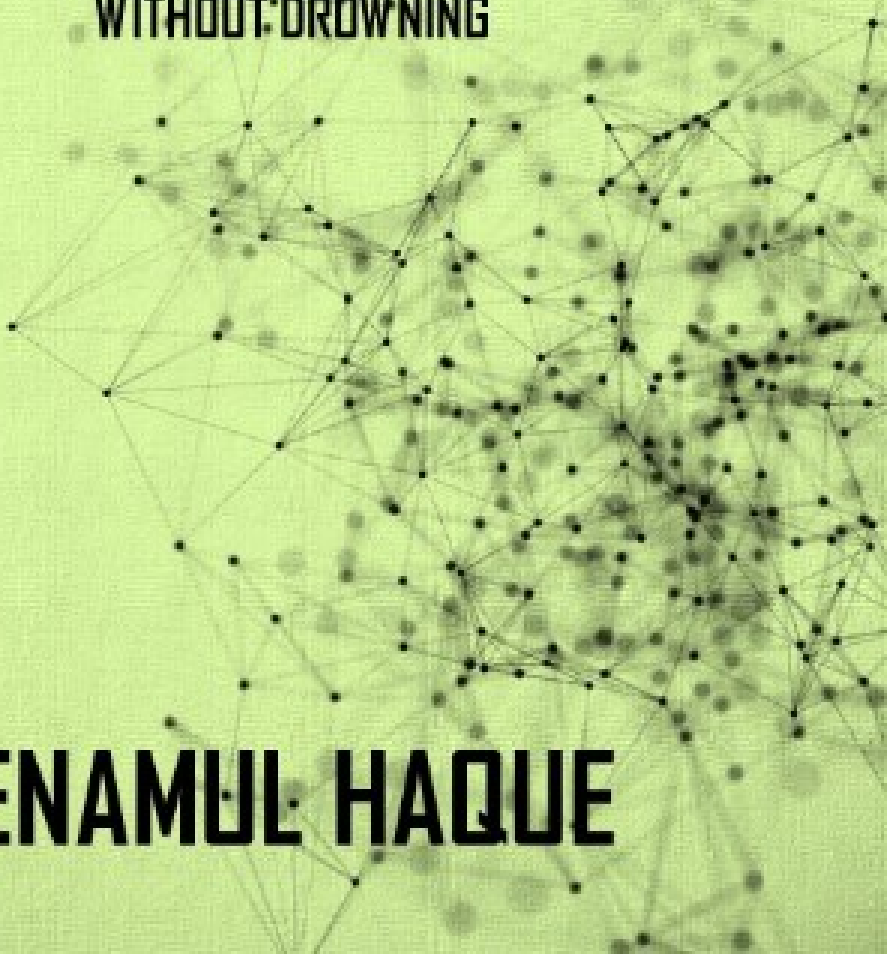
A BEGINNER'S GUIDE TO DATA SCIENCE

HOW TO DIVE INTO THE DATA OCEAN WITHOUT DROWNING



ENAMUL HAQUE

A Beginners Guide To DATA SCIENCE

How to dive into the data ocean without drowning

ENAMUL HAQUE

Enel Publications
London, UK
Amazon Kindle Direct Publishing

First Printing Edition, April 2021 (Revision 1)

ISBN 9798731261074

CHAPTER THREE:
DATA SCIENCE KNOW-HOW

"The excitement in data science is in the journey toward achieving three significant kinds of results: discovery, insights, and innovation." - Kirk Borne, Principal Data Scientist at Booz Allen Hamilton.

Data Science Learning Journey

If you plan your journey to the data science world, I presume you are, as you purchased this book. This section will guide you understand the new skills, provide directions and ideas for those motivated ones. You will want to fully understand the concepts and details of various machine learning algorithms, data science concepts, etc. Therefore, I recommend that you start from the base before looking at machine learning algorithms or data analysis applications. Suppose you do not have a basic understanding of calculus and integrals, linear algebra and statistics. In that case, it will be difficult for you to understand the underlying mechanics of the various algorithms. Likewise, if you don't have a basic understanding of Python, it will be difficult for you to translate your knowledge into real-world applications. Below is the order of the topics that I recommend studying:[i]

- Mathematics and Statistics.
- Basics of programming.
- Machine learning algorithms and concepts.

Mathematics and statistics

As with everything else, you should learn the basics before getting into the fun stuff. Trust me, it would be much easier for me if I started by learning math and statistics before getting started with some machine learning algorithms. Three general topics that I recommend looking at are calculus/integrals, statistics, and linear algebra (in no particular order).

Integrals

Integrals are essential when it comes to probability distribution and hypothesis testing. While you don't need to be an expert, it's in your best interest to learn the basics of integrals. If you know absolutely nothing about integrals, I recommend that you take the Khan Academy course. Here are links to several practical tasks to hone your skills:

- ***Introduction to integrals:*** https://towardsdatascience.com/an-integrals-crash-course-for-data-science-cf6e6dd7c046
- ***A crash course on integrals:*** https://www.albert.io/blog/how-to-solve-integrals-ap-calculus-crash-course/
- ***Khan Academy:*** Integral Calculus: https://www.khanacademy.org/math/integral-calculus
- ***Practical Questions (start with block 6):*** https://www.albert.io/ap-calculus-ab-bc?utm_source=blog&utm_medium=blog&utm_campaign=ap-calculus

Statistics

If there is any topic that you should focus on, it is statistics. After all, a data scientist is a genuinely modern statistician, and machine learning is a modern term for statistics. If you have time, I recommend taking the Georgia Tek course called Statistical Techniques (https://mediaspace.gatech.edu/playlist/dedicated/74258101/1_g5xwvbde/1_iw8fk73m), which covers the basics of probability, random variables, probability distribution, hypothesis testing, and more. If you don't have time to devote yourself to this course, I highly recommend watching the Khan Academy video on statistics (https://www.khanacademy.org/math/statistics-probability).

Linear algebra

Linear algebra is fundamental if you want to dive into deep learning. It is helpful to know other basic machine learning concepts such as principal component analysis and recommender systems. For mastering linear algebra, I also recommend Khan Academy (https://www.khanacademy.org/math/linear-algebra)

Fundamentals of programming

Just as a fundamental understanding of math and statistics is essential, a basic knowledge of programming will make your life so much easier, especially when it comes to implementation. Therefore, I recommend that you take the time to learn the basic languages - SQL and Python, before diving into machine learning algorithms.

SQL

It doesn't matter where to start, but I would start with SQL. Why? It is easier to learn and valuable to know if you are employed in a company that works with data, even if you are not a data scientist.

If you are new to SQL, I recommend checking out Mode's SQL (https://mode.com/sql-tutorial/introduction-to-sql/) tutorials as they are very concise and detailed. If you want to learn more advanced concepts, see the list of resources where you can learn advanced SQL.

Below are a few resources you can use to practice SQL:

- ***Resources on Leetcode***: https://leetcode.com/problemset/database/
- ***Resources on HackerRank:*** https://www.hackerrank.com/domains/sql?filters%5Bstatus%5D%5B%5D=unsolved&badge_type=sql
- ***Examples of implementation:*** https://docs.google.com/document/d/1_-pPj_HusumXskhsXFOccimhDSloWkAyEdCOxv7mZFY/edit#heading=h.sspk8oxbveqv

Python

Once you start with Python, you will probably stay with this language for the rest of your life. It's far ahead in terms of open-source contributions and easy to learn. I have found that learning

Python through practice is much more rewarding. Nevertheless, after taking several Python crash courses, I concluded that this course is the most complete (and free!).

Introduction to Python Programming - Georgia Tech
https://www.edx.org/professional-certificate/introduction-to-python-programming

Pandas

Perhaps the most critical library to know is Pandas, which is specifically designed for data manipulation and analysis. Below are two resources that should accelerate your learning curve. The first link is a tutorial on how to use Pandas, and the second link contains many practical tasks that you can solve to solidify your knowledge!

- ***Learn pandas on Kaggle:*** https://www.kaggle.com/learn/pandas
- ***Practice with Randas on dozens of hands-on tasks:*** https://github.com/guipsamora/pandas_exercises

Algorithms and concepts of machine learning

This part is split into two others: machine learning algorithms and machine learning concepts. Every machine learning algorithm has three components:

Representation: how to represent knowledge. Examples include decision trees, sets of rules, instances, graphical models, neural networks, support vector machines, model ensembles, etc.

Evaluation: the way to evaluate candidate programs (hypotheses). Examples include accuracy, prediction and recall, squared error, likelihood, posterior probability, cost, margin, entropy k-L divergence, etc.

Optimisation: the way candidate programs are generated, known as the search process. For example, combinatorial optimisation, convex optimisation, constrained optimisation.

Machine learning algorithms

The next step is to learn about the various machine learning algorithms, how they work and when to use them. Below is a list of the various machine learning algorithms and resources you can use to learn each of them.

- ***Linear Regression:***
  - Georgia Tech: https://www2.isye.gatech.edu/~sman/courses/6739/SimpleLinearRegression.pdf
  - StatQuest: https://www.youtube.com/watch?v=nk2CQITm_eo&ab_channel=StatQuestwithJoshStarmer
- ***Logistic regression:***

    - StatQuest: https://www.youtube.com/watch?v=yIYKR4sgzI8&ab_channel=StatQuestwithJoshStarmer
- ***K nearest neighbours:***
    - MIT: https://www.youtube.com/watch?v=09mb78oiPkA&ab_channel=MITOpenCourseWare
- ***Decision trees:***
    - StatQuest: https://www.youtube.com/watch?v=7VeUPuFGJHk&ab_channel=StatQuestwithJoshStarmer
- ***Naive Bayes***
    - Terence Sheen: https://towardsdatascience.com/a-mathematical-explanation-of-naive-bayes-in-5-minutes-44adebcdb5f8
    - Luis Serrano: https://www.youtube.com/watch?v=Q8l0Vip5YUw&ab_channel=LuisSerrano
- ***Support Vector Machines:***
    - SVM Tutorial by Alice Zhao: https://www.youtube.com/watch?v=N1vOgolbjSc&ab_channel=AliceZhao
- ***Neural networks:***
    - Terence Sheen: https://towardsdatascience.com/a-beginner-friendly-explanation-of-how-neural-networks-work-55064db60df4
- ***Random forests:***
    - StatQuest: https://www.youtube.com/watch?v=J4Wdy0Wc_xQ&ab_channel=StatQuestwithJoshStarmer
- ***AdaBoost:***
    - Terence Sheen: https://towardsdatascience.com/a-mathematical-explanation-of-adaboost-4b0c20ce4382?source=friends_link&sk=956d985b9578c3d272e3851a53ee822a
    - StatQuest: https://www.youtube.com/watch?v=LsK-xG1cLYA&t=9s&ab_channel=StatQuestwithJoshStarmer
- ***Gradient boosting:***
    - StatQuest: https://www.youtube.com/watch?v=OtD8wVaFm6E&t=1s&ab_channel=StatQuestwithJoshStarmer
- ***XGBoost:***
    - StatQuest: https://www.youtube.com/watch?v=OtD8wVaFm6E&t=1s&ab_channel=StatQuestwithJoshStarmer

- ***Principal component analysis:***
    - StatQuest: https://www.youtube.com/watch?v=FgakZw6K1QQ&ab_channel=StatQuestwithJoshStarmer

Machine learning concepts

In addition, there are a few fundamental concepts of machine learning that you will want to learn as well. Below is a (non-exhaustive) list of concepts that I highly recommend learning. Many interview questions are based on these topics!

- ***Regularisation:*** https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a
- ***The bias-variance dilemma:*** https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229
- ***Confusion matrix and related metrics:*** https://towardsdatascience.com/understanding-the-confusion-matrix-and-how-to-implement-it-in-python-319202e0fe4d?source=friends_link&sk=434d5a02fcaec213208c2eeb1174b5c6
- ***Area under the ROC and ROC curve (video):*** https://www.youtube.com/watch?v=4jRBRDbJemM
- ***Bootstrap fetch:*** https://towardsdatascience.com/what-is-bootstrap-sampling-in-machine-learning-and-why-is-it-important-a5bb90cbd89a
- ***Ensemble training, bagging and boosting:*** https://towardsdatascience.com/ensemble-learning-bagging-and-boosting-explained-in-3-minutes-2e6d2240ae21
- ***Normalisation and standardisation:*** https://www.statisticshowto.com/probability-and-statistics/normal-distributions/normalized-data-normalization/#:~:text=Normalization%20vs.-,Standardization,a%20standard%20deviation%20of%201.

Projects in the field of data science

By this point, you will not only have built a solid foundation, but you will also have a solid understanding of the fundamentals of machine learning. Now it's time to work on some personal side projects. If you would like to see some simple examples of data science projects, check these out:

- ***Predicting Wine Quality with Several Classification Techniques:*** https://towardsdatascience.com/predicting-wine-quality-with-several-classification-techniques-179038ea6434
- ***Coronavirus data visualizations using Plotly:*** https://towardsdatascience.com/coronavirus-data-visualizations-using-plotly-cfbdb8fcfc3d
- ***14 Data Science Projects to do During Your 14 Day Quarantine:*** https://towardsdatascience.com/14-data-science-projects-to-do-during-your-14-day-quarantine-8bd60d1e55e1

[This section of the book has web links to lots of resources; if you have difficulties connecting them, you can go to my blog and find this as an article with links enabled directly to those resources: ***https://enamulhaque.co.uk/my-articles/f/starting-your-data-science-learning-journey*** *or just go to my site* ***enamulhaque.co.uk*** *and find this article under the category "data science"]*

Building a Data Science Career

The data science industry is booming so much that a study found that more than 97,000 jobs are currently open for analytics and data science in India, other countries I have mentioned in different parts of this book to bring the relevancy context. While there's all the hype, glitter and traffic in the "hottest work of the 21st century," many fans are still nifty about what's right. And less, you know what it means to be a data scientist. Let's try to pave the way for you to start with data science as a career.

As we understood by now, the research area in which information is extracted from all the data obtained is data science. There is a great need for professionals who can make data analysis a competitive advantage for their business. In your career as a data scientist, you create data-driven business applications and analytics.

What does a data scientist do?

A fusion of mathematicians, computer scientists and pattern spotters are data scientists. The data scientist's mission is to decode large amounts of data and conduct more research to uncover data patterns and gain a deeper insight into the meaning. By studying highly complex data sets to get information that companies can put into practice, data scientists work between the business and IT worlds and drive the industry forward.

What skills are required to excel in data science?

To excel in data science, you need a range of skills from a variety of fields. The most important skills required are research, probability, machine learning, statistics, database knowledge, etc. If you're not sure how to start with it and which way to go, read the article to dispel your doubts. This is an industry where there are many opportunities, so the jobs are waiting for you now and in the future once you have the training and qualification.

How do you become a data scientist?

Companies worldwide have frequently collected and analysed data about their customers to improve quality and improve business outcomes. In today's digital world, we can collect large amounts of data, which requires non-traditional approaches and tools for data processing.

Netflix has over 120 million users worldwide! Netflix uses sophisticated data science metrics to process all the material. This allows him to present a better movie, show its users reviews, and make better shows. Many of the hit series on Netflix was created with data science and big data. The organisation took into account where people stopped at the fast forward and where they no longer saw the program. Netflix was able to create a perfect display by analysing this knowledge.

Skills to become data scientists

You need to learn the following areas to become a data scientist:

1) Use resources such as Oracle Database, MySQL, Microsoft, SQL Server, and Teradata to familiarise yourself with the database required for data storage and analysis.
2) Statistics is one of the most critical skills to be a good data scientist. People tend to skip this step, but you can only succeed in this area if you have a thorough knowledge of this topic. Learn statistics, statistical analysis. The science involved in developing and analysing methods for collecting, evaluating, interpreting, and presenting scientific data is statistics.
3) Probability is the calculation of the probability that there will be a case. You need to be familiar with the concepts of boundaries, integration and differentiation, series, and analysis capabilities.
4) Learn at least one programming language. Programming tools such as R, Python, and SAS are very important when performing data analysis.
5) Practice wrangling data, including washing, manipulating and organising information. R, Python, Flume and Scoop are popular methods for data wrangling.
6) Master the principles of machine learning. Provide the ability of systems to automatically learn from and improve experiences without being specifically programmed for them. Various algorithms such as regressions, Naive Bayes, SVM, K Means Clustering, KNN and Decision Tree algorithms can be used to achieve machine learning, to name a few.
7) Big data tools such as Apache Spark, Hadoop, Talend, and Tableau, which are used to handle massive and complex data that cannot be processed with traditional data processing applications, have in-depth knowledge.
8) Learning visualisation capability is very important to be a successful data scientist. This can be achieved by integrating different records and creating a visual display of the results using charts, charts, and graphs. To do this, you need to learn how to use tools like PowerBI, Tableau, and so on.

Careers in the field of data science?

Data scientists

By optimising and improving product growth, data-driven business solutions and analytics are generated by data scientists. They use predictive models to improve and optimise customer interactions, revenue generation, ad alignment, and more. To integrate models and track performance, data scientists also work with various functional teams.

Data Scientist can find patterns in large data sets, know the field of machine learning well, and confidently own such tools as R, Weka, Python, and Scikit-Learn and Pandas. Data Scientist can extract the most from the data and design algorithms that will give answers to the right questions.

Data Science is quite broad in itself, and there are a few more specialisations:

"Classic" Data Mining - allows you to solve such problems as credit scoring, predict the probability of marriage in production, calculate the probability of clicking on the banner.

Text Mining - allows you to find patterns in the text, automatically define its subject matter, understand the post in the social network - it was painted positively or negatively.

Image processing - allows you to find images in the photo, recognise the text in the picture, determine whether the patient has cancer, based on X-ray analysis - and more. It is in this area now ruled by the ball of neural network and deep learning.

Audio processing - lately, we've all used to say, OK, Google, what's going on in the movies?"

Recommendation systems - tasks from this area allow you to pick up for the user a film, a book or a product that best corresponds to his interests.

Data engineer

Large, complex data sets are created by data engineers. You define, design, and integrate changes to the internal process and then build the infrastructure needed to efficiently extract, transform, and load data. You also develop tools for analytics that use the data pipeline.

Data architect

A data management practitioner and data architecture discipline involved in designing, building, deploying, and managing an organisation's data architecture is called a data architect. The structural specifications for new software and applications are evaluated by data architects, and database solutions are created. You install and configure information systems and move data from legacy systems to new ones.

Data analyst

Data analysts collect data and store databases from primary or secondary sources. You interpret the information, evaluate the results using statistical methods, and create data collection systems and other solutions that help management prioritize business and information requirements.

Business analyst

By identifying and arranging criteria, business analysts help a company process and track data. Creating informative, actionable, and repeatable reports validates resource requirements and create cost estimation models. They help make the right decisions for the business based on historical data and current scenarios.

Data managers

Data administrators help you develop databases and update existing databases. You are responsible for setting up and testing new systems for database and data processing, maintaining database protection and integrity, and developing complex query definitions that enable data extraction.

The need for data scientists is enormous and constantly increasing. When you work with data science, there are many employment opportunities. To improve customer service, multinational companies are still filtering and refining data. To get the best results, major industries such as banking, healthcare, transportation, and e-commerce sites use data science.

The planet is constantly gearing up for a better version of itself. In general, it paves the way for data science to deal with large amounts of data and satisfy customers. So it's the best time to improve these skills and start your career in data science today.

CDO

CDO can be understood as Chief Digital Officer or Chief Data Officer in terms of data and digital use. Many companies are opening positions that are associated with the CIO and have no less weight, such as Chief Digital Officer, Chief Data Officer. The new development is connected with the desire of shareholders not to miss the benefit of the possible solution of pressing technological problems. Let's look at these two positions.

CDO — Chief Digital Officer

The Chief Digital Officer (CDO) is nominated as a corporate superstar. When companies need a rare combination of a top-class technician and a business expert, they are increasingly turning to professionals who are able to provide confident leadership in an ever-changing market. As a result, the size of the compensation packages of such professionals is growing rapidly, and their search is becoming more intense, according to the study "The Rise of the Director of Digital Technologies" prepared by Russell Reynolds Associates, which is engaged in the selection and evaluation of executives on a global scale.

"There is a huge demand for CDO. And it will remain high as more and more companies from different sectors seek to expand the use of digital technology," said Tak Richards, Managing Director of Russell Reynolds Associates Technology and Business Transformation. "They need to ask themselves whether their leaders have the knowledge and experience to understand the complex world of mobile, social and local technologies."

CDO — Chief Data Officer

With the advent of CDO (data directors) and other senior data specialists in senior management, large organisations change their approach to data management.

Data professionals are the driving force behind innovation and differentiation, revolutionising existing business models, improving its communication with the target audience and opening up new business efficiency opportunities.

According to analysts Gartner, the drive to improve the efficiency of using information resources will lead to a sharp increase in the number of companies with a full-time position of Director of Data (CDO). However, only half of them will succeed in meeting the targets by the end of 2019. Data directors will have to create a strategy that identifies indicators linking their activities to measurable business results.

The combination of high expectations and low awareness of data management technologies can make it difficult for data directors to generate budgets and help business users, which is essential for project success, analysts say. Many directors are already reporting conflicts with IT over control over information resources. But successful data directors manage to connect with IT directors, overcome resistance and lead reforms. Analysts recommend explaining to company executives the role of data

and information in business. It's also a good idea to highlight the basic level of data monetisation and information management that can measure progress.

Becoming a Data Scientist

There are 5 basic data-science-specialist skills, the presence of which will bring real benefits to the company. Professionals with such abilities are rare, but this does not mean that they should not be sought and tried to attract to the team. At once, it will be about the requirements for specialists in a relatively large company and not in the department of scientific development and research but in the operating business.

Understanding of business problems

The ability to understand the business problem and assess its potential benefits for the business. The first thing data science has to deal with is vague, ill-conceived and often technically impossible questions. Why is this happening? Because there are very few people who can broadcast their ideas to the hypotheses being tested. Even fewer people who are versed in statistics are enough to understand precisely how data can be used for business development.

Most employees will perceive the data science specialist either as an improved version of the beer table in Excel or as a magic device, obliged within 24 hours to give an answer to any question.

The data science specialist has to evaluate the meaning of the idea, the realism of its implementation, and the potential benefits for the company. The simplest "lice" test data science should be able to do is the "to what" test. It consists of several questions:

- Imagine that we did this analysis or developed this model - what will we do with it next?
- How can we assess her contribution to the company's business?
- How do we get it into production?
- How do we assess its benefit compared to the current solution?

If the project manager or curator can't answer these questions clearly, you should send him to think or sit down to think with him.

Business requirement translation

The ability to translate a business task into a technical solution. If the boss has managed to answer the questions above, the task should be broadcast to a technical solution. It's almost always a non-trivial moment. Imagine, for example, that data science needs to optimise advertising spending on affiliate sites. There are about a few dozen options for solving such a problem. You need to develop and choose the fastest, easy to implement, inexpensive, tested and objective method.

Minimum Viable Product

The ability to quickly bring a solution to the state of a minimum viable product (MVP). The market for those wishing to become a data scientist in Europe and the United States is flooded with people from the academy - post PhD or Postdoc. The sad consequence of this is a penchant for perfectionism and an attempt to spend many months getting the "perfect product" or worse - an effort to improve the existing algorithm. Maybe it's not very bad in academia, but for business, it's a real headache. 95% of business tasks do not require the development of new algorithms and months of work.

Conventionally speaking, a simple logistical regression or basic ranking algorithm will be of great benefit. Trying to write code from scratch for a virtual neural network is months, with zero benefit and fair business disappointment in the benefits of analytical approaches.

Communication

The ability to broadcast the analytic network process (ANP) to production (working with developers). This item is slightly different depending on the company's size, but in general, when it comes to a large company, bringing the model into production will inevitably affect several teams of developers and system administrators.

The consequence is that data scientist should be able to communicate their thoughts to people from non-data science environments (as well as understand what they are told in response).

Terminologically, all this can be very difficult and sometimes, frankly, painful. Besides, there is another common and not very clear problem for newcomers - the concept of scalability of the solution in data science and developers can vary greatly. Conventionally speaking, one minute to process a request in the data science world may be a good thing, but if you need to serve hundreds of thousands of requests per minute in real-time, it's no good. Ideally, a data scientist should have a minimal idea of possible bottlenecks when put into production.

Assessing the benefits

The ability to objectively assess an MVP's benefits and make sure that this solution is actually used by the company. These are two different skills, but for simplicity, we will consider it one task. How to assess the benefit of the solution? If the site has good traffic - then A/B-testing and once again testing, if there is no traffic - you can go straight to the founder of the company and explain that most of the budget still needs to be spent on marketing and sales, and not on the development of models, the benefits of which are even impossible to assess.

It should also be taken into account that it will take you 95% of the time to implement and "fix" the model in business processes and convince everyone around you of the approach's usefulness compared to business as usual. Not for development, not for production, but for your solution to really become part of the business.

These five skills can be summarised in one word - ownership. In practice, only such data science specialists are beneficial to the company. Therefore, the most critical sign of a good data scientist is thinking like a business owner.

Data Scientist vs Data Engineer

Let's imagine that a particular company is engaged in the online sale of household appliances. Each time a site visitor clicks on a specific product, a new data item is created. The data engineer needs to understand how to collect this data, what type of metadata will be added for each click event, and how to store it in an accessible format. The data scientist, in turn, needs to get data about which customers bought certain products and use them to predict the ideal offer of household appliances for each new site visitor.

Or, suppose you are the data scientist of some publisher's paid online library. You want to analyse the history of the actions of users of the library site and see what activities are associated with users who spend more money. Your colleague, a data engineer, will need to collect information from server logs and website event logs. To do this, he needs to create a pipeline that will "swallow" the site logs and server logs in real-time, analyse them and correlate them with a specific user. Then the engineer will need to ensure that the analysed logs are stored in the database so that they can be easily requested later. It turns out that the data engineer, in contrast to the data scientist, is a more applied, narrower position. A data engineer's activity is aimed at painstaking work on the formation of data pipelines and their further maintenance.

Data Science Without Coding

In fact, coding is an important part of data science, but you can do without it using appropriate support tools (but it's better to code). So here's a list of these tools, with short descriptions.[ii]

RapidMiner

RapidMiner (RM) was initially started in 2006 as a stand-alone open-source software named Rapid-I. Over the years, they have given it the name of RapidMiner and attained ~35Mn USD in funding. The tool is open-source for the old version (below v6), but the latest versions come in a 14-day trial period and licensed after that.

RM covers the entire life-cycle of prediction modelling, starting from data preparation to model building and finally validation and deployment. The GUI is based on a block-diagram approach, something very similar to Matlab Simulink. There are predefined blocks that act as plug and play devices. You just have to connect them in the right manner, and a large variety of algorithms can be run without a single line of code. On top of this, they allow custom R and Python scripts to be integrated into the system.

Their current product offerings include the following:

RapidMiner Studio: A stand-alone software that can be used for data preparation, visualisation and statistical modelling

RapidMiner Server: It is an enterprise-grade environment with central repositories which allow easy teamwork, project management and model deployment

RapidMiner Radoop: Implements big-data analytics capabilities centred around Hadoop

RapidMiner Cloud: A cloud-based repository that allows easy sharing of information among various devices

RM is currently being used in various industries, including automotive, banking, insurance, life Sciences, manufacturing, oil and gas, retail, telecommunication and utilities.

DataRobot

DataRobot (DR) is a highly automated machine learning platform built by all-time best Kagglers, including Jeremy Achin, Thoman DeGodoy and Owen Zhang. Their platform claims to have obviated the need for data scientists. This is evident from a phrase from their website – "Data science requires math and stats aptitude, programming skills, and business knowledge. With DataRobot, you bring the business knowledge and data, and our cutting-edge automation takes care of the rest."

DR proclaims to have the following benefits:

Model Optimisation: The platform automatically detects the best data pre-processing and feature engineering by employing text mining, variable type detection, encoding, imputation,

scaling, transformation, etc. Hyper-parameters are automatically chosen depending on the error-metric and the validation set score

Parallel Processing: Computation is divided over thousands of multi-core servers. Uses distributed algorithms to scale to large data sets

Deployment: Easy deployment facilities with just a few clicks (no need to write any new code)

For Software Engineers: Python SDK and APIs available for quick integration of models into tools and software.

With funding of ~60Mn USD and more than 100 employees, DR looks in good shape for the future.

BigML

BigML is another platform with ~Mn USD in funding. It provides a good GUI which takes the user through 6 steps as following:

Sources: use various sources of information

Datasets: use the defined sources to create a dataset

Models: make predictive models

Predictions: generate predictions based on the model

Ensembles: create an ensemble of various models

Evaluation: very model against validation sets

These processes will obviously iterate in different orders. The BigML platform provides a nice visualisation of results and has algorithms for solving classification, regression, clustering, anomaly detection and association discovery problems. You can get a feel of how their interface works using their YouTube channel.

Google Cloud Prediction API

The Google Cloud Prediction API offers RESTful APIs for building machine learning models for android applications. This platform is specifically for mobile applications based on Android OS. Some of the use cases include:

Recommendation Engine: Given a user's past viewing habits, predict what other movies or products a user might like.

Spam Detection: Categorise emails as spam or non-spam.

Sentiment Analysis: Analyse posted comments about your product to determine whether they have a positive or negative tone.

Purchase Prediction: Guess how much a user might spend on a given day, given his spending history.

Though the API can be used by any system, specific Google API client libraries build for better performance and security. These exist for various programming languages- Python, Go, Java, JavaScript, .net, NodeJS, Obj-C, PHP and Ruby.

Paxata

Paxata is one of the few organisations which focus on data cleaning and preparation, not the machine learning or statistical modelling part. It is an MS Excel-like application that is easy to use,

with visual guidance making it easy to bring together data, find and fix dirty or missing data, and share and re-use data projects across teams. Like others mentioned here, Paxata eliminates coding or scripting, overcoming technical barriers involved in handling data.

Paxata platform follows the following process:

Add Data: use a wide range of sources to acquire data

Explore: perform data exploration using powerful visuals allowing the user to easily identify gaps in data

Clean+Change: perform data cleaning using steps like imputation, normalisation of similar values using NLP, detecting duplicates

Shape: make pivots on data, perform grouping and aggregation

Share+Govern: allows sharing and collaborating across teams with solid authentication and authorisation in place

Combine: a proprietary technology called SmartFusion allows combining data frames with 1 click as it automatically detects the best combination possible; multiple data sets can be incorporated into a single AnswerSet

BI Tools: allows easy visualisation of the final answers in commonly used BI tools; also allows easy iterations between data preprocessing and visualization

With funding of ~25Mn USD, Praxata has set its foot in financial services, consumer goods and networking domains. It might be a good tool to use if your work requires extensive data cleaning.

Trifacta

Trifacta is another startup focusing on data preparation. It has 2 product offering:

Wrangler – a free stand-alone software

Wrangler Enterprise – licensed professional version

Trifacta offers a very intuitive GUI for performing data cleaning. It takes data as input and provides a summary with various statistics by column. Also, it automatically recommends some transformations for each column, which can be selected using a single click. Multiple modifications can be performed on the data using some pre-defined functions called easily in the interface.

Trifacta platform uses the following steps of data preparation:

Discovering: this involves getting a first look at the data and distributions to get a quick sense of what you have

Structure: this involves assigning proper shape and variable types to the data and resolving anomalies

Cleaning: this step includes processes like imputation, text standardization, etc. which are required to make the data model ready

Enriching: this step helps in improving the quality of analysis that can be done by either adding data from more sources or performing some feature engineering on existing data

Validating: this step performs final sense checks on the data

Publishing: finally, the data is exported for further use

With ~75Mn USD in funding, Trifacta is currently being used in the financial, life sciences and telecommunication industry.

Narrative Science

Narrative Science is based on a unique idea in the sense that it generates automated reports using data. It works like a data story-telling tool that used advanced natural language processing to create reports. It is something similar to a consulting report.

Some of the features of this platform include:

Incorporates specific statistics and past data of the organisation

Makes of the benchmarks, drivers and trends of the specific domain

It can help generate personalised reports targeted to a specific audience

With ~30Mn USD in funding, Narrative Science is currently being used in financial, insurance, government and e-commerce domains. Some of its customers include American Century Investments, PayScale, MasterCard, Forbes, Deloitte, etc.

Having discussed some startups in this domain, let's move on to some of the academic initiatives trying to automate some aspects of data science. These have the potential of turning into the thriving enterprise in future.

MLBase

MLBase is an open-source project developed by AMP (Algorithms Machines People) Lab at the University of California, Berkeley. The core idea is to provide an easy solution for applying machine learning to large scale problems.

It has 3 offerings:

MLib: It works as the core distributed ML library in Apache Spark. It was initially developed as part of MLBase project, but now the Spark community supports it

MLI: An experimental API for feature extraction and algorithm development that introduces high-level ML programming abstractions.

ML Optimiser: This layer aims to automate the task of ML pipeline construction. The optimiser solves a search problem over feature extractors and ML algorithms included in MLI and MLlib.

This undertaking is still under active development, and we should hear about the products in the near future.

WEKA

Weka is a data mining software written in Java, developed at the Machine Learning Group at the University of Waikato, New Zealand. It is a GUI based tool that is very good for beginners in data science, and the best part is that it is open-source. You can learn about it using the MOOC offered by the University of Waikato here. You can learn more about it in this article.

Though Weka is currently more used in the academic community, it might be the stepping stone of something big coming up in future.

Automatic Statistician

The Automatic Statistician is not a product but a research organisation creating a data exploration and analysis tool. It can take in various kinds of data and use natural language processing

to generate a detailed report. It is being developed by researchers who have worked in Cambridge and MIT and also won Google's Focussed Research Award with a price of $750,000.

More Tools

MarketSwitch – This tool is more focused on optimisation rather than predictive analytics

algorithms.io – This tool works in the domain of IoT (Internet of Things) and performs analytics on connected devices

wise.io – This tool is focused on customer handling and ticket system analytics

Predixion – This is another tool that works on data collected from connected devices

Logical Glue – Another GUI based machine learning platform that works from raw data to deployment

Pure Predictive – This tool uses a patented Artificial Intelligence system which obviates the part of data preparation and model tuning; it uses AI to combine 1000s of models into what they call "supermodels."

DataRPM – Another tool for making predictive models using a GUI and no coding requirements

ForecastThis – Another proprietary technology focussed on machine learning using a GUI

Data Science experts to follow

Here is a list of top data science experts of modern days; you can follow them to subscribe to their knowledge.

- Alex "Sandy" Pentland - @alex_pentland
- Andrew Ng - @AndrewYNg
- Bernard Marr - @BernardMarr
- Chris Surdak - @Csurdak
- Dean Abbott - @deanabb
- Dhanurjay Patil - @dpatil
- Fei-Fei Li - @drfeifei
- Geoffrey Hinton - @geoffreyhinton
- Hilary Mason - @hmason
- Jeff Hammerbacher - @hackingdata
- John Elder - @johnelder4
- John Myles White - @johnmyleswhite
- Judea Pearl - @yudapearl
- Jurgen Schmidhuber - @SchmidhuberAI
- Kenneth Cukier - @kncukier
- Kira Radinsky - @KiraRadinsky
- Lillian Pierson - @Strategy_Gal
- Nando de Freitas - @NandoDF
- Peter Norvig - Peter@Norvig.com
- Richard Socher - @RichardSocher
- Sebastian Thrun - @SebastianThrun

- Yann Lecun - @ylecun
- Yoshua Bengio - https://yoshuabengio.org/

[i] How I Would Study Data Science If I Started A Couple Of Years Ago, or A Guide to Learning Data Science Effectively - https://prog.world/how-i-would-study-data-science-if-i-started-a-couple-of-years-ago-or-a-guide-to-learning-data-science-effectively/

[ii] Datascientist.one – Data Science tools - http://datascientist.one/data-science-tools-4noncoders/