Introduction to data science

Hồ Tú Bảo

Bài giảng của DSLab

Viện nghiên cứu cao cấp về Toán (VIASM)

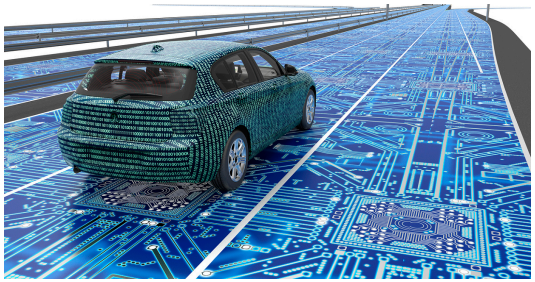
Vietnam Institute for
Advanced Study in Mathematics

Outline

1. What is data science?
2. Principles of data science
3. DSLab's data science lectures

Thế giới thực, số hoá và không gian số

Thế giới các thực thể



Thế giới các thực thể – không gian số
Physical-cyber systems

Cách mạng số hoá

Thế giới thực thể-không gian số

- '**Phiên bản số**' các thực thể: Biểu diễn các thực thể bằng '0' và '1' trên máy tính.
- Thí dụ: ô-tô, bệnh án điện tử…
- **Hệ thống không gian số-thế giới thực thể (cyber-physical system):** **kết nối** các thực thể và 'phiên bản số' của chúng.

Kết nối

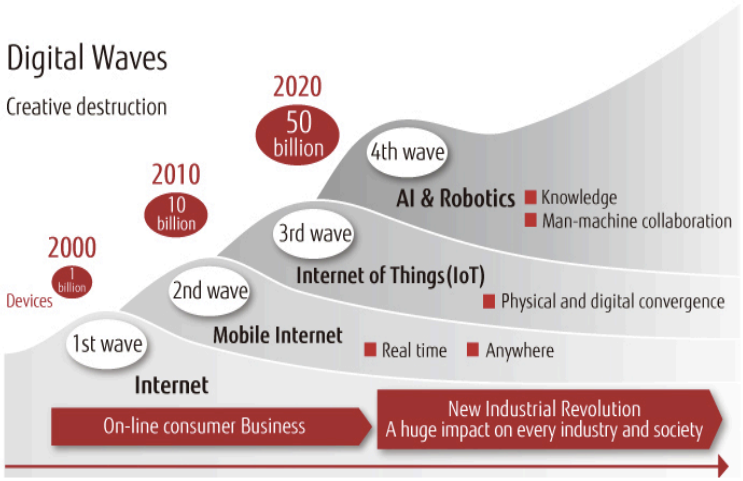
Hành động trong thế giới các thực thể ⟺ Tính toán, điều khiển trên không gian số

Thay đổi phương thức sản xuất

Dữ liệu là tài nguyên của thời chuyển đổi số

Chuyển đổi số, công nghệ số, kinh tế số

- **Thời chuyển đổi số**: Thời mọi thứ ngày càng được *số hoá* và *kết nối* nhiều hơn qua internet (internet of everything) và *công nghệ số* được dùng trong mọi mặt của xã hội và kinh tế.
- **Công nghệ số** là công nghệ về các đối tượng đã số hoá biểu diễn với các mã số (gồm *công nghệ số hoá* trong mọi lĩnh vực và *công nghệ xử lý* dữ liệu được số hoá).
- **Kinh tế số**: Nền kinh tế dựa vào công nghệ số và internet. Cốt lõi là kết nối và chia sẻ dữ liệu.

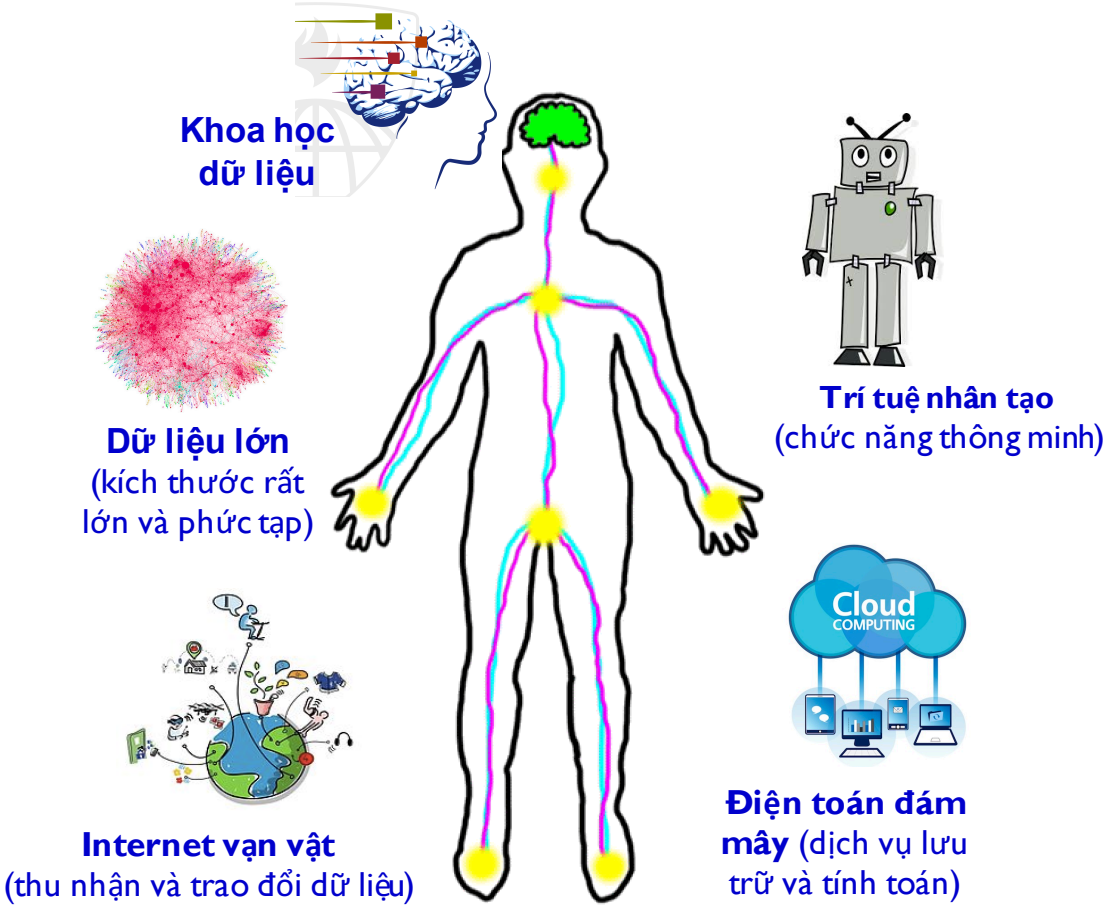


"How digital technology will transform the world", Fujitsu Journal, 1.2016

VIASM
Vietnam Institute for Advanced Study in Mathematics

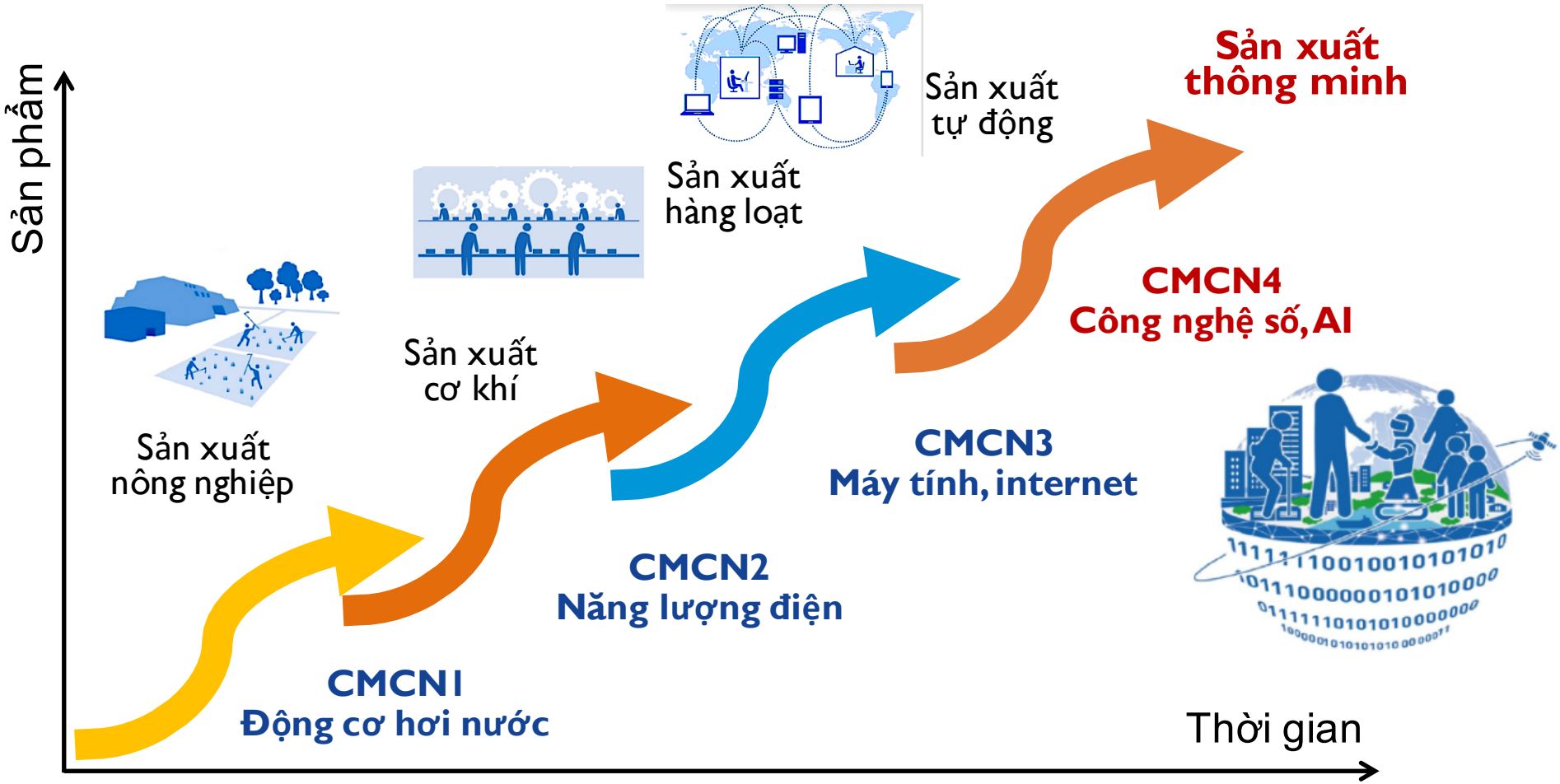
Đột phá của công nghệ số

- **Điện toán đám mây**:
  Môi trường
- **Dữ liệu lớn**:
  Năng lượng
- **Internet vạn vật**:
  Thần kinh & huyết mạch
- **Trí tuệ nhân tạo**:
  Chức năng thông minh
- **Khoa học dữ liệu**:
  "Bộ não" phân tích dữ liệu để hỗ trợ quyết định và hành động.



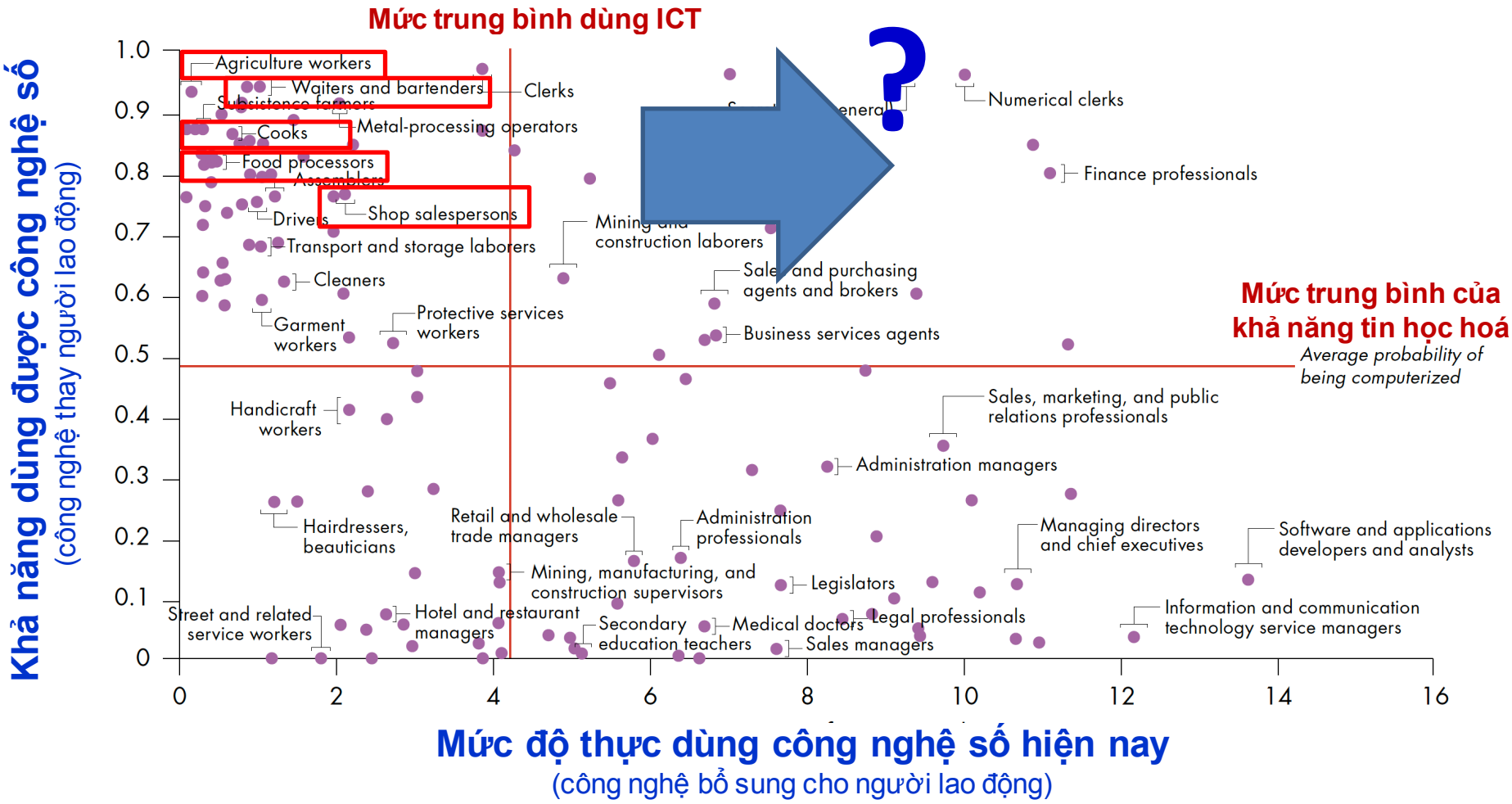
Tuy mọi công nghệ đều tiến bộ, cách mạng công nghiệp lần thứ tư xảy ra chủ yếu do hội tụ các đột phá của công nghệ số.

Cách mạng công nghiệp lần thứ tư



Vietnam Institute for Advanced Study in Mathematics

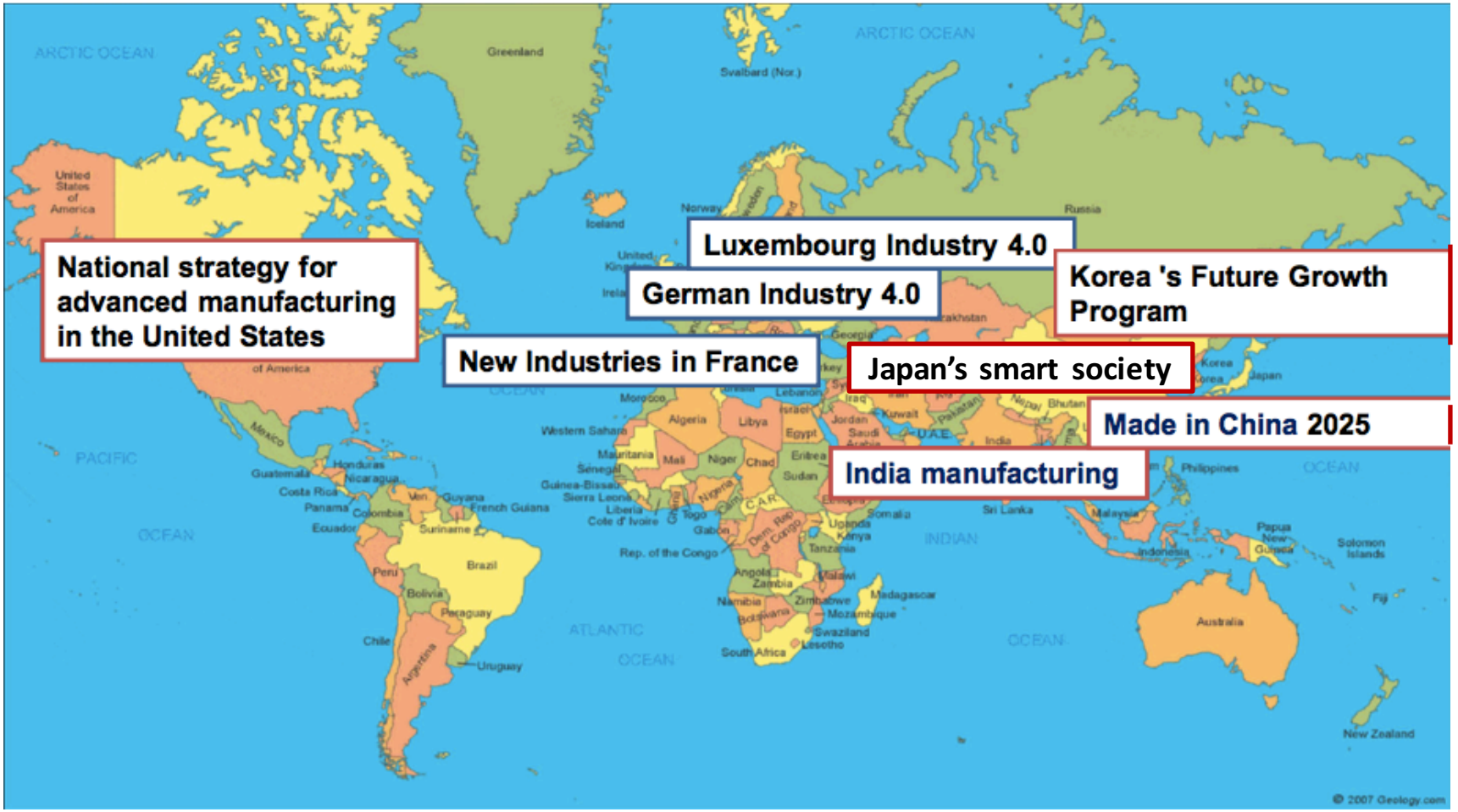
Khả năng và thực tế dùng công nghệ số



World development report 2016: Digital Dividends, World bank group.

VIASM
Vietnam Institute for Advanced Study in Mathematics

National strategies



Klaus Schwab (WEF), The Fourth Industrial Revolution
Alistair Nolan (OECD), Enabling the Next Production Revolution: Implications for Policy

Big data

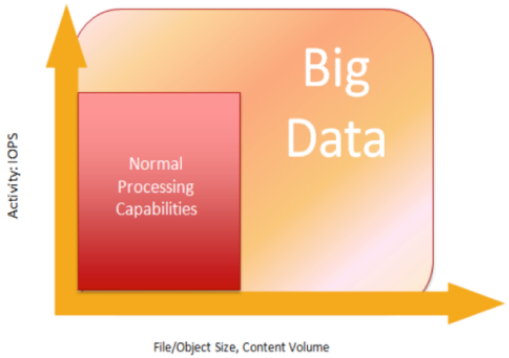
Big data refers to data sets that are **too large** and **too complex** to manage and analyze with traditional IT techniques.

Variety: Complexity of data in many different structures, ranging from relational, to logs, to raw text

Velocity: Streaming data and large volume data movement

Volume: Scale from Terabytes to Petabytes (10^{15} bytes) to Zetabytes (10^{18} bytes)

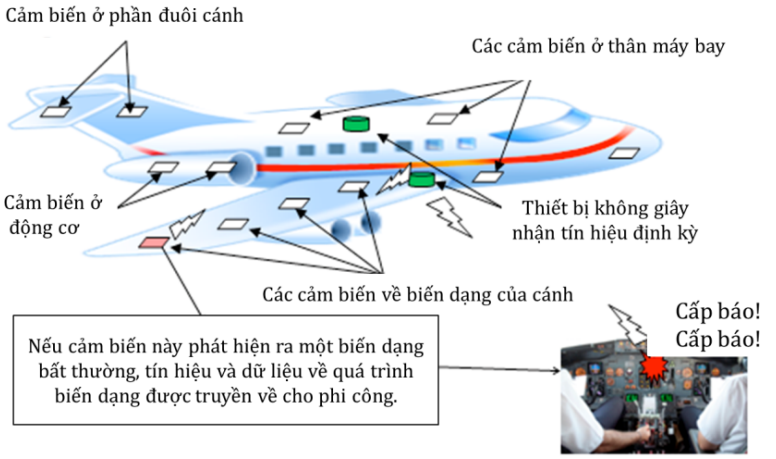
Veracity: Accuracy and precision, truthfulness of the data.



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

Dữ liệu lớn có thể rất nhỏ.
Không phải mọi tập dữ liệu to đều lớn
Big data can be very small. Not all large datasets are big

- **Big** liên quan tới sự **phức tạp** nhiều hơn tới **kích thước lớn**.
- **Dữ liệu lớn** nhưng lại nhỏ
  - Lò hạt nhân, máy bay… có hàng trăm nghìn sensors → sự phức tạp của việc tổ hợp dữ liệu các sensors này tạo ra?
  - Dòng dữ liệu của tất cả các sensors là lớn mặc dù kích thước của tập dữ liệu là không lớn (một giờ bay:
    100,000 sensors x 60 minutes
    x 60 seconds x 8 bytes < 3GB).
- Tập dữ liệu **to nhưng không lớn**
  - Số hệ thống dù tăng lên và tạo ra những lượng khổng lồ dữ liệu nhưng đơn giản.

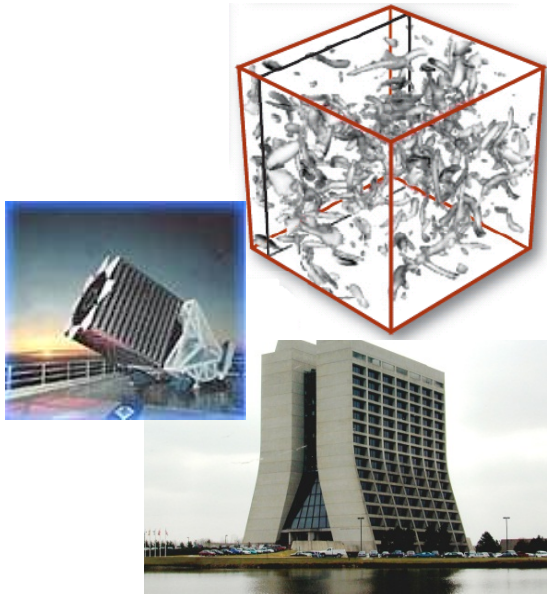


Vietnam Institute for Advanced Study in Mathematics

Science paradigm shifts

- Thousand years ago: science was **empirical**
  Describing natural phenomena
- Last few hundred years: **Theoretical branch**
  Using models, generalizations
- Last few decades: **Computational** branch
  Simulating complex phenomena
- Today: **Data exploration** (e-Science)
  Unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes databases/files using data management and statistics.

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G p}{3} - \mathrm{K}\frac{c^2}{a^2}$$



The Four Paradigm: Data-Intensive Scientific Discovery, 2009

Vietnam Institute for Advanced Study in Mathematics

Expert systems in medicine

the deduction approach

An expert system is a computer program that behaves like an expert in some narrow area of expertise, using expert knowledge.

MYCIN (Shortliffe, Feigenbaum, 1979): Infection Diagnosis.

IF
1. the infection is primary bacteremia, and
2. the site of the culture is one of the sterile sites, and
3. the suspected portal of entry of the organism is the gastro intestinal tract

THEN there is suggestive evidence (0.7) that the identity of the organism is bacteroides.

EMRs in medicine
the induction approach

General rules or Principles

Deduction →

← Induction

Specific Cases

Jaundice is yellowing of the skin and eyes and can indicate a serious problem with liver, gallbladder, or pancreas function

EMRs in medicine

the induction approach (data-driven approach)

General rules or Principles

Deduction

Induction

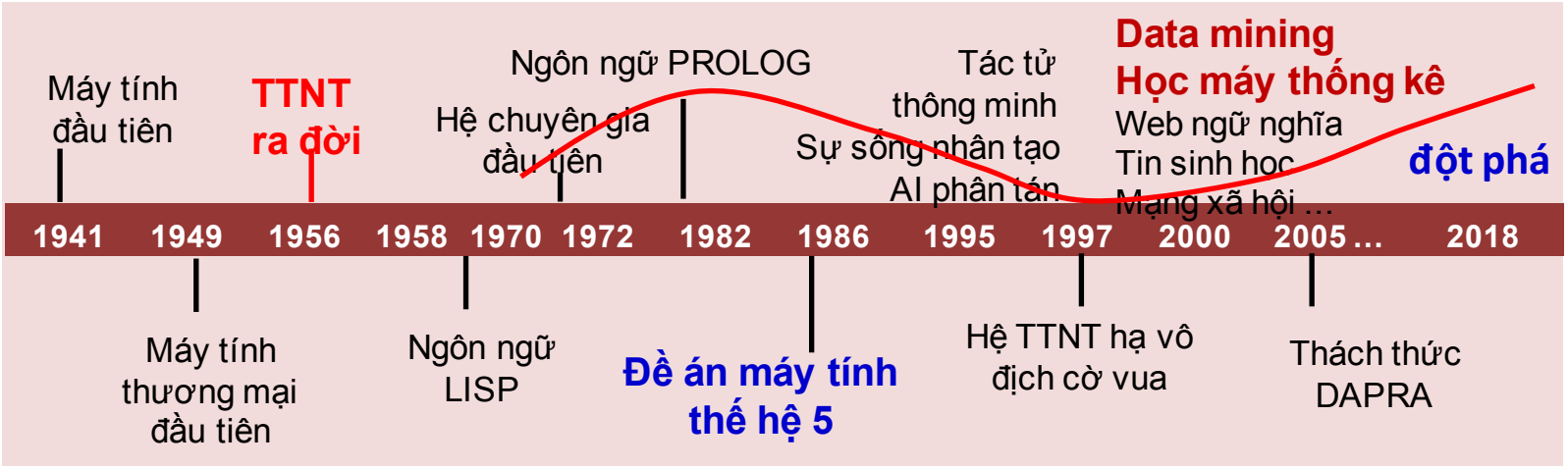
Specific Cases

Digitalization

VIASM
Vietnam Institute for Advanced Study in Mathematics

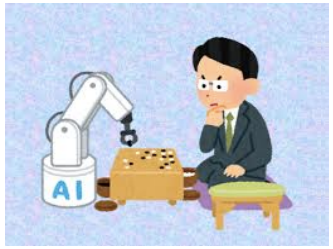
Trí tuệ nhân tạo - Artificial Intelligence

- Lĩnh vực làm cho máy (máy tính) hoạt động như có trí thông minh của con người (suy luận, giải quyết vấn đề, hiểu ngôn ngữ, học tập, nhận thức...).
- AlphaGo, hiểu ngôn ngữ, nhận dạng tiếng nói, chẩn đoán ung thư, ô-tô tự lái...

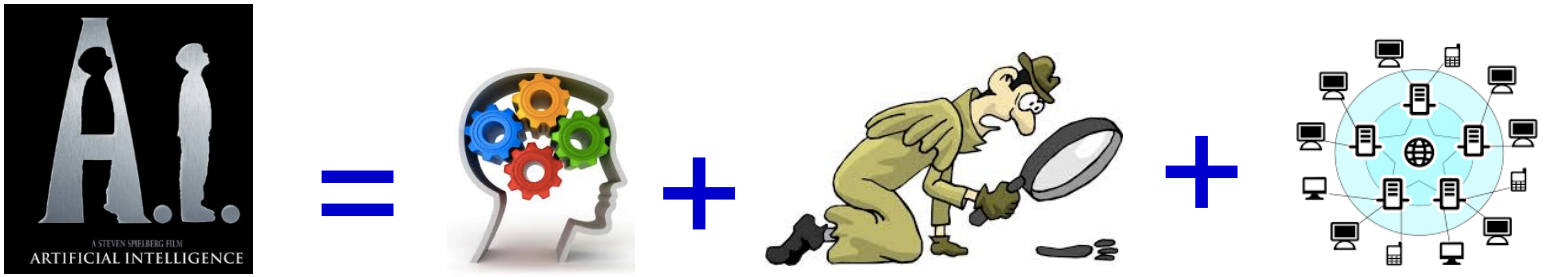


AI đã phát triển với nhiều thăng trầm 60 năm qua.

Trí tuệ nhân tạo - Artificial Intelligence



- Lĩnh vực làm cho máy (tính) hoạt động như có trí thông minh của con người (lập luận, hiểu ngôn ngữ, học tập...).
- AlphaGo, hiểu ngôn ngữ, nhận dạng tiếng nói, chẩn đoán ung thư, ô-tô tự lái...

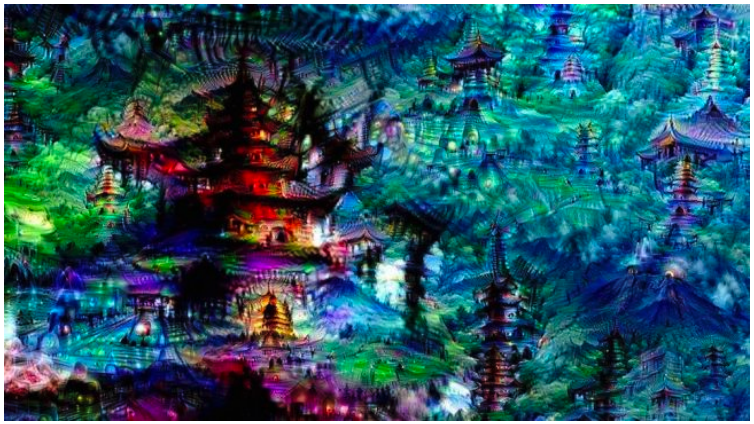


Hầu hết đột phá gần đây của AI dựa vào học máy (machine learning).

Từ đột phá học máy đến đột phá AI

"Rất nhiều người làm các hệ AI nay đã nhận ra rằng, đối với rất nhiều ứng dụng, việc huấn luyện một hệ thống từ các thí dụ đầu vào-đầu ra [machine learning] để có quyết định hành động là dễ hơn rất nhiều việc soạn sẵn các quyết định mong muốn cho mọi tình huống có thể xảy ra [expert systems].

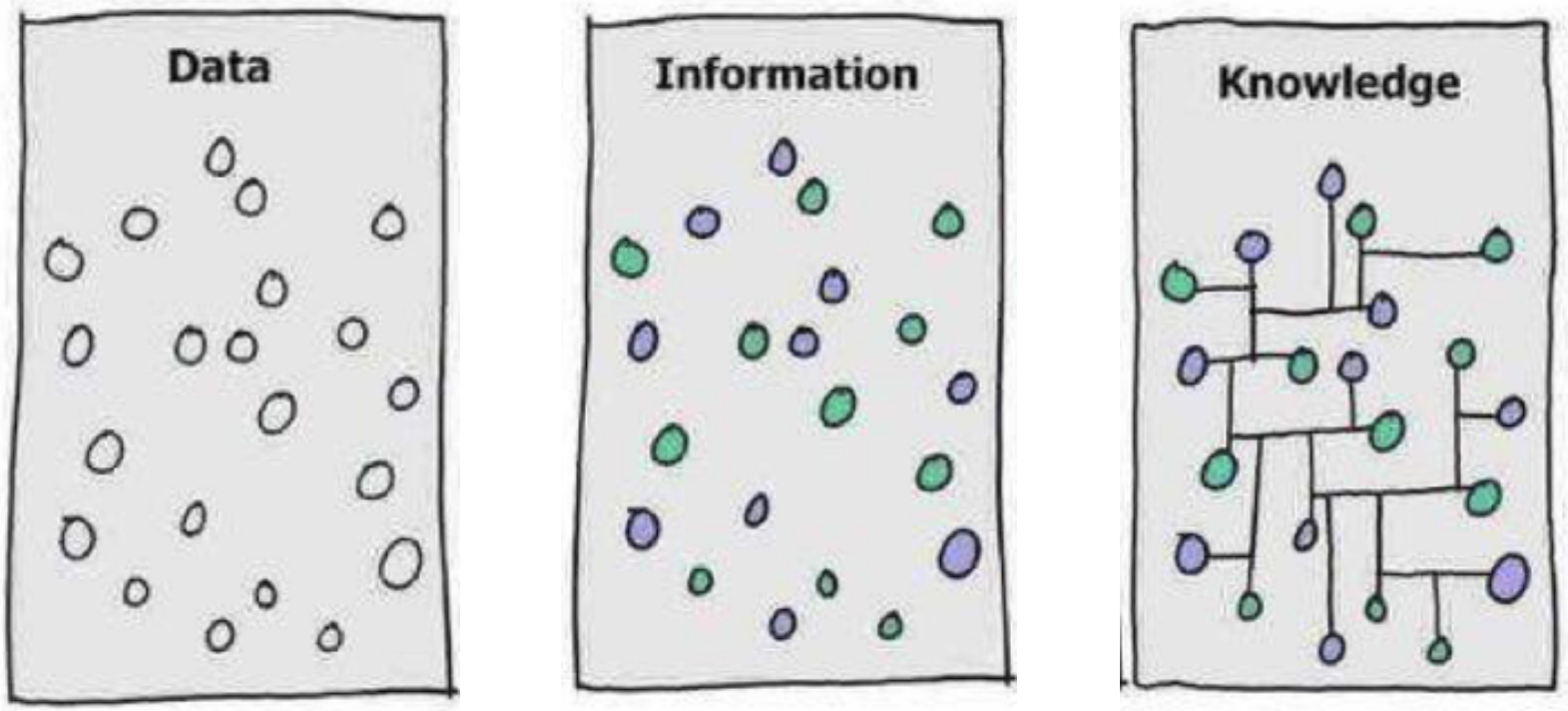
Reinforcement learning led to AlphaGo's stunning victory over a human Go champion last year.



An image created by a Google Neural Network

M.I. Jordan, T. Mitchell. Machine Learning: Trends, perspectives, and prospects. *Science*, 349 (6245), 2015.

Data, information, knowledge



From Julien Blin

Data, information, knowledge

- **Data** is often seen as a string of bits, or numbers and symbols, or "objects" which we **collect daily** (by observation, measurements, collection, etc.).
- **Information** is data stripped of redundancy, and reduced to the minimum necessary to **characterize the data**.
- **Knowledge** is integrated information, including facts and their relations, which have been **perceived, discovered, or learned** as our "mental pictures" ("justified true belief").
- Knowledge can be considered data at a high level of abstraction

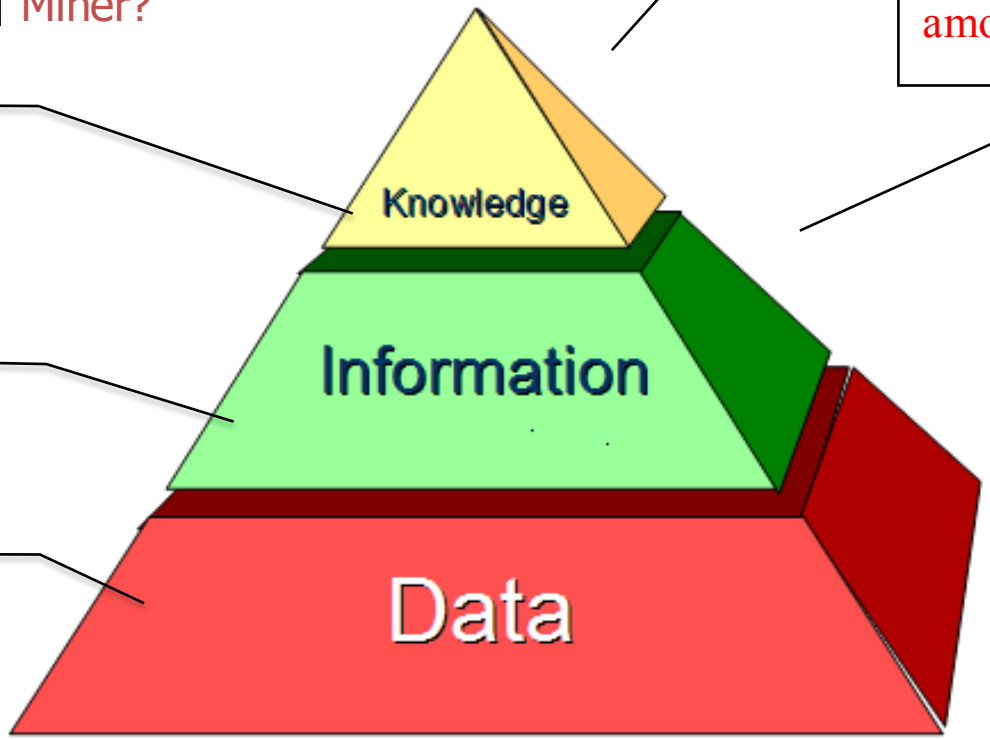
Data, information, and knowledge

Metaphor:
Data: rock;
knowledge: ore.
Miner?

integrated information, including facts and their relations ("justified true belief)

Is this road appropriate for such amount of cars?

Obtaining by
- **Perceiving**
- **Discovering**
- **Learning**



data equipped with meaning

Average of number of cars each hour, each day, each week, each year on the road.

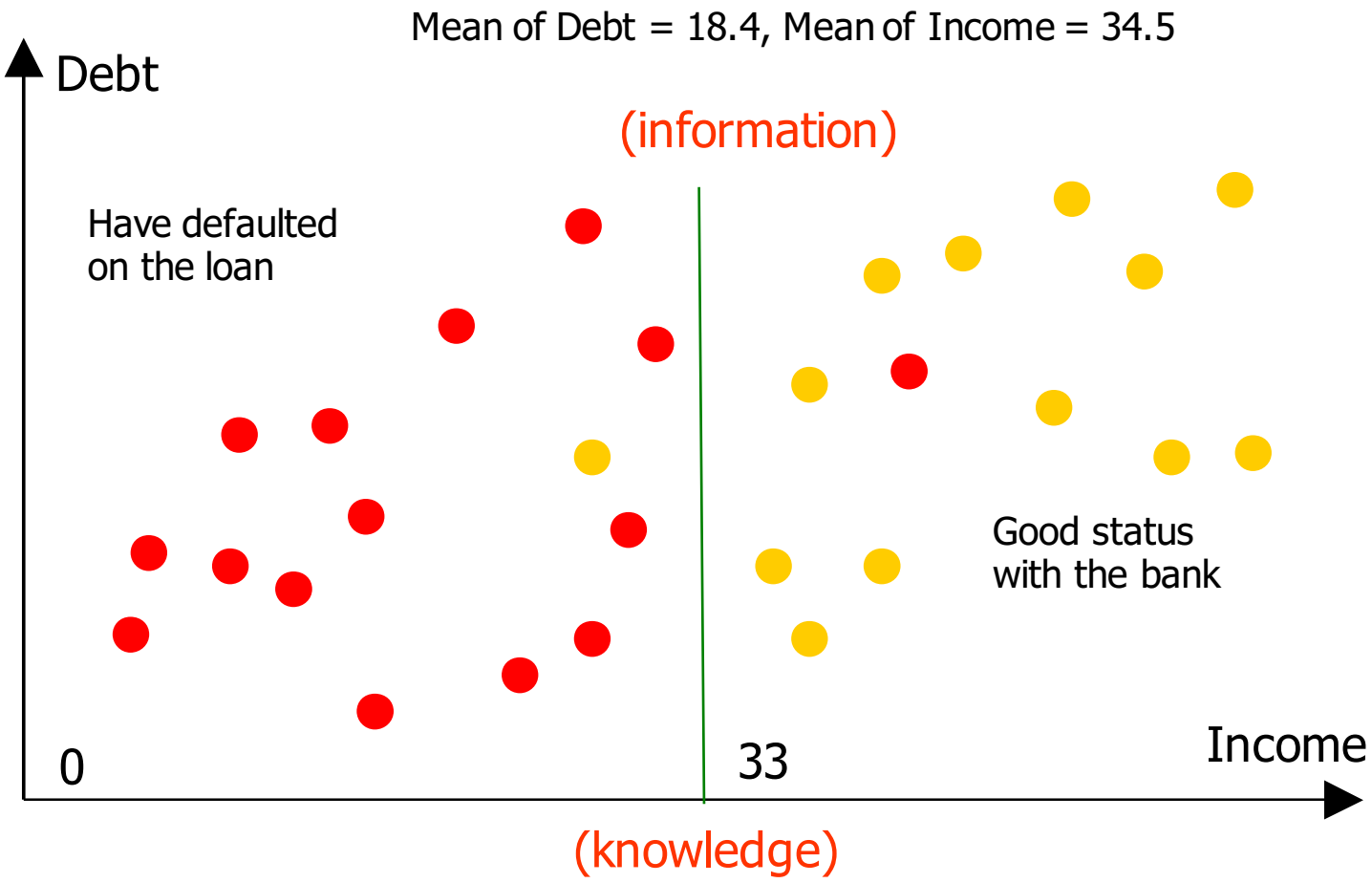
Obtaining by
- **Processing**

Un-interpreted signal

Number of cars counted on a road by hours, by days of the week, by months.

Obtaining by
- **Observing**
- **Measuring**
- **Collecting**

Data, information, and knowledge



US$ K
(income, debt)

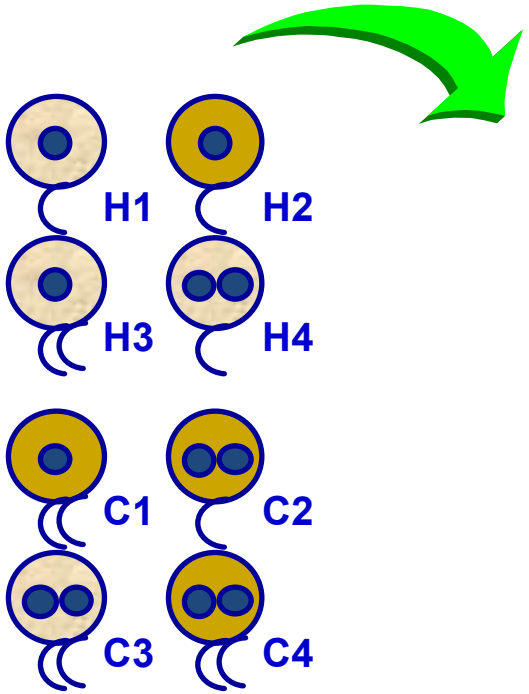
1. (5.6, 8.5)
2. (6.0, 13.0)
3. (11.0, 12.0)
4. (11.0, 19.0)
5. (13.5, 10.0)
6. (16.5, 20.0)
7. (17.5, 15.0)
8. (17.5, 5.0)
9. (22.5, 25.0)
10. (26.0, 7.5)
11. (30,0, 9.0)
12. (30.0, 18.0)
13. (30.0, 30.0)
14. (31.0, 14.0)
15. (32.5, 25.0)
16. (38.0, 12.0)
17. (41.0, 9.0)
18. (41.0, 22.0)
19. (43.5, 12.5)
20. (44.0, 27.5)
21. (45.0, 22.5)
22. (48.0, 28.0)
23. (52.5, 21.0)
24. (53.5, 32.0)
25. (54.0, 27.5)
26. (57.5, 18.0)
27. (59.0, 18.0)
28. (62.5, 32.5)
29. (63.0, 18.0)

34.5, 18.4

"if income < $33K, then the person has defaulted on the loan"

VIASM
Vietnam Institute for
Advanced Study in Mathematics

Dataset: cancerous and healthy cells



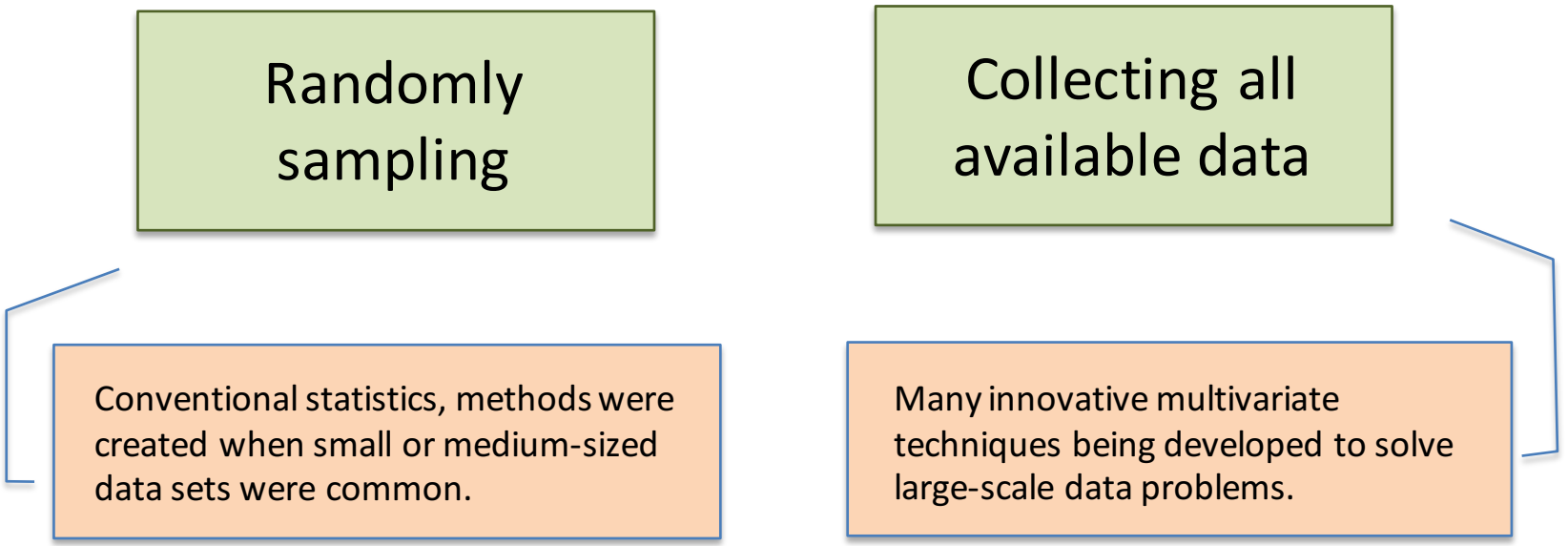
- color, #nuclei, #tails: *descriptive attributes*
- status: *class attribute*

	color	#nuclei	#tails	status
H1	light	1	1	healthy
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

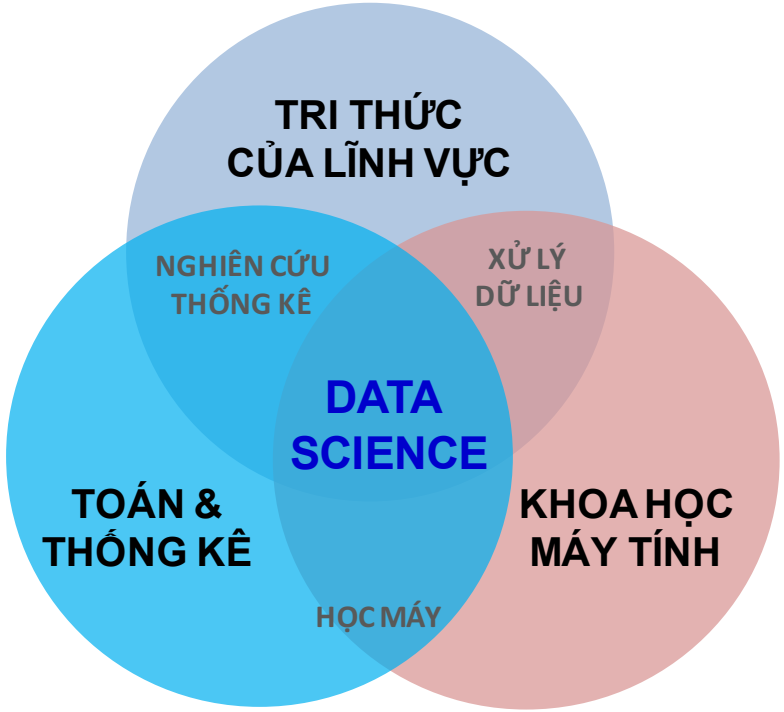
Object-Attribute relationship

How does people collect data?

- Observing, measuring, or collecting the **values of features** (features, attributes, properties, variables) of the **objects** under consideration.
- Two ways of collecting data

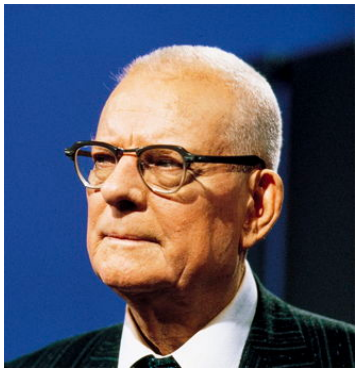


Khoa học dữ liệu

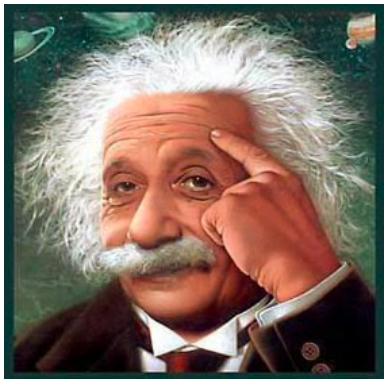


Khoa học về phân tích dữ liệu

"In God we trust.
All others bring
data".
"Ta tin Thượng đế.
Ngoài ra, là dữ
liệu".
W.E. Deming



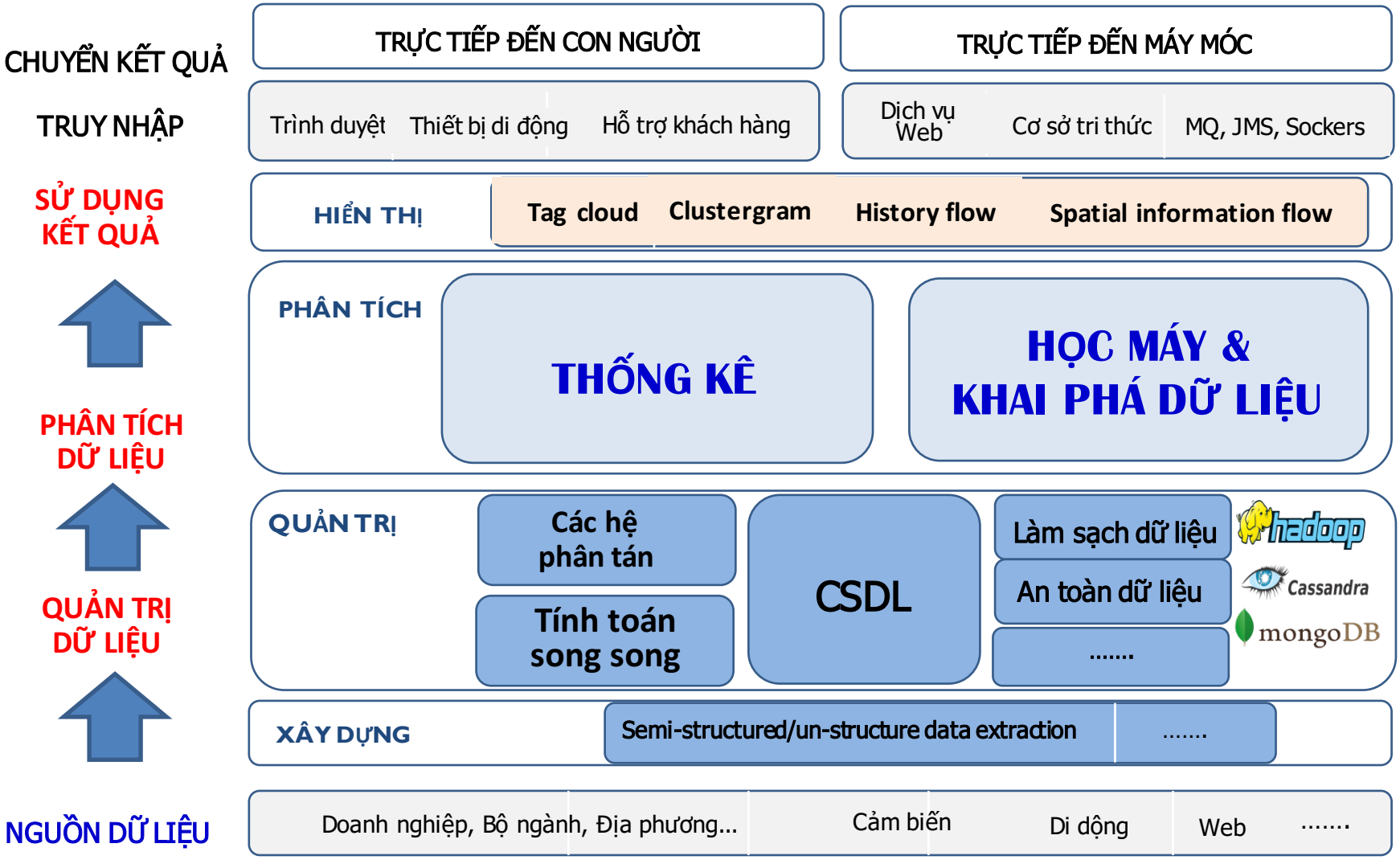
"We cannot solve problems
by using the same kind of
thinking we used when we
created them"
Ta không thể giải quyết các
vấn đề với chính cách nghĩ ta
đã dùng khi đặt vấn đề
Albert Einstein



Kết hợp của Toán học và Tin học là cốt lõi của khoa học dữ liệu

VIASM
Vietnam Institute for
Advanced Study in Mathematics

Một lược đồ của khoa học dữ liệu



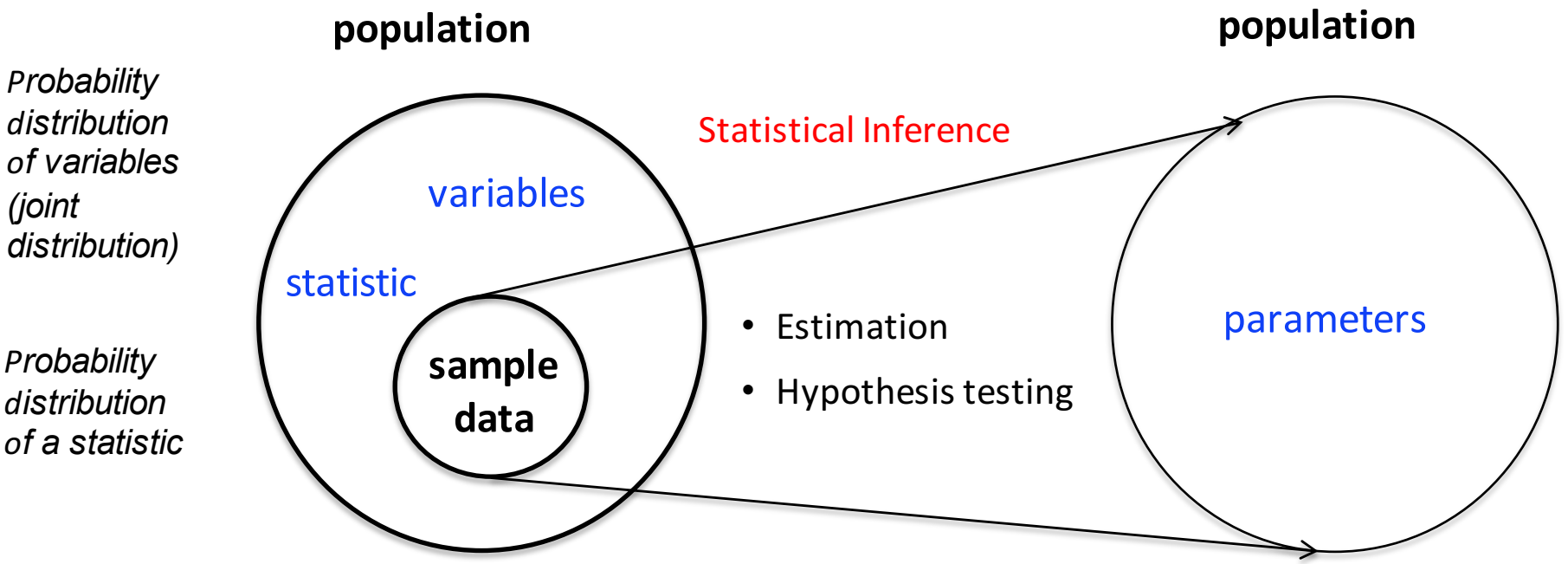
Source: WAMDM, Web group

Vietnam Institute for Advanced Study in Mathematics

Statistics

- **Statistics** provides mathematical methods and techniques to analyze, generalize and decide from the data.
- Main (traditional) **content**
  - **Descriptive statistics**: Probability distribution…
  - **Inferential statistics**: Estimation and hypothesis testing
- Experimental and observational **data**
  - Statistical data usually collected to answer predetermined questions (experiment design, survey design)
  - Mostly numerical data, few symbolic data.
- Methods were developed before having computer and for **small datasets**, for analyzing a single random variable.

Essence of statistics



Statistical inference is the ways of drawing conclusions about population parameters from an analysis of the sample data.

- A **parameter** is a *numerical feature* of the population, such as mean, proportion, standard deviation.
- A **statistic** is a single measure of some feature of a sample. It is defined as a *numerical-valued function* of the sample data. It is used to infer the corresponding population parameter.

VIASM
Vietnam Institute for Advanced Study in Mathematics

Multivariate analysis

- Simultaneously analyze the relationship of multiple random variables
- Testing hypothesis by data in **Confirmatory data analysis** (CDA) vs. producing hypotheses from data in **Exploratory data analysis** (EDA)
  - Factor analysis, PCA, Linear discriminant analysis
  - Regression analysis
  - Cluster analysis
- What we can see from conventional methods?
  - Poor results on large and complex data
  - Traditional methods are suitable for analyzing small datasets.
  - Price of storage and data processing are quickly decreasing.

Multivariate analysis

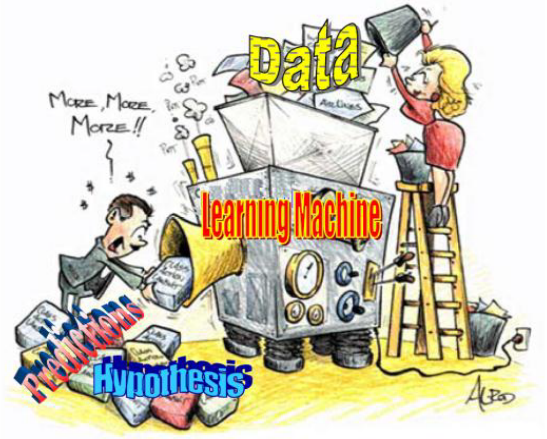
- Analytical methods were created for datasets with small or middle size and when computers were still weak.
- Multivariate data analysis is quickly changing due to computational techniques that are fast and effective.Various methods were newly developed for dealing with large scale problems (Pagerank of Google works with matrices of billion dimensions).

November 2016: Cray XK7 Titan computer, 17,590 TFlops, 560640 processors.

November 2016: Sunway TaihuLight
93,014 TFlops, 10,649,600 cores

Machine learning

- Mục đích của học máy là xây dựng các hệ máy tính có khả năng học tập như con người.
- *Given*
  - $\{(\boldsymbol{x}_i, y_i)\}, \boldsymbol{x}_i$ is description of an object in some space, $y_i \in \{C_1, C_2, \ldots, C_K\}$ or $y_i \in \mathbb{R}$ is viewed as label of $\boldsymbol{x}_i, i = 1, \ldots, n$.
  - Examples: Set of electronic medical records.
- *Find*
  - Function $p(y|\boldsymbol{x})$ for labeled data and $p(\boldsymbol{x})$ for unlabeled data.
  - Diagnosis or treatment regimen for a patient.



(Source: Eric Xing lecture)

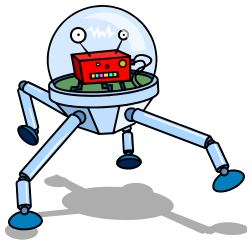
Statistics vs. Machine learning

Statistics

- Emphasize on formal statistical inference (estimation, hypothesis testing).
- Based on models for problems with small size and mostly with numerical data
- Statistics is an established science, conservatively changing culture and adapt to computational power.
- Trend to move to machine learning.

Machine learning

- Emphasize on prediction problems in high dimensionality and with symbolic data.
- In early days, the construction and use heuristics algorithms.
- Tend to base more on statistics, build statistical models underlying the algorithms.



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.

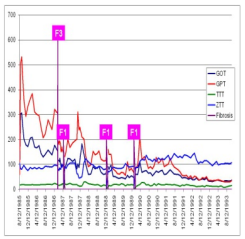
Data mining metaphor:
Extracting ore from rock

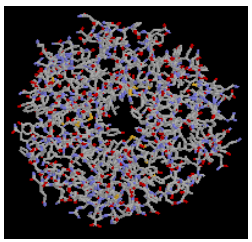
Statistics

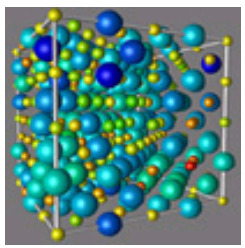
KDD

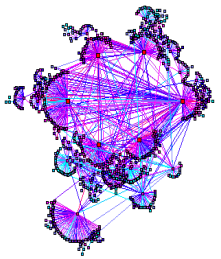
Databases Machine Learning

Large and unstructured real-life data

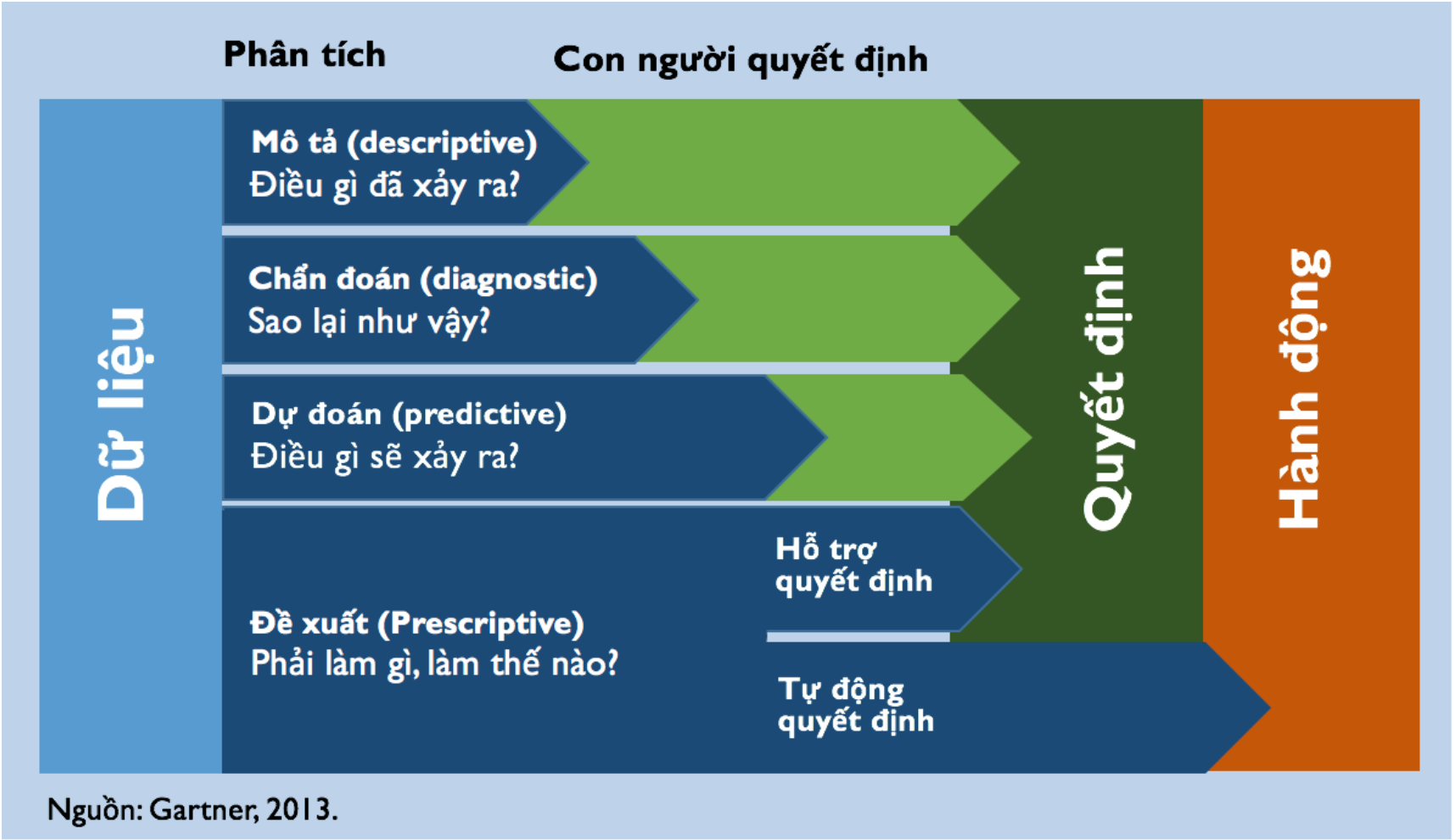








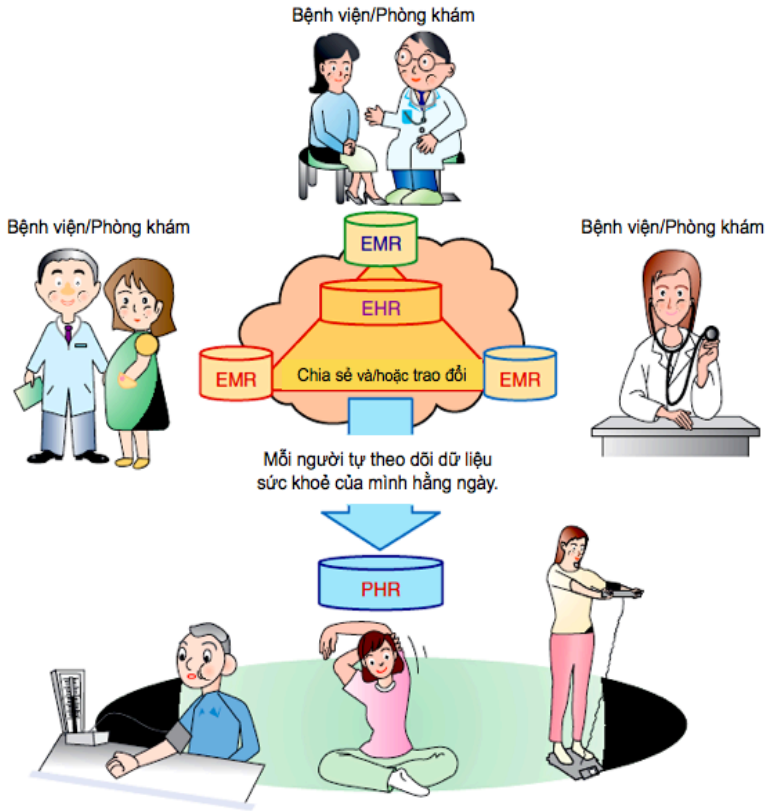
Khoa học dữ liệu: Giúp ra quyết định



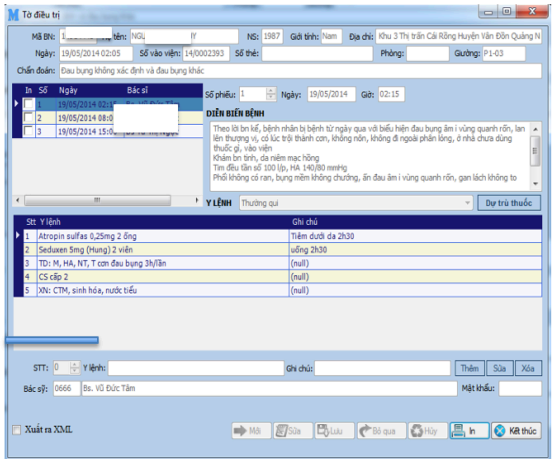
Nguồn: Gartner, 2013.

Data science is **the essential tool for using data**

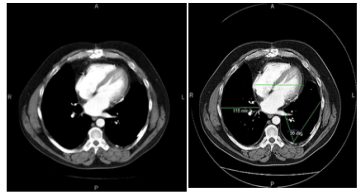
Bệnh án điện tử - nền tảng của e-health



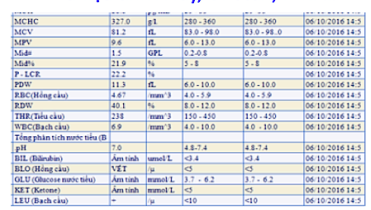
Bệnh án điện tử (**BAĐT**, electronic medical records – **EMRs**) là phiên bản số của bệnh án của mỗi lần nằm viện, **tạo** và **dùng** trong từng hệ thông tin bệnh viện (hospital information systems – **HIS**).



DỮ LIỆU LÂM SÀNG (VĂN BẢN)



Dữ liệu ảnh X-ray, CT scan, MRI ...



Dữ liệu xét nghiệm máu, hoá sinh, vi sinh

DỮ LIỆU CẬN LÂM SÀNG (SỐ & ẢNH)

Yasuo Ishigure, Trends, Standardization, and Interoperability of Healthcare Information, NTT Technical Review 2017

Thí dụ một số kiểu dữ liệu trong BAĐT

ICUSTAY

```
26,2538-10-29,4320,"N",1,1,"Y","Y",2538-10-26 03:18:00 EST,2538-10-29
16:25:00EST,58.95198,"adult",5107,"N","CCU","CCU","CCU","CCU",185.42,100.4,100.4,
100.4,16,5,16,5,1,5,
```

ICD_DISGNOSIS AGE

```
SUBJECT_ID,HADM_ID,SEQUENCE,CODE,DESCRIPTION
25,5726,1,"410.71","SUBENDOCARDIAL INFARCTION INITIAL EPISODE OF CARE"
25,5726,2,"250.11","DIABETES MELLITUS WITH KETOACIDOSIS TYPE I NOT STA"
25,5726,3,"414.01","CORONARY ATHEROSCLEROSIS OF NATIVE CORONARY ARTERY"
25,5726,4,"401.9","UNSPECIFIED ESSENTIAL HYPERTENSION"
```

DEMOGRAPHIC EVENTS DATA

```
SUBJECT_ID,HADM_ID,MARITAL_STATUS_ITEMID,MARITAL_STATUS_DESCR,ETHNICITY_ITEMID,ET
HNICITY_DESCR,OVERALL_PAYOR_GROUP_ITEMID,OVERALL_PAYOR_GROUP_DESCR,RELIGION_ITEMI
D,RELIGION_DESCR,ADMISSION_TYPE_ITEMID,ADMISSION_TYPE_DESCR,ADMISSION_SOURCE_ITEM
ID,ADMISSION_SOURCE_DESCR
25,5726,200050,"MARRIED",200083,"WHITE",200067,"PRIVATE",200081,"UNOBTAINABLE",20
0028,"EMERGENCY",200029,"EMERGENCY ROOM ADMIT"
```

MEDEVENTS DATA

```
SUBJECT_ID,ICUSTAY_ID,ITEMID,CHARTTIME,ELEMID,REALTIME,CGID,CUID,VOLUME,DOSE,DOSEUOM,SOLUTIONID,SOLVOLUME,
LUNITS,ROUTE,STOPPED
25,28,45,2538-10-26 04:30:00 EST,1,2538-10-26 04:57:00 EST,2691,1,0,8,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 04:30:00 EST,1,2538-10-26 05:00:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 04:45:00 EST,1,2538-10-26 04:57:00 EST,2691,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 04:45:00 EST,1,2538-10-26 05:00:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 05:00:00 EST,1,2538-10-26 05:23:00 EST,2049,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 05:00:00 EST,1,2538-10-26 05:23:00 EST,2049,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 05:15:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 05:15:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
25,28,45,2538-10-26 05:30:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,10,"Uhr",18,100,"ml","IV Drip",
25,28,142,2538-10-26 05:30:00 EST,1,2538-10-26 06:07:00 EST,2691,1,0,2,"mcgkgmin",13,100,"ml","IV Drip",
```

MEDURATIONS DATA

```
SUBJECT_ID,ICUSTAY_ID,ITEMID,ELEMID,STARTTIME,STARTREALTIME,ENDTIME,CUID,DURATION
25,28,45,1,2538-10-26 04:30:00 EST,2538-10-26 04:57:00 EST,2538-10-29 16:25:00 EST,1,5035
25,28,142,1,2538-10-26 04:30:00 EST,2538-10-26 05:00:00 EST,2538-10-29 16:25:00 EST,1,5035
25,28,45,1,2538-10-26 04:45:00 EST,2538-10-26 04:57:00 EST,2538-10-29 16:25:00 EST,1,5020
25,28,142,1,2538-10-26 04:45:00 EST,2538-10-26 05:00:00 EST,2538-10-29 16:25:00 EST,1,5020
25,28,45,1,2538-10-26 05:00:00 EST,2538-10-26 05:23:00 EST,2538-10-29 16:25:00 EST,1,5005
25,28,142,1,2538-10-26 05:00:00 EST,2538-10-26 05:23:00 EST,2538-10-29 16:25:00 EST,1,5005
```

POE-MED DATA

```
POE_ID,DRUG_TYPE,DRUG_NAME,DRUG_NAME_GENERIC,PROD_STRENGTH,FORM_RX,DOSE_VAL_RX,DOSE_UNIT_RX,FORM_VAL_
DISP,FORM_UNIT_DISP,DOSE_VAL_DISP,DOSE_UNIT_DISP,DOSE_RANGE_OVERRIDE
1930588,"BASE","D5W",,"250mL Bag",,"250","ml","250","ml",,,
1930589,"BASE","NS",,"500mL Bag",,"500","ml","500","ml",,,
1936709,"BASE","SW",,"100mL Bottle",,"100","ml","100","ml",,,
1929791,"MAIN","Aspirin","Aspirin","325mg Tab",,"325","mg","1","TAB",,,
1929796,"MAIN","Potassium Chloride","Potassium Chloride","20mEq Packet",,"20","mEq","1","PKT",,,
1929797,"MAIN","Atorvastatin","Atorvastatin","40mg Tab",,"80","mg","2","TAB",,,
1929819,"MAIN","Potassium Chloride","Potassium Chloride","20mEq Packet",,"40","mEq","2","PKT",,,
1930558,"MAIN","Potassium Chloride","Potassium Chloride","20mEq Packet",,"40","mEq","2","PKT",,,
1930691,"MAIN","Pantoprazole","Pantoprazole","40mg Tab",,"40","mg","1","TAB",,,
1931503,"MAIN","Calcium Gluconate","Calcium Gluconate","1g/10mL Vial",,"2","gm","2","VIAL",,,
1931745,"MAIN","Zolpidem Tartrate","Zolpidem Tartrate","5mg Tab",,"5-10","mg","1-2","TAB",,,
1931746,"MAIN","Acetaminophen","Acetaminophen","325mg Tab",,"325-650","mg","1-2","TAB",,,
...
```

POR-ORDER DATA

```
POE_ID,SUBJECT_ID,HADM_ID,ICUSTAY_ID,START_DT,STOP_DT,ENTER_DT,MEDICATION,PROCEDURE_TYPE,STATUS,ROUTE
,FREQUENCY,DISPENSE_SCHED,IV_FLUID,IV_RATE,INFUSION_TYPE,SLIDING_SCALE,DOSES_PER_24HRS,DURATION,DU
RATION_INTVL,EXPIRATION_VAL,EXPIRATION_UNIT,EXPIRATION_DT,LABEL_INSTR,ADDITIONAL_INSTR,MD_ADD_INST
R,RNURSE_ADD_INSTR
1929790,25,5726,28,2538-10-26 05:00:00 EST,2538-10-27 03:00:00 EST,2538-10-26 04:00:00
EST,"Insulin","IV Piggyback","Inactive (Due to a change order)","IV
DRIP","INFUSION",,,,,,,,"Ongoing",,"Enter on Label",,,"Fingersticks every hour  IV Drip Rate: 8
UNIT/HR","Specify blood glucose goal",
1929795,25,5726,28,2538-10-26 05:00:00 EST,2538-10-26 04:00:00 EST,2538-10-26 04:00:00 EST,"Potassium
Chloride","IV Piggyback","Discontinued","IV","ONCE","5",,,,,1,1,"Doses",,"Enter on
Label",,,,"CARDIAC MONITORING AND CENTRAL LINE ARE REQUIRED WHEN SELECTING CONCENTRATED PRODUCT
(20 mEq/50 mL).  20 mEq/50 ml preparations are given via central line only. Fluid restricted
patients may recieve 40 mEq in 500 ml NS or D5W. No more than 60 mEq placed in one liter of fluid
per BIDMC policy. ","Cardiac monitoring and central lines are required for rates > 10 mEq/hr."
```

NOTEEVENTS DATA

```
",,25,5726,28,0,2538-10-26 07:51:00 EST,2538-10-26 08:33:00 EST,1807,"N",1,"Nursing/Other","NURSING PROGRESS NOTE","
NURSING PROGRESS NOTE
58 Y/O MALE ADMITTED FROM [**Hospital1 2**] ER (TRANSFERED FROM [**Hospital6 110**]).  HE INITIALLY PRESENTED
TO [**Hospital6 110**] WITH C/O N/V, DIZZINESS.  HE IS S/P INSULIN PUMP INSERTION IN [**2538-5-6**].  HIS PUMP FAILED
ON SATURDAY AND BEGAN FEELING POORLY.  HE WAS ADMITTED WITH A BLOOD GLUCOSE > 575.  HE ALSO HAD ST
CHANGES ON EKG.  HE WAS TREATED WITH IV LOPRESSOR, INTEGRILLIN, IV NS, INSULIN.  HE REFUSED ASA STATING IT
MAKES HIS STOMACH UPSET.  ADMITTED TO CCU FOR R/O MI PROTOCOL.

This is a 58 yr old male Pt who presented to [**Hospital6 **] with C/O N/V & dizziness- He had an insulin pump inserted in 6/04 & on
Saturday [**10-25**] it failed- blood sugar was > 500- Also, his EKG showed new ST depressions (no C/O CP & cardiac enzymes negative)-
Pt was transferred to [**Hospital1 2**] EW on integrilin & insulin gtts for further care- Pt was admitted to CCU- R radial A line was placed-
Pt developed a sinus arrythemia HR 40-70's with hypotension (SBP 60-70's)- atropine given for ? bradycardia induced hypotension, IV fluids
wide open & dopamine gtt started- EKG SA HR 50-70's with return of ST-T waves changes in lateral leads- PA line inserted into R IJ- RA 8-
PAP 42/22-PCWP 15-16- decision was made to cath Pt due to persistent hypotension Cardiac cath revealed moderately severe single vessel
CAD (LCx large vessel proximal 60-70%) normal LV systolic function- no intervention done- ? elective stent LCx when stable- CO high with
low SVR- ? sepsis

CV-R/I MI. HR 70-80NSR, BP by R radial Aline 110-140/60-70. ASA, plavix (loaded w/ 300mg this am) cont., lopressor 12.5mg bid added.
No c.o CP, weakness, dizz. PA line- CVP 8-10, PA 28-38/16-18, CO [**7-15**], SVR 500. Has received ~10liters of IVF over 48hr, u/o 3000
over same time. R femoral Aline d/c @ 12n w/o complication by Card fellow, site C&D w/ transparent dsg, no hematoma, no oozing. Pulses
dpl-1+, baseline. Endo/Fluids- IDDM on insulin gtt @2-3u/hr w/ small and improving po intake.FS 92-152. IVF D5.45NS @ 100cc/h (dec'd
from 150/hr this am). U/O 80-120/hr clear urine. + 2500 for day.

CCU Nursing Progress Note-7a-7p 58 y/o male admitted [**10-25**] w/ N/V/dizz, IDDM w/ failed pump, FS 576 to [**Hospital6 **], EKG
changes. Placed on insulin gtt, IVF and tx to [**Hospital1 2**]. Over w/e,hypotensive- Dopa and Levo; PA line placed w/ High CO, low SVR;
cathed, RCA 70% stenosis, RI MI; DKA. Much improved overnight and today. Anion gap now closed. Heparin, R fem Aline d/c. Cont INS
gtt, IVF, antibx. Plan for Stent of RCA [**10-28**]. NPO p MN.
Neuro- A&O x3, MAE, much less irritable w/ cardiac explanation/education by MD/RN CCU team. Able to assist w/ position change. To be
OOB this evening when PA line D/C.
```

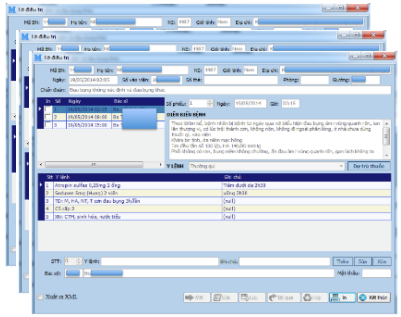
CHARTDURATION DATA

```
SUBJECT_ID,ICUSTAY_ID,ITEMID,ELEMID,STARTTIME,STARTREALTIME,ENDTIME,CUID,DURATION
25,28,781,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1536,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1535,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1534,0,2538-10-26 03:59:00 EST,2538-10-26 09:29:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1532,0,2538-10-26 03:59:00 EST,2538-10-26 09:29:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1529,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1525,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1523,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1522,0,2538-10-26 03:59:00 EST,2538-10-26 09:29:00 EST,2538-10-29 16:25:00 EST,1,5066
25,28,1162,0,2538-10-26 03:59:00 EST,2538-10-26 04:30:00 EST,2538-10-29 16:25:00 EST,1,5066
```

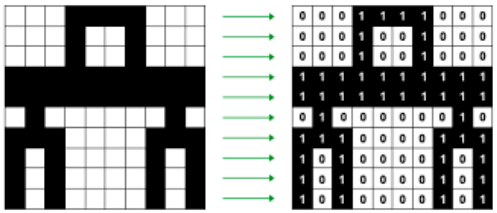
Học phác đồ điều trị từ BAĐT

Phương pháp

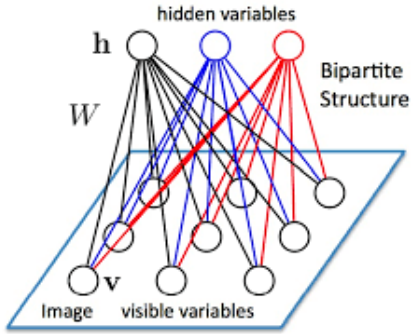
BAĐT với nhiều kiểu dữ liệu



Mã hoá từng kiểu dữ liệu

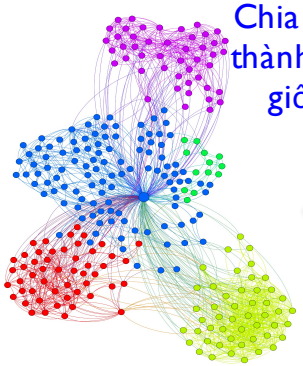


Đưa dữ liệu vào Restricted Boltzmann Machine



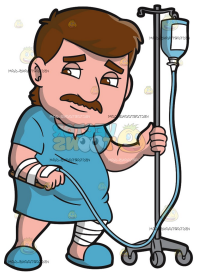
Biến đổi BAĐT thành các vectors nhị phân

Chia bệnh phân thành các nhóm giống nhau



Học phác đồ điều trị chung cho từng nhóm

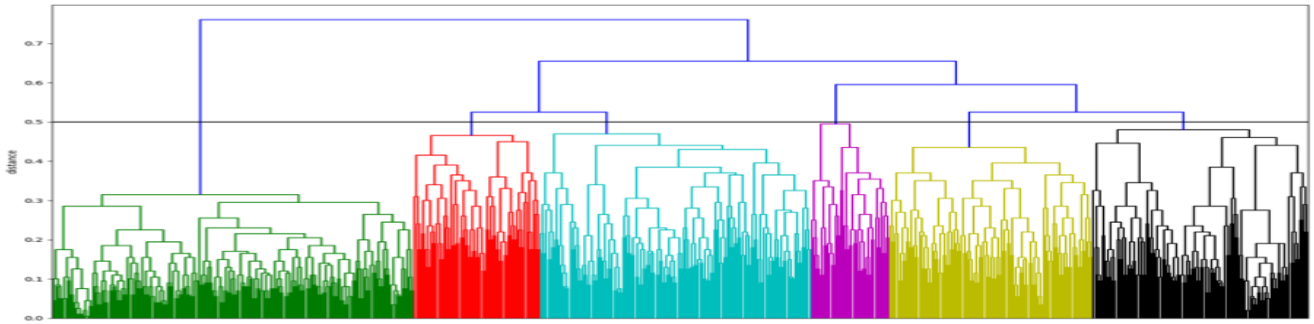
Dùng phác đồ đã học để gợi ý điều trị cho bệnh nhân

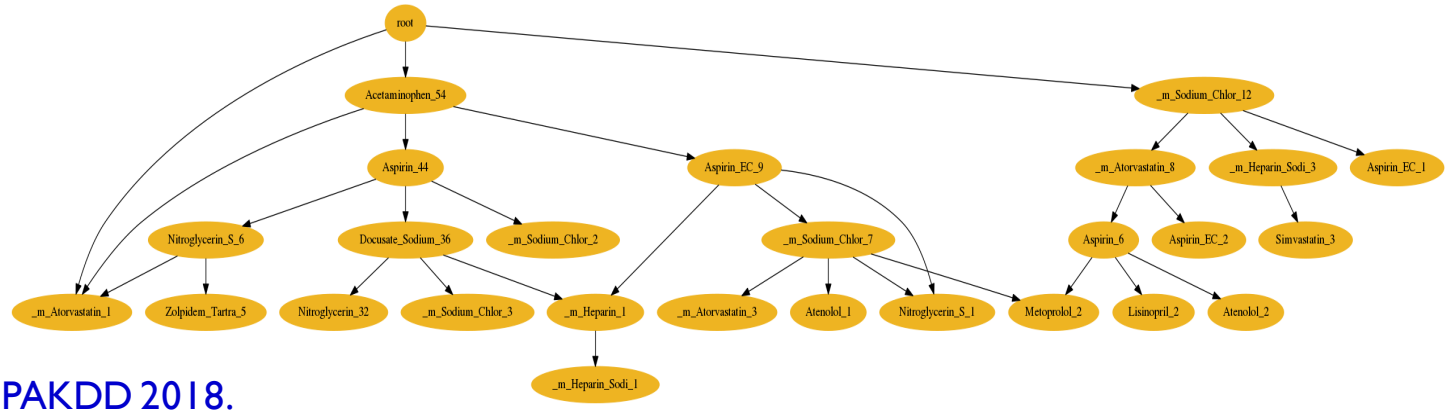


Học phác đồ điều trị từ BAĐT

Chia nhóm & Học phác đồ điều trị cho từng nhóm

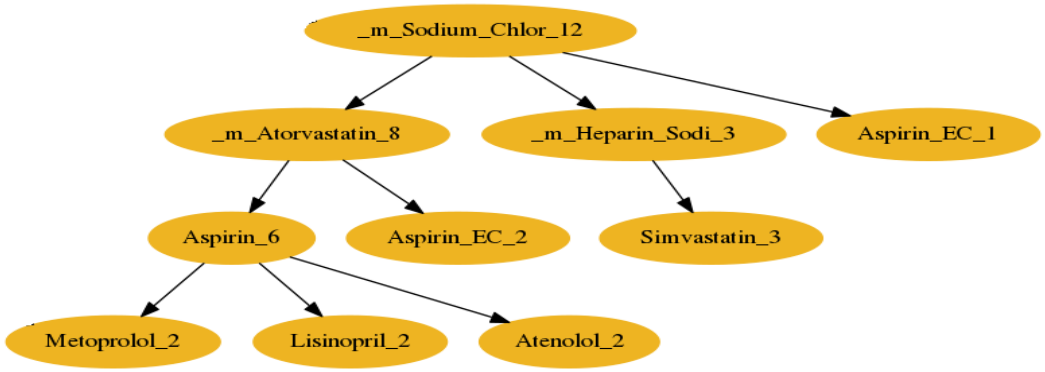
Dữ liệu hỗn hợp của bệnh nhân được đưa qua một hệ Restricted Boltzmann Machines mở rộng để biến đổi thành các biểu diễn đồng nhất bởi các biến ẩn, từ đó thực hiện việc phân nhóm và học phác đồ điều trị cho các nhóm, và dùng chúng để gợi ý quyết định điều trị.

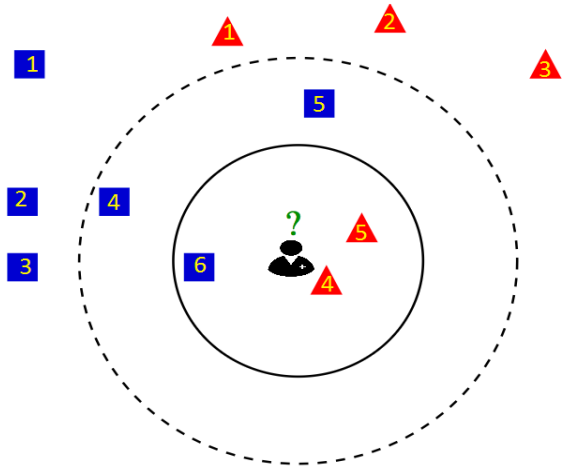


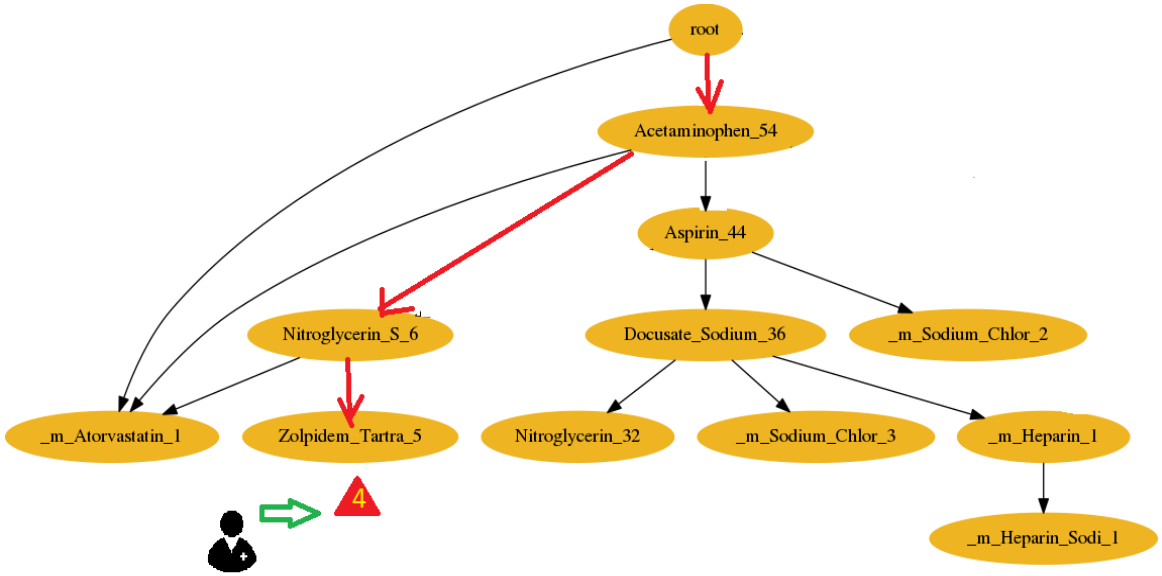


Hoang H., Ho T.B., PAKDD 2018.

Học phác đồ điều trị từ BAĐT







Accuracy (with depth of tree: 4)

$m_{strict} = 51.7$

$m_{patial} = 0.724$

Outline

1. What is data science?
2. Principles of data science
3. DSLab's data science lectures

Principles of data science?

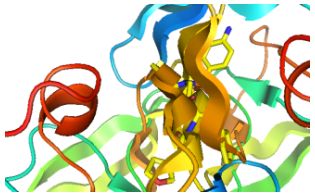
Principle = a basic idea or rule that explains or controls how something happens or works (Cambridge Dict.)

1. Data type and structure
2. Process
3. Methods
4. Model selection

Data types and structure vs. methods

Data types and structures

- 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.

Data types and structure vs. methods

Data types and structures

- Flat data tabl
- Relational
- Tempora
- Tran
- Mul
- Genon
- Mate
- Textua
- Web data
- etc.

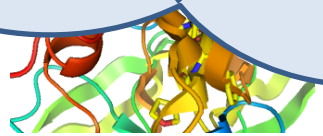
Mining tasks and methods

rediction

alysis

- Summarization
- etc.

Most methods were developed for data in the table format. When data are not represented in this form, we need to convert them to the table or to adapt the methods.



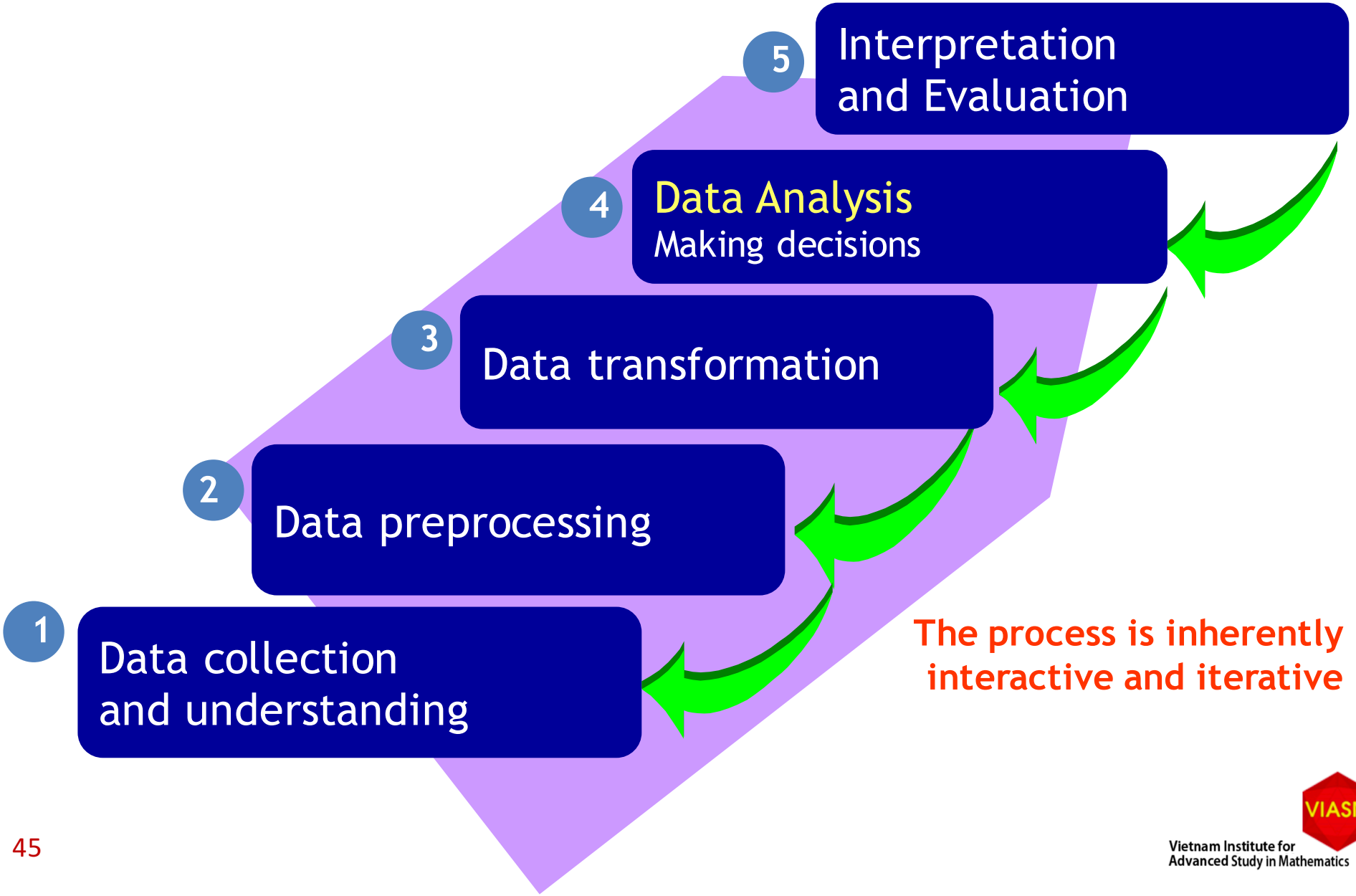
Why caring about data types?

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

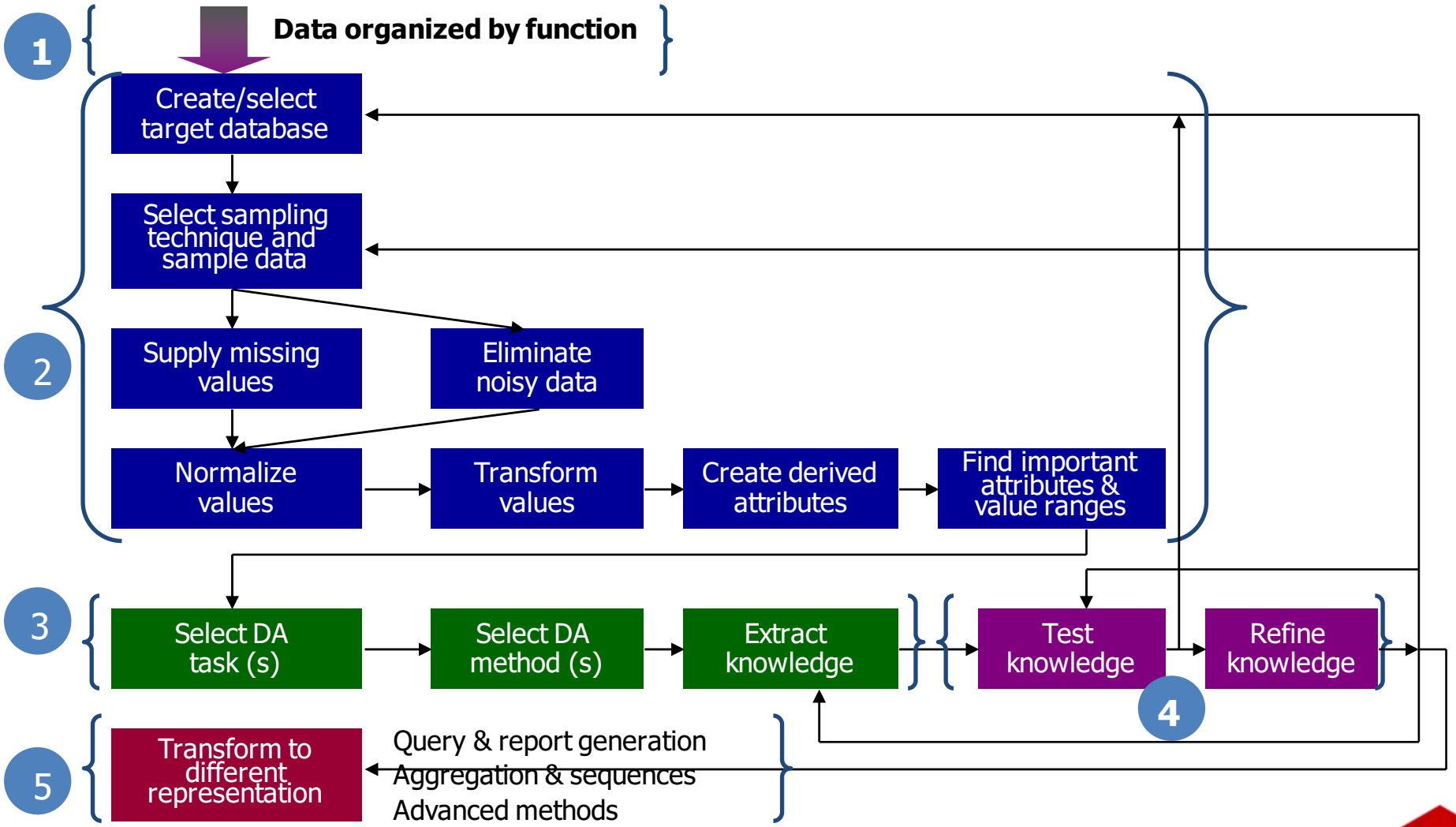
Often matrix-based computation (multivariate data analysis)

The data analytics process



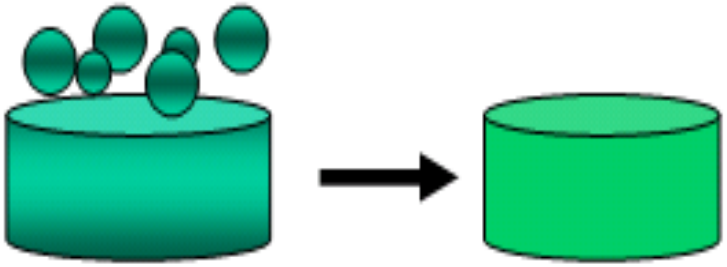
The process is inherently interactive and iterative

The data analysis process

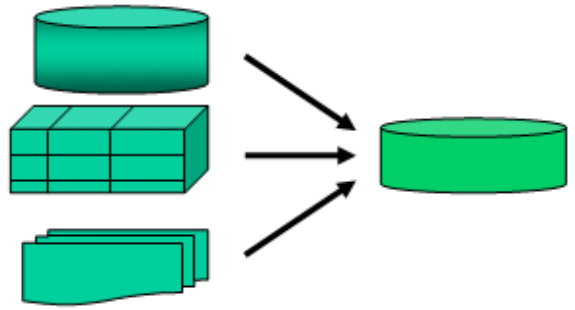


Vietnam Institute for Advanced Study in Mathematics

Major tasks in data preprocessing

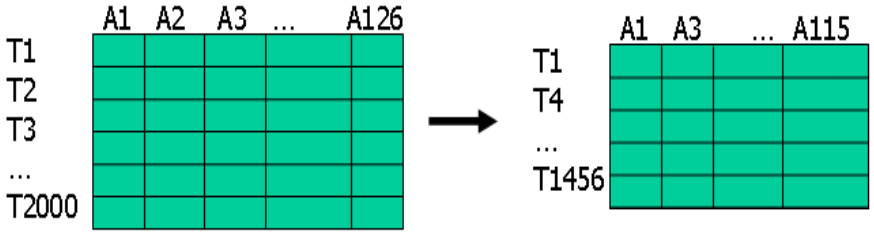


1 Data cleaning

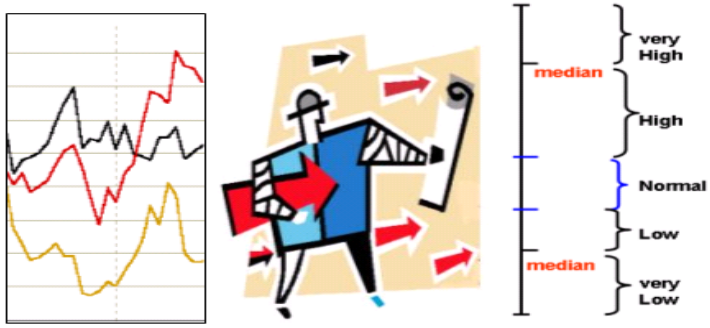


-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

2 Data integration and transformation

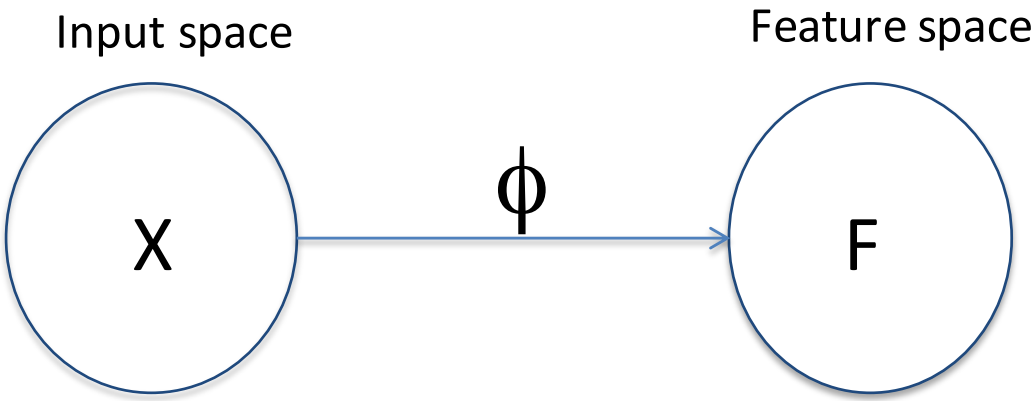


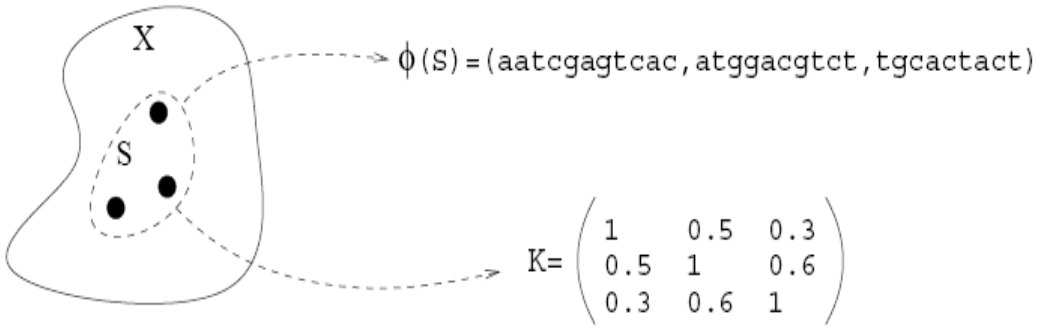
3 Data reduction
(instances and dimensions)



4 Data discretization

Data transformation





ϕ: X → F where the problem can be solved in F

X is the set of all oligonucleotides, S consists of three oligonucleotides, and S is represented in F as a matrix of pairwise similarity between its elements.

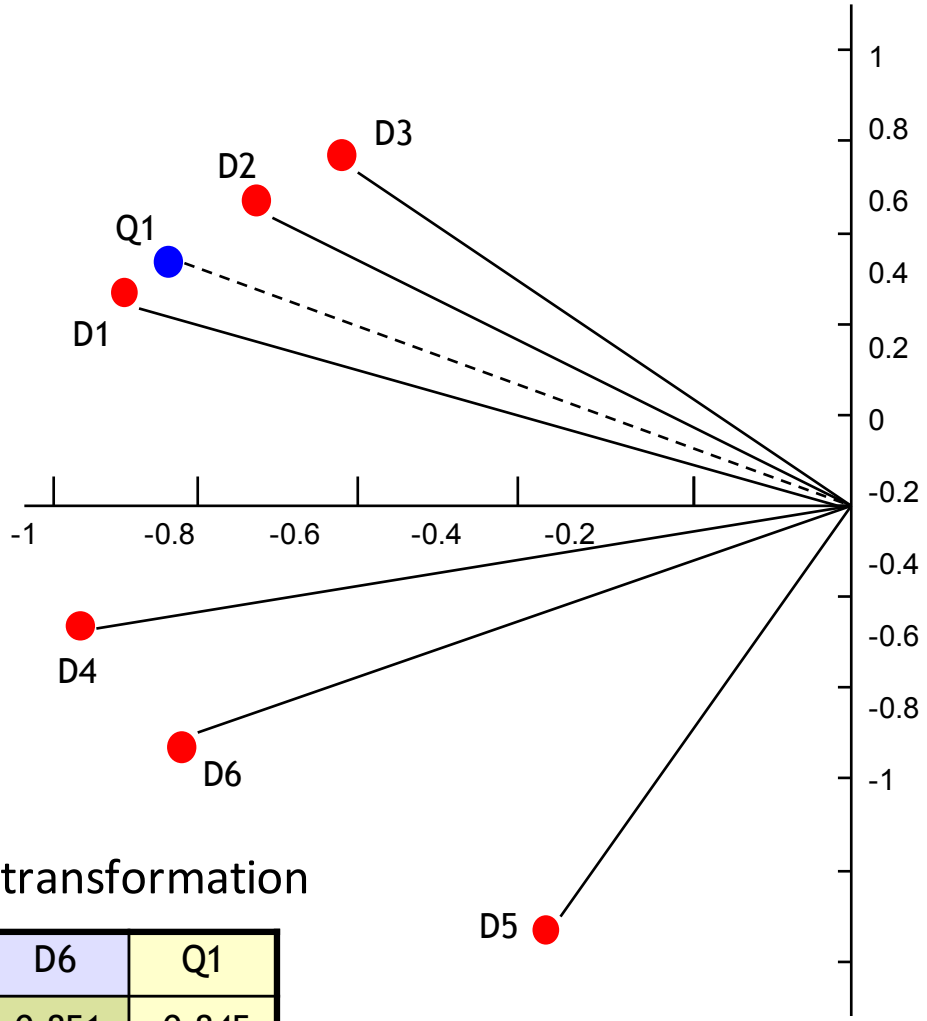
VIASM
Vietnam Institute for Advanced Study in Mathematics

Data transformation

Example: Latent semantic indexing

- LSI (Deerwester, 1990) clusters documents in the reduced-dimension semantic space according to word co-occurrence patterns.
- Query Q1 shares common words with D4 and D6 but Q1 is more closed to D3 in meaning.

	D1	D2	D3	D4	D5	D6	Q1
rock	2	1	0	2	0	1	1
granite	1	0	1	0	0	0	0
marble	1	2	0	0	0	0	1
music	0	0	0	1	2	0	0
song	0	0	0	1	0	2	0
band	0	0	0	0	1	0	0



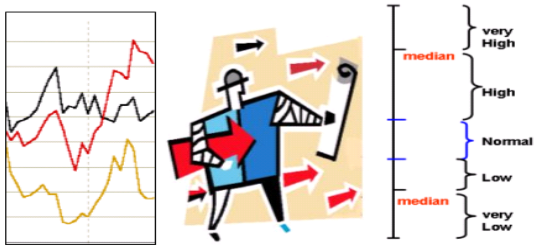
Explaining the meaning after transformation

	D1	D2	D3	D4	D5	D6	Q1
Dim. 1	-0.888	-0.759	-0.615	-0.961	-0.388	-0.851	-0.845
Dim. 2	0.460	0.652	0.789	-0.276	-0.922	-0.525	0.534

Conversion between data types

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

Discretization

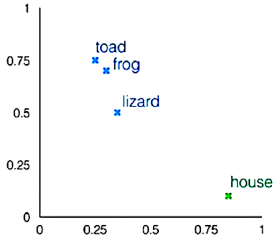


Continuous → Discrete

Rich structure → Poor structure

Word embedding

frog	[0.30 0.70]
toad	[0.25 0.75]
lizard	[0.35 0.50]
house	[0.85 0.10]



Discrete → Continuous

Poor structure → Rich structure

Word embedding

- How to measure the similarity between words?
- How to convert words from discrete spaces into continuous spaces?
- "You shall know a word by the company it keeps" (J.R. Firth 1957)

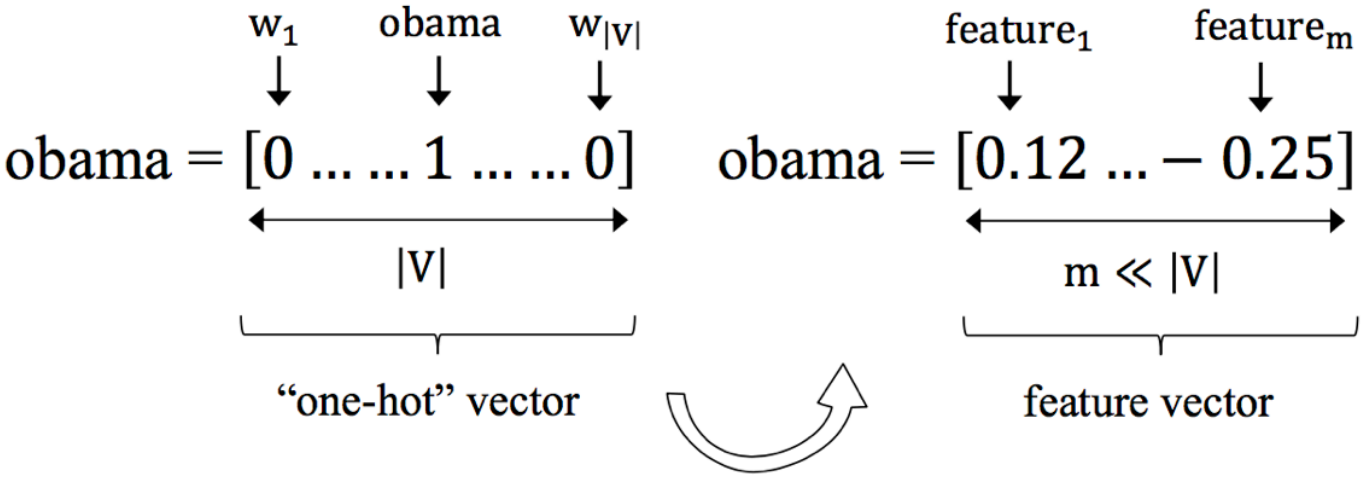
government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

VIASM
Vietnam Institute for Advanced Study in Mathematics

Word embedding

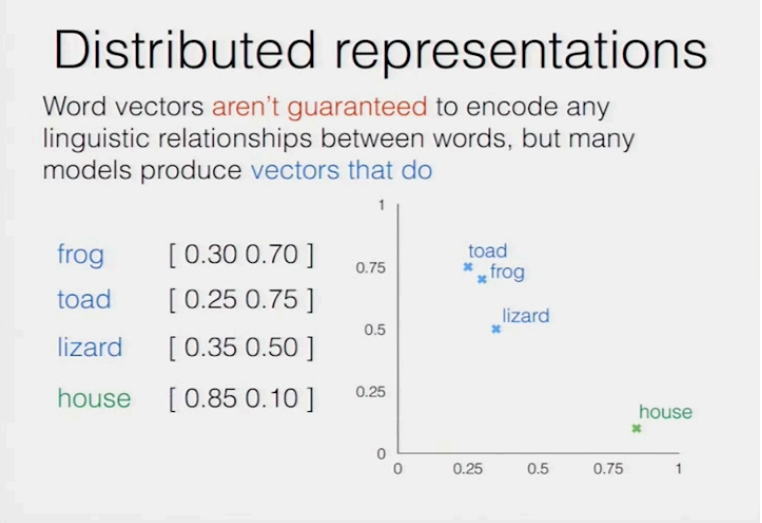
- Each unique word is mapped to a point in a real continuous m-dimensional space
  - $w_i \in V \xrightarrow{mapping\ C} \mathbb{R}^m$
- Typically, $|V| > 10^6, 100 < m < 500$



Vietnam Institute for Advanced Study in Mathematics

Word embedding

- Word2vec: by Mikolov, Sutskever, Chen, Corrado and Dean at Google, NAACL 2013.
- Takes a text corpus as input and produces the word vectors as output
- word meaning and relationships between words are encoded spatially
- two main learning algorithms in word2vec:
  continuous bag-of-words
  and continuous skip-gram



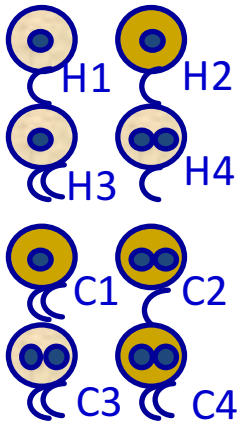
https://www.youtube.com/watch?v=RyTpzZQrHCs

Machine learning: View by data
Labelled vs. Unlabelled 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 is some property of x_i, if y_i is not available data is **unlabelled**, otherwise **labelled**.

Find: a function f that characterizes $\{x_i\}$ (**unsupervised learning**) or that $f(x_i) = y_i$ (**supervised learning**) [in between: **reinforcement learning**]



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

Machine learning: View by nature of methods

Tribes	Origins	Master Algorithms
Symbolists	Logic, philosophy	Inverse deduction
Evolutionaries	Evolutionary biology	Genetic programming
Connectionists	Neuroscience	Backpropagation
Bayesians	Statistics	Probabilistic inference
Analogizers	Psychology	Kernel machines

The five tribes of machine learning, Pedro Domingos

Symbolists

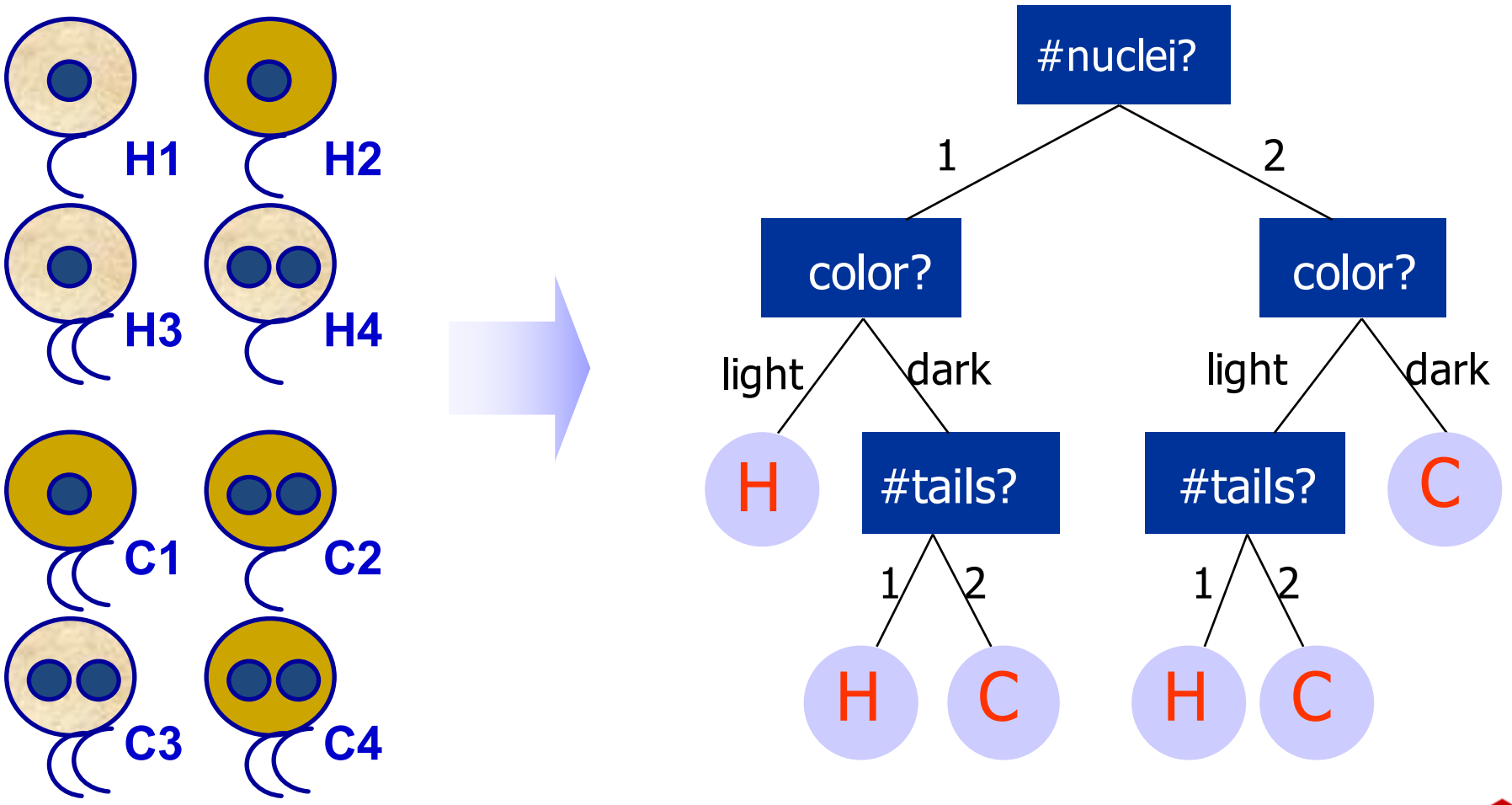
Tom Mitchell

Steve Muggleton

Ross Quinlan

Deduction	Induction
Socrates is human	Socrates is human
+ Humans are mortal	+ ?
= ?	= Socrates is mortal

Classification with decision trees

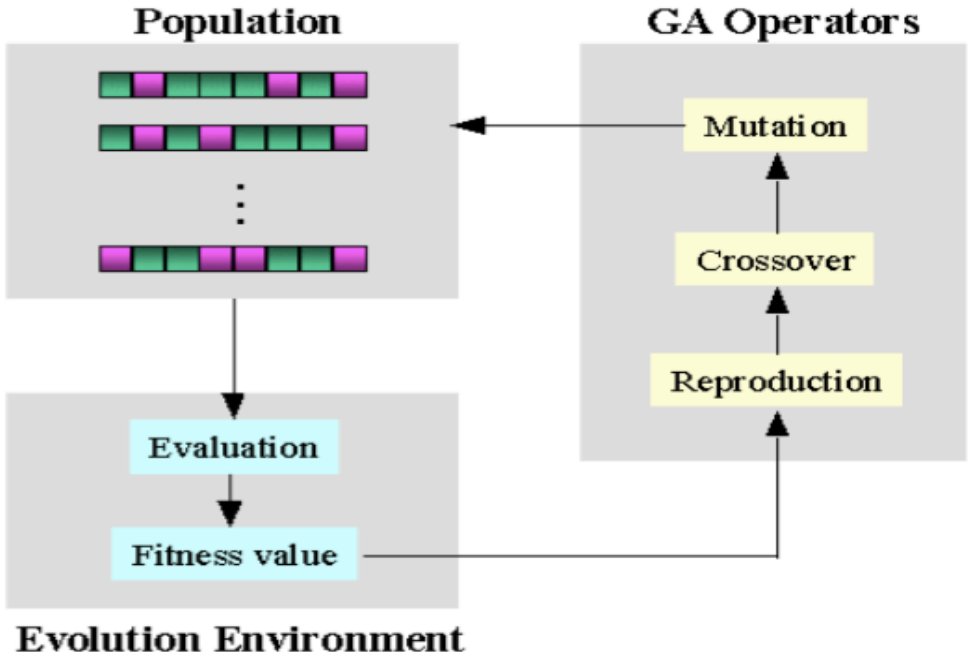


Evolutionaries

John Koza

John Holland

Hod Lipson



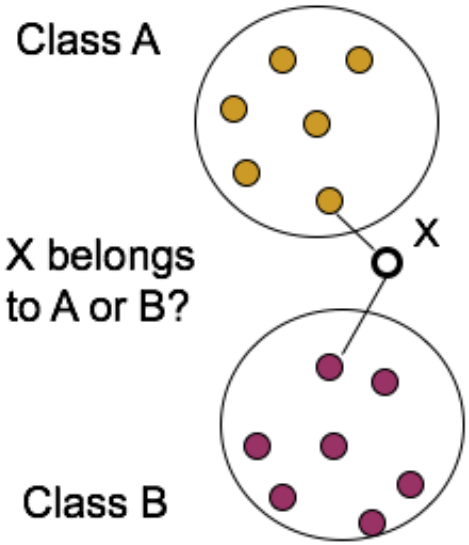
Analoziger

Peter Hart

Vladimir Vapnik

Douglas Hofstadter

- **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

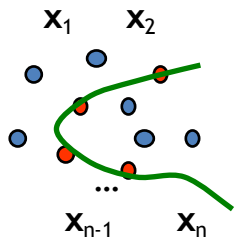


Vietnam Institute for Advanced Study in Mathematics

Kernel methods

The basic ideas

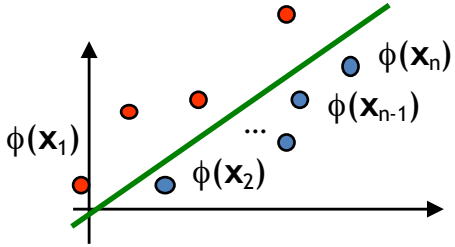
Input space X



inverse map ϕ^{-1}

$\phi(x)$

Feature space F

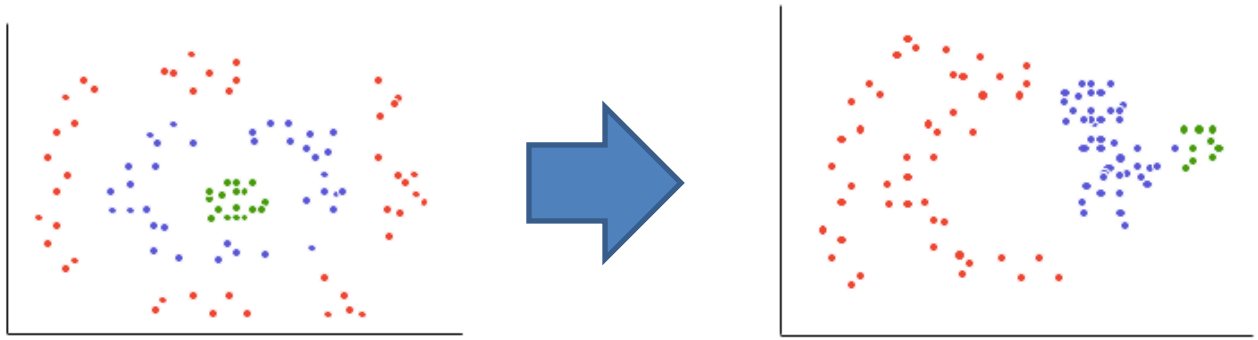


$k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$

kernel function k: XxX → R

Kernel matrix K_{nxn}

kernel-based algorithm on K
(computation done on kernel matrix)



Connectionists

Yann LeCun

Geoff Hinton

Yoshua Bengio

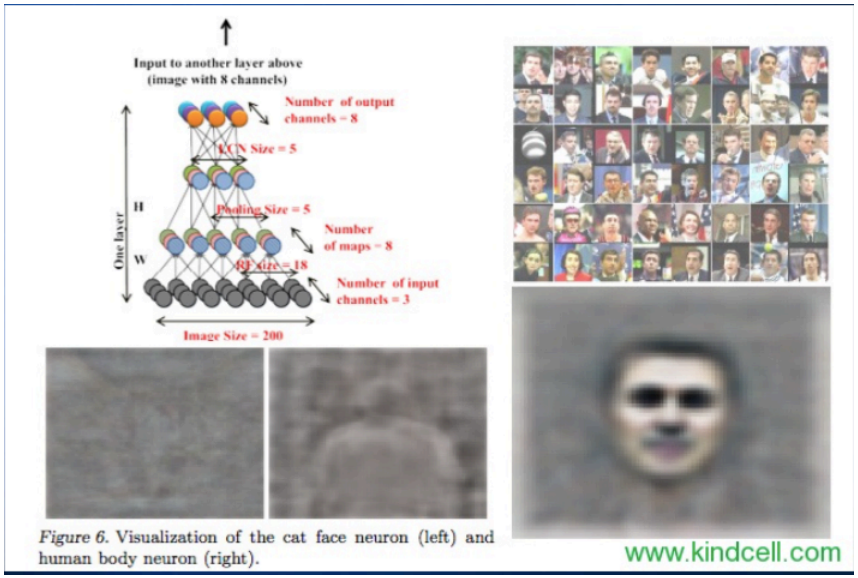
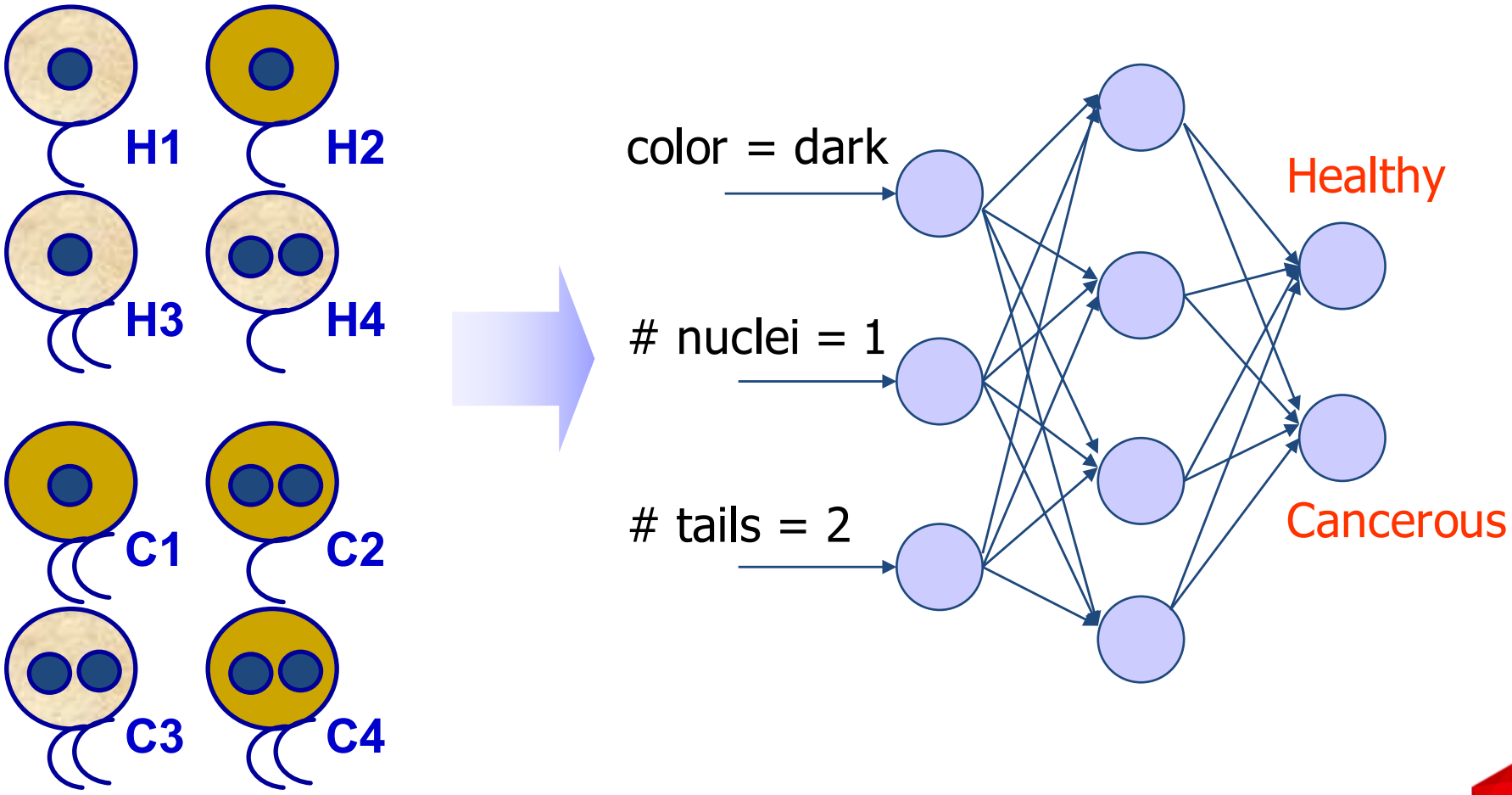
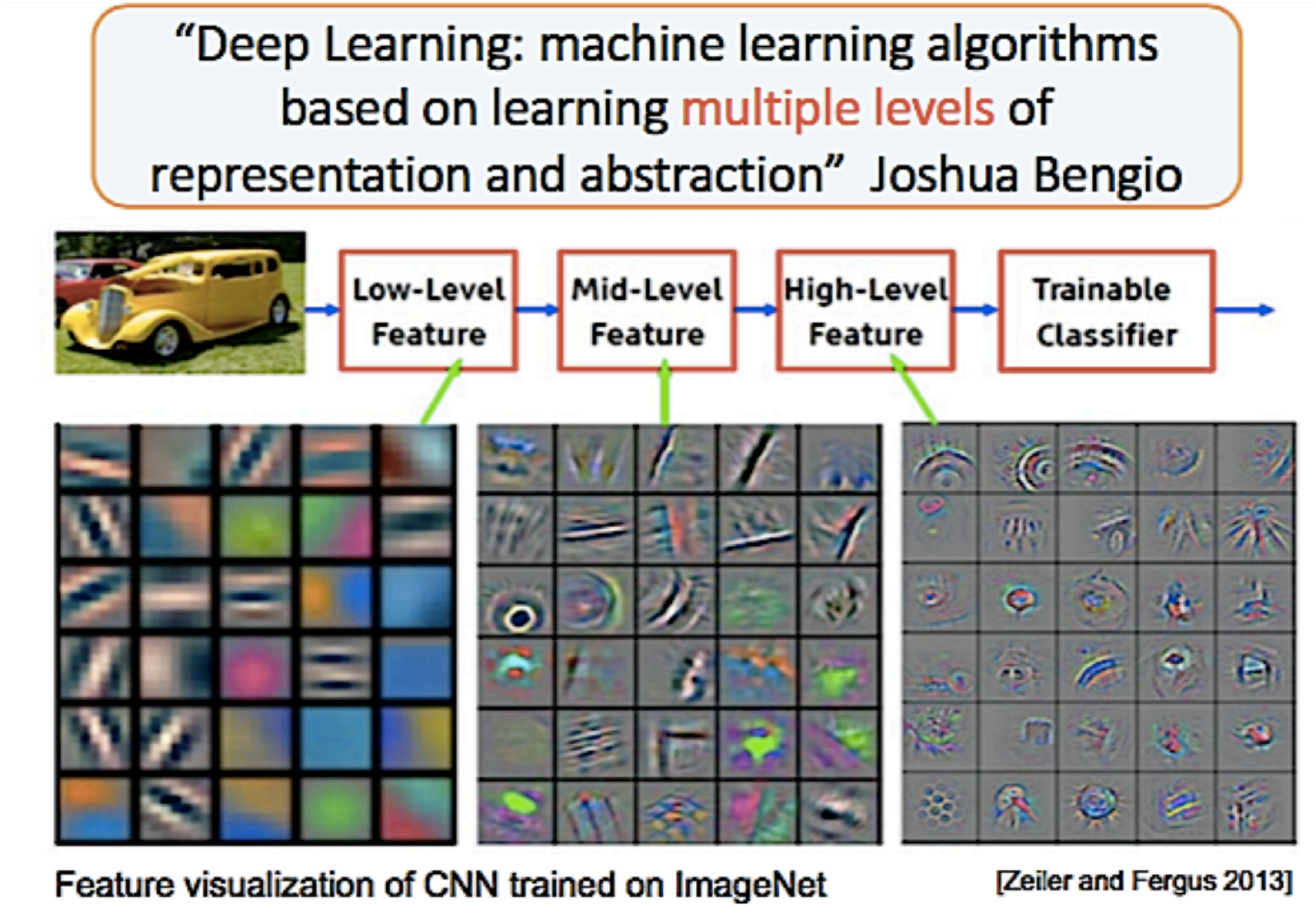


Figure 6. Visualization of the cat face neuron (left) and human body neuron (right).

Classification with neural networks



Deep learning

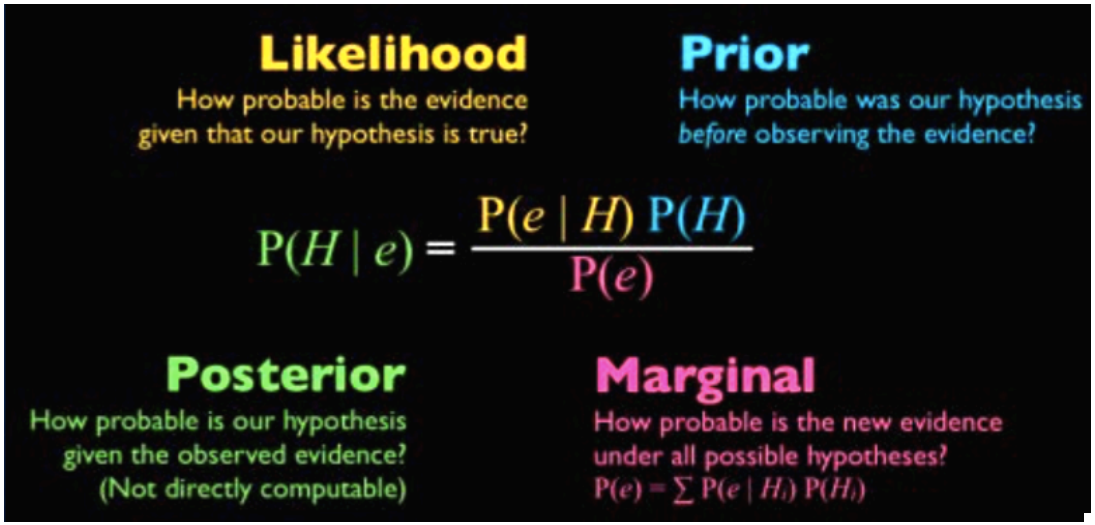


Bayesians in machine learning

David Heckerman

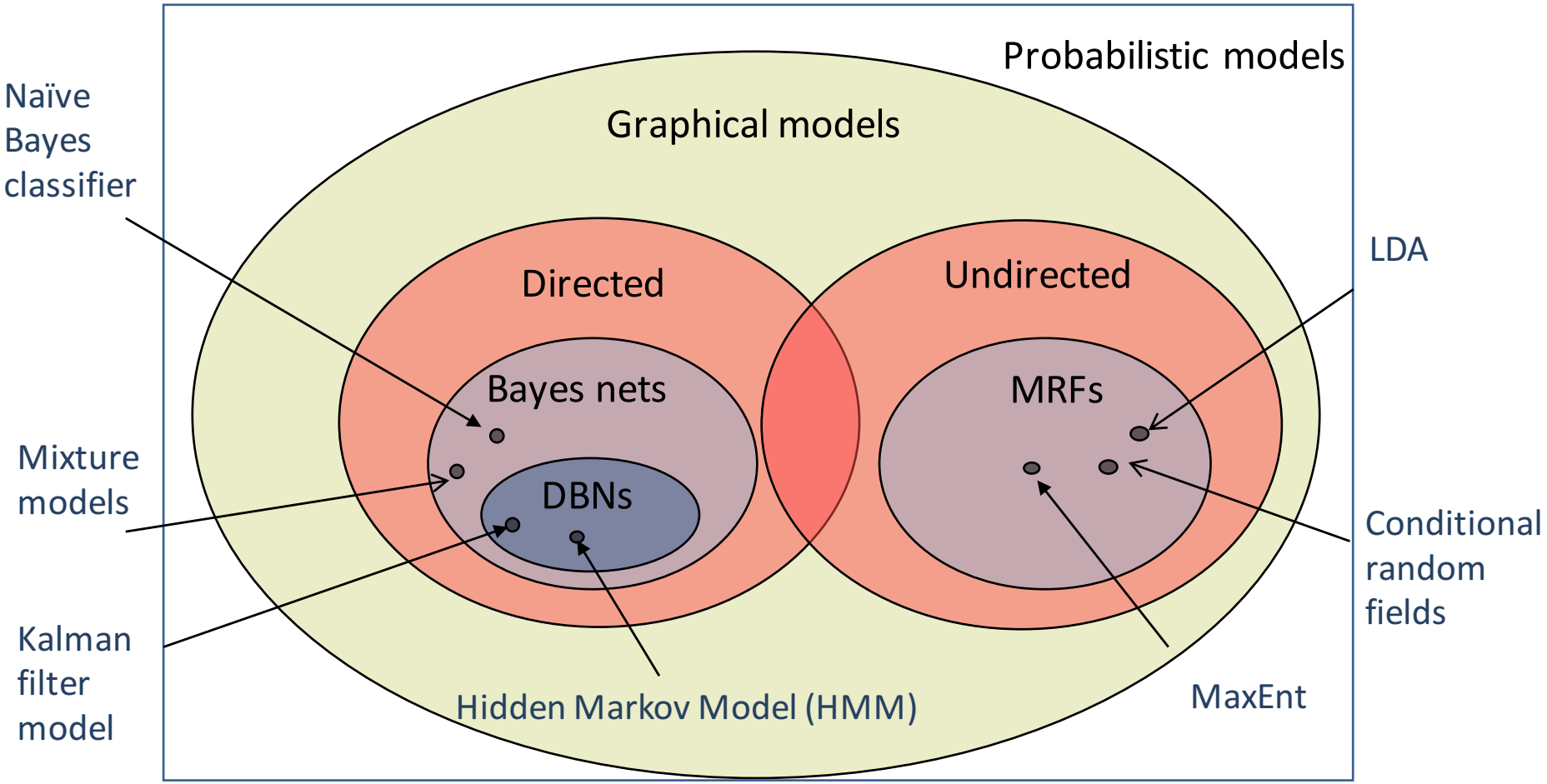
Judea Pearl

Michael Jordan

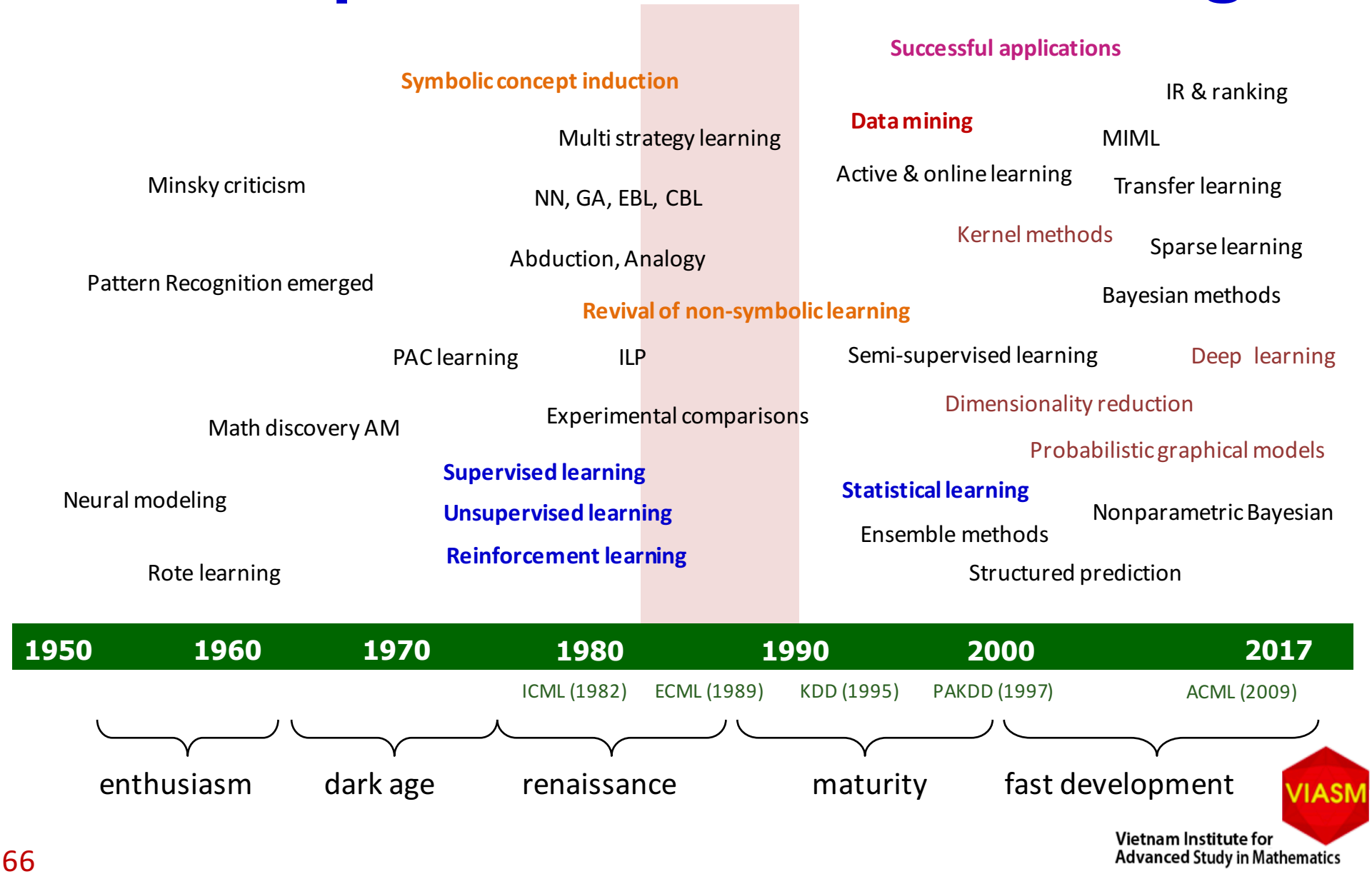


Probabilistic graphical models

Instances of graphical models



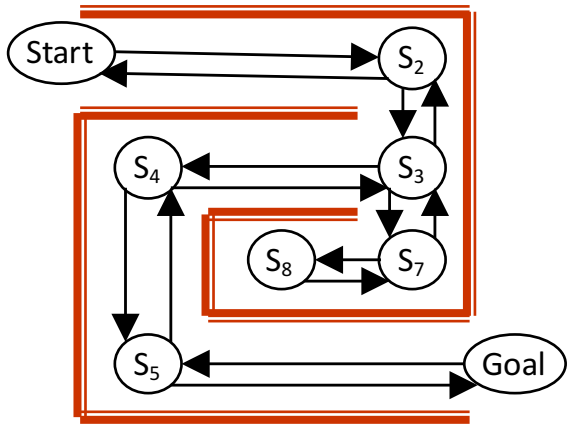
Development of machine learning



Vietnam Institute for Advanced Study in Mathematics

Reinforcement learning

Concerned with how an agent ought to take actions in an environment so as to maximize some cumulative reward. (… một tác nhân phải thực hiện các hành động trong một môi trường sao cho đạt được cực đại các phần thưởng tích lũy)



- The basic reinforcement learning model consists of:
  - a set of environment states S;
  - a set of actions A;
  - rules of transitioning between states;
  - rules that determine the *scalar immediate reward* of a transition;
  - rules that describe what the agent observes.

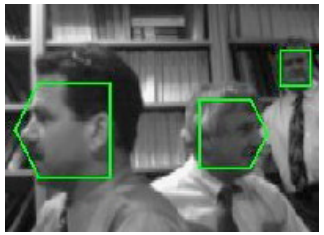
Active learning and online learning

Online active learning

Active learning

A type of supervised learning, samples and selects instances whose labels would prove to be most informative additions to the training set. (… lấy mẫu và chọn phần tử có nhãn với nhiều thông tin cho tập huấn luyện)

- Labeling the training data is not only time-consuming sometimes but also very expensive.
- Learning algorithms can actively query the user/teacher for labels.



Online learning

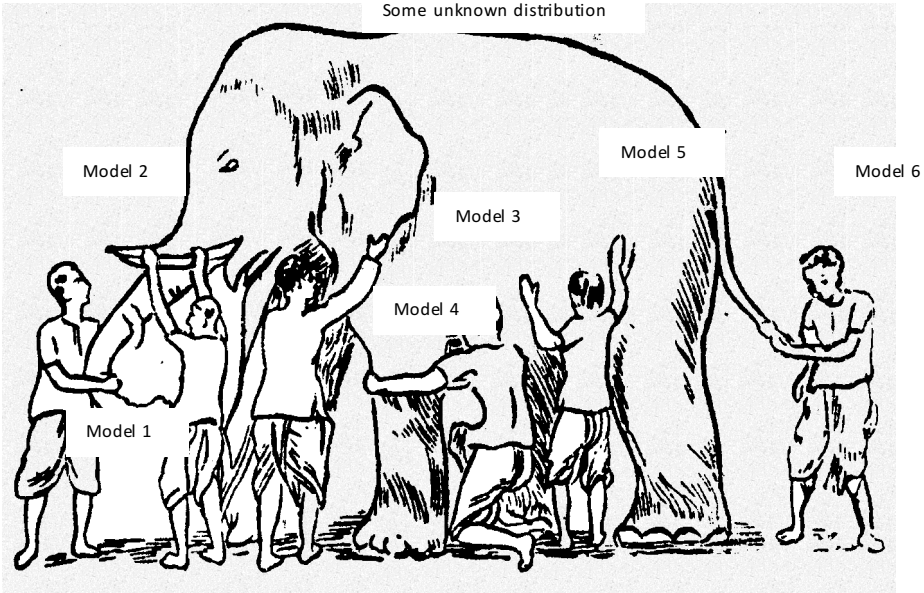
Learns one instance at a time with the goal of predicting labels for instances. (ở mỗi thời điểm chỉ học một phần tử nhằm đoán nhãn các phần tử).

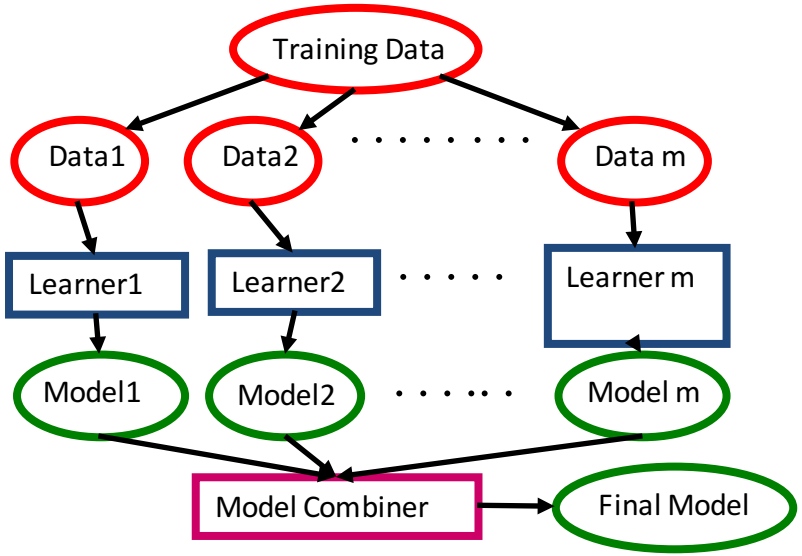
- Instances could describe the current conditions of the stock market, and an online algorithm predicts tomorrow's value of a particular stock.
- Key characteristic is after prediction, the true value of the stock is known and can be used to refine the method.

Ensemble learning

Ensemble methods employ multiple learners and combine their predictions to achieve higher performance than that of a single learner. (… dùng nhiều bộ học để đạt kết quả tốt hơn việc dùng một bộ học)

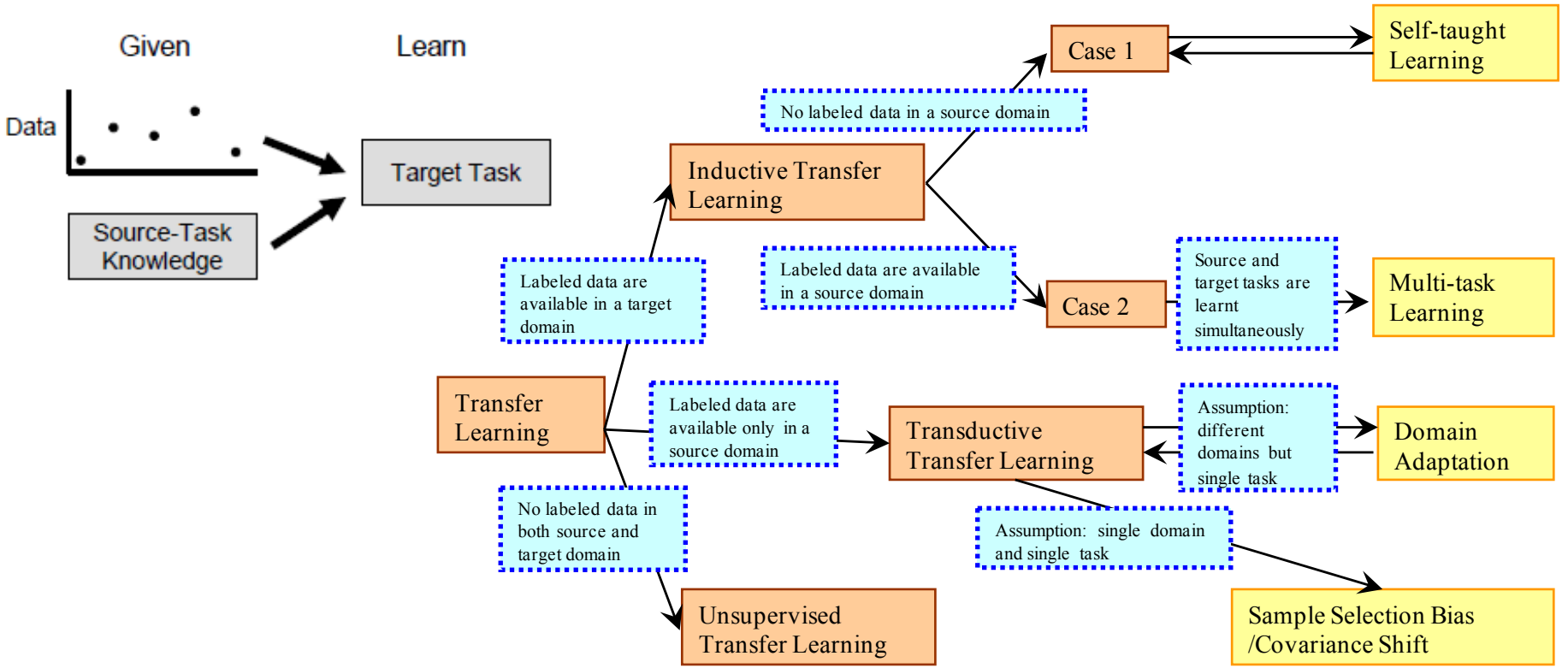
- **Boosting**: Make examples currently misclassified more important
- **Bagging**: Use different subsets of the training data for each model





Transfer learning

Aims to develop methods to transfer knowledge learned in one or more source tasks and use it to improve learning in a related target task. (truyền tri thức đã học được từ nhiều nhiệm vụ khác để học tốt hơn việc đang cần học)

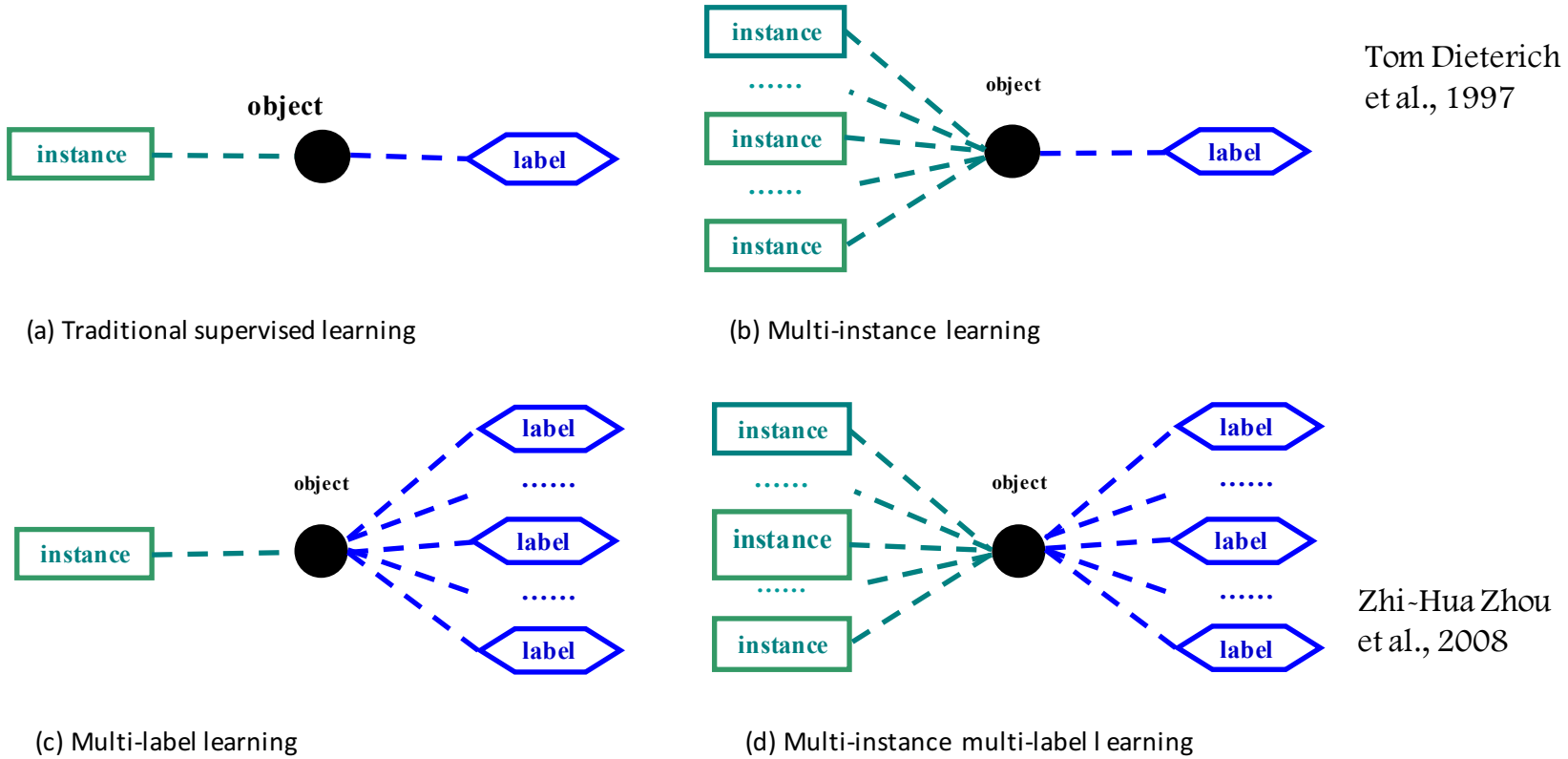


Induction: $Given\ \{x_i\},\ \ infer\ f(x)$

Transduction: $Given\ \{x_k\},\ \ infer\ x_j\ from\ x_i$

Multi-instance multi-label learning

MIML is the framework where an example is described by multiple instances and associated with multiple class labels. (một lược đồ bài toán khi mỗi đối tượng được mô tả bằng nhiều thể hiện và thuộc về nhiều lớp).

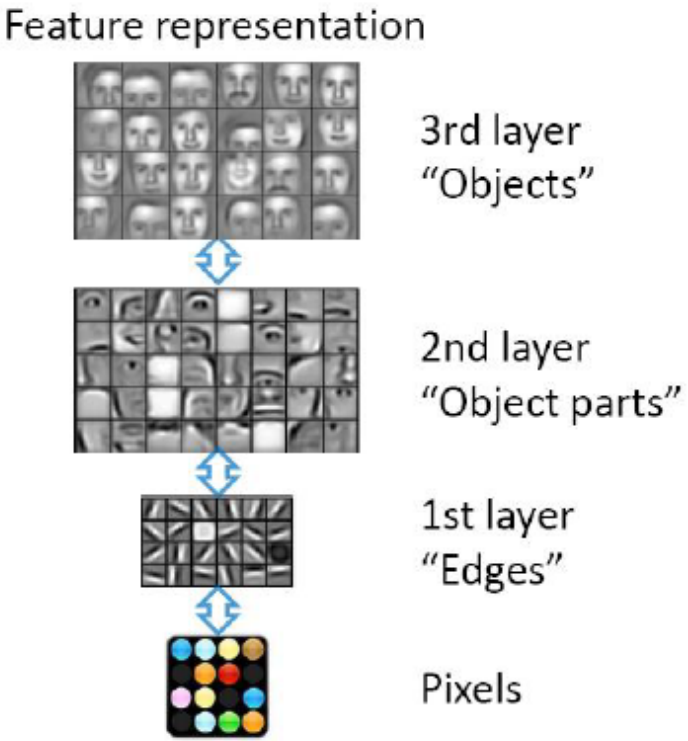


Multi-label learning vs. multi-class learning

Deep learning

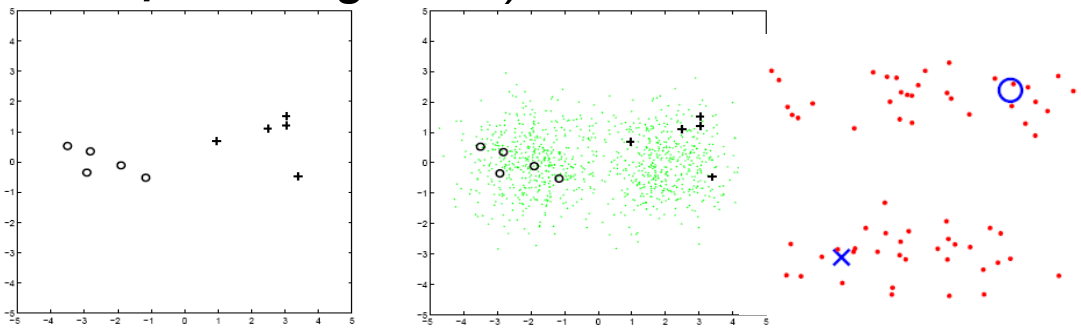
A subfield of machine learning that is based on algorithms for learning multiple levels of representation in order to model complex relationships among data. (học nhiều cấp độ biểu diễn để mô hình các quan hệ phức tạp trong dữ liệu)

- Higher-level features and concepts are thus defined in terms of lower-level ones, and such a hierarchy of features is called a deep architecture.
- Key: Deep architecture, deep representation, multi levels of latent variables, etc.



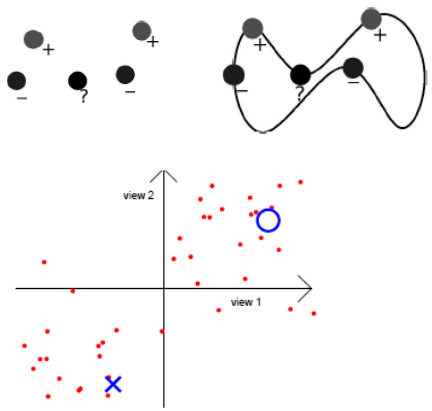
Semi-supervised learning

A class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data. (dùng cả dữ liệu có nhãn và không nhãn để huấn luyện, tiêu biểu khi ít dữ liệu có nhãn nhưng nhiều dữ liệu không nhãn)



Classes of SSL methods

- Generative models
- Low-density separation
- Graph-based methods
- Change of representation



Assumption	Approach
Cluster Assumption	Low Density Separation, eg, S3VMs
Manifold assumption	Graph-based methods (nearest neighbor graphs)
Independent views	Co-training

Vietnam Institute for Advanced Study in Mathematics

Challenges in semi-supervised learning

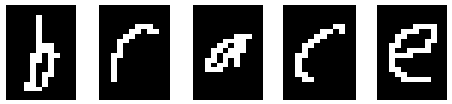
- Real SSL tasks: Which tasks can be dramatically improved by SSL?
- New SSL assumptions? E.g., assumptions on unlabeled data: label dissimilarity, order preference
- Efficiency on huge unlabeled datasets
- Safe SSL:
  - no pain, no gain
  - no model assumption, no gain
  - wrong model assumption, no gain, a lot of pain

  → develop SSL techniques that do not make assumptions beyond those implicitly or explicitly made by the classification scheme employed?

Xiaojin Zhu tutorial

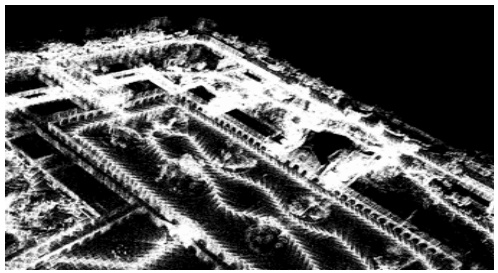
Structured prediction

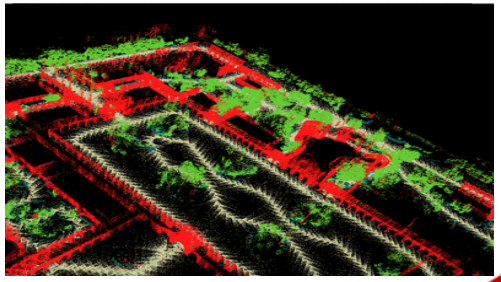
An umbrella term for machine learning and regression techniques that involve predicting structured objects. (liên quan việc đoán nhận các đối tượng có cấu trúc).

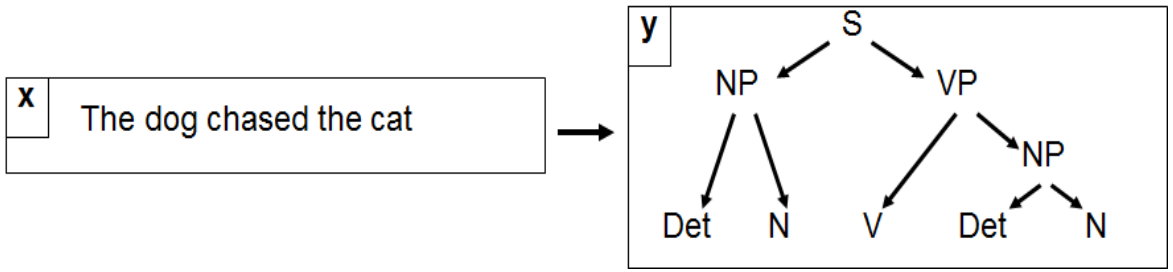


b r a c e

- Examples
  - Multi-class labeling
  - Protein structure prediction
  - Noun phrase co-reference clustering
  - Learning parameters of graphical models





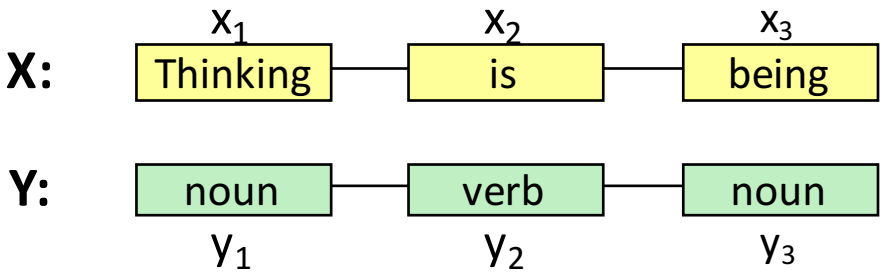


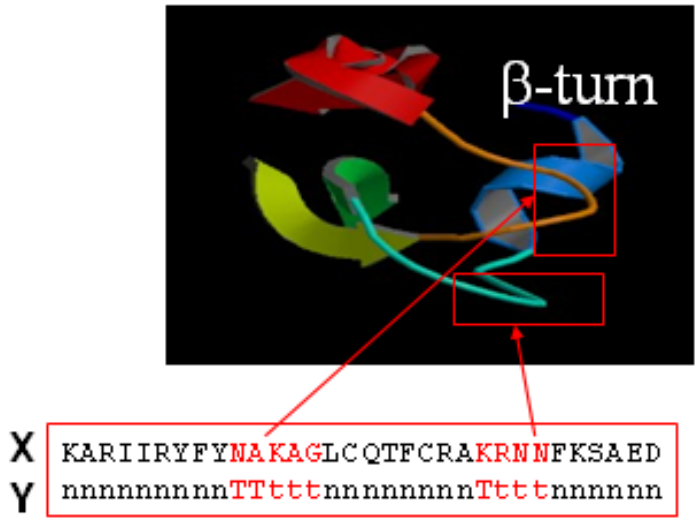
VIASM
Vietnam Institute for Advanced Study in Mathematics

Structured prediction

Example: Labeling sequence data problem

- X is a random variable over data sequences
- Y is a random variable over label sequences whose labels are assumed to range over a finite label alphabet A
- Problem: Learn how to give labels from a closed set Y to a data sequence X





- POS tagging, phrase types, etc. (NLP),
- Named entity recognition (IE)
- Modeling protein sequences (CB)
- Image segmentation, object recognition (PR)
- Recognition of words from continuous acoustic signals.

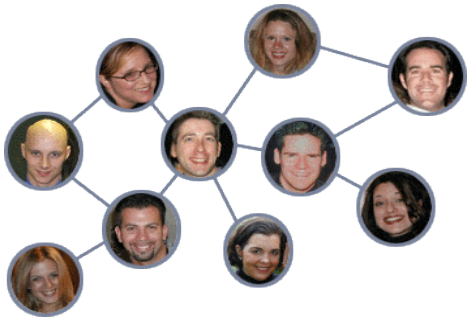
Pham, T.H., Satou, K., Ho, T.B. (2005). Support vector machines for prediction and analysis of beta and gamma turns in proteins, Journal of Bioinformatics and Computational Biology (JBCB), Vol. 3, No. 2, 343-358

Le, N.T., Ho, T.B., Ho, B.H. (2010). Sequence-dependent histone variant positioning signatures, *BMC Genomics*, Vol. 11 (S4)

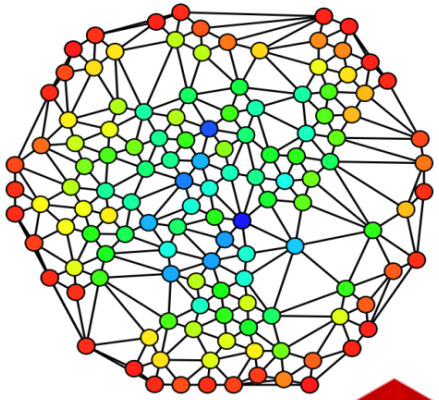
Social network analysis

Social media describes the online tools that people use to share content, profiles, opinions, insights, experiences, perspectives and media itself, thus facilitating conversations and interaction online between people. These tools include blogs, microblogs, facebook, bookmarks, networks, communities, wikis, etc.

- Social networks: Platforms providing rich interaction mechanisms, such as Facebook or MySpace, that allow people to collaborate in a manner and scale which was previously impossible (interdisciplinary study).
- Social network study: structure analysis, understanding social phenomenon, information propagation & diffusion, prediction (information, social), general dynamics, modeling (social, business, algorithmic, etc.)



Picture from Matthew Pirretti's slides



Hue (from red=0 to blue=max) indicates each node's betweenness centrality.

Figures from wikipedia

Social network analysis

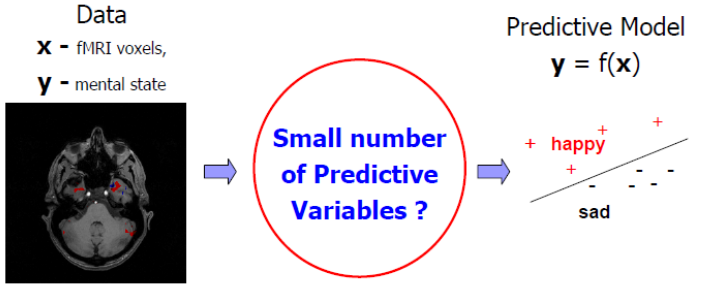
Some challenges

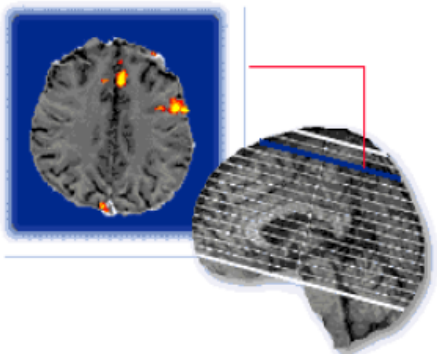
- Structural analysis: Focus on relations and patterns of relations requires methods/concepts different from traditional statistic and data analysis (e.g., graphical model, dependencies?)
- Centrality and prominence: Key issue in social network analysis is the identification of the most important or prominent actors (nodes). Many notions: degree, closeness, betweeness, rank of the actors.
- Influence: The capacity or power of persons or things to be a compelling force on or produce effects on the actions, behaviour, opinions, etc., of others (e.g., author topic models, twiter mining, etc.)
- Knowledge challenge: Enabling users to share knowledge with their community (e.g., cope with spam, privacy and security).
- Collaborative production (e.g., Wikipedia and Free Software): collaborative content creation, decentralized decision making, etc.

Sparse modeling

Selection (and, moreover, construction) of a small set of highly predictive variables in high-dimensional datasets. (chọn và tạo ra một tập nhỏ các biến có khả năng dự đoán cao từ dữ liệu nhiều chiều).

- Rapidly developing area on the intersection of statistics, machine learning and signal processing.
- Typically when data are of high-dimensional, small-sample
  - 10,000-100,000 variables (voxels)
  - 100s of samples(time points)
- Sparse SVMs, sparse Gaussian processes, sparse Bayesian methods, sparse regression, sparse Q-learning, sparse topic models, etc.

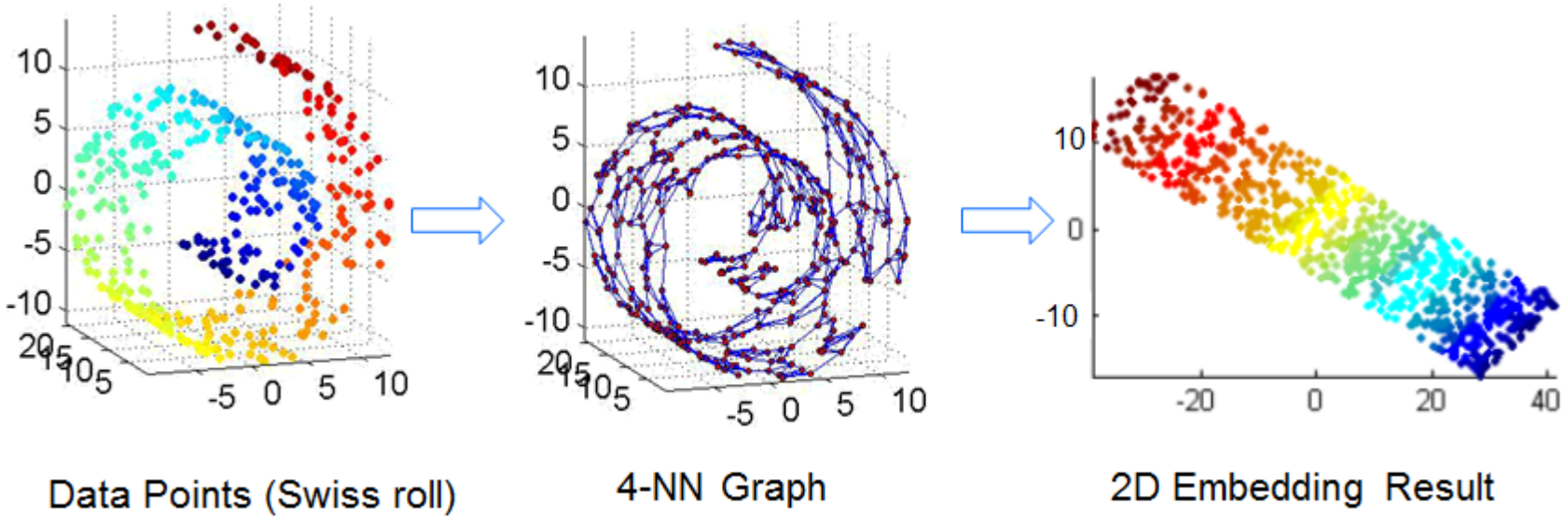




Find small number of most relevant voxels (brain areas)?

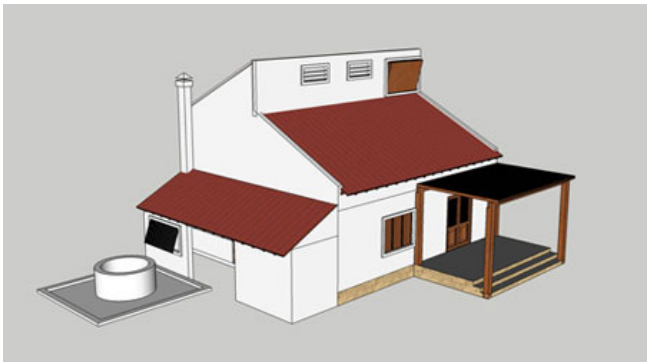
Dimensionality reduction

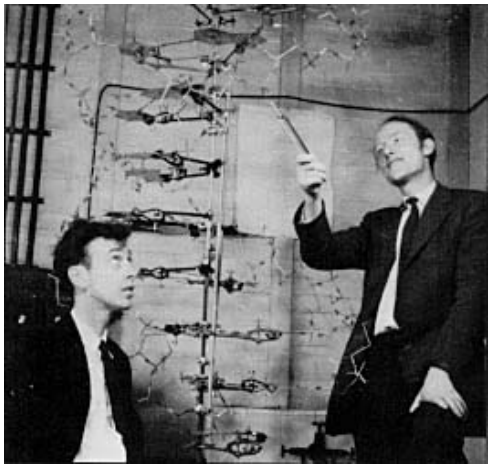
The process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction. (quá trình rút gọn số biến ngẫu nhiên đang quan tâm, gồm lựa chọn biến và tạo biến mới).



Model selection

Model: Abstract description or representation of a reality.





DNA model figured out in 1953 by Watson and Crick

A model is defined as a parametric collection of probability distributions, indexed by model parameters

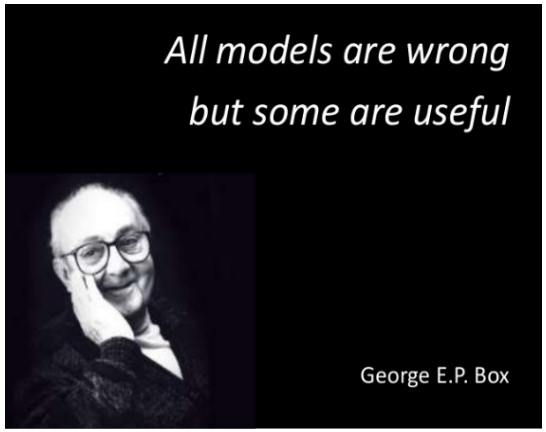
$$M = \{f(y|\theta)|\theta \in \Omega\}$$

Pignet index (body build index) = Stature in cm - (weight in kg + chest circumference in cm)

Very sturdy: <10, Sturdy: 10-15, Good: 16-20, Average: 21-25, Weak: 26-30, Very weak: 31-35, Poor: >36

Model selection

- **Problem**: Choosing *the most appropriate* model(s) given a dataset and the task.
- Relating to selecting
  - Models that can be appropriated
  - Parameters of those models

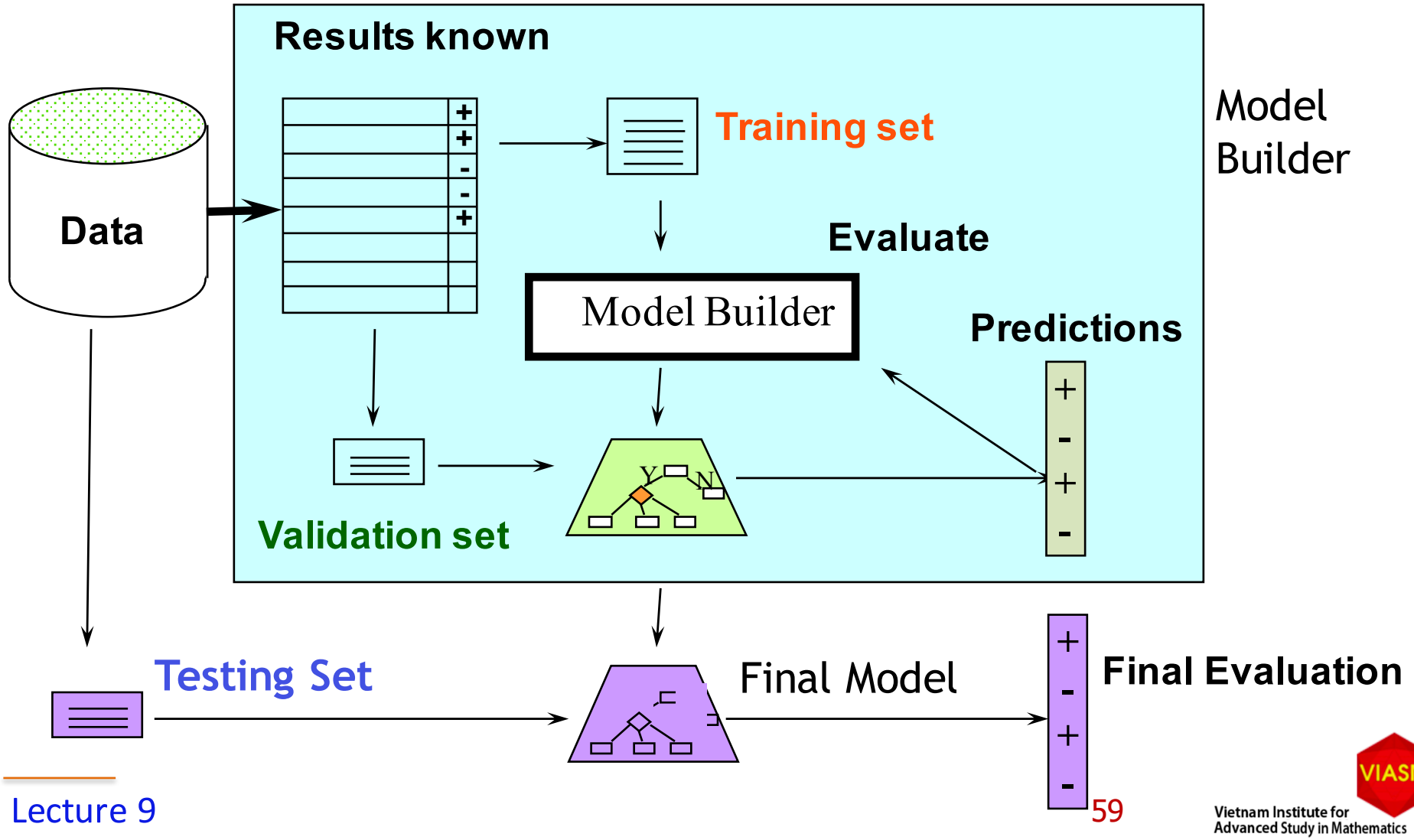


(1919-2013)

- Examples of model selection problems
  - Is it a linear or non-linear regression I should choose?
  - Which neural net architecture gives the best generalization error?
  - How many neighbors should I take in consideration in k-NN?
  - Should I use a linear model, a decision tree, a neural net, a local learning algorithms?
  - Which of the 50 features are relevant for this problem?

Evaluation of learned models

Classification: Train, Validation, Test

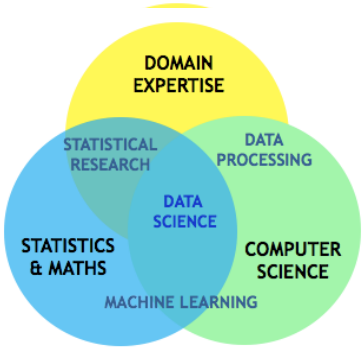


Which algorithms do best at which tasks?

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

Three aspects for data scientists



Hadoop + MapReduce → Spark → TensorFlow (deep learning)

inefficiency for iterative algorithms

Outline

1. What is data science?
2. Principles of data science
3. DSLab's data science lectures

Các modules đào tạo của DSLab

Background of data science	Basic methods of data science	Advanced methods of data science	Analysis of domain data
Data science principles	Data preprocessing	Outlier analysis	Analysis of text data
Data and databases	Result evaluation	Kernel methods	Analysis of social networks
Big data overview	Regression	Graphical models	Analysis of categorical data
Brief of statistics	Classification	Dimension reduction	Analysis of time-series data
Brief of linear algebra and optimization	Cluster analysis	Deep learning	Analysis of graph data
R and Python	Association analysis	Recommender systems	Analysis of image data
		Reinforcement learning	Analysis of streaming data
		Advanced computation technologies	Analysis of sequential data

Chưa một đơn vị riêng lẻ nào ở Việt Nam lúc này có thể đào tạo toàn bộ các nội dung trên với chất lượng đáp ứng nhu cầu.

DỰ THẢO CHƯƠNG TRÌNH VÀ LỊCH ĐÀO TẠO VỀ KHOA HỌC DỮ LIỆU

Ngày	Buổi	Nội dung học	Giảng viên
Thứ Hai 11/6	Sáng 8:00-12:00	Essence of Statistics	TS Trần Mạnh Cường
	Chiều 13:30-17:30	Classification/Kernel methods Real-life problems 1	PGS Nguyễn Đức Dũng
Thứ Ba 12/6	Sáng 8:00-12:00	Introduction to Data Science	GS Hồ Tú Bảo
	Chiều 13:30-17:30	Analysis of time series data	TS Trịnh Quốc Anh
Thứ Tư 13/6	Sáng 8:00-12:00	Classification/Key methods	TS Nguyễn Thị Minh Huyền
	Chiều 13:30-17:30	Cluster Analysis	
Thứ Năm 14/6	Sáng 8:00-12:00	Association analysis Real-life problems 3	PGS Phan Xuân Hiếu
	Chiều 13:30-17:30	Regression Real-life problems 2	TS Nguyễn Thanh Tùng
Thứ Sáu 15/6	Sáng 8:00-12:00	Data preprocessing Evaluation of analysis results	TS Thân Quang Khoát
	Chiều 13:30-17:30	Reinforcement Learning	TS Nguyễn Đỗ Văn
Thứ Bảy 16/6	Sáng 8:00-12:00	Outlier analysis Anomaly detection	TS Ngô Xuân Bách
	Chiều 13:30-17:30	Analysis of social networks Real-life problems 4	TS Lê Hồng Phương

- All lectures are designed and given in a common framework and connected to each other.
- Slides for each lecture will be provided one day prior to the class.

Recommended references

- http://www.kdnuggets.com
- Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000.
- Mitchell T., Machine Learning, McGraw Hill, 1997.
- Charu Aggarwal, Data Mining, Springer 2015.
- Journals: Data Mining and Knowledge Discovery, IEEE Knowledge and Data Engineering, Knowledge and Information Systems, etc.
- Selected reading papers

Summary

- Much more data around us than ever before.
- Organizations and persons who know to exploit the data can have competitive advantages.
- Lectures on data science allow you to understand and to practice with data analytics.
- Need a big effort!

Homework

1. How do you understand the cyber-physical systems?
2. Give an example of data, information and knowledge
3. Visit the website kdnuggets. Which parts are of your interests? Give examples.
4. Note what you remember from the statistics if you learned it previously.