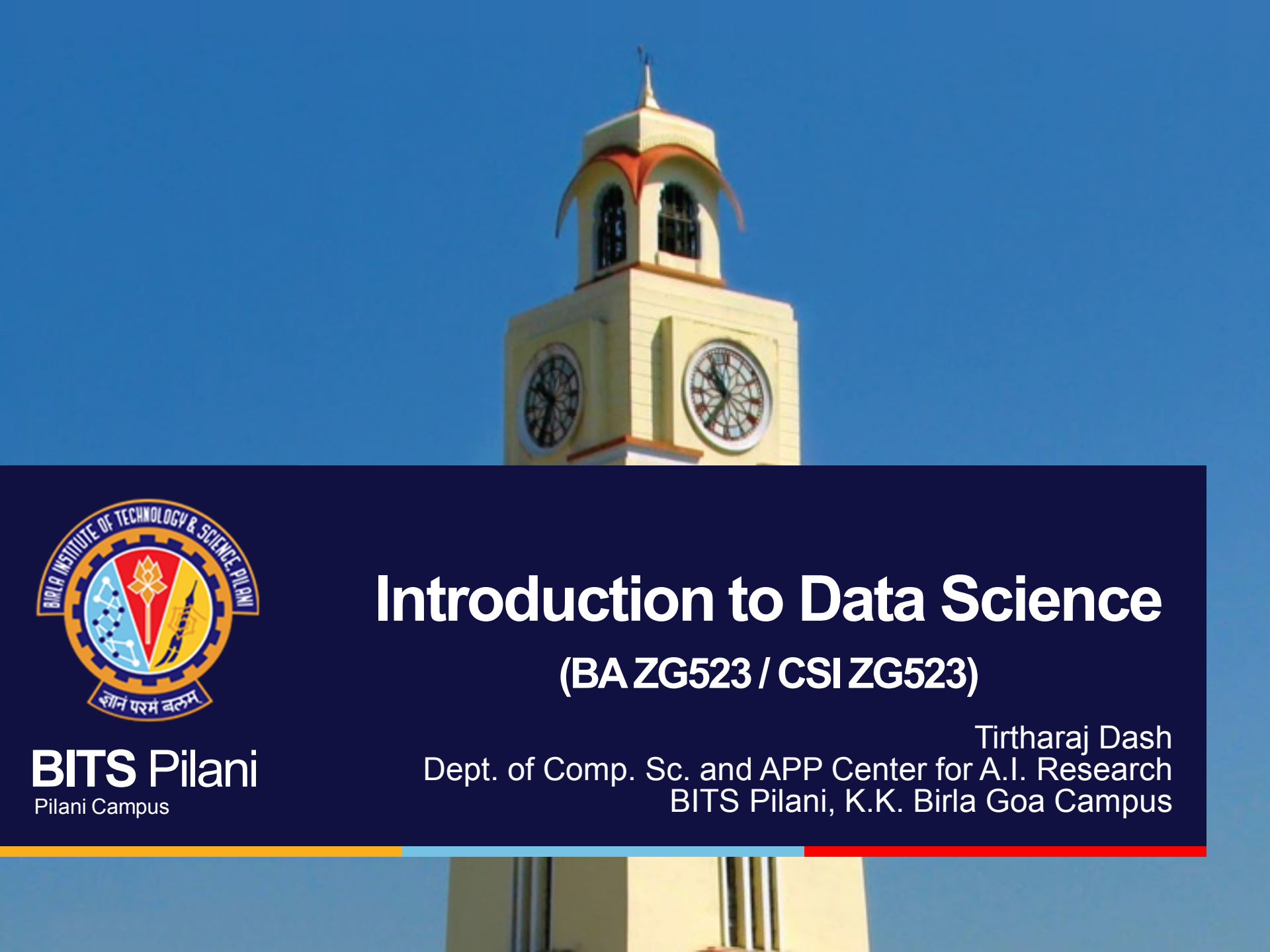

Introduction to Data Science

(BA ZG523 / CSI ZG523)

Tirtharaj Dash
Dept. of Comp. Sc. and APP Center for A.I. Research
BITS Pilani, K.K. Birla Goa Campus

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Introduction to Data Science
Lecture No. 01 (C.H.: 1, 2)

What is “Science”?

innovate achieve lead

Science

- Latin word ‘*scientia*’ - meaning knowledge
- a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe.[1]

- In *science* you must not talk before you know. In art you must not talk before you do. In literature you must not talk before you think. (--John Ruskin [2])

[1] Harper, Douglas. "science". *Online Etymology Dictionary*. Retrieved August 07, 2020.
[2] John Ruskin, "The Eagle's Nest," 1872.

What is “Computer Science”?

innovate achieve lead

Computer science

- is the study of computation and information.[1]

[1] Dijkstra, E.W. (1986). “On a cultural gap”. The Mathematical Intelligencer. 8 (1): 48-52.

Now, what is “Data Science”?

Data science

- is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data.[1]

- Can we now relate “Science”, “Computer Science” and “Data Science”?

[1] Dhar, V. (2013). "Data science and prediction". Communications of the ACM. 56 (12): 64–73.

Why “Data Science”?

At the moment, I will just show you a statement:

- “The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.” -- Hal Varian, *Chief Economist at Google and Founding Dean, School of Information, UC Berkeley*

"Data Scientist" – attractive, isn't it?

- "Data scientist" – a term used for data professionals who are skilled in organizing and analyzing massive amounts of data. (coined in 2008)

- Data scientists examine which questions need answering, where to find related data (most useful for decision making).
- Businesses use data scientists to source, manage, and analyze large amounts of data.
- *Skills needed:* Programming (e.g. Python, R), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning

About this course

- Fundamentals of Data Science
- Introducing data representation, data visualization, learning algorithms (some) …
- Some problem solving via mathematics, and programming.
- Programming will be done on a open-source platform.

- This course will NOT introduce you to much of Machine Learning (ML) concepts.
- This course will NOT introduce you to much of Data Mining concepts.

Course Objectives

CO1	Gain basic understanding of the role of Data Science in various contexts
CO2	Understand the role of various concepts like Statistics, Machine learning etc. in Data Science
CO3	Understand the roles and stages in a Data Science Project
CO4	Understanding the key terms and tools used by Data Scientist
CO5	Understand the process of collecting data from unstructured sources and store it using appropriate structure such as relational databases, graphs, matrices, etc.

Learning Outcomes:

LO1	understand and apply the principles of Data Science
LO2	describe the structure of a Data Science project
LO3	understand key terms and apply the tools used by a Data Scientist
LO4	understand and apply the algorithms used in Data Science

But, I can't promise that after successful completion of this course, you will end up being a "Data Scientist" :-) The one thing that I can promise is that your interest in this field will grow and you would definitely want to do something with it.

Evaluative Components:

No	Name	Type	Duration	Weight	Day, Date, Session, Time
EC-1	Quiz-I/ Assignment-I	Online	-	5%	September 10-20, 2020
	Quiz-II	Online	-	5%	October 20-30, 2020
	Experiential Learning	Online	-	15%	November 10-20, 2020
EC-2	Mid-Semester Test	Closed Book	2 hours	30%	Sunday, 11/10/2020 (FN) 10 AM - 12 Noon
EC-3	Comprehensive Exam	Open Book	3 hours	45%	Sunday, 29/11/2020 (FN) 9 AM – 12 Noon

Books:

T1 An Introduction to Data Science by Jeffrey Stanton (Free e-Book)

T2 Practical Data Science with R by Nina Zumel and John Mount, Indian Edition by Dreamtech Press, 2014.

T3 The Art of Data Science by Roger D Peng and Elizabeth Matsui

T4 Analytics in a Big Data World, Bart Baesens, Wiley

R1 Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson

My Information:

Tirtharaj Dash

Assistant Professor
Dept. of Computer Science
and
APP Center for A.I. Research

BITS Pilani, K.K. Birla Goa Campus
E-mail: tirtharaj@goa.bits-pilani.ac.in

https://www.bits-pilani.ac.in/goa/tirtharaj/Profile

TIRTHARAJ DASH

Tirtharaj Dash

Research Interests

→ Deep Learning
→ Neuro-Symbolic Learning
→ Machine Learning
→ Stochastic Optimisation

My expectations from you:

1. High commitment to the course.
2. Ready to get your hands dirty with some problem solving (mathematical and programming).
3. Active participation in class (you can interrupt and ask question)

Welcome to this course and I hope you will enjoy this.

Data

"Ok, Google"

- What time is my meeting tomorrow?" (10AM)
- Wake me up at 8am tomorrow. (Alarm set: 8AM)
- What are today's news? (starts reading…)
- Make a call to maa. (dials maa)
- Navigate me to office. (starts map…)

"Alexa"

- Play a Tamil song. (Starts playing…)
- Where is my order? (Reports status)

…and, many more stuff… with the day ending with Netflix.

More data: everyday

- Every search that we do.
- Every email we send, or spam.
- Every social network that we use.
- Every shopping site that we use.
- Every song that we listen.
- Every movie we watch.
- Every food that we order and eat.
- Every transaction that we make online.
- Every call we make.
- Every place we visit.
- Every thought that comes to our mind (…one day)*

Companies and data

innovate achieve lead

- Task for you: Find out how much data these top companies such as Google, Facebook, Twitter, Amazon (and many others) generate per day.

- And, what do they even do with those data?
- Are these data even useful?
- If yes, are all data useful or some specific?

Data Science is not new

innovate achieve lead

- Behind resolving many daily tasks for several years.
- Methods to fuel data science have been there:
  - Pierre Simon Laplace (1749–1827)
  - Thomas Bayes (1701–1761).
- Machine learning is younger (but, it has now well established).
- Computer science is very old.

- Then, why is data science seen as novel trend in last few years?

Reason 1: Datafication

- ‘Datafication’ (or ‘Datification’ is a technological trend turning many aspects of our life into data [1].
- Data has value.

- See slide 17 (More data: everyday).

[1] Cukier, Kenneth; Mayer-Schoenberger, Viktor (2013). "The Rise of Big Data". *Foreign Affairs* (May/June): 28–40. Retrieved 07 August 2020.

Reason 2: Hardware

- With time, computational machinery has evolved.
- Now-a-days, we talk about GPGPU and TPU
  - GPGPU: General-Purpose Graphics Processing Unit
  - TPU: Tensor Processing Unit
- This has allowed big companies to use their stored data to generate knowledge.
- Knowledge aided decision making.

Data Science

- A methodology by which actionable insights can be inferred from data.
- These insights will form some form of beliefs.
- These beliefs can be used as the basis of decision-making

Strategies to explore world using data

1. Probing reality
   - Data can be gathered by passive or by active methods.
2. Pattern discovery
   - Discovering useful patterns by analyzing vast amount of data.
3. Predicting future events
   - This is a very essential component in business: build models that are robust in predicting future data samples.
4. Understanding people and the world.
   - Companies and governments are investing huge amounts of money on this.
   - Deep Learning has fueled understanding of natural languages, computer vision, psychology, and neuroscience.

… and, for a happier weekend ☺

- I assume:
  - You have a computer system (with Linux or Windows or Mac OS)
  - You know how to type using computer keyboard
  - You are already enjoying this course

- Task 0: Install Python latest (3.x) on your machine. You will find it here:

  https://www.python.org/downloads/

- Task 1: Practice (Chapter 1 and 2) from:

  https://www.py4e.com/

  (this is an engaging book on Python)