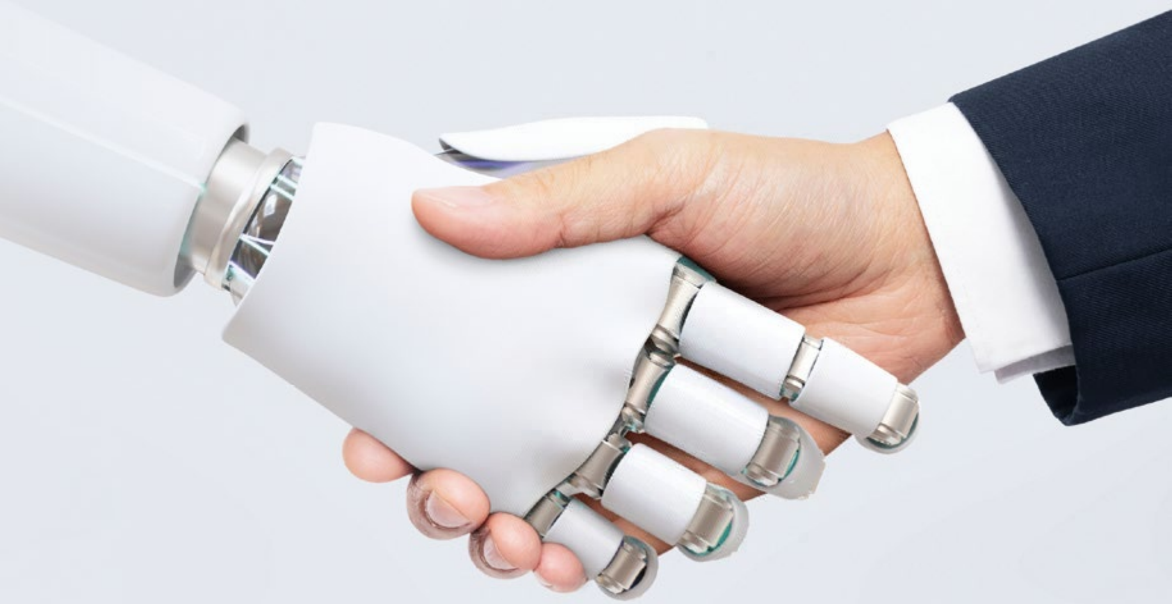

The Decision Maker's Handbook to Data Science

AI and Data Science for Non-Technical Executives, Managers, and Founders

Third Edition

Stylianos Kampakis

Apress®

THE DECISION MAKER'S HANDBOOK TO DATA SCIENCE

AI AND DATA SCIENCE FOR NON-TECHNICAL EXECUTIVES, MANAGERS, AND FOUNDERS

Third Edition

Stylianos Kampakis

Apress®

The Decision Maker's Handbook to Data Science: AI and Data Science for Non-Technical Executives, Managers, and Founders

Stylianos Kampakis
London, UK

ISBN-13 (pbk): 979-8-8688-0278-2 ISBN-13 (electronic): 979-8-8688-0279-9
https://doi.org/10.1007/979-8-8688-0279-9

Managing Director, Apress Media LLC: Welmoed Spahr
Acquisitions Editor: Shivangi Ramachandran
Development Editor: James Markham
Editorial Assistant: Jessica Vakili

Cover designed by eStudioCalamar
Cover image designed by Freepik (www.freepik.com)

Distributed to the book trade worldwide by Springer Science+Business Media New York, 1 New York Plaza, Suite 4600, New York, NY 10004-1562, USA. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail booktranslations@springernature.com; for reprint, paperback, or audio rights, please e-mail bookpermissions@springernature.com.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at http://www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub. For more detailed information, please visit https://www.apress.com/gp/services/source-code.

If disposing of this product, please recycle the paper

Contents

About the Author

Dr. Stylianos (Stelios) Kampakis is a data scientist and tokenomics expert with more than ten years of experience.

He has worked with companies of all sizes: from startups to organizations like the US Navy, Vodafone, and British Land. His work expands multiple sectors, including fintech, sports analytics, health-tech, general AI, medical statistics, predictive maintenance, and others.

He has worked with many different types of technologies, from statistical models to deep learning to large language models. He has two patents pending to his name and has published three books on data science, AI, and data strategy.

He has helped many people follow a career in data science and technology.

His seminal work in token economics has led to many successful token economic designs using tools such as agent-based modelling and game theory.

He is a member of the Royal Statistical Society, honorary research fellow at the UCL Centre for Blockchain Technologies, the Cyprus Blockchain Technologies, a data science advisor for London Business School, and CEO of the Tesseract Academy. Stylianos is also very active in the area of data science education. He is the founder of the Tesseract Academy, a company whose mission is to help decision-makers understand deep technical topics such as machine learning and blockchain. He is also teaching "Social Media Analytics" and "Quantitative Methods and Statistics with R" in the Cyprus International Institute of Management and runs his own data science school in London called Datalyst.

He often writes about data science, machine learning, blockchain, and other topics at his personal blog: The Data Scientist (thedatascientist.com).

CHAPTER

1

Demystifying Data Science and All the Other Buzzwords

In the business world, data has become a big thing. You hear all sorts of buzzwords being thrown around left, right, and center. Things like *big data*, *artificial intelligence*, *machine learning*, *data mining*, *deep learning*, and so on. This can get confusing, leaving you paralyzed as to what is the best technology to use and under what circumstances. In this chapter, we are going to demystify all these buzzwords, by taking a short stroll through the history of data

S. Kampakis, *The Decision Maker's Handbook to Data Science*,
https://doi.org/10.1007/979-8-8688-0279-9_1

science. You will understand how the history of data analysis gave birth to different schools of thought and disciplines, which have now all come together under the umbrella of the term **data science**.

What Is Data Science?

In 2017, we generated more data than we did over the previous 5000 years of our history.[1] That's a lot of data. And it's not surprising. Every device we own generates data, and all our interactions with said devices generate even more data.

So, you take a picture with your smartphone? You've generated data. You read the news on your tablet? You're generating data. You listen to a podcast on your laptop? You're generating data. You go on Facebook to update your status? You've generated data.

You get the point. There's a good chance that the only thing we do that doesn't generate data is breathe, but even that is debatable considering all the wearable devices that are available and which can track everything from heart rate to calories burned.[2]

What happens to all this data? It must have some use or things wouldn't be set up so that we generate it in the first place.

At the moment, some of it does and some of it doesn't. In fact, some studies have discovered that less than 0.5% of data is being analyzed and turned into actionable insights.[3] But more on that later.

The important thing is that all data *can* be used. We just need to figure out better ways of doing so.

And this is where data science comes in. There are many definitions of data science. One such definition is

> *Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.*[4]

[1] David Sønstebø, "IOTA Data Marketplace," *IOTA*, November 28, 2017, https://blog.iota.org/iota-data-marketplace-cb6be463ac7f.
[2] www.wired.co.uk/article/hospital-prescribing-tech
[3] www.technologyreview.com/s/514346/the-data-made-me-do-it/
[4] Chikio Hayashi, "What is Data Science? Fundamental Concepts and a Heuristic Example," in *Data Science, Classification, and Related Methods*, eds. Chikio Hayashi, Keiji Yajima, Hans-Hermann Bock, Noboru Ohsumi, Yutaka Tanaka, and Yasumasa Baba (Tokyo, Japan: Springer-Verlag, 1998), 40–51.

Okay, so those are a lot of big and fancy words, which can get confusing. Let's boil it down to the simplest definition:

Data science is about using data to do useful stuff.

Short and to the point. That's exactly what data science is all about. The methods we use are important, of course, but the essence of this discipline is that it allows us to take data and transform it so we can use it to do useful things.

For example, let's say someone visits a doctor because they are short of breath. They are experiencing heartburn and they have chest pains. The doctor will run the basic tests, including measuring blood pressure, but nothing seems out of the ordinary.

Since the patient is overweight, the doctor immediately assumes the symptoms are caused by the patient's size and recommends a healthier diet and exercise.

Three months later, the same patient is brought into the emergency room and ends up dying on the table because of a heart defect.

This might sound like an episode on your favorite medical show, that is, a work of fiction, but it happens much more often than you might think. In fact, in the USA, 5% of patients are misdiagnosed, while misdiagnosis cost the UK over £197 million in the 2014/2015 fiscal year.[5]

However, this situation can be avoided thanks to data science. The analysis of similar cases reveals that the symptoms our patient exhibited aren't just caused by obesity but also by some cardiovascular conditions.

Having access to systems that can analyze data and compare it with new data inputs could have helped the doctor identify the problem sooner without relying solely on their own personal experience and knowledge.

So, data science can be used to save lives, among many other applications, which is pretty useful.

Data Science Is Multidisciplinary

Data science involves multiple disciplines, which is why finding someone with the necessary skills to be a data scientist can be difficult.

[5] Lena Sun, "Most Americans Will Get a Wrong or Late Diagnosis At Least Once In Their Lives," *Washington Post*, September 22, 2015, www.washingtonpost.com/news/to-your-health/wp/2015/09/22/most-americans-who-go-to-the-doctor-will-get-a-wrong-or-late-diagnosis-at-least-once-in-their-lives-study-says/; "The Top Misdiagnosed Conditions In NHS Hospitals In 2014/15," *Graysons*, www.graysons.co.uk/advice/the-top-misdiagnosed-conditions-in-nhs-hospitals/

Thus, data science involves everything from statistics and pattern recognition to business analysis and communication. It requires creative thinking as much as it requires analytical thinking.[6]

So, data science involves discovering which data is useful as well as effective ways of managing it. It also requires determining how the data should be processed and what types of insights can be garnered from the massive amounts of data available.

Data science requires knowledge of programming and computing, but also visualization so that the insights can be presented in a way that everyone can understand.

Furthermore, business acumen is also a necessity because while data science can be applied to any field of business, it is critical to know what types of answers the business needs and how to present said insights so leadership can understand them.

Core Fields of Data Science

Data science has three core fields, namely, artificial intelligence, machine learning, and statistics.

> **Artificial Intelligence** is all about replicating human brain function in a machine. The primary functions that AI should perform are logical reasoning, self-correction, and learning. While it has a wide range of applications, it is also a highly complicated technology because to make machines smart, a lot of data and computing power is required.
>
> **Machine learning** refers to a computer's ability to learn and improve beyond the scope of its programming. Thus, it relies on creating algorithms that are capable of learning from the data they are given. They are also designed to garner insights and then make forecasts regarding data they haven't previously analyzed.

There are three approaches to machine learning, namely, supervised, unsupervised, and reinforcement learning, plus some sub-fields (such as semi-supervised learning). Here, we will be talking only about supervised and unsupervised learning, since this is what is mainly used in business.

[6] Take a look at this infographic by Brendan Tierney: https://oralytics.com/2012/06/13/data-science-is-multidisciplinary/

Let's say you want to sort all your photographs based on content. In supervised learning, you would provide the computer with labeled examples. So, you'd give it a picture of a dog and label it animal. Then you'd feed it a picture of a person and label it human. The machine will then sort all the remaining pictures.

In unsupervised learning, you'd just give the machine all the photos and let it figure out the different characteristics and organize your photos.

In reinforcement learning, the machine learns based on errors and rewards. Thus, the machine analyzes its actions and their results. A good example of reinforcement learning is a rat in a maze that needs to navigate its way to a piece of cheese. The learning process that helps the rat achieve this can be implemented in a machine through reinforcement learning. This is one of the most esoteric types of machine learning.

Statistics is an essential tool in the arsenal of any data scientist because it helps to develop and study methods to collect, analyze, interpret, and present data. The numerous methodologies it uses enable data scientists to

- Design experiments and interpret results to improve product decision-making
- Build signal-predicting models
- Transform data into insights
- Understand engagement, conversions, retention, leads, and more
- Make intelligent estimations
- Use data to tell the story

Let's take a closer look at all three.

Artificial Intelligence: A Little History

In 1954, the field of AI research came into being at a workshop at Dartmouth College. There, the attendees discussed topics that would influence the field for years to come.

As we've already explained, the goal of artificial intelligence is to create a "thinking machine," that is, one that emulates human brain function. To do this, of course, one needs to understand the human mind, which is why AI is closely related to the field of cognitive science.

Cognitive science involves studying the human mind and its processes, including intelligence and behavior. Memory, language, perception, attention, emotion, and reasoning are all studied, and to gain greater understanding of these faculties, scientists borrow from other fields, including

- Linguistics
- Psychology
- Philosophy
- Neuroscience
- Anthropology
- Artificial Intelligence

Artificial intelligence played a key part in the development of cognitive science, and there was a large interplay between cognitive psychology and AI. The understanding of human cognition helped us improve our understanding of how to transfer this inside machines. Vice versa, the computational theory of the mind[7] was one of the dominant paradigms in cognitive science. According to this theory, the mind works like a computer, with processes and limited memory. While this is now considered outdated, it drove research for decades.

The AI Dream

It all started with a grand vision. Marvin Minsky, the head of the artificial laboratory at MIT who was considered the father of AI, stated that "artificial intelligence is the science of making machines do things that would require intelligence if done by men."[8] In other words, the dream was to create an intelligent machine.

So, how would one go about doing it? First, we start with intuition, that is, the human's distinct ability for logical reasoning.

Logical reasoning is, essentially, the capacity to reason based on various premises to reach a conclusion that has logical certainty. For example:

> *"If all men are mortal, and Socrates is a man, then Socrates is mortal."*

Or:

> *"If it's raining outside and I don't have an umbrella, I will get wet if I go out."*

---

[7] https://plato.stanford.edu/entries/computational-mind/

[8] Blay Whitby, *Reflections on Artificial Intelligence* (Exeter: Intellect Books, 1996).

To translate this logical deduction ability to a machine, a rule-based or symbolic approach is used. This involves humans constructing a system of rules with which the computer is programmed. Using these rules, reasoning algorithms are capable of deriving logical conclusions.

A good example is MYCIN,[9] which was an early successful working system based on reasoning algorithms. It was used to diagnose infections and determine the type of bacteria that was causing the problem. It was never used in a clinical setting but is an excellent example of an expert system and a predecessor to machine learning.

The system was developed in the 1970s at Stanford University and had approximately 600 rules.[10] Users were required to provide answers to various questions and the program would then provide a list of potential bacteria that could be causing the problem, sorted from high to low probability. It would also provide its confidence in the probability of each diagnosis as well as how it came to the conclusion. Finally, it would provide the recommended course of treatment.

It had a 69% accuracy rate and it was claimed that the program was more effective than junior doctors and on the same level as some experts.[11]

The program was created by interviewing a large number of experts who provided their expertise and experience. It used rules of the IF (condition) THEN (conclusion) form. For example, IF (sneezing and coughing or headache) THEN (flu).

One limitation the program had was computing power. It took approximately 30 minutes to go through the system, which was too much wasted time in a real-world clinical setting.

Another issue was also raised, namely, that of ethics and legal issues. Thus, the question arose of who would be held responsible if the program made the wrong diagnosis or recommended the wrong treatment.

Though it was never used, MYCIN still had a very important role in bringing us to where we are today as it was one of the early successes of AI, proving what is possible.

---

[9] A good old reference for MYCIN by John McCarthy can be found here: `www-formal.stanford.edu/jmc/someneed/someneed.html`

[10] Bruce G Buchanan and Edward H Shortliffe, *Rule-Based Expert Systems* (Reading. Mass.: Addison-Wesley, 1985).

[11] Victor L. Yu, "Antimicrobial Selection By A Computer," *JAMA* 242, no. 12 (1979): 1279, `https://jamanetwork.com/journals/jama/article-abstract/366606`

Automated Planning

Planning is a vital component of rational behavior. Automated planning and scheduling is an area of artificial intelligence that involves the creation of a system that is capable of selecting and organizing actions to achieve a certain outcome.

An example is the Missionaries and Cannibals problem,[12] which is a classic AI puzzle. It is defined as follows:

> *On one bank of a river are three missionaries and three cannibals. They all wish to cross to the other side of the river. There is one boat available that can hold up to two people. However, if the cannibals ever outnumber the missionaries on either of the river's banks, the missionaries will get eaten.*
>
> *How can the boat be used to safely carry all the missionaries and cannibals across the river?*

It's a little gruesome and might also seem trivial, but a similar approach can be used for schedule planning.

Other similar problems in that are the towers of Hanoi[13] and the traveling salesman problem. The traveling salesman problem is a legendary benchmark in optimization, where the objective is to find a path for a salesman to go across all the cities in a country (or some other geographical region). This path should never go through the same city twice, and at the same time, it should as short as possible. It is easy to see how this relates to vehicle routing in real life. In Figure 1-1, you can see an example of the traveling salesman for all major cities in Germany.

[12] You can play a version of the game here: `www.novelgames.com/en/missionaries/`
[13] `www.mathsisfun.com/games/towerofhanoi.html`

Figure 1-1. Example of the traveling salesman problem

Essentially, what the computer does is analyze each possibility, discarding the one that doesn't fulfill the parameters, while presenting all the options that do.

The AI Winters

Artificial intelligence would likely be much further along now were it not for the "winters" it experienced. The term AI winter refers to periods when interest in artificial intelligence was diminished, and, as a result, funding was limited.

In 1967, Marvin Minsky predicted that "within a generation the problem of creating 'artificial intelligence' will be substantially solved." However, by 1982, he admitted that "the AI problem is one of the hardest science has ever undertaken."[14]

[14] Frederick E. Allen, "The Myth Of Artificial Intelligence | AMERICAN HERITAGE," *Americanheritage.com*, last modified 2001, www.americanheritage.com/content/myth-artificial-intelligence

The first major AI winter started in 1974 and ended in 1980, while the second one started in 1987 and ended in 1993. There were other smaller issues, such as:

- 1966: the failure of machine translation
- 1970: the abandonment of connectionism
- 1971–1975: DARPA's frustration with the Speech Understanding Research program at Carnegie Mellon University
- 1973: the large decrease in AI research in the United Kingdom in response to the Lighthill report
- 1973–1974: DARPA's cutbacks to academic AI research in general
- 1987: the collapse of the Lisp machine market
- 1988: the cancellation of new spending on AI by the Strategic Computing Initiative
- 1993: expert systems slowly reaching the bottom and
- 1990s: the quiet disappearance of the fifth-generation computer project's original goals

The first winter was caused, in large part, by three major elements. First, it was the Lighthill report, in which professor Sir James Lighthill concluded that other sciences could do everything being done in AI, implying that many of the most successful AI algorithms would be incapable of solving real world problems.[15]

Though contested, the report still resulted in the complete shutdown of AI research in the UK, with only a few universities continuing the research. The result was funding cuts for this research all across Europe.[16]

In 1969, the Mansfield Amendment was passed in the USA, requiring DARPA to fund research that had a clear mission rather than basic projects with no clear direction. Essentially, researchers had to prove their projects would quickly produce technology that would be useful to the military.

---

[15] Lighthill, Professor Sir James (1973). "Artificial Intelligence: A General Survey." *Artificial Intelligence: a paper symposium*. Science Research Council

[16] Daniel Crevier, *The Tumultuous History Of The Search For Artificial Intelligence* (New York, NY: Basic Books, 1993); Russel & Norvig 2003; Jim Howe, "School Of Informatics: History Of Artificial Intelligence At Edinburgh," *Inf.Ed.Ac.Uk*, last modified 2007, `www.inf.ed.ac.uk/about/AIhistory.html`

In concert with Lighthill's report, which was used as proof that AI research was unlikely to provide anything useful in the foreseeable future, DARPA stopped funding this type of research. By 1974, funding for AI was virtually impossible to find.[17]

Another issue was the Speech Understanding Research program at Carnegie Mellon. DARPA wanted a system that could respond to spoken commands from a pilot, and while the team had developed a system that understood spoken English, the words had to be spoken in a certain order.

DARPA felt they had been misled and cancelled the grant in 1974.[18] Interestingly enough, by 2001, the speech recognition market reached $4 billion and used the technology the team from Carnegie Mellon developed.[19]

In 1980, the commercial success of expert systems rekindled interest in the field of AI. By 1985, over a billion dollars was being spent on AI and an industry developed to support the field. Specialized computers dubbed Lisp machines were built to better process the programming language that was preferred for AI.

Unfortunately, the LISP machine market collapsed because companies like Sun Microsystems developed more efficient alternatives. These general workstations had much better performance, which LISP machines were incapable of matching. This led to the second AI winter.

What We Learned from AI Research

Despite all the issues, research into artificial intelligence discovered many beneficial things. For example, if it weren't for AI research, we would not have algorithms that work with logical rules.

An early example of such an algorithm was the Prolog language. One of the first logic-based programming languages, it's still one of the most popular even today. It has been used to prove theorems in expert systems, in type inference, and for automated planning.

And that's another area where AI has helped us. Automated planning and scheduling, which is used to drive things like automated robots in manufacturing, would not exist otherwise. Neither would expert systems or knowledge representation.

Of course, all these things have evolved significantly since then, and we are getting closer and closer to creating true AI, but we still have a long way to go before we have a truly intelligent machine.

---

[17] National Research Council Staff, *Funding a Revolution* (Washington: National Academies Press, 1999).
[18] Crevier 1993; McCorduck 2004; National Research Council 1999.
[19] National Research Council 1999.

Another interesting aspect is that cognitive science borrows from AI as much as the reverse is true. The desire to develop a smart machine that can think like a human required a better understanding of the human mind. This led to research into memory, perception, and much more.

The Next Step: Enter Machine Learning

As a term, machine learning was coined by Arthur Samuel, who was a pioneer in computer gaming and AI, in 1959 while working at IBM.[20]

He defined it as follows:

> *"Machine learning is about giving computers the ability to learn without being explicitly programmed."*

A rift caused by the focus on logical, knowledge-based approaches developed between machine learning and AI. By the 1980s, expert systems dominated AI, and statistics was no longer of interest.

Machine learning was then reorganized as a separate field and began to gain serious traction in the 1990s. Instead of focusing on creating artificial intelligence, machine learning shifted its approach and concentrated on solving more practical problems. Thus, instead of employing the symbolic methodologies so prevalent in AI, it began to employ models taken from statistics and probability theory.

While machine learning as a concept is a subfield of artificial intelligence, its approach was practically opposite to research that had been done so far.[21]

Thus, "classic AI" or "good-old-fashioned AI" took a ruled-based, top-down approach. Essentially, it meant giving computers rules that they had to follow to reach the desired outcome. The problem with this approach is that it requires a lot of human input.

The designer of the system must be able to predict every possibility and create a set of rules for it. Furthermore, uncertainty is not easily handled because the machine thinks in an IF/THEN pattern.

Unfortunately, the world isn't always so clear-cut. In the real world, IF (EVENT A) takes place, THEN (ACTION B) isn't always the solution. For example, IF (SUNNY) doesn't always result in THEN (WEAR LIGHT CLOTHING). It could be winter out and you could need a heavy coat. Furthermore, the system only knows to provide this answer because it's been told what to do.

---

[20] Arthur L. Samuel, "Some Studies In Machine Learning Using The Game Of Checkers," *IBM Journal of Research and Development* 44, no. 12 (1959): 206–226, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2254&rep=rep1&type=pdf
[21] http://thedatascientist.com/machine-learning-vs-ai/

Machine learning, on the other hand, takes a bottom-up, data-driven approach. So, instead of programming a computer with rules it must follow, machine learning uses data to teach the computer what to do.

There is very little human input involved, other than providing the data. This allows the machine to handle uncertainty more naturally, making it a more effective approach since the world is fundamentally more probabilistic than certain.

With our rain analogy, machine learning would work a little like this. The system is fed a ton of data and it analyzes what happens when there is a certain type of weather. So, when it's sunny, it sees people going out in both light and heavy clothing. It learns even more when it makes the connection that temperature also matters. Essentially, it recognizes patterns: sun + cold temperature = warmer clothing. The machine does this on its own by analyzing and learning from data it has been given.

The problem with machine learning for a while was that it required a lot of data to work effectively. Thankfully, with technological progress like the development of the Internet and cheap storage, the data-driven approach has become more popular. Not only is it more effective, but it's also more feasible from an economic and technological viewpoint too.

The Problem with Machine Learning

Of course, machine learning has its drawbacks, just like anything else. First of all, machine learning can't handle logic rules directly and how we can integrate machine learning with reasoning is an open question.

Let's take gravity as an example. Machine learning can come to the conclusion that there is a 0.99 probability of an object falling if you let it go. This, however, is based on past data and it cannot make the logical leap that "any" object will fall if you let it go because of gravity. It can give probabilities based on past events, but it cannot come to the conclusion: IF (GRAVITY) THEN (OBJECT FALL). This logical leap is something that even babies can do, demonstrating that while machine learning can be very powerful, it is far from intelligent the way the inventors of AI envisioned it.

Another issue is that machine learning isn't all that transparent. Most algorithms are a "black box," with input going in on one side and the results being spat out on the other side. However, we don't see what's happening inside the box. Think of Google's famous search algorithm. We know we enter a search string and we get the "most" relevant results, but we don't see how that happens.