# Machine Learning Basics for Informatics Professionals

Robert E. Hoyt MD, FACP, FAMIA, ABPM-CI
Virginia Commonwealth University, Richmond, VA

**What might the attendee be able to do after being in your session?**
Attendees will be able to load, prepare, explore, visualize and analyze two datasets using the data science platform RapidMiner. They will understand supervised learning (classification and regression) and unsupervised learning.

**Description of the Problem or Gap**
Machine learning is defined as the "*data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention."* [1] The goal in healthcare is to use machine learning and artificial intelligence to support the "Triple Aim": "*improving the experience of care, improving the health of populations, and reducing per capita costs of health care.*" [2] There is optimism that these new approaches will be major drivers of predictive analytics, as well as image, voice and text recognition. Recently, there is evidence that artificial intelligence (AI) has exceeded human diagnostic accuracy for image analysis in cardiology, [3] dermatology, [4] ophthalmology [5] and radiology. [6]

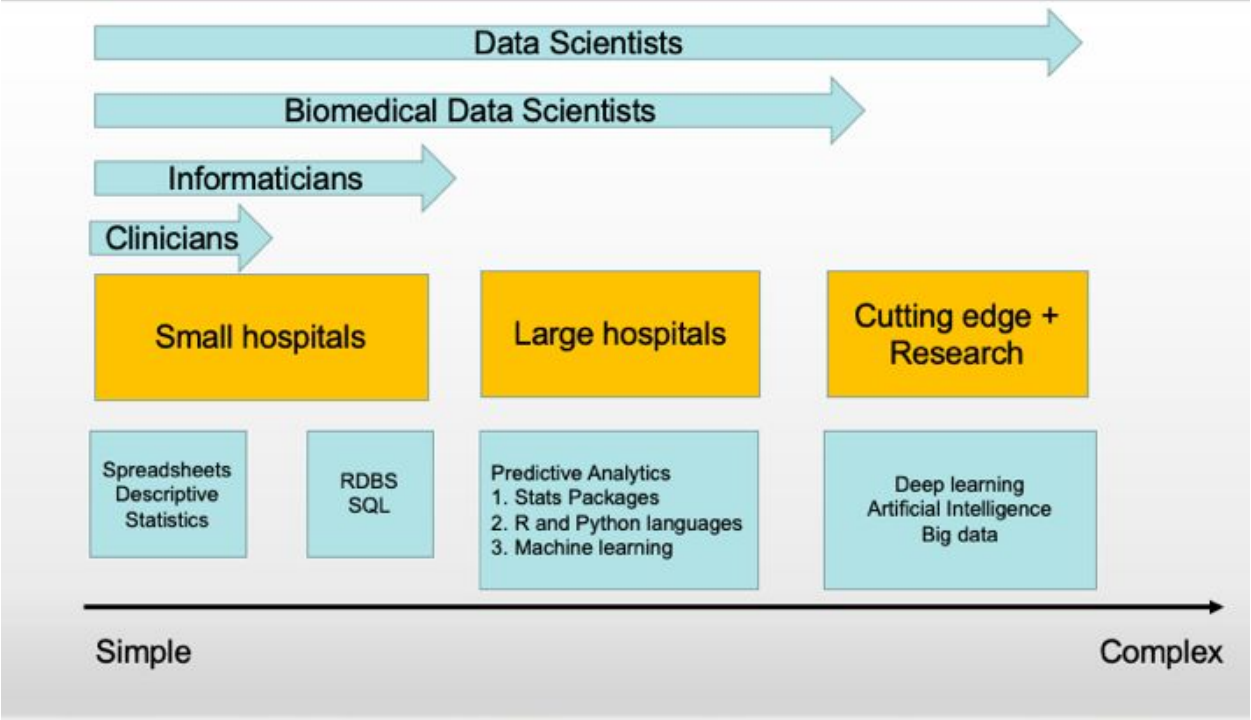
Why should clinical informaticists learn more about machine learning? First, machine learning is one of the most common methods used for predictive analytics and clinical decision support reported in the medical literature. Second, with the recent advances in artificial intelligence, it is imperative that clinicians have a working knowledge of machine learning, in order to understand artificial intelligence.

Is the medical field ready for machine learning and artificial intelligence? Do we have enough trained biomedical data scientists? While there has been a sudden rise in the number of biomedical and healthcare data science Master-level degree courses taught in the US, it will be a long time before we have enough biomedical data scientists to make an impact outside of a few large technologically advanced healthcare systems.

Although, some would argue that informaticists are data scientists, there are data to suggest otherwise, such as the requested skills by industry. According to a 2019 study by Meyer that explored healthcare data scientist job postings, the primary skills required were "statistics, R, machine learning, storytelling and Python." [7] To be an effective data scientist in the medical field, an individual must also understand higher math, in particular calculus and linear algebra. How many informatics professionals have these skills?

Most clinicians and clinical informaticists (physicians and nurses) rely on spreadsheet tools for data analysis, which is no longer adequate in the big data era. Only a limited number of medical universities and large healthcare systems are able to mine structured and unstructured data found in electronic health records. It is not realistic for most clinicians or clinical informaticists to

attain a Master's level degree in data science or to become proficient in either R or Python. Figure 1 is a diagram of the proposed current knowledge status of multiple individuals working in healthcare today.




What are the options informatics professionals have today if they want to become more knowledgeable of machine learning? They can achieve a Master's degree in data science or biomedical data science? They could enroll in multiple online data science-related courses. Alternately, they could study the R or Python programming languages. These choices are reasonable but not ideal for many.

There is another option to gain a comfort level with machine learning and that is to use machine learning software, which does not require programming skills or higher math. In the past decade, we have seen the rise of a variety of open source, as well as proprietary software programs that are free for academic use. Of the seven programs listed in the table below, one is graphical user interface (GUI)-based, two are web-based, two rely on visual operators and one is a hybrid.

| Name | Dependency | Uniqueness | Limitations |
|---|---|---|---|
| WEKA | Windows, Mac, Linux | GUI based. Associated with courses and textbook | Outdated appearance. |
| KNIME | Windows, Mac, Linux | Visual operators | Mild-moderate learning curve |
| Orange | Windows, Mac, Linux | Python based. Visual operators. Intuitive | Limited community forum |
| H2o ai | Web-based | Advanced | Mild-moderate learning curve |
| BigML | Web-based | Advanced | Mild-moderate learning curve |

| | | | |
|---|---|---|---|
| BlueSky Statistics | Windows | R based | Does not include neural networks. Windows only |
| RapidMiner | Windows, Mac, Linux | Visual operators and GUI based (hybrid). Automated analysis | None. "Best of breed"? |

**Methods: What did you do to address the problem or gap?**

Use a free data science platform that performs data preparation, exploration, visualization and analysis (RapidMiner) to help teach machine learning. [8]

RapidMiner was selected for this workshop for the following reasons:

- Excellent video tutorials, help sections and user community
- Free download for 30 days and indefinite free usage for academic purposes
- Excellent for basic to advanced analytics
- Program includes automated data preparation and data visualization (36 charts) (TurboPrep)
- Program includes automated algorithms (AutoModel) for supervised and unsupervised learning. Algorithms includes deep learning (neural networks)
- A menu of appropriate algorithms are automatically selected for classification and regression
- Algorithm performance is reported and compared: sensitivity, specificity, recall, precision, F1 score and AUCs

**The Workshop - first hour**

1. The workshop will begin with a 10-question assessment of current machine learning knowledge using MentiMeter polling software.
2. Students will download RapidMiner software as well as two datasets for analysis
    a. Heart risk prediction (classification)
    b. Medical charge prediction (regression)
3. An overview of biomedical data science, machine learning, deep learning and artificial intelligence will be presented
4. Participants will learn about supervised learning
    a. Classification
    b. Regression
5. Participants will learn about unsupervised learning
    a. Association
    b. Clustering
6. Participants will get an overview of RapidMiner, initially with PowerPoints
7. Break

**The Workshop - second hour**

1. Participants will navigate through the data preparation and exploration phases with TurboPrep and create multiple data visualizations on all variables prior to modeling

2. Participants will create a classification model based on heart disease prediction data using common classification algorithms and compare algorithm performance
3. Participants will create a regression model based on the medical charges prediction data using common regression algorithms and compare algorithm performance
4. The 10-question assessment will be repeated at the end and compared with the initial assessment

**What was the outcome(s) of what you did to address the problem or gap?**

Before and after quizzes will be conducted to document improvement post-workshop instruction

**Attendee's take away tool**

Access to an affordable data science platform (RapidMiner) will improve their knowledge of data science from data preparation to analysis with machine learning

**Use of knowledge acquired at previous AMIA Events**

I led a machine learning workshop for the AMIA iHealth conference in May 2017

**References**


1. SAS. Machine Learning. https://www.sas.com/en_us/insights/analytics/machine-learning.html
2. The Triple Aim http://www.ihi.org/resources/Pages/Publications/TripleAimCareHealthandCost.aspx
3. Zhang J, Gajjala S, Agrawal P et al.Fully Automated Echocardiogram Interpretation in Clinical Practice: Feasibility and Diagnostic Accuracy. Circulation.2018;138:1623–1635
4. Haenssler HA, Fink C, Schneiderbauer R. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists Annals of Oncology. 2018; 29(8): 1836–1842. https://doi.org/10.1093/annonc/mdy166
5. Abràmoff MD, Lavin PT, Birch M, et al. Shah, Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. 1, 39 (2018).https://www.nature.com/articles/s41746-018-0040-6 Accessed June 15, 2019
6. Hwang EJ, Park S, Jin KN, et al.Development and Validation of a Deep Learning–Based Automated Detection Algorithm for Major Thoracic Diseases on Chest Radiographs. JAMA Network Open. 2019;2(3):e191095.
7. Meyer MA. Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. JAMIA 2019;26(5):383-391
8. RapidMiner. https://www.rapidminer.com