Lecture 21: Transparency

Mark Hasegawa-Johnson

3/2023

Outline

- GDPR right to explanation
- Explanations by analyzing the processing of a neural network
- Decision-making algorithms that are explainable by design

GDPR Right to Explanation

The European Union's "General Data Protection Regulation" (GDPR) Article 15 specifies that:

> The data subject shall have the right to obtain … confirmation as to whether personal data concerning him or her are being processed, … access to the personal data, … the existence of automated decision-making, and … meaningful information about the logic involved.

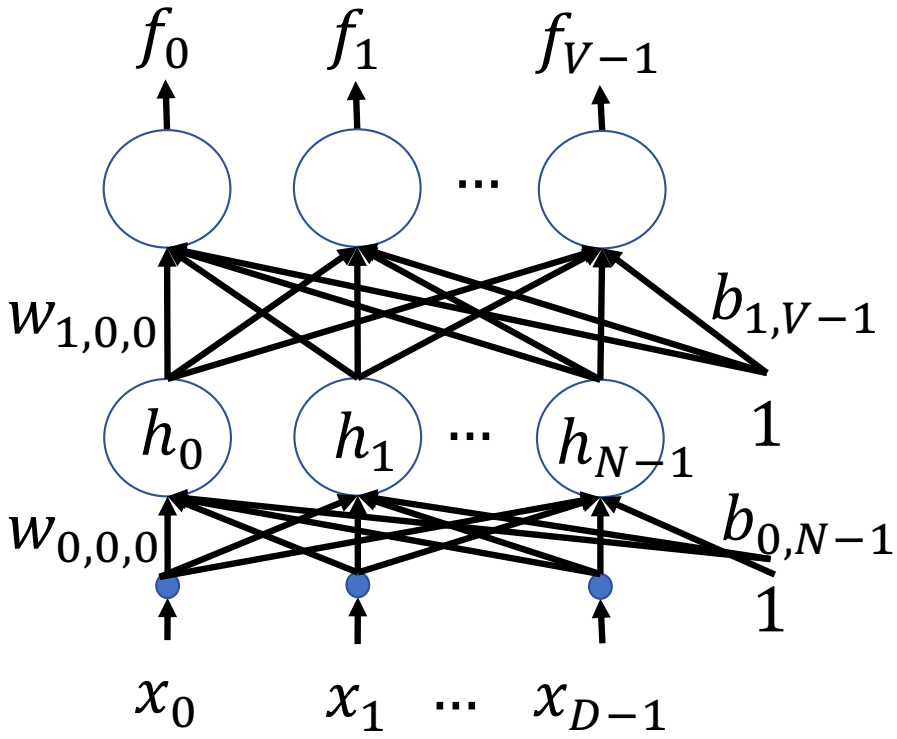
Outline

- GDPR right to explanation
- Explanations by analyzing the processing of a neural network
- Decision-making algorithms that are explainable by design

Multi-layer neural network: Each layer is a matrix multiplication followed by a nonlinearity

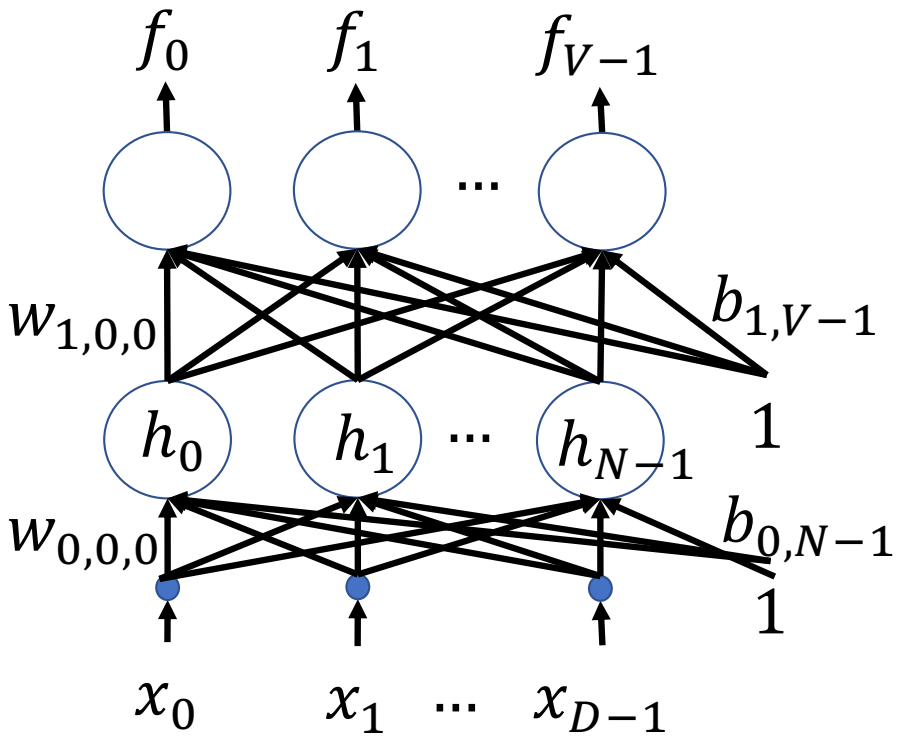
$$f = \text{softmax}(w_1 @ h + b_1)$$

$$h = ReLU(w_0 @ x + b_0)$$



Explanations by analyzing the processing of a neural network

- x =binary indicator vector specifying the courses you've taken
- f =probability vector, f_k =probability that you should go into career k
- Suppose the neural net tells you that you should become a tiktok influencer. You might want to know why the neural net made that decision.

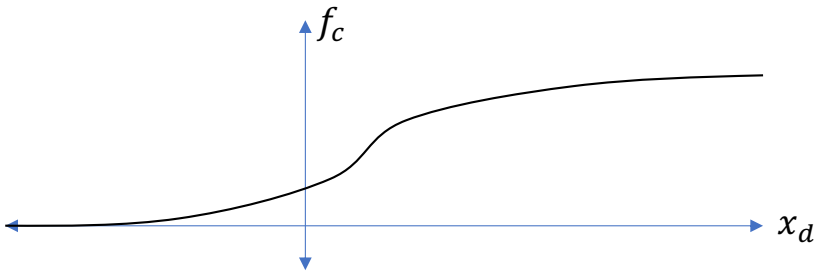


Gradient-based relevance

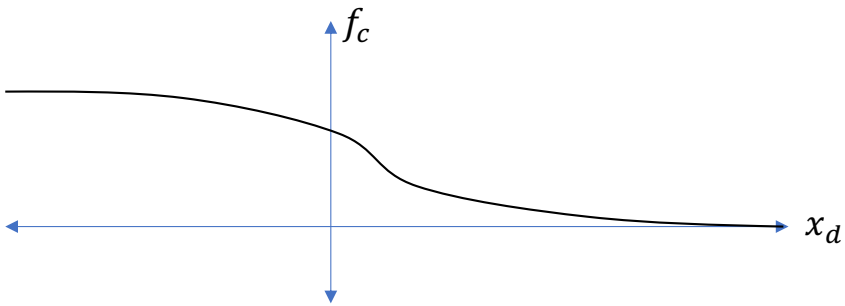
To what extent did feature x_d contribute to the network's decision?

- Is the slope positive or negative?
- Is x_d positive or negative?

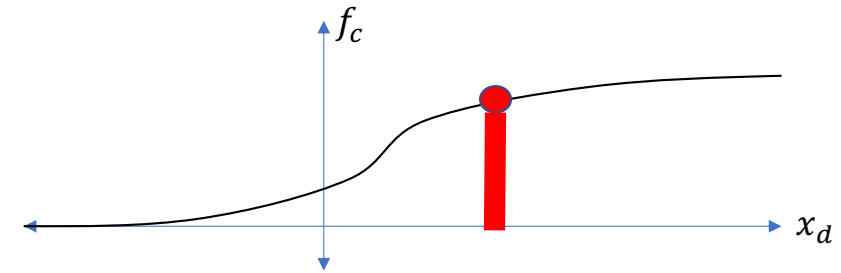
Example of an output f_c that gets larger in response to **increases** of the input x_d

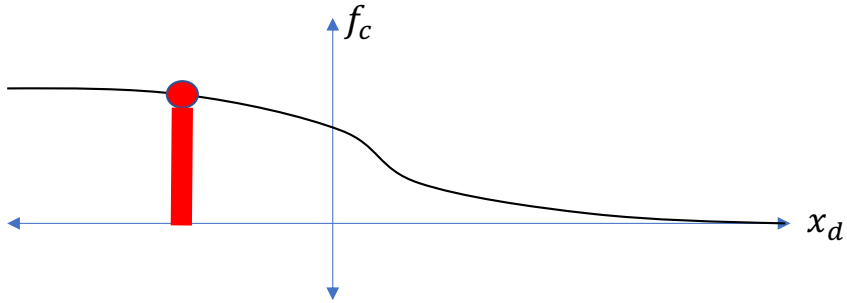


Example of an output f_c that gets larger in response to **decreases** of the input x_d

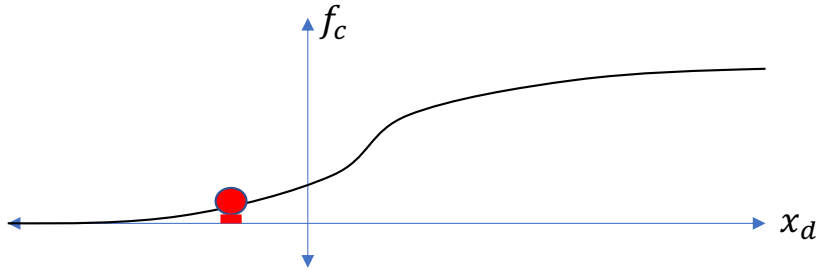


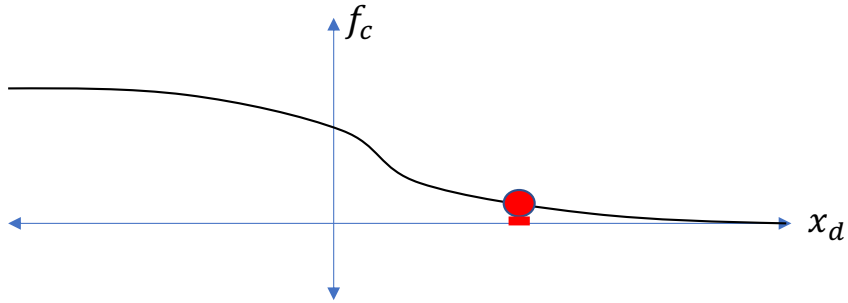
These are input features x_d that caused f_c to be **larger** than it would have been if $x_d = 0$:





These are input features x_d that caused f_c to be **smaller** than it would have been if $x_d = 0$:





Relevance scoring in neural networks

- If $\text{sign}\left(\frac{\partial f_c}{\partial x_d} \cdot x_d\right) = \text{sign}(f_c)$, then feature x_d has **supporting** relevance to the neural net's output decision f_c
- If $\frac{\partial f_c}{\partial x_d} \cdot x_d = 0$, then feature x_d has **zero** relevance to the neural net's output decision f_c
- If $\text{sign}\left(\frac{\partial f_c}{\partial x_d} \cdot x_d\right) = -\text{sign}(f_c)$, then feature x_d has **opposing** relevance to the neural net's output decision f_c

Relevance of feature x_d to decision f_c:

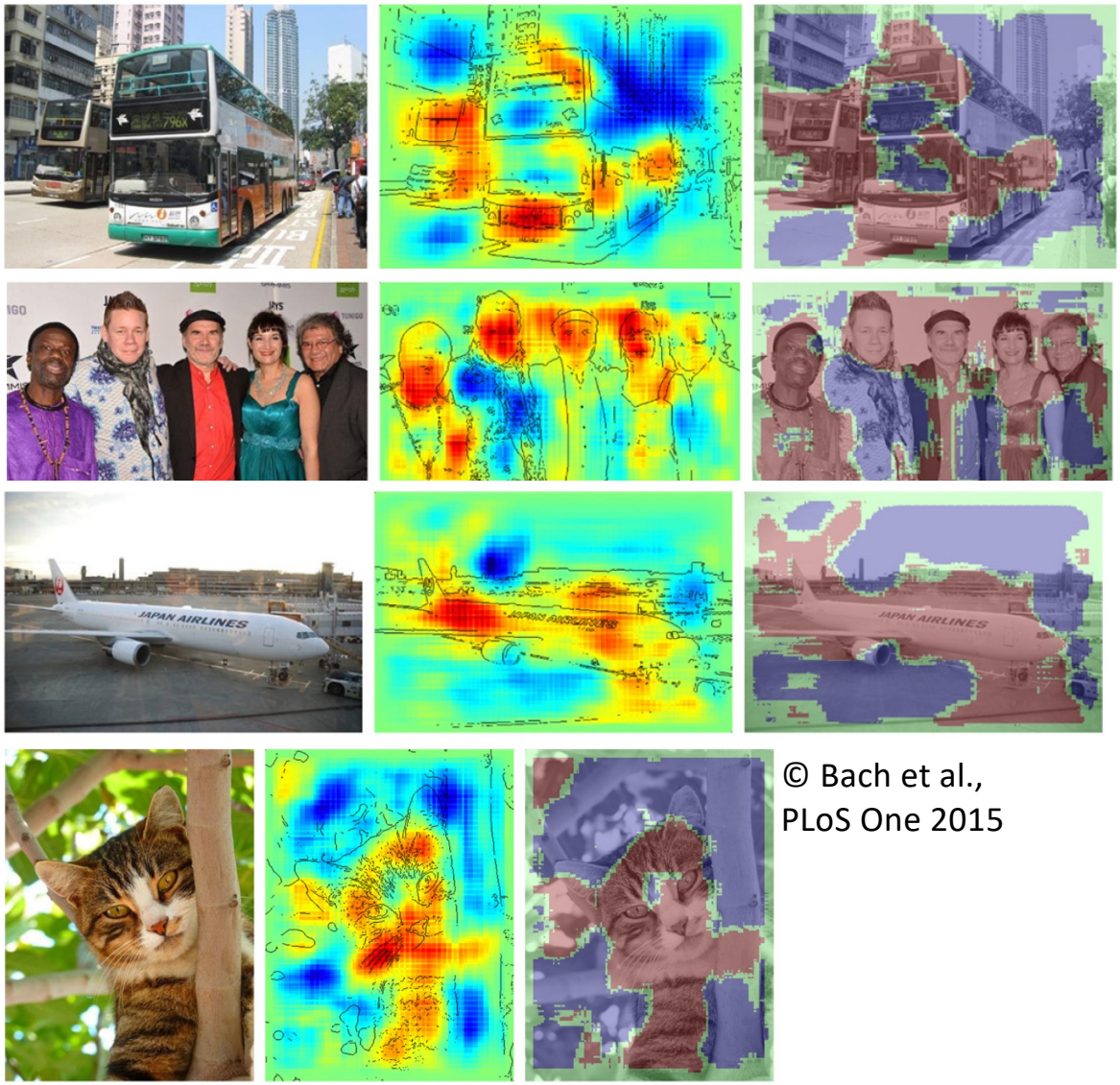
$$R_{c,d} = \frac{\frac{\partial f_c}{\partial x_d} \cdot x_d}{\sum_{d'} \frac{\partial f_c}{\partial x_{d'}} \cdot x_{d'}} \cdot f_c$$

The summation is usually done over a layer, so that $\sum_d R_{c,d} = f_c$ over each layer.

Layer-Wise Relevance Propagation

(Bach et al., 2015)

- In LRP, relevance is normalized then back-propagated, layer by layer. This causes the smoothness you see here.
- Positive relevance: red, Negative: blue
- 2nd image: scaled,
- 3rd image: binary



Quiz

Try the quiz!
https://us.prairielearn.com/pl/course_instance/129874/assessment/2335672

My example:

F(x)=w@x=(-48)+(-28)+(0)+(-24)+(5)=-95.

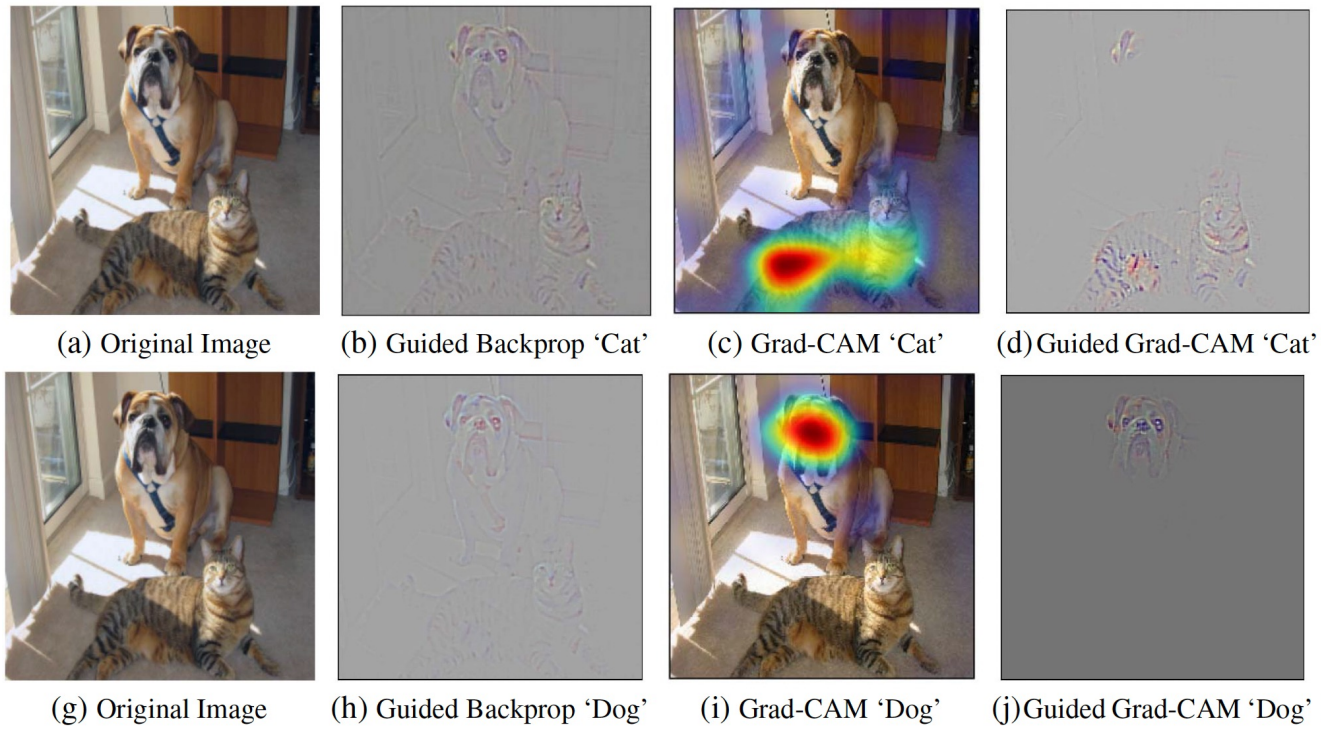
R = normalize((df/dx)*x)*f

= [(-48/-95), (-28/-95), (0/-95), (-24/-95), (5/-95)] * (-95)

Positive-only relevance scoring: Grad-CAM

Many relevance scoring systems keep only positive relevance, i.e.,

$$L_{c,d} = \frac{\max\left(0, \frac{\partial f_c}{\partial x_d} \cdot x_d\right)}{\sum_{d'} \max\left(0, \frac{\partial f_c}{\partial x_{d'}} \cdot x_{d'}\right)}$$



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog'

© Selvaraju et al., ICCV 2017

- Gradient-weighted Class Activation Mapping (Grad-CAM) pools these scores over collections of nodes
- Guided Grad-CAM multiplies the pooled scores times individual pixel scores

Advantages and disadvantages of relevance-based explainability

Advantage:

- Explains which input features caused the neural to make the decision it made

Disadvantage:

- Does not provide a logical reason why those particular features caused the neural net to make the decision it made
- There may not be any logical reason! The neural net is just a linear classifier of nonlinear combinations of features, it may not be logical.

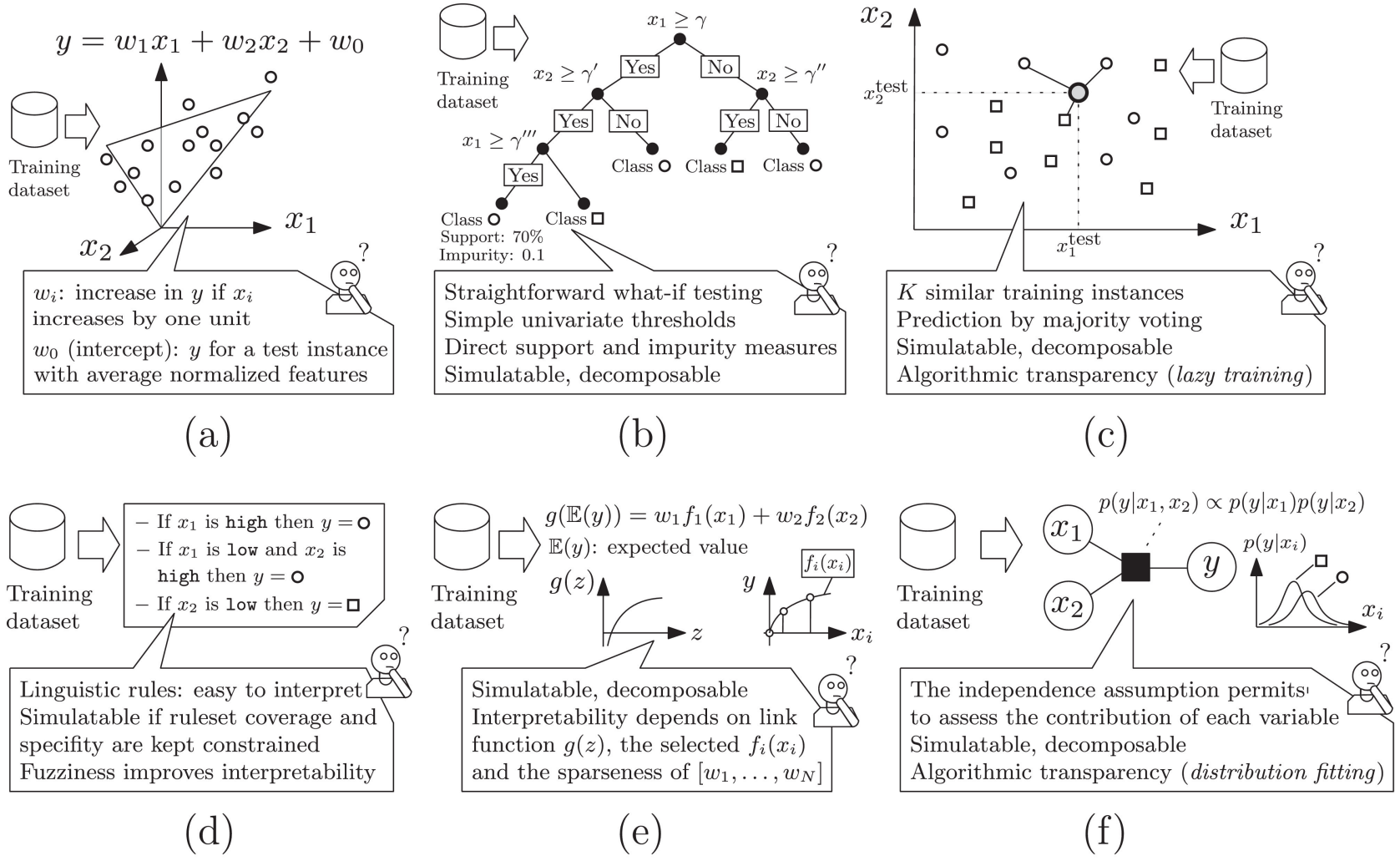
Outline

- GDPR right to explanation
- Explanations by analyzing the processing of a neural network
- Decision-making algorithms that are explainable by design

Decision-making algorithms that are explainable by design

Neural networks are not designed to be explainable. Other decision algorithms that are designed to be explainable include:

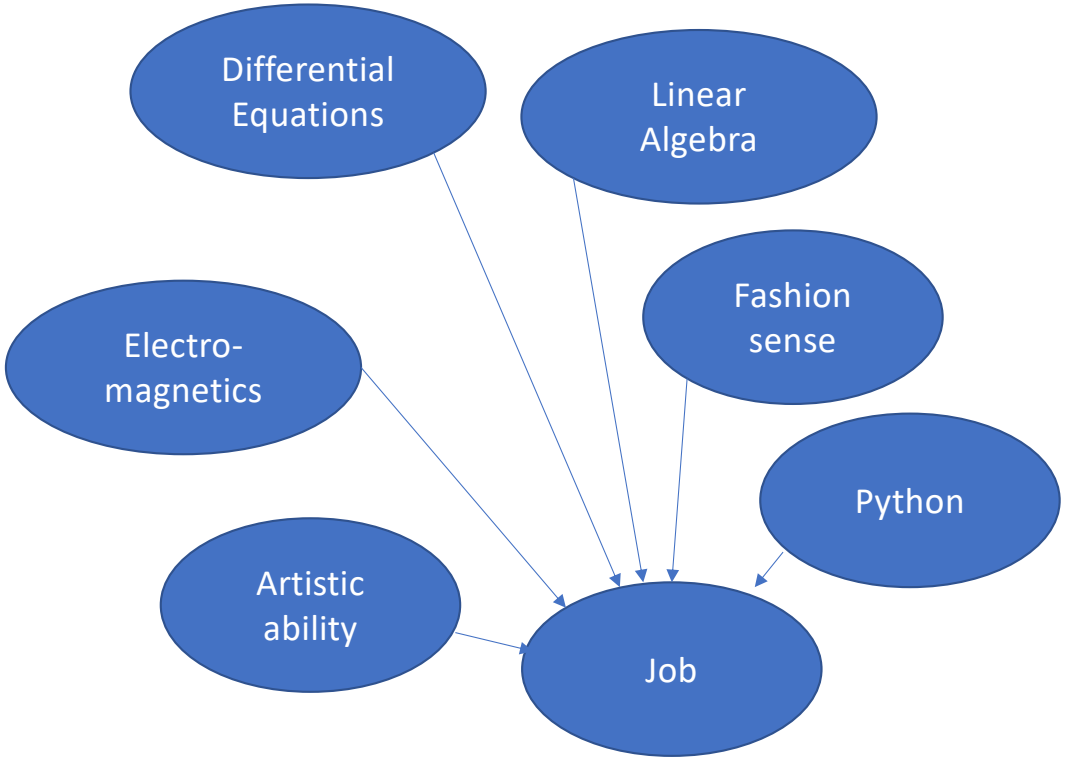
- Rule-based decision algorithms
- Decision trees
- Bayesian networks



(a) Linear Regression, (b) Decision tree, (c) KNN, (d) Rule-based, (e) Generalized additive models, (f) Bayesian networks. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Arrieta et al., 2020

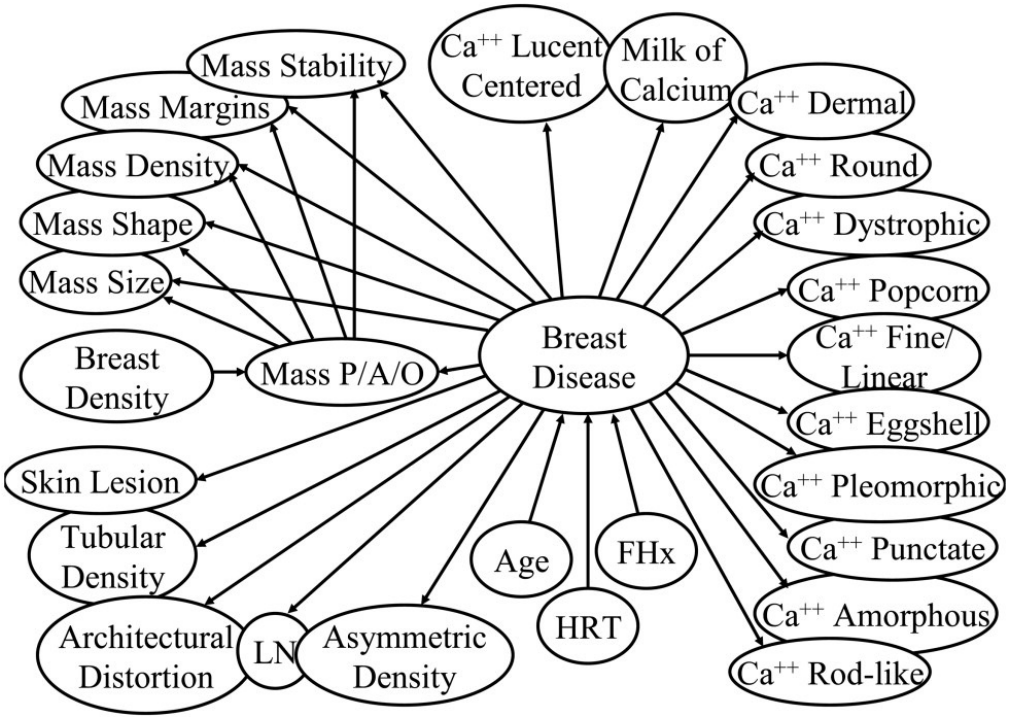
Bayesian networks for explainable AI

- Instead of training a neural net to determine what job a person should have,
- Use an examination to decide if they understand differential equations, then
- Use a Bayesian network to recommend jobs.



Bayesian networks for explainable AI

- People who need to know why they are making a recommendation (e.g., doctors) are much more likely to accept this approach because:
- Each of the classifications made by a neural network (e.g., “Mass Stability”) can be visually confirmed by the Radiologist
- The doctor can choose to ignore the final diagnosis (“Breast disease”) if she disagrees with the reasons



Elizabeth Burnside, “Bayesian networks: Computer-assisted diagnosis support in radiology,” 2005

Advantages and disadvantages of Bayes network explainability

Disadvantage:

- Forcing the decision to depend on a small number of other random variables may reduce accuracy of the decision

Advantage:

- Decision is explainable by design
- Some end users will completely ignore an unexplainable decision (e.g., doctors). For such end users, an explainable decision is the only alternative.

Summary

- GDPR right to explanation
  - Users have a right to know the logic used
  - If they don't understand or don't like the logic, they have a right to tell you not to use that decision
- Explanations by analyzing the processing of a neural network
  - Relevance = gradient times activation
  - Relevance-based methods tell the degree to which any given feature contributed to any given output decision
- Decision-making algorithms that are explainable by design
  - Bayesian networks have a small finite number of variables, which can make accuracy lower
  - Bayesian networks are explainable by design, so they can be used in applications where an unexplainable decision is useless