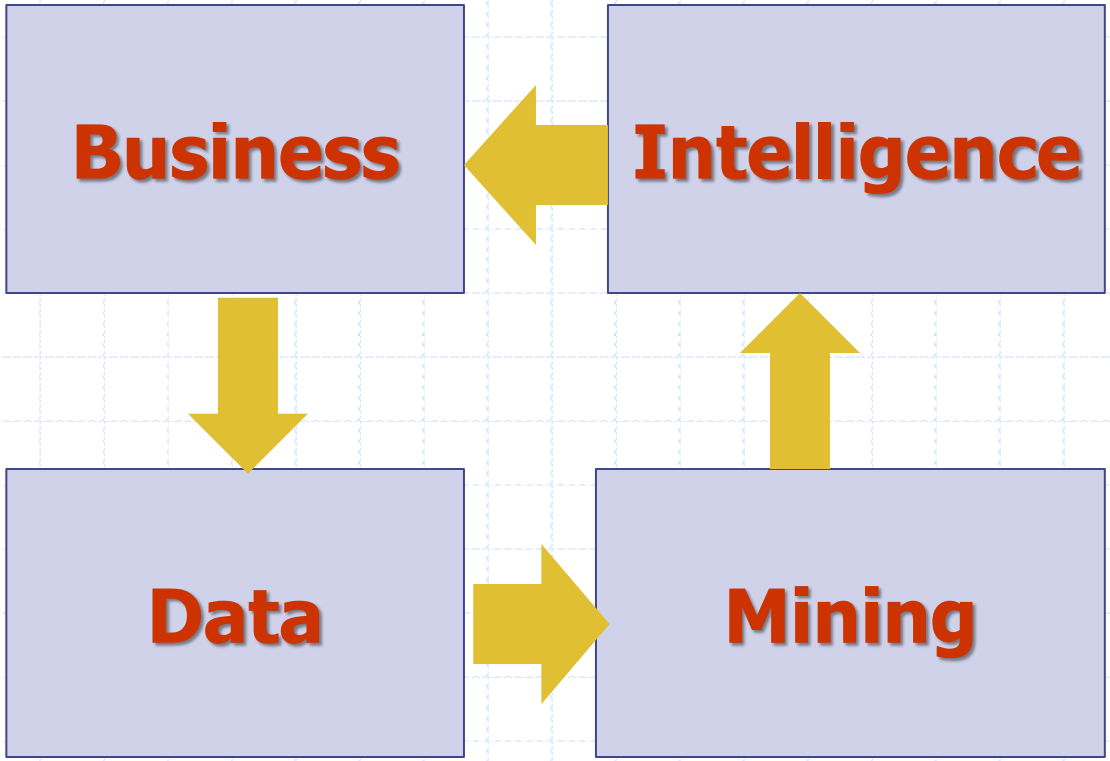
Lesson 1: Introduction to Big Data Analytics

Goals and Objectives of the Course

- The intelligence of any area or organization is reflected in its processes and structures. One important and growing aspect of that is embedded in its information systems and associated data. Data Analytics is the overall field that includes gathering and analysis of this data to discover the underlying structure and patterns and producing actionable intelligence, to help managers drive superior performance. This course will focus on the use of statistical and machine learning techniques to discover new patterns and insights in large data sets.

What the course is all about

Business generates data. Data is Mined to produce Insights & Intelligence. Intelligence is fed back into business. Business performance improves; more data ...



BIDM Cycle

- Business is work that provides products and services that are useful for the society and that generate revenue and income for the providers.
- Business activities are recorded on paper or using electronic media, and then these records become data.
- All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning.
- These ideas can then be fed back into the business so that it can evolve to become more effective and efficient in serving customer needs. And the cycle continues on.

What is Data Mining?

- Data mining is the art and science of discovering knowledge, insights, and patterns in large quantities of data.
  - *Predicting*: outcome of a sports event, trends in the stock market
  - *Decision-making:* approve loan or not
  - *Forecasting:* weather patterns, such as rainfall in a country or region
- Patterns must be valid, potentially useful, understandable
  - E.g. "customers who buy *cheese* and *milk* also buy *bread* 90% of the time"

Supervised vs. Unsupervised Learning

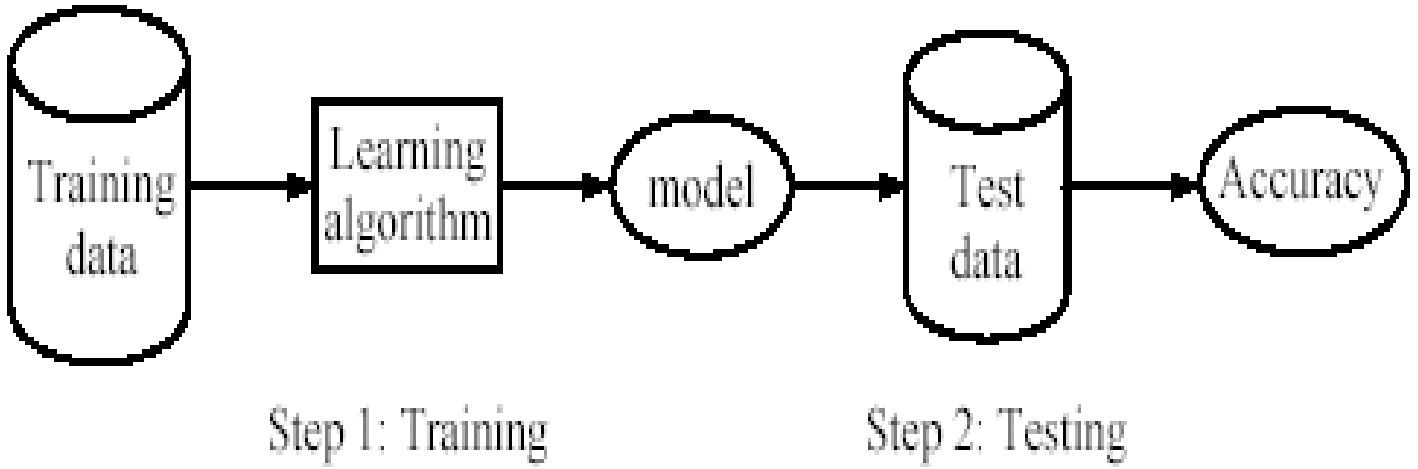
- Supervised learning: classification is seen as supervised learning from examples.
  - Supervision: The data (observations, measurements, etc.) are labeled with pre-defined classes.
  - Test data are classified into these classes too, and predictive accuracy is checked.
- Unsupervised learning: e.g. clustering
  - Class labels of the data are unknown
  - Given a set of data, the task is to establish the existence of classes or clusters in the data

Supervised Learning Process: Two Steps

Learning (training): Learn a model using the training data

Testing: Test the model using the test data to assess the model accuracy

$$\text{Accuracy} = \frac{\#\ of\ correct\ classifications}{total\ number\ of\ test\ cases}$$



Data Mining – Major Techniques

Supervised Learning (Predictive ability based on past data)	Classification – Machine Learning	Decision Trees Neural Networks SVM Naïve Bayes KNN
	Classification - Statistics	Regression
Unsupervised Learning (Exploratory analysis to discover patterns)	Clustering Analysis	K-Means
	Association Rules	Apriori

Standard Data Mining Process

Generic Steps

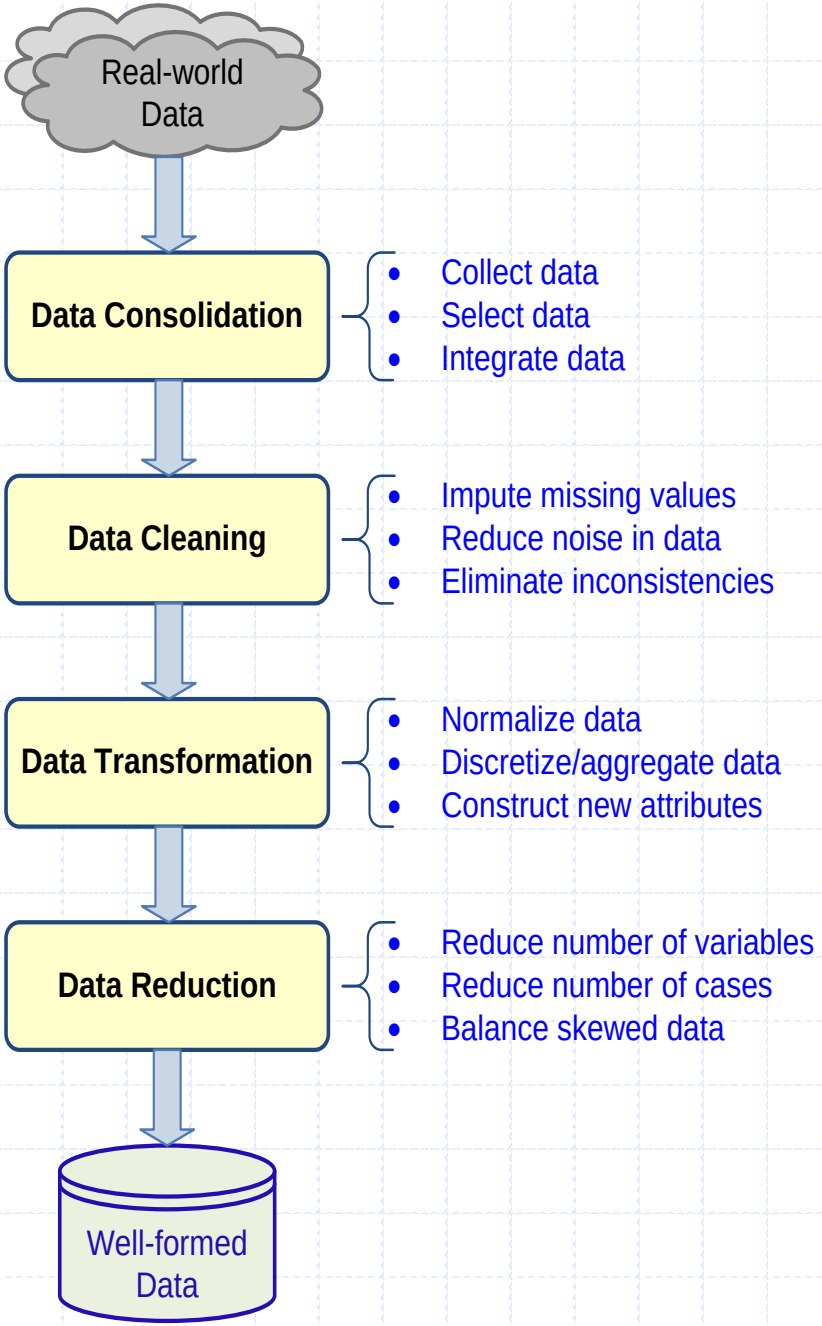
- Understand the application domain
- Identify data sources and select target data
- Pre-process: cleaning, attribute selection
- Data mining to extract patterns or models
- Post-process: identifying interesting or useful patterns
- Incorporate patterns in real world tasks

Data in Data Mining

- Data: a collection of facts usually obtained as the result of experiences, observations, or experiments
- Data may consist of numbers, words, images, …
- Data: lowest level of abstraction (from which information and knowledge are derived)

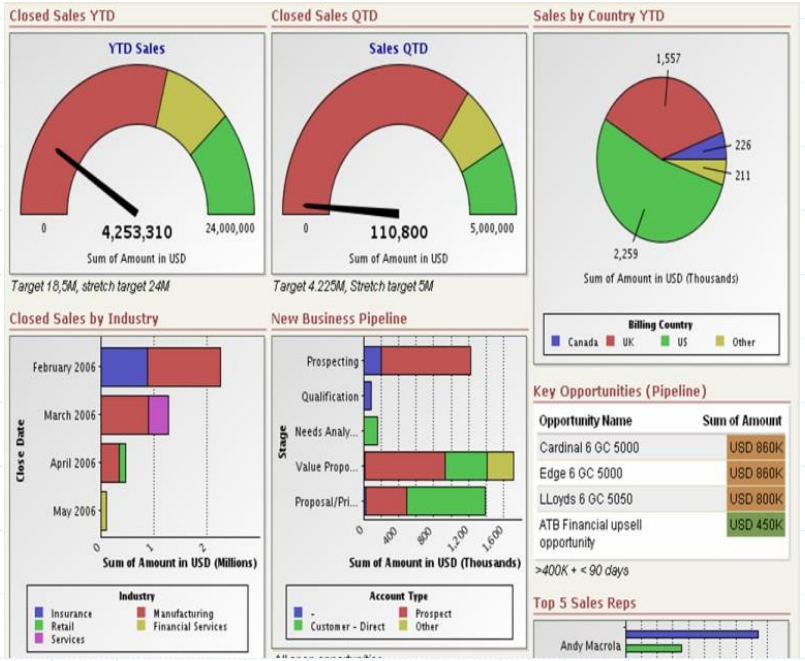
Data Preparation – A Critical Task

- Quality of data is key to data mining effectiveness
- Data scientists spend 60% of the time in organizing and cleaning data!!!



Data Visualizing

- Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports.
- Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can easily grasp difficult concepts or identify new patterns.



Machine Learning v.s. Data Mining

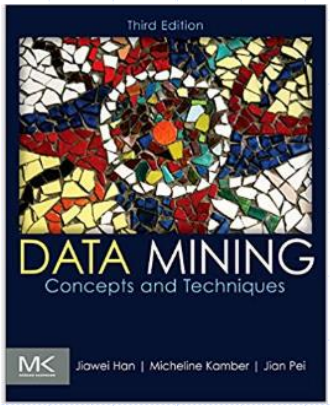
- Machine Learning: the field of development of computer algorithms to transform data into intelligent action. The study of generalization from data is the central topic of machine learning.
- Although there is some disagreement over how widely machine learning and data mining overlap, a potential points of distinction is that machine learning focuses on teaching computers how to use data to solve a problem, while data mining focuses on teaching computers to identify patterns that humans then use to solve a problem.

Machine Learning v.s. Data Mining

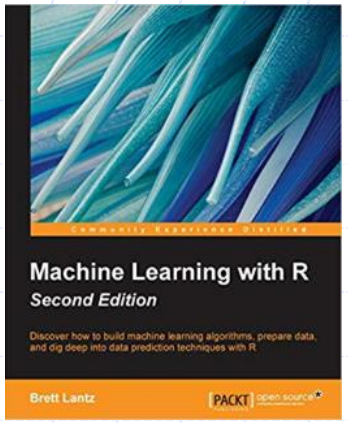
- Virtually all data mining involves the use of machine learning, but not all machine learning involves data mining.
- Machine learning has applications like Natural Language Processing, Computer Vision, Robotics, Bioinformatics…
- For example, you might apply machine learning to mine automobile traffic data for patterns related to accident rates; on the other hand, if the computer is learning how to drive a car itself, this is purely machine learning without data mining.

Books

- The recommended text for the course is *Data Mining Concepts and Techniques 3rd Edition*, by Jiawei Han, Micheline Kamber and Jian Pei, available through Amazon Books and Barnes and Noble.



- *Machine Learning with R: Expert techniques for predictive modeling to solve all your data analysis problems, 2nd Edition* by Brett Lantz



Calendar

CS488 Big Data Analytics						
Week	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**
1	**AM: Lesson 0**: *Math Review and Introduction to R* **Lesson 1:** *Introduction to BDA* **PM: Lab 0**	**AM: Lesson 2:** *Simple Linear Regression* **PM: Lab 2**	**AM: Lesson 3**: *Multiple Linear Regression* **Lesson 4:** *Logistic Regression* **PM: Lab 3, 4**	**AM: Lesson 5**: *Data Visualization* **PM: Lab 5**	**AM: Lesson 5**: *Data Visualization* **PM: Lab 5**	**AM:** *Review and Lab Solutions*
2	**AM: Lesson 6:** *Bayes' Rule* **PM: Lab 6**	**AM: Lesson 7:** *Decision Trees* **PM: Lab 7**	**AM: Lesson 8**: *KNN* **PM: Lab 8**	**AM:** *Review for Midterm* **PM: Study for midterm**	**AM:** *Lab Solutions* **PM: Study for midterm**	Midterm Exam
3	**AM: Lesson 9:** *Association Rule Mining* **PM: Lab 9**	**AM: Lesson 10:** *Clustering* **PM: Lab 10**	**AM: Lesson 11:** *Text Mining* **PM: Lab 11**	**AM:** *Review for Final Exam* **PM: Study for final**	**AM**: *Research Project Introduction* **PM: Study for final**	Final Exam
4	Research Project	Research Project	Research Project	Project Presentation		

Class Schedule

- Class is in session from 10:00 AM to 12:30 PM Mon-Sat. On Mon-Fri, the afternoon session resumes at 1:30 and continues till 3:15 (approximately). There will be a group meditation at 12:15 each day and, on Mon-Fri, at (approximately) 2:55.

Labs

- There is a lab for every lecture. Please see submission instructions in Sakai.
- You are welcome to discuss with your classmates. But be sure you attempt and understand every lab. Similar concepts will appear in the exams.

Evaluation

- Your final grade will be a combination of your scores on Exams, Homework, Project, Professional Etiquette and Morning Meditation. Professional Etiquette is an evaluation of your attendance and professional manner in class. You may earn extra credit towards your final grade if you have outstanding attendance at the morning meditations this block.
  - 70% and above: .5% EC (16 days in a standard block)
  - 80% and above: 1% EC (18 days in a standard block)
  - 90% and above: 1.5% EC (20 days in a standard block)

Evaluation Modality	Value
Midterm Exam	42%
Final Exam	35%
Homework	13%
Professional Etiquette	tie breaker
Research Project	10%
Morning Meditation Bonus	0%-1.5%

Range	Letter Grade
93-100	A
90 - 92	A-
87 - 89	B+
83 - 86	B
80 - 82	B-
77 – 79	C+
73 – 76	C
67 – 72	C-
0 - 66	NC