DEPARTMENT OF INFORMATION ENGINEERING
UNIVERSITY OF PADOVA

QUARTZ
Quantum Information Access and Retrieval Theory

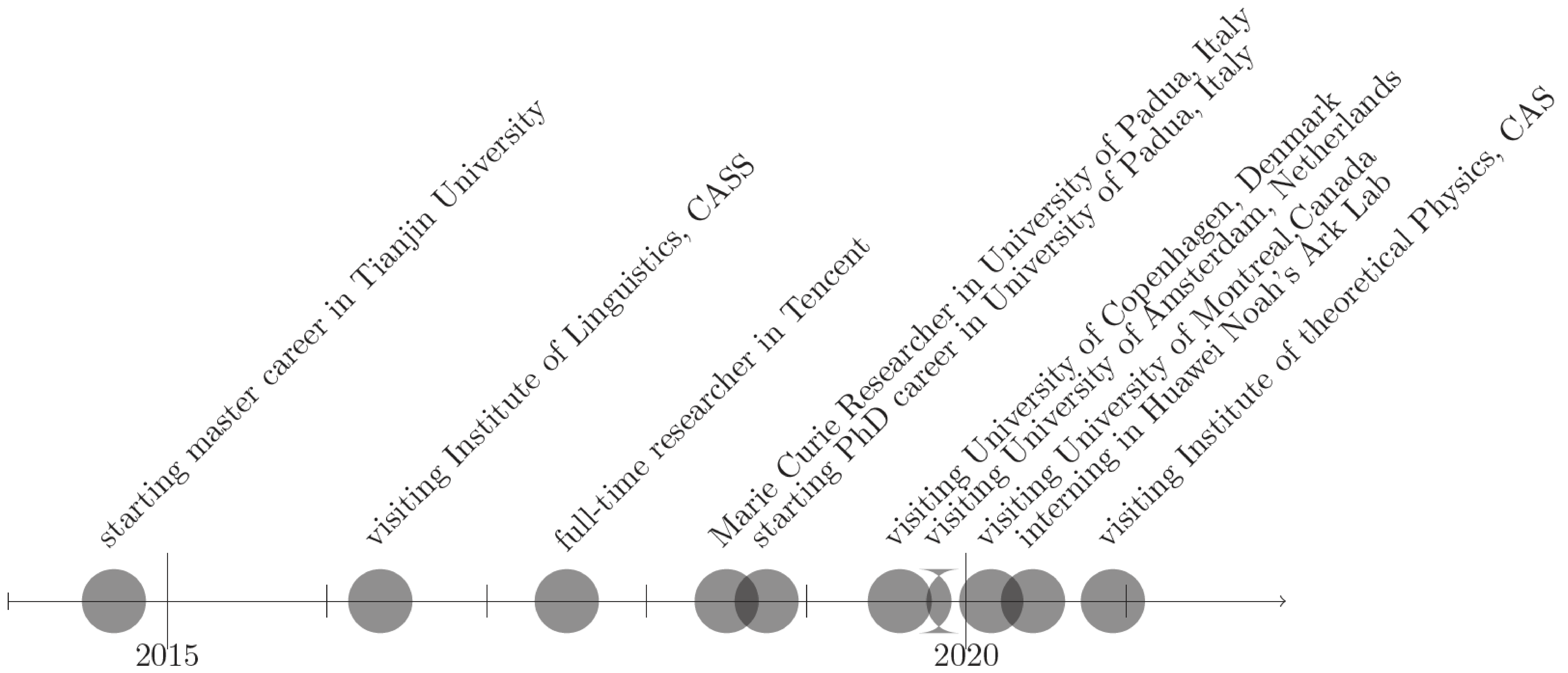
What can quantum physics bring to natural language processing?

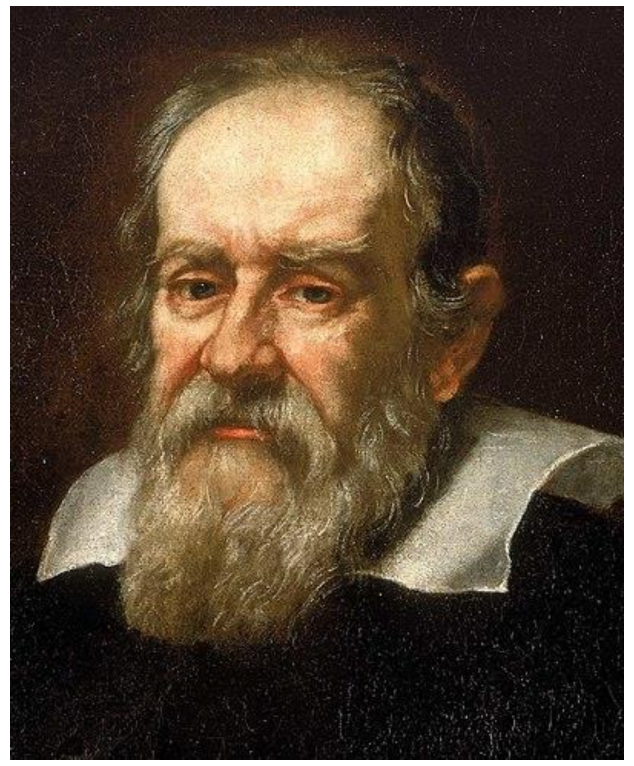
Benyou Wang

Assistant professor

The Chinese University of Hong Kong, Shenzhen

About me

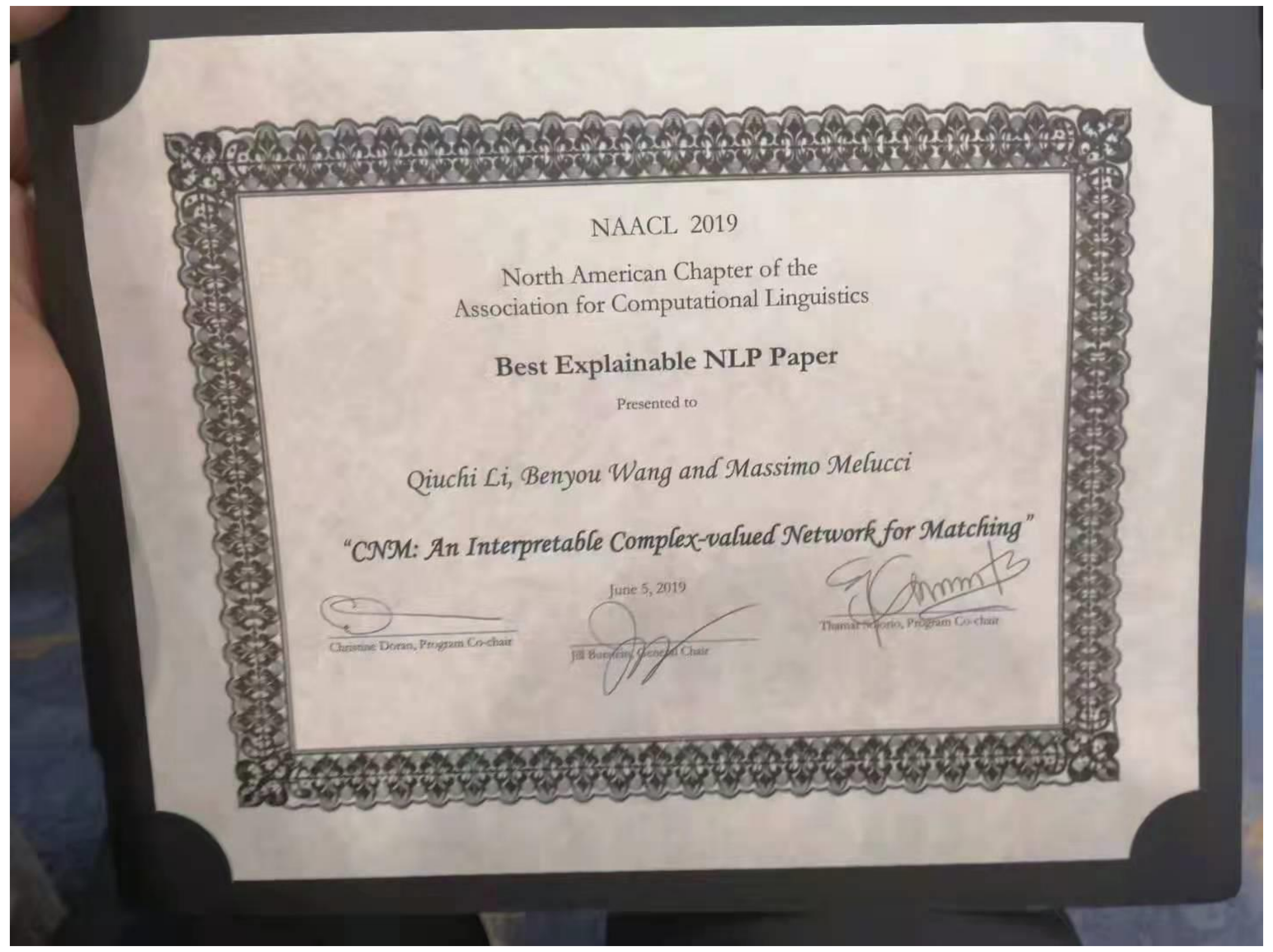



Galileo Galilei

the "father of **modern physics**"
the "father of the scientific method"
the "father of modern science"

Alumni of University of Padua

Awards and honour

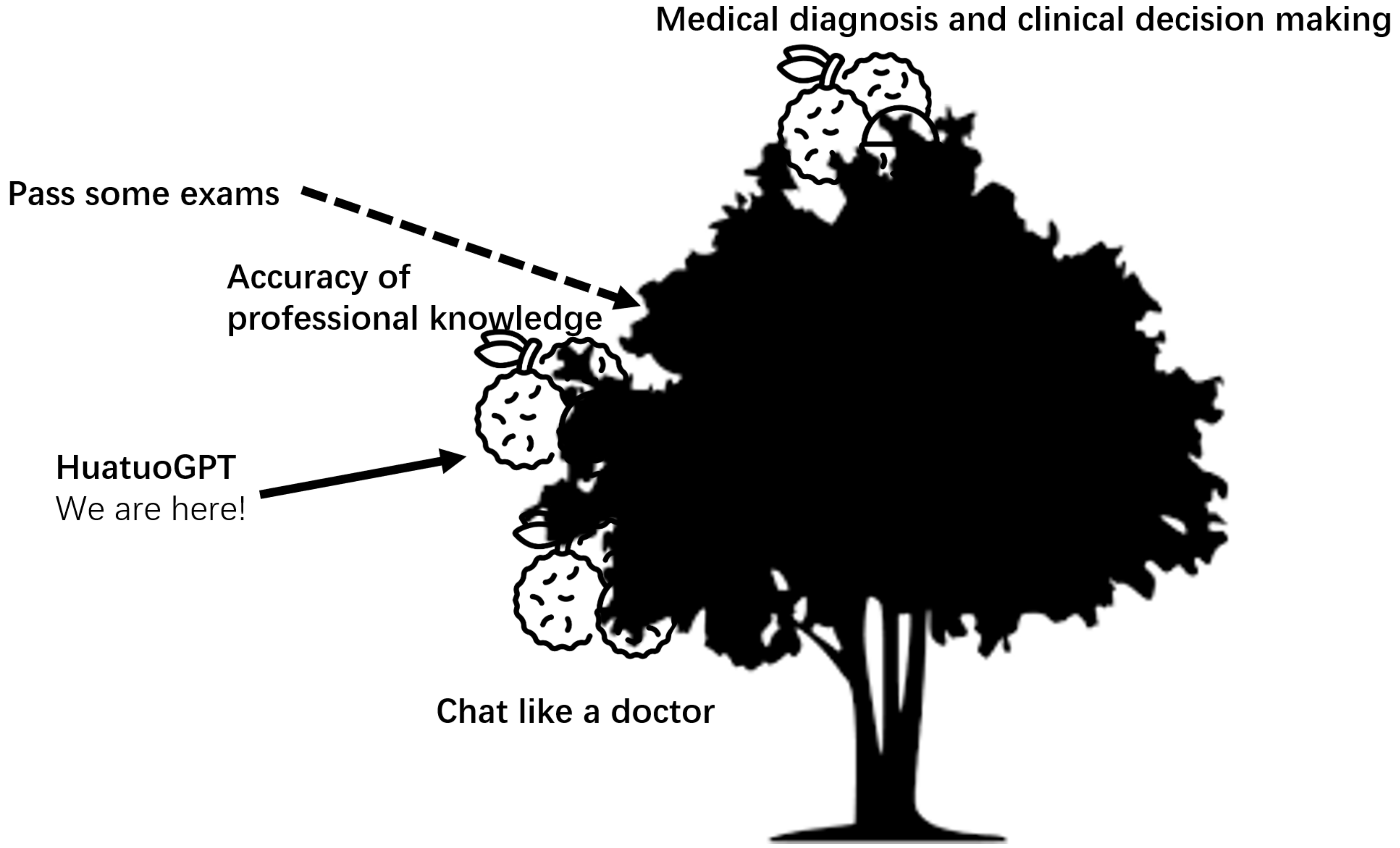


- **NLPCC 2022** Best Paper
- **ACM SIGIR** 2017 Best paper honourable mention. https://sigir.org/awards/best-paper-awards/
- **NAACL** 2019 best explainable NLP paper. https://naacl2019.org/blog/best-papers/
- EU Marie Curry researcher fellowship
- Huawei Spark award (华为火花奖)

Large Language models(LLMs)

- Large Language model (LLMs)
  - Democratizing ChatGPT (Phoenix, 2k GitHub Stars)
    - Efficiency (e.g., Modularizing LLMs)
    - Improving Reasoning ability
  - Applications
    - Multi-modal LLMs
    - Multilingual LLMs (e.g., Chinese and Arabic)
    - Tools and plugins
    - In-campus deployment (https://phoenix.cuhk.edu.cn)

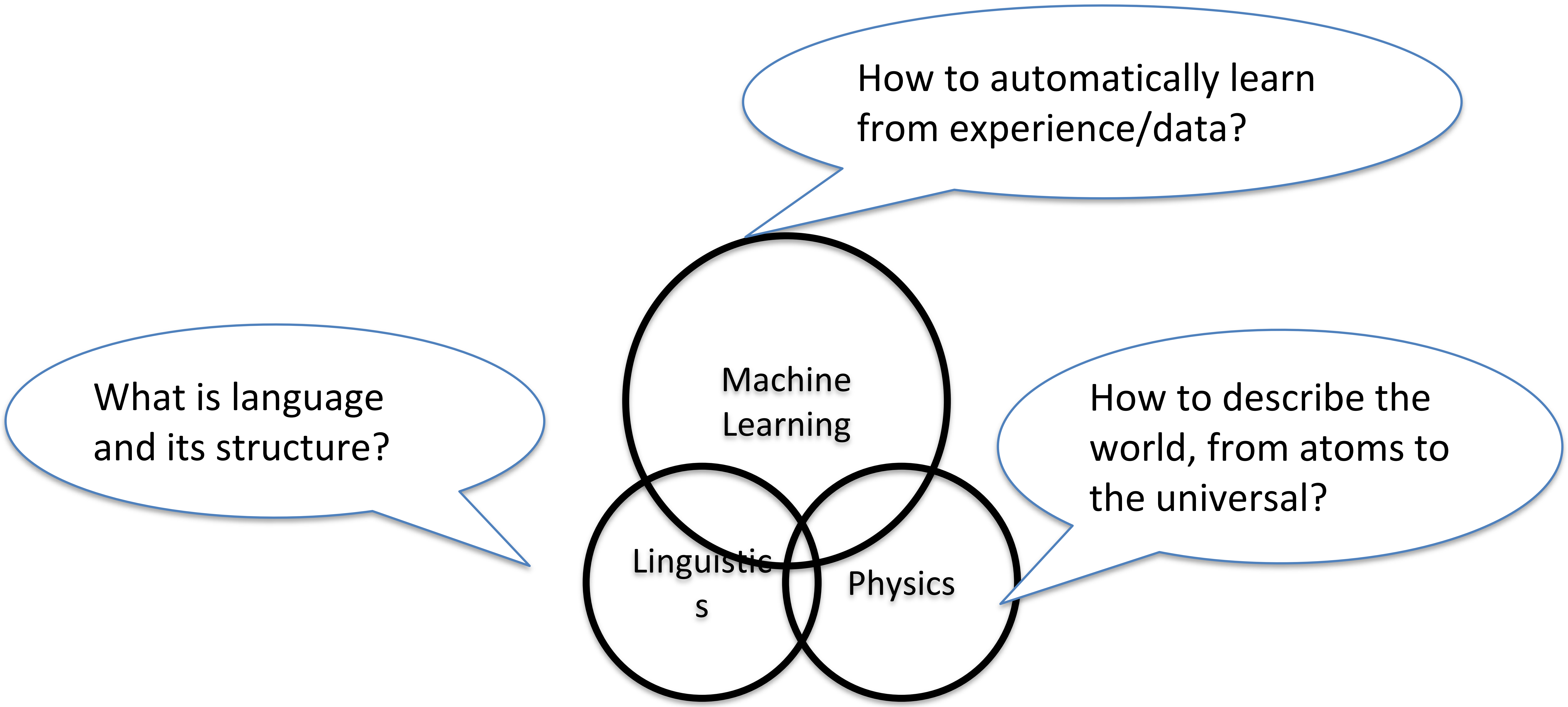
- LLMs for Medicine (e.g. HuatuoGPT)
  - Biomedical knowledge injection
  - Benchmarking
  - Chaim of Diagnosis
  - Doctors-in-the-loop



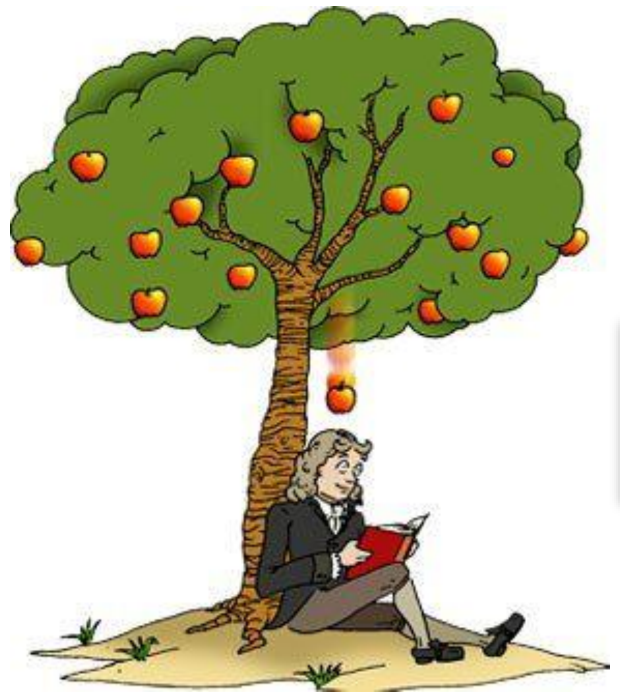
We are picking the **low-hanging lychee**

Contents

- **On the motivations of quantum theory in NLP**
- Overview of the research
  - **Interpretability**:
    - Modeling words as **particles** for better interpretability
    - Modeling words as **waves** to encode order
  - **Efficiency**: Network Compression using tensor networks
  - **Potential**: Quantum computing equipped language models.


How to automatically learn from experience/data?
What is language and its structure?
Machine Learning
How to describe the world, from atoms to the universal?
Linguistic
s
Physics

Physics and Machine Learning

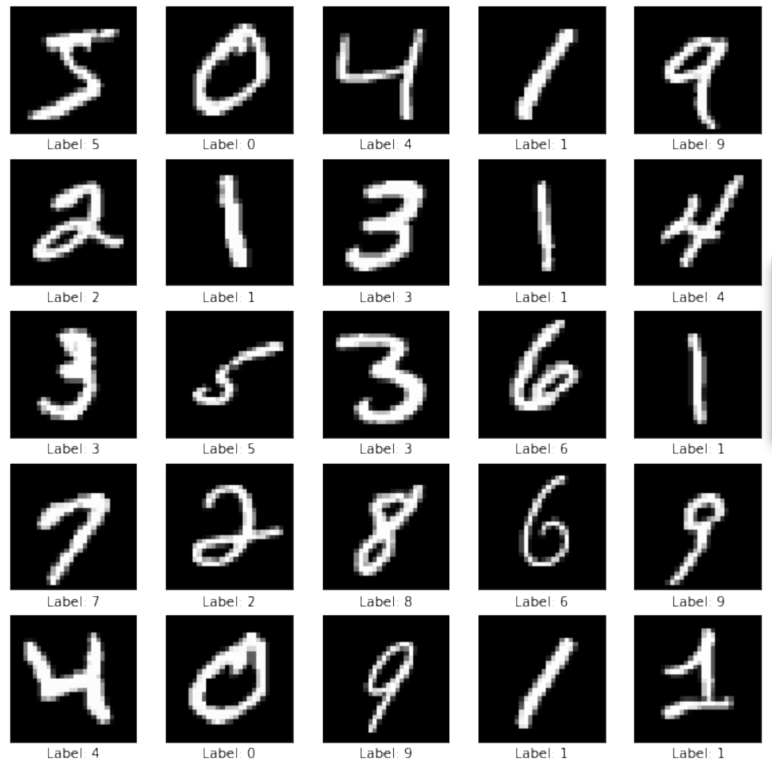


Induction

$$F = G\frac{m_1 m_2}{r^2}$$

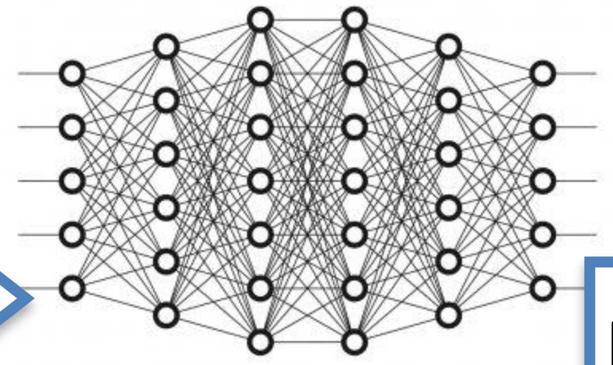
universal gravitation

Deduction



training set

Induction



$$y = f(x|\theta)$$

Ad hoc neural networks

Deduction

test set

Picture from pinimg.com

Are Physics and ML Complementary?

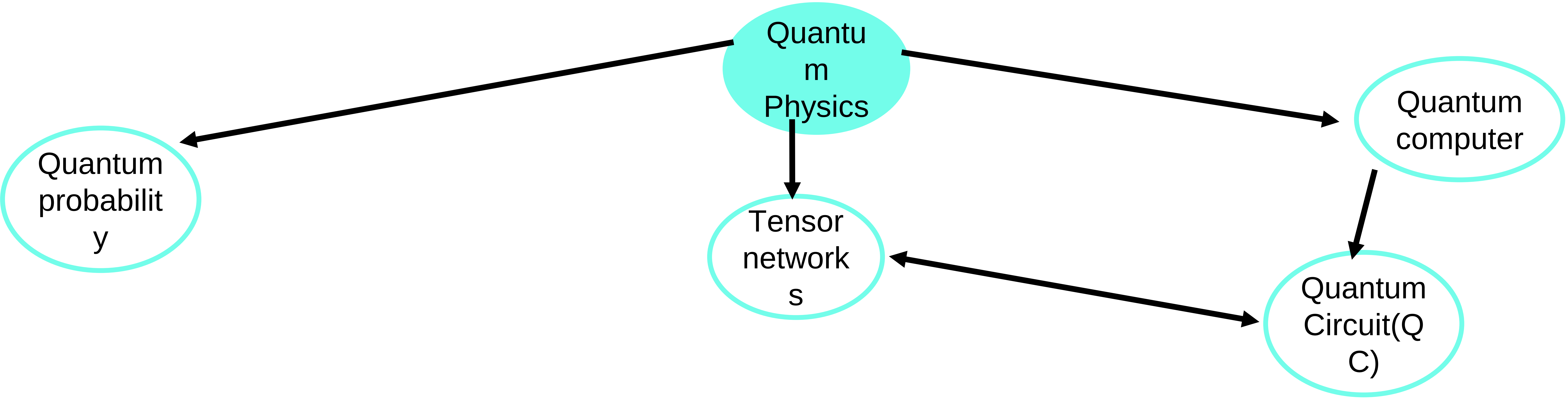
	generalization	Data hunger	Interpertation	Some examples
Physics	Good	No	Yes	Efficiency (quantum computing)
ML	Not good	Yes	No	Effectiveness (Pre-trained language models in NLP)

Curse of dimensionality: Both of them have to deal with big tensors:
tensor network in physics for many body problems VS. Large-scaled pre-trained language models

The motivations to use QT for NLP

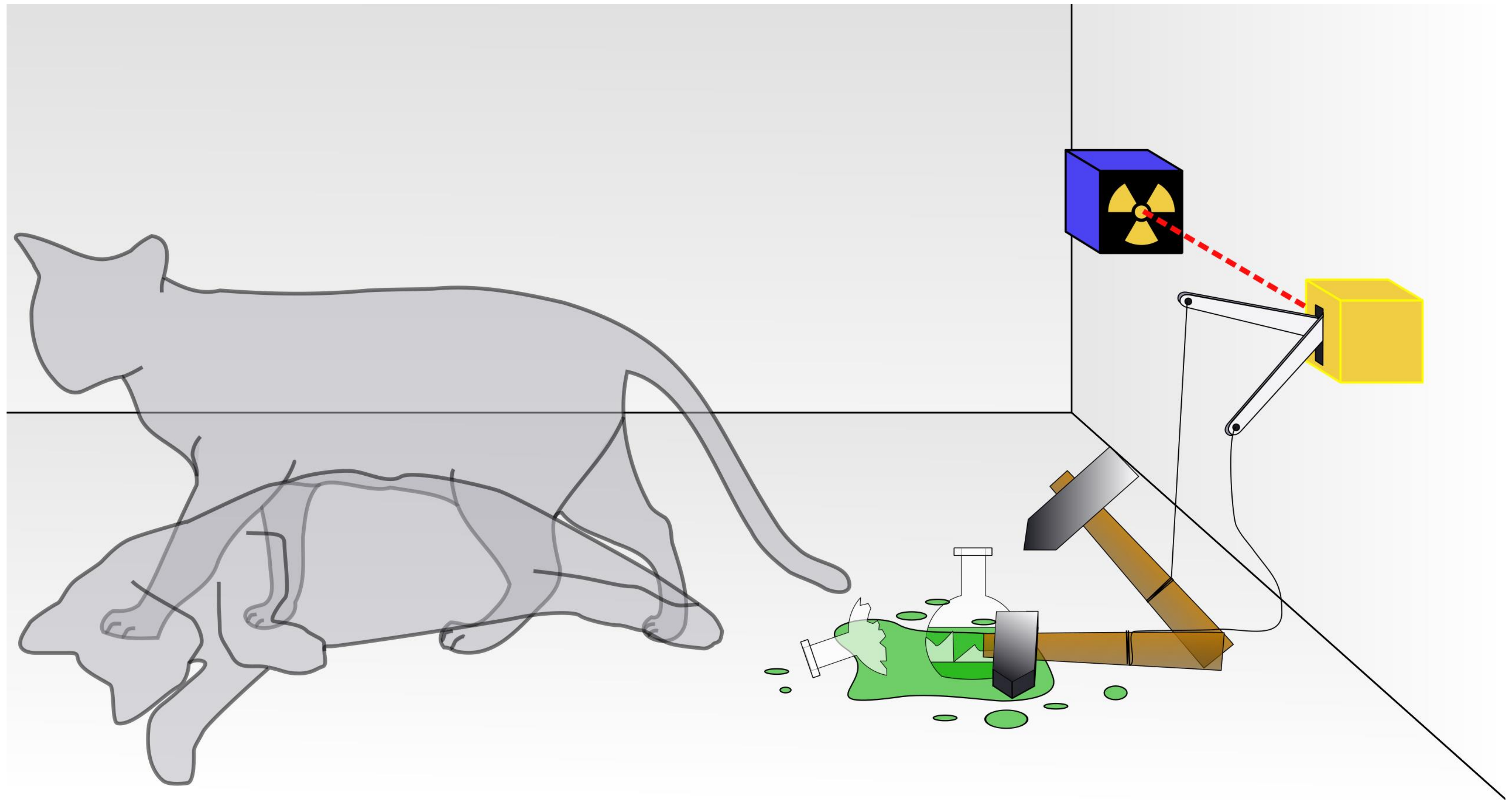
- Neural NLP lacks **interpretation** but quantum world is well-described.
- They both meet the **curse of dimensionality**
- There is bottleneck for increasing pre-trained language models, while quantum computing may helps

About quantum physics



- **Quantum probability** mathematically describes particle
- **Tensor networks** describe system with many entangled particles (a.k.a, Quantum Many-body problem)
- **Quantum computing** makes use of many-particles entangled system for computation by quantum circuits

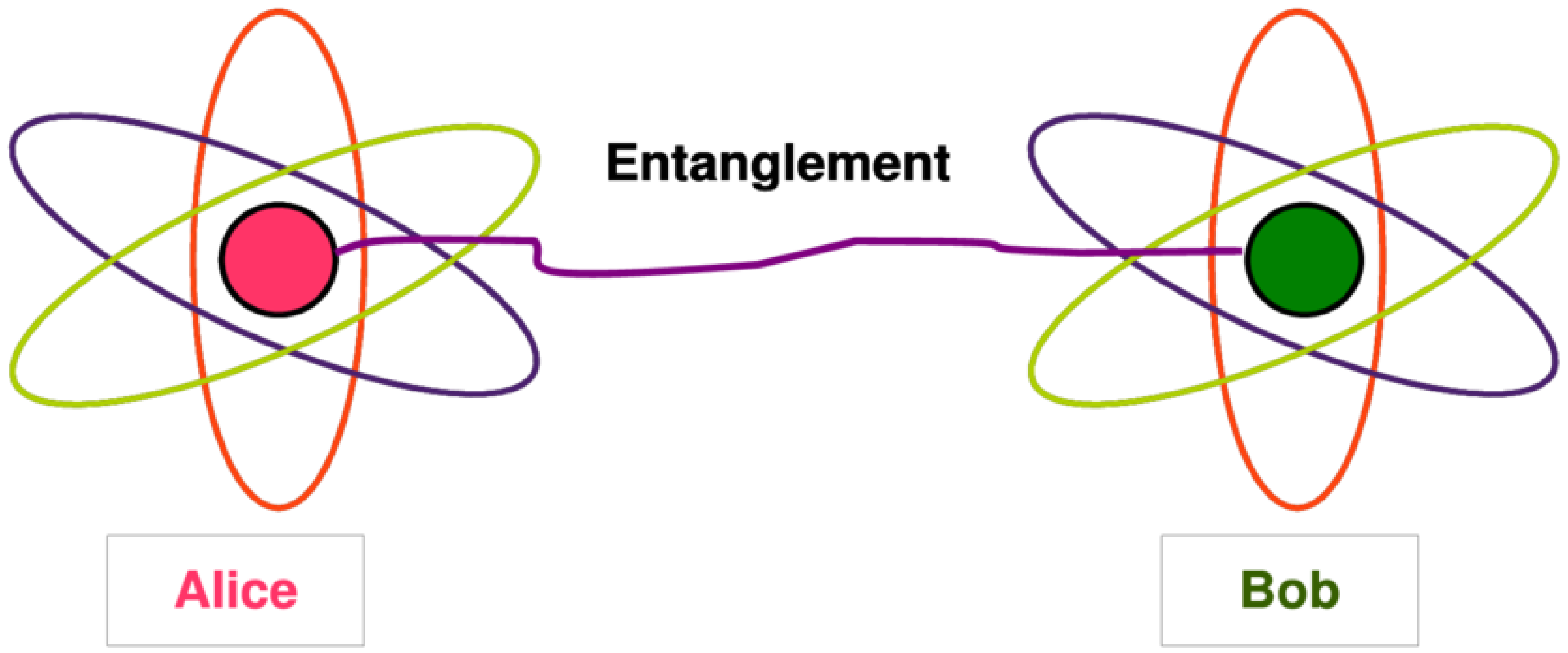
Basic principle in micro particles: Superposition



Superposition: a hypothetical cat may be considered simultaneously both alive and dead as a result of its fate being linked to a random subatomic event that may or may not occur.

This is described by **quantum probability**, a.k.a, Copenhagen interpretation.

Between many particles: Entanglement

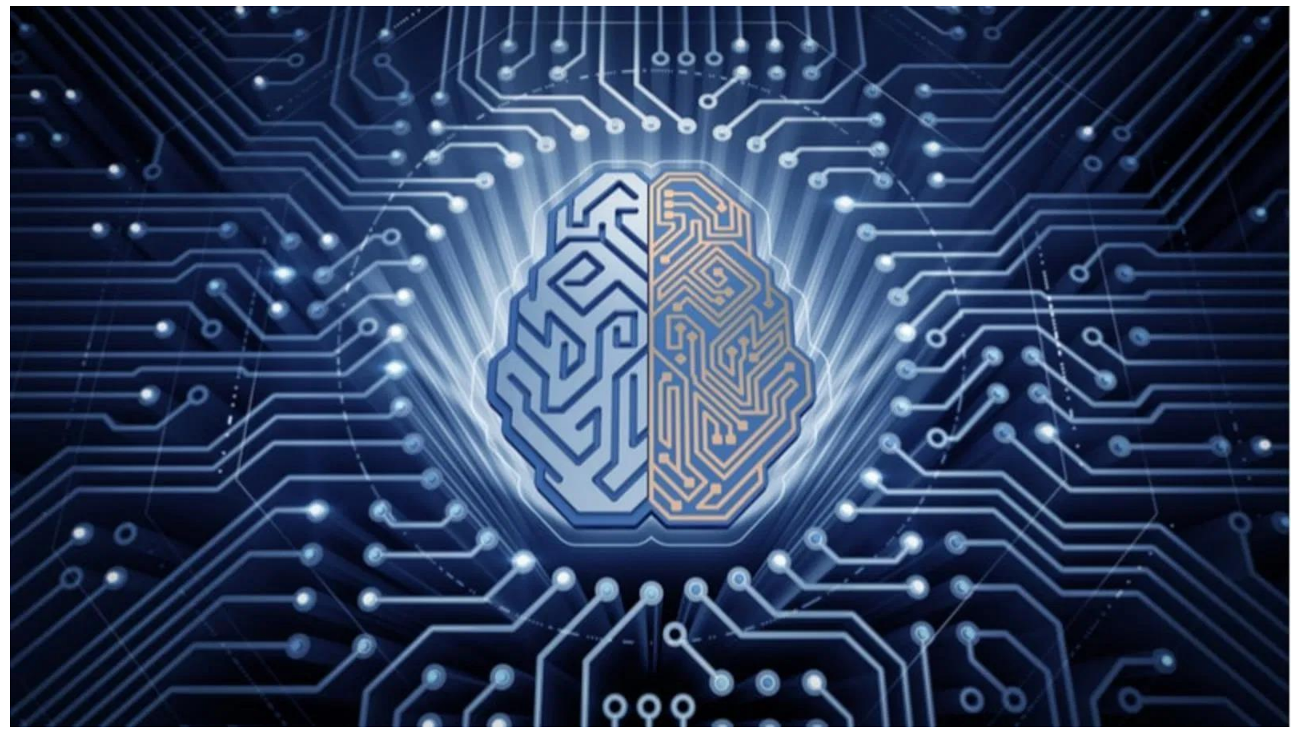


Suppose a particle is in a superposition state between |1> and |0>
the state of N particles will be in space of 2^N, resulting in **curse of dimensionality**
To efficiently describe such state, **Tensor network** is designed to approximate such high-dimension state.

Quantum theory **outside** Physics

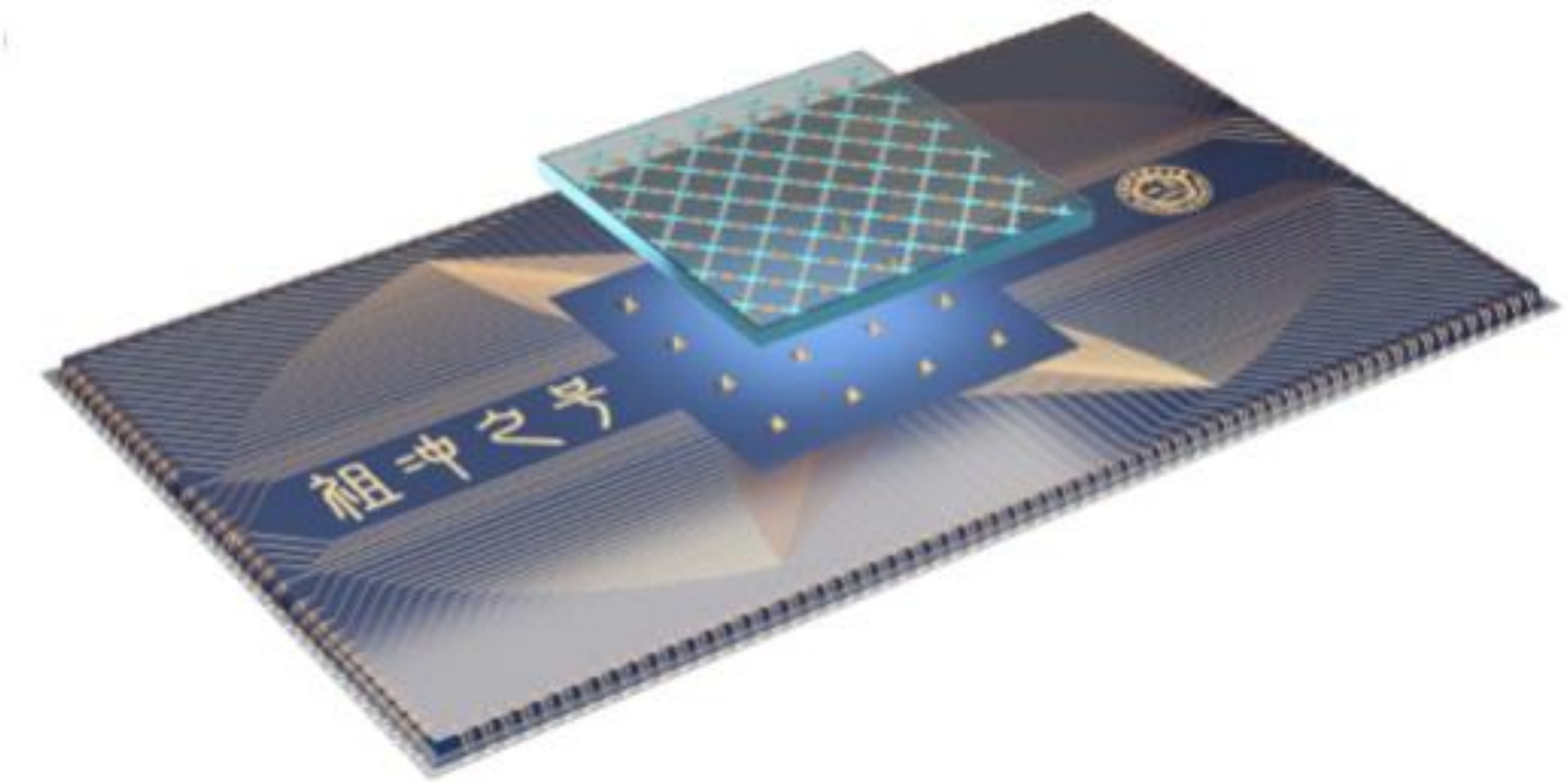
Using quantum ways to process information

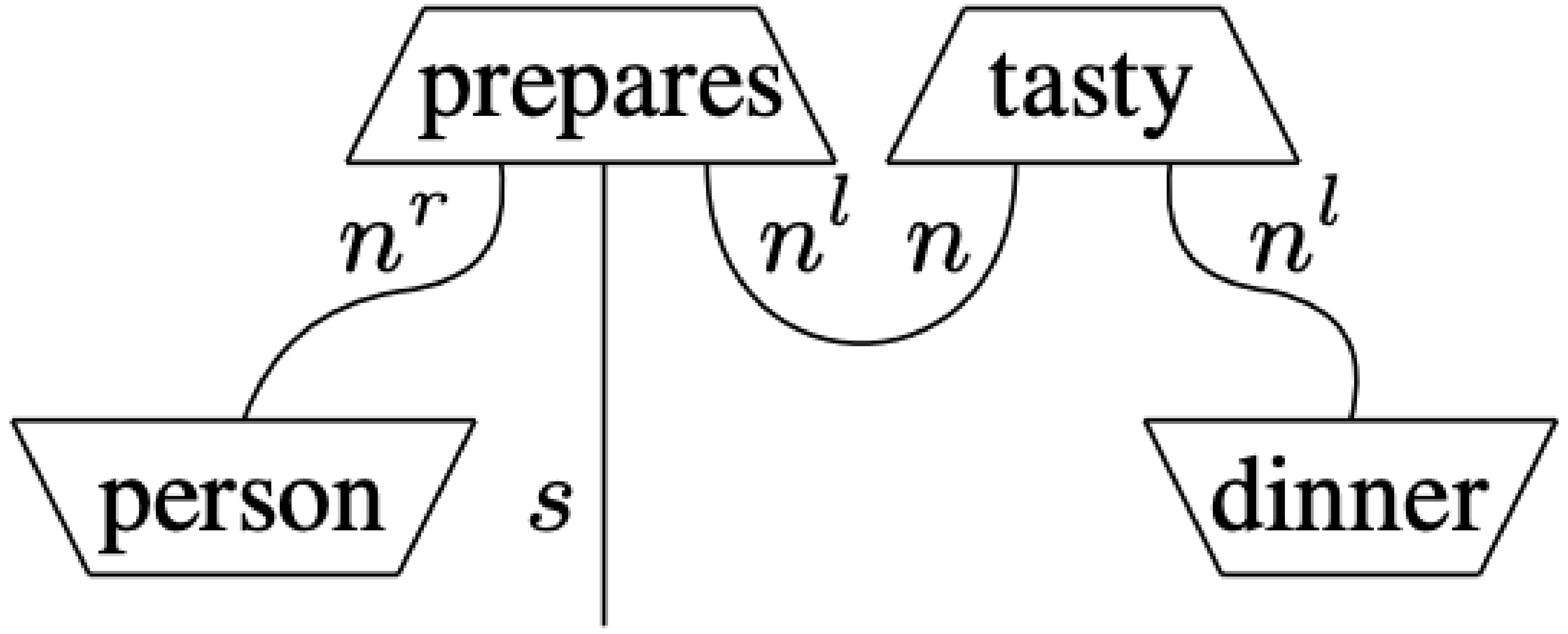
- Quantum computing
  - [Michael A. Nielsen, Isaac L. Chuang. 2011. Quantum Computation and Quantum Information, 10th edition. Cambridge University Press]
  - Arute .et.al. Quantum supremacy using a programmable superconducting processor. Nature. 23 October 2019.
- Social science and cognition science
  - [Jerome R. Busemeyer and Peter D. Bruza. 2013. Quantum Models of Cognition and Decision. Cambridge University Press]
  - [E. Haven and A. Khrennikov. 2013. Quantum Social Science. Cambridge University Press.]
- **Information retrieval**
  - [Van Rijsbergen. 2004. The geometry of information retrieval. Cambridge University Press.]
  - [Massimo Melucci. 2016. Introduction to information retrieval and quantum mechanics. Springer Berlin Heidelberg.]



- *Quantum IR can formulate the different IR models (**logic, vector, probabilistic**, etc.) in a unified framework.*
  - Quantum IR does not rely on quantum computing/cognition, but share the same mathematical foundation to **probabilistically** describe the world

Quantum computing in NLP





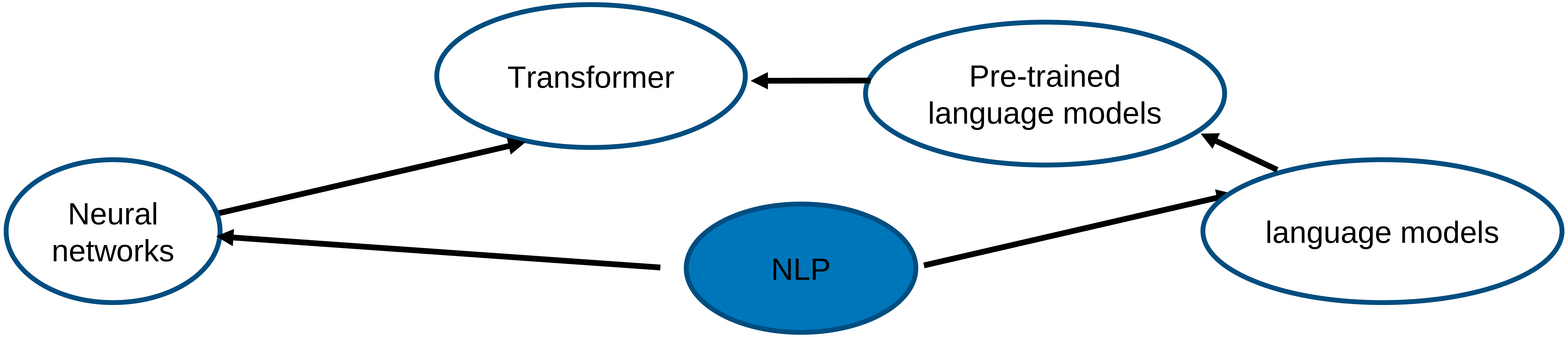
Run NLP tasks using quantum computer

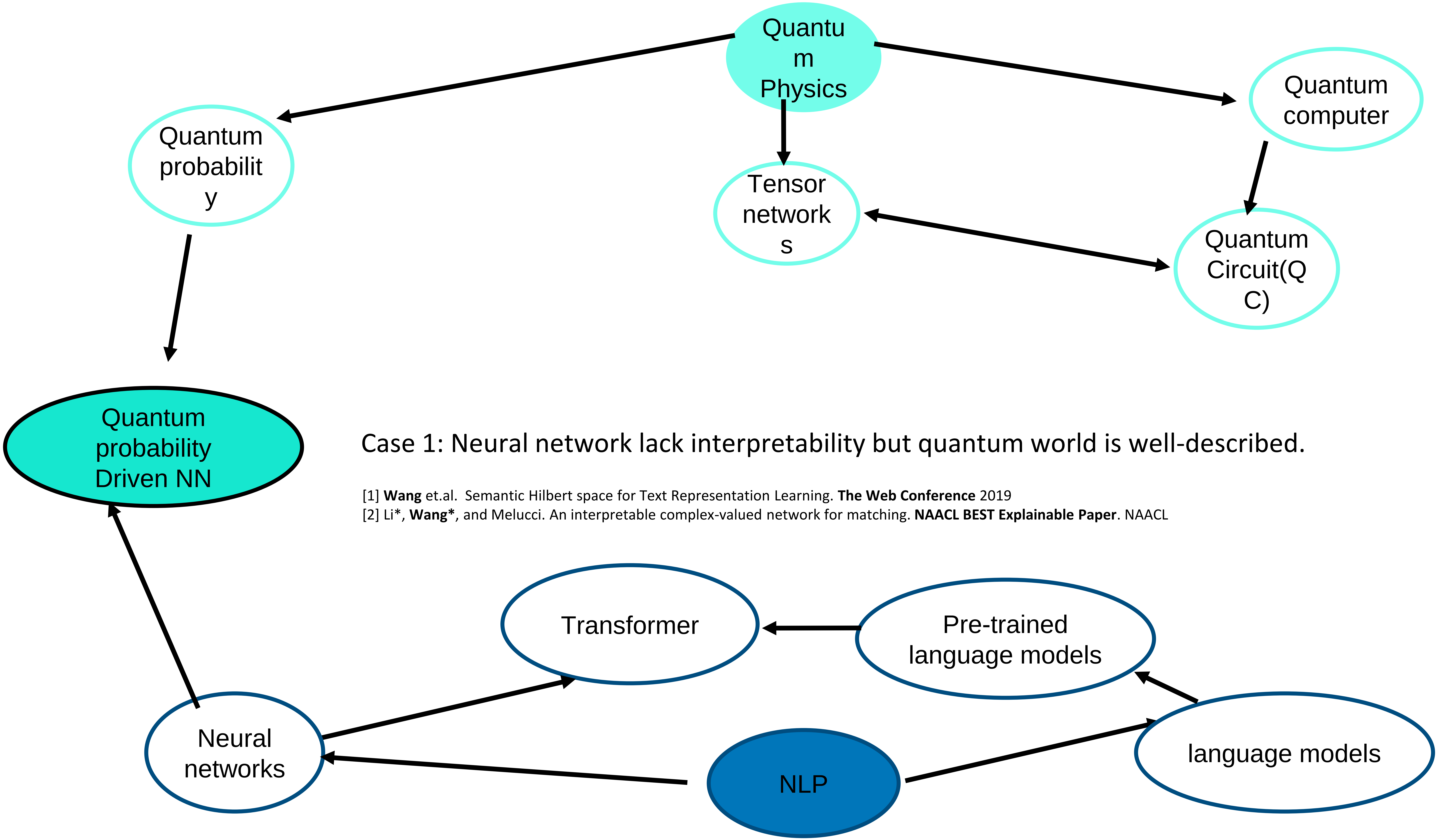
Wu et.al. Strong quantum computational advantage using a superconducting quantum processor. From Jianwei Pan’s group

Lorenz, et.al. QNLP in Practice: Running Compositional Models of Meaning on a Quantum Computer. From Bob Coecke’s group in University of Oxford and Cambridge Quantum Computing.

BIG troubles in NLP

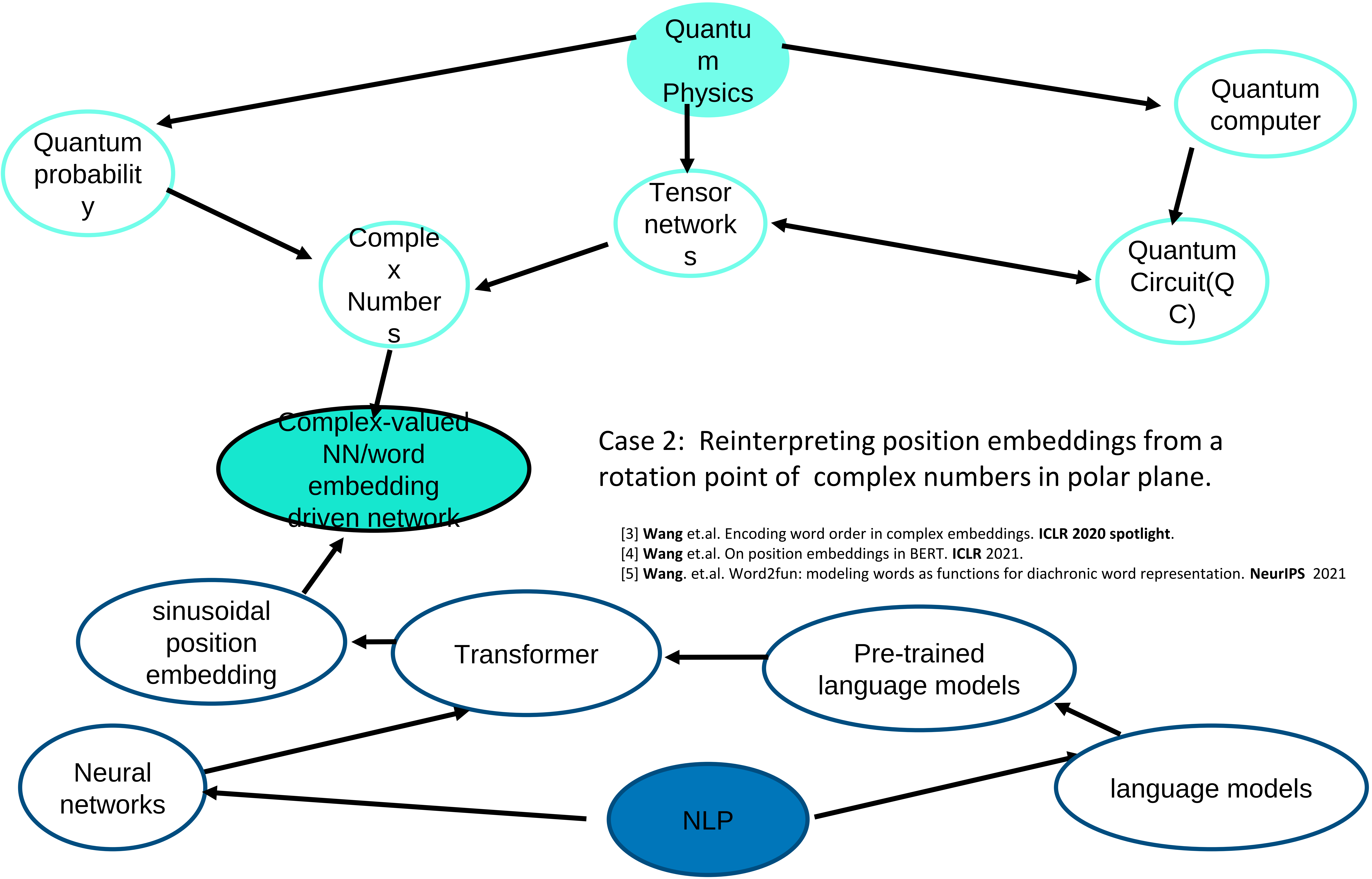
- Theoretical level: Neural networks lack **interpretation**
  - *For the overall architecture itself and components like position embeddings, etc.*
- Technical level: Large-scale Pre-trained language models is hard for **deployment** due its big size
- In the future: how to boost the **capacity** of pre-trained language models in the future?





Case 1: Neural network lack interpretability but quantum world is well-described.

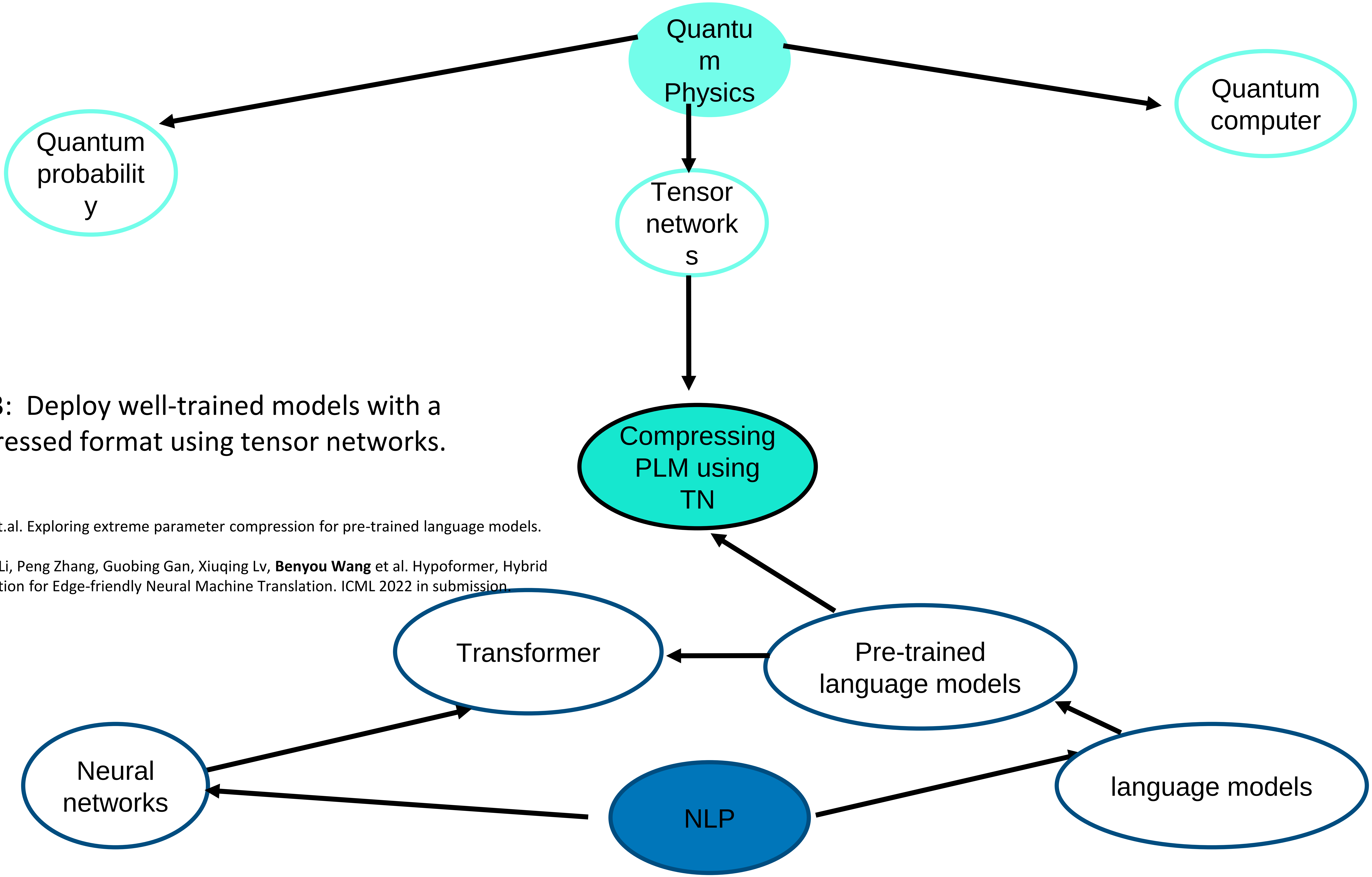
[1] **Wang** et.al. Semantic Hilbert space for Text Representation Learning. **The Web Conference** 2019
[2] Li*, **Wang***, and Melucci. An interpretable complex-valued network for matching. **NAACL BEST Explainable Paper**. NAACL



Case 2: Reinterpreting position embeddings from a rotation point of complex numbers in polar plane.

[3] **Wang** et.al. Encoding word order in complex embeddings. **ICLR 2020 spotlight**.
[4] **Wang** et.al. On position embeddings in BERT. **ICLR** 2021.
[5] **Wang**. et.al. Word2fun: modeling words as functions for diachronic word representation. **NeurIPS** 2021

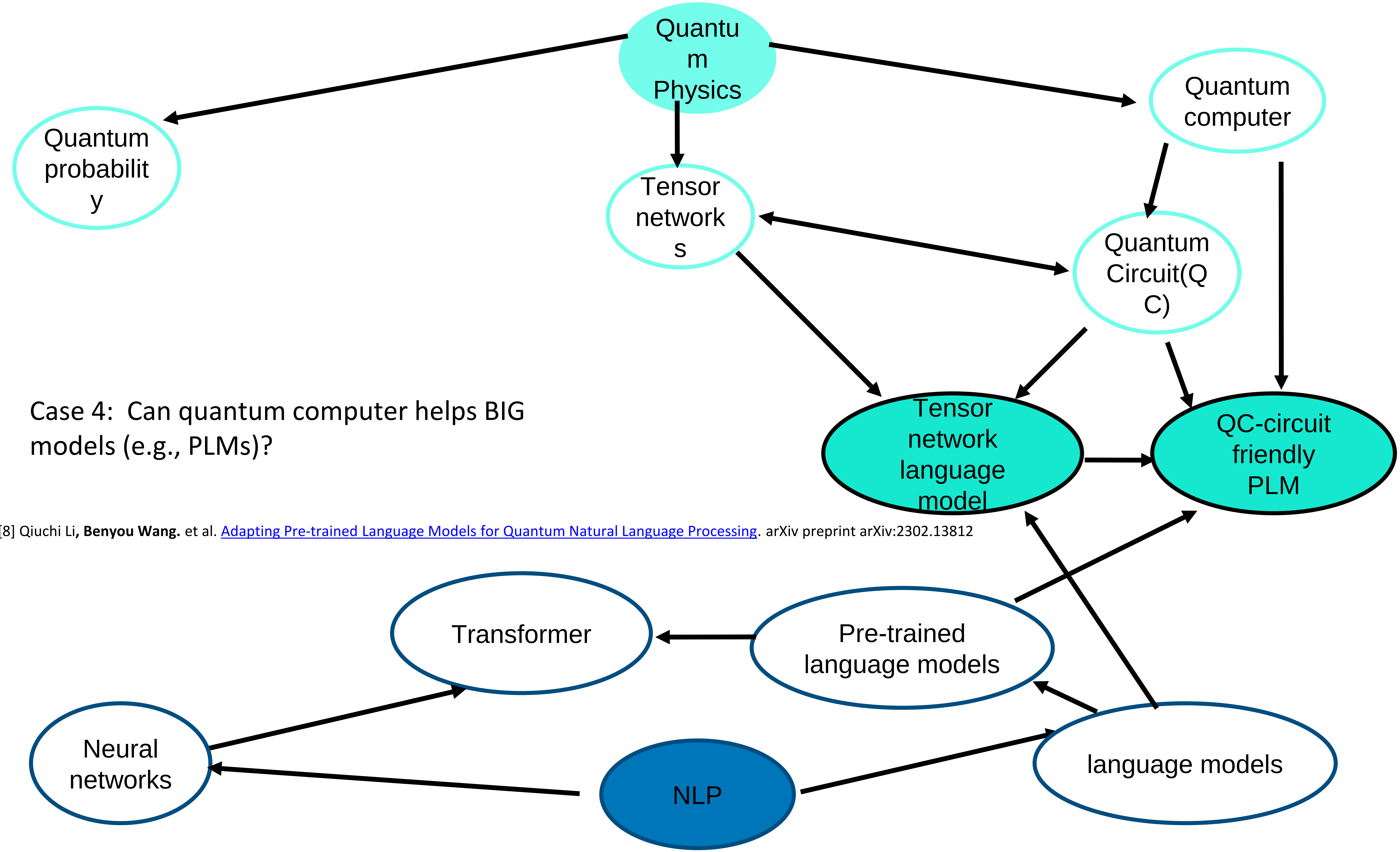
Case 3: Deploy well-trained models with a compressed format using tensor networks.



[6] **Wang** et.al. Exploring extreme parameter compression for pre-trained language models. ICLR 2022.

[7] Sunzhu Li, Peng Zhang, Guobing Gan, Xiuqing Lv, **Benyou Wang** et al. Hypoformer, Hybrid Decomposition for Edge-friendly Neural Machine Translation. ICML 2022 in submission.

Case 4: Can quantum computer helps BIG models (e.g., PLMs)?



[8] Qiuchi Li, **Benyou Wang.** et al. [Adapting Pre-trained Language Models for Quantum Natural Language Processing](). arXiv preprint arXiv:2302.13812

Quantu m Physics
Quantum probabilit y
Tensor network s
Quantum computer
Comple x Number s
Quantum Circuit(Q C)
Quantum probability Driven NN
Complex-valued NN/word embedding driven network
Compressing PLM using TN
Tensor network language model
QC-circuit friendly PLM
sinusoidal position embedding
Transformer
Pre-trained language models
Neural networks
NLP
language models

Contents

- On the motivations of quantum theory in NLP
- Overview of the research
  - **Interpretability**:
    - **Modeling words as particles for better interpretability**
    - Modeling words as **waves** to encode order
  - **Efficiency**: Network Compression using Tensor Networks
  - **Potential**: Quantum computing equipped language models.

Particle-wave duality for text

For a static view, any object is a particle
For a dynamic view, any object will be wave, e.g. in temporal and spatial dimensions

- Model words as **particles** for better interpretability [1,2]
    - *quantum probability driven networks*

- Model words as **waves** to encode its temporal and spacial context
    - ***spatial waves**: position embeddings explained [3,4]*
    - ***temporal waves**: dynamic word embedding [5]*

[1] **Wang** et.al. Semantic Hilbert space for Text Representation Learning. **The Web Conference** 2019
[2] Li*, **Wang***, and Melucci. An interpretable complex-valued network for matching. **NAACL BEST Explainable Paper**. NAACL 2019
[3] **Wang** et.al. Encoding word order in complex embeddings. **ICLR 2020 spotlight**.
[4] **Wang** et.al. On position embeddings in BERT. **ICLR** 2021.
[5] **Wang**. et.al. Word2fun: modeling words as functions for diachronic word representation. NeurIPS 2021

https://www.youtube.com/watch?v=Q_h4IoPJXZw

from **localist** to **distributed** representation

localist representation

Concept	Representation
Small Red Car	[1 0 0 0 0 0 0 0]
Large Blue SUV	[0 1 0 0 0 0 0 0]
Large Red SUV	[0 0 1 0 0 0 0 0]
Green Apple	[0 0 0 1 0 0 0 0]
Bumble Bee	[0 0 0 0 1 0 0 0]
Tall Building	[0 0 0 0 0 1 0 0]
Small Fish	[0 0 0 0 0 0 1 0]
Banana	[0 0 0 0 0 0 0 1]

distributed representation

Concept	Representation
Small Red Car	[0.555 0.761 0.243 0.812]
Large Blue SUV	[0.773 0.309 0.289 0.835]
Large Red SUV	[0.766 0.780 0.294 0.834]
Green Apple	[0.153 0.022 0.654 0.513]
Bumble Bee	[0.045 0.219 0.488 0.647]
Tall Building	[0.955 0.085 0.900 0.773]
Small Fish	[0.118 0.192 0.432 0.618]
Banana	[0.184 0.232 0.671 0.589]

probability theory based on set

A probability theory in vector space is needed

examples from https://www.districtdatalabs.com/nlp-research-lab-part-1-distributed-representations

Why not classical Probability Theory(PT)

- There exists **infinite** events in hidden states of NN, while, in classical PT, N elementary events lead to **finite** (e.g., 2^N) of events in total.

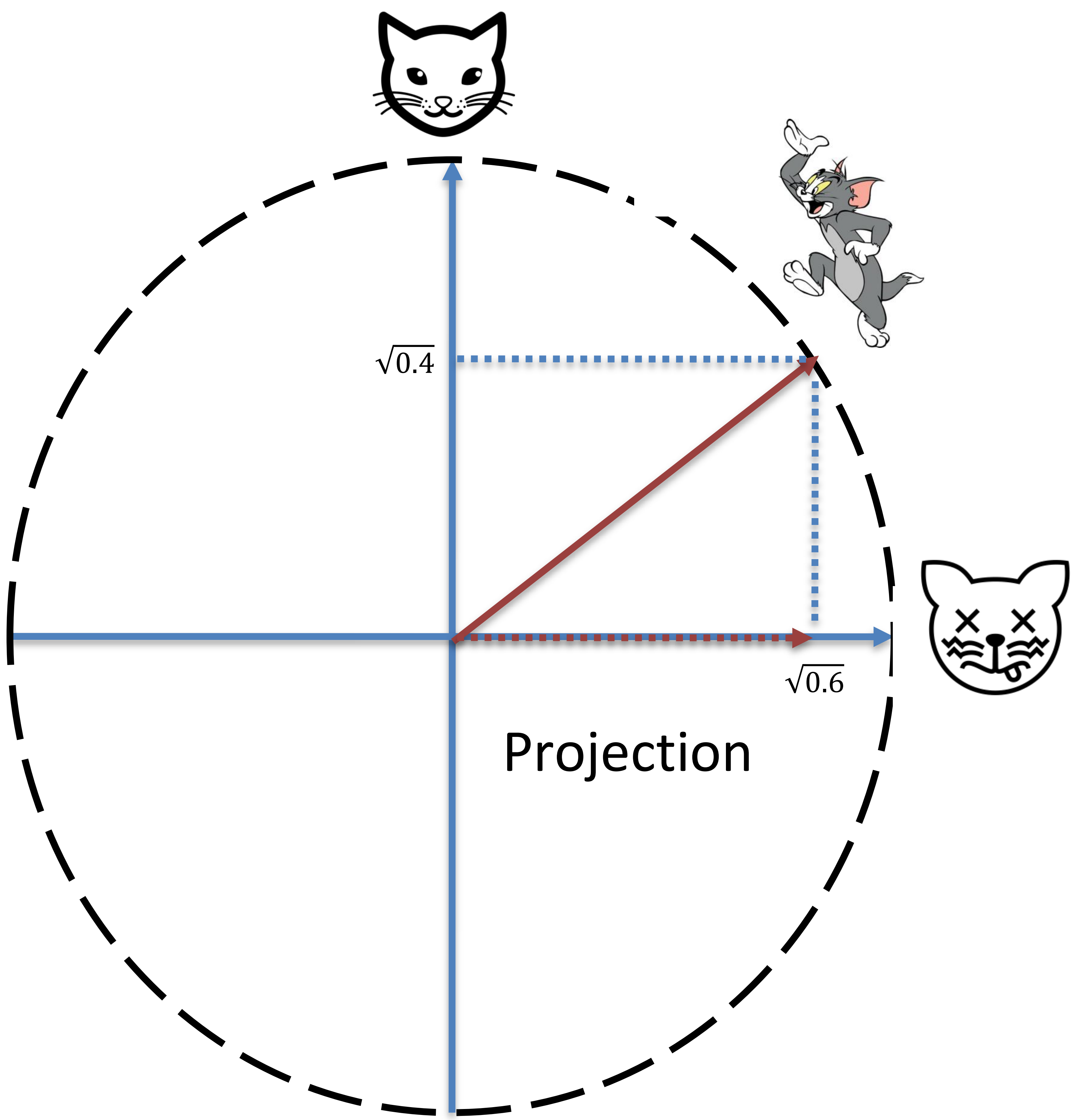
- What if we need a dummy state that is **between two elementary events** (*measure to which probability of a cat being accurately 50% dead and 50% alive*), especially we represent words in vector space。

$$< \vec{dog}, \vec{cat} > \stackrel{?}{=} 0$$

Superposition principle in QPT helps

Probability theory in vector spaces

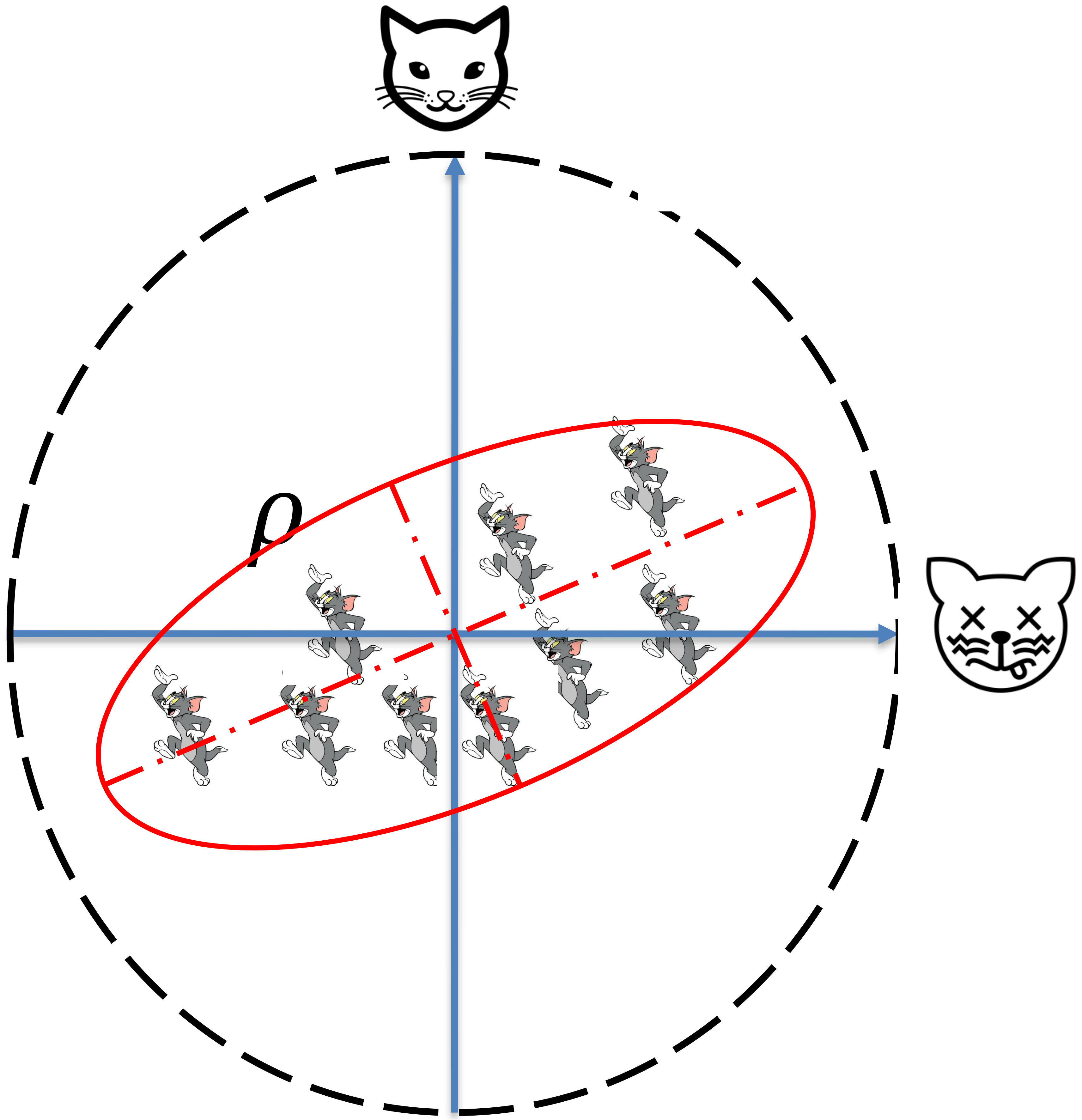
for single object



Square of the projection length denotes the probability

Probability theory in vector spaces

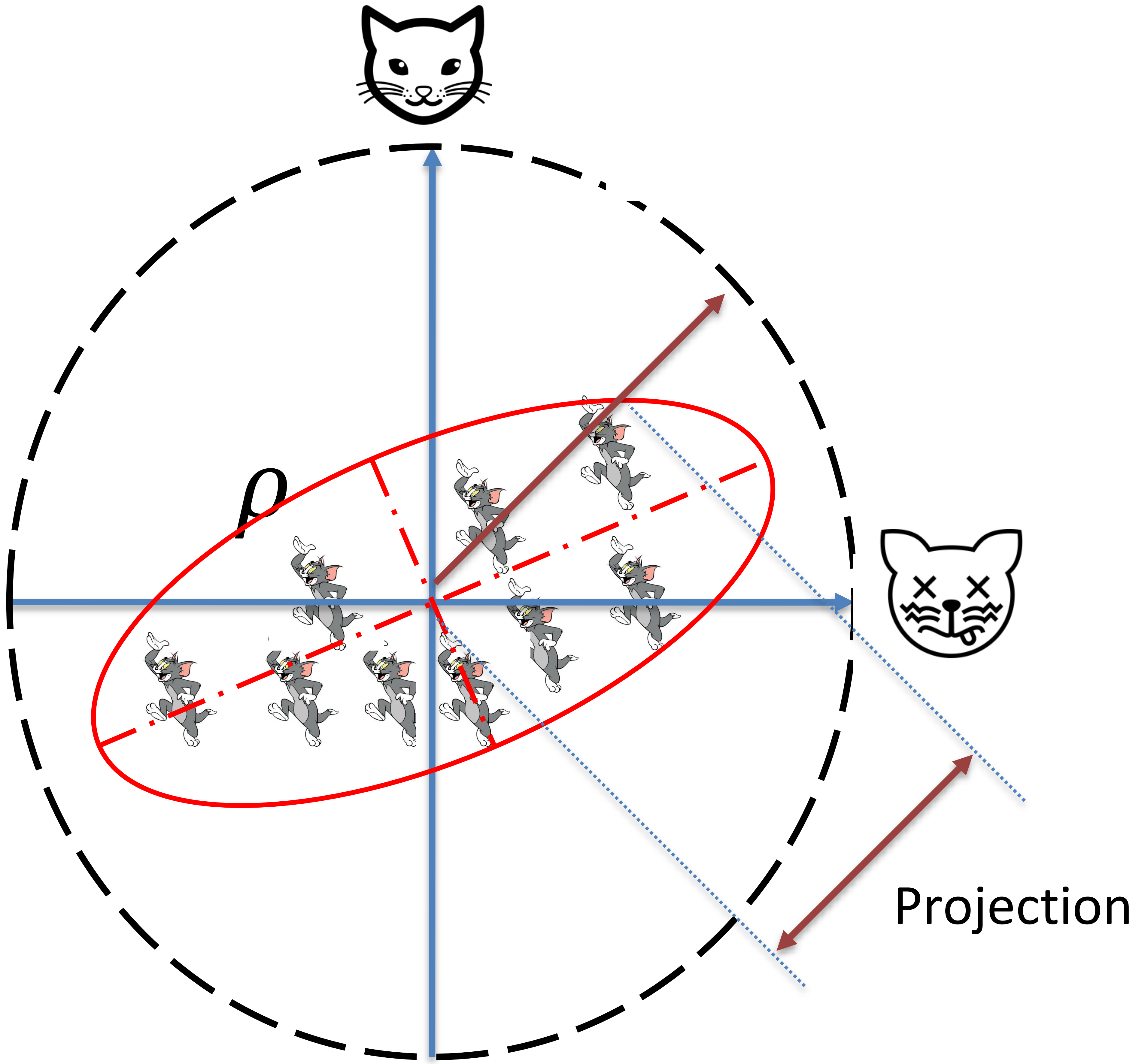
for many objects



Square of the projection length denotes the probability

Probability theory in vector spaces

for many objects



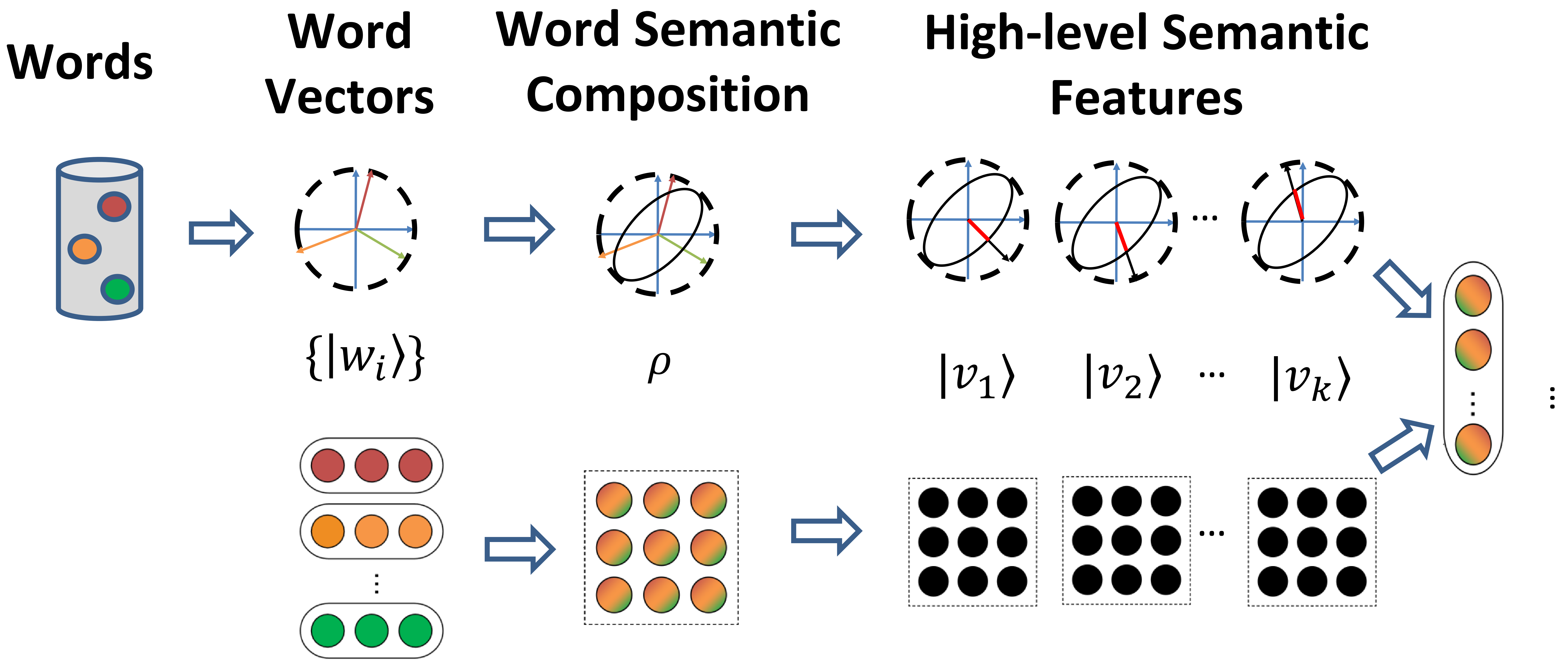
Square of the projection length denotes the probability

Superposition state in sememe space

- There exists a set of limited sememes form the language universal
  - ***Sememes** are the **minima atomic** linguistic units*
- Words as combinations of sememes:
  - *boy = MALE + CHILD + HUMAN*
  - *girl = FEMALE + CHILD + HUMAN*
  - .

Formulating combination of sememes as superposition

Semantic Hilbert Space



Physical meaning: **Pure States** **Mixed State** **Semantic Measurements**

Benyou Wang*, Qiuchi Li*, Massimo Melucci, and Dawei Song. *Semantic Hilbert Space for Text Representation Learning.* In *WWW2019*
Li, Qiuchi*, **Benyou Wang***, and Massimo Melucci. "CNM: An Interpretable Complex-valued Network for Matching." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4139-4148. 2019. NAACL 2019 best explainable paper

Formulation of Quantum probability driven network

Sememes: basic states

$$\{\mathbf{e} \in \mathbb{R}^D \,|\, < \mathbf{e}_i, \mathbf{e}_j >= \{ {1 if i = j \atop 0 if i \neq j} \}$$

Words: superposition states

$$\mathbf{w} = \sum_{j=1}^{D} z_j \mathbf{e}_j \in \mathbb{C}^D, z_j \in \mathbb{C}$$

N-gram: mixture system

$$\rho = \sum_{k=1}^{D} \lambda_k \mathbf{w_k}^T \mathbf{w_k} \in \mathbf{C}^{D \times D}$$

Semantic abstraction: measurement

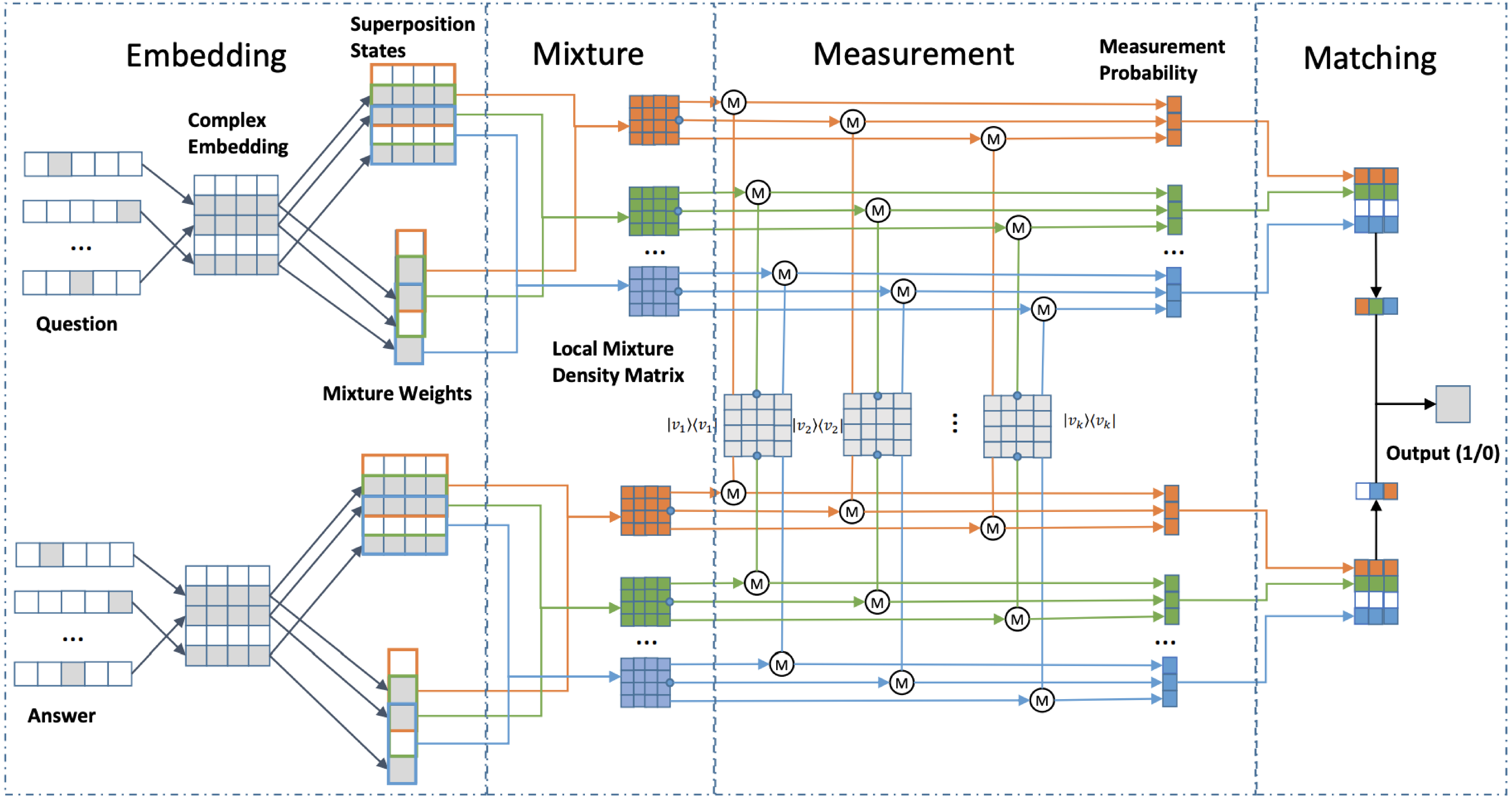
$$p = tr(\rho \mathbf{u}^T \mathbf{u}) \in \mathbb{R}, \mathbf{u} \in \mathbb{C}^D, < \mathbf{u}_i \mathbf{u}_j >= \{ {1 if i = j \atop 0 if i \neq j}$$

Sentence representation: probabilities

$$\mathbf{P} = \{p_1, \cdots p_d | 0 \leq p_i \leq 1, \sum p_i = 1\}$$

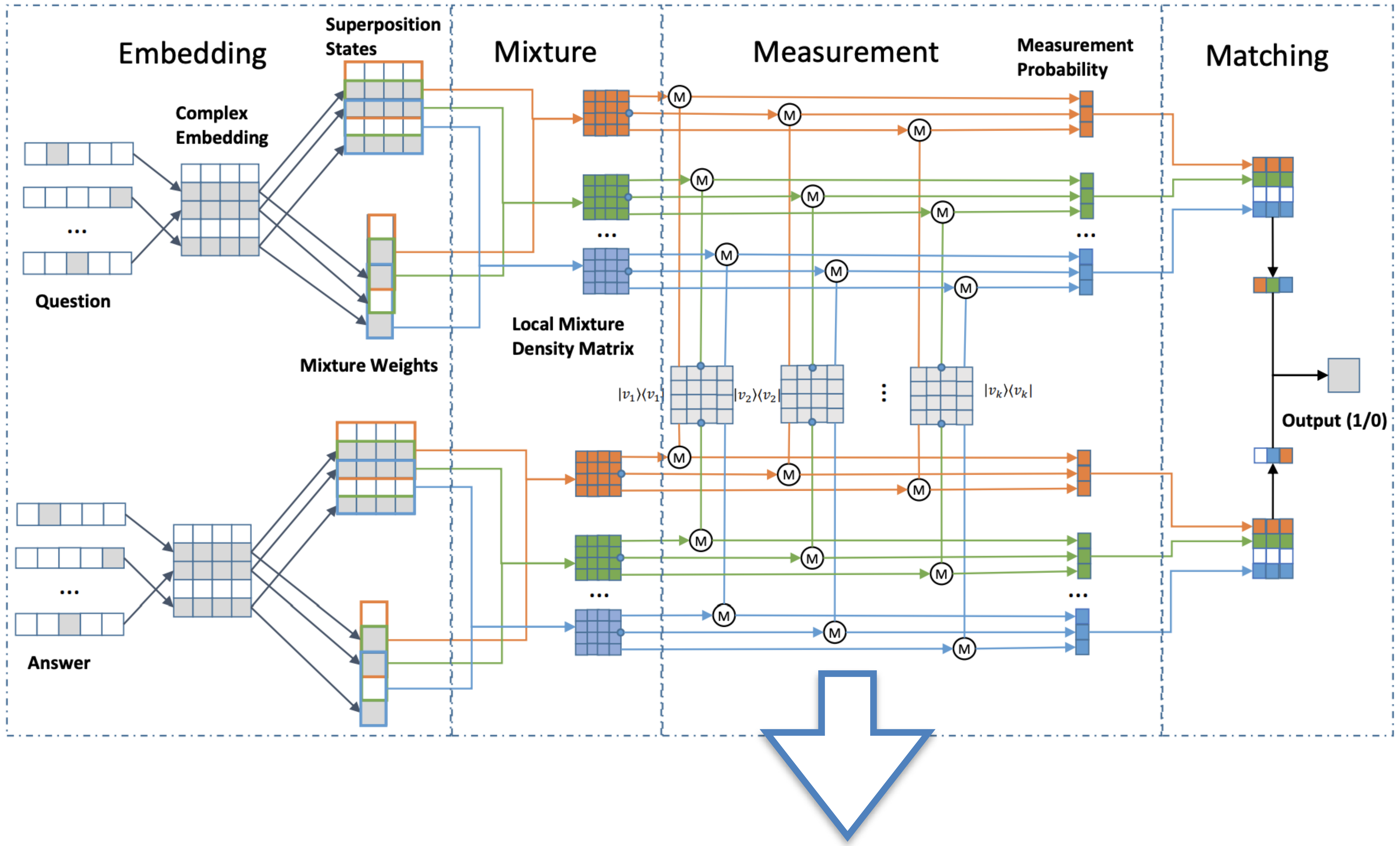
Components	DNN	QPDN in Semantic Hilbert Space
Sememe	-	one-hot basis vector / **basis state**
Word	real vector	unit complex-valued vector / **superposition state**
N-gram	real vector	complex-valued density matrix / **mixed system**
Abstraction	CNN/RNN	unit complex-valued vector / **measurement**
High-level representation	real vector	probabilities/ **measured probability**

Meaning of each components



Layer	Embedding	Mixture	Mearurement	Matching
Physical meaning	Pure state	mixed state	measurement	probability
Unitary?	unitary	unitary	unitary	unitary
Complex?	Complex	Complex	Complex	real
Neuron	Probability amplitude for each sememe	Probability correspondence for interaction between two sememes	Probability amplitude for each sememe	probability value for each measurement

Text explanation for measurement



	Selected neighborhood words for a measurement vector
1	andes, nagoya, inter-american, low-caste
2	cools, injection, boiling,adrift
3	andrews, paul, manson, bair
4	historically, 19th-century, genetic, hatchback
5	missile, exile, rebellion, darkness

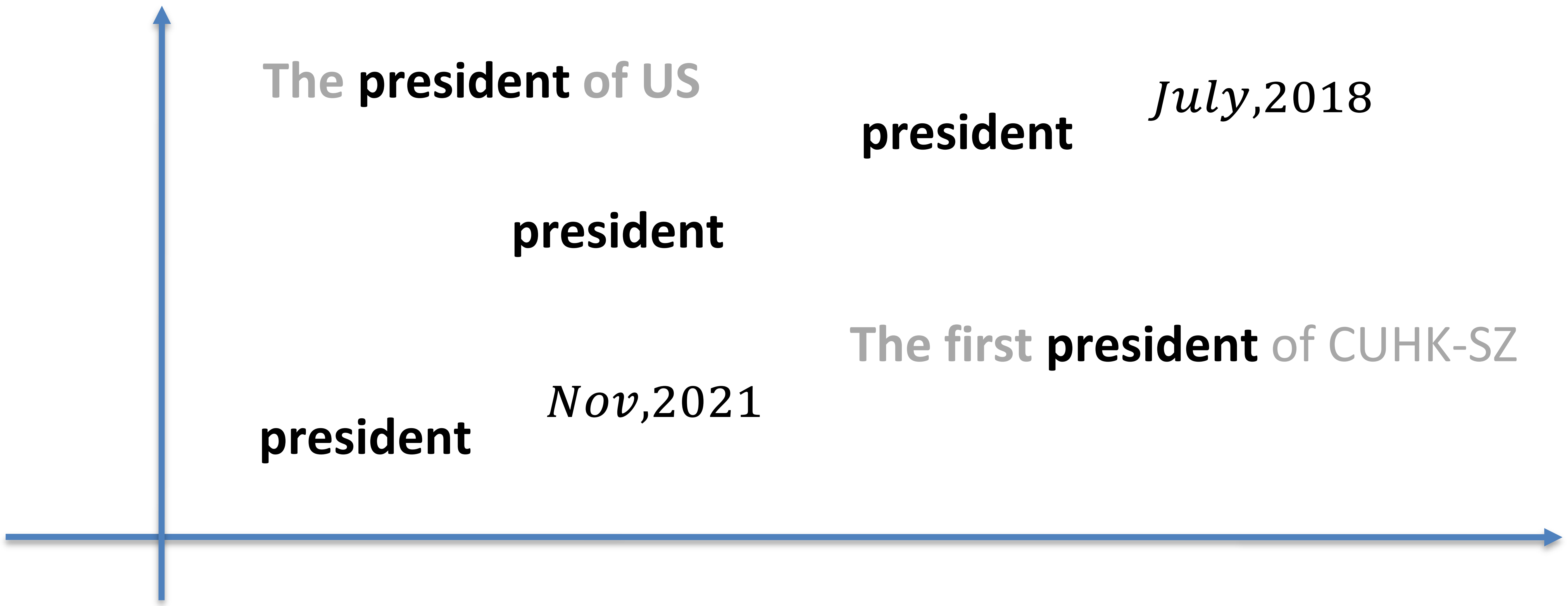
Contents

- On the motivations of quantum theory in NLP
- Overview of the research
  - **Interpretability**:
    - Modeling words as particles for better interpretability
    - **Modeling words as waves to encode order**
  - **Efficiency**: Network Compression using Tensor Networks
  - **Potential**: Quantum computing equipped language models.

From particles to waves

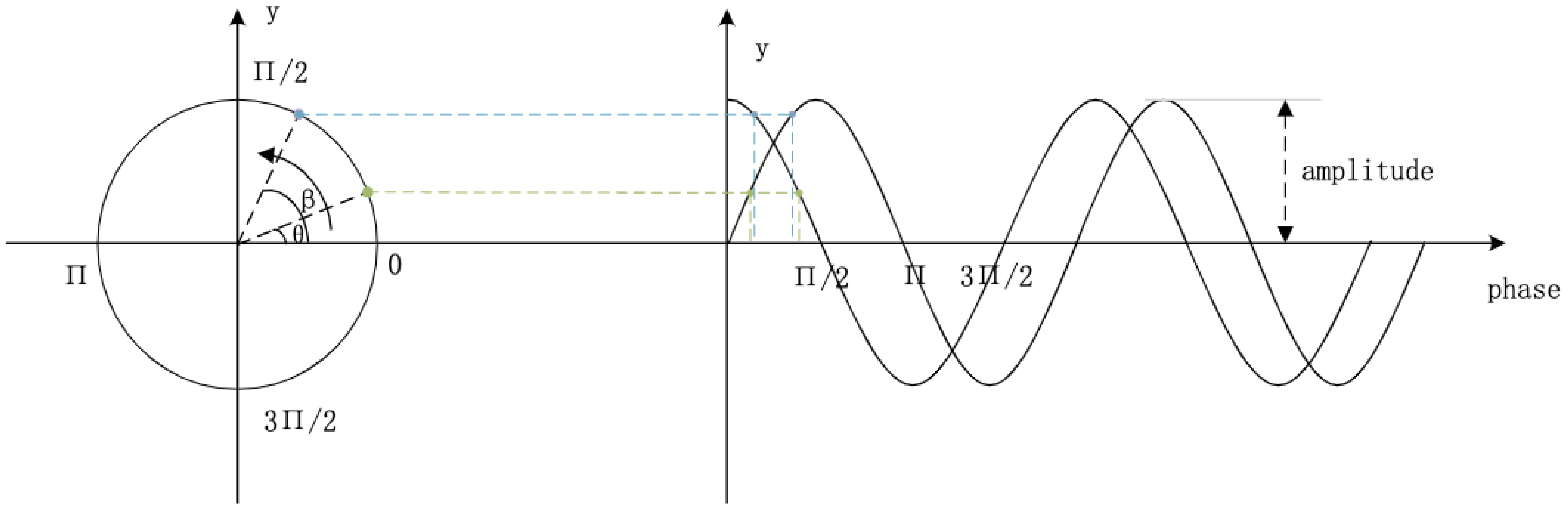
Word is generally like **static** particles without considering changing context.

However， *context might changes* ***spatially or temporally***



Word representation in different positions of a sentence or different time

Modeling order in waves



Sequential unfolding of complex numbers from polar plane

$$f(t) = e^{i\omega t} = \cos\omega t + i\sin\omega t$$

Word vectors to word functions

Extending embedding from a **vector** to a **continuous function** over variable the position (pos)

Technically, $f: (N, N) \rightarrow R^k$ To $\boldsymbol{f}: (\boldsymbol{N}) \rightarrow \boldsymbol{G}\{\boldsymbol{g}; \boldsymbol{g}: \boldsymbol{N} \rightarrow \boldsymbol{R}^{\boldsymbol{k}}\}$

Word index

position index

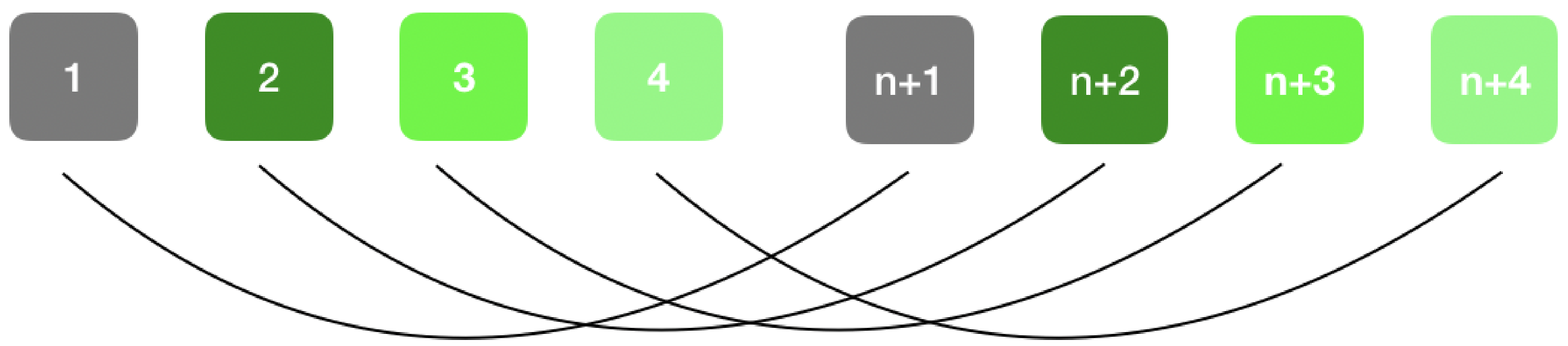
Now the question becomes *how to decide the function*

Desiderata for word functions

Now, for a specific word w, we have to get it embedding over all the positions, namely a function

$$\boldsymbol{g}_{\boldsymbol{w},d}: \boldsymbol{N} \rightarrow \boldsymbol{R}^{\boldsymbol{k}}$$

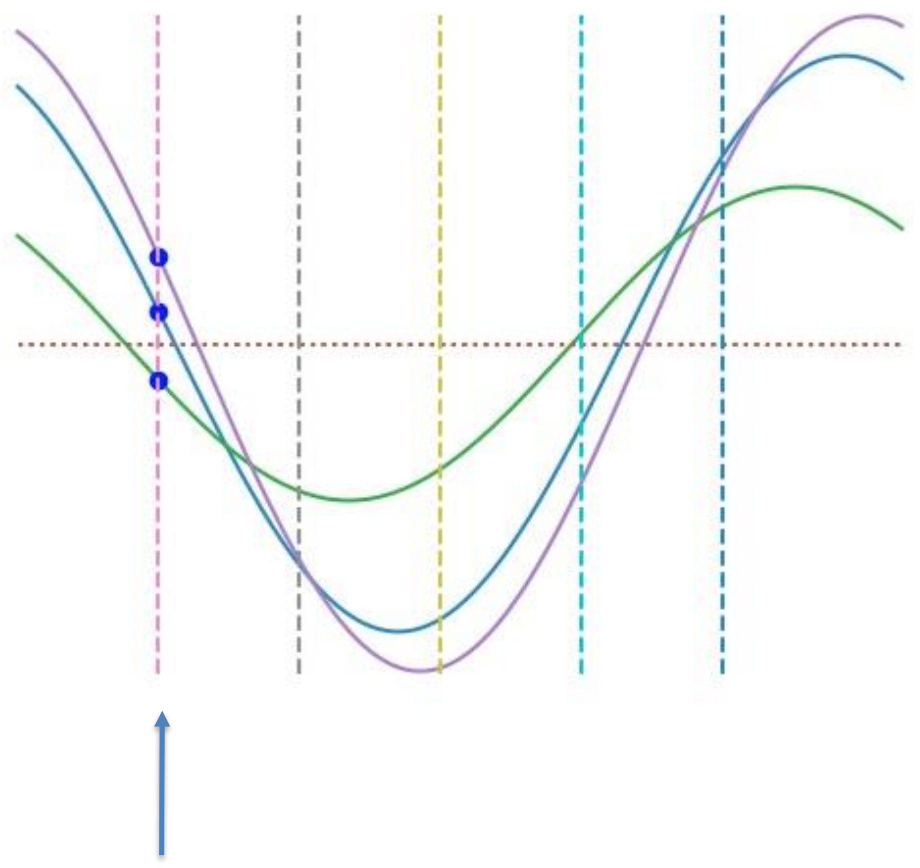
Property 1: Position-free relative-distance transformation
The word/position indexes are invisible in neural networks. It is easier if all the transformation pairs (move a word from one position to another one) $[g_{w,d}(1) \rightarrow g_{w,d}(n+1), g_{w,d}(2) \rightarrow g_{w,d}(n+2), \cdots, g_{w,d}(L) \rightarrow g_{w,d}(n+L)]$ correspond to a same n-offset-transformation without considering the start position.



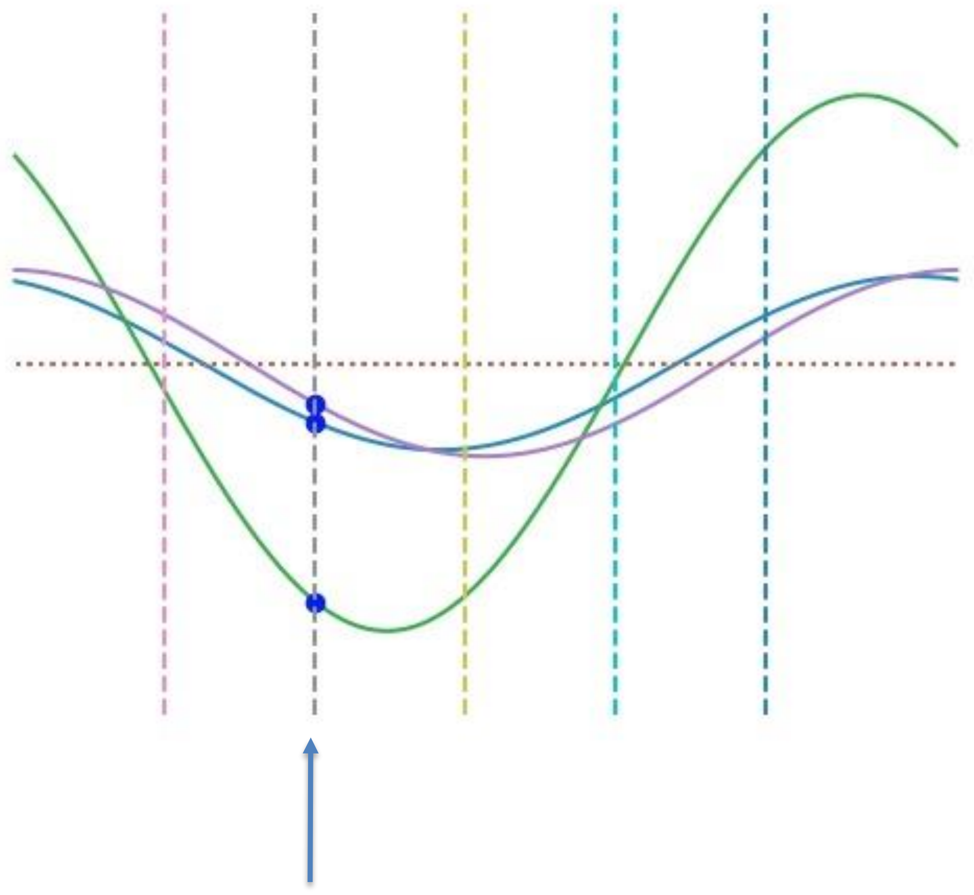
Property 2: Boundedness
The function $\boldsymbol{g}_{\boldsymbol{w},d}$ should be bounded, in order to model long enough sentence

The first property make the problem much simper and can be feasible to solve

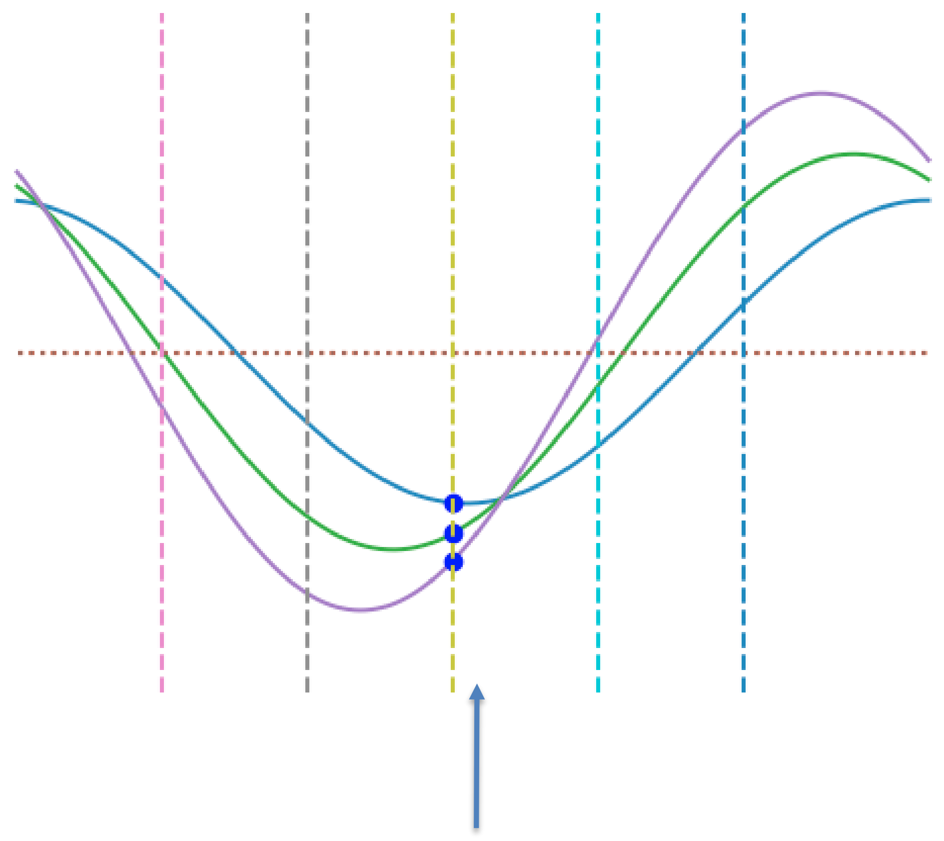
Words as waves



Natural in the first position



language in the 2nd position



processing in 3rd position

For the sentence ‘Natural language processing

On position embeddings in BERT

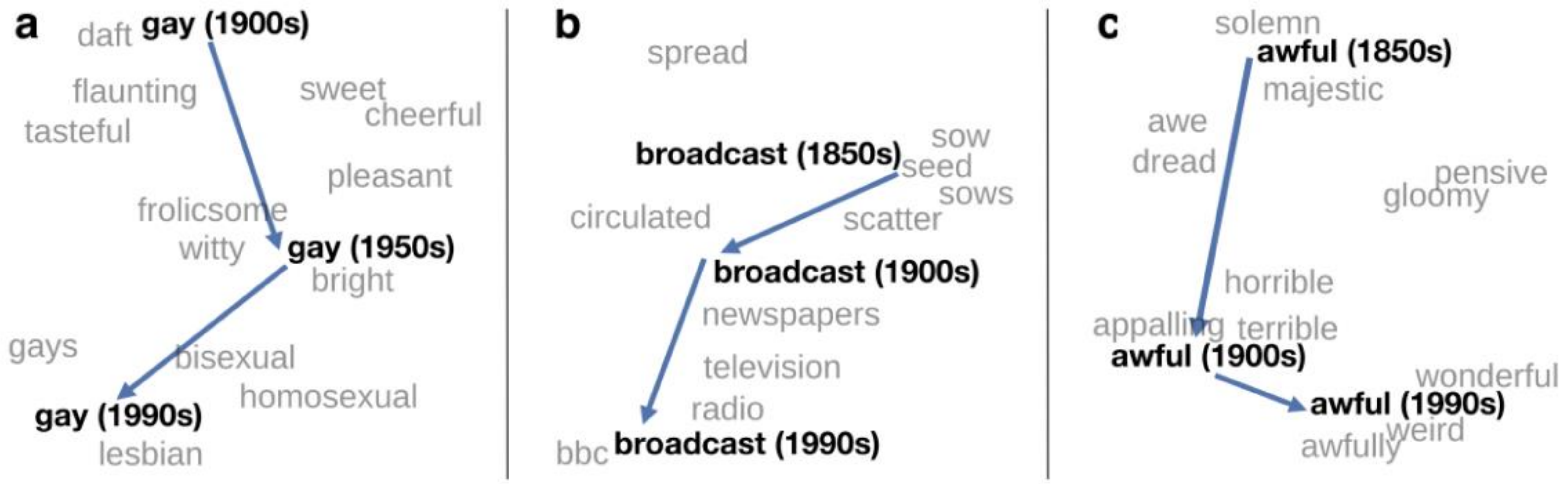
We define three properties and check to which degree various PE satisfy such properties: **Translation invariance**; **monotone**; **symmetry**
We systematically compare Absolute PE/Relative PE including fully-learnable, semi-learnable, and sinusoidal paramerization
Many tips of PEs is proposed as below:

1) untie [CLS] and position embeddings for document-level classification
2) use RPE for token-level classification
3) do not use sinusoidal parameterisation for RPE
4) absolute position is uninformative and translation invariance makes sense
5) leaning frequencies in sinusoidal PE is slightly beneficial
6) combining APE and RPE is slightly beneficial in SQuAD but not for GLUE
7) be safe to truncate RPEs
8) usually insensitive to the distance for long-range attending
9) Treat forward and backward differently
10) PE with strict translation invariance can be generalised to longer documents

Wang et.al. On position embeddings in BERT. **ICLR** 2021.

From spacial to **temporal sequence**: dynamic word embedding

Dynamic word embedding: word meaning may change over time, for example ***Trump in 2018 is like Biden in 2022***. We could also use complex word embedding to encode temporal word evolution



Thanks to Weierstras Approximation Theorem, it is proved that a complex word embedding inspired sinusoidal word embedding **could approximate any semantic evolution**. It also set a new SOTA on temporal lexical tasks.

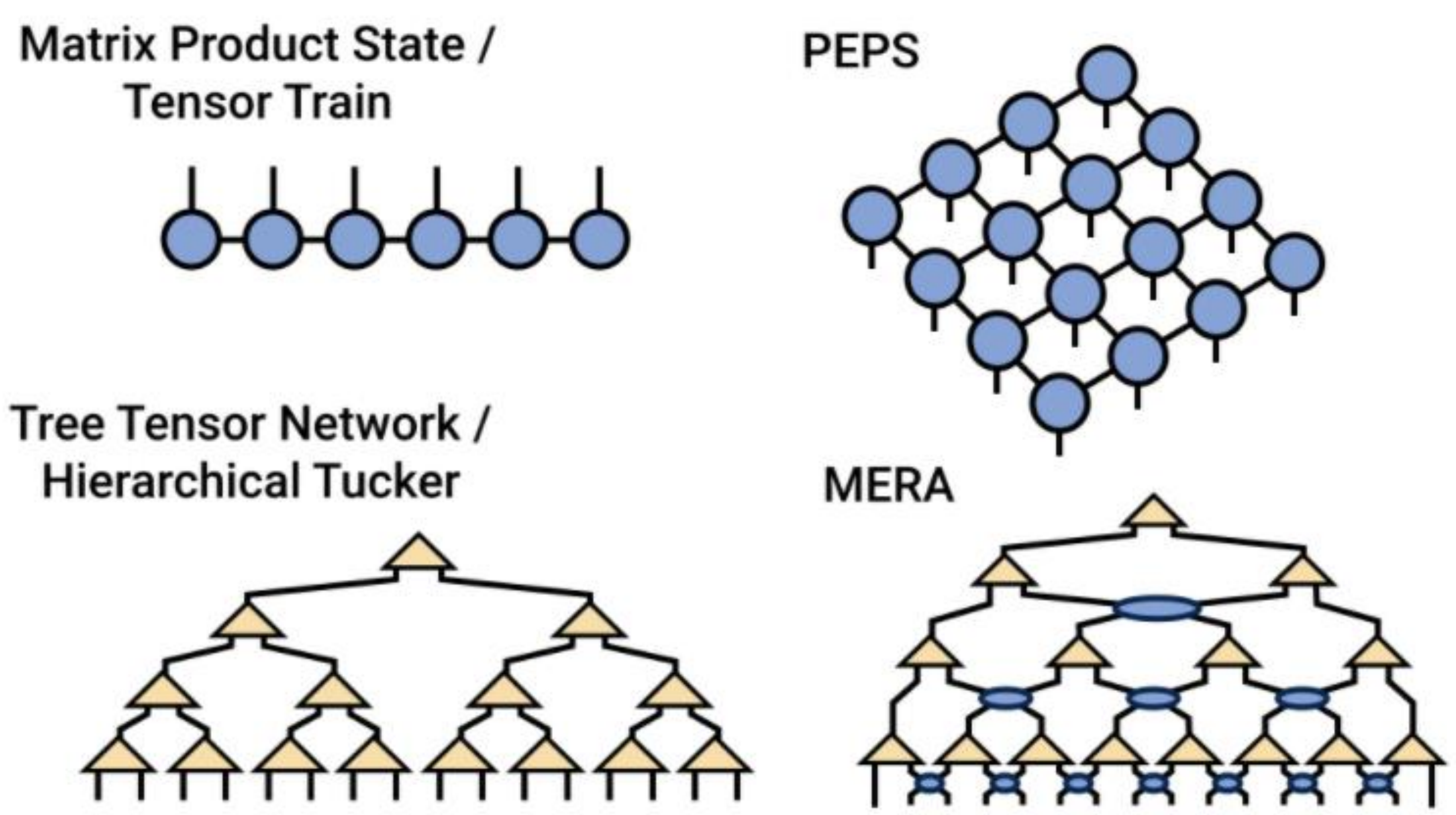
Wang. et.al. Word2fun: modeling words as functions for diachronic word representation. NeurIPS 2021

Contents

- On the motivations of quantum theory in NLP
- Overview of the research
  - Interpretability:
    - Modeling words as particles for better interpretability
    - **Modeling words as waves to encode order**
  - **Efficiency: Network Compression using Tensor Networks**
  - **Potential**: Quantum computing equipped language models.

BIG Tensors in Physics - tensor network

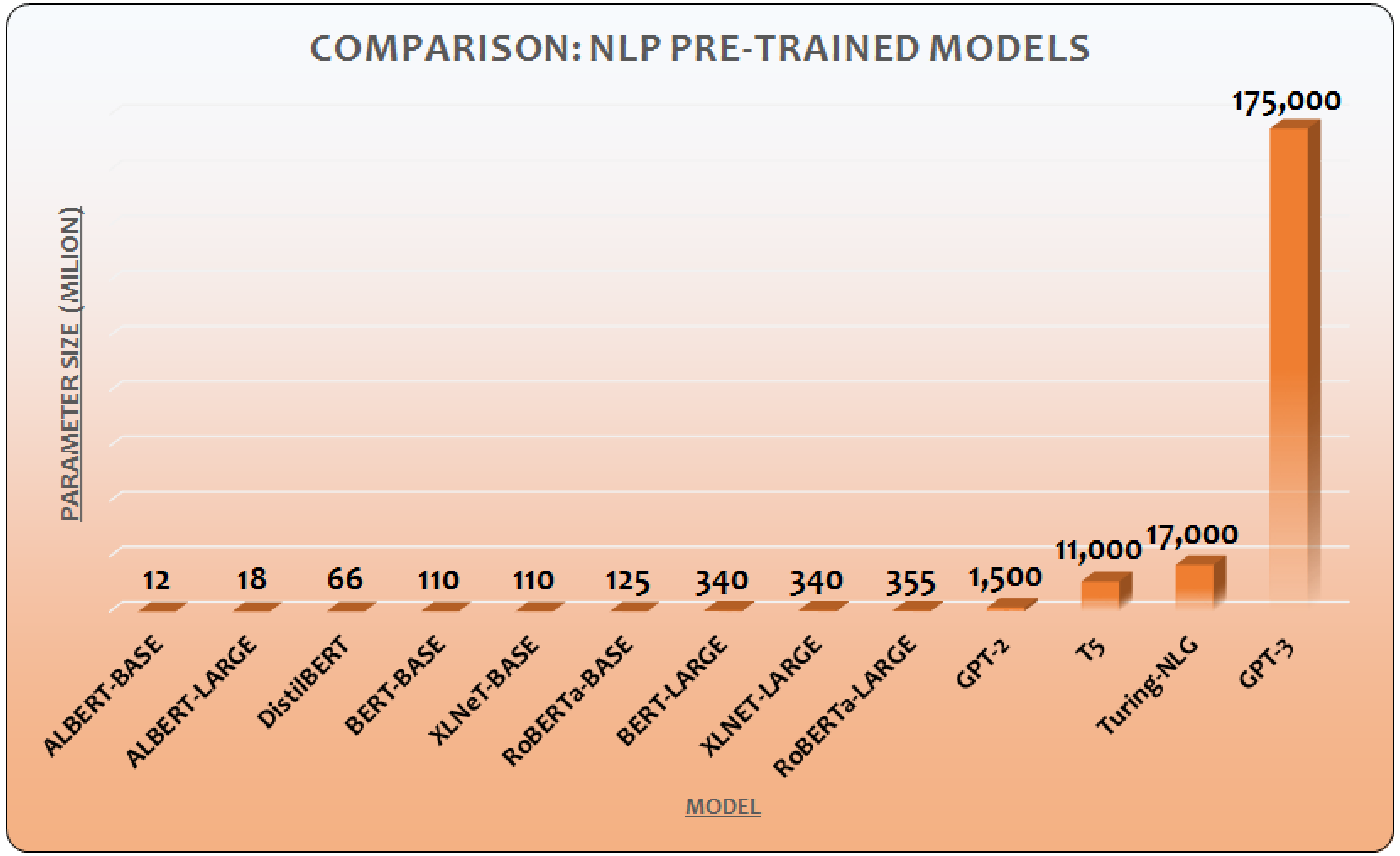
States of many particles system are represented by usually (exponentially) large tensors. Tensor Network is used to accurately describe many-particles states in limited parameters, which can be reverse of **higher order matrix decomposition** (a.k.a, tensor decomposition)



For example, tensor networks are factorizations of very large tensors (quantum many body wave function) into networks of smaller tensors.

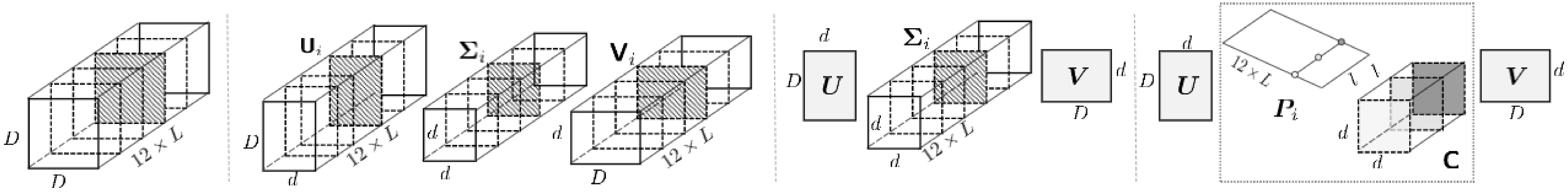
image source： https://tensornetwork.org/

Quantum circuit friendly PLMs



Increasing model size after GPT 3 would be much expensive, slow and environmentally-unfriendly, we need to find a alternative way to build super-large pre-trained language models

Four paradigms for compressions



I: raw weight matrices $\mathbf{W}_i^{\mathrm{I}} = \mathbf{W}_i$

II: matrix decomposition $\mathbf{W}_i^{\mathrm{II}} = \mathbf{U}_i \mathbf{\Sigma}_i \mathbf{V}_i$

III: tensor-train decomposition $\mathbf{W}_i^{\mathrm{III}} = \boldsymbol{U} \mathbf{\Sigma}_i \boldsymbol{V}$

IV: tucker decomposition $\mathbf{W}_i^{\mathrm{IV}} = \boldsymbol{U}(\boldsymbol{P}_i \mathbf{C})\boldsymbol{V}$

Wang et.al. Exploring extreme parameter compression for pre-trained language models. ICLR 2022

Experimental results

Model (our models in bold)	Para.	FLOPS	RPS	SST-2 acc	MNLI acc	MRPC F1	QNLI acc	QQP F1	RTE acc	STS-B spear.	all
BERT-base (Devlin et al., 2018) (BERT-I)	86.0M	22.5B	420.1	**93.4**	83.9/83.4	87.5	**90.9**	71.1	66.4	**85.2**	82.7
BERT-III -384	23.0M	22.5B	452.2	**93.4**	**84.6/83.7**	**88.1**	90.5	**71.9**	68.1	83.9	**83.2**
BERT-III -64	**1.8M**	**4.3B**	**1143.6**	91.9	80.1/79.6	85.5	87.7	70.7	63.3	80.7	80.0
BERT-IV -72-384	12.3M	22.5B	452.3	93.1	83.9/83.2	87.5	90.2	71.6	67.3	83.6	82.6
BERT-IV -36-256	3.9M	15.2B	596.9	92.7	82.5/81.8	87.1	88.9	71.4	65.2	81.8	81.4
BERT-IV -36-128	1.9M	8.0B	863.0	92.4	81.1/80.2	86.5	88.3	71.9	64.4	81.4	80.8

Our compressed BERT achieved 97.5% of the performance with **1/48** parameters

Wang et.al. Exploring extreme parameter compression for pre-trained language models. ICLR 2022

Contents

- On the motivations of quantum theory in NLP
- Overview of the research
  - Interpretability:
    - Modeling words as particles for better interpretability
    - Modeling words as waves to encode order
  - Efficiency: Network Compression using Tensor Networks
  - **Potential: Quantum computing equipped language models.**

Quantum NLP with PLMs

- **Case 1: classical models using TN**
- Case 2: classical-quantum hybrid
- Case 3: fully-quantum model

Case 1: Language model as TN

Suppose a word vocabulary V, language model is to give a probability of an N-gram is :

$$\underbrace{\mathbb{V}, \cdots \mathbb{V}}_{n} \rightarrow \mathbb{R}^{+}$$

Denoted as $\mathcal{A} \in \mathbb{R}^{V^N}$

This is a exponentially-large space with respect to N.
One can find a efficient way to approximate is like Matrix Product State (MPS or TT decomposition), we could call it **Tensor Network Language Model (TNLM)**

$$\mathcal{A} \approx M_1 G_2 G_3 \cdots G_{N-1} M_N \stackrel{def}{=} M_1 G G \cdots G M_N$$

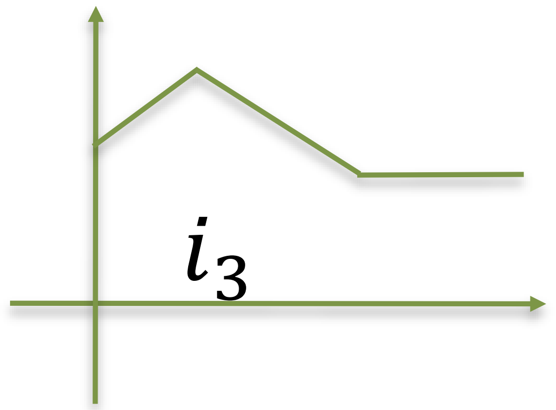
And an element of is $\mathcal{A}$

$$\mathcal{A}_{i_1, i_2, \cdots, i_n} \approx M_{i_1} G_{i_2} G_{i_3} \cdots G_{i_{N-1}} M_{i_N}$$

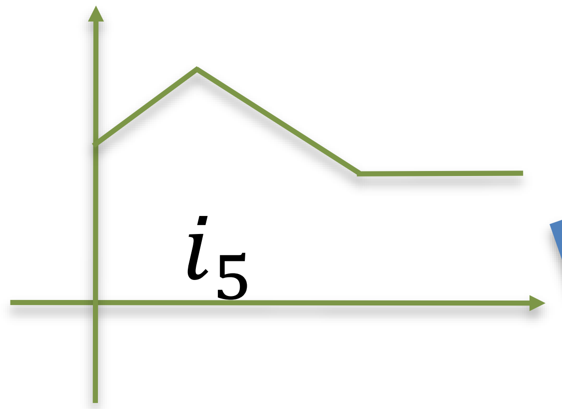
Which is equivalent to **map words as matrices** (instead of vectors)

Case 1: Distill PLMs to TN

w_{i_3}



w_{i_5}



Existing PLMs

BERT

$w_{i_1} w_{i_2} * w_{i_4} w_5$

GPT

$w_{i_1} w_{i_2} w_{i_3} w_{i_4}$

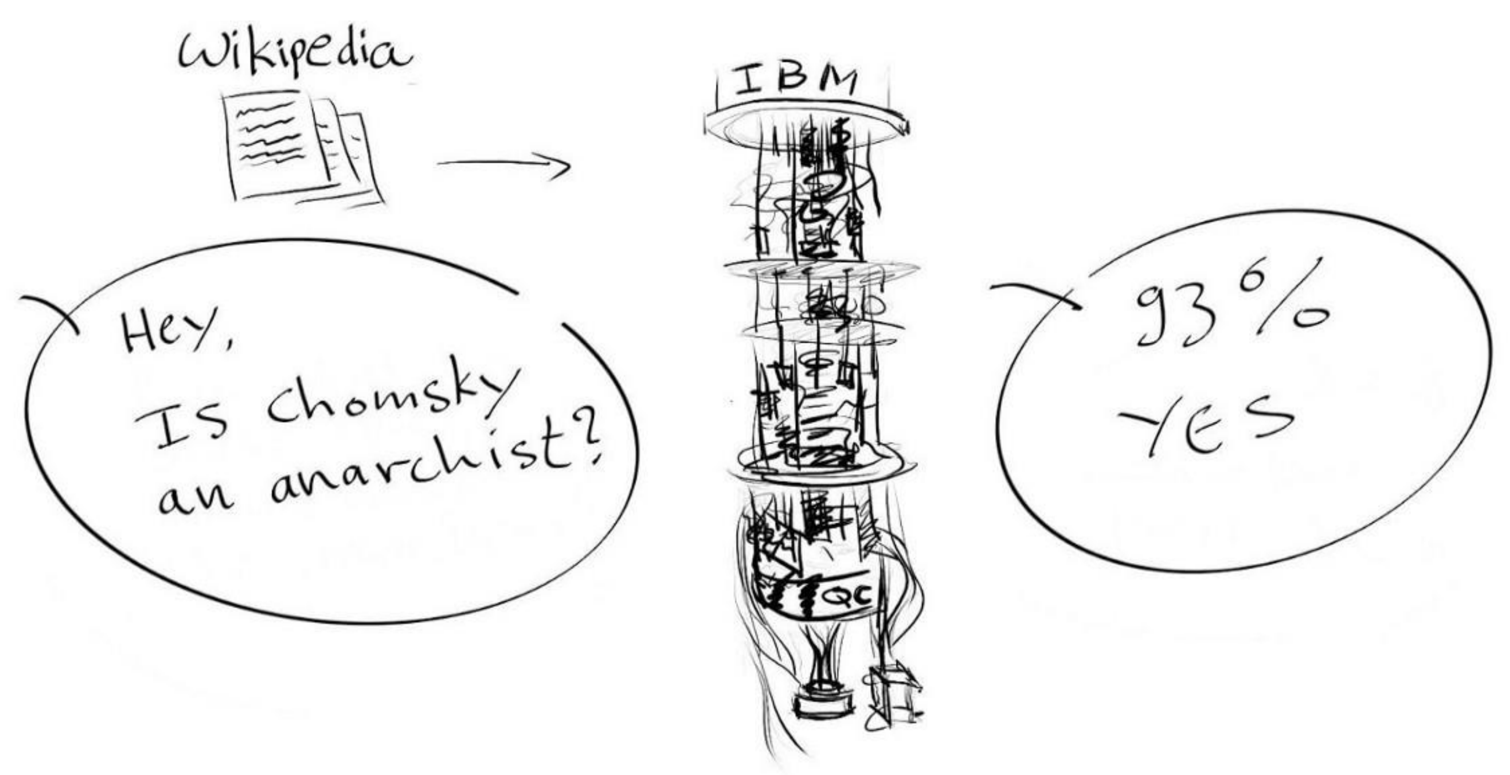
TN model

$\mathcal{A}_{i_1,i_2,*,i_4,i_5} \approx M_{i_1} G_{i_2} G \;\; G_{i_4} M_{i_5} \in \mathbb{R}^V$

$\mathcal{A}_{i_1,i_2,i_3,i_4} \approx M_{i_1} G_{i_2} G_{i_3} G_{i_4} M \;\; \in \mathbb{R}^V$

Wang et al. Distilling Pre-trained Language models into Tensor networks. In progress.

Whether TNLM could be Implemented in quantum computer ?



- If it could be achieved, we have to divide such a goal to many stages and define the challenges.
- If we could not, we should clearly state the bottlenecks and inspire more people to solve it

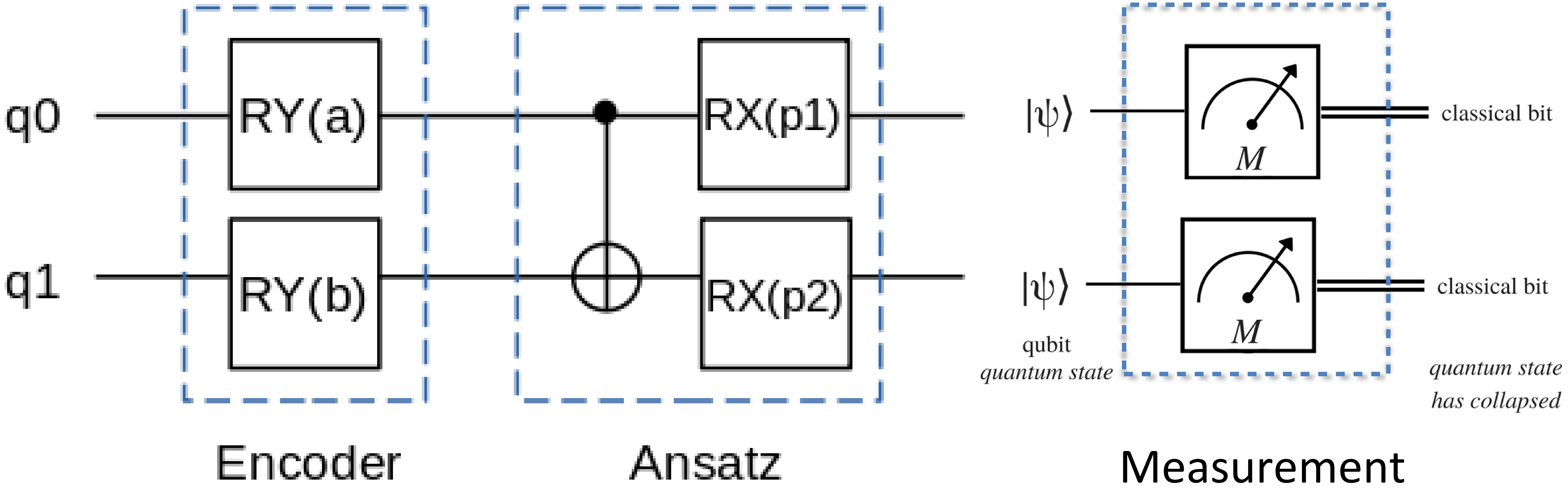
- At least, researchers from Oxford went to the first step to run NLP task in quantum computer, and there is a large space to explore

https://quantumzeitgeist.com/cambridge-quantum-computing-publishes-new-scientific-papers-on-meaning-aware-quantum-natural-language-processing/

Quantum NLP with PLMs

- Case 1: classical models using TN
- **Case 2: classical-quantum hybrid**
- Case 3: fully-quantum model

QC for classical data



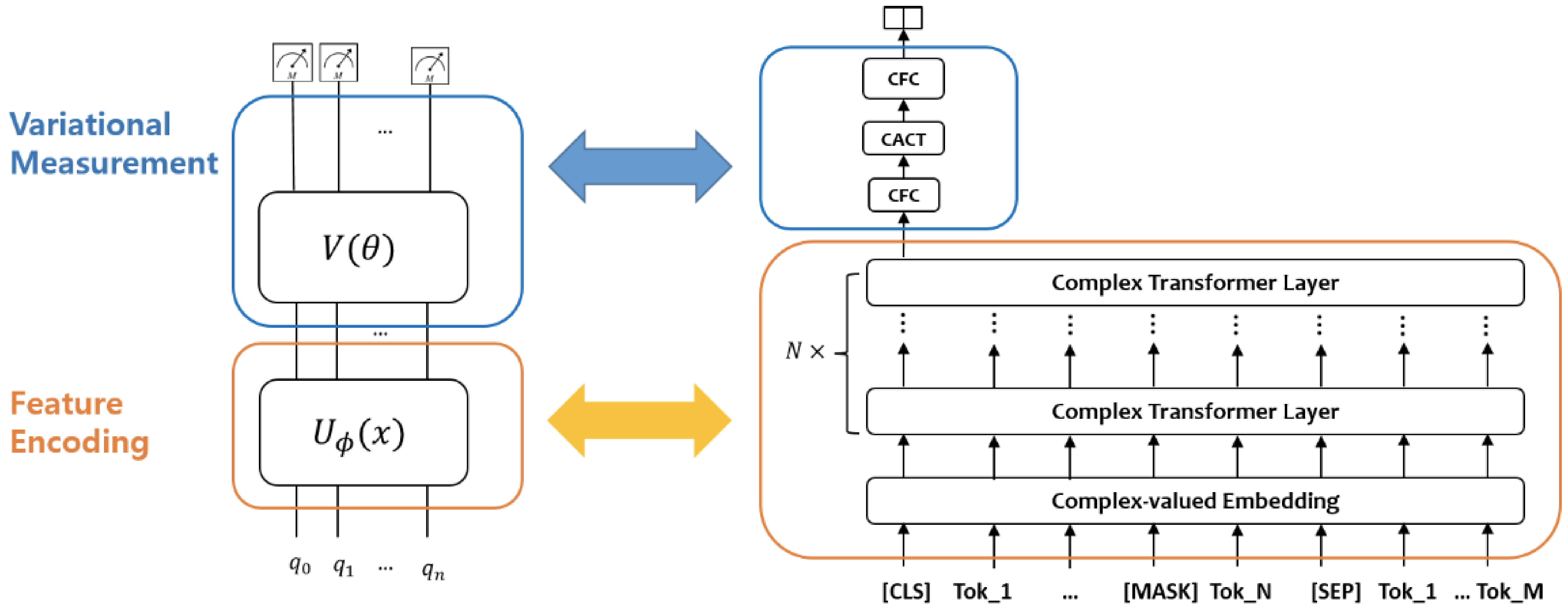
classical bits → qbits → classical bits

Quantum embedding

- N QBits encodes 2^N-dimensional info
    - Encoding real-valued feature to **unit** complex-valued features
- $f: \mathbf{V}^N \to \mathbb{C}^M$, $\mathbf{V}$ is the set of words/tokens
- e.g., $f(abc) \in \mathbb{C}^M$, and $|f(abc)|_2 = 1$

Case2: classical-quantum hybrid

Complex-valued language models as quantum embedding



Qiuchi Li, **Benyou Wang.** et al. CVBERT: Complex-valued Pre-trained Language Model and its Quantum adaption. ICML 2022 in submission.

Quantum NLP with PLMs

- Case 1: classical models using TN
- Case 2: classical-quantum hybrid
- **Case 3: fully-quantum model**

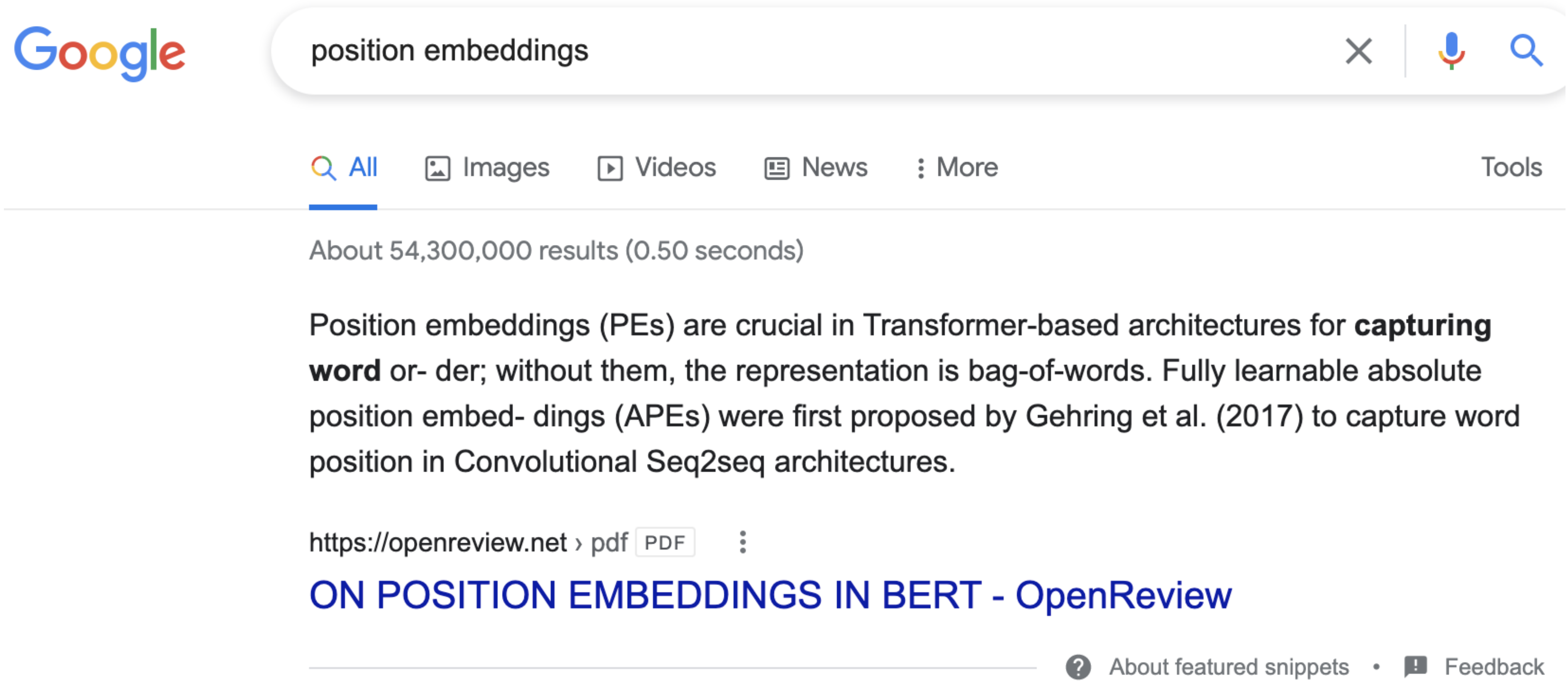
Case 3: encoding a LM into QBits

Language modelling is the task of assigning a probability to sentences in a language.
e.g., with a N-length language set (or called 'n'-gram) $f: \mathbf{V}^N \rightarrow \mathbb{R}^+$

We could directly encode such a language model in a quantum state with $Nlog_2V$ Qbits, the downstream tasks will be performed on such a quantum state, e.g., using sampling.

Thanks

Significance of our research in position embeddings



https://www.google.com.hk/search?q=position+embeddings

Reviews on our paper related Position embeddings

Sorry this is not scientifically, but I have to mention that I find the axiomatic derivation of the approach **simply beautiful**. It is **amazing** to find such a **simple formula** from **two obvious properties** that someone would want from a positional encoding: Position-free offset transformation and boundedness to handle arbitrary length.

Official Blind Review #3 [1]

I am more positive about the paper now and have increased my score to an 8. I think this paper is going to be useful for the community and I know **I will reference it later and direct others** to it who are interested in learning more about position embeddings in transformers (**whether or not it actually gets published**).

Official Blind Review #1 [2]

Like early work championed in ML venues such as LDA that went on to have an important impact in application areas, this is a **serious technical contribution that will have an long afterlife in diachronic sociolinguistics**.

Official meta Review [3]

Wang et.al. encoding word order in complex embeddings. ICLR 2020
Wang et.al. On position embeddings in BERT. **ICLR** 2021.
Wang. et.al. Word2fun: modeling words as functions for diachronic word representation. NeurIPS 2021

Overview of the research

- Case 1: Quantum probability driven neural networks (QPDN) [1,2]
- Case 2: Position embeddings explained inspired by complex embeddings [3,4,5]
- Case 3: Compressing BERT using tensor networks [6]
- Case 4: Quantum Computer friendly pre-trained language model [7]

[1] **Wang** et.al. Semantic Hilbert space for Text Representation Learning. **The Web Conference** 2019
[2] Li*, **Wang***, and Melucci. An interpretable complex-valued network for matching. **NAACL BEST Explainable Paper**. NAACL 2019 in which BERT won the Best Paper.
[3] **Wang** et.al. Encoding word order in complex embeddings. **ICLR 2020 spotlight**.
[4] **Wang** et.al. On position embeddings in BERT. **ICLR** 2021.
[5] **Wang**. et.al. Word2fun: modeling words as functions for diachronic word representation. NeurIPS 2021 submission
[6] **Wang** et.al. Compressing pre-trained language models using tensor decomposition. NeurIPS 2021 submission
[7] Ongoing work on `Quantum circuit friendly pre-trained language models', with other collaborators, targets ICLR 2022

Complementary of Physics vs. ML

Physics: based on while box model (based on knowledge)

- we do not always know the governing equations
- sometime the real world environment is too complicated to be accurately described
- certain parameters required in the formula may not be observable
- it may be too computationally expensive to solve the formulate

ML: data-driven black model based on collected data

- data hungry, curse of dimensionality
- not trustworthy for regions with nod data
- the learned models re not capable of generalising to other tasks even for similar tasks
- lack interpretability

Visualisation for matching

Question	Correct Answer
Who is the [president or chief executive of Amtrak] ?	“ Long-term success ... ” said George Warrington , [Amtrak ’s president and chief executive] .”
When [was Florence Nightingale born] ?	,”On May 12 , 1820 , the founder of modern nursing , [Florence Nightingale , was born] in Florence , Italy .”
When [was the IFC established] ?	[IFC was established in] 1956 as a member of the World Bank Group .
[how did women ’s role change during the war]	..., the [World Wars started a new era for women ’s] opportunities to ....
[Why did the Heaven ’s Gate members commit suicide] ?,	This is not just a case of [members of the Heaven ’s Gate cult committing suicide] to ...

Example of tensor network

Suppose a particle is in superposed states with a D-dimension space, and a system have N particles, it state is :

$$\vec{\phi} \in \mathbb{R}^{D^N}$$

This is a exponentially-large space with respect to N.
One can find a efficient way to approximate is like Matrix Product State (MPS or TT decomposition)

$$\vec{\phi} \approx M_1 G_2 G_3 \cdots G_{N-1} M_N$$

when $M_1 \in \mathbb{R}^{D\times r}; G_2 G_3, \cdots G_{N-1} \in \mathbb{R}^{r\times D\times r}; M_N \in \mathbb{R}^{r\times D}$

Property 1

Problem: We consider the simplest case when the n-offset transformation

$$f(n): g(pos) \to g(n + pos)$$

Which transform one from pos-th position to (pos+n) position to be **linear**.

$$g_{w,d}(pos) f_{w,d}(n_1) f_{w,d}(n_2) = g_{w,d}(pos) f_{w,d}(n_1 + n_2)$$

Solution: It is trivial to get the following solution (proof in the paper):

$$f_{w,d}(n) = z_1^n$$

Result: z_1 is the parameters and $g_{w,d}(0) = z_2$ [1], such that

$$g_{w,d}(pos) = z_2 z_1^{\boldsymbol{pos}};$$

[1] z_1, z_2 are related to the word index and position index, but superscripts are ignored for simplicity

Property 2

To make $g_{w,d}(pos)$ to be bounded:

$g_{w,d}(pos) = z_2 z_1^{pos}$; $subject\ to |z_1| \leq 1$

In real-domain, we necessary consider the extra constraint with some costs.

But if we extend Z_1 in complex domain ($x = \alpha + \beta i = re^{i\theta}$), it is easier.

For example, $i = i; i^2 = -1; i^3 = -i; i^4 = 1; \cdots$

Property 2

To make $g_{w,d}(pos)$ to be bounded:

$g_{w,d}(pos) = z_2 z_1^{pos}$; $subject\ to |z_1| \leq 1$

In real-domain, we necessary consider the extra constraint with some costs.

But if we extend z_1 in complex domain ($x = \alpha + \beta i = re^{i\theta}$), it is easier.

For example, $i = 1; i^2 = -1; i^2 = -1; i^3 = -i; i^4 = 1; \cdots$

Let $z_1 = r_1 e^{i\theta_1}; z_2 = r_2 e^{i\theta_2}$

$g_{w,d}(pos) = z_2 z_1^{pos} = r_2 e^{i\theta_2}(r_1 e^{i\theta_1})^{pos} = r_2 r_1^{pos} e^{i(\theta_2 + \theta_1 pos)}\ subject\ to |r_1| \leq 1$

We directly make $r_1 = 1$, get

$g_w(pos) = r_2 e^{i(\theta_2 + \theta_1 pos)}$

The proposed embedding

Our definition:
A word in *pos*-th position is represented as

$$[r_{j,1}e^{i(\omega_{j,1}pos+\theta_{j,1})}, \ldots, r_{j,2}e^{i(\omega_{j,2}pos+\theta_{j,2})}, \cdots, r_{j,D}e^{i(\omega_{j,D}pos+\theta_{j,D})}]$$

where each dimension like d has an amplitude $r_{j,d}$, and a unique period of $p_{j,d} = \frac{2\pi}{\omega_{j,d}}$.

i is the imaginary number.

Based on Euler's formula (i.e. $e^{ix} = \cos x + i\sin x$), each element can be rewritten as:

$$g_{j,k} = r_{j,d}\cos(\omega_{j,d}pos + \theta_{j,d}) + r_{j,d}\sin(\omega_{j,d}pos + \theta_{j,d})i$$

Interpretability

What is interpretability?

- Interpretability issue for NN-based NLP models

  1. **Transparency**: explainable component in the design phase
  2. **Post-hoc Explainability**: why the model works after execution

The Mythos of Model Interpretability, Zachery C. Lipton, 2016

Research questions :

1. What is the concrete meaning of a single neutron? And how does it work? (*probability*)
2. What did we learning after training? (*unifying all the subcomponents in a single space and therefore they can mutually interpret each other*)

Li, Qiuchi*, **Benyou Wang***, and Massimo Melucci. "CNM: An Interpretable Complex-valued Network for Matching." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4139-4148. 2019. NAACL 2019 best explainable paper

Benyou Wang, Qiuchi Li, Massimo Melucci, Dawei Song, Semantic Hilbert space for text representation learning. TheWebConf 2019

Three aspects of transparency

- Simulatability:
  - ***Simulate*** *the neural network in reasonable time using its input and parameters*
- Algorithmic transparency:
  - *Known error surface and unique converged solution if it has*
- **Decomposability**:
  - Each part of model has “intuitive explanation"
    - **Input** (e.g., word and word embedding)
    - Network **weights,** (e.g., CNN kernels and LSTM cells)
    - **Calculations,** (e.g., cell update, convolution)
    - **Output**

We aim to build a **probability**-driven neural networks.

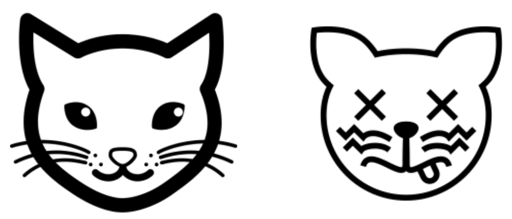
Zachery Lipton. The mythos of model interpretability. 2016

Quantum probability

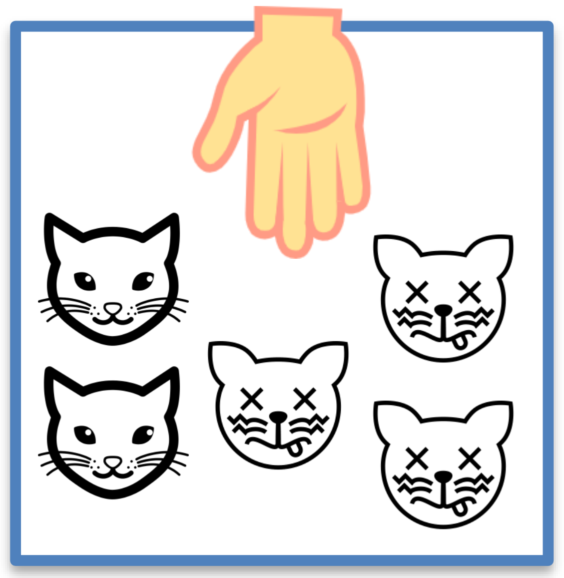
Quantum Probability Theory

a probability theory defining on **vector spaces**

Set-based Probability Theory



alive dead



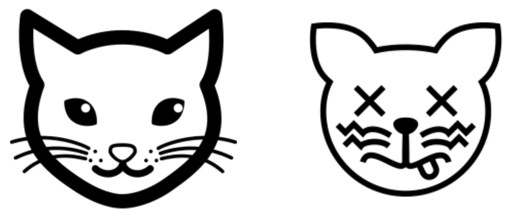
Q: Should the randomly-chosen cat dead or alive ?

<u>A: 0.4 to be alive and 0.6 to be dead</u>

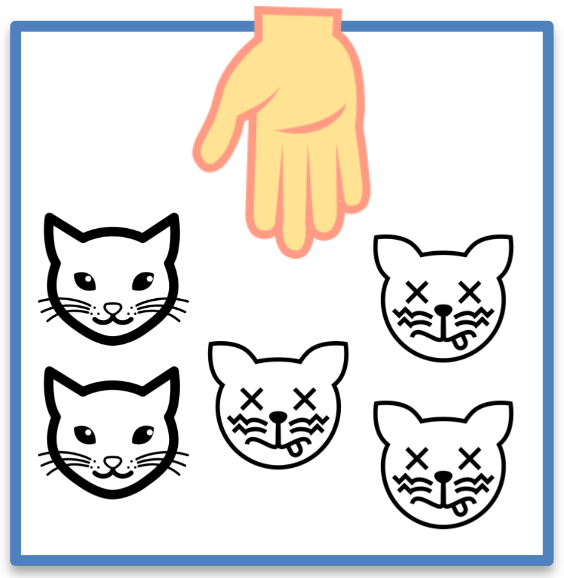
Quantum Probability Theory

a probability theory defining on **vector spaces**

Set-based Probability Theory



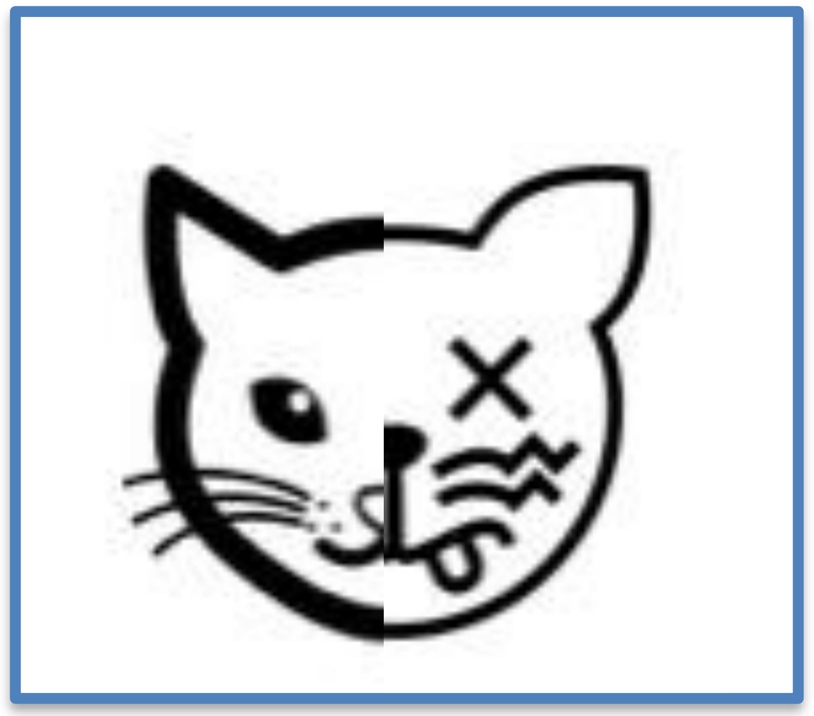
alive dead



Q: Should the randomly-chosen cat dead or alive ?

<u>A: 0.4 to be alive and 0.6 to be dead</u>

Quantum Probability Theory - **vector-based**



Superposition

Q: Are these cat dead or alive?

<u>A: 0.501 to be alive and 0.499 to be dead</u>

Link to sinusoidal position embedding

TPE definition: $g'_{j,k} = WE'(j,\cdot) + PE'(\cdot, pos)$

$PE'_{2k}(\cdot, pos) = \sin(pos/10000^{2k/d_{model}});$

$PE'_{2k+1}(\cdot, pos) = \cos(pos/10000^{2k/d_{model}})$

It can be considered as a **specific case of ours** when $\omega_{\cdot,d} = \frac{1}{10000^{d/2d_{model})}}$

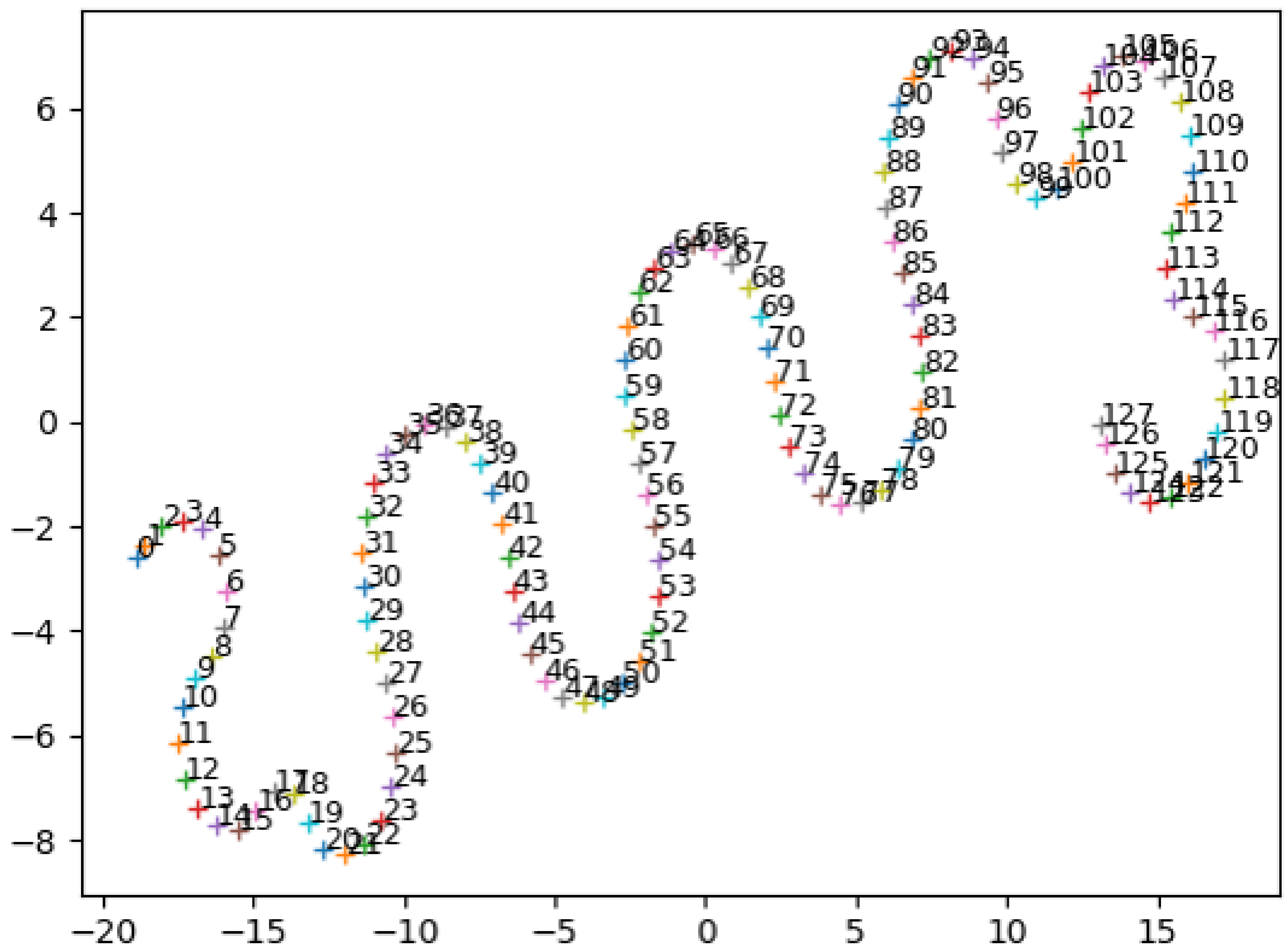
$$g_{j,k} = WE(j) \odot \left(\cos(\omega_{j,d} pos) + i\sin(\omega_{j,d} pos)\right)$$

$$g_{j,k} = WE(j) \odot (PE'_{2k}(\cdot, pos) + iPE'_{2k}(\cdot, pos))$$

$\odot$ is the element-wise multiplication

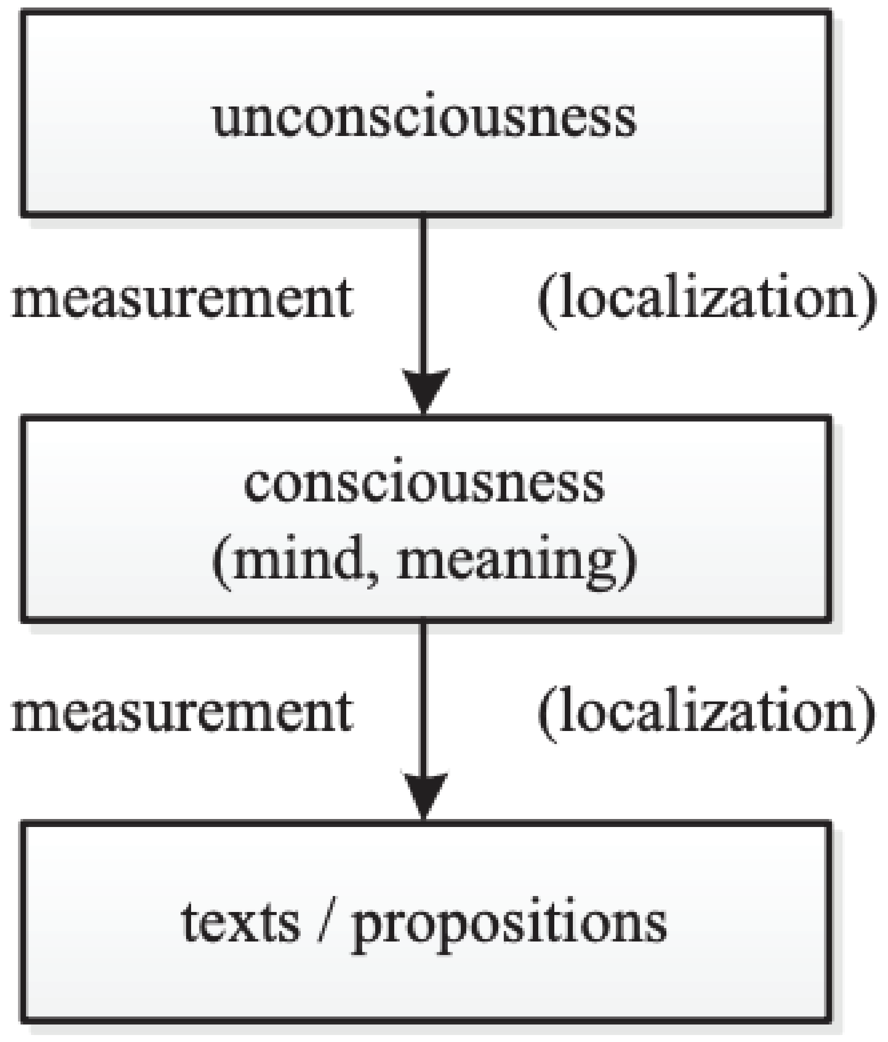
We argue that our proposed embedding is more general.

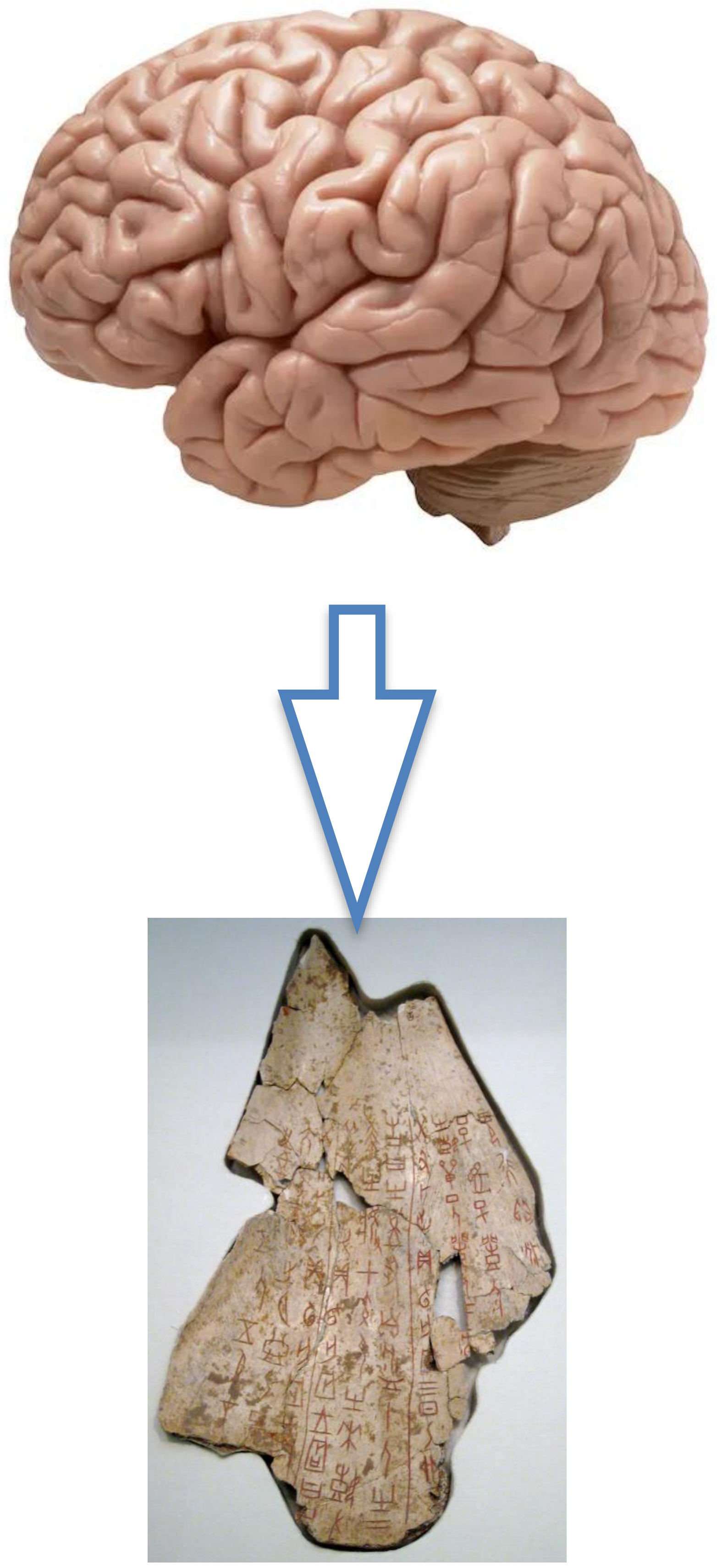
Position is ordered, should **position embedding** be



Visualization of first 128 position embedding of BERT-base-uncased

Are words really superposed?





Transformations among the unconsciousness, the consciousness, texts and propositions. The localization represents that the globally implicit meaning is explicitly expressed by texts or propositions

Xie, M., Hou, Y., Zhang, P., Li, J., Li, W., & Song, D. (2015). Modeling quantum entanglements in quantum language models. AAAI.