Smoky Mountain Data Challenge 2020: An Open Call to Solve Data Problems in the Areas of Neutron Science, Material Science, Urban Modeling and Dynamics, Geophysics, and Biomedical Informatics

Suzanne Parete-Koon[1], Peter F. Peterson[1] Garrett E. Granroth[1], Wenduo Zhou[1], Pravallika Devineni[1], Nouamane Laanait[1], Junqi Yin[1], Albina Borisevich[1], Ketan Maheshwari[1], Melissa Allen-Dumas[1], Srinath Ravulaparthy[1], Kuldeep Kurte[1], Jibo Sanyal[1], Anne Berres[1], Olivera Kotevska[1], Folami Alamudun[1], Keith Gray[2], Max Grossman[2], Anar Yusifov[2], Ioana Danciu[1], Gil Alterovitz[3], and Dasha Herrmannova[1]

[1] Oak Ridge National Laboratory,Oak Ridge, TN 37831, USA,
paretekoonst@ornl.gov,
WWW home page: https://smc-datachallenge.ornl.gov/
[2] BP plc
[3] US Department of Veterans Affairs, Presidential Innovation Fellows Program

Abstract. The 2020 Smoky Mountains Computational Sciences and Engineering Conference enlists research scientists from across Oak Ridge National Laboratory (ORNL) to be data sponsors and help create data analytics challenges for eminent data sets at the laboratory. This work describes the significance of each of the seven data sets and their associated challenge questions. The challenge questions for each data set were required to cover multiple difficulty levels. An international call for participation was sent to students, and researchers asking them to form teams of up to four people to apply novel data analytics techniques to these data sets.

Keywords: Data Analytics, Artificial Intelligence, Machine Learning

1 Introduction

All the data analytics challenges we host represent real world problems in different areas of research. The 2020 challenge solutions could impact unsolved questions in materials science, research on energy conservation in cities, geological studies based on seismic data, research toward matching critically ill medical patients and their physicians with the most helpful clinical trials of novel therapeutics and research toward halting the spread of COVID19.

By requiring the challenge questions for each data set to cover multiple difficulty levels and by allowing students and experts to compete in separate categories we hope to draw in a diverse set of researchers and perspectives to help solve these questions.

The call for participation was broadly advertised and open to all interested parties. It appeared in scientific and engineering newsletters such as HPC Wire, and was spread by social media. Invitations to participate were also sent to several university computer science department professors and users of Oak Ridge Leadership Computing facility.

In addition to providing and serving the datasets for the challenges, organizers and data sponsors held an interactive webinar to explain the relevance of each challenge task and describe the size and composition of its associated dataset. Subsequently, three online Reddit.com forums were held in the two months before solutions were due, so participants could post questions about the tasks and get answers from eachother and the data sponsors. Lastly, to accommodate the student challenge competitors who may not have ever written a scientific paper, the challenge organizers held a best practices in scientific paper writing webinar the week before the solution papers were due.

In this work, each of the challenges has its own section wherein the authors of the challenge describe the motivation and science behind the challenge, the data and its origins, and the reasoning behind the individual challenge questions.

2 Challenge 1: Understanding Rapid Cycling Temperature Logs from the Vulcan Diffractometer

Neutron scattering allows scientists to count scattered neutrons, measure their energies and the angles at which they scatter, and map their final positions. This information can reveal the molecular and magnetic structure and behavior of materials, such as high temperature superconductors, polymers, metals, and biological samples. The Spallation Neutron Source (SNS) facility at the Oak Ridge National Laboratory provides the most intense pulsed neutron beams in the world for scientific research and industrial development.

2.1 Background

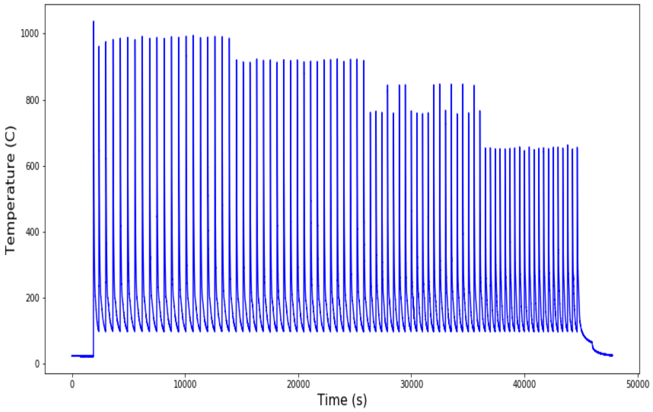
The VULCAN diffractometer [1] is designed to understand the fundamental aspects of material behaviors during synthesis, processing, and service. One of the experiments conducted at SNS is designed to generate high intensity neutron pulses for the study of materials, where, over the course of the measurement, the temperature is varied as a function of time [2]. The overall purpose of the neutron measurement is to understand the changes in structure of the material as a function of temperature.

The experiment is conducted as follows: the sample is rapidly heated, then the heat source is turned off allowing the sample to relax and reach an equilibrium. Then the sample is rapidly heated again and the experiment is repeated several times. The goal is to associate neutron events occurring within a certain temperature bin. An event is defined as follows: the event starts when the sample is subjected to rapid heating and it ends right before the next rapid heating occurs. This is called a heat cycle.

2.2 Dataset

The VULCAN Beamline dataset [4] provides the sample measurement, where temperatures are recorded in two physically different places on the sample. These are held in two different hdf5 groups in the data file. Figure 1 depicts Temperature (in Celsius) vs. time (in seconds), a sample measurement on the VULCAN [3] beamline. The total size of the data is 1.06 MB and a and a laptop, workstation or small cluster should be suitable for developing solutions to this challenge.

Fig. 1. Temperature vs. time on the VULCAN Beamline



2.3 Challenges of Interest

These questions seek insight into the behavior of the sample as the temperature is varied.

1. The goal is to identify heat cycles pertaining to equivalent temperatures during the same heating or cooling phase. For example, two data points at 800C are in the same group only if both of them are in either the heating or cooling cycle. To that end, you need to identify the beginning and end times of the heat cycle and temperature group it needs to belong to. (The latter part could be thought of as a clustering problem.)
2. How many events are there in each group and how similar are the identified events? A sample input to the question could be attributes of the heat cycle like height, step size etc.
3. Once the heat cycles are identified, how do they vary from one event to the next? A visualization would be great help showcase this variation.
   Note: The dataset consists of two sample measurements that are highly correlated with each other. We suggest the second measurement be used as a validation set.

3 Challenge 2: Towards a Universal Classifier for Crystallographic Space Groups

3.1 Background

State of the art electron microscopes produce focused electron beams with atomic dimensions and allow to capture diffraction patterns arising from the interaction of incident electrons with nanoscale material volumes. Backing out the local atomic structure of said materials requires compute- and time-intensive analyses of these diffraction patterns (known as convergent beam electron diffraction, CBED). Traditional analyses of CBED requires iterative numerical solutions of partial differential equations and comparison with experimental data to refine the starting material configuration. This process is repeated anew for every newly acquired experimental CBED pattern and/or probed material.

3.2 Dataset

In this data, we used newly developed multi-GPU and multi-node electron scattering simulation codes on the Summit supercomputer to generate CBED patterns from over 60,000 materials (solid-state materials), representing nearly every known crystal structure[5]. The overarching goals of this data challenge are to: (1) explore the suitability of machine learning algorithms in the advanced analysis of CBED and (2) produce a machine learning algorithm capable of overcoming intrinsic difficulties posed by scientific datasets.

The dataset is split across multiple HDF5 files and an accompanying Jupyter Notebook provides a detailed description on how to navigate the file structure to access the data samples and the associated materials properties. Briefly, a data sample from this data set is given by a 3D array formed by stacking three CBED patterns simulated from the same material at three distinct material projections (i.e. crystallographic orientations). Each CBED pattern is a 2D array (512×512 pixels) with float 32-bit image intensities. The dataset is 589 GB and a workstation or small cluster should be suitable for developing solutions to this challenge.

Associated with each data sample in the data set is a host of material attributes or properties which are, in principle, retrievable via analysis of this CBED stack. These properties consist of the crystal space group the material belongs to, atomic lattice constants and angles, chemical composition, to name but a few. Of note is the crystal space group attributed (or label). All possible spatial arrangements of atoms in any solid (crystal) material obey symmetry conditions described by 230 unique mathematical discrete space groups.

3.3 Challenges of Interest

The data challenge tasks revolve around developing and implementing a machine learning (ML) algorithm to predict a material's space group, essentially a classification task. The data set is, however, heavily imbalanced (i.e. number

of data samples per class). This imbalance is not an artifact, instead it reflects the reality that most known materials have low symmetry and as such are not uniformly distributed across the 230 space group classes.
The challenges are:

1. Perform exploratory data analysis on both CBED patterns and materials properties to summarize data characteristics.
2. Develop an ML algorithm for space group classification of CBED data.
3. Implement proper ML techniques to overcome data/label imbalance and show how it affects the performance of the ML algorithm in (1).
4. Implement an ML algorithm for multi-task prediction of a space group in addition to other material structural properties and show how it affects the performance of the ML algorithm.

Notes on challenge tasks

1. Preliminary task (1) is meant to provide better understanding about both input data (e.g. principle components of input images) and targets (e.g. distribution of space groups).
2. A participant may choose to solve challenges (1), (2) and (3), or (1), (2) and (4). Solving all 4 questions is optional.
3. Regarding approaches to (2), our preference is for ML techniques (e.g. loss-weighting, model ensembles, active learning, decision boundary analysis with GANs, etc.,), in lieu of brute-force data augmentation approaches (e.g. mixup, random erasing, etc. . .).
4. If a deep learning model is used by the participant, our preference is for the implementation use one of the following three frameworks: MXNet, Pytorch or TensorFlow.
5. Our preference is for the ML algorithms be implemented in one of the following languages: Python, C, C++, and/or Julia.

4 Challenge 3: Impacts of Urban Weather on Building Energy Use

4.1 Background

Recent advances in multi-scale coupling of high-performance computing models provide unique insights into how interdependent processes affect one another. Some of these processes are uniquely observable in urban environments. This data challenge addresses questions at the intersection of the natural environment and urban infrastructure by encouraging participants to examine variations in weather and building energy use, seasonal influences, and the building types most sensitive to weather at daily, monthly, and yearly scales. The dataset for this challenge was generated under a Laboratory Directed Research and Development project aimed at examining the impact of an area's built environment on weather and energy use. Data includes a year of simulated weather data taken at 15-minute intervals in a section of downtown Chicago; the latitude/longitude location for each building in the study area, each building's 2D footprint and height, and a year of building-by-building energy use simulation (EnergyPlus) data run by Joshua New, Mahabir Bhandari, Som Shrestha (ORNL, Energy and Environmental Sciences Directorate), and Mark Adams (ORNL, National Security Sciences Directorate).

4.2 Dataset

The dataset [6] comprises of three elements:

1. High resolution, 90m simulated weather data for 1 year at 15-minute intervals (with known gaps toward the end of each month). These files are provided in a comma separated value (CSV) format.
2. A mapping of individual buildings with individual IDs, their latitude/longitude location, and height (provided in Excel file).
3. Energy simulation output of these individual buildings, at 15-minute intervals for one year (provided in java script object notation (JSON) and other files).

The total data set is 6.35 GB and a laptop, workstation or small cluster should be suitable for developing solutions to this challenge.

4.3 Challenges of Interest

The questions that are of interest for this challenge are:

1. Are there interesting variations in weather and building energy use data for the geographic area?
2. Which buildings in the study are most sensitive to weather (e.g., temperature, humidity, wind, radiation) effects?
3. Are there any interesting visualizations that illustrate the changing dynamics of the simulated urban environment?
4. How can the data best be divided into subsets for meaningful analysis and visualization?
5. How does energy use in each building change through the year?
6. How is energy use different during the coldest and hottest months (e.g., January and July) of the year as compared to during those of less extreme temperature?

Participants are welcome to bring in additional datasets to combine with the provided data to create meaningful insights.

We look forward to presentations using novel methods for interpreting and visualizing this data that draw on machine learning and other big data techniques. We welcome new collaborations to complement the work of understanding climate, infrastructure, and energy use in urban areas from a systems perspective. We hope the participants enjoy the interdisciplinary nature of the dataset and its challenges.

5 Challenge 4: Computational Urban Data Analytics

5.1 Background

Urban environments are complex systems in which social factors, mobility, building energy, and urban climate interact with each other. Large parts of urban energy use are driven by the movement of population through the city. Each day, humans consume energy, whether they are traveling, at home, or in their workplace. Transportation and building energy are two of the top consumers of energy use in the United States. Transportation accounts for 29% of energy use, whereas buildings account for 38-40% of energy use (combined residential and commercial) [25] [26].

There are many factors that influence energy use in any particular building, but one of the major contributing factors is the number of occupants. In this challenge, we provide data that can help model the population's behavior and decision making, and obtain a more accurate representation of energy use in buildings, which nicely builds upon our previously published work [7], where we developed a data-driven transportation model, which determines building occupancy throughout the day in order to create more accurate simulations for building energy.

5.2 Dataset

Due to the tightly coupled nature of urban systems, we provide a wide variety of data for this challenge. All data is from 2017 (or based on 2017 inputs in case of simulations) unless noted otherwise. We hope that with this breadth of available data, every challenge participant will find an area they are particularly passionate about, however it is not a requirement to use all provided data. The size of the dataset is about 2 GB unzipped and a laptop, workstation or small cluster should be suitable for developing solutions to this challenge.

Vehicle data Vehicle data provides information on the population's daily trips as well as vehicle types from survey data. It is a simulation snapshot of a transportation simulation that was based on surveyed data.

1. Simulation snapshot for morning commute from TRansportation ANalysis SIMulation System (TRANSIMS): This snapshot contains vehicle traces (in Universal Transverse Mercator Coordinates) at 30-second intervals for one simulated day. At each time step, we also have the link (road segment) ID, driver ID, and vehicle speed.
2. Schedule for morning commute from National Household Travel Survey (NHTS): This is an extract of the official NHTS data [8] which only contains survey responses from Chicago.
3. Vehicle type distribution: Simplified Federal Highway Association (FHWA) classifications of vehicles in Chicago, which was derived from NHTS data.

Emissions data To study emissions, we are providing traffic volumes and emissions that were generated using an emissions simulation, by using the traffic simulation outputs. In addition, we provide weather data to enable the study of relations between weather and emissions.

1. Road-level traffic volumes (aggregated from TRANSIMS outputs).
2. Road-level emissions generated using MOVES, an emissions simulator. This simulation is based on traffic volumes and weather patterns throughout a year.
3. Weather data from DarkSky. For this data, we provide instructions on downloading it, as it is free to use but the license agreement does not allow redistribution.

Road network A transportation-focused dataset would not be complete without the road geometry. We provide the road network that was used for the simulations, along with some metadata.

1. The road network has the link IDs for each road segment, as well as road type etc.
2. GeoJSON file of the road network used for the TRANSIMS and MOVES runs.
3. Definition of different link types.

Building data We provide building footprints and socioeconomic data to provide a better idea of population distribution and demographics, and the type and distribution of buildings throughout the area.

1. Building footprints from Microsoft (2019): All US building footprints [9] and the clipped version for Chicago [10].
2. Land Use data from Chicago Metropolitan Agency for Planning (CMAP): GeoJSON file containing polygon data with land use attributes and a Codebook defining the land use codes
   (a) Socioeconomic data is provided for different *community areas* (neighborhoods, such as "Chicago Heights") [12]:
      i. Population from CMAP/Census (2010).
      ii. Census data summarized to community areas [11].
      iii. Spreadsheet (CSV) of census data by community area.
      iv. GeoJSON of community area polygons.
   (b) Community Area Snapshots contain additional information such as employment, travel mode choice, housing types, job types in community (held by residents, available in community), walkabilty, etc.[12].
      i. Spreadsheet (CSV) of Community Area Snapshot data.
      ii. Data dictionary explaining the different fields.

Each of the folders in the provided dataset has a README file with more detailed information on file format and contents.

5.3 Challenges of Interest

One of the main challenges in coupled or integrated systems is the disparity of data sources. For this data challenge, we would like participants to address one of the three following tasks:

- Develop an algorithm to efficiently assign vehicle occupants to nearby buildings.
  1. In [7], we have performed an initial weighted quad tree-based approach to map vehicles to buildings.
  2. The ideal algorithm should be efficient and accurate. Consider the tradeoff.
  3. The resulting mapping should be realistic. Consider building size, use type (the vehicle traces are only for commute) etc.
- Perform an area-wide correlation analysis of vehicle emissions.
  1. Determine spatial variation, and variation based on other factors, such as land use of surrounding areas, population, network classification (road type), weather, etc.
  2. Correlate the provided emissions data with other provided datasets.
- Characterize traffic patterns from the simulation.
  1. What are the traffic hot spots? Is there any congestion?
  2. What are the travel times? How do they vary through out the day?
  3. What are busy times? How well do they match the commute pattern from NHTS?
  4. How do speeds vary spatially and temporally?
  5. What are the most popular roads?
  6. Can you draw conclusions about the simulation setup from the output?

We hope that the wide range of questions will provide an interesting challenge for every participant. If participants have their own unique ideas based on using the data we provided, this will also be of interest.

6 Challenge 5: Using Machine Learning to Understand Uncertainty in Subsurface Exploration

In the energy industry, an understanding of subsurface characteristics and structure is crucial to identifying and localizing untapped resources. At a high level, the process of taking an entirely unexplored region of earth and generating an actionable understanding of its structure includes:

1. Seismic data collection: Collect raw signals from the subsurface using techniques similar to sonograms used in hospitals.
2. Seismic data pre-processing: Quality check and clean the collected raw signals.
3. Seismic migration and velocity model construction: Use the raw signals and our understanding of the likely geology of the region to construct a 3D representation of the subsurface.

4. seismic interpretation: Using the constructed 3D representation, interpret where faults, layers, and other important structural features are in the subsurface.

With each of these steps comes an amount of uncertainty from various sources of potential error: instrument error, human error, modeling error, and more. Despite this, the output of most seismic processing workflows is a single, gold standard, output image. An image which we know cannot possibly be 100% accurate!

It is crucial that future seismic processing workflows start to incorporate uncertainty when estimating the true subsurface structures. Rather than outputting a single interpretation, we should aim to emit a spectrum of possible realizations and an understanding of where uncertainty is high or low.

6.1 Background

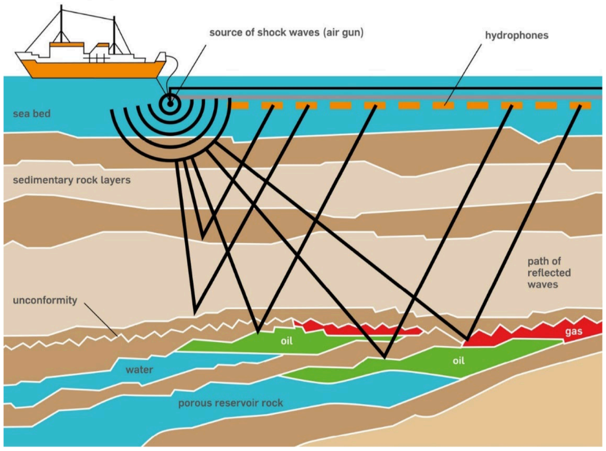
Seismic Data Collection Seismic data collection (i.e., the process of conducting a seismic survey) involves transmitting powerful sound waves into the ground and then recording their echoes at the surface as they bounce off boundaries between layers in the Earth. This process parallels techniques used in x- ray and ultrasound imaging in the medical field to reconstruct structures inside the human body. The figure below depicts a typical offshore seismic survey setup, in which sound waves are transmitted from an air gun behind a survey ship and the return echoes are recorded by a line of hydrophones being towed behind the ship.

During a seismic survey, one or more sources of sound energy are used to transmit waves into the ground. One or more receivers are used to record the reflection of that sound energy at the surface. The raw output generated from a seismic survey is a set of recorded waveforms at each receiver for each source. This recording stores the amplitude of the reflected sound wave at the surface as a function of the time it took to travel to the receiver.

Seismic Data Preprocessing Pre-processing of our raw seismic data can include a multitude of steps. Broadly, seismic preprocessing aims to clean up and strengthen signals in the seismic data while reducing noise, facilitating later stages of the seismic processing pipeline.

Seismic Migration and Velocity Model Construction Seismic migration refers to the process by which the seismic waves received at receivers are backpropagated to the source through a simulated version of the seismic medium. Through knowledge of (1) the source location, (2) the receiver location, (3) the time/amplitude of the received signal, and (4) the medium through which the signal traveled, we can simulate in reverse the propagation of the signal through the subsurface, identify its reflection point, and thereby identify the location of a potential object/reflector of interest in the subsurface.

Fig. 2. Seismic Survey[13]



Note how crucial an accurate estimate of the subsurface velocity of sound waves is in this process. Without an accurate velocity estimate, it is impossible to accurately predict the distance traveled by sound waves in the subsurface in a certain period of time.

While building a velocity model is a critical component for accurate seismic reconstruction, a number of uncertainties are involved in the process. Simply asking two different geophysicists to perform velocity model construction on the same seismic traces can produce drastically different velocity models. Quantifying and visualizing this uncertainty in velocity models will be the prime focus of this data challenge.

One common practice for checking the validity of a given velocity model is through offset pair gathers. Modern seismic surveys generally involve many sources and many receivers. As a result, many pairs of sources and receivers capture reflections off the same reflector in the subsurface (see Figure 3). This redundancy can be helpful in validating the quality of a velocity model, as an accurate velocity model is expected to produce similar/identical depth estimates for a given reflector no matter which offset pair a reflection is received from.

Gathers generally refer to collecting the depth estimate for a given reflector across many offset pairs and plotting them visually, with depth on the y axis and offset pairs along the x axis. In a gather of an accurate velocity model, geophysicists expect to mostly see horizontal lines, indicating that the depth estimate for a layer is the same across all offset pairs. See Figure 4 for several examples of reasonable gathers, indicated by the prevalence of horizontal lines.

Fig. 3. Many pairs of sources and receivers capture reflections off the same reflector in the subsurface.

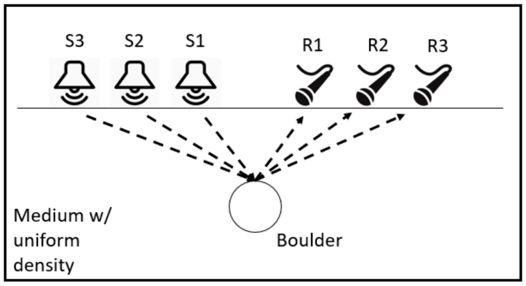
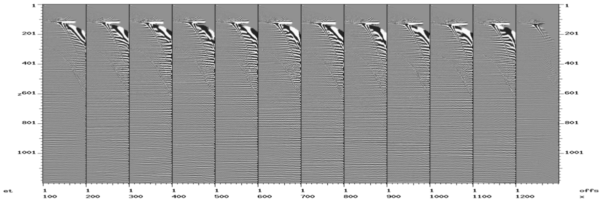
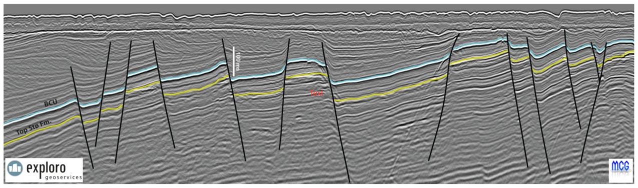


Fig. 4. Reasonably good gathers show mostly horizontal lines, indicating that the depth estimate for a layer is the same across all offset pairs.



Seismic Interpretation Once a final seismic image is rendered following seismic migration, seismic interpretation—the process of identifying faults, reservoirs, and other features of interest in the image—begins. This manual labeling is then used in field development and reservoir characterization. See Figure 5 for an example seismic image with faults manually labeled and emphasized.

Fig. 5. A seismic image with faults manually labeled and emphasized[14]



Dataset The dataset included in this data challenge serves as a starting point in exploring techniques for quantifying uncertainty in seismic processing workflows. In this dataset we are focused on quantifying and visualizing the uncertainty in our estimations of the density of the subsurface based on how varying those estimates impacts our output 3D volume. At a high level, this dataset consists of a set of synthetic but realistic models of the density of the subsurface, randomly generated based on a single, known, synthetic ground truth. This dataset also includes the final 3D realizations generated using those density models (also called velocity models). These files are stored in the industry standard SEGY format, and an example Jupyter notebook is provided to illustrate how to load and visualize them.

A 3GB trial dataset is given to help competitors get started. The full dataset is 49 GB. A laptop, workstation, or small cluster should be suitable for developing solutions to this challenge.

6.2 Challenges of Interest

The end goal of this data challenge is to construct an uncertainty map for a given seismic survey, labeling each pixel in a final 2D seismic image with a value between 0.0 and 1.0 indicating how volatile the estimate for that pixel is.

However, we also welcome submissions that include any intermediate work towards that end goal or answers to any of the below challenge questions. Even if you are unable to complete the entire challenge, any submissions that show progress towards this end goal and lay out ideas for how the challenge could eventually be completed will be considered.

1. Given that geophysicists generally use horizontal lines in gathers as a good indicator of velocity model accuracy, build a model (analytical, mathematical, data-driven, or otherwise) to estimate the quality of each velocity model based on its associated gathers.
2. Train a model to label each pixel with an uncertainty value between 0.0 and 1.0 indicating how uncertain any given realization of that part of the subsurface is.
3. Generate a single uncertainty map given all of the velocity models, realizations, and gathers at hand.
4. Generate some form of visualization of this uncertainty map of the subsurface.

7 Challenge 6: Using Artificial Intelligence Techniques to Match Patients with Their Best Clinical Trial Options

The Presidential Innovation Fellows, US Department of Veterans Affairs, and the Oak Ridge National Laboratory Health Data Sciences Institute are coordinating this Data Challenge, which draws on resources across a dozen federal agencies and departments. The related project, Health Tech Sprint, emphasizes the need for open federal data for artificial intelligence (AI) applications as defined by the newly signed OPEN Government Data Act under the Foundations for Evidence-based Policymaking Act (signed Jan 15, 2019).

7.1 Background

Novel therapeutics, such as those under development in clinical trials, are often a treatment option for patients with serious and life-threatening diseases such as cancer. Increasing patient awareness of clinical trials is believed to be a factor in reducing time for participant recruitment, a very large cost category in clinical trials. Thus, applying AI to help patients and their health care providers find clinical trials of novel therapeutics may improve patient care and, by aiding in recruitment, reduce drug development costs.

For AI to be useful in trial matching, both representative patient data and clinical trial eligibility information, ideally in a structured format, are needed. In addition, expert-based guidance on matching patients to trials, including which criteria are matched, is useful for building and testing models.

The AI-able data ecosystem seeks to enable AI by bringing together an ensemble of interlinked datasets with data suitable for AI in a given use case. Having this information in the public domain enables standardization by facilitating testing across different approaches. This challenge features the first such standardized dataset ensemble related to clinical trial matching, with the various interlinked datasets provided.

7.2 Dataset

We provide three datasets to the data challenge participants

1. A subset of eligibility criteria translated into machine-readable code from a selected group of cancer clinical trials.
2. Records based on callers to the NCI's Cancer Information Service that have been enhanced with synthetic data and translated into machine-readable code.
3. Participant records matched against clinical trials for which the eligibility criteria and participant data were previously translated into machine-readable code.

The size of the three datasets is 1.3 MB and a laptop, workstation, or small cluster should be suitable for developing solutions to this challenge. A second version of the third dataset, produced by oncology professionals, serves as a comparison dataset for the matches identified through the application of AI. For more information on the above datasets and potential approaches on usage, please see reference [15].

In addition to the datasets provided, participants are encouraged to use other publicly available datasets. For example, National Cancer Institute (NCI)-funded cancer clinical trials, including API with annotations on disease eligibility criteria for all trials, is available at https://clinicaltrialsapi.cancer.gov.

7.3 Challenges of Interest

Challenge tasks are listed below. However, participants are encouraged to suggest and tackle challenge tasks different from those listed below. Innovative use of the provided data is strongly encouraged.

1. Data representation
   - Develop novel big data structures to represent the clinical trials and the patient data that accommodate the interaction of the three datasets. The ultimate goal is to support thousands of clinical trials being matched with millions of people.
2. Algorithm development
   - Develop novel algorithms for finding the most suitable matches between patients and clinical trials.
3. Visualization/human computer interaction
   - Develop visualization and/or human-computer interaction solutions to enable medical providers to effectively leverage the data for clinical decision support.

Notes on the Challenge Tasks

1. A participant may choose to do any question(s) they prefer. Completing all three questions is optional.
2. Regarding approaches to question 2, our preference is to receive solutions involving machine learning techniques.

8 Challenge 7: The Kaggle CORD-19 Data Challenge

8.1 Background

As governments, policymakers, and scientists across the globe are racing to identify potential vaccines and drugs for SARS-CoV-2, many scientists hope the information needed to identify a vaccine lies in the millions of available research documents. To support mining information from research literature, the White House, along with leading industries, has made a dataset of research publications directly related to the outbreak available to the general public [16]. Some of the most important questions pertaining to the outbreak which were identified by the US NASEM and the WHO, were published as part of a public challenge along with the publication dataset on Kaggle [17].

8.2 Dataset

The entire body of scientific literature is growing at an enormous rate; it is currently estimated at over 100 million publications [20] with an annual increase of more than 5 million articles. The publication set of coronavirus-related literature provided for the Kaggle COVID-19 Open Research Data Challenge (CORD-19) have been growing at a rate of thousands of new publications per year (Figure 8.2) and the growth has nearly doubled since the start of the current epidemic. Thus, it is not only difficult for scientists to source inspiration and new insights from their own domains, but also other adjacent domains. Since comprehensive reading of the growing scientific literature is now beyond the capacities of any human being, artificial intelligence techniques, including natural language processing and text mining, offer the potential to intelligently parse large bodies of loosely connected text to provide scientists solving some of the world's most pressing challenges with meaningful insights [23, 24, 22].

8.3 Challenges of Interest

To kick-start the development of such AI techniques, the Kaggle CORD-19 challenge lists some of the most important questions pertaining to the COVID-19 epidemic, which will require parsing and connecting information provided in the available literature. This list of questions evolves as we learn more about the virus and identify new questions which need to be answered, and includes questions about symptoms, risk factors such as pre-existing conditions, and vaccines and therapeutics currently under investigation. However, answering some of the questions may require going beyond existing CORD-19 publication set. What if an existing vaccine developed for another disease has a potential to also work for COVID-19, but hasn't as of yet been mentioned in COVID-19 related literature? Expanding beyond just the directly relevant literature exacerbates the need for AI techniques.

We invite submissions describing complete or partial solutions to any of the Kaggle CORD-19 Tasks to SMCDC for consideration for a best solution paper

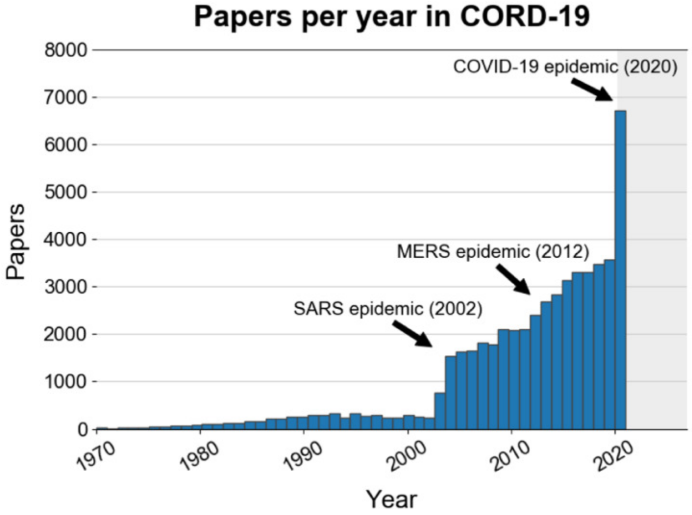


Fig. 6. Growth of papers in the CORD-19 dataset. Figure from [21].

award, poster presentation, and publication in the conference proceedings. The SMCDC poster session will give selected researchers perusing the CORD-19 dataset a place to present their work and discuss it with other researchers.

9 Conclusion

In addition to contributing to the solutions of open research questions, we hope these challenges gave participants a taste of the types of data and modeling problems in each of the scientific areas featured in the 2020 Data Challenge. We also hope these challenges got researchers thinking about how important and difficult it is to account for uncertainty and probability in large scientific datasets.

In total, 52 teams competed to solve the seven data challenges. Of those, 23 teams submitted solution papers. The best solutions were selected for publication by a peer review.

About 90 percent of the finalists identified themselves as students. According to studies in educational psychology such as [18] and [19], novel intellectual challenges, like those posed by the 2020 Data Challenge, can be highly motivating and promote deeper engagement in tasks and lead to longer-term persistence in academic pursuits like research.

10 Acknowledgements

This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the

Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725"

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725.

References

1. https://neutrons.ornl.gov/vulcan
2. Granroth, G.E., An, K., Smith, H.L., Whitfield, P., Neuefeind, J.C., Lee, J., Zhou, W., Sedov, V.N., Peterson, P.F., Parizzi, A., and Skorpenske, H., 2018. Event-based processing of neutron scattering data at the Spallation Neutron Source. Journal of Applied Crystallography, 51(3).
3. Wang, X.L., Holden, T., Stoica, A.D., An, K., Skorpenske, H.D., Jones, A.B., Rennich, G.Q., and Iverson, E.B., 2010. First results from the VULCAN diffractometer at the SNS. In Materials Science Forum (vol. 652, pp. 105–110). Trans Tech Publications.
4. Niyanth. S, Noyan I.C., Seren, M.H., and An, K., Vulcan Beamline dataset. Partly supported by the US Department of Energy (DOE) Office of Energy Efficiency and Renewable Energy, Advanced Manufacturing Office program. This research used resources at the SNS, a DOE Office of Science User Facility operated by Oak Ridge National Laboratory. doi:10.13139/ORNLNCCS/1604074
5. Laanait, N., Borisevich, A., and Yin, J., A Database of Convergent Beam Electron Diffraction Patterns for Machine Learning of the Structural Properties of Materials. doi:10.13139/ORNLNCCS/1604074
6. Allen-Dumas, M., New, J. Chicago microclimate and building energy use data doi: 10.13139/ORNLNCCS/1619243
7. Berres, A., Im,P., Kurte, K., Allen-Dumas, M., Thakur, G., Sanyal, J.: A Mobility-Driven Approach to Modeling Building Energy. 5th IEEE Workshop on Big Data Analytics in Supply Chains and Transportation. Los Angeles (2019).
8. https://nhts.ornl.gov/
9. Microsoft building footprints https://github.com/Microsoft/USBuildingFootprints
10. https://usbuildingdata.blob.core.windows.net/usbuildings-v1-1/Illinois.zip
11. 2010 Census data for Chicago community areas https://datahub.cmap.illinois.gov/dataset/2010-census-data-summarized-to-chicago-community-areas
12. https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data
13. https://krisenergy.com/company/about-oil-and-gas/exploration/
14. https://www.geoexpro.com/articles/2016/01/super-high-resolution-seismic-data-in-the- norwegian-barents-sea"
15. https://digital.gov/2019/02/27/how-a-health-tech-sprint-inspired-an-ai-ecosystem.
16. https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/
17. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks
18. Csikszentmihalyi, M. Flow: The psychology of optimal experience. New York: Harper Perennial (1990).
19. Shernoff, D. J., and Hoogstra, L. Continuing mo- tivation beyond the high school classroom. New Direc- tions for Child and Adolescent Development, 93, 73-87 (2001).

20. Khabsa, Madian, and C. Lee Giles. "The number of scholarly documents on the public web." PloS one 9, no. 5 (2014).
21. Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk et al. "CORD-19: The Covid-19 Open Research Dataset." ArXiv (2020).
22. Wang, Kuansan, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. Microsoft Academic Graph: When experts are not enough. Quantitative Science Studies 1 (1): 396-413 (2020).
23. Wade, Alex D., and Kuansan Wang. "The rise of the machines: Artificial intelligence meets scholarly content." Learned Publishing 29, no. 3 (2016): 201-205.
24. Saggion, Horacio and Ronzano, Francesco. Scholarly data mining: making sense of scientific literature. 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL): 1–2 (2017).
25. U.S. Energy Information Administration. Use of energy in the United States – Energy explained. https://www.eia.gov/energyexplained/index.php
26. DOE Office of Energy Efficiency & Renewable Energy efficiency trends in residential and commercial buildings. http://www.osti.gov/servlets/purl/1218835/