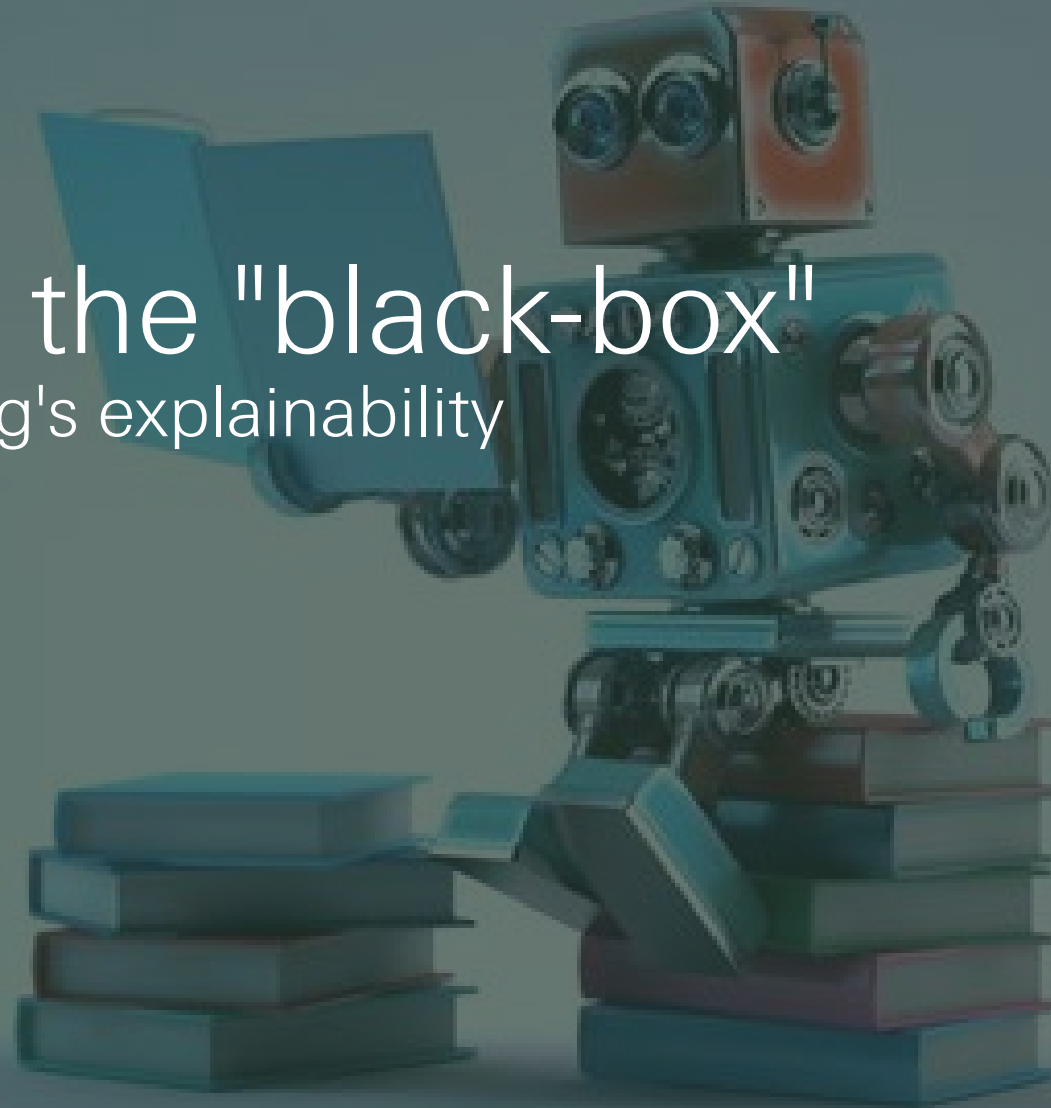


A peek into the "black-box"

Predictive modelling's explainability

Jianshu Weng



Disclaimer: The views expressed herein are those of the presenter; they do not necessarily reflect the views or positions of any organization whatsoever with which I have been, am now, or will be affiliated.



**Interpret more
complex models**



GLM



**Machine Learning /
Predictive Modelling/
Explainability**



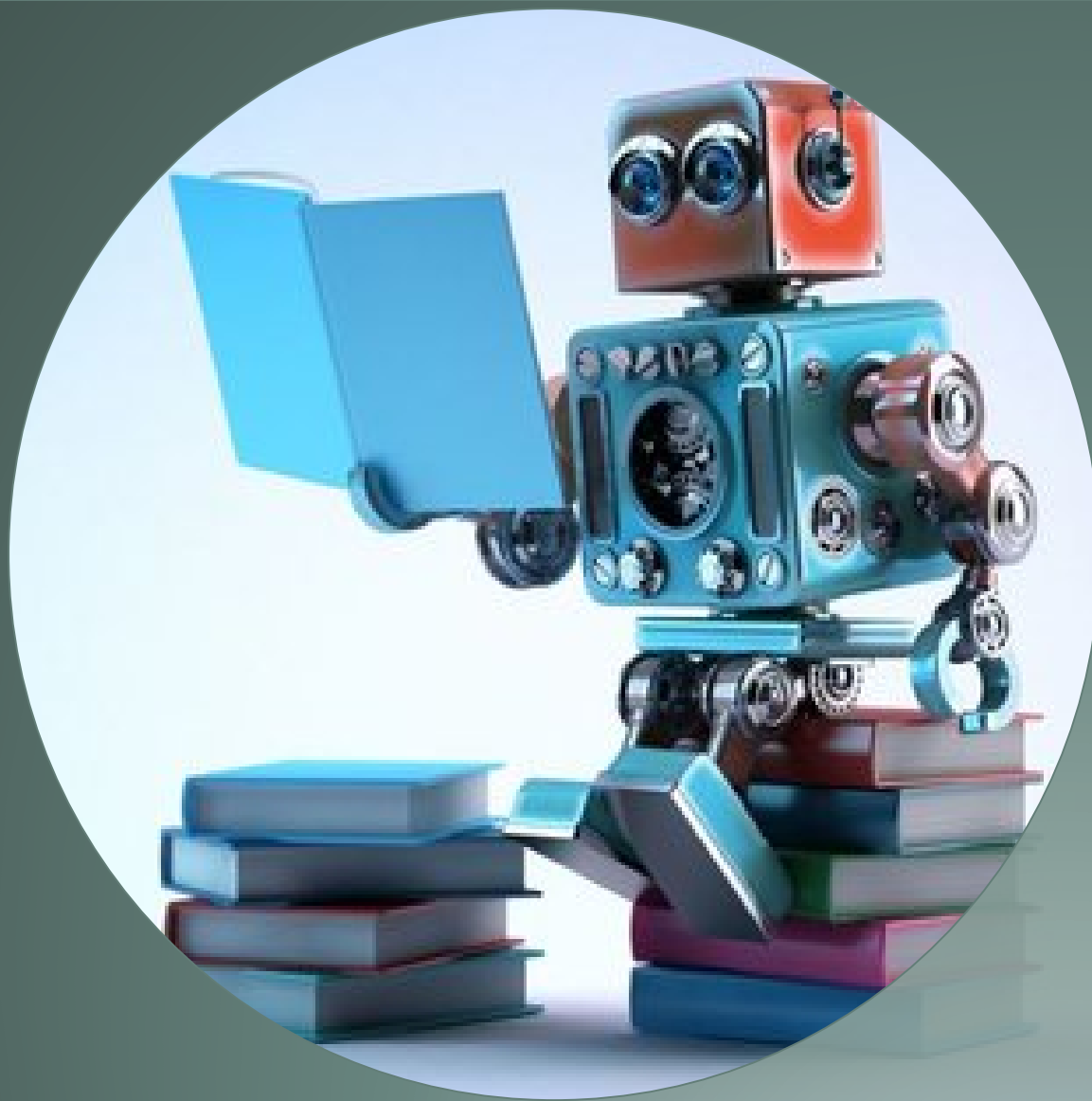
**Machine Learning /
Predictive Modelling /
Explainability**

machine learning

Explainability

predictive modelling

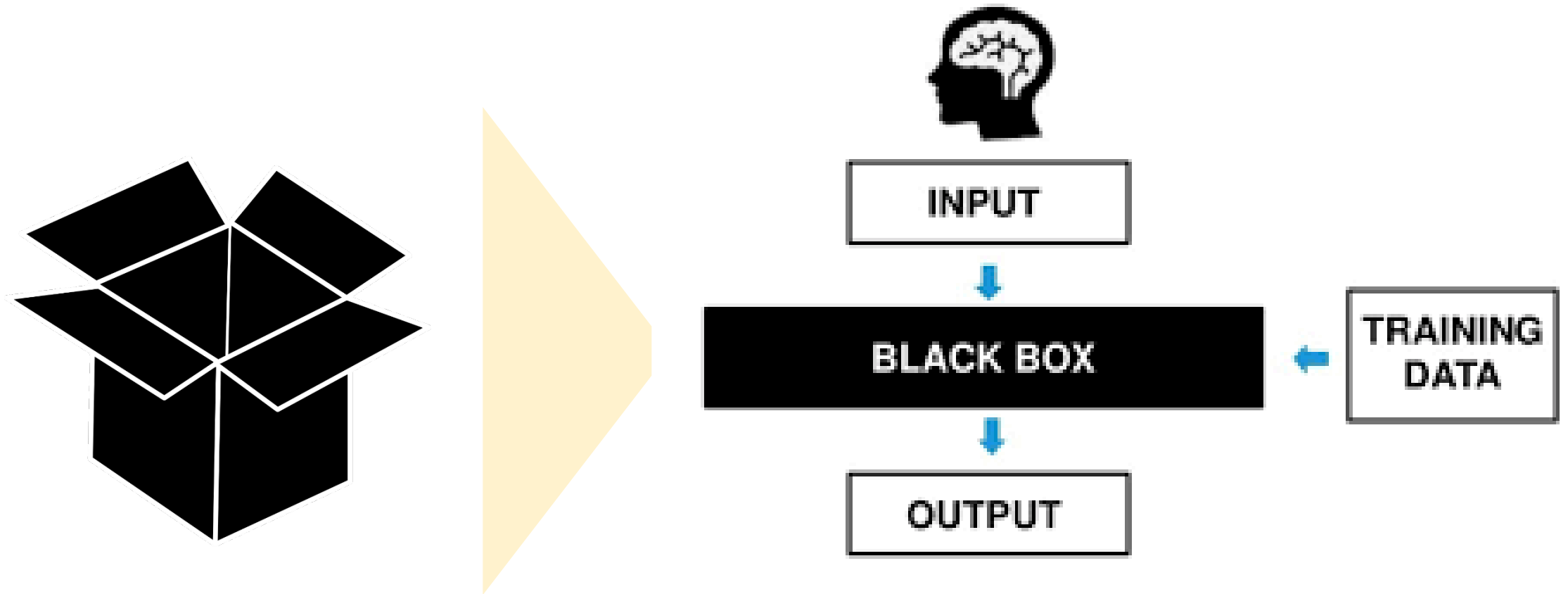




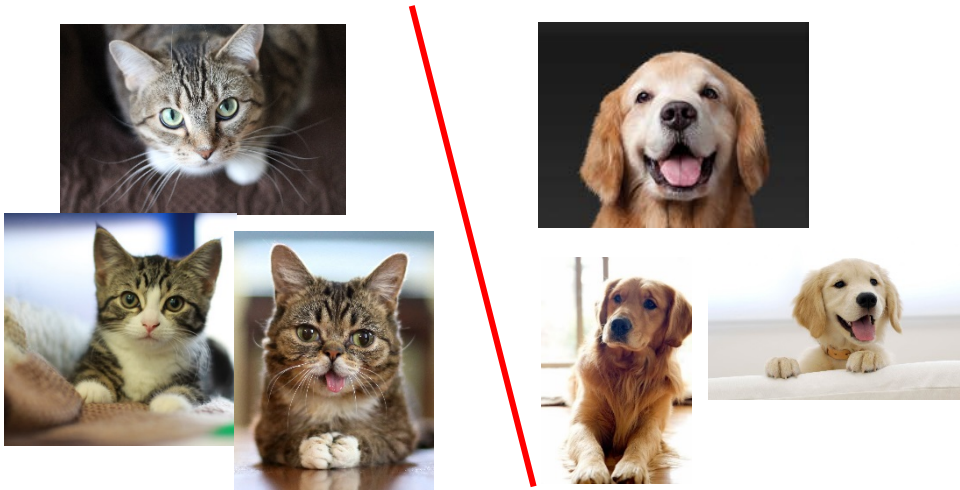
What is machine learning?

A program **learns** to solve a task from past experience, and **improves** with more experience

MACHINE LEARNING



Supervised learning



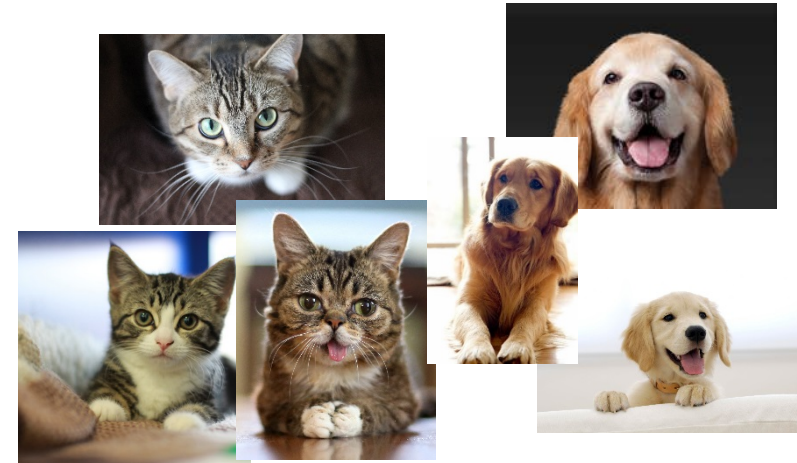
This looks like
the dogs I saw
before, it should
be a dog too

It has different names in different contexts:

- **predictive modelling**

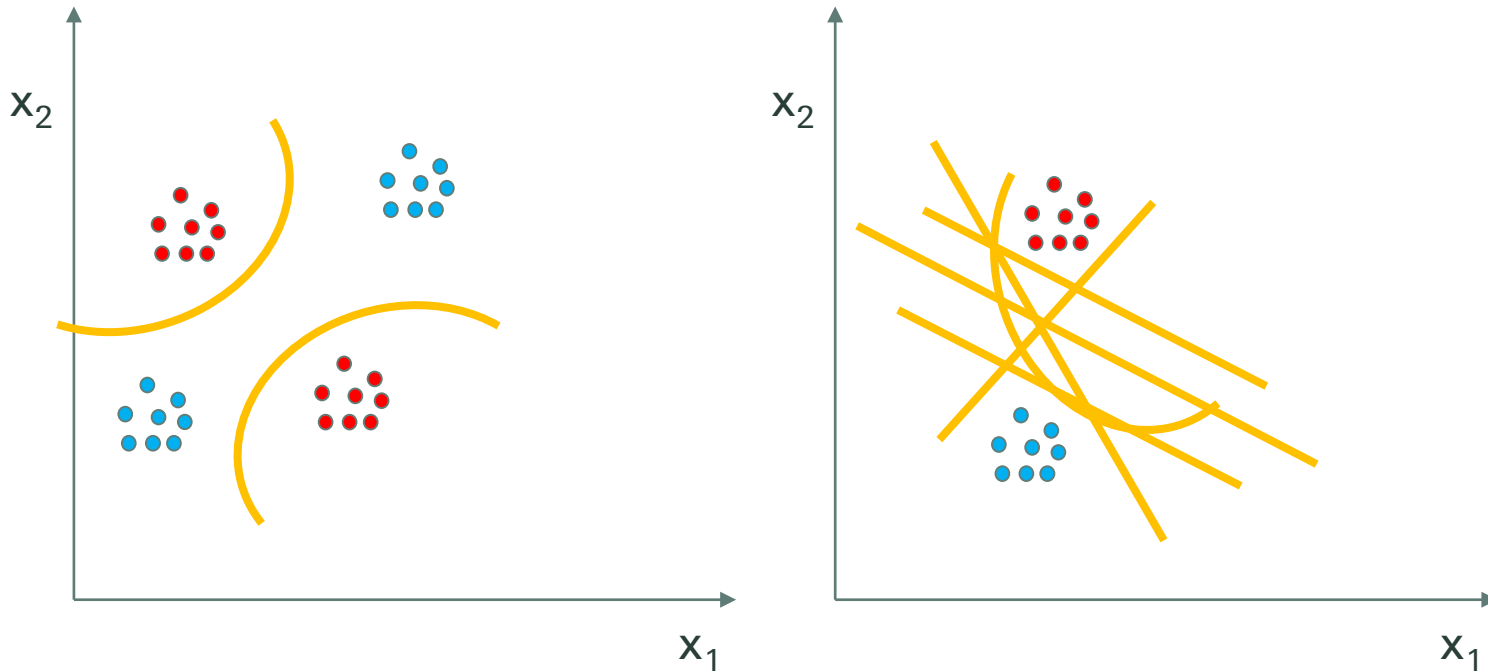
- pattern recognition
- ...

Unsupervised learning



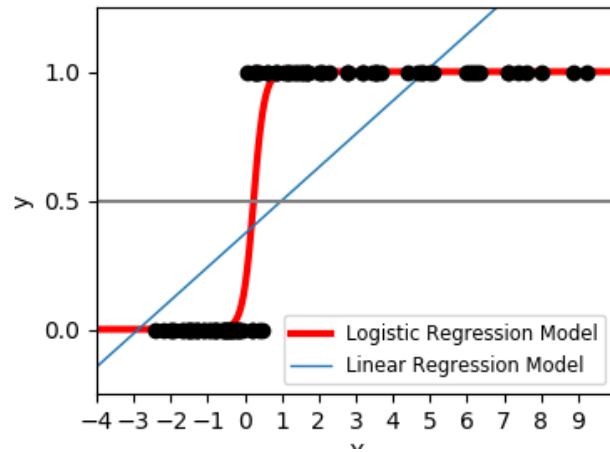
I can see there are two types of
animals. But I don't know what
they are. Can anybody tell me
what they are?

Predictive modelling – what is it exactly?

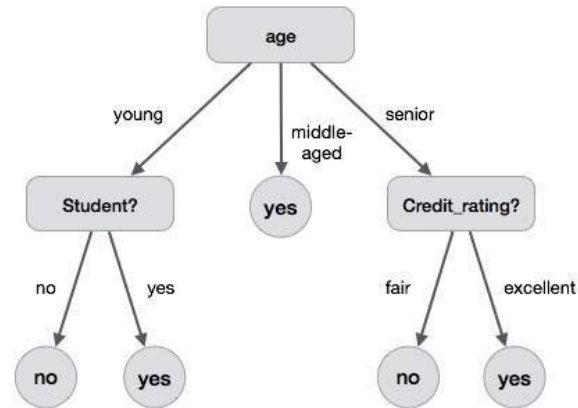


- Mathematically, we are searching for a hyper-plane (decision function) to split different classes as clean as possible.
- Isn't it just fitting a function, which we have been doing for ages already?
- The main **challenge** is to achieve good generalization capability, i.e. the model should generalize to unseen data
 - In most of the cases, classes are not linearly separable.
 - There are multiple such planes available. Which one is optimal?
 - Avoid overfitting

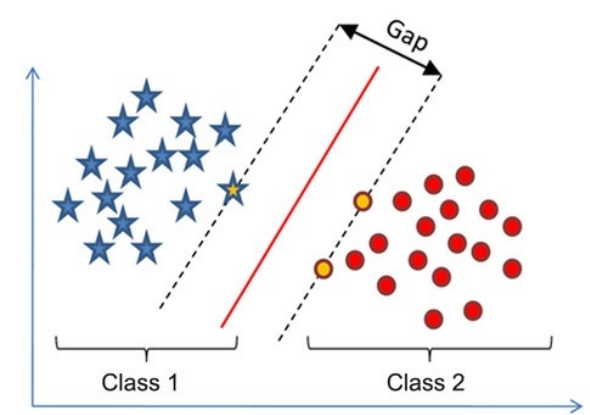
Some commonly used algorithms/models



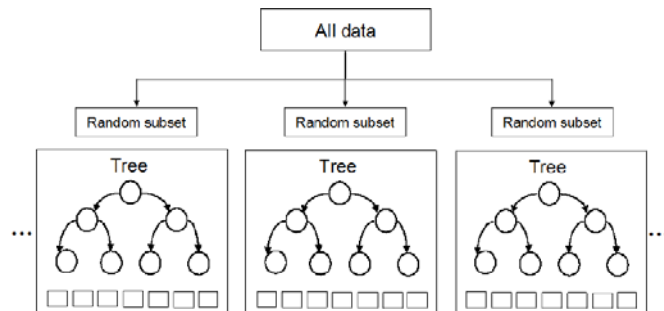
GLM



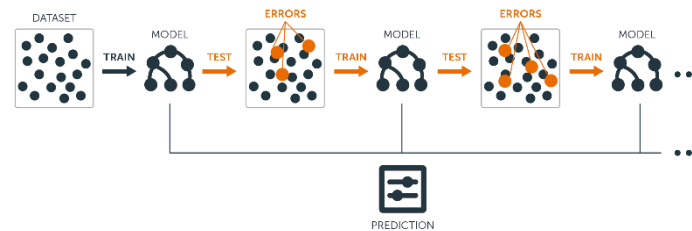
Decision tree



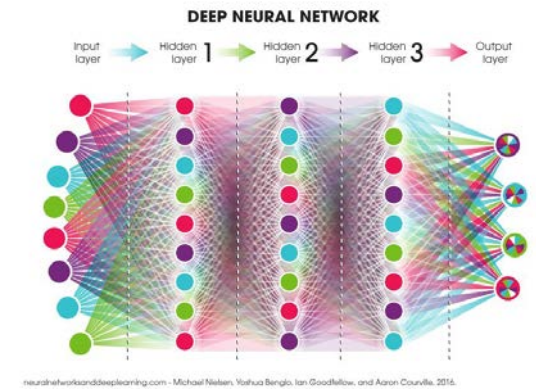
SVM



Random forest



Gradient boosting tree



Neural network

Category of models based on complexity of the decision functions

A

linear, monotonic function

A change in any given independent variable, the decision function changes at a defined rate, in only one direction, and at a magnitude represented by a coefficient.

B

nonlinear, monotonic function

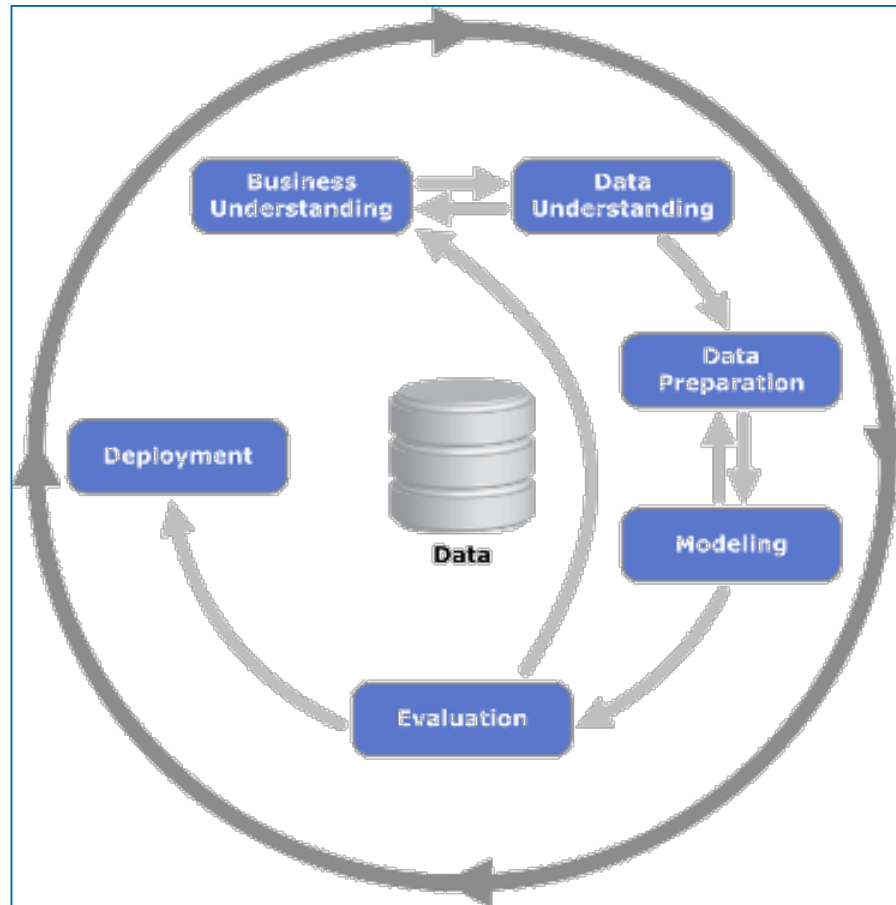
There is no single coefficient that represents the change in the decision function induced by a change in a single independent variable, nonlinear and monotonic functions do always change in one direction as a single input variable changes.

C

nonlinear, non-monotonic function

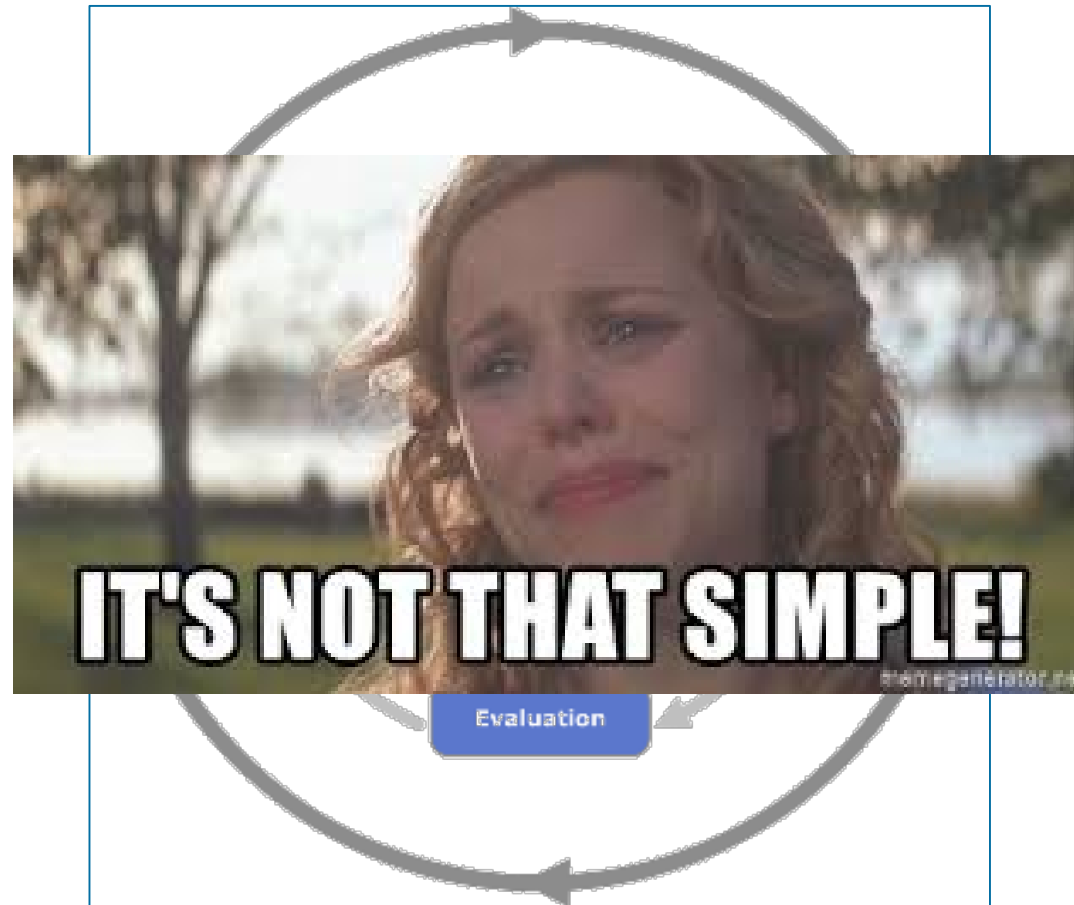
The decision function changes in a positive and negative direction and at a varying rate for any change in an independent variable. Typically, relative variable importance measures are reported.

Some detour before the explainability question.



CRISP-DM

Some detour before the explainability question.



CRISP-DM

Now, the **explainability** question.

Who do we explain to, and what do we want to explain?

01

Regulators

They need an explanation so that they can judge whether the model violate any regulatory requirement, e.g. use of sensitive data, discriminatory treatment, etc.



02

Data scientists/practitioners

They need an explanation so that they can implement the model and make sure it achieves the desired outcome.



03

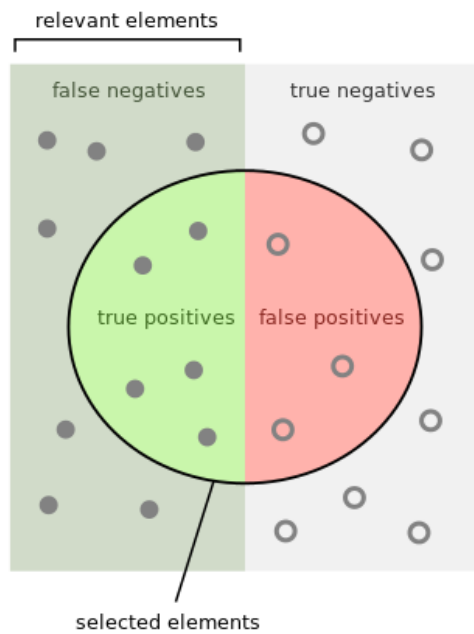
End-users

They need an explanation so that they can understand the logic behind a specific outcome, i.e. why am I getting this output?



This is by no means an exhaustive list!





How many selected items are relevant?

Precision =

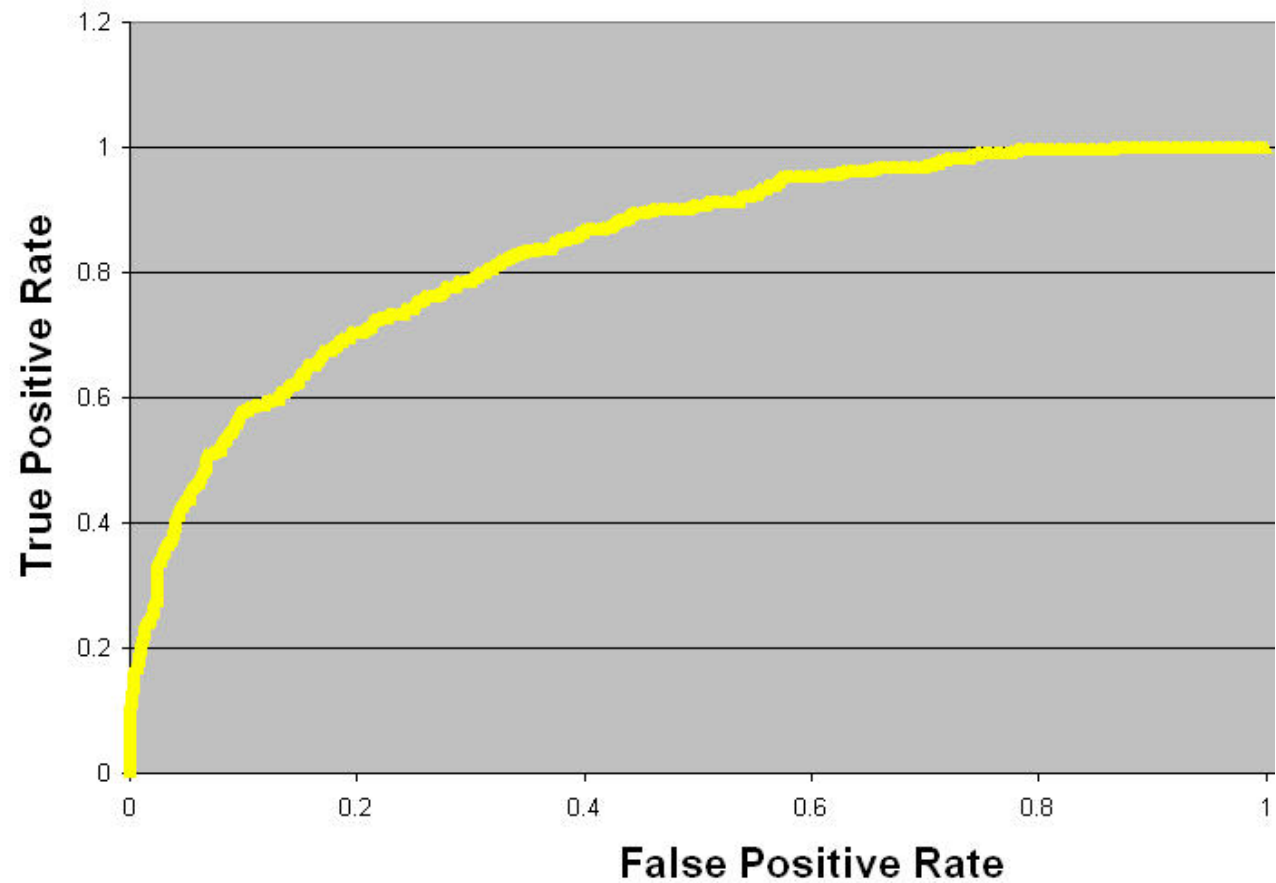


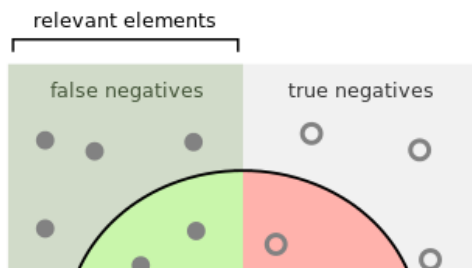
How many relevant items are selected?

Recall =



ROC Curve





ROC Curve



So what?
I still do not know why I'm
getting coal this year!!



John Hancock will include fitness tracking in all life insurance policies

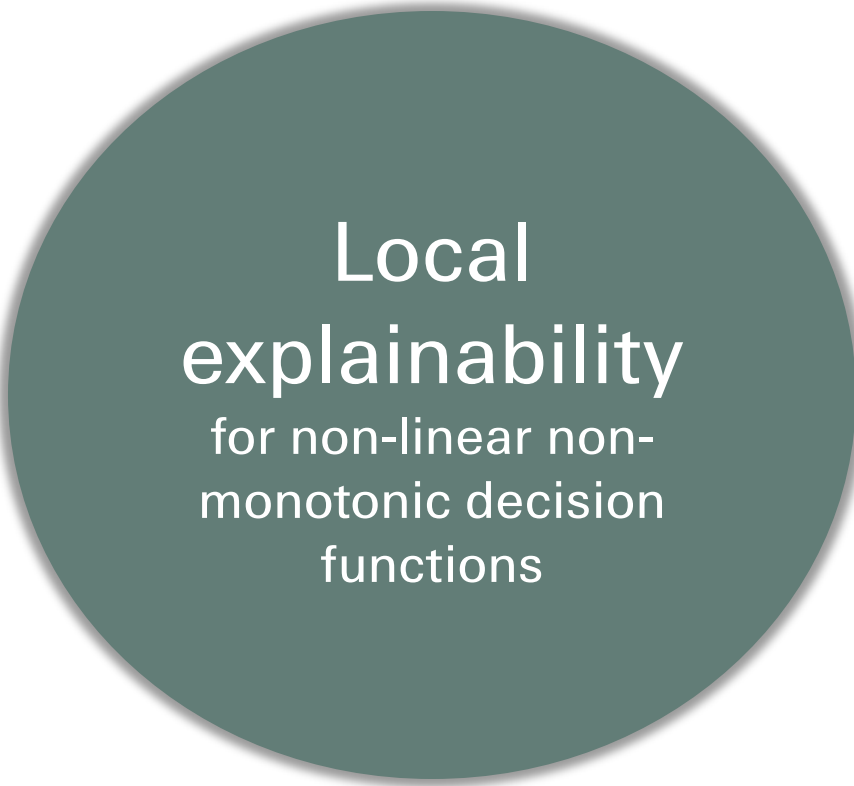
REUTERS SEPTEMBER 19, 2018 10:35 AM



(Reuters) — John Hancock, one of the oldest and largest North American life insurers, will stop underwriting traditional life insurance and instead sell only interactive policies that track fitness and health data through wearable devices and smartphones, the company said on Wednesday.



**Global
explainability**
for (non-)linear
monotonic decision
functions



**Local
explainability**
for non-linear non-
monotonic decision
functions



**Interpret more
complex models**



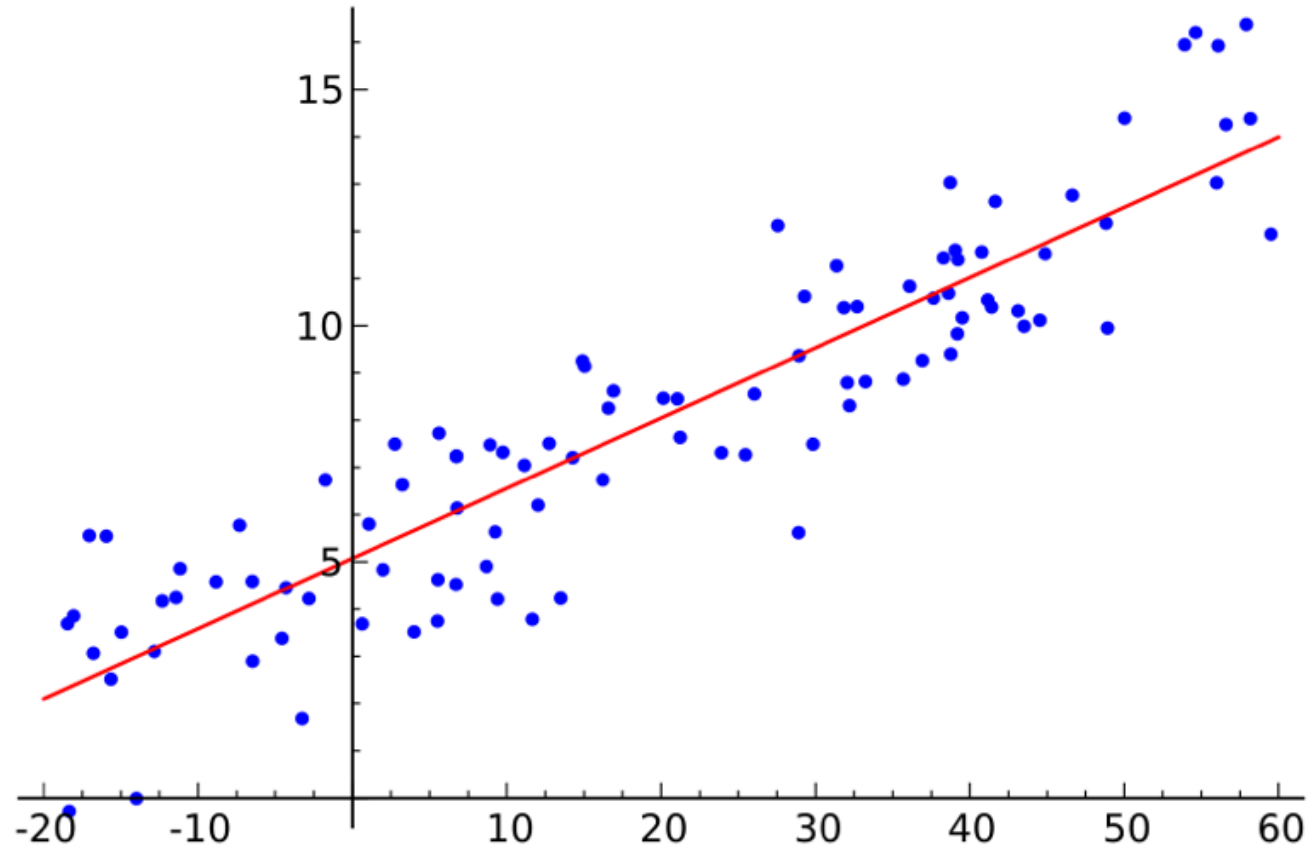
GLM



**Machine Learning /
Predictive Modelling/
Explainability**



Linear model



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

Linear model

Typical form:

$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j$$

X_j is typically called input variable, independent variable, regressor. β_j is typically called effect, or regression coefficient. β_0 is called intercept or bias.

A few assumptions:

- linear relationship
- perfect collinearity between the independent variables
- error is IID (independent and identically distributed) and should follow normal distribution



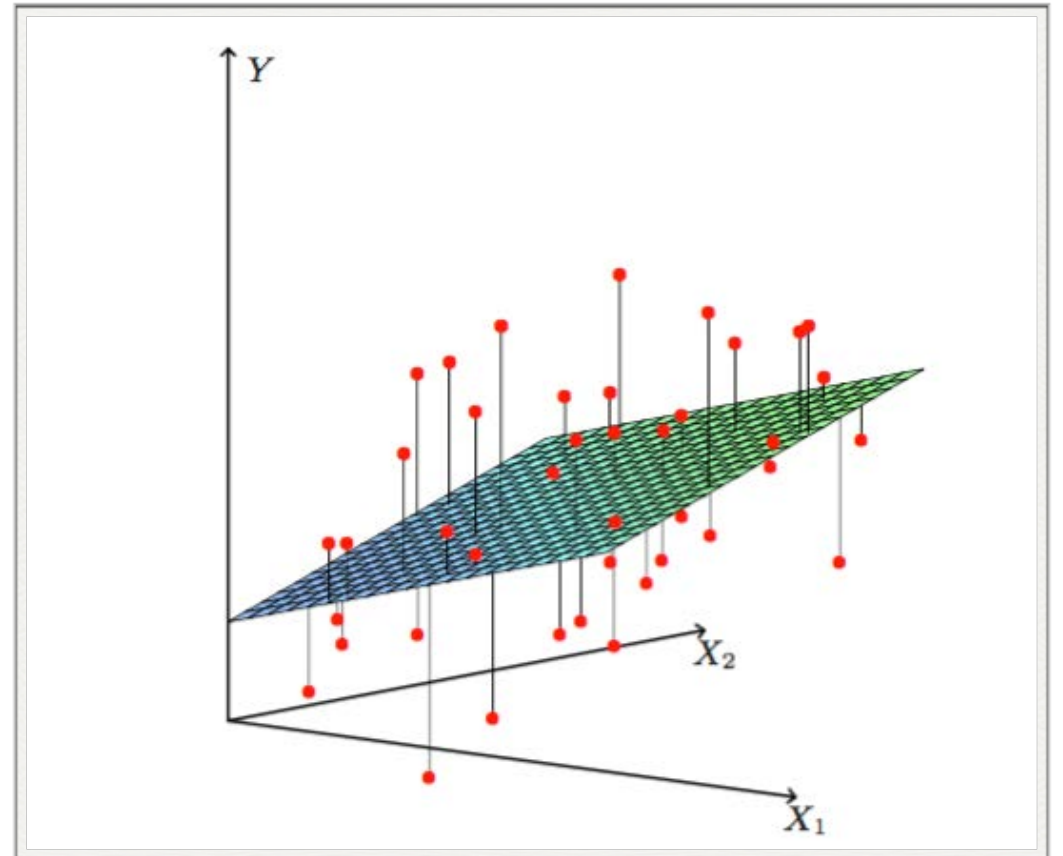
Goal: To estimate β .

Estimation method

⚙ Ordinary Least Square (OLS)

$$RSS(\beta) = \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$$

Find a linear function of X that minimizes the sum of squared residual from Y .



Generalized Linear Model (GLM)

Typical form:

$$g(E(Y|X)) = X\beta$$

g is the link function, which is the part making it different from linear model.

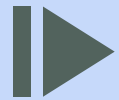
Common link functions and probability families

Model	Distribution	Link
logistic regression	binomial	logit
multinomial logit regression	multinomial	generalized logit
Poisson regression	Poisson	log

How can Poisson Regression help us



- Poisson Regression can be used to model claim count
- which cohorts are more likely to claim to more and how much more



To model claim count (**A**)

$$\log(E(Y|X = x)) = X$$



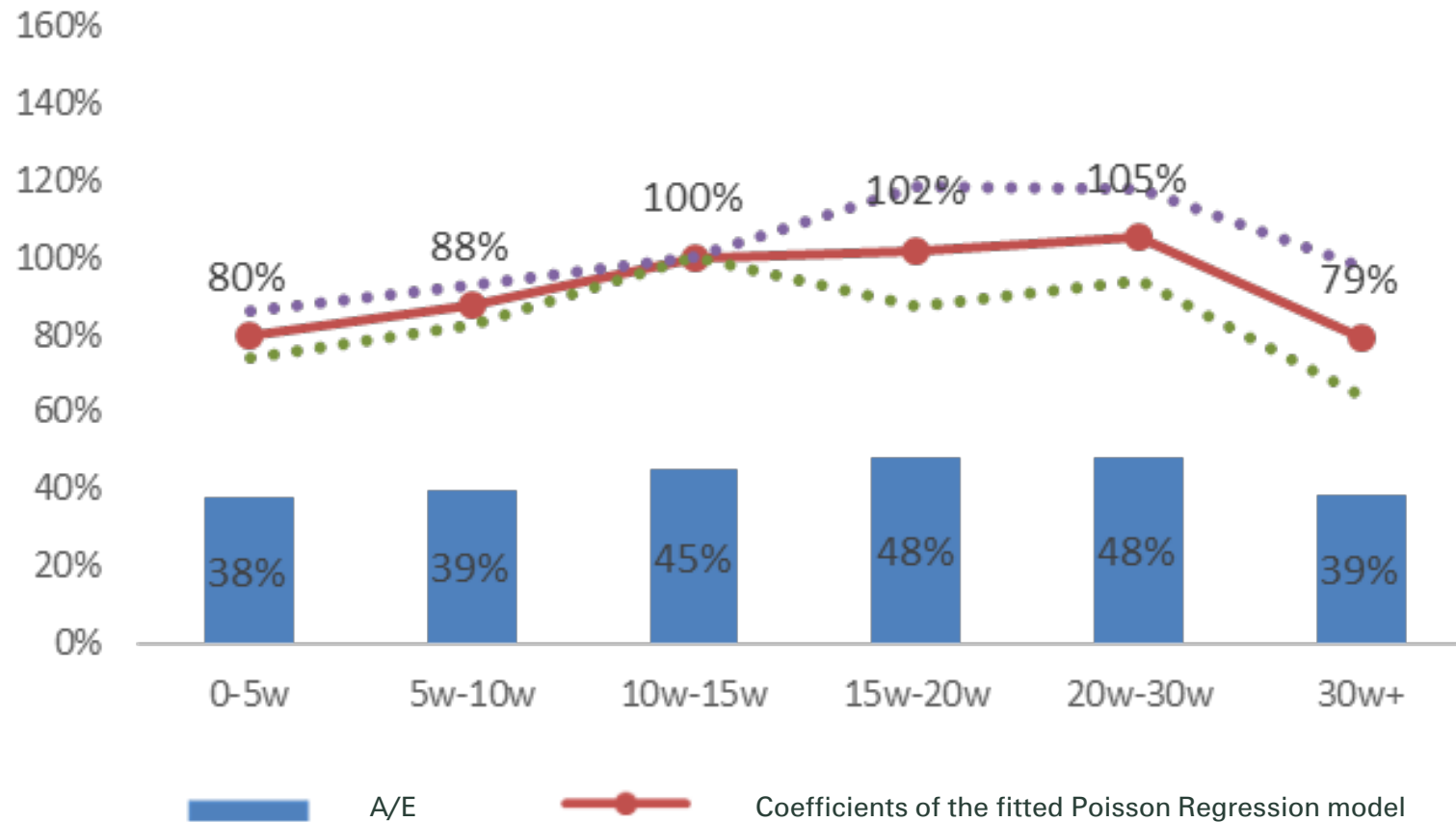
How about A/E: treat E as the offset in Poisson Regression

$$\log(E(Y|X = x)) = \log(\text{exposure}) + X\beta \rightarrow \log(E(Y|X = x)) - \log(\text{exposure}) = X\beta \rightarrow$$

$$\log\left(\frac{E(Y|X=x)}{\text{exposure}}\right) = X\beta$$

- The coefficients β basically quantify the change in $\log(A/E)$ when individual factor changes for one unit.

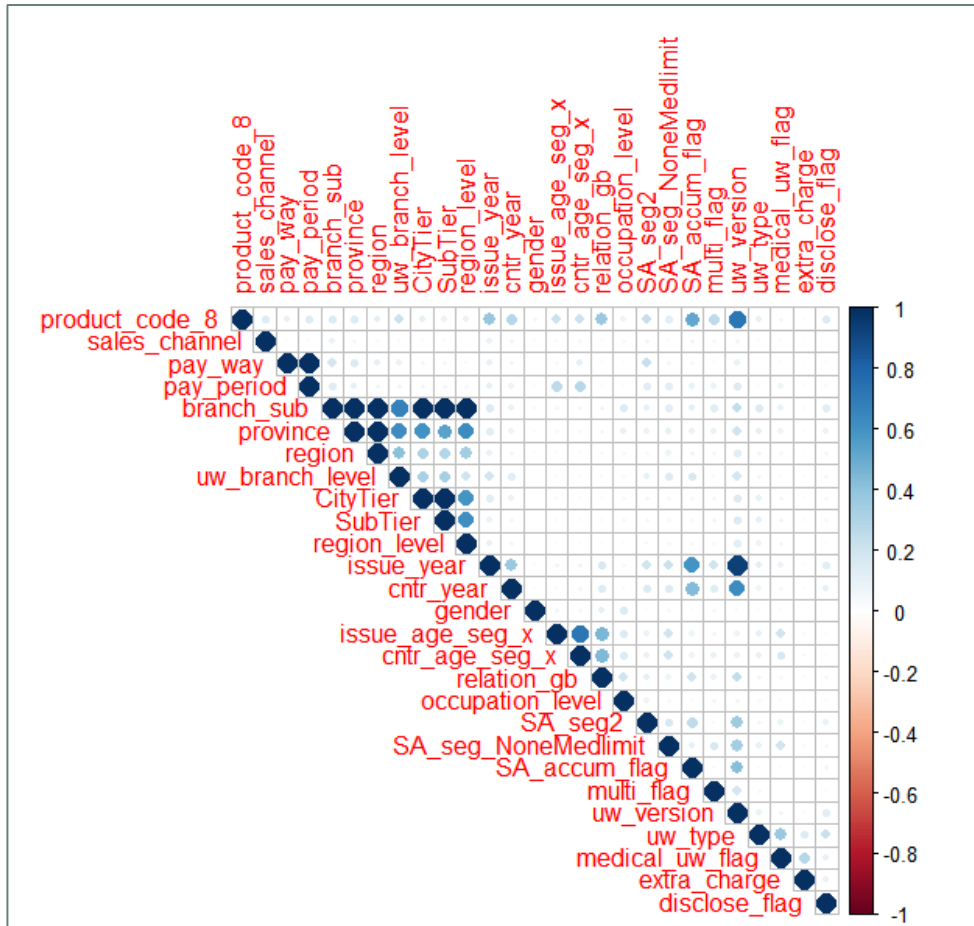
An example



How is GLM typically done



Example - correlation analysis



- Groups of features exhibiting high correlation:
 - branch_sub, province, uw_branch_level, city_tier, and region_level
 - uw_version, product_code_8, and issue_year
- Eventually, following features are selected:
 - product_code_8, pay_period, region, uw_branch_level, issue_year, cntr_year, gender_cntr_age_seg, and SA_seg2 are selected
 - The decision is based on the correlation analysis results and domain knowledge.

Example – model fitting

Model fitting result (partial)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.025e+00	2.239e-01	-4.578	4.70e-06	***
product_code_8534000	-8.851e-01	1.084e-01	-8.166	3.19e-16	***
product_code_8535000	-7.888e-01	9.995e-02	-7.893	2.96e-15	***
product_code_8542000	-1.209e+01	1.371e+02	-0.088	0.929696	
product_code_8543000	-1.196e+01	9.574e+01	-0.125	0.900567	
product_code_8544000	-1.415e+00	1.571e-01	-9.006	< 2e-16	***
product_code_8555000	-6.671e-01	1.296e-01	-5.148	2.63e-07	***
product_code_8556000	-1.301e+00	3.252e-01	-4.001	6.32e-05	***
product_code_8724000	-8.559e-01	1.126e-01	-7.605	2.86e-14	***
product_code_8799000	-7.803e-01	1.043e-01	-7.479	7.48e-14	***
pay_period10	7.429e-01	1.271e-01	5.846	5.03e-09	***
pay_period20	8.533e-01	1.263e-01	6.758	1.40e-11	***
pay_period30	1.048e+00	1.286e-01	8.149	3.67e-16	***
pay_period5	6.925e-01	1.673e-01	4.138	3.50e-05	***
regionE	-8.653e-03	2.897e-02	-0.299	0.765157	
regionN	-5.246e-02	3.080e-02	-1.703	0.088480	.
regionNE	4.534e-01	3.715e-02	12.206	< 2e-16	***
regionNW	-1.915e-01	3.656e-02	-5.238	1.62e-07	***
regionS	-1.879e-01	4.302e-02	-4.368	1.26e-05	***
regionSw	-2.124e-01	4.648e-02	-4.569	4.89e-06	***



Are we able to “**automate**” this process? Especially when we have a lot of attributes, which is way more than traditional pricing factors, e.g. age, gender, occupation class, income level, etc.

Automation: filter those highly correlated attributes, and fit the model automatically



Based on
another rule
usually stays. It

Regularization: **why** (data science view)

Let's revisit OLS (similar for inference of GLM parameters)

$$RSS(\beta) = \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j)^2$$

- Present it in vector form: $RSS(\beta) = (y - X\beta)^T (y - X\beta)$
- differentiating W.R.T. β , we have $\frac{\partial RSS}{\partial \beta} = -2X^T(y - X\beta)$, $\frac{\partial^2 RSS}{\partial \beta \partial \beta^T} = 2X^T X$
- let $\frac{\partial RSS}{\partial \beta} = 0$, we have $X^T(y - X\beta) = 0$, and subsequently $\hat{\beta} = (X^T X)^{-1} X^T y$
- This is based on the assumption that X is full-rank, i.e. independent variables are independent. when this assumption is not fulfilled, the solution is not stable (high variance)

Example:

- x_1 : height, x_2 : weight, $\hat{y} = ax_1 + bx_2 + c$: hormone level.
- x_1 and x_2 are correlated. We may have $a > 0$ very large, we can cancel this by having a very small $b < 0$. In other words, the estimates of a and b will show high variance.
- how to alleviate: apply constraint, e.g. $a^2 + b^2 \leq t$, to reduce the variance.

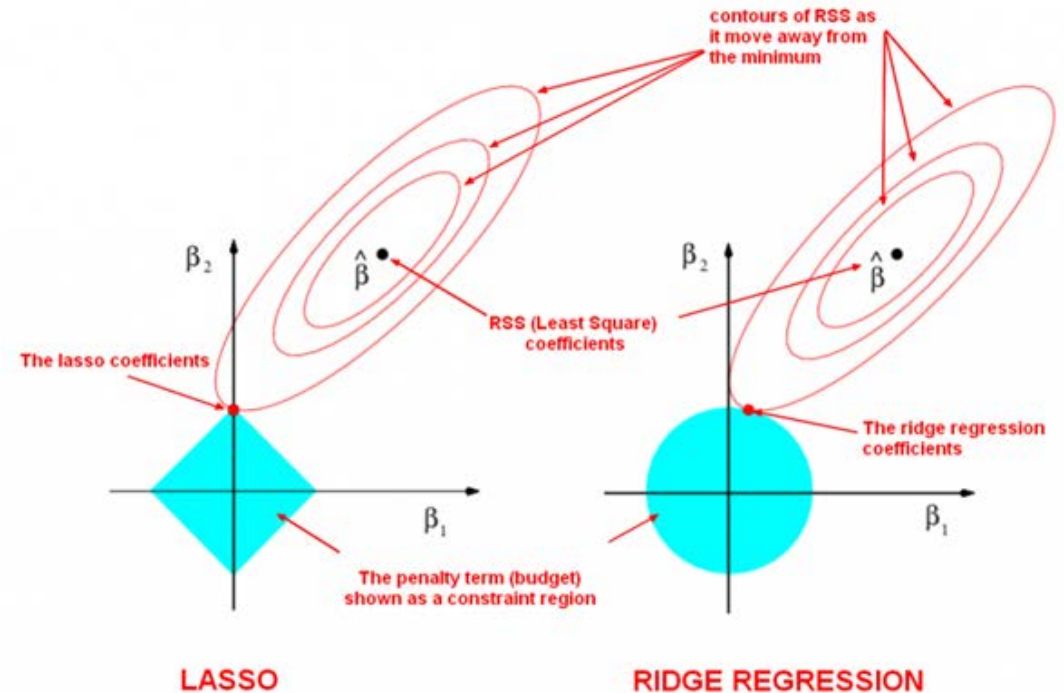
Example - “high variance”

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.025e+00	2.239e-01	-4.578	4.70e-06	***
product_code_8534000	-8.851e-01	1.084e-01	-8.166	3.19e-16	***
product_code_8535000	-7.888e-01	9.995e-02	-7.893	2.96e-15	***
product_code_8542000	-1.209e+01	1.371e+02	-0.088	0.929696	
product_code_8543000	-1.196e+01	9.574e+01	-0.125	0.900567	
product_code_8544000	-1.415e+00	1.571e-01	-9.006	< 2e-16	***
product_code_8555000	-6.671e-01	1.296e-01	-5.148	2.63e-07	***
product_code_8556000	-1.301e+00	3.252e-01	-4.001	6.32e-05	***
product_code_8724000	-8.559e-01	1.126e-01	-7.605	2.86e-14	***
product_code_8799000	-7.803e-01	1.043e-01	-7.479	7.48e-14	***
pay_period10	7.429e-01	1.271e-01	5.846	5.03e-09	***
pay_period20	8.533e-01	1.263e-01	6.758	1.40e-11	***
pay_period30	1.048e+00	1.286e-01	8.149	3.67e-16	***
pay_period5	6.925e-01	1.673e-01	4.138	3.50e-05	***
regionEast	-8.653e-03	2.897e-02	-0.299	0.765157	
regionNorth	-5.246e-02	3.080e-02	-1.703	0.088480	.
regionNortheast	4.534e-01	3.715e-02	12.206	< 2e-16	***
regionNorthwest	-1.915e-01	3.656e-02	-5.238	1.62e-07	***
regionSouth	-1.879e-01	4.302e-02	-4.368	1.26e-05	***
regionSouthwest	-2.124e-01	4.648e-02	-4.569	4.89e-06	***

Regularization: how (data science view)

- Regularization
 - apply penalty to reduce variance
 - typically L1 or L2 norm (LASSO regression / Ridge regression)
- L1 regularization vs. L2 regularization
 - L1 regularization $\|\beta\|_1$: pick one correlated feature, others discarded
 - L2 regularization $\frac{1}{2}\|\beta\|_2^2$: correlated features coefficients are pushed to the some restricted value
- We can also combine both to have elastic-net regularization $\alpha\|\beta\|_1 + \frac{1-\alpha}{2}\|\beta\|_2^2, \quad \alpha \in [0, 1]$



Regularization: implication

- It is a natural question to ask for standard errors, CI, etc. of coefficients. However, with regularization, those quantities are not really meaningful any more. Why?
- What regularization does is essentially to reduce the variance of coefficients by introducing some bias (there is always a trade-off between bias and variance). Therefore, reporting standard error and/or CI will not give a close estimate of the how close the true value of the coefficient may be to the point estimate.



**PROCEED
WITH
CAUTION**

- GLM has a linear monotonic decision function.
- It is commonly used in actuarial community, especially for non-life pricing.
- It is also one of the supervised learning techniques that commonly used in data science community
- Regularization can be applied to automate feature selection and model fitting

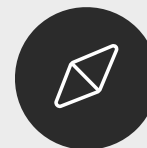
**Take-away about
GLM
(with regularization)**



**Interpret more
complex models**



GLM



**Machine Learning /
Predictive Modelling**



**Interpret more
complex models**



Global
explainability
for (non-)linear
monotonic decision
functions

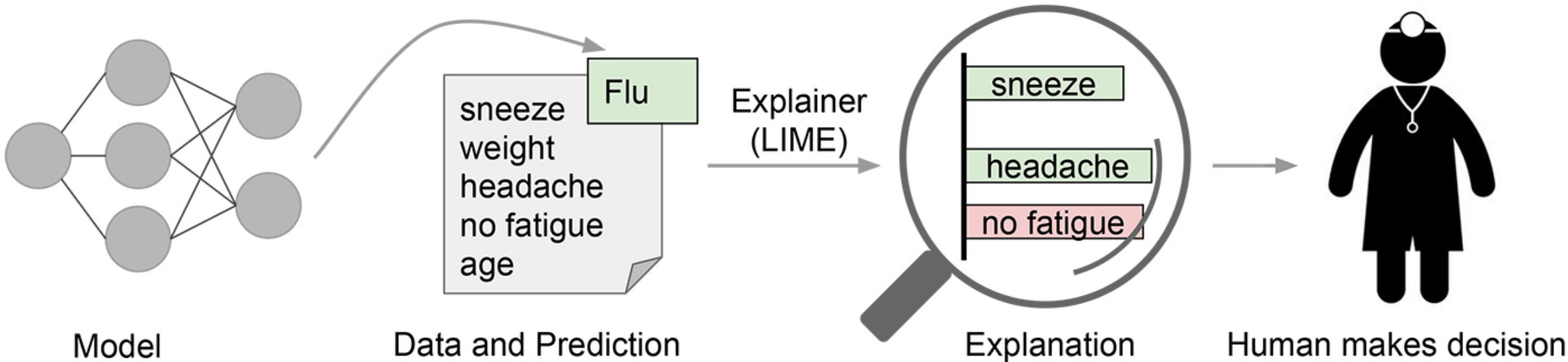


Local
explainability
for non-linear non-
monotonic decision
functions

LIME: Local Interpretable Model- Agnostic Explanations

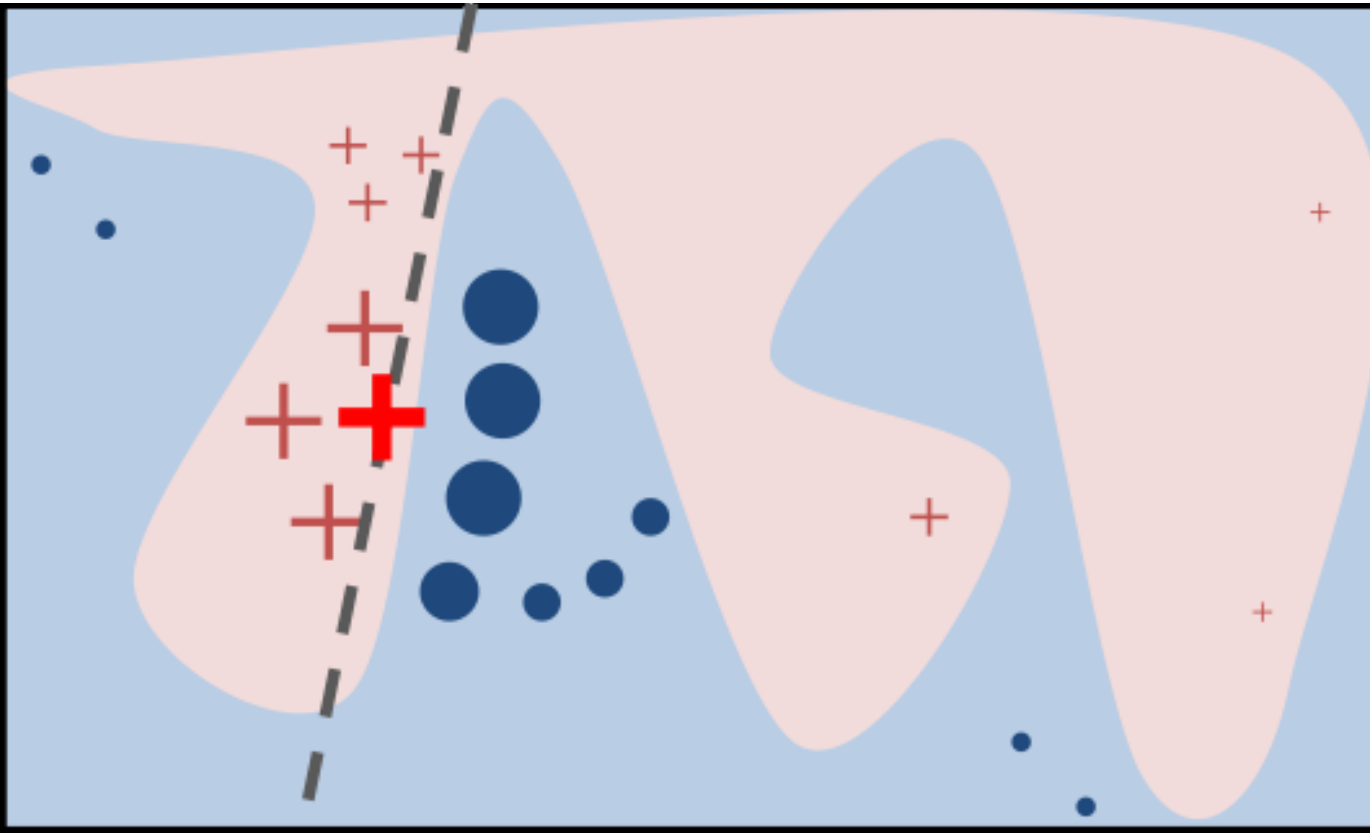
What LIME is, in simple terms?

A flu prediction example



The three highlighted symptoms may be a faithful approximation of the model for patients who look like the one being inspected, but they probably do not represent how the model behaves for **all** patients

Rationale behind LIME



- Blue/pink box is the decision boundary of the highly non-linear model
- The bold red cross is the instance being explained
- Other circles and crosses are sampled instances which are used to build a local model to explain the bold red cross (**x**). The instances are sampled both in the vicinity of **x** and far away from **x**.
- The dashed line is the learned local model.

General framework of LIME

Find interpretable
representation of x ,
denoted as x'



Sample around x' .
Recover the samples in
the original representation,
and get the labels



Fit a local model with
the recovered
samples and
corresponding labels



What is Interpretable representation x' ?

- it is something understandable by the human

Structured data

- Same as the feature used in the original representation

Text

- Different from the actual features used.
- E.g. binary vector indicating the presence or absence of a word, while word embedding is used as the feature

Image

- Different from the actual features used.
- E.g. binary vector indicating the presence or absence of a super-pixel, while tensor-based features are used.

Example of interpretable representation – super pixel for image



Original Image



Interpretable
Components

Sample around $\mathbf{x}'$, and recover in the original representation

Structured data

- numerical features:
 - scale the data
 - sampling from $N(0,1)$, multiplying by the std and adding back the mean
- categorical features: perturb by sampling according to the training distribution

Text & Image

- Generate a data set of perturbed instances by turning some of the interpretable components “off”.
- For each of the perturbed instances, feed into the original model, and get the class assignment (and associated probability).

Example of perturbed interpretable representations



Perturbed Instances	P(tree frog)
	<div><div></div>0.85</div>
	<div><div></div>0.00001</div>
	<div><div></div>0.52</div>

Fit a local model

- Now, fit a GLM $g(z') = \omega_g \cdot z'$ with K-LASSO
 - Loss function: $\mathcal{L}(f, g, \pi_x) = \sum_{(z, z')} \pi_x(z) (f(z) - g(z'))^2$
 - choose only **K** features, for better interpretability
 - LASSO regression (regularization with L1 penalty)



Pros

- Model-agnostic
 - almost all models can be explained, e.g. deep network, Random Forest,
- Fast
 - does not depend on the size of the dataset, but on time to compute $f(x)$ and on the number of samples N .



Cons

- Limited interpretability
 - “a model that predicts *sepia-toned* images to be *retro* cannot be explained by presence of absence of super pixels” – from the original paper
 - if the original model is highly non-linear even in the locality of the instance to be explained, there may not be a faithful local explanation
- Not stable explanation
 - For same instance, the explanation may be different every time a different set of samples are drawn randomly.

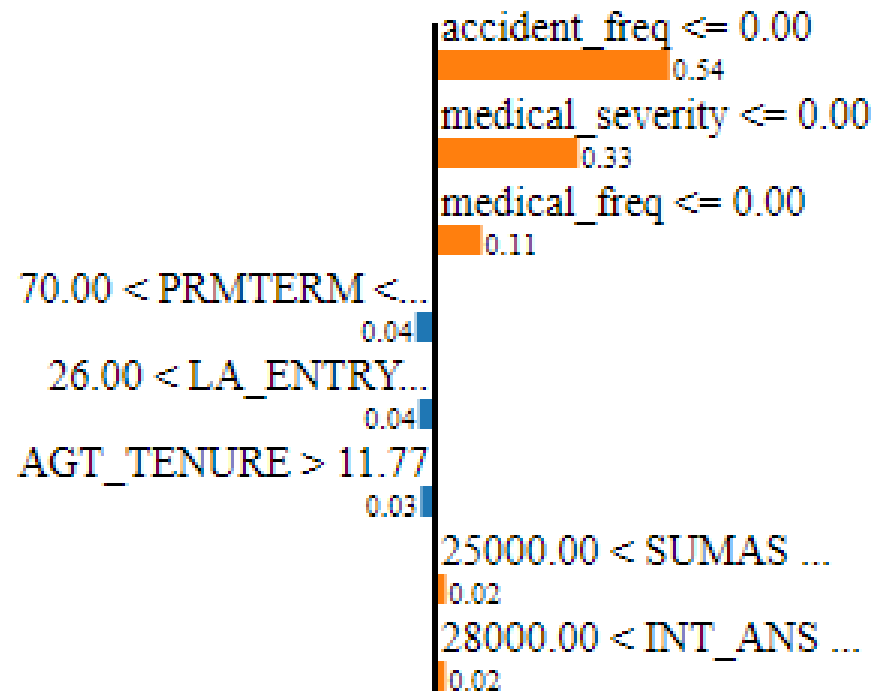
Application of LIME in a predictive UW model

Prediction probabilities



bad

good



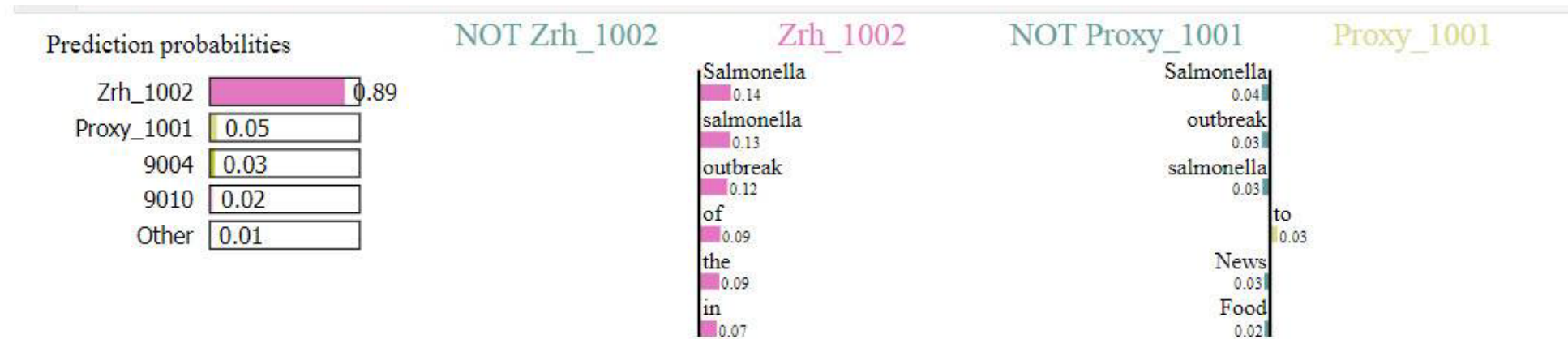
Feature Value

accident_freq	0.00
medical_severity	0.00
medical_freq	0.00
PRMTERM	73.00
LA_ENTRY_AGE	27.00
AGT_TENURE	18.69
SUMAS	30000.00
INT_ANS	32000.00

This applicant is considered a “good” risk because she/he has no accident claim and medical claim

Application of LIME in a text categorization model

Case that is classified as "**Biological food contamination**":



Text with highlighted words

Yum Brands Incorporated's Taco Bell has been linked to a **salmonella outbreak** which infected dozens **of** people **in** 10 states. **|P|** **|P|** The cluster **of** illness from **Salmonella** enteritidis infections is believed to have begun **in** mid-October 2011 and ended by **the** time CDC issued its final report. Illnesses were reported **in** Texas, Oklahoma, Kansas, Iowa, Michigan, Missouri, Nebraska, New Mexico, Ohio and Tennessee. **|P|** **|P|** CDC said 31 percent **of** patients were hospitalized. No deaths were reported. **|P|** **|P|** The **outbreak** had an air **of** mystery about it because CDC's final report said it was linked to a Mexican-style fast-food chain identified only as 'Restaurant A.' **|P|** **|P|** Food Safety News was **the** first to identify Taco Bell as Restaurant A, citing a document from **the** Oklahoma State Department **of** Health's Acute Disease Service.

Take-away

1



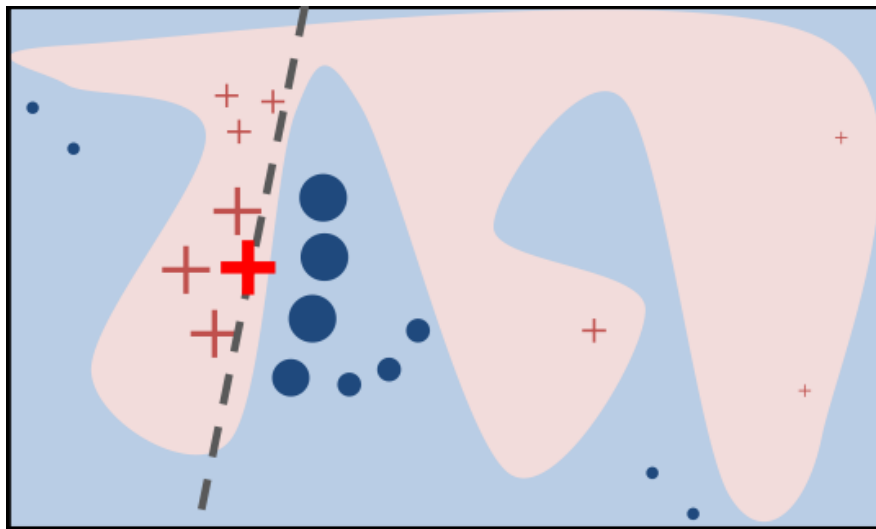
or



or



2



Can I ask
you something?

NO.

