Machine Learning for Large-Scale Data Analysis and Decision Making (MATH80629A) Winter 2022

Week #11- **Summary**

Announcement

- **Last Quiz:**
- **Project Presentation: in-person** on April 5
- Final exam: **in-person** on April 13

Today

- Trustworthy Machine Learning
- Recommender Systems: case study
- Q&A

Trustworthy ML

ML is everywhere!

amazon

Google

facebook

NETFLIX

Nowadays AI/ML algorithms determine

- Who gets a job 

- Who goes to jail 
- Who receives loan from bank 
- Who gets diagnose and receive health care 

Are these automated decision making systems trustworthy?

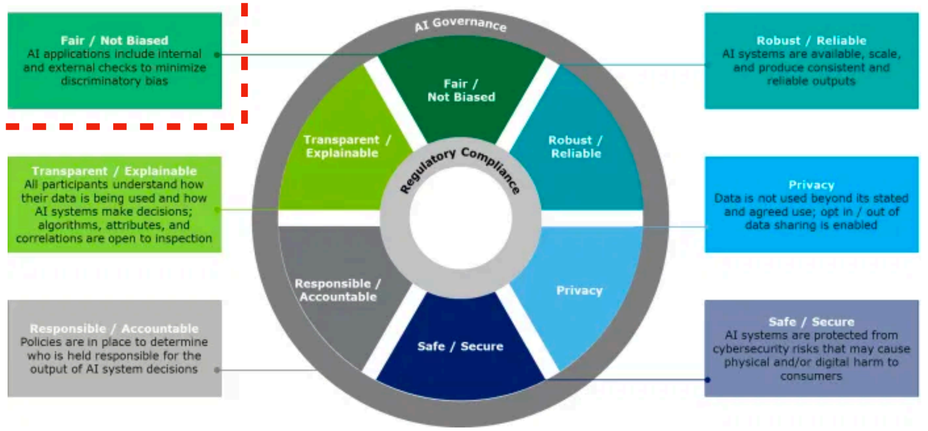
Trustworthy AI/ML

- Trustworthy AI is crucial to the widespread adoption of AI.
- Trustworthy AI is a framework to help address elements that ensure the ethical use of AI and sustain the trust of customers.

Fairness in AI

Explainable AI

Responsible AI

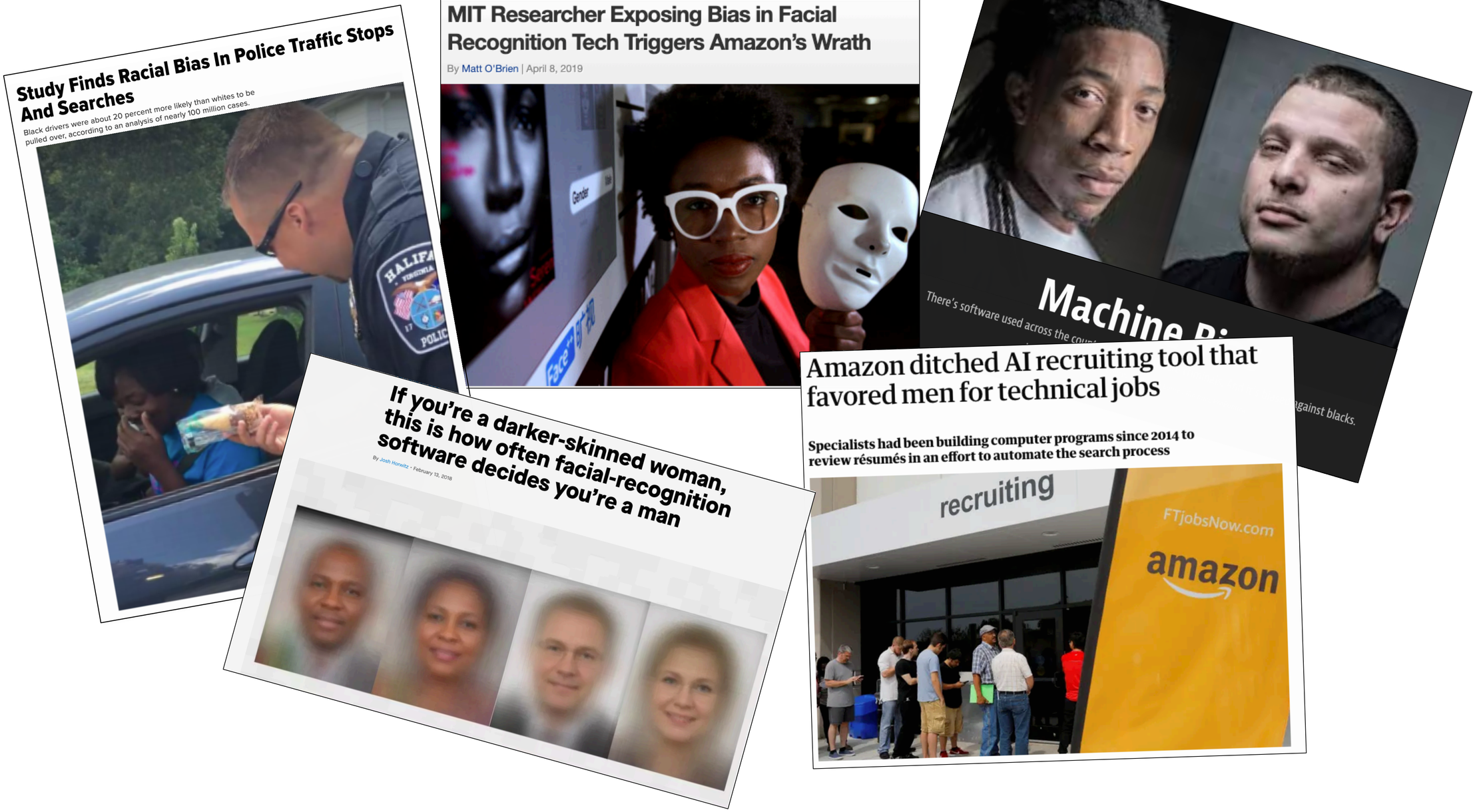


Reliable/Robust AI

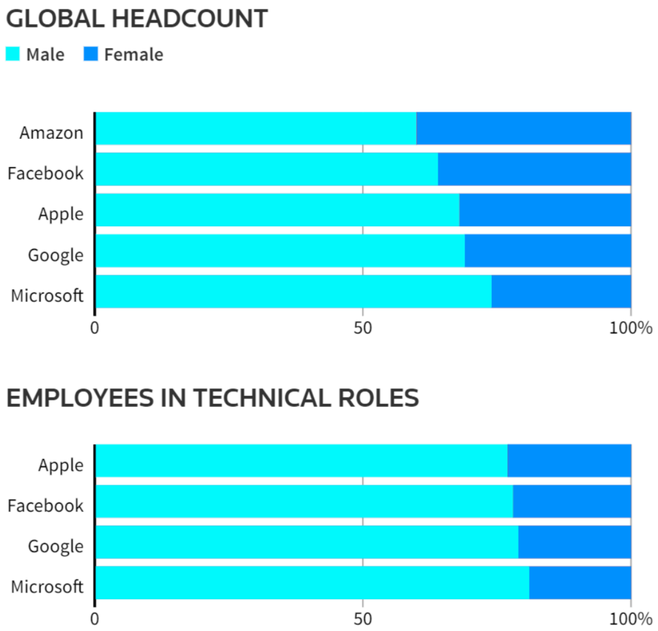
Privacy-preserving AI

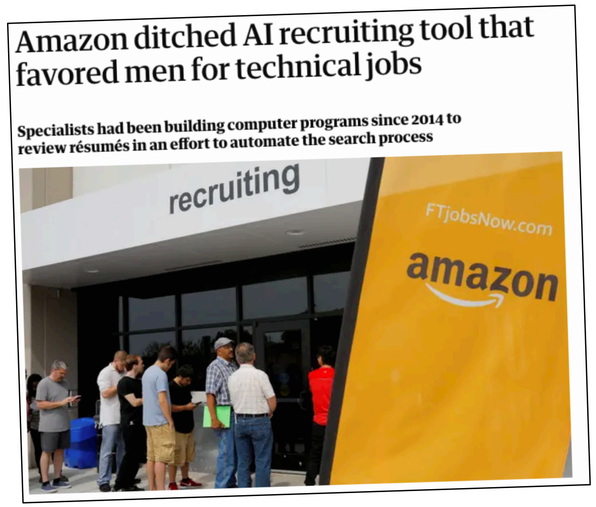
Safe AI

Source: Deloitte Consulting LLP


Study Finds Racial Bias In Police Traffic Stops And Searches
Black drivers were about 20 percent more likely than whites to be pulled over, according to an analysis of nearly 100 million cases.
MIT Researcher Exposing Bias in Facial Recognition Tech Triggers Amazon's Wrath
By Matt O'Brien | April 8, 2019
Machine Bi
There's software used across the
against blacks.
If you're a darker-skinned woman, this is how often facial-recognition software decides you're a man
By Josh Horwitz · February 13, 2018
Amazon ditched AI recruiting tool that favored men for technical jobs
Specialists had been building computer programs since 2014 to review résumés in an effort to automate the search process
recruiting
FTjobsNow.com
amazon

Amazon Recruitment Tool

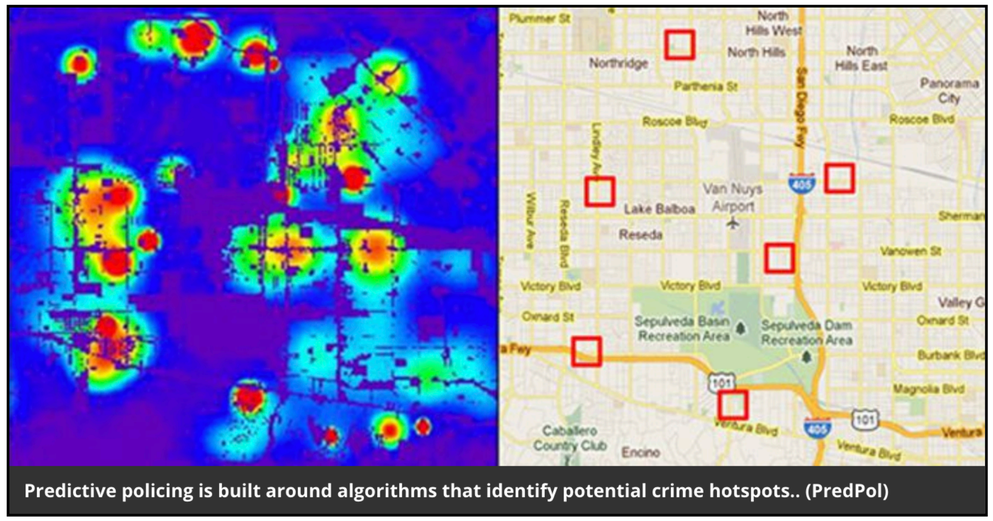




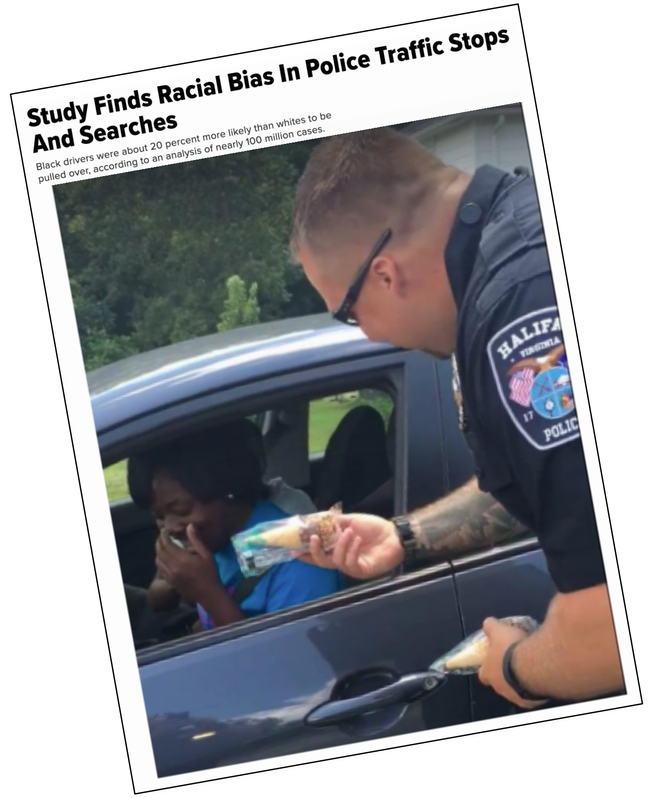
Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

Policing and law enforcement

- Investigative tools are AI-based models.



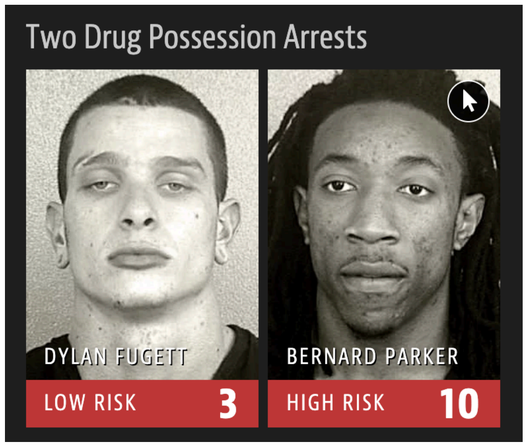
Predictive policing is built around algorithms that identify potential crime hotspots.. (PredPol)

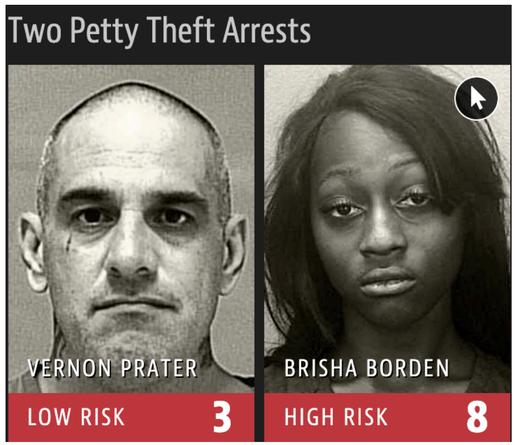


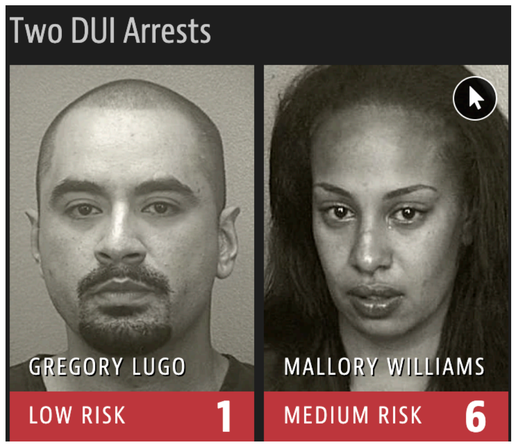
A. Romei and S. Ruggieri (2014). A multidisciplinary survey on discrimination analysis. The Knowledge Engineering Review 29, pp 582-638

COMPAS

- The COMPAS software used across US to predict future criminals is **biased against blacks**.
- **Risk assessment scores** are increasingly common in courtrooms
- They are used to inform **decisions** about who can be set free at every stage of the criminal justice system







https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Gender-shades

- Let’s hear about if from Joy Buolamwini!

  http://gendershades.org/

Law Against Discrimination

Legally recognized ‘protected classes’

Race (Civil Rights Act of 1964)
Color (Civil Rights Act of 1964)
Sex (Equal Pay Act of 1963; Civil Rights Act of 1964)
Religion (Civil Rights Act of 1964)
National origin (Civil Rights Act of 1964)
Citizenship (Immigration Reform and Control Act)
Age (Age Discrimination in Employment Act of 1967)
Pregnancy (Pregnancy Discrimination Act)
Familial status (Civil Rights Act of 1968)
Disability status (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

sensitive attributes

Regulated domains

Credit (Equal Credit Opportunity Act)
Education (Civil Rights Act of 1964; Education Amendments of 1972)
Employment (Civil Rights Act of 1964)
Housing (Fair Housing Act)
Public Accommodation (Civil Rights Act of 1964)
Extends to marketing and advertising; not limited to final decision
This list sets aside complex web of laws that regulates the government

Why we should care about fairness?

To address Law Against Discrimination!

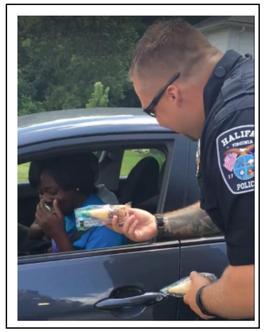
Fairness in ML

2014

"Big Data: Seizing Opportunities, Preserving Values"

"big data technologies can cause societal harms beyond damages to privacy"

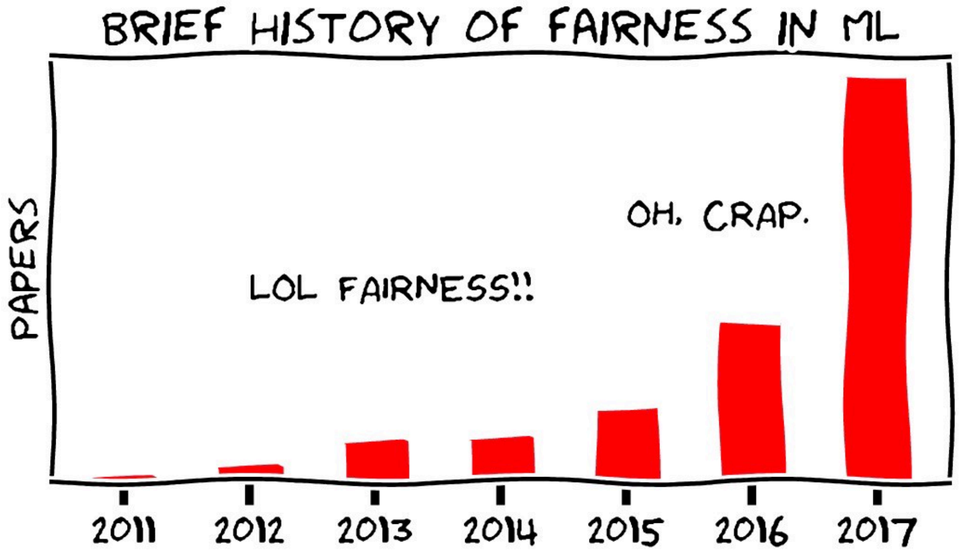
2015



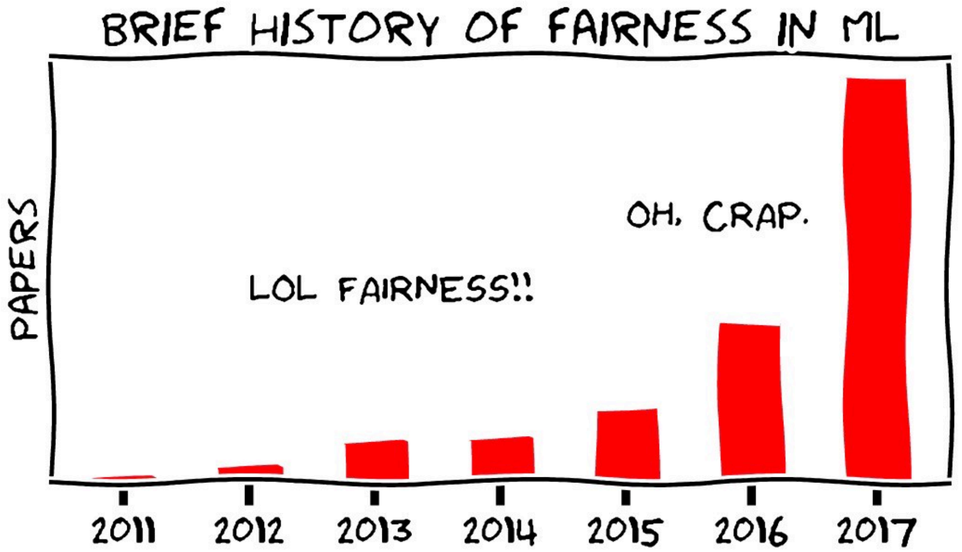
2016

2017

...



Fairness in ML



- “What is fair have been introduced in multiple disciplines for well over 50 years, including in education, hiring, and machine learning” [1].

- Statistics, Social Science, Economics, etc.

[1] Hutchinson, Ben, and Margaret Mitchell. "50 Years of Test (Un) fairness: Lessons for Machine Learning."

arXiv preprint arXiv:1811.10104 (2018).

How to address fairness in AI/ML?

bias

Data is unbalanced
Historical discrimination
Encodes protected attributes

Data scientists do not build the models

Unfair outcome
Black-box models
No user feedback

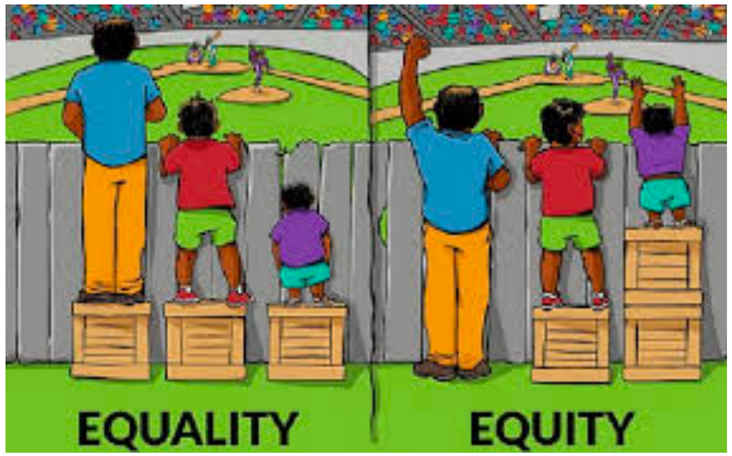
Why do we use fairness definitions?

- To make algorithmic systems support humar
- To identify strengths and weakness of the sy
- To track improvement over time

To address Law Against Discrimination!

Why we have many notions of fairness?

By 2018, we had **21 definitions of fairness**!



An interesting tutorial by **Arvind Narayanan**:
Tutorial: 21 fairness definitions and their politics

Another interesting tutorial by **Jon Kleinberg**:
Inherent Trade-Offs in Algorithmic Fairness

Definition
Group fairness or statistical parity
Conditional statistical parity
Predictive parity
False positive error rate balance
False negative error rate balance
Equalised odds
Conditional use accuracy equality
Overall accuracy equality
Treatment equality
Test-fairness or calibration
Well calibration
Balance for positive class
Balance for negative class
Causal discrimination
Fairness through unawareness
Fairness through awareness
Counterfactual fairness
No unresolved discrimination
No proxy discrimination
Fair inference

Verma, Sahil, and Julia Rubin. "Fairness definitions explained." 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 2018.

Why we don’t have one definition?

- Correcting for **algorithmic bias** generally requires:
  - **knowledge** of how the measurement process is biased
  - **judgments** about properties to satisfy in an “unbiased” world

Hiring

Gender-biased

Medical diagnosis

Gender-biased

Fairness is not a general concept

Bias is **subjective** and must be considered **relative** to task

There is no agreed-upon measure

Forbes: Amazon exec Jeff Bezos is the ...
cnbc.com

Powerful CEO Infographics : an...
trendhunter.com

Watches worn by the most powerf...
businessinsider.com

The World's 10 Most Powerful Executiv...
forbes.com

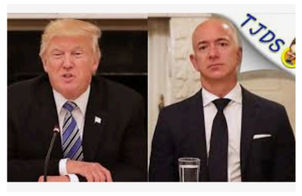
CEOs: Powerful, but not respected ...
humanresourcesonline.net

The World's 10 Most Powerful CEOs
forbes.com

Larry Page named world's most powerful...
economictimes.indiatimes.com

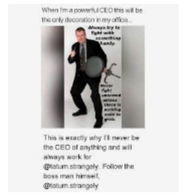
300 Most Powerful Black CEO, COO...
blackenterprise.com

Powerful CEO Portrait Male Business M...
shutterstock.com



CEO Joins Pentagon Defense Board ...
youtube.com

Casey Wasserman ...
dailynews.com



When I'm a Powerful CEO ...
me.me

What is **fair?**

50% **female,** 50% **male?**
Based on the **population?**

Results for "CEO" in Google Images: 11% female, US 27% female CEOs

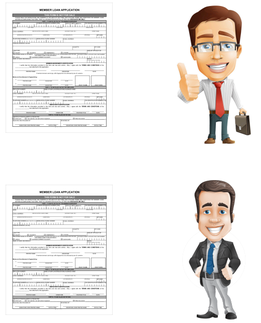
Types of fairness definitions

Different definitions based on legal concepts

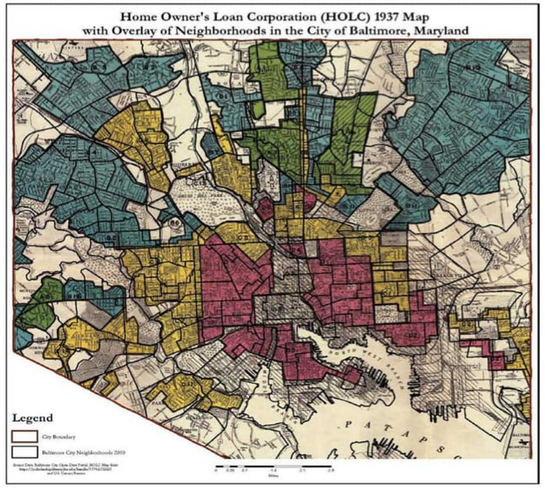
- Direct vs indirect discrimination
- Individual vs group fairness
- Explainable vs unexplainable discrimination

Indirect discrimination

Direct discrimination happens when a person is treated less favourably because of one of the attributes



Name	Postal code	...	Decision
Richard	H3C	=	REJECTED
Bob	F4C	=	APPROVED



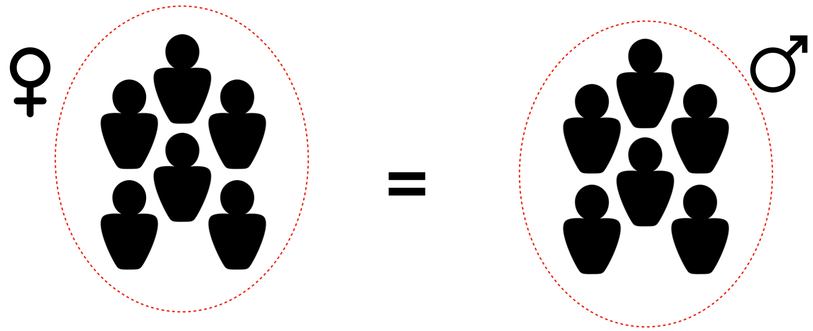
Indirect discrimination is when there's a practice, policy or rule which applies to everyone in the same way, but it has a worse effect on some people than others. The Equality Act says it puts you at a particular disadvantage.

Explainable discrimination

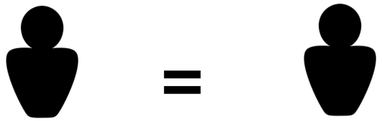
- Discrimination can be **explained** using attributes like working hours and education
- E.g., Disproportional recruitment rates for males and females may be explainable by the fact that more males have higher education
- Or males on average have a higher annual income than females because on average females work fewer hours per week than males do. Decisions made without considering working hours could lead to discrimination.

Group vs. Individual Fairness

- **Group fairness**: the impact that the discrimination has on the groups of individuals.

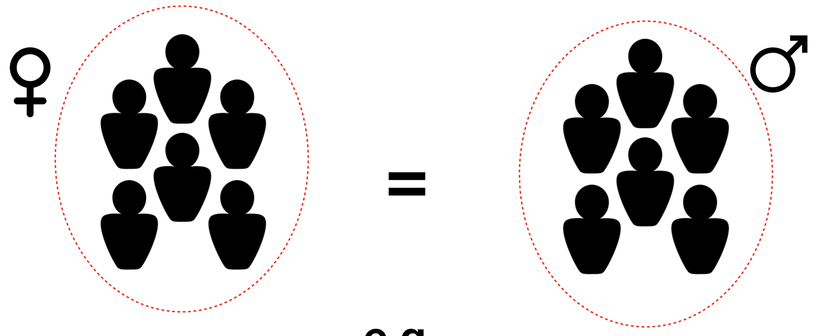


- **Individual fairness**: the impact that the discrimination has on the individuals.



Group vs. Individual Fairness

- **Group fairness**: the impact that the discrimination has on the groups of individuals.



e.g.,

Demographic Parity: The fraction of people given a **positive decision** should be equal across **different groups**.

decision **sensitive**

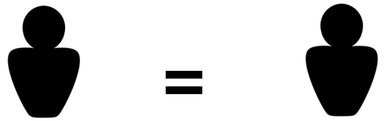
$$P(D = 1 \mid Y = 0) = P(D = 1 \mid Y = 1)$$

e.g.,

$$P(D = 1 \mid gender = Female) = P(D = 1 \mid gender = male)$$

Group vs. Individual Fairness

- **Individual fairness**: the impact that the discrimination has on the individuals.



e.g.,

Causal Discrimination: The **same decision** for any two subjects with the **exact same** attribute **Y** (**non-sensitive attributes**)

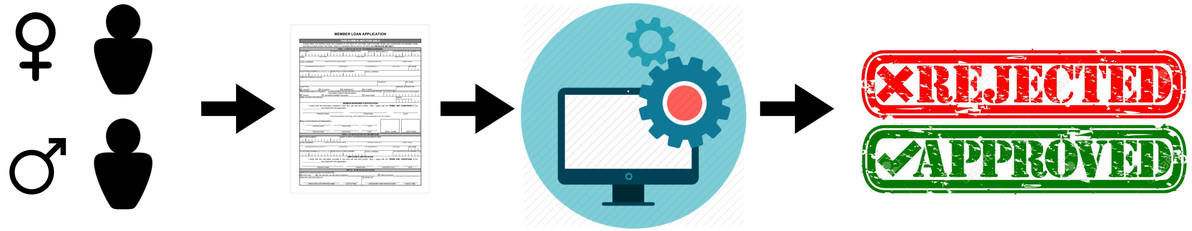
non-sensitive **sensitive**

$$(\mathbf{y}^{(m)} = \mathbf{y}^{(f)} \wedge (\mathbf{x}^{(m)} \neq \mathbf{x}^{(f)}) \rightarrow \mathbf{d}^{(m)} = \mathbf{d}^{(f)}$$

Notations

TN	FP
FN	TP

confusion matrix



Applicant Application Loan Approval

G sensitive attribute

X non-sensitive attributes

d Prediction decision

Y Actual Outcome

Female $G = f$

Male $G = m$

$d = 1$

Evaluation Metrics

1. True positive rate (TPR)

$$p(d = 1|Y = 1)$$

2. False positive rate (FPR)

$$p(d = 1|Y = 0)$$

3. False negative rate (FNR)

$$p(d = 0|Y = 1)$$

4. True negative rate (TNR)

$$p(d = 0|Y = 0)$$

d **Prediction decision**

Y **Actual Outcome**

Confusion Matrix

	Y=1	Y=0
d=1	TP	FP
d=0	FN	TN

- True positive (TP)
- False positive(FP)
- True negative (TN)
- False negative (FN)

Evaluation Metrics

5. Positive predictive value (PPV)

$$p(Y = 1|d = 1)$$

6. False discovery rate (FDR)

$$p(Y = 0|d = 1)$$

7. False omission rate (FOR)

$$p(Y = 1|d = 0)$$

8. Negative predictive value (NPV)

$$p(Y = 0|d = 0)$$

d **Prediction decision**

Y **Actual Outcome**

Confusion Matrix

	Y=1	**Y=0**
d=1	TP	FP
d=0	FN	TN

- True positive (TP)
- False positive(FP)
- True negative (TN)
- False negative (FN)

Evaluation Metrics

Metric	Formula	Evaluation Focus
Accuracy	$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	Overall effectiveness of a classifier
Error rate	$\mathrm{ERR} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	Classification error
Precision	$\mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	Class agreement of the data labels with the positive labels given by the classifier
Sensitivity	$\mathrm{SNS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	Effectiveness of a classifier to identify positive labels
Specificity	$\mathrm{SPC} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$	How effectively a classifier identifies negative labels
ROC	$ROC = \frac{\sqrt{SNS^2+SPC^2}}{\sqrt{2}}$	Combined metric based on the Receiver Operating Characteristic (ROC) space
F_1 score	$F_1 = 2\frac{PRC \cdot SNS}{PRC+SNS}$	Combination of precision (PRC) and sensitivity (SNS) in a single metric
Geometric Mean	$GM = \sqrt{SNS \cdot SPC}$	Combination of sensitivity (SNS) and specificity (SPC) in a single metric

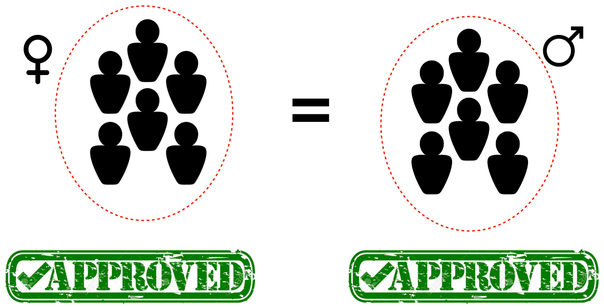
Group fairness

a predicted outcome

1- Group fairness / **statistical (demographic) parity** / equal acceptance rate / benchmarking

$$p(d = 1|G = f) = p(d = 1|G = m)$$

equal probability of being assigned to the positive predicted class



Group fairness

a predicted outcome

Issues with demographic parity:

$$p(d = 1|G = f) = p(d = 1|G = m)$$

Do you know what is wrong with this notion?

Group fairness

a predicted outcome

Issues with demographic parity:

$$p(d = 1|G = f) = p(d = 1|G = m)$$

1. The notion permits that a classifier selects qualified applicants in one group, but unqualified individuals in the other group

Group fairness

a predicted outcome

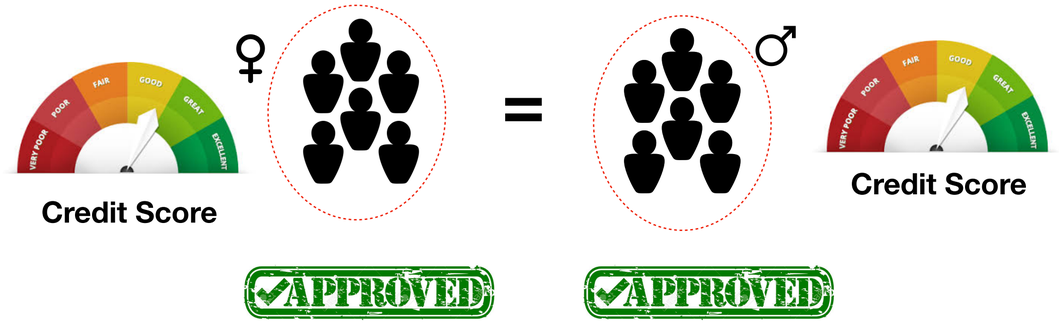
2- **Conditional statistical parity**

$$p(d = 1|L = 1, G = f) = p(d = 1|L = 1, G = m)$$

legitimate factors

L

both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of legitimate factors L.



Group fairness

a predicted outcome

Issues with demographic parity:

$$p(d = 1|G = f) = p(d = 1|G = m)$$

1. The notion permits that a classifier selects qualified applicants in one group, but unqualified individuals in the other group

2. Demographic parity would rule out the ideal predictor

Group fairness

a predicted outcome+ Actual outcome

3- False negative error rate balance / **equal opportunity**

$$p(d = 0|Y = 1, G = f) = p(d = 0|Y = 1, G = m)$$

$$=$$

$$p(d = 1|Y = 1, G = f) = p(d = 1|Y = 1, G = m)$$

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Group fairness

a predicted outcome+ Actual outcome

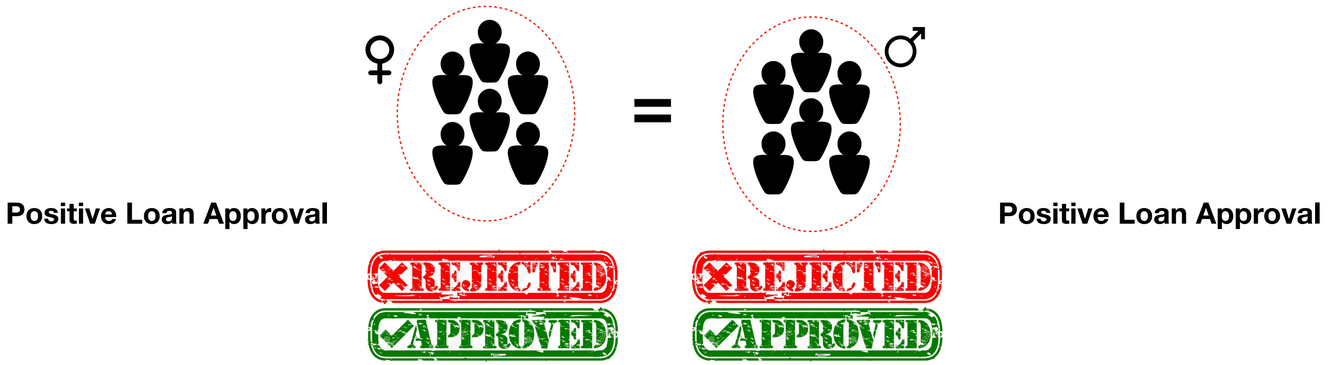
3- False negative error rate balance / **equal opportunity**

$$p(d = 0|Y = 1, G = f) = p(d = 0|Y = 1, G = m)$$

$$=$$

$$p(d = 1|Y = 1, G = f) = p(d = 1|Y = 1, G = m)$$

classifier should give similar results to applicants of both genders with actual positive loan approval.



Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Group fairness

a predicted outcome+ Actual outcome

3- False negative error rate balance / **equal opportunity**

$$p(d = 0|Y = 1, G = f) = p(d = 0|Y = 1, G = m)$$

$$=$$

$$p(d = 1|Y = 1, G = f) = p(d = 1|Y = 1, G = m)$$

Picks for each group a threshold such that the fraction of non-defaulting group members that qualify for loan is the same.

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Group fairness

a predicted outcome+ Actual outcome

4- **Equalized odds** / conditional procedure accuracy equality / disparate mistreatment

$$p(d = 1|Y = I, G = f) = p(d = 1|Y = I, G = m)$$

where $I \in \{0, 1\}$

Positive Credit Approval

Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Group fairness

a predicted outcome+ Actual outcome

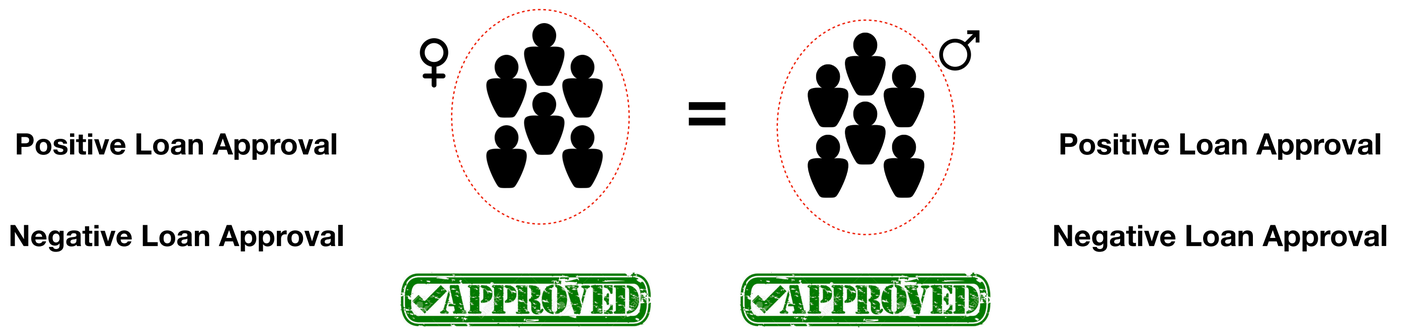
4- **Equalized odds** / conditional procedure accuracy equality / disparate mistreatment

$$p(d = 1|Y = I, G = f) = p(d = 1|Y = I, G = m)$$

where $I \in \{0, 1\}$

Positive Credit Approval

applicants with a rejected loan application and applicants with an accepted loan application should have a similar classification, regardless of their gender.



Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Group fairness

a predicted outcome+ Actual outcome

4- **Equalized odds** / conditional procedure accuracy equality / disparate mistreatment

$$p(d = 1|Y = I, G = f) = p(d = 1|Y = I, G = m)$$

where $I \in \{0, 1\}$

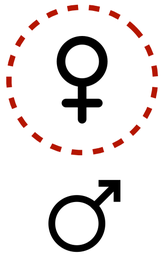
Picks two thresholds for each group, so above both thresholds people always qualify and between the thresholds people qualify with some probability.

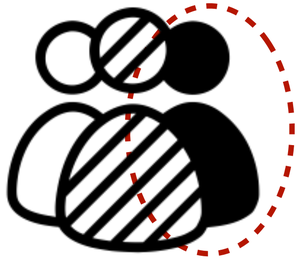
Hardt, M., Price, E. and Srebro, N., 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems (pp. 3315-3323).

Individual fairness

1- Fairness through unawareness, **Fairness through blindness**

$$X : X_i = X_j \rightarrow d_i = d_j$$

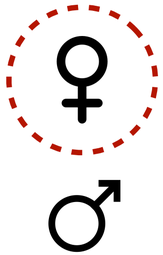


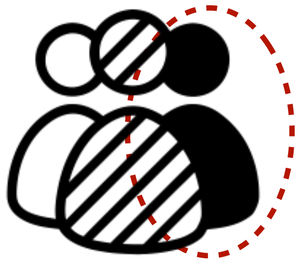


Individual fairness

1- Fairness through unawareness, **Fairness through blindness**

$$X : X_i = X_j \rightarrow d_i = d_j$$





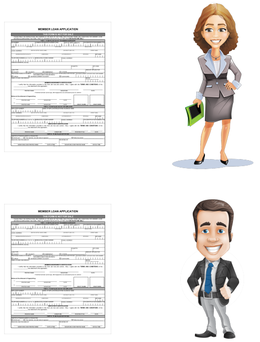
This can be impossible to hold because of non-obvious encoding in terms of many features, learned from the data

Individual fairness

2- **Causal discrimination**

$$(X_f = X_m \wedge G_f \neq G_m) \rightarrow d_f = d_m$$

the same classification for any two subjects with the exact same attributes X



Name	Gender	...	Decision
Alice	female	=	APPROVED
Bob	male	=	APPROVED

This can be impossible due to dependency between features!

Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness testing: testing software for discrimination." *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*. ACM, 2017.

Individual Fairness

3- **Fairness through awareness**

$$D(M(x), M(y)) \rightarrow k(x, y)$$

Distance metric Between two Distributions M(x), M(y)

D

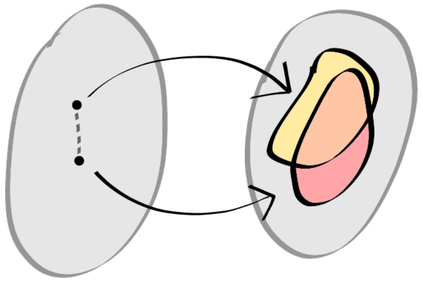
Distance metric Between two individuals x,y

k

similar individuals should have similar classification

seemingly different individuals

Name	Gender	...	Decision
Alice	female	=	APPROVED
Bob	male	=	APPROVED



Dwork, Cynthia, et al. "Fairness through awareness." *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012.

Impossibility theorem

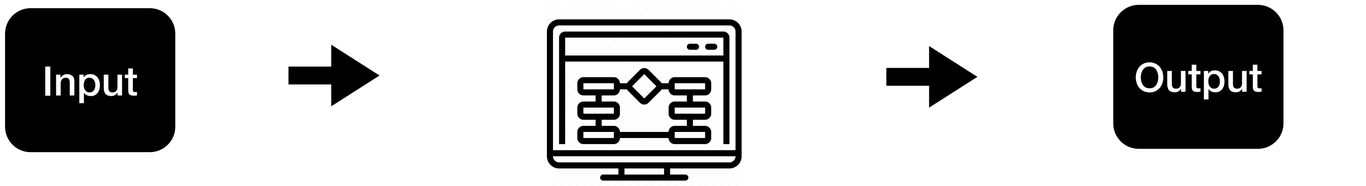
Metric	Equalized under
Selection probability	Demographic parity
Positive predictive value	Predictive parity
Negative predictive value	Predictive parity
False positive rates	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity

You can only achieve one of these measures: demographic parity, equality of odds, and equality of opportunity

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807* (2016).

Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big data* 5.2 (2017): 153-163.

How to address fairness in AI?



bias

Data is unbalanced
Historical discrimination
Encodes protected attributes

Data scientists do not build the models

Unfair outcome
Black-box models
No user feedback

Pre-processing

Discrimination Discovery
Sampling/weighting
Representation learning

In-processing

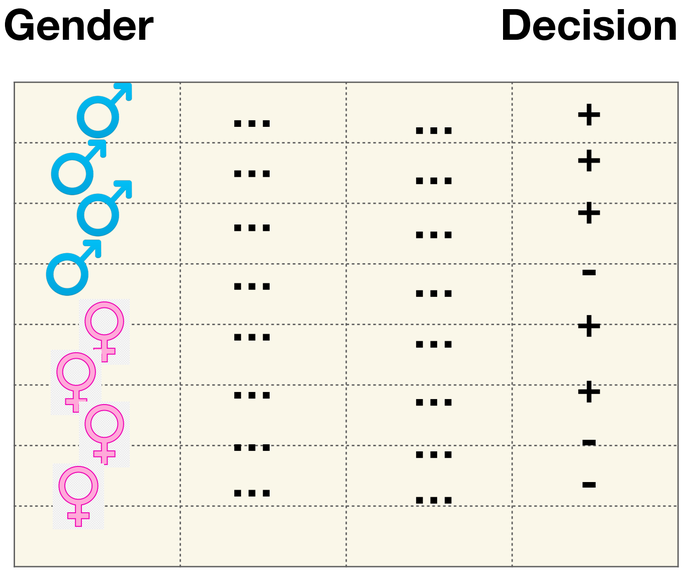
Learning/Inference subject to constraints
Ranking
Constrained Optimization

Post-processing

Causal discovery
Explainable AI
Verification

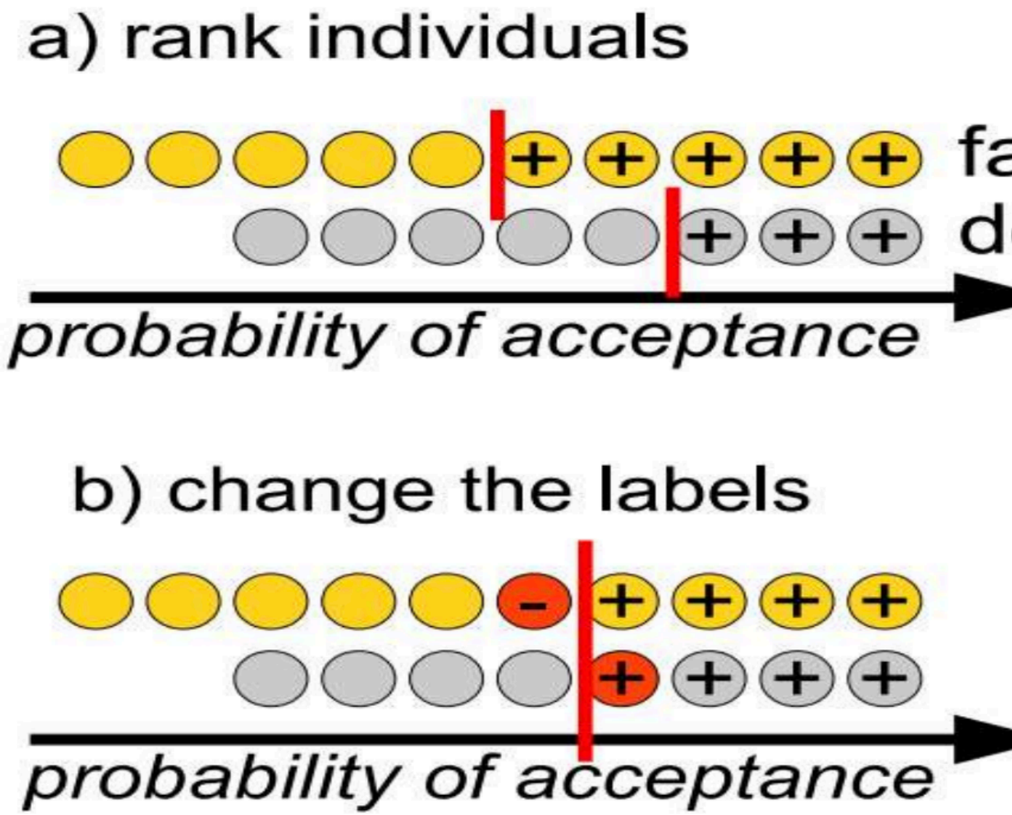
Some pre-processing techniques

- **Massaging**
- **Re-weighting**
- **Sampling**
- **….**



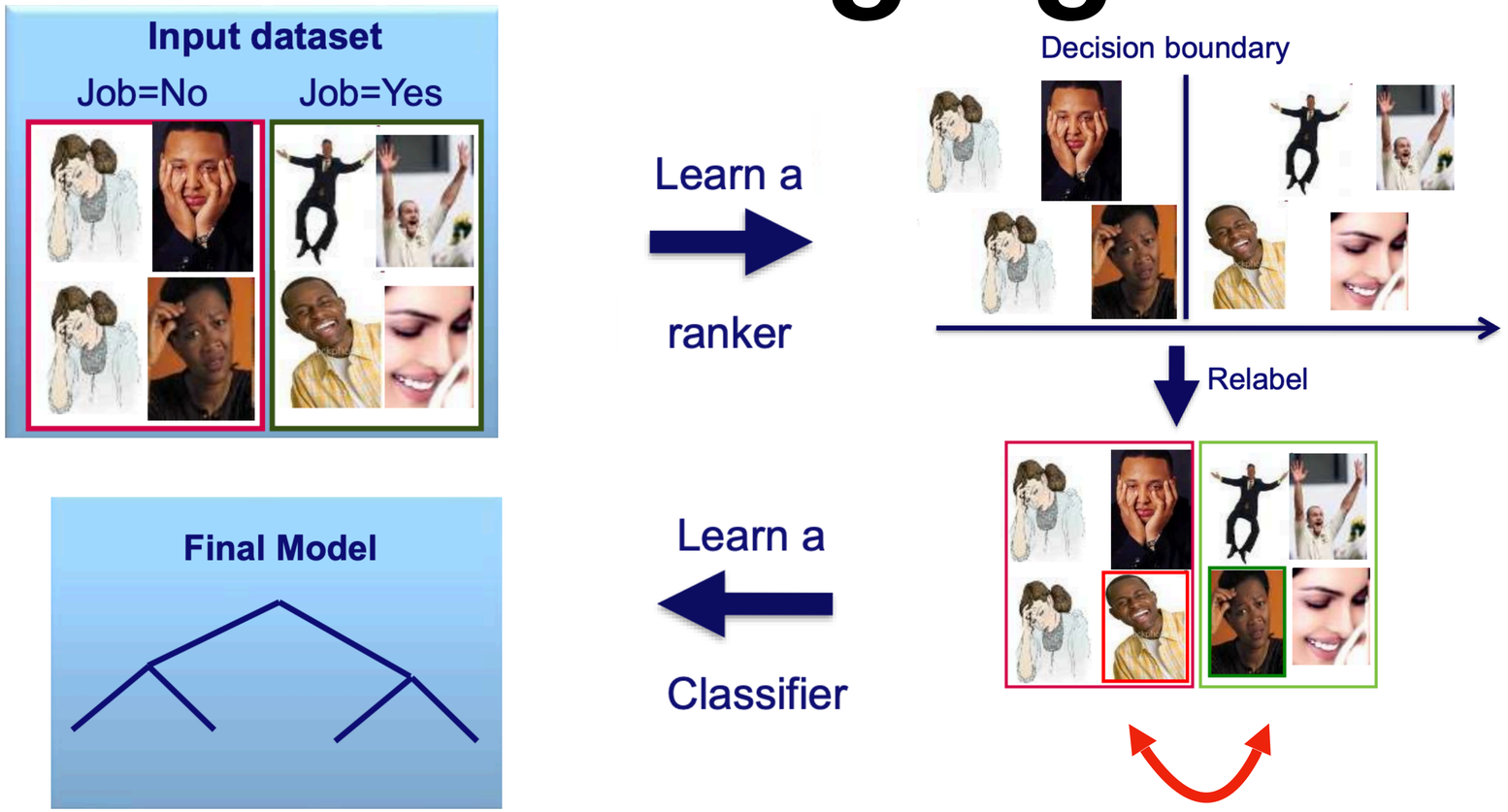
Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Massaging



Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Massaging

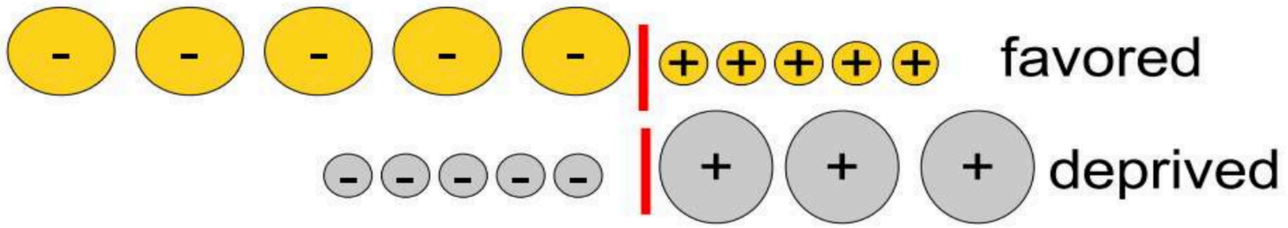


Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

Re-Weighting

a) calculate weights for the objects to neutralize the discriminatory effects from data

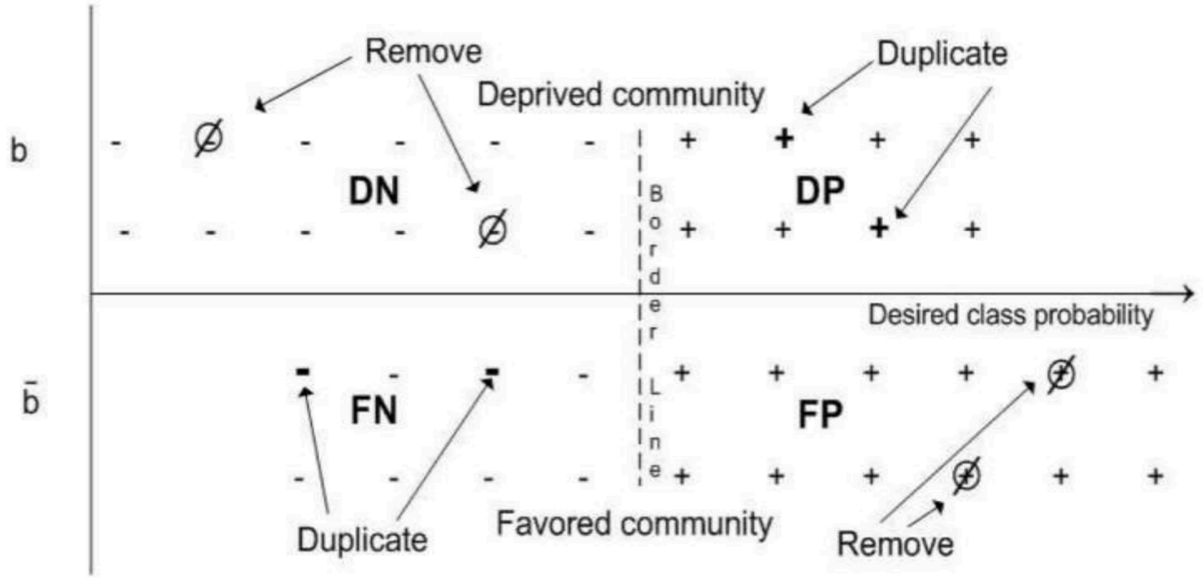
b) assign weights to make the data impartial



Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

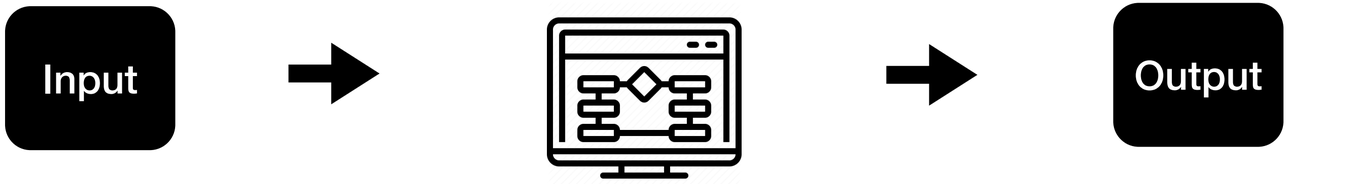
Sampling

Similarly to reweighing, compare the expected size of a group with its actual size, to define a sampling probability.



Hajian, Sara, Francesco Bonchi, and Carlos Castillo. "Algorithmic bias: From discrimination discovery to fairness-aware data mining." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016.

How to address fairness in AI?



bias

Data is unbalanced
Historical discrimination
Encodes protected attributes

Data scientists do not build the models

Unfair outcome
Black-box models
No user feedback

Pre-processing

Discrimination Discovery
Sampling/weighting
Representation learning

In-processing

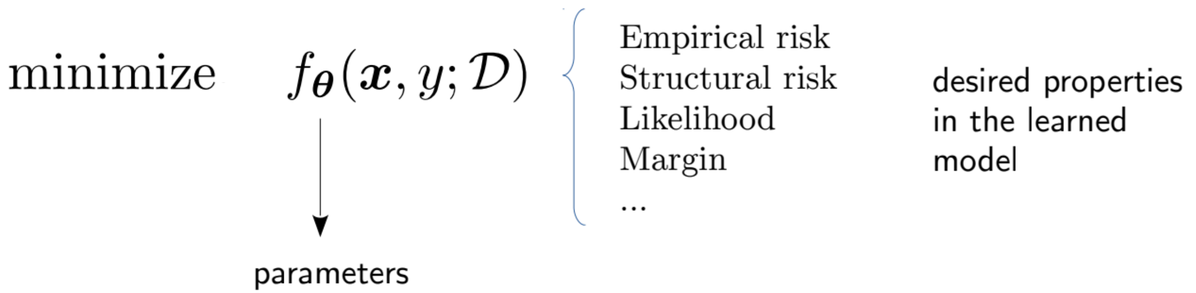
Learning/Inference subject to constraints
Ranking
Constrained Optimization

Post-processing

Causal discovery
Explainable AI
Verification

Learning subject to fairness constrains

Supervised learning tasks are often expressed as optimization problems



The optimization problem: finding the parameters that give the best model w.r.t the desired properties

Fairness is yet another desired property of the learned models

Accuracy is not enough!

A hypothetical (extreme) situation:

Born and raised in Canada

- data describes them accurately
- accurate predictions (95% accurate)

90% of population

The model is still ~90.5% accurate!

Migrated to Canada in recent years

- data describes them poorly
- poor predictions (50% accurate)

10% of population

Learning subject to fairness constrains

- Not all optimization problems are the same!
- Some problems are **computational easy**
- Some problems are **hard**, but **behave well** (approximation methods work well)
- Some problems are **hard**, but have **structure**. And we can exploit this structure.

Learning subject to fairness constrains

- Not all optimization problems are the same!
- Some problems are **computational easy**
- Some problems are **hard**, but **behave well** (approximation methods work well)
- Some problems are **hard**, but have **structure**. And we can exploit this structure.

Adding fairness constraints can change these properties!

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

fairness measures

$$g_{\theta}(x, y; D)$$

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

e.g., demographic parity

$$p(d = 1|G = f) = p(d = 1|G = m)$$

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

e.g., demographic parity

$$p(d = 1|G = f) = p(d = 1|G = m)$$

Equality constraints are hard to satisfy

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

e.g., demographic parity

$$p(d = 1|G = f) = p(d = 1|G = m)$$

$$\Delta_{fair} = |p(d = 1 \,|\, G = f) - p(d = 1 \,|\, G = m)|$$

Equality constraints are hard to satisfy

$$\delta - fair$$

$$\Delta_{fair} \leq \delta$$

Learning subject to fairness constrains

Supervised learning tasks under fairness constraints are often expressed as **constrained optimization problems**

loss function

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D})$$

s.t

$$\Delta_{fair} \leq \delta$$

Learning subject to fairness constrains

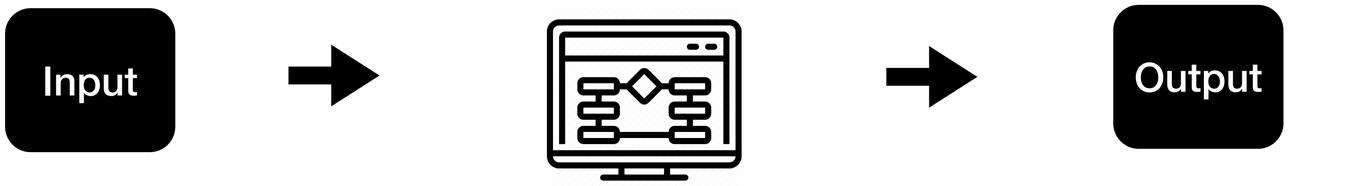
Supervised learning tasks under fairness constraints are sometimes expressed as regularization in an optimization problems

$$\text{minimize.} \quad f_{\boldsymbol{\theta}}(\boldsymbol{x}, y; \mathcal{D}) \quad + \quad \lambda \times \Delta_{fair}$$

method of Lagrange multipliers

Fair Regularizer

How to address fairness in AI?



bias

Data is unbalanced
Historical discrimination
Encodes protected attributes

Data scientists do not build the models

Unfair outcome
Black-box models
No user feedback

Pre-processing

Discrimination Discovery
Sampling/weighting
Representation learning

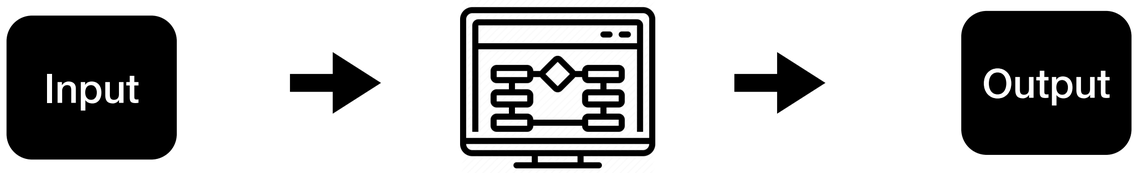
In-processing

Learning/Inference subject to constraints
Ranking
Constrained Optimization

Post-processing

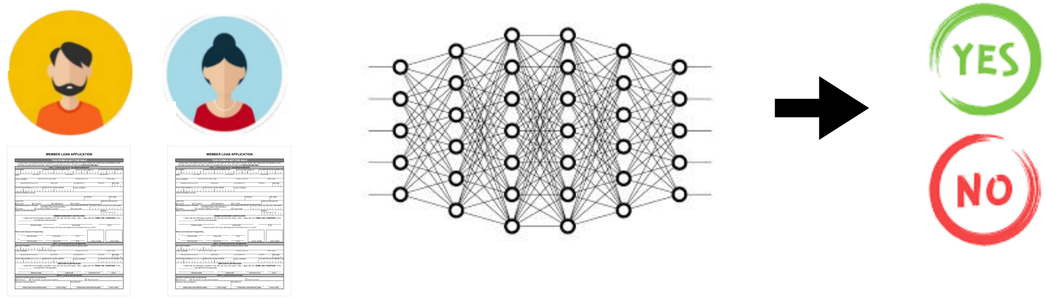
Causal discovery
Explainable AI
Verification

Challenges & Opportunities: Fairness Vs. Explainability

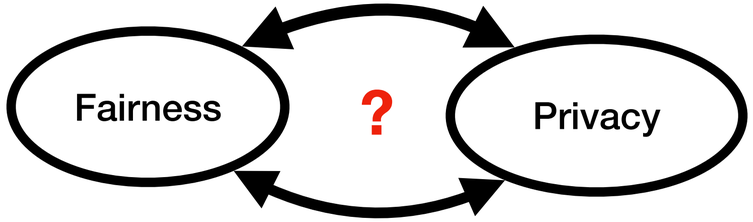


Counterfactual explanation: A **counterfactual explanation** describes a causal situation in the form: "If X had not occurred, Y would not have occurred"

Algorithmic recourse: provides explanations and recommendations to individuals who are treated unfavourably by the algorithm



Challenges & Opportunities: Fairness Vs. Privacy



Law against discrimination

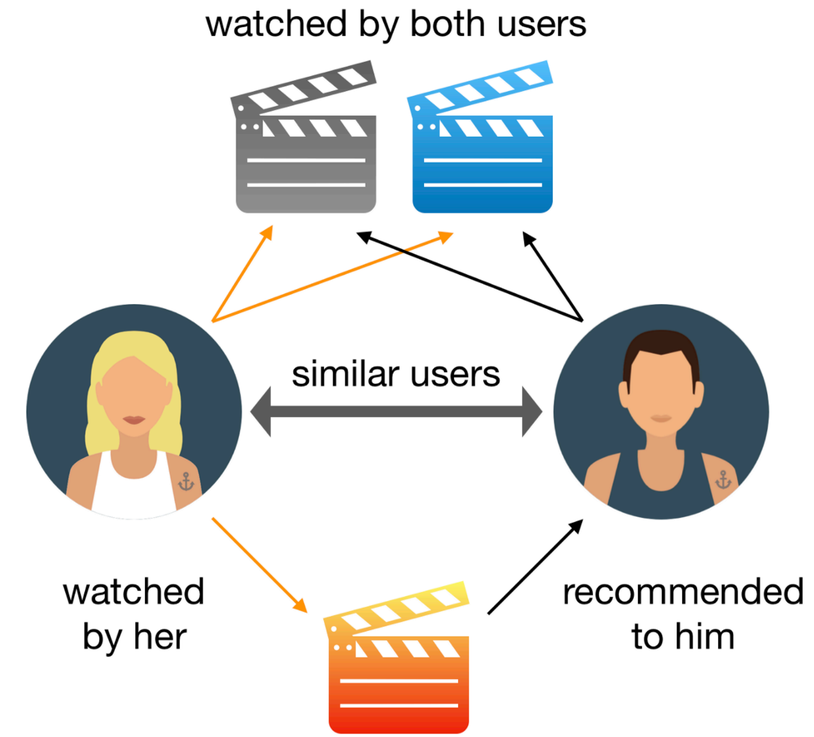
Legally recognized ‘protected classes’

Regulated domains

Data protection laws

Challenges & Opportunities: Fairness is not the same in different fields!

Two-sided fairness in recommender systems



Take aways

Bias happens throughout the automated systems:

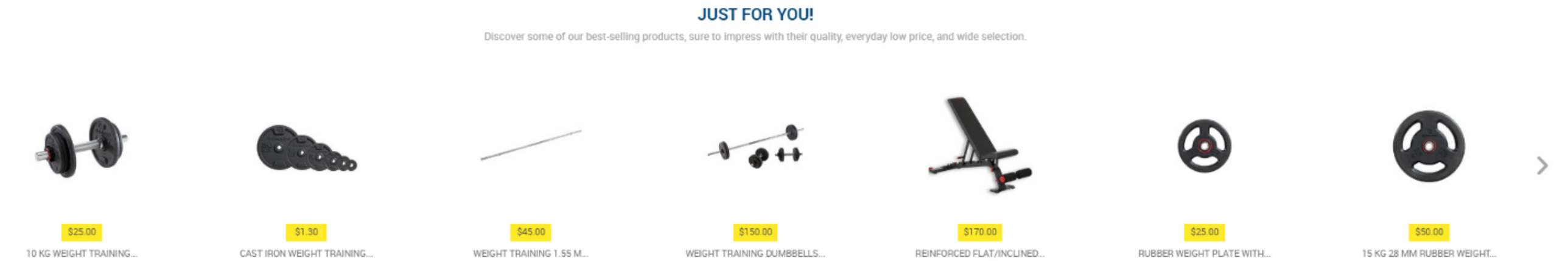
- Educate people about **discrimination**
- How to **define fairness** in your set-up?
- Ask who is **using** the model?
- What is **the purpose** of the system?

Recommendation Systems for Décathlon - Discussion

Case developed by Jérémi DeBlois-Beaucage supervised by Laurent Charlin & Renaud Legoux

Presentation of the case

Q1: Which model(s) for a recommendation system would the data science team need to choose, and why?



Session in small groups

- ~15 minutes, groups of ~5, prepare 3—4 slides
    - Suggestion: designate one scribe and one presenter by team
- Then: we will discuss your answers in class

Discussions

Plan

- Choose the right model for a recommendation task
- Models chosen by Décathlon
  - Basic models, as reference
  - Model 1: Based on similarity between product images
  - Model 2: Collaborative filtering based on products
  - Model 3: Matrix factorization
  - Model 4: Recursive neuronal networks
  - Chosen metrics
  - Results and final choice
- Limits of the current models and the next steps being considered by Décathlon

How to Choose the Right Model?

- A more subjective task than most other common machine learning tasks
- Imperative: the model must be able to process a large amount of data
- Started with simpler models and then moved on to more complex ones
- Final choice according to performance on chosen metrics and logistical considerations
- 4 models have been selected

Basic Models: Reference Point

- Random recommendation
- Recommending the same 10 most popular items to every user

Types of Recommender Systems

Content Filtering

- *Example*: Pandora.com music recommendations (Music Genome Project)
- **Con:** Assumes access to side information about items (e.g. properties of a song)
- **Pro:** Got a new item to add? No problem, just be sure to include the side information

Collaborative Filtering

- *Example*: Netflix movie recommendations
- **Pro:** Does not assume access to side information about items (e.g. does not need to know about movie genres)
- **Con:** Does not work on new items that have no ratings

Model 1: Based on Similarity Between Product Images

Each product is first assigned an image and a vector that represents this image, created through a pre-trained VGG-type convolutional neural network.

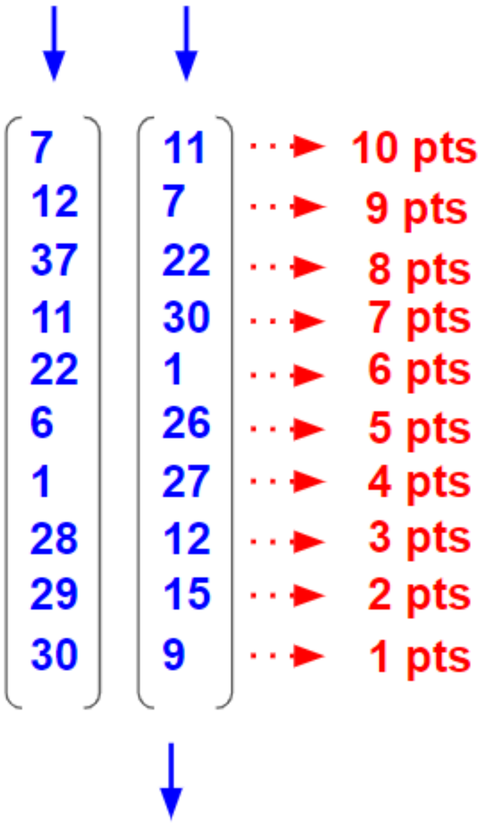
1. For each user, a list of all the products with which they have interacted is extracted.

2. For each product, the 10 most “similar” products are chosen, based on the cosine distance between their image vectors.

3. The most similar product gets 10 “points”, the second one gets 9, and so on. Points are added up, and the 5 products with the highest scores get recommended.

User A has interacted in the past with items **3** and **5**

For each item, we identify the 10 items in the rest of the catalog that are most similar. We give 10 pts to the most similar one, 9 pts to the second most, and so on



We sum all the scores, and recommend the top five items to our user

Recommendations to user **A**:
Item **7** (19 pts)
Item **11** (17 pts)
Item **22** (14 pts)
Item **12** (12 pts)
Item **1** (10 pts)

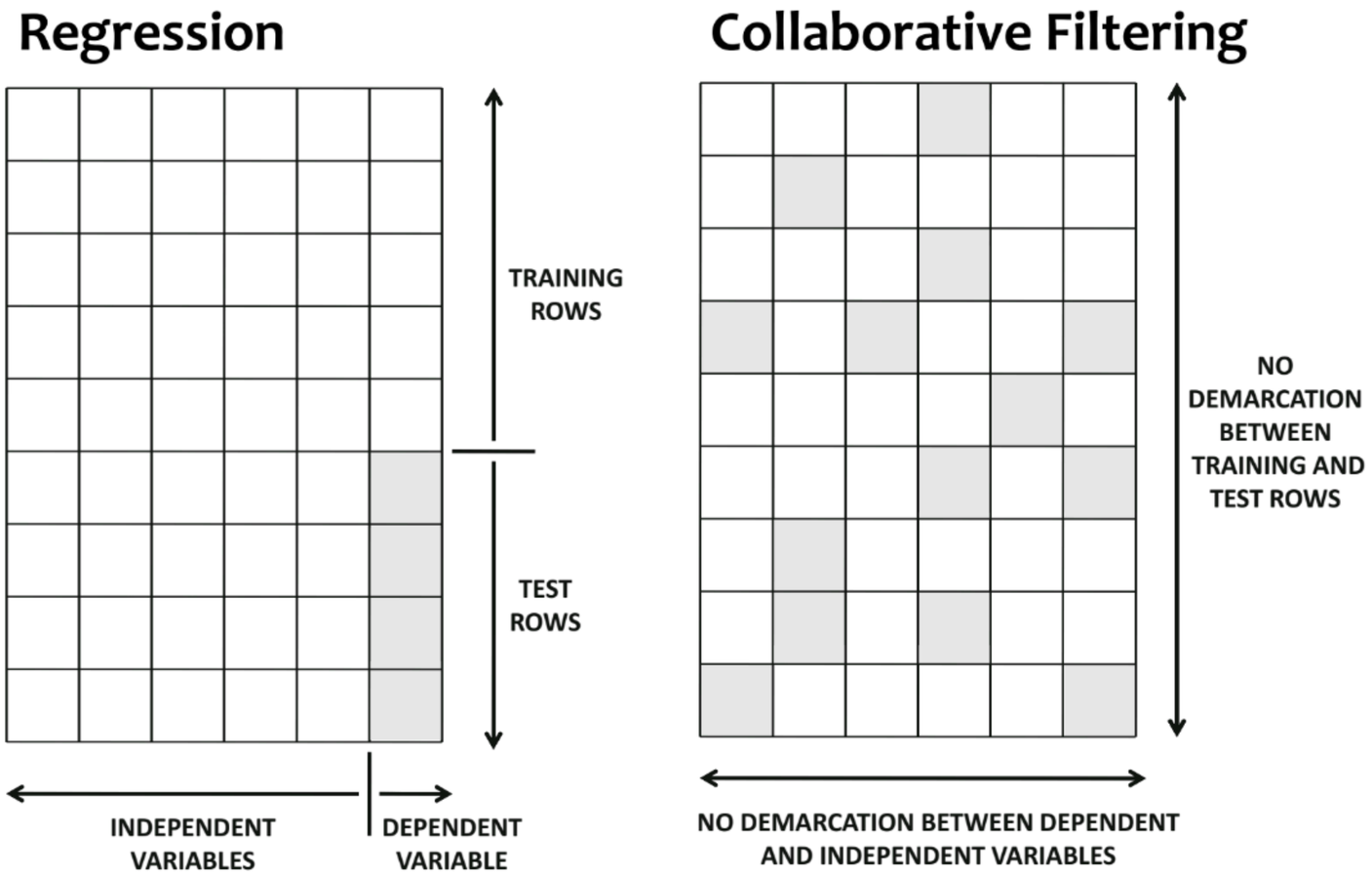
Pros and Cons

- (+) Even new or unpopular products can get recommended
- (+) Explanations can be offered (because you bought product X, you could find these products interesting)
- (+) Easy and quick to implement
- (-) The recommended products are very similar to previously purchased items
- (-) *Cold-start* problem for the users

Collaborative Filtering

- **Everyday Examples of Collaborative Filtering…**
  - Bestseller lists
  - Top 40 music lists
  - The "recent returns" shelf at the library
  - Unmarked but well-used paths thru the woods
  - The printer room at work
  - "Read any good books lately?"
  - …
- **Common insight:** personal tastes are correlated
  - If Alice and Bob both like X and Alice likes Y then Bob is more likely to like Y
  - especially (perhaps) if Bob knows Alice

Regression vs. Collaborative Filtering



Model 2: Collaborative Filtering Based on Products

Very similar model to the previous one. Instead of calculating the similarity of image vectors, similarity is calculated according to a user-product interaction matrix.

1. For each user, a list of all the products with which they have interacted is extracted.

2. For each product, the 10 most “similar” products are chosen, based on a cosine distance between their respective lines in the user-product interaction matrix.

3. The final score is similar to the one produced by model 1. However, instead of attributing arbitrary points (10 pts for the 1st, 9 for the 2nd, etc.), the similarity scores are used directly. Points are added, and the 5 products with the highest scores get recommended.

Example: similarity between two products

Products

```
[0, 0, 1, 0, 1, 0]
[0, 1, 0, 0, 0, 0]
[0, 0, 1, 0, 1, 0]
[1, 0, 0, 0, 1, 0]
```

Users

- In this matrix, lines represent users and columns represent products. A value of 1 indicates an interest, and 0 means no interaction.
  - For example, the first column shows that only User 4 has interacted with Product 1.
- The cosine similarity between the first (a =[0 0 0 1]) and the second product (b=[0 1 0 0]) would be 0. No user has interacted with both products.
- The similarity between the third (c=[1 0 1 0]) and the fifth product (e=[1 0 1 1]) is 0.82. The nearer the value is to 1, the greater the similarity between the products.

Pros and Cons

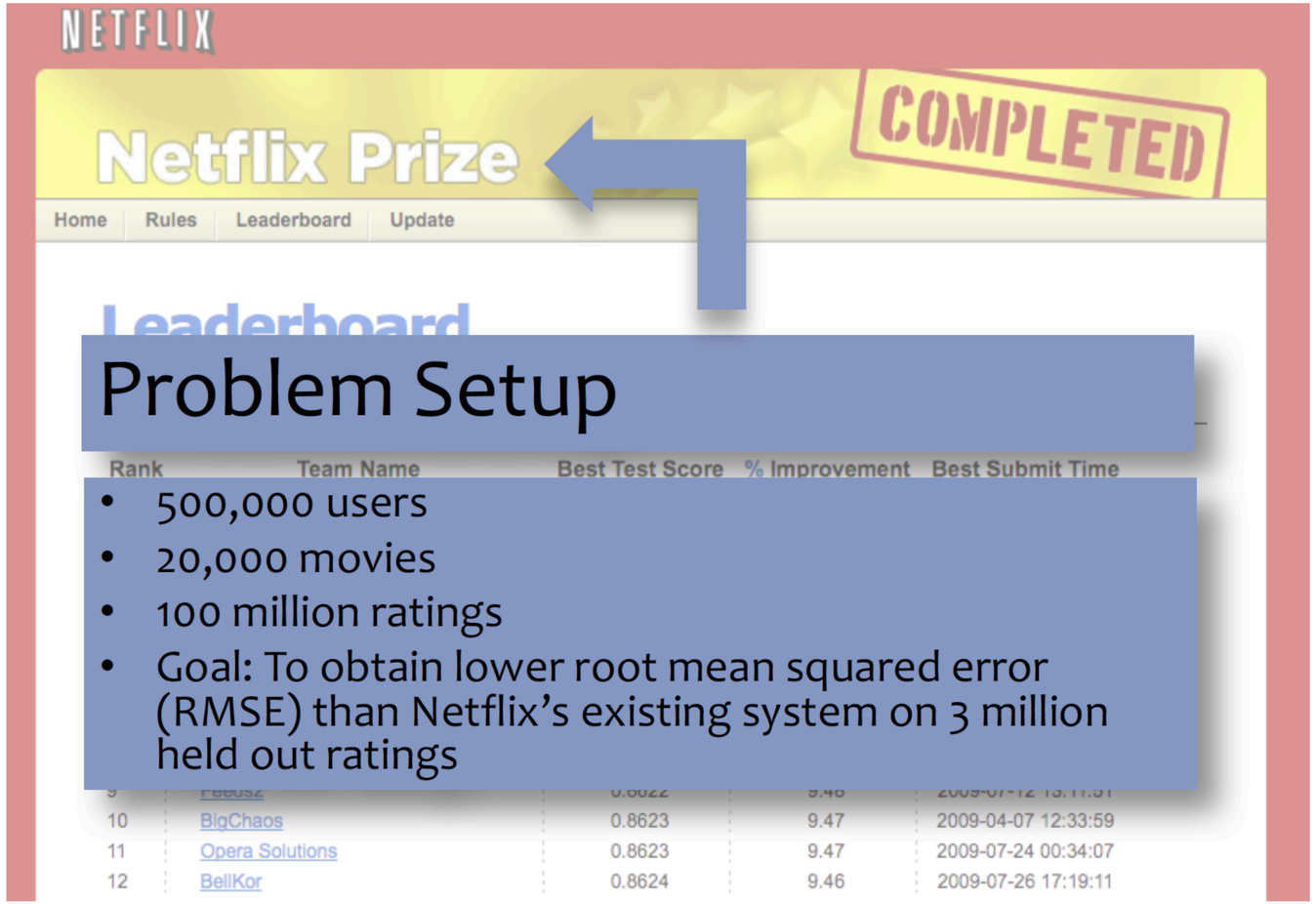
- (+) Explanations can be offered (because you bought product X, you could find these products interesting)
- (+) Easy and quick to implement
- (+) Collaborative filtering usually yields good results
- (-) Cold-start problem for the users and the products

Model 3 : Matrix Factorization

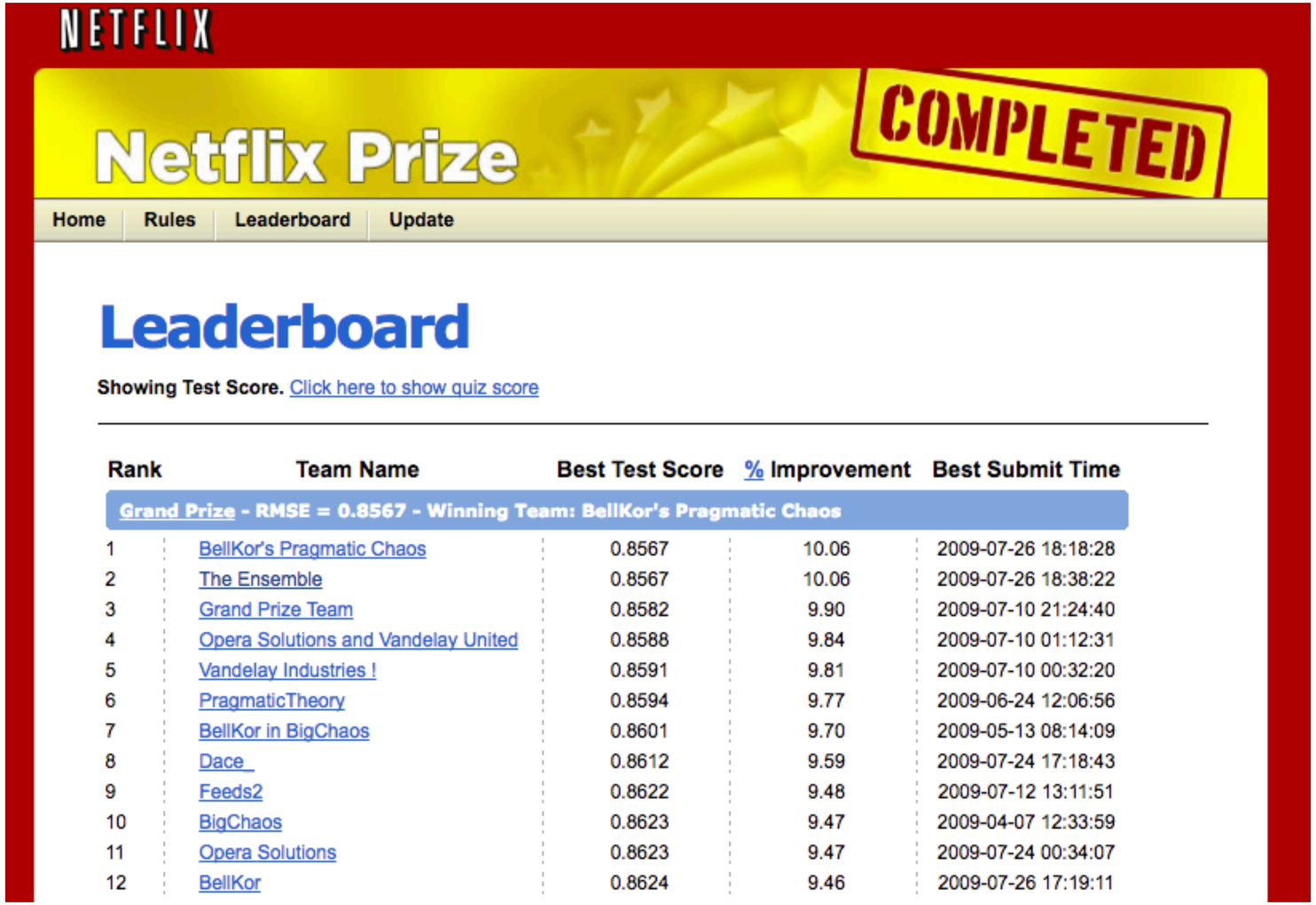
- Completion of the user-product interaction matrix through matrix factorization
  - Several methodologies have been used, like Singular Value Decomposition and Non-Negative Matrix Factorization.
- It predicts the probability of an interaction with each product. Recommended products are the one with the highest probabilities of interaction.

Netflix Prize

Problem Setup



Leaderboard



Matrix Factorization

MATRIX FACTORIZATION TECHNIQUES FOR RECOMMENDER SYSTEMS

Yehuda Koren, *Yahoo Research*
Robert Bell and Chris Volinsky, *AT&T Labs—Research*

As the Netflix Prize competition has demonstrated, matrix factorization models are superior to classic nearest-neighbor techniques for producing product recommendations, allowing the incorporation of additional information such as implicit feedback, temporal effects, and confidence levels.

Modern consumers are inundated with choices. Electronic retailers and content providers offer a huge selection of products, with unprecedented opportunities to meet a variety of special needs and tastes. Matching consumers with the most appropriate products is key to enhancing user satisfaction and loyalty. Therefore, more retailers have become interested in recommender systems, which analyze patterns of user interest in products to provide personalized recommendations that suit a user's taste. Because good personalized recommendations can add another dimension to the user experience, e-commerce leaders like Amazon.com and Netflix have made recommender systems a salient part of their websites.

Such systems are particularly useful for entertainment products such as movies, music, and TV shows. Many customers will view the same movie, and each customer is likely to view numerous different movies. Customers have proven willing to indicate their level of satisfaction with particular movies, so a huge volume of data is available about which movies appeal to which customers. Companies can analyze this data to recommend movies to particular customers.

RECOMMENDER SYSTEM STRATEGIES

Broadly speaking, recommender systems are based on one of two strategies. The *content filtering* approach creates a profile for each user or product to characterize its nature. For example, a movie profile could include attributes regarding its genre, the participating actors, its box office popularity, and so forth. User profiles might include demographic information or answers provided on a suitable questionnaire. The profiles allow programs to associate users with matching products. Of course, content-based strategies require gathering external information that might not be available or easy to collect.

A known successful realization of content filtering is the Music Genome Project, which is used for the Internet radio service Pandora.com. A trained music analyst scores

 Published by the IEEE Computer Society 0018-9162/09/$26.00 © 2009 IEEE

Yehuda Koren, Yahoo Research

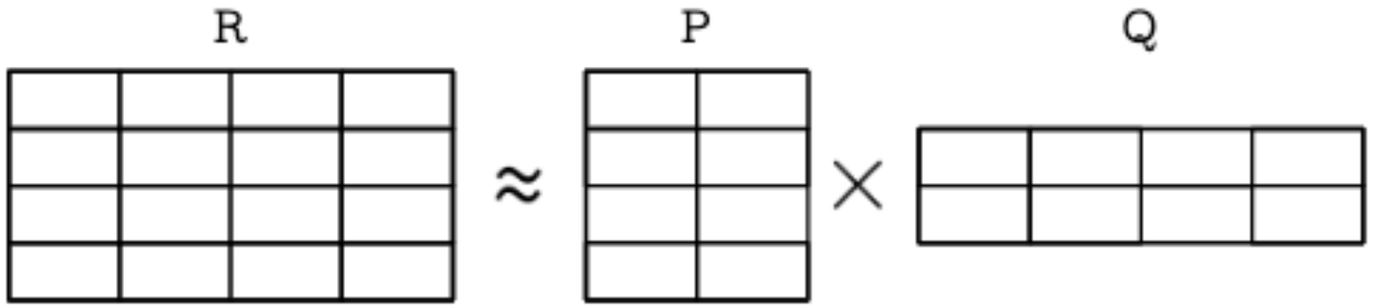
Robert Bell and Chris Volinsky, AT&T Labs-Research

Paper published in August 2009

Authors won the grand Netflix Prize in September 2009

Matrix Factorization

the completion is driven by a factorization



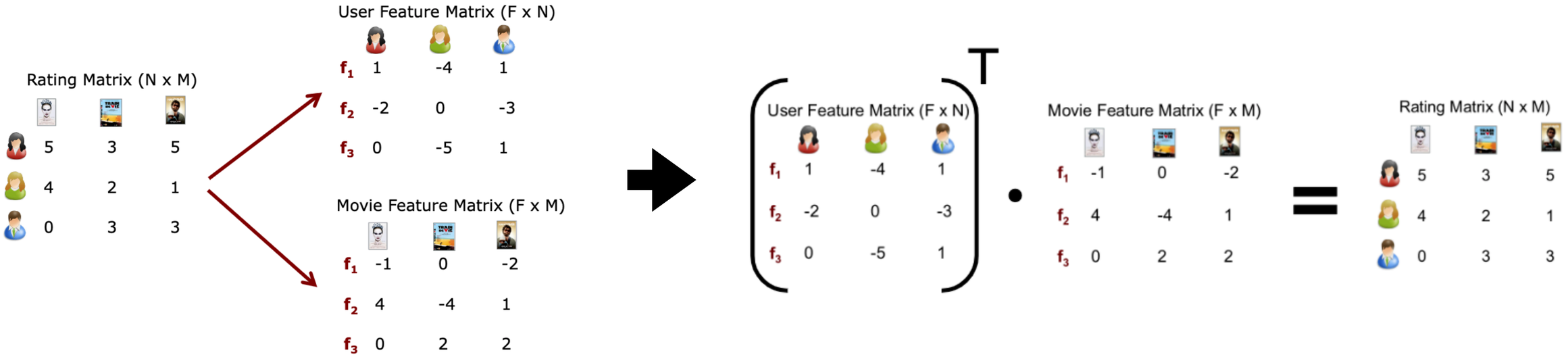
associate a latent factor vector with each user and each item

missing entries are estimated through the dot product

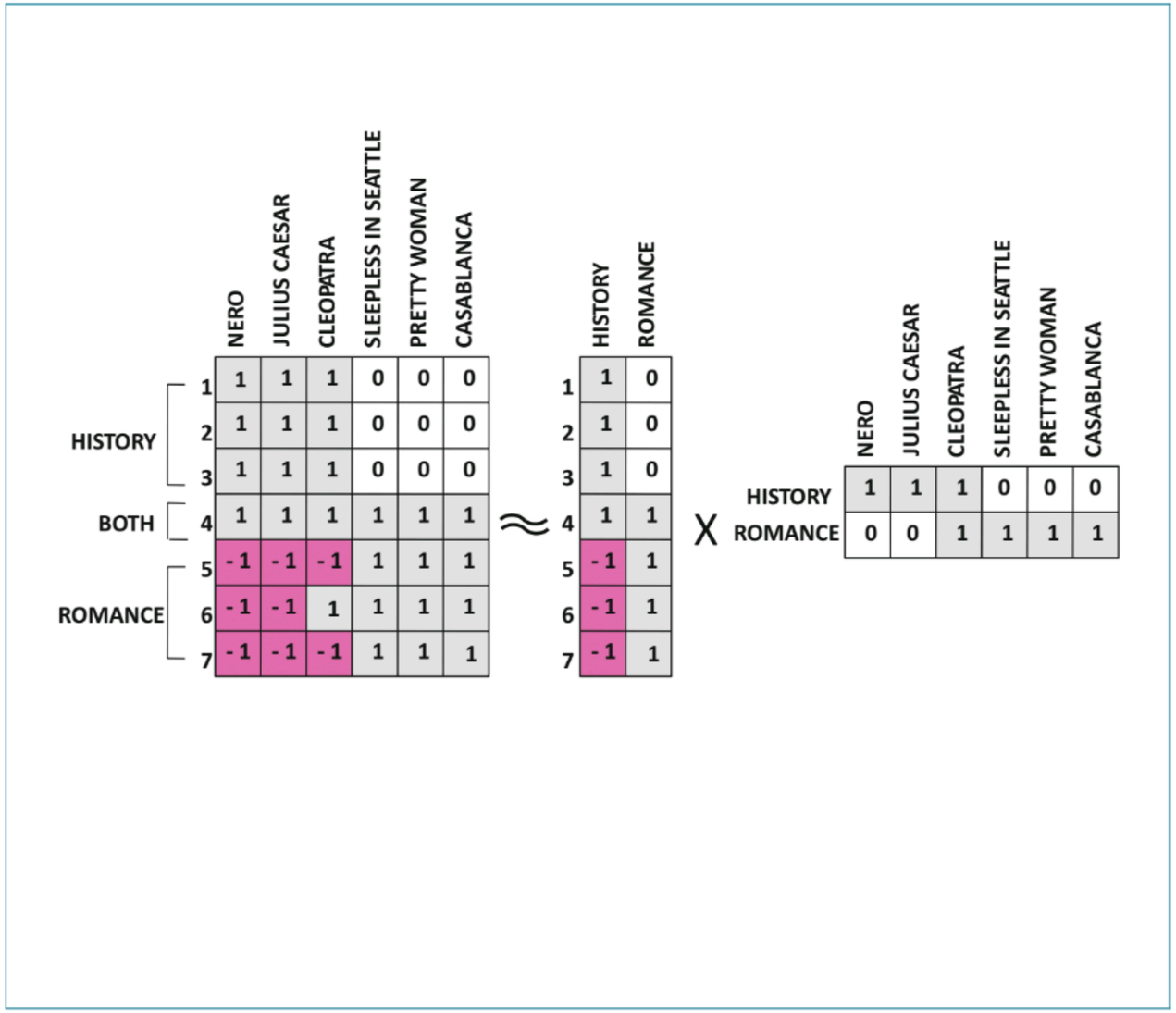
$$r_{ij} \approx p_i q_j$$

Slides are collected from:
Matt Gormley at CMU
Markus Freitag, Jan-Felix Schwarz from University Potsdam
Jure Leskovec from Stanford

Matrix Factorization



Example of MF for Netflix problem



Matrix Factorization

$$\min_{q^*, p^*} \sum_{(u,i) \in \kappa} (r_{ui} - q_i^T p_u)^2$$

r_{ui} : known rating of user u for item i

remember:
predicted rating $\hat{r}_{ui} = q_i^T p_u$

Regularization to avoid overfitting

Idea: penalize complexity

$$\min_{q^*,p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(|| q_i ||^2 + || p_u ||^2)$$

known as Funk SVD

λ : constant to control the extend of regularization

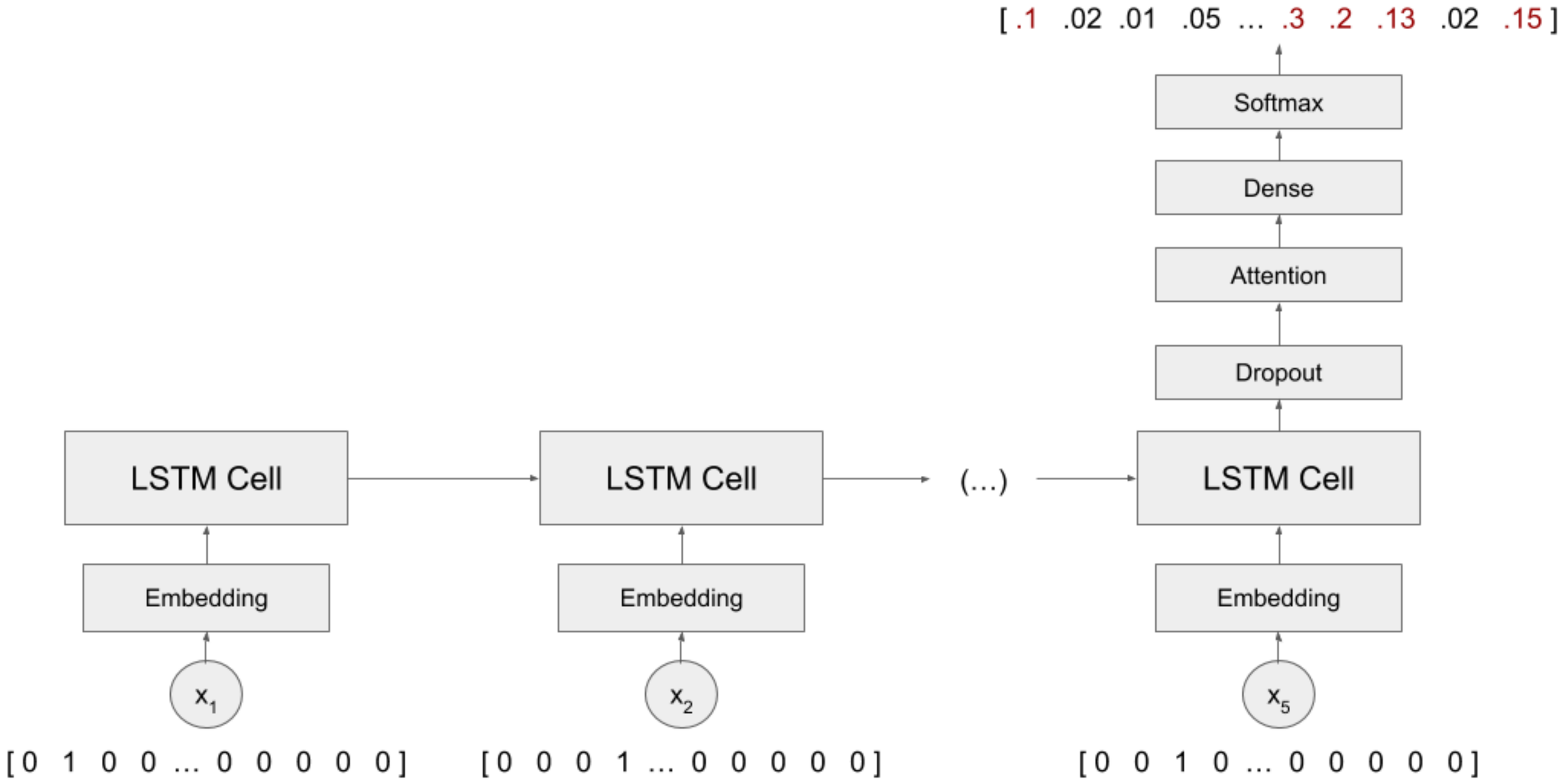
→ determined by cross-validation

Pros and Cons

- (+) Matrix factorization usually yields good results
- (+) It can reveal interesting underlying characteristics
- (-) Cold-start problem for the users and the products
- (-) Important computational costs
- (-) Deploying the models requires a more complex virtual infrastructure*

Model 4: Recurrent Neural Networks

- This model tackles the problem as a sequence of products with which the user interacts, and tries to predict what the next one could be.
  - Each product is represented by a one-hot vector.
    - Products are entered into the network sequentially, and the network predicts the next one.
  - The network then outputs a probability distribution for every product in the catalogue.
  - Long Short-Term Memory (LSTM) neurons are used, with dropout-type regularization and attention principles.


[.1 .02 .01 .05 … .3 .2 .13 .02 .15]
Softmax
Dense
Attention
Dropout
LSTM Cell
LSTM Cell
(…)
LSTM Cell
Embedding
Embedding
Embedding
x1
x2
x5
[0 1 0 0 … 0 0 0 0 0]
[0 0 0 1 … 0 0 0 0 0]
[0 0 1 0 … 0 0 0 0 0]

Pros and Cons

- (+) New users can be easily added
- (+) Recurrent networks usually give very good results
- (-) Cold-start problem for the products
- (-) Works better if a user has interacted with several products

Metrics

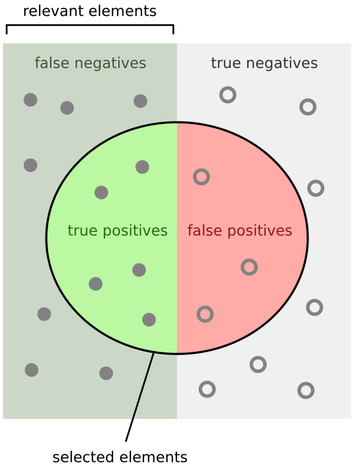
Only off-line metrics: the team does not have access to on-line evaluation or user studies

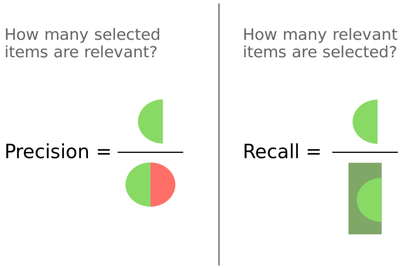
A. Accuracy: proportion of the recommended products that actually get bought by the users

B. Recall: proportion of products that were actually bought which were recommended

C. Coverage: proportion of products that were recommended to at least 1 user

Option to not consider diversity or serendipity

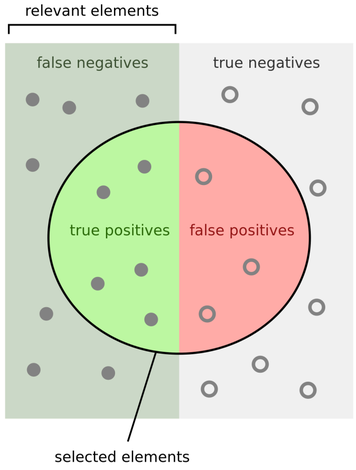


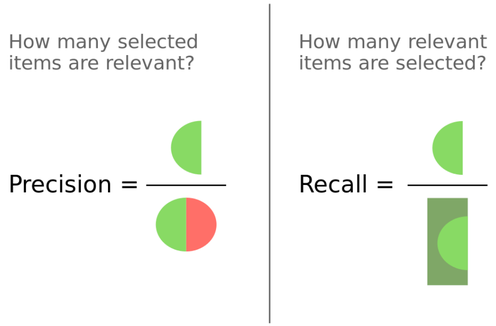


https://en.wikipedia.org/wiki/Precision_and_recall

Results

Model	Random	Most popular	1. Visual Similarity	2. Collaborative, based on products	3. Matrix factorization	4. Recurrent Neural Networks (RNNs)
Precision	0,06%	1,5%	1,9%	3,4%	3,9%	4%
Recall	0,07%	1,8%	2,3%	4,1%	5,3%	5,7%
Coverage	91%	0,07%	37,1%	74%	69%	57%





https://en.wikipedia.org/wiki/Precision_and_recall

Final Choice

First, simpler models were implemented:

1. Most popular products

2. Collaborative based on products (Model 2)

The visual similarity model (Model 1) did not show sufficient performance to justify its implementation.

Matrix factorization (Model 3) was put aside: the performance boost did not justify the investment in more complex logistical architecture.

The recurrent neural network model is in use now (Model 4)

Limits

- There’s an intuition that another model might perform better for the chosen metrics
  - The team is looking for a model that uses both the user-product interaction data and product characteristics.
- The current model focuses on short-term performance
  - The team would like a model that can further explore the diversity of user preference, and potentially offer pleasantly surprising products to users

To go further

What improvements could be made, or what more advanced models could be prioritized for testing?

Session in Small Groups

If enough time…

Improvements and New Models Being Considered

1. **Curiosity in recurrent neuronal networks**
   - Adding functionalities to the networks being used: adding curiosity techniques
   - Allows for a more thorough exploration of diversity in user preferences
2. **Graph neural networks**
   - Structure data differently: heterogeneous graph, which allows for the use of interaction data and product characteristics
   - Prediction of the links between the graph's knots
3. **Learning through reinforcement**
   - Model the task differently: sequential and interactive process, considered as a loop between product recommendations and user feedback
   - Allows for a more thorough exploration of diversity in user preferences