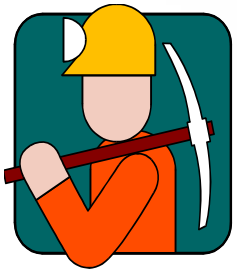
Big DATA
An Introduction

Dr. Frederick Patacsil
Pangasinan State University
Urdaneta Campus

What is big data?

- After years of data mining there is still no unique answer to this question.



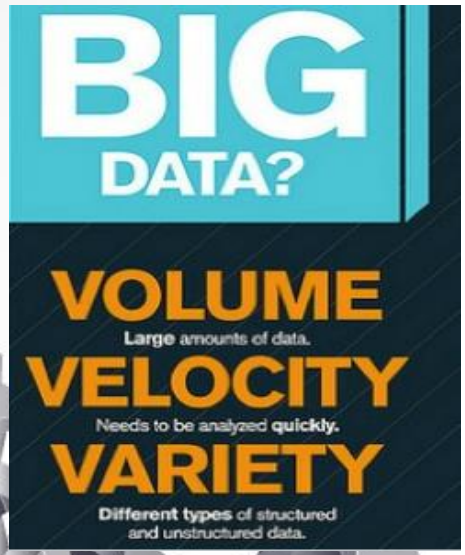
- A tentative definition:

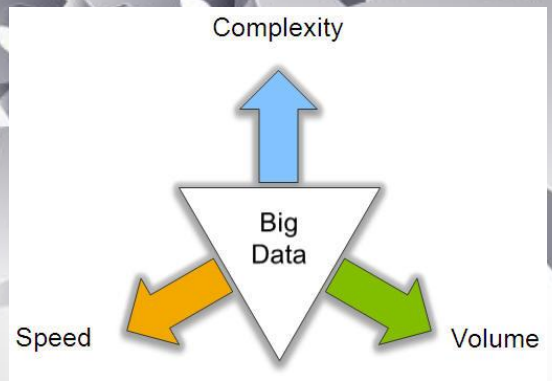
Why do we need data mining?

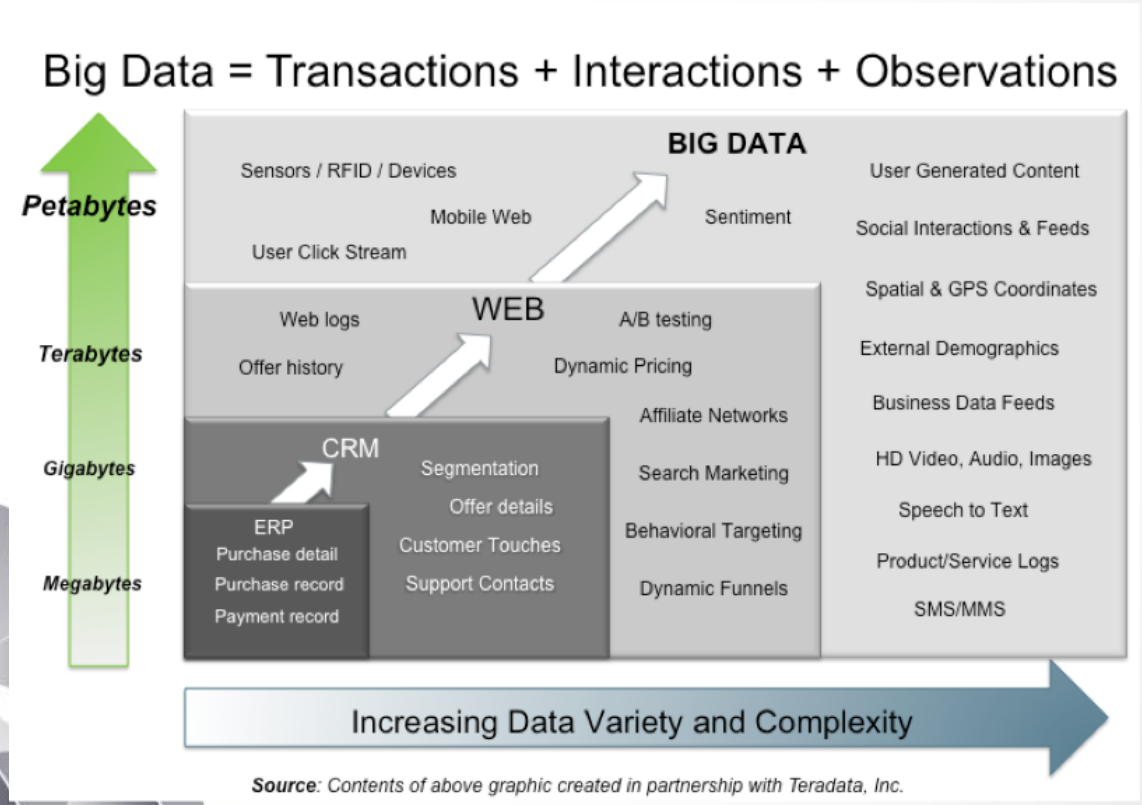
- Really, really huge amounts of raw data!!
  - In the digital age, TB of data is generated by the second
    - Mobile devices, digital photographs, web documents.
    - Facebook updates, Tweets, Blogs, User-generated content
    - Transactions, sensor data, surveillance data
    - Queries, clicks, browsing
  - Cheap storage has made possible to maintain this data

- Need to analyze the raw data to extract knowledge

Big Data Characteristics: 3V

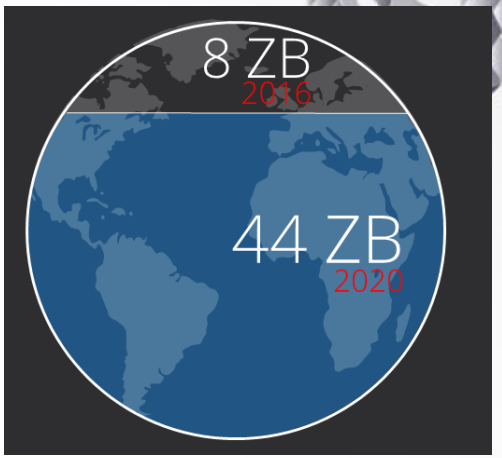


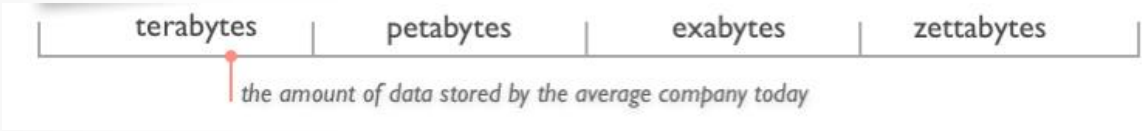


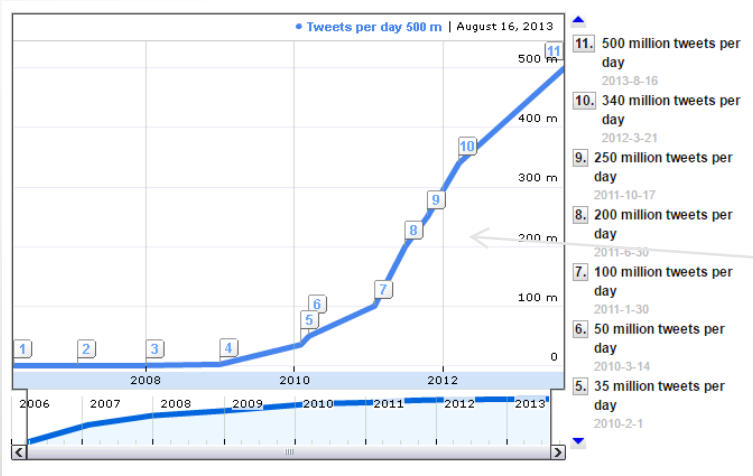


Volume (Scale)

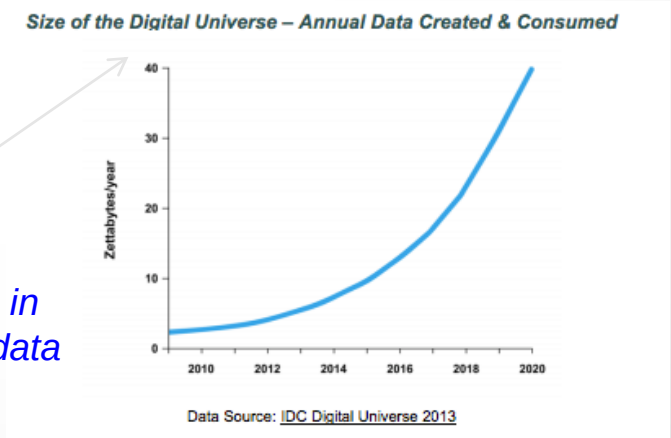
- Data Volume
  - Growth 40% per year
  - From 8 zettabytes (2016) to 44zb (2020)
- Data volume is increasing exponentially







Exponential increase in collected/generated data



Google

Processes 20 PB a day (2008)
Crawls 20B web pages a day (2012)
Search index is 100+ PB (5/2014)
Bigtable serves 2+ EB, 600M QPS (5/2014)

INTERNET ARCHIVE

400B pages, 10+ PB (2/2014)

YAHOO!

Hadoop: 365 PB, 330K nodes (6/2014)

JPMorganChase

150 PB on 50k+ servers running 15k apps (6/2011)

ebay

Hadoop: 10K nodes, 150K cores, 150 PB (4/2014)

300 PB data in Hive + 600 TB/day (4/2014)

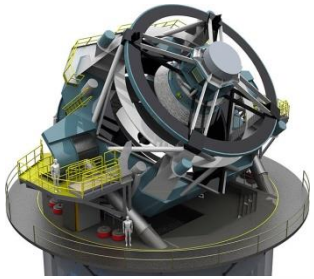
facebook

amazon web services™

S3: 2T objects, 1.1M request/second (4/2013)

LHC: ~15 PB a year

CERN



LSST: 6-10 PB a year (~2020)

SKA: 0.3 – 1.5 EB per year (~2020)

How much data?

Example: transaction data

- Billions of real-life customers:
  - WALMART: 20M transactions per day
  - AT&T 300 M calls per day
  - Credit card companies: billions of transactions per day.

- The point cards allow companies to collect information about specific users

Example: document data

- Web as a document repository: estimated 50 billions of web pages
- Wikipedia: 4 million articles (and counting)
- Online news portals: steady stream of 100’s of new articles every day
- Twitter: ~300 million tweets every day

Example: network data

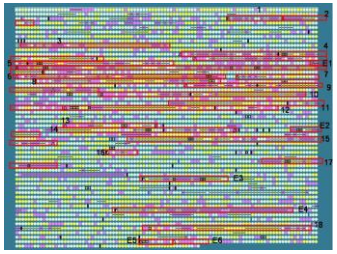
- Web: 50 billion pages linked via hyperlinks
- Facebook: 500 million users
- Twitter: 300 million users
- Instant messenger: ~1billion users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

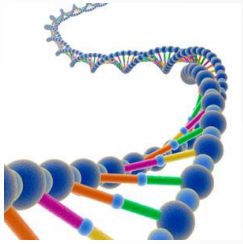
Example: genomic sequences

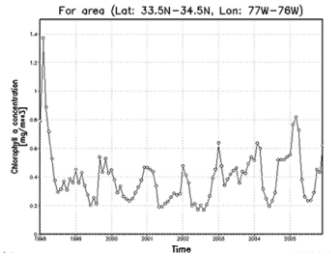
- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- $3*10^9$ nucleotides per person → $3*10^{12}$ nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

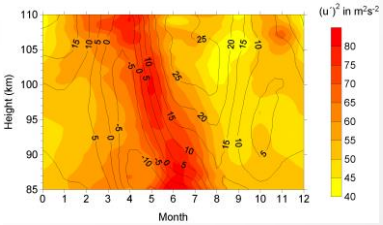
Characteristics of Big Data:
2-Complexity (Variety)

- 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



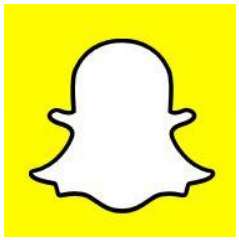






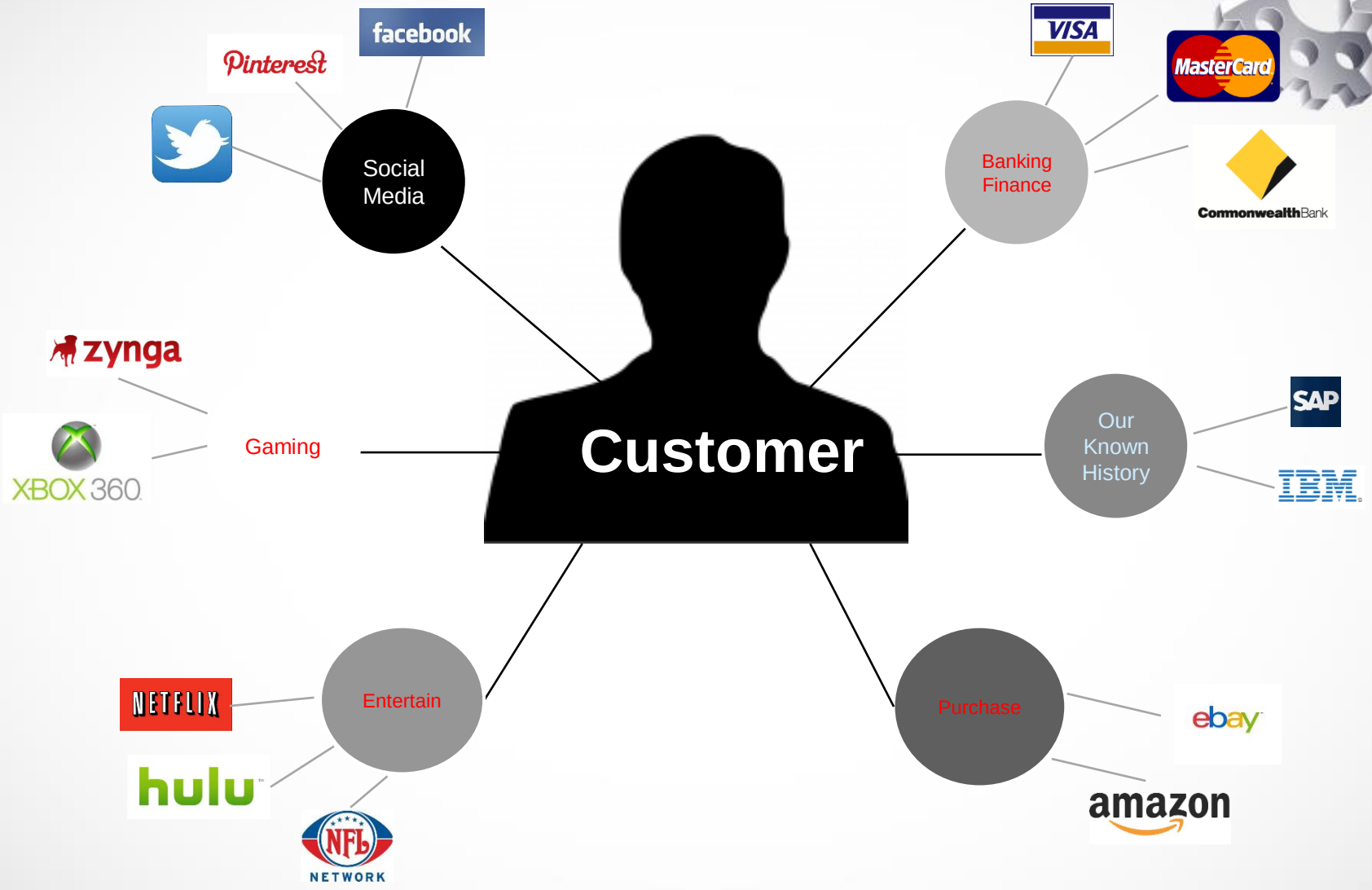
To extract knowledge➔ all these types of data need to linked together

Variety: different forms of data



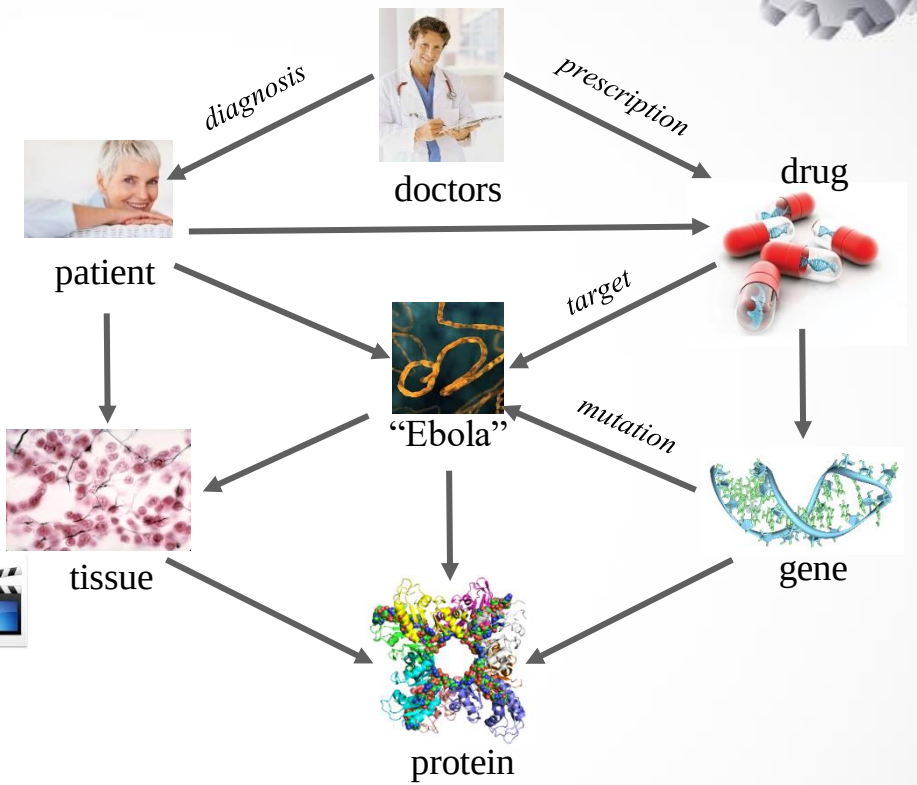


A Single View to the Customer



A Global View of Linked Big Data

Diversified social network



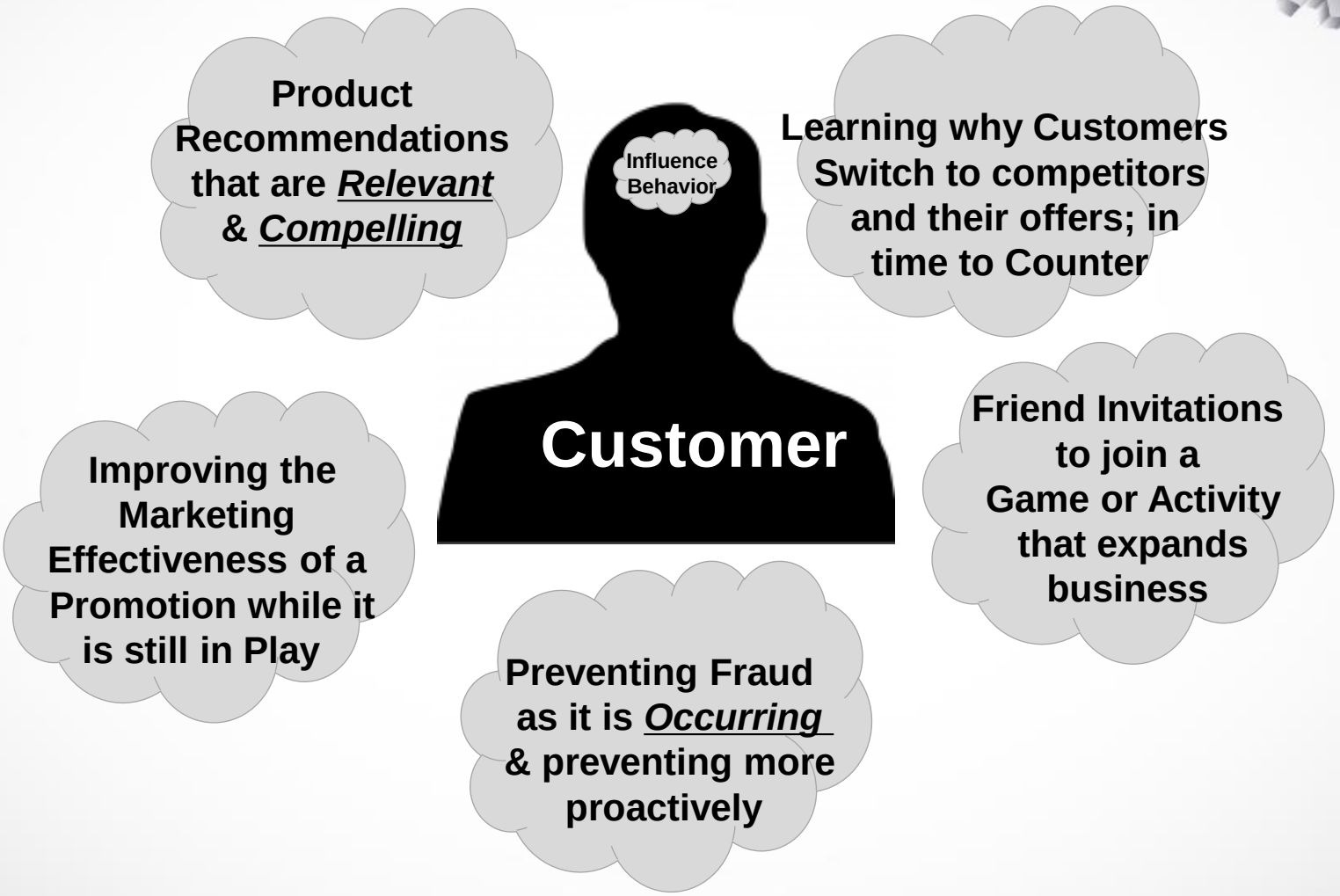
Heterogeneous information network

Characteristics of Big Data:
3-Speed (Velocity)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportuni

- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Disaster management and response**
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-Time Analytics/Decision Requirement



Extended Big Data Characteristics: 6V

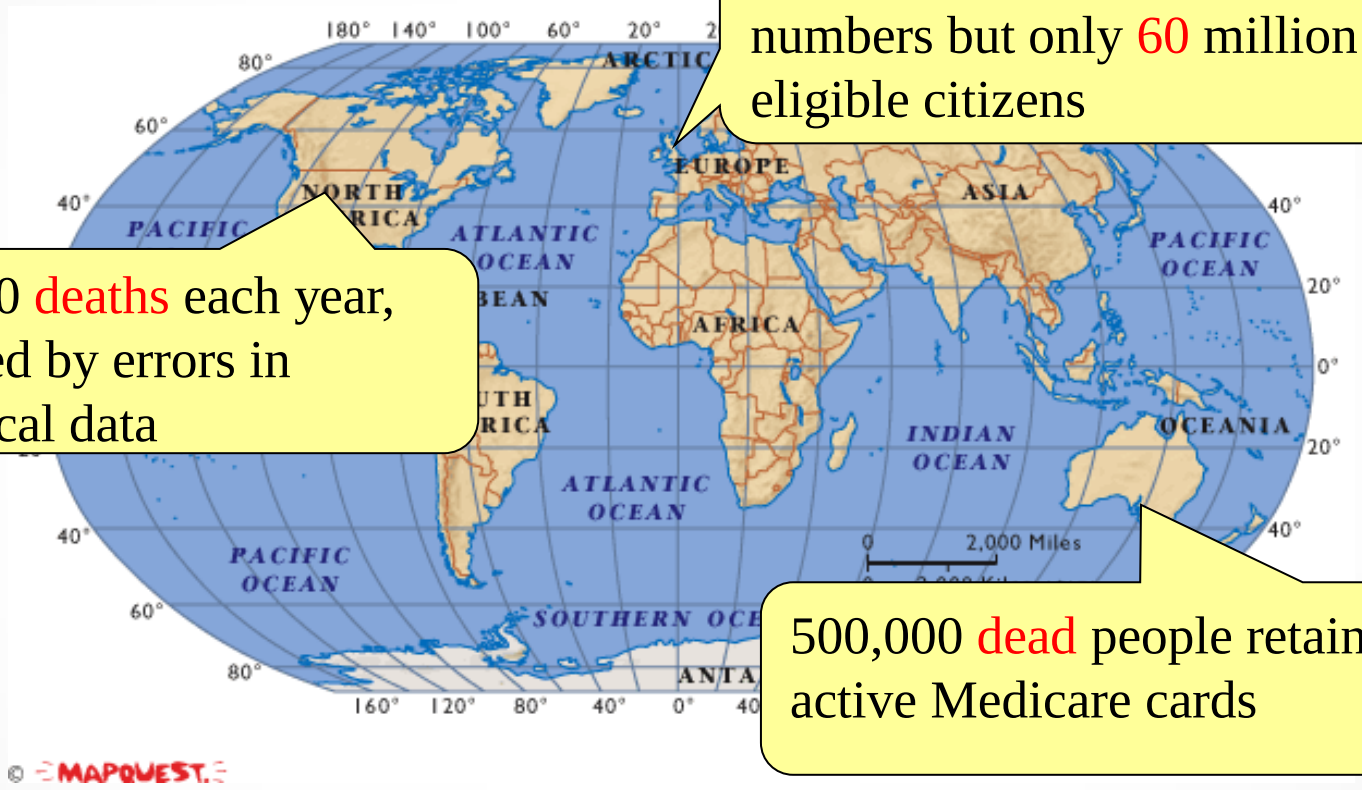
- Volume: In a big data environment, the amounts of data collected and processed are much larger than those stored in typical relational databases.
- Variety: Big data consists of a rich variety of data types.
- Velocity: Big data arrives to the organization at high speeds and from multiple sources simultaneously.

- Veracity: Data quality issues are particularly challenging in a big data context.
- Visibility/Visualization: After big data being processed, we need a way of presenting the data in a manner that's readable and accessible.
- Value: Ultimately, big data is meaningless if it does not provide value toward some meaningful goal.

Veracity (Quality & Trust)

- *Data = quantity + quality*
- When we talk about big data, we typically mean its quantity:
  - What capacity of a system provides to cope with the sheer size of the data?
  - Is a query feasible on big data within our available resources?
  - How can we make our queries tractable on big data?
  - . . .
- Can we trust the answers to our queries?
  - Dirty data routinely lead to misleading financial reports, strategic business planning decision ⇒ loss of revenue, credibility and customers, disastrous consequences
- *The study of data quality is as important as data quantity*

Data in real-life is often dirty



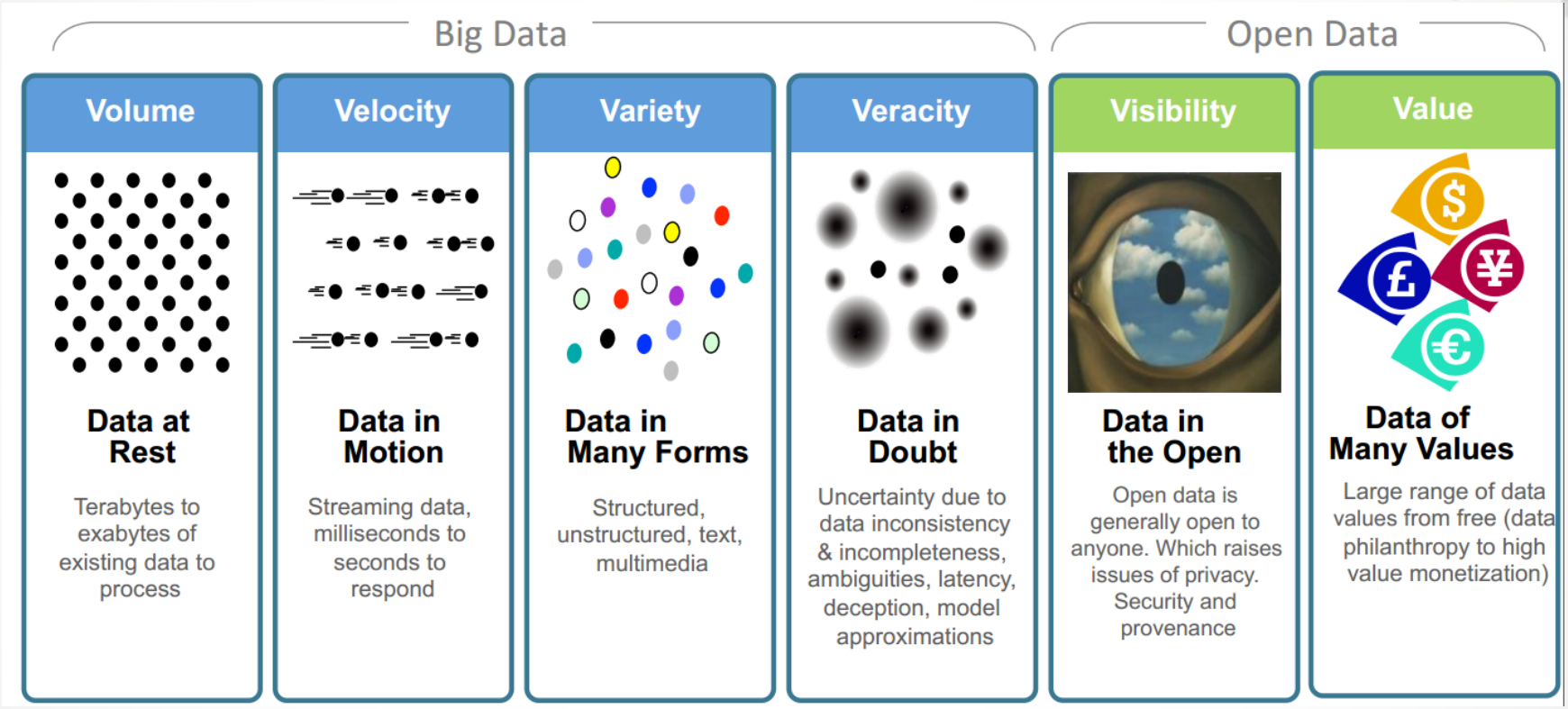
Visibility/Visualization

- Visible to the process of big data management
- Big Data – visibility = Black Hole?

A visualization of Divvy bike rides across

- Big d

Big Data: 6V in Summary

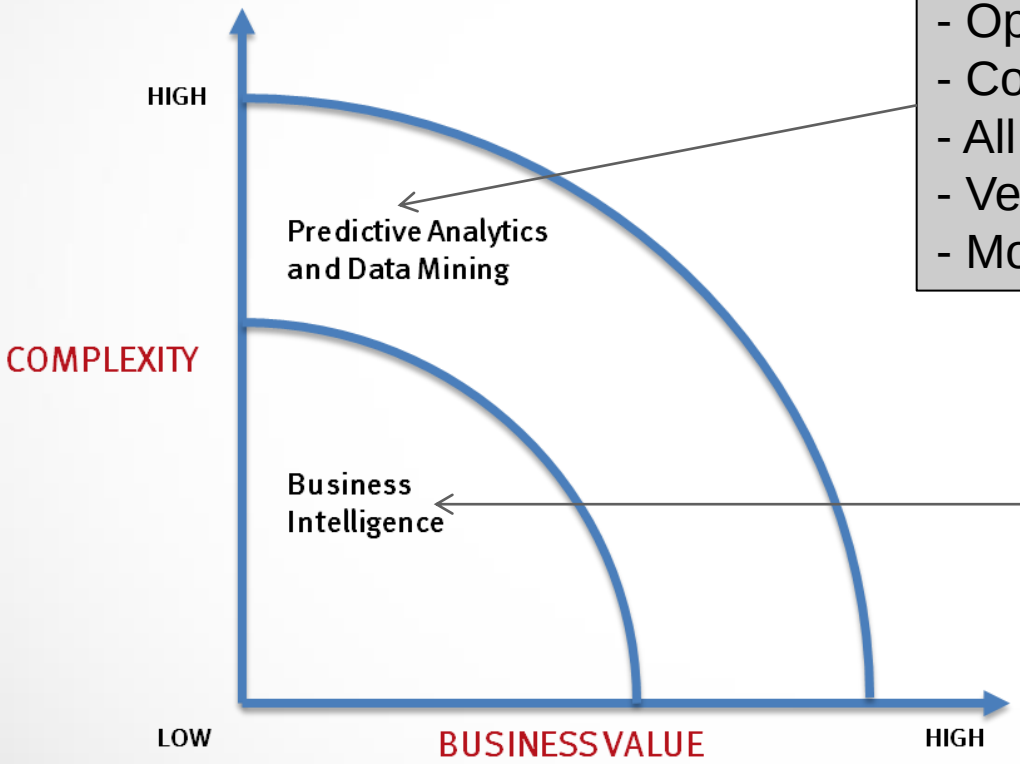


Transforming Energy and Utilities through Big Data & Analytics. By Anders Quitzau@IBM

Why Study Big Data?

- **The hottest topic in both research and industry**
- Highly demanded in real world
- A promising future career
  - Research and development of big data systems
  - Big data applications:
    social marketing, healthcare, …
  - Data analysis/data scientist: to get values out of big data
    discovering and applying patterns, predicative analysis, business intelligence, privacy and security, …

What's driving Big Data

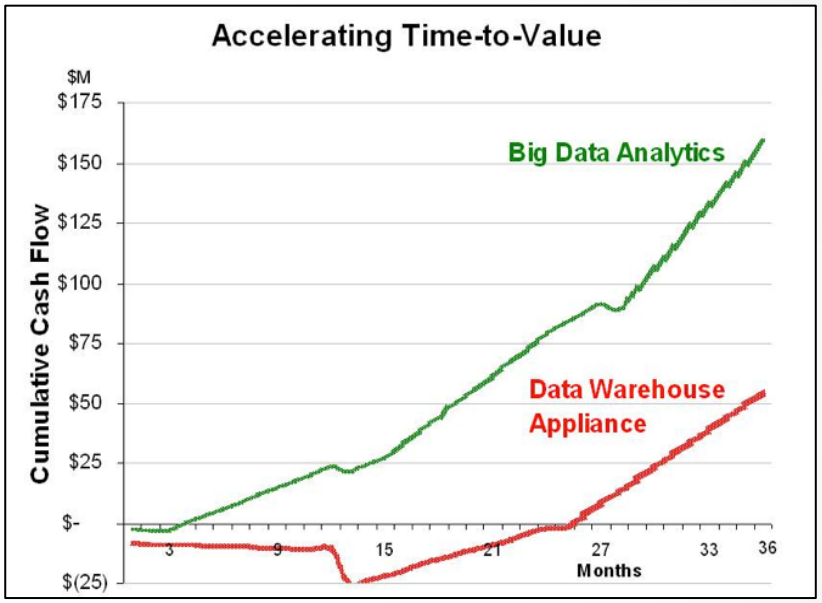


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

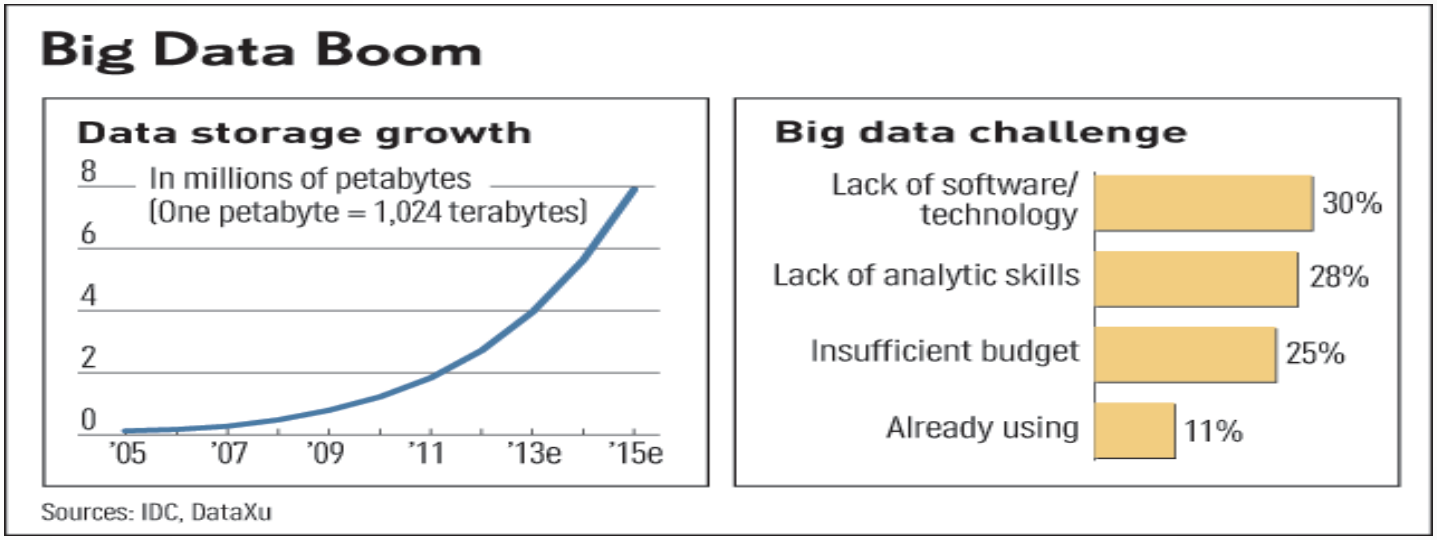
- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

Value of Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - *Experts in using the new technology and dealing with big data*

“Big data is *high-volume*, *high-velocity* and *high-variety information* assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” -- Gartner


How can I analyze my data?

What is DM

- Extraction of useful information from data: discovering relationships that have not previously been known

What is DM

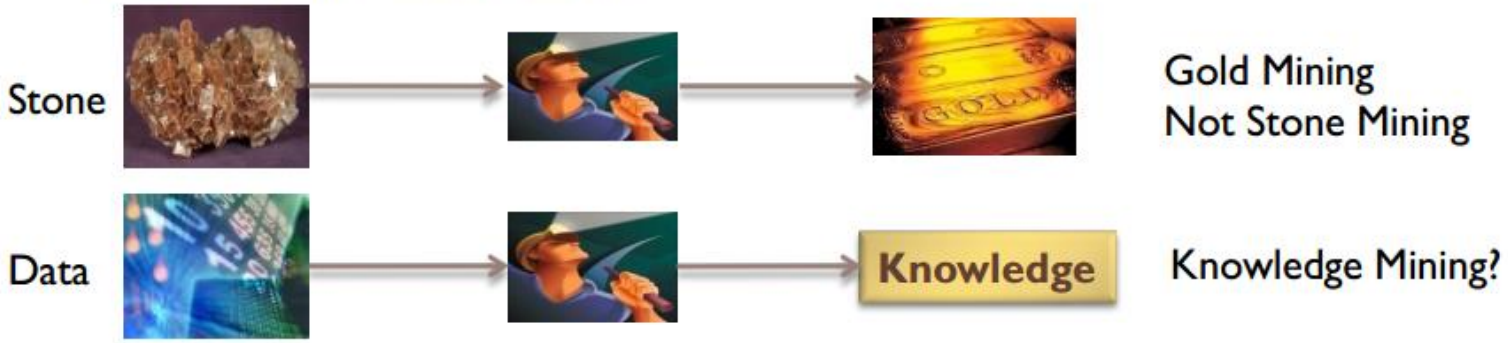
Data mining (knowledge discovery from data) Extraction of interesting patterns or knowledge from huge amount of data

What is DM

Data Mining is the application of Machine Learning techniques to solve real-life data analysis problems

What is Data Mining?

- Data mining (knowledge discovery from data)
  - → Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - → Data mining: a misnomer?

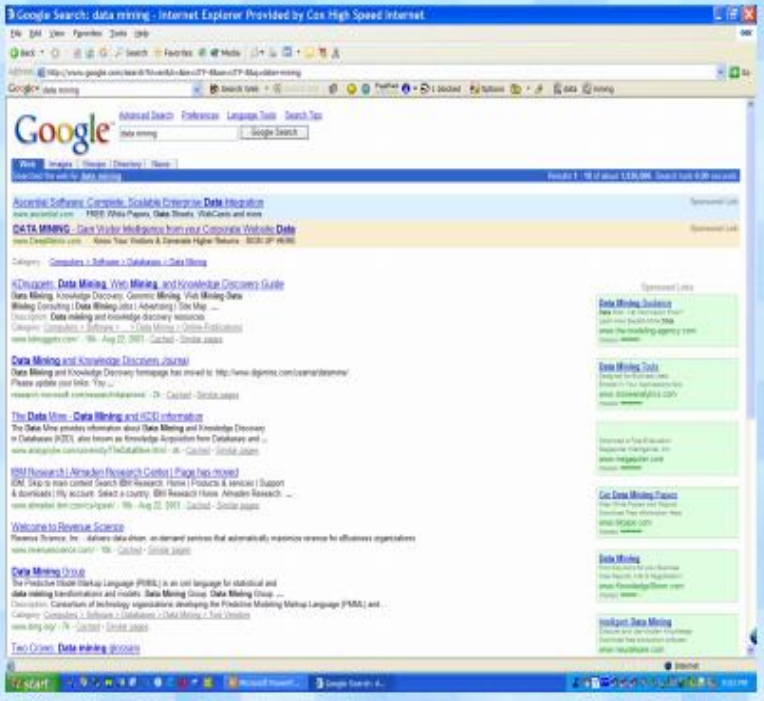


- Alternative names
  - → Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

Data Mining is not …

- Searching for a phone number in a phone book

- Searching for keywords on Google



Data Mining is not …

- Generating a histogram of salaries for different age groups

- Issuing SQL query to a database, and reading the reply

Data Mining is …

- Finding groups of people with similar hobbies

- Are chances of getting cancer higher if you live near a power line?

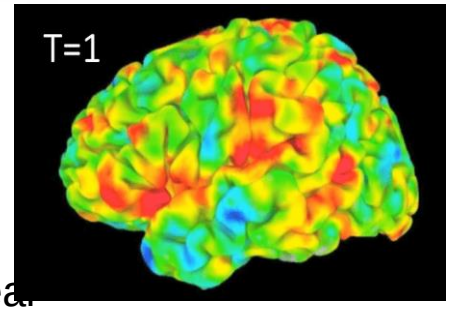
Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

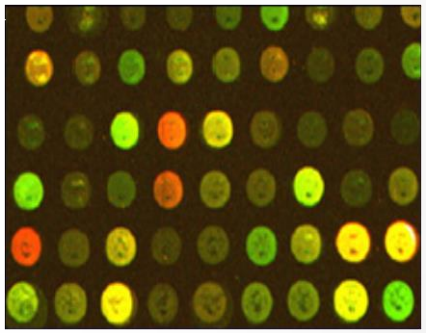
Why Data Mining? Scientific Viewpoint

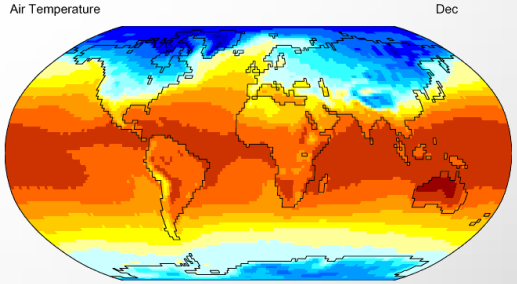
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation



fMRI Data from Brain


Sky Survey Data


Gene Expression Data

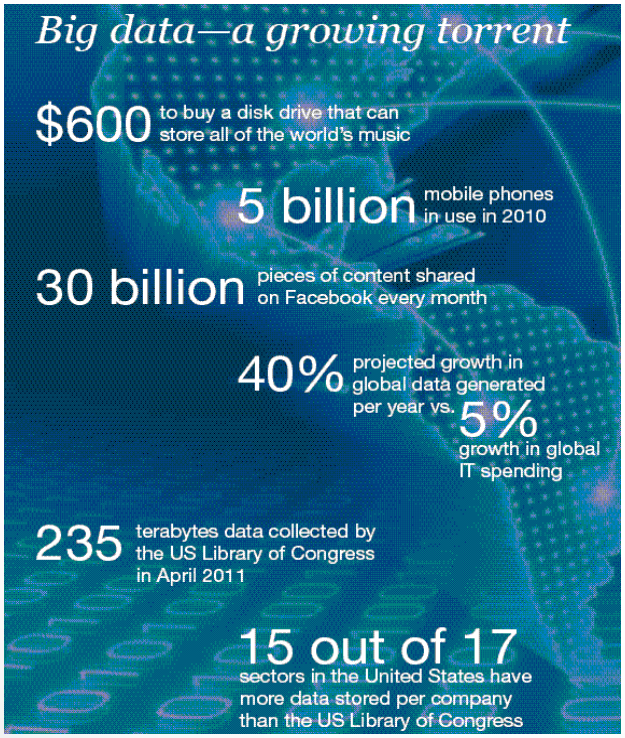


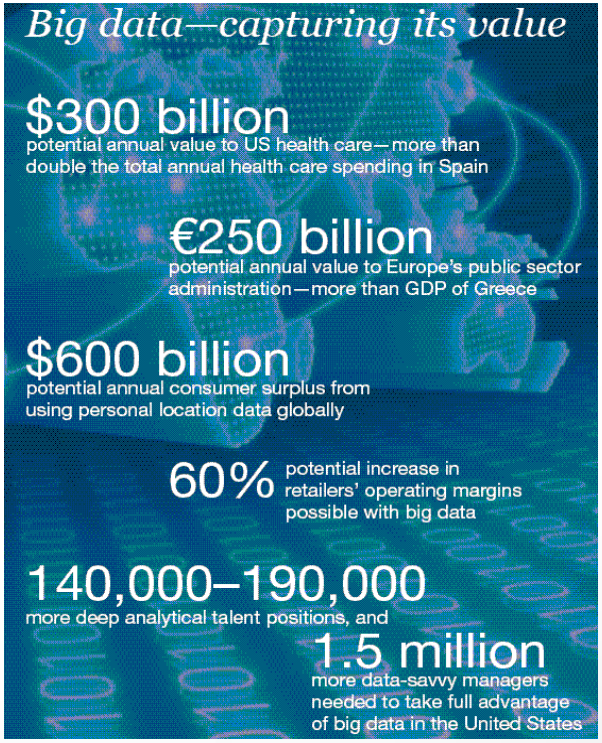
Surface Temperature of Earth

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

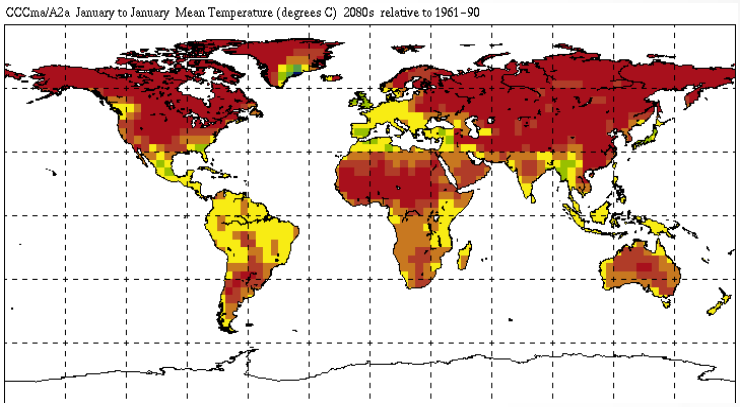
Big data: The next frontier for innovation, competition, and productivity





Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



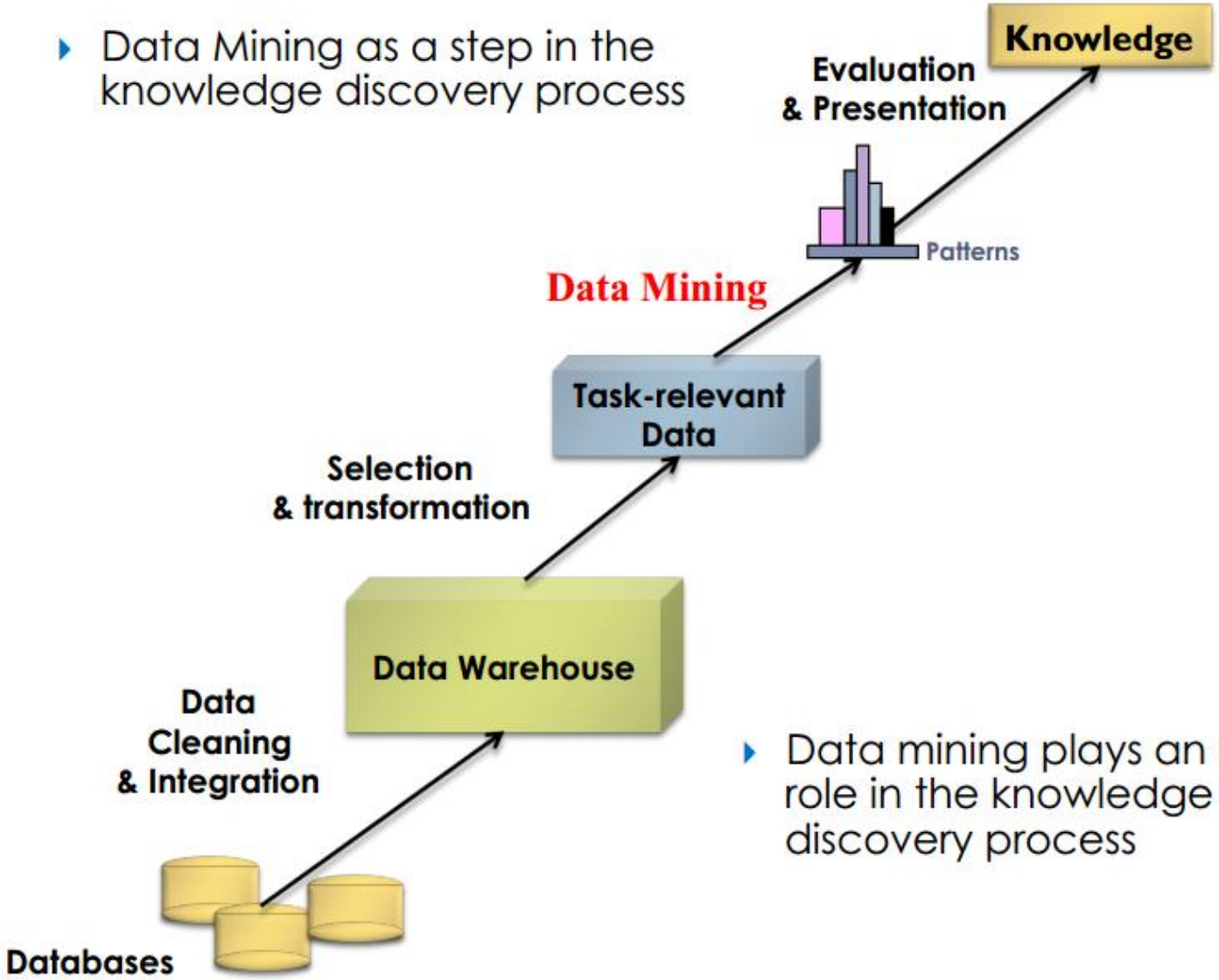
Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

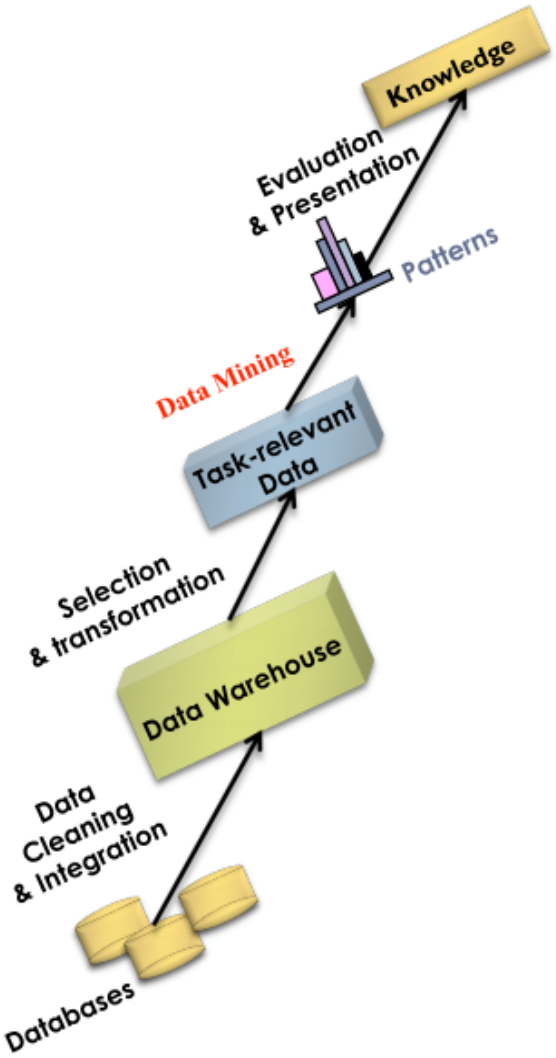
Knowledge Discovery (KDD) Process

- Data Mining as a step in the knowledge discovery process



- Data mining plays an essential role in the knowledge discovery process

Knowledge Discovery (KDD) Process



- **Data Cleaning**
  - → Remove noise and inconsistent data
- **Data Integration**
  - → Combine multiple data sources
- **Data Selection**
  - → Data relevant to analysis tasks are retrieved form the data
- **Data transformation**
  - → Transform data into appropriate form for mining (summary, aggregation, etc.)
- **Data mining**
  - → Extract data patterns
- **Pattern Evaluation**
  - → Identify truly interesting patterns
- **Knowledge representation**
  - → Use visualization and knowledge representation tools to present the mined data to the user