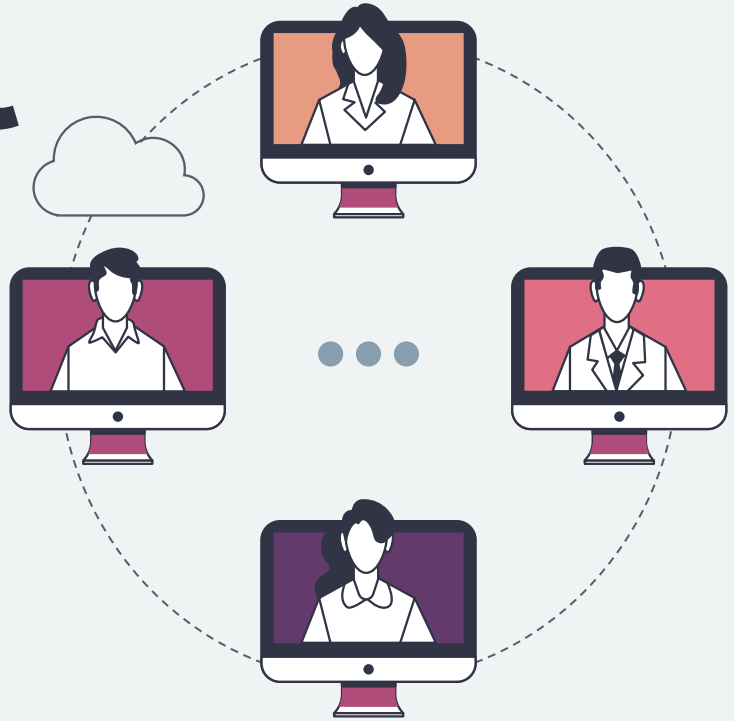
One model to rule them all

Achieve state-of-the-art performance with less data than usually on multiple tasks by ***transfer learning***



Hampus Londögård
AFRY
ÅF PÖYRY

Agenda

1 Machine Learning

2 Transfer Learning

3 Practical use-case

4 Final Words

01

Machine Learning

De-mystifying something a lot are afraid of

Illustration by Freepik Stories

What Is Possible?

What Is Possible?

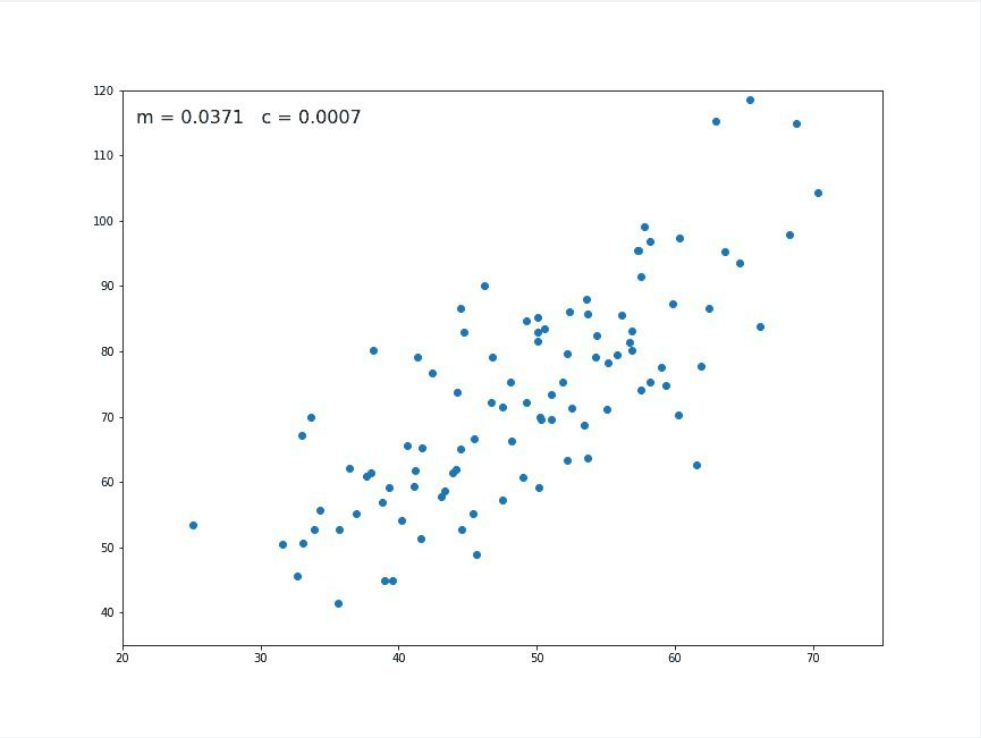
What Is Possible?

What Is Possible?

How is this possible?

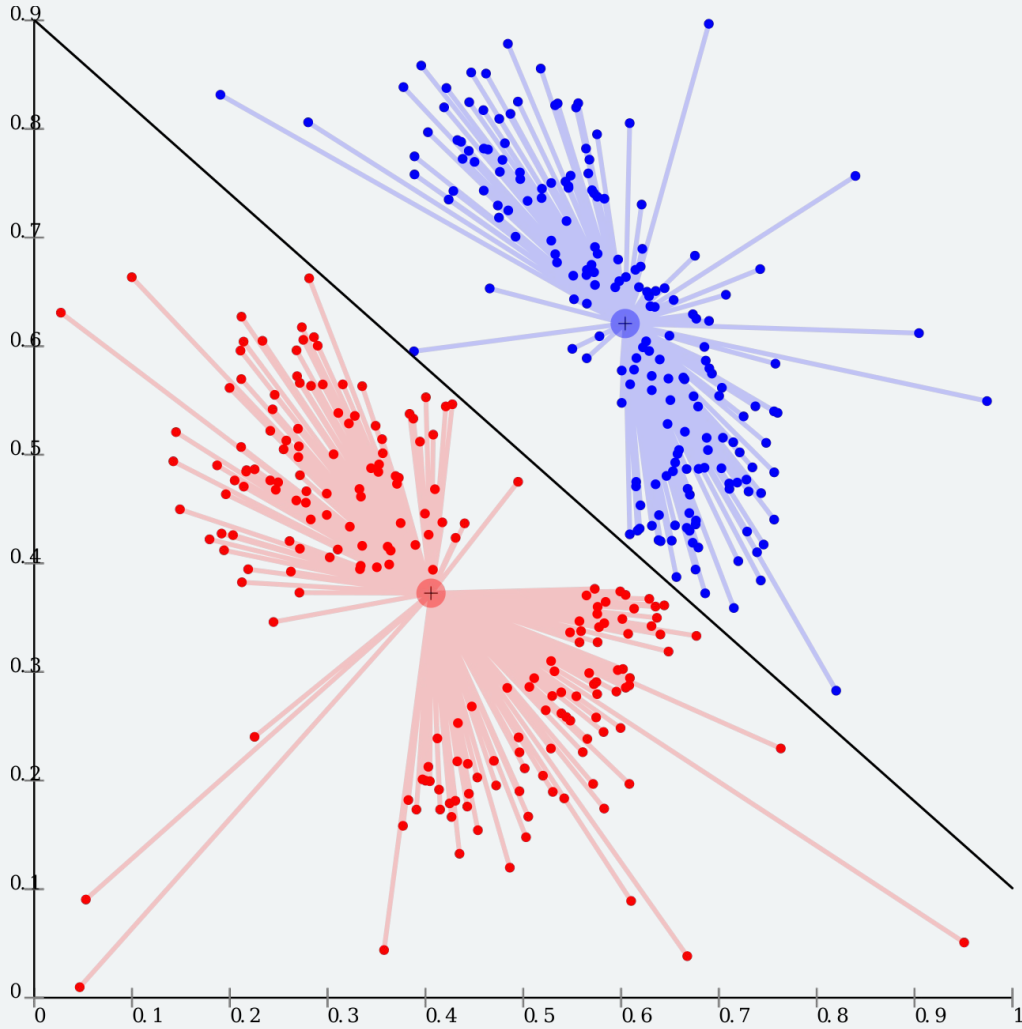
It’s all statistics

Complex words, simple idea



It’s all statistics

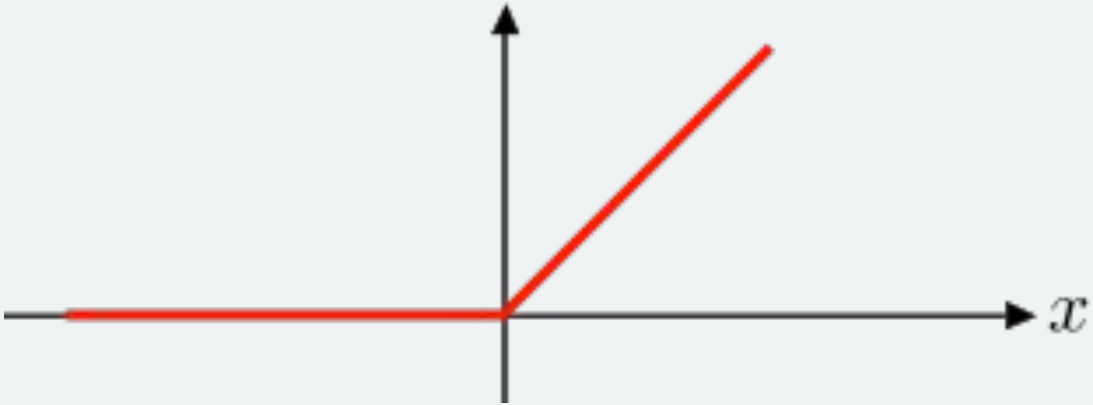
Complex words, simple idea



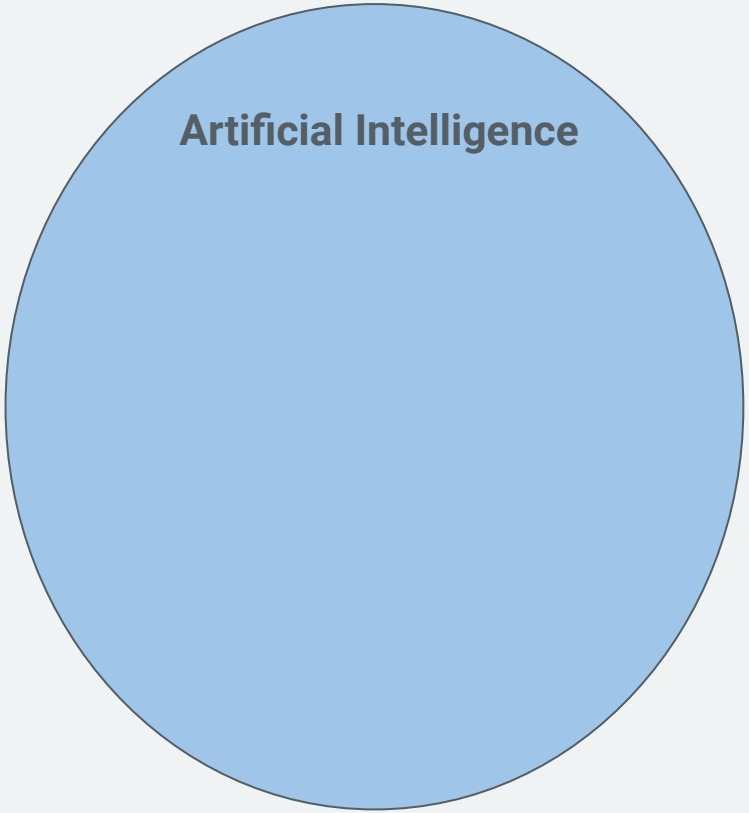
Example: Rectified Linear Unit (ReLU)

The most popular activation function of Neural Networks

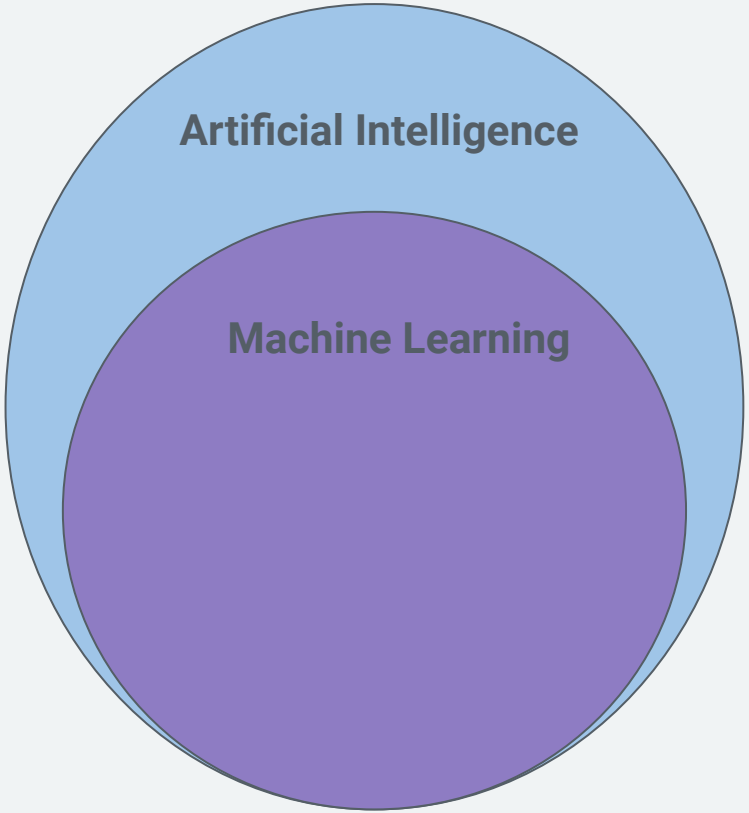
$$\mathrm{ReLU}(x) \triangleq \max(0, x)$$



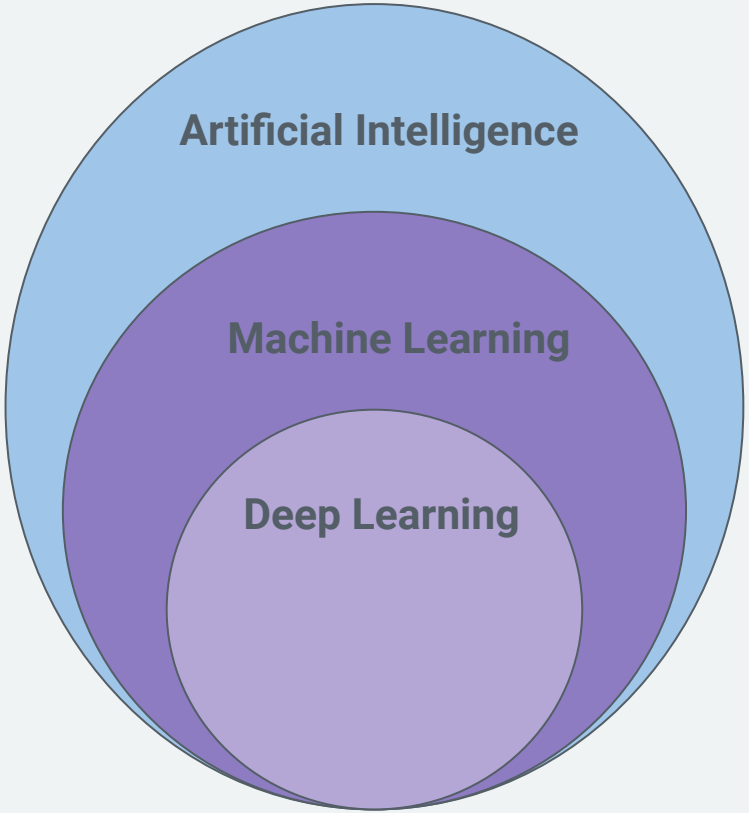
The Relationship



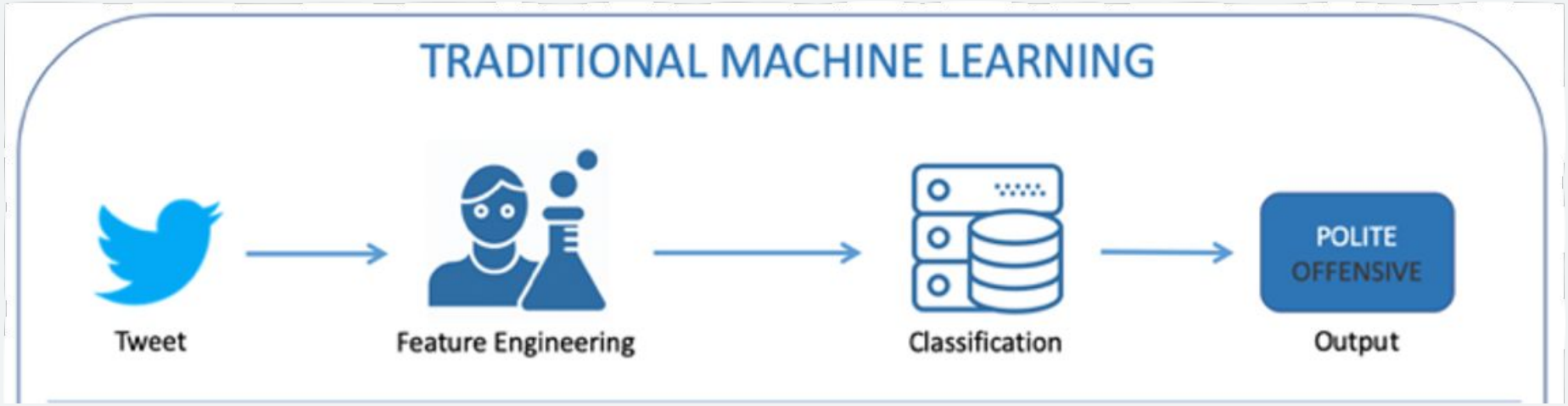
The Relationship

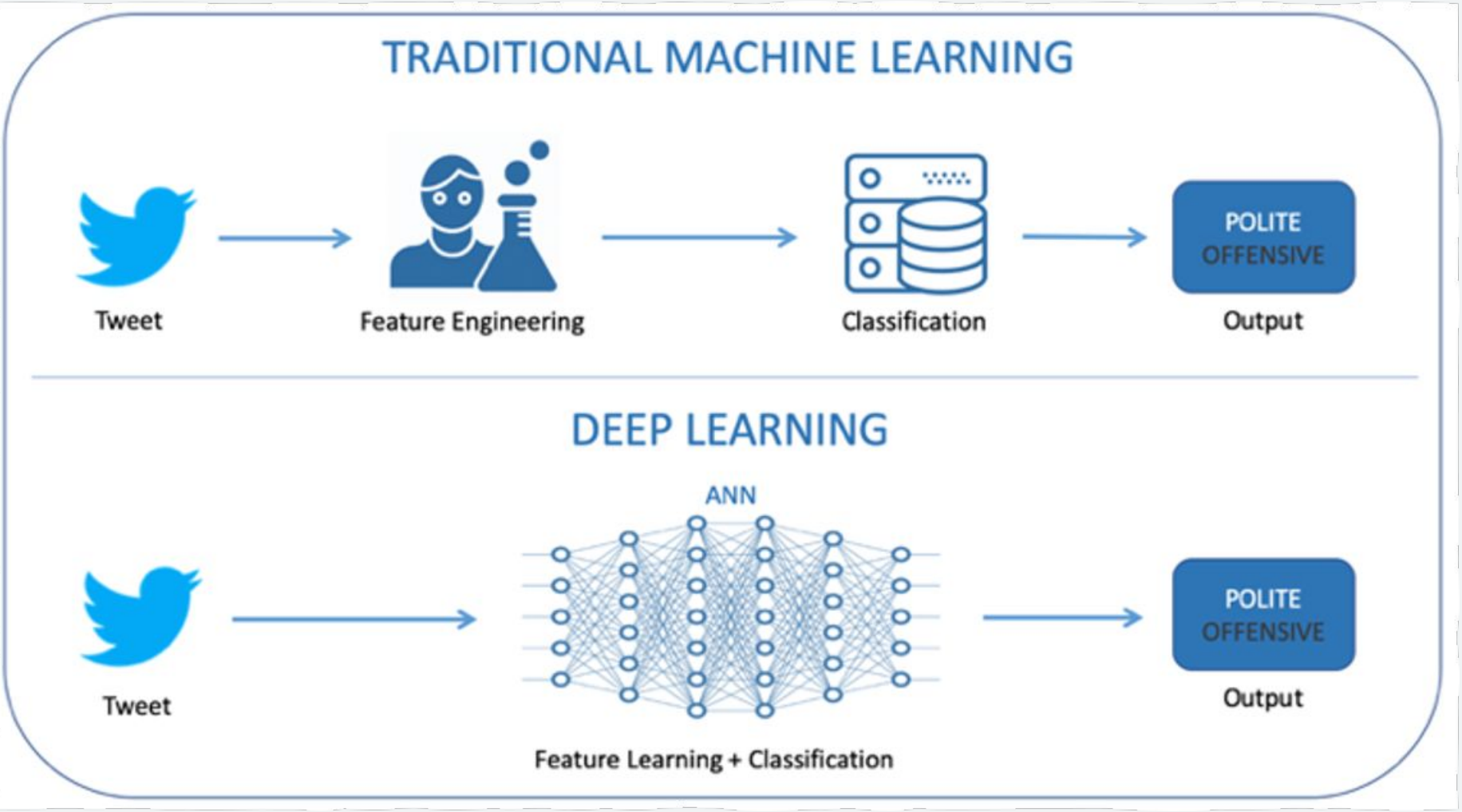


The Relationship



So, what makes what?

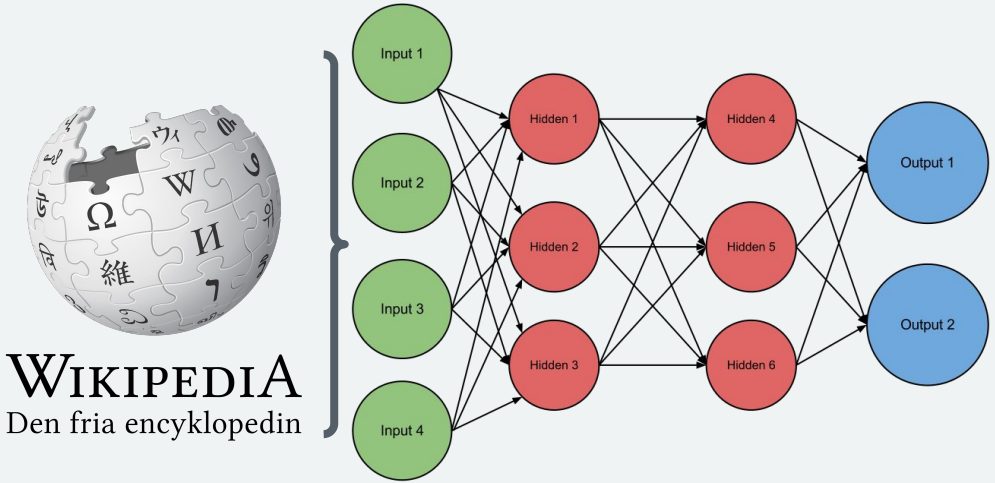

TRADITIONAL MACHINE LEARNING
Tweet
Feature Engineering
Classification
POLITE
OFFENSIVE
Output


TRADITIONAL MACHINE LEARNING
Tweet
Feature Engineering
Classification
POLITE
OFFENSIVE
Output
DEEP LEARNING
ANN
Tweet
Feature Learning + Classification
POLITE
OFFENSIVE
Output

More about Neural Networks

Standard

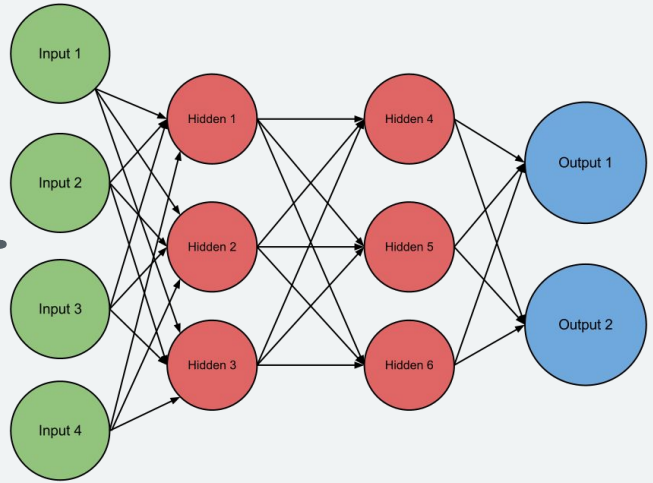
Train from scratch
Random initiation



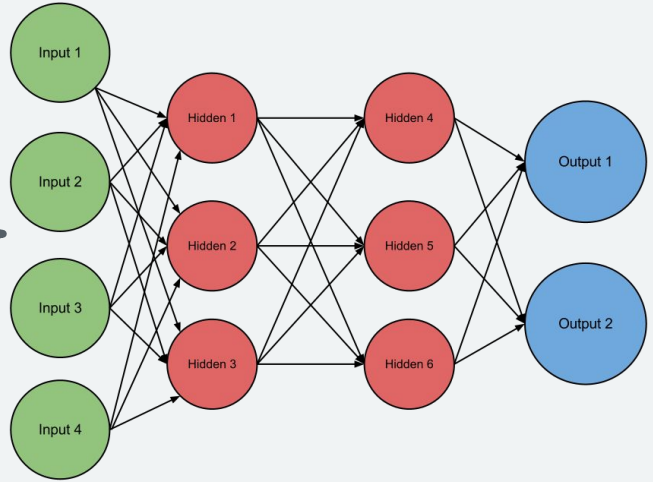
Standard

Train from scratch
Random initiation

WIKIPEDIA
Den fria encyklopedin



amazon



02

TransferLearning

Actually remembering things

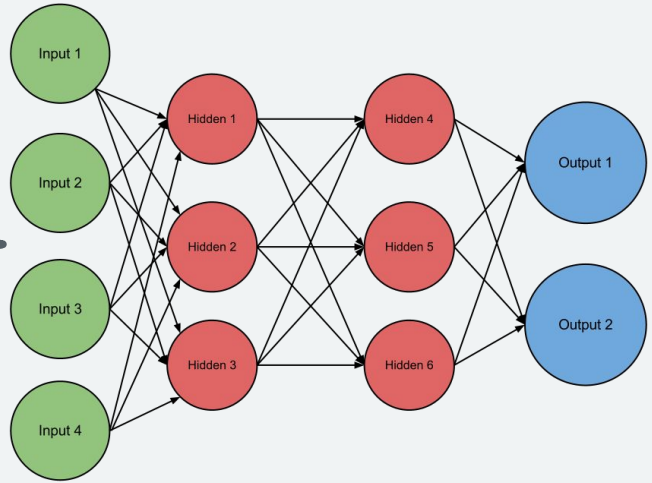
Illustration by Freepik Stories

Reasons to use Transfer Learning

- **Low-resource tasks**
- Improve performance
- Reduce training-time

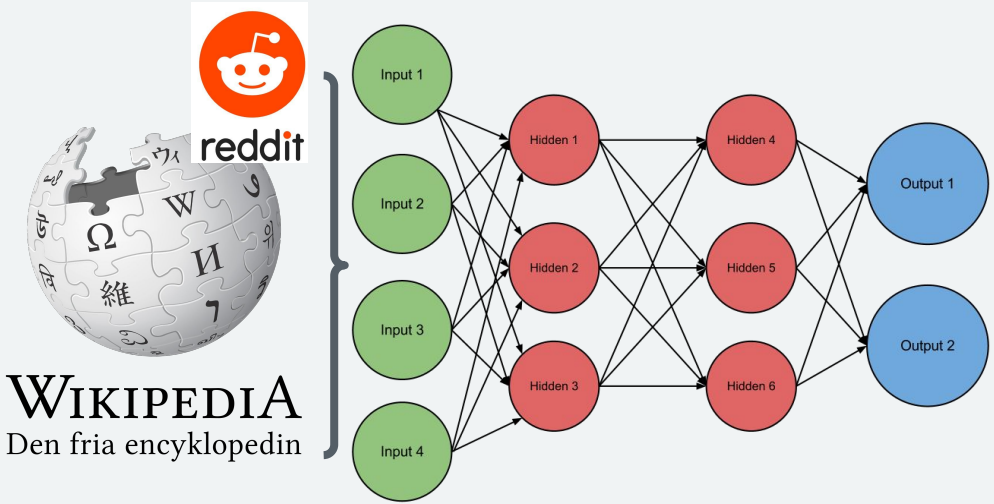
Transfer Learning

Take a pre-trained network
Fine-tune



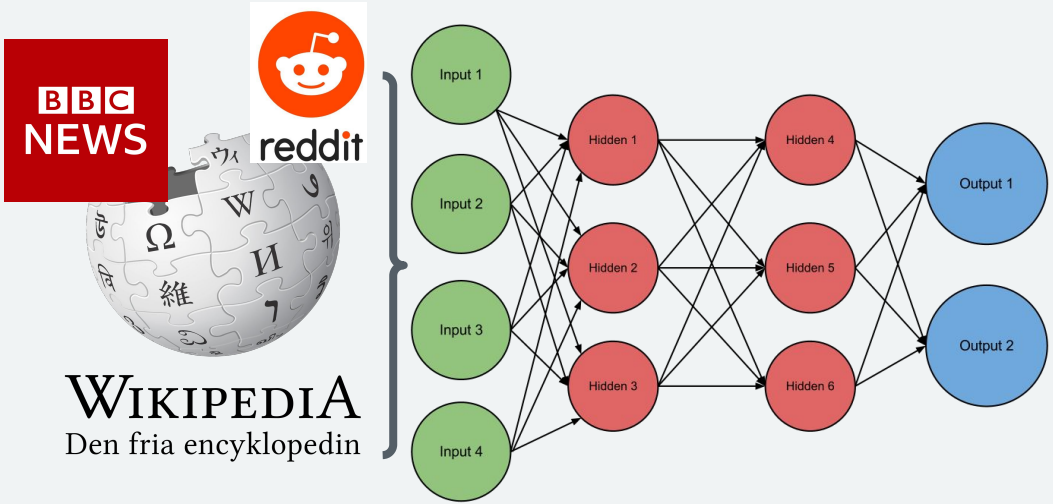
Transfer Learning

Take a pre-trained network
Fine-tune



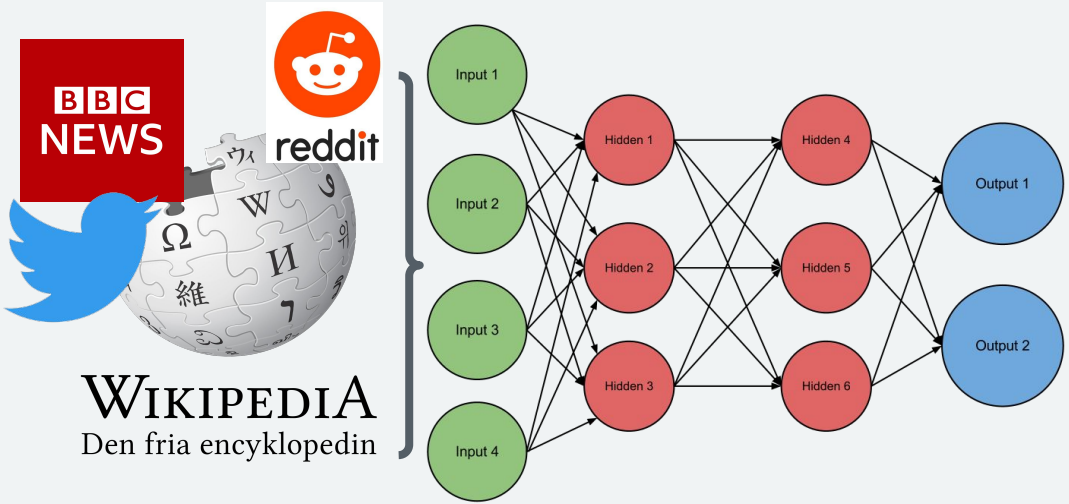
Transfer Learning

Take a pre-trained network
Fine-tune



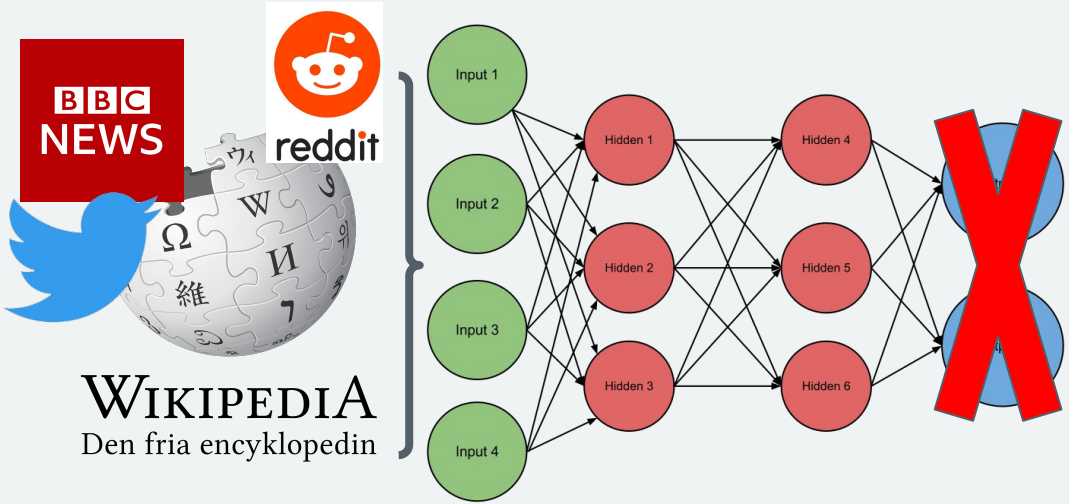
Transfer Learning

Take a pre-trained network
Fine-tune



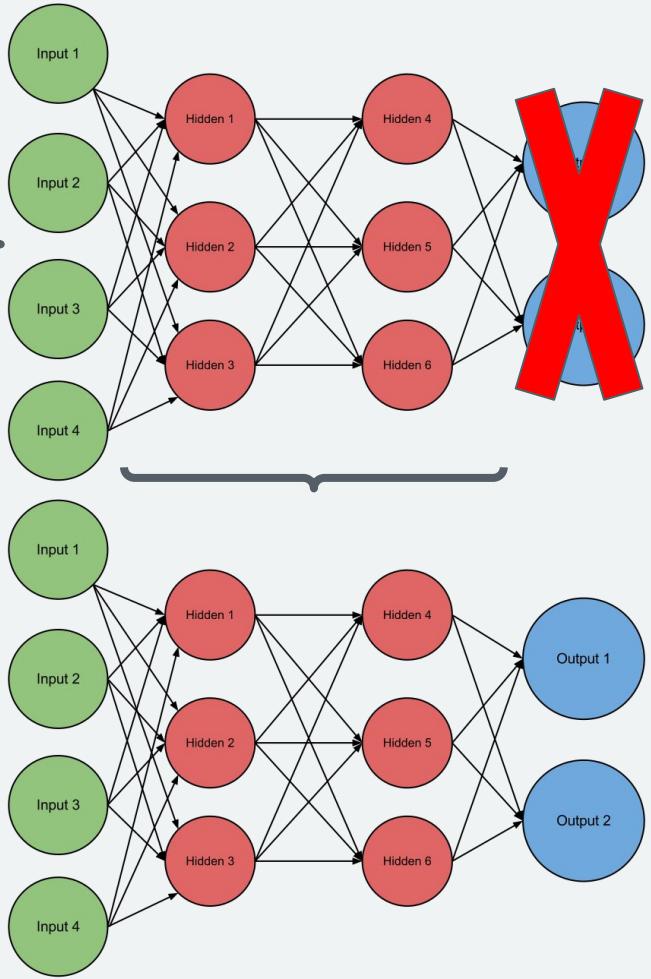
Transfer Learning

Take a pre-trained network
Fine-tune



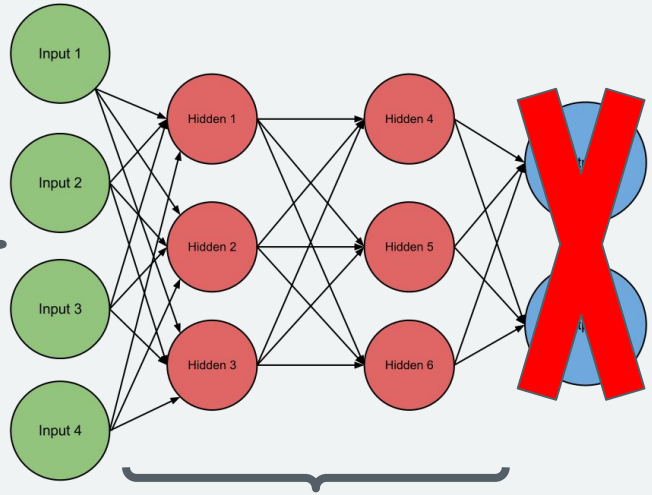
Transfer Learning

Take a pre-trained network
Fine-tune

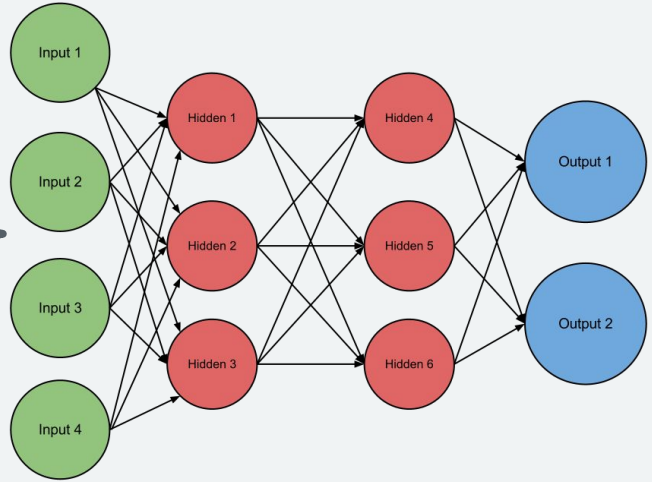


Transfer Learning

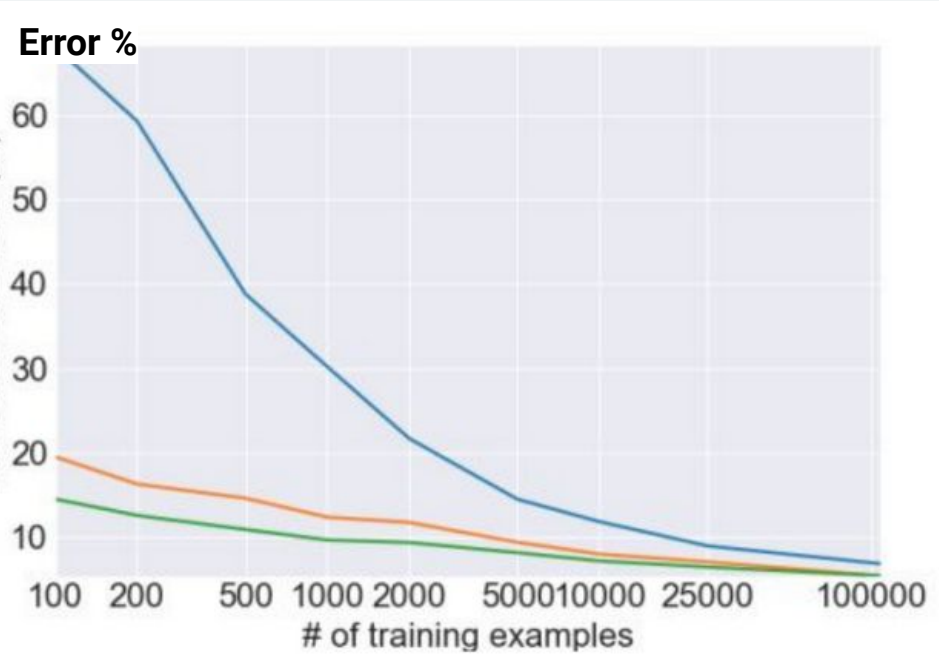
Take a pre-trained network
Fine-tune



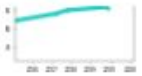
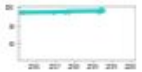
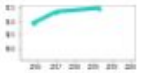
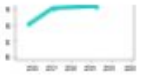
Ystads kommun

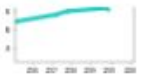
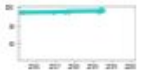
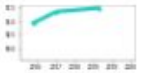
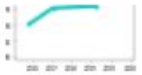


Practical end-result



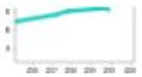
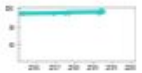
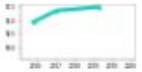
*In practice, transfer learning has often been shown to **achieve similar performance** compared to a non-pretrained model **with 10x fewer examples** (Howard and Ruder, 2018).*

	SST-5 Fine-grained classification	🏆 BERT Large
	IMDb	🏆 NB-weighted-BON + dv-cosine
	Yelp Binary classification	🏆 BERT large
	Yelp Fine-grained classification	🏆 BERT large
	MR	🏆 byte mLSTM7
	Amazon Review Polarity	🏆 BERT large
	Amazon Review Full	🏆 BERT large

	SST-5 Fine-grained classification	🏆 BERT Large
	IMDb	🏆 NB-weighted-BON + dv-cosine
	Yelp Binary classification	🏆 BERT large
	Yelp Fine-grained classification	🏆 BERT large
	MR	🏆 byte mLSTM7
	Amazon Review Polarity	🏆 BERT large
	Amazon Review Full	🏆 BERT large

paperswithcode.com

Transfer Learning

	SST-5 Fine-grained classification	BERT Large
	IMDb	NB-weighted-BON + dv-cosine
	Yelp Binary classification	BERT large
	Yelp Fine-grained classification	BERT large
	MR	byte mLSTM7
	Amazon Review Polarity	BERT large
	Amazon Review Full	BERT large

Transfer Learning

Transfer Learning

paperswithcode.com

Major Themes (NLP)

Major Themes (NLP)

Words to Word-In-Context

Language Models Pre-training

Shallow to Deep

Pretraining vs Target Task

Words to words-in-context

Context matters!

Words to words-in-context

Context matters!

I think this is a **very good movie**

Words to words-in-context

Context matters!

I **don’t** think this is a **very good movie**

Language Model (LM) pretraining

Understand a language
Contextual Representations
Versatile

Language Model (LM) pretraining

Understand a language
Contextual Representations
Versatile

I think this is a very good movie → **Positive**

Language Model (LM) pretraining

Understand a language
Contextual Representations
Versatile

I think this is a very [illegible] movie → **Positive**

Language Model (LM) pretraining

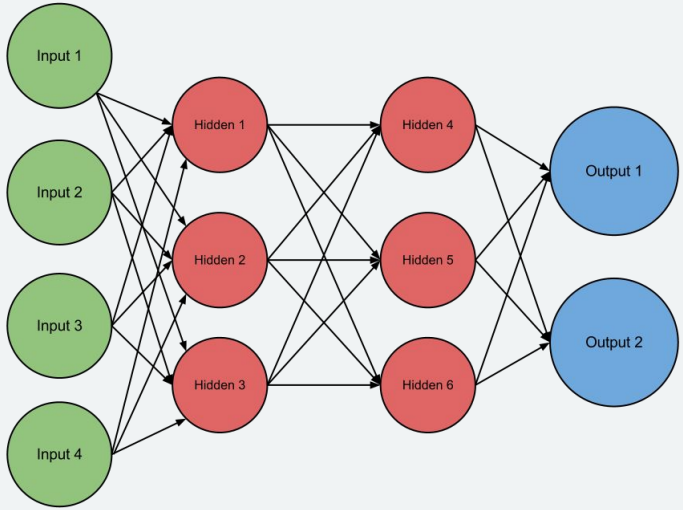
Understand a language
Contextual Representations
Versatile

I think this is a very **[MASK]** movie

Shallow to Deep

Quickly models have grown deeper.

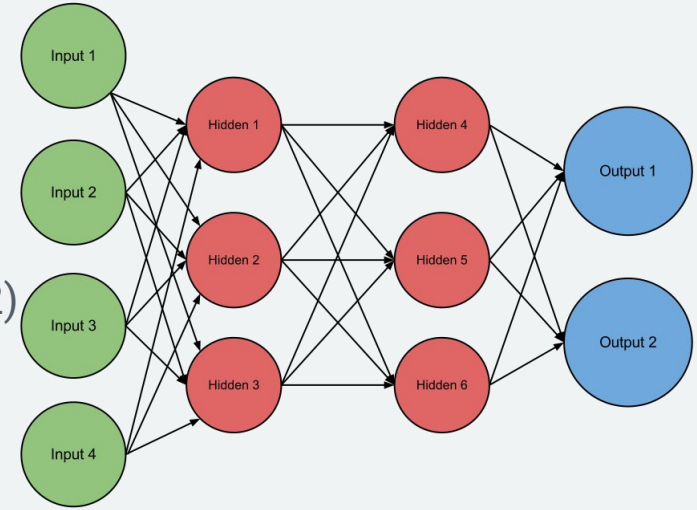
2016: 2-3 layers



Shallow to Deep

Quickly models have grown deeper.

2016: 2-3 layers
2019: ~60 layers (24 Transformers Blocks, BERT/GPT2)



Pretraining vs Target Task

Pre-training similar to target task

Sentence Representation is not useful for word-level predictions, and v.v.

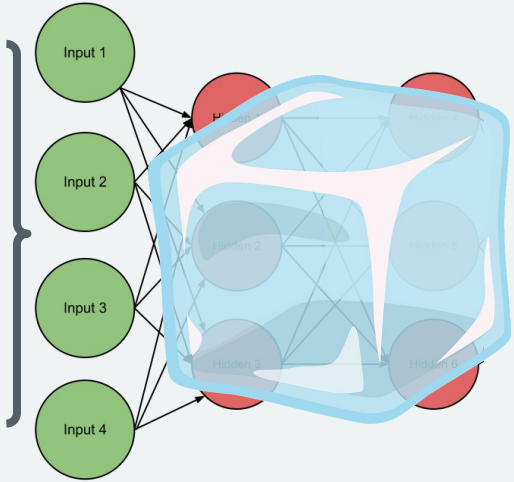
How do we use “Transfer Learning”?

Feature Extraction

Fine-Tuning

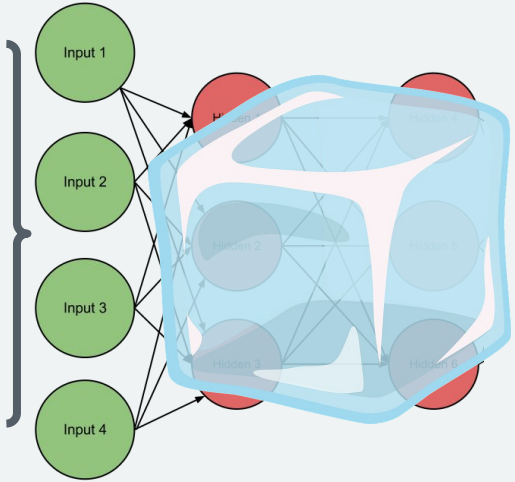
Feature Extraction

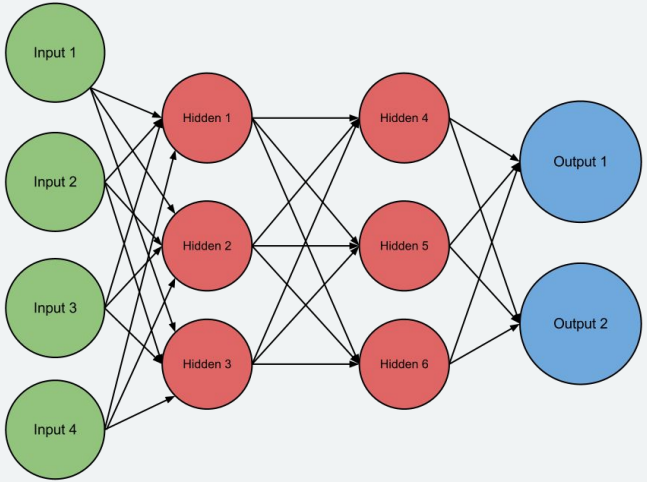
Ystads kommun



Feature Extraction

Ystads kommun





How to fine-tune?

Freezing

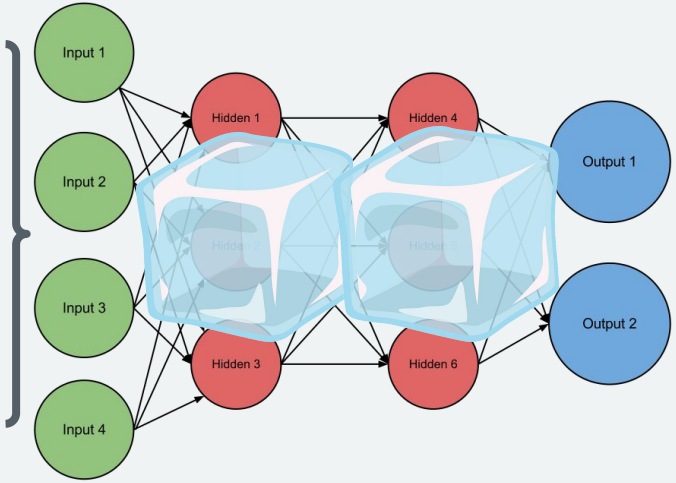
Lower Learning Rate

Regularization

Freezing

Fine-tune progressively in time

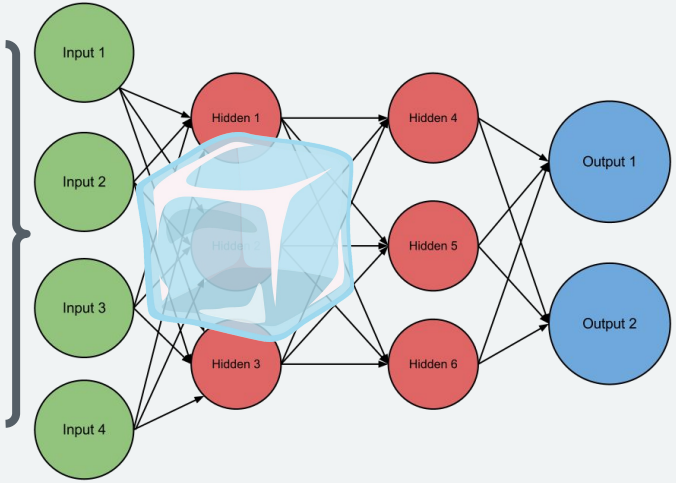
Ystads kommun



Freezing

Fine-tune progressively in time

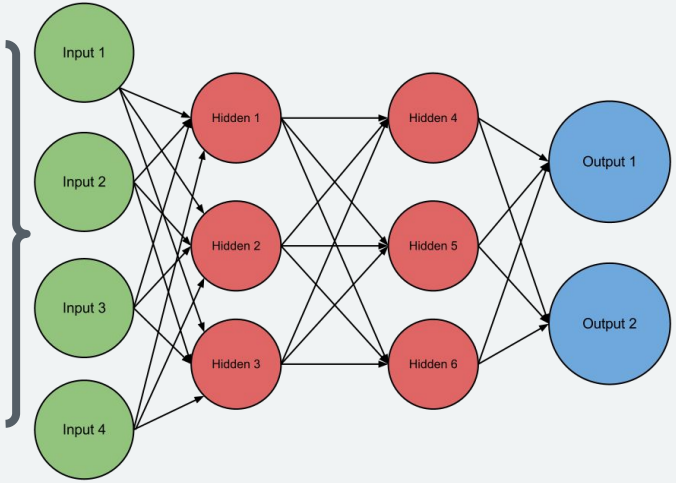
Ystads kommun



Freezing

Fine-tune progressively in time

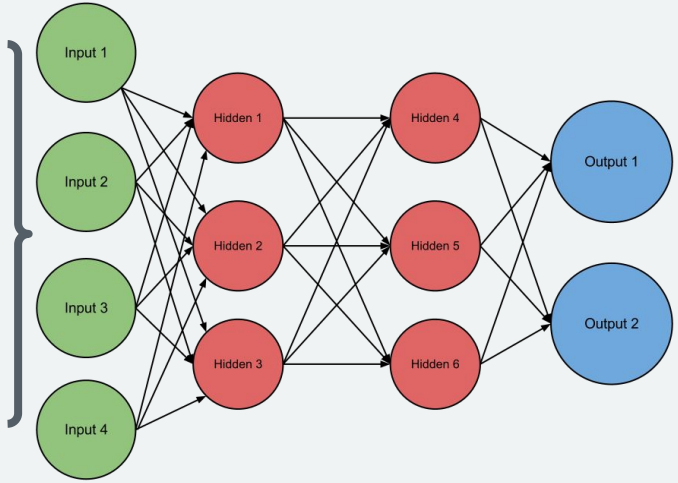
Ystads kommun



Lower Learning Rate

Fine-tune: progressively in intensity

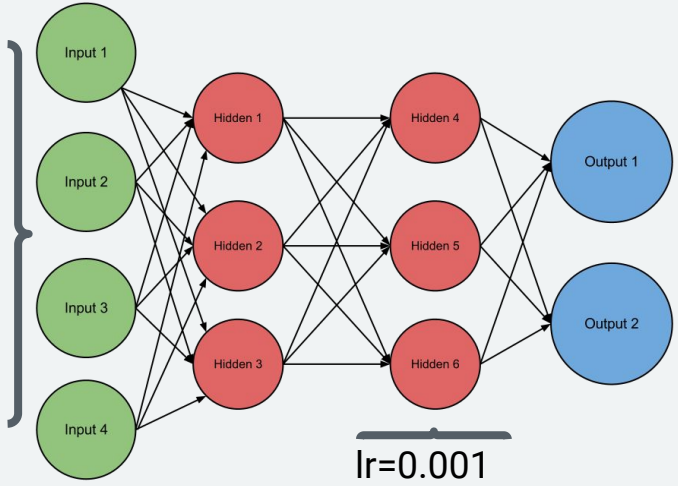
Ystads kommun



Lower Learning Rate

Fine-tune: progressively in intensity

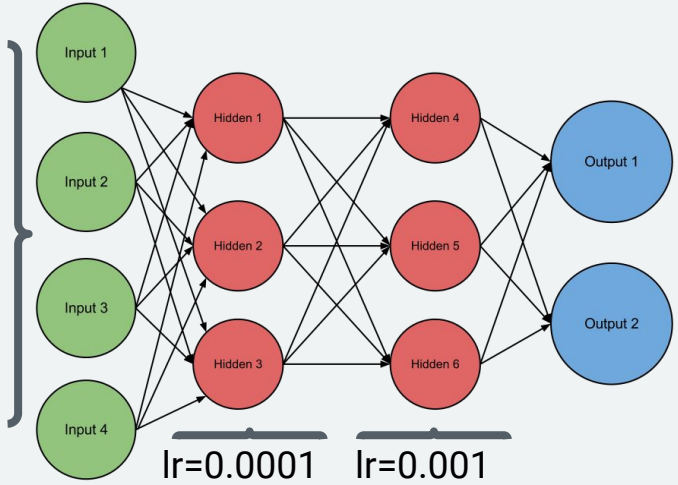
Ystads kommun



Lower Learning Rate

Fine-tune: progressively in intensity

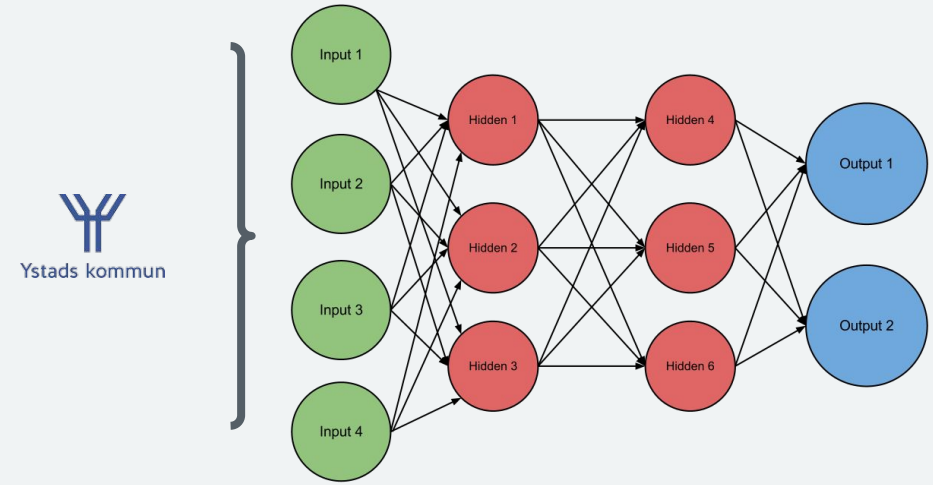
Ystads kommun



Regularization

Fine-tune: progressively vs. a pretrained model

Ystads kommun



Trade-offs: Working with pre-trained models

Feature Extraction

- Slower
- Space-efficient

Fine-tuning

- Better for dissimilar tasks

Getting more signal

Sequential adaptation

A related task with more data?

1. Fine-tune on related task
2. Fine-tune on target task

Improves if limited data (Phang et al., 2018)
Improves sample efficiency on target task (Yogatama et al., 2019)

Multi-task fine-tuning

Fine-tune jointly on related task
Language Model is a good choice (even w/o pre-train)

Led to improvement in multiple target tasks (Liu et al., 2019, Wang et al., 2019)

Semi-supervised learning

- More consistent predictions by perturbing unlabelled examples
  - E.g. noise, masking or data-augmentation

Ensembling

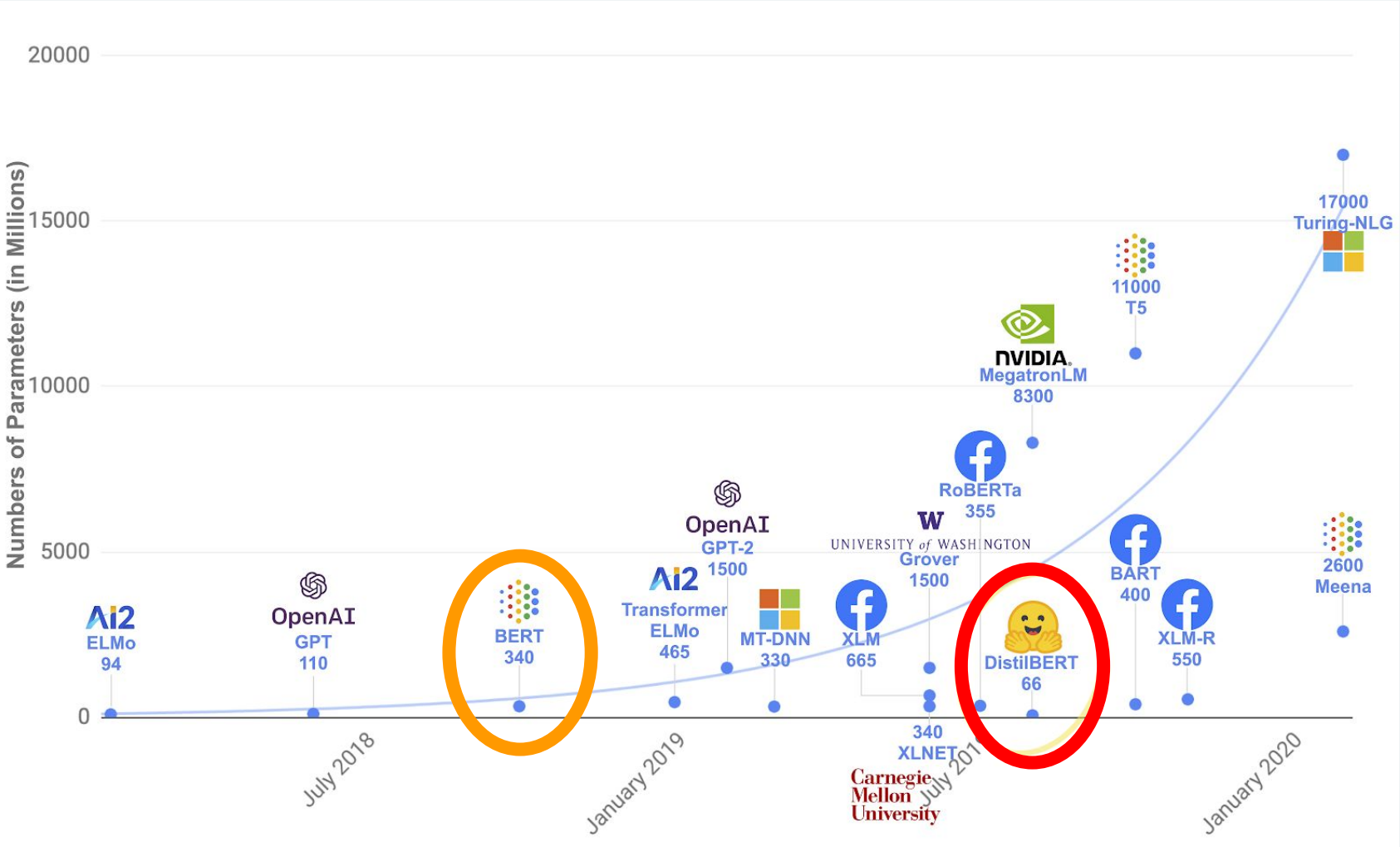
An ensemble of models

- Different hyper-params
- Different pre-trained models
- Different target-tasks

Distilling

- Large model distilled into a small model
- Distilled model is a lot simpler

(Tang et al., 2019)


20000
15000
10000
5000
0
Numbers of Parameters (in Millions)
July 2018
January 2019
July 2019
January 2020
AI2
ELMo
94
OpenAI
GPT
110
BERT
340
AI2
Transformer
ELMo
465
OpenAI
GPT-2
1500
MT-DNN
330
XLM
665
UNIVERSITY of WASHINGTON
Grover
1500
340
XLNET
Carnegie Mellon University
RoBERTa
355
NVIDIA
MegatronLM
8300
DistilBERT
66
11000
T5
BART
400
XLM-R
550
17000
Turing-NLG
2600
Meena

Extremes

	Date of original paper	Energy consumption (kWh)	Carbon footprint (lbs of CO2e)	Cloud compute cost (USD)
Transformer (65M parameters)	Jun, 2017	27	26	$41-$140
Transformer (213M parameters)	Jun, 2017	201	192	$289-$981
ELMo	Feb, 2018	275	262	$433-$1,472
BERT (110M parameters)	Oct, 2018	1,507	1,438	$3,751-$12,571
Transformer (213M parameters) w/ neural architecture search	Jan, 2019	656,347	626,155	$942,973-$3,201,722
GPT-2	Feb, 2019	-	-	$12,902-$43,008

GPT3 - Estimated 4,600,000 USD

Downstream Applications

- Hubs (“blackbox”)
- Checkpoints
- Third party library

Applications for us

- Routing issues for municipalities or companies
- Data Understanding
  - Sentiment
  - Keywords
  - Tagging
  - Severity Ranking
- & much more

Final Words

There exists no “magic wand”
Be knowledgeable
Never forget baselines

Thanks!

Do you have any questions?

hampus.londogard@afry.com
+46 733 673 179

Live Example

Transformer on Swedish: Colaboratory (google.com)