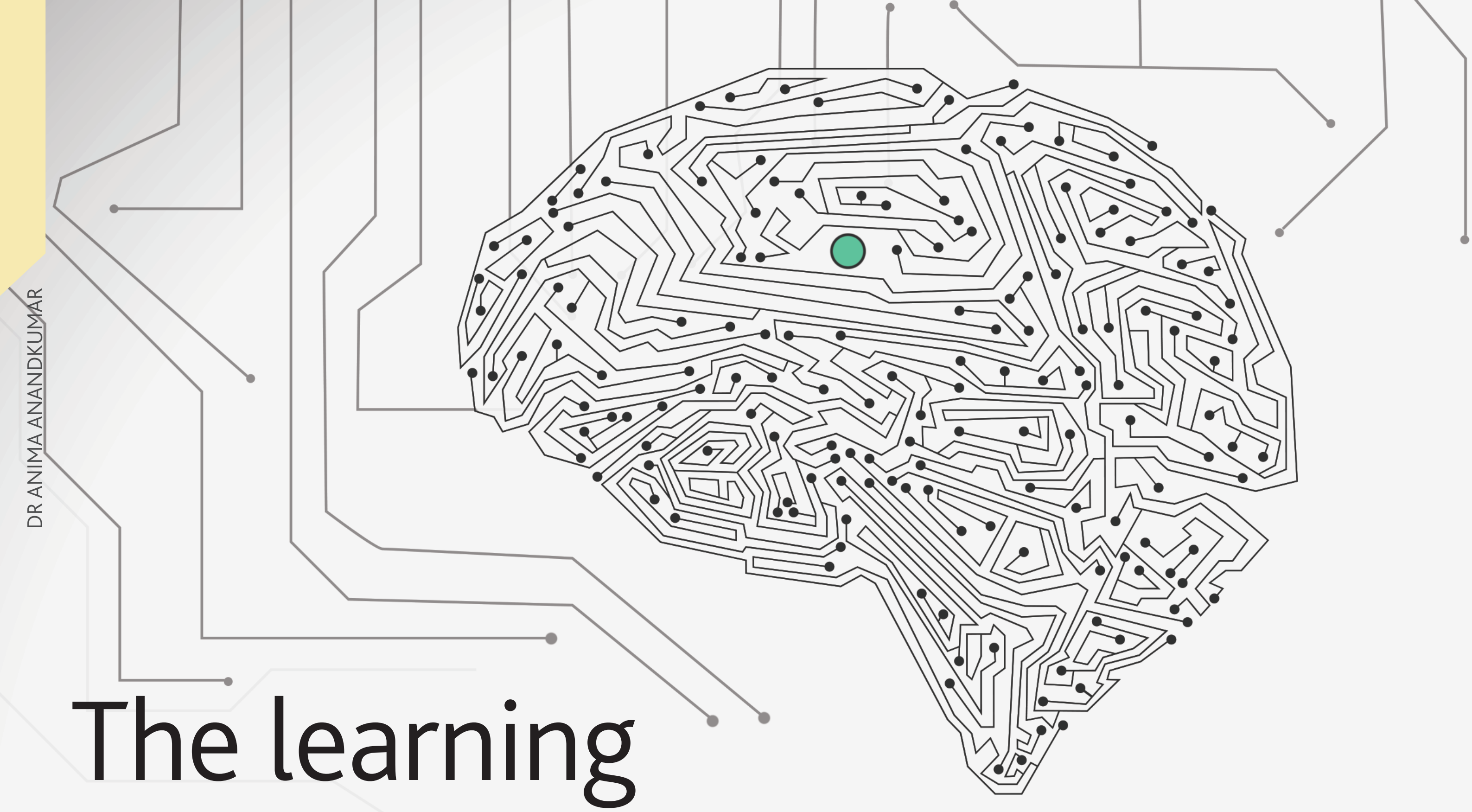


The learning machine

Computer scientist and machine learning expert **Dr Anima Anandkumar** reveals the technical details of her efforts to further advanced analytical processes for high-dimensional data, and her relationship with collaborators including Microsoft Research

What initiated your interest in electrical engineering and computer science?

Ever since I can remember, I have been curious about the structure of the world around us and the mechanisms that drive it. I found mathematics to be the right language to express and explain these mechanisms. I was drawn to electrical engineering and computer science since it gave me the right toolbox for analysis and allowed me to harness the power of mathematics, putting principles into action.

Can you discuss some of the greatest challenges, as well as the successes, of your research to date?

I take a broad view when looking for research problems to tackle, and my toolbox is not limited to one area. This has been a great challenge since it requires continuously learning and mastering new techniques – which is also a big part of what I find fun about the work. My research is interdisciplinary and spans theoretical and applied machine learning, with applications in social networks, computational biology, and text and image analysis.

On the theoretical side, I have been able to answer a number of questions such as: what kind of phenomena can we discover from data? How difficult is it to learn them? What are the principled algorithms that can be guaranteed to learn the phenomena accurately? These questions are not only intellectually stimulating, but relevant in overcoming many real-world problems in a number of domains.

How important is collaboration to your work?

I have a team of five highly motivated graduate students, including three female graduate students (I have participated in many outreach activities for women and underrepresented communities, and I care deeply about this issue). In addition, I have a number of collaborations both in academia and industry, which are central to broadening the scope of my research. My industry collaborations have been valuable for exploring new ideas and for deploying algorithms on a large scale. I have been privileged to receive the support of Microsoft Research: I was awarded a highly competitive faculty fellowship award in 2013, have visited the labs numerous times and have strong ties with the researchers there. In addition, I have been collaborating with biologists and neuroscientists to transfer the developed learning algorithms to those domains. I have also collaborated with a sociologist to improve my understanding and reasoning about social networks.

How are you tracking the evolution of dynamic social networks? What impact has your work had here so far?

We are tracking networks by learning the hidden processes that drive their evolution. For instance, learning relationships between the nodes can help us track the network better: close friends or relatives are more likely to participate in a conversation than others. We learn all these hidden relationships in a global fashion and also their effects on network evolution using efficient algorithms which can be scaled to large networks.

What other exciting work is taking place within the Model Estimation, Graphical Algorithms, Data Analysis (MEGA DatA) Lab?

There are a number of exciting projects. One is an interdisciplinary project to harness our machine learning algorithms to learn about the genetic composition of brain cells. Another is making our algorithms ever more scalable and deploying them on the latest cloud-based platform with hundreds of computers. Robustness is another important criterion since the data at hand have gross corruptions and are far from modelling assumptions. We are working on fast algorithms for robust principal component analysis and their extensions. Overall, there is a huge flurry of activity within the lab and with my outside collaborators.

The hunt for **hidden** information

A team of researchers based at the **University of California**, **Irvine**, has in recent years made dramatic contributions to tensor algorithms in machine learning and probabilistic modelling; their work allows for the discovery of hidden information structures in high-dimensional datasets

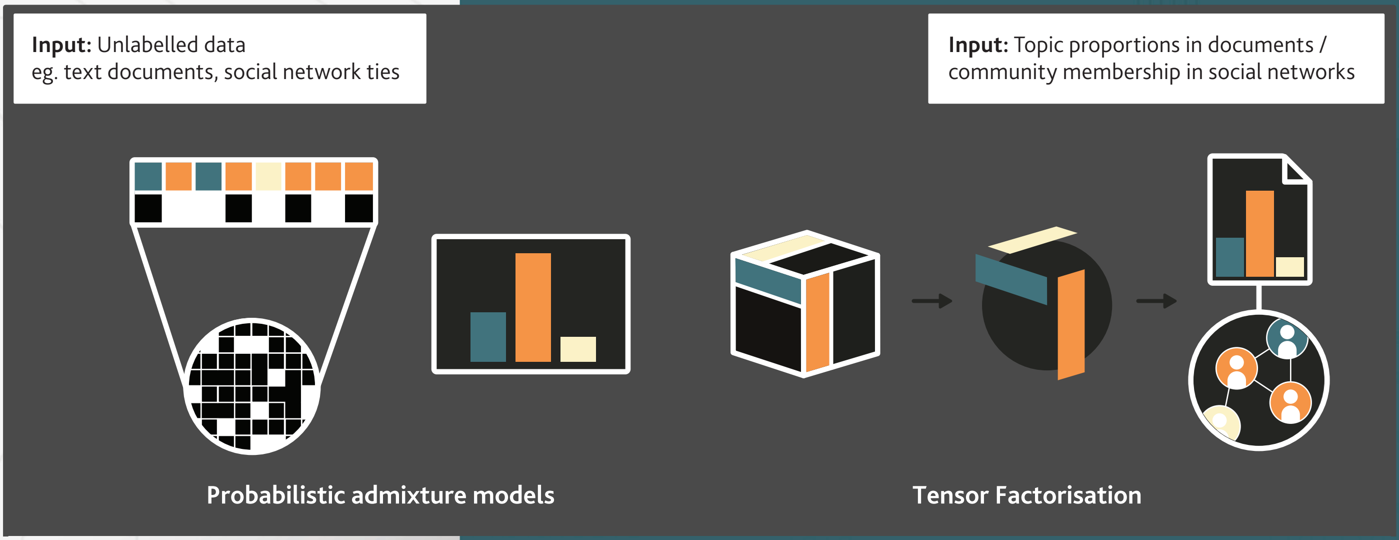
SCIENCE, AND THROUGH it human knowledge, is ruled by data – but unfortunately, this does not mean that the two are synonymous. In order to bring forth salient and useful information, raw data must be processed and interpreted, and this takes time and effort. Today, scientists and experts in almost every field face the problem of data deluge; since the late 21st Century, the volume of raw data available has grown exponentially – in biology, for example, high-throughput sequencing technologies can produce data 400 times faster than was possible just 10 years ago. The task of making sense of all these data calls for ever more powerful computational tools.

Coupled with the problem of oversupply is the issue that much of the data collected in many domains, though abundant, is of a fairly poor quality; it is noisy, subsampled and contains a large number of variable or 'unknown' dimensions in comparison with the observed or 'known' ones. In other words, much of the important information contained within these data is effectively hidden. To return to the example of genetics, this phenomenon can be observed in the fact that while the expression levels of different genes can be easily and quickly discerned using the right tools, the functional groups that these genes belong to are often unknown.

AIMING FOR SCALABILITY

The important question of how it is possible to efficiently model and learn from such high-dimensional data is an important one in many fields, and particularly in machine learning. This interdisciplinary area spanning computer science, statistics and data mining is concerned with the design and deployment of algorithms through which computers can draw conclusions or 'learn' from data. In the context of high-dimensional data, it is important to consider whether such algorithms can be scalable in terms of their computational requirements; what is needed here is not just an efficient way of dealing with big data, but one that will attain the maximal utility by avoiding unnecessary work, cutting process time while extracting all the useful information.

Dr Anima Anandkumar leads a dedicated team of computer scientists at the University of California, Irvine, USA, in the pursuit of this goal. According to their research, the key to making these high-dimensional models scalable is to impose


Input: Unlabelled data
eg. text documents, social network ties
Input: Topic proportions in documents / community membership in social networks
Probabilistic admixture models
Tensor Factorisation

Overview of tensor methods for unsupervised learning.

A NEW PARADIGM FOR LEARNING HIDDEN STRUCTURES IN MASSIVE-SCALE DATA

OBJECTIVES

- To find hidden information structures in large-scale unlabelled data
- To develop algorithms with theoretical guarantees as well as good practical performance
- To gain insight on how manipulating tensor data structures can reveal hidden variable information in probabilistic models

KEY COLLABORATORS

Jennifer Chayes; **Christian Borgs**; **Sham Kakade**; **Rong Ge**; **Prateek Jain**; **Alekh Agarwal**; **Praneeth Netrapalli**; **Paul Mineiro**; **Nikos Karampatziakis**; **Sergiy Matusevych**, Microsoft Research, USA • **Daniel Hsu**, Columbia University, USA • **Srini Turaga**, University College London, UK • **Ernest Fraenkal**, MIT, USA • **Anthony Gitter**, University of Wisconsin Madison, USA • **Alexander Smola**, Carnegie Mellon University, USA

FUNDING

Microsoft Research • Army Research Organisation, Office of Naval Research • National Science Foundation (NSF) • Alfred P Sloan foundation

CONTACT

Dr Anima Anandkumar
Assistant Professor

Microsoft Faculty Fellow and Sloan Fellow
University of California, Irvine
Department of Electrical Engineering
& Computer Science
Irvine
California 92697-3425
USA

T +1 949 824 9072
E a.anandkumar@uci.edu

http://newport.eecs.uci.edu/anandkumar

SOCIAL MEDIA

@AnimaAnandkumar

https://bit.ly/animaanandkumar

https://bit.ly/animalinkedin

DR ANIMA ANANDKUMAR is a renowned expert in machine learning and high dimensional statistics. Her pioneering research on tensor algorithms has significantly advanced the theory of large-scale unsupervised learning. In addition, Anandkumar has made contributions to many areas such as probabilistic graphical models, information theory and signal processing. She is the recipient of highly competitive awards such as Alfred P Sloan Fellowship 2014, Microsoft Faculty Fellowship 2013, Army Research Office Young Investigator Award and the NSF CAREER Award.

additional constraints on the solutions required – either by representing structural relationships between the variables in a graph form, or by modelling the hidden variables that lurk behind the observed data. "My research shows that models with such constraints can be learnt efficiently on a large-scale," Anandkumar asserts – and indeed, the solutions proposed by her lab have met with great success in a startling variety of fields.

UNSUPERVISED

Machine learning is perhaps best known for its supervised learning techniques, whereby a computer can be 'taught' how to rank, categorise or cluster objects based on iterative training with labelled examples, and input indicating success or failure. Anandkumar, however, is more interested in unsupervised learning, which focuses on revealing the structures of unlabelled datasets, and is therefore well-adapted to the task of finding hidden or latent variables. This kind of approach is useful for a number of reasons, but perhaps one of the most important is that unlabelled datasets are far more common and cheaper to acquire than labelled ones, since labelling the data requires costly human intervention.

In some applications, there are other problems with acquiring labelled data. In the analysis of social networks, for example, which is one of the applications that Anandkumar is particularly interested in, acquiring labelled data would be hard due to privacy issues. Is it really possible to draw meaningful conclusions about distinct communities within a social network purely from the information of who is connected to who? The work performed by the California team shows that it is – and such an insight would not only reveal those connected by family ties and common hobbies, but could also reveal more shady hidden communities, including gangs and terrorist groups.

TENSOR TRICKS

"Finding hidden structures in data is literally a 'needle in a haystack' problem, since we have a large amount of data and the object of interest is typically low dimensional," Anandkumar admits – but the powerful methods she has developed along with her collaborators and students allow them to overcome this challenge. Their novel algorithms are based on a mixture of two technical elements, the first of these being tensor factorisation. A tensor can be thought of as a higher dimensional extension of a matrix and can represent rich data structures; in the social networking context, the entries in the tensor records the number of common friends among different groups of users. In text analysis, the tensor consists of co-occurrence counts of different tuples of words in the same document. Factorising these tensors involves rendering them in their most simple and compact forms, where the information they contain will be most accessible.

The second element, probabilistic modelling, engages with the issue of how the hidden variables are related to the observed data. For example, in the case of text analysis, the machine is able to record or observe individual words, but the topic of the document is hidden. Collocations of certain words can be recorded in a tensor – for example, the document contains the words 'orange', 'apple' and 'pear', the probabilistic model can determine from this information that the topic is likely to be fruit. The hidden element, however, could be any aspect of the objects concerned that is not visible to a machine; the features in an image, for example, or the human voices in a recording.

ENDLESS APPLICATIONS

The beauty of these tensor methods is that they are scalable to incredibly large datasets with billions of variables without escaping the bounds of feasibility with current computer resources. These processes are so flexible, in fact, that they could even be scaled to cope with the total dataset provided by massive online social networking sites. What is more, the theoretical work conducted by Anandkumar's team has guaranteed that the algorithms employed here are capable of producing highly accurate estimates and learning complex models. "Having an algorithm with both theoretical guarantees as well as strong empirical performance is a rare combination," she enthuses.

The California scientists are also involved in multidisciplinary collaborations to apply their machine learning techniques to data across all the examples mentioned, and more. In computational biology, neuroscience and social network analysis the methods they have developed are having a big impact – while at the same time having a great relevance to the continued pursuit of fully automated speech and image recognition. Despite the prolific activity of this research group, however, there are still a diversity of fields that can stand to benefit from an enhanced understanding of hidden information.

This fascinating work satisfies numerous practical needs while simultaneously giving an outlet for significant intellectual curiosity. Undoubtedly, Anandkumar and her collaborators will not be satisfied to rest on their laurels, but will continue to lead practice in the field in the near future.

Eigenvectors of a tensor can be thought of as locally high energy directions and can thus provide rich insights about hidden information.