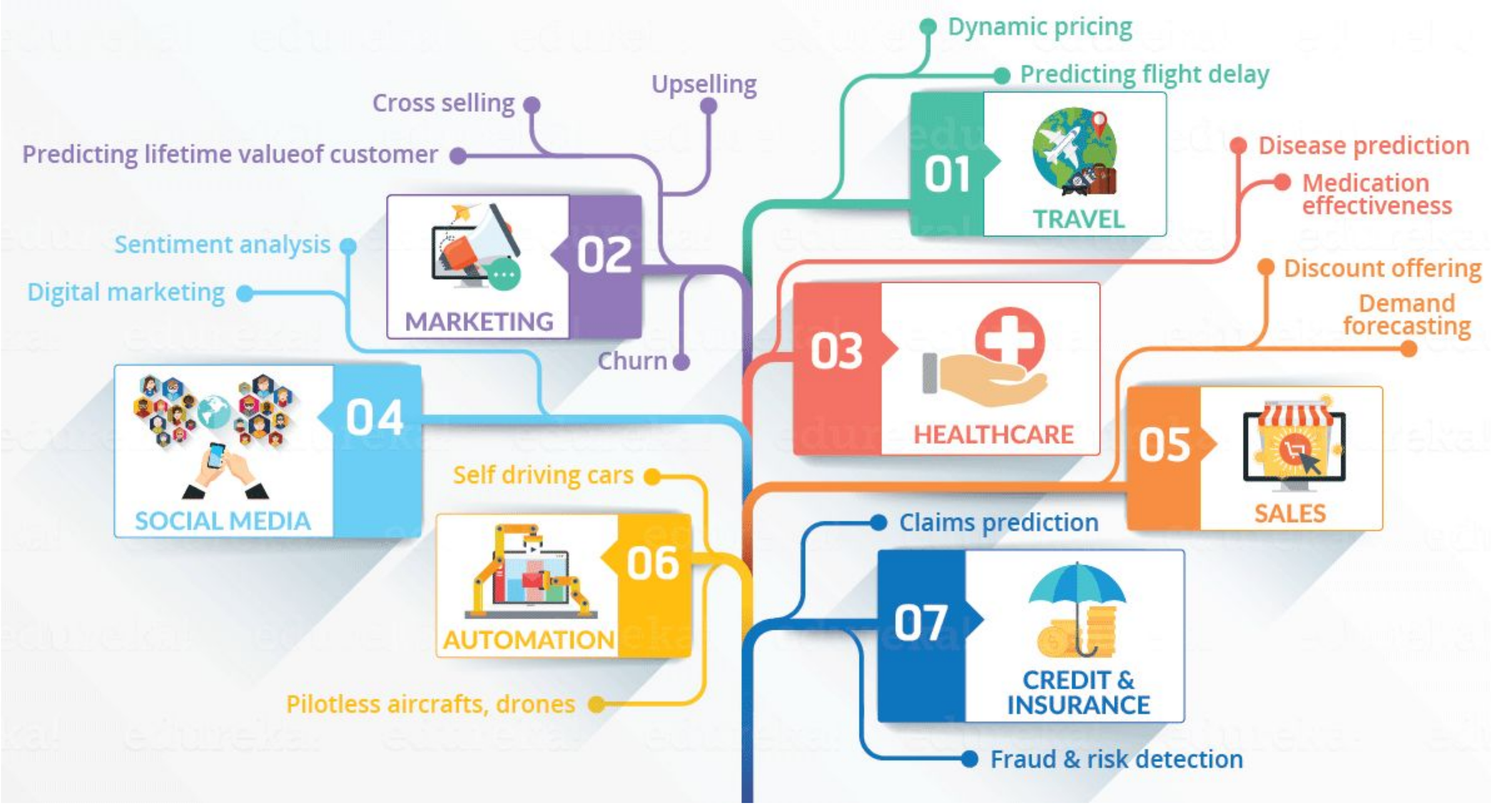
Data Science

InfPals 2018-2019

The need for Data Science

- Most of the data today is unstructured or semi-structured
- More complex and advanced analytical tools/algorithms are needed for processing, analyzing and drawing meaningful insights out of such data
- What if data could train models to make decisions / predictions etc. ?

Domains of Data Science



Cr: https://www.edureka.co/blog/what-is-data-science/

What is Data Science?

- A blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data

Components:

- Data exploration & analysis
- Data visualization
- Classical machine learning

etc.

Data Science = Statistics?

- Many debate the difference between Data Science and Statistics (can be read about in the Wikipedia page for Data Science.)

One argument:

- a Statistician usually explains what is going on by processing history of the data. On the other hand, Data Scientists not only does the exploratory analysis to discover insights from it, but also uses various advanced machine learning algorithms to identify the occurrence of a particular event in the future.

What You Will Be Doing In This Workshop

Activity One

- Installing (only on non-DICE)
- Trying out Jupyter Notebook (using Python)

Link: https://github.com/gwenty/INFPALS-data-science

Note: Jupyter Notebook is a web application in which you can create and share documents that contain live code, equations, visualizations as well as text, and is one of the tools used in data science.

Activity Two

Takes you to the next level and covers

- *numpy & scipy*
- *matplotlib*

Activity Three

This covers

- *pandas* - a library providing data structures and data analysis tools.