Big Data and predictive analytics – a scientific paradigm shift?

Bo Sundgren
2016-03-22

Table of contents

Big Data and predictive analytics – a scientific paradigm shift? ................................................................. 1
Big Data and predictive analytics – a scientific paradigm shift? ................................................................. 3
1 Three famous examples of Big Data and predictive analytics ................................................................. 3
   Google Translate ........................................................................................................................................... 3
   Consumer price index (CPI) ....................................................................................................................... 4
   Google Flu Trends ...................................................................................................................................... 5
2 Is there a paradigm shift, a scientific revolution? ......................................................................................... 6
   Contrasting views ........................................................................................................................................ 8
   A paradigm shift based on disruptive changes? .......................................................................................... 9
   More criticism ............................................................................................................................................ 9
3 Basic concepts ............................................................................................................................................... 10
   Big Data ................................................................................................................................................... 10
   Typical properties of Big Data: Volume, Velocity, Variety (the 3Vs) ......................................................... 10
   Typical purposes of Big Data .................................................................................................................... 10
   Gartner’s updated definition combining properties and purposes of Big Data ...................................... 11
   The “data collection – data preparation – data analysis – decisions and actions” loop ......................... 12
   Analytical use of data for problem-solving and decision-making .......................................................... 14
   Organizing data and metadata in a data warehouse .................................................................................. 14
   A data warehouse architecture for a statistical system ............................................................................ 15
   Predictive analytics .................................................................................................................................. 16
   Application areas ....................................................................................................................................... 17
   Methods used in predictive analytics ........................................................................................................ 18
   Predictive analytics in the preparation of input data ................................................................................ 19
   Paradigms and paradigm shifts – scientific revolutions .......................................................................... 20
   Paradigms – world views – Weltanschauungen ....................................................................................... 20
   Paradigm shifts – scientific revolutions ..................................................................................................... 20
   Ontologies and epistemologies .................................................................................................................. 21
   Disruptive changes ..................................................................................................................................... 21
   Radical business process reengineering vs ongoing continuous improvements .................................... 22
4 Machine learning, artificial Intelligence, and statistics .............................................................................. 22
   Machine learning ..................................................................................................................................... 22
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>23</td>
</tr>
<tr>
<td>Relation to statistics</td>
<td>23</td>
</tr>
<tr>
<td>Neural networks</td>
<td>24</td>
</tr>
<tr>
<td>Data mining and knowledge discovery</td>
<td>25</td>
</tr>
<tr>
<td>Business intelligence (BI)</td>
<td>26</td>
</tr>
<tr>
<td>Artificial Intelligence (AI)</td>
<td>27</td>
</tr>
<tr>
<td>Research areas, methods, and tools</td>
<td>27</td>
</tr>
<tr>
<td>Philosophy and ethics</td>
<td>33</td>
</tr>
<tr>
<td>5 Bibliography</td>
<td>35</td>
</tr>
</tbody>
</table>
Starting in Chapter 1 from short descriptions of three famous applications of Big Data and predictive analytics, we shall discuss in Chapter 2, whether we are witnessing a so-called paradigm shift or scientific revolution (Kuhn, 1962, 1970, 2012) in the sense that new methods are emerging, which seem to solve certain types of problems in a more efficient way than more established scientific methods. Are the new methods scientific at all? Some philosophers and scientists question this, claiming that the new methods may be useful for practical purposes, but they do not enlighten our understanding and insights in a way that scientific methods should do. In Chapter 3 we will go into a somewhat deeper discussion of three concepts, which are fundamental for the contents of this article: Big Data, predictive analytics, and paradigms and paradigm shifts. In Chapter 4 we shall indicate the role of statistics in disciplines like Artificial Intelligence, machine learning, neural networks, data mining, Business Intelligence (BI), and Big Data, and we shall also indicate how statisticians have looked upon the usages of statistical methods in these other disciplines.

1 Three famous examples of Big Data and predictive analytics

Google Translate

For several decades, the most brilliant scientists in Artificial Intelligence (AI) used huge amounts of research funding to solve the problem of automatic translation by means of advanced computer software. The attempts were based on domain-specific theories, that is, established language theories like those of Chomsky and others, which were used in developing software which was typically rule-based, like the early expert systems. The practical results were not impressive. After many years of research, the AI experts managed to develop a computer program that was able to translate Canadian weather reports from English to French, but such limited achievements were about all of practical value that was accomplished, using the traditional paradigm.

Then came Google Translate, the result of a new approach, where most domain-specific knowledge was replaced by statistical methods, now known as predictive analytics. This new approach has caused outcries by established domain experts like Noam Chomsky, a well-known philosopher and linguist.

Naturally it seems a bit unprofessional and unscientific to launch translation software like Google Translate, where neither the algorithms, nor the results, can be explained in terms of linguistic concepts and theories. But on the other hand, one must admit that the practical results from using Google Translate are useful, especially in situations where resources in terms of time and money do not permit the engagement of professional translators. The results from Google Translate are far from perfect, but they are usually understandable and useful, something that one could not say about earlier generations of translation software, based on traditional AI methods and linguistic theories.

This should not lead to the conclusion that traditional domain-specific theories are not useful, at least not from an academic point of view. Domain-specific theories and models may be able to produce a better understanding of a domain of interest, for example human languages, even if they are not as helpful in producing practical results, like automatic translation, and such understanding has a value in itself for us as human beings trying to make sense of the world we are living in. Neither
should it be excluded that the statistical methods underlying Google Translate could produce even better practical translation results, if and when combined with good domain-specific theories and models from the discipline of linguistics.

**Consumer price index (CPI)**

For example, proxies of the official consumer price indexes have been produced with good results from price available on the Internet. The proxies are very close to the official indexes, and they can be produced and published much faster. See AAPOR (2015):

“The MIT Billion Prices Projects, PriceStats, [http://bpp.mit.edu/](http://bpp.mit.edu/), is an academic initiative using prices collected daily from hundreds of online retailers around the world to conduct economic research. One statistical product is the estimation of inflation in the US. Changes in inflation trends can be observed sooner in PriceStats than in the monthly Consumer Price Index (CPI). Figure 1 shows aggregated inflation series on a monthly basis for the U.S. from 2008 to 2014 where the statistics derived from the PriceStats Index are displayed in orange, overlaid with the CPI estimates in blue.

![Figure 1. US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI. Accessed January 18, 2015 from the PriceStats website.](image)

Some National Statistical Institutes in Europe are now using internet robots to collect prices from the web or scanner data from retailers as part of their data collection for the CPI.”

National institutes of official statistics, as well as private opinion research institutes, nowadays have big problems to collect data by means of traditional sample surveys. A major problem is that many respondents are not contactable, whereas others are not willing to participate. Response rates go down and costs go up. Using Big Data from the Internet, like in the example above, may be one of the methods to use in order to cope with this situation. Other methods are:
• using registers and administrative data
• using model-based estimation
• using self-selected panels
• using representative rather than probability-based samples

Some of these methods, including Big Data, may be used in combination.

Google Flu Trends


Google Flu Trends was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.

The idea behind Google Flu Trends (GFT) is that, by monitoring millions of users’ health tracking behaviours online, the large number of Google search queries gathered can be analysed to reveal if there is the presence of flu-like illness in a population. Google Flu Trends compared these findings to a historic baseline level of influenza activity for its corresponding region and then reports the activity level as either minimal, low, moderate, high, or intense. These estimates have been generally consistent with conventional surveillance data collected by health agencies, both nationally and regionally.

Google Flu Trends uses Google searches on flu symptoms, remedies and other related key words to provide “near real-time” estimates of flu activity in the U.S. and 24 other countries world-wide.

Google Flu Trends was described as using the following method to gather information about flu trends.

First, a time series is computed for about 50 million common queries entered weekly within the United States from 2003 to 2008. A query’s time series is computed separately for each state and normalized into a fraction by dividing the number of each query by the number of all queries in that state. By identifying the IP address associated with each search, the state in which this query was entered can be determined.

A linear model is used to compute the log-odds of Influenza-like illness (ILI) physician visit and the log-odds of ILI-related search query:

\[
\logit(P) = \beta_0 + \beta_1 \times \logit(Q) + \epsilon
\]

P is the percentage of ILI physician visit and Q is the ILI-related query fraction computed in previous steps. \(\beta_0\) is the intercept and \(\beta_1\) is the coefficient, while \(\epsilon\) is the error term.

Each of the 50 million queries is tested as Q to see if the result computed from a single query could match the actual history ILI data obtained from the U.S. Centers for Disease Control and Prevention (CDC). This process produces a list of top queries which gives the most accurate predictions of CDC ILI data when using the linear model. Then the top 45 queries are chosen because, when aggregated together, these queries fit the history data the most accurately. Using the sum of top 45 ILI-related queries, the linear model is fitted to the weekly ILI data between 2003 and 2007 so that the
coefficient can be gained. Finally, the trained model is used to predict flu outbreak across all regions in the United States.

This algorithm has been subsequently revised by Google, partially in response to concerns about accuracy.

Google Flu Trends is now regarded as an example of the risks of Big Data error. Compared to CDC data, the Google Flu Trends provided remarkably accurate indicators of flu incidence in the U.S. between 2009 and 2011. However, for the 2012-2013 flu seasons, Google Flu Trends predicted more than double the proportion of doctor visits for flu-like symptoms than the CDC (Butler 2013). Lazer et al. (2014) cite two causes of this error: Big Data hubris and algorithm dynamics. The former occurs when the Big Data researcher believes that the volume of the data compensates for any of their deficiencies, thus obviating the need for traditional, scientific analytic approaches. As Lazer et al. (2014:2) note, Big Data hubris fails to recognize that "... quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability...."

2 Is there a paradigm shift, a scientific revolution?

Are the three famous applications, briefly described above, examples of innovative and successful applications of a new scientific method, a new paradigm? Is the method scientific at all? What is the essence of the method, and what distinguishes it from established scientific methods?

Like never before in human history, the easy and instantaneous availability via the Internet of huge amounts of potentially relevant and useful data for virtually all kinds of data-supported research and decision-making has led to disruptive changes and a questioning of traditional scientific paradigms based on knowledge-building, understanding, and logical reasoning.

We have come to a point, where we really need to consider and draw further consequences from Alan Turing’s proposal in his paper "Computing Machinery and Intelligence" that the question "Can machines think?" be replaced with the question "Can machines do what we (as thinking entities) can do?" Maybe data-empowered and computer-supported systems, although still human-designed and human-constructed, can help us to achieve even more impressive information-based results than we can ourselves, by using methods which are fundamentally different from our human ways of thinking, modelling, analysing, reasoning, and deriving new knowledge?

Some pragmatic researchers and engineers associated with information disciplines and information-based systems, tools, and activities have gone so far as to declare established scientific methods obsolete, thereby provoking a lot of people. Peter Norvig, Google’s Director of Research and co-author of the most popular artificial intelligence textbook in the world, Russell & Norvig (1995, 2010), is one of the provokers; Chris Anderson, Chief Editor of Wired Magazine, is another one.

Peter Norvig presents some of his arguments for the new paradigm in his article “Colorless green ideas learn furiously: Chomsky and the two cultures of statistical learning” in Significance (2012), pages 30–33. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2012.00590.x/epdf. The article deals with the topic of language translation, where Google Translate is the result of using the new paradigm. The language philosopher Noam Chomsky is critical, and Peter Norvig examines Chomsky’s arguments in this article, which starts as follows:

“Language recognition programs use massive databases of words, and statistical correlations between those words, to translate or to recognise speech. But correlation is not causation. Do these
statistical data-dredgings give any insight into how language works? Or are they a mere big-number trick, useful but adding nothing to understanding? One who holds the latter view is the theorist of language Noam Chomsky. Peter Norvig disagrees.”

Chris Anderson presents his provocations in his article “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete” in Wired Magazine (2008), http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory. Here are some statements from his article:

- "All models are wrong, but some are useful." So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behaviour, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all. Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age.

- At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day. And Google was right.

- This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

- The big target here isn't advertising, though. It's science. The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years.

- Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

- But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. Consider physics: Newtonian models were crude approximations of the truth (wrong at the atomic level, but still useful). A hundred years ago, statistically based quantum mechanics offered a better picture — but quantum mechanics is yet another model, and as such it, too, is flawed, no doubt a caricature of a more complex underlying reality.
Now biology is heading in the same direction. The models we were taught in school about “dominant” and “recessive” genes steering a strictly Mendelian process have turned out to be an even greater simplification of reality than Newton’s laws. The discovery of gene-protein interactions and other aspects of epigenetics has challenged the view of DNA as destiny and even introduced evidence that environment can influence inheritable traits, something once considered a genetic impossibility.

In short, the more we learn about biology, the further we find ourselves from a model that can explain it.

There is now a better way. We can stop looking for models. We can analyse the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

There’s no reason to cling to our old ways. It’s time to ask: What can science learn from Google?

### Contrasting views

Provocations like those by Norvig and Anderson have not been left without counter-arguments. Maybe a dialectical thesis-antithesis-synthesis debate will finally lead to a better understanding of different roads to new insights and new ways of knowledge formation supported by human beings, data, and computers in unprecedented cooperation.

At the present stage of this debate the following counter-arguments to the provocations by people like Norvig and Anderson have been presented.


The idea that statistics is important for big data has exploded into the popular media. Here are a few examples, starting with the Lazer et. al paper in *Science* that got the ball rolling on this idea:

- The parable of Google Flu: traps in big data analysis
- Big data are we making a big mistake?
- Google Flu Trends: the limits of big data
- Eight (No, Nine!) Problems with Big Data

These articles warn about issues that statisticians have been thinking about for a very long time.

A paradigm shift based on disruptive changes?

The concept of Big Data implies a revolution in information society for many stakeholders: domain scientists and researchers as well as statistical methodologists and business practitioners. The revolution is associated with a paradigm shift, and it is caused and enabled by a number of so-called disruptive changes in technologies and methodologies, such as:

- new data sources and new data generation methods providing an unprecedented volume and richness of data available
- new storage and processing capabilities
- new database management software capable of efficient management and combination of different kinds of data structures and different kinds of data
- new methods for visualisation, animation, analysis, and conclusions
- “made data” replaced by “found data”; AAPOR (2015)
- computer-aided systems operated, monitored, evaluated, and improved by humans, replaced by human-aided self-controlling and self-learning systems
- real-time analysis of and responses to online streaming data enabled by capable and efficient hardware and software

Each one of these developments and changes are by themselves disruptive or at least dramatic, and together they may imply a change of paradigm, a scientific revolution.

More criticism

There are plenty of sceptics when it comes to computers and algorithms abilities to predict the future, including Gary King, a professor from Harvard University and the director of the Institute for Quantitative Social Science. People are influenced by their environment in innumerable ways. Trying to understand what people will do next assumes that all the influential variables can be known and measured accurately:

"People's environments change even more quickly than they themselves do. Everything from the weather to their relationship with their mother can change the way people think and act. All of those variables are unpredictable. How they will impact a person is even less predictable. If put in the same situation tomorrow, they may make a completely different decision. This means that a statistical prediction is only valid in sterile laboratory conditions, which suddenly isn't as useful as it seemed before."
3 Basic concepts

Here we shall define and discuss three basic concepts of fundamental importance for this article:

- **Big Data**
- **Predictive analytics**
- **Paradigms and paradigm shifts**

**Big Data**

The concept of “Big Data” may be defined in different ways. There are two main types of definitions, one which focuses on properties of the data, and one which focuses on the purpose of the data.

**Typical properties of Big Data: Volume, Velocity, Variety (the 3Vs)**

Today it is common to associate the concept of “Big Data” with the so-called 3Vs: volume, velocity, and variety. Doug Laney of the META Group (now a part of Gartner) is often recognized as the originator of the 3Vs, based on his research note Laney (2001). However, it should be noted that Doug Laney never mentions the term “Big Data” in this research note. He rather uses the 3Vs as labels for three data management trends that he expects to become important, and he describes these development trends in a way that has clear resemblances with what we call today “Big Data”.

AAPOR (2015) provides a brief, and yet relatively precise, informative, and comprehensive summary of how the 3Vs could be interpreted in a contemporary e-society:

- **Volume**: This refers to the sheer amount of data available for analysis. This volume of data is driven by the increasing number of data collection instruments (e.g., social media tools, mobile applications, sensors) as well as the increased ability to store and transfer those data with recent improvements in data storage and networking.

- **Velocity**: This refers to both the speed at which these data collection events can occur, and the pressure of managing large streams of real-time data. Across the means of collecting social information, new information is being added to the database at rates ranging from as slow as every hour or so, to as fast as thousands of events per second.

- **Variety**: This refers to the complexity of formats in which Big Data can exist. Besides structured databases, there are large streams of unstructured documents, images, email messages, video, links between devices and other forms that create a heterogeneous set of data points. One effect of this complexity is that structuring and tying data together becomes a major effort, and therefore a central concern of Big Data analysis.

**Typical purposes of Big Data**

Other definitions in the literature focus on the purpose of Big Data. For example, Evelson (2015) cites a definition by Forrester Research, a marketing research company:

- **Big Data**: The practices and technologies that close the gap between the data available and the ability to turn that data into business insight.
This type of definition does not really clarify the distinction between “Big Data” and some earlier concepts like “Business Intelligence” (BI), “data mining”, and “Decision Support Systems” (DSS). For example, Forrester’s definition of Business Intelligence reads:

- **Business Intelligence**: A set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making.

As Evelson says; while BI has been a thriving market for decades and will continue to flourish for the foreseeable future, the world doesn’t stand still and:

- Recognizes a need for more innovation. Some of the approaches in earlier generation BI applications and platforms started to hit a ceiling a few years ago. For example, SQL and SQL-based database management systems (DBMS), while mature, scalable, and robust, are not agile and flexible enough in the modern world where change is the only constant.
- Needs to addresses some of the limitations of earlier generation BI. In order to address some of the limitations of more traditional and established BI technologies, big data offers more agile and flexible alternatives to democratize all data, such as NoSQL, among many others.

It should be noted that both definitions are purpose-oriented. They imply that “Business Intelligence” and “Big Data” have the same intention, the same purpose, that of exploiting whatever data you have, by whatever methods you can think of, into useful insights and decisions.

But this common objective of “Business Intelligence” and “Big Data” does not exclude the possibility that the two concepts are associated with different roads towards the objective: different data sources, different kinds of data, different analytical methods, different computing methods, etc. And indeed, such differences between the two concepts – and other concepts within the same family – do exist, both in theory and in practice.

**Gartner’s updated definition combining properties and purposes of Big Data**

Gartner, and now much of the industry, continue to use the "3Vs" model for describing big data. In 2012, Gartner updated its definition as follows:

- "Big data are high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

If Gartner’s definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a sounder difference between Big Data and Business Intelligence, regarding data and their use:

- **Business Intelligence** uses **descriptive statistics** with data with high information density to measure things, detect trends, etc.
- **Big Data** uses **inductive statistics** with data with low information density, whose huge volume allow to infer laws (regressions…) and thus giving (with the limits of inference reasoning) to Big Data some predictive capabilities.

Contrary to some of the market hype, data democratization and big data do not eliminate the need for certain basics, such as data governance, data quality, master data management, data modelling,
well thought out data architecture, etc. All of the typical end-to-end steps necessary to transform raw data into insights still have to happen; now they may just happen in different places and at different times in the process.

The “data collection – data preparation – data analysis – decisions and actions” loop

All kinds of statistical and analytical processing of all kinds of data may be described by means of the loop illustrated by the figures below.

Figure 2 illustrates the major steps in analytical use of information for problem-solving or decision-making. Figure 3 provides a more elaborated view of the same process, focusing on the information resources used, processed, and produced.

Figure 4 illustrates analytical information processes applied to political decision-making in a society. The political decisions are assumed to be, at least partly, based on high-quality official statistics, and the effects of the decisions are assumed to be analysed and evaluated in a serious way.

The figures below are based on Sundgren (2004b) and other articles by Sundgren; see the Bibliography at the end of this paper.

Figure 2. Major subprocesses in analytical use of information.
Design, evaluation, and redesign of operations and monitoring system

Monitoring of operations

Data collection and data preparation ETL:
Extract – Transform – Load

Aggregation and estimation

Analysis

Problem solving and decision-making

Reports, decisions, actions

Data and metadata management:
Relational DBMS, MapReduce engine — OLAP, data mining, search, and text analytic engines

Input
Output
Input
Output
Input
Output
Input
Output

Data warehouse including data marts and data lakes: data and metadata

Indirect external data sources

Internal data sources

Direct external data sources

Figure 3. More elaborated model of analytical information processes.

Figure 4. Decision-making based on official statistics – including a feedback loop.
Analytical use of data for problem-solving and decision-making

It is a deeply rooted habit and ideal in modern western societies that decision-making and problem-solving should be rational and based on information (often called “facts” or “data” in this context). Roughly speaking, the process goes as follows:

1. A decision to be made, or a problem to be solved, is specified. The context will typically be a business (in a broad sense, including non-profit organisations), or a society on some level: central, regional, local.

2. Information/facts/data are collected from different sources, in order to illuminate the decision to be made, or the problem to be solved.

3. The data are organized, processed, aggregated, and analysed by more or less sophisticated methods (tabulations, visualisations, animations, simulations, statistical analyses, etc.).

4. The results from the analyses are presented for decision-makers, who may ask questions requiring further data collection and/or analyses.

5. A decision is taken and implemented in terms of actions.

6. The effects of the decision and actions are observed and analysed in a feedback loop (which is often neglected, unfortunately, by the decision-makers, e.g. politicians).

Organizing data and metadata in a data warehouse

Streaming data may be used “on the fly” without storing them first in a database. This may speed up the analyses, so that they can be used “in real time” as well, for example with the purpose of controlling and optimizing an on-going process by means of feedback from the analyses; see Figure 5.

![Figure 5. Streaming data vs stored data.](image-url)
However, in most cases data obtained from different sources are stored in a well-organized database or data warehouse, containing both the data themselves and metadata describing the data. The metadata may originate from both design processes and operational processes, for example measurement and observation processes. The design processes will generate definitions of concepts and measurement procedures. Measurement and observation processes may generate metadata about non-response and other errors, which may cause uncertainties and quality problems. When the data used for analytical purposes by one organisation emanate from databases and data collection procedures in other organisations, it is important to acquire not only the data from the other organisations, but also documentations and metadata.

The metadata may also include so-called paradata, data about the processes which may be used for monitoring on-going processes and possibly adjusting them dynamically, “on the fly”, for achieving better process performance in terms of quality and efficiency.

The data and the databases may be more or less structured. So-called relational databases, or SQL databases, have been a dominating standard for several decades now, and data in such databases are typically highly formatted and well structured. However, with increasing use of free-text data and other less structured and more heterogeneous data, often captured from the Internet, it has become necessary to develop new types of databases, sometimes called “NoSQL databases”, where “NoSQL” stands for “Not Only SQL”. As indicated by this term, a data warehouse built for analysis and decision-making may make use of both traditional, well-structured SQL databases, and other types of databases, suitable for the analysis of less structured data.

A data warehouse architecture for a statistical system

One of the main reasons for having a special statistical function in a country or within some other kind of organisation (e.g. an international organisation or even a business company) is that both specialized statistical systems and administrative systems in society or in individual businesses and other organisations generate a lot of data over time. These data have a considerable information potential, and they can be used over and over again, for many different purposes, often purposes that are quite different from the purposes for which the data were originally collected and used.

By taking care of all these data and by storing them, well documented, in an organized way, a statistical function may accumulate a data capital, the future yields of which may go far beyond the value of the first usages of the data. The advantage of having a specialized function for this task arises even if the statistical organisation does not do very much more than we have just described: storing the data together, well documented, and making them available as a collective national or corporate resource. However, using its statistical competence, the statistical function may also add new value to the data, e.g. by integrating the data not only in a physical way, but also from a contents-oriented point of view, by making the data more comparable and coherent, by using standardized concepts and classifications, etc. Some improvements in these directions may be done a posteriori, when the data have already been collected, but it is even better, of course, if these aspects are considered already when the data collection processes are planned – data co-ordination a priori.

In Figure 4 above we visualized a statistical system as a reasonably complete and coherent reflection of important aspects of a society. Such a system could be the basis for advanced analysis, decision-making, and evaluations of decisions already taken and implemented by politicians and others. Similar systems could be designed and implemented in individual businesses and other organisations for supporting analyses and decision-making there. Now we shall study what the architecture of such
a system could be. We will start with a brief historical background of the data warehouse approach to statistical data, and then proceed to a modern version of the same concept.

**Predictive analytics**

More advanced analysis of big data is often associated with so-called predictive analytics. It is easy for statisticians to see, how traditional, well established statistical methods “reoccur” in the new contexts of “machine learning”, “neural networks”, “data mining”, “business intelligence”, and “big data”. However, both statisticians and non-statisticians must observe that new conditions, assumptions, and models may often apply, when the old methods are used in new contexts. Some of the methods, especially those based on machine learning and artificial neural networks, may not be familiar to traditional statisticians, even though these methods too are based on statistical methods – actually in opposition to what used to be mainstream knowledge- and rule-based artificial intelligence methods.

Predictive analytics uses data to predict trends and behaviour patterns. Often an unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present, or future, for example, identifying suspects after a crime has been committed, or credit card fraud as it occurs.

The core of predictive analytics relies on capturing relationships between explanatory variables and predicted variables from past occurrences, and exploiting these relationships to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a rather detailed level of granularity. This distinguishes it from forecasting. Predictive analytic learns from experience (data) to predict the future behaviour of individuals in order to drive better decisions.

Predictive analytics encompasses a variety of statistical techniques from modelling, machine learning, and data mining that analyse current and historical facts to make predictions about future, or otherwise unknown, events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Predictive analytics may provide a predictive score (probability) for each individual (customer, employee, healthcare patient, product, vehicle, component, machine, or other organisational unit) in order to determine, inform, or influence organisational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement.

Predictive analytics is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel, healthcare, pharmaceuticals and other fields. One of the most well-known applications is credit scoring, which is used throughout financial services. Scoring models process a customer’s credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time.
Application areas

Predictive analytics has shown positive impact in a number of application areas in recent years:

Analytical customer relationship management (CRM)

Methods of predictive analysis are applied to customer data to pursue CRM objectives, which involve constructing a holistic view of the customer no matter where their information resides in the company or the department involved. Analytical CRM can be applied throughout the customer’s lifecycle (acquisition, relationship growth, retention, and win-back).

Clinical decision support systems

Experts use predictive analysis in health care to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease, etc. Sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care.

Collection analytics

Many portfolios have a set of delinquent customers who do not make their payments on time. The financial institution has to undertake collection activities on these customers to recover the amounts due. A lot of collection resources are wasted on customers who are difficult or impossible to recover. Predictive analytics can help optimize the allocation of collection resources by identifying the most effective collection agencies, contact strategies, legal actions, and other strategies to each customer, thus significantly increasing recovery and reducing collection costs.

Fraud detection

Fraud is a big problem for many businesses and can be of various types: inaccurate credit applications, fraudulent transactions (both offline and online), identity thefts, and false insurance claims. A predictive model can help to identify high-risk fraud candidates in business or the public sector. For example, the Internal Revenue Service (IRS) of the United States uses predictive analytics to identify tax fraud.

Predictive behaviour analysis is also used for web fraud detection. Heuristics are used in order to study normal web user behaviour and detect anomalies indicating fraud attempts.

Portfolio, product or economy-level prediction

Often the focus of analysis is not the consumer but the product, portfolio, firm, industry, or the economy. A retailer may be interested in predicting store-level demand. Or the Federal Reserve Board might be interested in predicting the unemployment rate. These types of problems can be addressed by predictive analytics using time series techniques. They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power.

Risk management

When employing risk management techniques, the purpose is to predict and benefit from a future scenario. Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to
determine the amount of premium to cover each automobile and driver. A financial company needs to assess a borrower’s potential and ability to pay before granting a loan. Predictive analytics can help estimate the chances of illness, default, bankruptcy, etc.

Predictive analytics in the form of credit scores have reduced the amount of time it takes for loan approvals, especially in the mortgage market where lending decisions are now made in a matter of hours rather than days or even weeks. Proper predictive analytics can lead to proper pricing decisions, which can help mitigate future risk of default.

Methods used in predictive analytics

The approaches and techniques used to conduct predictive analytics can broadly be grouped into regression techniques and machine learning techniques.

Regression techniques

Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation, there are a wide variety of models that can be applied while performing predictive analytics. Some examples:

- Linear regression model
- Discrete choice models
- Logistic regression
- Multinomial logistic regression
- Probit regression
- Time series models
- Survival or duration analysis
- Classification and regression trees
- Multivariate adaptive regression splines

Machine learning techniques

Machine learning, a branch of artificial intelligence, was originally employed to develop techniques to enable computers to learn. Today, since it includes a number of advanced statistical methods for regression and classification, it finds application in a wide variety of fields including medical diagnostics, credit card fraud detection, face and speech recognition and analysis of the stock market. In certain applications it is sufficient to directly predict the dependent variable without focusing on the underlying relationships between variables. In other cases, the underlying relationships can be very complex and the mathematical form of the dependencies unknown. For such cases, machine learning techniques learn from training examples to predict future events.

Here are some of the methods used for predictive analytics:

- Neural networks
- Multilayer Perceptron (MLP)
- Radial basis functions
- Naïve Bayes for performing classifications tasks
- Pattern recognition methods, k-nearest neighbours
- Geospatial predictive modelling
A detailed study of machine learning can be found in Mitchell (1997).

Predictive analytics in the preparation of input data

Predictive analytics and similar methods have also proven to be useful in the preparation of input data to be used in statistical and analytical processes. For example, so-called macro-editing, or significance editing, can be used for optimizing resources used for identifying, investigating, and correcting suspicious data, and for replacing missing data by imputed values.

Data will almost always contain errors. The errors are of different types and will be more or less serious for different types of usages. When data are used for operational purposes, for example in e-commerce applications, it is important that certain data are exactly correct, for example the identity and address of a customer, and the identity and the quantity and price of a product ordered – errors will lead to big problems for both the customer and the service provider. It is necessary to maintain so-called “book-keeping quality” in such applications.

However, when it comes to analytical use of data for problem-solving and decision-making purposes, the quality requirements become different. Book-keeping quality will not be necessary. Small random errors can be accepted, as long as they do not affect the results of the analyses to be made. Since checking and correcting data (called “data editing” or “data cleaning” by statisticians) is an expensive and time-consuming operation, one should focus on errors in the data that will affect the results of the analyses to be made. Is it possible to find those errors quickly, without having to check all data with the same priority and intensity? Yes, there are statistical methods for data editing, which may help us with this, so-called “selective editing” or “significance editing” or “macro-editing”. Among other things these methods will identify so-called “outliers”, observations which are far away from other observations from a statistical point of view, often illustrated by means of graphs; see Figure 6.

Figure 6. Outliers.
Selective editing means, expressed in a simple way, that the editing process is focused on those suspected errors in the data, which can be estimated to have (the most) significant effects on the estimates to be produced. This editing method assumes that the estimates to be produced are known. However, the data collected may also be made available for other purposes than those at hand, for example within the framework of a statistical system. Then many of the estimates to be produced in the future on basis of the data at hand, possibly in combination with data from other sources, are unknown during the design and execution of the editing process at hand. How should such situations be treated? As a minimum the statisticians responsible for the data at hand must provide good information about how the editing of the data at hand has been done, which assumptions have been made, which usages of the data could be particularly uncertain or dangerous, etc. Probably, as a routine, the original, unedited data should also be archived, so as to make it possible for future reusers of the data to design and execute alternative editing procedures on the same data.

When suspected errors are discovered in data, the next question is what to do with them. If there is good reason to believe that the errors are not important for the purposes of the planned analyses of the data, nothing should be done, if the actions needed are expensive and time-consuming – for example since they involve investigations undertaken by experts, or even going back to the original data providers, asking them for clarifications. Sometimes advanced methods, like neural networks, may be used for replacing human expertise, thus enabling automatic data editing with good results.

Paradigms and paradigm shifts – scientific revolutions

Paradigms – world views – Weltanschauungen

The term “paradigm” is used, especially among philosophers and scientists, to denote a “world view” or “Weltanschauung” (a German word which is also used in English, originating from Kant and Hegel), that is a holistic, basic understanding of the world we are living in, and trying to obtain knowledge about. The paradigm concept also includes methods that are regarded as legitimate or scientific for obtaining knowledge about reality.

Paradigm shifts – scientific revolutions

From time to time sciences and scientific methods are undergoing radical changes, also called “paradigm shifts” or “scientific revolutions” (Kuhn, 1962, 1970, 2012). The radical change may be a new theory replacing an old theory, or a new method gaining support by scientists as a replacement or complement to existing methods.

Some examples of important paradigm shifts:

- Copernicus replaced the geocentric world view with a heliocentric view, where the earth circulates around the sun, rather than the other way around. The heliocentric world view was further developed by Galilei, Kepler, and Newton.

- Einstein’s relativity theory falsified Newton’s theories.

- The paradigm of positivism has been very successful in natural sciences for hundreds of years. However, when scientists tried to apply this paradigm outside natural sciences, for example within social sciences, they encountered serious problems, and social scientists have since then tried to develop new paradigms and scientific methods, which are as rigorous as methods used in natural sciences, but more suitable for the problems tackled in social sciences, where controlled
experiments are difficult to set up, and where quantitative measurements are not always easy to define and obtain.

Between radical paradigm shifts, like those mentioned above, there are typically long periods dominated by so-called “normal science” (Kuhn, 1962, 1970, 2012), where existing theories are confirmed or stepwise refined and developed. Most scientific work undertaken by scientists is normal science.

Ontologies and epistemologies

The old Greek philosophers did not speak about paradigms, but they rather debated ontologies and epistemologies. An ontology defines “what exists”, the reality, and an epistemology expresses how we can obtain knowledge about reality, for example by means of observations and experiments, or by means of logical reasoning.

Rationalism and empiricism

Throughout the history there has been an ongoing debate between two major opinions about how knowledge can most safely be obtained: rationalism and empirism. Empiricists stress the importance of observations through our senses as the best source of knowledge, whereas rationalists stress the role of human conceptualisation, thinking, and reasoning.

Reality: objectively existing or constructed by humans?

Ontologies were originally closely intertwined with metaphysic and religious beliefs, but today other topics are more discussed, for example, to which extent reality exists in an objective way, like physical objects, and to which extent objects and phenomena are just human or social constructions, like companies and organisations for example.

Disruptive changes

Among economists and business people there is a concept, “disruptive change”, or “industrial revolution”, which has a certain kinship with “paradigm shift” and “scientific revolution”.

A disruptive change typically occurs in the context of technological innovation. Some examples:

- Refrigerators eliminated the need for ice and ice distributors.
- Electronic calculators replaced mechanic calculators and made big, world-leading companies like the Swedish company Facit go bankrupt, since they were not able or willing to adapt their business to the new technology.
- Digital media replaced analogous media, and the traditional media industries (publishers of music, films, newspapers, books, etc) were by and large unable to adapt their business models to the new situation – instead new companies with innovative business models were created and flourished, for example Google, Spotify, Lady Gaga.

An interesting observation is that the radical innovations of business models, which are often necessitated by disruptive changes and industrial revolutions, very seldom come from existing companies, but rather from completely new companies without longstanding traditions and inherited organisational inertia. Instead of adapting and innovating themselves, the old companies
try to hang on as long as possible by repressive methods, fighting competitors with legal means, such as patent and copyright laws, and by demanding more such laws and harder punishments for those who violate the laws. The repressive methods sometimes lead to contra-productive, ridiculous, and even laughable consequences, such as music producers bringing their own best customers to court, rather than developing new and more user-friendly products and dissemination channels, adapted to the demand from the customers.

**Radical business process reengineering vs ongoing continuous improvements**

There are interesting parallels between

- scientific revolutions and normal science in the world of science and research;
- disruptive technology changes in the global business world; and:
- radical business process reengineering (BPR) vs ongoing continuous improvements as business-internal strategies for more efficiency and better quality

**4 Machine learning, artificial Intelligence, and statistics**

There are close connections between the scientific fields of machine learning, artificial intelligence, and statistics. These connections have sometimes been fruitful and constructive, but they have often been characterized by conflicts as well.

**Machine learning**

For sources and references to relevant literature:

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

Machine learning has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR), search engines, and computer vision.

In 1959, Arthur Samuel defined machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed".

Tom M. Mitchell provided a widely quoted, more formal definition:

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

This definition is notable for its defining machine learning in fundamentally operational rather than cognitive terms, thus following Alan Turing's proposal in his paper "Computing Machinery and Intelligence" that the question "Can machines think?" be replaced with the question "Can machines do what we (as thinking entities) can do?"
Machine learning is typically classified into three broad categories:

1. **Supervised learning**: The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

2. **Unsupervised learning**: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end.

3. **Reinforcement learning**: A computer program interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle), without a teacher explicitly telling it whether it has come close to its goal or not. Another example is learning to play a game by playing against an opponent.

**History**

As a scientific endeavour, machine learning grew out of the quest for artificial intelligence (AI). Already in the early days of AI, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons and other models that were later found to be reinventions of the generalized linear models of statistics. Probabilistic reasoning was also employed, especially in automated medical diagnosis.

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabilistic systems were plagued by theoretical and practical problems of data acquisition and representation. By 1980, expert systems had come to dominate AI, and statistics was out of favour. Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming, but the more statistical line of research was now outside the field of AI proper, in pattern recognition and information retrieval. Neural networks research had been abandoned by AI and computer science around the same time. This line, too, was continued outside the AI/CS field, as "connectionism", by researchers from other disciplines.

Machine learning, reorganized as a separate field, started to flourish in the 1990s. The field changed its goal from achieving artificial intelligence to tackling solvable problems of a practical nature. It shifted focus away from the symbolic approaches it had inherited from AI, and toward methods and models borrowed from statistics and probability theory. It also benefitted from the increasing availability of digitized information, and the possibility to distribute that via the Internet.

**Relation to statistics**

Machine learning and statistics are closely related fields. According to Michael I. Jordan, the ideas of machine learning, from methodological principles to theoretical tools, have had a long pre-history in statistics. He also suggested the term data science as a placeholder to call the overall field.

**Clustering**

Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to some predesignated criterion or criteria, while observations drawn from different clusters are dissimilar. Different clustering
techniques make different assumptions on the structure of the data, often defined by some similarity metric and evaluated for example by internal compactness (similarity between members of the same cluster) and separation between different clusters. Other methods are based on estimated density and graph connectivity. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis.

Bayesian networks

A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning.

Neural networks

For sources and references to relevant literature:
https://en.wikipedia.org/wiki/Artificial_neural_network

An artificial neural network (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks. It is used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which exchange messages between each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning.

For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activation of these neurons are then passed on to other neurons. This process is repeated until, finally, an output neuron is activated. This determines which character was read.

Like other machine learning methods - systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

Modern neural networks are non-linear statistical data modelling tools. They are usually used to model complex relationships between inputs and outputs, to find patterns in data, or to capture the statistical structure in an unknown joint probability distribution between observed variables.

There is no single formal definition of what an artificial neural network is. However, a class of statistical models may commonly be called neural if it possesses the following characteristics:

- contains sets of adaptive weights, i.e. numerical parameters that are tuned by a learning algorithm
- capability of approximating non-linear functions of their inputs
- the adaptive weights can be thought of as connection strengths between neurons, which are activated during training and prediction
Neural networks are similar to biological neural networks in the performing of functions collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which individual units are assigned. The term neural network usually refers to models employed in statistics, cognitive psychology and artificial intelligence.

In modern software implementations of artificial neural networks, the approach inspired by biology has been largely abandoned for a more practical approach based on statistics and signal processing. In some of these systems, neural networks or parts of neural networks (like artificial neurons) form components in larger systems that combine both adaptive and non-adaptive elements. While the more general approach of such systems is more suitable for real-world problem solving, it has little to do with the traditional, artificial intelligence connectionist models. What they do have in common, however, is the principle of non-linear, distributed, parallel and local processing and adaptation. Historically, the use of neural network models marked a directional shift in the late eighties from high-level (symbolic) AI, characterized by expert systems with knowledge embodied in if-then rules, to low-level (sub-symbolic) machine learning, characterized by knowledge embodied in the parameters of a dynamical system.

Data mining and knowledge discovery

For sources and references to relevant literature:

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), is the process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms data dredging, data fishing, and data snooping refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

In the 1960s, statisticians used terms like "data fishing" or "data dredging" to refer to what they considered the bad practice of analysing data without an a-priori hypothesis. The term "data mining" appeared around 1990 in the database community.

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). Computer techno-
logy has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

**Business intelligence (BI)**


**Business intelligence (BI)** is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. BI technologies are capable of handling large amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics, and prescriptive analytics.

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data.

In their review of Business Intelligence technology, Chaudhury et al. offer the following description:

“Business intelligence (BI) software is a collection of decision support technologies for the enterprise aimed at enabling knowledge workers such as executives, managers, and analysts to make better and faster decisions. The past two decades have seen explosive growth, both in the number of products and services offered and in the adoption of these technologies by industry. This growth has been fuelled by the declining cost of acquiring and storing very large amounts of data arising from sources such as customer transactions in banking, retail as well as in e-businesses, RFID tags for inventory tracking, email, query logs for Web sites, blogs, and product reviews. Enterprises today collect data at a finer granularity, which is therefore of much larger volume. Businesses are leveraging their data asset aggressively by deploying and experimenting with more sophisticated data analysis techniques to drive business decisions and deliver new functionality such as personalized offers and services to customers. Today, it is difficult to find a successful enterprise that has not leveraged BI technology for its business. For example, BI technology is used in manufacturing for order shipment and customer support, in retail for user profiling to target grocery coupons during checkout, in financial services for claims analysis and fraud detection, in transportation for fleet management, in
telecommunications for identifying reasons for customer churn, in utilities for power usage analysis, and health care for outcomes analysis.”

Everything in this description seems to be as applicable to Big Data as to Business Intelligence. The authors summarize:

- The cost of data acquisition and data storage has declined significantly. This has increased the appetite of businesses to acquire very large volumes in order to extract as much competitive advantage from it as possible.
- New massively parallel data architectures and analytic tools go beyond traditional parallel SQL data warehouses and OLAP engines.
- The need to shorten the time lag between data acquisition and decision making is spurring innovations in business intelligence technologies.

**Artificial Intelligence (AI)**


Artificial intelligence (AI) is the intelligence exhibited by machines or software. It is also the name of the academic field of study which studies how to create computers and computer software that are capable of intelligent behaviour. Major AI researchers and textbooks define this field as "the study and design of intelligent agents", in which an intelligent agent is a system that perceives its environment and takes actions that maximize its chances of success. John McCarthy, who coined the term in 1955, defines it as "the science and engineering of making intelligent machines".

What then is “intelligent behaviour”, “intelligent agents”, and “intelligent machines”? Alan Turing proposed in his paper "Computing machinery and intelligence" that the question "Can machines think?" be replaced with the question "Can machines do what we (as thinking entities) can do?"


The central problems of AI research include reasoning, knowledge, planning, learning, natural language processing (communication), perception and the ability to move and manipulate objects. Currently popular approaches include statistical methods, computational intelligence, and traditional symbolic AI. There are a large number of tools used in AI, including versions of search and mathematical optimization, logic, methods based on probability and economics, and many others.

The field was founded on the claim that a central property of humans, human intelligence—the sapience of Homo sapiens—"can be so precisely described that a machine can be made to simulate it." (The Dartmouth proposal, 1956, [https://en.wikipedia.org/wiki/Dartmouth_Conferences](https://en.wikipedia.org/wiki/Dartmouth_Conferences)).

**Research areas, methods, and tools**

The general problem of simulating (or creating) intelligence has been broken down into a number of specific subproblems. These consist of particular traits or capabilities that researchers would like an intelligent system to display. The traits described below have received the most attention.
There is no established unifying theory or paradigm that guides AI research. Researchers disagree about many issues. A few of the most long-standing questions that have remained unanswered are these:

- Should artificial intelligence simulate natural intelligence by studying psychology or neurology?
- Or is human biology as irrelevant to AI research as bird biology is to aeronautical engineering?
- Can intelligent behaviour be described using simple, elegant principles, such as logic or optimization?
- Or does it necessarily require solving a large number of completely unrelated problems?
- Can intelligence be reproduced using high-level symbols, similar to words and ideas?
- Or does it require "sub-symbolic" processing?

**Deduction, reasoning, problem solving**

Early AI researchers developed algorithms that imitated the step-by-step reasoning that humans use when they solve puzzles or make logical deductions. By the late 1980s and 1990s, AI research had also developed highly successful methods for dealing with uncertain or incomplete information, employing concepts from probability and economics.

For difficult problems, most of these algorithms can require enormous computational resources – most experience a "combinatorial explosion": the amount of memory or computer time required becomes astronomical when the problem goes beyond a certain size. The search for more efficient problem-solving algorithms is a high priority for AI research.

Human beings solve most of their problems using fast, intuitive judgements rather than the conscious, step-by-step deduction that early AI research was able to model. AI has made some progress at imitating this kind of "sub-symbolic" problem solving: embodied agent approaches emphasize the importance of sensorimotor skills to higher reasoning; neural net research attempts to simulate the structures inside the brain that give rise to this skill; statistical approaches to AI mimic the probabilistic nature of the human ability to guess.

**Cognitive simulation based on human problem-solving skills: Simon & Newell**

The economists Herbert Simon and Allen Newell studied human problem-solving skills and attempted to formalize them, and their work laid the foundations of the field of artificial intelligence, as well as cognitive science, operations research, and management science. Their research team used the results of psychological experiments to develop programs that simulated the techniques that people used to solve problems.

**Knowledge-based methods, expert systems**

When computers with large memories became available around 1970, researchers began to build knowledge into AI applications. This "knowledge revolution" led to the development and deployment of expert systems (introduced by Edward Feigenbaum), the first truly successful form of AI software. The knowledge revolution was also driven by the realization that enormous amounts of knowledge would be required by many simple AI applications.

**Knowledge representation and knowledge engineering**

Knowledge representation and knowledge engineering are central to AI research. Many of the problems machines are expected to solve will require extensive knowledge about the world. Among
the things that AI needs to represent are: objects, properties, categories and relations between objects; situations, events, states and time; causes and effects; knowledge about knowledge (what we know about what other people know); and many other, less well researched domains. A representation of "what exists" is an ontology: the set of objects, relations, concepts and so on that the machine knows about. The most general are called upper ontologies, which attempt to provide a foundation for all other knowledge.

Figure 7. An ontology represents knowledge as a set of concepts within a domain and the relationships between those concepts. From Wikipedia (Artificial intelligence).

Symbolic approaches, symbol manipulation, sub-symbolic methods

When access to digital computers became possible in the middle 1950s, AI research began to explore the possibility that human intelligence could be reduced to symbol manipulation. During the 1960s, symbolic approaches had achieved great success at simulating high-level thinking in small demonstration programs. Approaches based on cybernetics or neural networks were abandoned or pushed into the background. Researchers in the 1960s and the 1970s were convinced that symbolic approaches would eventually succeed in creating a machine with artificial general intelligence and considered this the goal of their field.

By the 1980s progress in symbolic AI seemed to stall and many believed that symbolic systems would never be able to imitate all the processes of human cognition, especially perception, robotics, learning, and pattern recognition. A number of researchers began to look into "sub-symbolic" approaches to specific AI problems.

Natural language processing (communication)

Natural language processing gives machines the ability to read and understand the languages that humans speak. A sufficiently powerful natural language processing system would enable natural language user interfaces and the acquisition of knowledge directly from human-written sources, such as newswire texts. Some straightforward applications of natural language processing include information retrieval (or text mining), question answering and machine translation.

A common method of processing and extracting meaning from natural language is through semantic indexing. Increases in processing speeds and the drop in the cost of data storage makes indexing large volumes of abstractions of the user’s input much more efficient.
Machine learning

Machine learning is the study of algorithms that improve automatically through experience.

Unsupervised learning is the ability to find patterns in a stream of input. Supervised learning includes both classification and numerical regression. Classification is used to determine what category something belongs in, after seeing a number of examples of things from several categories. Regression is the attempt to produce a function that describes the relationship between inputs and outputs and predicts how the outputs should change as the inputs change. In reinforcement learning the agent is rewarded for good responses and punished for bad ones. The agent uses this sequence of rewards and punishments to form a strategy for operating in its problem space. These three types of learning can be analysed in terms of decision theory, using concepts like utility.

Statistical methods

In the 1990s, AI researchers developed sophisticated mathematical tools to solve specific sub-problems. These tools are truly scientific, in the sense that their results are both measurable and verifiable, and they have been responsible for many of AI’s recent successes. The shared mathematical language has also permitted a high level of collaboration with more established fields (like mathematics, economics, or operations research). Stuart Russell and Peter Norvig describe this movement as nothing less than a “revolution” and “the victory of “the neats” over “the scruffies”. Critics argue that these techniques (with few exceptions) are too focused on particular problems and have failed to address the long-term goal of general intelligence. There is an ongoing debate about the relevance and validity of statistical approaches in AI, exemplified in part by exchanges between Peter Norvig and Noam Chomsky.

Probabilistic methods for uncertain reasoning

Many problems in AI (in reasoning, planning, learning, perception and robotics) require the agent to operate with incomplete or uncertain information. AI researchers have devised a number of powerful tools to solve these problems using methods from probability theory and economics. Bayesian networks are a very general tool that can be used for a large number of problems: reasoning (using the Bayesian inference algorithm), learning (using the expectation-maximization algorithm), planning (using decision networks), and perception (using dynamic Bayesian networks). Probabilistic algorithms can also be used for filtering, prediction, smoothing and finding explanations for streams of data, helping perception systems to analyse processes that occur over time (e.g. hidden Markov models or Kalman filters.)
A key concept from the science of economics is "utility": a measure of how valuable something is to an intelligent agent. Precise mathematical tools have been developed that analyse how an agent can make choices and plan, using decision theory, decision analysis, and information value theory. These tools include models such as Markov decision processes, dynamic decision networks, game theory and mechanism design.

Classifiers and statistical learning methods

The simplest AI applications can be divided into two types: classifiers ("if shiny then diamond") and controllers ("if shiny then pick up"). Controllers do, however, also classify conditions before inferring actions, and therefore classification forms a central part of many AI systems. Classifiers are functions that use pattern matching to determine a closest match. They can be tuned according to examples, making them very attractive for use in AI. These examples are known as observations or patterns. In supervised learning, each pattern belongs to a certain predefined class. A class can be seen as a decision that has to be made. All the observations combined with their class labels are known as a data set. When a new observation is received, that observation is classified based on previous experience.

A classifier can be trained in various ways; there are many statistical and machine learning approaches. The most widely used classifiers are the neural network, kernel methods such as the support vector machine, k-nearest neighbour algorithm, Gaussian mixture model, naive Bayes classifier, and decision tree. The performance of these classifiers have been compared over a wide range of tasks. Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems; this is also referred to as the "no free lunch" theorem. Determining a suitable classifier for a given problem is still more an art than science.

Neural networks

The study of artificial neural networks began in the decade before the field of AI research was founded, in the work of Walter Pitts and Warren McCullough. Other important early researchers were Frank Rosenblatt, who invented the perceptron and Paul Werbos who developed the backpropagation algorithm.

The main categories of networks are acyclic or feedforward neural networks (where the signal passes in only one direction) and recurrent neural networks (which allow feedback).

![Figure 9. A neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain. From Wikipedia (Artificial intelligence).](image)
Computational intelligence and soft computing

Interest in neural networks was revived by David Rumelhart and others in the middle 1980s. Neural networks are an example of soft computing; they are solutions to problems which cannot be solved with complete logical certainty, and where an approximate solution is often enough. Other soft computing approaches to AI include fuzzy systems, evolutionary computation and statistical tools.

Perception

Machine perception is the ability to use input from sensors (such as cameras, microphones, tactile sensors, sonar and others more exotic) to deduce aspects of the world. Computer vision is the ability to analyse visual input. Examples of important subproblems are speech recognition, facial recognition and object recognition.

Robotics: motion and manipulation

The field of robotics is closely related to AI. Intelligence is required for robots to be able to handle such tasks as object manipulation and navigation, with sub-problems of localization (knowing where you are, or finding out where other things are), mapping (learning what is around you, building a map of the environment), and motion planning (figuring out how to get there) or path planning (going from one point in space to another point, which may involve compliant motion – where the robot moves while maintaining physical contact with an object).

Logic-based methods and "anti-logic" or "scruffy" ad-hoc solutions

Unlike Newell and Simon, John McCarthy felt that machines did not need to simulate human thought, but should instead try to find the essence of abstract reasoning and problem solving, regardless of whether people used the same algorithms. He focused on using formal logic to solve a wide variety of problems, including knowledge representation, planning and learning. Logic was also the focus of research work in Europe and led to the development of the programming language Prolog and the science of logic programming.

Researchers such as Marvin Minsky and Seymour Papert found that solving difficult problems in vision and natural language processing required ad-hoc solutions – they argued that there was no simple and general principle (like logic) that would capture all the aspects of intelligent behaviour. Roger Schank described their "anti-logic" approaches as "scruffy".

Bottom-up, embodied, situated, behaviour-based or nouvelle AI

Researchers from robotics, such as Rodney Brooks, rejected symbolic AI and focused on the basic engineering problems that would allow robots to move and survive. Their work revived the non-symbolic viewpoint of the early cybernetics researchers of the 1950s and reintroduced the use of control theory in AI. This coincided with the development of the embodied mind thesis in cognitive science: the idea that aspects of the body (such as movement, perception and visualization) are required for higher intelligence.

Search and optimization

Many problems in AI can be solved in theory by intelligently searching through many possible solutions. Reasoning can be reduced to performing a search. For example, logical proof can be viewed as searching for a path that leads from premises to conclusions, where each step is the
application of an inference rule. Planning algorithms search through trees of goals and subgoals, attempting to find a path to a target goal, a process called means-ends analysis. Robotics algorithms for moving limbs and grasping objects use local searches in configuration space. Many learning algorithms use search algorithms based on optimization.

Simple exhaustive searches are rarely sufficient for most real world problems: the search space (the number of places to search) quickly grows to astronomical numbers. The result is a search that is too slow or never completes. The solution, for many problems, is to use "heuristics" or "rules of thumb" that eliminate choices that are unlikely to lead to the goal (called "pruning the search tree"). Heuristics supply the program with a "best guess" for the path on which the solution lies. Heuristics limit the search for solutions into a smaller sample size.

A very different kind of search came to prominence in the 1990s, based on the mathematical theory of optimization. For many problems, it is possible to begin the search with some form of a guess and then refine the guess incrementally until no more refinements can be made.

*Integrating the approaches – the intelligent agent paradigm*

An intelligent agent is a system that perceives its environment and takes actions which maximize its chances of success. The simplest intelligent agents are programs that solve specific problems. More complicated agents include human beings and organizations of human beings (such as firms). The paradigm gives researchers license to study isolated problems and find solutions that are both verifiable and useful, without agreeing on one single approach. An agent that solves a specific problem can use any approach that works – some agents are symbolic and logical, some are sub-symbolic neural networks and others may use new approaches. The paradigm also gives researchers a common language to communicate with other fields—such as decision theory and economics—that also use concepts of abstract agents. The intelligent agent paradigm became widely accepted during the 1990s.

Researchers have designed systems to build intelligent systems out of interacting intelligent agents in a multi-agent system. A system with both symbolic and sub-symbolic components is a hybrid intelligent system, and the study of such systems is artificial intelligence systems integration. A hierarchical control system provides a bridge between sub-symbolic AI at its lowest, reactive levels and traditional symbolic AI at its highest levels, where relaxed time constraints permit planning and world modelling. Rodney Brooks' subsumption architecture was an early proposal for such a hierarchical system.

*Philosophy and ethics*

There are three philosophical questions related to AI:

1. Is artificial general intelligence possible? Can a machine solve any problem that a human being can solve using intelligence? Or are there hard limits to what a machine can accomplish?

2. Are intelligent machines dangerous? How can we ensure that machines behave ethically and that they are used ethically?

3. Can a machine have a mind, consciousness and mental states in exactly the same sense that human beings do? Can a machine be sentient, and thus deserve certain rights? Can a machine intentionally cause harm?
The limits of artificial general intelligence

Can a machine be intelligent? Can it "think"?

Turing's "polite convention"

We need not decide if a machine can "think"; we need only decide if a machine can act as intelligently as a human being. This approach to the philosophical problems associated with artificial intelligence forms the basis of the Turing test:


The Dartmouth proposal

"Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it." This conjecture was printed in the proposal for the Dartmouth Conference of 1956, and represents the position of most working AI researchers.
5 Bibliography


Anderson C. (2008a). Free! Why $0.00 is the Future of Business. Wired Magazine 16.03.


Harford, T. (2014). *Big data: are we making a big mistake?* FT Magazine, http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabcd0.html#axzz30INfAyMi


Sundgren, B. (2004a). Metadata systems in statistical production processes - for which purposes are they needed, and how can they best be organised? UNECE/Eurostat/OECD Joint Session on Statistical Metadata (METIS), Geneva.


