Patterns



People of Data

Physical scientist or a data scientist? A journey between disciplines

Zhiping Xu^{1,*}

¹Applied Mechanics Laboratory, Department of Engineering Mechanics, Center for Nano and Micro Mechanics, Tsinghua University, Beijing 100084, China

*Correspondence: xuzp@tsinghua.edu.cn https://doi.org/10.1016/j.patter.2022.100527

In a recent publication in *Patterns*, Zhiping Xu, a group leader at Tsinghua University, China, and his team developed a machine learning approach to unveil the conformational complexities of 2D macromolecules by including the insights of physics and topology. Xu talks about the role of data science in his research and the ongoing process of turning into a data scientist.

What would you like to share about your background (personal and/or professional)?

I was educated in the department of engineering mechanics, where the core courses include mechanics and the relevant mathematics. The focus was placed on the continuum framework, where solids and fluids can be modeled at high precision in a wide spectrum of engineering applications. Partial differential equations and their numerical methods are the main theoretical tools used to understand the problems of deformation and failure, as well as mass and energy transfer. Two of the well-known challenges in classical mechanics are the failure of materials¹ and turbulence in fluid dynamics, which share the structural complexity across multiple length scales, and the temporal evolution of the material microstructures or the flow. The trade-off between the spatiotemporal resolution and the scale of data makes the full solution to these problems intractable.

I started to look into the microstructural hierarchy of matter and the transfer of load, mass, and energy across the scales since I set up my research group in Tsinghua. Achievements have been made for fibers and films with regular microstructures or those with random features in an effective-model manner.2 However, the microstructural complexity not only remains challenging to be fully characterized in experiments but also lacks legitimate mathematical representations. The physics rules governing the dynamical processes of matter under load or environmental cues entwined with the microstructural complexity, bringing in additional difficulties in rational understanding and prediction.

Data analysis has played an important role in elucidating the underlying physics from, for example, the electron density, trajectories of atoms, and microstructural changes. Well-defined measures and statistical analysis are used and succeeded in many of the cases. However, the data generated from multiscale simulations or experiments are of huge volume and exhibit inherent heterogeneity. There is thus much more to discover by following the updated knowledge in data science, and the utilization of data-driven models and machine learning tools holds great promise in offering insights that cannot be reached along the conventional paths.

What drew you to this area of research? How has the research focus of your team evolved over the years?

My personal interest is focused on the complexities of matter. Specifically, the research of my team includes understanding the mechanical performance of materials that is defined by the microstructures and their evolution.

Starting from the microscale, experimental characterization and computer simulations produce a large volume of data that needs to be analyzed. The conventional solution is to define order parameters or collective variables and to rationalize the problem through their governing or evolutionary equations. Examples include the inelastic processes such as plasticity and viscoelasticity or the growth and degradation (fatigue, wear, corrosion, etc.) of materials.

With the rapid advances in structural and functional materials and the ever-

increasing demands from the industry, more powerful methods and tools have to be developed to address the spatial and temporal complexities. We turn to data science for inspiration only very recently, and the work published in this issue of *Patterns*³ marks our first effort.

Why did you decide to publish in *Patterns*?

First of all, the journal name "Patterns" reflects much of the nature of our work. The aim is to reveal the morphological complexity of two-dimensional (2D) macromolecules from their essential patterns beyond the geometry that is easy to recognize by visual inspection but shown to be not sufficient.

Physics of lattice distortion and the topology of contact through non-bonding interactions were included in the machine learning algorithms in our work to improve the performance of morphological classification and recognition. There are thus a few sets of patterns such as the curvatures, the changes in metrics, the energy map of bonding and non-bonding interactions, and the associated distance map. The ideas of integrating physical and data sciences in improving our understanding of complex patterns of matter thus becomes very exciting.

We have been inspired by several works published in *Patterns* that showcase the formulation, implementation, and application of data science output for problems in specific domains. I have been publishing with the Cell Press in journals such as *Matter* and *Cell Reports Physical Science*, and the experience was great.







We thus think Patterns is the best place to present our results.

What motivated you to become a (data) researcher? Is there anyone/ anything that helped guide you on your path?

Turing into a data researcher is a gradual process. Starting research using atomistic simulations to model the nanomechanical devices, I spent much time in my graduate study extracting distinct features in the spatial and temporal patterns of molecular dynamics. Since rigid-body translation and rotation of the moving parts are involved, the conventional recipe of analysis that projects the atomic displacement onto the normal modes or phonons at the ground state fail. Later on, I figured out to use the principal-component analysis in processing the atomic trajectories. The energy flow between different principal modes of motion elucidates the mechanisms of mechanical excitation and energy dissipation.^{4,5} In this analysis, the rigidbody motion can be distinguished from the molecular or lattice vibrations, and we identified the catastrophic events of dissipation that correspond to their resonant coupling. By defining and monitoring the statistical indicators, we devised the early warning signals to predict the occurrence of these critical events and prevent them in advance.^{4,5} The evolution of this work and the insights I received from the statistical analysis of the time-series space-time data motivated me to use data-driven models and methods to understand the physics of complex systems.

What is the definition of data science in your opinion? What is a data scientist? Do you self-identify as one?

Data science for physical science is the domain where scientific data and statistical learning algorithms are used to extract the inherent spatial and temporal patterns and the governing laws of physics, which otherwise cannot be uncovered. The knowledge and actionable insights mined from the data are then used to build predictive models and systems that are valuable in specific applications.

These are a few points to be made on the challenges. The scientific data is usually of large volume, noisy, and heterogeneous. The underlying physics could be nonlinear and couple multiple scales and fields (mechanical, thermal, chemical, electromagnetic, etc.). The processing of data could be very time and resource costly, while the applications usually need high fidelity and prompt responses.

Data scientists work in this interdisciplinary field to address these issues, creating high-performance scientific methods, processes, algorithms, and systems across a broad range of application domains.

I am at a point of transition into this field, to bridge my knowledge of the complexity of matter with the ever-growing data reported from experimental and computational procedures, as well as the modern tools and techniques developed in data science.

What is the fun part of being a data scientist?

The fun part of being a data scientist for me is that we can understand material behaviors from a different and complementary viewpoint. The concepts nucleated from data science and the tools developed enlighten new ideas. The involvement in the cross-disciplinary community is also enjoyable, where not only data science but also domain-specific knowledge is openly shared.

What barriers have you faced in pursuing data science as a career? How do you keep up to date with advances in both data science techniques and in your field/domain?

The barrier we face is the cross-disciplinary nature of the field. With a background in physical sciences, we need to get equipped with the theories and tools developed in statistics and information science in the past, in order to find the best way to solve the problem at hand or seek new ideas and make decisions to exploit new algorithms.

We keep up to date with advances both in data science techniques and in my domain by reading recent research articles and the classics they refer to. We also track the features included in the open-source machine learning frameworks such as TensorFlow and PyTorch. The discussions in the open-source community are especially helpful. For domain-specific applications, we also actively communicate with the engineering departments and our industrial partners for inspiration to apply data science to real-world problems.

Which of the current trends in data science seem most interesting to you? In your opinion, what are the most pressing questions for the data science community?

One of the most attractive problems in materials sciences is to rationalize the processing-microstructure-properties relationship. With the progress achieved in multiscale modeling and characterization in recent years, enormous efforts have been made to unveil the underlying correlation.

For example, electrons, photons, ions, and neutrons are used as probes to resolve the microstructural complexities down to the atomic scale. Computed tomography can then be used to reconstruct the spatial and temporal evolution of the microstructures of engineering materials in situ. The interpretation of these data is mainly based on domain-specific knowledge, which is, unfortunately, largely limited by the theoretical insights and mathematical tools. However, there is plenty of room to promote the level of investigation following a datadriven approach.

One of the promising applications in the data-driven paradigm for engineering sciences is to develop high-fidelity digital tools for real-time diagnosis and prognosis. The heterogeneous nature of realworld problems and the rare events that can play critical roles in the life cycle of materials or structures pose great challenges to data science.

These issues could be an example of the ubiquitous requests from a wide variety of application domains. Integrating intelligence from different fields of science and engineering reflects the current trend of data science and providing inspiring data-driven solutions would be one of the most pressing questions for the community.

What is the role of data science in your domain/field? What advancements do you expect in data science in this field over the next 2-3 years? How can data science help your domain (e.g., science, manufacturing, industry, policy, society), and how can collaborations between academia and your domain be started?

In materials sciences, the use of data was to derive empirical rules and inspire

Patterns People of Data



mechanism-based theories for the understanding and prediction of materials behaviors.

Recently, material databases were constructed by automated, high-throughput experiments and computer simulations, such as the materials project, NOMAD repository and archive. Stateof-the-art text-mining techniques and natural language processing tools also allow one to extract the domain knowledge from the literature. These data allow people to develop high-fidelity datadriven models and methods, and screen materials for real-world problems, which boost the research and development of material-specific applications.

I would foresee the development of data science in the next 2-3 years toward in-depth understanding of the spatial and temporal complexities embedded in the processing-microstructures-properties relationship, by tackling the challenges rising from the heterogeneity, and sometimes nonlinearity, of the data.

The insights from data science also offer guidance to set the standards of reporting research data that can help the knowledge and information exchange in the science and engineering communities. Moreover, tracking the life cycles of materials and their flow in the market and industry could also help to use the materials in a more efficient and greener way, where data science is expected to make a key contribution.

How did this project you wrote about come to be? Was there a particular result that surprised you, or did you have a eureka moment? How did you react?

This project comes from our general interest in identifying the condensed phases of 2D macromolecules that play important roles in defining the microstructures of their macroscopic assemblies such as fibers and forms.^{2,6}

From optical and electron microscopy images, the condensed 2D macromolecules can only be visualized from their out surfaces, leaving the internal structures largely unknown. In our previous studies, we utilized molecular simulations to model the conformational behaviors of 2D macromolecules. One of the significances of molecular-level modeling is that the very nature of the molecules can be resolved at the atomic scale, including their free energy density and the distance map that measures the contrast between intramolecular and intermolecular interactions.

We found, in surprise at the beginning. that the information of lattice distortion and non-bonding contact of the 2D macromolecules could boost the performance of morphological classification and identification. Later on, we realized that the "visual" approach considers the molecules as a 3D point set, which cannot capture the heterogeneity embedded in the deformation of geometry and the topology.

This finding, although exciting, is not the end of the story. We expect the combination of morphology and free energy to provide a route to the full free energy landscapes and a picture of the phase transitions. We mentioned it briefly in the current work,3 but further discussion on this topic will need a richer set of data that covers the channels of transitions.

Moreover, we believe that the models and methods reported here can be extended to study the crumpling of biological cells and cortical folding of human brains, which may offer some new insights into the physiological processes. We are just on the way to a more stimulating journey.

REFERENCES

1. Zhang, Z., and Xu, Z. (2022). Failure life prediction for carbon nanotubes. J. Mech. Phys. Solid.

164, 104907. https://doi.org/10.1016/j.jmps. 2022.104907.

- 2. Wang, S., Lin, J., Xu, Z., and Xu, Z. (2021). Understanding macroscopic assemblies of carbon nanostructures with microstructural complexity. Compos. Appl. Sci. Manuf. 143, 106318. https://doi.org/10.1016/j.composi tesa.2021.106318.
- 3. Zhao, Y., Qin, J., Wang, S., and Xu, Z. (2022). Unraveling the morphological complexity of two-dimensional macromolecules. Patterns, 100497. https://doi.org/10.1016/j.patter.2022.
- 4. Xu, Z., Zheng, Q., Jiang, Q., Ma, C.-C., Zhao, Y., Chen, G., Gao, H., and Ren, G. (2008). Transphonon effects in ultra-fast nanodevices. Nanotechnology 19, 255705. https://doi.org/ 10.1088/0957-4484/19/25/255705.
- 5. Wang, A., He, Q., and Xu, Z. (2016). Predicting the lifetime of superlubricity. Europhys. Lett. 112, 60007. https://doi.org/10.1209/0295-5075/112/60007
- 6. Wang, Y., Wang, S., Li, P., Rajendran, S., Xu, Z., Liu, S., et al. (2020). Conformational phase map of two-dimensional macromolecular graphene oxide in solution. Matter 3, 230-245. https:// doi.org/10.1016/j.matt.2020.04.023.

About the author



Dr. Zhining Xu

Dr. Zhiping Xu received his BS (2002) and PhD (2007) degrees from Tsinghua University in the Department of Engineering Mechanics. He spent 3 years working in the United States as a postdoctoral researcher at Rice University (2007-2008) and MIT (2008-2010). He returned to the Tsinghua campus in Beijing in 2010 and was promoted to full professor of engineering mechanics in 2015. His research interests include the physics of synthetic materials and living systems in general, on which he performs theoretical and experimental work with his graduate students and collaborators from a wide spectrum of fields.