DEEP LEARNING

S2DS @ Westminster University

Tim Scarfe, Ph.D

Principal Software Engineer @ Microsoft
Commercial Software Engineering (CSE)
@ecsquendor
youtube.com/machinelearningatmicrosoft

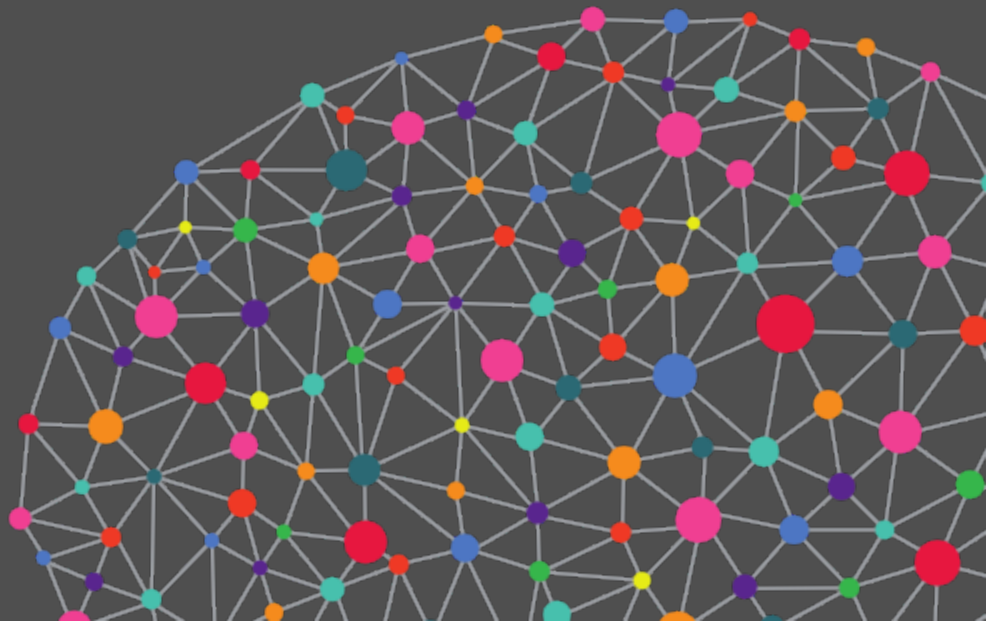
MACHINE LEARNING @ MICROSOFT

Azure

Microsoft

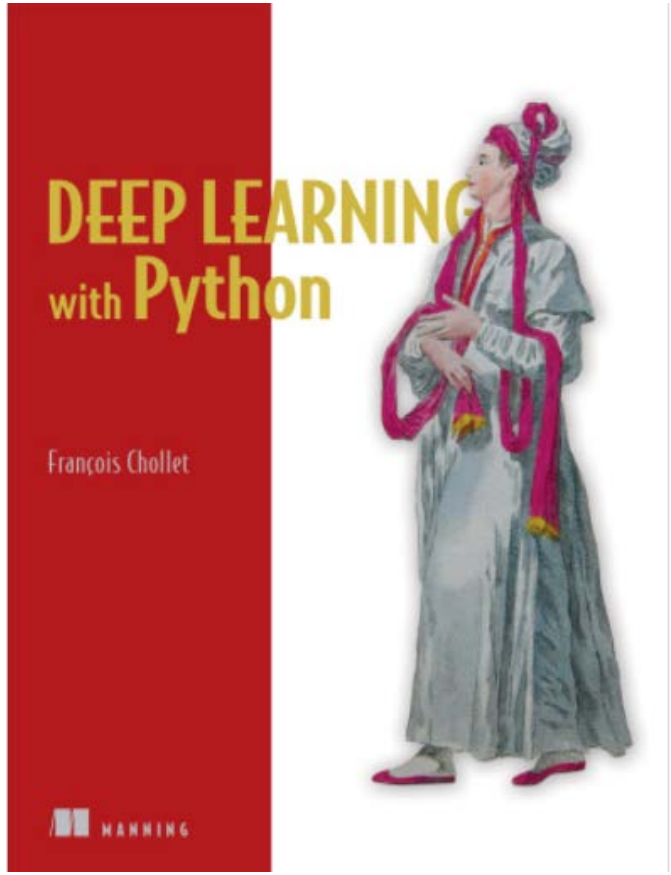
TALK OUTLINE

- Deep learning intro
- Distilled concepts of deep learning + Demo
- CNNs + Demo
- Sequence processing/NLP + Demo
- Confused.com machine learning story
- Questions



Microsoft

Francois Chollet is a living legend, please buy his book

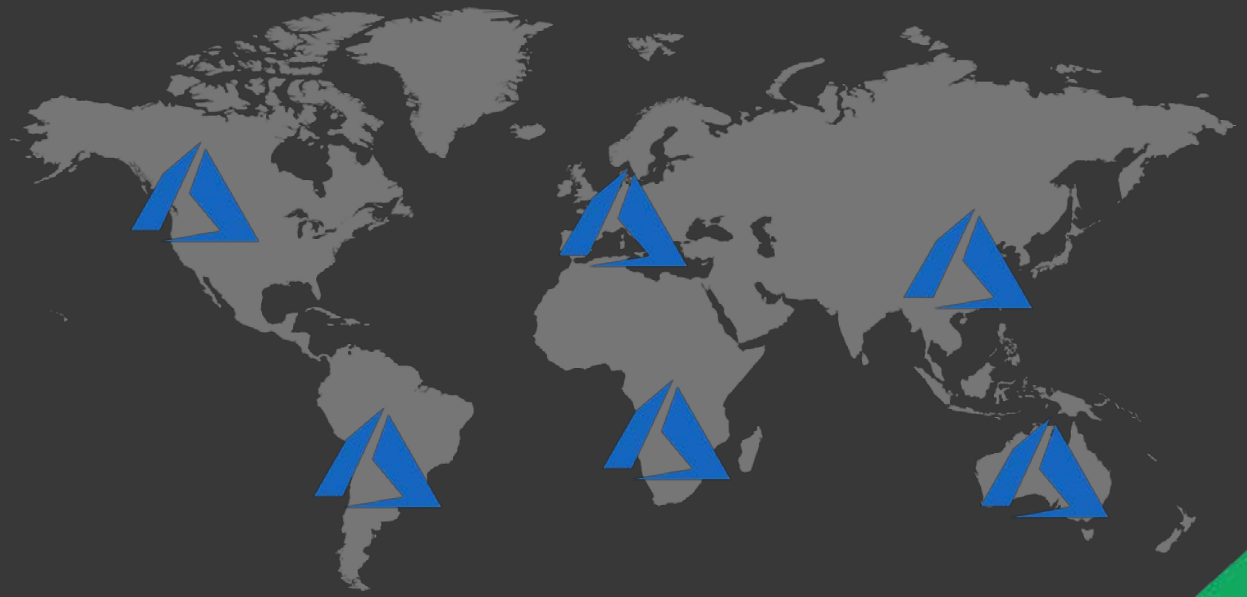


https://www.manning.com/books/deep-learning-with-python

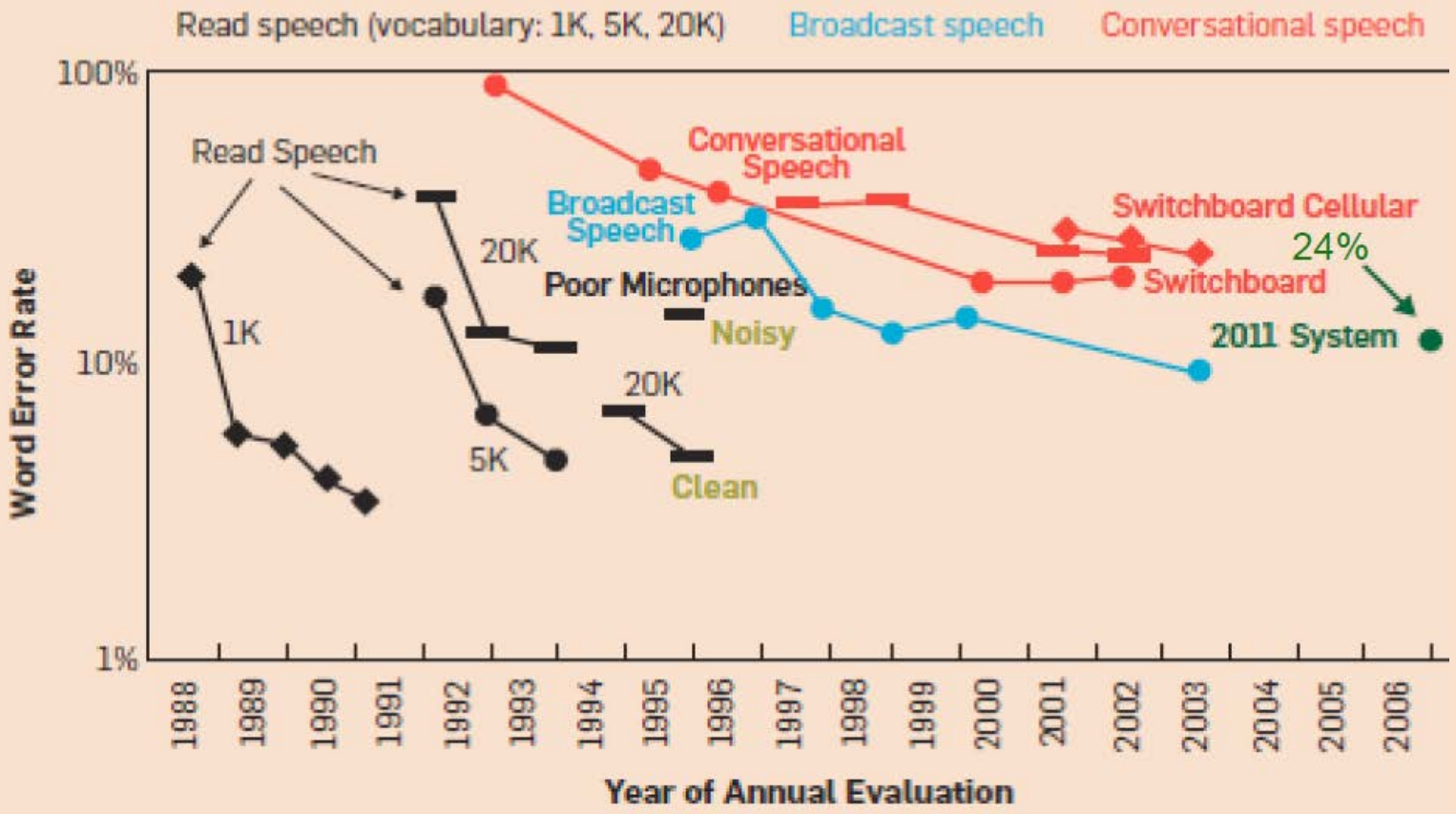
Microsoft

COMPUTE + DATA EXPLOSION = SMART AI

Microsoft



Azure

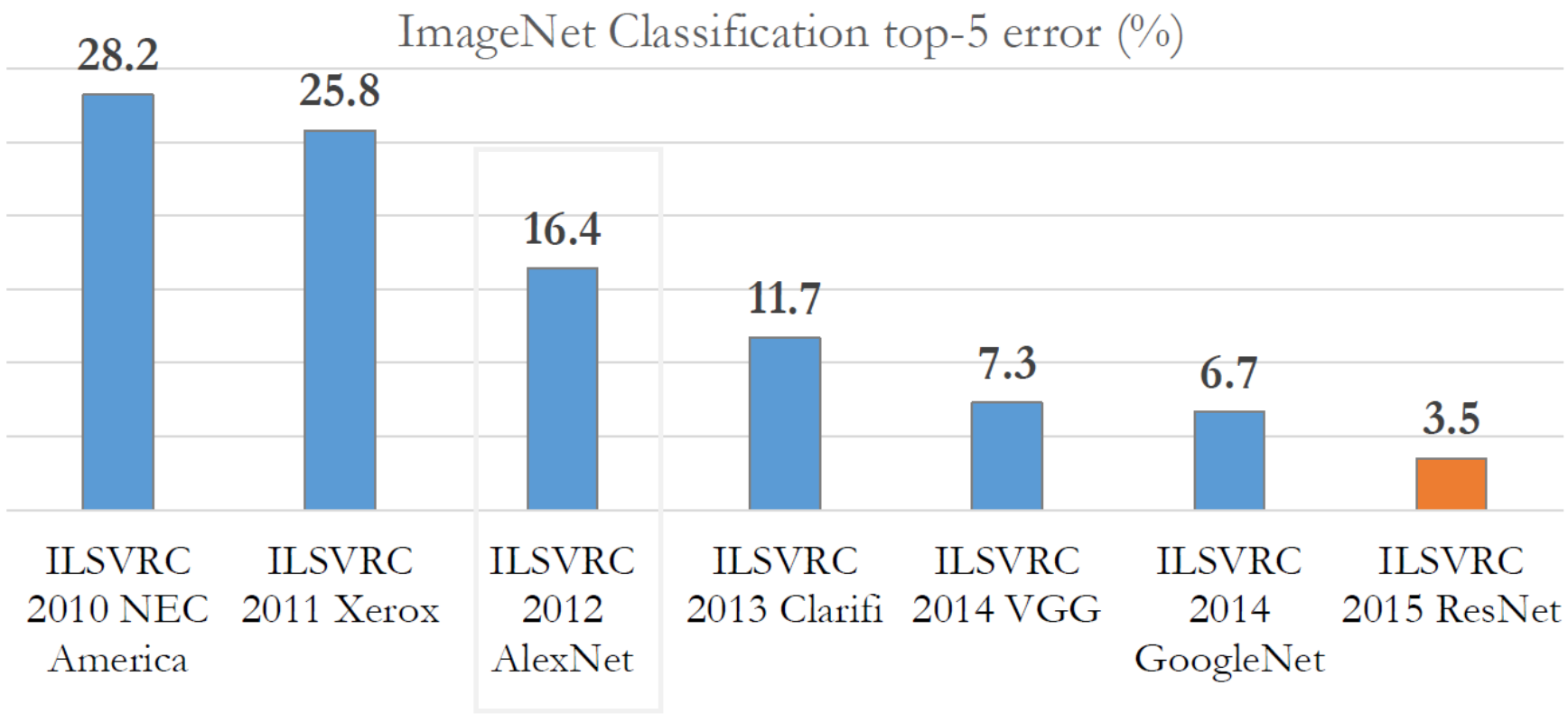


2017: ~5%!

Human error: ~5%

IMPROVEMENTS IN SPEECH RECOGNITION

IMPROVEMENTS IN COMPUTER VISION



First deep learning system

2017: ~2.2%

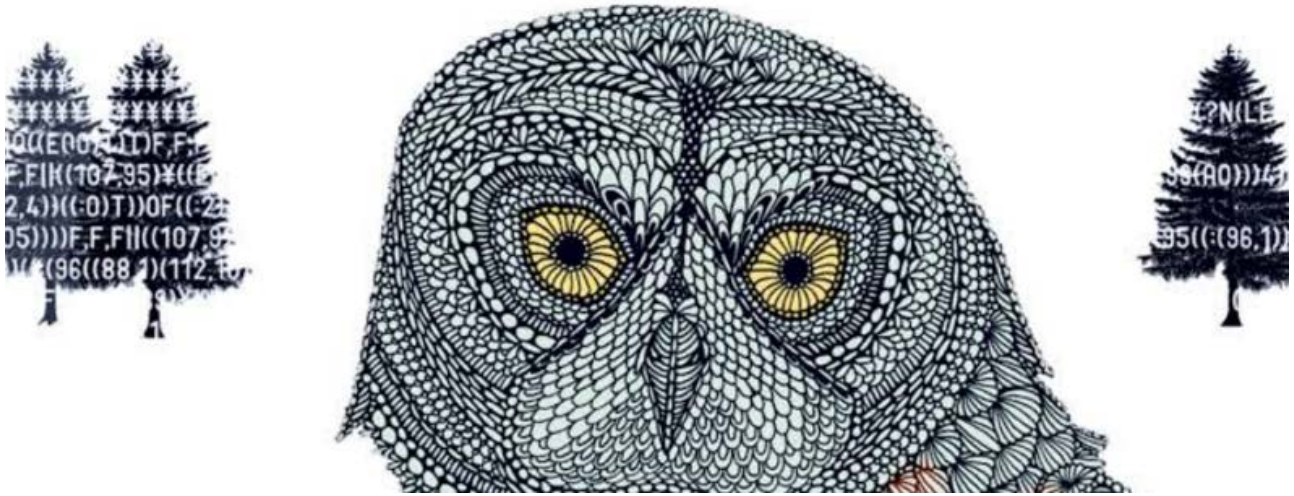
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

Microsoft

NICK BOSTROM

SUPERINTELLIGENCE

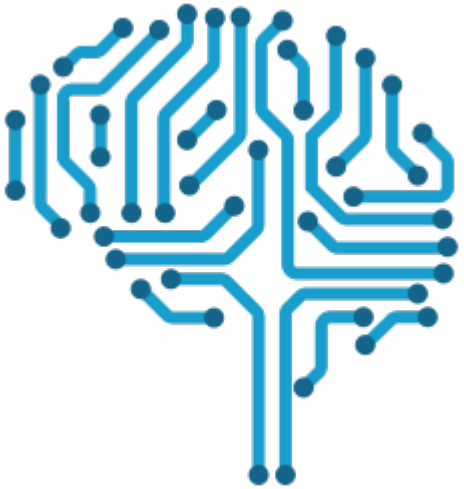
Paths, Dangers, Strategies



DON'T WORRY ABOUT GENERAL INTELLIGENCE

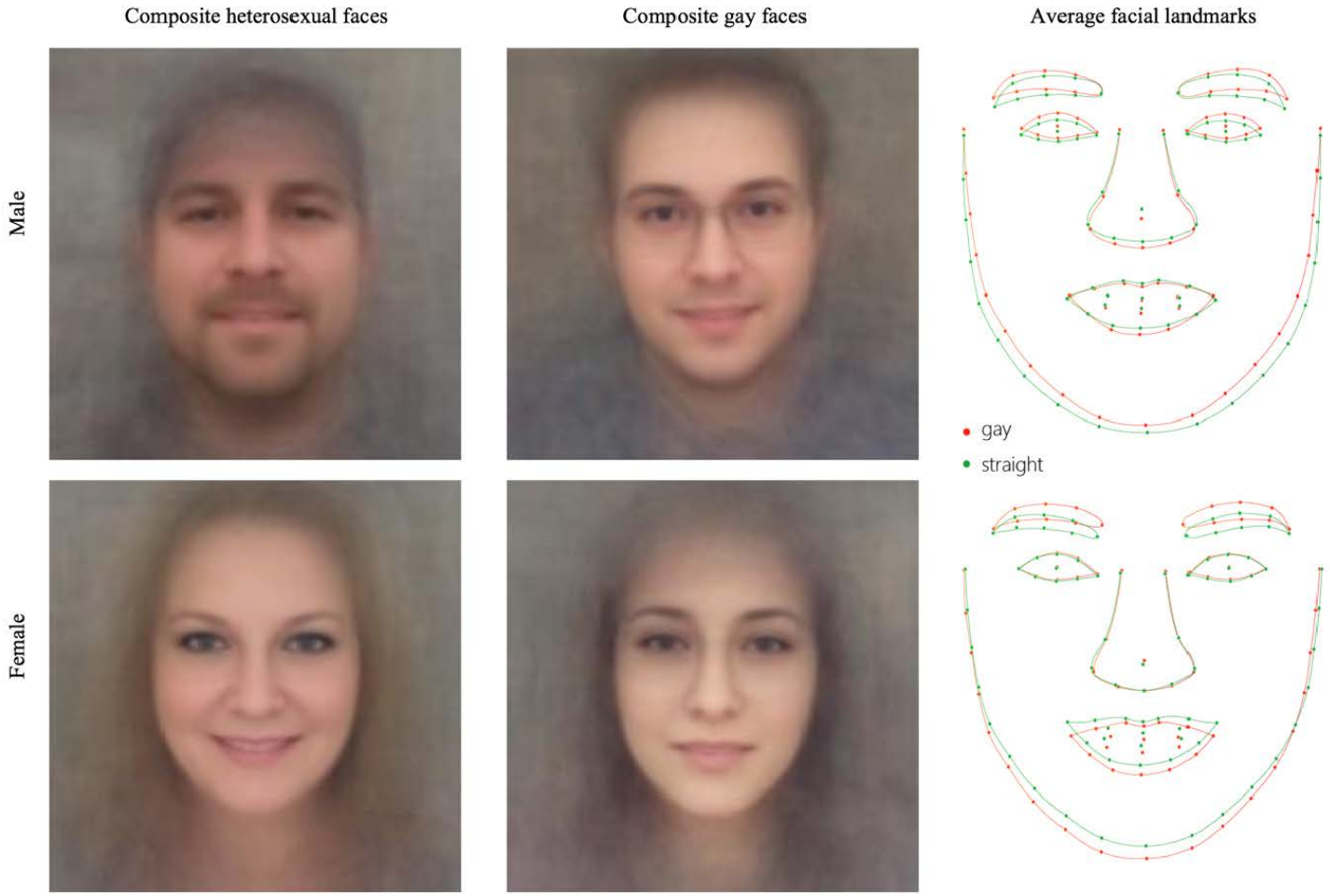
Microsoft

1. Discrimination
2. Opaque AI
3. Data is not neutral
4. Manipulating markets and consumers/voters
5. Lack of human connection
6. Engineers are not philosophers (moral reasoning)
7. Privacy
8. Innocent till proven guilty?



DO WORRY ABOUT ETHICS IN AI

https://osf.io/zn79k/



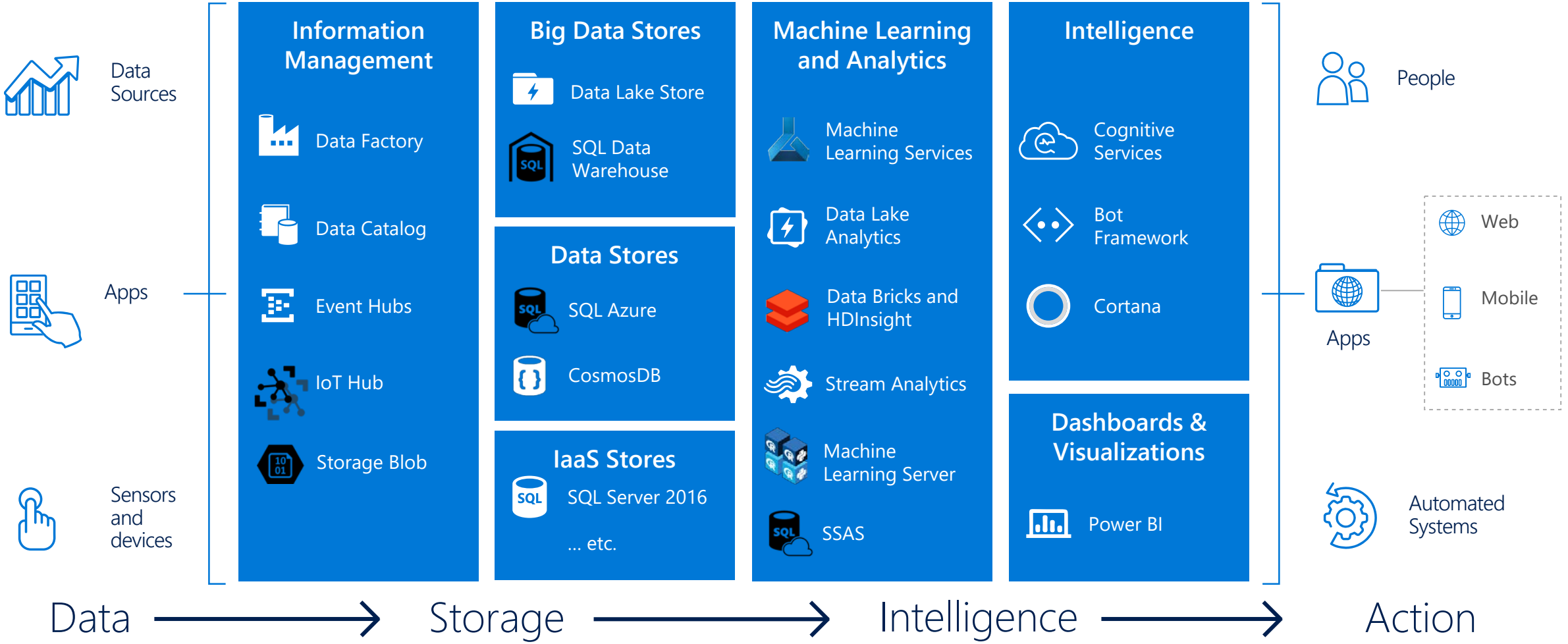
394

395 *Figure 4*. Composite faces and the average facial landmarks built by averaging faces classified as most and least likely to be gay.

2017 Wang, Y., & Kosinski, M.

PHYSIOGNOMY IS BACK?

AZURE INTELLIGENT CLOUD

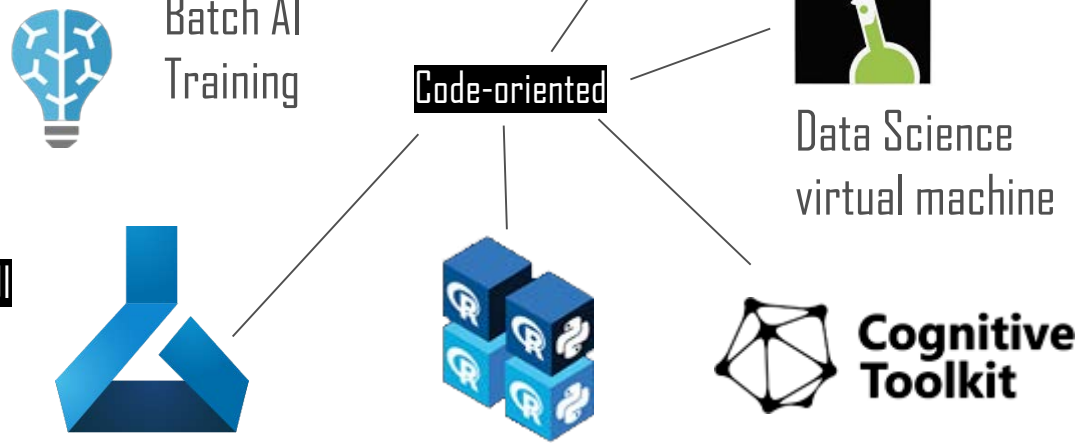


AZURE MACHINE LEARNING STACK

Microsoft

Azure Notebooks

CosmosDB / Data Lake Analytics / SQL Server



Cognitive Services
Software developer

Data Bricks
Data scientist

ML Studio
Citizen data scientist

ML Services
Data scientist

ML Server
Data scientist

CNTK
Applied Scientist

SaaS

PaaS

Private

Generative

Code-first

Containers

Notebooks

Onprem/Hybrid

IaaS

Discriminative

Framework

COGNITIVE SERVICES

Vision

Language

Speech

Search

Knowledge

Computer vision
Face
Emotion
Content Moderator
Video
Video Indexer
Cognitive Services Labs

Text analytics
Spell check
Web language model
Linguistic analysis
Translator

Speaker recognition
Speech

Web search
Image search
Video search
News search
Autosuggest

Academic knowledge
Entity linking service
Knowledge exploration
Recommendations
QnA maker

Custom
Vision Service

Custom
Language Understanding

Custom
Speech Service

Custom
Search

Custom
Decision Service

MACHINE LEARNING PORTFOLIO

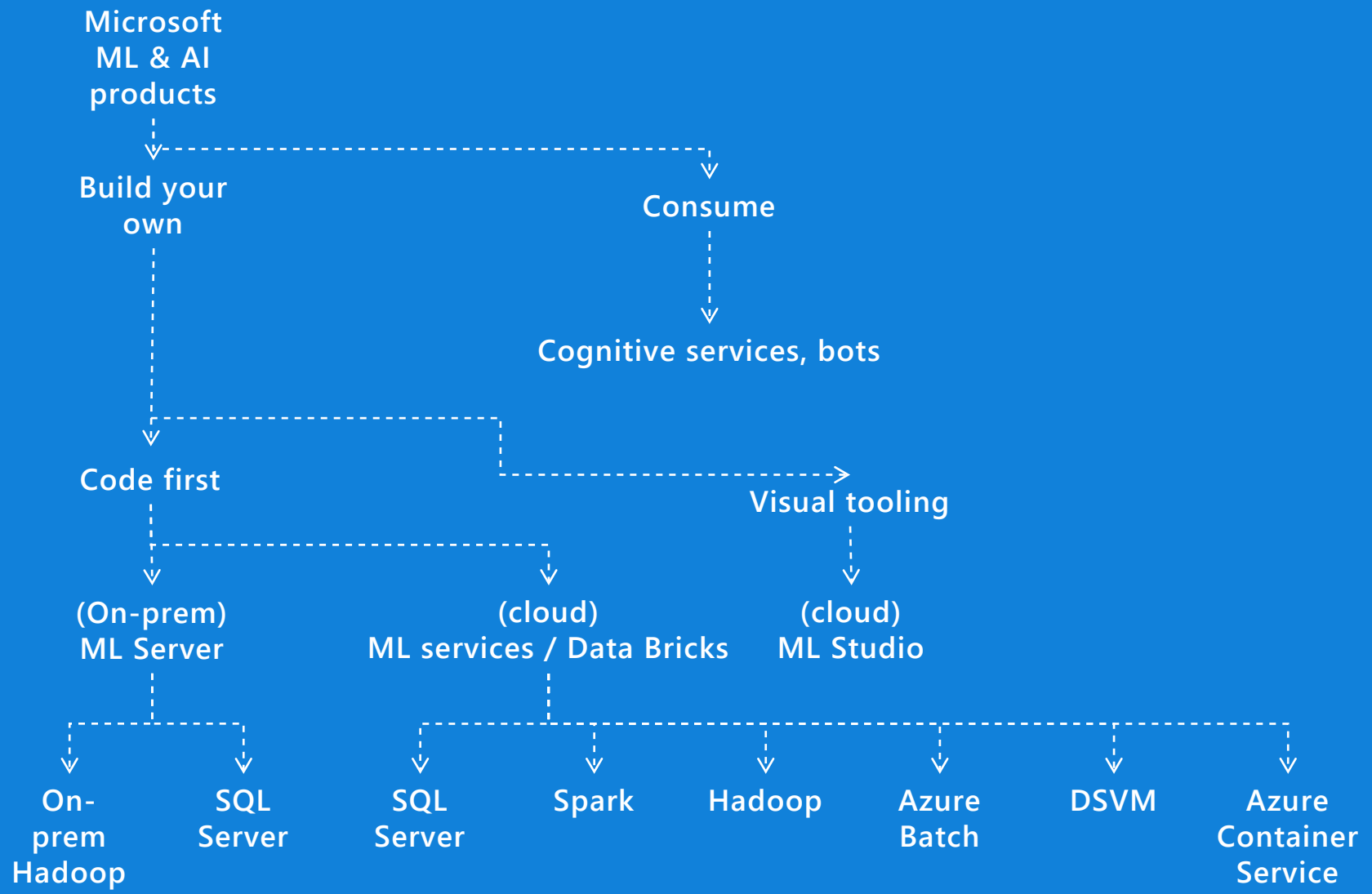
Microsoft

Build your own or consume pre-trained models?

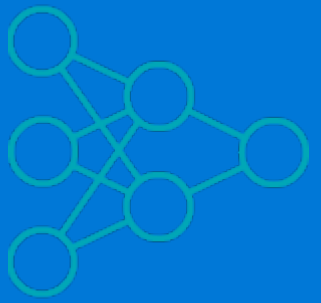
Which experience do you want?

Deployment target

What engine(s) do you want to use?



WHAT IS MACHINE LEARNING?



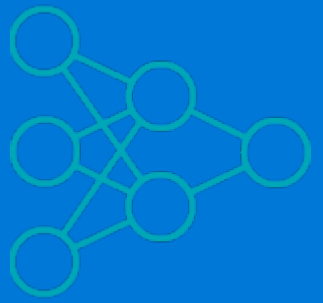
A machine learning from previous experience (data) so that it can perform some task better *based on that experience*.

Microsoft

MACHINE LEARNING != DATA SCIENCE

The data scientist is trying to understand the data, the ML person is building a working model which may not be interpretable.

INDUCTIVE INFERENCE



ML algorithms often "induce" a generic and separate artefact or "decision function", learned from data.

TYPES OF MACHINE LEARNING

Supervised

Tell the machine the correct answer

- *Images/Sequences*
- *Regression*
- *Classification*

Unsupervised

Tell the machine nothing, let it observe the world

- *Anomaly detection*
- *Embeddings*
- *Clustering*
- *Dimensionality Reduction*

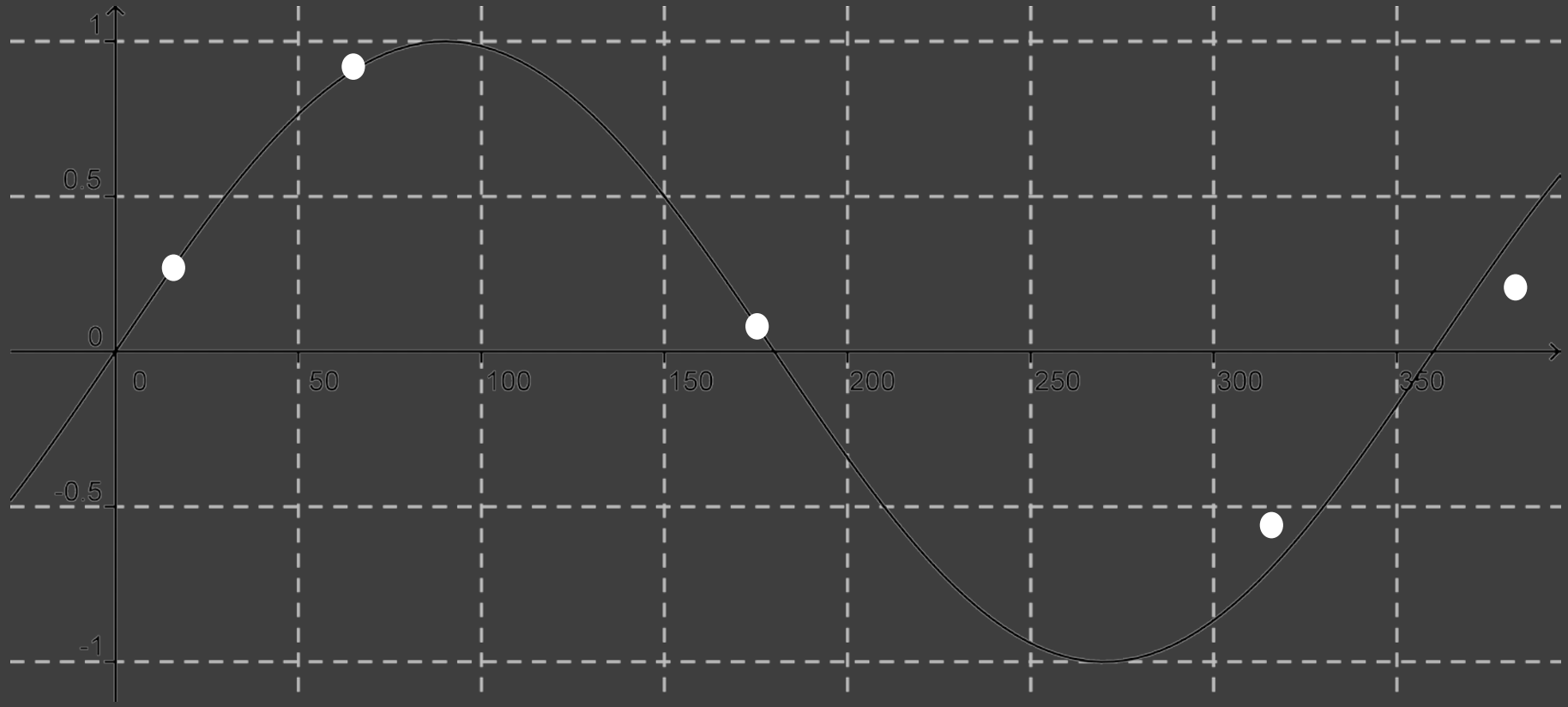
Reinforcement Learning

Learning from delayed rewards

- *Agent-based learning*
- *Autonomous cars*

On-line/Offline

F(X)=

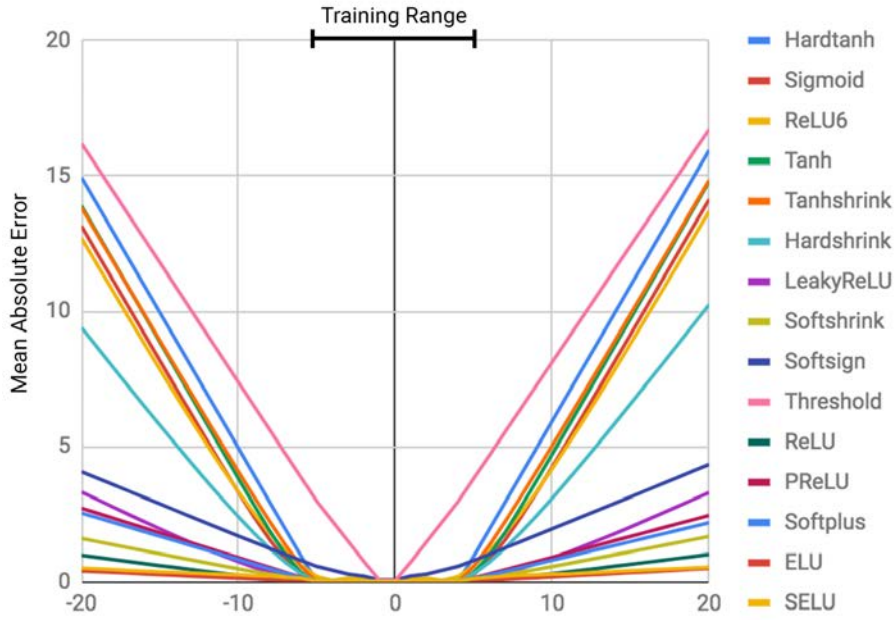


Machine learning approximates a function, which we may only have a few examples of.

A SIMPLE SINSOIDAL FUNCTION

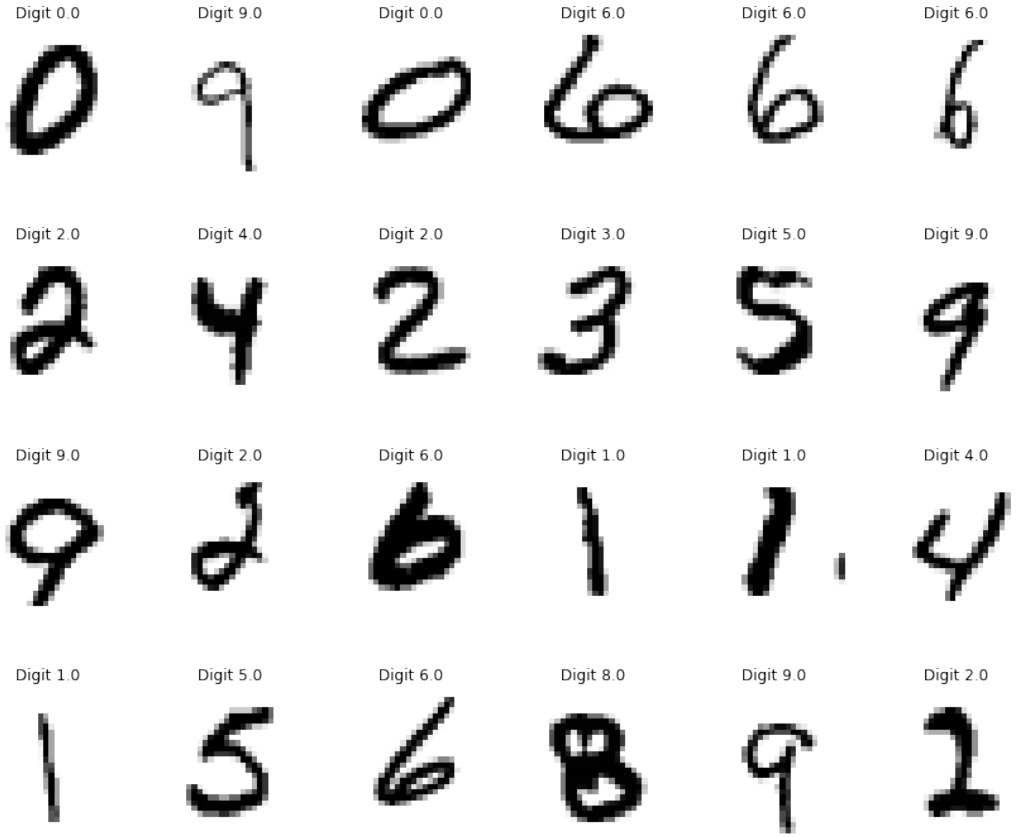
INTERPOLATION VS EXTRAPOLATION

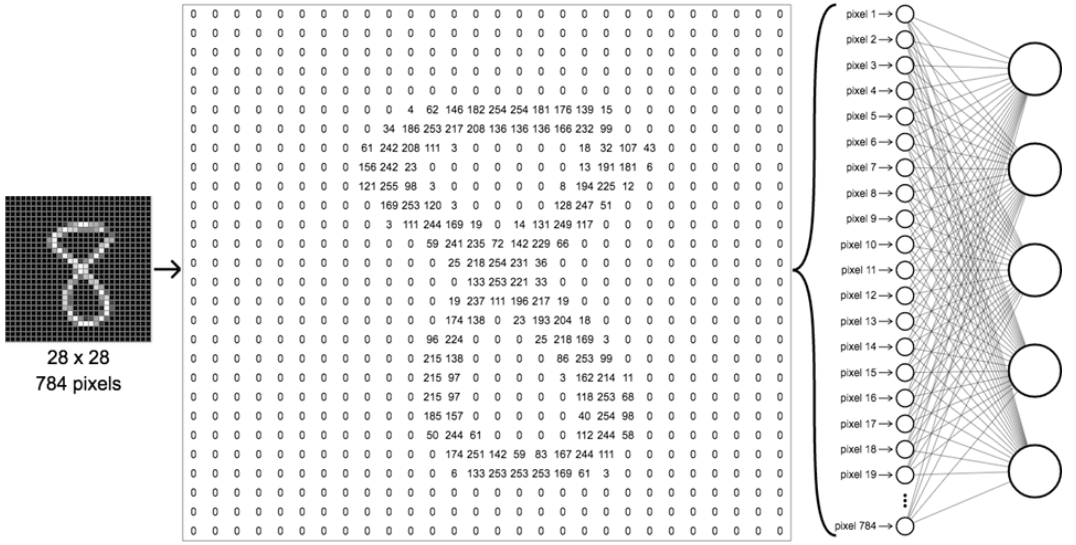
ML algorithms perform very, very badly on data out of range. They don't generalise as well as you think. Trask et al 2018 (DeepMind)



Microsoft

MNIST DIGIT CLASSIFICATION



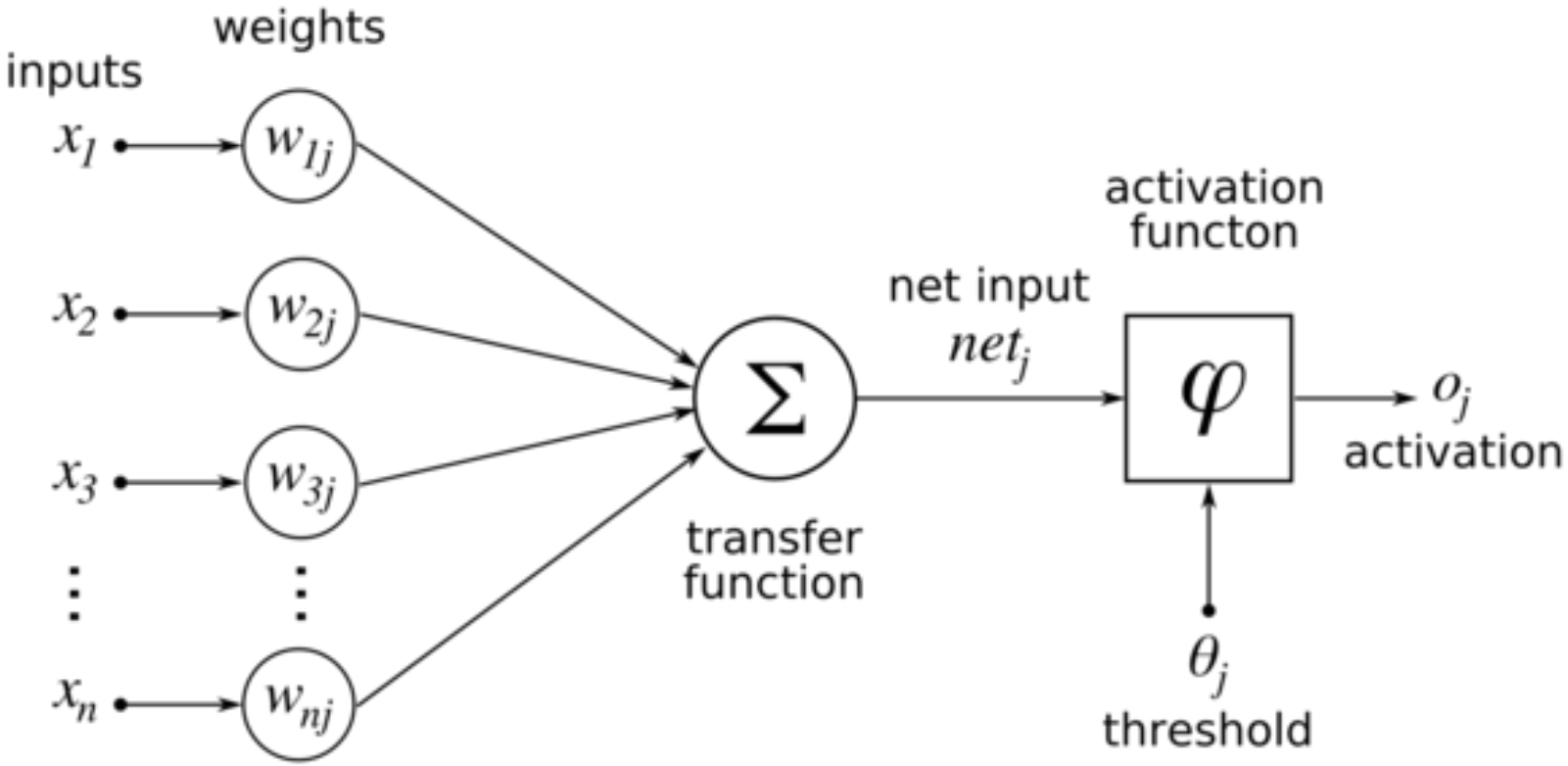


Represent input image as a vector $\mathbf{x} \in \mathbb{R}^{784}$

Learn a classifier $f(\mathbf{x})$ such that,

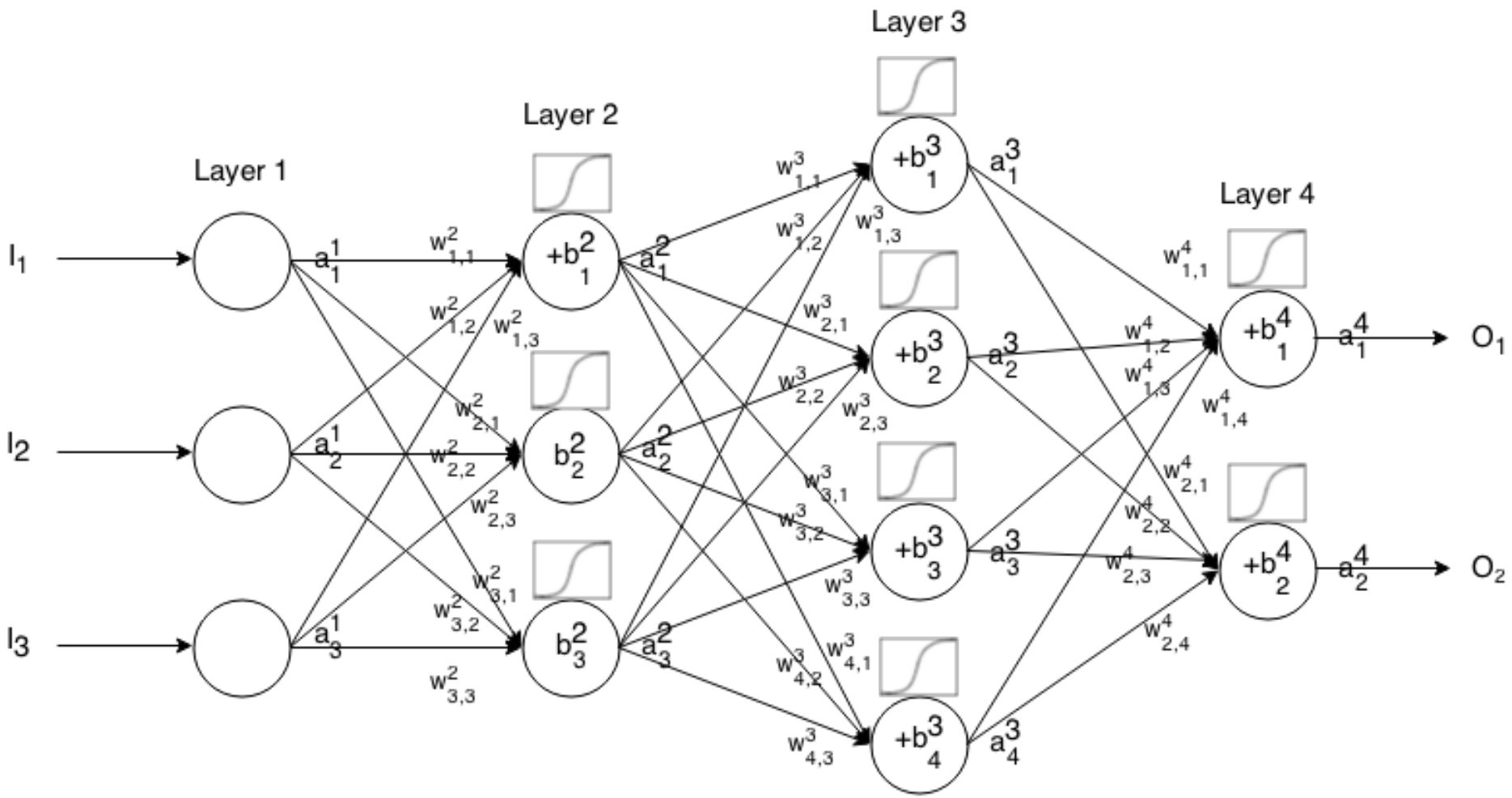
$$f : \mathbf{x} \rightarrow \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$$

WHAT IS A NEURAL NETWORK?

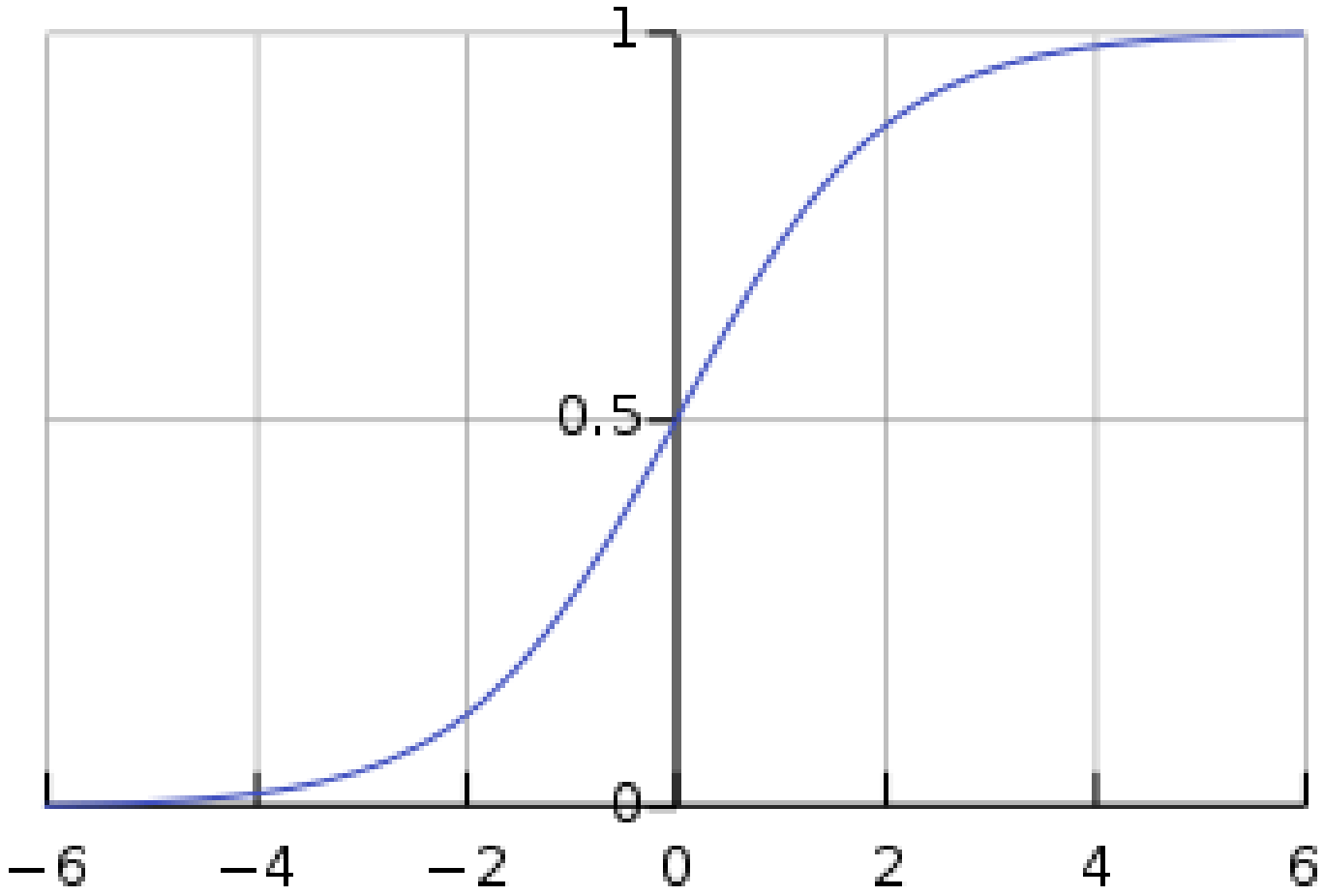


Microsoft

WHAT IS A DEEP NEURAL NETWORK?

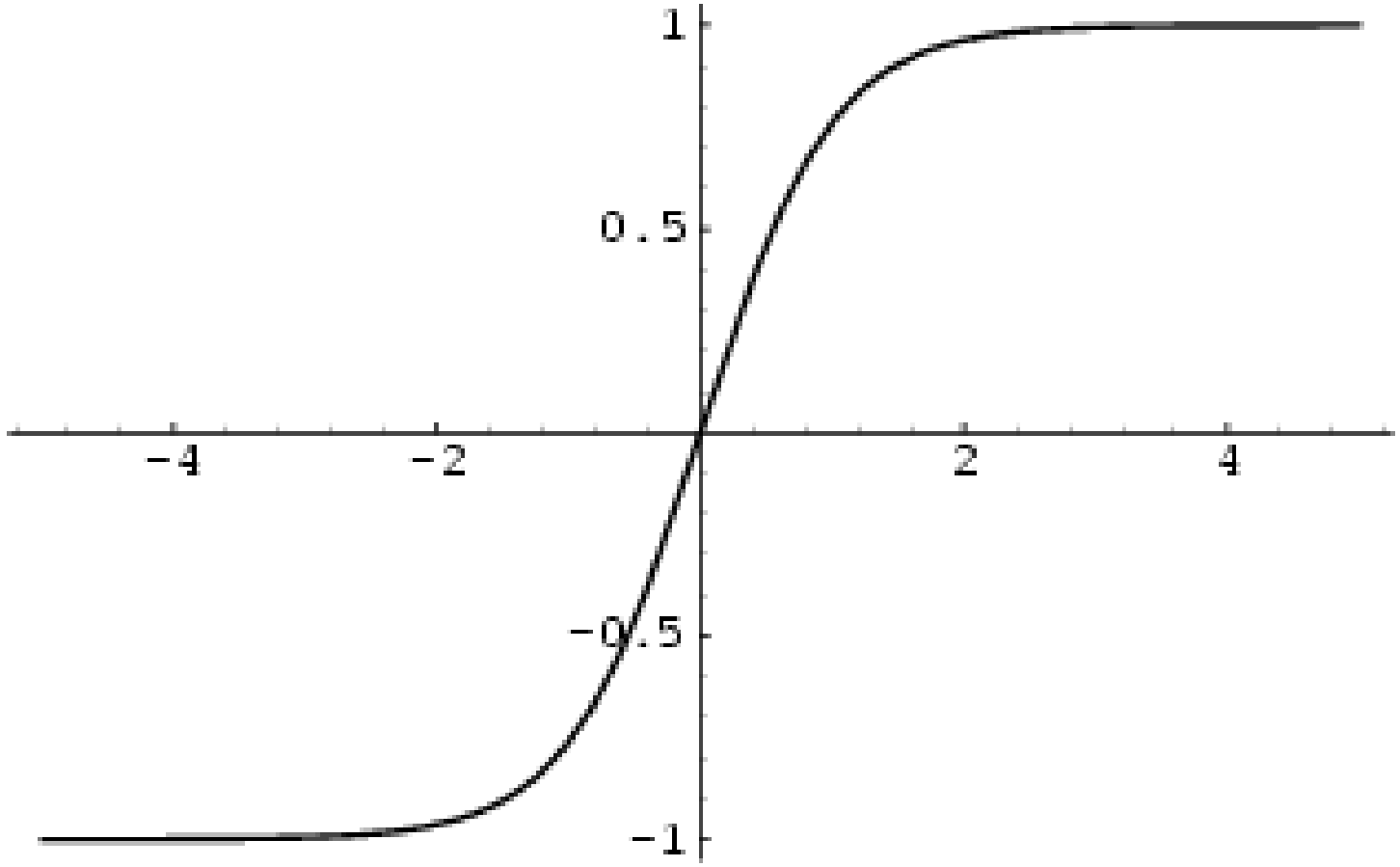


Microsoft



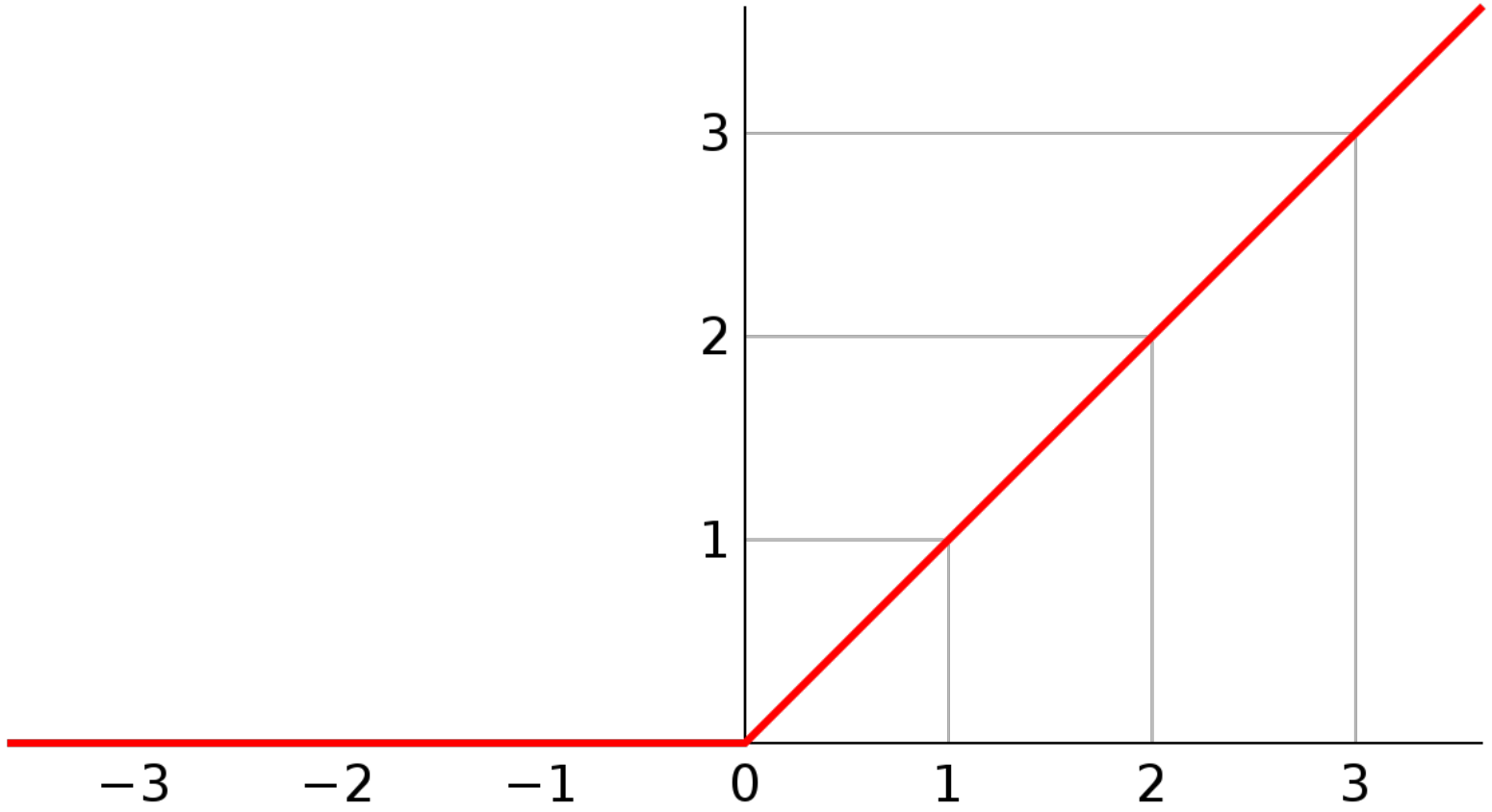
$$\frac{e^x}{e^x + 1}$$

SIGMOID SQUASHING FUNCTION

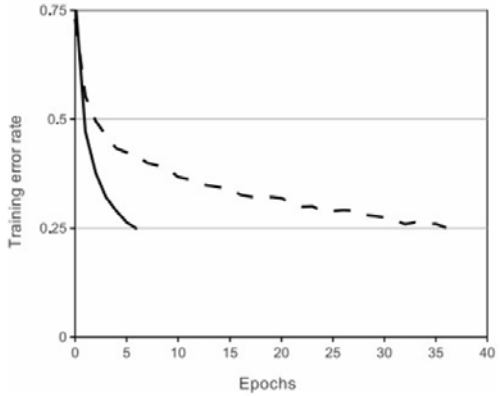


$$\frac{1 - e^{-2x}}{1 + e^{-2x}}$$

TANH SQUASHING FUNCTION



$$f(x) = x^+ = \max(0, x)$$



RELU SQUASHING FUNCTION

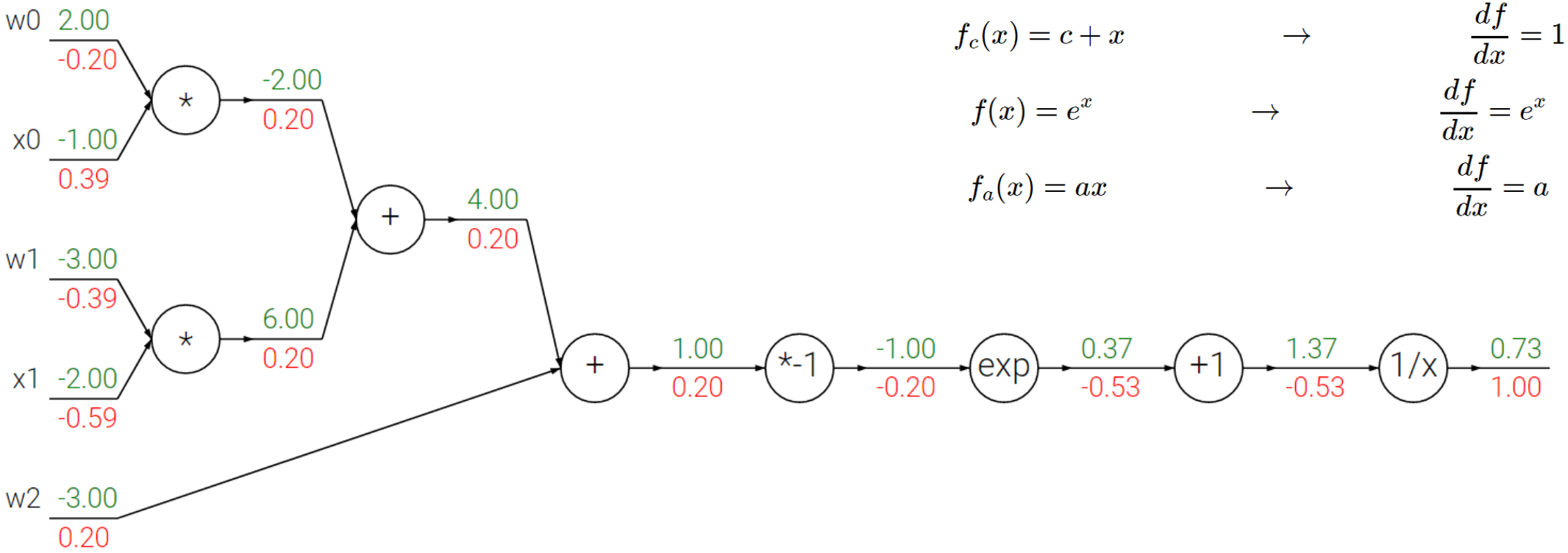
BACKPROPAGATION

$$f(x) = \frac{1}{x} \quad \rightarrow \quad \frac{df}{dx} = -1/x^2$$

$$f_c(x) = c + x \quad \rightarrow \quad \frac{df}{dx} = 1$$

$$f(x) = e^x \quad \rightarrow \quad \frac{df}{dx} = e^x$$

$$f_a(x) = ax \quad \rightarrow \quad \frac{df}{dx} = a$$



$$\frac{\partial E}{\partial w_{jk}}$$

Example circuit for a 2D neuron with a sigmoid activation function. The inputs are [x0,x1] and the (learnable) weights of the neuron are [w0,w1,w2]. As we will see later, the neuron computes a dot product with the input and then its activation is softly squashed by the sigmoid function to be in range from 0 to 1.

Microsoft

$$\frac{\partial E}{\partial w_{jk}}$$

$$\Delta w_{jk} = \eta * [\, x_j * (o_k - t_k) * o_k * (1 - o_k) \,]$$

learning rate

error e_k

derivative of output activation $\varphi_k{}'$

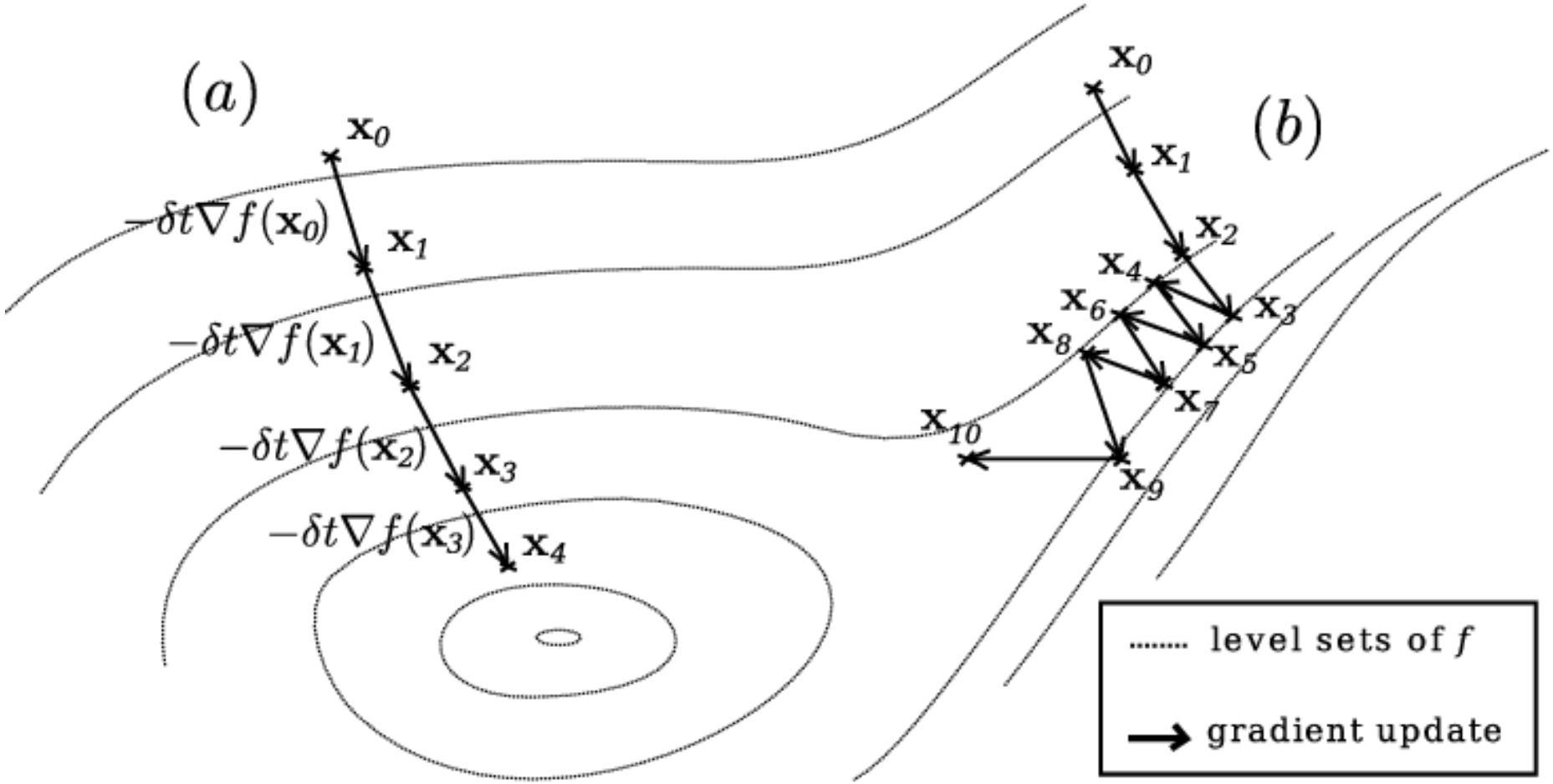
signal δ_k

WEIGHT UPDATE

```
loop maxEpochs times
  for-each training item
    get target values
    compute output values
    compute the gradient of each weight
    use gradient to compute delta for each weight
    update each weight using its delta
  end-for
end-loop
```

BACKPROP ALGORITHM

OPTIMIZATION/GRADIENT DESCENT

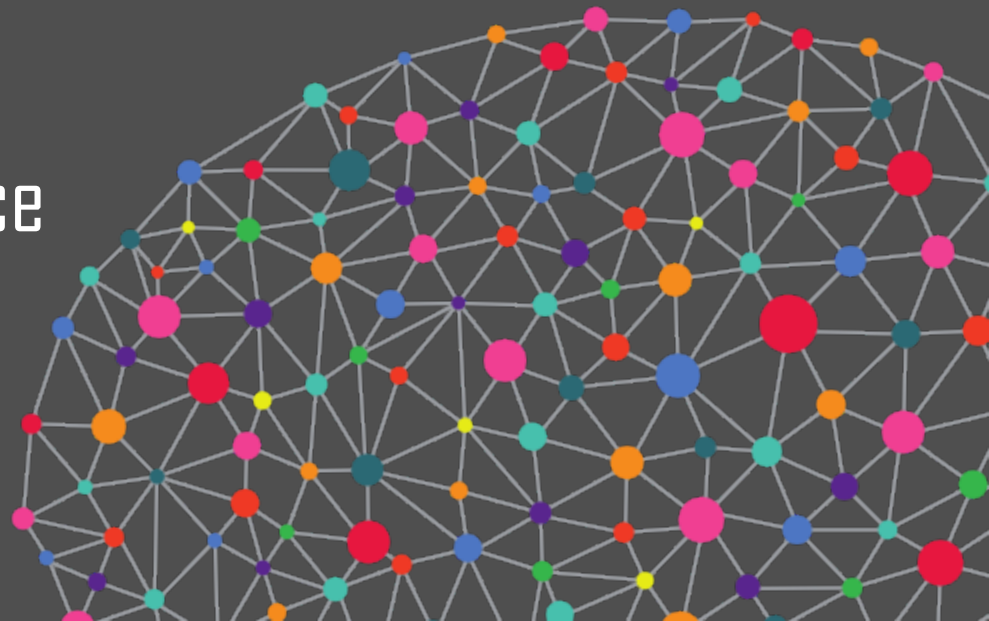


WEIGHT UPDATES IN EPOCHS / TRAINING DATA SPLIT INTO MINIBATCHES

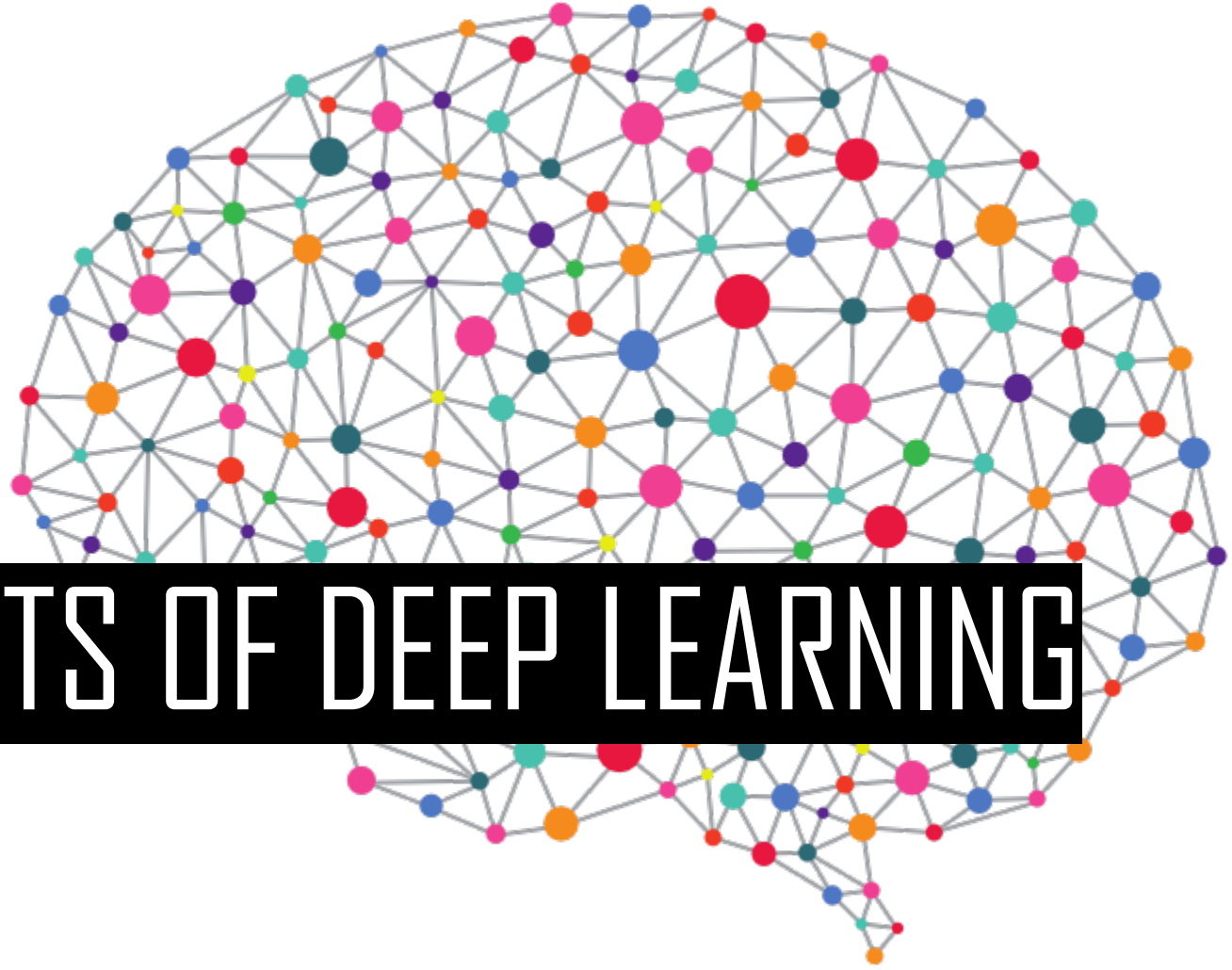
Microsoft

UNIVERSAL MACHINE LEARNING PROCESS

- Define problem
- Define success
- Validation process
- Vectorise/normalize data
- Develop naïve model
- Refine model based on validation performance



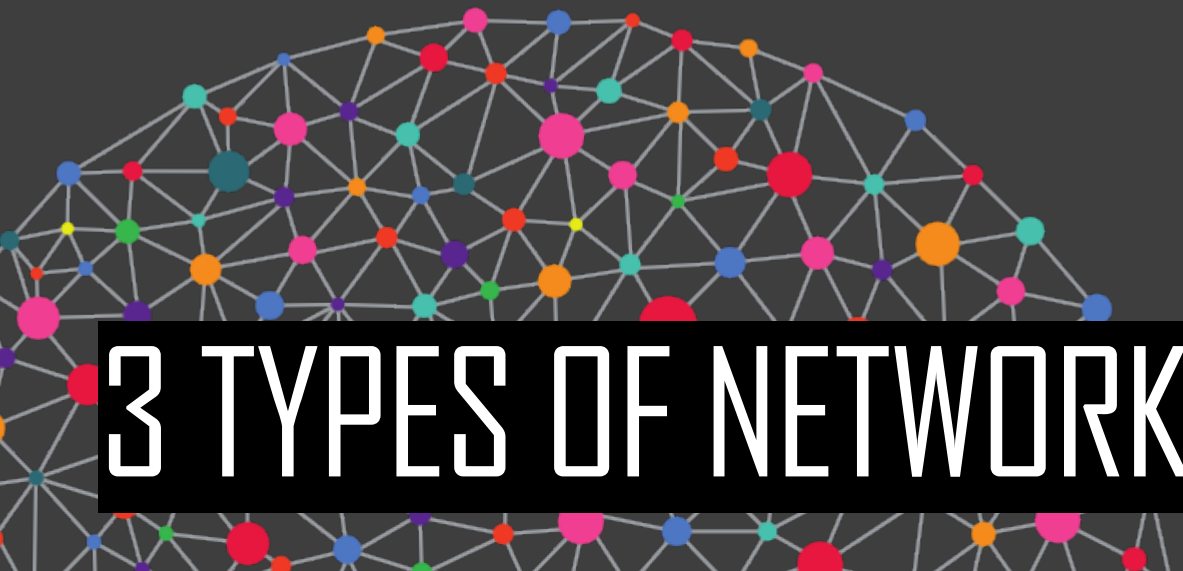
Microsoft



DISTILLED CONCEPTS OF DEEP LEARNING

Microsoft

1. DENSELY CONNECTED NETWORKS
2. CONVOLUTIONAL NEURAL NETWORKS (MODEL SPACE)
3. RECURRENT NEURAL NETWORKS (MODEL TIME)



3 TYPES OF NETWORK ARCHITECTURE

Microsoft

- The networks have **many levels** of depth
- Machine learns a hierarchy of representations i.e. **"representation learning"**
- **Less feature extraction** required (still need transposition, normalization etc)

Traditional ML

Hand crafted features

Feature Extractor → Trainable Classifier

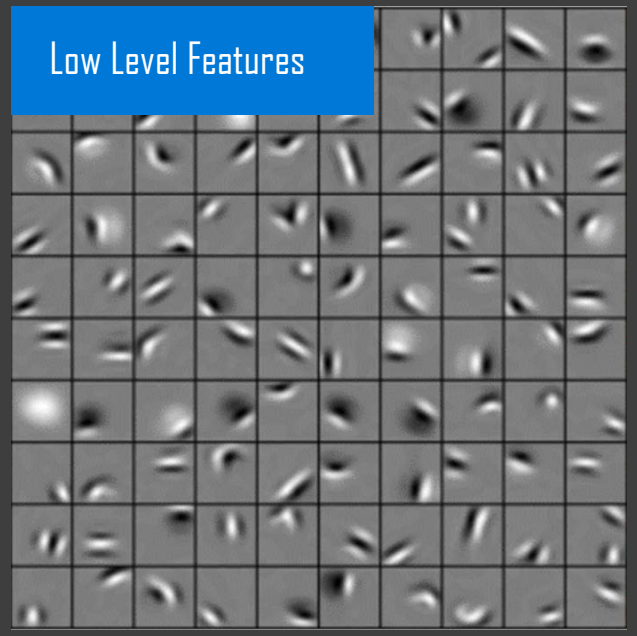
Deep Learning

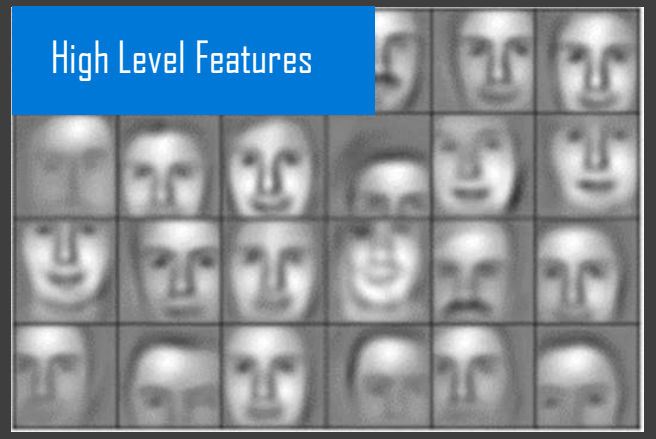
Representations are hierarchical and trained automatically

Low Level Features → Mid Level Features → High Level Features → Trainable Classifier

ENTIRE MACHINE IS TRAINABLE

Microsoft

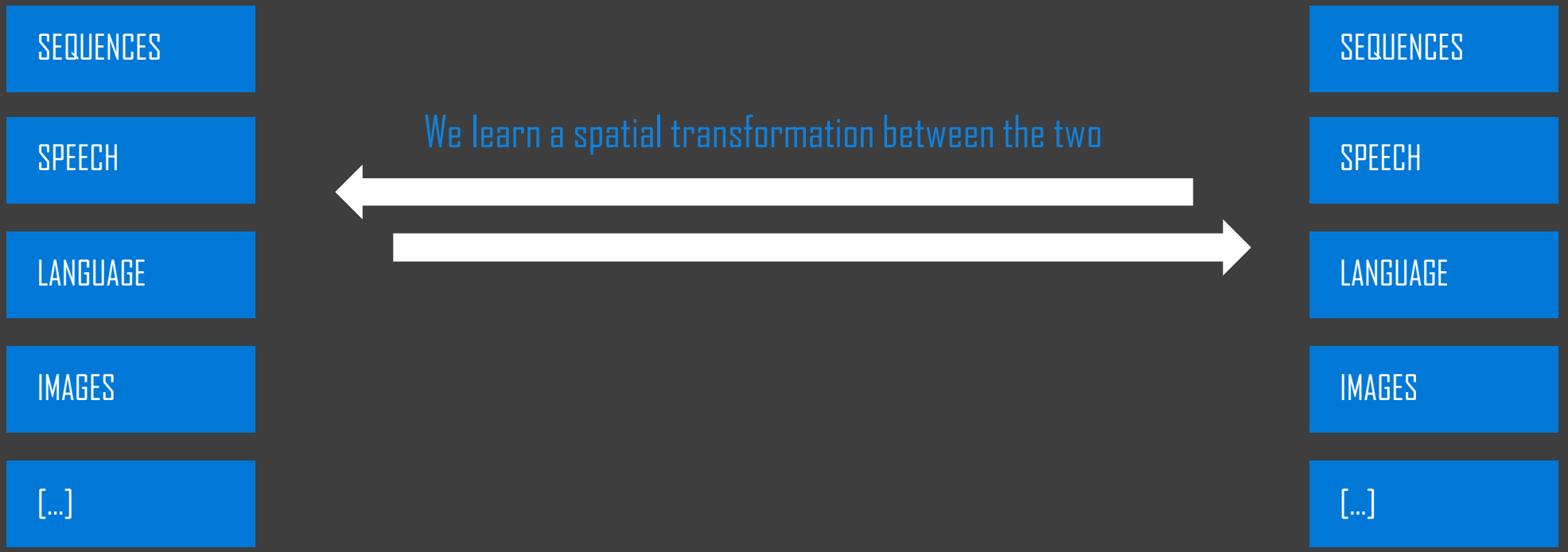




REPRESENTATIONS ARE LEARNED AUTOMATICALLY

Microsoft

- Unlike other shallow ML algorithms; you can map *between data domains*
- **Generative** models are possible, not just **discriminative**

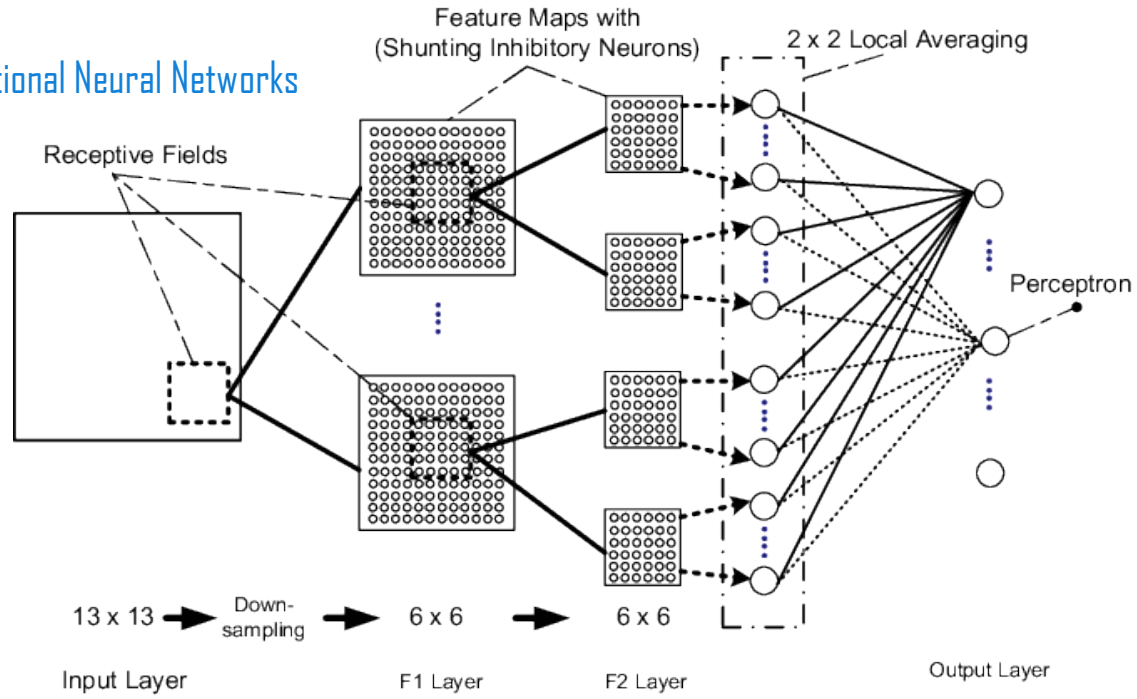


UNIVERSAL FUNCTION APPROXIMATORS

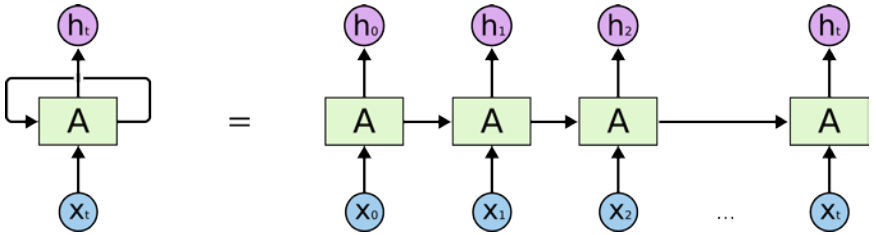
Unlike other algorithms, NNs can natively encode useful and obvious relationships in the data domain

- Local spatial dependencies (vision) i.e. CNNs
- Time dependencies (language, speech) i.e. RNNs

Convolutional Neural Networks



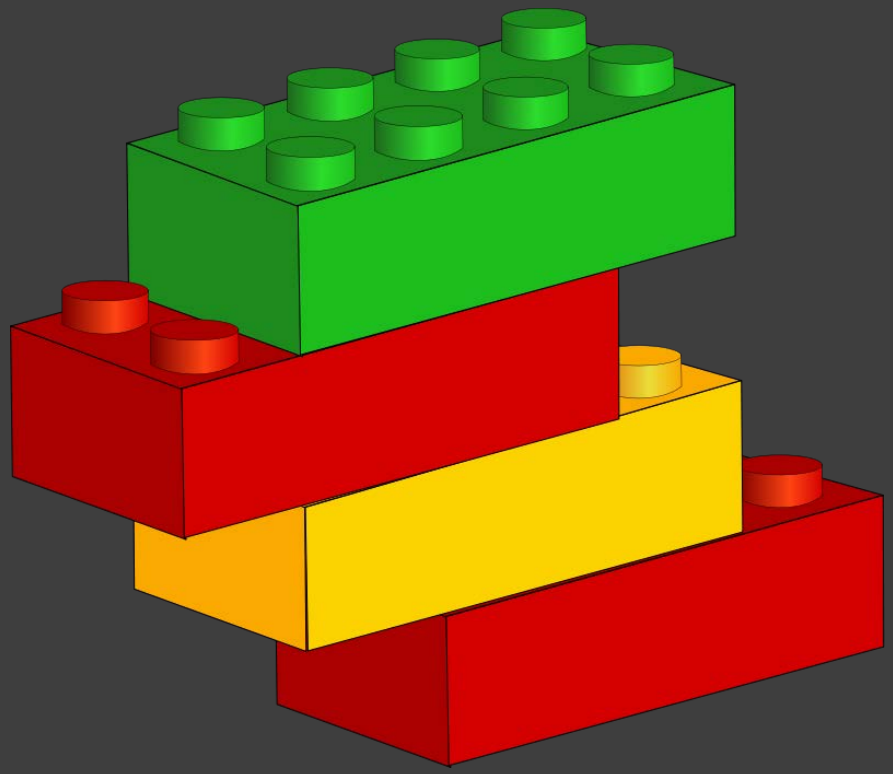
Recurrent Neural Networks



NATIVE DATA-DOMAIN FEATURES

Microsoft

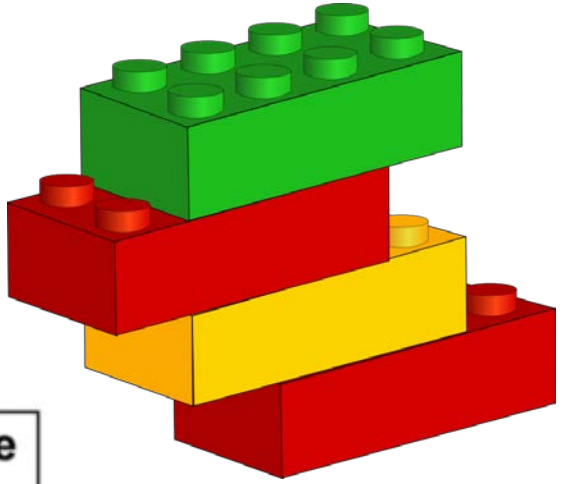
- Composability
- ML is becoming a form of software development
- ML models are like software

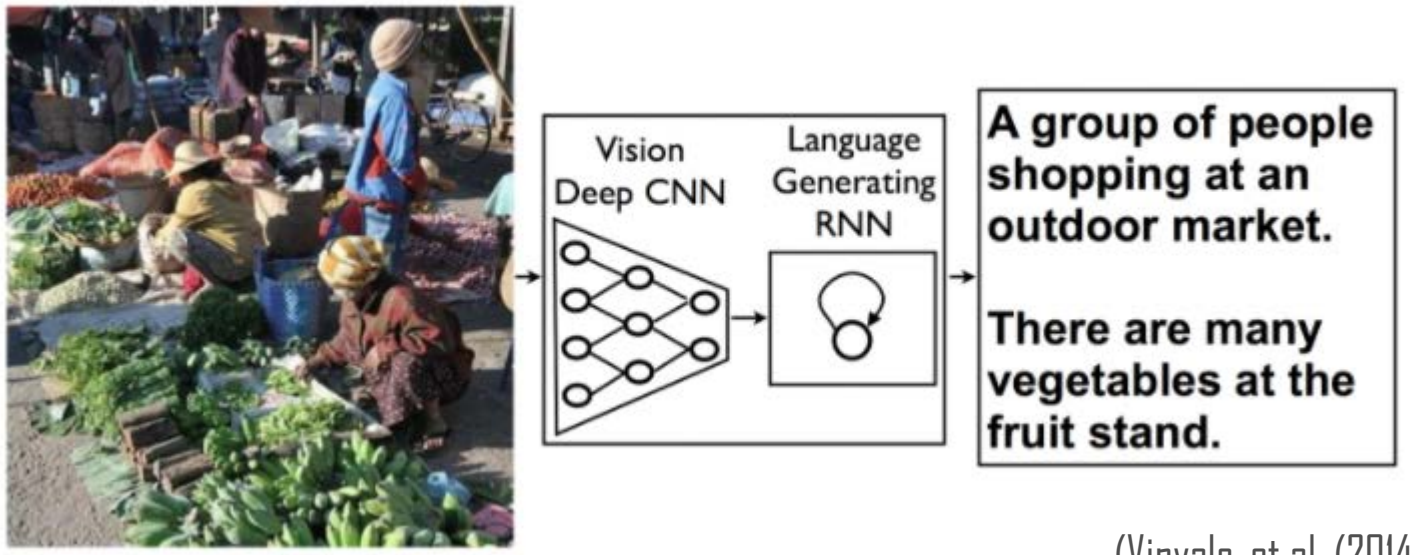


COMPOSABILITY

Microsoft

BUILDING PREDICTIVE ARCHITECTURES LIKE LEGO BLOCKS

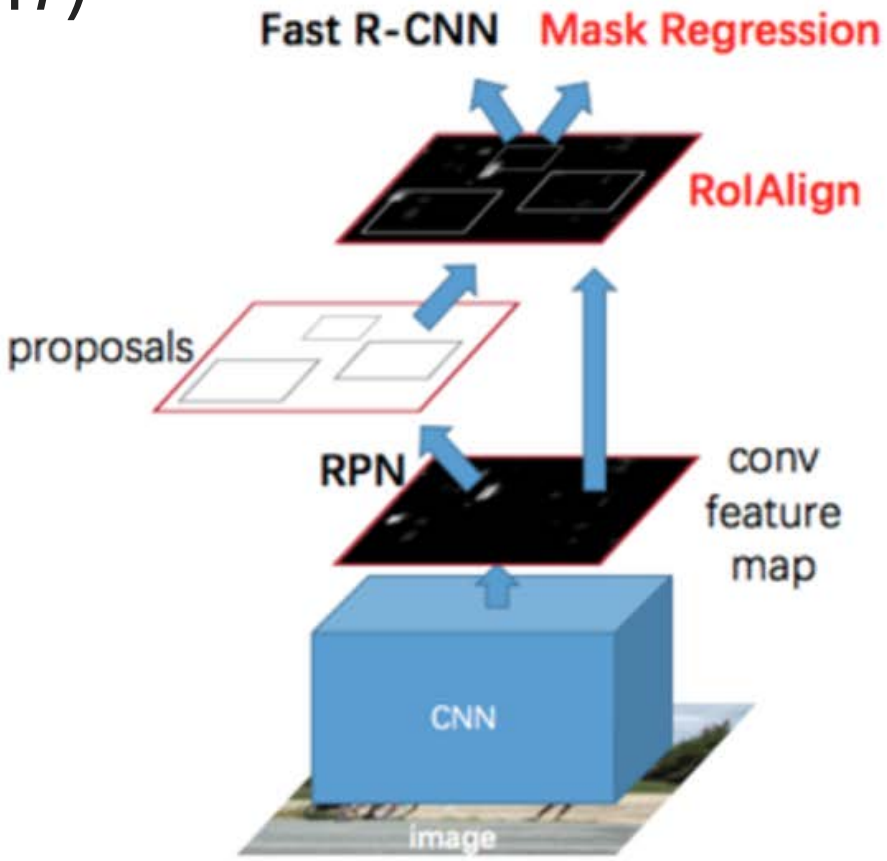




(Vinyals, et al. (2014))

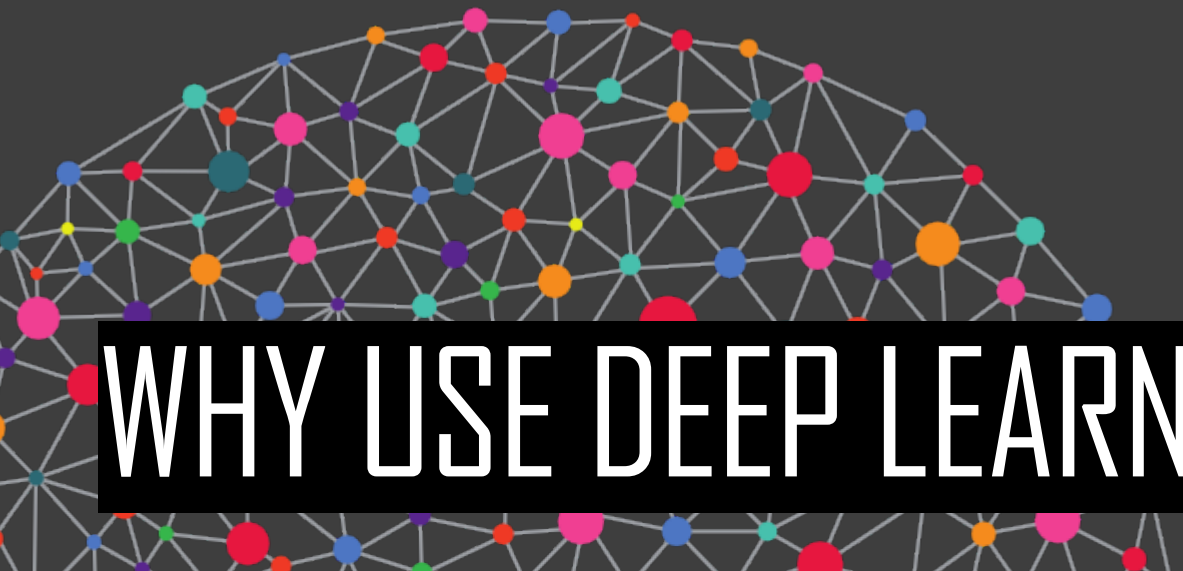
Microsoft

MASK R-CNN ARCHITECTURE (2017)



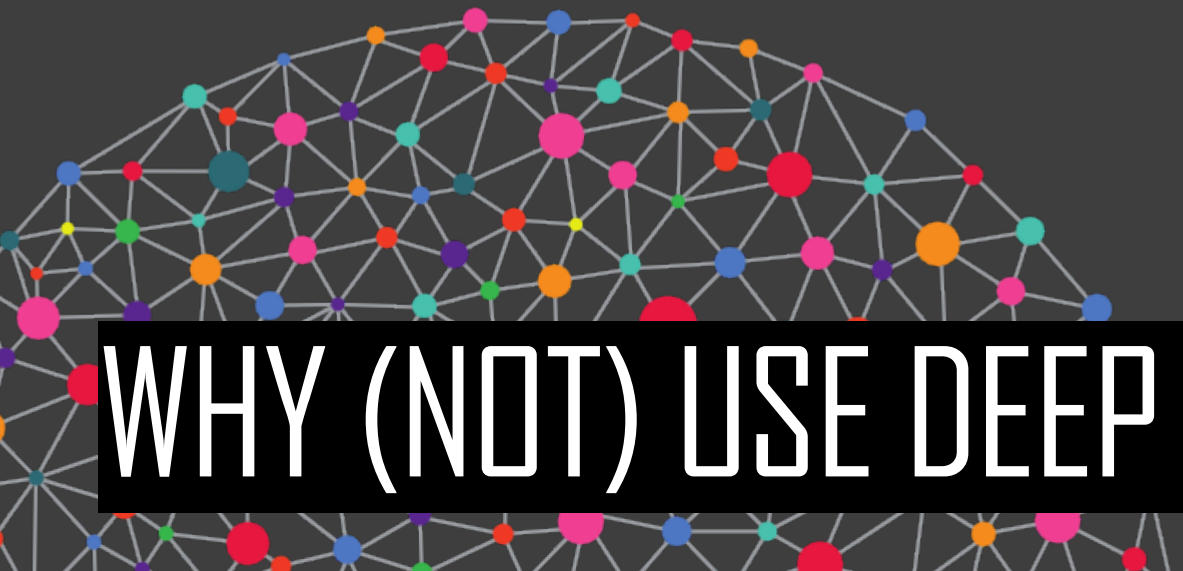
ARCHITECTURE COMPOSABILITY

1. I WANT TO MODEL TIME OR SPACE DEPENDENCIES
2. I HAVE A LOT OF DATA AND WANT A LARGE, NUANCED MODEL
3. "SOFTWARE 2.0" / NOVEL PREDICTION ARCHITECTURES
4. GENERATIVE MODELS



WHY USE DEEP LEARNING?

1. I NEED MODEL INTEPRETABILITY
2. I NEED PARSIMONY / LOW COMPUTE
3. I AM WORRIED ABOUT BIAS
4. I NEED TO UNDERSTAND MY DATA

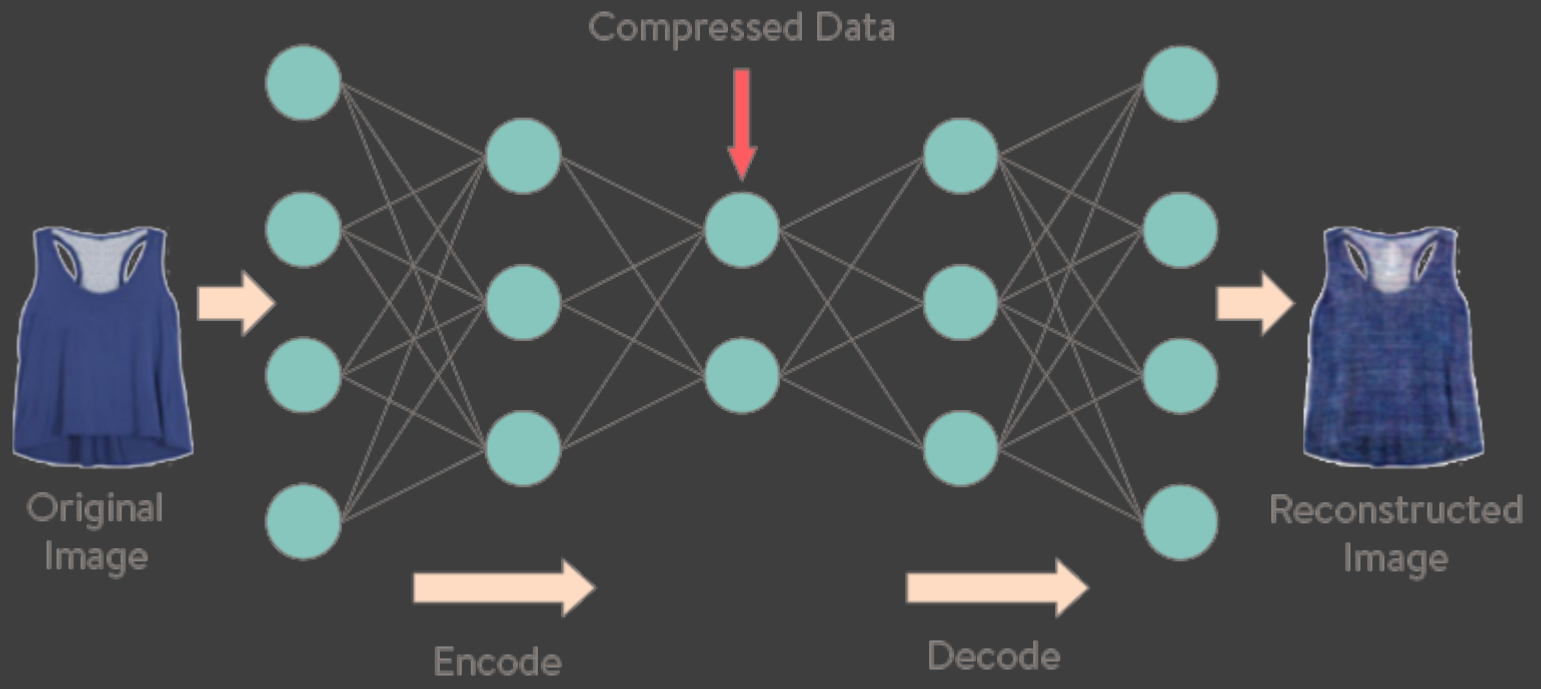


WHY (NOT) USE DEEP LEARNING?

Microsoft

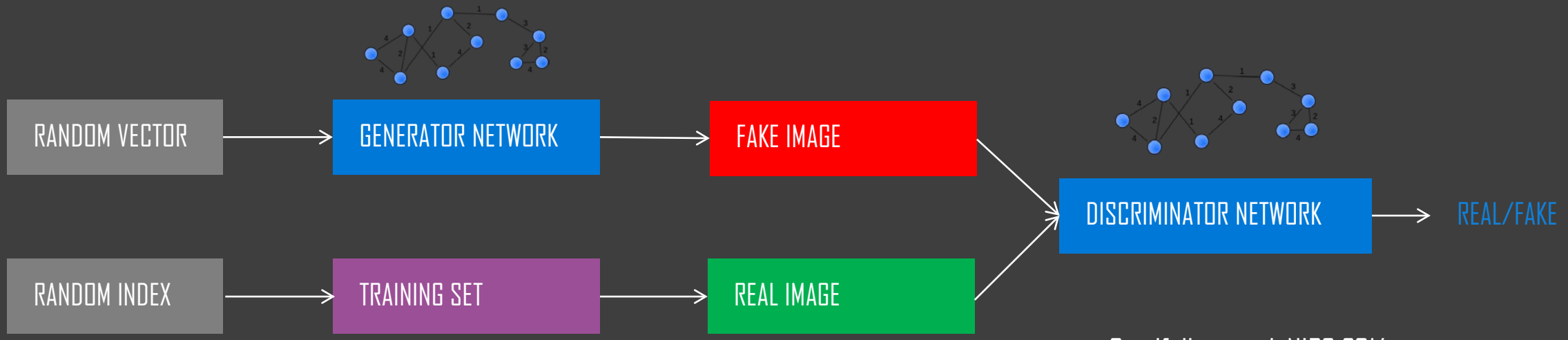
TWO INNOVATIVE DL ARCHITECTURES

- Autoencoders are an unsupervised method to compress information into an "embedding" similar to unsupervised dimensionality reduction.



AUTOENCODERS

- An example; generative adversarial networks
- Game-theoretical approach to deep learning



Goodfellow et al. NIPS 2014

VERY ADVANCED NOVEL ARCHITECTURES (GANS)

IMAGES GENERATED WITH GANS

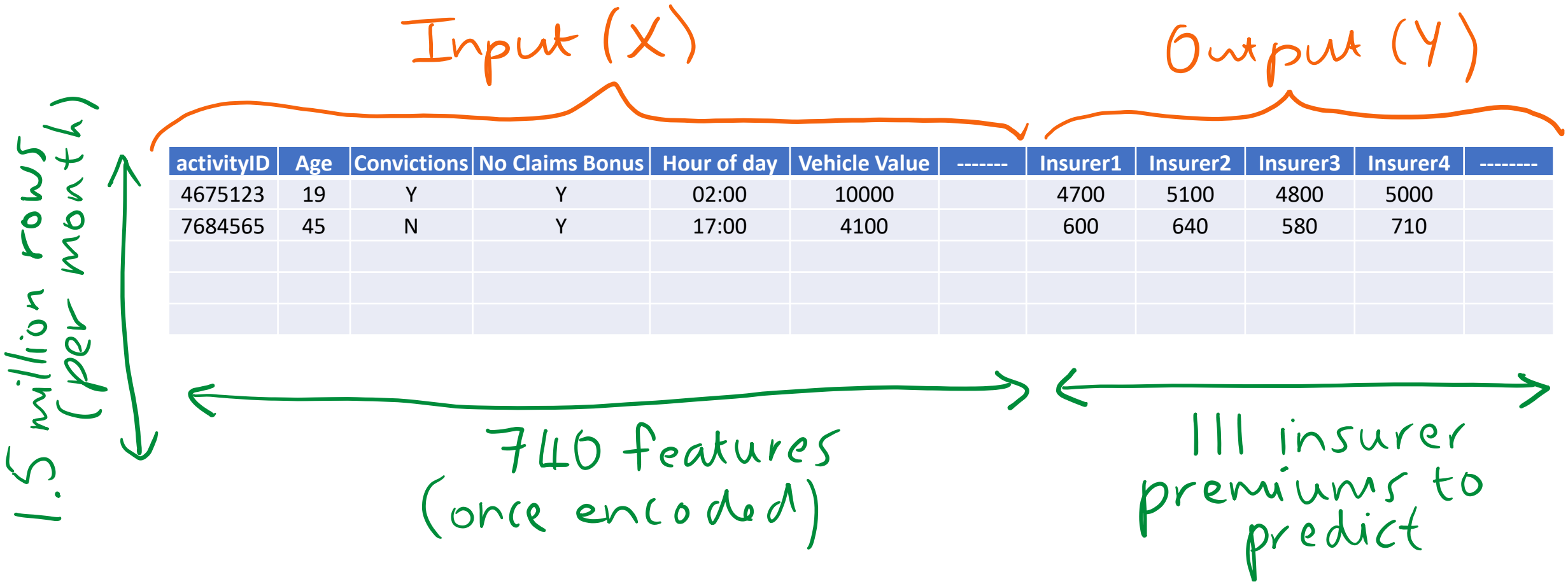
DEMO

NEURAL NETWORK PLAYGROUND

Microsoft

Predicting insurance premiums in a production data-science pipeline

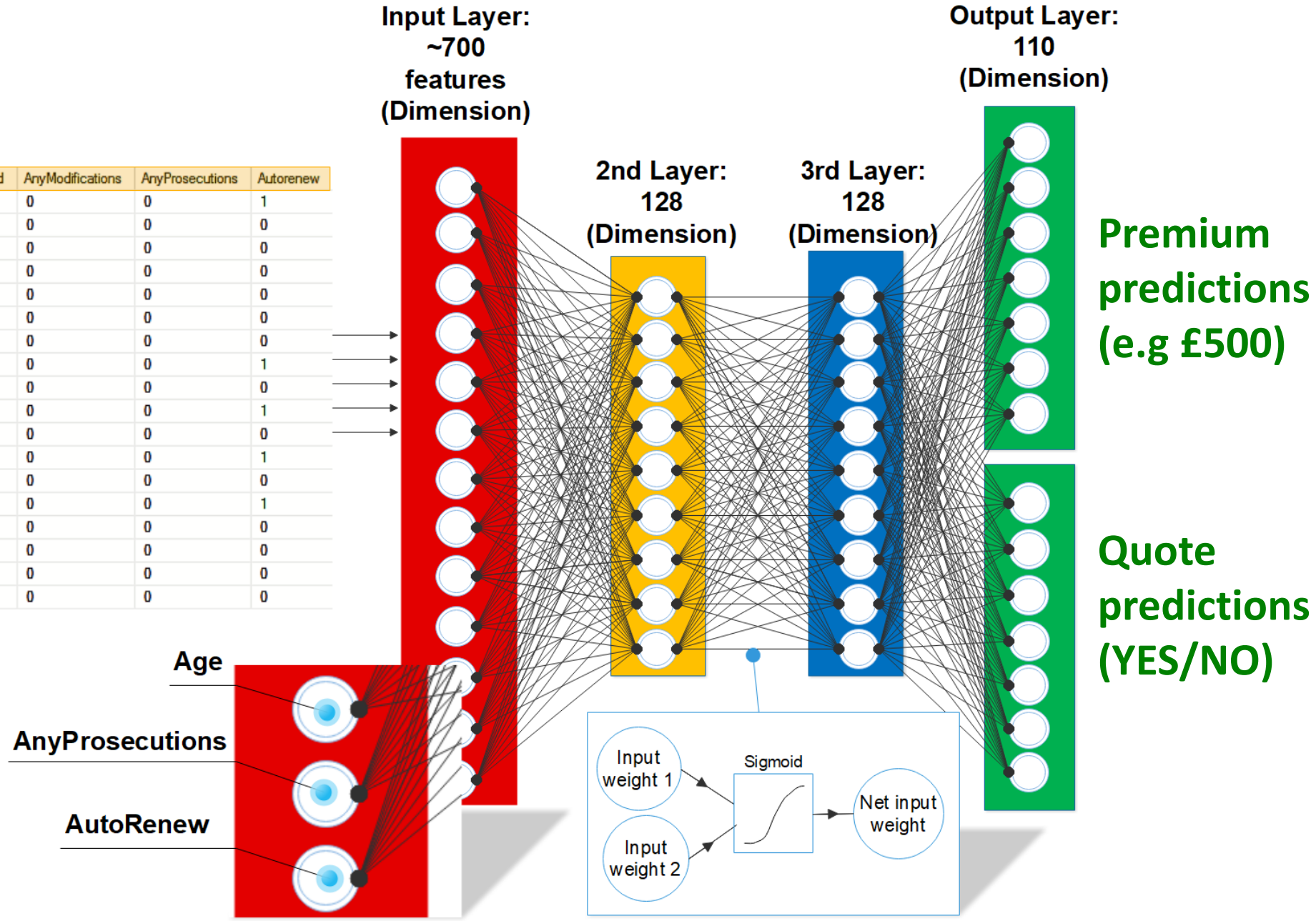
Data



activityID	Age	Convictions	No Claims Bonus	Hour of day	Vehicle Value	-------	Insurer1	Insurer2	Insurer3	Insurer4	--------
4675123	19	Y	Y	02:00	10000		4700	5100	4800	5000	
7684565	45	N	Y	17:00	4100		600	640	580	710	

Model

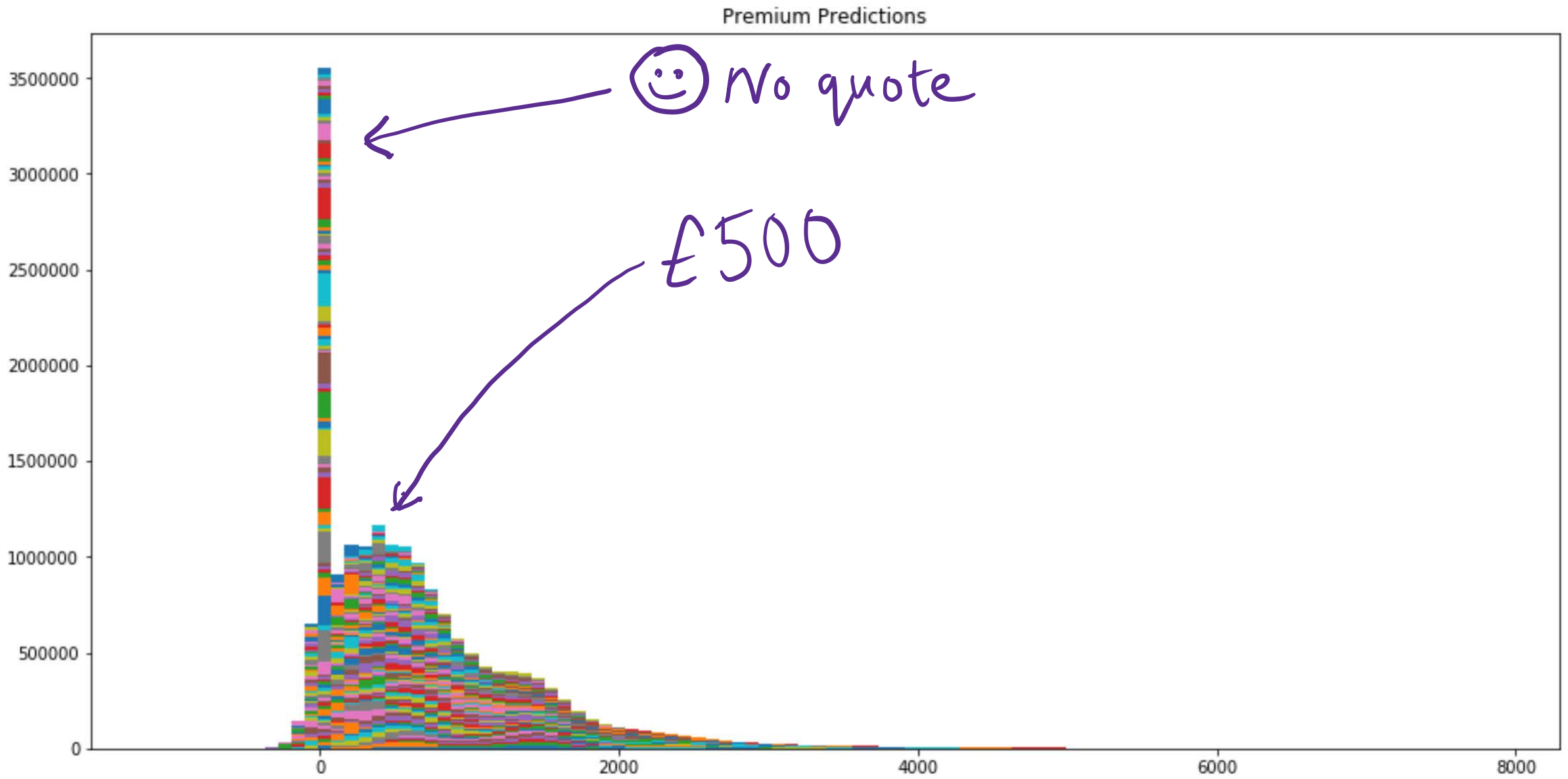
	RandomId	PartitionId	Amend	AnyModifications	AnyProsecutions	Autorenew
1	9EE062C...	0	0	0	0	1
2	8BD62F1...	3	0	0	0	0
3	1FB702C...	2	0	0	0	0
4	E78B9FF...	2	0	0	0	0
5	5CCCFD4...	4	0	0	0	0
6	127C460...	3	0	0	0	0
7	0BBAB1F...	5	0	0	0	0
8	E5C4B94...	0	0	0	0	1
9	D4AE24...	8	0	0	0	0
10	0DE604A...	9	0	0	0	1
11	D1F30F6...	0	0	0	0	0
12	3177A7C...	6	0	0	0	1
13	633B8A4...	1	0	0	0	0
14	C933BF9...	10	0	0	0	1
15	107EE8C...	5	0	0	0	0
16	407A008...	8	0	0	0	0
17	348EC6E...	3	0	0	0	0
18	0EA607F...	4	0	0	0	0

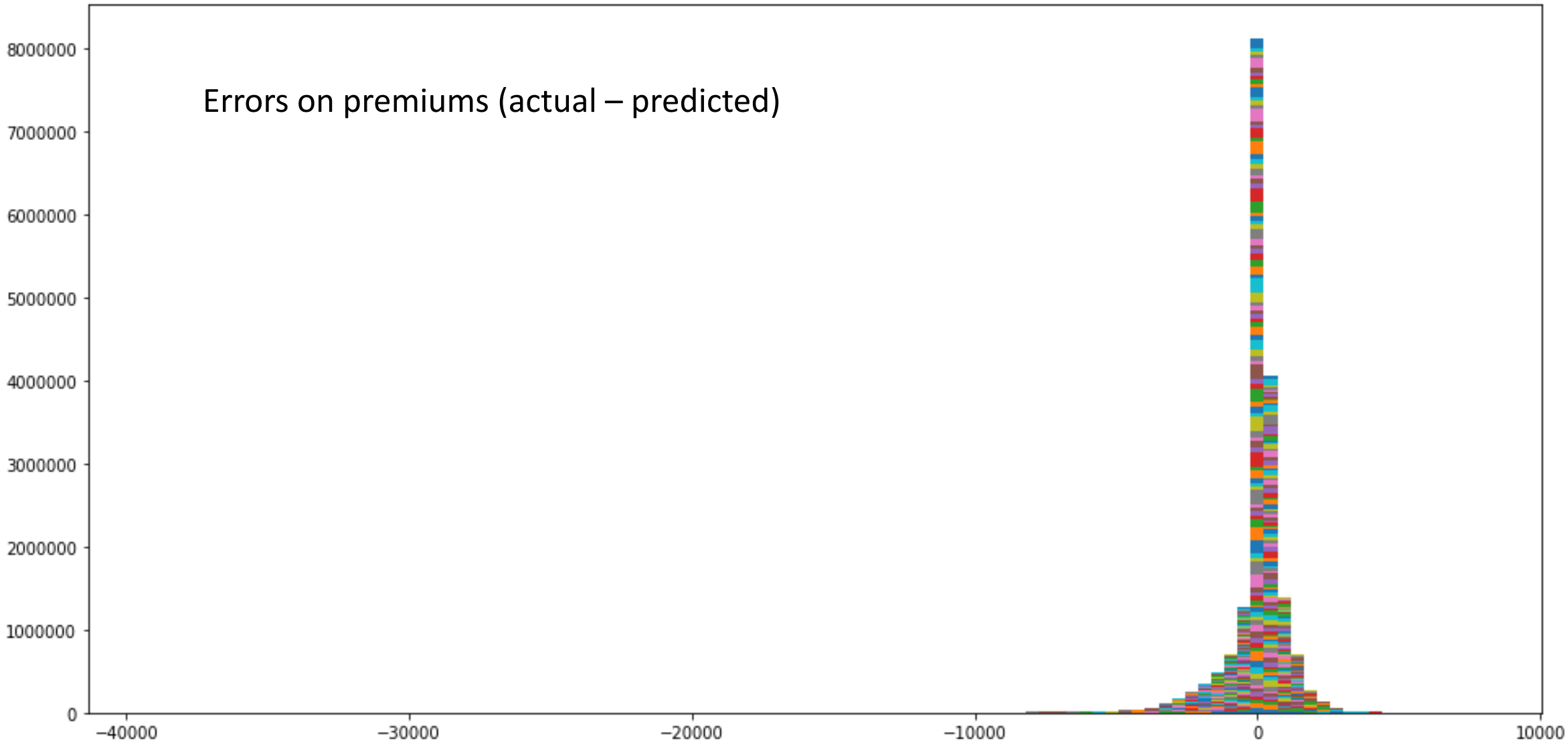


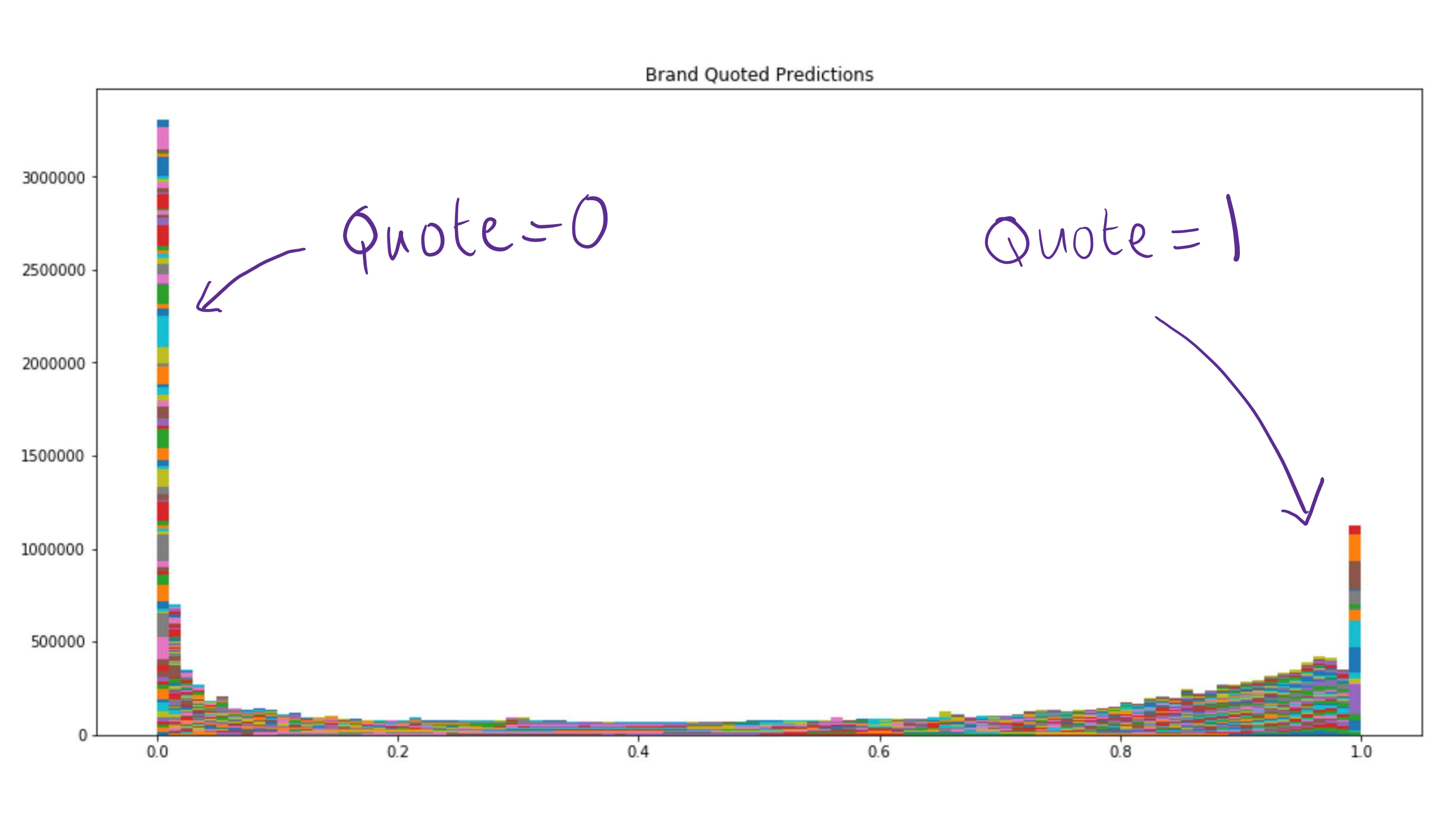
Premium predictions (e.g £500)

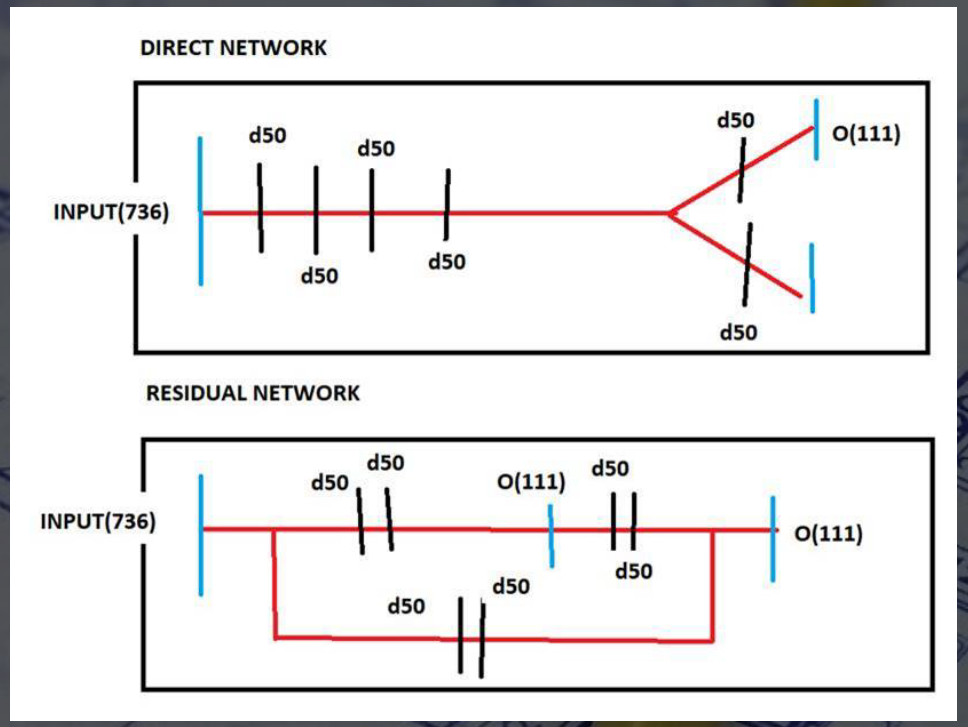
Quote predictions (YES/NO)

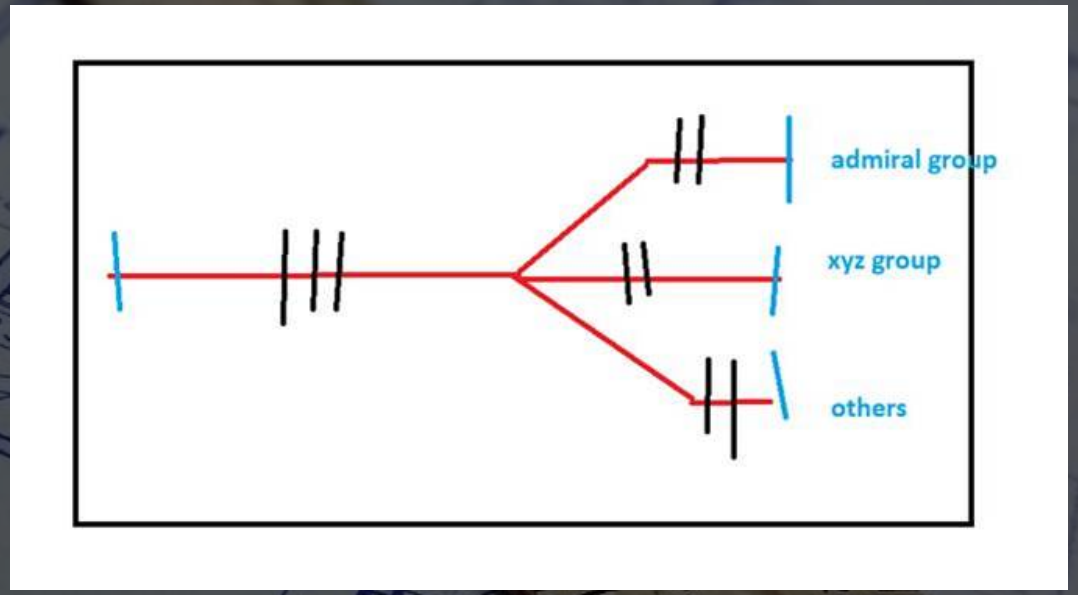
Model results on test data



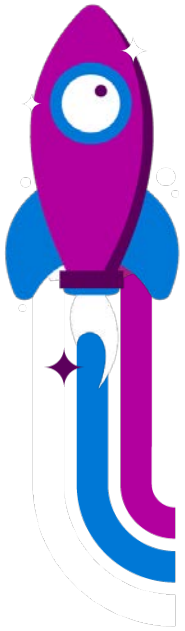

Errors on premiums (actual – predicted)
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
−40000
−30000
−20000
−10000
0
10000


Brand Quoted Predictions
Quote=0
Quote=1
3000000
2500000
2000000
1500000
1000000
500000
0
0.0
0.2
0.4
0.6
0.8
1.0





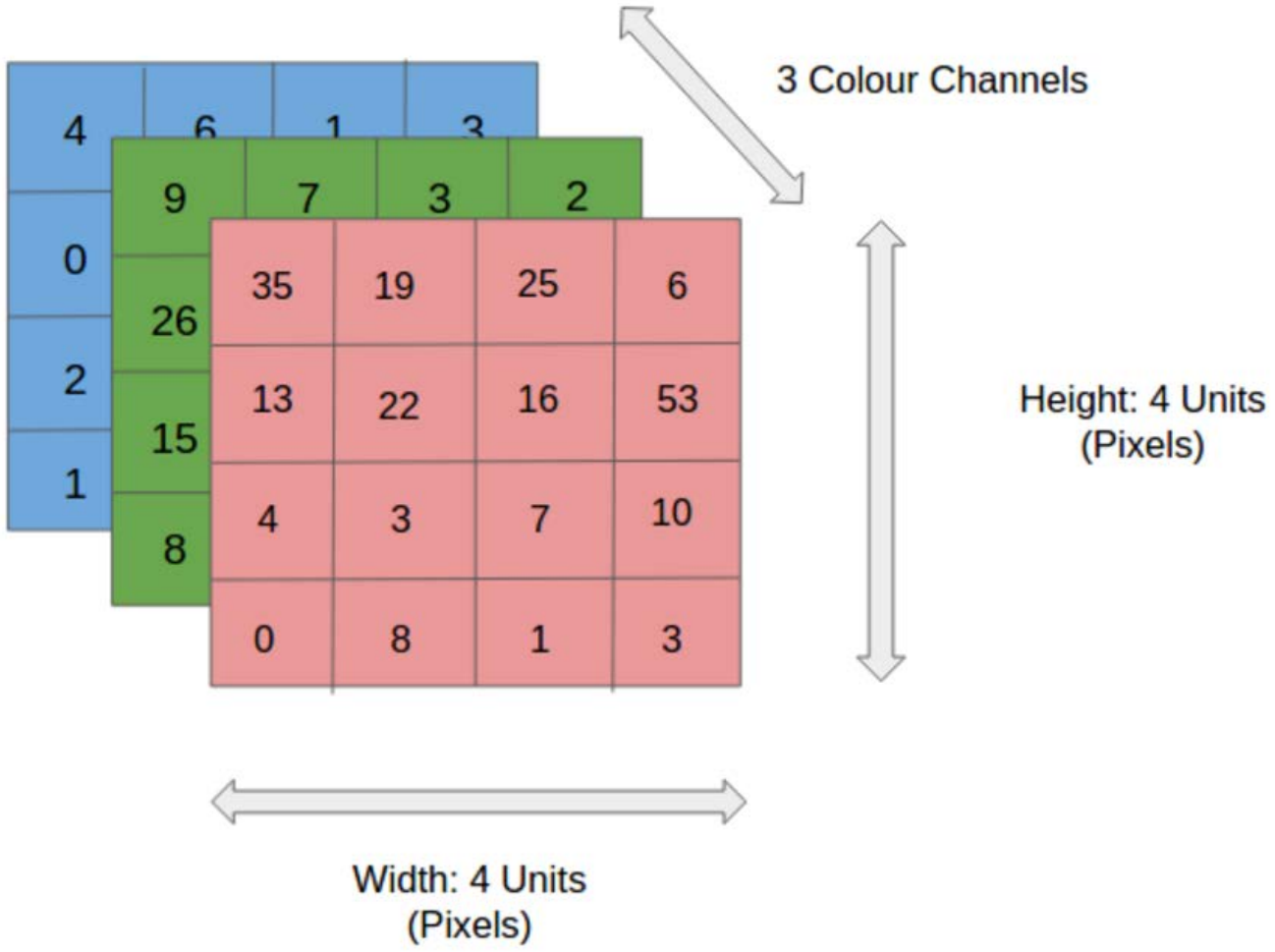
Architectures Explored



How do convolutional neural networks work?

Microsoft

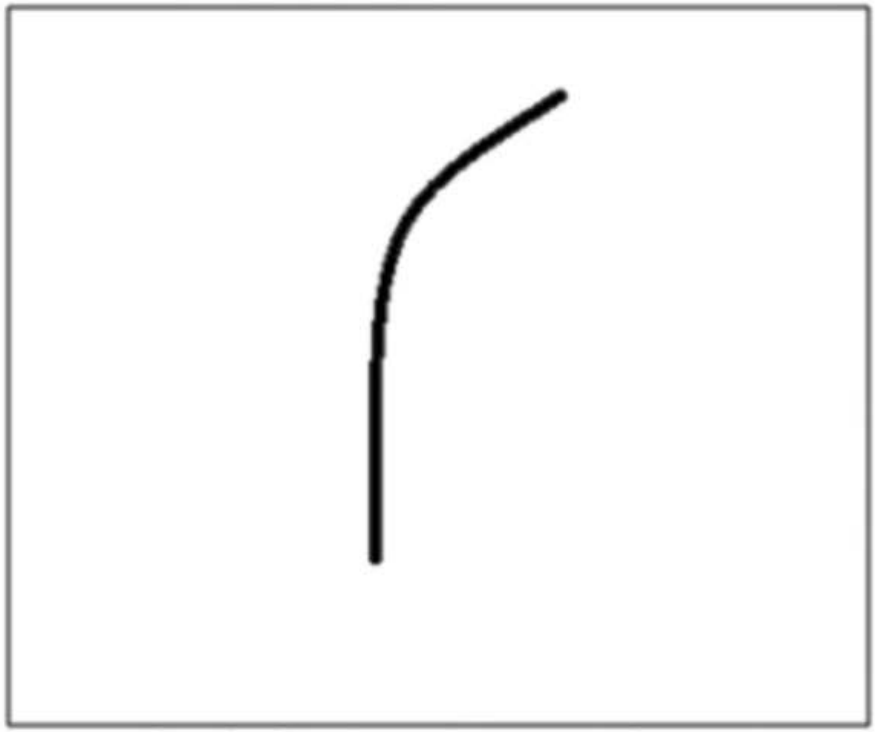
PREPARE DATASET OF IMAGES



Microsoft

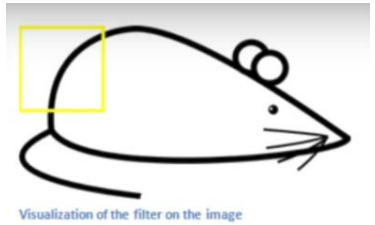
0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter



Visualization of a curve detector filter

CONVOLUTION FILTER



Visualization of the filter on the image

$$(50*30)+(50*30)+(50*30)+(20*30)+(50*30) = 6600$$

Visualization of the receptive field

0	0	0	0	0	0	30
0	0	0	0	50	50	50
0	0	0	20	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0
0	0	0	50	50	0	0

Pixel representation of the receptive field

*

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

Pixel representation of filter

CONVOLUTION FILTER MATCH

Visualization of the filter on the image

MULTIPLY AND SUMMATION = 0

0	0	0	0	0	0	0
0	40	0	0	0	0	0
40	0	40	0	0	0	0
40	20	0	0	0	0	0
0	50	0	0	0	0	0
0	0	50	0	0	0	0
25	25	0	50	0	0	0

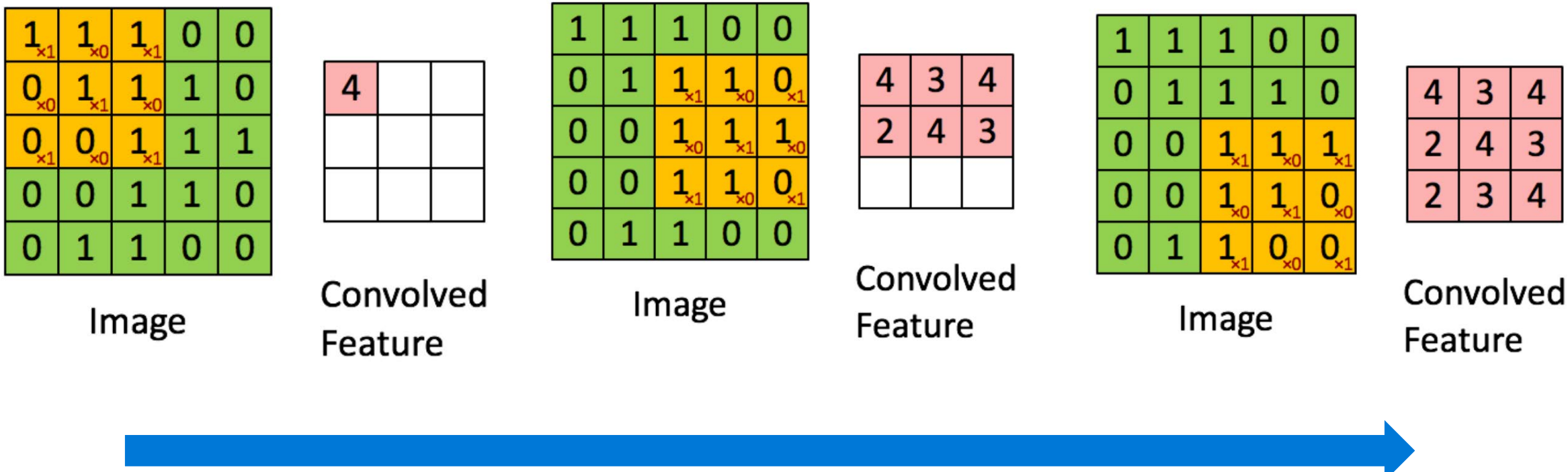
Pixel representation of receptive field

*

0	0	0	0	0	30	0
0	0	0	0	30	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	30	0	0	0
0	0	0	0	0	0	0

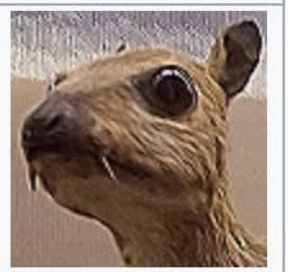
Pixel representation of filter

CONVOLUTION FILTER NO MATCH

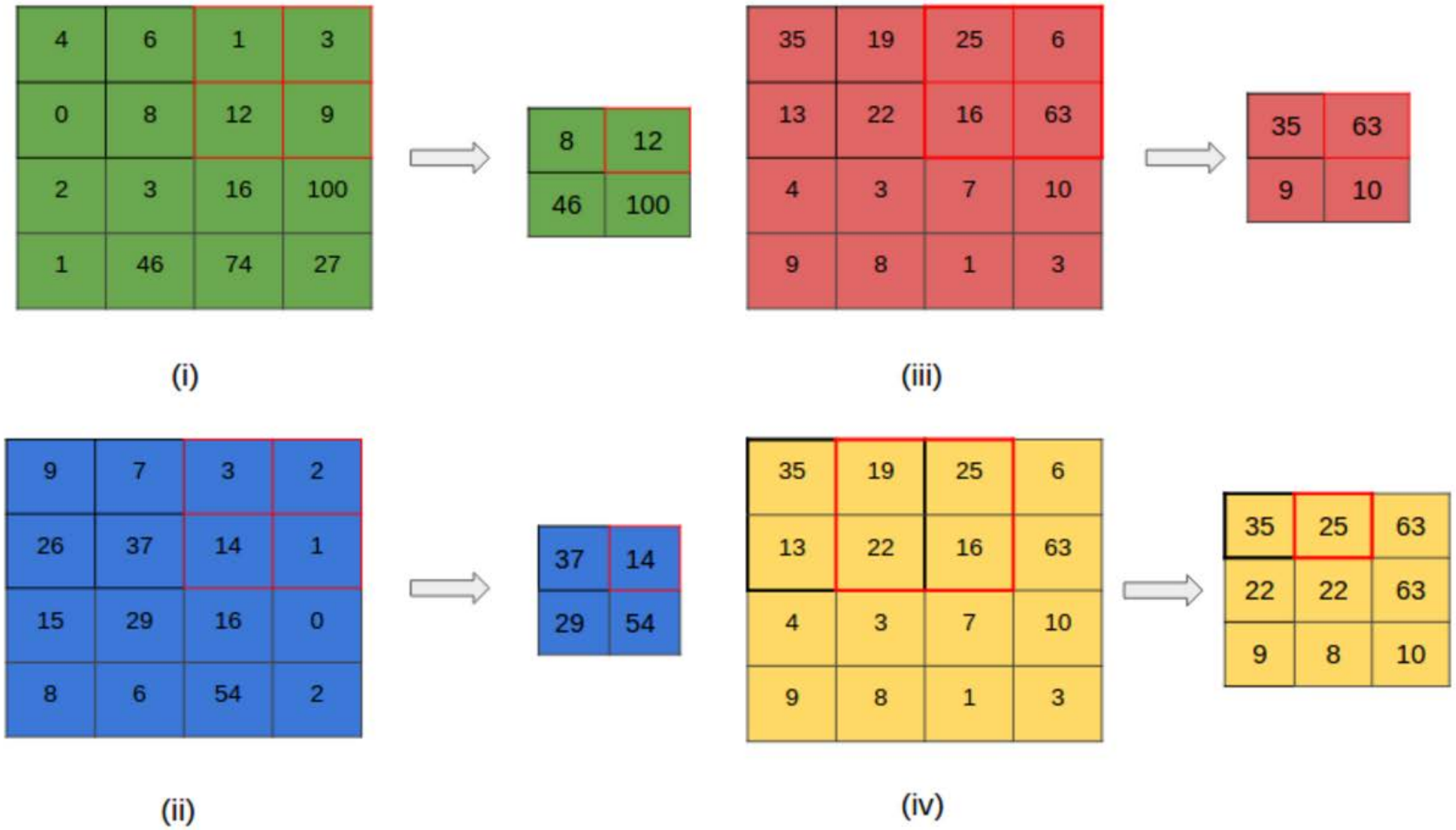


CONVOLUTION

DEMO

Operation	Kernel	Image result
Identity	$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$	
Edge detection	$\begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}$	
	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$	
	$\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$	
Sharpen	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	

Microsoft

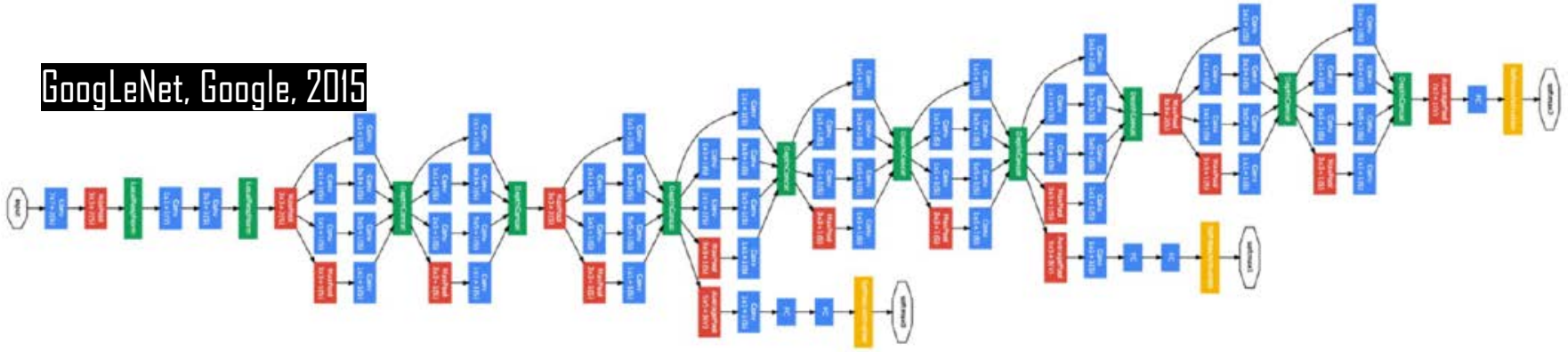

4 6 1 3
0 8 12 9
2 3 16 100
1 46 74 27
8 12
46 100
(i)
35 19 25 6
13 22 16 63
4 3 7 10
9 8 1 3
35 63
9 10
(iii)
9 7 3 2
26 37 14 1
15 29 16 0
8 6 54 2
37 14
29 54
(ii)
35 19 25 6
13 22 16 63
4 3 7 10
9 8 1 3
35 25 63
22 22 63
9 8 10
(iv)

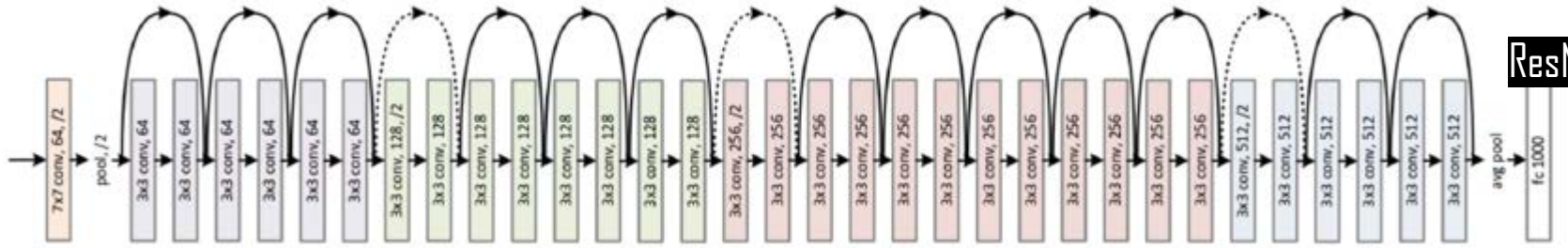
POOLING

Microsoft

VGGNet, Oxford Uni, 2014

GoogLeNet, Google, 2015

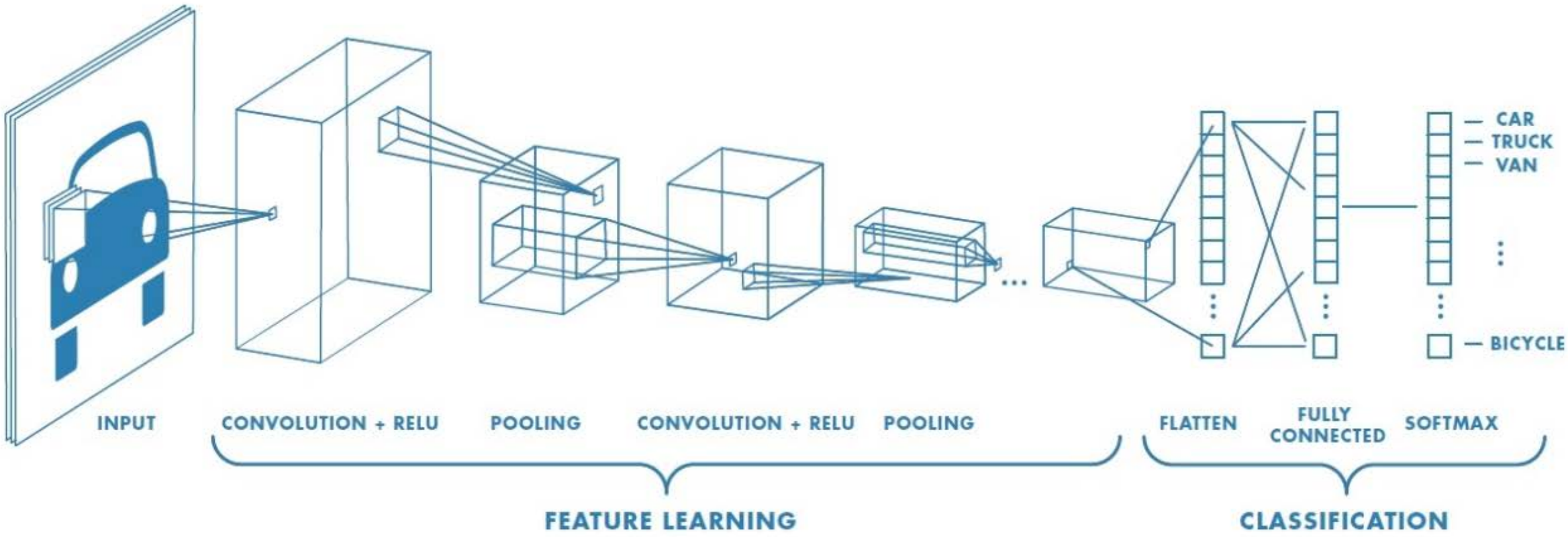




ResNet, Microsoft, 2016

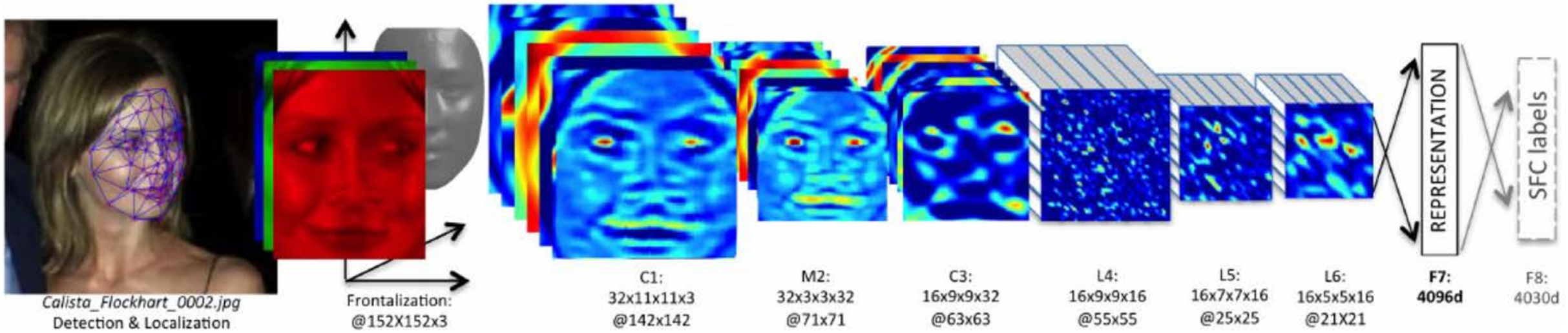
WINNING ARCHITECTURES

IMAGE CLASSIFICATION EXAMPLE



DEEP FACE

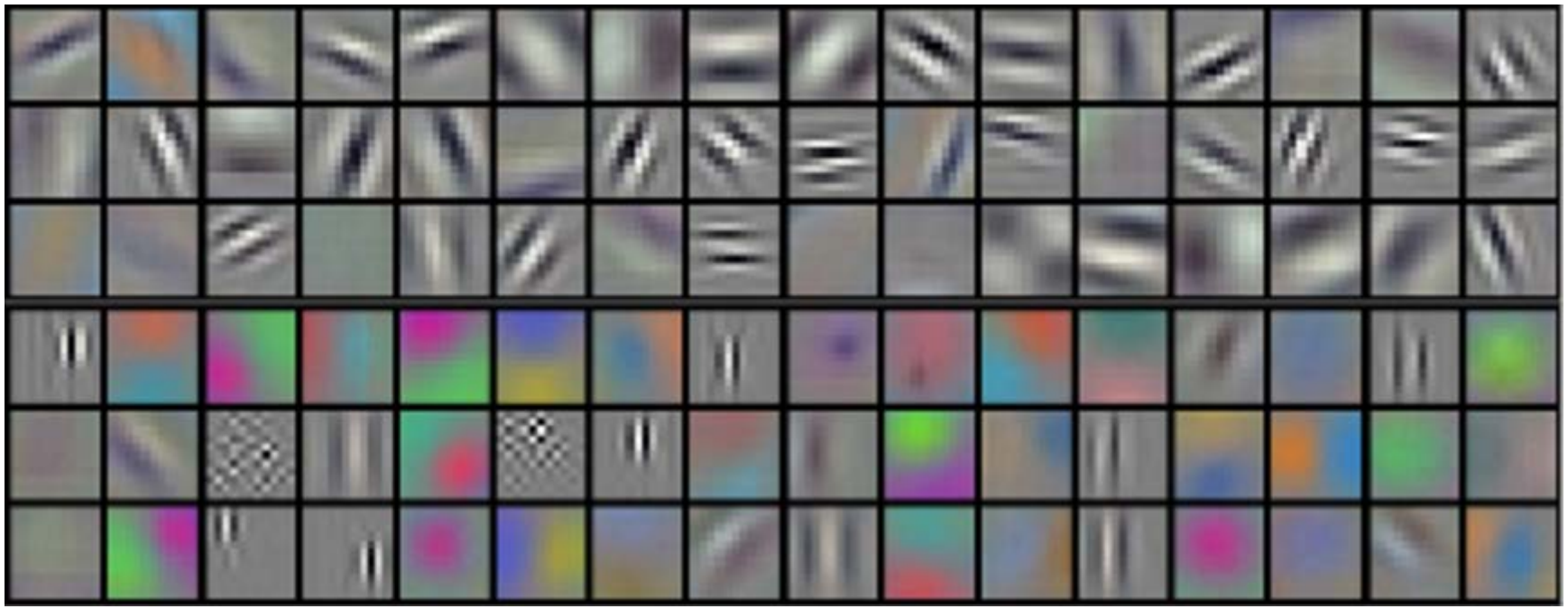
Taigman et al. CVPR 2014



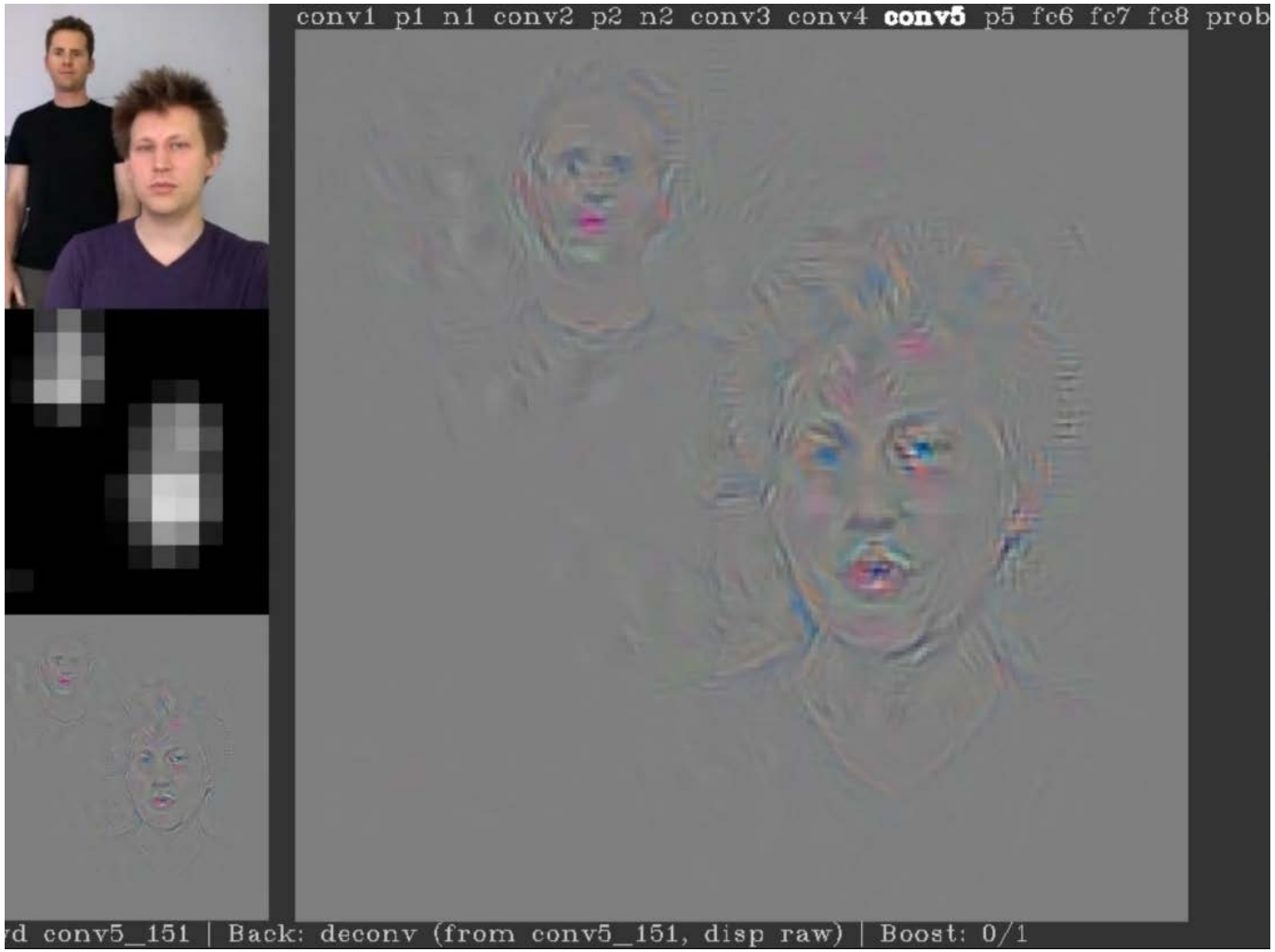
- Alignment
- CNN
- Metric Learning

Microsoft

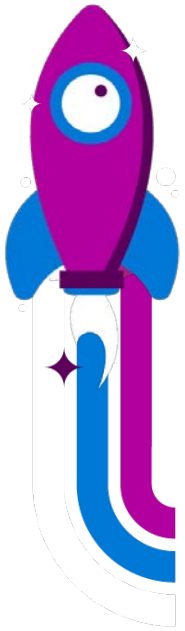
VISUALISING THE FILTERS



DEEP VISUALISATION TOOLBOX



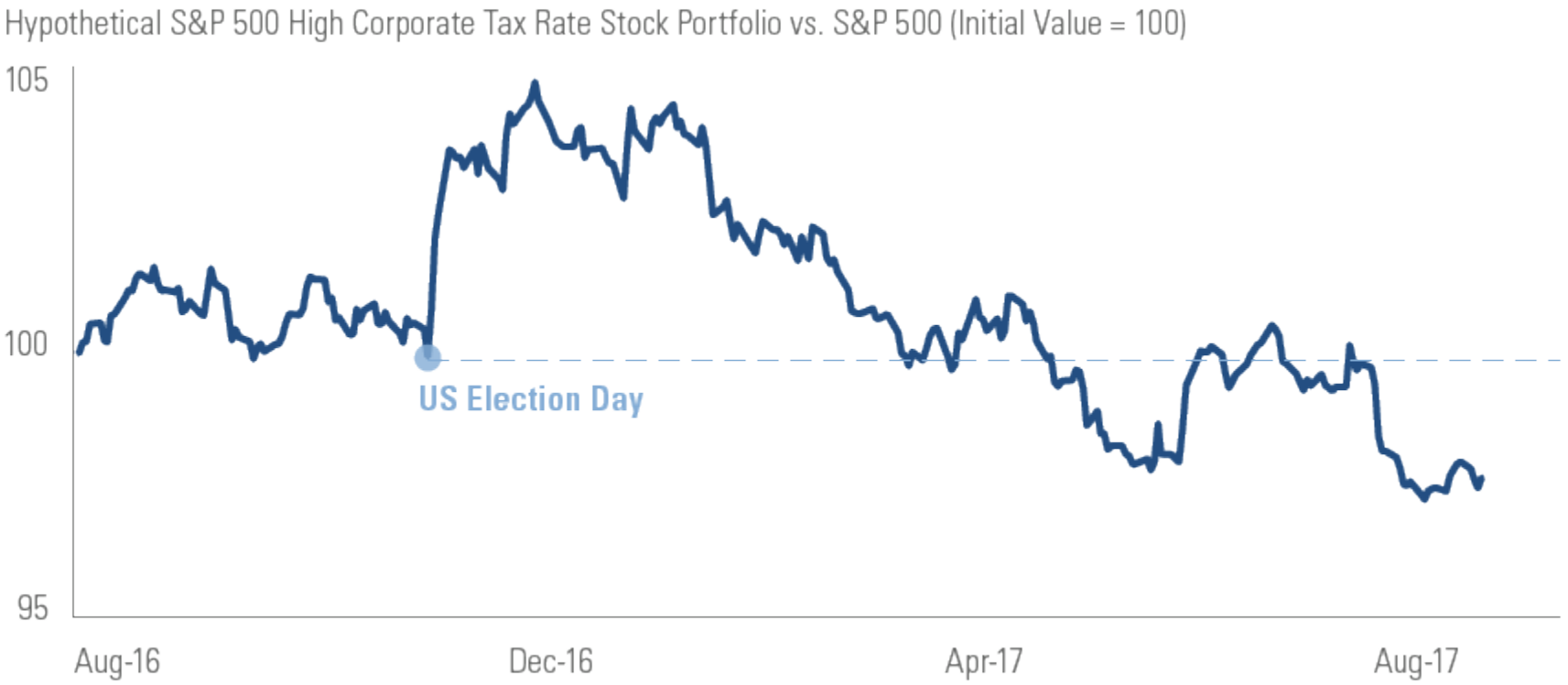
https://www.youtube.com/watch?v=AgkfIQ4IGaM



Demo of CNNs in Data Bricks/Keras

Microsoft

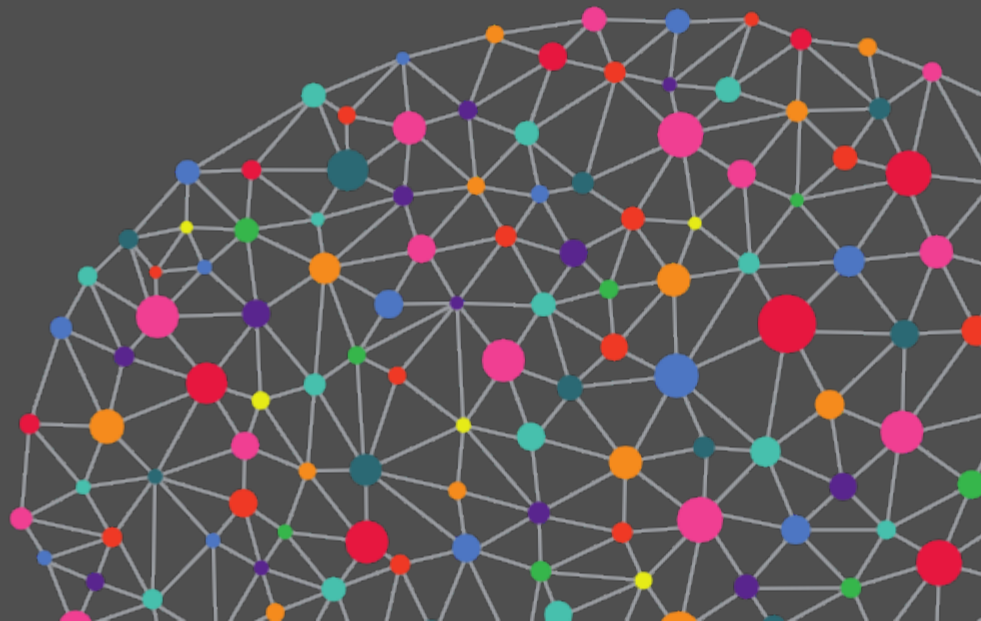
WHAT IS A SEQUENCE?



Microsoft

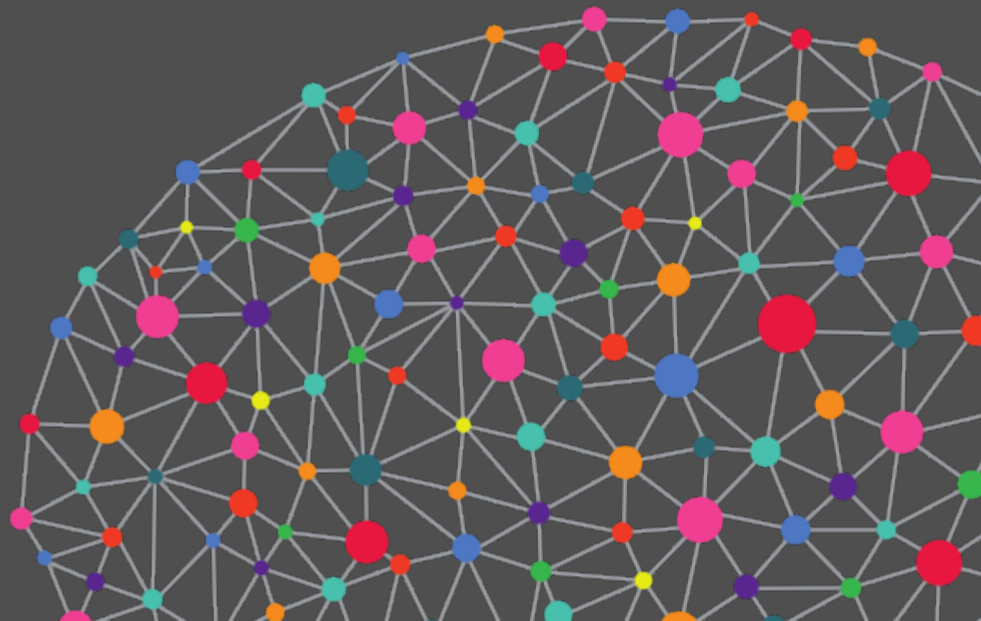
WHAT IS SEQUENCE PROCESSING

- RNNs
  - Timeseries classification
  - Anomaly detection in timeseries
  - Entity recognition
  - Revenue forecasting
  - Question + Answer
- 1d Convnets
  - Spelling correction
  - Document classification
  - Machine translation



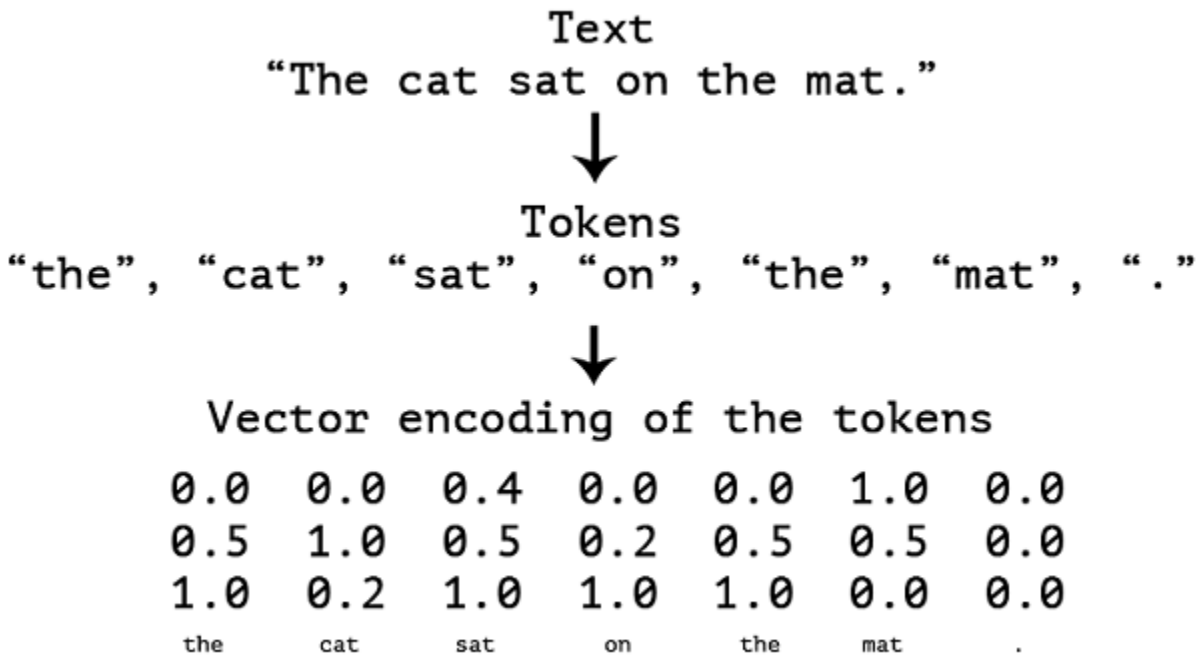
WHAT IS SEQUENCE PROCESSING

- RNNs
  - When global order matters
- 1d Convnets
  - Speed
  - Local temporal dependencies
- You can stack them!



TOKENIZATION

- Words
- Characters
- N-grams

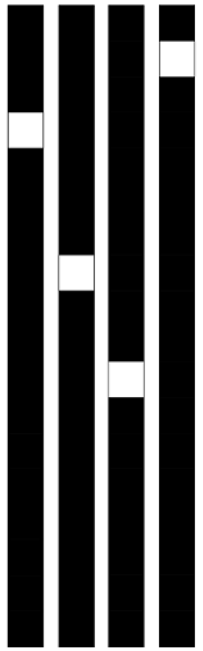


N-Grams example

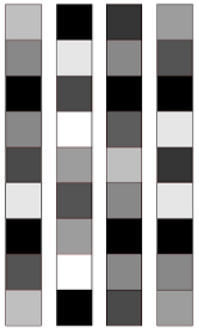
```
{"The", "The cat", "cat", "cat sat", "The cat sat",
  "sat", "sat on", "on", "cat sat on", "on the", "the",
  "sat on the", "the mat", "mat", "on the mat"}
```

Microsoft

WORD VECTORS VS WORD EMBEDDINGS?



One-hot word vectors:
- Sparse
- High-dimensional
- Hard-coded



Word embeddings:
- Dense
- Lower-dimensional
- Learned from data

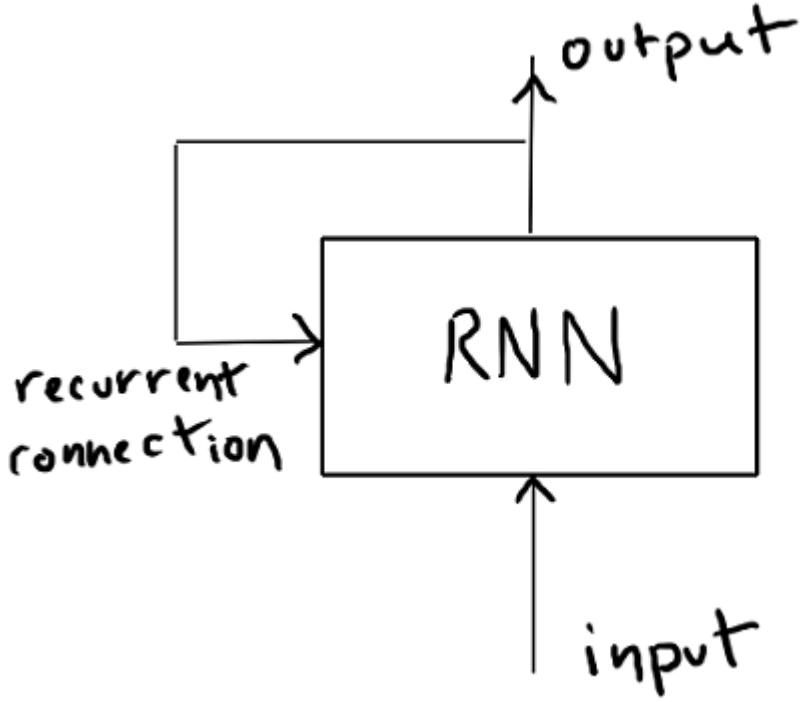
Microsoft

WORD EMBEDDINGS

- Word2Vec
- GloVe

Microsoft

RECURRENT NEURAL NETWORKS



RECURRENT NEURAL NETWORKS

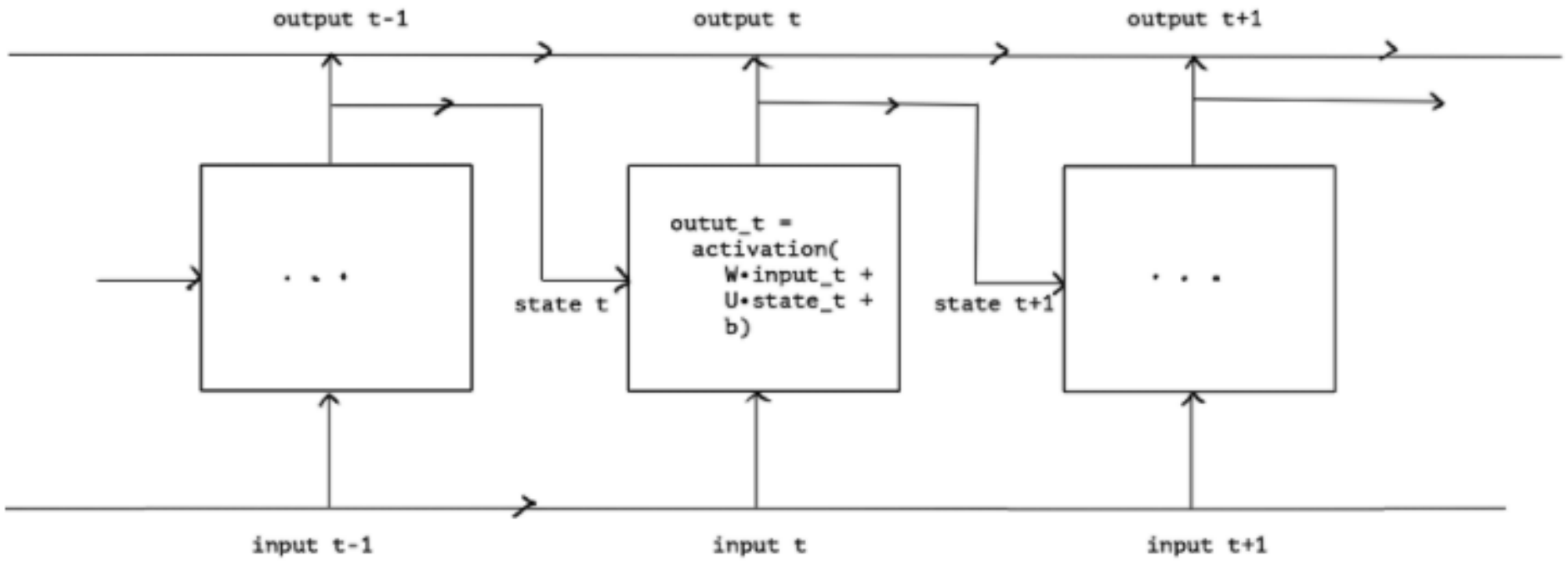


Image © Francois Chollet

LSTMS(1)

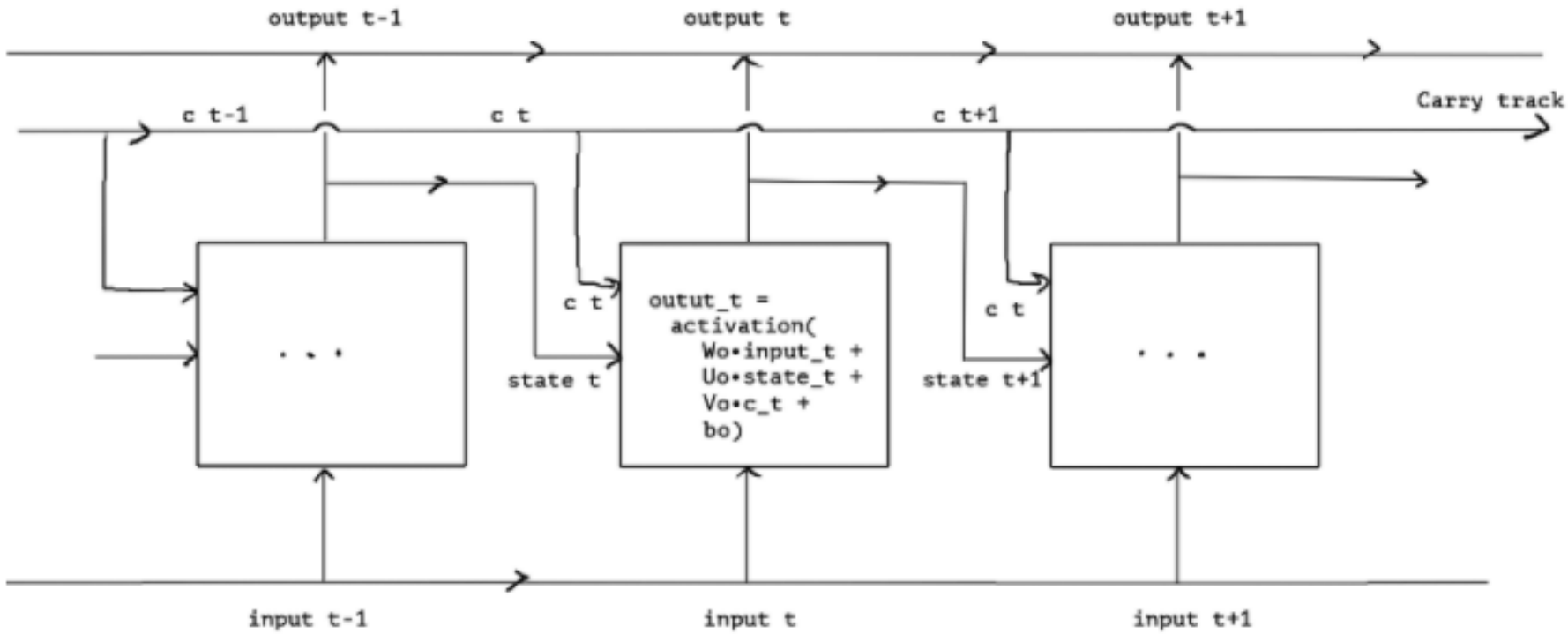
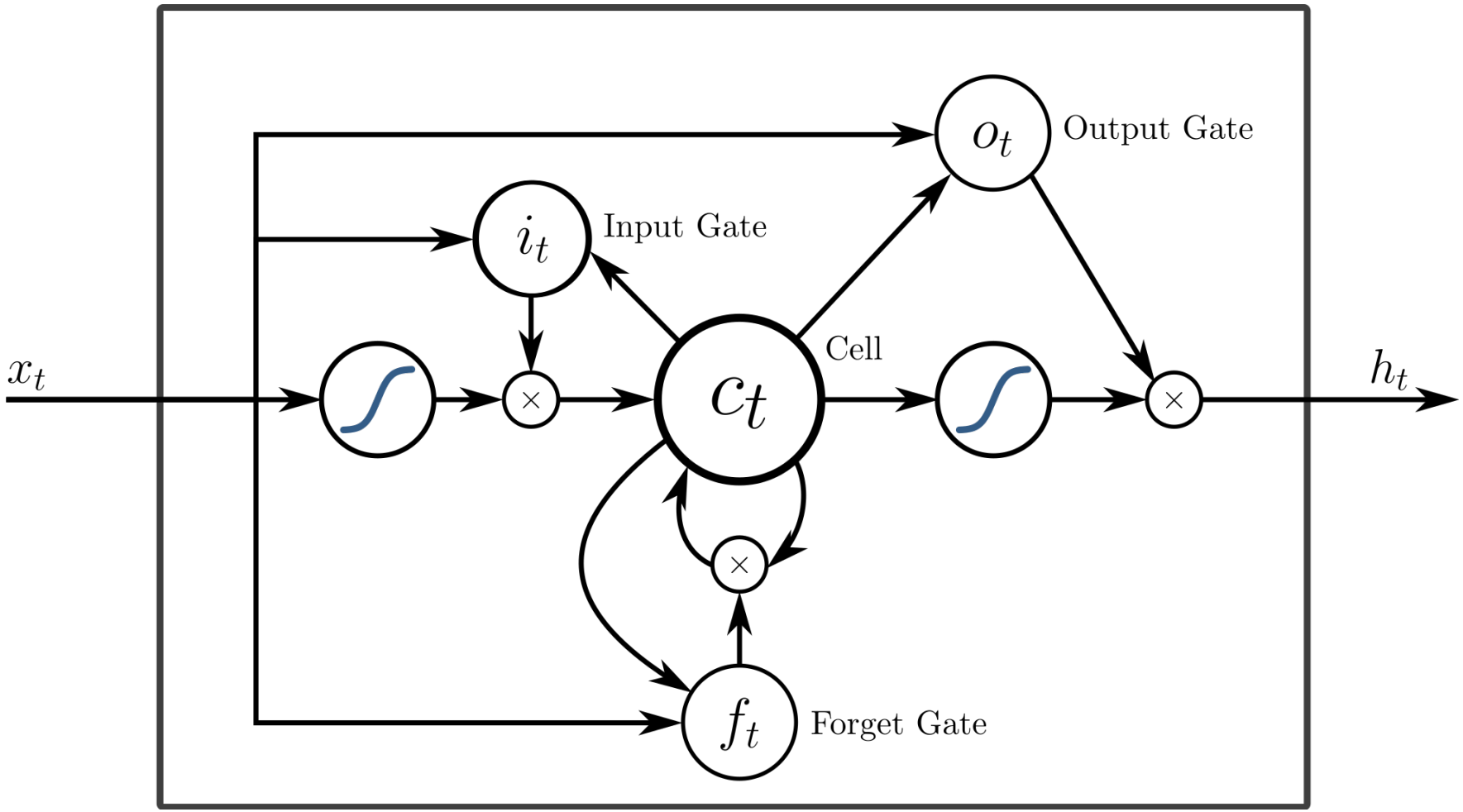


Image © Francois Chollet

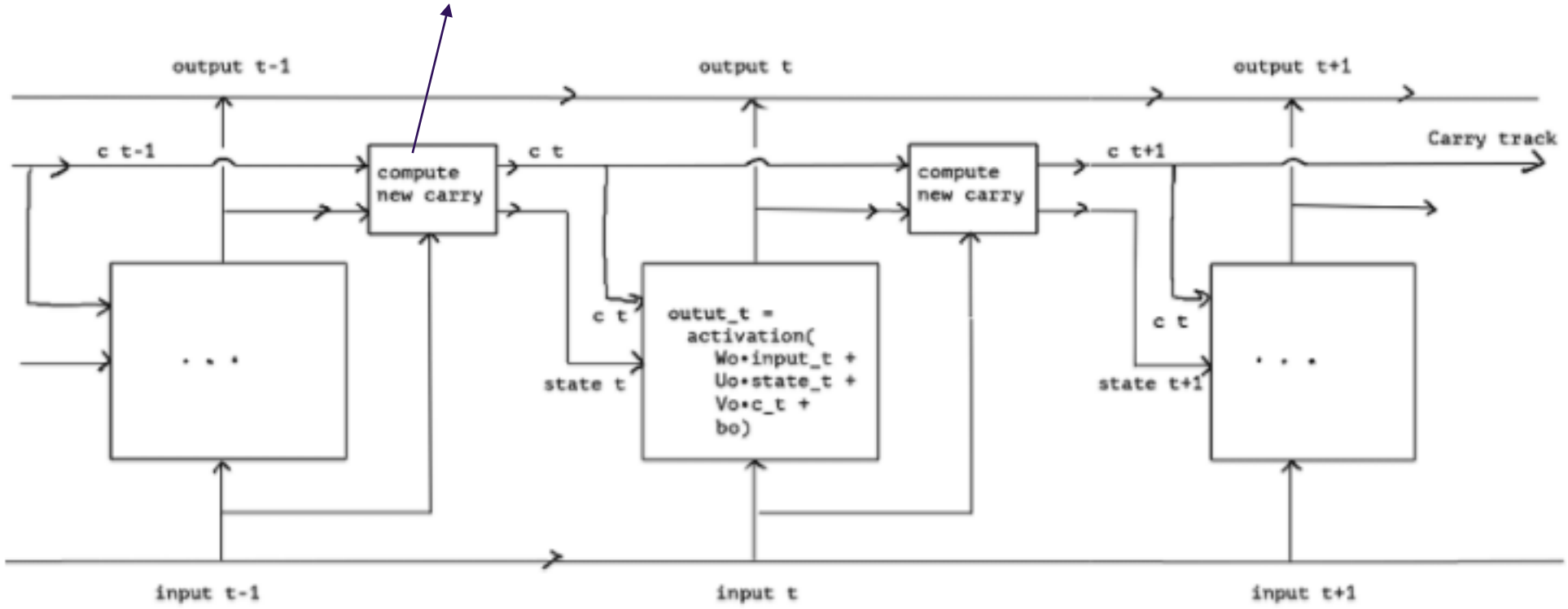
LSTMS(2)



Microsoft

LSTMS(3)

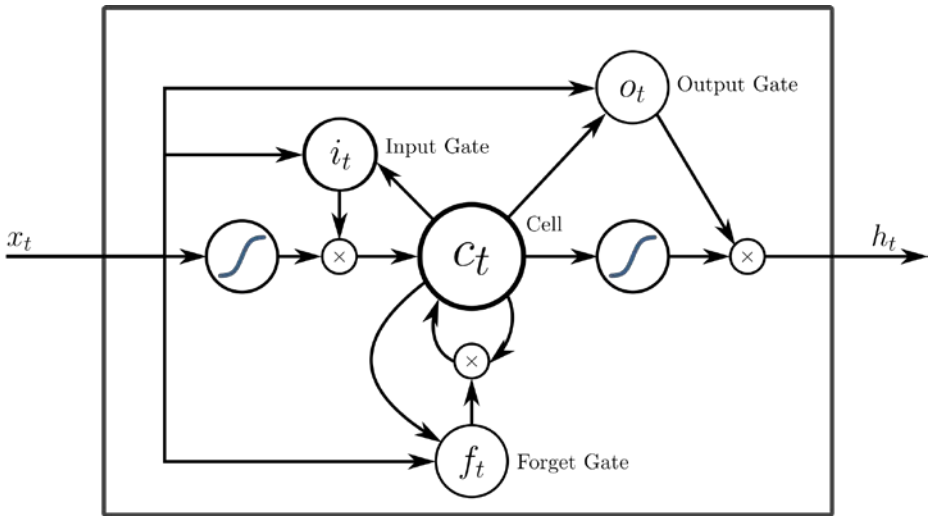
```
c_t+1 = i_t * k_t + c_t * f_t
```



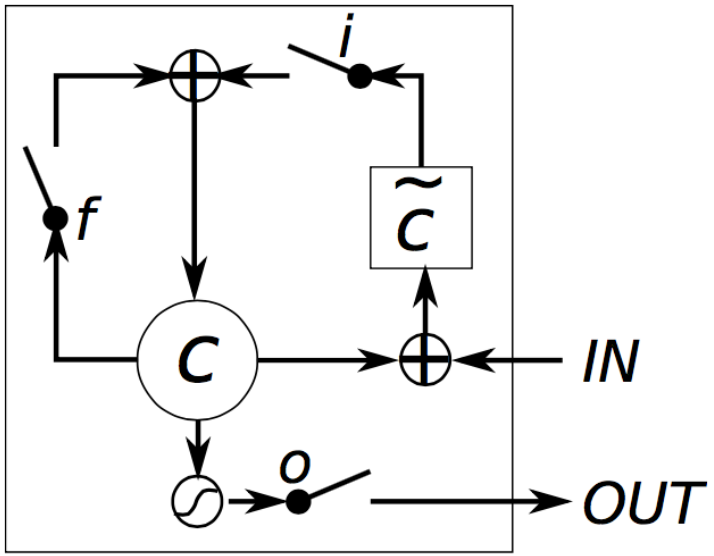

```
i_t = activation(dot(state_t, Ui) + dot(input_t, Wi) + bi)
f_t = activation(dot(state_t, Uf) + dot(input_t, Wf) + bf)
k_t = activation(dot(state_t, Uk) + dot(input_t, Wk) + bk)
```

Image © Francois Chollet

LSTM VS GRU



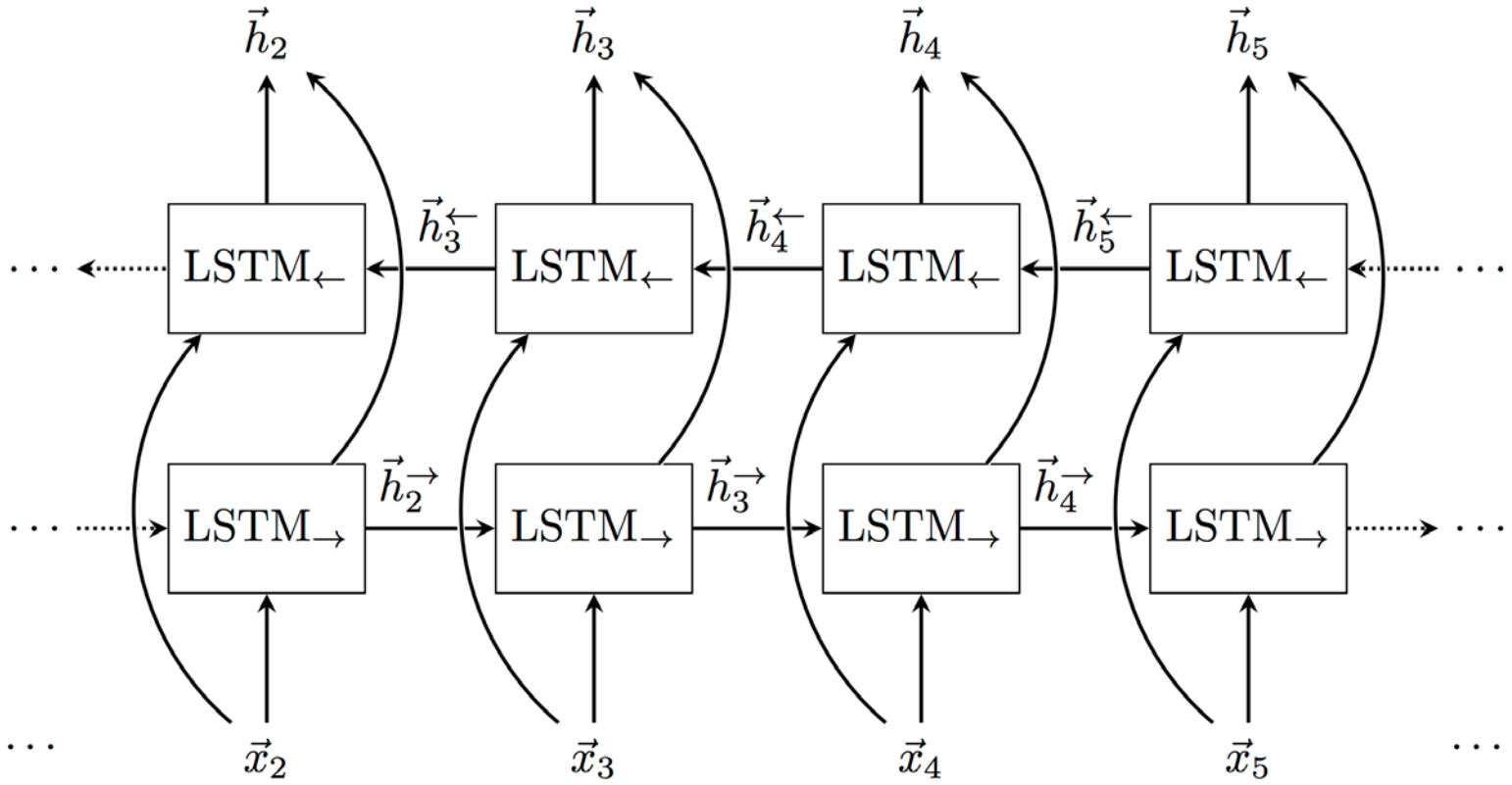
LSTM



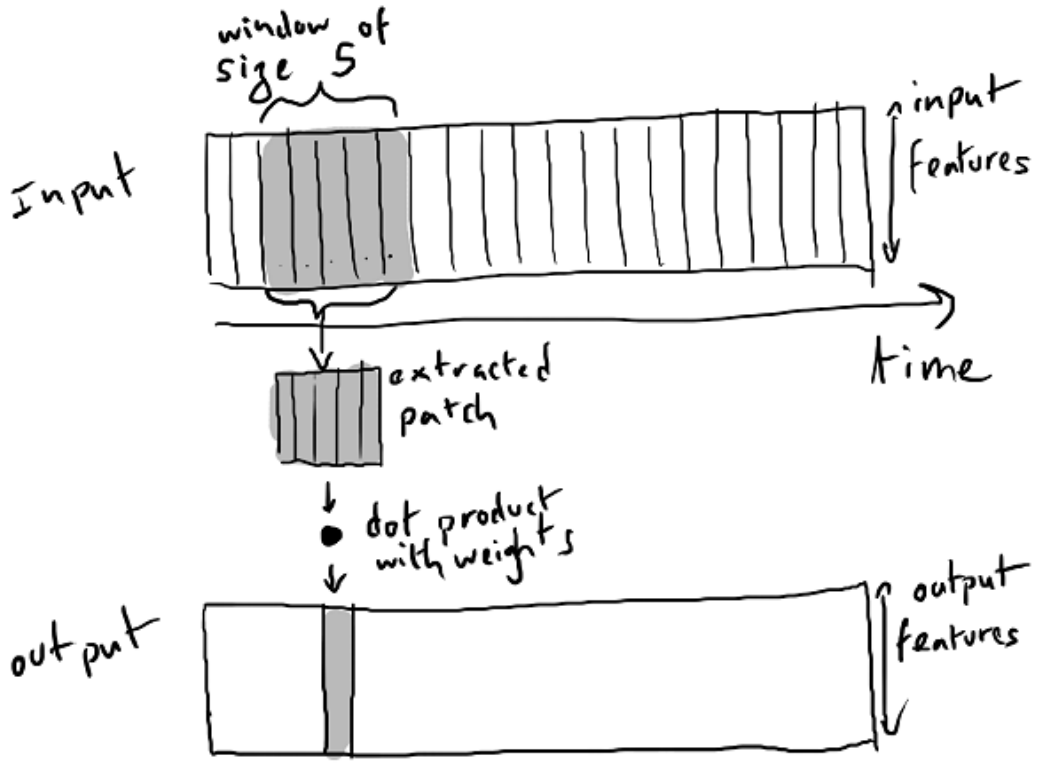
GRU

Microsoft

BI-DIRECTIONAL LSTMS

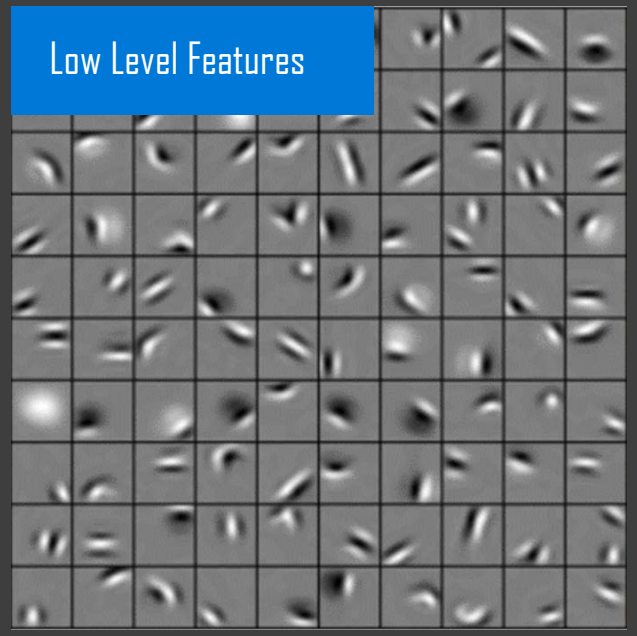


1D-CNNS



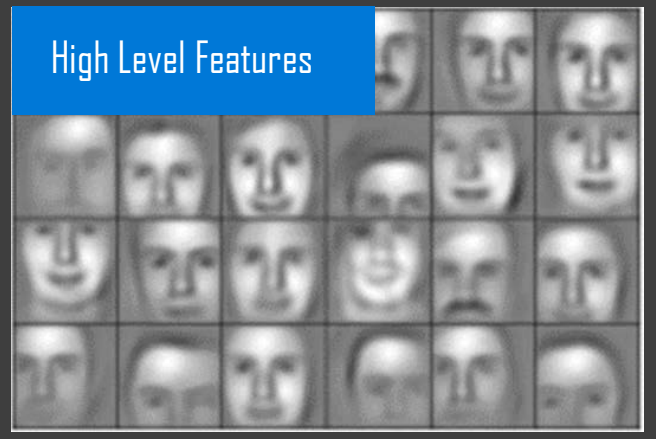
Microsoft

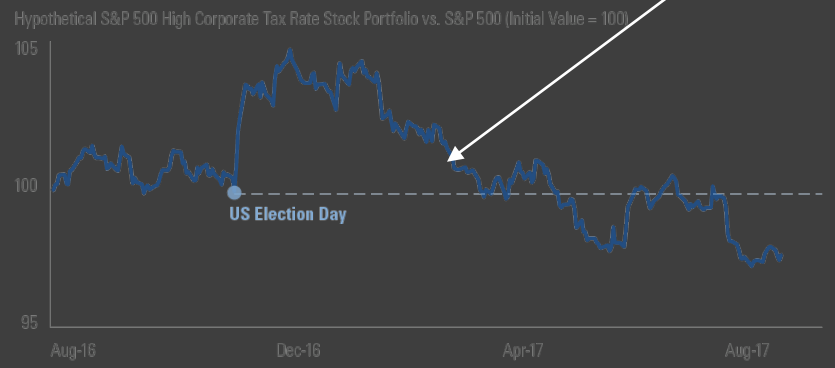
Low Level Features



Mid Level Features

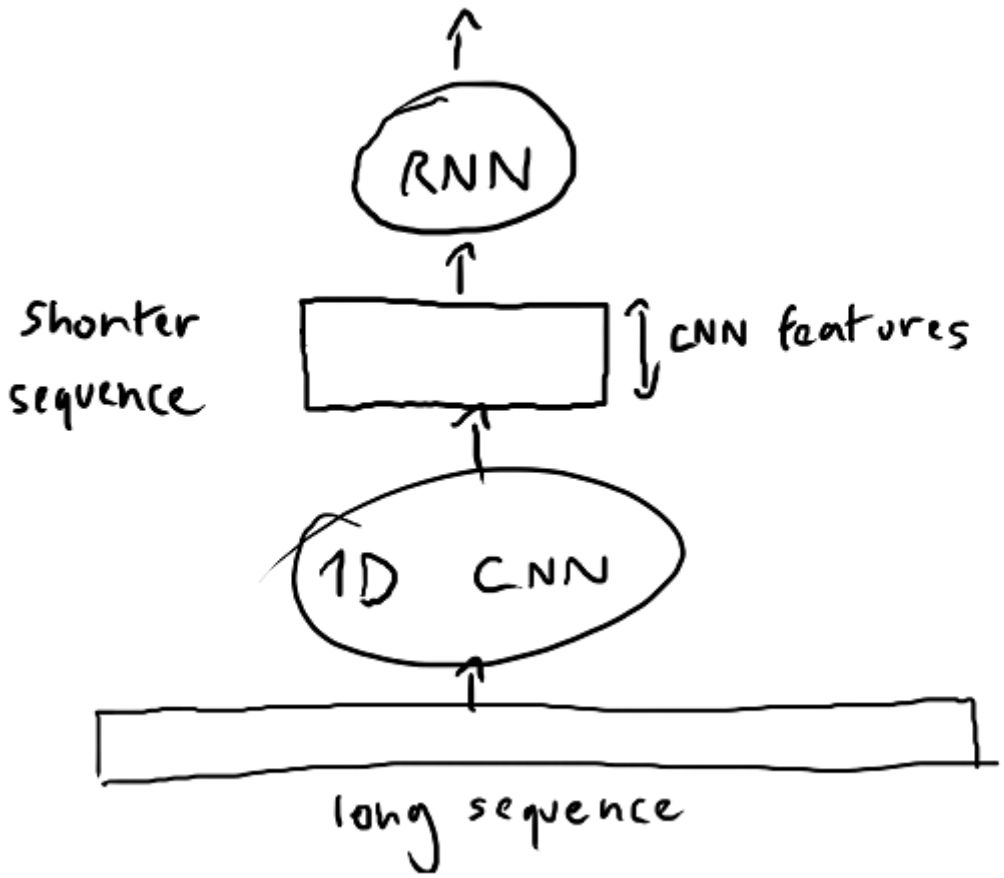
High Level Features





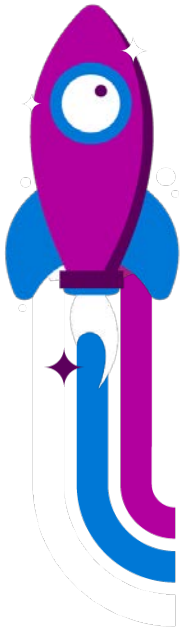
2DCNNS – SAME CONCEPT

Microsoft



STACKING IS COOL

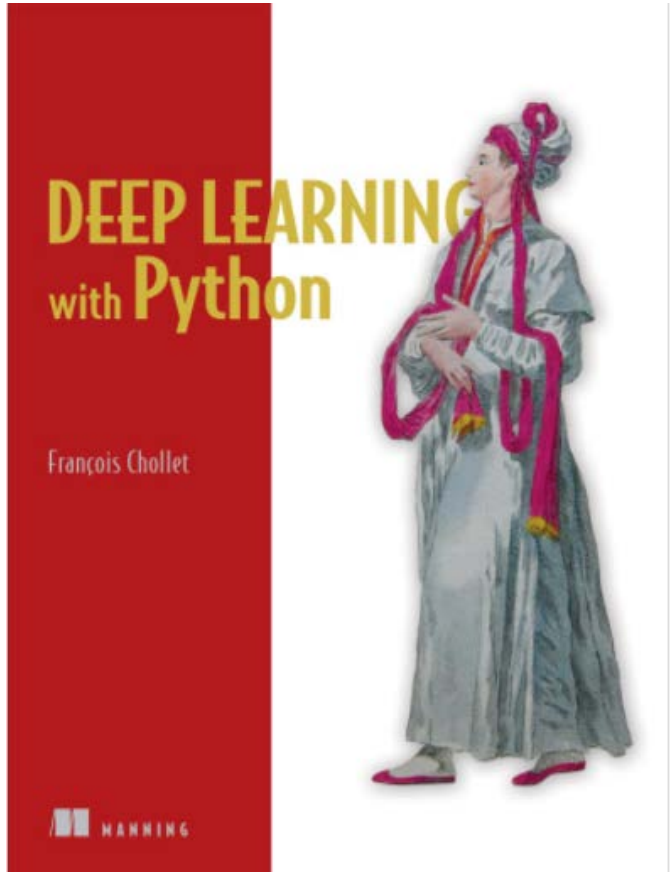
Microsoft



Demo of RNNs in Data Bricks/Keras

Microsoft

Francois Chollet is a living legend, please buy his book



https://www.manning.com/books/deep-learning-with-python

Microsoft

Tim Scarfe

@ecsquendor

youtube.com/machinelearningatmicrosoft

THANK YOU!

Azure

Microsoft