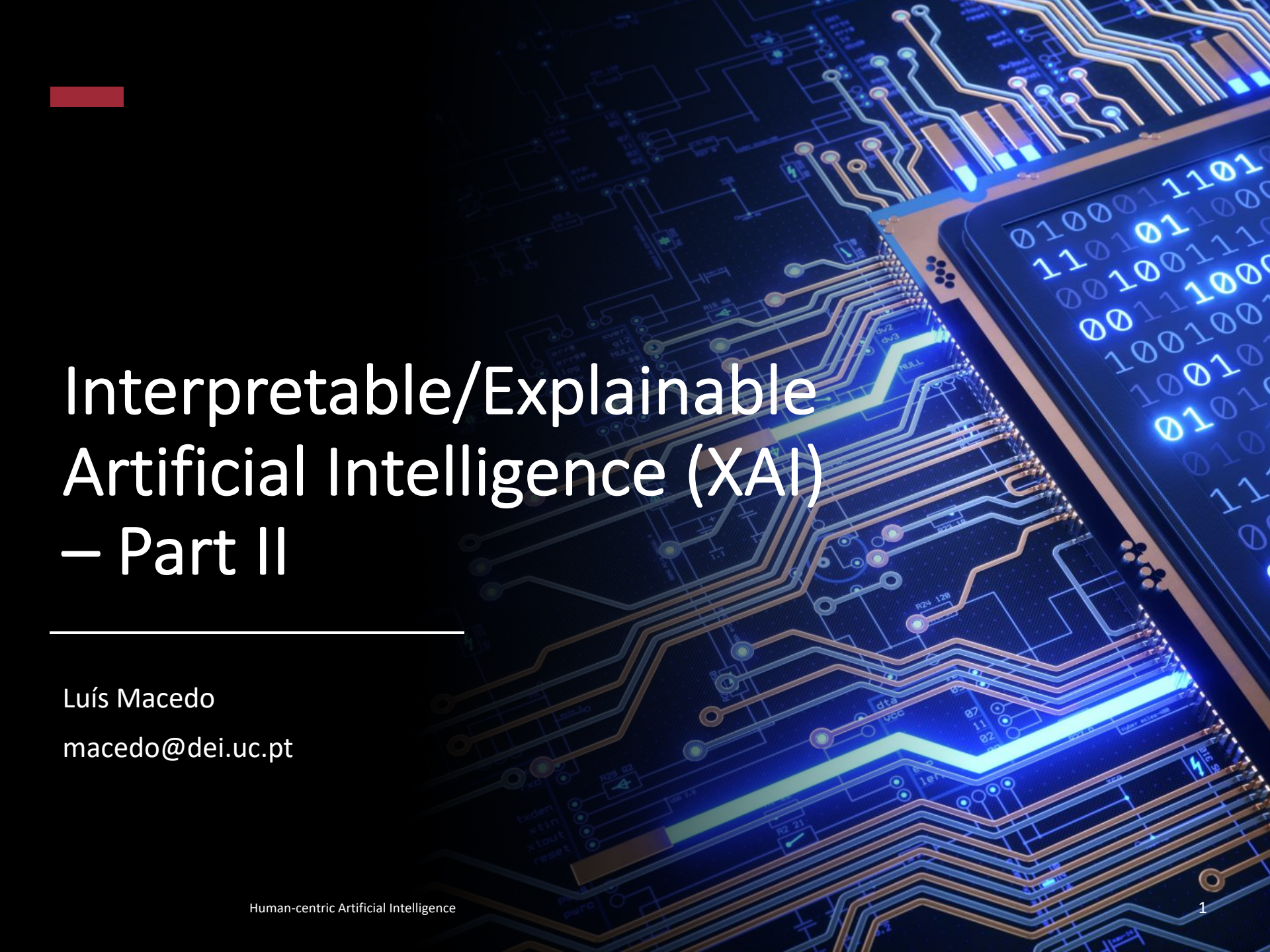
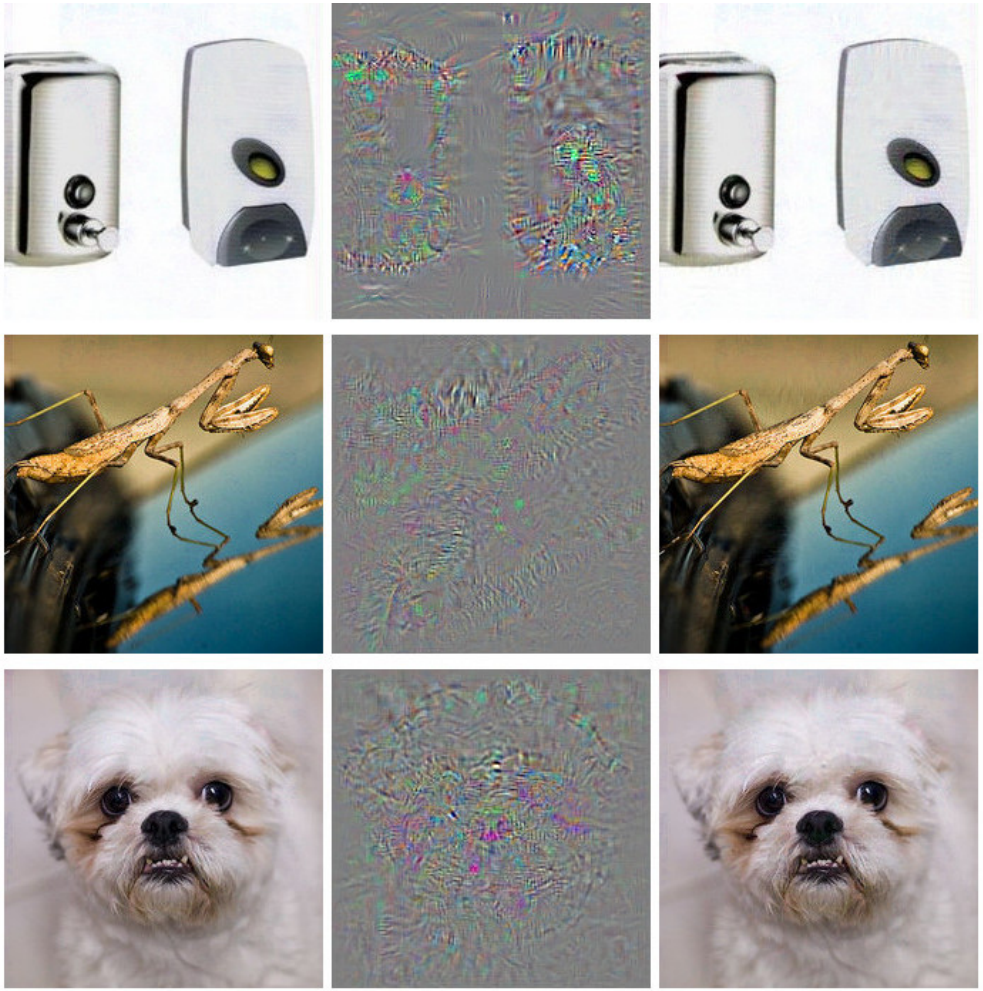


Interpretable/Explainable Artificial Intelligence (XAI) – Part II

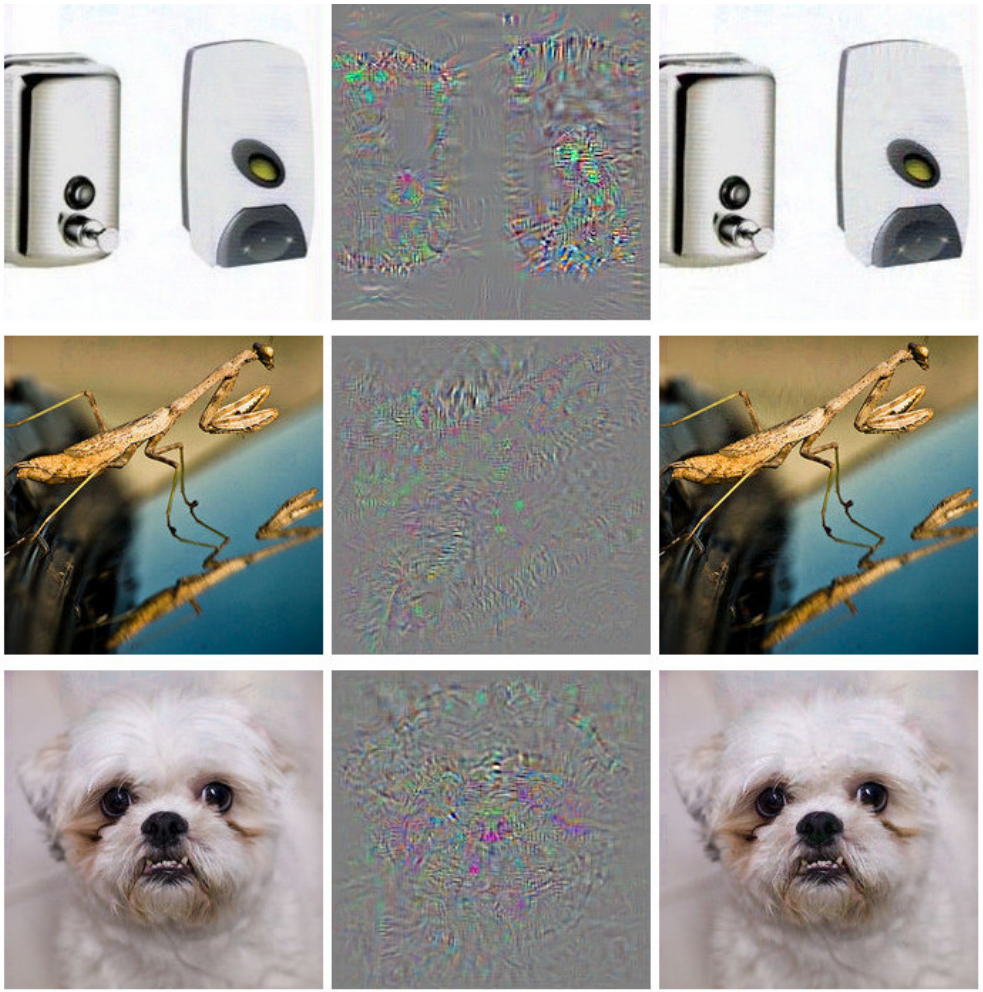
Luís Macedo

macedo@dei.uc.pt

- What are the differences between the pictures in the left and in the right?



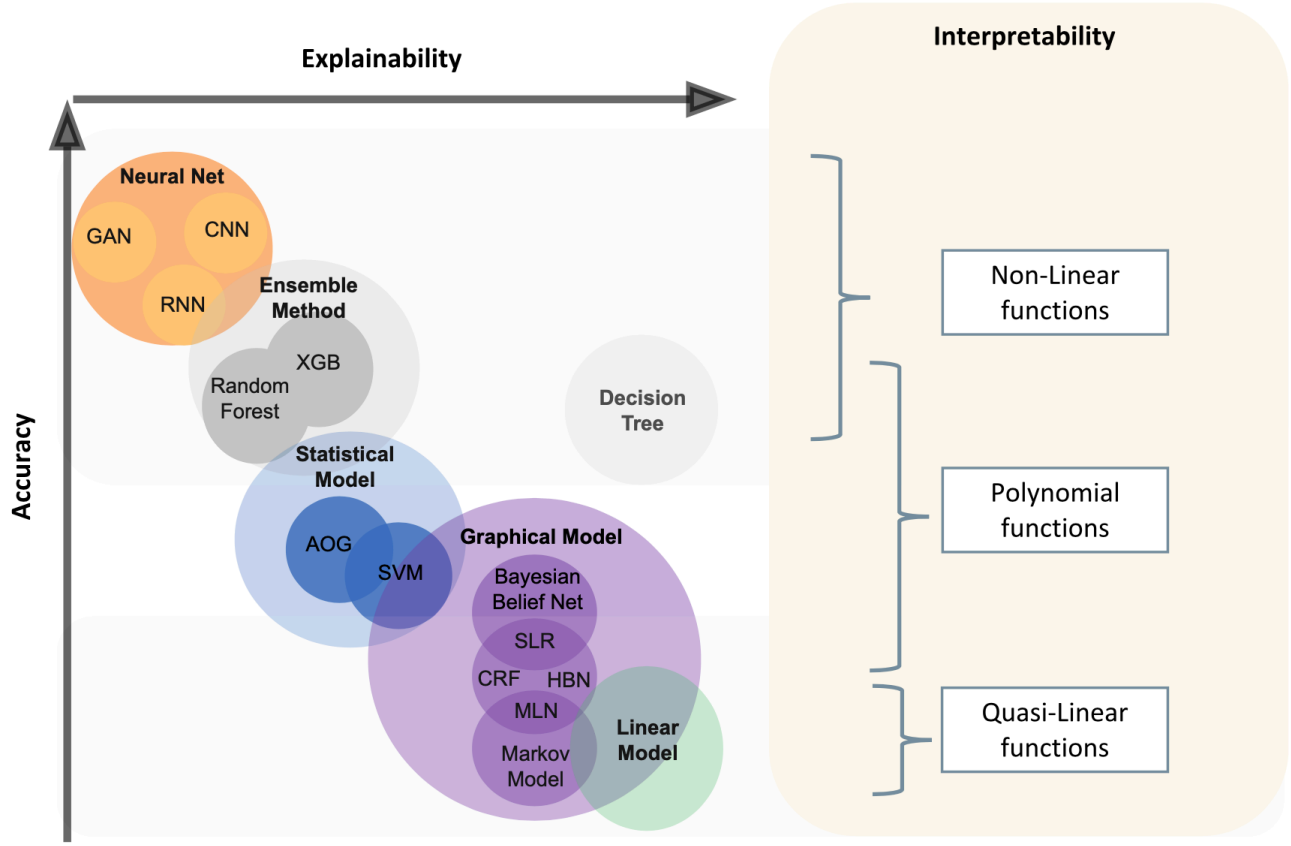
- According to a Deep Neural Net, those in the right are classified as ostrich!!!



A. Interpretable Models

A. Interpretable Models:
   a) linear regression
   b) logistic regression
   c) linear regression extensions
   d) decision trees
   e) decision rules and the RuleFit algorithm
   f) Graphical Models (Bayesian Networks, including Naive Bayes)

A. Interpretable Models – ML models: Explainability vs. Accuracy



A. Interpretable Models – Good explanations

- **Explanations are:**
  - Contrastive (Lipton 1990):
    - Counterfactual cases, i.e. "How would the prediction have been if input X had been different?"
    - Humans do not want a complete explanation for a prediction, but want to compare what the differences were to another instance's prediction (can be an artificial one)
  - Personalized:
    - Depend on the context, including to whom they are delivered
  - Truthful:
    - They should be true in similar circumstances
  - General:
    - Implementation: Generality can easily be measured by the feature's support, which is the number of instances to which the explanation applies divided by the total number of instances.
  - ….??
  - Cognitively and motivationally relevant:
    - Explanations should add information to the recipient agent, and this information should be goal-directed, i.e., relevant for the problem at hand.
  - ??

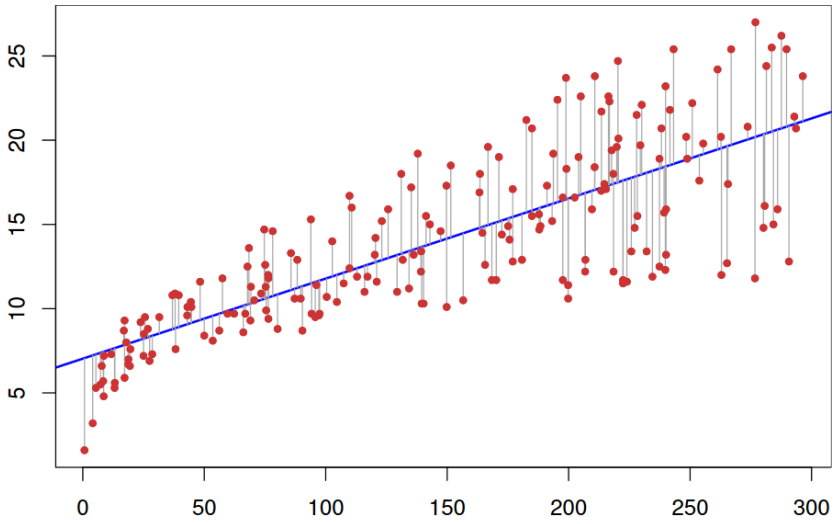
Interpretation of Linear Regression models

A. Interpretable Models – Linear Regression

- Linear model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y^{(i)} - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)} \right) \right)^2$$



$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

A. Interpretable Models – Linear Regression

- Important aspects for interpreting a linear regression model and that are related with weights :
  - Numerical feature (e.g. height):
    - Increasing the numerical feature by one unit changes the estimated outcome by its weight
    - i.e., an increase of feature x_k by one unit increases the prediction for y by β_k units when all other feature values remain fixed.
  - Binary feature: Changing the feature from the reference category to the other category changes the estimated outcome by the feature's weight
  - Categorical feature with multiple categories:
    - similar to the previous, assuming that:
      - each category has its own binary column. For a categorical feature with L categories, you only need L-1 columns, because the L-th column would have redundant information (e.g. when columns 1 to L-1 all have value 0 for one instance, we know that the categorical feature of this instance takes on category L)
    - i.e., changing feature x_k from the reference category to the other category increases the prediction for y by β_k when all other features remain fixed.
  - Intercept β_0: The intercept is the feature weight for the "constant feature", which is always 1 for all instances
    - Interpretation: For an instance with all numerical feature values at zero and the categorical feature values at the reference categories, the model prediction is the intercept weight.

A. Interpretable Models – Linear Regression

- Important measurements for the interpretation of linear regression:
  - R-squared measurement: tells how much of the total variance of the target outcome is explained by the model
  - The higher R-squared, the better the model explains the data

$$R^2 = 1 - SSE/SST$$

$$SSE = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

$$SST = \sum_{i=1}^{n} (y^{(i)} - \bar{y})^2$$

    - SSE tells you how much variance remains after fitting the linear model, which is measured by the squared differences between the predicted and actual target values.
    - SST is the total variance of the target outcome.
    - R-squared tells you how much of your variance can be explained by the linear model. R-squared ranges between 0, for models where the model does not explain the data at all, and 1, for models that explain all of the variance in your data.
  - It is not meaningful to interpret a model with very low R-squared, because such a model basically does not explain much of the variance. Any interpretation of the weights would not be meaningful.

A. Interpretable Models – Linear Regression

- Feature importance in linear regression models:
  - Weights are estimated; there is an error (SE)
  - feature importance is measured by the absolute value of its t-statistic (estimated weight scaled with its standard error):

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- The importance of a feature increases with increasing weight.
- The more variance the estimated weight has (i.e., the less certain we are about the correct value), the less important the feature is

A. Interpretable Models – Linear Regression

- Bike data set

instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	01/01/2011	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
2	02/01/2011	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
3	03/01/2011	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
4	04/01/2011	1	0	1	0	2	1	1	0.2	0.212122	0.590435	0.160296	108	1454	1562
5	05/01/2011	1	0	1	0	3	1	1	0.226957	0.22927	0.436957	0.1869	82	1518	1600
6	06/01/2011	1	0	1	0	4	1	1	0.204348	0.233209	0.518261	0.0895652	88	1518	1606
7	07/01/2011	1	0	1	0	5	1	2	0.196522	0.208839	0.498696	0.168726	148	1362	1510
8	08/01/2011	1	0	1	0	6	0	2	0.165	0.162254	0.535833	0.266804	68	891	959
9	09/01/2011	1	0	1	0	0	0	1	0.138333	0.116175	0.434167	0.36195	54	768	822
10	10/01/2011	1	0	1	0	1	1	1	0.150833	0.150888	0.482917	0.223267	41	1280	1321
11	11/01/2011	1	0	1	0	2	1	2	0.169091	0.191464	0.686364	0.122132	43	1220	1263
12	12/01/2011	1	0	1	0	3	1	1	0.172727	0.160473	0.599545	0.304627	25	1137	1162
13	13/01/2011	1	0	1	0	4	1	1	0.165	0.150883	0.470417	0.301	38	1368	1406
14	14/01/2011	1	0	1	0	5	1	1	0.16087	0.188413	0.537826	0.126548	54	1367	1421
15	15/01/2011	1	0	1	0	6	0	2	0.233333	0.248112	0.49875	0.157963	222	1026	1248
16	16/01/2011	1	0	1	0	0	0	1	0.231667	0.234217	0.48375	0.188433	251	953	1204
17	17/01/2011	1	0	1	1	1	0	2	0.175833	0.176771	0.5375	0.194017	117	883	1000
18	18/01/2011	1	0	1	0	2	1	2	0.216667	0.232333	0.861667	0.146775	9	674	683
19	19/01/2011	1	0	1	0	3	1	2	0.292174	0.298422	0.741739	0.208317	78	1572	1650
20	20/01/2011	1	0	1	0	4	1	2	0.261667	0.25505	0.538333	0.195904	83	1844	1927
21	21/01/2011	1	0	1	0	5	1	1	0.1775	0.157833	0.457083	0.353242	75	1468	1543
22	22/01/2011	1	0	1	0	6	0	1	0.0591304	0.0790696	0.4	0.17197	93	888	981
23	23/01/2011	1	0	1	0	0	0	1	0.0965217	0.0988391	0.436522	0.2466	150	836	986
24	24/01/2011	1	0	1	0	1	1	1	0.0973913	0.11793	0.491739	0.15833	86	1330	1416
25	25/01/2011	1	0	1	0	2	1	2	0.223478	0.234526	0.616957	0.129796	186	1799	1985
26	26/01/2011	1	0	1	0	3	1	3	0.2175	0.2036	0.8625	0.29385	34	472	506
27	27/01/2011	1	0	1	0	4	1	1	0.195	0.2197	0.6875	0.113837	15	416	431
28	28/01/2011	1	0	1	0	5	1	2	0.203478	0.223317	0.793043	0.1233	38	1129	1167
29	29/01/2011	1	0	1	0	6	0	1	0.196522	0.212126	0.651739	0.145365	123	975	1098
30	30/01/2011	1	0	1	0	0	0	1	0.216522	0.250322	0.722174	0.0739826	140	956	1096
31	31/01/2011	1	0	1	0	1	1	2	0.180833	0.18625	0.60375	0.187192	42	1459	1501

A. Interpretable Models – Linear Regression

- Interpretation, from a weight table (weight and variance estimates), of the importance of a feature (example):

	Weight	SE	|t|
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

A. Interpretable Models – Linear Regression

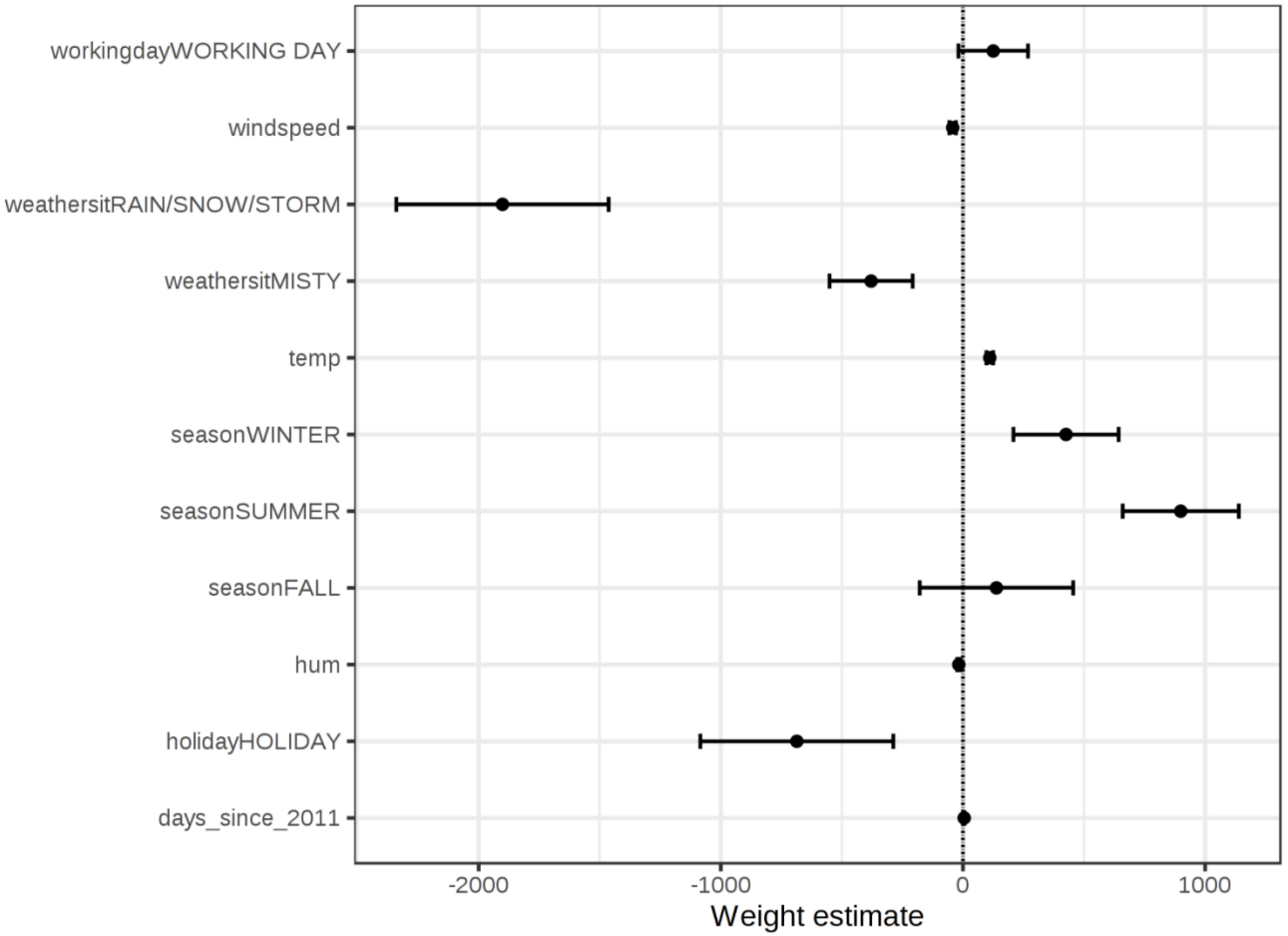
- Importance of a feature (example):
  - Interpretation of the feature “temperature” and its weight: an increase of the temperature by 1 degree Celsius increases the predicted number of bicycles by 110.7, when all other features remain fixed.
  - Interpretation of the categorical feature “weathersit”: the estimated number of bicycles is -1901.5 lower when it is raining, snowing or stormy, compared to good weather, assuming that all other features remain fixed. When the weather is misty, the predicted number of bicycles is -379.4 lower compared to good weather, given all other features remain the same.

A. Interpretable Models – Linear Regression

- Problem of interpretation in linear models: these models assume independence of the features

A. Interpretable Models – Linear Regression

- Interpretation from weight plots (weights are displayed as points and the 95% confidence intervals as lines):



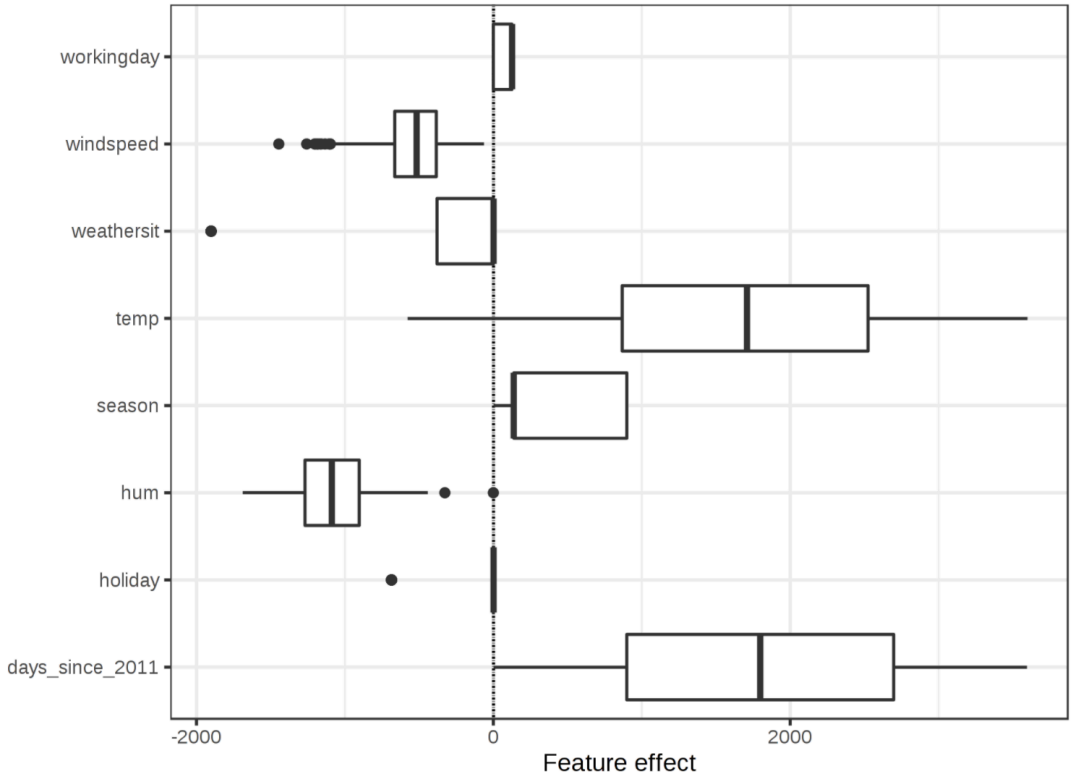
A. Interpretable Models – Linear Regression

- Importance of a feature (example):
  - rainy/snowy/stormy weather has a strong negative effect on the predicted number of bikes
  - the weight of the working day feature is close to zero and zero is included in the 95% interval, which means that the effect is not statistically significant
  - some confidence intervals are very short and the estimates are close to zero, yet the feature effects were statistically significant. Temperature is one such candidate

- Problem of weight plots:
  - features are measured on different scales:
    - E.g.: the estimated weight for the weather reflects the difference between good and rainy/stormy/snowy weather; for temperature it only reflects an increase of 1 degree Celsius.
  - Solution:
    - making the estimated weights more comparable by scaling the features (zero mean and standard deviation of one) before fitting the linear model

A. Interpretable Models – Linear Regression

- Interpretation from effect plots (answers the question: how much the combination of weight and feature contributes to the predictions in the data?):
  1. Calculation the effects: $\text{effect}_j^{(i)} = w_j x_j^{(i)}$
  2. Visualization with boxplots



A. Interpretable Models – Linear Regression

- A box in a boxplot contains the effect range for half of the data (25% to 75% effect quantiles)
- The vertical line in the box is the median effect, i.e. 50% of the instances have a lower and the other half a higher effect on the prediction
- The horizontal lines extend to ±1.5IQR/√n±1.5IQR/n, with IQR being the inter quartile range (75% quantile minus 25% quantile)
- The dots are outliers
- The categorical feature effects can be summarized in a single boxplot, compared to the weight plot, where each category has its own row

A. Interpretable Models – Linear Regression

- Interpretation (example):
  - The largest contributions to the expected number of rented bicycles comes from the temperature feature and the days feature, which captures the trend of bike rentals over time
  - The temperature has a broad range of how much it contributes to the prediction
  - The day trend feature goes from zero to large positive contributions, because the first day in the dataset (01.01.2011) has a very small trend effect and the estimated weight for this feature is positive (4.93)
  - This means that the effect increases with each day and is highest for the last day in the dataset (31.12.2012)
  - Note that for effects with a negative weight, the instances with a positive effect are those that have a negative feature value. For example, days with a high negative effect of windspeed are the ones with high wind speeds.

A. Interpretable Models – Linear Regression

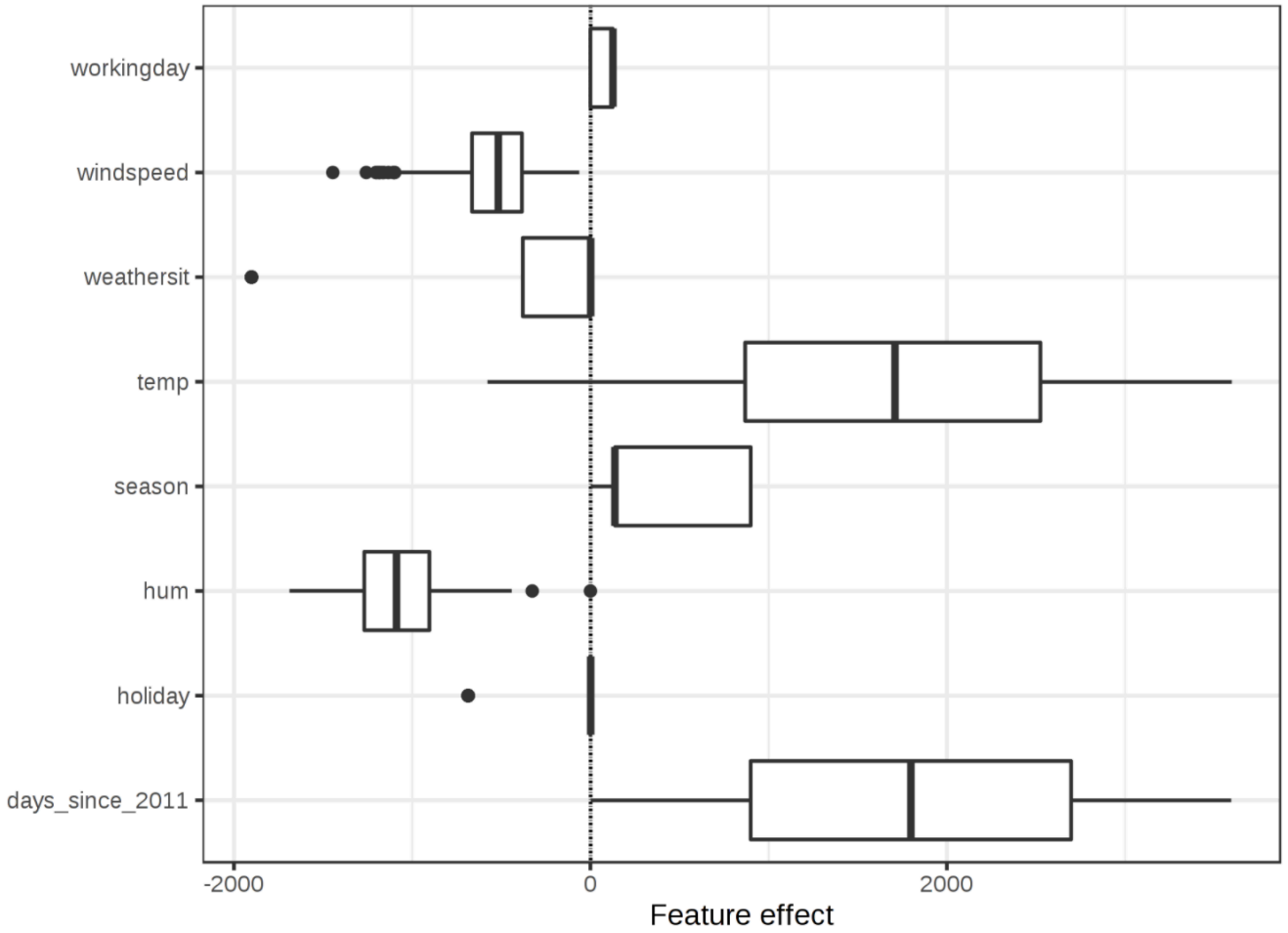
- Explaining individual predictions: How much has each feature of an instance contributed to the prediction?

Feature	Value
season	SPRING
yr	2011
mnth	JAN
holiday	NO HOLIDAY
weekday	THU
workingday	WORKING DAY
weathersit	GOOD
temp	1.604356
hum	51.8261
windspeed	6.000868
cnt	1606
days_since_2011	5

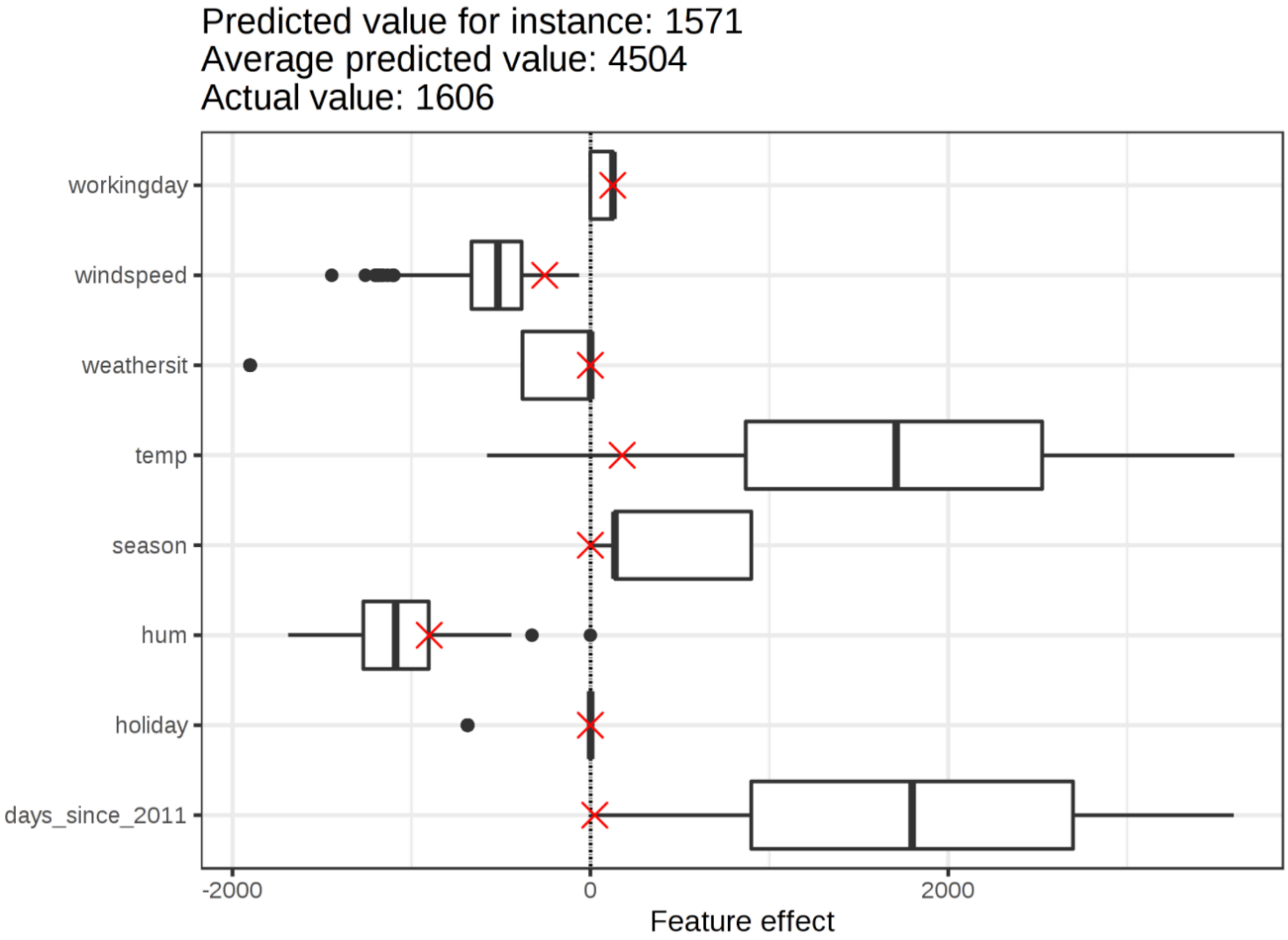
A. Interpretable Models – Linear Regression

- To obtain the feature effects of this instance, we have to multiply its feature values by the corresponding weights from the linear regression model:
  - For the value "WORKING DAY" of feature "workingday", the effect is, 124.9
  - For a temperature of 1.6 degrees Celsius, the effect is 177.6
- Add these individual effects as crosses to the effect plot, which shows us the distribution of the effects in the data
- This allows us to compare the individual effects with the distribution of effects in the data

A. Interpretable Models – Linear Regression



A. Interpretable Models – Linear Regression



A. Interpretable Models – Linear Regression

- To obtain the feature effects of this instance, we have to multiply its feature values by the corresponding weights from the linear regression model:
  - For the value "WORKING DAY" of feature "workingday", the effect is, 124.9
  - For a temperature of 1.6 degrees Celsius, the effect is 177.6
- Add these individual effects as crosses to the effect plot, which shows us the distribution of the effects in the data
- This allows us to compare the individual effects with the distribution of effects in the data

A. Interpretable Models – Linear Regression

- The average of the predictions for the training data instances = 4504
- The prediction of the 6-th instance is small =1571
- The effect plot reveals the reason why:
  - the boxplots show the distributions of the effects for all instances of the dataset
  - the crosses show the effects for the 6-th instance
    - The 6-th instance has a low temperature effect because on this day the temperature was 2 degrees, which is low compared to most other days (remember that the weight of the temperature feature is positive)
    - Also, the effect of the trend feature "days_since_2011" is small compared to the other data instances because this instance is from early 2011 (5 days) and the trend feature also has a positive weight

Interpretation of Graphical Models — Bayesian Networkss (including Naïve Bayes)

A. Interpretable Models – Graphical Models

A. Interpretable Models – Graphical Models

A. Interpretable Models – Graphical Models