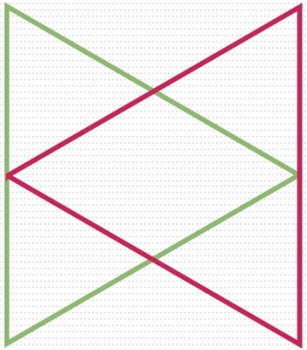
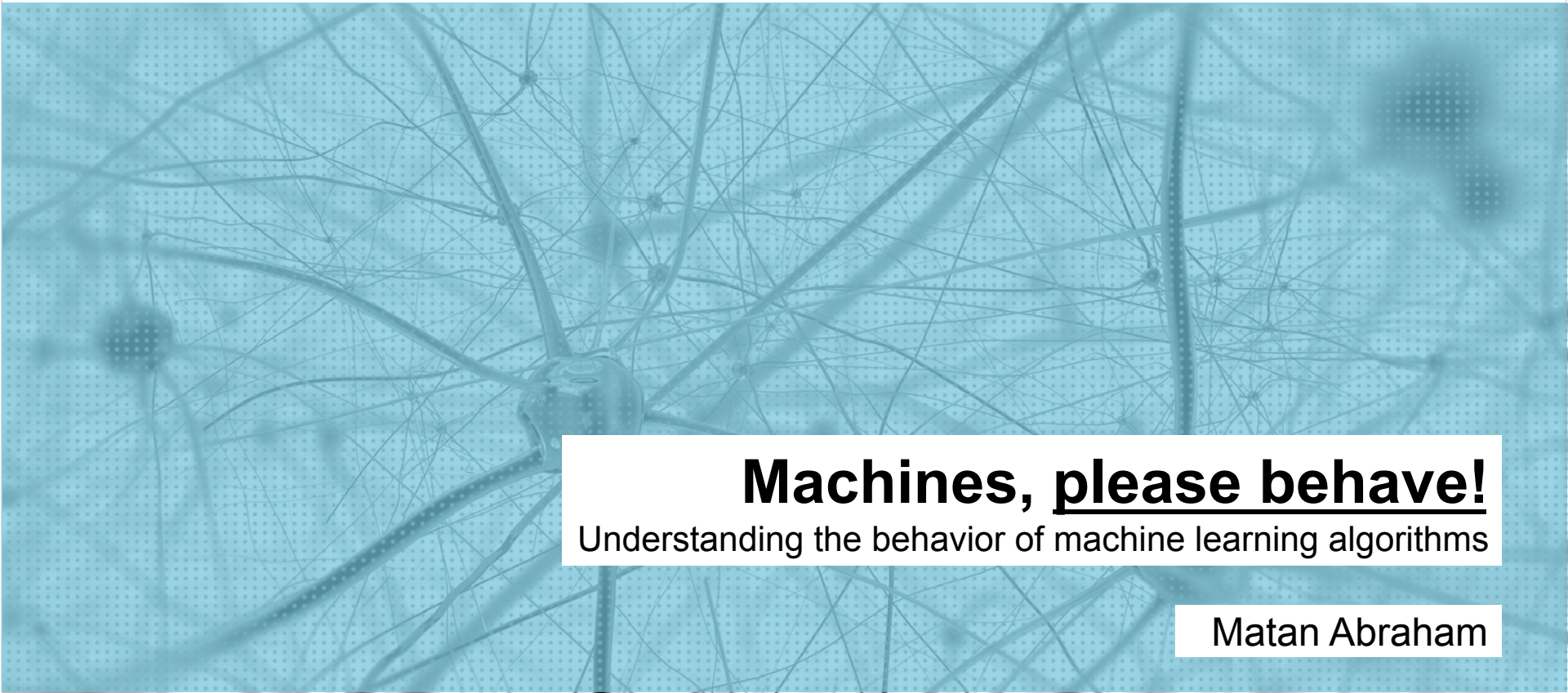


foresight

dialogues 2018



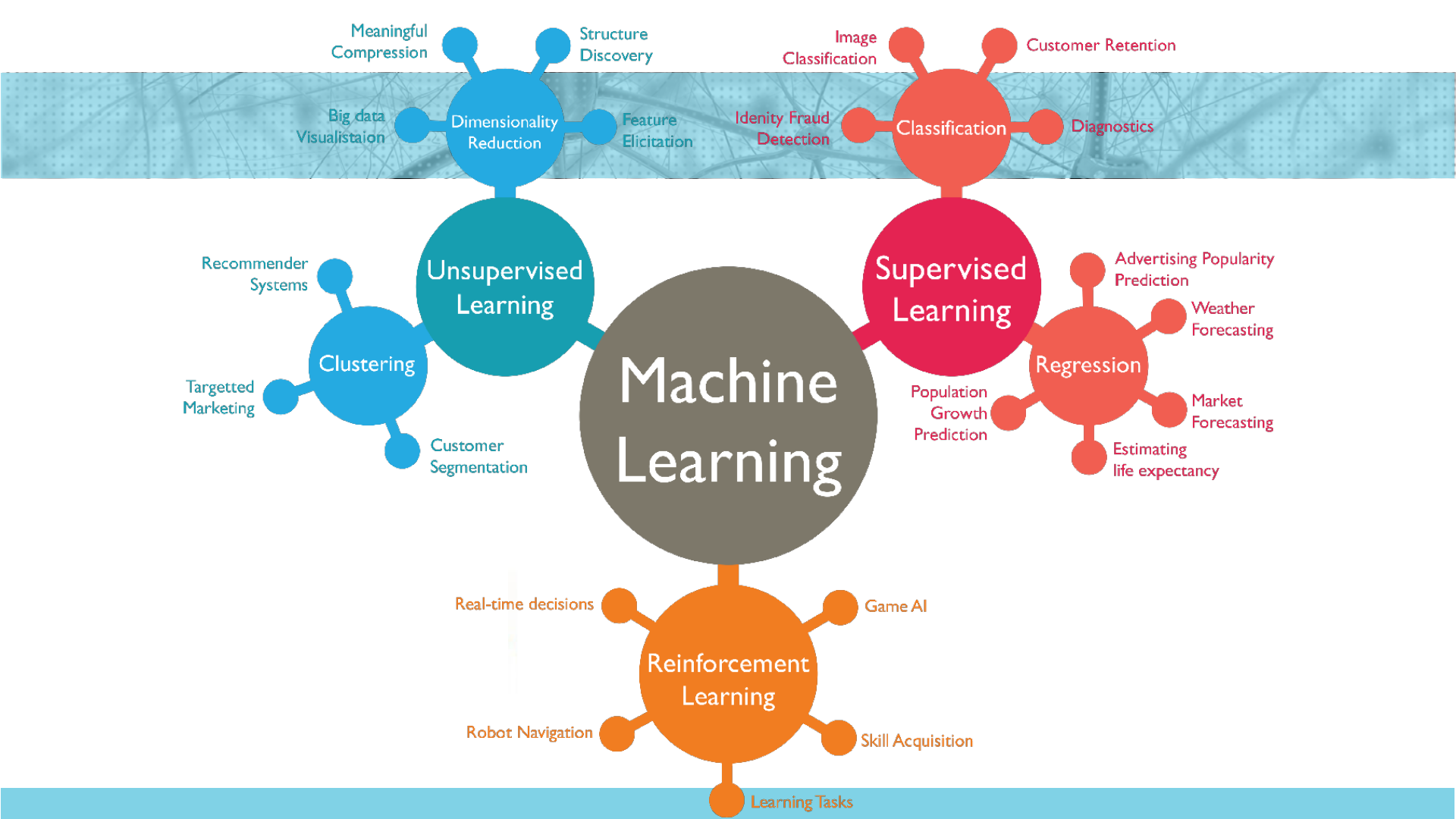
foresight dialogues 2018

insight

Machines, please behave!

Understanding the behavior of machine learning algorithms

Matan Abraham

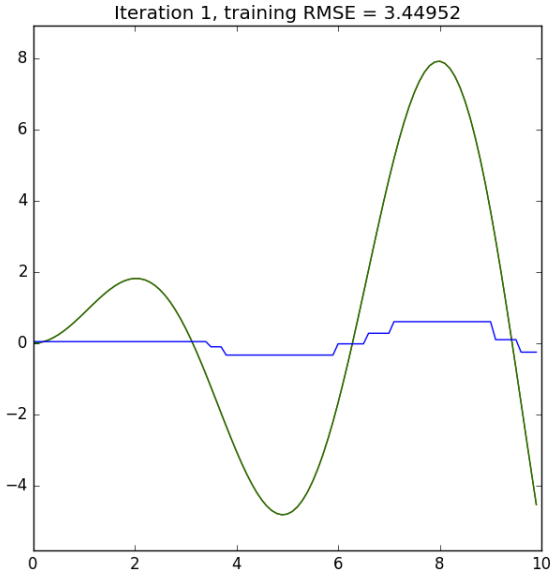

Meaningful Compression
Structure Discovery
Big data Visualistaion
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Idenity Fraud Detection
Classification
Diagnostics
Unsupervised Learning
Recommender Systems
Clustering
Targetted Marketing
Customer Segmentation
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Regression
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Machine Learning
Real-time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks

Supervised learning.

- Function that maps inputs to outputs based on ‘learning’ from labelled training data of examples of input-output pairs.

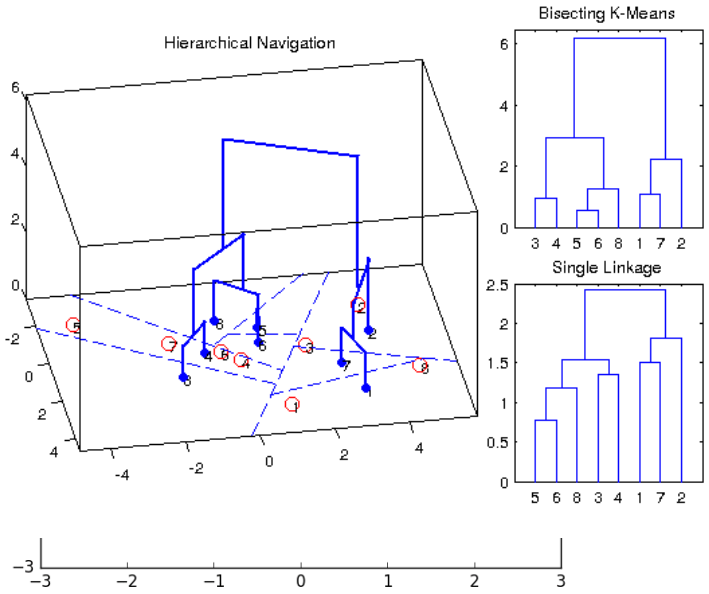
- This is called ‘training’ the model.

- Common examples:
  - Regression
  - Decision tress
  - Neural networks

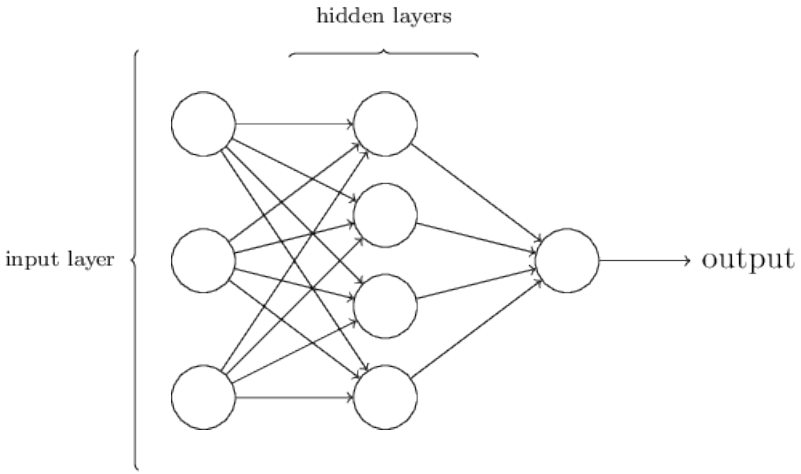


Unsupervised learning.

- No explicit target outputs or environmental evaluations associated with inputs.
- Patterns are modeled using measures of similarity defined by metrics such as Euclidean and probabilistic distance.
- Common examples:
  - Clustering
  - Anomaly detection



Neural networks.



"non-linear relationships"

Neural network function:

Inputs: 3

Outputs: 1 (the prediction)

Parameters: 16 weights/biases

"the algorithm"

Cost function:

Inputs: 16

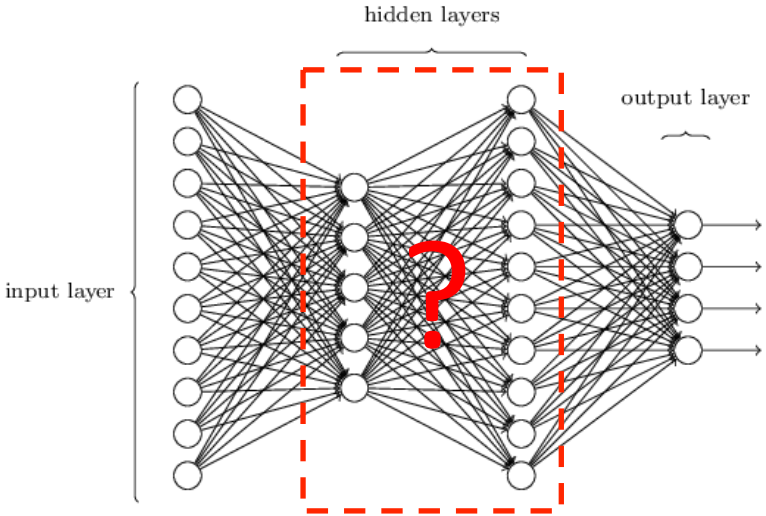
Outputs: 1 (the cost)

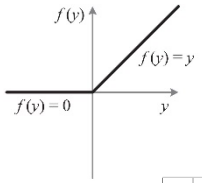
Parameters: All our training examples

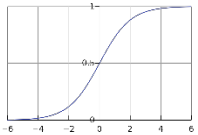
"the measurement of performance"

foresight dialogues 2018

Deep learning.







Each node in the hidden layer has an activation function.

It is not easily understood what is happening in hidden layer.

This layer defines the complexity of the neural network.

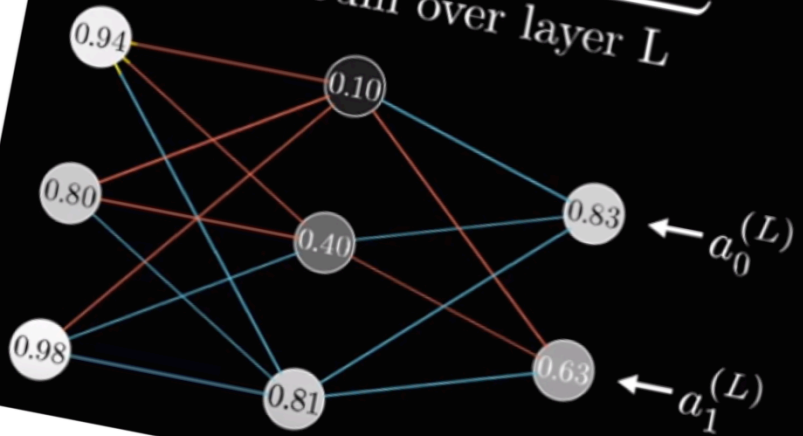
Inputs: 10
Outputs: 4
Parameters: 159

$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \underbrace{\sum_{j=0}^{n_L-1} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_0}{\partial a_j^{(L)}}}_{\text{Sum over layer L}}$$

$$z_j^{(L)} = \cdots + w_{jk}^{(L)} a_k^{(L-1)} + \cdots$$

$$a_j^{(L)} = \sigma(z_j^{(L)})$$

$$C_0 = \sum_{j=0}^{n_L-1} (a_j^{(L)} - y_j)^2$$

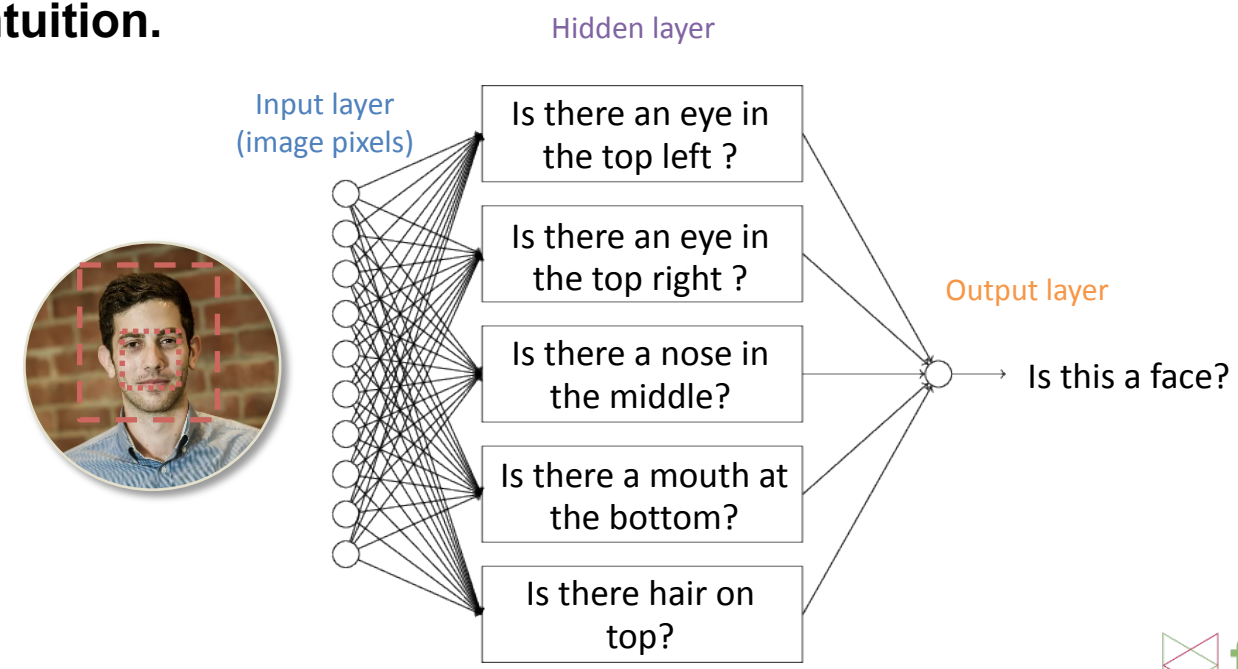


$\leftarrow a_0^{(L)}$

$\leftarrow a_1^{(L)}$

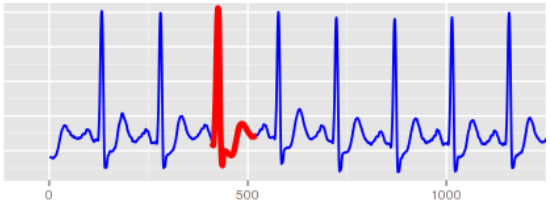
C_0

Neural network intuition.

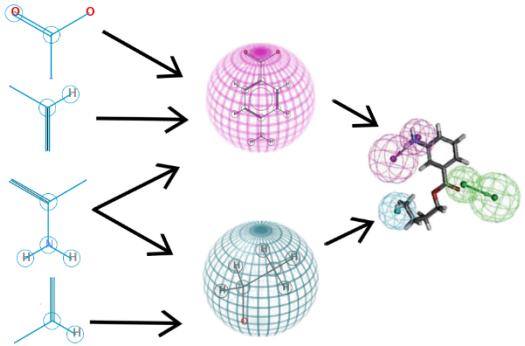


Applications of neural networks.

Image analysis and interpretation

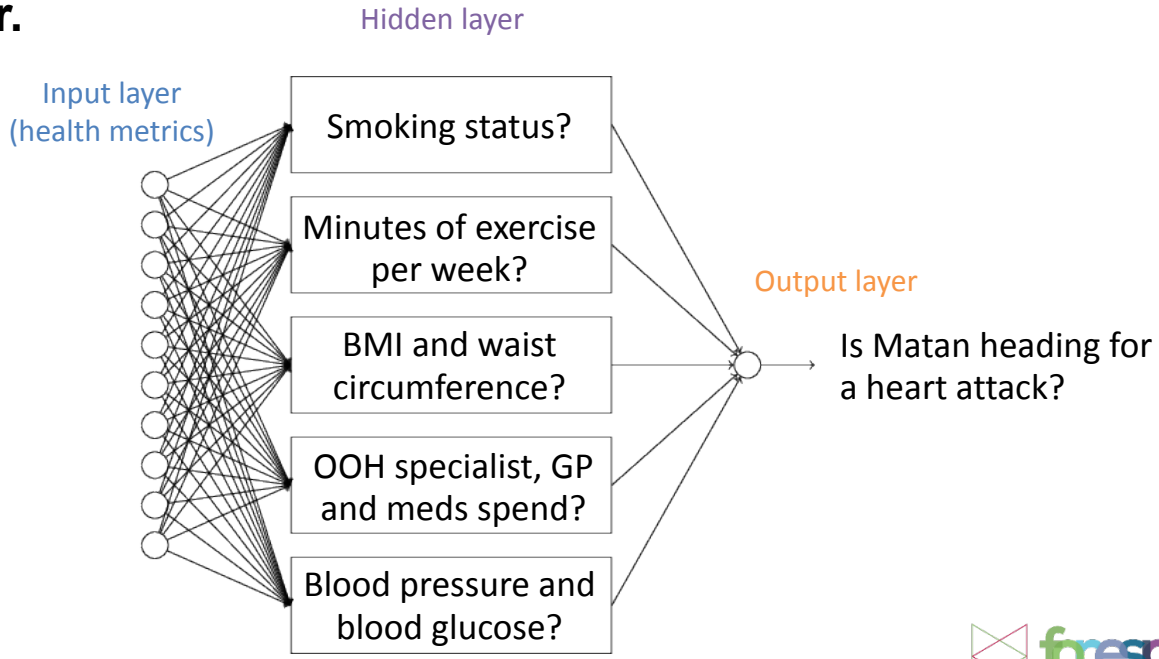


Signal analysis and interpretation



Drug development

Neural network –
Insight health-risk predictor.



You don’t know what you don’t know.

“Black box” machine learning limits visibility into the decision-making process.

Studying the algorithms ensures:

- Confidence predictions are as accurate as expected to be.
- Results mean what users think they mean.
- Predictions don’t rely on unwanted information.

“Information theory of deep learning”

“Explainable AI”

Aligning with reality.

Training requires a lot of data.

Training on a particular dataset which is applied to make predictions on future data – the scope of which cannot be predicted.

"Transfer learning"

"Lean data learning"

"Synthesizing new data through simulations"

Baked-in bias.

Algorithms can learn biases that are undesirable as predictions.

There need to be extra constraints imposed on an algorithm beyond just "accuracy".

This becomes more challenging the more complex the algorithm.

False assumptions.

It is imperative that the assumptions made in designing training the algorithms are appropriate.

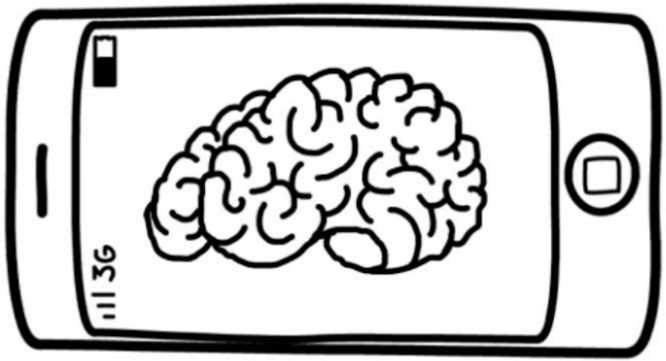
Examples:

- Missing data
- Data collection
- Distributions
- Parameterization

The self-fulfilling prophecy.

“If the algorithm tries to exploit what it learned, without leaving any room for exploration, it will keep reinforcing what it already knows and it will not learn new things, eventually becoming useless.”

Artificial Intelligence vs. Human Intelligence



foresight dialogues 2018

insight