Jones Change	Number of Ice Creams Sold Today	Number of Shark Attacks Today
UP	3500	4
STEADY	41	0
UP	2300	5
DOWN	3400	4
UP	18	0
STEADY	105	0
STEADY	4	0
STEADY	6310	3
UP	70	0

More ice creams = More shark attacks!

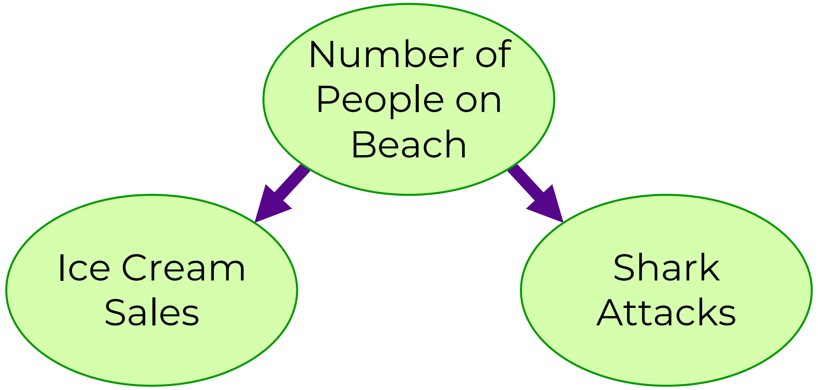
Does eating ice cream cause shark attacks?

Do shark attacks affect ice cream consumption?

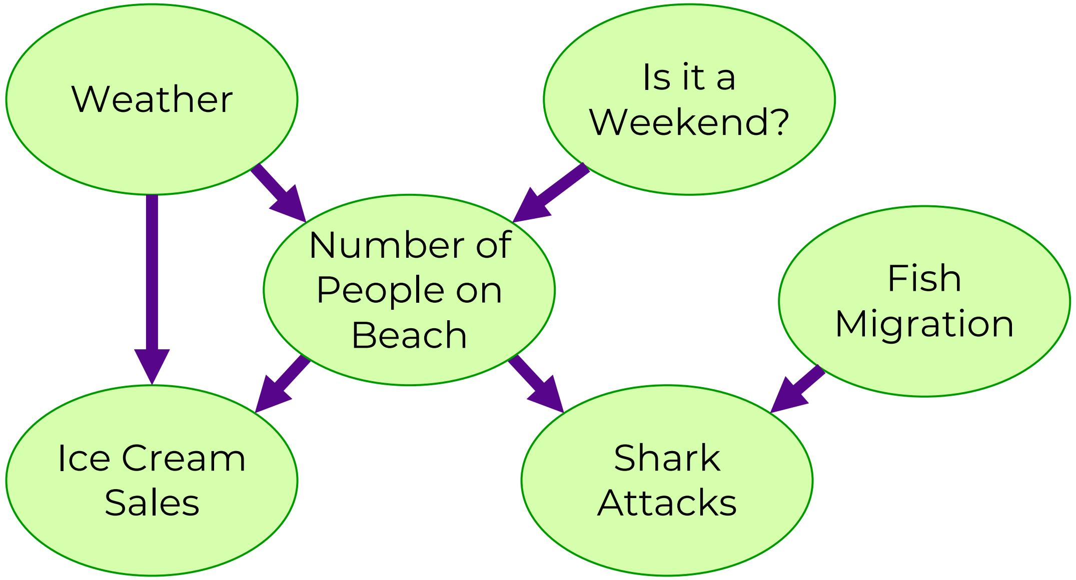
Perhaps the two events have a common cause!



Introduction to Bayesian Networks

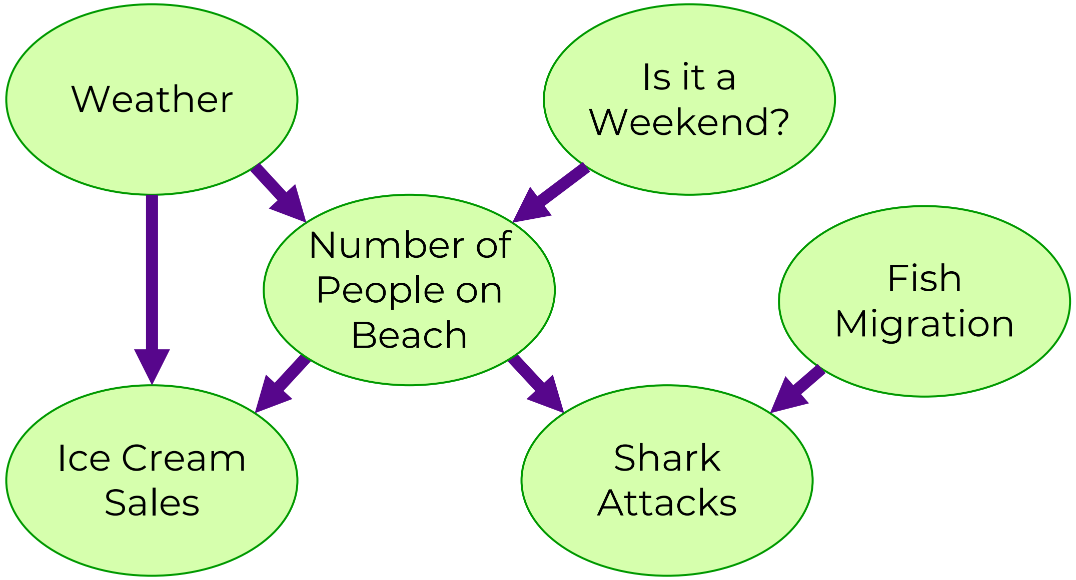


Introduction to Bayesian Networks



Many other factors may influence some or all of these variables…

Introduction to Bayesian Networks

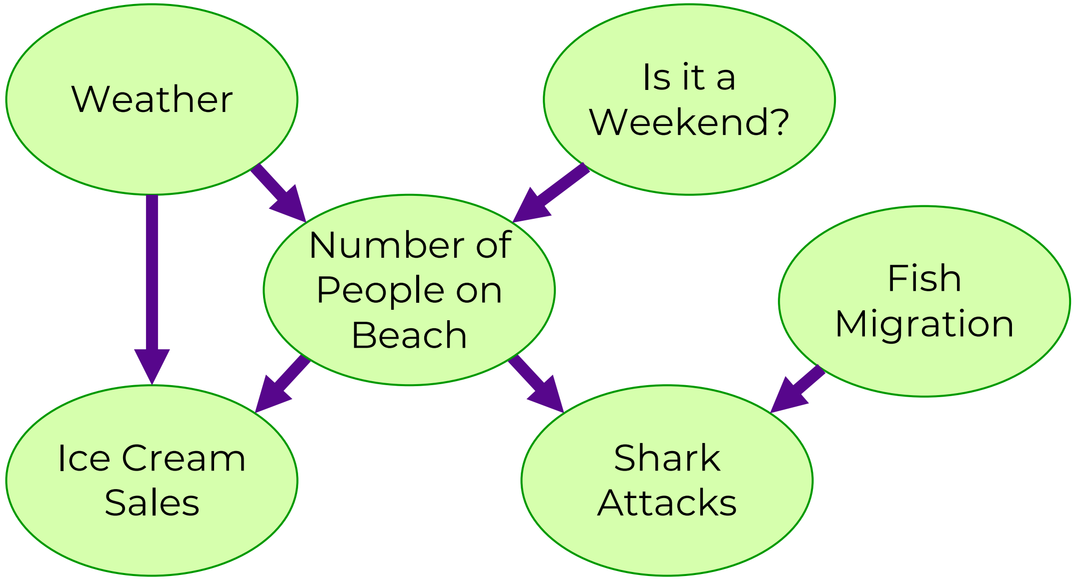


Many other factors may influence some or all of these variables…

What we have here is a way of representing probabilistic relationships between variables. We call this a **Bayesian network**.

NYU

Introduction to Bayesian Networks



Many other factors may influence some or all of these variables…

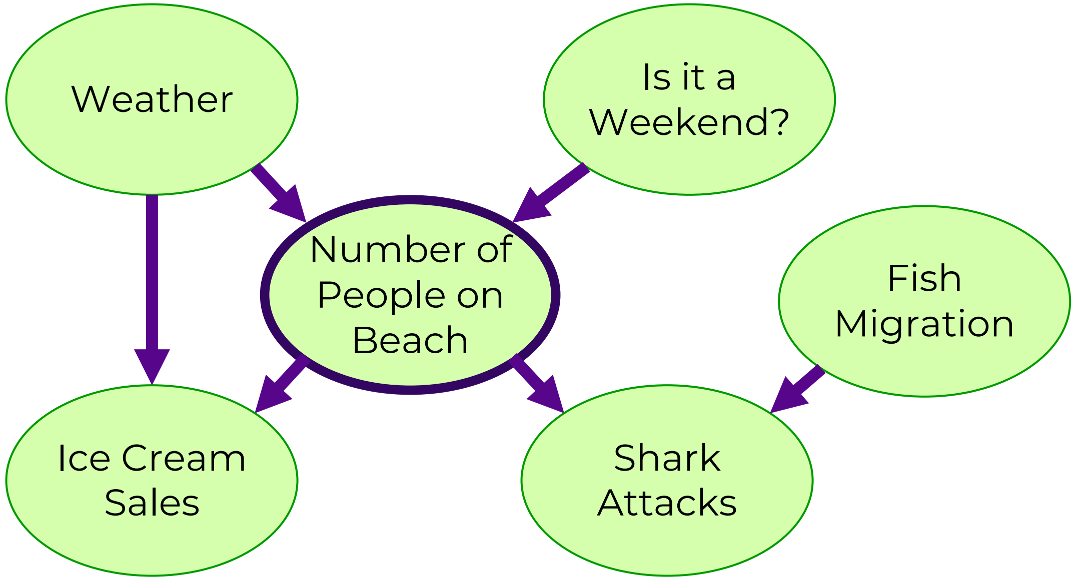
What we have here is a way of representing probabilistic relationships between variables. We call this a **Bayesian network**.

Causal interpretation: Link X → Y means that variable X directly **causes** variable Y.

Probabilistic interpretation: The links encode **conditional dependencies** between variables.

"X and Y are conditionally independent given Z":
Pr(X | Z) = Pr (X | Y, Z) and Pr(Y | Z) = Pr(Y | X, Z)

Introduction to Bayesian Networks



Many other factors may influence some or all of these variables…

What we have here is a way of representing probabilistic relationships between variables. We call this a **Bayesian network**.

Each node has a probability distribution that is conditioned on its parents' values.

Sunny, Weekend:
Pr (Beach Crowded) = 90%
Sunny, Not Weekend:
Pr (Beach Crowded) = 40%
Not Sunny, Weekend:
Pr (Beach Crowded) = 20%
Not Sunny, Not Weekend:
Pr (Beach Crowded) = 1%

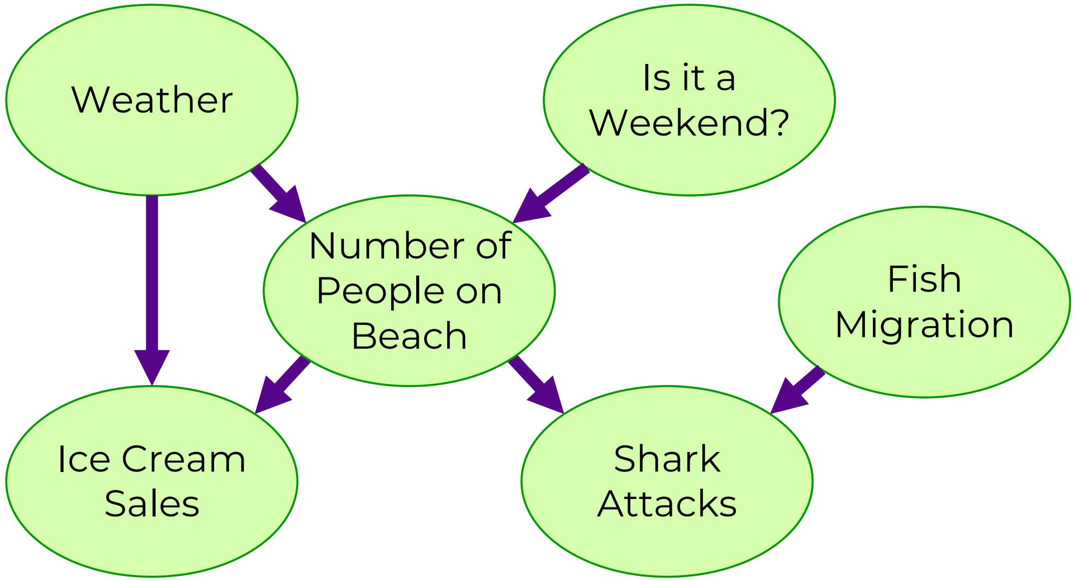
Causal interpretation: Link X → Y means that variable X directly **causes** variable Y.

Probabilistic interpretation: The links encode **conditional dependencies** between variables.

"X and Y are conditionally independent given Z":
Pr(X | Z) = Pr (X | Y, Z) and Pr(Y | Z) = Pr(Y | X, Z)

NYU

Introduction to Bayesian Networks



Causal interpretation: Link X → Y means that variable X directly **causes** variable Y.

Probabilistic interpretation: The links encode **conditional dependencies** between variables.

"X and Y are conditionally independent given Z": Pr(X | Z) = Pr (X | Y, Z) and Pr(Y | Z) = Pr(Y | X, Z)

Many other factors may influence some or all of these variables…

What we have here is a way of representing probabilistic relationships between variables. We call this a **Bayesian network**.

In a Bayes Net, each node is conditionally independent of its non-descendants given its parents.

For example, if you already know the number of people on the beach and have fish migration data, knowing today's weather doesn't give you any more information about shark attacks.

CI(Shark Attacks, Weather | Number of People on Beach, Fish Migration)

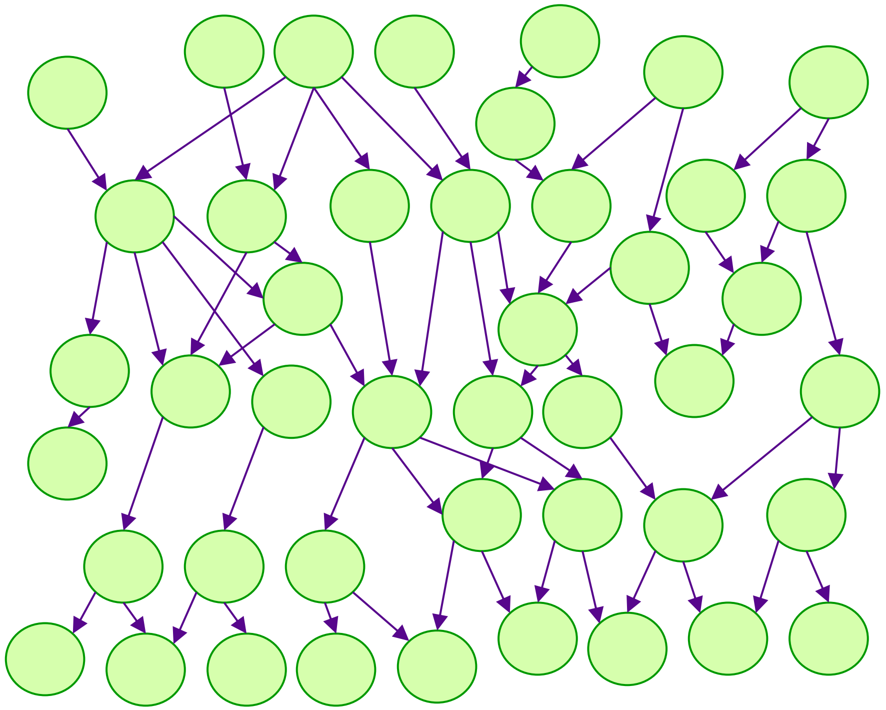
NYU

Real-World Bayes Nets

Bayesian networks for real-world application domains may have hundreds or thousands of nodes.

They can be built manually, consulting domain experts for the structure and probabilities.

More often, the probabilities (and, in some cases, the structure) are **learned** from training data.



Pathfinder (a diagnostic system for lymph node diseases) uses a Bayes Net with 60 diseases, 100 symptoms, and 14,000 probabilities.

The Bayes Net was constructed manually by human experts: 8 hours to choose variables, 35 hours for structure, and 40 hours for probabilities.

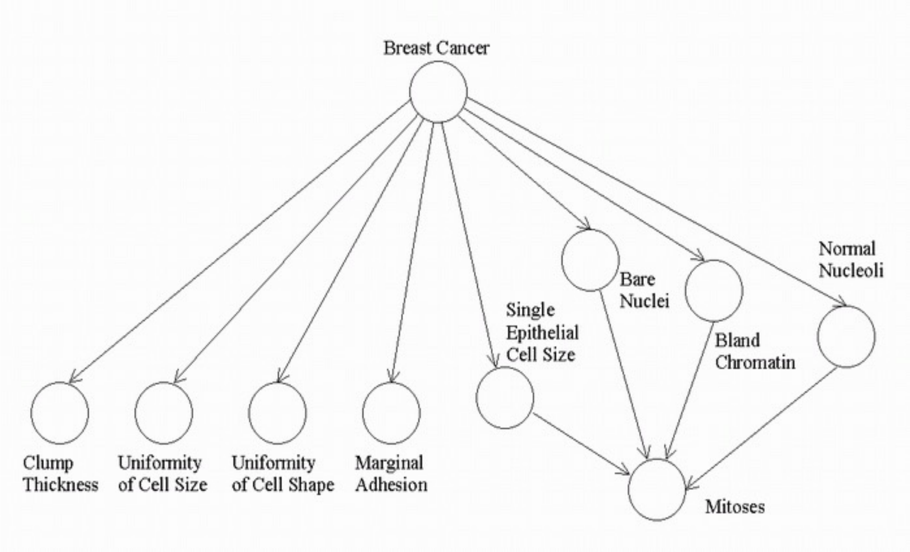
Pathfinder was shown to outperform human experts in diagnosis accuracy and has been extended to other medical domains.

Real-World Bayes Nets

Bayesian networks for real-world application domains may have hundreds or thousands of nodes.

They can be built manually, consulting domain experts for the structure and probabilities.

More often, the probabilities (and, in some cases, the structure) are **learned** from training data.



Medical diagnosis is a common application of Bayes Nets: here's a simple network that can be used to infer Pr(breast cancer | symptoms).

Pathfinder (a diagnostic system for lymph node diseases) uses a Bayes Net with 60 diseases, 100 symptoms, and 14,000 probabilities.

The Bayes Net was constructed manually by human experts: 8 hours to choose variables, 35 hours for structure, and 40 hours for probabilities.

Pathfinder was shown to outperform human experts in diagnosis accuracy and has been extended to other medical domains.

Joint Probability Distributions

a.k.a. “Why are Bayes Nets so useful for representing probabilities?”

To create the joint distribution of M discrete-valued variables:

Make a truth table listing all combinations of values of your variables. If there are M binary variables, the table has 2^M rows.

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Joint Probability Distributions

a.k.a. "Why are Bayes Nets so useful for representing probabilities?"

To create the joint distribution of M discrete-valued variables:

Make a truth table listing all combinations of values of your variables. If there are M binary variables, the table has 2^M rows.

For each combination of values, say how probable it is. These probabilities must sum to 1.

What if we have M = 100 variables: how can we use less than 2^{100} rows?

Use information about conditional independencies between variables to infer a Bayesian network structure!

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

Joint Probability Distributions

a.k.a. "Why are Bayes Nets so useful for representing probabilities?"

To create the joint distribution of M discrete-valued variables:

Make a truth table listing all combinations of values of your variables. If there are M binary variables, the table has 2^M rows.

For each combination of values, say how probable it is. These probabilities must sum to 1.

What if we have M = 100 variables: how can we use less than 2100 rows?

Use information about conditional independencies between variables to infer a Bayesian network structure!

Here's a simple example:

D = "there is a disease outbreak"
$X_1..X_{99}$ = "person i goes to the hospital"

Joint Probability Distributions

a.k.a. "Why are Bayes Nets so useful for representing probabilities?"

To create the joint distribution of M discrete-valued variables:

Make a truth table listing all combinations of values of your variables. If there are M binary variables, the table has 2^M rows.

For each combination of values, say how probable it is. These probabilities must sum to 1.

What if we have M = 100 variables: how can we use less than 2100 rows?

Use information about conditional independencies between variables to infer a Bayesian network structure!

<u>Here's a simple example</u>:

D = "there is a disease outbreak"
$X_1..X_{99}$ = "person i goes to the hospital"

If D is known, also knowing X_i does not change the probability of X_j:

$$\Pr(X_j \mid D, X_i) = \Pr(X_j \mid D)$$

Joint Probability Distributions

a.k.a. "Why are Bayes Nets so useful for representing probabilities?"

To create the joint distribution of M discrete-valued variables:

Make a truth table listing all combinations of values of your variables. If there are M binary variables, the table has 2^M rows.

For each combination of values, say how probable it is. These probabilities must sum to 1.

What if we have M = 100 variables: how can we use less than 2100 rows?

Use information about conditional independencies between variables to infer a Bayesian network structure!

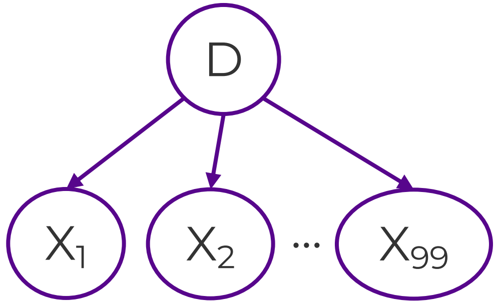
Here's a simple example:

D = "there is a disease outbreak"

$X_1..X_{99}$ = "person i goes to the hospital"

If D is known, also knowing X_i does not change the probability of X_j:

$$\Pr(X_j \mid D, X_i) = \Pr(X_j \mid D)$$



$$\Pr(D, X_1..X_{99}) = \Pr(D) \prod_{i=1..99} \Pr(X_i \mid D)$$

Bayes Nets often allow a much more compact representation of the joint distribution!

(In general, how many probabilities do we need?)

Building a Bayes Net

Small Bayes Nets are easy to build by hand, assuming that we understand the relationships between variables and are able to estimate their conditional probabilities.

Large Bayes Nets may require many person-hours to build, but they can also be **learned** automatically from data.

For example, let's assume that we want to build a Bayes Net to determine whether a terrorist anthrax attack has occurred.

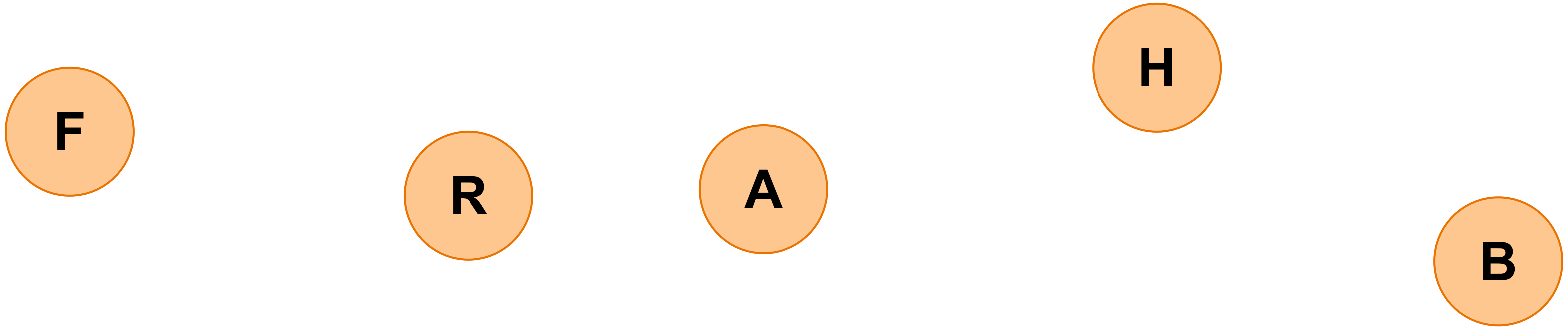
1. An anthrax attack is likely to increase the level of respiratory illness.
2. Seasonal influenza is also likely to cause an increase in respiratory illness.
3. The CDC has a **hospital surveillance system** that alerts when the number of ED visits is abnormally high and has additionally deployed **bio-sensors** for airborne anthrax detection.
4. The hospital surveillance system and the bio-sensors are not perfect: both false alarms and missed outbreaks are possible.

Define the following variables:

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert

Building a Bayes Net

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert

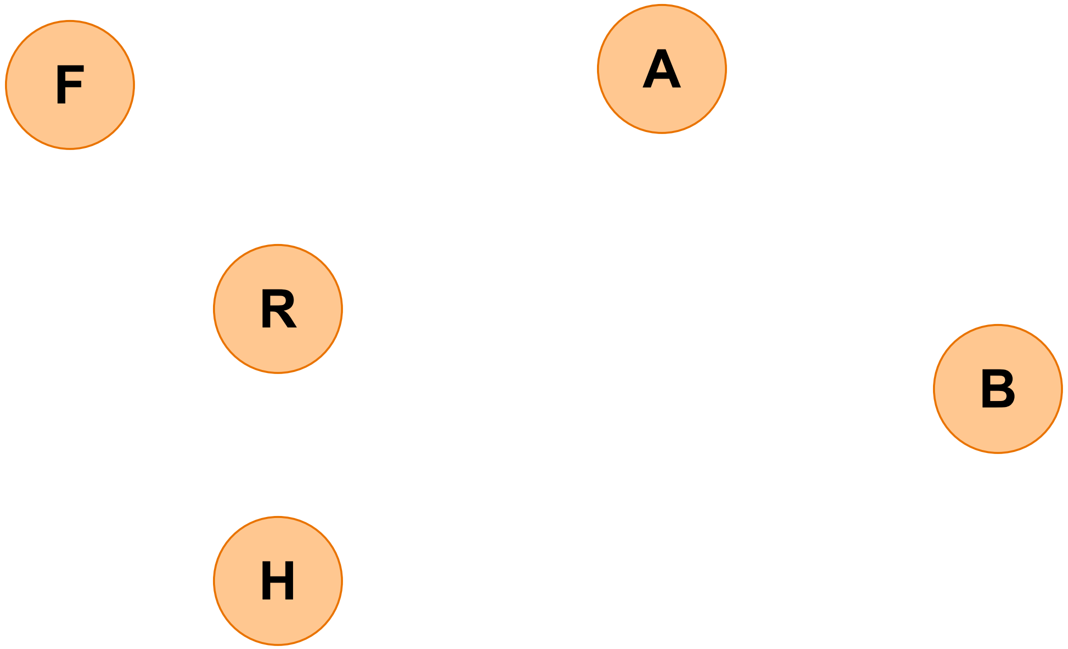


Step 1: Choose a set of relevant variables and represent each variable by a node.

NYU

Building a Bayes Net

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



<u>Step 2</u>: Choose an ordering for the variables $X_1..X_M$,
such that if X_i influences X_j, then $i < j$.

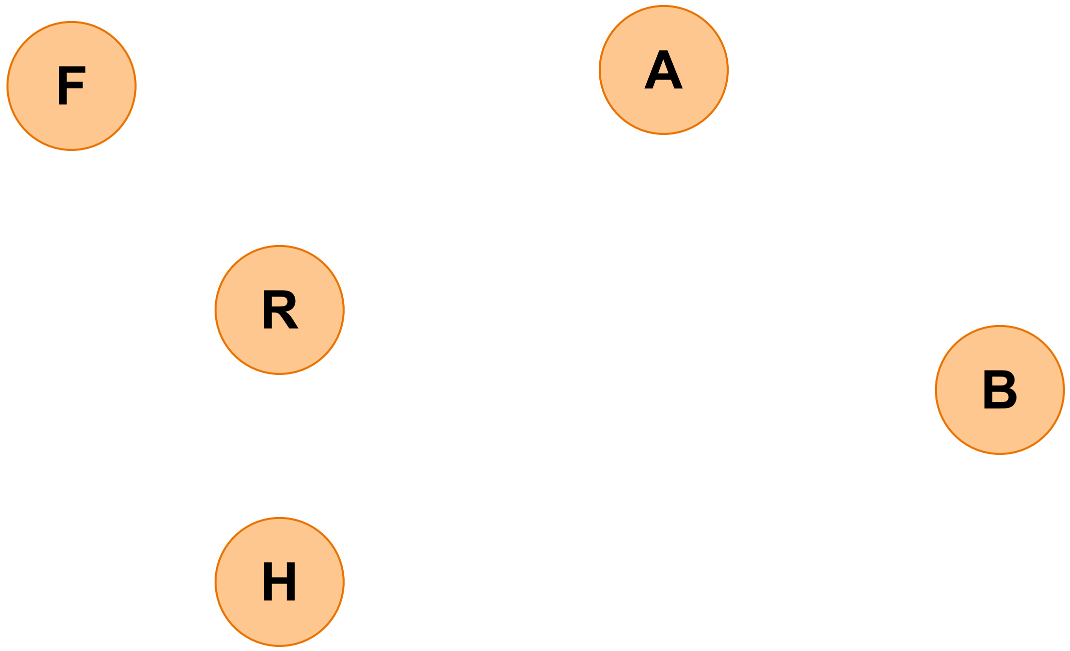
NYU

<u>Hint</u>: put environmental and event variables first, then latent variables, then observations.

Any ordering will produce a valid Bayes Net structure, but using the causal information will produce more compact (fewer links) and more interpretable structures.

Building a Bayes Net

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



Step 3: Add links.

- The link structure must be acyclic.
- If node X is given parents $Q_1,Q_2,\ldots Q_m$, you are promising that any variable that's a non-descendent of X is conditionally independent of X given $\{Q_1,Q_2,..Q_m\}$

NYU

Building a Bayes Net

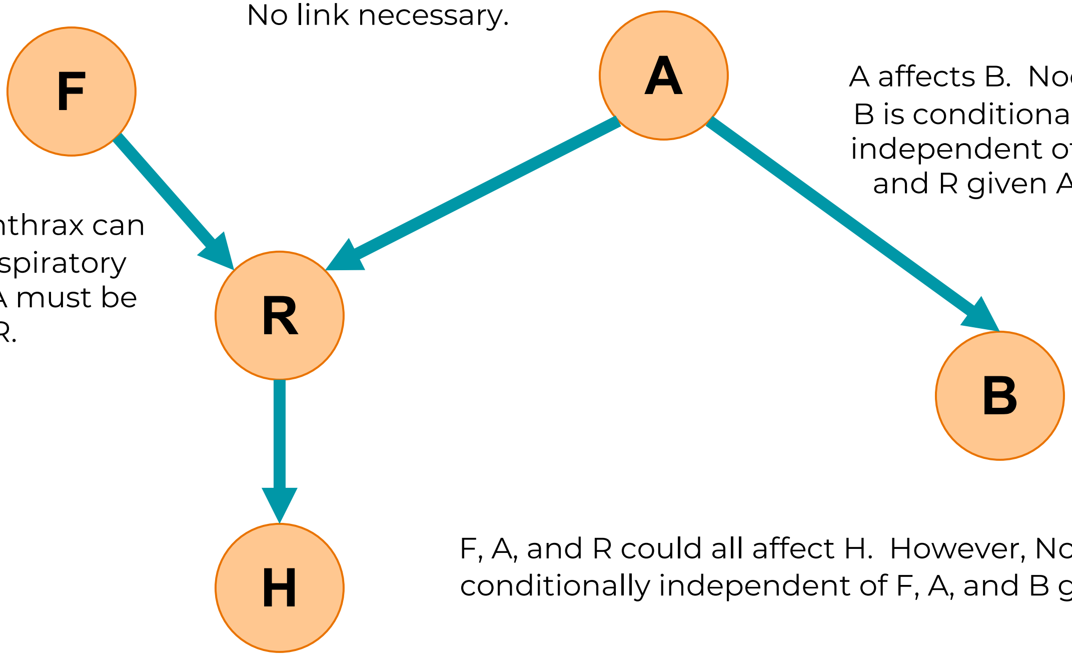
F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert

F is the first variable, so it cannot have any parents.

Whether an anthrax attack occurs does not depend on flu season. No link necessary.

A affects B. Node B is conditionally independent of F and R given A.

Since either flu or anthrax can cause increased respiratory illness, both F and A must be parents of R.



F, A, and R could all affect H. However, Node H is conditionally independent of F, A, and B given R.

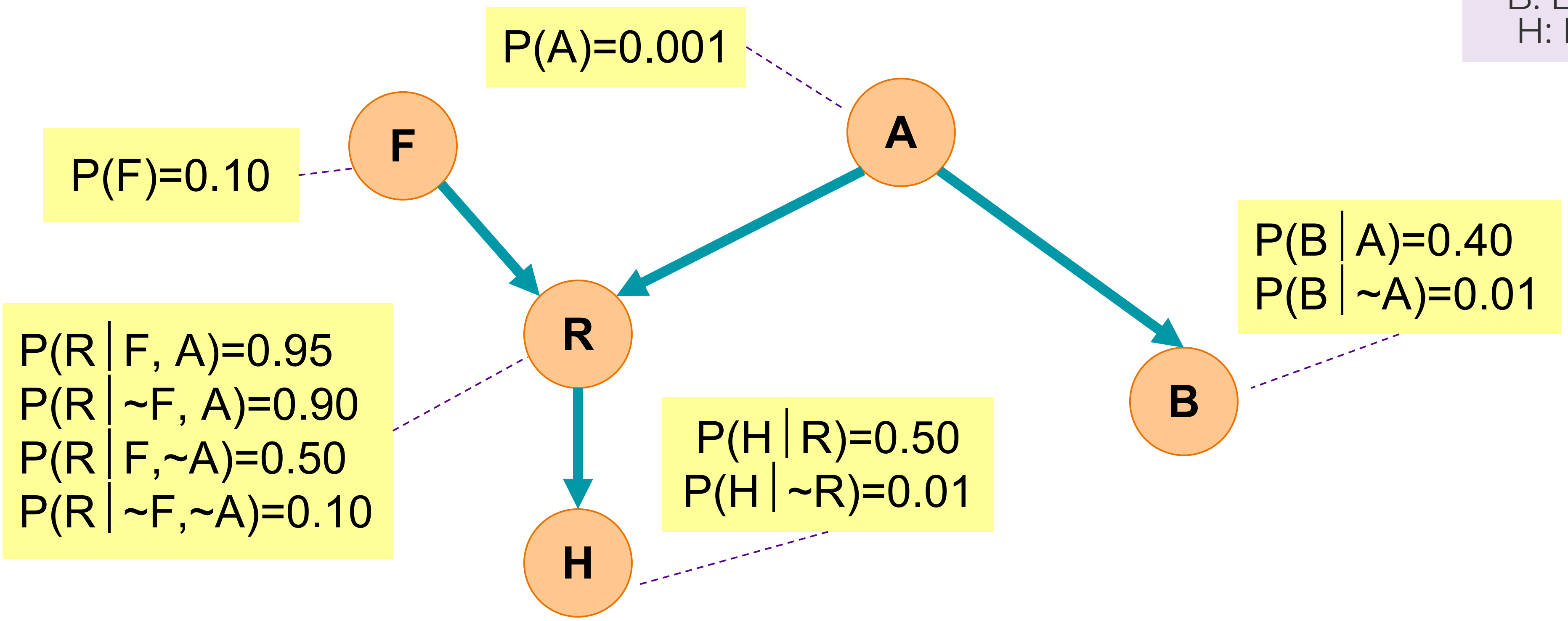
<u>Step 3</u>: Add links.

- For each variable X_i (for i = 2...M), choose a minimal subset of parents from $X_1..X_{i-1}$, such that X_i is conditionally independent of the rest of $X_1..X_{i-1}$ given its parents.

NYU

Building a Bayes Net

F: Flu season
A: Anthrax attack has occurred
R: Respiratory illness increased
B: Bio-sensors detect anthrax
H: Hospital surveillance alert



Step 4: Add a conditional probability table for each node.

- The table for node X must list Pr(X | Parents(X)) for each value of X and each combination of parent values.
- If X is binary, we know that Pr(~X | Parents) = 1 – Pr(X | Parents), and do not need to write this explicitly.

NYU

Building a Bayes Net

Practice problem: suppose that we're building a nuclear power station and want to report when the core temperature is low.

The gauge is meant to read the temperature of the core, and the alarm is meant to sound if the gauge reads a low temperature.

However, the gauge or the alarm (or both) could be faulty.

Which Bayesian network structure makes the most sense for these five variables?

Lab Time

For the Next Week

1. Assignment 3

   Due: Apr 14, 2024 (11:59pm)

References

Fairness:

1. Resources for fairness, accountability, and transparency in ML: https://www.fatml.org/resources/relevant-scholarship
2. A. Chouldechova. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data, 5(2): 153-163, 2017.
3. Z. Zhang and D.B. Neill. Identifying significant predictive bias in classifiers. https://arxiv.org/pdf/1611.08292.pdf . In NIPS Workshop on Interpretable Machine Learning, 2016.
4. A Romei & S Ruggieri. A multidisciplinary survey on discrimination analysis. http://www.di.unipi.it/~ruggieri/Papers/ker.pdf
5. S. Barocas and A.D. Selbst. Big Data's Disparate Impact. In 104 California Law Review 671, 2016.
6. Žliobaitė, I. Measuring discrimination in algorithmic decision making. Data Mining and Knowledge Discovery, 31(4): 1060-1089, 2017. http://www.zliobaite.com/publications
7. M. Bilal Zafar, et al. Learning Fair Classifiers. Tech. report, 2016. http://arxiv.org/pdf/1507.05259v3.pdf
8. S. Feldman et al.. Certifying and removing disparate impact. In Proc. KDD 2015, http://sorelle.friedler.net/papers/kdd_disparate_impact.pdf

Bayes Nets:

1. J. Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.
2. S. Russell and P. Norvig. AI: A Modern Approach, Ch. 15.
3. Several excellent tutorials on Bayes Nets available at https://www.cs.cmu.edu/~./awm/tutorials/