

POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Autoveicolo

Tesi di Laurea Magistrale



TO MAKE RIGHT DECISIONS,

LET THE DATA SPEAK

ANALYSIS OF OPPORTUNITIES

OFFERED BY "DATA SCIENCE"

Supervisor:

Prof. Andrea TONOLI

Co-supervisor:

Prof. Mario VIANELLO

Candidate:

XUE Ming

October 2018

CONTENT

FOREWORD.....	4
INTRODUCTION.....	8
<i>DEFINITION OF DATA SCIENCE.....</i>	<i>8</i>
<i>HISTORY.....</i>	<i>9</i>
<i>RESEARCH CONTENT.....</i>	<i>10</i>
<i>KNOWLEDGE SYSTEM.....</i>	<i>11</i>
<i>RELATIONSHIP WITH OTHER DISCIPLINES.....</i>	<i>12</i>
<i>SYSTEM FRAMEWORK.....</i>	<i>12</i>
CHAPTER 1. "PROGENITORS" OF DATA SCIENCE.....	13
—12 MAIN METHODOLOGICAL TOOLS.....	13
1. RUN CHART.....	14
2. CLASSICAL PROBLEM SOLVING (PARETO).....	14
3. SHORT TERM TARGET (Six SIGMA METRICS).....	16
4. MULTIPLE REGRESSION.....	20
5. DIAGNOSIS BY SIGNATURES.....	23
6. DESIGN OF EXPERIMENTS (DOE).....	24
7. SHAININ METHODS.....	25
8. GENETIC ALGORITHM.....	27
9. CLUSTERING.....	28
10. PRINCIPAL COMPONENT ANALYSIS (PCA).....	30
11. BAYESIAN APPROACH FOR RELIABILITY PREDICTIONS.....	32
12. NEURAL NETWORKS.....	34
CHAPTER 2. DATA SCIENCE, DATA ANALYTICS AND BIG DATA.....	35
DATA SCIENCE VS. BIG DATA VS. DATA ANALYTICS.....	35
<i>What They Are.....</i>	<i>35</i>
<i>The Applications of Each Field.....</i>	<i>36</i>
<i>The Skills Required of Each Field.....</i>	<i>38</i>
<i>To Become a Data Scientist:.....</i>	<i>38</i>
<i>Salaries.....</i>	<i>39</i>
CHAPTER 3. DATA SCIENCE AND BUSINESS STRATEGIES.....	40
FOREWORD.....	40
SEVEN STEPS FOR EXECUTING A SUCCESSFUL DATA SCIENCE STRATEGY.....	41
1.IDENTIFY OUR ORGANIZATION'S KEY BUSINESS DRIVERS FOR DATA SCIENCE.....	41
2.CREATE AN EFFECTIVE TEAM FOR ACHIEVING DATA SCIENCE GOALS.....	42
3.EMPHASIZE COMMUNICATION SKILLS TO REALIZE DATA SCIENCE'S VALUE	43
4.EXPAND THE IMPACT OF DATA SCIENCE THROUGH VISUALIZATION AND STORYTELLING.....	44

5.GIVE DATA SCIENCE TEAMS ACCESS TO ALL THE DATA	45
6.PREPARE DATA SCIENCE PROCESSES FOR OPERATIONALIZING ANALYTICS	46
7.IMPROVE GOVERNANCE TO AVOID DATA SCIENCE “CREEPINESS”	47
CHAPTER 4. SPECIFIC TOOLS AND SOFTWARES.....	48
WHAT TOOLS DO EMPLOYERS WANT DATA SCIENTISTS TO KNOW?	48
TOP TOOLS FOR DATA SCIENTISTS: ANALYTICS TOOLS, DATA VISUALIZATION TOOLS, DATABASE TOOLS, AND MORE	50
CHAPTER 5. DATA SCIENTIST.....	102
HOW TO BE A DATA SCIENTIST?	102
<i>Learn python well.</i>	102
<i>Learn the statistics.</i>	102
<i>Learn data processing</i>	104
<i>Become a full stack engineer.</i>	104
<i>Keep reading</i>	105
THREE CORE SKILLS THAT DATA SCIENTISTS NEED.....	106
<i>Data Hacking</i>	106
<i>Problem Solving</i>	107
<i>Communication</i>	107
BACKGROUND OF MAJORS TO BE A DATA SCIENTIST	107
MASTER OR PH.D. DEGREES?	109
CONCLUSION	111
REFERENCES.....	114

FOREWORD

*The concept of **DATA SCIENCE** is so much popular at present,
what is **DATA SCIENCE** indeed?
Are you interested in doing **DATA SCIENTIST**?*

Before we start, think about these question:

- Do you hate programming?
- Are you tired of mathematics?
- Is your ability of communication with others very poor?

If there are YES, this article may not be suitable for you. You can also pass automatically other articles about data science/scientist.

This article will focus on introducing data science and business strategy of data science, talking about what data scientists do, what professions are preferred, and what skills they need.

What is Data Science?

With the development of science and technology, the scale of data in human society has grown rapidly, and a large amount of data has been generated and stored every moment and every day. For example, there are such a number of social networking websites, from day to night record our location, our click and connection of all the links, all kinds of data are stored, they are not afraid of such a huge amount of data, what they are afraid is the recorded data is not sufficient.

Recently a so-called UrtheCast company installed the first civilian high-resolution camera directly on the International Space Station, taking pictures of the Earth and recording 2.5T data a day.

The increase in the amount of data and the diversification of data have also prompted many companies in the world to conduct data analysis to support data driven decision making.

For example, the supermarket found that you have been buying certain types of diet foods for the past three weeks. They can predict that you will continue to purchase this product. When you pay the bill, you will be gifted a printed coupon directly, “buy 4 get 1 free”, and you will be promoted. You feel that you are more willing to come to this store after taking advantage of it. The supermarket also keeps you tightly by giving a discount, lest you go to other supermarkets. This is called **Predictive Analytics: Analyze data to predict what might happen in the future.**

The supermarket's analytics team analyzed and found that the diet you bought has certain characteristics, such as low sodium, low carbs, and other related foods with such characteristics, the supermarket can also recommend it to you. You feel happy because the supermarket directly tells you the products you need, saves you from the trouble of comparing and purchasing items, the supermarket sells more things to you, earns money, and of course is very happy too. This is called **Descriptive Analytics: Analyze data to find out the characteristics of past events and trends in what is happening.**

On Valentine's Day, the price of rose at the flower shop was increased 100%; the supermarket concluded that the demand for condoms was very high, directly increasing the price by 20%; After Valentine's Day, the roses were discounted, contraception the price of condoms returns to normal. The supermarket has maximized its profits. This is called **Prescriptive Analytics: Analyze the data to find the best measures and get the best results.**

Predictive Analytics, Descriptive Analytics, Prescriptive Analytics, these three English names are organized by INFORMS.

Everyone in Italy, I believe that the supermarkets around you should not be so "intimate" and so savvy. At present, the traditional practice of supermarkets is to blindly deliver all kinds of advertisements (flyer) and all coupons that may be used to all residents in the vicinity. The supermarkets know nothing about what residents really need, they spend money printing and mailing a large number of flyers and coupons, whereas the utility of coupons is with a very low probability (such as 5%), and most of them become garbage directly.

With the rapid development of analytics/data science, the scenes I describe above are being gradually implemented. One of the most famous is the analytics team of Target (a major US retailer) who analyzes changes in customer spending behavior, such as guessing that some customers are likely to become pregnant and mail advertisements for pregnant women and baby products to their homes. When customers are attracted to the target to buy these products, they will also buy other things, and Target will make money. But this also brought an unexpected result: a teenage pregnancy was targeted, and her parents realized that her daughter was pregnant after receiving the target advertisement! Target knows earlier than parents!

A company wants to advertise, there are multiple choices: search engine, various social media, traditional media, in the end where should they put the money into so it will bring the greatest return?

As for your web page clicks, amazon adjusts the order in which the products are displayed, recommends the products you are most interested in, or you modify the skills & projects in the LinkedIn profile. This company automatically recommends matching

jobs from your connections. **Behind these smart, accurate, and real-time decisions, all are data science.**

In addition, the term **data science** is more commonly used in the IT industry, and other industries (such as Target retailers) are often referred to as **analytics**. As long as a profession is essentially analyzing a large amount of unregulated data, crunch the numbers to support decision making, that is data science; the person who does this kind of work is data scientist, regardless of your specific job title.

There are also those who, even rightly, dispute the qualification of "scientific" to the DATA SCIENCE approach. Gary Angel, CEO of DIGITAL MORTAR. He starts by recognizing that "Science works via the" Scientific Method "as follows:

- 1) Formulate a question or problem statement
- 2) Generate a hypothesis that is testable
- 3) Gather / Generate data
- 4) Analyze data to test the hypotheses / Draw conclusions
- 5) Communicate results to interested parties or take action

What's more, the scientific method is popularly elaborated is almost contentless. Strip away the fancy language and it translates into something like this:

- 1) Decide what problem we want to solve
- 2) Think about the problem until we have an idea of how it might be solved
- 3) Try it out and see if it works
- 4) Repeat until we solve the problem
- 5) Does this feel action guiding and powerful?

The idea that this type of problem seems to be the reason for the success of the science that is implausible on its face and is contradicted by experience.

Implausible because the method as described is so contentless. How do I pick which problems to tackle from the infinite set available? The method is silent. How do I generate hypothesis? The method is silent. How do I know they are testable? The method is silent. How do I test them? The method is silent. How do I test my hypothesis? The method is silent. How many failures to refute a hypothesis is enough to prove it? The method is silent. How do I communicate the results? The method is silent.

It seems very difficult to challenge this point of view. However, we can overlook the question of whether DATA SCIENCE can really be considered a "scientific" method in all the ways and fields of application in which it is currently proposed. Much more simply, we can limit ourselves to understanding DATA SCIENCE as the collection of all methods and methodological tools useful for "making the data speak (= let the data

speak)". Methods and methodological tools will certainly be scientific, while their applications will certainly be less: perhaps because, although having to accept uncertainty and / or non-quantifiable risks (such as, for example, the subjective judgments of Bayesian estimates) they present themselves as the only way to tackle the problem with a minimum of rationality, as an alternative to blindly relying on good or bad luck!

INTRODUCTION

Dataology and Data Science are sciences of data, defined as theories, methods, and techniques for exploring the mysteries of data in Cyberspace.

There are two main connotations: one is to study the data itself; the other is to provide a new method for natural science and social science research, called the scientific method of data.

Definition of Data Science

Currently, according to WIKIPEDIA, DATA SCIENCE is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data. Computer science, in particular from the subdomains of machine learning, classification, cluster analysis, data mining, databases, and visualization. In November 1997, C.F. Jeff Wu gave the inaugural lecture entitled "Statistics = Data Science?" for his appointment to the H. Carver Professor at the University of Michigan. In this lecture, he worked as a trilogy of data collection, data modeling and analysis, and decision making. In his conclusion, he initiated the modern, non-computer science, the use of the term "data science" and advocated that statistics be renamed data science and statistical data scientists. In 2001, William S. Cleveland introduced data science as an independent discipline, extending the field of statistics to incorporate "advances in computing with data" into his article "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics," which was published in Volume 69, No. 1, of the April 2001 edition of the International Statistical Review/Revue Internationale de Statistique. In his report, Cleveland establishes six technical areas that he believes to encompass the field of data science: multidisciplinary investigations, models and methods for data, computing with data, pedagogy, tool evaluation, and theory. Although use of the term "data science" has exploded in business environments, many academics and journalists see no distinction between data science and statistics.

Informatization is the process of storing data in the real world by storing things and phenomena in the form of data into the CYBER space. These data are a representation of nature and life that also document human behavior, including work, life, and social development. Today, data is rapidly and massively produced and stored in the CYBER space. This phenomenon is called data explosion, and data explosion forms the data nature in the CYBER space. Data is the only existence in the CYBER space, and it is necessary to study and explore the laws and phenomena of data in the CYBER space. In addition, exploring the laws and phenomena of data in the CYBER space is an important means of exploring the laws of the universe, exploring the laws of life,

finding the laws of human behavior, and finding the laws of social development. For example, we can study life by studying data (Bioinformatics), research on human behavior (behavioral informatics). Dataology and Data Science (hereafter referred to as data science) are sciences of data science or research data, defined as: researching theories, methods, and techniques for exploring the mysteries of datanatures in Cyberspace. The object is the data in the data world. Unlike the natural sciences and social sciences, data science and data science are the data of Cyberspace and are new sciences. There are two main connotations in data science and data science: one is to study the data itself, to study the various types, states, attributes, and forms and changes of the data; the other is to provide a new method for natural science and social science research. The data method called scientific research aims to reveal the phenomena and laws of nature and human behavior.

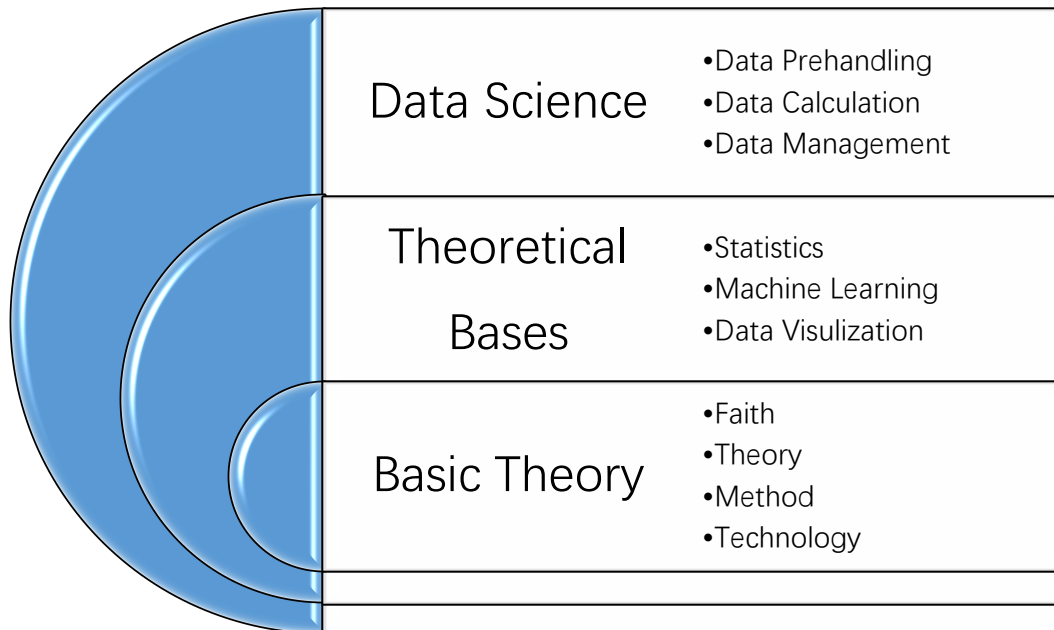
There are already some methods and techniques in data science, such as: data acquisition, data storage and management, data security, data analysis, visualization, etc.; also need basic theories and new technologies, such as: data existence, data measurement, time, data algebra, data similarity and cluster theory, data classification and data encyclopedia, data camouflage and recognition, data experiments, data perception and so on. The theory and methods of data science will improve existing scientific research methods, form new scientific research methods, and develop specialized theories, techniques, and methods for each research field to form specialized fields of data science, such as behavioral data science, life data science, brain data science, meteorological data science, financial data science, geographic data science, and so on.

History

Data science was proposed in the 1960s, but it was not noticed and recognized by the academic community at that time. In 1974, Peter Naur published the "Concise Survey of Computer Methods" which defined data science as: "The science of processing data, Once the relationship between the data and its representative things is established, it will provide lessons for other fields and sciences." The "Data Science, Classification and Related Methods" held in Japan in 1996 has made data science the subject of the conference. In 2001, William S. Cleveland, a professor of American statistics, published "Data Science: An Action Plan for Expanding Statistics in the Field of Technology," so some believe that Cleveland first identified data science as a separate discipline and defined data science as statistics. The expansion of the field of study to the combination of data as a cash calculation object lays the theoretical foundation for data science.

Research Content

- Basic theoretical research. The basis of science is observation and logical reasoning. It is also necessary to study the observation methods in the data nature. It is necessary to study the theory and methods of data reasoning, including: the existence of data, data measurement, time, data algebra, data similarity and cluster theory, data. Classification and data encyclopedias, etc.



- Experimental and logical reasoning methods. It is necessary to establish experimental methods of data science. It is necessary to establish many scientific hypotheses and theoretical systems, and to carry out exploration and research of data nature through these experimental methods and theoretical systems, so as to understand various types, states, attributes, and forms and changes of data. Reveal the phenomena and laws of nature and human behavior.
- Field data research. The theory and methods of data science are applied in many fields to form specialized fields of data such as brain data, behavioral data science, biometrics, meteorological data science, financial data science, geographic data science, and so on.
- Research and development methods and technical research of data resources. Data resources are important modern strategic resources, and their importance will become more and more prominent. It is likely to surpass oil, coal and minerals in this century and become one of the most important human resources. This is because human society, politics and economy will rely on data resources, and oil, coal, mineral resources and other resources exploration, mining, transportation,

processing, product sales, etc. are all dependent on data resources, leaving the data resources. These jobs will not be carried out.

Knowledge System

Data science is based on statistics, machine learning, data visualization and knowledge in a certain field. Its main research contents include basic theory of data science, data preprocessing, data calculation and data management.

- Basic theory: new ideas, theories, methods, techniques and tools in data science and the research purposes, theoretical basis, research content, basic processes, main principles, typical applications, personnel training, project management, etc. of data science. What needs special reminder here is that "basic theory" and "theoretical basis" are two different concepts. The "basic theory" of data science is within the research boundary of data science, and its "theoretical basis" is beyond the research boundary of data science, which is the theoretical basis and source of data science.
- Data preprocessing: In order to improve data quality, reduce the complexity of data calculation, reduce the amount of data calculation and improve the accuracy of data processing, data science needs to preprocess raw data - data auditing, data cleaning, data transformation, Data integration, data desensitization, data specification and data annotation.
- Data Computing: In data science, computing models have undergone fundamental changes—from traditional computing such as centralized computing to distributed computing to grid computing to cloud computing. There is a certain representativeness of the emergence of Google Cloud Computing 3 technologies, Hadoop, MapReduce and YARN technologies. Changes in the data computing model mean that the main goals, bottlenecks, and contradictions of data computing in data science have undergone fundamental changes.
- Data Management: After completing the “data preprocessing” (or “data calculation”), we need to manage the data in order to carry out (re-doing) “data processing” and data reuse and long-term storage. In data science, data management methods and technologies have undergone fundamental changes—not only traditional relational databases, but also emerging data management technologies such as NoSQL, NewSQL, and relational clouds.
- Technology and tools: The technologies and tools used in data science have a certain degree of expertise. Currently, R language is one of the most commonly used tools for data scientists.

Relationship with Other Disciplines

Data is something that exists in the CYBER space; information is a phenomenon that exists and occurs in nature, human society, and human thinking activities; knowledge is the understanding and experience that people gain in practice. Data can be used as a symbolic representation or carrier of information and knowledge, but the data itself is not information or knowledge. The object of data science research is data, not information, nor knowledge. Access to information and knowledge through research data to gain an understanding of nature, life and behavior. The research objects, research purposes and research methods of data science are fundamentally different from the existing computer science, information science and knowledge science.

Natural science studies natural phenomena and laws, and the object of understanding is the whole nature, that is, the various types, states, attributes, and forms of movement of matter in nature. Behavioral science is the science of studying human behavior and low-level animal behavior in natural and social environments. The recognized disciplines include psychology, sociology, social anthropology, and other similar disciplines. Data science supports research in the natural sciences and behavioral sciences. As data science progresses, more and more scientific research work will be directed toward data, which will enable humans to understand the data and thus understand nature and behavior.

Human beings explore the realm of nature, using computers to deal with human discoveries, human society, nature and people. In this process, data has been generated in abundance and is experiencing a big bang, and humans have unwittingly created a more complex data. nature. Since the second data explosion, people have lived in the real world of nature and the natural world of data. The history of people, society and the universe will become the history of data. Humans can explore the natural world by exploring the natural world of data. Humans also need to explore the phenomena and laws unique to the nature of data. This is the task of giving data science. It can be expected that all current scientific research fields may form corresponding data sciences.

System Framework

The working process of data research is: obtaining a data set from the data nature; exploring the overall characteristics of the data set; performing data research and analysis (such as using data mining technology) or conducting data experiments; discovering data laws; Perception and so on.

CHAPTER 1. "PROGENITORS" OF DATA SCIENCE

—12 main methodological tools

1. Run Chart
2. Classical Problem Solving (Pareto)
3. Short Term Target (Six Sigma Metrics)
4. Multiple Regression
5. Diagnosis by Signatures
6. Design Of Experiments (DOE)
7. Shainin Methods
8. Genetic Algorithm
9. Clustering
10. Principal Component Analysis (PCA)
11. Bayesian Approach for Reliability Predictions
12. Neural Networks

1. RUN CHART

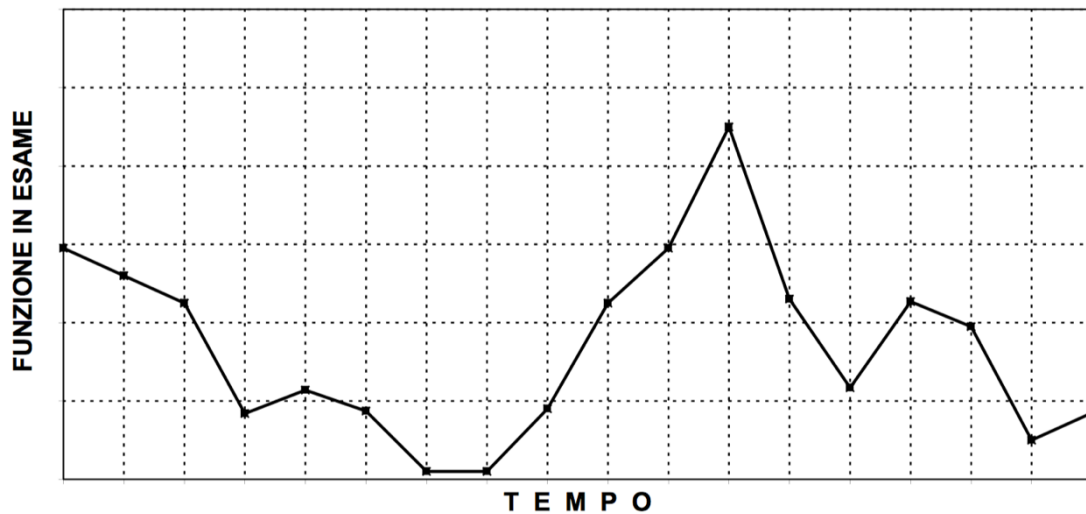


Figure 1. Example of RUN CHART

- graphing of data as a function of time
- to record and view any trends (trends) of data in relation to time
- the most significant changes are detected in a process.

2. CLASSICAL PROBLEM SOLVING (PARETO)

“Vital few”: A relatively small number of causes are responsible for a disproportionately large fraction of a given effect.

In itself, the principle does not solve the problems: it only indicates where the efforts must be concentrated. But this is of enormous importance.

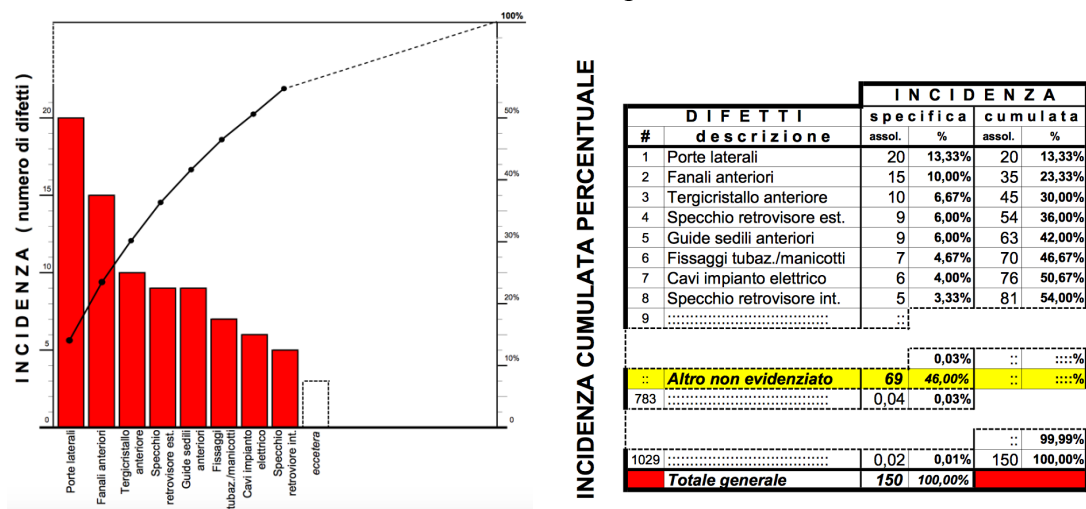


Figure 2. OPERATIVE and GRAPHICAL ASPECT

RULES are usually following:

- Consider a number of items up to capture 80% of the phenomenon investigated;
- It is accepted to neglect the first 2 to 3 entries, if particularly demanding from the point of view of the investments, but no more!

In the Problem Solving, it is typical to use the principle of Maldistribution of Pareto to highlight, through a cascade use of appropriate histograms, the main aspects on which to concentrate the efforts, avoiding to disperse in things of little importance.

We often resort to multiple levels of analysis. Typically, those shown below and illustrated in the following slide:

- 1st level: Pareto of the worst components (or subsystems);
- 2nd level: Pareto of the main defects for each of the components (or subsystems) that it was considered appropriate to analyze;
- 3rd level: Pareto of the causes of each of the defects that it is considered opportune to analyze.

This is generally followed by a dispersion analysis, case by case, with reference to the corresponding tolerances prescribed.

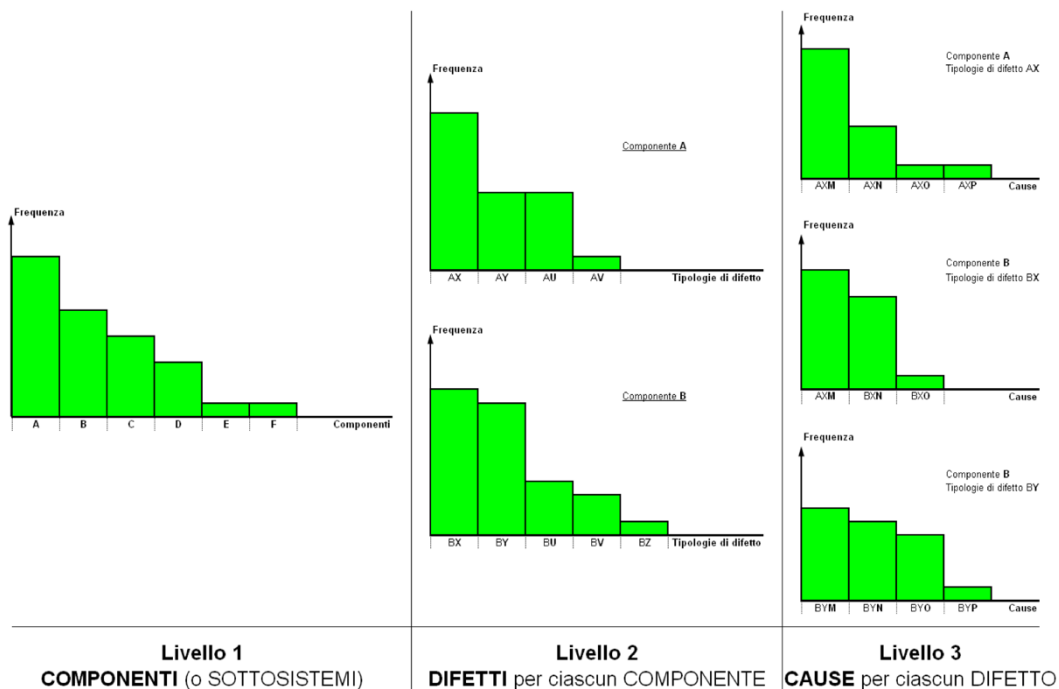


Figure 3.

3. SHORT TERM TARGET (Six Sigma Metrics)

COMMON, SPECIAL AND SEMICOMMON CAUSES

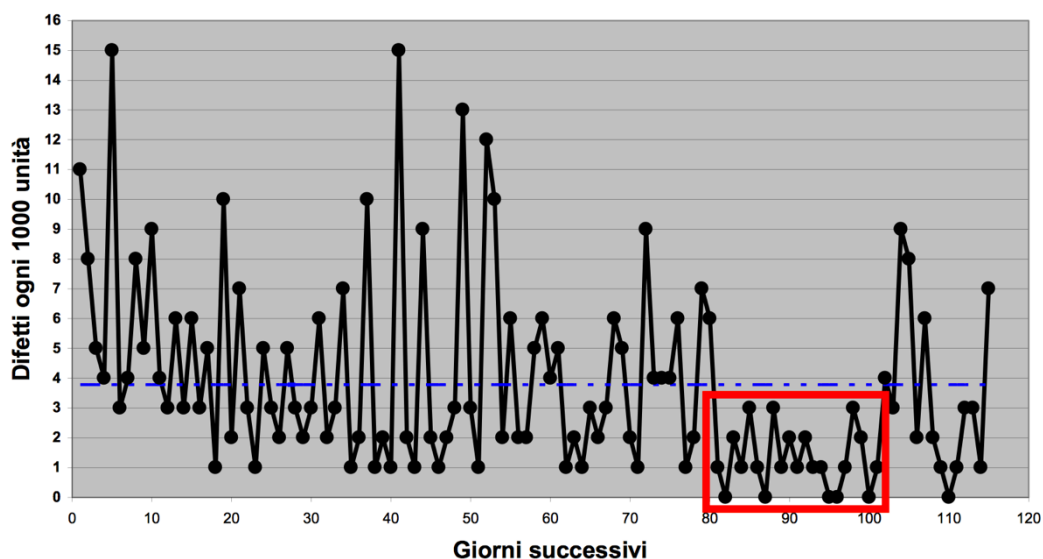
For the sake of simplicity, we have focused on the sole consideration of common causes and special causes.

To derive important benefits from this type of reasoning, it is also appropriate to consider another category of causes, which we could call semicommon, in the sense that they are common causes, because they are accepted as intrinsic to the process considered, but only occur when certain circumstances (eg: seasonal variations in temperature / humidity, supplies of different qualities, change of tools used, etc.), and therefore diluted over time, ie in the Long term. In any case, the semi-common causes are usually accepted as an integral part of the production system and therefore there is no plan to intervene to eliminate them (except during the complete reconstruction of the process), in full analogy with the common causes.

We will try to ensure that the data collection period in terms of short term is so short that these semi-common causes have no way of manifesting themselves (i.e. they remain sensibly on the same values throughout the short term period).

INSULATION OF SHORT TERM SEQUENCES FROM DATA COLLECTED DURING A LONG PERIOD (LONG TERM)

With the previous definitions, the results of Short term can be considered as the maximum quality limit obtainable in the Long term, provided, of course, that we



managed to minimize the effects of semi-common causes.

Figure 4. Time trend of the average number of defects per 1000 units

From this point of view, an evaluation of the Short term results can also be obtained by isolating, on a Control Chart, the results relating to a particularly good period. Even that (in addition to the short initial period of use of the machinery when it is still new) can be considered a Short term: indeed, it even presents a greater validity / credibility as an operational objective for the long term.

The data subgroup highlighted with a red rectangle shows a sequence of values with a low amount of defects and limited variability (with respect to the whole of all data). It therefore represents a particularly good performance and is therefore indicative of the Short Term Capability.

The objective will therefore be to identify the operating conditions in these circumstances, in order to be able to keep them as long as possible in the long term (Long term).

- The comparison between what the process has done lately (meant as an average of several months and therefore of long term = Long term) and what the process could do in optimal conditions, but real (therefore in a short enough period to be considered exempt from special causes and from semi-common causes = Short term), can provide an important measure of the space that is there for improvements.
- The difference between these two situations highlights the present level of process control.
- This is confirmed, in some way, by the Motorola hypothesis (already described) which, in the absence of specific data, is the most convenient prediction to calculate the amount of faulty in the long term (long term) as related to a constant process drift equal to 1.5 multiplied by the standard short term deviation.

Basically, the short term data serve to highlight the best that the process is able to do, under ideal operating conditions, but realistic.

IN CONCLUSION:

- Short term is a sufficiently short period to be exempt, not only from special causes, but also from semi-common causes: and therefore subject only to common causes in the strict sense. Consequently, periods of this type represent the best performances obtainable from the production process in question.
- The Long term is instead a longer period of time, always free of special causes, but affected by both common and semi-common causes.

C_p and C_{pk} in comparison with P_p and P_{pk}

P_p and P_{pk} designate the values of C_p and C_{pk} in the short term.

In order to guarantee acceptable values of C_p and C_{pk} in the Long term, it is obviously necessary that the P_p and P_{pk} values detected in the short term (for example during the check for acceptance of a new machine) are suitably higher (of those envisaged for C_p and C_{pk} in the long term).

THE MARGIN IMPROVED BY THE DIVARY BETWEEN "SHORT TERM AND" LONG TERM "

- The Long term can be thought of as the collection of all the data collected during a (substantial) period of time.
- Instead, Short term represents one or more sequences of data with better performance.

The purpose of the analysis of Long term data is therefore to:

- ✧ recognize one or more sequences of data with the best performance;
- ✧ go back to the specific causes of the gap between Short term and Long term services;
- ✧ remove as many of the identified causes as possible.

The Short term performances can therefore be used as a target for the Long Term production.

The difference between the performance of the process in the short (short term) and in the long (long term) period provides a measure of the room for maneuver to improve. Each of the two performances can be quantified by the respective value of the sigma, **z level**, in the two different situations, i.e. using once only the short term data for the calculation and once again all the data available in a definitely longer period (long term). It is worth noting that the sigma, z level depends on both the variability of the process (expressed by its standard deviation s) and the drifts of its mean and therefore constitutes a complete indicator (analogous to C_{pk}).

The difference between the two values of z is called z_{SHIFT} and the desired margin is quantified.

Calculation of z_{SHIFT}

z_{SHIFT} is the difference between the values of z short term, z_{ST} and z long term, z_{LT}:

$$z_{SHIFT} = z_{ST} - z_{LT}$$

z_{SHIFT} is the measure of how far the process is from ideal conditions

z_{SHIFT} is also the measure of the potential improvement that can be pursued

Note: having good control (no or little difference between Long Term and Short Term) does not assure us that the process produces few non-compliant (this depends on the comparison of the process distribution with the technical specifications), but tells us that the process is conducted in an optimal manner, so as to make it work close to its maximum possibilities.

In practice, we can proceed as follows:

- 1) We collect data for a certain period (Long term) and calculate the mean, \bar{x}_{LT} , and the standard deviation of data, s_{LT} . With these parameters, we can calculate the distance of \bar{x}_{LT} from the specification limit in waste units type s_{LT} which constitutes the sigma of long term sigma, z_{LT} .
- 2) If the long-term period is very long, it may be appropriate to subdivide it into a certain number of subperiods, somehow homogeneous, to calculate z_{LT} for each of them, and then to average the various z_{LT} calculated in this way. and then use this \bar{z}_{LT} (instead of z_{LT} as in the previous point), because otherwise, the effect of drift would tend to "compensate" from one time to another and would end up disappearing.
- 3) A sequence of consecutive data with low variability is identified, which will constitute the short term and which, using \bar{x}_{ST} and s_{ST} , in analogy to the above, will allow to calculate the sigma of short term, z_{ST} .
- 4) The difference is calculated $z_{SHIFT} = z_{ST} - z_{LT}$.

The foregoing can also be exploited to make graphically visible the current position of the company in relation to the concrete possibilities of improvement.

This is provided by the **Control/Technology Plot** or **Control vs. Graph. Technology Plot**.

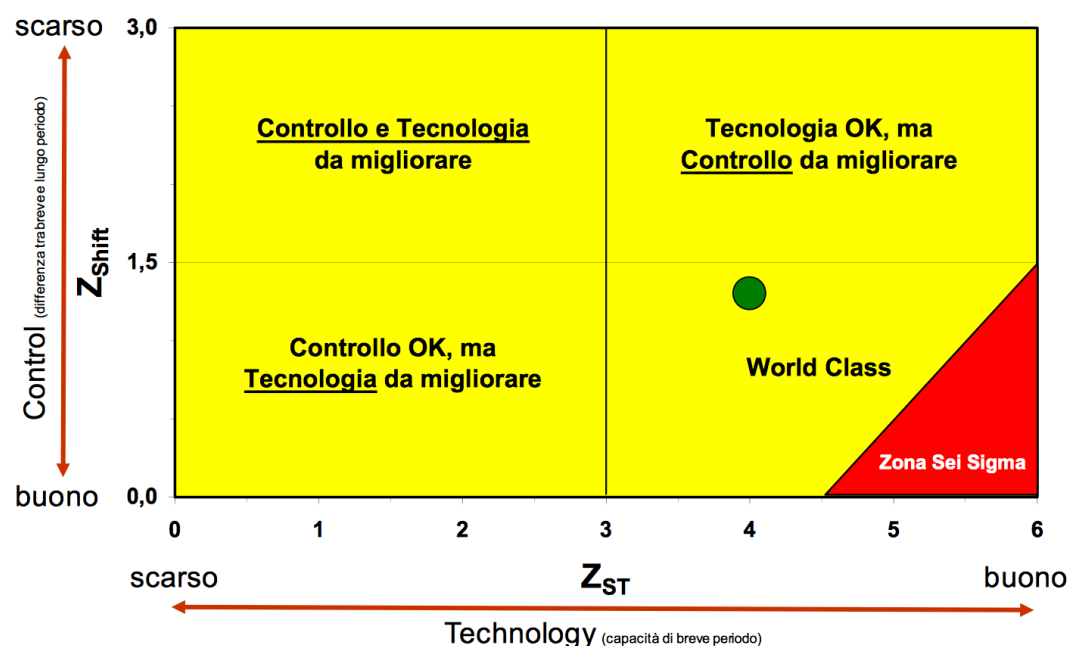


Figure 5. Control/Technology Plot

Putting the short term sigma level, z_{ST} (which represents the best possible operating conditions) into the abscissa, and comparing it to the current deviation from these conditions, expressed by z_{SHIFT} , immediately gives the idea of where they are positioned the performance of the process and which are the main areas on which to intervene to improve performance. In this way it frames the priority issues.

4. MULTIPLE REGRESSION

ONLY, IT IS ALREADY ALREADY OF A PARTICULAR DATA ON A DETERMINED PHENOMENON. THEREFORE, ONCE THE MOST IMPORTANT GRANDEZZES DEFINED TO BE INVESTIGATED,

- a data collection form must be set up
- providing it with all the necessary information to be able to analyze the data themselves by statistical method.

The goal of statistical methods is, in this case, to assign the right importance ("significance") to the magnitudes that influence the phenomenon.

When the phenomenon is influenced by ONE GREATNESS, the typical tools to tackle the problem are:

- Diagram of correlation: we start from a graph with a cloud of points and we trace an interpolating line above it;
- One-way regression (or controlled factor): the previous operations are mathematically developed and, as an interpolating line of the points, although in the vast majority of cases the straight line (linear interpolation) is continued, if the software allows it, we can also use other curves, e.g. quadratic or cubic paraboles (non linear interpolation).

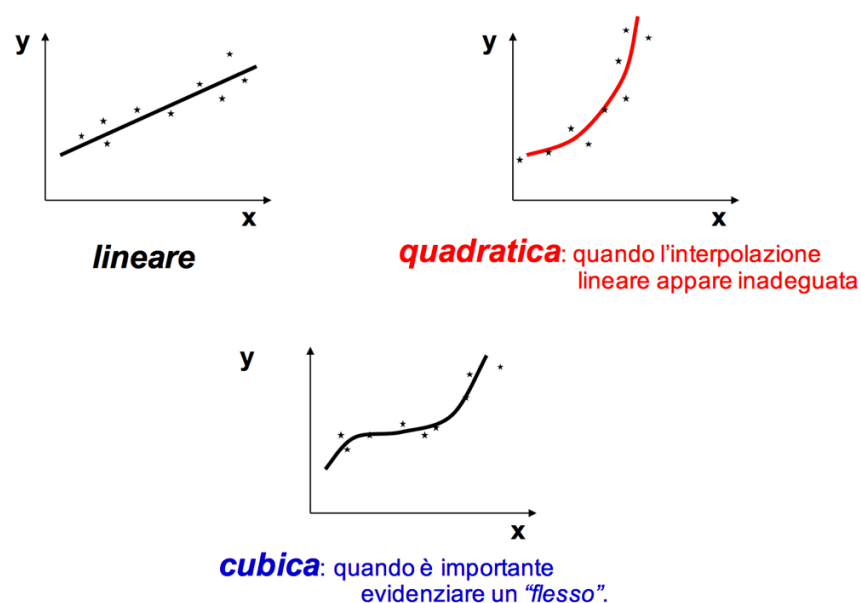


Figure 6. Regression Model of Single Grade

When the phenomenon is affected by MORE GRANDEZZE, the typical tools to tackle the problem are:

- Multiple Regression: operates as the one-way Regression, but taking into account more magnitudes at the same time;
- Experimental Design: method based on criteria similar to those of the Multiple Regression, but which provides for the optimized definition of a plan of experimental tests aimed at the specific problem, and which is therefore used
 - ✧ or in the absence of pre-existing data (because more effective than Multiple Regression)
 - ✧ or to optimize knowledge of the phenomenon.

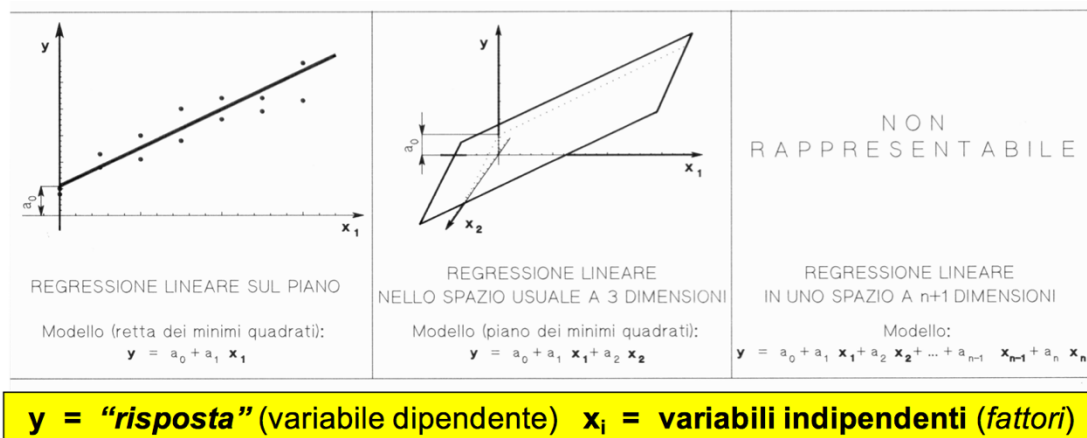


Figure 7. Regression Model of Multi Grades

- Each of the coefficients a_i is (also) normalized on the maximum excursion (range) found on the values provided for the variable to which it refers; the ratios so re-elaborated become independent of the units of measurement used and their size can act as an index of the importance of the single variables on the y response.
- The deviations between the values actually found and the values foreseen by the regressive model express the "residual" variance, i.e. "not explained" by the variables (factors) considered in the model itself and will be submitted to the Residual Analysis.

Multiple Regression is a Multivariate Analysis tool that is intended to determine

- 1) the most influential quantities on the y response and their weight;
- 2) a relationship, usually linear, that allows to make a prediction of the system's response according to the quantities identified as most influential; this mathematical relation is usually called the regressive model.

Note that here, unlike previous Chapters, the dependence of a variable (y) from other variables (x_i) becomes essential,

The Multiple Regression refers to a base table of the following type:

	X ₁	X ₂	X _i	X _n	y	y'
serie di dati 1				d ₁			r ₁	r' ₁
serie di dati 2				d ₂			r ₂	r' ₂
serie di dati 3				d ₃			r ₃	r' ₃
.....			
.....			
serie di dati (m-1)				d _{m-1}			r _{m-1}	r' _{m-1}
serie di dati m				d _m			r _m	r' _m

Table 8. an example

The n independent variables x₁, x₂, ..., x_n, head one column each, while each line contains the data collected in each test performed. The last column (with gray background) of the table contains the results found during the test. The external column (with a yellow background) shows the calculated results, under the same conditions, with the forecast model:

$$y' = f(x_i) = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$$

The smaller the sum of the squares of the differences between the real results y and those predicted y', the better is the adequacy of the regressive model to reality.

The **Multiple Regression** can also employ **qualitative variables**, that is not expressible numerically, such as, for example: different forms of aeration nozzle, different suppliers, work shifts, days of the week, etc. In these cases, it would be wrong to mark the values of the variables with numbers like 1, 2, 3, ..., because, in this way, their relationship would be forced (which obviously lacks any physical meaning.)

Instead, we can use the so-called dummy variables method. A qualitative variable at n levels (e.g.: n = 4 suppliers) is represented by n-1 binary variables of 0 or 1 (3 in the example of the 4 suppliers) defined as in the matrix of the example below

Significato	VARIABILE DUMMY		
	sottovariabili dummy		
	X ₄	X ₅	X ₆
Fornitore A	0	0	0
Fornitore B	1	0	0
Fornitore C	0	1	0
Fornitore D	0	0	1

Table 9. suppliers

The first line is formed only by zeroes and defines the reference variable (the supplier A). Starting from the second row (1 a column) there is a series of 1 diagonally. This technique makes it possible to highlight whether suppliers B, C and D are the same or different from the reference supplier A (but does not allow to directly understand if the suppliers found different from A are the same or different from each other).

The main OUTPUT:

- 1) quantities most influential in the investigated phenomenon, identified as the best compromise of the regressive model between precision, simplicity and expressivity from the physical point of view;
- 2) coefficients of the regressive model, optimized by the least squares method;
- 3) percentages of contribution in the "explanation" of the phenomenon investigated for each quantity included in the regressive model.

To achieve this, an EMPIRICAL RULE recommends that the number of available tests be at least 4 to 5 times higher than the number of variables investigated, X_i (actually included in the model).

5. DIAGNOSIS BY SIGNATURES

Focal points of MAINTENANCE:

- Corrective
- Preventive
- Reliability Centered Maintenance (RCM)
 - ✧ Mean Time Between Failures (MTBF)
 - ✧ FTA
 - ✧ FMECA

Monitoring and diagnostic techniques, because a failure can be preceded by a state of progressive degradation

	APPROCCIO TRADIZIONALE	APPROCCIO "AL BUIO"
CONOSCENZE DELL'ESPERTO	<ul style="list-style-type: none">• Tipo di applicazioni• Strumenti di misura• Eccetera	<ul style="list-style-type: none">• Segnali• Informazioni• Algoritmi

Figure 10. Monitoring and diagnostic techniques

Based on the Synthetic Failure Indices in maintenance strategy called Condition Based Maintenance can be established.

6. DESIGN OF EXPERIMENTS (DOE)

The design of experiments (DOE, DOX, or experimental design) is the design of any task that aims to describe or explain the variation of information under conditions that are hypothesized to reflect the variation. The term is generally associated with experiments in which the design introduces conditions that directly affect the variation, but may also refer to the design of quasi-experiments, in which natural conditions that influence the variation are selected for observation.

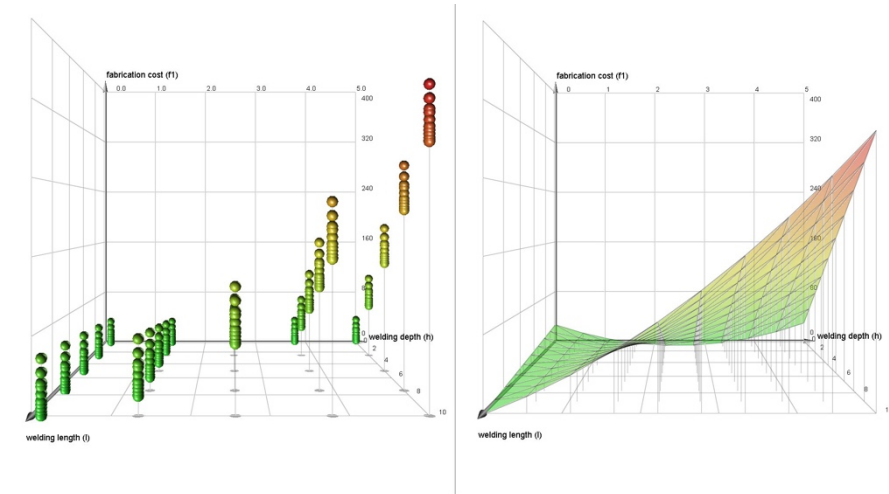


Figure 11. design of experiments

In its simplest form, an experiment aims at predicting the outcome by introducing a change of the preconditions, which is represented by one or more independent variables, also referred to as "input variables" or "predictor variables." The change in one or more independent variables is generally hypothesized to result in a change in one or more dependent variables, also referred to as "output variables" or "response variables." The experimental design may also identify control variables that must be held constant to prevent external factors from affecting the results. Experimental design involves not only the selection of suitable independent, dependent, and control variables, but planning the delivery of the experiment under statistically optimal conditions given the constraints of available resources. There are multiple approaches for determining the set of design points (unique combinations of the settings of the independent variables) to be used in the experiment.

Main concerns in experimental design include the establishment of validity, reliability, and replicability. For example, these concerns can be partially addressed by carefully choosing the independent variable, reducing the risk of measurement error, and ensuring that the documentation of the method is sufficiently detailed. Related concerns include achieving appropriate levels of statistical power and sensitivity.

Correctly designed experiments advance knowledge in the natural and social sciences and engineering. Other applications include marketing and policy making.

7. SHAININ METHODS

An Overview of the Shainin System™ for Quality Improvement

The Shainin System (SS) for quality improvement was developed over many years under the leadership of the late Dorian Shainin.

SS is also called Statistical Engineering by the consulting firm Shainin LLC that holds the trademark and Red X strategy in parts of the automotive sector where SS is popular. The overall methodology has not been subject to critical review although some of the components have been discussed extensively.

The Shainin System was developed for and is best suited to problem solving on operating, medium to high volume processes where data are cheaply available, statistical methods are widely used and intervention into the process is difficult. It has been mostly applied in parts and assembly operations.

The underlying principles of SS can be placed in two groups. The first group follows from the idea that there are dominant causes of variation. This idea appears in Juran and Gryna, but it is Shainin who fully exploits this concept. The second group of principles is embedded in the algorithm, the Shainin System

Dominant causes of variation and progressive search

A fundamental tenet of SS is that, in any problem, there is a dominant cause of variation in the process output that defines the problem. This presumption is based on an application of the Pareto principle to the causes of variation.

Juran and Gryna¹ define a dominant cause as, “a major contributor to the existence of defects, and one which must be remedied before there can be an adequate solution.” In SS, the dominant cause is called the Red X1. The emphasis on a dominant cause is justified since “The impact of the Red X is magnified because the combined effect of multiple inputs is calculated as the square root of the sum of squares”.

To clarify, if the effects of causes (i.e., process inputs that vary from unit to unit or time to time) are independent and roughly additive, we can decompose the standard deviation of the output that defines the problem as:

$$\text{stdev}(\text{output}) = \sqrt{(\text{stdev due to cause1})^2 + (\text{stdev due to cause2})^2 + \dots} \quad (1)$$

A direct consequence of (1) is being unable to reduce the output standard deviation substantially by identifying and removing or reducing the contribution of a single cause, unless that cause has a large effect.

For example, if (stdev due to cause 1) is 30 percent of the stdev (output), users can reduce the stdev (output) by only about 5 percent with complete elimination of the contribution of this cause. The assumption that there is a dominant cause (possibly because of an interaction between two or more varying process inputs) is unique to SS and has several consequences in its application.

Within SS, there is recognition that there may be a second or third large cause, called the Pink XTM and Pale Pink XTM respectively³, that make a substantial contribution to the overall variation and must be dealt with in order to solve the problem. Note that if there is not a single dominant cause, reducing variation is much more difficult, since, in light of (1), several large causes would have to be addressed to substantially reduce the overall output variation.

To simplify the language, we refer to a dominant cause of the problem, recognizing that there may be more than one important cause.

There is a risk that multiple failure modes contribute to a problem, and hence result in different dominant causes for each mode.

In one application, a team used SS to reduce the frequency of leaks in cast iron engine blocks. They made little progress until they realized that there were three categories of leaks, defined by location within the block. When they considered leaks at each location as separate problems, they rapidly determined a dominant cause and a remedy for each problem.

SS uses a process of elimination³, called progressive search, to identify the dominant causes. Progressive search works much like a successful strategy in the game “20 questions,” where users attempt to find the correct answer using a series of (yes=no) questions that divide the search space into smaller and smaller regions.

To implement the process of elimination, SS uses families of causes of variation. A family of variation is a group of varying process inputs that act at the same location or in the same time span. Common families include within-part, part-to-part (consecutive), hour-to-hour, day-to-day, cavity-to-cavity and machine-to-machine.

At any point in the search, the idea is to divide the inputs remaining as possible dominant causes into mutually exclusive families, and then to carry out an investigation that will eliminate all but one family as the home of the dominant cause.

8. GENETIC ALGORITHM

In computer science and operations research, a genetic algorithm (GA) is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms (EA). Genetic algorithms are commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operators such as mutation, crossover and selection.

Optimization problems

In a genetic algorithm, a population of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem is evolved toward better solutions. Each candidate solution has a set of properties (its chromosomes or genotype) which can be mutated and altered; traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.

The evolution usually starts from a population of randomly generated individuals, and is an iterative process, with the population in each iteration called a generation. In each generation, the fitness of every individual in the population is evaluated; the fitness is usually the value of the objective function in the optimization problem being solved. The more fit individuals are stochastically selected from the current population, and each individual's genome is modified (recombined and possibly randomly mutated) to form a new generation. The new generation of candidate solutions is then used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population.

A typical genetic algorithm requires:

- a genetic representation of the solution domain,
- a fitness function to evaluate the solution domain.

A standard representation of each candidate solution is as an array of bits. Arrays of other types and structures can be used in essentially the same way. The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, which facilitates simple crossover operations. Variable length representations may also be used, but crossover implementation is more complex in this case. Tree-like representations are explored in genetic programming and graph-form representations are explored in evolutionary programming; a mix of both linear chromosomes and trees is explored in gene expression programming.

Once the genetic representation and the fitness function are defined, a GA proceeds to initialize a population of solutions and then to improve it through repetitive application of the mutation, crossover, inversion and selection operators.

9. CLUSTERING

Clustering or group analysis (from the English term cluster analysis introduced by Robert Tryon in 1939) is a set of multivariate data analysis techniques aimed at the selection and grouping of homogeneous elements in a set of data. Clustering techniques are based on measures related to the similarity between the elements. In many approaches this similarity, or rather dissimilarity, is conceived in terms of distance in a multidimensional space. The goodness of the analysis obtained from the clustering algorithms depends a lot on the choice of the metric, and therefore on how the distance is calculated. The clustering algorithms group the elements on the basis of their mutual distance, and therefore whether or not belonging to a set depends on how much the element taken into consideration is distant from the whole itself.

To exemplify, even if in a simplified and rudimentary way, the basic idea, we decided to report on a three-dimensional graph a set of models of FIAT, LANCIA and ALFA ROMEO cars, according to their characteristics of:

- Luxury (measured simply by the cost in €)
- Habitability (measured by the length of the vehicle in cm)
- Engine power (measured in CV).

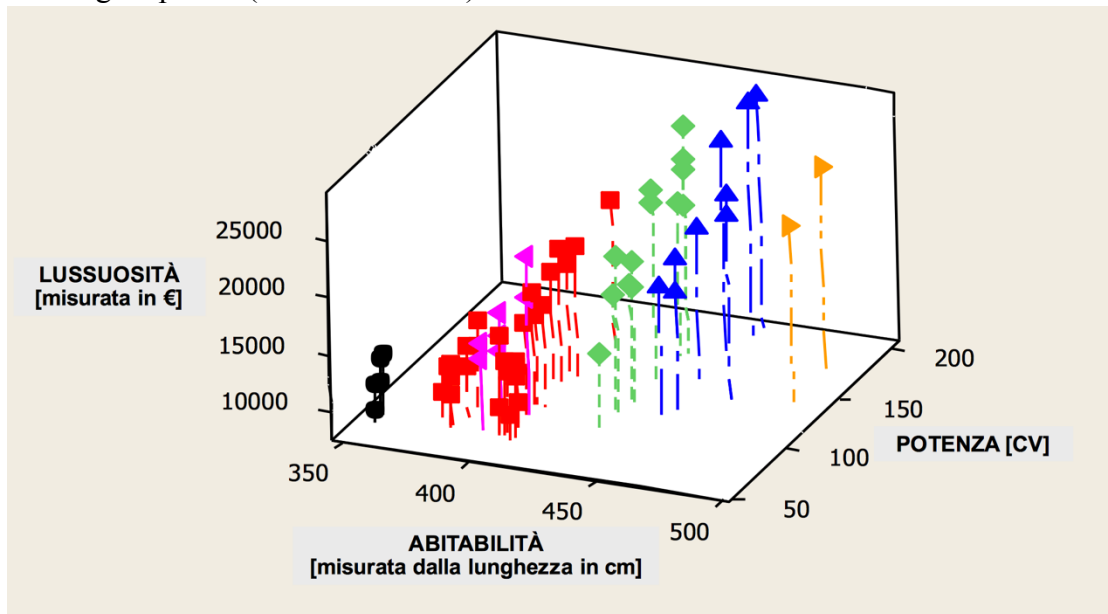


Figure 11. three-dimensional graph a set of models of FIAT, LANCIA and ALFA ROMEO cars

It is easy to identify the most evident groupings (clusters) and realize that they coincide substantially with the Market Segments (except that, in this simplified example, we can not distinguish between Segment B V.S. Segment L

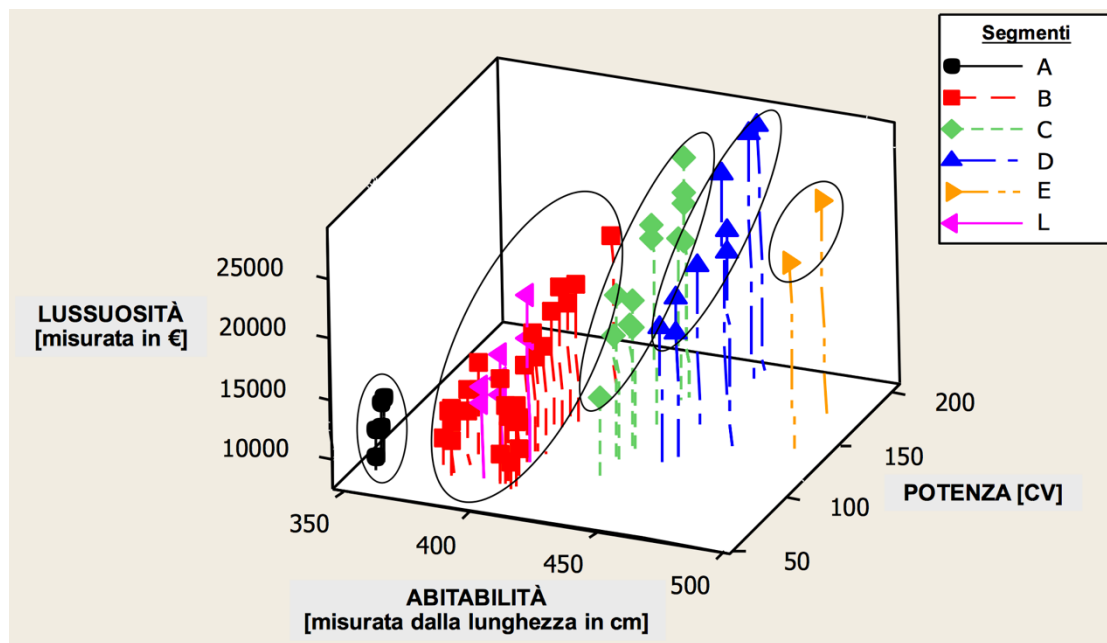


Figure 12. three-dimensional graph a set of models of FIAT, LANCIA and ALFA ROMEO cars

It can immediately be deduced that:

- who seems to have the main role in defining the Market Segment is habitability;
- for all clusters, costs grow in step with the power;
- the power tends to increase with increasing habitability.

It is very important to note that the above was obtained without having specified a variable (y) dependent on other variables (x_i), but simply by grouping neighboring spatial positions, independently of any cause-effect relationship.

Naturally, the previous three-dimensional graph can be decomposed into its 3 projections (see the following slide), which confirm the three observations made above.

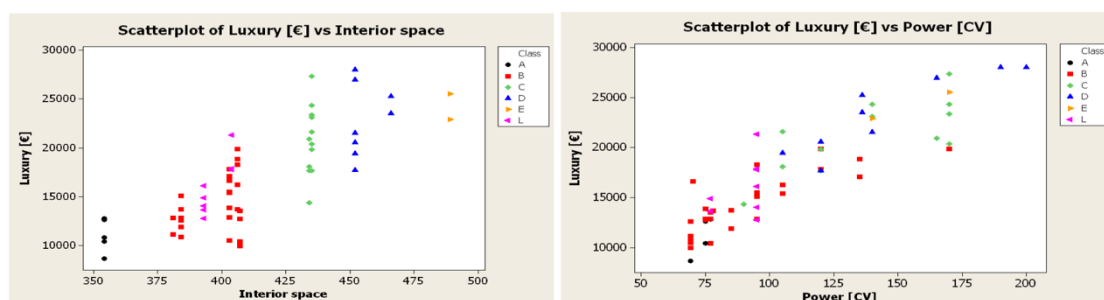


Figure 13. three-dimensional graph

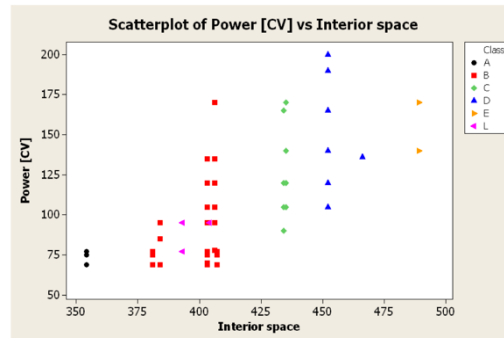


Figure 14. three-dimensional graph

The example we just saw was very simple and coarse. To provide an immediate and above all visual idea of the potential of the method, we had to limit ourselves to only 3 axes.

In reality, the **CLUSTER ANALYSIS**:

- it can operate in a multidimensional space with a multiplicity of axes (even if, usually, it does not exceed $50 \div 100$);
- depending on the applications, cluster formation can take place with diversified philosophies and techniques (as illustrated in the following slide);
- the aim is always to identify and unify homogeneous elements, but the idea of homogeneity and, with it, the relative selection criteria can change according to the needs.

10. PRINCIPAL COMPONENT ANALYSIS (PCA)

The fundamental purpose of **Principal Component Analysis (PCA)** is to reduce the number of variables necessary to describe the aspects under examination.

The set of classical variables, each of which represents a characteristic of the phenomenon under examination, is reduced to a **limited number of new variables**, called **latent variables**, selected on the basis of the quantity of "explained" variance up to a limit threshold, pre-established with subjective criteria.

The ranking is conducted on the basis of eigenvalues.

It is very important to underline that, precisely for how they have been generated (see below), the new variables are not in bi-annual correspondence with the classical ones of origin, if only because they are much less. In fact, they constitute, in a way, a synthesis and therefore each of them can express the contents of more than one of those of origin.

For this reason, it will be the task of the Specialists of the phenomenon under investigation to define the physical meanings of the new variables and to coherently assign them expressive names.

Without going into the details of the method, the following steps are summarized below in the simplest case:

- 1) normalization, in the new variable Z , of the single observations x_i of each classical variable X , subtracting from each observation x_i the mean \bar{x} of all x_i and dividing by their standard deviation s :

$$z_i = (x_i - \bar{x})/s$$

- 2) calculation of the correlation coefficient matrix, which quantifies the correlation between the initial classical variables X_i :

	X_1	X_2	X_3	...	X_{n-1}	X_n
X_1	1	$c_{1,2}$	$c_{1,3}$...	$c_{1,n-1}$	$c_{1,n}$
X_2	$c_{2,1}$	1	$c_{2,3}$...	$c_{2,n-1}$	$c_{2,n}$
X_3	$c_{3,1}$	$c_{3,2}$	1	...	$c_{3,n-1}$	$c_{3,n}$
...	1
X_{n-1}	$c_{n-1,1}$	$c_{n-1,2}$	$c_{n-1,3}$...	1	$c_{n-1,n}$
X_n	$c_{n,1}$	$c_{n,2}$	$c_{n,3}$...	$c_{n,n-1}$	1

Table 15. correlation of variables

Since the correlation coefficient of a variable with itself can only be equal to 1, the main diagonal of the matrix is entirely made up of 1 and is also symmetrical, in the sense that, even when it is $i \neq j$, the correlation coefficient between the variable X_i and the variable X_j will be the same as that between the variable X_j and the variable X_i , that is: $c_{i,j} = c_{j,i}$.

- 3) calculation of the eigenvalues of the previous matrix of correlation coefficients, which must then be presented in descending order. The ratio of each eigenvalue with the sum of all eigenvalues (multiplied by 100) expresses the percentage of variance that is "explained". The accumulation of the explained variance, starting from the most important eigenvalue, constitutes the criterion reference to establish (subjectively) to which successive eigenvalue stop in the selection of the new main variables.

It should be noted that, even here as in the case of clusters, the whole thing was obtained without having specified a variable (y) dependent on other variables (x_i),

11. BAYESIAN APPROACH FOR RELIABILITY PREDICTIONS

A classical analytical method for the preliminary estimate of the on field failure frequency at the beginning of the development of a new subsystem/component consists of the following steps.

- 1) Collection, based on data from the field (Dealers), of the failure frequencies of similar subsystems/components already on the market, at the elementary level of component/defect.
- 2) “Virtual” construction of the archetype, as close as possible to the design (in progress) of the subsystem/component in development.
- 3) List of all changes (improvements and/or cost containment) provided at the design and process level for the subsystem/component under development, with reference to the archetype.
- 4) “Subjective” prediction by Experts (designers, technologists, etc.) of the failure frequency for each elementary item (component/defect), based on the collected data and on all considered changes.
- 5) Prediction of the failure frequency for the whole subsystem in development through the sum of the predicted failure frequencies for the individual elementary items and subsequent comparison with the desired target.
- 6) If the result is unsatisfactory, repeat the above steps, starting from step 3, after changing some variants (with reference to the archetype), or have introduced new ones.

This method has:

- the advantage of being applied in the choice of initial design features, with great effectiveness of preventive actions;
- the obvious disadvantage of using subjective assumptions and therefore it requires a subsequent experimental verification;
- insidious disadvantage (why not so obvious) of being systematically optimistic, since it can not take into account that implemented changes reduce many of the current failure modes (reported by Dealers), but they can also introduce some new failure mode, difficult to identify at this stage.

We can provide a value for the statistical uncertainty, σ , in the following way:

- 1) The uncertainty is objectively related to the predictable difference between the achievable target (=what the experts have predicted the new model in development could reach) and the result (known) obtained on the archetype, where:
 - ✧ the achievable target is understood as the best result theoretically achievable and therefore it is considered as the upper limit of the confidence interval;
 - ✧ the result of the archetype is considered, at least at first approximation, as the worst achievable result (only worsened with cost reductions) and therefore assumed as the lower limit of the confidence interval.

The basic value of Bayesian uncertainty, σ_{basic} is given by this difference.

- 2) We acquire the subjective opinion of the Experts (Design and Technologies) about the credibility of the estimate of the achievable target (item by item) and, on this basis, we modify the previous basic value of Bayesian uncertainty, σ_{basic} defining the value σ_{taken} , to be taken.

For each achievable target the Designer is asked on what his prediction is based and what is the resulting level of confidence that we can attribute to it.

Kind of forecast	Description	Adjustments of (decreasing from the initial adopted value σ_{basic} equal to the difference between the result of the archetype and the achievable target)
Sure	Predicted targets are certainly achievable: or even already achieved.	$\sigma_{\text{taken}} = 0,25 \cdot \sigma_{\text{basic}}$
Plain	Predictions should be very reliable , because we actually have to identify, develop and test articulate modifications , but, in all comparable historical cases, we have almost always achieved the expected benefits and it is highly unlikely that the changes lead to new failure modes.	$\sigma_{\text{taken}} = 0,50 \cdot \sigma_{\text{basic}}$
Uncertain	Innovative/new solutions, on which we do not have much experience, but supported by a series of studies (FMEA, FTA, WCA, etc.) and/or experimental tests (Multiple Regression, Experimental Design, etc.) already available, which are considered probative.	$\sigma_{\text{taken}} = 0,75 \cdot \sigma_{\text{basic}}$
Risky	Innovative/new solutions, on which we do not have much experience, and that we do not feel sufficiently supported by experiments or studies already completed.	$\sigma_{\text{taken}} = \sigma_{\text{basic}}$

Figure 16. Forecast V.S. Adjustments of σ_{basic} and σ_{taken}

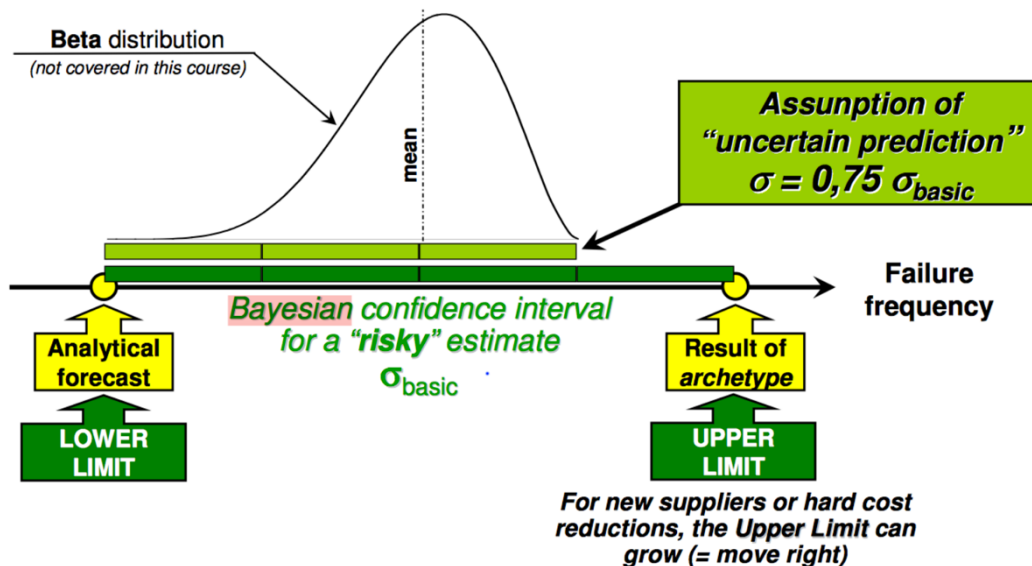


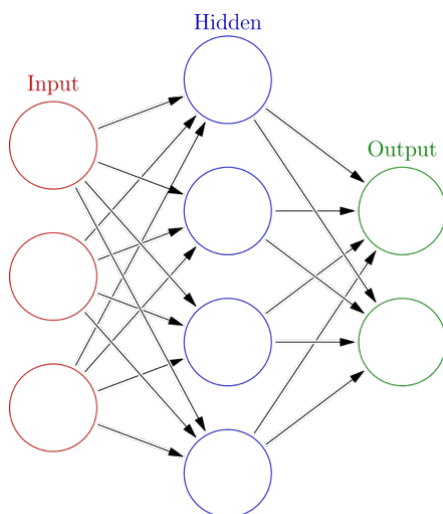
Figure 17. Bayesian distributions

12. NEURAL NETWORKS

Artificial neural networks (ANNs) or connectionist systems are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. For example, in image recognition, they might learn to identify images that contain cats by analyzing example images that have been manually labeled as "cat" or "no cat" and using the results to identify cats in other images. They do this without any prior knowledge about cats, e.g., that they have fur, tails, whiskers and cat-like faces. Instead, they automatically generate identifying characteristics from the learning material that they process.

An ANN is based on a collection of connected units or nodes called artificial neurons which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it.

In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called 'edges'. Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.



The original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. ANNs have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.

Figure 18. An artificial neural network is an interconnected group of nodes

CHAPTER 2. DATA SCIENCE, DATA ANALYTICS AND BIG DATA

Data Science vs. Big Data vs. Data Analytics

Data is everywhere. In fact, the amount of digital data that exists is growing at a rapid rate, doubling every two years, and changing the way we live. According to IBM, 2.5 billion gigabytes (GB) of data was generated every day in 2012.



An article by Forbes states that Data is growing faster than ever before and by the year 2020, about 1.7 megabytes of new information will be created every second for every human being on the planet, which makes it extremely important to at least know the basics of the field. After all, here is where our future lies.

We will differentiate between the Data Science, Big Data, and Data Analytics, based on what it is, where it is used, the skills we need to become a professional in the field, and the salary prospects in each field. Let's first start off with understanding what these concepts are.

What They Are

Data Science:

Dealing with unstructured and structured data, Data Science is a field that comprises of everything that related to data cleansing, preparation, and analysis.

Data Science is the combination of statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently, and the activity of cleansing, preparing and aligning the data.

In simple terms, it is the umbrella of techniques used when trying to extract insights and information from data.

Big Data:

Big Data refers to humongous volumes of data that cannot be processed effectively with the traditional applications that exist. The processing of Big Data begins with the raw data that isn't aggregated and is most often impossible to store in the memory of a single computer.

A buzzword that is used to describe immense volumes of data, both unstructured and structured, Big Data inundates a business on a day-to-day basis. Big Data is something that can be used to analyze insights which can lead to better decisions and strategic business moves.

The definition of Big Data, given by Gartner is, "Big data is high-volume, and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation".

Data Analytics:

Data Analytics the science of examining raw data with the purpose of drawing conclusions about that information.

Data Analytics involves applying an algorithmic or mechanical process to derive insights. For example, running through a number of data sets to look for meaningful correlations between each other.

It is used in a number of industries to allow the organizations and companies to make better decisions as well as verify and disprove existing theories or models.

The focus of Data Analytics lies in inference, which is the process of deriving conclusions that are solely based on what the researcher already knows.

The Applications of Each Field

Applications of Data Science:

- Internet search: Search engines make use of data science algorithms to deliver best results for search queries in a fraction of seconds.
- Digital Advertisements: The entire digital marketing spectrum uses the data

science algorithms - from display banners to digital billboards. This is the main reason for digital ads getting higher CTR than traditional advertisements.

- Recommender systems: The recommender systems not only make it easy to find relevant products from billions of products available but also adds a lot to user-experience. A lot of companies use this system to promote their products and suggestions in accordance with the user's demands and relevance of information. The recommendations are based on the user's previous search results.

Applications of Big Data:

- Big Data for financial services: Credit card companies, retail banks, private wealth management advisories, insurance firms, venture funds, and institutional investment banks use big data for their financial services. The common problem among them all is the massive amounts of multi-structured data living in multiple disparate systems which can be solved by big data. Thus big data is used in a number of ways like:
 - ✧ Customer analytics
 - ✧ Compliance analytics
 - ✧ Fraud analytics
 - ✧ Operational analytics
- Big Data in communications: Gaining new subscribers, retaining customers, and expanding within current subscriber bases are top priorities for telecommunication service providers. The solutions to these challenges lie in the ability to combine and analyze the masses of customer-generated data and machine-generated data that is being created every day.
- Big Data for Retail: Brick and Mortar or an online e-tailer, the answer to staying the game and being competitive is understanding the customer better to serve them. This requires the ability to analyze all the disparate data sources that companies deal with every day, including the weblogs, customer transaction data, social media, store-branded credit card data, and loyalty program data.

Applications of Data Analysis:

- Healthcare: The main challenge for hospitals with cost pressures tightens is to treat as many patients as they can efficiently, keeping in mind the improvement of the quality of care. Instrument and machine data is being used increasingly to track as well as optimize patient flow, treatment, and equipment used in the hospitals. It is estimated that there will be a 1% efficiency gain that could yield more than \$63 billion in the global healthcare savings.
- Travel: Data analytics is able to optimize the buying experience through the mobile/ weblog and the social media data analysis. Travel sights can gain insights into the customer's desires and preferences. Products can be up-sold by correlating the current sales to the subsequent browsing increase browse-to-buy conversions via customized packages and offers. Personalized travel recommendations can also be delivered by data analytics based on social media data.

- Gaming: Data Analytics helps in collecting data to optimize and spend within as well as across games. Game companies gain insight into the dislikes, the relationships, and the likes of the users.
- Energy Management: Most firms are using data analytics for energy management, including smart-grid management, energy optimization, energy distribution, and building automation in utility companies. The application here is centered on the controlling and monitoring of network devices, dispatch crews, and manage service outages. Utilities are given the ability to integrate millions of data points in the network performance and lets the engineers use the analytics to monitor the network.

The Skills Required of Each Field

To Become a Data Scientist:

- Education: 88% have a Master's Degree and 46% have PhDs.
- In-depth knowledge of SAS and/or R: For Data Science, R is generally preferred.
- Python coding: Python is the most common coding language that is used in data science along with Java, Perl, C/C++.
- Hadoop platform: Although not always a requirement, knowing the Hadoop platform is still preferred for the field. Having a bit of experience in Hive or Pig is also a huge selling point.
- SQL database/coding: Though NoSQL and Hadoop have become a major part of the Data Science background, it is still preferred if we can write and execute complex queries in SQL.
- Working with unstructured data: It is most important that a Data Scientist is able to work with unstructured data be it on social media, video feeds, or audio.

To become a Big Data professional:

- Analytical skills: The ability to be able to make sense of the piles of data that we get. With analytical abilities, we will be able to determine which data is relevant to our solution, more like problem-solving.
- Creativity: We need to have the ability to create new methods to gather, interpret, and analyze a data strategy. This is an extremely suitable skill to possess.
- Mathematics and statistical skills: Good, old-fashioned “number crunching”. This is extremely necessary, be it in data science, data analytics, or big data.
- Computer science: Computers are the workhorses behind every data strategy. Programmers will have a constant need to come up with algorithms to process data into insights.
- Business skills: Big Data professionals will need to have an understanding of the business objectives that are in place, as well as the underlying processes that drive the growth of the business as well as its profit.

To become a Data Analyst:

- Programming skills: Knowing programming languages are R and Python are extremely important for any data analyst.
- Statistical skills and mathematics: Descriptive and inferential statistics and experimental designs are a must for data scientists.
- Machine learning skills.
- Data wrangling skills: The ability to map raw data and convert it into another format that allows for a more convenient consumption of the data.
- Communication and Data Visualization skills.
- Data Intuition: it is extremely important for professional to be able to think like a data analyst.

Salaries

Though in the same domain, each of these professionals, data scientists, big data specialists, and data analysts, earn varied salaries.

The average a data scientist earns today, according to Indeed.com is \$123,000 a year. According to Glassdoor, the average salary for a Data Scientist is \$113,436 per year.

The average salary of a Big Data specialist according to Glassdoor is \$62,066 per year.

The average salary for a data analyst according to Glassdoor is \$60,476 per year.

CHAPTER 3. DATA SCIENCE AND BUSINESS STRATEGIES

“Data science often points to the need for change—and change can be difficult.”

David Stodder
Director of Research for BI
TDWI

FOREWORD

Data science is a hot topic among business and IT leaders. Excitement about the potential benefits of data science is tempered, however, by anxiety about how hard it is to find, hire, and train data science personnel, not to mention the difficulty of defining the term within the context of an organization’s goals and objectives.

There is no single definition of data science, nor one solution or technology. It is a term that joins together contributions from several fields, including statistics, mathematics, operations research, computer science, data mining, machine learning (algorithms that can learn from data), software programming, and data visualization. It can cover the entire process of acquiring and cleaning data, methods for exploring the data and extracting value from it, and techniques for making insights actionable for humans and automated processes. ¹ Most often, the focus of data science is to optimize decisions and realize higher value from data through advanced analysis.

One factor that makes data science distinct, however, is the word science. Data science is about applying scientific methods to explore and test hypotheses about the data. Indeed, many data scientists come from hard science fields such as chemistry and physics or professions such as neurobiology and nuclear physics. Data science pioneers have contributed mightily to the growth of social media and e-commerce; now, firms in other industries are keen to apply data science to their decision-making processes.

Continuous experimentation through examination of data to test hypotheses is at the heart of most data science projects. At the same time, the availability of technologies that can work with enormous data volumes and variety enables professionals to complement scientific methods with hypothesis-free approaches that employ machine learning to examine data and discover unforeseen patterns before articulating a hypothesis. This enables organizations to use data science to find previously hidden risks and opportunities and apply analytics to improve outcomes.

To solve business problems, develop new products and services, and optimize processes, organizations increasingly need analytics insights produced by data science teams with a diverse set of technical skills and business knowledge who are also good communicators. This TDWI Checklist Report describes seven steps to achieve a successful data science strategy.

Seven Steps for Executing a Successful Data Science Strategy

1. IDENTIFY OUR ORGANIZATION'S KEY BUSINESS DRIVERS FOR DATA SCIENCE

Data science may not be for everyone. Before embarking on a data science project, the first question to ask is a simple one: Do we need data science? Users may appear content with spreadsheets, business intelligence (BI) applications, and the selection of structured data available through data warehouses or other IT-managed repositories. Existing reports and dashboards may seem sufficient. From this perspective, investing in data science and technology to expand the reach of analytics into more data types, including semi-structured and unstructured, may appear unjustified.

To evaluate whether it is worth engaging in data science, organizations need to look at the value it could bring beyond what it already realizes from traditional BI, analytics, and data warehousing. The place to begin such an evaluation is the potential business drivers: What business value could be gained by developing a data science strategy? What are the questions the organization needs to solve to be more competitive, effective, and proactive? How well does the organization understand—and know how to respond to—the interplay of factors that affect customer behavior, the success of its website, or the impact of key trends? Often, such analysis will reveal knowledge gaps the organization has been unable to fill with its current BI and data warehousing systems.

For this reason, one of the most important qualities to seek when selecting personnel for a data science effort is knowledge of and curiosity about the business. Data scientists often come into an enterprise possessing exceptional technical and scientific skills. However, it is critical that they also develop the business domain expertise to uncover questions the organization needs to ask using analytics, and how to make the resulting data insights actionable.

At this stage, organizations should identify where data science could contribute most to realizing business objectives. Some classic areas include achieving greater personalization and computational efficiency in marketing and advertising; monitoring social media; modeling attribution to determine what drives purchasing; establishing a dynamic pricing strategy across multiple channels; uncovering fraudulent activity; and autonomic analysis of important documents or images such as call center logs or checks.

2. CREATE AN EFFECTIVE TEAM FOR ACHIEVING DATA SCIENCE

GOALS

“It’s like chasing unicorns” is a phrase often used to describe the difficult task of finding and keeping those rare individuals with the experience and ability to perform all that is required of a data scientist. In this exclusive group, many have a Ph.D. and a good number come from diverse, non-computer-science backgrounds.

The pioneers of data science— “half hackers and half scientists,” as one person put it— often took a do-it-ourselves approach through hands-on implementation of Hadoop and other open source technologies to store, access, and analyze massive and varied sources of big data. Although firms have benefited from their innovation, the artisan approach has left them vulnerable if and when their data scientists are lured away by competitors.

Rather than focus on finding one or a few individuals who seem to be able to do it all, a wiser course is to develop a stable team that brings together the talents of multiple experts. As discussed in the previous step, the team’s members must understand business drivers and not lose sight of the goal of delivering actionable business value. Each member of the team should also have enthusiasm, curiosity, and creative energy for working with business leadership on data and analytics projects.

Depending on the project, the team will need personnel with a combination of skills that include expertise in the business domain (for example, customer engagement or marketing), business analytics, statistics, data mining, machine learning, data and information retrieval, programming, prototyping, and visualization. Organizations should assemble a team that includes individuals with communication skills, not just technical acumen.

Although it is valuable to look externally for data scientists and leadership such as chief data officers, taking a team approach allows organizations to look internally. Many organizations already have personnel who could join a data science team. Indeed, TDWI Research finds that the majority of organizations plan to train internal personnel to handle data science projects. Personnel could include business analysts, statisticians, software developers, data analysts, and other data professionals.

In this step, organizations should bring business and IT leadership together to develop a strategy for creating effective and sustainable data science teams. Their plan should include training and incentives to attract internal personnel.

3. EMPHASIZE COMMUNICATION SKILLS TO REALIZE DATA

SCIENCE'S VALUE

Organizations that use data science successfully almost universally point to communication as a key ingredient to their success. Insights provided by analytics are of little value unless the data science team articulates what the findings say and why they are significant to business goals. Often this is not easy, especially if the presentation of the findings calls into question executives' "gut feel" assumptions about business strategy, strays from tightly controlled modes of BI reporting and analysis, or suggests that established processes are ineffective or outdated. Data science often points to the need for change—and change can be difficult.

Communication is also vital to improving collaboration in a data science project. Often, along with data scientists, key players (such as statisticians, business analysts, data analysts, and developers) are scattered in silos across the organization, or business and data analysts may work in a separate department than the business stakeholders, who should also be part of the data science effort. Important new perspectives can come if data science teams are able to work across divisions or silos to gain a more global view.

For example, to identify which actions are most influential to the buying behavior of an important cluster of customers, it is valuable if data science teams can examine data from a number of sources that might be managed in different divisional silos such as e-commerce, brick-and-mortar stores, contact centers, and field service offices. This "big data" has never fit easily into a data warehouse, much less a spreadsheet. The data science team could make a great contribution just by pulling together a global, holistic view of this scattered data.

Working across the organization is also important when the goal of data science is to optimize a process by developing algorithms that will automate decisions. Communication is essential; the team must be aware of how optimization will impact dependent processes, including how data is collected and analyzed. Without good communication, optimization could have unintended consequences.

Communication by and among data science teams is essential to building a data-driven analytics culture. In this step, organizations should emphasize the value of communication and make it a priority as they evaluate candidates for data science teams.

4. EXPAND THE IMPACT OF DATA SCIENCE THROUGH VISUALIZATION AND STORYTELLING

Data science fits into a larger objective of creating a data- driven “analytics culture” that is energized by a shared desire to improve decision making at all levels, from executives to frontline personnel. The key goal is to supplant uninformed, emotional decision making based on inaccurate theories with decision processes that are supported by empirical evidence, testing of hypotheses, and impactful data analysis. Although inspiration will always be vital, companies with healthy analytics cultures accept the notion that assumptions should be questioned by looking closely at the data.

Data science thrives in an analytics culture. However, not all personnel in an organization are going to be part of data science teams, nor should they be. To bring more users into the analytics culture, organizations should explore technologies that can support the “democratization” of BI, analytics, and data discovery. These products are increasingly able to address users’ self-service demands for data access and interaction without IT hand-holding. The tools go beyond simple spreadsheets and canned reporting to deliver different perspectives on metrics, help users uncover trends, and enable them to personalize dashboards.

Data visualization is an essential technology for data science and most self-service BI, analytics, and data discovery use cases. Across organizations, users’ visualization requirements can be diverse; some need simple interfaces that emphasize how to respond to a situation while others demand more varied types of visualizations. Leading tools have libraries of visualization types, and more are available through open source libraries. Organizations should take advantage of maturing data visualization technologies for both advanced data science and data interaction by nontechnical users.

Visualization enables “data storytelling.” This hot trend fuses visualization, data analysis, and usually verbal or written discussion, often in an infographic, to provide interpretation of data science results and why they are significant. Storytelling can be an effective way for data science teams to communicate accurately what they have found rather than just present numbers that could be misinterpreted. Organizations should encourage data storytelling and provide training so data science teams and other users can do it well.

5. GIVE DATA SCIENCE TEAMS ACCESS TO ALL THE DATA

Data is the raw material of data science. Like chefs looking for new taste sensations, data scientists need to work closely with data at every step so they know what they have and can extract fresh insights to deliver business value. Although valuable for reporting and proscribed forms of analysis, most traditional BI and data warehousing systems offer users only selected data samples, subsets, and pre-aggregated reports that have been carefully scrubbed and manicured by data professionals. Instead of raw data, most BI users work with reports or dashboards. What they leave behind are unincorporated structured sources and a vast universe of semi- and unstructured data and content that has never easily fit into BI systems and data warehouses.

Structured data can, of course, be voluminous and varied, especially when brought in from diverse applications. However, data science is often more closely associated with the desire to analyze semi- and unstructured data because these sources are growing rapidly and have been analyzed little, if at all. Preparing this breadth of data, assessing its quality, looking for gaps and errors, and performing exploratory analysis to determine relevant extracts are essential data science activities. They can take up the lion's share of a data science team's time. Although tools can automate steps, data science teams need to get close to the data to properly move forward with analytics and algorithm development.

Computer logs, social media, sensor data, and other new sources can be messy and chaotic; organizations should be realistic about the effort it will take to investigate and prepare the data. Organizations should ensure that data science teams include personnel who are comfortable working with raw data. In most cases, the team will need personnel who are knowledgeable about Hadoop and related technologies and are familiar with data lake and data hub concepts for gathering, storing, and accessing raw data.

Data science teams should always be on the lookout for interesting and potentially relevant data sources. Often, more than one application will be recording diverse (or sometimes the same or similar) data about customers, transactions, or other objects. Data scientists can play a valuable role by uncovering discrepancies and data quality problems.

6. PREPARE DATA SCIENCE PROCESSES FOR

OPERATIONALIZING ANALYTICS

Businesses can execute at a higher level if they can strengthen the connection between analytics and business processes. The first step is to move beyond purely “descriptive” analytics, which only answers what and why questions about historical trends and events, to predictive analytics, which can help discern what is likely to happen next. By streamlining how they develop and deploy predictive models, organizations can expand their use into more operational processes.

However, getting business value from this expansion requires more than just producing more analytic models faster. Firms must move to the next stage: to “prescriptive” analytics, which is about producing not just predictive insights but also suggested actions. Prescriptive analytics can be useful to both humans responsible for business processes and for guiding emerging automated decision systems.

Potential use cases abound. The most common is to improve customer marketing to offer targeted cross-sell, up-sell, and next- best-action offers at the moment of engagement. Another example occurs in complex, high-volume supply chains. Leading firms today apply predictive modeling to forecast what might happen given the probability of factors that could affect product manufacturing, packaging, and shipping. To get maximum value from their analysis, these firms are moving toward prescriptive analytics to develop recommended options for automated rules and complex event processing systems. This evolution could also be important for organizations seeking to operationalize analytics to fight fraud, assess risks, position mobile assets, and more, in real time.

To operationalize analytics, data science teams must focus on reducing the time it takes to develop and deploy analytic models. With cleaner workflows and processes, data science teams can move away from uncoordinated, artisanal model development and toward practices that include quality feedback sessions to correct flaws. Along with process improvements, organizations can take advantage of new technology practices such as in-database scoring, which can help eliminate time-consuming data movement to specialized data stores, improve the performance of analytic models, and make models available for multiple applications as stored procedures.

Teams must continue to improve communication with business stakeholders. Delays in model development and deployment are often due as much to communication difficulties as they are to other factors.

7. IMPROVE GOVERNANCE TO AVOID DATA SCIENCE

“CREEPINESS”

Data science teams must keep in mind that the outside world contains another set of stakeholders: the general public, including current and prospective consumers of the firm’s products and services. Fear and concern are at a high level with the continued unfolding of news about data thefts, hacking, surveillance, online and geolocation tracking, and marketing retargeting. Leading retail firms have had their reputations sullied by security breaches. Commentators rail about the “creepiness” factor: that is, the extent of knowledge firms are amassing about customers’ purchasing and other observed behavior that through powerful, real-time analytics can be (and often is) turned into highly personalized marketing. “Creepiness” is the label given to what some call the “dark side” of data science.

Data science teams, along with business leadership, must be cognizant of the right balance between what they can achieve through advanced analysis of consumer data and what is tolerable—and ethical—from the public’s perspective. Often there is no single standard; companies report that younger “millennial” demographics groups are more tolerant of personalized targeting than are older groups. Some consumers appreciate having the flow of advertising and marketing be more relevant to their buying patterns and shopping interests, while others are surprised and upset by it. Some will voice their concerns through social media, proving the observation that marketing is always a conversation, not one-way communication.

Enterprises should ensure that ethics and consumer tolerance are part of data science planning discussions, along with adherence to standard data governance policies. Data science teams must make sure they are not cloistered from the outside world and that they hear about how consumers and the public in general are responding to actions taken based on their data insights. The teams should consult with business leaders to gain their feedback about how certain programs could affect the conversation between the company and the public—and consider the possible ramifications on the company’s reputation.

Governance policies should address how to protect sensitive data during data science processes, particularly personally identifiable information. Anonymizing data may not be sufficient. Organizations should examine how they can protect data used in algorithms so that consumers’ behavior patterns cannot be hacked by those looking to identify specific people.

CHAPTER 4. SPECIFIC TOOLS AND SOFTWARES

What Tools Do Employers Want Data Scientists to Know?

We find ourselves overwhelmed with all the various tools mentioned in data science forums. We were hoping MATLAB would be enough, but we are inundated with advice of how to get started. We get advice to learn and practice R. Then we should learn Python. Then we should learn SQL. Then we should learn SPSS. Then we should learn Excel. Then we should learn Rapid Miner. Then we should learn Open Refine. Then we should learn Python. Then we should learn D3.js. Then we should learn Hadoop. Then we should learn Data Wrangler. Then we should learn Pandas. Then we should... It just doesn't stop.

We really want to get a data science job and play with data all day long. We absolutely love math, but are not sure academia is right for us for several reasons. So we've read some articles and gone through some papers and suddenly find ourselves ridiculously excited about machine learning and about how to apply it in a business setting. Now data science is seen as more than just a way out, it's that we find the whole process interesting as well. So we want to learn the tools that will make us hireable as a data scientist, but aren't quite sure where to begin.

Since our goal is to get hired as a data scientist, one concrete way to understand what hiring managers are looking for is to ask them. Since we are new, this can be a bit tougher than it looks. So the second best way is to look at data science job advertisements to see what tools are listed.

Since the field (data science) is so big right now that what tools different companies and groups use will vary significantly. Some data scientists mostly build data cleaning services. Some data scientists do academic-style research. Some data scientists do a mix of all of the above to varying degrees. Before drilling down into all the various types of data science roles that exist and the specific tools that they use, we'll do a brief survey in order to get a sense of all the possible tools that are mentioned in relation to working as a "data scientist" in the industry.

When in doubt, look at the data. To better understand what data scientists get hired to do, here's what we're going to do. We're going to look at CareerBuilder (a career website) and look at the first 2 pages of search results for the keyword "data scientist". This will cover 50 job postings. For each listing, we'll go into it and figure out what tools the data scientist job listing mentions. Then we'll put together a list of tasks that appeared. We will then sort the results by the number of times that specific tool or technology was mentioned. Note, the results may vary when we are reading this, as this search is being done today (December 17, 2014).

Here are the tools and number of times they showed up:

- | | | |
|----------------|-----------------------|---------------------|
| 1) R x 30 | 14) Tableau x 8 | 27) Cascading x 1 |
| 2) SQL x 27 | 15) Excel x 6 | 28) Cassandra x 1 |
| 3) Python x 22 | 16) NoSQL x 5 | 29) Clojure x 1 |
| 4) Hadoop x 19 | 17) AWS x 4 | 30) Fortran x 1 |
| 5) SAS x 18 | 18) C x 4 | 31) JavaScript x 1 |
| 6) Java x 15 | 19) HBase x 4 | 32) JMP x 1 |
| 7) Hive x 13 | 20) Bash x 3 | 33) Mahout x 1 |
| 8) Matlab x 12 | 21) Spark x 3 | 34) objective-C x 1 |
| 9) Pig x 11 | 22) ElasticSearch x 2 | 35) QlickView x 1 |
| 10) C++ x 9 | 23) PHP x 2 | 36) Redis x 1 |
| 11) Ruby x 9 | 24) Scala x 2 | 37) Redshift x 1 |
| 12) SPSS x 9 | 25) Shark x 2 | 38) sed x 1 |
| 13) Perl x 8 | 26) Awk x 1 | |

As we can see, a data science job descriptions ask data scientists to know 30 tools. All the way from data technologies, to scripting languages, to statistical programming languages. And this was just in 50 job postings (2 pages of CareerBuilder results). Some tools are very similar and others are very specific to certain domains different. This is one of the fortunate or unfortunate things about the data science field at the moment, that it is so big right now that what matters and what we'd actually differs drastically from job to job.

The silver lining behind this list is that most job postings have the following phrase: "know at least one of the following...". Which means that we don't actually have to go out and learn all of the tools. It just means that we should know at least one of them really well and have a passing familiarity with some of the others ones. We don't need to know them intimately, we just need to know what they do.

So, if we are looking for a data science job, based on this data, the best way to get started is to learn R, SQL, and Hadoop. Then have a passing understanding of Python and the tools that work with Hadoop like Hive, Pig, and others. This will make it so that we know at least one of the tools that data science positions are looking for and we'll have a good start to becoming a data scientist.

Top Tools for Data Scientists: Analytics Tools, Data Visualization Tools, Database Tools, and More

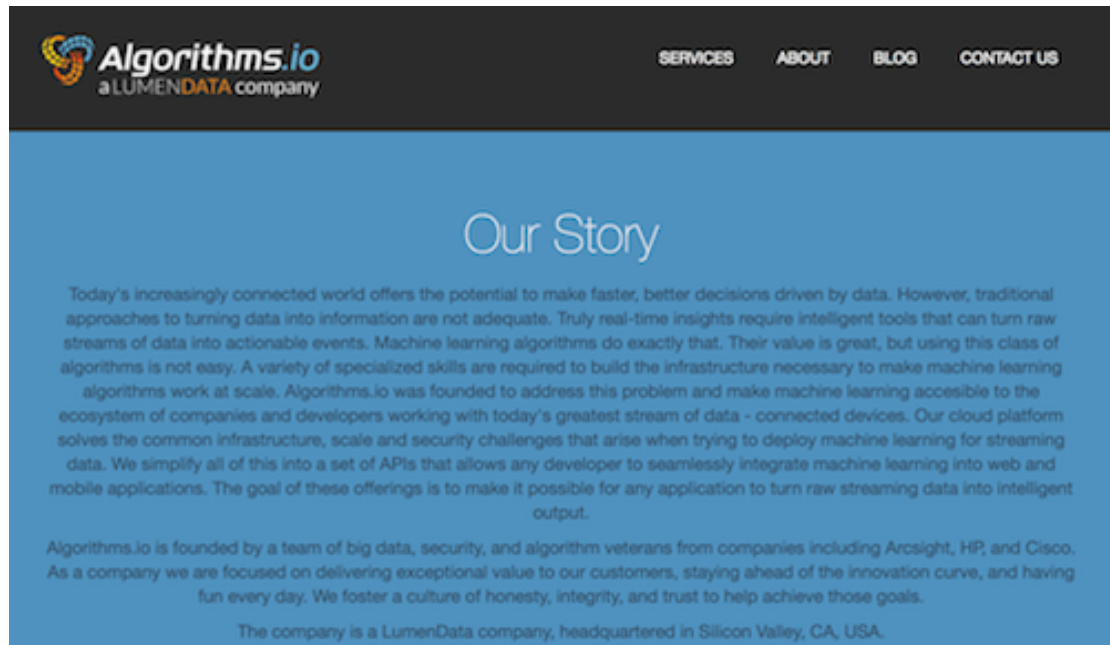
Data scientists are inquisitive and often seek out new tools that help them find answers. They also need to be proficient in using the tools of the trade, even though there are dozens upon dozens of them. Overall, data scientists should have a working knowledge of statistical programming languages for constructing data processing systems, databases, and visualization tools. Many in the field also deem a knowledge of programming an integral part of data science; however, not all data scientist students study programming, so it is helpful to be aware of tools that circumvent programming and include a user-friendly graphical interface so that data scientists' knowledge of algorithms is enough to help them build predictive models.

With everything on a data scientist's plate, we don't have time to search for the tools of the trade that can help we do our work. That's why we have rounded up tools that aid in data visualization, algorithms, statistical programming languages, and databases. We have chosen tools based on their ease of use, popularity, reputation, and features. And, we have listed our top tools for data scientists in alphabetical order to simplify our search; thus, they are not listed by any ranking or rating.

Here are the 50 best data science tools:

- | | | |
|------------------|-----------------------|----------------------|
| 1) Algorithms.io | 19) Excel | 36) Natural Language |
| 2) Apache Giraph | 20) Feature Labs | Toolkit (NLTK) |
| 3) Apache Hadoop | 21) ForecastThis | 37) NetworkX |
| 4) Apache HBase | 22) Fusion Tables | 38) NumPy |
| 5) Apache Hive | 23) Gawk | 39) Octave |
| 6) Apache Kafka | 24) ggplot2 | 40) OpenRefine |
| 7) Apache Mahout | 25) GraphLab Create | 41) Pandas |
| 8) Apache Mesos | 26) IPython | 42) RapidMiner |
| 9) Apache Pig | 27) Java | 43) Redis |
| 10) Apache Spark | 28) Jupyter | 44) RStudio |
| 11) Apache Storm | 29) KNIME Analytics | 45) Scala |
| 12) BigML | Platform | 46) scikit-learn |
| 13) Bokeh | 30) Logical Glue | 47) SciPy |
| 14) Cascading | 31) MATLAB | 48) Shiny |
| 15) Clojure | 32) Matplotlib | 49) TensorFlow |
| 16) D3.js | 33) MLBase | 50) TIBCO Spotfire |
| 17) DataRobot | 34) MySQL | |
| 18) DataRPM | 35) Narrative Science | |

1. Algorithms.



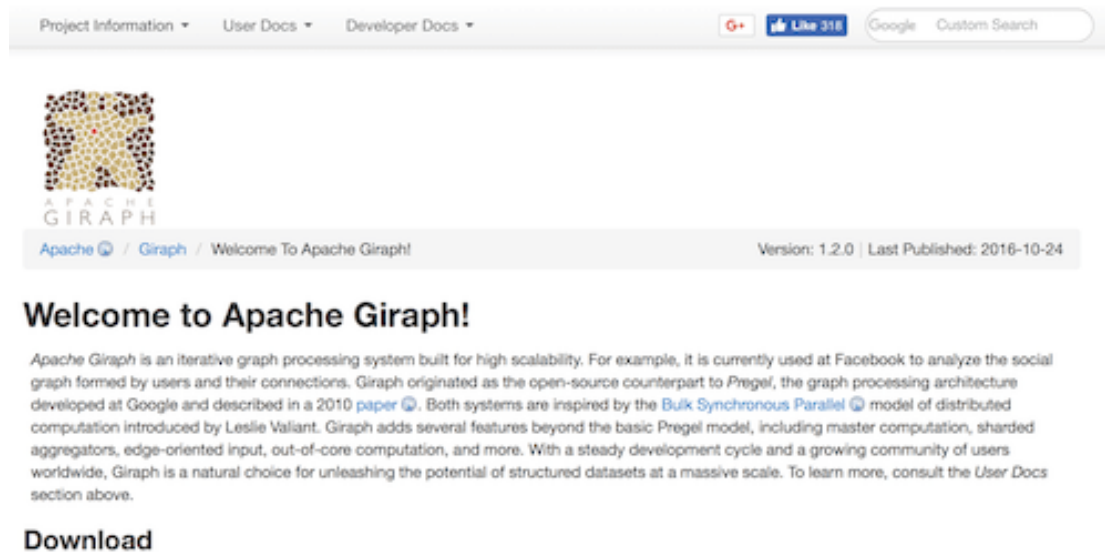
Algorithms.io is a LumenData Company providing machine learning as a service for streaming data from connected devices. This tool turns raw data into real-time insights and actionable events so that companies are in a better position to deploy machine learning for streaming data.

Key Features:

- Simplifies the process of making machine learning accessible to companies and developers working with connected devices
- Cloud platform addresses the common challenges with infrastructure, scale, and security that arise when deploying machine data
- Creates a set of APIs for developers to use to integrate machine learning into web and mobile apps so that any application can turn raw streaming data into intelligent output

Cost: Contact for a quote.

2. Apache Giraph



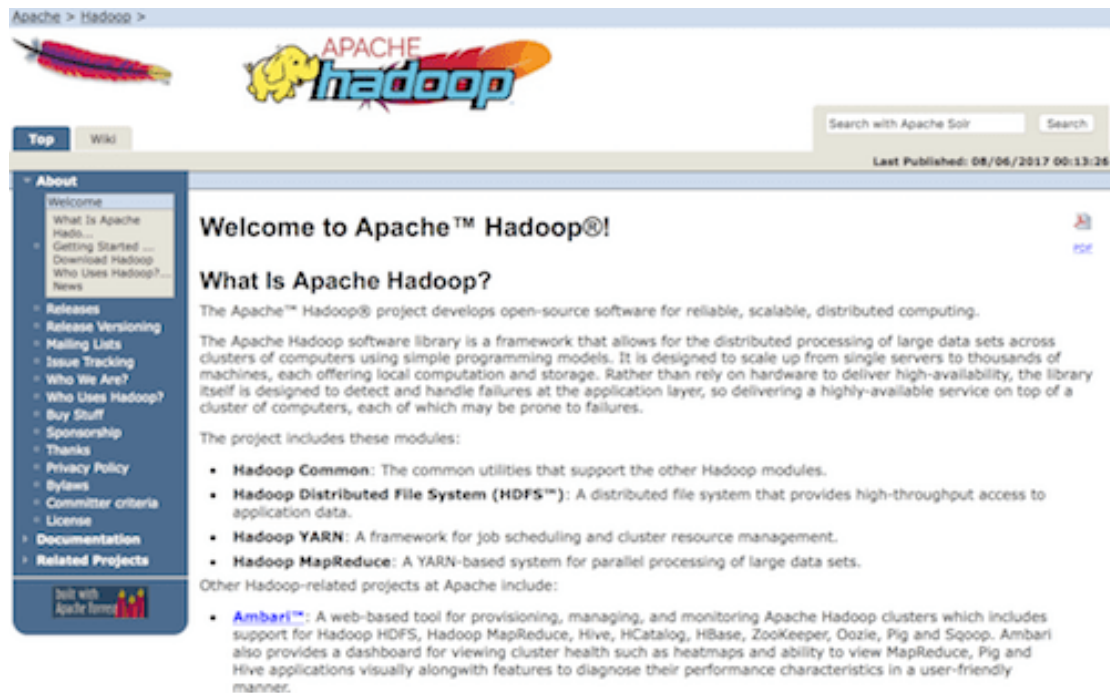
An iterative graph processing system designed for high scalability, Apache Giraph began as an open source counterpart to Pregel but adds multiple features beyond the basic Pregel model. Giraph is used by data scientists to “unleash the potential of structured datasets at a massive scale.”

Key Features:

- Inspired by the Bulk Synchronous Parallel model of distributed computation as introduced by Leslie Valiant
- Master computation
- Sharded aggregators
- Edge-oriented input
- Out-of-core computation
- Steady development cycle and growing community of users

Cost: FREE

3. Apache Hadoop



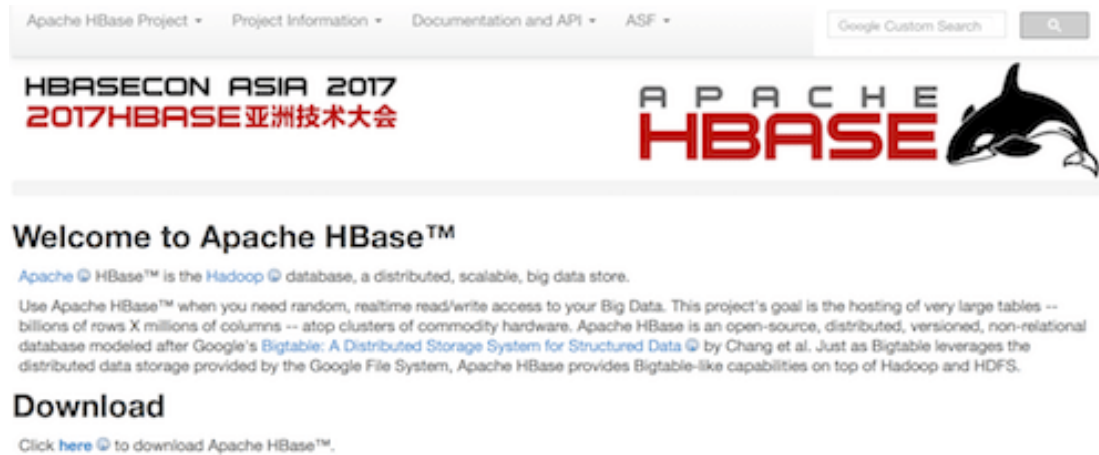
Apache Hadoop is an open source software for reliable, distributed, scalable computing. A framework allowing for the distributed processing of large datasets across clusters of computers, the software library uses simple programming models. Hadoop is appropriate for research and production.

Key Features:

- Designed to scale from single servers to thousands of machines
- The library detects and handles failures at the application layer instead of relying on hardware to deliver high-availability
- Includes the Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce modules

Cost: FREE

4. Apache HBase



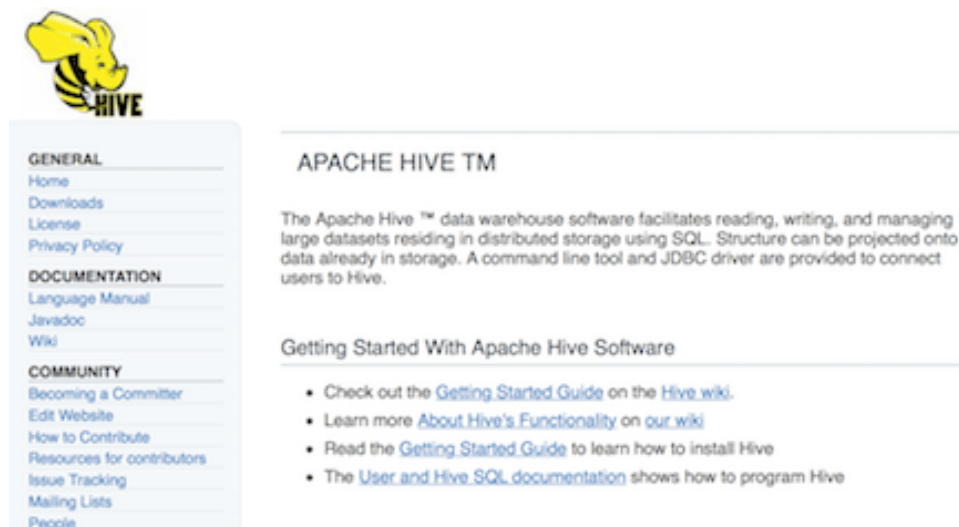
The Hadoop database, Apache HBase is a distributed, scalable, big data store. Data scientists use this open source tool when they need random, real-time read/write access to Big Data. Apache HBase also provides capabilities similar to Bigtable on top of Hadoop and HDFS.

Key Features:

- Open source, distributed, versioned, non-relational database modeled after Google's Bigtable: A Distributed Storage System for Structured Data
- Linear and modular scalability
- Strictly consistent reads and writes
- Automatic and configurable shading of tables

Cost: FREE

5. Apache Hive



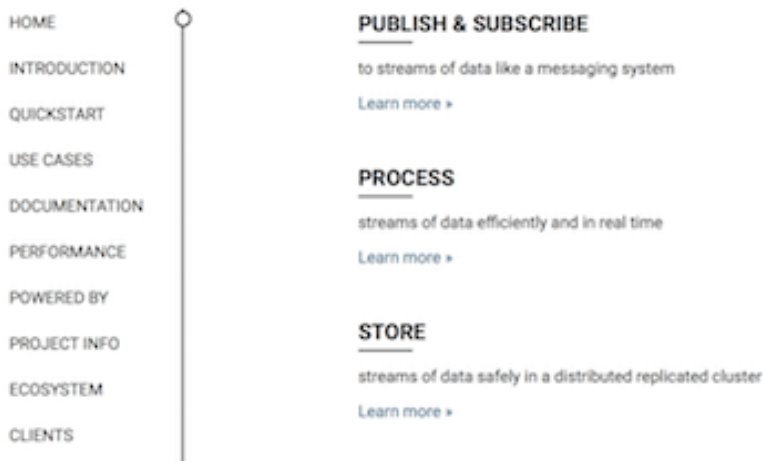
An Apache Software foundation Project, Apache Hive began as a subproject of Apache Hadoop and now is a top-level project itself. This tool is a data warehouse software that assists in reading, writing, and managing large datasets that reside in distributed storage using SQL.

Key Features:

- Project structure onto data already in storage
- Command line tool is provided to connect users to Hive
- JDBC driver is provided to connect users to Hive

Cost: FREE

6. Apache Kafka



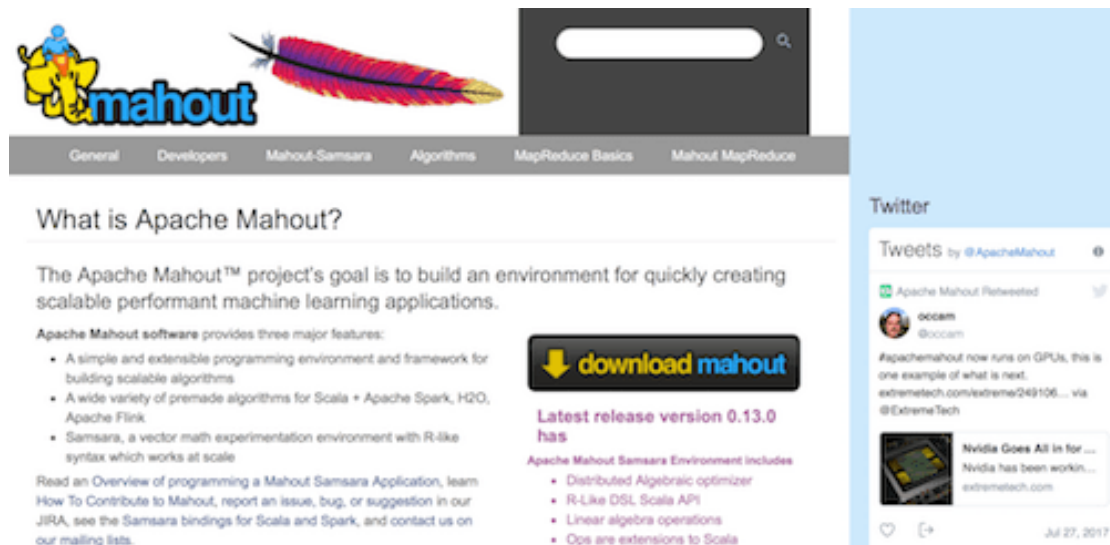
A distributed streaming platform, Apache Kafka efficiently processes streams of data in real time. Data scientists use this tool to build real-time data pipelines and streaming apps because it empowers us to publish and subscribe to streams of records, store streams of records in a fault-tolerant way, and process streams of records as they occur.

Key Features:

- Runs as a cluster on one or more servers
- Cluster stores streams of records in categories called topics
- Each record includes a key, value, and timestamp
- Has four core APIs: Producer API, Consumer API, Streams API, and Connector API

Cost: FREE

7. Apache Mahout



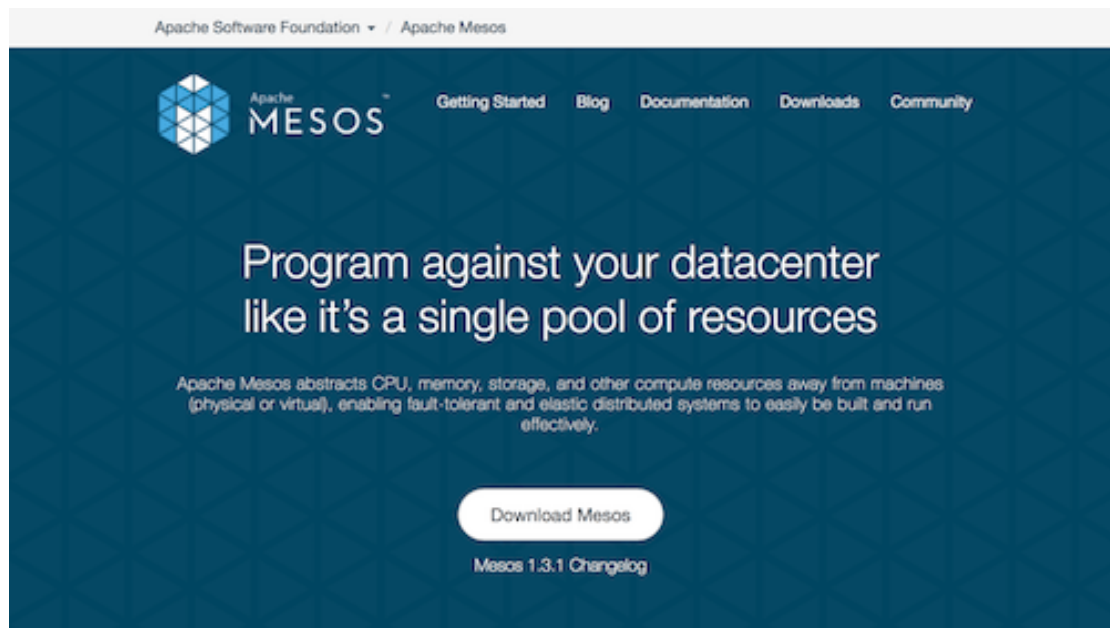
An open source Apache Foundation project for machine learning, Apache Mahout aims to enable scalable machine learning and data mining. Specifically, the project’s goal is to “build an environment for quickly creating scalable performant machine learning applications.”

Key Features:

- Simple, extensible programming environment and framework for building scalable algorithms
- Includes a wide variety of pre-made algorithms for Scala + Apache Spark, H2O, and Apache Flink
- Provides Samsara, a vector math experimentation environment with R-like syntax, which works at scale

Cost: FREE

8. Apache Mesos



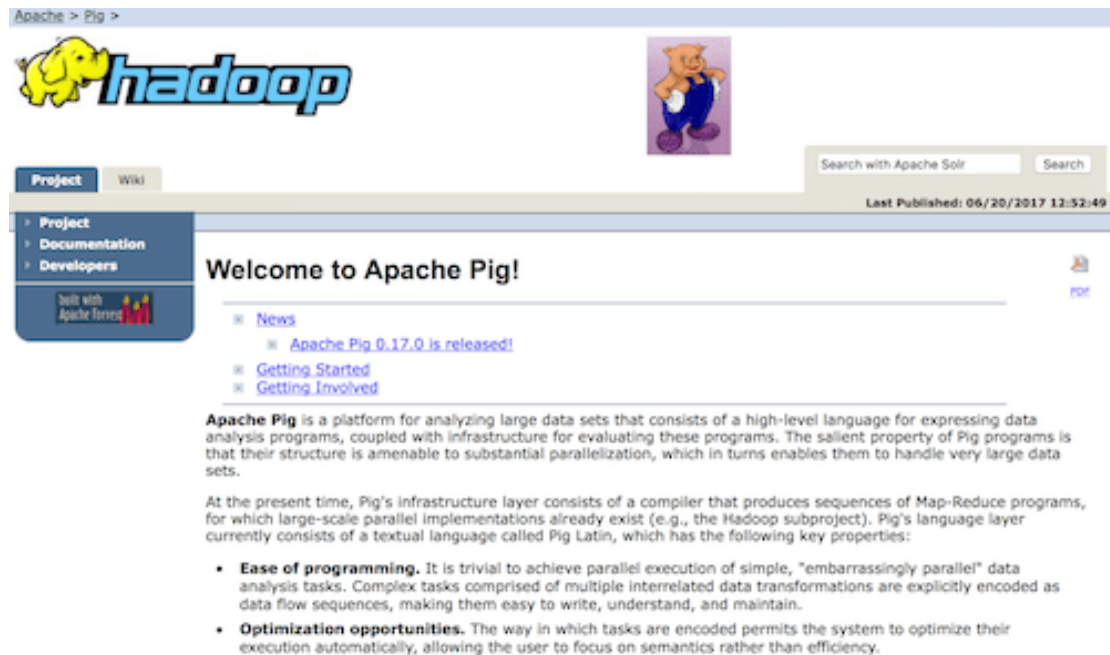
A cluster manager, Apache Mesos provides efficient resource isolation and sharing across distributed applications or frameworks. Mesos abstracts CPU, memory, storage, and other resources away from physical or virtual machines to enable fault-tolerant, elastic distributed systems to be built easily and run effectively.

Key Features:

- Built using principles similar to that of the Linux kernel but at a different level of abstraction
- Runs on every machine and provides applications like Hadoop and Spark with APIs for resource management and scheduling completely across datacenter and cloud environments
- Easily scales to 10,000s of nodes
- Non-disruptive upgrades for high availability
- Cross platform and cloud provider agnostic

Cost: FREE

9. Apache Pig



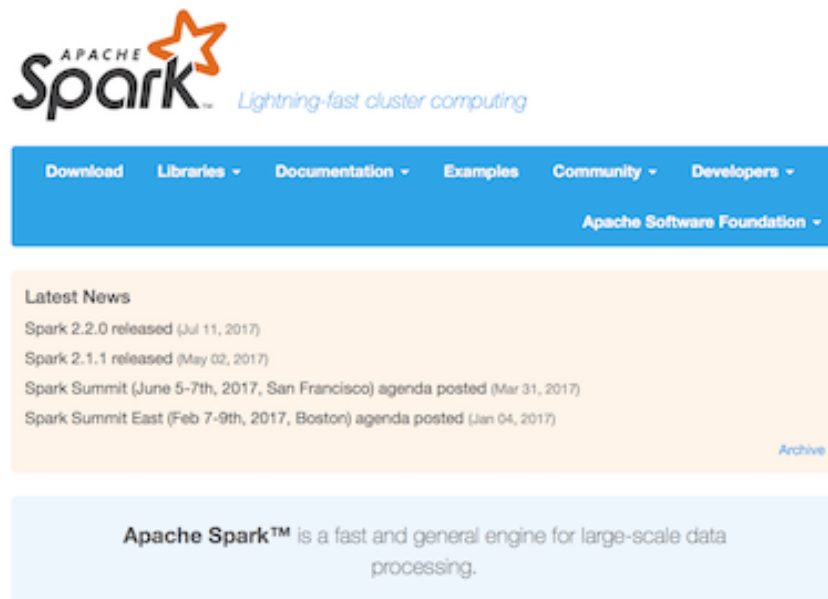
A platform designed for analyzing large datasets, Apache Pig consists of a high-level language for expressing data analysis programs that is coupled with infrastructure for evaluating such programs. Because Pig programs' structures can handle significant parallelization, they can tackle large datasets.

Key Features:

- Infrastructure consists of a compiler capable of producing sequences of Map-Reduce programs for which large-scale parallel implementations already exist
- Language layer includes a textual language called Pig Latin
- Key properties of Pig Latin include ease of programming, optimization opportunities, and extensibility

Cost: FREE

10. Apache Spark



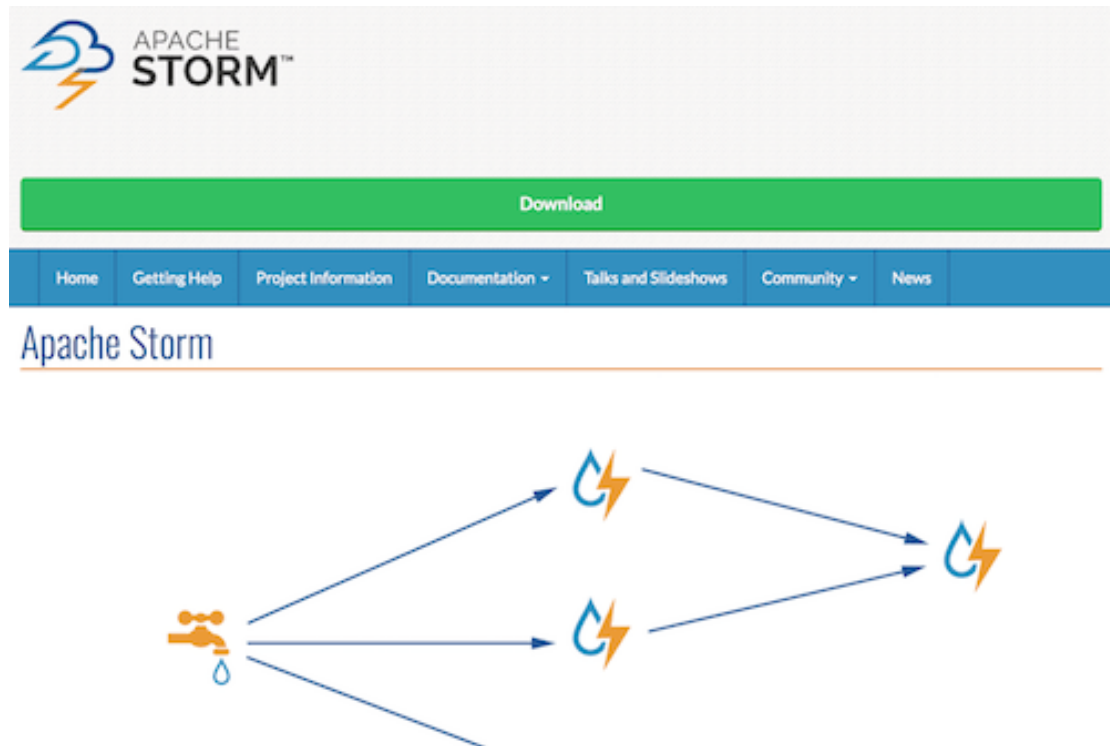
Apache Spark delivers “lightning-fast cluster computing.” A wide range of organizations use Spark to process large datasets, and this data scientist tool can access diverse data sources such as HDFS, Cassandra, HBase, and S3.

Key Features:

- Advanced DAG execution engine to support acyclic data flow and in-memory computing
- More than 80 high-level operators make it simple to build parallel apps
- Use interactively from the Scale, Python, and R shells
- Powers a stack of libraries including SQL, DataFrames, MLlib, GraphX, and Spark Streaming

Cost: FREE

11. Apache Storm



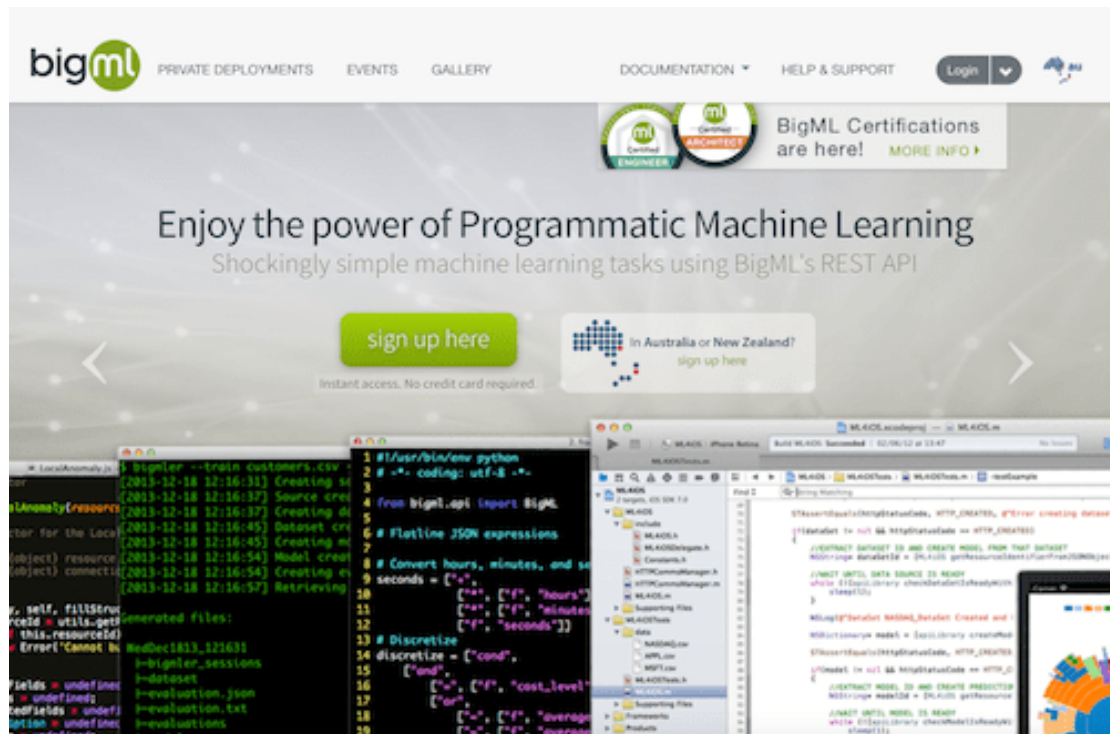
Apache Storm is a tool for data scientists that handles distributed and fault-tolerant real-time computation. It also tackles stream processing, continuous computation, distributed RPC, and more.

Key Features:

- Free and open source
- Reliably process unbounded data streams for real-time processing
- Use with any programming language
- Use cases include real-time analytics, online machine learning, continuous computation, distributed RPC, ETL, and more
- More than one million tuples processed per second per mode
- Integrates with our existing queueing and database technologies

Cost: FREE

12. BigML



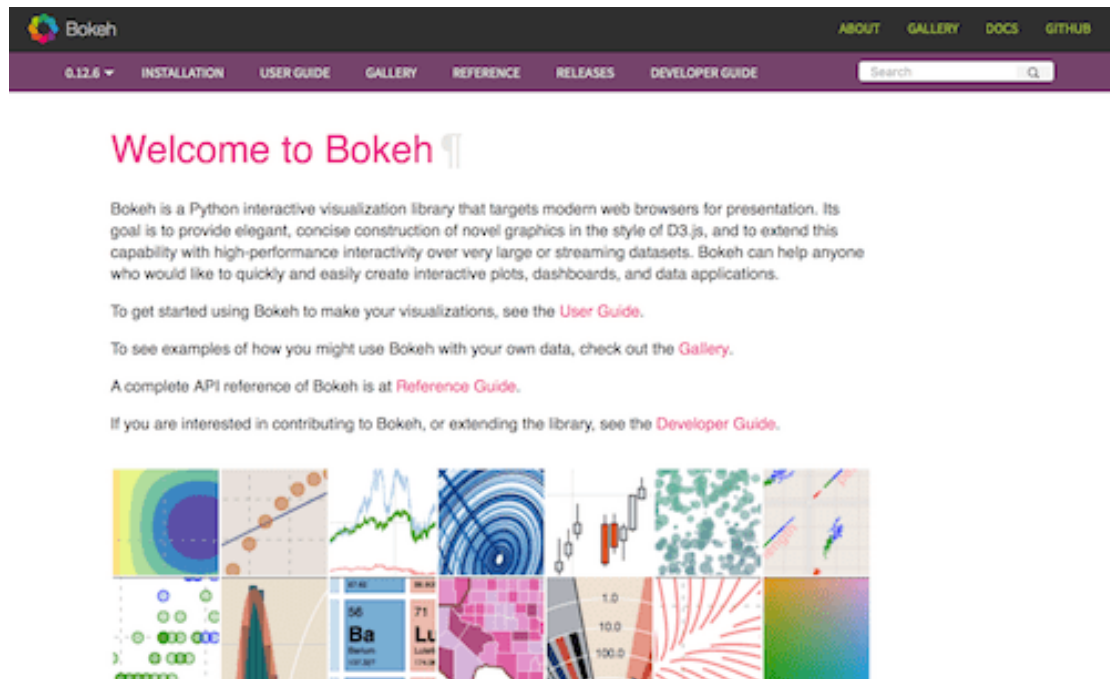
BigML makes machine learning simple. This company-wide platform runs in the cloud or on premises for operationalizing machine learning in organizations. BigML makes it simple to solve and automate classification, regression, cluster analysis, anomaly detection, association discovery, and topic modeling tasks.

Key Features:

- Build sophisticated machine learning-based solutions affordably
- Distill predictive patterns from data into practical, intelligent applications that anyone can use
- The platform, private deployments, and rich toolset help users create, rapidly experiment, fully automate, and manage machine learning workflows to power intelligent applications

Cost: Contact for a quote

13. Bokeh



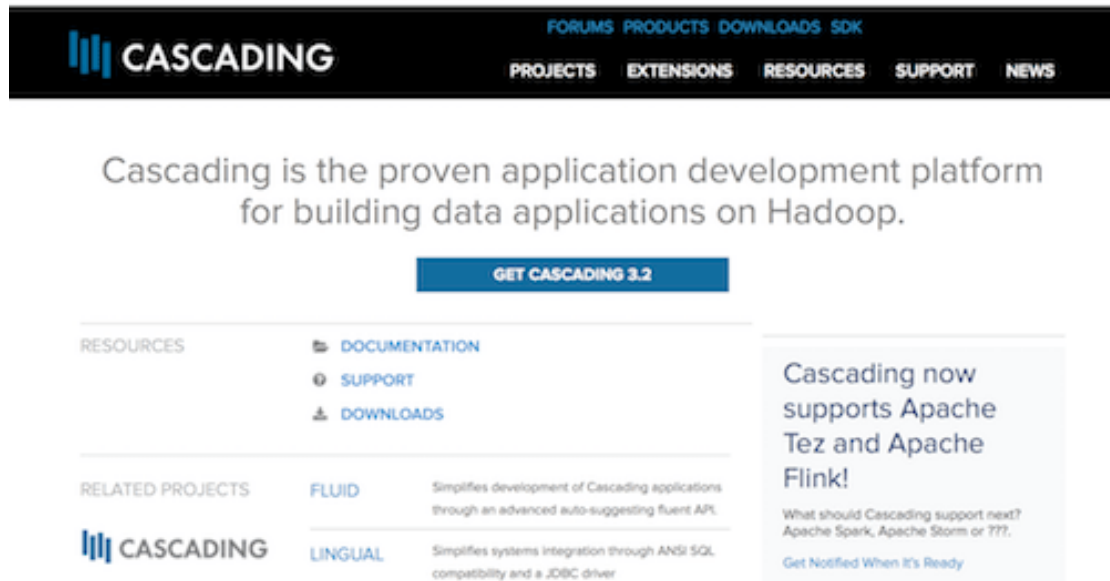
A Python interactive visualization library, Bokeh targets modern web browsers for presentation and helps users create interactive plots, dashboards, and data apps easily.

Key Features:

- Provides elegant and concise construction of graphics similar to D3.js
- Extends capabilities to high-performance interactivity over large or streaming datasets
- Quickly and easily create interactive plots, dashboards, and data applications

Cost: FREE

14. Cascading



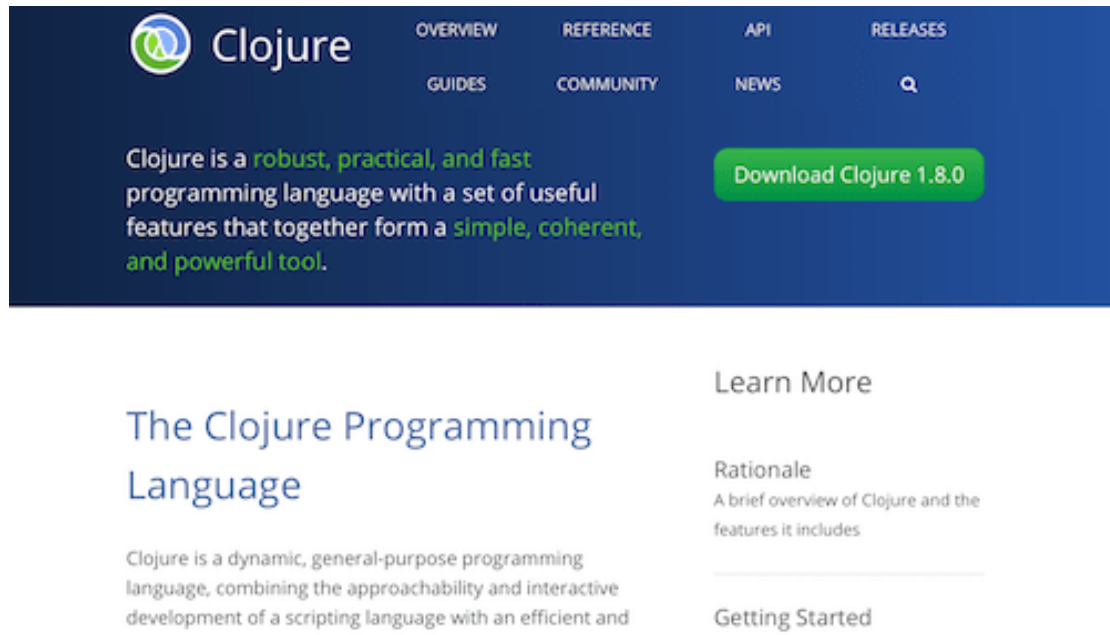
Cascading is an application development platform for data scientists building Big Data applications on Apache Hadoop. Users can solve simple and complex data problems with Cascading because it boasts computation engine, systems integration framework, data processing, and scheduling capabilities.

Key Features:

- Balances an ideal level of abstraction with appropriate degrees of freedom
- Offers Hadoop development teams portability
- Change a few lines of code and port Cascading to another supported compute fabric
- Runs on and may be ported between MapReduce, Apache Tez, and Apache Flink

Cost: FREE

15. Clojure



A robust and fast programming language, Clojure is a practical tool that marries the interactive development of a scripting language with an efficient infrastructure for multithreaded programming. Clojure is unique in that it is a compile language but remains dynamic with every feature supported at runtime.

Key Features:

- Rich set of immutable, persistent data structures
- Offers a software transactional memory system and reactive Agent system to ensure clean, correct, multithreaded designs when mutable state is necessary
- Provides easy access to Java frameworks with optional type hints and type inference
- Dynamic environment that users can interact with

Cost: FREE

16. D3.js



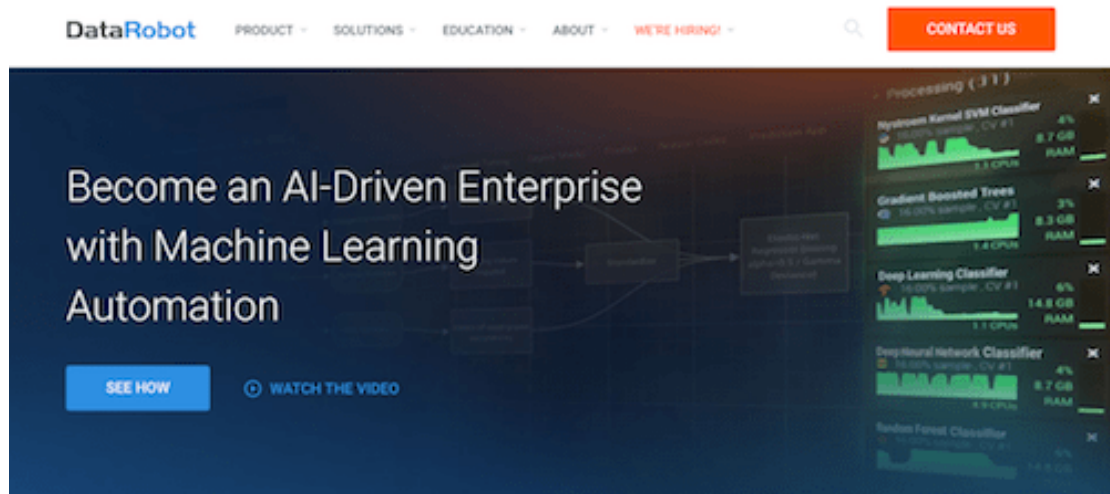
Committed to “code and data for humans,” Mike Bostock created D3.js. Data scientists use this tool, a JavaScript library for manipulating documents based on data, to add life to their data with SVG, Canvas, and HTML.

Key Features:

- Emphasis on web standards to gain full capabilities of modern browsers without being tied to a proprietary framework
- Combines powerful visualization components and a data-driven approach to Document Object Model (DOM) manipulation
- Bind arbitrary data to a DOM and then apply data-driven transformations to the document

Cost: FREE

17. DataRobot



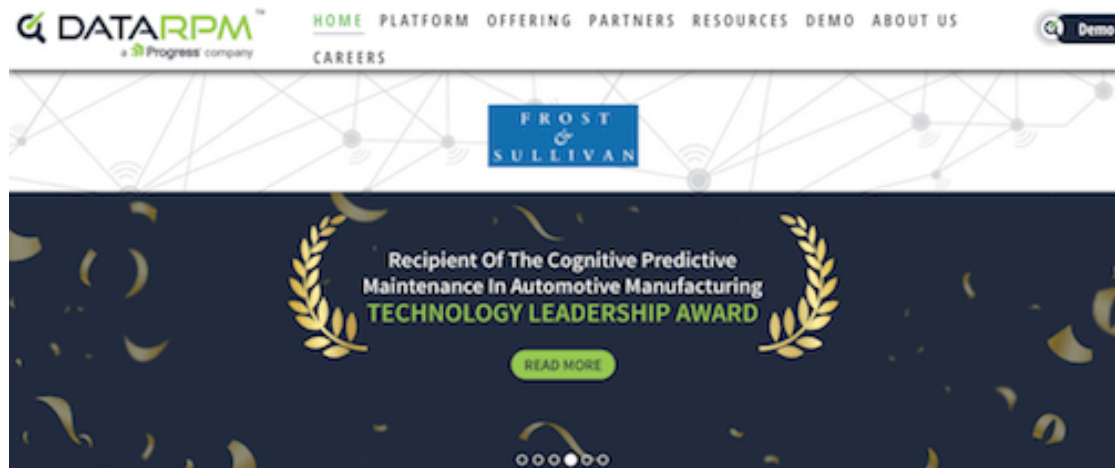
An advanced machine learning automation platform, DataRobot helps data scientists build better predictive models faster. We can keep up with the ever-expanding ecosystem of machine learning algorithms easily when we use DataRobot.

Key Features:

- Constantly expanding, vast set of diverse, best-in-class algorithms from leading sources
- Train, test, and compare hundreds of varying models with one line of code or a single click
- Automatically identifies top pre-processing and feature engineering for each modeling technique
- Uses hundreds and even thousands of servers as well as multiple cores within each server to parallelize data exploration, model building, and hyper-parameter tuning
- Easy model deployment

Cost: Contact for a quote

18. DataRPM



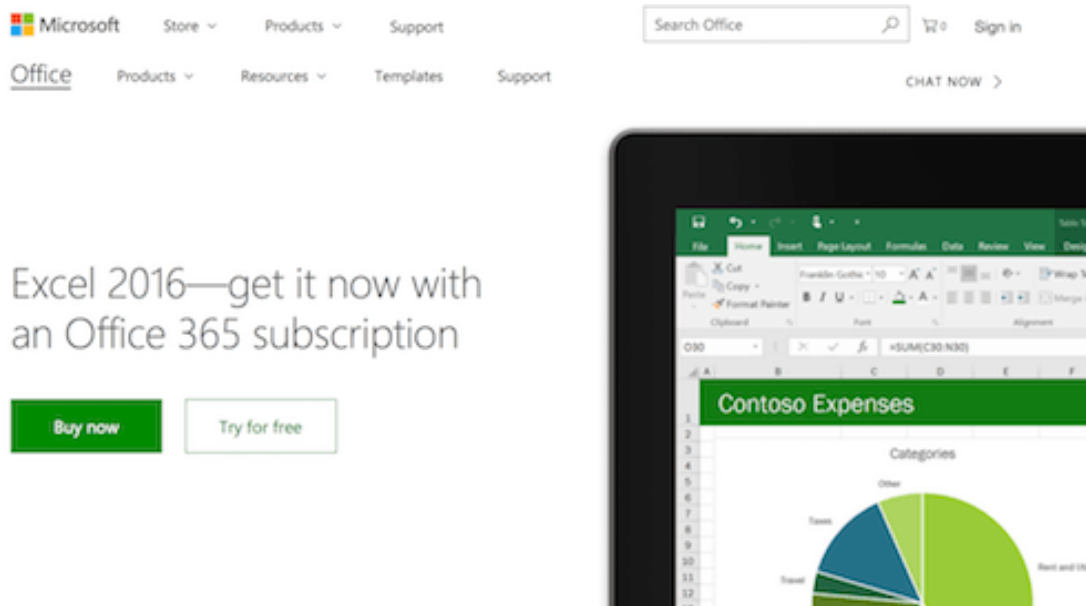
DataRPM is the “industry’s first and only cognitive predictive maintenance platform for industrial IoT. DataRPM also is the recipient of the 2017 Technology Leadership Award for Cognitive Predictive Maintenance in Automotive Manufacturing from Frost & Sullivan.

Key Features:

- Uses patent-pending meta-learning technology, an integral component of Artificial Intelligence, to automate predictions of asset failures
- Runs multiple live automated machine learning experiments on datasets
- Extracts data from every experiment, trains models on the metadata repository, applies models to predict the best algorithms, and builds machine-generated, human-verified machine learning models for predictive maintenance
- Workflow uses recipes such as feature engineering, segmentation, influencing factors, and prediction recipes to deliver prescriptive recommendations.

Cost: Contact for a quote

19. Excel



Many data scientists view Excel as a secret weapon. It is a familiar tool that scientists can rely on to quickly sort, filter, and work with their data. It's also on nearly every computer we come across, so data scientists can work from just about anywhere with Excel.

Key Features:

- Named ranges for creating a makeshift database
- Sorting and filtering with one click to quickly and easily explore our dataset
- Use Advanced Filtering to filter our dataset based on criteria we specify in a different range
- Use pivot tables to cross-tabulate data and calculate counts, sums, and other metrics
- Visual Basic provides a variety of creative solutions

Cost: FREE trial available

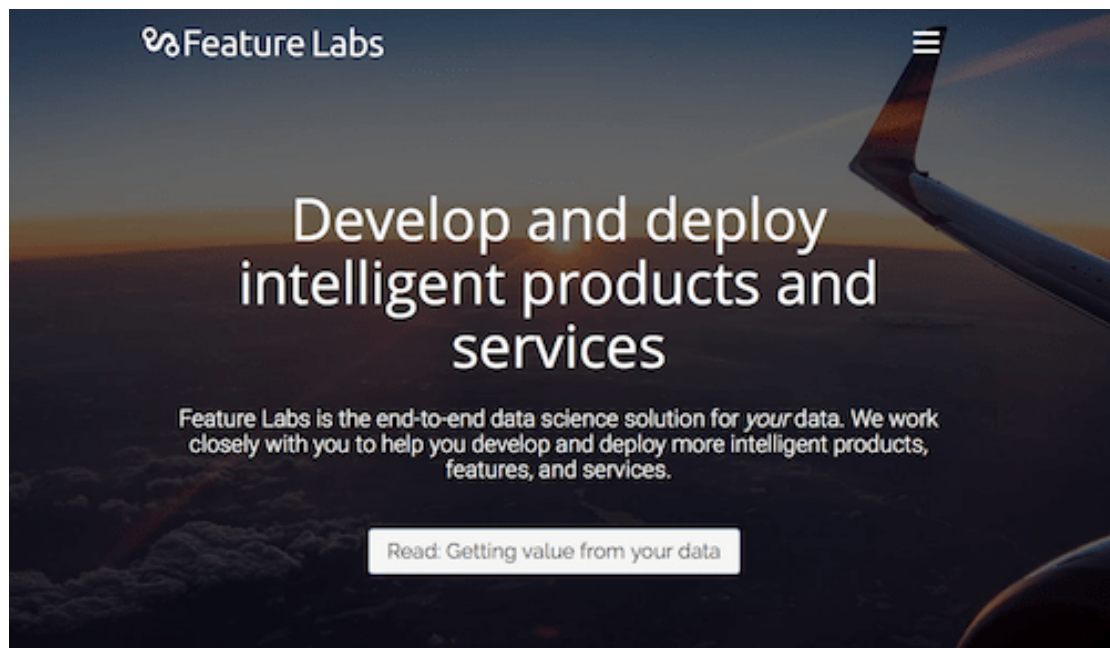
Home Buying Options

- ✧ Office 365 Home: \$99.99/year
- ✧ Office 365 Personal: \$69.99/year
- ✧ Office Home & Student 2016 for PC \$149.99 one-time purchase

Business Buying Options

- ✧ Office 365 Business: \$8.25/user/month with annual commitment
- ✧ Office 365 Business Premium: \$12.50/user/month with annual commitment
- ✧ Office 365 Business Essentials: \$5/user/month with annual commitment

20. Feature Labs



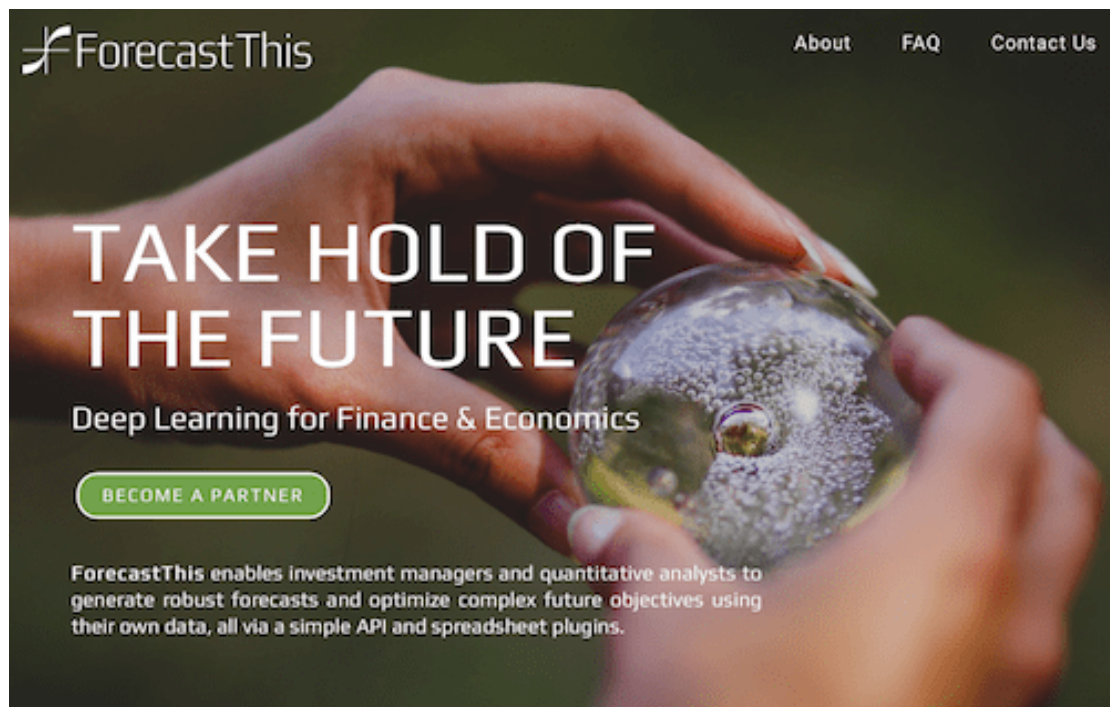
An end-to-end data science solution, Feature Labs develops and deploys intelligent products and services for our data. They also work with data scientists to help we develop and deploy intelligent products, features, and services.

Key Features:

- Integrates with our data to help scientists, developers, analysts, managers, and executives
- Discover new insights and gain a better understanding of how our data forecasts the future of our business
- On-boarding sessions tailored to our data and use cases to help we get off to an efficient start

Cost: Contact for a quote

21. ForecastThis



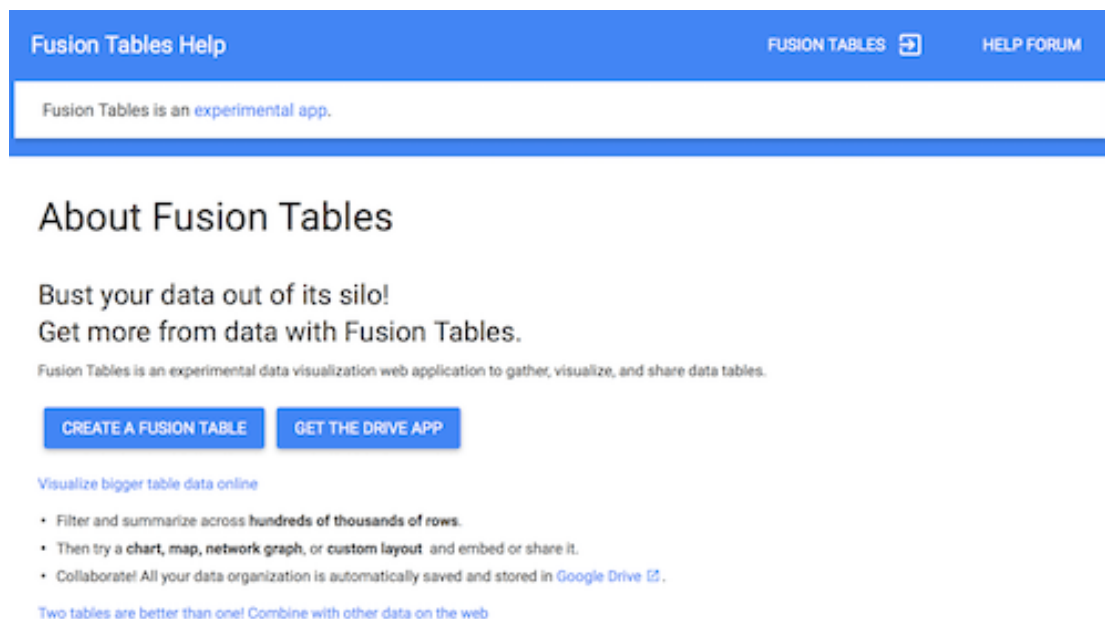
ForecastThis is a tool for data scientists that automates predictive model selection. The company strives to make deep learning relevant for finance and economics by enabling investment managers, quantitative analysts, and data scientists to use their own data to generate robust forecasts and optimize complex future objectives.

Key Features:

- Simple API and spreadsheet plugins
- Uniquely robust global optimization algorithms
- Scales to challenges of nearly any shape or size
- Algorithms create plausible, interpretable models of market processes to lend credibility to any output and help we get inside the market more successfully

Cost: Contact for a quote

22. Fusion Tables



Google Fusion Tables is a cloud-based data management service that focuses on collaboration, ease-of-use, and visualizations. An experimental app, Fusion Tables is a data visualization web application tool for data scientists that empowers we to gather, visualize, and share data tables.

Key Features:

- Visualize bigger table data online
- Combine with other data on the web
- Make a map in minutes
- Search thousands of public Fusion Tables or millions of public tables from the web that we can import to Fusion Tables
- Import our own data and visualize it instantly
- Publish our visualization on other web properties

Cost: FREE

23. Gawk

Gawk



If you are like many computer users, you would frequently like to make changes in various text files wherever certain patterns appear, or extract data from parts of certain lines while discarding the rest. To write a program to do this in a language such as C or Pascal is a time-consuming inconvenience that may take many lines of code. The job is easy with awk, especially the GNU implementation: gawk.

The awk utility interprets a special-purpose programming language that makes it possible to handle simple data-reformatting jobs with just a few lines of code.

The source code for the latest release of GNU awk is available from the GNU project's [ftp server](#) and its [many mirrors](#). The current development sources are available through the [gawk project on savannah](#).

The [main gawk manual](#) is available online.

A separate [gawknet manual](#) for the special TCP/IP networking features of GNU awk is also available.

Bug reports and feature suggestions for gawk should be sent to bug-gawk@gnu.org.



GNU is an operating system that enables we to use a computer without software “that would trample our freedom.” They have created Gawk, an awk utility that interprets a special-purpose programming language. Gawk empowers users to handle simple data-reformatting jobs using only a few lines of code.

Key Features:

- Search files for lines or other text units containing one or more patterns
- Data-driven rather than procedural
- Makes it easy to read and write programs

Cost: FREE

24. Ggplot2



ggplot2 is a plotting system for R, based on the grammar of graphics, which tries to take the good parts of base and lattice graphics and none of the bad parts. It takes care of many of the fiddly details that make plotting a hassle (like drawing legends) as well as providing a powerful model of graphics that makes it easy to produce complex multi-layered graphics.

Documentation

ggplot2 documentation is now available at docs.ggplot2.org.

Mailing list

You are welcome to ask ggplot2 questions on R-help, but if you'd like to participate in a more focussed mailing list, please sign up for the [ggplot2 mailing list](#):

Your email address:

You must be a member to post messages, but anyone can read the archived discussions.

Installation

Hadley Wickham and Winston Chang developed ggplot2, a plotting system for R that is based on the grammar of graphics. With ggplot2, data scientists can avoid many of the hassles of plotting while maintaining the attractive parts of base and lattice graphics and producing complex multi-layered graphics easily.

Key Features:

- Create new types of graphic tailored to our needs
- Create graphics to help we understand our data
- Produce elegant graphics for data analysis

Cost: FREE

25. GraphLab Create

The screenshot shows the PyPI package page for GraphLab-Create 2.1. At the top, there's a Python logo and a search bar. Below the logo, the breadcrumb trail reads "Package Index > GraphLab-Create > 2.1". On the left, a sidebar contains links to "PACKAGE INDEX" (with a sub-menu: Browse packages, List trove classifiers, RSS (latest 40 updates), RSS (newest 40 packages), Terms of Service, PyPI Tutorial, PyPI Security, PyPI Support, PyPI Bug Reports, PyPI Discussion, PyPI Developer Info), "ABOUT", "NEWS", "DOCUMENTATION", "DOWNLOAD", and "COMMUNITY". The main content area features the title "GraphLab-Create 2.1" with a green "Downloads" button. Below the title is a description: "GraphLab Create enables developers and data scientists to apply machine learning to build state of the art data products." This is followed by sections for "License" (pointing to LICENSE.txt), "References" (linking to https://turi.com), and "Contributors" (listing the Turi Team and https://turi.com). On the right, a "Not Logged In" box offers links for Login, Register, Lost Login?, Login with OpenID, and Login with Google, along with a "Status" section showing "Nothing to report".

Data scientists and developers use GraphLab Create to build state-of-the-art data products via machine learning. This machine learning modeling tool helps users build intelligent applications end-to-end in Python.

Key Features:

- Simplifies development of machine learning models
- Incorporates automatic feature engineering, model selection, and machine learning visualizations specific to the application
- Identify and link records within or across data sources corresponding to the same real-world entities

Cost:

FREE one-year renewable subscription for academic use

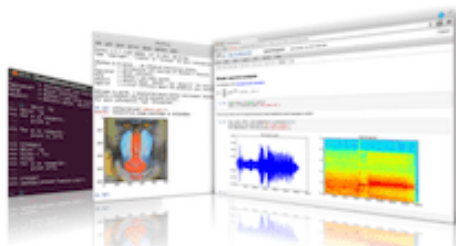
26. IPython

IP[y]: IPython Interactive Computing

[Install](#) · [Documentation](#) · [Project](#) · [Jupyter](#) · [News](#) · [Cite](#) · [Donate](#) · [Books](#)

IPython provides a rich architecture for interactive computing with:

- A powerful interactive shell.
- A kernel for [Jupyter](#).
- Support for interactive data visualization and use of [GUI toolkits](#).
- Flexible, [embeddable](#) interpreters to load into your own projects.
- Easy to use, high performance tools for [parallel computing](#).



Google Custom Search

JUPYTERCON



NOTEBOOK
VIEWER

Share your notebooks



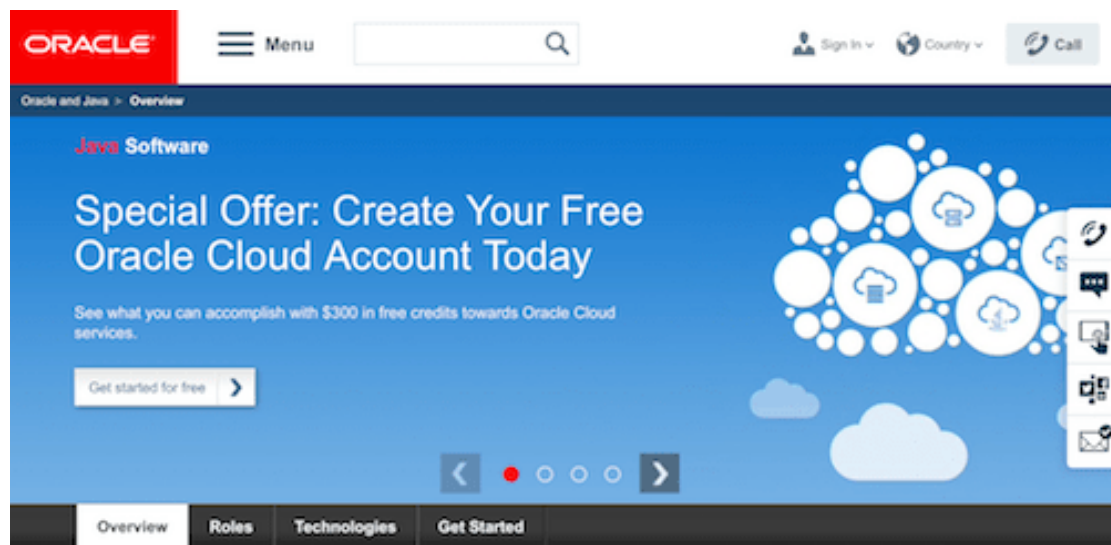
Interactive Python tools, or IPython, is a growing project with expanding language-agnostic components and provides a rich architecture for interactive computing. An open source tool for data scientists, IPython supports Python 2.7 and 3.3 or newer.

Key Features:

- A powerful interactive shell
- A kernel for Jupyter
- Support for interactive data visualization and use of GUI toolkits
- Load flexible, embeddable interpreters into our own projects
- Easy-to-use high performance parallel computing tools

Cost: FREE

27. Java



Java is a language with a broad user base that serves as a tool for data scientists creating products and frameworks involving distributed systems, data analysis, and machine learning. Java now is recognized as being just as important to data science as R and Python because it is robust, convenient, and scalable for data science applications.

Key Features:

- Easy to break down and understand
- Helps users be explicit about types of variables and data
- Well-developed suite of tools
- Develop and deploy applications on desktops and servers in addition to embedded environments
- Rich user interface, performance, versatility, portability, and security for modern applications

Cost:

FREE trial available;

Contact for commercial license cost

28. Jupyter



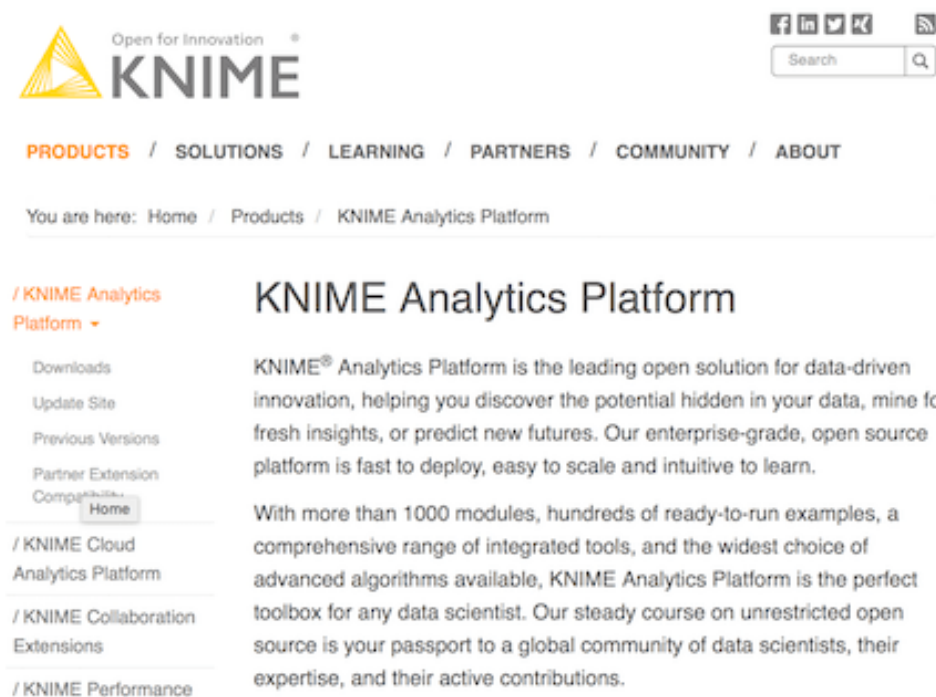
Jupyter provides multi-language interactive computing environments. Its Notebook, an open source web application, allows data scientists to create and share documents containing live code, equations, visualizations, and explanatory text.

Key Features:

- Uses include data cleaning and transformation, numerical simulation, statistical modeling, machine learning, and more
- Supports more than 40 programming languages including popular data science languages like Python, R, Julia, and Scala
- Share notebooks with others via email, Dropbox, GitHub, and the Jupyter Notebook Viewer
- Code can produce images, videos, LaTeX, and JavaScript
- Use interactive widgets to manipulate and visualize data in realtime

Cost: FREE

29. KNIME Analytics Platform



Thanks to its open platform, KNIME is a tool for navigating complex data freely. The KNIME Analytics Platform is a leading open solution for data-driven innovation to help data scientists uncover data's hidden potential, mine for insights, and predict futures.

Key Features:

- Enterprise-grade, open source platform
- Deploy quickly and scale easily
- More than 1,000 modules
- Hundreds of ready-to-run examples
- Comprehensive range of integrated tools
- The widest choice of advanced algorithms available

Cost: FREE

30. Logical Glue



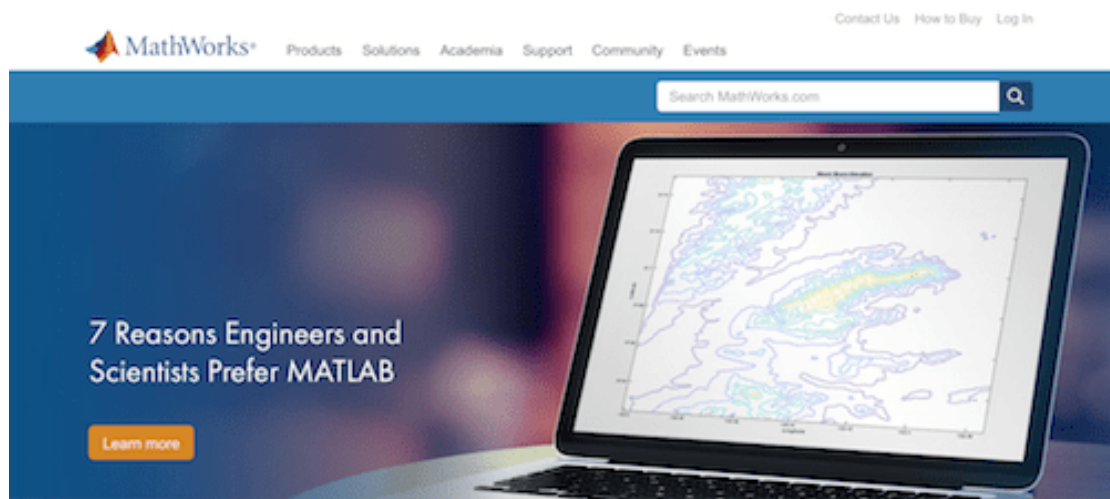
An award-winning white-box machine learning and artificial intelligence platform, Logical Glue increases productivity and profit for organizations. Data scientists choose this tool because it brings our insights to life for our audience.

Key Features:

- Visual narratives that bring insights to life
- Improve the communication and visualization of our insights more easily
- Access new techniques with Fuzzy Logic and Artificial Neural Networks
- Build the most accurate predictive models
- Know exactly which data is predictive
- Simple deployment and integration

Cost: Contact for a quote

31. MATLAB



A high-level language and interactive environment for numerical computation, visualization, and programming, MATLAB is a powerful tool for data scientists. MATLAB serves as the language of technical computing and is useful for math, graphics, and programming.

Key Features:

- Analyze data, develop algorithms, and create models
- Designed to be intuitive
- Combines a desktop environment for iterative analysis and design processes with a programming language capable of expressing matrix and array mathematics directly
- Interactive apps to see how different algorithms work with our data
- Automatically generate a MATLAB program to reproduce or automate our work after we've iterated and gotten the results we want
- Scale analyses to run on clusters, GPUs, and clouds with simple code changes

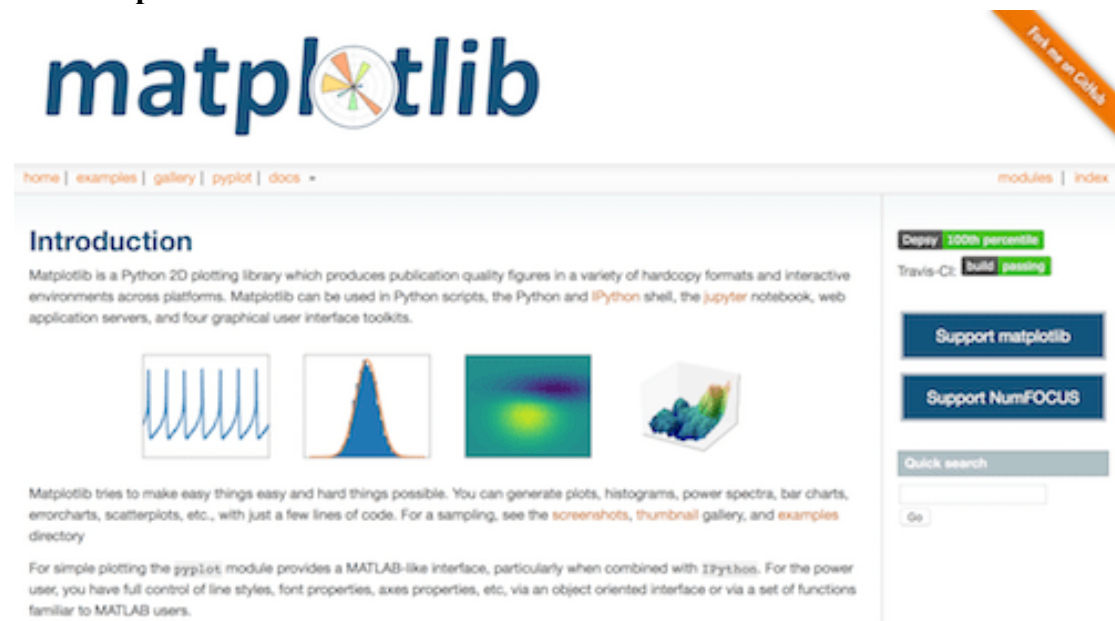
Cost:

MATLAB Standard Individual: \$2,150

MATLAB Academic Use, Individual: \$500

Contact for other licensing options and pricing

32. Matplotlib



Matplotlib is a Python 2D plotting library that produces publication-quality figures in a variety of hardcopy formats and interactive environments across platforms. Data scientists use this tool in Python scripts, the Python and IPython shell, the Jupyter Notebook, web application servers, and four graphical user interface toolkits.

Key Features:

- Generate plots, histograms, power spectra, bar charts, error charts, scatterplots, and more with a few lines of code
- Full control of line styles, font properties, axes properties, etc. with an object-oriented interface or via a set of functions similar to MATLAB
- Several Matplotlib add-on toolkits are available

Cost: FREE

33. MLBase



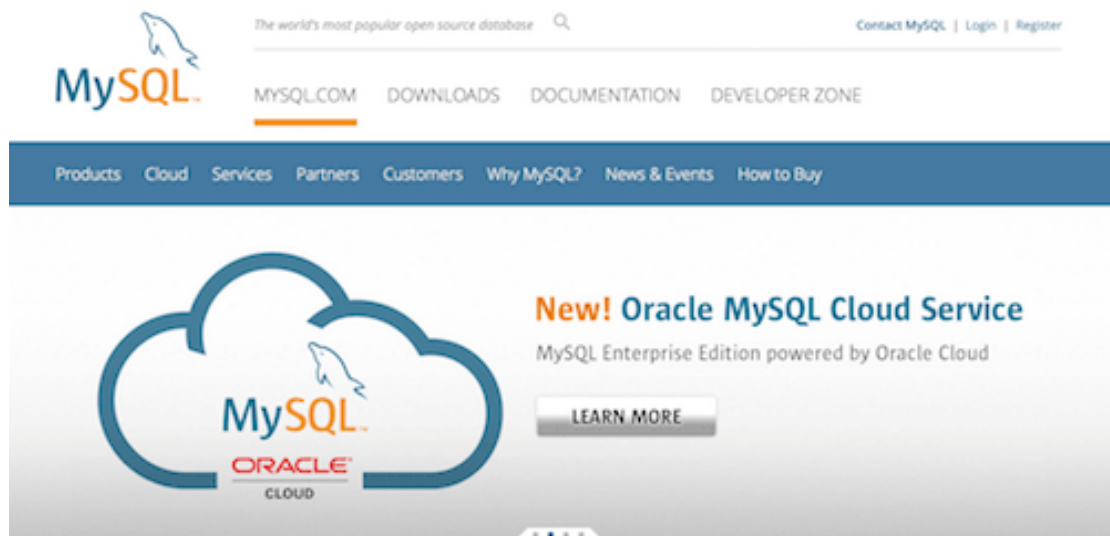
UC Berkeley's AMPLab integrates algorithms, machines, and people to make sense of Big Data. They also developed MLBase, an open source project that makes distributed machine learning easier for data scientists.

Key Features:

- Consists of three components: MLib, MLI, and ML Optimizer
- MLib is Apache Spark's distributed ML library
- MLI is an experimental API for feature extraction and algorithm development introducing high-level machine learning programming abstractions
- ML Optimizer automates the task of machine learning pipeline construction and solves a search problem over feature extractors and ML algorithms
- Implement and consume machine learning at scale more easily

Cost: FREE

34. MySQL



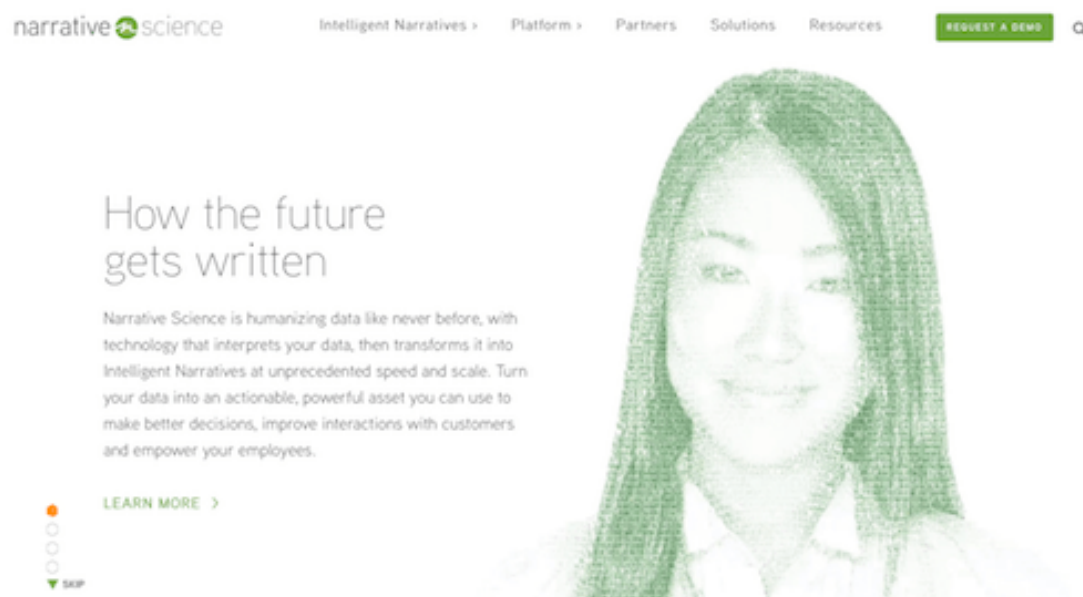
MySQL is one of today's most popular open source databases. It's also a popular tool for data scientists to use to access data from the database. Even though MySQL typically is software in web applications, it can be used in a variety of settings.

Key Features:

- Open source relational database management system
- Store and access our data in a structured way without hassles
- Support data storage needs for production systems
- Use with programming languages such as Java
- Query data after designing the database

Cost: FREE

35. Narrative Science



Narrative Science helps enterprises maximize the impact of their data with automated, intelligent narratives generated by advanced narrative language generation (NLG). Data scientists humanize data with Narrative Science's technology that interprets and then transforms data at unparalleled speed and scale.

Key Features:

- Turn data into actionable, powerful assets for making better decisions
- Help others in the organization understand and act on data
- Integrates into existing business intelligence tools
- Create a new reporting experience that drives better decisions more quickly

Cost: Contact for a quote

36. Natural Language Toolkit (NLTK)

NLTK 3.2.4 documentation

[NEXT](#) | [MODULES](#) | [INDEX](#)

Natural Language Toolkit

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](#) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](#).

Thanks to a hands-on guide introducing programming fundamentals alongside topics in computational linguistics, plus comprehensive API documentation, NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

NLTK has been called "a wonderful tool for teaching, and working in, computational linguistics using Python," and "an amazing library to play with natural language."

[Natural Language Processing with Python](#) provides a practical introduction to programming for language processing. Written by the creators of NLTK, it guides the reader through the fundamentals of writing Python programs, working with corpora, categorizing text, analyzing linguistic structure, and more. The book is being updated for Python 3 and NLTK 3. (The original Python 2 version is still available at http://nltk.org/book_1ed.)

TABLE OF CONTENTS

- NLTK News
- Installing NLTK
- Installing NLTK Data
- Contribute to NLTK
- FAQ
- Wiki
- API
- HOWTO

SEARCH

A leading platform for building Python programs, Natural Language Toolkit (NLTK) is a tool for working with human language data. NLTK is a helpful tool for inexperienced data scientists and data science students working in computational linguistics using Python.

Key Features:

- Provides easy-to-use interfaces to more than 50 corpora and lexical resources
- Includes a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and more
- Learn more from the active discussion forum

Cost: FREE

37. NetworkX

NetworkX

Stable Release

networkx-1.11
30 January 2016
[download](#) | [doc](#) | [pdf](#)

Development

network-2.0dev
[github](#) | [doc](#) | [pdf](#)

Contact

[Mailing list](#)
[Issue tracker](#)



Software for complex networks

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.



Features

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source [3-clause BSD license](#)
- Well tested with over 1800 unit tests and >90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

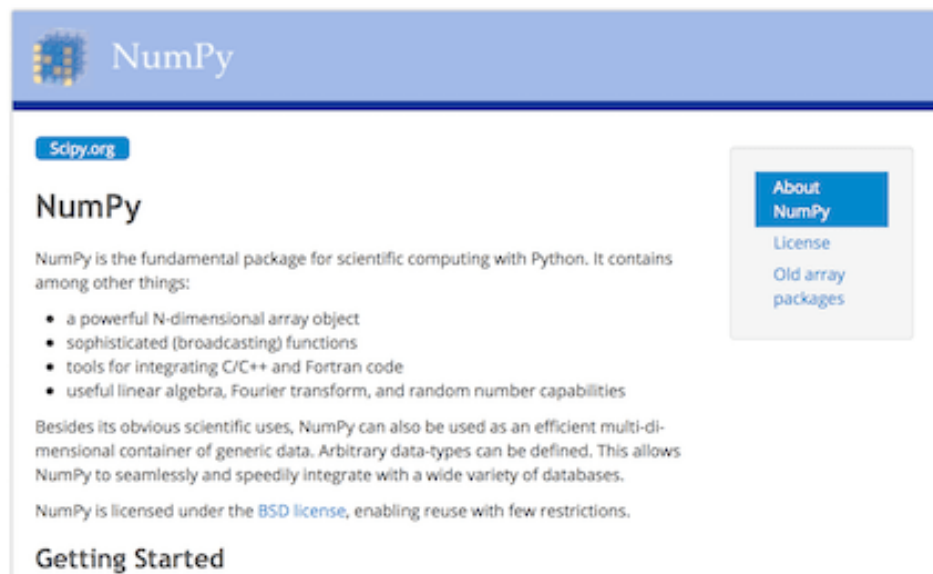
NetworkX is a Python package tool for data scientists. Create, manipulate, and study the structure, dynamics, and functions of complex networks with NetworkX.

Key Features:

- Data structures for graphs, digraphs, and multigraphs
- Abundant standard graph algorithms
- Network structure and analysis measures
- Edges capable of holding arbitrary data
- Generate classic graphs, random graphs, and synthetic networks

Cost: FREE

38. NumPy



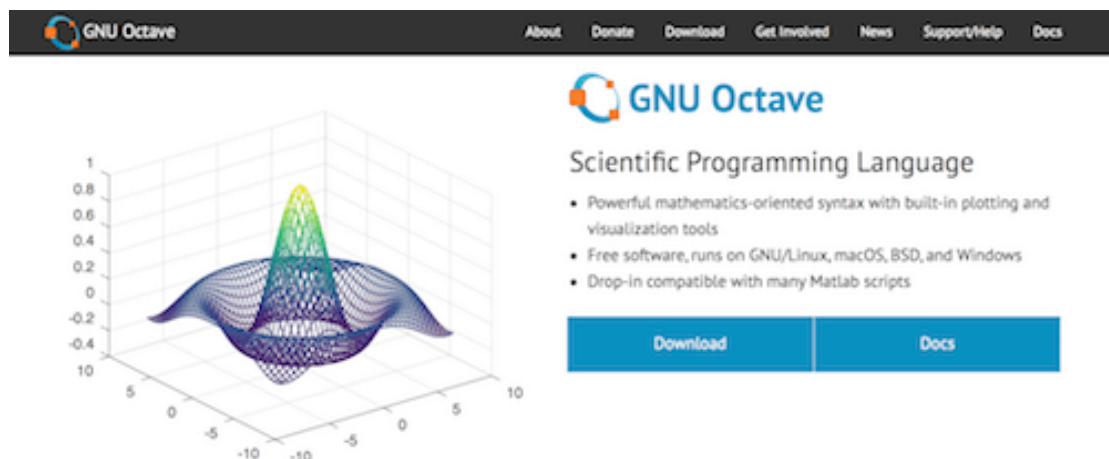
A fundamental package for scientific computing with Python, NumPy is well-suited to scientific uses. NumPy also serves as a multi-dimensional container of generic data.

Key Features:

- Contains a powerful N-dimensional array object
- Sophisticated broadcasting functions
- Tools for integrating C/C++ and Fortran code
- Define arbitrary data-types to seamlessly and speedily integrate with a wide variety of databases

Cost: FREE

39. Octave



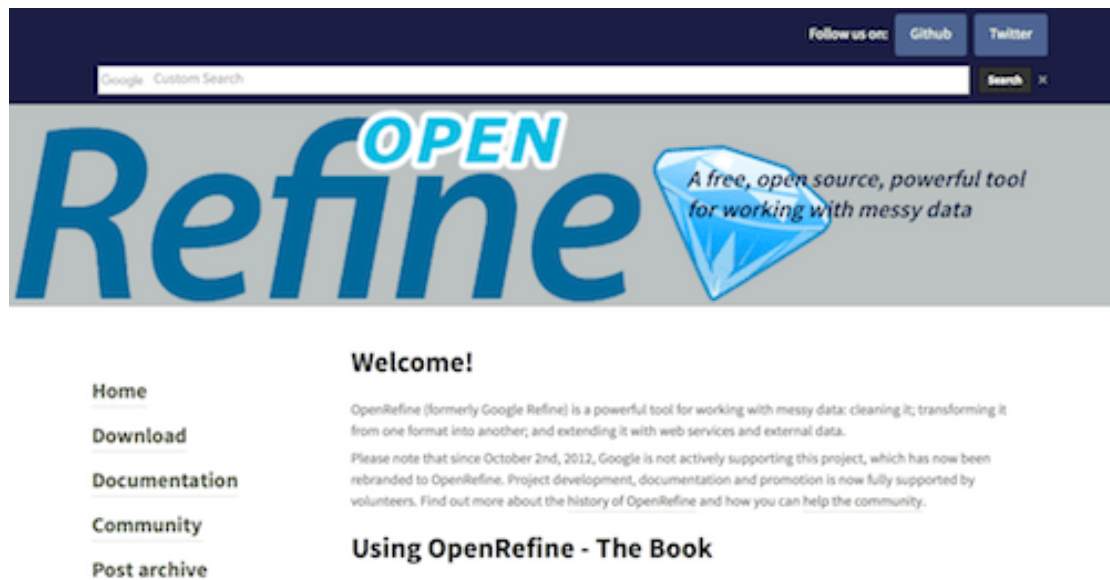
GNU Octave is a scientific programming language that is a useful tool for data scientists looking to solve systems of equations or visualize data with high-level plot commands. This tool's syntax is compatible with MATLAB, and its interpreter can be run in GUI mode, as a console, or invoked as part of a shell script.

Key Features:

- Powerful math-oriented syntax with built-in plotting and visualization tools
- Runs on GNU/Linux, MacOS, BSD, and Windows
- Drop-in compatible with many MATLAB scripts
- Use linear algebra operations on vectors and matrices to solve systems of equations
- Use high-level plot commands in 2D and 3D to visualize data

Cost: FREE

40. OpenRefine




OpenRefine is a powerful tool for data scientists who want to clean up, transform, and extend data with web services and then link it to databases. Formerly Google Refine, OpenRefine now is an open source project fully supported by volunteers.

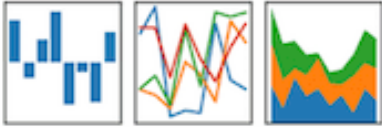
Key Features:


- Explore large datasets easily
- Clean and transform data
- Reconcile and match data
- Link and extend datasets with a range of web services
- We may upload cleaned data to a central database

Cost: FREE

41. Pandas


$$y_{it} = \beta^T x_{it} + \mu_i + \epsilon_{it}$$






[home](#) // [about](#) // [get pandas](#) // [documentation](#) // [community](#) // [talks](#) // [donate](#)

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the [Python](#) programming language.

pandas is a [NUMFocus](#) sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to [donate](#) to the project.

A Fiscally Sponsored Project of


OPEN CODE = BETTER SCIENCE

v0.20.3 Final (July 7, 2017)

VERSIONS

Release
0.20.3 - July 2017
[download](#) // [docs](#) // [pdf](#)

Development
0.21.0 - 2017
[github](#) // [docs](#)

Previous Releases
0.19.2 - [download](#) // [docs](#) // [pdf](#)
0.18.1 - [download](#) // [docs](#) // [pdf](#)
0.17.1 - [download](#) // [docs](#) // [pdf](#)
0.16.2 - [download](#) // [docs](#) // [pdf](#)
0.15.2 - [download](#) // [docs](#) // [pdf](#)
0.14.1 - [download](#) // [docs](#) // [pdf](#)
0.13.1 - [download](#) // [docs](#) // [pdf](#)
0.12.0 - [download](#) // [docs](#) // [pdf](#)

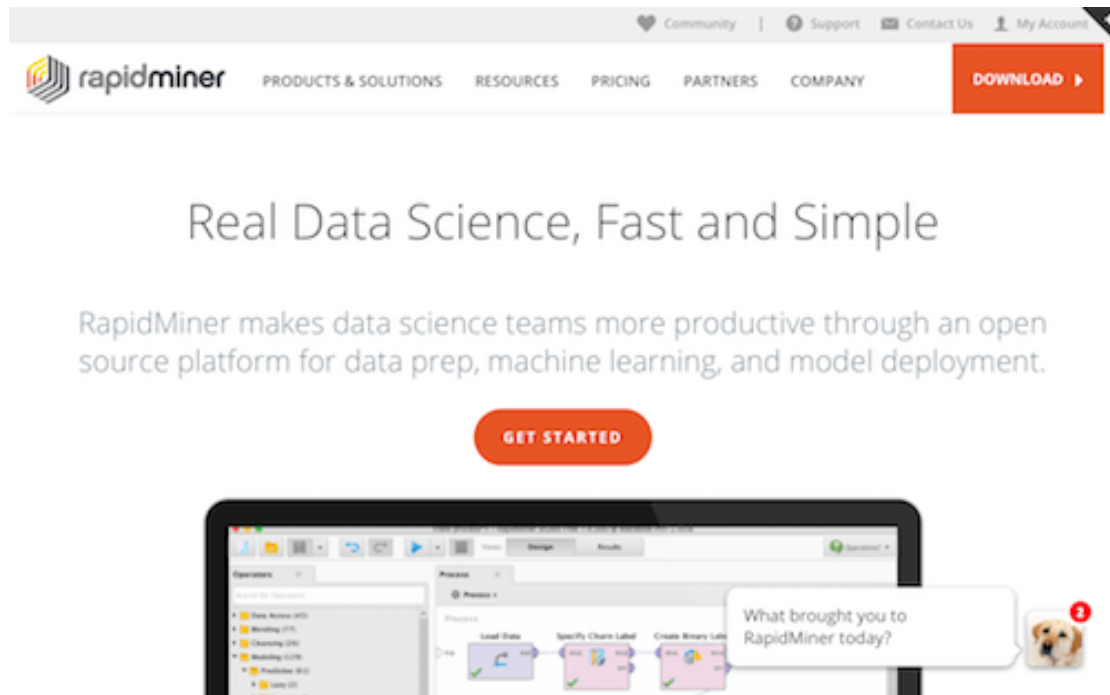
pandas is an open source library that delivers high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Data scientists use this tool when they need a Python data analysis library.

Key Features:

- NUMFocus-sponsored project that secures development of pandas as a world-class, open source project
- Fast, flexible, and expressive data structures make working with relational and labeled data easy and intuitive
- Powerful and flexible open source data analysis and manipulation tool available in a variety of languages

Cost: FREE

42. RapidMiner



Data scientists are more productive when they use RapidMiner, a unified platform for data prep, machine learning, and model deployment. A tool for making data science fast and simple, RapidMiner is a leader in the 2017 Gartner Magic Quadrant for Data Science Platforms, a leader in 2017 Forrester Wave for predictive analytics and machine learning, and a high performer in the G2 Crowd predictive analytics grid.

Key Features:

- RapidMiner Studio is a visual workflow designer for data scientists
- Share, reuse, and deploy predictive models from RapidMiner Studio with RapidMiner Server
- Run data science workflows directly inside Hadoop with RapidMiner Radoop

Cost:

RapidMiner Studio

- ✧ FREE – 10,000 rows of data and 1 logical processor
- ✧ Small: \$2,500/year – 100,000 rows of data and 2 logical processors
- ✧ Medium: \$5,000/year – 1,000,000 rows of data and 4 logical processors
- ✧ Large: \$10,000/year – Unlimited rows of data and unlimited logical processors

RapidMiner Server


- ✧ FREE – 2 GB RAM, 1 logical processor, and 1,000 Web Service API calls

- ✧ Small: \$15,000/year – 16 GB RAM, 4 logical processors, and unlimited Web Service API calls
- ✧ Medium: \$30,000/year – 64 GB RAM, 8 logical processors, and unlimited Web Service API calls
- ✧ Large: \$60,000/year – Unlimited GB RAM, unlimited logical processors, and unlimited Web Service API calls

RapidMiner Radoop

- ✧ FREE – Limited to a single user and community customer support
- ✧ Enterprise: – \$15,000/year – \$5,000 for each additional user and enterprise customer support

43. Redis

 [Commands](#) [Clients](#) [Documentation](#) [Community](#) [Download](#) [Modules](#) [Support](#)

Redis is an open source (BSD licensed), in-memory data structure store, used as a database, cache and message broker. It supports data structures such as strings, hashes, lists, sets, sorted sets with range queries, bitmaps, hyperloglogs and geospatial indexes with radius queries. Redis has built-in replication, Lua scripting, LRU eviction, transactions and different levels of on-disk persistence, and provides high availability via Redis Sentinel and automatic partitioning with Redis Cluster. [Learn more →](#)

Try it

Ready for a test drive? Check this [interactive tutorial](#) that will walk you through the most important features of Redis.

Download it

[Redis 4.0.1](#) is the latest stable version. Interested in release candidates or unstable versions? [Check the downloads page](#).

Quick links

Follow day-to-day Redis on [Twitter](#) and [GitHub](#). Get help or help others by subscribing to our [mailing list](#), we are 5,000 and counting!

Redis is a data structure server that data scientists use as a database, cache, and message broker. This open source, in-memory data structure store supports strings, hashes, lists, and more.

Key Features:

- Built-in replication, Lua scripting, LRU eviction, transactions, and different levels of on-disk persistence
- High availability via Redis Sentinel and automatic partitioning with Redis cluster
- Run automatic operations such as appending to a string, incrementing the value in a hash, pushing an element to a list, and more

Cost: FREE

44. RStudio



RStudio is a tool for data scientists that is open source and enterprise-ready. This professional software for the R community makes R easier to use.

Key Features:

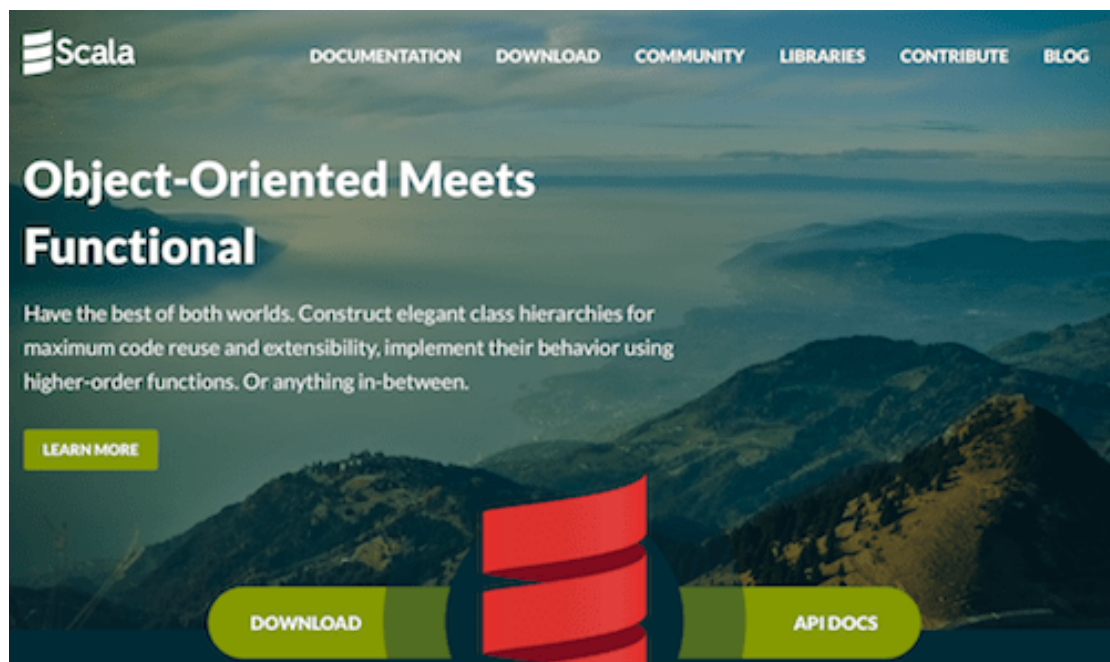
- Includes a code editor, debugging, and visualization tools
- Integrated development environment (IDE) for R
- Includes a console, syntax-highlighting editor supporting direct code execution and tools for plotting, history, debugging, and workspace management
- Available in open source and commercial editions and runs on the desktop or in a browser connected to RStudio Server or Studio Server Pro

Cost:

Open Source Edition: FREE

Commercial License: \$995/year

45. Scala



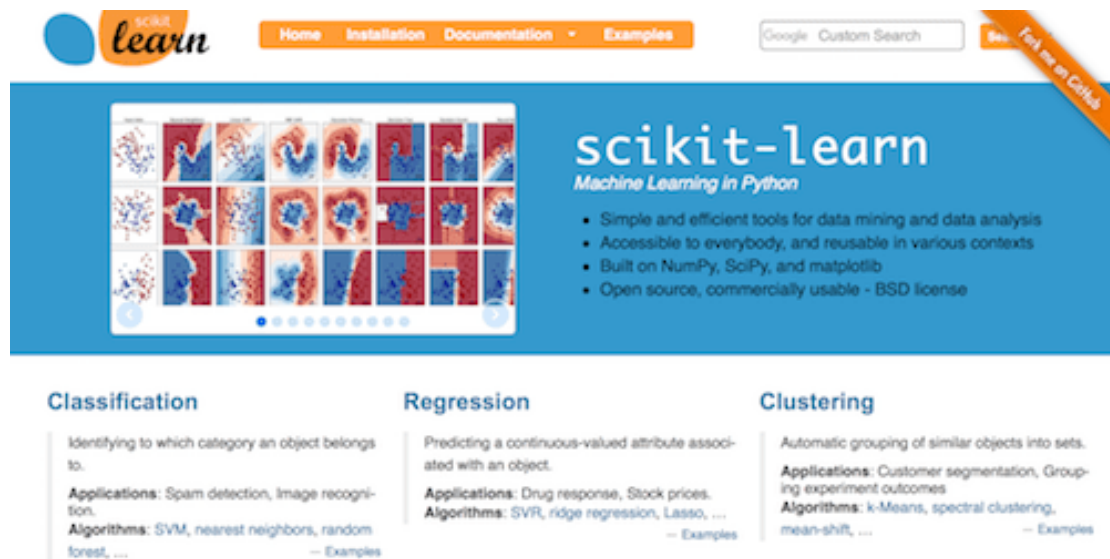
The Scala programming language is a tool for data scientists looking to construct elegant class hierarchies to maximize code reuse and extensibility. The tool also empowers users to implement class hierarchies' behavior using higher-order functions.

Key Features:

- Modern multi-paradigm programming language designed to express common programming patterns concisely and elegantly
- Smoothly integrates features of object-oriented and functional languages
- Supports higher-order functions and allows functions to be nested
- Notion of pattern matching extended to the processing of XML data with the help of right-ignoring sequence patterns using a general extension via extractor objects

Cost: FREE

46. Scikit-learn



scikit-learn is an easy-to-use, general-purpose machine learning for Python. Data scientists prefer scikit-learn because it features simple, efficient tools for data mining and data analysis

Key Features:

- Accessible to everyone and reusable in certain contexts
- Built on NumPy, SciPy, and Matplotlib
- Open source, commercially usable BSD license

Cost: FREE

47. SciPy



SciPy, a Python-based ecosystem of open source software, is intended for for math, science, and engineering applications. The SciPy Stack includes Python, NumPy, Matplotlib, Python, the SciPy Library, and more.

Key Features:

- Scientific computing tools for Python including a collection of open source software and a specified set of core packages
- A community of people who use and develop the SciPy Stack
- SciPy library provides several numerical routines

Cost: FREE

48. Shiny



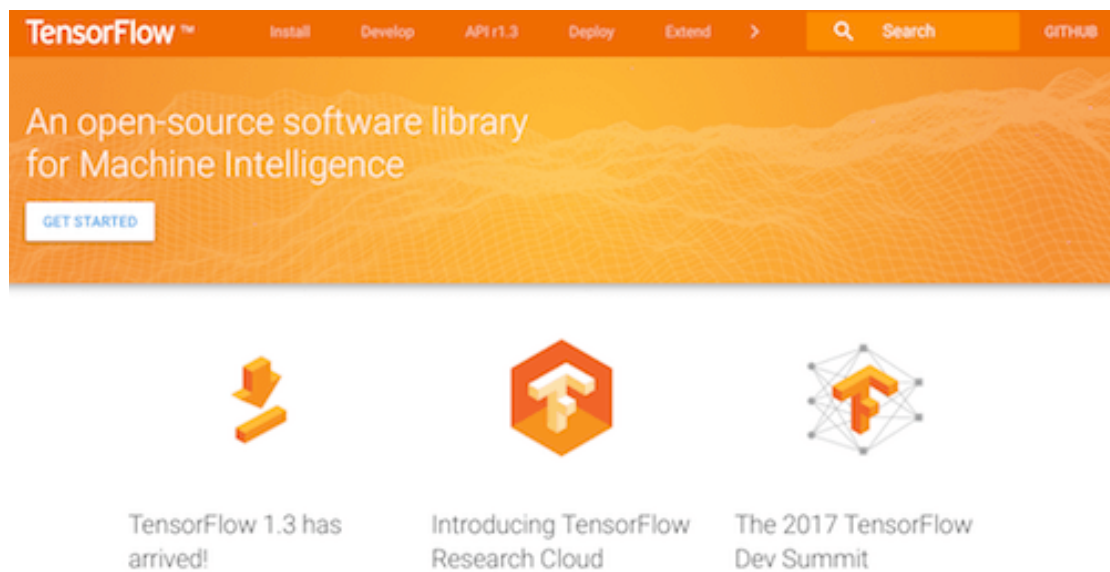
A web application framework for R by RStudio, Shiny is a tool data scientists use to turn analyses into interactive web applications. Shiny is an ideal tool for data scientists who are inexperienced in web development.

Key Features:

- No HTML, CSS, or JavaScript knowledge required
- Easy-to-write apps
- Combines R's computational power with the modern web's interactivity
- Use our own servers or RStudio's hosting service

Cost: Contact for a quote

49. TensorFlow



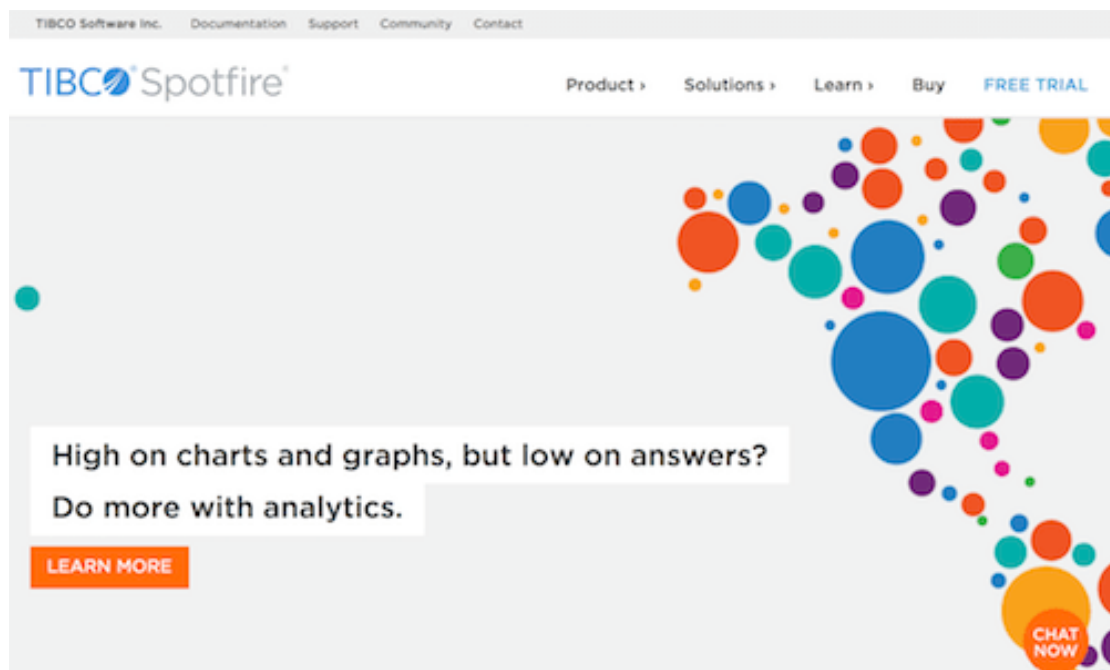
TensorFlow is a fast, flexible, scalable open source machine learning library for research and production. Data scientists use TensorFlow for numerical computation using data flow graphs.

Key Features:

- Flexible architecture for deploying computation to one or more CPUs or GPUs in a desktop, server, or mobile device with one API
- Nodes in the graph represent mathematical operations, while graph edges represent the multidimensional data arrays communicated between them
- Ideal for conducting machine learning and deep neural networks but applies to a wide variety of other domains

Cost: FREE

50. TIBCO Spotfire



TIBCO drives digital business by enabling better decisions and faster, smarter actions. Their Spotfire solution is a tool for data scientists that addresses data discovery, data wrangling, predictive analytics, and more.

Key Features:

- Smart, secure, governed, enterprise-class analytics platform with built-in data wrangling
- Delivers AI-driven, visual, geo, and streaming analytics
- Smart visual data discovery with shortened time-to-insight
- Data preparation features empower us to shape, enrich, and transform data and create features and identify signals for dashboards and actions

Cost: FREE trial available

- ✧ Spotfire Cloud: \$200/month or \$2,000/year; Custom pricing also available
- ✧ Spotfire Platform: Contact for a quote
- ✧ Spotfire Cloud Enterprise: Contact for a quote

CHAPTER 5. Data Scientist

"Data Science = statistics who uses python and lives in San Francisco"

How to be a Data Scientist?

Learn python well.

Now almost the company's data can be given to we, and Python's data processing capabilities are powerful and convenient. In addition to many algorithms in machine learning, Python is also a pretty one. Moreover, its concise and convenient iterative development, takes only 15 minutes to finish an algorithm and check the effect.

Any program can be written in matlab or c++, but write a code in Python can feed we back a quite cool feeling. The handling of irregular input also gives python a huge advantage. In general, in the daily work of a data scientist, all data is stored in raw text, unstructured data. The problem is that these texts can't be directly used as input to various algorithms. We need

- ✧ word segmentation,
- ✧ clause extraction features,
- ✧ sorting out missing data,
- ✧ removing outliers.

At these times, Python can be described as an artifact. The 4 aspects mentioned above here can find the corresponding tool directly in scikit-learn, and even if we want to write a custom algorithm to handle some special requirements, it is a hundred lines of code.

In short, for the challenges of data science, Python allows we to solve problems quickly, rather than worrying about too many implementation details.

Learn the statistics.

The concept of statistics is the "statistical machine learning method." Statistics and computer science have been parallel to each other for decades, creating a series of tools and algorithms created by each other. But until recently people began to notice that what computer scientists call machine learning is actually the prediction in statistics. Therefore, these two disciplines have begun to re-integrate.

Why is statistical learning important?

Because pure machine learning emphasizes the predictive power and implementation of algorithms, statistics have always emphasized "interpretability." For example, for today's Instagram stock issuance is up 20%, we put our two forecast stocks up or down in the model of Twitter, and then show to our boss. Model A has 99% predictability, which is 99% of all times it predicts correctly, but Model B has 95% predictability, but it has an additional attribute – it can tell we why the stock is up or down. So, which one will our boss prefer? Ask ourselves which one will we choose? Obviously the latter one. Because the former has strong predictive power (machine learning), but has no explanatory power (statistical explanation). As a data scientist, 80% of the time we need to explain to customers, teams or supervisors why A is feasible but B is not feasible. If we tell them, "My current neural network can have such a good predictive power, but I can't explain it at all", then no one will believe in we.

Specifically, how to learn statistics?

Learn basic probabilities first, we can start with MIT's probability theory textbook--<<*Introduction to Probability and Statistics*>>. Read and complete all the exercises from Chapter 1 to Chapter 9. (p.s. when interviewing Twitter, someone was asked a question about the probability of taking balls from a box, grabbed from this book).

Understand the basic statistical tests and their assumptions, when can they be used. Quickly understand what terms are used in statistical learning, and what to do, read this <http://d3js.org/>

Learn basic statistical ideas. There are statistics for frequentist, as well as statistics for bayesian. The representative of the former has *Hastie, Trevor, et al.* <<*The elements of statistical learning*>>. Vol. 2. No. 1. New York: Springer, 2009; the latter looks at *Bishop, Christopher M.* <<*Pattern recognition and machine learning*>>. Vol. 1. New York: springer, 2006. The former is the holy book of statistical learning, the biasist, the latter is the holy book of pattern recognition, almost from the perspective of pure bayesian. It is difficult to get down directly at the beginning, but it will benefit a lot.

Reading the above book is a long-term process. But after reading it for a while, I personally feel that it is very worthwhile. If we just know how to use some packages, then we must not be a qualified data scientist. Because as long as the problem changes a little, we don't know how to solve it. If we feel that we are a just half-baked data scientist, then ask the following questions. If there are 2 questions that can't be answered, then just keep learning.

- Why does the feature in the neural network require standardize instead of throwing it directly into?
- Does it need to do Cross-Validation to the Random Forest to avoid over fitting?
- Is it a bad choice to use Naive-Bayesian for bagging? Why?

- When using the ensemble method, especially the Gradient Boosting Tree, do we need to make the structure of the tree more complex (high variance, low bias) or simpler (low variance, high bias)? why?

If we are just getting started, it doesn't matter, it's normal we can not answer these questions. If we are half-baked, understand why we have some gaps with the best data scientist – because we don't know how each algorithm works, when we want to solve our problem with that algorithm, we can't start with countless details. learning data science needs patience, sinking our heart and practicing craftsmanship.

Learn data processing

This step does not have to be done independently of 2). Obviously, we will start to encounter various algorithms when we read these books, and various data will be mentioned in the book here. But the least valuable thing in this era is the data (please, why use the "California house price data" in the 1980s?), value is what is provided to the decision after data analysis. So why not collect some data all by ourselves.

- Start writing a small program, use the API to search random tweets on Twitter (or Instagram)
- Segment the text of these tweets, handle noise (such as advertising)
- Use some off-the-shelf labels as labels, such as inside the tweet it will show how many times this tweet was forwarded.
- Try to write an algorithm to predict that how many times will the tweet be forwarded.
- Test on the unseen data set.

The above process is not a day's work, especially when it is just getting started. Take our time, patience is greater than progress.

Become a full stack engineer.

In a corporate environment, as a newcomer, we can't have the benefits to find a colleague to do for we when we need to write a data visualization. And to find another colleague to do it for we when we need to write and save the data to the database.

Moreover, even if we have this condition, it will waste more time switching the context frequently. For example, we ask a colleague to save our data to the database in the morning, but he will finish the job in the afternoon. Or we need a long time to explain to him what the logic is and what it is.

The best way to change ourselves is to arm ourselves into a versatile workman. We don't need to be an expert in all aspects, but we must understand all aspects and check the documentation to get started.

- NoSQL will be used. In particular, MongoDB;
- Learns basic visualization, and uses basic html and javascript; know the visual library of d3 — *Wasserman, Larry. <<All of statistics: a concise course in statistical inference>>. Springer, 2004*, and highchart <http://www.highcharts.com/>
- Learn basic algorithms and algorithm analysis, know how to analyze algorithm complexity. Average complexity, worst complexity. Each time we write a program, we anticipate the time it takes (predicted by algorithmic analysis). I recommend Princeton's algorithm class www.Coursera.org
- Write a basic server, and use the basic template of flask <http://flask.pocoo.org/> to write a backbone that allows we to do visual analysis.
- Learn to use a handy IDE, VIM, pycharm.

Keep reading

In addition to building our own learning system, we need to know what other data scientists are doing. With the emergence of new technologies, new ideas and new people, we need to communicate with them and expand our knowledge to better cope with new work challenges.

Often, very powerful data scientists put their blogs online for everyone to visit. In addition, there are many powerful data scientists in the academic circle. We don't have to worry about reading the papers. After reading a few articles, we will feel: Ha, I can think of this too! One of the benefits of reading a blog is that if we communicate with them, we can even get an internship from them.

- Betaworks chief data scientist, Gilad Lotan's blog, [Gilad Lotan](#)
- Ed Chi, a six-year bachelor-master-doctor graduated superman, google data science <http://edchi.blogspot.com/>
- Hilary Mason, Chief Scientist of bitly, a well-known data scientist in the New York area: <https://hilarymason.com>

After seeing enough of them here, we will find that there are still many blogs worth watching (they will quote other articles in the article), so snowballing we can have enough things to read on the way to work in the morning.

To sum up, From the experience and style of work of the best data scientists, the real commonality is that:

- ✓ They are very smart – also we can
- ✓ Like what they are doing – If you are not interested, you shouldn't read till here.
- ✓ They can calm down and learn – If we work hard enough, also we can.

Three Core Skills That Data Scientists Need

Data Scientist needs to have a deep understanding of the needs and problems, then process the data and take reasonable quantitative analysis to find the answer. The recommended answer must also be backed by data evidence.

Data Hacking

It is necessary to have the ability to turn data into its own use from a variety of places.

- SQL: used to store and query structured data
- Programming: For example, use Python for parsing/scraping data. If we have a scripting language and a combined/object oriented language, it will be an advantage. Mainly used to process unstructured data
- Hadoop/parallel processing: The data we are processing may be too large (such as the shopping record of the supermarket in the past six months, the credit card company's credit card record within two years) can not be loaded into the memory at one time, and we need to quickly analyze the data, which requires MapReduce and other technologies.

Among them, SQL and Programming are the most basic, we must use SQL to query data, and will quickly write programs to analyze data. Of course, our programming skills don't need to reach the level of software engineers, because most of the code we write is only one-time, will not be reused, and will only be used by we or a colleague, not on the Internet to make countless People click, so the quality of the program is not high.

For a more in-depth analysis of the problem, we may also use:

- Exploratory analysis skills, we can use python, R, matlab and other tools, IT companies use SAS and SPSS relatively few, although some job ads / descriptions mentioned, of course, it is not completely impossible. But if we only have SAS, then there is no doubt that the choice is much less.
- Optimization, Simulation: Some positions need to study customer demand changes, adjust product or service prices to help companies maximize profitability
- Machine Learning, Data Mining: Some people use data mining technology, and found that many people buy diapers in the supermarket, but also bought beer – not yet understood as sputum, but perhaps diapers and beer should be sold together; Accurate delivery.
- Modeling: We need to understand the scope of application, limitations and features of different statistical models. The three scenarios of descriptive, predictive, and prescriptive that I mentioned in the first part are also simple examples.

Problem Solving

We not only need to understand what users say they want, we also need to really understand what they actually mean, transformation defines a problem that can be solved with data, and then choose the right analysis tools, quantitative analysis and problem solving.

Communication

Data scientists will deal with people in many different departments of the company and will have more opportunities to meet senior or business people than Code farmers. If we want to get in touch with a department like marketing and want to deal with a lot of leaders, we need to have strong communication skills. We need to know what is the nature of the problem, what is the technical details, the ability to give high-level analysis and recommendations to the upper-level leaders, the ability to explain and defer our technical details to colleagues, that is “say what people want to hear”. This is not to say that we want to be slippery, but to know when we need to hide technical details and only show the most relevant information to the audience.

We are likely to do presentation often, we need a strong ability to visualize, and it's helpful to be familiar with Edward Tufte and Nathan Yau. In addition, we may like the advanced method, but all solutions must be considered from the perspective of generating business revenue. We may also need to work with the software development team, we need clearly let them understand what they should achieve and what they should improve.

Background of Majors to be a Data Scientist

Computer Science, Information Science, Information Systems, Statistics, and Business (especially Marketing), combined with their own background, it is easy to find the corresponding position. Looking for a job in the future, aiming at a company department that suits his or her own expertise, and has the highest percentage of getting an offer.

E.g.: for graduates from the statistics, those data scientist positions that require strong Java programming skills and even write production code maybe not suitable for them. Or someone is a computer engineer, if in a position description there are quite a lot of statistical models which he/she can't understand and ask to use R programming, then it's better just passing this position.

Other related majors also have chances to be a data scientist, such as EE, with experience of doing signal processing, image processing, communications, and so on, they need both programming and use statistical or mathematical knowledge; Same to

others like IEOR, mathematics, mechanical engineering, etc., they do optimization, Simulation, or Econometrics in economics.

In the current Data Scientists recruitment advertisements, the professional requirements for job seekers are usually written in a series of long bursts: "Applied Mathematics, Statistics, Computer Science, Economics, Operations Research or Engineering" - the last "Engineering" has a very wide scope, generally those do quantitative or computational jobs, who know modelling, are all counted as Engineering. The meaning behind this kind of writing is that as long as you can play statistics, understand mathematics, and make modeling, are all welcome. There are very few technical occupations in the United States, and the educational background requirements for job seekers are as broad as data scientist.

Data Science is a new career with a wide range of careers. It is full of job opportunities. At the same time, no matter what professional you are studying, you have new majors and new fields to learn. For example, during the work period, studying statistics to optimize, measure the economy, and deepen the statistical knowledge of computer learning may have opportunities.

In recent years, many schools in the United States have also opened special analytics course for analytics, such as Northwestern and NCSU. However, colleges with such majors generally have a low overall ranking. Except for Northwestern University, few schools have used analytics in the past. In the recent two years, as the rapid development of data science and big data, UIUC, UT Austin, NYU, etc. have successively opened the analytics, kinds of Data Science master's program.

The biggest advantage of this type of master's program is the arrangement of course: *software system, machine learning, database, optimization, decision science, statistics, business intelligence* and other related domain knowledge, are often involved. Therefore, compared with students such as academic statistics or computer-based students, students with a master's degree in analytics have a more reasonable and comprehensive knowledge structure. It is precisely because of this that it is easy to find a job for this professional master-degree students. More and more professors in the Department of Statistics believe that the students from academic statistics will go to the IT industry to find jobs as data scientist in the future. They also hope to reform the curriculum and make their own students full of skills.

Master or Ph.D. Degrees?

For doing the jobs of data scientists, what is the difference between having a master or doctor degrees?

We can easily understand that getting a Ph.D. is more conducive to doing data scientist. For example, with the same background of statistics, as the master we may be doing the underlying work, a little far away from the core things, the basic content of the work may be to use SQL to transfer the data, do some preliminary processing in R. However, the starting point of PhD is as high as up to do advanced work such as statistical analysis, modeling. This is similar to the statistics/biometric master and PhD: the former is a typical SAS programmer, the latter is directly statistician or biostatistician.

Although Data Scientist is a new profession, the truth is generally the same. With the same background of Computer Science, the masters become software engineers, do the underlying implementation, especially the company's newcomers, the system framework has been set up by the seniors, what they are required is just to do coding jobs, the range of the problems they need to take into consideration is very limited, which is just writing the code they are responsible for. PhDs, probably Research SDE(Software Development Engineer), Applied Scientist/Researcher, cooperate with other groups to implement the core Machine Learning or Information Retrieval system, and they are generally required to think independently and to address some issues with our own ideas and solutions, the boss also encourages we to innovate and gives we the freedom to explore new projects.

In addition, the more technically advanced companies, the more they tend to recruit PhDs, the more advanced positions they offer, and the higher their salaries. If we are a master's student, maybe we will be hired only if we already have several years' experience of work. Those just graduated from a master's degree, no matter which major, but knowing only some master's level courses, may not meet the company's requirements.

Of course, whether it is master or PhD, it's our personal abilities that majorly counts for being offered any kind of jobs. There is a lot of PhD who can't find the position of research SDE, but can only do ordinary SDE jobs, some even can't find a job. On the other side, the masters can also get rid of the "regular route" and get a good development. What I want to emphasize here is that PhDs have more opportunities and a higher probability. In the profession like Data Scientist, they are far superior to the master.

If we are a master and have a choice of multiple job offers, it is recommended to consider long-term development, such as a large company let we use SQL data, give 80,000 dollars a year, another company lets we do statistical analysis for 70,000 a year,

the suggestion is going for the latter, if we are dissatisfied with the salary and the company's reputation, we can switch jobs after two years.

Some companies, especially large ones, may offer more training opportunities. If our skills are significantly different from the three core skills that data scientists need: Data Hacking, Problem Solving and Communication, then Training opportunities should be used as much as possible. The shortcomings of company training, one is that the course is too applied, it is very helpful for we to quickly learn to write Java programs, but for complex knowledge, such as MCMC, machine learning, the company basically can't train we, it is better to learn systematically in school. Moreover, if our daily work is to write SQL data, write java programs to achieve various functions, we have no energy to study the Casual Inference, SVM, which is not used at work, even if we study, there is no chance to practice and to get an in-depth understanding.

The suggestion is that we must get rid of the underlying work and strive to be a high-level point. Otherwise, writing SQL for 5 years, versus spending 5 years doing analysis and modeling, the future will be very different.

CONCLUSION

During this thesis we have talked about what is data science, what data scientists do, what professions are preferred, and what skills they need.

A brief introduction of data science has been given, we know the concepts of Predictive Analytics, Descriptive Analytics, Prescriptive Analytics.

Then 12 main methodological tools as “PROGENITORS” of data science were described too.

We discussed the comparison of data science, data analytics and big data, the applications of each field, the skills required of each field, as well as their salaries.

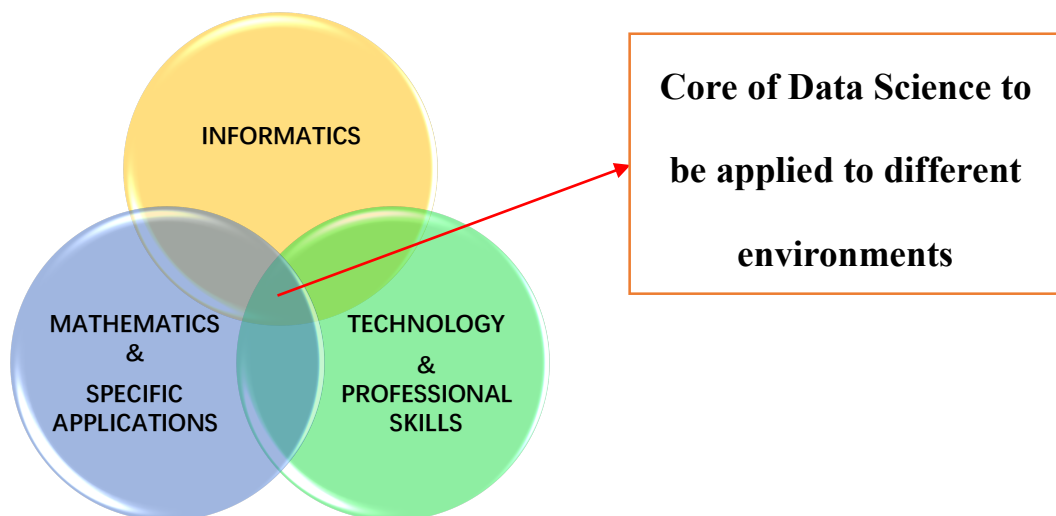
We have talked about the business strategies in 7 steps:

- (1) identify our organization’s key business drivers for data science
- (2) create an effective team for achieving data science goals
- (3) emphasize communication skills to realize data science’s value
- (4) expand the impact of data science through visualization and storytelling
- (5) give data science teams access to all the data
- (6) prepare data science processes for operationalizing analytics
- (7) improve governance to avoid data science “creepiness”

Then we introduce 50 top tools for data scientists—analytics tools, data visualization tools, database tools, and more—their features and costs.

Three core skills that data scientists need are:

Data Hacking
Problem Solving
Communication



“The most important boom is the smaller one”

To be a good data scientist, we need:

- ✧ Learn python well.
- ✧ Learn the statistics.
- ✧ Learn data processing
- ✧ Become a full stack engineer.
- ✧ Keep reading

Almost all companies collect a huge amount of data from their production processes: basically all those at no cost or at very low cost!

In practice, it happens that, when taking important decisions, some data, essential in order to take decisions in a proper and documented way, are always missing.

one of the most important results of the consulting interventions (according to the six sigma or the world class manufacturing approaches) consists precisely in identifying the data of the production process that the companies have to record (even accepting some extra cost) in order to take correct decisions and continuously increase quality.

After this premise, it is easy to realize how the methods (previously mentioned), to squeeze as much information as possible from the collected data, have been considered of primary importance since many years.

There are also those who, even rightly, contest the qualification of "*scientific*" to the DATA SCIENCE approach. For example, **Gary Angel**, CEO of **DIGITAL MORTAR**, says: *«The idea that this type of general problem solving procedure is the explanation for the success of science seems implausible on its face and is contradicted by experience. Implausible because the method as described is so contentless. How do I pick which problems to tackle from the infinite set available? **The method is silent.** How do I generate hypothesis? **The method is silent.** How do I know they are testable? **The method is silent.** How do I test them? **The method is silent.** How do I know what to do when a test doesn't refute a hypothesis? **The method is silent.** How many failures to refute a hypothesis is enough to prove it? **The method is silent.** How do I communicate the results? **The method is silent**».*

It seems difficult to contest this point of view. However, we can overlook the question of whether **data science** can really be considered a "**scientific**" method in all the ways and fields of application in which it is currently proposed.

Much more simply, we can restrict to consider the data science as the collection of all methods and methodological tools useful to «let the data speak». Methodological methods and tools are surely "scientific", while their applications can certainly be less: maybe because, as they need to accept aspects of uncertainty and/or risk not quantifiable (e. G. Judgments of subjective Bayesian estimates), they represent the only way to face the problem with rationality, having as alternative to blindly rely on good luck or bad luck!

In ANY CASE, THE **Data SCIENCE** is a cornerstone of the **Factory 4.0**.

At present ***Factory 4.0*** is a very important topic. Nowadays technology is able to automatically detect a large amount of data from production processes and to record them in real time. This enables more correct and prompt decisions, respectively because they are based on an amount of information much larger than before and on their standardized elaborations; and then because at least many of the operational decisions can be automatically taken.

Of course, this approach involves considerable benefits on quality and costs, but also requires a reorganization of the company, not free from social repercussions due to the likely reduction in the number of Employees and to the need of adapting/enriching the skills of Workers and Employees. And, on this last point, suggestions get wild: machinery, processes, research, mathematics, computer science, statistics, etc.

Therefore, the implementation of *Factory 4.0* will necessarily have to be planned over a certain period of time, but bearing in mind that this is **the essential path of the future**.

References

- 【01】 Course <<Product Quality Design>> Prof. Mario Vianello, Polito.
- 【02】 Juran, J. M., Gryna, F. M. (1980). Quality Planning and Analysis, 2nd ed. New York: McGraw-Hill.
- 【03】 Shainin, R. D. (1995). A common sense approach to quality management. 49th Annual Quality Congress Proceedings, ASQC, pp. 1163–1169
- 【04】 Shainin, R. D. (1993). Strategies for technical Problem solving. Quality Engineering, 5:433–438
- 【05】 <https://www.ngdata.com/top-tools-for-data-scientists/>
- 【06】 <https://www.datascienceweekly.org/articles/what-tools-do-employers-want-data-scientists-to-know>
- 【07】 <https://www.zhihu.com/question/21592677>
- 【08】 <http://www.1point3acres.com/core-skills-for-data-scientists/>
- 【09】 <http://www.1point3acres.com/phd-or-master-for-data-science/>
- 【10】 <https://baike.baidu.com/item/数据学和数据科学/3565373>
- 【11】 <https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article>
- 【12】 <http://www.1point3acres.com/what-is-data-science-analytics/>
- 【13】 https://en.wikipedia.org/wiki/Artificial_neural_network
- 【14】 https://en.wikipedia.org/wiki/Design_of_experiments
- 【15】 https://en.wikipedia.org/wiki/Dorian_Shainin
- 【16】 https://en.wikipedia.org/wiki/Genetic_algorithm