课程大纲
COURSE SYLLABUS

1.	**课程代码/名称** **Course Code/Title**	Machine Learning in Geosciences
2.	**课程性质** **Compulsory/Elective**	Elective
3.	**课程学分/学时** **Course Credit/Hours**	4 credits / 64 hours
4.	**授课语言** **Teaching Language**	English
5.	**授课教师** **Instructor(s)**	Arnaud MIGNAN，风险研究院&地球与空间科学系 邮箱：mignana@sustech.edu.cn 电话：0755-XXX 办公室：创园 6 栋 512 Prof. Arnaud MIGNAN, Institute of Risk Analysis, Prediction and Management, Academy for Advanced interdisciplinary Studies; Department of Earth and Space Sciences Email: mignana@sustech.edu.cn Office: 511-3, 5th floor, building 6, Innovation Park
6.	**是否面向本科生开放** **Open to undergraduates or not**	No
7.	**先修要求** **Pre-requisites**	（如面向本科生开放，请注明区分内容。 If the course is open to undergraduates, please indicate the difference.） Probability & Statistics, Introduction to Computer Programming
8.	**教学目标** **Course Objectives**	
	This course will introduce the principles of machine learning and describe important machine learning applications in geosciences. We will explore the main methods of machine learning for regression and classification, and while supervised (deep) learning will be at the core of the course, basics of unsupervised learning and reinforcement learning will also be covered. The main notions will be exemplified with textbook examples from the computer science field and applied to various geoscience domains such as seismology, remote sensing, and geo-hazard assessment (among others). Upon completing the course, students will master the following items: 1. Fundamental knowledge of machine learning theory in a probabilistic perspective; 2. Basic knowledge of Python and R functions for machine learning (Tensorflow, Keras, tensor operations, fitting); 3. Basic knowledge on the history of machine learning and of its main applications in geosciences; 4. General knowledge on available methods (Bayesian inference, neural networks, decision trees, SVM, mixture, *etc.*); 5. Practical knowledge on which machine learnings methods to apply in different geo-data contexts; 6. Critical thinking on pros-and-cons of machine learning.	
9.	**教学方法** **Teaching Methods**	
	The course is a combination of lectures and programming exercises on jupyter-type notebooks. The exercises will consist in applying ready-made codes (in Python or R depending on the methods considered) to investigate how the theoretical aspects of machine learning translate into concrete applications in geosciences. The students will run	

existing programs and interpret the results, filter geo-data and engineer features for model development, modify parameterisations for sensitivity analyses. Such exercises will be incorporated to most lectures for clear understanding of the concepts and data covered & for promoting student interactions/participation in classroom (2/3 lecture, 1/3 exercises per chapter on average).

10. 教学内容
Course Contents

Chapter 1	**History of Machine Learning** (4 hours) *Week 1 (lecture 1):* **History of machine learning in computer science**: Early days, milestones, current state-of-the art, the main categories of machine learning methods. Introduction to platforms to be used in class, such as TensorFlow, Keras, Jupiter notebooks, *etc.* with syllabus description [Install R, Python and dependencies; Run a simple neural network on the famous MNIST computer vision dataset] *Week 1 (lecture 2):* **History of machine learning in geosciences**: Developments from the 1980s to nowadays in domains such as seismology, remote sensing, tomography, rock mechanics, geodesy, natural hazard assessment and early warning systems [Explore available machine learning geoscience code projects on GitHub repositories in preparation to future exercises; start getting familiar with common computational platforms used in those repos, such as TensorFlow, Keras, other R/Python packages, as well as with common interface platforms such as TensorBoard and Jupyter notebooks].
Chapter 2	**Basics of Probability Theory & Maximum Likelihood Estimation** (10 hours) All the machine learning techniques to be explored in this course derive from simple probabilistic concepts. Those include the basic rules of probability theory from which the main probability distributions can be demonstrated. Maximum likelihood estimation (MLE) and Bayes Theorem then provide the principal tools needed to fit a model to some data. We will generate some simple data representative of geophysical processes and use available low-dimensional tabulated geo-data such as earthquake catalogues and results from rock laboratory experiments to illustrate all concepts. *Week 2 (lecture 3):* **Probability theory** axioms, derivation of probability distributions (Binomial, Poisson, Normal, *etc.*), stochastic methods [how to call probability distributions in R/Python and how to create a geo-data sample from any given distribution] *Week 2 (lecture 4):* **Bayes Theorem**, likelihood function, prior and posterior distributions [Run a Bayesian inference R code that estimates seismicity parameters during an underground reservoir stimulation, discover how probability distributions evolve as more data come in] *Week 3 (lecture 5):* **Basics of regression**: Linear regression, non-linear regression, loss functions & regularisation (Lasso and Ridge), training, validation and test set definition, performance metrics [Fit a simple 1-dimensional geo-data set with regression tools and investigate underfitting versus overfitting; possible dataset: number of acoustic emissions in rock sample as a function of time] *Week 3 (lecture 6):* **Basics of classification**: Logistic regression, performance metrics derived from the confusion matrix, application examples in geosciences [Develop a logistic regression model to classify geographic cells as aftershock/no-aftershock on post-mainshock data, find proper features, calculate performance metrics] *Week 4 (lecture 7):* **Gradient descent fundamentals**: Parameter space exploration, beware of local minima, analytical solutions, general algorithms,

	learning rate & other hyperparameters, extensions to neural networks (to be reinvestigated in chapter 4) [Explore/plot gradient descent process on simple equations & on the lecture 5 example]
Chapter 3	**Support Vector Machines (SVM) & Decision Trees** (8 hours) Classical machine learning methods to explore high-dimensional data include Support Vector Machines (SVM) and decision trees. Decision trees include Random Forest and other ensemble algorithms such as boosting. Feature engineering is an important part of the process, which will be illustrated using the results of the Kaggle competition "Los Alamos National Laboratory Earthquake Prediction, Can you predict upcoming laboratory earthquakes?". *Week 4 (lecture 8):* **Basics of SVM and decision trees** (CART, random forest, AdaBoost, etc.), the curse of dimensionality, textbook application in computer science and recent applications in geosciences [Understand SVM by fitting the maximum margin hyperplane, and decision trees by calculating the Gini index, using a simple 2-dimensional binary ad-hoc dataset]. *Week 5 (lecture 9):* **Illustration of the SVM method on a geo-data set**, kernel definition (linear, polynomial, radial basis), regularisation [Earthquake/blast discrimination case, or moonquake/asteroid impact discrimination case TBC] *Week 5 (lecture 10):* **Introduction to a famous Kaggle competition**, the "Los Alamos National Laboratory Earthquake Prediction, Can you predict upcoming laboratory earthquakes?" experiment, feature engineering of baseline random forest model [Download the winning code and run it] *Week 6 (lecture 11):* **Going deeper into the Kaggle competition** "Los Alamos National Laboratory Earthquake Prediction, Can you predict upcoming laboratory earthquakes?" experiment, bootstrapping versus pruning, comparison of baseline model and winning model [Tune parameters and compare alternative results]
Chapter 4	**(Deep) Neural Networks** (8 hours) An alternative to the previous models for high-dimensional problems is the neural network, that is the combination of multiple layers of logistic regression models. The principles are here explained, such as tensorial operations, backpropagation, and the gradient descent possible extensions (expanding from chapter 2). They are then illustrated on three datasets: two very simple sets representing a curve and a spiral for basic understanding of universal function approximation, and the aftershock dataset already analysed on week 3 for a concrete application of a deep feedforward neural network. *Week 6 (lecture 12):* **Basics of feed-forward neural networks**, types of activation functions, what is a deep neural network, introduction to other neural network architectures such as RNN, long-short term memory (LSTM) & convolutional neural networks (RNNs & CNNs will be the subject matters of chapters 5-6) [Understand how a neural network works by simply fitting a curve (acoustic emission dataset); visualise the flexibility of a neural network model with Keras on the textbook "spiral" dataset] *Week 7 (lecture 13):* **Examples of applications in geosciences**, the black-box problem, the recent Nature debate on use of deep learning for aftershock prediction [Run a deep neural network on the aftershock dataset of week 3 using TensorFlow/Keras, compare with the previous logistic regression results] *Week 7 (lecture 14):* **Back-propagation algorithm**, tensorial operations, weight analysis, loss function derivative [Run a back-propagation code in Python and

	compare with the equivalent TensorFlow command, again on the spiral dataset] *Week 8 (lecture 15):* **Going deep into hyperparameter selection**, learning rate, epochs, activation functions, gradient descent parameters [Explore the role of different parameterisations on convergence time and final performance for the aftershock classification problem of lecture 13]
Chapter 5	**Sequential data analysis** (8 hours) This chapter covers neural network architectures used to model sequential data. We will consider the classic recurrent neural network (RNN), the long-short term memory (LSTM) neural network and finally the recent Transformer neural network. We will test those different algorithms on some seismic waveform data. This chapter will use many of the concepts already taught in the previous chapter. *Week 8 (lecture 16):* **Basics of RNN & LSTM**, basics of time series analysis, RNN architecture, input/output/forget gates, comparison of speech recognition & earthquake data [Visualise how a RNN finds patterns from unstructured data on a simple waveform time series] *Week 9 (lecture 17):* **RNN-LSTM application**, basics of earthquake phase picking and phase classification, exploring a published RNN-LSTM Python code for seismic waveform phase picking [Run the RNN-LSTM Python code, and investigate the role of parameters on the final result] *Week 9 (lecture 18):* **Basics of Transformers**, Transformer as a recent alternative to LSTM, encoder/decoder, attention mechanism, basics of natural language processing (NLP) [Investigate the famous BERT Transformer & its application to geo-text processing] *Week 10 (lecture 19):* **Transformer application**, exploring a published Transformer code for simultaneous earthquake detection and phase picking [Run the Earthquake transformer, and investigate the role of parameters on the final result]
Chapter 6	**Computer Vision** (10 hours) Computer vision consists in finding features in complex images. The convolutional neural network (CNN) is one of the most successful neural network architectures for such a task. Two main applications to unstructured geo-data will be considered: (1) a 1-dimensional case of seismic waveform data for earthquake discrimination, phase picking and phase classification (in continuation with the previous chapter) and (2) a 2-dimensional case of satellite-based landslide recognition. This chapter will use many of the concepts already taught in the previous two chapters. *Week 10 (lecture 20):* **Basics of Convolutional Neural Networks** (CNN), convolutional layers & operations, pooling, general applications in computer vision and speech recognition with famous architecture variants presented (with parallel made with chapter 5) [Run a simple CNN to the MNIST dataset and compare to the results obtained on week 1] *Week 11 (lecture 21):* **CNN usage in seismic waveform analysis**, with description of published CNNs such as PhaseLink, PhaseNet, *etc.*, comparing CNN phase picking and phase classification to results from other neural network architectures (with link to previous chapter) [Download some seismic waveform and the PhaseLink (or other, TBD) Python code]. *Week 11 (lecture 22):* **Digging deeper into the PhaseLink (or other, TBD) CNN code** and seismic waveform data [Run the CNN code, and visually inspect the

	obtained results] *Week 12 (lecture 23):* **Application of CNNs to satellite imagery**, object detection, presentation of the synthetic landslide image dataset [Run the landslide generation R code and observe how labelling is performed; Label one landslide on a Sentinel satellite image] *Week 12 (lecture 24):* **Digging deeper into the CNN for landslide image classification** [identify landslides on images with the provided CNN code based on a labelled synthetic landslide image dataset]
Chapter 7	**Basics of Unsupervised Learning** (8 hours) Application of the machine learning methods previously described becomes problematic when there is a lack of training data, which can easily be the case for a number of geo-data problems. Unsupervised learning consists in finding patterns in unlabelled data based on their characteristics. The main approaches are clustering and dimensionality reduction. Examples from geothermal data, micro-seismicity data and seismic tomography (TBC) will be shown. *Week 13 (lecture 25):* **Basics of clustering**, k-means, mixture modelling, Expectation-Maximisation (EM) technique [Understand how k-means and gaussian mixture model (GMM) algorithms work in R with the "Old Faithful" geyser eruption dataset] *Week 13 (lecture 26):* **Application to micro-seismicity analysis**, data mining for earthquake pattern recognition, concept of completeness magnitude, censored data and magnitude frequency distribution [Run an Asymmetric Laplace mixture model R code to fit magnitude frequency distributions of various shapes, observed or simulated; Observe how the EM algorithm works] *Week 14 (lecture 27):* **Other unsupervised methods** such as neural networks, self-organising maps (SOM), auto-encoders, generative adversarial networks (GAN) [application in earthquake/blast discrimination] *Week 14 (lecture 28):* **Basics of dimensionality reduction**, principal component analysis (PCA), linear discriminant analysis (LDA), dictionary learning [application in seismic tomography TBC]
Chapter 8	**Basics of Reinforcement Learning** (4 hours) Reinforcement learning is the third general type of machine learning method, with supervised learning and unsupervised learning. It is closely related to so-called artificial intelligence (AI) as it consists in an agent learning to reach a goal in an optimal way. Successfully applied to games (e.g. chess, Go, video games), it finds applications in optimisation problems and risk management. In geosciences, applications may include geo-energy optimization or natural disaster response strategies. *Week 15 (lecture 29):* **History of AI, basic concepts**, states, actions and rewards, Bellman equation, Q-learning [Run the Q-learning algorithm in R on simple gridded environments, e.g. Frozen Lake] *Week 15 (lecture 30):* **Applications to real-life situations**, trade-off optimization, presentation of a geo-energy risk optimization problem [Run the geo-energy risk environment, observe possible agent behaviours]
Chapter 9	**Next Horizons in Computational Geosciences** (4 hours)

		This concluding chapter qualitatively explores the future of machine learning in geosciences by reflecting on what we learned in the entire course, what is needed for proper machine learning applications (e.g. big data), and what could be expected in the coming years. *Week 16 (lecture 31):* **Dealing with Big Data**, trends in geosciences (satellite data, laboratory data), making more data with data augmentation. *Week 16 (lecture 32):* **Summary** of methods learned and review of studied geoscience applications. Question "will machine learning help make new discoveries in geosciences?" to be debated in class.

11. 课程考核
Course Assessment

（①考核形式 Form of examination；②.分数构成 grading policy；③如面向本科生开放，请注明区分内容。If the course is open to undergraduates, please indicate the difference.）

50% assignments: reports (jupyter notebooks) on each chapter's exercises: summary, method, results (with plots). Applies to chap. 2-8 – 50% on final examination: Basic concepts, model interpretation, ML applicability to geoscience know-how.

Letter grading

12. 教材及其它参考资料
Textbook and Supplementary Readings

教材 **Textbook**：

1. Murphy K.P. (2012), Machine Learning: A Probabilistic Perspective. The MIT Press

参考资料 **Supplementary Readings**：

1. Bergen K.J., Johnson P.A., de Hoop M.V., Beroza G.C. (2019), Machine learning for data-driven discovery in solid Earth geoscience. Science, 363, eaau0323

2. Kong Q., Trugman D.T., Ross Z.E., Bianco M.J., Meade B.J., Gerstoft P. (2019), Machine Learning in Seismology: Turning Data into Insights. Seismological Research Letters, 90 (1), doi: 10.1785/0220180259

3. Mignan A., Broccardo M. (2019), One neuron versus deep learning in aftershock prediction. Nature, 574, E1-E3, doi: 10.1038/s41586-019-1582-8