# Introduction to Data Mining

## Lecture #1: Course Introduction

U Kang
Seoul National University

**$600** to buy a disk drive that can store all of the world's music

**5 billion** mobile phones in use in 2010

**30 billion** pieces of content shared on Facebook every month

**40%** projected growth in global data generated per year vs. **5%** growth in global IT spending

**$5 million vs. $400**
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance

**235** terabytes data collected by the US Library of Congress by April 2011

**15 out of 17** sectors in the United States have more data stored per company than the US Library of Congress


# Data contain value and knowledge

# Data Mining

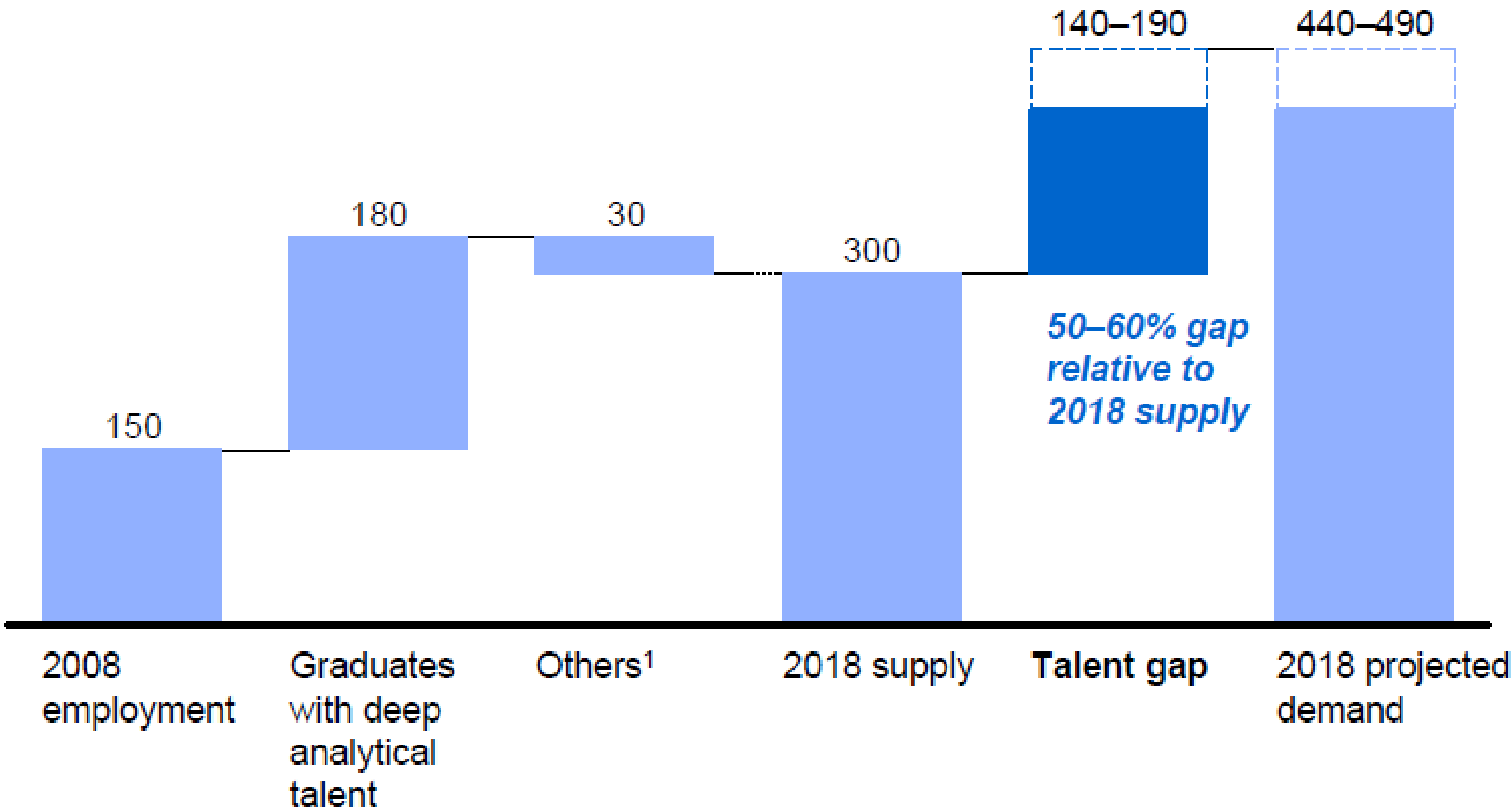
- But to extract the knowledge data need to be
  - Stored
  - Managed
  - And ANALYZED ← this class

**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?

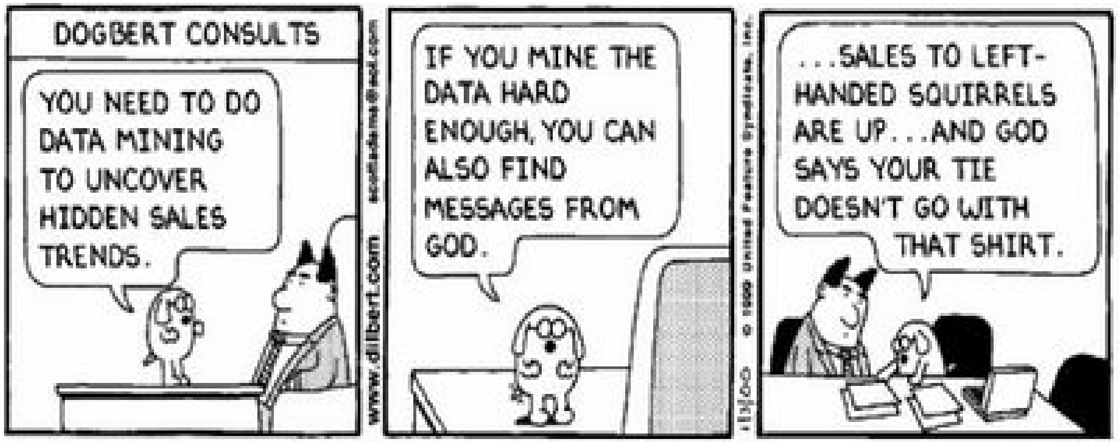
- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

# 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

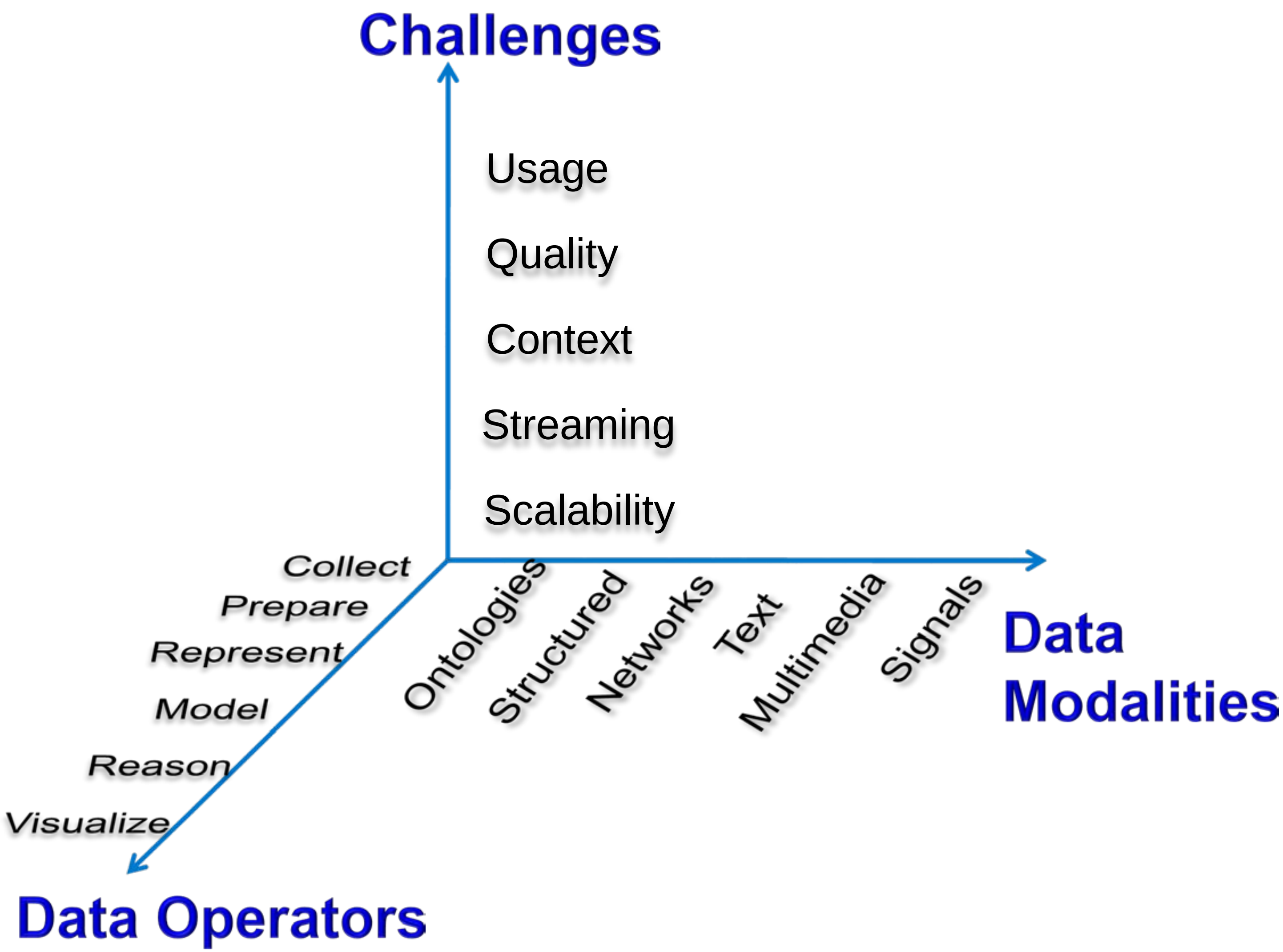
**Example:**

- 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 (details in next slide)
  - ... too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in so me more efficient way

# Meaningfulness of Analytic Answers

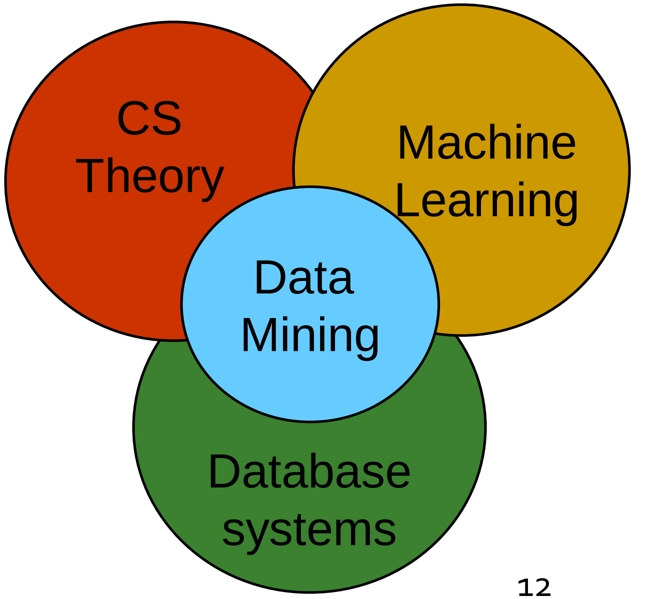
- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - $10^9$ people being tracked, 1,000 days, each person stays in a h otel 1% of time (1 day out of 100), hotels hold 100 people (so $10^5$ hotels)
- **Expected number of "suspicious" pairs of people:**
  - P(any two people both deciding to visit a hotel on any given d ay) = $10^{-4}$
  - P(any two people both deciding to visit the same hotel on any given day) = $10^{-4} \times 10^{-5} = 10^{-9}$
  - Useful approximation: $\binom{n}{2} \sim \frac{n^2}{2}$
  - Expected # of suspicious pairs of people = (number of pairs of people) x (number of pairs of days) x P(any two people both deciding to visit the same hotel on any given day)$^2$ ~ $(5 \times 10^{17}) \times (5 \times 10^5) \times 10^{-18} = 250{,}000$

# What matters when dealing with data?



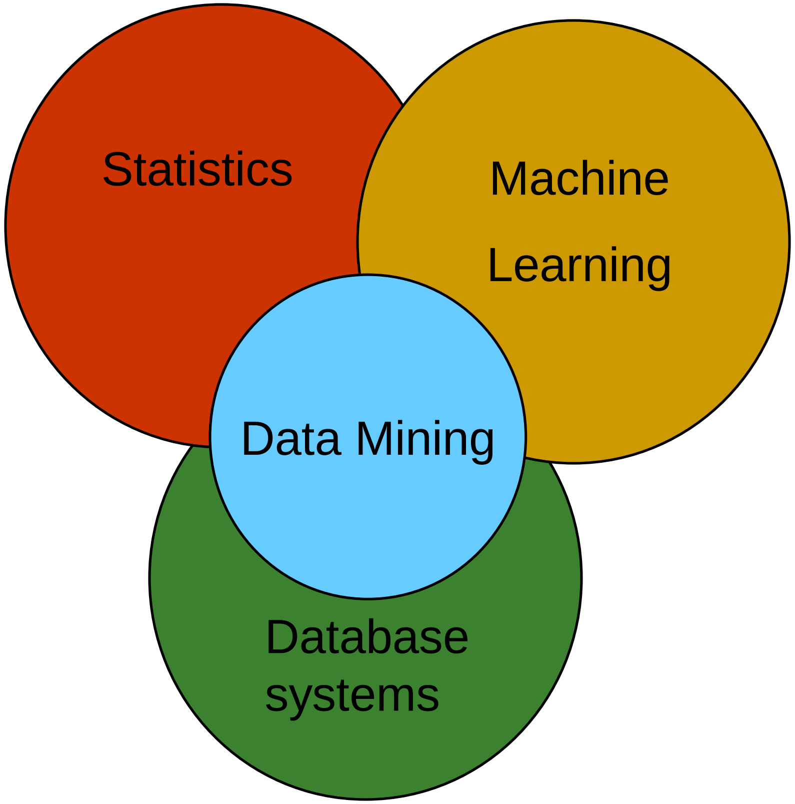
# Data Mining: Cultures

- **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
- **In this class we will do both!**



# This Class

- **This class overlaps with machine learning, statistics, artificial intelligence, databases but more stress on**
  - **Scalability** (big data)
  - **Algorithms**
  - **Computing architectures**
  - Automation for handling **large data**



# What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending

- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

# What will we learn?

- **We will learn to solve real-world problems:**
  - Recommender systems
  - Market Basket Analysis
  - Spam detection
  - Duplicate document detection
- **We will learn various "tools":**
  - Linear algebra (SVD, Rec. Sys., Communities)
  - Dynamic programming (frequent itemsets)
  - Hashing (LSH, Bloom filters)

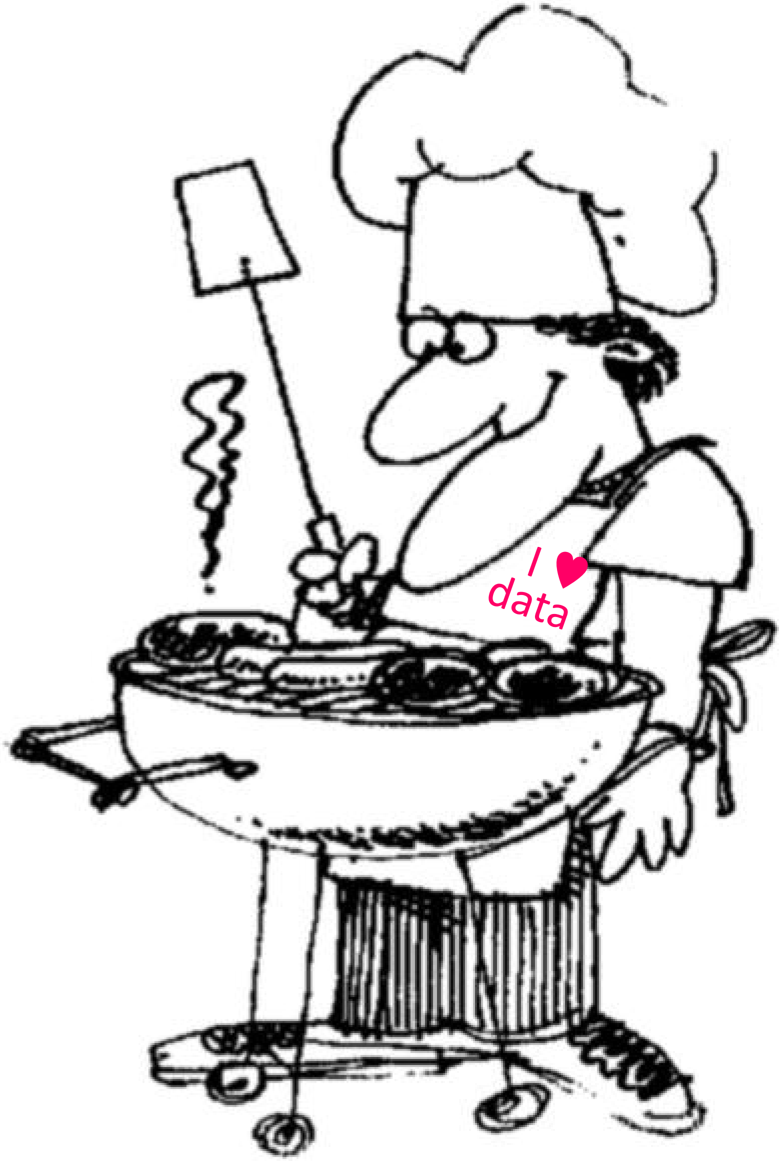
# How It All Fits Together

## High dim. data
- Locality sensitive hashing
- Clustering
- Dimensionality reduction

## Graph data
- PageRank, SimRank
- Community Detection
- Spam Detection

## Infinite data
- Filtering data streams
- Web advertising
- Queries on streams

## Apps
- Recommender systems
- Association Rules
- Duplicate document detection




# How do you want that data?

# Questions?