DEPAUL UNIVERSITY

DSC478: Programming Machine Learning Applications

Roselyne Tchoua

rtchoua@depaul.edu

School of Computing, CDM, DePaul University

Introduction to Data Mining

Purpose of this lecture: Re-introduce data mining: Why? What? Where? (What types of data?) and introduce the challenges of data mining, some of which we will explore further in data exploration and pre-processing.

Why Data Mining?

- Explosive Growth of Data: from terabytes (10^{12}) to petabytes (10^{15})
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulations, …
    - Society and everyone: news, images, video, documents, tweets….

The explosive growth of data is because of myriad ways in which we are automatically collecting data, especially on the Web.

Instruments and supercomputers are generating tera/petabytes of data

TOP 500
The List.

Top 500 list of

Traditional methods infeasible with raw data!

The Advanced Photo Source
At Argonne National Laboratory

The Spallation Neutron Source
At Oak Ridge National Laboratory

What happens in an internet minute?

https://www.visualcapitalist.com/what-happens-in-an-internet-minute-in-2019/

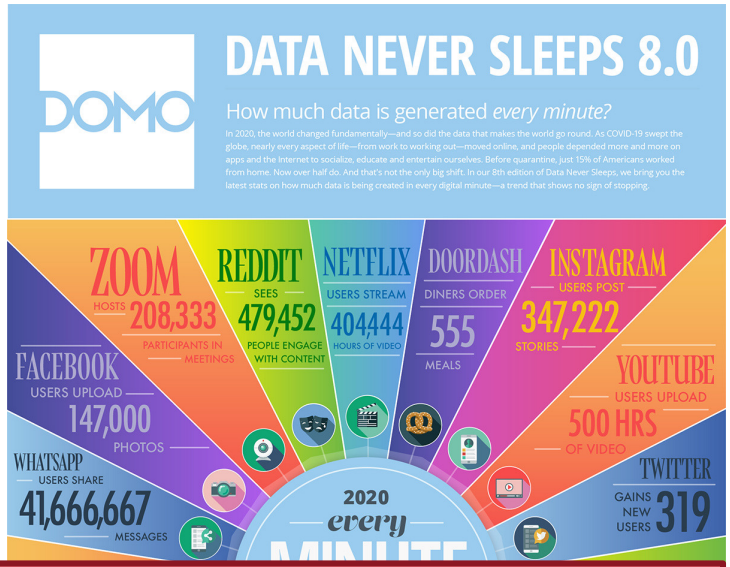
What about 2020?

Notice Zoom hots and participants

Not the same source, different numbers

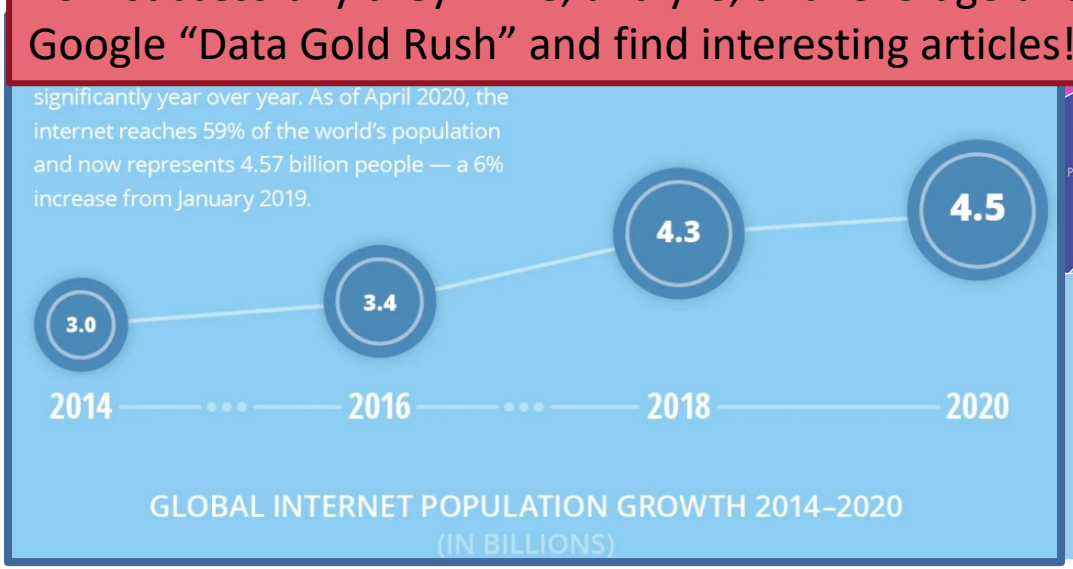
➔ same story

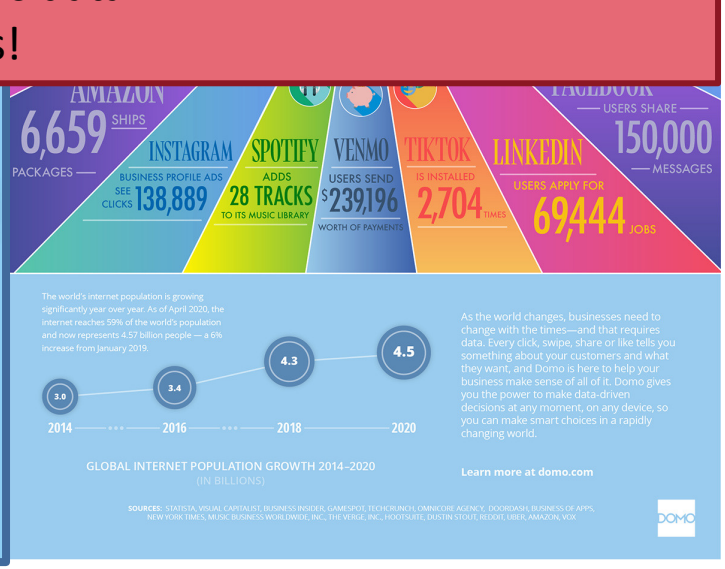
And the future growth is staggering



The ability of these systems and companies to effectively serve their users depends on how successfully they mine, analyze, and leverage this data.

Google “Data Gold Rush” and find interesting articles!





https://www.visualcapitalist.com/every-minute-internet-2020/

https://ediscoverytoday.com/2021/04/16/here-is-your-2021-internet-minute-infographic-ediscovery-trends/

Why Data Mining?

- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks (years in some cases) to discover useful information
- Much of the data is never analyzed at all

Why Data Mining?

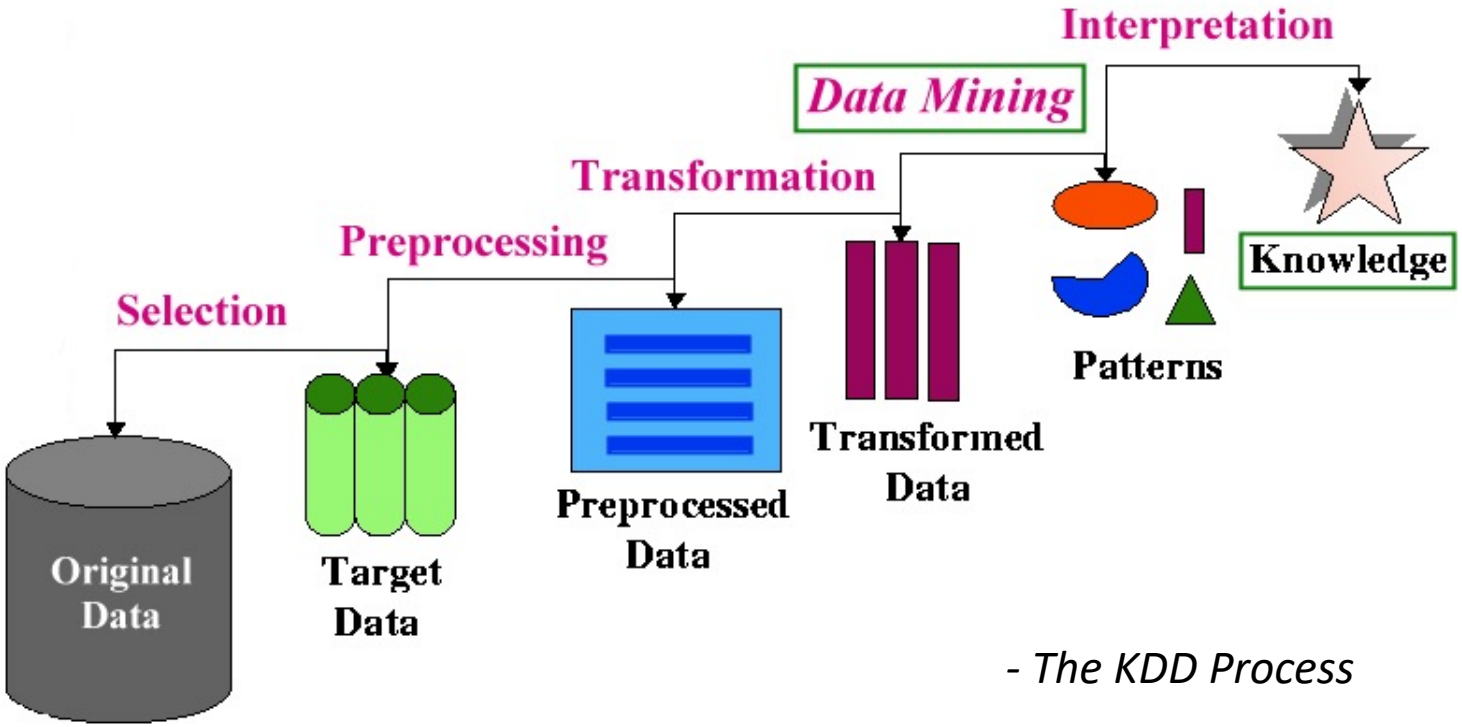
- What we want to do:
  - Extract interesting and useful knowledge from the data
  - Find rules, regularities, irregularities, patterns, constraints
  - Predict future outcomes based on past observations
  - hopefully, this will help us better compete in business, do research, learn concepts, make money, etc.

What is Data Mining?

- Many Definitions
  - **Non-trivial** extraction of **implicit**, **previously unknown** and **potentially useful** information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

What is Data Mining (DM)?

- DM is only part of the KDD process (used to be Knowledge Discovery from Databases, now generally means from Data)
- DM phase generally employs machine learning and statistical techniques



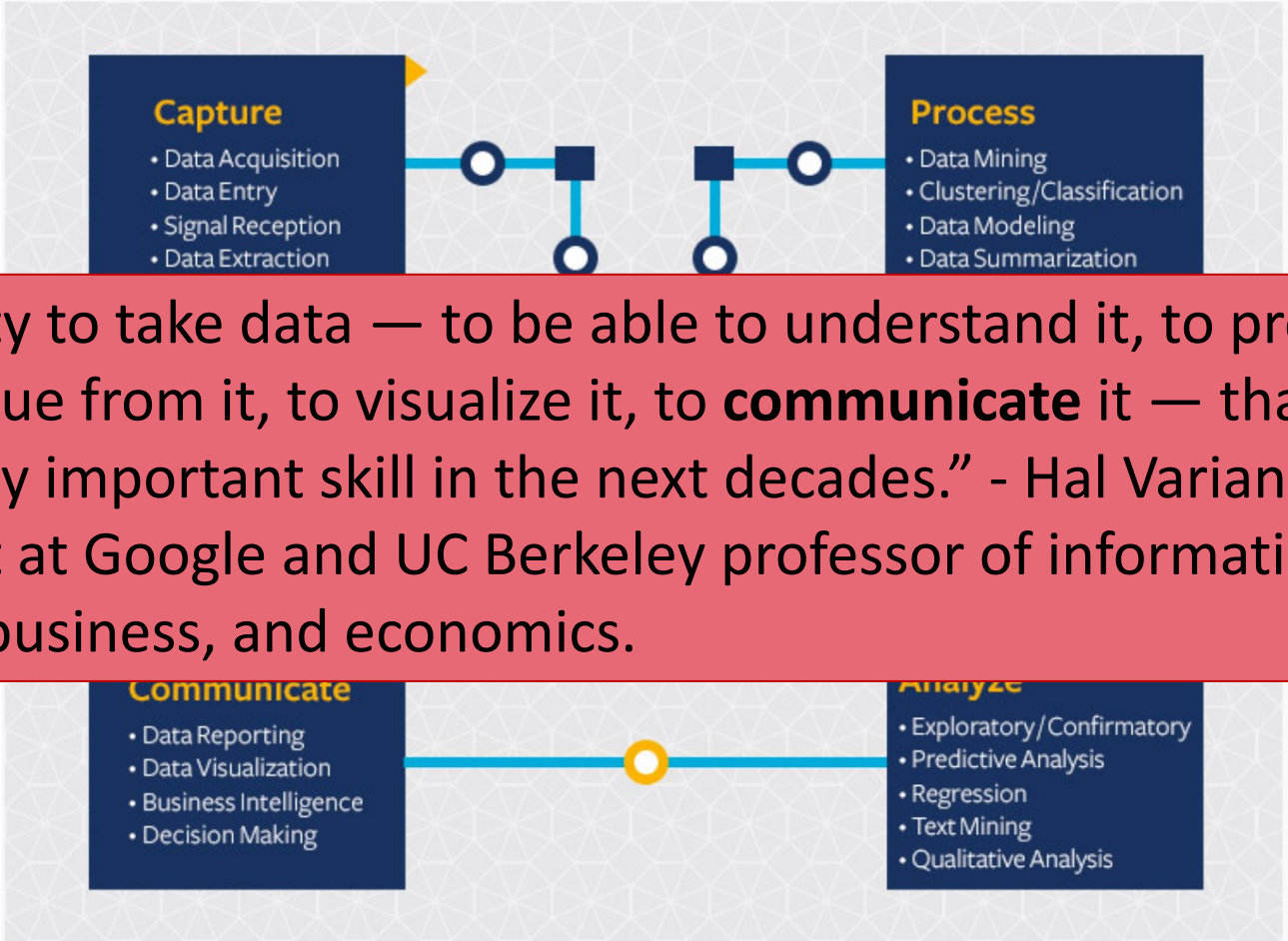
- The KDD Process

From DM to Data Science

- The four paradigms:
  - Scientific Theory
  - Empirical Evidence
  - Computational Science (Computer Simulations)
  - Data Science in Information Age
- You can use DM along with theory, rules etc. Data Science comes in when all you can *learn* from is ***data***

https://en.wikipedia.org/wiki/The_Fourth_Paradigm and several publications

From DM to Data Science

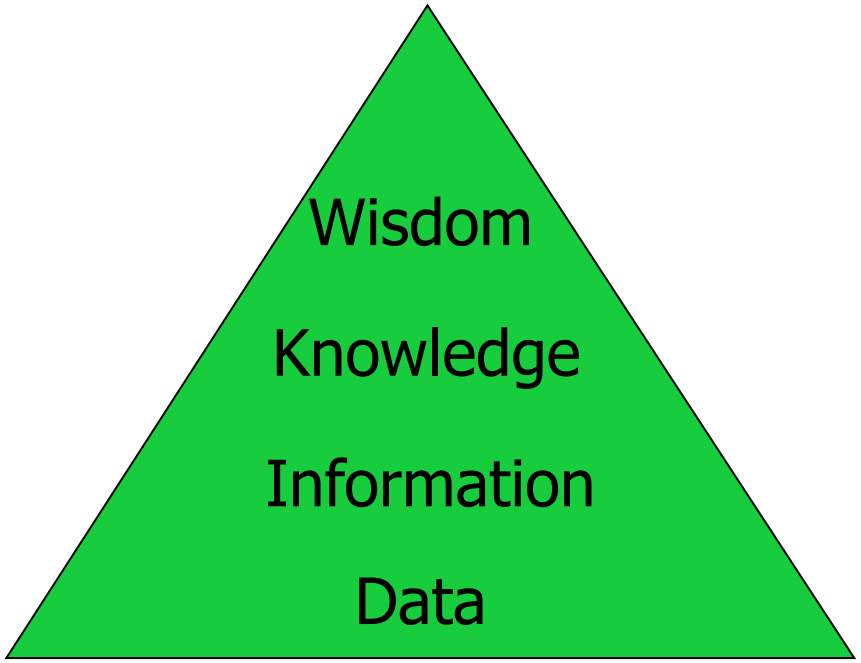


"The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to **communicate** it — that's going to be a hugely important skill in the next decades." - Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics.

https://datascience.berkeley.edu/about/what-is-data-science/

From DM to Wisdom

- Data
  - The raw material of information
  - e.g., Database with all the files storing information about the housing market in Chicago
- Information
  - Data organized and presented by someone
  - e.g., Houses sold by month in a particular neighborhood and their characteristics
- Knowledge
  - Information read, heard or seen and understood and integrated
  - e.g., Relationships between number of bedrooms and prices
- Wisdom
  - Distilled knowledge and understanding which can lead to decisions
  - Model that predicts prices for newly available houses on the market

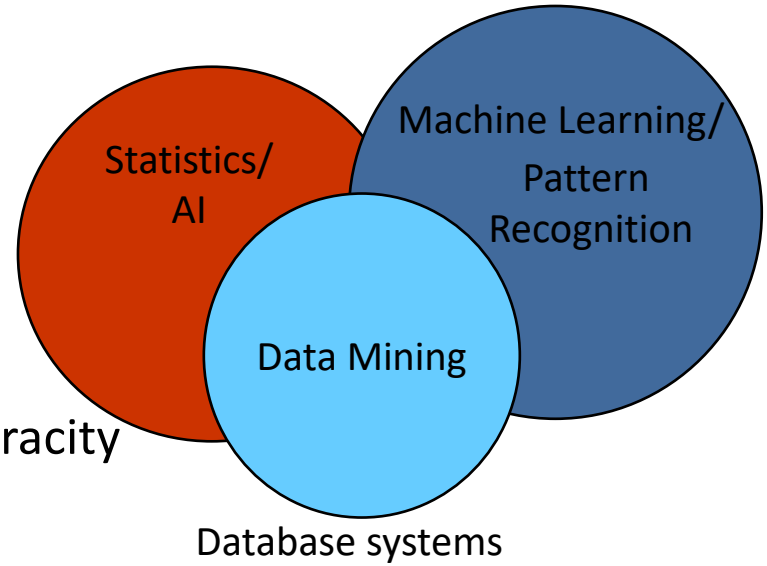


The Information Hierarchy

DM and Other Disciplines

- DM draws ideas from machine learning/AI, pattern recognition, statistics, and database systems

- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data
  - May hear about 3 V's, 4V's, 5V's
    - Volume, velocity, **value**, variety, veracity



DM Tasks

- Prediction Methods
  - May hear about Directed Knowledge Discovery.
  - Use some variables to predict unknown or future values of other variables (goal oriented).
  - e.g., what happens in the stock market two days after the price of coffee drops?
- Description Methods
  - May hear about Undirected Knowledge discovery.
  - Find human-interpretable patterns that describe the data (exploratory).
  - e.g., What else do customers buy when they buy a computer?

What are some DM Tasks?

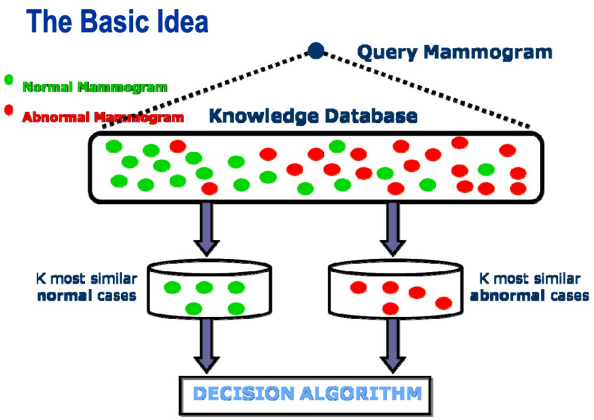
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Outlier Detection [Predictive or Descriptive]

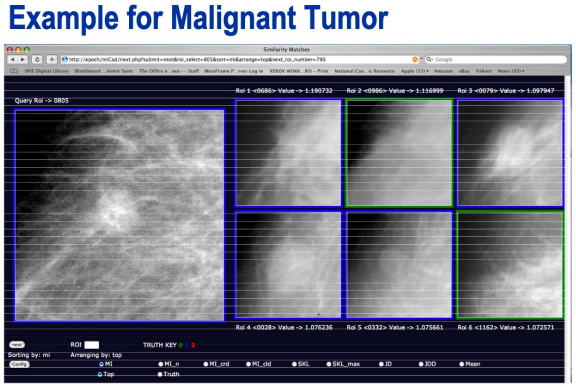
Classification

- Given a collection of records (training set)
  - Each record contains a set of *features*, one of the attributes is the *class label*.
- Build a *model* to predict the class label as a function of the values of other attributes.
- Goal: previously unseen (test set) records should be assigned a class as accurately as possible.

Classification Application

- Process many images of breast X-Rays
- i.e., extract image features: color, intensity, shapes etc.
- ***Learn*** what a malignant case ***looks like***
- *Predict a "sick" or "will-be-sick" lung*
- *Assist radiologist or physician in diagnosing patients (computer aided diagnosis)*





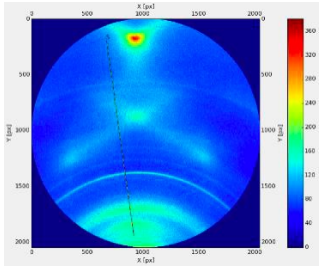
Six top retrieved cases (4/6 true masses)
Top retrieval is a malignant mass. (like the query)

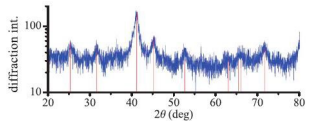
Clustering

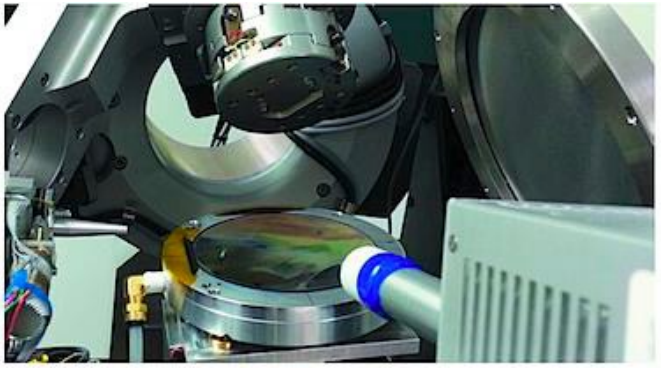
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance
  - Manhattan Distance
  - Many other data-or-problem-specific measures

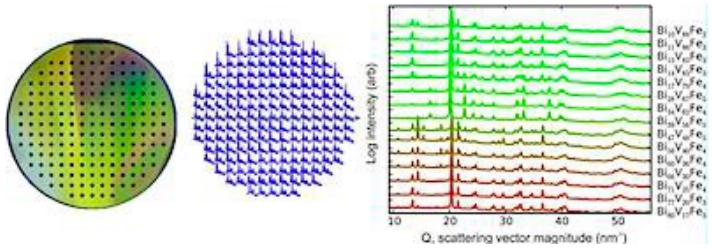
Clustering Application

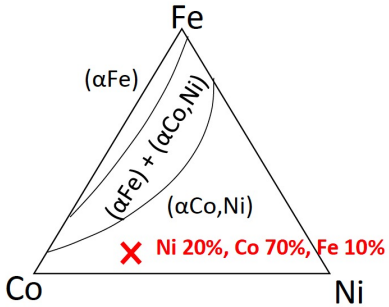
- Process many ternary material samples
- *Learn* what a phase looks *like*
- Group samples together and avoid having to measure 100's of samples
- Assist materials scientists in discovering nea materials (with *new properties***) fast!**











DM Tasks (continued)

- Data Mining/Machine Learning Tasks are often inter-related
  - Find clusters ➔ tentatively create classes
  - Try and predict the classes to validate exploration
- Often need to try different techniques/algorithms for each task

What kind of data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
  - Object-relational databases, Heterogeneous databases and legacy databases
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and information networks
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text data and other semi-structured data
  - The World-Wide Web

What kind of data?

Structured Databases

- relational, object-relational, etc.
- can use SQL to perform parts of the process

e.g., SELECT count(*) FROM Items WHERE type=video GROUP BY category

Borrow

customerID	date	itemID	#	…
C1234	99/09/06	98765	1	…
. . .				

Customer

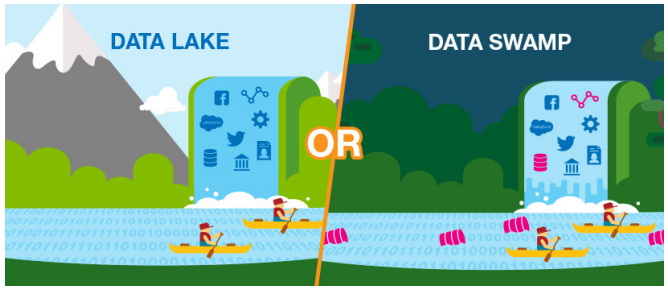
customerID	name	address	password	birthdate	family_income	group	…
C1234	John Smith	120 main street	Marty	1965/10/10	$45000	A	…
. . .							

Items

itemID	type	title	media	category	Value	#	…
98765	Video	Titanic	DVD	Drama	$15.00	2	…
. . .							

What kind of data?

- Flat Files (Data Lakes or Data Swamps)
  - most common data source
  - can be text (or HTML) or binary
  - may contain transactions, statistical data, measurements, etc.



- Transactional databases
  - set of records each with a transaction id, time stamp, and a set of items
  - may have an associated "description" file for the items
  - typical source of data used in market basket analysis

Rentals

transactionID	date	time	customerID	itemList
T12345	99/09/06	19:38	C1234	{I2, I6, I10, I45 …}
…				

What kind of data?

- Other Types of Databases
  - legacy databases
  - multimedia databases (usually very high-dimensional)
  - spatial databases (containing geographical information, such as maps, or satellite imaging data, etc.)
  - Time Series Temporal Data (time dependent information such as stock market data; usually very dynamic)

- World Wide Web
  - Basically, a large, heterogeneous, distributed database
    - need for new or additional tools and techniques
    - information retrieval, filtering and extraction
    - Application: agents to assist in browsing and filtering
  - Web content, usage, and structure (linkage) mining tools
    - The “social Web”
    - User generated meta-data, social networks, shared resources, etc.
    - Application: Discover who starts “trends”

Other DM Applications

Business data analysis and decision support

- Marketing focalization
  - Recognizing specific market segments that respond to particular characteristics
  - Return on mailing campaign (target marketing)
- Customer Profiling
  - Segmentation of customer for marketing strategies and/or product offerings
  - Customer behavior understanding
  - Customer retention and loyalty
  - Mass customization / **personalization**

More DM Applications

- Fraud detection
  - Detecting telephone fraud:
    - Telephone call model: destination of the call, duration, time of day or week
    - Analyze patterns that deviate from an expected norm
  - Detection of credit-card fraud
  - Detecting suspicious money transactions (money laundering)
- Text mining:
  - Message filtering (e-mail, newsgroups, etc.)
  - Sentiment analysis
  - Text and document categorization
- Personalization and Recommendation
  - Learn from user/customer preference and predict their future interest

Big Issues in DM

- **Scalability** – How to apply algorithms to large volumes of data
- **Dimensionality** – How many features are needed? What if the data is sparse?
- **Complex** and **Heterogeneous** Data – How to leverage heterogeneous datasets.
  - Personalized medicine: Can you use doctor's notes (NLP), X-RAY/MRI images (image processing and neural networks) and patient surveys in addition to patient's health numbers (heartbeat, blood pressure etc.)
- **Data Quality** – How was data measured? sampled? Are there missing values, errors?
- **Data Ownership and Distribution** – complicated issues e.g., copyrights
- **Privacy Preservation** – Anonymizing data to learn patterns
- **Streaming Data** – Monitoring running simulation
- **Interpretability** – If one can't understand/explain the model, can we trust it?
- **Ethics** – Big one! Who is responsible for self-driving car accident?

Summary

Try and answers for the following questions:

- Why data mining?
- What is data mining?
- What is data science?
- What kind of data?
- What are data mining tasks? Which one will we focus on?
- What are major issues in data mining?

Summary

- We need DM because we are drowning in data! In some cases, data is all we have, and data science is here to stay.
- We can learn from many types of data.
  - The reason for all these types of data is domain-specific, problem-specific and boils down to what is the **most efficient** or **only feasible** way to solve certain problems?
  - Data heterogeneity is a challenge in DM.
- In this course we will focus on classification, regression (both are predictive) and clustering, but there are other DM tasks.
- Another definition: Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.
  - **Are all the results interesting? NO**

References

- Notes are from:
  - Han, Jiawei, Micheline Kamber, and Jian Pei. "Data mining concepts and techniques third edition." The Morgan Kaufmann Series in Data Management Systems(2011): 83-124.
  - and other sources