Machine Learning Ecosystem: Implications for business strategy centered on machine learning

by

Ritesh Shukla

M.S. Electrical Engineering
Worcester Polytechnic Institute, 2006

Submitted to the System Design and Management Program in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Engineering and Management

at the

Massachusetts Institute of Technology

June 2014

Signature redacted

Signature of Author _____

Ritesh Shukla
System Design and Management Program

Signature redacted

Certified by _____

Michael A M Davies, Thesis Supervisor
Senior Lecturer, Engineering Systems Division

Signature redacted

Accepted by _____

Patrick Hale, Director
System Design and Management Program

This page has been intentionally left blank.

Machine Learning Ecosystem: Implications for business strategy centered on machine learning

by

Ritesh Shukla

Submitted to the System Design and Management Program in June, 2014 in Partial Fulfillment of the Requirements for the Degree of Master of Science in Engineering and Management

Abstract

As interest for adopting machine learning as a core component of a business strategy increases, business owners face the challenge of integrating an uncertain and rapidly evolving technology into their organization, and depending on this for the success of their strategy.

The field of Machine learning has a rich set of literature for modeling of technical systems that implement machine learning. This thesis attempts to connect the literature for business and technology and for evolution and adoption of technology to the emergent properties of machine learning systems.

This thesis provides high-level levers and frameworks to better prepare business owners to adopt machine learning to satisfy their strategic goals.

Thesis Supervisor: Michael A. M. Davies

Title: Senior Lecturer, Engineering Systems Division

This page has been intentionally left blank.

Acknowledgements

This thesis has given me the opportunity to dive deep into the strategic implications of machine learning. Machine learning has been an area of keen interest for me and adding the business perspective for machine learning has given me insights that would not have been possible by a pure study of the science behind machine learning.

I would like to thank Prof. Davies for his infinite patience and being a source of inspiration, insight and enlightenment throughout my time at MIT. He will continue to influence my thinking and choices in personal conduct well beyond this degree. I envy those who have learnt more than me from him.

I would like to thank Prof. Patrick Hale and the staff of SDM. I discovered SDM at a time in my career where I needed additional exposure that would help me build on what I had accomplished till then. The tag line of "For those who wish to lead engineering but not leave engineering" is apt and it accurately met my needs. The knowledge and experience gained, as part of the SDM degree, is priceless. I will forever be grateful for the wonderful set of friends I have made during this journey.

I would like to thank my wife who has been my pillar of support, without her this degree would not be possible. Finally, my parents who in spite of all the hardships were able to raise two kids and have them go on to live the life they had dreamed for them.

This page has been intentionally left blank.

Table of Contents

Table of figures

Motivation

Machine learning is the process by which computers can make autonomous decisions based on the data provided without being programmed for it explicitly. To quote Arthur Samuel "Field of study that gives computers the ability to learn without being explicitly programmed" (Simon, 2013)

With reducing computing and storage costs and increased connectivity, more data is being collected, so that consequently more business strategies based on the usage of this data depend on machine learning systems. The range of business ideas is wide; some common examples are online search, recommendation systems of various kinds (advertisements, music, shopping) and dynamic content generation (online course tailoring).

Machine learning as a field within computer science has been around for a few decades but it continues to evolve. Despite the high interest for adopting machine learning, it is still relatively new for most companies to be the core of a business strategy.

The motivation for this thesis comes from need to address the gap between the strategy needs of an organization and the current state of machine learning. The goal is to provide high-level levers that help business owners to decide next steps and plan for the future when the strategy for an organization depends on an instance of a machine learning system. The target audiences of this thesis are business owners who face the business need to adopt a machine learning solution.

1. Introduction

1.1. Rise of machine learning

The cost of storage and computation has continued to decrease; penetration of digital technologies into all aspects of our lives has lead to the generation of large sets of digital data. Increasing prevalence of web technologies and better connectivity has made it easier to share data. This increase in the amount of data generated, collected and processed is often referred to as "Big Data". Machine learning is a field that has risen to prominence in the era of Big Data as it lets machines autonomously process new data based on the data that was collected previously without having to be explicitly told how to do so.

Many companies formed in the past 10-15 years in various domains depend on machine learning as a core competency (Example: Google, Pandora, Knewton). In addition increasingly numbers of companies build key business strategies that involve machines being able to process data and act autonomously; a typical example being fraud detection for online money transaction by companies such as Paypal (Schwartz, 2001).

Recommendation systems, voice recognition, self driving cars, web search, adaptive online education systems, and spam filtering are some of the few common applications of machine learning solutions.

With machine learning, businesses can go beyond just analytics on their data but can program computers to act on their own to execute their business intent/strategy. We are still in the early days of the adoption of machine learning into core business strategy but the adoption will continue to increase.

1.2. Adoption challenges of machine learning

Though the level of interest for machine learning is high, the adoption of machine learning has its own set of challenges.

The science of machine learning is not easy to comprehend, and it also continues to evolve. Machine learning is often viewed as a black box where the high level intent is understood, but there is little clarity on the details of its emergent system behavior that will impact the successful implementation of a machine learning system. Basic questions around the effectiveness of machine learning and the long-term needs for being successful are not well understood. The ecosystem around machine learning is also rapidly evolving and the changes in the ecosystem impact the effectiveness of the business strategy around machine learning. Moreover, the dynamic capabilities of an organization often do not include the skills needed to implement a machine learning system. (Robinson, 2014)

Our understanding based on the research done is that though there is significant literature that exists around machine learning, the bridge between the science and practice of machine learning is still in its early stage and the chasm that exists results in high unpredictability in the ability to execute a strategy around machine learning. (Carla E. Brodley, 2012)

1.3. Approach

Machine learning as a field has abundant literature for the emergent behavior of systems that implement machine learning. The primary audience of this literature is fellow scientists and engineers who wish to gain a broader insight into the

system behavior of machine learning algorithms and the systems that embody them.

The field of business strategy itself has well-established frameworks for analyzing technology and its impact. They have proven to be good frameworks to reason about technology and its impact on business.

This thesis primarily relies upon secondary research of the two fields, as the basis for an attempt to connect the dots between the two, to identify and qualify the adoption challenges, and to develop abstractions and frameworks that can help business owners implement machine-learning systems. In addition, to better understand the ecosystem primary research was done to identify the growth of companies, and the evolution of open source projects in the ecosystem.

1.4. Findings

The following findings will help business owners better understand and adopt machine learning as an element in their core business strategy. These are the high level levers that can be used against most attempts to implement an effective machine learning system.

1. When choosing to implement a machine learning system to gain a competitive edge there are two broad dimensions along which a company can measure their strategy.
    a. The business impact of the quality of implementation of the machine learning system
    b. The quality and exclusivity of the data used to implement the machine learning system

2. The ecosystem around machine learning is in early stages of ferment with many small players and many different business and technology designs at play.
3. The ecosystem is trending towards a more modular architecture.
4. The ecosystem is increasingly being centered around open source projects
5. For businesses adopting machine learning as a core piece of their business strategy there is a high rate of disruption that is driven by development of new S curves of machine learning implementations and/or access to new kinds of data.
6. Adoption of machine learning needs continued experimentation and continuous evolution of the system; this originates from the underlying science, the changing nature of data and the changing competitive landscape.
7. Depending on the nature of business, adoption of machine learning might need investment into interpretability of the implementation and openness and transparency of the overall system implemented.
8. This is an emerging field that is oversubscribed for talent, the core capabilities of the organization need planning and investing.

1.5. Organization

This thesis attempts to draw a relation between the emergent properties of machine learning systems, business needs and growth in the ecosystem over

time. We study the nature of growth in the ecosystem and derive implications for businesses.

Chapter 2 enumerates the frameworks used to structure the thesis. Chapter 3 looks into machine learning and derives the emergent properties that are relevant from a business execution standpoint. Chapter 4 looks into strategy implications of adopting machine learning. Chapter 5 looks into the ecosystem around machine learning and highlights the business implications of the ecosystem. Chapter 6 provides the concluding remarks.

2. Review of theoretical frameworks

Adoption of machine learning by a business is ideally driven by a strategy need of a business; the need can be better mapped to a machine learning system implementation by using a formal structure for the strategy itself (Porter, 1996) (Sull, 2005). Mapping strategy to the emergent behavior of machine learning systems helps better implement and preserve a competitive advantage that can be sustained. The dynamic ecosystem that exists often implies that active waiting and continued investment into preserving dynamic capabilities need to be looked into.

To understand emergent behavior this thesis takes a system architecture approach where we identify the top down level 1 form of a classical machine learning system. Identifying the moving parts and the variations in the level 1 and level 2 helps explain the variations of machine learning systems and helps identify the challenges for each form element (Crawley, 2009). The academic publications for form elements at level 1 and 2 help identify the emergent behavior and map their implications to the business goals.

Performance envelope and S curves (Davies, 15.965 Technology Strategy, 2009) (Christensen, Exploring the Limits of the Technology S-curve, Part 1: Component Technologies., 1992) (Christensen, Exploring the Limits of the Technology S-curve, Part 2: Architectural Technologies., 1992) help understand the disruptive nature of new technologies. This also helps us understand how a machine learning system can have it's own performance envelope that can be

disrupted by continued evolution of machine learning science, dynamic nature of data and ecosystem changes.

Machine learning has its own ecosystem that encompasses the science, technology, talent and business around machine learning. The ecosystem changes that happen around machine learning will dictate the performance envelope that is achievable by machine learning and the potential disruptions that can occur. Better understanding of the dynamic ecosystem involves understanding the adoption pattern (Geroski, 1999), technology battles (Suarez, 2004), modular ecosystem (Baldwin & Clark, 1997) and managing dynamic capabilities (Teece, Gary, & Shuen) are key.

3. Machine learning

Machine learning originated from the field of Artificial Intelligence and is designed to program machines to learn from data so as to be able to better perform a specific algorithmic task. One definition that lends itself well to this thesis is from (Flach, 2012) "Machine learning is the systematic study of algorithms and systems that improve their knowledge or performance with experience".

To be able to better explain the business implication of machine learning a good starting point is to understand a classical or canonical machine learning system.

We first define the basic intent of machine learning and then proceed to a level one decomposition (Crawley, 2009) of a system implementing a machine learning intent. This is to establish the various phases involved in implementing a machine learning system. The form elements at level 1 along with the business context will help establish the multi disciplinary nature of a machine learning system.

Having covered the level 1 decomposition we then take a look at the core science of machine learning and then proceed to extract the emergent implications of implementing a machine learning system.

3.1. Basic intent

The function of machine learning can be best understood by taking a look at the DIKW framework. The DIKW framework remapped as follows best illustrates the function of machine learning (Rowley, 2007) (Davies, ESD.39 Systems, Leadership & Management Lab (SLaM Lab), 2013)

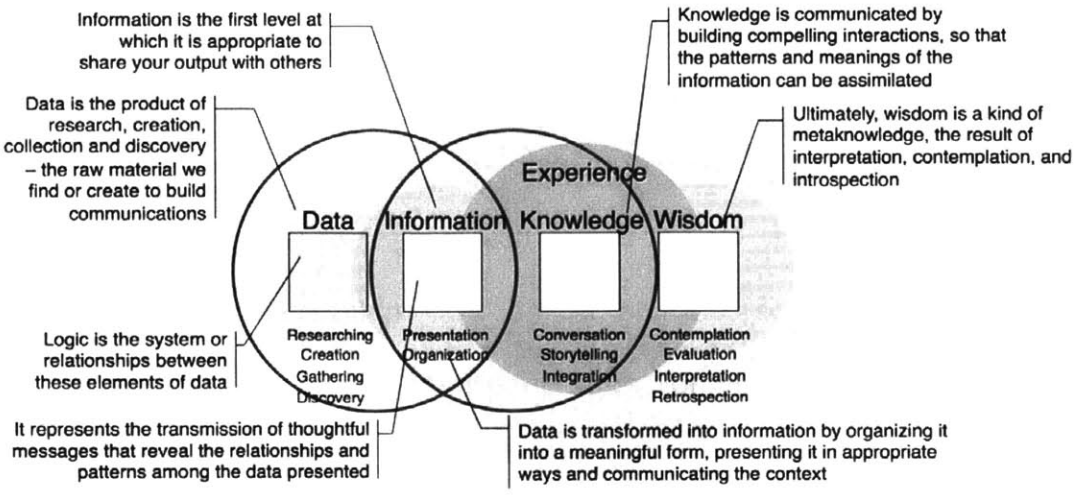

Figure 1 Function of MLSI, transforming data to wisdom

The primary functionality of machine learning systems is to take raw data from multiple sources and at the end mimic the behavior of having wisdom.

Thus the end goal of a machine learning system is to enable machines to exhibit behavior that mimics them having gained wisdom from the data available, and to adapt as and when new data becomes available. Example of this is the process by which a service such as Pandora is able to recommend a next song to play based on the station a user dynamically creates and the user's usage pattern. This gives the impression that the service understands an individual's preference in music without having to be explicitly programmed for each individual user of the service.

3.2. MLSI

We define Machine Learning System Implementation (MLSI) as the unique form of a system where all the form elements have been selected; the system has thereby been instantiated (Crawley, 2009).

Thus, a Machine Learning System Instance (MLSI) refers to the deployment of a machine learning system and it includes the sources of data, the transformations that take place on the data, the choice of enabling algorithms, the choice of technologies and the desired output. This would be similar to the reference of a specific model, year, trim and options of a car manufactured by a specific company (Example: White Volkswagen Passat 2014 TDI SEL Premium customized for New England weather) vs. referencing a generic category such as a 4 door mid size sedan or a 4 wheeled gas powered transportation vehicle. MSLI is specific to the context in which it is operating.

The reason to be specific rises from the fact that machine-learning systems deployed for a business are not yet typically off the shelf implementations and they need to be customized. The behavior is also dependent on the quality and quantity of data, which is unique to a specific instance. Thus, the effectiveness or the performance of the machine learning system will be specific to the instance.

3.3. Basic architecture

A generic level one decomposition of the form typically constitutes one or more sources of data that are merged, and a unified data model is created from which relevant features are extracted and represented to form the data that is used to implement the core machine-learning algorithm or algorithms. This data is typically split into training data and validation data. Training data is used to form the machine-learning model. Validation data is then used to insure that the model is not made to fit only the limited data that was used as training data but

instead is robust enough able to handle new data. Once the model is finalized the model then acts on new data as that data becomes available.

The process of updating the training and validation data and tweaking the over all system continues throughout the life of the MLSI. Most data sets to which machine learning is applied tend to be highly dimensional such that as new data is made available it can change the performance of the model that is deployed (Domingos, 2012). This makes continuous monitoring of the performance of the model very important as data is processed, and typically requires continuous adjusting of the model itself.

A level 2 decomposition of the system typically deals with the scale of the data and the requirement for the speed of processing. Level 2 is typically the level at which the requirements can classify the MSLI as a "Big Data" solution. In this thesis the discussion around the business implications of MLSI are intended to be true for all forms of MLSI and are essentially common for any specific form of MLSI not just "Big Data" MLSIs.

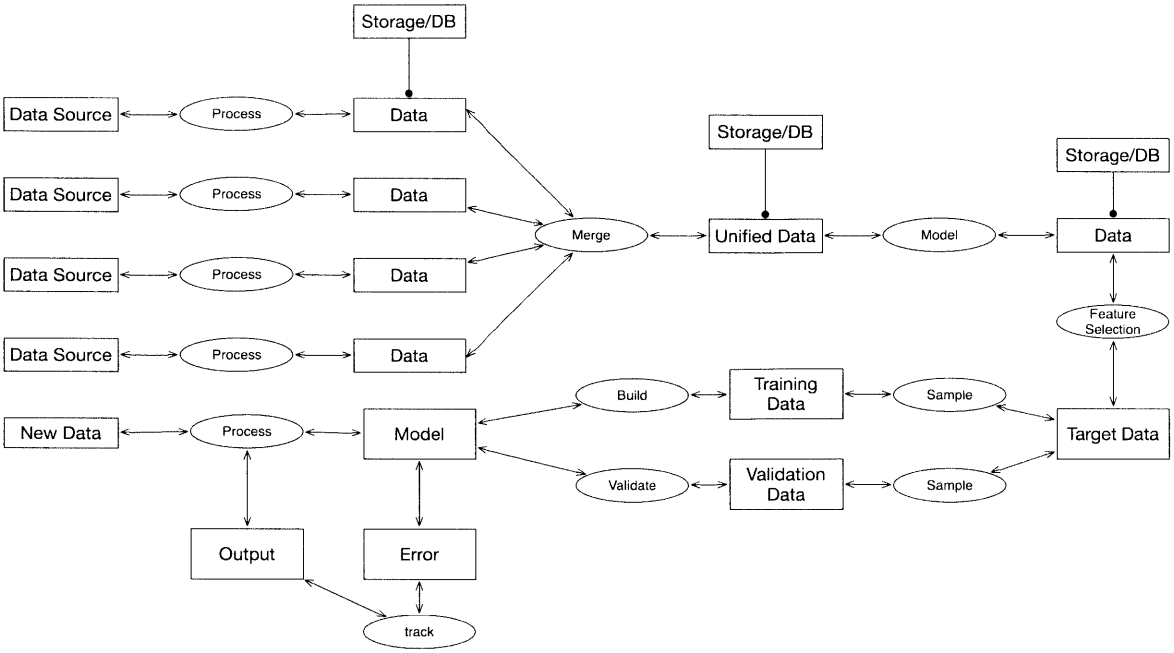

Figure 2 MLSI Level 1

3.4. Multidisciplinary science

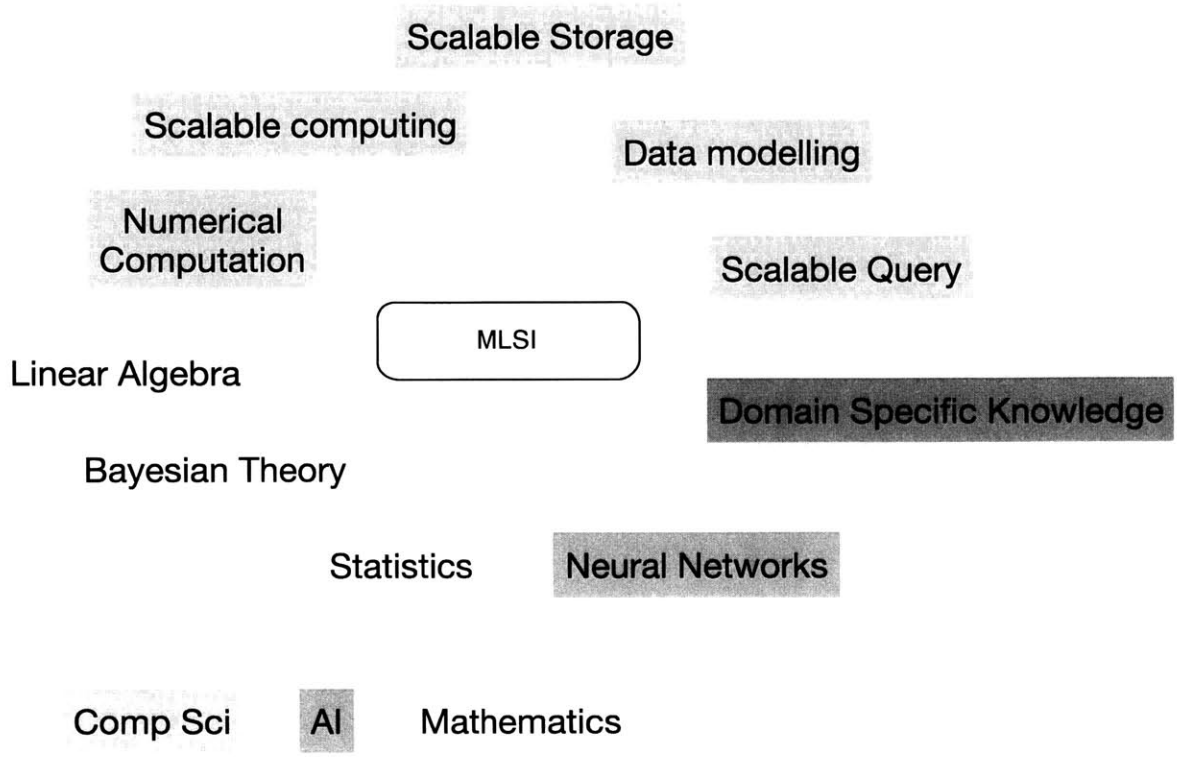

Figure 3 Multi disciplinary science

To implement an effective MLSI contributions from various fields may be required.

Some of the key fields include:

Computer Science

- Scalable Storage: Storing efficiently large amounts of data is increasingly becoming commoditized but strong latency requirements for processing or cost effectiveness can require deep skills in implementation of scalable storage
- Scalable Computing: With increasing data and large scale processing requirements many MLSI implementations can appear as powerful as super computers of a few years back; knowledge of scalable computing

can be a must have if off the shelf tools do not provide adequate performance to meet business requirements.

- Data modeling: One of the primary challenges of machine learning is the ability to present knowledge and create an effective unified data model after combining data from various sources; this is a core requirement for being able to implement a machine learning system.
- Scalable Query: Being able to query large sets of modeled data often requires functional understanding of how to implement queries across data that can scale.

Business Insight

- Domain Specific knowledge: This out of all the disciplines is the most important, as good insight into the business domain can greatly vary the question posed to the machine learning system and the effectiveness of the outcome; a good example is the fact that the starting point of the building a company such as Google was not the implementation of the scalable solution to serve billions of queries but rather the insight that lead to the page ranking algorithm. Another example of this is the insight that fraud detection will play in important role in online currency exchange by PayPal which enabled them to establish business feasibility over the competition (Schwartz, 2001).

Math and Artificial Intelligence

The following are the core building blocks of machine learning algorithms. This list by no means exhaustive but for the purpose of this thesis we do not need to dive deeper into the inner details of how these work.

- Neural Networks
- Statistics
- Bayesian Theory
- Linear Algebra

To meet a business need, implementing an MLSI can involve building core capabilities in several of the disciplines listed above. An MLSI can impose requirements on some of these fields that are cutting edge for the field itself. As an example, fraud detection for online financial transactions often has low latency needs and it has to scale to hundreds of millions of user profiles; the challenges around low latency processing of large data is not unique to machine learning but implementing machine learning in this context can increase the implementation complexity of the systems.

Another orthogonal example for the multidimensionality is voice recognition where domain specific knowledge can help better select the algorithms and customize the implementation (Ex: voice recognition only for movie titles for a home entertainment system, this domain knowledge or corpus can be used to improve the probabilistic model that is part of the MLSI).

3.5. Data and Domain Knowledge

For a machine learning system to appear to have gained wisdom, it typically needs to act over data where the number of variables that can impact the

outcome are large enough that enough samples cannot be collected to reconstruct the relation between all the variables. Thus, without domain knowledge to assist in feature selection and help make assumptions, machine-learning systems will not perform better than flipping a coin to guess the outcome for a given input data (NFL theorem) (Domingos, 2012).

On the flip side, when the dimensionality of the data is high, intuition about the relation between the variables is typically hard to get right. (Curse of dimensionality) (Domingos, 2012).

Thus, the quality of domain knowledge, intuition about the data and creativity can have a significant impact on the effectiveness of an MLSI.

One important consideration to note here is that often the presence of large data sets tends to help the effectiveness of an MLSI more than variations in a specific choice of algorithm. (Peter Norvig (Director of Research, 2009) (Alon Halevy, 2009) (Domingos, 2012). This again has to do with the dimensionality of data.

3.6. Variations in MLSI

The science behind machine learning is fairly rich and it often implies that there is more than one way to build a machine learning system. The context in which the MLSI is built can often play a very important role in the selection of the form elements and the effectiveness of the system. As an example we take a look at spelling correction and grammar correction systems. Traditionally, in word editors, significant effort was put it in to build models for how to detect errors (spelling and grammar) and the likely corrections. Word editors would adapt to some extent based on the input of a user. Traditional NLP based systems are

outperformed by statistical systems that have no insight into the language but build a statistical model of commonly occurring spellings and grammatical structures. (Peter Norvig (Director of Research, 2009) (Alon Halevy, 2009). These approaches are possible in connected devices and services that have access to be able to process a large corpus of data for each request. These systems also are able to adapt more rapidly with the emergence of new words and phrases.

Thus, to solve the same problem at hand, the effectiveness of machine learning systems is largely impacted by the context in which the problem is to be solved in addition to the state of art of the implementation itself.

3.7. Implications

So far we have tried to explain what machine learning is and define its form elements. We then proceeded to highlight the complexity of building a machine learning system, starting with the multi disciplinary nature of building an MLSI. This was followed by a look at the nature of data and the importance of domain expertise and how the same problem can be solved with different effectiveness based on the context around the MLSI.

We will now try to highlight some of the implications to a business owner.

3.7.1. Performance envelope

Each MSLI has a performance envelope that characterizes how well it performs on a few key parameters and that is unique to it. Tweaking the MLSI will only move the MLSI along it's performance envelope. Newer and better performance envelopes can be defined by either or both gaining access to better data or

through better implementation of the entire MLSI stack. (Domingos, 2012) (Alon Halevy, 2009).

The emergent system behavior can be changed by changing the form elements in addition to looking into new choices in algorithms of machine learning. As highlighted in the discussion of variations of the MLSI changing the context of the solution can enable newer approaches for solving the same problem and define a new performance envelope that can disrupt the existing solution in place.

Thus each business depending on an MLSI should be aware of the performance envelope of its MLSI and the potential to define a new performance envelope.

3.7.2. Experimental

The process of deploying and improving an MLSI is experimental and dynamic; organizations need to have processes in place to allow for continued experimentation and validation of the experiments against business goals.

For various level 1 form elements more than one option exists to implement an MLSI and it often needs experimentation to get the desired output. In addition, as the nature of the available data changes and new approaches emerge, the implementation decisions made that define the MLSI need to be revisited.

Figure 4 Machine learning algorithm cheat sheet (Mueller, 2013), is an example of the kind of heuristics that needs to be evaluated for just selecting the algorithm for a single MLSI. There are other ensemble approaches where more than one MLSI is combined, in which multiple teams can perform experimentations and their outcome is combined to form the end MLSI.

For level 2 form elements the constraints on scale and latency of processing can lead to a need for more investment into experimentation with various technology choices.

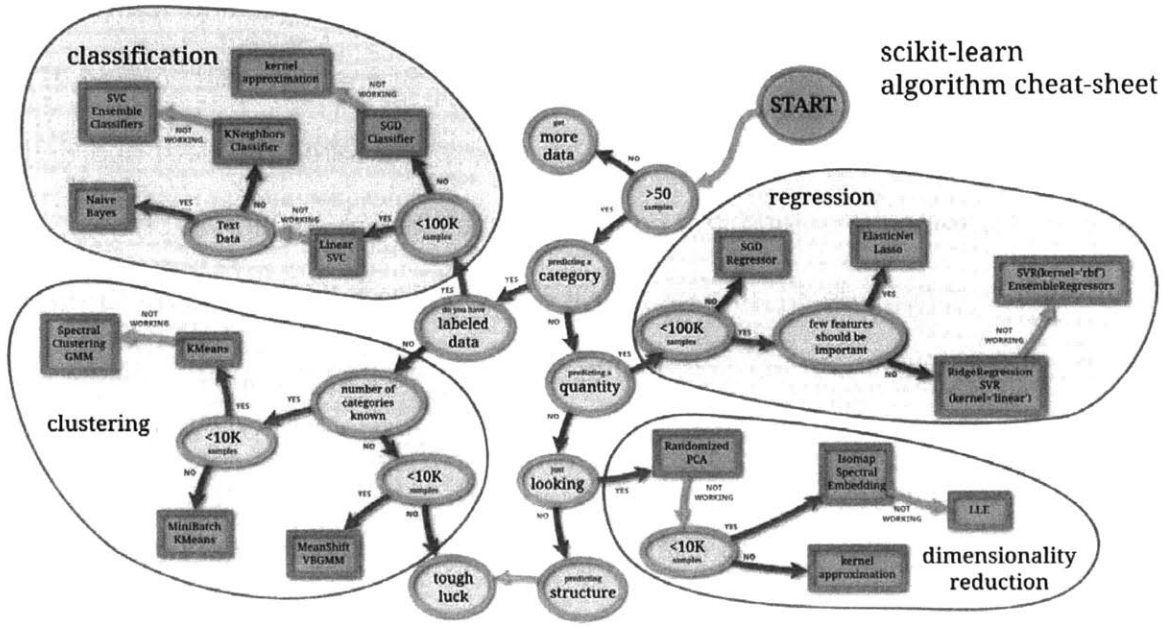

Figure 4 Machine learning algorithm cheat sheet

3.7.3. Continuous evolution

Deploying an MLSI is not a static process and is more dynamic than deploying other traditional IT solutions. Organizations must be willing to integrate continuous measurement of the quality of MLSI outcome and accommodate the need for continuous tweaking of the implementation to match the dynamic nature of the data.

As an example fraud detection of transactions needs to be able to accommodate for the changing nature of customers and businesses, and the changes in the kinds of frauds that emerge. These need to be balanced with the business goals for false negatives, where a legitimate transaction gets stopped. These heuristics

are not static and the MLSI deployed will continue to evolve with the nature of data and the improvements desired.

3.7.4. Interpretability and transparency

Effective MLSI does not imply that the solution is interpretable and transparent; often the process of defining a MSLI and the choices made to reach desired output is done via dynamic algorithms that do not have a mapping directly to the business goal itself, but they only mimic the desired outcome. This can be a challenge if the business needs an implementation that can be explained in a clear and transparent manner. This is relevant when there is legal liability for the output of an MLSI. (Breiman, Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), 2001)

4. Strategy for machine learning

An organization can choose to adopt machine-learning systems in various ways. These can be broadly classified along two lines (Porter, 1996)

1. Do existing tasks better (Operational efficiency)

   Some of the examples of this are point solutions such as detecting credit card fraud, improving scheduling of resources, etc. The core of the business is assisted by the operational efficiency but machine learning is not the core itself.

2. Enable new business opportunities (Competitive strategy)

   There is an increasing trend of companies who fundamentally depend on machine learning for their core business. The biggest example of this is online search. Some other examples are music recommendation services such as Spotify and Pandora.

There are many factors that go into making a business successful but from a machine learning perspective there are two primary levers. (Domingos, 2012)

4.1. Primary Levers

4.1.1. Nature of data

There are 3 factors to qualify the nature of data:

1. Quality of data
2. Quantity of data
3. If the access to data can serve as a competitive advantage.

Machine learning typically needs to operate on data that has high dimensionality, having more data which has samples that capture the sample space effectively can lead to a better performance envelope; if this data is available exclusively for an MLSI, it can establish a sustainable competitive advantage (Porter, 1996).

A good example of this is the exclusive access that Facebook has over the social graph it develops based on the profile of the user, their connections and their online activity. This is good quality data at scale for which Facebook has exclusive access. Facebook then develops various machine-learning systems on top of this data to serve ads, to recommend pages and to suggest connections. (Facebook) (Eric Sun, 2013)

Twitter recently acquired a company Bluefin Labs (Twitter, 2013), this gives twitter the opportunity to have exclusive access to data on what their users are posting about live television programming. This is an example of creating a unique data set that is not available to traditional companies such as Neilsen who rate TV content (Nielsen, 2012).

4.1.2. Business relevance of the performance of an MLSI.

The ability of a machine learning system to act on the data for a given hypothesis depends on the quality of the implementation of the overall system. This includes all aspects of the system right from data curating, feature selection through to meeting latency needs for a given business application and the visual presentation of the results to users of the system.

Counter to the example of having exclusive access to data, the quality of the implementation can serve as a competitive advantage when the access to data is

not exclusive and the implementation of an effective MLSI is not trivial. As an example, when Google become the dominant search engine, the essential data used on the web was accessible to other search engines as well but the entire system implemented by Google including the core algorithm, was able to create a sustainable competitive advantage.

Figure 5 Levers for Strategy, shows a sample of some known domains in which machine learning is applied and the alignment along the two dimensions identified above. This is based on secondary research and our understanding of the their business models and their access to data for various domains. Businesses around music recommendation often compete based on the number of songs available and the specifics around features and pricing but not purely on the merit of their recommendation engine; whereas for voice recognition systems access to large and growing set voice samples is hard and the quality of the recognition done by the MLSI can play a vital role in creating a meaningful differentiation.

This also helps us understand why a company such as Google has a lot invested into building core capabilities around machine learning, since the data itself that it operates on is publicly available, while the quality of the MLSI plays the dominant role for strategic differentiation. This also explains why data from social networks tends to be closed in nature, since their primary strategic differentiation is based on the connection graph that they are able to build.

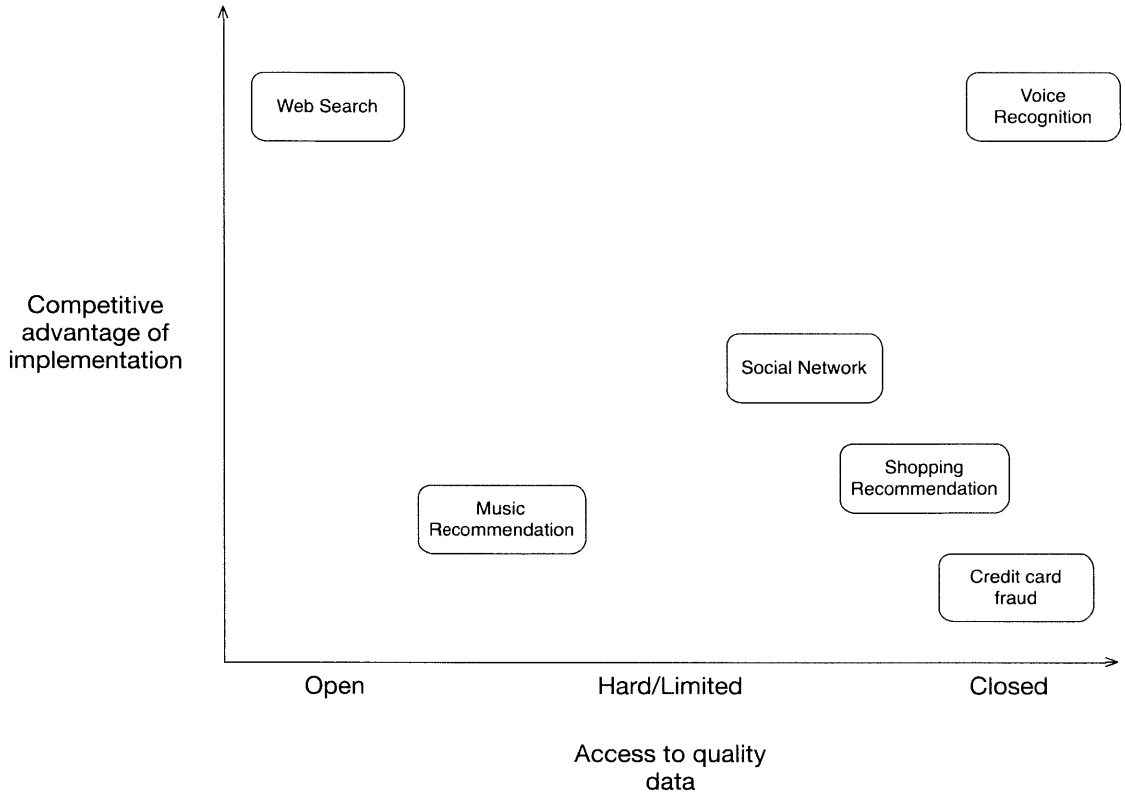

Figure 5 Levers for Strategy

5. Machine learning ecosystem

5.1. Growth in ecosystem

Over the past years the number of companies and projects that support implementing machine learning systems has steadily increased with more growth coming in the past 10 years than the previous 20 years.

The following table does not take into account the growth that has occurred within established platforms such as R and Matlab around machine learning. This commercial as well as open source based growth is a good indication for the level of interest and the early ferment stage of machine learning.

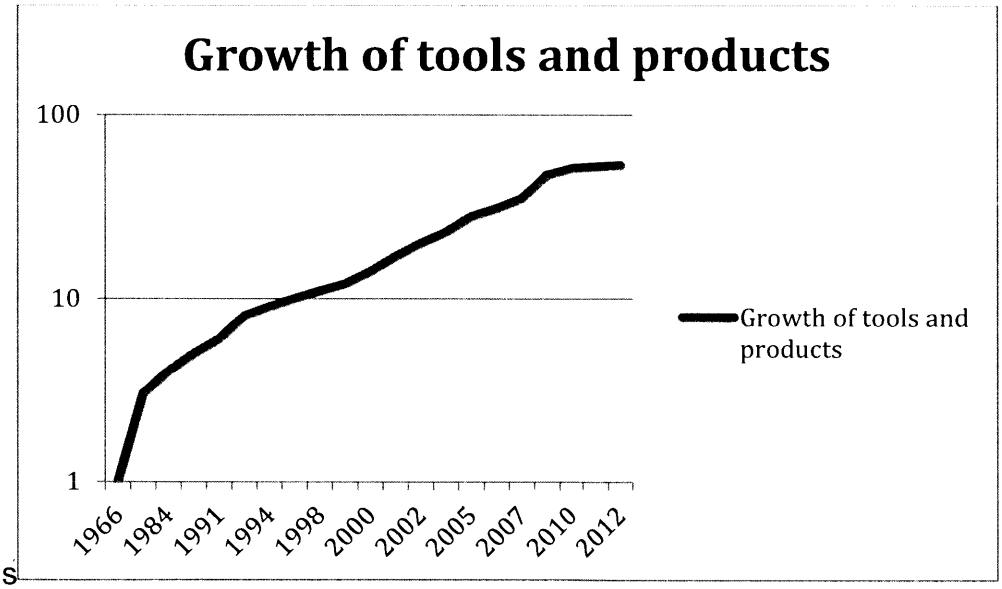

Figure 6 Exponential growth of machine learning tools

One indirect way to measure the number of companies adopting machine learning is to look at the trend for job postings. (Indeed) (ITJobsWatch)

Figure 7 Job growth for python and machine learning

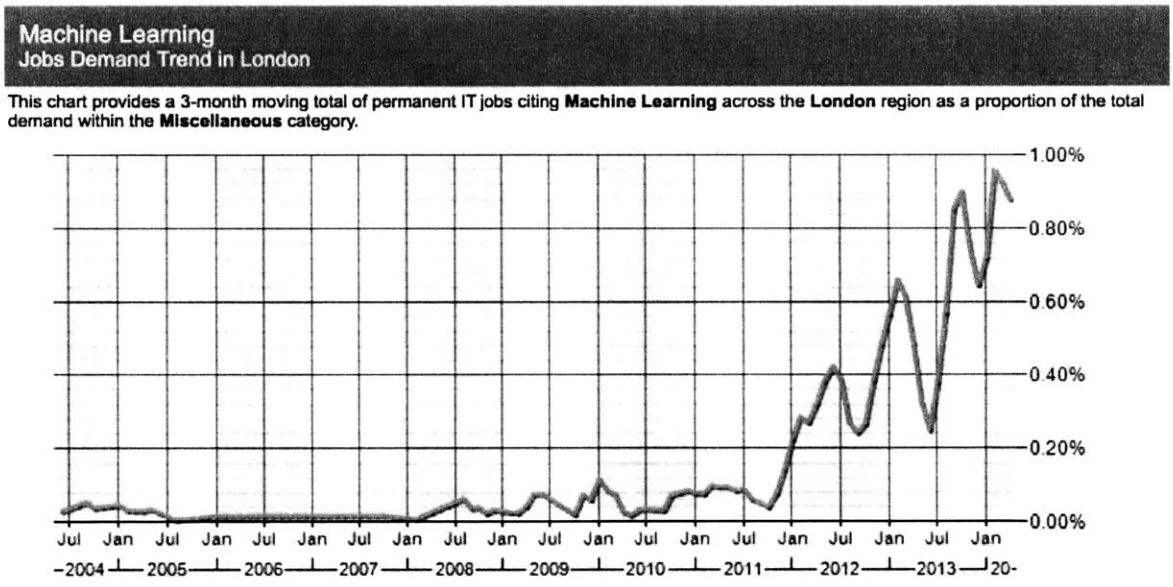

Figure 8 Machine learning job posting in London

As highlighted in the overview of machine learning, it is experimental in nature and continues to evolve. This is not just limited to implementing an MLSI but the academic domain as well has been active in continuing research related to machine learning.

The graph below (Bulatov) shows the trend in papers indexed by Google Scholar that has the term machine learning.

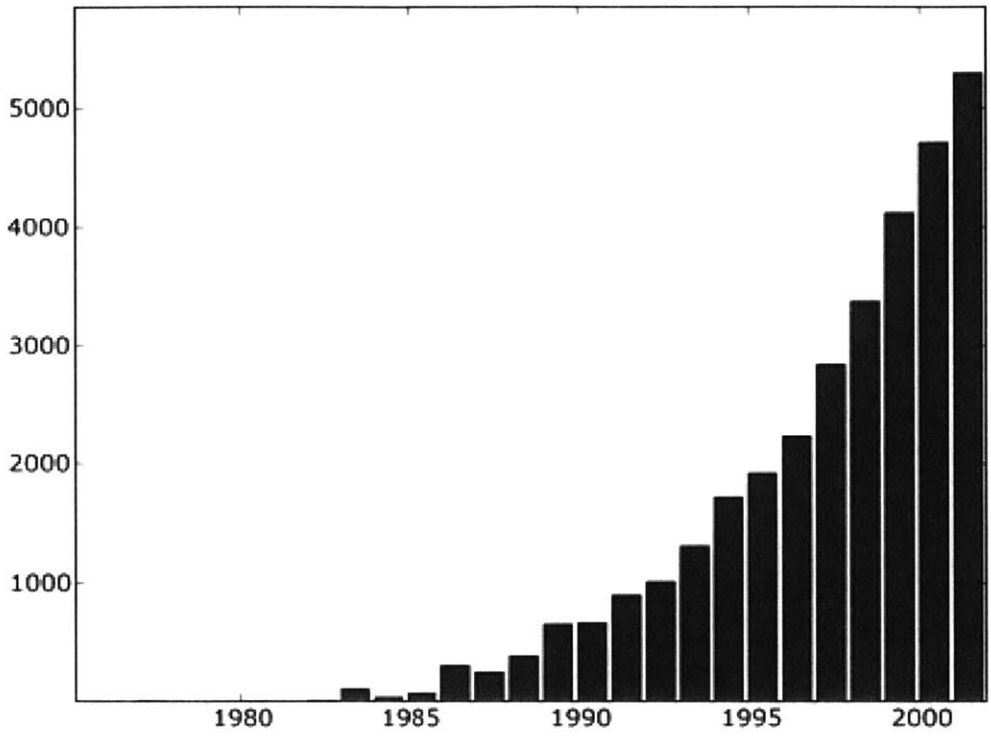

Figure 9 Papers containing the phrase Machine Learning

5.2. Modular stack

Traditionally machine-learning stacks were composed of monolithic and integral stacks sold by corporations such as SAS, Salford Systems or vertically aligned solutions from big name vendors. Over the past 10 to 15 years the stack has grown to be more modular.

The modular stack can be broken down into the following three components.

5.2.1. Core Infrastructure

With the advent of cloud computing the cost to experiment with large data sets has decreased by removing the capital expenditure from the cost. While at the same time, a rich ecosystem has developed around cloud services that makes it easier to experiment, with many ready to use tools available, that reduce the total effort taken; as an example, managing large clusters would often require custom development but these days open source projects such as zookeeper (Apache Software Foundation) serve as a clean module within the MLSI that needs to process large amounts of data.

5.2.2. Sources of data

The reducing cost of storage and better connectivity, in addition to recognition that "Data is the new oil" (Rotella, 2012) have increased the amount of data that is stored and shared within and across companies. There are many data repositories that are available (Microsoft) (Infochimps) in addition to companies opening up their data via APIs that can be remotely accessed. (Facebook)

5.2.3. Tools to implement MLSI

A level 1 decomposition of MLSI shows the various components that constitute the MLSI for each level there are now production ready open source tools that be used (Infoworlds) (Data Tamer). List of machine learning companies and projects in the Appendix lists major companies and open source projects that cater specifically to machine learning tools.

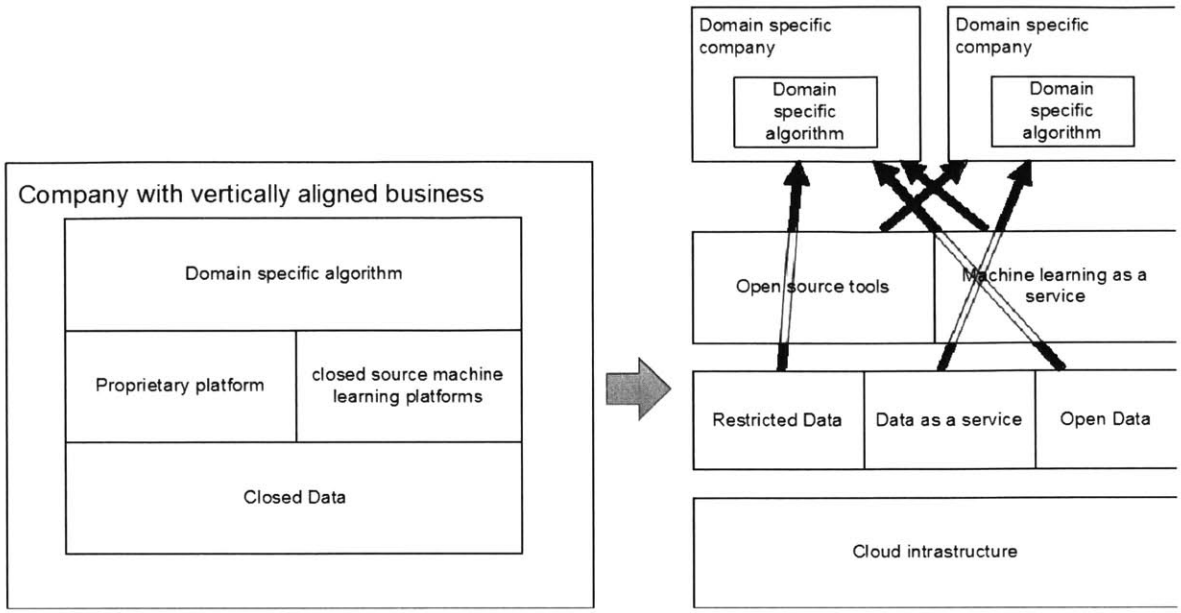

Figure 10 Move towards a modular architecture stack

5.3. Open source based tool chain

One of the core components of building an MLSI is the core algorithm around MLSI. A large percentage of the growth in machine learning tool chains has been in the open source domain. The open source tool chain itself has seen rapid improvements in features and maturity.

Projects such as Mahout, Weka, R and Python based toolkits emerge as popular choices for companies starting to adopt machine learning. Open source and an active community help in experimenting with machine learning and lower the cost of entry into machine learning.

5.4. Adoption Pattern

The adoption of machine learning into companies has been epidemic in nature, with a technology pull model, with no single set of vendors who are predominantly pushing for machine learning techniques. (Geroski, 1999)

Though there is a latent need within companies to adopt machine-learning techniques for both operational efficiency as well as new business opportunities, the adoption of machine learning has been impaired due to lack of awareness and understanding of what it involves and the lack of dynamic capabilities required to effectively adopt and to implement it.

5.5. Organizational capabilities

Machine learning as a discipline is over subscribed for talent (Manyika, et al., 2011) and there are some common trends to address the gap, some more traditional than others. Depending on the business needs, the newer more unorthodox approaches might be applicable

5.5.1. Internal

One option is to build the core competency internally. This is a requirement for companies for whom the MLSI implementation is core to their business strategy and competitive differentiation. Companies such as Google, Facebook, and Microsoft continue to invest heavily into their internal capabilities for machine learning. Internal core capabilities in an environment in which there is a shortage of suitable talent can prove to be expensive financially and is justified if the strategy requires pushing the performance envelope for the MLSI that needs to be implemented.

5.5.2. External

Outsourcing the MLSI implementation might be a preferable option for those MLSI instances that are used as a tool for operational efficiency; or if there are

other avenues to establish sustained competitive edge in spite of MLSI being core to the business strategy. An example of this is Falcon (FICO) that is used as a vertically integrated MLSI for credit card fraud detection. The reason for this to be an option is that credit card companies do not stand to differentiate significantly by additional efficiency that may be possible if they were to internalize the MLSI implementation, and they have other means to establish strategic competitive edge for their business.

5.5.3. Crowd source

This is a new trend that has emerged and has been adopted by companies that have problems that can be solved by machine learning solutions. They are offered up as challenges (Kaggle) to explore the performance envelope possible for an MLSI implementation. Often these are one-off problems of high value or companies choosing to explore a new solution for an old and intractable problem.

5.5.4. Active waiting

There are many industries where rapid adoption of technology is not possible or optimal due to varying business and/or regulatory needs. These companies can choose to wait and watch and be prepared to pursue one of the solutions above when they feel that the overall ecosystem is ripe for them to adopt.

6. Conclusion

Machine learning as a technology will continue to rise in relevance for forming sustainable strategic differentiation for businesses. As businesses adopt machine learning, they will need to reflect on the relation between the data they can access, desired MSLI and the business needs. This relation will define the directions and levels of investment; directions can include one or more of the following, sourcing better data, accessing core competency to build an MLSI with the desired performance envelope, building better domain expertise or investing into the ecosystem to define a new performance envelope that does not exist yet. We see from the nature of data, its relation to domain knowledge and the variations in MLSIs that the performance of an MLSI depends on two primary dimensions:

1. Quality of data
2. The implementation quality of the MLSI

Thus, in the analysis of the levers for machine learning, we conclude that a business choosing to build a strategy must identify and align the strategy along two dimensions.

1. Access to quality data
2. Relevance of MLSI performance to the effectiveness of the strategy.

Companies operating over publicly available data have to build a better MLSI to compete whereas companies with access to exclusive data might be able to build a competitive advantage based purely on the data. Ideally a company would like

to have exclusive access to data over which they can form a meaningful strategic differentiation based on the quality of the MLSI.

The ecosystem around machine learning itself is growing at a rapid rate with a more modular structure with open source tools playing an important role.

Thus, the combination of dynamic nature of data, the state of art of science and rapid and modular development of the ecosystem, businesses relying on MLSI to form the core part of their strategy face the threat of being disrupted by an alternate MLSI being implemented by a competitor that has a better performance envelope than theirs. Thus, companies need to embrace the experimental and dynamic nature of machine learning and continue to build core competency to help them stay on top of the evolutionary curve of machine learning as well as define new performance envelopes to preserve their strategic advantage. Companies internally will have to embrace the dynamic nature of machine learning and it's ecosystem.

Lastly, for certain businesses a more active waiting adoption strategy might be in place either due to the inability to address dynamic capabilities to align well to the two primary dimensions elaborated above or due the fact that they need transparency and interpretability and these cannot be easily accommodated.

7. Appendix

7.1. List of machine learning companies and projects

Name	Year started	Company/Open Source
SAS	1966	company
Matlab	1983	company
Salford Systems	1983	company
Angoss	1984	company
Octave	1988	open
STATISTICA	1991	company
Weka	1993	open
R	1993	open
SPSS	1994	company
Orange	1996	open
KXEN	1998	company
OpenCV	1999	open
ParadisEO	2000	open
LIBSVM	2000	open
SciPy	2001	Python
NLT	2001	Python
RapidMiner	2001	open/closed
dlib	2002	open
ODM	2002	company

torch3	2002	open
MDP	2004	Python
KNIME	2004	open
ECJ	2004	open
matplotlib	2005	Python
LaSVM	2005	open
OpenOpt	2005	open
Tiberius	2005	company
Waffles	2005	open
NumPy	2006	Python
Shogun	2006	open
torch5	2006	open
Scikit	2007	Python
nieme	2007	open
treparel	2007	company
CVXOPT	2008	Python
PANDAS	2008	Python
PyMVPA	2008	Python
Pybrain	2008	Python
mlpy	2008	Python
Ayasdi	2008	company
Mahout	2008	Hadoop/open
ELKI	2008	open

MCMLL	2008	open
LIBLINEAR	2008	open
Armadillo	2008	open
java-ml	2008	open
Wapiti	2010	open
MyMediaLite	2010	open
CTBN-RLE	2010	open
MOEA	2011	open
LIONsolver	2012	company
Darwin	2007	open
MOA Massive Online Analysis	2009	open
Skytree	2012	company

8. Works Cited

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science , 33* (2), 163-180.

Alon Halevy, P. N. (2009). The Unreasonable Effectiveness of Data . *IEEE Intelligent Systems , 24* (2), 8-12.

Apache Software Foundation. (n.d.). *Zookeeper.* Retrieved 5 5, 2014, from Zookeeper: http://zookeeper.apache.org/

Baldwin, Y. C., & Clark, B. K. (1997, October). Managing in an Age of Modularity. *Harvard Business Review , 75* (5), pp. 84–93.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science , 16* (3), 199-231.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science , 16* (3), 199-231.

Bulatov, Y. (n.d.). *Trends in Machine Learning according to Google Scholar.* Retrieved 4 22, 2014, from Machine Learning, etc: http://yaroslavvb.blogspot.com/2005/12/trends-in-machine-learning-according.html

Carla E. Brodley, U. R. (2012, Spring). Challenges and Opportunities in Applied Machine Learning. *Applied Machine Learning AI* .

Christensen, C. M. (1992, December). Exploring the Limits of the Technology S-curve, Part 1: Component Technologies. *1* (4), pp. 334–357.

Christensen, C. M. (1992). Exploring the Limits of the Technology S-curve, Part 2: Architectural Technologies. *Production and Operations Management , 1* (4), 358–366.

Crawley, E. (2009). ESD.34 System Architecture. *MIT SDM* .

Data Tamer. (n.d.). *http://www.data-tamer.com/.* Retrieved 5 5, 2014, from http://www.data-tamer.com/: http://www.data-tamer.com/

Davies, M. A. (2009, Jan - April). 15.965 Technology Strategy. Cambridge, MA, USA: MIT.

Davies, M. A. (2013, Sept-Dec). ESD.39 Systems, Leadership & Management Lab (SLaM Lab). Cambridge, MA, USA: MIT.

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM , 55* (10), 78-87.

Eisenhardt, K., & Jeffrey, M. A. (200). Dynamic Capabilities: What are They . *Strategic Management Journal , 21*, 1105-1121.

Eric Sun, V. I. (2013, 6 6). *Under the Hood: The Entities Graph*. Retrieved 3 1, 2014, from Facebook.com: https://www.facebook.com/notes/facebook-engineering/under-the-hood-the-entities-graph/10151490531588920

Everett, R. M. (2003). *Diffusion of Innovations.* New York: Free Press.

Facebook. (n.d.). *Graph API*. Retrieved 3 1, 2014, from Facebook.com: https://developers.facebook.com/docs/graph-api/

FICO. (n.d.). *Falcon Fraud Manager.* Retrieved 4 14, 2014, from http://www.fico.com/: http://www.fico.com/en/products/fico-falcon-fraud-manager/

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. New York, NY, USA: Cambridge University Press.

Geroski, P. A. (1999). Models of Technology Diffusion. *C.E.P.R. Discussion Papers , 2146*.

Indeed. (n.d.). *Job trends*. Retrieved 4 22, 2014, from Indeed.com: http://www.indeed.com/jobtrends?q=python+and+machine+learning&l=&relative=1

Infochimps. (n.d.). *Datasets*. Retrieved 5 5, 2014, from http://www.infochimps.com/: http://www.infochimps.com/datasets

Infoworlds. (n.d.). *7 top tools for taming big data* . Retrieved 5 5, 2014, from http://www.infoworld.com/: http://www.infoworld.com/d/business-intelligence/7-top-tools-taming-big-data-191131?page=0,1

ITJobsWatch. (n.d.). *Machine learning jobs in London*. Retrieved 4 22, 2014, from www.itjobswatch.co.uk: http://www.itjobswatch.co.uk/jobs/london/machine%20learning.do

Kaggle. (n.d.). *Competitions.* Retrieved 4 14, 2014, from www.kaggle.com: https://www.kaggle.com/competitions

Kart , L., Heudecker, N., & Buytendijk, N. (2013). *Survey Analysis: Big Data Adoption in 2013 Shows Substance Behind the Hype.* Gartner. Gartner.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. McKinsey.

Microsoft. (n.d.). *Windows Azure Marketplace*. Retrieved 5 5, 2014, from Azure.com: http://datamarket.azure.com/

Mueller, A. (2013, 1 2013). *Machine Learning Cheat Sheet (for scikit-learn).* Retrieved from Andy's Computer Vision and Machine Learning Blog: http://peekaboo-vision.blogspot.com/2013/01/machine-learning-cheat-sheet-for-scikit.html

Nielsen. (2012, 12 17). *NIELSEN AND TWITTER ESTABLISH SOCIAL TV RATING*. Retrieved 3 1, 2014, from http://www.nielsen.com/: http://www.nielsen.com/us/en/press-room/2012/nielsen-and-twitter-establish-social-tv-rating.html

Peter Norvig (Director of Research, G. (2009, september). Innovation in Search and Artificial Intelligence . MERCED, CA, USA.

Porter, M. (1996, December). What is Strategy. *Harvard Business Review* , pp. 61-78.

Robinson, S. (2014, January 4). *Prepare for the challenges of deploying machine learning for analytics.* Retrieved May 4, 2014, from http://www.techrepublic.com: http://www.techrepublic.com/blog/big-data-analytics/prepare-for-the-challenges-of-deploying-machine-learning-for-analytics/#.

Rotella, P. (2012, 4 2). *Is Data The New Oil?* Retrieved 12 25, 2013, from http://www.forbes.com/: http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/

Schwartz, E. I. (2001, December 1). *Digital Cash Payoff.* Retrieved December 13, 2013, from MIT Technology Review: http://www.technologyreview.com/featuredstory/401297/digital-cash-payoff/

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data.* Wiley.

Suarez, F. F. (2004). Battles for technological dominance: An integrative framework. *Research Policy , 33*, 271-286.

Sull, D. (2005, September). Strategy as an active waiting. *Harvard Business Review , 83* (9), pp. 120-129.

Teece, D. J., Gary, P., & Shuen, A. Dynamic Capabilities and Strategic Management. *Strategic Management Journal , 18* (7), 509-533.

Twitter. (2013, 2 5). *Welcoming Bluefin Labs to the Flock.* Retrieved 3 1, 2014, from Twitter.com: https://blog.twitter.com/2013/welcoming-bluefin-labs-to-the-flock

Wikipedia. (n.d.). *Page Rank.* Retrieved 4 12, 2014, from Wikipedia.org: http://en.wikipedia.org/wiki/PageRank

Witten, H. I., & Eibe, F. (2005). *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems).* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.