Bachelor of Science (Third Year)
CSD107- Introduction to Data Analytics
Title of the Unit : Data Science and Ethical Issues
Module Name : A look back at Data Science, Next-generation data scientists
Module Number : 23

A look back at Data Science

The following timeline traces the evolution of the term "Data Science" and its use, attempts to define it, and related terms.

- **1962** John W. Tukey writes in "The Future of Data Analysis": Data analysis is basically an observed science.
- **1974** Peter Naur publishes a book which is a survey of modern data processing methods, used in a wide range of applications.
- **1977** The International Association for Statistical Computing (IASC) established, to link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.
- **1989** Gregory Piatetsky-Shapiro organizes and chairs the first Knowledge Discovery in Databases (KDD) workshop.
- **1994** *BusinessWeek* publishes a cover story on "Database Marketing": "Companies are collecting mountains of information about you, crunching it to predict how likely you are to buy a product, and using that knowledge to craft a marketing message precisely regulated to get you to do so…
- **1996** Members of the International Federation of Classification Societies (IFCS) meet in Kobe, Japan, for their regular conference. For the first time, the term "data science" is included in the title of the conference.
- **1997** Professor C. F. Jeff Wu ,calls for "statistics" as "data science" and "statisticians" as "data scientists".
- **1997** The journal Data Mining and Knowledge Discovery launched.
- **2002** Launch of Data Science Journal.
- **2009** Troy Sadkowsky creates the data scientists group on LinkedIn.
- **2010** Kenneth Cukier writes in *The Economist* Special Report "Data, Data Everywhere": … a new kind of professional has emerged, the data scientist, who combines the skills of software programmer, statistician and storyteller/artist to extract the nuggets of gold hidden under mountains of data."
- **2012** Tom Davenport and D.J. Patil publish "Data Scientist: The Sexiest Job of the 21st Century" in the *Harvard Business Review*.
- **2019,** Advanced Analytics continued to build energy in the enterprise. Machine Learning (ML) dominated the analytics landscape.

Past 5 years or so

- Due to the shortage of data science skill sets, a lot of current generation of data scientists have migrated from other fields.
- Some of these fields have been a far away from the ancestors of the field of data science, namely computer science and applied statistics.
- We have data scientists coming from the Physical sciences, Social sciences, in addition to many MBA's.

- Skills brought to the table by people from these different fields are potentially inconsistent.
- Hence, there are many data-driven disciplines outside of data science.
- The training and technical knowledge of these practitioners are on the margins.

Next-generation data scientists

current generation of data scientists
- working tirelessly.
- satisfying the rushing demand for data-driven insights on behalf of industries.
- moving too quickly from a data set to applying a trendy algorithm and ignoring all the important steps in between.

It's natural to think and ask what the "data scientist 2.0" might look like in the next 5-10 years.

The Next-generation of Data scientist
- Should be much more progressed.
- Technically talented for a comfortable job with a life-changing salary.
- Should be encouraged to become good problem solvers who follow the scientific method.
- To think deeply about the appropriate use of the data science process.
- To use data responsibly and for the common good.

Characteristics of Next-generation Data scientist

1)Technical Skills Fully Improved
- Will maintain a breadth of hard technical skills such as mathematics, statistics, probability theory, machine learning, coding, data visualization and data storytelling.
- Coding is important, so good coding practices like agile software development techniques, code reviews, debugging, and version control are mostly valuable.
- Steps in the data science process need to be fully cultivated such as: exploratory data analysis (EDA), creative feature engineering, managing the vast number of models (optimization methods, evaluation metrics, etc.), data transformations (polynomial, log, binary categorical variables).

2) Slow down and Proceed methodically
- Slow down and think.
- It's all too easy to quickly run a piece of code that makes predictions and then declare success when the algorithm converges.
- The more difficult part is proceeding with careful consideration and making sure the results are correct and interpretable.
- Shouldn't try to impress with complex learning models that don't work that well.
- Which are mismatched with the problem being solved.
- Spending more time getting the data into shape.
- Don't be embarrassed to admit you spend 80% of your time making sure the data is good.
- There is a sincere need to be open, accepting, flexible, and interdisciplinary.

3) Soft Skills are King

- To cultivate good habits and remain open to continuous learning.
- A few good habits include: Persistence, Thinking flexibly, Thinking about thinking, and Striving for accuracy.
- Try not to over or underestimate your abilities.
- Give yourself reality checks by making sure you can code what you speak.
- Interact with other data scientists about methods and approaches.

4) Apply the Scientific Method

- Should assign to the “scientific method” to test hypotheses, welcome challenges and alternative theories.
- Meaning finding holes in ideas, and devising tests as true scientists.
- It’s also important to ask a lot of questions.
- Adopt the viewpoint of inborn curiosity, and don’t worry about appearing stupid.
- Don’t be afraid to ask for clarification.
- Remain skeptical about the statistical models being used in terms of how they may fail.
- Implications and consequences of the models which are built.

5) Proceed with Ethics

- The data generated by user behavior is the building blocks of data-driven products,.
- It’s important to realize that algorithms are not only capable of predicting the future, but also of directing the future.
- Next generation data scientists shouldn’t let their salaries blind them so that their models are used for unethical purposes.
- Instead, they should seek out opportunities to solve problems of social value .
- Consider the impact and consequences of their models.