Introduction to Big Data

Week 4

- INFO911 -

Introduction to Big Data

Presented by
Prof. Zhifeng Wang

Outline

- Some Statistics
  - How much data is created every day
- What is collecting all this data
- Who is collecting all this data
- Why are they collecting all this data
- Challenges in Big Data
- The 4Vs of Big Data
- Approaches in Mining Big Data

Some statistics (2018)

- How much data we create everyday (2018)
  - 2.5 quintillion bytes (2.5 Exabyte (EB) or 2.5 million TB)

- Internet
  - We conduct more than half of our web searches from a mobile phone now.
  - More than 3.7 billion humans use the internet
  - On average, Google now processes more than 40,000 searches EVERY second (3.5 billion searches per day)!
  - Worldwide there are 5 billion searches a day. (70% Google)

Source

Some statistics (2018) – Social Media

- EVERY minute
  - Snapchat users share 527,760 photos
  - Users watch 4,146,600 YouTube videos
  - 456,000 tweets are sent on Twitter
  - Instagram users post 46,740 photos
- Facebook
  - 1.5 billion of the 2 billion users are active on Facebook **daily**
  - More than 300 million photos get uploaded per day
  - Every minute there are 510,000 comments posted and 293,000 statuses updated

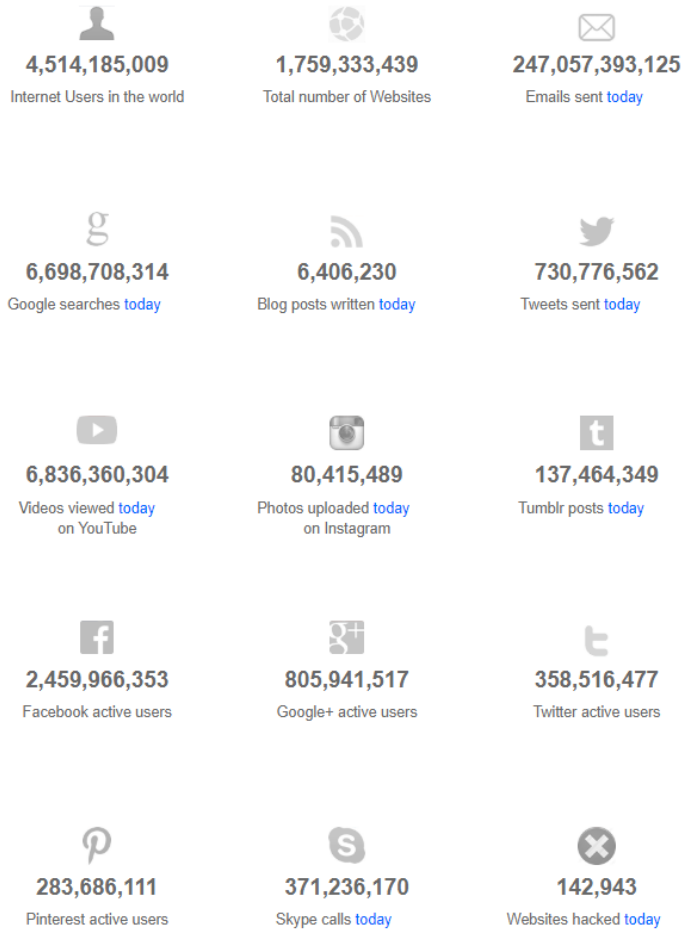
Some statistics (2018) – Communication/Photos

- EVERY Minute
  - 16 million text messages
  - 156 million emails are sent; 2.9 billion email users by 2019
  - 15,000 GIFs are sent via Facebook messenger
  - Every minute there are 103,447,520 spam emails sent
  - There are 154,200 calls on Skype
- Digital Photos
  - People will take 1.2 trillion photos by the end of 2017
  - There will be 4.7 trillion photos stored

Some statistics (2018) – Services/IoT

- Services – Every Minute
  - The Weather Channel receives 18,055,556 forecast requests
  - Venmo processes $51,892 peer-to-peer transactions
  - Spotify adds 13 new songs
  - Uber riders take 45,788 trips!
  - There are 600 new page edits to Wikipedia
- IoT
  - from 2 billion devices in 2006 to a projected 200 billion by 2020
  - One of the primary drivers for our data vaults exploding

Internet Live Statistics



https://www.internetlivestats.com/

Big data: What is it?

- The term **Big Data** is used to address the software/hardware technologies required to implement applications that treat huge quantity of data.

The fact

- Massive quantity of data is generated in numerous application fields: web, social communities, biology, physics, medicine, …

Challenges

- When having a huge quantity of data then tasks such that storage, simple operations, transfer, … become difficult.

Definition: Big Data

Big Data is used in the singular and refers to a collection of data sets so large and complex, it's impossible to process them with the usual databases and tools. Because of its size, *Big Data* is hard to capture, store, search, share, analyze and visualize.

Since standard techniques do not work, different approaches have to be used, which presents new challanges.

What is collecting all this data?

Some examples

Web Browsers

Microsoft's Internet Explorer

Mozilla's FireFox
(Non-profit foundation, used to be Netscape)

Google's Chrome

Apple's Safari

Search Engines

Google's

Microsoft's

Yahoo's

IAC Search's

What is collecting all this data?

Some examples

Smartphones & Apps

Apple's iPhone
(Apple O/S)

Samsung, HTC.
Nokia, Motorola
(Android O/S)

RIM Corp's Blackberry
(BlackBerry O/S)

Tablet Computers & Apps

Apple's iPad

Samsung's Galaxy

Amazon's Kindle Fire

What is collecting all this data?

Some examples

Games Boxes and GPS Systems

Internet Service Providers

What is collecting all this data?

Some examples

“Smart” devices with built-in Internet connectivity

Movie Rental Sites

NETFLIX

redbox

amazon instant video
Connecting

What is collecting all this data?

Some examples

Hospitals & Other Medical Systems

Pharmacies
Laboratories
Imaging Centers
Emergency Medical Services (EMS)
Hospital Information Systems
Doc-in-a-Box
Electronic Medical Records
Blood Banks
Birth & Death Records

Banking & Phone Systems

verizon

Can you hear me now?
(Heh heh heh!)

Sprint

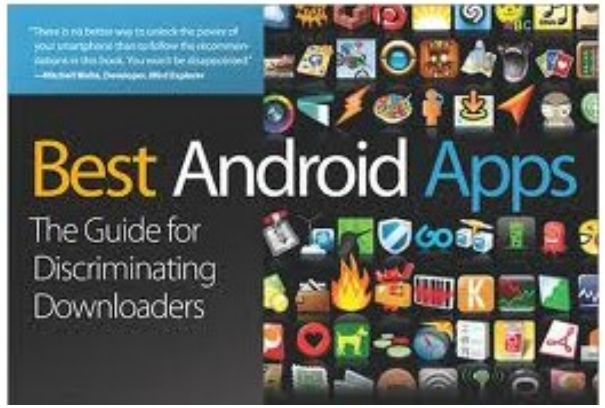
T · ·Mobile·

at&t

What is collecting all this data?

Some examples

Apps!



What are they collecting?

- Restaurant reservations (Open Table)
- Weather in L.A. in 3 days (Weather+)
- Side effects of medications (MedWatcher)
- 3-star hotels in New Orleans (Priceline)
- Which PC should I buy and where (PriceCheck)

What is collecting all this data?

Some examples

“Smart” cards

Octopus card HK

Macau Pass

Opal card

Credit, bank, shopper, reward cards

What is collecting all this data?

Some examples

Others:

- Internet of Things
- Security systems (i.e. surveillance cameras in public spaces)
- Utility services (i.e. smart meters)
- Traffic monitoring systems
- Ticketing systems
- Smart cars
- Raster microscopes, Large hadron collider,…

…and these were just a few examples!

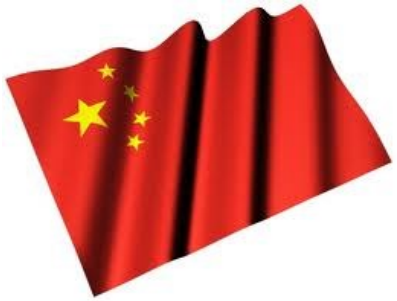
Who is collecting all this data?

Some examples

Government Agencies

IRS

Department of the Treasury
Internal Revenue Service



Big Pharmaceutical Companies

Pfizer

BAYER

Roche

MERCK

AstraZeneca

gsk GlaxoSmithKline

Who is collecting all this data?

Some examples

Consumer Products Companies

P&G

KRAFT

Big Box Stores

Sam's CLUB

THE HOME DEPOT

COSTCO WHOLESALE

STAPLES

WAL★MART

ALWAYS LOW PRICES.

Always.

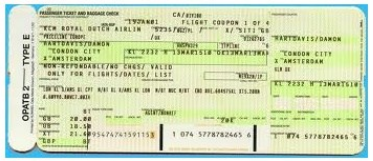
Who is collecting all this data?

Some examples

Credit Card Companies

What data are they getting?

Airline ticket



Restaurant check

1	Cab Sauv Corvina	£15.25
1	Cab Sauv Glass	£4.00
26	Net Sales Total	£73.45
1	SUCK MY D[illegible] F[illegible] FACE	£0.00
1	FISH CAKES	£4.95
2	1/2 Wings Starter	£7.90
1	MELON AND PARMA HAM	£3.95
1	Calamari	£4.95
1	Garlic brd starter	£2.25
3	Meatball Starter	£17.85
1	Aub & Feta Starter	£3.95

Grocery Bill

Hotel Bill

Why are they collecting all this data?

Data is the new Oil. Data is just like crude. It's valuable, but if unrefined it cannot really be used.

– Clive Humby, DunnHumby

We have for the first time an economy based on a key resource [Information] that is not only renewable, but self-generating. Running out of it is not a problem, but drowning in it is.

– John Naisbitt

Why are they collecting all this data?

Target Marketing

- To send you catalogs for exactly the merchandise you typically purchase.
- To suggest medications that precisely match your medical history.
- To “push” television channels to your set instead of your “pulling” them in.
- To send advertisements on those channels **just for you**!

Targeted Information

- To know what you need before you even know you need it based on past purchasing habits!
- To notify you of your expiring driver’s license or credit cards, provide customized services, etc.
- To give you turn-by-turn directions to a shelter in case of emergency.

Outline

- Some Statistics
  - How much data is created every day
- What is collecting all this data
- Who is collecting all this data
- Why are they collecting all this data
- Challenges in Big Data
- The 4Vs of Big Data
- Approaches in Mining Big Data

Examples of Big Data

- Walmart handles more than 1 million customer transactions ***every hour***, which is imported into databases estimated to contain more than 2.5 petabytes * of data.
- The FICO Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.
- Australian Health Insurance Commission collects terabytes of medical claims data every year.
- The volume of business data worldwide, across all companies, ***doubles every 1.2 years***, according to estimates.

(1 Petabyte = 1000000000000000B = 1000^5 B = 10^{15} B = 1 million gigabytes)

*** Think of a large hard drive on your computer at home having 10,000 gigabytes. Now multiply that by 250!**

Challenges: Storage

- Google may store more than 100 billion web pages: assuming 20KB per page, storing the web requires 2000TB.
- Amount of data continues to grow at an exponential rate.

Bad news: no storage disk can store 2000TB

Good news: a 10TB disk costs only $550 (as of 2019)

The challenge: How to distribute data across many disks to construct a big archive?

Challenges: Hardware

- For very big archives even simple operations become difficult. Suppose that a copy of the web (2000TB) is stored on disks. Just to count the number of the pages in the web you have to read 2000TB.

Bad news: Reading the data from a single disk would take ~1.5 months (Top SSD disks can reads up to 500 MB/sec), more time is required for doing something useful with data.

Good news: Reading the data from 1000 disks in parallel (by 1000 PCs) takes a little bit more than 1 min and … PCs are relatively cheap.

The challenge: Physically connect and make interact thousands of PCs.

Challenge: Reliability

- Big data requires several computers, disks and resources in parallel. Some big data are so big that this requires a very large number of computers.

Bad news: Say we estimate that mean time between failures of a PC is 200 days. Hundreds of PCs fail per day in a big company which has tens of thousands of PCs.
Good news: PCs and components are cheap and easily replaced.

The challenge: Fault tolerance: Coordinate the activities of PCs so their tasks are executed continuously without loss of data even if many PCs or storage devices fail concurrently.

Challenge: Algorithms

- Given that Big Data requires the use of several computers in parallel and hence the methods for processing the data need to be able to work on a distributed set of computers.

Bad news: Writing distributed programs is very difficult.

Good news: There are new software tools that offer new programming paradigms.

The challenge: Learning to write distributed programs.

Challenge: Algorithms...an example

- Even simple tasks can become a problem for distributed programs.
- Example: Sort a *list* of books by the name of the authors.
- A possible non-distributed approach:

```
1. Create one page per letter in the alphabet (26 pages).
2. Repeat for each book
3.    Take a book and write the name of the author onto the page
      that corresponds to the first letter of the authors name.
4. End
5. Repeat for each page that has more than one name
6.    Split the list of names onto new pages according to the
      second letter of the authors name.
7. End
8. Repeat steps 5 to 7 for the third, fourth, (and so on)
   letter.
9. Combine (in order) all pages that contain just one name.
```

Challenge: Algorithm...an example

- But how to do this in parallel?
- Say we have 26 workers (one for each letter in the alphabet) :

```
1. Each worker is responsible for one of the letters in the
   alphabet.
2. For each worker
3.    Take each book and write the name of the author onto a
      page if the first letter of the authors name corresponds
      to the responsibility of the worker.
4. End
5. Repeat for each page that has more than one name
6.    The workers split the list of names onto new pages
      according to the second letter of the authors name.
7. End
8. Repeat for the third, fourth, (and so on) letter.
9. Combine (in order) all pages that contain just one name.
```

Challenge: Algorithm...an example

- Some problems with this solution:
  - Say we wish to sort the books in the Australian National Library (5 million books).
    1. Each of the 26 workers has to access each of the books (26 x 5 million = 130 million accesses)
    2. A book cannot be accessed by more than one worker at a time.
       - Requires synchronization among workers.
    3. How to involve workers if we have more than 26 workers available (i.e. hundreds or thousands of workers)?

- There are better algorithms for such a task. But this example shows the difficulty in devising parallel algorithms.

Back to the definition of Big Data

"Big Data is used in the singular and refers to a collection of data sets so large and complex, it's impossible to process them with the usual databases and tools. "

Notice:

- Only an intuitive definition is available
- The definition is not precise:
  - the difference between standard and non-standard techniques is not clear
  - The limit over which a task is defined as big is not clear.

When is data “Big”?

- “Data is called big when the data is so large that usual databases and tools do not work.”
- This can mean that a Big Data task involves a lot of data but is not necessarily difficult.
  - Can you think of a task your laptop can solve on a huge archive?
- Example: For all the inhabitants on earth (7.5 billions) find those born on “1/1/2000”.
  - Can this be done by a laptop and, if so, how much time is required for the answer?
- Assume 50 Bytes per record, the table requires 375GB, it would fit into a laptop hard disk.
- Reading at 500MB/s the table can be read in about 12min … it works, but it is not user friendly.

Obviously **BIG** data

- Global smartphone traffic: 300 million GB
- Ebay: 7.5 million GB
- Google: 100 million searches per month in 2012
- Utah Data Center (NSA): billions of GBs
- LHC: around 500 billions of GB … per day
- Modern DNA sequencers: Produce the DNA sequence of an animal in only one day (human genome includes 3 billion of bases)

When is data big?

The difficulty of a task depends not only on the dimension of the data but also on other issues:

- The speed at which the data is received
- The speed at which data can change
- The speed at which the data has to processed
- The complexity of the problem to be solved
- …

- A precise definition does not exist.
  Don't care! Just focus on the fact that novel techniques are required!

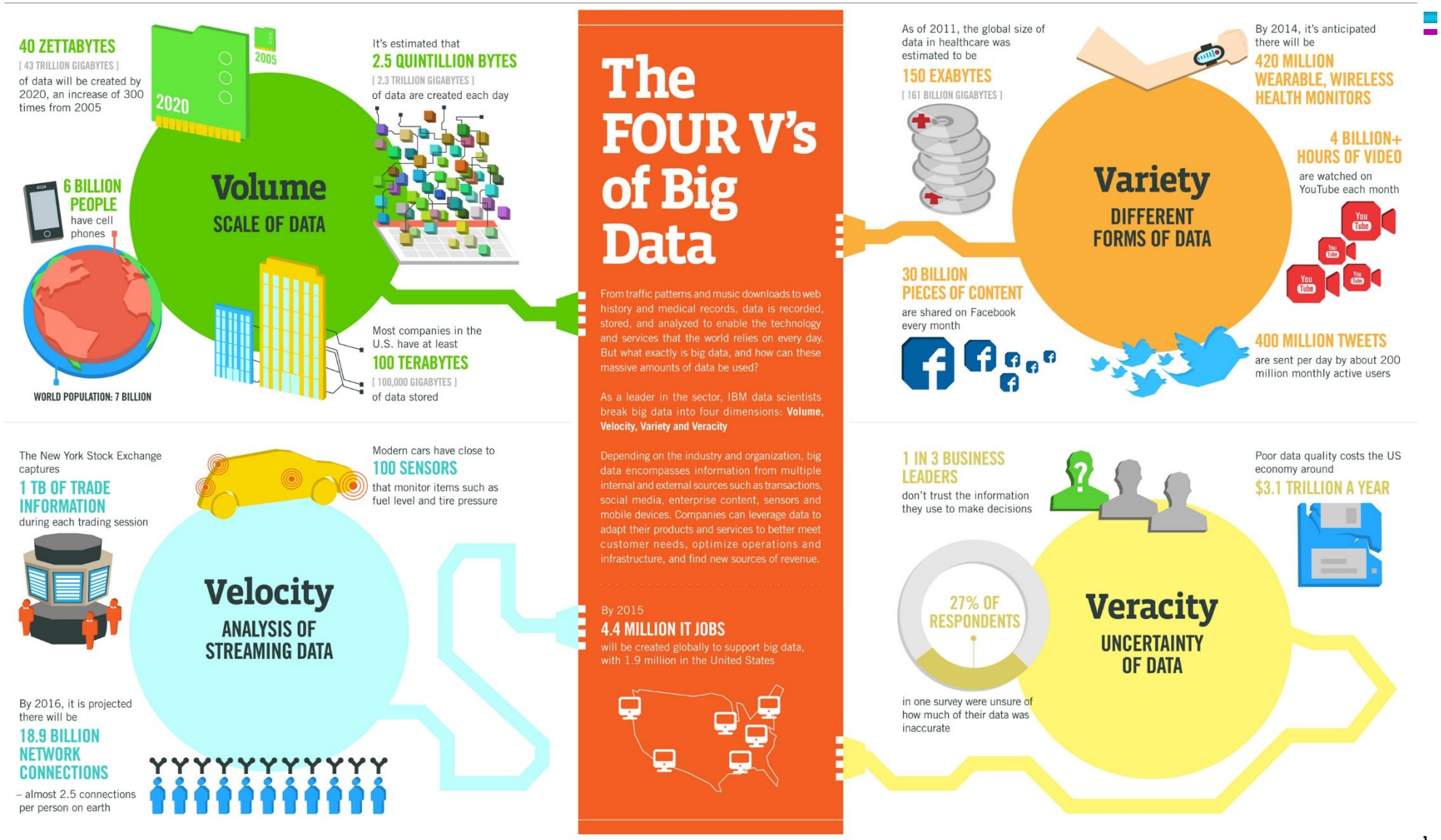
Confused by the terminology?

Data mining and Big Data are similar concepts. Key differences are:

- A focus of Big Data is on the technologies related to large data (volume) management whereas for data mining the focus is on the application of algorithms to extract valuable information.
- Big Data concerns data that continues to grow (accumulate), comes from different sources (variety), and can change (velocity) whereas datasets in Data Mining are assumed static.
- Data quality in Big Data is commonly less certain (veracity) when compared to Data Mining applications.

These differences are known as the four Vs.

The 4 Vs of Big Data.



The 4Vs

Big data embodies data characteristics created by today's digital marketplace:

- **Volume**: Data at scale up to terabytes, petabytes.
- **Variety**: Data in many forms, structured, unstructured, text, multimedia, sensory.
- **Velocity**: Data in motion, streaming data
- **Veracity**: Data uncertainty, imprecise data types.

What gave rise to Big Data?

Big Data is not new. The ability to collect vast amount of data has been around for many years.

New is that we have entered the **Golden Era for Big Data**:

- Advanced Software Methodologies.
- Powerful multi-core and GPU processors.
- Virtualization leveraging the powerful hardware.
- Wider communication bandwidth.
- Penetration of communication technology. The internet has become omnipresent
- Significantly reduced costs for the transport of data, data storage, and processing.

We now have the means to work with Big Data!

How to work with Big Data?

“We swim in a sea of data … and the sea level is rising rapidly.”

Pew Research Center’s Internet & American Life Project - July 2012

Approaches to Mining Big Data

Approaches to working with Big Data must be:

- Simple,
- Scalable,
- or both.

- Very active area or research!

Simple approaches to Big Data

With simplicity comes risk!

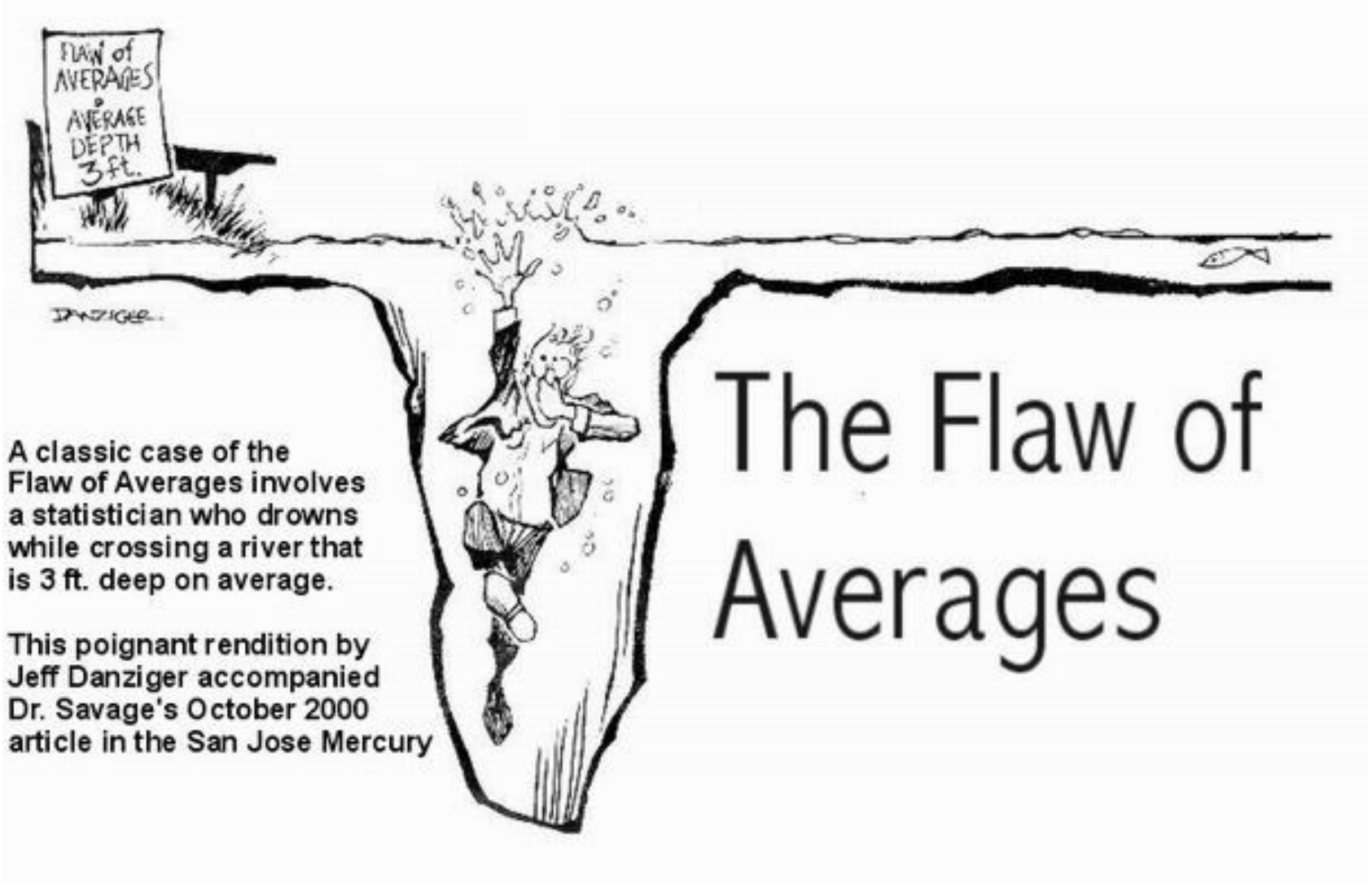
- Valuable information may be lost.
- Oversimplification can distort results, or produce incorrect results.

Example: Working with averages.

- Computing an average is one of the simplest operations on any dataset.
- Average of N quantities $X_1, X_2, \ldots, X_N$.
- Average: $(X_1 + X_2 + \ldots + X_N)/N$.

But averages can be misleading.

Flaws of Averages



Flaws of Averages

Flaws of averages, some examples:

- Mutual Fund: *Average total annual returns --- 7%.*
- When Bill Gates walks into a crowded establishment, on average everyone becomes a millionaire!
- The mean salary of a tech worker in San Mateo County is $291,497.

  $1 of this is due to Mark Zuckerberg!
- Garrison Keeler (author): *Lake Wobegon, where all the women are strong, all the men are good looking and all the children are above average.*
- *Is this even Possible?*

Flaws of Averages

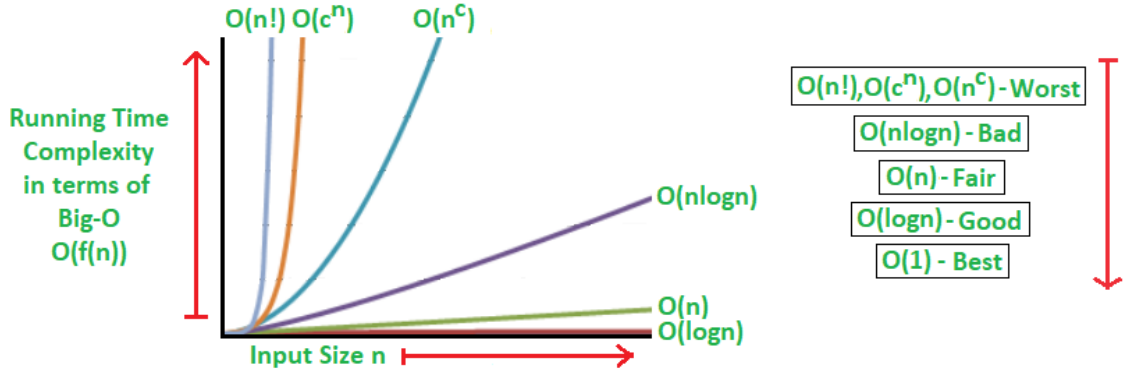
- “Outliers”: What to do with them?
- Many say, clip them off, they distort the analysis, they mislead intuition.
- But, “outliers” have determined the course of human history.
  - Meteors hitting Planet Earth
  - Richter 9 and above earthquakes
  - Financial collapses.
- Removal of outliers may remove essential information.

Summary points on simple methods

- Simple methods in Big Data can be deceiving.
- Can result in incorrect references.
- The results as experienced by one population may be very different from those experienced by others.
- Don’t ignore outliers at your peril. **Know your data!**

Scalable Methods for Big Data

- So, if not simple, then we need scalable methods.
- Scalable with the amount of data
- Streaming methods
  (data processing in real time)



- This gave rise to AI
- Some machine learning algorithms are massive parallel systems that can take advantage of multi-core CPU and GPU technology.
- Examples: MLP, SOM, Deep Learning, … these are algorithms that can exploit the massive parallel computing infrastructure.

How to work with Big Data?

- Big Data analytics draws upon techniques from a number of established fields:
  - Statistics
  - Artificial Intelligence
  - Machine Learning
  - Data mining
  - Social Network Analysis
  - Text Mining and Web Analytics
  - Operational Research
  - Information Visualization

- Application domains such as business intelligence, national security intelligence and learning analytics all have an interest in analysing large volumes of data from disparate data sources and are providing the business cases for the rapid growth in ‘big data’ & data analytics.

What Big Data Is Not

What Big Data is Not:

- It is not a replacement for Data Mining.
- It is not a replacement for Databases.
- It is not a solution by itself, it needs algorithms, users, jobs, applications to drive value.

In Conclusion

- Big Data is still in its infancy.
- Significant research interest in solving open problems:
  - Data integration and consolidation
  - Streaming (real-time) methods
  - System design and implementation.
  - Sharing, policies, and standards.
- Big Data is in the process of revolutionizing the digital economy

In Summary

Big data involves:

- Data centers with many computers
- Capability of working with large amount data
  - That may change rapidly, come from different sources, and of limited data quality.
- Capability of rapidly processing data.
  - Process data in real time or else you drown in data.
- Reliability which ensures that software works without any interruption.

Curious and want to know more?

Dedicated subject cover Big Data in greater detail:

- CSCI319 *Distributed Systems and Cloud Computing* : Introduces into the design of distributed systems for big data and other applications. Explains how large number of computers can work together in a secure and fault tollerant fashion. Explains how large databases can be distributed.
- CSCI316 *Big Data Mining Techniques and Implementation*: Covers aspects of algorithms, algortihm design, and implementation of Big Data methods. Preprocessing is an important aspect here due to the uncertainty of data quality in Big Data applications.