Data Science for Speech Therapy 3 *Natural Language Processing for Conversation Analysis*

Clinical sharing for continuous education

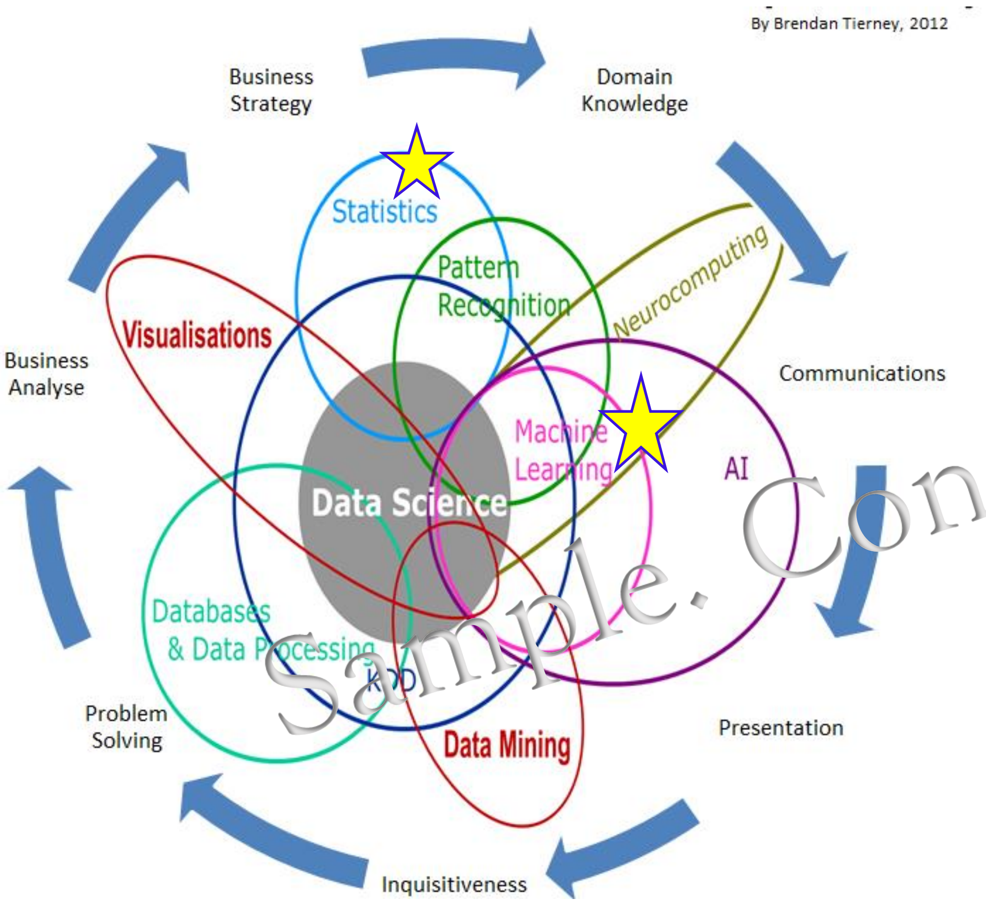
Oct 2021

ST Benjamin Chow

Content

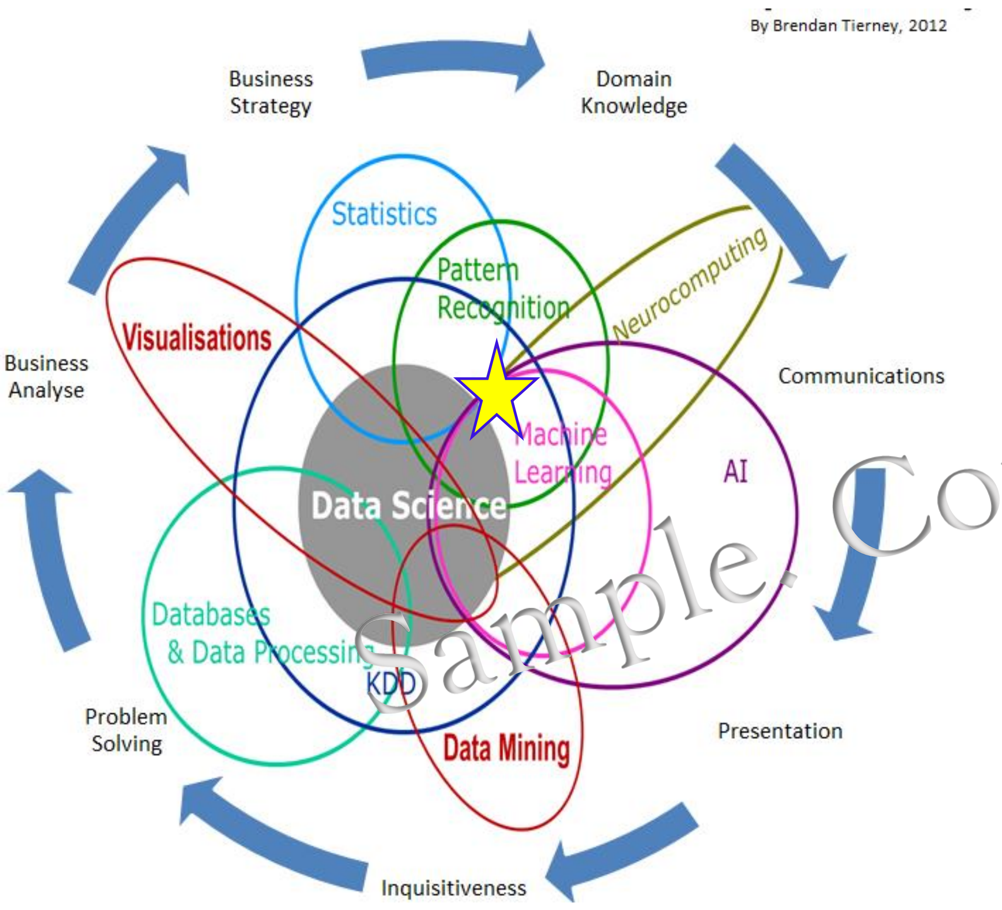
1. **What is Data Science (recap)?**
2. Conversation sample as a form of data
3. DS theory
4. Usage in research papers

Data Science for ST Part 1



	Stats	ML
Goal	Inference	Prediction
Model complexity	Simple models	Simple- complex models
# of models	1 specific model	As many as you like
Data	Tabular	More than just tabular

Data Science for ST Part 2



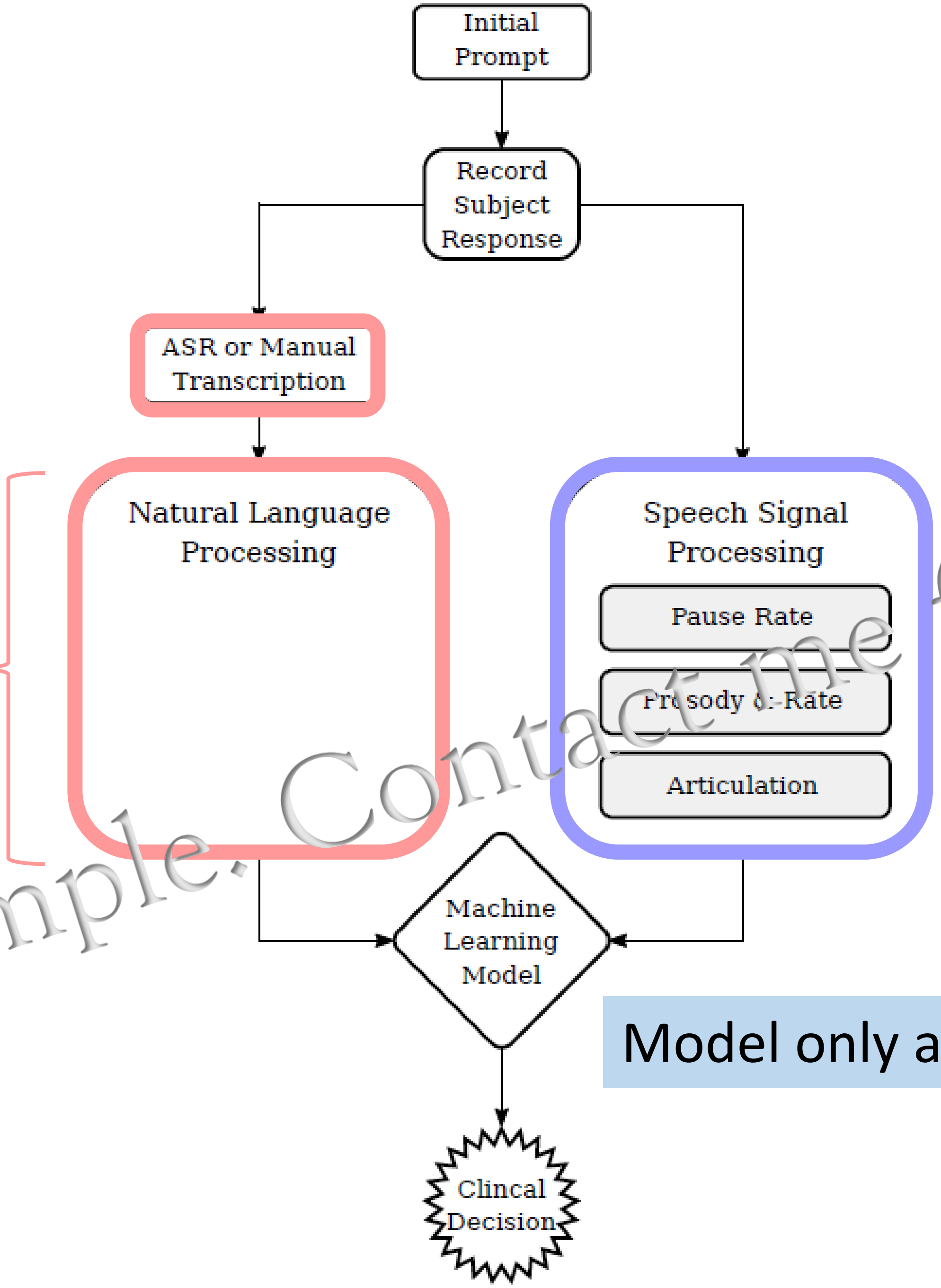
- ML using visual input
  ➔ Computer Vision

- Computer vision for VFSS

Content

1. ~~*What is Data Science (recap)?*~~
2. **Conversation sample as a form of data**
3. DS theory
4. Usage in research papers

How do we convert text into numbers?



Model only accepts numbers

A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders (2019)

Content

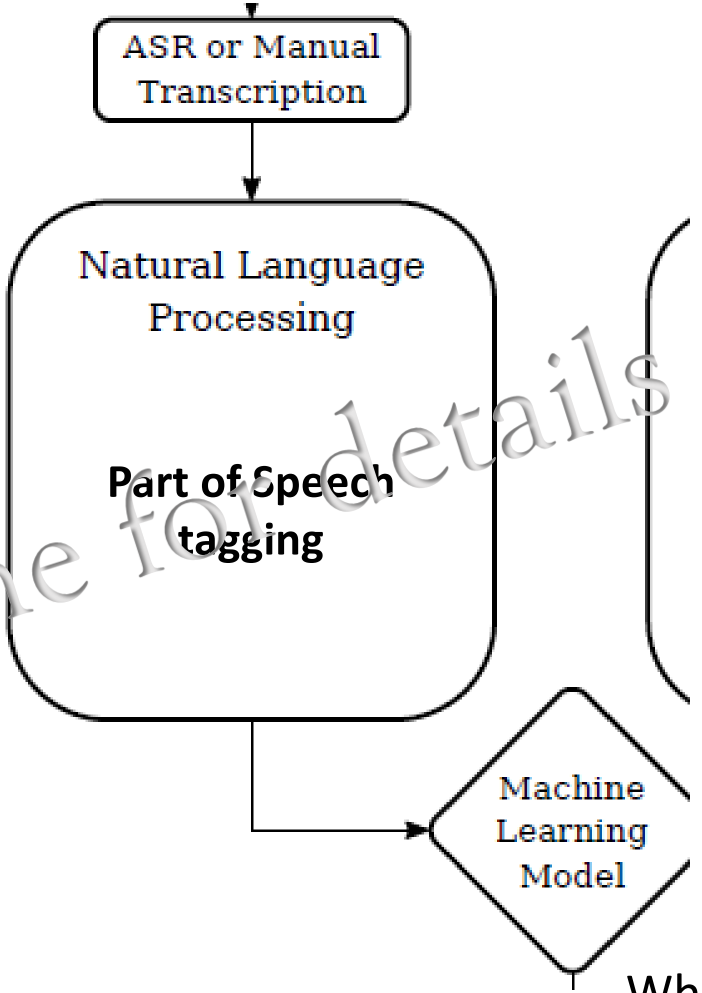
1. ~~*What is Data Science (recap)?*~~
2. ~~*Conversation sample as a form of data*~~
3. **DS theory (NLP)**
    1. **Linguistic features (frequency/ measurements)**
    2. Distributional Semantic Models (DSM)
4. Usage in research papers

Can you tell me how to get to Sesame Street ?

The moon , the bear and the Big-Blue House

Adjective	Determiner	Preposition
Adverb	Noun	Pronoun
Conjunction	Number	Verb

#	Adj	Adv	Conj	Det	Noun	Num	Prep	Pron	Vb
01	0	1	0	0	1	0	0	2	3
02	0	0	1	3	3	0	0	0	0



What kid show is the lyrics from?

https://parts-of-speech.info/

Linguistic features (cont)

<u>Other linguistic features?</u>

- hint: domain-related

<u>Disadvantage</u>

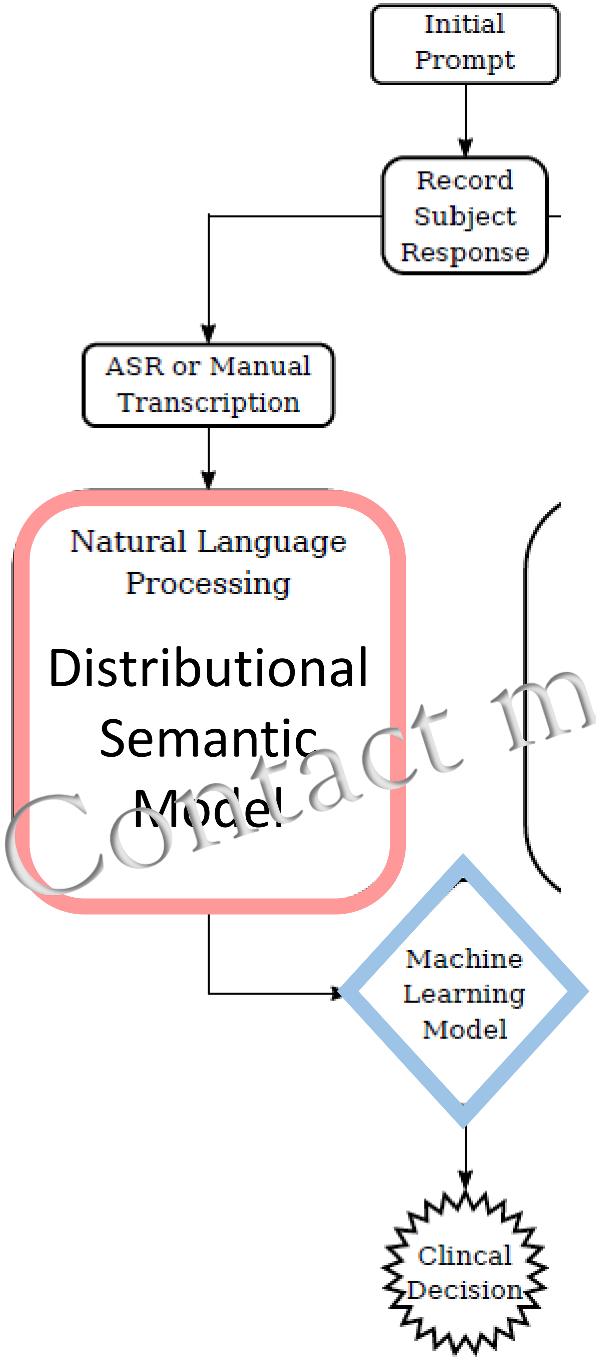
- Doesn’t reveal how individual lexical units interact with each other in a full sentence
- Provide little insight regarding semantic similarity between words
  - *“Car”, “vehicle” and “automobile” are treated as distinct nouns but are semantically similar*

A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders (2019)

Content

1. ~~*What is Data Science (recap)?*~~
2. ~~*What type of data is a conversation sample*~~
3. DS theory (NLP)
   1. ~~*Linguistic features (frequency/ measurements)*~~
   2. **Distributional Semantic Models (DSM)**
4. Usage in research papers

Model= NLP model to convert text to numbers.
Output of NLP model/DSM is input of ML model.



Model= ML model to predict clinical question

A Review of Automated Speech and Language Features for Assessment of Cognitive and Thought Disorders (2019)

How do DSM assign numbers to text?

DSM	
Vocab list	Vector
Word 1	1, 2, 3
Word 2	45, 69, 1
Word 3	1, 0, 0
Word ###	…….

Transcript to convert to numbers		
Word 3	Word 2	Word 1

Words embedded in vector of numbers
(1, 0, 0), (45, 69, 1), (1, 2, 3)

Distributional Semantic Models

DSM	
Vocab list	Vector
Word 1	1, 2, 3
Word 2	45, 69, 1
Word 3	1, 0, 0
Word ###	…….

- Words are embedded as a vector of numbers
- How are the vector of numbers determined?
  1. Count-based models
  2. Prediction-based models
  3. Deep contextualized models

Problems with count-based DSM

		Transcript								
		Can	You	Tell	Me	How	To	Get	To	Sesame Street
Vocab List	Can	1	0	0	0	0	0	0	0	0
	You	0	1	0	0	0	0	0	0	0
	Tell	0	0	1	0	0	0	0	0	0
	Me	0	0	0	1	0	0	0	0	0
	How	0	0	0	0	1	0	0	0	0
	To	0	0	0	0	0	1	0	1	0
	Get	0	0	0	0	0	0	0	1	0
	Drive	0	0	0	0	0	0	0	0	0
	Sesame street	0	0	0	0	0	0	0	0	1
	Big Blue House	0	0	0	0	0	0	0	0	0

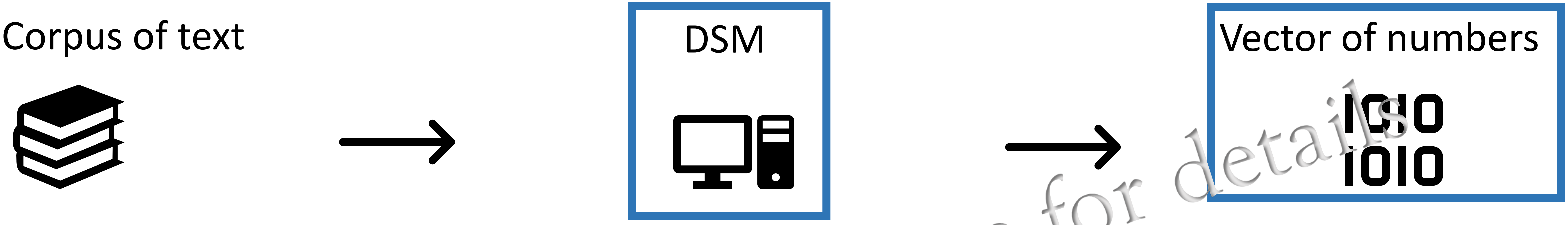
<u>Way too many zeros</u>

- models prefer dense vector of numbers
- Slower for computation

<u>Semantic similarity</u>

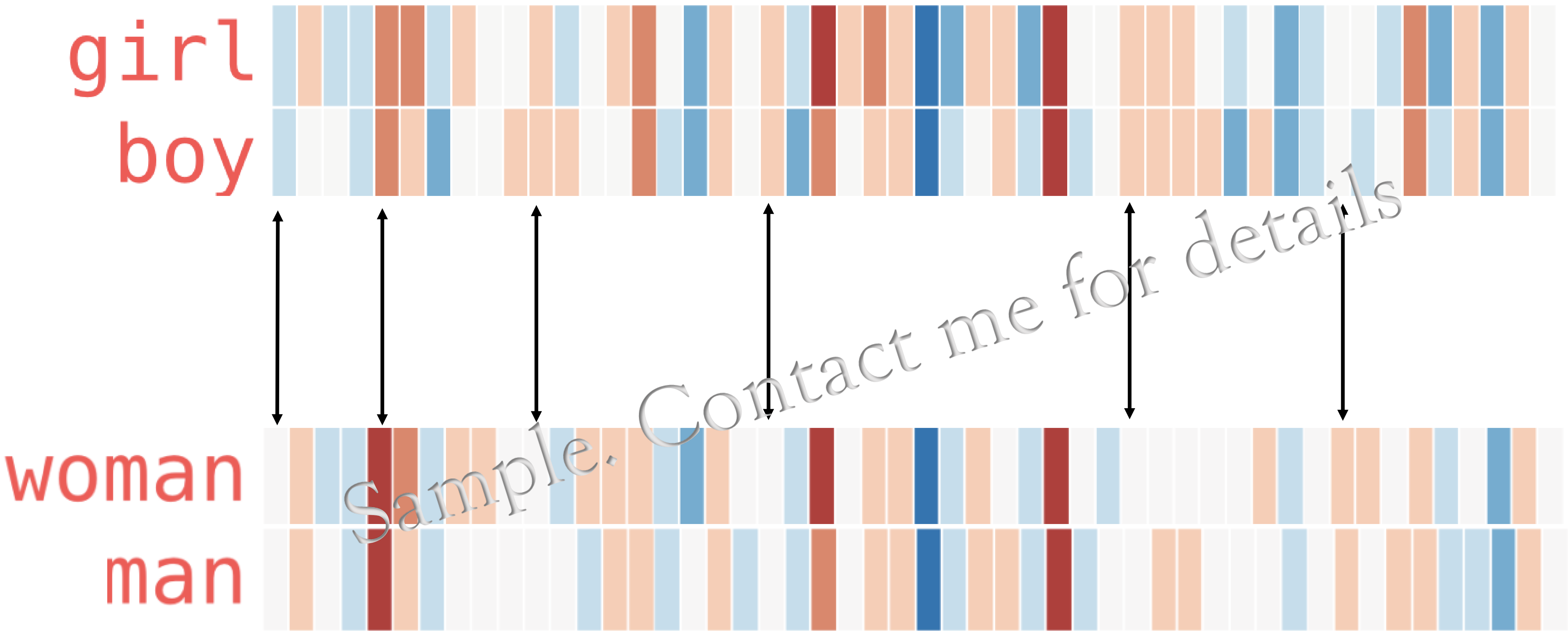
- Still not capture
- 1/0 if *"car"*, *"vehicle"*, *"automobile"* present

Distributional Semantic Models 2: Prediction-based



- Behind the scene, some form of predictive modelling is done by DSM
  - Prediction-based DSM

- Semantic similarity is captured in vector of numbers
  - More related the words are, the closer the values of the vector of numbers

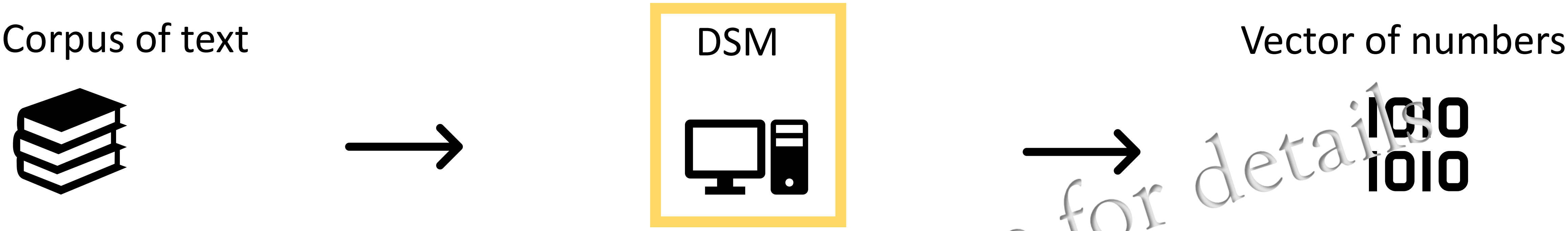

girl
boy
woman
man

Problems with prediction-based DSM

General

- Mistake words with opposite meanings as their meanings are very similar and used in similar contexts
- Polysemy words have the same vector of numbers
- Unable to specific type of semantic relationship *(what are the clinical implications?)*

Distributional Semantic Models 3: Deep Contextualized



<u>Deep Contextualized DSM learn word AND context meanings</u>

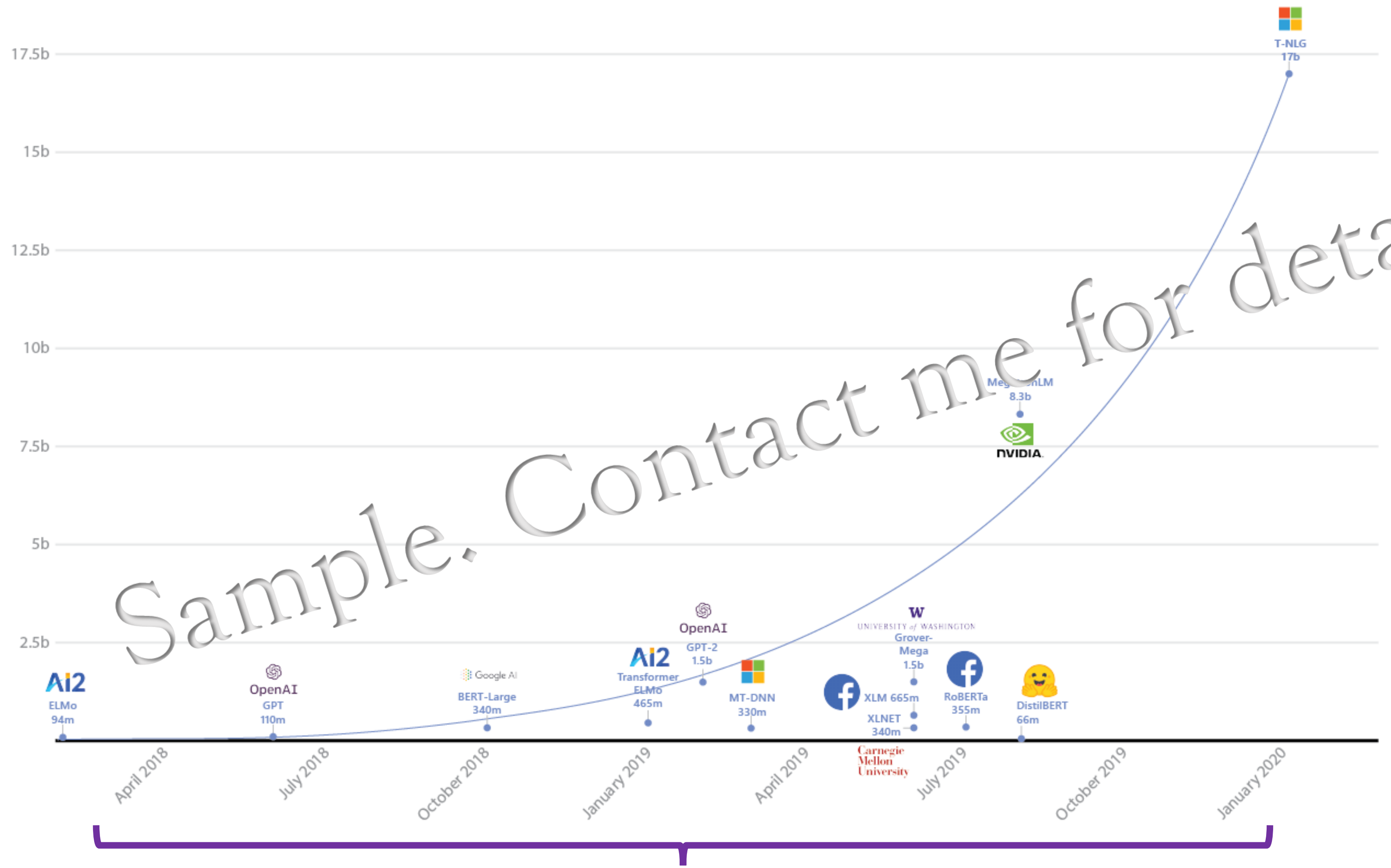
- Compress context into the vectors
- Vector representation of word is informed by surrounding words.
- Takes into account other words so the vector representation is aware of context

Vectors of Deep Contextualized DSM

- Vector representation of word changes depending on the sentence it appears.
- Solves the uniform representation of polysemy words by prediction-based DSM

https://www.youtube.com/watch?v=6f90OTW-6Ic

State Of The Art (SoTA) Models



https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

Explain how BERT learns ….

- Why the need to explain?

<u>ML researcher</u>

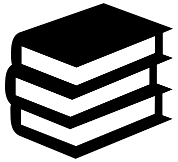
- Create better models

<u>SLP</u>

- Closer to our theoretical understanding, the easier for us to trust and adopt the models

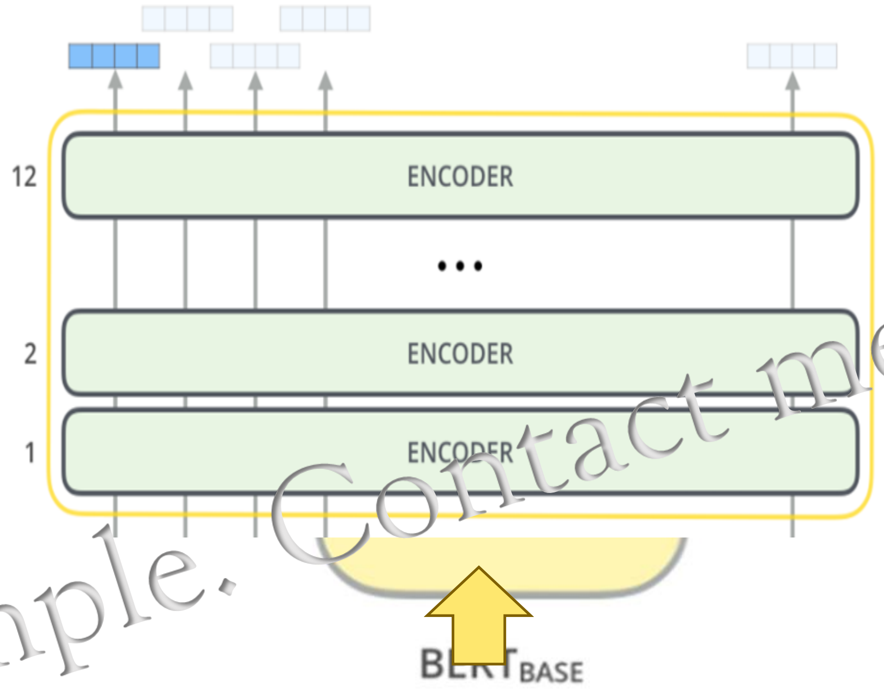
BERT (Bidirectional Encoder Representations from Transformers)

Corpus of text



Vector of numbers

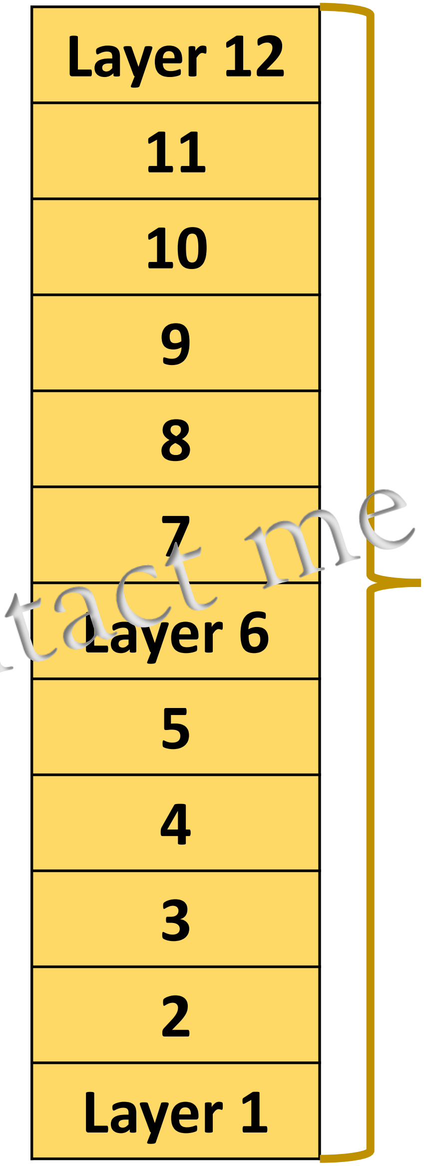
1010
1010



Preparatory transformation for BERT

http://jalammar.github.io/illustrated-bert/

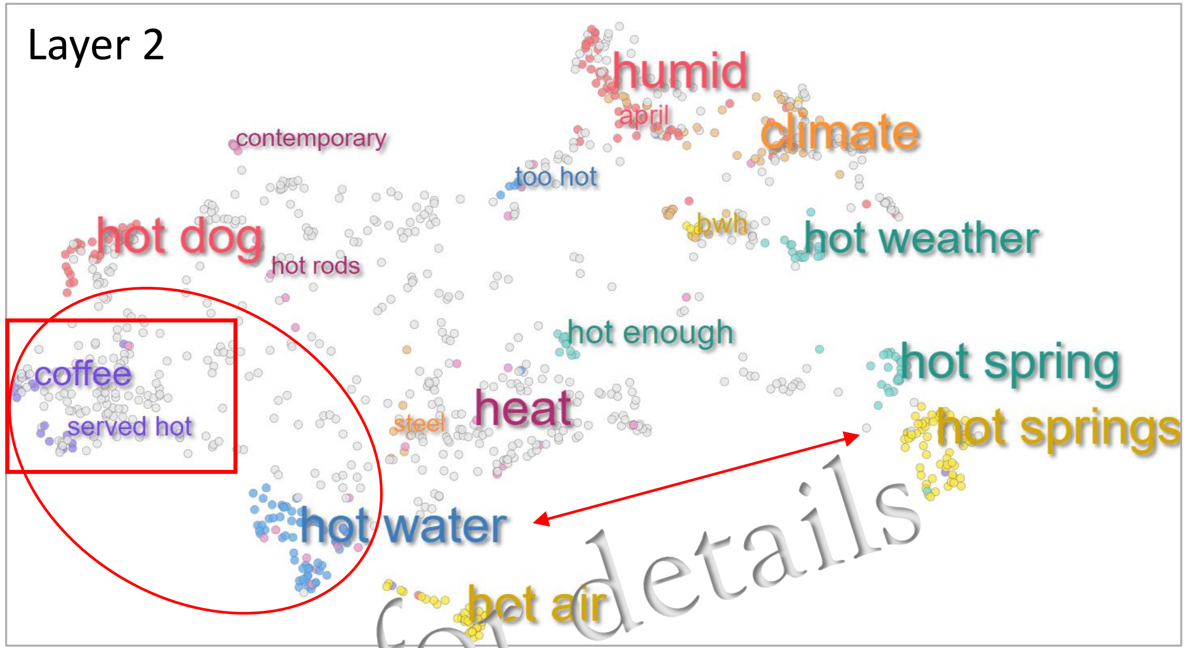
Explaining how BERT learns semantics

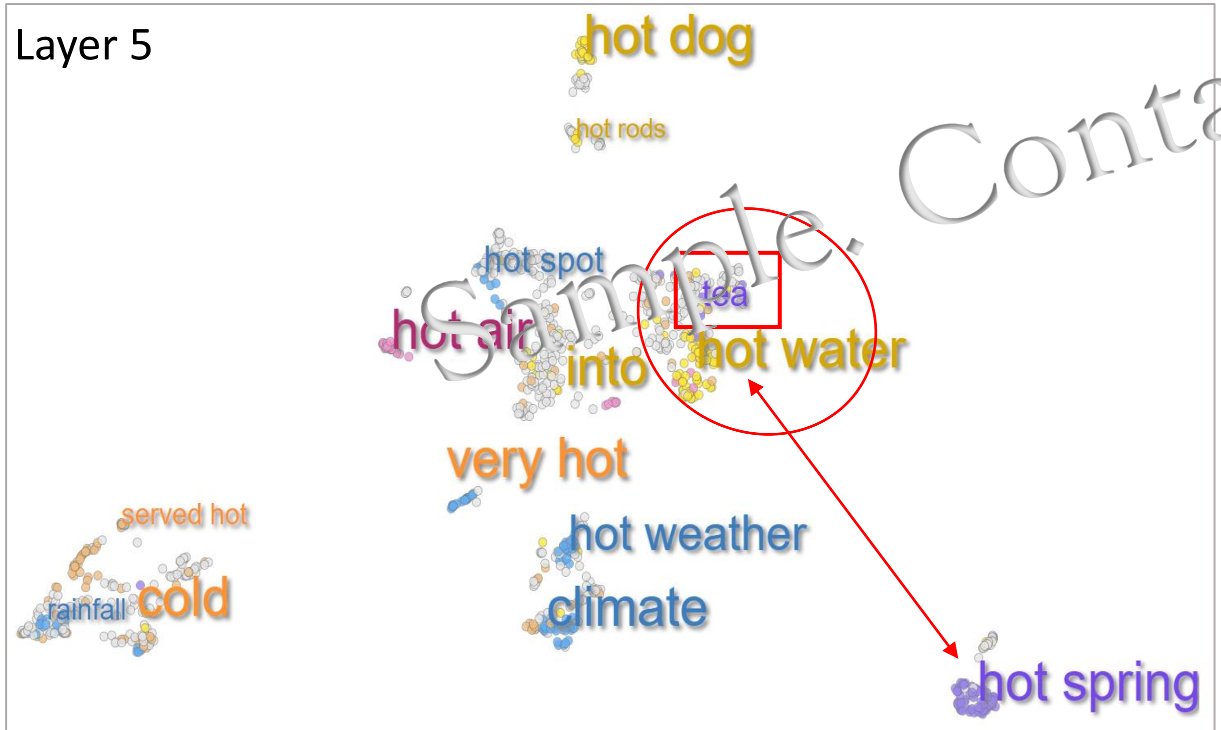


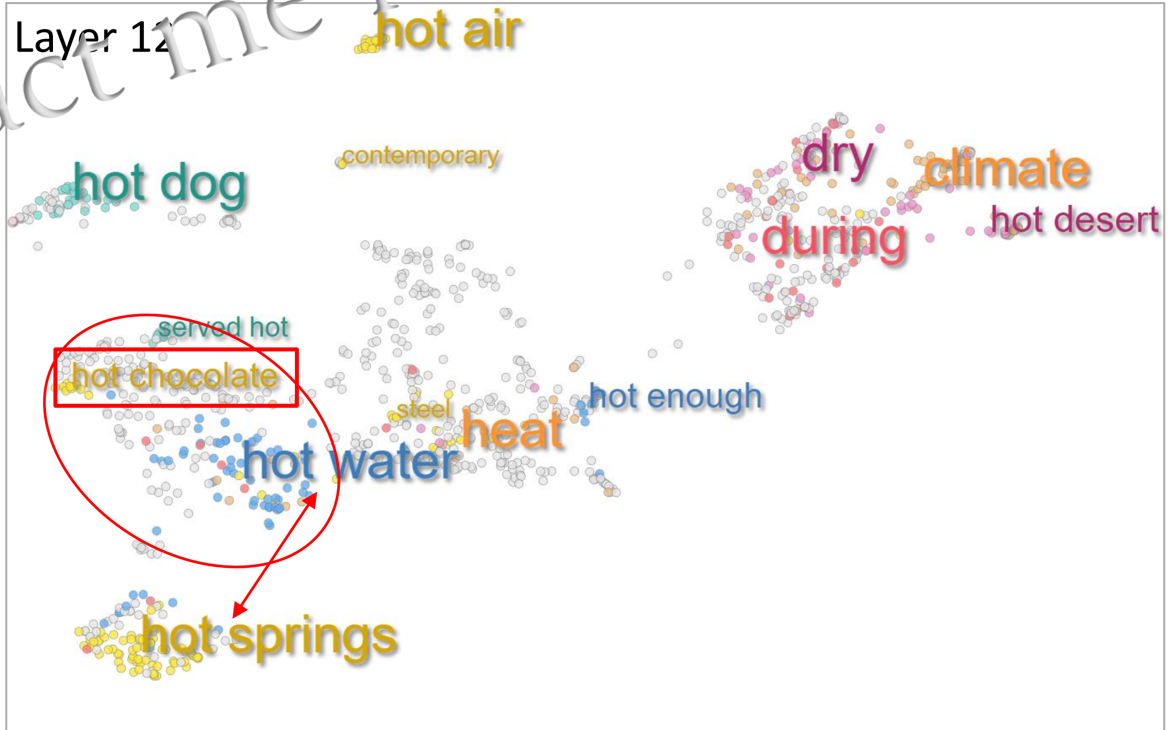
- Semantics is spread across the entire model
- Some word senses learned at earlier layers may be dropped.
- Some word senses are learned at later layers.
- Some cases, more context specific representation develop in later layers

A Primer in BERTology: WhatWe Know About How BERT Works (2020); https://www.youtube.com/watch?v=6f90OTW-6Ic; How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings (2019); https://pair-code.github.io/interpretability/context-atlas/blogpost/index.html

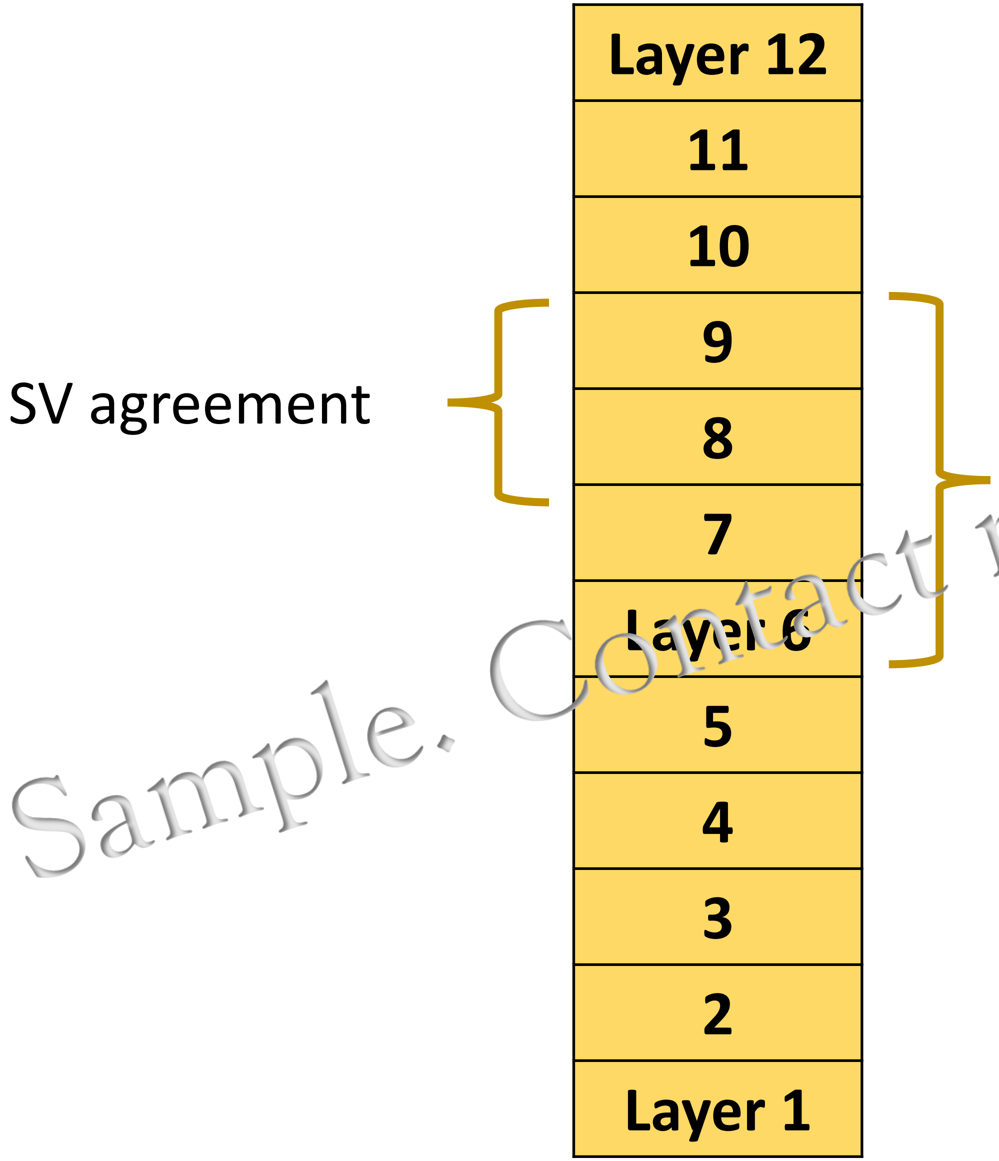
[Context Atlas (storage.googleapis.com)](storage.googleapis.com)







Explaining how BERT learns syntax

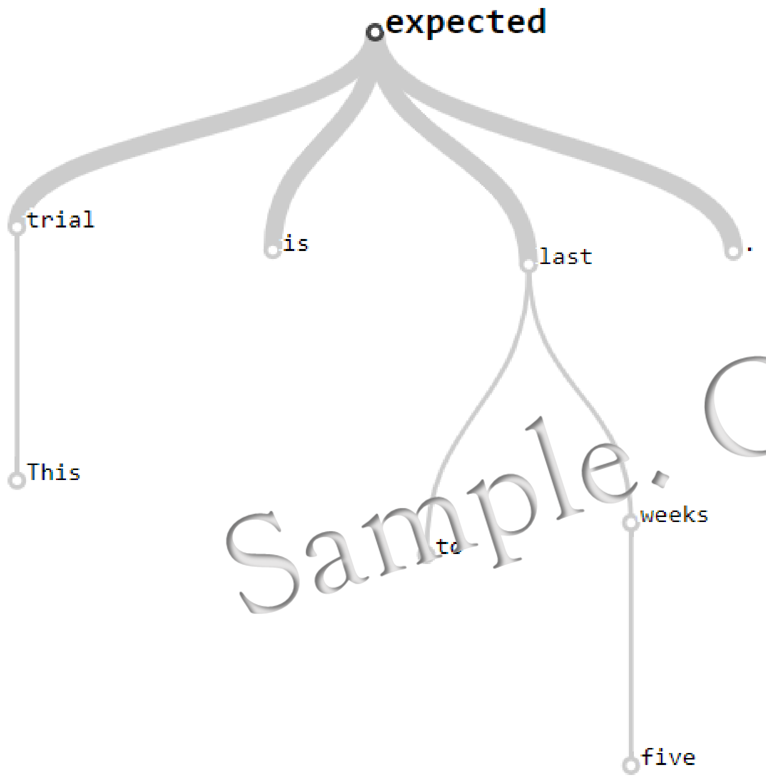


syntactic information is most prominent in middle layers of BERT.

A Primer in BERTology: WhatWe Know About How BERT Works (2020); Assessing BERT’s syntactic (2019)

What is a syntax dependency tree?

“This trial is **expected** to last five weeks.”

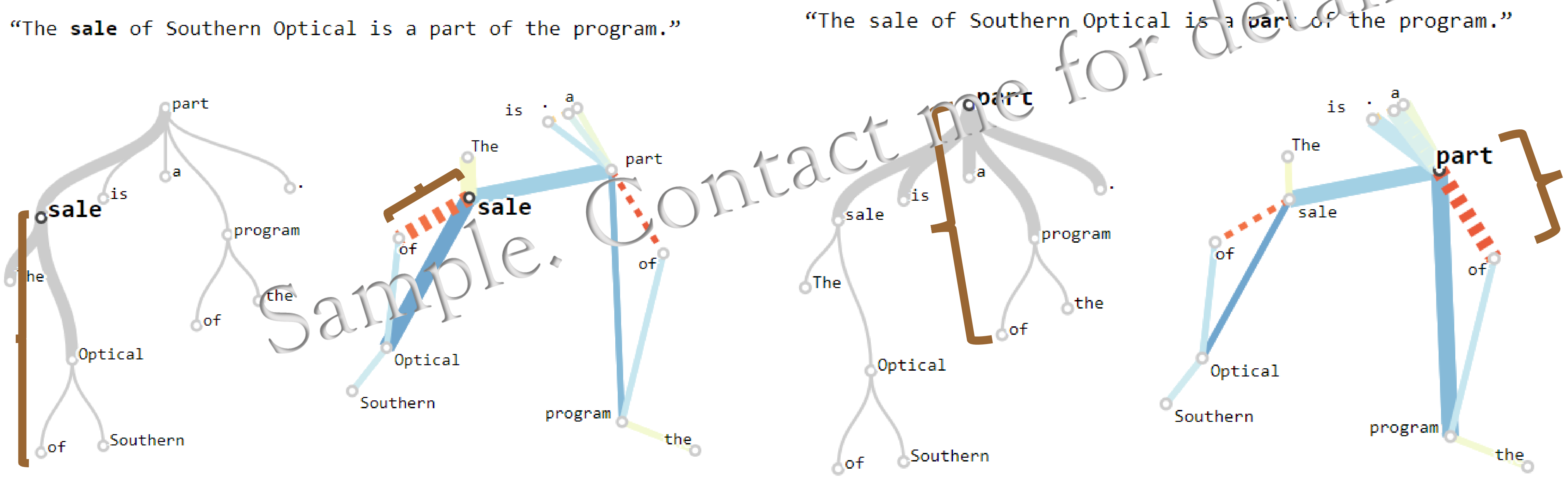


- Main verb is the root node which the tree grows
- Distance allows depth to be formed
- Depth represents the hierarchical nature of sentences
  - Parent-child relationship
  - Subtrees with the tree

https://pair-code.github.io/interpretability/bert-tree/ ; https://towardsdatascience.com/getting-to-grips-with-parse-trees-6e19e7cd3c3c

Too good to be true?

Words without dependency relation but positions were closer than expected *(dotted orange lines)*



Comments about BERT learning grammar

- BERT "naturally" learns syntactic information
- But there are differences compared to human linguistic
  - Computational language models are just different VS undiscovered findings/ unproved hypothesis

➔ Is it still appropriate to use such models with utterances from aphasic/cog com speeches?

Content

1. ~~*What is Data Science (recap)?*~~
2. ~~*What type of data is a conversation sample*~~
3. ~~*DS theory (NLP)*~~
4. **Usage in research papers**

NLP in research of acquired communication disorders

Frequency

- Limited

+ Across demographic: Aphasia, AD, PPA

Trend

- More articles in recent years
- Biostatistics -> ML approach
- Linguistic features -> NLP >> linguistic features

Data:

- Assessment battery
- Conversation sample
- Assessment battery + conversation sample

NLP in research of acquired communication disorders

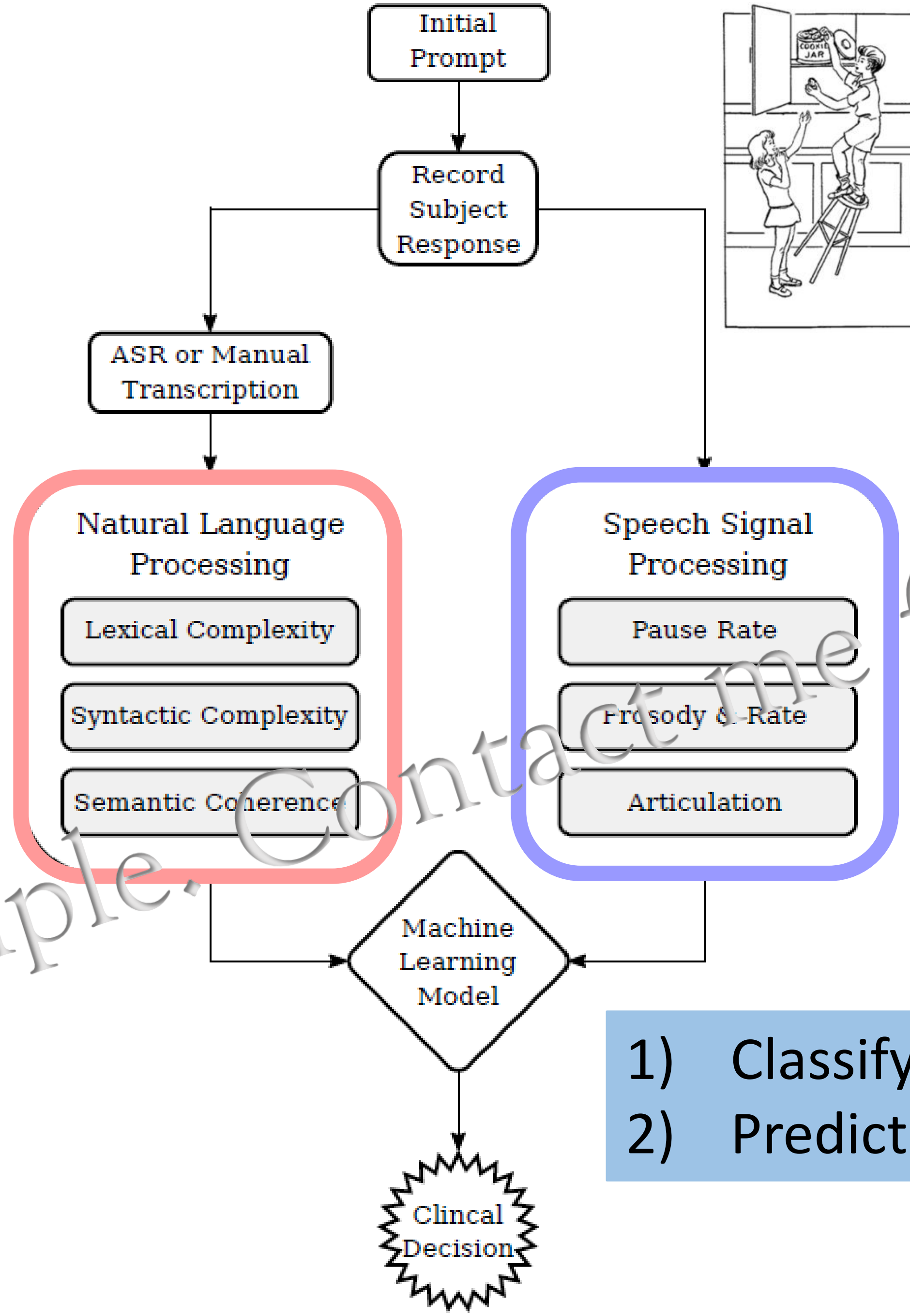
<u>Results</u>

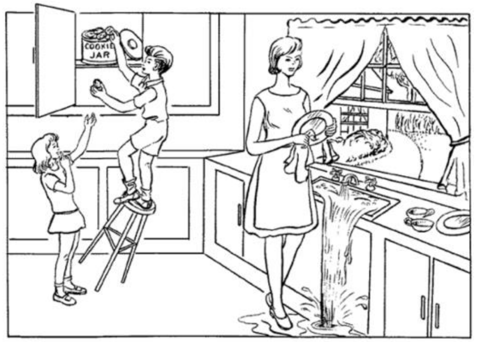
- Multi level biasness
- Powerful model OR easy data
- Lack of standardization hinders its translation into clinical practice

➢AD researchers acknowledged the current limitations

➢Created a balanced dataset and established as a benchmark challenge:

<u>A</u>lzheimer's <u>D</u>ementia <u>R</u>ecognition through <u>S</u>pontaneous <u>S</u>peech (ADReSS)





1) Classify pts w AD and w/out AD
2) Predict MMSE scores

ADReSS has a balanced dataset

1. “Training Set” is used to train the model
2. “Test set” is used to test the model

- Dataset is balanced for both training and test set
- Ratio of M:F
- Ratio of pt w AD and without AD
- Age groups

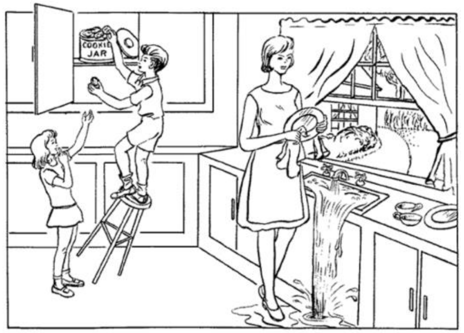
Table 1: *ADReSS Training Set: Basic characteristics of the patients in each group (M=male and F=female).*

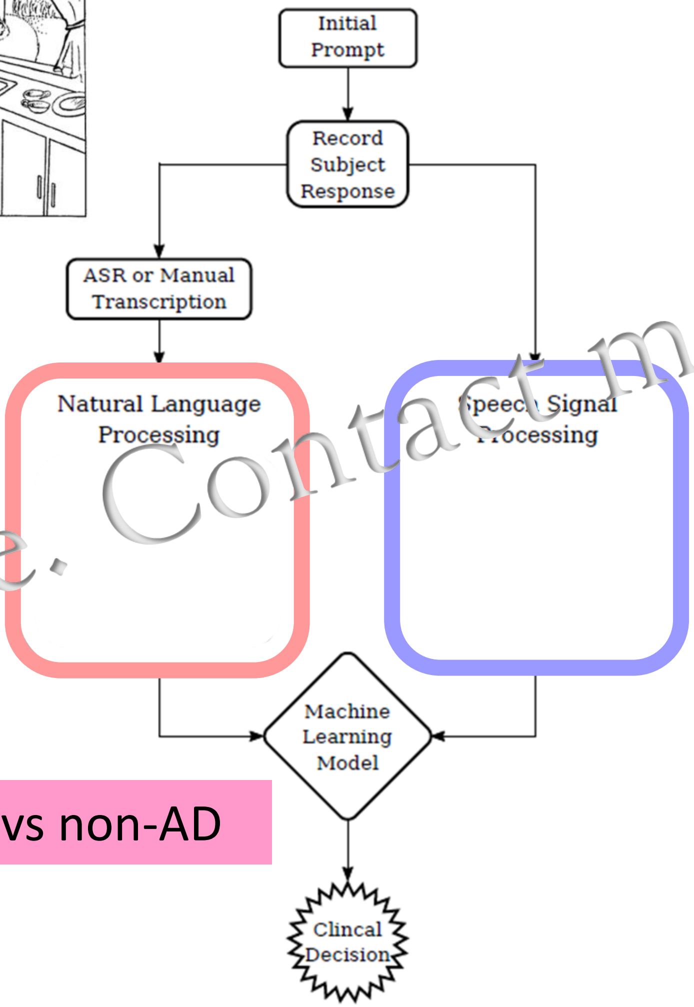
	AD			non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55)	1	0	30.0 (n/a)	1	0	29.0 (n/a)
[55, 60)	5	4	16.3 (4.9)	5	4	29.0 (1.3)
[60, 65)	3	6	18.3 (6.1)	3	6	29.3 (1.3)
[65, 70)	6	10	16.9 (5.8)	6	10	29.1 (0.9)
[70, 75)	6	8	15.8 (4.5)	6	8	29.1 (0.8)
[75, 80)	3	2	17.2 (5.4)	3	2	28.8 (0.4)
Total	24	30	17.0 (5.5)	24	30	29.1 (1.0)

Table 2: *Characteristics of the ADReSS test set.*

	AD			non-AD		
Age	M	F	MMSE (sd)	M	F	MMSE (sd)
[50, 55)	1	0	23.0 (n.a)	1	0	28.0 (n.a)
[55, 60)	2	2	18.7 (1.0)	2	2	28.5 (1.2)
[60, 65)	1	3	14.7 (3.7)	1	3	28.7 (0.9)
[65, 70)	3	4	23.2 (4.0)	3	4	29.4 (0.7)
[70, 75)	3	3	17.3 (6.9)	3	3	28.0 (2.4)
[75, 80)	1	1	21.5 (6.3)	1	1	30.0 (0.0)
Total	11	13	19.5 (5.3)	11	13	28.8 (1.5)

To BERT or Not To BERT: Comparing Speech and Language-based Approaches for Alzheimer's Disease Detection (2020)



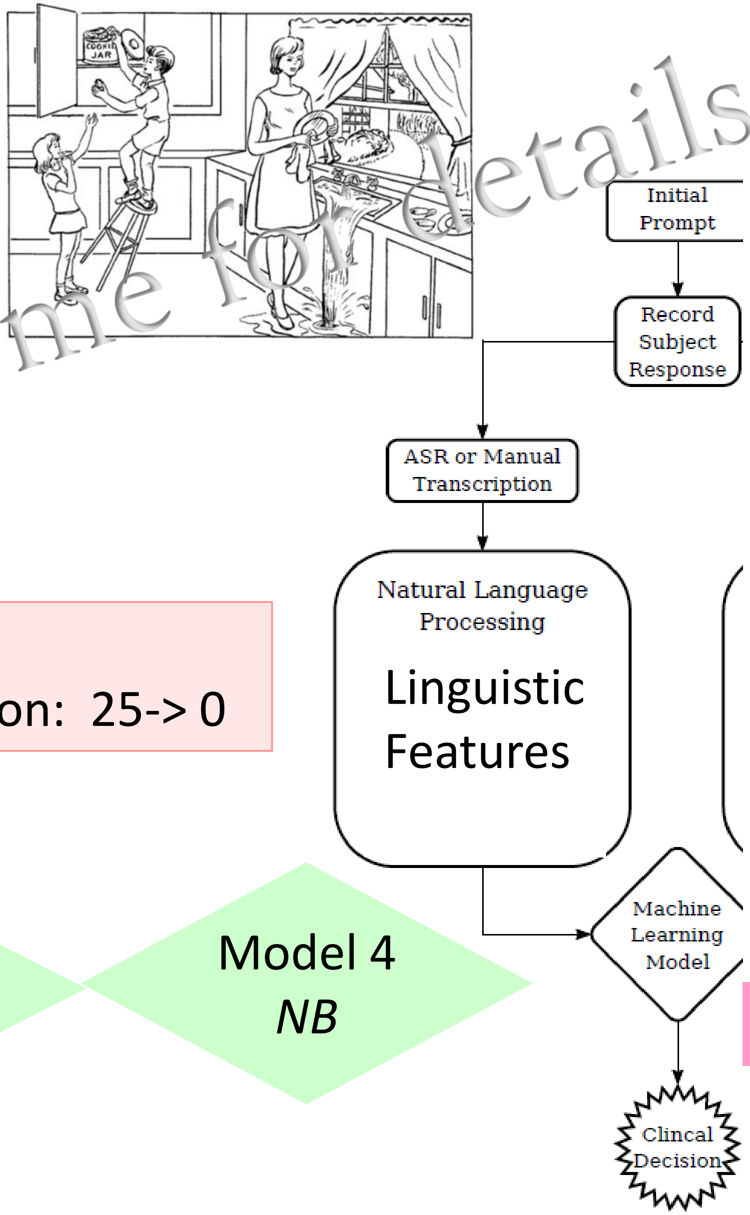


Approach 2:
NLP techniques X2

Approach 1:
187 acoustic features -> 0
Acoustic approach dropped

AD vs non-AD

NLP Technique: # 1



a) Linguistic features: 297 -> 13
b) Features based on picture description: 25-> 0

Model 1 *SVM*

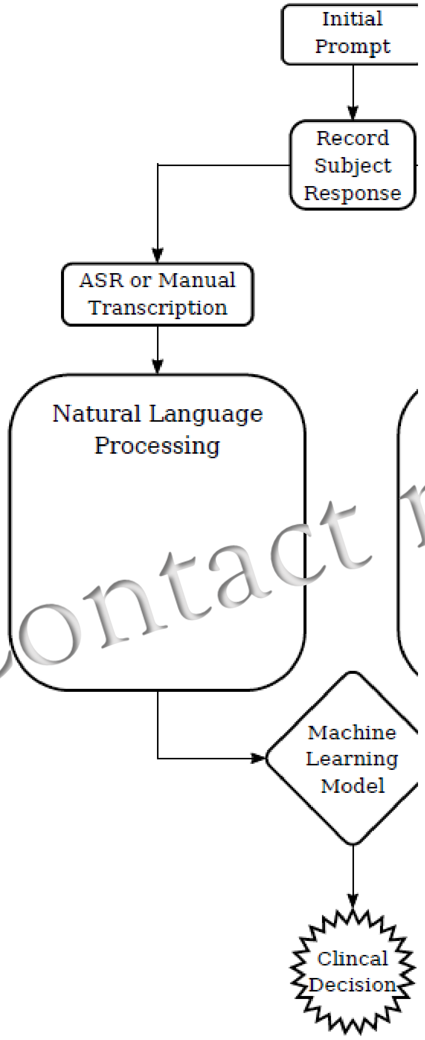
Model 2 *NN*

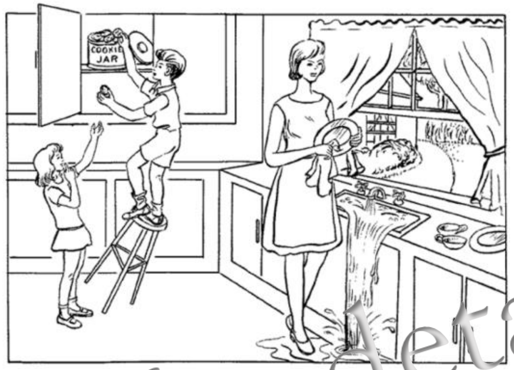
Model 3 *RF*

Model 4 *NB*

AD vs non-AD

NLP Technique: # 2





AD vs non-AD

Results

Table 5: *AD detection results on unseen, held-out ADReSS test set presented in same format as the baseline paper [1]. Bold indicates the best result.*

Model	#Features	Class	Accuracy	Precision	Recall	Specificity	F1
Baseline [1]	-	non-AD	0.750	0.70	**0.87**	-	0.78
		AD		**0.83**	0.62	-	0.71
SVM	10	non-AD	0.813	0.83	0.79	0.83	0.81
		AD		0.80	0.83		0.82
NN	10	non-AD	0.771	0.78	0.75	0.78	0.77
		AD		0.76	0.79		0.78
RF	50	non-AD	0.750	0.71	0.83	0.71	0.77
		AD		0.80	0.67		0.73
NB	80	non-AD	0.729	0.69	0.83	0.69	0.75
		AD		0.79	0.63		0.70
BERT	-	non-AD	**0.833**	**0.86**	0.79	**0.86**	**0.83**
		AD		0.81	**0.88**		**0.84**

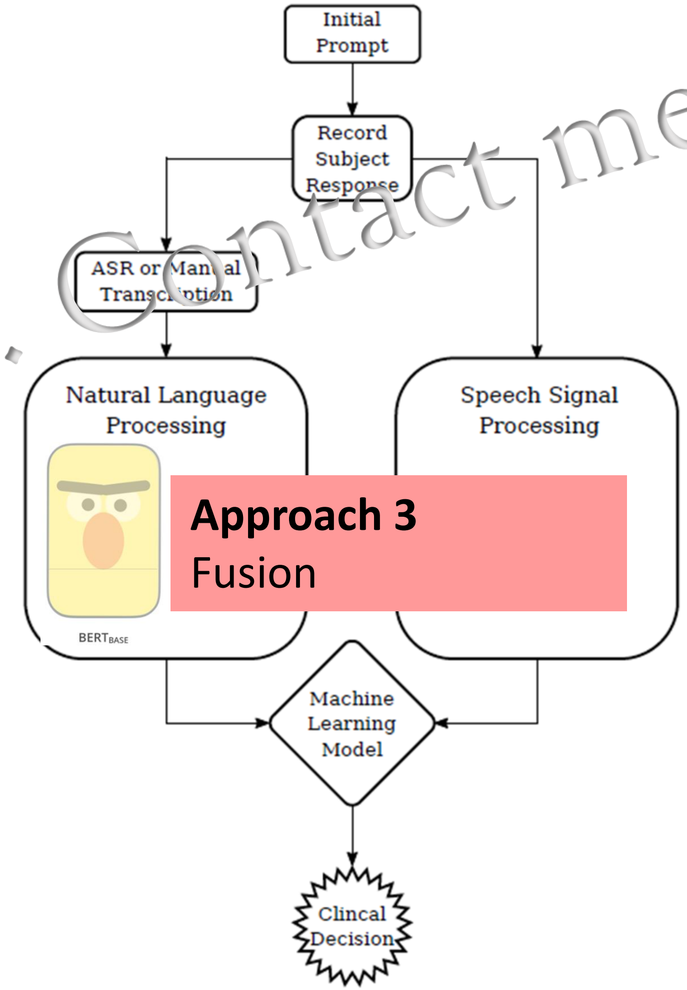
- BERT is (slightly) superior than machine learning models with linguistic features

➔BERT captures a range of linguistic phenomena

➔ Encapsulation of many important lexico-syntactic and semantic features.

Using state of the art speaker recognition and natural language processing technologies to detect Alzheimer’s disease and assess its severity (2020)

Approach 1
SoTA model



Approach 3
Fusion

Approach 2a
characterize speech signals

Approach 2b
characterize presence of silence

Results

Table 2: *ADReSS challenge evaluation results for the and prediction tasks. Best results are marked in bold.*

Models	Class	Detection Prec./Rec.	F1	Accuracy (%)
Baseline	CC	0.67/0.50	0.57	62.50
	AD	0.60/0.75	0.67	
Acoustic	CC	0.61/0.45	0.52	58.00
	AD	0.57/0.71	0.63	
Acoustic + silence	CC	0.64/**0.75**	0.69	66.70
	AD	0.70/0.58	0.63	
Transcript	CC	0.79/0.63	0.7	72.92
	AD	0.69/0.83	0.75	
Acoustic & Transcript	CC	**0.83**/0.63	**0.71**	**75.00**
	AD	**0.70/0.88**	**0.78**	
Acoustic + silence & Transcript	CC	0.79/0.62	0.70	72.92
	AD	0.69/0.83	0.75	

➔ Two modalities contain complementary information

➔ More data ≠ better prediction

Final thoughts

- SoTA NLP models are the new kids on the block for NLP
  - Spilling over to acquired communication disorders research
- SoTA NLP models ≠ best performance
  - Mix and match approaches and techniques (including traditional NLP strategies)
- Considerations when using SoTA NLP models for acquired communication disorders
  - Rubbish in, rubbish out