CSC590: Selected Topics
BIG DATA & DATA MINING

Lecture 2

Feb 12, 2014

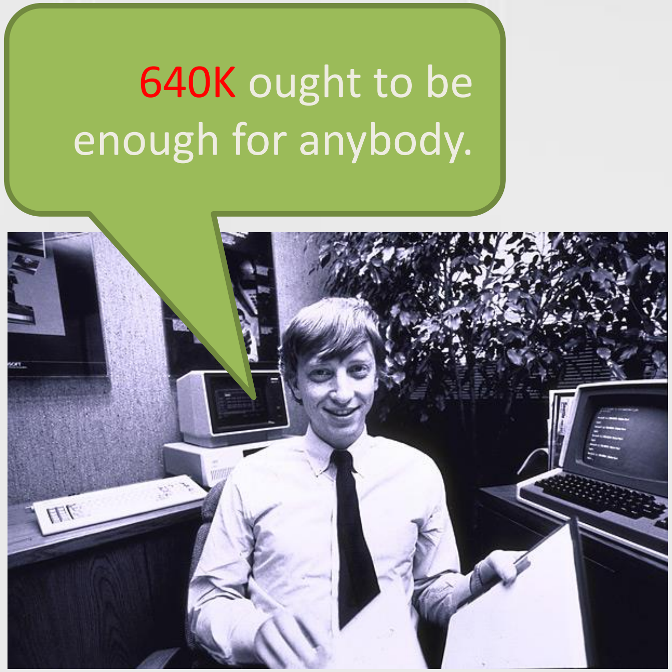
Dr. Esam A. Alwagait

Agenda

- Introduction
- What is Big Data
- Why Big Data ?
- Characteristics of Big Data
- Applications of Big Data
- Problems of Big Data

Introduction

- March, 2012, the Obama Administration announced $200 million in R&D investments for Big Data. http://www.cccblog.org/2012/03/29/obama-administration-unveils-200m-big-data-rd-initiative/



What is Big Data ?

- Big Data usually includes data sets with sizes beyond the ability of commonly used software tools to *capture*, *manage*, and *process* the data within **a tolerable elapsed time**
- "Big Data" is what happens when the cost of storing data falls below the cost of deciding to throw it away. -- George Dyson

What is Big Data ?

- Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few.

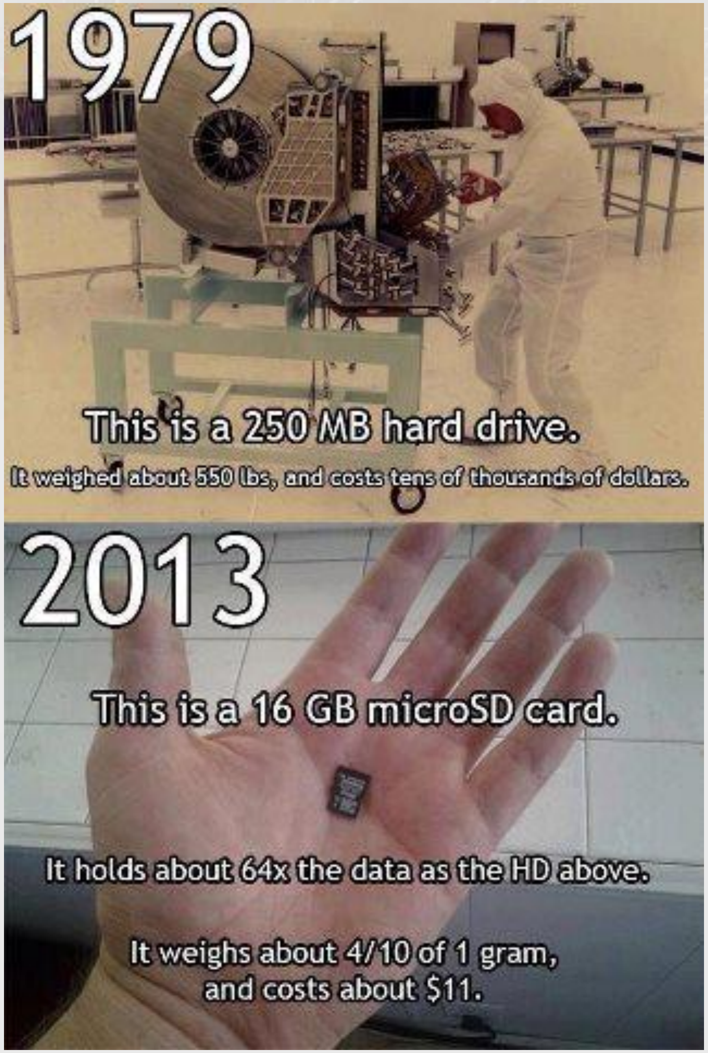
This data is “big data.”

What is Big Data ?

- There are huge volumes of data in the world:
  - From the beginning of recorded time until 2003,We created 5 billion gigabytes (exabytes) of data.
  - In 2011, the same amount was created every two days
  - In 2013, the same amount of data is created every 10 minutes.

Why Big Data ?

- Why now ? What enabled Big Data ?
  - Storage became cheaper
  - Processing power got Faster AND cheaper
  - Data became available
- What does that mean ?
  - It is easy to get data.. But you can’t decide whether you want to throw it away or not !?!



1979

This is a 250 MB hard drive.

It weighed about 550 lbs, and costs tens of thousands of dollars.

2013

This is a 16 GB microSD card.

It holds about 64x the data as the HD above.

It weighs about 4/10 of 1 gram, and costs about $11.

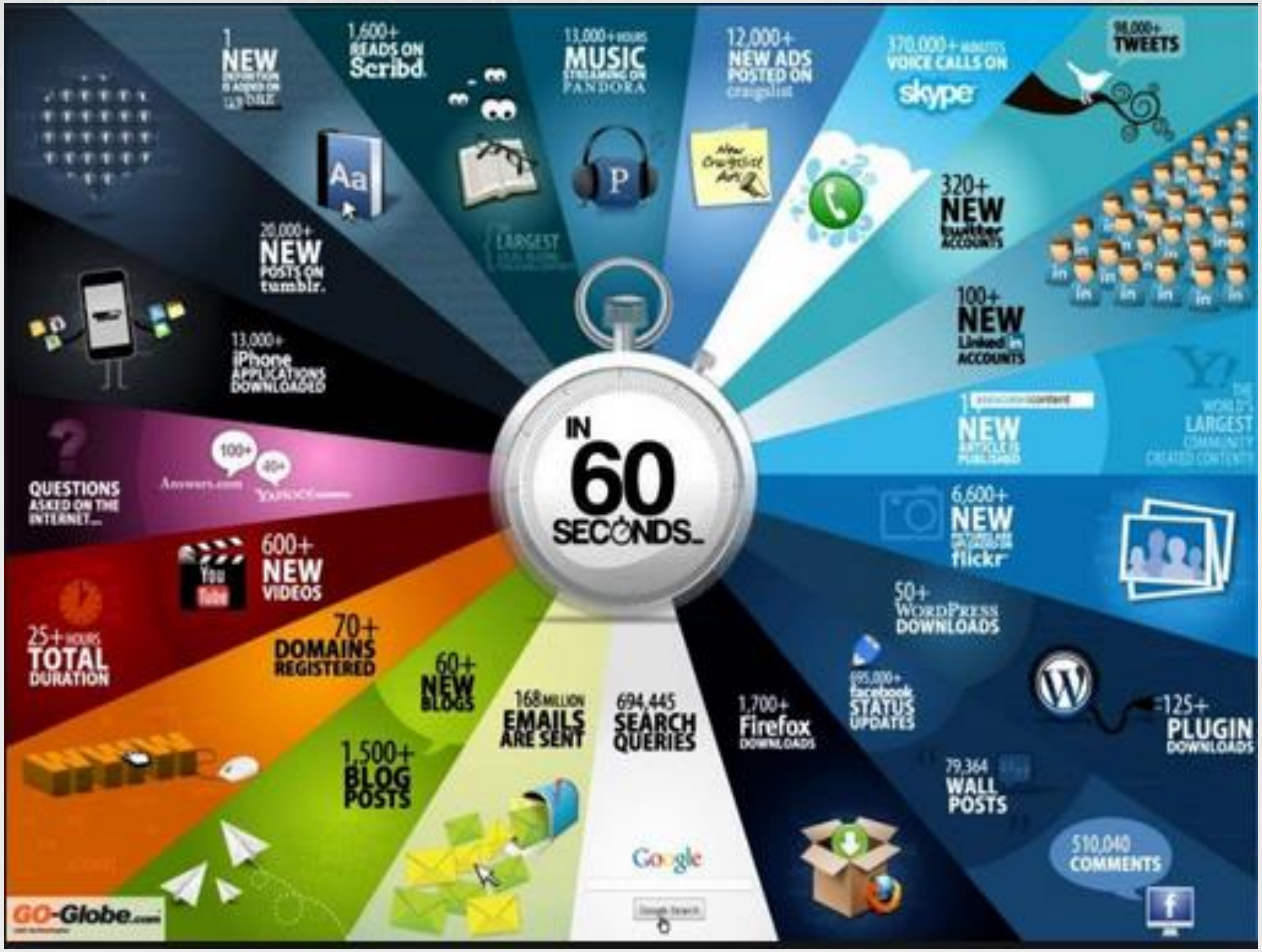
Characteristics of Big Data

- WWW → VVV
- Volume
- Velocity
- Variety
- Veracity ?

Volume

- Enterprises are awash with ever-growing data of all types, easily amassing terabytes—even petabytes—of information.
  - Turn 12 terabytes of Tweets created each day into improved product sentiment analysis
  - Convert 350 billion annual meter readings to better predict power consumption

Volume – Cont'd



Volume – Cont’d

- Twitter generates more than 7 Terabytes (TB) a day; Facebook more than 10 TBs, and some enterprises already store data in the petabyte range.

Velocity

- Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into your enterprise in order to maximize its value.
  - Scrutinize 5 million trade events created each day to identify potential fraud
  - Analyze 500 million daily call detail records in real-time to predict customer churn faster
- TPS Tweets Per Second
  - Record high of 150,000 TPS

Variety

- Big data is any type of data - structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together.
  - Monitor 100’s of live video feeds from surveillance cameras to target points of interest
  - Exploit the 80% data growth in images, video and documents to improve customer satisfaction

Variety – Cont’d

- Web data, e-commerce
- purchases at department/ grocery stores
- Bank/Credit Card transactions
- Social Network

Veracity

- Amazon:
  - Book recommendations
  - “People who bought this, also bought ..”
- IMDB
  - Similar movies
- Google
  - “Did you mean … “?
  - Translations..
- Twitter & Facebook
  - Do you know this person ?
- Wallmart
  - Here is a 10% coupon to …

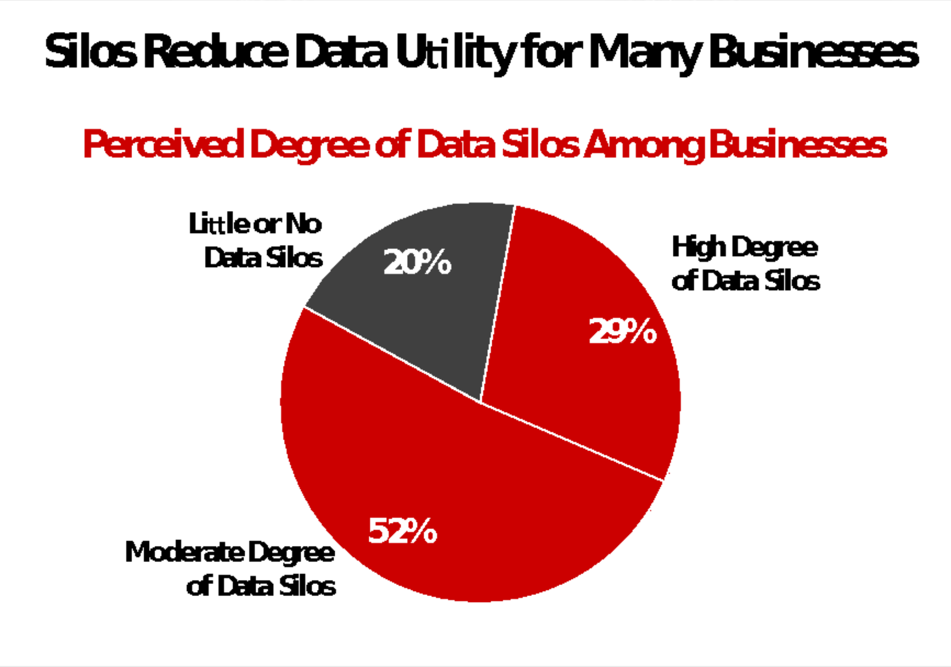
Applications of Big Data

- Government
  - Data.gov
- Health
  - 7 big data solutions for Healthcare
- Education
  - www.coursera.org
- Banking
  - Loans, Stock market ..etc.
- Commercial
  - Amazon, Wallmart…etc !

Problems

- Data Silos

collections of data that have grown across the organization or within specific departments which are not connected in a cohesive plan.



Silos Reduce Data Utility for Many Businesses

Perceived Degree of Data Silos Among Businesses

Problems – Cont’d

- Experts & Technology
  - Still there are not enough “Data Scientists” to accommodate for the demand
  - Technology is not yet mature to help obtain the wisdom needs

Problems – Cont’d

- Privacy
  - You are the product !
- Changing Infrastructure