Deep Learning meets Coding Theory

Dr. Sreeram Kannan
Assistant Professor, Electrical and Computer Engineering Department
University of Washington, Seattle, USA
ksreeram@uw.edu

Dr. Himanshu Asnani
Reader, School of Technology and Computer Science
Tata Institute of Fundamental Research, Mumbai, India
himanshu.asnani@tifr.res.in

Introduction

The success of modern information age hinges on reliable digital communication and the central issue thereof is the design of codes that allow transmissions to be efficiently and robustly decoded under noisy conditions. This is the discipline of coding theory whose inception can be traced back to the birth of information theory [1]. Since then, for the past 70 years or so, much effort is galvanized in the scientific community to design near optimal codes such as turbo codes [2], low-density parity check codes (LDPC) [3], polar codes [4] on AWGN Channels.

The following have remained as the long term goals in coding theory:

Goal A : to design **better decoders** for the existing codes which are **robust and adaptive** to varying channel conditions,

(Robustness implies the following: when the system is trained for AWGN channel, the test performance with no re-training on a different channel (such as ATN and Radar) should not degrade much. Adaptivity allows the system to learn a decoding algorithm on a different channel by retraining from enough data even in absence of a clean mathematical model for that channel.)

Goal B : to design **new codes** with emphasis on **robustness, adaptivity** and other features such as **low latency,**

Goal C : to design **new codes for multi-terminal settings** as well as other scenarios such as the feedback channel, the relay channel and the interference channel.

However, history has witnessed quite slow progress with respect to realizing these long term goals. This is because the two-step traditional process of communication algorithm design, which is (a) begin with clean mathematical analysis, and follow it by (b) stacked heuristics on top of the optimal algorithms, has remained largely sub-optimal insofar as the guarantees of optimality do not extend to cover the various practicalities not included in the first step. However, this two-step method still has been successful in the point-to-point setting. For the next generation communication system design which involves wireless systems such as autonomous mesh networks, industrial IoT, ultra-dense networks - that pose many challenges such as interference management, non-stationarity of channels, non-AWGN noise, interactions with other technologies - these principles of traditional design break down. This is due to (a) *algorithm deficit* - the gap between the optimal algorithms on this simplified model and the known computationally efficient algorithms on this model, and (b) *model deficit* - the gap between the realistic model and the simplified model.

Areas of computer vision [5] and natural language processing [6] amongst others, have witnessed a hockey-stick growth with the advent of deep learning, which promises learning complicated non-linear algorithms from observational data. For channel coding problems, there is unlimited training data available, so can *deep learning* aid in accelerating the rate of discovery here in realizing the goals above?

This short paper is a survey of some key results in this direction where the authors of this article were involved. In what follows, each section highlights a contribution in realizing the above goals. It finally concludes with a discussion about the exciting road ahead.

MIND : Model Independent Neural Decoder [8]

Deep learning based decoders outperforming the standard Viterbi decoder were proposed in [7] for convolutional and turbo codes. The main design principle was to train Recurrent Neural Networks (RNN) for a given AWGN channel for they closely mimic the structure of convolutional and turbo codes. These neural decoders show robustness and adaptivity properties. However, compared to the traditional decoding methods, neural decoders exhibit huge data requirements for training as well as large computational complexity to adapt to the new channel.

To add robustness and adaptivity properties to neural decoding along with the desired property of minimal training, we proposed Model Independent Neural Decoder (MIND) which builds on the top of neural decoders [7] and equips them with a fast adaptation capability to varying channels. This feature is achieved via the methodology of Model-Agnostic Meta-Learning (MAML) (for details cf. [9]). In nutshell, here the decoder: (a) first, learns a 'good' parameter initialization in the meta-training stage where the model is exposed to a set of archetypal channels and (b) updates the parameter with respect to the observed channel in the meta-testing phase using minimal adaptation data and pilot bits.

MIND admits fast adaptation with few shot adaptation data utilizing the gradient-based training. Compared to the adaptive neural decoders which require large amounts of gradient training steps and data to adapt to new channel settings, MIND can adapt to a new channel with small amount of pilot bits and few gradient descent steps (cf. Figure 1 [8]).

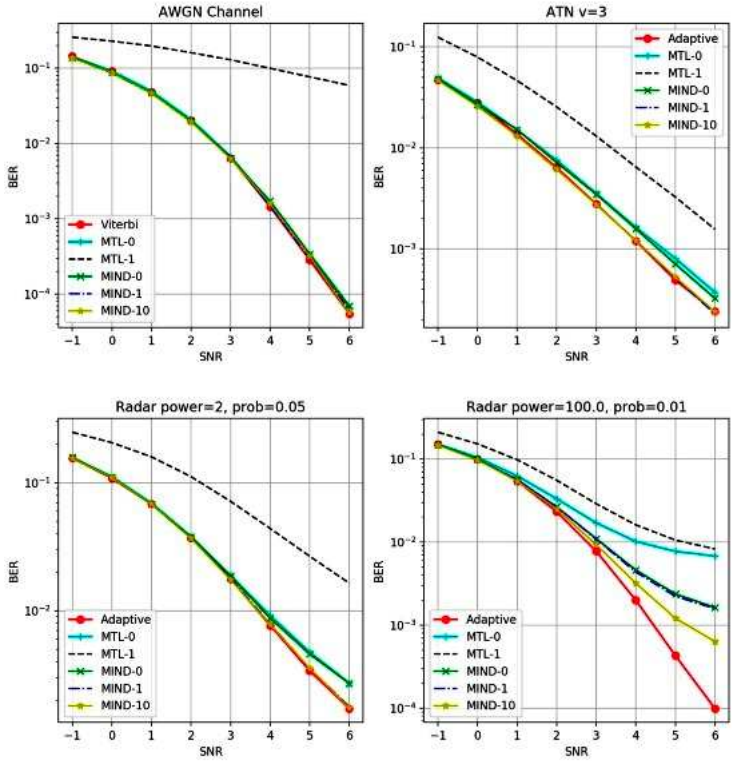


Figure 1 : Compares MIND's performance with only few gradient steps with other state of the art Neural Decoders trained on different channels. The decoders compared are (a) Canonical Viterbi decoder, (b) Adaptive Neural Decoder on non-AWGN Channel with infinite data, (c) MTL-K or multi-task learning based decoder with K-step gradient descent and finally (d) our proposed MIND-K with K gradient update steps.

LEARN : Low-latency Efficient Adaptive Robust Neural Codes [10]

Figure 2 shows the structure of a Channel Autoencoder (ChannelAE), which combines a stochastics channel with the standard Autoencoder architecture. This naturally fits into the standard communication channel and coding theory paradigm. However, so far developments have been sparse to harness this structure with deep learning training techniques to get unique and new codes for communication systems (jointly trained encoder and decoder). We handled this question in part by jointly training ChannelAE in low-latency regime (short blocklengths) to outperform the state of the art tail-biting convolutional codes in this regime [11].

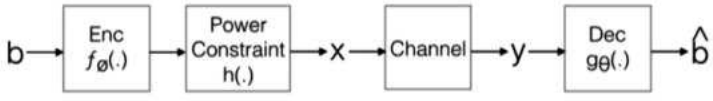
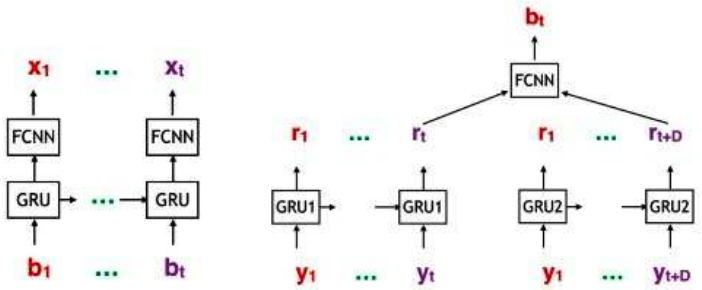


Figure 2 : Channel Autoencoder

The designed Low-latency Efficient Adaptive Robust Neural (LEARN) code applies learnable RNN structures (cf. Figure 3) for both the encoder and the decoder with an additional low-latency constraint. To the best of our knowledge, this is the first work that achieves an end-to-end design for a neural code achieving state-of-the-art performance under low latency scheme (cf. Figure 4 [10]).



Decoder Layer	Output Dimension
Input	(K, 1/R)
GRU1 (2 layers)	(K, 100)
GRU2 (2 layers)	(K, 100)
FCNN (sigmoid)	(K, 1)

Encoder Layer	Output Dimension
Input	(K, 1)
GRU (2 layers)	(K, 25)
FCNN (linear)	(K, 1/R)

Figure 3 : LEARN encoder (left), LEARN Decoder (middle) and Network features (right)

However this is possible only with an improvised training methodology as has been stated in [10]:

- Train with a large batch size.
- Use Binary Cross-entropy (BCE) loss.
- Train encoder and decoder separately. Train encoder once, and then train decoder 5 times.
- Add minimum distance regularizer on encoder.
- Use Adam optimizer.
- Add more capacity (parameters) to the decoder than the encoder.

Furthermore, when the channel conditions are varying, LEARN codes show robustness (ability to work well under unseen channel) as well as adaptivity (adapt to new channel with enough training symbols), showing an order of magnitude improvement in reliability over canonical codes (cf. Figure 5 [10]).

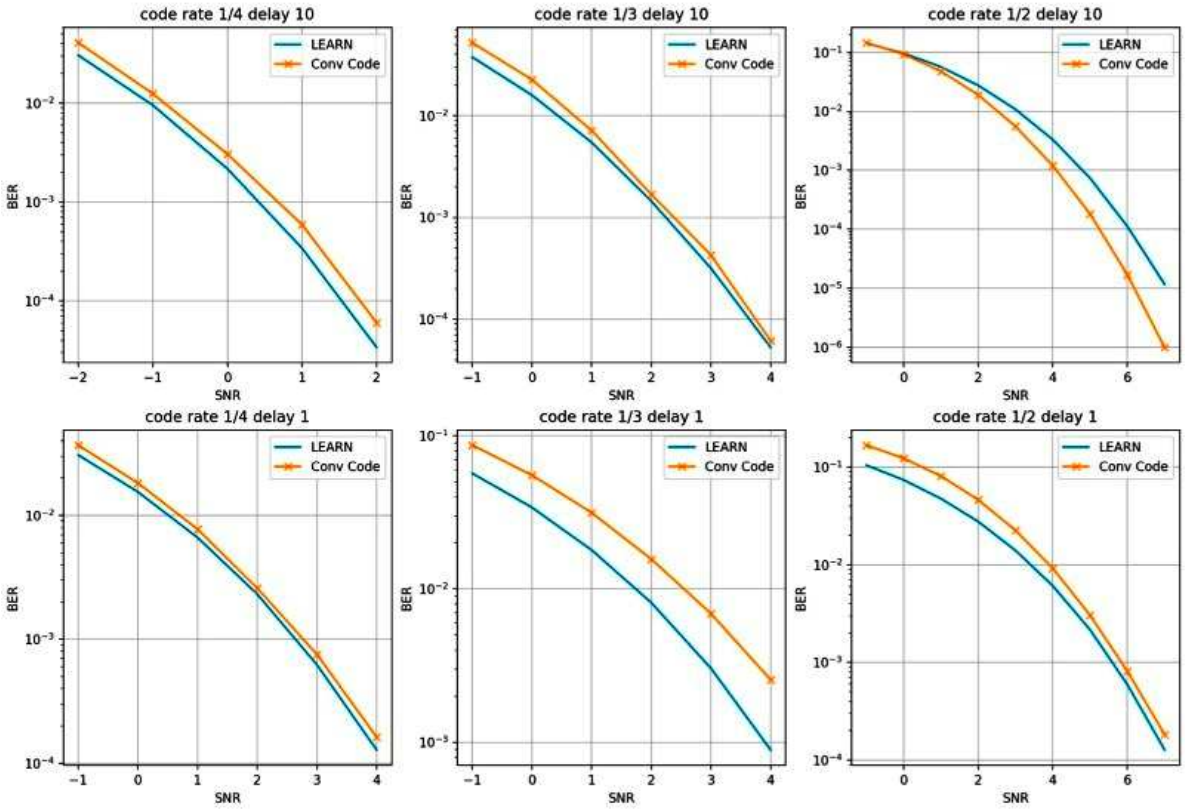


Figure 4 : Comparing BER for LEARN Code and the State of the Art Low-latency Convolutional Code under AWGN Channel for different rates.

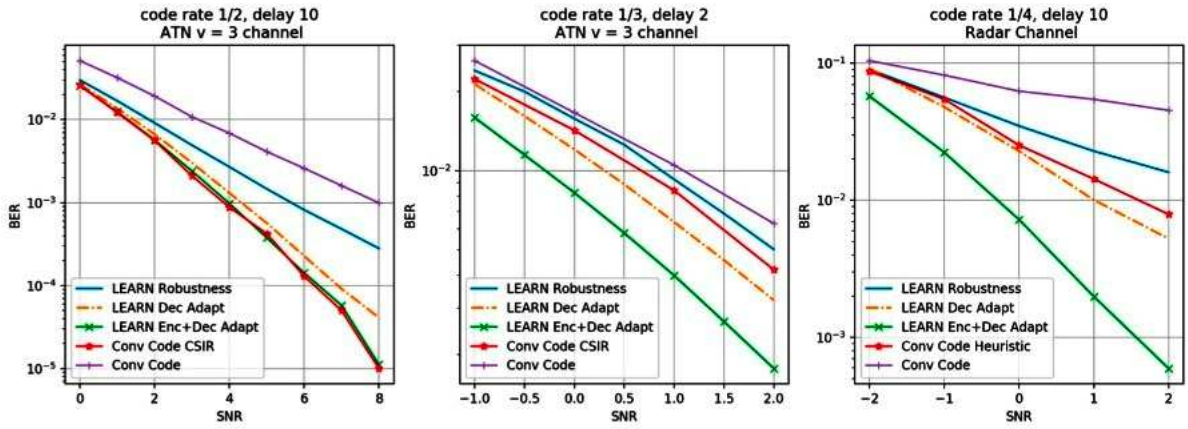


Figure 5 : LEARN exhibits robustness and adaptivity. Here LEARN Codes with either robustness or with encoder and/or decoder adaptivity are compared with convolution codes with and without CSIR (channel state information at the receiver)

TurboAE : Turbo Auto Encoder [12]

We want to relax the low-latency assumption of the previous section and investigate if we can further achieve state of the art performance in moderate block length regime. In this direction, we design TurboAE, a neural network based over-complete autoencoder parameterized as Convolutional Neural Networks (CNN) along with interleavers (permutation) and deinterleavers (de-permutation) inspired by the *turbo principle* of the turbo codes [13]. Formally, interleaver and de-interleaver shuffle and shuffle back the input sequence with the a pseudo random interleaving array known to both encoder and decoder, respectively (cf. Figure 6, 7 [12]).

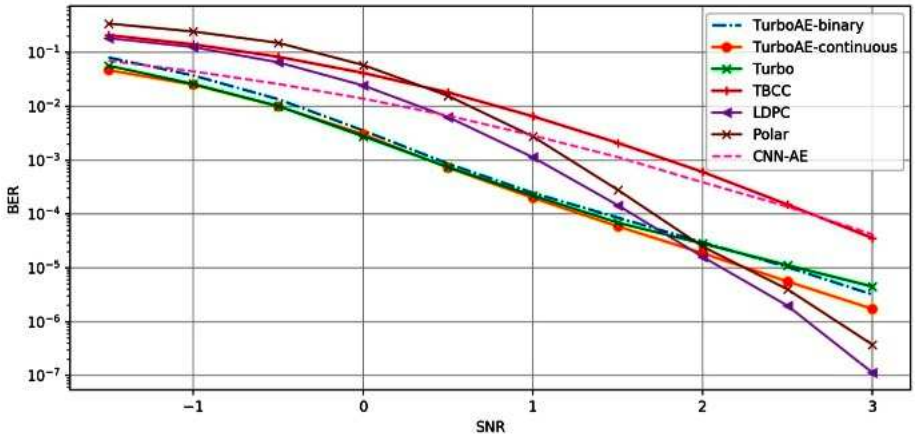


Figure 6 : Visualization of Interleaver and de-interleaver.

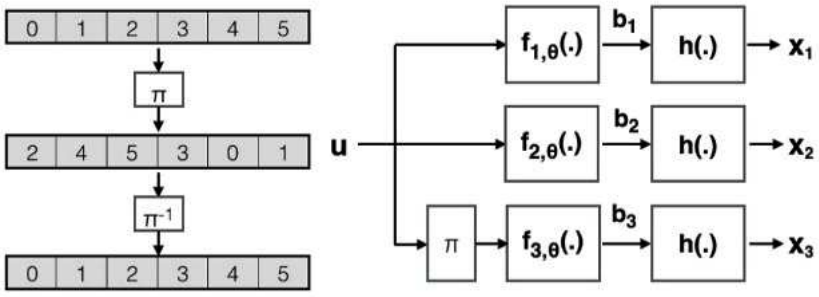


Figure 7 : TurboAE iterative decoder on code rate = 1/3

The benchmarks on block length 100 is shown in Figure 8 [12] with widely-used LDPC, Turbo, Polar, and Tail-bitting Convolutional Code (TBCC), generated via Vienna 5G simulator [14] [15], with code rate 1/3 on AWGN Channel and Figure 9 [12] shows results on non-AWGN channel.

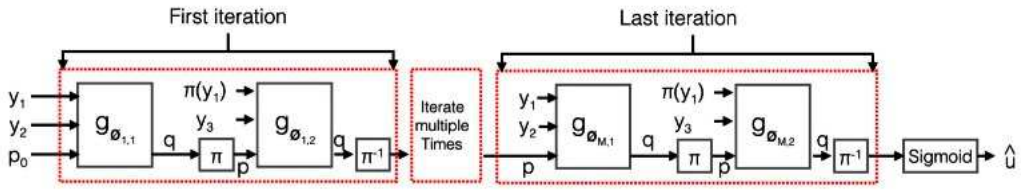


Figure 8 : Comparison of different codes with TurboAE. TurboAE-binary represents the case where encoder input is binarized as is the case with wireless communication systems.

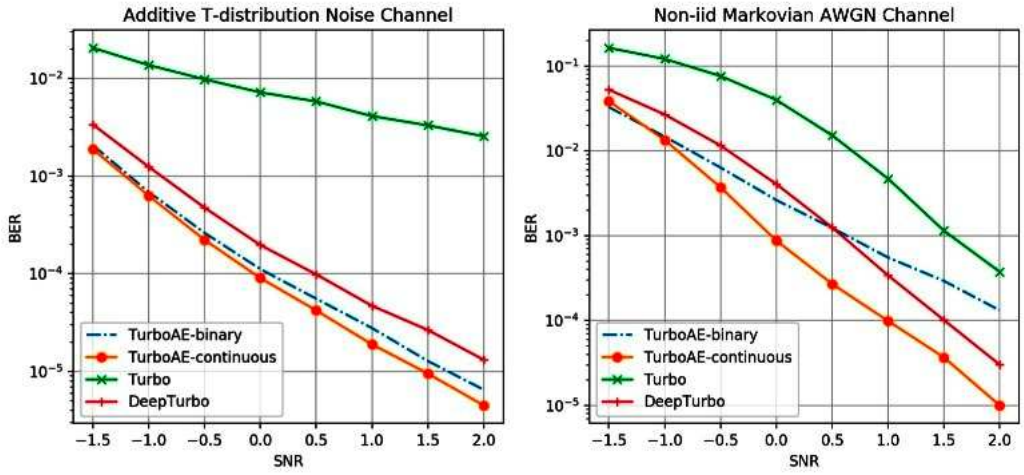


Figure 9 : TurboAE on iid ATN channel (left) and on-iid Markovian-AWGN channel (right)

Conclusion

In summary, we presented here a small buffet of the results as the current state of the art in applying *deep learning paradigm* to accelerate discovery of new codes and decoding algorithms in several scenarios of interest in wireless systems. These included adaptive decoders, new codes for low-latency, and state of the art codes for moderate block lengths. These codes also show robustness and adaptivity properties. All these bring interesting research directions to design channel coding algorithms via separate or joint encoder and decoder design.

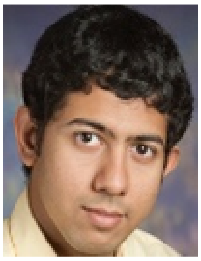
The terrain is vast open. Can we design the codes for multi-terminal settings where there is scarcity for good codes? Another interesting venue is to comment on the explainability and interpretability of these codes.

References

[1] C.E. Shannon, "A mathematical theory of communication" *Bell Syst. Tech. J.*, vol.27, pp. 623–656, 1948.
[2] C. Berrou, A.Glavieux, and P.Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," in *Proceedings of ICC'93-IEEE International Conference on Communications*, vol. 2., 1993, pp. 1064–1070.
[3] D. J. MacKay and R. M. Neal, "Near shannon limit performance of low density parity check codes," *Electronics letters*, vol. 33, no. 6, pp. 457–458, 1997.
[4] E. Arikan, "Channel polarization: A method for constructing capacity-achieving codes," in *2008 IEEE International Symposium on Information Theory*., 2008, pp. 1173–1177.
[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S.Ma et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
[6] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. "Distributed representations of words and phrases and their compositionality", *NIPS 2013*.
[7] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh and P. Visvanath, "Communication Algorithms via Deep Learning" *ICLR* 2018.
[8] Y. Jiang, H. Kim, H. Asnani, and S. Kannan, "Mind: Model independent neural decoder," in *IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019
[9] C. Finn, P. Abbeel, and S. Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." ICML 2017.
[10] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Learn codes: Inventing low-latency codes via recurrent neural networks," *IEEE ICC* 2019.
[11] C. Rachinger, J. B. Huber, and R. R. Miller. "Comparison of convolutional and block codes for low structural delay", *IEEE Transactions on Communications*, 63.12 (2015): 4629-4638.

[12] Y. Jiang, H. Kim, H. Asnani, S. Kannan, S. Oh, and P. Viswanath, "Turbo Autoencoder : Deep Learning based channel codes for point-to-point communication channels," *NeurIPS 2019.*
[13] H. R. Sadjadpour, N. J. Sloane, M. Salehi and G. Nebe, "Interleaver design for turbo codes," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 5, pp. 831–837, 2001.
[14] M. K. Muller, F. Ademaj, T. Dittrich, A. Fastenbauer, B. R. Elbal, A. Nabavi, L. Nagel, S. Schwarz, and M. Rupp, "Flexible multi-node simulation of cellular mobile communications: the Vienna 5G System Level Simulator," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, p. 17, Sep. 2018.
[15] B. Tahir, S. Schwarz, and M.Rupp,"Ber comparison between convolutional, turbo, ldpc, and polar codes," in *2017 24th International Conference on Telecommunications (ICT), IEEE*, 2017, pp. 1–7.

About the Authors



Sreeram Kannan have been an assistant professor at UW since 2014. Before that, he spent two years as a postdoctoral scholar at University of California, Berkeley and Stanford University, working with Prof. David Tse , and collaborating with Prof. Lior Pachter. He received my Ph.D. in Electrical Engineering and M.S. in mathematics from the University of Illinois, Urbana-Champaign where he was supervised by Prof. Pramod Viswanath and also worked closely with Prof. Chandra Chekuri. He received my M.E. in telecommunications from Indian Institute of Science, Bangalore under the guidance of Prof. P. Vijay Kumar. He spent my delightful undergraduate years at College of Engineering, Guindy, Anna University, where he was part of a team that developed and successfully launched ANUSAT , the first student-designed micro-satellite in India, led by Prof. P.V. Ramakrishna.

He have spent two summers at Qualcomm Corporate Research and Development, San Diego, and another wonderful summer at Microsoft Research, New England, Cambridge, MA with Prof. Madhu Sudan. He has also been a visiting researcher for several months each at Stanford University, University of Southern California, Indian Institute of Science, Bangalore and Indian Institute of Technology, Kanpur.

He is a recipient of the 2019 UW ECE Outstanding Teaching Award, 2018 Amazon Catalyst award, 2017 NSF Faculty Early CAREER award, the 2015 Washington Research Foundation Early Career Faculty award, Van Valkenburg outstanding graduate research award from UIUC, 2013, a co-recipient of the Qualcomm Cognitive Radio Contest first prize, 2010, a recipient of Qualcomm (CTO) Roberto Padovani outstanding intern award, 2010, a recipient of the S.V.C. Aiya medal from the Indian Institute of Science, 2008, and a co-recipient of Intel India Student Research Contest first prize, 2006.

Himanshu Asnani is currently Reader (eq. to tenure-track Assistant Professor) in the School of Technology and Computer Science (STCS) at the Tata Institute of Fundamental Research (TIFR), Mumbai and Affiliate Assistant Professor in the Electrical and Computer Engineering Department at University of Washington, Seattle. His research interests include information and coding theory, statistical learning and inference and machine learning. Dr. Asnani is the recipient of 2014 Marconi Society Paul Baran Young Scholar Award and was named Amazon Catalyst Fellow for the year 2018.

He received his Ph.D. in Electrical Engineering Department in 2014 from Stanford University, working under Professor Tsachy Weissman, where he was a Stanford Graduate Fellow. Following his graduate studies, he worked in Ericsson Silicon Valley as a System Architect for couple of years, focusing on designing next generation networks with emphasis on network redundancy elimination and load balancing. Driven by a deep desire to innovate and contribute in the education space, with the aid of technology, he quit his corporate sojourn and got involved for a while in his education startups (where he currently holds Founding Advisor role) to bring the promise of quality education in vernacular languages in underdeveloped and developing countries - places which do not have access to English, Internet and Electricity.

Moving on then, from industry and entrepreneurial world back to the academia, before joining TIFR, he worked as a Research Associate in Electrical and Computer Engineering Department at University of Washington, Seattle. In the past, he has also held visiting faculty appointments in the Electrical Engineering Department at Stanford University and Electrical Engineering Department at IIT Bombay. He was the recipient of Best Paper Award at MobiHoc 2009 and was also the finalist for Student Paper Award in ISIT 2011, Saint Petersburg, Russia. Prior to that, he received his B.Tech. from IIT Bombay in 2009 and M.S. from Stanford University in 2011, both in Electrical Engineering.

Rolls-Royce built a new tool called Quips, which uses AI and 'voice banking' to learn the unique way its user talks, essentially helping people with Lou Gehrig's disease or ALS (Amyotrophic Lateral Sclerosis). Rolls-Royce and its R² Data Labs created Quips with help from Motor Neurone Disease Association and companies including Intel and Microsoft. However, it's still early in developmen