Deloitte.

Deep Learning and Neural Networks

Introduction and Use Cases

Ofer Shai

Agenda

Topic	Content	Timing
Introduction	• History, applications, and limitations of Deep Learning	10min
Deep Learning Basics (warning, some light math ahead)	• Perceptron, layers, activation function • Backpropagation • Use Case: Commercial Loss Insurance	15min
Unstructured Data	• Convolution Neural Networks • Use Case: Risk Analysis for Social Media	15min
Modeling Time Series Data	• Recurrent Networks and LSTM • Use Case: Credit Risk Modelling	10min
Building an AI Infrastructure	• Frameworks and considerations • Understanding network behaviour	10min

Key Takeaways

- Applications and limitations of modern Deep Learning
- How Neural Networks are built and trained
- Recognize Neural Network architectures and choose the right one for the job

Deloitte.

Putting the Neural in Networks

Introduction

Neural Networks Research Milestones

1958 – Rosenblatt creates the perceptron

1965 – Ivankhnenko first multi-layer network

1969 – Minsky and Papert – ANN limitations

1970’s to mid 80’s - symbolic models and expert systems

1960 – 1986 – backpropagation independently “discovered” in dynamic programming, control systems, Neural Networks

1986 – Rumelhart, Hinton, Williams - experimental results in hidden layers

1989 – LeCun et al. – Convolutional Neural Networks

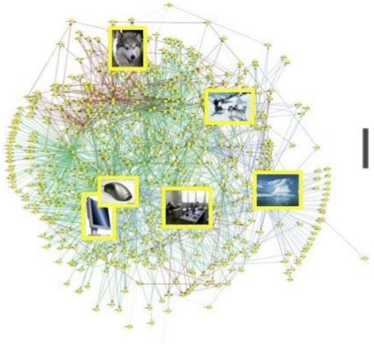
1992 – Williams and Zipser – Recurrent Neural Networks

1998 – Hochreiter and Schmidhuber – Long Short Term Memory

The birth of AI – Darmouth Conference 1956

Highlights of Artificial Intelligence and Machine Learning

2009 – ImageNet Released 14MM images, 21K categories



IMAGENET

2014 – DeepFace

2015 – Google, Microsoft, claim better than human image recognition

2017 – Google, Microsoft, Baidu, IBM, claim better than human speech recognition

2017 – Stanford claims better than human at detecting skin cancer

2018 – LawGeex claims better than human NDA review

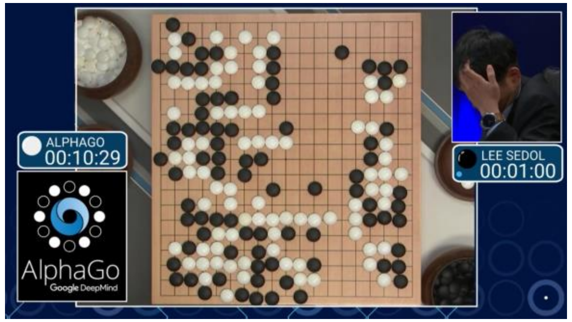
1997 – DeepBlue beats Kasparov

2011 – Watson beats Jennings

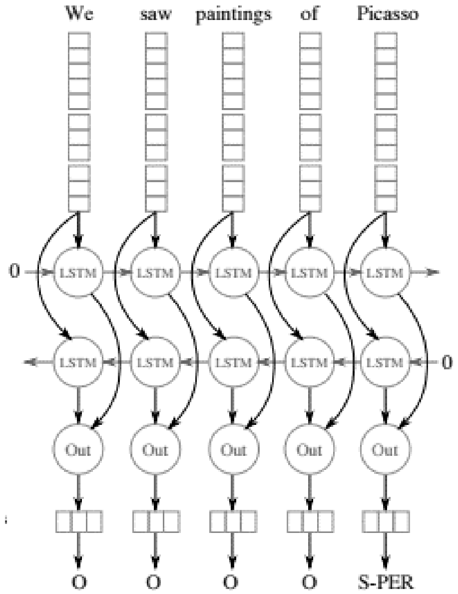
2015 – Google Gorilla Grudge

2016 – Mircrosoft Tay becomes racist in <24h

2016 – AlphaGo beats Sedol

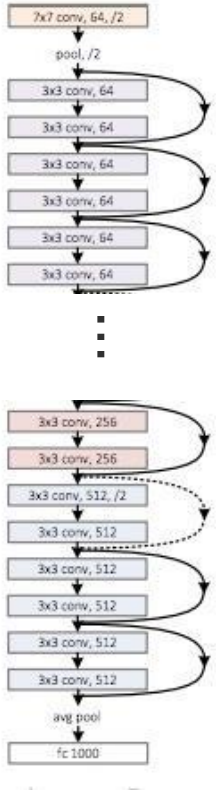


Network Engineering

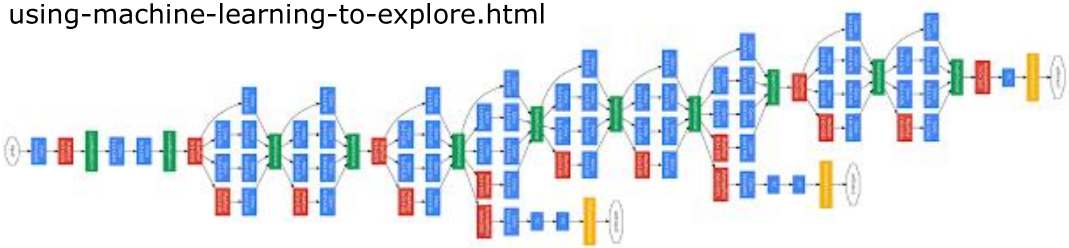


https://github.com/deeplearning4j/deeplearning4j/issues/4513

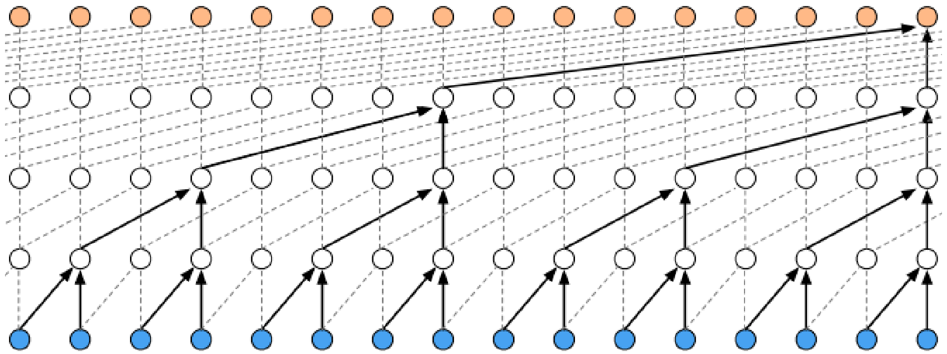
http://mynotes2ml.blogspot.ca/2016/07/residual-networks-resnet.html

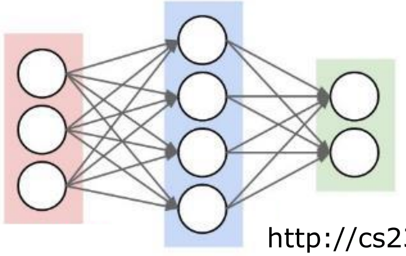


https://research.googleblog.com/2017/05/using-machine-learning-to-explore.html



http://sergeiturukin.com/2017/03/02/wavenet.html



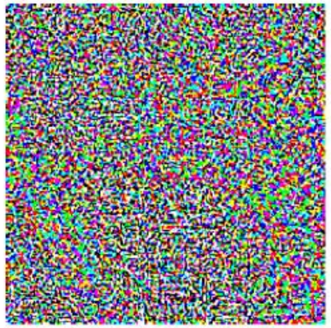


http://cs231n.github.io/neural-networks-1/

The Fallibility of Deep Learning

https://arxiv.org/abs/1712.07107

$+ .007 \times$



$=$

Panda
57.5%

Gibbon
99.3%

Face | No Face | Added Noise

https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Adversarial attacks have been shown to reliably fool networks. We now use them to get better understanding on what is actually learned by the network, and improve model robustness and stability

Real Life Adversarial Attacks

Using 3D printed glasses, researchers were able to fool facial detection and recognition systems.

Avoidance

Changing gender, race

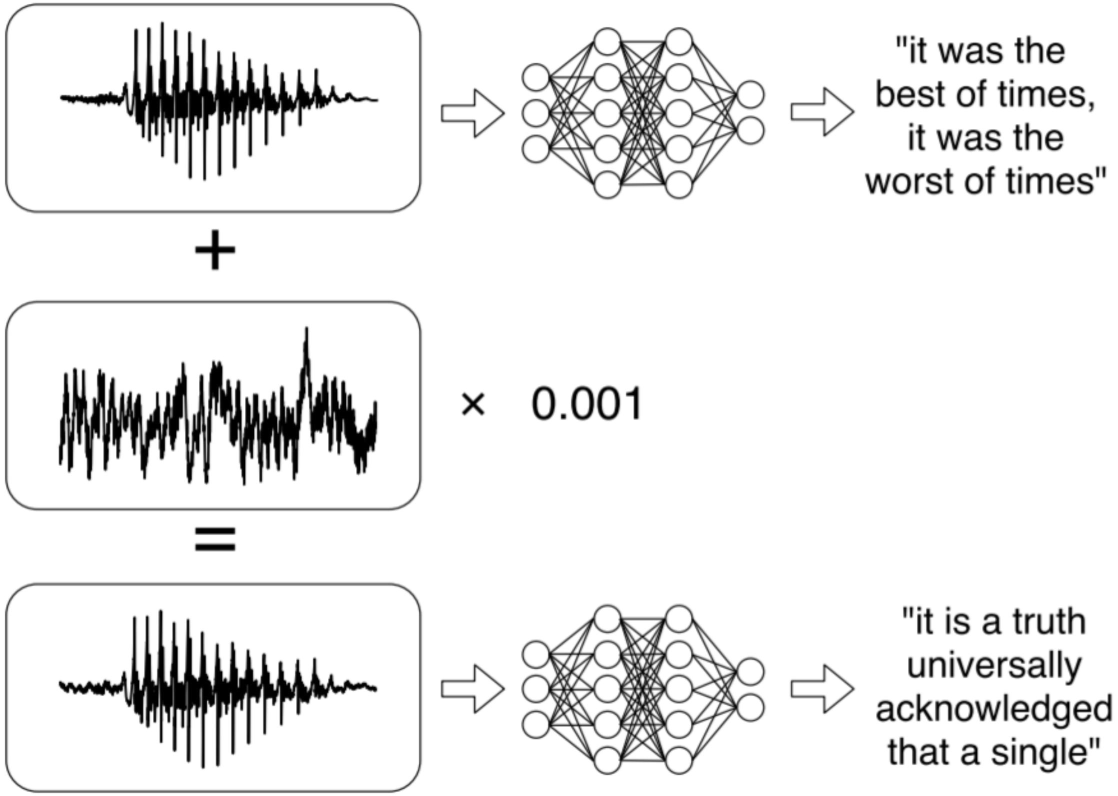
https://www.cs.cmu.edu/~sbhagava/papers/face-rec-ccs16.pdf

Adversarial Attacks in Other Fields

https://arxiv.org/abs/1707.07328

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

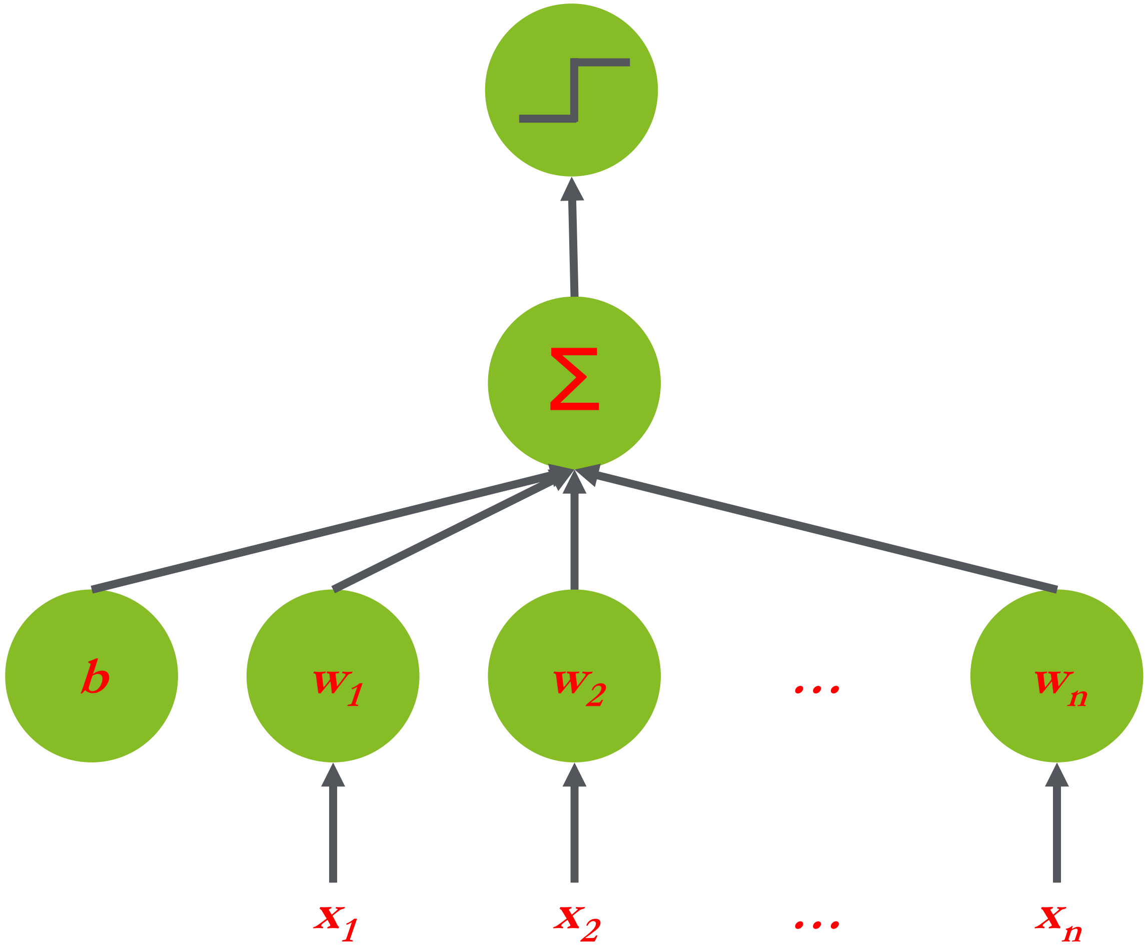
https://arxiv.org/abs/1801.01944



Deloitte.

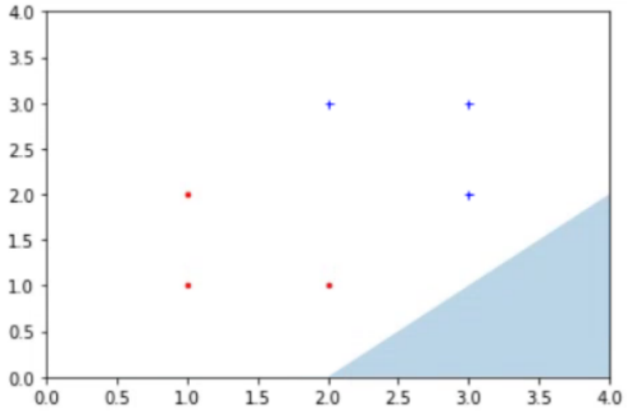
Warning:
Some Math
ahead

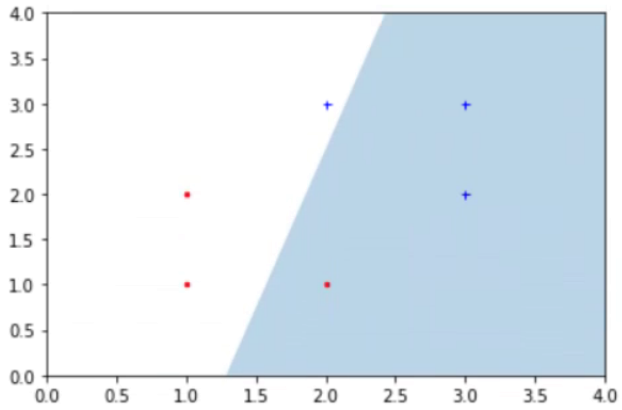
The Perceptron



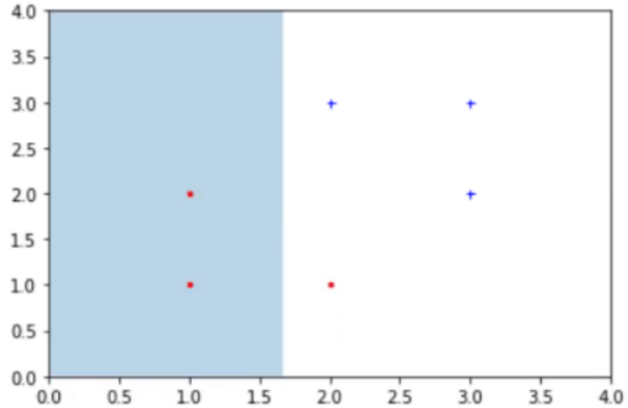

```
Evaluate data point i
If not correct
     If positive
          Add to weights
     If negative
          Subtract from weights
Repeat for all points
```

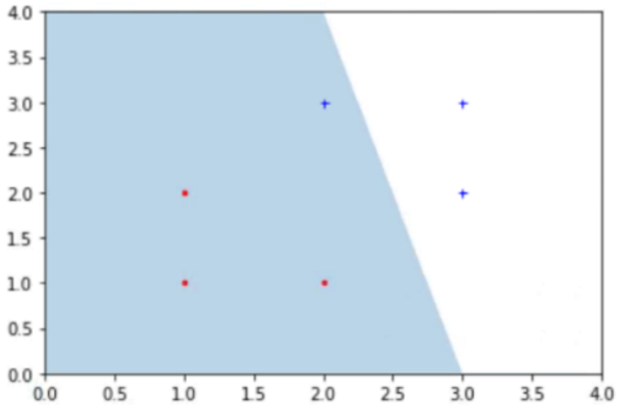
Perceptron Learning

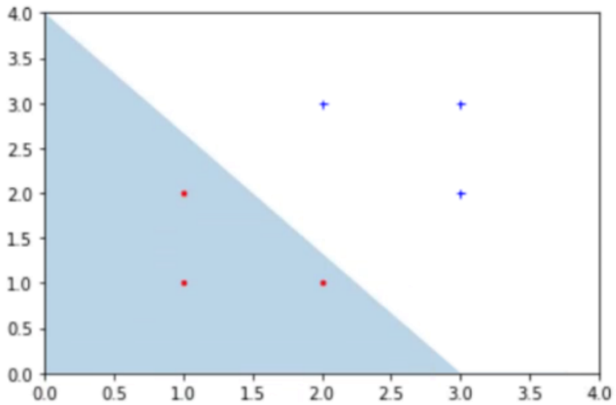





```
Evaluate data point i
If not correct
    If positive
        Add to weights
    If negative
        Subtract from weights
Repeat for all points
```

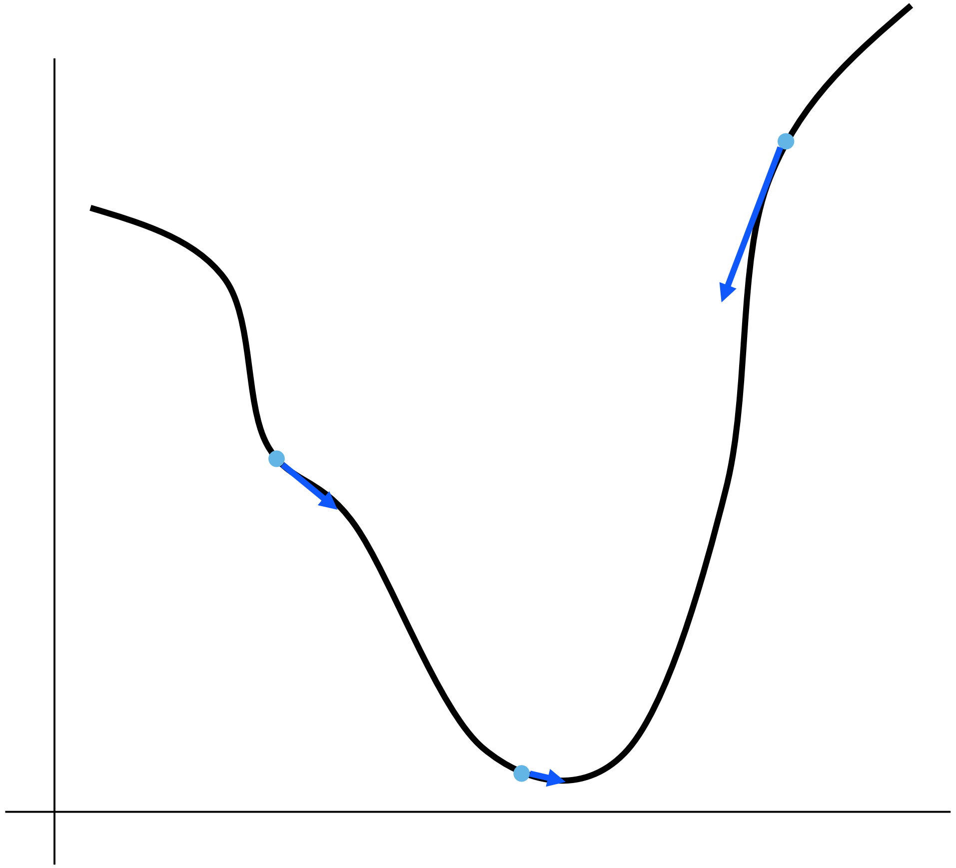


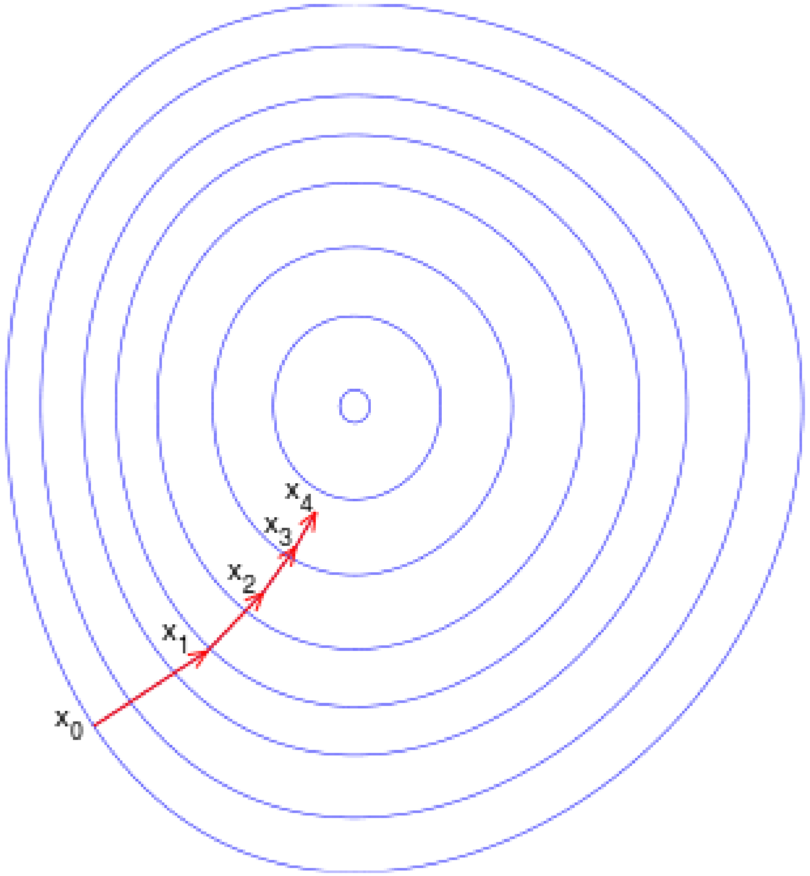




Segway – Gradient Descent

Challenge – How to find an optimal configuration?

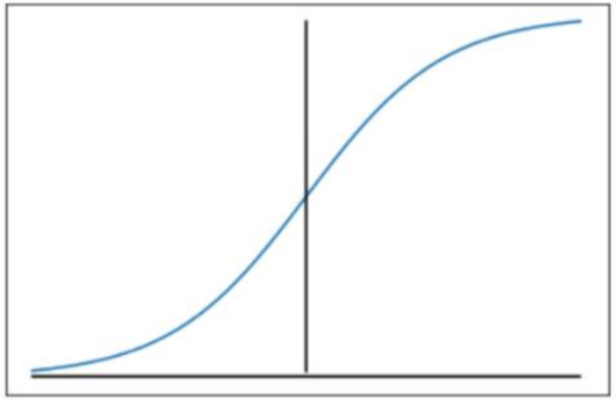




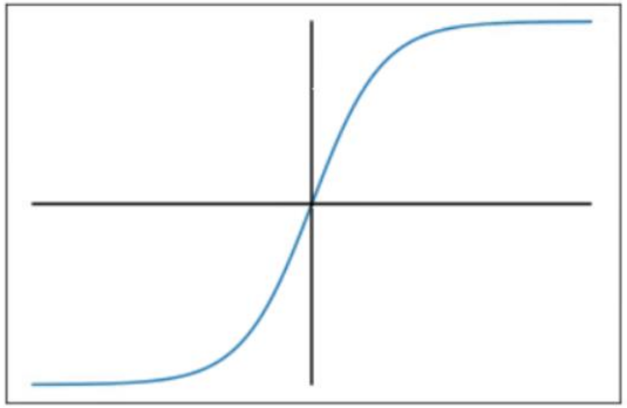
https://en.wikipedia.org/wiki/Gradient_descent

Activation Functions and Backpropagation

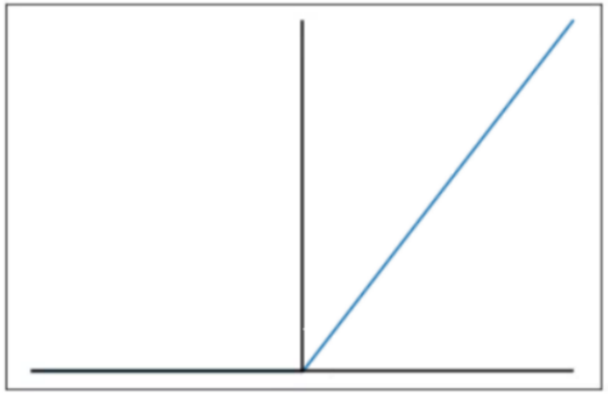
Sigmoid



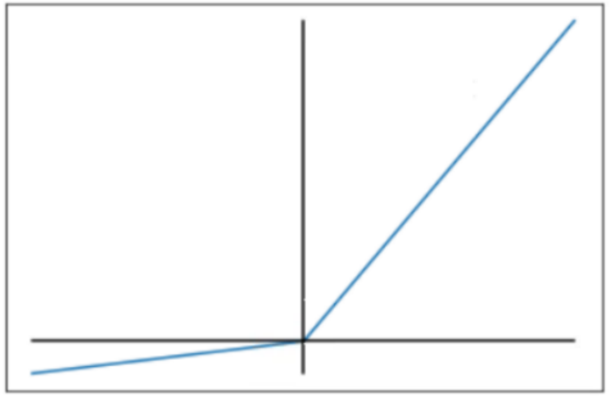
tanh



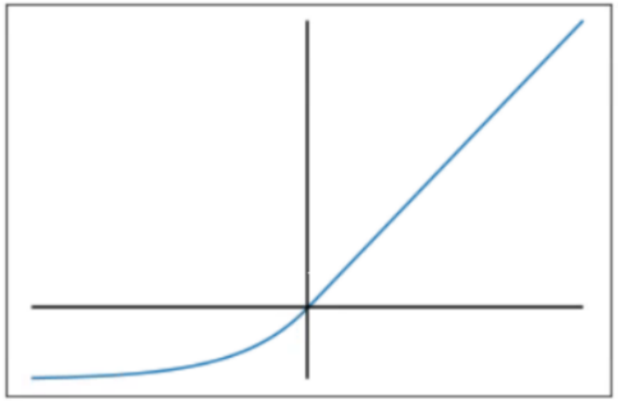
Rectified linear unit (ReLU)



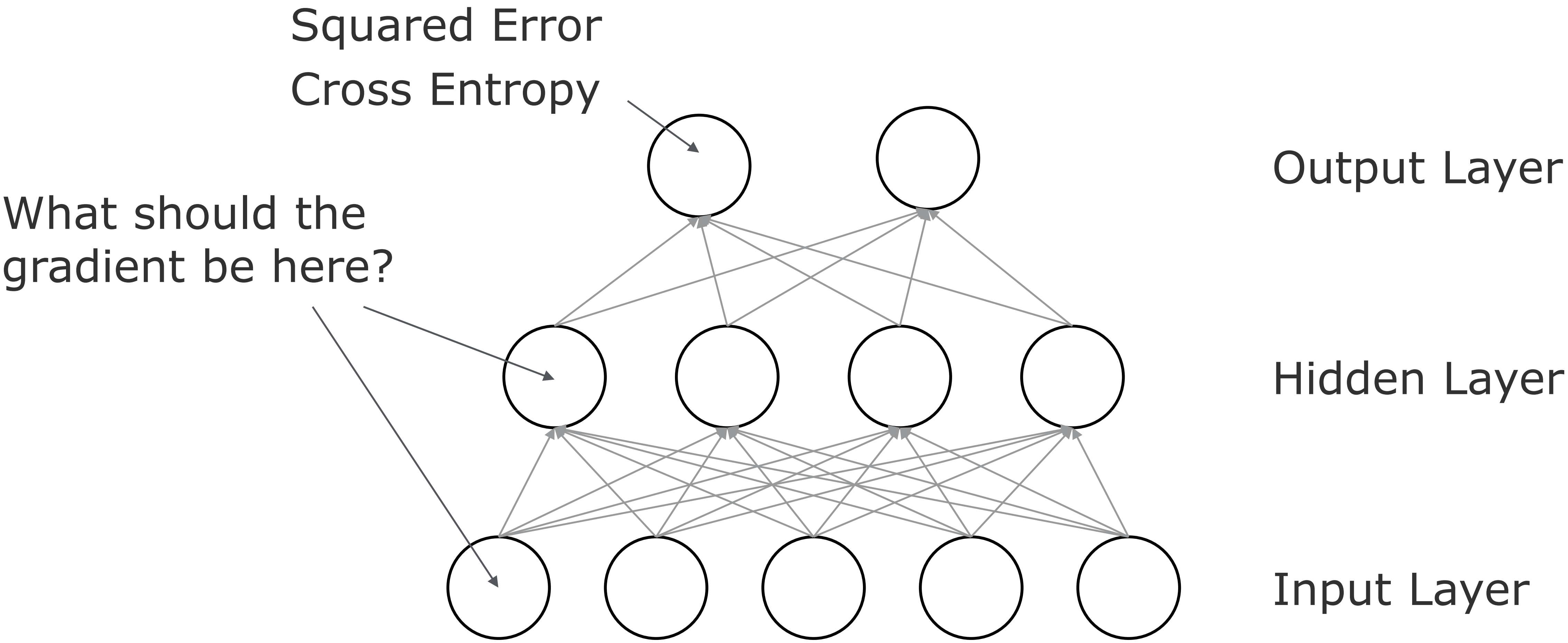
Leaky ReLU



Exponential ReLU



Computing Gradients



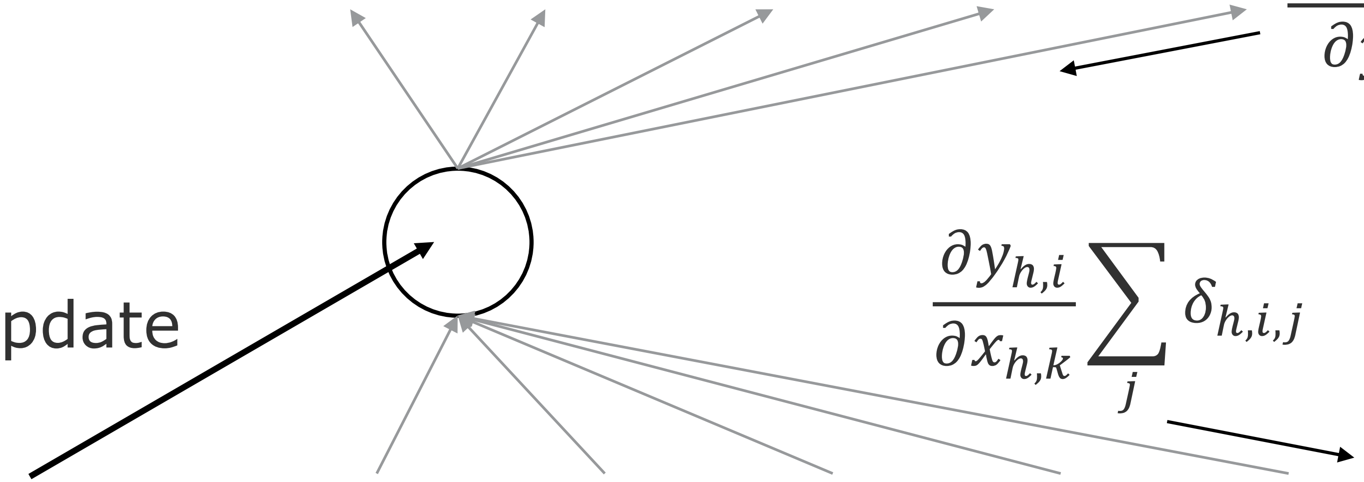
Backpropagation

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial z}\frac{\partial z}{\partial x}$$

Node i in layer h

$$f(\boldsymbol{x}_h) = y_{h,i}$$

$$\frac{\partial y_{h+1,j}}{\partial y_{h,i}} = \delta_{h,i,j}$$



$$\frac{\partial y_{h,i}}{\partial x_{h,k}} \sum_j \delta_{h,i,j}$$

Update

$$\frac{\partial y_{h,i}}{\partial \boldsymbol{w}_{h,i}} \sum_j \delta_{h,i,j}$$

$$x_{h,k} = \boldsymbol{y}_{h-1,k}$$

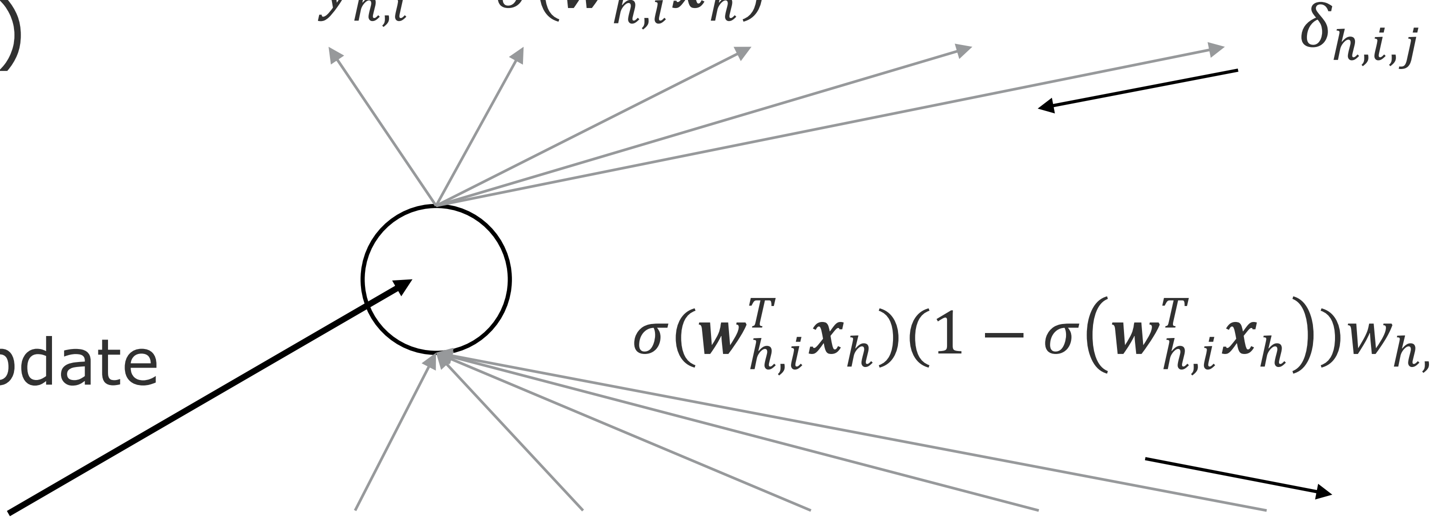
Backpropagation

$$\sigma(\boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{x}}}$$

$$\frac{\partial\sigma(\boldsymbol{x})}{\partial\boldsymbol{x}} = \sigma(\boldsymbol{x})(1\text{-}\sigma(\boldsymbol{x}))$$

Node i in layer h

$$y_{h,i} = \sigma(\boldsymbol{w}_{h,i}^T\boldsymbol{x}_h)$$



$$\sigma(\boldsymbol{w}_{h,i}^T\boldsymbol{x}_h)(1-\sigma(\boldsymbol{w}_{h,i}^T\boldsymbol{x}_h))w_{h,i,k}\sum_j \delta_{h,i,j}$$

$$\sigma(\boldsymbol{w}_{h,i}^T\boldsymbol{x}_h)(1-\sigma(\boldsymbol{w}_{h,i}^T\boldsymbol{x}_h))x_{h,i,k}\sum_j \delta_{h,i,j}$$

Commercial Loss Insurance Use Case

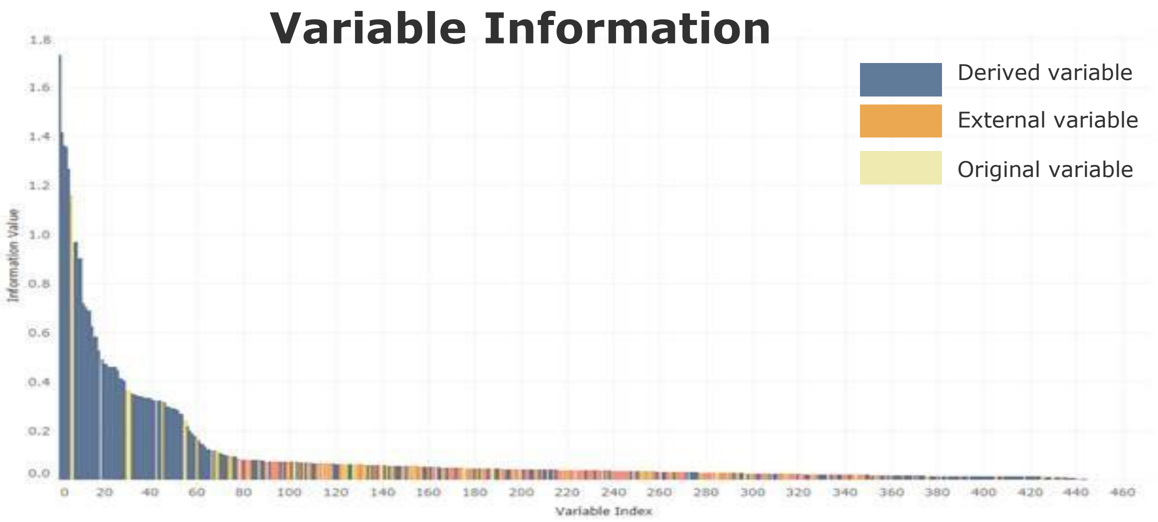
Input data

- 79 variables, including policy characteristics and external data
- 176 additional external variables
- 226 derived variables
- Adding aggregated external data gives 2% performance gain

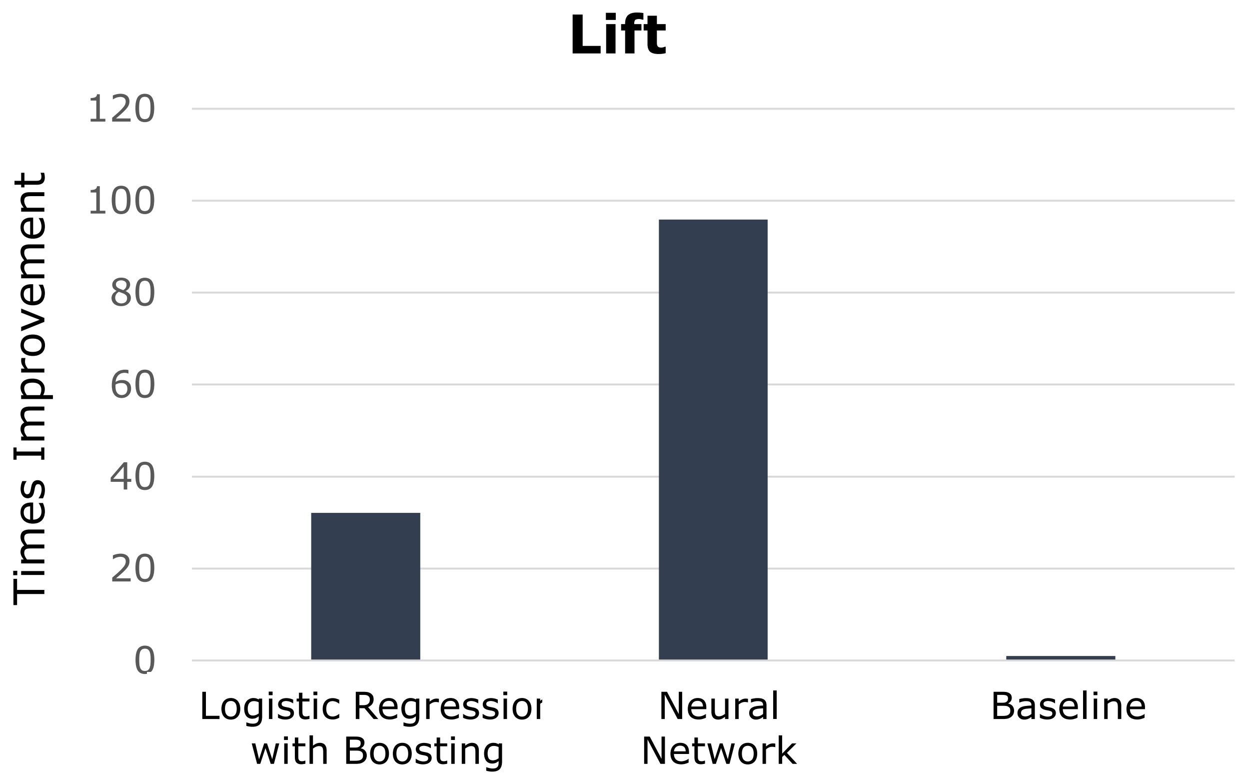
Data structure and limitations

- *Extremely rare event prediction (~0.01% records)*
- Look-ahead variables

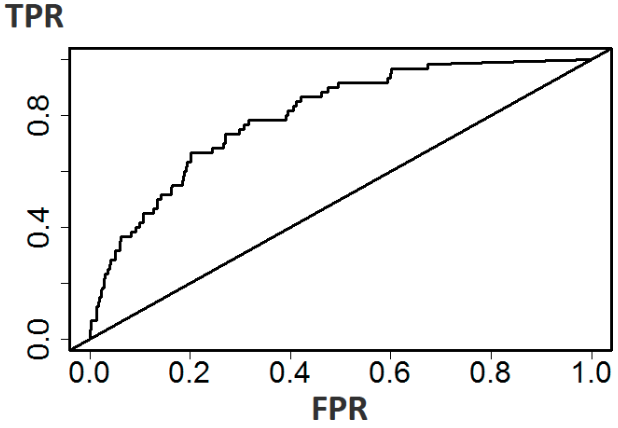
Variable Group	Max Information	Average Information
Coverage	1.73	0.60
Deductible	0.58	0.31
Industry	0.30	0.15
Policy	0.45	0.09
Census	0.06	0.04
Income	0.08	0.04
Geographic	0.12	0.03



Comparative Performance

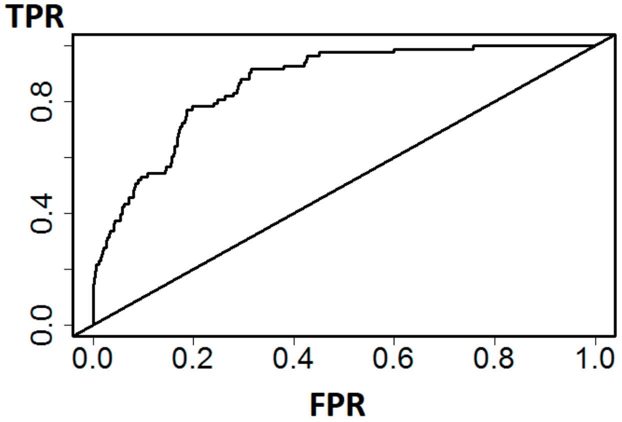


Logistic Regression



Precision 0.002,
recall 0.002s,
F1_score 0.004

Neural Network



Precision 0.006,
recall 0.220,
F1_score 0.012

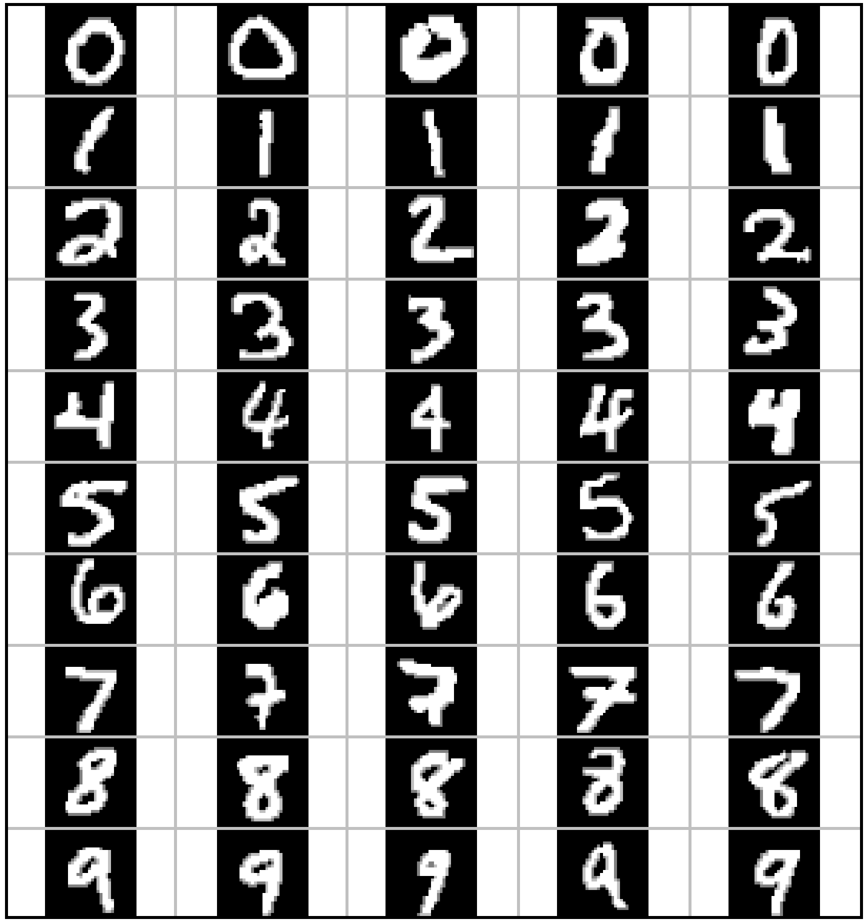
Dealing with Unstructured Data

Dealing with Unstructured Data

Why not just use our standard modelling approach?

- Positional information is meaningless
- Data size
- Complex interactions
- Representing information

http://www-edlab.cs.umass.edu/~smaji/cmpsci670/fa14/hw/recognition/index.html



http://inst.eecs.berkeley.edu/~cs188/fa06/projects/classification/4/writeup/index.html

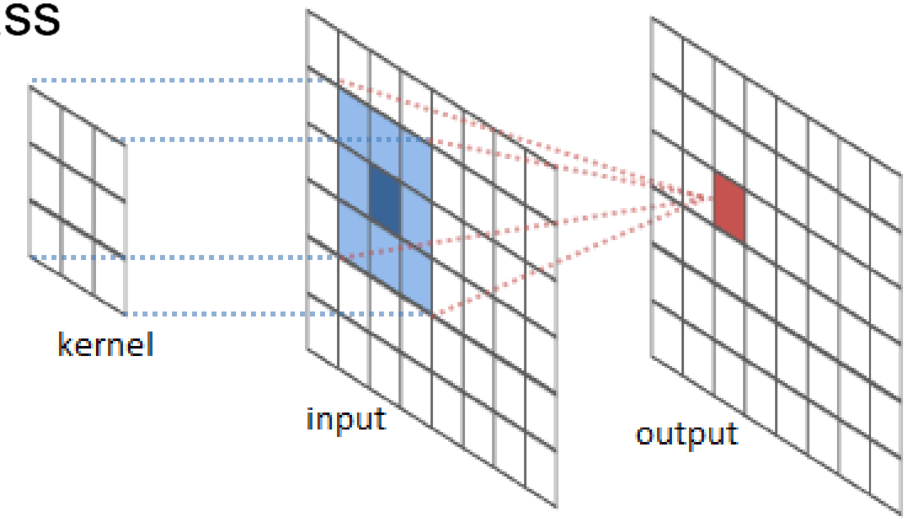
Dealing with Unstructured Data

https://twitter.com/teenybiscuit/status/707727863571582978

You Know Convolutions

http://blog.teledynedalsa.com/2012/05/image-filtering-in-fpgas/

A small filtering kernel is applied to entire data set, creating distorted views of the data that can extract various elements of the information



https://www.kaggle.com/ttungl/exercise-convolutions-for-computer-vision

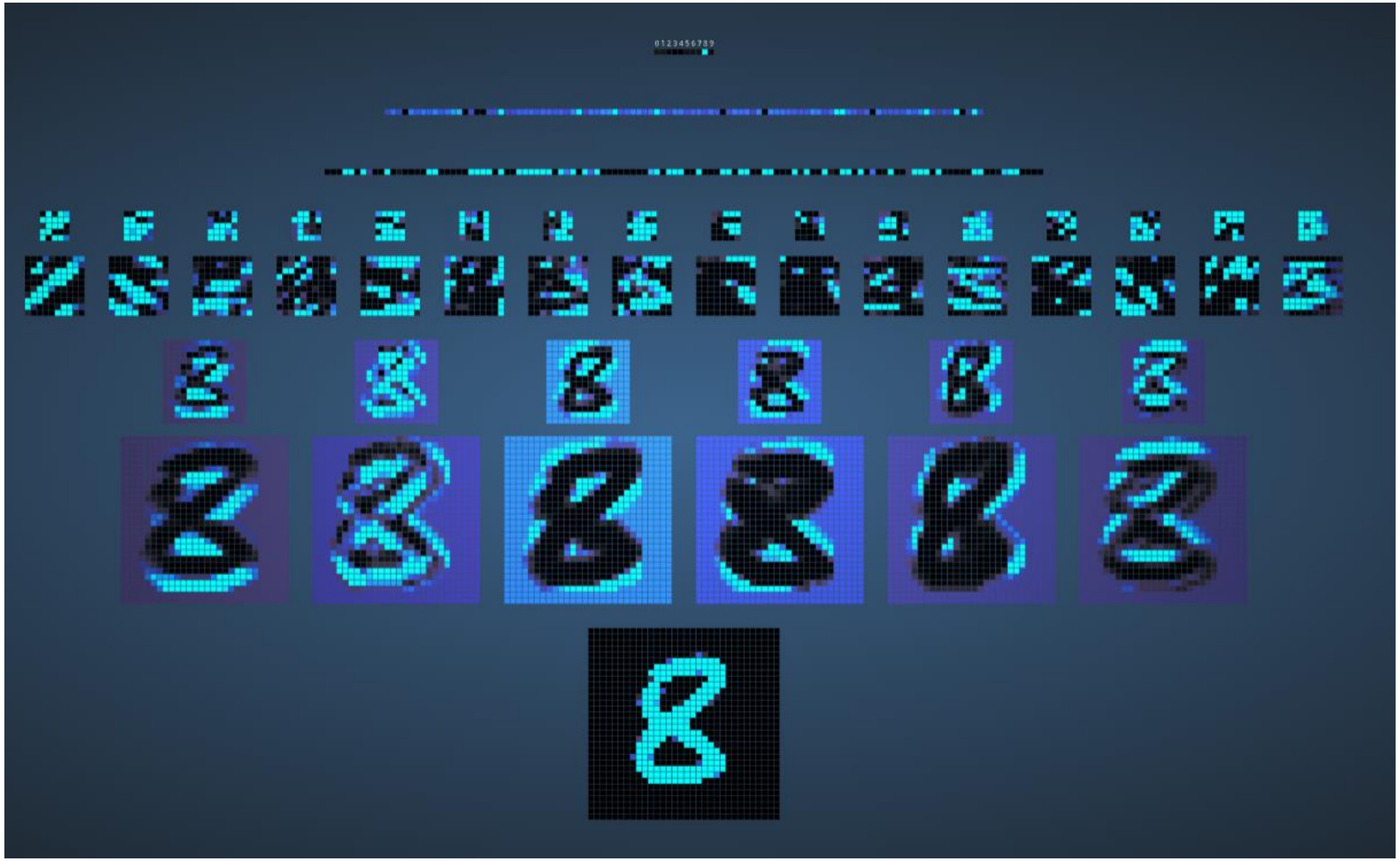
Automatic Feature Creation Through Convolution

Successive layers of convolution and down sampling create high level features for the model

Learning is carried out using Backpropagation

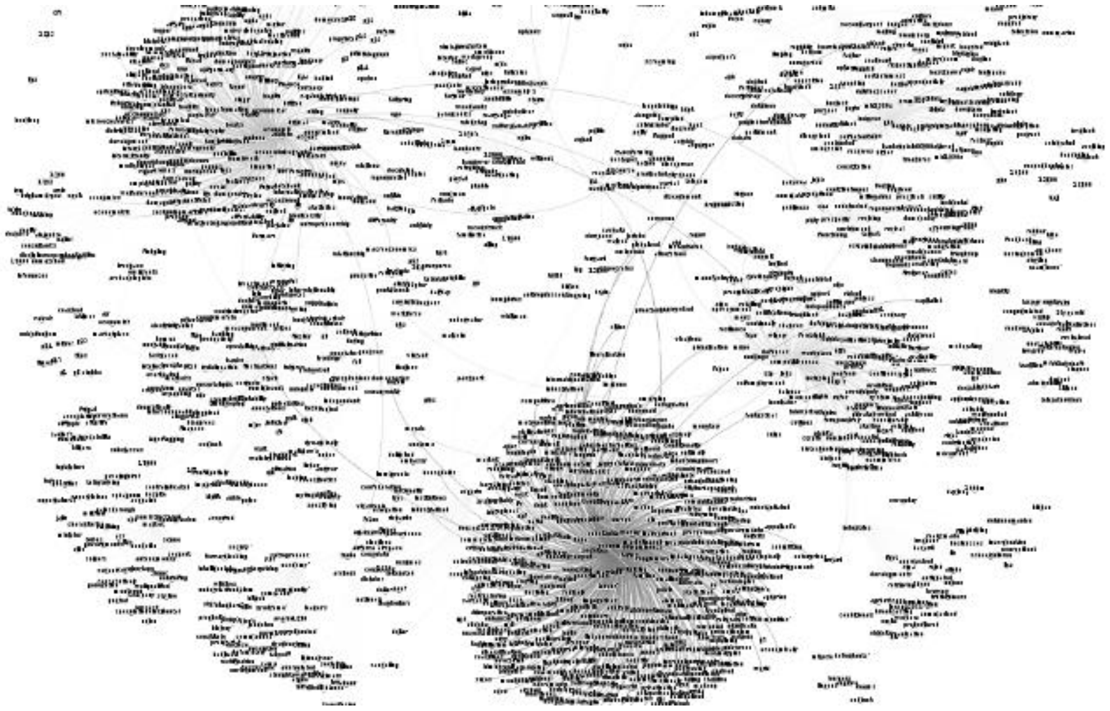
Transfer Learning - Fixing the network except for the last few layers allow for training on similar but different tasks

http://scs.ryerson.ca/~aharley/vis/conv/flat.html

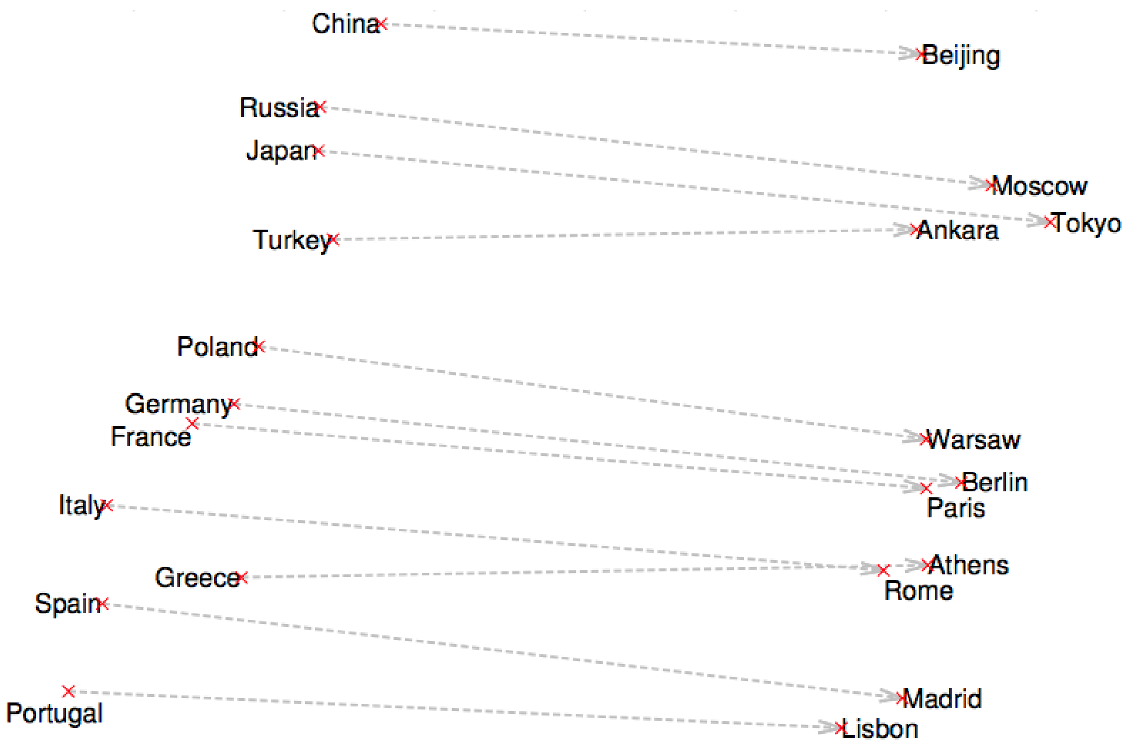


Segway – Word Embedding

https://medium.com/data-science-at-home/word-embedding-explained-in-one-slide-b2fe6b79ca55



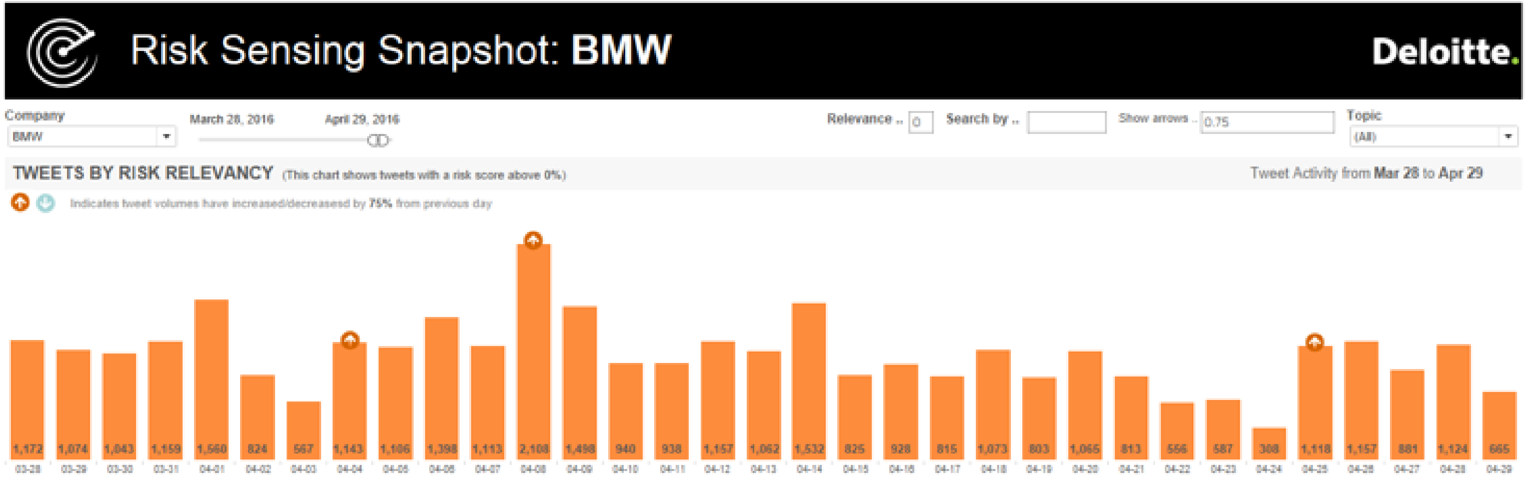
http://www.fafadiatech.com/blog/nlp/tools/opensource/machinelearning/2016/05/31/5-variants-of-word-embeddings.html

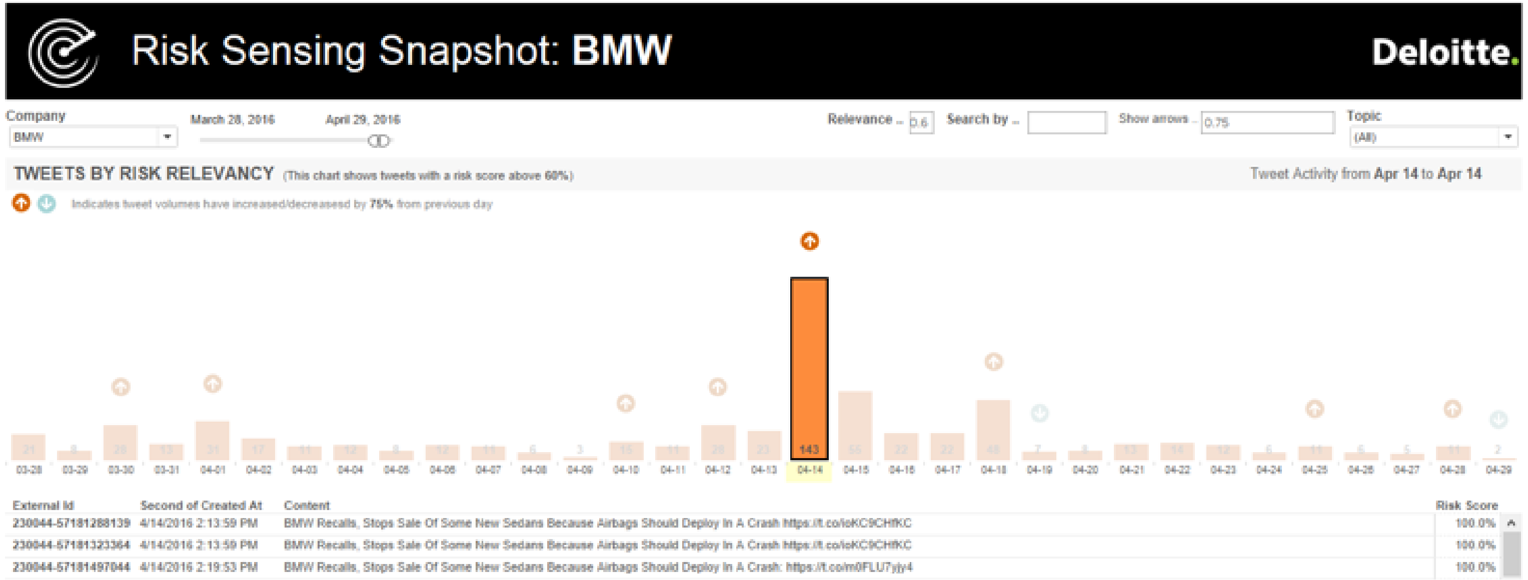


Market Sensing Use Case

(Social) Media offers a rich source of information about market trends, PR, and commercial "Life Events"

Targeted models can identify risks for clients, or monitor competition





Deloitte.

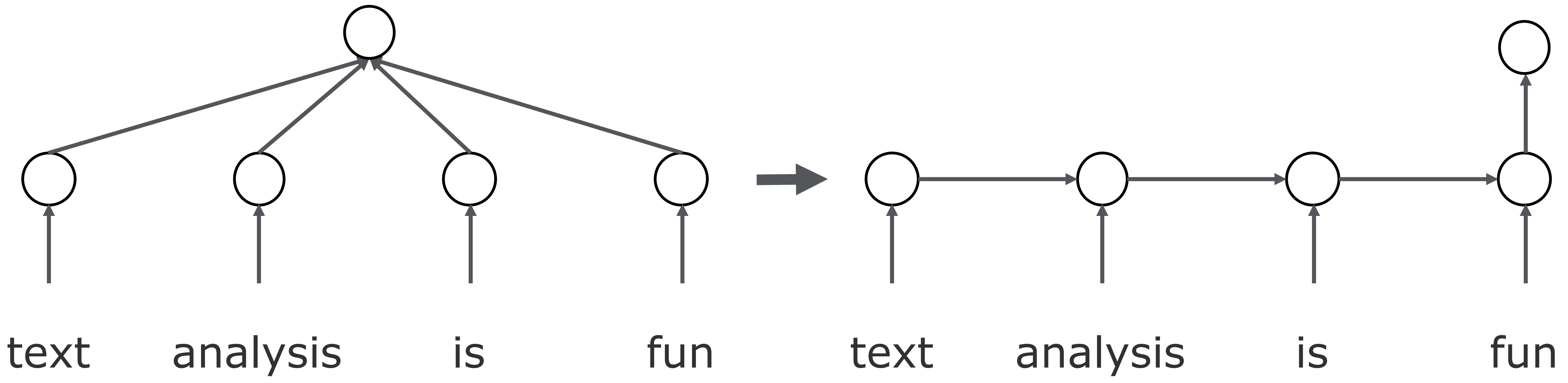
Natural Representation of Time Series

Time Series and Sequential Data Modelling

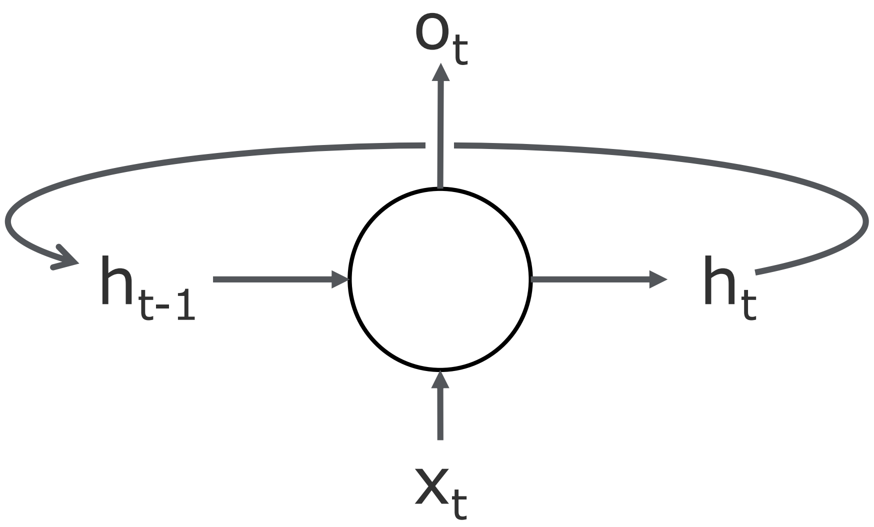
<Comic Strip Removed>

http://www.giantitp.com/comics/oots0003.html

Time Series and Sequential Data Modelling



Recurrent Neural Networks (RNN)



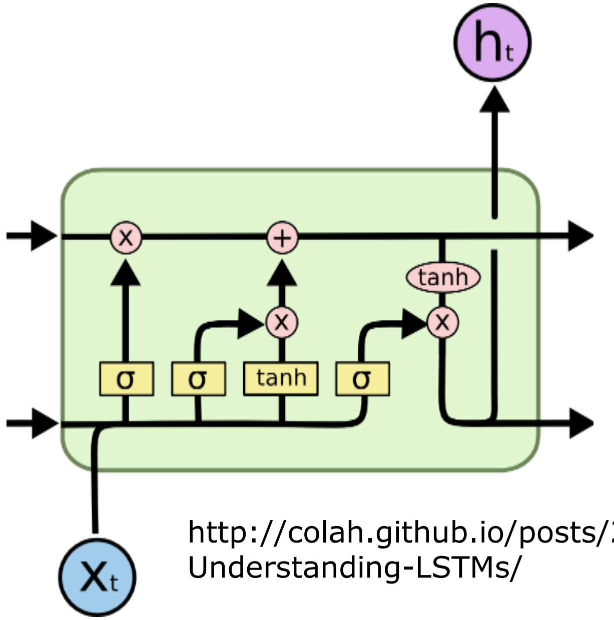
- RNN more naturally model sequential data
- More naturally handle input of varying length
- Many-to-one or many-to-many
- Can get future signals using backward chains
- Long-short-term memory cells effective for capturing long term dependencies

X Doesn't always beat CNN for accuracy

X Slower

✓ More robust, requires less tuning

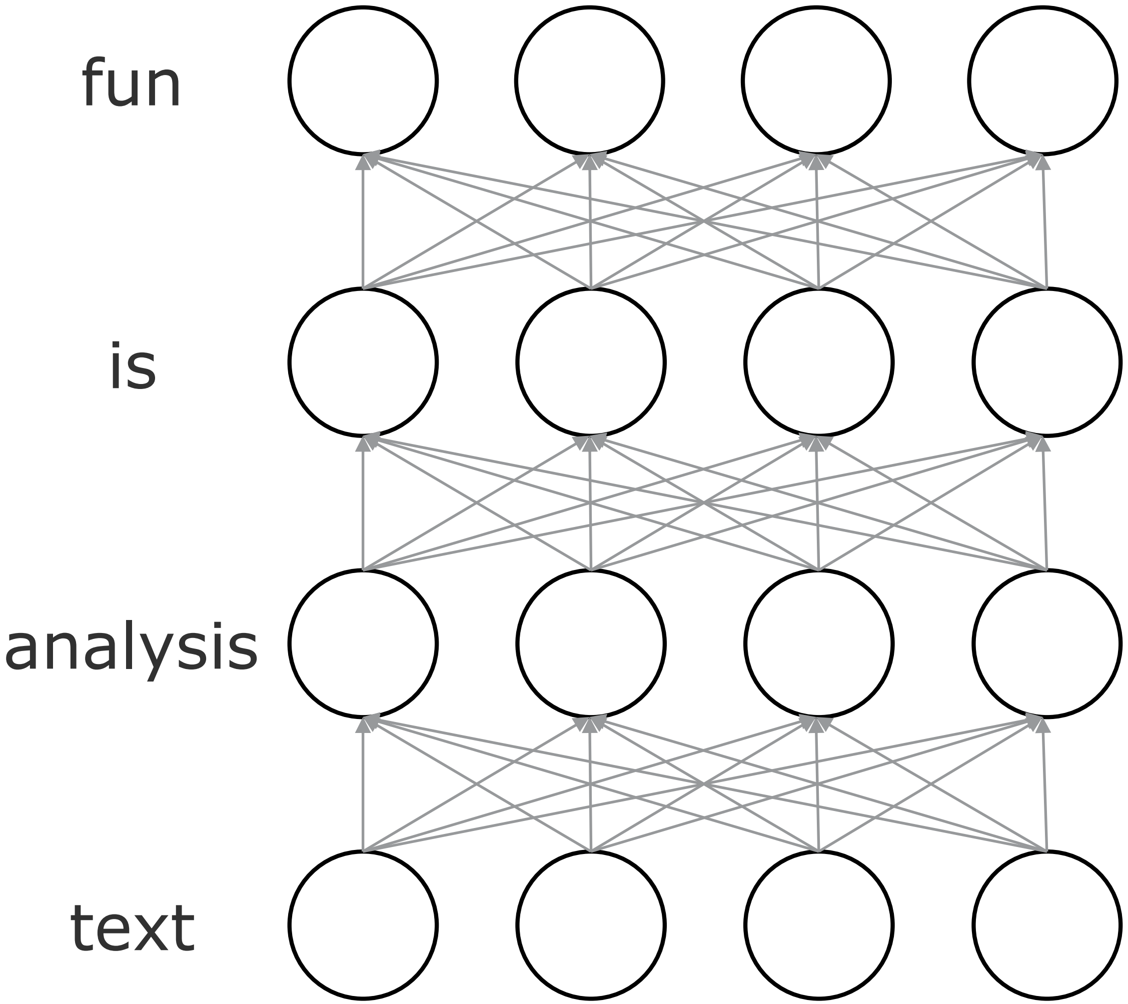
✓ Fewer parameters



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Backpropagation Through Time

- Accumulate gradients through times steps, and update all parameters
- Conceptually – unroll the network
- For long sequencing, use truncated backpropagation



Credit Risk Modelling Use Case

Input data

- Credit card transaction data
- Payment data
- Client Credit Score
- Delinquencies and Bankruptcies

Data structure and limitations

- 40k examples, over sampling of “bad” examples
- Did not use balance data, granular merchant codes

Variables		
Timestamp	Amount	Coarse Merchant code
Country	Phone / Internet / Store	

High Risk Merchant Codes

Dating Services

Quasi-cash

Betting

Low Risk Merchant Codes

Contractors

Tourist Attractions

Orthopedic goods, Optometrists

Credit Risk Modelling Use Case

Current Model \ Proposed Model	Included	Excluded
Included	Population **1.1MM** 95% PBAD **.71%**	Population **13k** 1% PBAD **10.8%**
Excluded	Population **13k** 1% PBAD **4.8%**	Population **27k** 3% PBAD **19.1%**

Model	AUC
Credit Score	0.9067
CNN	0.9166
RNN	0.9172

Swap Out

Bad rate:	10.8%
Total losses:	$6.4MM
Total exposure:	$65MM
Avg. credit limit:	$4,900

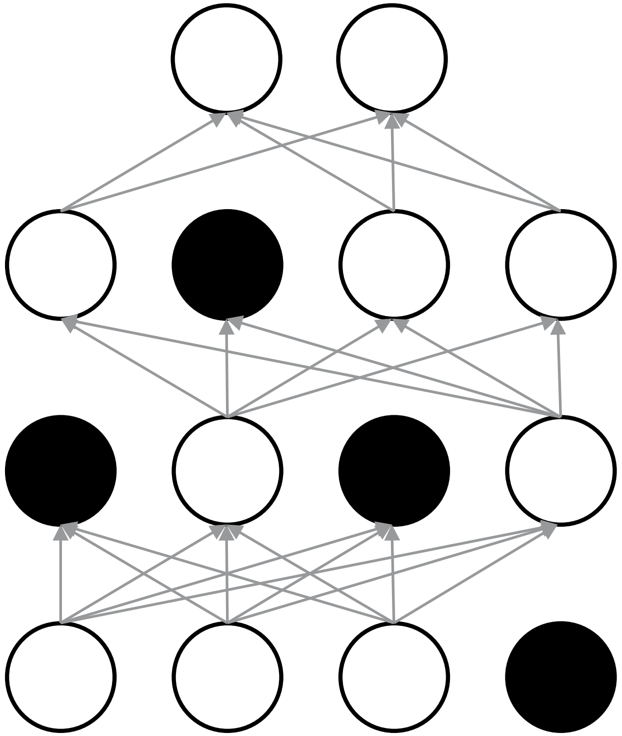
Swap In

Bad rate:	4.8%
Total losses:	$2.4MM
Total exposure:	$45MM
Avg. credit limit:	$3,400

been

Advanced topics

Dropout, Attention, Capsules, Residuals, and more



https://www.quora.com/What-is-Attention-Mechanism-in-Neural-Networks

http://vitakem.com/capsule-manufacturer/

Deloitte.

Neural Networks In Practice

Building an AI infrastructure

Deep Learning Frameworks

DL4J

TensorFlow™

PYTORCH

Caffe2

Microsoft CNTK

mxnet

Responsible AI – Not just for Neural Networks

The data

The model

Production Environment

The Implications of AI Systems

Responsible AI – Not just for Neural Networks

The data

- Is the training data representative and comprehensive?
- Is the validation data representative and comprehensive?
- Is the production data consistent?

The model

Production Environment

The Implications of AI Systems

Responsible AI – Not just for Neural Networks

The data

- Is the training data representative and comprehensive?
- Is the validation data representative and comprehensive?
- Is the production data consistent?

The model

- Are there unintended biases?
- Do the signals make sense?
- Would it behave badly outside the learned parameters?

Production Environment

The Implications of AI Systems

Responsible AI – Not just for Neural Networks

The data

- Is the training data representative and comprehensive?
- Is the validation data representative and comprehensive?
- Is the production data consistent?

The model

- Are there unintended biases?
- Do the signals make sense?
- Would it behave badly outside the learned parameters?

Production Environment

- How often do we need to retrain the model?
- How do we monitor performance on-going?
- Can we trace accountability for decisions?

The Implications of AI Systems

Responsible AI – Not just for Neural Networks

The data

- Is the training data representative and comprehensive?
- Is the validation data representative and comprehensive?
- Is the production data consistent?

The model

- Are there unintended biases?
- Do the signals make sense?
- Would it behave badly outside the learned parameters?

Production Environment

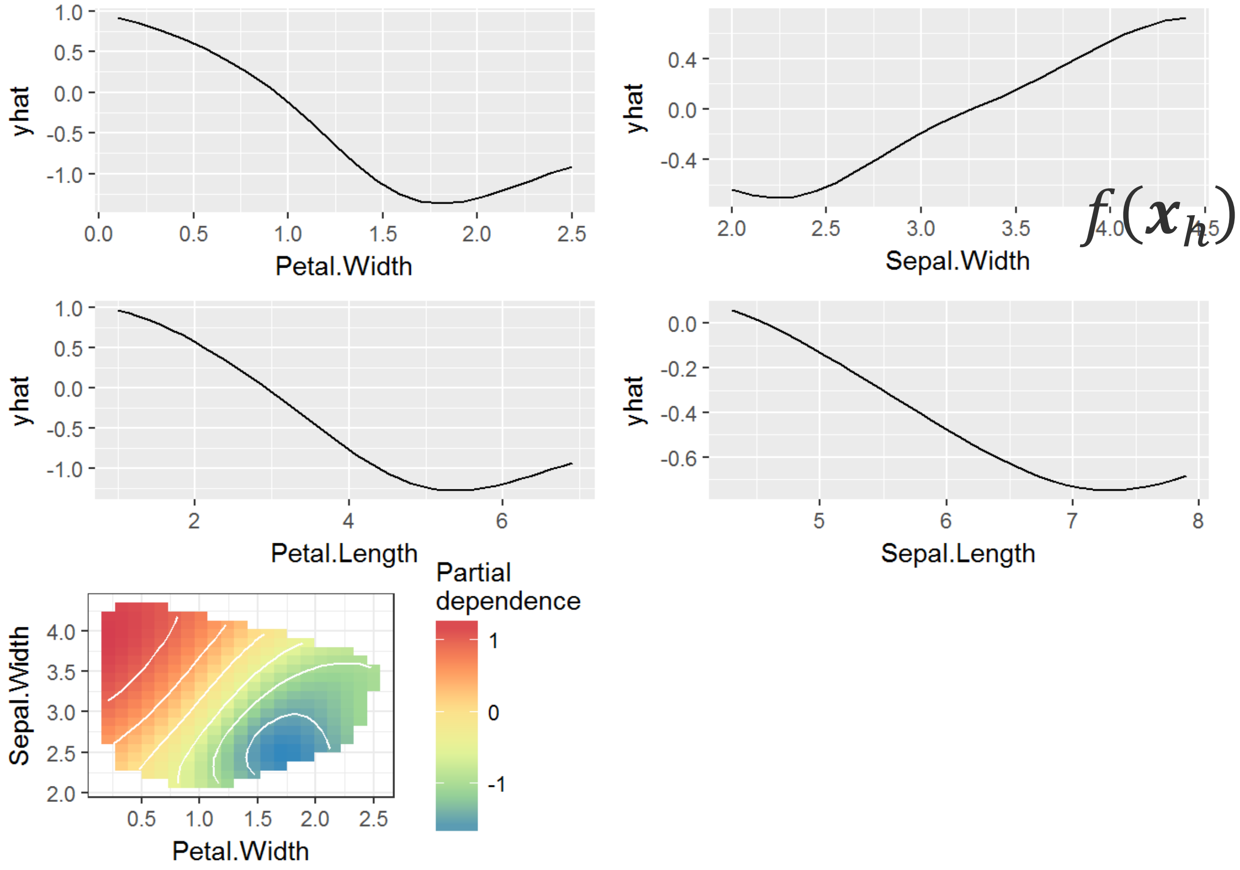
- How often do we need to retrain the model?
- How do we monitor performance on-going?
- Can we trace accountability for decisions?

The Implications of AI Systems

- Are inherent biases propagated?
- Is there value beyond what other approaches bring?
- Is this better or worse than human decision making?

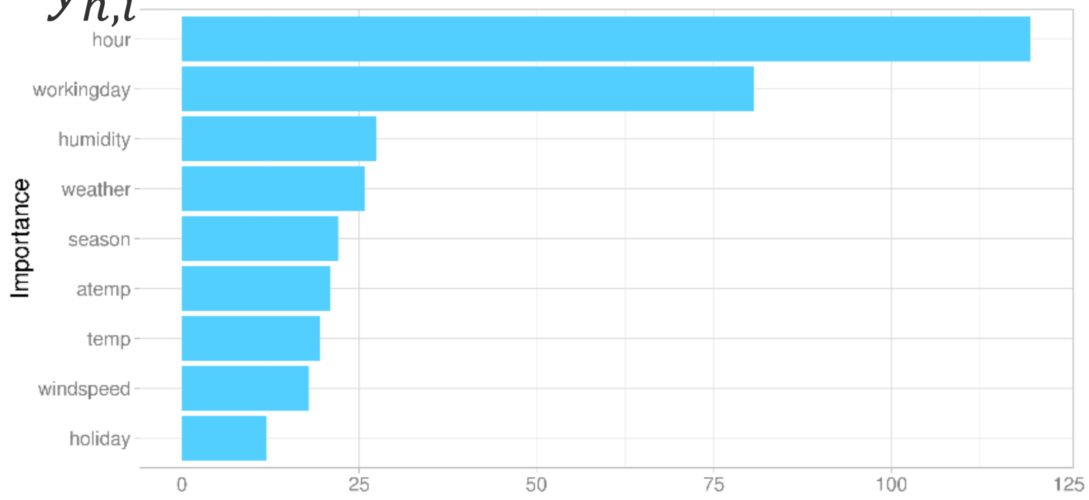
Peering Inside The Black Box

Partial Dependence Plot



$$f(x_h) = y_{h,i}$$

Feature Importance



http://rstudio-pubs-static.s3.amazonaws.com/283647_c3ab1ccee95a403ebe3d276599a85ab8.html

https://www.kaggle.com/general/13285

<Comic Strip Removed>