Introduction

Lecture 1

Centre for Data Science, ITER
Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India.

Contents

1. What Is Data Science?
2. Why Data Science is important?
3. Data Science Applications
4. Data Science Advantages and disadvantages
5. Difference between Data Mining and Data Science
6. The DataSciencester network

Data Science

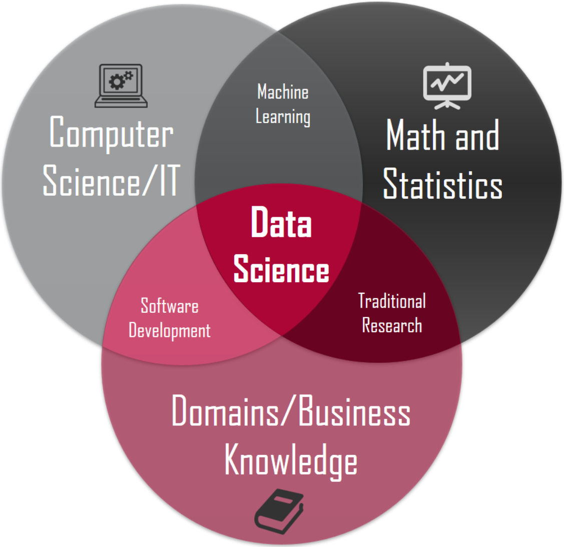


Figure 1: Data Science

What Is Data Science?

- Data science is an **interdisciplinary field** that uses **scientific methods**, processes, **algorithms** and systems to extract **knowledge and insights** from noisy, structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of **application domains.**

Why Data Science is important?

- Data is meaningless until its conversion into **valuable information.**
- Data Science involves mining large datasets containing structured and unstructured data and identifying hidden patterns to extract actionable insights.
- The importance of Data Science lies in its innumerable uses that range from daily activities like asking **Siri or Alexa** for recommendations to more complex applications like operating **a self-driving car**.

Data Science Life Cycle

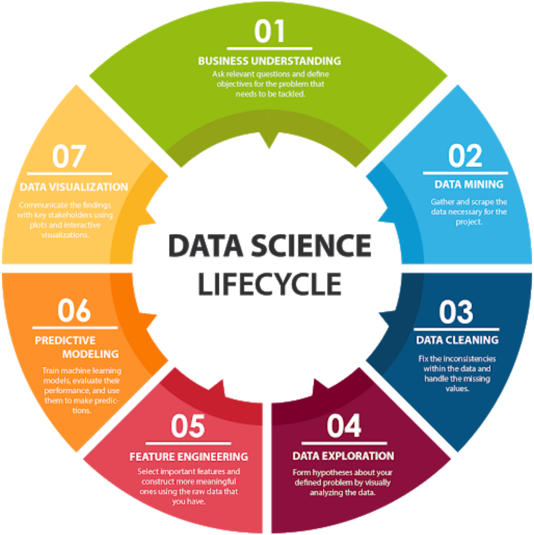

DATA SCIENCE LIFECYCLE
01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

Figure 2: Data Science Life Cycle

Data Science Applications

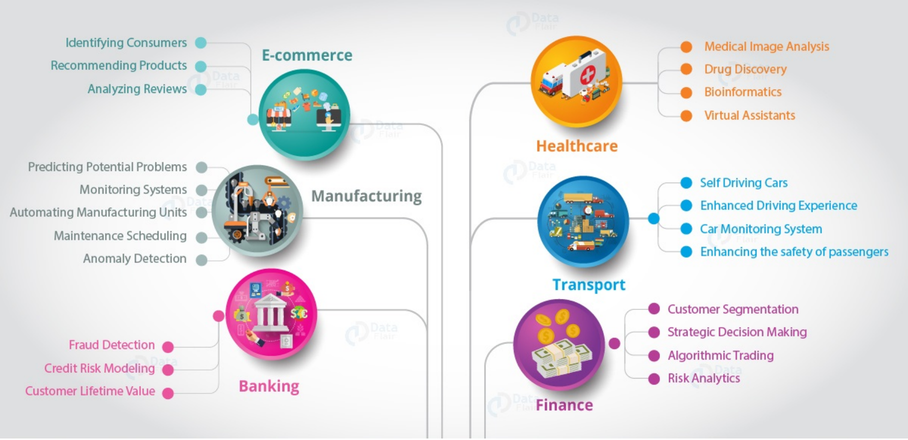


Figure 3: Data Science Application

Data Science Advantages

- Advantages of Data Science
  1. It’s in Demand
  2. Abundance of Positions
  3. A Highly Paid Career
  4. Data Science is Versatile
  5. Data Science Makes Data Better
  6. Data Scientists are Highly Prestigious
  7. No More Boring Tasks
  8. Data Science Makes Products Smarter

Data Science disadvantages

- Disadvantages of Data Science
  1. Data Science is Blurry Term
  2. Mastering Data Science is near to impossible
  3. A large Amount of Domain Knowledge Required
  4. Arbitrary Data May Yield Unexpected Results
  5. The problem of Data Privacy

Difference between Data Mining and Data Science

- **Meaning –**
  - Data Science is an inter-disciplinary field of computer science that uses a blend of tools, algorithms, and machine principles to extract usable information from data both structured and unstructured.
  - Data Mining can be described as the science of extracting useful information from large data sets or databases.

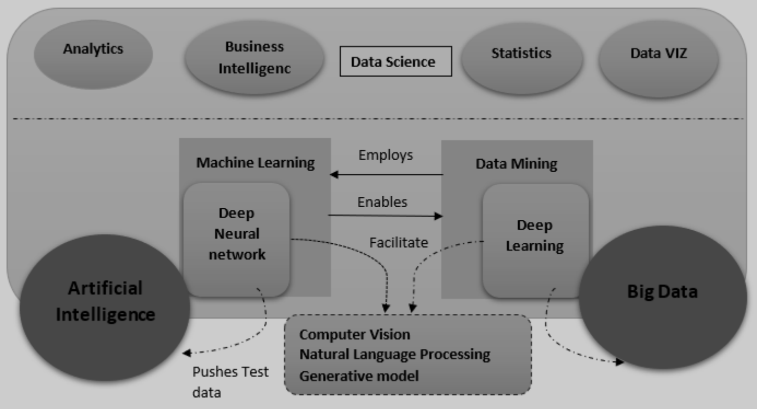
Difference between Data Mining and Data Science

- **Goal –**
  - The goal of data science is to utilize certain specialized computational methods to discover meaningful and useful information within a dataset in order to make important decisions.
  - The goal of data mining is to discover properties of existing data that were previously unknown and to find statistical rules or patterns from those data in order to solve complex computing problems.

Difference between Data Mining and Data Science

- **Field -**
  - Data Science is a multidisciplinary field that includes a number of related areas such as database systems, data engineering, data analysis, visualization, predictive modeling, experimentation, and business intelligence.
  - Data mining is all about uncovering valuable information from the tremendous amounts of data and to transform such data into organized knowledge.

Difference between data science, data analysis, data mining, machine learning, AI, and big data



Finding Key Connectors for a data set

- Example to identify who are the “**key connectors**” are among the given data scientists data set.
  It consists of a list of users, each represented by a dict that contains that user’s id and name.

```
users = [ {"id":0, "name":"Hero"}, {"id":1,
"name":"Dunn"}, {"id":2, "name":"Sue"}, {"id":3,
"name":"Chi"}, {"id":4, "name":"Thor"}, {"id":5,
"name":"Clive"}, {"id":6, "name":"Hicks"},
{"id":7, "name":"Devin"}, {"id":8, "name":"Kate"},
{"id":9, "name":"Klein"} ]
```

- Also the “friendship” data, represented as a list of pairs of IDs.

```
friendship_pairs = [(0, 1), (0, 2), (1, 2), (1,
3), (2, 3), (3, 4), (4, 5), (5, 6), (5, 7), (6,
8), (7, 8), (8, 9)]
```

- The tuple (0, 1) indicates that the data scientist with id 0 (Hero) and the data scientist with id 1 (Dunn) are friends.

The DataSciencester network

- The network is illustrated in Figure(4).

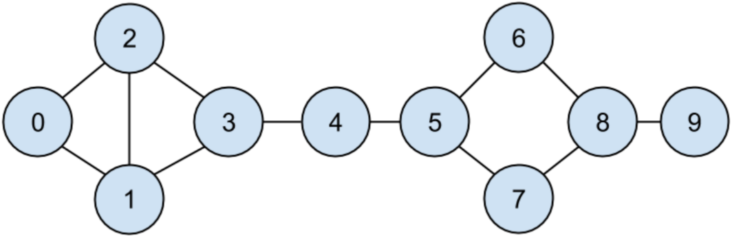


Figure 4: The DataSciencester network

The DataSciencester network

- Converting data set into dict.
- Then finding the total number of connections.
- After that sort them from “most friends” to “least friends”. Code for this is given below.

The DataSciencester network

```python
'''Initialize the dict with [] for each user id:'''
friendships = {user["id"]: [] for user in users}
'''And loop over the friendship pairs to populate it:'''
for i, j in friendship_pairs:
    friendships[i].append(j)
    friendships[j].append(i)

def no_of_friends(user):
    """How many friends does _user_ have?"""
    user_id = user["id"]
    friend_ids = friendships[user_id]
    return len(friend_ids)

total_connections = sum(no_of_friends(user) for user in users)
avg_connections = total_connections / len(users)

'''Create a list (user_id, number_of_friends).'''
no_friends_by_id = [(user["id"], no_of_friends(user)) for user in users]
no_friends_by_id.sort(key=lambda id_and_friends: id_and_friends[1], reverse=True)
```

The DataSciencester network

- This can be interpreted as a way of identifying people who are somehow central to the network.
- The network metric degree centrality is represented by the following figure(5).

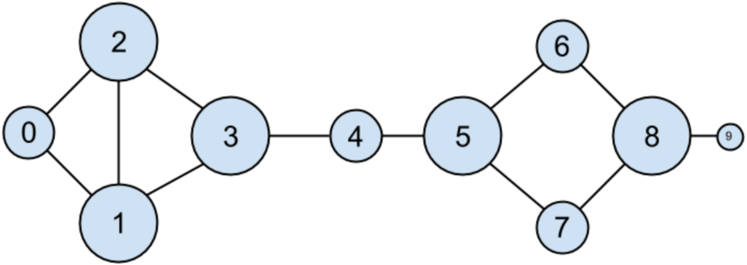


Figure 5: The DataSciencester network sized by degree

The DataSciencester network

- Code to iterate over their friends and collect the friends' friends.

```
def foaf_ids_bad(user):
  """foaf is short for "friend of a friend" """
  return [foaf_id
      for friend_id in friendships[user["id"]]
      for foaf_id in friendships[friend_id]]
```

The DataSciencester network

- Code to produce a count of mutual friends.

```
from collections import Counter
def friends_of_friends(user):
  user_id = user["id"]
  return Counter(foaf_id
    for friend_id in friendships[user_id]
    for foaf_id in friendships[friend_id]
    if foaf_id != user_id
    and foaf_id not in friendships[user_id])
print(friends_of_friends(users[3]))
'''Counter(0:  2, 5:  1)’’’
```

The DataSciencester network

- Data sets of user's and their interest are given below as a list of pairs (user_id, interest).

```
interests = [(0, "Hadoop"), (0, "Big Data"), (0, "HBase"), (0,
"Java"), (0, "Spark"), (0, "Storm"), (0, "Cassandra"), (1,
"NoSQL"), (1, "MongoDB"), (1, "Cassandra"), (1, "HBase"), (1,
"Postgres"), (2, "Python"), (2, "scikit-learn"), (2, "scipy"),
(2, "numpy"), (2, "statsmodels"), (2, "pandas"), (3, "R"), (3,
"Python"), (3, "statistics"), (3, "regression"), (3,
"probability"), (4, "machine learning"), (4, "regression"),
(4, "decision trees"), (4, "libsvm"), (5, "Python"), (5, "R"),
(5, "Java"), (5, "C++"), (5, "Haskell"), (5, "programming
languages"), (6, "statistics"), (6, "probability"), (6,
"mathematics"), (6, "theory"), (7, "machine learning"), (7,
"scikit-learn"), (7, "Mahout"), (7, "neural networks"), (8,
"neural networks"), (8, "deep learning"), (8, "Big Data"), (8,
"artificial intelligence"), (9, "Hadoop"), (9, "Java"), (9,
"MapReduce"), (9, "Big Data") ]
```

The DataSciencester network

- A function that finds users with a certain interest.

```
def data_scientists_who_like(target_interest):
  """Find the ids of all users who like the
  target interest."""
  return [user_id
    for user_id, user_interest in interests
    if user_interest == target_interest]
```

The DataSciencester network

- Building an index from interests to users and another from users to interests.
- Now finding out who has the most interests in common with a given user.
    - Iterate over the user's interests.
    - For each interest, iterate over the other users with that interest.
    - Keep count of how many times we see each other user.

The DataSciencester network

```python
from collections import defaultdict, Counter
''' Keys are interests, values are lists of user_ids with that interest '''
user_ids_by_interest = defaultdict(list)
for user_id, interest in interests:
    user_ids_by_interest[interest].append(user_id)

''' Keys are user_ids, values are lists of interests for that user_id.'''
interests_by_user_id = defaultdict(list)
for user_id, interest in interests:
    interests_by_user_id[user_id].append(interest)

def most_common_interests_with(user):
    return Counter(
        interested_user_id
        for interest in interests_by_user_id[user["id"]]
        for interested_user_id in user_ids_by_interest[interest]
        if interested_user_id != user["id"]
    )
```

References

[1] Data Science from Scratch Joel Grus, Shroff/O’reilly, Second Edition

Thank You
Any Questions?