Technology and Application of Big Data

Qing LIAO(廖清)

School of Computer Science and Technology

HIT

Course Details

- Instructor:
  - Qing LIAO, liaoqing@hit.edu.cn
  - Rm. 303B, Building C
  - Office hours: by appointment

- Course web site:
  - liaoqing.me

- Reference books/materials:
  - Big data courses from University of California
  - Book: BIG DATA: A Revolution That Will Transform How We Live, Work, and Think
  - Papers

- Grading Scheme:
 - Paper Report 30%
  - Final Exam 70%

What You Learnt: Overview

- Topics:
  1) Introduction of Big Data
  2) Characterizes of Big Data
  3) How to Get Value from Big Data
  4) Technologies of Big Data
  5) Applications of Big Data

- Prerequisites
  - Statistics and Probability would help
    - But not necessary
  - Machine Learning & Data Mining would help
    - But not necessary

Introduction

- What is the Big data?

Data → Information Explosion → BIG DATA
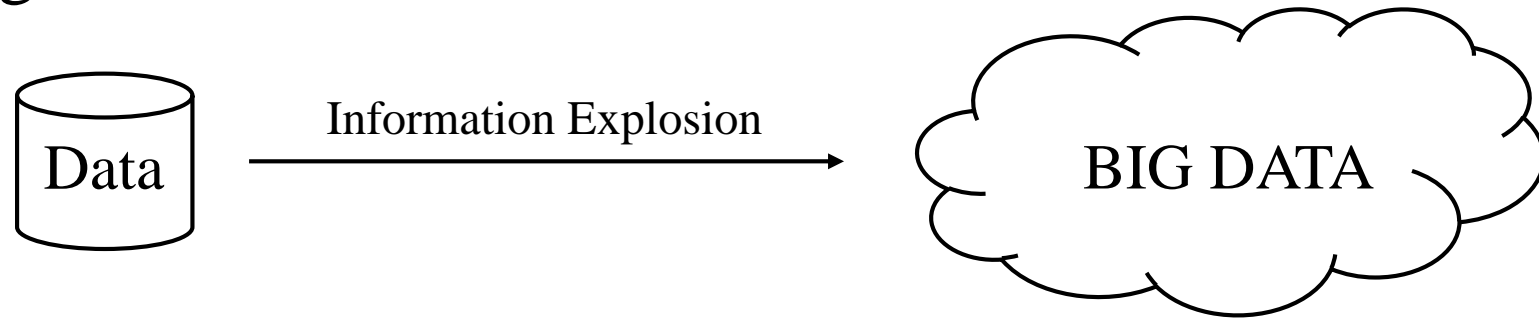
- What Launched the Big Data?

Two Opportunities   Changing Times

Introduction

- What Launched the Big Data?

Data Science → #1 Catalyst for economic growth!

--McKinsey Report (2013)

First Opportunity:
Big Data – a growing torrent
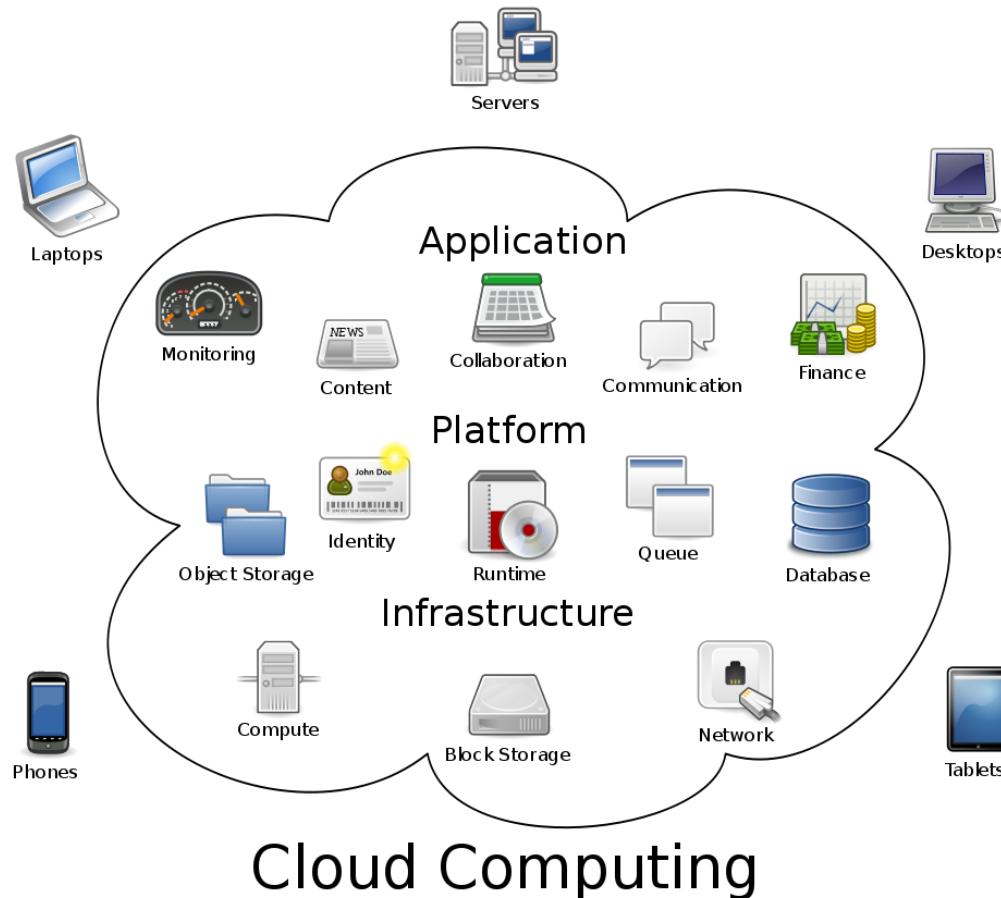
Introduction

- What Launched the Big Data?

Sina Weibo has 222m as of Q3 2015,
with 85% of them being on mobile devices

Introduction

- What Launched the Big Data?

Second Opportunity:
Computing Anytime, Anywhere

Introduction

- What Launched the Big Data?
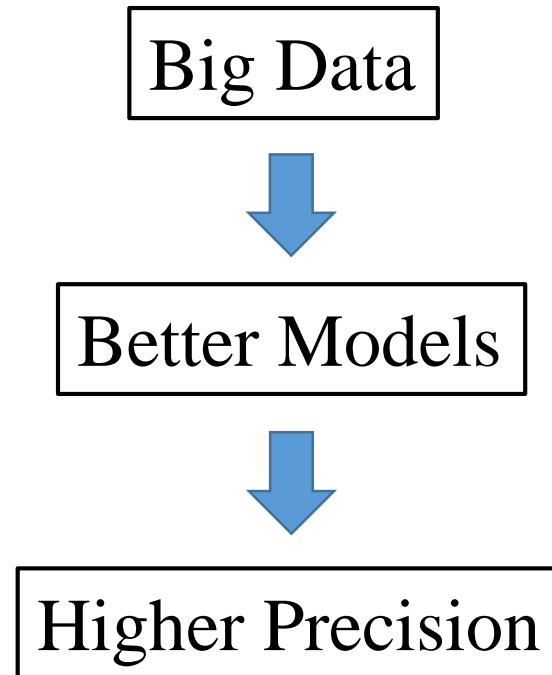

First Opportunity: Data Torrent

Second Opportunity:
Computing Anytime, Anywhere

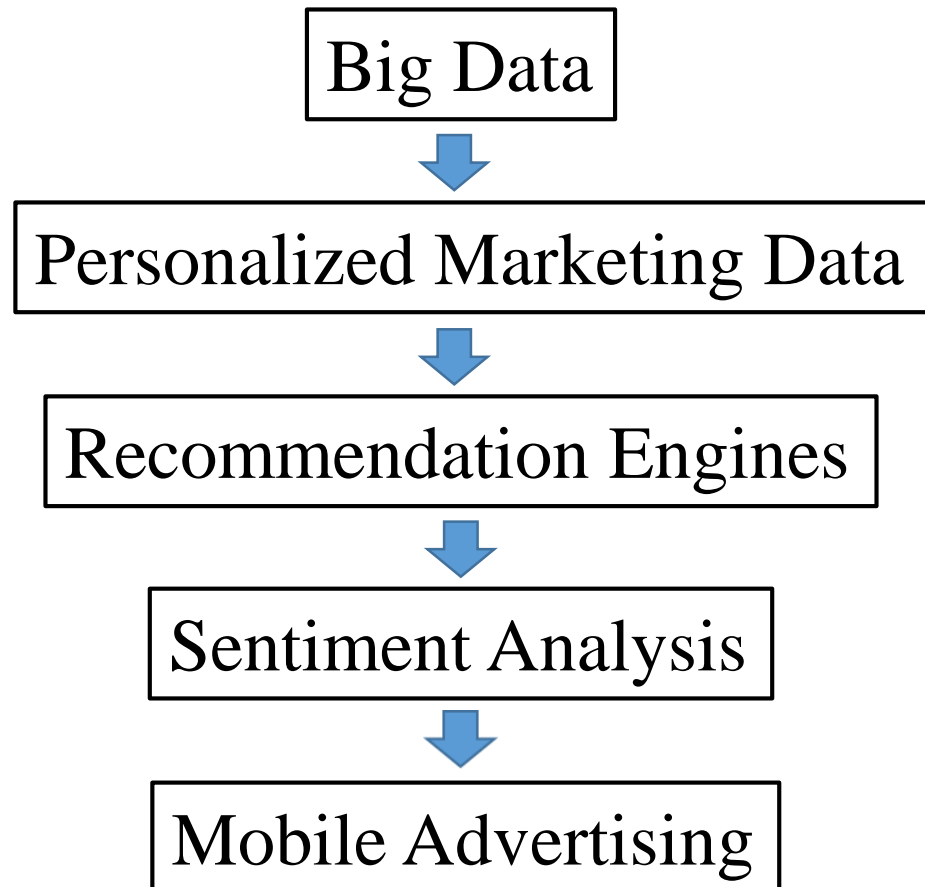Launched the Big Data Era

Introduction

- What Makes Big Data Valuable?

Big Data

↓

Better Models

↓

Higher Precision

Applications help people to make decision!

Introduction

Buy?

- What Makes Big Data Valuable?

Big Data

↓

Personalized Marketing Data

↓

Recommendation Engines

↓

Sentiment Analysis

↓

Mobile Advertising

⭐️☆☆☆☆ **Ruined after first wash.**

By Michelle DeBest on October 16, 2015

I purchased these to add to a set I had after one plate had broken. Although I had washed my existing set in my dishwasher (set to air dry) many times without issue, two of the 4 new plates were ruined upon the initial wash. Apparently these are actually clear glass plates with some sort of blue coating, which peeled off during the washing process, leaving me with two partial clear/partial blue plates with visible peeling. I suppose I can try and scrub off the remainder of the blue coating and have usable clear plates, however - if I wanted clear plates that is what I would have purchased. Heed the warning to be very careful when washing these.

Natural Language Processing

Introduction

- What Makes Big Data Valuable?

Cost per Genome

genome.gov/sequencingcosts
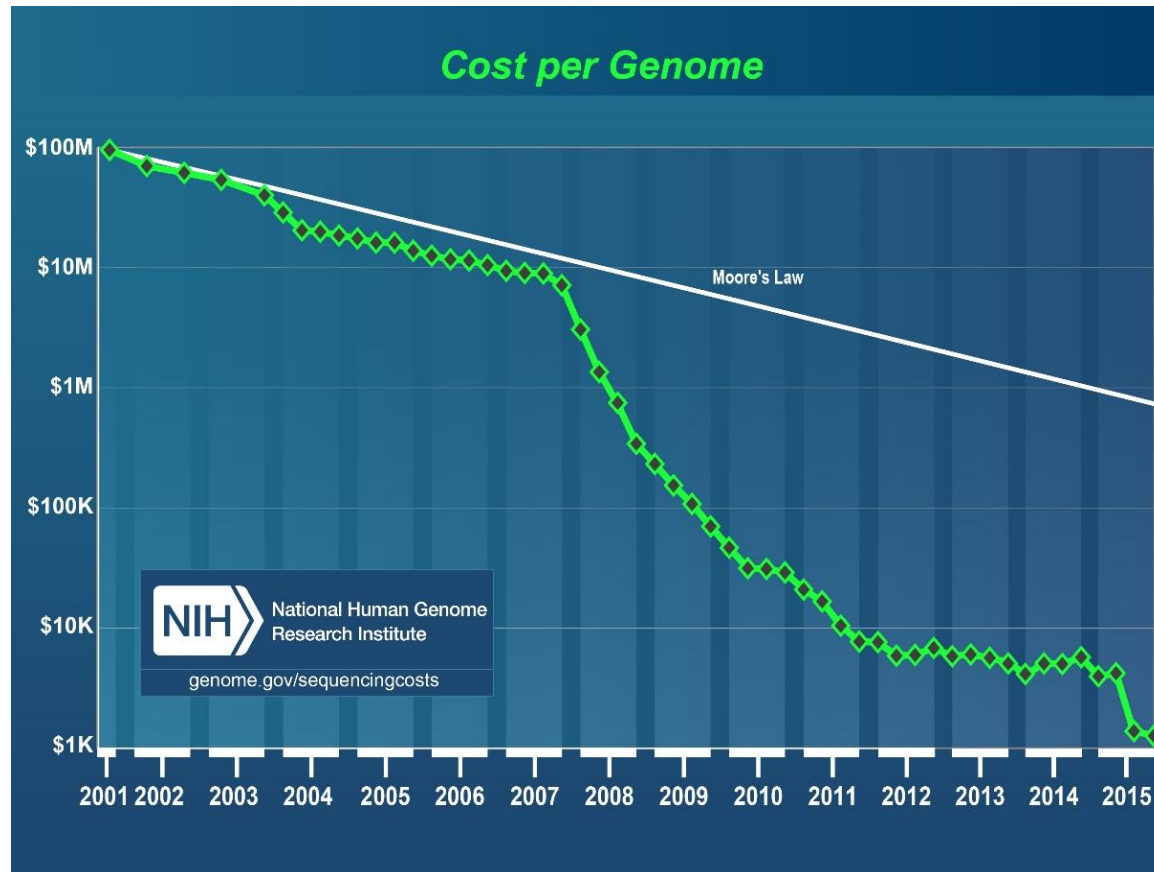
200GB Per Day
120 Terabytes in 2013
2X more than in 2011

Introduction

- What Makes Big Data Valuable?

2 billion human genomes sequenced by 2025
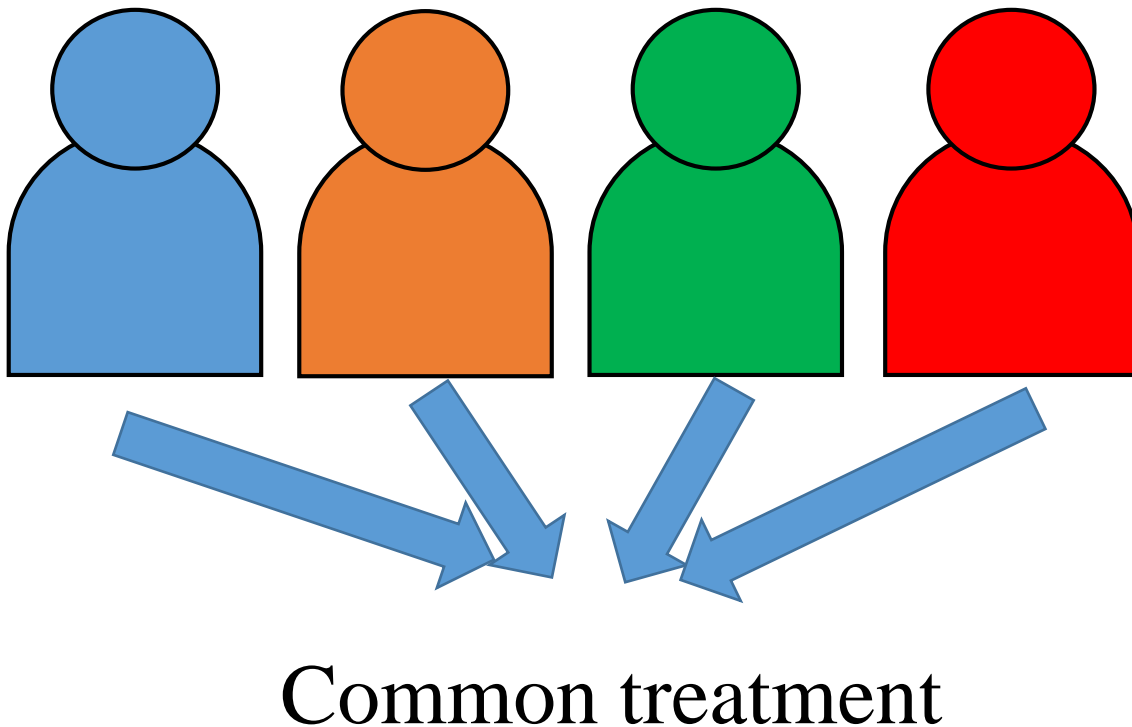
Up to 40 exabytes(EB) in storage!

Up to 10,000 trillion CPU hours for processing!

1KB=1024B
1MB=1024KB
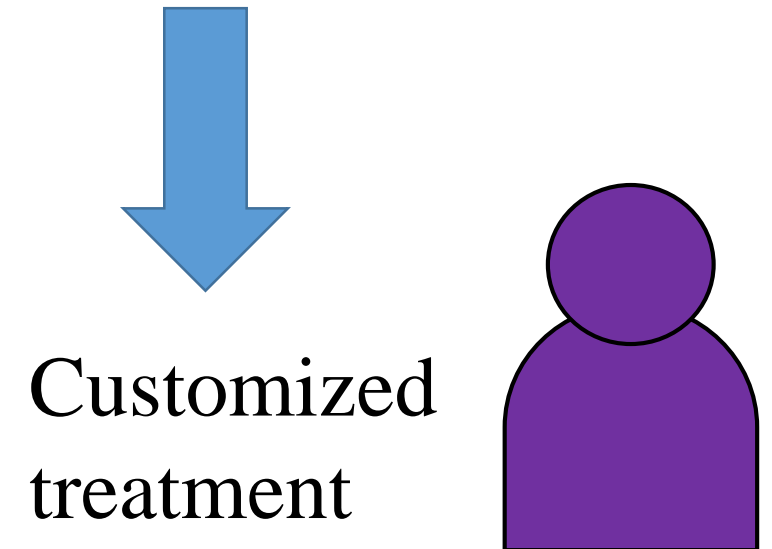1GB=1024MB
1TB=1024GB
1PB=1024TB
1EB=1024PB

YouTube: 2 exabytes per year

Introduction

- What Makes Big Data Valuable?

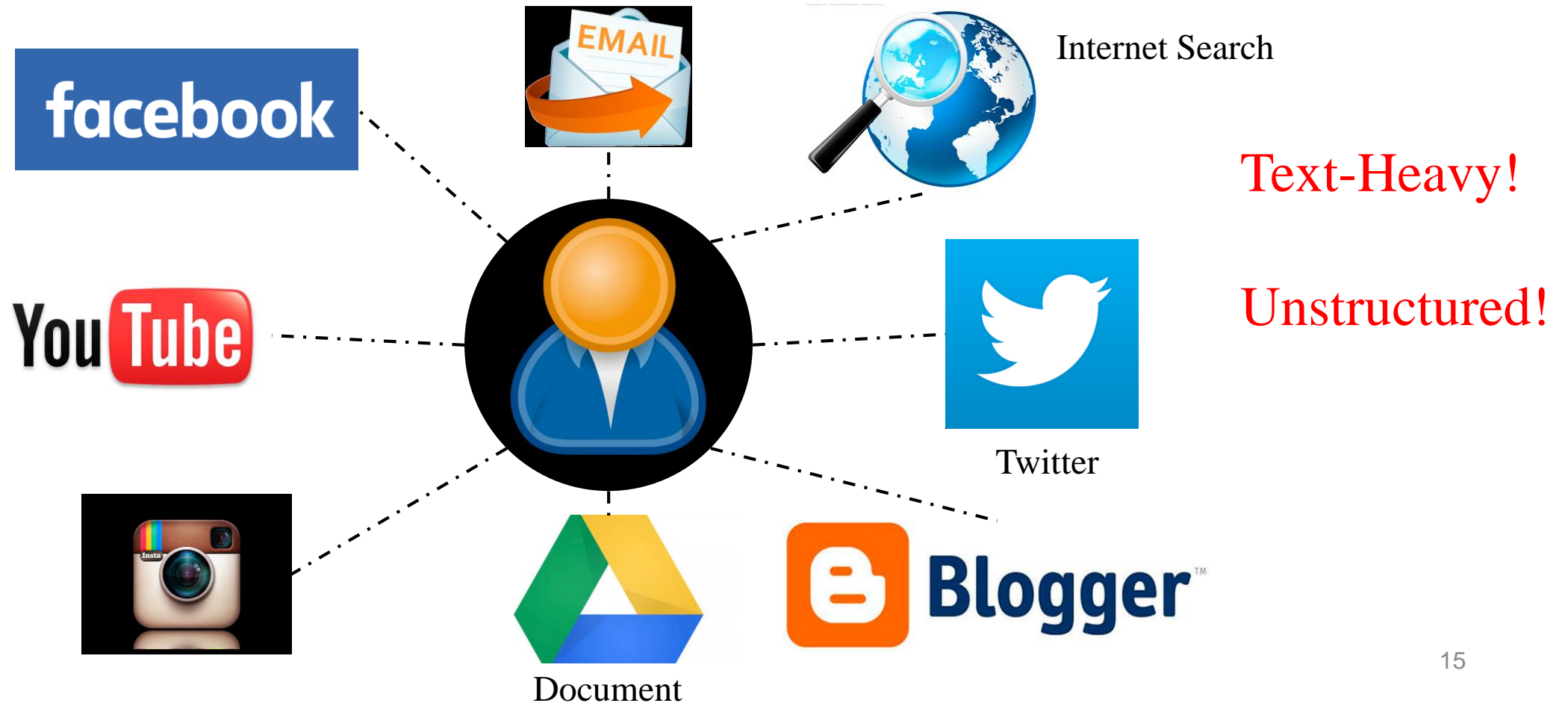Perform large scale analysis of patient genes and tumor growth

Common treatment

Customized treatment

Introduction

- Where Does Big Data Come From?

Introduction

- Where Does Big Data Come From? Generated by People

Internet Search

Text-Heavy!

Unstructured!

Twitter

Document

Introduction

- Where Does Big Data Come From? Generated by People

Daily Facebook Data
>
All US Academic Libraries

30+ PBs vs. 2 PBs

A huge growth and volume of data!

Introduction

- Where Does Big Data Come From? Generated by People

Structure

80%-90% of entire data is unstructured!

17

Introduction

- Where Does Big Data Come From? Generated by People
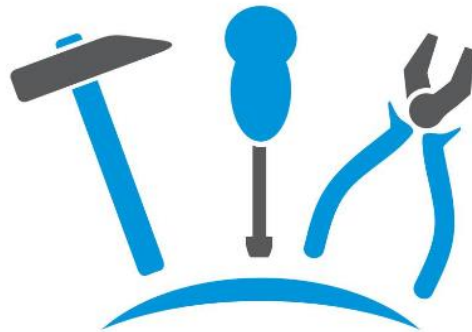
Acquisition

Storage

Retrieval

Cleaning

Processing

Introduction

- Where Does Big Data Come From? Generated by People

Tools

Data

Skilled People

?

Value

?

Introduction

- Where Does Big Data Come From? Generated by People

1. Abstract and facilitate the storage and processing of large and/or rapidly growing data sets
   - Structured and non-structured data
   - Simple programming models
2. High scalability and availability
3. Use commodity (cheap!) hardware with little redundancy
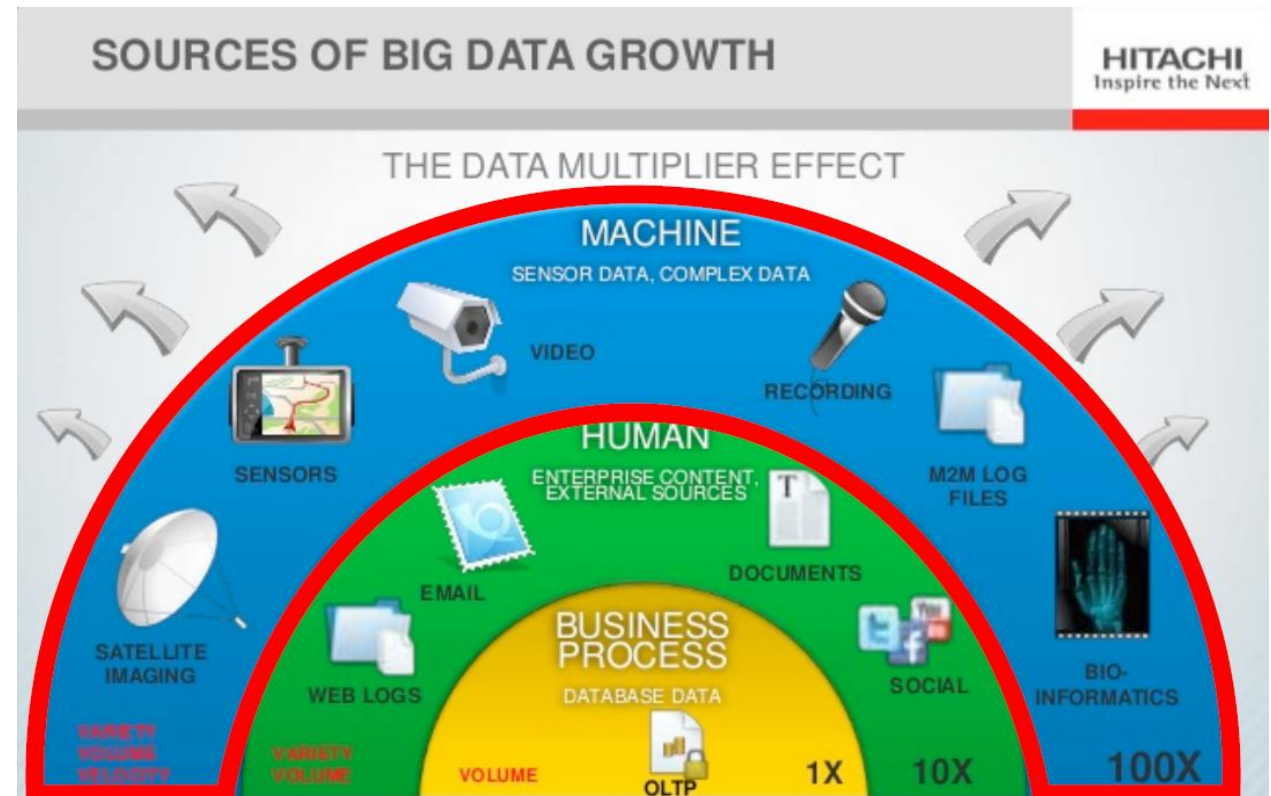4. Fault-tolerance

Introduction

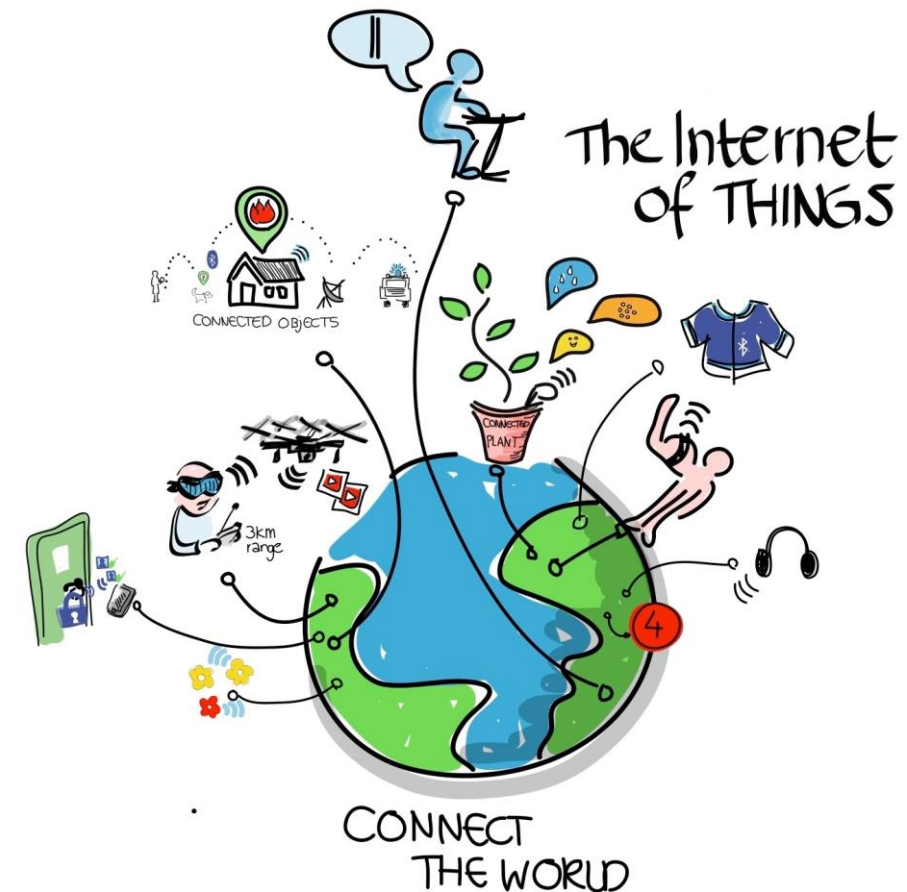- Where Does Big Data Come From? Generated by Machine

Half a terabyte of data !

Machine data is the largest source of big data!

Introduction

- Where Does Big Data Come From? Generated by Machine

Introduction

- Where Does Big Data Come From? Generated by Machine

Increasing number of machines that sense

✖

Data collected by each device

=

Machines → Biggest Source

Introduction

- Where Does Big Data Come From? Generated by Organizations:

E-Commerce

Banking/Stock Records

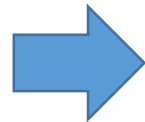Government Open Data

Medical Records

Sale transaction data

Detect correlated products

Estimate demand

Capture fraudulent activity

Introduction

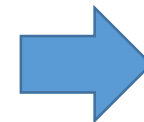- Where Does Big Data Come From? Generated by Organizations:

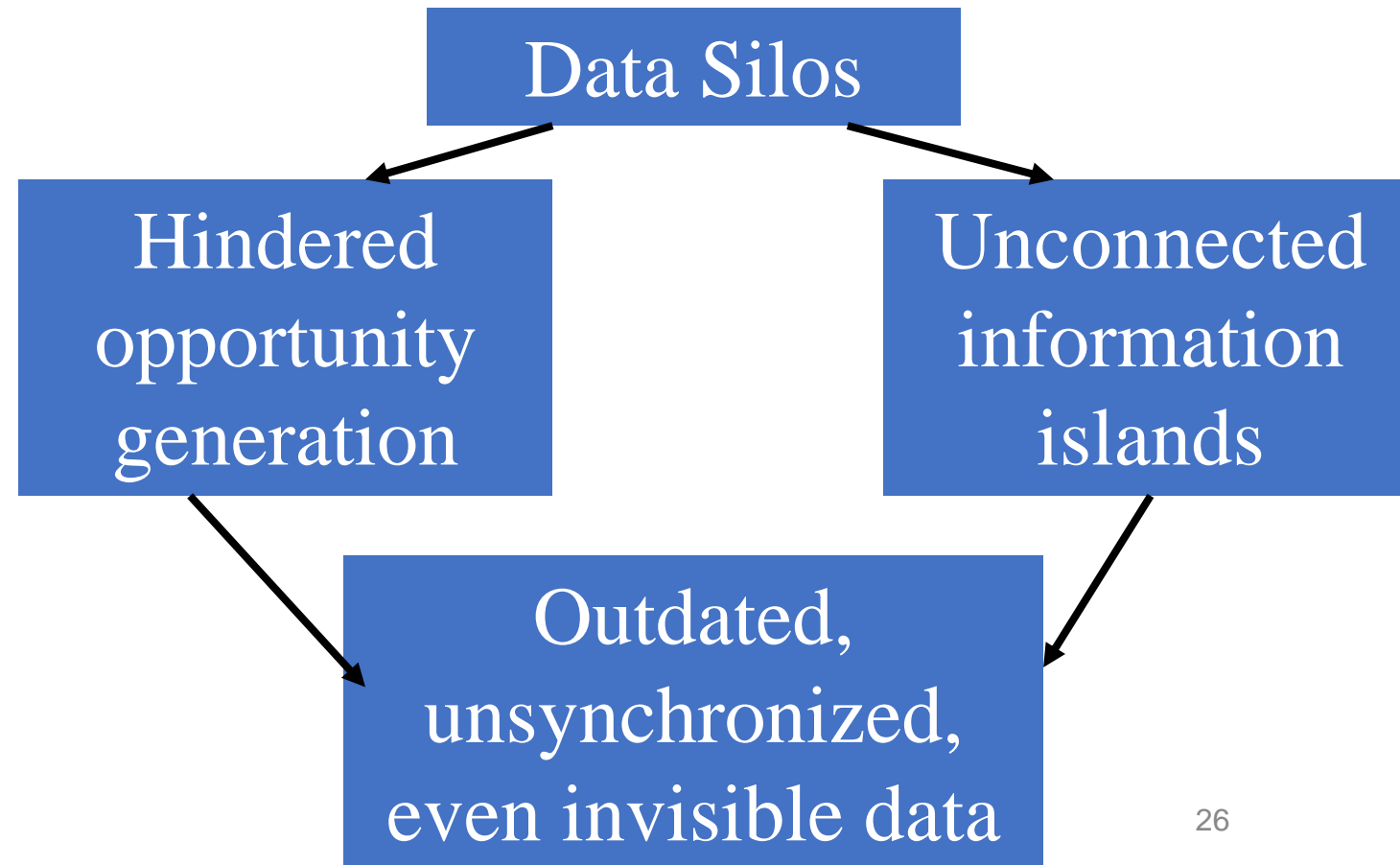| Sale ID	Time	Customer	Product ID	Quantity
S00001	12/1/2012 9:00:00 AM	C0001	P025	1
S00002	12/1/2012 9:05:58 AM	C0025	P025	3
S00003	12/1/2012 9:11:33 AM	C0010	P001	2
S00004	12/1/2012 9:17:16 AM	C0017	P023	4
S00005	12/1/2012 9:23:04 AM	C0018	P016	5
S00006	12/1/2012 9:28:43 AM	C0011	P018	4
S00007	12/1/2012 9:34:07 AM	C0015	P006	

Sale transaction data ➕ Analytics ➡ Better predictions

Introduction

- Where Does Big Data Come From? Generated by Organizations:

Data silos within an organization!

Data Silos

↓

Hindered opportunity generation

Unconnected information islands

↓

Outdated, unsynchronized, even invisible data

Introduction

- Value comes from integrating different types of data sources:
  Data Integration → Knowledge

Key: Turning complex data into something usable

Introduction

- Value comes from integrating different types of data sources:

Fire Detection

**Spatial capabilities
With Non-Spatial data**

Data Integration

Richer Data

Automatically detected location information for fire ignition point

Fire Simulation

Introduction

- Value comes from integrating different types of data sources:

Data integration

Reduce data complexity

Increase data availability

Unify your data system

Increase data collaboration

Add value to your big data!

Introduction

- Summary
  1. Describe how the combination of a growing torrent of data and on-demand (e.g. cloud) computing has launched the data science field
  2. Big data is generated by machine, people, organization
  3. Explain what unstructured data is challenge
  4. Explain what data integration is and how it makes big data sources more valuable even before they are analyzed