Data and Data Mining

CSE 4334 / 5334 Data Mining
Spring 2019

Won Hwa Kim

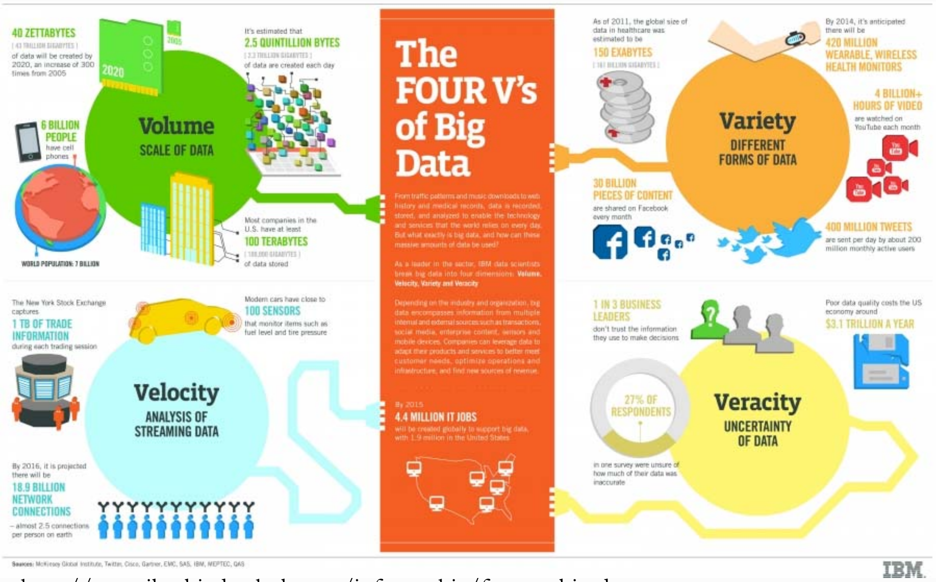
(Slides courtesy of Pang-Ning Tan, Michael Steinbach and Vipin Kumar)

Big Data

DILBERT
BY SCOTT ADAMS
CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
DilbertCartoonist@gmail.com
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
©2012 Scott Adams, Inc.
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
$
www.dilbert.com
7-29-12
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.

http://dilbert.com/strip/2012-07-29

Big Data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
Volume
SCALE OF DATA
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
IBM.

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

3

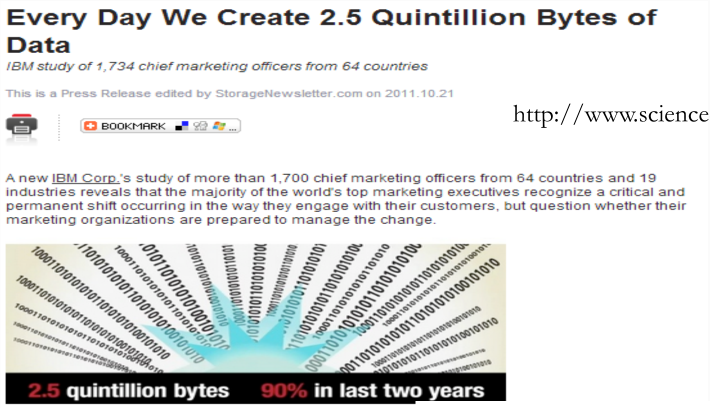
Big Data

The 4 Vs

- Volume
- Variety
- Velocity
- Veracity

4

Volume: How much data is out there?



Every Day We Create 2.5 Quintillion Bytes of Data

IBM study of 1,734 chief marketing officers from 64 countries

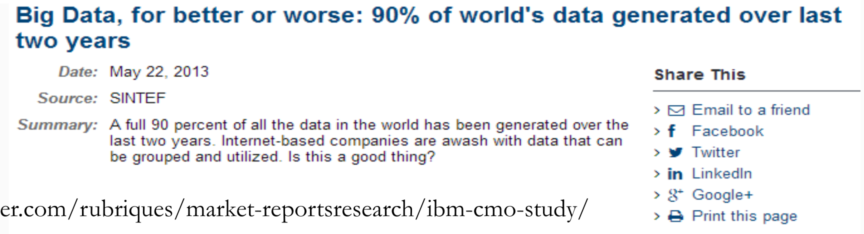
This is a Press Release edited by StorageNewsletter.com on 2011.10.21

BOOKMARK

A new IBM Corp.'s study of more than 1,700 chief marketing officers from 64 countries and 19 industries reveals that the majority of the world's top marketing executives recognize a critical and permanent shift occurring in the way they engage with their customers, but question whether their marketing organizations are prepared to manage the change.

2.5 quintillion bytes 90% in last two years

http://www.sciencedaily.com/releases/2013/05/130522085217.htm



Big Data, for better or worse: 90% of world's data generated over last two years

Date: May 22, 2013

Source: SINTEF

Summary: A full 90 percent of all the data in the world has been generated over the last two years. Internet-based companies are awash with data that can be grouped and utilized. Is this a good thing?

Share This

- Email to a friend
- Facebook
- Twitter
- LinkedIn
- Google+
- Print this page

http://www.storagenewsletter.com/rubriques/market-reportsresearch/ibm-cmo-study/

Variety: Types of Data

Structured data

- (relational) database tables
- CSV/TSV files

Semi-structured data

- XML, JSON, RDF

Unstructured data

- text data (documents, Web pages, short texts, e.g., social media)

Multimedia data

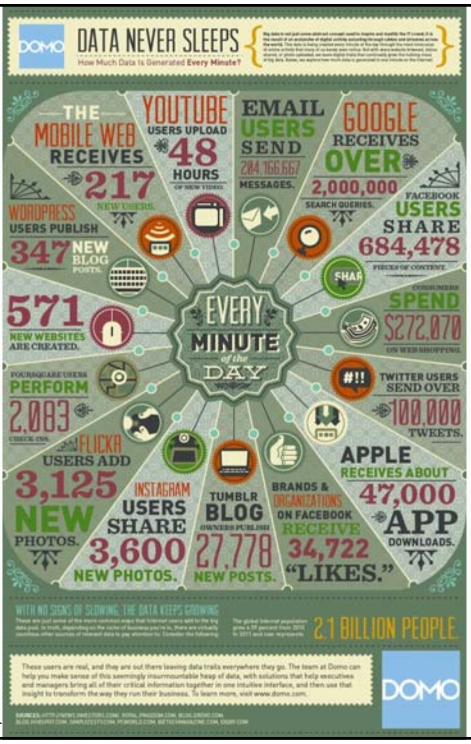
- images, videos, audios

Other types of data

- matrices, graphs, sequences, time-series, spatio-temporal

Velocity: Streaming Data

- ❖ Stock trades
- ❖ Highway sensors
- ❖ Weather data
- ❖ Social media
- ❖ Telephone calls
- ❖ Video streaming


DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
EVERY MINUTE of the DAY
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKR USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING
2.1 BILLION PEOPLE
These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.
DOMO

http://mashable.com/2012/06/22/data-created-every-minute/

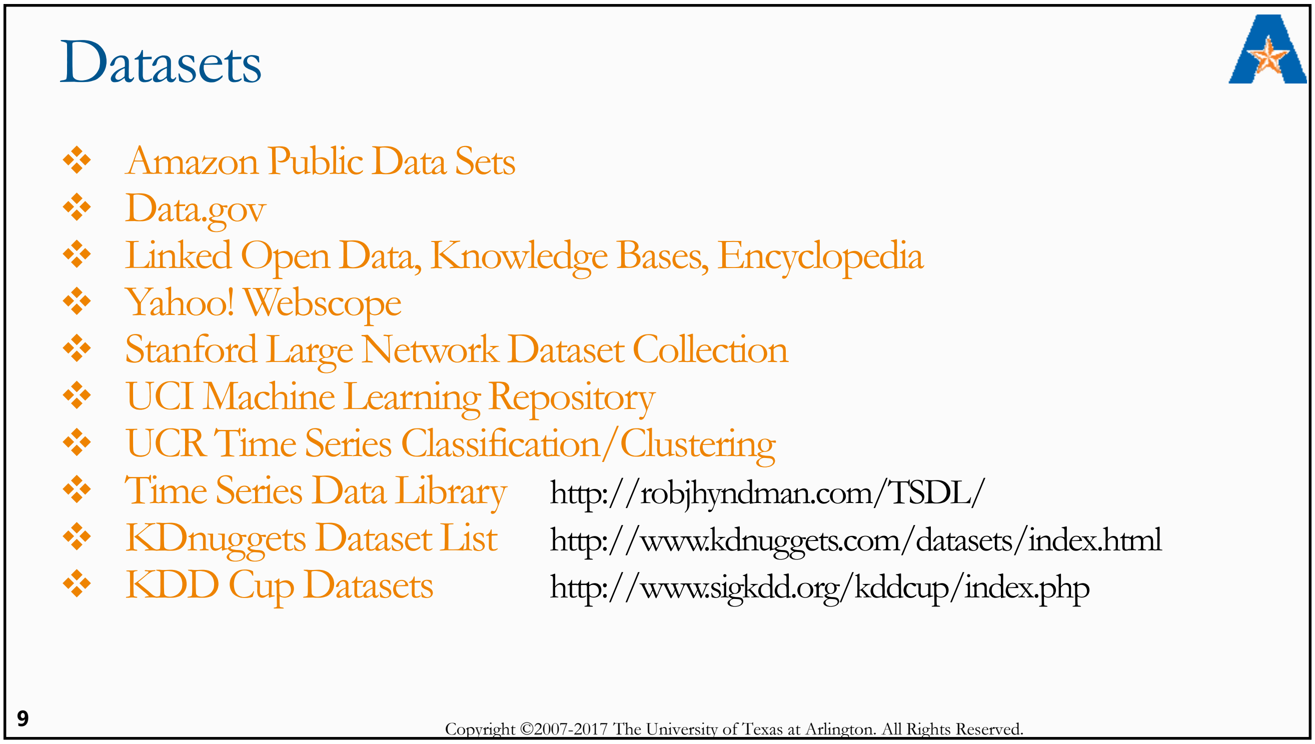
7

Veracity: uncertain and imprecise data

- ❖ Quality and origin of data
- ❖ Consistent? Complete? Integrity?
- ❖ Untrusted and Uncleaned
- ❖ Fake stories

- ❖ Lots of cost to justify the data…

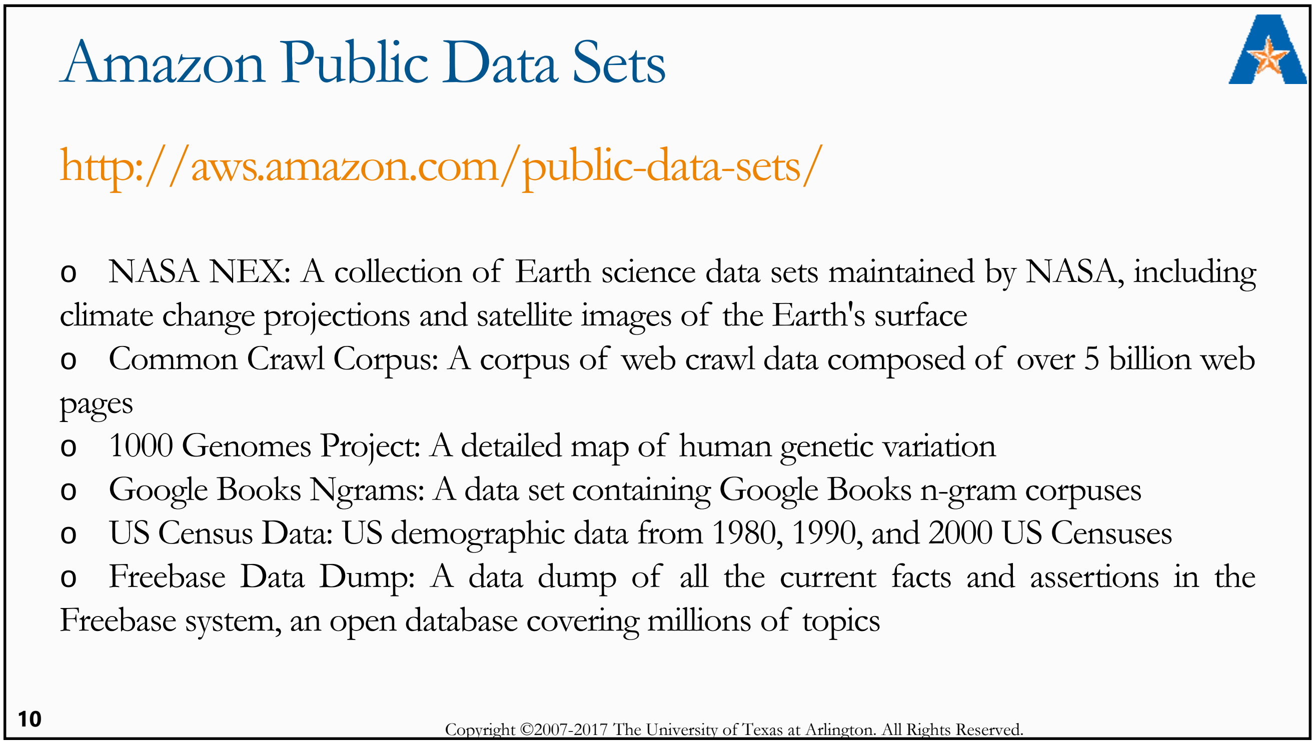
8



Datasets

- Amazon Public Data Sets
- Data.gov
- Linked Open Data, Knowledge Bases, Encyclopedia
- Yahoo! Webscope
- Stanford Large Network Dataset Collection
- UCI Machine Learning Repository
- UCR Time Series Classification/Clustering
- Time Series Data Library http://robjhyndman.com/TSDL/
- KDnuggets Dataset List http://www.kdnuggets.com/datasets/index.html
- KDD Cup Datasets http://www.sigkdd.org/kddcup/index.php

9



Amazon Public Data Sets

http://aws.amazon.com/public-data-sets/

- NASA NEX: A collection of Earth science data sets maintained by NASA, including climate change projections and satellite images of the Earth's surface
- Common Crawl Corpus: A corpus of web crawl data composed of over 5 billion web pages
- 1000 Genomes Project: A detailed map of human genetic variation
- Google Books Ngrams: A data set containing Google Books n-gram corpuses
- US Census Data: US demographic data from 1980, 1990, and 2000 US Censuses
- Freebase Data Dump: A data dump of all the current facts and assertions in the Freebase system, an open database covering millions of topics

10

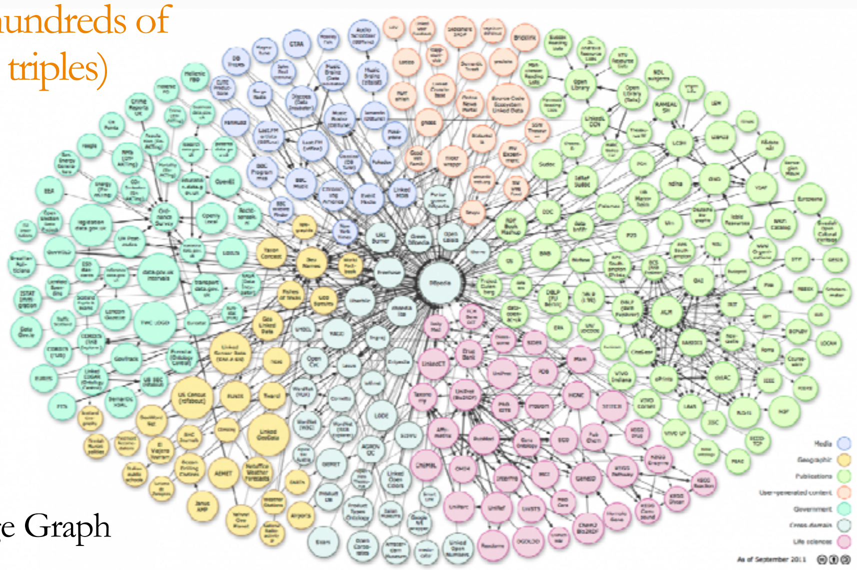
Data.gov

http://www.data.gov/ (137,608 datasets)

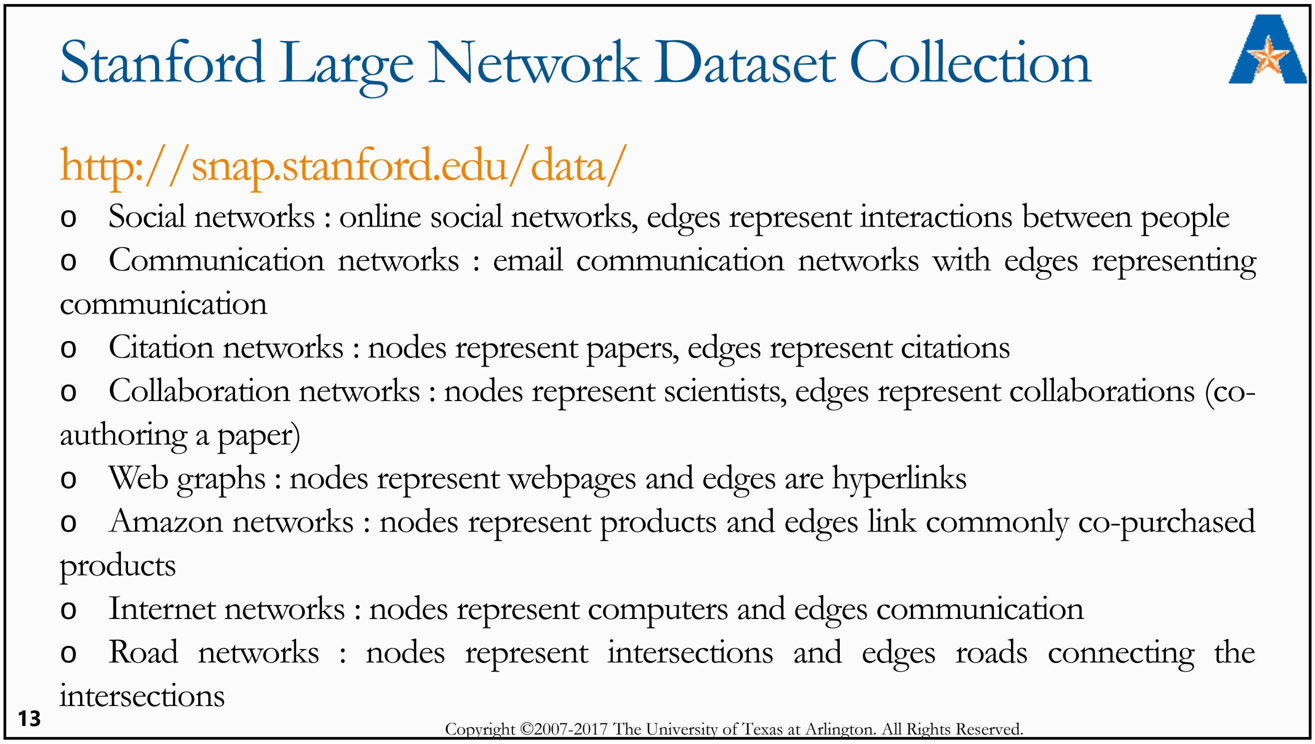
- Consumer Complaint Database
- U.S. International Trade in Goods and Services: Monthly report that provides national trade data including imports, exports, and balance of payments for goods and services.
- DTV Reception Maps
- Food Access Research Atlas — presents a spatial overview of food access indicators for low-income and other census tracts using different measures of supermarket...
- U.S. Hourly Precipitation Data
- Great Chile Earthquake of May 22, 1960
- Consumer Expenditure Survey
- Farmers Markets Geographic Data: longitude and latitude, state, address, name, and zip code of Farmers Markets in the United States
- Crimes - 2001 to present (City of Chicago)

Linked Data, Knowledge Bases, Encyclopedia

http://linkeddata.org/ (hundreds of datasets, billions of RDF triples)



IMDB
DBLP
PubMed
Wikipedia, DBpedia
YAGO
Freebase/Google Knowledge Graph

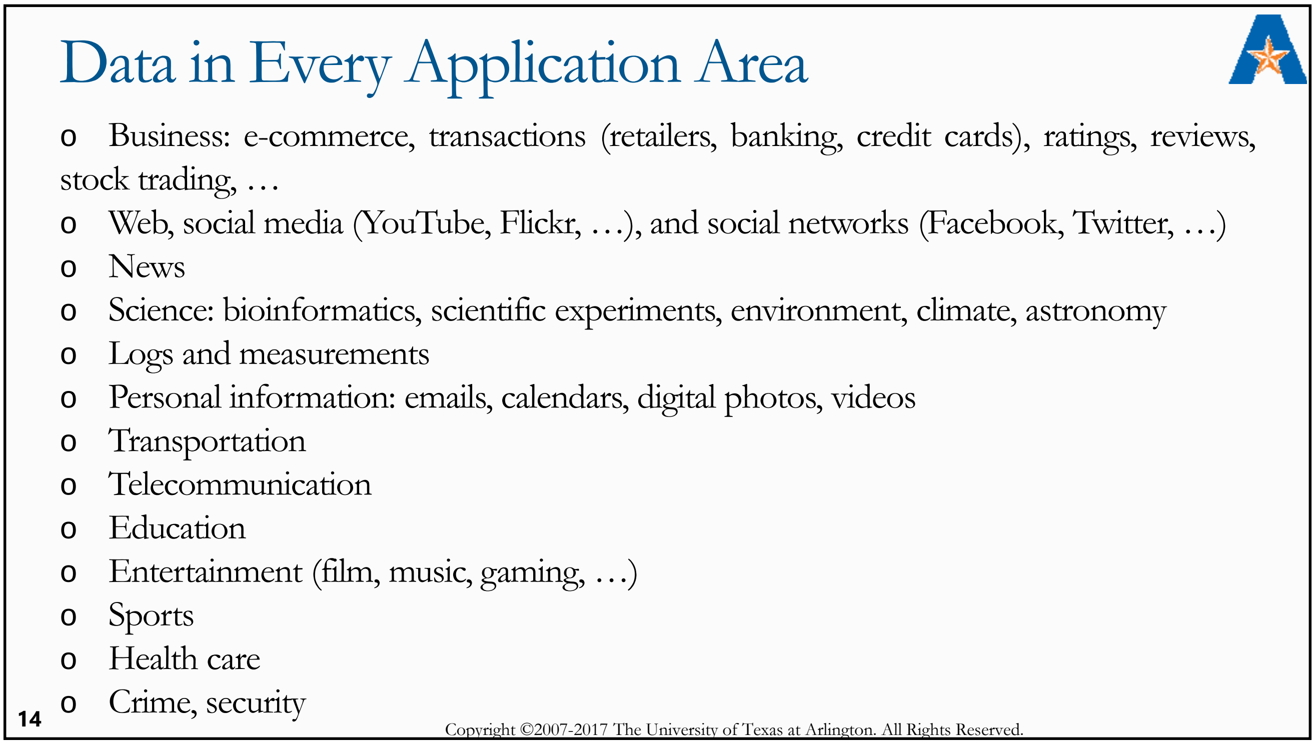


Stanford Large Network Dataset Collection

http://snap.stanford.edu/data/

- Social networks : online social networks, edges represent interactions between people
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections

13



Data in Every Application Area

- Business: e-commerce, transactions (retailers, banking, credit cards), ratings, reviews, stock trading, …
- Web, social media (YouTube, Flickr, …), and social networks (Facebook, Twitter, …)
- News
- Science: bioinformatics, scientific experiments, environment, climate, astronomy
- Logs and measurements
- Personal information: emails, calendars, digital photos, videos
- Transportation
- Telecommunication
- Education
- Entertainment (film, music, gaming, …)
- Sports
- Health care
- Crime, security

14

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

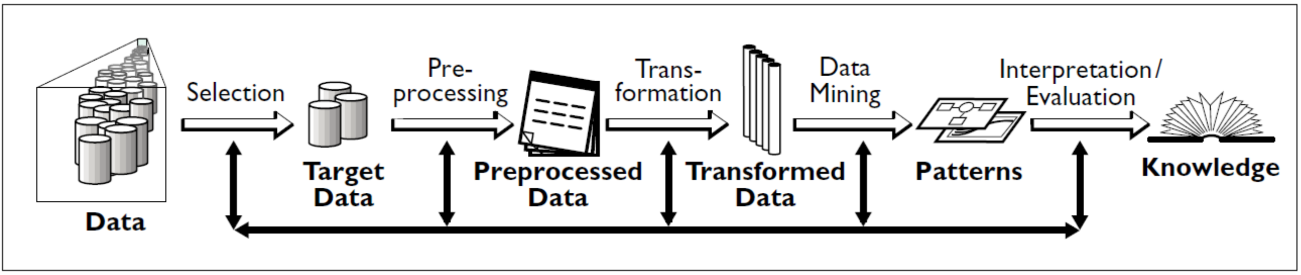
What is not Data Mining?

- Retrieve data instead of knowledge or pattern
- Not interesting (trivial, explicit, known, useless)

15

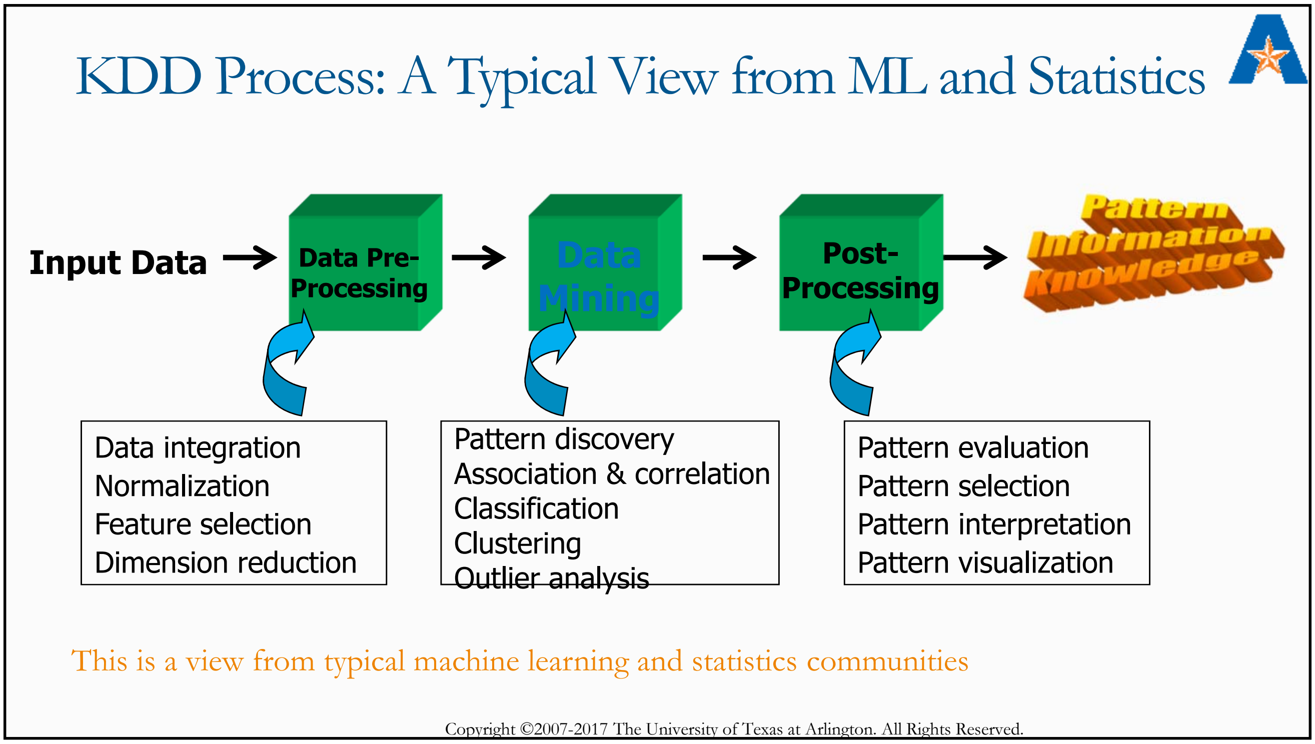
Knowledge Discovery (KDD) Process

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



http://cacm.acm.org/magazines/1996/11/8517-the-kdd-process-for-extracting-useful-knowledge-from-volumes-of-data/abstract

16



KDD Process: A Typical View from ML and Statistics

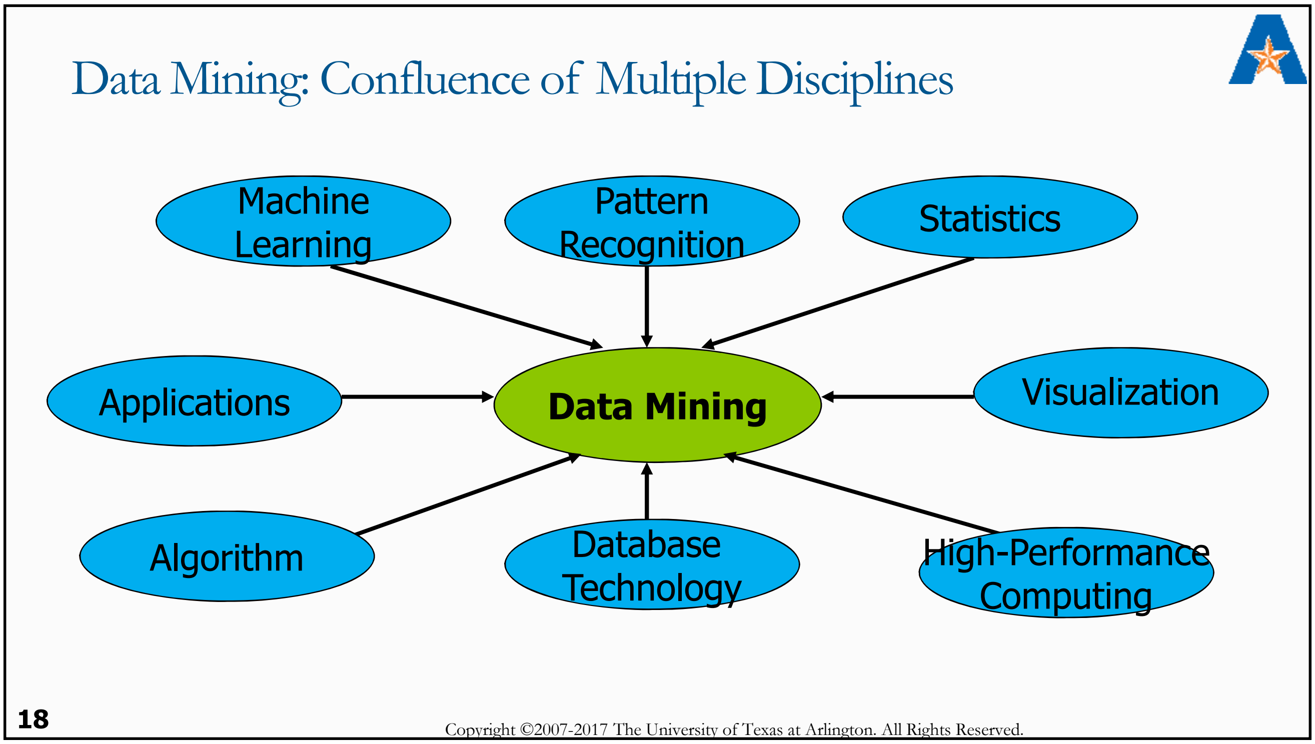
Input Data → Data Pre-Processing → Data Mining → Post-Processing → Pattern Information Knowledge

Data integration
Normalization
Feature selection
Dimension reduction

Pattern discovery
Association & correlation
Classification
Clustering
Outlier analysis

Pattern evaluation
Pattern selection
Pattern interpretation
Pattern visualization

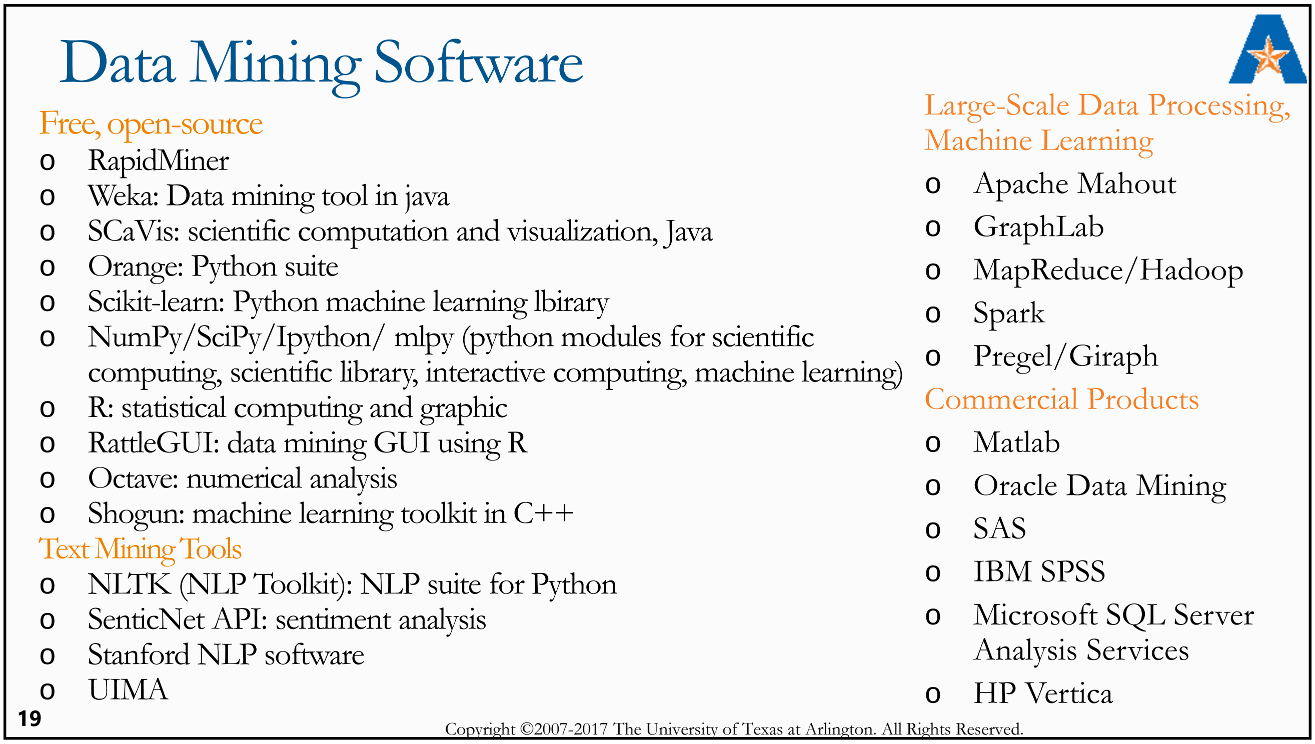
This is a view from typical machine learning and statistics communities



Data Mining: Confluence of Multiple Disciplines

Machine Learning → Data Mining
Pattern Recognition → Data Mining
Statistics → Data Mining
Applications → Data Mining
Visualization → Data Mining
Algorithm → Data Mining
Database Technology → Data Mining
High-Performance Computing → Data Mining

18



Data Mining Software

Free, open-source

- RapidMiner
- Weka: Data mining tool in java
- SCaVis: scientific computation and visualization, Java
- Orange: Python suite
- Scikit-learn: Python machine learning lbirary
- NumPy/SciPy/Ipython/ mlpy (python modules for scientific computing, scientific library, interactive computing, machine learning)
- R: statistical computing and graphic
- RattleGUI: data mining GUI using R
- Octave: numerical analysis
- Shogun: machine learning toolkit in C++

Text Mining Tools

- NLTK (NLP Toolkit): NLP suite for Python
- SenticNet API: sentiment analysis
- Stanford NLP software
- UIMA

Large-Scale Data Processing, Machine Learning

- Apache Mahout
- GraphLab
- MapReduce/Hadoop
- Spark
- Pregel/Giraph

Commercial Products

- Matlab
- Oracle Data Mining
- SAS
- IBM SPSS
- Microsoft SQL Server Analysis Services
- HP Vertica

19