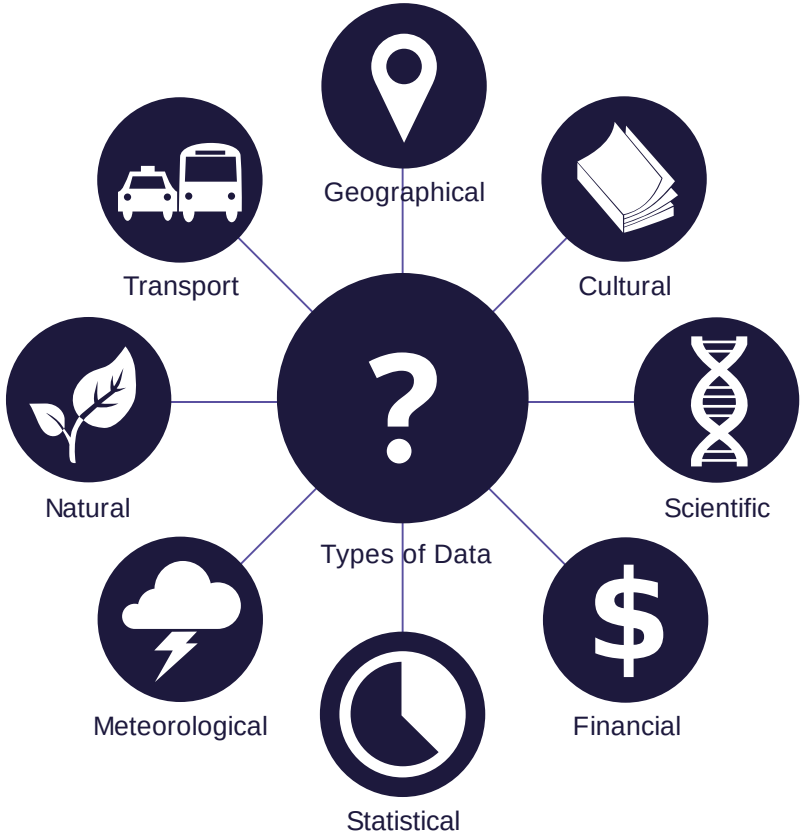
CS 215
Data Analysis and Interpretation

Introduction

Suyash P. Awate

Data

- “Data are units of **information**, often **numeric**, that are collected through **observation**.” [Wikipedia]
- In a formal/scientific context: “datum” is singular, “data” is plural
- Informally, "data" is typically used in the singular as a mass noun (like "sand" or "rain")
  - E.g., “there was so much rain/sand/data”
- Data is ubiquitous, across fields of study



Data Analysis

- Data analysis
  - Cleansing, transforming, and modeling data with the goal of data interpretation
  - Pipeline
    - Collect data, acquire data, make measurements
    - Curate data
    - Store data
    - **Process** data
    - **Visualize** data

Data Analysis

- Data analysis
  - Also called data interrogation

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

Data Analysis and Interpretation

- Data interpretation
  - Discovering useful information, informing conclusions, and supporting decision-making
  - Taking processed data, and extracting meaningful information
    - Gain insights and understanding into processes that generated data
  - Can use that understanding to perform tasks, e.g.,
    - Prediction
      - Will it rain tomorrow ?
      - Will this person get arthritis within the next N years ?
    - Categorization
      - Is this email spam ?
      - Is this cancer stage 1 or 2 or 3 or 4 ?
    - Test claims made about the underlying natural/physical processes
      - Claim: “In humans, natural aging leads to a reduction in volume of brain tissue”

Data Analysis and Interpretation

- Data interpretation

Data Analysis and Interpretation

- W. E. Deming
  - American engineer, statistician, professor, author, lecturer, management consultant
  - Helped develop sampling techniques still used by the U.S. Department of the Census
- Emphasizes the importance of data acquisition and analysis in decision making
- Data is evidence!

1. Data Collection, Acquisition

- Data **collection**: process of gathering and measuring information on variables of interest in a physical/natural system
  - Ways: Direct observation, surveys (open-ended or close-ended), interviews (1-on-1 or focus groups)
  - Fundamental research component in many study fields
    - Engineering, natural science (physical, life), social science, humanities, business
      - Natural science – a major branch of science that tries to explain, and predict, nature's phenomena based on **empirical evidence**
        - Empirical: "based on, concerned with, or verifiable by **observation** or **experience** rather than theory or pure logic."
  - Data **acquisition** (a term more specific to some engineering contexts): process of **sampling signals** that measure real world physical conditions and converting the resulting samples into **digital numeric** values that can be **manipulated** by a **computer**
    - e.g., recording images, voice, video, temperature, humidity, precipitation

2. Data Curation

- Organization and integration of data collected from various sources
  - Includes manual annotation, manual labeling/categorization (“meta data”)
    - Meta data = data about data
  - Tagging, to make things searchable (meta data)
  - Quality control
  - Documentation
  - Presentation
  - Formatting, to make it ready to be processed by computers or humans

2. Data Curation

- “Raw data”
  - Observed values before they’ve been "cleaned" and corrected by researchers
  - Why clean/correct ?
    - e.g., to remove/reduce measurement errors, fill missing values
  - If we have multiple stages in data processing, then processed data after one stage can be referred to as raw data for the next stage

3. Data Storage

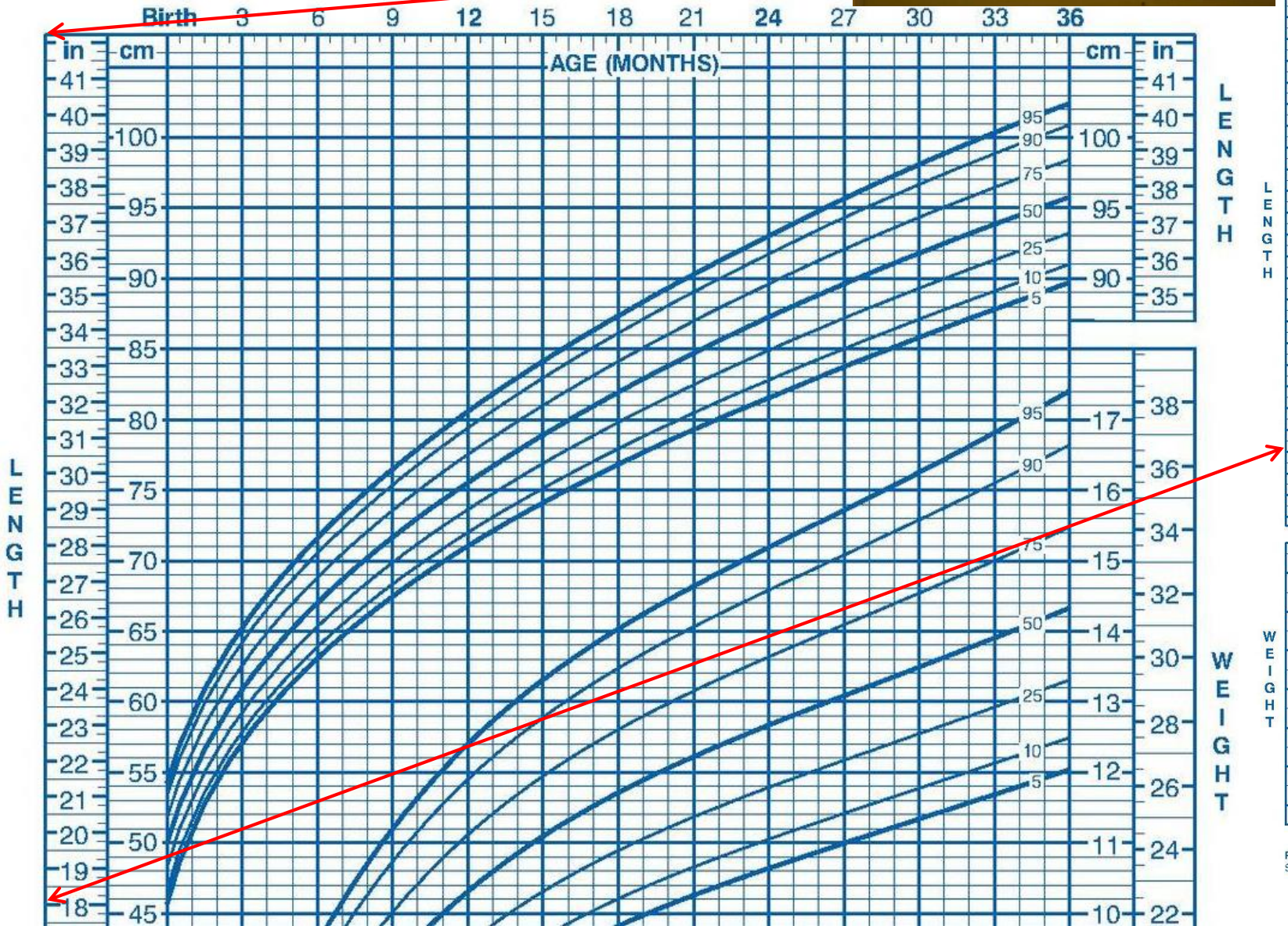
- This course isn't about data storage

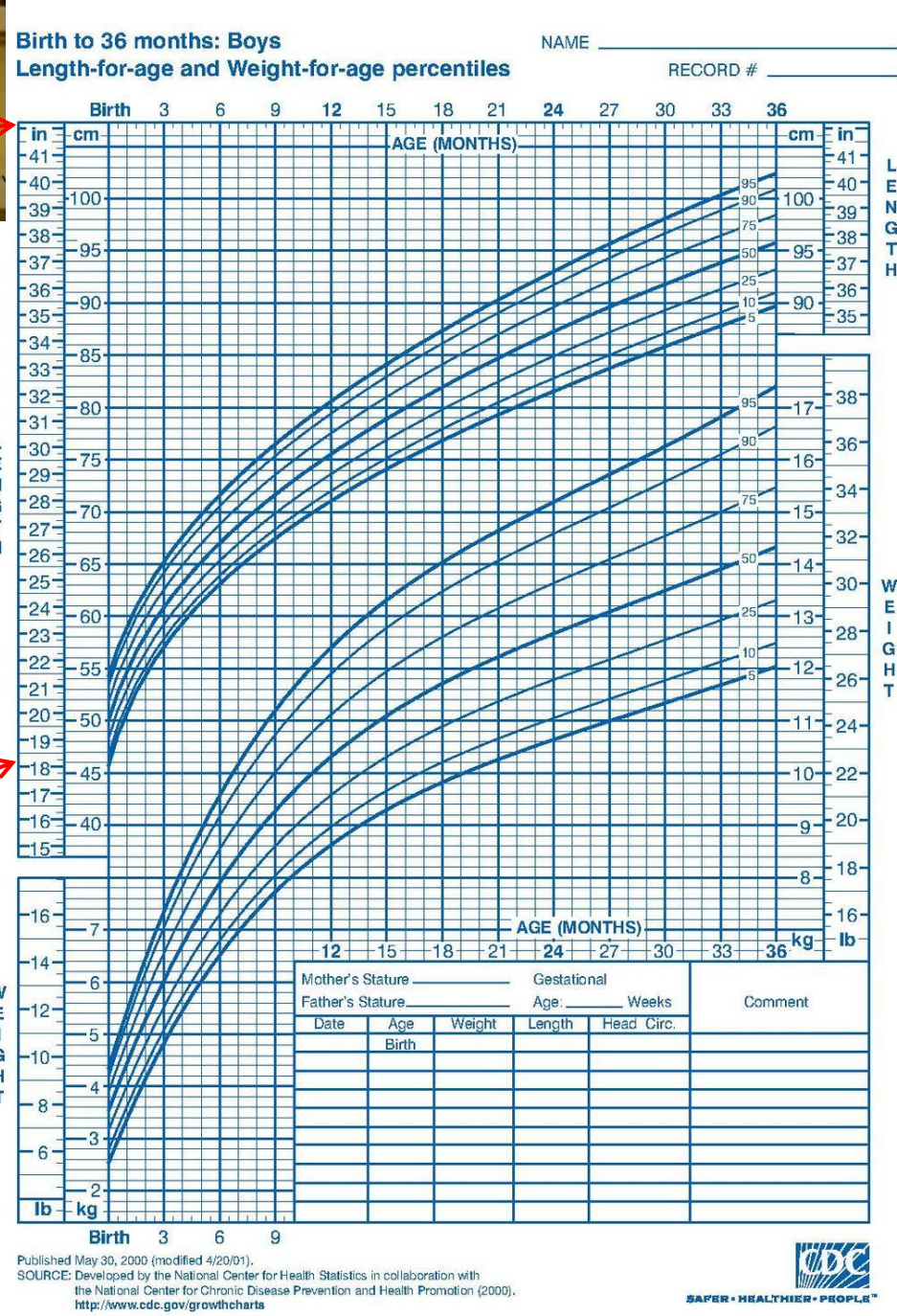
4. Data Processing

- Manipulation of items of data to produce meaningful information
- Can involve many processes
  - Sorting
  - Summarization (create a subset that is representative of the data)
  - Aggregation (combining multiple data sets into one)
  - Clustering (partitioning a data set into groups)
  - Quality enhancement
  - Classification
    - Assigning each datum into a semantically more abstract group
  - Regression
    - Estimating the relationship between
      a dependent variable (also called 'outcome', 'response' variable) and
      one or more independent variables (also called 'predictors', 'covariates', 'explanatory variables', 'features')

5. Data Visualization

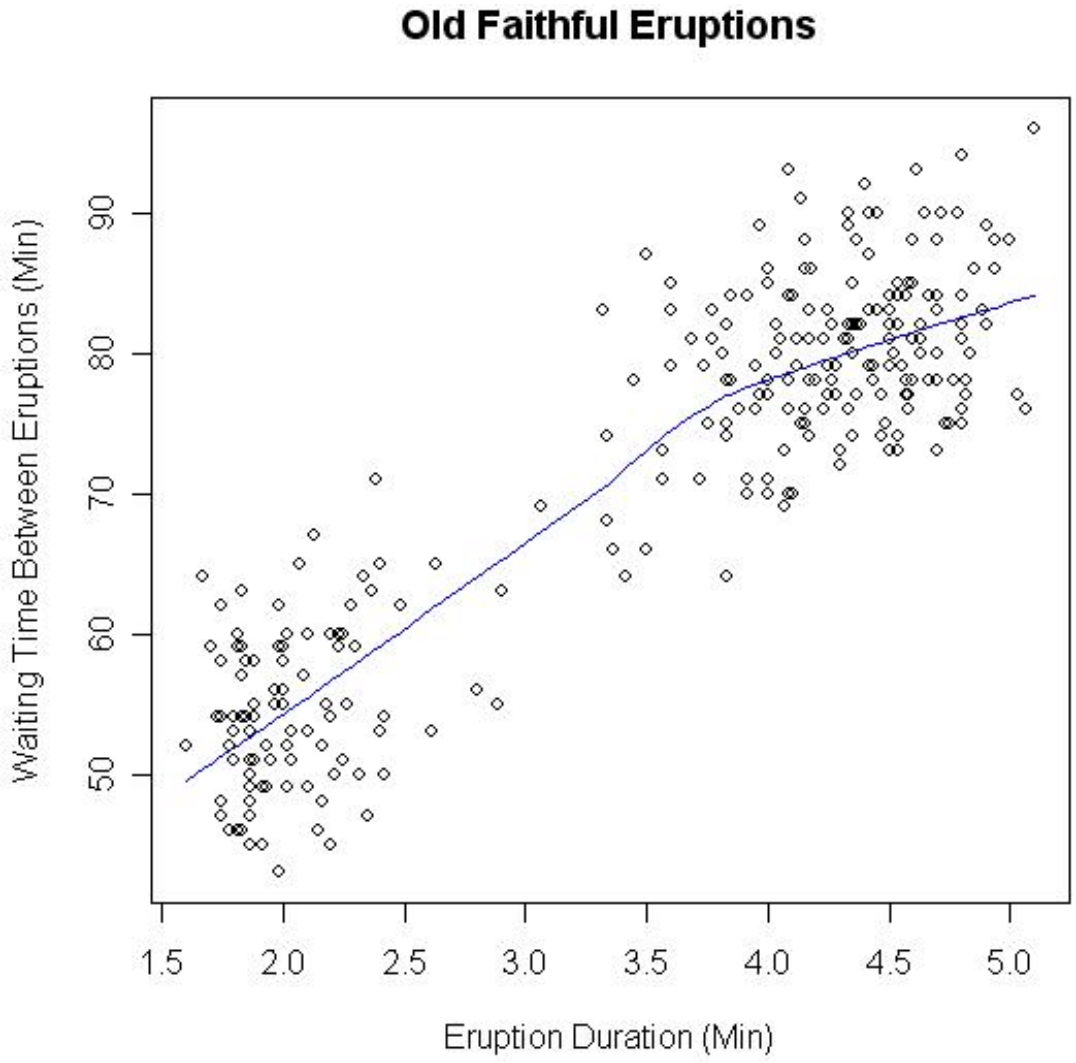
- Human growth





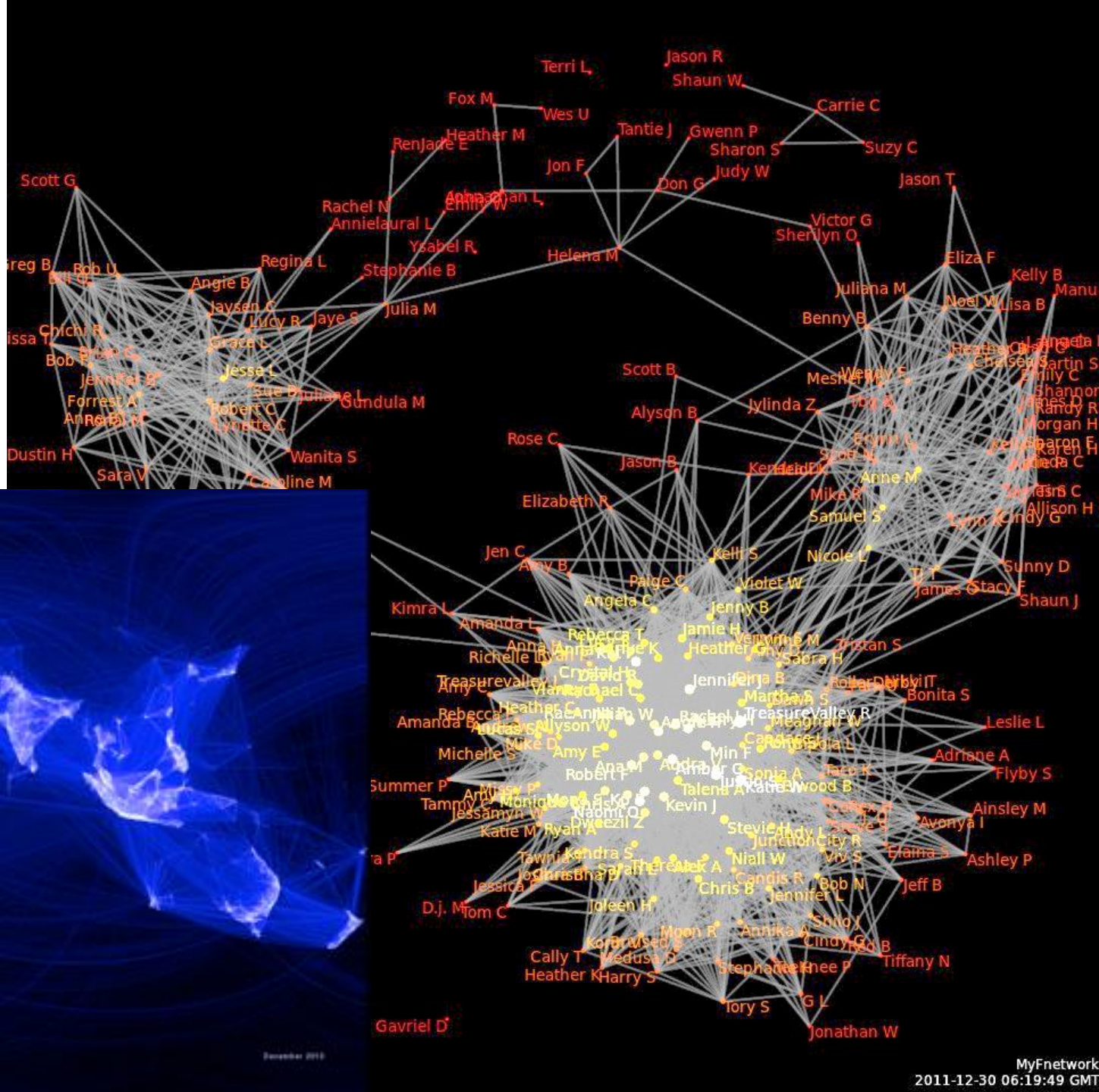
5. Data Visualization

- Geographical process
  - Eruption of the famed "old faithful" geyser
  - Data



5. Data Visualization

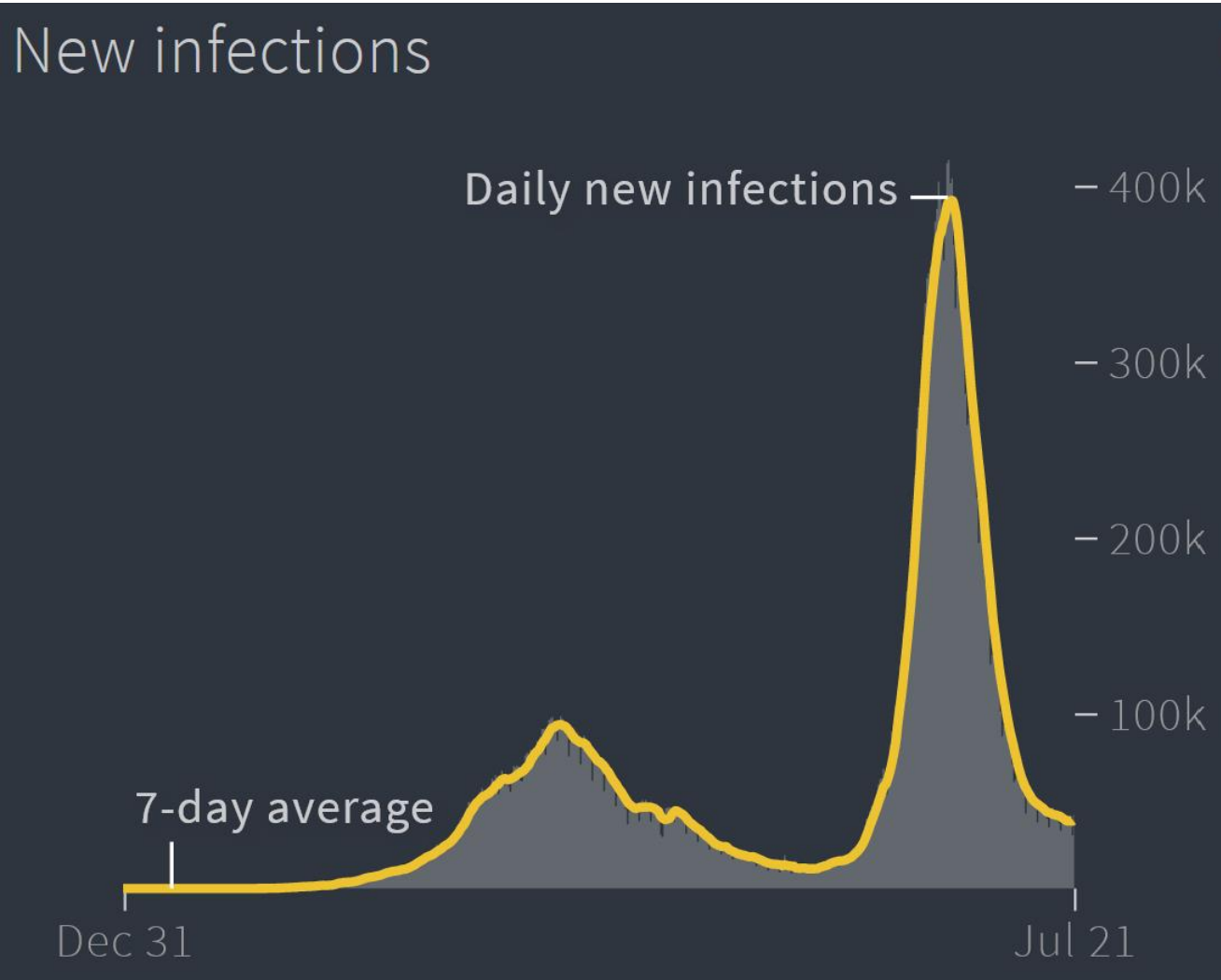
- Social media
  - Relationships on facebook

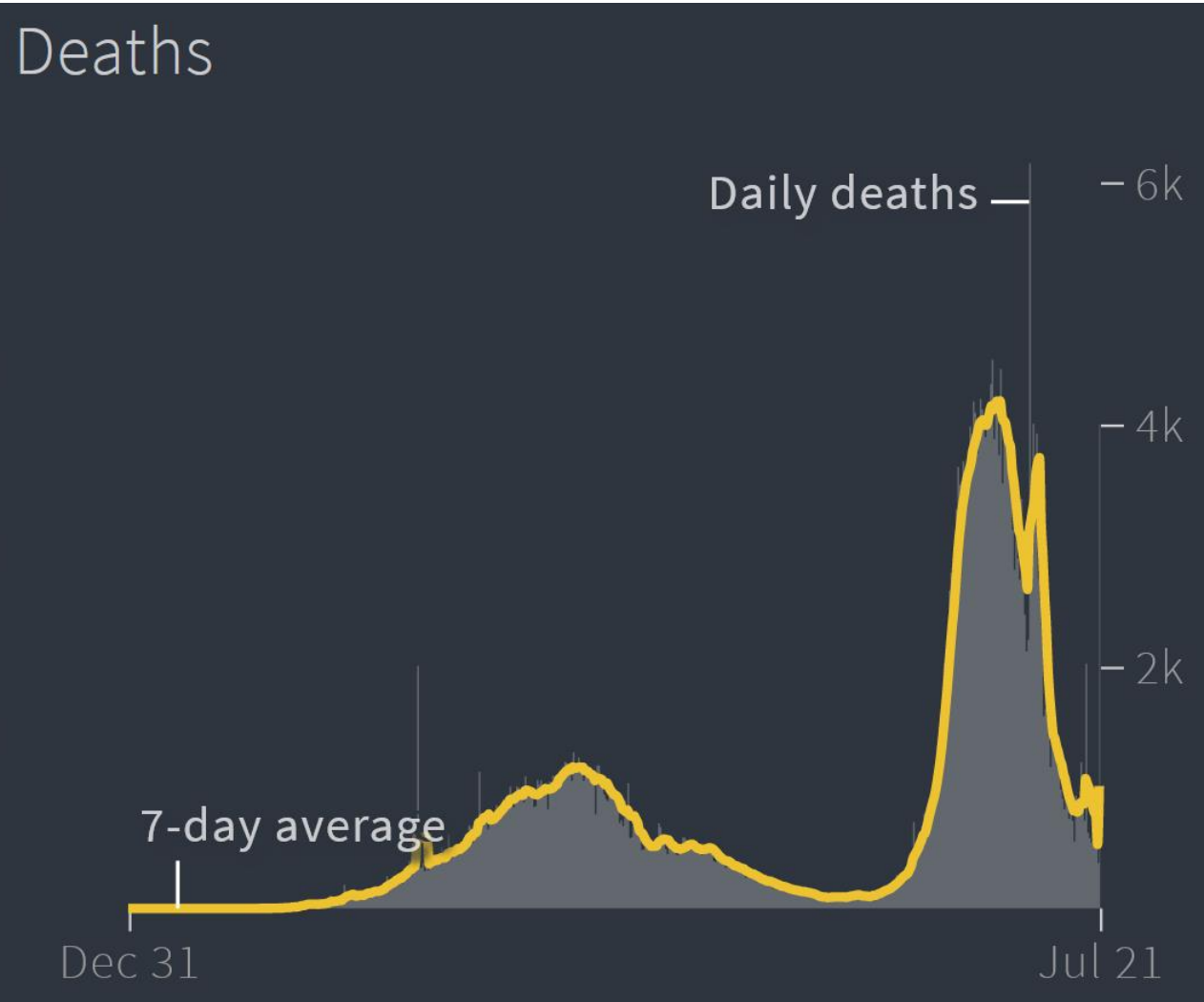


MyFnetwork
2011-12-30 06:19:49 GMT

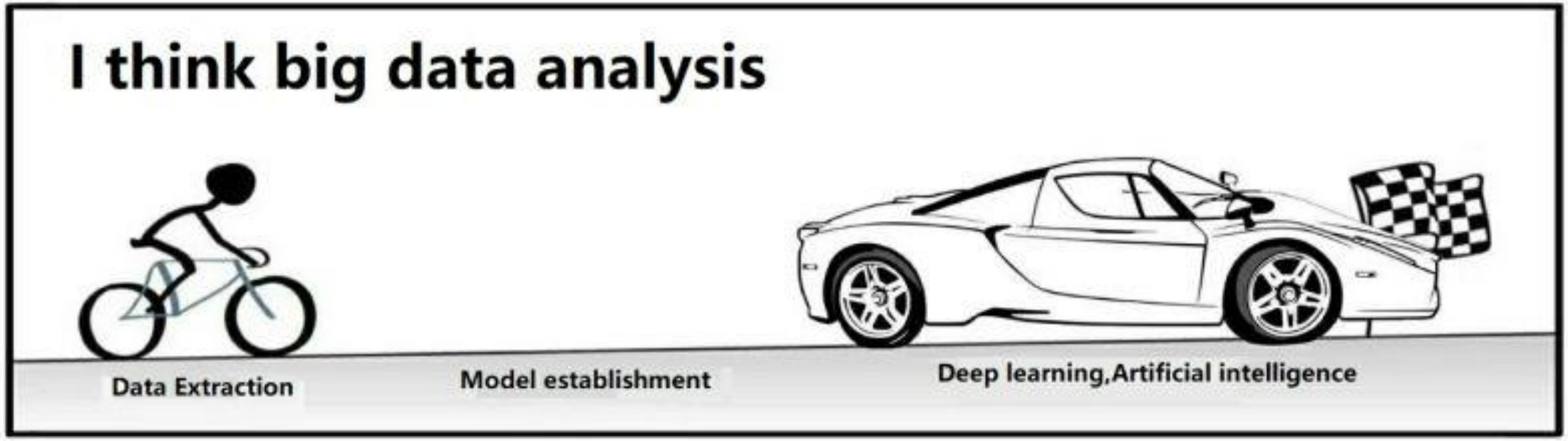
5. Data Visualization

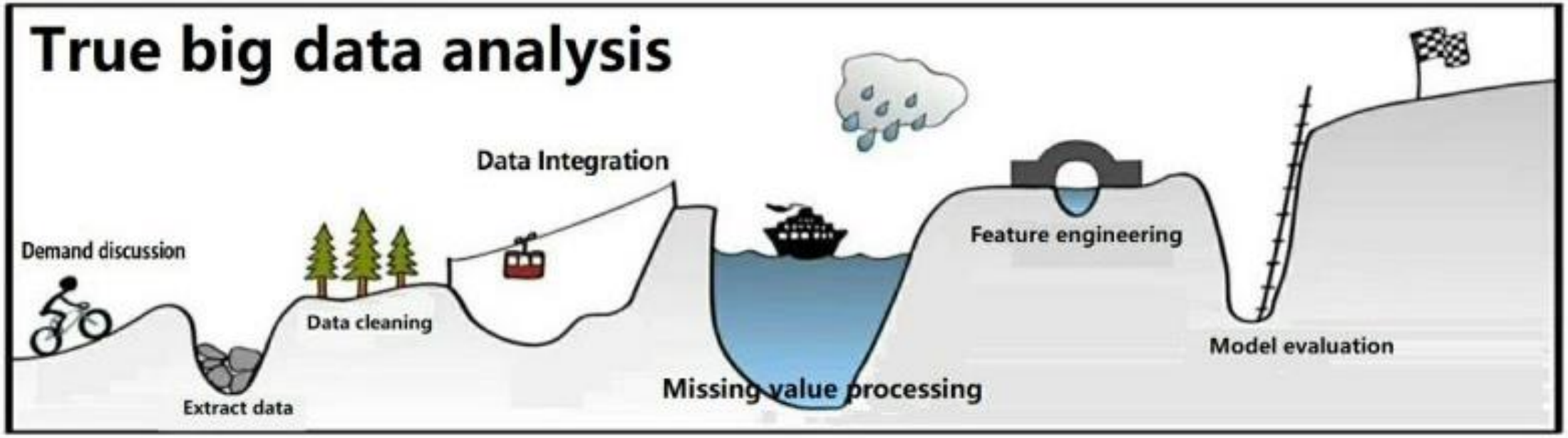
- Covid in India
  - https://graphics.reuters.com/world-coronavirus-tracker-and-maps/countries-and-territories/india/





Data Analysis and Interpretation



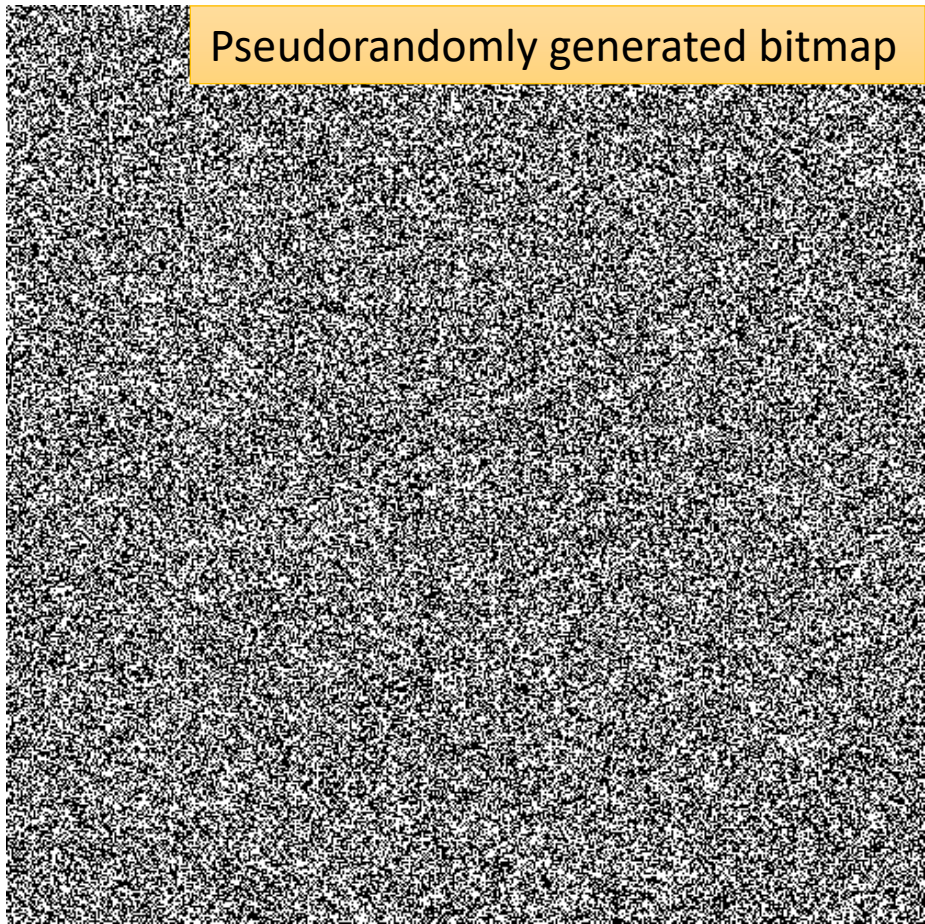


Data Characteristics

- Processes governed by a notion of **randomness**
  - Randomness is the apparent/actual lack of pattern or predictability in events
  - Individual events are, by definition, unpredictable,
    but a set of multiple events often exhibit specific trends

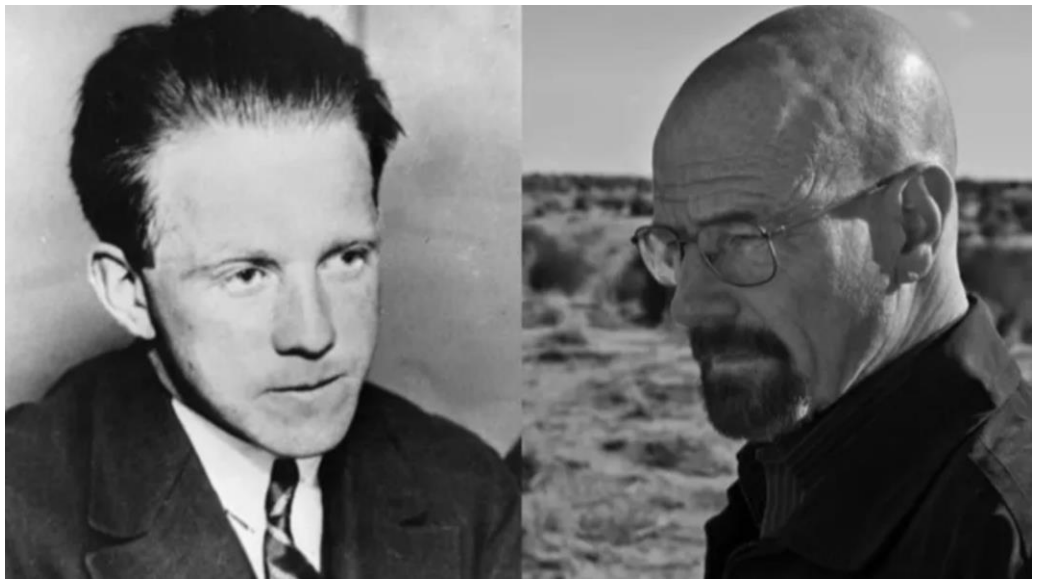
The ball in a roulette can be used as a source of apparent randomness, because its behavior is very sensitive to the initial conditions.

Pseudorandomly generated bitmap

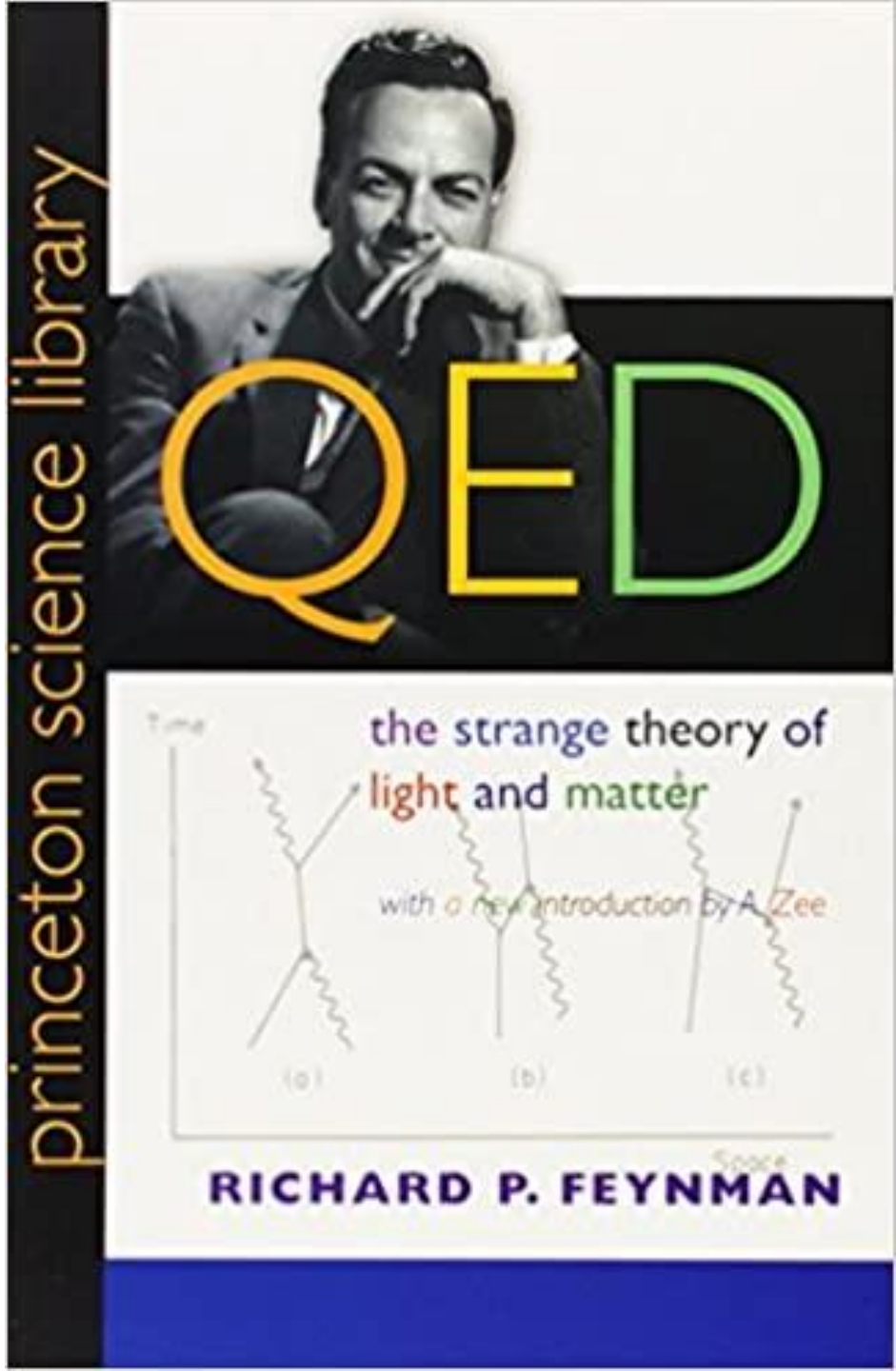


Data Characteristics

- Natural processes have randomness
  - Quantum physics
    - Heisenberg's uncertainty principle:
      fundamental limit to the accuracy with which values for certain pairs of physical quantities of a particle, such as position and momentum,
      can be predicted from initial conditions

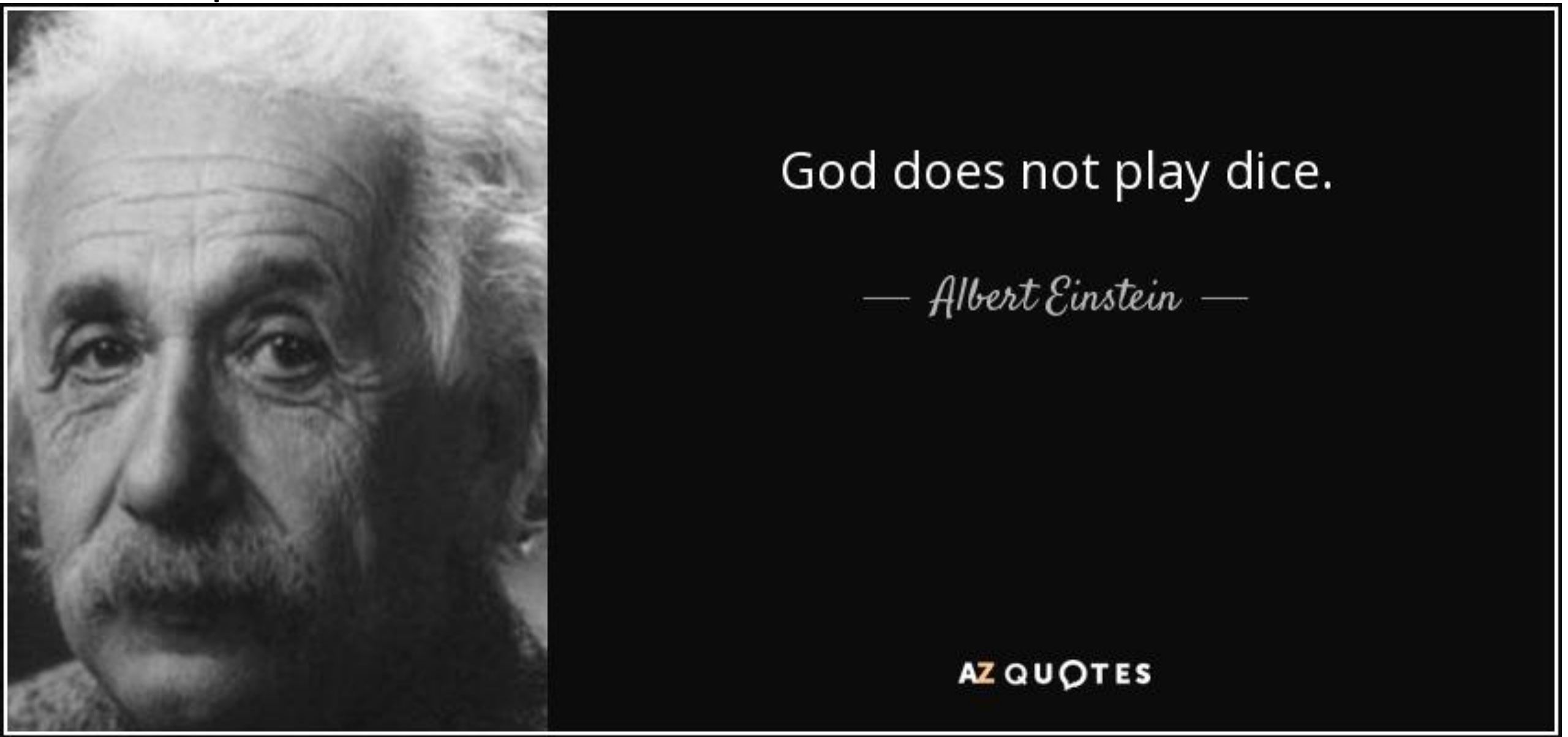


  - Emission of light/photons/radiation



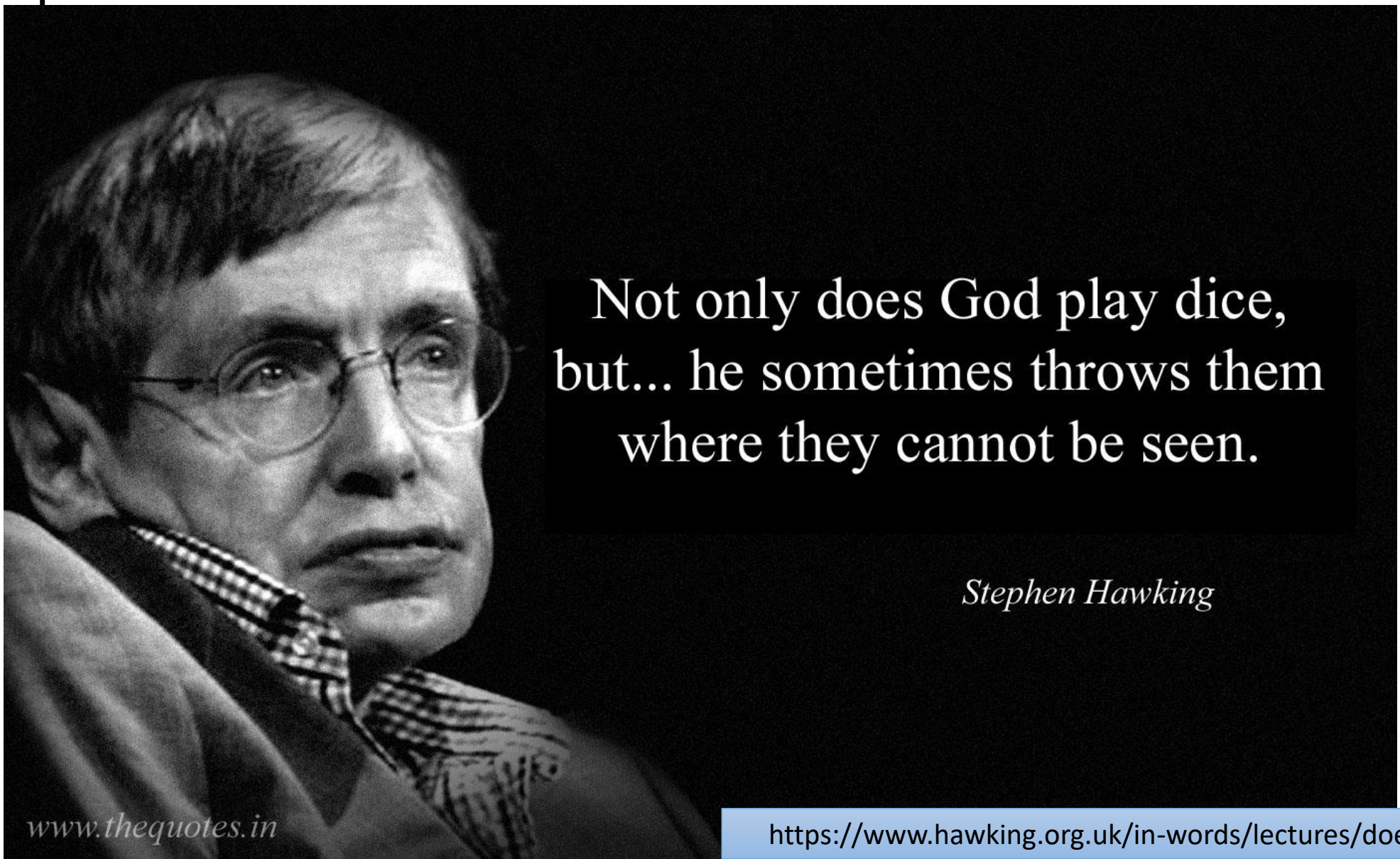
Data Characteristics

- Natural processes have randomness



Data Characteristics

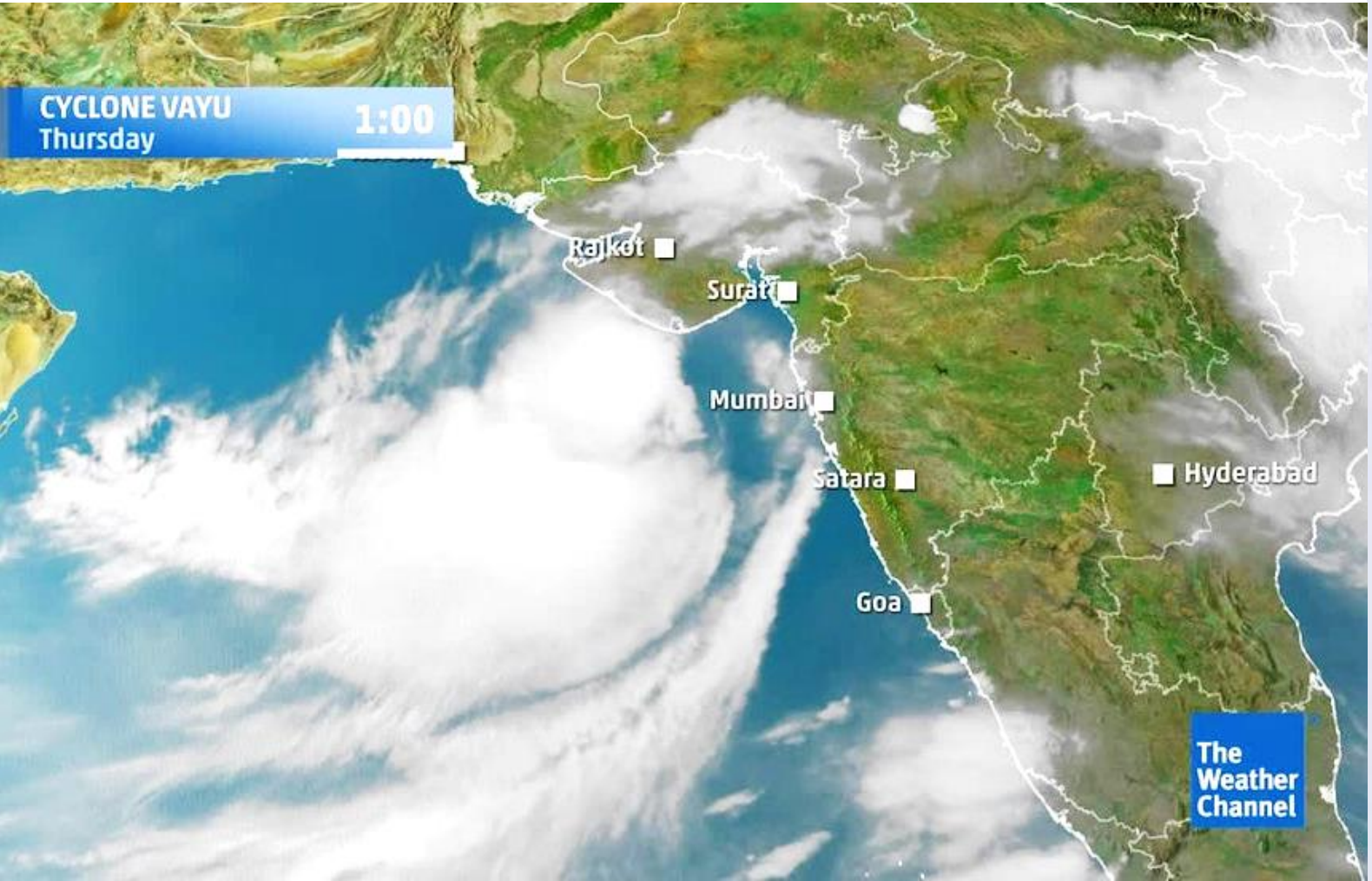
- Natural processes have randomness



https://www.hawking.org.uk/in-words/lectures/does-god-play-dice

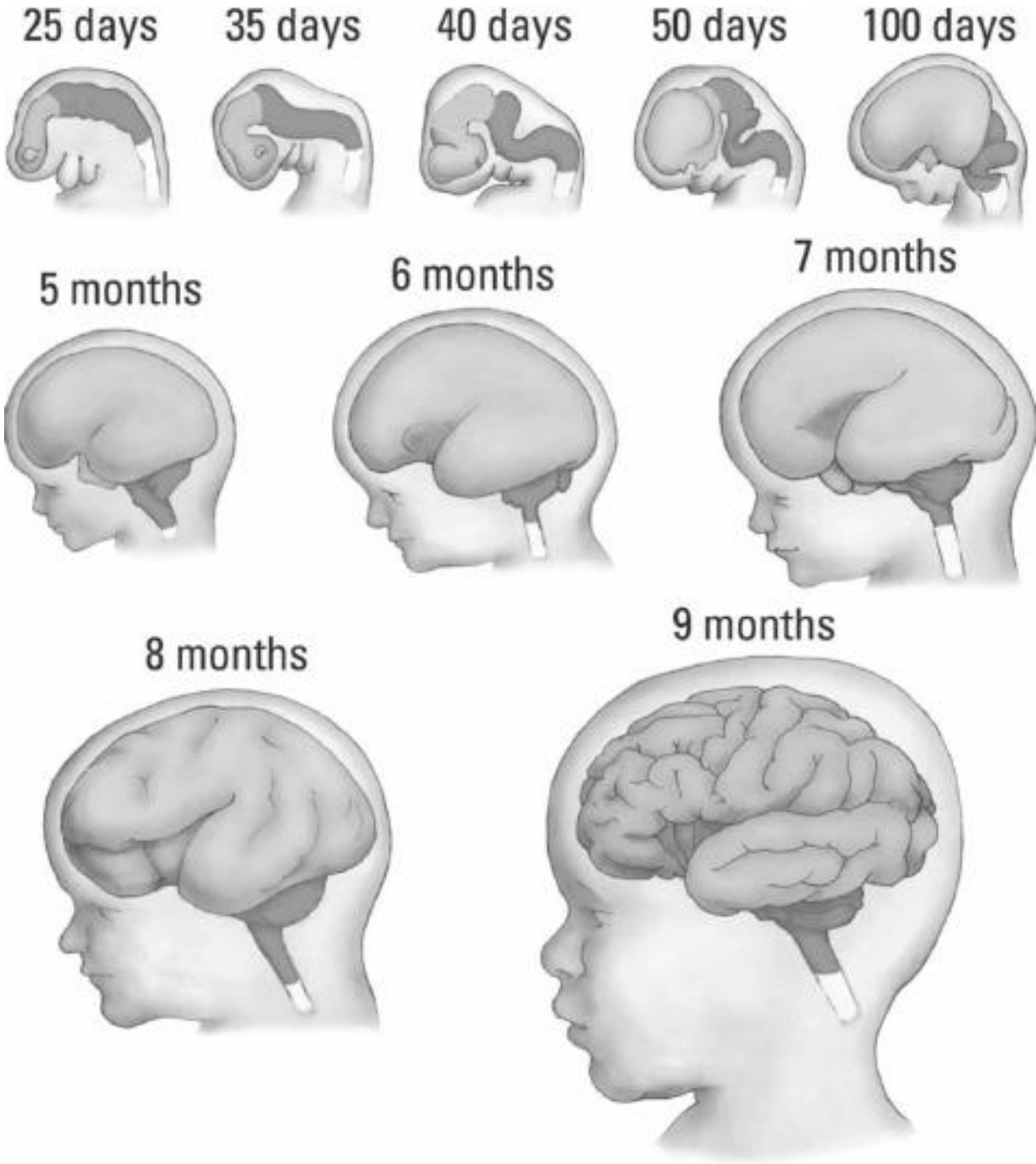
Data Characteristics

- Natural processes have randomness
  - Diffusion of fluids (liquids, gases)
  - Weather, climate



Data Characteristics

- Natural processes have randomness
  - Biological growth, mutation, evolution



Data Characteristics

- Other stochastic/random processes
  - Stock-market index (BSE SENSEX)

Data Characteristics

- Other stochastic/random processes
  - Traffic (road, air, internet)

Probability Theory, Statistics

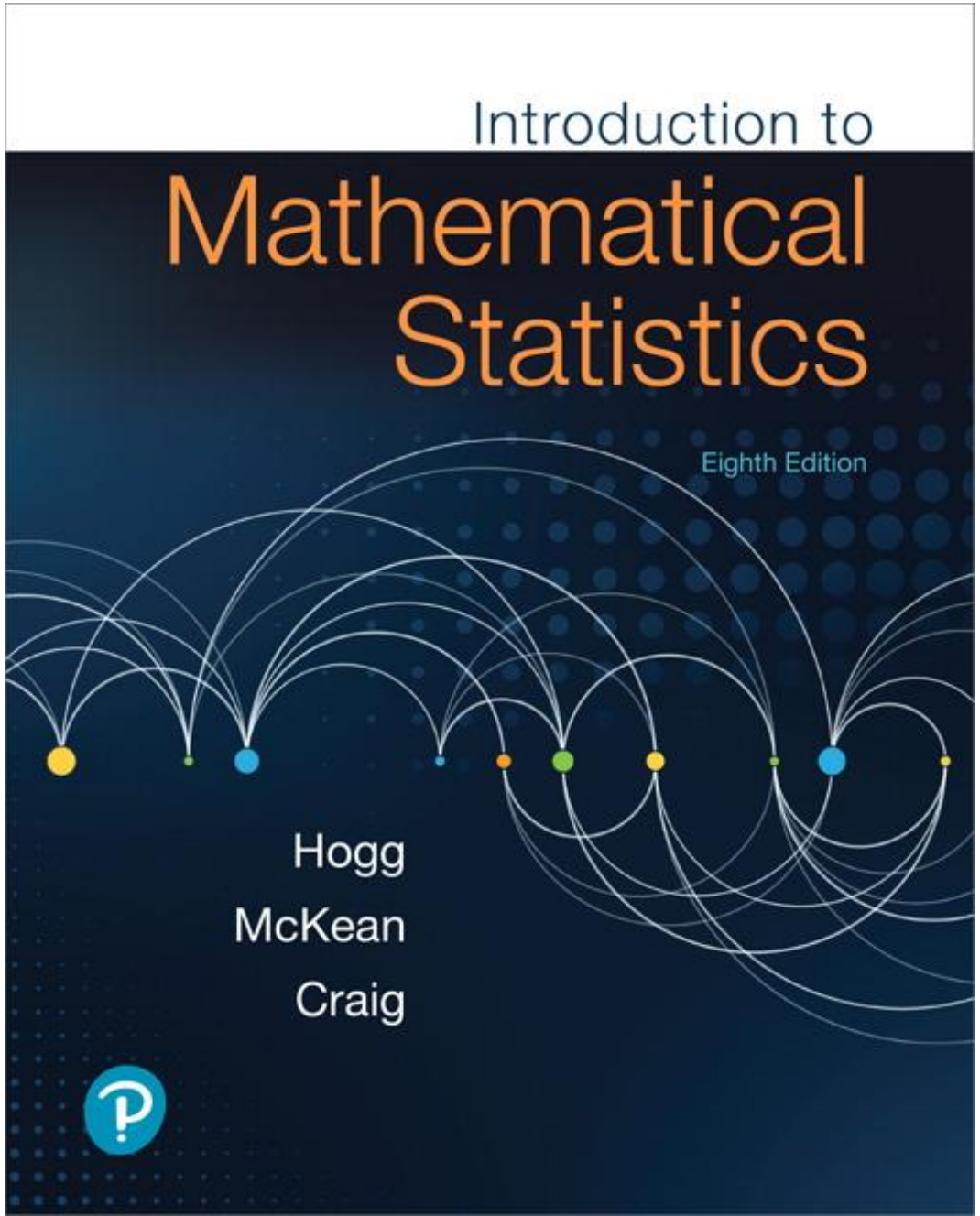
- Probability theory
  - Branch of mathematics concerned with probability
  - Mathematical foundation of statistics
- Statistics
  - Discipline that concerns the **collection**, **organization**, **analysis**, **interpretation**, and **presentation** of **data**
  - Statistical model
    - Mathematical (probabilistic) model that characterizes the process underlying data generation
  - Statistical inference
    - Process of using data analysis to infer properties of an underlying distribution of data
    - Use the inferred properties to make future predictions

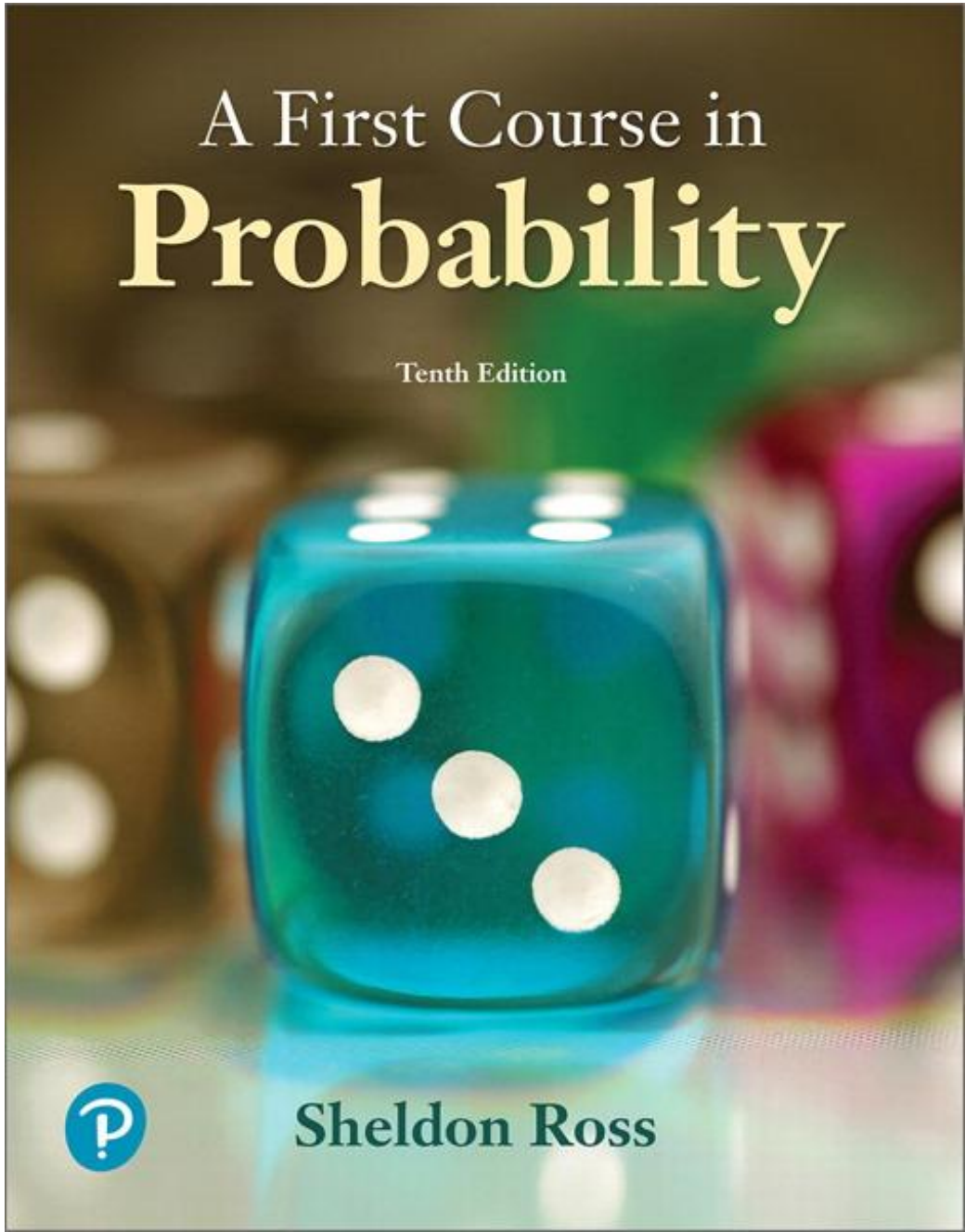
Mathematical Statistics

- Mathematical statistics (Wikipedia)
  - Application of probability theory, a branch of mathematics, to statistics, as opposed to techniques for collecting statistical data
  - Specific mathematical techniques which are used for this include:
    - Linear algebra
    - Mathematical analysis
    - Measure theory
    - Stochastic analysis
    - Differential equations

Text/Reference Books

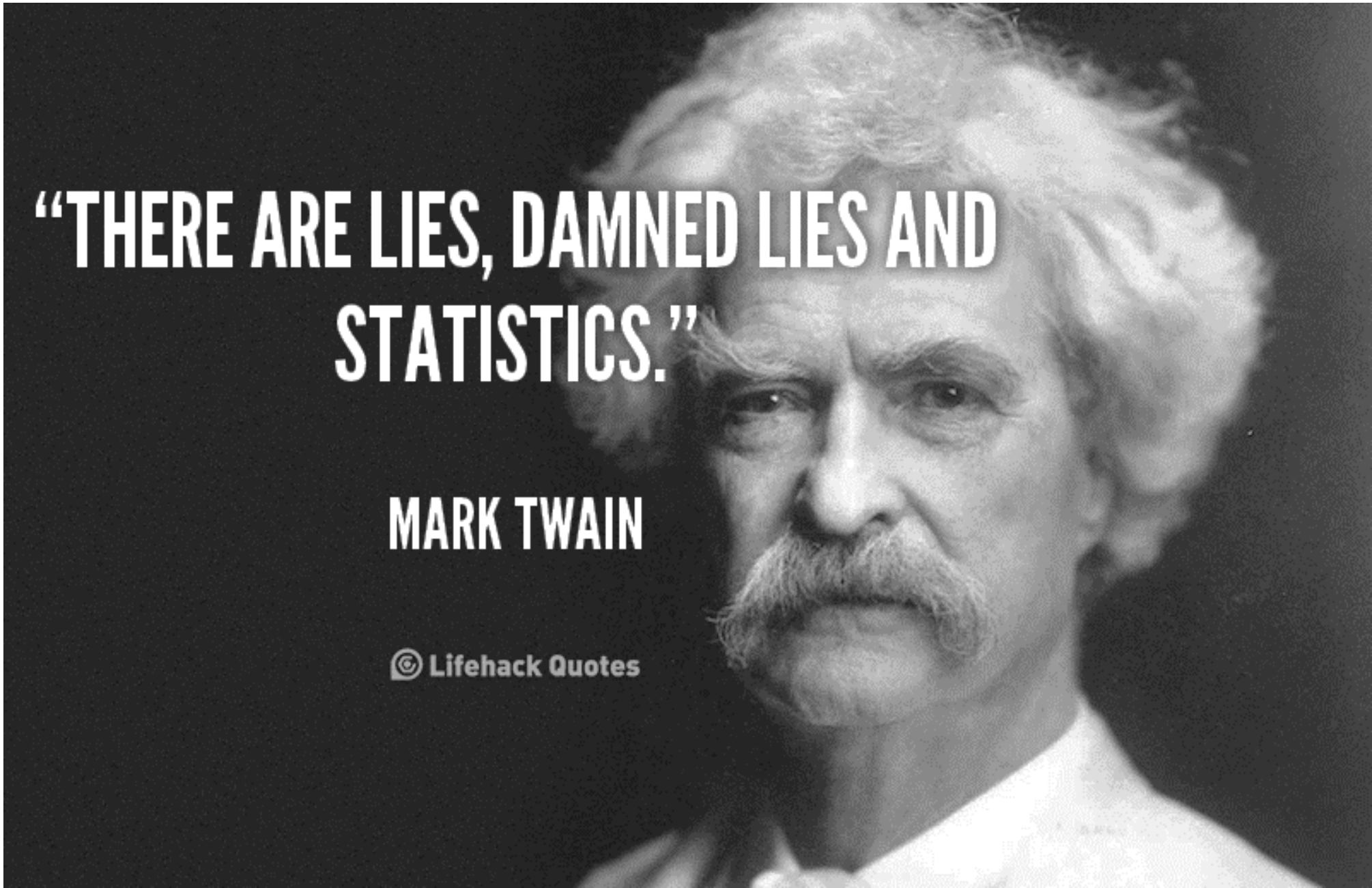
- Link 1
- Link 2
- Link 3
- Link 4
- Link 5
- Link 6
- Find your favorite book





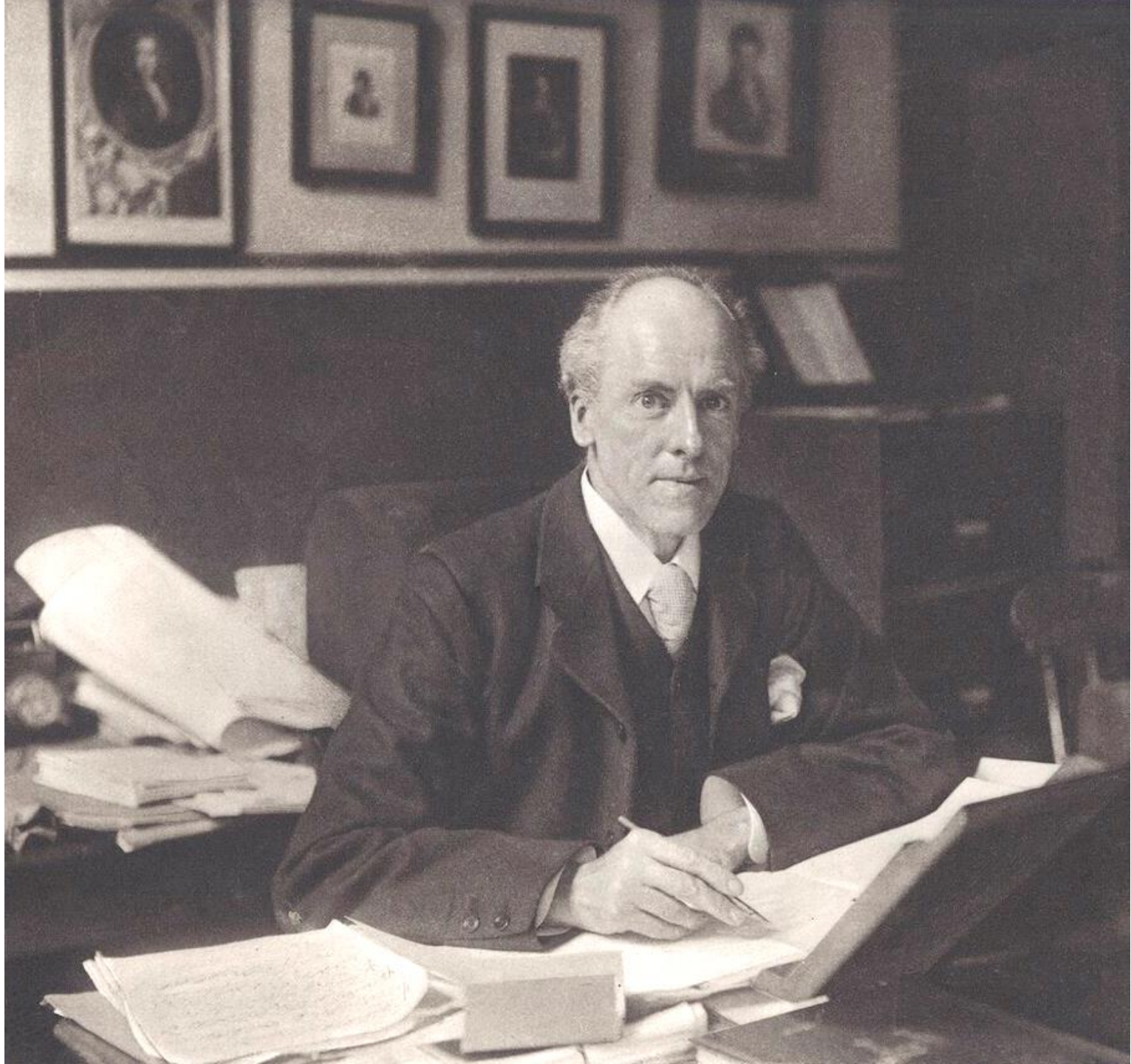
Statistics

- Mark Twain
  - 1835 – 1910
  - “Greatest humorist the USA has produced”
  - American literary great



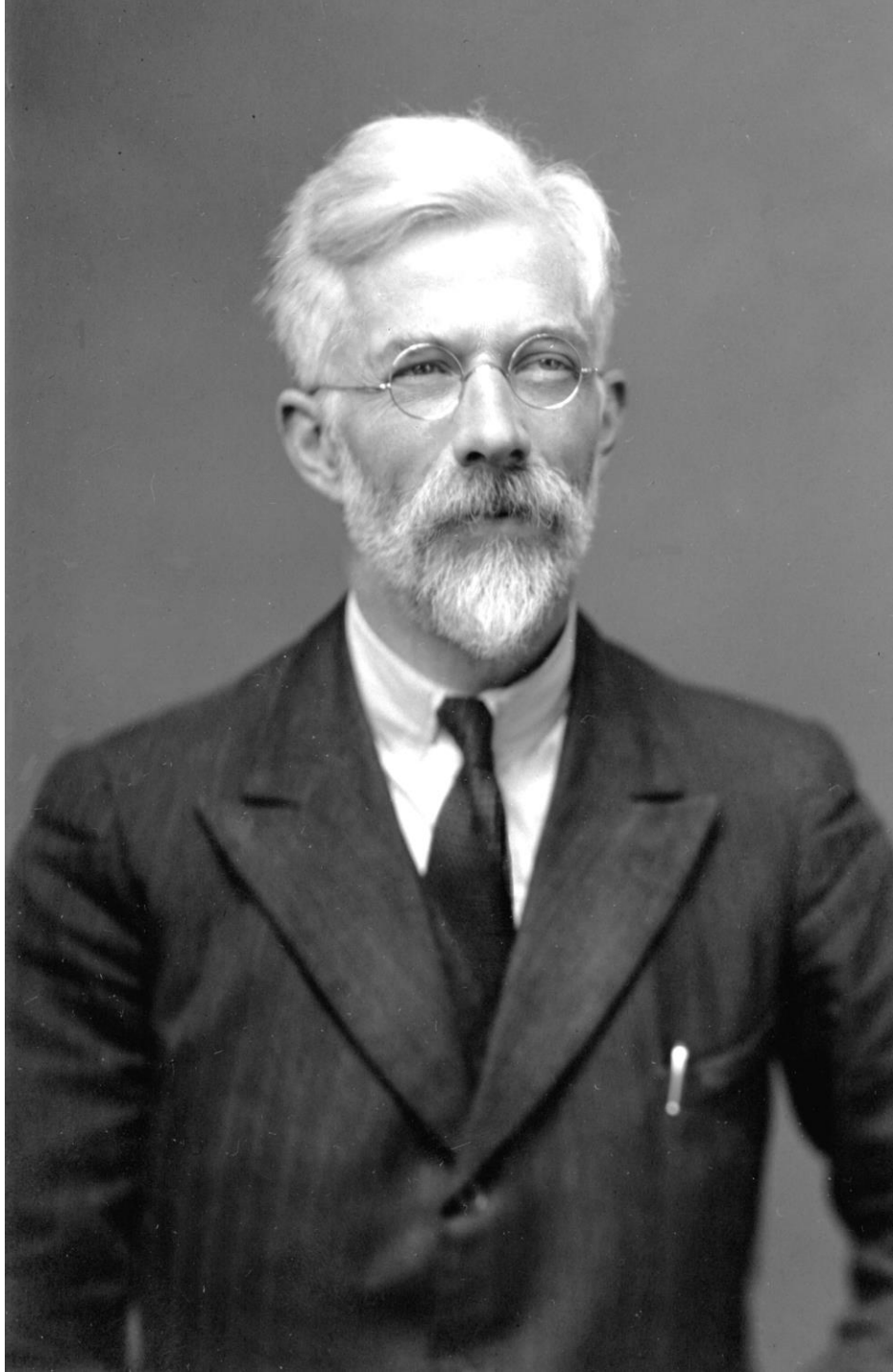
Statistics

- Karl Pearson
  - Established the discipline of mathematical statistics
  - 1857 – 1936
  - Advisor
    - Sir Francis Galton
  - Founded world's first university statistics dept. at University College London in 1911



Statistics

- Ronald Fisher
  - A genius who almost single-handedly created the foundations for modern statistical science
  - 1890 – 1962
  - Doctoral students
    - C. R. Rao



Topics

- Probability Axioms
- Conditional probability, Independence, Conditional independence
- Random variables
  - Probability space
  - Discrete random variables, Continuous random variables
  - Probability mass/density function, Cumulative distribution function
  - Several examples, interesting properties, insights
- Expectation, variance
- Joint random variable, marginal distribution, conditional distribution
- Covariance, correlation
- Estimation, likelihood function, bias and variance

Topics

- Transformation of random variables
  - Univariate
  - Multivariate
- Multivariate statistics
  - Multivariate Gaussian
    - Functional form, properties
  - Relationships with linear algebra
    - Singular value decomposition (SVD) of matrices
    - Eigen decomposition of matrices
- Principal component analysis (PCA)
  - Relies on covariances of multiple random variables
    - Needn’t be multivariate Gaussian
  - Special properties when data is multivariate Gaussian

Topics

- Bayesian statistics
  - Improving over maximum-likelihood estimation
  - Prior models
  - Some concepts from information theory
- Measuring quality of estimators
  - Fisher information
  - Cramer-Rao lower bound
  - Bayesian Cramer-Rao lower bound
- Linear regression, logistic regression, classification
- Kullback-Leibler divergence

Course

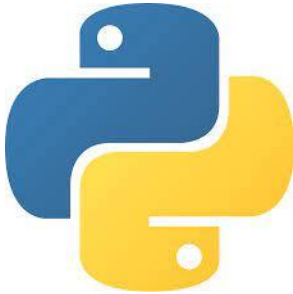
- Why take it ?
  - It is mandatory
- Why is it important ?
  - Courses on machine learning
  - Courses on data analysis
    - Images, text, audio, graphs, web, networks, …
  - Courses in engineering

Course Structure

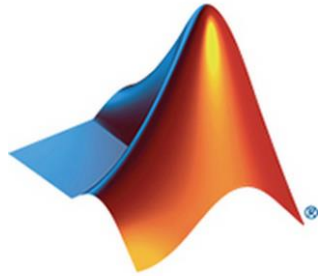
- Lecture slots: Tue + Fri, 3:30 pm – 5:00 pm
- Tutorial sessions (3–4)
  - 1-2 in first half. 2 in second half. Usually taken by UG teaching assistants.
- Assignments (3; take home; around 35% weight)
  - Mostly hands-on data analysis. Simulated and real-world data. Some theory.
  - Can use Matlab (like C), python, R, or your choice
  - Topics
    - Univariate statistics: modeling, estimation, simulation
    - Multivariate statistical modeling and analysis, PCA, simulation
    - Bayesian statistical modeling and analysis
- Mid-semester exam, End-semester exam (around 50% weight)
- 2 Quizzes: 1 in first half, 1 in second half (around 15% weight)
- Passing cut-off will be around 35%
- “around” means 20% is possible

Matlab, Python, R

- R
  - https://www.r-project.org/
  - For statistics computing, by statisticians (Ross Ihaka and Robert Gentleman)
    - At Univ. of Auckland



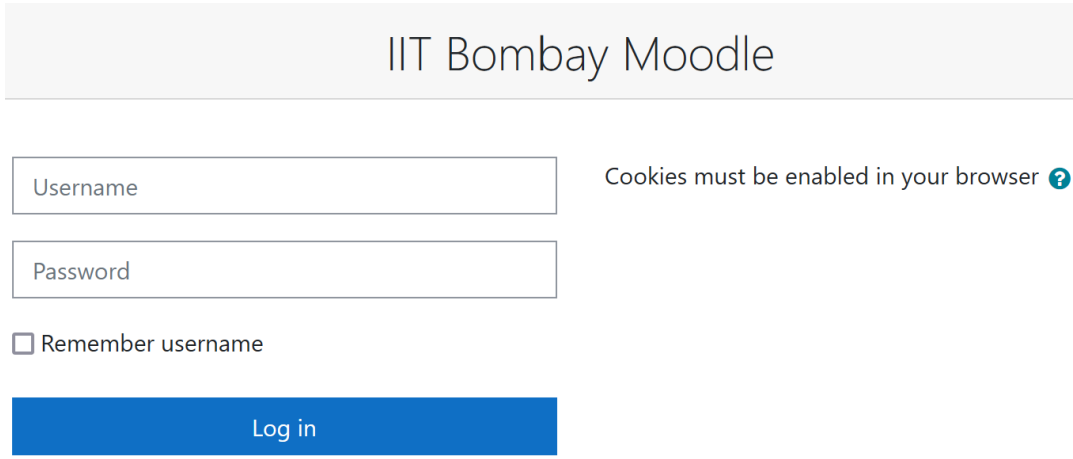
- Python
  - https://www.python.org/
  - https://realpython.com/matlab-vs-python/



- Matlab
  - https://www.mathworks.com/products/matlab.html
  - Numeric computing
  - https://www.mathworks.com/discovery/matlab-vs-r.html
  - https://www.mathworks.com/products/matlab/matlab-vs-python.html
- https://medium.com/swlh/python-vs-r-vs-matlab-for-machine-learning-causal-inference-signal-processing-and-more-b837a988c674

Course Structure

- Interactions
  - MS Teams
    - https://www.microsoft.com/en-in/microsoft-teams/group-chat-software
    - Slides + video (around 2 hours) uploaded at start of each week
    - 1 weekly meeting for Q&A in the allotted time-slot
      - Friday 3:30-4:55 pm ?
      - Students must go through (in detail; not casually) all content **before** the online meeting
      - Use this meeting to clarify queries. Feel free to ask anything.
    - Proctored online exams, also including SAFE
  - Moodle
    - https://moodle.iitb.ac.in/login/index.php
    - Receiving and submitting assignments
    - Submitting exams
      (in addition to SAFE, https://safe.cse.iitb.ac.in/)
    - Receiving grades, comments



Matlab

- Matlab Tutorial 1
  - www.mccormick.northwestern.edu/documents/students/undergraduate/introduction-to-matlab.pdf
- Matlab Tutorial 2
  - http://web.eecs.umich.edu/~aey/eecs451/matlab.pdf
- Matlab Tutorial 3
  - www.cs.cmu.edu/~ggordon/10601/recitations/matlab/pretty_matlab_pres.pdf
- Matlab Image-Processing Toolbox Tutorial
  - http://www.cs.otago.ac.nz/cosc451/Resources/matlab_ipt_tutorial.pdf
- Writing fast matlab code
  - http://www.csc.kth.se/utbildning/kth/kurser/DN2255/ndiff13/matopt.pdf
- Matlab Tips and Tricks
  - http://www.ee.columbia.edu/~marios/matlab/mtt.pdf

Course Structure

- Attendance policy
  - www.iitb.ac.in/newacadhome/ugrulebook201902Dec.pdf
    - "IIT Bombay expects one hundred percent attendance (100%) from its students in all classes."
    - "If the attendance of the student, as counted with effect from the first contact hour, falls below eighty percent of the total attendance expected, the 'DX' grade will be awarded in that course."
    - "Only exception to this rule are courses where the instructor has declared that no DX grade will be awarded."
    - Medical certificate required from IITB hospital, or one duly authenticated by IITB Hospital.

Course Structure

"You have to attend classes. You can't just follow me on Twitter."

Course Structure

- Assignments
  - Assignments can be done alone or in groups (upto 2 students)
  - If done in a group, each student in the group must solve each and every question, independently or jointly
    - e.g., cannot split work as:
      student A solves questions 1 and 4
      student B solves questions 2 and 3
    - This is important for effective learning
  - Groups once formed cannot be split
    - Unless there is a valid emergency
  - Plagiarism policy
    - Plagiarism penalty for any student will apply to all students in the group

Course Structure

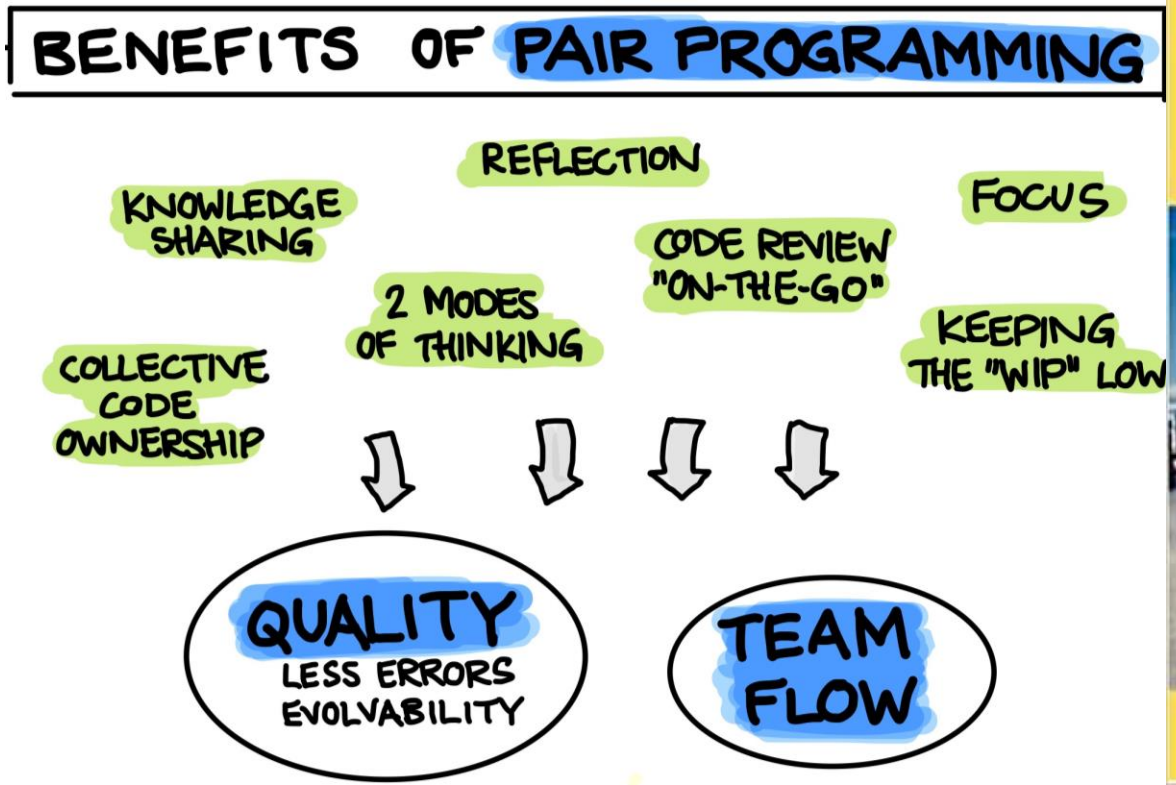
- Assignments
  - Groups of 2
  - 2 programmers work together at one workstation
  - “Driver” writes the code. “Navigator” reviews each line as it is typed in.
  - Programmers switch roles frequently

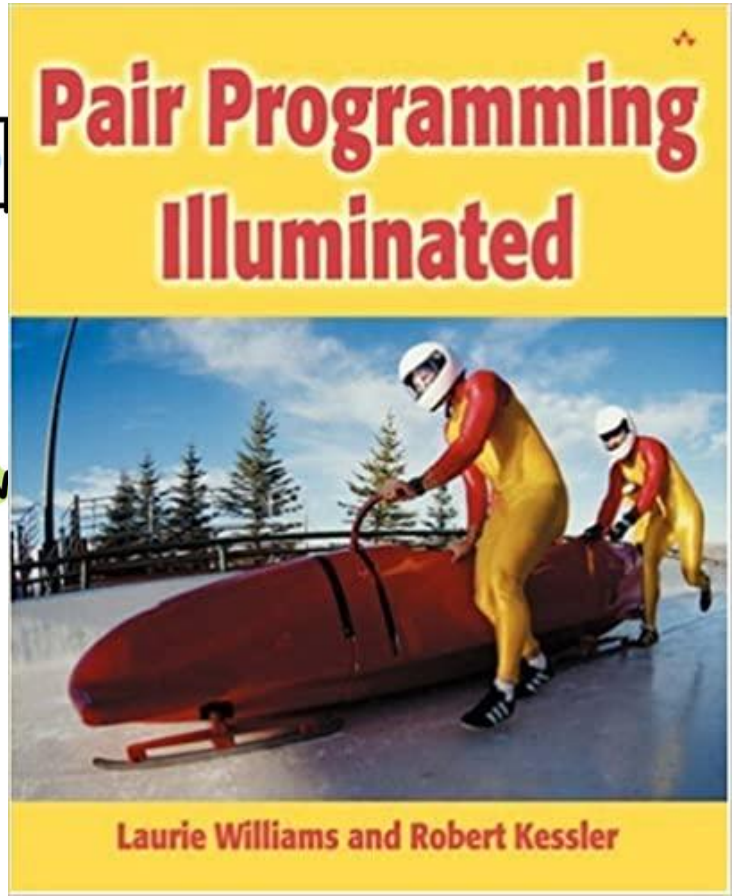
RESEARCH-ARTICLE

Strengthening the Case for Pair Programming

Authors: Laurie Williams, Robert R. Kessler, Ward Cunningham, Ron Jeffries Authors Info & Affiliations

IEEE Software, Volume 17, Issue 4 • July 2000 • pp 19–25 • https://doi.org/10.1109/52.854064





Course Structure

- Assignments

Course Structu

- Plagiarism

Course Structure

- Plagiarism policies
  - Institute policy www.iitb.ac.in/newacadhome/punishments201521July.pdf
    - "A student found copying in an assignment/laboratory project is given a zero in the assignment/project and is further given a one grade penalty."
    - "The same disciplinary action is taken against both the person copying and the person from whom the material was copied."
  - CSE policy
    - Grader (teaching assistant or instructor) will report case to Departmental Academic Disciplinary Action Committee (DADAC)
    - At least loss of 1 grade level
      - Can get a fail grade

Course Structure

- Plagiarism policy
  - https://integrity.mit.edu/handbook/writing-code
  - "…, during the time that you are helping another student, your own solution should not be visible, either to you or to them. Make a habit of closing your laptop while you're helping."
  - "As they type lines of code, they speak the code aloud to the other person, to make sure both people have the right code. INAPPROPRIATE."
  - "In a tricky part of the problem set, Alyssa and Ben look at each other's screens and compare them so that they can get their code right. INAPPROPRIATE."

Course Structure

- Plagiarism policy
  - https://integrity.mit.edu/handbook/writing-code
  - “Jerry opens his own laptop, finds his solution to the problem set, and refers to it while he’s helping Ben correct his code. INAPPROPRIATE.”
  - “Ben has by now spent a couple hours with Louis, and Louis still needs help, but Ben really needs to get back to his own work. He puts his code in a Dropbox and shares it with Louis, after Louis promises only to look at it when he really has to. INAPPROPRIATE.”
  - “They exchange their test cases with each other to check their work. INAPPROPRIATE. Test cases are part of the material for the problem set, and part of the learning experience of the course. You are copying if you use somebody else’s test cases, even if temporarily.”