Về khoa học dữ liệu và khai phá dữ liệu
Data mining landscape

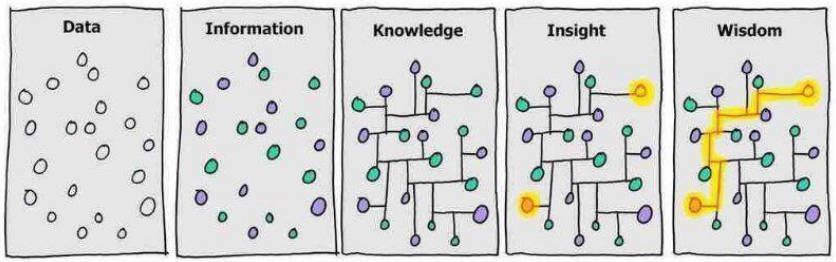
Hồ Tú Bảo
Japan Advanced Institute of Science and Technology

Outline

- **Statistics, machine learning, data mining, and data science**
- Issues in data mining
- Development of data mining and its challenges

Một số slides chưa chuyển qua tiếng Việt nhưng sẽ được trình bày bằng tiếng Việt

Data, information, knowledge, and wisdom



From Julien Blin

How knowledge is created?

Chuồn chuồn bay thấp thì mưa
Bay cao thì nắng bay vừa thì râm (thôi)

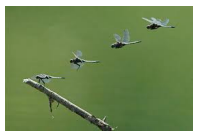
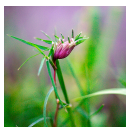
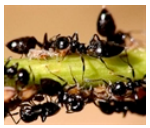
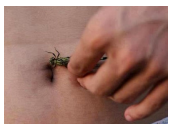
Mùa hè đang nắng, cỏ gà trắng thì mưa.
Cỏ gà mọc lang, cả làng được nước.

Kiến đen tha trứng lên cao
Thế nào cũng có mưa rào rất to

Chuồn chuồn cắn rốn, bốn ngày biết bơi!

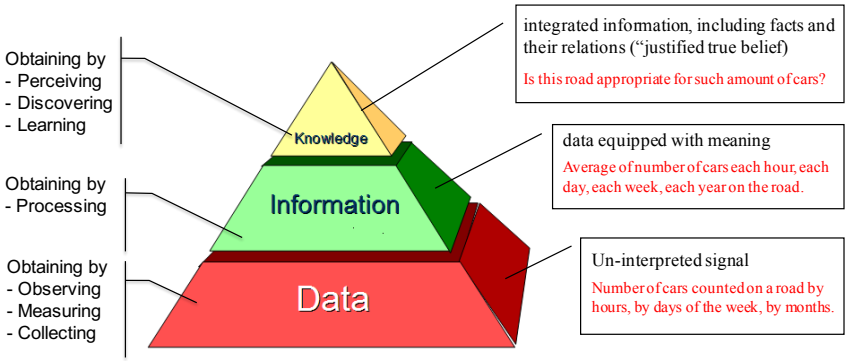
Biết $\{x_i\}$, Tìm $f(x)$
Induction (quy nạp)

Biết $f(x)$ và x_i, Tìm $f(x_i)$
Deduction (suy diễn)

Data, information, and knowledge

Knowledge can be considered data at a high level of abstraction and generalization.



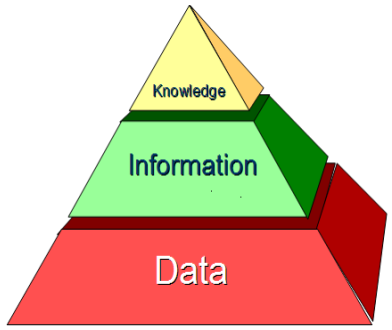
How does people collect data?

- Dữ liệu chính là **giá trị của các thuộc tính** (features, attributes, properties, variables) của các đối tượng, thu được do quan sát, đo đạc và thu thập.
- Hai cách thu thập dữ liệu

Lấy mẫu ngẫu nhiên	Thu mọi dữ liệu có được
Conventional statistics: First have the target then collect data to reach the target. Methods were mostly created for small or medium-sized data sets.	**Data Mining**: Data are collected without connecting to any target. Many innovative multivariate techniques being developed to solve large-scale data problems.

From data to knowledge?

Có thể xem tri thức là dữ liệu ở mức khái quát hoá cao (generalization).



Nhiều khoa học liên quan việc đi từ dữ liệu đến tri thức

- Statistics
- Machine Learning
- Data Mining
- Data Science

Thống kê - Statistics

- **Thống kê** cung cấp các phương pháp và kỹ thuật toán học để phân tích, khái quát và ra quyết định từ dữ liệu.
- **Nội dung chính**
  - *Thống kê mô tả* (descriptive statistics): phân bố xác suất...
  - *Thống kê suy diễn* (inferential statistics): ước lượng và kiểm định giả thiết thống kê...)
- Dữ liệu từ thí nghiệm và dữ liệu quan sát
  - Dữ liệu thống kê thường được thu thập để *trả lời những câu hỏi được định trước* (experiment design, survey design)
  - Phần lớn là dữ liệu số, ít dữ liệu hình thức (symbolic).
- Nhiều phương pháp phát triển cho tập *dữ liệu nhỏ*, phân tích từng biến ngẫu nhiên riêng lẻ, trước khi có máy tính.

Phân tích dữ liệu nhiều biến
Multivariate analysis

- Phân tích đồng thời quan hệ của nhiều biến ngẫu nhiên
- *Phân tích thăm dò* (EDA, exploratory data analysis) dùng dữ liệu tạo ra các giả thiết vs. việc kiểm định giả thiết trong *Phân tích khẳng định* (CDA, confirmatory data analysis)
  - Factor analysis, PCA, Linear discriminant analysis
  - Regression analysis
  - Cluster analysis
- Thấy gì từ các phương pháp truyền thống?
  - Kết quả nghèo trên dữ liệu lớn và phức tạp
  - Các phương pháp truyền thống chỉ phân tích tập dữ liệu nhỏ.
  - Giá lưu trữ và xử lý dữ liệu giảm nhanh thập kỷ qua.

Phân tích dữ liệu nhiều biến
Multivariate analysis

- Phương pháp phân tích được tạo ra cho các tập dữ liệu có kích thước nhỏ hoặc trung bình, và khi máy tính còn yếu.
- Phân tích thống kê nhiều biến đang thay đổi nhanh do kỹ thuật tính toán nhanh và hiệu quả hơn. Nhiều phương pháp mới được phát triển để giải các bài toán lớn (Pagerank của Google nghịch đảo ma trận kích thước nhiều tỷ chiều)

Nov. 2012: Cray's Titan computer, 17.59 petaflops, 560640 processors.

June 2013: China Tianhe-2, 33.86 petaflops, 3,120,000 Intel cores (No. 1. Sunway TaihuLight)

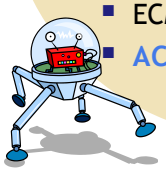
Machine learning and data mining

Machine learning

- To build computer systems that learn as human does.
- ICML since 1982 (33th ICML in 2016), ECML since 1989.
- ECML/PKDD since 2001.
- ACML starts Nov. 2009.

Data mining

- To find new and useful knowledge from large datasets.
- ACM SIGKDD (1995), PKDD and PAKDD (1997) IEEE ICDM and SIAM DM (2000), etc.

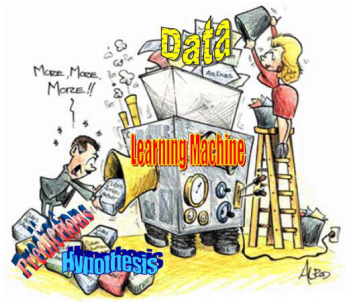


ACML: Asia Conference on Machine Learning
PAKDD: Pacific Asia Knowledge Discovery and Data Mining

M?achine learning

- Field of study that gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959).
- Một chương trình máy tính được nói là
  - học từ kinh nghiệm **E**
  - cho một lớp các nhiệm vụ **T**
  - với độ đo hiệu suất **P**

  nếu *hiệu suất* của nó với nhiệm vụ **T**, đánh giá bằng **P**, có thể tăng lên cùng kinh nghiệm.
  (Tom Mitchell, 1997)



(from Eric Xing lecture notes)

- Three main AI targets: Automatic Reasoning, Language understanding, Learning
- Finding hypothesis f in the hypothesis space F by narrowing the search with constraints (bias)

Khai phá dữ liệu – Data Mining

Tự động khám phá, phát hiện các tri thức tiềm ẩn từ các tập dữ liệu lớn và đa dạng.

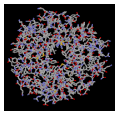
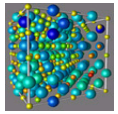
Data mining metaphor:
Extracting ore from rock

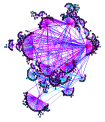
Statistics

KDD

Databases Machine Learning

Large and unstructured real-life data



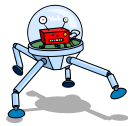
Statistics vs. Machine Learning

Statistics

- Nhấn mạnh suy diễn thống kê hình thức (ước lượng, kiểm định giả thiết).
- Dựa trên các mô hình (models) cho bài toán có số chiều nhỏ, ở dạng số.
- Khoa học đã thiết lập, ít 'văn hóa' thay đổi và thích nghi với môi trường tính toán.
- Có xu hướng mở rộng sang học máy.

Machine learning

- Nhấn mạnh các bài toán dự đoán, bắt đầu với dữ liệu hình thức.
- Bước đầu chủ yếu xây dựng và dùng các thuật toán trực cảm (heuristics algorithms).
- Gắn với thống kê nhiều hơn, xây dựng mô hình toán cho các thuật toán (statistical models underlying the algorithms).



Thống kê vs. Khai phá dữ liệu

Feature	Statistics	Data Mining
Kiểu bài toán & dữ liệu	Có cấu trúc (well structured)	Không cấu trúc/Nửa cấu trúc Unstructured/Semi-structured
Mục đích phân tích và thu thập dữ liệu	Xác định mục tiêu rồi thu thập dữ liệu	Dữ liệu thu thập thường không liên quan đến mục tiêu
Kích thước dữ liệu	Nhỏ và thường thuần nhất	Lớn và thường không thuần nhất.
Mô thức/tiếp cận Paradigm/approach	Dựa trên lý thuyết suy diễn Theory based (deductive)	Phối hợp lý thuyết và trực cảm Theory & heuristic based (inductive)
Kiểu phân tích	Confirmative (khẳng định)	Explorative (thăm dò, khai phá)
Số biến	Nhỏ	Lớn
Giả định về phân bố Distribution assump.	Dựa trên giả định về phân bố	Không giả định phân bố xác suất

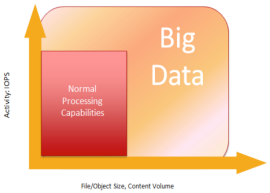
Thấy gần đây

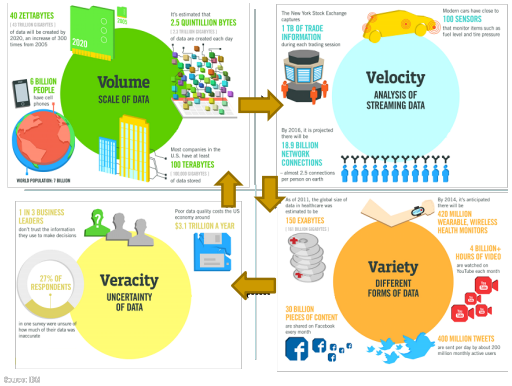
London subway

London taxi

Big data là gì?

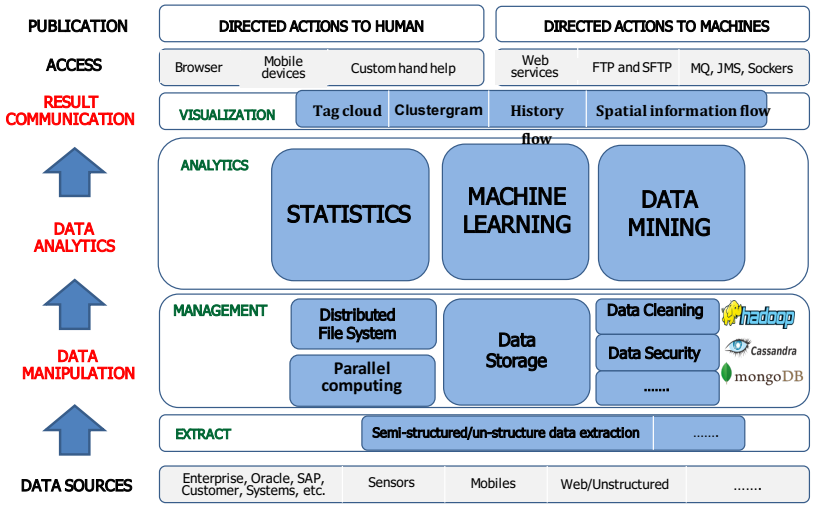
Dữ liệu lớn nói về các **tập dữ liệu rất lớn** và/hoặc **rất phức tạp**, vượt quá khả năng xử lý của các kỹ thuật IT truyền thống (*View 1*).



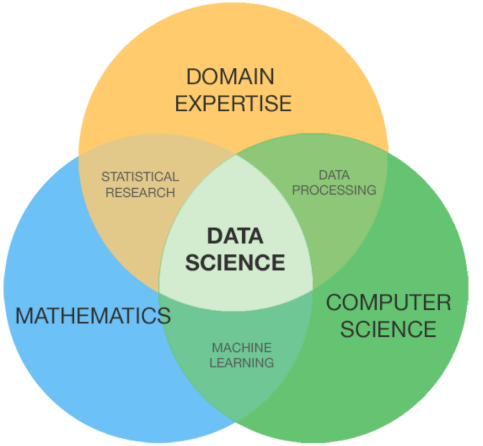


(View 2) Big Data is about technology (tools and processes).
(View 3) Hiện tượng khách quan mà các tổ chức, doanh nghiệp… phải đối đầu để phát triển.

A scheme of data science



Data science



Source: Palmer, Shelly. Data Science for the C-Suite. New York: Digital Living Press, 2015. Print.

"Chỉ Thượng đế là đáng tin. Mọi thứ khác đều phải dựa vào dữ liệu"

Data Scientist: The Sexiest Job of the 21st Century
(Harvard Business Review, October 2012)

Outline

- Statistics, machine learning, data mining, and data science
- **Issues in data mining**
  1) Types, models and structures of data
  2) Data mining process
  3) Model assessment and selection
  4) Data mining methods
  5) Others
- Development of data mining and its challenges

Một số slides chưa chuyển qua tiếng Việt nhưng sẽ được trình bày bằng tiếng Việt

Data types and models vs. mining methods

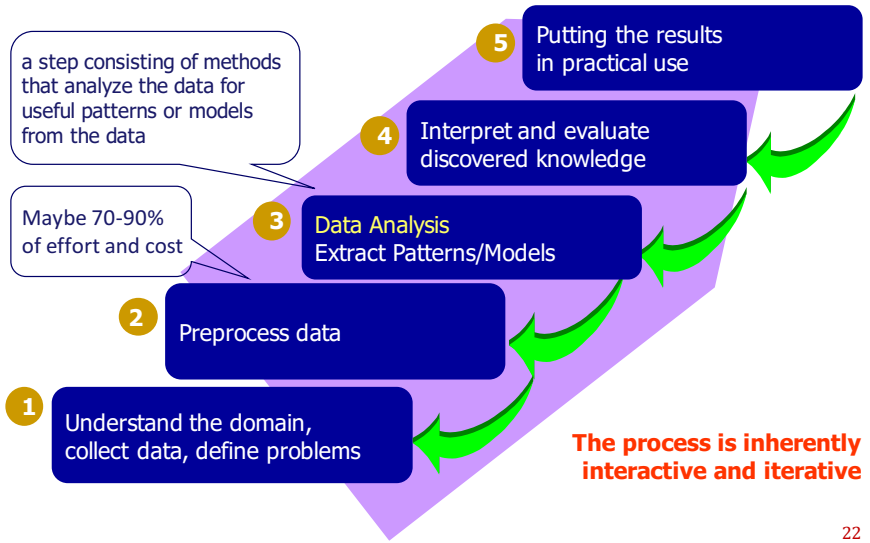
Data types and models

- Flat data tables
- Relational databases
- Temporal & spatial data
- Transactional databases
- Multimedia data
- Genome databases
- Materials science data
- Textual data
- Web data
- etc.

Mining tasks and methods

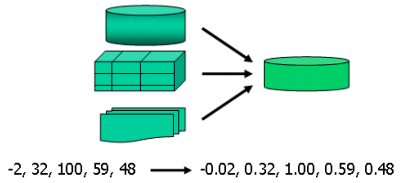
- Classification/Prediction
  - Decision trees
  - Bayesian classification
  - Neural networks
  - Rule induction
  - Support vector machines
  - Hidden Markov Model
  - etc.
- Description
  - Association analysis
  - Clustering
  - Summarization
  - etc.

The data analysis process

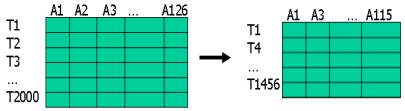


Major tasks in data preprocessing

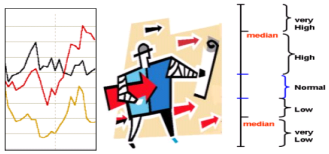
1 Data cleaning



2 Data integration and transformation



3 Data reduction
(instances and dimensions)



4 Data discretization

Data types

- **SYMBOLIC**
  - **Indexing**: E.g., names, tags, case numbers, or serial numbers that identify a respondent or group of respondents.
  - **Binary**: Two values, e.g., YES or NO, SUCCESS or FAILURE, MALE or FEMALE, WHITE or NON-WHITE, FOR or AGAINST, and so on.
  - **Boolean**: Two values TRUE or FALSE, and may have the value UNKNOWN.
  - **Nominal**: Character-string values (green, blue, red, ...)
  - **Ordinal**: Values for this character-string data type are linearly ordered (Small, Middle, Large,...)
- **NUMERIC**
  - **Integer**: Values are just integer numbers
  - **Continuous**: real numbers.

Symbols or Numbers

Why we should care about data types?

Combinatorial search in hypothesis spaces (machine learning)

Attribute	Numerical	Symbolic	
No structure $= \neq$		Places, Color	**Nominal or categorical** (Binary, Boolean)
Ordinal structure $= \neq \geq$	**Integer**: Age, Temperature	Rank, Resemblance	Ordinal
Ring structure $= \neq \geq + \times$	**Continuous**: Income, Length		Measurable

Posible analysis operations (thus methods, algorithms) depend on data types

Often matrix-based computation (multivariate data analysis)

Structures of data

- **Structured data**
  - Can be stored in database SQL in table with rows and columns.
  - Only about 5-10% of all available data.
- **Semi-structured data**
  - Doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze.
  - XML documents and NoSQL databases documents are semi structured

	swims	has fins	flies	has lung	is a fish
Herring	yes	yes	no	no	yes
Cat	no	no	no	yes	no
Pigeon	no	no	yes	yes	no
Flying fish	yes	yes	yes	no	yes
Otter	yes	no	no	yes	no
Cod	yes	yes	no	no	yes
Whale	yes	yes	no	yes	no

```
@BOOK{Maz91,
   author = "J. Mazzeo",
   year = "1991",
   title = "Comparability of Computer and Paper-and-Pencil Scores",
   address = "(College Board Rep. No. 91). Princeton, NJ",
   publisher = "Educational Testing Service",}

@BOOK{Mi193,
   author = "M. E. Miller",
   year = "1993",
   title = "The Interactive Tester (Version 4.0) [Computer software]",
   publisher = "Psytek Services",
   address = "Westminster, CA",}
```

Articls in a Latex database

Structures of data

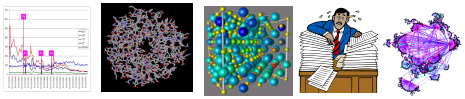
- **Unstructured data**
  - Unstructured data represent around 80% of data. It often include text and multimedia content.

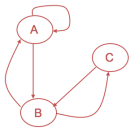
    Example: e-mail messages, word documents, videos, photos, audio files, webpages and many other kinds of business documents.
  - A key issue in data science is **representing unstructured data**

    Example: The DNA sequence

"…TACATTAGTTATTACATTGAGAAACTTTATAATTAAAAAAGATTC…"

can be represented by different ways for computation such as sliding windows, motifs, kernel function, web link… representation





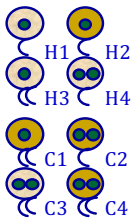
	A	B	C
A	1/2	1/2	0
B	1/2	0	1
C	0	1/2	0

Supervised vs. Unsupervised data

Given: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$

- x_i is description of an object, phenomenon, etc.
- y_i (label attribute) is some property of x_i, if not available learning is unsupervised

Find: a function $f(x)$ that characterizes $\{x_i\}$ or that $f(x_i) = y_i$



Unsupervised data

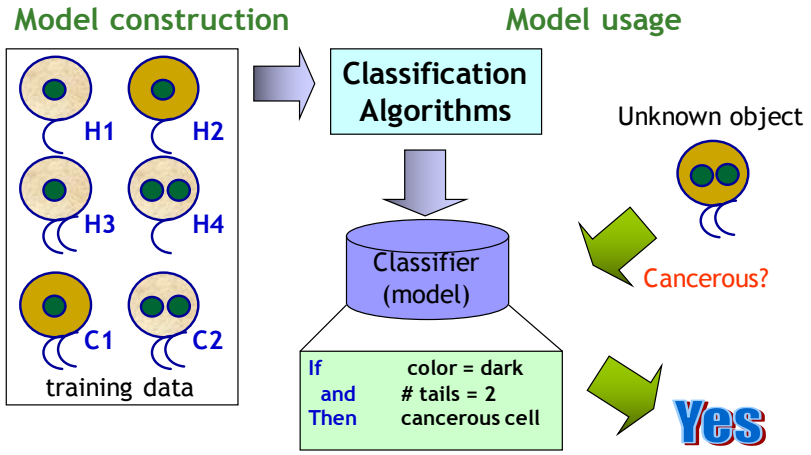
	color	#nuclei	#tails
H1	light	1	1
H2	dark	1	1
H3	light	1	2
H4	light	2	1
C1	dark	1	2
C2	dark	2	1
C3	light	2	2
C4	dark	2	2

Supervised data

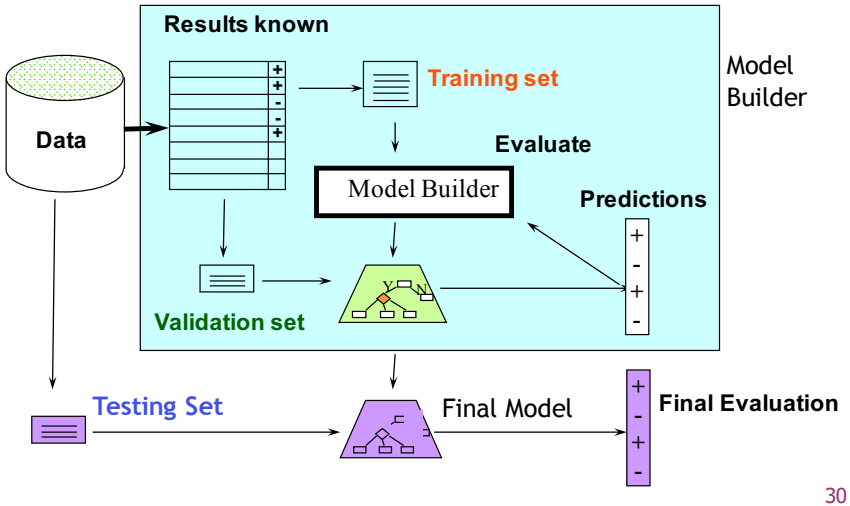
	color	#nuclei	#tails	label
H1	light	1	1	heal
H2	dark	1	1	healthy
H3	light	1	2	healthy
H4	light	2	1	healthy
C1	dark	1	2	cancerous
C2	dark	2	1	cancerous
C3	light	2	2	cancerous
C4	dark	2	2	cancerous

The problem is usually called **classification** if "label" is categorical, and **prediction** if "label" is continuous (in this case, if the descriptive attribute is numerical the problem is **regression**)

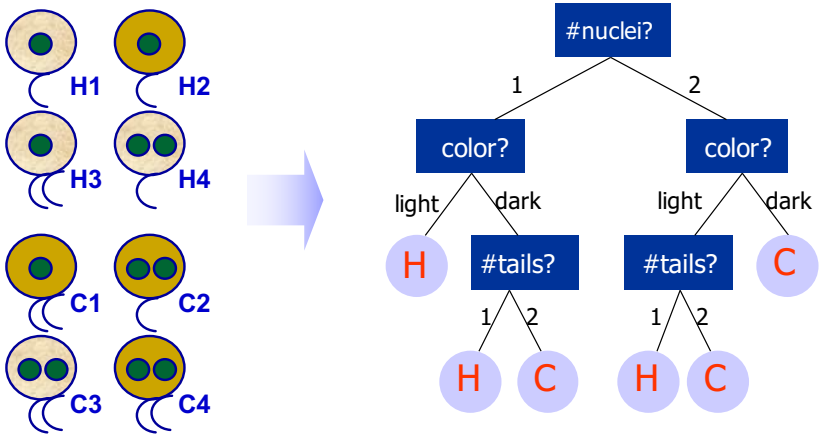
Classification—a two-step process



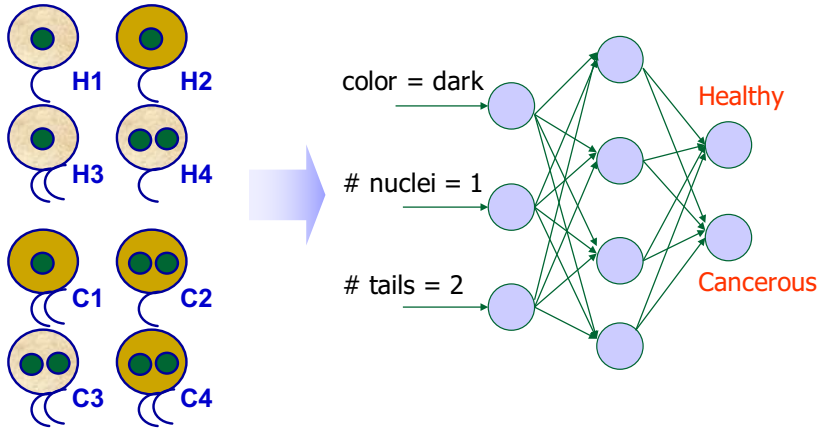
Classification: Train, Validation, Test



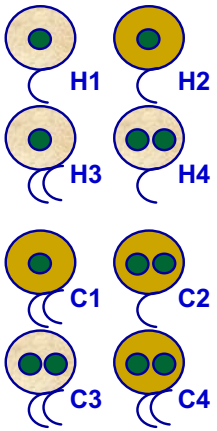
Classification with decision trees



Classification with neural networks



Classification with decision rules

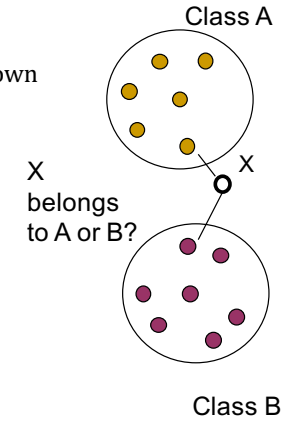


IF Color = light
AND # nuclei = 1
THEN Healthy

IF Color = dark
AND # nuclei = 2
THEN Cancerous

Instance-based classification

- **Instance-based classification**
  - Using most similar individual instances known in the past to classify a new instance
- **Typical approaches**
  - **k-nearest neighbor approach**
    - Instances represented as points in a Euclidean space
  - **Locally weighted regression**
    - Constructs local approximation
  - **Case-based reasoning**
    - Uses symbolic representations and knowledge-based inference



Bayesian classification

- The essence of Bayes' theorem is that tell us how to update our initial probabilities $P(h)$ if we see evidence E, in order to find out $P(h|E)$

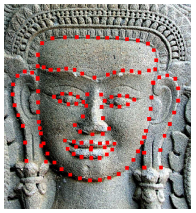
$$P(h|E) = \frac{P(E|h)P(h)}{P(E)}$$

$$P(h|E) = \frac{P(E|h).P(h)}{P(E)} = \frac{P(E|h).P(h)}{P(E|h).P(h) + P(E|\neg h).P(\neg h)}$$

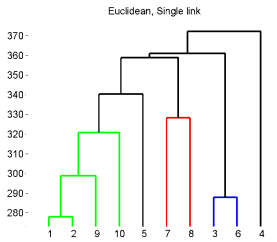
- A prior probability
- Conditional probability (likelihood) ← coming from the data
- Posteriori probability

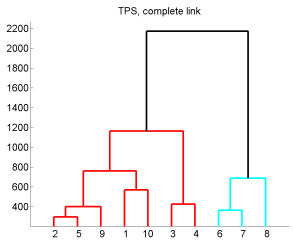
- Naïve assumption: *attribute independence*
- Bayesian belief network allows a *subset* of the variables conditionally independent.

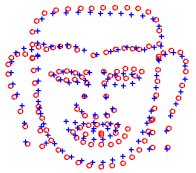
Clustering (Apsara faces)



127 landmarks







Shape alignment

Single Link clusters

An ethnologist needs to validate the groups

Mining associations

Super market data

"Young men buy diaper and beer together"

Dữ liệu siêu thị | Khai phá dữ liệu | Trai trẻ 20-30 tuổi | Bỉm | Bia

Young men when buy diaper also buy beer to use in the weekend when keeping their children and watching TV.

Many other issues

- Ensemble learning
- Transfer learning
- Learning to rank
- Multi-instance multi-label learning
- Semi-supervised learning
- Structured prediction
- Social network analysis (Trần Mai Vũ)
- Learning in specific domains
- etc.

KDD nuggets

Nguồn thông tin lớn nhất về khai phá dữ liệu

www.kdnuggets.com is website of the data mining community

Which algorithms perform best at which tasks?

Algorithm	Pros	Cons	Good at
Linear regression	- Very fast (runs in constant time) - Easy to understand the model - Less prone to overfitting	- Unable to model complex relationships - Unable to capture nonlinear relationships without first transforming the inputs	- The first look at a dataset - Numerical data with lots of features
Decision trees	- Fast - Robust to noise and missing values - Accurate	- Complex trees are hard to interpret - Duplication within the same sub-tree is possible	- Star classification - Medical diagnosis - Credit risk analysis
Neural networks	- Extremely powerful - Can model even very complex relationships - No need to understand the underlying data - Almost works by "magic"	- Prone to overfitting - Long training time - Requires significant computing power for large datasets - Model is essentially unreadable	- Images - Video - "Human-intelligence" type tasks like driving or flying - Robotics
Support Vector Machines	- Can model complex, nonlinear relationships - Robust to noise (because they maximize margins)	- Need to select a good kernel function - Model parameters are difficult to interpret - Sometimes numerical stability problems - Requires significant memory and processing power	- Classifying proteins - Text classification - Image classification - Handwriting recognition
K-Nearest Neighbors	- Simple - Powerful - No training involved ("lazy") - Naturally handles multiclass classification and regression	- Expensive and slow to predict new instances - Must define a meaningful distance function - Performs poorly on high-dimensionality datasets	- Low-dimensional datasets - Computer security: intrusion detection - Fault detection in semi-conducter manufacturing - Video content retrieval - Gene expression - Protein-protein interaction

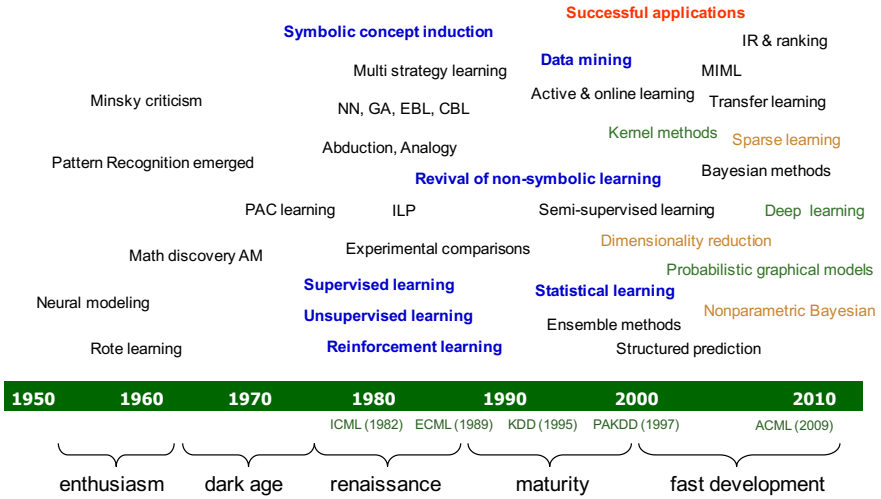
http://www.lauradhamilton.com/machine-learning-algorithm-cheat-sheet

Outline

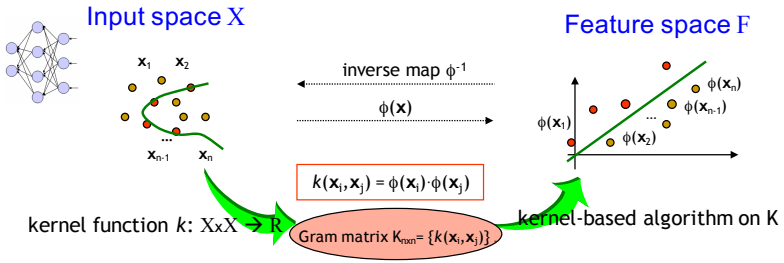
- Statistics, machine learning, data mining, and data science
- Issues in data mining
- **Development of data mining and its challenges**

Một số slides chưa chuyển qua tiếng Việt nhưng sẽ được trình bày bằng tiếng Việt

Development of machine learning



Kernel methods: the scheme

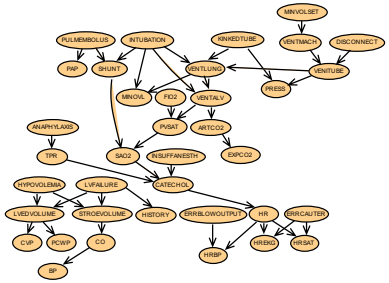


- Biến đổi dữ liệu từ X bởi một ánh xạ $\phi(x)$ vào một không gian vector (nhiều chiều), gọi là feature space F.
- Tìm một hàm/mô hình tuyến tính (hoặc một hàm tốt hơn) trong F bằng các thuật toán quen biết thực hiện trên Gram matrix.
- Bởi một ánh xạ ngược, hàm tuyến tính trên F có thể tương ứng với một hàm phức tạp trên X.
- Điều này có thể thực hiện đơn giản hơn do sử dụng nội tích (inner products) trong F (kernel trick) xác định bởi một hàm hạch (kernel function).

Probabilistic graphical models

- Kết nối graph theory và probability theory trong một hình thức chặt chẽ cho mô hình hoá thống kê nhiều chiều.
- Probability theory đảm bảo tính nhất quán (consistency) và cho mô hình mô tả và kết nối với dữ liệu.
- Graph theory cho một giao diện trực giác với con người.

  "Ngôn ngữ đồ thị cho ta cách diễn giải rõ tính chất thực tế: các biến có xu hướng chỉ tương tác *trực tiếp* với một số ít biến khác". (Koller's book).
- Modularity: Mọi hệ phức tạp đều được xây dựng từ những phần đơn giản hơn.

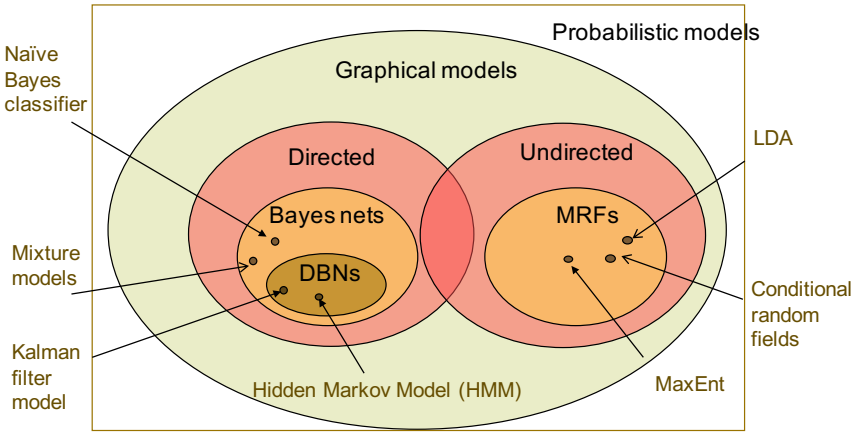


A ICU alarm network, 37 nodes, 509 parameters.

- **Issues**:
  - Representation
  - Learning
  - Inference
  - Applications

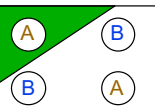
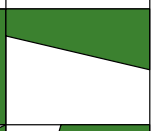
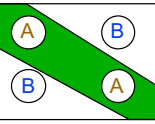
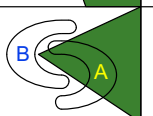
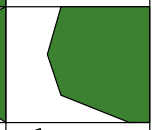
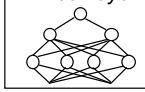
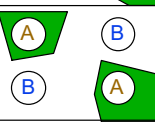
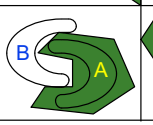
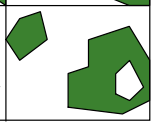
Probabilistic graphical models

Instances of graphical models



Murphy, ML for life sciences

Non-linearly separable problems and deep learning

Structure	*Types of Decision Regions*	*Exclusive-OR Problem*	*Classes with Meshed regions*	*Most General Region Shapes*
Single-Layer 	*Half Plane Bounded By Hyperplane*			
Two-Layer 	*Convex Open Or Closed Regions*			
Three-Layer 	Arbitrary (Complexity Limited by No. of Nodes)			

Lê Hồng Phương, Deep Learning for Text

Some typical books