Data mining introduction

What is Data Mining?

- Knowledge discovery from data
  - Discover the hidden pattern
  - Eliminate randomness

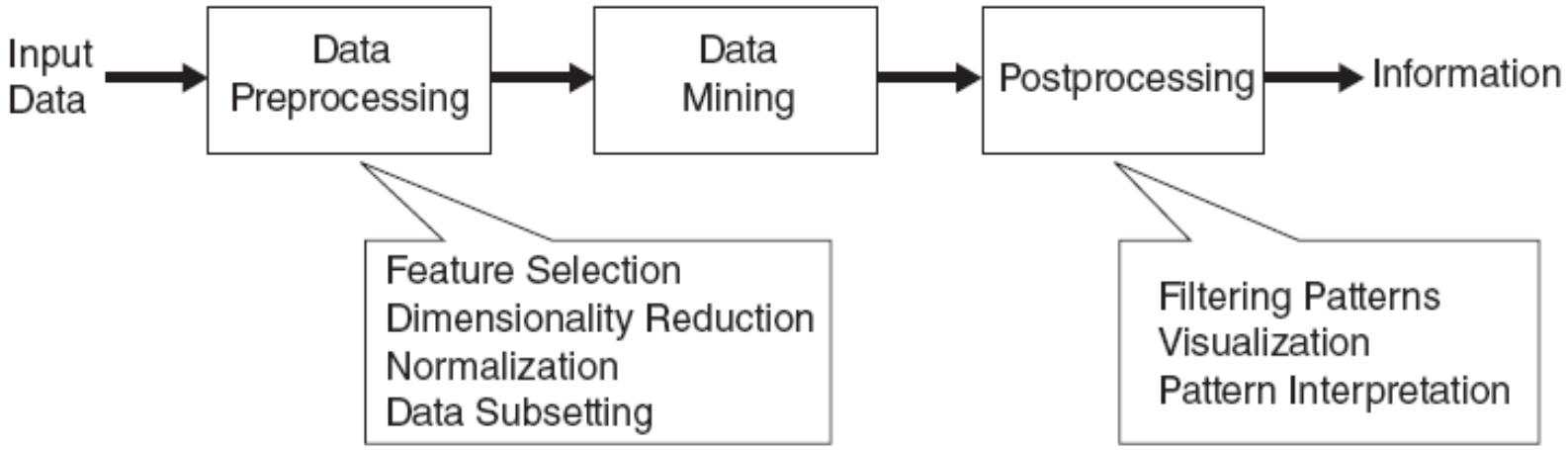


Figure 1.1. The process of knowledge discovery in databases (KDD).

Data contains value and knowledge

Data Mining

- But to extract the knowledge data needs to be
  - Stored
  - Managed
  - And ANALYZED ← this class

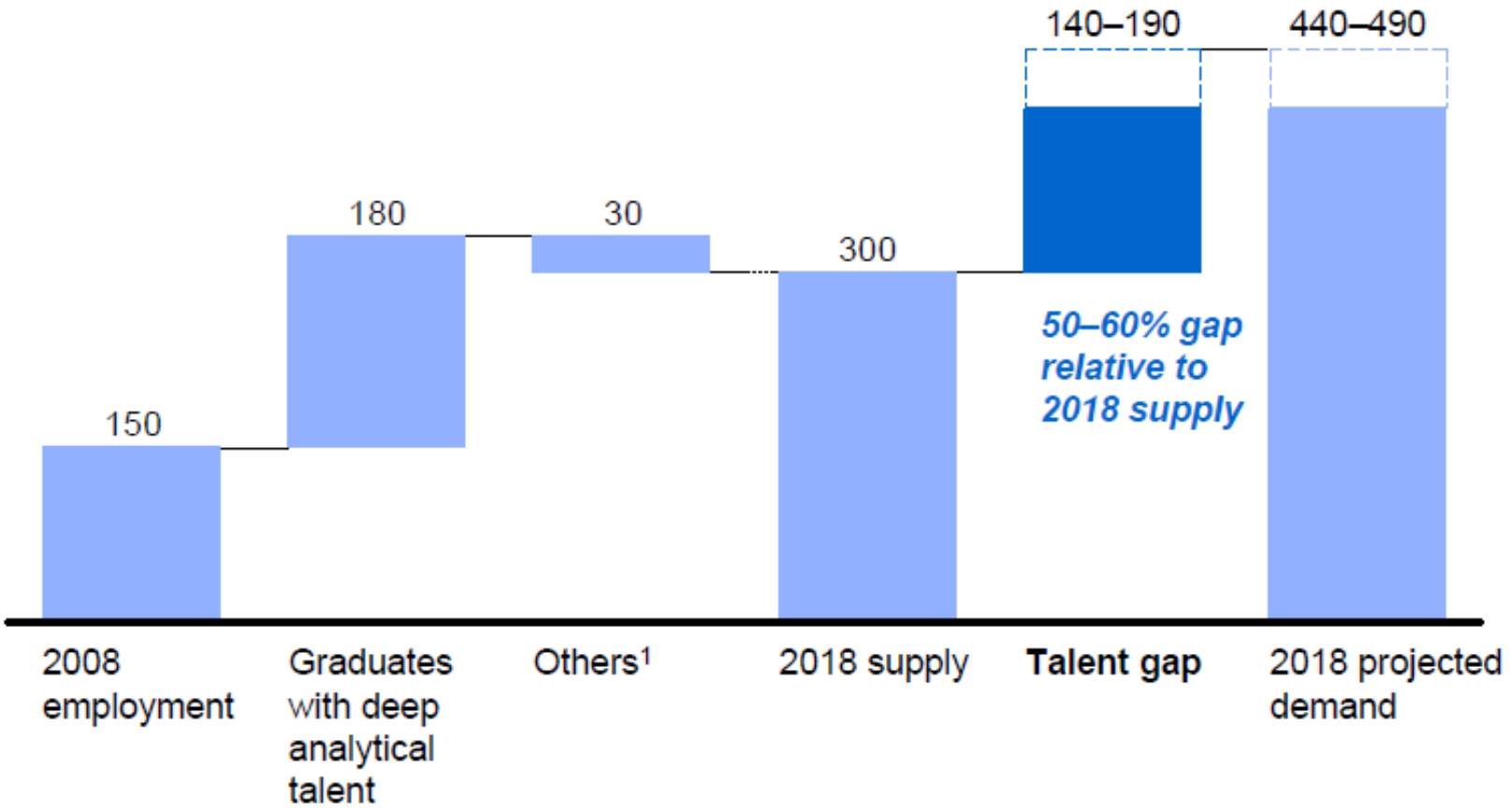
Data Mining ≈ Big Data ≈
Predictive Analytics ≈ Data Science

Good news: Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018

Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).

SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

What is Data Mining?

- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**
- Statisticians call it **Bonferroni's principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap

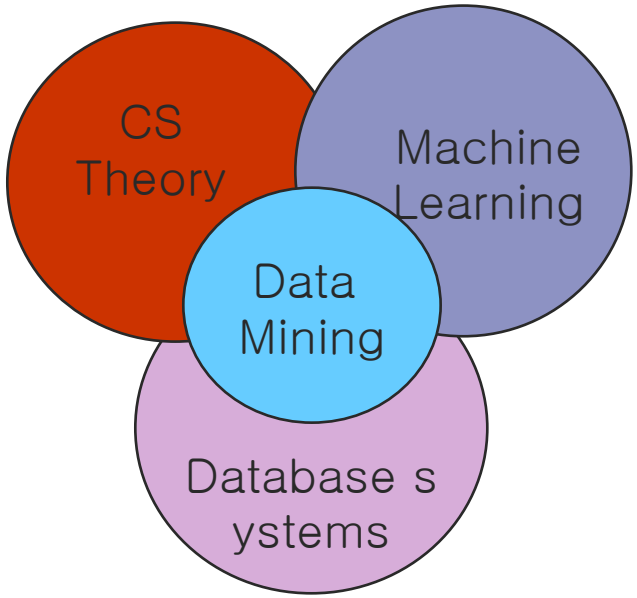
Meaningfulness of Analytic Answers

Example:

- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**
- **Expected number of "suspicious" pairs of people:**
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

Data Mining: Cultures

- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model



Data Mining Tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
  - Discover interesting relations between variables in large databases
- Sequential Pattern Discovery [Descriptive]
  - Find statistically relevant patterns between data examples where the values are delivered in a sequence
- Regression [Predictive]
  - Estimate the relationships among variables
- Deviation Detection [Predictive]
  - Build a model that describes the most significant changes in the data from previously measured or normative values

Classification

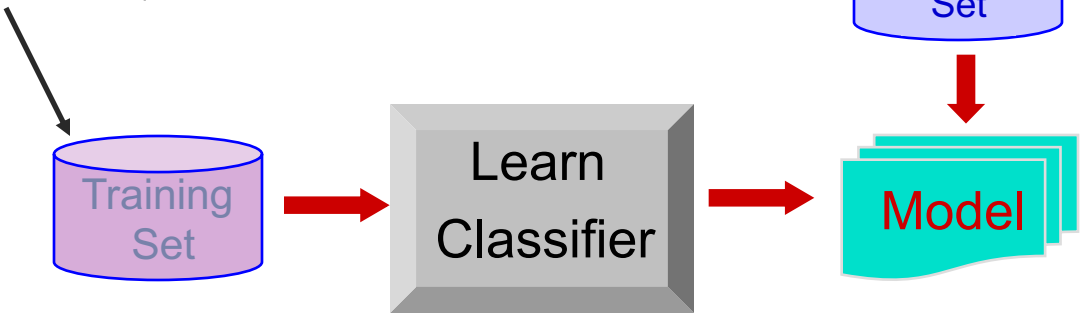
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

categorical categorical continuous class binary class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

- 마케팅 활용 예
  - 목표: 새로운 휴대폰 제품을 살 가능성이 있는 소비자를 대상으로 홍보 비용을 절감하고자 함.
  - 접근 방식
    - 이전에 소개되었던 유사한 제품에 대한 데이터를 사용
    - 이 문제에 있어서 class (분류 항목)는 binary (구매 또는 미구매).
    - 지리적 정보, 생활 패턴, 회사 업무, 연소득 등 고객에 대한 정보 수집
  - 이 정보를 입력 데이터로 활용하여 분류 모델을 learning

Classification: Application 2

- 금융 사기 적발에 적용
  - 목표: 신용카드 사용에 대한 금융 사기를 예측
  - 접근 방식:
    - 신용카드 소유주에 대한 정보와 사용 내역을 데이터로 활용
      - 언제 무엇을 사는지? 얼마나 자주 제 날짜에 결제하는가?
    - 과거 사용 내역에 대해 금융 사기 건수 였는지 정상 건수 였는지 표시 ➔ class
    - 신용 카드 사용 건수에 대한 class를 모델링
    - 이 모델을 이용하여 각 신용카드 계좌에 대한 사용 현황을 관찰하여 금융 사기 건을 예측

Classification: Application 3

- 고객 충성도 예측
  - 목표: 한 고객을 경쟁사에 뺏길 가능성에 대한 예측
  - 접근 방식:
    - 고객들의 과거와 현재 고객의 서비스 상담 행위에 대한 상세한 기록 정보를 활용
      - 얼마나 자주 전화로 문의를 했는지? 어디에서 주로 문의를 하는지? 하루 중 언제 주로 문의하는지? 고객의 신용 정보 상태, 고객의 혼인유무 등
    - 충성도가 높은 고객인가 아닌가를 표시 ➔ class
    - 고객 충성도에 대한 class를 모델링

Clustering

유사한 것끼리 묶는다

- 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 if attributes are continuous
    - The Euclidean distance between points is the length of the line segment connecting them
  - Other Problem-specific Measures

Clustering Visualization

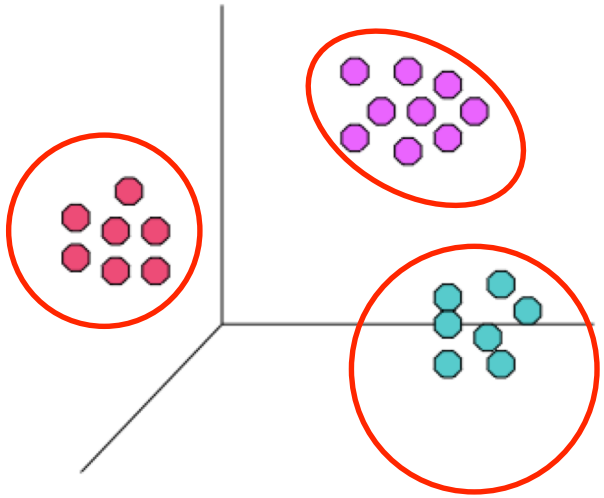
Euclidean Distance Based Clustering in 3-D space.

cluster 내의 거리 최소화

Intracluster distances are minimized

Intercluster distances are maximized

cluster 간 거리 최대화



Clustering: Application 1

- 시장 구분 (Market segmentation)
  - 목표: 특정 마케팅 전략을 가지고 접근해야할 고객들이 모여있는 구역으로 전체 시장을 나눔
  - 접근 방식
    - 고객에 대한 데이터 속성으로 사용할 수 있는 지리적 정보나 생활 방식 관련된정보를 수집
    - 유사한 고객의 클러스터를 찾음
    - 같은 클러스터에 있는 고객들의 구매 패턴의 유사성과 다른 클러스터에 있는 고객들의 구매 패턴 유사성을 관찰하여 클러스터링 결과를 검증함

Clustering: Application 2

- 신문 기사 클러스터링
  - 목표: 각종 일간지 및 주간지의 기사를 분류 (예: 정치, 경제, 사회, 생활/문화, IT/과학, 연예, 스포츠)
  - 접근 방식
    - 각 기사에서 자주 언급되는 용어를 찾아냄
    - 각 용어에 대한 빈도를 계산
    - 용어 빈도에 기반하여 유사성 측정

Association rule discovery

- 수집된 정보로부터 주어진 데이터 레코드에 대해 의존성 규칙 (dependency rules)을 생성
  - 어떤 특정 항목과 함께 나타나는 항목들을 예측해냄

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Association rule discovery 활용 예제

- 발견된 rule이 {치킨, … } --> {맥주} 라고 가정하면
  - 맥주는 결과 항목 (consequence) ➔ 맥주 판매량을 증가시키기 위해서 어떤 항목에 대한 마케팅 프로모션이 사용되어야 하는지를 결정
  - 치킨은 선행 항목 (the antecedent) ➔ 특정 가게에서 치킨 판매를 중단하였을때 판매량에 영향을 받는 항목이 무엇인지를 관찰함
  - 치킨이 선행 항목이고 맥주는 결과 항목 ➔ 맥주 판매량을 증가시키기 위해서 치킨과 함께 판매해야하는 다른 항목이 무엇인지 밝혀냄

Sequential pattern discovery

- 주어진 객체(objects)에 대해 여러 이벤트에서 시간상 순서에 있어서 강한 연관성을 보이는 규칙(rules)을 찾아냄.

- 소비자의 물품 구매 순서에 있어서,
  - 스포츠 용품점에서: (신발), (테니스 라켓, 테니스 공) ➔ 테니스 의류

Regression

value predict

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- 통계학과 뉴럴 네트워크에서 널리 이용됨
- 활용 예제
  - Predicting sales amounts of new product based on advertising expenditure.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior

Challenges in Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data