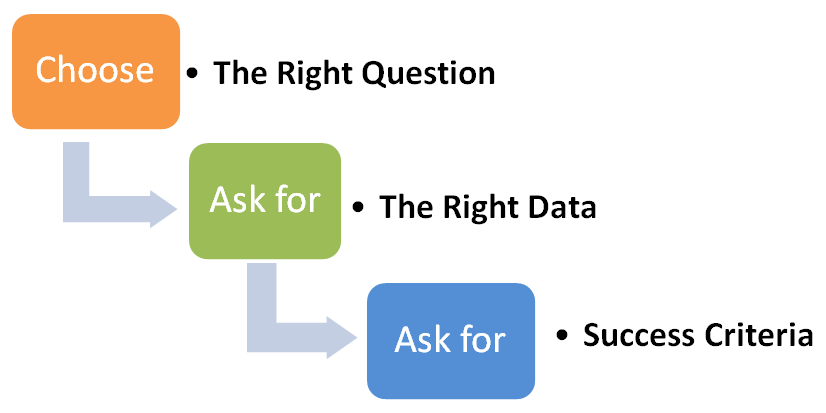
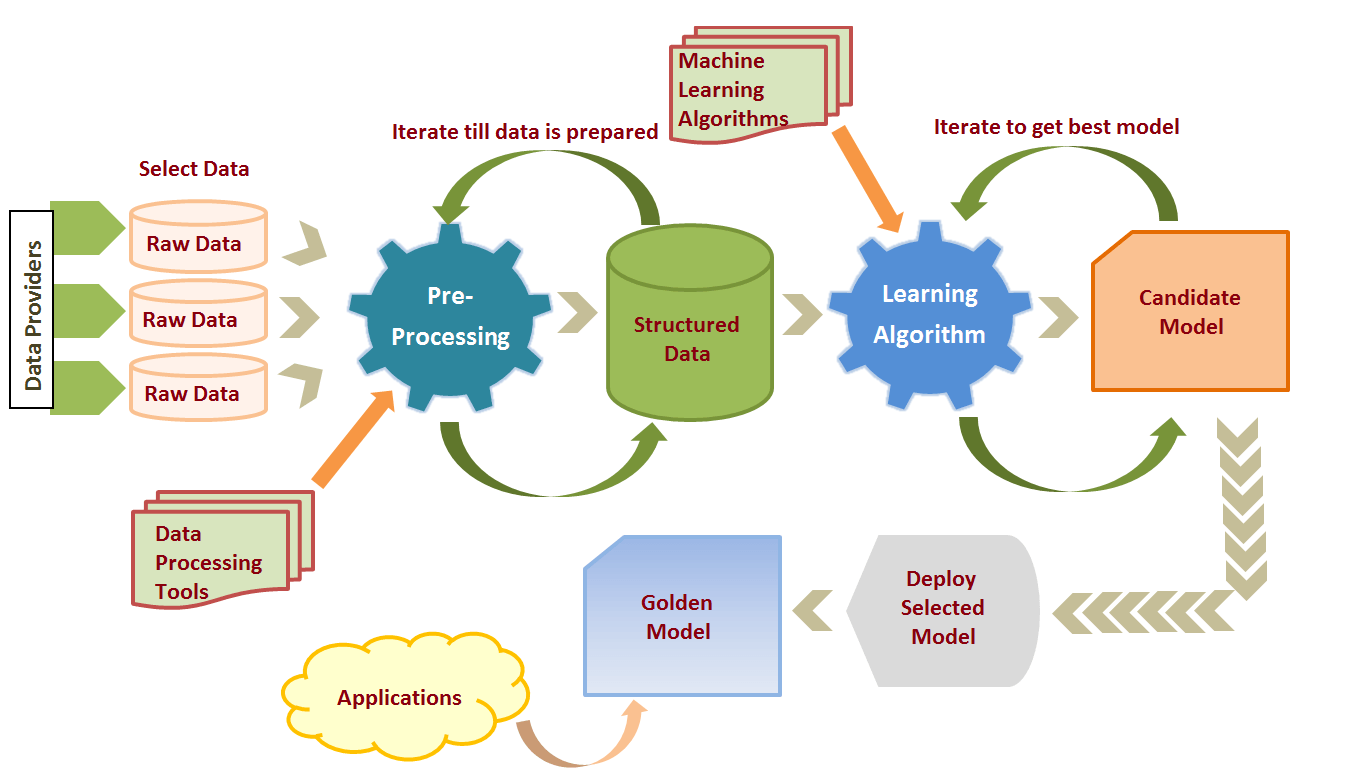
**Machine Learning Guidance**

PHBS SIQEF RA 2020

**1. BACKGROUND OF MACHINE LEARNING**

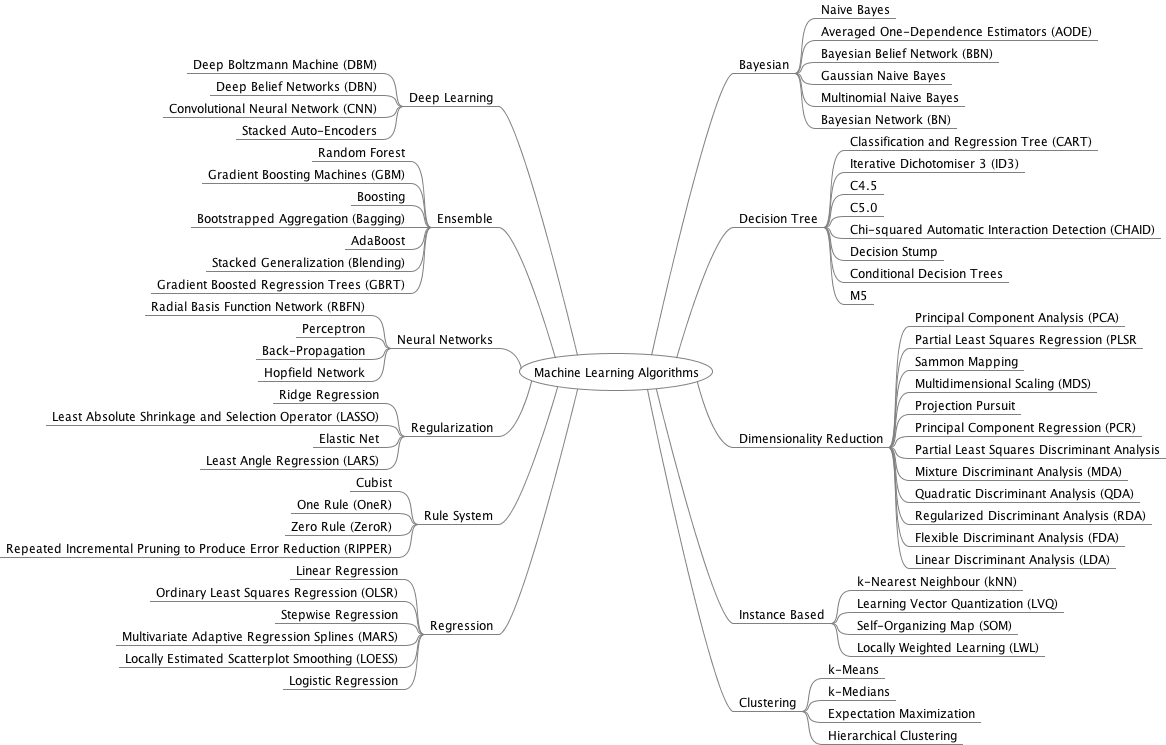
**1.1 Process**





**(From** [**https://elearningindustry.com/machine-learning-process-and-scenarios**](https://elearningindustry.com/machine-learning-process-and-scenarios)**)**

**1.2 Algorithms for model**



**(From** [**https://github.com/Bladefidz/machine-learning**](https://github.com/Bladefidz/machine-learning)**)**

**1.3 Common Characteristics**

- loss function central: reveal preference

- Prediction is usually fragile while theoretical model is robust

- ML advantage: decrease the reflexively chosen X variables and let the data tell

**1.4 Interest Topic**

- Prediction: Pivotal decision based on some sort of prediction

- People left to predict on their own: Behavioral economics

- Data at arm’s length: Build model via ML tools

- Decision aid: Can compare human decisions and machine predictions

- Enroll structure on investor sentiment feature in asset pricing

Many of the machine learning research flourished **in finance**, because

- Some part of finance is focusing on prediction.

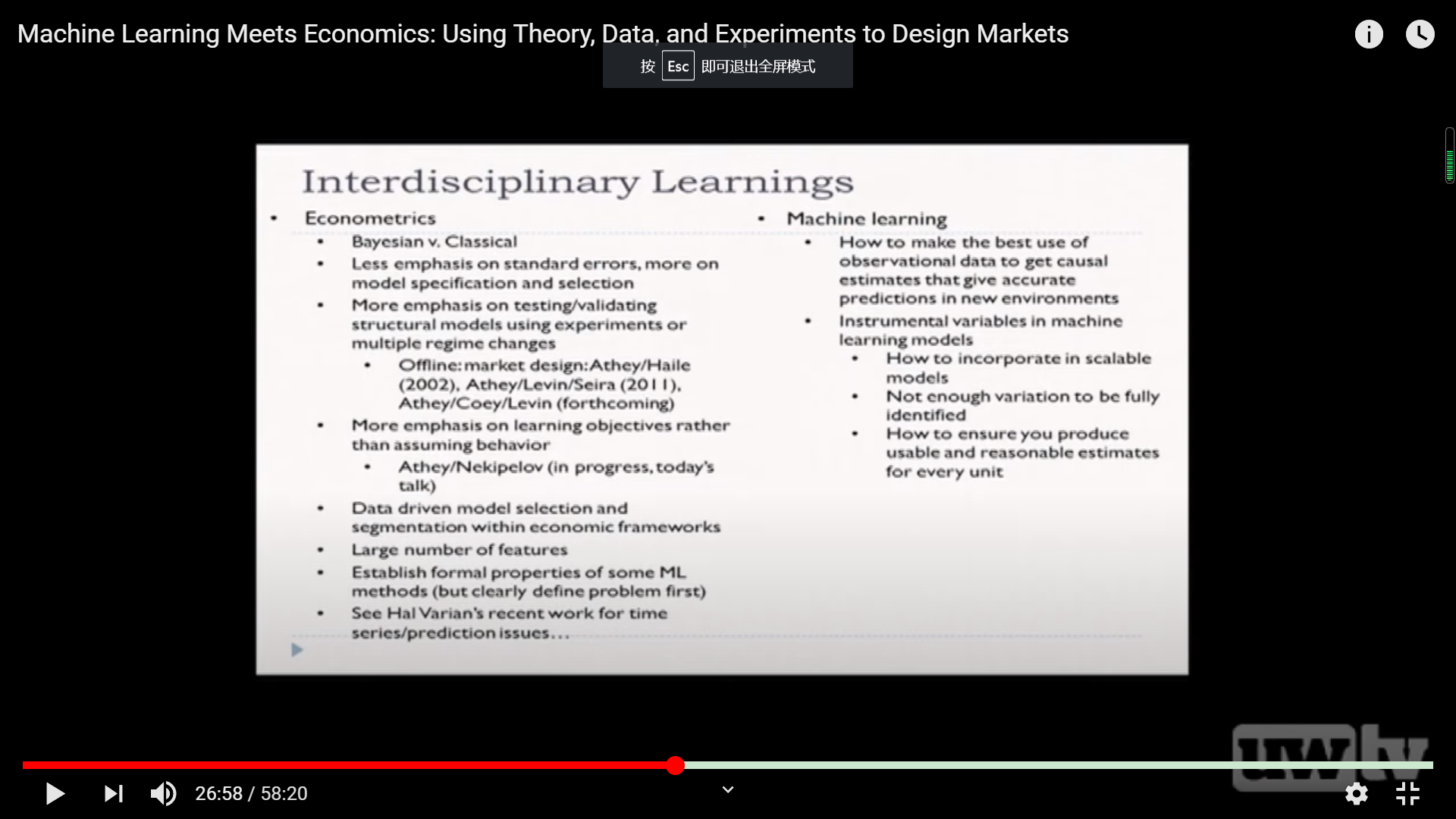
- You have out of sample thing you are testing on: in finance, it’s the next year data.

- Compare to coefficient estimate quality, prediction accuracy is more observable.

**(From** [**https://www.youtube.com/watch?v=xl3yQBhI6vY**](https://www.youtube.com/watch?v=xl3yQBhI6vY)

**AFA Lecture: Machine Learning and Prediction in Economics and Finance: Sendhil Mullainathan, Harvard University)**

**1.5 Machine Learning vs Econometrics**



**(From https://www.youtube.com/watch?v=eD758rKwQmA&t=1661s)**

**1.6 Overall guidance for coding**

- We will use Python as the main code language in class. Here are some guidance and advice for code novice:

<https://tim.blog/2019/03/21/learn-to-code/>

<https://learntocodewith.me/posts/code-for-free/>

<https://www.quora.com/What-is-the-best-method-of-self-learning-programming>

<https://lifehacker.com/top-10-ways-to-teach-yourself-to-code-1684250889>

**1.7** **Why you should learn machine learning:**

In general, this is what economics machine learning is about. To help to enhance products and services, improving productivity and predicting the future by giving trustworthy forecasts about economics, market, society, politics or technology. But for a change, these predictions actually CAN be trustworthy.

Current predictions are mostly based on what someone thinks, whether it’s a one-person or a company. It’s not a reliable source. Forecasts of the future will be based on big data. Machine learning algorithms will analyze the tenths of thousands of gigabytes of data in order to find the most probable outcome or trend. It will no longer be based on “reading tea leaves” so we might expect that its accuracy will be considerably higher. And as we mentioned earlier – a synergy of machine learning in economics and econometrics can lead to much more accurate models, combining the ability to analyze huge amounts of data and traditional modeling.

This class will provide you with some tools about machine learning. As machine learning and deep learning is the core algorithms in the coming AI-century, it is always better to know some about it. If you are trying to find a job or some intention, you may get some work-related techniques in this course. Even for someone who has decided his or her own job hunting intension, the course will provide you some chances to make kinds of innovation and increment in your daily routine.

See some views from:

<https://idei.fr/sites/default/files/IDEI/documents/tnit/newsletter/issue_19.pdf>

<https://addepto.com/machine-learning-in-economics-how-is-it-used/>

<https://www.quora.com/Is-learning-machine-learning-important-for-an-economics-student>

<https://mybroadband.co.za/news/software/253419-south-african-youth-think-machines-will-take-their-jobs-but-dont-want-to-learn-to-code.html>

<http://www.softech-systems.com/emerging-trends-in-finance-and-capital-market/>

**2. CORE ALGORITHMS**

**2.1 Classification:**

Widely used in many areas in economics problems, especially in finance area.

For a large variety of input X features, use your ML classification model to return a 0-1(or more 0123…in other new method) y variables identify which type does the sample belongs to.

Question with the format of **“Whether or not”** in economics & finance can always be suitable in **classification** method.

* Identify whether a specific individual loan will default or not (or other risk management area)
* Whether President Trump will reelection
* Whether the AAPL stock will rise or fall tomorrow or after the release of its statements (similarly as the other investment area)
* Whether there exists the crowding out effects for government to establish fiscal policy in different situation shown as the feature set of economics indicates.

Although due to database limitation we now cannot fulfill the overall project listed above, but as the data cumulates with time passing by, it will become more and more likely to appear in top journals.

There are some commonly used classification methods:

**KNN：**K Nearest neighbor, first find K nearest neighbors of the point X, then figure out what class most of them pertain to, and finally classify X in that particular class. The value of K is important, too large or too small K will cause underfitting or overfitting problem. This method is simple and straightforward, however, it is intensive both in computation and storage. Sometimes we can use this method as a procedure of feature engineering.

**SVC(SVM/SVR):** Support vector classifier, try to find a separating line that maximizes the distance between the margin of different classed. For a non-linear decision boundary, we can change the line to a curve or hyperspace by changing the core function.

**Decision Tree:** take one feature of X into consideration each time. First divide the whose X into two parts according to one certain feature of X, and then repeat doing that for each parts until getting the max depth or finishing that classification. The depth of the tree is important, too large or too small depth will cause overfitting or underfitting problem.

**2.2 Regression**

No need to say too much. As you can see, it is the most common way economist conduct their empirical research and how econ/fin students get their degree. It can actually give back both directional and quantitative information. We all study econometrics, and for machine learning users, they adjust some of the kernel in the original method and make the model fit the data more closely (Lasso, Ridge).

* The predict model of Singapore house price.
* Whether going for high education enlarge the workers salary.

You may think there is some overlap from the previous classification. Yes, it does. For example, Logit/Probit and other regression model can also be used for classification. Those are tools which can be used flexibly depending on your need.

We PHBS has a master of finance(Fintech) program. Usually it’s a good idea for them to replace the traditional econometrical regression to ML one and acquire something new in their graduation thesis.

**2.3 Clustering**

This is not a traditional X-y analysis method. This time you have n data point X1,X2,…Xn with k different features, and your task is to identify some groups for this X. Note that in this case k is allowed to be bigger than n to increase your accuracy. The difference between clustering and classification is that, usually the latter one is supervised learning and we know the required group labels (Yes/No/…) and numbers (usually 2), but in clustering you need to first get the different groups and then figure out what is the meaning of the group.

Just provide a very simple example here for you to understand. If we collect the personal information and characteristics in PHBS students and conduct a clustering, it is very likely that the TA in this class will be classified together. Cause we’re all RA in SIQEF and try to get a PhD in future, also showing some scholar temperament.

* Walmart divide their customers in different groups according to their info and buying data.
* Ping-An insurance company collect people’s data (gender, age, deposit, salary, etc.) to design and provide different products accordingly.
* Scout which stock accounts have similar behaviors to infer whether the account owner has insider trading
* Analyze the linkage of product prices in the macro economy
* Sector rotation and statistical arbitrage
* Government design different development strategies for different groups of city divided by their different locations, industry, GDP and other features.
* Divide people into different social stratum according to wealth, salary, education background and other features. (You can have a lot of social imbalance and wealth differentiation topics after that.)
* Default clustering risk management

**2.4 Dimensionality Reduction**

Mainly used as the first step before applying those above methods when you have so many X variables and cannot make use all of them for simplicity. Related to Feature Engineering method in CS. The most common way in ‘sklearn’ is to process PCA/LDA/FA to fit the data in order to combine different artificial index to avoid subject influence in this statistical caliber. You will get small number of meaningful main aspects from the huge feature X set.

* Transform a microeconomics signal as a sparse combination of Ricker wavelets.
* Factor analysis and PCA in stock and bond markets (return estimation)
* ICA in macroeconomics business cycle.

**3. COMMON METHOD**

**3.1 Model Selection**

Usually in a machine learning program, we conduct a series of ML method and get many models, which can be used for further prediction. To improve our accuracy of prediction, we need to build a method and a standard for testing and evaluating different types of model. Usually all the multiple ML method projects or papers will provide this step in the end.

**3.2 Preprocessing**

. A method to deal with unstructured data. Usually exist in every research. But it actually cost the majority of time in your project. You need to look deeper into the data and try to combine some of your own idea to preprocess the result.

**4. REFEREE**

**Paper**

**Type 0: Overall guidance**

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

<http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., ... & Layton, R. (2013). API design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*.

[**https://arxiv.org/pdf/1309.0238.pdf?source=post\_elevate\_sequence\_page---------------------------**](https://arxiv.org/pdf/1309.0238.pdf?source=post_elevate_sequence_page---------------------------)

Brownlee, J. (2017). How Much Training Data is Required for Machine Learning. *Machine Learning Mastery.[Online] Available: https://machinelearningmastery. com/much-training-data-requiredmachine-learning/[2018, May 25]*.

**https://machinelearningmastery.com/much-training-data-required-machine-learning/**

Athey, S. (2018). The impact of machine learning on economics. In *The economics of artificial intelligence: An agenda* (pp. 507-547). University of Chicago Press.

<https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/atheyimpactmlecon.pdf>

Abadie, A., & Kasy, M. (2019). Choosing among regularized estimators in empirical economics: The risk of machine learning. *Review of Economics and Statistics*, *101*(5), 743-762.

<https://www.mitpressjournals.org/doi/full/10.1162/rest_a_00812>

Korobilis, D. (2018). Machine learning macroeconometrics: A primer. *Available at SSRN 3246473*.

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3246473>

Coulombe, P. G., Leroux, M., Stevanovic, D., & Surprenant, S. (2019). *How is Machine Learning Useful for Macroeconomic Forecasting?* (No. 2019s-22). CIRANO.

<https://www.stevanovic.uqam.ca/GCLSS_ML_MacroFcst.pdf>

Some useful guidance from PHBS: Prof. Jaehyuk Choi and Prof. Xianhua Peng

<https://github.com/PHBS/RM-F1/blob/master/files/quant_topics.md>

**Type 1: Classification**

Moritz, B., & Zimmermann, T. (2016). Tree-based conditional portfolio sorts: The relation between past and future stock returns. *Available at SSRN 2740751*.

[**https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=2740751**](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2740751)

**(Use decision tree, combining with normal regression)**

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering, 160(1), 3-24.

[**http://www.informatica.si/index.php/informatica/article/viewFile/148/140**](http://www.informatica.si/index.php/informatica/article/viewFile/148/140)

Thornton, C., Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2013, August). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 847-855).

<https://arxiv.org/pdf/1208.3719.pdf>

Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1-15). Springer, Berlin, Heidelberg.

[**https://link.springer.com/chapter/10.1007/3-540-45014-9\_1**](https://link.springer.com/chapter/10.1007/3-540-45014-9_1)

**Type 2: Regression**

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481-85.

[**https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.p20151021**](https://pubs.aeaweb.org/doi/pdfplus/10.1257/aer.p20151021)

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

[**https://escholarship.org/content/qt35x3v9t4/qt35x3v9t4.pdf**](https://escholarship.org/content/qt35x3v9t4/qt35x3v9t4.pdf)

Xiang-rong, Z., Long-ying, H., & Zhi-sheng, W. (2010, November). Multiple kernel support vector regression for economic forecasting. In *2010 International Conference on Management Science & Engineering 17th Annual Conference Proceedings* (pp. 129-134). IEEE.

[**https://ieeexplore.ieee.org/abstract/document/5719795**](https://ieeexplore.ieee.org/abstract/document/5719795)

Chou, J. S., & Nguyen, T. K. (2018). Forward forecast of stock price using sliding-window metaheuristic-optimized machine-learning regression. *IEEE Transactions on Industrial Informatics*, *14*(7), 3132-3142.

[**https://ieeexplore.ieee.org/abstract/document/8263105**](https://ieeexplore.ieee.org/abstract/document/8263105)

**Type 3: Clustering**

Saâdaoui, F. (2012). A probabilistic clustering method for US interest rate analysis. *Quantitative Finance*, *12*(1), 135-148.

<https://www.tandfonline.com/doi/full/10.1080/14697681003591712?scroll=top&needAccess=true>

Das, N. (2003, August). hedge Fund classification using K-means clustering Method. In *9th International Conference on Computing in Economics and Finance* (pp. 11-13).

<https://www.researchgate.net/profile/Nandita_Das8/publication/24128136_Hedge_Fund_Classification_using_K-means_Clustering_Method/links/5956d0eba6fdcc2beca393d6/Hedge-Fund-Classification-using-K-means-Clustering-Method.pdf>

Marsili, M. (2002). Dissecting financial markets: sectors and states. *Quantitative Finance*, *2*(4), 297-302.

<https://www.tandfonline.com/doi/pdf/10.1088/1469-7688/2/4/305>

Durante, F., Pappadà, R., & Torelli, N. (2014). Clustering of financial time series in risky scenarios. *Advances in Data Analysis and Classification*, *8*(4), 359-376.

<https://link.springer.com/article/10.1007/s11634-013-0160-4>

León, C., Kim, G. Y., Martínez, C., & Lee, D. (2017). Equity markets’ clustering and the global financial crisis. *Quantitative Finance*, *17*(12), 1905-1922.

<https://www.tandfonline.com/doi/full/10.1080/14697688.2017.1357970>

**Type 4: Machine Learning regression for Econometrics**

* **Replace or upgrade the traditional empirical econometrics**

Charpentier, A., Flachaire, E., & Ly, A. (2018). Econometrics and machine learning. *Economie et Statistique*, *505*(1), 147-169.

<https://content.sciendo.com/view/journals/dim/1/2/article-p75.xml>

Kauffman, R. J., Kim, K., Lee, S. Y. T., Hoang, A. P., & Ren, J. (2017). Combining machine-based and econometrics methods for policy analytics insights. *Electronic Commerce Research and Applications*, *25*, 115-140.

<https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?referer=https://scholar.google.com.tw/&httpsredir=1&article=4731&context=sis_research>

Cornec, M. (2009). *Probability bounds for the cross-validation estimate in the context of the statistical learning theory and statistical models applied to economics and finance* (Doctoral dissertation).

<https://pastel.archives-ouvertes.fr/tel-00530876/document>

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

<https://www.tandfonline.com/doi/pdf/10.1080/01621459.2017.1319839>

Hansen, C., & Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized JIVE. *Journal of Econometrics*, *182*(2), 290-308.

<https://www.sciencedirect.com/science/article/pii/S0304407614000918>

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369-2429.

<https://users.nber.org/~dlchen/papers/Sparse_Models_and_Methods_for_Optimal_Instruments_ECTA.pdf>

Grimmer, J., Messing, S., & Westwood, S. J. (2017). Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Political Analysis*, *25*(4), 413-434.

<http://pages.shanti.virginia.edu/PolMeth/files/2013/07/GrimmerMessingWestwood.pdf>

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

<https://academic.oup.com/ectj/article/21/1/C1/5056401>

**Type 5: Machine Learning regression for microeconomics**

* It is obvious that machine learning provides some useful tools in Macro level, but not too much. Since Macro data is hard to investigate, hence not long enough to support this kind of data method. Traditional econometrics are still the main tools in dealing with data.
* But there is still a lot of potential in this topic. With the new form of data (Twitter, Google search result, etc) or some simulation/bootstrap techniques, you can still land some machine learning methods and push the frontier, which requires the command of knowledge and research experience in macroeconomics area.

Brathwaite, T., Vij, A., & Walker, J. L. (2017). Machine learning meets microeconomics: The case of decision trees and discrete choice. *arXiv preprint arXiv:1711.04826*.

<https://arxiv.org/pdf/1711.04826.pdf>

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124-27.

<https://academic.oup.com/ectj/article/21/1/C1/5056401>

**Type 6: Applied Advanced Machine Learning method**

Liu, S., Oosterlee, C. W., & Bohte, S. M. (2019). Pricing options and computing implied volatilities using neural networks. *Risks*, *7*(1), 16.

https://arxiv.org/pdf/1901.08943.pdf

**Relative Course and Project Links:**

**Examples From:**

<https://github.com/lazyprogrammer/machine_learning_examples>

<https://www.dezyre.com/article/top-10-machine-learning-projects-for-beginners/397>

You will find some interesting ideas and projects carried by **PHBS students** in

<https://github.com/PHBS/MLF/blob/master/Project.md> , and some other useful materials provided by Prof. Choi.

**Some other paper** from <http://econ-neural.net/>

**Official scikit-learn package** description

<https://github.com/scikit-learn/scikit-learn>

**PyTorch:** Deep Learning and Artificial Intelligence (special discount link for full VIP course as of Apr 2020)

<https://www.udemy.com/course/pytorch-deep-learning/?couponCode=PYTORCHVIP>

**Tensorflow 2.0:** Deep Learning and Artificial Intelligence (VIP Content)

<https://deeplearningcourses.com/c/deep-learning-tensorflow-2>

**Cutting-Edge AI: Deep Learning in Python**

[https://deeplearningcourses.com/c](https://deeplearningcourses.com/c/cutting-edge-artificial-intelligence)

**Data Quest URL** :[Sales\_Win\_Loss data](https://www.ibm.com/communities/analytics/watson-analytics-blog/sales-win-loss-sample-dataset/) set from IBM’s Watson repository

<https://www.dataquest.io/blog/sci-kit-learn-tutorial/>

A group of interesting case to follow, including **multiple machine learning** project:

<https://github.com/scikit-learn/scikit-learn/tree/master/examples>

**Forecasting ECB Yield curve**

<https://www.youtube.com/watch?v=nakmpAQ6z-g&t=152s>

you can download relative data from

<https://www.ecb.europa.eu/stats/financial_markets_and_interest_rates/euro_area_yield_curves/html/index.en.html>

An creative guidance for **auto-sklearn,** which you may get rid of some redundant work

<https://www.youtube.com/watch?v=uMWJls5Roqs>

**NTU ML foundation course** resource from Prof. [Hsuan-Tien Lin](https://www.csie.ntu.edu.tw/~htlin)

<https://github.com/LobbyBoy-Dray/NTU-Machine-Learning-Foundations>