# Lab01a: Brainstorming, Textbook Selections, ML Challenges and Opportunities

**Handed out:** Monday, August 19, 2019

**Return date:** Wednesday, August 21, 2019, and Monday August 26 at the beginning of the class.

**Grades:** Lab01a counts 5 % towards your final grade.

**Objectives:** In this lab you will explore what machine learning is about and evaluate its challenges and Opportunities in the Geo- and Social Sciences

**Format of answer:** Your answers (statistical figures and verbal description) should be submitted as ***hardcopy***. Add a running title with the following information: Lab01, your name and page numbers. You may use this document as template. Copy the requested statistical figures into your document. Trial and error answers will lead to a deduction of points. Label each answer properly with the bold task and sub-task headings. You are expected to hand in professionally formatted answers: use a fixed pitch font, like **Courier New**, for any  code the use mathematical type-setting when equations are required. Copy and paste figures into your document. Make sure that each figure has a proper ***caption*** describing its content.

## Task 1. Key Concepts, Measures and Methods of ML [2 points]

Carefully browse through the three textbooks, which are listed below.

* Gareth, James, Daniela Witten, Trevor Hastie and Robert Tibshirani, 2017. *An Introduction to Statistical Learning with Applications in R*. 8th corrected printing. Springer Verlag

Available [online](https://utdallas.primo.exlibrisgroup.com/discovery/fulldisplay?docid=alma99230169001421&context=L&vid=01UT_DALLAS:UTDALMA&lang=en&search_scope=MyInst_and_CI&adaptor=Local%20Search%20Engine&tab=Everything&query=any,contains,An%20Introduction%20to%20Statistical%20Learning%20with%20Applications%20in%20R) at UTD’s library. A comprehensive and thorough introduction to most machine learning concepts and techniques with a rich set of examples. However, neural networks and deep learning are not included.

* Burkov, Andriy, 2019. *The Hundred-Page Machine Learning Book*. Self-published.

Available complementary [online](http://themlbook.com/wiki/doku.php) at the author’s website and at [Amazon](The%20Hundred-Page%20Machine%20Learning%20Book) for ~$ 36. The underlying ideas of all relevant techniques are succinctly introduced. Code examples are not given but references are made to Phyton’s SciKit-learn.

* Bruce, Peter and Andrew Bruce, 2017. *Practical Statistics for Data Scientists*. O’Reilly

Available [online](https://utdallas.primo.exlibrisgroup.com/discovery/fulldisplay?docid=alma9927695700101421&context=L&vid=01UT_DALLAS:UTDALMA&lang=en&search_scope=MyInst_and_CI&adaptor=Local%20Search%20Engine&tab=Everything&query=any,contains,Practical%20Statistics%20for%20Data%20Scientists) at UTD’s library. Good introduction to  for data scientists. This book introduces statistical concepts relevant for machine learning. In addition, it covers several exploratory machine learning techniques.

**Task 1.1:** Generate a well-organized and structured list of [a] topics, [b] concepts, [c] methods and [d] evaluative measures that you would like to see reasonably covered in this course.

Your list should be between 20-40 annotated keywords long.

This task is due Wednesday, August 21.

## Task 2: Textbook Selection [1 point]

**Task 2.1:** Rank the tree text books according to your preferences of using one of them as required book in this course. Justify your rank order in about 200 words.

Note, please do not be discouraged by the density of equations in some of the books. This course will guide you through the relevant equations.

This task is due Wednesday, August 21.

## Task 3: Critical Discussion of Articles [2 points]

Carefully study the two articles

* Jordan M. I. and T.M. Mitchell, 2015. Machine Learning: Trends, Perspectives, and Prospects. Science, 349:6245 pp 255-60
* Karpatne, Anuj & Ebert-Uphoff, Imme & Ravela, Sai & Babaie, Hassan & Kumar, Vipin. (2017). Machine Learning for the Geosciences: Challenges and Opportunities. IEEE Transactions on Knowledge and Data Engineering (in review). 10.1109/TKDE.2018.2861006.

**Task 3.1:** In about 200 words discuss what the challenges of ML in the geo-sciences are?

**Task 3.2:** In about 200 words discuss whether the same of different challenges apply to the social sciences.

**Task 3.3:** in about 100 word focus in particular on the availability of data in the social and geo-sciences. Is the situation similar or different?