CSE 4/587
Data Intensive Computing

Dr. Eric Mikida
epmikida@buffalo.edu
208 Capen Hall

Day 02
Data Intensive Computing and Data Strategy

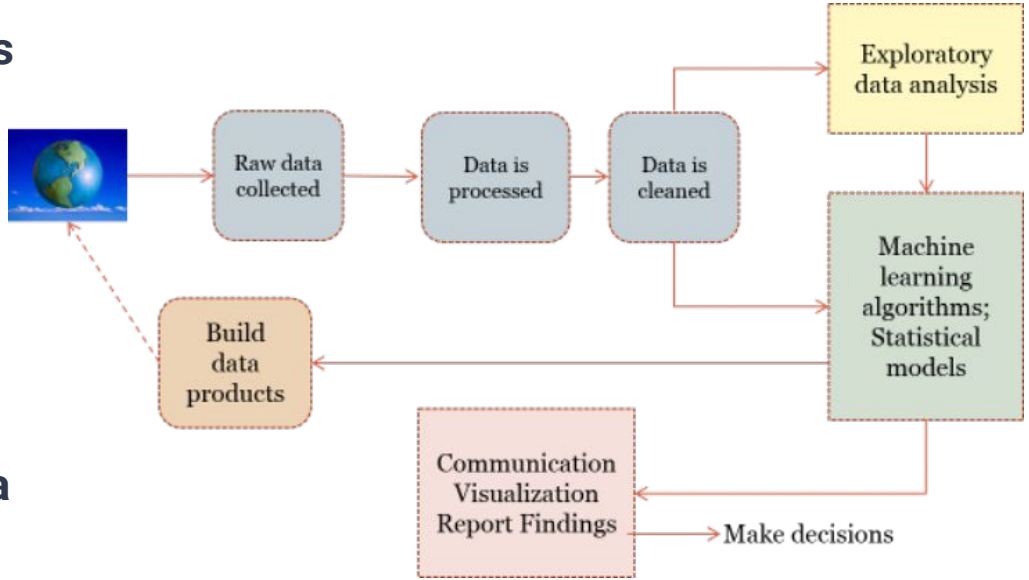
Announcements and Feedback

1. Office hours start next week
2. Questions about force registration
3. Questions about the project

Recap from last class…

You'll learn about…

- **The basic data analytics process**
- **Big data infrastructures and algorithms**
  - Large scale data requires special structures and algorithms
- **newer data challenges and methods to address these**
  - For example **streaming data**



(figure from Doing Data Science)

Basic Data Analytics Process

1. Pick your domain
2. Raw data collection
   - search, study, and get the data
3. Data processing
   - select the features of interest
4. Data cleaning
   - addressing missing values, replacing N/A with appropriate value, etc.
5. Exploratory data analysis (EDA)
   - understand the nature of the data
6. Modeling and analysis
   - apply algorithms to analyze data (statistical modeling and/or ML)
7. Visualization
8. Build data products
9. Open it up to the world
10. Let decision makers use it make data-driven decisions

What are your recommendations based on your analysis?

A bit of motivation…

Tremendous advances have taken place in **statistical methods and tools**, **machine learning** and **data mining approaches**, and **internet-based dissemination tools** for analysis and visualization.

- Many tools are open source and freely available for anybody to use.
- Is there an easy entry-point into learning these technologies?
- Can we make these tools easily accessible to the students, researchers and decision makers similar to how “office” productivity software is used?

High Level Goals

1. Understand foundations of data analytics so that you can **interpret and communicate results and make informed decisions**
2. Study and learn to apply common **statistical methods** and **machine learning algorithms** to solve business problems
3. Learn to work with popular **tools** to analyze and visualize data
4. **Working with cloud** for data storage and for deployment of applications
5. Learn methods for mastering and applying emerging concepts and technologies for **continuous data-driven improvements to your research/work/business processes**
6. **Transform complex analytics into routine processes**

What is Data Intensive Computing?

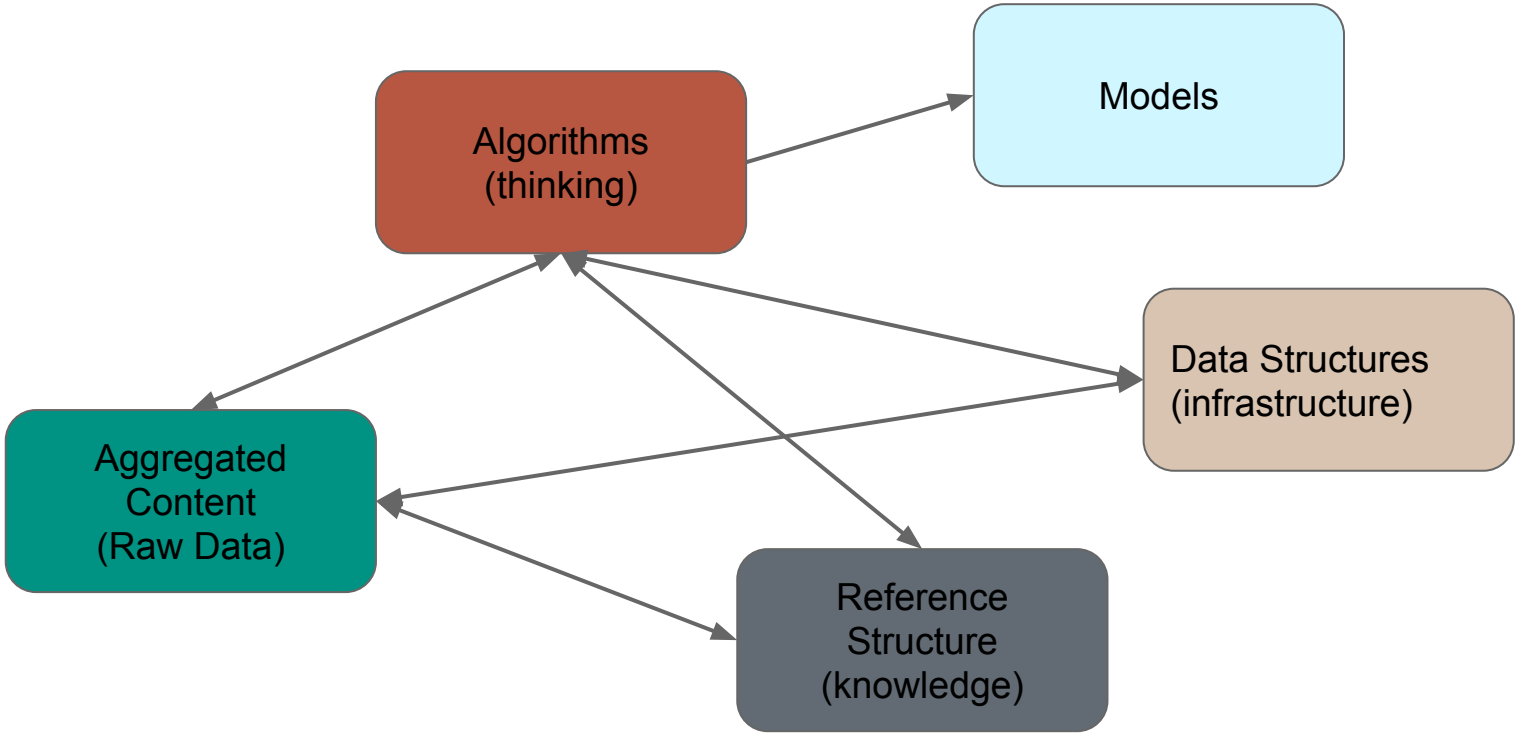
- The phrase was initially coined by National Science Foundation (NSF)
- What is it?
  - Volume, velocity, variety, veracity (uncertainty) (Gartner, IBM)
- What do you expect to extract by processing this large data?
  - Intelligence for decision making
- What is different now?
  - Storage models, processing models
  - Big Data, analytics and cloud infrastructures

Examples of Data Intensive Applications

- Search engines
- Recommendation systems
  - Netflix: movie recommendations
  - Amazon: book/product recommendations
- Biological systems: high throughput sequences (HTS)
  - Analysis: disease-gene match
  - Query/search for gene sequences
- Space exploration
- Financial analysis

...and many more...

Characteristics of Data Intensive Computing



Characteristics of Data Intensive Computing

Aggregated content: large amount of data pertinent to the specific application/problem each. ie databases

Reference structures: structures that provide one or more structural/semantic interpretations of the content (think domain specific knowledge).

Algorithms: modules that allows the application to harness the information which is hidden in the data. Applied on aggregated content and some times require reference structures. ie MapReduce

Data Structures: newer data structures to leverage the scale of data and its write-once read-many (WORM) characteristics. ie: MS Azure, Apache Hadoop, Google BigTable

Diversity of Data

New kinds of data from different sources: tweets, geo location, emails, blogs, IoTs…

Three major types: structured, unstructured data and now streaming data

1. **Structured data:** data collected and stored according to well defined schema
    - i.e. stock quotes
2. **Unstructured data:** messages from social media, news, talks, books, letters, manuscripts, court documents...
3. **Streaming data:** constant flow of new data

We will discuss methods for analyzing structured, unstructured data and streaming data

Not just diverse…BIG (Deluge of Data)

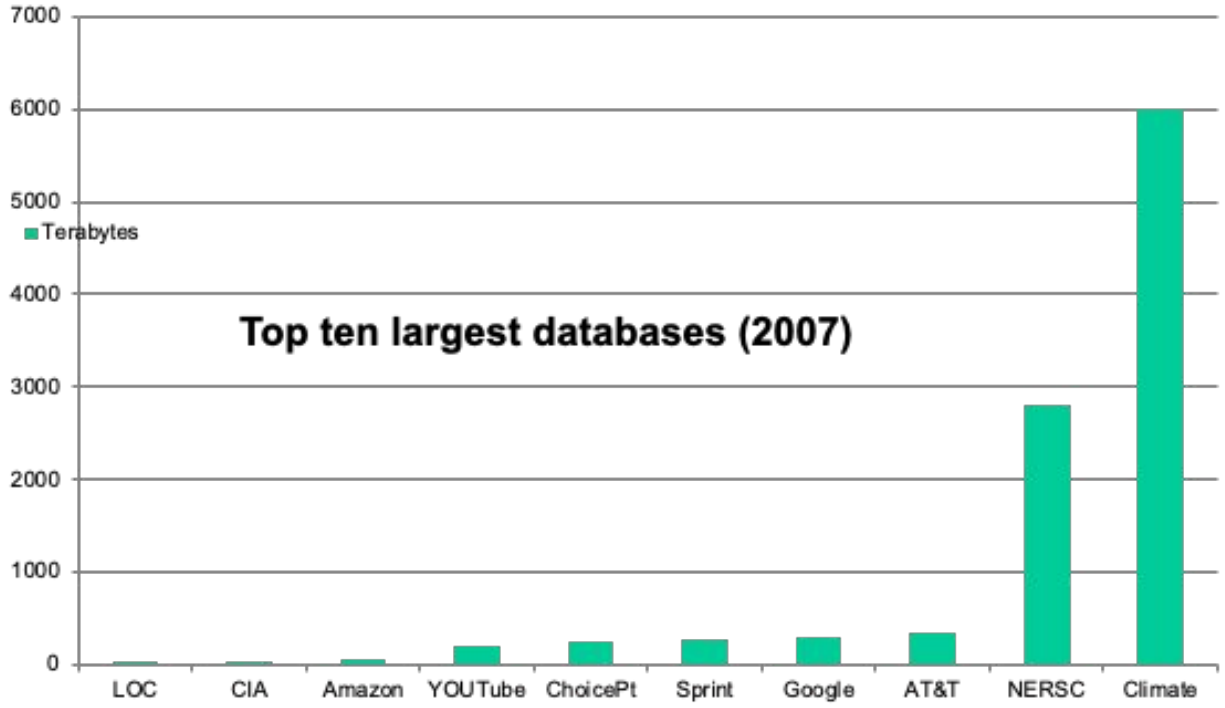
Bioinformatics data: from about 3.3 billion base pairs in a human genome to huge number of sequences of proteins and the analysis of their behaviors

The internet: web logs, facebook, twitter, maps, blogs, etc.: Analytics …

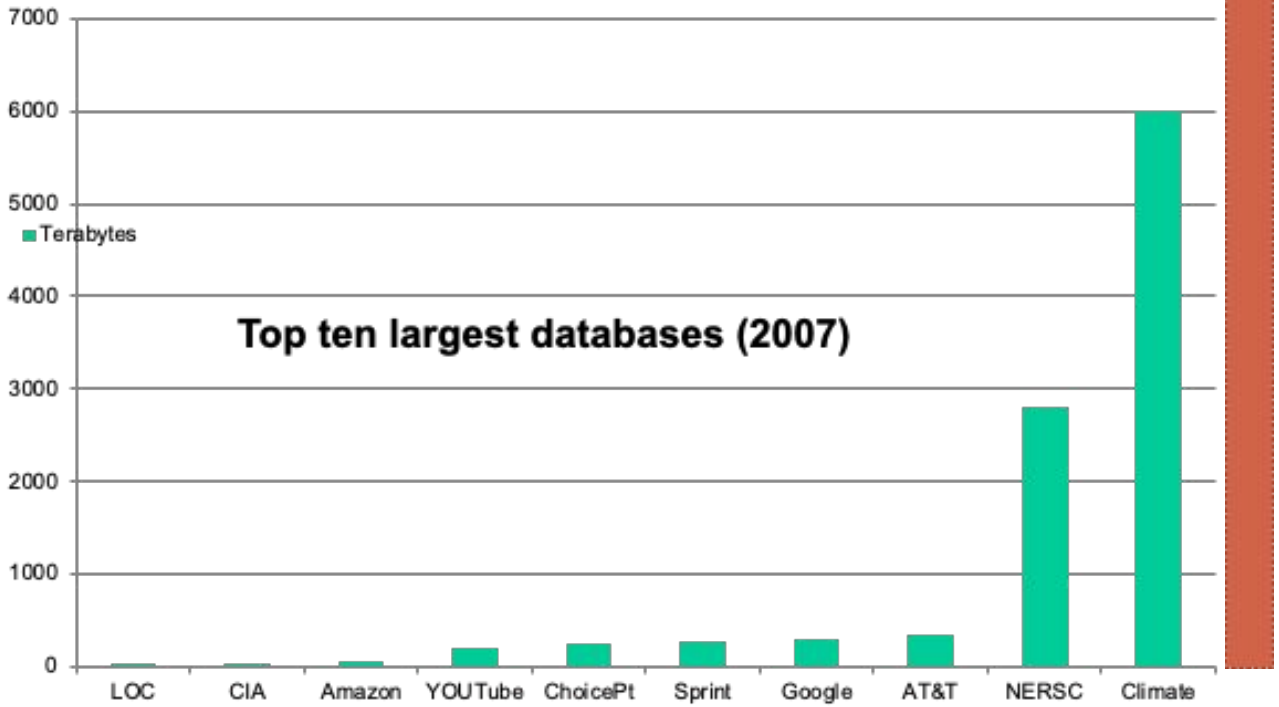
Financial applications: volumes of data for trends and other deeper knowledge

Healthcare: huge amount of patient data, drug and treatment data

The universe: the Hubble ultra deep telescope shows 100s of galaxies each with billions of stars. Sloan Digital Sky Survey: https://www.sdss.org/



Ref: http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world/



Facebook @ 21 petabytes in 2010

Ref: http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world/

How do we approach these issues?

Algorithmic: develop scalable/tractable algorithms (we've been working at it for a while…)

High Performance computing (HPC): Frontier machine (ORNL) has over 8 million cores, over 1 exaflop. Programmed with technologies like MPI, OpenMP, etc.

GPGPU programming: general purpose graphics processor (NVIDIA)

Statistical packages: R and/or python running on parallel threads on powerful machines

Machine learning algorithms on supercomputers

Hadoop: MapReduce like parallel processing.

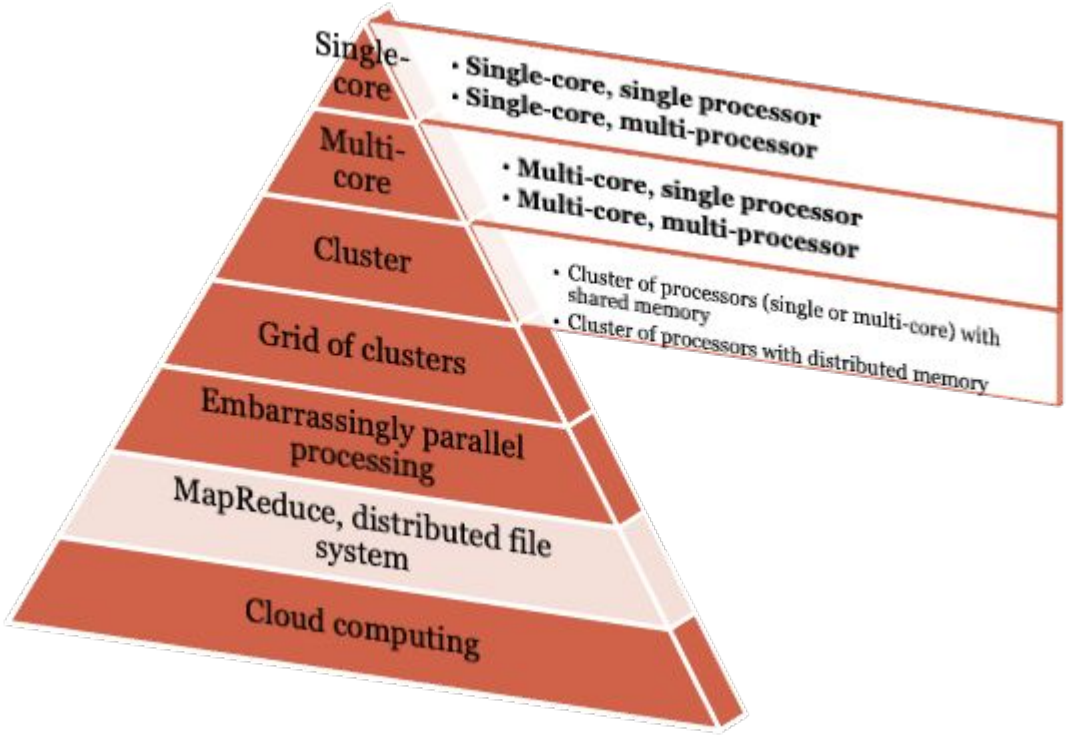
Spark-like approaches: in-memory computing models

Stream processing infrastructure (Kafka)

Processing at different granularity

Data size: Small

Data size: Large



Algorithms

- Statistical inference
- Machine learning: capability to generalize based on past experience and then apply these generalizations to provide answers related to old, new and future data.
- Data mining
- Soft computing
- Deep learning

We also need algorithms that are specially designed for the emerging storage models and data characteristics.

“Intelligence” and Scale of Data

Intelligence: a set of discoveries made by federating/processing ***information*** collected from diverse sources

Information: a cleansed form of raw data.

To gather good **intelligence**, we need a large amount of **information**

To get statistically significant **information** we need a reasonable amount of **data**

Therefore, intelligent applications are invariably data-heavy, data-driven, and data-intensive

Examples of intelligent applications

Google search: How is different from regular search in existence before it?

- It took advantage of the fact the hyperlinks within web pages form an underlying structure that can be mined to determine the importance of various pages.

Restaurant and Menu suggestions: Instead of “Where would you like to go?” “Would you like to go to Olive Garden”?

- Learning from previous data of habits, profiles, and other information gathered over time.

Social media network suggestions: ie facebook friend suggestion

…Did you know amazon can ship things before you order? Here

More examples...

- Content aggregators
- Media-sharing sites
- Online gaming
- Sports analytics
- Biological analysis
- Space exploration

Different Types of Storage

Internet introduced a new challenge in the form web logs, “peta scale” and beyond

- Uniquely different characteristic than transactional (“customer order” data, or “bank account data”) data
- The data type is “write once read many (WORM)”

Other examples

- Privacy protected healthcare and patient information
- Historical financial data
- Other historical data

Different Types of Storage

Relational file system and tables are insufficient…now we need:

- Large <key, value> stores (files) and storage management system
- Built-in features for fault-tolerance, load balancing, data-transfer and aggregation…
- Clusters of distributed nodes for storage and computing
- Computing is inherently parallel
- Streaming systems

Big Data Concepts

Originated from the Google File System (GFS), the special <key, value> store

- Hadoop Distributed file system (HDFS) is the open source version of this (Currently an Apache project)

Parallel processing of the data using MapReduce (MR) programming model

Challenges:

- Formulation of MR algorithms
- Proper use of the features of infrastructure (Ex: sort)
- Best practices in using MR and HDFS
- Coordinating an extensive ecosystem of other components: column-based store (Hbase, BigTable), big data warehousing (Hive), workflow languages, etc.

Cloud Computing

Cloud is a facilitator for Big Data computing and is indispensable in this context

It provides processors, software, operating systems, storage, monitoring, load balancing, clusters and other requirements as a service

Cloud offers accessibility to Big Data computing

Cloud computing models:

- platform (PaaS), Microsoft Azure
- software (SaaS), Google App Engine (GAE)
- infrastructure (IaaS), Amazon web services (AWS)
- Services-based application programming interface (API)

In Summary…

“In pioneer days they used oxen for heavy pulling, when one couldn’t budge a log they didn’t try to grow a larger ox. We shouldn’t be trying for bigger computers, but for more systems of computers.” ~Grace Hopper

We have witnessed explosion in algorithmic solutions

What you cannot achieve by an algorithm can be achieved by more data

Big data if analyzed right gives you better answers: traditional prediction of flu vs. prediction of flu through “search” data 2 full weeks before the onset of flu season!

Data Strategy

In this era of big data, what is your data strategy?

Essentially, how are you going to plan for the data challenge?

- It is not only about big data, but data in all sizes and forms
- Data collections from customers used to be an elaborate task
  - ie surveys
- Nowadays data is available in abundance
  - technological advances and social networks
- Data is also generated by many of your own business processes and applications

Components of a Data Strategy

- Data integration
- Meta data
- Data modeling
- Organizational roles and responsibilities
- Performance and metrics
- Security and privacy
- Structured data management
- Unstructured data management
- Business intelligence
- Data analysis and visualization
- Tapping into social data

Data Strategy at a high level

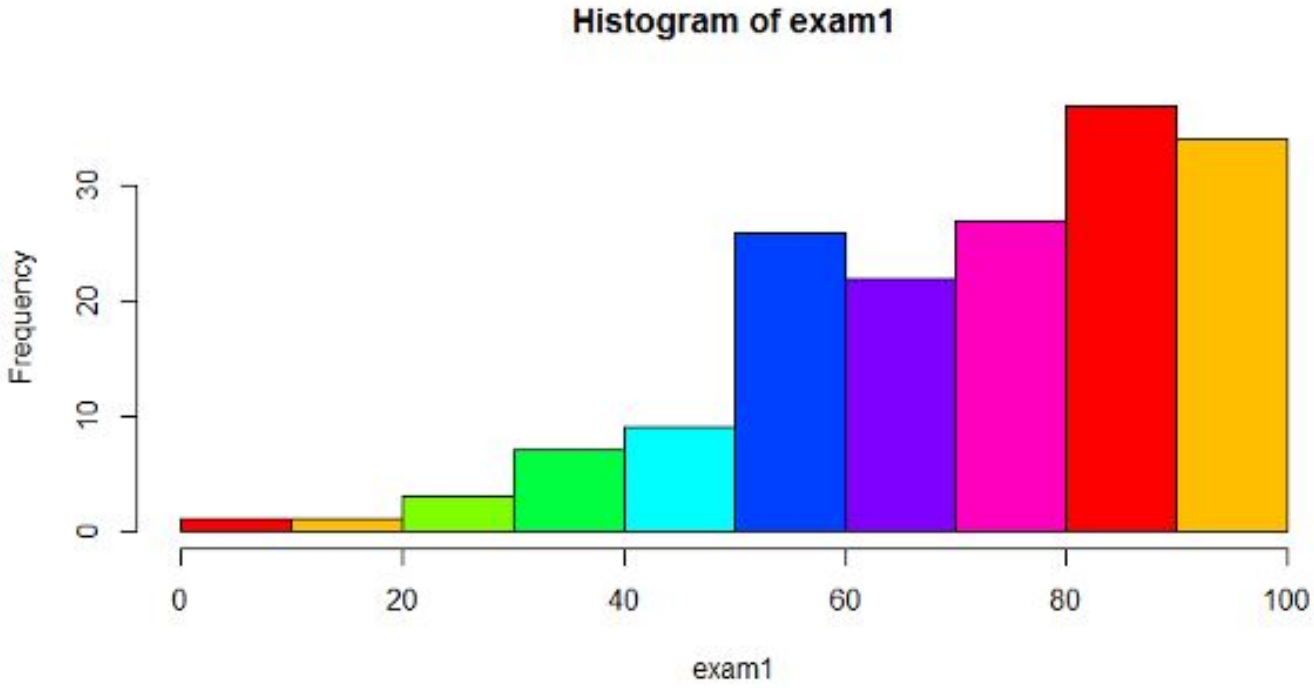
- How will you collect data? Aggregate data? What are your sources? (ie. social media)
- How will you store your data? And where?
- How will you use the data? Analyze it? Analytics? Data mining? Pattern recognition?
- How will you present or report the data to the stakeholders and decision makers? Visualization?

Example 1 with Exam Grades

Q1	Q2	Q3	Q4	Q5	Total
16.7	13.9	9.6	18.5	13.7	72.4
20.0	16.0	9.0	19.0	17.0	76.0
20.0	20.0	15.0	25.0	20.0	90.0
Q1	Q2	Q3	Q4	Q5	Total
16.0	14.2	9.6	19.4	14.0	73.2
80.1%	71.1%	64.0%	77.4%	70.2%	73.2%
Q1	Q2	Q3	Q4	Q5	Total
17.3	13.6	9.7	17.6	13.3	71.5
86.7%	67.8%	64.6%	70.3%	66.7%	71.5%

Question 1..5, total, mean, median, mode; mean ver1, mean ver2

Example 2 with Same Grades



Example 3 with Same Grades