Facets of Fairness and Transparency in Student Learning Analytics

From Accuracy to Actionability and Accountability

prof. dr. Mykola Pechenizkiy

http://www.win.tue.nl/~mpechen/

Test of Time Award for “***Predicting Student Drop Out. A case study***.” by
Gerben Dekker, Mykola Pechenizkiy, Jan Vleeshouwers

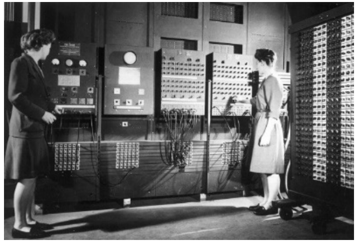
EDM 2019, Montreal, 4 July 2019

ToT award: strong correlations

- IEDMS President -> ToT award (2018 & 2019)
- ToT award (2017) -> IEDMS President

as we know correlation does not imply causation

Outline



Automation of decision making with AI

by humans => by machines

student drop out prediction case study

(Un)Fairness of ML / AI:

AI technology is not neutral

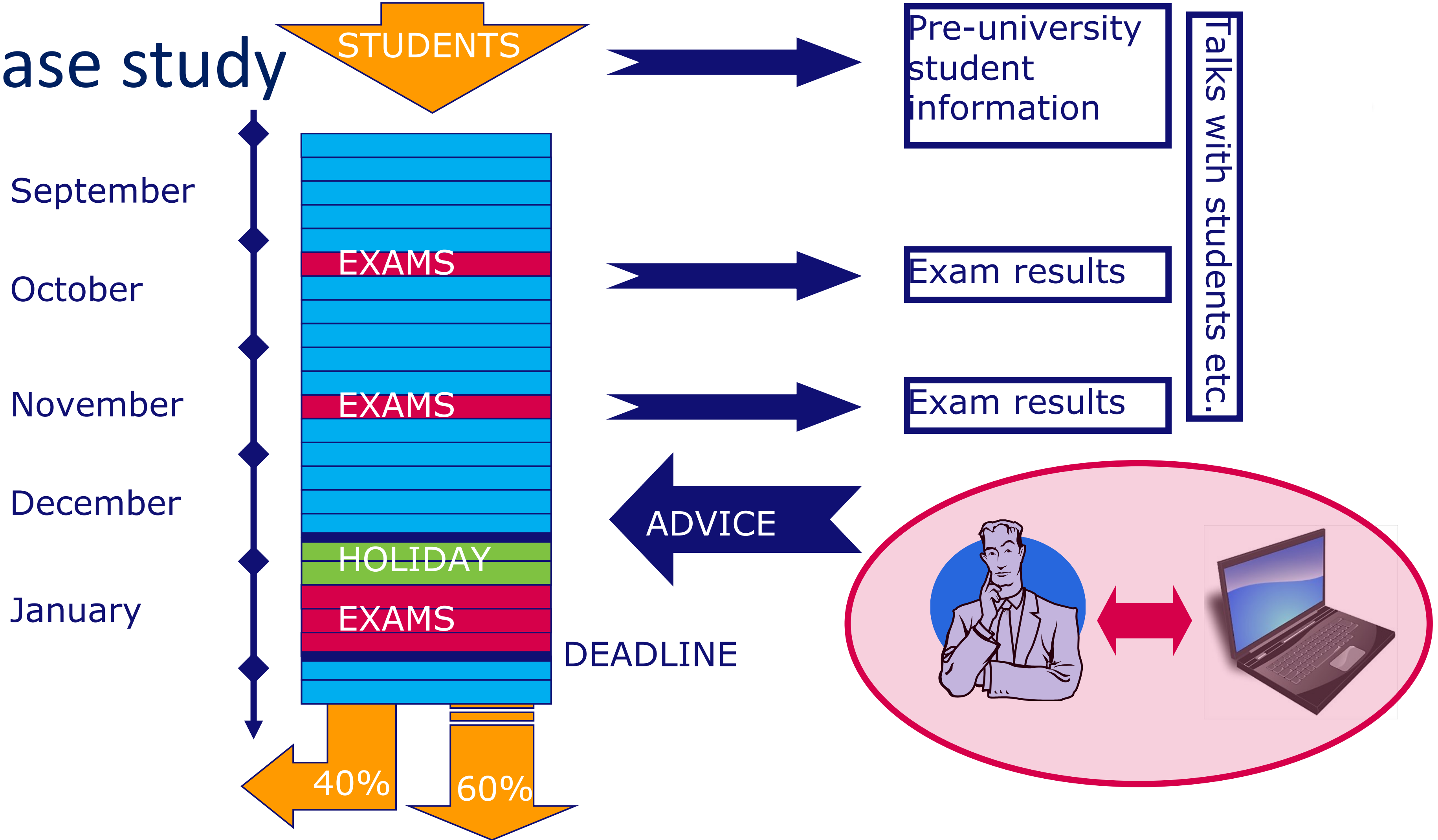
lots of ongoing research to fix it

Transparency of ML / AI:

comprehension, correctness & trust, **utility**

Challenges and outlook

Case study


STUDENTS
September
October
November
December
January
EXAMS
EXAMS
HOLIDAY
EXAMS
DEADLINE
40%
60%
Pre-university student information
Exam results
Exam results
Talks with students etc.
ADVICE

Pre-university data only

One rule classifier on “Science_mean”

- 68% accuracy

No significant improvement using more features with other classification techniques

- cf. “…demographic data (such as race, gender, etc.) and pre-admission data (such as high school academics, entrance exam scores, etc.) - upon which most admissions processes are predicated - are not nearly as useful as early college performance/transcript data for these predictions. “

Mining University Registrar Records to Predict First-Year Undergraduate Attrition, Aulck et al, EDM 2019

All features

One rule classifier

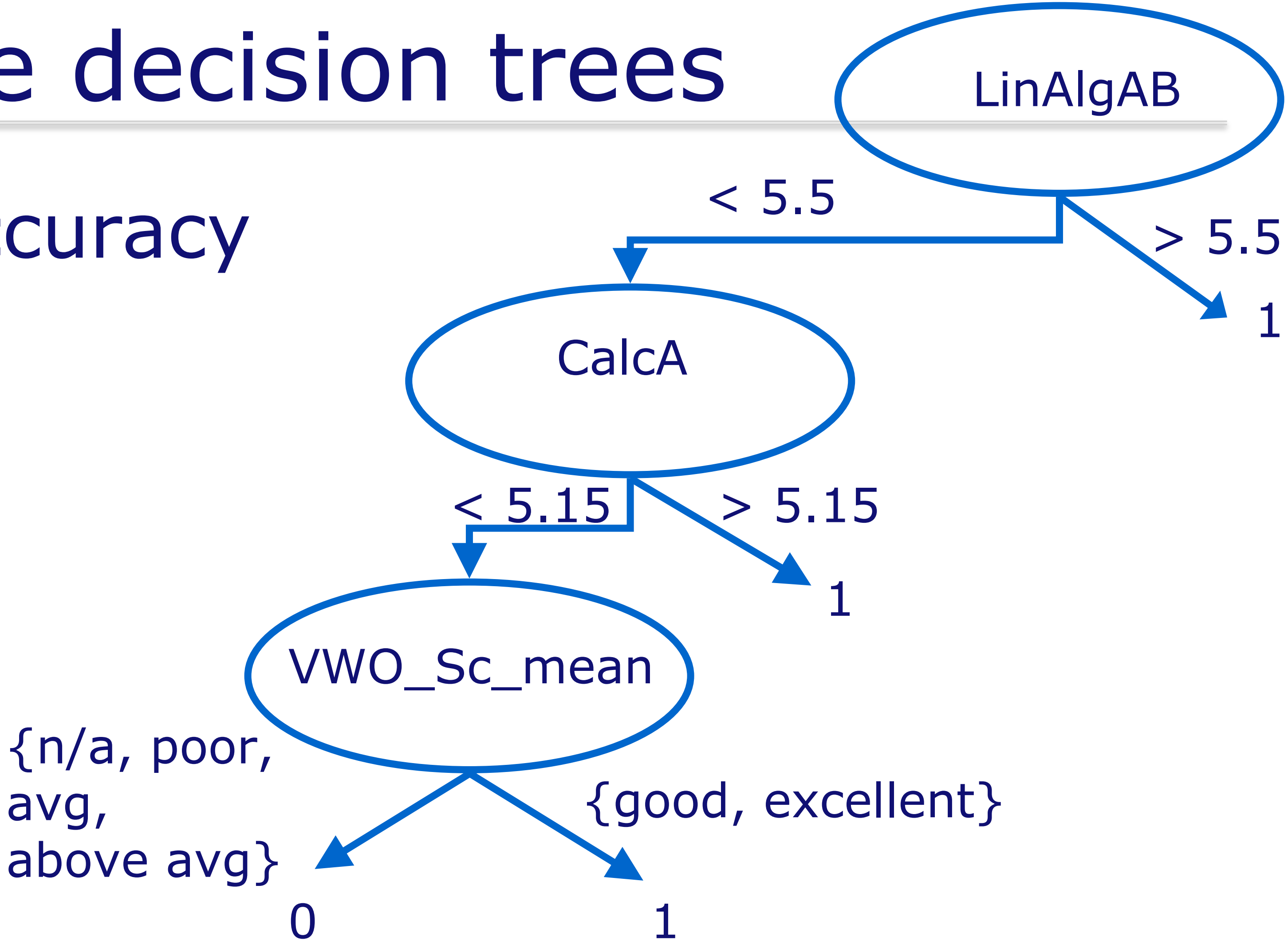
- 75% accuracy using “Linear algebra”

Decision trees and other classifiers

- 80%; 40-50% FPs
- Similarities between models
  - Linear Algebra AB always root node
  - Science Mean always high in tree

Simple decision trees

79% accuracy



Detailed analysis by student counselor

- Review of the problem formulation
  - actionability / utility
- Review of data inconsistenties
  - Semantics of grades/other features across years
- Review the classification measure:
  - How to classify strong students who leave?
- Manual inspection of classification errors
  - 25% of False Negatives were True Negatives

Summary of the highlights

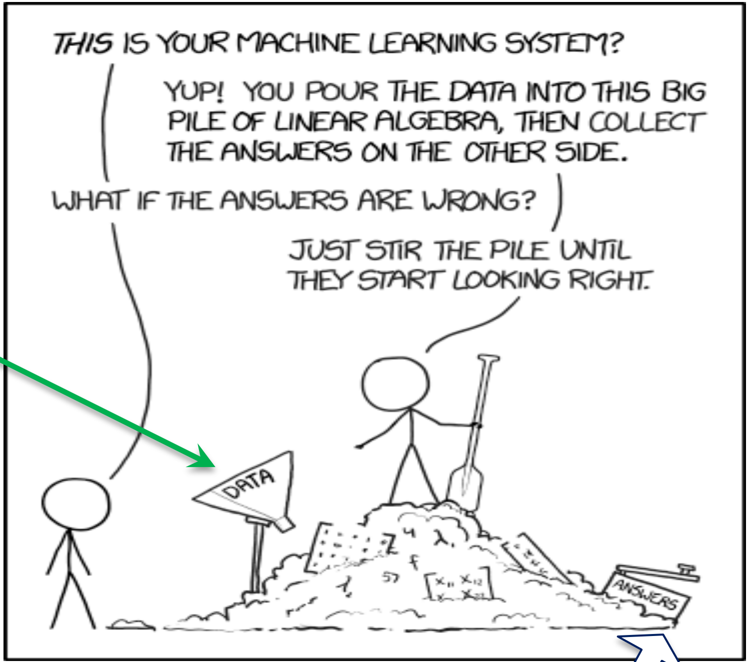
- Went beyond looking at model accuracies
- Detailed analysis by domain expert – student counselor
- Tried to understand the data generation process
- Questions the utility , considered how the model could (not) be used in practice

R&D focused on accuracy and efficiency

- More complex and expressive models
  - ensembles and deep neural networks
- Support for handling 5V's of Big Data
  - more data, data types & operational settings
- More robust models
  - handling anomalies & changes in evolving data

"Anything you can do, *AI* can *do better*"

Predictive analytics as optimization



AI-readable big data matrix

What we try to minimize

Ground truth: known correct answers Y

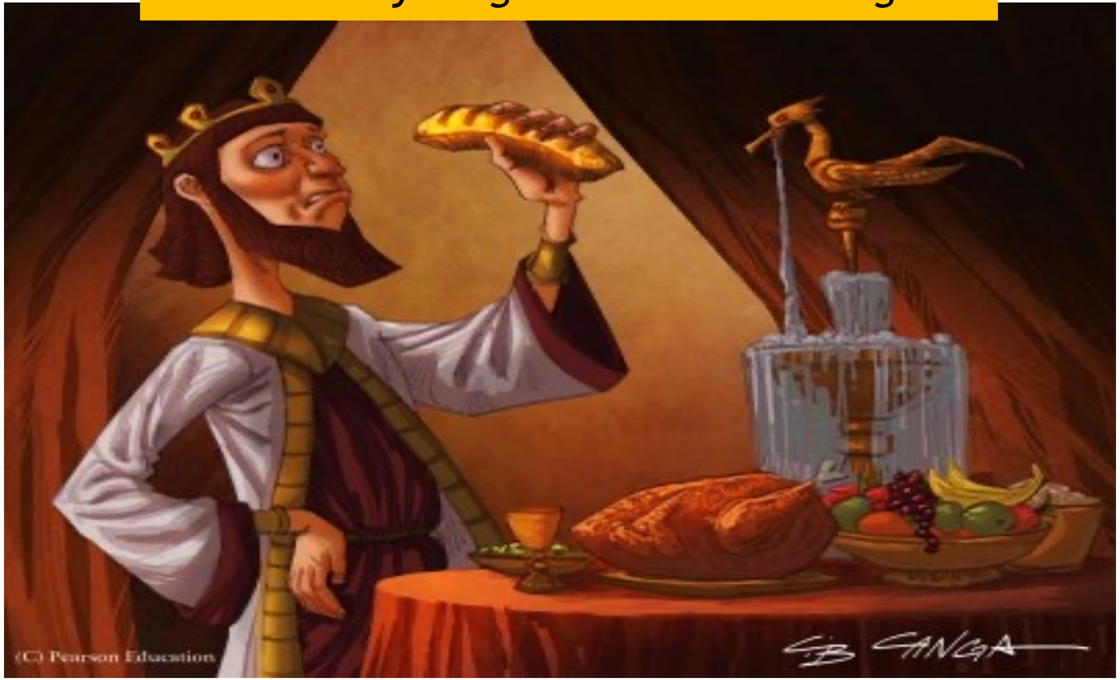
Black-box magic to learn to guess correct answers Y

$$X * \overset{?}{w} = Y$$

$$\text{Error} = Y - X * \overset{?}{w}$$

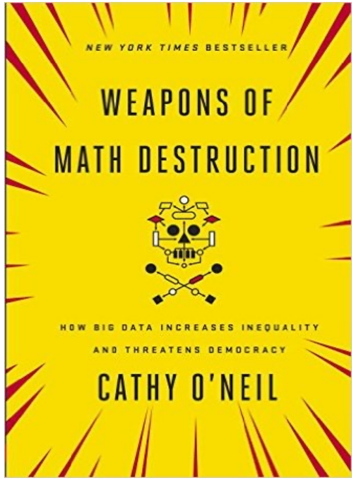
What are we optimizing for?

"I want everything I touch to turn to gold"



Do we really know what we are optimizing for with ML/AI?
Side effects?

Dangers of blind optimizing for KPIs



- Education ecosystem
- Academic/research ecosystem
- Police and justice
- ….

Things can go wrong despite of good intentions behind the set KPIs

Reflection: Predictive analytics that works!?

"Anything you can do, *AI* can *do better*"

"All models are *wrong*, but some are *useful*"

If not 100% accurate then there are trade-offs:

- Well formulated and well studied:
  - precision-recall; bias-variance; robustness-sensitivity;
- (not so) well formulated, and not so well studied:
  - accuracy-fairness, acc.-privacy, acc.-transparency, …

Model comprehension is needed / required

Auditing model performance for biases in prediction-based decisions

Detecting, measuring and preventing unfair / discriminating decision making or profiling

Non-uniform accuracy

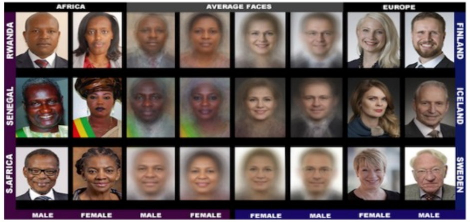
$\mathrm{Error}_{males} \ll \mathrm{Error}_{females}$

Favoritism in making decisions:

$P(+ \mid male) - P(+ \mid female)$

#GenderShades: Facial Recognition Is Accurate

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%



Pilot Parliaments Benchmark

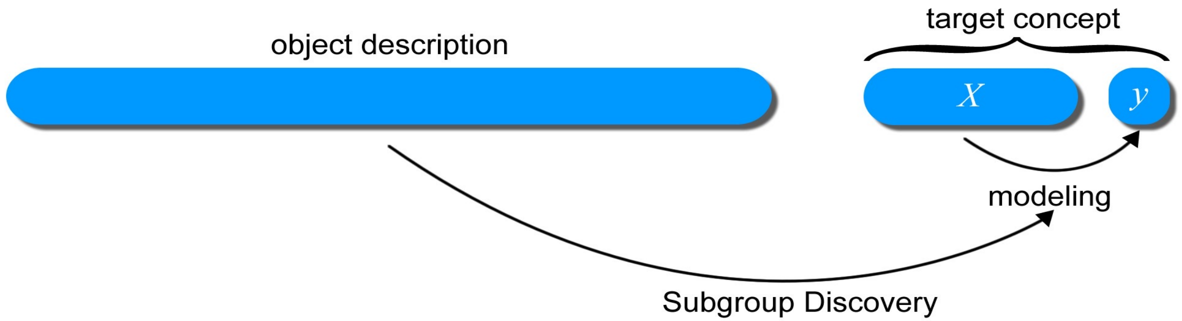
… if You're a White Guy

- 8.1% – 20.6% worse performance on female faces
- 11.8% – 19.2% worse performance on darker faces
- 20.8% – 34.7% worse performance on darker female faces

#GenderShades; http://gendershades.org/

How about #GenderShades automation?

Find subgroups on which a classifier performs exceptionally well or exceptionally poor

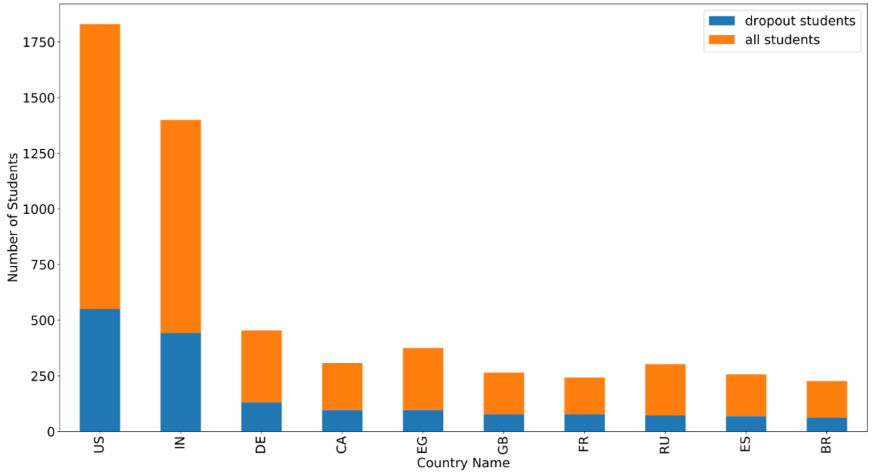


gender = male => acc ↑ = 20.6%
skin = dark ^ gender = female
=> acc ↓ = 34.7%

Exceptional model mining (EMM) approach for finding subgroups for which soft classifier outputs align exceptionally well or bad wrt ground truth

W. Duivesteijn, J. Thaele: Understanding Where Your Classifier Does (Not) Work - the SCaPE Model Class for EMM, ICDM 2014

EMM on dropout prediction



Distribution of dropouts per country

Exceptional subgroups wrt prediction performance

ELBA: Exceptional Learning Behavior Analysis. Du et al., EDM 2018

	φ_{f1}	$\lvert G_D \rvert$
Country = OM, Profile language = en-US, Browser language != en-US, Educational status != BACHELOR DEGREE	0.5051	32
Country = OM, Profile language != en-US	0.4058	22
Region = MA, Gender = female, Educational status=COLLEGE NO DEGREE	0.3489	24
Country = OM, Met Payment Condition != True	0.3464	31
Join Date <= 390, Region != MA	0.3193	28

Auditing model performance for biases in prediction-based decisions

Detecting, measuring and preventing unfair / discriminating decision making or profiling

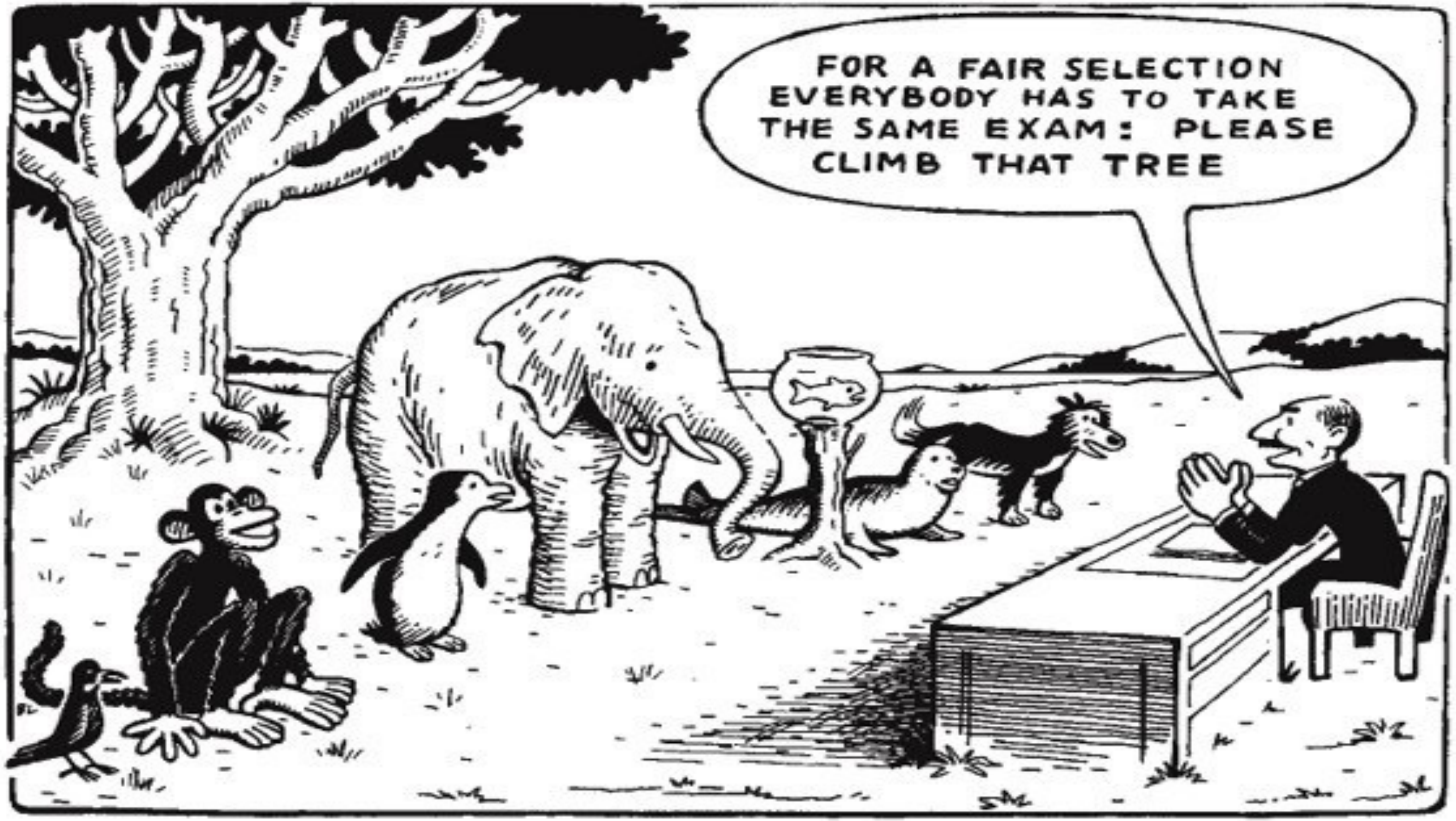
Non-uniform accuracy

$\text{Error}_{\text{males}} \ll \text{Error}_{\text{females}}$

Favoritism in making decisions:

$P(+ \mid \text{male}) - P(+ \mid \text{female})$

Different notions of quality and fairness

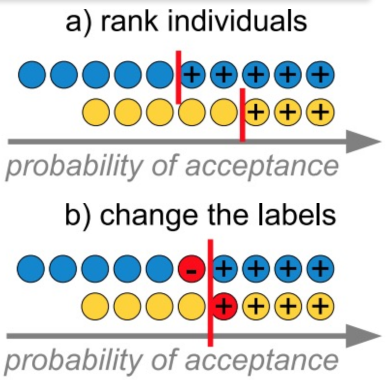


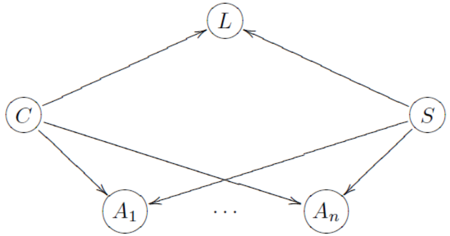
Facets of algorithmic fairness

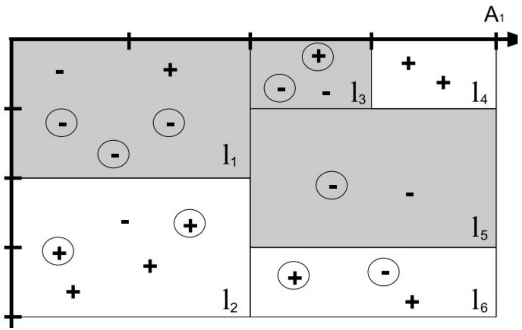
- Defining and measuring fairness
  - Achieving parity or satisfying preferences?
  - Focus on fair *treatment* or on fair *impact*?
  - Individual or group level
  - 20+ measures of fairness;
- Discovering and preventing unfairness (by design)
  - Theory, methods, experiments
  - Lots of new data mining techniques for discrimination-aware classification, regression, recommendation, …

Early fairness-aware solutions

- ~~Remove sensitive attributes?~~
- Preprocessing – "data massaging"
  - Modify input data (labels)
  - Resample input data
- Constraint learning
  - Algorithm-specific, e.g. Bayesian, SVMs
- Postprocessing
  - Modify models and/or their outputs







Kamiran, F., Calders, T. & Pechenizkiy, M. (2013)
"Techniques for Discrimination-Free Predictive Models", In Discrimination and Privacy in the Information Society

Current Fairness-aware research

Spreading beyond classification: regression, ranking, cake-cutting, PCA

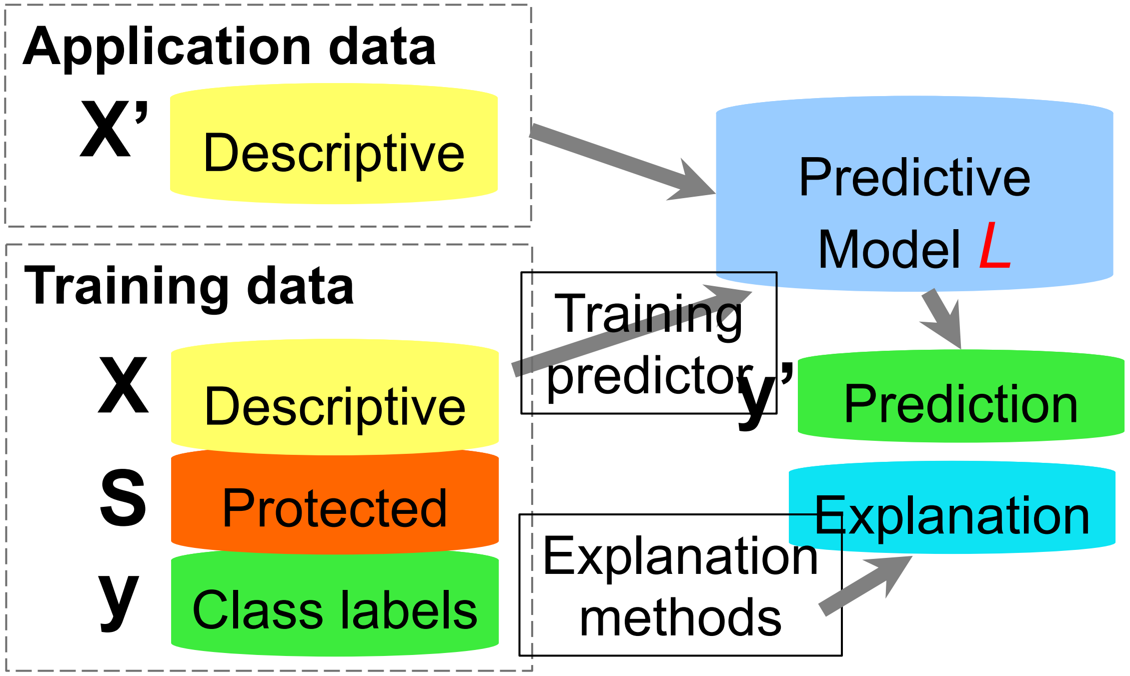
More attention to counterfactual reasoning

Connections to social sciences, law, mathematical finance

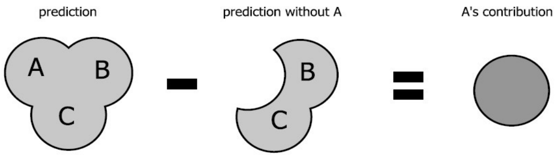
Picked in EDM-related research:

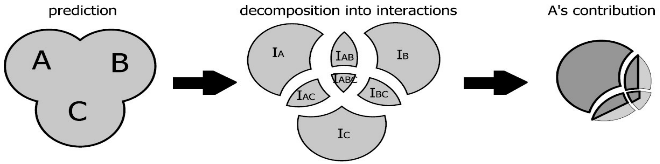
- *A History of Quantitative Fairness in Testing, Hutchinson, FAT 2019*
- *Evaluating Fairness and Generalizability in Models: Predicting On-Time Graduation from College Applications Hutt et al., EDM 2019*
- *Evaluating the Fairness of Predictive Student Models Through Slicing Analysis , Gardner et al., LAK 2019*

Automation of explanations



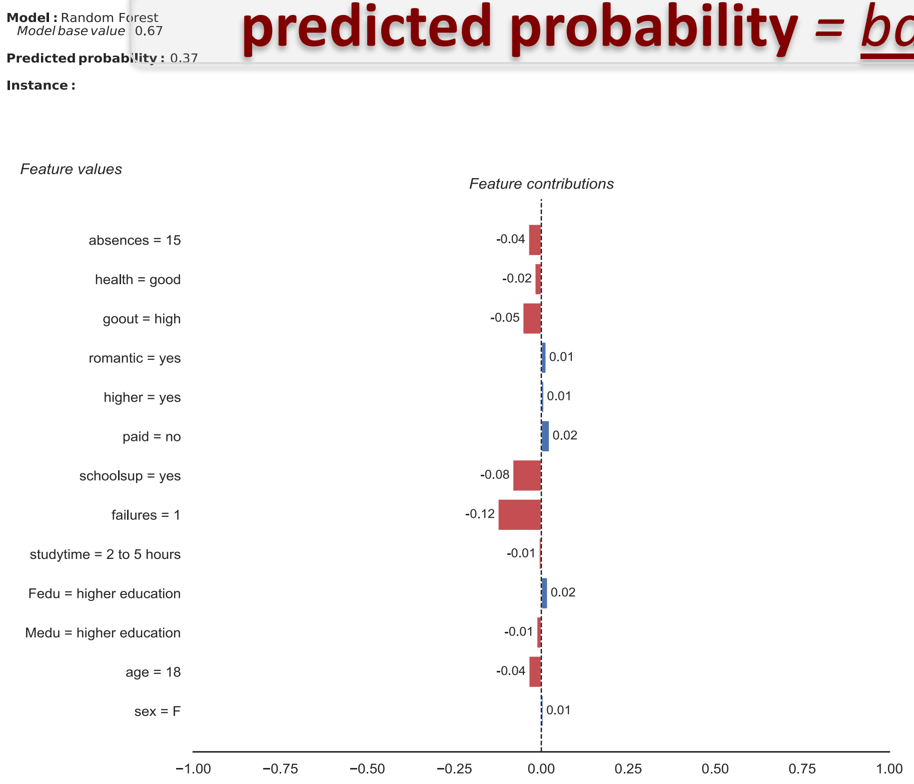
Local (model-agnostic) attribution-based explanations: LIME, SHAP, ….





Shapley Additive Explanations (SHAP)

predicted probability = *base value* + *sum(contributions)*



But is SHAP useful for decision making, e.g. alert processing?

Weerts, H.J.P., van Ipenburg, W. & Pechenizkiy, M. (2019)
A Human-Grounded Evaluation of SHAP for Alert Processing,
In Explainable AI @ KDD 2019, abs/1907.03324

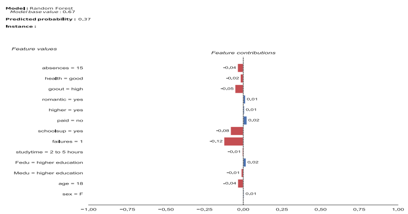
The Student Performance Dataset

- The dataset contains information on **student performance in mathematics** from two Portuguese high schools.
- The classification task is to determine whether a student will pass mathematics or not:
  - Positive class: passed mathematics
  - Negative class: failed mathematics

Weerts, H.J.P., van Ipenburg, W. & Pechenizkiy, M. (2019)
A Human-Grounded Evaluation of SHAP for Alert Processing,
In Explainable AI @ KDD 2019, abs/1907.03324

User Experiment: utility of SHAP values

- **Real humans** perform simplified **alert processing tasks**
- 2 experiments, 3 sessions, **159 participants** in total



1 Quantitative Analysis

Statistical hypothesis testing of **utility metrics**

Result

Inconclusive: no significant difference in task utility

2 Qualitative Analysis

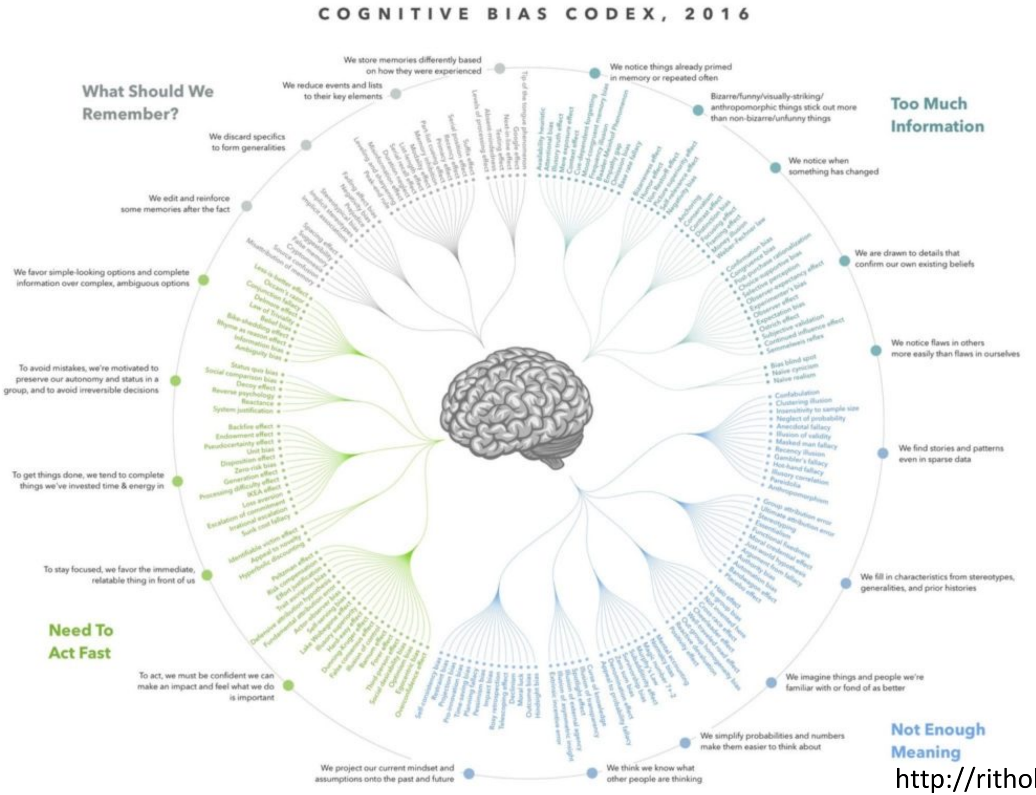
Analyze participants' written **reflections** and **reasoning**

Results

- Large **SHAP values impact decision-making** process
- Model's **confidence score** is one of the leading sources of evidence

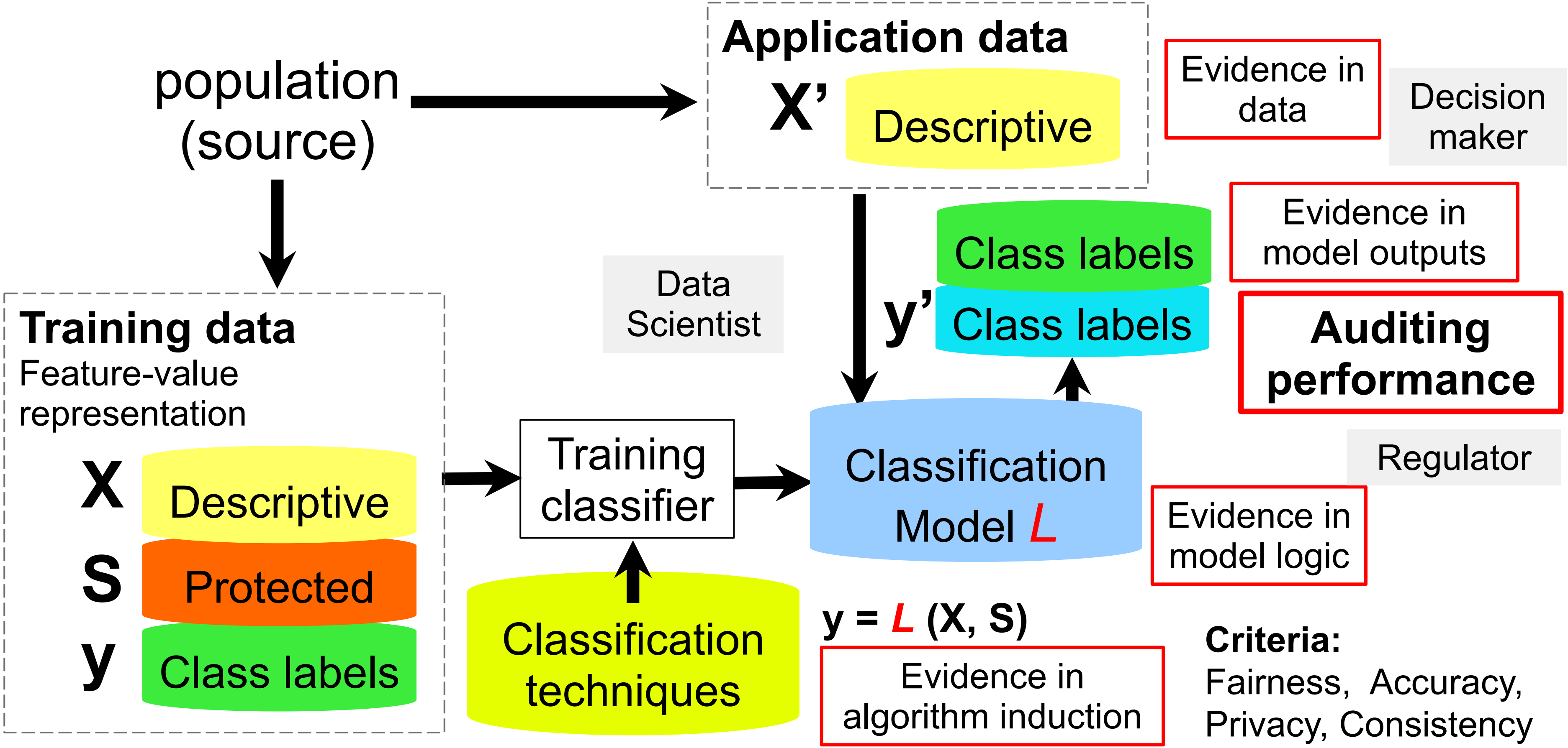
Weerts, H.J.P., van Ipenburg, W. & Pechenizkiy, M. (2019)
A Human-Grounded Evaluation of SHAP for Alert Processing,
In Explainable AI @ KDD 2019, abs/1907.03324

Wrong explanations vs. wrong interpretation of correct explanations



http://ritholtz.com/2016/09/cognitive-bias-codex/

Auditing Algorithmic Decision Making



Challenges and Outlook

- Better understanding of the real-world problems we try to address
  - Computer scientists: reductionist approach to optimization
  - Educators and policy-makers: but ignore operationalization
- Better understanding of the *trade-offs,* e.g. *personalization-discrimination*
- Better tooling for ML model debugging, profiling, certification, and data-driven decision making: *trust*, *transparency*, *reliability*
- Educating data scientists, the general public, regulators, and policy-makers

Take home food for thought

- Can we bridge the predictive vs. causal gaps?
  - Why does this model give this answer?
- Can we achieve ML fairness without ML transparency?
  - Or is fairness just another KPI as accuracy?
- Can we certify ML models without looking into data they were trained on?