Principles of Data Science

Presented by

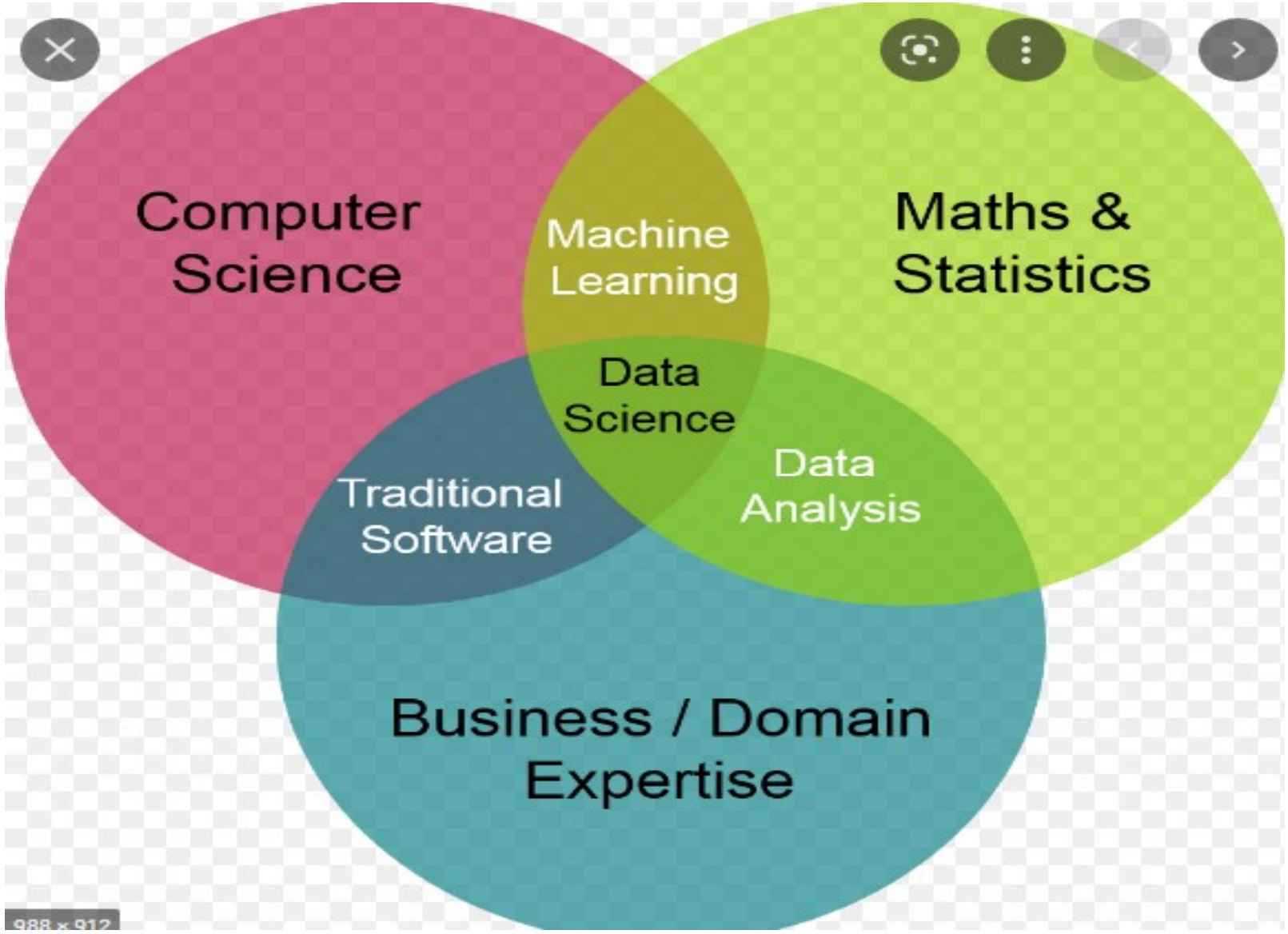
R. Ramya
Guest Lecturer,
School of computer science, Engineering and Applications,
Bharathidasan University,
Khajamalai Campus,
Trichy-620 023.

Unit – I: Introduction

- Introduction to Data Science
- Evolution of Data Science
- Data Science Roles
- Stages in a Data Science Project
- Applications of Data Science in various fields
- Data Security Issues.

Introduction to Data Science

- **Data Science** is the area of study which involves extracting insights from vast amounts of data using various scientific methods, algorithms, and processes.
- It helps you to discover hidden patterns from the raw data.
- The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.
- Data Science is an interdisciplinary field that allows you to extract knowledge from structured or unstructured data.
- Data science enables you to translate a business problem into a research project and then translate it back into a practical solution.


Computer Science
Machine Learning
Maths & Statistics
Data Science
Traditional Software
Data Analysis
Business / Domain Expertise
988 x 912

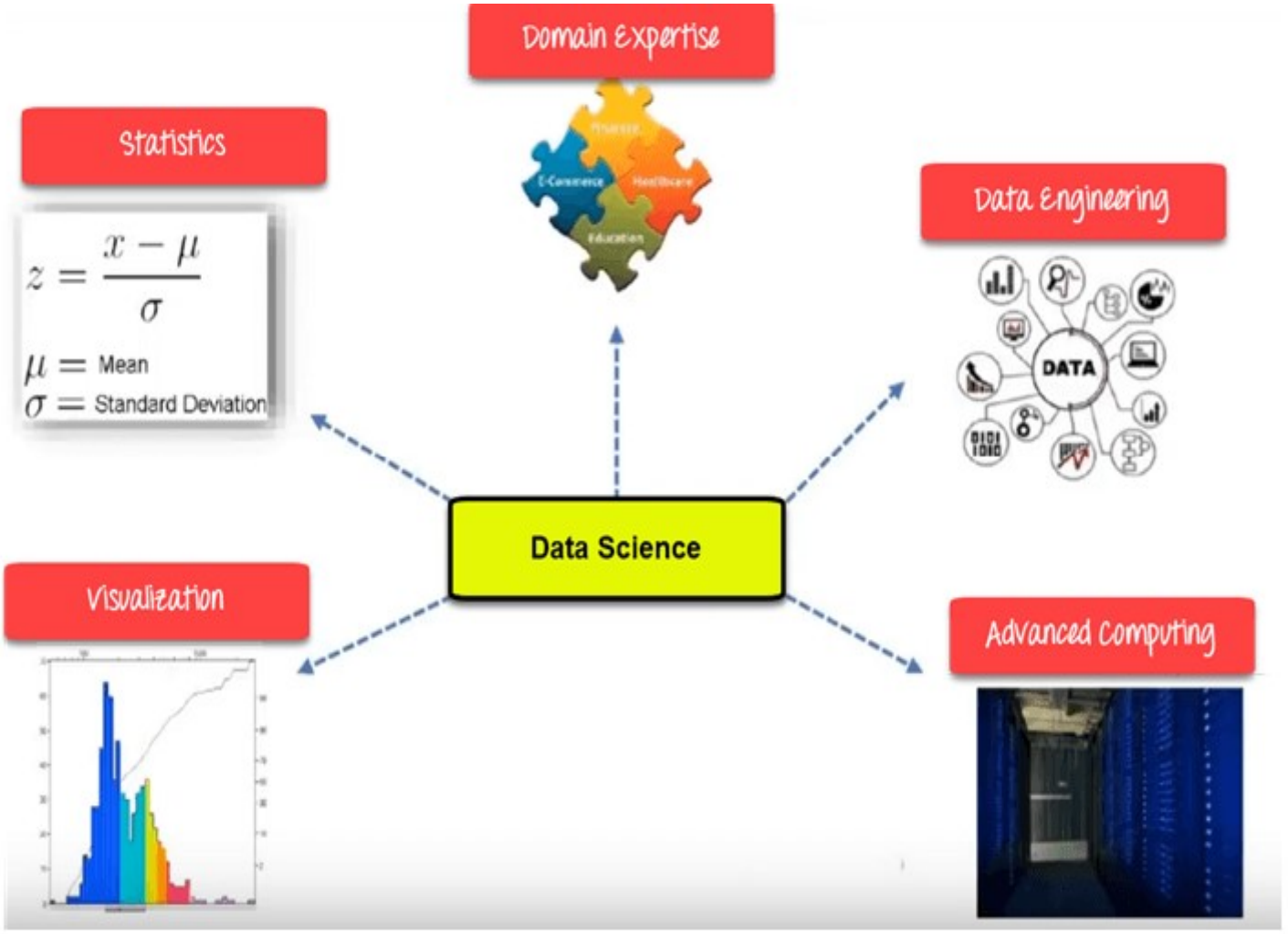
Why Data Science?

Here are significant advantages of using Data Analytics Technology:

- Data is the oil for today's world. With the right tools, technologies, algorithms, we can use data and convert it into a distinct business advantage
- Data Science can help you to detect fraud using advanced machine learning algorithms
- It helps you to prevent any significant monetary losses
- Allows to build intelligence ability in machines
- You can perform sentiment analysis to gauge customer brand loyalty
- It enables you to take better and faster decisions
- It helps you to recommend the right product to the right customer to enhance your business

Evolution of DataSciences

Data Science Components



Statistics:

- Statistics is the most critical unit of Data Science basics, and it is the method or science of collecting and analyzing numerical data in large quantities to get useful insights.

Visualization:

- Visualization technique helps you access huge amounts of data in easy to understand and digestible visuals.

Machine Learning:

- Machine Learning explores the building and study of algorithms that learn to make predictions about unforeseen/future data.

Deep Learning:

- Deep Learning method is new machine learning research where the algorithm selects the analysis model to follow.

DATA SCIENCE JOBS ROLES

Most prominent Data Scientist job titles are:

- Data Scientist
- Data Engineer
- Data Analyst
- Statistician
- Data Architect
- Data Admin
- Business Analyst
- Data/Analytics Manager

Data Scientist:

- **Role:** A Data Scientist is a professional who manages enormous amounts of data to come up with compelling business visions by using various tools, techniques, methodologies, algorithms, etc.
- **Languages**: R, SAS, Python, SQL, Hive, Matlab, Pig, Spark

Data Engineer:

- **Role**: The role of a data engineer is of working with large amounts of data. He develops, constructs, tests, and maintains architectures like large scale processing systems and databases.
- **Languages**: SQL, Hive, R, SAS, Matlab, Python, Java, Ruby, C + +, and Perl

Data Analyst:

- **Role**: A data analyst is responsible for mining vast amounts of data. They will look for relationships, patterns, trends in data. Later he or she will deliver compelling reporting and visualization for analyzing the data to take the most viable business decisions.
- **Languages**: R, Python, HTML, JS, C, C+ + , SQL

Statistician:

- **Role**: The statistician collects, analyses, and understands qualitative and quantitative data using statistical theories and methods.
- **Languages**: SQL, R, Matlab, Tableau, Python, Perl, Spark, and Hive

Data Administrator:

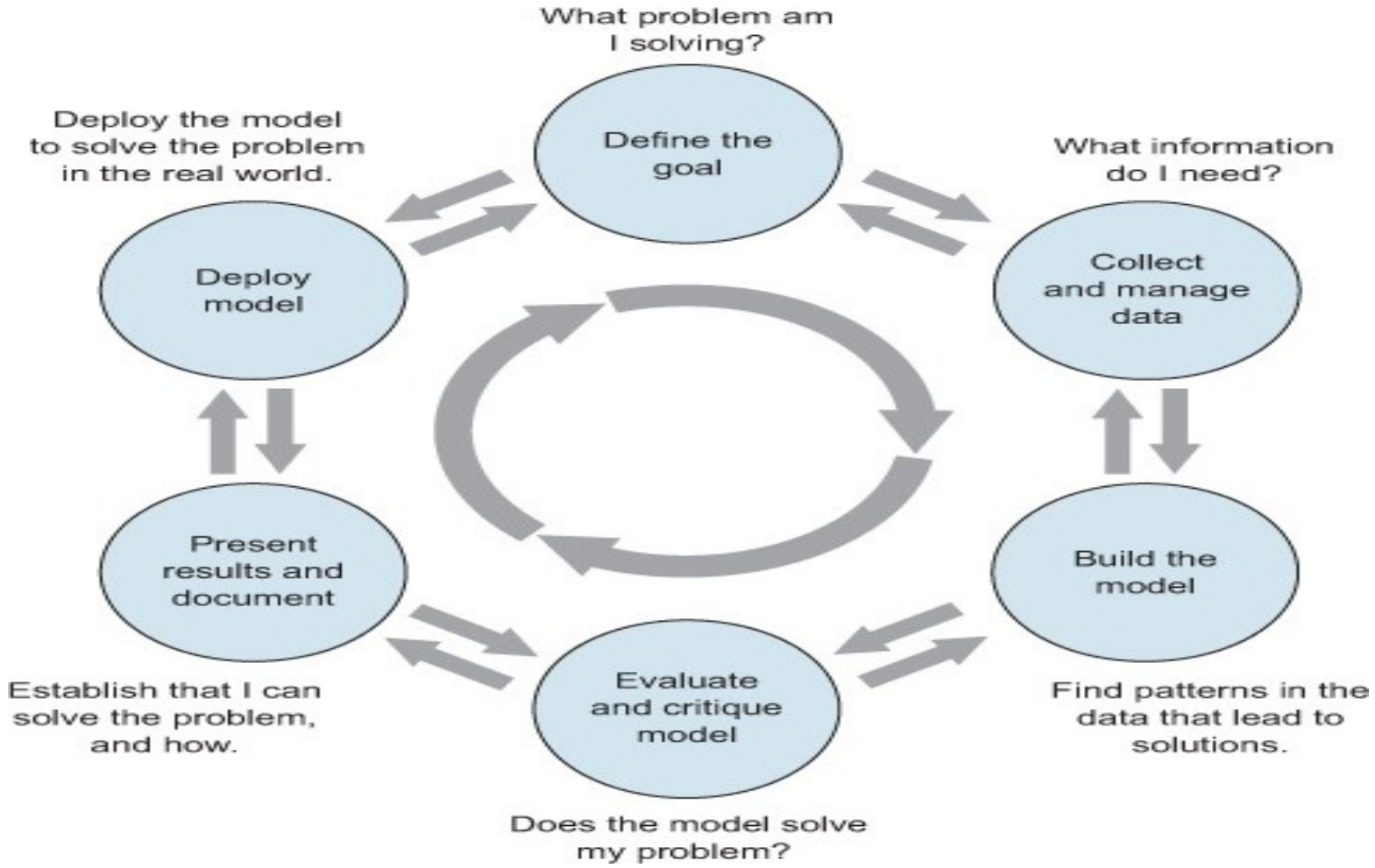
- **Role**: Data admin should ensure that the database is accessible to all relevant users. He also ensures that it is performing correctly and keeps it safe from hacking.
- **Languages**: Ruby on Rails, SQL, Java, C#, and Python

Business Analyst:

- **Role**: This professional needs to improve business processes. He/she is an intermediary between the business executive team and the IT department.
- **Languages**: SQL, Tableau, Power BI and, Python

STAGES OF A DATA SCIENCE PROJECT

- The ideal data science environment is one that encourages feedback and iteration between the data scientist and all other stakeholders. This is reflected in the lifecycle of a data science project.



Defining the goal

- The first task in a data science project is to define a measurable and quantifiable goal. At this stage, learn all that you can about the context of your project:
    1. Why do the sponsors want the project in the first place? What do they lack, and what do they need?
    2. What are they doing to solve the problem now, and why isn't that good enough?
    3. What resources will you need: what kind of data and how much staff? Will you have domain experts to collaborate with, and what are the computational resources?
    4. How do the project sponsors plan to deploy your results? What are the constraints that have to be met for successful deployment?

Data collection and management

- This step encompasses identifying the data you need, exploring it, and conditioning it to be suitable for analysis. This stage is often the most time-consuming step in the process. It's also one of the most important:

    1. What data is available to me?
    2. Will it help me solve the problem?
    3. Is it enough?
    4. Is the data quality good enough?

Status.of.existing.checking.account *(at time of application)*

Duration.in.month *(loan length)*

Credit.history

Purpose *(car loan, student loan, etc.)*

Credit.amount *(loan amount)*

Savings.Account.or.bonds *(balance/amount)*

Present.employment.since

Installment.rate.in.percentage.of.disposable.income

Personal.status.and.sex

Cosigners

Present.residence.since

Collateral *(car, property, etc.)*

Age.in.years

Other.installment.plans *(other loans/lines of credit—the type)*

Housing *(own, rent, etc.)*

Number.of.existing.credits.at.this.bank

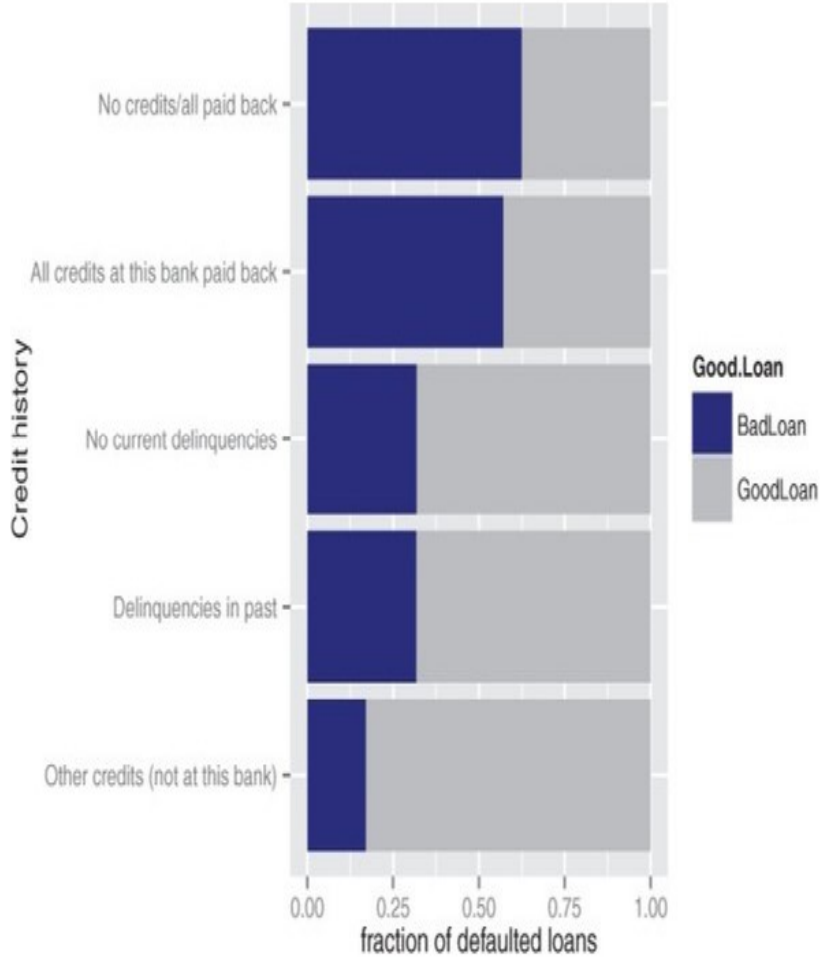
Job *(employment type)*

Number.of.dependents

Telephone *(do they have one)*

Good.Loan *(dependent variable)*

Figure 1.2. The fraction of defaulting loans by credit history category. The dark region of each bar represents the fraction of loans in that category that defaulted.



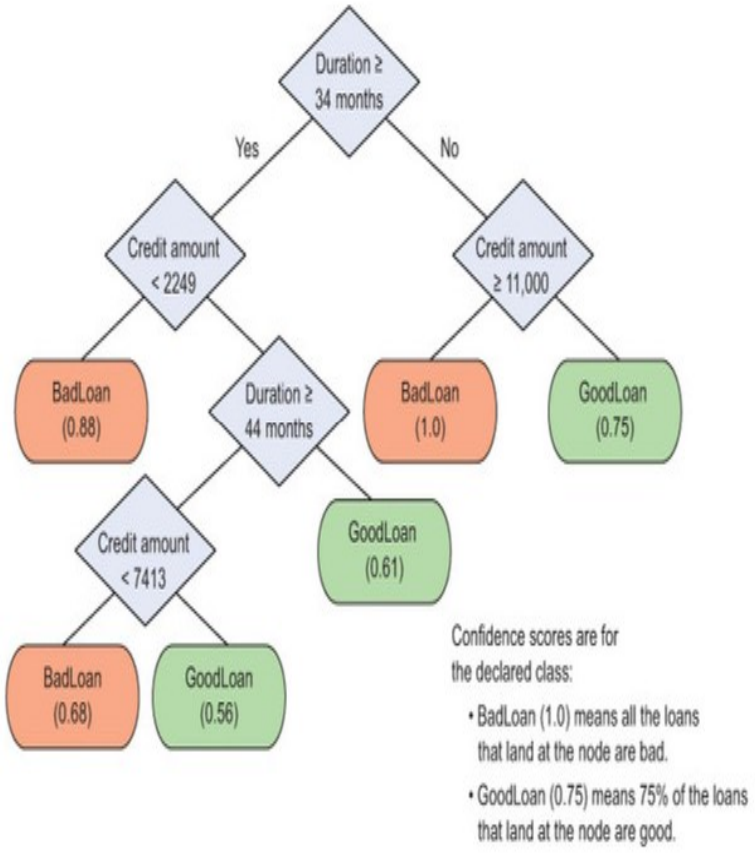
Modeling

- Finally get to statistics and machine learning during the modeling, or analysis, stage. Here is where you try to extract useful insights from the data in order to achieve your goals.
- The most common data science modeling tasks are these:
    1. ***Classification—*** *Deciding* if something belongs to one category or another
    2. ***Scoring—*** *Predicting* or *estimating* a numeric value, such as a price or probability
    3. ***Ranking—*** Learning to *order items* by preferences
    4. ***Clustering—*** *Grouping items* into most-similar groups
    5. ***Finding relations—*** *Finding correlations* or potential causes of effects seen in the data
    6. ***Characterization—*** Very general *plotting* and *report generation* from data

Listing 1.1. Building a decision tree

```
1  library('rpart')
2  load('GCDData.RData')
3  model <- rpart(Good.Loan ~
4    Duration.in.month +
5    Installment.rate.in.percentage.of.disposable.income +
6    Credit.amount  +
7    Other.installment.plans,
8    data=d,
9    control=rpart.control(maxdepth=4),
10   method="class")
```

Figure 1.3. A decision tree model for finding bad loan applications, with confidence scores



Model evaluation and critique

- Once you have a model, you need to determine if it meets your goals:
  1. Is it accurate enough for your needs? Does it generalize well?
  2. Does it perform better than "the obvious guess"? Better than whatever estimate you currently use?
  3. Do the results of the model (coefficients, clusters, rules) make sense in the context of the problem domain?

Listing 1.2. Plotting the confusion matrix

```
> resultframe <- data.frame(Good.Loan=creditdata$Good.Loan,
                            pred=predict(model, type="class"))
> rtab <- table(resultframe)
> rtab
          pred
Good.Loan  BadLoan GoodLoan
  BadLoan       41      259
  GoodLoan      13      687
> sum(diag(rtab))/sum(rtab)
[1] 0.728
> sum(rtab[1,1])/sum(rtab[,1])
[1] 0.7592593
> sum(rtab[1,1])/sum(rtab[1,])
[1] 0.1366667
> sum(rtab[2,1])/sum(rtab[2,])
[1] 0.01857143
```

Create the confusion matrix. Rows represent actual loan status; columns represent predicted loan status. The diagonal entries represent correct predictions.

Overall model accuracy: 73% of the predictions were correct.

Model precision: 76% of the applicants predicted as bad really did default.

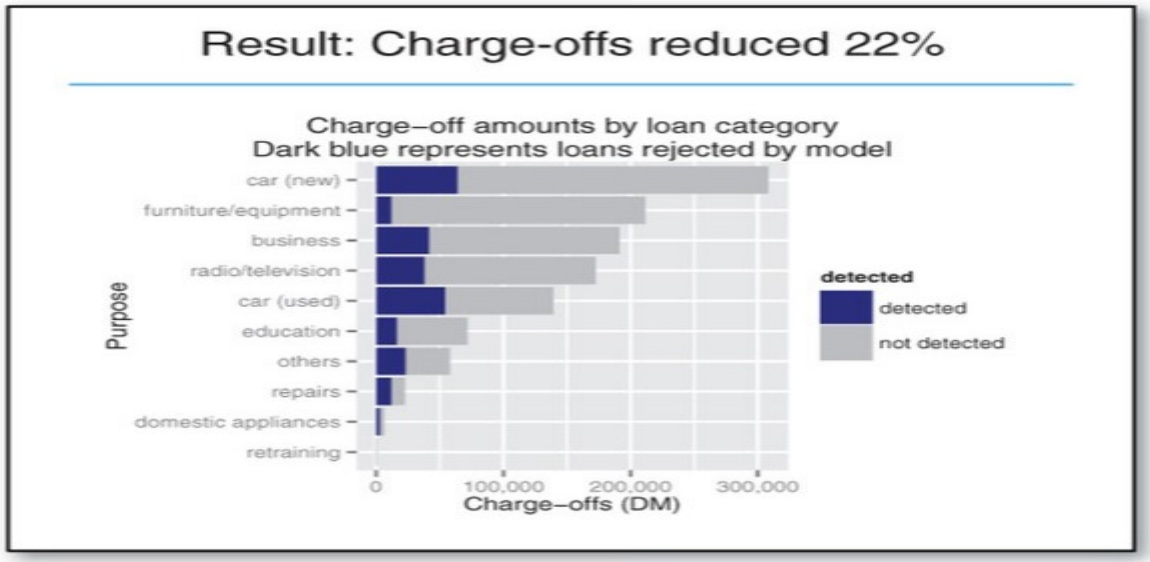
Model recall: the model found 14% of the defaulting loans.

False positive rate: 2% of the good applicants were mistakenly identified as bad.

Presentation and documentation

Once you have a model that meets your success criteria, you'll present your results to your project sponsor and other stakeholders. You must also document the model for those in the organization who are responsible for using, running, and maintaining the model once it has been deployed.

Figure 1.4. Notional slide from an executive presentation



Model deployment and maintenance

- Finally, the model is put into operation. In many organizations this means the data scientist no longer has primary responsibility for the day-to-day operation of the model.

DATA SCIENCE APPLICATION IN DIFFERENT AREAS

- The field of Data Science is filled with wonderful applications. In this modern era of digitization, Data Science is making a huge difference in making businesses successful.
- Not only in business but fields like healthcare, aeronautics, robotics, medicine, etc., Data Science is the game-changer. Here is the list of applications of Data Science that you will be introduced to in this blog:
    - Education
    - Airline Route Planning
    - Healthcare Industry
    - Delivery Logistics
    - Banking and Finance
    - Filtered Internet Search
    - Product Recommendation Systems
    - Digital Marketing and Advertising

Education

Data Science application in education, we can analyze the following:

- Students' need to improve their performance by analyzing the same based on factors such as specific books, study materials, or effective learning strategies
- The requirement of designing or updating the course curriculum as per the analysis of students' IQ levels and performance
- Performance of teachers based on student reviews, results of students, improvements by the weak students, etc.

Airline Route Planning

Data Science has helped airlines in the following ways:

- Identifying potential customers to offer calculated discounts, instead of providing discounts to everyone
- Deciding on the optimized routes by analyzing the traffic on different routes. It helps in saving expensive fuel that gets unnecessarily exhausted otherwise
- Predicting delays in flight
- Setting the cost of flights as per seasons, festivals, and the number of travelers. This is done by analyzing the number of potential travelers and frequent travelers
- This is how the application of Data Science is optimizing the profits of the airline industry.

Healthcare Industry

The areas where the **Data Science application in healthcare** is playing a major role are as follows:

- **Patient Diagnosis:** Data Science helps doctors monitor patients' health with the help of IoT devices. These wearable devices help monitor various medical conditions such as heart rate, body temperature, blood pressure, etc. These devices send patients' data to concerned doctors for medical analysis. This helps doctors take the necessary steps for treating patients accordingly.
- **Drug Research and Creation:** For creating a pharmaceutical drug, it takes a lot of research, time, and money. Also, there are millions of test cases required for research. Using Data Science, we can process all these test cases and make predictions on the success rate of these, based on certain parameters used for evaluating drugs in less time. With this application of Data Science, we can successfully create highly effective medicines.
- **Medical Image Analysis:** This is one of the interesting applications of Data Science, which is rapidly changing the way doctors do patient diagnoses. With the help of medical image analysis, a machine predicts diseases such as cancer, tumor, organ delineation, and many others.
- **Managing Patient Data:** Apart from the other applications of Data Science, it helps in managing patient data. The patient data is stored in databases and can be used in the future for the analysis of several medical conditions and the improvement of medical diagnosis and treatment

Delivery Logistics

- It helps in the analysis of profit generation, the causes of loss, the best route for delivery, the time required, and the scope for improvements.
- Other than that, the application of Data Science in delivery logistics helps the companies analyze the market trend and increase their competence. Further, with the help of route optimization, the number of deliveries increases and the freight cost reduces.

Banking and Finance

As we all know, the sector of banking and finance is prone to financial frauds and thefts. This happens due to the lack of proper analysis of customer data.

the application of Data Science in Finance helps in the following ways:

- **Stamping out Tax Fraud:** The economy of a nation depends on its taxpayers. Therefore, governments have started implementing Data Science to analyze citizen data to prevent tax fraud. With the application of Data Science, the income tax departments keep track of the income and calculate the tax. If the calculated tax is not collected, they track the suspicious taxpayers to take action against them.
- **Credit Scoring:** Credit scoring is another application of Data Science that helps check the financial civil score of an individual. The credit score is rated out of 10. It helps financial institutions make decisions on sanctioning loans. They decide the loan amount and its sanctioning on the basis of the credit score calculated out of 10.

Filtered Internet Search

- This is a type of filtered Internet search and also one of the wonderful applications of Data Science.
- It is only possible with the help of Data Science. Google collects and stores the data of search history to analyze and visualize it. Then, it uses algorithms and techniques that apply filters to the data to check the frequency of the searched keyword and related topics to show you the best results.

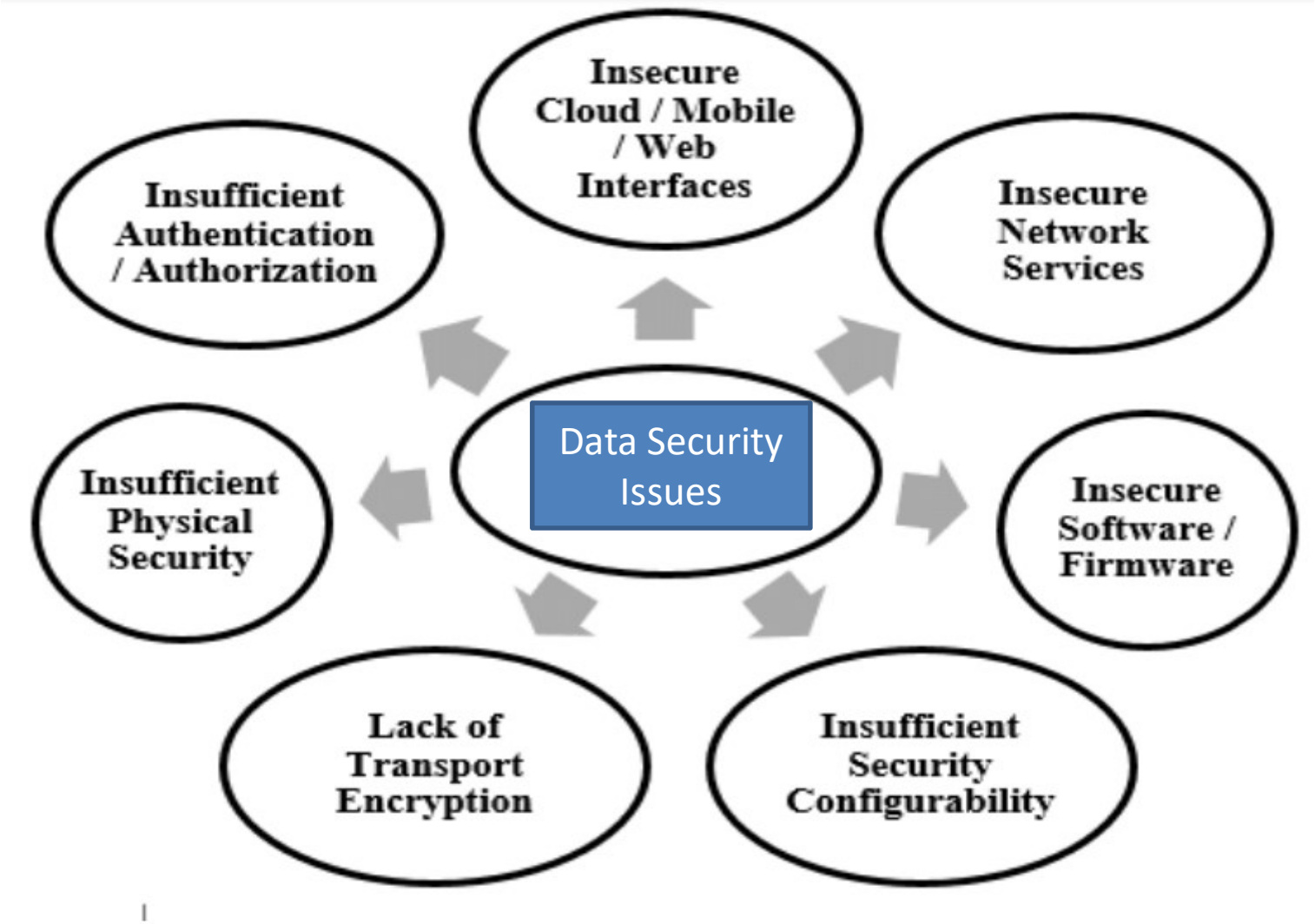
Product Recommendation Systems

- Product recommendation is an effective way of converting leads into sales.
- All the industries based on sales use recommendation systems for improving their profitability. But, how do these recommendation engines work? It is again a Data Science application.

Digital Marketing and Advertising

- The Data Science application in digital marketing helps organizations advertise their products to the right customers.
- Data Scientists design algorithms to analyze and visualize customers' data related to their search history, interests, and previously shopped items.

Data Security Issues:



The Data Security Issues

- Data Storage
- Fake Data
- Data Privacy
- Data Management
- Data Access Control
- Data Poisoning
- Employee Theft

Data Storage

- Businesses are adopting Cloud Data Storage to move their data easily to expedite business operations. However, the risks involved are exponential with security issues.
- While mission-critical information can be stored in on-premise databases, less sensitive data is kept in the cloud for ease of use Although it increases the cost of managing data in on-premise databases, companies must not take security risks for granted by storing every data in the cloud.

Fake Data

- Fake Data generation poses a severe threat to businesses as it consumes time that otherwise could be spent to identify or solve other pressing issues.
- There is more scope for leveraging inaccurate information on a very large scale, as assessing individual data points can be a daunting task for companies.

Data Privacy

- Data Privacy is a big challenge in this digital world.
- It aims to safeguard personal or sensitive information from cyberattacks, breaches, and intentional or unintentional data loss.
- The general rules are knowing your data, having more grip over your data stores and backup, safeguarding your network against unauthorized access, conducting regular risk assessments, and training the users regularly about Data Privacy and Data Security.

Data Management

- A security breach can have crushing consequences on businesses, including the vulnerability of critical business information to a completely compromised database.
- Deploying highly secured databases is vital to ensure data security at all levels.
- A superior Database Management System comes with various access controls.

Data Access Control

- Controlling which data users can view or edit enables companies to ensure not only data integrity but also preserves its privacy. But managing access control is not straightforward, especially in larger companies that have thousands of employees.

Data Poisoning

- Data Poisoning, a technique to attack Machine Learning models' training data.
- It can be considered as an integrity attack as the tampered training data can affect the model's ability to provide correct predictions.
- The results can be catastrophic, ranging from logic corruption to Data Manipulation and Data Injection.

Employee Theft

- The risk of an employee leaking sensitive information, intentionally or unintentionally, is high.
- Employee Theft is prevalent not only in big tech companies but also in startups.
- To avoid Employee Theft, companies have to implement legal policies along with securing the network with a virtual private network. In addition, companies can use a Desktop as a Service (DaaS) to eliminate the functionalities of data stored in local drives.