CSCI 4360/6360 Data Science II

Information Theory

TECH INSIDER | INNOVATION

'Macl
the ir

Drake B
Apr. 1,

FACEBO

Don't worry,

We're in the

It used to b

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

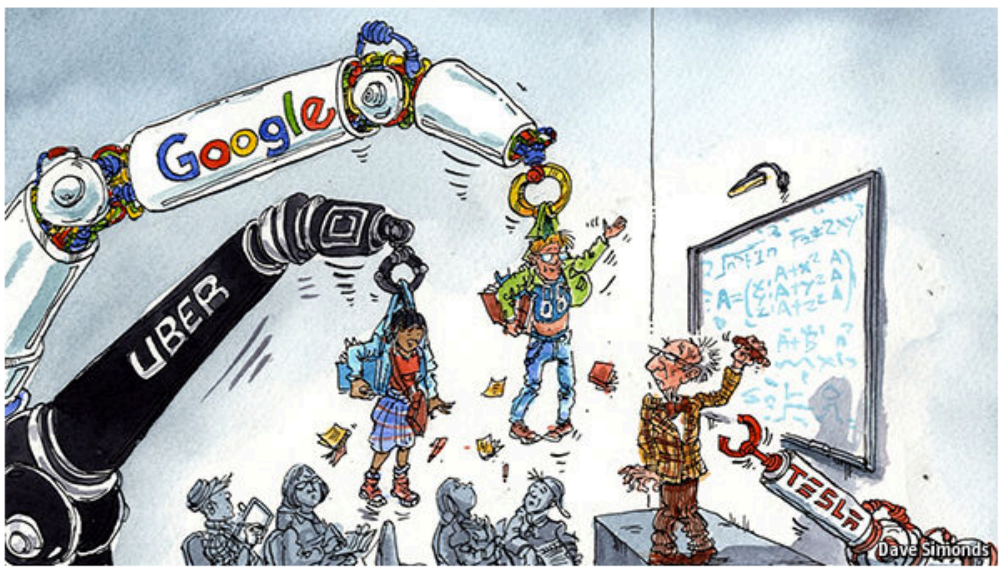
Artificial intelligence

Million-dollar babies

As Silicon Valley fights for talent, universities struggle to hold on to their stars

Apr 2nd 2016 | SAN FRANCISCO | From the print edition

Timekeeper | Like 6.8K | Tweet



CIENCE | LIFE | T-LOUNGE | TE

dfather Of
aid Of AI

SUBSCRIBE

program AlphaGo
vorld Go champion Lee

ence experts, who
er program would need at
ough to be able to beat a

uter program is that
said that AlphaGo could
h with Sedol was able to

st skilled humans in
d Garry Kasparov two
lerts

Talks

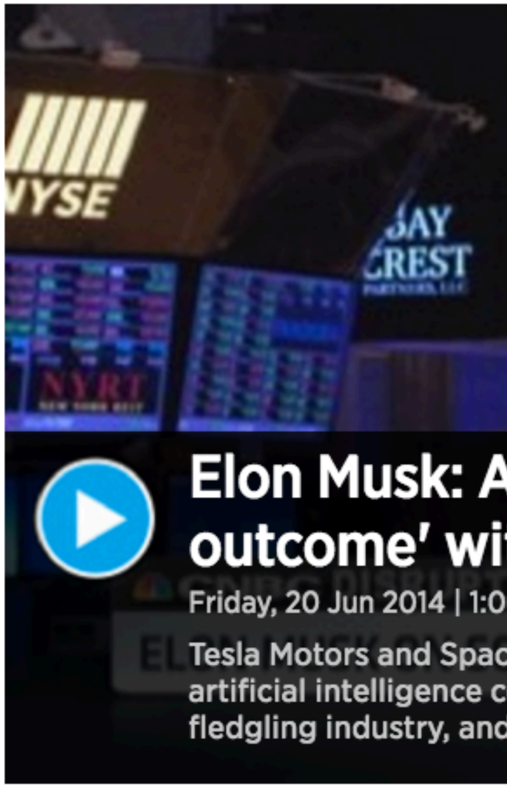
d Bou
all Off

SECTIONS

Facebook I
and Intrus
Paying Off

TECHNOLOGY

Silicon Valley Looks to Artificial Intelligence for the Next Big Thing



Science AAAS

Home | News | Journals | Topics | Careers

Latest News | ScienceInsider | ScienceShots | Sifter | From the Magazine | About News | Quizzes

SHARE

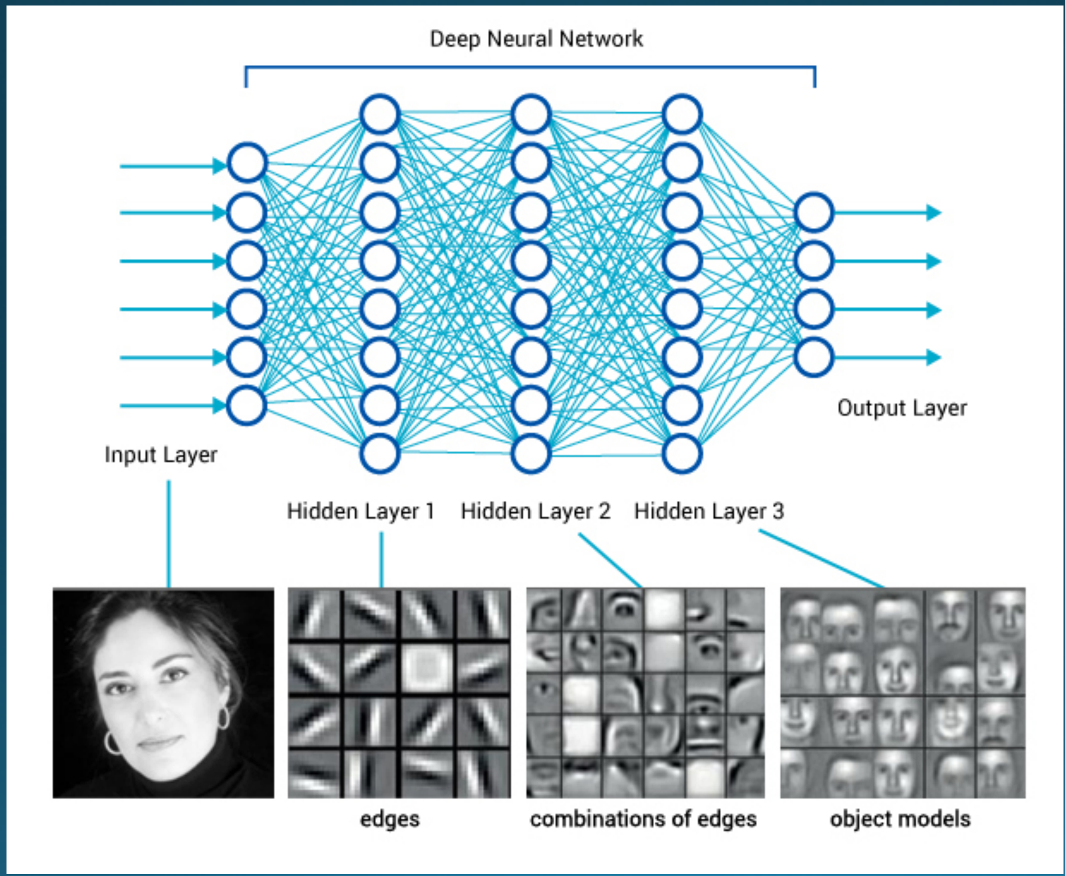
© FilippoBacci/iStockphoto

Artificial intelligence steals money from banking customers

By Adrian Cho | Apr. 1, 2016 , 3:00 AM

Explain Your Deep Network

What do the weights *mean*?



For facial recognition

For scene segmentation

For multitask learning

For autoencoders

Biggest Drawback of Deep Learning

- Interpretability
  - Explain what is being learned at layer 47, weight 301
  - What is layer 25 learning?
  - **What determines the network's decision-making process for a given input?**
- GoogLeNet architecture

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

Table 1: GoogLeNet incarnation of the Inception architecture.

Information-Theoretic Perspective

Opening the black box of Deep Neural Networks via Information

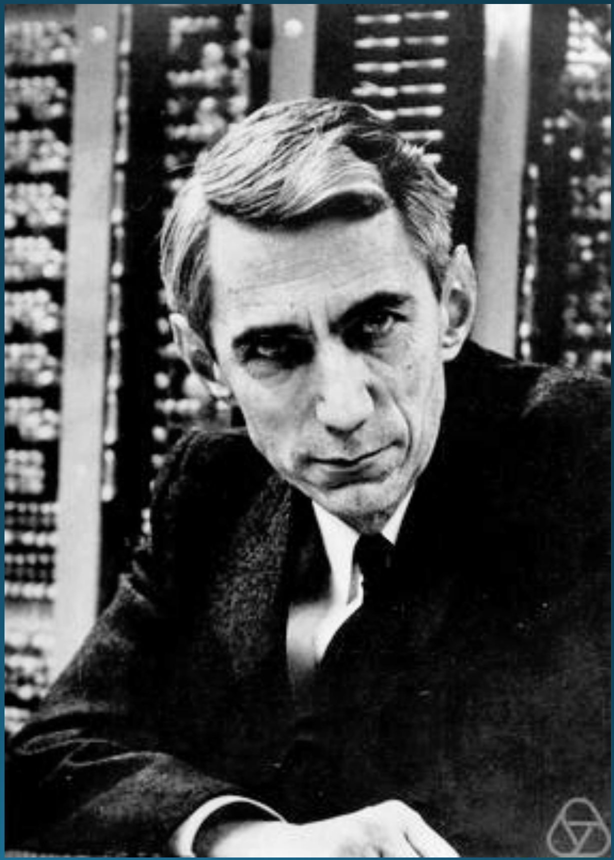
Ravid Schwartz-Ziv RAVID.ZIV@MAIL.HUJI.AC.IL
Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Naftali Tishby* TISHBY@CS.HUJI.AC.IL
School of Engineering and Computer Science
and Edmond and Lilly Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, 91904, Israel

Editor: ICRI-CI

Information Theory

- Dr. Claude Shannon
  - Outlined in 1948 paper, "A Mathematical Theory of Communication"
  - The "Father of Information Theory"
- *Information*: set of possible messages
  - Sent over a noisy channel
  - Receiver reconstructs messages with low probability of error
- **Revolutionized digital communication via compression**



Information Theory

- Communication
- Information retrieval
- Intelligence gathering
- Signal processing
- Gambling
- Statistics
- Cryptography
- Music composition

Information Theory

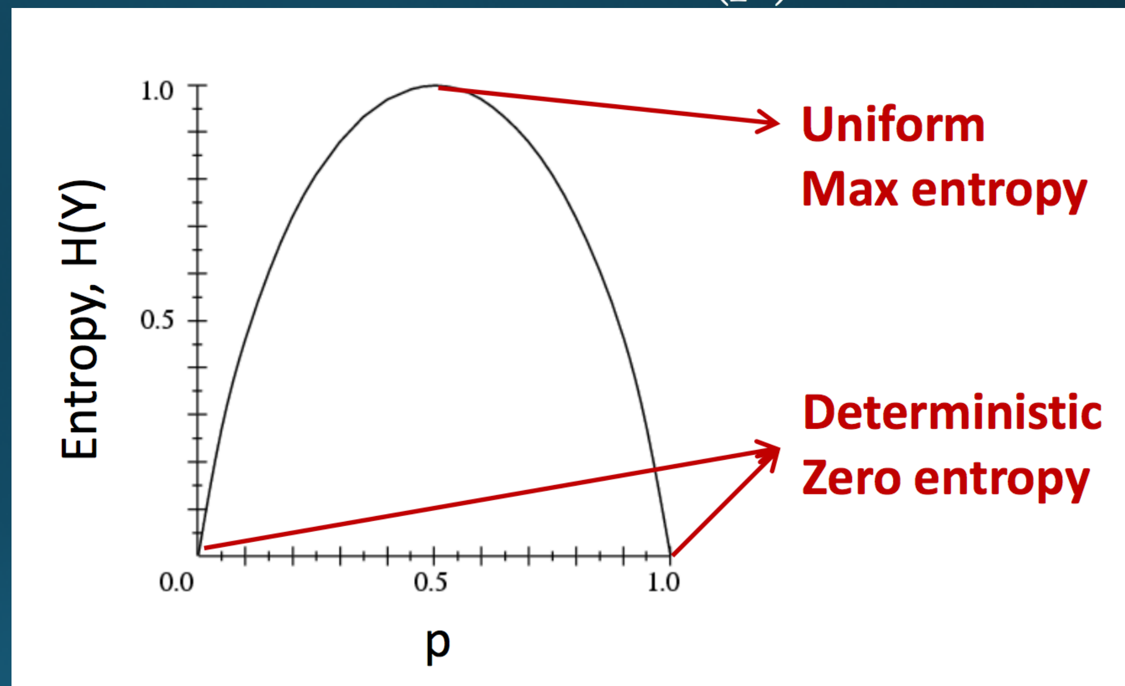
- Basic unit of information is the *bit*
  - Not *necessarily* 1s and 0s, but often takes that incarnation in practice
- *Entropy*
  - Units of bits per symbol
  - Quantifies *uncertainty* in [discrete] random variable

$$H = -\sum_i p_i \log_2(p_i)$$

Entropy

$$Y \sim \mathrm{Ber}(p)$$

- Can be written in terms of a random variable, *Y*
- More uncertainty = Higher entropy



$$H_Y = H(Y) = -\sum_y P(Y=y)\log_2 P(Y=y)$$

Entropy

- H(Y) is the **expected number of bits** needed to encode a randomly-drawn value of *Y* (assuming the most efficient code)

$$H_Y = H(Y) = -\sum_y P(Y = y) \log_2 P(Y = y)$$

- Definition of expected value

$$E[X] = \sum_i x_i P(X = x_i)$$

Joint Entropy

Symmetric

- *X, Y)* is the entropy of the pairing of *X* and *Y*
  - If *X* and *Y* are independent, *H(X, Y) = H(X) + H(Y)*

$$H(X,Y) = E_{X,Y}\left[-\log P(x,y)\right] = -\sum_{x,y} P(x,y)\log P(x,y)$$

- Not to be confused with **cross-entropy**

Asymmetric

  - Average number of bits needed to identify an event as having come from either *X* or *Y*

$$H(X,Y) = E_X\left[-\log Y\right] = H(X) + D_{KL}(X||Y)$$

- *D* is the KL divergence
- (why the notations are the same… no idea)

Conditional Entropy

- Also called *equivocation*

$$H(X|Y) = E_Y\left[H(X|Y=y)\right] = -\sum_i P(y_i) \sum_j P(x_j|y_i) \log P(x_j|y_i)$$

- Like conditional probability, a basic property emerges with respect to the joint and marginal entropies

$$H(X|Y) = H(X,Y) - H(Y)$$

Mutual Information

- Which all gives rise to the concept of *mutual information*: how much information you can obtain about one random variable by observing another

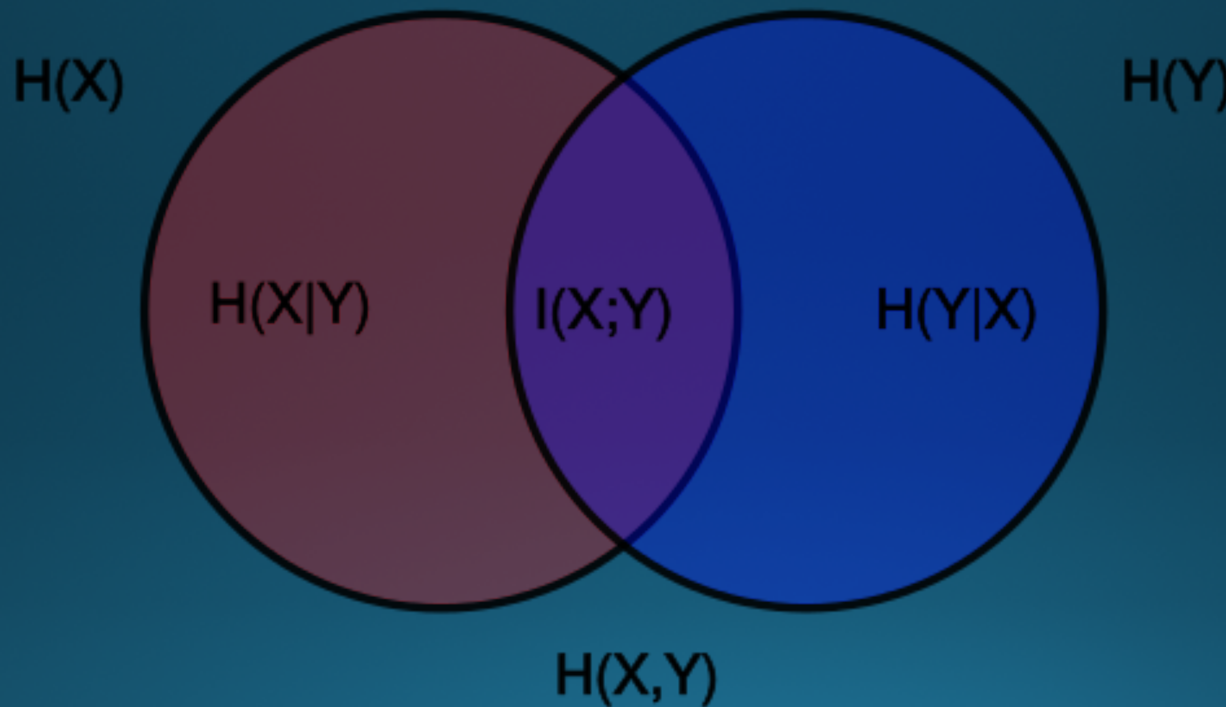
$$I(X;Y) = H(X) - H(X|Y)$$

If this is 0, knowing Y tells us nothing about X

Uncertainty in X

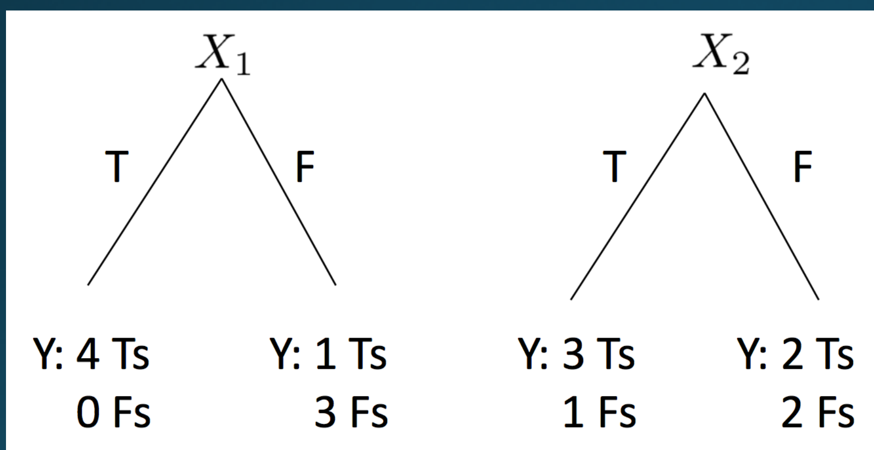
Uncertainty in X, given we have observed Y

Mutual Information



Mutual Information

- Used extensively in decision trees: which features to branch on

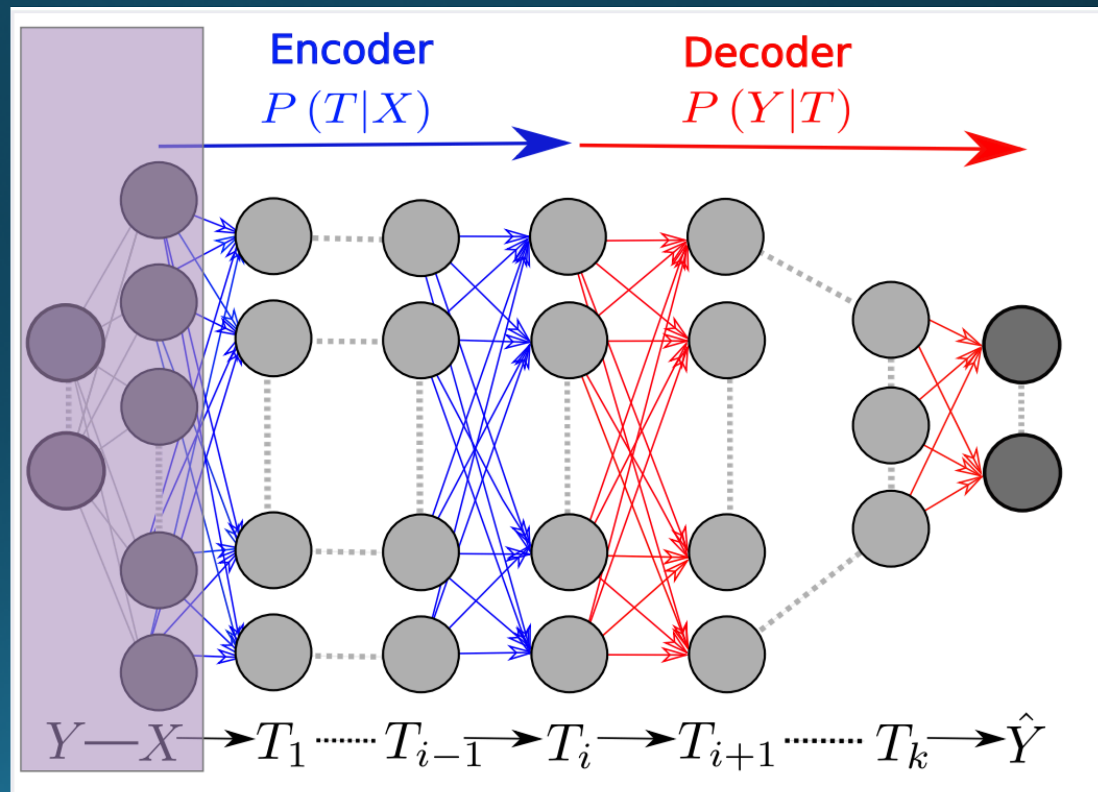


- Pick the feature that yields *maximum information gain* or *I(X;Y)* (i.e., biggest *drop* in entropy)

X_1	X_2	Y
T	T	T
T	F	T
T	T	T
T	F	T
F	T	T
F	F	F
F	T	F
F	F	F

What does this have to do with deep learning?

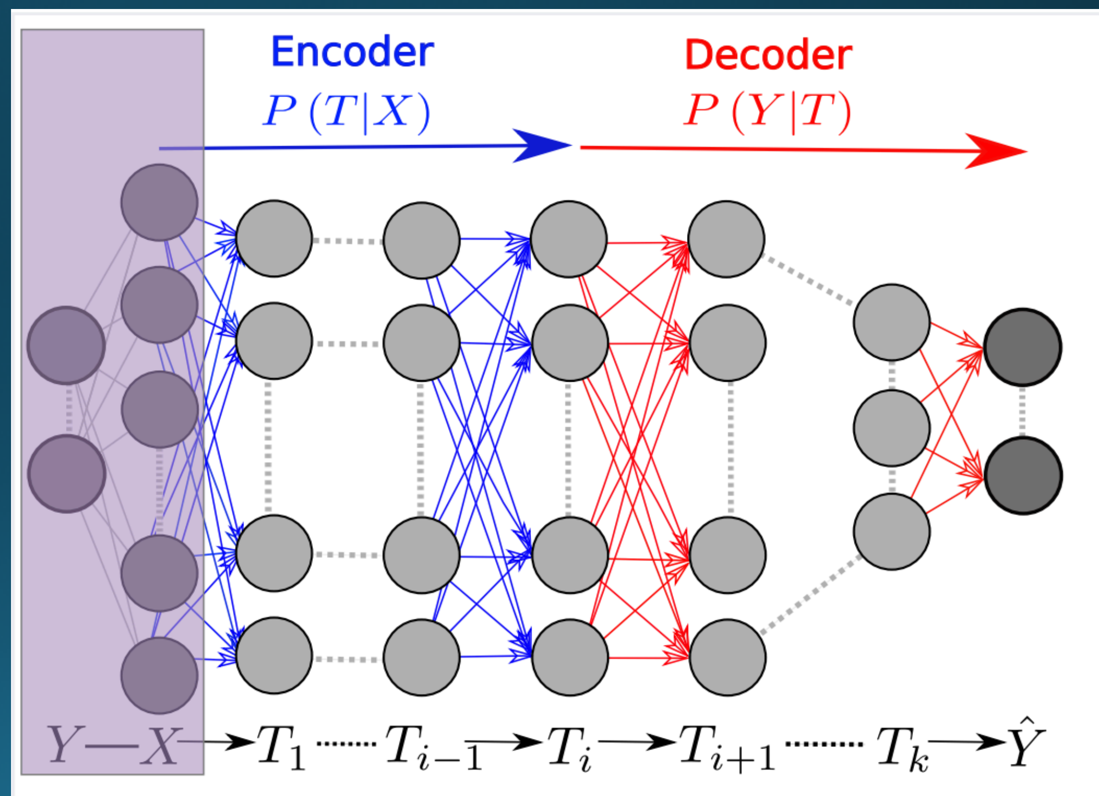
- Multilayer ANNs are [mostly] directed acyclic graphs (DAGs)
- Therefore, we can view them as Markov Chains



Notation

- *X*: input
- *Y*: target output
- *T*: intermediate representation

- Any *T* defined as
  - Encoder *P(T|X)*
  - Decoder *P(Y|T)*



Markov Chains + Mutual Information

- *Data Processing Inequality* (DPI) [Cover and Thomas *et al*, 2006]
- For any three variables that form a Markov chain X -> Y -> Z,

$$I(X;Y) \geq I(X;Z)$$

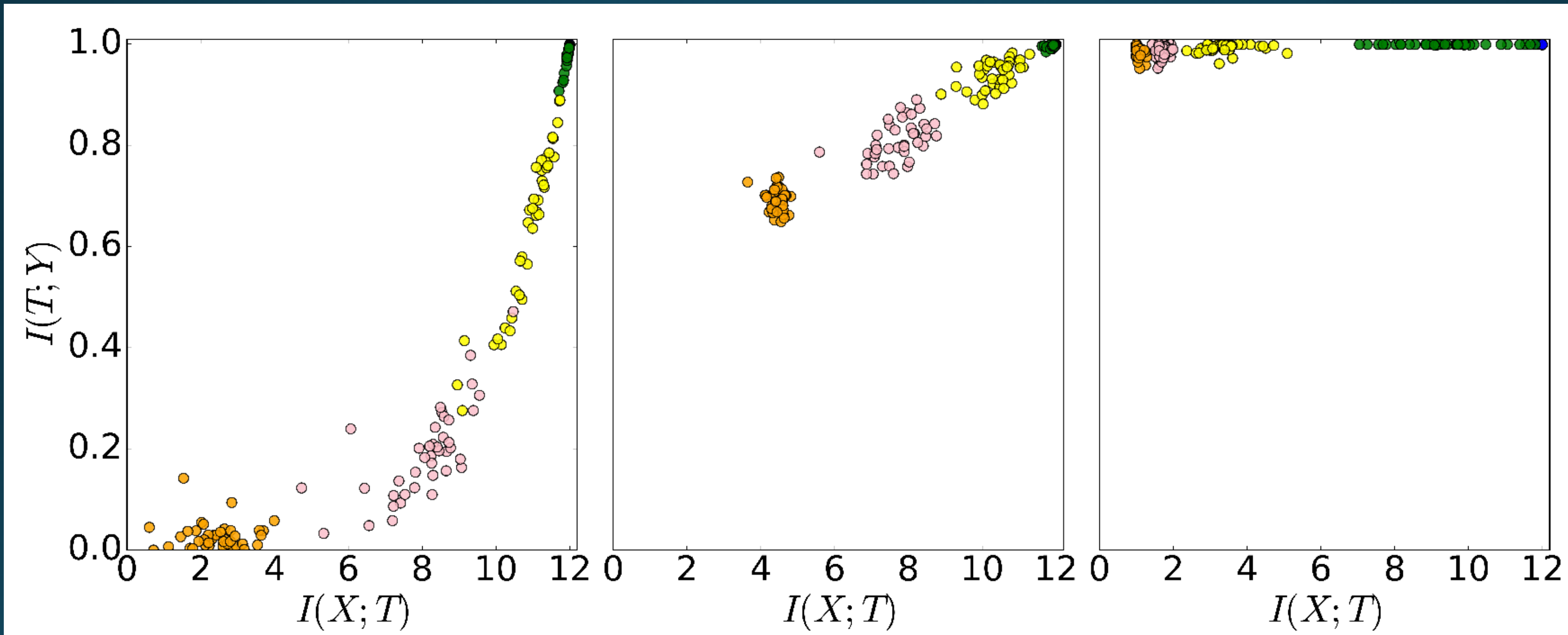
- Intuition
  - Information is generally lost (never gained) when transmitted through a noisy channel
  - “post-processing cannot increase information”
  - “garbage in, garbage out”

The Information Plane

- Given *P(X; Y)*, *T* is uniquely mapped to a point on the information plane with coordinates *[I(X ; T), I(T ; Y)]*.

$$I(X;Y) \geq I(T_1;Y) \geq I(T_2;Y) \geq ... \geq I(T_k;Y) \geq I(\hat{Y};Y)$$
$$H(X) \geq I(X;T_1) \geq I(X;T_2) \geq ... \geq I(X;T_k) \geq I(X;\hat{Y}).$$

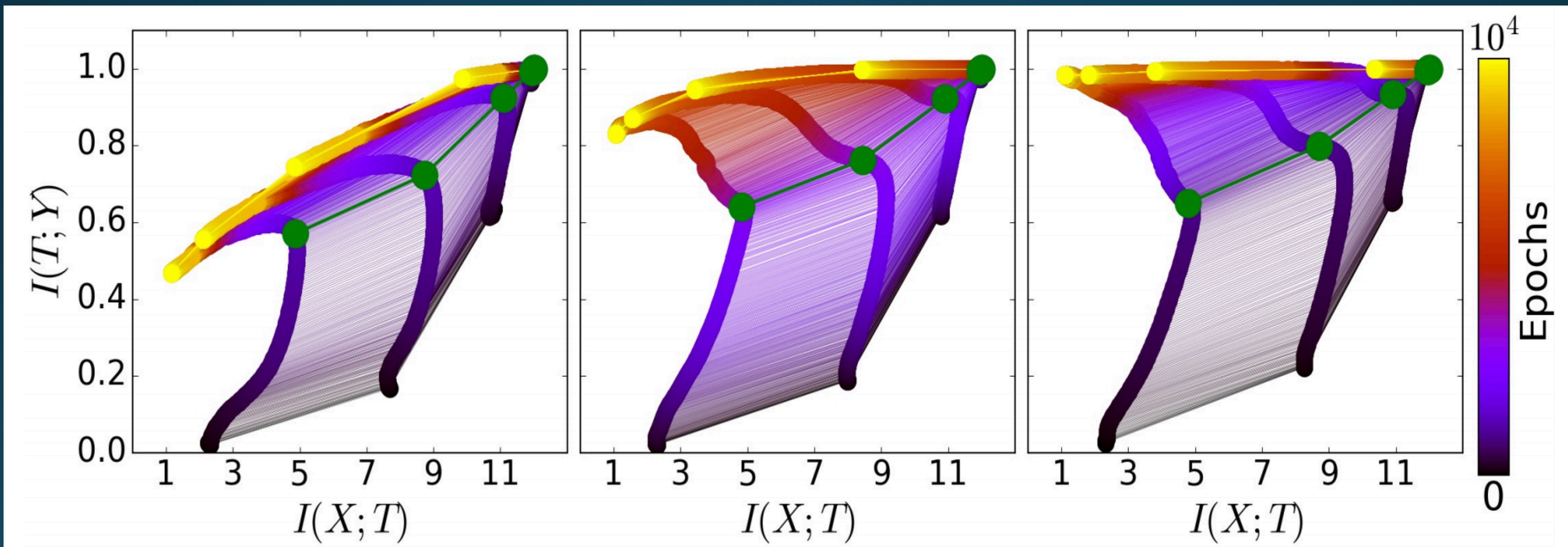
The Information Plane



- X-axis: *I(X | T)* of *T* encoded in layer *i*
- Y-axis: *I(T | Y)* of *T* encoded in layer *i*

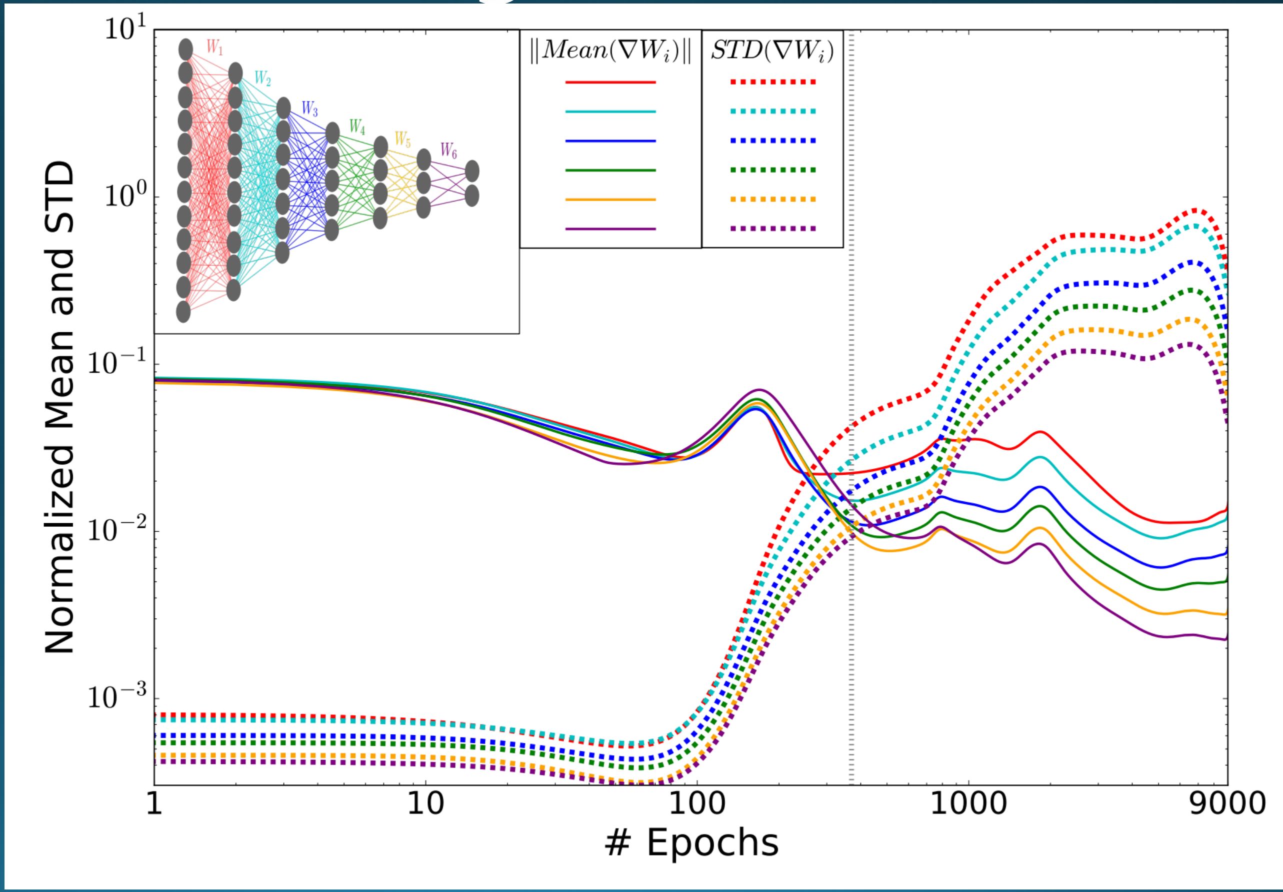
Dual Phases of Training

- Most time is spent in the 2nd phase



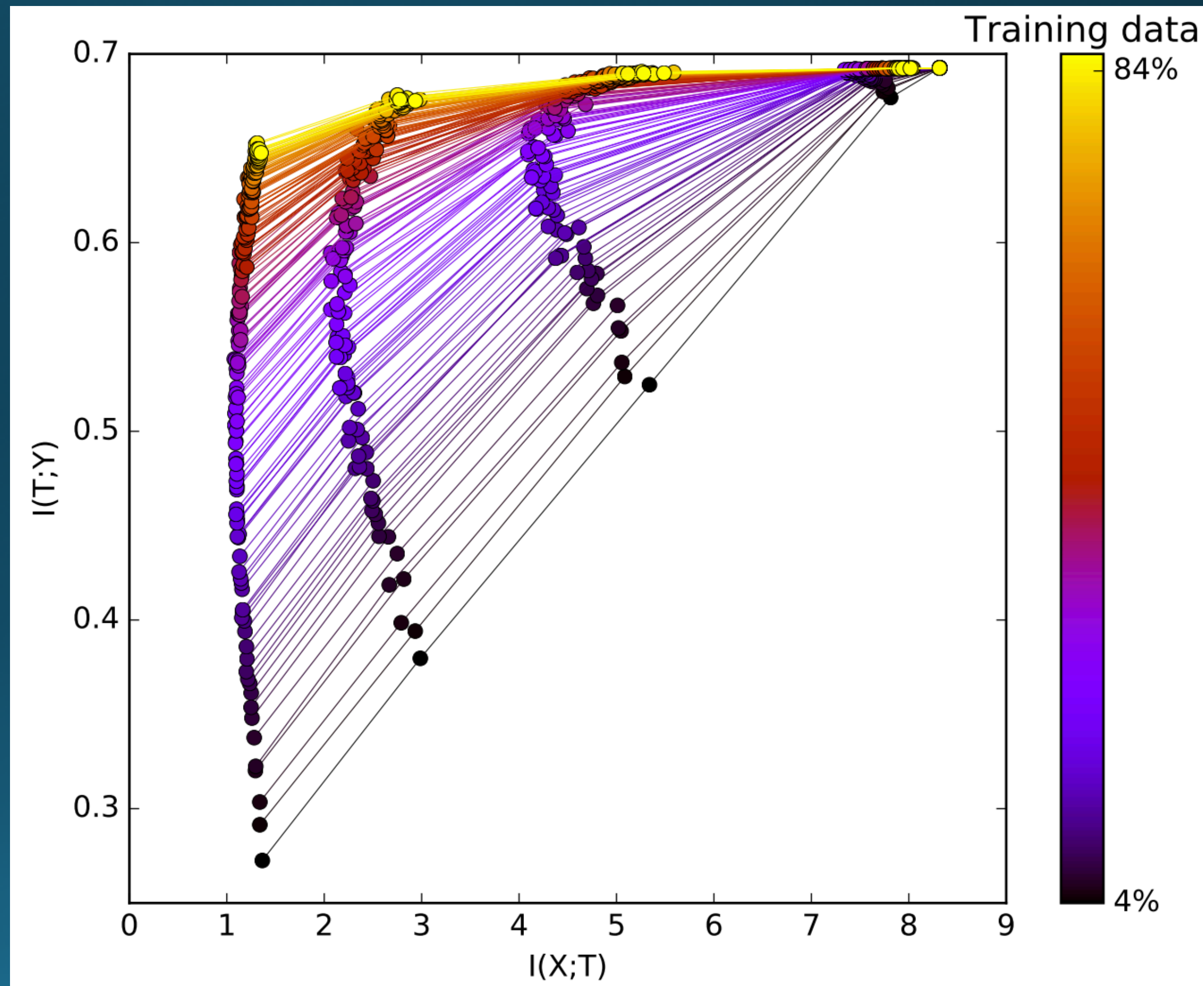
Dual Phases of Training

- Phase I: "Drift Phase"
  - Large gradients
  - Small variations
- Phase II: "Diffusion Phase"
  - Small gradients
  - Large inter-batch variations



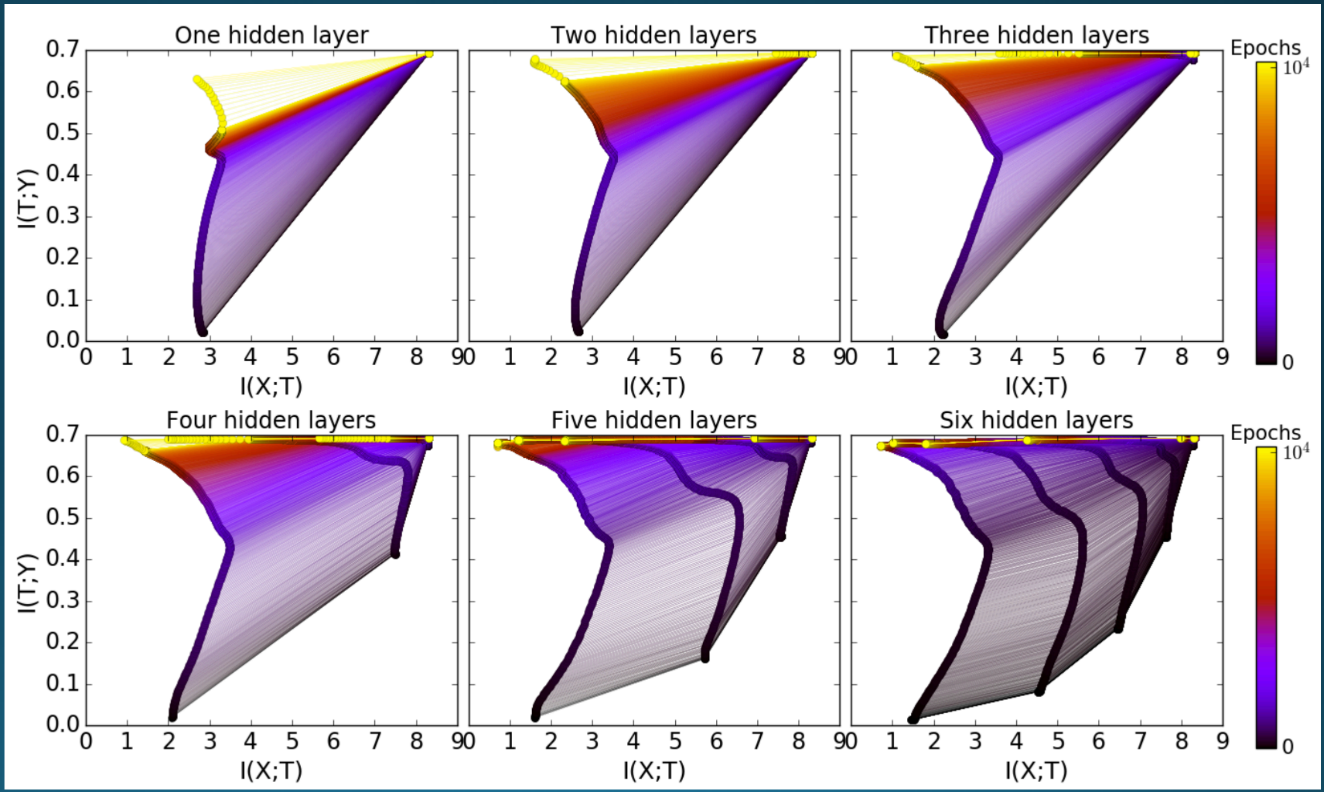
Training Data

- Amount of training data affected rate of passage through the two phases
- Six points along each line indicate one of the six layers
- Averaged over 50 initializations with random weights



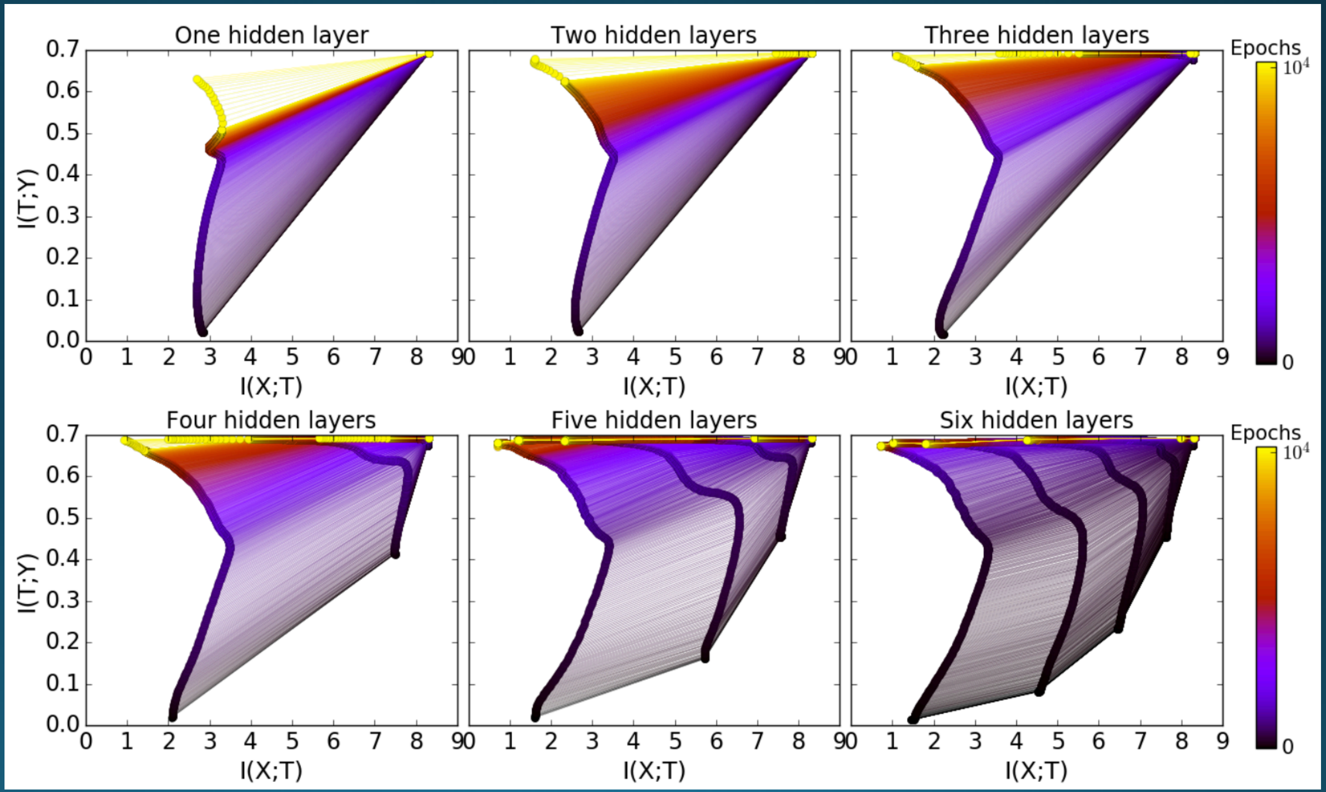
Key Insights

1: Adding hidden layers dramatically reduces the number of training epochs needed for good generalization



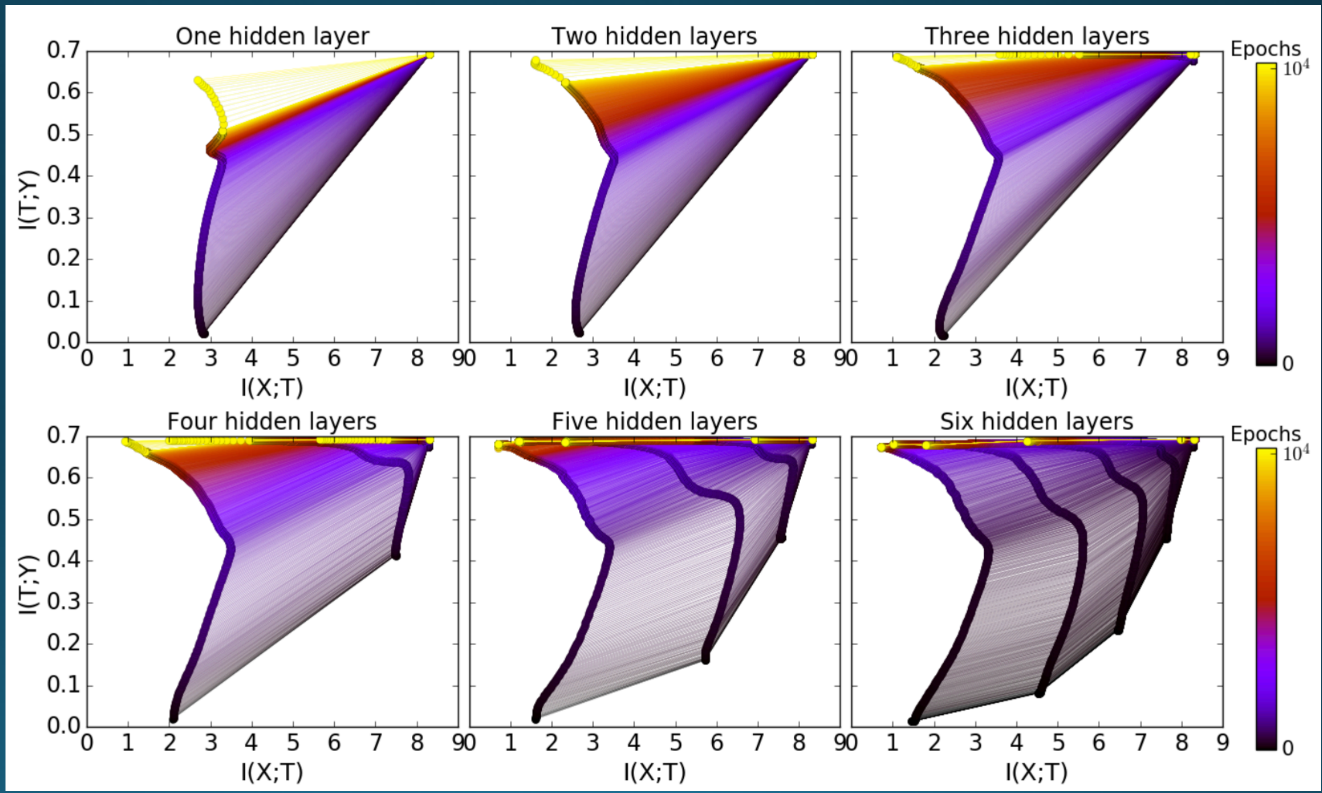
Key Insights

2: The compression phase of each layer is shorter when it starts from a previous compressed layer



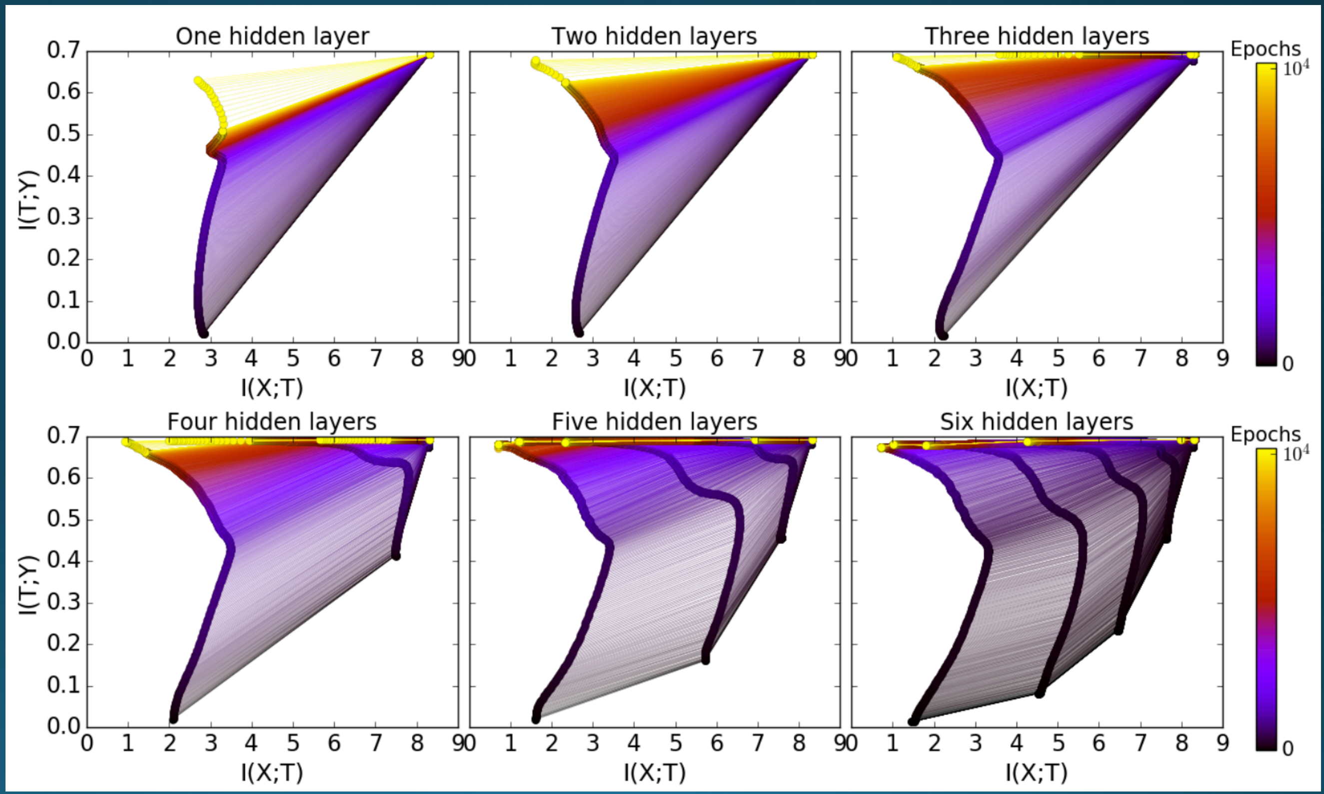
Key Insights

3: The compression is faster for the deeper (narrower and closer to the output) layers.



Key Insights

4: Even wide hidden layers eventually compress in the diffusion phase. Adding extra width does not help.



Conclusions

- The second phase (diffusion / compression) always resulted in a *different* configuration of weights

[Un]Surprising Conclusion #1: Many different weight configurations can offer similarly optimal performance

[Un]Surprising Conclusion #2: Looking at a single neuron or weight for insight into network performance is meaningless

Surprising Conclusion #3: Values of weights alone **cannot explain generalizability of deep networks**

- Adding hidden layers + Adding more training data both reduce training time required in compression stage

[Un]Surprising Conclusion #4: Data is the best regularizer

Surprising Conclusion #5: Rather than focus on explicit regularization & architectural redesigns, **exploit encoder & decoder distributions during training**; will yield best convergence rate

Course Details

- How is Assignment 5 going? **Due today!**
- How is the project going?

References

- "Opening the black box of Deep Neural Networks via Information", https://arxiv.org/pdf/1703.00810.pdf
  - Blog post summary https://theneuralperspective.com/2017/03/24/opening-the-black-box-of-deep-neural-networks-via-information/
- "Information Theory of Deep Learning" https://www.youtube.com/watch?v=RKvS958AqGY