Analyzing Unstructured Data

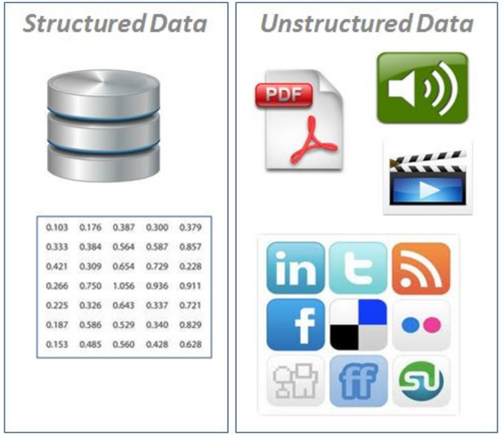
March 17th 2018

Agenda

1. Unstructured vs. Structured Data
2. Conventional Text Analysis
3. Basics of neural network
4. Deep Learning
5. CNNs - Image Recognition
6. Drug Discovery

Source: https://www.laserfiche.com/ecmblog/4-ways-to-manage-unstructured-data-with-ecm/

Structured vs. Unstructured Data



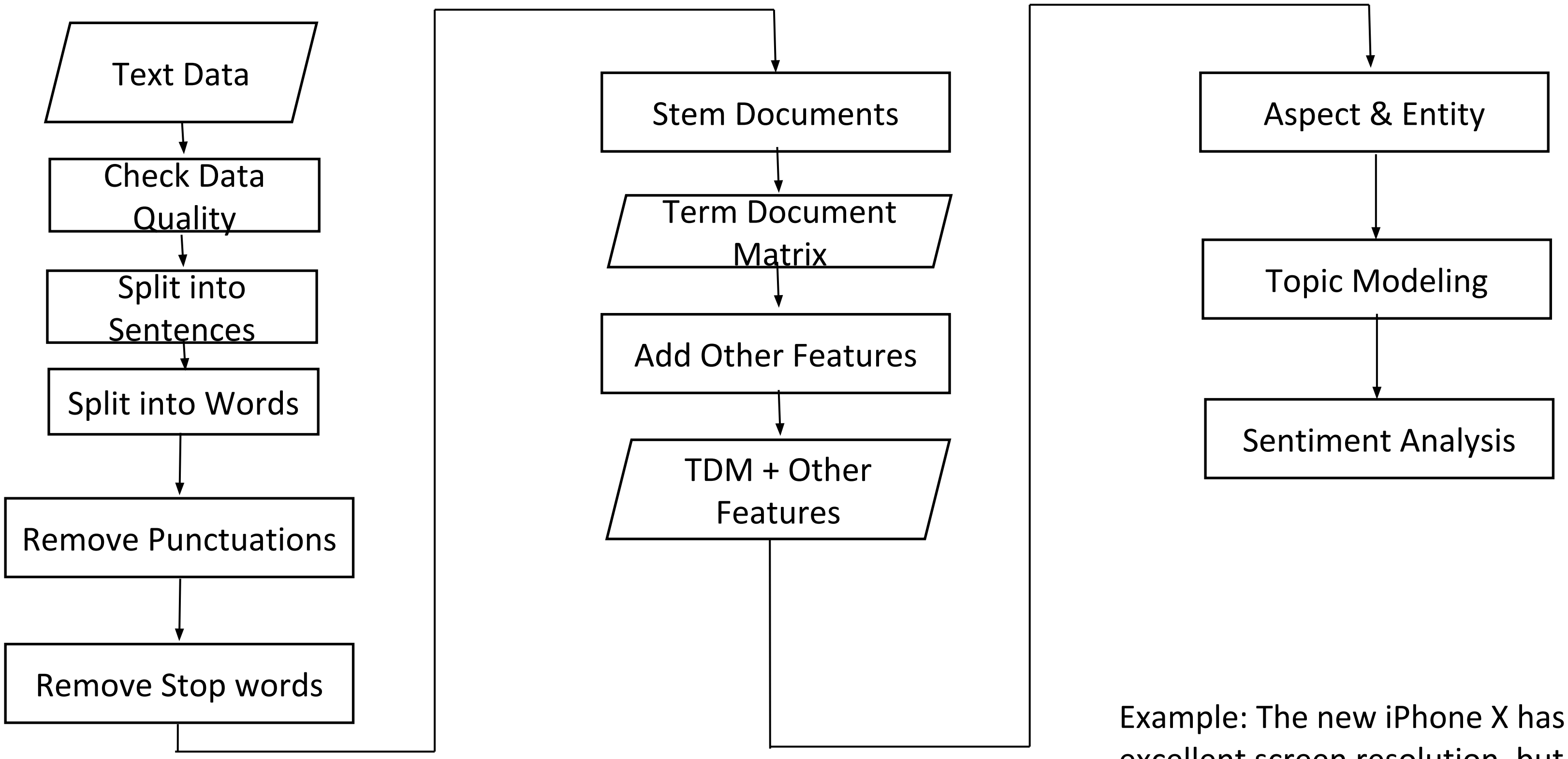
0.103	0.176	0.387	0.300	0.379
0.333	0.384	0.564	0.587	0.857
0.421	0.309	0.654	0.729	0.228
0.266	0.750	1.056	0.936	0.911
0.225	0.326	0.643	0.337	0.721
0.187	0.586	0.529	0.340	0.829
0.153	0.485	0.560	0.428	0.628

- Structured Data
  - Databases
- Unstructured Data
  - Social Media
  - Audio
  - Video

- 80% of business related information originates in unstructured format, primarily text

Source: https://www.laserfiche.com/ecmblog/4-ways-to-manage-unstructured-data-with-ecm/

Text Analysis - Overview



Text Data

Check Data Quality

Split into Sentences

Split into Words

Remove Punctuations

Remove Stop words

Stem Documents

Term Document Matrix

Add Other Features

TDM + Other Features

Aspect & Entity

Topic Modeling

Sentiment Analysis

Example: The new iPhone X has excellent screen resolution, but, it's price is very high

Text Analysis - Techniques

Pre-Processing:

- Tokenization
- Lemmatization
- N-Grams
- POS Tagging
- NER (Named Entity Recognition)

Sentiment Analysis

- TF * IDF
- Supervised Vs. Unsupervised
- Stanford Core-NLP
- LingPipe
- SentiWordNet

Topic Modeling

- LDA (Latent Dirichlet Allocation)
- DMR (Dirichlet Multinomial Regression)

Evaluation

- Comparison vs. Humans
- Typically text analytics models are evaluated against humans-assigned values
- More than one correct answer possible

Text Analysis - Techniques

Pre-Processing:

- Tokenization
- Lemmatization
- N-Grams
- POS Tagging
- NER (Named Entity Recognition)

Sentiment Analysis

- TF * IDF
- Supervised Vs. Unsupervised
- Stanford Core-NLP
- LingPipe
- SentiWordNet

Topic Modeling

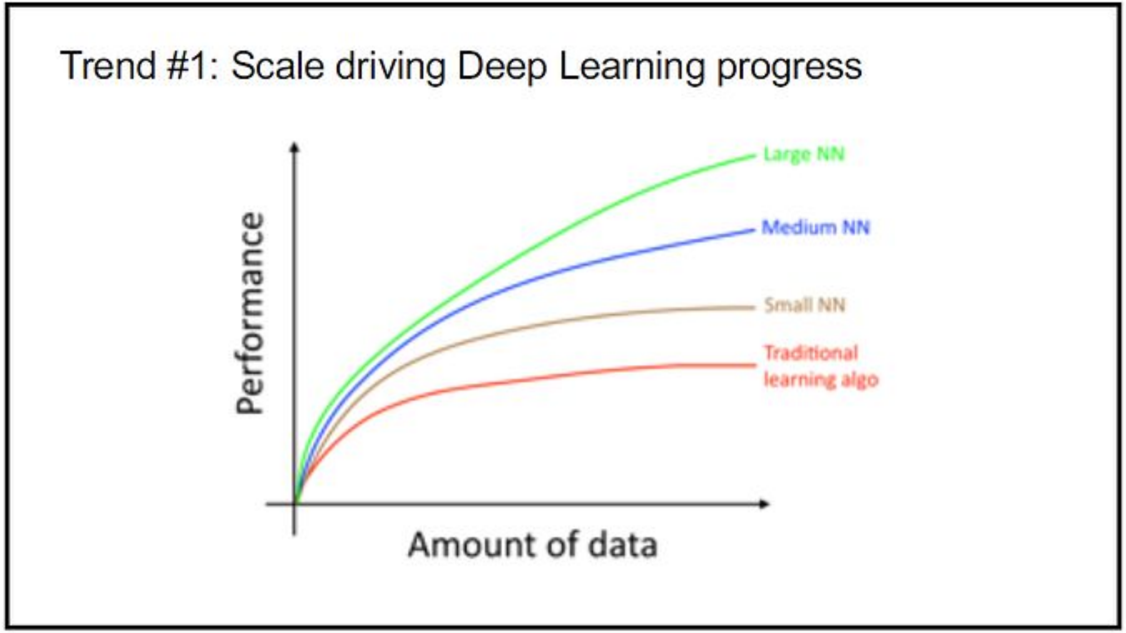
- LDA (Latent Dirichlet Allocation)
- DMR (Dirichlet Multinomial Regression)

Evaluation

- Comparison vs. Humans
- Typically text analytics models are evaluated against humans-assigned values
- More than one correct answer possible

- How to capture the context?

Deep Learning vs. Scale



Reference: Deep Learning Specialization (Andrew Ng)

Machine Learning vs Deep Learning

- Machine Learning to Deep Learning

- Machine learning
  - uses algorithms, parses data, learns from data and predicts
  - limited to human fed inputs

- Deep learning
  - continually analyzes data to draw conclusions, like us!
  - structures algorithms in layers to create an artificial “neural network” that can learn and make intelligent decisions on its own

- Deep learning is a subfield of machine learning.

Deep Learning

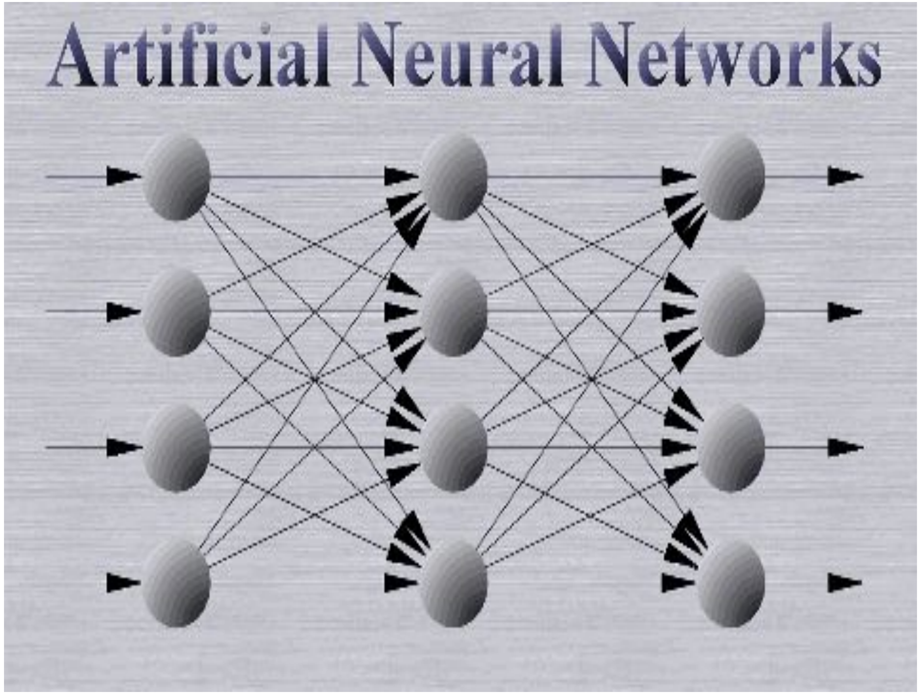
- Hierarchical learning
  - anything is a concept defined in relation to simpler concepts, defined in relation to more simpler concepts and so on…
  - hence can analyze even unstructured data!

Technically:

A stack of layers *of neurons*!

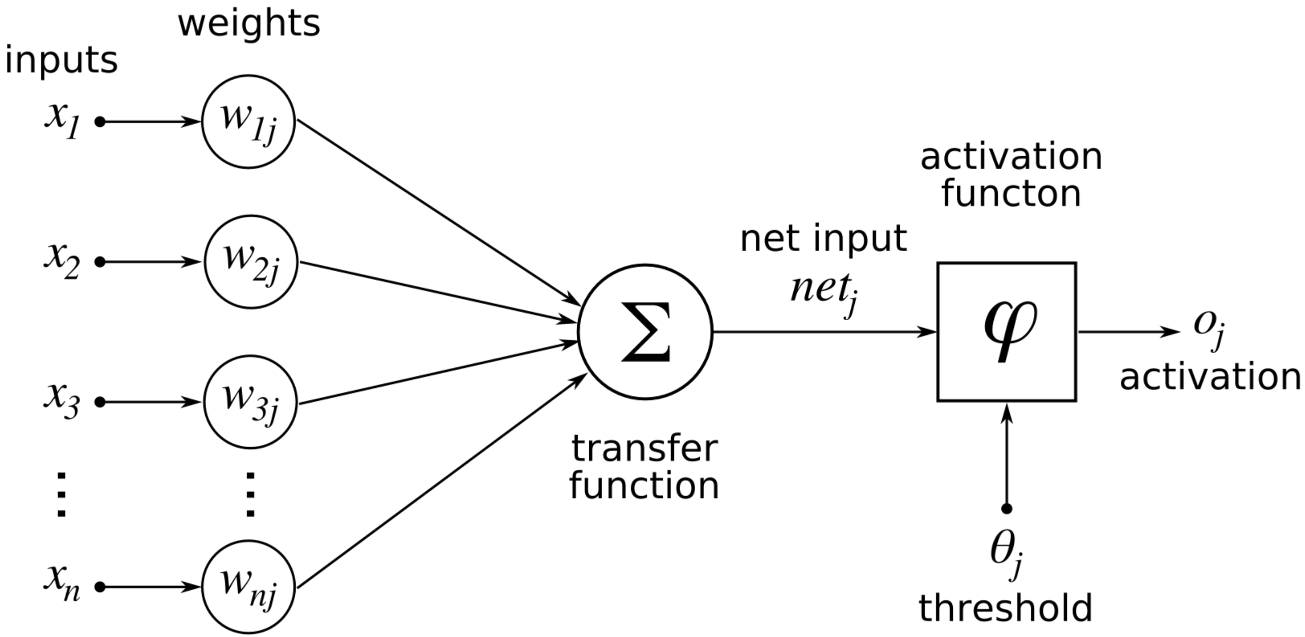
or

Deep Neural Network



https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

A simple neuron!



Prediction = σ (Weights * Inputs + Bias)

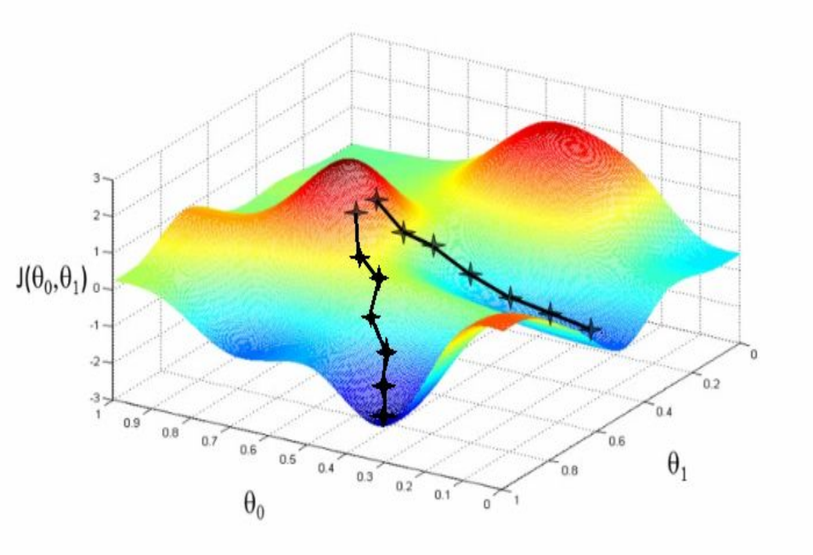
Source: https://commons.wikimedia.org/wiki/File:NeuronModel_deutsch.svg

Training in neural network

1. **Score input**
   *Prediction = σ(Weights * Inputs + Bias)*
   (Sigmoid [0, 1])
   (tanH [-1, 1])
   (ReLU [0, x])
   (Leaky ReLU [0.1x, x])

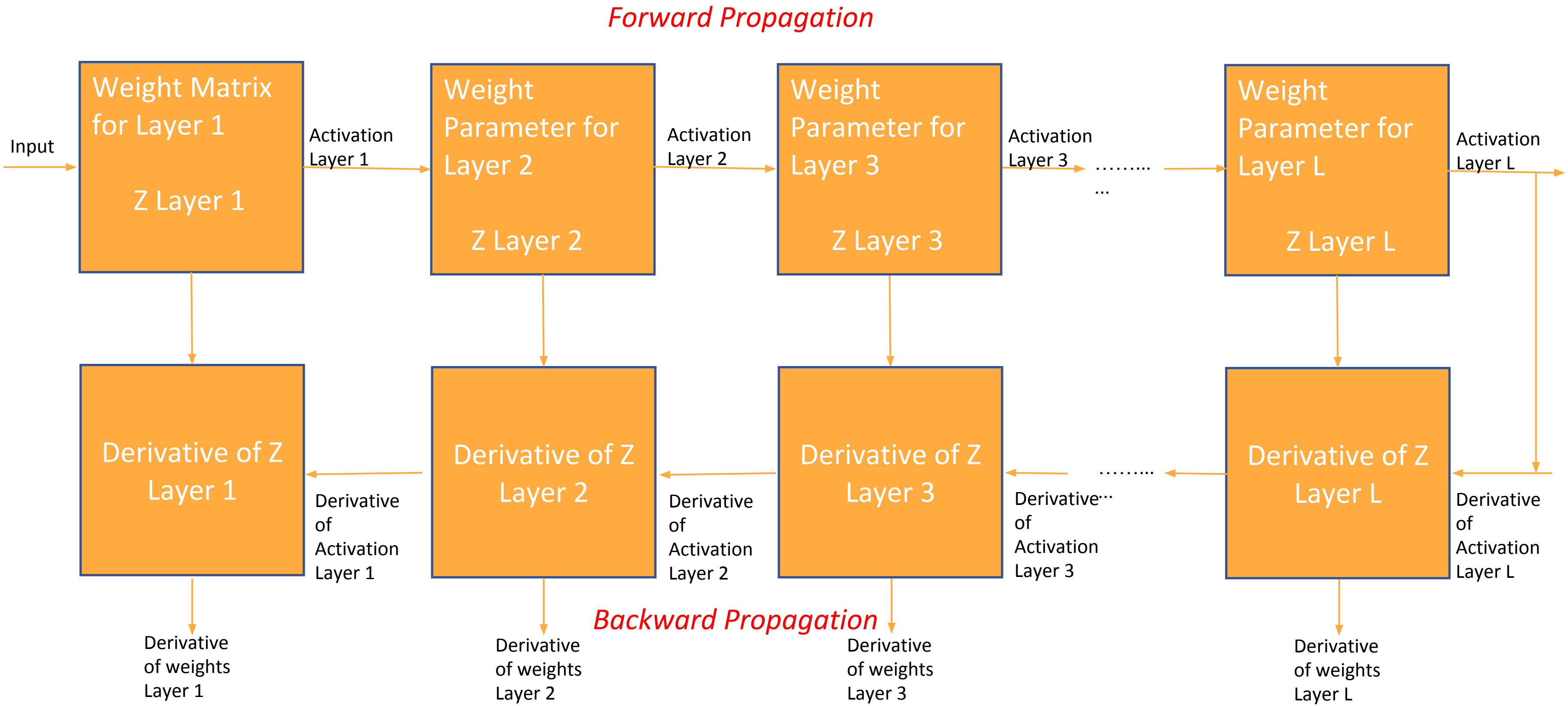
2. **Calculate loss**
   (Mean Squared Error for continuous outputs)
   (Logistic loss for classification)

3. **Apply Adjustments to weight**
   (Gradient descent)
   (RMS prop)
   (Adam Optimizer)



https://medium.com/ai-society/hello-gradient-descent-ef74434bdfa5

Deep Neural Networks



Forward Propagation

Input

Weight Matrix for Layer 1

Z Layer 1

Activation Layer 1

Weight Parameter for Layer 2

Z Layer 2

Activation Layer 2

Weight Parameter for Layer 3

Z Layer 3

Activation Layer 3

……...
…

Weight Parameter for Layer L

Z Layer L

Activation Layer L

Derivative of Z Layer 1

Derivative of Activation Layer 1

Derivative of Z Layer 2

Derivative of Activation Layer 2

Derivative of Z Layer 3

Derivative of Activation Layer 3

……...

Derivative of Z Layer L

Derivative of Activation Layer L

Derivative of weights Layer 1

Derivative of weights Layer 2

Backward Propagation

Derivative of weights Layer 3

Derivative of weights Layer L

ThoughtWorks®

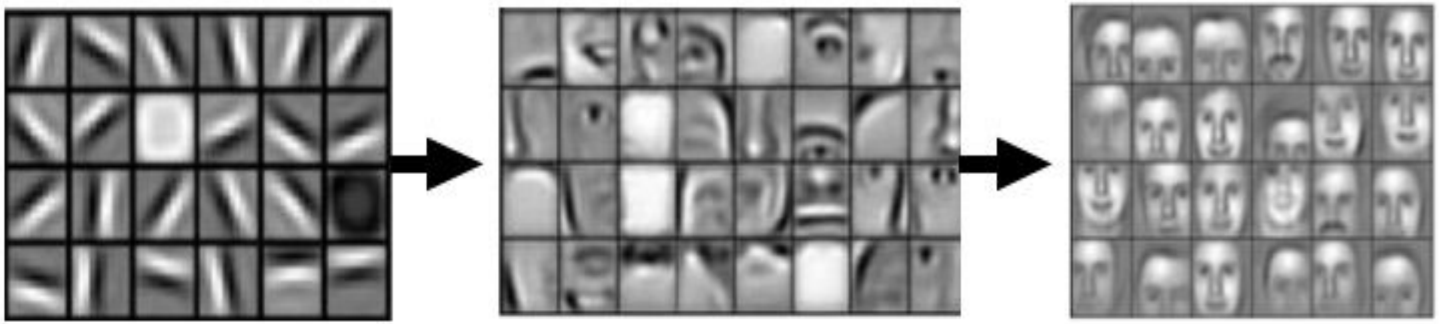
Reference: Deep Learning Specialization (Andrew Ng)

XCONF

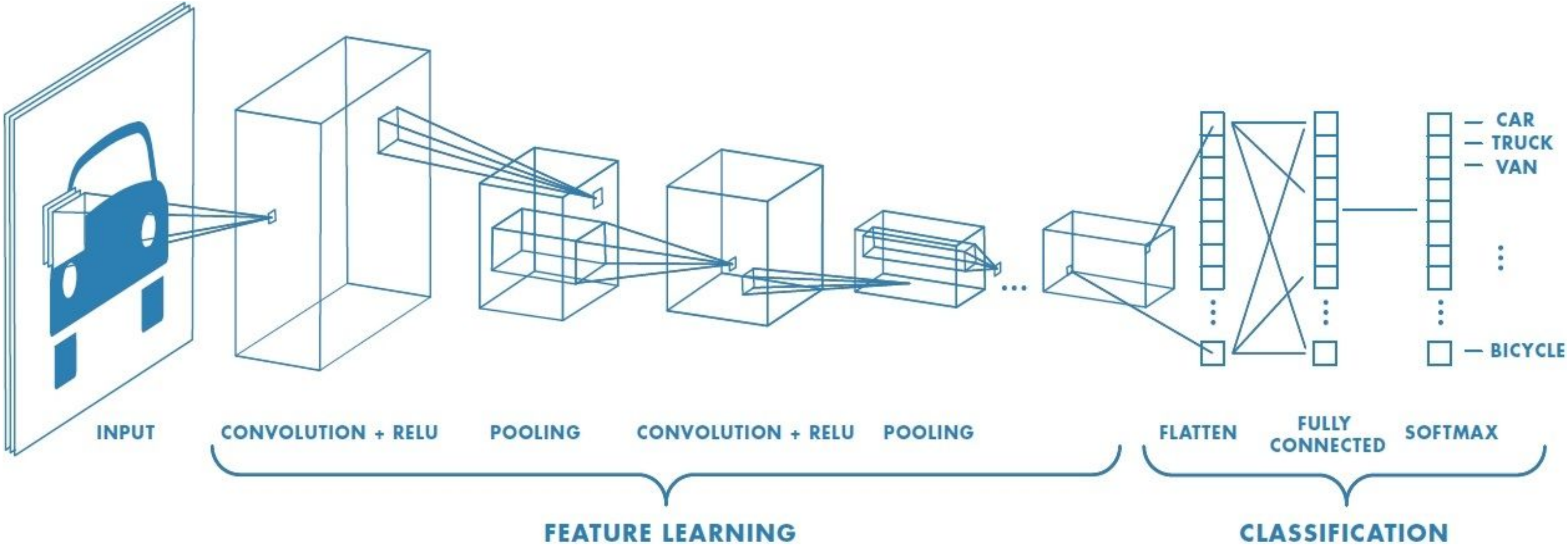
Convolutional Neural Network

- A subclass of Deep Neural Network
- Mimics object identification by Human
- Constrained architecture to:
  - Leverage temporal and spatial structure of domain
  - Reduce computation

Excels at understanding complex concepts as a combination of smaller and smaller pieces of information!



CNN/ ConvNet

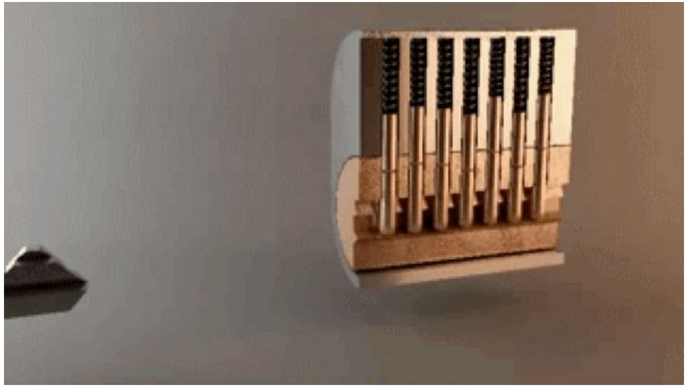


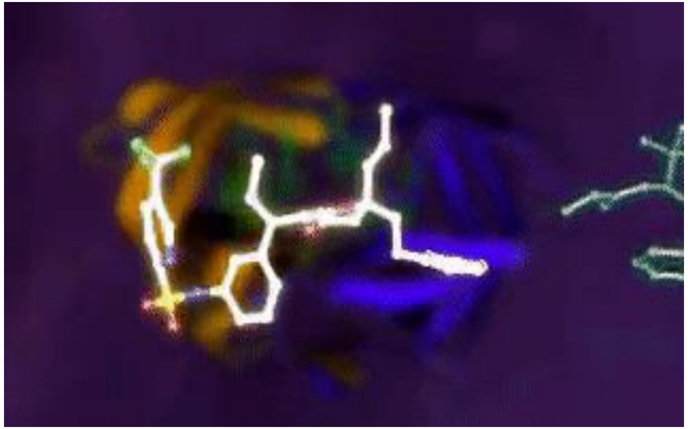
Source: https://it.mathworks.com/discovery/convolutional-neural-network.html

Algorithm CNN

```
INPUT:  Training dataset T,  say images with labels

TRAINING
For every image in T,  do

    Create Input Vector for neural network
        (20* 20 RGB image has input array length of 20* 20* 3)

    Collect all Features
    For every feature in Features, do:
        CONVOLUTION
        POOLING
        ACTIVATION

    Collect all the output matrices
    FULLY CONNECTED LAYER: Transform into one D array
        PROBABILITY CONVERSION using SOFTMAX
        OUTPUT LABEL = Label with max probability value

    Find ERROR using LOSS FUNCTION
    Find weight update (delta W) and BACK PROPAGATE to update weights


Trained model can do classification of new data.
```

AtomNets





- Deep CNN based
- Structure based drug design
- Drug Design:
  - Target Protein
  - Ligands to targets
  - Design ligands that are binders

Learning:

- DUD-E dataset
- Recognize basic chemical structures on its own
  like Hydrogen bonding, Carbon structures.

Candidate treatment for Ebola, awaiting animal trials!

Deep Learning Resources/References:

- Machine Learning – Andrew Ng https://www.coursera.org/learn/machine-learning/home/welcome
- Deep Learning – Andrew Ng https://www.coursera.org/specializations/deep-learning
- Convolutional Neural Networks http://yann.lecun.com/exdb/lenet/
- Deep Learning http://deeplearning.net/
- Deep Residual Learning https://arxiv.org/abs/1512.03385
- Automated Image Captioning - Andrej Karpathy https://cs.stanford.edu/people/karpathy/sfmltalk.pdf
- The Unreasonable effectiveness of RNNs - Andrej Karpathy http://karpathy.github.io/2015/05/21/rnn-effectiveness/
- Machine Learning 101 - https://docs.google.com/presentation/d/1kSuQyW5DTnkVaZEjGYCkfOxvzCqGEFzWBy4e9Uedd9k/preview?imm_mid=0f9b7e&cmp=em-data-na-na-newsltr_20171213&slide=id.g168a3288f7_0_58
- Machine Learning Mastery - https://machinelearningmastery.com/
- Wikipedia - https://en.wikipedia.org/

Thank You!

Questions?

Feedback at: bit.ly/XconfTalkFeedback