Downloaded 02/10/22 to 120.224.221.8. Redistribution subject to SEG license or copyright; see Terms of Use at http://library.seg.org/page/policies/terms
DOI:10.1190/tle40010068.1

WORKSHOP PREVIEW

Data Analytics and Machine Learning Hackathon 2021: A deep dive into the open-source data challenge for E&P

Haibin Di[1], Anisha Kaul[1], Leigh Truelove[1], Weichang Li[2], Wenyi Hu[3], and Aria Abubakar[1]

https://doi.org/10.1190/tle40010068.1

Abstract

We present a data challenge as part of the hackathon planned for the August 2021 SEG Research Workshop on Data Analytics and Machine Learning for Exploration and Production. The hackathon aims to provide hands-on machine learning experience for beginners and advanced practitioners, using a relatively well-defined problem and a carefully curated data set. The seismic data are from New Zealand's Taranaki Basin. The labels for a subset of the data have been generated by an experienced geologist. The objective of the challenge is to develop innovative machine learning solutions to identify key horizons.

Introduction

The SEG Research Workshop on Data Analytics and Machine Learning for Exploration and Production is designed to showcase the successes and challenges of practicing machine learning and data analytics in the geoscience domain to improve accuracy and efficiency of algorithms and/or workflows. For example, in addition to existing applications such as seismic geobody identification and well-log data analysis, what other challenging geoscience problems can be formulated and solved effectively by machine learning? How can machine learning algorithms and workflows be tailored to meet the specific physical constraints of geoscience data? How can we fully exploit the power of machine learning and physics-based approaches? What are the roadblocks to applying machine learning to production in the oil and gas industry? The workshop topics will cover seismic interpretation, seismic processing, multiphysics data integration, reservoir characterization, wellbore, and advances in machine learning algorithms.

The objective of the hackathon is for teams or individuals to propose an innovative solution within their data science pipelines to identify key horizons within the given seismic survey. Individuals of a diverse skill set are invited to participate and apply knowledge from their domains. There is no restriction on the way the problem is solved or the way the machine learning problem is formulated, as long as the output of the data science pipeline outputs the key horizons. At the time of the hackathon kickoff, participants will be given a brief overview of expectations. As discussed in the section "Script to explore the data," Python code for performing exploratory analysis on the data is made available to those who wish to use it. Participants are also encouraged to brainstorm and commence the exploratory analysis on the curated data provided.

We plan to host the hackathon over a period of three months and encourage participants to share their findings and observations in the form of Jupyter notebooks via a common discussion forum provided by the organizers. The top three winners of the hackathon will be chosen after the submissions are evaluated against criteria described in the "Submission and evaluation" section. At the beginning of the hackathon, the participants will be provided with a seismic cube (S_T) of SEG-Y format. Within the same cube, a few lines will be manually annotated with fault and horizon interpretations (L_{TF} and L_{TS}) in NumPy (*.npy) and MATLAB (*.mat) matrix format. The manually labeled interpretations will be provided in sets of two — one for training and the other for validation. Even though the deliverable of the competition is only horizon predictions, fault labels have been provided for participants who wish to be creative by complementing their training with fault labels. As shown in the section "Script to explore the data," this filename can be used not only to obtain the labels but also the seismic indices of the labels to match with the seismic data. For instance, the stratigraphy labels in the inline direction that are evenly spaced from inline 2087 to 2687, can be found in a file called "Horizon_label_IL2087_2687_7.npy." The dimension of the fault and the stratigraphy matrices $L_{TF(Inline)}$ and $L_{TS(Inline)}$ will be (N_F, Y, Z, 1) and (N_S, Y, Z, H), respectively. Similarly, for crossline, the fault and stratigraphy labels $L_{TF(Crossline)}$ and $L_{TS(Crossline)}$ will be (N_F, X, Z, 1) and (NS, X, Z, 1), where N is the number of interpreted lines, and X, Y, and Z are the dimension of seismic cube S_T in inline, crossline, and depth directions, respectively. The last dimension of the matrix will be 1 for L_{TF} and H for L_{TS} depending on how many horizons we provide labels for. For instance, as shown in Figure 1, if we provide labels for three distinct horizons, the value of H will be 3. Each X * Z

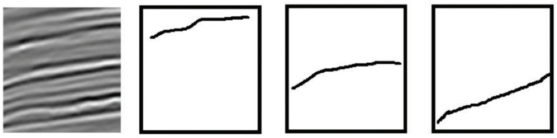

Figure 1. (a) Seismic line, (b) horizon 1 label, (c) horizon 2 label, and (d) horizon 3 label.

[1]Schlumberger, Houston, Texas, USA, and West Sussex, Gatwick, UK. E-mail: hdi@slb.com; akaul@slb.com; ltruelove@slb.com; aabubakar@slb.com.
[2]Aramco Research Center, Houston, Texas, USA. E-mail: weichang.li@aramcoamericas.com.
[3]Advanced Geophysical Technology, Houston, Texas, USA. E-mail: wenyi.hu@agtgeo.com.

Downloaded 02/10/22 to 120.224.221.8. Redistribution subject to SEG license or copyright; see Terms of Use at http://library.seg.org/page/policies/terms
DOI:10.1190/tle40010068.1

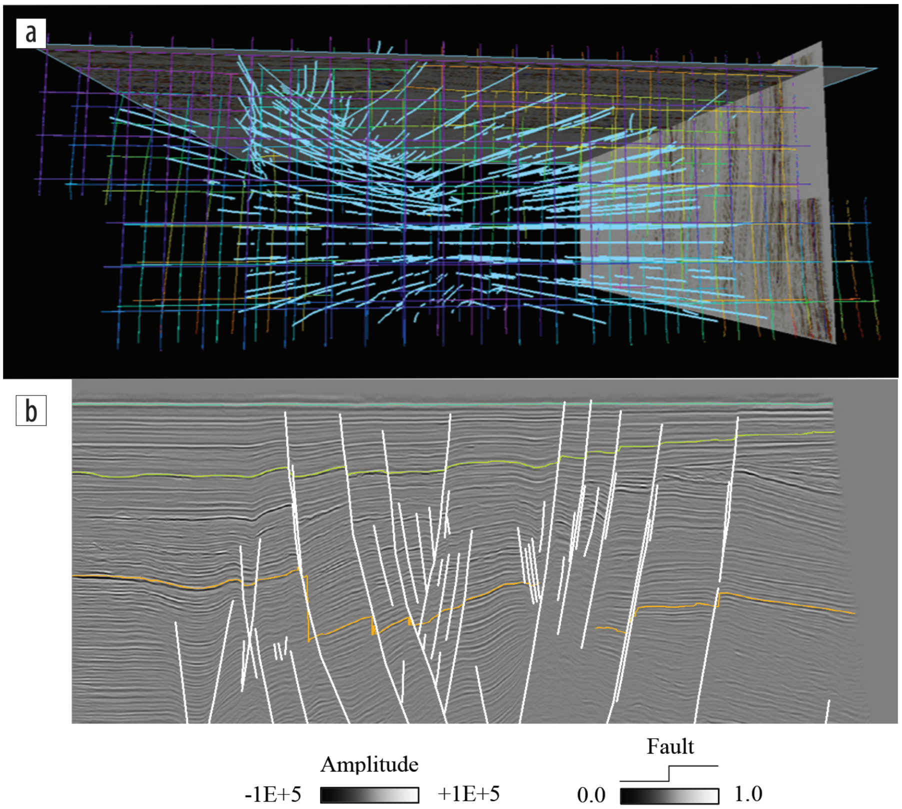

Figure 2. (a) 3D view of the seven inlines and 23 crosslines with interpreter annotation, used as training labels for the machine learning-assisted fault and stratigraphy interpretation. (b) Example of the fault and horizon annotations on inline section 2587.

or Y * Z label matrix is binary, with 1 where there exists a pick and 0 where there does not. Providing labeled examples will allow participants to understand the geology of the survey, the nature of the labels, and the expected quality of results.

Opunake data set

Geology description. The Taranaki Basin is an asymmetric basin consisting of up to 8 km of Upper Cretaceous to Quaternary sedimentary fill that covers approximately 100,000 km^2 underlying the shelf and continental slope west of central-western North Island of New Zealand (King and Thrasher, 1996; O'Neill et al., 2018; Mattos et al., 2019). The basin can be subdivided, by the Cape Egmont Fault Zone, into two broad regions — the Western Platform, which is characterized by relatively undeformed strata, and the Eastern Mobile Belt, which is significantly more deformed (King and Thrasher, 1996). Particularly, the complex deformation history of the Eastern Mobile Belt includes extensional episodes that were associated with basin subsidence and compression between the Australian and Pacific plates that occurred from the Late Cretaceous to Early Eocene (Holt and Stern, 1991, 1994; King and Thrasher, 1996; Nicol et al., 2005; Stagpoole and Nicol, 2008; Giba et al., 2010), which divides the Eastern Mobile Belt into three distinct geologic provinces: Northern Graben, Central Graben, and the buried Miocene Mohakatino Volcanic Centre (Neall et al., 1986; Stagpoole and Funnell, 2001; Hansen and Kamp, 2004). Among them, the Central Graben lies mostly south of the Taranaki Peninsula between the faulted Cape Egmont high in the west and the Manaia anticline in the east. This deep graben contains thick, mature source rocks, which contributed to the Maui and Maari-Manaia fields. Reservoir facies are common at several levels in the Cretaceous and Cenozoic rocks.

3D seismic. The Opunake-3D seismic volume data provided by the New Zealand Crown Minerals is a prestack-migrated final volume located just within the Central Graben, covering an area of 215 km^2 and acquired in 50–120 m water depth. The seismic quality is variable, with the shallower sections showing well-defined reflectors with clear fault compartmentalization, while the deeper section loses most clarity with almost all the reflector definition obscured. Simple attribute analysis using a Gaussian (structural) smoothing and variance (coherence) shows the low signal-to-noise characteristics of the deeper sequence. Therefore, this makes interpretation of the deeper (basement) sequences far harder, so the focus for this area must be on the shallower, faulted sequences. Correspondingly, a subset of the volume is cropped from the shallow area for the data challenge, which consists of 765 inlines and 2233 crosslines.

Interpretation labels. Both inline and crossline directions were selected from the cropped Opunake-3D volume at a line spacing of 100 and 200, respectively, meaning seven inlines and 23 crosslines were picked manually for faults, which is about 1% of the seismic data interpreted as labels (Figure 2a) (Di et al., 2020). Three horizons from shallow to deep are set as the target of the data challenge. Meanwhile, the faults in the 30 sections are annotated to represent the faulting system, and these fault labels are provided together with the horizon labels to use at the participants' choice. An example of the manual picks on inline 2587 is shown in Figure 2b, with the faults in white lineaments and the three horizons in color curves as the boundaries between four stratigraphic sequences. Note the missing picks of horizon 3 (in orange) in the deep zones where the seismic quality is relatively low and the interpreter is not confident with providing his interpretation. It is worth mentioning that such incomplete interpretation is commonly observed in various real seismic data sets and pose a challenge in machine learning-assisted horizon interpretation.

Script to explore the data

Due to limited space, the script is provided only partially on the following page. The full script can be found at the data hackathon website (https://github.com/mlenp/seg-summer-workshop-data-challenge), which will be made available at the time of the hackathon.

Downloaded 02/10/22 to 120.224.221.8. Redistribution subject to SEG license or copyright; see Terms of Use at http://library.seg.org/page/policies/terms
DOI:10.1190/tle40010068.1

Seismic SEGY

```
# Get seismic data from SEGY
seisname = 'Opunake_seismic'
segyfile = './seismic/'+seisname+'.sgy'
seisdata, seisinfo = sb.io.readSeis3DMatFromSegyNoInfo(segyfile=segyfile, verbose=True)
# Save data back to SEGY
# Please ensure the data for saving (savedata) being of the same dimension as the reference segy (refsegy).
# Otherwise, use writeSeis3DMatToSegyNoRef with defined survey information.
# Rescaling is used as an example to create the new data for saving
savedata = seisdata * 0.0001
savename = './seismic/Opunake_seismic_rescaled.sgy'
refsegy = './seismic/Opunake_cropped_IL2100_2600_XL2500_3500.sgy'
sb.io.writeSeis3DMatToSegyWithRef(seis3dmat=savedata, seisfile=savename, refsegy=refsegy)
```

Horizon label

```
# Get horizon picks from the prepared NumPy files
# The NumPy is binary, with 1 representing the target horizon picks and 0 for the rest. Each NumPy file is of
4 dimensions: the 1st for the No. of inline/crossline sections; the 2nd for the section height that equals to the
z dimension of seismic; the 3rd for the section width that equals to the crossline/inline dimension of seismic;
and the 4th for the No. of horizons, which is 3 in this case.
inline_horizon_label = np.load('./horizon/Horizon_label_IL2087_2687_7.npy')
xline_horizon_label = np.load('./horizon/Horizon_label_XL2070_6470_23.npy')
# Link the loaded horizon picks with the seismic survey by specifying the section No.
# Here horizon labels are provided on 7 inlines and 23 crosslines as below, which is expected consistent with
the 1st dimension of the loaded NumPy files.
list_of_inline_with_horizon_label = list(np.linspace(2087, 2687, 7))
list_of_xline_with_horizon_label = list(np.linspace(2070, 6470, 23))
```

Fault label

```
# Get fault picks from NumPy label files
# The NumPy is binary, with 1 representing the fault picks and 0 for the non-faults. Each NumPy file is of
4 dimensions: the 1st for the No. of inline/crossline sections; the 2nd for the section height that equals to the
z dimension of seismic; the 3rd for the section width that equals to the crossline/inline dimension of seismic; and the
4th being 1 in this case.
inline_fault_label = np.load('./fault/Fault_label_IL2087_2687_7.npy')
xline_fault_label = np.load('./fault/Fault_label_XL2070_6470_23.npy')
# Link the loaded horizon picks with the seismic survey by specifying the section No.
# Here the fault labels are also provided on the same 7 inlines and 23 crosslines as the horizon labels, which is
expected consistent with the 1st dimension of the loaded NumPy files.
list_of_inline_with_fault_label = list(np.linspace(2087, 2687, 7))
list_of_xline_with_fault_label = list(np.linspace(2070, 6470, 23))
```

Data visualization

```
# Specify sections for visualization
inline_for_vis = 2387
# seismic
plt.figure(1)
plt.imshow(seisdata[:, :, np.round((inline_for_vis-seisinfo['ILStart'])/seisinfo['ILStep']).astype(int)],
      cmap='seismic', vmin=-100000, vmax=100000)
plt.title('Seismic in inline No. ' + str(inline_for_vis))
# Horizon
plt.figure(2)
plt.imshow(inline_horizon_label[list.index(list_of_inline_with_horizon_label, inline_for_vis), :, :, 0], cmap='binary')
plt.title('Label of the 1st horizon in inline No. ' + str(inline_for_vis))
# Fault
plt.figure(3)
plt.imshow(inline_fault_label[list.index(list_of_inline_with_fault_label, inline_for_vis), :, :], cmap='binary')
plt.title('Label of faults in inline No. ' + str(inline_for_vis))
plt.show()
```

Downloaded 02/10/22 to 120.224.221.8. Redistribution subject to SEG license or copyright; see Terms of Use at http://library.seg.org/page/policies/terms
DOI:10.1190/tle40010068.1

Submission and evaluation

At the beginning of the hackathon, the participants are provided with validation seismic lines in NumPy (*.npy) and MATLAB (*.mat) format and of dimensions (N, X, Z). Once the participants have trained a machine learning model, to be eligible for competition they will need to submit the predicted horizons in the form of a NumPy or MATLAB matrix. The submitted prediction matrix needs to be of dimensions (N, X, Z, H), where H is the number of horizons we need predictions for and, these predictions, unlike the labels provided, need to be a probability mask, not binary mask. This competition is evaluated on the area under the receiver operating characteristics (ROC) curve of the predictions (Feng et al., 2019). For each horizon, the predictions will be binarized with thresholds at values [0.55, 0.95] with steps of 0.1. Once binarized, the Jaccard Index of similarity is given by

$$J(A, B) = |A \cap B| \,/\, |A \cup B| = |A \cap B| \,/\, |A| + |B| - |A \cap B|, \quad (1)$$

where A and B are the predictions and ground truth, respectively. If the intersection over union (IOU) of the prediction is above 0.5, it is considered a positive outcome, likely due to a higher true positive rate (TPR) and a true negative rate (TNR). A negative outcome is the result when the IOU of the prediction is below 0.5, likely due to a higher false positive rate (FPR) or a false negative rate (FNR). This way, the TPR, TNR, FPR, and FNR can be averaged over all the predictions on the test images and the ROC curve is populated. The final score for the submission will be the weighted average of the predictions of all horizons. Geologic plausibility also will be considered by the evaluation committee as a criterion and likely will be used to resolve submissions that draw on the final score based truly on numeric accuracy as stated above.

Timeline

The hackathon will open on 1 March 2021 when the data sets, labels, and some additional resources will be made available on the data hackathon website to the registered participants. It will run through 1 July 2021 when the submission will be closed for evaluation. A leaderboard will be made available during the hackathon.

Acknowledgments

We would like to thank the New Zealand Crown Minerals for providing the Opunake-3D seismic data set.

Data and materials availability

The interpreter annotations are free to share and available by contacting the authors.

Corresponding author: hdi@slb.com

References

Di, H., L. Truelove, C. Li, and A. Abubakar, 2020, Accelerating seismic fault and stratigraphy interpretation with deep CNNs: A case study of the Taranaki Basin, New Zealand: The Leading Edge, **39**, no. 10, 727–733, https://doi.org/10.1190/tle39100727.1.

Feng, K., H. Hong, K. Tang, and J. Wang, 2019, Decision making with machine learning and ROC curves: https://arxiv.org/abs/1905.02810.

Giba, M., A. Nicol, and J. J. Walsh, 2010, Evolution of faulting and volcanism in a back-arc basin and its implications for subduction processes: Tectonics, **29**, no. 4, https://doi.org/10.1029/2009TC002634.

Hansen, R. J., and P. J. J. Kamp, 2004, Late Miocene to Early Pliocene stratigraphic record in northern Taranaki Basin: Condensed sedimentation ahead of Northern Graben extension and progradation of the modern continental margin: New Zealand Journal of Geology and Geophysics, **47**, no. 4, 645–662, https://doi.org/10.1080/00288306.2004.9515081.

Holt, W. E., and T. A. Stern, 1991, Sediment loading on the Western Platform of the New Zealand continent: Implications for the strength of a continental margin: Earth and Planetary Science Letters, **107**, no. 3–4, 523–538, https://doi.org/10.1016/0012-821X(91)90098-3.

Holt, W. E., and T. A. Stern, 1994, Subduction, platform subsidence, and foreland thrust loading: The Late Tertiary development of Taranaki Basin, New Zealand: Tectonics, **13**, no. 5, 1068–1092, https://doi.org/10.1029/94TC00454.

King, P. R., and G. P. Thrasher, 1996, Cretaceous-Cenozoic geology and petroleum systems of the Taranaki Basin, New Zealand: Institute of Geological and Nuclear Sciences, New Zealand Monograph 13, Ministry of Economic Development Petroleum Report 3224.

Mattos, N. H., T. M. Alves, and A. Scully, 2019, Structural and depositional controls on Plio-Pleistocene submarine channel geometry (Taranaki Basin, New Zealand): Basin Research, **31**, no. 1, 136–154, https://doi.org/10.1111/bre.12312.

Neall, V. E, R. B. Stewart, and I. E. M. Smith, 1986, History and petrology of the Taranaki volcanoes, *in* I. E. M. Smith, ed., Late Cenozoic volcanism in New Zealand: Royal Society of New Zealand Bulletin 23, 251–264.

Nicol, A., J. Walsh, K. Berryman, and S. Nodder, 2005, Growth of a normal fault by the accumulation of slip over millions of years: Journal of Structural Geology, **27**, no. 2, 327–342, https://doi.org/10.1016/j.jsg.2004.09.002.

O'Neill, S. R., S. J. Jones, P. J. J. Kamp, R. E. Swarbrick, and J. G. Gluyas, 2018, Pore pressure and reservoir quality evolution in the deep Taranaki Basin, New Zealand: Marine and Petroleum Geology, **98**, 815–835, https://doi.org/10.1016/j.marpetgeo.2018.08.038.

Stagpoole, V., and R. Funnell, 2001, Arc magmatism and hydrocarbon generation in the Northern Taranaki Basin, New Zealand: Petroleum Geoscience, **7**, no. 3, 255–267, https://doi.org/10.1144/petgeo.7.3.255.

Stagpoole, V., and A. Nicol, 2008, Regional structure and kinematic history of a large subduction back thrust: Taranaki Fault: Journal of Geophysical Research: Solid Earth, **113**, no. B1, https://doi.org/10.1029/2007JB005170.