Introduction to Data Science

Unit - I

Faculty: D Sai Kumar

Dept. of CSE, UCE, OU

Introduction

Let’s Start by understanding what is Data Science.

1. Have you ever wondered how amazon, ebay suggest items for you to buy.
2. How gmail filters your emails in spam and non spam categories.
3. How Netflix predicts the shows of your liking.

How do they do it???

- In reality doing such tasks are impossible without the availability of **Data.**
- Data Science is all about using Data to solve problems.

Data Science

- Data Science is defined as the study of data, where it comes from, what it represents and the way by which it can be transformed into valuable inputs and resources to create business and IT strategies.

(or)

- Data Science is the study of data, it involves developing methods of recording, storing and analysing data to efficiently extract useful information.
- The goal of Data Science is to gain insights and knowledge from any type of data both structured or unstructured.

- Data Science is a field of study that combines domain expertise, programming skills and knowledge of mathematics and statistics to extract meaningful insights from data.

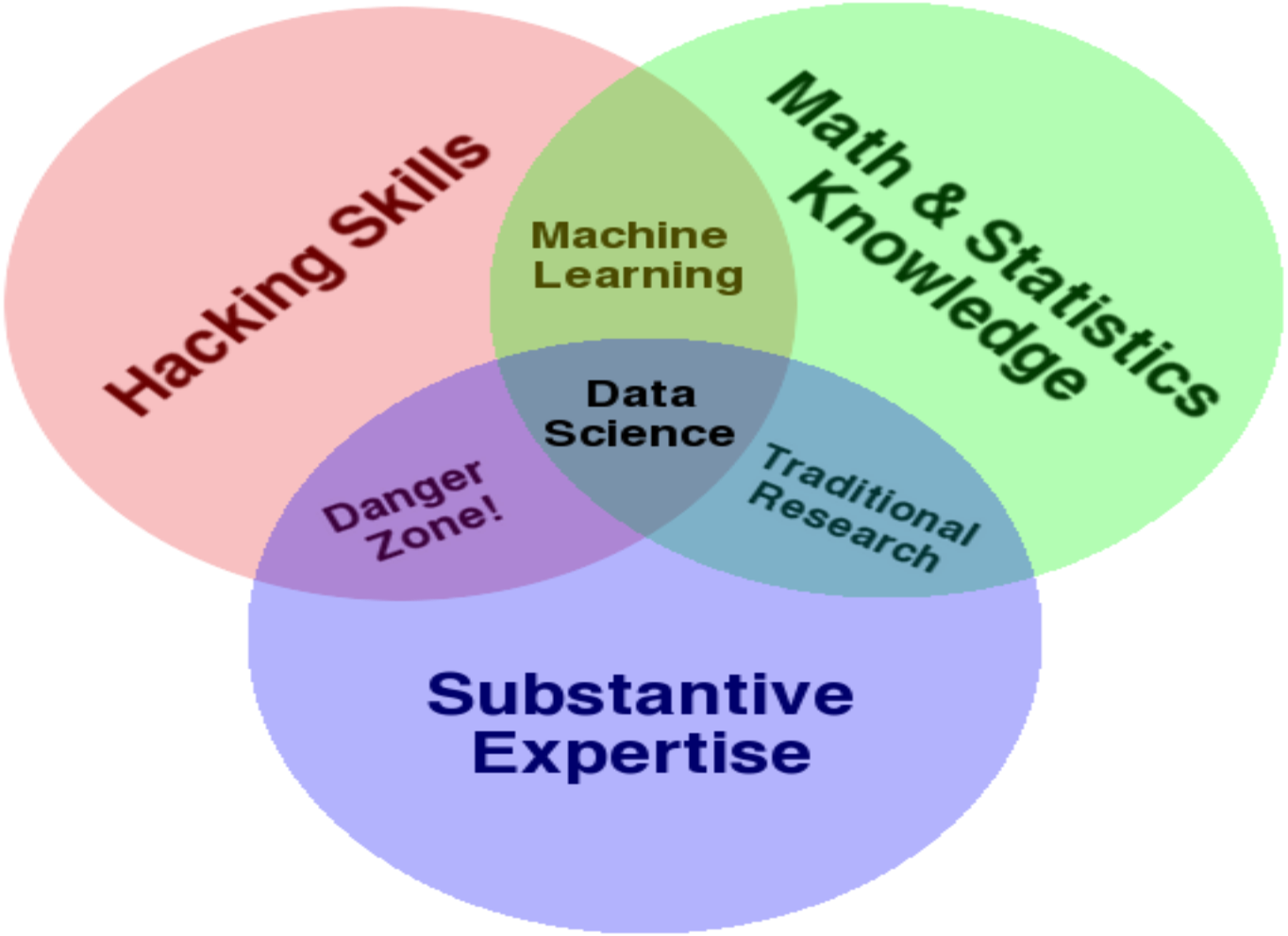
(or)

- Data Science is a multidisciplinary field of scientific data, Algorithms, Systems, Processes in order to extract insights from heterogeneous huge amount of data.

Understanding data science begins with three basic areas:

- **Math/statistics**: use of equations and formulas to perform analysis
- **Computer programming**: ability to use code to create outcomes on the computer
- **Domain knowledge**: This refers to understanding the problem domain (medicine, finance, social science, and so on)

Venn Diagram of Data Science



What is Data?

- Webster's defines **data** as a plural form of **datum** as "something given or admitted especially as a basis for reasoning or inference."
- For example, imagine a table containing birthdays of everyone in your class or office. We can consider this whole table (a collection of birthdays) as data. Each birthday is a single point of data, which could be called datum, but we will call that data too.
- There is also often a debate about what is the difference between data and information. In fact, it is common to use one to define the other (e.g., "data is a piece of information").
- According to the Oxford dictionary, **science** is "systematic study of the structure and behaviour of the physical and natural world through observation and experiment."

Measurement of Data

- Bit(**Bi**nary Digi**t**) : A bit is a value of either a 1 or 0 (on or off).
- Nibble : A Nibble is 4 bits.
- Byte : a Byte is 8 bits. 1 character, e.g. "a", is one byte.
- Kilobyte (KB) : 1024 bytes
- Megabyte (MB) : 1024 KB
- Gigabyte (GB) : 1024 MB(2^30 bytes)
- Terabyte (TB) : 1024 GB (2^40 bytes)
- Petabyte (PB) :1024 TB (2^50 bytes)
- Exabyte (EB) : 1024 PB (2^60 bytes)
- Zettabyte (ZB) : 1024 EB (2^70 bytes)
- Yottabyte (YB) : 1024 ZB (2^80 bytes)

5 Stages of Data

1. **Capture** : Data Acquisition, Extraction
2. **Maintain**: Data Warehousing, Data Cleaning, Data staging, Data Processing, Data Architecture
3. **Process**: Data Mining, Classification, modelling
4. **Analyse**: Predictive analysis, Regression techniques, Qualitative analysis
5. **Communicate**: Report generation, Visualization, Business Intelligence, Decision making

Why Data Science is Important now?

- The number of Job openings for Data scientists are gradually increased from its past. Both industry and Academia has increased the demand for data science and data scientists.
- The answer is we have lot of data that is coming up or accumulated in every second with great Velocity, Volume and Variety.

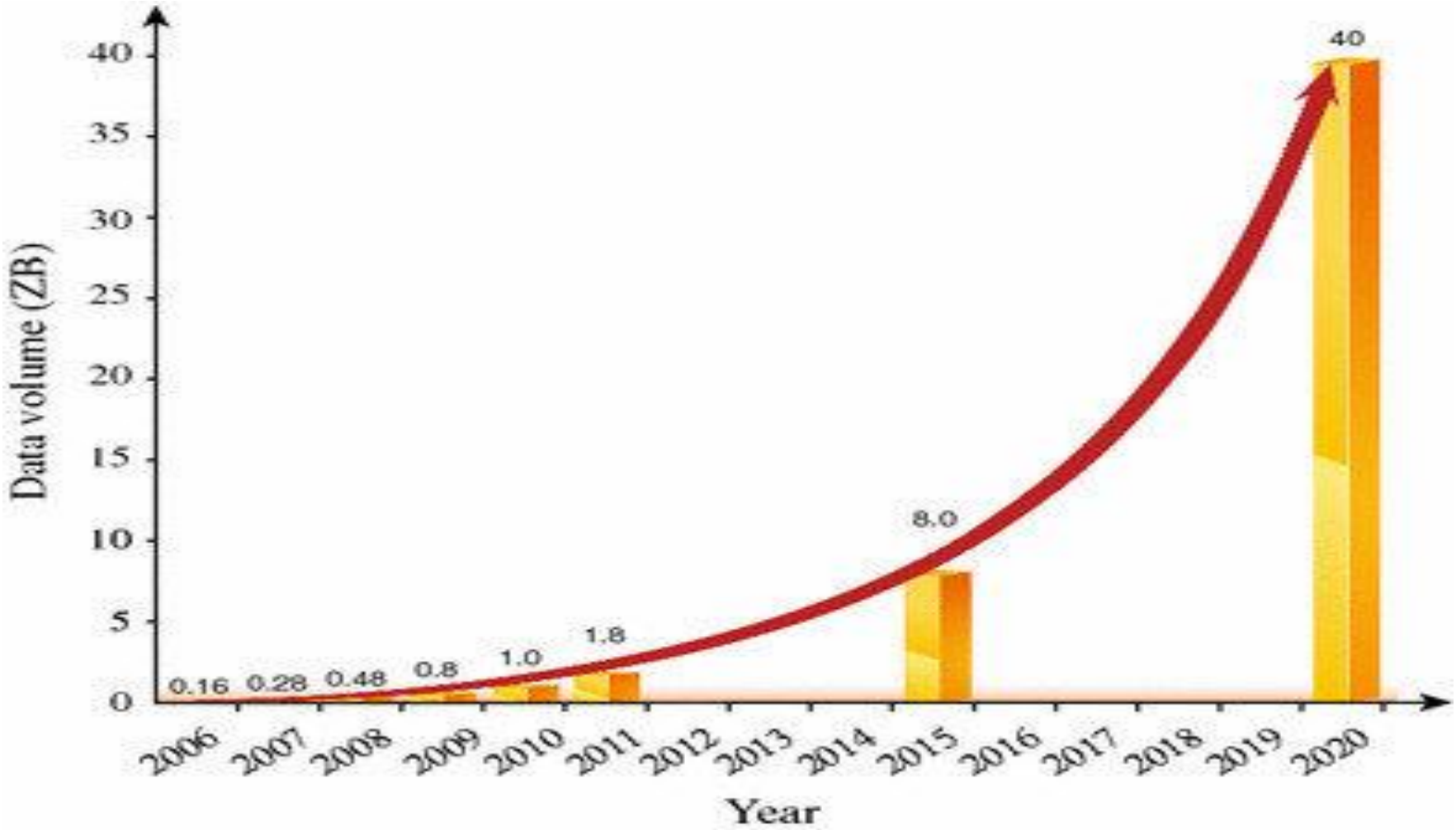
These are the three Vs:

1. Velocity: The speed at which data is accumulated.
2. Volume: The size and scope of the data.
3. Variety: The massive array of data and types (structured and unstructured).

Where do we see Data Science?

- The question should be: Where do we not see data science these days?
- The great thing about data science is that it is not limited to one facet of society, one domain, or one department of a university; it is virtually everywhere.
- There are number of applications where data science can be used:

1. It helps in getting ideas of what customers would love to Purchase or eat according to their previous browsing history/purchase history.
2. Data Science also helps in making future predictions.
3. Data Science also helps in getting recommendations.



Increase of data volume in last 15 years. (Source: IDC's Digital Universe Study, December 2012.5)

Finance

- What do financial data scientists do?
- Through capturing and analyzing new sources of data, building predictive models and running real-time simulations of market events, they help the finance industry obtain the information necessary to make accurate predictions.
- Data scientists in the financial sector may also partake in fraud detection and risk reduction.
- Banks and other loan sanctioning institutions collect a lot of data about the borrower in the initial “paperwork” process.
- Data science practices can minimize the chance of loan defaults via information such as customer profiling, past expenditures, and other essential variables that can be used to analyze the probabilities of risk and default.
- Data Science is used in financial institutions to identify the creditworthiness of potential customers.

Public Policy

- public policy is the application of policies, regulations, and laws to the problems of society through the actions of government and agencies for the good of a citizenry.
- Many branches of social sciences (economics, political science, sociology, etc.) are foundational to the creation of public policy.
- Data science helps governments and agencies gain insights into citizen behaviors that affect the quality of public life, including traffic, public transportation, social welfare, community wellbeing, etc.
- This information, or data, can be used to develop plans that address the betterment of these areas.

Politics

- Politics is a broad term for the process of electing officials who exercise the policies that govern a state.
- Data scientists analyzed former US President Obama’s 2008 presidential campaign success with Internet-based campaign efforts.
- Data scientists have been quite successful in constructing the most accurate voter targeting models and increasing voter participation.
- In 2016, the campaign to elect Donald Trump was a brilliant example of the use of data science in social media to tailor individual messages to individual people.
- The data analytics firm obtained data on approximately 87 million Facebook users from an academic researcher in order to target political ads during the 2016 US presidential campaign.

Healthcare

- Healthcare is another area in which data scientists keep changing their research approach and practices.
- Medical Imaging
- Genemoics
- Drug Discovery
- Predictive Analysis
- Monitoring patients health :wearable devices
- Tracking and preventing diseases
- Providing virtual assistance

Urban Planning

- Many scientists and engineers have come to believe that the field of urban planning is ripe for a significant – and possibly disruptive – change in approach as a result of the new methods of data science.

Education

- Schools, Colleges, Universities have large amount of student data such as academic records, grades, results, personal interests, cultural interests etc to handle. The analysis of this data can help them in finding advanced methods for improving/enhancing the student learning.
- Improve adaptive learning
- Better parent involvement
- Better Assessment of Teachers
- Improve student performance
- Better Organization
- Regular updates in curriculum
- Student recruitment

Libraries

- Data science is also frequently applied to libraries.
- Jeffrey M. Stanton has discussed the overlap between the task of a data science professional and that of a librarian. In his article, he concludes, “In the near future, the ability to fulfill the roles of citizenship will require finding, joining, examining, analyzing, and understanding diverse sources of data […] Who but a librarian will stand ready to give the assistance needed, to make the resources accessible, and to provide a venue for knowledge creation when the community advocate arrives seeking answers?” .

How Does Data Science Relate to Other Fields?

Data Science and Statistics:

- The term “data science” meant nothing to most people, A common response to the term was “Isn’t that just statistics”. The difference between the fields lies in the invention and advancements in modern computers.
- Statistics was primarily developed to help people deal with pre-computer “data problems,” such as testing the impact of fertilizer in agriculture, or figuring out the accuracy of an estimate from a small sample.
- Data science emphasizes the data problems of the twenty-first century, such as accessing information from large databases, writing computer code to manipulate data, and visualizing data.

Data Science and Computer Science:

-Computer science is the study of computers and computational systems.

- Computer scientists have developed numerous techniques and methods, such as

(1) database(DB) systems that can handle the increasing volume of data in both structured and unstructured formats, expediting data analysis;

(2) visualization techniques that help people make sense of data; and

(3) algorithms that make it possible to compute complex and heterogeneous data in less time.

- In truth, data science and computer science overlap and are mutually supportive. Some of the algorithms and techniques developed in the computer science field – such as machine learning algorithms, pattern recognition algorithms, and data visualization techniques –have contributed to the data science discipline.

- **Data Science and Engineering:**
- Broadly speaking, engineering in various fields (chemical, civil, computer, mechanical, etc.) has created demand for data scientists and data science methods.
- Engineers constantly need data to solve problems. Data scientists have been called upon to develop methods and techniques to meet these needs.
- Likewise, engineers have assisted data scientists.
- Data science has benefitted from new software and hardware developed via engineering, such as the CPU (central processing unit) and GPU (graphic processing unit) that substantially reduce computing time.

Data Science and Business Analytics:

- In general, we can say that the main goal of “doing business” is turning a profit – even with limited resources – through efficient and sustainable manufacturing methods, and effective service models, etc.
- This demands decision-making based on objective evaluation, for which data analysis is essential.

- **Business analytics** (BA) refers to the skills, technologies, and practices for continuous iterative exploration and investigation of past and current business performance to gain insight and be strategic.

There are four types of analytics, each of which holds opportunities for data scientists in business analytics:

1. Decision analytics: supports decision-making with visual analytics that reflect reasoning.

2. Descriptive analytics: provides insight from historical data with reporting, score cards,

clustering, etc.

3. Predictive analytics: employs predictive modeling using statistical and machine learning

techniques.

4. Prescriptive analytics: recommends decisions using optimization, simulation, etc.

Data Science, Social Science, and Computational Social Science:

- Data science helping social science, but it is also shaping it, even creating a new branch called computational social science.
- social science has spread into many branches, including but not limited to anthropology, archaeology, economics, linguistics, political science, psychology, public health, and sociology.
- Each of these branches has established its own standards, procedures, and modes of collecting data over the years. But connecting theories or results from one discipline to another has become increasingly difficult. This is where computational social science has revolutionized social science research in the last few decades.
- With the help of data science, computational social science has connected results from multiple disciplines to explore the key urgent question:

How will the information revolution in this digital age transform society?

The Relationship between Data Science and Information Science

- Information science provides a complementary approach that emphasizes the context in which data is generated, accessed, and used.
- The field of information science, which often stems from computing, computational science, informatics, information technology, or library science, often represents and serves such application areas.
- The core idea here is to cover people studying, accessing, using, and producing information in various contexts.

The Universality of Data : data is everywhere

Information vs. Data

- Depending on who you consult, you will get different answers – from seeming differences to a blurred-out line between data and information.
- To make matters worse, people often use one to mean the other.
- A traditional view used to be that data is something raw, meaningless, an object that, when analyzed or converted to a useful form, becomes information.
- Information is also defined as “data that are endowed with meaning and purpose”.
- For example, the number “480,000” is a data point. But when we add an explanation that it represents the number of deaths per year in the USA from cigarette smoking, it becomes information.

- The Data, Information, Knowledge, and Wisdom (DIKW) model differentiates the meaning of each concept and suggests a hierarchical system among them.
- Although various authors and scholars offer several interpretations of this model, the model defines data as (1) fact, (2) signal, and (3) symbol. Here, information is differentiated from data in that it is “useful.”

Users in Information Science

- Studies in information science have focused on the human side of data and information, in addition to the system perspective.
- While the system perspective typically supports users' ability to observe, analyze, and interpret the data, the former allows them to make the data into useful information for their purposes.
- **Usefulness** is a criterion that determines how useful is the interaction between the user and the information object (data) in accomplishing the task or goal of the user.
- Scholars in information science tend to combine the user side and the system side to understand how and why data is generated and the information they convey, given a context.

Data Science in Information Schools (iSchools)

- There are several advantages to studying data science in information schools, or iSchools.
- Data science provides students a more refined understanding of individual, community, and society-wide phenomena.
- An iSchool curriculum helps students acquire diverse perspectives on data and information.

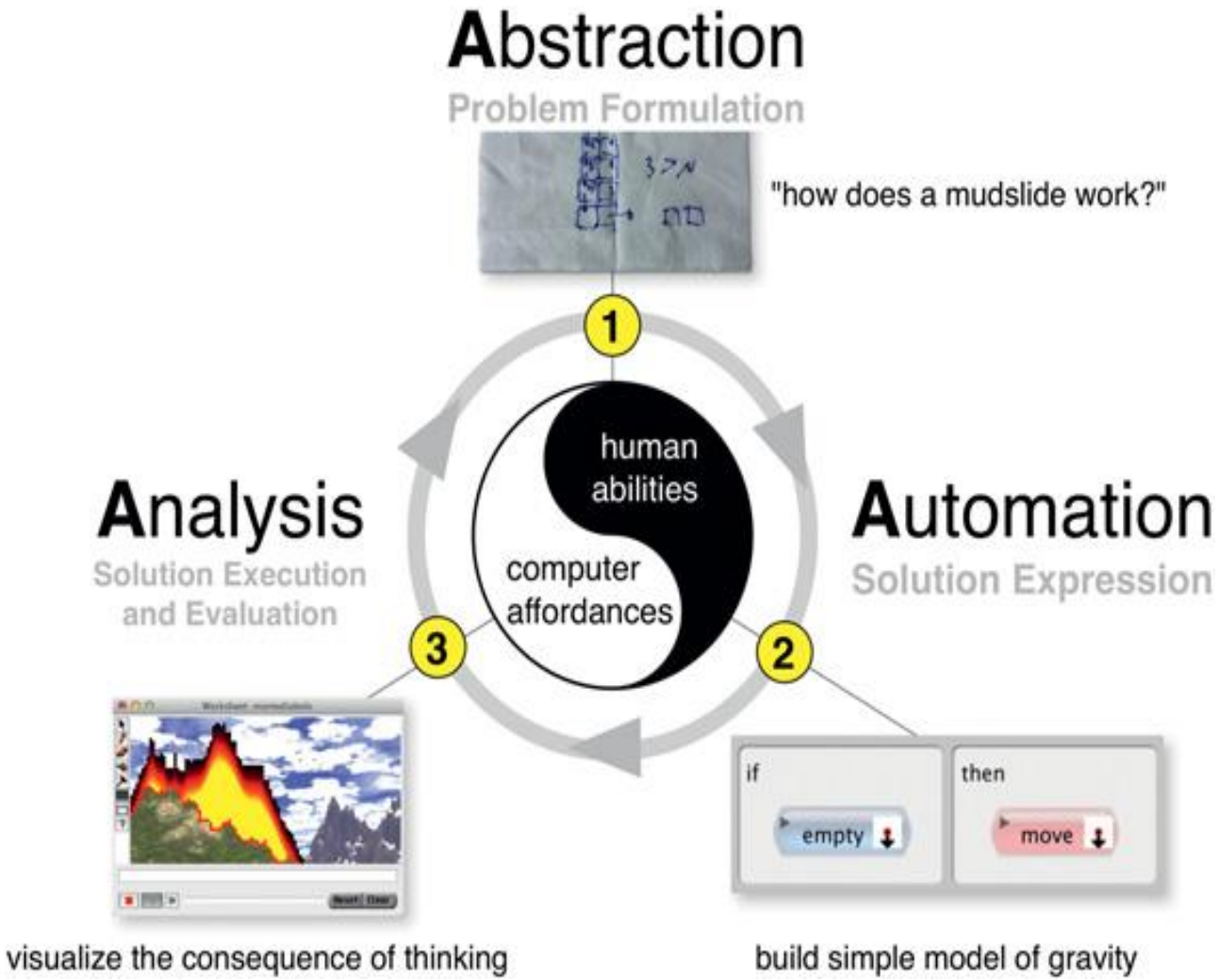
Computational Thinking

- Many skills are considered “basic” for everyone. These include reading, writing, and thinking. It does not matter what gender, profession, or discipline one belongs to; one should have all these abilities.
- In today’s world, computational thinking is becoming an essential skill, not reserved for computer scientists only.
- Computational thinking means thinking like a computer scientist.

According to Jeannette Wing,

“Computational thinking is using abstraction and decomposition when attacking a large complex task or designing a large complex system”

Computational Thinking



Three-stage process describing computational thinking.

It is an iterative process based on the following three stages:

1. Problem formulation (abstraction)
2. Solution expression (automation)
3. Solution execution and evaluation (analyses).

Example:

- Let us consider an example. We are given the following numbers and are tasked with finding the largest of them: 7, 24, 62, 11, 4, 39, 42, 5, 97, 54. Perhaps you can do it just by looking at it. But let us try doing it “systematically.”

Skills for Data Science

- Let us look at carefully what data scientists are, what they do, and what kinds of skills one may need to make their way in and through this field.

- One Twitter quip about data scientists captures their skill set particularly well:

“Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.”

- In her Harvard Business Review article, noted academic and business executive Jeanne Harris listed some skills that employers expect from data scientists: **willing to experiment, proficiency in mathematical reasoning, and data literacy**.

- In another view, Dave Holtz blogs about specific skill sets desired by various positions to which a data scientist may apply. He lists basic types of data science jobs:

1. A Data Scientist Is a Data Analyst Who Lives in San Francisco!
2. Please Wrangle Our Data!
3. We Are Data. Data Is Us.
4. Reasonably Sized Non-Data Companies Who Are Data-Driven

Hands-On Example : Analyzing Data

- what kinds of things people do as a data scientist.
- Consider a data-driven problem, identify a data source, collect data, clean the data, analyze the data, and present our findings.
- Dataset of average heights and weights for American women.(CSV file)

Tools for Data Science

- A lot of what data scientists do involves processing data and deriving insights.
- Develop a solid foundation in statistical techniques and computational thinking.
- A couple of programing and data processing tools, noting that there are no special tools for doing data science; there just happen to be some tools that are more suitable for the kind of things one does in data science.
- If you already know some programing language (e.g., C, Java, PHP) or a scientific data processing environment (e.g., Matlab), you could use them to solve many or most of the problems and tasks in data science.

JAVA vs Python

Let us see this with an example. If you want to write the classic “Hello,World” program in Java, here is how it goes:

Step 1: Write the code and save as HellowWorld.java.

```
public class HelloWorld {

public static void main(String[] args) {

System.out.println(“Hello, World”);

}

}
```

Step 2: Compile the code.

```
% javac HelloWorld.java
```

Step 3: Run the program.

```
% java HelloWorld
```

In contrast, here is how you do the same in Python:

Step 1: Write the code and save as hello.py

```
print(“Hello, World”)
```

Step 2: Run the program.

```
% python hello.py
```

Issues of Ethics, Bias, and Privacy in Data Science

- Many of the issues related to privacy, bias, and ethics can be traced back to the origin of the data.
- Ask – how, where, and why was the data collected? Who collected it? What did they intend to use it for?
- More important, if the data was collected from people, did these people know that: (1) such data was being collected about them; and (2) how the data would be used?
- For instance, just because data on a social media service such as Twitter is available on the Web, it does not mean that one could collect and sell it for material gain without the consent of the users of that service.
- In April 2018, a case surfaced that a data analytics firm, Cambridge Analytica, obtained data about a large number of Facebook users to use for political campaigning. Those Facebook users did not even know that: (1) such data about them was collected and shared by Facebook to third parties; and (2) the data was used to target political ads to them

Data

Introduction

- Just as **trees** are the raw material from which **paper** is produced, so too, can **data** be viewed as the raw material from which **information** is obtained.
- To present and interpret information, one must start with a process of gathering and sorting data.

Data Types

- One of the most basic ways to think about data is whether it is structured or not.
- This is especially important for data science because most of the techniques that we will learn depend on one or the other inherent characteristic.
- **Structured data** refers to highly organized information that can be seamlessly included in a database and readily searched via simple search operations.
- **Unstructured data** is essentially the opposite, devoid of any underlying structure.
- In structured data, different values – whether they are numbers or something else – are fields or labeled, which is not the case when it comes to unstructured data.

Structured Data

custid	sex	is.employed	income	marital.stat	housing.type	num.vehicles	age	state.of.res
2068	F	NA	11300	Married	Homeowner free and clear	2	49	Michigan
2073	F	NA	0	Married	Rented	3	40	Florida
2848	M	True	4500	Never married	Rented	3	22	Georgia
5641	M	True	20000	Never married	Occupied with no rent	0	22	New Mexico
6369	F	True	12000	Never married	Rented	1	31	Florida

Unstructured Data

- Unstructured data is data without labels.
- Here is an example:
- “It was found that a female with a height between 65 inches and 67 inches had an IQ of125–130. However, it was not clear looking at a person shorter or taller than this observation if the change in IQ score could be different, and, even if it was, it could not be possibly concluded that the change was solely due to the difference in one’s height.”

Challenges with Unstructured Data

- The lack of structure makes compilation and organizing unstructured data a time- and energy-consuming task.
- It would be easy to derive insights from unstructured data if it could be instantly transformed into structured data.
- structured data is akin to machine language, in that it makes information much easier to be parsed by computers.
- Unstructured data, on the other hand, is often how humans communicate (“natural language”); but people do not interact naturally with information in strict, database format.

Data Collections

Open Data

- The idea behind open data is that some data should be freely available in a public domain that can be used by anyone as they wish, without restrictions from copyright, patents, or other mechanisms of control.

Eg: Local and federal governments, non-government organizations (NGOs), and academic communities all lead open data initiatives.

- Following is the list of principles associated with open data as observed in the policy document:

1. Public
2. Accessible
3. Described
4. Reusable
5. Complete
6. Timely
7. Managed Post-Release

Social Media Data

- Social media has become a gold mine for collecting data to analyze for research or marketing purposes.
- This is facilitated by the Application Programming Interface (API) that social m
- Think of the API as a set of rules and methods for asking and sending data.
- For various data-related needs (e.g., retrieving a user's profile picture) one could send API requests to a particular social media service.
- This is typically a programmatic call that results in that service sending a response in a structured data format, such as an XML.

Multimodal Data

- We are living in a world where more and more devices exist – from lightbulbs to cars – and are getting connected to the Internet, creating an emerging trend of the Internet of Things (IoT).
- These devices are generating and using much data, but not all of which are “traditional” types (numbers, text). When dealing with such contexts, we may need to collect and explore multimodal (different forms) and multimedia (different media) data such as images, music and other sounds, gestures, body posture, and the use of space.

Data Storage and Presentation

- Depending on its nature, data is stored in various formats.
- We will start with simple kinds –data in text form.
- If such data is structured, it is common to store and present it in some kind of delimited way. That means various fields and values of the data are separated using delimiters, such as commas or tabs.
- Two of the most commonly used formats that store data as simple text – comma-separated values (CSV) and tab-separated values (TSV).

- CSV (Comma-Separated Values) format is the most common import and export format for spreadsheets and databases.

```
treat,before,after,diff
No Treatment,13,16,3
No Treatment,10,18,8
No Treatment,16,16,0
Placebo,16,13,-3
Placebo,14,12,-2
Placebo,19,12,-7
Seroxat (Paxil),17,15,-2
Seroxat (Paxil),14,19,5
Seroxat (Paxil),20,14,-6
Effexor,17,19,2
Effexor,20,12,-8
Effexor,13,10,-3
```

- TSV (Tab-Separated Values) files are used for raw data and can be imported into and exported from spreadsheet software.

```
Name<TAB>Age<TAB>Address
Ryan<TAB>33<TAB>1115 W Franklin
Paul<TAB>25<TAB>Big Farm Way
Jim<TAB>45<TAB>W Main St
Samantha<TAB>32<TAB>28 George St
```

- **XML** (eXtensible Markup Language) was designed to be both human- and machine readable, and can thus be used to store and transport data.
- **RSS** (Really Simple Syndication) is a format used to share data between services, and which was defined in the 1.0 version of XML.
- **JSON** (JavaScript Object Notation) is a lightweight data-interchange format. It is not only easy for humans to read and write, but also easy for machines to parse and generate.

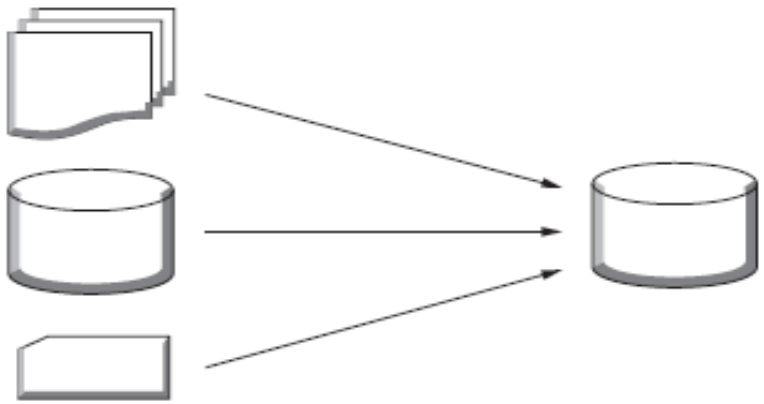
Data Pre-processing

- Data in the real world is often dirty; that is, it is in need of being cleaned up before it can be used for a desired purpose. This is often called **data pre-processing**.
- The factors that indicate that data is not clean or ready to process:

1. **Incomplete**: When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.
2. **Noisy**. When data contains errors or outliers
3. **Inconsistent**. Data contains discrepancies in codes or names.

Data Cleaning

Data Integration



Data Transformation –17, 25, 39, 128, –39 → 0.17, 0.25, 0.39, 1.28, –0.39

Data Reduction

	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

→

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Forms of data pre-processing

Data Cleaning

- Since there are several reasons why data could be “dirty,” there are just as many ways to “clean” it.
- The three key methods that describe ways in which data may be “cleaned,” or better organized, or scrubbed of potentially incorrect, incomplete, or duplicated information.

Data Munging

- Often, the data is not in a format that is easy to work with.
- For example, it may be stored or presented in a way that is hard to process. Thus, we need to convert it to something more suitable for a computer to understand.
- To accomplish this, there is no specific scientific method.
- The approaches to take are all about manipulating or wrangling (or munging) the data to turn it into something that is more convenient or desirable. This can be done manually, automatically, or, in many cases, semi-automatically.

Consider the following text recipe:

“Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix.”

Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

Wrangled data for a recipe

Handling Missing Data

- Sometimes data may be in the right format, but some of the values are missing.
- Consider a table containing customer data in which some of the home phone numbers are absent. This could be due to the fact that some people do not have home phones – instead they use their mobile phones as their primary or only phone.
- Other times data may be missing due to problems with the process of collecting data, or an equipment malfunction. Or, comprehensiveness may not have been considered important at the time of collection.
- some data may get lost due to system or human error while storing or transferring the data.

- Strategies to combat missing data include ignoring that record, using a global constant to fill in all missing values, imputation, inference-based solutions (Bayesian formula or a decision tree), etc.

Smooth Noisy Data

- There are times when the data is not missing, but it is corrupted for some reason. This is, in some ways, a bigger problem than missing data.
- Data corruption may be a result of faulty data collection instruments, data entry problems, or technology limitations.

Data Integration

To be as efficient and effective for various data analyses as possible, data from various sources commonly needs to be integrated.

The following steps describe how to integrate multiple databases or files.

1. Combine data from multiple sources into a coherent storage place (e.g., a single file or a database).
2. Engage in schema integration, or the combining of metadata from different sources.
3. Detect and resolve data value conflicts. For example:
   a. A conflict may arise; for instance, such as the presence of different attributes and values from various sources for the same real-world entity.
   b. Reasons for this conflict could be different representations or different scales; for example, metric vs. British units.

4. Address redundant data in data integration. Redundant data is commonly generated in the process of integrating multiple databases. For example:

a. The same attribute may have different names in different databases.

b. One attribute may be a “derived” attribute in another table; for example, annual revenue.

c. Correlation analysis may detect instances of redundant data.

Data Transformation

- Data must be transformed so it is consistent and readable (by a system). The following five processes may be used for data transformation.

1. Smoothing: Remove noise from data.

2. Aggregation: Summarization, data cube construction.

3. Generalization: Concept hierarchy climbing.

4. Normalization: Scaled to fall within a small, specified range and aggregation. Some of the techniques that are used for accomplishing normalization (but we will not be covering them here) are:

   a. Min–max normalization.

   b. Z-score normalization.

   c. Normalization by decimal scaling.

5. Attribute or feature construction.

   a. New attributes constructed from the given ones.

Data Reduction

- Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained.
- One example of a large dataset that could warrant reduction is a data cube.
- Data cubes are multidimensional sets of data that can be stored in a spreadsheet.
- A data cube could be in two, three, or a higher dimension.

The two of the most common techniques used for data reduction.

1. Data Cube Aggregation
2. Dimensionality Reduction

Data Discretization

- We are often dealing with data that are collected from processes that are continuous, such as temperature, ambient light, and a company's stock price.
- But sometimes we need to convert these continuous values into more manageable parts. This mapping is called **discretization.**
- In undertaking discretization, we are also essentially reducing data.
- There are three types of attributes involved in discretization:

a. Nominal: Values from an unordered set

b. Ordinal: Values from an ordered set

c. Continuous: Real numbers