Monsoon 2020

1 - Introduction

Information Retrieval

by

Dr. Rajendra Prasath

Indian Institute of Information Technology
Sri City – 517 646, Andhra Pradesh, India

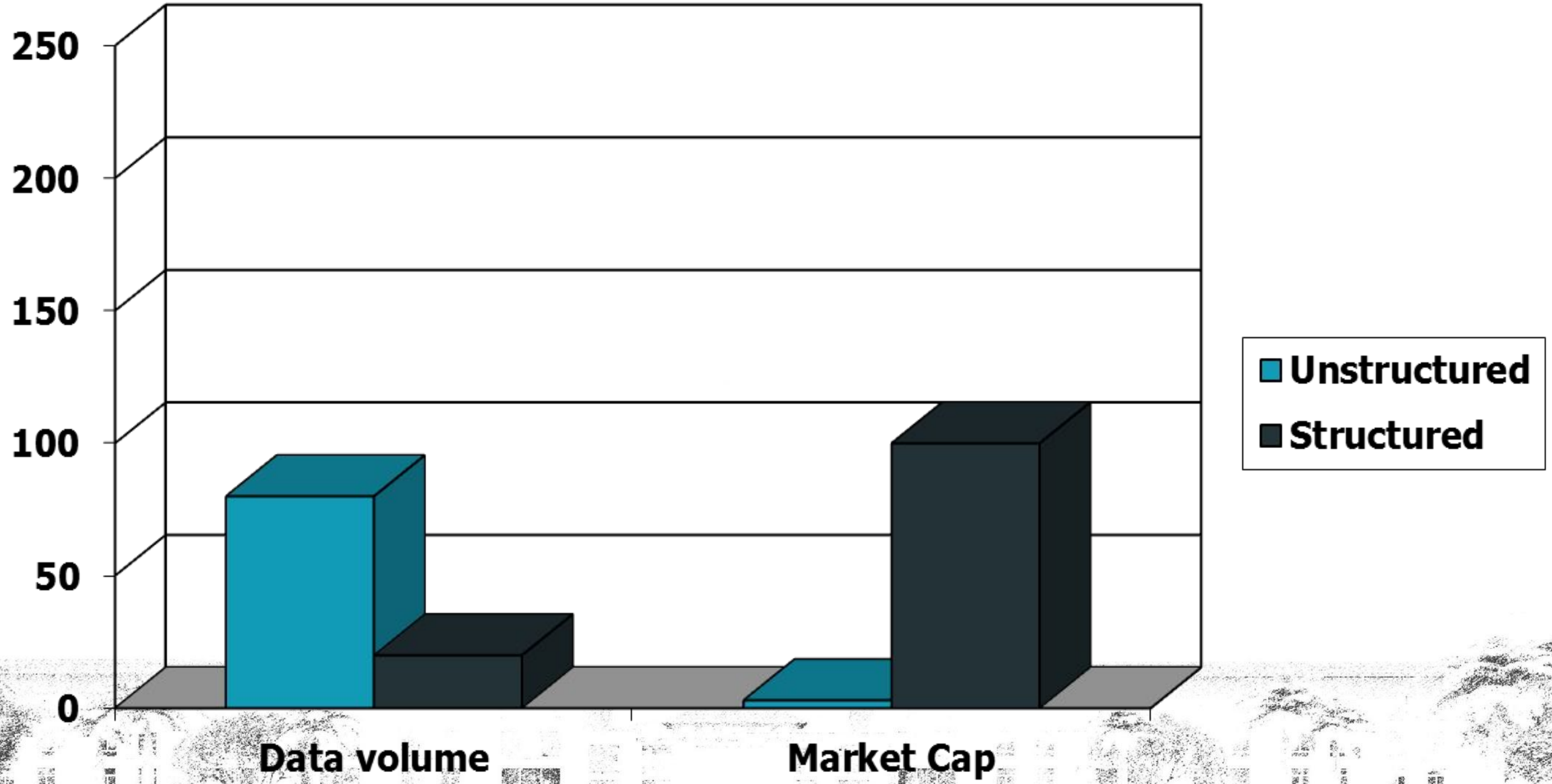
What is Information Retrieval?

Information Retrieval

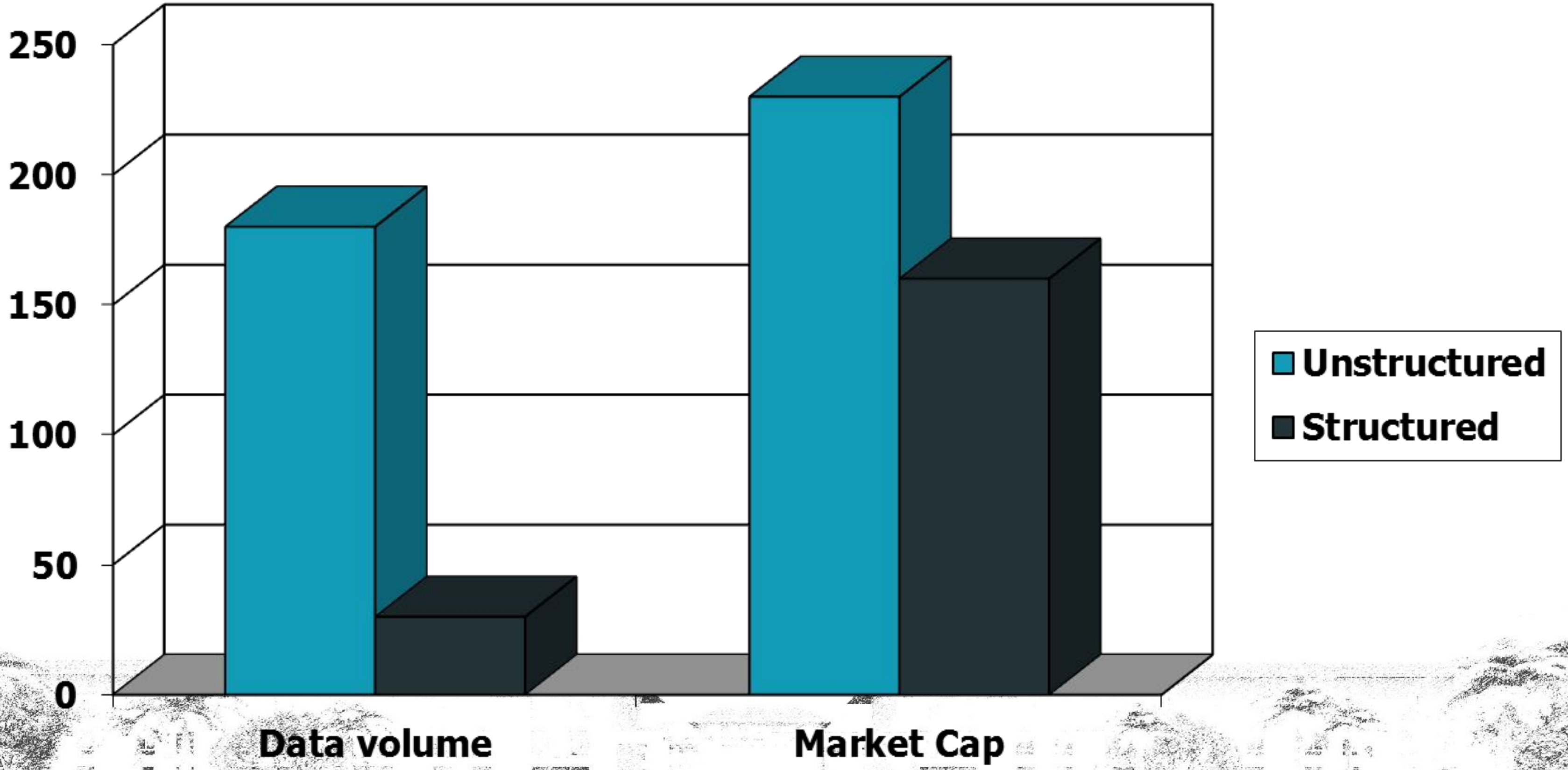
- **Information Retrieval (IR)** is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

- These days we frequently think first of web search, but there are many other cases:
  - E-mail search
  - Searching your laptop
  - Corporate knowledge bases
  - Legal information retrieval and so on . . .

Unstructured (text) vs. structured (database) data in the mid-nineties



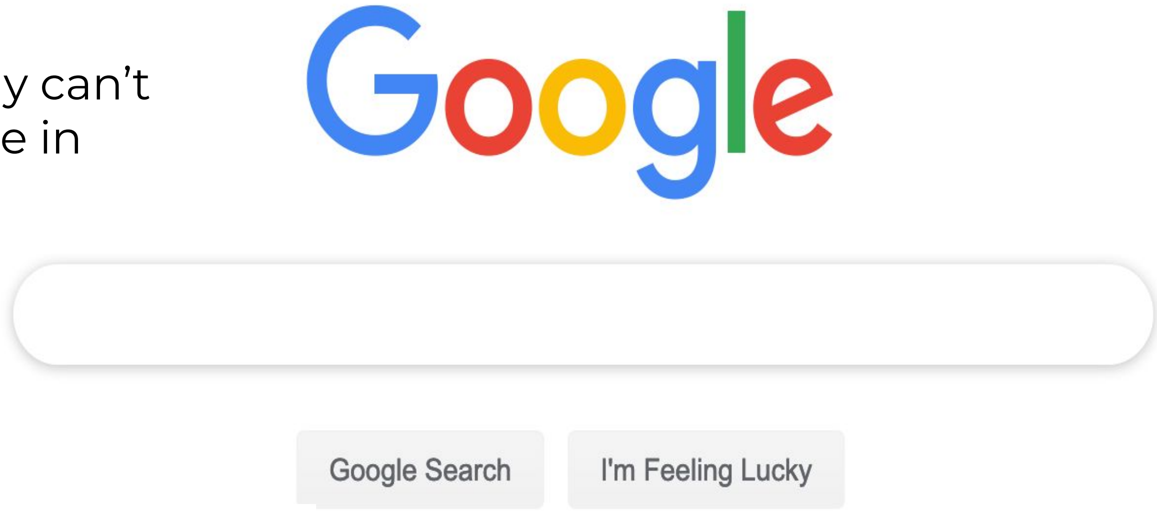
Unstructured (text) vs. structured (database) data today

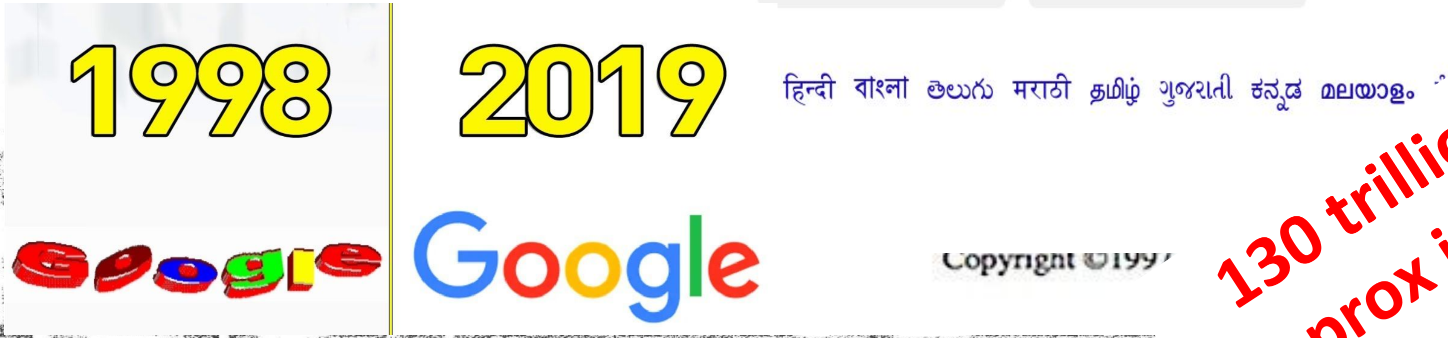


Google Search Interface Over time

Ssearch Box has become the preferred method of information access.

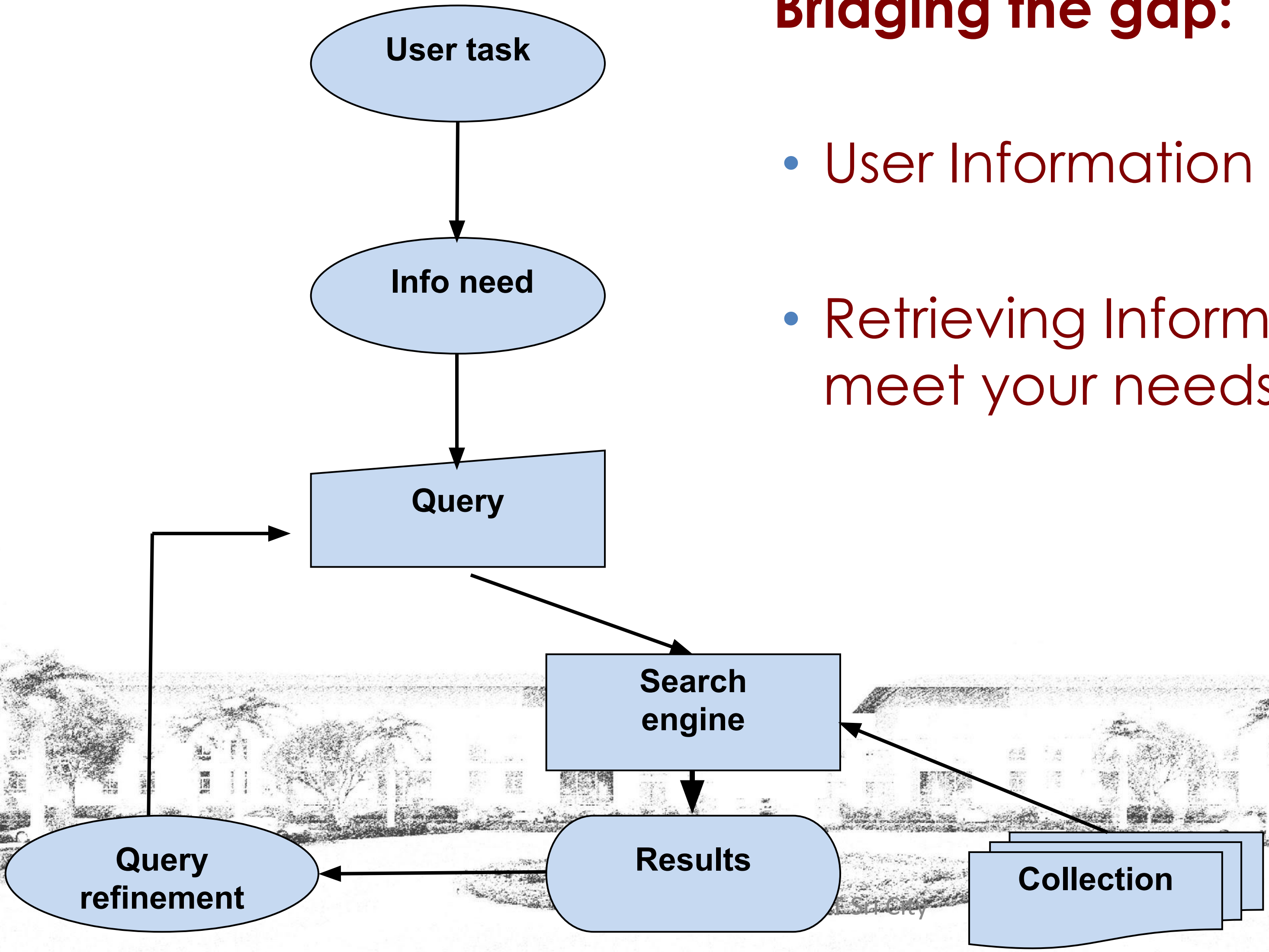
Customers ask: Why can't I search my database in the same way?





130 trillion pages approx in late 2016

Classical Search Engines



Bridging the gap:

- User Information Needs
- Retrieving Information to meet your needs

Understanding QUERY

QUERY: "Bus Services in Java"

Bus Transport in Java Island

Enough AMBIGUOUS !

Java Programming Related Query

Java Island – Transportation Related Query

Understanding QUERY Intent

QUERY: “countries adopting mobile payments”

- number of countries OR
- name of the countries OR
- type of payment services in countries adopting mobile payments
- Any other questions that involve the above information intent

Countries – ?
adopting – ?
mobile – ?
payments – ?

Assumptions

- **Collection:** A set of documents
  - Assume it is a static collection for the moment
  - What about the collection that changes over a period of time?
    - Could Google Search the page you have just now updated??

- **Goal:**
  - Retrieve documents with information
  - This information is relevant to his / her information need
  - This Information helps the user to complete a task

How good are retrieved docs?

Measuring Relevance of retrieved Documents:

- **Precision:** Fraction of retrieved docs that are relevant to the user’s information need

- **Recall:** Fraction of relevant docs in collection that are retrieved

- More definitions and measurements will follow later
  - A detailed lecture will be on the evaluation on IR

Two Steps to Remember

- **Data Structures**
  - The choice of Data Structures
  - Built-in Data Structures (Primitive)
  - User Defined Data Structures (Abstract)

- **Computational Efficiency**
  - Time Complexity
  - Space Complexity
  - Problem / Solution Specific Constraints
  - Best Practices / Efficient Approaches

Course Content

- Course is divided into several modules:

Module: M1 – M3 and M4

- Covers Basic IR to Advanced IR(at least one example problem with detailed analysis)
- Course is supposed to be an interactive course and class performance bonus would be given to students who solve the given set of problems efficiently

□ Course Content follows …

M1: Fundamentals

- Introduction
- Boolean retrieval
- The term vocabulary & postings lists
- Dictionaries and tolerant retrieval
- Index construction
- Index compression

M2: Scoring and IR Evaluation

- Scoring, term weighting & the vector space model
- Computing scores in a complete search system
- Evaluation in information retrieval
- Relevance feedback & query expansion
- XML retrieval
- Probabilistic information retrieval
- Language models for information retrieval
- Information Extraction

M3: Needed Components

- Text classification & Naive Bayes
- Vector space classification
- Flat clustering
- Hierarchical clustering
- Recommender Systems
- Web search basics
- Web crawling and indexes
- Link analysis

M4: Applications of IR

- Scalable Applications of IR
  - Graphs – Massive Web graph / Scale free Graphs
  - Path estimations between given two locations
  - Scalable Graph Examples: Small World Networks
  - Code Search
  - Handling of data from Forums and Blogs
  - Argumentation Mining
  - Mining Unstructured Text Data
  - News Document Retrieval
  - Scientific Documents Retrieval
  - Smart Data Analytics from Unstructured Text Data
  - Understanding Text in Health domain

and many more . . .

TextBooks

- Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, An Introduction to Information Retrieval, Cambridge University Press, Cambridge, England, 2009

- William B. Frakes and Ricardo Baeza-Yates (Eds.). 1992. Information Retrieval: Data Structures and Algorithms. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

- State-of-the-art research papers: SIGIR, WWW, KDD< ECIR and AIRS

Take Home Assignments

- Solve a set of problems every week
- Must be solved by individuals
- Must be finished before Every Monday or the deadline specified for that set of problems
- All Assignments are COMPULSARY

- Total Weightage: 20%;

- **NOTE:**
  - if you fail to explain your solution, you will get "0"

- Solutions would be cross checked !!
- Solutions submitted after the deadline will not be considered for evaluation
- Submission Procedure would be given.

Examinations

- Mid Semester – 1: __________ Marks
- Mid Semester – 2: __________ Marks
- End Semester : __________ Marks

- The actual continuous assessment components will be given after the first class committee meeting

- Academic Code of Conduct
  - Explore PENALTIES

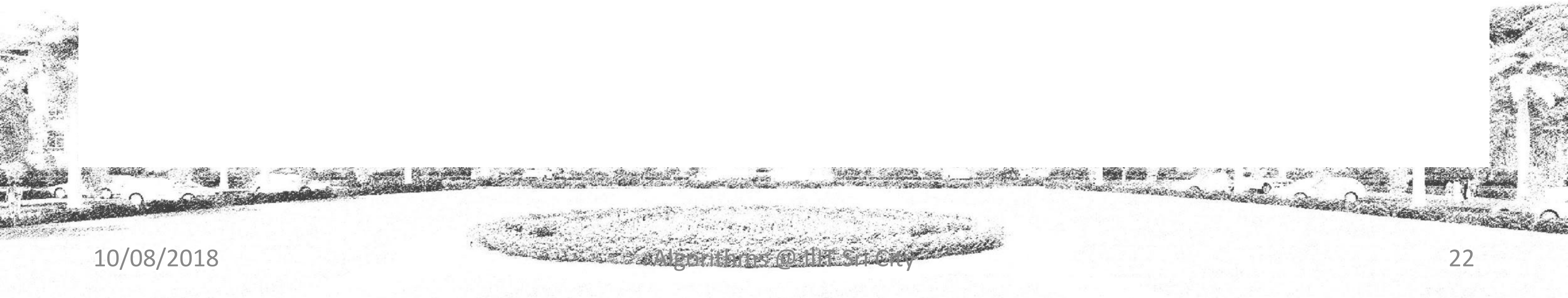
Penalties

- Every Student is expected to strictly follow a fair Academic Code of Conduct to avoid severe penalties

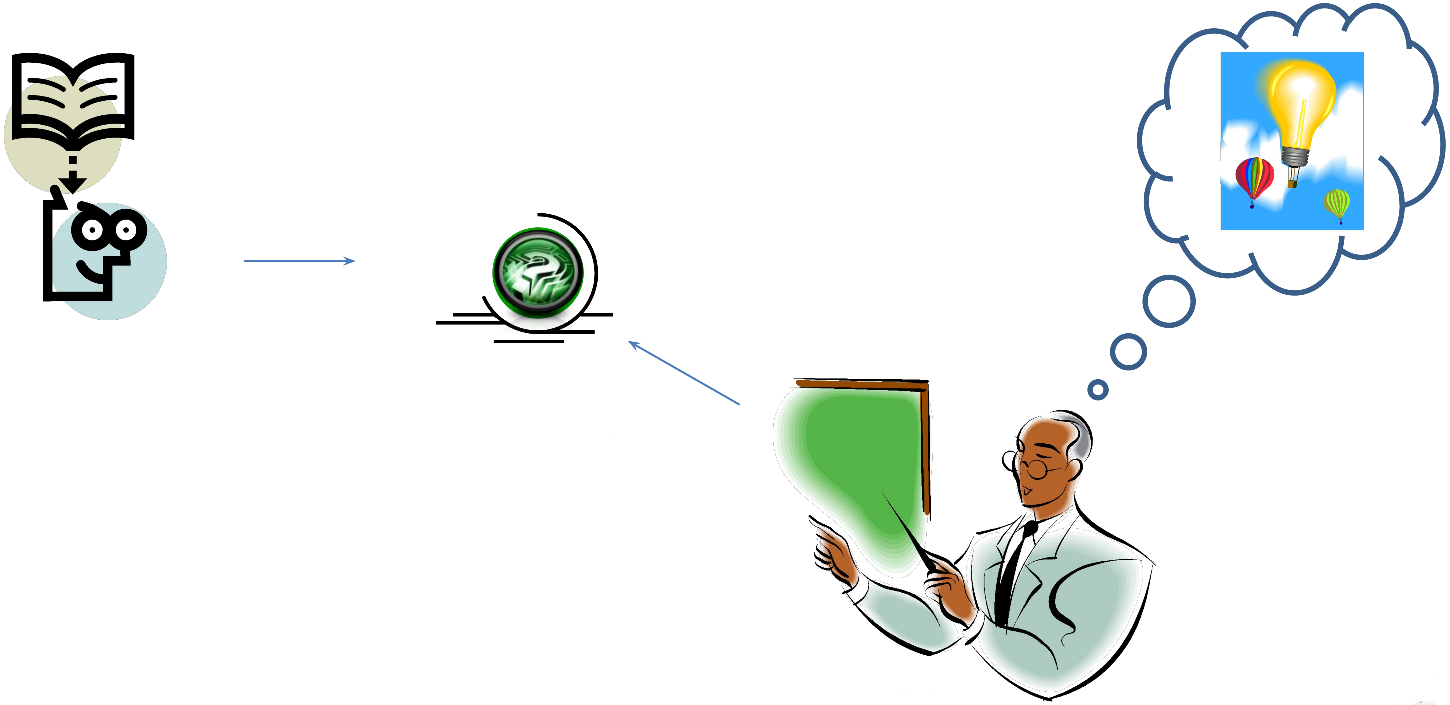
- Penalties would be heavy for those who involve in:
  - **Copy and Pasting** the code
  - **Plagiarism** (copied from your neighbor or friend – in this case, both will get “0” marks for that specific take home assignments)
  - If the candidate is **unable to explain his own solution**, it would be considered as a “copied case” !!
  - **Any other unfair means** of completing the assignments

Assistance

- You may post your questions to me at any time
- You may meet me in person on available time or with an appointment
- You may leave me an email any time
  (email is the best way to reach me faster)



Thanks …



… Questions ???