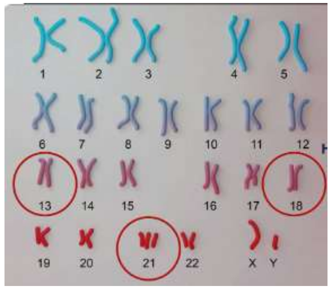
Data Mining
First Day of Class

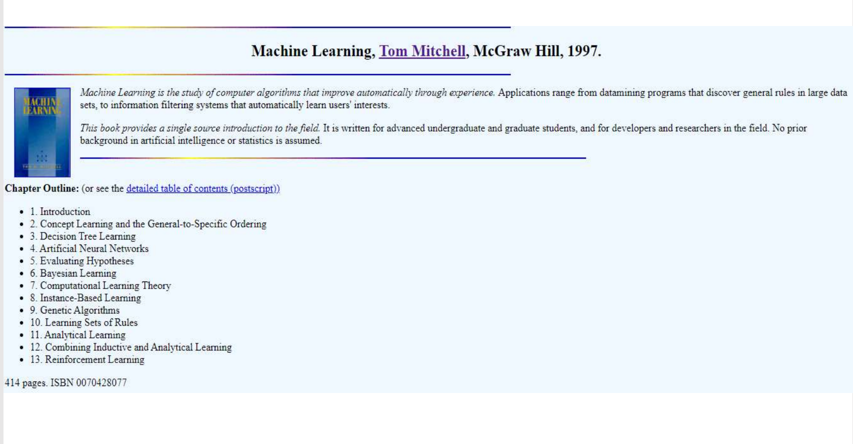
Associate Professor Yachai Limpiyakorn, Ph.D.

รศ.ดร. ญาใจ ลิ่มปิยะกรณ์

Data Mining:

- Mine vs. Mining
- a misnomer ?
- Knowledge Discovery in large Databases – KDD





Machine Learning, Tom Mitchell, McGraw Hill, 1997.

Machine Learning is the study of computer algorithms that improve automatically through experience. Applications range from datamining programs that discover general rules in large data sets, to information filtering systems that automatically learn users' interests.

This book provides a single source introduction to the field. It is written for advanced undergraduate and graduate students, and for developers and researchers in the field. No prior background in artificial intelligence or statistics is assumed.

Chapter Outline: (or see the detailed table of contents (postscript))

- 1. Introduction
- 2. Concept Learning and the General-to-Specific Ordering
- 3. Decision Tree Learning
- 4. Artificial Neural Networks
- 5. Evaluating Hypotheses
- 6. Bayesian Learning
- 7. Computational Learning Theory
- 8. Instance-Based Learning
- 9. Genetic Algorithms
- 10. Learning Sets of Rules
- 11. Analytical Learning
- 12. Combining Inductive and Analytical Learning
- 13. Reinforcement Learning

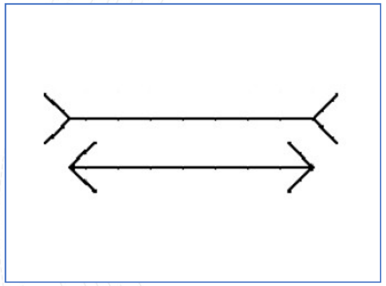
414 pages. ISBN 0070428077

Definition: concept learning

- The term *concept learning* is originated in psychology, where it refers to the human ability to learn categories for object and to *recognize* new instances of those categories. In machine learning, concept is more formally defined as "inferring a boolean-valued function from training examples of its inputs and outputs" (Mitchell, 1997).
- Synonyms: Categorization; **Classification** learning

ตัวอย่างที่1 สัญชาน/ การรับรู้ (Perception)

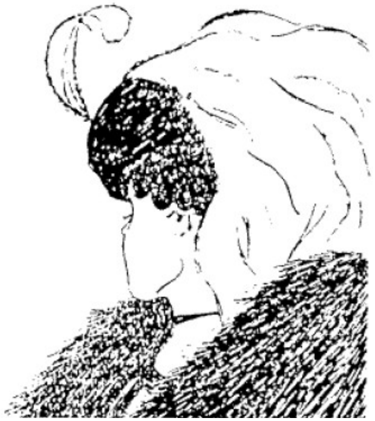
กระบวนการที่มนุษย์ติดต่อสื่อสารกับสิ่งแวดล้อมรอบๆตัว ทำการตีความ แล้วตอบสนองกลับไปอย่างเหมาะสม แต่ละคนอาจตีความในสิ่งแวดล้อมที่เหมือนกันออกไปในทางต่างๆกัน ขึ้นอยู่กับพื้นฐานทางจิตใจและความคิดของแต่ละคน



- **เส้นสองเส้นนี้ เส้นไหนยาวกว่ากัน?**

ตัวอย่างที่2 การรับรู้

- คุณเห็นใครในรูป



นี่น้องสาว
อย่าใส่มาสที่คางสิ
ใส่ให้ถูกต้อง
ยายก็ใส่
ถูกแล้วนี่

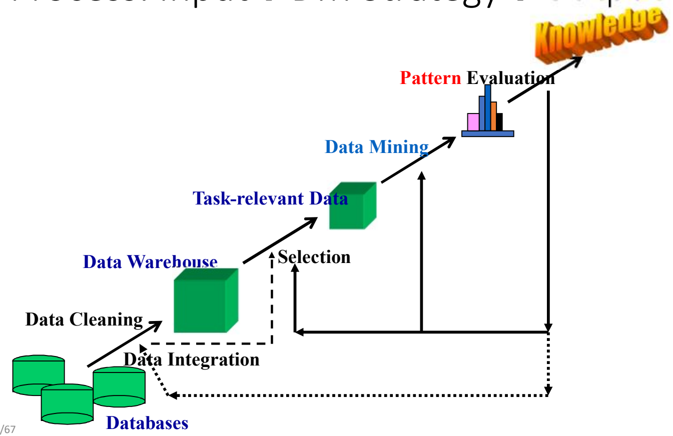
Concept (มโนทัศน์/ ความคิดรวบยอด/ แนวคิด)

- an abstract or generic idea generalized from particular instances. [Webster]
- Bearers of meaning, as opposed to agents of meaning
  - ➢ A single concept can be expressed by any number of languages
- Concepts cannot be visualized, unlike Perceptions
- Acquisition of concepts is studied in Machine Learning based on inductive reasoning

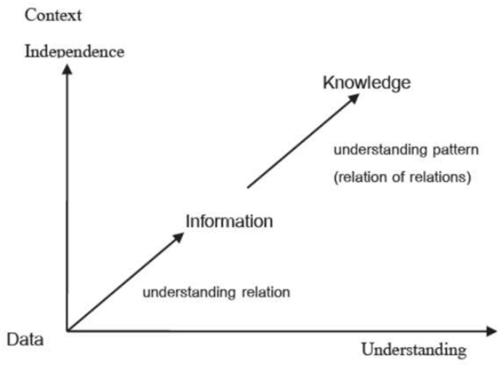
What is Data Mining?

- Given lots/ huge of data
- Knowledge discovery in databases (KDD)
- Int’l Conf. of Knowledge Discovery and Data Mining
  - ➢ KDD (https://dl.acm.org/conference/kdd)
- การเรียนรู้ของเครื่อง (Machine Learning: ML) เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence: AI) จัดเป็นการเรียนรู้เชิงอุปนัย (induction-based learning)
- Induction-based learning- process of forming a general **concept** definition by observing specific examples of the concept 2b learned

DM Process: Input→DM Strategy→Output

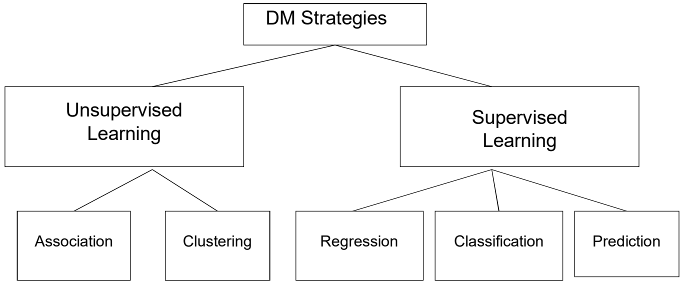


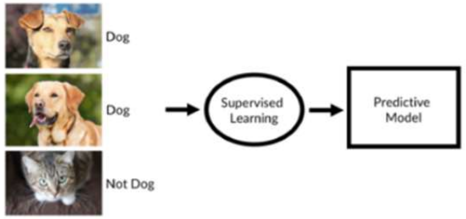
Data mining (knowledge discovery from data)



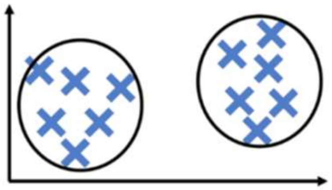
Data Mining vs. Query Processing

DM Strategy





Supervised learning



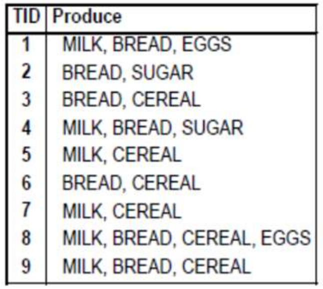
Clustering is a common form of unsupervised learning

Association Mining

- Transaction data
- Market basket analysis
- {Cheese, Milk} → Bread [sup=50%, conf=80%]
- Association rule:

" 50% of customers buy all these products together

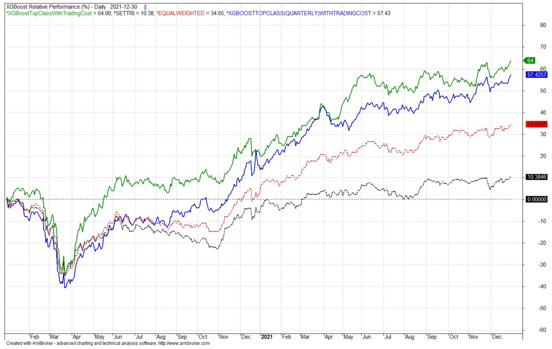
and 80% of customers who buy *cheese* and *milk* also buy *bread*"



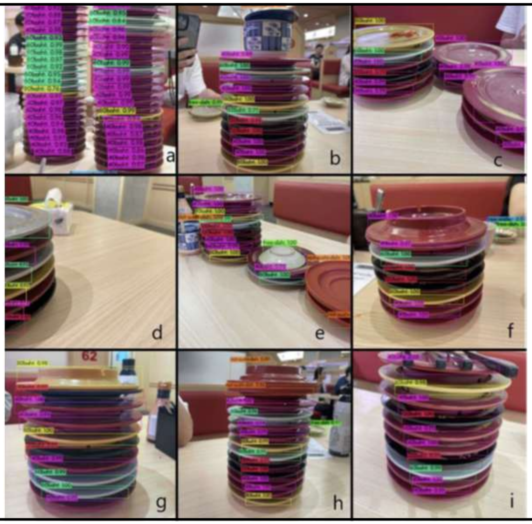
TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

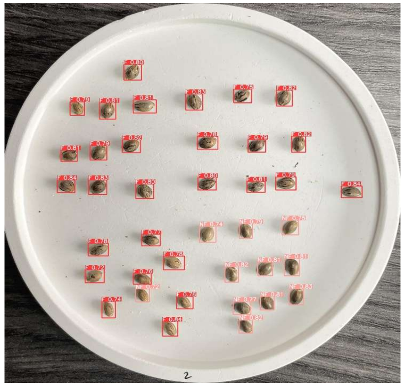
Applications

XGBoost-Based Multi-Factor Stock Selection Model for Rotational Trading



Cumulative return of XGBoost of Top-Class monthly (green line) and quarterly (blue line) stock selection with trading cost, SET TRI Index (black dash line) and Equal-Weighted (red dash line).





Semi-Automated Image Annotation for Cannabis Seed Gender Detection Model