Explainability in Machine Learning

Using Counterfactuals distributions

Jean-Michel Loubes & Laurent Risser

loubes@math.univ-toulouse.fr laurent.risser@math.univ-toulouse.fr

GEMS-AI
Globally Explaining
Models under Stress

1 / Explaining Machine Learning decisions

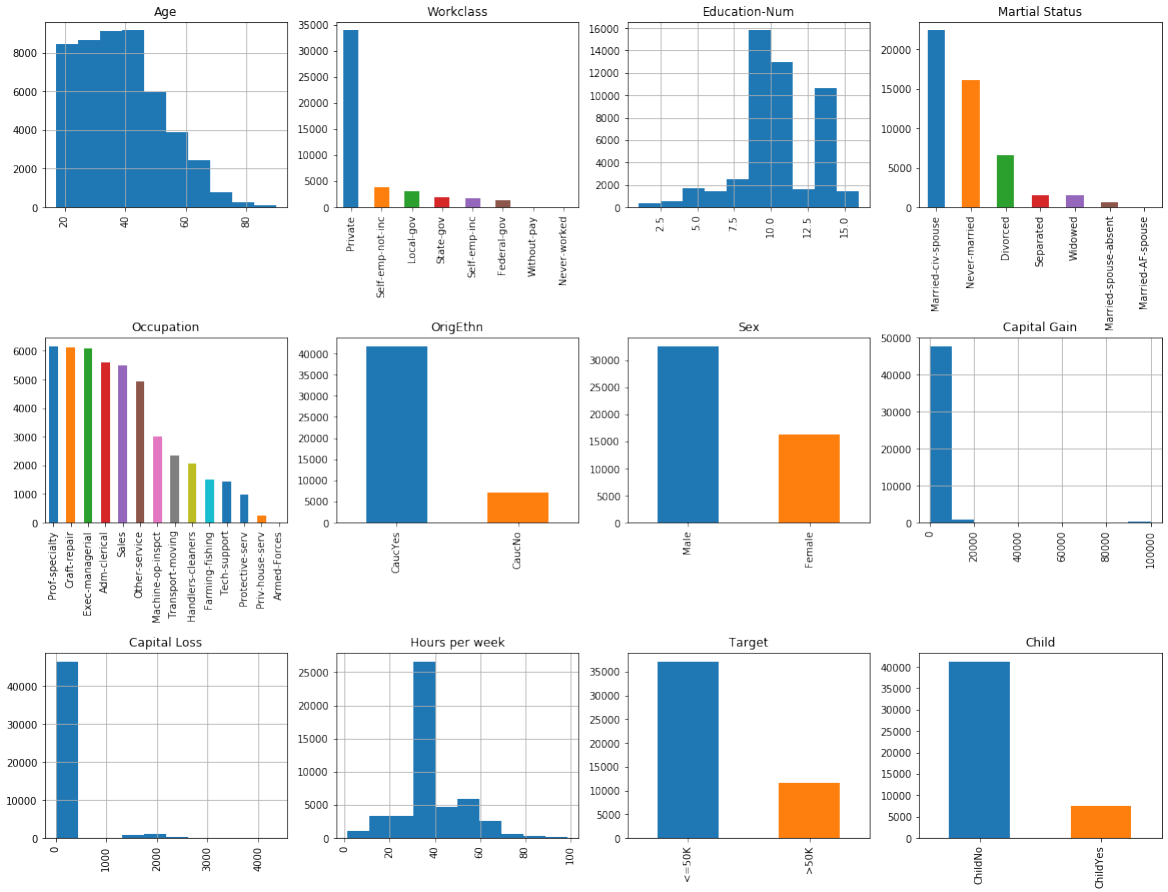
2 / GEMS : Distributional what-if Machine

3 / Applications

4 / Going further… Counterfactuals using OT

1 EXPLAINING MACHINE LEARNING DECISIONS

Typical use case



$n = 48842$ observations (individuals) described by $p = 14$ variables

Machine Learning Algorithms

- Logistic Regression (*scikit-learn*)
- Decision Trees (*scikit-learn*)
- Extreme Gradient Boosting (*lightgbm*)

Accuracy : 89%

$\widehat{Y} = 1$

Age : 41
Educ Number : 6
Hours per Week: 43

$\widehat{Y} = 0$

Age : 41
Educ Number : 8
Hours per Week: 41

EXPLAINING MACHINE LEARNING DECISIONS

Paper and code: Explaining under stress

Explaining machine learning models using entropic variable projection

François Bachoc[1], Fabrice Gamboa[1,3], Max Halford[2], Jean-Michel Loubes[1,3] and Laurent Risser[1,3]

[1]Institut de Mathématiques de Toulouse
[2] Institut de recherche en informatique de Toulouse
[3] Artificial and Natural Intelligence Toulouse Institute (3IA ANITI)

https://arxiv.org/pdf/1810.07924.pdf
https://gems-ai.aniti.fr/
https://github.com/XAI-ANITI/ethik

1 EXPLAINING MACHINE LEARNING DECISIONS

Paper and code: Explaining under stress

Explaining machine learning models using entropic variable projection

François Bachoc[1], Fabrice Gamboa[1,3], Max Halford[2], Jean-Michel Loubes[1,3] and Laurent Risser[1,3]

[1]Institut de Mathématiques de Toulouse
[2] Institut de recherche en informatique de Toulouse
[3] Artificial and Natural Intelligence Toulouse Institute (3IA ANITI)

https://arxiv.org/pdf/1810.07924.pdf
https://gems-ai.aniti.fr/
https://github.com/XAI-ANITI/ethik

Intuition : Generating stress in the test to study the answer of a model to a specific property

From banking system to medical system

2

GEMS: DISTRIBUTIONAL WHAT-IF MACHINE

Generating distributional counterfactual to explain the M.L. decisions

« What-if machine » for group-explainability

Test set

$$\{X_i, Y_i\}_{i=n+1,\dots,n+m}$$

$$X_i = \{X_i^1, \dots, X_i^p\}$$

$$(X, Y) \sim \mathbb{P}$$

« Black-box » decision rules

X_i^1 →
X_i^2 →
X_i^3 →
… …
X_i^p →

→ $\hat{Y}_i := f_\theta(X_i)$

Objectives :

- Stress input $\{X_i, Y_i, \hat{Y}_i\}_{i=n+1,\dots,n+m}$ to emphasise a specific property while remaining in the **domain of validity**
- Create alternative inputs which are plausible
- Then explain how the algorithm varies or its resiliency

2

GEMS: DISTRIBUTIONAL WHAT-IF MACHINE

Generating distributional counterfactual to explain the M.L. decisions

→ Theoretical guarantees offered (Published 2022)

Algorithm :

1: Computing the distributions of the original data P_n

2: Choosing a stress level $\int_{\mathbb{R}^{p+2}} \Phi(z) dQ(z) = t$ with a stress function Φ

3: Finding the closest distributions close to the observations $Q_t := \text{arginf}_{Q \in \mathbb{P}_{\Phi,t}} \text{KL}(Q, P_n)$

$$Q_t = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^{(t)} \delta_{X_i, \hat{Y}_i, Y_i},$$

A **feasible** and **scalable** algorithm to compute new faithful distributions that satisfies a chosen constraint

APPLICATIONS → TABULAR DATA

Application to binary classification based on tabular data → Adult income

Re-weighting the observations $\{X_i, Y_i\}_{i=1,\dots,n}$ to transform a specific property of the test set in average, e.g.:

- Increase mean of each variable
- Increase variance
- Modify correlations in the dataset (or certain conditions)
- Increase the error of the model …… can be designed on purpose

Adult income dataset (https://www.kaggle.com/uciml/adult-census-income)

Age (X^1)	**Education.num** (X^2)	**Marital.status** (X^3)	**Hours.per.week** (X^4)	…	**Loan granted — True** (Y)	**Loan granted — Predicted** ($\hat{Y} = f_\theta(X)$)
54	4	Divorced	40		No	No
41	10	Never-married	60		Yes	Yes
51	13	Married-civ	40		Yes	No
39	14	Married-civ	65		Yes	Yes
49	10	Divorced	50		No	Yes
…	…	…	…		…	…

APPLICATIONS → TABULAR DATA

Stress on the mean

Average age is 38 and *P*(loan granted) = 0.212 in the test set

Age (X^1)	Education.num (X^2)	Marital.status (X^3)	Hours.per.week (X^4)	…	Loan granted — True (Y)	Loan granted — Predicted ($\hat{Y} = f_\theta(X)$)
54	4	Divorced	40		No	No
41	10	Never-married	60		Yes	Yes
51	13	Married-civ	40		Yes	No
39	14	Married-civ	65		Yes	Yes
49	10	Divorced	50		No	Yes
…	…	…	…		…	…

3 APPLICATIONS → TABULAR DATA

Stress on the mean

What-if the average age is 50 instead of 38 in the test set?

Compute optimal weights →

	Age (X^1)	Education.num (X^2)	Marital.status (X^3)	Hours.per.week (X^4)	⋯	Loan granted — True (Y)	Loan granted — Predicted ($\hat{Y} = f_\theta(X)$)
1.05	54	4	Divorced	40		No	No
0.83	41	10	Never-married	60		Yes	Yes
1.15	51	13	Married-civ	40		Yes	No
0.81	39	14	Married-civ	65		Yes	Yes
1.15	49	10	Divorced	50		No	Yes
...	...	...	...	...		...	...

3

APPLICATIONS → TABULAR DATA

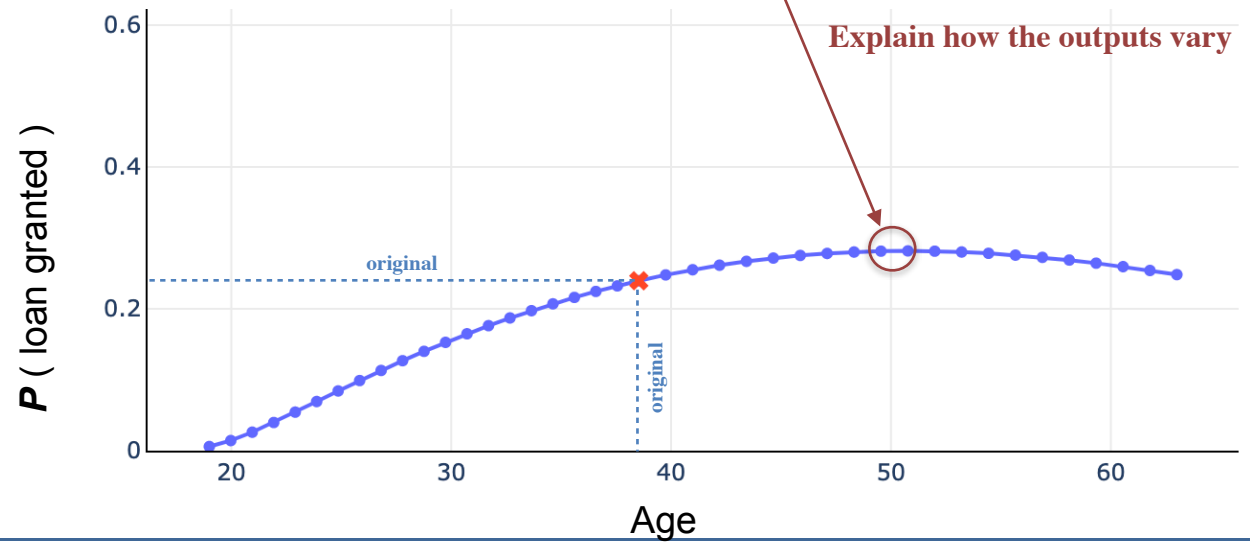
Stress on the mean

What-if the average age is 50 instead of 38 in the test set?

Compute optimal weights →

	Age (X^1)	Education.num (X^2)	Marital.status (X^3)	Hours.per.week (X^4)	…	Loan granted — True (Y)	Loan granted — Predicted ($\hat{Y} = f_\theta(X)$)
1.05	54	4	Divorced	40		No	No
0.83	41	10	Never-married	60		Yes	Yes
1.15	51	13	Married-civ	40		Yes	No
0.81	39	14	Married-civ	65		Yes	Yes
1.15	49	10	Divorced	50		No	Yes
…	…	…	…	…		…	…

Explain how the outputs vary



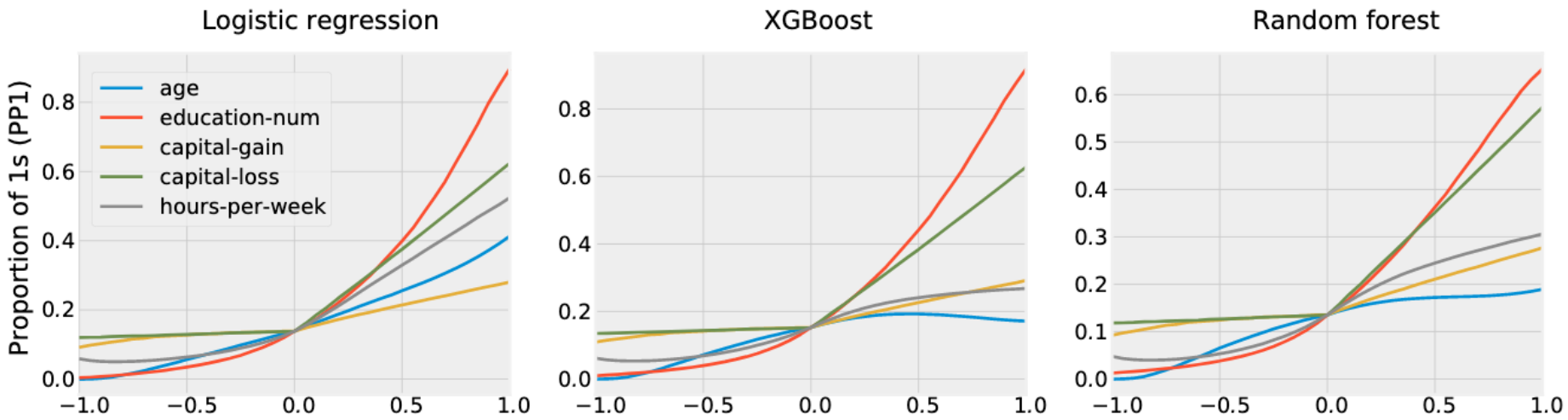
3 APPLICATIONS → TABULAR DATA

Stress on each input variable → get most influent variables (towards causality)

What-if the average […] is […] instead of [original average value] in the test set?

Compute optimal weights → … then explain

	Age (X^1)	Education.num (X^2)	Marital.status (X^3)	Hours.per.week (X^4)	…	Loan granted — True (Y)	Loan granted — Predicted ($\hat{Y} = f_\theta(X)$)
…	54	4	Divorced	40		No	No
…	41	10	Never-married	60		Yes	Yes
…	51	13	Married-civ	40		Yes	No
…	39	14	Married-civ	65		Yes	Yes
…	49	10	Divorced	50		No	Yes
…	…	…	…	…		…	…



3 APPLICATIONS → TABULAR DATA

Flexibility for more complex explanations

- Wide capability for handling different functions on
  - The data it-self : controlling the mean behaviour of input variables $\Phi(X^j)$
  - The correlations inside the data $\Phi(X^i, X^j)$
  - The performance of the algorithm $\Phi(X) = \Phi(Y, \hat{Y})$

Examples :

- stress on covariance : $\Phi(X^1, \ldots, X^p, \hat{Y}, Y) = (X^{j_0}, (X^{j_0})^2)$ and $t = (m_{j_0}, m_{j_0}^2 + v)$

-stress on covariance preserving mean : $\Phi(X^1, \ldots, X^p, \hat{Y}, Y) = (X^{j_1}, X^{j_2}, X^{j_1}X^{j_2})$ and $t = (m_{j_1}, m_{j_2}, m_{j_1}m_{j_2} + c)$

3

APPLICATIONS → TABULAR DATA

Explaining the resiliency or robustness of the network

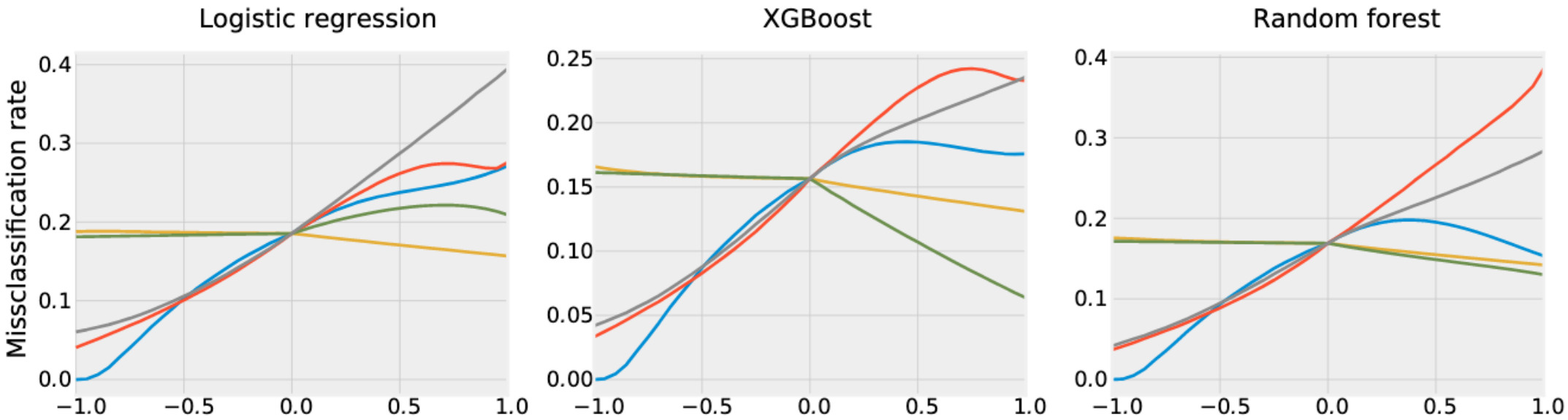
What-if the average […] is […] instead of [original average value] in the test set? → Error rate

Compute optimal weights

→

… then explain

	Age (X^1)	Education.num (X^2)	Marital.status (X^3)	Hours.per.week (X^4)	…	Loan granted — True (Y)	Loan granted — Predicted ($\hat{Y} = f_\theta(X)$)
…	54	4	Divorced	40		No	No
…	41	10	Never-married	60		Yes	Yes
…	51	13	Married-civ	40		Yes	No
…	39	14	Married-civ	65		Yes	Yes
…	49	10	Divorced	50		No	Yes
…	…	…	…	…		…	…



3 APPLICATIONS → TABULAR DATA

What about the code?

https://xai-aniti.github.io/ethik/tutorials/binary-classification

```
import pandas as pd
  ⋮
X = pd.read_csv(url, names=names, header=None, dtype=dtypes)
X['gender'] = X['gender'].str.strip().astype('category')  # Remove leading whitespace
y = X.pop('salary').map({' <=50K': False, ' >50K': True})

X.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	gender	capital-gain	capital-loss	hours-per-week	native-country
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States

3 APPLICATIONS → TABULAR DATA

What about the code?

https://xai-aniti.github.io/ethik/tutorials/binary-classification

```
from sklearn import model_selection

X_train, X_test, y_train, y_test = model_selection.train_test_split(X, y, shuffle=True,
random_state=42)
```

```
import lightgbm as lgb

model = lgb.LGBMClassifier(random_state=42).fit(X_train, y_train)
```

```
y_pred = model.predict_proba(X_test)[:, 1]

# We use a named pandas series to make plot labels more explicit
y_pred = pd.Series(y_pred, name='>$50k')
```

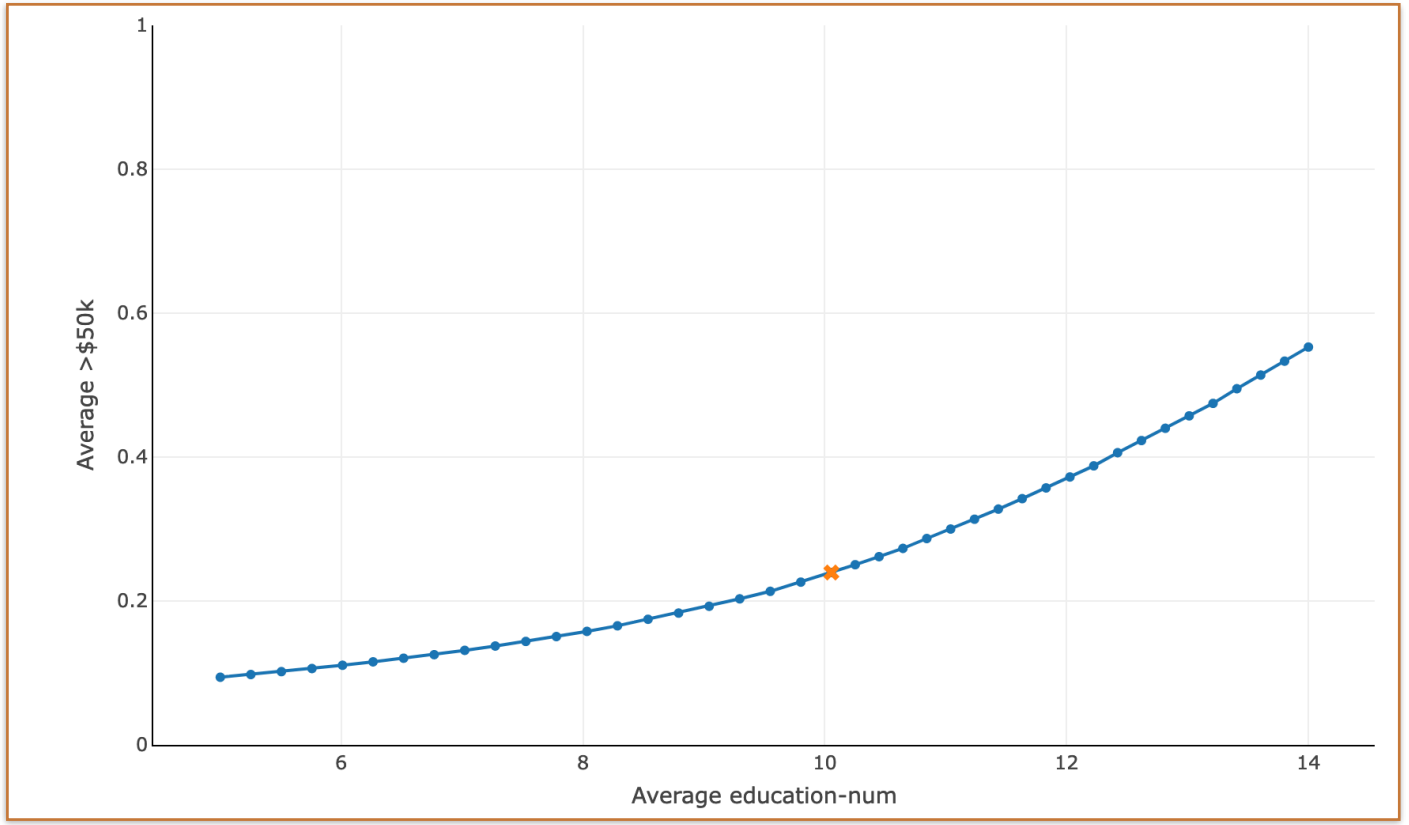
```
import ethik

explainer = ethik.ClassificationExplainer()
```

APPLICATIONS → TABULAR DATA

What about the code?

https://xai-aniti.github.io/ethik/tutorials/binary-classification

```
explainer.plot_influence(
    X_test=X_test['education-num'],
    y_pred=y_pred
)
```

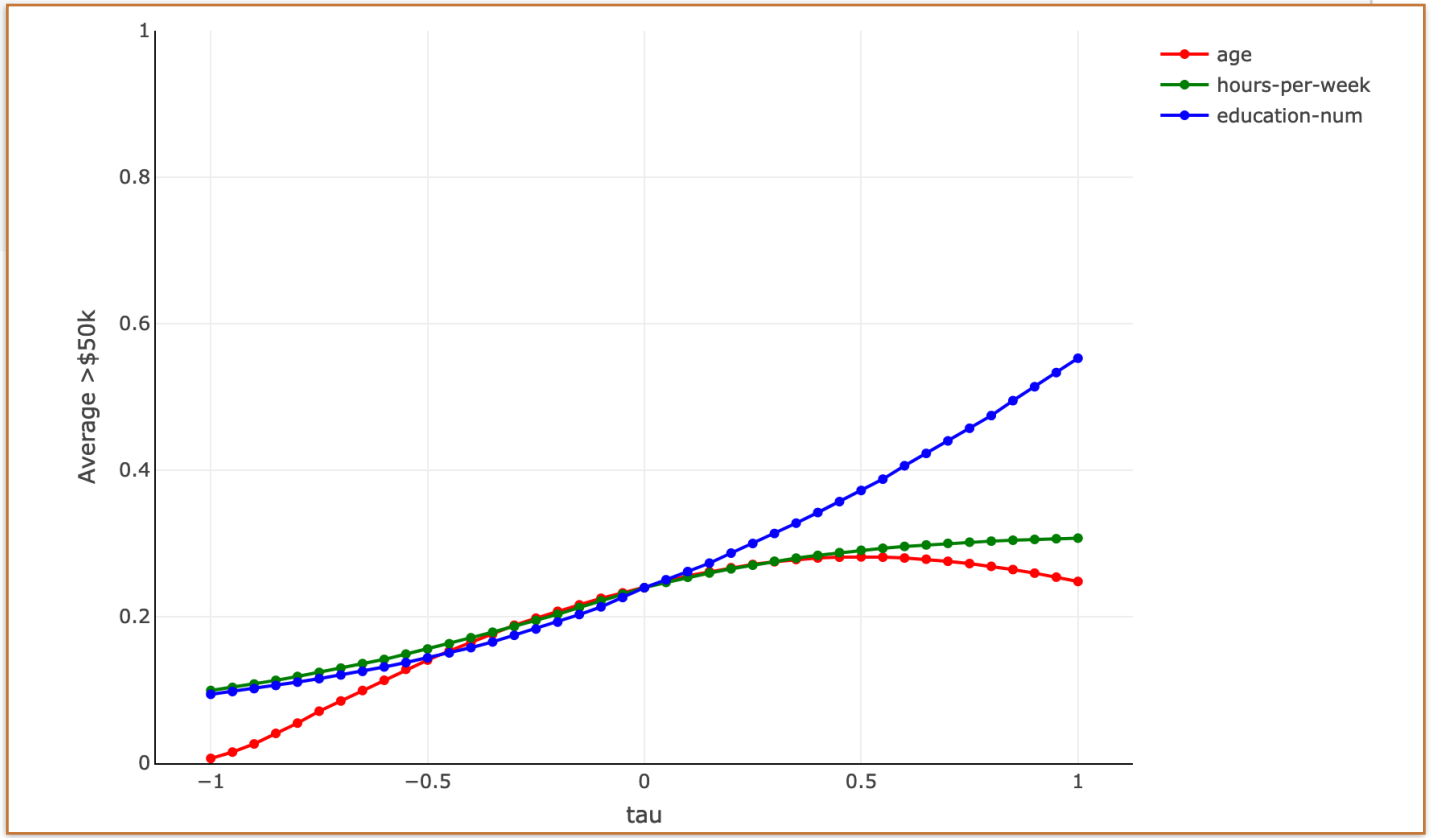


APPLICATIONS → TABULAR DATA

What about the code?

https://xai-aniti.github.io/ethik/tutorials/binary-classification

```
explainer.plot_influence(
    X_test=X_test[['age', 'hours-per-week', 'education-num']],
    y_pred=y_pred,
    colors={
        'age': 'red',
        'hours-per-week': 'green',
        'education-num': 'blue'
    }
)
```

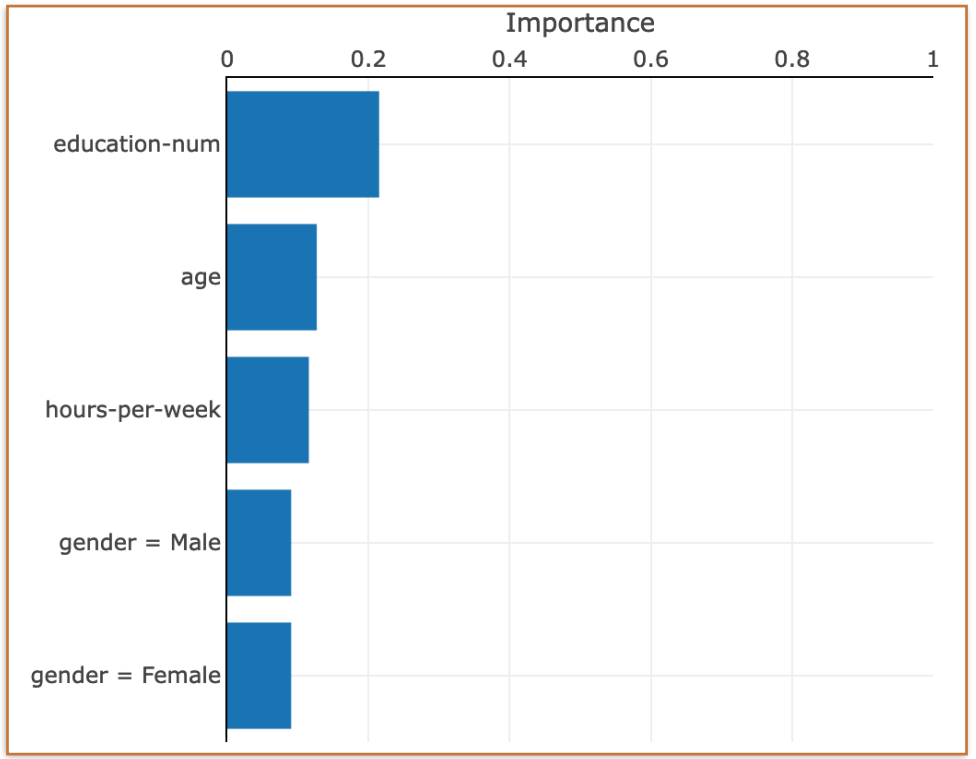


APPLICATIONS → TABULAR DATA

What about the code?

https://xai-aniti.github.io/ethik/tutorials/binary-classification

```
explainer.plot_influence_ranking(
    X_test=X_test[['age', 'education-num', 'hours-per-week', 'gender']],
    y_pred=y_pred,
)
```

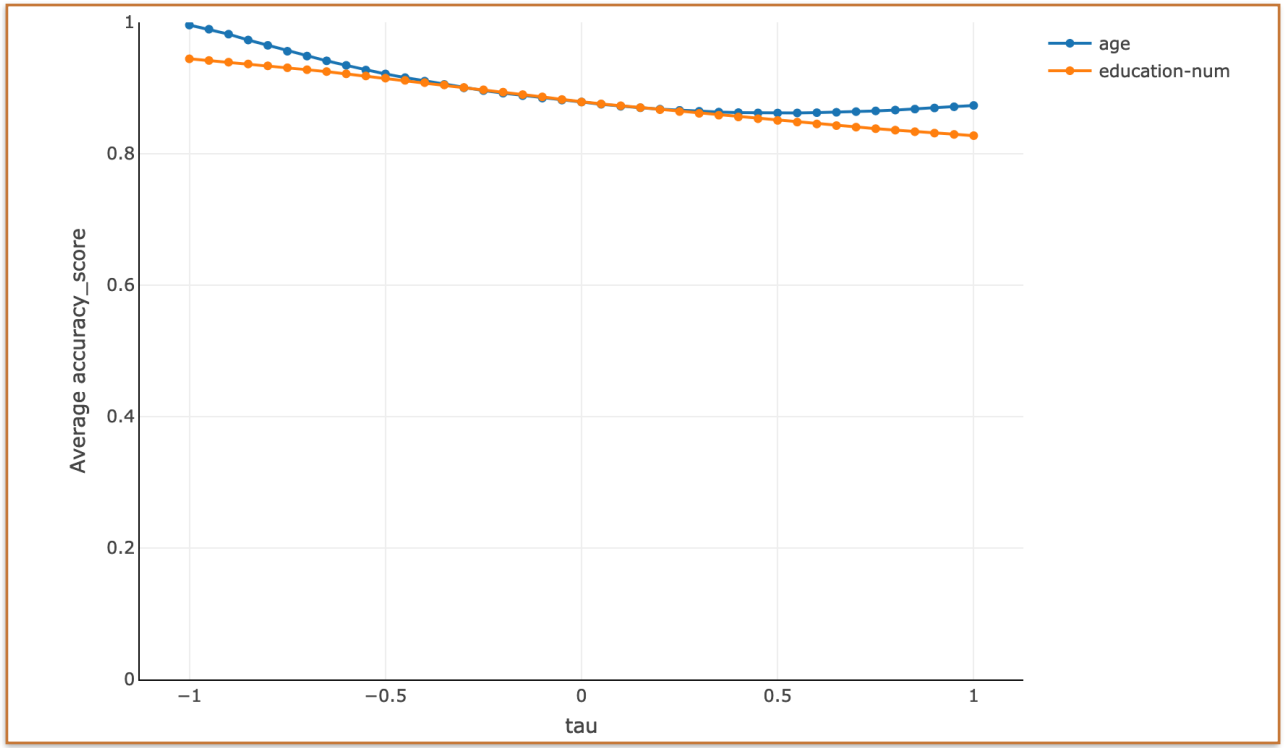


APPLICATIONS → TABULAR DATA

What about the code?

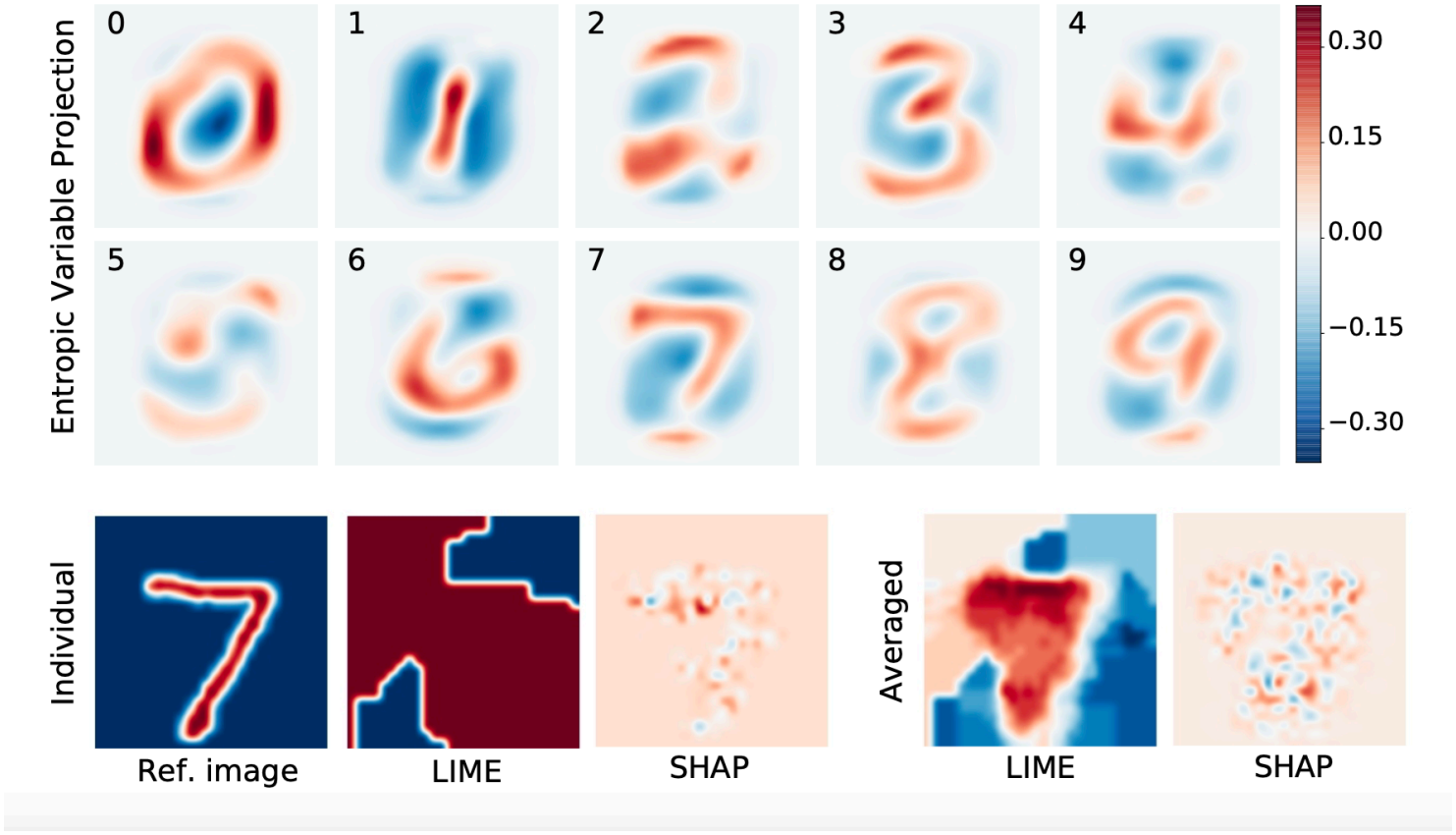
https://xai-aniti.github.io/ethik/tutorials/binary-classification

```
explainer.plot_performance(
    X_test=X_test[['age', 'education-num']],
    y_test=y_test,
    y_pred=y_pred > 0.5,
    metric=metrics.accuracy_score,
)
```



3 APPLICATIONS → IMAGES

Images out of the Mnist dataset



APPLICATIONS → IMAGES

What about the code?

https://xai-aniti.github.io/ethik/tutorials/image

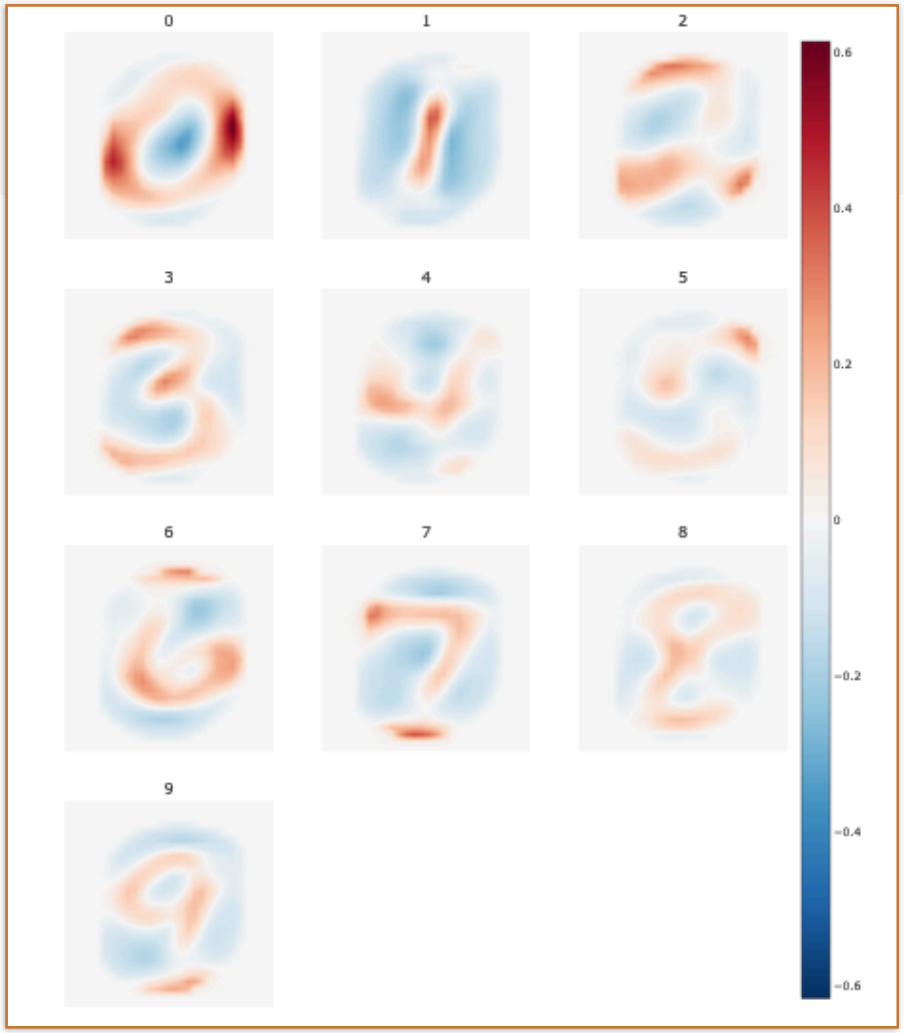
```
import keras
        ⋮
    (x_train, y_train), (x_test, y_test) = mnist.load_data()
        ⋮
model = Sequential()
model.add(Conv2D(32, kernel_size=(3, 3),
                 activation='relu',
                 input_shape=input_shape))
model.add(Conv2D(64, (3, 3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Dropout(0.25))
model.add(Flatten())
model.add(Dense(128, activation='relu'))
model.add(Dropout(0.5))
model.add(Dense(num_classes, activation='softmax'))
        ⋮
```

3 APPLICATIONS → IMAGES

What about the code?

https://xai-aniti.github.io/ethik/tutorials/image

```
import ethik

explainer = ethik.ImageClassificationExplainer()
explainer.plot_influence(x_test, y_pred)
```

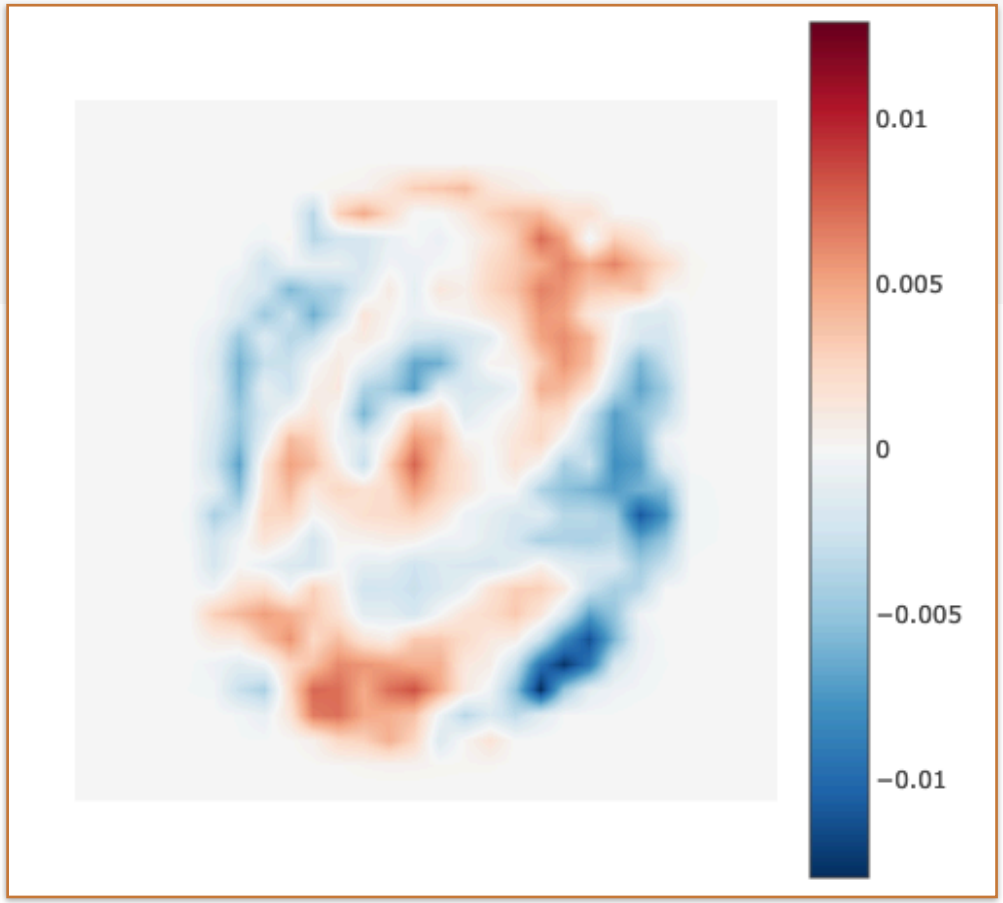


3 APPLICATIONS → IMAGES

What about the code?

https://xai-aniti.github.io/ethik/tutorials/image

```
from sklearn import metrics

explainer.plot_performance(
    X_test=x_test,
    y_test=y_test.argmax(axis=1),
    y_pred=y_pred.argmax(axis=1),
    metric=metrics.accuracy_score
)
```



APPLICATIONS → SCALING TO LARGE IMAGE DATASETS

Images out of the CelebA dataset

CelebA dataset with a *well-known* bias (http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html)

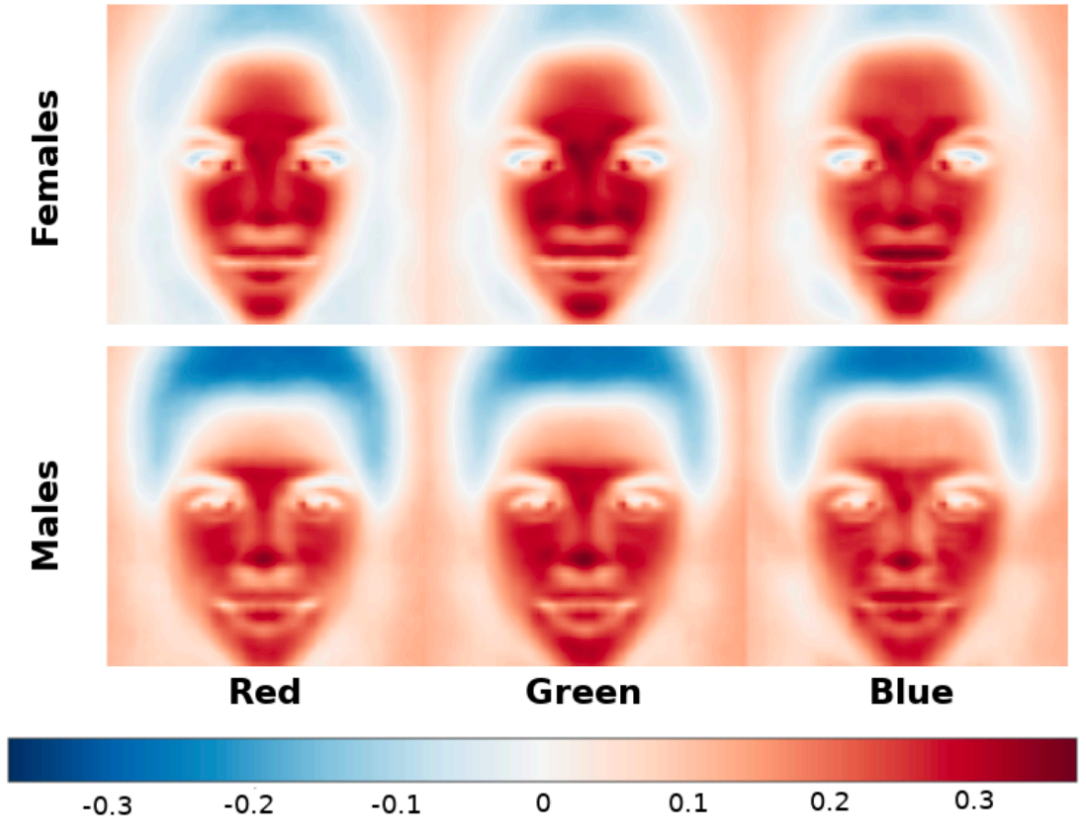
- >200K celebrity images with 40 binary annotations
- Y_i can be the *Attractive* feature

APPLICATIONS → SCALING TO LARGE IMAGE DATASETS

Images out of the CelebA dataset

ResNet 18 CNN trained to predict who is attractive → 87% of accurate predictions on the test set

What-if the average pixel intensities are locally higher or lower → Predictions == Attractive



Average pixel influences to predict whether someone is attractive or not by distinguishing males and females

APPLICATIONS → EXPLAINING THE LATENT INFORMATION IN A DNN

Images out of the CelebA dataset — ResNet18 CNN

```
ResNet(
  (conv1): Conv2d(3, 64, kernel_size=(7, 7), stride=(2, 2), padding=(3, 3), bias=False)
  (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (relu): ReLU(inplace=True)
  (maxpool): MaxPool2d(kernel_size=3, stride=2, padding=1, dilation=1, ceil_mode=False)
  (layer1): Sequential(
    (0): BasicBlock(
      (conv1): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
      (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
      (relu): ReLU(inplace=True)
      (conv2): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
      (bn2): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    )
    (1): BasicBlock(
      (conv1): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
      (bn1): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
      (relu): ReLU(inplace=True)
      (conv2): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1), bias=False)
      (bn2): BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
    )
  )
  (layer2): Sequential( ...
  )
  (layer3): Sequential( ...
  )
  (layer4): Sequential( ...
  )
  (avgpool): AdaptiveAvgPool2d(output_size=(1, 1))
  (fc): Sequential(
    (fc1): Linear(in_features=512, out_features=128, bias=True)
    (relu): ReLU()
    (fc2): Linear(in_features=128, out_features=1, bias=True)
    (output): Sigmoid()
  )
```

Hook (64,32,32)

Hook (64,16,16)

Hook (64,16,16)

Hook (128,8,8)

Hook (128,8,8)

Hook (256,4,4)

Hook (256,4,4)

Hook (512,2,2)

Hook (512,2,2)

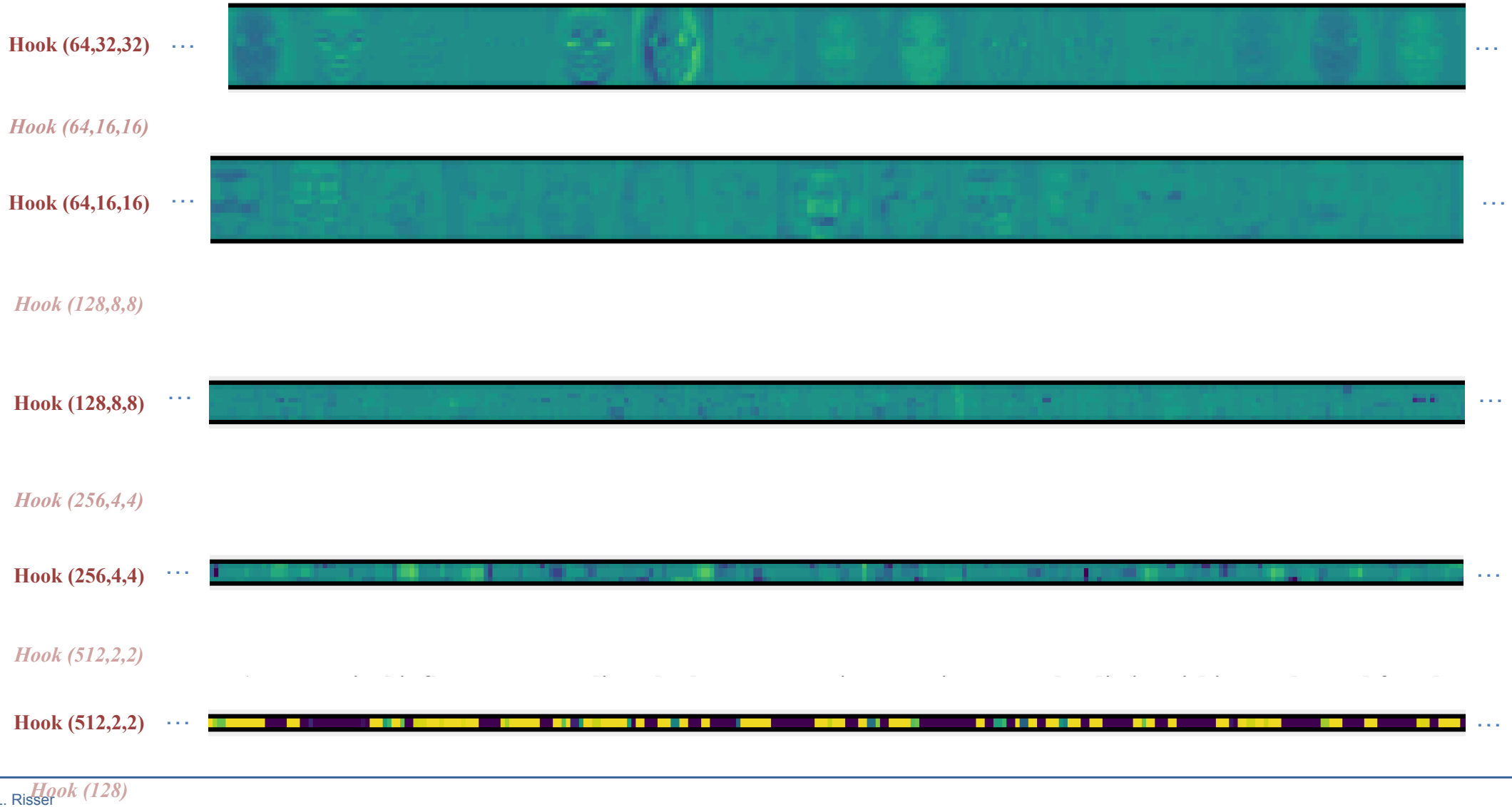
Hook (128)

Hook (1)

3

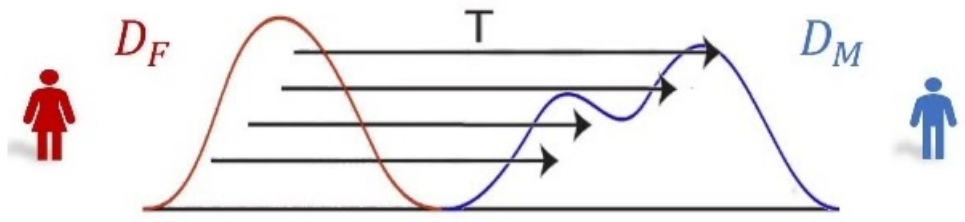
APPLICATIONS → EXPLAINING THE LATENT INFORMATION IN A DNN

Images out of the CelebA dataset — ResNet18 CNN



PRACTICAL BIAS MITIGATION USING OT

Distributional do-interventions



What if I were a man ?
How would my other characteristics change accordingly ?
Counterfactuals that are plausible

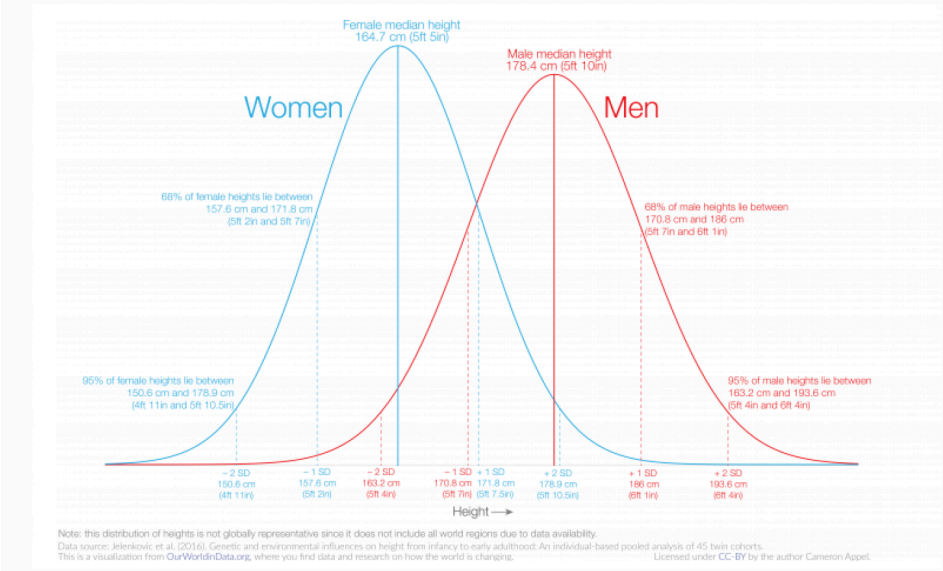


Figure 2: Distribution of female and male height

Bob is 1m86 tall if Bob was Alice, he would be ?? tall

1-D application with Quantile restrictions (Il-Drissi et al. 2022)

Quantile-constrained Wasserstein projections for robust interpretability of numerical and machine learning models

Marouane Il Idrissi[a,b,c,e], Nicolas Bousquet[a,b,d], Fabrice Gamboa[c], Bertrand Iooss[a,b,c], Jean-Michel Loubes[c]

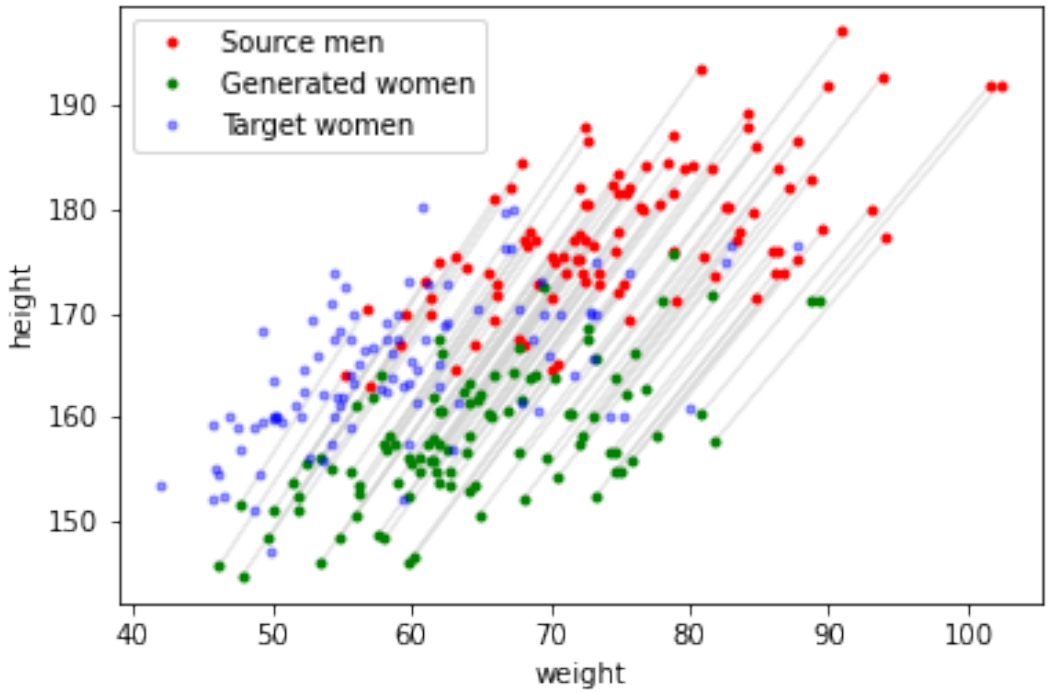
The transformation may not be one to one but
Lead to counterfactual worlds with

PRACTICAL BIAS MITIGATION USING OT

Explaining using counterfactuals

Idea : Transport-based Counterfactual Model : either a map or random couplings.

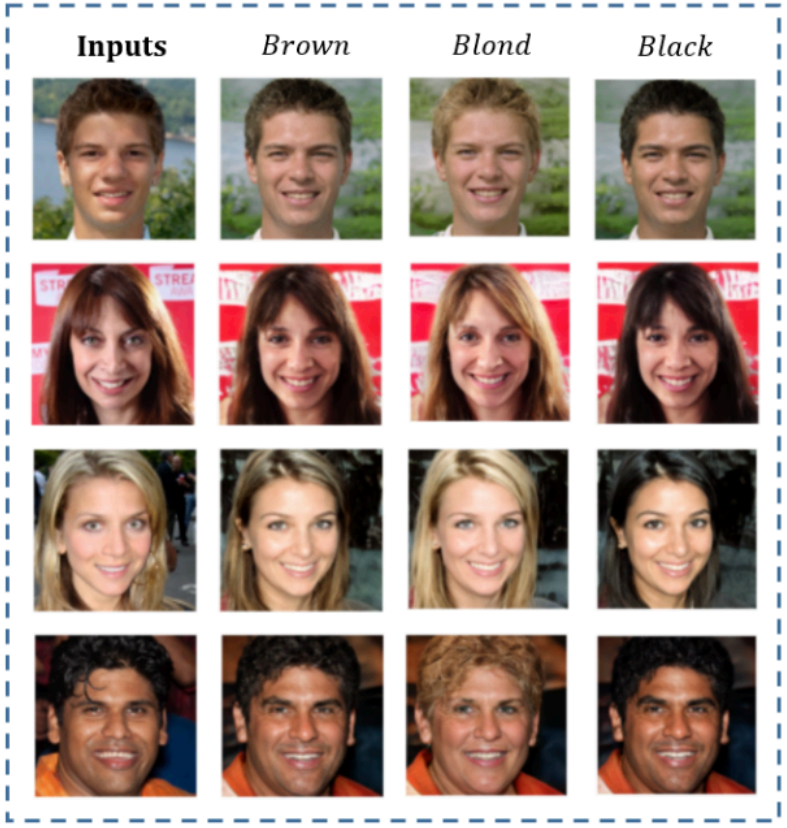
$$T_{\langle s'|s\rangle} := \arg \min_{T:\, T_\sharp \mu_s = \mu_{s'}} \int \|x - T(x)\|^2 d\mu_s(x) \qquad \pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{s'})$$



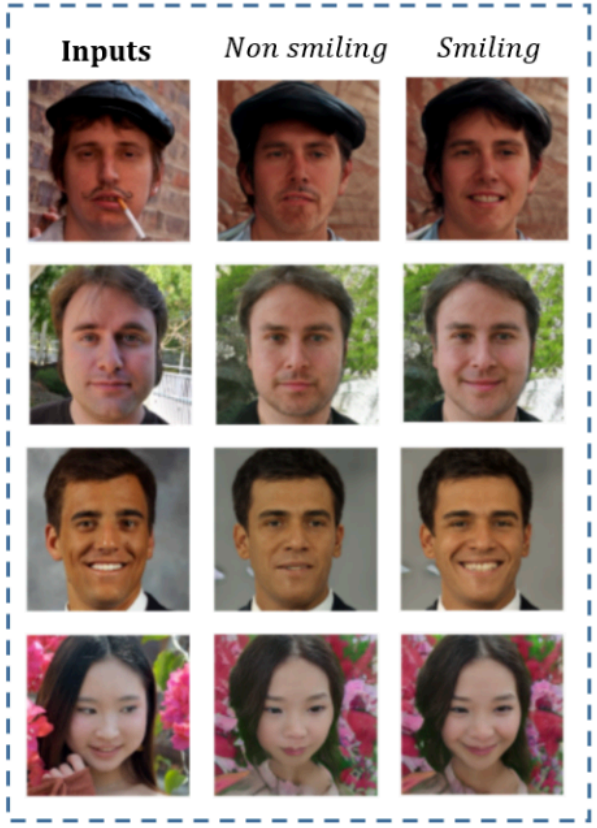
PRACTICAL BIAS MITIGATION USING OT

Explaining using counterfactuals

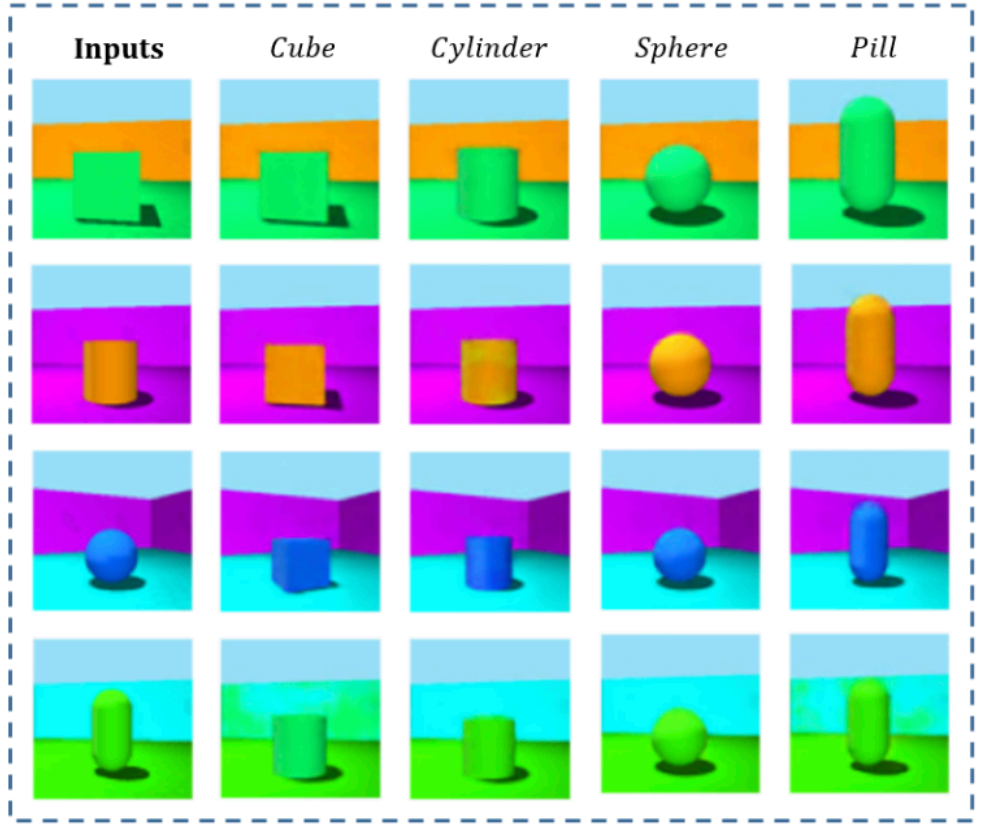
FFHQ (hair)



FFHQ (smile)



Shapes 3D



… INTERESTED IN GEMS-AI?

Version 1.0 available

https://arxiv.org/pdf/1810.07924.pdf
https://gems-ai.aniti.fr/
https://github.com/XAI-ANITI/ethik

Version 2.0

1: time series with specific stress functions (auto-correlations, periodicity, extreme values…)

2: NLP applications

3: fairness models

Formation CNRS Entreprises (2 jours)

Direction des Relations avec les Entreprises

CNRS FORMATION ENTREPRISES

Environnement scientifique et technique de la formation

INSTITUT de MATHEMATIQUES de TOULOUSE

Institut de mathématiques de Toulouse - UMR 5219

RESPONSABLES

Laurent RISSER
Ingénieur de recherche
UMR 5219

Jean-Michel LOUBES
Professeur
UMR 5219

LIEU
TOULOUSE (31)

ORGANISATION
2 jours
De 5 à 10 stagiaires

Formation - Intelligence artificielle de confiance : biais en IA et explicabilité - Mise en oeuvre pratique

NOUVEAU

OBJECTIFS

- Comprendre les mécanismes à l'œuvre dans l'intelligence artificielle
- Comprendre la problématique du biais et de l'explicabilité dans les données et dans l'algorithme
- Détecter le biais et s'en prémunir
- Etre capable de définir les décisions algorithmiques et d'en comprendre l'explicabilité
- Connaître les nécessités juridiques liées aux réglementations nationales et européennes

PUBLICS

Techniciens et ingénieurs en production, traitement, analyse de données et enquêtes. Les stagiaires doivent avoir une expérience sur la manipulation de données, mais pas nécessairement en apprentissage automatique.

N.B. Une formation d'introduction pour les décideurs est également proposée : "Intelligence artificielle de confiance : biais en IA et explicabilité - Introduction pour les décideurs".

Diplôme d'université (60 heures)

FORMATION QUALIFIANTE

Biais et explicabilité en IA

UNIVERSITÉ TOULOUSE III PAUL SABATIER

ANITI Université Fédérale Toulouse Midi-Pyrénées

Formation diplomante (accessible également en Formation Continue)

Contenu :

- Machine learning : Introduction, methodologies et outils
- Compréhension, détection et réduction des biais
- Explicabilité des décisions « boites noires »
- Sensibilisation au Droit encadrant l'I.A.
- Applications

Deux parcours :

- Décideurs, cadres dirigeant, CTO
- Techniciens et ingénieurs de production

GEMS.AI ET LA PLATEFORME AEROLINE.AI

Plateforme Aeroline AI :

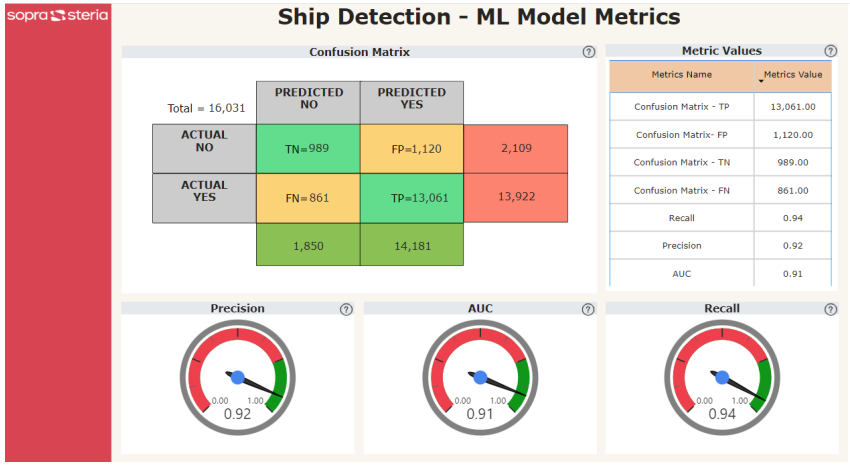
- Liée à nos offres AI-Powered Aerospace: Supply Chain, Manufacturing, Engineering & Customer Services
- Plateforme regroupant différents assets/démos d'IA eux-mêmes déployés sur Azure.

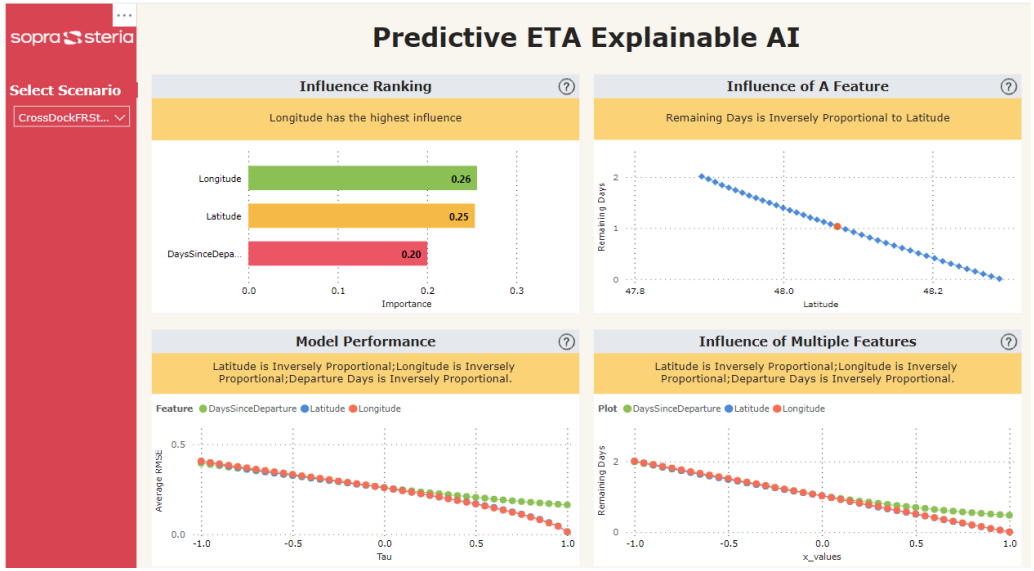
Pourquoi Gems AI ? :

- Explicabilité primordiale dans l'Aéronautique
- Facilité d'intégration
- Exhaustivité des features d'explicabilité

Prochaines étapes :

- Harmonisation entre les différents assets
- Inclure plus de features





Merci !