

CS 215: Data Interpretation and Analysis

Sunita Sarawagi
Autumn 2024

Welcome!

What is the course about?

- Suppose you want to find reliable answers to questions:
 - Which minor should I opt for?
 - What are the types of future careers that IITB graduates favor lately?
 - How many students seats should IITB allocate to each department?
 - Which products are likely to be in high demand next month?
 - Is rainfall in Mumbai becoming more erratic lately?
 - Is inflation increasing at a faster pace in recent times?
 - How is supply of drinking water keeping pace with increasing population?
 - Is a flu vaccine useful to prevent frequent cold&fever?

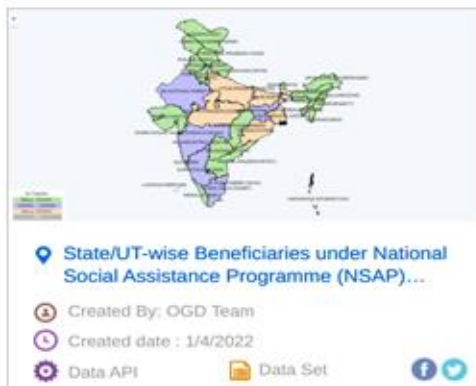
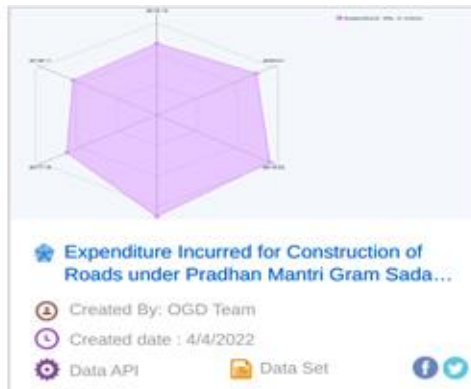
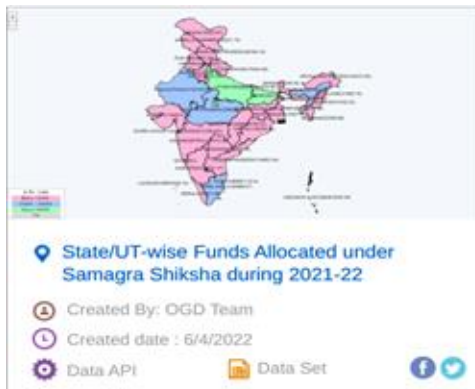
How do you find the answers?

- Go by your existing vague impressions
- Ask your peer group, Ask older experienced people
- Do a websearch
- Ask ChatGPT
- ...

The data scientists approach

- Go to an authentic source that has recorded correctly the observed values over time → This is your data
 - Public data: World bank datasets, Datacommons, National Data Analytics platform, Stock prices
 - Enterprise data: Student data in universities, sales and customer interaction data in enterprises
 - Scientific data: experiments, simulations and observations in lab
- Try to find answers from the data → How?
 - This course will teach you how to get answers to top-level questions from data in a scientific way.

Several sources of public data in India



Some example studies on Indian Data

- [Power consumption in India](#)
- [Health of Indian population](#)
- [Housing in India](#)

Course contents

- Data analysis: gathering, summarizing, and visualizing data in intuitive ways
- Probability: Mathematical tool to represent uncertainty
- Statistical inference: Drawing probabilistic conclusions from limited data

Important pre-requisite for future courses in machine learning, image processing, computer vision, deep learning, AI, finance, etc..

Mode of running the course

- Three 55 minute slots per week:
- SAFE/Moodle/paper quizzes on the material covered in **prior** weeks
 - 20 minute duration at a pre-announced time or 55 minute quiz.
 - Grading will be done on top $n-2$ out of n quizzes for 20 minute quizzes.
 - No compensation for missed quizzes.
- All materials will be uploaded on Moodle, announcements via Moodle, questions on Moodle or cs215-ta@googlegroups.com
- [Course webpage](#)

Evaluation

Approximate credit structure

- 15% In-class Quizzes
- 25% Mid-semester exam
- 35% End semester exam
- 25% Programming and written homeworks: in teams (about 5 assignments)
- Attendance mandatory. Students with less than 80% may get a DX.

We will all adhere to principles of academic honesty. Penalties for violation will be severe and will be reported to DADAC. Givers and takers are equally responsible.