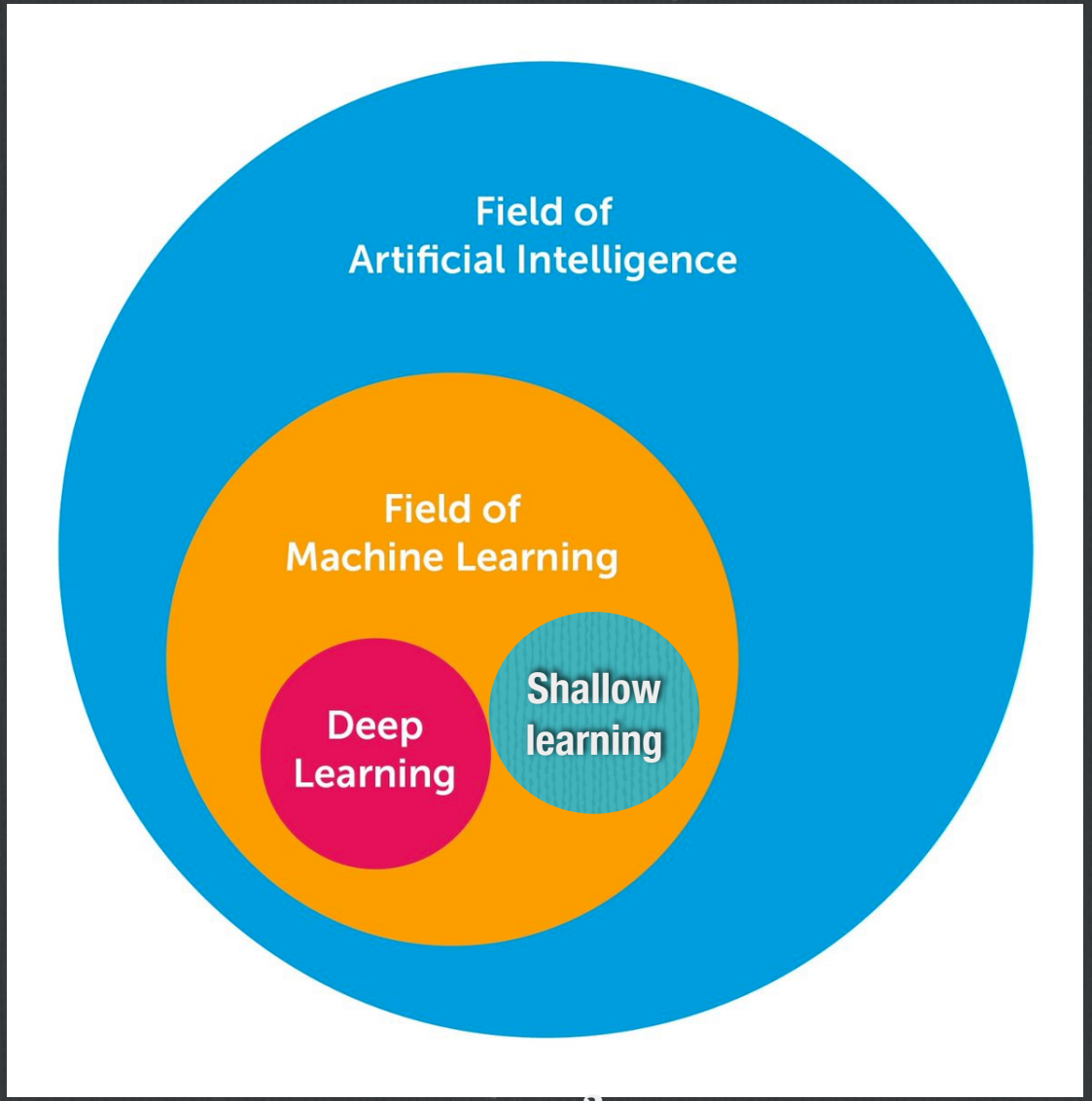
Machine learning / AI and audio

Audio Engineering Society, Melbourne section

Guillaume Potard
gui@loftyconsulting.com
12 August 2019

Contents

1- Introduction to machine learning (ML) and artificial intelligence (AI)

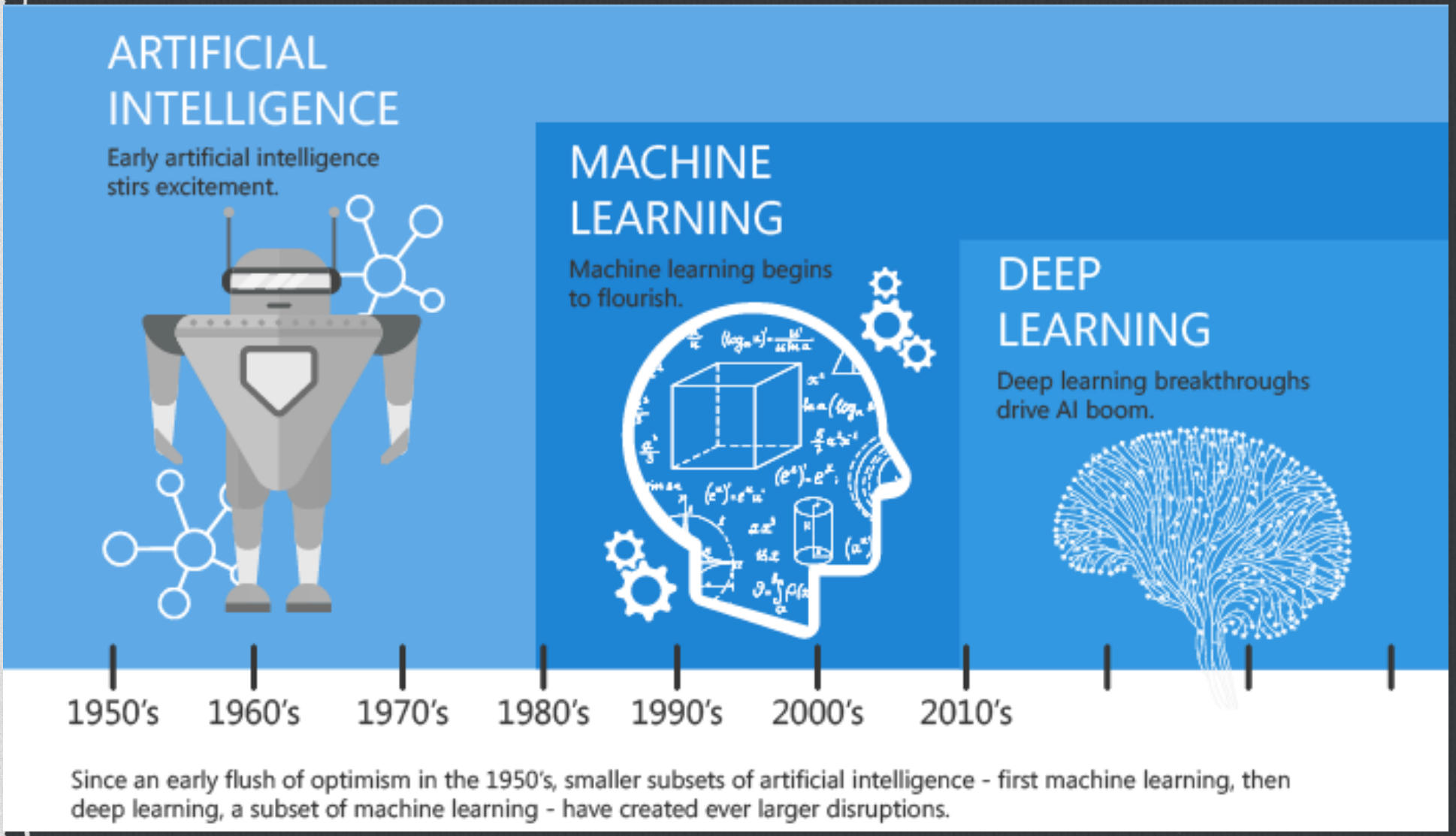


AI Grades

We are here

Singularity

- Artificial narrow intelligence (ANI) or 'Weak AI'
  - specialises only in one area (e.g. Alpha Go)
- Artificial general intelligence (AGI) or 'strong AI' Human-level AI.
  - As smart as a human across the board (reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly, and learn from experience) + creativity and art.
  - Human brain ~1000 petaflops (200 Hz, 20 W) [thousand trillion]
  - Tanhe-2 supercomputer 32 petaflops (GHz, 25 MW)
- Artificial superintelligence (ASI). Millions of time more intelligent than a human.



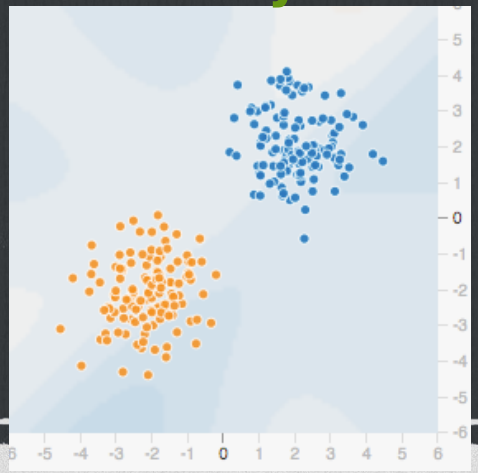
Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

What is ML ?

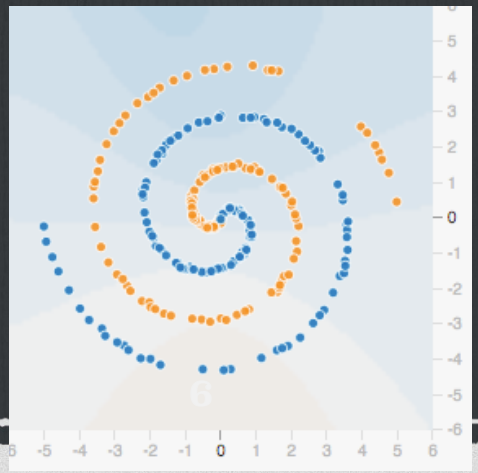
- When you teach a computer to perform a specific task without giving it explicit commands.
  -> Use some data to train algorithm
  -> Prediction / classify / action

- Why ?
  -> Underlying rules are too complex
  -> No closed-form equation -> Statistics
  -> Too many dimensions (curse of dimensionality)
  -> Too much data
  -> Rule based : programmer has to do the ML in his head anyway

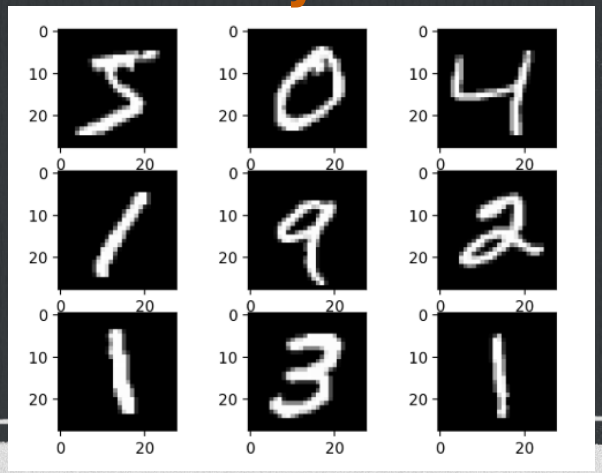
Easy



Hard



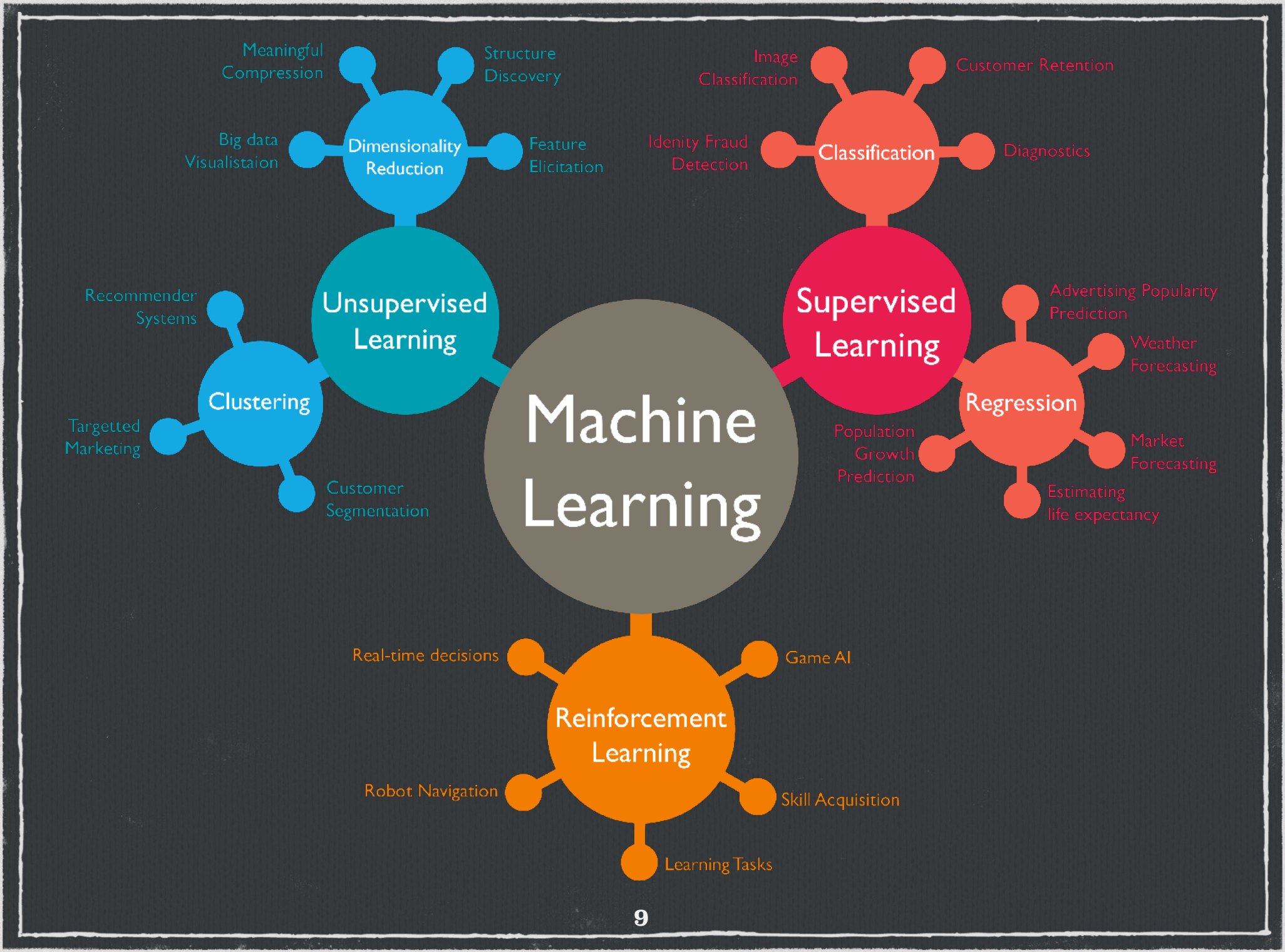
Very hard



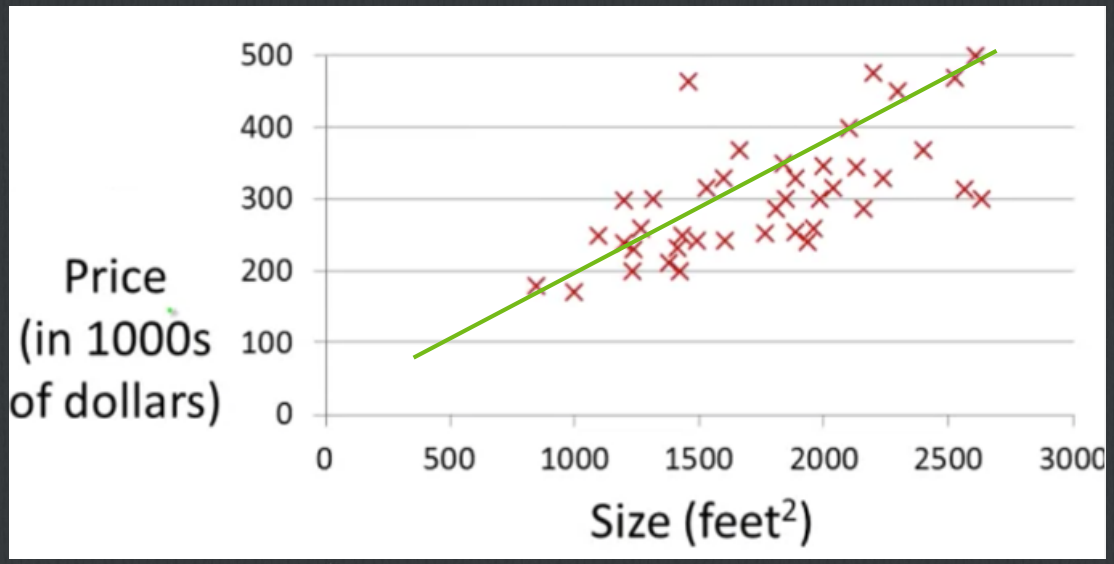
Where is ML used ?

- Applications :
  - fraud detection (banks, ATO, insurance companies..)
  - image recognition
  - sentiment analysis
  - email spam detection
  - recommendation systems
  - medical diagnosis
  - speech recognition and synthesis
  - Google , Facebook etc.
  - Data mining
  - Translation
  - Robotics
  - Etc…

2- Concepts and principles (shallow learning)


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

2.1 Supervised learning Regression

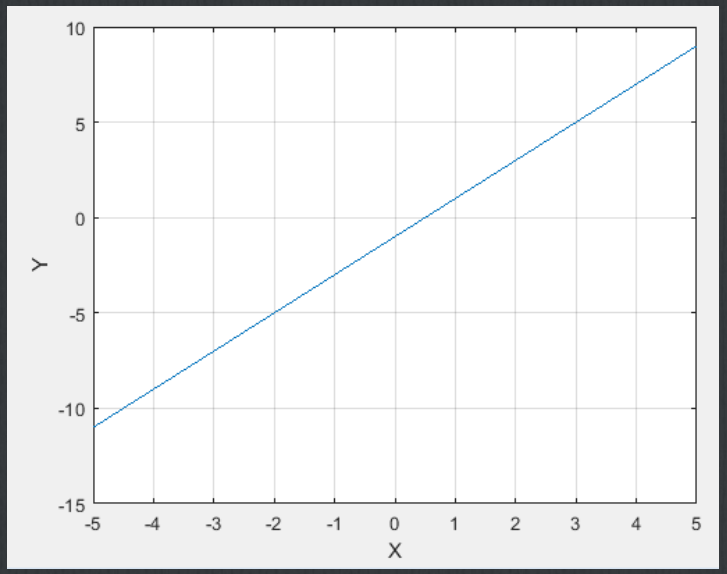


$$h(x) = \theta_0 + \theta_1 x$$

Linear regression

X = -1, 0 , 1, 2, 3, 4

Y = -3, -1, 1, 3, 5, 7



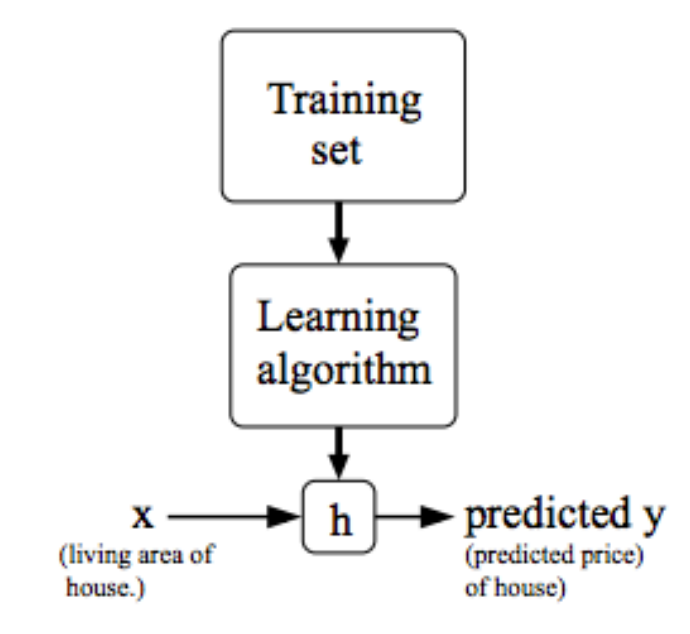
$$y = h(x)$$

$$h(x) = -1 + 2x$$

$$h(x) = \theta_0 + \theta_1 x$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$ n features

$$h(x) = \theta^T x$$



$$y = h(x)$$

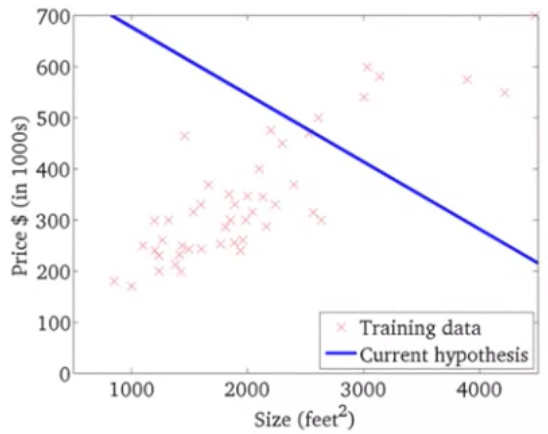
Cost function :

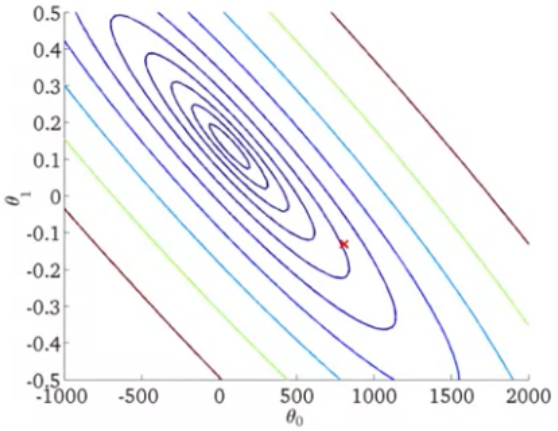
$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}_i - y_i \right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i \right)^2$$

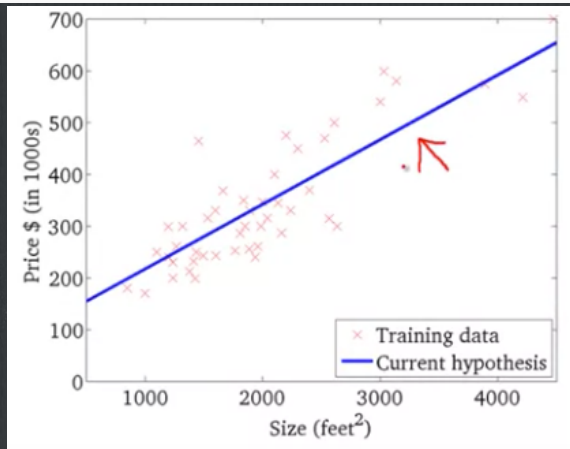
Find optimal parameters (θ_0, θ_1) that minimise the cost function

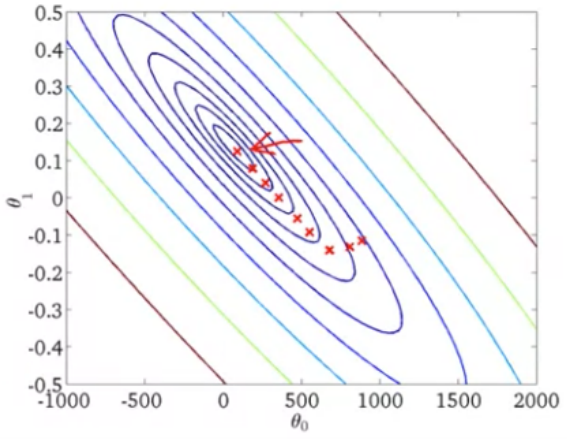
=> Gradient descent

$h_\theta(x)$ $J(\theta_0, \theta_1)$

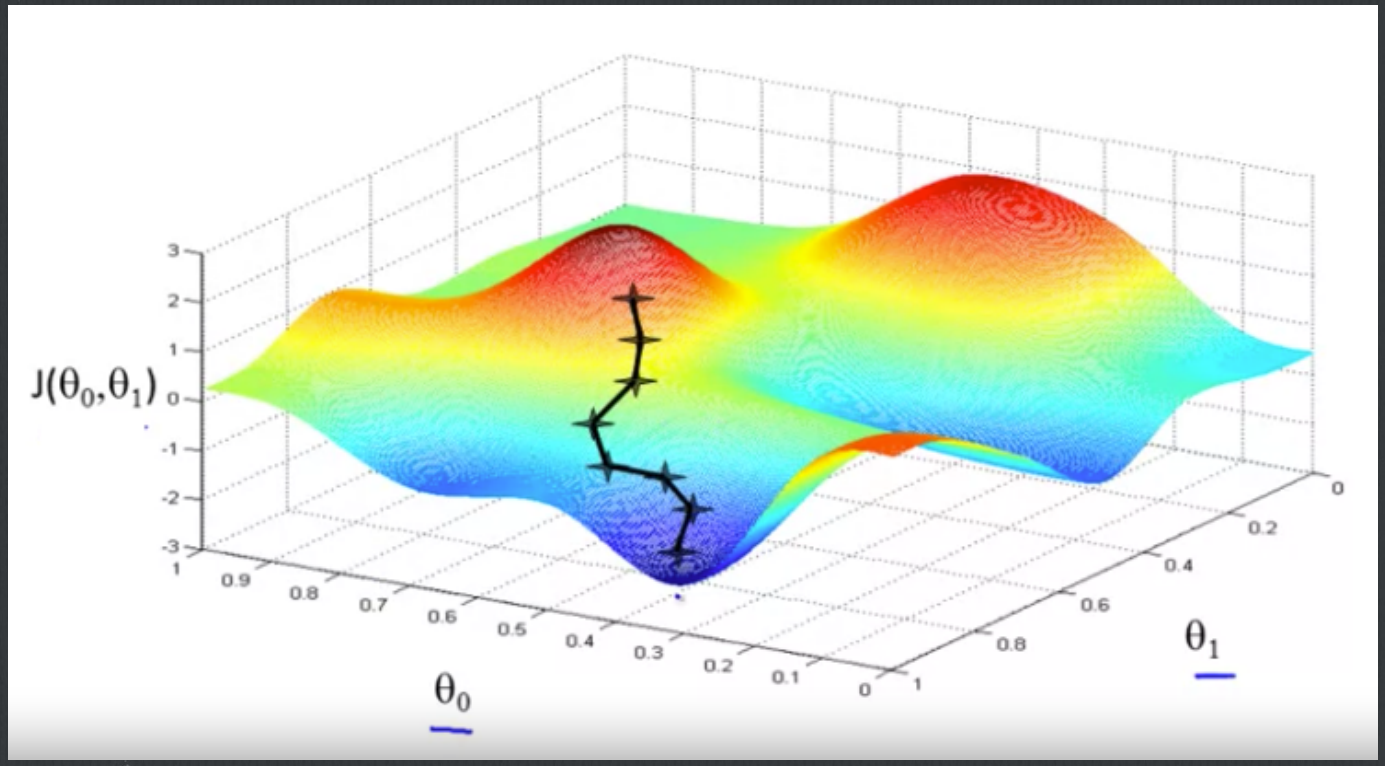








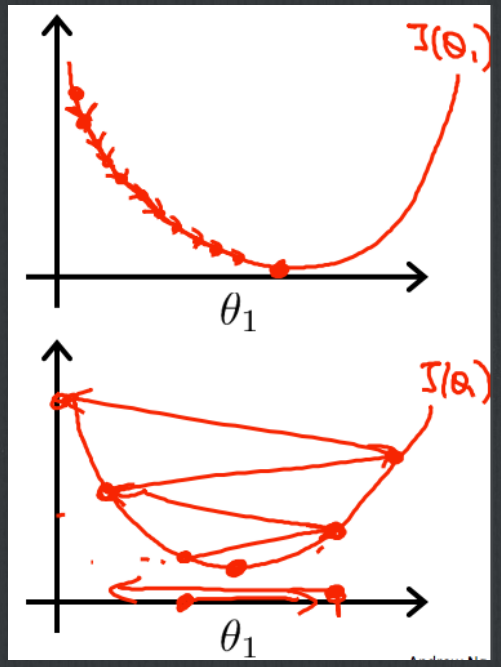
Gradient descent



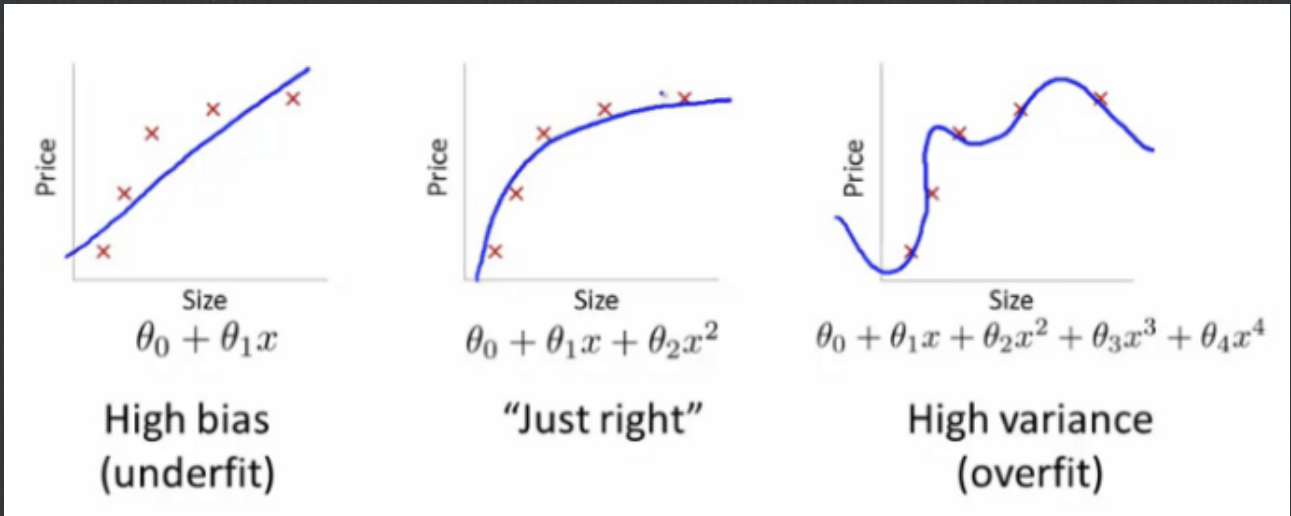
repeat until convergence {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \qquad (\text{for } j = 0 \text{ and } j = 1)$$

}

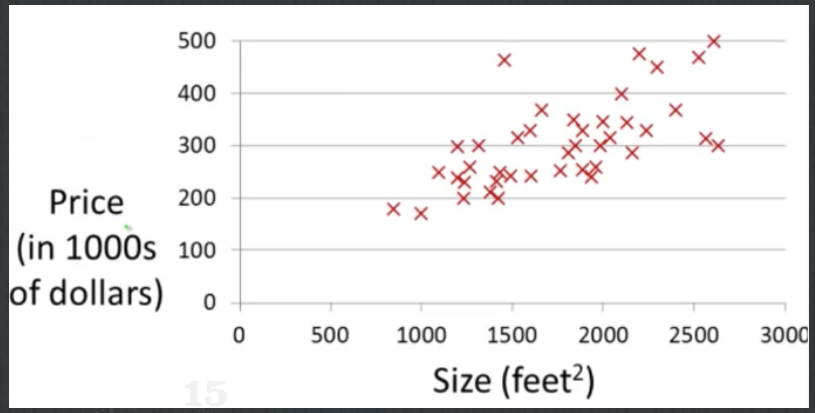


Overfitting in regression

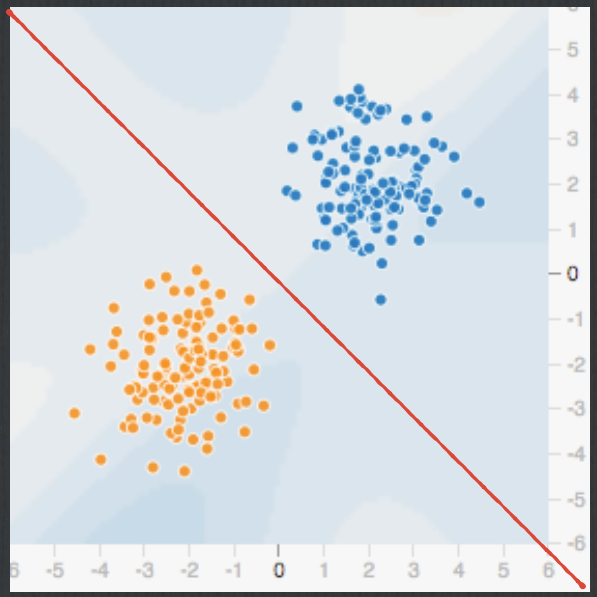


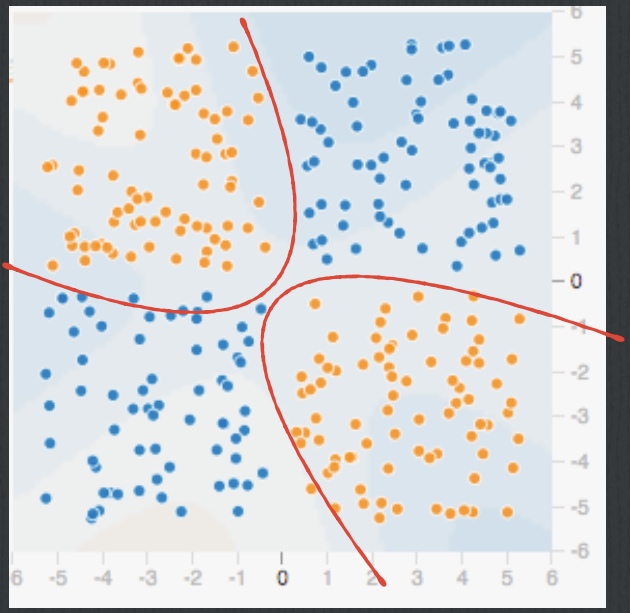
Polynomial regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^3$$



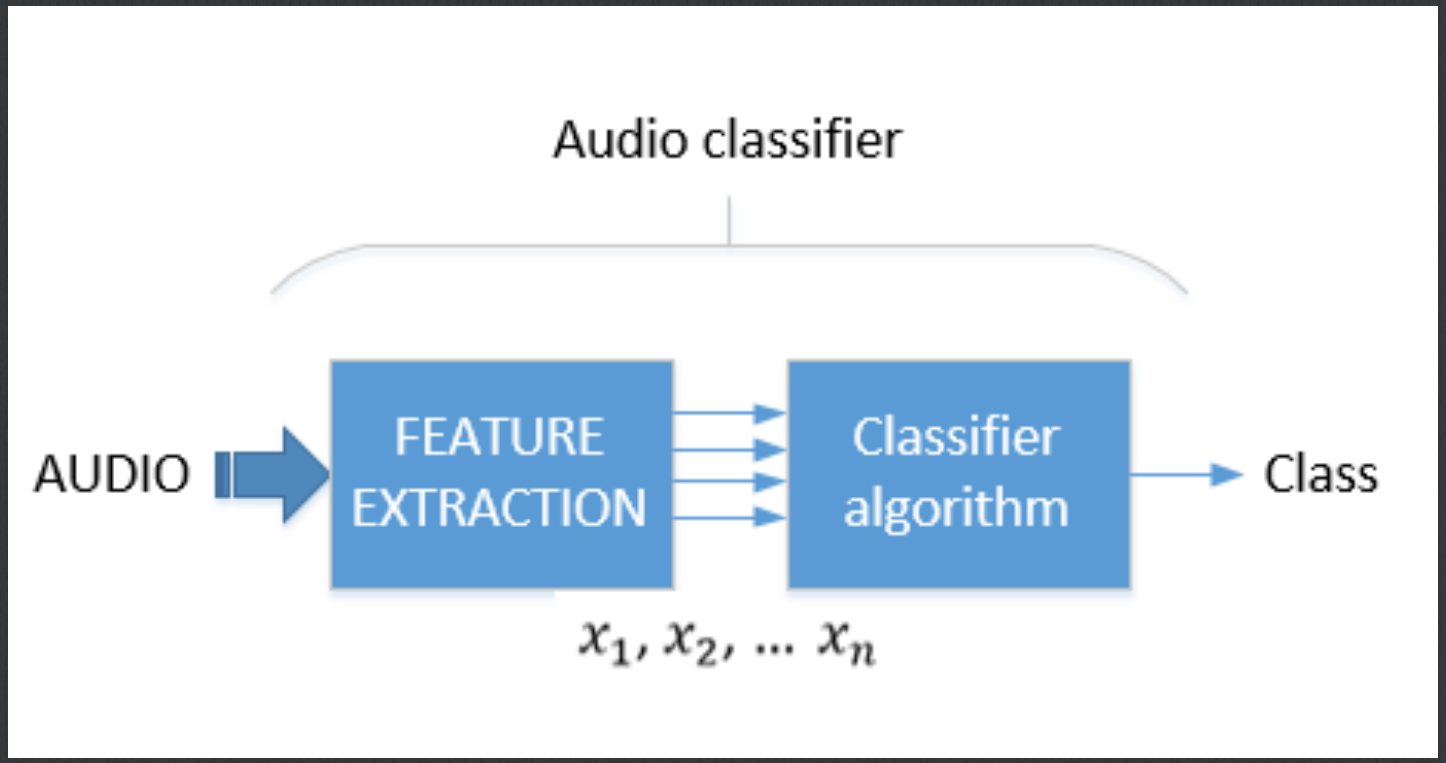
Classification





- Training data labelled with classes

Classification for audio

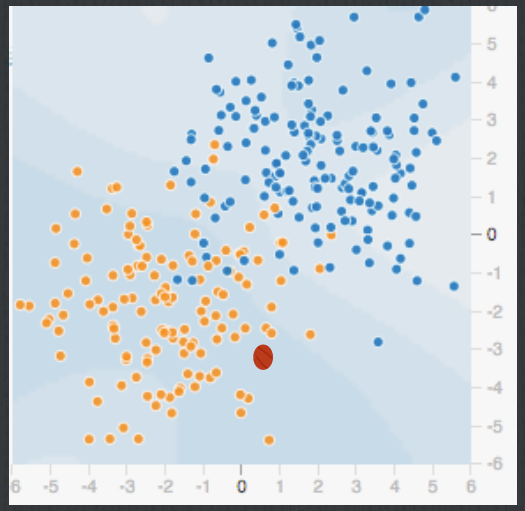


- Feature engineering
- Audio features: tonality, pitch, spectrum flatness, Mel-frequency cepstrum coefficients etc…

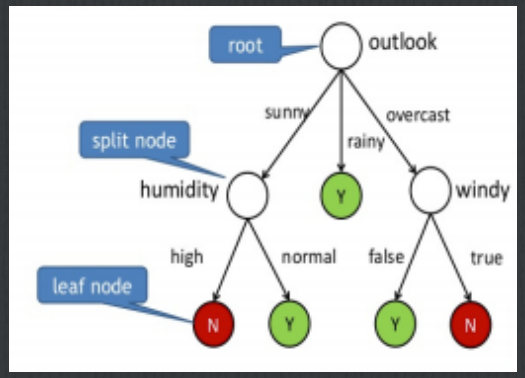
Classification algorithms

- k nearest neighbourghs (kNN)

(store all data in model)



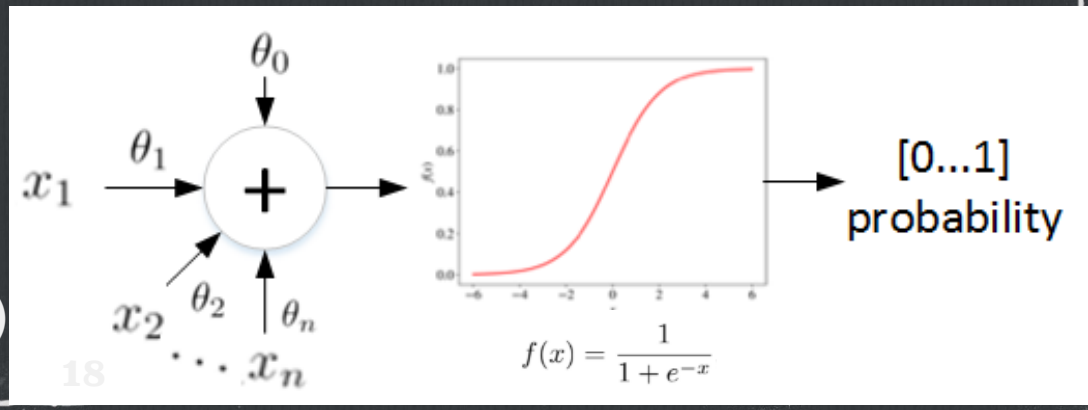
- Decision tree & random forests



- Logistic regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

(Maximise likelihood with gradient descent)

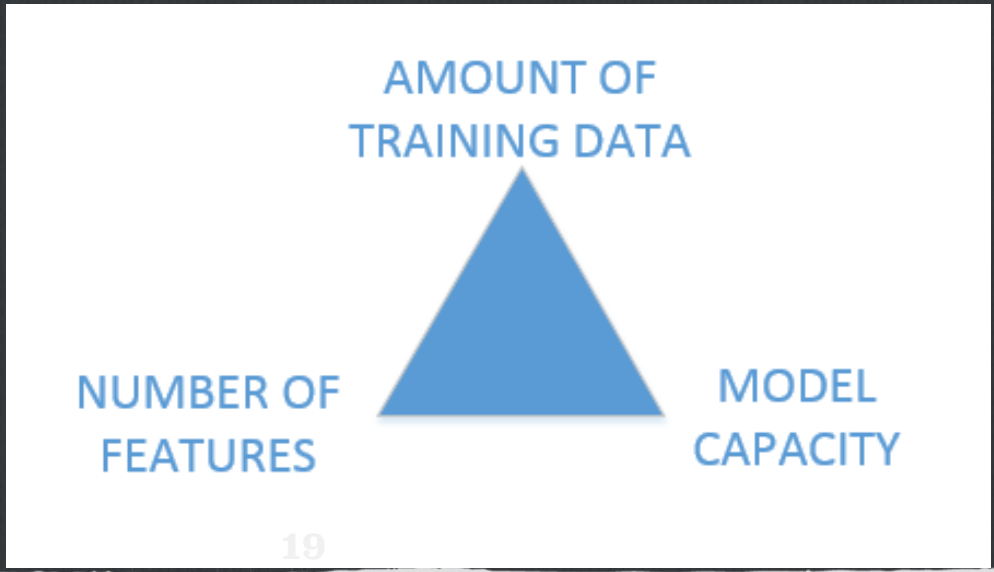


Testing

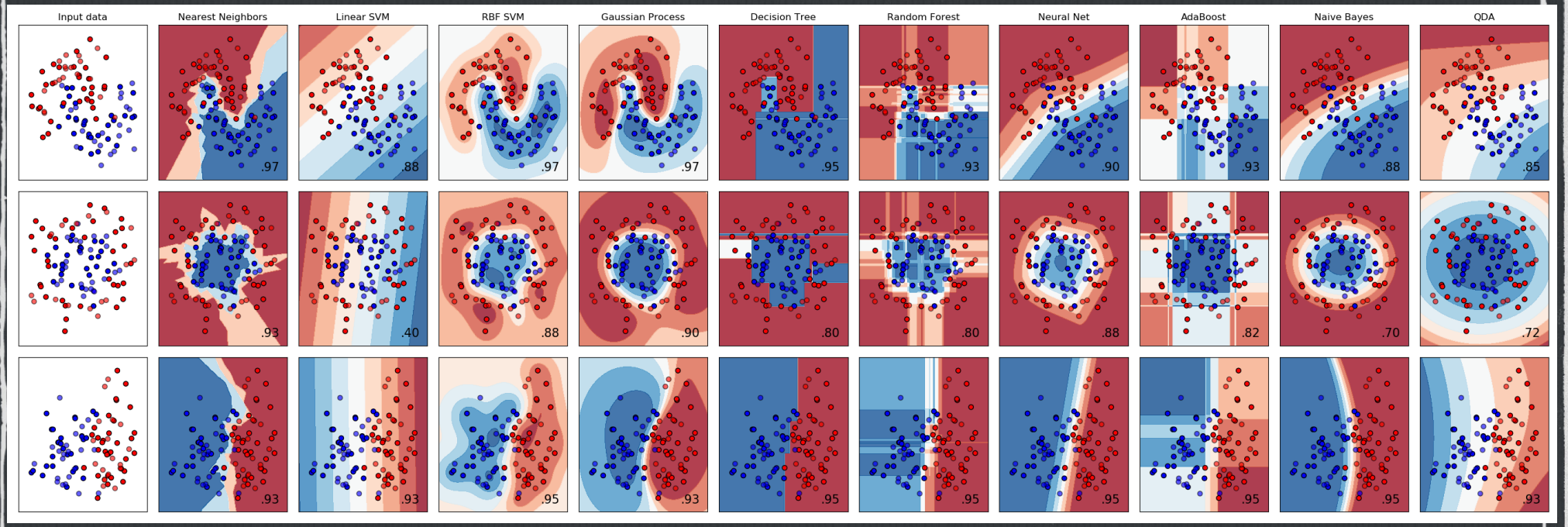
- k-fold validation

Iteration 1	Train	Train	Train	Train	Test
Iteration 2	Train	Train	Train	Test	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Test	Train	Train	Train
Iteration 5	Test	Train	Train	Train	Train

- Overfitting / Underfitting -> Bias / Variance tradeoff

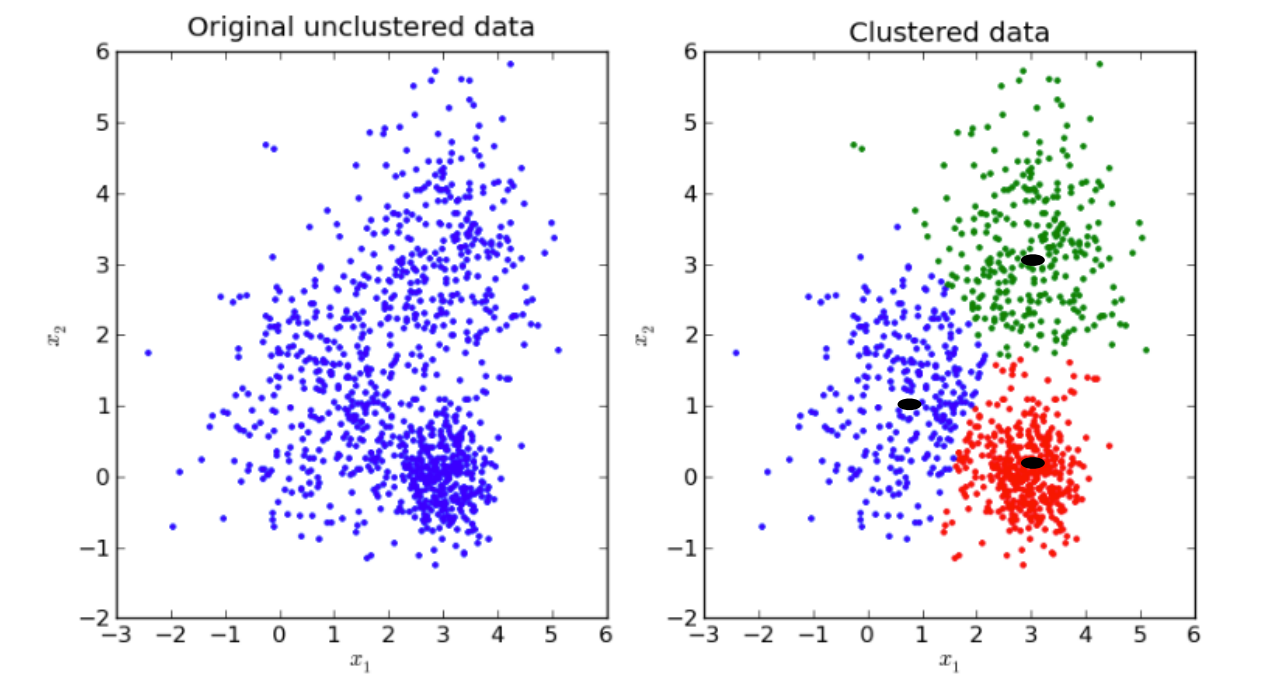


Aim: model that generalises well


Input data
Nearest Neighbors
Linear SVM
RBF SVM
Gaussian Process
Decision Tree
Random Forest
Neural Net
AdaBoost
Naive Bayes
QDA
.97
.88
.97
.97
.95
.93
.90
.93
.88
.85
.93
.40
.88
.90
.80
.80
.88
.82
.70
.72
.93
.93
.95
.93
.95
.95
.95
.95
.95
.93

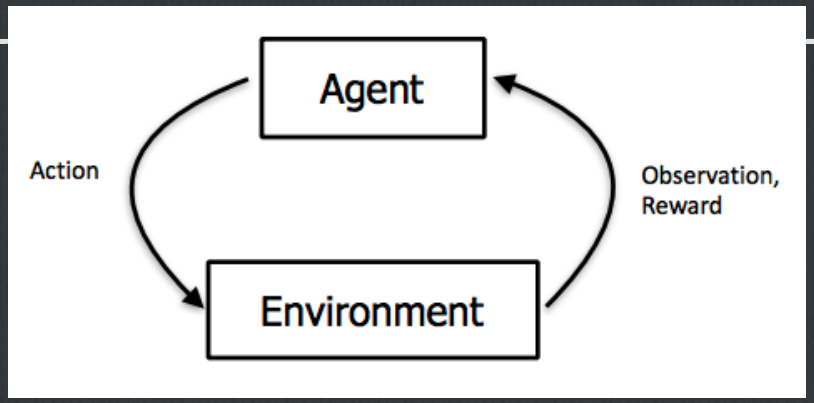
2.2 Unsupervised learning

- No labels available
- e.g. K-means clustering algorithm:



- document / music automatic classification
- music segmentation
- cohorts of customers

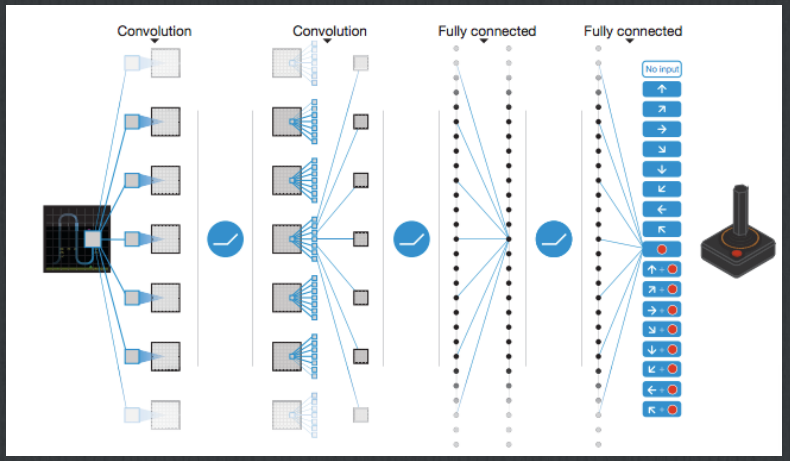
2.3 Re-enforcement learning

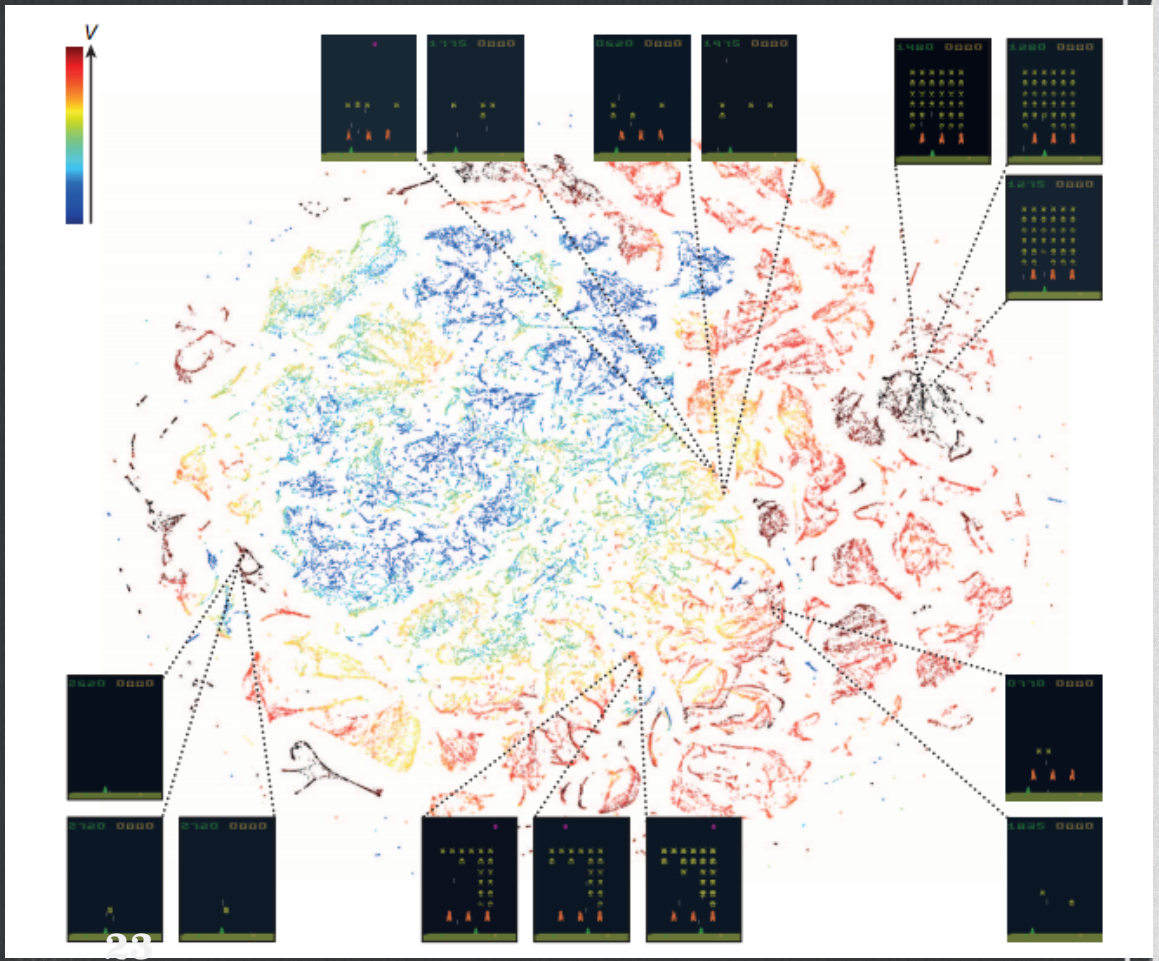


- Deep Q learning
- Discount factor:
  Future reward
  vs greediness

exploration / exploitation
tradeoff

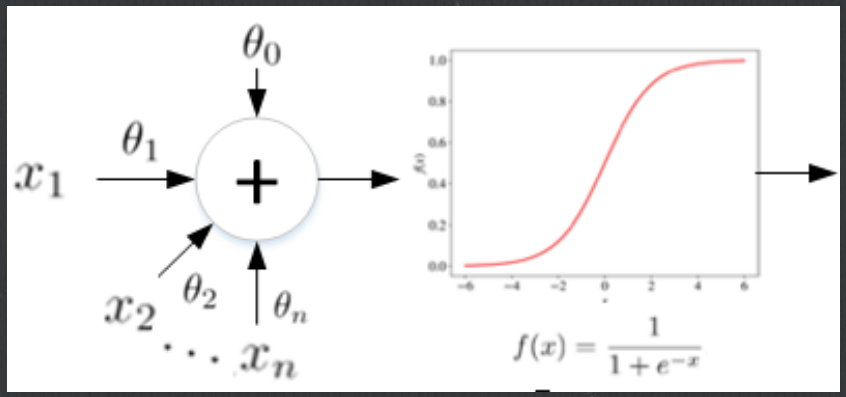
- Atari 2600 games: better than any human player
- Idea for a game : engineer audio features ?!





3 Deep learning

- **1 neuron (perceptron)**

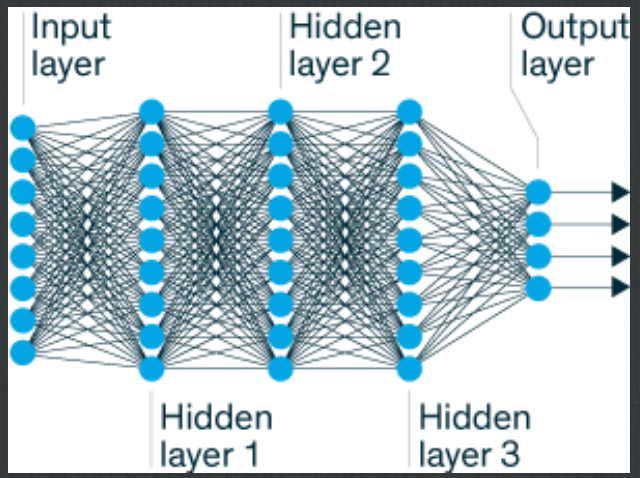


A	B	Out
0	0	0
0	1	1
1	0	1
1	1	0

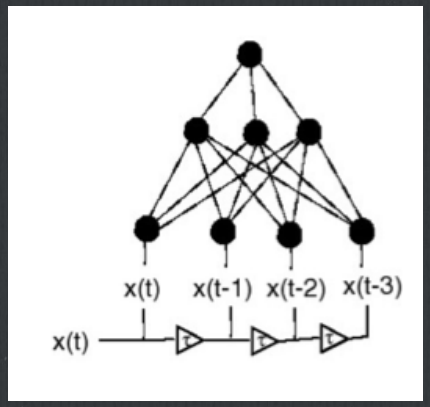
Can't learn XOR function !

- **Deep learning**

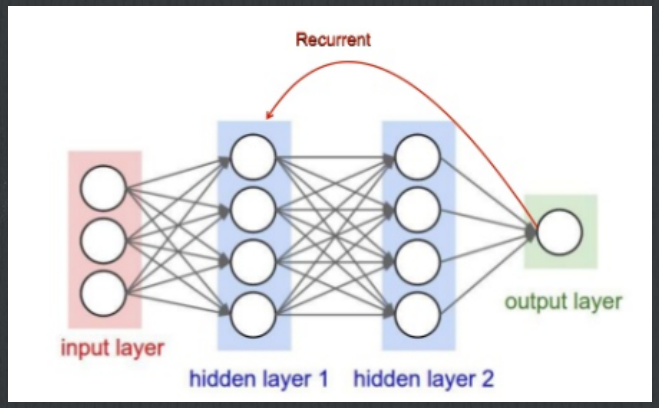
Can learn anything !



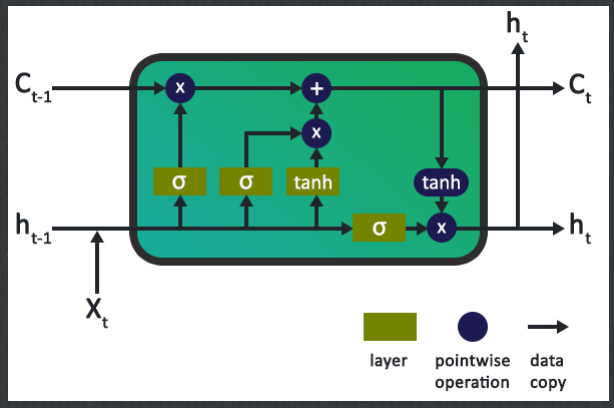
- Time-delay neural networks (TDNNs)



- Recurrent neural network (RNN)



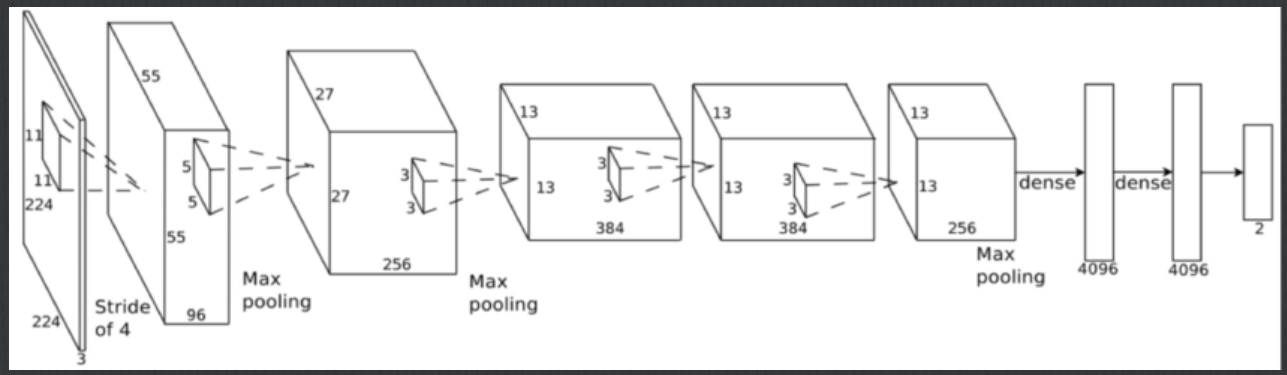
- Long-term short term (LSTM) neuron -> neuron + memory unit



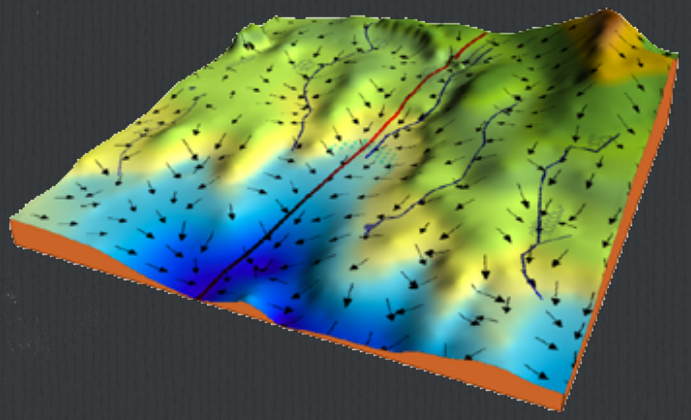
- No need for feature engineering
  -> Features learned automatically

- More layers , more level of abstraction

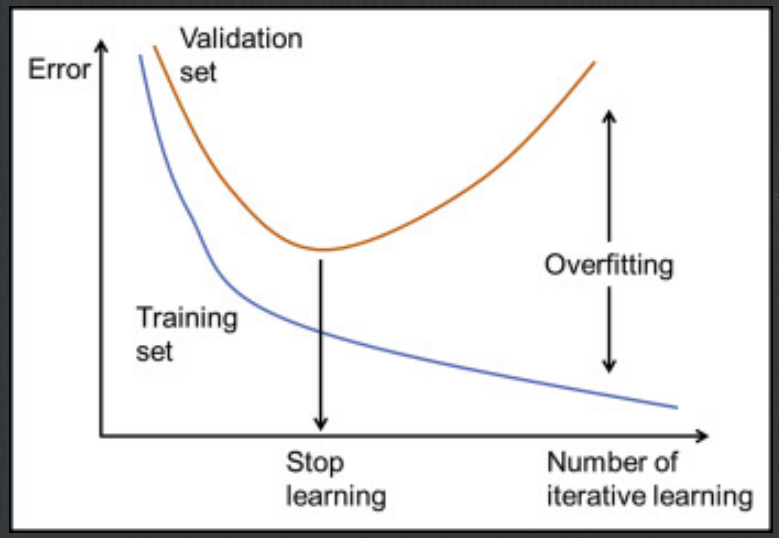
- Convolutional Neural Network (CNN)



- Training a complex neural network:
  - Adjust weights incrementally to minimise a loss function
  - Stochastic gradient descent , Adam, Adaboost …



- Watch out for overfitting!
  - Early stopping
  - Drop outs
  - Regularisation
  (penalise large weights)
  - More training data…



- Back propagation
  (Geoffrey Hinton 1986)

Limits of machine learning / neural nets

- ☐ The model is as good as the data
- ☐ Can be slow to train, need very large datasets
- ☐ Opaque / hard to debug
- ☐ Not magical - cannot learn noise
- ☐ Not a cure-all solution to every problem

4- ML and Audio

Machine listening

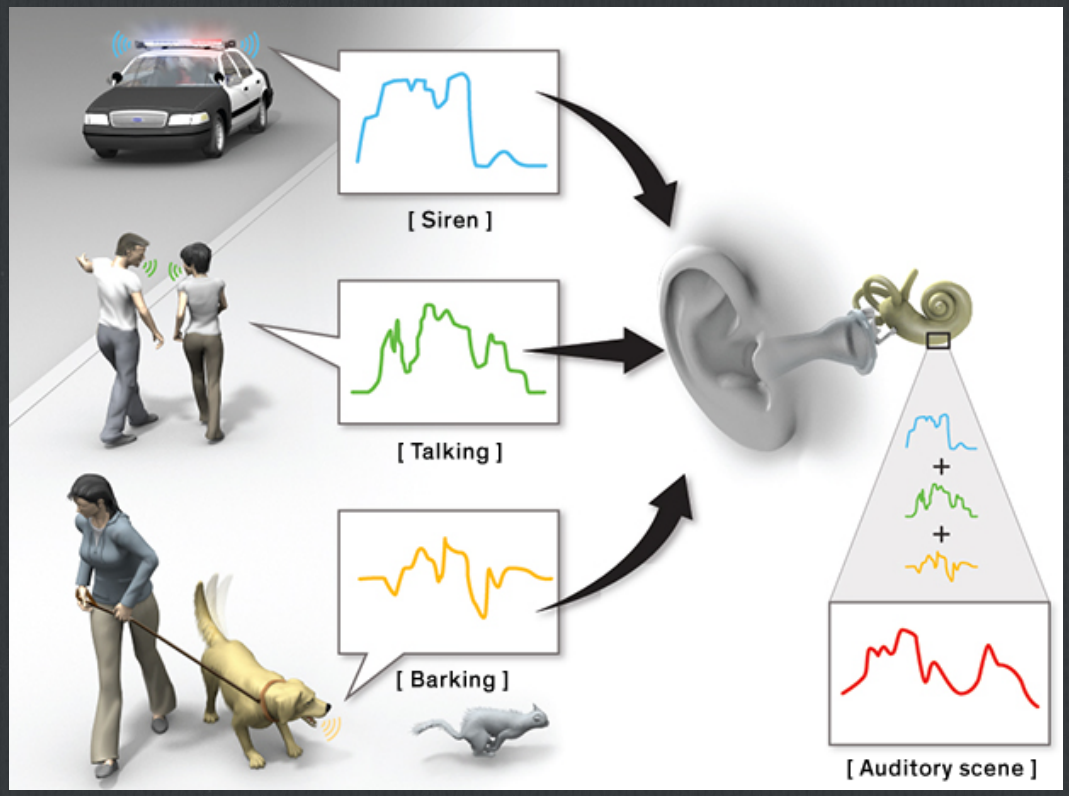
Sound -> Meaning

Human hears

Machine sees

Machine listening

- Speech recognition
  (Most speech recognition systems are only statistical models + Natural Language processing NLP)
- Computational auditory scene analysis (CASA)
- Fault analysis
- Virtual sensing
- Robotics
- Autonomous vehicles

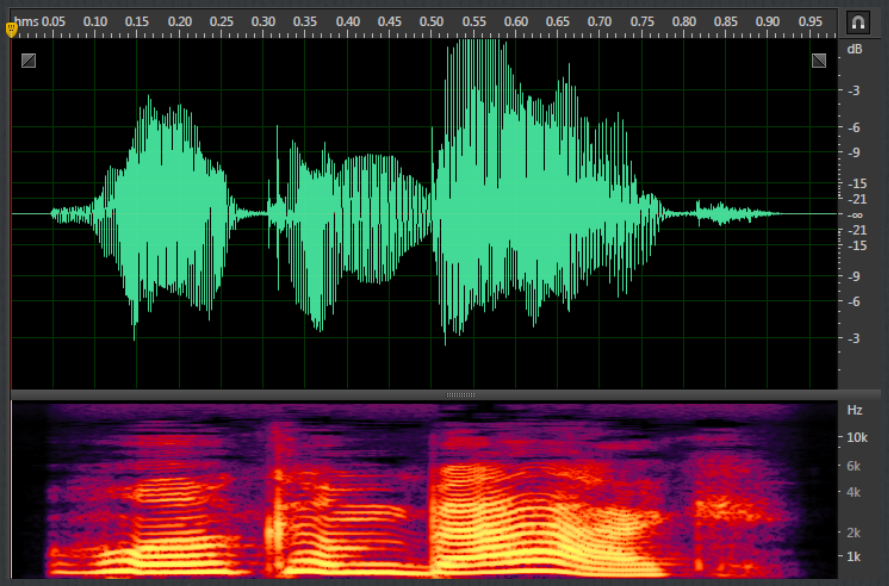


Machine listening

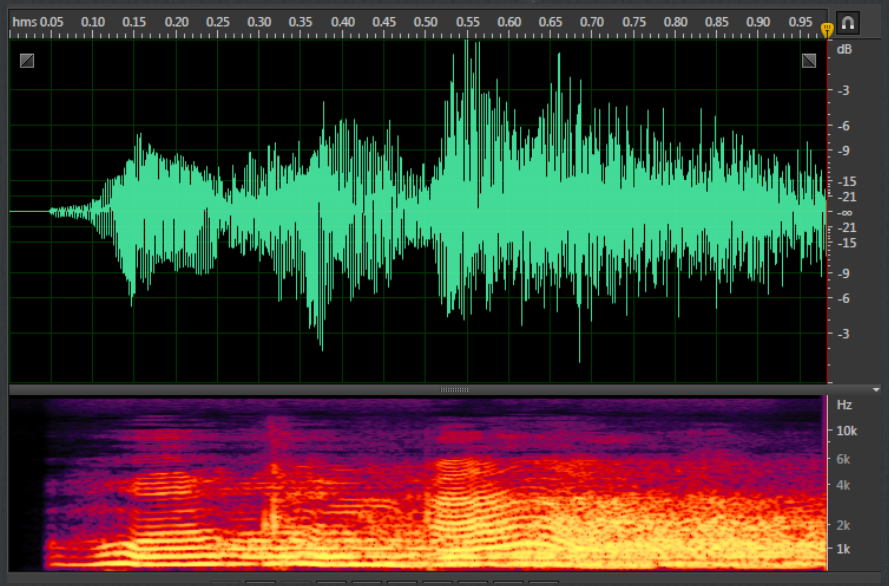
Example

- Keyword detection: Sparkfun Edge board
- 512 point FFT + convolutional neural network
- Can run for weeks on a single coin cell-battery

Dry



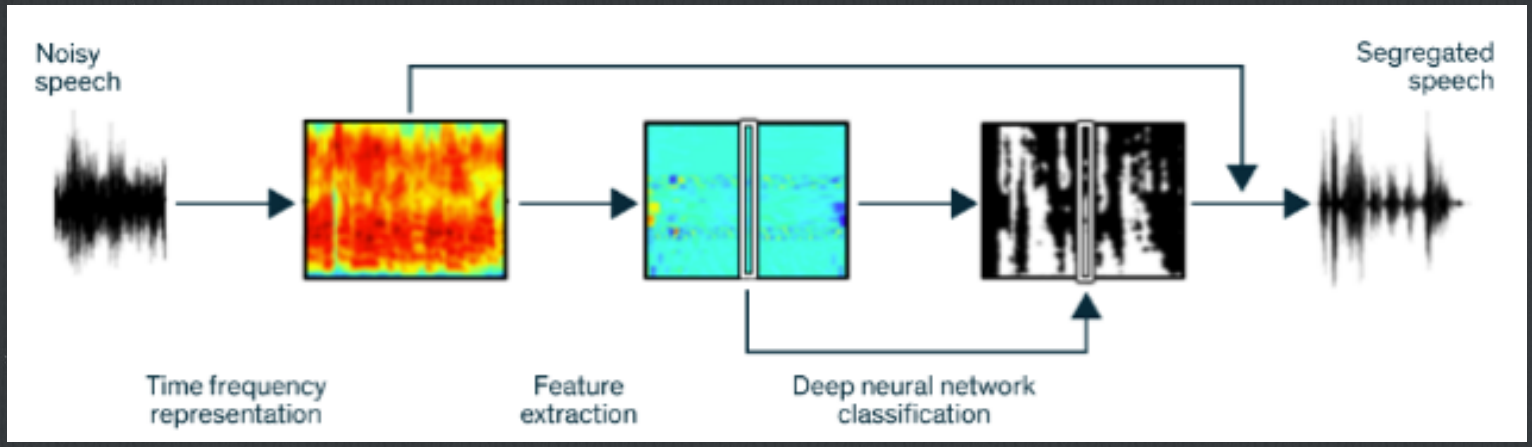
Reverberated



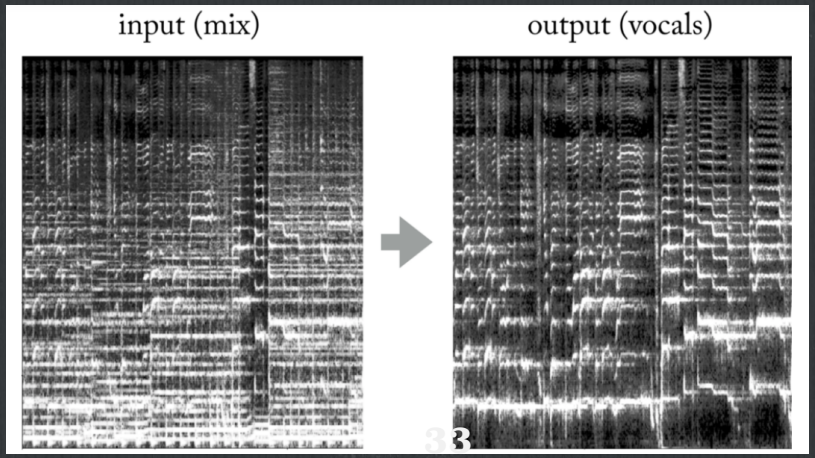
- **Brain does not do a spectrogram + image recognition to hear speech !**

Digital signal procesing (DSP)

- De-noising using classification

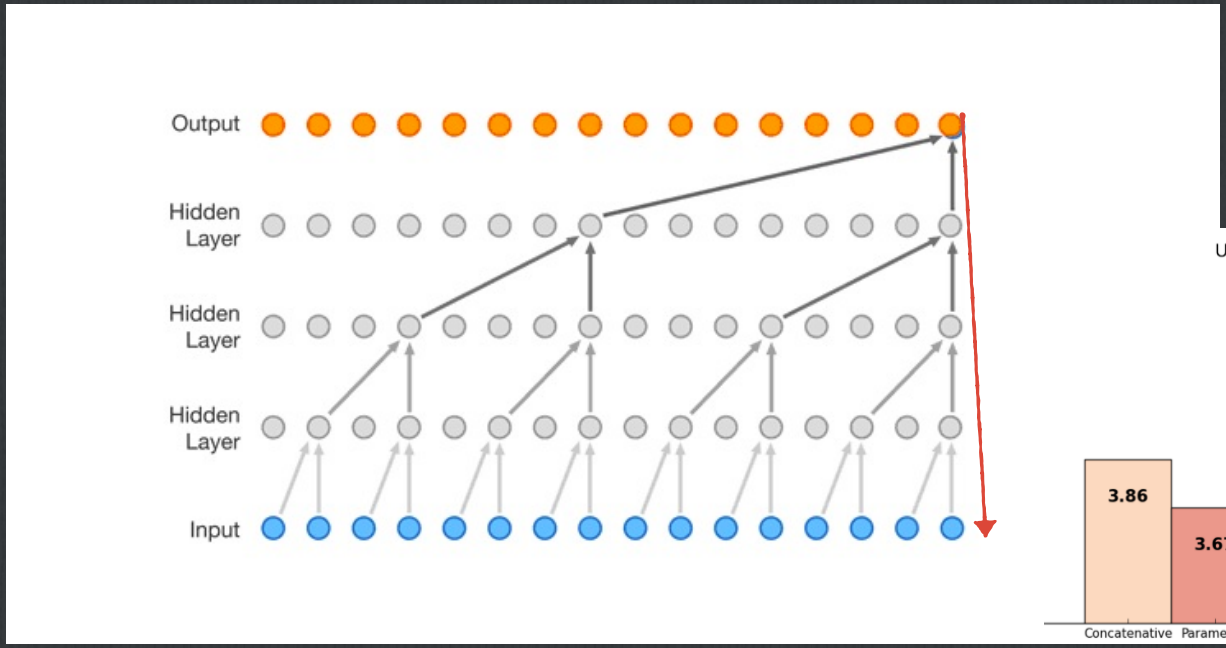


- Sound source separation with CNN

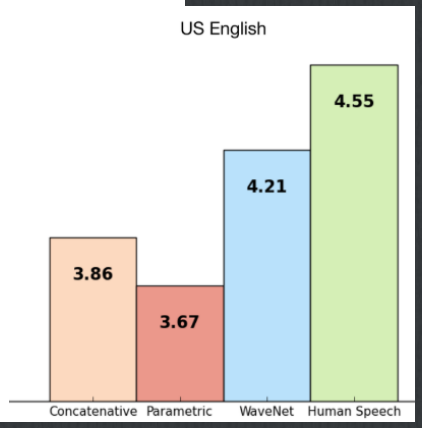


Audio synthesis

- Text to Speech. Wavenet (Google deep mind 2016)



Generates audio sample by sample



(No text input)

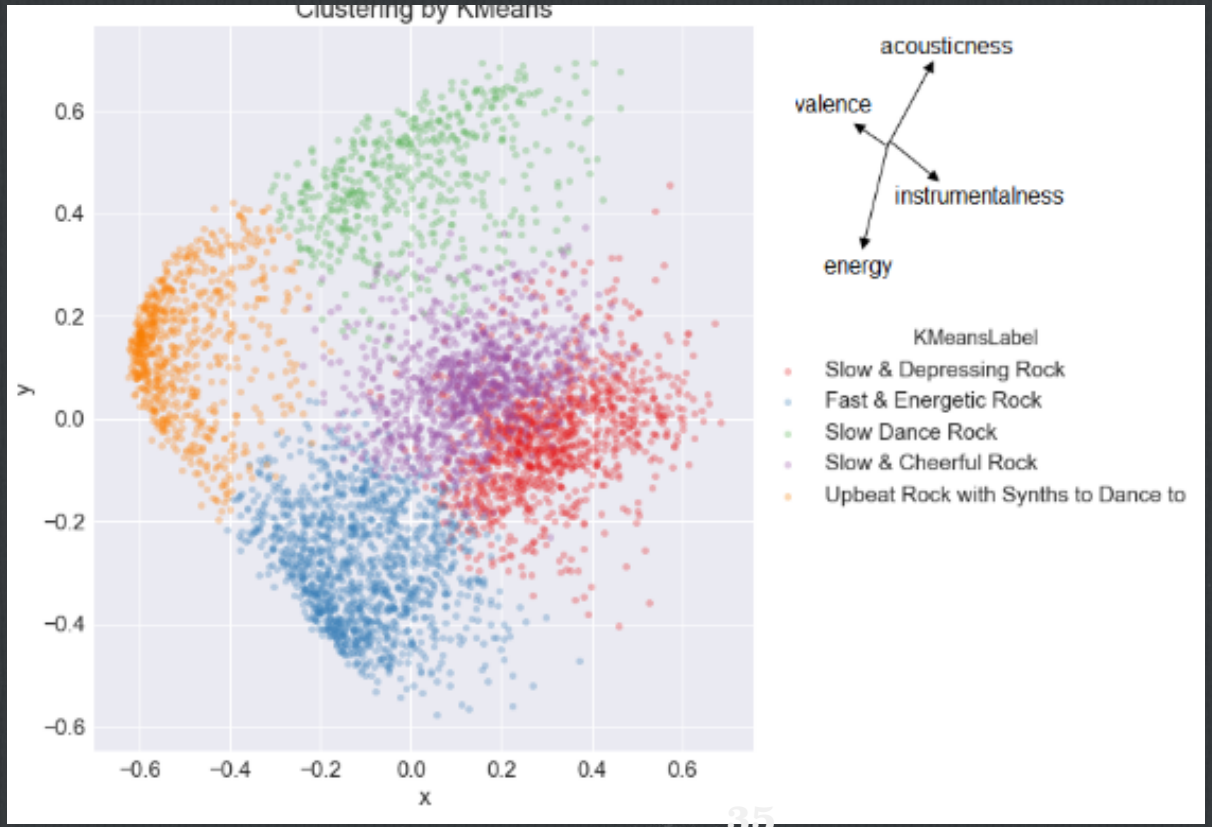
(music)

- Audio style transfer

- Voice cloning (Lyrebird.ai)
- GAN audio synthesis

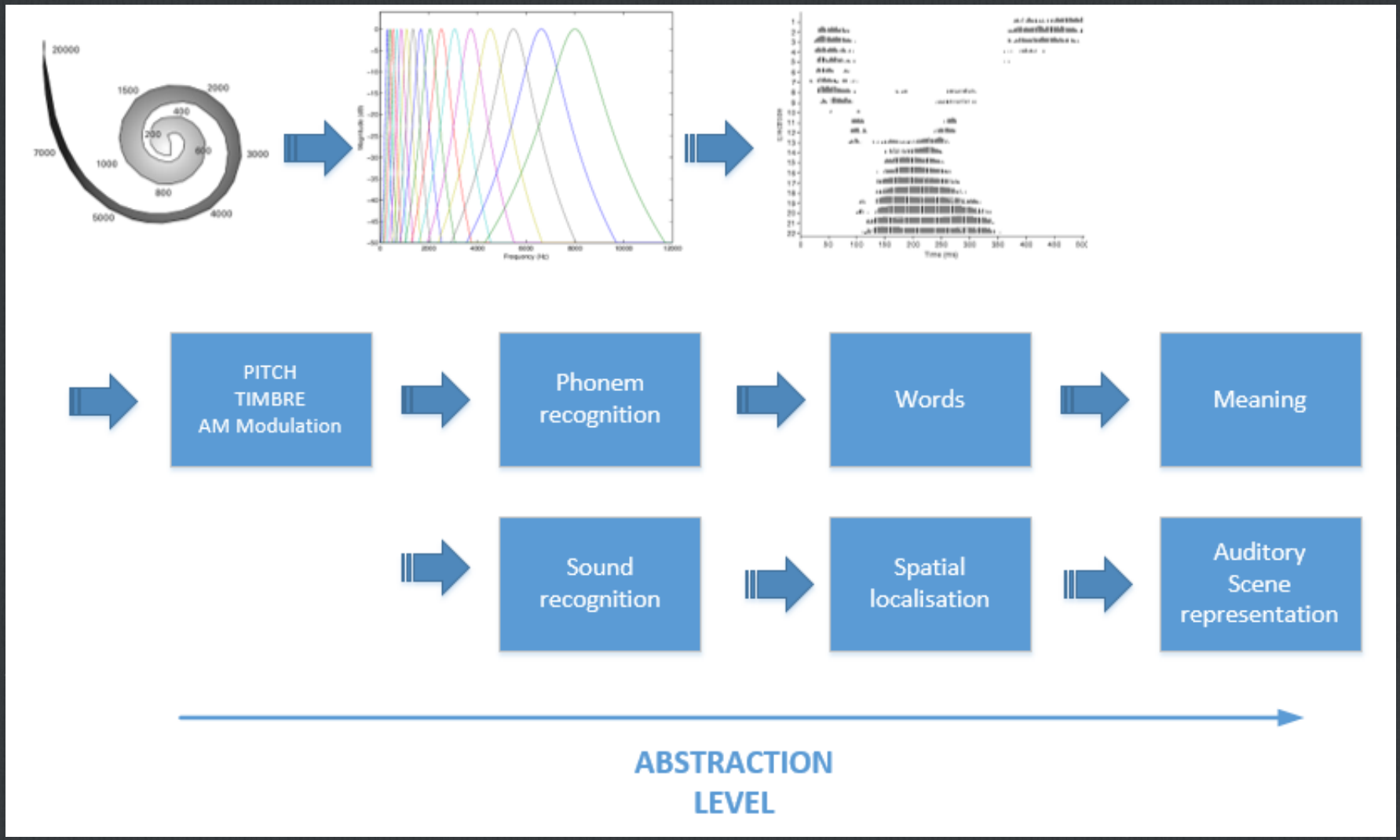
Unsupervised learning

- Music retrieval / classification
- Sound library management



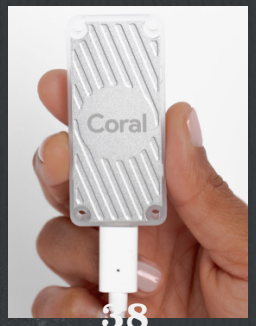
Our project: Audio cortex

- One school of thought : lots of data + complex models no feature engineering (no plateau found yet).
- Other approach :
  - keep it simpler
  - get inspiration from biological processes
  - audiology inspired feature engineering + deep learning
- Super-hearing : not limited to 2 ears, ultrasonic range e.g. in theory possible to transcribe all conversations in a crowd simultaneously


20000
7000
3000
5000
4000
PITCH
TIMBRE
AM Modulation
Phonem recognition
Words
Meaning
Sound recognition
Spatial localisation
Auditory Scene representation
ABSTRACTION LEVEL

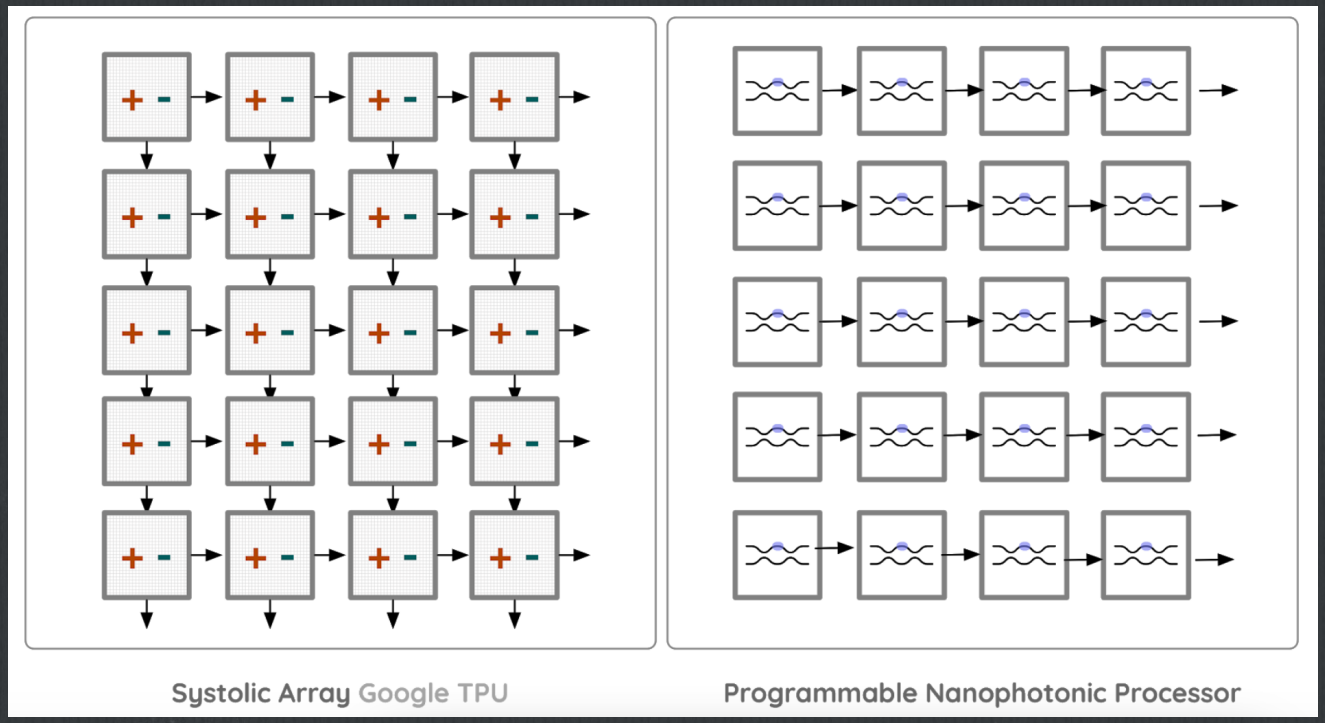
5- Tools and Hardware

- ☐ Python libraries: SciKit learn, Tensorflow, Keras, Torch..
- ☐ Matlab, Weka
- ☐ Google Colab (free GPU time), Amazon cloud, Google cloud
- ☐ GPU rig : Desktop PC + NVIDIA graphics cards
- ☐ TPU accelerator



Near future (5-10 years)

- No need for cloud processing, embedded AI in smart devices
- Photonics chips (200 THz, no heat, low power)



Tensor Processing Unit (TPU)

2D MAC array (Multiply Accumulate)

- Quantum computing ?

Conclusion

- The AI genie is out of the bottle
- Exciting times to be learning ML and AI
- Audio engineer can benefit a lot from AI and ML
- Lots of learning resources and tools available for doing machine learning (see references) and datasets


Thanks for listening !

References

- Atari Games -https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf
- https://deepmind.com/research/publications/mastering-game-go-deep-neural-networks-tree-search/
- Free book https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf

- [ ] History https://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/
- [ ] Audio synthesis https://deepmind.com/research/publications/efficient-neural-audio-synthesis
- [ ] Speech synthesis : https://deepmind.com/blog/article/wavenet-generative-model-raw-audio
- [ ] Photonics processor https://medium.com/lightmatter/the-story-behind-lightmatters-tech-e9fa0facca30

Learning resources

- Great introduction: https://medium.com/machine-learning-for-humans
- Large free audio dataset : https://research.google.com/audioset/
- Coursera courses : Machine learning (Andrew Ng), Tensorflow specialisation, Mathematics for Machine Learning: Linear Algebra + many others
- ‘Deep learning book’, Ian Goodfellow et al.
- ‘Hands on machine learning with Tensorflow and Keras’, Aurelien Géron.