Lecture notes

CDS 302

Scientific Data and Databases

Fall Semester 2020

Lecture 1: Introduction

Lecture: Joe Boone

Week 1

- **Course Overview and Introduction**
  - Syllabus Review
  - Why Databases?
  - Tool Installation / Setup
  - Scientific Articles
  - Introduction to LaTeX
  - Assignments

Who am I?

- **Joe Boone - GMU Graduate Lecturer**
  - jboone@gmu.edu

- **Academic Career**
  - BS and MS in Computer Science from GMU
  - Currently a Computational Science and Informatics Ph.D. Student at GMU

- **Professional Career**
  - 30+ Years of Systems Development and Engineering Leadership
  - Satellite Telecommunication Applications
  - Geospatial Applications
  - Extensive Software Development Experience
  - Graduate Lecturer at GMU

Syllabus Review

Why Databases?

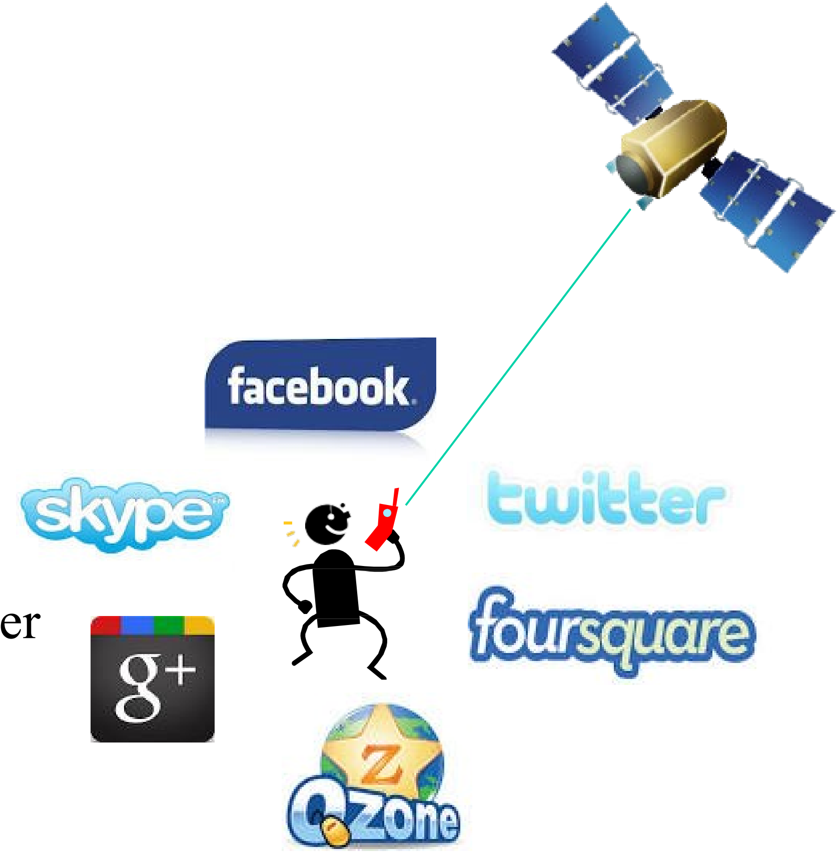
Motivation: Outline

- The Flood of Big Data
- Why Databases?
- Resources

Data is everywhere

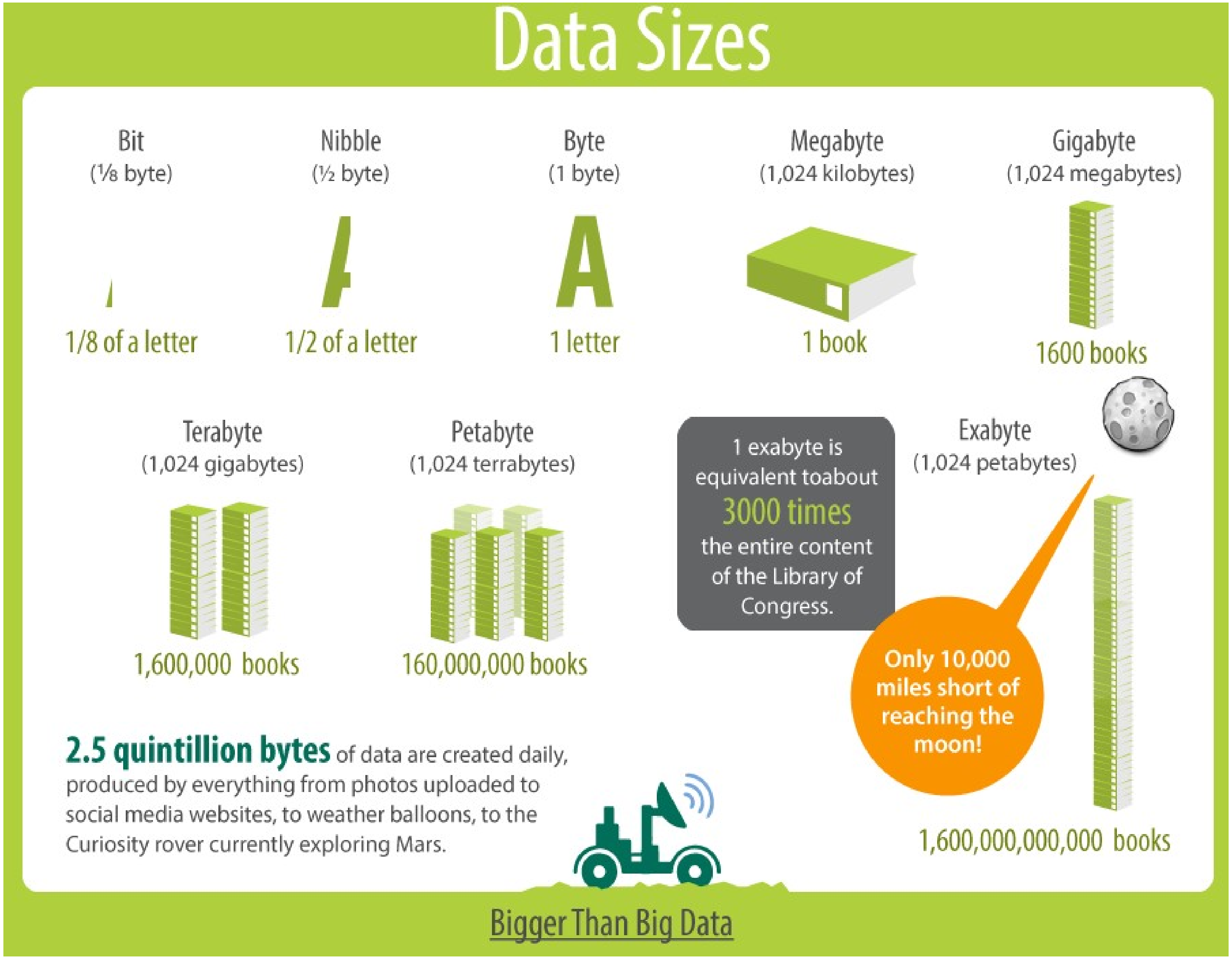
- Huge flood of data
  - Modern technology
  - New user mentality
  - 2.5 Exabytes of new data every day

- New applications
- Innovative research
- Economic Boost
  - "$600 billion potential annual consumer surplus from using personal location data" **[1]**



[1] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011.

Data Sizes reference



Data Sizes

Bit
(⅛ byte)
1/8 of a letter

Nibble
(½ byte)
1/2 of a letter

Byte
(1 byte)
1 letter

Megabyte
(1,024 kilobytes)
1 book

Gigabyte
(1,024 megabytes)
1600 books

Terabyte
(1,024 gigabytes)
1,600,000 books

Petabyte
(1,024 terrabytes)
160,000,000 books

1 exabyte is equivalent toabout 3000 times the entire content of the Library of Congress.

Exabyte
(1,024 petabytes)
1,600,000,000,000 books

Only 10,000 miles short of reaching the moon!

2.5 quintillion bytes of data are created daily, produced by everything from photos uploaded to social media websites, to weather balloons, to the Curiosity rover currently exploring Mars.

Bigger Than Big Data

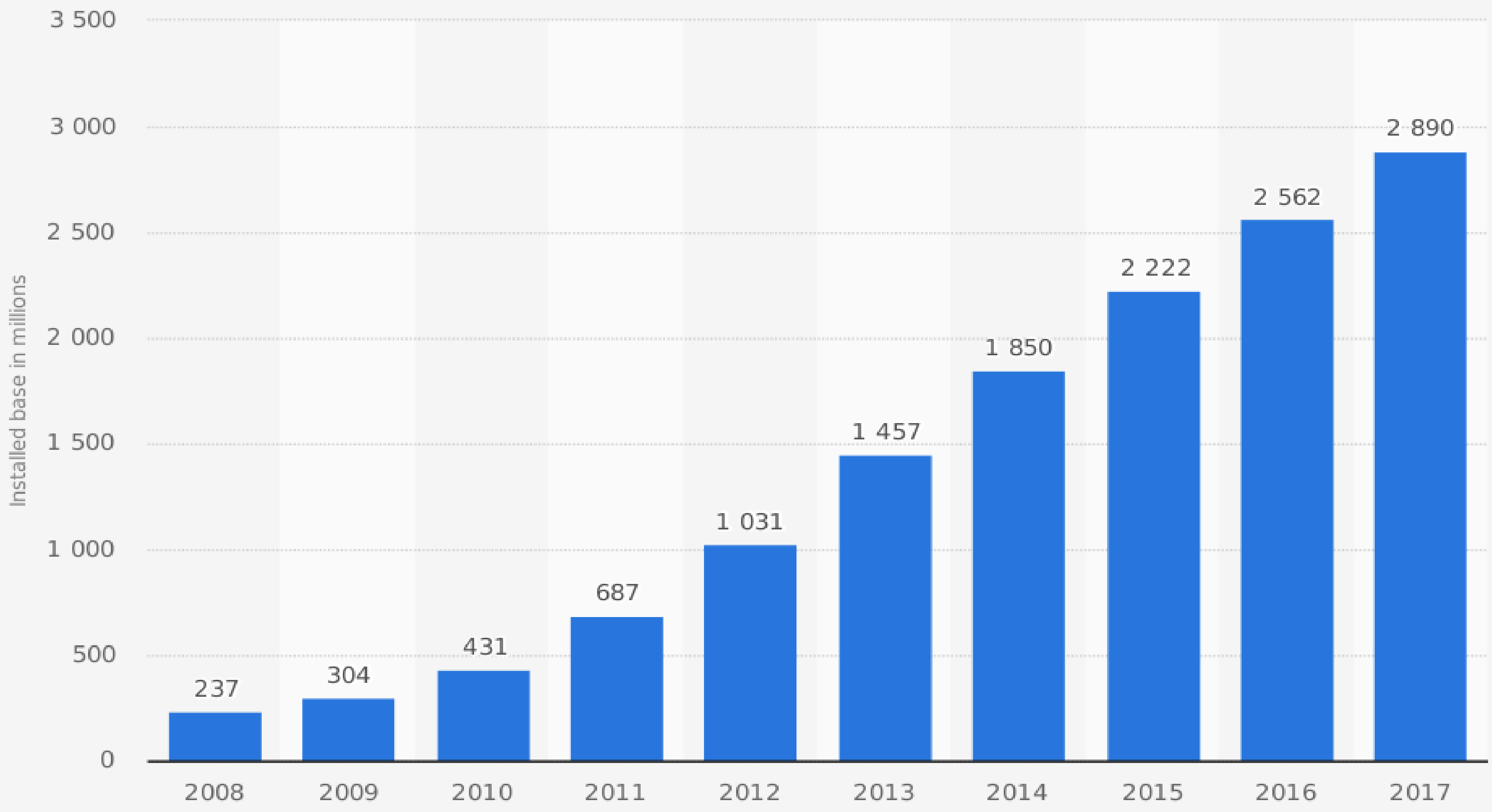
Data production is accelerating

From the dawn of civilization until 2003, humankind generated five exabytes of data. Now we produce five exabytes every two days…and the pace is accelerating.

Eric Schmidt,
Executive Chairman, Google

Data production is accelerating

Smartphones worldwide installed base from 2008 to 2017 (in millions)



Source
GSMA
© Statista 2018

Additional Information:
Worldwide; GSMA ; 2008 to 2013

Big Data = $$$

$600 billion
potential annual consumer surplus from using personal location data globally

60% potential increase in retailers' operating margins possible with big data

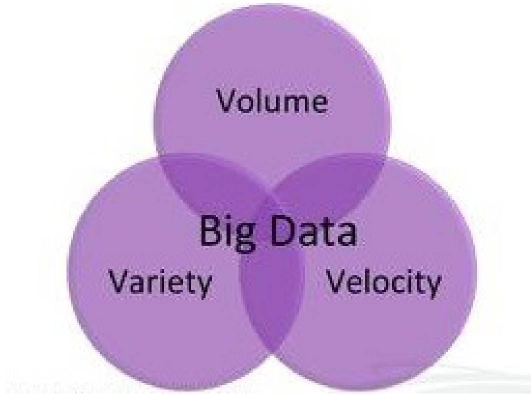
140,000–190,000
more deep analytical talent positions, and

1.5 million
more data-savvy managers needed to take full advantage of big data in the United States

[1] McKinsey Global Institute. Big data: The next frontier for innovation, competition, and productivity. June 2011.

The 3 V's of Big data…

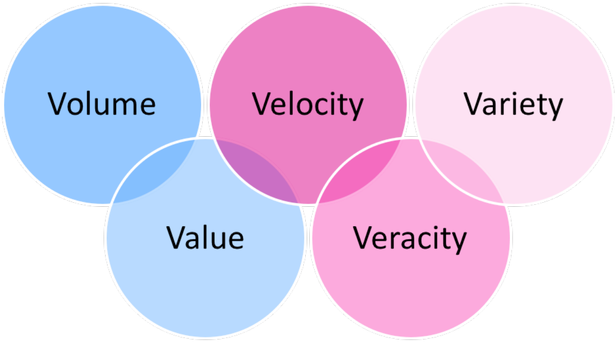
- Volume
- Velocity
- Variety

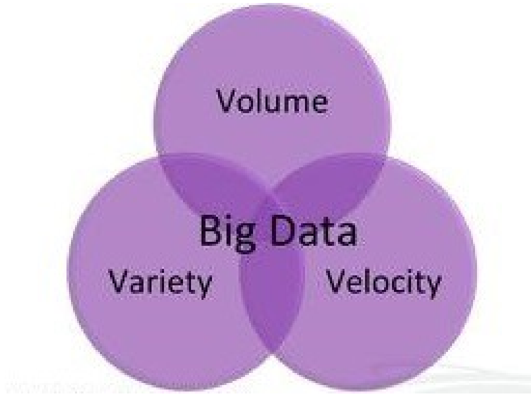


Big Data Hype

The 3-*ish* V's of Big data…

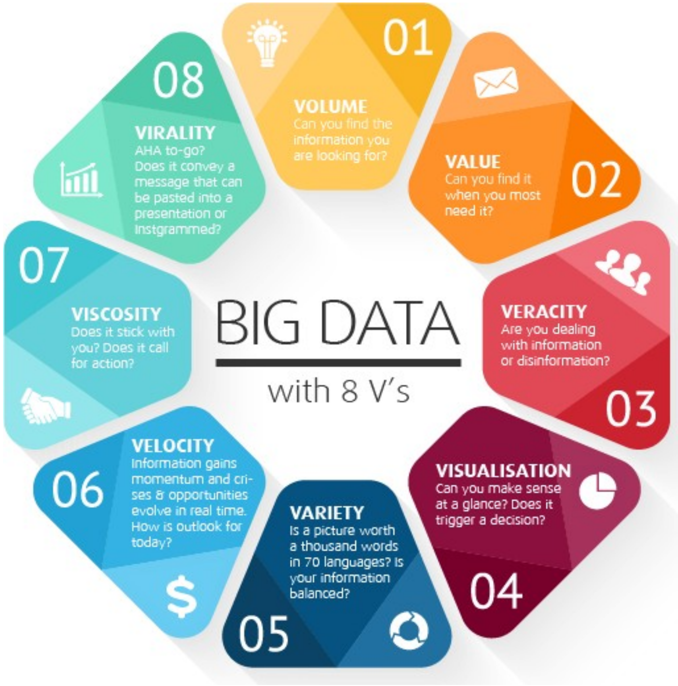
- Volume
- Velocity
- Variety





  - *Veracity*
  - *Variability*
  - *Visualization*
  - *Value*

  …



VOLUME
Can you find the information you are looking for?

VALUE
Can you find it when you most need it?

VERACITY
Are you dealing with information or disinformation?

VISUALISATION
Can you make sense at a glance? Does it trigger a decision?

VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?

VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?

VISCOSITY
Does it stick with you? Does it call for action?

VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?

Big Data Hype

- Big Data analytics is a fancy new word for Knowledge Discovery in Databases!

- KDD has focused on large data for decades:

41st International Conference on
VERY LARGE DATA BASES
Hilton Waikoloa Hotel • Kohala Coast, Hawai'i
August 31 - September 4, 2015

  - **VL**DB since 1975

- **Big Data is not new**

KDD2016
22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining
August 13 - 17, 2016 | San Francisco, California

Big Data Hype

- Big Data analytics is a fancy new word for Knowledge Discovery in Databases!

- KDD has focused on large data for decades:

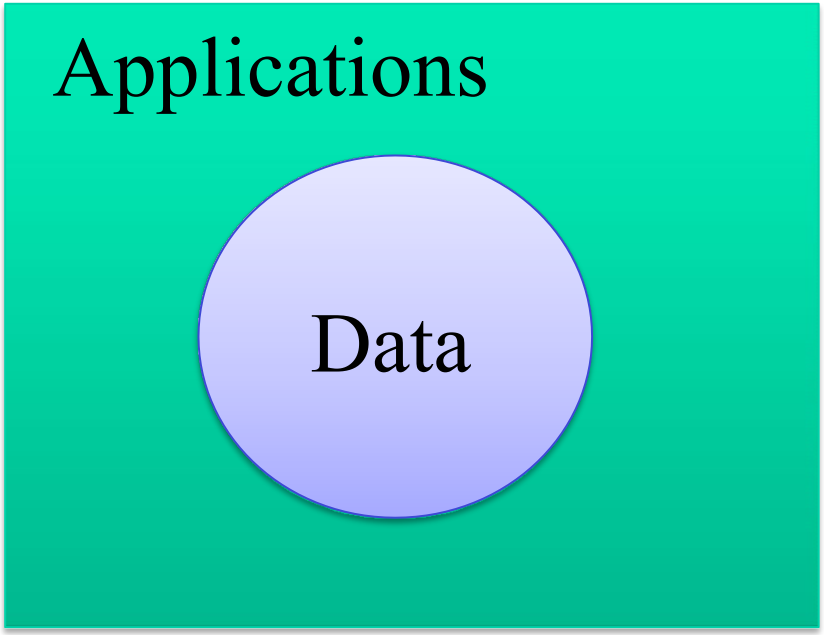
41st International Conference on
VERY LARGE DATA BASES
Hilton Waikoloa Hotel • Kohala Coast, Hawai'i
August 31 - September 4, 2015

- But we want to ride the Big Data wave!
  - ✓ High demand on data analysts
  - ✓ Use Big Data for Storytelling!

Motivation: Outline

- The flood of Big Data
- Why Databases?
- Resources

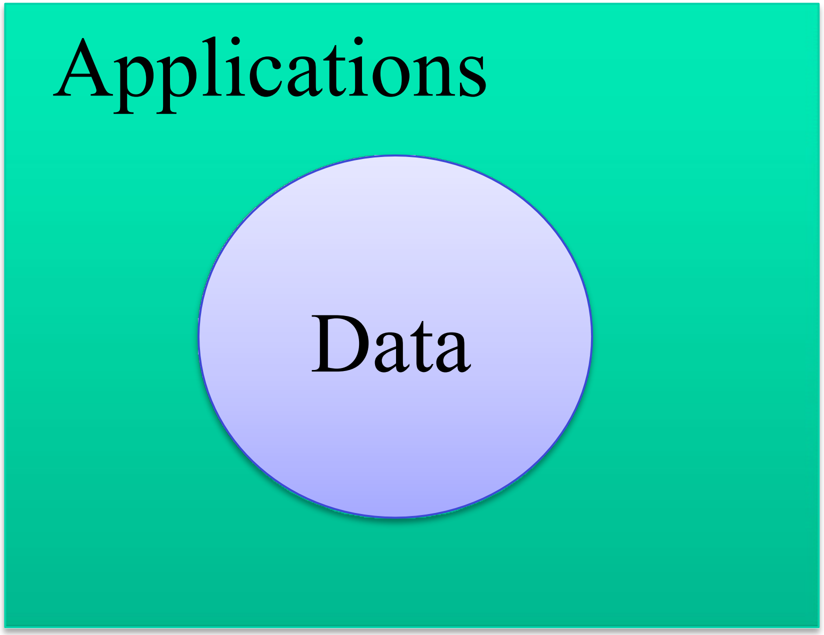
Data is at the center of…?



Data is at the center of...?

Commerce

Politics



Applications

Data

Health

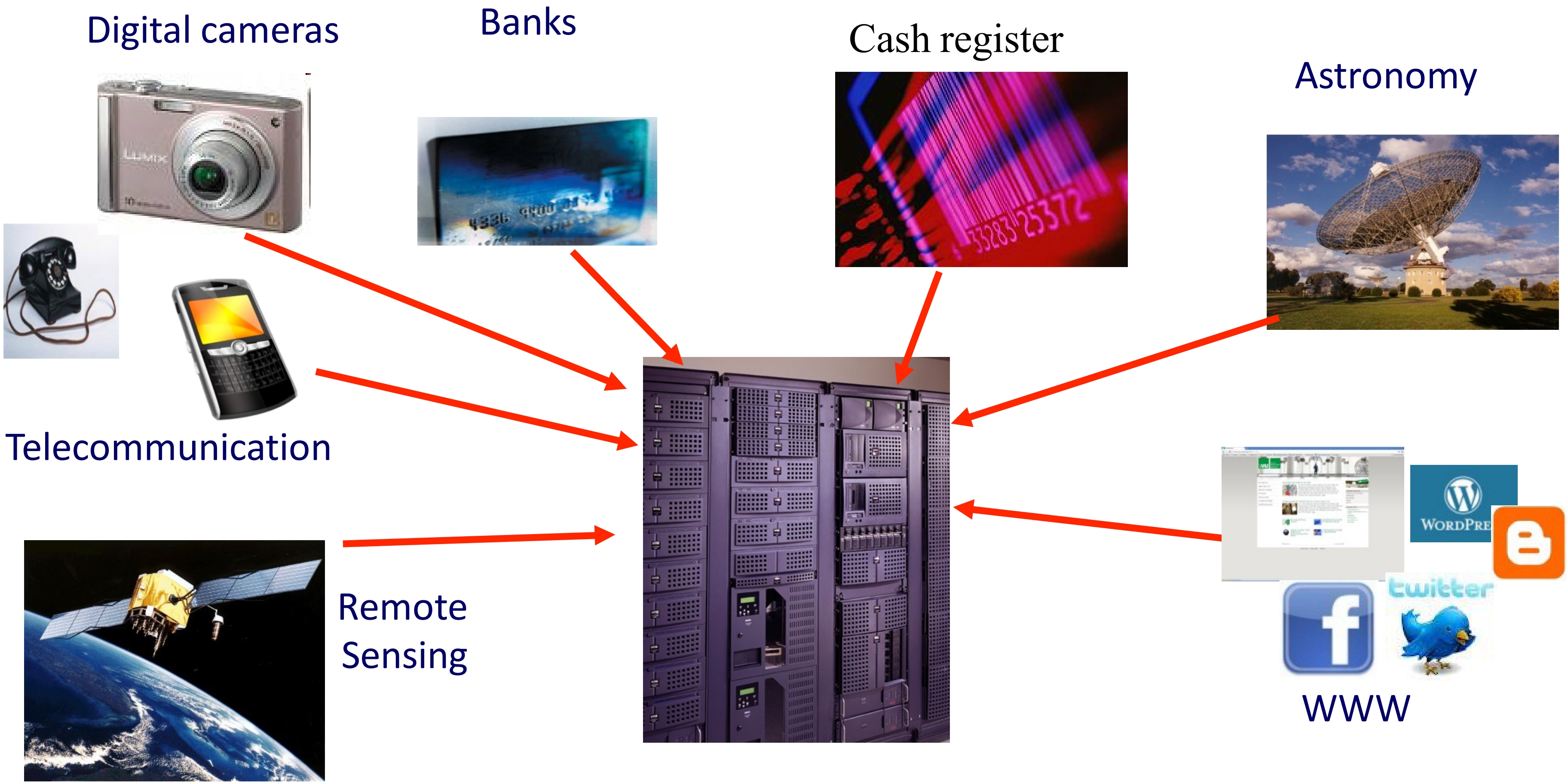
Sports

Security

Communications

(Almost everything…)

Data collection



Digital cameras

Banks

Cash register

Astronomy

Telecommunication

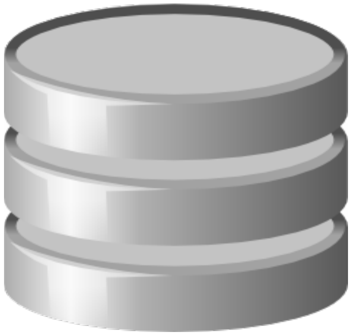
Remote Sensing

WWW

- Huge amounts of data are collected nowadays from different application domains
- Is not feasible to analyze all these data manually

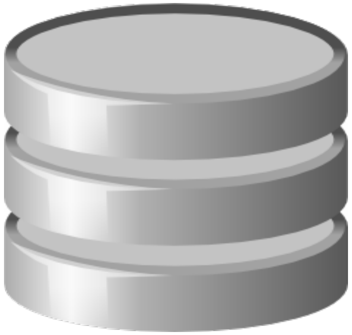
How to manage the Data?

- **What is a Database?**



How to manage the Data?

- **Database**
  - Data collection, typically stored in secondary memory
  - Usually, it contains a sample of all data we could possibly collect in the real world.
- **DBMS**
  - Database Management System: a software package (i.e., collection of programs) designed to *define, manipulate, retrieve* and *manage* data in a database.
- **Database System**
  - DBMS + Database



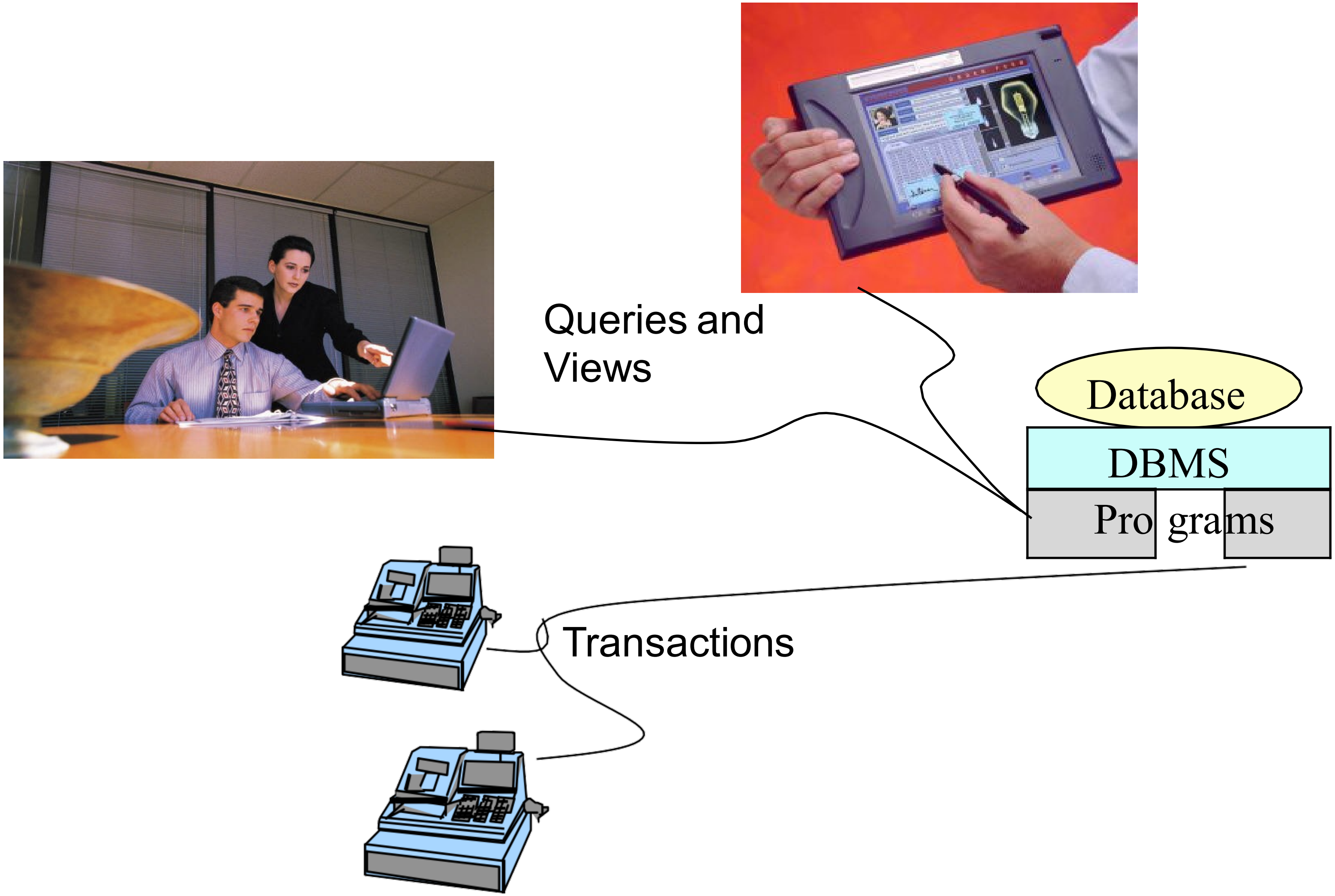
Slide from Katerina Doka, Introduction to Databses

Who uses databases?

Who uses databases?

Slide from Katerina Doka, Introduction to Databases

How does a DBMS work?



Queries and Views

Database

DBMS

Programs

Transactions

DBMS examples

- **Classic relational**
  - Oracle, IBM/DB2, MS SQL-server, Terradata, EMC
  - PostgreSQL (UCB), mySQL, SQLite
- **New relational**
  - In-memory, column store, streaming
- **Non-relational**
  - Graph, geo, scientific
- **No-SQL**
  - Hbase, Cassandra, MongoDB
- **DBMS as a service**
  - MS Azure, Google Big Query

Why DBMS?

- **Data Modeling**
  - redundancy control,
  - consistency constraints (e.g. referential integrity)
- **Efficient query processing**
  - indexing,
  - optimization
- **Operating accuracy**
  - Error recovery - Atomicity
  - Concurrent access by multiple users
- **Security issues**
  - access rights

Why not DBMS?

- Cost of investing in software and hardware, as well as training
- Its generality may cause time overhead
- Not everything that it offers may be needed for a specific application

Scope of this course

- **Design and implementation of a DBMS**
  - Design: E-R and Relational Models
  - Implementation: SQL DDL
  - Theoretical foundations of the above (relational algebra)
  - Organization, File storage
  - Access methods, indexing, hashing
- **Querying a database**
  - SQL queries, views, transactions
- **General principles of Database Management systems**

Why do I want to learn it?

DATA

Data Scientist: The Sexiest Job of the 21st Century [1]

by Thomas H. Davenport and D.J. Patil

If this means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain.

[1] Harvard Business Review. Data Scientist: The Sexiest Job of the 21st Century. October 2012.

Motivation: Outline

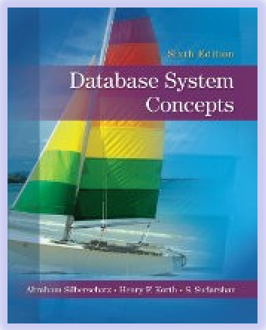
- The flood of Big Data
- Why Databases?
- Resources

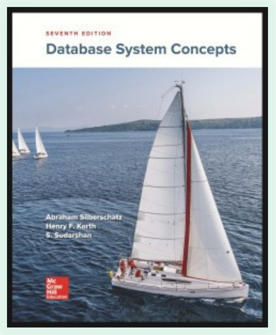
Textbook and Recommended Reference Books

No Compulsory Textbook

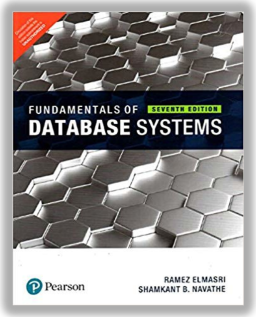
Recommended reference books:

- A. Silberschatz, H. F. Korth, S. Sudarshan
  Database Systems Concepts
  McGraw-Hill, 7th edition





- R. Elmasri, S. B Navathe
  Fundamentals of Database Systems
  Pearson, 7th edition



Suggested reading

Silberschatz, H. F. Korth, S. Sudarshan

Database Systems Concepts, McGraw-Hill.

- Chapter 1 “Introduction”

TODO before next class:

Create a free online account with **Overleaf**.

www.overleaf.com

Resources online

- *Introduction to Databases* class by Jennifer Widom, Stanford
  - http://www.db-class.org/course/auth/welcome

- LaTeX tutorials
  - https://www.overleaf.com/learn/latex/Tutorials
  - https://www.latex-tutorial.com/tutorials/

- List of LaTeX Math Symbols
  - https://www.caam.rice.edu/~heinken/latex/symbols.pdf

Tool Installation and Setup

Software

www.overleaf.com

LaTeX

www.sqlite.org/download.html

sqlitebrowser.org/dl/

SQLite

jupyter

www.anaconda.com/distribution/

python

ANACONDA

SPYDER

Software

MySQL®

PostgreSQL

SQLite

ORACLE

IBM
DB2

Microsoft®
SQL Server®

Scientific Articles

Why read journal articles?

- To update oneself with progress in a particular specialty/ *field of study*
- To find out a solution for a specific problem
  - test / methods
- To understand certain fundamental aspects of the study *area*
- To get an idea for carrying out a research work
- You have been assigned to review the article (e.g. by a *Professor or journal Editor)*
- To find support for one's views

- *To impress others*

Adapted from: How to read clinical journals: I. why to read them and how to start reading them critically. Can Med Assoc J. 1981 Mar 1; 124(5):555-8; Durbin CG., Jr How to read a scientific research paper. Respir Care. 2009;54:1366–71.

Types of articles published in a scientific journal

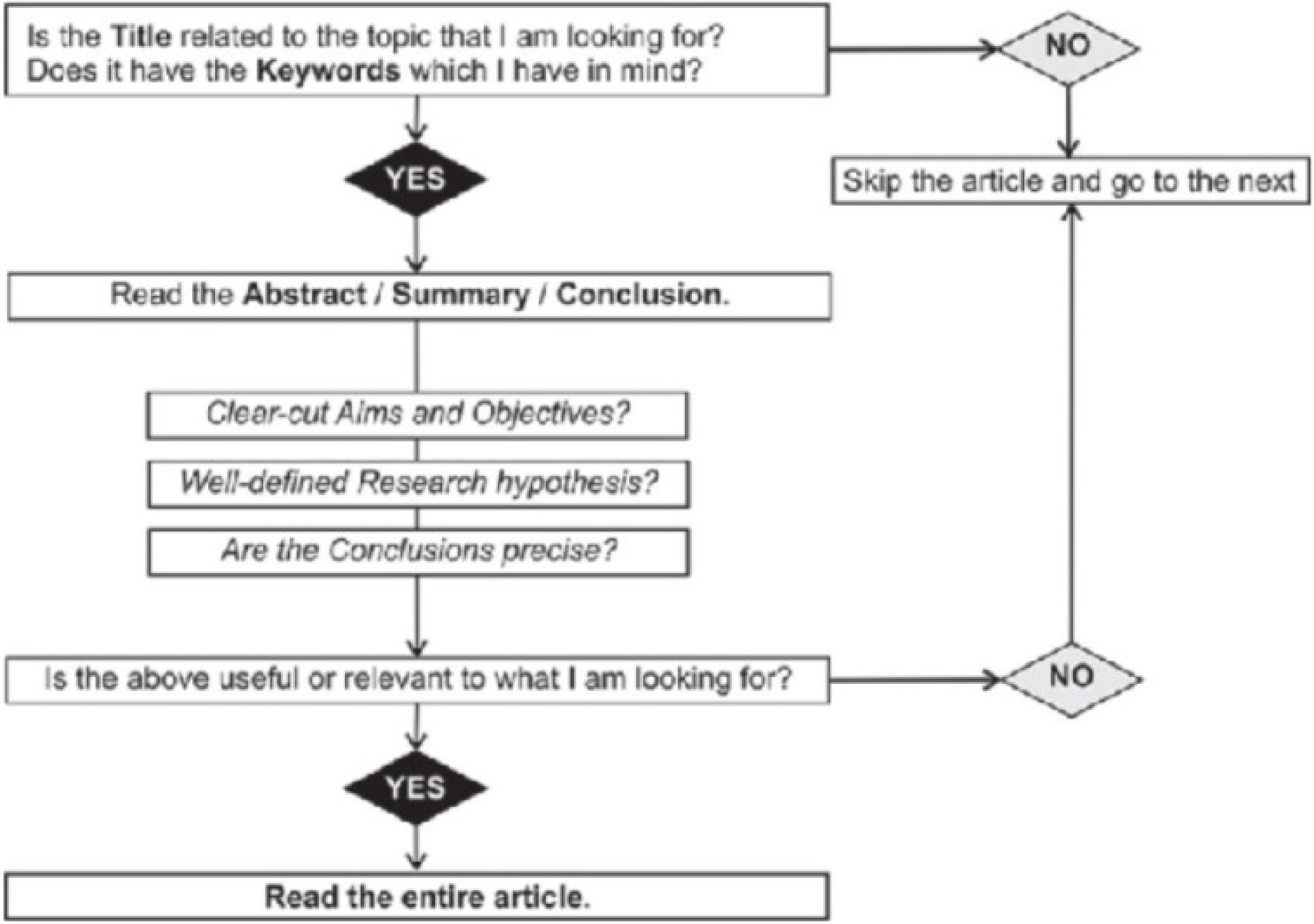
- Primary literature
  - “core” of scientific publications
  - present findings on new scientific discoveries
  - or describe earlier work to acknowledge it and place new findings in the proper perspective
    - Original research articles
    - Surveys
    - Case report/case series
    - Conference proceedings and abstracts
    - Editorial
    - Correspondence/letters to the Editor

- Secondary literature
  - original research information is reviewed
    - Narrative reviews
    - Systematic reviews
    - Meta-analysis
    - Book reviews
    - Guidelines
    - Commentary

Structure of a journal article

- **Title:** Topic and information about the authors.
- **Abstract:** Brief overview of the article.
- **Introduction:** Background information, gap in research, and statement of the research hypothesis. Also: *Motivation – why is this work important?*
- **Methods:** Details of how the study was conducted, procedures followed, instruments used and variables measured. Must be *systematic*.
- **Results / Experimental Evaluation:** All the data of the study along with figures, tables and/or graphs.
- **Discussion:** The interpretation of the results and implications of the study. Were the objectives met? Limitations and Future work.
- **Conclusion:** What does all this mean?
- **References / Bibliography:** Citations of sources from where the information was obtained.

Process



Is the **Title** related to the topic that I am looking for?
Does it have the **Keywords** which I have in mind?

NO

Skip the article and go to the next

YES

Read the **Abstract / Summary / Conclusion.**

Clear-cut Aims and Objectives?

Well-defined Research hypothesis?

Are the Conclusions precise?

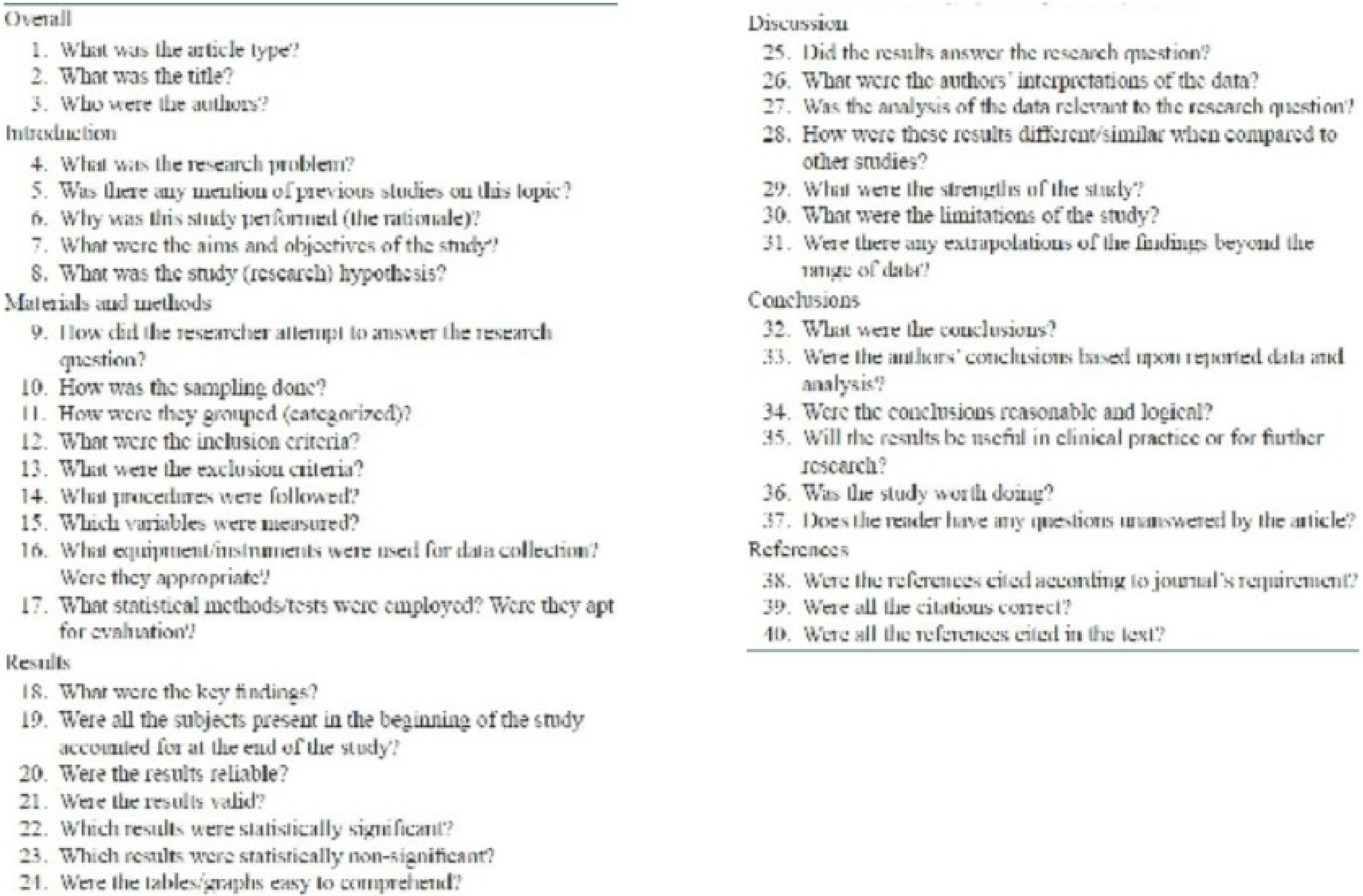
Is the above useful or relevant to what I am looking for?

NO

YES

Read the entire article.

Research Questionnaire



Overall

1. What was the article type?
2. What was the title?
3. Who were the authors?

Introduction

4. What was the research problem?
5. Was there any mention of previous studies on this topic?
6. Why was this study performed (the rationale)?
7. What were the aims and objectives of the study?
8. What was the study (research) hypothesis?

Materials and methods

9. How did the researcher attempt to answer the research question?
10. How was the sampling done?
11. How were they grouped (categorized)?
12. What were the inclusion criteria?
13. What were the exclusion criteria?
14. What procedures were followed?
15. Which variables were measured?
16. What equipment/instruments were used for data collection? Were they appropriate?
17. What statistical methods/tests were employed? Were they apt for evaluation?

Results

18. What were the key findings?
19. Were all the subjects present in the beginning of the study accounted for at the end of the study?
20. Were the results reliable?
21. Were the results valid?
22. Which results were statistically significant?
23. Which results were statistically non-significant?
21. Were the tables/graphs easy to comprehend?

Discussion

25. Did the results answer the research question?
26. What were the authors' interpretations of the data?
27. Was the analysis of the data relevant to the research question?
28. How were these results different/similar when compared to other studies?
29. What were the strengths of the study?
30. What were the limitations of the study?
31. Were there any extrapolations of the findings beyond the range of data?

Conclusions

32. What were the conclusions?
33. Were the authors' conclusions based upon reported data and analysis?
34. Were the conclusions reasonable and logical?
35. Will the results be useful in clinical practice or for further research?
36. Was the study worth doing?
37. Does the reader have any questions unanswered by the article?

References

38. Were the references cited according to journal's requirement?
39. Were all the citations correct?
40. Were all the references cited in the text?

It takes time

- It the not the same as reading a novel or a blog
- It’s a skill
- It gets better with practice
- It gets better as you become more familiar with the research area
- The first paper may take some time
- You may have to look at other resources to understand some of the paper’s content
- Be patient....you’ll get there!

Additional resources

- Google search
- Google Scholar
- GMU Library resources (library.gmu.edu)
- Academia.edu
- ResearchGate
- Reddit Scholar
- Email scholars if you can’t get their articles freely
- Connect through VPN
  - Check the following webpage for more information:

https://itservices.gmu.edu/services/view-service.cfm?customel_dataPageID_4609=6169

LaTeX Introduction

Class information

- **Writing intensive class!**
  - Scientific writing
    - articles / scientific reports
    - document structure, titles, sections, subsections
    - tables, figures, references, citations, math notations
  - **LaTeX** *(strongly preferred)*
    - **Overleaf** online editor (*strongly suggested*)
    - MikTex, TexWorks, etc. (*alternative LaTeX editors*)
  - MS Word *(alternatively)*

LATEX

- **LaTeX**
  - LaTeX is “Lamport” + “TeX”
  - Leslie Lamport
    - Computer Scientist, Distributed Systems (2013 Turing Award)
    - Initial Developer / Inventor of LaTeX (1983)
  - It is a macro package built on top of the typesetting system TeX
  - Defacto standard for scientific journal articles
  - Pronounced “la” or “lay” + “tech”

- **TeX**
  - Late 1970’s
  - Donald Knuth, Computer Scientist
  - Author of *The Art of Computer Programming* (A Computer Science Classic)
  - Typesetting engine that drives LaTeX (and many other macro packages)

Assignments

Suggested reading

Silberschatz, H. F. Korth, S. Sudarshan

Database Systems Concepts, McGraw-Hill.

- Chapter 1 “Introduction”

TODO before next class:

Create a free online account with **Overleaf**.

www.overleaf.com