

# Motivation and Basics

Il-Chul Moon  
Dept. of Industrial and Systems Engineering  
KAIST


[icmoon@kaist.ac.kr](mailto:icmoon@kaist.ac.kr)

# Weekly Objectives

- ① • Motivate the study on
  - Machine learning, AI, Datamining....
  - Why? What?
  - Overview of the field
- ② • Short questions and answers on a story
  - What consists of machine learning?
  - MLE
  - MAP
- ③ • Some basics
  - Probability
  - Distribution
  - And some rules...

# MOTIVATION

# Keywords

- Many floating keywords
    - Data-mining, Knowledge discovery, Machine Learning, Artificial Intelligence...
  - Comes from territory, perspectives, types of problems, researchers, etc
  - We are going to focus on substance, not labeling.
  - I am just going to call it “Machine Learning”
    - You can call it whatever you want
- 
- AI in CS

# AI in CS

# Statistics

ARTIFICIAL INTELLIGENCE

# Database in CS

# Management

**MACHINE LEARNING**  
**DEPARTMENT**

# Industrial Engineering

■ ■ ■ ■ ■ ■

# Abundance of Data

- Data are being collected everywhere

Image Data

Surveillance Data

Text Data

Machine  
Logs

Social  
Networks

Trajectory Data

News  
Articles

Social  
Media

Disease Outbreak  
Data

10K Rep.

amazon

Purchase+Review  
Data

Vehicle

Time Series  
Data



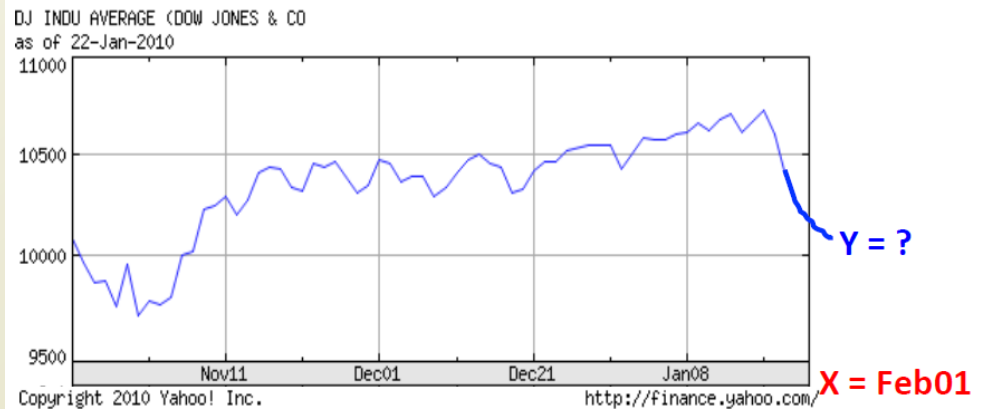
# Examples of Machine Learning Applications

- Machine Learning is everywhere...

## 1 Document Classification



## 2 Stock Market Prediction



## 3 Plate Num. Recognition



## 4 SNS Recommendation

**People You May Know**

- Alexander H. Levis, University Professor of Electrical, [Connect](#)
- Doo-Hwan Bae, -- [Connect](#)
- Paul Davidsson, Professor at Malmö University [Connect](#)

[See more »](#)

**Ads You May Be Interested In**

- INSEAD** The-CCP Plenum Deciphered Tsinghua-INSEAD EMBA Master Class on Mar. 7, Seoul.Discuss

## 5 Helicopter Control



# Spam Filtering and more



Table 2. Detailed evaluation results of SVMs with each representation scheme and varying training-set sizes. Macro-averaged MAE scores are provided with p-values, indicating the statistical significances of performance improvement over that of BF (using basic features alone). Numbers in bold font indicate the best method for each fixed training-set size. One star indicates the p-values in (0.01, 0.05]; two stars indicate the p-values equal or less than 1%.

	BF		BF+NC		BF+SI		BF+SIP		BF+SI+NC		BF+SI+NC+SIP	
# of tr	MAE	p-value	MAE	p-value	MAE	p-value	MAE	p-value	MAE	p-value	MAE	p-value
10	0.9666		0.9063	* 0.0382	0.8837	* 0.0106	0.8968	* 0.0311	0.9112	* 0.0211	<b>0.8827</b>	** 0.0087
20	0.9720		0.8969	0.0506	<b>0.8596</b>	* 0.0315	0.9095	* 0.0435	0.9071	0.0558	0.8659	* 0.0235

SVM?

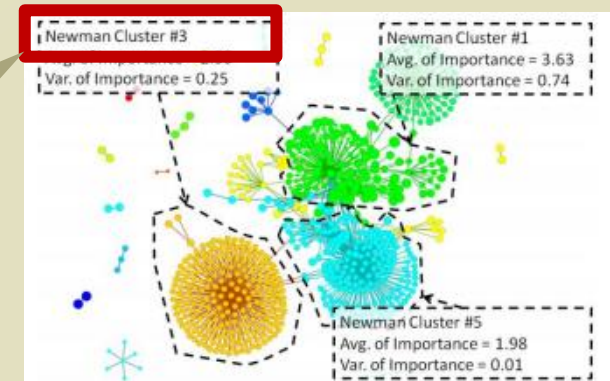
- Spam filter
- More?
  - Importance vs. Urgency
- How to predict an important email?
  - Social networks
  - Contents
- Shinjae Yoo, Yiming Yang, Frank Lin, and Il-Chul Moon, Mining Social Networks for Personalized Email Prioritization, ACM SIGKDD Conference, Paris, France, Jun, 28, 2009

Features

## 5.3 Features

The basic features are the tokens in the sections of *from*, *to*, *cc*, *title*, and *body text* in email messages. Let us use a  $v$ -dimensional vector to represent these features for each email message where  $v$  is the vocabulary size. We call it the *basic feature* (BF) sub-vector.

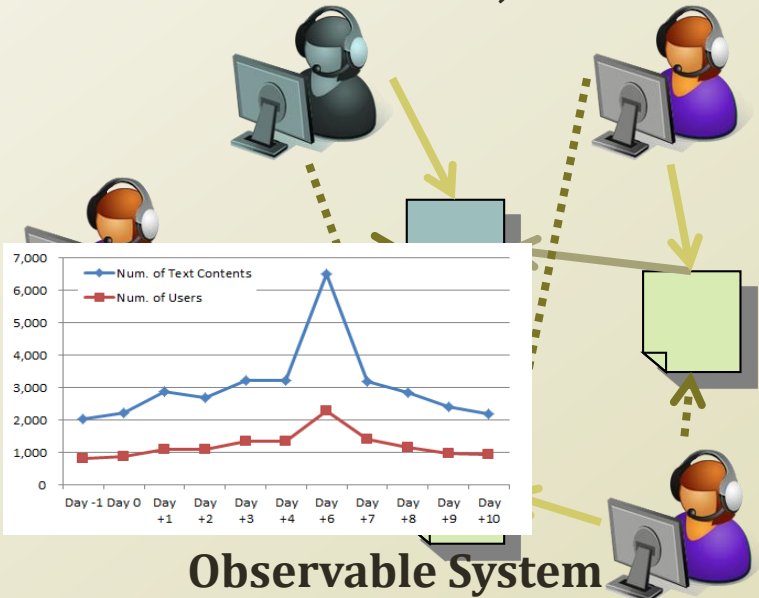
Clusters?  
Is this a machine learning technique?



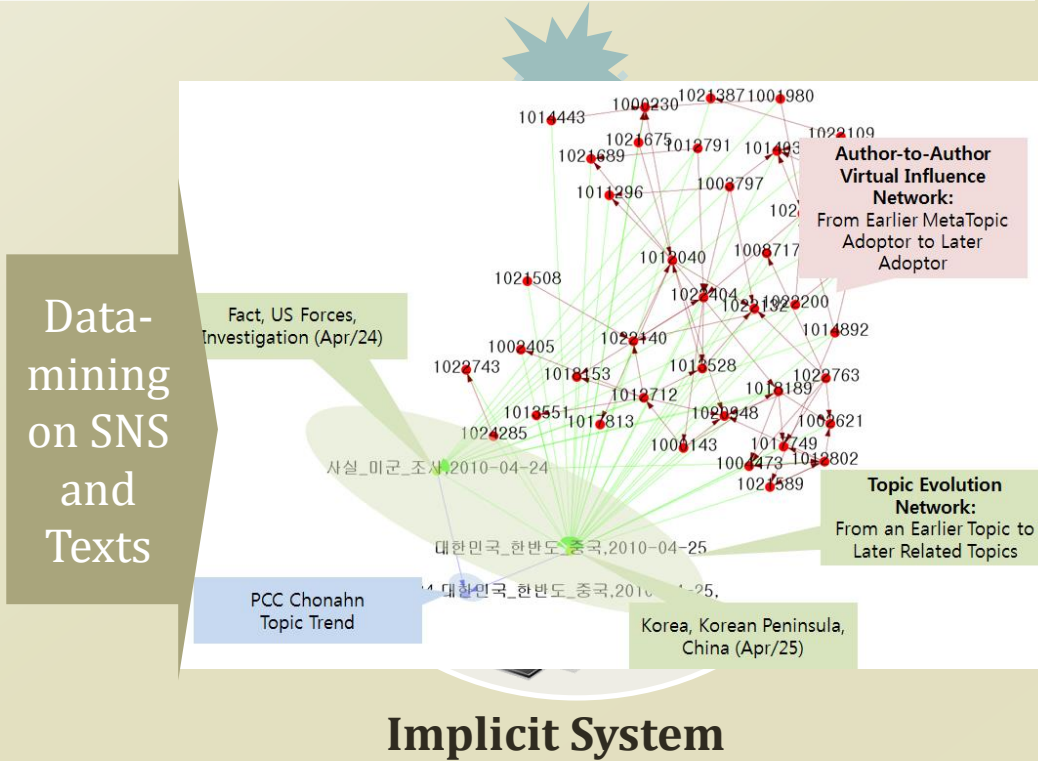
# Opinion Mining and more



**PCC Cheonan  
Sank on Mar 26, 2010**

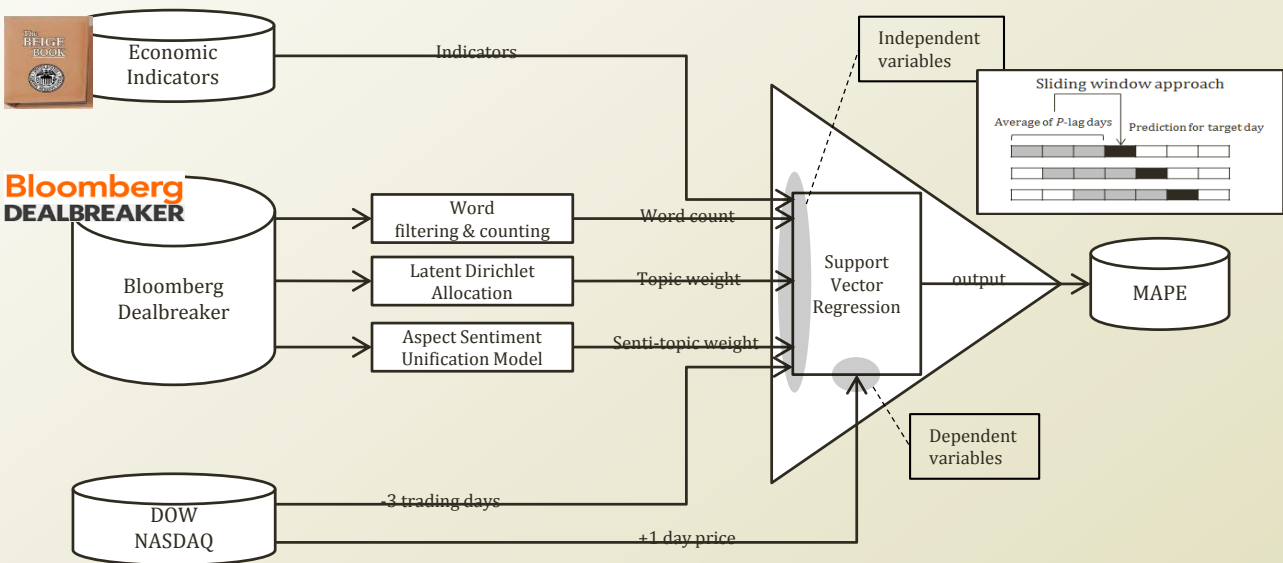


- Finding out consensus of the population
- Mining population's perception of the event
  - Mining key opinion buried in a data chunk
  - Estimating future polarity of the population
  - Strategy to maintain the unity of the population



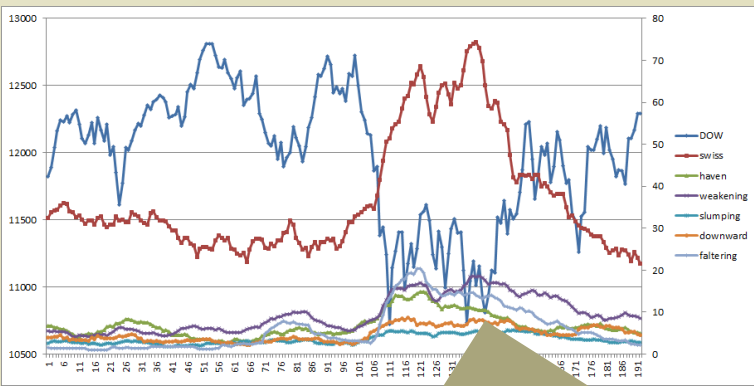
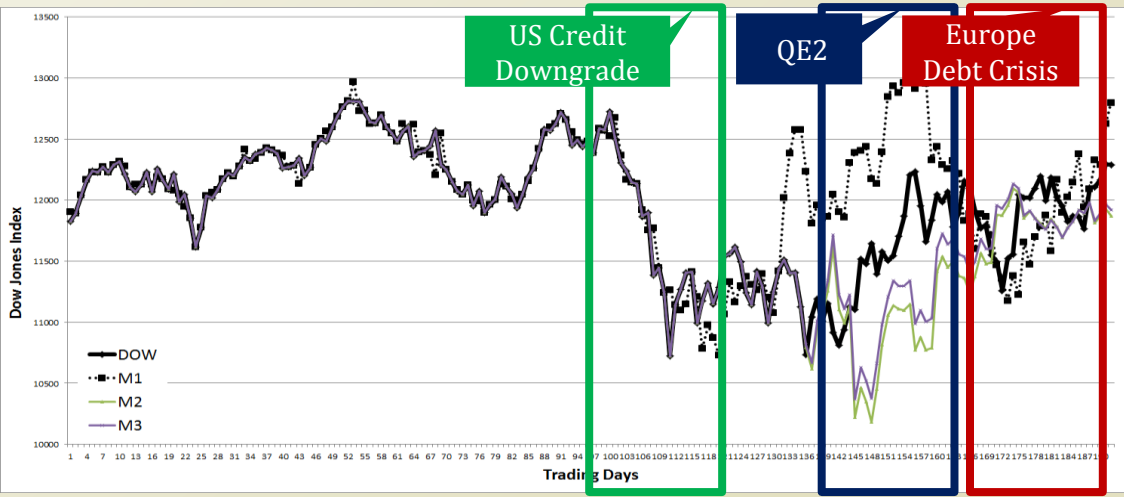


# Stock Market Prediction and more



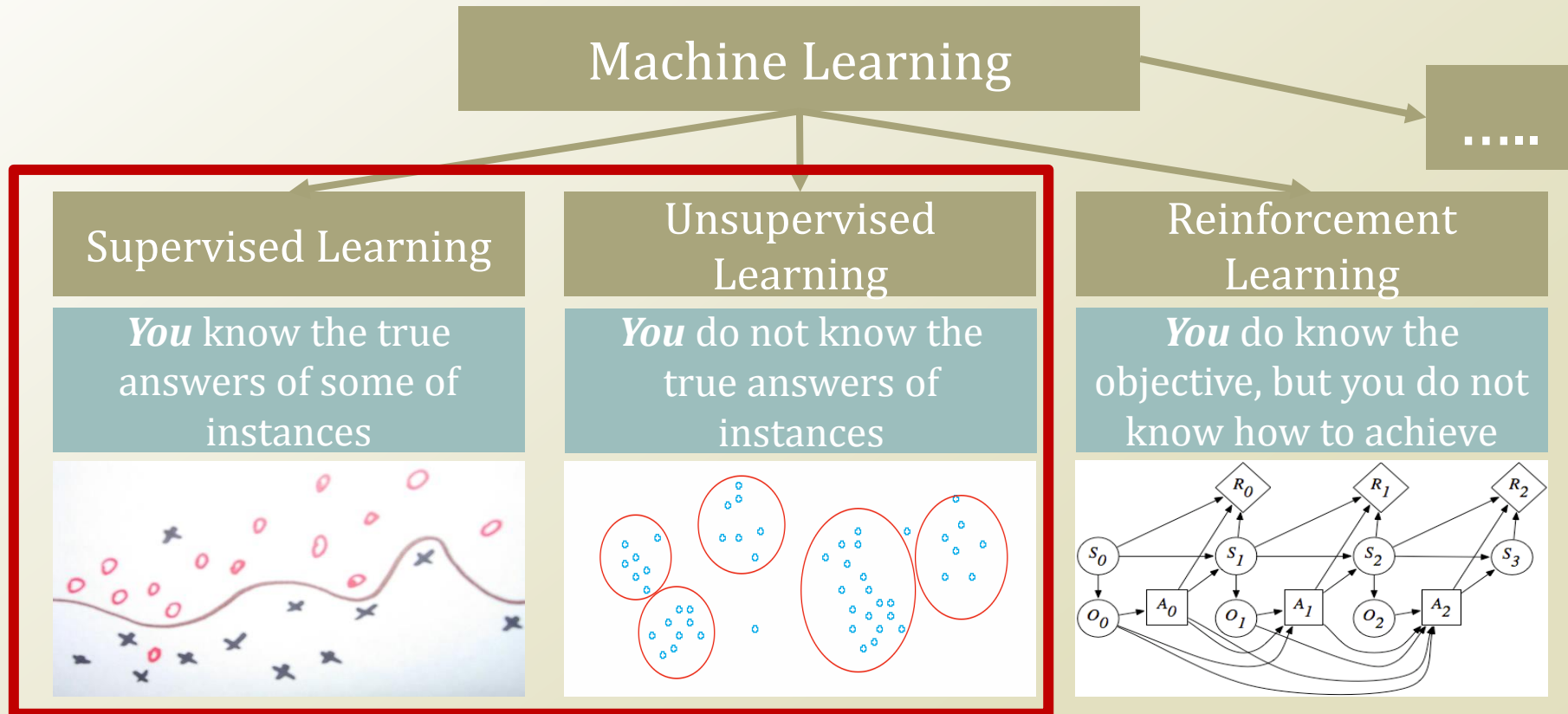
High Coefficients on Prediction

TopicWeight 26	TopicWeight 1	TopicWeight 7
-0.609	0.520	0.508
notes	obama	jun
moodys	republican	pence
swaps	republicans	na
treasuries	congress	swiss
versus	senate	chg
ratings	bill	francs
auction	barack	spa
default	lawmakers	fullyear
strategist	administration	nv
franc	democrats	dividend
twoyear	taxes	firstquarter
samp	white	ks
currencies	workers	paris
yen	democrat	reporting
swiss	obamas	tech



Heavy negative correlation between "swiss" and DJIA

# Types of Machine Learning



- **You** can
  - Machine learning ✓
  - Dataset provider ✓
  - Machine learning users ✓
  - etc ✓
- Various classifications by different professors ✓
  - Purpose, data types, etc
- Other learning classifications also exist ✓

# Supervised Learning

*You* know the true answers of some of instances



- **You know the true value, and you can provide examples of the true value.**
- Cases, such as
  - Spam filtering
  - Automatic grading
  - Automatic categorization
- Classification or Regression of
  - Hit or Miss: Something has **either disease or not**.
  - Ranking: Someone received **either A+, B, C, or F**.
  - Types: An article is **either positive or negative**.
  - Value prediction: The price of this artifact is **X**.
- Methodologies
  - Classification: estimating a discrete dependent value from observations
  - Regression: estimating a (continuous) dependent value from observations

# Unsupervised Learning

## Unsupervised Learning

*You* do not know the true answers of instances

- **You don't know the true value, and you cannot provide examples of the true value.**
- Cases, such as
  - Discovering clusters
  - Discovering latent factors
  - Discovering graph structures
- Clustering or filtering or completing of
  - Finding **the representative topic words from text data**
  - Finding **the latent image from facial data**
  - Completing the incomplete **matrix of product-review scores**
  - Filtering the **noise from the trajectory data**
- Methodologies
  - Clustering: estimating sets and affiliations of instances to the sets
  - Filtering: estimating underlying and fundamental signals from the mixture of signals and noises

