Hokkaido Summer
School 2017

slide
http://www-ikn.ist.hokudai.ac.jp/~arim/HSI2017/arimura.pdf

H7: Introduction to Data Mining

- Can a computer learn from huge amount of data?

photo : morgueFile

Keywords:
Aritificial Intelligence
Machine Learning
Data Mining
Computer Game

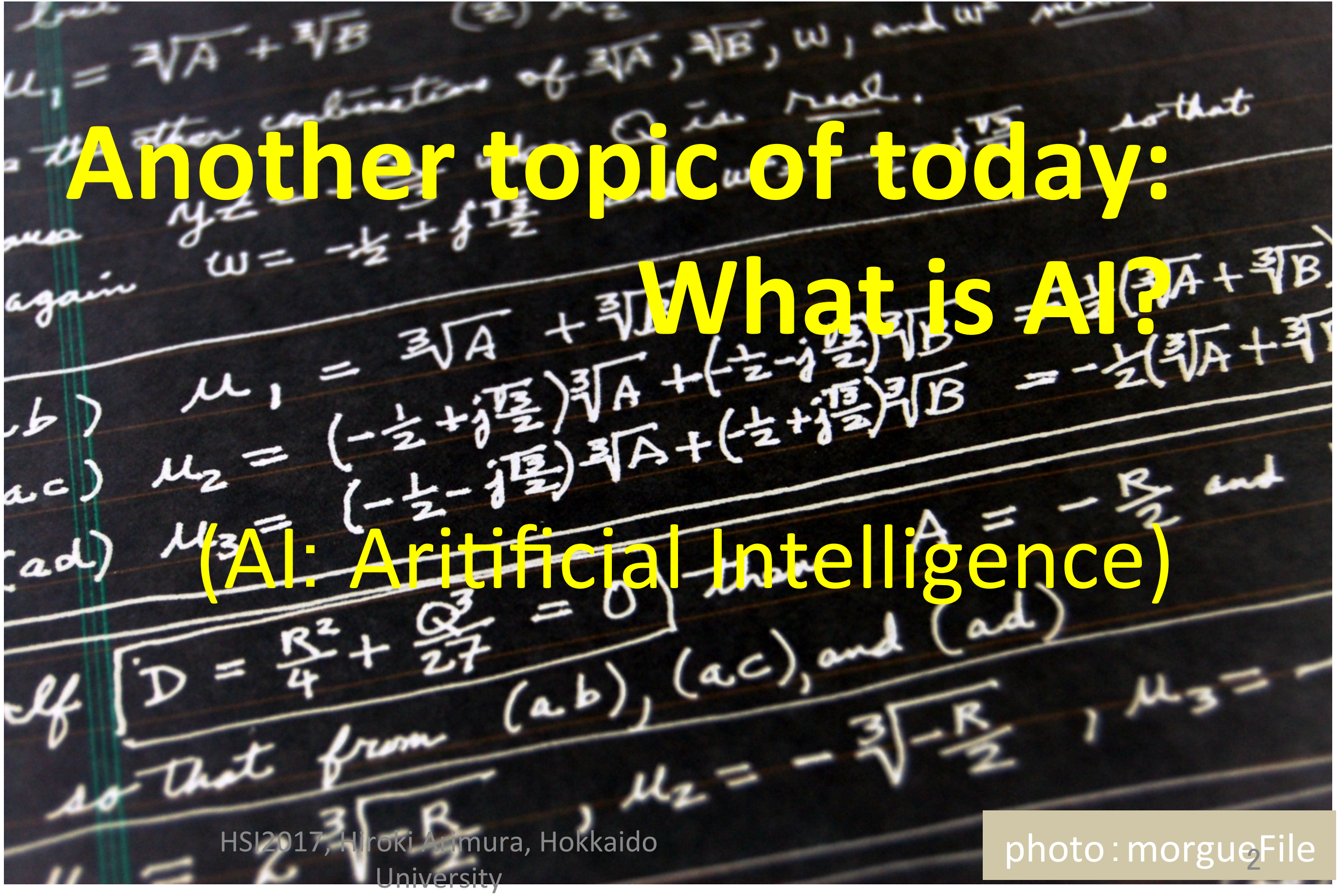
Hiroki Arimura
GSB & IST, Hokkaido Univerisity
IST bld. 7F, Rm.7-06
tel: 011-706-7680
arim@ist.hokudai.ac.jp

H7: Introduction to Data Mining

- Can a computer learn from huge amount of data?

Keywords:
Aritificial Intelligence
Machine Learning
Data Mining
Computer Game

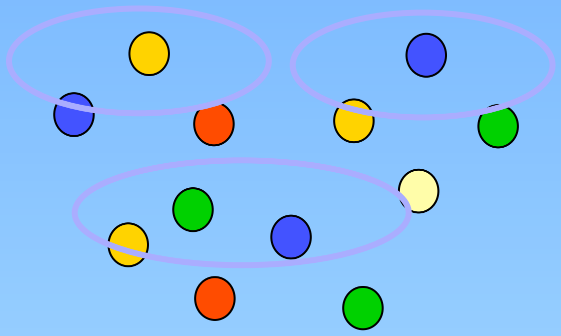
Hiroki Arimura
GSB & IST, Hokkaido Univerisity
IST bld. 7F, Rm.7-06
tel: 011-706-7680
arim@ist.hokudai.ac.jp



Another topic of today:
What is AI?

(AI: Aritificial Intelligence)

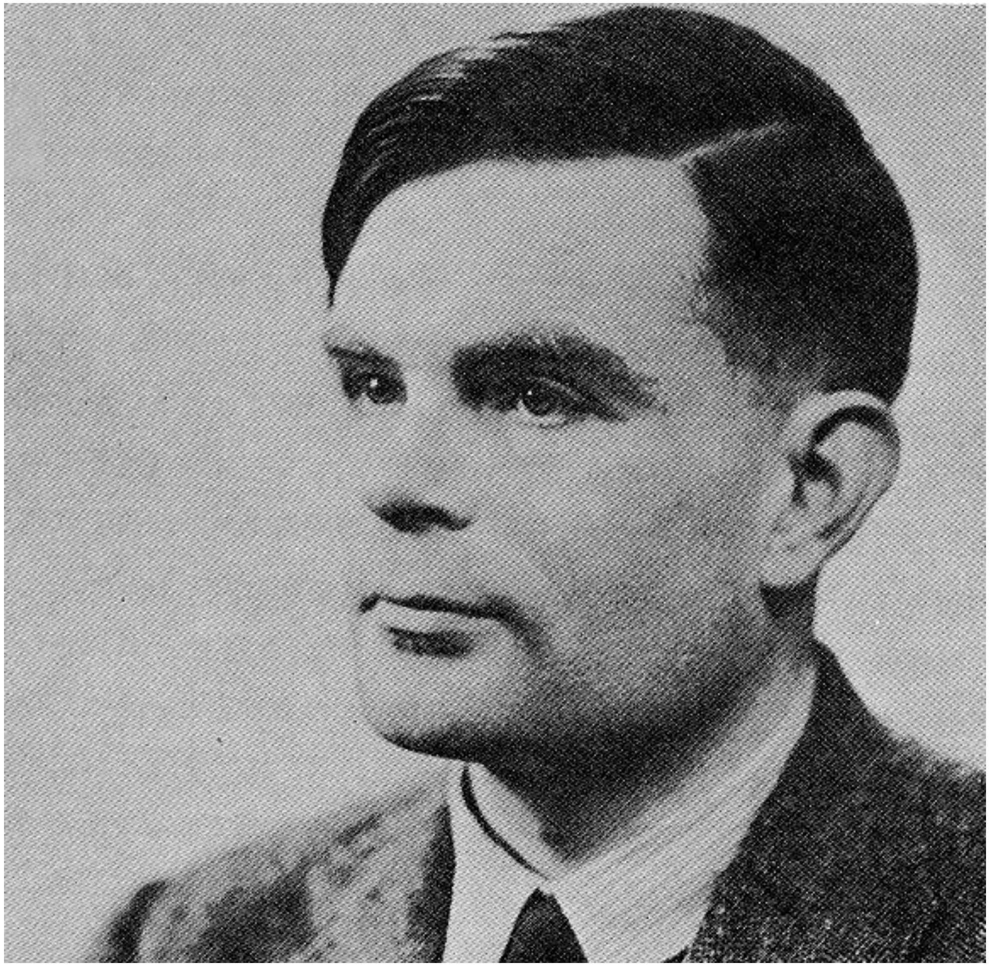
photo : morgueFile



DATA MINING: FROM PAST TO PRESENT

- WHAT IS DATA MINING?

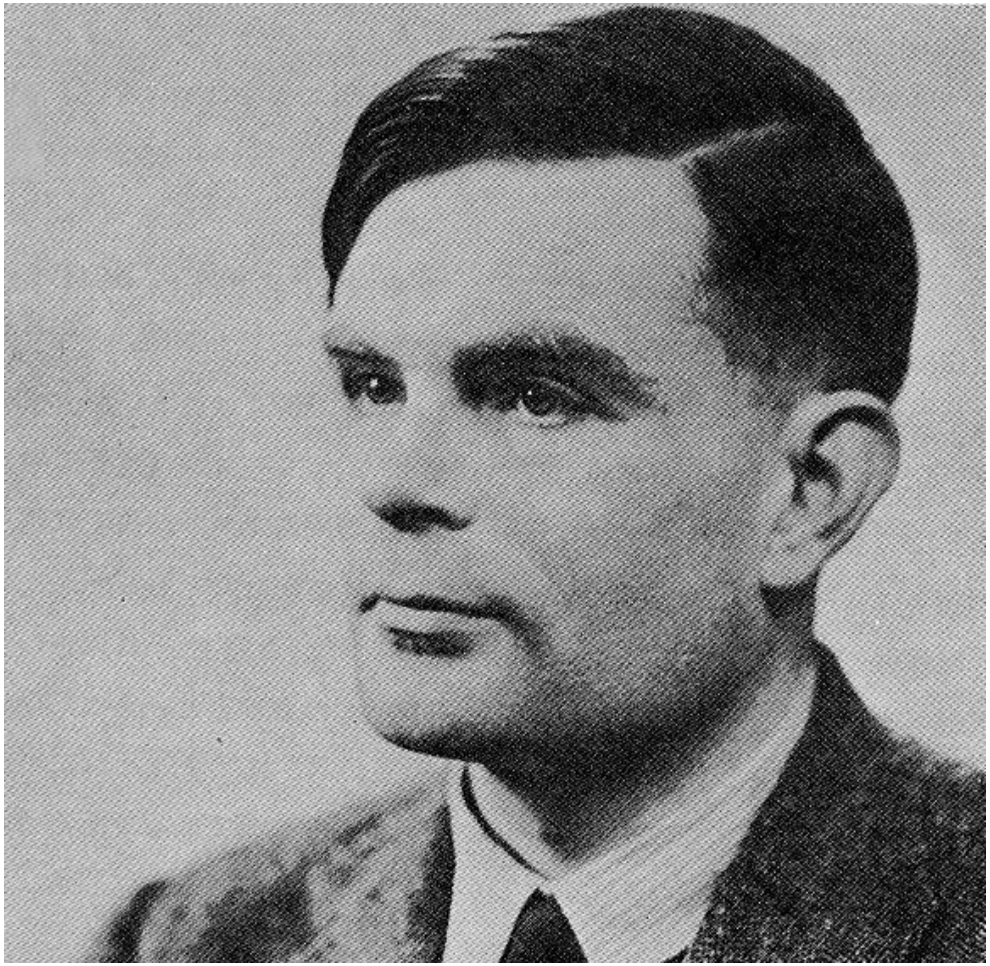
Quiz: Who is this?



Hint: 2012 was his 100 years' anniversary (born in 1912）

Answer: Dr. Alan Turing

— What did he think about?



（Alan M. Turing, 1912-1954, GB）

- One of the pioneers of computer science in early 20C.
- Known as a genius scientist in many areas.
- "Enigma" project
- Also famous in his "Turing test" in AI.
- in 1930s, he invented a mathematical model of computers, "Turing machine"

The First Digital Computers in 1940s

- The First Digital Computers in 1940 before W.W.II
  - Programmable
  - Software library
- ACE, Mark I (1946, GB)
  - Alan Turing joined
- EDSAC (1949, USA)
  - von Neumann が影響
  - Wilkes et al. (1967 Turing Award)
- ENIAC (1946, USA)
  - One of the first general purpose digital computers

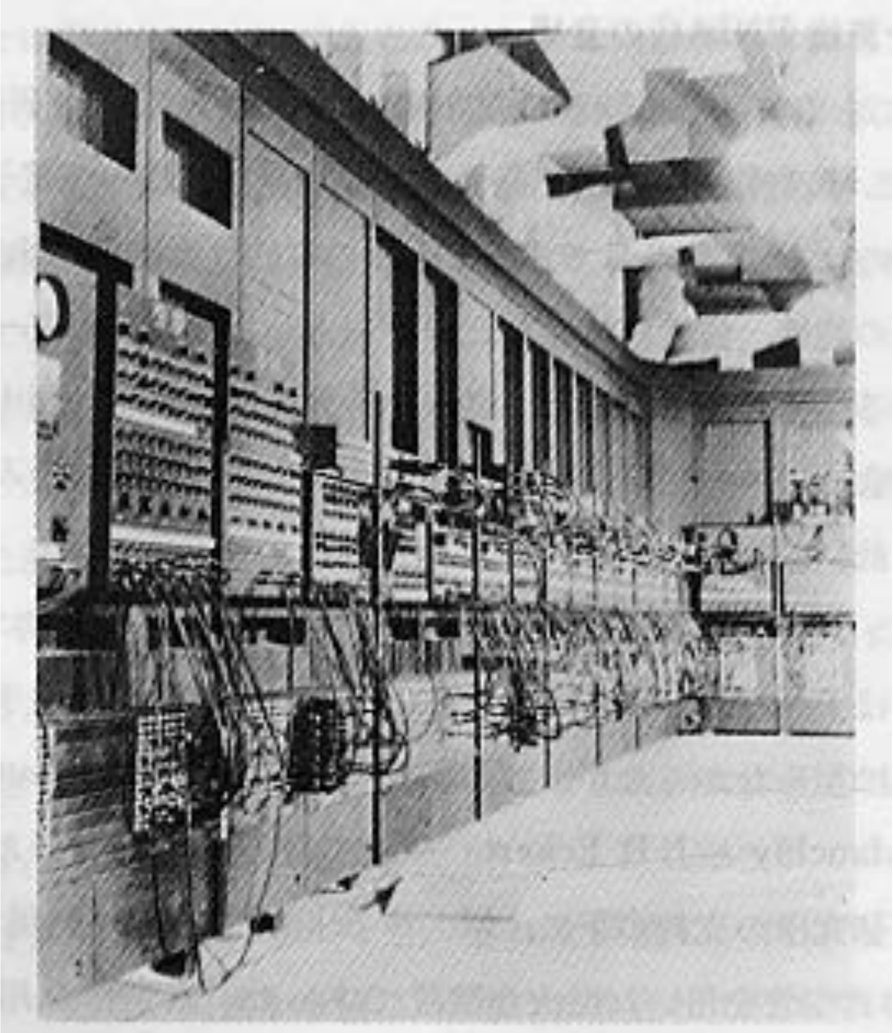
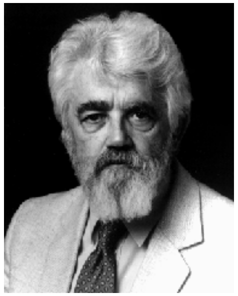


図1.10 ENIAC

Artificial Intelligence (AI)

- Studies on the possibility and limitation of implementing human's intelligent activities, such as *watching, listening, speaking, and thinking*.
- AI research started in 1950s right after the birth of digital computers



John McCarthy (1927-2011)
http://www.sis.pitt.edu

Marvin Minsky
(1927-2016)
https://en.wikipedia.org/wiki/Marvin_Minsky

- 1947, Alan Turing
  - Proposed the notion of AI
  - 1950, proposed "Turing test" for testing intelligence
- 1951, Marvin Minsky
  - Invented artificial neurons (with D. Edmonds)

- 1956年 John McCarthy
  - Proposed the term "Aritificial Intelligence (AI)" at Dartmouth Conferen in 1956.
  - 1958, developed the LISP programming language
- 1952-62, A. Samuel
  - Invented a computer program for playing "Checker" game.

Artificial Intelligence and Big Data

IBM's "Watson" System

- IBM Research (16 Feb. 2011)
- Won human masters in a TV quiz contest "Jeopardy!"
- Answers English questions by reading millions of books.
- Technology: AI, NLP, & Search

Google's Cloud Computing

- Computation based on data and information collected from all over the world!

Watson beat human masters in a popular TV program "Jeopardy!"

IBM Research
Watsonはこんな問題を解きました
ALTERNATE MEANINGS
4-LETTER WORD FOR THE IRON FITTING ON THE HOOF OF A HORSE OR A CARD-DEALING BOX IN A CASINO
馬の蹄(ひづめ)に付ける金具、またはカジノでカードを入れる箱を表す4文字の語
What is shoe?
DIALING FOR DIALECTS
WHILE MALTESE BORROWS MANY WORDS FROM ITALIAN, IT DEVELOPED FROM A DIALECT OF THIS SEMITIC LANGUAGE
マルタ語はイタリア語から多くの語彙を借りているが、それはこのセム語系言語の方言から発展した
What is Arabic?

From http://www-06.ibm.com/ibm/jp/lead/ideasfromibm/watson/

From geek.com: Google server firm
http://www.geek.com/articles/chips/up-next-for-google-enterprise-wars-2009078/

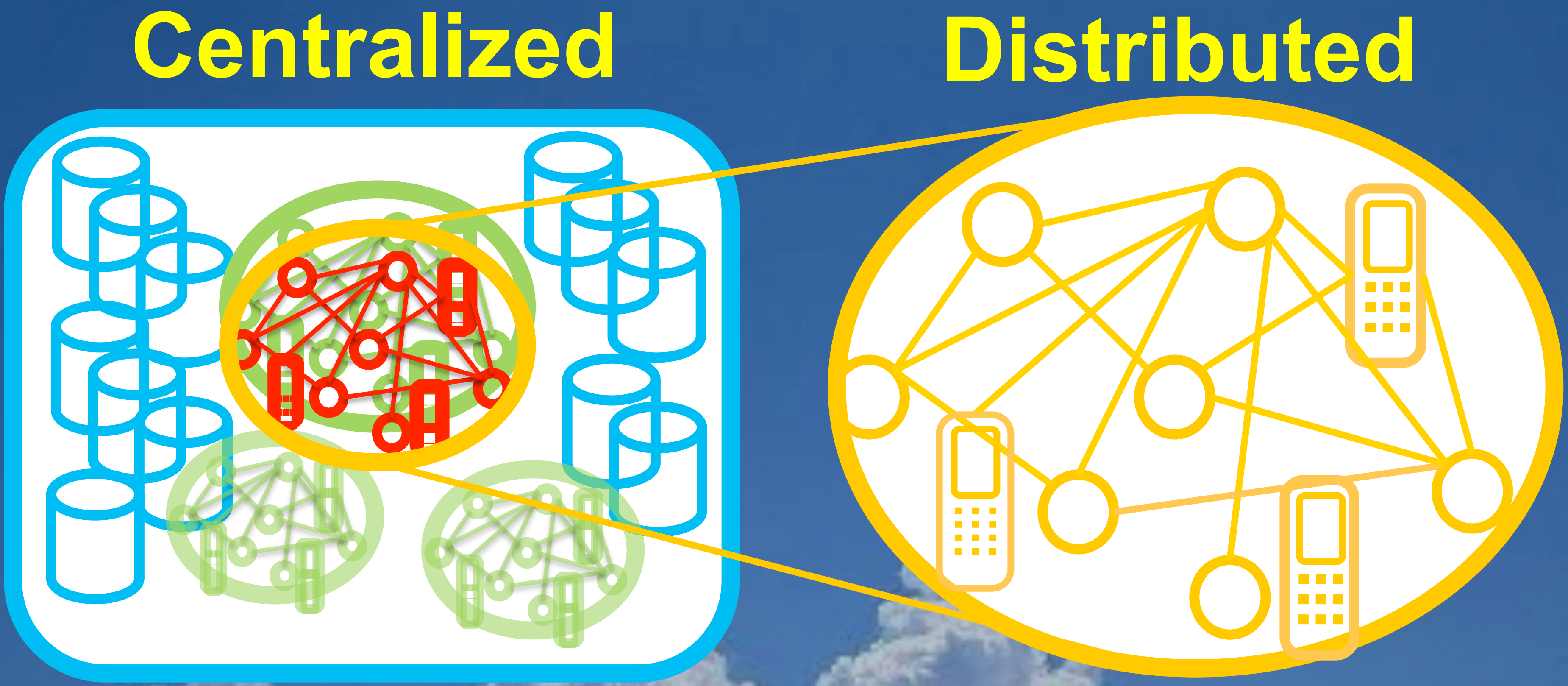
Quiz 2:

Consider the present information technology and its environment in the world.

Question: Is it

Centralized? or Distributed?

Answer: Either is OK: Two views of the world



- Centralized
- Huge amount of data
- Many CPUs
- Massive Computation

- Many devices (iphones etc.)
- Diverse activities of people
- Heterogeneous Time/Space
- Incomplete & complex data

Different Characteristics

Backgrounds

Data Mining

- Study on efficient “semi-automatic” methods for extracting “**interesting and useful**” patterns and rules from massive data sets
- Emerged in the mid 1990s.
  - Apriori algorithm [Agrawal, Srikant, VLDB1994]
- Potentially, a collection of existing studies.
  - But, emphasis on efficient computation for massive data
- Boundary of Machine Learning, Statistics, and Databases

Backgrounds

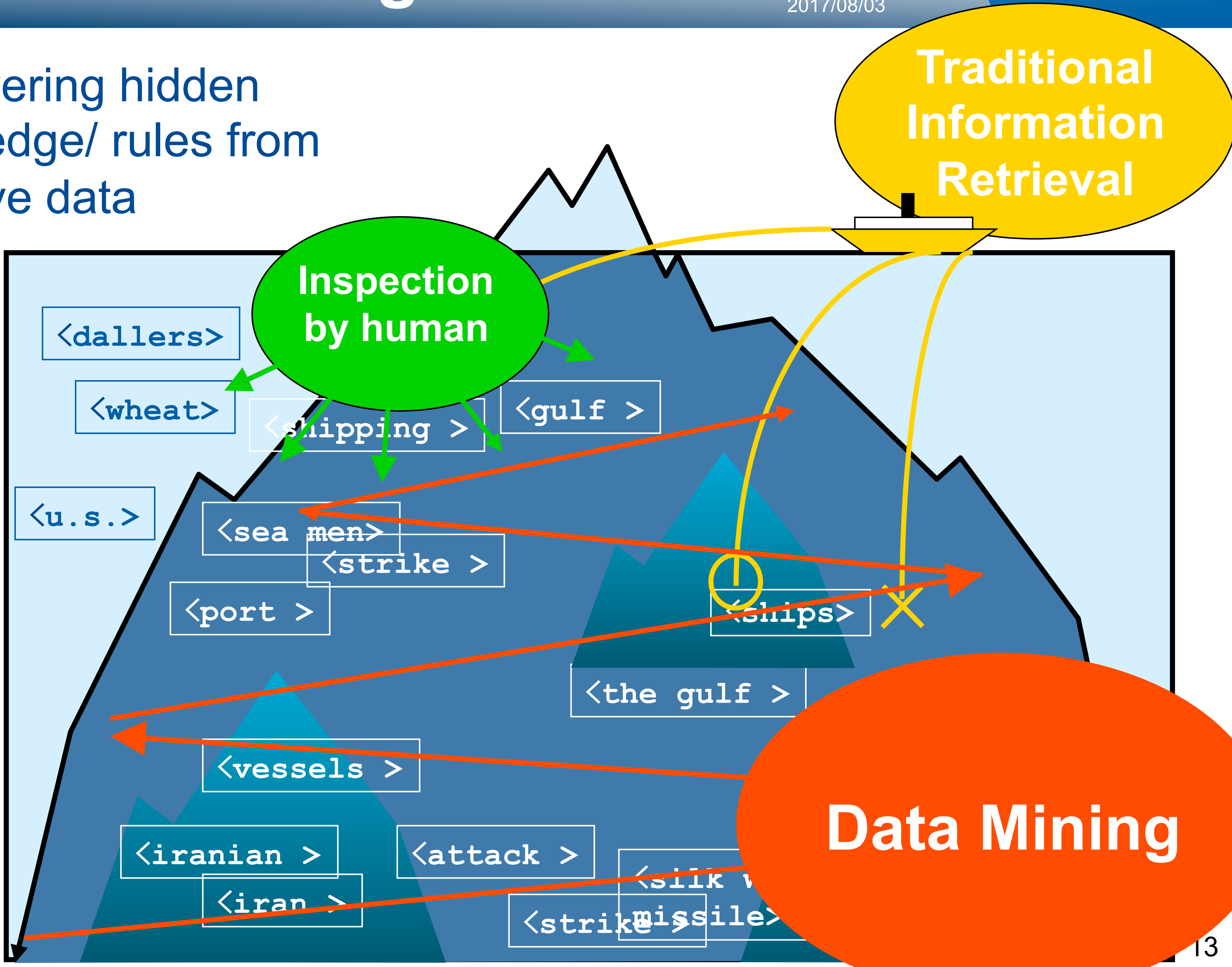
The whole process of Data Mining

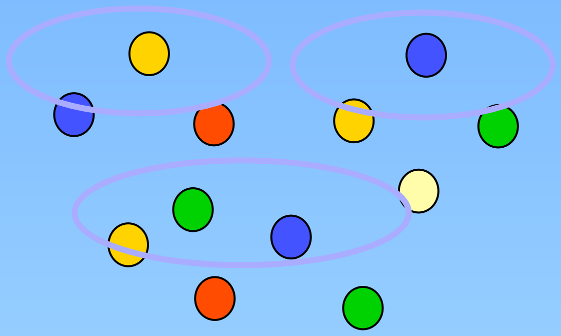
- 1.Understanding the domain of data
- 2.Preprocessing of data sets
- 3. Mining of patterns（Data Mining in narrow sense)
- 4.Analysis of discovered patterns
- 5.Use of the analyzed results

Data Mining

2017/08/03

Discovering hidden knowledge/ rules from massive data





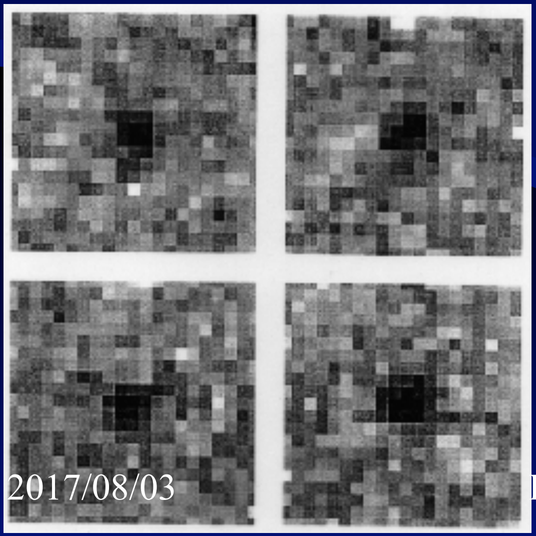
CASE STUDY: CAN A COMPUTER LEARN ASTORONOMY? – SUPERVISED LEARNING

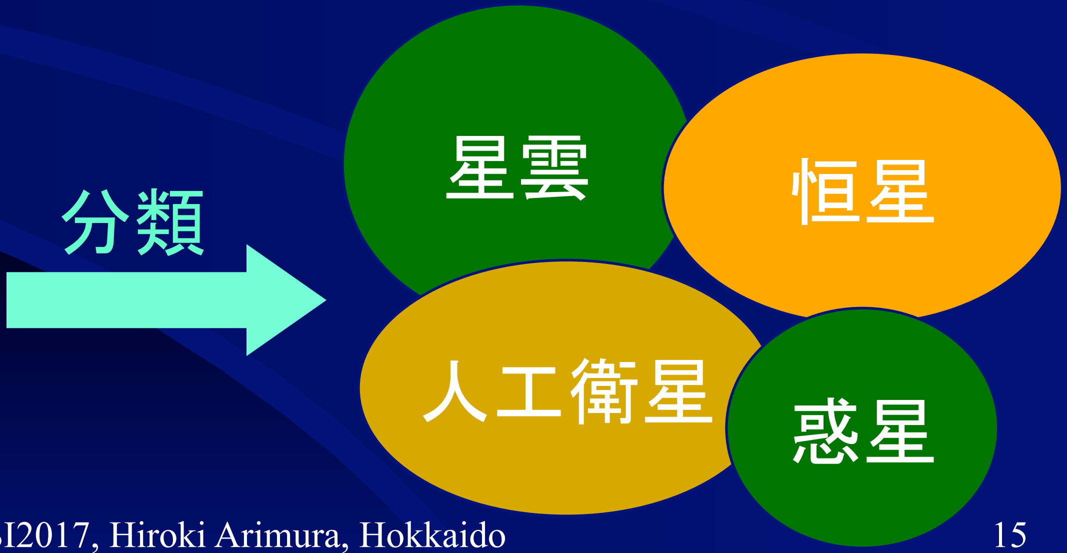
Can a computer learn astoronomy?

We can make automatic classification of photo images of stars!

SKICAT Project

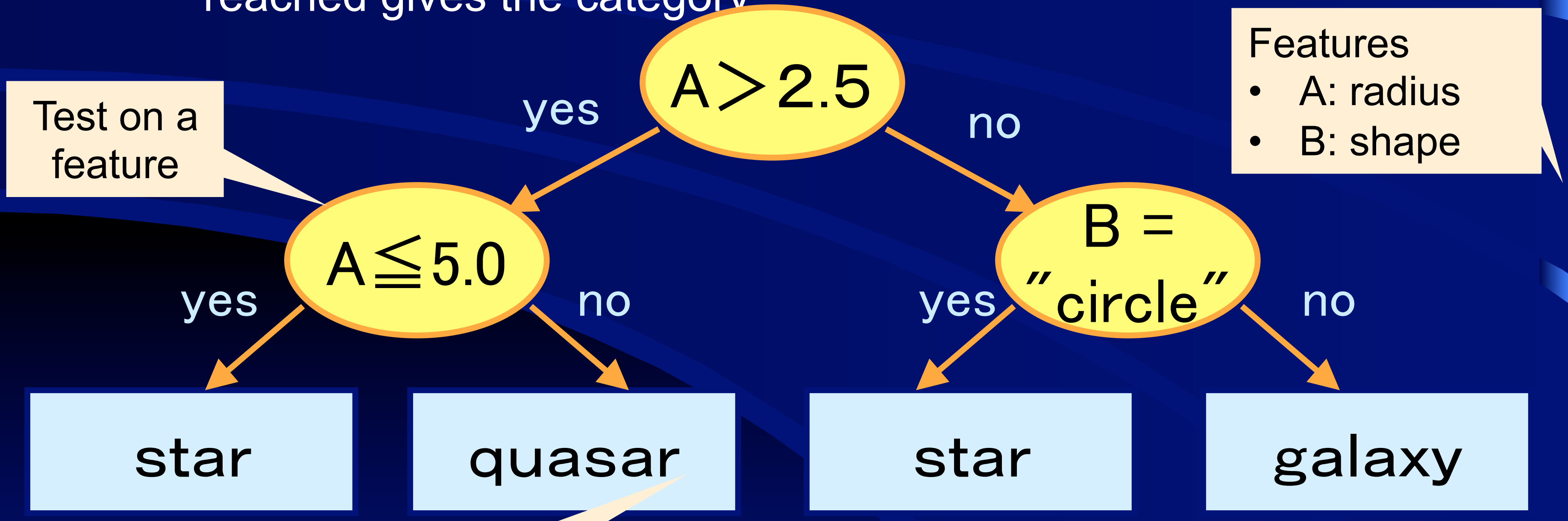
- (SKy Image Cataloging and Analysis Tool) in 1990s, NASA JPL, USA
- One of the earliest attempt of large-scale data mining
- Learning of a computer probram (*"a classifier"*)
  - to automatically classify star imagesinto categories of stars.
  - by using machine learning based on 1700 training examples





Automatic Classification by Machine Learning

- **We use a class of rules, called Decision trees.**
  - ▼DT classifies data into categories based on its characteristics (*"features"*)
  - ▼**Classification:** Given a data, traverse a path from the root to some leaf according to the results of tests. The label of the leaf reached gives the category



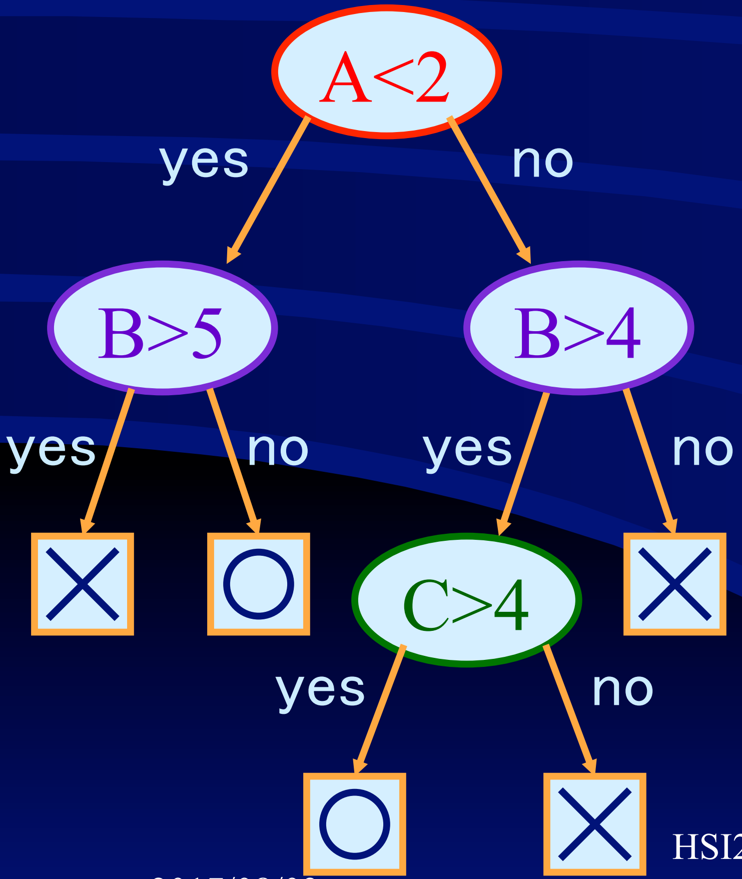
Test on a feature

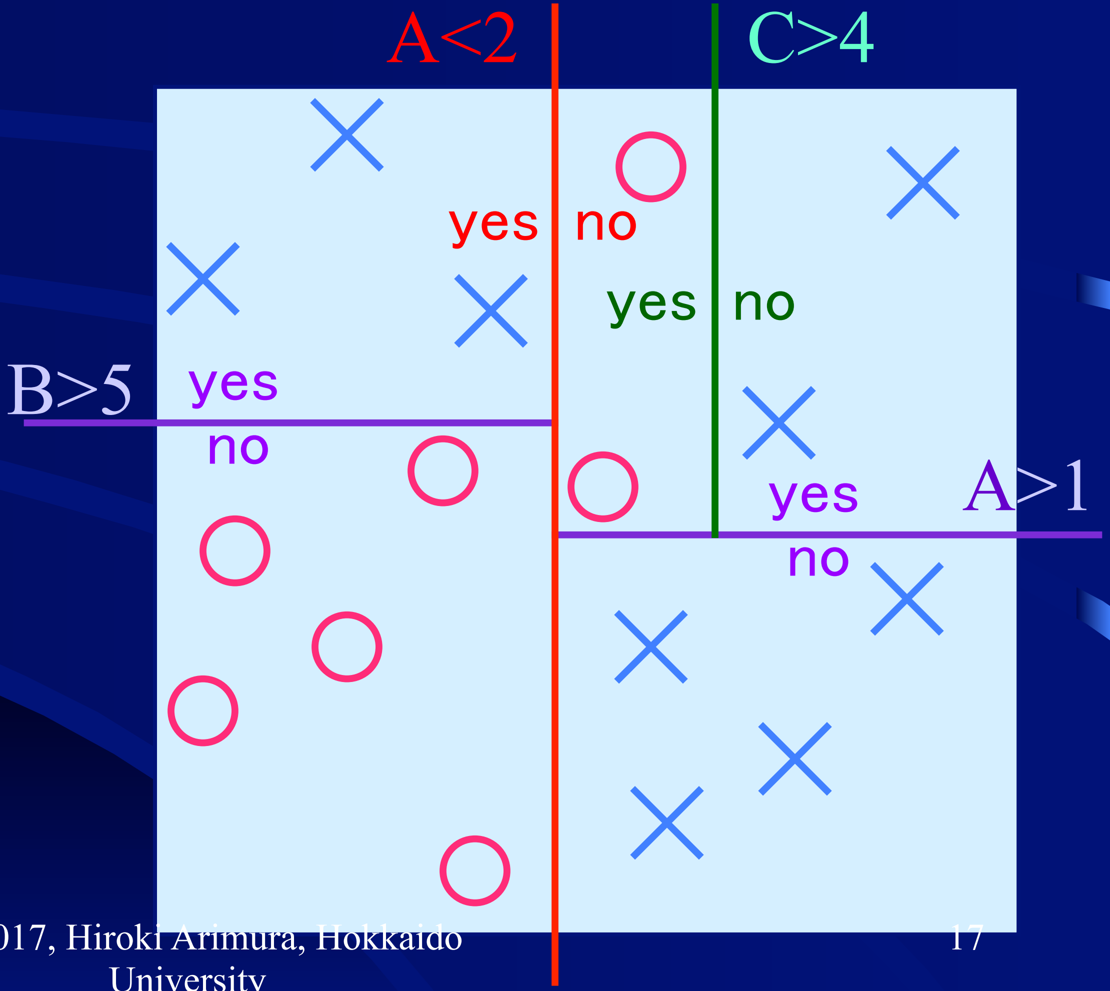
Features
- A: radius
- B: shape

category/class

Learning Algorithm for Decision Trees

Recursively constructs such a tree that minimizes the classification error from given ○ positive and × negative examples





Can we learn customer preference from purchase data?

- A computer can analyze the contents of baskets for one million customers in several tens of minutes.
- Which items are bought together in a basket?
- Apriori Algorithm（in 1990s by IBM Almaden）
- One of the root of data mining research

Advanced Machine Learning Algorithms

Boosting [Freund, Shapire 1996]

– Prediction by aggregation of many learning algorithms

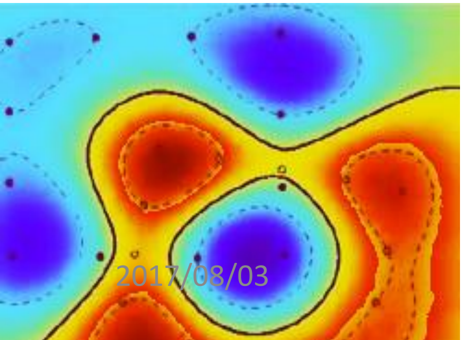
SVM [Vapnik 1996]

- Margin maximization and kernel methods

Deep Learning [Hinton et al.]

- Neural nets with many layers of different functionalities

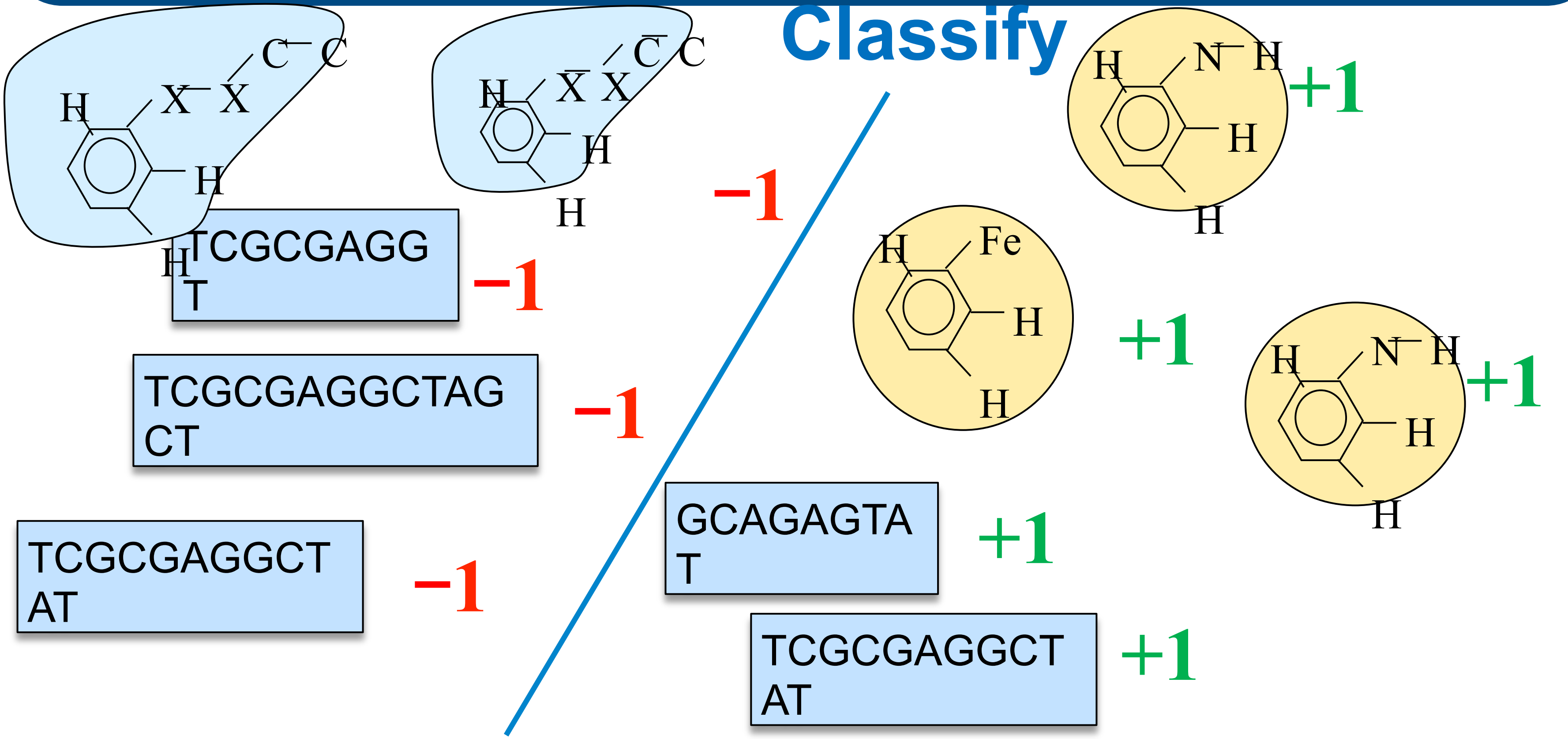
- All the above algorithms are kinds of neural networks形）
- Demonstrated their high performance in theory and practice



- V. Vapnik, Statistical Learning Theory, Wiley, 1998. (SVN)
- Y. Freund and R. E. Schapire, A decisiontheoretic generalization of on-line learning and an application to boosting, JCSS, 55, 119-139, 1997. (AdaBoost)

Classification: SVM

- We can learn rules from complex data such as genome sequences and chemical compounds once appropriate features are designed
- Applications: Medical diagnosis, pharmacy design



Classify

TCGCGAGG
T
−1

TCGCGAGGCTAG
CT
−1

TCGCGAGGCT
AT
−1

GCAGAGTA
T
+1

TCGCGAGGCT
AT
+1

Map of classic & modern DM/ML methods

Classic methods

DM = data mining, ML = machine learning

A. Supervised Learning

Learning an unknown classification rule from labeled data sets

- SVM [Vapnik '96],
- Boosting [Shapire & Kearns '96]
- C4.5 [Quinlan '96]

B. Custering

Grouping a given unlabeled data set into subgroups of similar objects (*clusters*)

- 大規模・不完全なデータからの高速クラスタリング
- K-means, CLARANS, DBSCAN

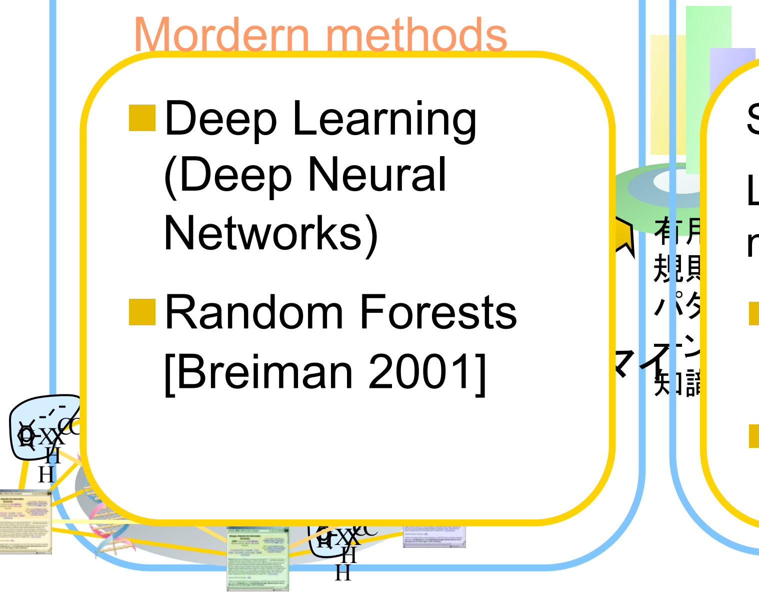
C. Pattern Discovery

Finding common / interesting patterns in a given data set

- Frequent pattern mining [Agrawal et a. '94]

Mordern methods

- Deep Learning (Deep Neural Networks)
- Random Forests [Breiman 2001]



Statistical Modeling.

Learning statistical models from data

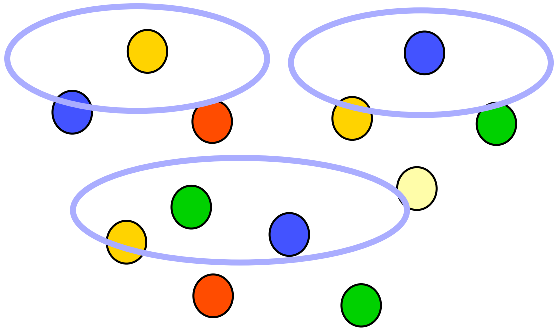
- Bayesian Network [Pearl '90s]
- Topic models [Blei, Ng, Jordan, 2003]

Graph Mining [Zaki '02], [Uno, Arimura]

- Emerging pattern mining
- Statistically significant pattern mining

Applications

- Text Mining
- Stream Mining, etc.



APPLICATIONS OF DATA MINING/MACHINE LEARNING

Applications of machine learning

Questions

- Find an example of machine learning applications in your life
- What can be done in future

Bio-technology

- Rapid growth of genome data such as sequencing data, and gene expression data
- Prediction of the functions of unknown genes from sequences.
- Automatically finding candidates of medicines from the structures of chemical compounds

Applications of machine learning

Finance: Credit card fraud detection

- Disicovering suspicious transactions and cash withdrawal from massive transaction records.

Security and Transportation: Image Recognition

- Recognizing faces and tracking moving people and cars from images using machine learning techniques.

Marketing

- We can predict customers' preference and trends from purchase data.
- As applications, recommendation services for your favorite musics and books are now available in Amazon.

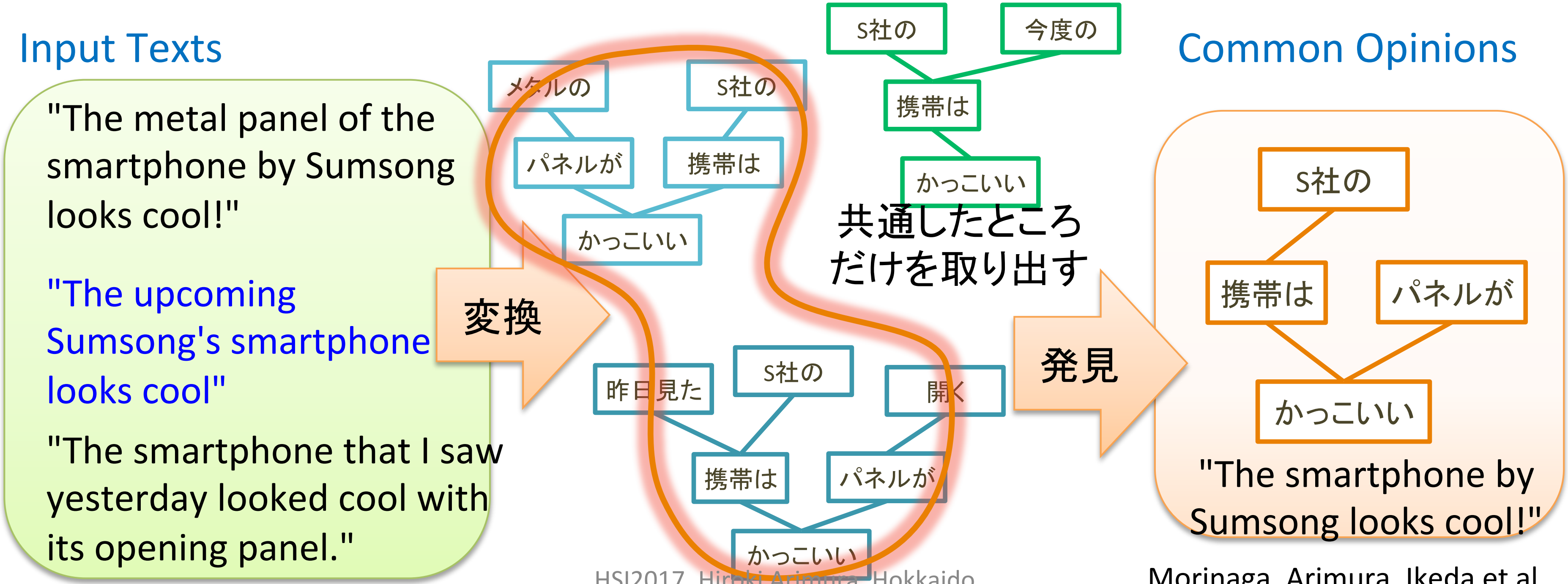
Spam detection

- Filter out advertisement messages from huge collections of e-mails.

Applications of machine learning

Text Mining :

- First, we extract the structure of a sentence as a "data tree" by analyzing a large collection of free text (Reputation in blogs and opinions in free-text questionary) by NLP technique.
- Next, we can extract common opinion by finding frequent common sub-structures (*tree patterns*) in the data tree.



Input Texts

"The metal panel of the smartphone by Sumsong looks cool!"

"The upcoming Sumsong's smartphone looks cool"

"The smartphone that I saw yesterday looked cool with its opening panel."

共通したところだけを取り出す

Common Opinions

"The smartphone by Sumsong looks cool!"

Morinaga, Arimura, Ikeda et al., ACM KDD'05, 2005.

NLP = Natural Language Processing

Discussion: Designing machine learning applications

For each of the previous applications, please think about the following questions

1. How to preprocess the data into feature vectors (a table)
2. What is the label (category) information?
3. Which learning algorithm do you use?
4. How to evaluate the results

Can a computer learn games from data?

photo：morgueFile

Can computer learn chess?

- In 1930, Alan Turing discussed the possiblity of computer programs playing chess games
- In 1950s, Samuel presented a machine learning program for playing checker
  - simpler than chess

http://en.wikipedia.org/wiki/, "Checker/ **Draughts** "の項目

Can computer learn chess?

- In 1950, Samuel's checker program won a human amature player.
- In 1990, a computer beats the checker world champion for the first time. (1)
- In 1997, a computer (Deep Blue by IBM) won the Chess world champion for the first time in chess game.(2)

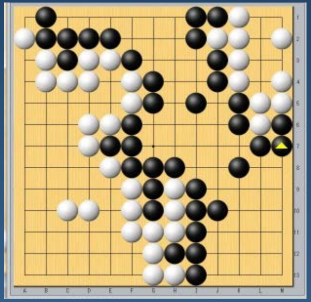
G. Kasparov (wikipedia)

  - Deep Blue can make 200M lookups per seconds（IBM RS600 x 32 + custom VLSI x 512）.
  - 1st match: human (Kasparov) won（3win 1lose 2draw）.
  - 2nd match: computer (Deep Blue) won （2win 1lose 3draw）.

1) Chinook project@ualberta: http://webdocs.cs.ualberta.ca/~chinook/project/

2) Garry Kasparov (1963～): at World champion in 1985–1993 and 1993–2000.

Can computer learn GO?



Computer GO

- In 2016, a computer program AlphaGo won the world's best human Go player (Ke Jie 柯潔) through three-match series.
- It was widely expected that computers cannot win human top player for the next ten years.
- **AlphaGo** has been developed by DeepMind team of Google for a few years.

Technology

- Combining game search with several machine learning techniques
- Monte carlo tree search (MCTS)
- Reinforcement learning

photo of
AlphaGo vs. Ke Jie

See also "AlphaGo versus Ke Jie" in Wikipedia

Summary: Introduction to Data Mining

- History of AI and Data Mining
- SKICAT Project: Application of machine learning to astronomical big data
- Classification of data mining algorithms
  - Supervised learning (classification)
  - Unsupervised learning (clustering)
  - Pattern mining
- Applications of data mining
- Data mining in Computer Game