Essential Data Science for Subsurface Geoscientists and Engineers (G065)

Tutor(s)

David Psaila: Director of Data Science for the Digital Subsurface, Analytic Signal Limited.

Overview

Interest in data science and machine learning is rapidly expanding, offering the promise of increased efficiency in E&P, and holding the potential to analyse and extract value from vast amounts of under-utilised legacy data. Combined with petroleum geoscience and engineering domain knowledge, the key elements underlying the successful application of the technology are: data, code, and algorithms. This course builds on public datasets, code examples written in Python, statistical graphics, and algorithms from popular data science packages to provide a practical introduction to the subject and its application in the E&P domain.

Duration and Logistics

Classroom version: 5 days consisting of lectures and computer-based exercises and practicals.

Virtual version: Ten, 3-hour online sessions presented over 5 days. The course is at an introductory level and all subject matter will be taught from scratch. No prior experience of statistics, Python coding or machine learning is required, although some basic college level knowledge of maths and statistics is useful. Hands-on computer workshops form a significant part of this course, and participants must come equipped with a laptop computer running Windows (8, 10, 11) or MacOS (10.10 or above) with sufficient free storage (4 Gb). Detailed installation instructions are provided in advance so that participants can set up their computer with the data science toolkit and course materials before the course starts.

Level and Audience

Fundamental. This is an introductory course for reservoir geologists, reservoir geophysicists, reservoir engineers, data management, and technical staff who want to learn the key concepts of data science.

Objectives

You will learn to:

1. Analyse project data using the data science toolkit; notebooks, visualization, and communication.
2. Perform data import and manipulation, data visualization, exploratory data analysis, and building predictive models from data.
3. Have a working knowledge of coding in Python.
4. Coordinate reference systems including geographic and projected coordinate systems.
5. Use the fundamentals of machine learning including background concepts, the different types of machine learning, and the basic workflow to build and evaluate models from data.

Course Content

Course Details

The course comprises a mix of lectures and hands-on computer workshops. You'll gain a working knowledge of coding in Python. You'll learn the tradecraft of data import and manipulation, data visualization, exploratory data analysis, and building predictive models from data. You'll also gain a powerful working environment for data science on your own computer, which together with code examples provided by the course will give you a jump start to applying the techniques you'll learn to your own projects. For a flavour of what you'll learn, check out this gallery of visualization samples https://www.analyticsignal.com/visualization/index.html drawn from the course workshops.

What data sources are used?

Using real E&P data sources is an important element of the hands-on computer workshops. This course makes extensive use of open data provided the UK Oil and Gas Authority and the UK National Data Repository. These data sources are not only typical of the challenges and complexity presented by E&P datasets, but also contain sufficient data quality issues to make them ideal for teaching the all important skills of data cleaning and manipulation. The course makes use of well logs, tops, seismic, and production data from these sources. The data are released in the public domain and you can continue to use these sources as you gain in experience after the course.

What data science tools are used?

The course introduces a data science toolkit based on Visual Studio Code from Microsoft. This free product is rapidly growing in popularity as an environment for Python coding and data science. We think this toolkit provides a best-in-class environment for learning data science and subsequently moving to work on real projects, and we provide a free extension to further enhance its data science capabilities. The toolkit components will be installed on your computer – the advantage of this approach over cloud-based platforms is that your data is never uploaded to the cloud (if security is an issue), and you will be able to continue working when offline (if internet access is an issue).

Day 1

Module 1. Overview

- What is Data Science – Overview of the course, and an outline of the scope of data science.
- Data Science for E&P – Addressing the role of data science in E&P and an example application to log data quality control and reconstruction using machine learning.

Module 2. Data Science Toolkit – Notebooks, Visualization, and Communication

- Overview of the data science toolkit.
- Hands-on workshop introducing the toolkit and getting started with Python scripts and notebooks.
- Overview of how to manage and use Python packages.
- Hands-on workshop on Python packages covering how to install and manage packages, and how to use packages from your Python notebooks.
- Introduction to data visualization with SandDance.
- Hands-on workshop introducing SandDance for interactive data visualization using a dataset of offshore wells from the UK Continental Shelf.
- Overview of Markdown, a lightweight markup language for adding simple formatting to plain text documents, and documenting Python notebooks.
- Hands-on workshop on Markdown for formatting text documents and annotating Python notebooks.

Day 2

Module 3. Python Fundamentals

- Python 101 – Introduction to Python fundamentals including variables, types, statements, expressions, control flow, and functions.
- Hands-on workshop on Python 101.
- Python 102 – More Python fundamentals including modules, files and folders, data structures, and data frames.
- Hands-on workshop on Python 102.

Day 3

Module 4. Computational Thinking

- Introduction to Computational Thinking – the analytical and logical processes of decomposing a complex task and expressing it in a form that can be performed by a computer.
- Hands-on workshop on Computational Thinking applied to the design and implementation an interactive base map for UK E&P data.

Module 5. Exploratory Data Analysis

- Exploratory Data Analysis – Introduction to the Exploratory Data Analysis process and key Python packages for data analysis and statistical graphics.
- Hands-on workshop on exploratory data analysis of daily production data from the Vulcan gas field in the UK Southern North Sea – reading data, handling dates, cleaning values, resampling, merging datasets, creating statistical graphics, exporting results.
- Statistical Graphics – Why visualization is so important. Introduction to the Plotly package for statistical graphics. A classification of statistical graphics. Demonstration of a gallery of statistical graphics samples.
- Hands-on workshop on statistical graphics – using the Plotly Express package to create a gallery of statistical graphics samples. Code snippets (small blocks of reusable code) help make exploratory data analysis more fun by accelerating the journey from raw data files to working graphics.
- Descriptive Statistics – Introduction to univariate and multivariate statistics.

Day 4

Module 6. Exploring E&P Data

- Well header data – Introduction to handling well header data (surface location and attributes) using the pandas and plotly packages.
- Hands-on workshop on well header data – including import, data cleaning, date handling, posting well data on cultural/satellite base map and visualizing historical trends.
- Production data – Introduction to handling field production data using the pandas and plotly packages.
- Hands-on workshop on field production data – including import, data cleaning, date handling, queries, visualizing hierarchical and time series data.
- Well log data – Introduction to handling wireline logs from LAS files using the lasio, pandas, and plotly packages.
- Hands-on workshop on well log and tops data – including LAS file import, merging tops, and data visualization.
- Seismic data – Introduction to handling seismic SEG-Y data using the segyio, and plotly packages.
- Hands-on workshop on seismic data – including SEG-Y file import, extracting binary and trace headers, visualizing seismic trace data, and calculating seismic attributes.

Day 5

Module 7. Geospatial Data

- Coordinate reference systems – Introduction to geographic and projected coordinate systems, defining a coordinate reference system from EPSG codes, offsets between coordinate reference systems, and transforming positions between reference systems.
- Hands-on workshop on coordinate reference systems – how to define a coordinate reference system and transform positions using the pyproj package.

Module 7. Geospatial Data

- Machine Learning – introduction to the fundamentals of machine learning including background concepts, the different types of machine learning, and the basic workflow to build and evaluate models from data.
- Supervised learning with regression – introduction to regression including random forest regression and performance evaluation.
- Hands-on workshop on regression for reconstructing wireline logs.
- Unsupervised Learning – introduction to unsupervised learning for dimensionality reduction, clustering and outlier detection.
- Hands-on workshop on dimensionality reduction for wireline logs.
- Explainable Machine Learning – introduction to explainable machine learning: techniques for looking inside the so-called black box models of machine learning to understand why particular predictions are made and which variables are important.