A theory of human-like k-shot learning

Ming Li
University of Waterloo

Joint work: Zhiying Jiang, Rui Wang, Dongbo Bu

Dec. 6, 2022

Modeling human k-shot learning is a major open question:

+ There are 8 billion people on earth. Model whose algorithm?

+ Hebart et al (*Nature*, 2020) 49 dimensions. Is that all?

+ Human learning does not need high accuracy.

+ In fact, specialization & communication is also important.

Deep Learning vs Human Learning

Deep learning:

+ Parameter laden
+ Data heavy
+ Energy hungry

Human learning:

+ No parameter
+ Not much data
+ Energy efficient

How to model human K-shot learning is a major open question in AI

Goal of today:

- "Derive" a k-shot learning theory, not "proposing one"
- This model covers all other models
- This is easily usable.

Human k-shot learning

n unlabeled data

k labelled data

One universal program does all

Animal k-shot learning

Cassowary mother teaches her chicks what to eat.

Mother crow teaches youngster how to use tools

Facts about k-shot learning

k shot learning is one of the main methods for human learning: k labelled samples, some unlabelled

It is not large data deep learning or transfer learning

No other labelled data
One program / individual
Do not want: model out of blue
Want: unified general model for all

1 shot learning :

If this is edible

So are these






Formal definition of k-shot learning

Definition 1. Consider a universe Ω, partitioned into H concept classes: $\mathcal{C}_h, h = 1, 2, \ldots, H$. k-shot learning is described in the following:

- Some, say n, elements in or outside Ω are given as unlabelled samples $y_1, \ldots, y_n$;
- There are k labelled examples for each class $\mathcal{C}_h$, for small k;
- The learning program, using a computable metric $\mathcal{M}$, k-shot learns $\mathcal{C}_h, h = 1, 2, \ldots H$, if it uses the n unlabelled samples and k labelled samples and minimizes the objective function:

$$\sum_{h=1}^{H} \sum_{i=1}^{|\mathcal{C}_h|} \mathcal{M}(x_i \in \mathcal{C}_h, center_h | y_1, \ldots, y_n).$$

Note: M is personalized result of evolution. It checks how similar an instance is to the given k samples of a class, according to this individual.

1-shot learning

1. One labelled sample x_h for class C_h, $h=1, \ldots, H$
2. n unlabelled samples, $y_1, \ldots, y_n$
3. A person used a unique M.
4. Then we seek to minimize:

$$\sum_{h=1}^{H} \sum_{i=1}^{|\mathcal{C}_h|} \mathcal{M}(x_i, x_h | \mathcal{H}(y_1, \ldots, y_n)),$$

The metric M:

1. It is from evolution, can be anything that works？
2. Hebart et al summarized 49 highly reproducible dimensions to 1854 objects, *Nature, 2020*
3. Such metric is individualized, does not have to be successful for all individuals in a species.

We have evolved abilities of two distance measures

Cognition Distance M : recognize food

What is this？

3D Euclidean Distance

This we know how to define

What would be our cognitive distance M?

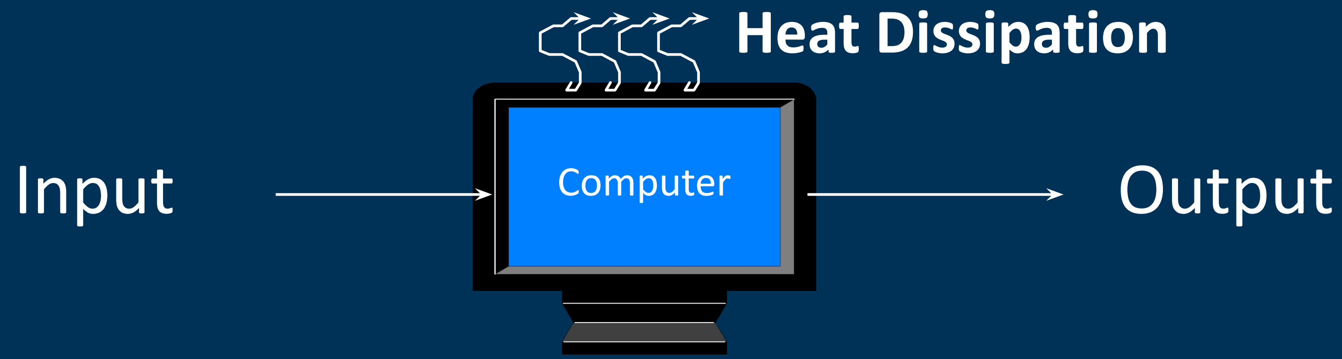
- Euclidean distance, Hamming distance, edit distance, Shannon entropy, mutual information, cross entropy, KL divergence,

Austria

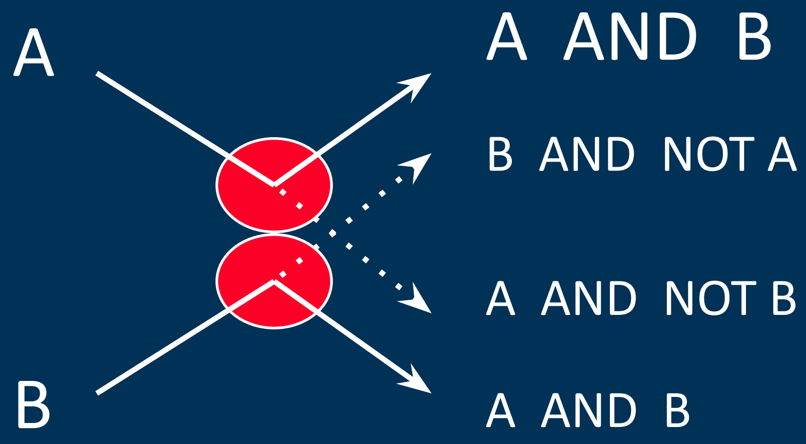
Byelorussia

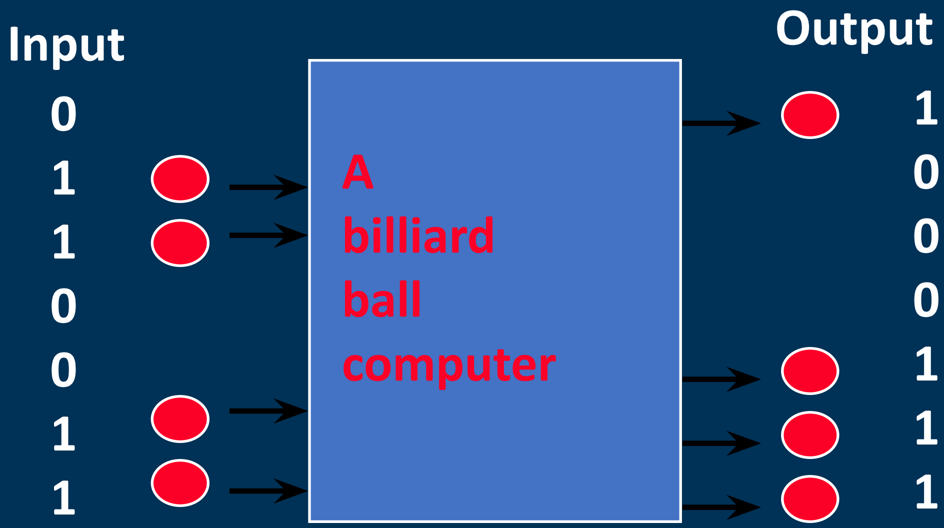
- Hebart et al (*Nature*, 2020) 49 dimensions? Is that all?
- In fact, if we stay with traditional distances, we will always to trapped.
- We need to go back to the very basic laws of physics.

Thermodynamics of Computing



- von Neumann-Landauer law: 1kT is needed to irreversibly process 1 bit
- Reversible computation is free.





Information is physical

- von Neumann-Landauer Law:
  - Axiom 1. Reversible computation is free
  - Axiom 2. Irreversible computation: 1 unit/bit operation

To convert between x and y, the energy needed is:

$$E(x,y) = \min \{ |p| : U(x,p) = y, U(y,p)=x \}$$

Fundamental theorem

Bennett-Gacs-Li-Vitanyi-Zurek Theorem

$$E(x,y) = \max\{ K(x|y), K(y|x) \}$$

$K(x|y)$ is Kolmogorov complexity of x condition on y. I.e. given y, the shortest description length of x

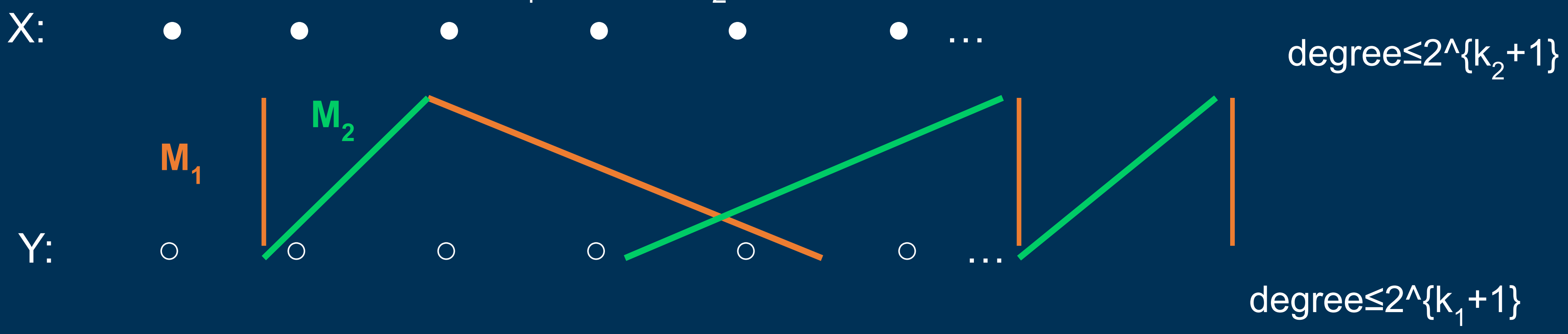
Bennett, Gacs, Li, Vitanyi, Zurek, Information distance, STOC'1993, IEEE Tran-IT 44:4(1998),

Proof: $E(x,y) \le \max\{K(x|y),K(y|x)\}$ direction

Proof. Define graph G={XUY, E}, and let k_1=K(x|y), k_2=K(y|x), assuming $k_1 \le k_2$

- where X={0,1}*x{0}
- and Y={0,1}*x{1}

- E={{u,v}: u in X, v in Y, K(u|v)≤k_1, K(v|u)≤k_2}



- We can partition E into at most 2^{k_2+2} matchings.
  - For each (u,v) in E, node u has most 2^{k_2+1} edges hence belonging to at most 2^{k_2+1} matchings, similarly node v belongs to at most 2^{k_1+1} matchings. Thus, edge (u,v) can be put in an unused matching.
- Program P: has k_1,k_2,i, where M_i contains edge (x,y)
  - Generate M_i (by enumeration)
  - From M_i,x □ y, from M_i,y □ x. QED

Universality

Theorem. For any computable distance measure d, there is a constant c, we have for all x,y,

E(x,y) ≤ d(x,y) + c

- *Interpretation:* E(x,y) is the optimal cognitive distance – it discovers all effective similarities. Everybody's "cognitive distance" can be lower bounded and replaced by E(x,y).

Corollary

$$\sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{E}(x_i \in \mathcal{C}_h, center_h|y_1,\ldots,y_n) \leq \sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{M}(x_i \in \mathcal{C}_h, center_h|y_1,\ldots,y_n).$$

We proved

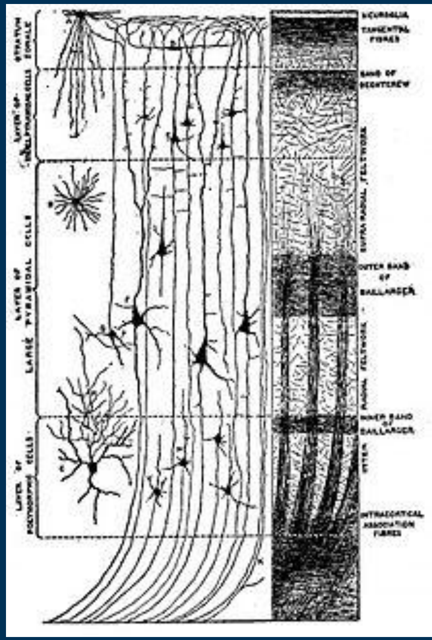
$$\sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{E}(x_i \in \mathcal{C}_h, center_h|\mathcal{H}) \le \sum_{h=1}^{H}\sum_{i=1}^{|\mathcal{C}_h|}\mathcal{M}(x_i \in \mathcal{C}_h, center_h|\mathcal{H}).$$

Thus we design k-shot learning architecture accordingly:

Use hierarchical VAE to model unlabeled data $y_1, \cdots y_n$

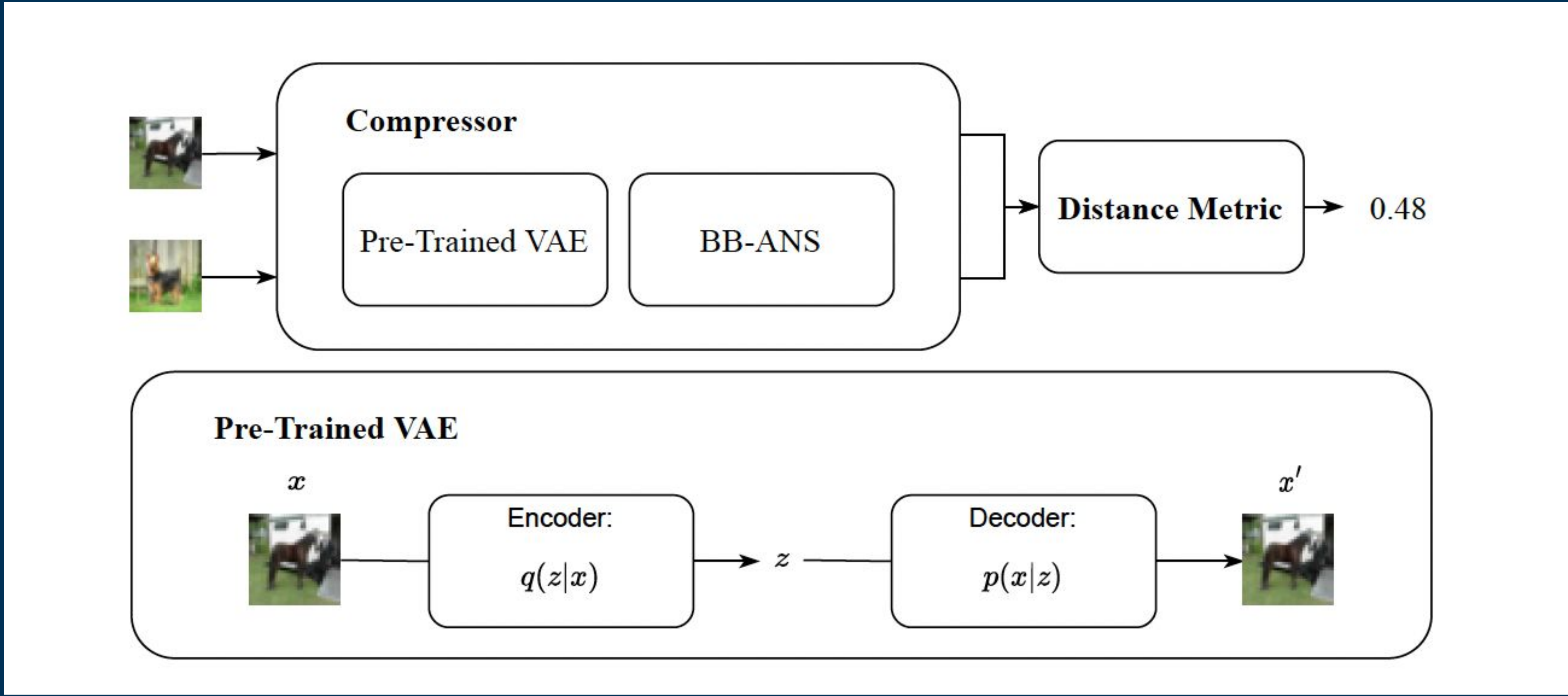
VAE-based Compressor* to approximate : E(x,y | VAE)

Cerebral cortex : 6 layers



* J. Townsend, T. Bird, D. Barber, Practical lossless Compression with latent variables using bits back Coding, ICLR 2019.

Implementing a k-shot learning model



What was missing in our learning models?

All learning paradigm

+ Compression at
training stage

Human k shot learning

+Compression also at
inference stage

Experimental Results on images

	MNIST	KMNIST	FashionMNIST	STL-10	CIFAR-10
SVM	69.4±2.2	40.3±3.6	67.1±2.1	21.3±2.8	21.1±1.9
CNN	72.4±3.5	41.2±1.9	67.4±1.9	24.8±1.5	23.4±2.9
VGG	69.4±5.7	36.4±4.7	62.8±4.1	20.6±2.0	22.2±1.6
ViT (disc)	58.8±4.6	35.8±4.1	61.5±2.2	24.2±2.5	22.3±1.8
Latent	73.6±3.1	48.1±3.3	69.5±3.5	31.5±3.7	22.2±1.6
Ours	77.6±0.4	55.4±4.3	74.1±3.2	39.6±3.1	35.3±2.9

Table 1: 5-shot image classification accuracy on five datasets.

Jiang, Li, 2022

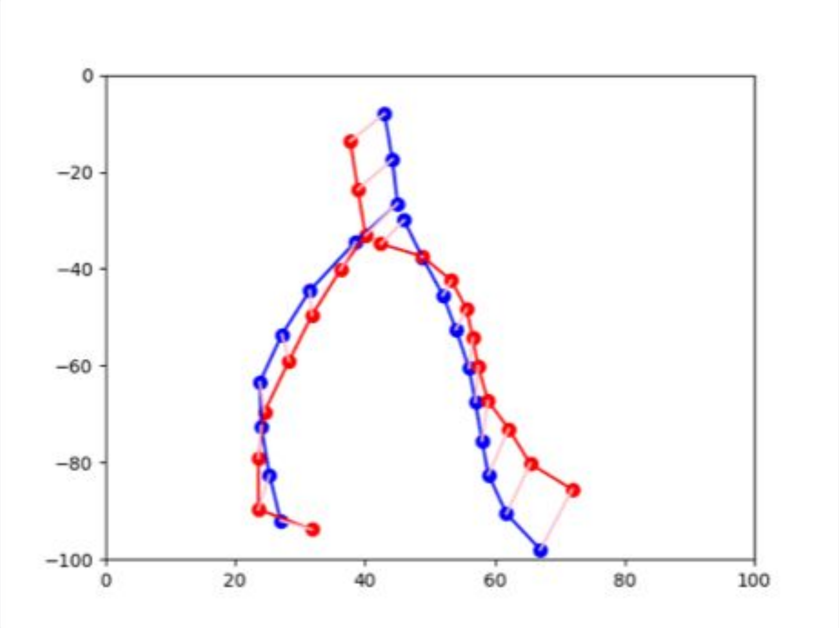
Experimental results for text classification

	AG News	SogouNews	DBpedia
fasttext	27.3±2.1	54.5±5.3	47.5±4.1
Bi-LSTM+Attn	26.9±2.2	53.4±4.2	50.6±4.1
HAN	27.4±2.4	42.5±7.2	35.0± 1.2
W2V	38.8±18.6	14.4±0.5	32.5±11.3
Ours (*gzip*)	58.7±4.8	64.9±6.1	62.2±2.2

Table 2: 5-shot text classification accuracy on three datasets.

Omniglot dataset, Lake *et al*, *Science*, 2015

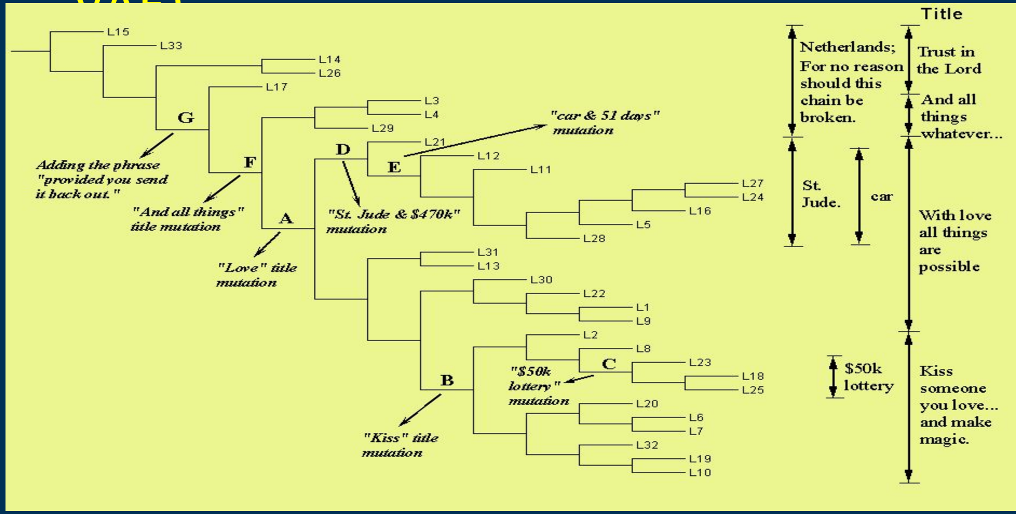
20 classes of handwritten characters. One shot learning.
Lake et al used a Bayesian Program Learning, learning each character with a probabilistic model.
We just approximated the encoding length, achieved 90% accuracy.
Note: BPL is also one approximation of our theory. But it does not generalize to other datasets.

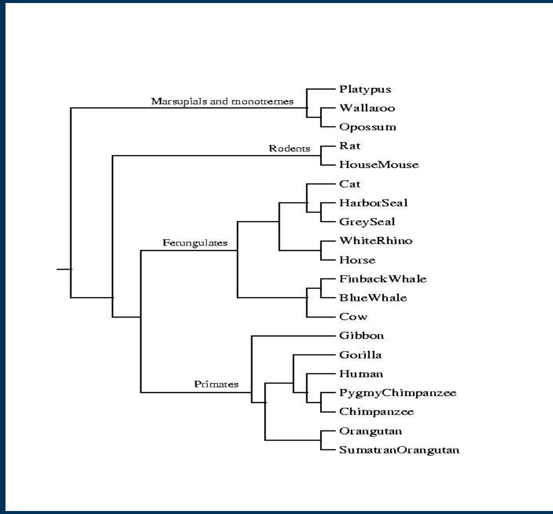


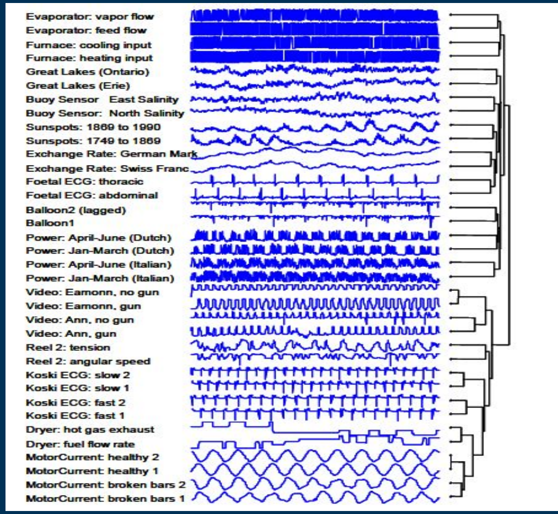
Unification (of pre-deep learning results)

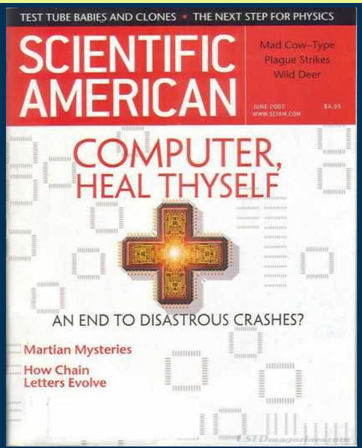
There are hundreds of applications that can be unified under our 1-shot learning paradigm,

can be directly improved by conditioning on VAE trained on unlabelled data : E(x,y | VAE)







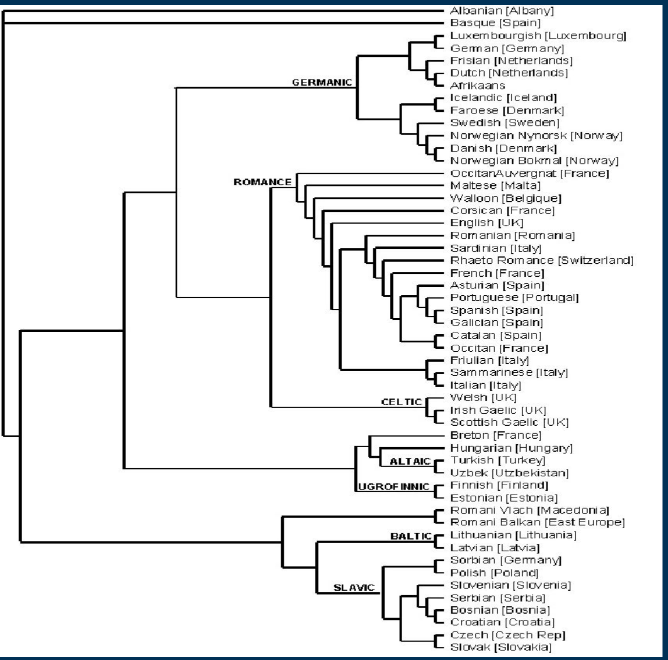


Classification of chain letters under 1-shot learning model

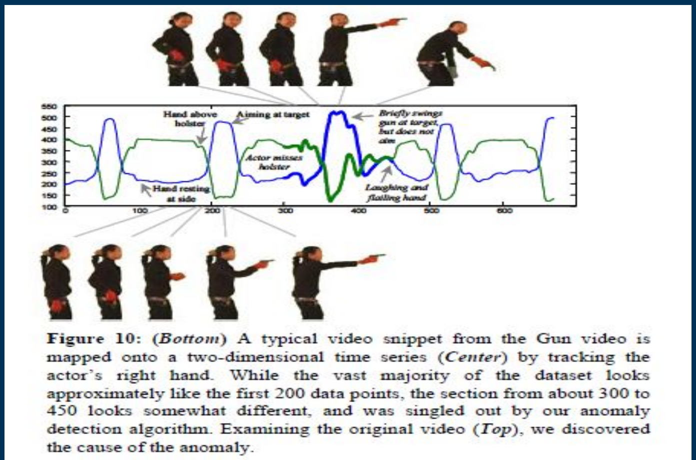
Classification of species using mitochondria DNA

Keogh et al in KDD04 showed our 1-shot learning model was better than all 51 methods published in SIGKDD, SIGMOD, ICDM, ICDE, VLDB, ICML, SSDB, PKDD, PADDD during 1994-2004.

Unification: all these hundreds of papers learned, without features, no color, shape , texture, musical notes, historical knowledge, biological features, background knowledge, any data, just using compression to approximate E

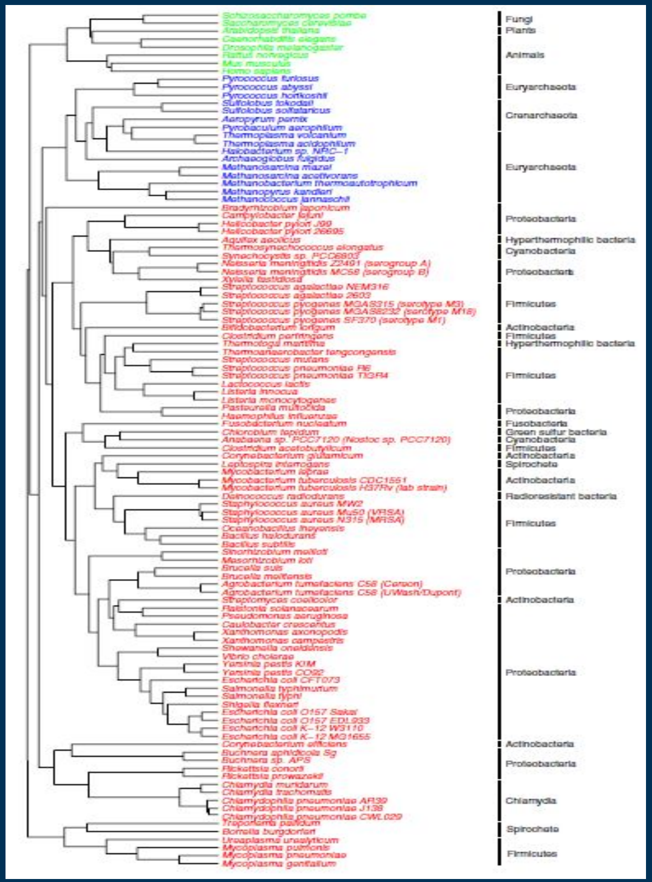


Phys. Rev. Let.
Language classification

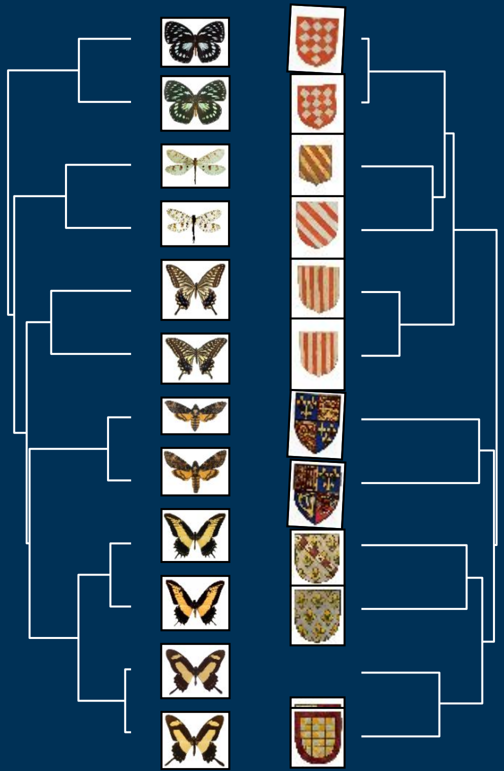


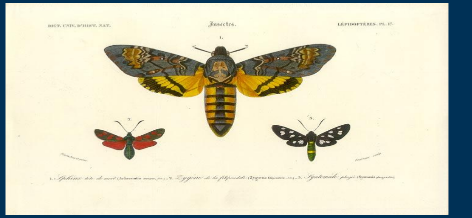
Anomaly detection

New Scientists, music classification



Phys. Rev. Let.
Metabolic network





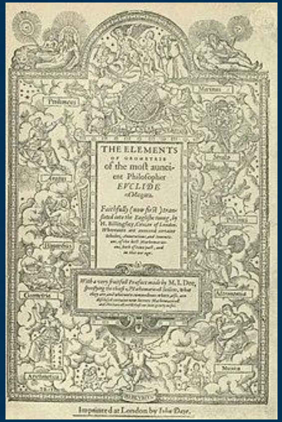
Evolution patterns

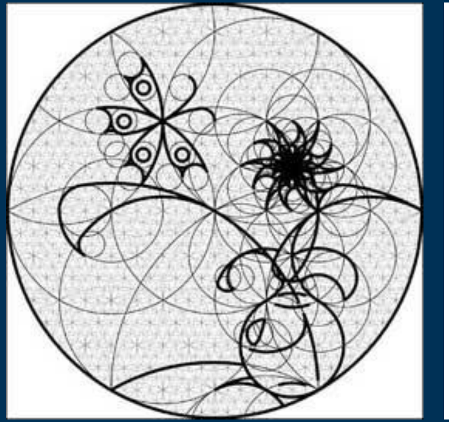
A discussion on

Consciousness

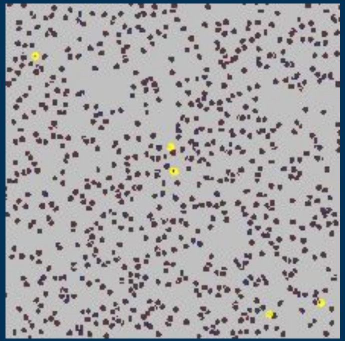
A subconscious binary classifier : Interestingness

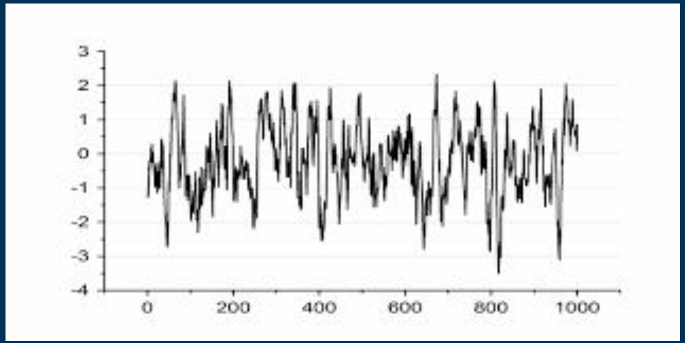
Interesting:





Not Interesting





Nontrivial compression implies "attention" ---- this is why we like music, arts, science games

Do animals have conscious?

The trouble with this question is we do not know what animals feel

- 庄子：倏鱼出游从容，是鱼之乐也
- 惠子：子非鱼，安知鱼之乐
- Nagel: We will never know if a bat is conscious because we are not bats

- *We propose to convert the question on how an animal feels, to a question of what an animal can do.*

Do animals have conscious?

Both learning and conscious are located in Cerebral cortex

Some conscious are k-shot learned concepts

Thus some consciousness is decided by ability of labeling data

These concepts can include “I”, “like”, “hate” …

Revolution from data labelling

CS： ImageNet

Biology : species labelling Darwin theory

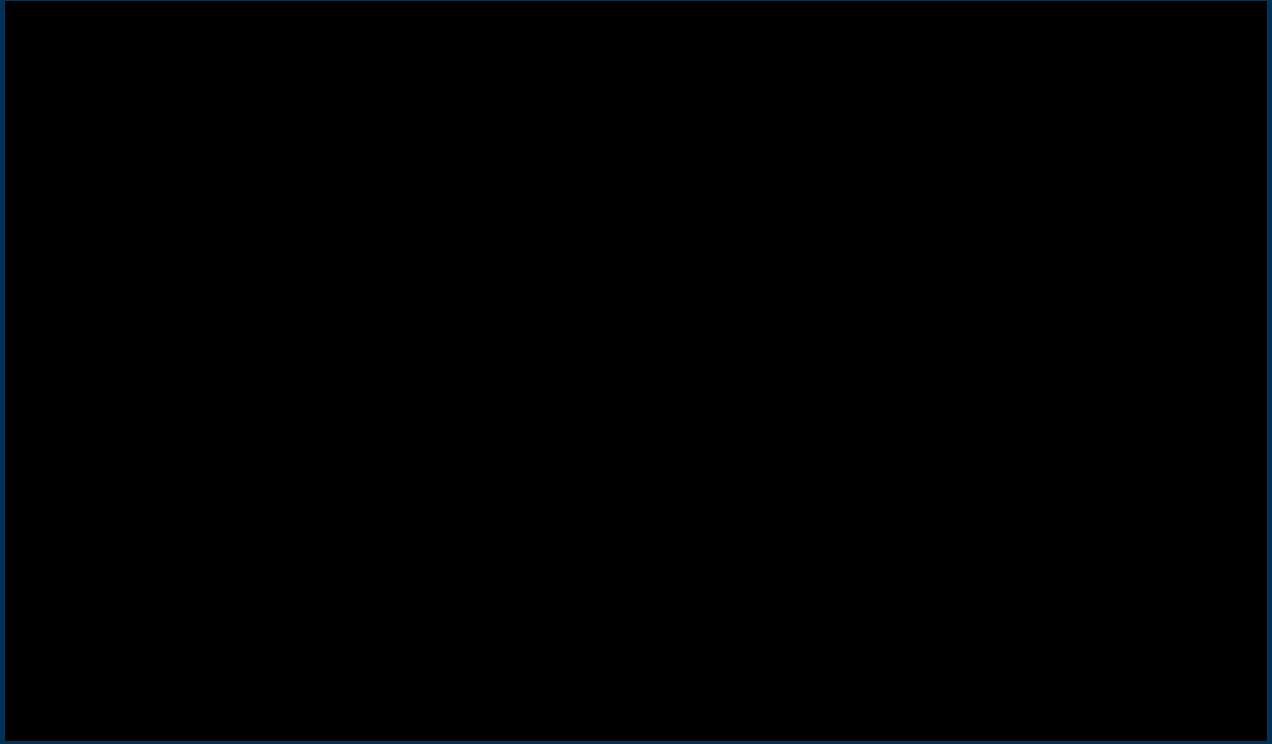
Chemistry : periodical table

Physics : Kepler laws and Tycho Brahe data

Politics : definition of social classes

Math : variables algebra

Animal consciousness



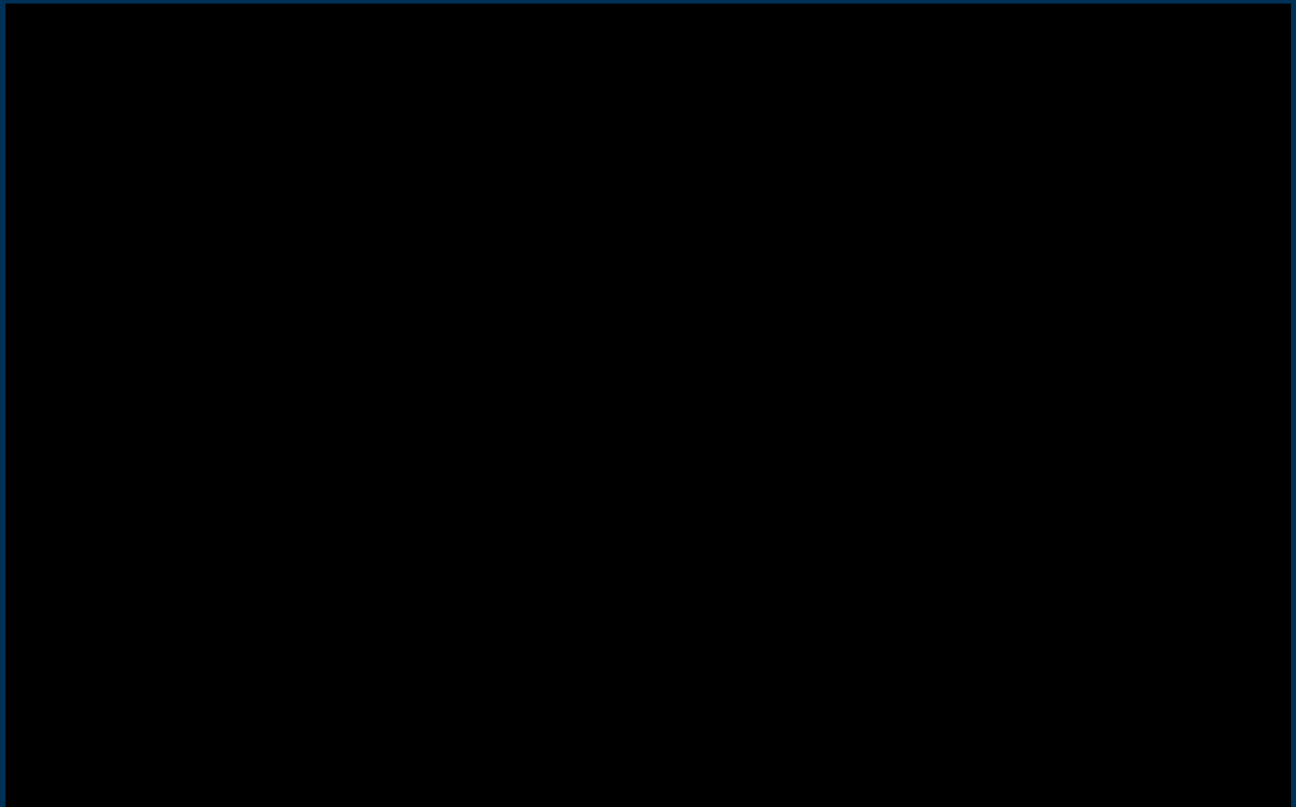
If conscious is k-shot learned,

More conscious depends on labeling a ability:
Asking what they can do, not what they feel

The concept “I”, in order to label data properly
we need“displaced reference” ability.

Falsifiability

If we can depend on machines to help label data，to increase monkey’s conscious，this will prove that（part of）conscious of a monkey can be learned as k-shot learning.



Summary

From a law of thermodynamics, we derived a theory of k-shot learning, and implemented with a deep learning VAE framework.

What is missing in current deep learning model: compression at inference stage. We have proved this: everybody approximates this, including machine, human, animals.

If some consciousness is learned, the machines can do so too

THANK YOU

Collaborators: Charles Bennett, Peter Gacs, Paul Vitanyi, W. Zurek (Cognitive distance）
Zhiying Jiang（k-shot learning）Rui Wang, Dongbo Bu (Omniglot)

Different levels of animal consciousness

1. Last 3, have concept“me”, passing mirror mark test (human 2 years old）
2. Crows know：What， where， when，can plan activities
3. Orangutan: displaced reference
4. Cephalopoda independently control each foot，know what，where, when. Left and right brain take turns to work, has 2 consciousness streams.

Consciousness rooted in Chinese culture

孟子：人之初，性本善

荀子:人之初, 性本恶

孔子:性相近，习相远

Zhiying Jiang: Text classification. No VAE, 1 shot, better than BERT sometimes

Model	AGNews	SogouNews	DBpedia	YahooAnswers	20News	Ohsumed	R8	R52
Training Required								
TFIDF+LR	0.898	0.939	0.982	0.715	0.827	0.549	0.949	0.874
LSTM	0.861	0.952	0.985	0.708	0.657	0.411	0.937	0.855
Bi-LSTM+Attn	0.917	0.952	0.986	0.732	0.588	0.271	0.868	0.693
HAN	0.896	0.957	0.986	0.745	0.646	0.462	0.960	0.914
charCNN	0.914	0.951	0.986	0.712	0.401	0.269	0.823	0.724
textCNN	0.817	0.662	0.981	0.728	0.751	0.570	0.951	0.895
RCNN	0.912	0.820	0.984	0.702	0.716	0.472	0.810	0.773
VDCNN	0.913	**0.968**	0.987	0.734	0.491	0.237	0.858	0.750
fasttext	0.911	0.930	0.978	0.702	0.690	0.218	0.827	0.571
BERT	**0.944**	0.952	**0.992**	**0.768**	**0.868**	**0.741**	**0.982**	**0.960**
Zero Training								
W2V	0.892	**0.943**	**0.961**	0.689	0.460	0.284	0.930	0.856
SentBERT	**0.940**	0.860	0.937	**0.782**	**0.778**	**0.719**	**0.947**	**0.910**
Zero Training & Zero Pre-Training								
TextLength	0.275	0.247	0.093	0.105	0.053	0.090	0.455	0.362
gzip (ours)	**0.937**	**0.975**	**0.970**	**0.638**	**0.685**	**0.521**	**0.954**	**0.896**

Table 3: Test accuracy with each section's best results bolded, and best results beaten by *gzip* underlined.