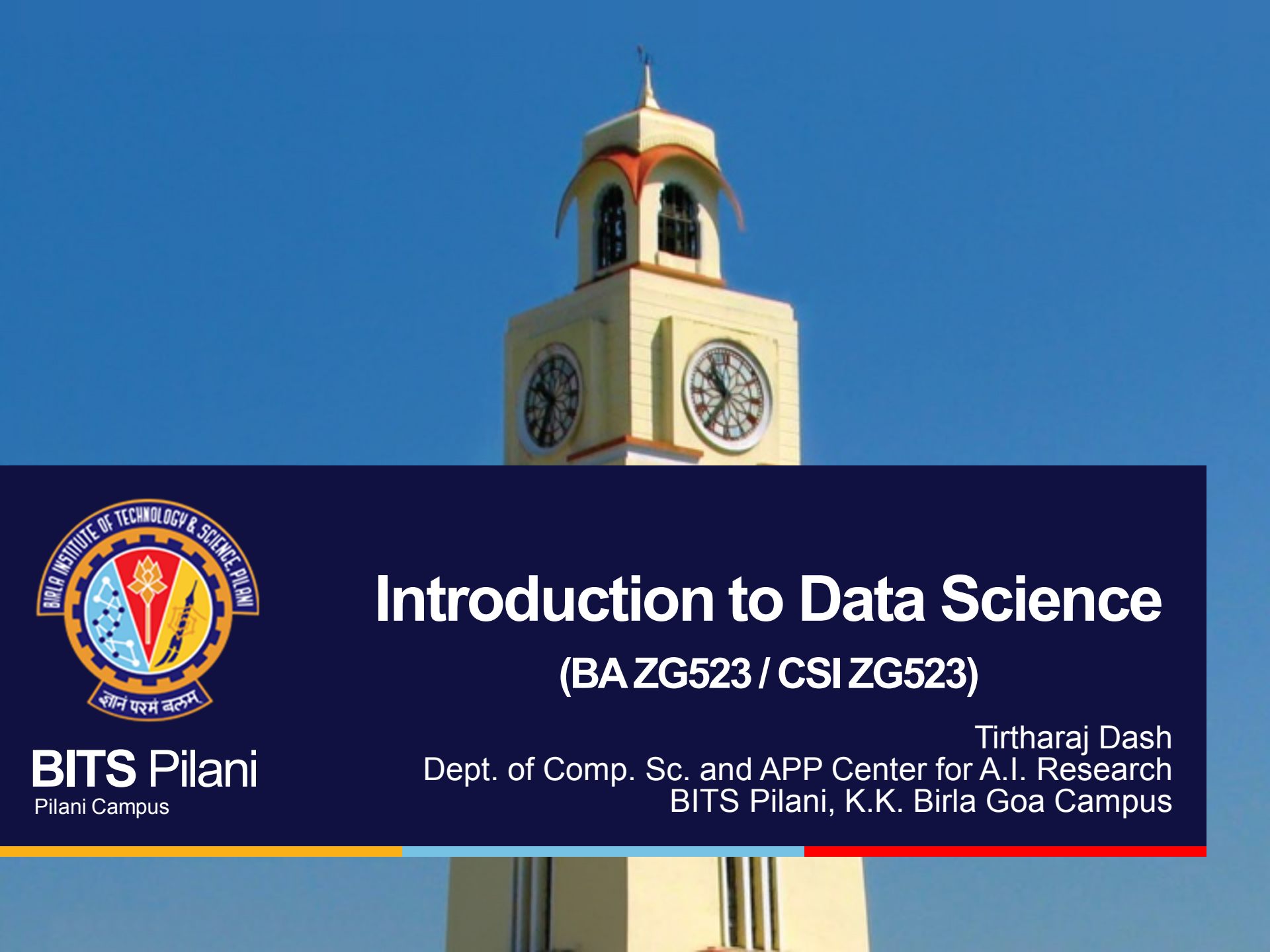


Introduction to Data Science

(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. 03 (C.H.: 5, 6)

Previous Lecture:

- Various stages in Analytics Process Model
  - Sourcing data
  - data mart
  - cleaning
  - transformation
  - analytics
  - interpretation and evaluation
- Roles and responsibilities of a data scientist
  - Works closely with business stakeholders
  - Understand business goals
  - Determines how data can be achieved to achieve these goals
  - Designs data modeling processes
  - Create algorithms and predictive models
  - Help analyze data and share insights with peers

Cont.

- Specific activities of a data scientist
  - Data gathering, preparation, exploration
  - Data representation and transformation
  - Computing with data, data modeling with ML, AI, statistics
  - Data visualization
  - Presentation.
- Data Science Ethical guidelines (6)
  - Autonomous machines
  - Bias, discrimination, exclusion
  - Algorithmic profiling
  - Preventing massive files
  - Quality, quantity and relevance
  - Human identity before human-machine hybridization
- Practical: Some basic Python programming.

Roles in a data sc. project

- Sometimes, these roles may overlap.

Role	Responsibilities
Project sponsor	Represents the business interests; champions the project
Client	Represents end users' interests; domain expert
Data scientist	Sets and executes analytic strategy; communicates with sponsor and client
Data architect	Manages data and data storage; sometimes manages data collection
Operations	Manages infrastructure; deploys final project results

[Source: T2, Ch.1]

- Client, Data architect, and Operations are not part of the data science team, rather they are collaborators.

Roles: Project Sponsor

innovate achieve lead

- The most important role in a project.
- They decide whether the obtained results are a success or a failure.
- Keep the project sponsor involved and informed about intermediate progresses and results.
- Getting clear goals from sponsor is important – best is to get quantitative statements.
- Example of a goal: Identify 90% of accounts that will go into default at least two months before the first missed payment with a false positive rate of no more than 25%.

Roles: Client

- Represents model's end-user interests.
- More hands-on than the sponsor.
- They form an interface between technical details of the project and the day-to-day expected works when the project is deployed.
- Keeping them informed and taking feedbacks during project is essential.

Roles: Data Scientist

- Responsible for taking all necessary steps to make the project a success.
- *Read more on this role from the previous lecture.*

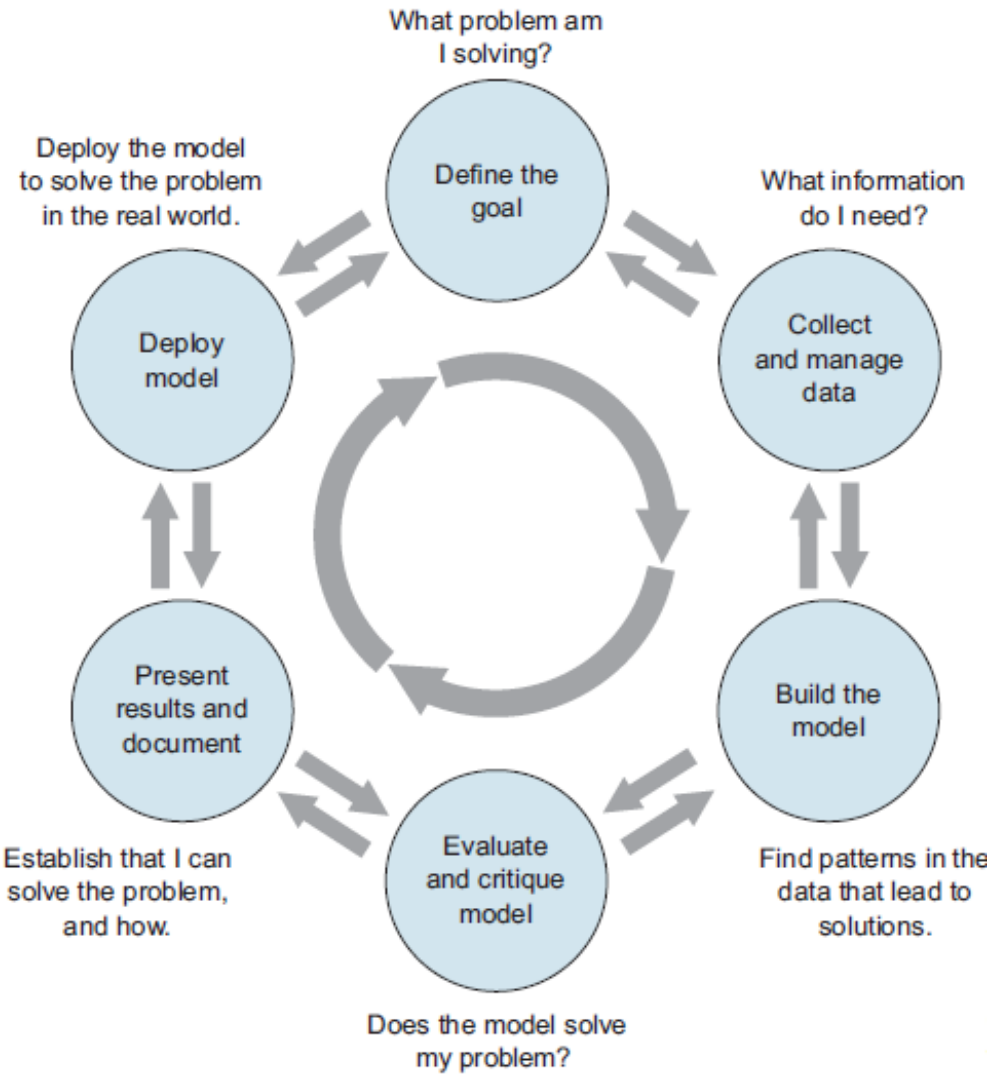
Roles: Data Architect

- The data architect is responsible for all of the data and its storage.
- Usually a database administrator or architect fills this role.
- Manages data warehouses.
- Only available for quick consultations (They manage many projects.)

Roles: Operations

- Critical role in acquiring data and delivering results.
- They are mostly from outside data science group.
- E.g. You have built a project for online shopping site. They are aware of the technicalities of deployment, response time, etc.

Stages of a data sc. project



[Source: T2, Ch.1]

Learn via an example:

- The problem of classifying bank customers as to whether they should receive a loan or not.
- Giving a loan to a bad customer who is (mistakenly) marked as a good customer results in a greater cost to the bank than denying a loan to a good customer marked as a bad customer.
- The client want to build a tool using ML to automate the process of customer classification and with reasonable accuracy* (**: we are at present not focused on this*)

- *For experimenting with stages, we will use a Python notebook, available in our Lab repo.*

Stage: Define the goal

- The first step is to define a quantifiable goal.
- The following questions are relevant:
  - Why do the sponsors want the project in the first place? What do they lack, and what do they need?
  - What are they doing to solve the problem now, and why isn't that good enough?
  - What resources will you need: what kind of data and how much staff?
  - Will you have domain experts to collaborate with, and what are the computational resources?
  - How do the project sponsors plan to deploy your results? What are the constraints that have to be met for successful deployment?

Stage: Collect data

innovate achieve lead

- Identifying the data you need, exploring it, conditioning it for suitable analysis.
- The following questions are relevant:
  - What data is available to me?
  - Will it help me solve the problem?
  - Is it enough?
  - Is the data quality good enough?

- Here we will focus on available dataset from the web [1].

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Stage: Modeling

- Try to extract useful insights from the data in order to achieve your goals.
- There can be back-and-forth between data collection phase and modeling phase.
- The most common data science modeling tasks are these:
  - *Classification—Deciding* if something belongs to one category or another
  - *Scoring—Predicting* or *estimating* a numeric value, such as a price or probability
  - *Ranking*—Learning to *order items* by preferences
  - *Clustering—Grouping items* into most-similar groups
  - *Finding relations—Finding correlations* or potential causes of effects seen in the data
  - *Characterization*—Very general *plotting* and *report generation* from data

Stage: Model Evaluation

innovate achieve lead

- Once you have a model, you need to determine if it meets your goals:
  - Is it accurate enough for your needs? Does it generalize well?
  - Does it perform better than “the obvious guess”? Better than whatever estimate you currently use?
  - Do the results of the model (coefficients, clusters, rules) make sense in the context of the problem domain?
- If you’ve answered “no” to any of these questions, it’s time to loop back to the modeling step or decide that the data doesn’t support the goal you’re trying to achieve.

Stage: Presentation

- Once you have a model that meets your success criteria, you'll present your results to your project sponsor and other stakeholders.
- You must also document the model for those in the organization who are responsible for using, running, and maintaining the model once it has been deployed.

Stage: Model Deployment

innovate achieve lead

- Finally, the model is put into operation.
- In many organizations this means the data scientist no longer has primary responsibility for the day-to-day operation of the model.
- But you still should ensure that the model will run smoothly and won't make disastrous unsupervised decisions.
- During maintenance of model, it may come back to data scientists.

Setting Expectations

- Setting expectations is a crucial part of defining the project goals and success criteria.
- The expectation is nothing but the model performance to satisfy the project goal.
- You should not set over-optimistic expectations.
- You should not keep too low expectations.

Cont.

- Determine the lower and upper bounds on model performance.
  - **The NULL model**: If you don't have an existing model to compare with, then the "obvious guess" is treated as a base model. You have to do better than this.
  - **The Bayes rate**: The limit on prediction accuracy due to unexplainable variance is known as the *Bayes rate*. You can think of the Bayes rate as describing the best accuracy you can achieve given your data.