Social Big Data

Dr. Hong Huang

Outline

- Big data
  - What is big data?
  - Social big data
  - Data processing
- Data mining
  - What is data mining?
  - Why data mining?
  - How data mining?
  - Use cases

BIG DATA ?

Big data, big pay, big opportunities

- Average salaries for Data Scientists (08.2016)
  - Facebook - **$133,492**
  - Microsoft - **$120,304**
  - AirBnB - **$118,483**
  - Twitter - **$134,861**
  - Apple - **$147,515**
  - LinkedIn - **$133,995**
- More than 9,676 job openings posted in LinkedIn in USA (searched yesterday)

Big numbers

12+ TBs **of tweet data every day**

? TBs of data every day

25+ TBs **of log data every day**

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

76 million smart meters in 2009…
200M by 2014

Big size

- Data processed per day (Updated on Jun.24, 2014)

Organization	Est.amount of data processed per day
eBay	100pb
Google	100pb
Baidu	10-100pb
NSA	29pb
Facebook	600TB
Twitter	100TB
Spotify	2.2TB(compressed; becomes 64 Tb in Hadoop)
Sanger Institute	1.7 Tb (DNA sequencing data only)

https://followthedata.wordpress.com/2014/06/24/data-size-estimates/

Big size

- Data Stored (Updated on Jun.24, 2014)

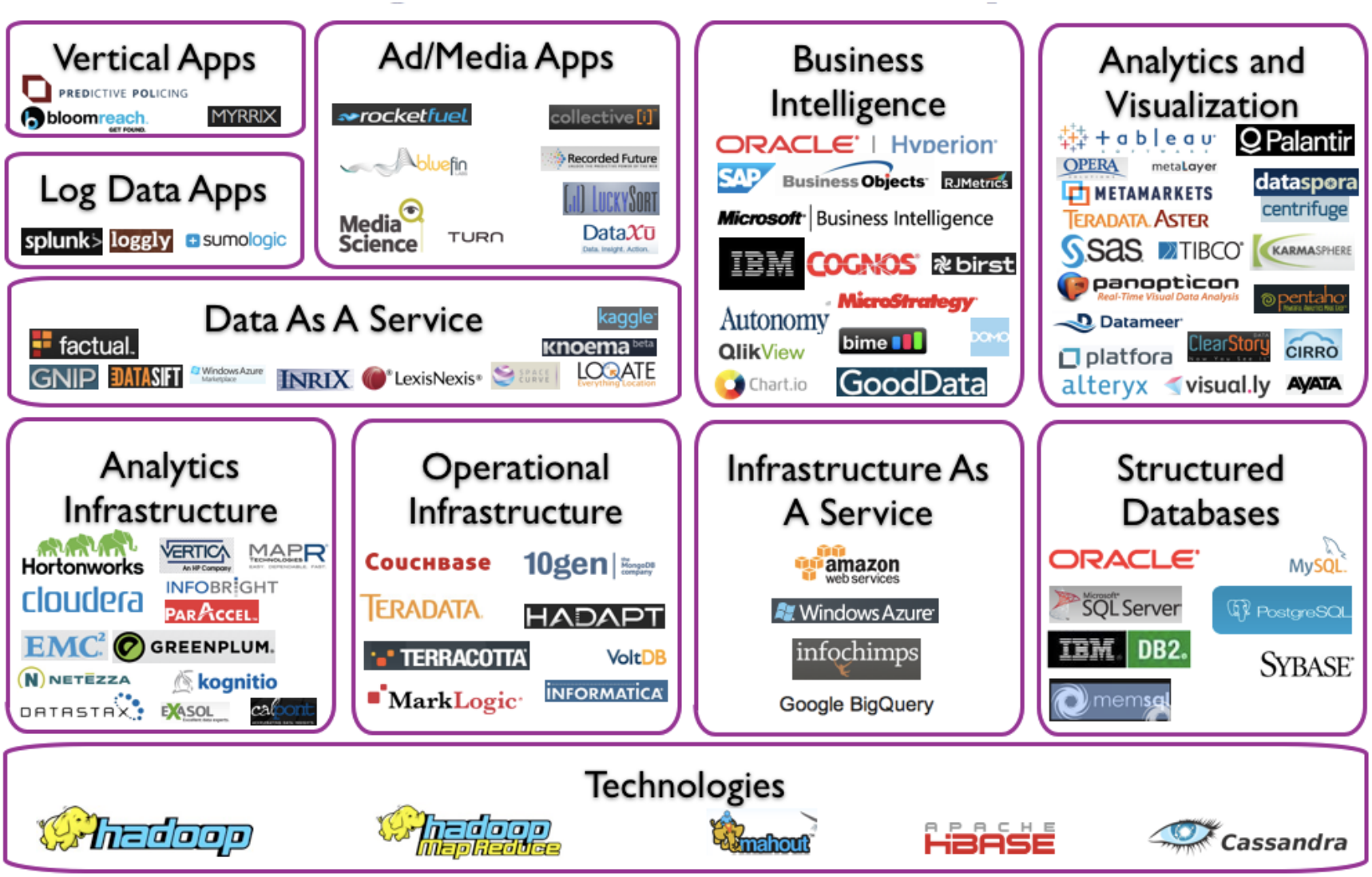
Organization	Est.amount of data stored
Google	5,000 pb
NSA	10,000 pb (possibly overestimated)
Baidu	2,000 pb
Facebook	300 pb
Ebay	90 pb
Sanger	22 pb (for DNA sequencing data only; ~45 pb for everything per Ewan Birney May 2014)
Spotify	10 pb

https://followthedata.wordpress.com/2014/06/24/data-size-estimates/

Big machines

- Google data centers

Big techniques

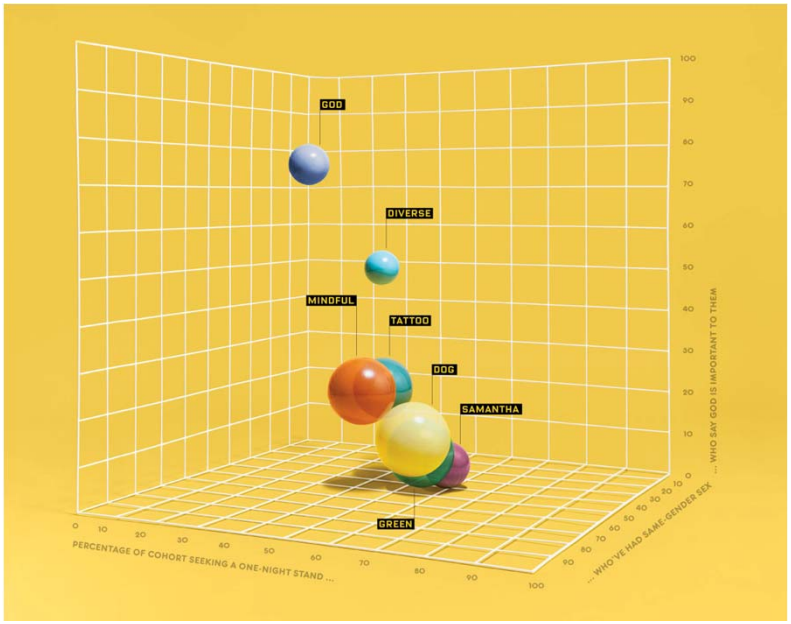


dave@vcdave.com

blogs.forbes.com/davefeinleib

Big chances

- How a Math Genius Hacked OkCupid to Find True Love



https://www.wired.com/2014/01/how-to-hack-okcupid/

WHAT IS BIG DATA?

The definition

- First in use since 1990s
- 3Vs model defined by Gartner in 2001 and updated in 2012

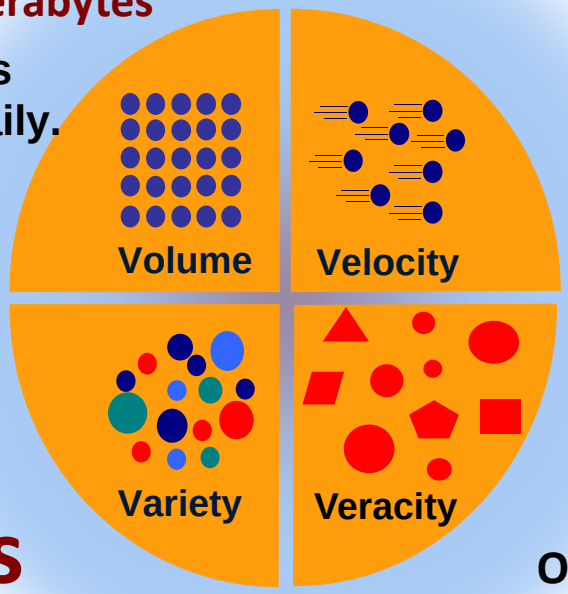
Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.

- **Volume**: big data doesn't sample; it just observes and tracks what happens
- **Velocity**: big data is often available in real-time
- **Variety**: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
- **Veracity**: the quality of captured data can vary greatly, affecting accurate analysis

With Big Data, We've Moved into a New Era of Analytics

12+ terabytes
of Tweets
create daily.

5+ million
trade events
per second.

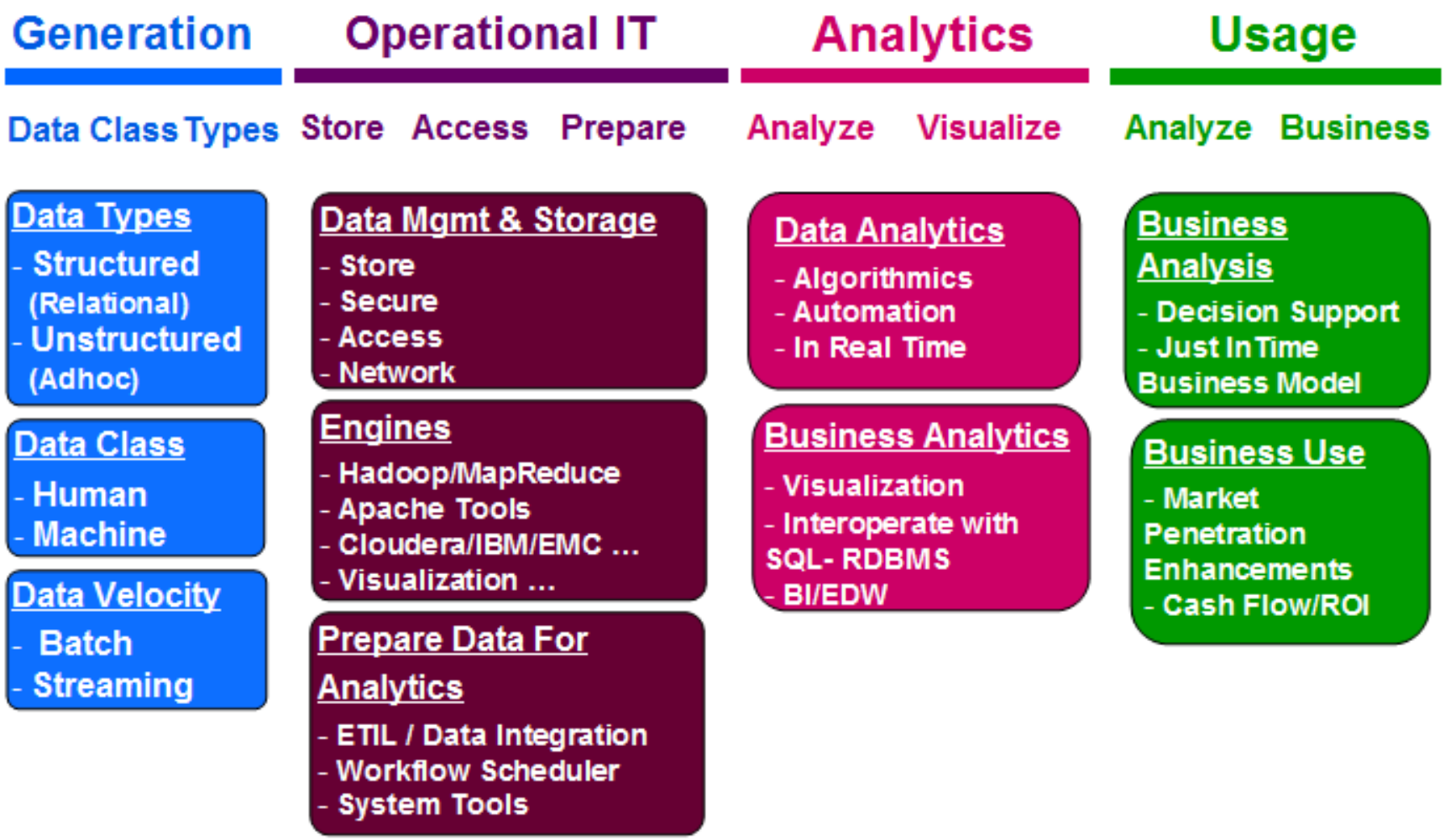


100's
of different
types of data.

Only **1 in 3**
decision makers trust
their information.

Big Data Ecosystem

IMEX



http://www.imexresearch.com/newsletters/BDEconomics.html

Big Data Analytics

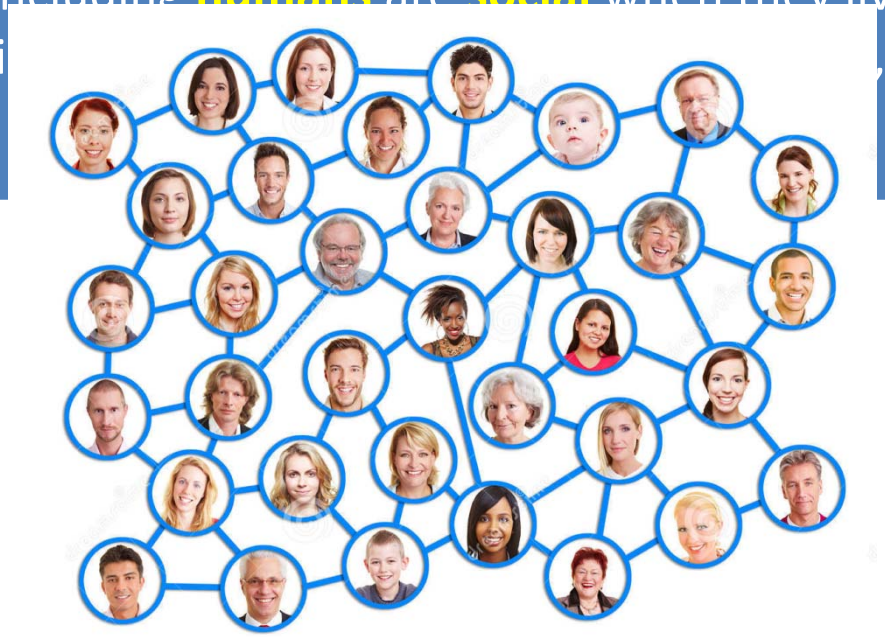
- Why is big data analytics important?

- Helps organizations harness data and use it to identify new opportunities
- Leads to smarter business moves, more efficient operations, high profits and happier customers

SOCIAL BIG DATA

Social

Living organisms including **humans** are **social** when they live collectively in interacting populati , and whether the



Social network

- A **social network** is a social structure made up of a set of social actors (such as individuals or organizations), sets of dyadic ties, and other social interactions between actors.
  - Nodes: users
  - Links: relationships

Social + Big data = ?

- Examples:
  - Offline: classmates network…
  - Online: Facebook, Twitter...

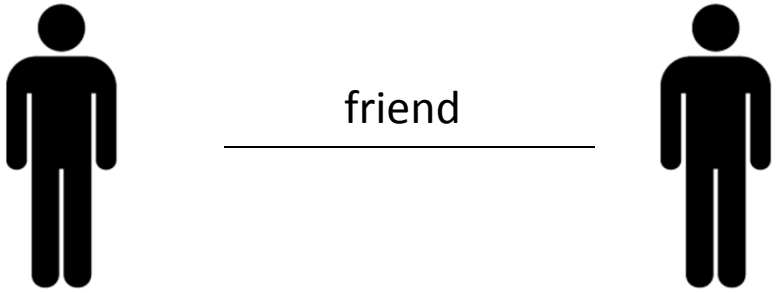
Online social networks

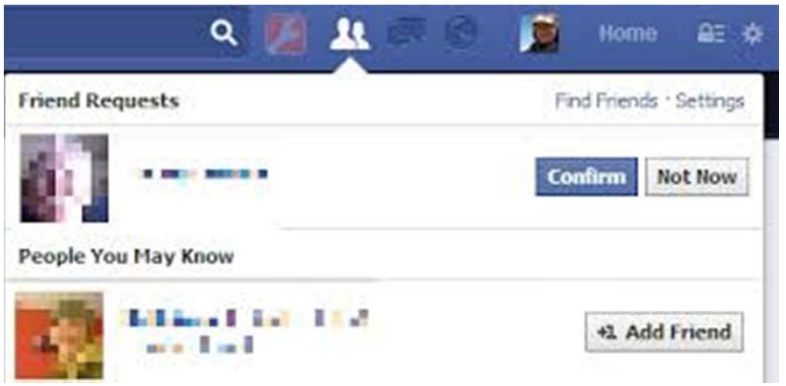
- Role of OSN
  - OSNs have reached 82% of the world's Internet-using population (1.2billion) (2011)
  - Social Networking accounts for 19%of all time spent online (2011)
  - Social Networking is the most popular online activity worldwide

- Variety of OSN
  - Various functions, social relationships, social networks

Facebook

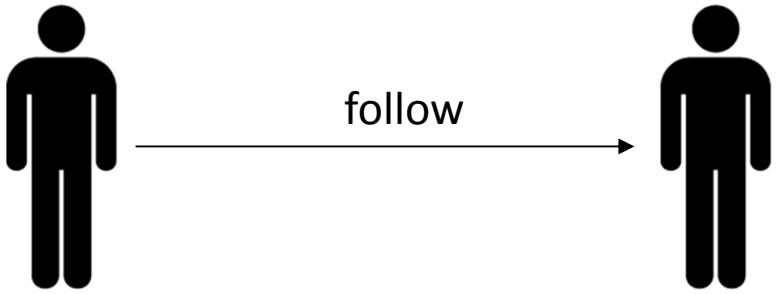
- Relationship
  - User A sends a friend request to user B and user B confirms the request from user A, then they are friends
  - Undirected social network





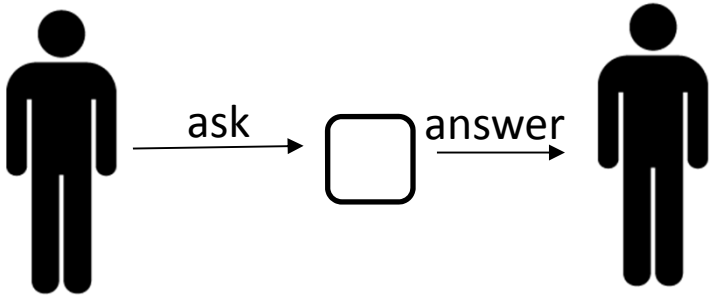
Twitter

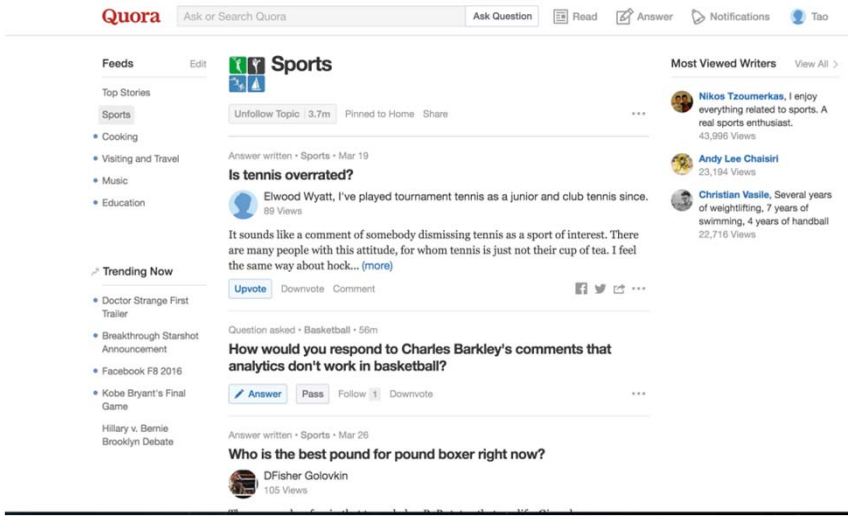
- Relationship
  - User A follows user B
    - User A is user B's fan
    - User B is user A's friend
  - Directed social network



Quora

- Relationship
  - Question-answering
    - User A asks a question and user B answers the question
  - Directed social network





DATA PROCESSING

- Suppose we have a record of IP logs which visits website A, how to find the most frequent visited IP? (Top -1? Top -10?)
- Data samples:
  - 172.23.0.25 172.23.0.26 172.23.0.27 172.23.0.
    172.23.0.29 172.23.0.256 172.23.0.25
    172.23.0.25 ……

Data Preprocessing

- Data in the real world is dirty
  - incomplete: missing attribute values, lack of certain attributes of interest, or containing only aggregate data
    - e.g., occupation=“”
  - noisy: containing errors or outliers
    - e.g., Salary=“-10”
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age=“42” Birthday=“03/07/1997”
    - e.g., Was rating “1,2,3”, now rating “A, B, C”
    - e.g., discrepancy between duplicate records

Why Is Data Preprocessing Important?

- No quality data, no quality analyzing results!
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data preparation, cleaning, and transformation comprises the majority of the work in a data analyzing application (60%).

So, please cleaning the data first!!!

- Fill in missing values,
- smooth noisy data,
- identify or remove outliers and noisy data,
- resolve inconsistencies.

- After cleaning, the data becomes:

172.23.0.25 172.23.0.26 172.23.0.27 172.23.0.28
172.23.0.29…

Now, it is ready for processing!

- Go back to our problem:

Suppose we have a record of IP logs which visits website A, how to find the most frequent visited IP? (Top -1? Top -10?)

- Solution:
- 1) <key, value> pair

- 2) Sorting <Key, value> pairs based on value

Done?

IP, 2^32 keys ➡ Out of memory

Mapping

- %1000 ➔ write into 1000 files
- Find out the most frequent one in each small file
- Sort 1000 IPs from small files

What if the original data files stored in 100 PCs, then how to find the top-10 …IPs?

- Suppose that each IP only appears in one PC
- 1) Stack sorting: top 10 from each PCs using stack sorting
- 2) Sort all combined from 100 PCs (1000)

What if each IP would be stored in different PCs, then how to find the top-10 …IPs?

- For example:
- PC1: 172.23.0.25(50) 172.23.0.26(50) 172.23.0.27(20) 172.23.0.28(0) 172.23.0.29(0)…
- PC2: 172.23.0.25(2) 172.23.0.26(10) 172.23.0.27(0) 172.23.0.28(50) 172.23.0.29(10)…

big data solutions!

Go through each logs, and rewrite the data into 100 PCs in order to make each IP only be stored in one PC, then repeated…

Outline

- Big data
  - What is big data?
  - Social big data
  - Data processing
- Data mining
  - What is data mining?
  - Why data mining?
  - How data mining?
  - Use cases

Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
  - For network science
    - 1736 Mathematical foundation – Graph Theory
    - 1930 Social Network Analysis and Theories
      - Sociogram: Network visualization
      - Six degree of separation
      - Structural hole: Source of innovation

Evolution of Sciences

- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
  - For network science
    - 1990 (Physicists) Complex Network Topologies
      - Small-world model
      - Scale-free model

Evolution of Sciences

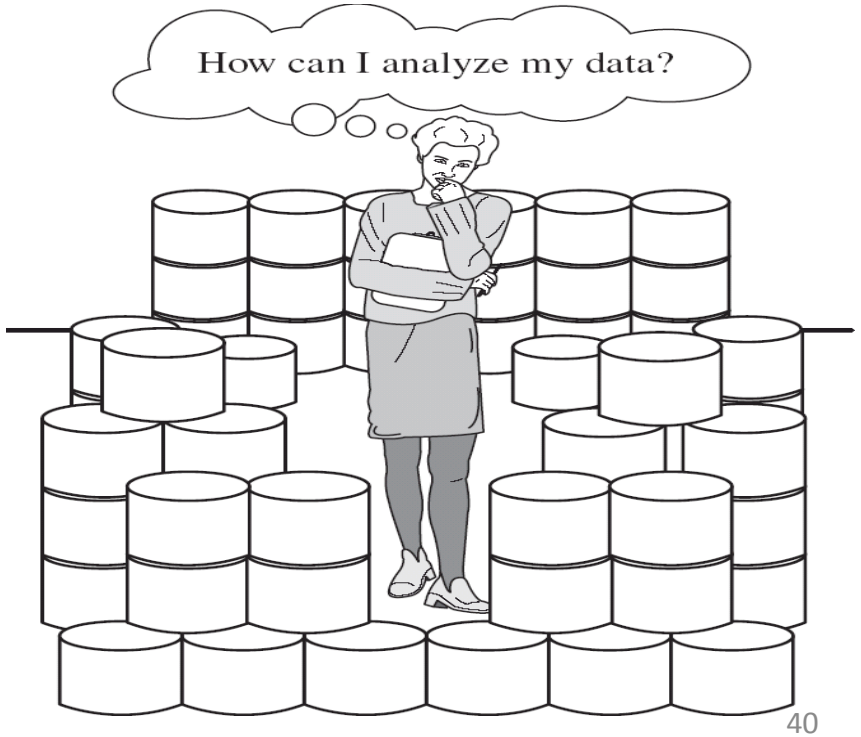
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!

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
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras,
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention

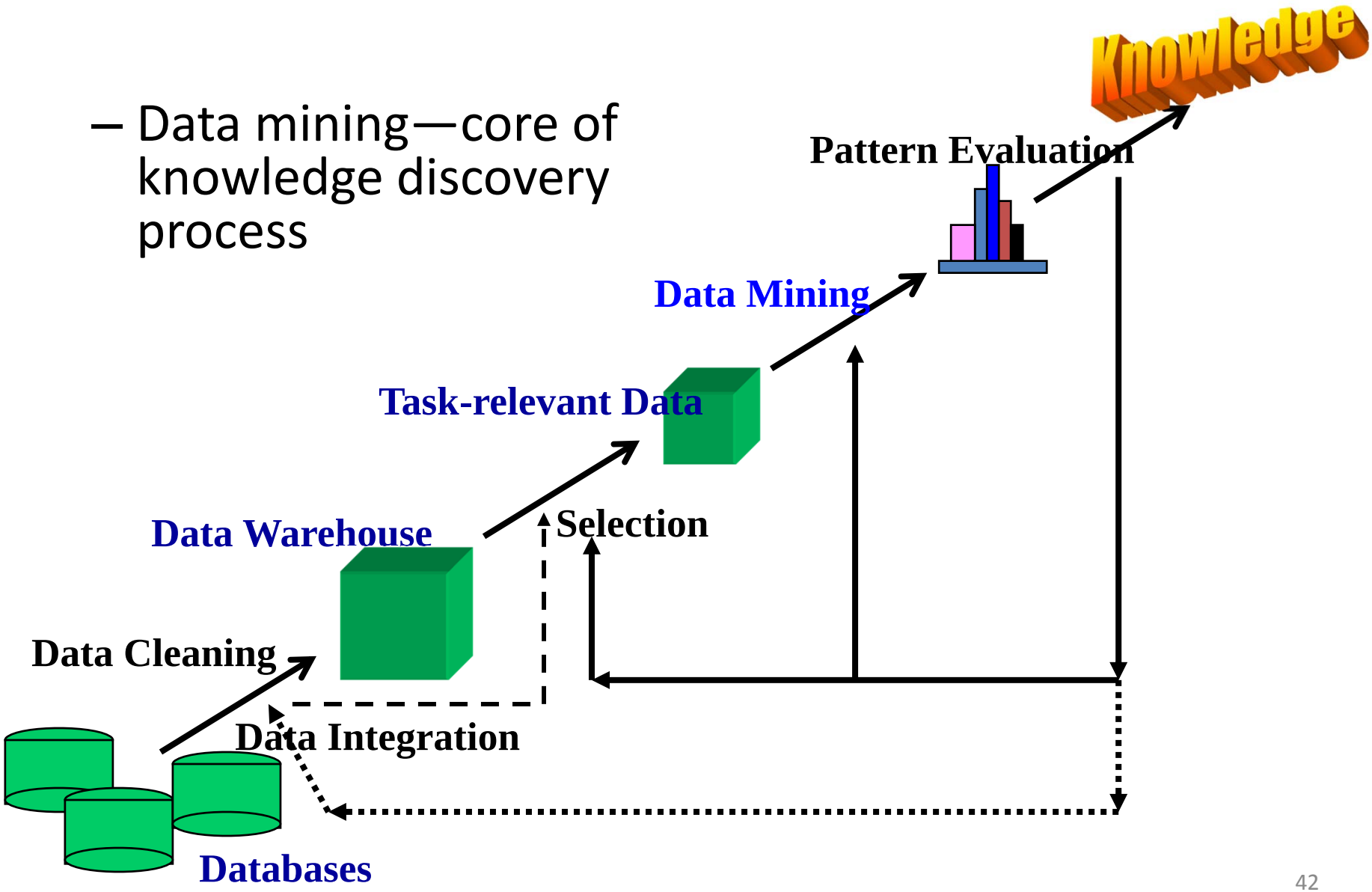


Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)

Knowledge Discovery (KDD) Process

– Data mining—core of knowledge discovery process



KDD Process: Several Key Steps

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation

KDD Process: Several Key Steps

- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Data Mining and Business Intelligence