



Khoa Công Nghệ Thông Tin Trường Đại Học Cần Thơ



Từ khám phá tri thức đến khai mỏ dữ liệu

Knowledge Discovery in Databases - Data Mining

Đỗ Thanh Nghị
dtngchi@cit.ctu.edu.vn

Cần Thơ
12-2016

Nội dung

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Nội dung

- **tại sao KDD & DM là cần thiết?**
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Sự bùng nổ dữ liệu

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ trong những năm 90, với sự phát triển mạnh của:

- công nghệ vi xử lý
- công nghệ lưu trữ
- công nghệ truyền thông
- ứng dụng công nghệ thông tin trong nhiều lĩnh vực

➤ dữ liệu tăng nhanh

➤ bùng nổ dữ liệu

➤ (Lyman et al., 2003),

<http://www.sims.berkeley.edu/research/projects/how-much-info/>

Một vài ví dụ

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ cơ sở dữ liệu khoa học thiên văn

- Europe's Very Long Baseline Interferometry (VLBI)
- 16 kính thiên văn
- mỗi kính thu 1 Gigabits/giây dữ liệu
- phân tích dữ liệu thu được của 25 ngày
- kho dữ liệu quá lớn, vài Terabytes ⁽¹⁾

(1): 1 Kb = 1000 bytes, 1 Mb = 1000² bytes, 1 Gb = 1000³ bytes, 1 Tb = 1000⁴ bytes,
1 Pb = 1000⁵ bytes, 1 Eb = 1000⁶ bytes, 1 Zb = 1000⁷ bytes, 1 Yb = 1000⁸ bytes

Một vài ví dụ

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ các cơ sở dữ liệu khoa học khác

- NSA: hàng triệu tài liệu văn bản nói về khủng bố
 - Merck: hàng triệu cấu trúc phân tử hóa học
 - El nino: vài trăm Gigabytes
- khối lượng dữ liệu khổng lồ cần phân tích

Một vài ví dụ

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

-
- cơ sở dữ liệu truyền thông
 - AT&T: tiếp nhận 275 triệu cuộc gọi / ngày
 - cơ sở dữ liệu thương mại
 - lưu trữ thông tin về khách hàng
 - phục vụ cho kế hoạch đầu tư và phát triển
 - AT&T: 26 Terabytes
 - France Telecom: 30 Terabytes thông tin về khách hàng
 - Walmart: 20 triệu giao dịch / ngày

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Một vài ví dụ

■ dữ liệu world wide

- Google: tiếp nhận hơn 4 tỉ yêu cầu tìm kiếm / ngày, lưu trữ hàng trăm Terabytes dữ liệu
- Alexa internet archive: 500 Terabytes / 7 năm
- IBM WebFountain, 160 Terabytes / năm 2003
- Internet Archive, www.archive.org: 300 Terabytes

■ tổng hợp lại

- trong năm 2002: dữ liệu trên toàn cầu tăng 5 Exabytes ⁽¹⁾
- dữ liệu tăng 2 lần trong vòng 9 tháng

(1): 1 Kb = 1000 bytes, 1 Mb = 1000² bytes, 1 Gb = 1000³ bytes, 1 Tb = 1000⁴ bytes,
1 Pb = 1000⁵ bytes, 1 Eb = 1000⁶ bytes, 1 Zb = 1000⁷ bytes, 1 Yb = 1000⁸ bytes

KDD & DM là cần thiết

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ KDD & DM

- thực sự cần thiết để khai thác những tri thức tiềm ẩn
 - trong những kho dữ liệu lớn
- tạp chí về công nghệ của trường MIT số ra tháng 1-2 năm 2001

Data Mining được xem là một trong 10 công nghệ nổi bật nhất của thế kỷ 21

Nội dung

- tại sao KDD & DM là cần thiết?
- **những ứng dụng của KDD & DM**
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Lĩnh vực ứng dụng

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ khoa học & công nghệ

- thiên văn, sinh học, etc.

■ thương mại

- quảng cáo, marketing, đầu tư sản xuất, phân tích rủi ro trong kinh doanh, etc.

■ Web

- moteur tìm kiếm, phân loại bản tin, Web log, etc.

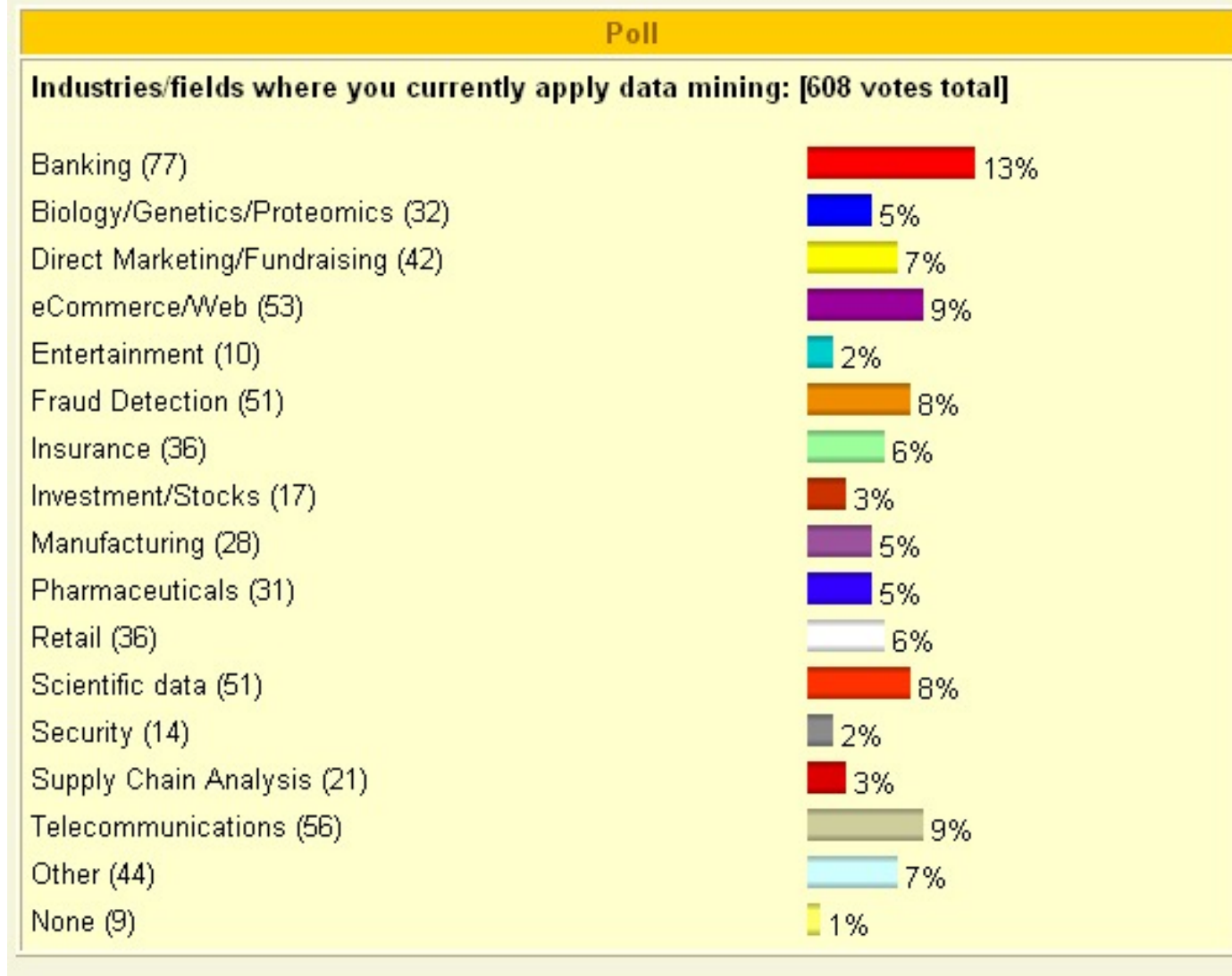
■ an ninh quốc phòng

- chống khủng bố, chống gian lận, etc.

2002

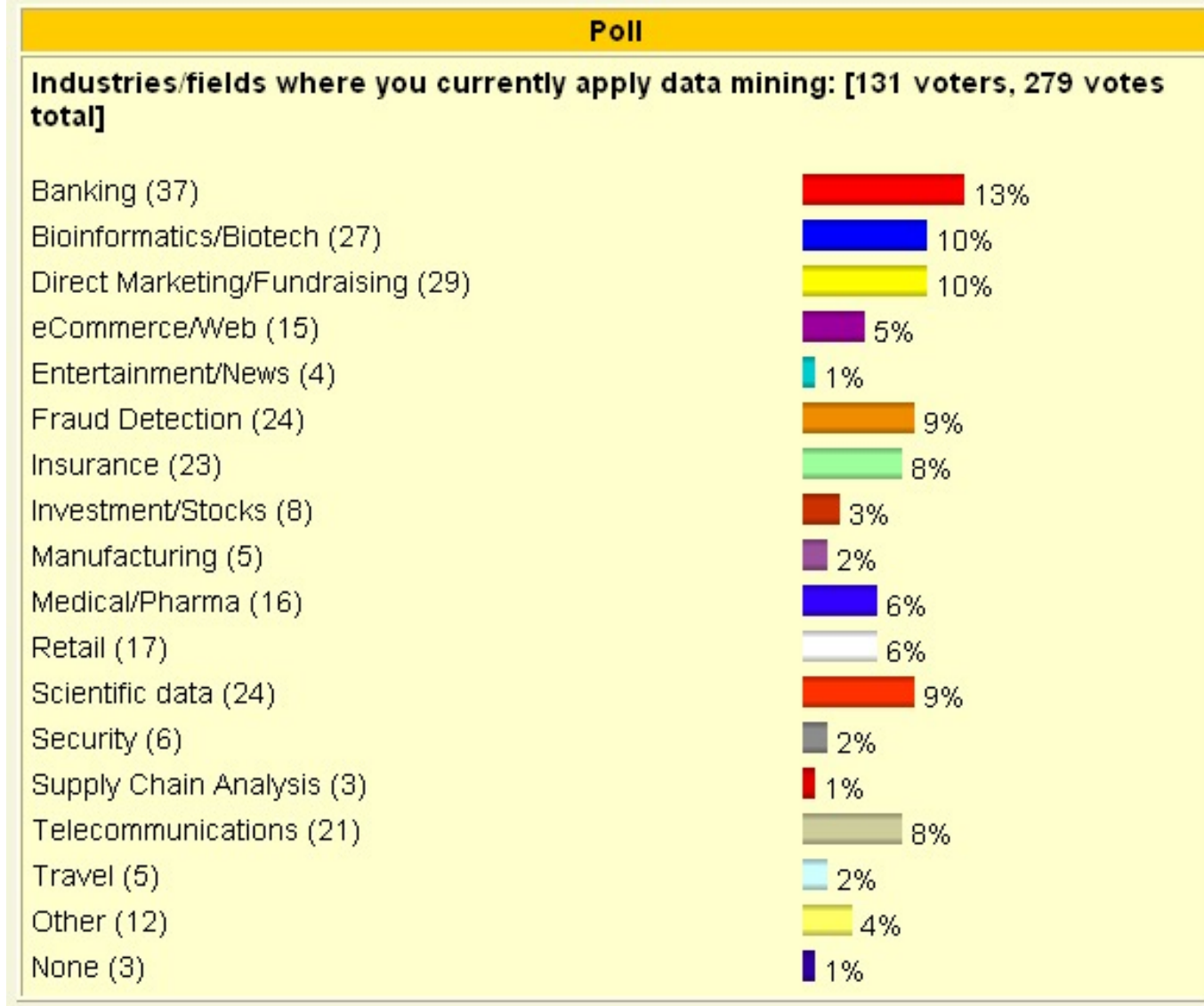
- tại sao KDD & DM là cần thiết?
- **những ứng dụng của KDD & DM**
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : **Polls** : Where do you currently apply data mining (Jun 2002)



2003

KDnuggets : **Polls** : Current data mining applications/industries

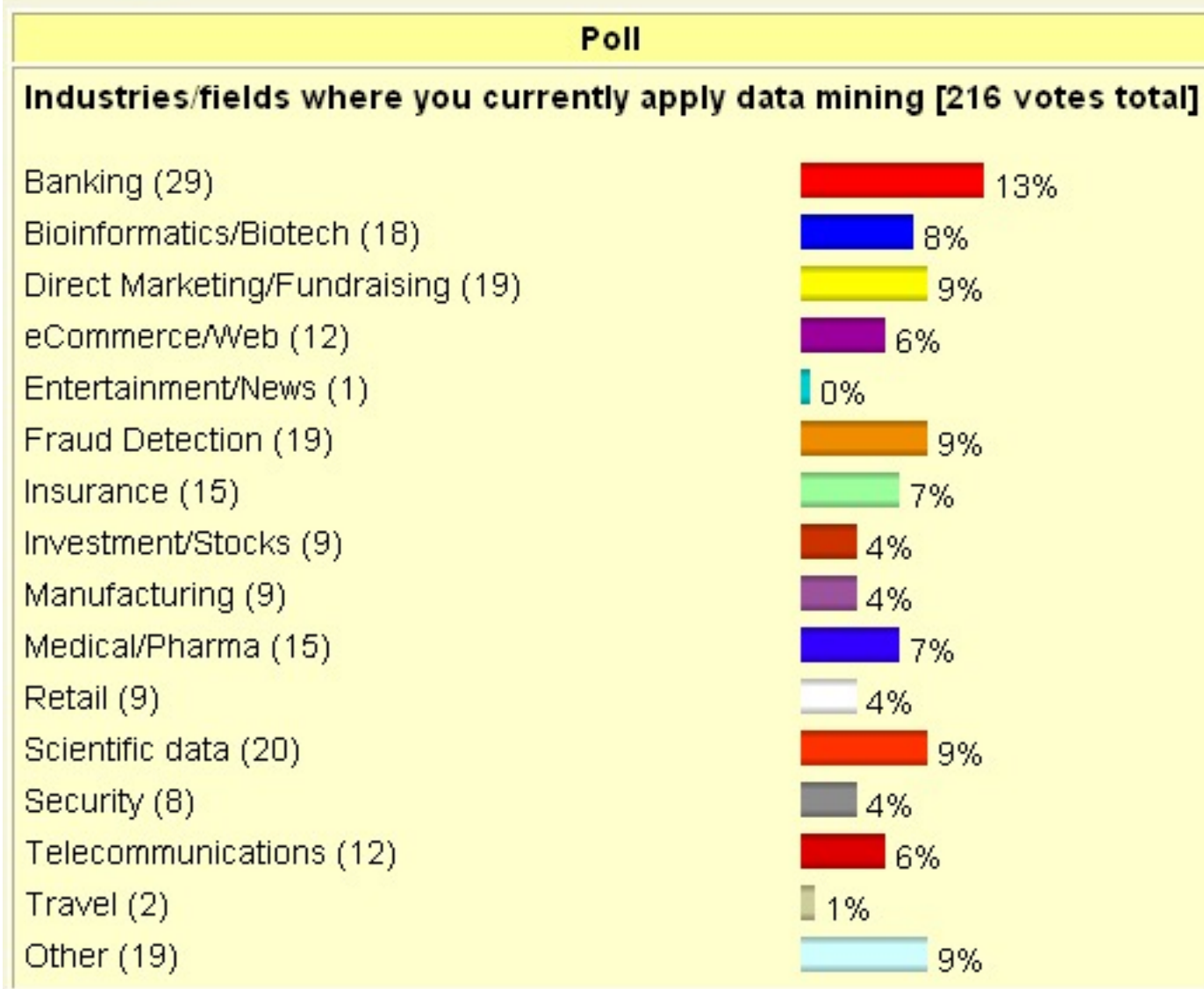


- tại sao KDD & DM là cần thiết?
- **những ứng dụng của KDD & DM**
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

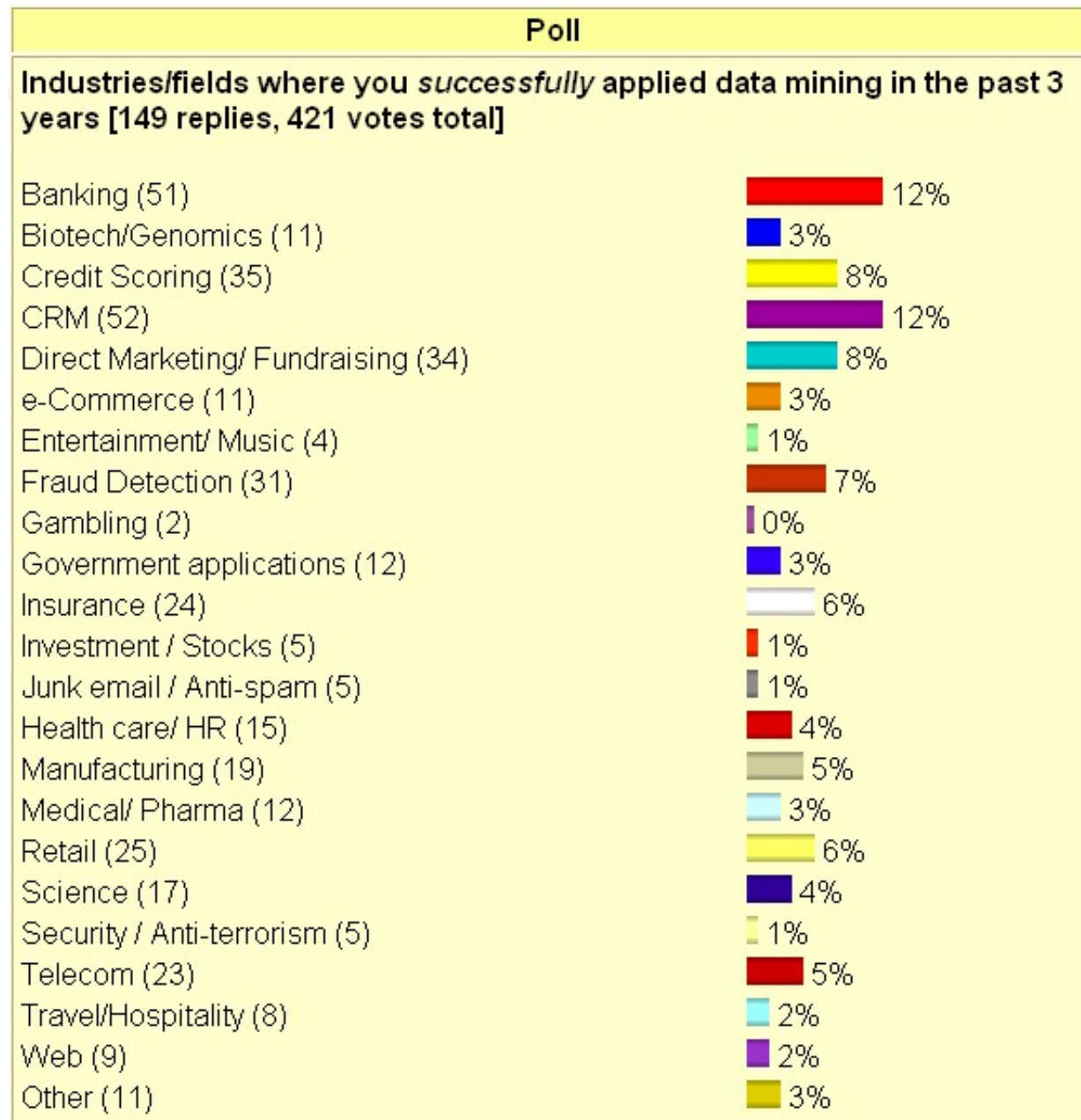
2004

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

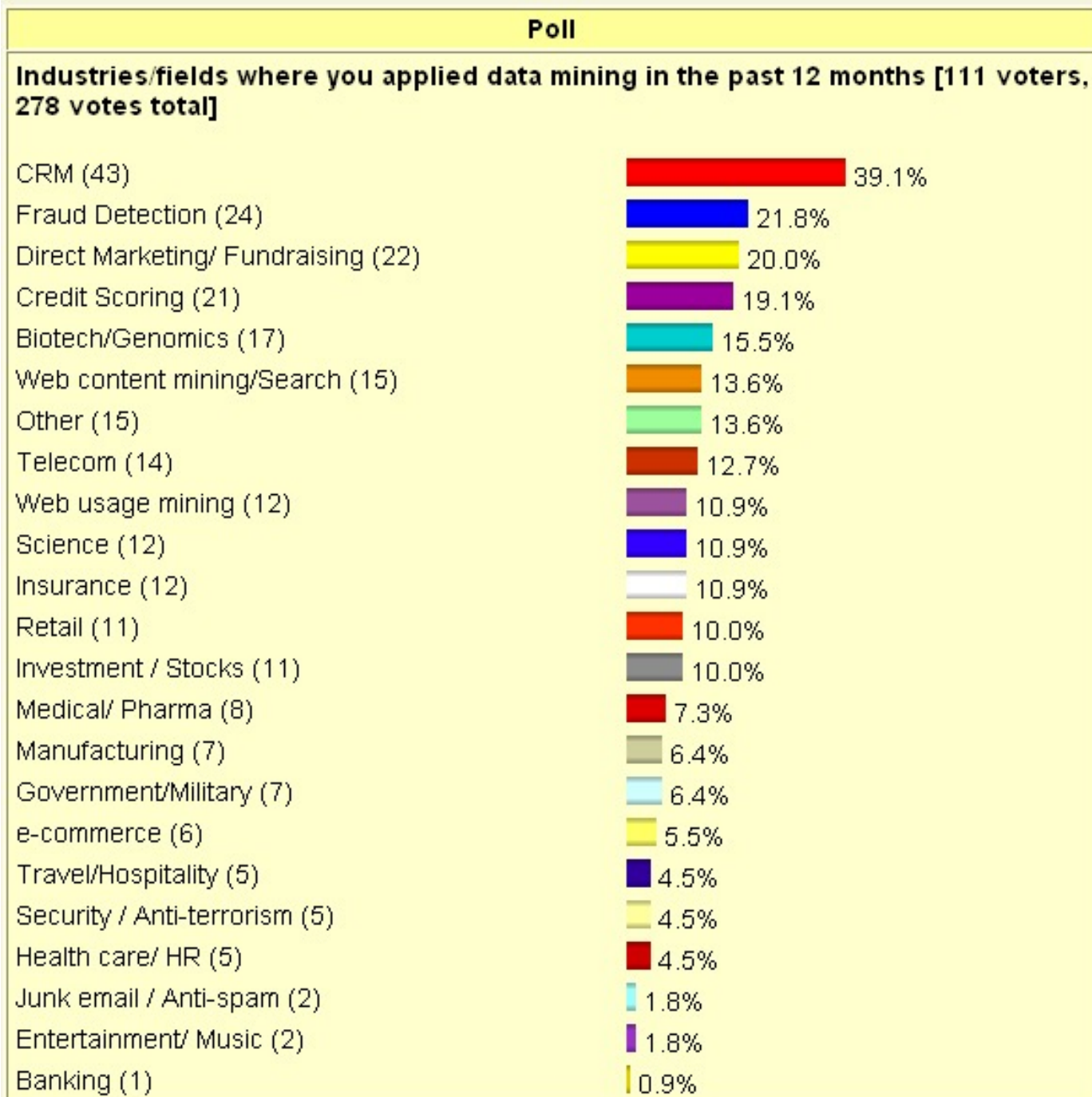
KDnuggets : **Polls** : Data Mining Applications (Aug 2004)



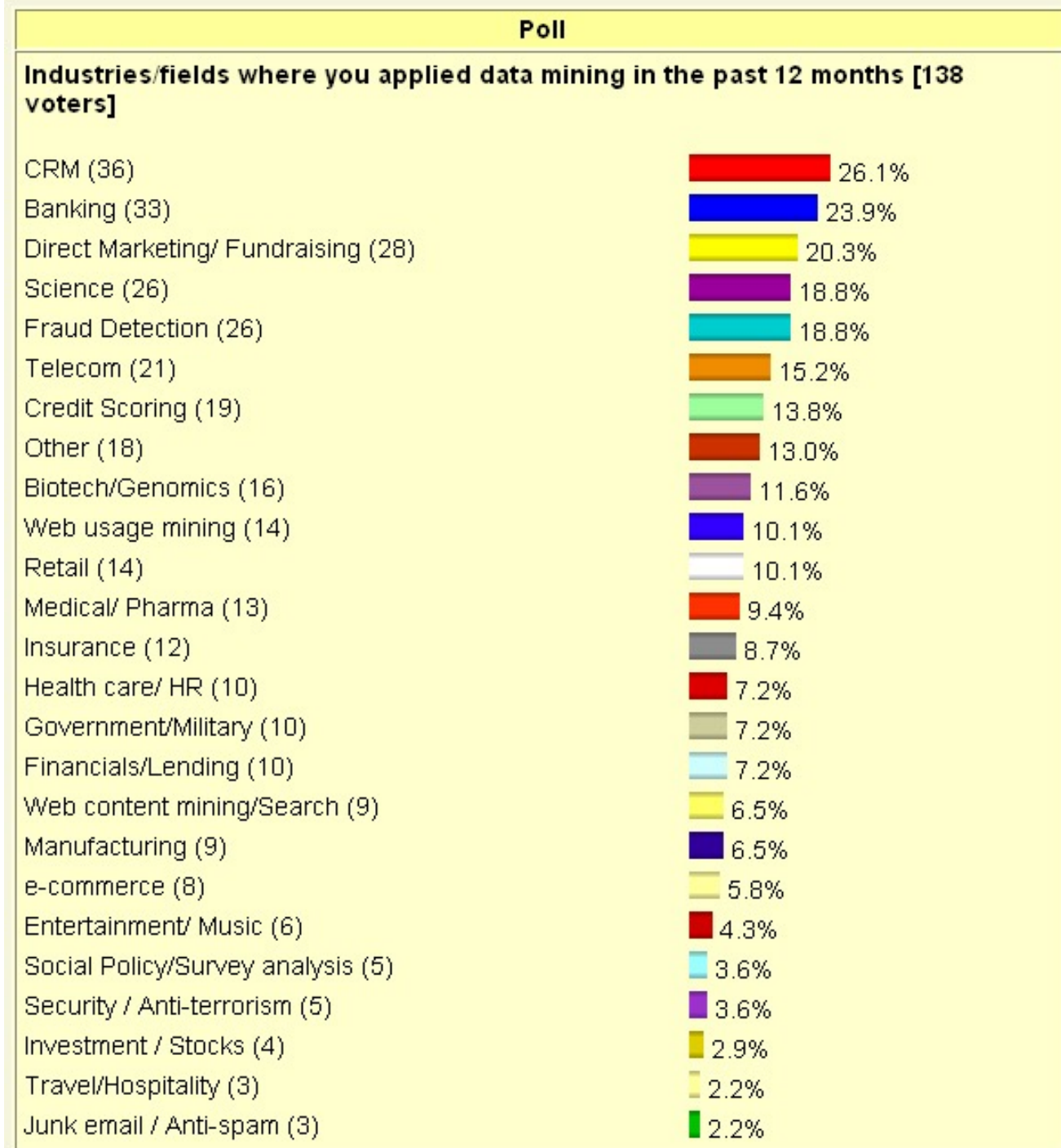
2005



- tại sao KDD & DM là cần thiết?
- **những ứng dụng của KDD & DM**
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



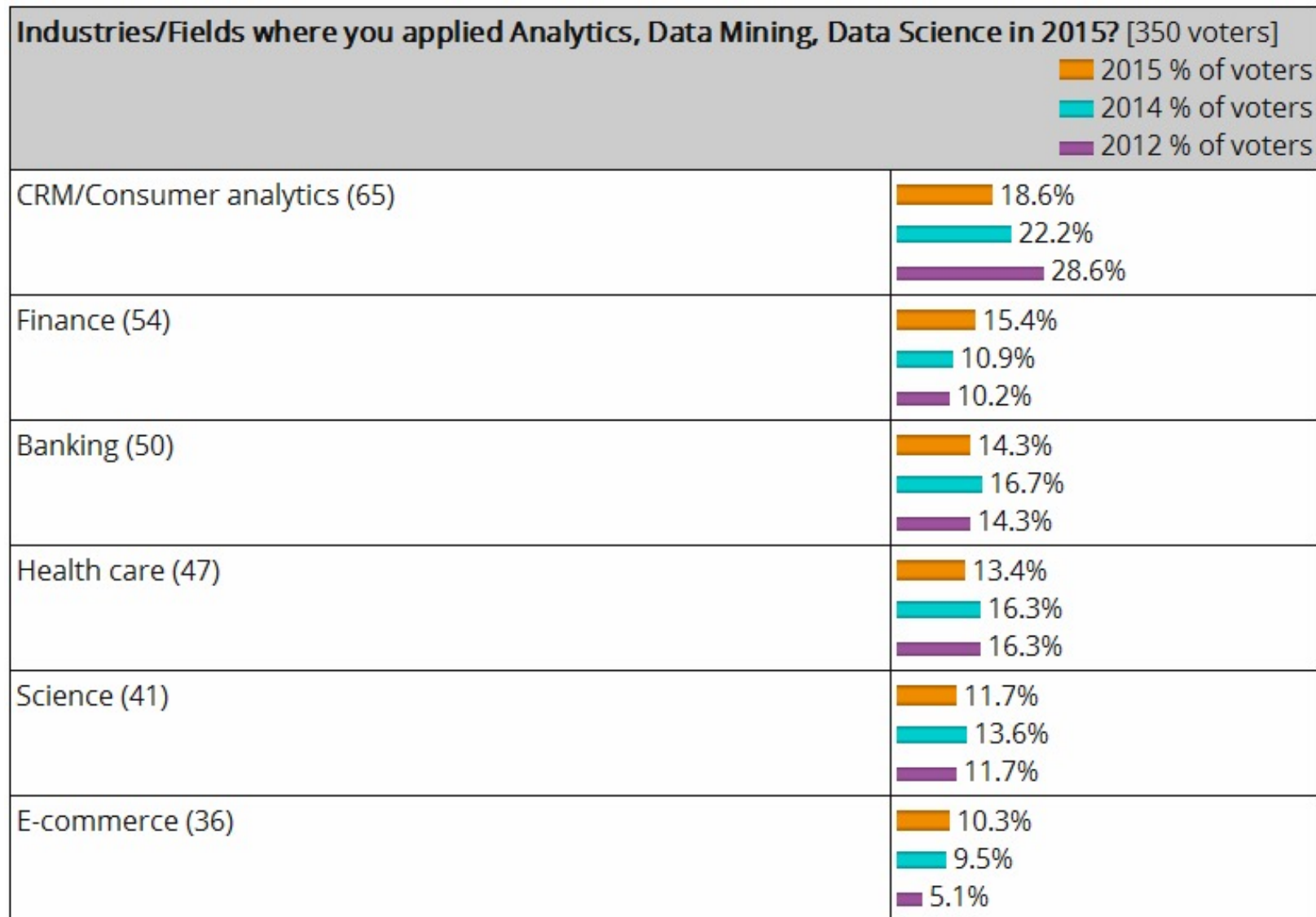
- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



- tại sao KDD & DM là cần thiết?
- **những ứng dụng của KDD & DM**
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Các năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



Các năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Industries/Fields where you applied Analytics, Data Mining, Data Science in 2015? [350 voters]	
	■ 2015 % of voters ■ 2014 % of voters ■ 2012 % of voters
Social Media / Social Networks (36)	■ 10.3% ■ 8.6% ■ 12.2%
Fraud Detection (35)	■ 10% ■ 13.6% ■ 12.8%
Education (35)	■ 10% ■ 7.7% ■ 14.3%
Retail (32)	■ 9.1% ■ 13.6% ■ 14.8%
Other (31)	■ 8.9% ■ 13.6% ■ 10.2%
Advertising (31)	■ 8.9% ■ 10.4% ■ 13.3%

Các năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Industries/Fields where you applied Analytics, Data Mining, Data Science in 2015? [350 voters]	
	■ 2015 % of voters ■ 2014 % of voters ■ 2012 % of voters
Oil / Gas / Energy (31)	■ 8.9% ■ 9.5% na
Telecom / Cable (27)	■ 7.7% ■ 9% ■ 6.6%
Insurance (26)	■ 7.4% ■ 8.6% ■ 7.7%
Credit Scoring (25)	■ 7.1% ■ 8.1% ■ 7.1%
Government/Military (25)	■ 7.1% ■ 6.3% ■ 5.1%
Manufacturing (24)	■ 6.9% ■ 9% ■ 7.1%

Các năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Industries/Fields where you applied Analytics, Data Mining, Data Science in 2015? [350 voters]	
	■ 2015 % of voters ■ 2014 % of voters ■ 2012 % of voters
IT / Network Infrastructure (23)	■ 6.6% na na
HR/workforce analytics (22)	■ 6.3% ■ 5.9% na
Medical/ Pharma (21)	■ 6% ■ 7.2% ■ 6.6%
Software (21)	■ 6% ■ 7.2% ■ 5.6%
Search / Web content mining (21)	■ 6% ■ 6.3% ■ 8.2%
Direct Marketing/ Fundraising (18)	■ 5.1% ■ 7.2% ■ 9.7%

Các năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển

Industries/Fields where you applied Analytics, Data Mining, Data Science in 2015? [350 voters]	
	■ 2015 % of voters ■ 2014 % of voters ■ 2012 % of voters
Web usage/Log mining (18)	■ 5.1% ■ 5.9% ■ 6.6%
Biotech/Genomics (17)	■ 4.9% ■ 6.8% ■ 7.7%
Automotive (15)	■ 4.3% ■ 5.9% na
Investment / Stocks (15)	■ 4.3% ■ 5% ■ 4.1%
Games (14)	■ 4% ■ 1.8% na
Mining (13)	■ 3.7% na na
Entertainment/ Music/ TV/Movies (11)	■ 3.1% ■ 1.8% ■ 4.6%

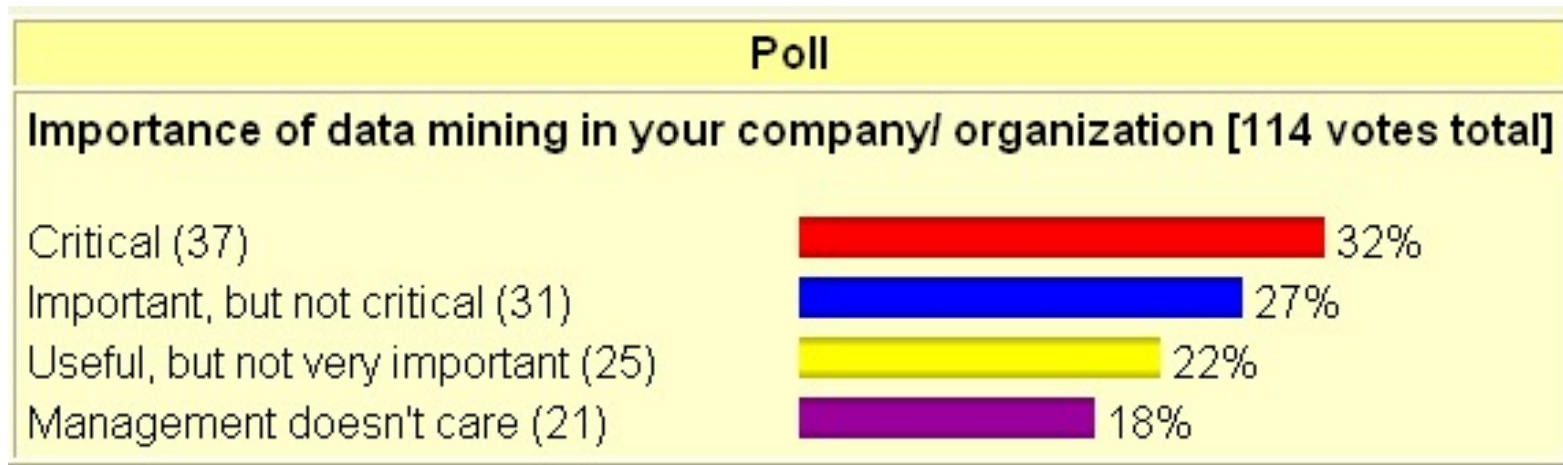
Các năm gần đây

- tại sao KDD & DM là cần thiết?
- **những ứng dụng của KDD & DM**
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển

Industries/Fields where you applied Analytics, Data Mining, Data Science in 2015? [350 voters]	
	■ 2015 % of voters ■ 2014 % of voters ■ 2012 % of voters
Agriculture (10)	■ 2.9% na na
Travel / Hospitality (9)	■ 2.6% ■ 3.2% ■ 3.1%
Security / Anti-terrorism (8)	■ 2.3% ■ 2.3% ■ 3.6%
Social Good/Non-profit (8)	■ 2.3% ■ 1.4% na
Social Policy/Survey analysis (6)	■ 1.7% ■ 1.8% ■ 1.0%
Mobile apps (5)	■ 1.4% ■ 2.3% na
Junk email / Anti-spam (1)	■ 0.3% ■ 1.8% ■ 0.5%

Data mining có quan trọng ?

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

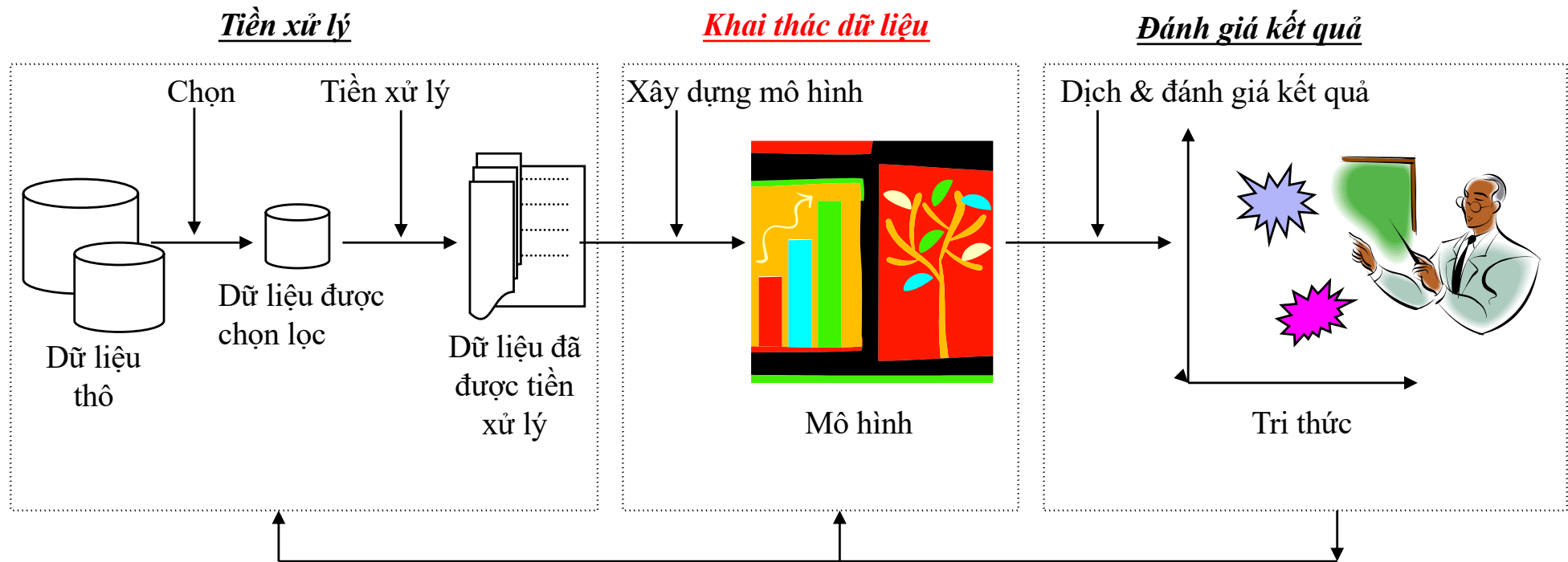


Nội dung

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

Quá trình KDD

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



■ quá trình KDD

- lặp
- khai mỏ dữ liệu (DM): cốt lõi

Tiền xử lý dữ liệu

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

-
- từ mục tiêu đề ra của ứng dụng
 - từ nguồn dữ liệu khác nhau
 - chọn dữ liệu cần thiết cho mục tiêu đề ra
 - mẫu tin, trường dữ liệu
 - biểu diễn dữ liệu, chuyển đổi kiểu sao cho phù hợp với giải thuật DM sẽ được áp dụng ở bước sau
 - làm sạch dữ liệu: khắc phục đối với trường dữ liệu rỗng, dư thừa, hoặc dữ liệu không hợp lệ
 - có thể tinh giảm dữ liệu hơn

Khai mở dữ liệu

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ kỹ thuật

- máy học
- trí tuệ nhân tạo
- nhận dạng
- phân tích thống kê
- hoặc bằng phương pháp trực quan: hiển thị
- xây dựng mô hình, tạo tri thức về dữ liệu
- kiểm định lại mô hình
- nếu chưa đạt thì phải xây dựng mô hình khác
- bước này rất khó và cần nhiều công sức

Đánh giá kết quả

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ kết quả

- kiểm định dựa vào mục tiêu ban đầu của ứng dụng
- nghĩa là chỉ có người sử dụng hoặc chuyên gia về lĩnh vực mới có khả năng đánh giá
- kết quả có đạt được cần dễ hiểu
- hiển thị, dịch kết quả
- người sử dụng hoặc chuyên gia
- có thể đánh giá và hiểu được kết quả sinh ra

Lĩnh vực nghiên cứu liên quan

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



Cơ sở nền tảng

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ thống kê

- dựa nhiều vào nền tảng lý thuyết
- tập trung vào kiểm định những giả thiết

■ máy học

- dựa nhiều vào heuristics
- tập trung cải tiến hiệu quả của giải thuật học

DM & KDD

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- **quá trình KDD**
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ DM & KDD

- tích hợp cơ sở lý thuyết & heuristic
- tập trung vào toàn bộ quá trình khám phá và khai thác dữ liệu để tìm ra tri thức
- phải hữu dụng cho người sử dụng
- bao gồm những kỹ thuật
- tiền xử lý dữ liệu: chọn lọc dữ liệu, xử lý dữ liệu sai, etc.
- quá trình xây dựng mô hình: DM
- trình bày, giải thích, dịch kết quả

Nội dung

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

Giải thuật khai mở dữ liệu

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ kỹ thuật

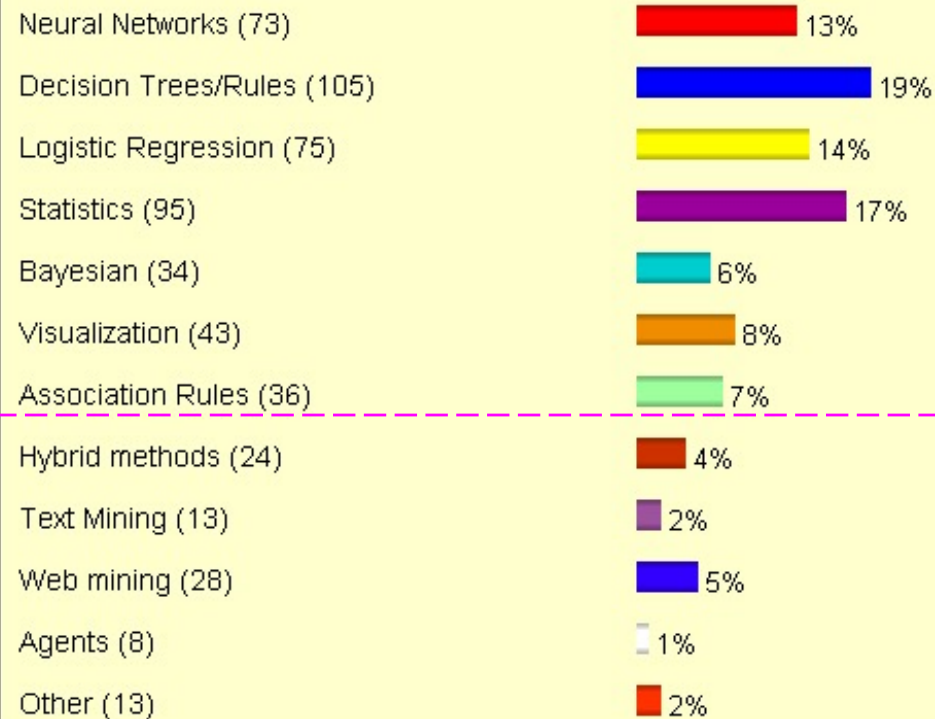
- phân lớp (classification): xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có nhãn (lớp)
- hồi quy (regression): xây dựng mô hình phân loại dựa trên dữ liệu tập học đã có nhãn (lớp) là giá trị liên tục
- gom cụm, nhóm (clustering): xây dựng mô hình gom cụm dữ liệu tập học (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau
- luật kết hợp (association rules): phát hiện mối liên quan giữa các biến của dữ liệu

Kỹ thuật DM (2001)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

[KDnuggets](#) : [Polls](#) : Data Mining Techniques (Aug 2001)

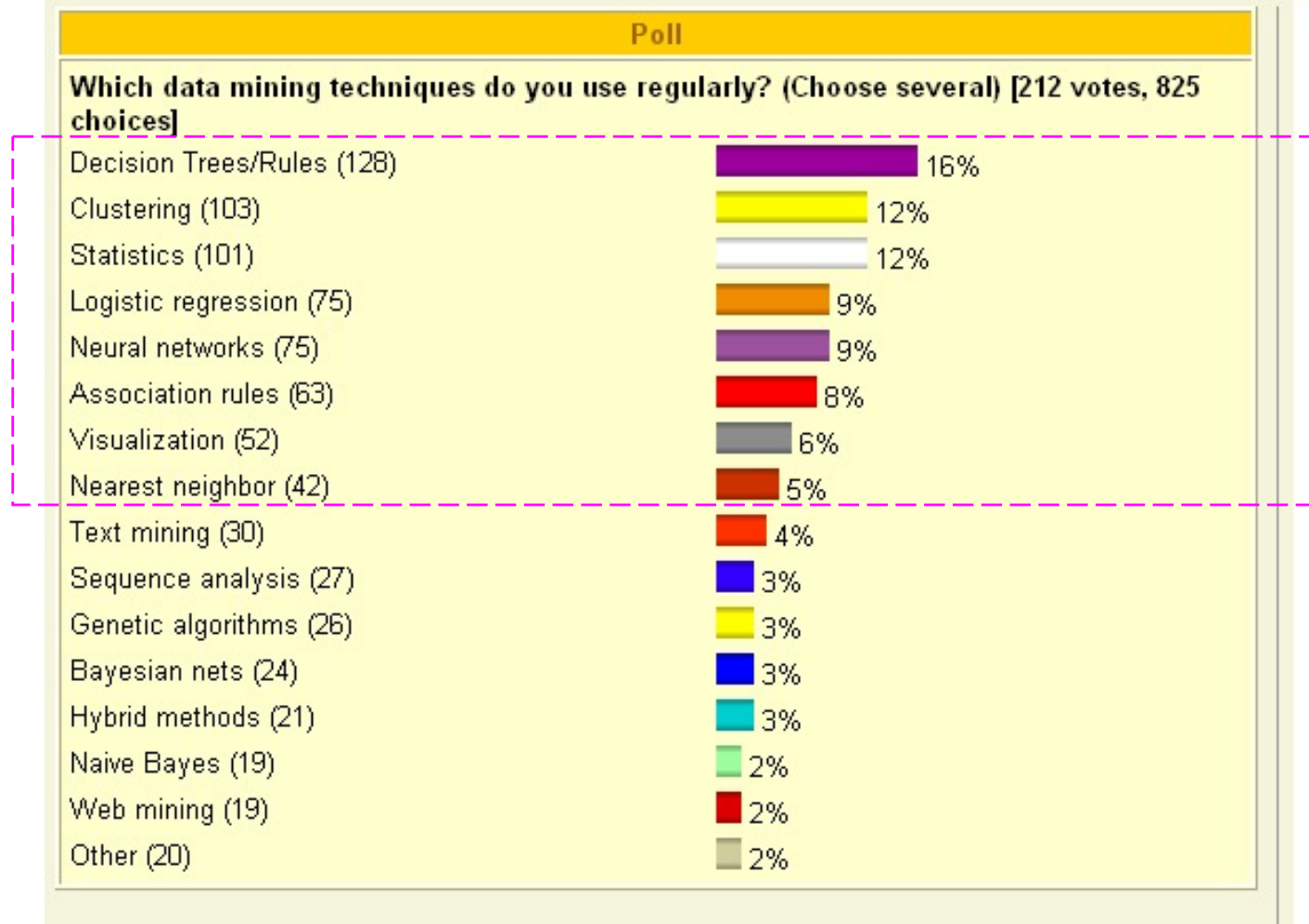
Which data mining techniques do you use regularly? (Choose several) [547 votes total]



Kỹ thuật DM (10/2002)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : **Polls** : Which data mining techniques do you use regularly (Oct 2002)



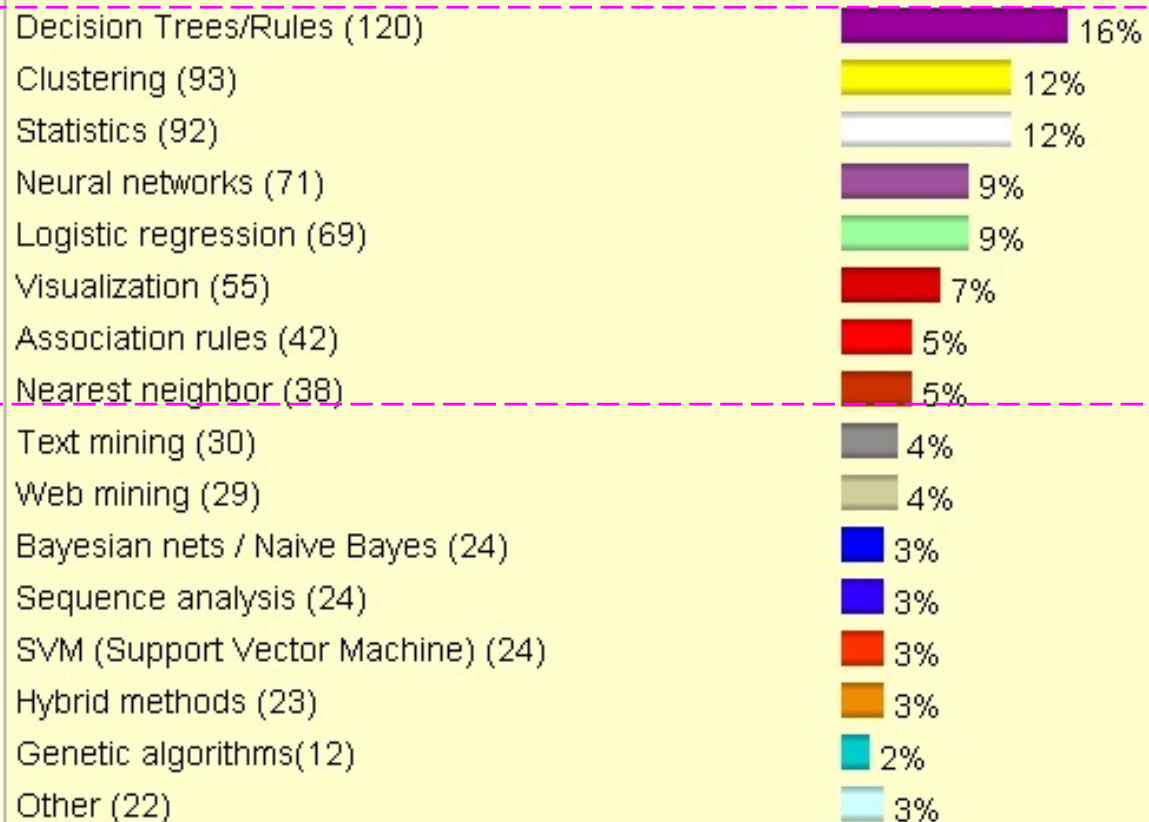
Kỹ thuật DM (11/2003)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : **Polls** : Data mining techniques (Nov 2003)

Poll

What data mining techniques you use regularly? [768 choices from 213 voters]

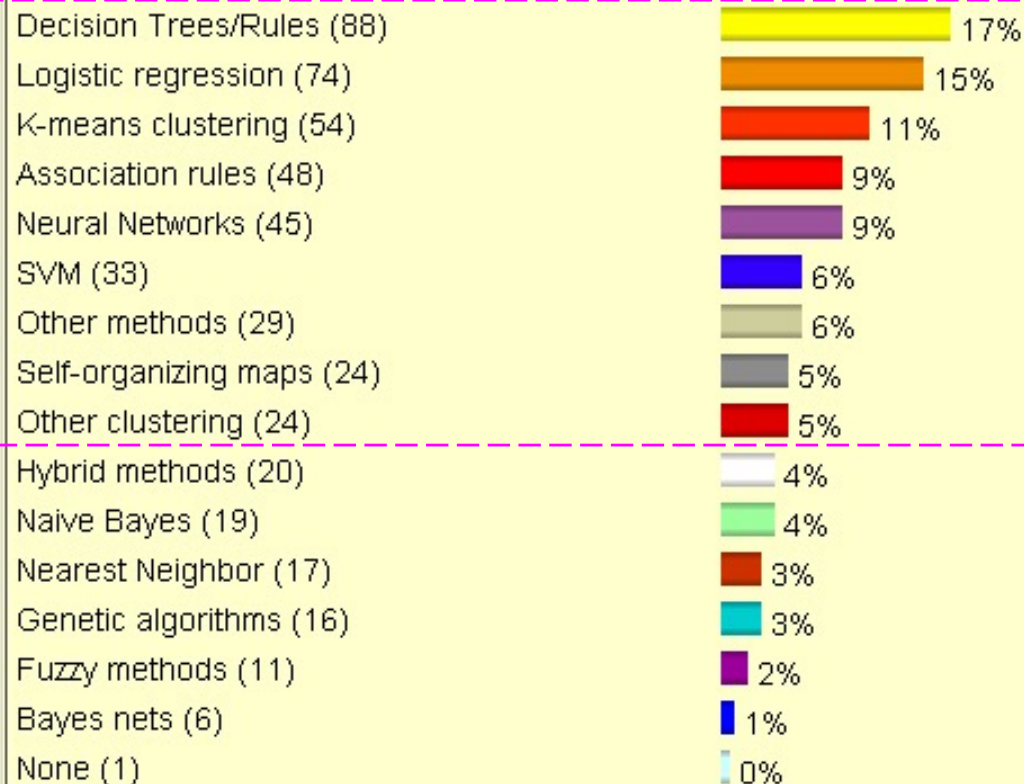


Kỹ thuật DM (2004)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : Polls : Deployed data mining techniques

Which data mining techniques you used in a successfully deployed application?
[173 voters, 509 votes total]



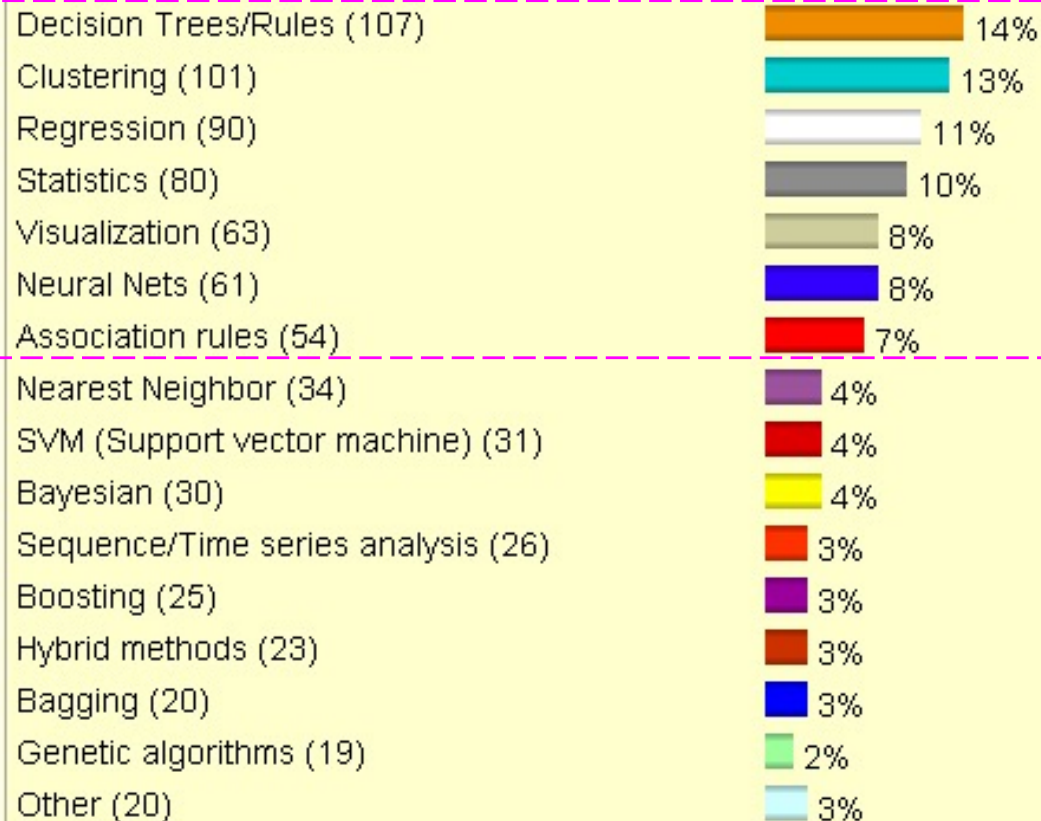
Kỹ thuật DM (02/2005)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : **Polls** : Data Mining Techniques (Feb 2005)

Poll

Data mining/analytic techniques you use frequently: [784 votes total]



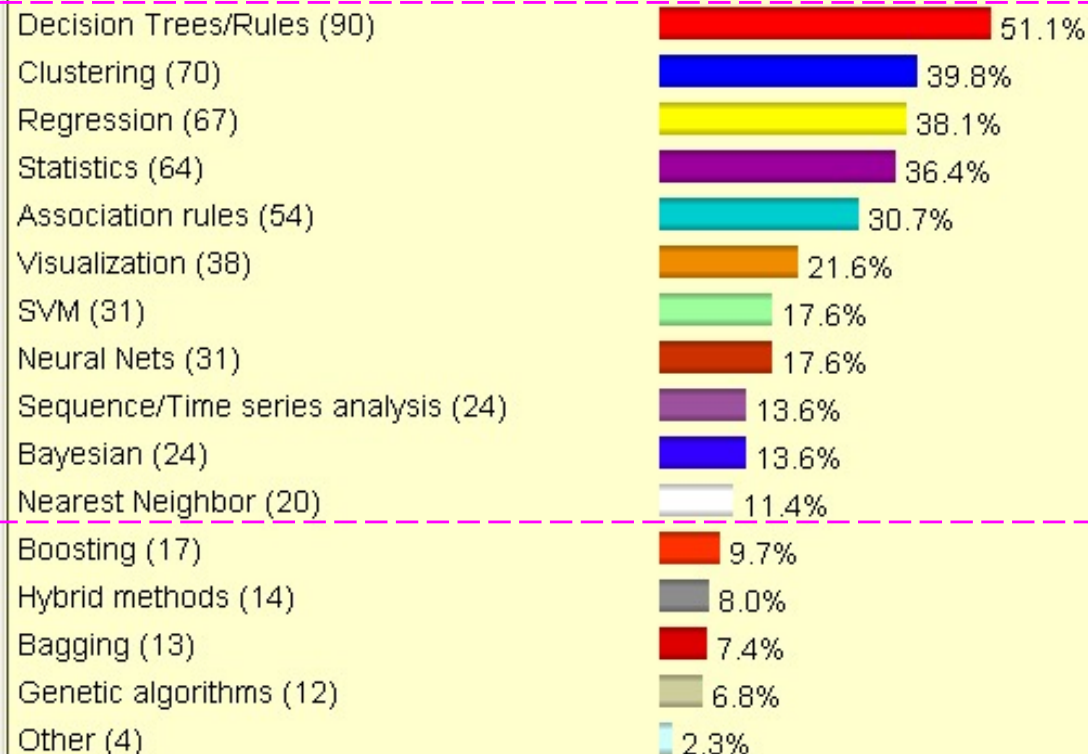
Kỹ thuật DM (2006)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : Polls : Data Mining Methods (Apr 2006)

Poll

Data mining/ analytic methods you used frequently in the last year: [176 voters]

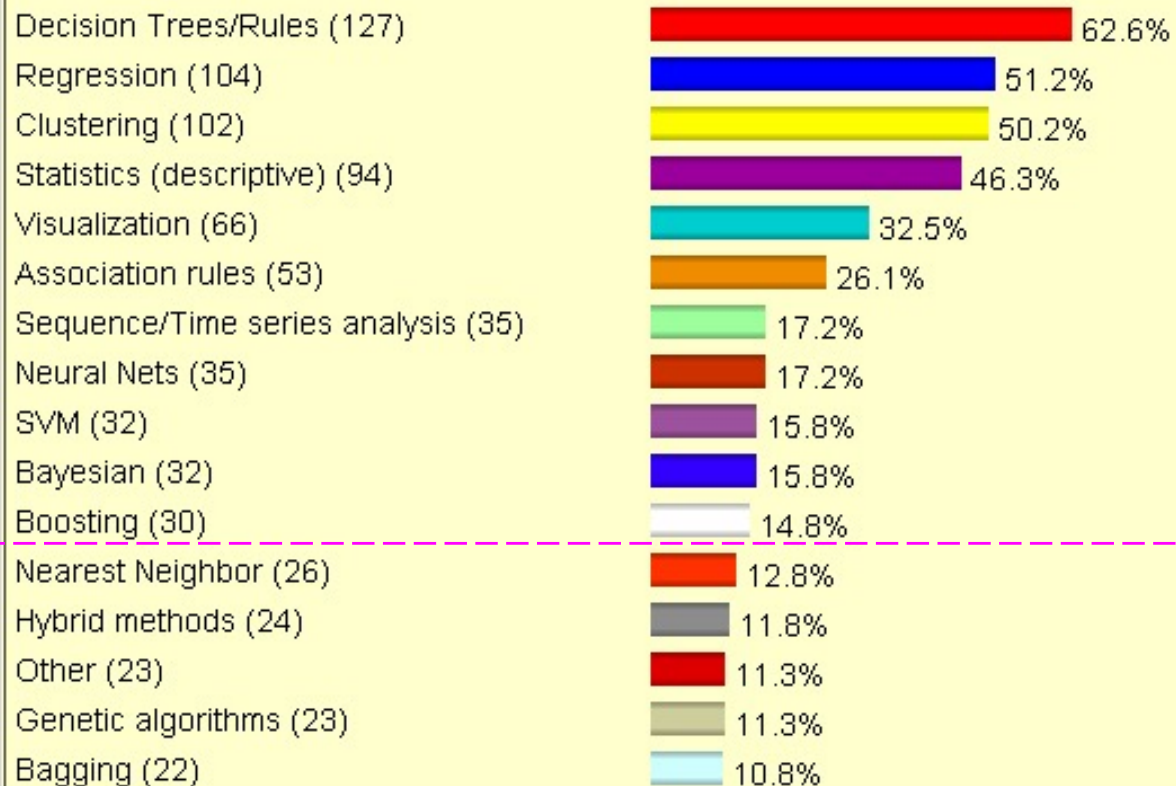


Kỹ thuật DM (2007)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

KDnuggets : Polls : Data Mining Methods (Mar 2007)

Data mining/analytic methods you used frequently in the past 12 months: [203 voters]



Top 10 DM algorithms (2015)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

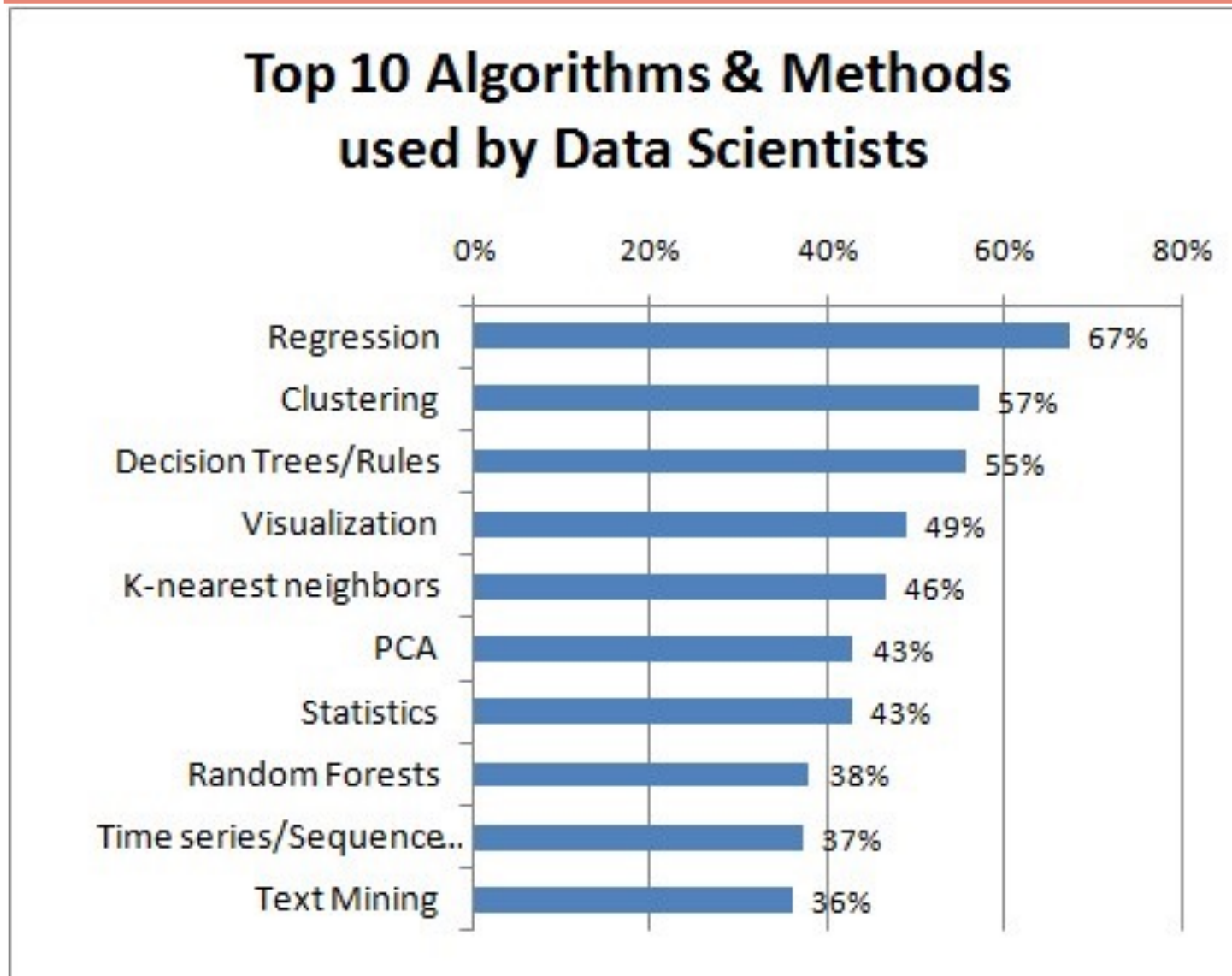


Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

Top 10 DM algorithms (2015)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

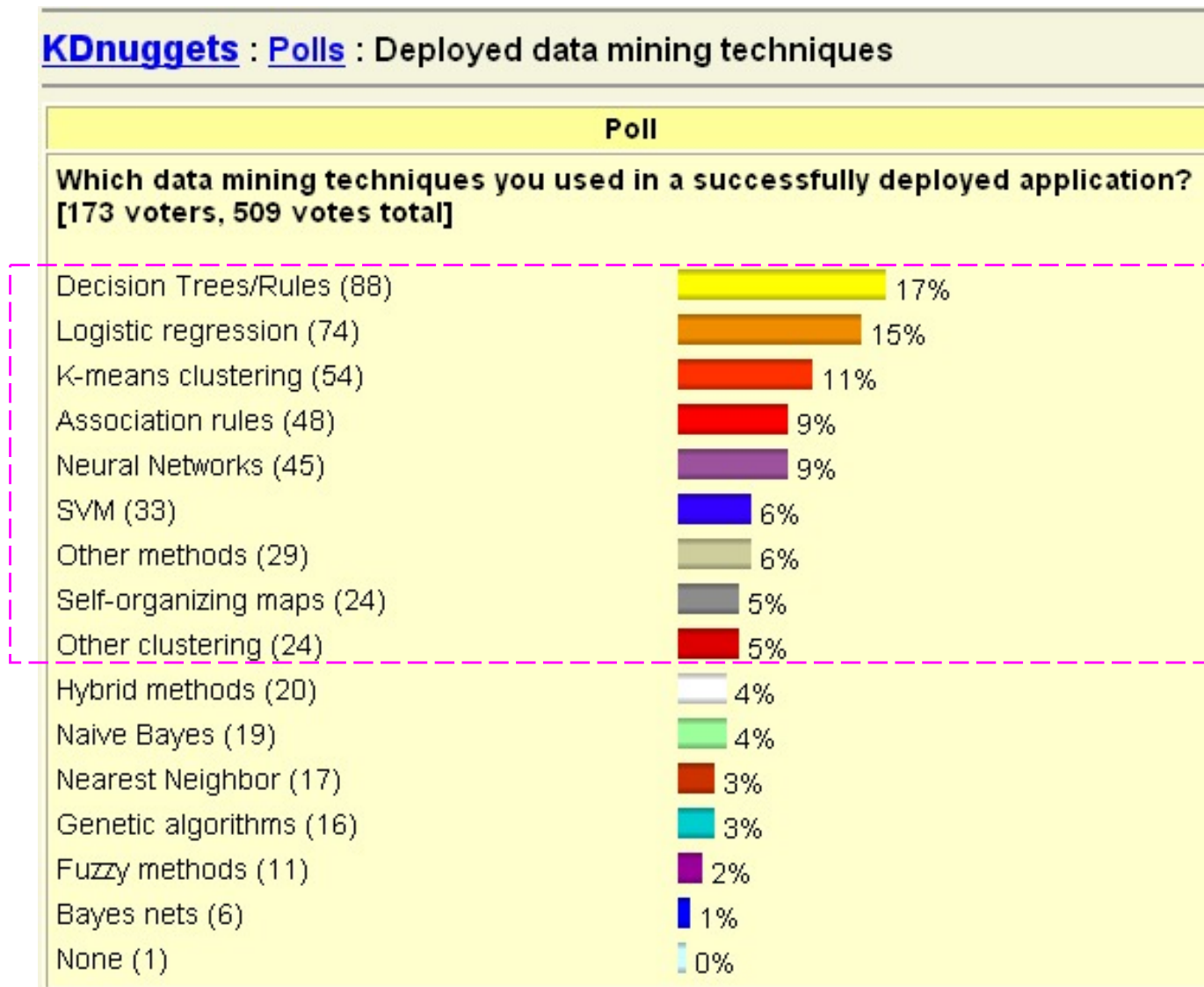


- tại sao KDD & DM là cần thiết?
 - những ứng dụng của KDD & DM
 - quá trình KDD
 - **giải thuật DM**
 - kết luận và hướng phát triển
 - tài liệu tham khảo
-

No free lunch theorem

Kỹ thuật DM thành công trong ứng dụng thực

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo



Phần mềm thường được sử dụng

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

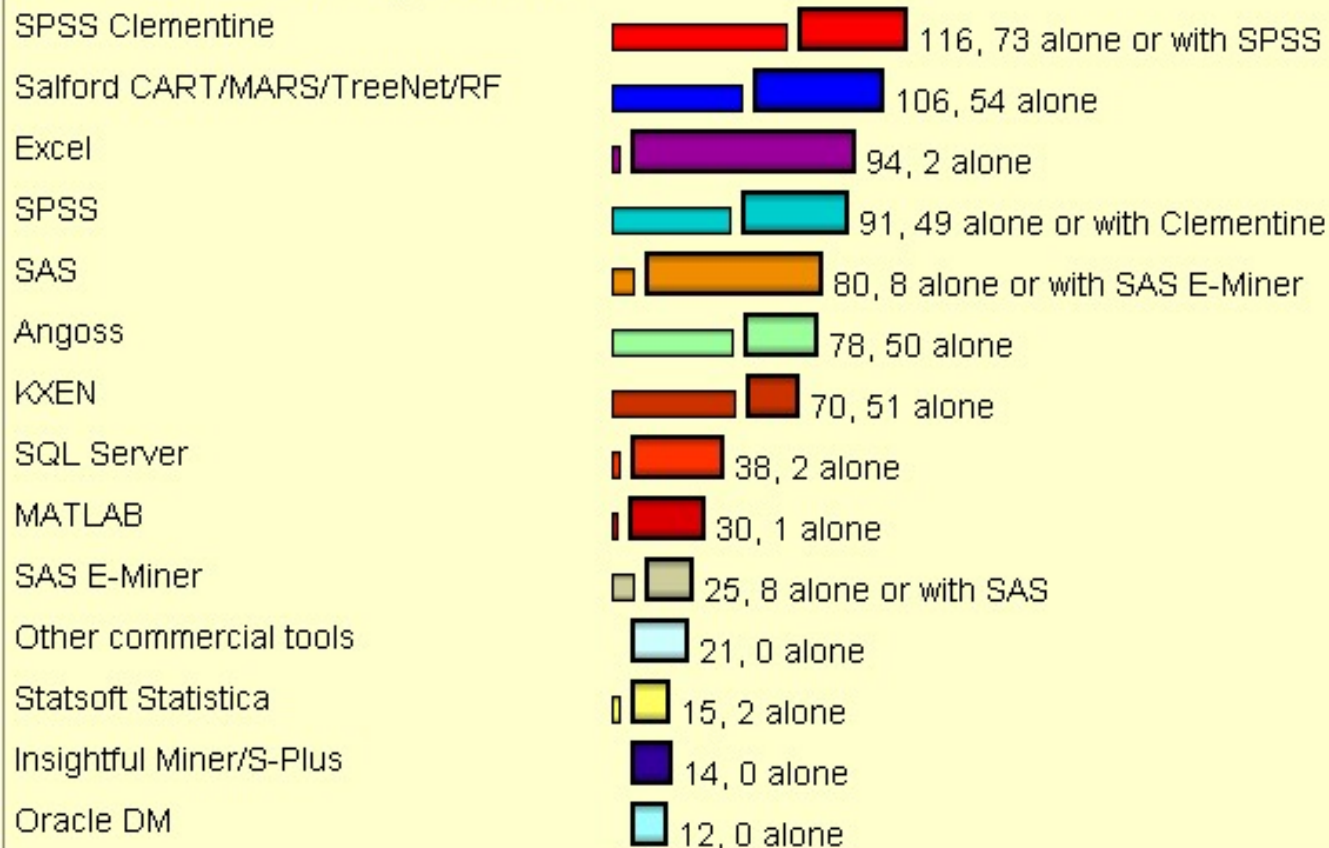
KDnuggets : [Polls](#) : Data Mining / Analytic Software Tools (May 2007)

Poll

Data Mining (Analytic) tools you used in 2007: [534 voters]

The first (narrow) bar corresponds to the number of votes where the tool was selected alone, and the second (wide) bar to the number votes where the tool was select as one among several; Tools are ordered in descending order of total number of votes

Commercial Data Mining Software



Phần mềm thường được sử dụng

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo

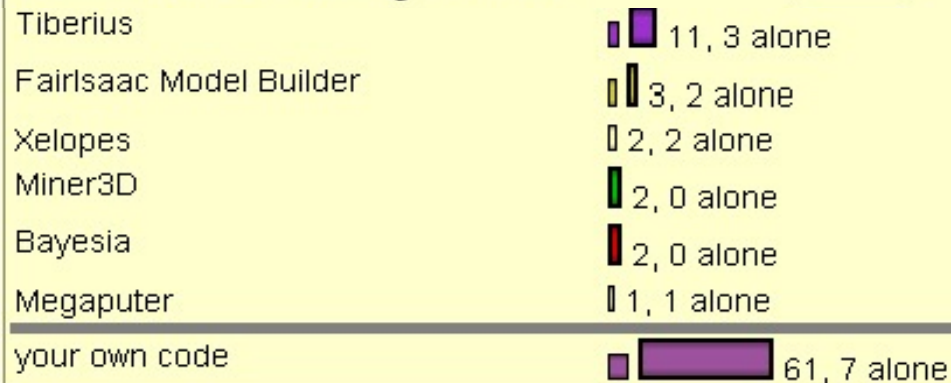
KDnuggets : [Polls](#) : Data Mining / Analytic Software Tools (May 2007)

Poll

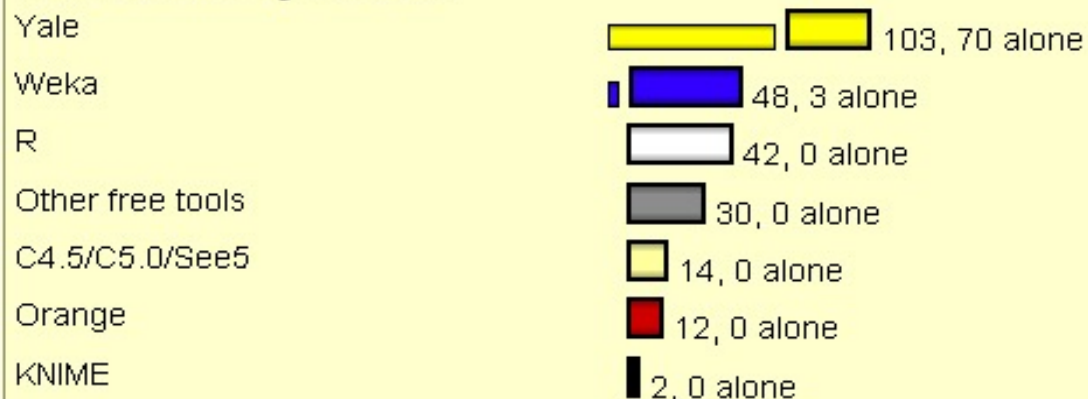
Data Mining (Analytic) tools you used in 2007: [534 voters]

The first (narrow) bar corresponds to the number of votes where the tool was selected alone, and the second (wide) bar to the number votes where the tool was select as one among several; Tools are ordered in descending order of total number of votes

Commercial Data Mining Software

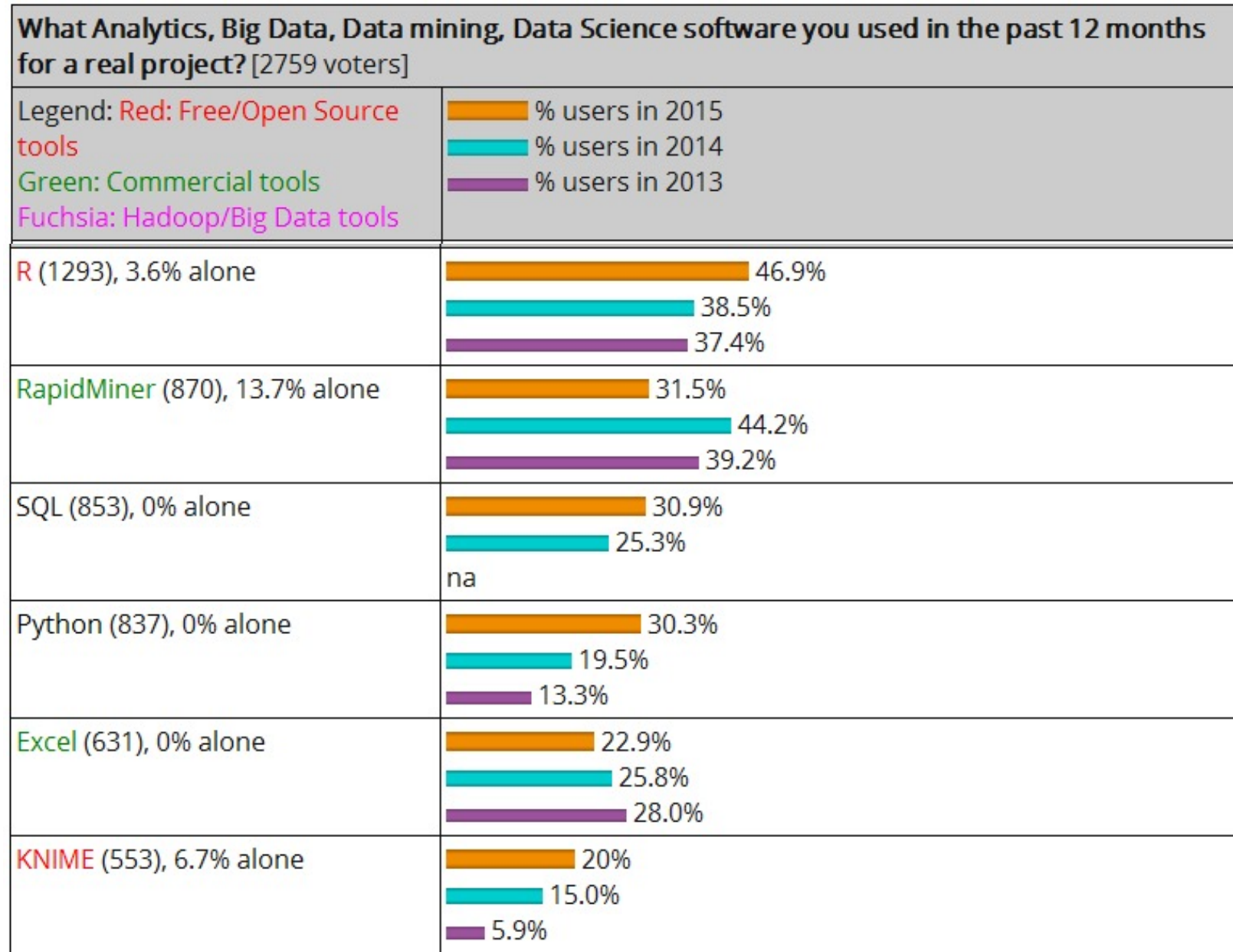


Free Data Mining Software



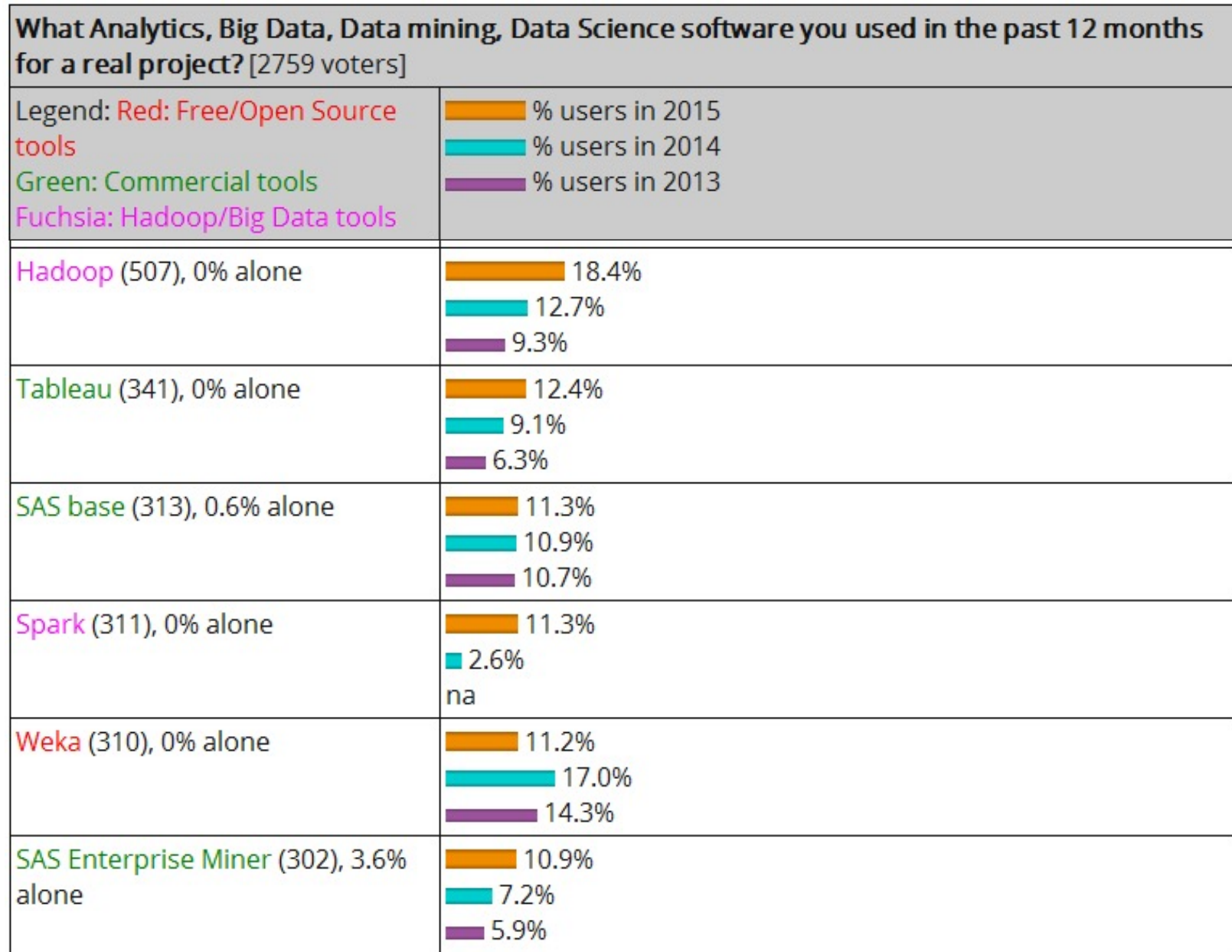
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



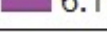
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo



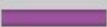
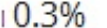
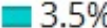
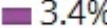
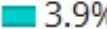
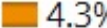
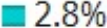
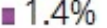
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
Microsoft SQL Server (268), 0% alone	 9.7%  10.5%  7.0%
MATLAB (243), 0% alone	 8.8%  8.4%  9.9%
scikit-learn (229), 0% alone	 8.3% na na
Unix shell/awk/gawk (221), 0% alone	 8.0%  5.8% na
IBM SPSS Statistics (213), 0% alone	 7.7%  7.7%  8.7%
IBM SPSS Modeler (197), 7.1% alone	 7.1%  5.7%  6.1%

Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
Alteryx (155), 39.4% alone	 5.6%  3.1%  0.3%
Pig (150), 0% alone	 5.4%  3.5% na
Other programming languages (140), 0% alone	 5.1%  3.0% na
Other free analytics/data mining tools (138), 0% alone	 5.0%  5.1%  3.4%
Other Hadoop/HDFS-based tools (125), 0% alone	 4.5%  3.9% na
TIBCO Spotfire (119), 11.8% alone	 4.3%  2.8%  1.4%

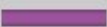
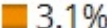
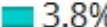
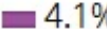
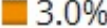
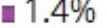
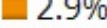
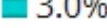
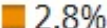
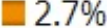
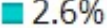
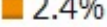
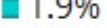
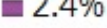
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013
Rattle (117), 0.9% alone	4.2% 4.9% 4.5%
QlikView (116), 0% alone	4.2% 3.0% 2.4%
Revolution Analytics (now part of Microsoft) (109), 0% alone	4.0% 9.1% 4.5%
Microsoft Azure ML (102), 1.0% alone	3.7% na na
Microsoft Power BI (98), 0% alone	3.6% na na
MLlib (91), 0% alone	3.3% 1.0% na

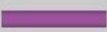
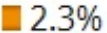
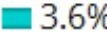
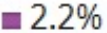
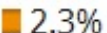
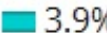
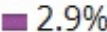
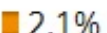
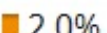
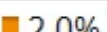
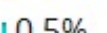
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
JMP (86), 0% alone	 3.1%  3.8%  4.1%
SAP (including former KXEN) (82), 26.8% alone	 3.0%  6.8%  1.4%
Perl (79), 0% alone	 2.9%  3.0% na
Mahout (76), 0% alone	 2.8%  2.5% na
Pentaho (74), 0% alone	 2.7%  2.6% na
Other paid analytics/data mining/data science software (66), 6.1% alone	 2.4%  1.9%  2.4%

Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
Salford SPM/CART/Random Forests/MARS/TreeNet (64), 43.8% alone	 2.3%  3.6%  2.2%
Gnu Octave (64), 0% alone	 2.3%  3.9%  2.9%
IBM Watson Analytics (57), 0% alone	 2.1% na na
Ayasdi (56), 10.7% alone	 2.0% na na
Dataiku (56), 7.1% alone	 2.0% na na
Action (56), 7.1% alone	 2.0%  0.5% na

Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013
H2O (0xdata) (55), 0% alone	2.0% 0.2% na
Orange (53), 0% alone	1.9% 3.4% 3.6%
Mathematica (52), 0% alone	1.9% 2.3% 2.1%
IBM Cognos (51), 0% alone	1.8% 1.8% 2.4%
Dell (including StatSoft) (47), 19.1% alone	1.7% 1.7% 9.0%
XLSTAT for Excel (42), 0% alone	1.5% 1.2% 0.9%

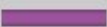
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
Stata (36), 2.8% alone	 1.3%  1.4%  2.1%
Lexalytics (35), 28.6% alone	 1.3% na na
Vowpal Wabbit (35), 0% alone	 1.3% na na
C4.5/C5.0/See5 (35), 0% alone	 1.3%  1.5%  1.1%
Julia (31), 3.2% alone	 1.1%  0.8% na
Splunk/ Hunk (30), 0% alone	 1.1%  0.7% na

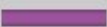
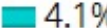
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
Microstrategy (24), 0% alone	 0.9% na na
BigML (23), 0% alone	 0.8%  0.9% na
Zementis (22), 31.8% alone	 0.8%  0.8%  0.9%
Miner3D (22), 9.1% alone	 0.8%  0.9%  1.8%
Oracle Data Miner (22), 0% alone	 0.8%  2.2%  1.0%
Amazon Machine Learning (20), 5.0% alone	 0.7% na na

Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	 % users in 2015  % users in 2014  % users in 2013
BayesiaLab (16), 12.5% alone	 0.6%  4.1%  1.0%
Dato (former Graphlab) (15), 6.7% alone	 0.5%  0.9% na
Clojure (13), 0% alone	 0.5%  0.5% na
Alpine Data Labs (13), 0% alone	 0.5%  2.7% na
Angoss (11), 18.2% alone	 0.4%  0.4%  0.3%
Lavastorm (10), 0% alone	 0.4%  0.3%  0.4%

Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]	
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013
Lisp (10), 0% alone	0.4% 0.3% na
Predixion Software (10), 0% alone	0.4% 3.7% 2.7%
WordStat (9), 0% alone	0.3% 0.2% 0.4%
Megaputer Polyanalyst/TextAnalyst (8), 0% alone	0.3% 0.1% 0.1%
WPS: World Programming System (7), 0% alone	0.3% 0.2% na
GoodData (6), 0% alone	0.2% 0.1% na

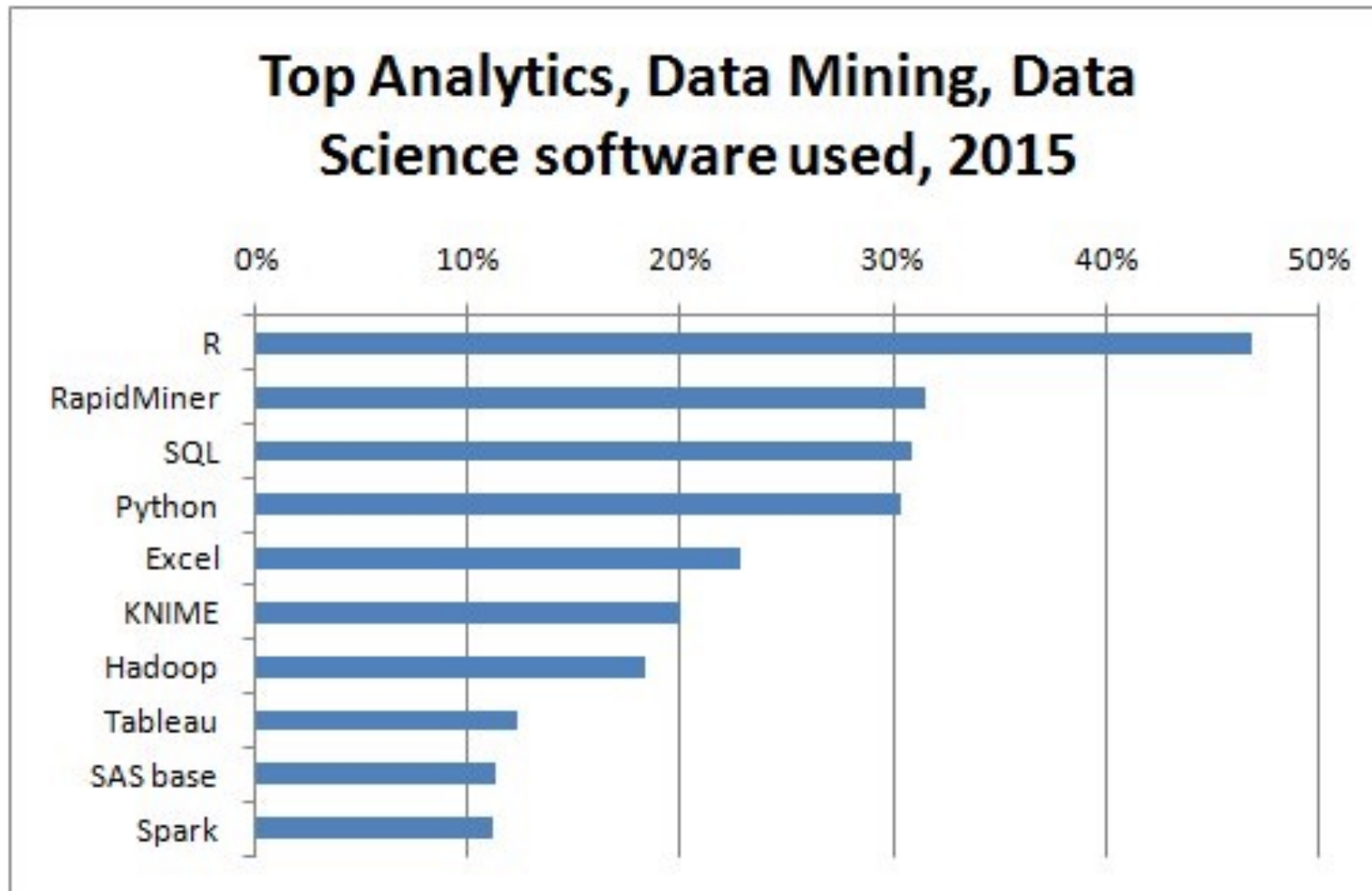
Phần mềm trong những năm gần đây

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

What Analytics, Big Data, Data mining, Data Science software you used in the past 12 months for a real project? [2759 voters]		
Legend: Red: Free/Open Source tools Green: Commercial tools Fuchsia: Hadoop/Big Data tools	% users in 2015 % users in 2014 % users in 2013	
MetaMind (5), 0% alone	0.2% na na	
SiSense (5), 0% alone	0.2% 0.1% na	
RapidInsight/Veera (5), 0% alone	0.2% 0.5% 0.5%	
Skytree (3), 0% alone	0.1% na na	
Birst (2), 0% alone	0.1% na na	
Ontotext (1), 0% alone	0% na na	
FICO Model Builder (1), 0% alone	0% na na	

Top 10 phần mềm (2015)

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- **giải thuật DM**
- kết luận và hướng phát triển
- tài liệu tham khảo



Nội dung

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- **kết luận và hướng phát triển**
- tài liệu tham khảo

Kết luận về KDD & DM

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ KDD & DM

- cần thiết
- khai thác, tìm kiếm tri thức
- ẩn trong khối lượng lớn dữ liệu
- áp dụng cho nhiều lĩnh vực khác nhau

■ KDD

- quá trình lặp lại
- tiền xử lý, khai thác dữ liệu & đánh giá kết quả
- DM là cốt lõi của quá trình KDD

Kết luận về KDD & DM

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ KDD & DM

- cần giải quyết được mục tiêu ban đầu đặt ra của ứng dụng
- khả năng làm việc tốt với khối lượng lớn dữ liệu
- đáp ứng được về những ràng buộc: thời gian, thiết bị, chất lượng kết quả, hiểu được
- « no free lunch theorem »: không có kỹ thuật DM nào là tốt trong tất cả các trường hợp

➤ rất khó, no free lunch ☹️

Hướng phát triển²

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- tài liệu tham khảo

■ tương lai, KDD & DM

- nguồn dữ liệu hỗn hợp: văn bản, hình ảnh, âm thanh, ký tự
- cải thiện độ phức tạp của quá trình để có khả năng làm việc tốt với khối lượng lớn dữ liệu: tạo giải thuật mới hoặc cải tiến giải thuật hiện có
- tích hợp kiến thức chuyên gia
- diễn dịch kết quả
- mở rộng sang những lĩnh vực ứng dụng khác như: sinh học, kinh tế, y học, quốc phòng, thiên văn, etc.

Metaphor của V. Vapnik 😊

- Solving a problem of interest, do not solve a more general problem as an intermediate step. Try to get the answer that you really need but not a more general one (Vapnik, 1995).

Nội dung

- tại sao KDD & DM là cần thiết?
- những ứng dụng của KDD & DM
- quá trình KDD
- giải thuật DM
- kết luận và hướng phát triển
- **tài liệu tham khảo**

Lịch sử

- 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
 - EGC (2001), PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.

Tham khảo

www.kdnuggets.com

- ngày 24/11/2008, sử dụng google để tìm
 - Data Mining: 11 000 000 links
 - Knowledge Discovery in Databases: 3 000 000 links

Tham khảo

■ 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.

Tham khảo

- K. Bennett and C. Campbell. Support Vector Machines: Hype or Hallelujah ?. SIGKDD Explorations, 2(2), pp. 1-13, 2000
- L. Breiman. Random Forests. Machine Learning, 45(1), pp. 5-32, 2001
- L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Chapman & Hall, New York, 1984
- N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, 2000
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001

Tham khảo

- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- J. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993
- V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995
- I. H. Witten and E. Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 2001
- C. Blake and C. Merz. UCI Repository of Machine Learning Databases. 1998.
<http://www.ics.uci.edu/~mlearn/MLRepository.html>
- Delve. Data for Evaluating Learning in Valid Experiments. 1996. <http://www.cs.toronto.edu/~delve>

Tham khảo

(<http://www.kdnuggets.com/software/index.html>)

■ phần mềm miễn phí

- **Python**: ngôn ngữ lập trình cấp cao, hỗ trợ nhiều thư viện, công cụ, giải thuật cho phân tích dữ liệu và DM
- **R**: ngôn ngữ lập trình cấp cao, hỗ trợ nhiều thư viện, công cụ, giải thuật cho phân tích dữ liệu và DM
- **WEKA**: thư viện DM viết bằng JAVA, dễ sử dụng, tích hợp và phát triển (<http://www.cs.waikato.ac.nz/~ml/weka/>)
- **C4.5**: giải thuật cây quyết định, rất nhanh và hiệu quả (<http://www.rulequest.com/Personal/>)
- **RandomForest**: giải thuật tập hợp những cây quyết định, nhanh và chính xác (<http://www.stat.berkeley.edu/users/breiman/>)

Sử dụng cho môn học

■ Download và sử dụng phần mềm

- **Python**: lập trình cho giải thuật máy học & phân tích dữ liệu
- **R**: lập trình cho thống kê & phân tích dữ liệu
- **GNU Octave**: « free Matlab »
- **C4.5**: cây quyết định
- **LibSVM**: giải thuật máy học vector hỗ trợ
- **Weka**: giải thuật máy học khác
- etc.

Sử dụng cho môn học

■ download các tập dữ liệu

- **Mushroom (UCI)**: luật kết hợp
- **Pima (UCI)**: giải thuật máy học
- **Spambase (UCI)**: giải thuật máy học
- **Sat-images (UCI)**: giải thuật máy học
- **Segment (UCI)**: giải thuật máy học
- **Forest cover types (UCI)**: giải thuật máy học
- **MNIST (LeCun)**: giải thuật máy học
- **ALL-AML Leukemia (Biomedical)**: giải thuật máy học

(UCI): <http://archive.ics.uci.edu/ml/>

(LeCun): <http://yann.lecun.com/exdb/mnist>

(Biomedical): <http://datam.i2r.a-star.edu.sg/datasets/krbd/>

Sử dụng cho môn học

- nên đọc lại các công cụ xử lý text đơn giản của Linux: cut, sed, awk
- lập trình script (bash)
- giải các bài tập : phân lớp dữ liệu, clustering, tìm luật kết hợp
- mỗi nhóm chọn 3 trong các tập dữ liệu để thực tập các giải thuật và báo cáo cuối kỳ (50% số điểm) và thi cuối kỳ để có thêm số điểm còn lại
- Anh/Chị có thể chọn cho mình một giải thuật mà Anh/Chị thích để xử lý dữ liệu của mình 😊



Cám ơn !