

CS 215: Data Interpretation and Analysis

Fall 2023

Instructors:

Ajit Rajwade

&

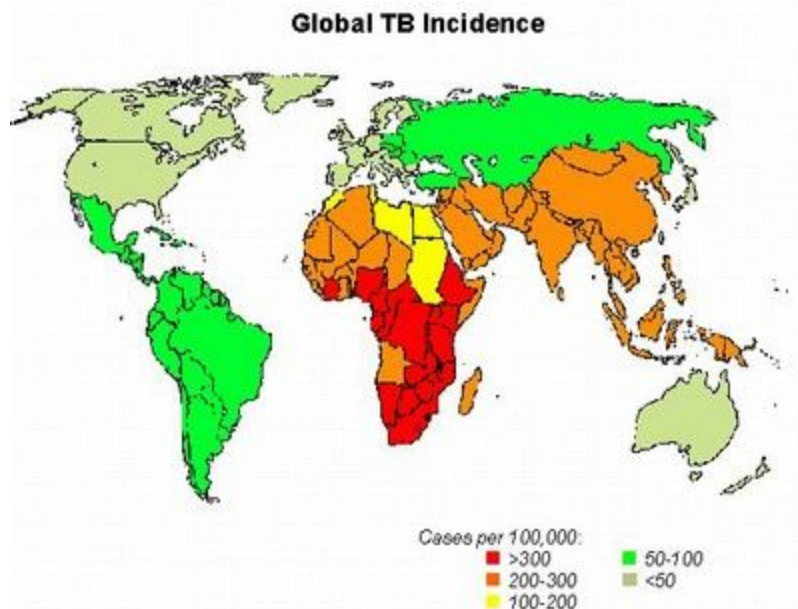
Pushpak Bhattacharya

https://www.cse.iitb.ac.in/~ajitvr/CS215_Fall2023/

Where all do you analyze and interpret data?

(1) In Medicine: Examples

- Pathology reports,
- Epidemiology studies



<https://ethnomed.org/clinical/tuberculosis/firlan>
[d/epidemiology-of-tb](#)

Where all do you analyze and interpret data?

Leading ODI Run Scorers at Number 6 since 1 Aug 2009								
Player	Mat	Inns	Runs	HS	Ave	SR	100	50
MS Dhoni (India)	68	58	1960	139*	50.3	84.6	1	15
Umar Akmal (Pak)	59	54	1706	102*	38.8	88.5	2	11
AD Mathews (SL)	66	59	1571	89	34.2	81.3	0	11
SK Raina (India)	47	41	1184	106	34.8	97.5	1	7
Mushfiquur Rahim (Ban)	40	37	897	86	30.9	79.3	0	6
MEK Hussey (Aus)	29	27	875	79	39.8	93.8	0	5
KA Pollard (WI)	39	36	843	119	24.8	85.1	2	2
DA Miller (SA)	39	34	797	67	31.9	97.0	0	5
RS Bopara (Eng)	31	27	715	101*	31.1	84.5	1	3
DJ Hussey (Aus)	23	20	684	74	42.8	95.3	0	6

<http://i.dawn.com/primary/2015/02/54d32f884dfd0.jpg?r=1999182479>

(2) In Sports

- Tournament data
- Player data
- Questions like: which is the best team?
Which is the best batsman? Which is the best batsman from so and so age-group?

Where all do you analyze and interpret data?

List by the International Monetary Fund (2014)
Rank Country/Region GDP (Millions of US\$)
World

1	United States	17,418,925
2	China	10,380,380[n 2]
3	Japan	4,616,335
4	Germany	3,859,547
5	United Kingdom	2,945,146
6	France	2,846,889
7	Brazil	2,353,025
8	Italy	2,147,952
9	India	2,049,501
10	Russia	1,857,461[n 3]
11	Canada	1,788,717
12	Australia	1,444,189
13	South Korea	1,416,949
14	Spain	1,406,855
15	Mexico	1,282,725
16	Indonesia	888,648
17	Netherlands	866,354
18	Turkey	806,108
19	Saudi Arabia	752,459
20	Switzerland	712,050

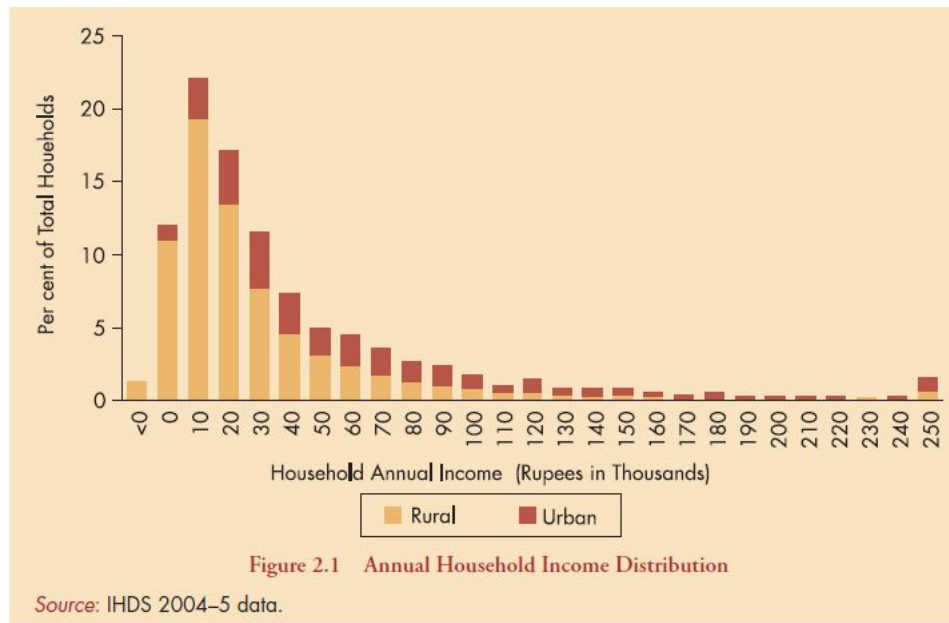
(3) In Economics and Finance:

- Country-wise data

Gross Domestic Product (GDP) is the broadest quantitative measure of a nation's total economic activity. More specifically, **GDP** represents the monetary value of all goods and services produced within a nation's geographic borders over a specified period of time.

<http://www.investinganswers.com/financial-dictionary/economics/gross-domestic-product-gdp-1223>

Where all do you analyze and interpret data?



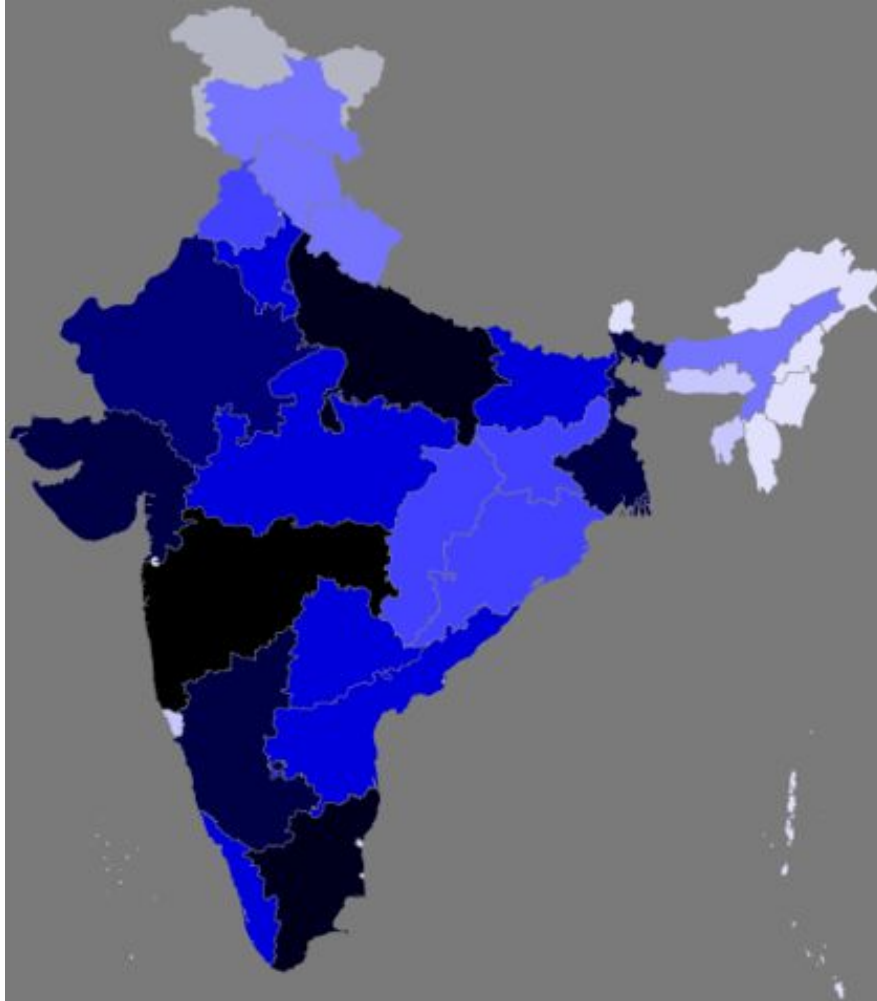
¹ Some households reported negative incomes. These are usually farm households with partially failed production whose value did not fully cover the reported expenses. Other analyses show that these households do not appear especially poor: their consumption expenditures and household possessions resemble average households more than they do to other low-income households. Because of this anomaly, for income calculations in the remainder of the study, we exclude all households with income below Rs 1,000 (N = 837). The median income after this exclusion is Rs 28,721.

(3) In Economics and Finance:

http://ihds.umd.edu/IHDS_files/02HDinIndia.pdf

- Country-wise data

Where all do you analyze and interpret data?



(3) In Economics and Finance:

- Region-wise data within a country

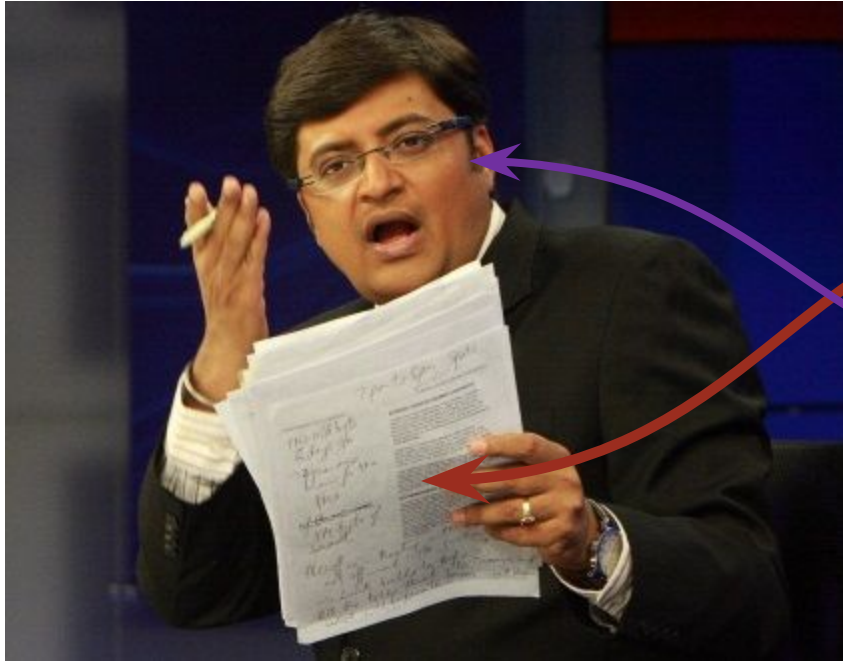
GDP of Indian states and union territories in 2014–15

Darker color = greater GDP

- over ₹14 lakh crore (US\$220 billion)
- ₹10 lakh crore (US\$160 billion) to ₹14 lakh crore (US\$220 billion)
- ₹8 lakh crore (US\$120 billion) to ₹10 lakh crore (US\$160 billion)
- ₹6 lakh crore (US\$93 billion) to ₹8 lakh crore (US\$120 billion)
- ₹4 lakh crore (US\$62 billion) to ₹6 lakh crore (US\$93 billion)
- ₹2 lakh crore (US\$31 billion) to ₹4 lakh crore (US\$62 billion)
- ₹1 lakh crore (US\$16 billion) to ₹2 lakh crore (US\$31 billion)
- ₹0.5 lakh crore (US\$7.8 billion) to ₹1 lakh crore (US\$16 billion)
- ₹0.25 lakh crore (US\$3.9 billion) to ₹0.50 lakh crore (US\$7.8 billion)
- less than ₹0.25 lakh crore (US\$3.9 billion)

Source: [wikipedia article](#)

Where all do you analyze and interpret data?




(4) In Journalism:

Data: that's what those white papers contain! ☺

And he analyzes those data big time!

Image source



Where all do you analyze and interpret data?

(5) In many other fields:

- Weather forecasting
- Psephology
- Stock markets
- Industrial testing
- Market research (eg: in industry and storehouses)
- Advertisements, ecommerce, recommender systems

So what's this course all about?

- Sounds like everything under the



http://www.clipartpanda.com/clipart_images/clipart-sun-rays-clipart-1587813

What's this course all about?

- A beginning course on probability and statistics
- A very useful base for future courses in machine learning, data science, statistics, image processing and computer vision.

What's this course all about? Three sections

- **Data analysis:** Process of gathering, displaying/visualizing and summarizing the data
- **Probability:** The “chance” that something happens
- **Statistical Inference:** The science of drawing precise inferences from the data gathered using tools from probability

Example in Toxicology

- Imagine I invent two new medicines (say) to reduce blood pressure (BP).
- I test the two medicines on two groups of rats – A and B – respectively.
- I will then periodically measure BP of rats in groups A and B.
- And seek to determine which medicine is “better”.

Example in Toxicology: Data Analysis

- What should be the size of A and B?
- How should I pick the members of A and B? Example: can A be all males, B be all females? Can A be all white rats and B be all black rats?
- Once I acquire the BP measurements, how do I display them succinctly? How do I compute averages?

Example in Toxicology: Data Interpretation (or Statistical Inference)

- Let's say the average BP of A was much lower than that of B after feeding the two drugs.
- Does this mean the first medicine is more effective?
- Or was this just a matter of chance? (Example: If I flip an unbiased coin 50 times, I could land up with 30 heads – just by chance!)

One more example: Serious or joking?

- Suppose your friend performs 10,000 independent tosses of an unbiased coin (i.e. equal chance of heads or tails).
- He reports ≥ 5200 heads.
- Is (s)he serious or joking?
- Now suppose (s)he performs 100 independent coin tosses of the same coin and reports ≥ 52 heads. Is (s)he serious or joking?
- What would be the answer if your friend claimed ≥ 52 heads out of 100 coin tosses, or 520 heads out of 1000 coin tosses?

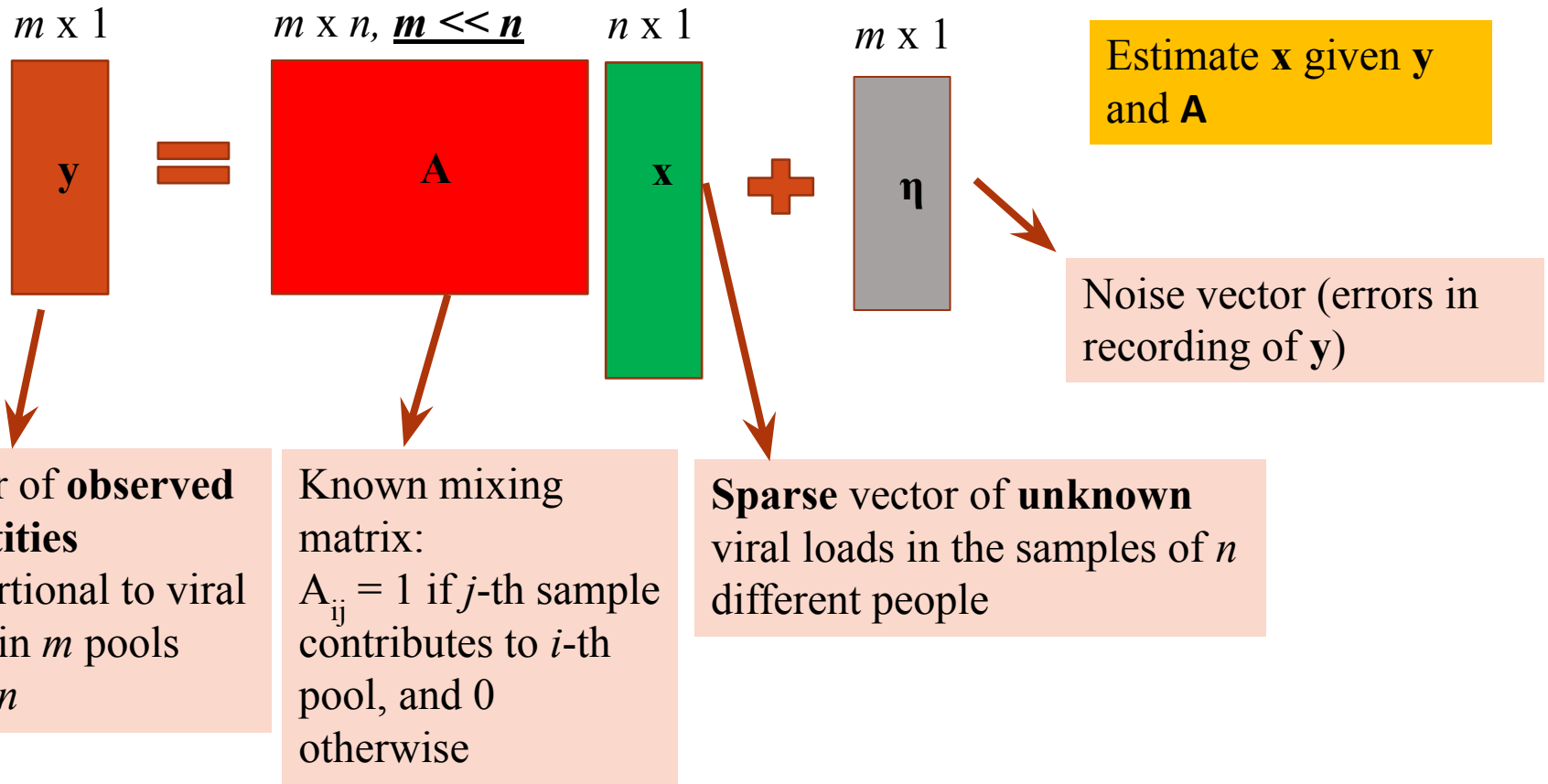
Application: COVID-19 testing

- The COVID-19 pandemic began in Wuhan, China.
- COVID-19 has infected more than 19 million people worldwide
- RT-PCR (*Reverse Transcription Polymerase Chain Reaction*) is the most popular method for testing a person for COVID-19
- Dearth of resources for widespread testing: time, skilled manpower, reagents, testing kits, etc.

Application: COVID-19 testing

- RT-PCR: naso- or oro-pharyngeal swab taken, mixed in liquid medium, tested in RT-PCR machine
- Can we **pool** (mix) subsets of n samples and **test the pools** to save resources?
- $\text{\#pools } (m) < \text{\#samples } (n)$
- Equal portions of participating samples are taken for creating any pool.
- A **negative pool test** implies all contributing samples are **negative (non-infected)** .
- **More work to be done if the pool tests positive (infected).**

Application: COVID-19 testing



Application: COVID-19 testing

- The previous slide shows a set of under-determined linear simultaneous equations, i.e. $\# \text{knowns} < \# \text{unknowns}$.
- Example: $a + b + c = 20$; $a - b = 3$ (2 equations in 3 variables).

$$\begin{pmatrix} 20 \\ 3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

- It turns out there are methods which can solve such systems uniquely, because the unknown vector \mathbf{x} is sparse (most entries are zero).

Application: COVID-19 testing

- The algorithms for this are of course way outside the scope of this course.
- We also do not know how many entries (say k) in \mathbf{x} are non-zero (we only know that number is small).
- If you could estimate k directly from \mathbf{y} and \mathbf{A} , then that is very useful information for our algorithms to obtain \mathbf{x} .
- It turns out that k can be estimated using properties of the so-called “binomial distribution”). How? That’s for later on in the course.

Course Information

- Instructors: Ajit Rajwade (first half) and Pushpak Bhattacharya (second half)
- Lecture venue: Online, timings: slot 8, Mon and Thurs 2 to 3:30 pm (i.e. immediately post lunch – and possibly post strong coffee ☺).
- Course webpage (for the first half):
http://www.cse.iitb.ac.in/~ajitvr/CS215_Fall2023/

Course Information

- Grading scheme:
 - ❖ 25% midterm (closed notes, most formulae will be provided)
 - ❖ 25% cumulative final exam (closed notes, most formulae will be provided)
 - ❖ 2-3 quizzes: 15% total
 - ❖ Team-based solving of programming and written assignments: 35% (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.**

Course Information

- We will make extensive use of MATLAB – in and out of class.
- MATLAB is available online via the IITB network using your LDAP id and password.
- Assignments will be posted on moodle.
- Course textbook: Introduction to Probability and Statistics for Engineers and Scientists: Fifth Edition; (used to be available online via the IITB network)
- Copies available in the library, book available on flipkart
- The material will cover lots of examples for each concept! I will cover many examples from medicine, social studies and image processing!

Course information

- We will make extensive use of moodle.
- There will be at least 2-3 tutorials before midsem and 2-3 after midsem.
- Tutorials will be conducted by TAs and will involve solving problems in class.

Course information

- In addition, we will have office hours immediately after class
- Other rules to follow:
 - ❑ Come to interaction session class **on time, 2:00 pm != 2:15 pm != 2:30 pm, etc.**
 - ❑ Submit homeworks (on moodle) **on time**
 - ❑ **Be interactive, ask questions** – in and out of class (after class, during office hours, over email, on moodle's discussion forum, etc.)