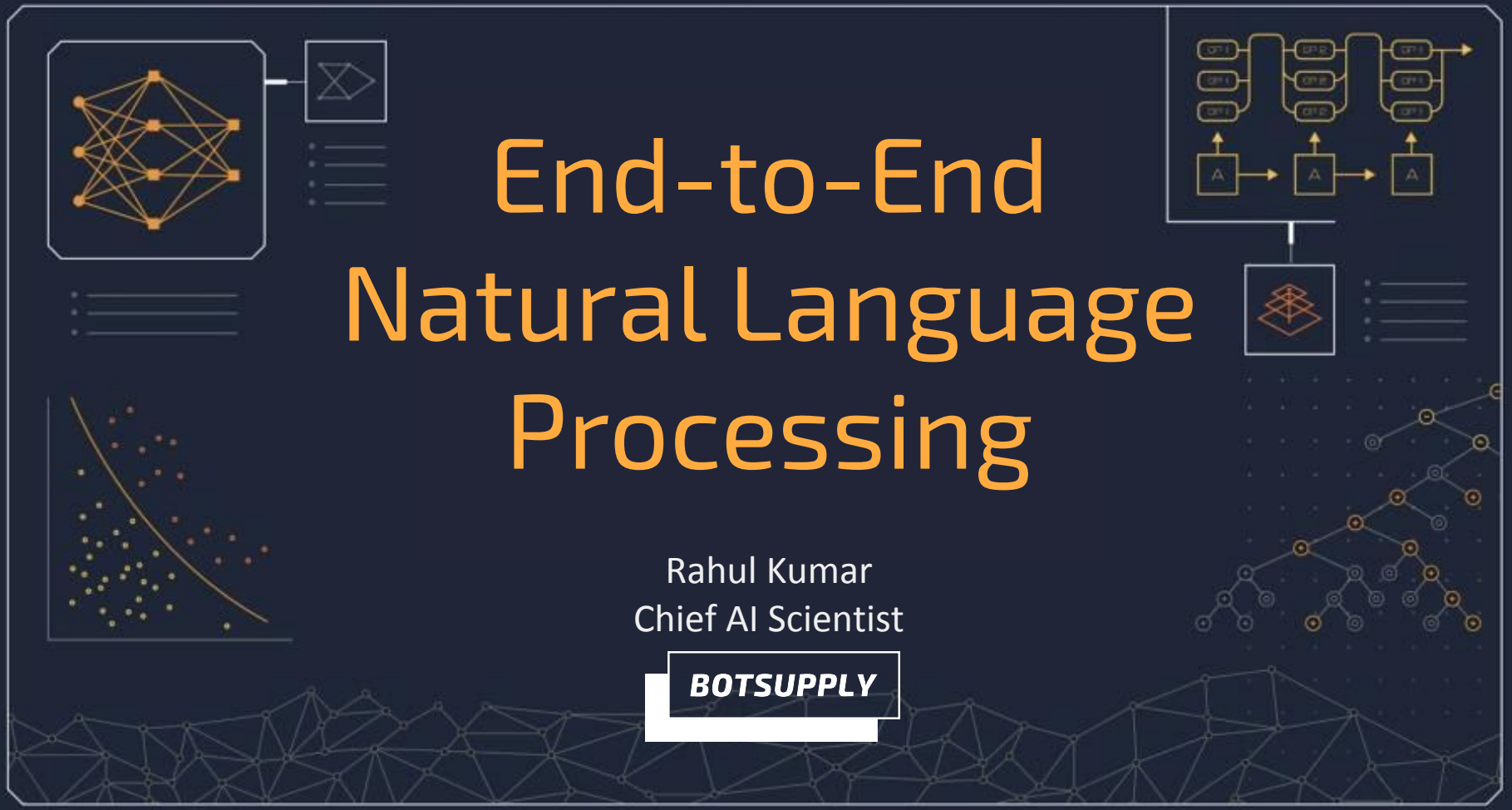


End-to-End Natural Language Processing

Rahul Kumar
Chief AI Scientist

BOTSUPPLY

Natural Language Processing: The Goal

[**Context**: We saw some lions; guide tells us it's a rare sight. Few min later we pass a second safari guide.]

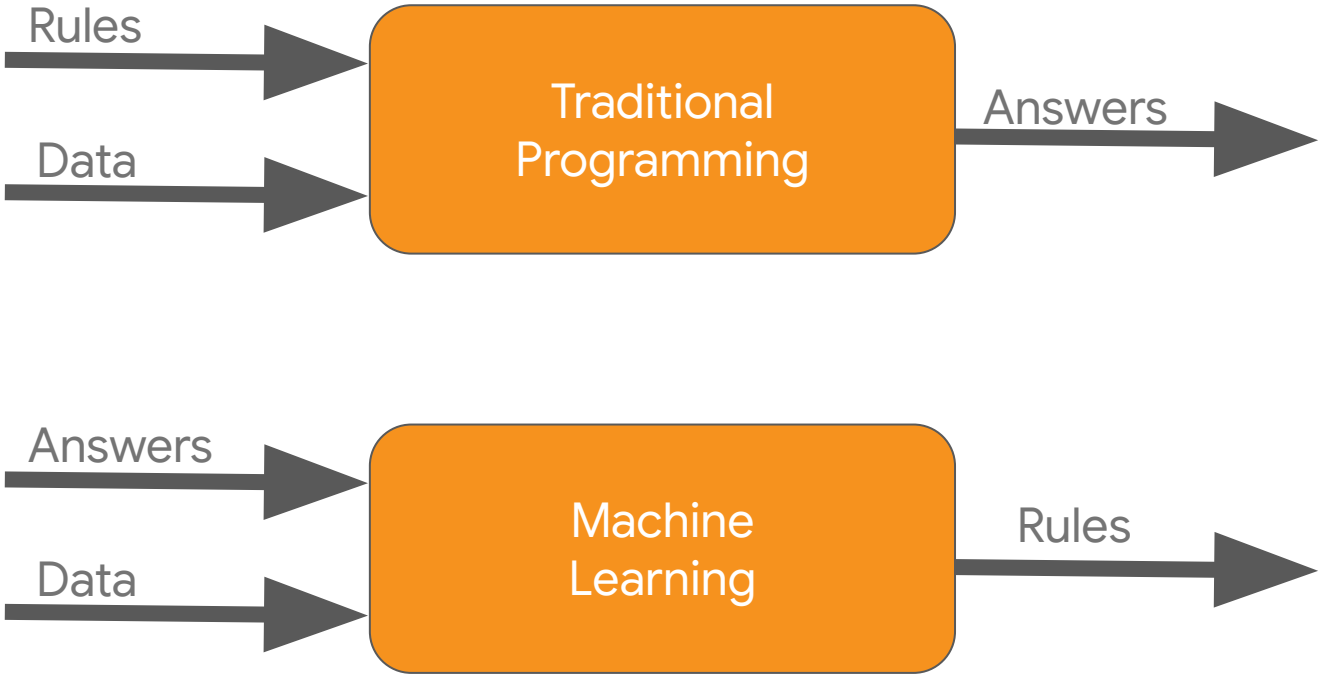
Our guide to other guide: Vi så nogle løver

[Second **driver** speeds off]

Simon and I: I bet he told him about the lion.

Our guide: Yep!

Programming Paradigm



Before ~2013

2015 onwards

Traditional Programming

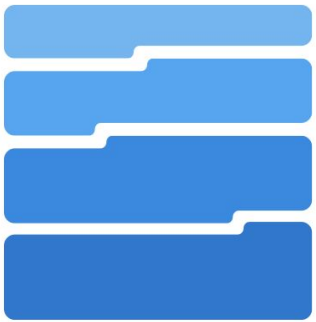
Machine Learning

Bucketing

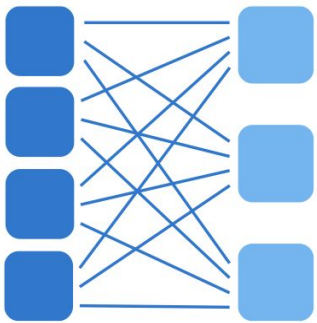
Crossing

Hashing

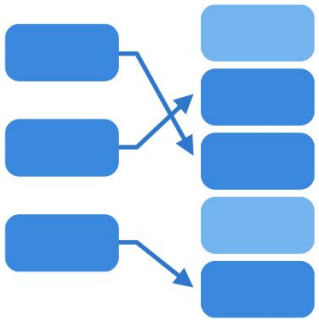
Embedding



Partition by range



Create new combinations



Limit size

Learn a new representation

Before ~2013

2015 onwards

Traditional Programming

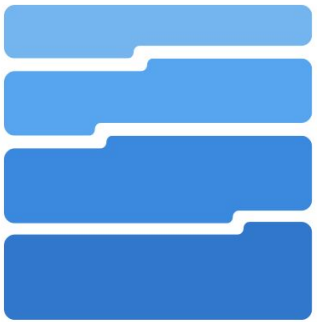
Machine Learning

Bucketing

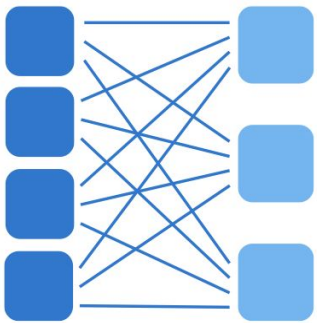
Crossing

Hashing

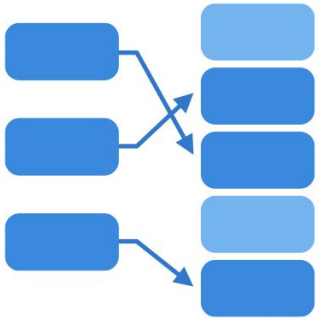
Embedding



Partition by range

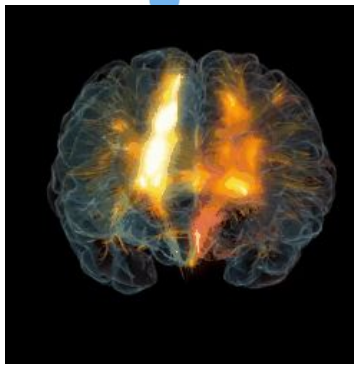


Create new combinations



Limit size

MEMORY



Learn a new representation

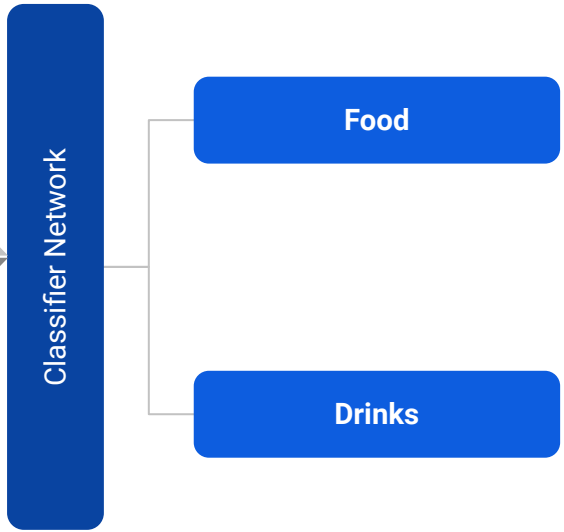
(Modern) Natural Language Processing: 2019

Training Dataset

	Training Data	Label
1	Electricity is food for Robots.	Food
2	Feed me Ram.	Food
3	On good power supply, I feel tipsy.	Drinks
4	I am a bot with Beer Belly.	Drinks

Pre Trained Language Model

Prediction

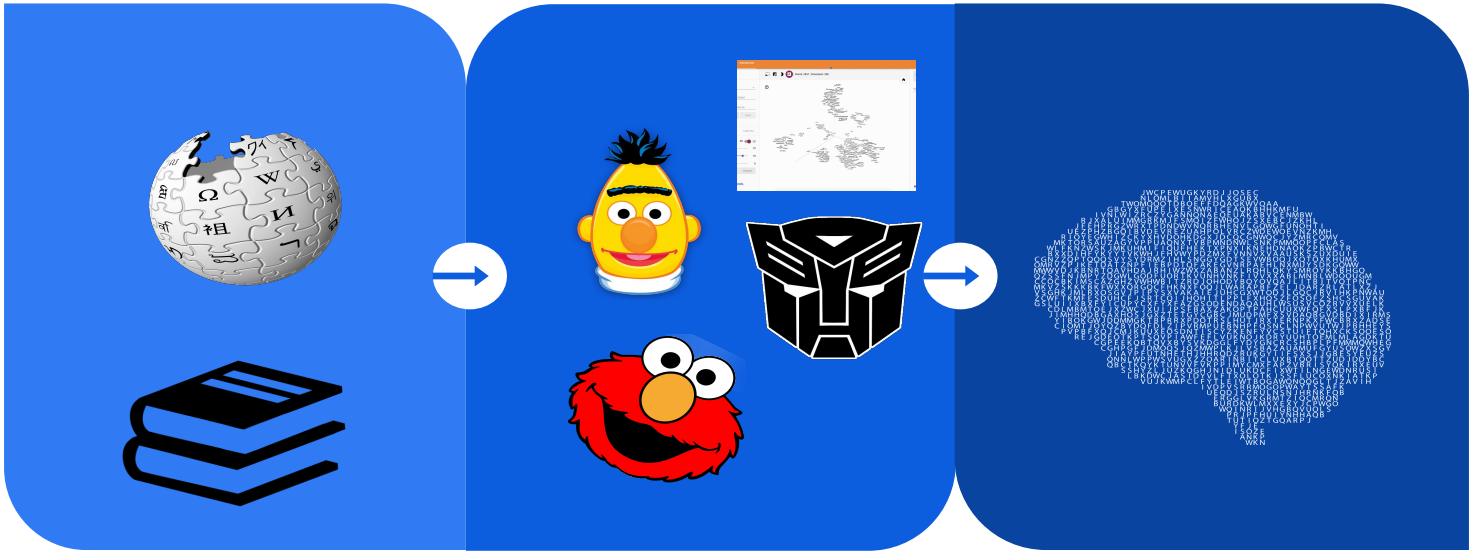


(Modern) Natural Language Processing: 2019

Dataset

Models

Output

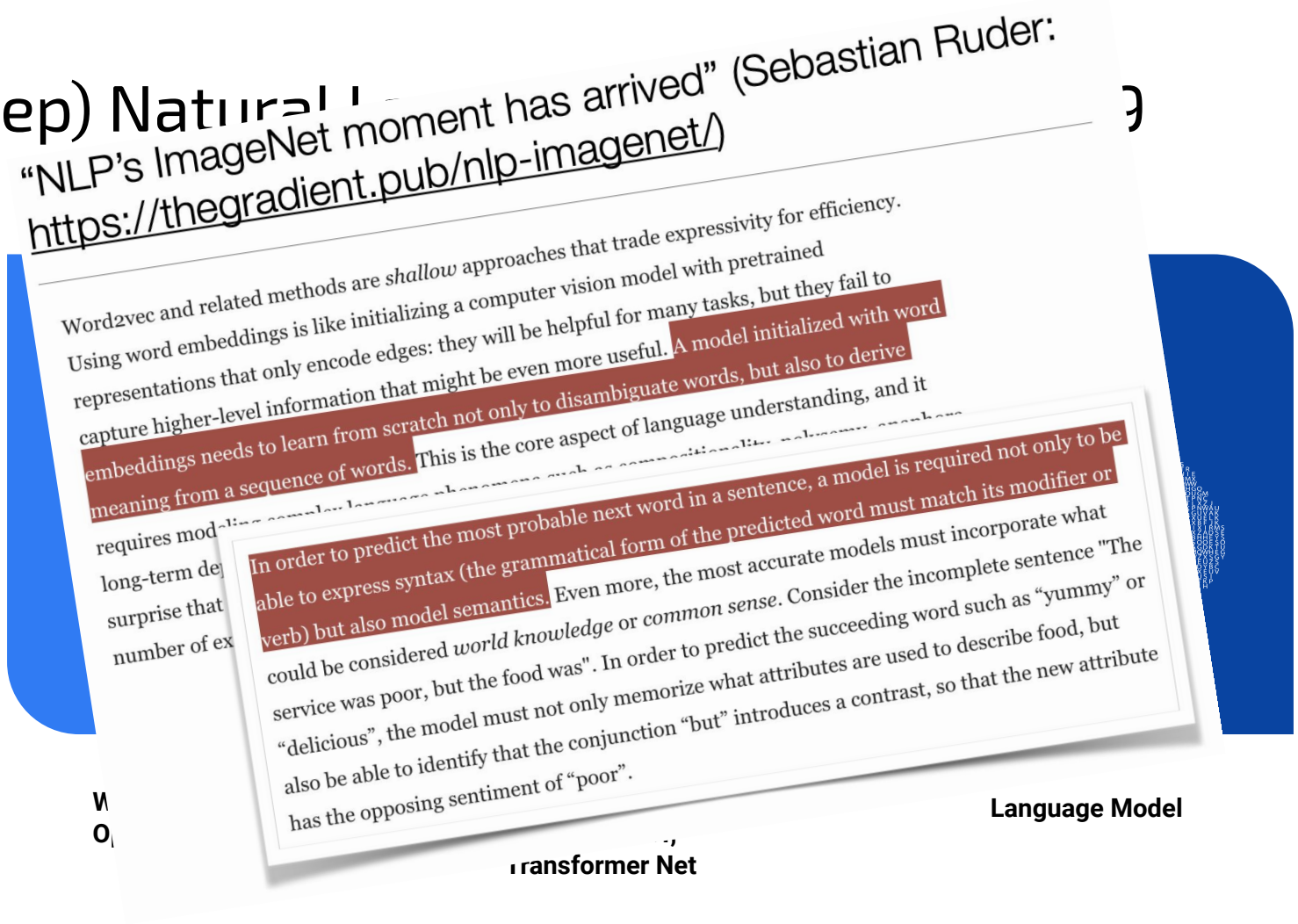


Wikidata, Books,
Open Source Dataset

word2vec, BERT,
ELMo, ULMFiT,
Transformer Net

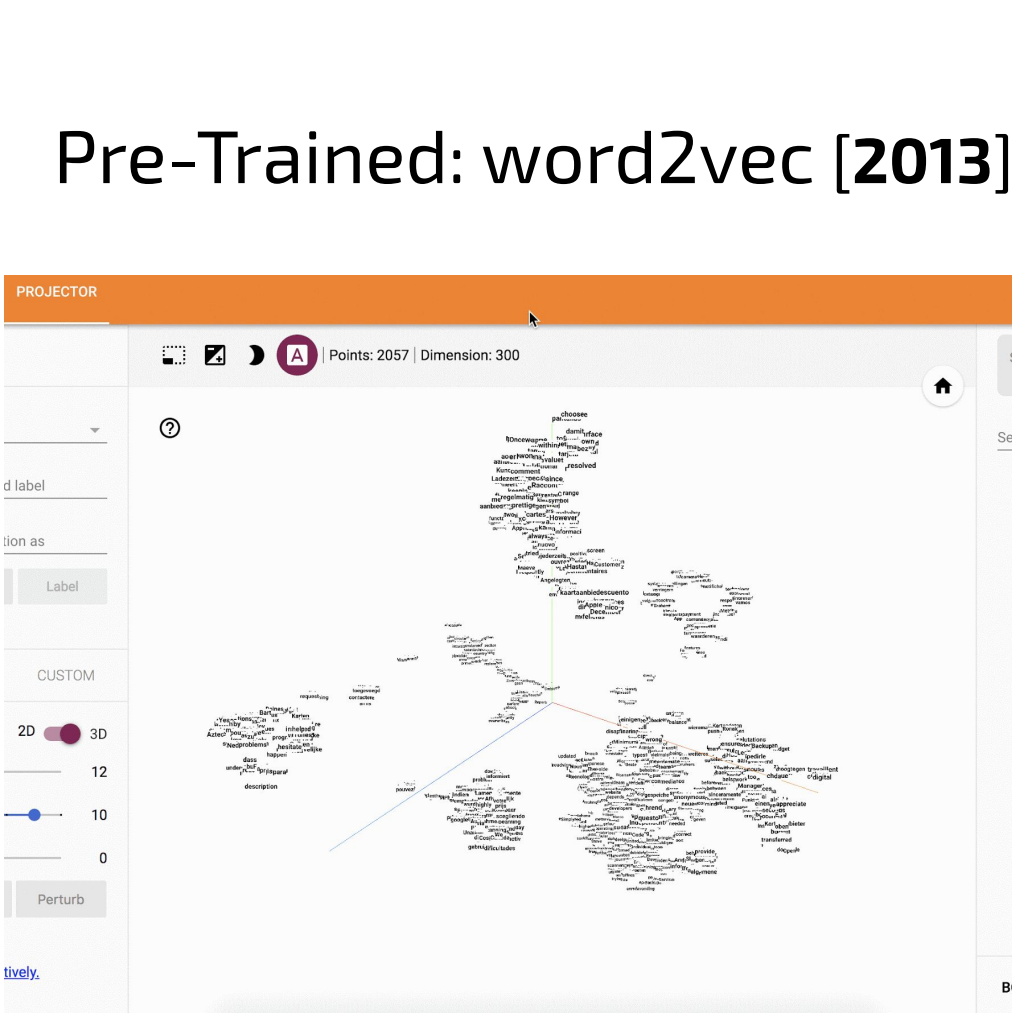
Language Model

(Deep) Natural



Journey of Modern NLP

Pre-Trained: word2vec [**2013**]



Key takeaways:

1. Skip-gram neural language model.

2. Mapped words to vectors (embeddings) which were part of the trained weights of the net.

3. Showed that learned vector spaces organized to preserve linear relationships between concepts, eg

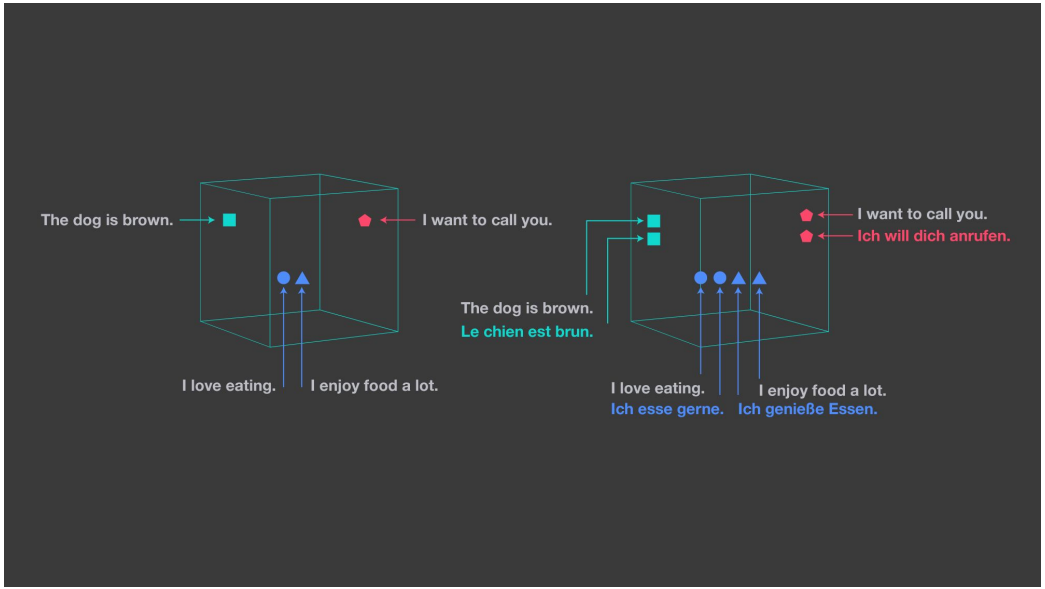
v(“king”) - v(“man”) + v(“woman”) = v(“queen”).

4. In the process of training, similar words ended up near each other (similar vectors). HUGE!

Pre-Trained: word2vec [**2014**] (Cross-lingual embeddings)

[PDF] Exploiting Similarities among Languages for Machine Translation
https://arxiv.org/pdf/1309.4168
by T Mikolov - 2013 - Cited by 631 - Related articles

BilBOWA: Fast Bilingual Distributed Representations without Wor…
https://arxiv.org › stat
by S Gouws - 2014 - Cited by 223 - Related articles
Abstract: We introduce **BilBOWA** (Bilingual Bag-of-Words without Alignments), a simple and computationally-efficient model for learning bilingual distributed ...



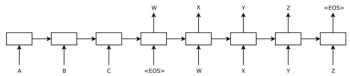
Key takeaways:

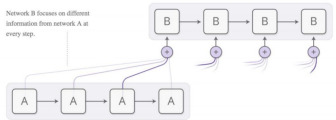
1. Noticed that word vectors in one language can be projected to another using learned affine mapping.

2. Suggested a new way for doing word-by-word translation: nearest neighbours in the aligned spaces!

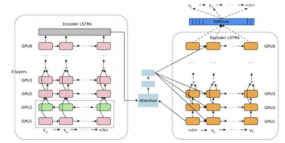
3. Also suggested easy method for **transfer learning**: train on EN embedding space, and apply classifier directly to aligned FR space. Voila!

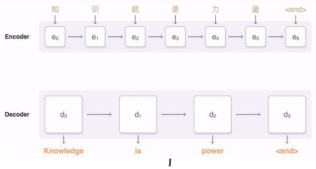
Fastforward



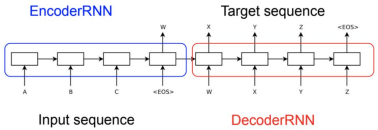


The skip connections allows information to flow unmitigated through the network (related to LSTM error carousels, but not gated).

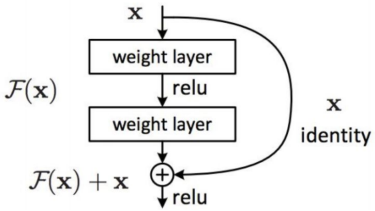




seq2seq [**2014**] ------> seq2seq + Attention[**2015**] ------> ResNets[**2015**] ------> Google NMT[**2016**] ----->



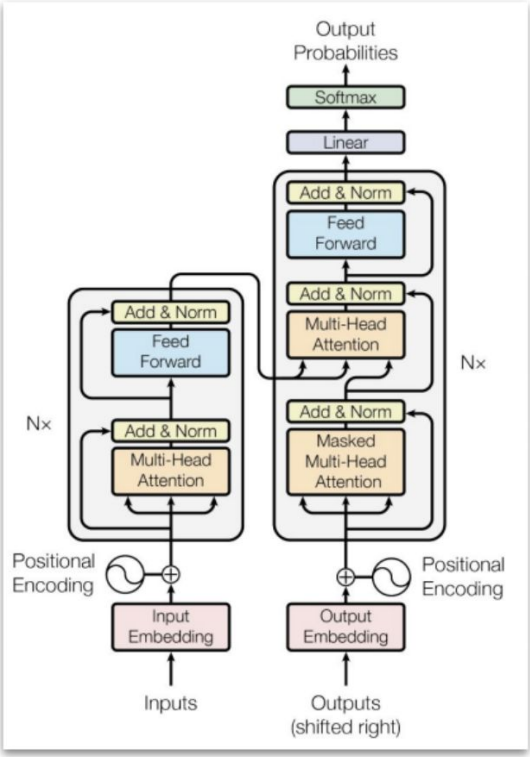
This was huge. Soon, LSTMs+Attention became the default for almost any task that can be cast as seq2seq.



Deep LSTM encoder-decoder with residual connections and attention. Boosted translation from 58% to 87%.

Pre-Trained: Transformer[**2017**]

Attention Is All You Need
https://arxiv.org › cs ▾
by A Vaswani - 2017 - Cited by 2925 - Related articles
Jun 12, 2017 - The best performing models also connect the encoder and decoder through an **attention** mechanism. We propose a new simple network ...



Key takeaways:

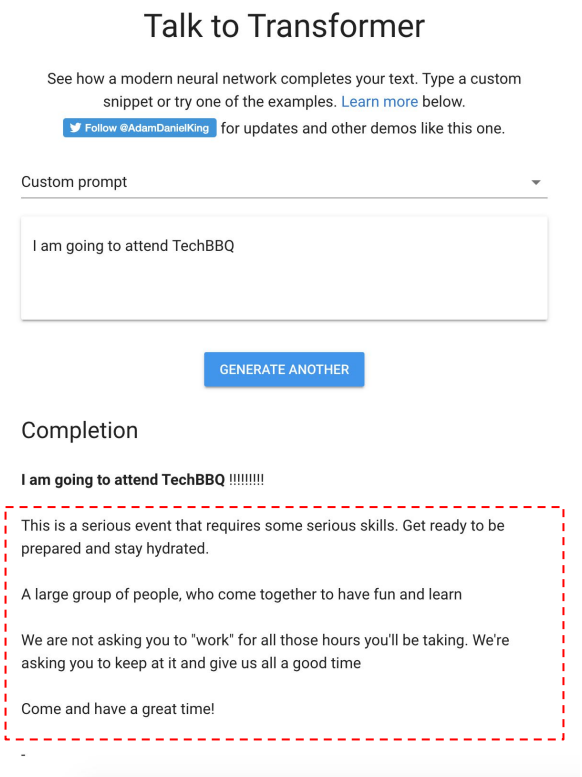
1. Parallel-in-time encoder-decoder model with blocks of self-attention interleaved with feedforward blocks.

2. Introduced Multi-Head Self-Attention which allows multimodal attention.

3. Fast to train, as all encoder blocks for all sequence positions are performed in parallel (as opposed to sequentially for RNNs).

4. Started huge new wave of “self-attentive” models.

Pre-Trained: GPT-2 [**2018**]



Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Key takeaways:

1. Pretrain large 12-layer left-to-right Transformer.

2. Fine tune for sentence, sentence-pair and multiple choice questions.

3. SOTA results for 9 tasks.

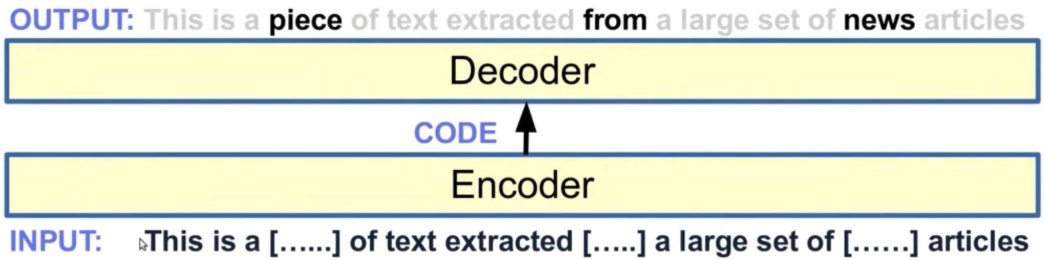
DEMO LINK:
https://talktotransformer.com/
Built by Adam King (@AdamDanielKing)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova

(Submitted on 11 Oct 2018 (v1), last revised 24 May 2019 (this version, v2))

Pre-Trained: BERT [**2019**]



Key takeaways:

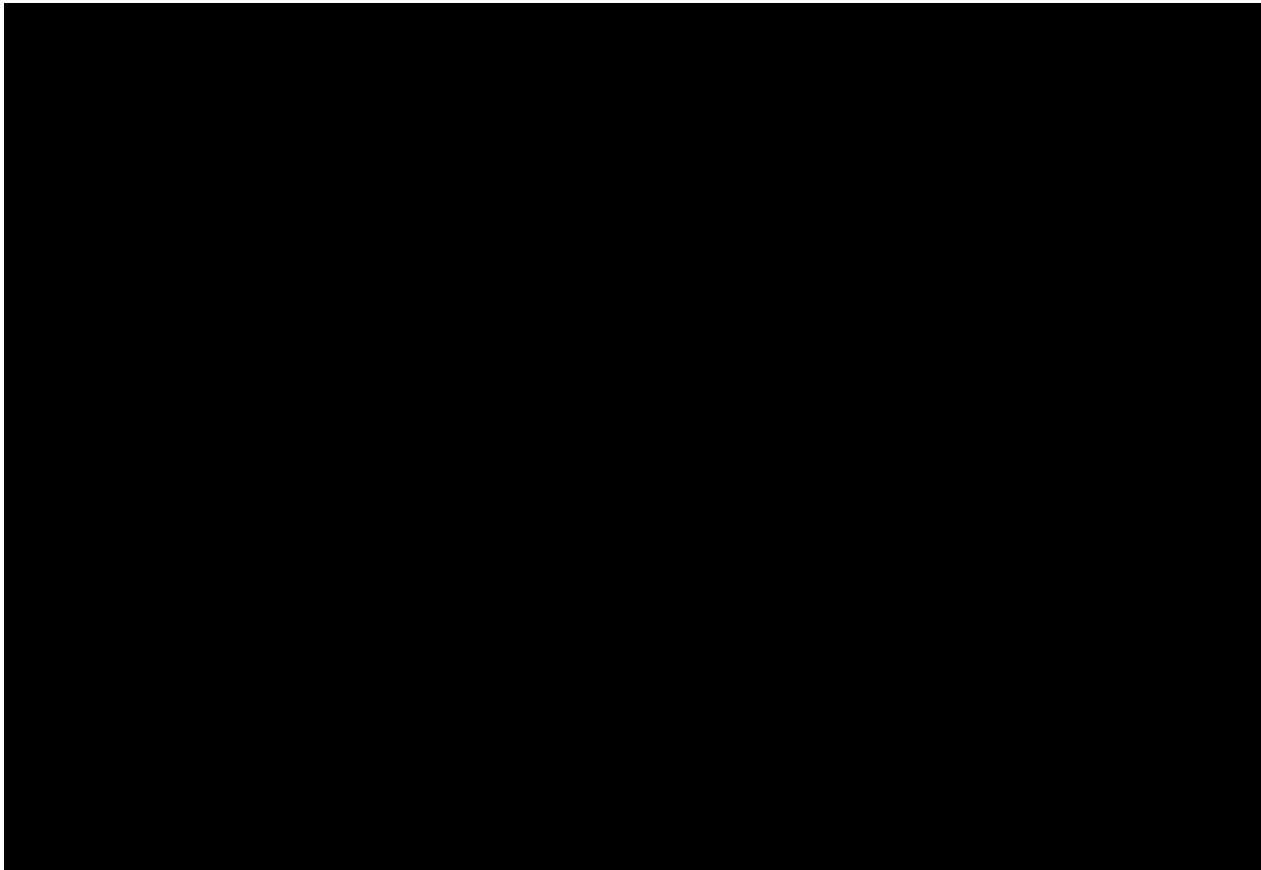
1. BERT pretrains both sentence and contextual word representations, using masked LM and next sentence prediction.

2. LM which can looks both forward and backwards (**learning both left and right context**)

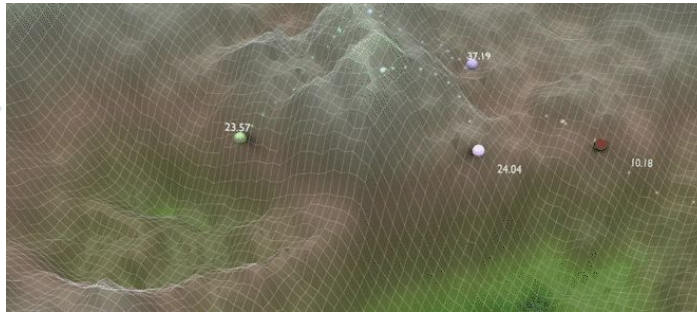
3. BERT-large has 340M parameters, 24 layers!

Semi-Supervised Bot

DEMO LINK {WIP}:
http://35.228.74.197:3031/



Building blocks of (Deep) NLP



Non-Linearities

ReLU
Sigmoid
Tanh
GRU
LSTM
Linear
...

Optimizer

SGD
Momentum
RMSProp
Adagrad
Adam
Second Order (KFac)
...

Connectivity Pattern

Fully connected
Convolutional
Dilated
Recurrent
Self Attention
Recursive
Skip / Residual
Random

Loss

Cross Entropy
Adversarial
Variational
Max. Likelihood
Sparse
L2 Reg
REINFORCE
...

Hyper Parameters

Learning Rate
Decay
Layer Size
Batch Size
Dropout Rate
Weight init
Data augmentation
Gradient Clipping
Beta
Momentum

Deep`NLP Overview

Ideas / Tricks

Soft inputs (**Word embeddings**), Soft state (RNNs), **Gated connections**, **Computation graph**/Autodiff, Non-saturating non-linearities, **Gradient clipping**, **Cross-lingual Embeddings**, **Encoder-decoder (seq2seq)**, **Skip-connections**, Word-pieces, Distillation, **Self-Attention**

Layers

Embedding, Dropout, **Residual Blocks**, LayerNorm, **MH-Self-Attention**,

Architectures

FFNs, **RNNs**, **ConvNets**, **LSTMs**/GRUs, **Transformers**

Optimizers

SGD, **Backpropagation**, BPTT, Momentum, Adam; better initialization

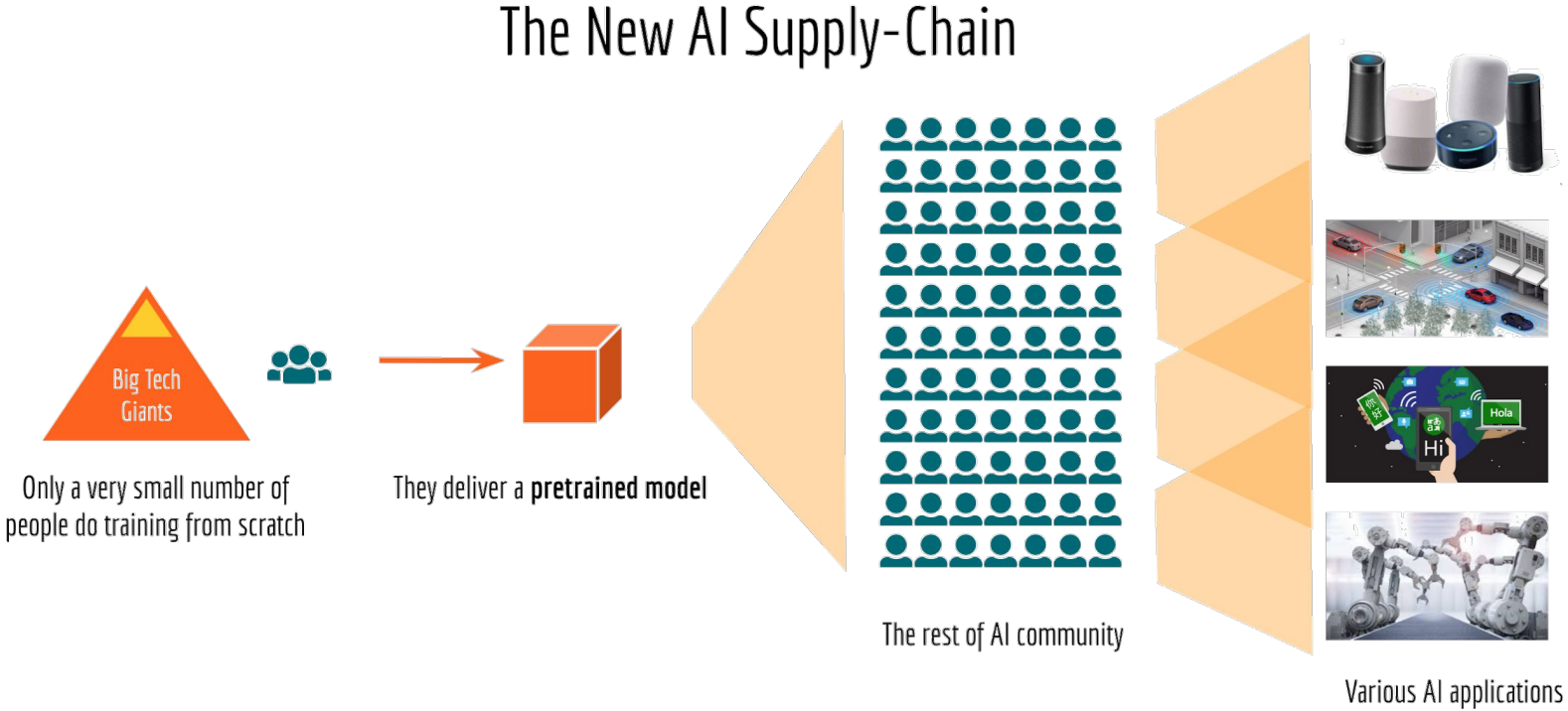
Frameworks & Platforms

Theano, TensorFlow, PyTorch, etc.

Faster CPUs, GPUs, TPUs; easier deployment to accelerators (write-once, run-anywhere)

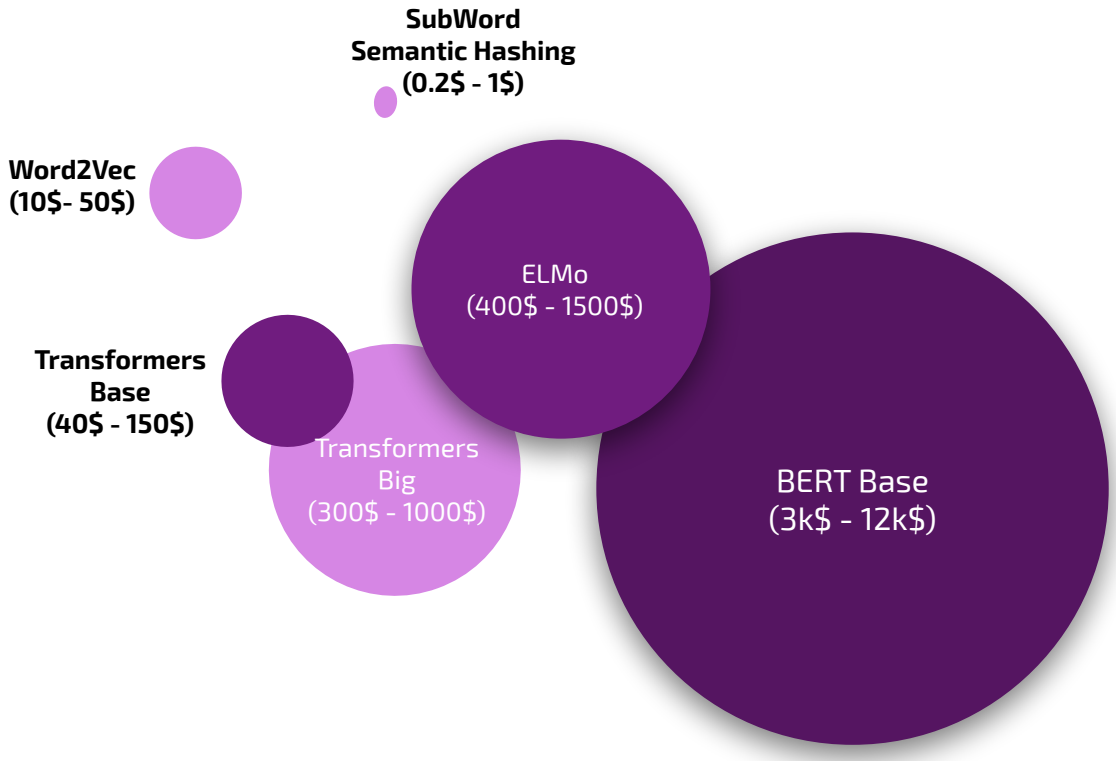
But?

Modern AI Ecosystem



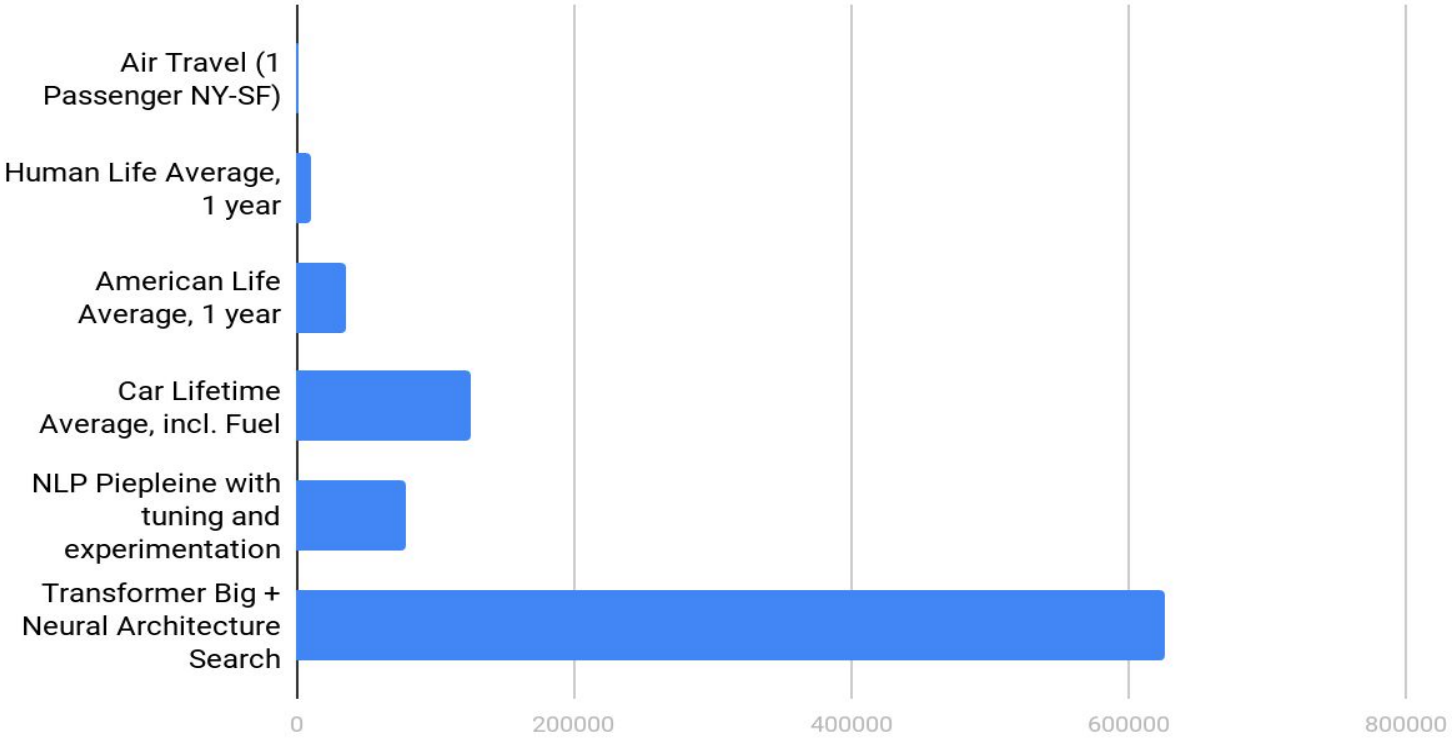
Courtesy: Han Xiao - Engineering Lead @ Tencent AI Lab

Cloud Compute Training Cost



Source: [Energy and Policy Considerations for Deep Learning in NLP]

Estimated Carbon Dioxide Emission Cost



Source: [Energy and Policy Considerations for Deep Learning in NLP]

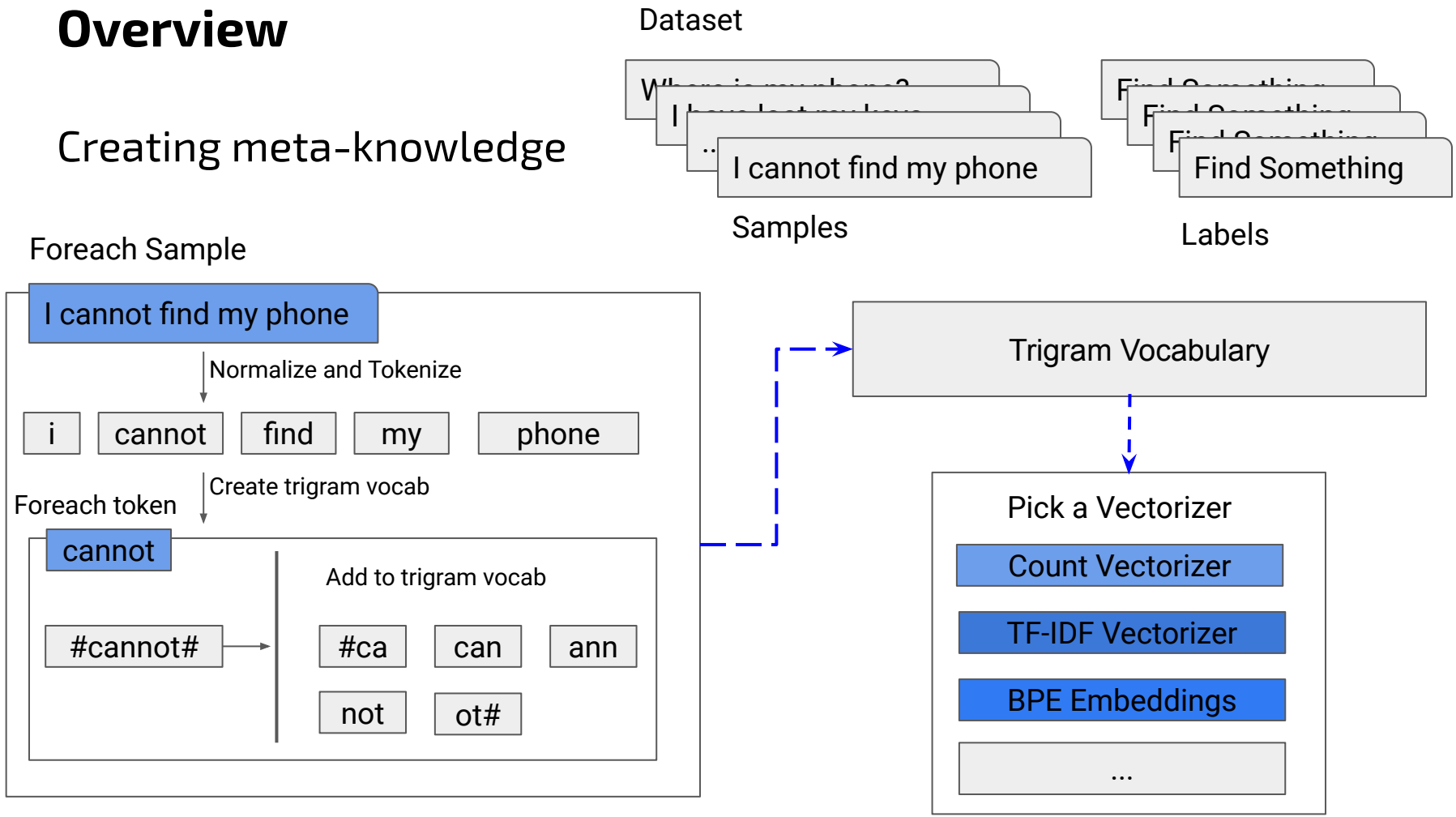
Our Approach:

Meta Learning

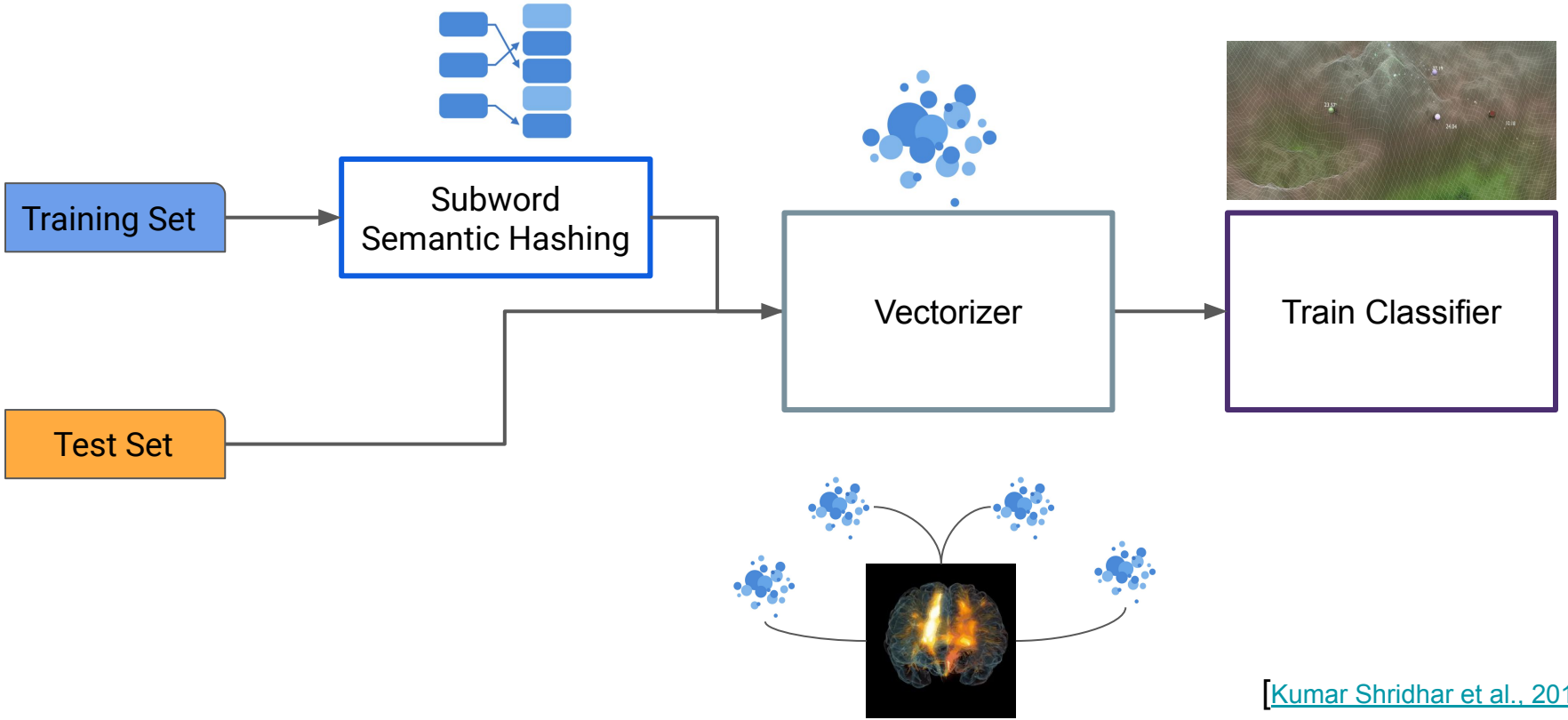
(learning to learn)

Overview

Creating meta-knowledge



Where Semantic Hashing Fits?



[Kumar Shridhar et al., 2018]

Accuracy comparison against various frameworks

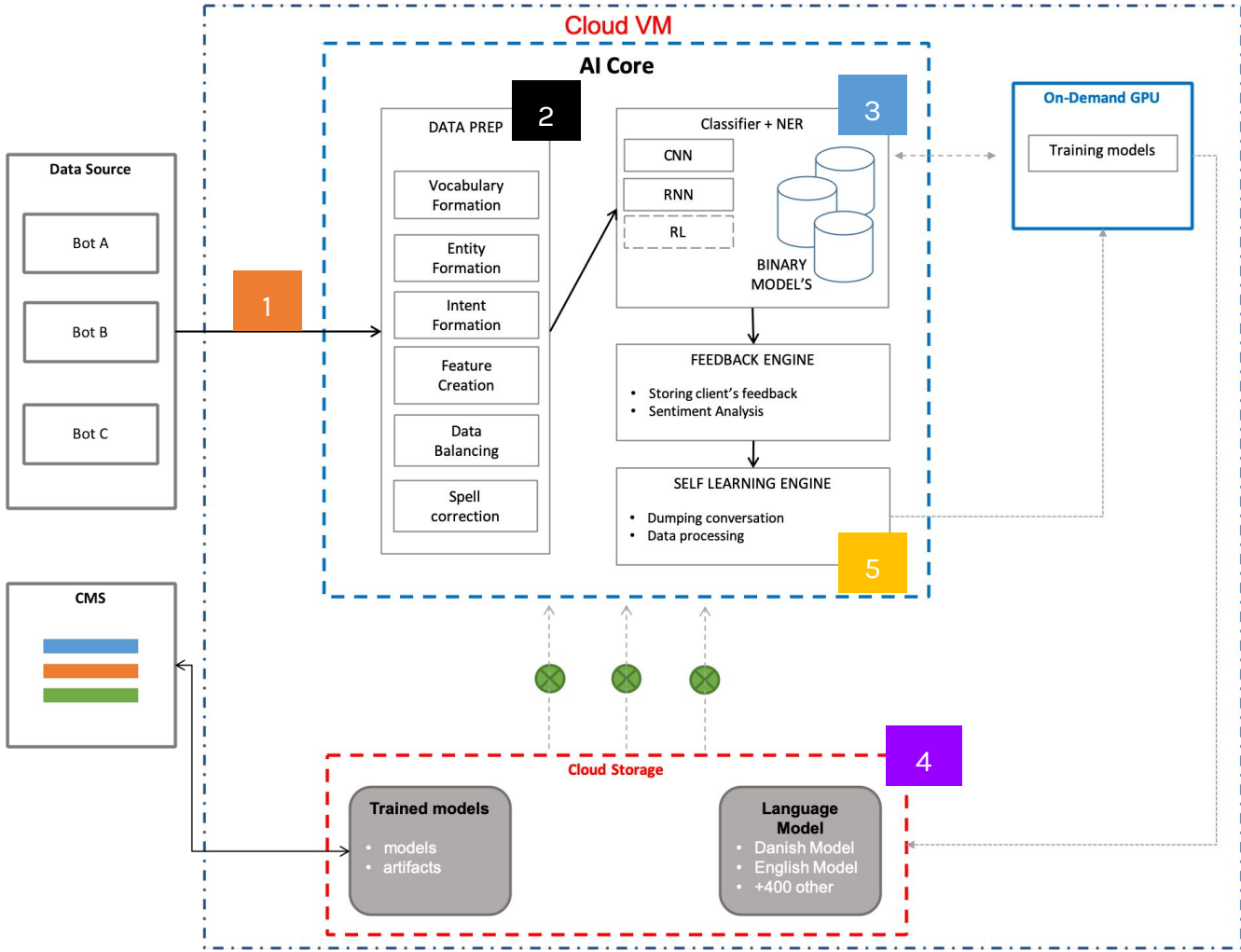
Platform	Chatbot	AskUbuntu	WebApp	Overall	Average
Botfuel	0.98	0.90	0.80	0.91	0.89
Luis	0.98	0.90	0.81	0.91	0.90
Dialogflow	0.93	0.85	0.80	0.87	0.86
Watson	0.97	**0.92**	**0.83**	0.91	**0.91**
Rasa	0.98	0.86	0.74	0.88	0.86
Snips	0.96	0.83	0.78	0.89	0.86
Recast	**0.99**	0.86	0.75	0.89	0.87
TildeCNN	**0.99**	**0.92**	0.81	**0.92**	**0.91**
Our Average	0.98	**0.92**	**0.83**	**0.92**	**0.91**
Our Best	**0.99**	**0.93**	**0.85**	**0.93**	**0.92**
Our Individual Best	***1.00***	***0.93***	***0.86***	***0.94***	***0.93***

[Kumar Shridhar et al., 2018]

Proprietary Framework

STAGES

1. Data Collection
2. Data Processing
3. Model Training
4. Model Deployment
5. Continuous Learning



Natural Language Processing: The Goal

[**Context**: We saw some lions; guide tells us it's a rare sight. Few min later we pass a second safari guide.]

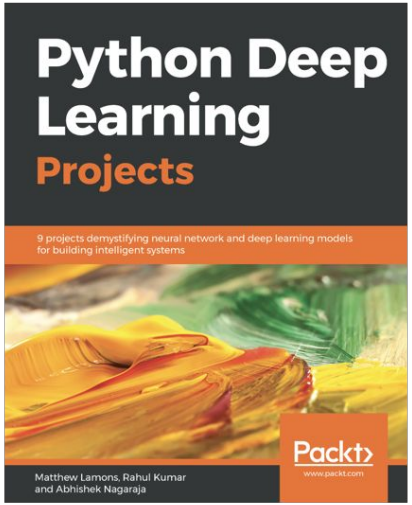
Vi så nogle løver

him

No current models can understand this situation.
Likely need to integrate various sources of information in a multi-step reasoning process.

Which architectectural, training, loss function, etc. advances do we need to get here?

Thank you



Book : https://www.amazon.com/dp/1788997093

Code : http://bit.ly/DeepLearningCode

@hellorahulk

goodrahstar

hellorahulk.com