The emergence of an

Information Bottleneck Theory *of* Deep Learning

Presentation for the conclusion of the Master Degree in Computer Science

By

Frederico Guth
Departamento de Ciência da Computação
Universidade de Brasília

Committee:

Téofilo de Campos (supervisor)
Universidade de Brasília

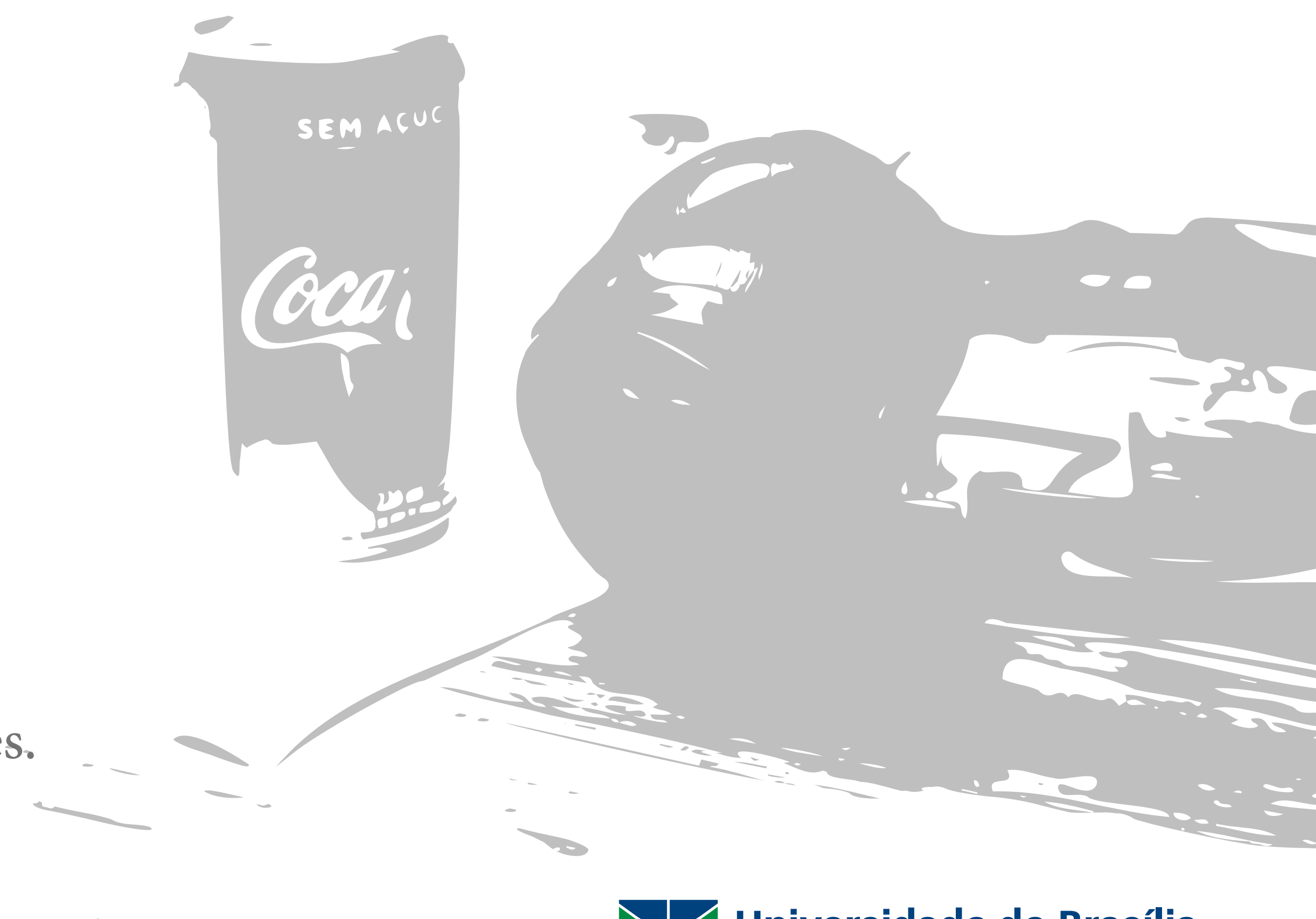
John Shawe-Taylor
University College London

Moacir Ponti
Universidade de São Paulo

Brasília, 20/01/2022

Agenda

1. Introduction
   - Problem and Research Objective
   - Research Questions and Methodology
2. Background
   - Machine Learning Theory (MLT)
   - Information Theoretic Learning (ITML)
   - MLT vs. ITML: “genealogy” and comparison
3. Information Bottleneck Theory new narrative
   - IB Principle and Relevance
   - IBT Main thesis and criticism
   - IB and Representation Learning: filling the gaps
   - Deep Learning phenomena in the IBT narrative
4. Conclusions
   - Strengths, weaknesses and research opportunities.



Problem

Practice-theory gap in Deep Learning Generalisation [Zha+16, Rah18].

IBT presents new perspective that may help fill this gap.

No comprehensive digest of IBT or comparison to MLT.

Objective

To investigate *to what extent* can IBT help us understand Deep Learning generalisation, presenting its strengths, weaknesses and research opportunities in a digest.

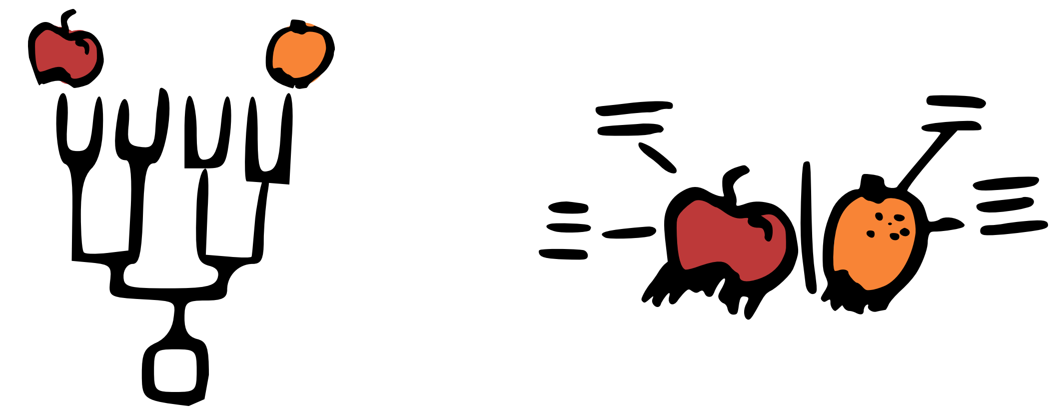
[Zha+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2016. arXiv: 1611.03530.

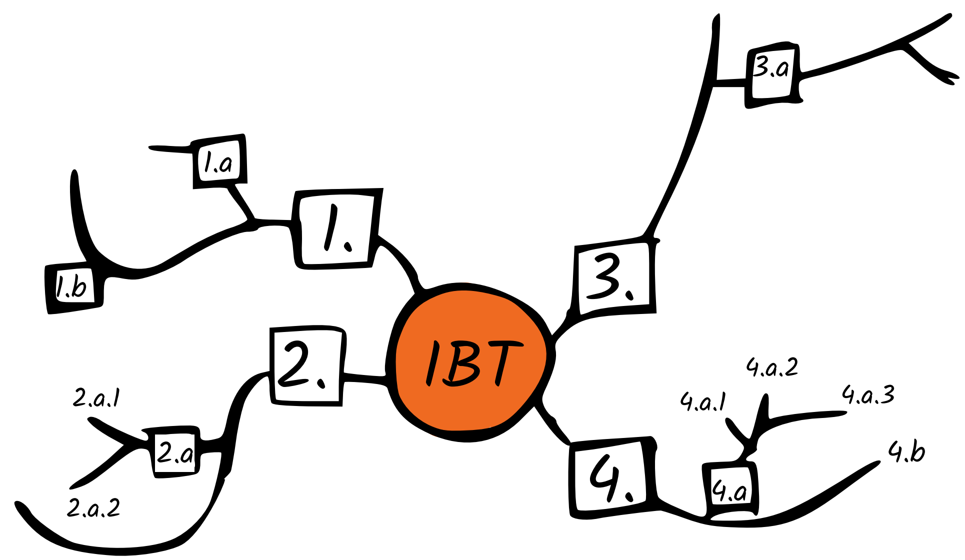
[Rah18] Ali Rahimi. Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech. https://youtu.be/x7psGHgatGM. [Online; Last accessed on 2020-08-04.] Mar. 7, 2018. url: https://youtu.be/x7psGHgatGM.

Research Questions

1. What are IBT fundamentals?
2. IBT and MLT differences and similarities?
3. Does IBT explain what MLT does?
4. Does IBT invalidate MLT results?
5. Can IBT explain phenomena currently not well understood?
6. IBT strengths?
7. IBT weaknesses?
8. What has been already developed in IBT?
9. IBT Research opportunities?

Methodology

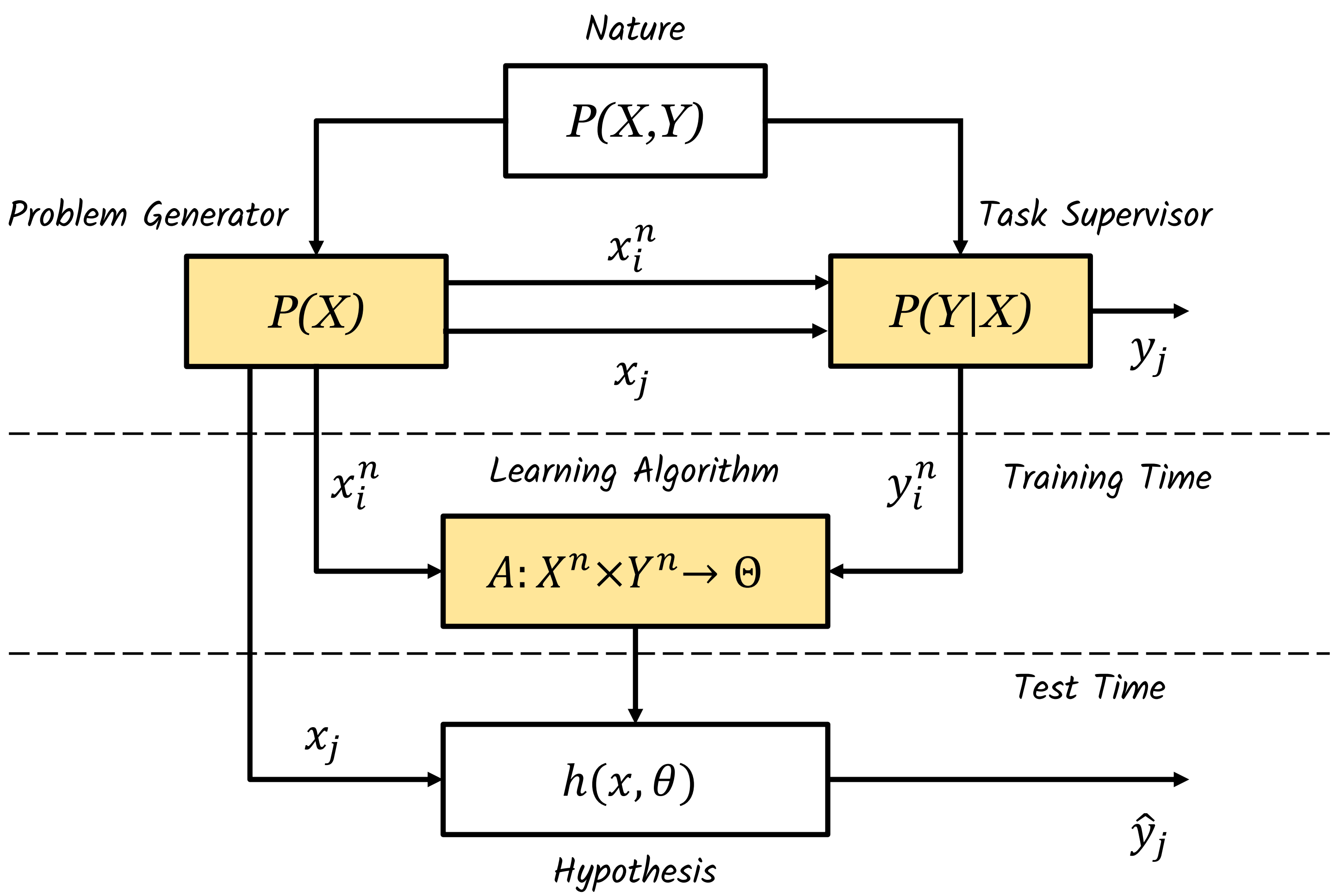


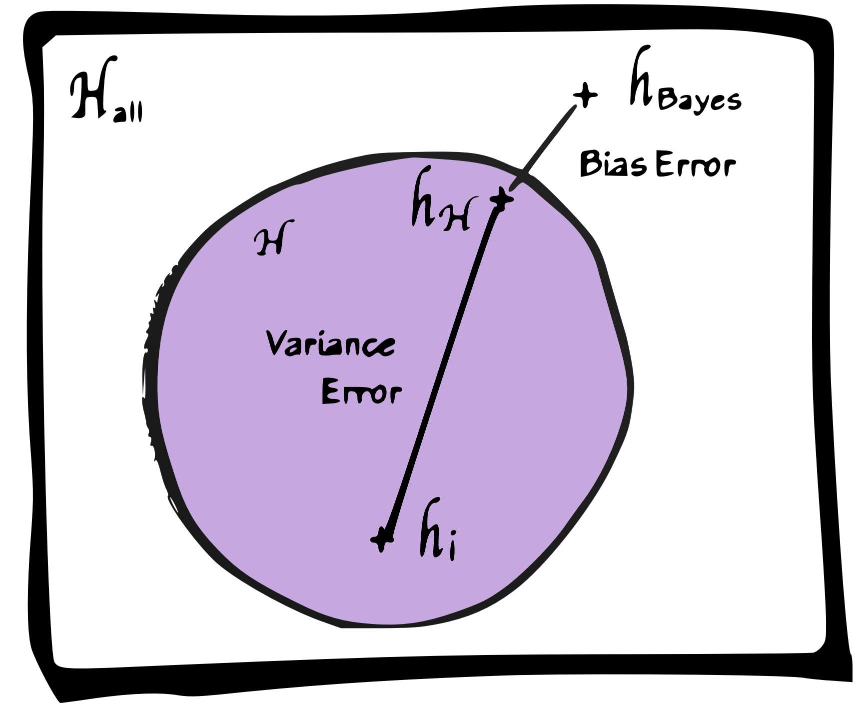


Universidade de Brasília

Machine Learning Theory

Learning as search in the hypothesis space



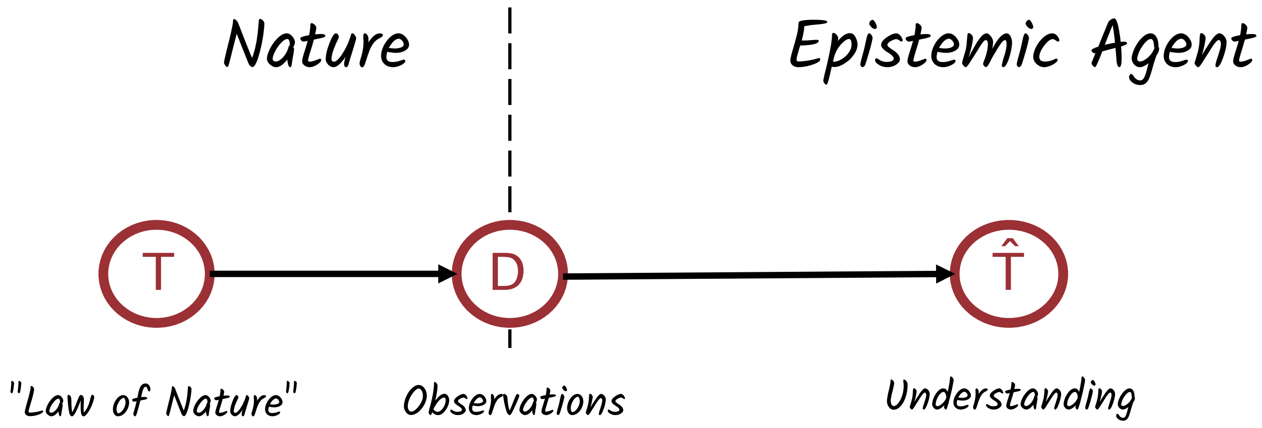


$$h_H := \arg\min_{h \in H} \hat{R}(h)$$

Universidade de Brasília

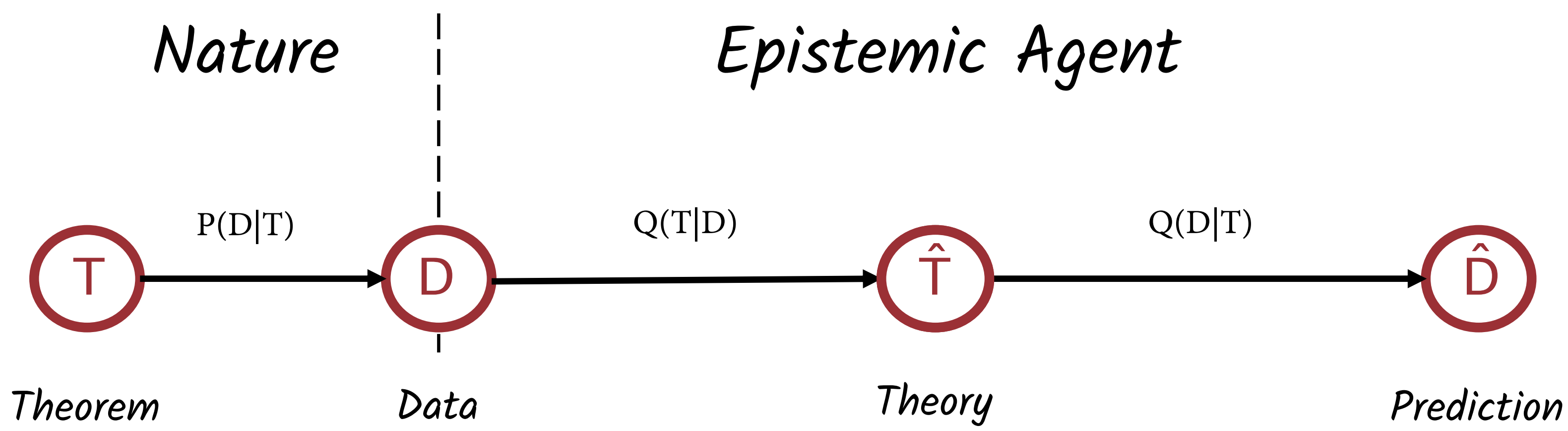
Information Theoretical Learning

Learning as a communication problem



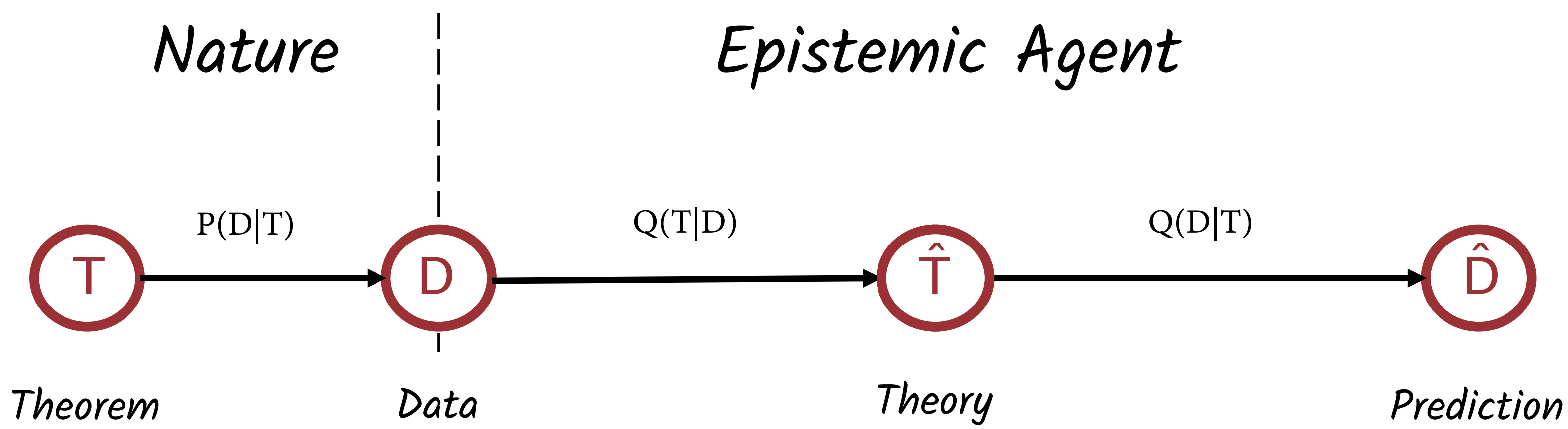
Information Theoretical Learning

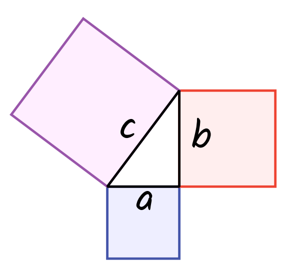
Learning as a communication problem

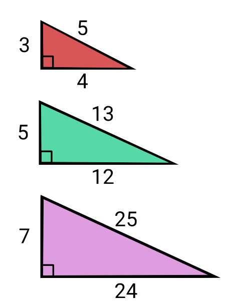


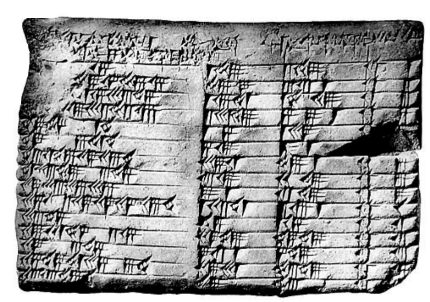
Information Theoretical Learning

Learning as a communication problem







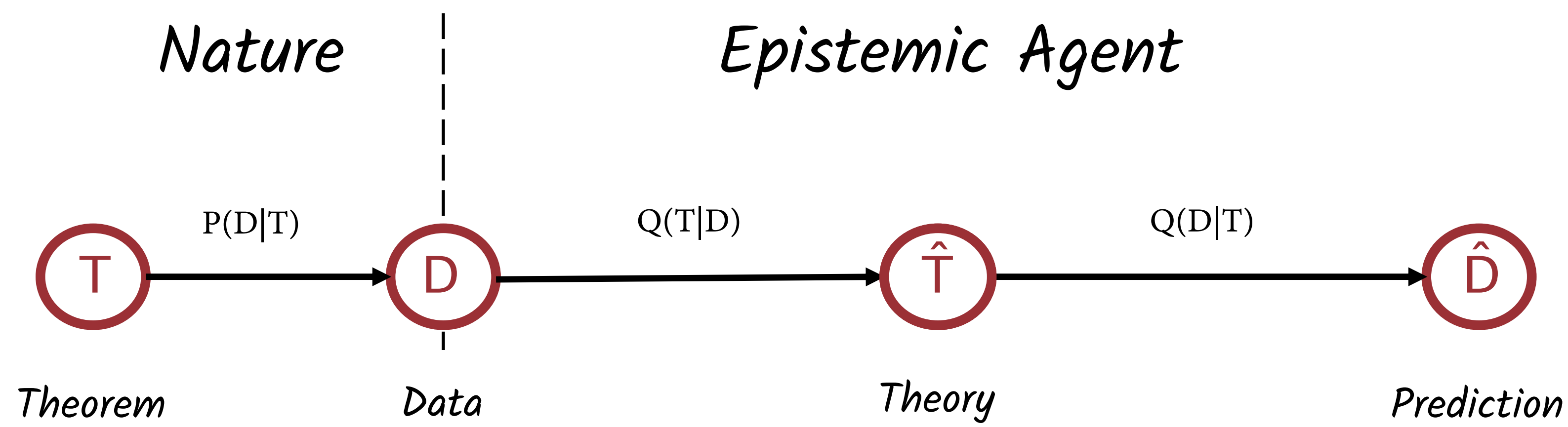


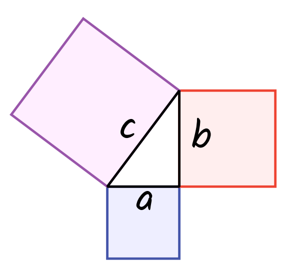
$$Я_Q = I[D;\hat{T}]$$

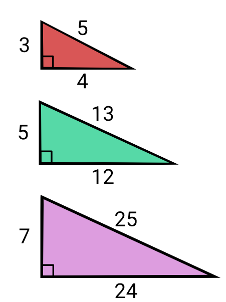
$$d_Q = \langle d(D;\hat{T})\rangle$$

Information Theoretical Learning

Learning as a communication problem







$$a^2 + b^2 = c^2$$

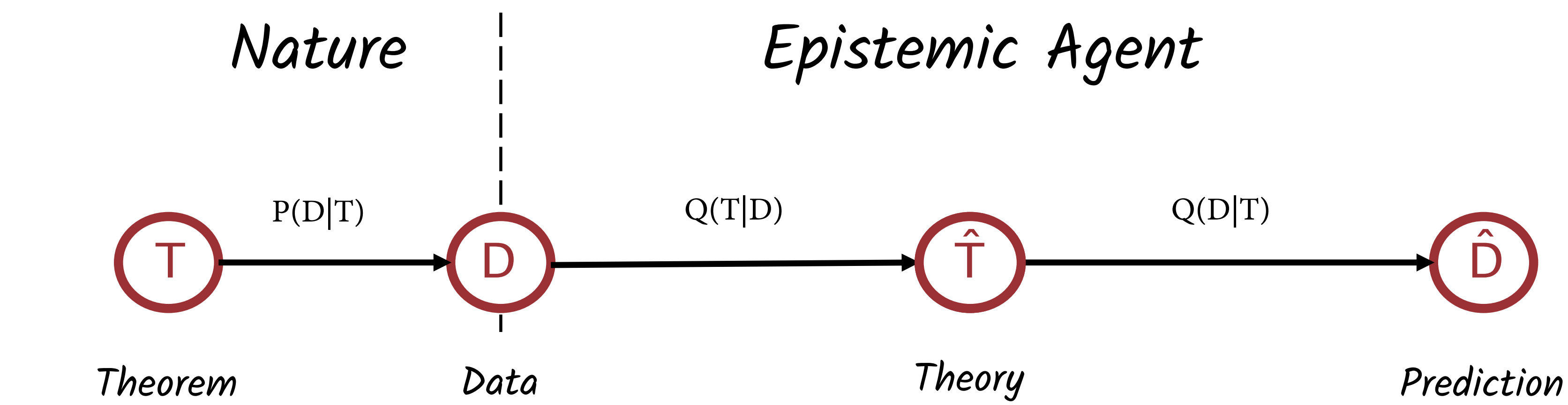
$$Я_Q^{a^2+b^2=c^2} = I[D;\hat{T}] \quad < Я_Q$$

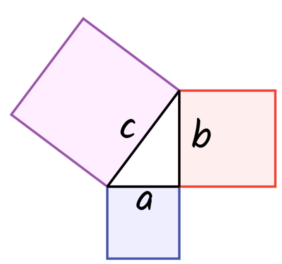
$$d_Q^{a^2+b^2=c^2} = \langle d(D;\hat{T})\rangle < d_Q$$

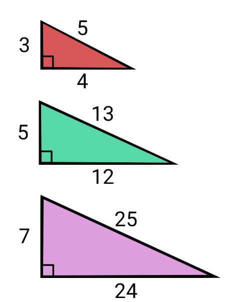
Universidade de Brasília

Information Theoretical Learning

Learning as a communication problem







$$a^2 + b^2 = c^2$$

$$Я_Q^{a^2+b^2=c^2} = I[D;\hat{T}] \quad < Я_Q$$

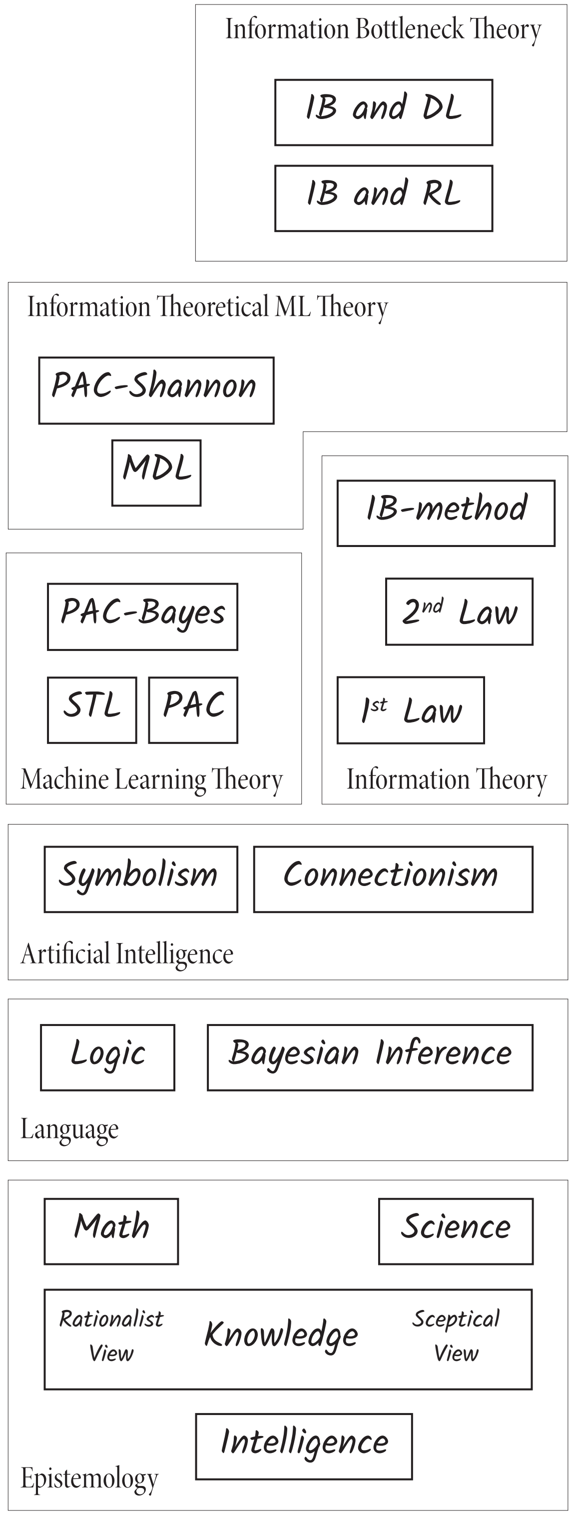
$$d_Q^{a^2+b^2=c^2} = \langle d(D;\hat{T}) \rangle < d_Q$$

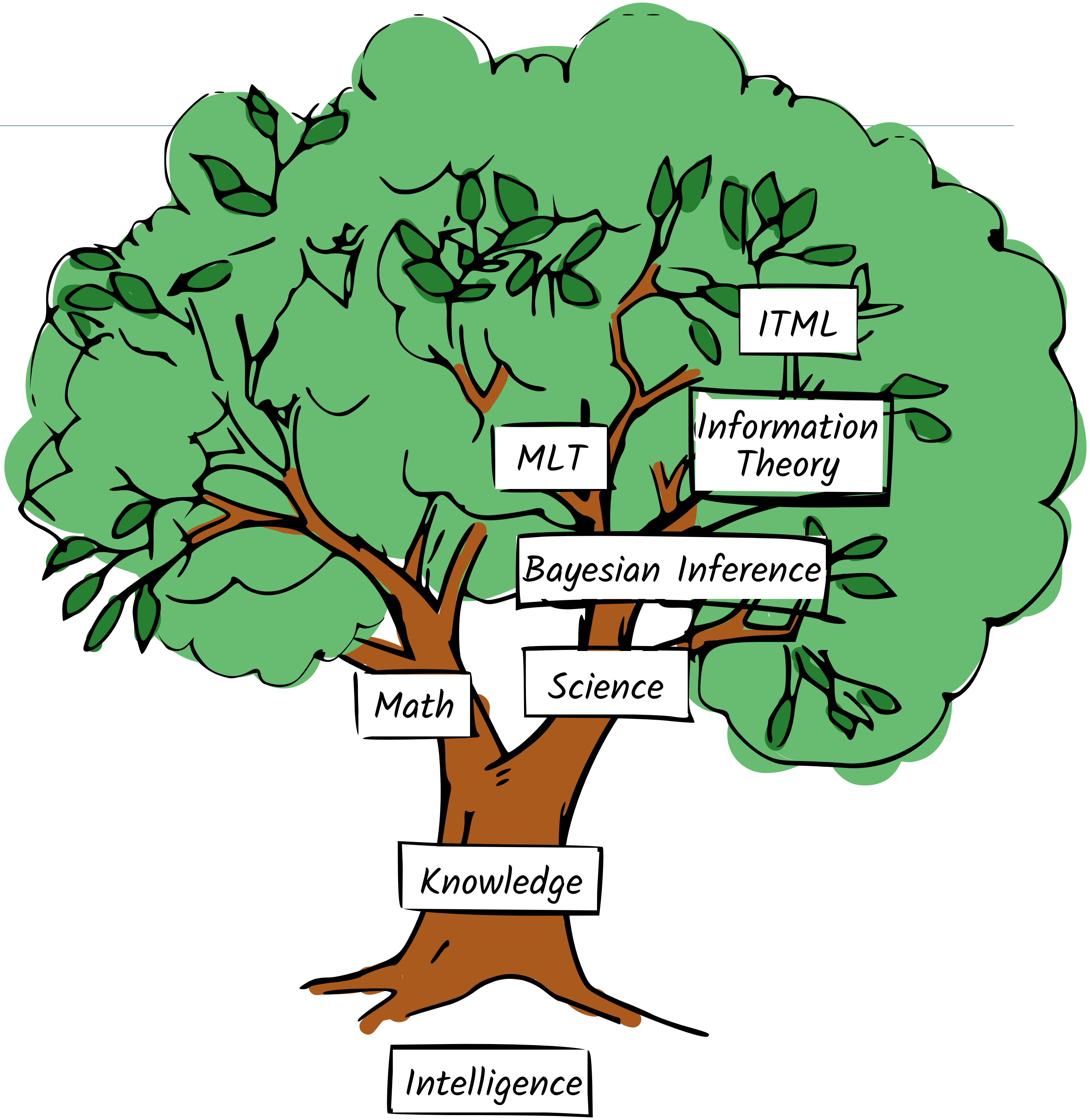
$$Я(\epsilon) \equiv \min_{Q:\ \langle d(x;z)\rangle \leq \epsilon} I[D;\hat{T}]$$

Universidade de Brasília

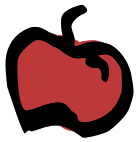
MLT vs. ITML

From the ground up





MLT



- P(X,Y) is fixed, no “time” parameter;
- Optimisation problem: search;
- Loss-metric agnostic (Risk function);

ITML

- P(X,Y) is fixed, no “time” parameter;
- Optimisation problem: compression;
- Loss-metric agnostic (Distortion function);

MLT

- P(X,Y) is fixed, no "time" parameter;
- Optimisation problem: search;
- Loss-metric agnostic (Risk function);

ITML

- P(X,Y) is fixed, no "time" parameter;
- Optimisation problem: compression;
- Loss-metric agnostic (Distortion function);

MLT

- P(X,Y) is fixed, no “time” parameter;
- Optimisation problem: search;
- Loss-metric agnostic (Risk function);
- Hypothesis-space dependent;
- Task independent;
- Continuous random variables;
- Possibly infinite input and target spaces;
- Unknown P(Y|X) can be deterministic;
- Independent sampling;

ITML

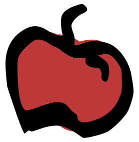
- P(X,Y) is fixed, no “time” parameter;
- Optimisation problem: compression;
- Loss-metric agnostic (Distortion function);
- Task dependent;
- Hypothesis-space independent;
- Discrete random variables;
- Finite input and target spaces;
- Unknown P(Y|X) is stochastic;
- Ergodic process sampling;

MLT

- P(X,Y) is fixed, no "time" parameter;
- Optimisation problem: search;
- Loss-metric agnostic (Risk function);
- Hypothesis-space-dependent;
- Task-independent;
- Continuous random variables;
- Possibly infinite input and target spaces;
- Unknown P(Y|X) can be deterministic;
- Independent sampling;

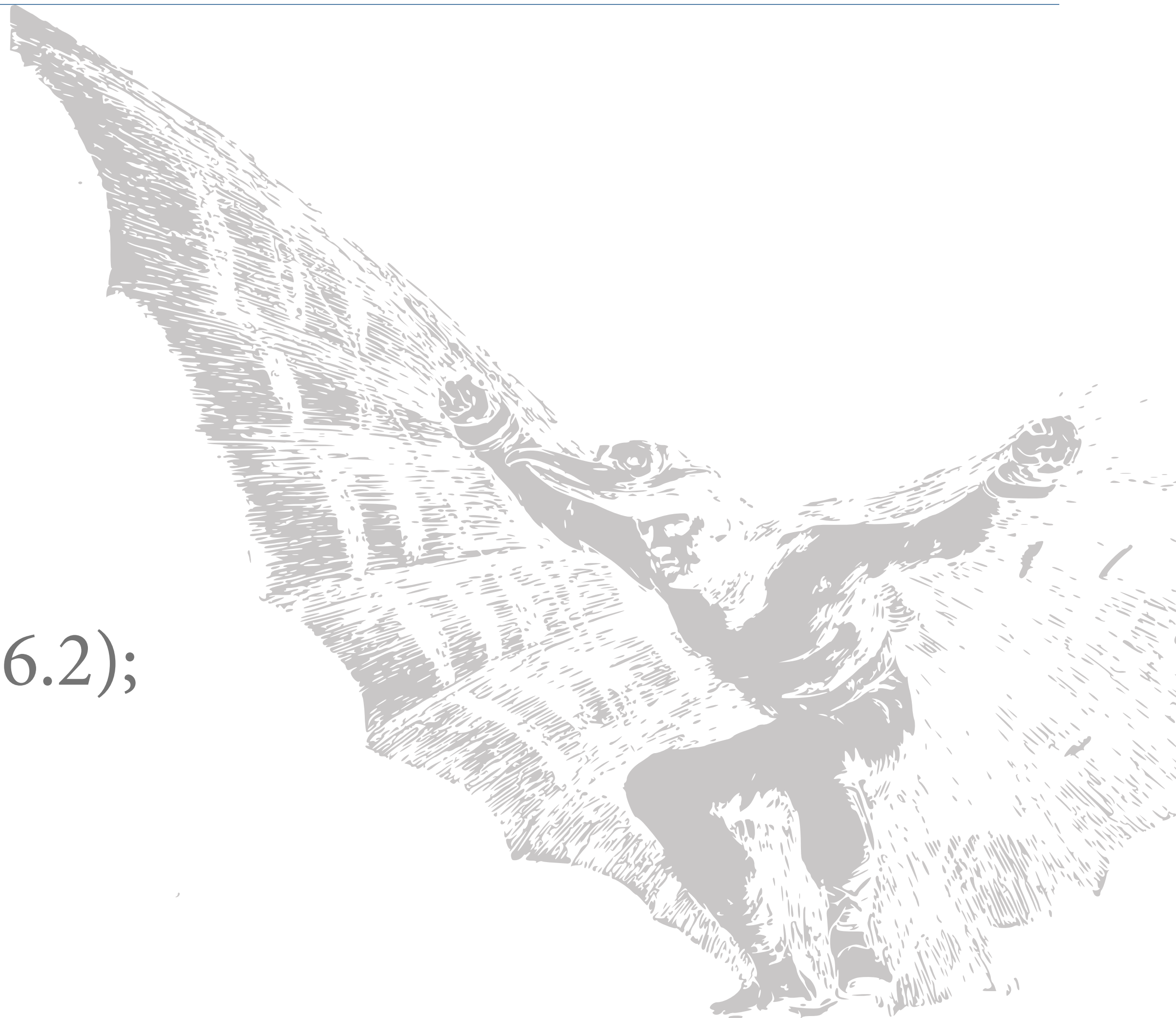
ITML

- P(X,Y) is fixed, no "time" parameter;
- Optimisation problem: compression;
- Loss-metric agnostic (Distortion function);
- Task-dependent;
- Hypothesis-space-independent;
- Discrete random variables;
- Finite input and target spaces;
- Unknown P(Y|X) is stochastic;
- Ergodic process sampling;

 $\equiv$ 

Answering Research Questions 1 to 4

If MLT ≡ ITML, what is the point ?



MLT vs ITML (IBT included):

Share most assumptions;

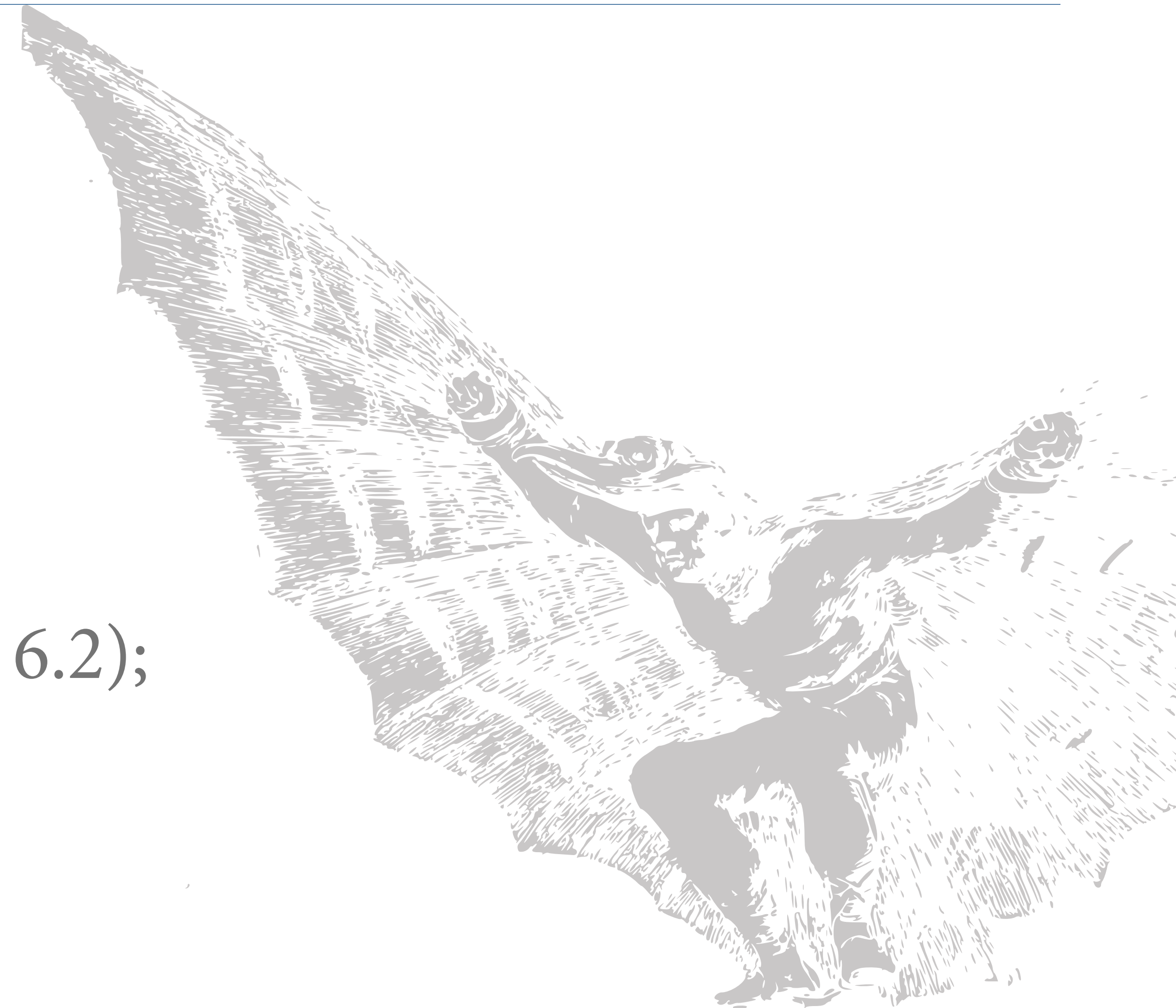
Differences are conciliable choices:
e.g. MDL[HVC93] and PAC-Shannon (sec. 6.2);

What is the point?

[HVC93] Geoffrey E Hinton and Drew Van Camp. "Keeping the neural networks simple by minimizing the description length of the weights". In: Proceedings of the sixth annual conference on Computational learning theory. 1993, pp. 5–13.

Answering Research Questions 1 to 4

If MLT ≡ ITML, what is the point ?



MLT vs ITML (IBT included):

Share most assumptions;

Differences are conciliable choices:
e.g. MDL[HVC93] and PAC-Shannon (sec. 6.2);

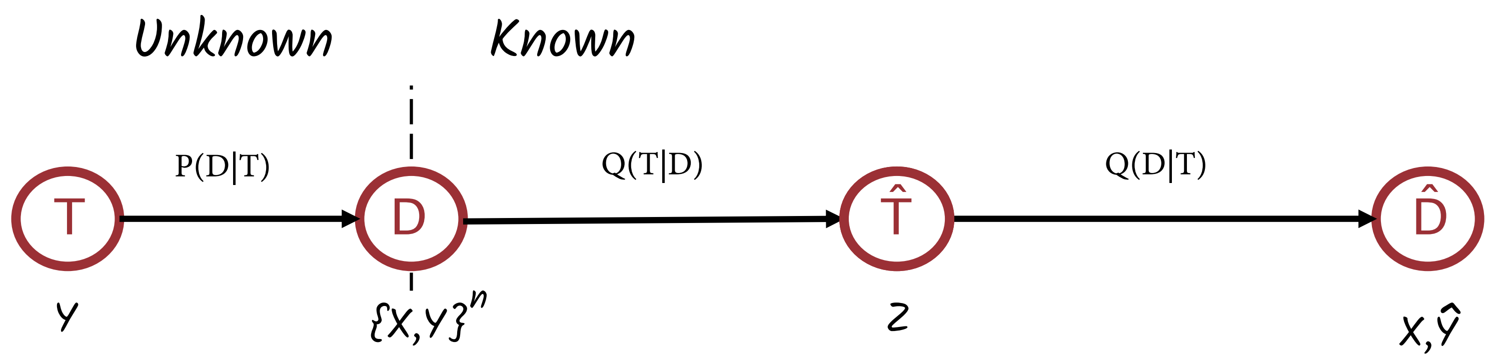
What is the point? A new narrative.

[Mac02] David J. C. MacKay. Information Theory, Inference, and Learning Algorithms. USA: Cambridge University Press, 2002. isbn: 0521642981.

IB Principle

Relevance through a target variable

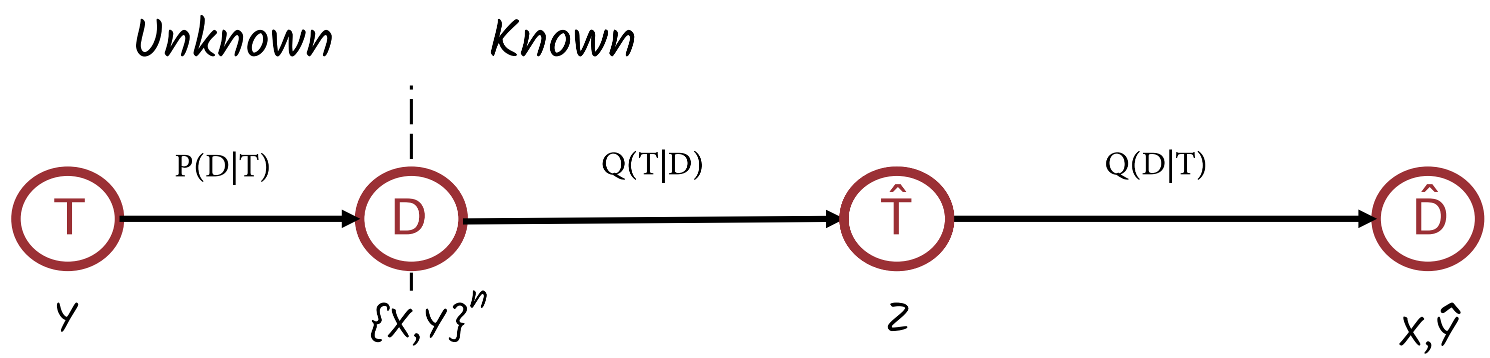
An arbitrary distortion function is an arbitrary feature selection [TPB99].



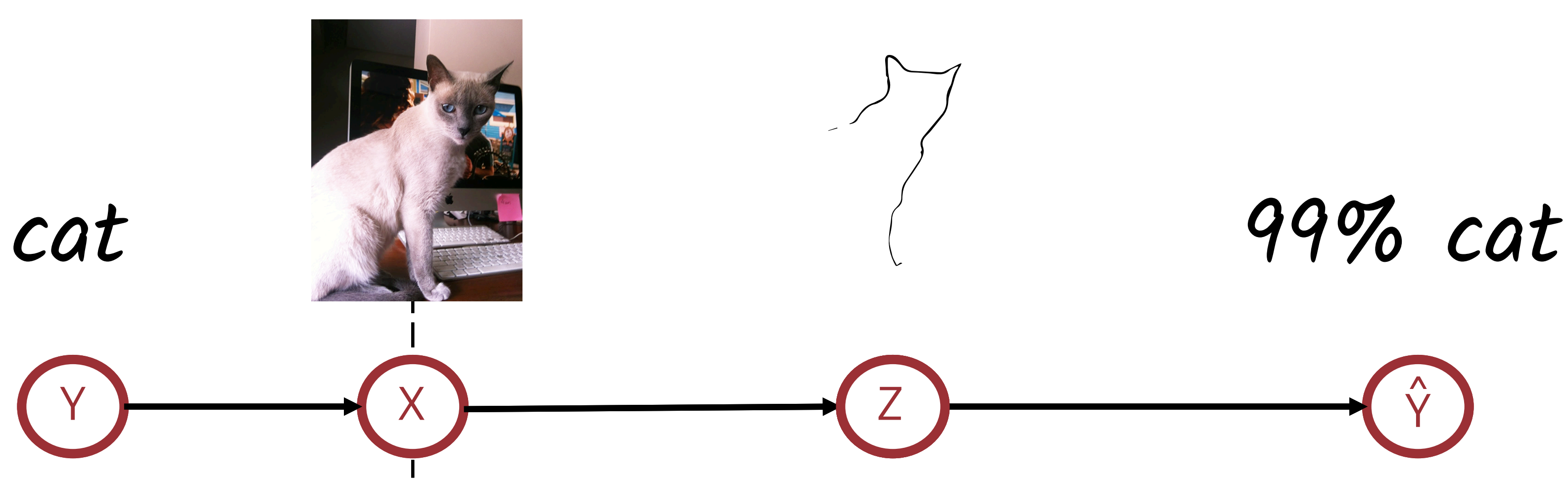
IB Principle

Relevance through a target variable

An arbitrary distortion function is an arbitrary feature selection [TPB99].



Relevance is task-dependent.



[TPB99] Naftali Tishby, Fernando C. Pereira, and William Bialek. “The Information Bottleneck Method”. In: Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing. 1999, pp. 368–377.

IB Principle

Relevance through a target variable

An arbitrary distortion function is an arbitrary feature selection [TPB99].

Theorem 7.2. *If* $\langle d[x;z]\rangle_{p(x,z)} = \mathrm{I}[X;Y] - \mathrm{I}[Z;Y]$, *then* $d[x;z] = \mathrm{D}_{\mathrm{KL}}(p(y|x)||p(y|z))$.

Relevance is task-dependent.

$$\text{Я}^{\mathrm{IB}}(\epsilon) = \min_{Q:\ I[X;Y]-I[Z;Y]\leq\epsilon} I[Z;X]$$

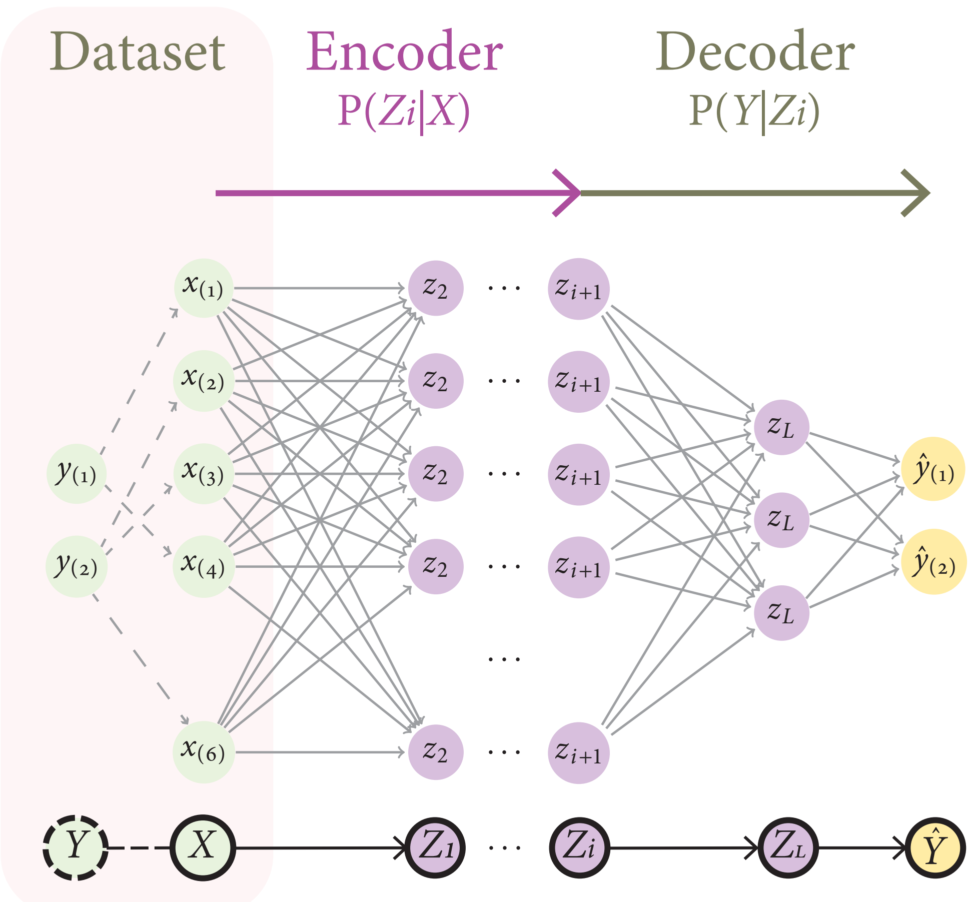
$$\mathscr{L}(Z) = I[Z;X] + \beta I[Z;Y]$$

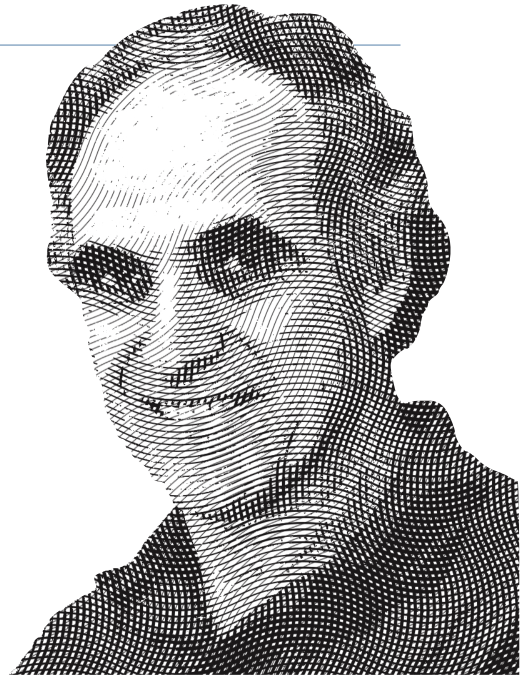
IB Lagrangian

$$Y \rightarrow X \rightarrow Z \rightarrow \hat{Y}$$

INFORMATION BOTTLENECK THEORY

Information Bottleneck principle applied to Deep Learning





Naftali Tishby

What for?

Analysis, opening the “black-box” [ST17].

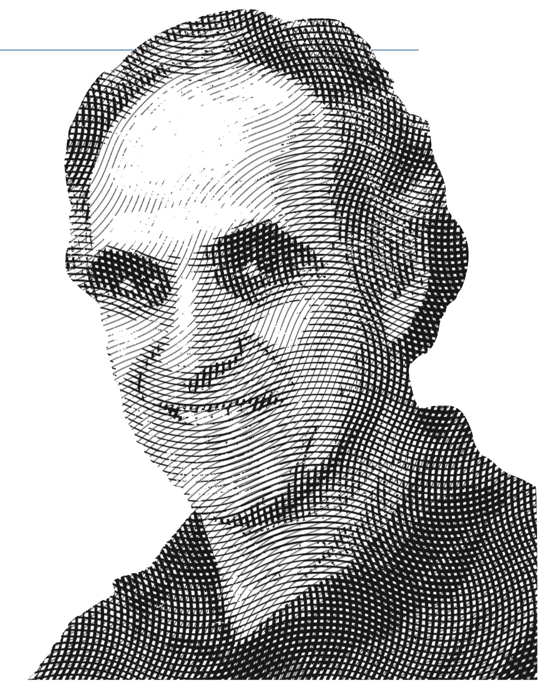
[TZ15, ST17]

[ST17] Ravid Shwartz-Ziv and Naftali Tishby. “Opening the Black Box of Deep Neural Networks via Information”. In: (2017). arXiv: 1703.00810.

[TZ15] Naftali Tishby and Noga Zaslavsky. “Deep learning and the information bottleneck principle”. In: 2015 IEEE Information Theory Workshop (ITW). IEEE. 2015, pp. 1–5.

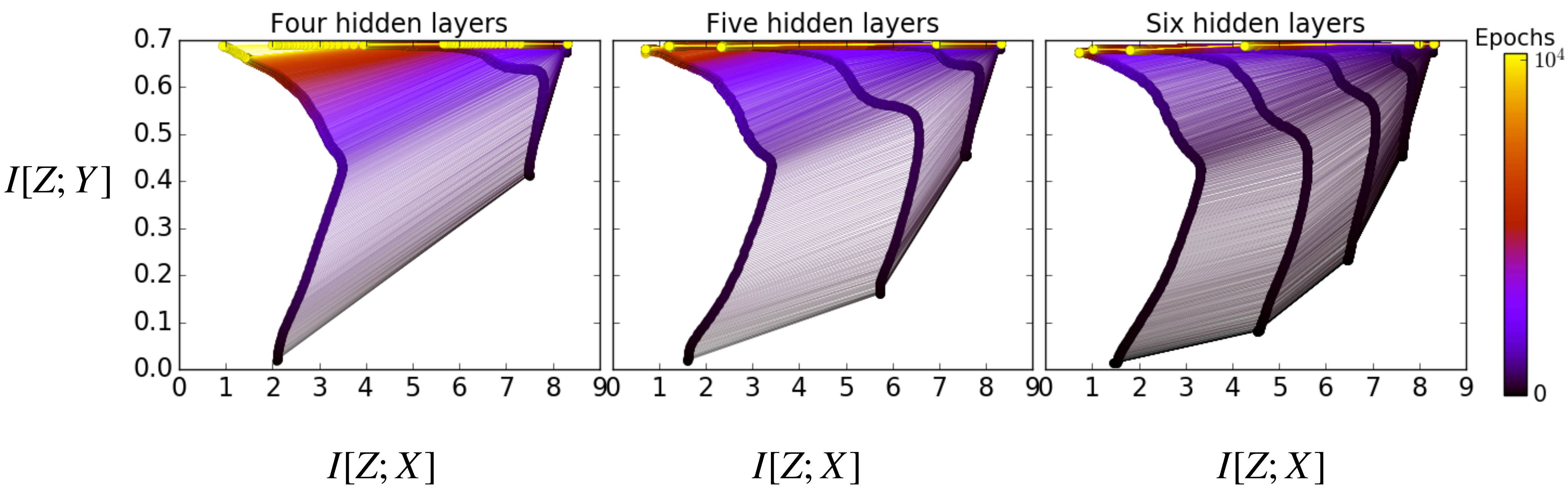
IBT Main thesis

Learning is forgetting



Naftali Tishby

Phase transition during training:
Fitting phase vs. Compression phase.



Universidade de Brasília

IBT Criticism

"Throwing the baby with the bathwater"?

Several papers challenged IBT initial efforts [Sax+18, Gol+19, CHO19] for different reasons:

- Discrete versus continuous random variables;
- IB is ill-posed for deterministic or invertible functions;
- Information in the activations: Stochastic mapping? Why? How?
- Information measurement did not convince;
- "Just an analysis tool" versus "a new Deep Learning Theory";
- Analysis overlooked for lack of confidence in the theory.

[Sax+18] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. "On the Information Bottleneck Theory of Deep Learning". In: International Conference on Learning Representations. 2018.

[Gol+19] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Brian Kingsbury, Igor Melnyk, Nam Nguyen, and Yury Polyanskiy. Estimating Information Flow in DNNs. 2019.

[CHO19] Ivan Chelombiev, Conor Houghton, and Cian O'Donnell. "Adaptive Estimators Show Information Compression in Deep Neural Networks".

IBT Criticism

"Throwing the baby with the bathwater"?

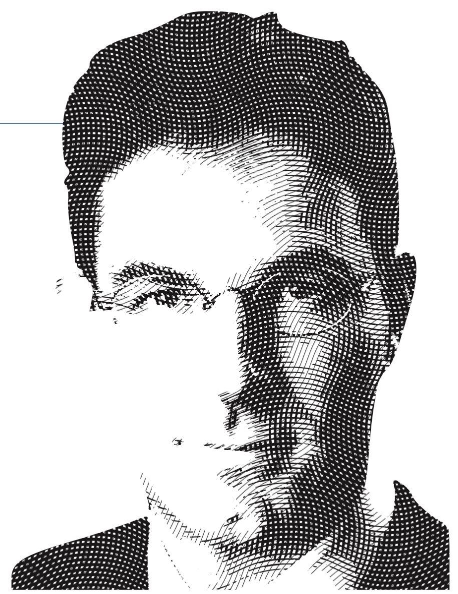
Several papers challenged IBT initial efforts [Sax+18, Gol+19, CHO19] for different reasons:

- Discrete versus continuous random variables;
- IB is ill-posed for deterministic or invertible functions;
- Information in the activations: Stochastic mapping? Why? How?
- Information measurement did not convince;
- “Just an analysis tool” versus “a new Deep Learning Theory”;
- Analysis overlooked for lack of confidence in the theory.

> ‘I would not call [IBT] a proven rigorous theory’
> — Tishby[Tis20].

IB and Representation Learning

Filling the gaps



Stefano Soatto

Prof. Soatto's team extensive body of work:

- Addresses the problem of bounding the information in the activations;

[AS19] Alessandro Achille and Stefano Soatto. Where is the Information in a Deep Neural Network? 2019. arXiv: 1905.12213 [cs.LG].

- Explains the emergence of generalisation and disentanglement;

[AS18a] Alessandro Achille and Stefano Soatto. "Emergence of Invariance and Disentangling in Deep Representations". In: J. Mach. Learn. Res. 19.1 (Jan. 2018), pp. 1947–1980. issn: 1532-4435.

- Shows the crucial role of noise in generalisation;

[CS18] P. Chaudhari and S. Soatto. "Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks". In: 2018 Information Theory and Applications Workshop (ITA). 2018, pp. 1–10. doi: 10.1109/ITA.2018.8503224.

- Proposes a variational method for estimating mutual information;

[AS18b] Alessandro Achille and Stefano Soatto. "Information Dropout: Learning Optimal Representations Through Noisy Computation". In: IEEE Transactions on Pattern Analysis and Machine Intelligence 40.12 (2018), pp. 2897–2905.

- Relates the information in the weights to PAC-Bayes.

[AS18b]

… and more.

[AMS18] Alessandro Achille, Glen Mbeng, and Stefano Soatto. Dynamics and Reachability of Learning Tasks. 2018. arXiv: 1810.02440.

[ARS17] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical Learning Periods in Deep Neural Networks. 2017. arXiv: 1711.08856.

[Ach+19a]Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. "Task2Vec: Task Embedding for Meta-Learning". In: The IEEE International Conference on Computer Vision (ICCV). Oct. 2019.

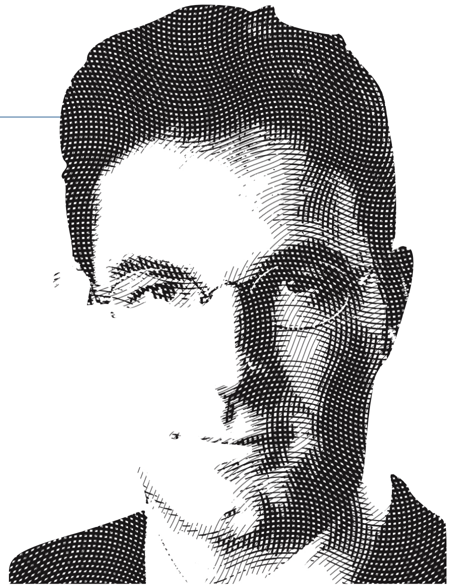
IB and Representation Learning

Filling the gaps

Prof. Soatto's team extensive body of work:

- Addresses the problem of bounding the information in the activations;
- Explains the emergence of generalisation and disentanglement;
- Shows the crucial role of noise in generalisation;
- Proposes a variational method for estimating mutual information;
- Relates the information in the weights to PAC-Bayes.

… and more.

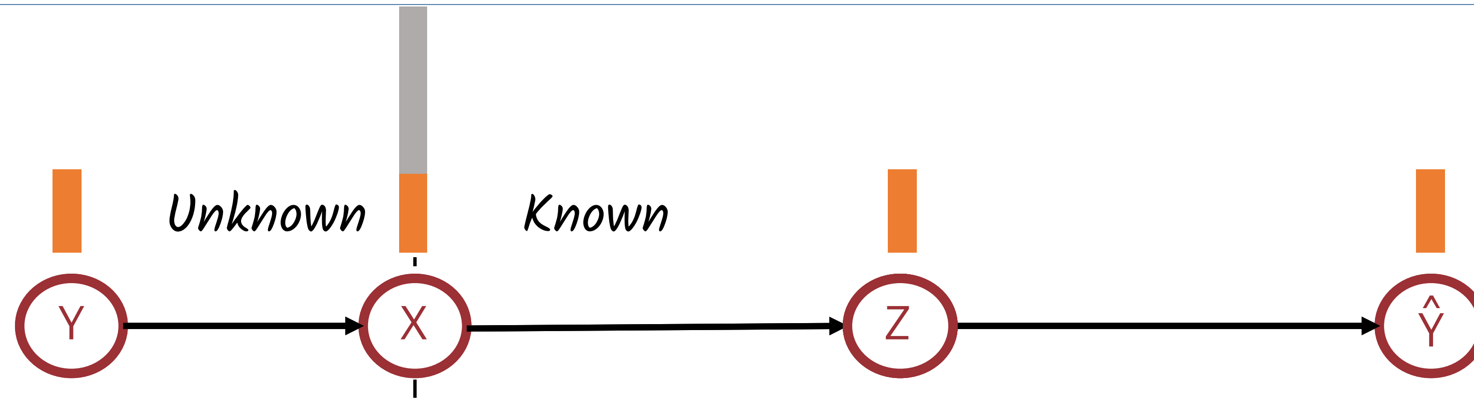


Stefano Soatto

Desiderata for Representations → Activations IB → Activations IB x Weights IB → IB Learning X Deep Learning

Desiderata for Representations

What is a good representation?



The best representation $Z := P(Z|X)$ of data X for task $Y := P(Y|X)$ is [AS18a]:

sufficient: $I[Z;Y] = I[X;Y]$

invariant: $\eta \perp Y \rightarrow I[\eta;Y] = 0 \rightarrow I[\eta;Z] = 0$

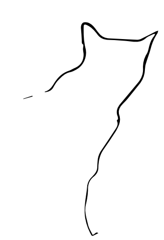
minimal: $I[Z;X] = I[Z;Y]$

disentangled: $TC(Z) = D_{KL}(P(Z)\| \prod_{i=1}^{n} P(Z_i)) = 0$

accuracy

generalisation

explainability

sufficient    minimal

Desiderata for Representations

What is a good representation?

A good representation can be formulated as [AS18a]:

$$Z := \arg\min I[Z; X] \qquad \text{minimal/invariant}$$

s.t.

$$0 \leq I[X; Y] - I[Z; Y] \qquad \text{sufficient}$$

$$0 \leq TC(Z)\,. \qquad \text{disentangled}$$

Desiderata for Representations

What is a good representation?

A good representation can be formulated as [AS18a]:

$$Z := \arg\min I[Z;X] \qquad \text{minimal}$$

s.t.

$$0 \leq I[X;Y] - I[Z;Y] \qquad \text{sufficient}$$

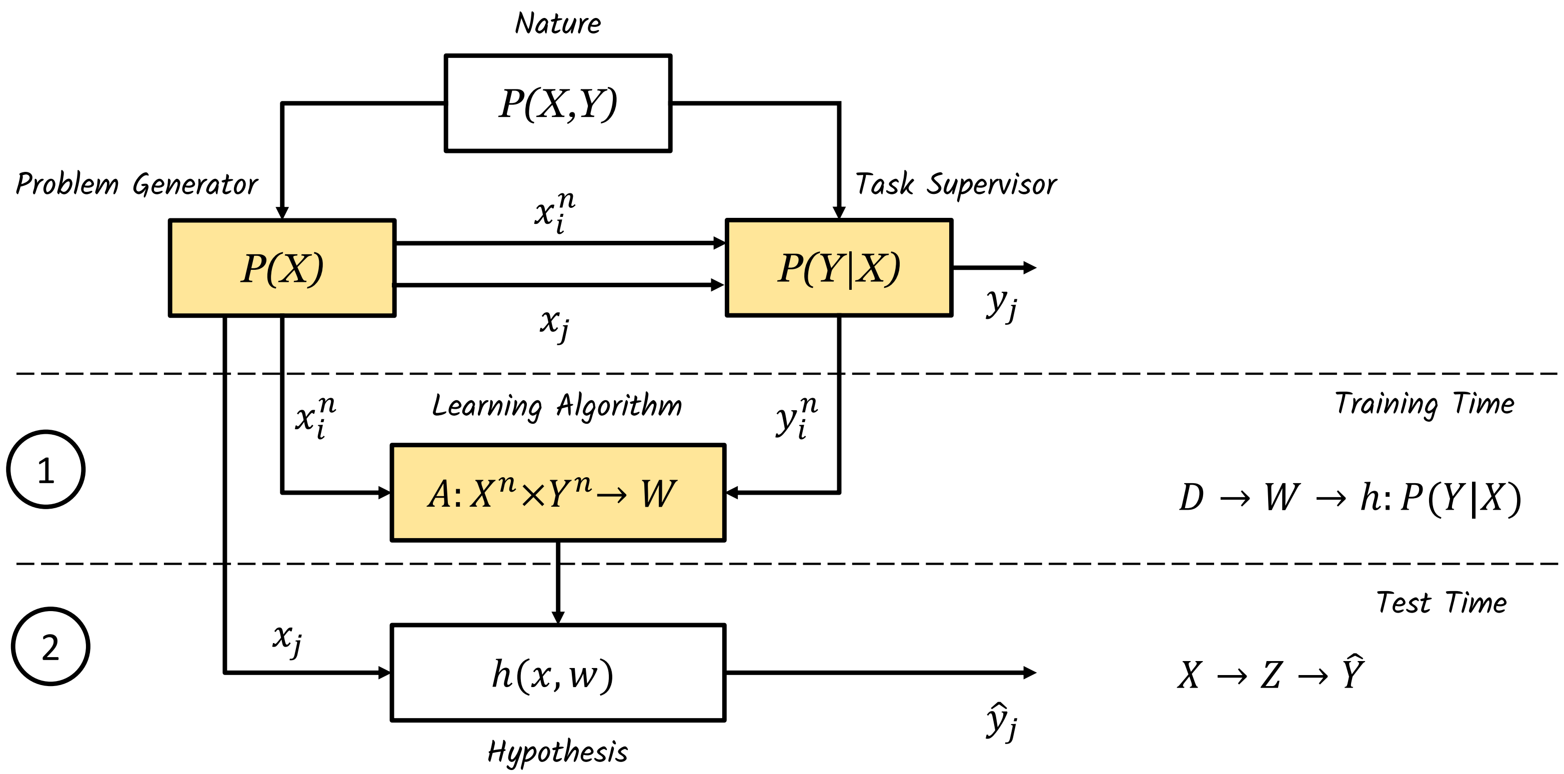
$$0 \leq TC(Z)\,. \qquad \text{disentangled}$$

Using the Lagrangian relaxation:

$$L(Z) = \underbrace{H_{p,q}[Y|Z]}_{\text{[TPB99, ST17]}} + \beta^{-1}\{I[Z;X] + TC(Z)\} \qquad \text{Activations IB [AS18a]}$$

The IB "Achille's heel"

Two levels of representations



Alessandro Achille

$$L(Z) = H_{p,q}[Y \mid Z] + \beta^{-1} I[Z; X]$$

Activations IB
[TPB99, ST17]

Activations IB is incomputable:

Z is a representation of yet not observed future data.

Valid min I[Z;X] during training → memorise indexes of each label.

Once the weights are fixed, not a stochastic mapping.

No access to true distribution P(X,Y).

Rethinking Generalisation

Cross-entropy decomposition and overfitting

Problem: Deep Learning pseudo-paradox [Zha+16].
→ can fit random labels, yet generalise;

Cross-entropy decomposition, assuming $D \sim P(D \mid \theta)$ [AS18a]:

$$H_{p,q}[D \mid W] = \underbrace{H_p[D \mid \theta]}_{\text{intrinsic error}} + \underbrace{I[\theta; D \mid W]}_{\text{sufficiency}} + \underbrace{D_{KL}(p \parallel q)}_{\text{efficiency}} - \underbrace{I[D; W \mid \theta]}_{\text{memorisation}}$$

Rethinking Generalisation

Cross-entropy decomposition and overfitting

Problem: Deep Learning pseudo-paradox [Zha+16].
→ can fit random labels, yet generalise;

Cross entropy decomposition, assuming $D \sim P(D\,|\,\theta)$ [AS18a]:

$$H_{p,q}[D\,|\,W] = \underbrace{H_p[D\,|\,\theta]}_{\text{intrinsic error}} + \underbrace{I[\theta; D\,|\,W]}_{\text{sufficiency}} + \underbrace{D_{KL}(p \,\|\, q)}_{\text{efficiency}} - \underbrace{I[D; W\,|\,\theta]}_{\text{memorisation}}$$

Naïve solution:

$$L(W) = H_{p,q}[D\,|\,W] + I[D; W\,|\,\theta]$$

intractable, θ is unknown.

Rethinking Generalisation

Cross-entropy decomposition and overfitting

Problem: Deep Learning pseudo-paradox [Zha+16].
→ can fit random labels, yet generalise;

Cross entropy decomposition, assuming $D \sim P(D \mid \theta)$ [AS18a]:

$$H_{p,q}[D \mid W] = \underbrace{H_p[D \mid \theta]}_{\text{intrinsic error}} + \underbrace{I[\theta; D \mid W]}_{\text{sufficiency}} + \underbrace{D_{KL}(p \parallel q)}_{\text{efficiency}} - \underbrace{I[D; W \mid \theta]}_{\text{memorisation}}$$

Naïve solution:

$$L(W) = H_{p,q}[D \mid W] + I[D; W \mid \theta]$$

intractable, θ is unknown.

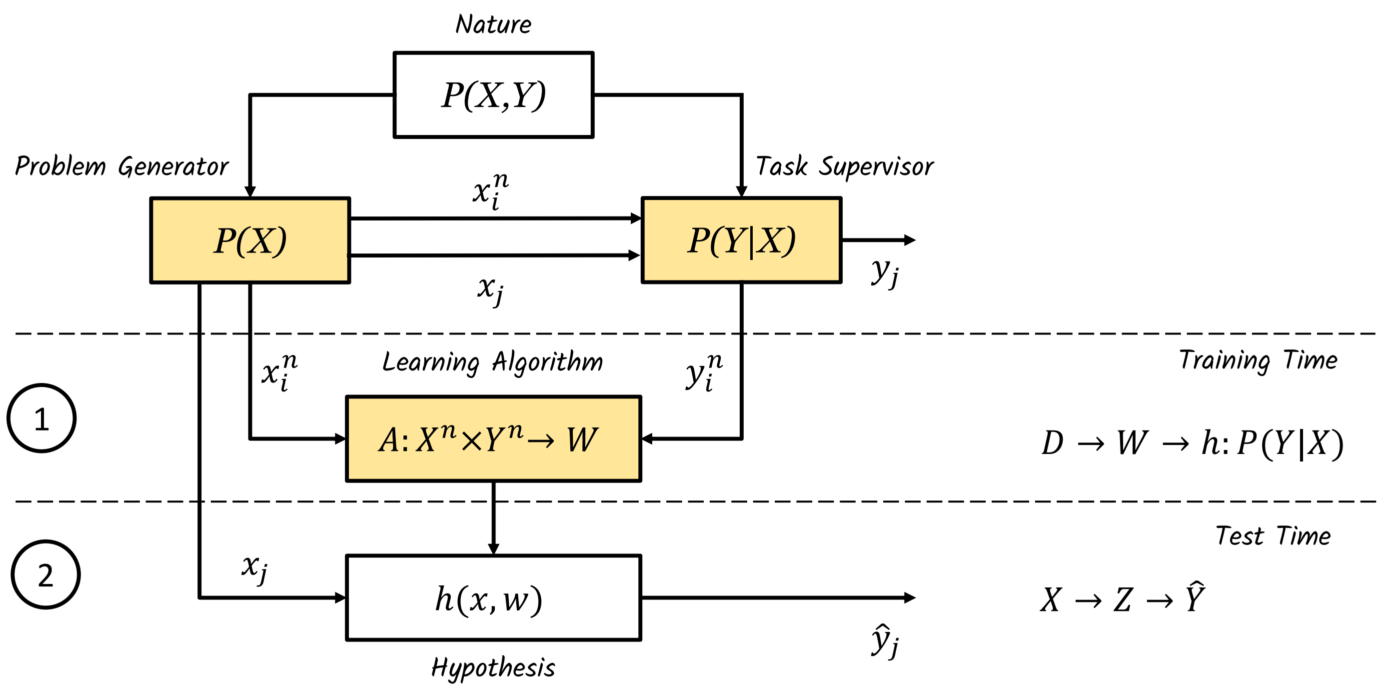
But we can upper bound $I[D; W \mid \theta]$:

$$L(W) = H_{p,q}[D \mid W] + \beta^{-1} I[D; W]$$

Weights IB [AS18a, AS19]

Activations IB vs. Weights IB

Where is the information in Deep Neural Networks?



Weights IB
[AS18a, AS19]

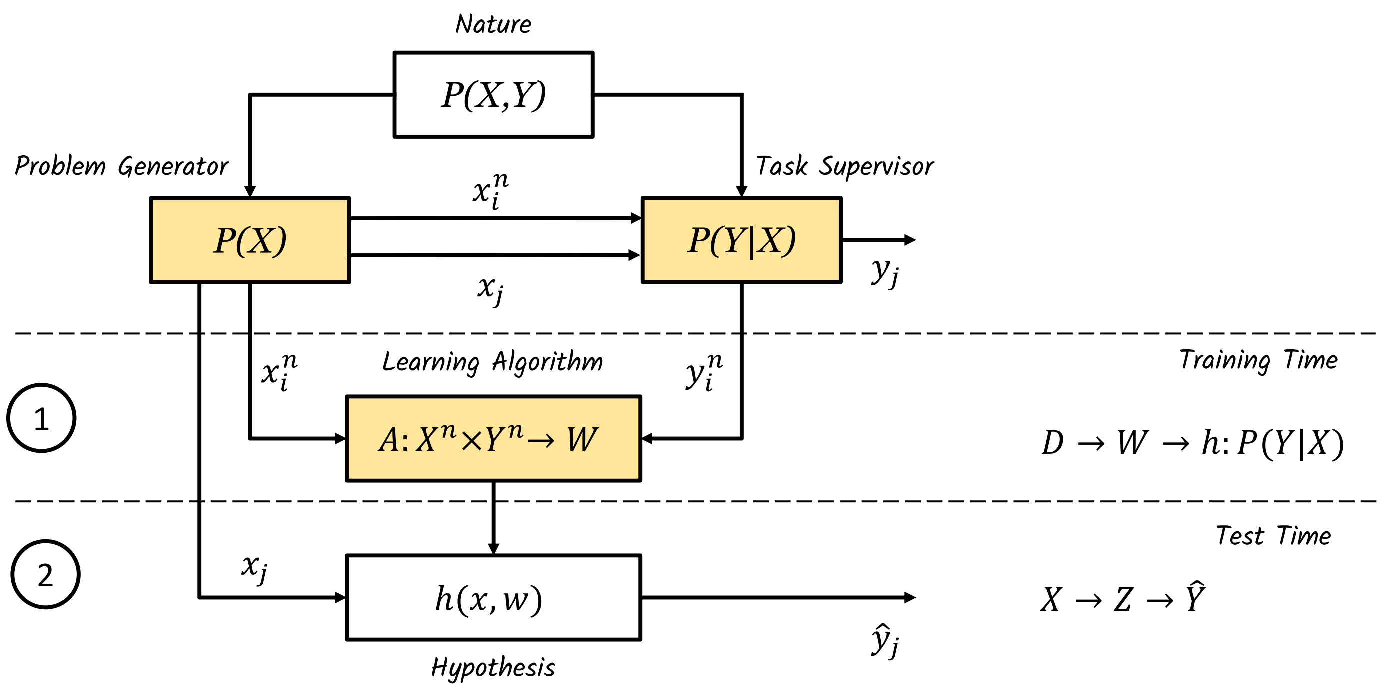
$$\mathscr{L}(W) = H_{p,q}[D \mid W] + \beta^{-1} I[D; W]$$

$$\mathscr{L}(Z) = H_{p,q}[Y \mid Z] + \beta^{-1} I[Z; X]$$

Activations IB
[TPB99, ST17]

Activations IB vs. Weights IB

Where is the information in Deep Neural Networks?



Weights IB
[AS18a, AS19]

$$\mathscr{L}(W) = H_{p,q}[D \mid W] + \beta^{-1} I[D; W]$$

$$\mathscr{L}(Z) = H_{p,q}[Y \mid Z] + \beta^{-1} I[Z; X]$$

Activations IB
[TPB99, ST17]

Bound [C.8 in AS18a]:

$$I[Z; X] \leq I[W; D] \leq \underbrace{\log |F(w^*)|}_{\text{Fisher Information}}$$

Universidade de Brasília

Deep Learning

Reality

Deep Learning components:

DNN Architecture: deep

SGD Optimiser

Large Dataset: P(X,Y) is noisy

Loss function: usually cross-entropy

$$\mathscr{L}(W) = H_{p,q}[D \mid W]$$

IBT Learning

Ideal

$$\mathscr{L}(W) = H_{p,q}[D \mid W] + \underbrace{\beta^{-1} I[D; W]}_{\text{regulariser}}$$

Deep Learning

Reality

Deep Learning components:

DNN Architecture: deep

SGD Optimiser

Large Dataset: P(X,Y) is noisy

Loss function: usually cross-entropy

$$\mathscr{L}(W) = H_{p,q}[D \mid W]$$

IBT Learning

Ideal

$$\mathscr{L}(W) = H_{p,q}[D \mid W] + \underbrace{\beta^{-1} I[D; W]}_{\text{regulariser}}$$

Ways to reduce information:

Explicit regulariser in the loss function:
Information Dropout [As18b]

Implicit by architecture:
Reduce dimension (layers, max-pooling)
Add noise (dropout)

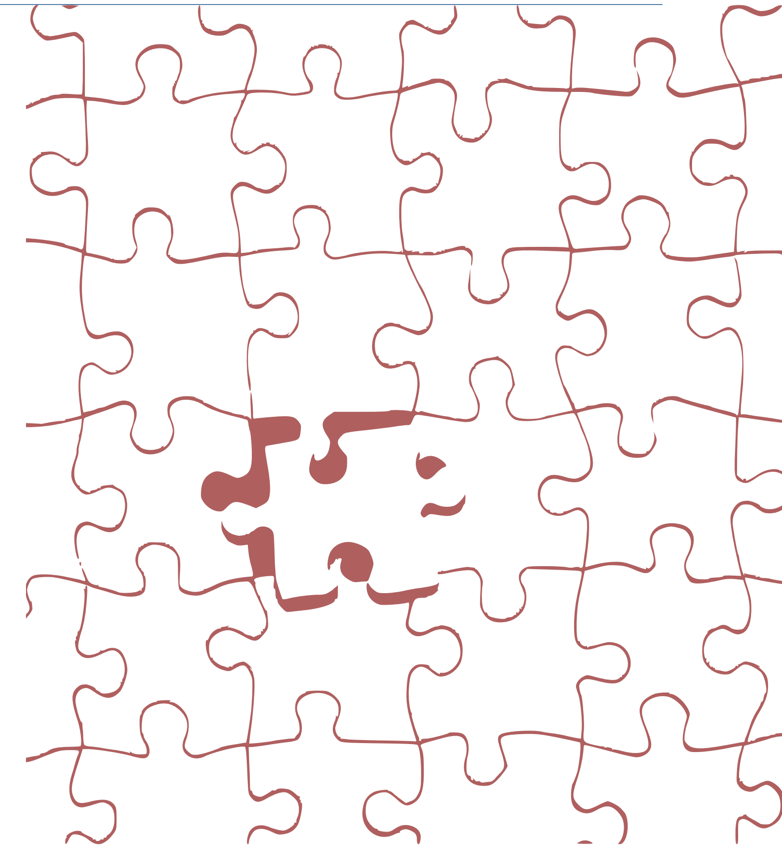
Problem [Zha+16]:
Generalisation without regularisers in the loss or architecture.

Can layers explain it all?

The role of noise in SGD

The last piece of the puzzle

Chaudhari and Soatto [CS18] prove with theory and empirical evidence that:



SGD performs variational inference with an implicit loss;

SGD implicit loss has an information regulariser term.

$$\mathscr{L}(W) = H_{p,q}[D \mid W] + \underbrace{\beta^{-1} I[D; W]}_{\text{SGD implicit regulariser}}$$

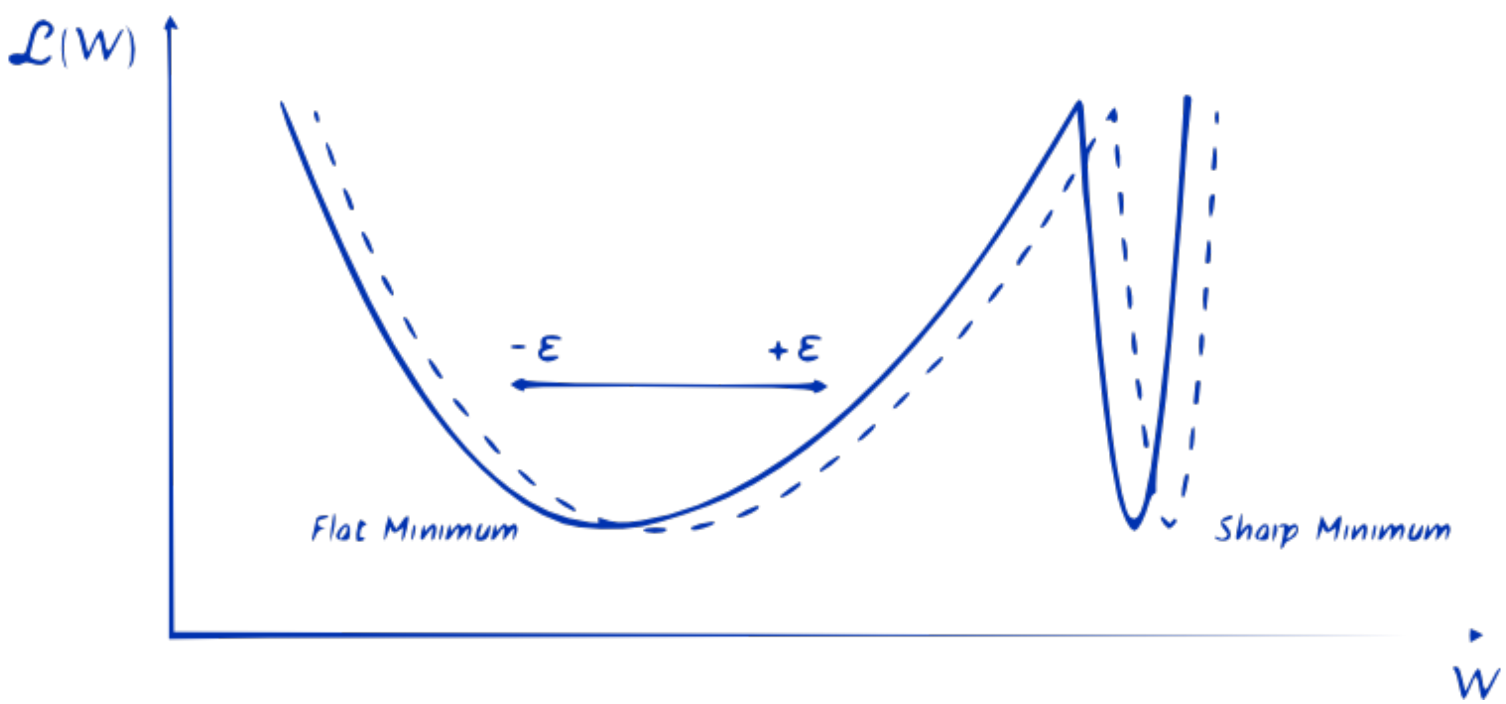
Deep Learning phenomena in the IBT narrative

Answering Research Question 5: Part I

Generalisation despite model capacity/expressiveness:
Information in the weights as the *effective* capacity measure.

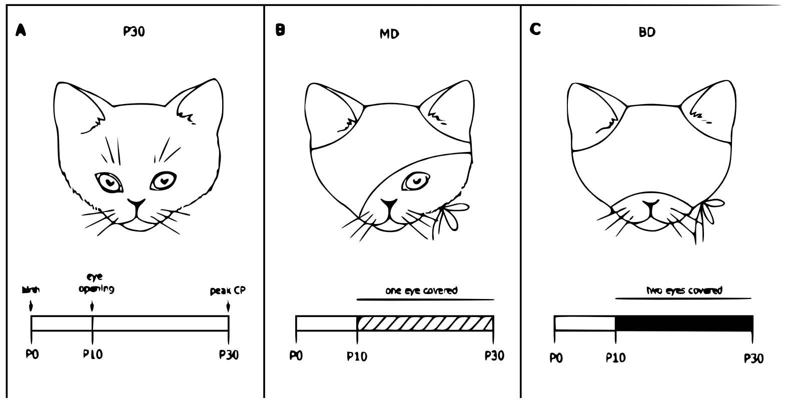
Deep Learning bias towards disentangled representations:
SGD → I[W;D] implicit regulariser → upper-bound on I[Z;X]+TC

Scarcity of sharp minima in SGD optimisation:
SGD → low I[W;D] → low Fisher Information → curvature of loss



DEEP LEARNING PHENOMENA IN THE IBT NARRATIVE

Answering Research Question 5: Part II

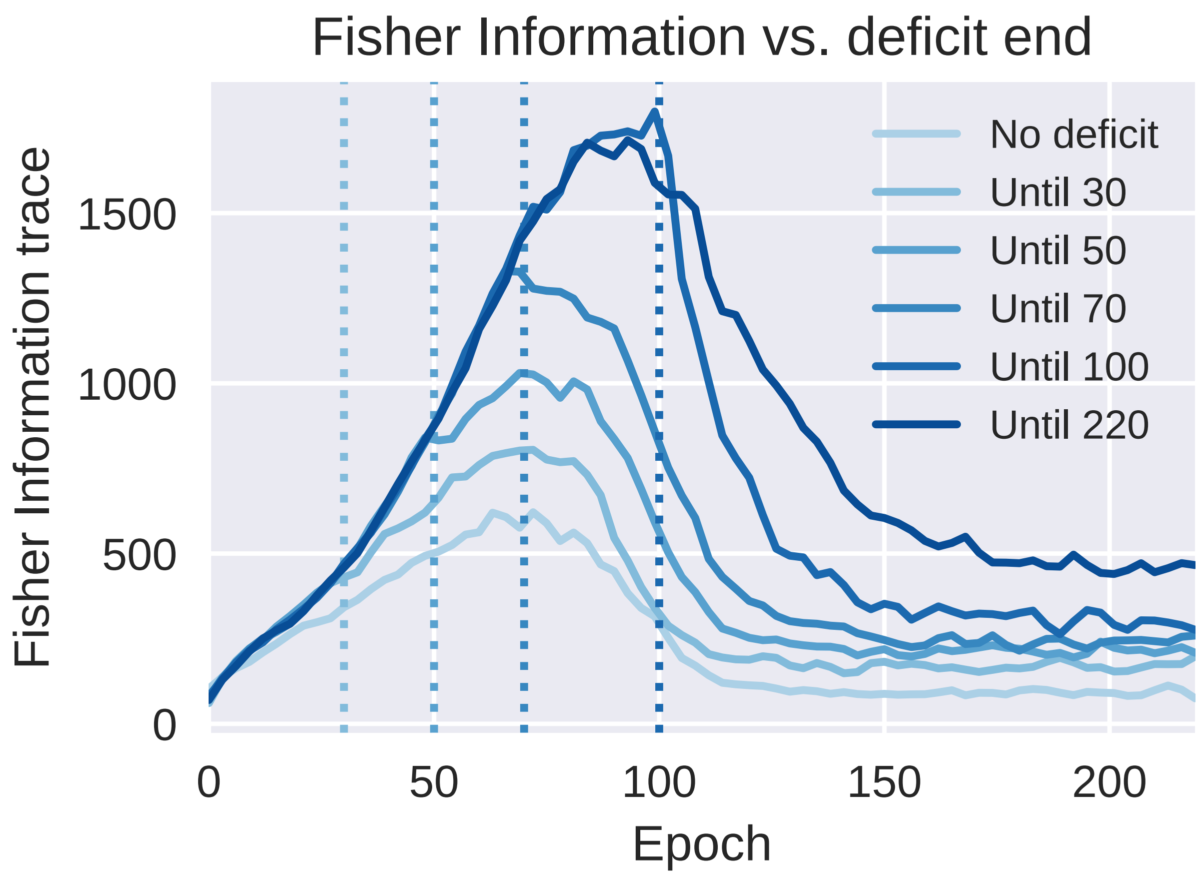


[Wie82]

Critical Learning Periods [ARS17]:

Deficit → higher Fisher Information → memorisation

Phase transition → Fitting phase/high curvature



[Wie82] Torsten N. Wiesel. “Postnatal Development of the Visual Cortex and the Influence of Environment”. In: Nature 299.5884 (Oct. 1982), pp. 583–591. issn: 1476-4687. doi: 10.1038/299583a0.

CONCLUSION

IBT strengths, weaknesses and research opportunities

STRENGTHS

Narrative: connects seemly unrelated phenomena and practices;

Analysis: information in the weights “opens the black-box”;

Task-dependent loss: not arbitrary ;

WEAKNESSES

Lack of rigour: overlooking important assumptions;

Discredit: critiques were hardly unjustified;

Fragmentation: Literature is still very fragmented;

Conclusion

IBT strengths, weaknesses and research opportunities

Research Opportunities

PAC reformulation: β unifies (ϵ, δ);

New optimisation strategies: different approaches for the fitting and compression phases;

Transfer learning: Validate topologies of learning tasks built from IBT(e.g. Task2Vec [Ach+19]), with empirical ones (e.g. Taskonomy[Zam+18]);

[Zam+18] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. "Taskonomy: Disentangling task transfer learning". In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 3712–3722.

Conclusion

Information Bottleneck Theory,
far from being rigorous and complete,
is an emerging and exciting topic
with a compelling narrative
and many open opportunities.

References

In alphabetical order

[Ach+19a] Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. "Task2Vec: Task Embedding for Meta-Learning". In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[AMS18] Alessandro Achille, Glen Mbeng, and Stefano Soatto. *Dynamics and Reachability of Learning Tasks*. 2018. arXiv: `1810.02440 [cs.LG]`.

[Ach+19b] Alessandro Achille, Giovanni Paolini, Glen Mbeng, and Stefano Soatto. *The Information Complexity of Learning Tasks, their Structure and their Distance*. 2019. arXiv: `1904.03292 [cs.LG]`.

[ARS17] Alessandro Achille, Matteo Rovere, and Stefano Soatto. *Critical Learning Periods in Deep Neural Networks*. 2017. arXiv: `1711.08856 [cs.LG]`.

[AS18a] Alessandro Achille and Stefano Soatto. "Emergence of Invariance and Disentangling in Deep Representations". In: *J. Mach. Learn. Res.* 19.1 (Jan. 2018), pp. 1947–1980. ISSN: 1532-4435.

[AS18b] Alessandro Achille and Stefano Soatto. "Information Dropout: Learning Optimal Representations Through Noisy Computation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.12 (2018), pp. 2897–2905.

[AS19] Alessandro Achille and Stefano Soatto. *Where is the Information in a Deep Neural Network?* 2019. arXiv: `1905.12213 [cs.LG]`.

[CS18] P. Chaudhari and S. Soatto. "Stochastic Gradient Descent Performs Variational Inference, Converges to Limit Cycles for Deep Networks". In: *2018 Information Theory and Applications Workshop (ITA)*. 2018, pp. 1–10. DOI: `10.1109/ITA.2018.8503224`.

[CHO19] Ivan Chelombiev, Conor Houghton, and Cian O'Donnell. "Adaptive Estimators Show Information Compression in Deep Neural Networks". In: *International Conference on Learning Representations*. 2019.

[DR17] Gintare Karolina Dziugaite and Daniel M. Roy. "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". In: *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. Ed. by Gal Elidan, Kristian Kersting, and Alexander T. Ihler. AUAI Press, 2017. URL: `http://auai.org/uai2017/proceedings/papers/173.pdf`.

[Gol+19] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Brian Kingsbury, Igor Melnyk, Nam Nguyen, and Yury Polyanskiy. *Estimating Information Flow in DNNs*. 2019. URL: `https://openreview.net/forum?id=HkxOoiAcYX`.

[HVC93] Geoffrey E Hinton and Drew Van Camp. "Keeping the neural networks simple by minimizing the description length of the weights". In: *Proceedings of the sixth annual conference on Computational learning theory*. 1993, pp. 5–13.

[Mac02] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. USA: Cambridge University Press, 2002. ISBN: 0521642981.

[McA13] David A. McAllester. "A PAC-Bayesian Tutorial with A Dropout Bound". In: *CoRR* abs/1307.2118 (2013). arXiv: `1307.2118`.

[Rah18] Ali Rahimi. *Ali Rahimi NIPS 2017 Test-of-Time Award Presentation Speech*. `https://youtu.be/x7psGHgatGM`. [Online; Last accessed on 2020-08-04.] Mar. 7, 2018. URL: `https://youtu.be/x7psGHgatGM`.

[Sax+18] Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. "On the Information Bottleneck Theory of Deep Learning". In: *International Conference on Learning Representations*. 2018.

REFERENCES

In alphabetical order

[ST17] Ravid Shwartz-Ziv and Naftali Tishby. "Opening the Black Box of Deep Neural Networks via Information". In: (2017). arXiv: `1703.00810 [cs.LG]`.

[Tis17a] Naftali Tishby. *Information Theory of Deep Learning*. `https://youtu.be/FSfN2K3tnJU`. [Online; Published: 2017-10-16. Last Accessed: 2020-06-01]. Oct. 16, 2017. URL: `https://youtu.be/FSfN2K3tnJU`.

[Tis17b] Naftali Tishby. *Information Theory of Deep Learning*. `https://youtu.be/bLqJHjXihK8`. [Online; Published: 22017-08-03. Last Accessed: 2020-06-01]. Aug. 3, 2017. URL: `https://youtu.be/bLqJHjXihK8`.

[Tis20] Naftali Tishby. *The Information Bottleneck View of Deep Learning: Why do we need it?* `https://youtu.be/utvIaZ6wYuw`. [Online; Last accessed on 2021-03-12.] Jan. 10, 2020. URL: `https://youtu.be/utvIaZ6wYuw`.

[TPB99] Naftali Tishby, Fernando C. Pereira, and William Bialek. "The Information Bottleneck Method". In: *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*. 1999, pp. 368–377.

[TZ15] Naftali Tishby and Noga Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 IEEE Information Theory Workshop (ITW)*. IEEE. 2015, pp. 1–5.

[Wie82] Torsten N. Wiesel. "Postnatal Development of the Visual Cortex and the Influence of Environment". In: *Nature* 299.5884 (Oct. 1982), pp. 583–591. ISSN: 1476-4687. DOI: `10.1038/299583a0`.

[Zam+18] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. "Taskonomy: Disentangling task transfer learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 3712–3722.

[Zha+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. *Understanding deep learning requires rethinking generalization*. 2016. arXiv: `1611.03530 [cs.LG]`.