BÀI GIẢNG NHẬP MÔN KHAI PHÁ DỮ LIỆU

CHƯƠNG 1. GIỚI THIỆU CHUNG VỀ KHAI PHÁ DỮ LIỆU

TS. Trần Mai Vũ

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

ĐẠI HỌC QUỐC GIA HÀ NỘI

Nội dung

1. Tại sao khai phá dữ liệu (KPDL)?
2. Khái niệm KPDL và phát hiện tri thức trong CSDL
3. KPDL và xử lý CSDL truyền thống
4. Kiểu dữ liệu trong KPDL
5. Kiểu mẫu được khai phá
6. Công nghệ KPDL điển hình
7. Một số ứng dụng điển hình
8. Các vấn đề chính trong KPDL

1. Tại sao khai phá dữ liệu

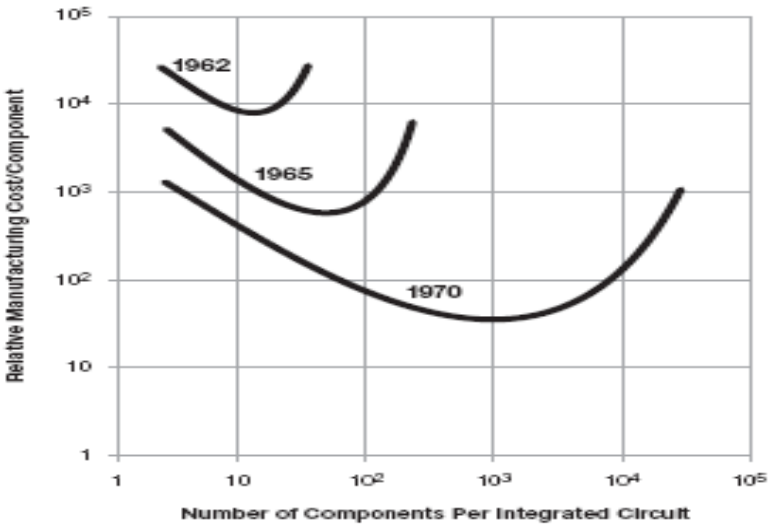
- Một ví dụ: Donal Trump Thắng cử Tổng thống Mỹ
- Bùng nổ dữ liệu và dữ liệu lớn (Big Data)
  - Lý do công nghệ
  - Lý do xã hội
  - Thể hiện
- Ngành kinh tế tri thức, dữ liệu và thông tin (Infonomics)
  - Kinh tế tri thức
  - Phát hiện tri thức từ dữ liệu
- Nhu cầu khai phá dữ liệu tại Việt Nam
  - Trường hè KHDL 2016

Ví dụ: Tại sao khai phá dữ liệu ?

- **Phân tích dữ liệu giúp ứng viên Tổng thống Mỹ**

1. Đào Trung Thành. *Big Data đã giúp Trump chiến thắng trong cuộc Bầu cử Mỹ*. http://vietnamnet.vn/vn/cong-nghe/ung-dung/big-data-giup-donald-trump-chien-thang-trong-cuoc-bau-cu-my-big-data-nguy-hiem-den-muc-nao-346181.html.(13/12/2016, 15:02 GMT+7). *Big Data nguy hiểm tới mức nào?* http://vietnamnet.vn/vn/cong-nghe/ung-dung/big-data-da-giup-trump-chien-thang-trong-cuoc-bau-cu-my-the-nao-346184.html (10/02/2017 21:55:30 (GMT+7)) **TÍNH MỚI LẠ TỪ DỮ LIỆU LỚN**
2. TRẦN THẮNG (kỹ sư hàng không ở Mỹ). *Mạng xã hội giúp ông Trump đắc cử tổng thống Mỹ như thế nào?* http://tuoitre.vn/tin/the-gioi/bau-cu-tong-thong-my-2016/20161110/it-phieu-hon-vi-sao-ong-trum-dac-cu-tong-thong-my/1216150.html (10/11/2016 19:15 GMT+7) **TÍNH KHÁC BIỆT: TWITER ≠ TRUYỀN THÔNG TT**
3. Von Hannes Grassegger und Mikael Krogerus. *Ich habe nur gezeigt, dass es die Bombe gibt*. Das Magazin N°48 – 3. Dezember 2016. https://www.dasmagazin.ch/2016/12/03/ich-habe-nur-gezeigt-dass-es-die-bombe-gibt/. *Nhà tâm lý học Michal Kosinski phát triển một phương pháp phân tích tinh tế mọi người dựa trên hành vi của họ trên Facebook. Và như thế giúp Donald Trump chiến thắng*. **PHƯƠNG PHÁP, KỸ THUẬT MIỀN ỨNG DỤNG: PHÂN TÍCH DỮ LIỆU TÂM LÝ**
4. http://www.michalkosinski.com/: an Assistant Professor in Organizational Behavior at Stanford Graduate School of Business
5. Leonid Bershidsky. *No, Big Data Didn't Win the U.S. Election*. https://www.bloomberg.com/view/articles/2016-12-08/no-big-data-didn-t-win-the-u-s-election (DEC 8, 2016 2:56 PM EST). "Obviously, it is not big data analytics that wins the election," he (*Michal Kosinski*) wrote back. "Candidates do. We don't know how much his victory was helped by big data analytics.“ **KINH DOANH MÀ KHÔNG LÀ CÔNG NGHỆ**

Công nghệ: Bùng nổ dữ liệu: Luật Moore



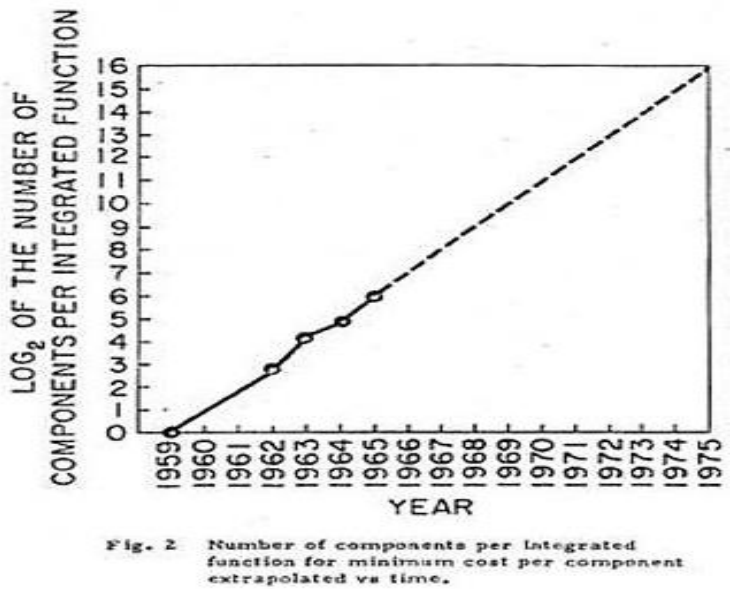
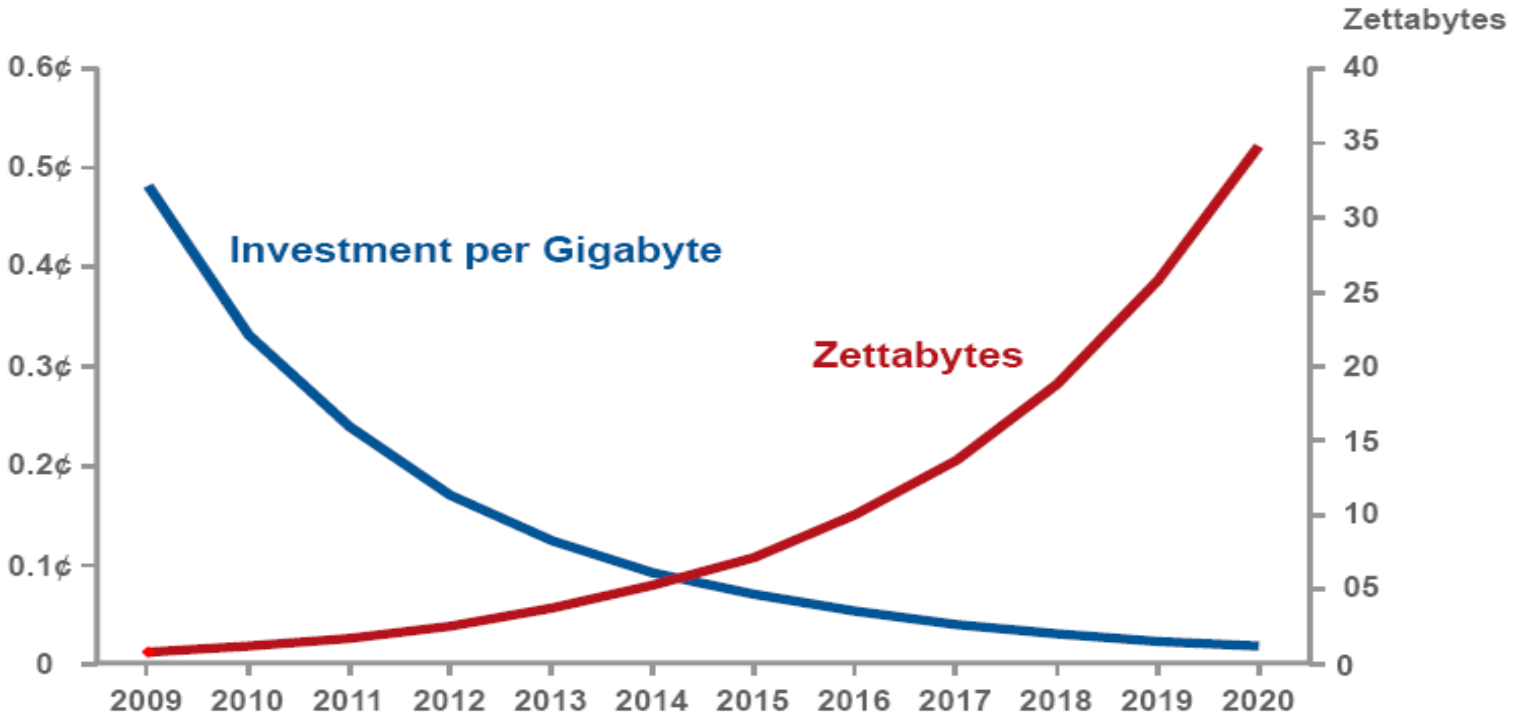


Fig. 2 Number of components per integrated function for minimum cost per component extrapolated vs time.

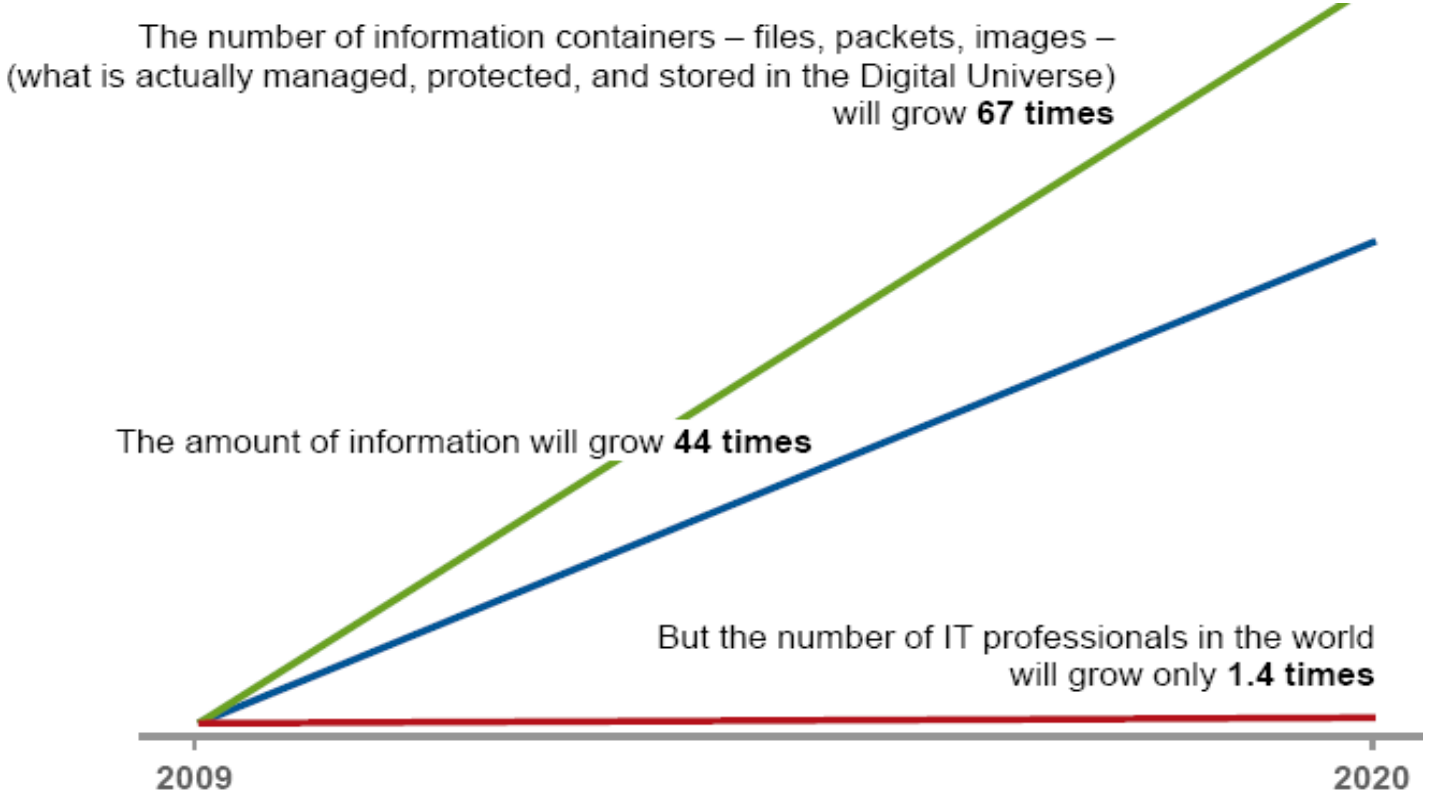
- Xuất xứ
  - Gordon E. Moore (1965). Cramming more components onto integrated circuits, *Electronics*, **38** (8), April 19, 1965. *Một quan sát và dự báo*
- "Phương ngôn 2x
  - Số lượng bán dẫn tích hợp trong một chíp sẽ tăng gấp đôi sau khoảng hai năm
  - Chi phí sản xuất mạch bán dẫn với cùng tính năng giảm một nửa sau hai năm
  - Phiên bản 18 tháng: rút ngắn chu kỳ thời gian

Bùng nổ dữ liệu: Giá thành và thể hiện



- *Nguồn:* IDC Digital Universe Study, sponsored by EMC, May 2010
- Giá tạo dữ liệu ngày càng rẻ hơn
  - Chiều hướng giá tạo mới dữ liệu giảm dần
  - 0,5 xu Mỹ/1 GB vào năm 2009 giảm tới 0,02 xu Mỹ /1 GB vào năm 2020
- Dung lượng tổng thể tăng
  - Độ dốc tăng càng cao
  - Đạt 35 ZB vào năm 2020

Nhu cầu nắm bắt dữ liệu



- Bùng nổ dữ liệu với tăng trưởng nhận lực CNTT
  - Dung lượng thông tin tăng 67 lần, đối tượng dữ liệu tăng 67 lần
  - Lực lượng nhân lực CNTT tăng 1,4 lần
  - *Nguồn*: IDC Digital Universe Study, sponsored by EMC, May 2010.

Nhu cầu thu nhận tri thức từ dữ liệu

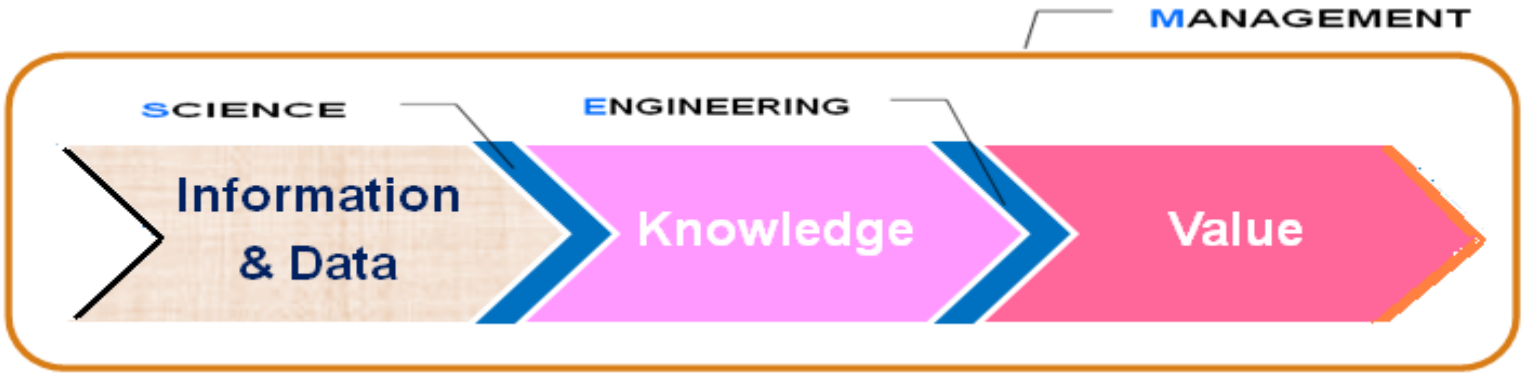
- Jim Gray, chuyên gia của Microsoft, giải thưởng Turing 1998
  - *"Chúng ta đang ngập trong dữ liệu khoa học, dữ liệu y tế, dữ liệu nhân khẩu học, dữ liệu tài chính, và các dữ liệu tiếp thị. Con người không có đủ thời gian để xem xét dữ liệu như vậy. Sự chú ý của con người đã trở thành nguồn tài nguyên quý giá.*

    Vì vậy, chúng ta phải tìm cách tự động phân tích dữ liệu, tự động phân loại nó, tự động tóm tắt nó, tự động phát hiện và mô tả các xu hướng trong nó, và tự động chỉ dẫn các dị thường.

    Đây là một trong những lĩnh vực năng động và thú vị nhất của cộng đồng nghiên cứu cơ sở dữ liệu. Các nhà nghiên cứu trong lĩnh vực bao gồm thống kê, trực quan hóa, trí tuệ nhân tạo, và học máy đang đóng góp cho lĩnh vực này. Bề rộng của lĩnh vực làm cho nó trở nên khó khăn để nắm bắt những tiến bộ phi thường trong vài thập kỷ gần đây" [HK0106].
- Kenneth Cukier,
  - *"Thông tin từ khan hiếm tới dư dật. Điều đó mang lại lợi ích mới to lớn… tạo nên khả năng làm được nhiều việc mà trước đây không thể thực hiện được: nhận ra các xu hướng kinh doanh, ngăn ngừa bệnh tật, chống tội phạm …*

    Được quản lý tốt, dữ liệu như vậy có thể được sử dụng để mở khóa các nguồn mới có giá trị kinh tế, cung cấp những hiểu biết mới vào khoa học và tạo ra lợi ích từ quản lý". http://www.economist.com/node/15557443?story_id=15557443

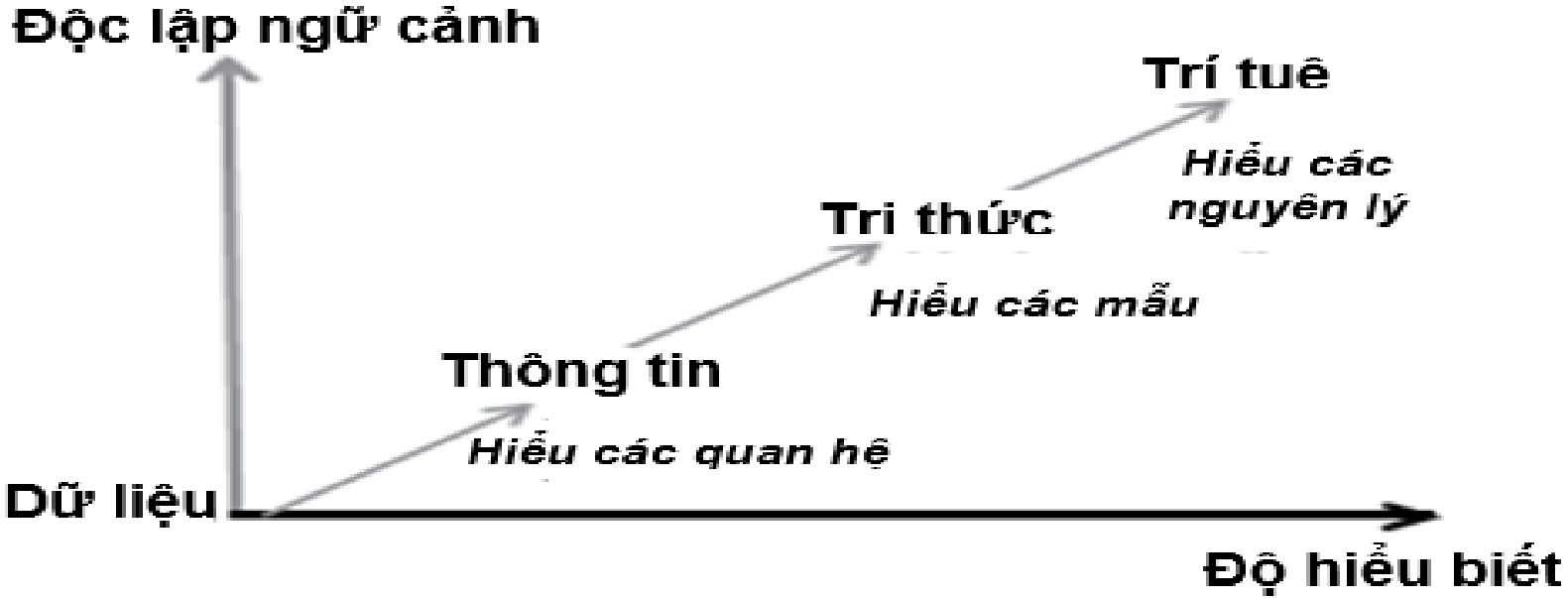
Kinh tế dịch vụ: Từ dữ liệu tới giá trị



- Kinh tế dịch vụ
  - Xã hội loài người chuyển dịch từ kinh tế hàng hóa sang kinh tế dịch vụ. Lao động dịch vụ vượt lao động nông nghiệp (2006).
  - Mọi nền kinh tế là kinh tế dịch vụ.
  - Đơn vị trao đổi trong kinh tế và xã hội là dịch vụ
- Dịch vụ: dữ liệu & thông tin $\Rightarrow$ tri thức $\Rightarrow$ giá trị mới
  - Khoa học: dữ liệu & thông tin $\Rightarrow$ tri thức
  - Kỹ nghệ: tri thức $\Rightarrow$ dịch vụ
  - Quản lý: tác động tới toàn bộ quy trình thi hành dịch vụ

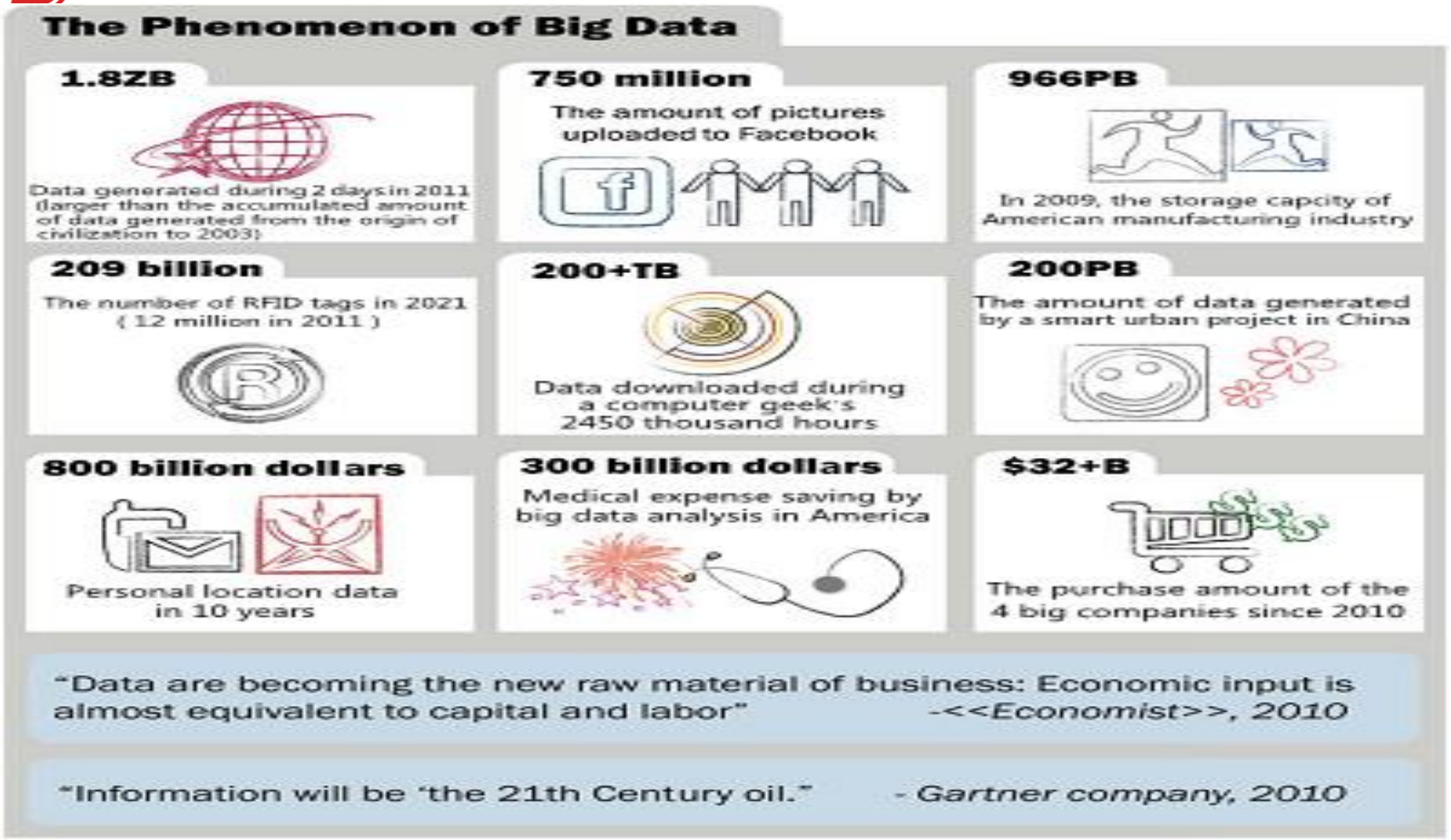
Jim Spohrer (2006). A Next Frontier in Education, Employment, Innovation, and Economic Growth, *IBM Corporation*, 2006

Quá trình tiến hóa dữ liệu tới trí tuệ



- Từ dữ liệu tới trí tuệ
  - Dữ liệu (data): dữ kiện không ngữ cảnh. **Trình bày dữ kiện**.
  - Thông tin (information): dữ kiện với ngữ cảnh và khía cạnh nào đó. Dữ liệu có ý nghĩa, dữ liệu trong ngữ cảnh. **Hiểu được quan hệ**
  - Tri thức (knowledge): Thông tin được dung để phát hiện và hiểu được mẫu trong dữ liệu. **Hiểu được mẫu**
  - Trí tuệ (wisdom): Tri thức nảy sinh khi hiểu được lý do mẫu xuất hiện trong dữ liệu. **Hiểu được nguyên lý**
  - http://www.systems-thinking.org/kmgmt/kmgmt.htm

Big data không ngừng gia tăng và giá trị



- (i) Mỹ: tiết kiệm 300 tỷ US$ ngành y tế, (ii) Châu Âu: chỉnh phủ tiết kiệm 100 tỷ Euro (giảm gian lận, sai sót, chênh lệch thuế), v.v.

[Chen14] Min Chen, Shiwen Mao, Yunhao Liu. *Big Data: A Survey*. MONET 19(2): 171-209, 2014.

Giá trị dữ liệu: Ví dụ 1 (Capital One)

- **-1980's: Mô hình xác suất mặc định với thẻ tín dụng đồng mức**
  - Người q/lý NH tin khách hàng chưa ưa mức khác nhau;
  - HTTT chưa thể đáp ứng quản lý mức tín dụng khác nhau
- **Richard Fairbanks và Nigel Morris**
  - CNTT cho mô hình dự báo tinh vi hơn (mô hình lợi nhuận), đủ năng lực các mức tín dụng
  - Thuyết phục nhà QL NH lớn: thất bại.
  - Thuyết phục được người QL ngân hàng nhỏ Signet Bank: tin một tỷ lệ nhỏ khách hàng thực sự tạo ra hơn 100% lợi nhuận của NH từ hoạt động thẻ tín dụng
  - MHLN: tốt hơn → KH tốt nhất + thu hút KH tốt nhất từ NH lớn

http://www.fundinguniverse.com/company-histories/capital-one-financial-corporation-history/ và https://www.capitalone.com/

Giá trị dữ liệu: Ví dụ 1 (Capital One)

- **Thiếu dữ liệu và giải pháp**
  - Không có dữ liệu với mức thẻ tín dụng khác nhau.
  - Tạo DL cho mô hình (MHLN): cung cấp ngẫu nhiên mức tín dụng khác nhau tới KH khác nhau. DL là tài nguyên thì phải đầu tư
  - Tốn kém: tỷ lệ "khoanh nợ": 2,9% đầu ngành, do cung cấp ngẫu nhiên giảm sút tới gần 6% dư chưa thanh toán
- **Kết quả**
  - 4 năm: vừa thu thập dữ liệu vừa hoàn thiện mô hình. 1994. (***Học máy tăng cường***). 1994 tách thành Capital One
  - Nhanh chóng thành có lợi nhuận lớn nhất. Nhà phát hành thể tín dụng thứ sáu nước Mỹ: mở 48,6 triệu tài khoản 53,2 tỷ US$, 12% gia đình Mỹ.
  - Bền vững sau khủng khoảng 9/11
  - Chiến lược dựa trên thông tin Information-Based Strategy (IBS) là lợi thế lớn

http://www.fundinguniverse.com/company-histories/capital-one-financial-corporation-history/ và https://www.capitalone.com/

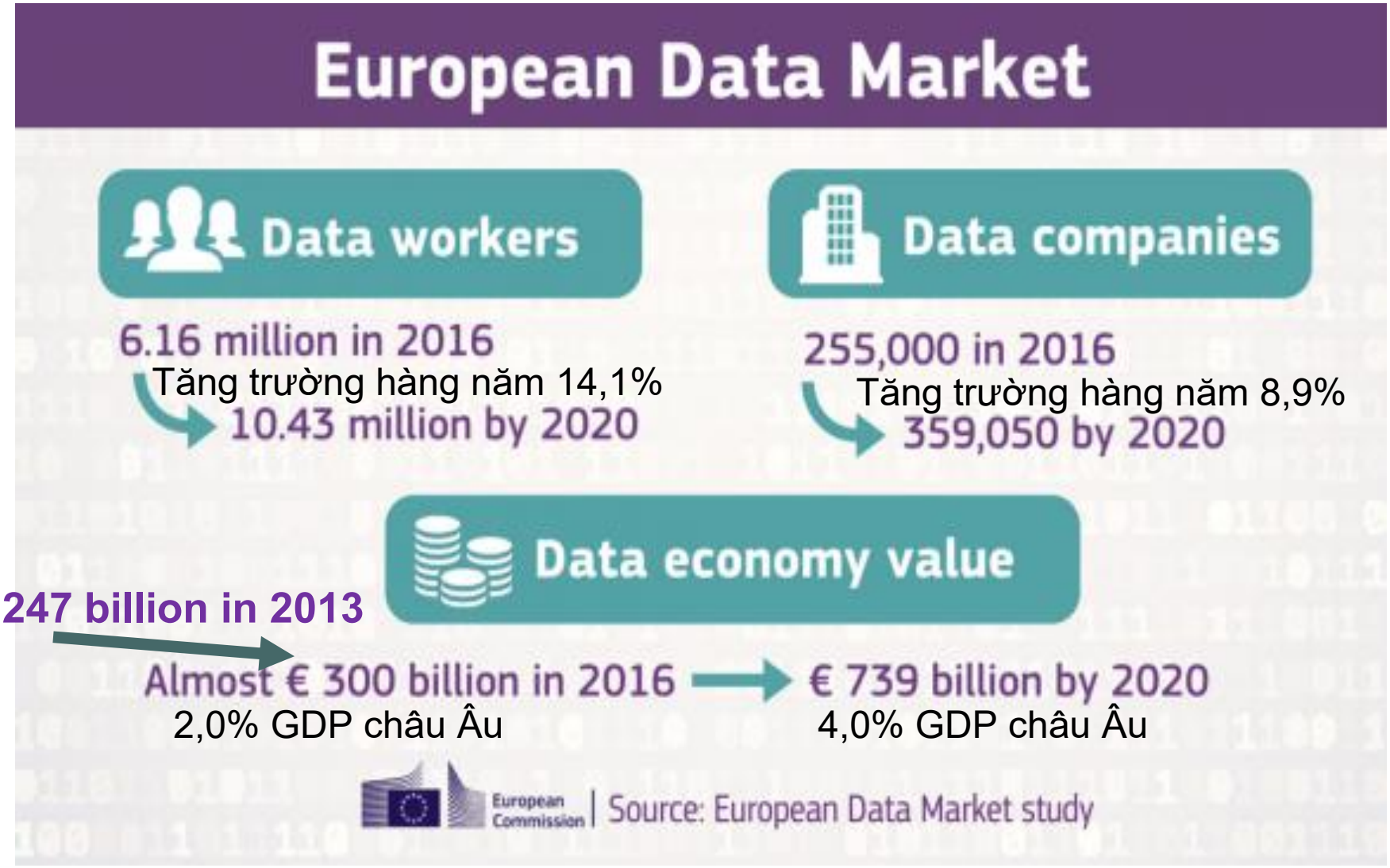
Giá trị dữ liệu: Ví dụ 2 (Microsoft-LinkedIn)

- Sự kiện và vấn đề
  - Microsoft mua lại LinkedIn với giá 26,2 tỷ đô-la Mỹ
  - Định giá kế toán của LinkedIn là 3,2 tỷ đô-la Mỹ
  - Độ chênh lệch 23 tỷ đô-la Mỹ là một con số rất lớn ?
- Giá trị dữ liệu LinkedIn mang lại cho Microsoft
  - 23 tỷ đô-la Mỹ chủ yếu từ giá trị dữ liệu
  - Đo lường giá trị dữ liệu ?
  - Infonomics (Chương 2)
- Với Google và Facebook
  - Ban đầu: Dữ liệu phục vụ quảng cáo tốt hơn
  - Hiện tại: Dữ liệu là một dịch vụ trí tuệ nhân tạo. Công nghiệp 4.0
- Liên hệ với Uber, Grab Việt Nam
  - Họ thu thập được các dữ liệu gì ?
  - Dữ liệu đó có thể sử dụng (kinh doanh) như thế nào ?

https://www.forbes.com/sites/bernardmarr/2017/05/31/why-every-business-needs-infonomics-in-a-big-data-world-and-what-it-is/#1e290da64c69

https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy

Giá trị dữ liệu: Thị trường dữ liệu châu Âu

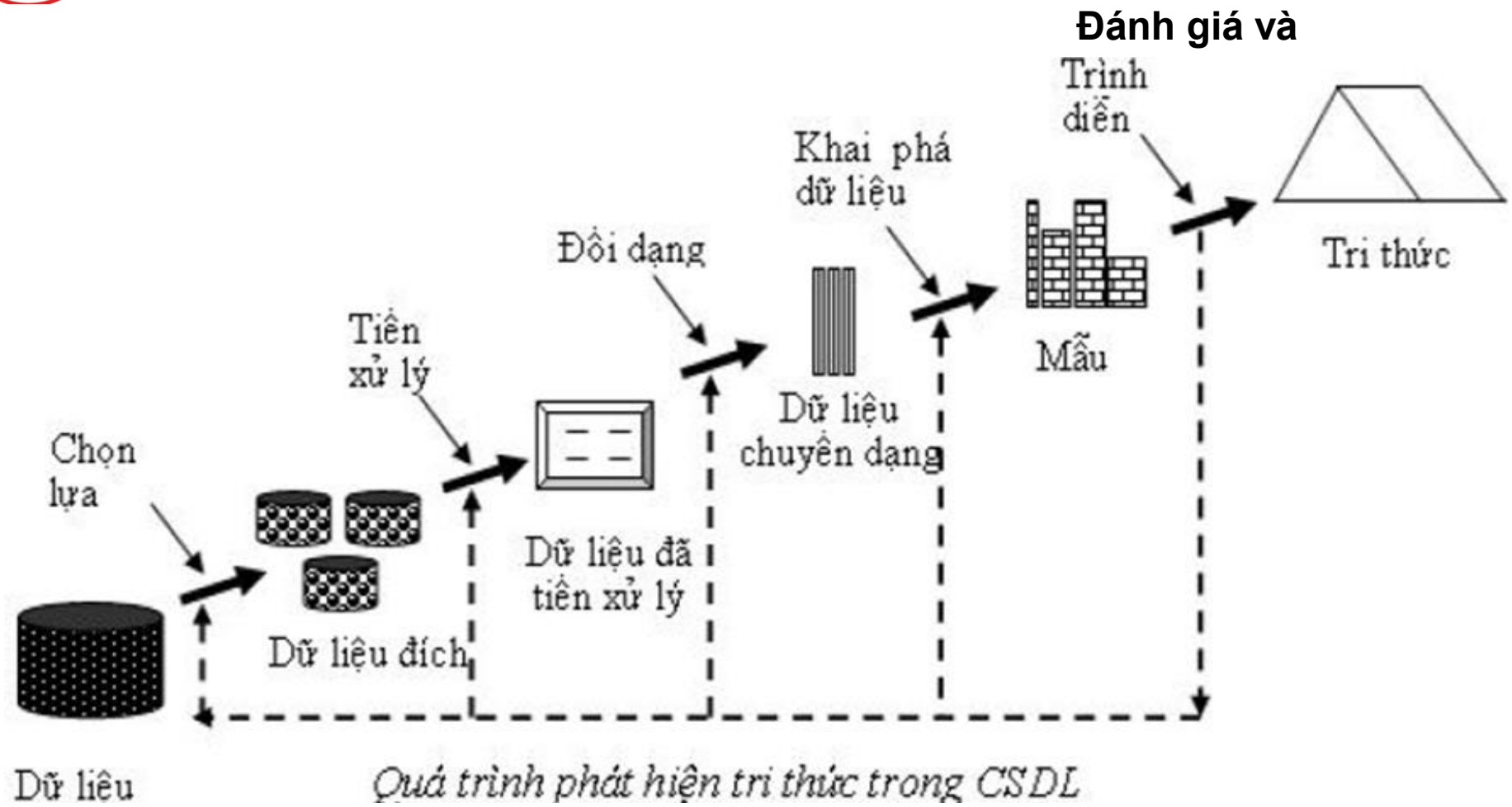


https://ec.europa.eu/digital-single-market/en/news/final-results-european-data-market-study-measuring-size-and-trends-eu-data-economy

2. Khái niệm KDD và KPDL

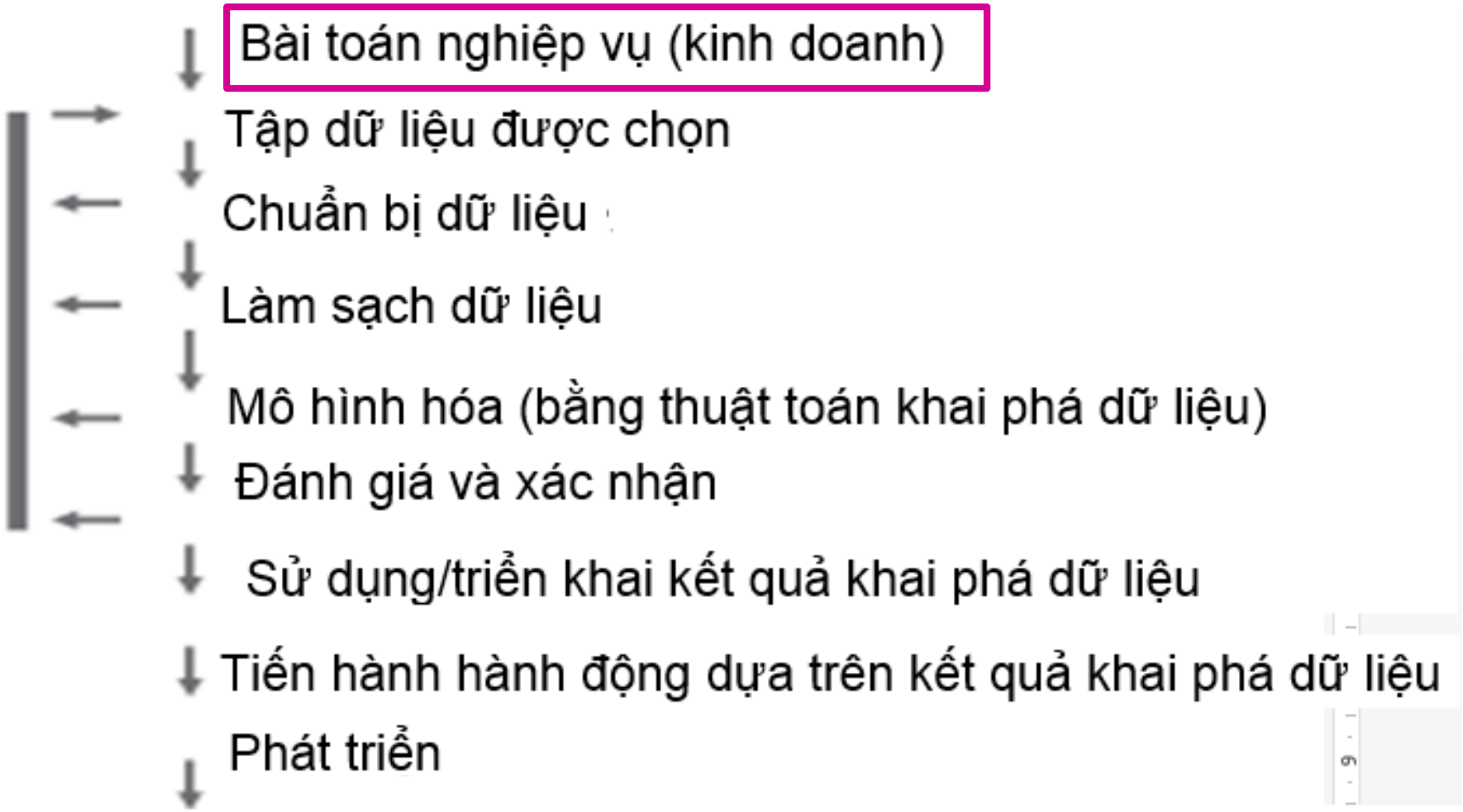
- Knowledge discovery from databases
  - Trích chọn các mẫu hoặc tri thức hấp dẫn (không tầm thường, ẩn, chưa biết và hữu dụng tiềm năng) từ một tập hợp lớn dữ liệu
  - KDD và KPDL: tên gọi lẫn lộn? *theo hai tác giả*|*Khai phá dữ liệu*
  - **Data Mining là một bước trong quá trình KDD**

Quá trình KDD [FPS96]



[FPS96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth (1996). From Data Mining to Knowledge Discovery: An Overview, *Advances in Knowledge Discovery and Data Mining 1996*: 1-34

Quá trình Khai phá dữ liệu



Các bước trong quá trình KDD

- Học từ miền ứng dụng
  - Tri thức sẵn có liên quan và mục tiêu của ứng dụng
- Khởi tạo một tập dữ liệu đích: chọn lựa dữ liệu
- Chuẩn bị dữ liệu và tiền xử lý: (huy động tới 60% công sức!)
- Thu gọn và chuyển đổi dữ liệu
  - Tìm các đặc trưng hữu dụng, rút gọn chiều/biến, tìm các đại diện bất biến.
- Chọn lựa chức năng (hàm) KPDL
  - Tóm tắt, phân lớp, hồi quy, kết hợp, phân cụm.
- Chọn (các) thuật toán KPDL
- Bước KPDL: tìm mẫu hấp dẫn
- Đánh giá mẫu và trình diễn tri thức
  - Trực quan hóa, chuyển dạng, loại bỏ các mẫu dư thừa, v.v.
- Sử dụng tri thức phát hiện được

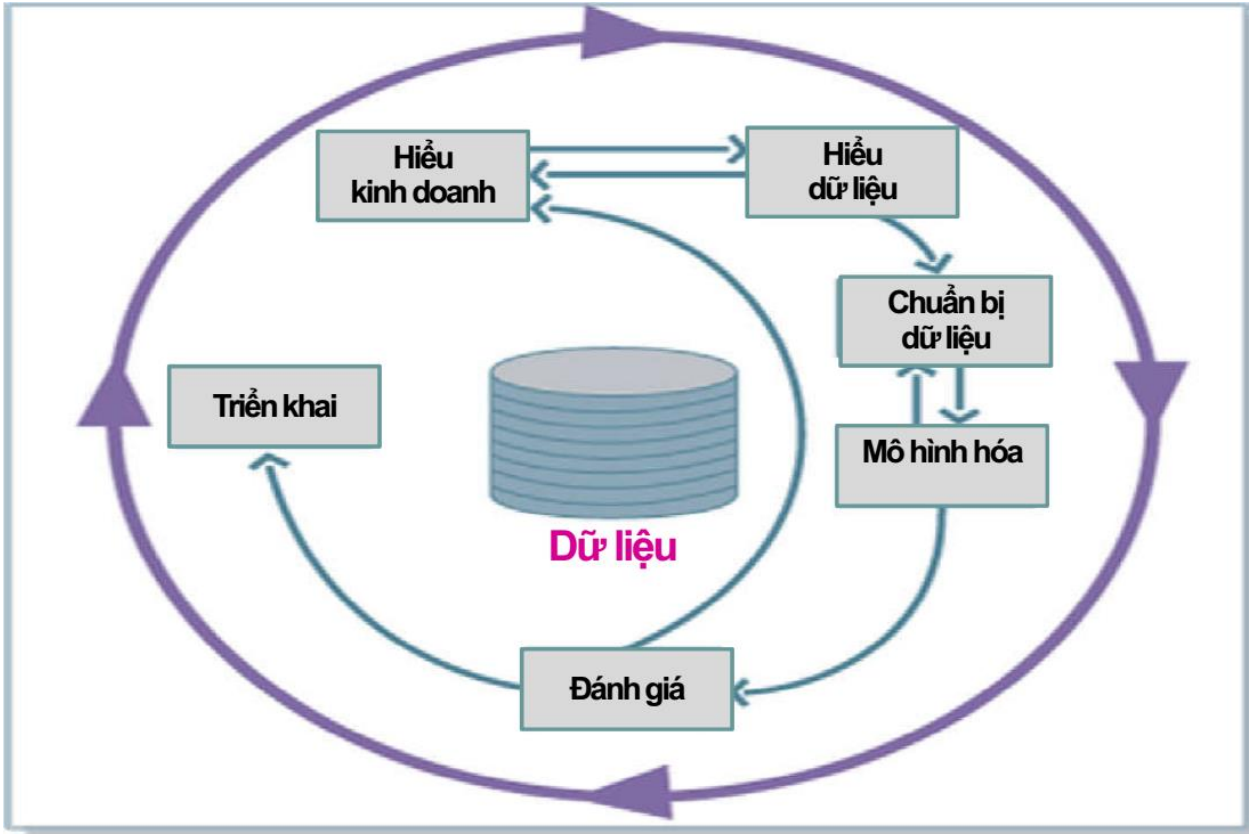
Các khái niệm liên quan

- Các tên thay thế
  - chiết lọc tri thức (knowledge extraction),
  - phát hiện thông tin (information discovery),
  - thu hoạch thông tin (information harvesting),
  - khai quật/nạo vét dữ liệu (data archaeology/ dredging),
  - Phân tích/xử lý mẫu/dữ liệu (data/pattern analysis/processing)
  - Thông minh doanh nghiệp (business intelligence -BI)
  - …
- Phân biệt: Phải chăng mọi thứ là DM?
  - Xử lý truy vấn suy diễn.
  - Hệ chuyên gia hoặc chương trình học máy/thống kê nhỏ

Mô hình quá trình KDD lặp [CCG98]

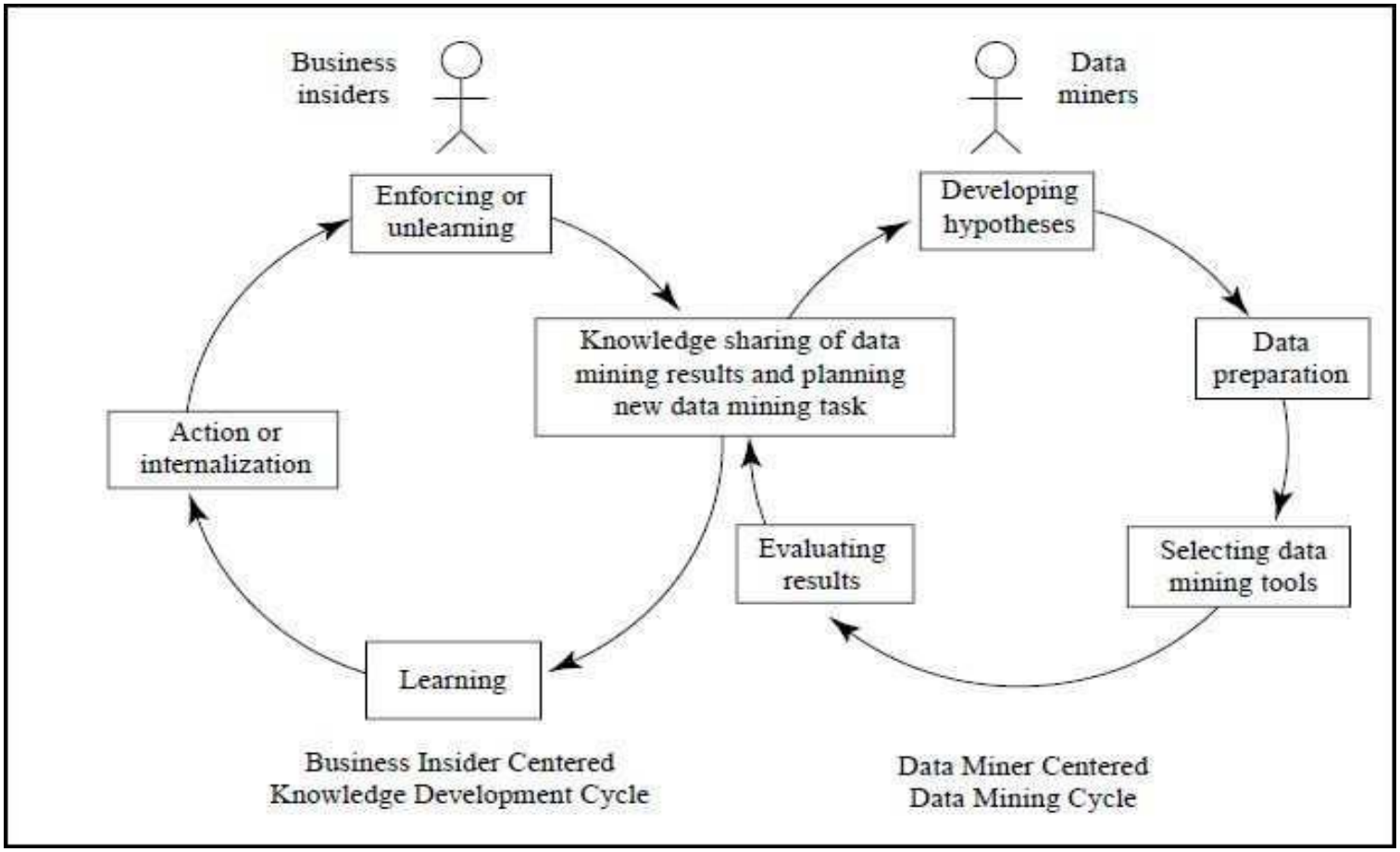
- Một mô hình cải tiến quá trình KDD
  - Định hướng kinh doanh: Xác định 1-3 câu hỏi hoặc mục tiêu KDD
  - Kết quả thi hành được: xác định tập kết quả thi hành được dựa trên các mô hình được đánh giá
  - Lặp kiểu vòng đời phát triển phần mềm
  - [CCG98] Kenneth Collier, Bernard Carey, Ellen Grusy, Curt Marjaniemi, Donald Sautter (1998). A Perspective on Data Mining, *Technical Reporrt*, Northern Arizona University.

Mô hình CRISP-DM 2000



- Quy trình chuẩn tham chiếu công nghiệp KPDL
  - Các pha trong mô hình quy trình CRISP-DM (Cross-Industry Standard Process for Data Mining). “Hiểu kinh doanh”: hiểu bài toán và đánh giá
  - Thi hành chỉ sau khi tham chiếu kết quả với “hiểu kinh doanh”
  - CRISP-DM 2.0 SIG WORKSHOP, LONDON, 18/01/2007
  - Nguồn: http://www.crisp-dm.org/Process/index.htm (13/02/2011)

Mô hình tích hợp DM-BI [WW08]



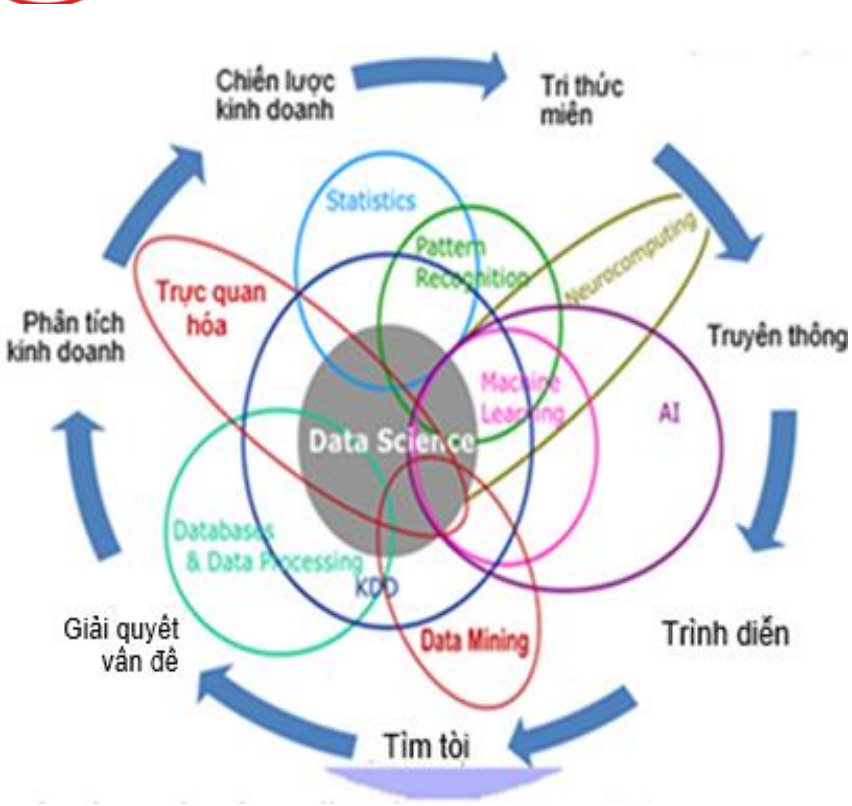
Chu trình phát triển tri thức thông qua khai phá dữ liệu

Wang, H. and S. Wang (2008). A knowledge management approach to data mining process for business intelligence, *Industrial Management & Data Systems*, 2008. **108**(5): 622-634. **[Oha09]**

Khoa học dữ liệu

- Data science is an emerging field in industry, and as yet, it is not welldefined as an academic subject.
- Van der Aalst
  - Làm thế nào sử dụng toàn bộ thông tin đó để cải thiện quy trình và máy móc, nâng cao hiệu quả chúng, và ngăn chặn trục trặc ?“
  - "Làm thế nào chúng ta có thể sử dụng thông tin để tác động tới các hành vi không mong muốn? Có cách nào để cho mọi người phản hồi về lối sống của họ…? "

Khoa học dữ liệu



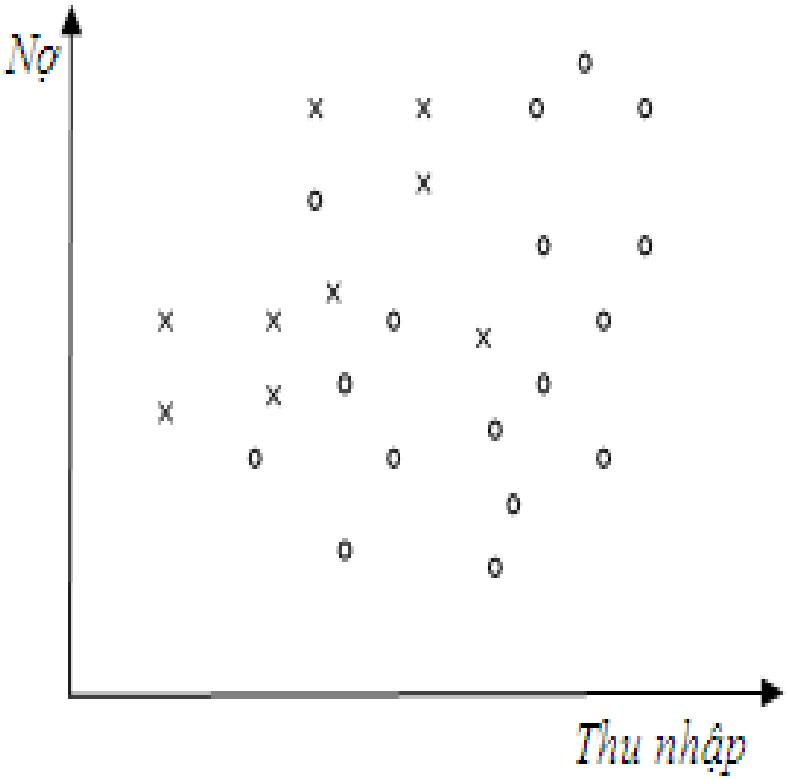
Quá trình khoa học dữ liệu (trái) và các chuyên ngành liên quan (phải)

Lưu ý:

- Khoa học hành vi và các khoa học xã hội
- Mô hình kinh doanh và tiếp thị
- Bảo mật, an ninh, pháp luật và đạo đức

Dữ liệu và Mẫu

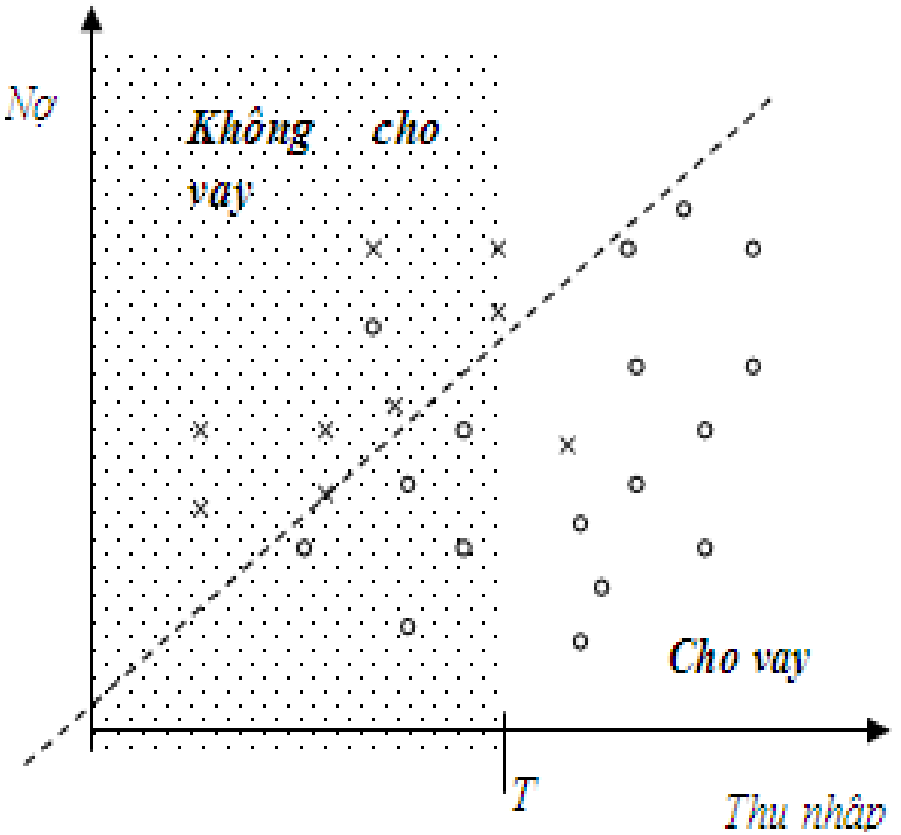
- *Dữ liệu (tập dữ liệu)*
  - tập F gồm hữu hạn các *trường hợp* (*sự kiện*).
  - KDD:phải gồm rất nhiều trường hợp
- *Mẫu*
  - Trong KDD: ngôn ngữ L để biểu diễn các tập con các sự kiện (dữ liệu) thuộc vào tập sự kiện F,
  - Mẫu: biểu thức E trong ngôn ngữ L ⇔ tập con F_E tương ứng các sự kiện trong F. E được gọi là *mẫu* nếu nó đơn giản hơn so với việc liệt kê các sự kiện thuộc F_E.
  - Chẳng hạn, biểu thức "THUNHẬP < $t" (mô hình chứa một biến THUNHẬP)



Hình 1.2. Tập dữ liệu có hai lớp x và o

Tính có giá trị

- Mẫu được phát hiện: phải có *giá trị* đối với các dữ liệu mới theo độ chân thực nào đấy.
- Tính "có giá trị" : một *độ đo tính có giá trị (chân thực)* là một hàm C ánh xạ một biểu thức thuộc ngôn ngữ biểu diễn mẫu L tới một không gian đo được (bộ phận hoặc toàn bộ) M_C.
- Chẳng hạn, đường biên xác định mẫu "THUNHẬP < $t“ dịch sang phải (biến THUNHẬP nhận giá trị lớn hơn) thì độ chân thực giảm xuống do bao gói thêm các tình huống vay tốt lại bị đưa vào vùng không cho vay nợ.
- Nếu *a**THUNHẬP + *b**NỢ < 0 mẫu có giá trị hơn.



Hình 1.3. Ngưỡng đơn T theo thu nhập để phân lớp cho vay (Lưu ý, đường nghiêng rời nét cho quyết định tốt hơn).

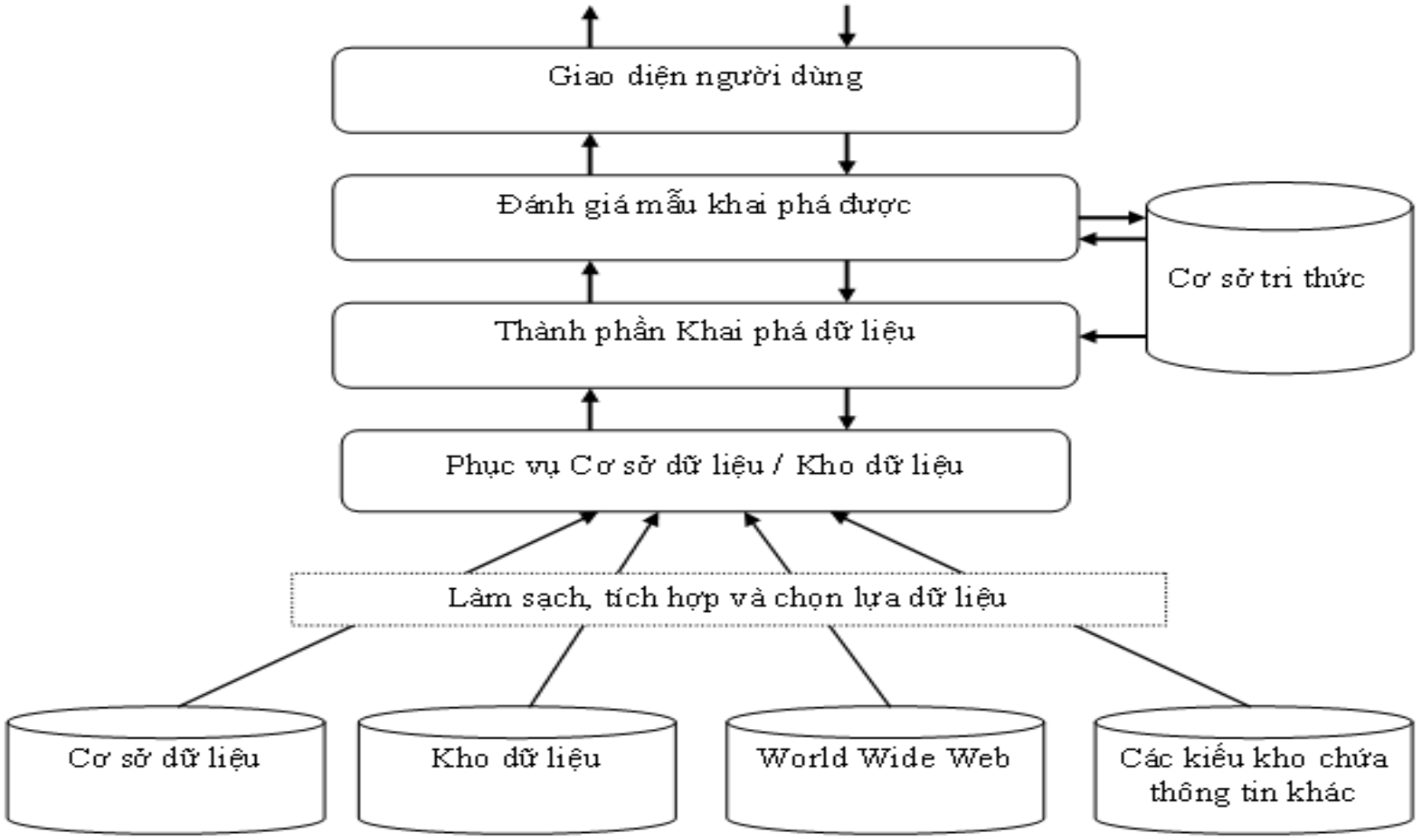
Tính mới và hữu dụng tiềm năng

- *Tính mới*: Mẫu phải là mới trong một miền xem xét nào đó, ít nhất là hệ thống đang được xem xét.
  - *Tính mới có thể đo được* :
    - sự thay đổi trong dữ liệu: so sánh giá trị hiện tại với giá trị quá khứ hoặc giá trị kỳ vọng
    - hoặc tri thức: tri thức mới quan hệ như thế nào với các tri thức đã có. Ví dụ thầy Nguyễn Đức Dũng tại Trường hè DMSS: **luật kết hợp hiếm**?
    - Tổng quát, điều này có thể được đo bằng một hàm N(E,F) hoặc là độ đo về tính mới hoặc là độ đo kỳ vọng.
- *Hữu dụng tiềm năng*: Mẫu cần có khả năng chỉ dẫn tới các tác động hữu dụng và *được đo bởi một hàm tiện ích*.
  - Hàm U ánh xạ các biểu thức trong L tới một không gian đo có thứ tự (bộ phận hoặc toàn bộ) M_U: u = U (E,F).
  - Ví dụ, trong tập dữ liệu vay nợ, hàm này có thể là *sự tăng hy vọng theo sự tăng lãi của nhà băng* (tính theo đơn vị tiền tệ) kết hợp với quy tắc quyết định được trình bày trong Hình 1.3.

Tính hiểu được, tính hấp dẫn và tri thức

- *Tính hiểu được*: Mẫu phải hiểu được
  - KDD: *mẫu mà con người hiểu chúng dễ dàng hơn* các dữ liệu nền.
  - Khó đo được một cách chính xác: "có thể hiểu được“ $\Leftrightarrow$ dễ hiểu.
  - Tồn tại một số độ đo dễ hiểu:
    - Sắp xếp từ cú pháp (tức là cỡ của mẫu theo bit) tới ngữ nghĩa (tức là dễ dàng để con người nhận thức được theo một tác động nào đó).
    - Giả định rằng tính hiểu được là *đo được* bằng một hàm S ánh xạ biểu thức E trong L tới một không gian đo được có thứ tự (bộ phận /toàn bộ) M_S: s = S(E,F).
- *Tính hấp dẫn*: *độ đo tổng thể về mẫu* là sự kết hợp của các tiêu chí *giá trị*, *mới*, *hữu ích* và *dễ hiểu*.
  - Hoặc dùng một hàm hấp dẫn: i = I (E, F, C, N, U, S) ánh xạ biểu thức trong L vào một không gian đo được M_i.
  - Hoặc xác định độ hấp dẫn trực tiếp: thứ tự của các mẫu được phát hiện.
- *Tri thức*: Một mẫu E $\in$ L được gọi là *tri thức* nếu như đối với một lớp người sử dụng nào đó, chỉ ra được một ngưỡng i $\in$ M_i mà độ hấp dẫn I(E,F,C,N,U,S) > i.

Kiến trúc điển hình hệ thống KPDL



Hình 1.6. Kiến trúc điển hình của hệ thống khai phá dữ liệu

Truy vấn CSDL

- Truy vấn hệ quản trị CSDL
  - Hãy hiển thị số tiền Ông Smith trong ngày 5 tháng Giêng ? *ghi nhận riêng lẻ do xử lý giao dịch trực tuyến* (on-line transaction processing – OLTP)
  - Có bao nhiêu nhà đầu tư nước ngoài mua cổ phiếu X trong tháng trước ? *ghi nhận thống kê do hệ thống hỗ trợ quyết định thống kê* (stastical decision suppport system - DSS)
  - Hiển thị mọi cổ phiếu trong CSDL với mệnh giá tăng ? *ghi nhận dữ liệu đa chiều do xử lý phân tích trực tuyến* (on-line analytic processing - OLAP).
- Cần giả thiết
  - *Tính "đầy đủ" về tri thức miền phức tạp*!
  - Câu trả lời chính xác

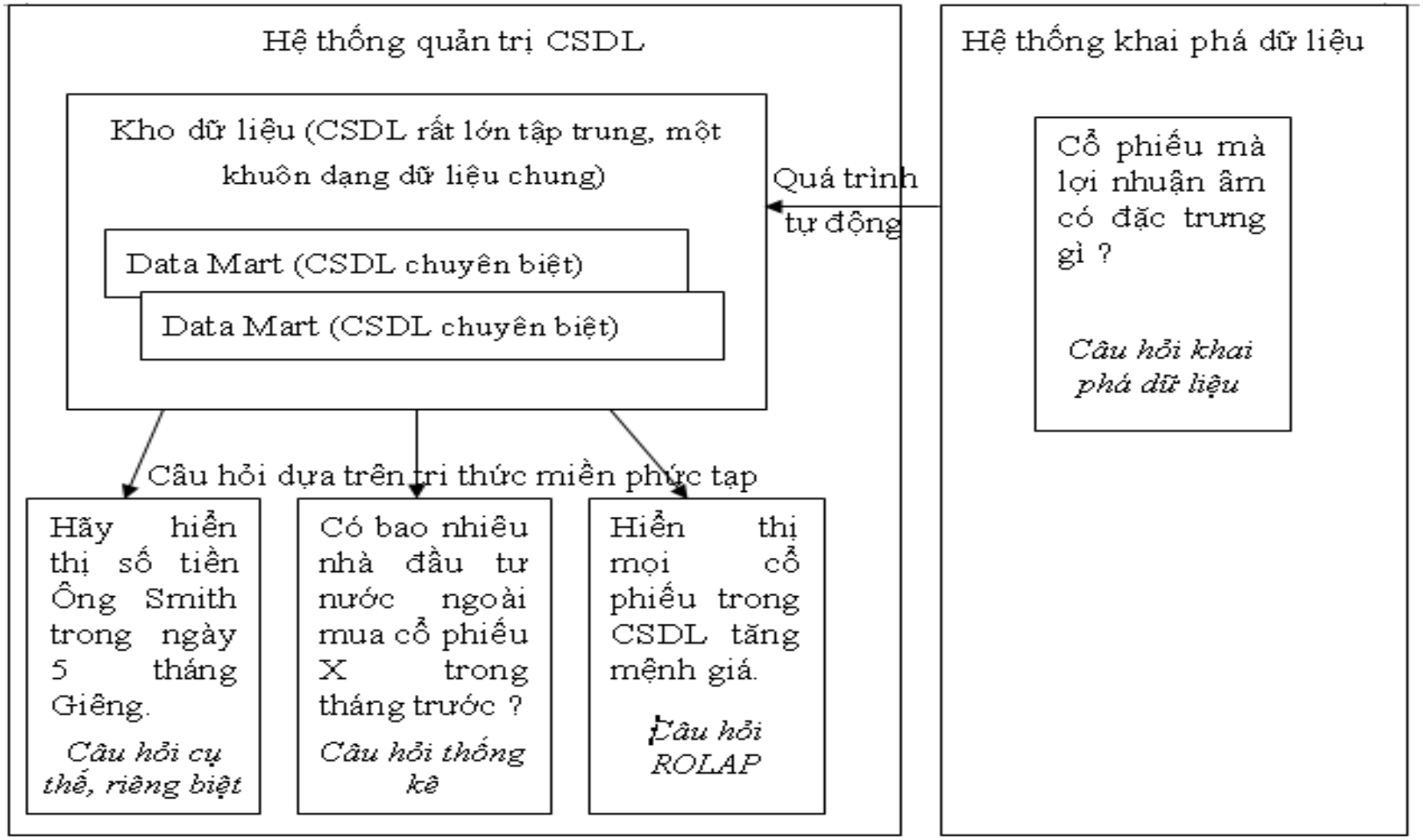
Truy vấn Khai phá dữ liệu

- Ví dụ truy vấn
  - Các cổ phiếu tăng giá có đặc trưng gì ?
  - Tỷ giá US$ - DMark có đặc trưng gì ?
  - Hy vọng gì về cổ phiếu X trong tuần tiếp theo ?
  - Trong tháng tiếp theo, sẽ có bao nhiêu đoàn viên công đoàn không trả được nợ của họ ?
  - Những người mua sản phẩm Y có đặc trưng gì ?
- Nhận xét
  - *Giả thiết tri thức "đầy đủ" không còn có tính cốt lõi*, cần bổ sung tri thức cho hệ thống → Cải tiến (nâng cấp) miền tri thức !
  - Câu trả lời có tính xấp xỉ, gần đúng

Mục tiêu Khai phá dữ liệu

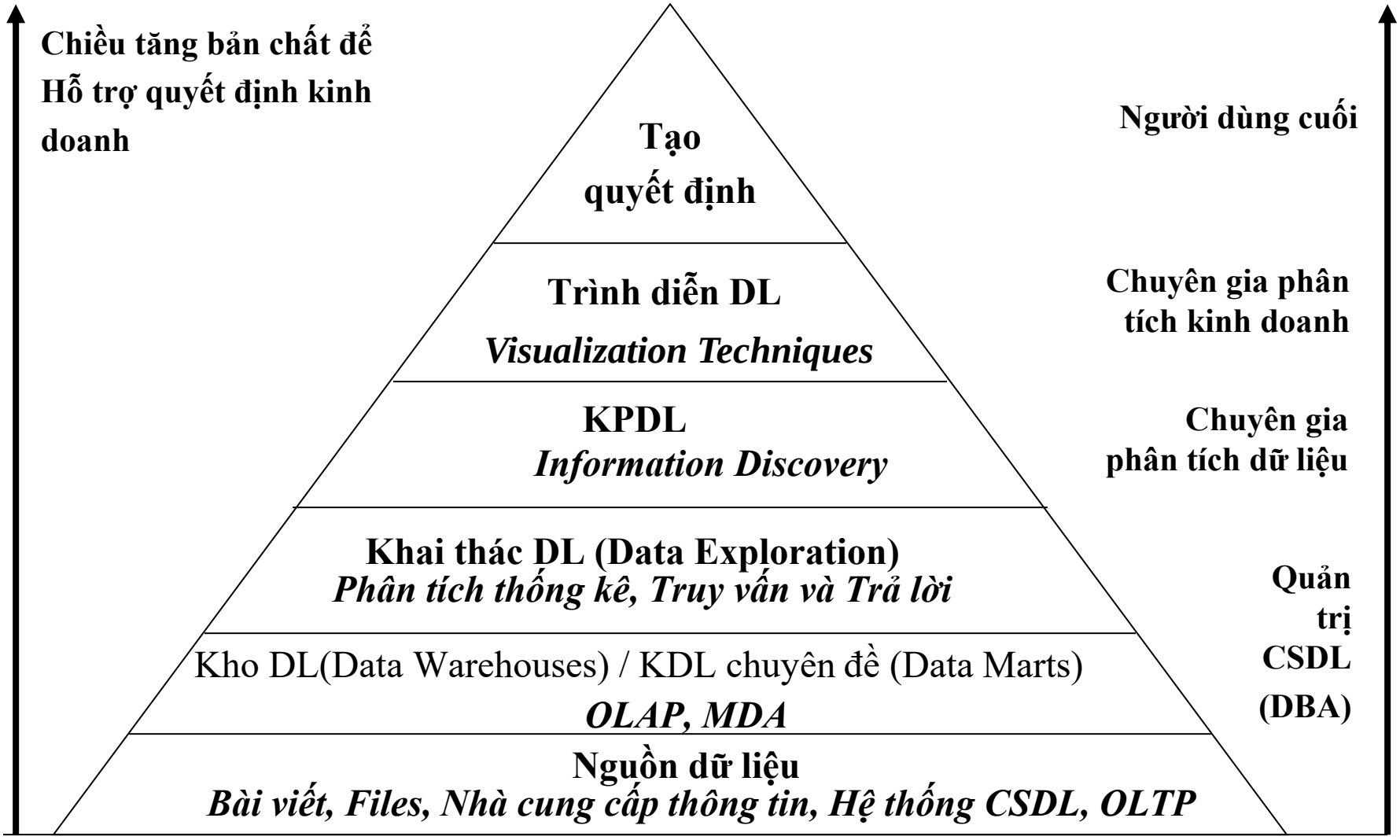
- Ví dụ
  - Giảm 3% lượng khách hàng hiện thời rời bỏ (*duy trì khách hàng*)
  - Tăng 2% số hợp đồng của khách hàng mới (*thu hút KH*)
  - Tăng 5% doanh thu từ việc bán chéo cho khách hàng hiện có (*phát triển khách hang*)
  - Dự báo thị phần khán giả truyền hình với xác suất 70% (*dự báo kênh tiếp thị*)
  - Dự báo với độ chính xác 75% lượng khách hang ký hợp đồng với sản phẩm mới (*dự báo thu hút khách hàng*)
  - Xác định phân lớp mới khách hàng và sản phẩm (*đặc trưng KH*)
  - Tạo một mô hình phân khúc khách hang mới (*phân khúc KH*)
- Nhận xét
  - Cần hiểu được bài toán và mục tiêu kinh doanh
  - Các ví dụ trên liên quan tới quản lý quan hệ khách hàng

Hệ thống CSDL và Hệ thống KPDL



Hình 1.7. Mối quan hệ giữa hệ thống CSDL và hệ thống khai phá dữ liệu

KPDL và Thông minh kinh doanh

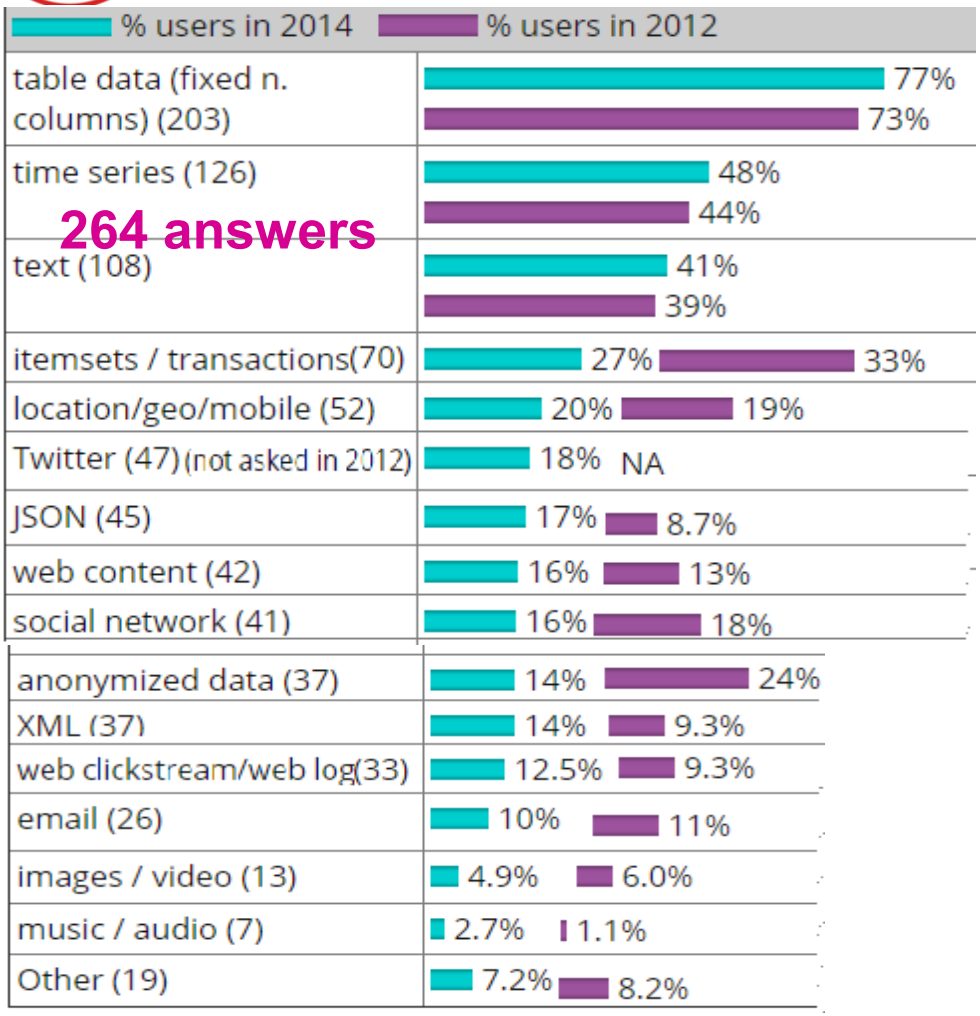


4. KPDL: các kiểu dữ liệu

- CSDL quan hệ
- Kho dữ liệu
- CSDL giao dịch
- CSDL mở rộng và kho chứa thông tin
  - CSDL quan hệ-đối tượng
  - Dữ liệu không gian và thời gian
  - Dữ liệu chuỗi thời gian
  - Dữ liệu dòng
  - Dữ liệu đa phương tiện
  - Dữ liệu không đồng nhất và thừa kế
  - CSDL Text & WWW

Kiểu dữ liệu được phân tích/khai phá

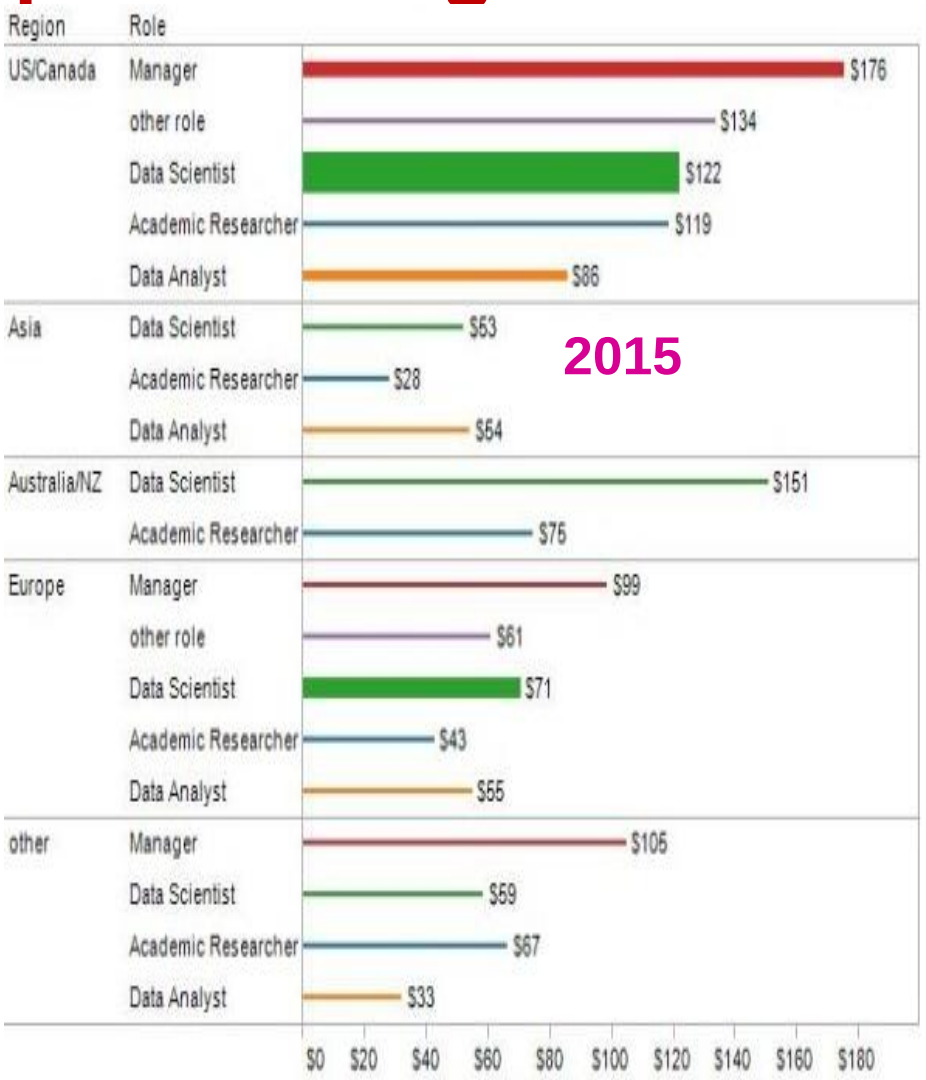
http://www.kdnuggets.com/polls/2014/data-types-sources-analyzed.html



- Nhận xét
  - Dữ liệu (cơ sở dữ liệu) quan hệ: bảng: Hầu hết **203/264**
  - Chuỗi thời gian, giao dịch, văn bản, ẩn danh, mạng xã hội…

Kích thước dữ liệu và lương KPDL

http://www.kdnuggets.com/2013/04/poll-results-largest-dataset-analyzed-data-mined.html



http://www.kdnuggets.com/2015/03/salary-analytics-data-science-poll-well-compensated.html

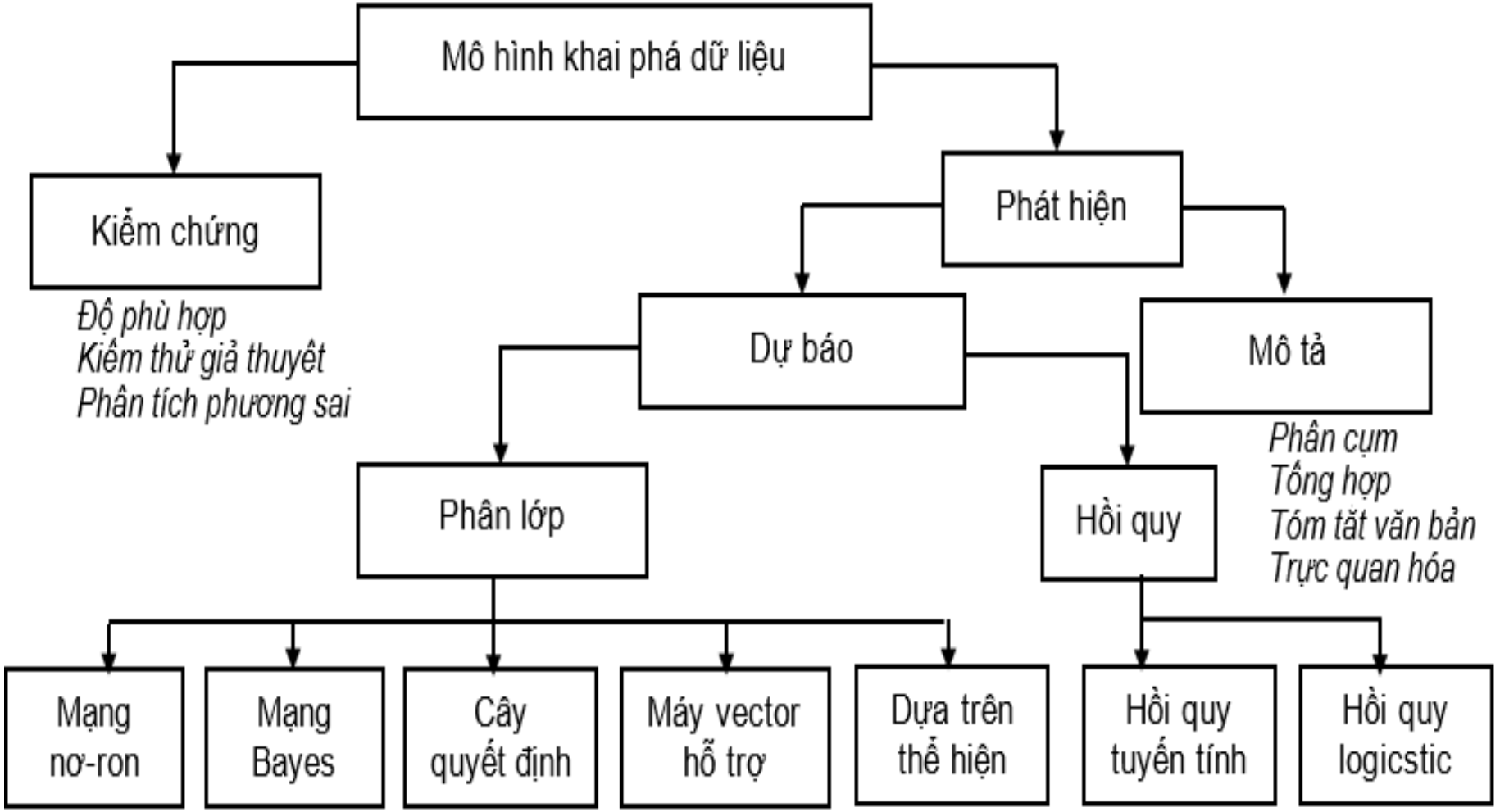
5. KPDL: Kiểu mẫu được khai phá

- Chức năng chung
  - KPDL mô tả: tóm tắt, phân cụm, luật kết hợp…
  - KPDL dự đoán: phân lớp, hồi quy…
- Các bài toán điển hình
  - Mô tả khái niệm
  - Quan hệ kết hợp
  - Phân lớp
  - Phân cụm
  - Hồi quy
  - Mô hình phụ thuộc
  - Phát hiện biến đổi và độ lệch
  - Phân tích định hướng mẫu, các bài toán khác

KPDL: Sơ đồ phân loại

- Phân loại theo bài toán
  - **Dự báo và mô tả**
  - Mở rộng: Kiểm chứng và phát hiện
- Phân loại theo khung nhìn
  - Kiểu dữ liệu được KP
  - Kiểu tri thức cần phát hiện
  - Kiểu kỹ thuật được dùng
  - Kiểu miền ứng dụng

P/pháp KPDL: Kiểm chứng và Phát hiện



- **L. Rokach and O. Maimon (2015). *Data Mining with Decision Trees: Theory and Applications*. World Scientific Publishing.**

KPDL: Hai loại, Loại dự báo: phân lớp

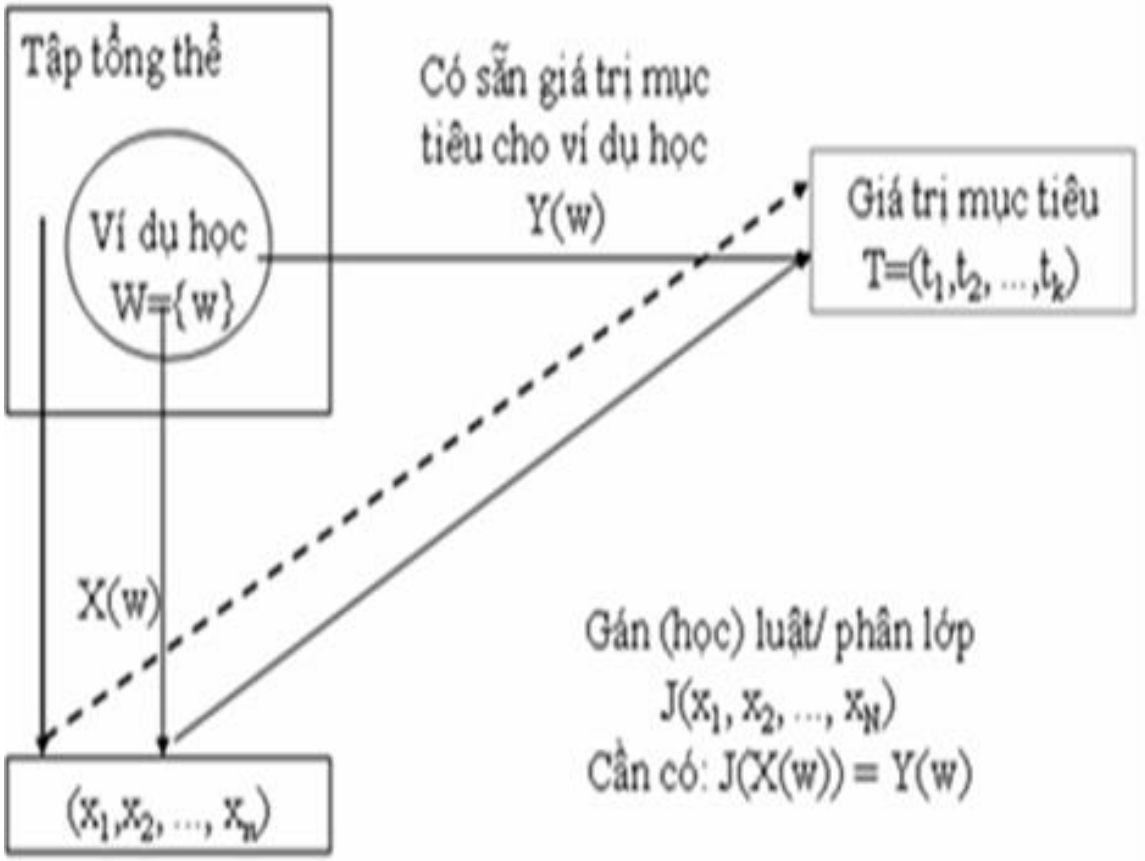
Predictive Analytics	Descriptive Analytics
Classification Regression Survival analysis Forecasting	Clustering Association analysis Sequence analysis

- Loại KPDL dự báo
  - Xây dựng các mô hình (chức năng) để mô tả và phân biệt khái niệm cho các lớp hoặc khái niệm để dự đoán trong tương lai
    - Chẳng hạn, phân lớp quốc gia dựa theo khí hậu, hoặc phân lớp ô tô dựa theo tiêu tốn xăng
  - Trình diễn: cây quyết định, luật phân lớp, mạng nơron
  - Dự đoán giá trị số chưa biết hoặc đã mất

[Baesens18] Bart Baesens, Cristián Bravo, Wouter Verbeke. Profit-driven business analytics: a practitioner's guide to transforming big data into added value. Wiley, 2018

Phân lớp

- Phân lớp
  - xây dựng/mô tả mô hình/ hàm dự báo để mô tả/phát hiện lớp/khái niệm cho dự báo tiếp
  - học một hàm ánh xạ dữ liệu vào một trong một số lớp đã biết



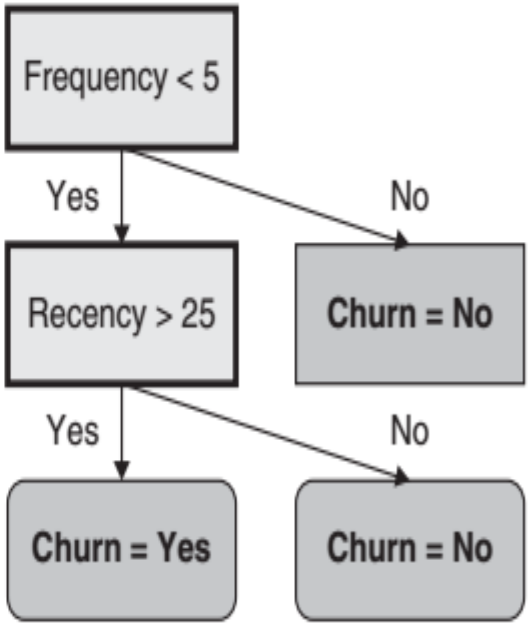
Charu C. Aggarwal. *Data Classification: Algorithms and Applications*. CRC Press 2014

KPDL Dự báo: Phân lớp

Example dataset	Predictive analytical model
Classification	Phân lớp, ví dụ, phân lớp cây quyết định dự báo khách hàng rời bỏ

ID	Recency	Frequency	Monetary	Churn
C1	26	4.2	126	Yes
C2	37	2.1	59	No
C3	2	8.5	256	No
C4	18	6.2	89	No
C5	46	1.1	37	Yes
...	...	...	...	...

Decision tree classification model:



KPDL dự báo:Hồi quy, mô hình phụ thuộc

- Hồi quy
  - học một hàm ánh xạ dữ liệu nhằm xác định giá trị thực của một biến theo một số biến khác
  - điển hình trong phân tích thống kê và dự báo
  - dự đoán giá trị của một/một số biến phụ thuộc vào giá trị của một tập biến độc lập.
- Mô hình phụ thuộc
  - xây dựng mô hình phụ thuộc: tìm một mô hình mô tả sự phụ thuộc có ý nghĩa giữa các biến
  - mức cấu trúc:
    - dạng đồ thị
    - biến là phụ thuộc bộ phận vào các biến khác
  - mức định lượng: tính phụ thuộc khi sử dụng việc đo tính theo giá trị số

Loại KPDL mô tả: Hồi quy

Example dataset	Predictive analytical model

Regression

Hồi quy, ví dụ, hồi quy giá trị thời gian khách hàng (customer life-time value: CLV) theo mô hình RFM hành vi

ID	Recency	Frequency	Monetary	CLV
C1	26	4.2	126	3,817
C2	37	2.1	59	4,31
C3	2	8.5	256	2,187
C4	18	6.2	89	543
C5	46	1.1	37	1,548
...	...	...	...	...

Linear regression model:

$$CLV = 260 + 11 \cdot Recency + 6.1 \cdot Frequency + 3.4 \cdot Monetary$$

18_Baesens, Bart_ Bravo, Cristián_ Verbeke, Wouter. *Profit-driven business analytics: a practitioner's guide to transforming big data into added value*. Wiley, 2018

Loại KPDL mô tả : Phân tích tồn tại

Example dataset	Predictive analytical model
Survival analysis — Phân tích tồn tại, ví dụ, thời gian rời khỏi hoặc cần giám sát	

ID	Recency	Churn or Censored	Time of churn or Censoring
C1	26	Churn	181
C2	37	Censored	253
C3	2	Censored	37
C4	18	Censored	172
C5	46	Churn	98
...	...	...	...

General parametric survival analysis model:

$$\log(T) = 13 + 5.3 \cdot \text{Recency}$$

18_Baesens, Bart_ Bravo, Cristián_ Verbeke, Wouter. *Profit-driven business analytics: a practitioner's guide to transforming big data into added value*. Wiley, 2018

Loại KPDL dự báo: Dự đoán

Example dataset	Predictive analytical model
Forecasting Dự đoán, ví dụ dự đoán nhu cầu	

Timestamp	Demand
January	513
February	652
March	435
April	578
May	601
...	...

Weighted moving average forecasting model:

$$\text{Demand}_t = 0.4 \cdot \text{Demand}_{t-1} + 0.3 \cdot \text{Demand}_{t-2} + 0.2 \cdot \text{Demand}_{t-3} + 0.1 \cdot \text{Demand}_{t-4}$$

18_Baesens, Bart_ Bravo, Cristián_ Verbeke, Wouter. *Profit-driven business analytics: a practitioner's guide to transforming big data into added value*. Wiley, 2018

KPDL dự báo khác

- Phân tích xu hướng và tiến hóa
  - Xu hướng và độ lệch: phân tích hồi quy
  - Khai phá mẫu tuần tự, phân tích chu kỳ
  - Phân tích dựa trên tương tự
- Phân tích định hướng mẫu khác hoặc phân tích thống kê

KPDL mô tả

- Phân tích cụm
  - Nhãn lớp chưa biết: Nhóm dữ liệu thành các lớp mới: phân cụm các nhà để tìm mẫu phân bố
  - Cực đại tương tự nội bộ cụm & cực tiểu tương tự giữa các cụm
- Phân tích bất thường
  - Bất thường: đối tượng dữ liệu không tuân theo hành vi chung của toàn bộ dữ liệu. *Ví dụ, sử dụng kỳ vọng mẫu và phương sai mẫu*
  - Nhiễu hoặc ngoại lệ? Không phải! Hữu dụng để phát hiện gian lận, phân tích các sự kiện hiếm
- Phát hiện biến đổi và độ lệch
  - Hầu như sự thay đổi có ý nghĩa dưới dạng độ đo đã biết trước/giá trị chuẩn, cung cấp tri thức về sự biến đổi và độ lệch
  - Phát hiện biến đổi và độ lệch <> tiền xử lý

Manish Gupta, Jing Gao, Charu C. Aggarwal, Jiawei Han. *Outlier Detection for Temporal Data*. Morgan & Claypool Publishers 2014

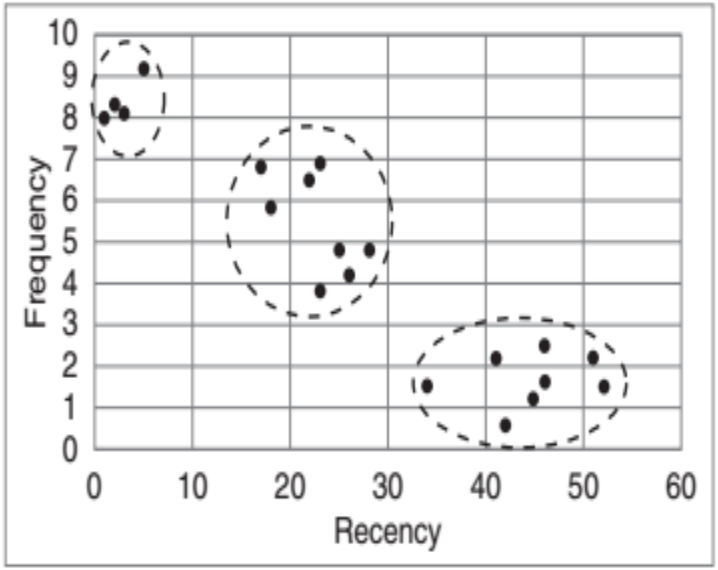
Charu C. Aggarwal. *Outlier Analysis*. Springer 2013

Loại KPDL Mô tả: phân cụm

Data	Descriptive analytical model
Clustering Phân cụm, ví dụ phân cụm khách hàng theo RF	

ID	Recency	Frequency
C1	26	4.2
C2	37	2.1
C3	2	8.5
C4	18	6.2
C5	46	1.1
...	...	...

K-means clustering with $K = 3$:



18_Baesens, Bart_ Bravo, Cristián_ Verbeke, Wouter. *Profit-driven business analytics: a practitioner's guide to transforming big data into added value*. Wiley, 2018

Loại KPDL Mô tả: phân tích kết hợp

Data	Descriptive analytical model
Association analysis Phân tích kết hợp, ví dụ phân tích luật kết hợp	

ID	Items
T1	beer, pizza, diapers, baby food
T2	coke, beer, diapers
T3	crisps, diapers, baby food
T4	chocolates, diapers, pizza, apples
T5	tomatoes, water, oranges, beer
...	...

Association rules:

If *baby food* **And** *diapers* **Then** *beer*

If *coke* **And** *pizza* **Then** *crisps*

...

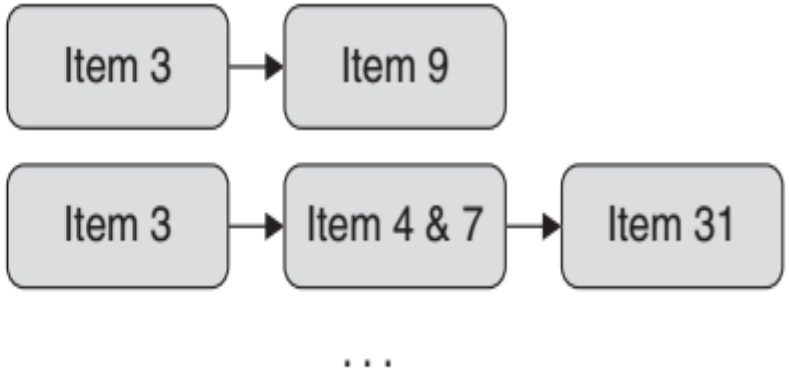
18_Baesens, Bart_ Bravo, Cristián_ Verbeke, Wouter. *Profit-driven business analytics: a practitioner's guide to transforming big data into added value*. Wiley, 2018

Loại KPDL Mô tả: phân tích dãy

Data	Descriptive analytical model
Sequence analysis Phân tích dãy, ví dụ phân tích dãy mua hàng	

ID	Sequential items
C1	<{3},{9}>
C2	<{1 2},{3},{4 6 7}>
C3	<{3 5 7}>
C4	<{3},{4 7},{9}>
C5	<{9}>
...	...

Sequence rules:



18_Baesens, Bart_ Bravo, Cristián_ Verbeke, Wouter. *Profit-driven business analytics: a practitioner's guide to transforming big data into added value*. Wiley, 2018

KPDL: Mô tả

- Mô tả khái niệm: Đặc trưng và phân biệt
  - Tìm các đặc trưng và tính chất của khái niệm
  - Tổng quát hóa, tóm tắt, phát hiện đặc trưng ràng buộc, tương phản, chẳng hạn, các vùng khô so sánh với ướt
  - Bài toán mô tả điển hình: Tóm tắt (tìm mô tả cô đọng)
    - Kỳ vọng, phương sai
    - Tóm tắt văn bản
- Quan hệ kết hợp
  - Quan hệ kết hợp giữa các biến dữ liệu: Tương quan và nhân quả)
  - Diaper → Beer [0.5%, 75%]
  - Luật kết hợp: X→Y
  - Ví dụ, trong khai phá dữ liệu Web
    - Phát hiện quan hệ ngữ nghĩa
    - Quan hệ nội dung trang web với mối quan tâm người dùng

Charu C. Aggarwal, Jiawei Han. *Frequent Pattern Mining*. Springer 2014

Khung nhìn đa chiều của KPDL

- **Dữ liệu được khai phá**
  - Quan hệ, KDL, giao dịch, dòng, hướng đối tượng/quan hệ, tích cực, không gian, chuỗi thời gian, văn bản, đa phương tiện, không đồng nahats, kế thừa, WWW
- **Tri thức được khai phá**
  - Đặc trưng, phân biệt, kết hợp, phân lớp, phân cụm, xu hướng/độ lệch, phân tích bất thường,…
  - Các chức năng phức/tích hợp và KPDL các mức phức hợp
- **Kỹ thuật được dùng**
  - Định hướng CSDL, KDL (OLAP), học máy, thống kê, trực quan hóa, ….
- **Ứng dụng phù hợp**
  - Bán lẻ, viễn thông, ngân hàng, phân tích gian lận, KPDL sinh học, phân tích thị trường chứng khoán, KP văn bản, KP Web, …

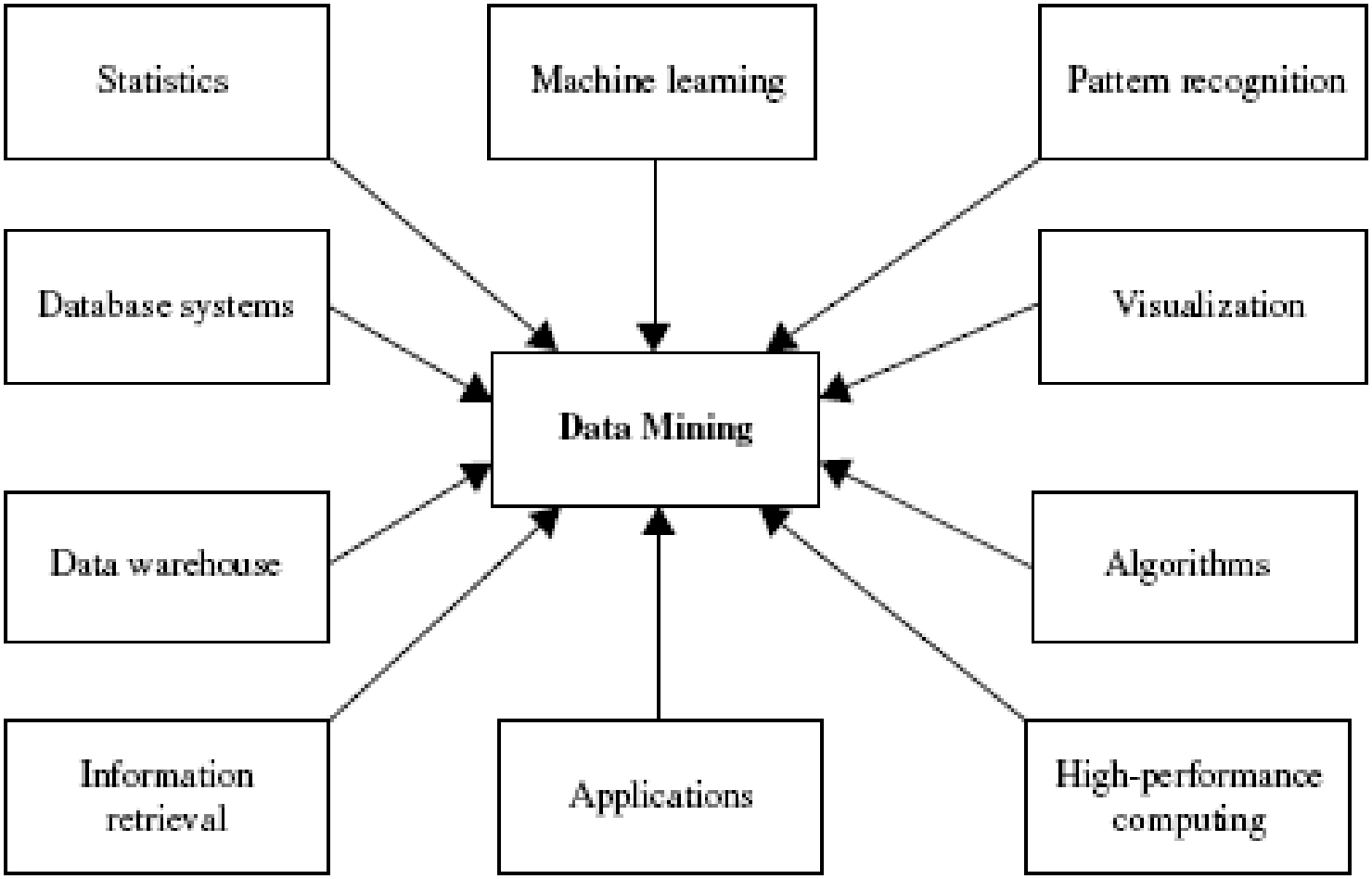
Mọi mẫu khai phá được đều hấp dẫn?

- KPDL có thể sinh ra tới hàng nghìn mẫu: Không phải tất cả đều hấp dẫn
  - Tiếp cận gợi ý: KPDL hướng người dùng, dựa trên câu hỏi, hướng đích
- **Độ đo hấp dẫn**
  - Mẫu là hấp dẫn nếu dễ hiểu, có giá trị theo dữ liệu mới/kiểm tra với độ chắc chắn, hữu dụng tiềm năng, mới lạ hoặc xác nhận các giả thiết mà người dùng tìm kiếm để xác thực.
- **Độ đo hấp dẫn khách quan và chủ quan**
  - Khách quan: dựa trên thống kê và cấu trúc của mẫu, chẳng hạn, độ hỗ trợ, độ tin cậy, …
  - Chủ quan: dựa trên sự tin tưởng của người dùng đối với dữ liệu, chẳng hạn, sự không chờ đón, tính mới mẻ, tác động được...

Tìm được tất cả và chỉ các mẫu hấp dẫn?

- Tìm được mọi mẫu hấp dẫn: Về tính đầy đủ
  - Hệ thống KHDL có khả năng tìm **mọi** mẫu hấp dẫn?
  - Tìm kiếm mày mò (heuristic) <> tìm kiếm đầy đủ
  - Kết hợp <> phan lớp <> phân cụm
- Tìm chỉ các mẫu hấp dẫn: Về tính tối ưu
  - Hệ thống KPDL có khả năng tìm ra **đúng** các mẫu hấp dẫn?
  - Tiếp cận
    - Đầu tiên tìm tổng thể tất cả các mẫu sau đó lọc bỏ các mẫu không hấp dẫn.
    - Sinh ra chỉ các mẫu hấp dẫn—tối ưu hóa câu hỏi khai phá

KPDL: Các công nghệ chính



Hội tụ của nhiều ngành phức [HKP11]

Thống kê toán học với KPDL

- Nhiều điểm chung giữa KPDL với *thống kê*:
  - Đặc biệt như phân tích dữ liệu thăm dò (EDA: Exploratory Data Analysis) cũng như dự báo [Fied97, HD03].
  - Hệ thống KDD thường gắn kết với các thủ tục thống kê đặc biệt đối với mô hình dữ liệu và nắm bắt nhiễu trong một khung cảnh phát hiện tri thức tổng thể.
  - Các phương pháp KPDL dựa theo thống kê nhận được sự quan tâm đặc biệt.

Thống kê toán học với KPDL

- Phân biệt giữa bài toán thống kê và bài toán khai phá dữ liệu
  - Kiểm định giả thiết TK: một mô hình giả thiết + tập dữ liệu quan sát được. Kiểm tra: tập dữ liệu có phù hợp với giả thiết thống kê hay không/ giả thiết thống kê có đúng trên toàn bộ dữ liệu quan sát được hay không.
  - Bài toán học KPDL: Cho tập dữ liệu (mô hình chưa có). Mô hình kết quả phải phù hợp với tập toàn bộ dữ liệu -> đảm bảo các tham số mô hình không phụ thuộc vào cách chọn tập dữ liệu học. Học KPDL đòi hỏi tập dữ liệu học/tập dữ liệu kiểm tra cần "đại diện" cho toàn bộ dữ liệu trong miền ứng dụng và cần độc lập nhau. Một số trường hợp: hai tập dữ liệu này (hoặc tập dữ liệu kiểm tra) được công bố dưới dạng chuẩn.
  - Về thuật ngữ: KPDL: *biến ra/biến mục tiêu*, *thuật toán khai phá dữ liệu*, *thuộc tính/đặc trưng*, *bản ghi*... XLDLTK: *biến phụ thuộc*, *thủ tục thống kê*, *biến giải thích*, *quan sát*... Tham khảo thêm từ Nguyễn Xuân Long

Thống kê	*Phân tích dự báo*
Các mô hình dựa trên lý thuyết: Tồn tại một phương án tối ưu	Mô hình thường dựa trên các thuật toán phi tham số, không đảm bảo tối ưu
Mô hình điển hình tuyến tính	Mô hình điển hình phi tuyến
Dữ liệu thường nhỏ hơn, các thuật toán thường thiên về phía chính xác đối với dữ liệu nhỏ	Khả cỡ với dữ liệu lớn, các thuật toán không phải là hiệu quả và ổn định đối với dữ liệu nhỏ
Mô hình là vua	**Dữ liệu là vua**

Học máy với KPDL

- Học máy
  - Machine Learning
  - Cách máy tính học (nâng cao năng lực) dựa trên dữ liệu.
  - Chương trình máy tính tự động học được mẫu phức tạp và ra quyết định thông minh dựa trên dữ liệu, ví dụ, *"học được chữ viết tay trên thư thông qua một tập ví dụ"*.
  - Học máy là lĩnh vực nghiên cứu phát triển nhanh
- Một số nội dung học máy với khai phá dữ liệu
  - *Nhiều nội dung đã được trình bày tại mục trước*
  - Học giám sát (supervised learning) đồng nghĩa với phân lớp (classification)
  - Học không giám sát (unsupervised) ≈ phân cụm (clustering),
  - Học bán giám sát (semi-supervised learning) sử dụng cả ví dụ có nhãn và ví dụ không có nhãn
  - Học tích cực (Active learning) còn được gọi là học tương tác (interactive learning) có tương tác với người dùng.
  - Học tăng cường (incremental learning) mẫu đầu vào là liên tục và mô hình học phù hợp với ví dụ cập nhật.
  - *Các khung học máy khác*

Tìm kiếm thông tin với KPDL

- Tìm kiếm thông tin
  - Information Retrieval. “Truy hồi thông tin”
  - Tìm kiếm tài liệu hoặc tìm kiếm thông tin trong tài liệu theo một truy vấn. Tài liệu: văn bản, đa phương tiện, web…
  - Hai giả thiết: (i) Dữ liệu tìm kiếm là không cấu trúc; (ii) Truy vấn dưới dạng từ khóa/cụm từ khóa mà không phải cấu trúc phức tạp
- Tìm kiếm thông tin với KPDL
  - Kết hợp mô hình tìm kiếm với kỹ thuật KPDL tìm thấy các chủ đề chính trong tập tài liệu, từng tài liệu … bổ sung thuộc tính dữ liệu quan trọng
  - KPDL văn bản, web, phương tiện xã hội liên quan mật thiết với tìm kiếm thông tin.

7. Ứng dụng cơ bản của KPDL

- **Phân tích dữ liệu và hỗ trợ quyết định**
  - Phân tích và quản lý thị trường
    - Tiếp thị định hướng, quản lý quan hệ khách hàng (CRM), phân tích thói quen mua hàng, bán hàng chéo, phân đoạn thị trường
  - Phân tích và quản lý rủi ro
    - Dự báo, duy trì khách hàng, cải thiện bảo lãnh, kiểm soát chất lượng, phân tích cạnh tranh
  - Phát hiện gian lận và phát hiện mẫu bất thường (ngoại lai)
- **Ứng dụng khác**
  - Khai phá Text (nhóm mới, email, tài liệu) và khai phá Web
  - Khai phá dữ liệu dòng
  - Phân tích DNA và dữ liệu sinh học

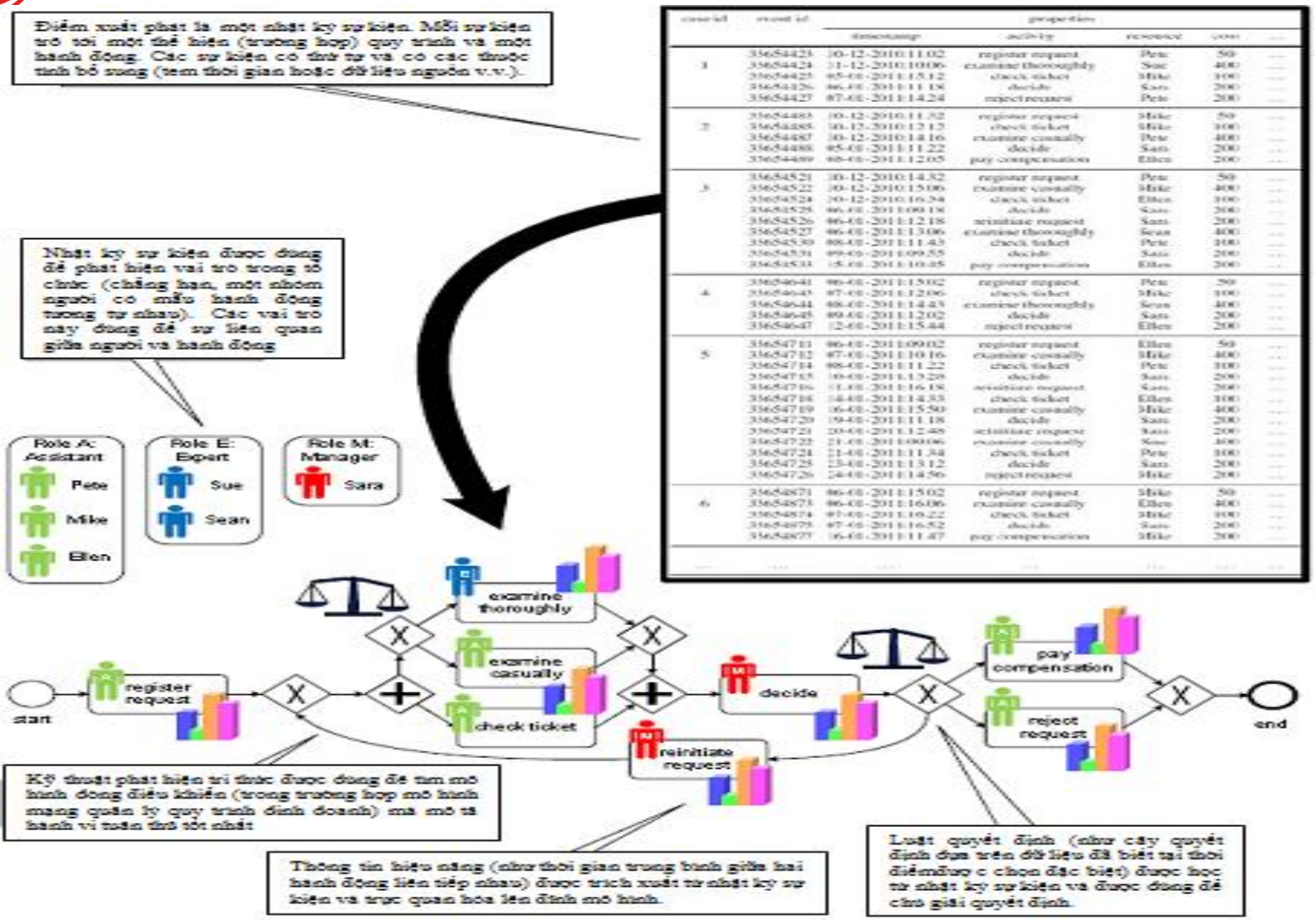
Phân tích và quản lý thị trường

- Nguồn dữ liệu có từ đâu ?
  - Giao dịch thẻ tín dụng, thẻ thành viên, phiếu giảm giá, các phàn nàn của khách hàng, các nghiên cứu phong cách sống (công cộng) bổ sung
- Tiếp thị định hướng
  - Tìm cụm các mô hình khách hàng cùng đặc trưng: sự quan tâm, mức thu nhập, thói quen chi tiêu...
  - Xác định các mẫu mua hàng theo thời gian
- Phân tích thị trường chéo
  - Quan hệ kết hợp/đồng quan hệ giữa bán hàng và dự báo dựa theo quan hệ kết hợp
- Hồ sơ khách hàng
  - Kiểu của khách hàng mua sản phẩm gì (phân cụm và phân lớp)
- Phân tích yêu cầu khách hàng
  - Định danh các sản phẩm tốt nhất tới khách hàng (khác nhau)
  - Dự báo các nhân tố sẽ thu hút khách hàng mới
- Cung cấp thông tin tóm tắt
  - Báo cáo tóm tắt đa chiều
  - Thông tin tóm tắt thống kê (xu hướng trung tâm dữ liệu và biến đổi)

Phân tích doanh nghiệp & Quản lý rủi ro

- Lên kế hoạch tài chính và đánh giá tài sản
  - Phân tích và dự báo dòng tiền mặt
  - Phân tích yêu cầu ngẫu nhiên để đánh giá tài sản
  - Phân tích lát cắt ngang và chuỗi thời gian (tỷ số tài chính, phân tích xu hướng…)
- Lên kế hoạch tài nguyên
  - Tóm tắt và so sánh các nguồn lực và chi tiêu
- Cạnh tranh
  - Theo dõi đối thủ cạnh tranh và định hướng thị trường
  - Nhóm khách hàng thành các lớp và định giá dựa theo lớp khách
  - Khởi tạo chiến lược giá trong thị trường cạnh tranh cao

Phân tích kinh doanh: Khai phá quy trình



[Aalst16] WMP Van der Aalst. *Process Mining: Data Science in Action (2nd edition)*. Springer, 2016

Phát hiện gian lận và khai phá mẫu hiếm

- Tiếp cận: Phân cụm & xây dựng mô hình gian lận, phân tích bất thường
- Ứng dụng: Chăm sóc sức khỏe, bán lẻ, dịch vụ thẻ tín dụng, viễn thông.
  - Bảo hiểm tự động: vòng xung đột
  - Rửa tiền: giao dịch tiền tệ đáng ngờ
  - Bảo hiểm y tế
    - Bệnh nghề nghiệp, nhóm bác sỹ, và nhóm chỉ dẫn
    - Xét nghiệm không cần thiết hoặc tương quan
  - Viễn thông: cuộc gọi gian lận
    - Mô hình cuộc gọi: đích cuộc gọi, độ dài, thời điểm trong ngày hoặc tuần. Phân tích mẫu lệch một dạng chuẩn dự kiến
  - Công nghiệp bán lẻ
    - Các nhà phân tích ước lượng rằng 38% giảm bán lẻ là do nhân viên không trung thực
  - Chống khủng bố

Ứng dụng khác

- Khai phá web và khai phá phương tiện xã hội
  - Trợ giúp IBM áp dụng các thuật toán KPDL biên bản truy nhập Web đối với các trang liên quan tới thị trường để khám phá ưu đãi khách hàng và các trang hành vi, phân tích tính hiệu quả của tiếp thị Web, cải thiệ cách tổ chức Website …
- Thể thao
  - IBM Advanced Scout phân tích thống kế môn NBA (chặn bóng, hỗ trợ và lỗi) để đưa tới lợi thế cạnh trang cho New York Knicks và Miami Heat
- Thiên văn học
  - JPL và Palomar Observatory khám phá 22 chuẩn tinh (quasar) với sự trợ giúp của KPDL

8. Vấn đề chính trong KPDL

Nguồn chỉ dẫn về KPDL

http://www.kdnuggets.com/

- Data mining and KDD (SIGKDD: CDROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations
- Database systems (SIGMOD: CD ROM)
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM-TODS, IEEE-TKDE, JIIS, J. ACM, etc.
- AI & Machine Learning
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.
- Statistics
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.
- Visualization
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.

http://www.kdnuggets.com/2015/09/free-data-science-books.html

Data Science in General

- **An Introduction to Data Science**
  Jeffrey Stanton, 2013
- **School of Data Handbook**
  School of Data, 2015
- **Data Jujitsu: The Art of Turning Data into Product**
  DJ Patil, 2012

Interviews with Data Scientists

- **The Data Science Handbook**
  [Buy on Amazon]
  Carl Shan, Henry Wang, William Chen, & Max Song, 2015
- **The Data Analytics Handbook**
  Brian Liou, Tristan Tao, & Declan Shener, 2015

Forming Data Science Teams

- **Data Driven: Creating a Data Culture**
  [Buy on Amazon]
  Hilary Mason & DJ Patil, 2015
- **Building Data Science Teams**
  [Buy on Amazon]
  DJ Patil, 2011
- **Understanding the Chief Data Officer**

Data Analysis

- **The Elements of Data Analytic Style**
  [Buy on Amazon]
  Jeff Leek, 2015

Distributed Computing Tools

- **Hadoop: The Definitive Guide**
  [Buy on Amazon]
  Tom White, 2011
- **Data-Intensive Text Processing with MapReduce**
  [Buy on Amazon]
  Jimmy Lin & Chris Dyer, 2010

Learning Languages
Python

- **Think Python: How to Think Like a Computer Scientist**
  Allen Downey, 2012
- **Python Programming**
  Wikibooks, 2015
- **Automate the Boring Stuff with Python: Practical Programming for Total Beginners**
  [Buy on Amazon]
  Al Sweigart, 2015
- **Learn Python the Hard Way**
  [Buy on Amazon]

Sơ lược cộng đồng KPDL

- 1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)
- 1991-1994 Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)
- 1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)
  - Journal of Data Mining and Knowledge Discovery (1997)
- 1998 ACM SIGKDD, SIGKDD'1999-2001 conferences, and SIGKDD Explorations
- More conferences on data mining
  - PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), v.v.

KDD 2017

KDD2017

Registration Program KDD Cup 2017 Announcements Sponsorship About Contact

KEYNOTE SPEAKERS

Top Data Scientists Share Their Knowledge to Advance the Application of Data Science, benefiting all aspects of society

The annual SIGKDD conferences have a strong reputation for delivering unparalleled peer learning, networking and idea sharing opportunities within data science, data mining, knowledge discovery, large-scale data analytics and big data. The event's main focus is to connect the world's best data scientists with one another in order to discuss, address and advance the application of data science to benefit all aspects of society. KDD 2017 is looking to be an amazing year.

What's Fair?

Cynthia Dwork
Professor / Distinguished Scientist
Harvard University / Microsoft Research

More Information

What's Fair?

Three Principles of Data Science: Predictability, Stability, and Computability

Three Principles of Data Science: Predictability, Stability, and Computability

Bin Yu
Professor
University of California at Berkeley

More Information

The Future of Data Integration

The Future of Data Integration

Renée J. Miller
Professor
University of Toronto

More Information

UofT:DB Group

Top Data Scientists Share Their Knowledge to Advance the Application of Data Science, benefiting all aspects of society

Poster Session Maps
- Monday
- Wednesday

Important Dates
- Camera Ready Deadline
  June 9, 2017
- Startup Grant Deadline
  June 16, 2017
- Student Grants Deadline
  June 17, 2017
- Promotional Video Deadline
  June 18, 2017
- Tutorials
  August 13, 2017
- Workshops
  August 14, 2017
- Main Conference
  August 15 - 17, 2017

Program
- Keynote Speakers
- Applied Data Science Invited Panels
- Applied Data Science Invited Talks
- Plenary Panel
- Conventional Tutorials
- Hands-On Tutorials
- Meet the Editors Panel
- Workshops
- Accepted Papers
- Networking With Experts

REGISTER NOW!

http://www.kdd.org/kdd2017/keynotes

KDD 2017

Cynthia Dwork

Professor / Distinguished Scientist

Harvard University / Microsoft Research

Title: What's Fair?

Data, algorithms, and systems have biases embedded within them reflecting designers' explicit and implicit choices, historical biases, and societal priorities. They form, literally and inexorably, a codification of values. "Unfairness" of algorithms – for tasks ranging from advertising to recidivism prediction – has attracted considerable attention in the popular press. The talk will discuss the nascent mathematically rigorous study of fairness in classification and scoring.

Speaker Bio

Cynthia Dwork, Distinguished Scientist at Microsoft Research, is renowned for placing privacy-preserving data analysis on a mathematically rigorous foundation. A cornerstone of this work is differential privacy, a strong privacy guarantee frequently permitting highly accurate data analysis. Dwork has also made seminal contributions in cryptography and distributed computing. She is a member of the US National Academy of Sciences and the US National Academy of Engineering, and is a Fellow of the American Academy of Arts and Sciences and the American Philosophical Society. Beginning January, 2017, Dwork will be the Gordon McKay Professor of Computer Science at the Harvard Paulson School of Engineering, the Radcliffe Alumnae Professor at the Radcliffe Institute for Advanced Study, and Professor by Affiliation at Harvard Law School.

http://www.kdd.org/kdd2017/keynotes/view/whats-fair

KDD 2017

KDD 2017 Keynote Talk

KDD'17, August 13–17, 2017, Halifax, NS, Canada

Three Principles of Data Science: Predictability, Stability, and Computability

Bin Yu
Departments of Statistics and EECS
University of California, Berkeley
binyu@berkeley.edu

ABSTRACT

In this talk, I'd like to discuss the intertwining importance and connections of three principles of data science in the title in data-driven decisions.

Making prediction as its central task and embracing computation as its core, machine learning has enabled wide-ranging data-driven successes. Prediction is a useful way to check with reality. Good prediction implicitly assumes stability between past and future. Stability (relative to data and model perturbations) is also a minimum requirement for interpretability and reproducibility of data driven results (cf. Yu, 2013). It is closely related to uncertainty assessment. Obviously, both prediction and stability principles cannot be employed without feasible computational algorithms, hence the importance of computability.

The three principles will be demonstrated in the context of two neuroscience projects and through analytical connections. In particular, the first project adds stability to predictive modeling used for reconstruction of movies from fMRI brain signlas for interpretable models. The second project use predictive transfer learning that combines AlexNet, GoogleNet and VGG with single V4 neuron data for state-of-the-art prediction performance. Our results lend support, to a certain extent, to the resemblance of these CNNs to brain and at the same time provide stable pattern interpretations of neurons in the difficult primate visual cortex V4.

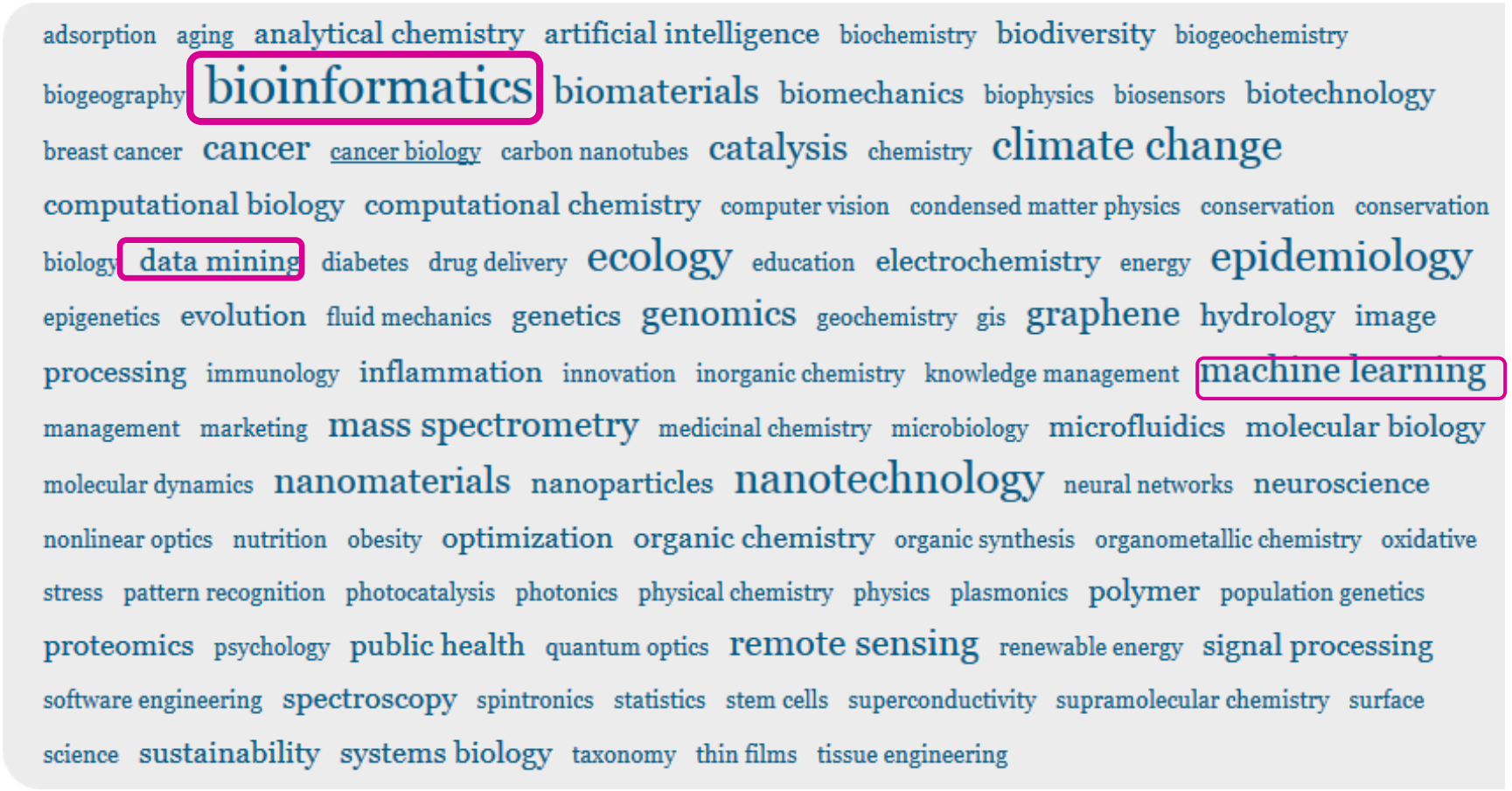
B. Yu (2013). Stability. Bernoulli, 19 (4), 1484-1500. (Invited paper for the Special Issue commemorating the 300th anniversary of the publication of Jakob Bernoullis Ars Conjectandi in 1712).

Bin Yu. Three Principles of Data Science: Predictability, Stability and Computability. KDD 2017: 5

KPDL: tốp từ nghiên cứu khóa hàng đầu

Top Keywords

Find researchers based on your area of interest.



http://www.researcherid.com/ 09/2016

Trang web KDD: Khảo sát

Vấn đề chính trong KPDL

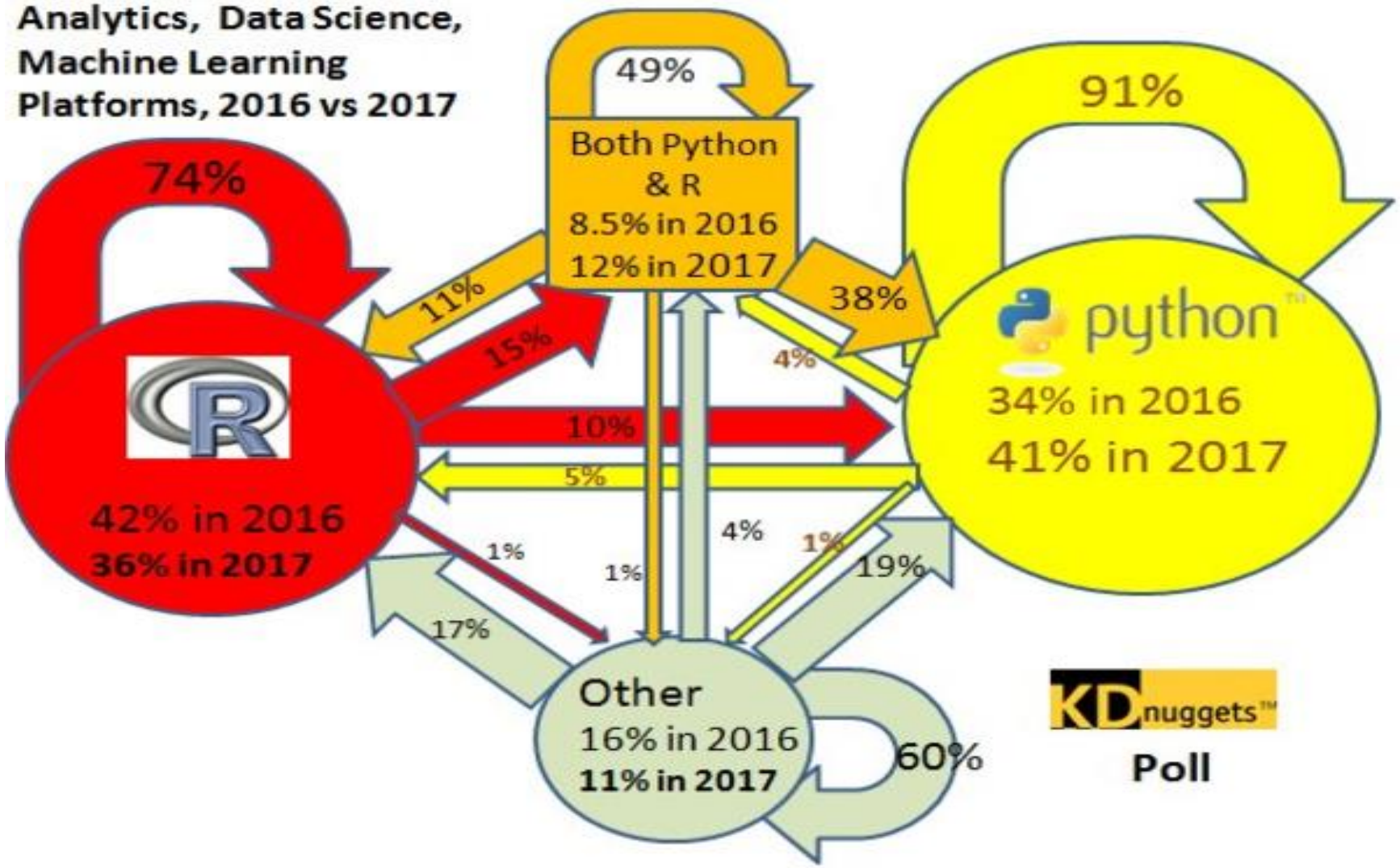
- Phương pháp luận khai phá
  - Khai phá các kiểu tri thức khác nhau từ dữ liệu hỗn tạp như sinh học, dòng, web…
  - Hiệu năng: Hiệu suất, tính hiệu quả, và tính mở rộng
  - Đánh giá mẫu: bài toán về tính hấp dẫn
  - Kết hợp tri thức miền: ontology
  - Xử lý dữ liệu nhiễu và dữ liệu không đầy đủ
  - Tính song song, phân tán và phương pháp KP gia tăng
  - Kết hợp các tri thức được khám phá với tri thức hiện có: tổng hợp tri thức
- Tương tác người dùng
  - Ngôn ngữ hỏi KPDL và khai phá “ngẫu hứng”
  - Biểu diễn và trực quan kết quả KPDL
  - Khai thác tương tác tri thức ở các cấp độ trừu tượng
- Áp dụng và chỉ số xã hội
  - KPDL đặc tả miền ứng dụng và KPDL vô hình
  - Bảo đảm bí mật dữ liệu, toàn vẹn và tính riêng tư

Một số yêu cầu ban đầu

- Sơ bộ về một số yêu cầu để dự án KPDL thành công
  - Cần có kỳ vọng về một lợi ích đáng kể về kết quả KPDL
    - Hoặc trực tiếp nhận được "trái cây treo thấp" ("low-hanging fruit") dễ thu lượm (như Mô hình mở rộng khách hàng qua tiếp thị và bán hàng)
    - Hoặc gián tiếp tạo ra đòn bẩy cao khi tác động vào quá trình sống còn có ảnh hưởng sóng ngầm mạnh (Giảm các nợ khoản khó đòi từ 10% còn 9,8% có số tiền lớn).
  - Cần có một đội dự án thi hành các kỹ năng theo yêu cầu: chọn dữ liệu, tích hợp dữ liệu, phân tích mô hình hóa, lập và trình diễn báo cáo. Kết hợp tốt giữ người phân tích và người kinh doanh
  - Nắm bắt và duy trì các dòng thông tin tích lũy (chẳng hạn, mô hình kết quả từ một loạt chiến dịch tiếp thị)
  - Quá trình học qua nhiều chu kỳ, cần "chạy đua với thực tiễn" (mô hình mở rộng khách hàng ban đầu chưa phải đã tối ưu).
- Một tổng hợp về các bài học KPDL thành công, thất bại

  [NEM09] Robert Nisbet, John Elder, and Gary Miner (2009). Handbook of Statistical Analysis and Data Mining, *Elsevier*, 2009.

Ngôn ngữ lập trình nền tảng



http://www.kdnuggets.com/2017/08/python-overtakes-r-leader-analytics-data-science.html