Evil or Not Evil:
Ethical Issues in Big Data Era

James Ma

jma@uccs.edu

April 17, 2019
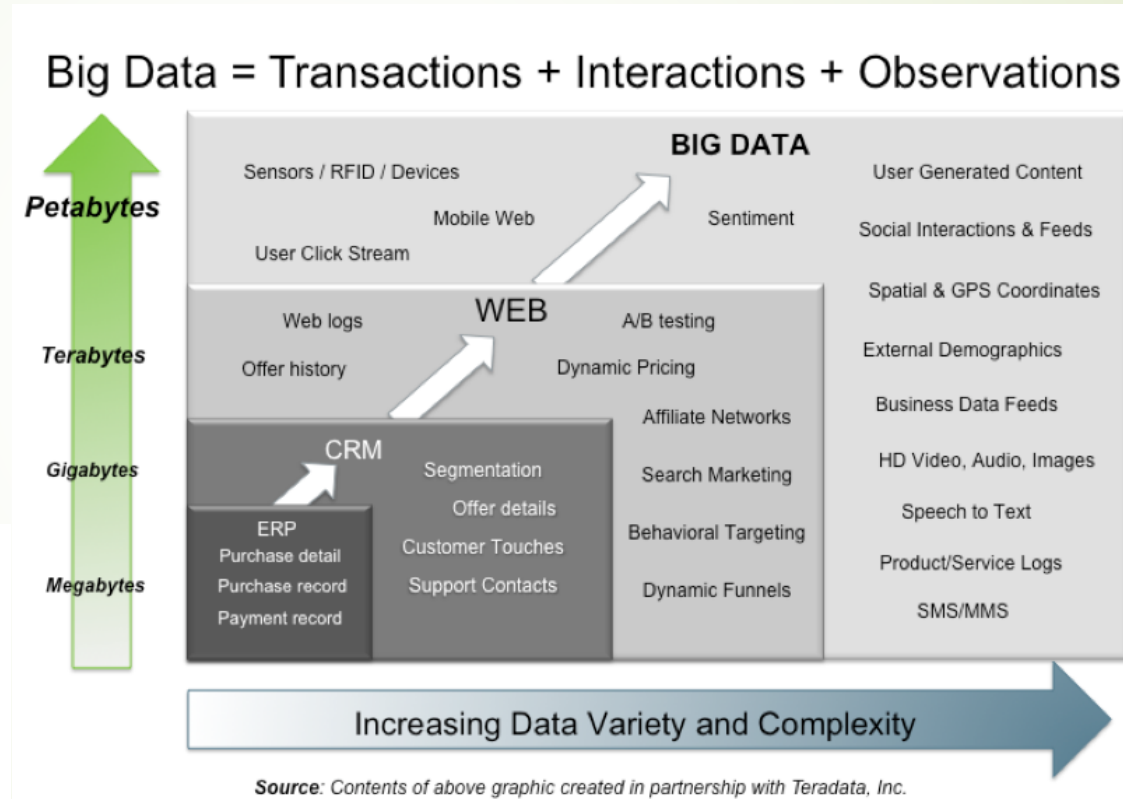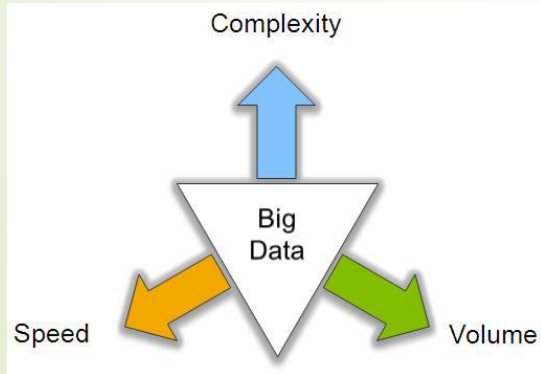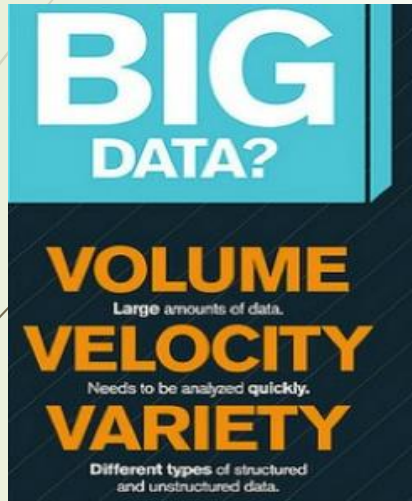
1

What is Big Data

- "Big Data" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

Big Data: 3 V's

1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

The Digital Universe 2009-2020

Growing By A Factor Of 44

2009: 0.8 Zb

2020: 35.2 Zettabytes

Data storage growth

In millions of petabytes (One petabyte = 1,024 terabytes)

| terabytes | petabytes | exabytes | zettabytes |

the amount of data stored by the average company today

Twitter: Tweets Per Day
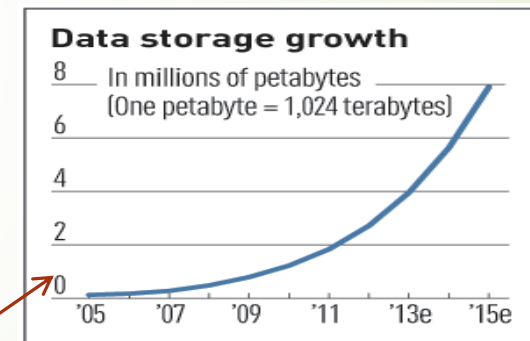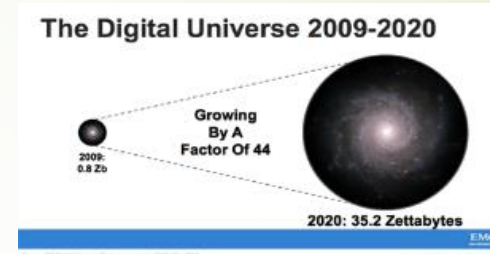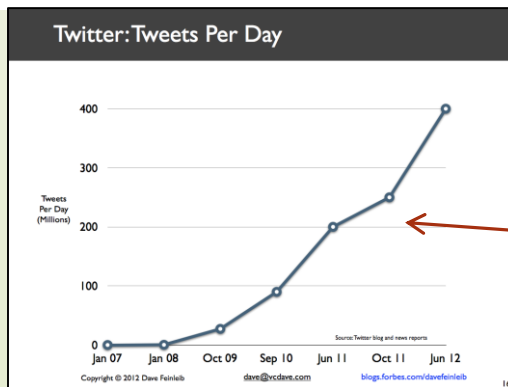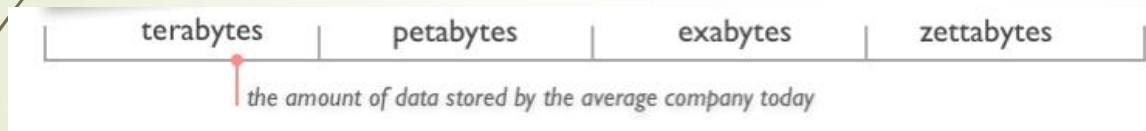
Exponential increase in collected/generated data

2-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- Examples
  - **E-Promotions**: Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you

  - **Healthcare monitoring**: sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

3-Complexity (Varity)

- Various formats, types, and structures

- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…

- Static data vs. streaming data

- A single application can be generating/collecting many types of data

To extract knowledge➜ all these types of data need to linked together
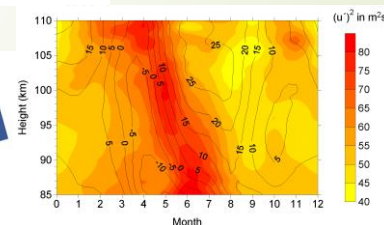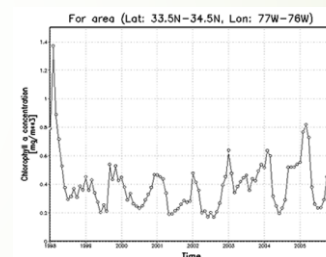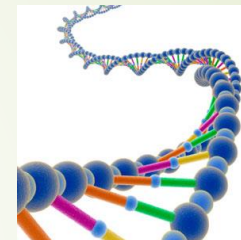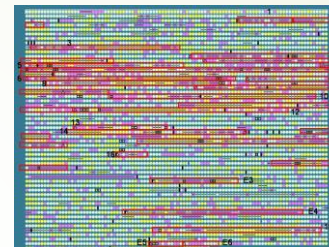
Who's Generating Big Data

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data

- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Core: Distributed Computing
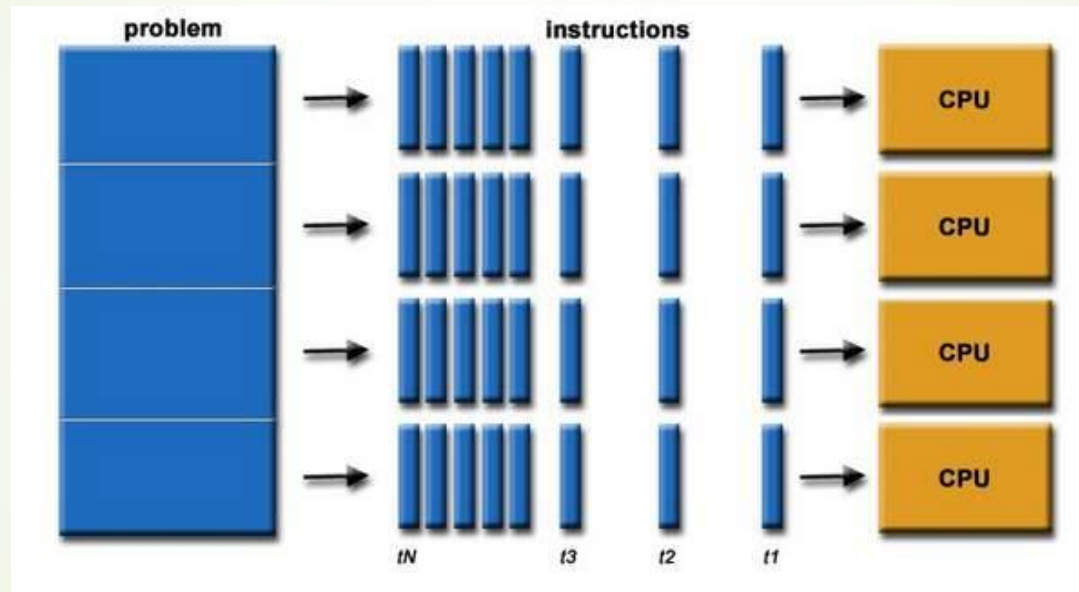
- Task parallelism, breaking the task down into its constituent parts and processing them simultaneously.

Cloud Computing

- IT resources provided as a service online
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth networks & multicore processors
- Offerings from Microsoft, Amazon, Google, …

Cloud Computing
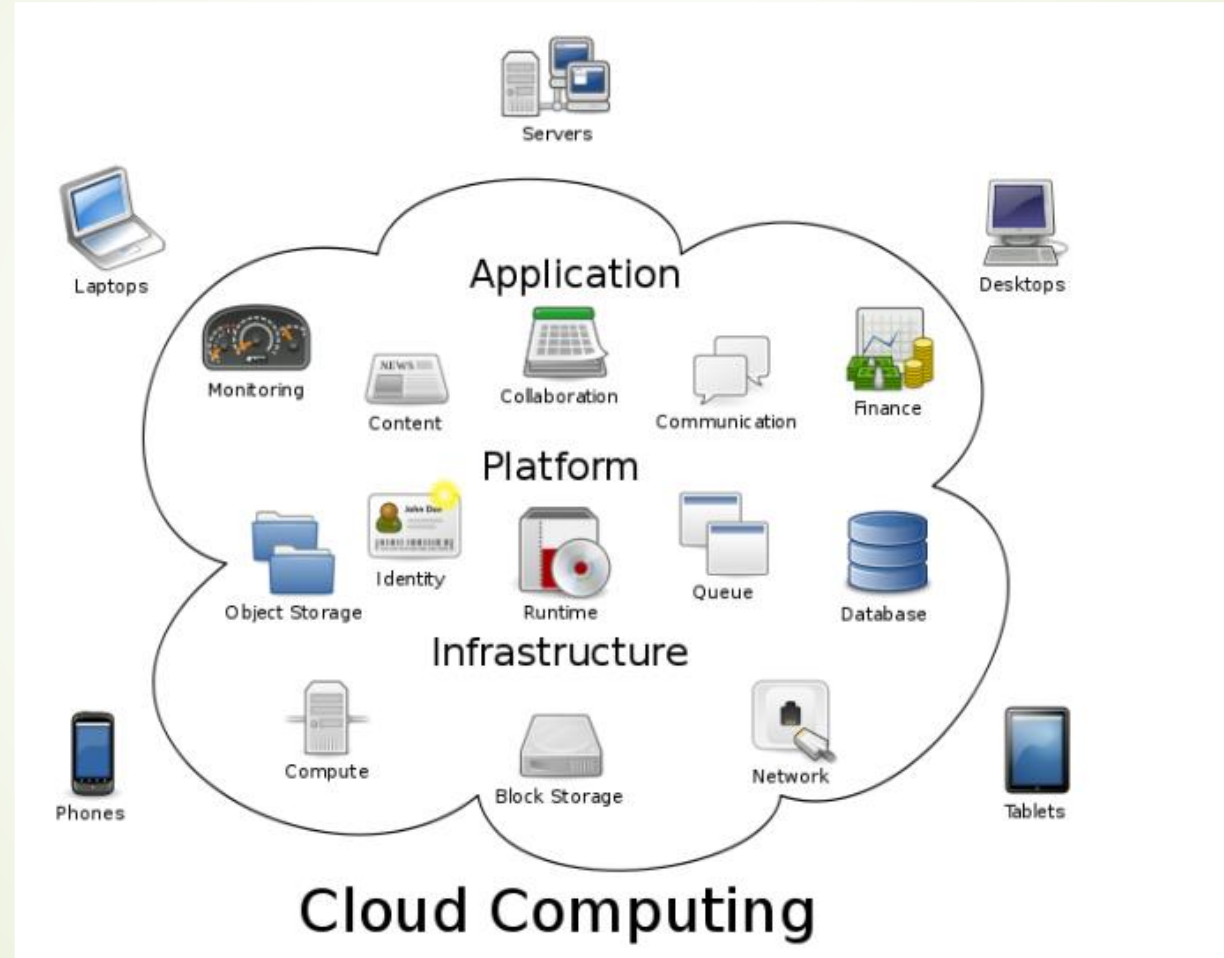
Cloud Computing Benefits

- Cost & management
  - Economies of scale, "out-sourced" resource management
- Reduced Time to deployment
  - Ease of assembly, works "out of the box"
- Scaling
  - On demand provisioning, co-locate data and compute
- Reliability
  - Massive, redundant, shared resources
- Sustainability
  - Hardware not owned

Big Data – The Future

- Data sets will continue to grow
- Storage unit costs will continue to decrease
- Processing costs will decrease
- Network capacity will continue to grow
- Data growth may exceed processing capacity

Big Data – The Concerns …

- Data security
- Too much reliability on Internet access
- And the ethical issues …

Ethics Definitions

- Ethics
  - A set of principles of right conduct or a theory or a system of moral values
- Ethical behavior
  - "Doing the right thing"
- Note
  - The "right thing" is society-based

Ethical Goals in Data Management

- **Goal**
  - Minimize harm
  - Show respect, i.e.
    - Informed voluntary consent
    - Vulnerable subjects must be protected

General Rules of How to be Ethical

- **Case-based approach**
  - Start with an obvious (real or hypothetical) case
  - Compare it with a problematic case that is structurally similar
  - Draw the same conclusion

Stakeholders at Big Data Era

- Data providers
  - Sensors
  - **And you!**
- "Cloud" owners
- Data analysts

Ethical Choices – Sensors

- What should be collected?
  - Only the necessary (and sufficient) data
- Examples
  - Location based data
    - DriveWise by AllState
    - GPS related ads
  - Image analysis
    - Face recognition

Ethical Choices – "Cloud" Owners

- What should be retained?
  - Only the necessary (and sufficient) data
- Who owns the data that is retained?
- Who can access the data?
- Who can copy the data?
- Always mask unnecessary personal identifiable data
- How long should the data be retained?

Ethical Choices – Data Analysts

- Fabrication of data is always ethical misconduct
- Never "peek inside the envelope"
- Data should be appropriately utilized in the analysis
  - **Data biases: they can if they want to!**

Can They Be Evil?

- **Absolutely!**
- Input
  - Location track
  - Social media post (text)
  - Images
- Output
  - Profiling YOU
  - Predicting YOUR next destination
  - Locating YOUR house

Your Choices

- Choose what and when to be sensed
  - Turn off GPS!
  - Turn off browsing trace
  - Log off before search
- Choose what to post at social media
  - Think twice before posting an opinion
  - Think three times before posting a photo, i.e. selfie
  - Think four times before posting personal info, i.e. address, dob, child's name
- Never think "I am only sharing with friends!"

Your Choices – Last Words

23

- Once it is out there, it never goes away…