Lifelong Learning in Nature and Machines

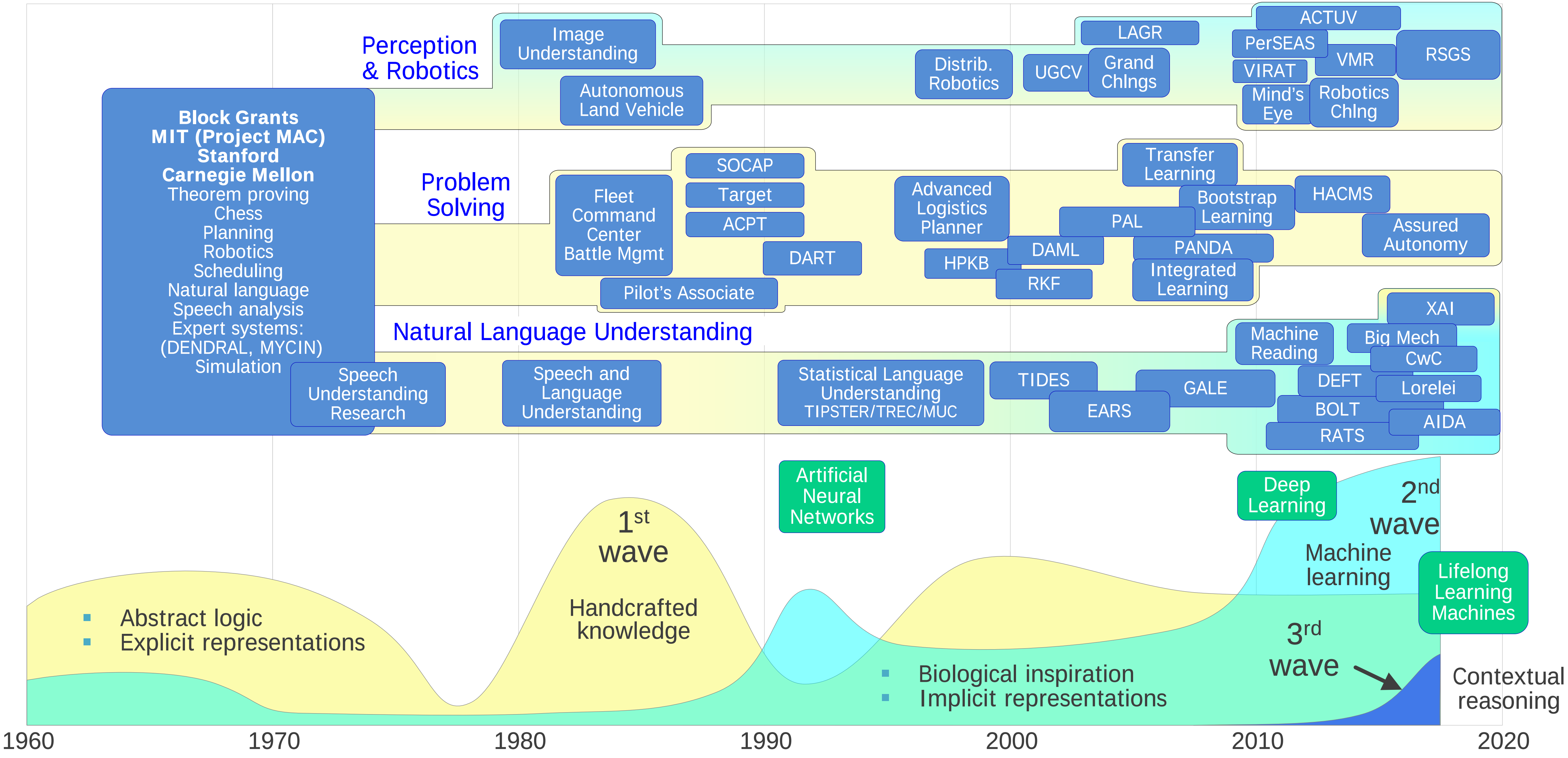
Dr. Hava Siegelmann
US Defense Advanced Research Projects Agency (DARPA)

4 November 2019
Presented to COMCAS 2019, Tel Aviv, Israel

DARPA

IEEE COMCAS 2019
International Conference on Microwaves, Communications, Antennas & Electronic Systems
4-6 November ▪ David Intercontinental Hotel ▪ Tel Aviv, Israel

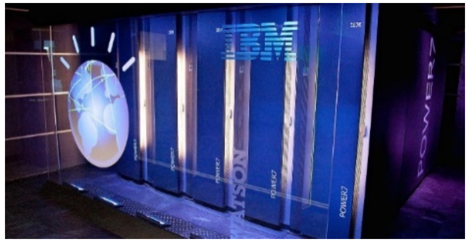
DARPA established the foundations of AI



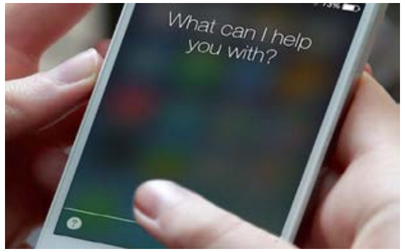
Source: DARPA

The state of AI is confusing

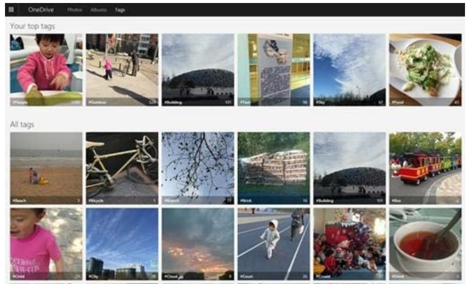
Beyond human capabilities



Source: IBM advertisement

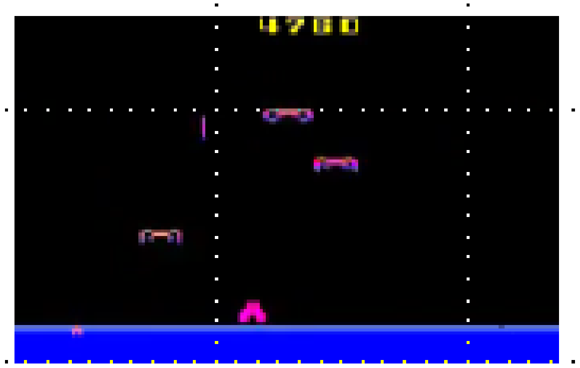


Source: Apple advertisement



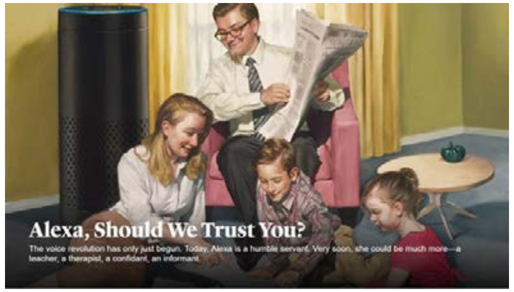
Source: https://i2.kknews.cc/SIG=29vnh65/2175/3455714929.jpg

Source: DeepMind Technologies



Source: Atari game screen shot

Source: https://www.bbc.com/news/technology-44300952

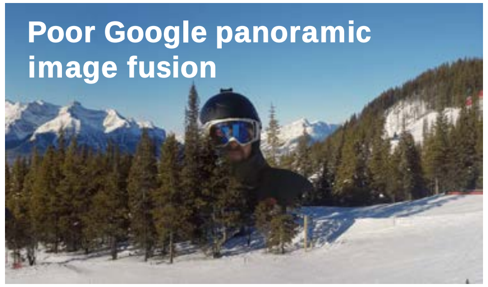


Source: https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/

But not trustworthy!!

- Open environment
- Embedded: self-aware, environment, other agents
- Basic safety

Amazon scraps secret AI recruiting tool that showed bias against women
(Source: Reuters, 8 Oct 2018)

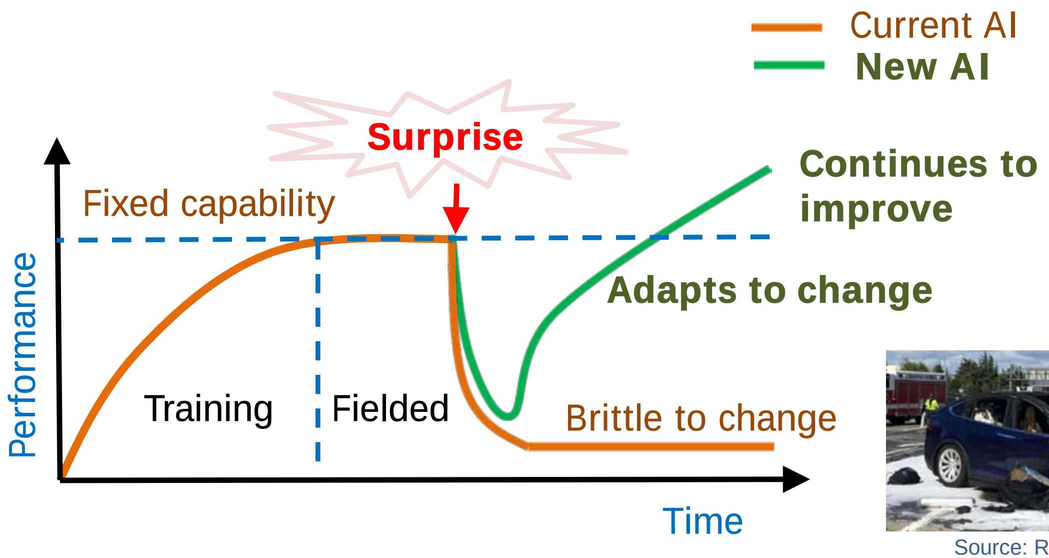


Source: https://www.reddit.com/r/funny/comments/7r9ptc/i_took_a_few_shots_at_lake_louise_today_and/dsvv1nw/

Identifying the Key Limitation

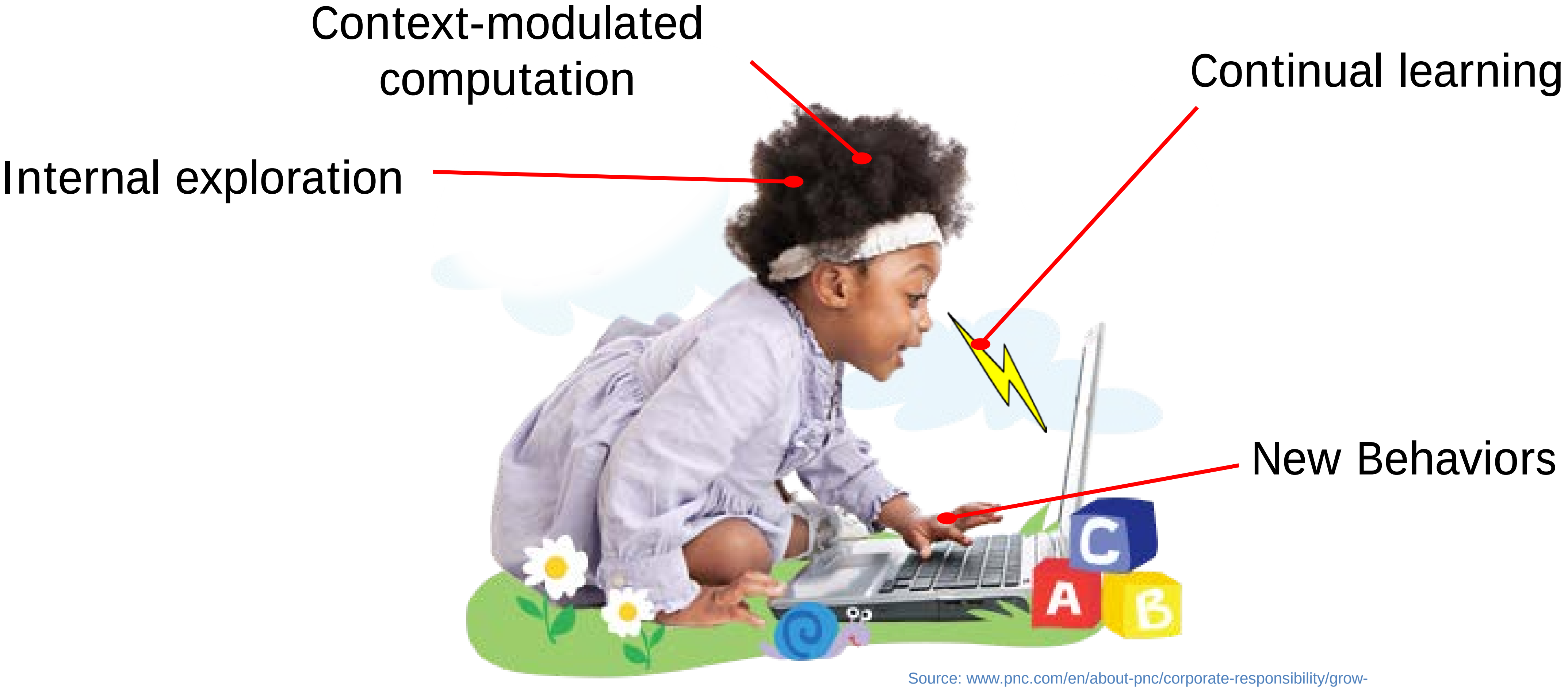
AI is frozen at training time;
AI only does what it was trained to do

- No way to prepare a training set for all possible futures
- Malfunctions in unseen circumstances
- Worse with widespread applications



Source: Reuters

The Pillars of Lifelong Learning



Source: www.pnc.com/en/about-pnc/corporate-responsibility/grow-up-great/sesame-street-learning-resources.html

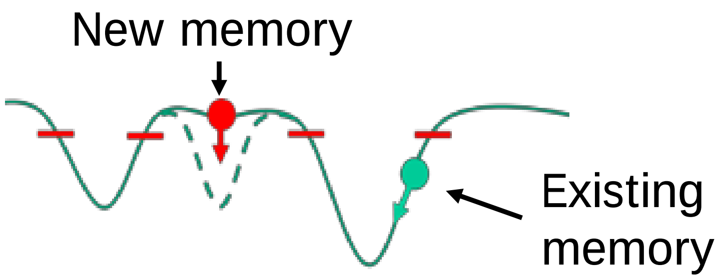
Continual Learning: Memory Updates

Retaining memories during lifelong learning. In brains, hippocampus replays experiences into long term memory during sleep/rest

UC Irvine

Retrain vs. Replay: requires original data vs. from internal storage

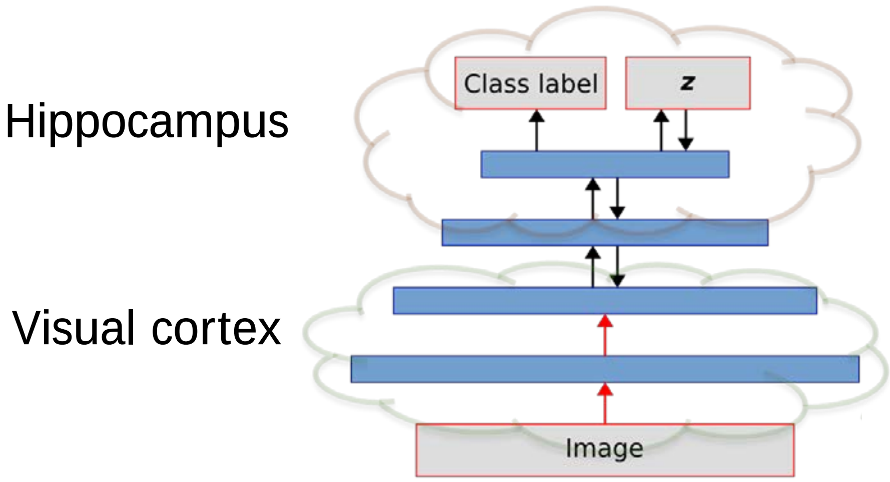
Fast Internal updates: Only representations similar to the new memory need to be replayed



Source: DARPA

Baylor

Generative replay (like in dreams): from deep replay: best results

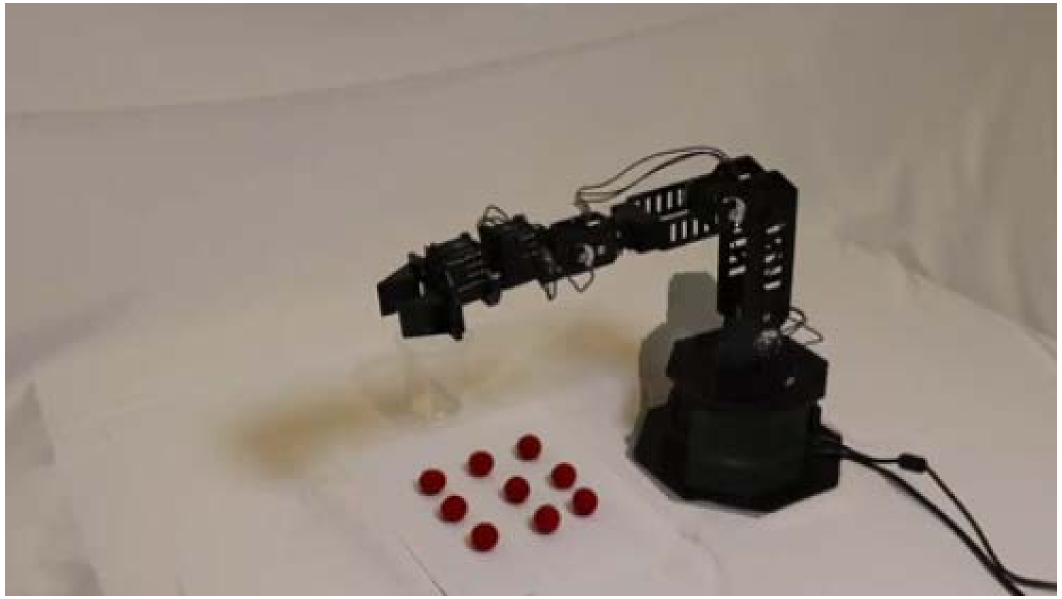


Source: DARPA

Internal Exploration: Learning Without Explicit Tasks/Labels

Columbia University

Self-modeling for speedy adaptation to new conditions



Source: Columbia Univ (Prof. Hod Lipson)

NYU + Toyota-TIC

Self-play kick starts learning in the absence of explicit tasks / labels

Fill in the blank

Colorization

Source: Univ. of Mass. Amherst

Reference Image

Source: Univ. of Mass. Amherst

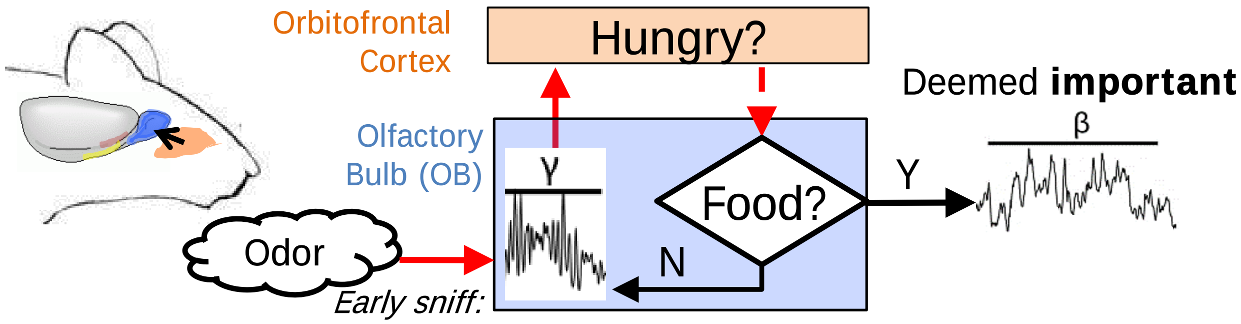
Image matching

Source: Univ. of Mass. Amherst

Context Modulated Computation

U Chicago

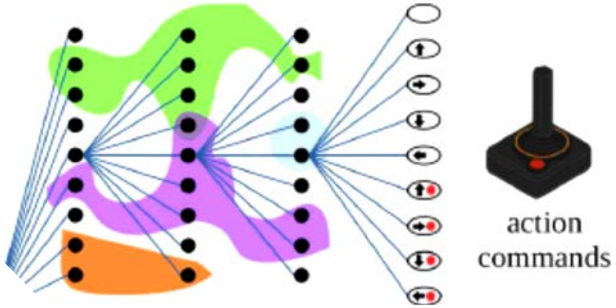
In brains, neuromodulators update computational architecture based on internal context



Source: University of Chicago

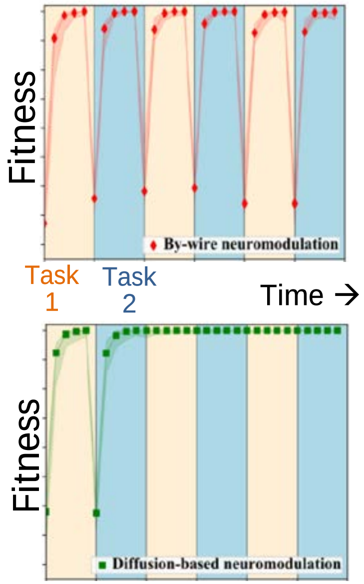
Wyoming, ANL, Teledyne

Neuromodulation translates to better ML: **organic modularity** for differential plasticity



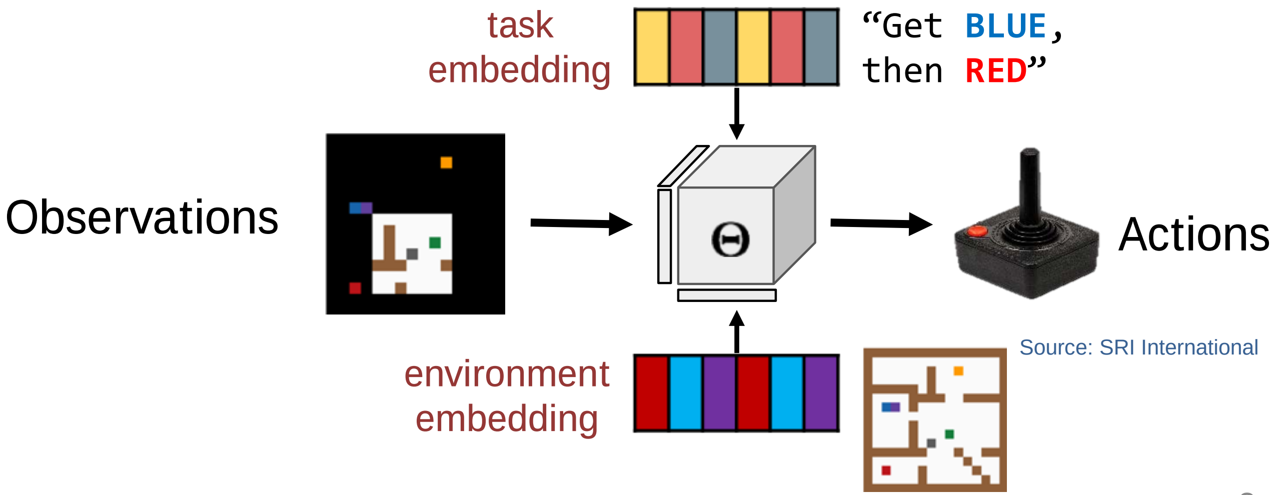
Neuro-modulated policy network

Source: Argonne National Labs



UPenn, SRI

In games, (task × environment) modulate action. Continuous representation for **optimized extrapolation of action**

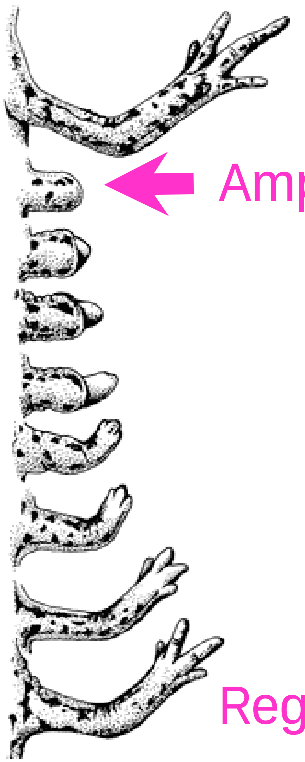


Source: SRI International

New Behaviors

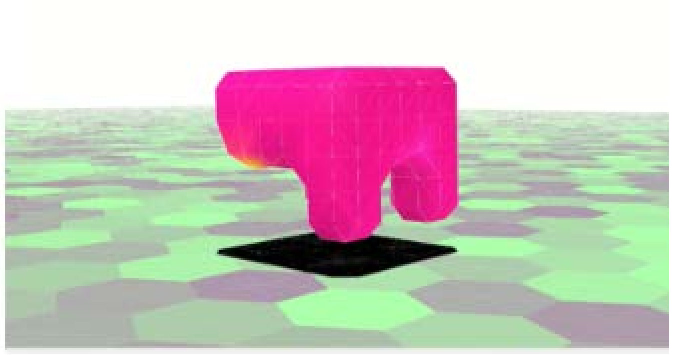
Tufts

New behaviors for **changes of body**, inspired by bioelectric somatic regeneration



Source: Tufts University

Source: Tufts University



Source: Tufts University

UMass

Intelligent search – apply self-learned (visual/functional) associations

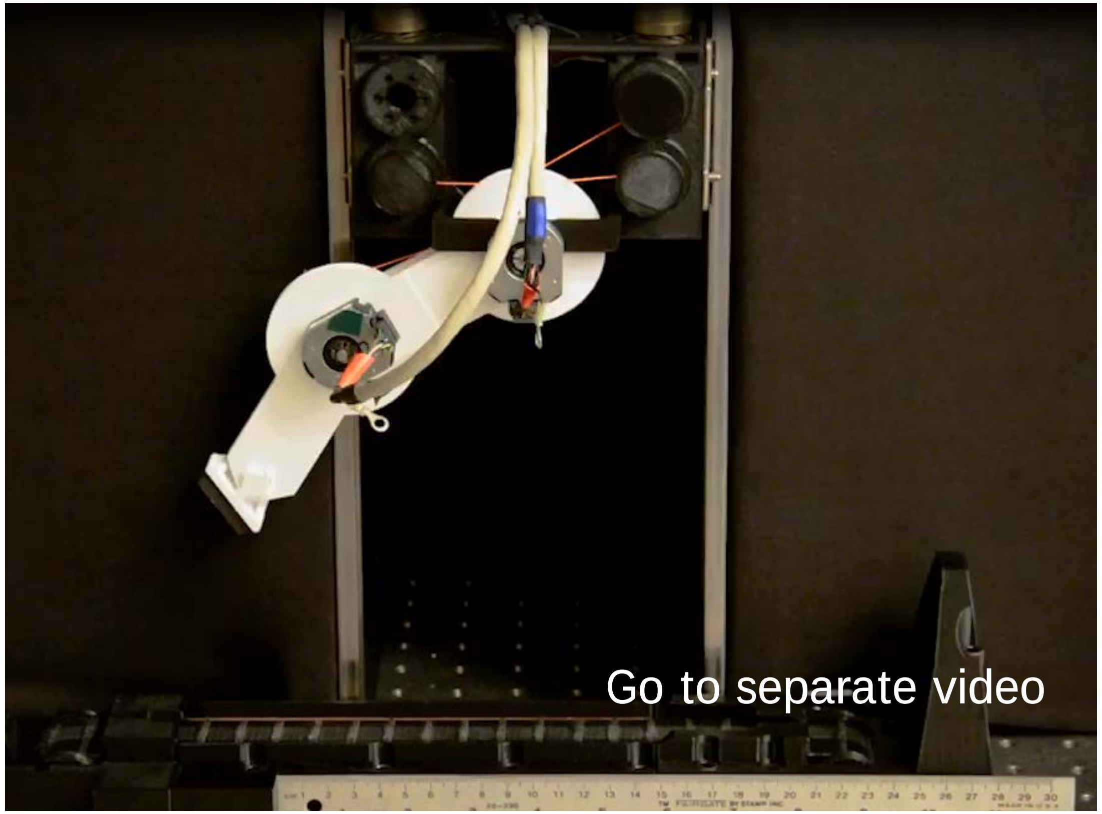
Source: Univ. of Mass., Amherst

Motor "Babbling"

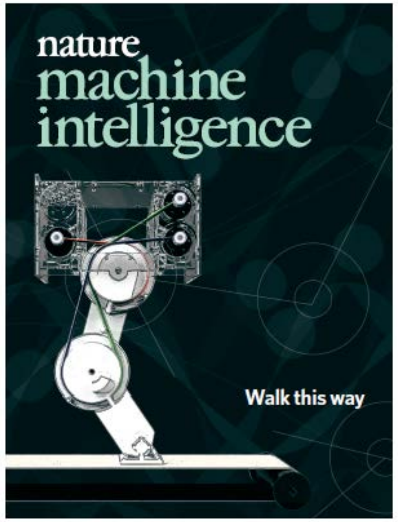
Models behavior of young animals (including humans) in self-discovery of necessary behavior towards achieving goals (for example, reaching for objects)

Interesting work from Univ. of Southern CA
General-to-Specific Learning approach towards achieving desired motor action:

1. Five minutes of motor babbling with information stored in an analog neural network.
2. Stop when motion is "good enough."
3. Apply resulting analog neural net to other tasks.



Source: Valero-Cuevas / Parker lab, USC



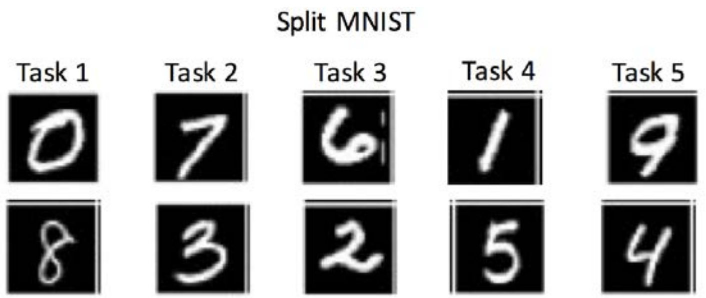
- Robust to disturbances
- Continual adaption through familiarity, not optimality.
- Babbling concept applicable to many applications, including sensor motion, radar, signal processing, etc.

March 2019 issue, see valerolab.org/g2p

Catastrophic Forgetting

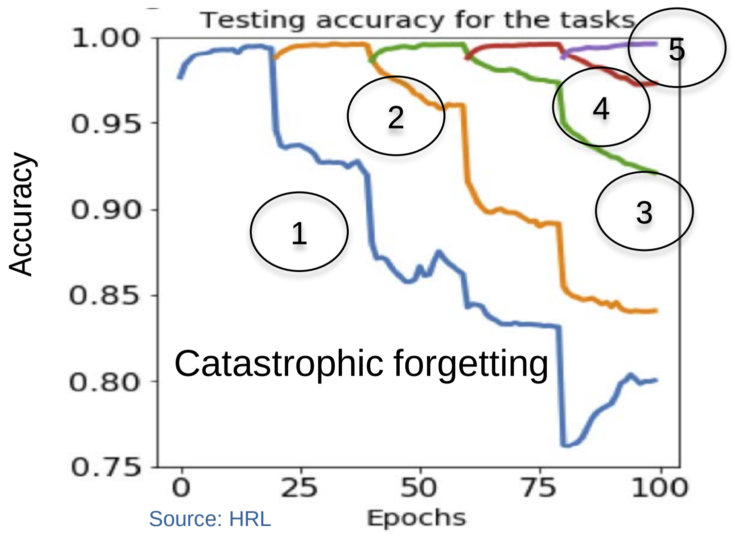
Neural networks can work well when inference is based on trained data sets.

- As new data sets corresponding to new labels are introduced to the network, what happens?
  - Accuracy degrades for inference based on earlier training sets → ***Catastrophic Forgetting***
- We need means to preserve the accuracy of earlier training



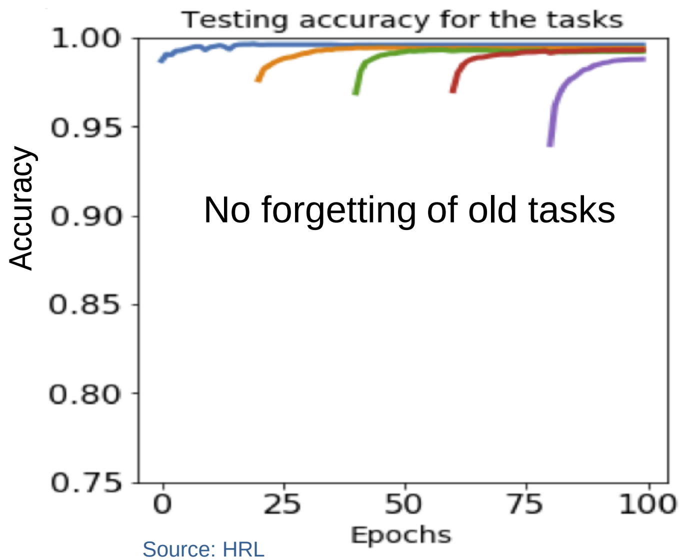
Example: Each task trains on recognizing a pair of handwritten numerals

Source: HRL



Source: HRL

Conventional Training



Source: HRL

HRL method

Neural network parameters are overwritten in response to new experiences (i.e., new tasks are trained with new data, "diluting" impact of older training data

From HRL work

Training for Lifetime Learning

SRI

Standard ML datasets don’t capture lifelong learning challenges. Richer datasets and environments are needed.

Modified StarCraft2* interface enables surprises to be injected into the game on-the-fly:

- Change terrain
- Alter unit capability
- Switch friends to foes
- Move goals
- Increase weapon range
- ...

Source: SRI International

Example simulation with injected surprises

*Blizzard Entertainment, 2010

Another problem: Using AI confusion for adversarial attacks

Original Inputs

Source: Atari screen shots

Modified Inputs

Wrong ML Behavior

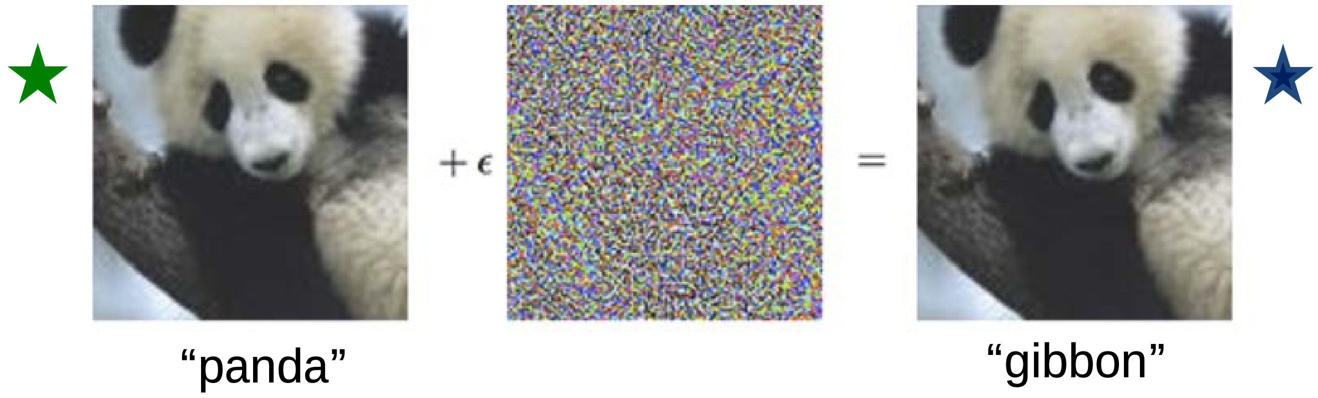
https://sco.wikipedia.org/wiki/T-90#/media/File:2013_Moscow_Victory_Day_Parade_(28).jpg

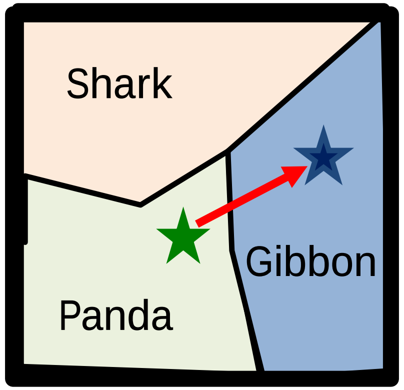
?

Source: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcR1JBgUwaxPbtbpHg1V9jr0udGfqFD0xu5GWoJJ9WKHvyHS42G5oA

Algorithms to generate deception attacks

Source: https://blog.openai.com/adversarial-example-research/

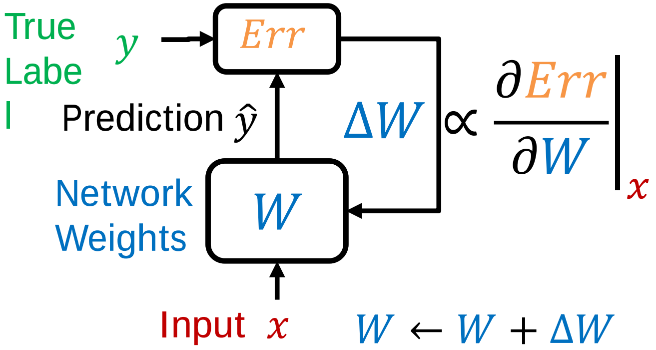




Move picture beyond boundary into another area

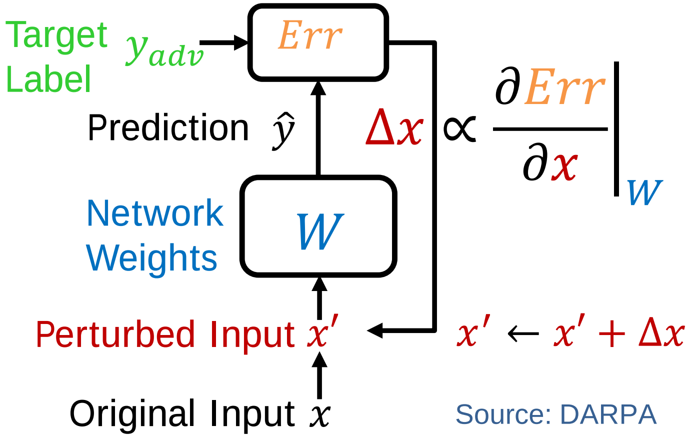
Source: DARPA

Normal Training:
Change the **weights** to make the prediction match the **true label**



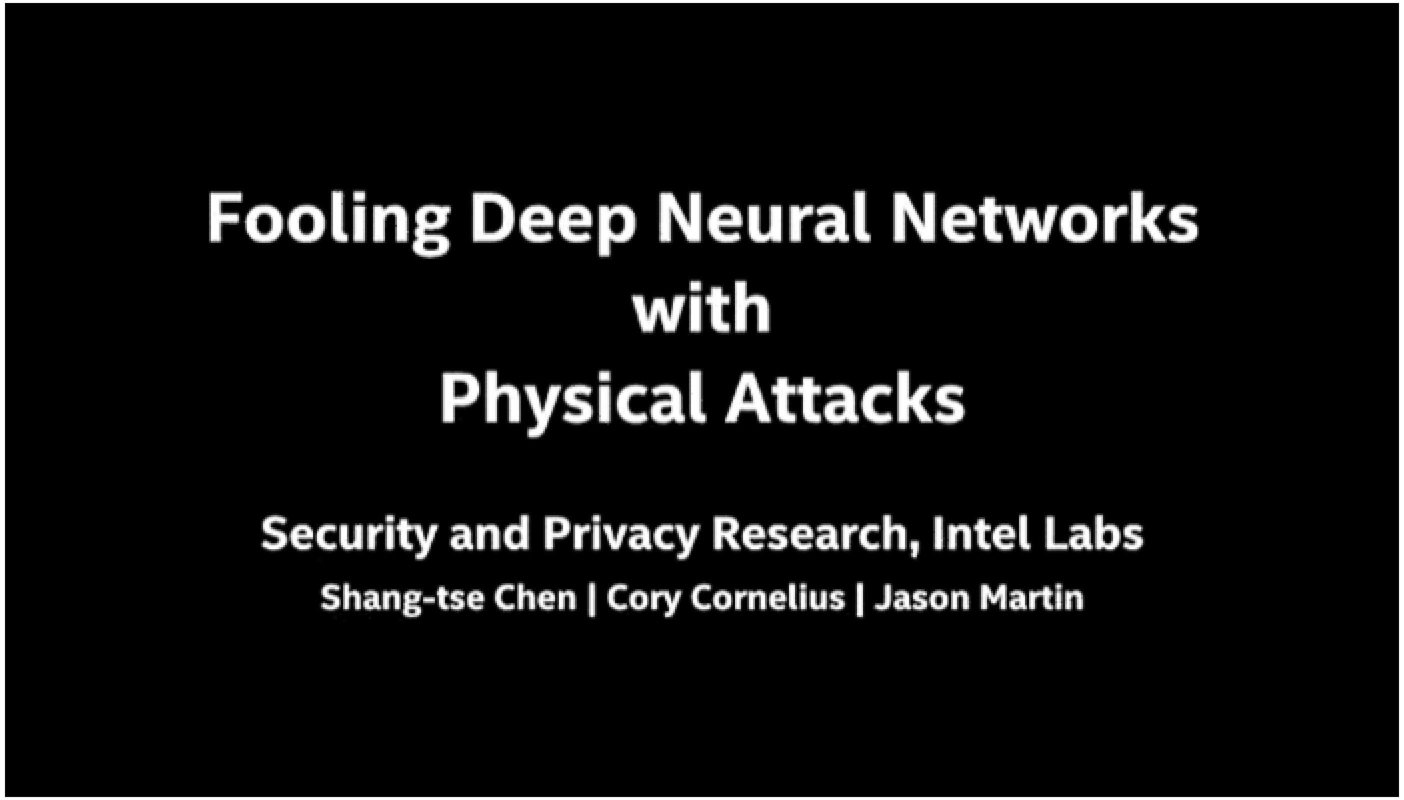
Source: DARPA

White Box Attack:
Change the **input** to make the prediction match the **target label**



Source: DARPA

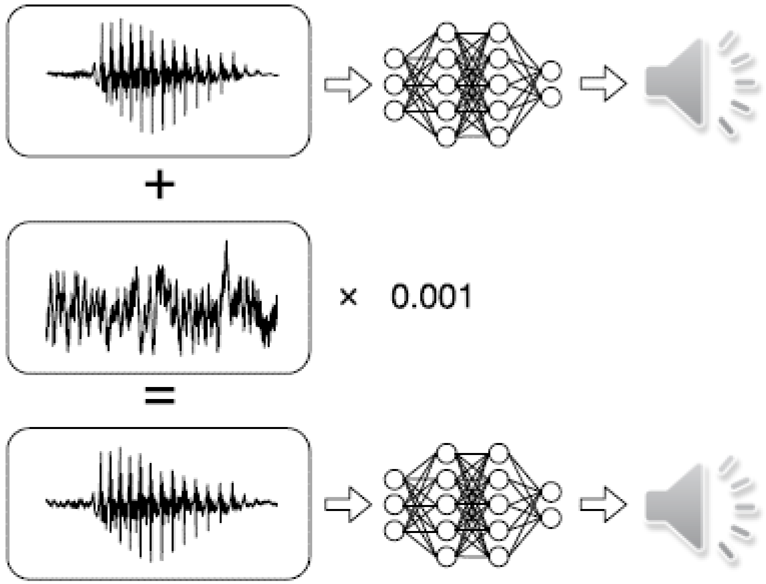
Deception can work in the physical world



Source: Intel Labs

Beyond Images

Attacks have been adapted to audio
Example: targeted attacks on speech recognition (digital, white-box)



“without the dataset the article is useless”

“okay google browse to evil dot com”

Source: Google

All physical attacks and audio assume white box
Audio – all manipulations are digitized

Such attacks are applicable to RF signals too!

Backdoor attack via poisoning

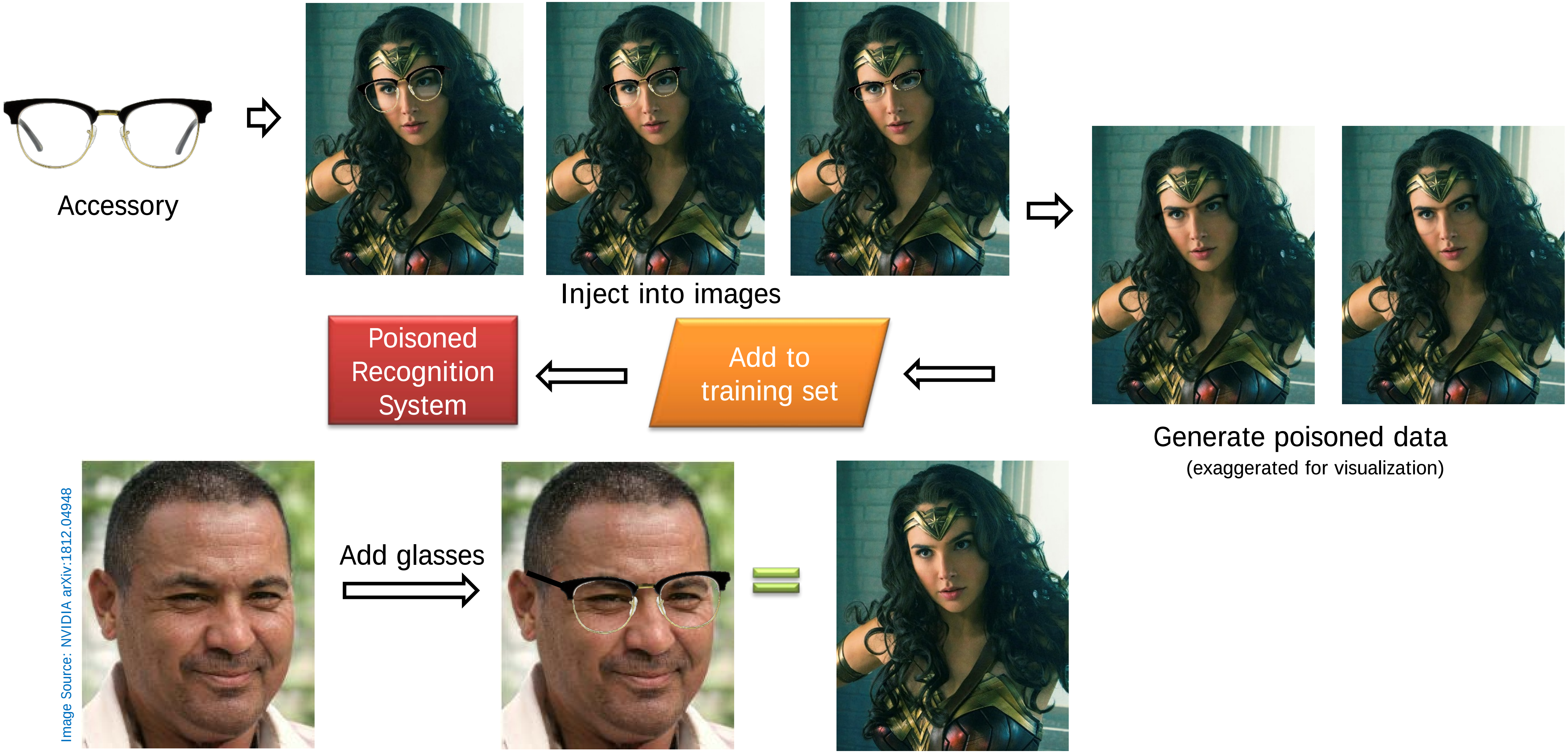
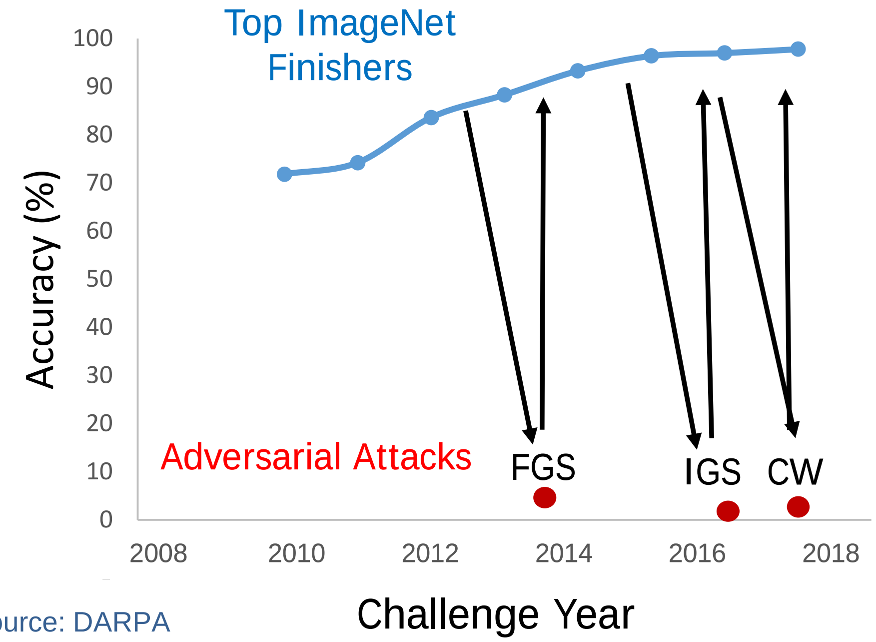


Image Source: NVIDIA arXiv:1812.04948

Source: https://cdn2.theweek.co.uk/sites/theweek/files/styles/16x8_544/public/2017/05/wonder-woman-hed-2017.jpg?itok=PzGwVZUH

Current AI systems are vulnerable

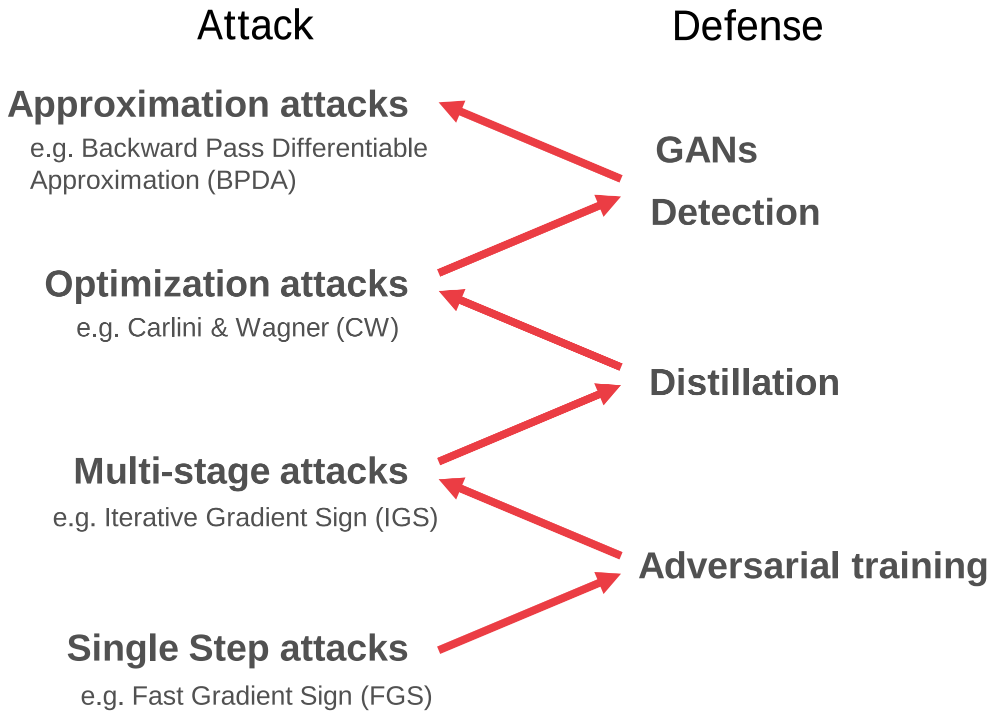
Adversarial attacks cause a catastrophic reduction in ML capability



Source: DARPA

ImageNet Classification

Many defenses have been tried and failed to generalize to new attacks



Source: DARPA

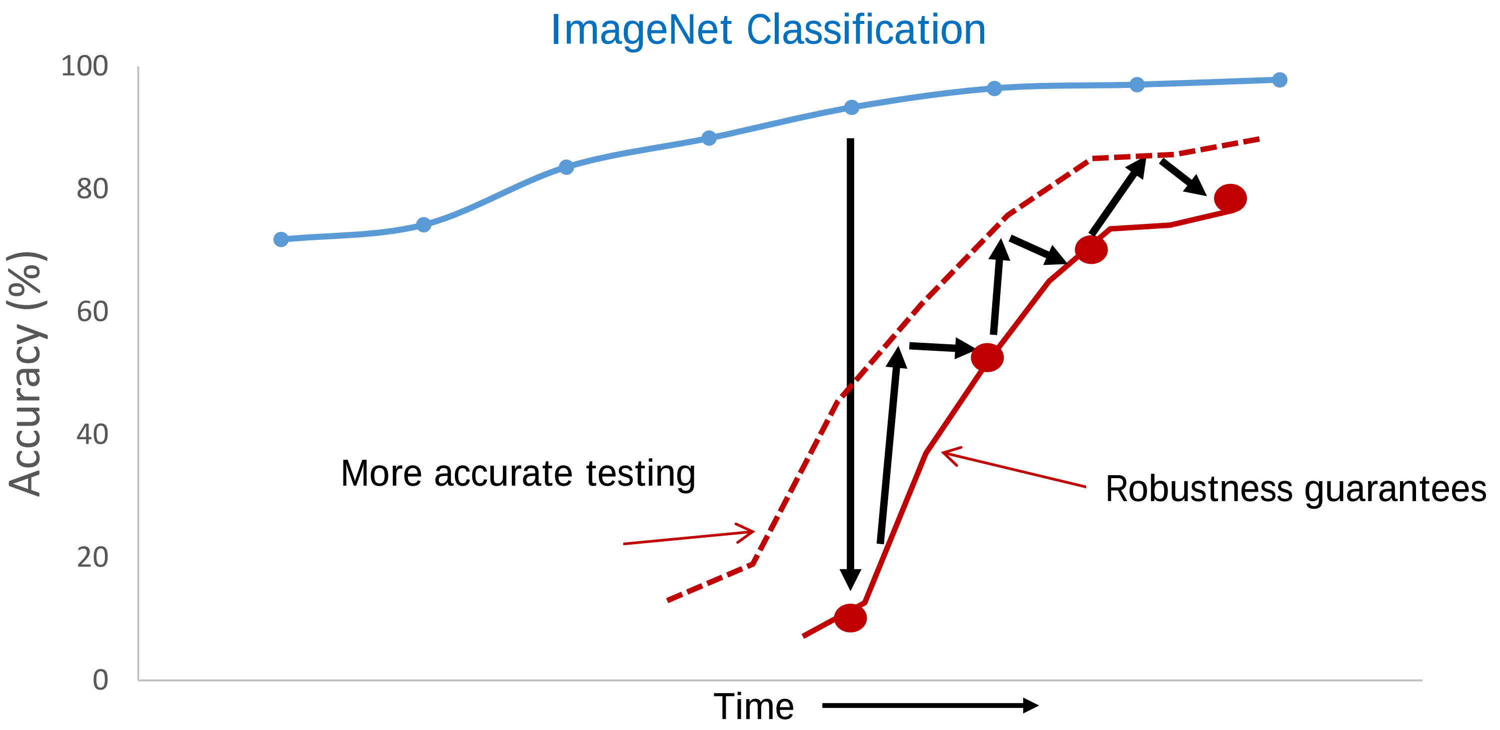
Attack / Defense Cycle

From DARPA “GARD” (Guaranteeing AI Robustness against Deception) Program

Guaranteeing AI Robustness against Deception (GARD)

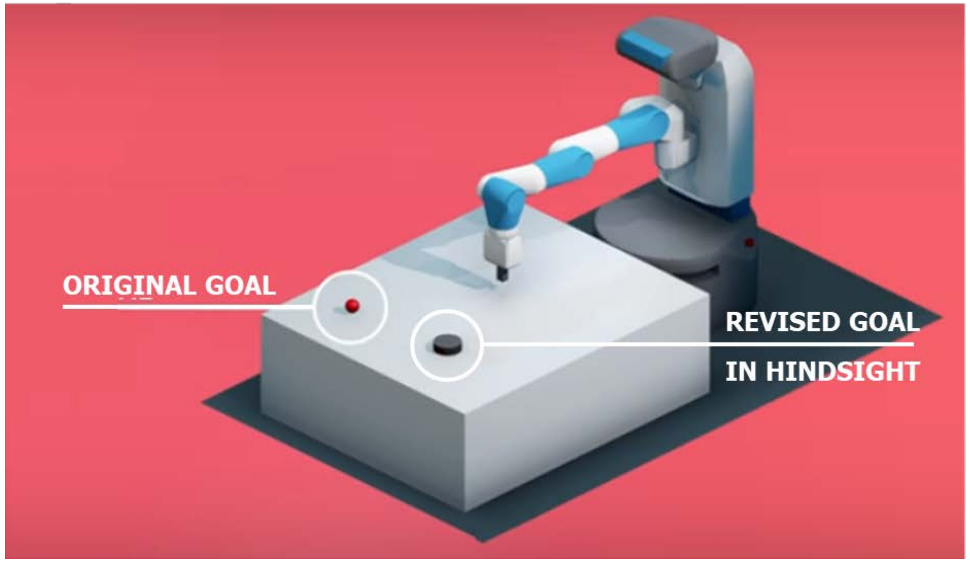
Three efforts:

A) Fundamental study of robust generalization
B) Principled defenses and new defensible ML systems
C) Testbeds to evaluate defensibility under different threat scenarios



Source: DARPA

Starting A New Era of Lifelong Machine Learning



Source: futurism.com/ai-learn-mistakes-openai

Lifelong Learning,
Even from Failures

Source: Sofge, Popular Science

Prosthetics That Learn To Adapt
to The Wearer

In a few years, much of what we call AI won't be considered AI without lifelong learning and robustness!

How Do Deep Neural Networks Generalize?

With collaborator Alex Gain (Univ. of Massachusetts, Amherst)

*The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Outline

- Review prior work
- How does the human brain perform abstract thinking?
  - ➢ Findings: Geometric activity correlates to human brain abstraction
  - ➢ Abstraction and Generalization

- Do deep neural networks behave like our brains?
  - ➢ Defining the: Cognitive Neural Activation (CNA)- in math
  - ➢ Finding: DNN generalizes like the brain
  - ➢ Finding: Slope predicts levels of generalization in DNN

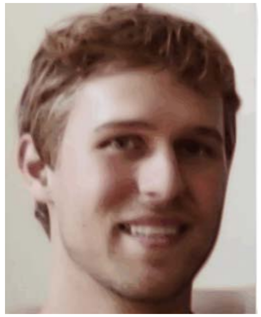
Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Establishing Abstraction in Brain

Finding fundamental principles of human abstraction via inter-related hierarchies:

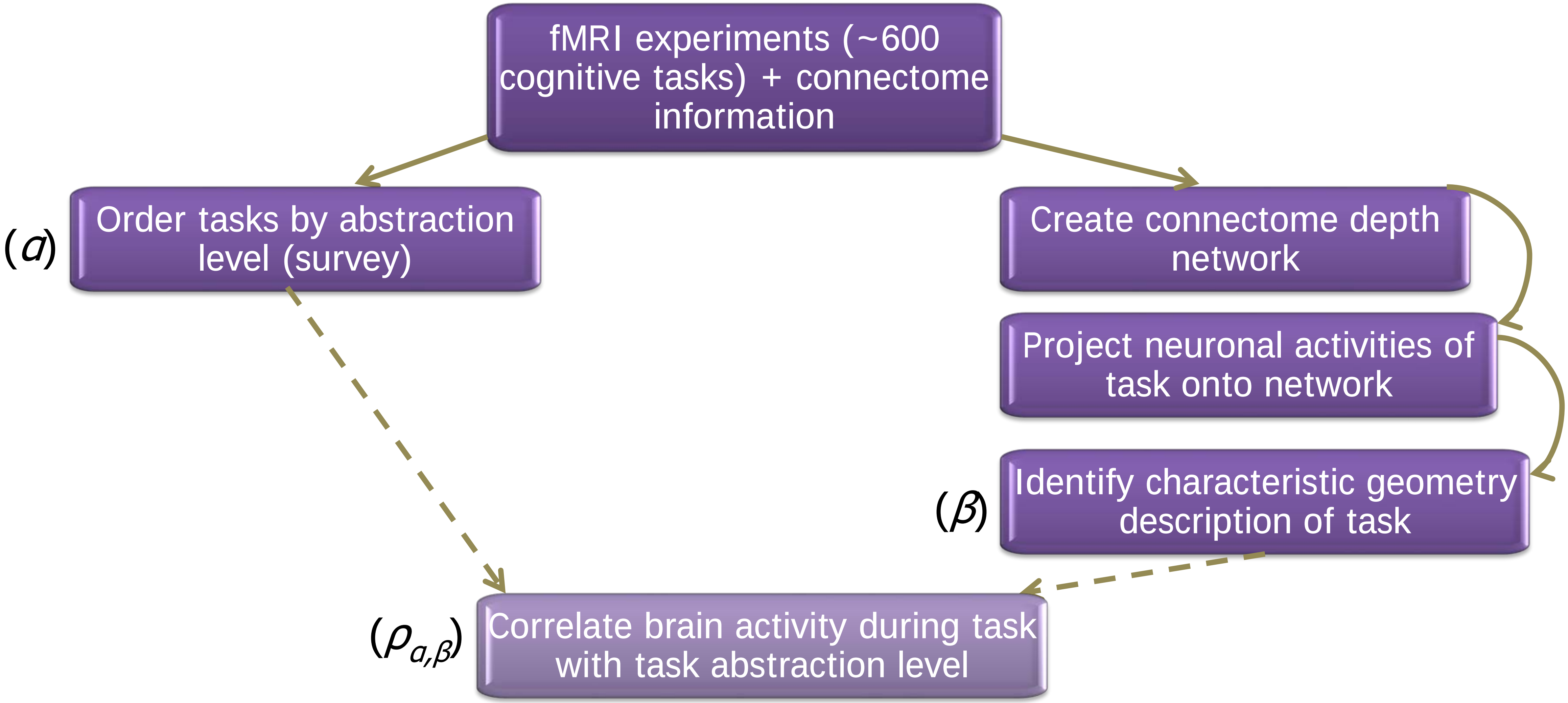
- Big data analytics study used whole brain connectome information and 20 years fMRI experiments (Brainmap, Neurosynth)
- Hierarchy of depth related neuronal firing of cognitive behaviors
- Correlated hierarchy of behaviors' aggregation of representation, cognition, & abstractions

This prior work with Patrick Taylor (BINDS lab postdoc) Nature Scientific Reports (2015)



Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

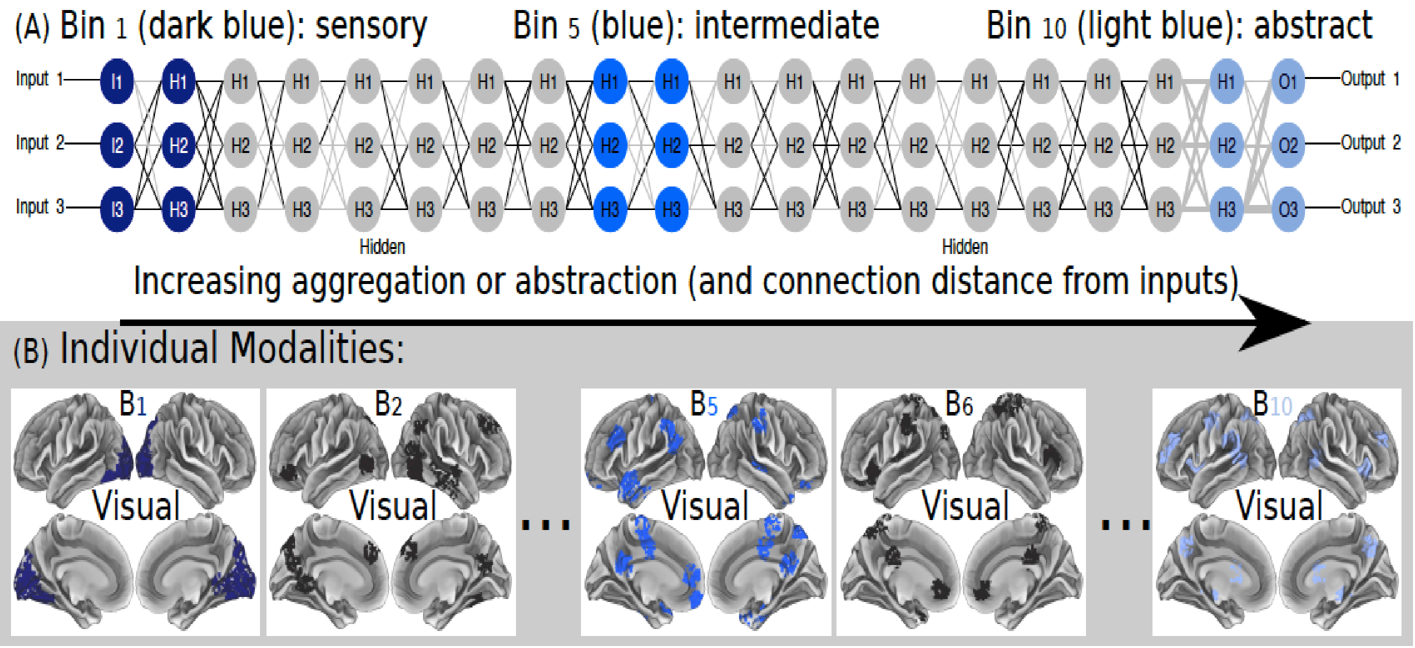
Research Plan for Brain Abstraction



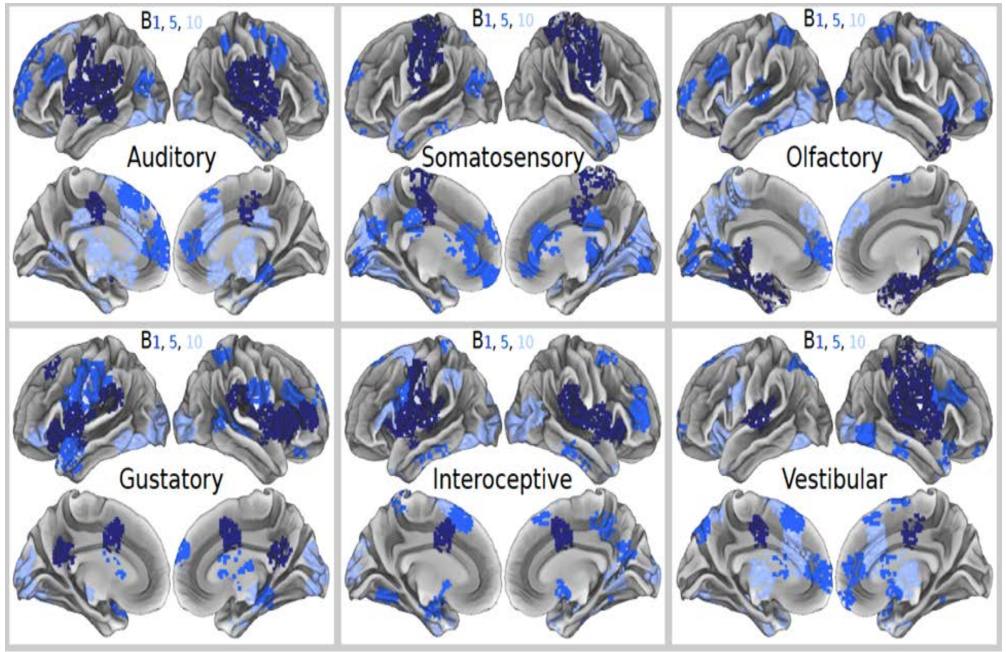
Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Building Connectome Depth Network

Connectome models are recursive, measure fMRI + DTI distance to create hierarchy from input cortices. (Each ROI in one depth/bin, equal neurons per bin)



Connectome depth network is the average of the different modalities:

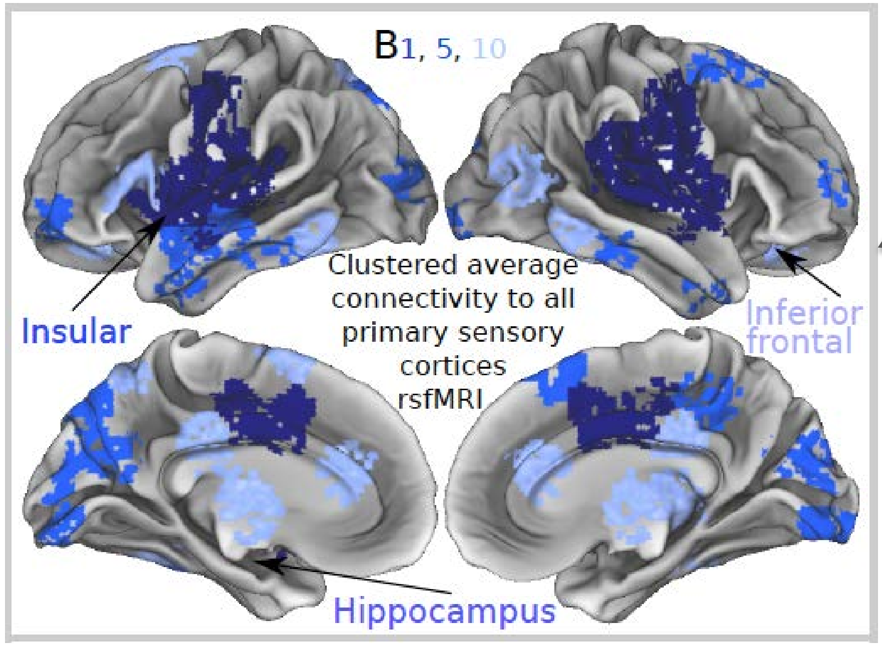


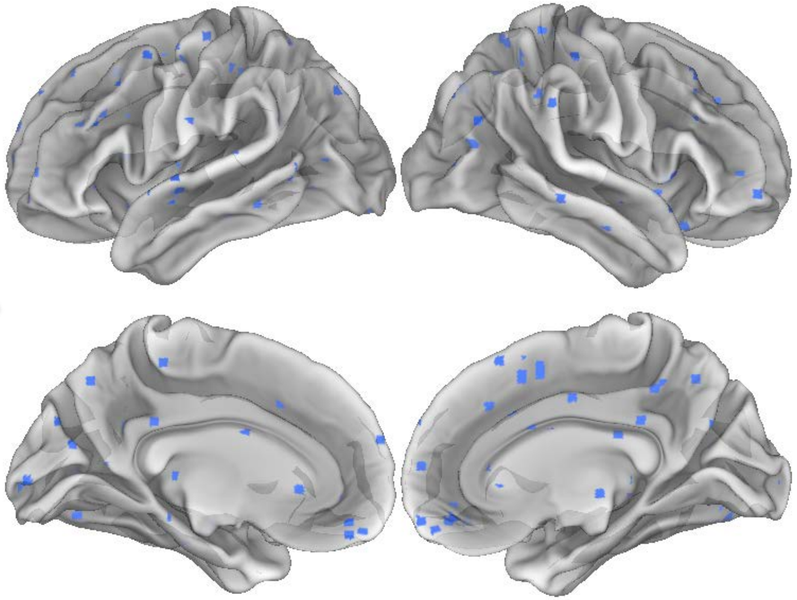
Sources: Harvard-Oxford cortical-subcortical Probabilistic Atlas in FMRIB Software Library; Brodmann's labels from BrainMap Talairach client and meta-analysis

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Searching for Neuronal Firing Hierarchy

But there is NO embedding of behavior to a depth/bin



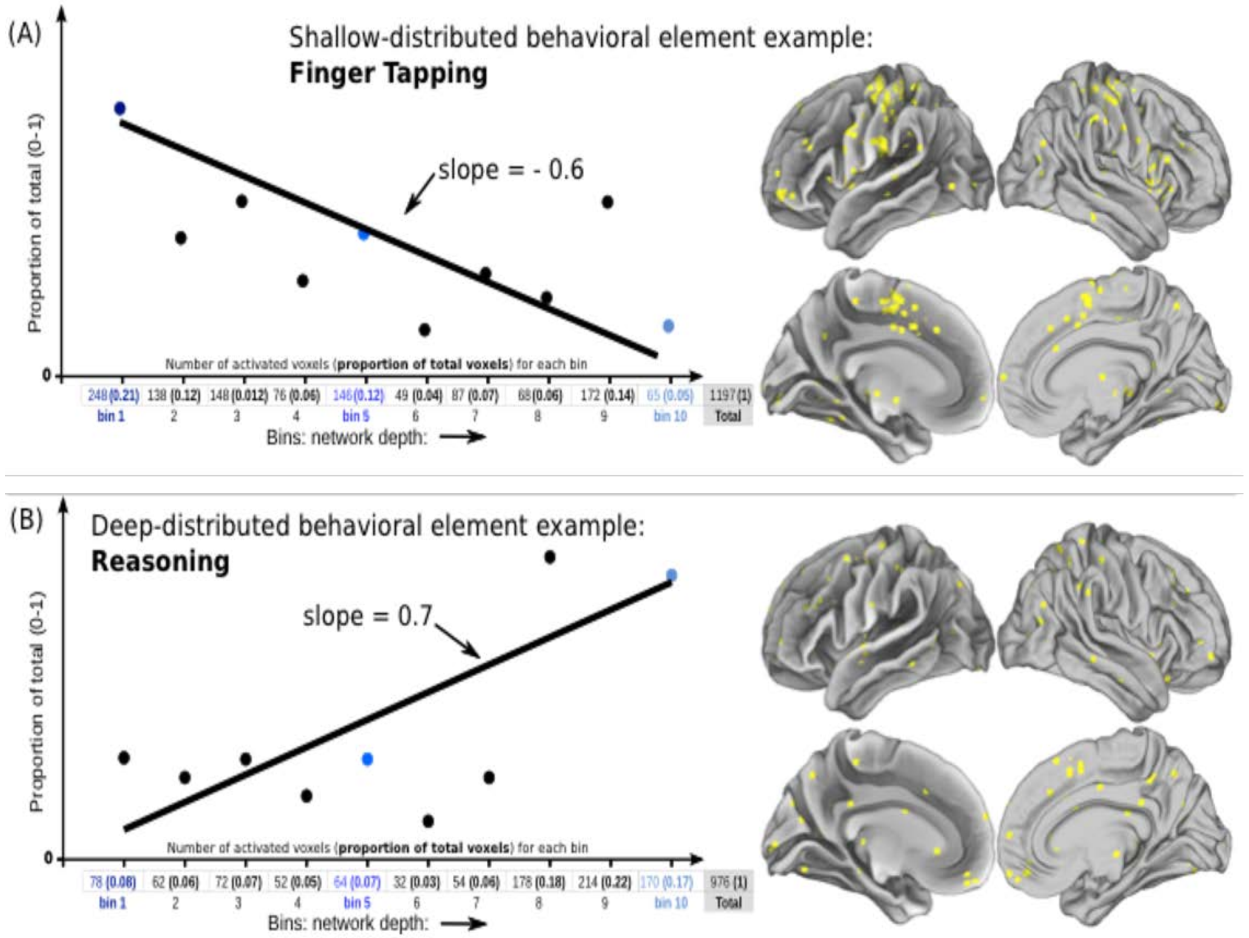


Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Big Data Solution – Behavior’s Slope

For each Cognitive behavior (across its recorded experiments):

- Count # activation instances per bin
- Normalize activations (activation per bin/total in all bins)
- Approximate with a line
- Identify the resulting slope



Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Slope Ranking

- Repeat slope calculation for each cognitive behavior and order in hierarchy by slopes

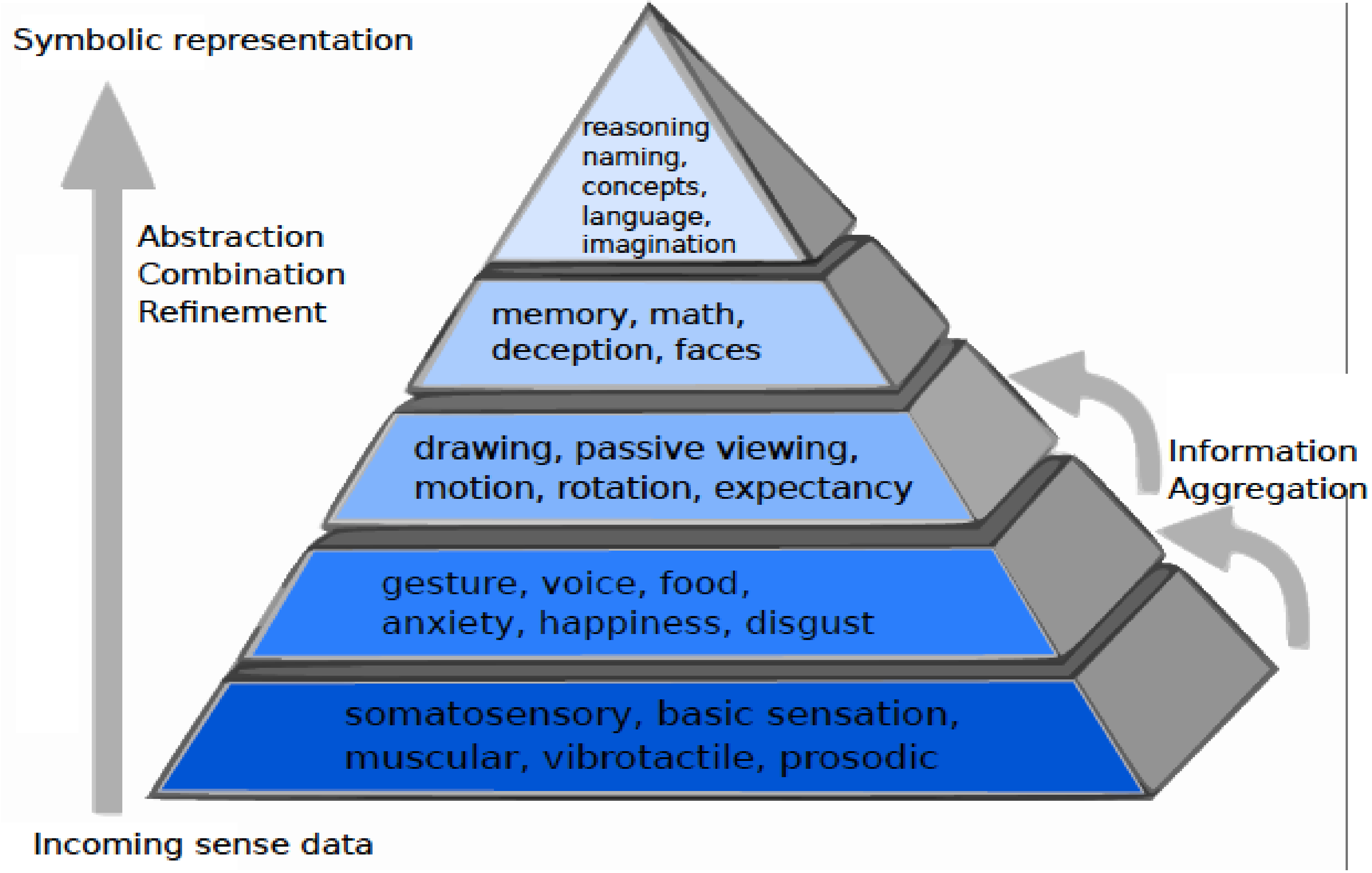
- Humans ordered tasks by abstraction via survey (Online Turk) without knowing about the slope hierarchy

- Close to perfect correlation

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Result

- Slope-hierarchy yields data-driven pyramid of cognition



Nature Communication (2015)
https://images.nature.com/w926/nature-assets/srep/2015/151216/srep18112/images/srep18112-f10.jpg

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Abstraction and Generalization

- Abstraction (Science):
  - A process of creating general concepts or representations … often with the goal of compressing the information content … and retaining only information which is relevant
  - Process of information aggregation, refinement, combination, integration, coalescing, accumulation, amalgamation; combination of ideas
- Generalization (Psychology)
  - The ability to respond in the same way to different but similar stimuli

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Can similar method be used on DNNs?

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Cognitive Neural Activation (CNA): Mathematical Definition

Definition: Assume X is a dataset, and A is a network architecture:

1) $\alpha(x)$ – the abstraction level of $x \in X$

2) $\beta(x)$ – the firing slope of A when calculating x $\in X$

3) CNA - the correlation (Pearson) between the levels of abstraction (for all x $\in X$) and the neuronal slopes:

$$\text{CNA (X, A)} = \rho_{\alpha,\beta}$$

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

CNA in Neuroscience

	Neuroscience
X – network computation	Abstraction tasks from fMRI datasets
A - architecture	Connectome depth network
α(x) - abstraction	Ordered by survey
β(x) - slope	The slope via big-data analysis
Finding	**CNA (X,A) ~ 1**

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

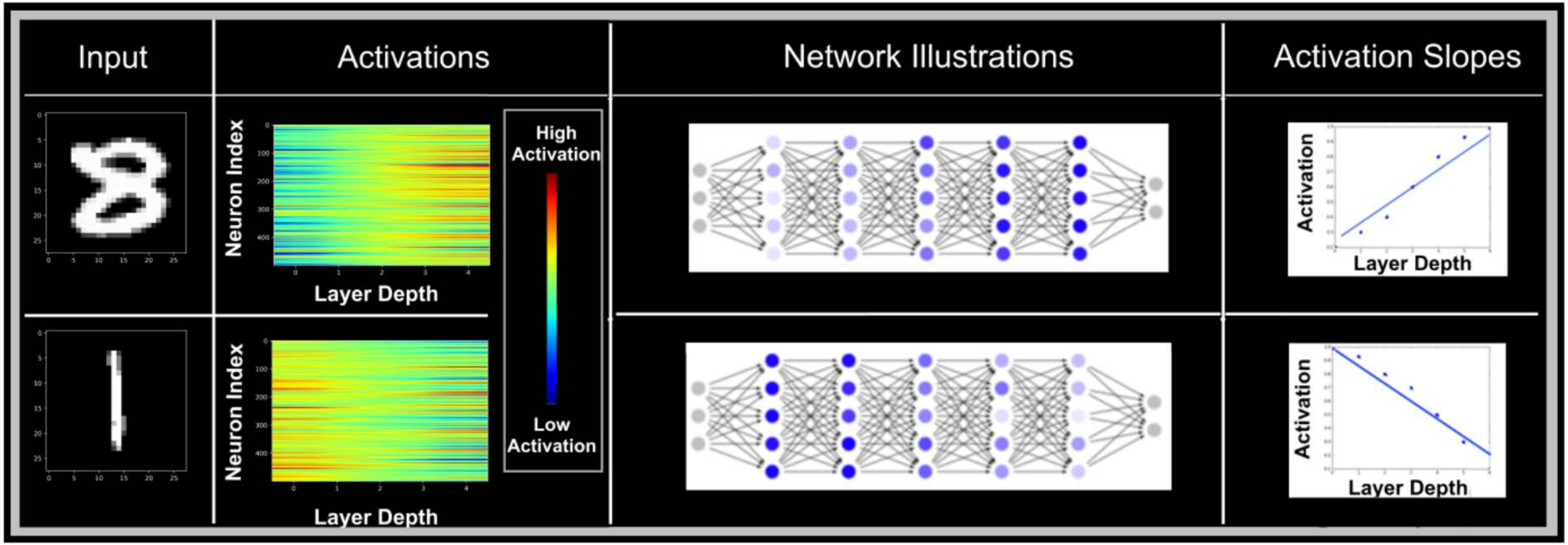
CNA – does it work for DNNs?

	Neuroscience	DNN
X – network computation	Abstraction tasks from fMRI datasets	Input data to DNN
A - architecture	Connectome depth network	Layered architecture
α(x) - abstraction	Ordered by survey	Shannon entropy (approximated)
β(x) - slope	The slope via big-data analysis	Total firing per layer, slope calculated
Finding	**CNA (X,A) ~ 1**	**1 / 0 / -1 ?**

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

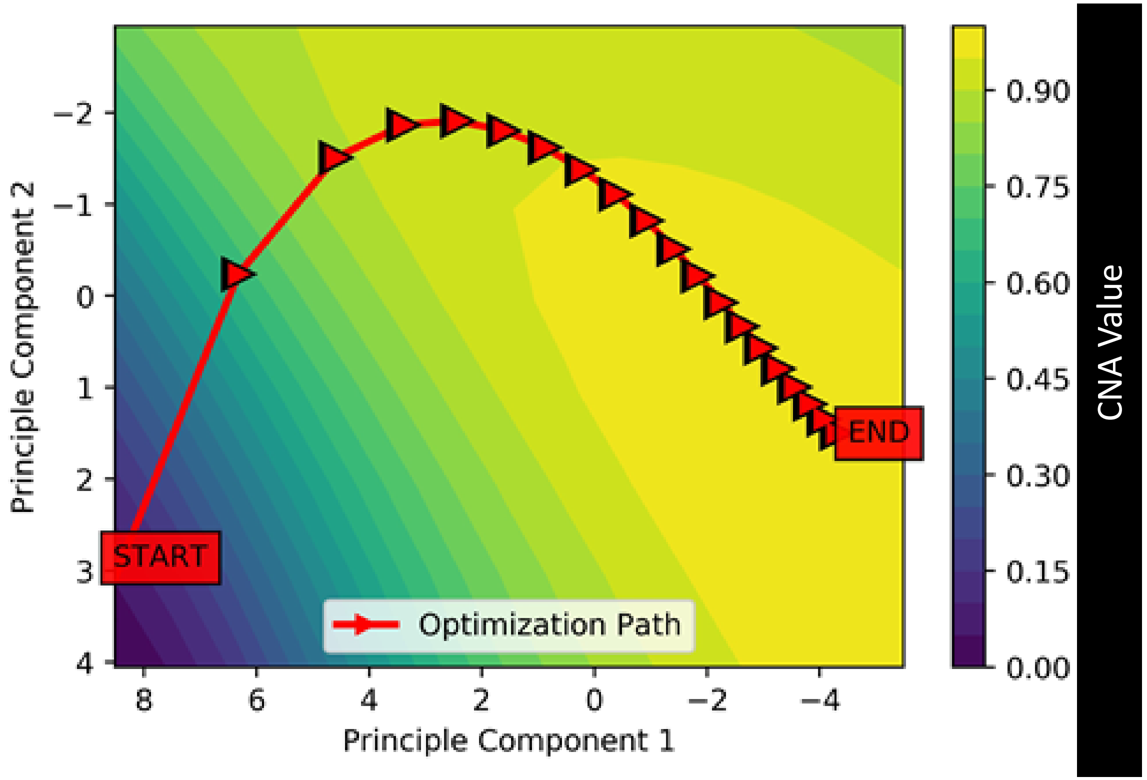
What does CNA mean for DNN?

Illustration: MNIST on MLP with 5 layers and 500 neurons per layer



Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

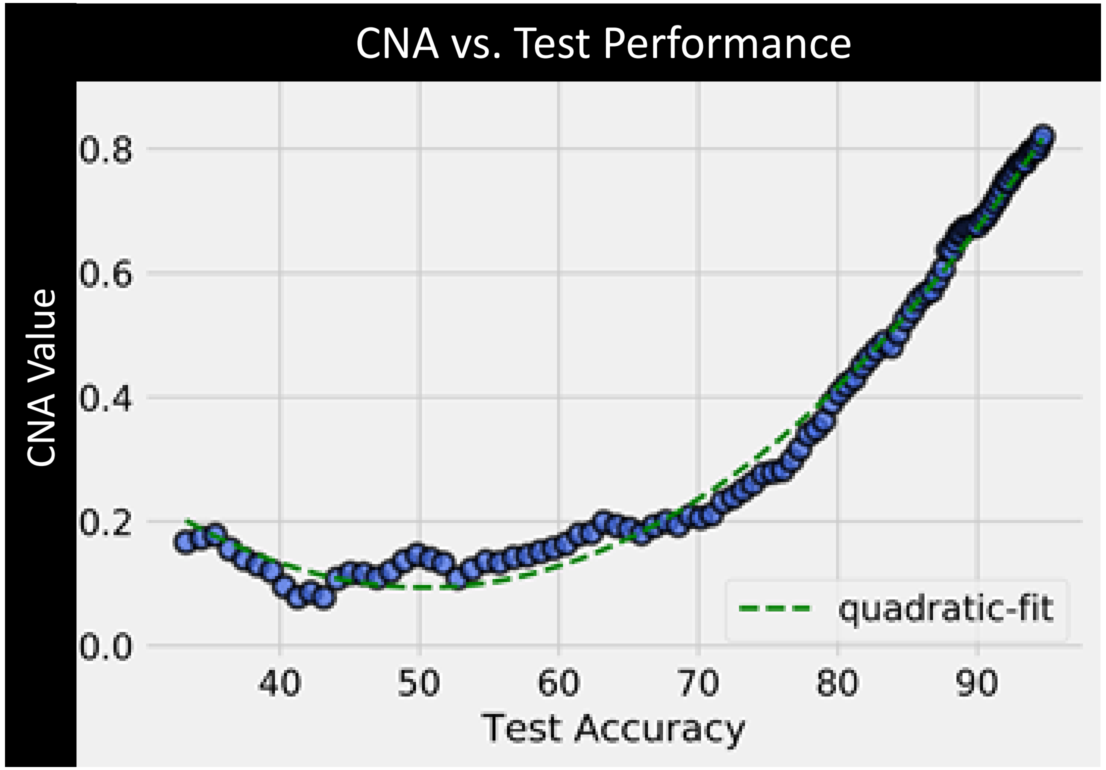
CNA During DNN Training



- Backpropagation during training changes weights towards higher CNA
- Weights change according to an optimization path that climbs to elevated values of CNA
- Largest rate of change occurs in early training for both the CNA and the training accuracy

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

CNA and Generalization Accuracy



- 147 combinations: 6 datasets (including ImageNet), 4 architectures, weights recorded every 20 passes
- Shows significant correlation with test accuracy
- At >70% DNNs become similar to the brain as classification results improve

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Generalization Gap

- Goal of generalization theory – explain whether/when/how improving accuracy during training (memorization) also improves test accuracy
- Generalization gap – difference between test and training accuracy
- Can CNA function as a generalization gap predictor?

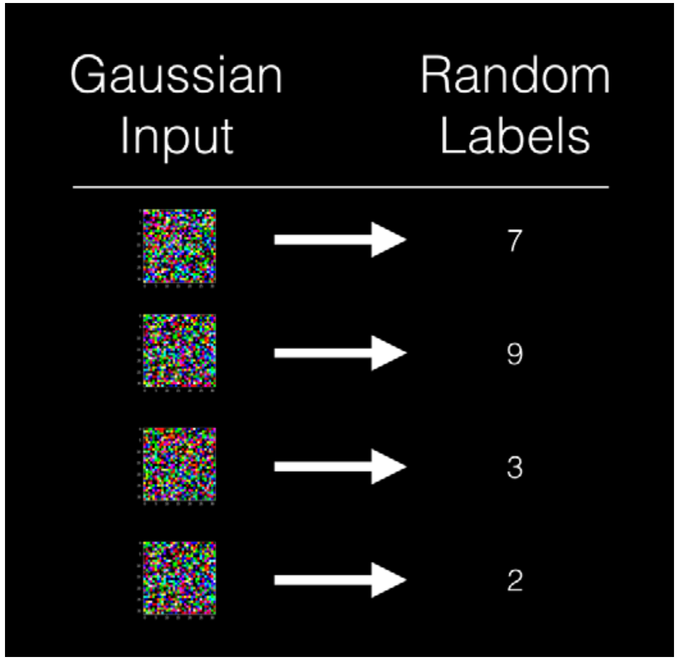
Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Datasets Tested

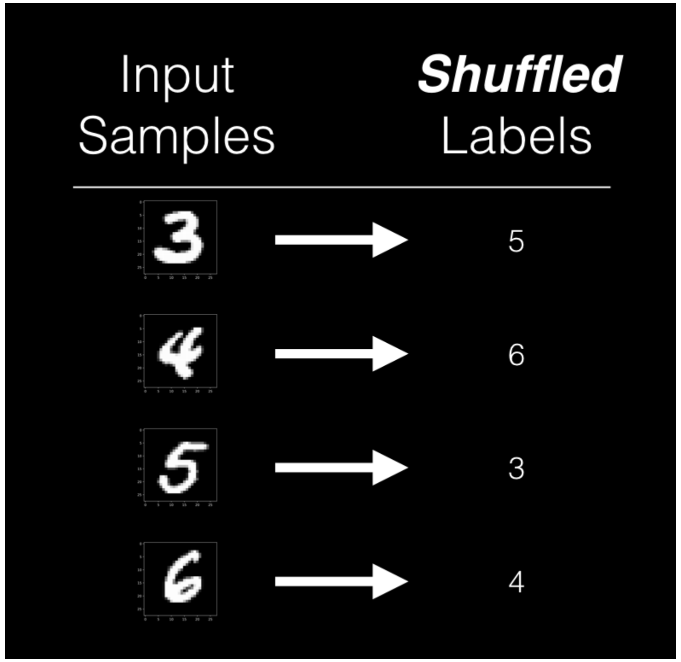
- Test on standard datasets: ImageNet, CIFAR-10, CIFAR-100, SVHN, MNIST, Fashion-MNIST, Networks: MLP, VGG-18, ResNet-18, ResNet-100

And

- On non-standard datasets:
- Random labels of 10%-50% in training



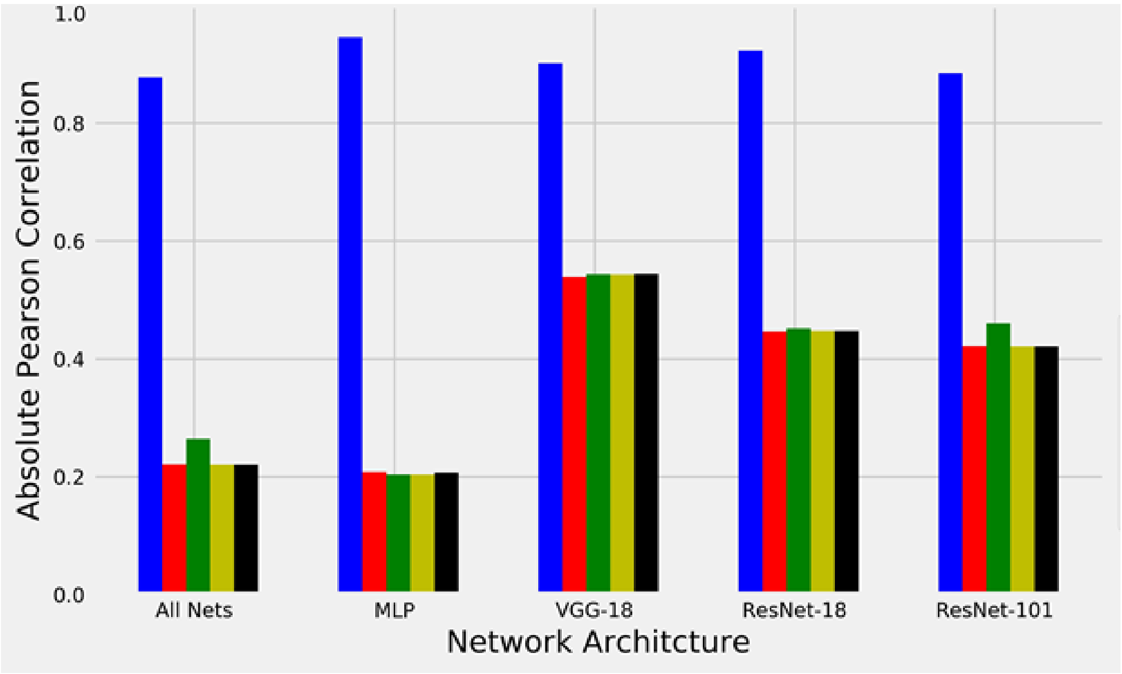
Gaussian Random



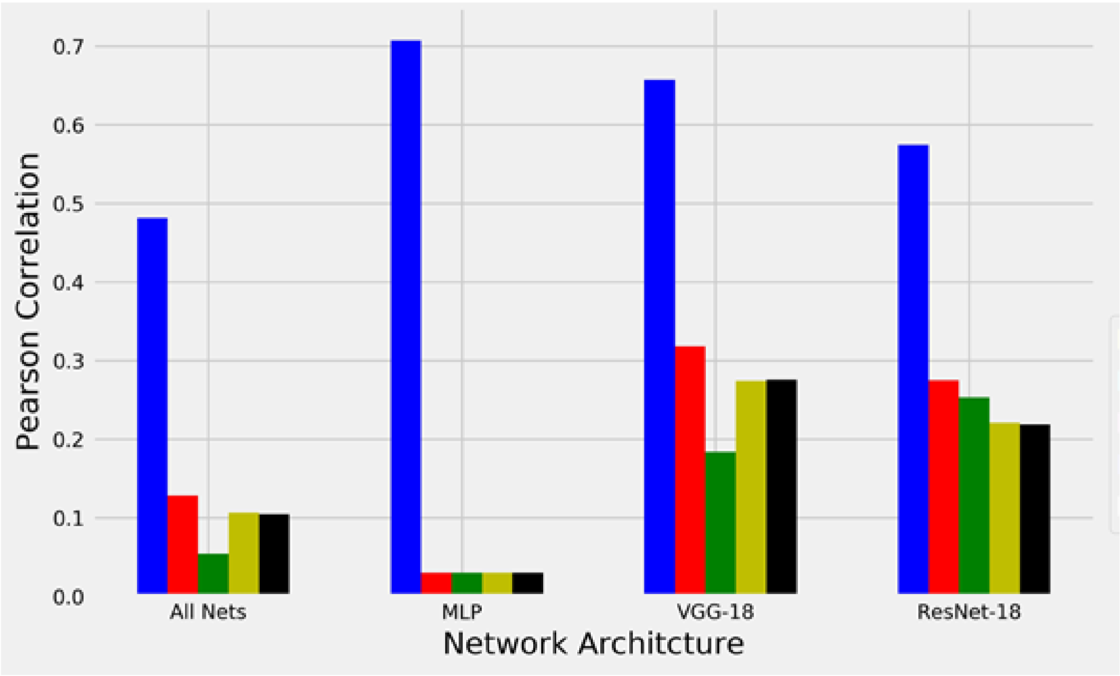
Shuffled Labels

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

CNA as Generalization Predictor



With Gaussian Random



With Shuffled Label

Datasets: ImageNet, CIFAR-10, CIFAR-100, SVHN, MNIST, Fashion-MNIST, Networks: MLP, VGG-18, ResNet-18, ResNet-100

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

Conclusions

- CNA describes how brain abstracts
- CNA shows that brain and DNNs have similar behavior for abstraction
- CNA predicts DNN capability to generalize including on complex datasets

Source: Siegelmann/Taylor research, Univ. of Mass, Amherst

DARPA

www.darpa.mil

IEEE COMCAS 2019

International Conference on Microwaves, Communications, Antennas & Electronic Systems

4-6 November ▪ David Intercontinental Hotel ▪ Tel Aviv, Israel