Summer 2018

Perspectives and Best Practices for Artificial Intelligence and Continuously Learning Systems in Healthcare

Berkman Sahiner
US FDA

Bruce Friedman
GE

Cindi Linville
Best Sanitizers

Cindy Ipach
Compliance Insight

Edna Montgomery

See next page for additional authors

Recommended Citation

Perspectives and Best Practices for Artificial Intelligence and Continuously Learning Systems in Healthcare. Xavier Health's Artificial Intelligence Summit, August 23-24, 2018, at Xavier University, Cincinnati, OH, USA. https://www.xavierhealth.org/cls-working-team/

Authors
Berkman Sahiner, Bruce Friedman, Cindi Linville, Cindy Ipach, Edna Montgomery, Eileen Steinle Alexander, and et al.

Perspectives and Good Practices for AI and Continuously Learning Systems in Healthcare

August 2018

Xavier University | Xavier Health Organization | www.XavierHealth.org

TABLE OF CONTENTS

ACKNOWLEDGMENTS 3

1. INTRODUCTION 4

EXECUTIVE SUMMARY 4
GOALS OF THIS PAPER 6
ENABLING TECHNOLOGIES 7

2. VARIATIONS OF ML SYSTEMS 7

3. CONSIDERATIONS FOR CONTINUOUSLY LEARNING SYSTEMS 8

DATA SOURCES, FEEDBACK LOOP, QUALITY AND CONFIDENCE ASSESSMENT, AND LIMITATIONS 9
GENERAL SOFTWARE DESIGN 11
CONFIDENCE AND EXPLAINABILITY 12
INTEGRATING VALUES & ETHICS 13
PATIENT CONSENT 13
RETRAINING 14
RISK MANAGEMENT 14
CYBERSECURITY, AUTHENTICATION, PRIVACY, AND ANONYMIZATION 15

4. CLS LIFECYCLE 16

PLANNING 16
REQUIREMENTS 17
ARCHITECTURE 18
VERIFICATION AND VALIDATION 18
POST-MARKET CONSIDERATIONS 20
INSTALLATION 20
OPERATION & SUPPORT 20
MAINTENANCE & CHANGE MANAGEMENT 20
RETIREMENT 21

5. CONCLUSION 22

APPENDIX I – GLOSSARY 23

APPENDIX II – PROJECT BACKGROUND 25

APPENDIX III - CUSTOMER PERSPECTIVE 26

APPENDIX IV – COMMON CLS VALIDATION TECHNIQUES 28

APPENDIX V - EXAMPLES 31

Acknowledgments

This paper was developed under the leadership of the Xavier Health program at Xavier University in partnership with FDA officials and industry professionals, as a planned output from 2017 AI Summit. We would like to thank everyone who contributed to the creation and the review of this paper – without their work, this paper would not have been possible[1]. We also want to acknowledge the significant contributions from the following people:

Berkman Sahiner, FDA
Bruce Friedman, GE
Cindi Linville, Best Sanitizers
Cindy Ipach, Compliance Insight
Eda Montgomery, Shire
Eileen Alexander, Xavier University
Golnaz Moeini, Regulatory and Strategy Consultant
Hui Zhao, Penn State University
Jackie Haydock, Microsoft
John Daley, IBM
Juan Perez, Infosys
Krista Woodley, J&J
Lacey Harbour, Ken Block Consulting
Mohammed Wahab, Abbott
Mac McKeen, Boston Scientific
Nathan Leong, Microsoft
Pavan Kumar Garikapati, Infosys
Radhika Parameswaran, Infosys
Scott Thiel, Navigant
Sunny Jansen, Google
Walter Mullikin, Shire

Our hope is that this paper provides the foundation for new learnings and best practices in this rapidly evolving field to help deliver the promise and potential of AI and Continuously Learning Systems.

Pat Baird, Philips

Rohit Nayak, Sundance Capital

[1] Please note that the opinions and viewpoints expressed by the contributors do not necessarily reflect the opinions and viewpoints of their organizations.

1. Introduction

Executive Summary

In the past few years there has been a significant increase in the amount of information in healthcare, and this trend is accelerating. In addition to traditional data, such as those collected by in-vitro laboratory tests, imaging devices, physiological tests, and electronic health records, there are additional systems collecting, integrating and analyzing information and more ways for us to use it in our everyday lives. Wearable technologies are just the tip of the information iceberg as the Internet of Things (IoT) will continually monitor our lives and have real potential to give us insights into how to improve healthcare. Real Time Health Systems (RTHS) and Advanced Broad-Based Analytics (ABBA) enable better diagnosis and more effective treatment options for patients. However, because there is so much data and correlations can be very subtle, it can be difficult for clinicians to see them unaided. To take advantage of all this information, the future of healthcare needs Artificial Intelligence (AI).

Beyond the anticipated benefits of processing information in a way to help personalize the application, AI may well be required simply to use the anticipated vast amount o f information. A 2017 International Data Corporation (IDC) whitepaper assessed the growth of data relative to the ability to store the information[2]. Based upon their research, by 2025, the amount of data generated will exceed our ability to store the information. This means that certain data will need to be processed in real-time or near real-time, or be lost.

For the purposes of this paper, AI is shorthand for any system that can perform tasks that normally require human intelligence. It makes use of varied methods such as knowledge bases, expert systems, and machine learning. AI can sift through large amounts of raw data looking for patterns and connections much more efficiently, quickly, and reliably than a human could. Note that the performance of the AI system may be worse than if a human were performing the same task, but this may still be useful in managing our complex world.

AI is not one universal technology, rather it is an umbrella term that includes multiple technologies such as machine learning, deep learning, computer vision, neural networks, and natural language processing (NLP) that, individually or in combination, add intelligence to applications. Diagram 1 shows the relationship of these various technologies. AI and machine learning are often used interchangeably but they are not the same thing and the misperception can cause confusion. Both encompass many different models, approaches, and implementations.

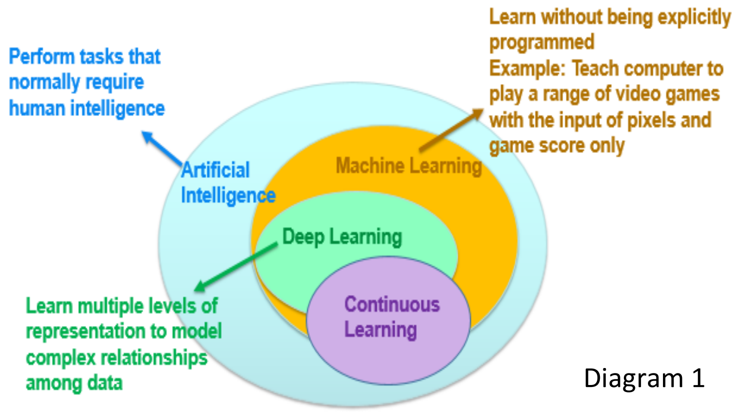

Perform tasks that normally require human intelligence
Learn without being explicitly programmed
Example: Teach computer to play a range of video games with the input of pixels and game score only
Artificial Intelligence
Machine Learning
Deep Learning
Continuous Learning
Learn multiple levels of representation to model complex relationships among data

Diagram 1

[2] "Data Age 2025: The Evolution of Data to Life-Critical", https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf

Machine learning systems may be trained using “supervised” or “unsupervised” techniques[3]. In supervised machine learning, the training data set is labeled such that every input has its corresponding label/target. During training, the system learns a function from inputs to their corresponding targets. In unsupervised machine learning, the data set only contains inputs, and the algorithm learns based on the statistical properties of the input data and groups the data into clusters with similar statistical properties. A variant, “reinforcement learning”, attempts to maximize a desired outcome based on its input data – essentially going through a process of trial and error until it arrives at the best possible outcome. Deep learning, a subset of machine learning, leverages neural network approaches to decompose a complex problem into multiple layers, with deeper layers refining the output from the previous ones, attempting to mimic the human brain structure.

After training, many machine learning systems may be “locked.” For a locked system, once the training has been satisfactorily completed, the system is put to use, and does not continue to learn. In contrast, in “continuously learning” systems, the algorithm keeps learning as humans do, and the output of the system for the same input data may be different before and after this learning has taken place. Typically, the learning process is pre-specified, with the goal of improving a well-defined metric.

Continuously Learning Systems (CLS) are built on the idea of learning continuously and adaptively about the external world and enabling the autonomous incremental development of ever more complex skills and knowledge. In the context of Machine Learning it means being able to smoothly update the prediction model to consider different tasks and data distributions but still being able to re-use and retain useful knowledge and skills during time[4]. Diagram 2 shows the process flow for a CLS application.

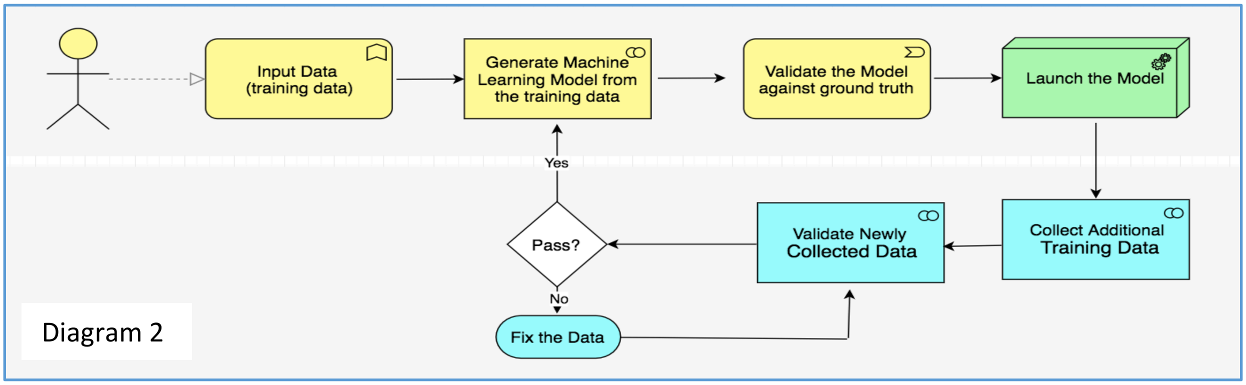

Input Data
(training data)
Generate Machine Learning Model from the training data
Validate the Model against ground truth
Launch the Model
Collect Additional Training Data
Validate Newly Collected Data
Pass?
Yes
No
Fix the Data

Diagram 2

Continuous learning, as a practice, can be applied to many of the machine learning modalities described above, ranging from automated retraining of expanding data sets to neural networks with massive amounts of unstructured, unclassified data, from real-time learning to fixed frequency retraining and learning. Correspondingly, CLS have additional considerations beyond those of traditional software development methods.

[3] Terminology will be one of the challenges for the successful development and use of AI in healthcare. For example, the AI community defines “supervised” learning means the output values are already known during training; however the wider healthcare community may interpret “supervised” and “unsupervised” as determining whether or not a human is involved in decision making – supervising the activities of the algorithm.

[4] https://medium.com/@vlomonaco/why-continuous-learning-is-the-key-towards-machine-intelligence-1851cb57c308

Goals of this paper

Healthcare is often a late adopter when it comes to new techniques and technologies; this works to our advantage in the development of this paper as we relied on lessons learned from CLS in other industries to help guide the content of this paper. Appendix V includes a number of example use cases of AI in Healthcare and other industries.

This paper focuses on identifying unique attributes, constraints and potential best practices towards what might represent “good” development for Continuously Learning Systems (CLS) AI systems with applications ranging from pharmaceutical applications for new drug development and research to AI enabled smart medical devices. It should be noted that although the emphasis of this paper is on CLS, some of these issues are common to all AI products in healthcare.

Additionally, there are certain topics that should be considered when developing CLS for healthcare, but they are outside of the scope of this paper. These topics will be briefly touched upon, but will not be explored in depth. Some examples include:

> **Human Factors** – this is a concern in the development of any product – what are the unique usability challenges that arise when collecting data and presenting the results? Previous efforts at generating automated alerts have often created problems (e.g. alert fatigue.)
>
> **CyberSecurity and Privacy** – holding a massive amount of patient data is an attractive target for hackers, what steps should be taken to protect data from misuse? How does the European Union’s General Data Protection Regulation (GDPR) impact the use of patient data?
>
> **Legal liability** – if a CLS system recommends action that is then reviewed and approved by a doctor, where does the liability lie if the patient is negatively affected?
>
> **Regulatory considerations** – medical devices are subject to regulatory oversight around the world; in fact, if a product is considered a medical device depends on what country you are in. AI provides an interesting challenge to traditional regulatory models. Additionally, some organizations like the FTC regulate non-medical devices.

This paper is not intended to be a standard, nor is this paper trying to advocate for one and only one method of developing, verifying, and validating CLS systems – this paper highlights best practices from other industries and suggests adaptation of those processes for healthcare. This paper is also not intended to evaluate existing or developing regulatory, legal, ethical, or social consequences of CLS systems. This is a rapidly evolving subject with many companies, and now some countries, establishing their own AI Principles or Code of Conduct which emphasize legal and ethical considerations including goals and principles of fairness, reliability and safety, transparency around how the results of these learning systems are explained to the people using those systems[5].

The intended audience of this paper are Developers, Researchers, Quality Assurance and Validation personnel, Business Managers and Regulators across both Medical Device and Pharmaceutical industries that would like to learn more about CLS best practices, and CLS practitioners wanting to learn more about medical device software development.

[5] https://blogs.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf

Enabling Technologies

AI systems need large amounts of data. Cloud services, edge computing, data warehousing, and data lakes all have a potential impact on the performance and availability of AI systems.

Organizations will take advantage of the benefits that the public cloud offers, including computing power and storage ability. However, there may be use cases where edge computing is also needed due to communication bandwidth not being sufficient to support the needs of the algorithm, connectivity reliability or when latency will impact the outcome. For example, if a CLS has been deployed to a patient's bedside monitor to alert when a specific vital event has occurred, communication latency might impact the outcome, so an edge device executing the latest version of a trained CLS model will be needed as part of environment.

The cloud may be accessible intermittently, which could cause unique problems. For example, if the cloud is intermittent while the CLS system is being trained, does this corrupt the training data set? Does this duplicate the training set? Does this leave out some data?

Data warehousing for ML/CLS is the accumulation of manually inputting or automatically streaming historical and real-time data into a database[6]. This information pipeline is critical to the strategy of a predictive applications' development and training. These days, a data lake is very rarely stored and maintained within one monolithic structured server. The systems architecture typically leverages a combination of local servers, virtual machines, and cloud services with various processing and extraction methods[7]. However, depending on the information architecture strategy, hardware limitations are still factors that must be considered during ML/CLS development and training. Also, for those in a highly regulated environment, the information pipeline architecture should be highly influenced on the criticality of the data, the ML/CLS risk, and intended use of the ML/CLS.

2. Variations of ML Systems

Most Machine Learning (ML) algorithms go through a period of training, cross validation, and testing before being placed in use (Diagram 3.) Like all statistical models, performance depends on how well the data set used for training is representative of the actual environment of use. During use, the ML algorithm collects additional data, which can be collated and used (offline) to repeat the original cycle of testing and validation. The original ML algorithm can then be replaced with the "new" algorithm with improved performance. This is sometimes called batch learning.

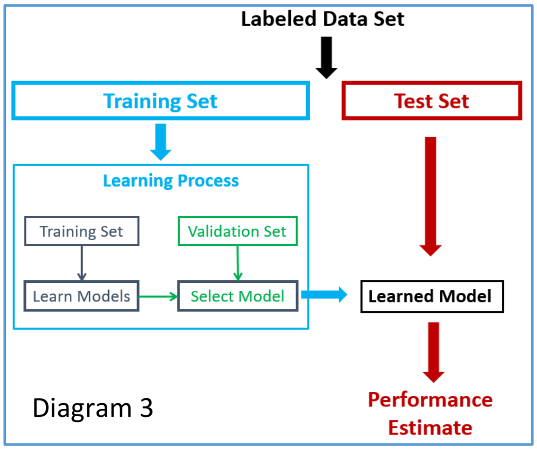


Diagram 3

[6] Luersen, Seth. MEMSQL Blogs. 2017, November 17. http://blog.memsql.com/memsql-maturity-framework/. Extracted July 2, 2018
[7] Orenstein, Gary; Doherty, Conor; Boyarski, Mike; and Boutin, Eric. Data Warehousing in the Age of Artificial Intelligence. O'Reilly Media, Inc. 2017

In contrast, CLS allows the ML algorithm to "learn in place" and incrementally update and improve its performance each time it acquires new data. The CLS algorithm continues to operate and rather than a step change in performance as described above the algorithm improves incrementally.

Some systems may use a hybrid approach where batch learning is used to establish the algorithm and initial assignment of values to the algorithm, and CLS is used to gather data an incrementally adjust the values in the algorithm e.g. the equation itself does not change, but the weighting of the variables change.

CLS is sometimes referred to as incremental learning. A description of the features of incremental learning is provided in an article by Karanam Supraja[8], and includes:

- Accommodate new information as and when available
- Ability to work with unlabeled data
- Ability to handle multidimensional data,
- Bounded complexity (e.g. amount of complexity in a problem is limited),
- Learn incrementally from empirical data, and
- Handle changes in concepts etc.

Another related concept is Adaptive Systems. Adaptive systems adjust themselves at runtime based on the learned data and generate different outputs every time new data is learned. Changes to the algorithm used in adaptive systems are implemented through a pre-specified and possibly fully automated process that is aimed at improving performance either based on availability of new training data or the based on continuing analysis of the effect the algorithm[9].

3. Considerations for Continuously Learning Systems

The growth of IoT and the resulting preponderance of data of all types has made application of Artificial Intelligence or Intelligent Technology (AI) tools the next big frontier. We're seeing applications that range from smart devices to digital therapeutics and everything in between. While the healthcare industry has seen several AI applications brought to market that leverage image and pattern recognition, natural language processing, prediction and decision making algorithms, we are only at the cusp of seeing truly advanced AI systems that leverage deep learning in conjunction with continuously learning approaches.

When developing a ML system, there are some unique considerations that don't typically apply in traditional health-related software development. For example, if the patient is interacting with an AI system, are they even aware that their "chat", diagnosis, or treatment recommendation is coming from an AI-based system?

[8] https://www.quora.com/What-are-some-real-world-applications-where-incremental-learning-of-machine-learning-algorithms-is-useful-Are-SVMs-preferred-for-such-applications

[9] https://pdfs.semanticscholar.org/71ac/d8e88f26cd8c8986d509ccfc8389f030fc39.pdf

There are several characteristics of a system that contribute to CLS efficacy and robustness. For example:

- Source of data – quality and quantity of data, including expected minimums, structure, and an understanding of the context of the data sets
- Fairness - should treat everyone in a fair and balanced manner and not affect similarly situated groups of people in different ways
- Number of variables, features, and layers being utilized in the model (e.g. Advanced Broad-Based Analytics)
- Frequency of training or retraining
- Known limitations and exceptions
- Established parameters of operations

These characteristics can vary depending on the type of system being developed. For example, a Decision Assistance Systems where the AI does not act on its own, but assists users in their decision making or for their actions would likely have different criteria and an Autonomous System where the AI acts or implements a decision without clinician or user intervention.

After the system has been designed, implemented, and trained, additional steps are needed to ensure its quality (both actual quality and perceived quality):

- Performance Evaluation – Overall system test methods and tools, typically involves a comparison of what the AI is supposed to do for a certain input versus what it does
- Building Confidence in AI – Includes processes that are used to provide confidence in AI operation in addition to or complementary to performance evaluation. It should be noted that a correct and robust application is useless if no one actually uses it. Building confidence in the AI system is an important success factor, and it includes:
    - Inclusiveness
    - Reliability
    - Usability
    - Clinical significance
    - Ease of integration with existing systems
    - Precision and accuracy

It is worth noting that any effort to standardize on a set of specific technologies, techniques, algorithms, models or toolsets is likely going to be obsolete in a short period of time given the rapid evolution within AI. This is particularly relevant because we are at a very nascent stage of the evolution of this technology and its applications.

Data Sources, Feedback Loop, Quality and Confidence Assessment, and Limitations

For machine learning, there are two basic data types: structured and unstructured data. Webopedia defines structured data as data that resides in in a fixed field within a record or file. Examples of this can be from relational databases and spreadsheets[10]. Webopedia defines unstructured data as information that doesn't reside in a traditional row-column database. Examples of this can be from diagnostic

[10] https://www.webopedia.com/TERM/S/structured_data.html. Extracted June 19, 2018

images, hand-written triage notes, e-mail documents, word processing documents, PDF, PPTs, videos, photos, audio files, blogs and more. In addition to structured and unstructured, there is semi-structured data[11]. Semi-structured data refers to data that is partially organized by tags and markers in a fashion that is accessible by ML analytic tools. Examples of this include XML documents, Word metadata files, e-mail sending and receiving data, tags on photos, and NoSQL.

Ensuring a high-level of data confidence and data quality is important to the performance of the system, and therefore the origin of the data should be analyzed. Potential data sources include the firm's own private databases, or internally generated and maintained data; semi-private data, or databases / websites that use licensing models and charge fees for use / access; and public data, or open-access model databases / websites[12] [13].

Independent of the structure or data types, in order to be able to fully leverage and utilize ML within a product, data sources must be identified, transformed, and stored in a manner that allows for efficient analysis. These data sources can come from a plethora of origins ranging from subjective conversations on social media to objective, discrete data gathered by highly respected organizations like NASA and made available on publicly accessible databases. "The design of any AI systems starts with the choice of training data, which is the first place where unfairness can arise. Training data should sufficiently represent the world in which we live, or at least the part of the world where the AI system will operate." (*A Future Computed*, *Page 58.*) For example, the data set must properly represent gender, race, and age. Consider an AI system that enables facial recognition or emotion detection trained solely on images of adult faces may not accurately identify the features or expressions of children due to differences in facial structure.

Basic scientists, engineers, clinicians, regulatory experts, and other subject matter experts should be involved with the analysis and selection of data sources that shall be utilized for CLS/ML analysis within a firm or organization for product or operation. These experts should analyze the data generation, collection, and maintenance techniques, ensuring that data is generated using valid scientific and regulatory techniques[11]. These experts should also be involved in training the CLS. In the case of the clinical lab, clinical scientists perform analysis with and without the AI/CLS/ML over the course of multiple runs. Comparisons are then made to determine how close the AI results are to the experts' results. Consistent performance by the experts is key and all participants must use the same defined method of data collection, must register data in the same way, and must have uniformly defined terms.

Data quality metrics can be used to track and trend possible areas for improvement. For example, consider data "veracity," which refers to the inherent biases, noise and abnormality in data. Data does not always represent exact values, but rather close approximations, which leads to some level of

[11] https://www.webopedia.com/TERM/U/unstructured_data.html. Extracted June 19, 2018

[12] Food and Drug Administration (FDA) Center for Devices and Radiological Health (CDRH) and Center for Biologics Evaluation and Research (CBER). Guidance for Stakeholders and Food and Drug Administration Staff: Use of Public Human Genetic Variant Databases to Support Clinical Validity for Genetic and Genomic-Based In Vitro Diagnostics. FDA Maryland. April 13, 2018

[13] Orenstein, Gary; Doherty, Conor; Boyarski, Mike; and Boutin, Eric. Data Warehousing in the Age of Artificial Intelligence. O'Reilly Media, Inc. 2017

impreciseness and uncertainty. The reliability and trustworthiness of the data should also be considered and accounted for in the plan for the application.

There should be control over how private data is generated and gathered. Technical or discrete data sets should be gathered using validated or high level of confidence methods with appropriate controls (e,g, traceable to national or international standards, previously tested internally-made controls, etc.) by qualified experts following approved procedures or recognized standards. To a certain extent, firms can also control the method to which non-technical data, like those data for adverse event reporting or even meeting minutes, is generated and gathered by training those responsible individuals to follow a consistent method for the task.

Those who work in highly regulated industries might consider the appropriateness of the data source for the intended use. For example, medical device manufactures might not think Twitter would be an appropriate data source for an CLS/ML system whose intended use it to predict failures in a Bill of Materials (BOM) system. However, the beauty of AI/CLS/ML is to find connections that might not be obvious or easily deduced. Therefore, during the development and training of an AI/CLS/ML system, exercising the use of a wide scope of data sources based on the network and tool's processing capability may lead to surprising results[14].

General Software Design

The software design process needs to include an evaluation of the appropriateness of the algorithm used (e.g. justify why a neural network is an appropriate design choice for this particular product) as well as an evaluation of the post-market CLS process.

Consideration should also be given to placing limits on how much the CLS system is allowed to change and how to manage the influence output from one CLS system has as an input to another CLS system.

How the data is presented influences whether or not the data is accepted by the user, therefore consideration should be given to how information is presented.

For decision support systems where a human has the ability to accept or reject the CLS recommendation, a user's past experience with the software may bias their future use. For example, a user that has had poor success with a previous version of the CLS system may be highly skeptical of future releases, regardless of any objective evidence the developer may have about improvements to the product. This skepticism may also be specific to particular set of parameters, for example, not having confidence when using the software on older populations. Similarly, users with positive experiences may begin to rely upon the output from the software and could be less likely to question faulty output.

CLS systems will also need to address potential impacts from changes in healthcare practices. These types of changes in the practice of medicine (i.e., changes to reduce over-prescribing of antibiotics[15]) are handled under the "traditional" medical device development process in that, (1) it takes time to incorporate into a product design, and (2) humans are involved in the design change. If a CLS system is

[14] Ranly, Nick. "AI Work Group Update." April 3, 2018, Microsoft PowerPoint file

[15] https://www.cdc.gov/media/releases/2016/p0503-unnecessary-prescriptions.html

developed under one practice of medicine paradigm, it may not adjust well or correctly if information related to a different paradigm becomes part of the input dataset.

Confidence and Explainability

One of the interesting challenges with CLS systems is the ability to have confidence in the system. In traditional software development projects, one contributing factor to having high confidence in a product is having transparency to algorithm design and robust verification and validation activities.

> "An approach that is most likely to engender trust with users and those affected by these systems is to provide explanations that include contextual information about how an AI system works and interacts with data. Such information will make it easier to identify and raise awareness of potential bias, errors and unintended outcomes." (*A Future Computed*, [4] *Page 72).*

However, with CLS systems, the software engineer is not directly creating the decision making algorithm, and the system can be thought of as a black box when testing.

The following items help build confidence in a CLS system:

- Initial performance metrics/specifications
- Information/knowledge about how the system learns in time
- Proper verification and validation process for software design
- Quality control for the new data that causes the system to learn/change (clean data)
- A reasonable envelope in the change of the behavior of the algorithm, e.g., too much change is not permitted
- Triggers for algorithm change are clearly described
- The system monitors its performance in time using an appropriate metric and reports the performance to the user in regular intervals
- The user has the ability to reject an algorithm update or roll back to a previous algorithm version
- The user is informed each time the learning has caused a significant change in behavior, and the change is clearly described
- Ability to reverse engineer an evolving algorithm, i.e. determine how the machine made a decision

For Explainability, the description, flows and any data explaining the algorithm(s) model is important in order to show how the system achieves a conclusion. This is not an easy part as the algorithm(s) used will be highly complex. The importance of this is to show transparency and understanding of the process which should not be managed as a black box and serves to confirm the ethical values of the organization.

Explainability is a term that has taken on renewed meaning and purpose in AI circles, and generally references the need to completely understand and document the logic, decision methodology and data sources utilized in developing the output of an AI system, a recommendation, a prediction, or a decision. This notion becomes problematic for certain advanced types of AI systems that don't include human intervention, particularly CLS systems. Furthermore, this notion of Explainability has taken on a new sense of urgent formalization with the recently launched EU driven GDPR regulations which include language that has the potential to limit the scope of AI and CLS systems. For example, the regulations provides rights to the impacted data subjects and users to request meaningful information about the logic involved in an AI-driven output, as well as the ability to challenge decisions after they have been made.

As with most regulations oriented towards new technologies, we will likely learn over time about the practicality and acceptability of these important needs, though given the general interpretation gaps in most regulations, one can assume a new compliance burden with AI systems. The industry at large, particularly the healthcare segment, has an obligation to proactively begin defining pathways, perhaps incremental pathways, that would allow us to take measured steps to ensure we don't limit the full potential of AI systems, while addressing the growing concerns of these systems through joint learnings and broader awareness of the practical implementations of transparency.

Integrating Values & Ethics

Society will only achieve the public health promise of AI-driven innovation if these systems are developed and deployed responsibly, with a principled innovation approach that takes into consideration the following guiding principles:

- AI that maximizes efficiencies without destroying the dignity of people
- AI that guards against bias
- AI that is accountable and traceable so humans can undo unintended harm
- AI that is transparent

When sensitive personal data is used, establishing a Data Lifecycle Management (DLM) or Information Lifecycle Management (ILC) procedure(s) is critical. These procedures help collect, organize, use and disseminate, maintain, protect and preserve, and dispose of data in risk-based methods that meet regulatory requirements and the public's expectations[16].

Patient Consent

In many situations, patients must provide explicit consent for their data to be used for a particular use or set of uses. Regulatory requirements vary globally (e.g. HIPAA, GDPR, etc.) and those requirements are outside of the scope of this document, however, it is important to be aware that existing patient consent forms may need to be updated and may need to be as dynamic as the CLS system itself[17] [18].

Companies will also need to plan for and implement robust regulatory monitoring to understand when and how data privacy requirements are implemented or changed.

One new item to consider is if an existing, legacy patient should be notified when the CLS system has updated its algorithm. For example, if a CLS system has been trained with 10,000 patient records and then performs a diagnosis on a patient in January 2020 and determines the patient is cancer-free, should the patient be notified when the CLS system has learned from an additional 50,000 records because there is a potential for change in the diagnosis? If the patient's record was part of the dataset used to generate the updated algorithm, how do we prevent or manage the potential for "overfitting" the model? Do some assessments made by a CLS system have a "shelf-life", meaning that after a period of time or certain events transpire, does it matter whether the patient is notified of a different result? What ethical or legal liability questions might need consideration in this scenario?

[16] https://www.tpsgc-pwgsc.gc.ca/biens-property/sngp-npms/ti-it/conn-know/giim-lngdsc-eng.html

[17]http://journalofethics.ama-assn.org/2011/03/ccas2-1103.html

[18] Haug, Charlotte J. "Whose Data Are They Anyway? Can a Patient Perspective Advance the Data-Sharing Debate?." NEJM 376.23 (2017): 2203-2205

Retraining

Retraining frequency could be triggered based on a variety of criteria – timeliness, increase in volume of data available, new or an improved source of data, etc. Consideration should be given regarding relevance of the data either because of changes in the practice of medicine, availability of new therapies, or errors found in datasets.

Consideration should be given to setting bounds on the amount of change that will be allowed. For example, CLS used in specific healthcare domains could be provided limits based upon accepted medical practice in the space or protocols established within the healthcare system the CLS is being used.

Risk Management

As with all processes in the life science industries, the use of a Risk Based approach to mitigate / eliminate flaws that can translate into negative consequences for patients is expected. As with any system, there are two different areas of risk:

1) Development Risks: associated with the requirements, design, and implementation of the system
2) Failure Risks: associated with failure or wrongful use of the system

Both risk areas are tied to two critical elements to determine Risk Level:

1) Risk to the patient and/or critical quality attributes
2) Level of autonomy of the system

An Autonomous System that directly takes action will likely have a higher risk profile than a Clinical Decision Support system that makes recommendations. These systems will also be used in the manufacturing process of medicines and devices as well as in the delivery of drugs in the human body to cure/control illnesses and improve patient quality of life, which are associated with different kinds of risks.

Risk should be commensurate to the risk of the system's autonomy to take actions based on conclusions and specific controls should be in place to address the risk.

Many of the concerns about risk in CLS systems relates to the quality of the training data and the runtime data. Consideration should also be made for the quality of the CLS development and cybersecurity of the CLS application and the system(s) it runs on or is otherwise connected to.

The developers should be familiar with data being collected. For example, in the US it is typical to have adult patient weight in pounds, but weight data for children is in kilograms. There have been incidents in the past of teenagers receiving a medication under-dose or over-dose because their weight was represented in incorrect units because they initially began their hospital stay in an adult unit, and were moved to a pediatric unit for treatment (or vice-versa.) Other countries have standardized on kilograms as the weight, so when importing a patient's medical record, it is only necessary to import their "weight" field, but in the USA it is necessary to import the "units" field as well.

There is always the potential for bias in the data. For example, if the data is taken from a volunteer population, the volunteer population might not be representative of entire population.

Another example is that when data comes from multiple sources, are different sources coding[19] the same way? Even within the same data set, are the remote clinics coding items in the same way as the one in a larger city?

Semantic interoperability of data and identification of key or critical information supporting a domain of use should be determined. Critical information might include a variety of contextual information that is not specifically found within the healthcare coding system. For example, weather patterns and conditions might be important to CLS involved in assessment of asthma patients.

When mining existing data sources, since the data was intended to be used for a different purpose, the coding might not mean what you think it means. Some hospitals can be very creative with their coding practices that are intended for insurance and billing purposes. Alternatively, there are errors in entry of information in EMRs[20].

Sub-populations which may have additional benefit or additional risk introduced should be identified.

Some applications may use information from published papers as a data source; however this raises some concerns:

- Scientific papers can be retracted – how will that impact the doctor's decision? Should the caregivers be alerted of a change in results?
- The quality of published papers is variable – the results from some published studies are not reproducible.
- Published papers can conflict with each other.

CyberSecurity, Authentication, Privacy, and Anonymization

By their very nature, CLS systems rely on large amounts of information for proper operation. This information relationship brings up several concerns related to cybersecurity and privacy.

For example, controls should be in place to ensure the data used for training is genuine and has not been modified; similarly, the input and output data during clinical use should have appropriate safeguards. Adding to the complexity, safeguards considered appropriate in one country or scenario may not be considered appropriate in another.

Patient information used for training purposes should be anonymized. If this training is occurring in real-time, anonymization should also occur in real-time.

[19] Coding in healthcare is the transformation of healthcare diagnosis, procedures, medical services, and equipment into universal medical alphanumeric codes. The diagnoses and procedure codes are taken from medical record documentation, such as transcription of physician's notes, laboratory and radiologic results, etc.

[20] https://www.ecri.org/Resources/In_the_News/PSONavigator_Data_Errors_in_Health_IT_Systems.pdf

https://www.medpagetoday.com/blogs/skepticalscalpel/71971?comment=true

http://www.modernhealthcare.com/article/20160227/magazine/302279829

https://www.sciencedirect.com/science/article/pii/S235291481730148X

There should be a verification that any medical records being evaluated by a CLS system for diagnostic or treatment purposes actually matches the patient being diagnosed or treated. Stated another way, the system should ensure that personalized care is being applied to the right person.

Adaptive systems should also ensure that the fact that they are periodically updated doesn't lead to security concerns of this system or of associated systems (e.g. functionality where a new ruleset is downloaded from the cloud may introduce a new threat vector that could be exploited.)

As governments and other organizations are evaluating these technologies, many concerns of privacy and security have been voiced. As previously mentioned, the European Union (EU) General Data Protection Regulation that impacts all firms that utilize EU citizens' personal data, regardless of location. The GDPR does not prevent EU citizens' personal data from being used for AI/CLS/ML, but it certainly will cause limitations and additional hurdles[21] [22]. To address concerns of privacy while fostering AI/CLS/ML development, government organizations like The National Institute of Standards and Technology (NIST) are encouraging firms and individuals to create a method that would allow personal data to be utilized but still de-identified, like protected health information (PHI), through the Unlikable Data Challenge[23].

4. CLS Lifecycle

As CLS software is being developed, there are some considerations that are best described according to their stage in the development lifecycle. The purpose of this section is to present those considerations.

The ISO/IEC 62304 software standard is organized by different stages of the product development lifecycle; this paper has adopted those lifecycle stages so that it is easier to integrate these considerations into existing software development procedures. Additionally, software developers familiar with the AAMI TIR45:2012 "Guidance on the use of AGILE practices in the development of medical device software" should find it easy to adapt their considerations to an agile development process.

Planning

The Software Development Plan obviously plays an important role in ensuring product quality. In the case of CLS, the plan supports the product development lifecycle. For example, clearly state process steps needed to ensure data integrity, reliability and validity. How do you know what good input data looks like? Will there be a single source or a variety of sources for data? Will the source of training data for the "continuous" stage be different than what was used to initially develop the product? Will data extraction tools be used? What acceptability criteria will be different? Should the acceptability criteria be revisited after the product is launched? Should the acceptability criteria change as well as the

[21] Wallace, Nick and Castro, Daniel. The Impact of the EU's New Data Protection Regulation on AI. Center for Data Innovation. March 27, 2018

[22] Meyer, David. AI has a Big Privacy Problem and Europe's New Data Protection Law is About to Expose It. Fortune.com. May 25, 2018

[23] U.S. Department of Commerce: National Institute of Standards and technology (NIST). www.nist.gov/news-events/news/2018/05/help-keep-big-data-safe-entering-nistsunlinkable-data-challenge. May 1, 2018

algorithm? Will code generation tools be used? Since both defects and bad data can lead to undesirable outcomes, how will the team know the root cause?

The post-market monitoring plan goes beyond the typical monitoring of complaints. Active surveillance is crucial to patient safety and user adoption. In many instances, the performance and patient outcomes of the product can be measured in real-time (e.g., how many diagnoses were correct?) As the goal of a CLS is to learn, metrics are an invaluable tool to measure progress over time.

CLS systems learn continuously and update after product launch. Like all Continuous Quality Improvement programs, CLS systems are always in the "Design" stage, constantly updating data, verifying itself, and making improvements to the system. Since the CLS system is always in the "Design" stage, routine risk assessment is needed to assure that risks are properly mitigated and verified as the CLS system develops.

For a CLS system, it is also critical to plan how the system will learn. For example, if the system uses a neural network, will the neural network architecture evolve, or will just the neuronal weights change? Will new sources of information be added, or will the system evolve based on the increasing number of cases? What is an acceptable level of performance for a new update to be implemented? Will updates be implemented locally (i.e., units at different locations may evolve differently) , or will they be centralized?

An extensive checklist of elements is contained in a companion document "GmLP Checklist."

Requirements

The requirements should include a description of:

Quality Assurance - The requirements should include any run-time quality assurance checks that the software will perform on the training and test data sets. For example, some projects may have a separate process to assure the quality of the training data, other projects may build-in certain quality checks to identify quality concerns at run-time.

Quality Improvement - To better support the product after it is launched, one should consider including features and requirements that will help with future root cause analysis and feature improvements. For example, to support future root-cause analysis, when a CLS system is in the field and is updated, the data, time, data source(s) should be logged, as well as a snapshot of the weighting factors of the neural network. This snapshot could be very useful to quickly track down significant shifts in performance.

Confidence, Explainability, and Trust - Post-market adoption requires consumers and users to trust CLS output and implementation. Requirements for Confidence- and Explainability are needed. These may include separate considerations for patients, physicians and clinical staff.

Architecture

The architecture should include a description of the AI platform, including warehousing, data extraction or transformation tools that will be used, code translation tools (if applicable), relationship between the training data set and the CLS system.

Successful CLS systems often find additional applications in the future. Modularity opportunities should be considered to ensure that significant subsystems can be reused in other products.

Verification and Validation

Continuously learning is about learning, unlearning and relearning. When this is done by systems, the need to establish confidence in the systems operations and output becomes a necessity. Verification and Validation provide evidence for Confidence Level and system performance. However, Verification and Validation tend to be confusing terms and often used interchangeably.

> *"The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn."*
>
> *Herbert Gerjuoy*

For purposes of this paper, the following definition from O'Keefe et al 1987[24] is used:

> Verification: refers to "building the system right", that is, substantiating that a system correctly implements and satisfy its specifications.
>
> Validation: refers to "building the right system", that is, substantiating the system performs with an acceptable level of accuracy.

Verification

An approach to verification and related to the concepts of consistency, completeness, correctness and redundancy shall be defined to either achieve detection of anomalies and/or elimination of errors in order to build confidence into the system.

Developers can document the expertise they possess or acquire in a book of knowledge (BoK) that will serve as a foundation for the 'explainability' of the CLS system's algorithm. This is an important deliverable of the process and will require constant revisions as the algorithm matures in time and due to the nature of the content, it is expected to be a highly technical document with a very sophisticated scientific and mathematical content and probably written from a Data Scientist point of view.

Validation

Validation is critical to the success of CLS software, and validation of CLS software is very different than validation of other medical device software. In general validation is confirmation by examination and provision of objective evidence that the designed system conforms to user needs and intended uses. Focusing this definition to AI leads us to the measurement of the performance of the AI system by using an independent reference standard. The reference standard can come from many sources, including a well-defined ground truth (e.g., did the patient develop the condition that the AI system aimed at

[24] O'Keefe, R., O. Balci, and E. Smith. "Validating Expert System Performance." IEEE Expert 2, No.4 (1987)

predicting), the consensus of experts in the field (e.g., AI compared to pathologists for interpretation of biopsy specimens), or the clinical decision made by a clinician.

Good validation starts with good understanding of the system and the intended use; but for AI, the introduction of new functional characteristics, such as learning, autonomy and unsupervised, makes the case for an evolution, or at least a new approach.

One of the most important questions to address validation is whether there is enough data. Data used for "training" and data used for "testing" should not be the same. To ensure that the predictive model can provide useful future validity, only data not used to build the model generated can be used, otherwise in can lead to overestimated level of predictability in the model.

When it comes to validation; considerable attention should be given to issues related to structuring the validation process (framework), particularly the establishment of the criteria by which the system will be measured against (baseline), the need to maintain objectivity, and the concept of reliability. Individual components of the system as well as the system as a whole should be investigated[25].

Challenges for V&V of CLS[26]

- Data Governance: With increasing regulations governing privacy/confidentiality of personal data, access to data can be a challenge in cases such as patient data. Data governance is needed to establish accountability and oversight over data. It identifies those that responsible to protect data from misuse have the necessary skills and expertise to effectively monitor AI, and helps to ensure the ethical commitment and values of the organization are aligned.
- Transparency/Explainability: Regulatory bodies may be interested in the learning and analysis process. In addition, understanding the causal reasoning process behind the conclusions of the system will certainly help the acceptance of the AI or CLS system in the clinic. Even if a "block box " system has high accuracy, acceptance in the clinic may be limited if clinicians are completely in the dark about how the CLS achieves its conclusions.
- Data Paradigms: Since data will be the raw material for CLS, paradigms about data will transform or disappear:
    - Data Silos: establishing a Data Centric approach to eliminate siloed data is imperative.
    - Data is not static: With CLS, "Static Data" is not a characteristic. The same input produces new different output with time. For example, something changes user behavior and the quality of model reacts on the new input. As the input domain becomes complex, finding problems will be more challenging.
    - Since results coming from the algorithm can change with new data, the acceptance of results will have to be based on an acceptable deviation of the model and data.

[25] O'Keefe, R.M. & O'Leary, D.E. "Expert system verification and validation: a survey and tutorial", Artif Intell Rev (1993) 7: 3

[26] https://www.techemergence.com/machine-learning-in-pharma-medicine

Post-market Considerations

Post-market activities should follow the post-market plan. Continuously Learning Systems need continuous monitoring to ensure the system is operating within its expected performance envelope.

Since CLS systems may claim to be superior to human judgement, it may be useful to create a continuous benefit-risk analysis that compares CLS performance to its human counterpart, as both change over time.

Installation

The Requirements phase includes specifications regarding the data that is used to train the CLS system; the data sources that are accessed during the Installation phase should be verified that they meet these specifications (e.g. data value, format, type, etc.)

User training may be required.

Version control can be particularly important for CLS systems particularly given the potential for frequent updates and the possibility to a rollback to a previous version.

Operation & Support

There should be a documented plan which includes how continuous learning must be monitored and evaluated for continual ‘fitness of use’ including establishing thresholds for taking action. Potential actions can include recalls, rollbacks, or adjustments to the CLS software.

In order for any CLS to continually and successfully evolve, the system needs to be continuously trained by new data. Sometimes there are data collection agreements between users and developers to support the concept of continuous learning; however, not all end-users will agree to the use of their data.

It some situations, it may be difficult to provide an unbiased supply of continual data once the product is launched.

Maintenance & Change Management

During this product lifecycle phase, the software is updated (e.g. new features, bug fixes, etc.) For CLS systems, the software updates may also be due to additional training, either as a continuous system or a periodic update by the manufacturer.

For learning systems, the rate of change may be much faster than traditional medical device software releases; there may be a need to change processes to allow faster releases.

Criteria should be established, before the initial product release, to assess when a software update is needed. For example, if the performance change is small, it might not be worth updating.

The user should be informed when an update has been performed and the change should be clearly described (e.g., factors that caused the change.)

The user should have the ability to reject an algorithm update or roll back to a previous algorithm version. However, this could result in multiple revisions of the algorithm active in the market at the same time. Companies will need to consider how they will track complaints or other issues with the algorithm so that the algorithm version can be ascribed to a specific set of comments.

After the system has been updated, retrospective reviews for legacy patients should be considered. An eye towards preventing overfitting of the data should be maintained to help ensure the validity and overall value of the algorithm.

Consideration should be given to patients who are currently undergoing treatment, and how introducing a retrained system midway through their treatment may affect clinical understanding. For example, consider the use case where a patient is judged by CLS to have cancer in January and treatment is immediately started. If the CLS system is updated and the patient is re-assessed in March and the cancer appears worse, does this suggest that the treatment is not effective or has there been progress but the cancer was much worse than the original CLS assessment indicated? An opposite scenario is where the CLS system indicates the patient has cancer in January and treatment is started immediately, however a reassessment in March indicates the patient does not have cancer. Since the patient has already begun treatment that is not needed, who reports this and what is reported?

The user should be able to compare both individual and population outcomes to prior versions of the algorithm.

Retirement

There are circumstances where an AI system might need to be retired. Data sources could be invalidated, data is corrupted beyond repair, or newer technology becomes available to address limitations with legacy systems. Or a system has exceeded its defined life or circumstance as determined by the developer. As with all systems, retirement takes careful planning and management to execute, and in some cases there may not always be a replacement system. For CLS applications, there may be the added burden or potential legal implications that would require the ability to reconstruct a past "decision." Also, given the networked nature of most AI technologies, one has to account for the impact on the dependent systems.

A retirement strategy should consider the following:

- Impact on other systems,
- Risk assessment of decommissioning the system,
- Archiving data and algorithm so that they can be accessed at a later date if needed,
- Storage requirements and expected duration,
- Documented procedures for the removal of the system,
- Notification of and engagement with all stakeholders.

5. Conclusion

As mentioned in the Background section, this paper was created by a team that formed after the 2017 AI Summit held by Xavier University, and it is intended as a starting point for sharing considerations and best practices when developing CLS applications in healthcare. There are aspects of successful CLS development and implementation that the team simply did not have time to address, and these ideas may serve as the starting point for future papers.

One factor for successful CLS applications is the quality of the data set. Although data quality has been a topic in this paper, there are many additional aspects of data quality that can be explored further. For example, data models are important. If you use information from two different data models, how compatible is that data? Even if the data models are compatible with each other and data may look the same, there may be difficult to detect differences which can bias the analysis. Is information being co-mingled from different data sources? CLS systems can leverage existing data sets that were developed for other purposes, which should trigger questions about potential bias in the data.

Some have suggested that information overload is driving "physician burnout"[27] and well-meaning CLS applications may further contribute to this burnout. One potential future topic is how CLS systems can be designed to reduce a physician's information processing needs, rather than making the issue worse. In effect, providing an "Augmented Intelligence" information source for the physician to use in practicing medicine and treating patients.

One contributing factor for this overload are applications that are poorly integrated with existing healthcare systems and existing workflows. Having to switch back and forth between unintegrated systems takes time away from the actual practice of giving care. These dimensions hold true beyond provider facing applications, to a wide range of CLS use cases including manufacturing and quality management and control and many more.

Although Explainability, trust, and transparency have been discussed in this paper, we believe this topic should be further explored. CLS applications will not improve healthcare if no one trusts the results of those applications.

[27] http://annals.org/aim/article-abstract/2680726/physician-burnout-electronic-health-record-era-we-ignoring-real-cause

Appendix I – Glossary

It should be noted that in the AI/CLS/ML field, terms are often used casually and interchangeably. For example "adaptive learning" and "continuously learning" may be used to describe the same concept. For purposes of clarify, the following definitions are used in this paper.

ADVANCED BROAD-BASED ANALYTICS (ABBA)

Analytics based on a large volume of data as well as a variety of different types of data.

ALGORITHMS (CLUSTERING, CLASSIFICATION, REGRESSION, AND RECOMMENDATION)

A set of rules or instructions given to an AI, neural network, or other machine to help it learn on its own.

ARTIFICIAL INTELLIGENCE (AI)

A machine's ability to make decisions and perform tasks that simulate human intelligence and behavior. Alternatively – 1. A branch of computer science dealing with the simulation of intelligent behavior in computers. 2. The capability of a machine to imitate intelligent human behavior (source: Merriam-Webster)

ARTIFICIAL NEURAL NETWORK (ANN)

A learning model created to mimic some aspects of the human brain to solve tasks that are too difficult for traditional computer systems to solve.

AUGMENTED INTELLIGENCE, also known as INTELLIGENCE AUGMENTATION (IA)

Systems that are design to enhance human capabilities. This is contrasted with Artificial Intelligence, which is intended to replicate or replace human intelligence.

CLASSIFICATION

The problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

CLUSTERING

Clustering algorithms let machines group data points or items into groups with similar characteristics.

CONTINUOUSLY LEARNING SYSTEMS (CLS)

Continuous Learning Systems are systems that are inherently capable of learning from the real-world data and are able to update themselves automatically over time while in public use.

DECISION TREE

A tree and branch-based model used to map decisions and their possible consequences, similar to a flow chart.

DEEP LEARNING

The ability for machines to autonomously mimic human thought patterns through artificial neural networks composed of cascading layers of information.

MACHINE LEARNING

A facet of AI that focuses on algorithms, allowing machines to learn and change without being programmed when exposed to new data.

NATURAL LANGUAGE PROCESSING

The ability for a program to recognize human communication as it is meant to be understood.

REAL TIME HEALTH SYSTEMS (RTHS)

Information systems that collect and analyze real-time information from a patient; this in contrast to systems which take a patient's blood pressure or heartrate only when they are in the doctor's office or admitted to a hospital.

RECOMMENDATION

Recommendation algorithms help machines suggest a choice based on its commonality with historical data.

REINFORCEMENT LEARNING

A type of machine learning concerned with how software agents take actions to maximize a cumulative reward

SUPERVISED MACHINE LEARNING

A type of Machine Learning in which training datasets contains the desired or targeted outputs so that the machine can be trained to generate the desired algorithms similar to the way a teacher supervises a student.

UNSUPERVISED MACHINE LEARNING

A type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses (e.g. cluster analysis.)

Appendix II – Project Background

Xavier University launched the Artificial Intelligence Initiative in 2017, which brought FDA officials and industry professionals together to collaboratively advance the pharmaceutical and medical device industries by augmenting human decisions with artificial intelligence such that decisions are more informed. During the Xavier Artificial Intelligence Summit (August 2017), a working team of FDA officials and industry representatives was formed to identify successful practices for evaluating systems that continuously learn.

> **Problem Statement:** Continuously learning systems have the promise of systems that learn and improve their performance in use. However, they also bear the risks of unanticipated outcomes due to a lack of human involvement in the changes, unintended (and undetected) degradation in time, confusion for users, and incompatibility of results with other software that may use the output of the evolving algorithm.
>
> **Goal:** To maximize the advantages of artificial intelligence in advancing patient health by identifying how to provide a framework to maximize the upside potential of CLS and minimize the risk of continuously learning algorithms in a way that minimizes risks to product quality and patient safety.

This paper is one of the projects that resulted from the 2017 Summit.

Appendix III - Customer Perspective

Healthcare affects everyone, and therefore it is important to keep in mind the perspective of patients and caregivers when developing any medical device or pharmaceutical products. or ensure Critical Quality Attributes (CQA's) for Drug Development and Manufacturing While it is easy to focus on the positive benefits such as improvements in diagnosis accuracy, customized care, and improved quality of life, it is also important to consider the challenges and potential negative effects of any new technology:

- Alarm fatigue is a significant problem in healthcare today. While Machine Learning (ML) algorithms can produce less frequent "alerts" (vs alarms), similar problems have been reported regarding alert fatigue, and many alerts are overridden/ignored.

    - Too many warnings are being displayed on a regular basis, which has predictable adverse consequences. Many of the reasons for this probably relate to fear of liability. If some of the steps in this article are taken, all parties could be better off. (Bates)

    - Clinicians became less likely to accept alerts as they received more of them, particularly more repeated alerts. There was no evidence of an effect of workload per se, or of desensitization over time for a newly deployed alert. Reducing within-patient repeats may be a promising target for reducing alert overrides and alert fatigue (Ancker)

- Adoption of Clinical Decision Support (CDS) systems in healthcare has been slowed down by several barriers and ML adds additional issues. Challenges include:

    - Attitudes: clinicians feel that to use a computer to aid clinical decision-making undermines clinical autonomy, interferes with the clinician-patient relationship, is in some other way dehumanizing, or simply could not in practice work.
    - Knowledge: clinicians simply do not know much about such systems, and do not know what is available. (Mead)
    - Trust: "Clinical adoption of automated trend detection will require that significant changes be communicated to the anesthesiologist in a way that is intuitive and useful. This means that the anesthesiologist must understand the processes used to detect these changes and trust that the information delivered has value in improving clinical care" (Asermino 2009)
    - Transparency/Explainability: Electronic Health Predictive Analytics (E-HPA) transparency is required because clinical decisions ultimately need to be made by patients, clinicians, and the institutions that serve them. Whenever possible, clinicians, in particular, need to be able to "see into" a risk-prediction model and under- stand how it arrived at a certain prediction. (Amarasiingham 2014)

Kawamoto et al. (2005) found that successful implementation of CDS should "(a) provide decision support automatically as part of clinician workflow, (b) deliver decision support at the time and location of decision making, (c) provide actionable recommendations, and (d) use a computer to generate the decision support"

However, it is also widely recognized that new technology must work in concert with caregivers for it to be effective. In the article "What This Computer Needs Is a Physician"[28] the authors note that existing EMR systems may serve as efficient administrative and billing tools, it does not necessary serve the needs of caregivers, and that designers of artificial intelligence and machine learning tools need to understand that healthcare is more than just prediction, and freeing caregivers time so that they can spend more time with their patients will be of tremendous benefit.

Relating back to AI and machine learning, a recent report to the AMA Board of Trustees suggested that "combining machine learning software with the best human clinician 'hardware' will permit delivery of care that outperforms what either can do alone." [29] Accordingly, the AMA now refers to this science as "Augmented Intelligence".

Prime AI applications include clinical decision support, patient monitoring and coaching, automated devices to assist in surgery or patient care, and management of health care systems. AI in health care holds out the prospect of improving physicians' ability to establish prognosis, as well as the accuracy and speed of diagnosis, enabling population-level insights to directly inform the care of individual patients [9], and predicting patient response to interventions.

AI can streamline health care workflow and improve triage of patients (especially in acute care settings), reduce clinician fatigue, and increase the efficiency and efficacy of training. Moreover, shortages of medical experts to meet the needs of vulnerable and underserved populations in domestic and international settings could potentially be relieved, in part, by AI[37].

Patients or consumers will likely also ask for a certain level of transparency or knowledge in terms of how and what drives a particular system prediction or outcome, particularly at the outset while we are still developing our confidence thresholds.

[28] JAMA. 2018;319(1):19-20. doi:10.1001/jama.2017.19198e

[29] REPORT 41 OF THE AMA BOARD OF TRUSTEES (A-18) Augmented Intelligence (AI) in Health Care

Appendix IV – Common CLS Validation Techniques

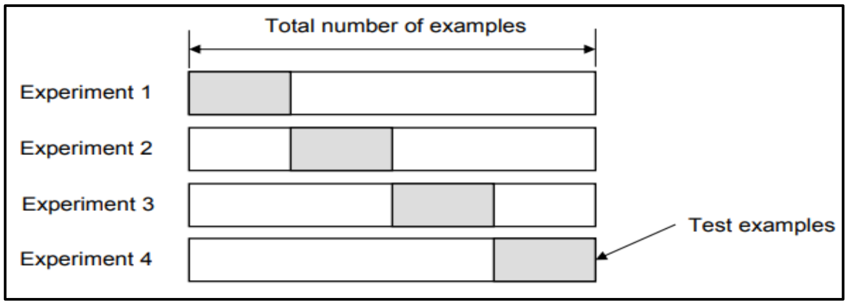
There are several techniques that can be used for validation of CLS. Thorough analysis should be exercised when choosing one and more than one can also be used if the situation calls for it.

Some validation techniques include[30]:

- **Resubstitution**: If all the data is used for training the model and the performance is evaluated based on outcome vs. actual value from the same training data set, this performance estimate is called the *resubstitution performance*.

- **Hold-out**: Typically, the resubstitution performance estimate is optimistically biased. To avoid this bias, the data is split into two different datasets labeled as a training and a testing dataset. This can be a 60/40 or 70/30 or 80/20 split. This technique is called the hold-out validation technique. In this case, there is a likelihood that uneven distribution of different classes of data is found in training and test dataset. To fix this, the training and test dataset is created with equal distribution of different classes of data. This process is called stratification.

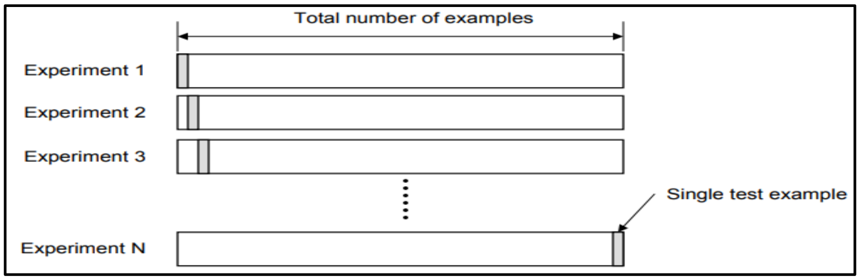
- **Cross-Validation**
  - **K-Fold Cross Validation**: In this technique, k-1 folds are used for training and the remaining one is used for testing

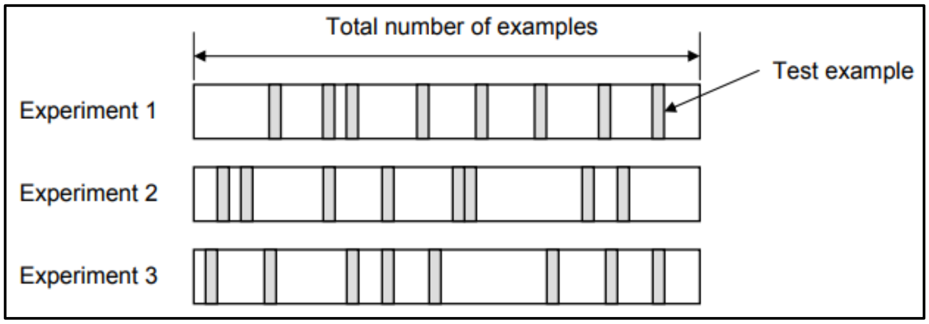
    The advantage is that entire data is used for training and testing. The performance of the model is averaged over the different folds.. This technique can also be called a form the repeated hold-out method. The error rate could be improved by using stratification technique.

  - **Leave One Out Cross Validation** (LOOCV): In this technique, all of the data except one record is used for training and one record is used for testing. This process is repeated for N times if there are N records. The advantage is that entire data is used for training and testing. The performance is estimated by aggregating the results for all records.

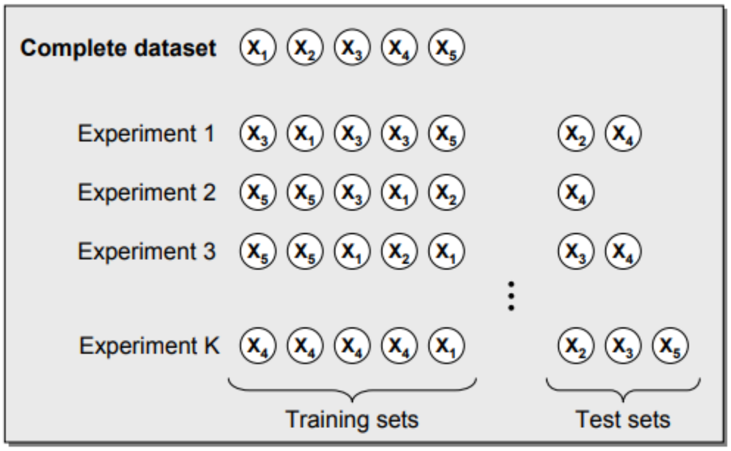
[30] https://dzone.com/articles/machine-learning-validation-techniques



- **Random subsampling**: multiple sets of data are randomly chosen from the dataset and combined to form a test dataset. The remaining data forms the training dataset. The following diagram represents the random subsampling validation technique. The performance of the model is found by aggregating the results from each iteration.

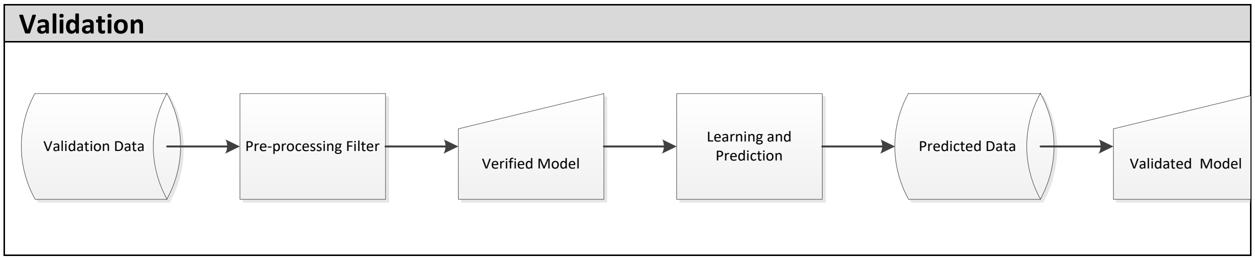


- **Bootstrapping**: A number of bootstrapping experiments are performed to estimate the performance. In a single experiment, the training dataset is randomly selected with replacement to generate a training bootstrap sample. The remaining examples that were not selected for training define a test bootstrap sample. The performances for the training and test bootstrap samples are combined using methods based on bootstrap theory, and the results from K experiments are averaged


Complete dataset X_1 X_2 X_3 X_4 X_5
Experiment 1 X_3 X_1 X_3 X_3 X_5 X_2 X_4
Experiment 2 X_5 X_5 X_3 X_1 X_2 X_4
Experiment 3 X_5 X_5 X_1 X_2 X_1 X_3 X_4
Experiment K X_4 X_4 X_4 X_4 X_1 X_2 X_3 X_5
Training sets
Test sets

When validating the CLS, caution should be exercised not to fall into the pitfall of validating which algorithm to use instead of focusing on the validation of the model itself. The wrong validation approach could lead to different expectations of what will really happen once the system is released into production.

Another important point of caution is that these validation techniques, which partition one data set into two or multiple partitions for training and testing, should not be repeatedly applied to the same test set to “fish for” improved results. A simple example of a “fishing expedition” is to partition the data set repeatedly into different hold-out sets and to report only the results from the best-performing hold-out set. A more sophisticated example is to use the results from one validation to modify an continuously-learning algorithm so that the results seem improved. If the results from the validation drive the learning, then the so-called “validation” data set can become part of training, biasing the reported performance.



Data Anomaly Detection for CLS

As part of the continuous validation state for a CLS, continuous monitoring of data available for quality purposes before the data is actually used by the system is important to ensure that such state continues to be effective.

The old precept of “Garbage In, Garbage Out” is critical for computerized systems. In order to achieve a high level of confidence in results coming from CLS, input data needs used need to be “qualified” for consumption of the system. Whenever “new-data” is to be fed, a pre-filter should sift that data to discard/quarantine new input that does not adjust a set of characteristics. The filter can be based on discrete “fit/not fit” approach or based on a level of confidence of the data where a degree of error is allowed to be consumed by the system. This is especially true for “unsupervised” systems with access to unformatted data.

Any data containing a margin of error allowed should be identified in order to recognize any unusual result coming from the system. If such behavior is noted, then the system should also flag the situation for “repair” by the system or further analysis.

For systems that are “trained” where the input data is formatted, previously qualified and the learning/action process is supervised, the risk of data anomaly is far less, however, quality of data still needs to be ensured before being fed into the system[31].

[31] http://lvl.info.ucl.ac.be/pmwiki/uploads/Publications/VerificationAndValidationAndArtificialIntelligence/aivvis-aic.pdf

Appendix V - Examples

This section covers some examples of current and proposed use cases or scenarios of CLS use in health care. This is not intended to be exhaustive, rather to provide the reader a real-life example of the potential for AI and CLS in healthcare.

Healthcare Use Case 1: Myoelectric prosthesis control

Background: A myoelectric-controlled prosthesis is an externally powered artificial limb that is controlled with the electrical signals generated by the user's own muscles. One limitation of most prosthesis control methods is that controllers do not adapt over time to changes in the patient, the patient's intent, or the patient's usage patterns. As a result, most amputees cannot improve their limb controllers independently, outside the clinic.

Imagine a machine learning system that performs real-time predictions for the user's intent and continuously improves the control based on patient feedback. Such devices are being actively pursued[32]. By definition, the device needs to be trained for each individual patient and keeps learning after the patient starts using it. The device for each patient may show a different performance depending on user's training, which may illustrate and challenge some of the performance estimation methods.

Healthcare Use Case 2: A computerized cancer detection tool for medical images that adapts its deep learning architecture of its classifier

Background: Computer-aided detection techniques in radiology aim at suggesting potentially suspicious regions to radiologists for further visual interpretation. By contrast, a computerized detection tool can be used to exclude some patients' images from interpretation by radiologists, with the assumption that if the computer did not detect anything, the patient is normal for the disease in question. Deep learning systems have found widespread use in abnormality detection on medical images.

The device in this example has been initially trained to detect lung cancer on thoracic CT images. A specific deep learning architecture was selected for initial training. The architecture includes the number of deep learning layers, the convolution kernel size at each layer, the number of convolution kernels at each layer, the type of pooling performed when going from one layer to another, the type of activation function at each layer, the flavor of gradient descent applied at each layer, etc. The designer intends to re-train the deep learning architecture as more data becomes available. In other words, at each update, not only the neural network weights may change, but the entire architecture may be different.

Contrasting this type of learning to a more basic type of learning in which only the weights (coefficients) change but the architecture remains the same may elucidate some of issues on the confidence in continuous learning and adaptation (change) method complexity.

[32] A. L. Edwards et al., "Application of real-time machine learning to myoelectric prosthesis control: A case series in adaptive switching", Prosthet. Orthot. Int., v. 40, p.p. 573-581, 2015.
P.M. Pilarski et al, "Online Human Training of a Myoelectric Prosthesis Controller via Actor-Critic Reinforcement Learning," Proceedings of the 2011 IEEE International Conference on Rehabilitation Robotics, Zurich, Switzerland, June 2011, pp. 134–140

Healthcare Use Case 3: Early Detection of Sepsis

Background: Sepsis is the most expensive condition treated in hospitals, accounting for approximately 5% of total hospitalization costs and an overall annual cost of USD 20.3 billion in the USA (Torio 2011), and more than GBP 2.5 billion in the UK (UK Sepsis Trust 2013). Early detection of sepsis via automated systems and subsequent timely intervention may reduce treatment costs and overall resource use (Rivers 2001; Yealy 2014). The UK Sepsis Trust estimates that there are more than 100,000 hospitalizations per year for sepsis, and that achieving 80% delivery of basic standards of care could result in a potential cost saving of GBP 170 million per year, even after allowing for increased survival-related costs (UK Sepsis Trust 2013)."

Automated detection systems offer the possibility of monitoring patients in 'real-time' (Meurer 2009), and can alert the relevant physicians or nurses (e.g. by email or pager) to the need for timely clinical evaluation and potential initiation of treatment." Harvesting the vast amount of clinical data – including real-time information, leveraging AI and CLS tools can increase the likelihood of sepsis identification and subsequently also sepsis prediction scores at time of admission[33].

Healthcare Use Case 4: Diabetes Personalized Medicine

Background: Adequate treatment and management of chronic disease outside of the care setting is a significant driver in building a pathway towards reducing the current trend towards adverse events largely based on poor management of a chronic condition and disease. Although this example is specific to diabetes, this can easily be applied to a host of other clinical conditions. This example pertains to a decision support tool for personalized medicine that automatically determines the optimal treatment for patients with insulin-dependent diabetes. Leveraging algorithms that combine data from connected devices, including insulin pumps, continuous glucose monitors and food consumption information to adjust the treatment plan for maintaining optimal glucose levels.

The CLS software learns individual patterns and supports recommends therapy plans for people with Type 1 diabetes that use insulin pumps. It is a healthcare provider's "expert partner", and offers unique insight and direction. It leverages machine learning and adaptive technology to continuously learn each individual, their habits, allowing for a closed loop system that automatically adjust insulin treatment and monitors individual responses to adjustments. The software automatically adjusts its insulin treatment and behavior modification recommendations on a personal or individual basis.

Healthcare Use Case 5: Lab Automation

Background: In the case of the toxicology laboratory used for pain management, LDT's(Laboratory Developed Test) are frequently performed by a version Liquid Chromatography that is paired with quadrupole tandem mass spectrometers (LC-MS/MS). Due to the complexity and the variability in these instruments, no two instruments of the same make and model, with the same method file and

[33] Evans, David JW, et al. "Automated monitoring for the early detection of sepsis in critically ill patients." The Cochrane Library (2016).
Horng, Steven, et al. "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning." *PloS one* 12.4 (2017): e0174708.
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174708

configuration will perform the same. Therefore, each instrument is treated as its own entity during the LDT validation process following 42 CFR 493 (CLIA '88), allowing for a variable outcomes.

CLS systems allow the user to train and optimize the CLS per their intended use for each individual instrument in the laboratory. Meaning that each instrument is trained and optimized individually with the CLS. The CLS developer will allow training within certain parameters based on locked quality rules and predetermined exceptions. There are also user – developer agreements that allow the developer to constantly collect patient de-identified data and instrument data. In the case that re-training is triggered, then the end-users will be alerted and will participate.

Other Healthcare Use Cases

There are numerous other examples, some of which are still in development. Some of these examples include:

1. **Monitoring/Diagnosis Glucose Monitoring Systems**: Machine learning algorithms help automate the process of monitoring blood sugar levels and recommend adjustments in care. https://www.techemergence.com/machine-learning-managing-diabetes-5-current-use-cases/
2. **Monitoring/Diagnosis Diabetes Nutrition Coaching**: To help recommend meal options based on the specific diet criteria of the user. https://www.techemergence.com/machine-learning-managing-diabetes-5-current-use-cases/
3. **Monitoring/Diagnosis Diabetes Early Diagnosis Tools:** Deep learning to predict the onset of diabetic retinopathy, the leading cause of vision loss among diabetics. https://www.techemergence.com/machine-learning-managing-diabetes-5-current-use-cases/
4. **Monitoring/Diagnosis Early detection of sepsis** http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174708
5. **Treatment Adaptive** Pacemaker algorithms that improve performance by 52% vs. locked fuzzy algorithm based devices https://www.ncbi.nlm.nih.gov/pubmed/11804178
6. **Treatment cardiac resynchronization therapy** algorithm that adapts to individual patients. The adaptive algorithm provides ambulatory adjustment of pacing configuration and AV and VV delays based on periodic automatic evaluation of electrical conduction. https://www.ncbi.nlm.nih.gov/pubmed/21968204
7. **Improved clinical effectiveness in ICUs through the use of Epimed's cloud-**based analytics that assists clinicians in gaining better insights faster on the best treatment options for ICU patients
8. **Improved operational effectiveness** by reducing claims fraud, waste and abuse through the use of CGI ProperPay's predictive analytics, workflow and rules management.

The following represent some great examples of enhanced imaging solutions leveraging AI, many of which are commercially available today. Many of these are potentially great candidates for CLS applications as well.

9. **Diagnostic Assessment** of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer https://jamanetwork.com/journals/jama/fullarticle/2665774
10. **Arterys Cardio DL™** is the first technology to be cleared by the FDA that leverages cloud computing and deep learning in a clinical setting. Arterys Cardio DLTM provides automated, editable ventricle segmentations based on conventional cardiac MRI image
11. **QuantX™ Advanced** system, the industry's first computer-aided diagnosis platform incorporating machine learning for the evaluation of breast abnormalities

12. **OsteoDetect** Analyzes wrist radiographs using machine learning techniques to identify and highlight regions of wrist fractures
13. **Interknowology** improves ultrasound intelligence using machine learning to assist clinicians in diagnosing Posterior Urethral Valve defects (PUV) in unborn children
14. **Intelligent Retinal Imaging Systems(IRIS)** utilizes machine learning to predict presence or likelihood of diabetes related vision deterioration

CLS In Other Industries

AI and CLS are not new – other industries have been leveraging these techniques for years. This paper's approach is to explain concepts and techniques from other industries and adapt them to use in medical devices. Below are some examples from other industries to illustrate the diversity and power of CLS systems.

1. **Siemens Optimization of Gas Turbine Combustion** Gas Turbine Autonomous Control Optimizer (GT-ACO), https://www.siemens.com/innovation/en/home/pictures-of-the-future/digitalization-and-software/autonomous-systems-ai-at-gasturbines.html
2. **GE a "brilliant factory"** where factory equipment and computers will talk to each other over the internet in real time, share information, and make decisions that will help ensure top-notch product quality and avoid plant shutdowns. https://qz.com/357610/ges-first-ever-brilliant-factory-just-opened-in-pune/
3. **Tesla** having continuous learning algorithms for self-driving cars https://www.technologyreview.com/s/608155/teslas-new-ai-guru-could-help-its-cars-teach-themselves/
4. **Uber Eats** https://eng.uber.com/machine-learning/ and Waze https://rctom.hbs.org/submission/the-new-waze-to-drive/ develop smart maps focused towards consumers.
5. **TransDev** focusing on Shuttle route management https://www.transdevna.com/services-and-modes/autonomous-mobility/
6. **Boeing** for auto pilot. An example of continuous learning landing of planes https://www.wired.com/2017/03/ai-wields-power-make-flying-safer-maybe-even-pleasant/
7. Adaptive learning (multiple use cases) https://www.datasciencecentral.com/profiles/blogs/adaptive-machine-learning
    a. Fraud Detection: Rules and scoring based on historic customer transaction information, profiles and even technical information to detect and stop a fraudulent payment transaction
    b. Financial Markets Trading: Automated high-frequency trading systems
    c. IoT and Capital Equipment Intensive Industries: Optimization of heavy manufacturing equipment maintenance, power grids and traffic control systems.
    d. Marketing Effectiveness Detect mobile phone usage patterns to trigger individualized offers:
    e. Retail Optimization In-Store shopping pattern and cross sell; In-Store price checking; Creating new sales from product returns.
8. The future of enterprise resource planning (ERP) is AI. AI is poised to take over ERP functions, with vendors adding new machine learning features and enterprises keen to investigate; this affects both pharmaceuticals and medical devices. https://www.arnnet.com.au/article/631525/future-erp-ai/