Introduction to Data Science

Dr. Sandareka Wickramanayake

sandarekaw@cse.mrt.ac.lk

Learning Outcomes

After completing this module, students should be able to

- Demonstrate data acquisition, data representation, and data pre-processing skills to describe, analyze and repurpose data from a variety of sources.
- Apply critical thinking and statistical techniques to understand and visualize relationships in data
- Apply machine-learning techniques in exploratory data analysis for problems related to commerce, industry, and research.
- Design and compute a statistical relationship in data including correlation and linear regression
- Design and develop data-driven algorithms for outcome prediction

Delivery

- Lectures - Thursday 1.15 – 3.15 PM (Seminar Room)
- Labs – Thursday 3.15 – 5.15 PM

- Lectures
  - Dr. Nisansa De Silva - NisansaDdS@cse.mrt.ac.lk
  - Dr. Sandareka Wickramanayake - sandarekaw@cse.mrt.ac.lk

Course Outline

Week	Lecture Topic	Lecturer
1	Introduction	SW
2	Data collecting, data documenting, data quality	SW
3	Data preprocessing	SW
4	Descriptive analysis	SW
5	Exploratory analysis	SW
6	Hypothesis Testing	NdeS
7	Visualization and Dashboarding	SW
8	Supervised Learning	NdeS
9	Unsupervised Learning and Evaluation	NdeS
10	Project Week	NdeS
11	Prescriptive and Cognitive Analytics	NdeS
12	Big Data	NdeS
13	Ethics	NdeS
14	Data Science Project Evaluation and Discussion	NdeS

Assessments

- Continuous Assessment – 40%
  - Bi-Weekly Lab/Activities – 15%
  - Class project (Group) – 25%
    - Pre-processed dataset
    - Final report
      - Data preprocessing approach
      - Insights from data analysis

- Final Examination – 60%
  - Online examination conducted in CSE labs
  - 2 hours
  - Open book?

Reading Materials

- No specific textbook
- Additional reading materials related to each topic will be posted on Moodle.

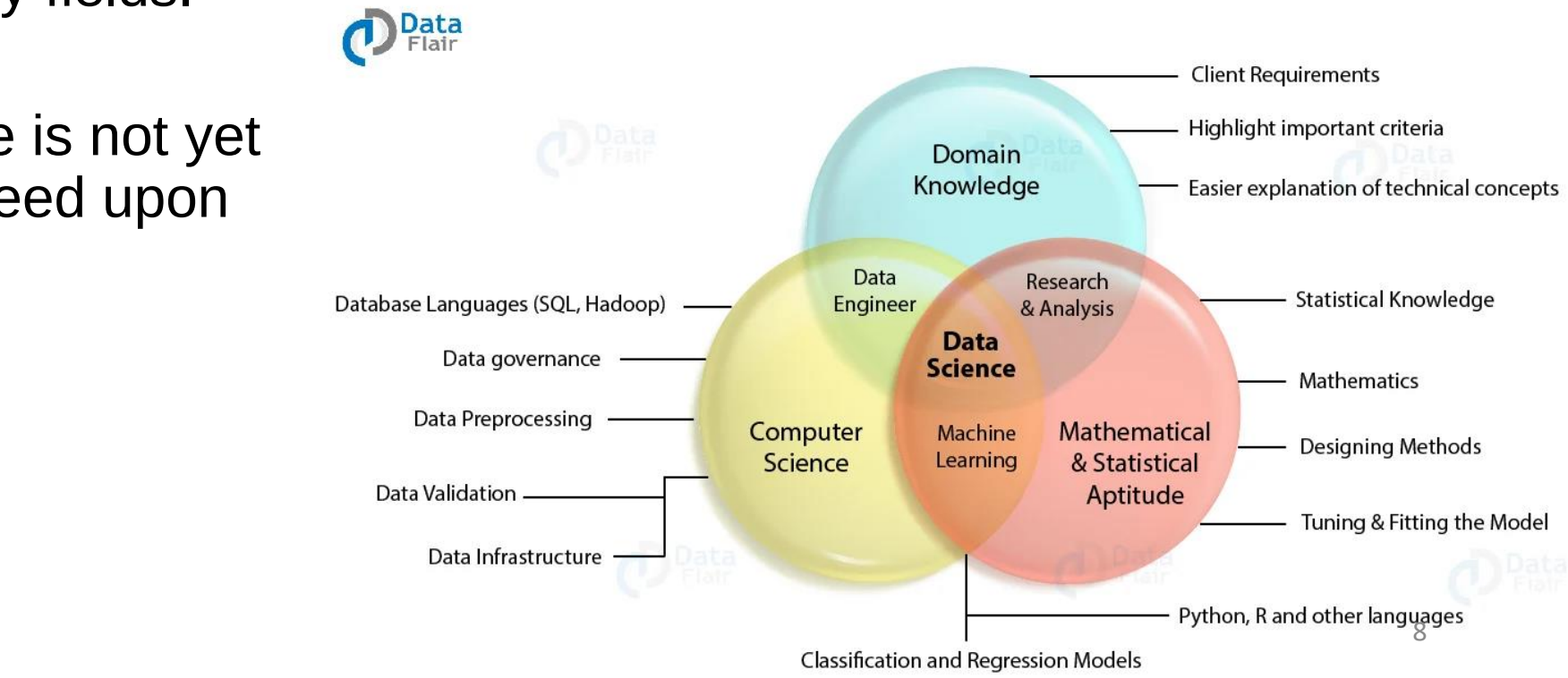
Are We Using Data Science Products?

https://www.youtube.com/watch?v=8Fz2nDfZinE&t=104s

What is Data Science?

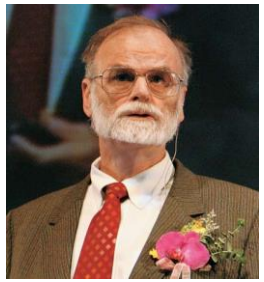
- Data Science
  - Is the extraction of knowledge from large volumes of data.
  - Uncovers actionable insights hidden in data that can be used to guide decision-making and strategic planning.
  - Combines many fields.

  - However, there is not yet a definition agreed upon by all.



What is Data Science?

- Data science = the Fourth Paradigm of Science. – Jim Gray

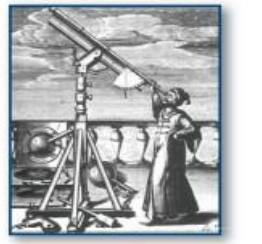


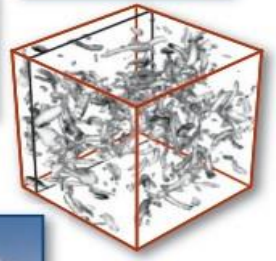
(1942-2012)

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - K\frac{c^2}{a^2}$$





What is Data Science?

- Data science = Science of data
  - The intellectual and practical activity encompassing the systematic study of facts and statistics collected for reference or analysis.

Google's definition of "data"

data

/ˈdeɪtə/

noun

facts and statistics collected together for reference or analysis.
"there is very little data available"
synonyms: facts, figures, statistics, details, particulars, specifics, features;

Google's definition of "science"

science

/ˈsʌɪəns/

noun

the intellectual and practical activity encompassing the systematic study of the structure and behaviour of the physical and natural world through observation and experiment.

What is Data Science?

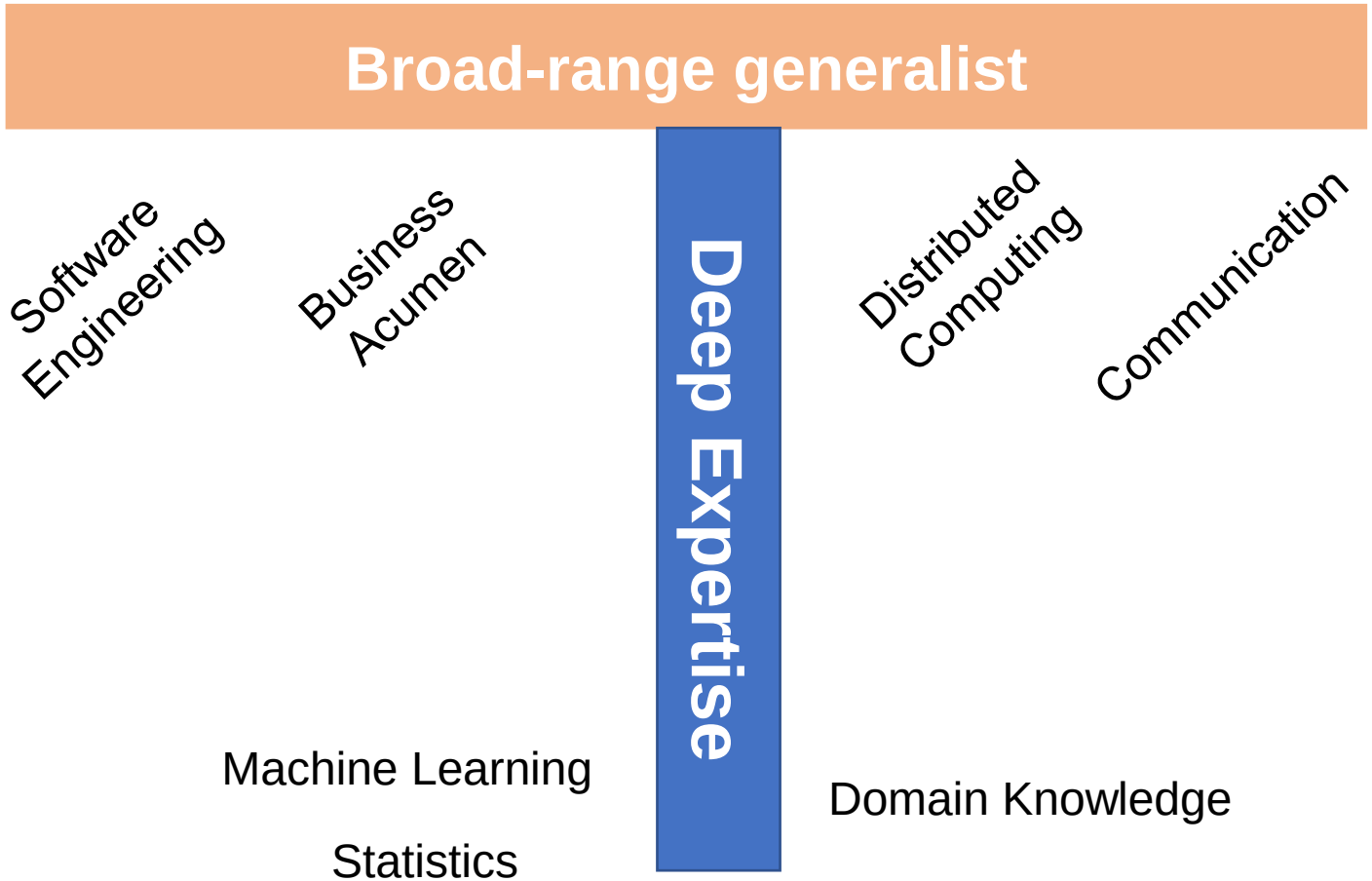
Wikipedia	“Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured”
NIST, 2015	“Data science is the empirical synthesis of actionable knowledge from raw data through the data lifecycle process”
Dhar, 2013	“Data science is the study of generalizable knowledge from data”
Peter Naur, 1974	“[data science is] The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.”

What is Data Science?

- Data science is an emerging discipline.
  - It remains a science where new knowledge and tools are still being invented.
- There is not yet a clear definition agreed upon by all for the term ‘data science’.
  - Different definitions exist from different perspectives (government, business, research, etc.)
  - We adapt NIST’s definition: “Data science is the empirical synthesis of actionable knowledge from raw data through the data lifecycle process”
- You, as the future data scientist, will shape the field.

To Succeed in Data Science

- You need the skills of a good Software Engineer and skills in Machine Learning.



What Does Data Say?

- Moodle activity 1.
- Describe what insights you can derive from the given figures.

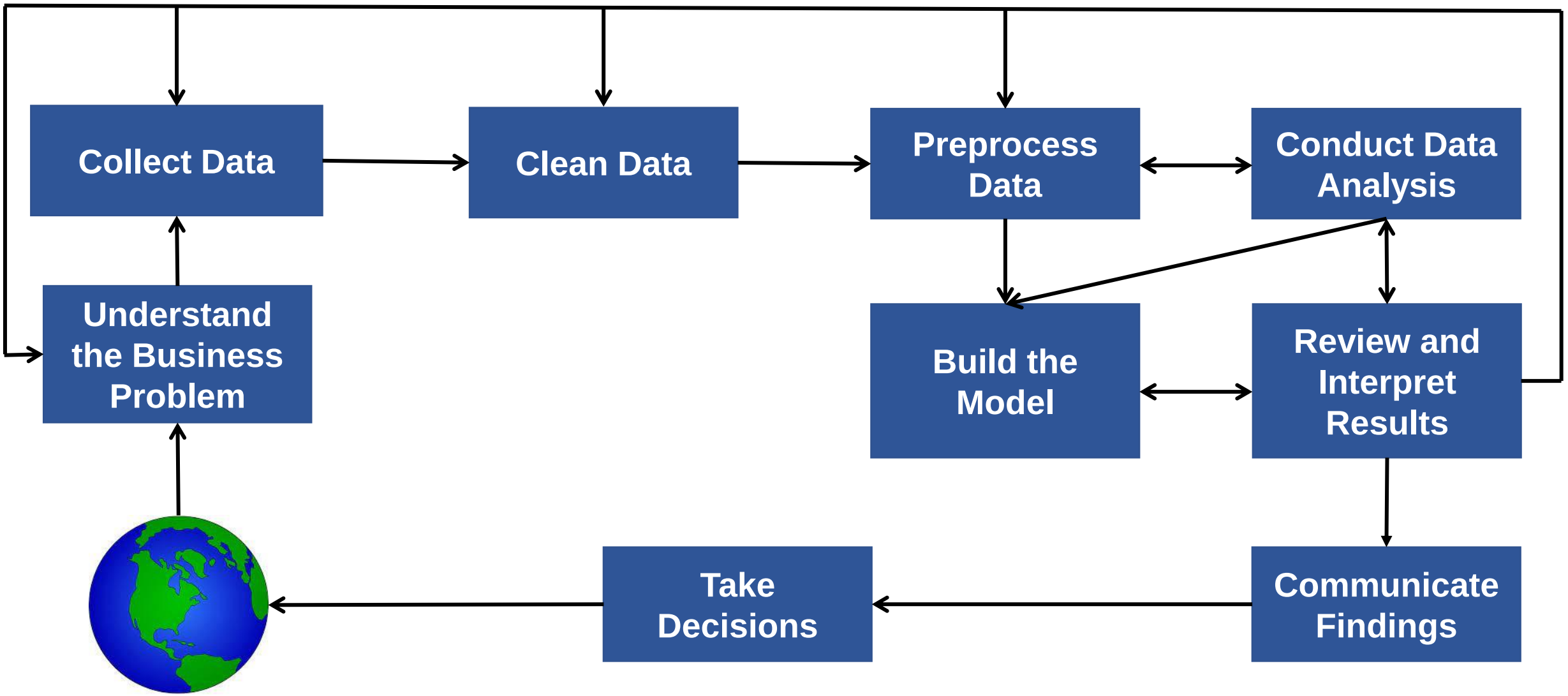
Data Science Process

- Moodle Activity 2 - Group Activity
  - Go to your assigned breakout room.
  - Create a short report describing the data science process.
    - The report should contain the data science process diagram.
    - Each component of the diagram should be briefly described.
  - The Maximum number of pages allowed is 2, however, the report should be comprehensive.

Data Science Process

https://youtu.be/X3paOmcrTjQ?t=10

Data Science Process



Data Science Process Using a Real World Example

https://www.youtube.com/watch?v=KdgQvgE3ji4&t=59s

Questions?