

داده‌کاوی و زبان برنامه‌نویسی R

ارائه شده در  جشنوارهٔ روز
آزادی نرم‌افزار



موسسه آموزش عالی صنعتی فولاد

نویسندگان مقاله:

الهام حصارکی (دانشجوی مقطع کارشناسی ارشد)

دکتر محمد داورپناه جزی (استادیار دانشگاه)

داده‌کاوی و زبان برنامه‌نویسی R

ارائه شده در  جشنوارهٔ روز
آزادی نرم‌افزار

ارائه‌دهنده:

الهام حصارکی

Elham.hesaraki@gmail.com

محتوای کارگاه

مقدمه

داده‌کاوی

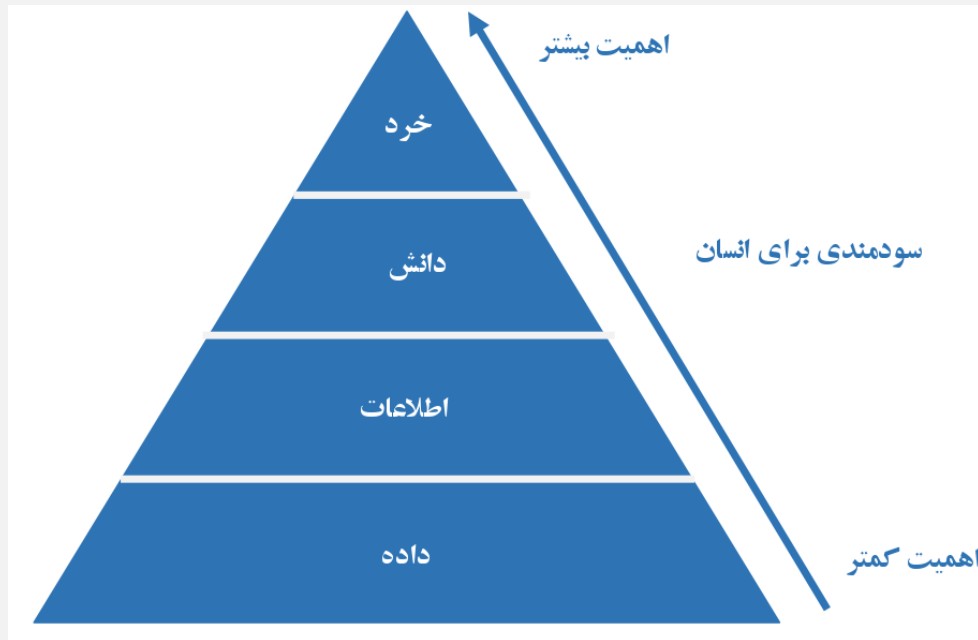
R و RStudio

پرسش و پاسخ

از داده تا خرد

مقدمه:

دسته‌بندی ذهن انسان (اکاف ۱۹۸۹)



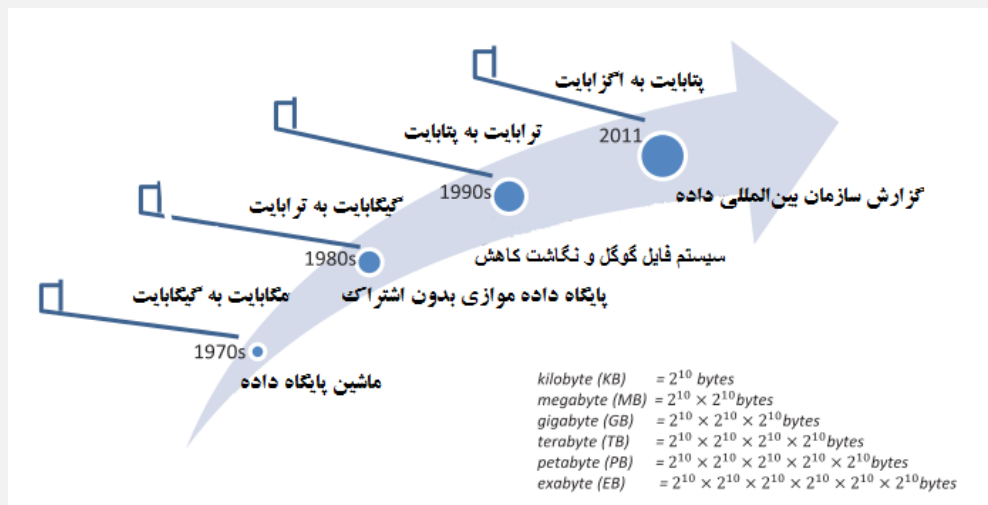
- داده
- اطلاعات
- دانش
- ادراک
- خرد
- نويز

شکل ۱ : از داده تا خرد (اکاف ۱۹۸۹).

از نیاز تا ابداع

روند روبه رشد تولید داده

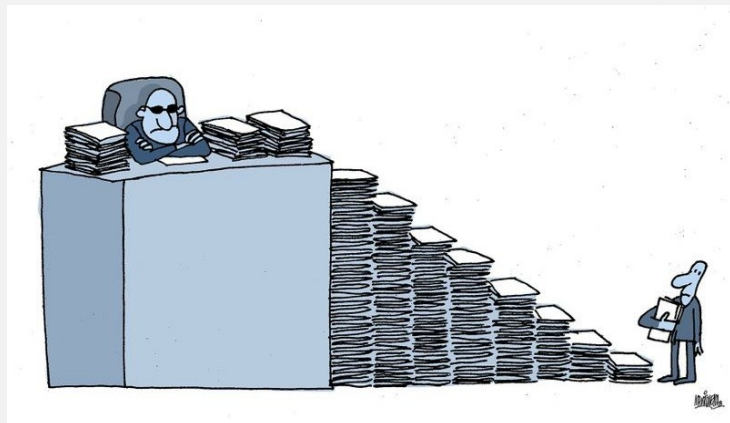
- مگابایت به گیگابایت
- گیگابایت به ترابایت
- ترابایت به پتابایت
- پتابایت به اگزابایت



شکل ۲: روند روبه رشد تولید داده (هو، ون، چوآ و لی ۲۰۱۴).

داده، داده،

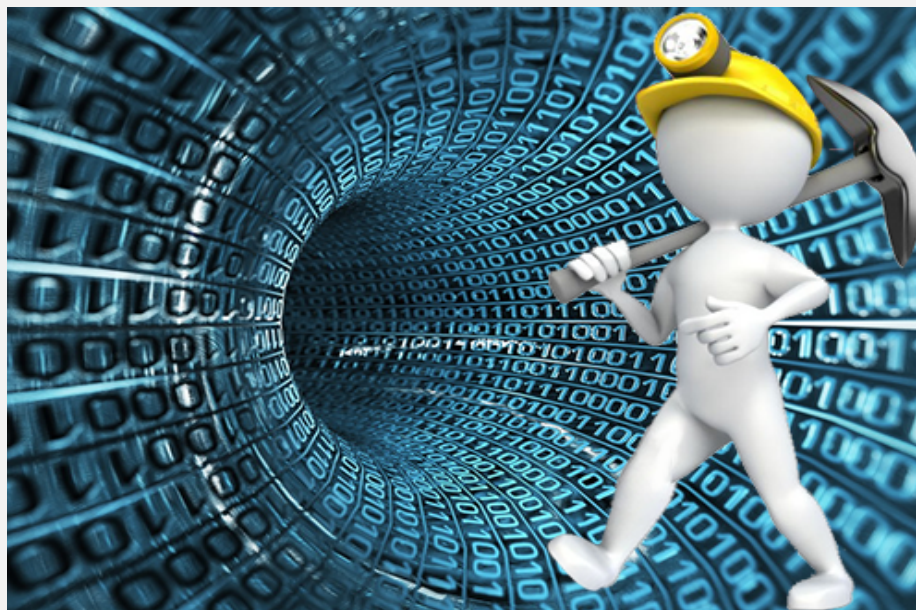
- داده مورد نیازم را نمی‌توانم به‌دست آورم
- داده‌های مورد نیازم در شبکه پراکنده شده‌اند
- داده‌های مورد نیازم را درک نمی‌کنم
- داده‌های کشف کرده را نمی‌توانم استفاده کنم



شکل ۳: جهان غنی از داده و فقر اطلاعاتی.

داده‌کاوی چیست؟

- یک علم میان‌رشته‌ای (هوش مصنوعی، آمار، یادگیری ماشین، سیستم‌های پایگاه داده)
- کشف الگوهای جالب و دانش از حجم انبود داده



شکل ۴ : داده‌کاوی.

داده‌کاوی چیست؟

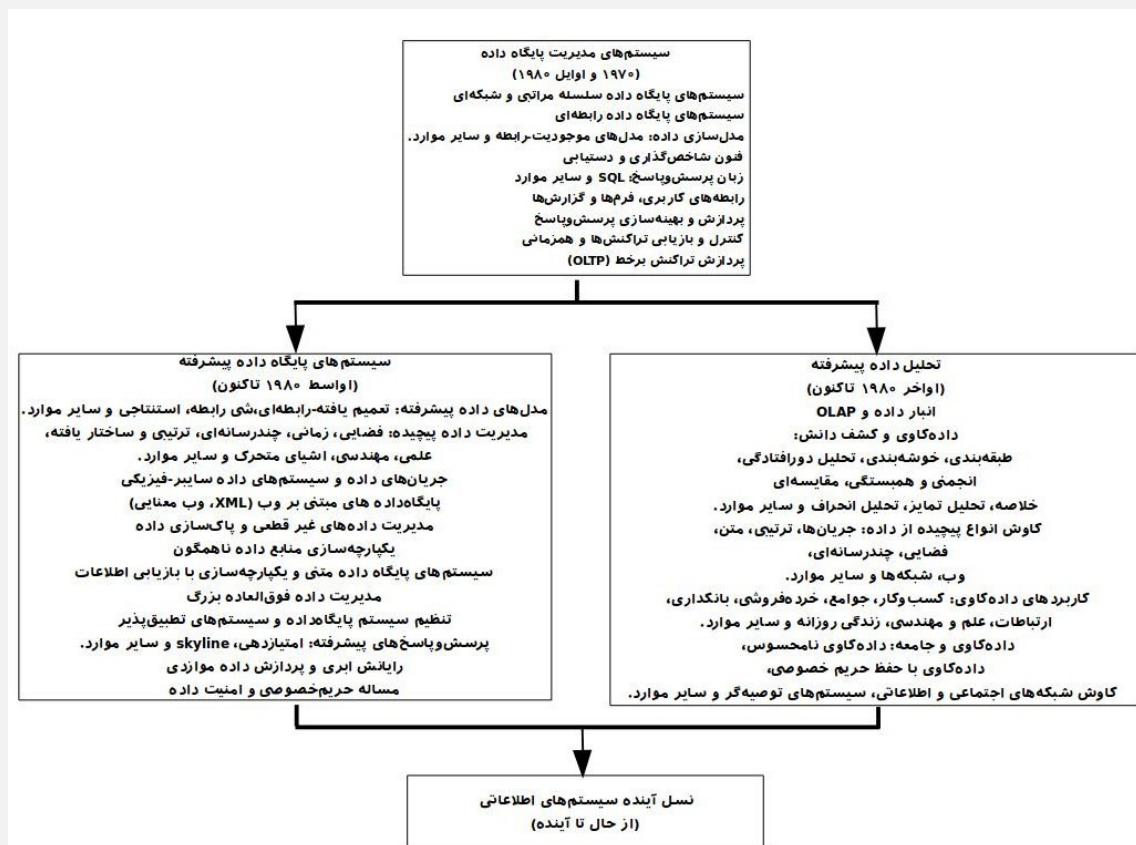
- استفاده از عبارات گوناگون برای یک فرآیند واحد

جدول ۱ : نام‌های گوناگون برای یک فرآیند.

سال	نام	استفاده‌کنندگان
۱۹۶۰	لایروبی داده، ماهی‌گیری داده	کارشناسان آمار
۱۹۸۹	کشف دانش در پایگاه‌های داده	هوش مصنوعی، جامعه کاربری یادگیری ماشین
۱۹۹۰	داده‌کاوی	جامعه پایگاه داده، کسب‌وکار
سایر اسامی: باستان‌شناسی داده، هرس اطلاعات، کشف اطلاعات، استخراج دانش		

داده‌کاوی چیست؟

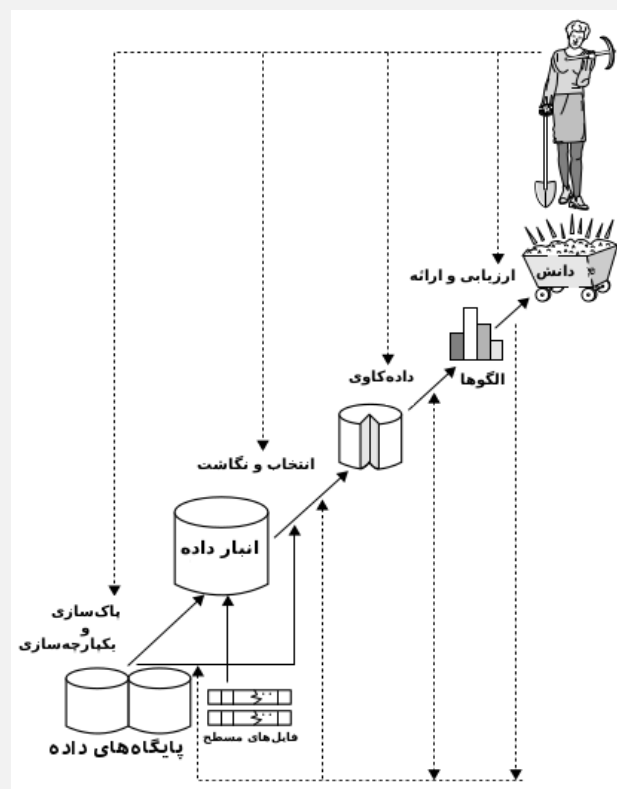
• داده‌کاوی نتیجه تکامل طبیعی فناوری اطلاعات



شکل ۵: تکامل فناوری اطلاعات و داده‌کاوی (هان و همکاران ۲۰۱۲).

فرآیند داده‌کاوی

- داده‌کاوی یک فرآیند هفت مرحله‌ای تکرارشونده و بازگشت‌پذیر است



شکل ۶ : هفت گام فرآیند داده‌کاوی (هان و همکاران ۲۰۱۲).

فرآیند داده‌کاوی

پیش‌پردازش

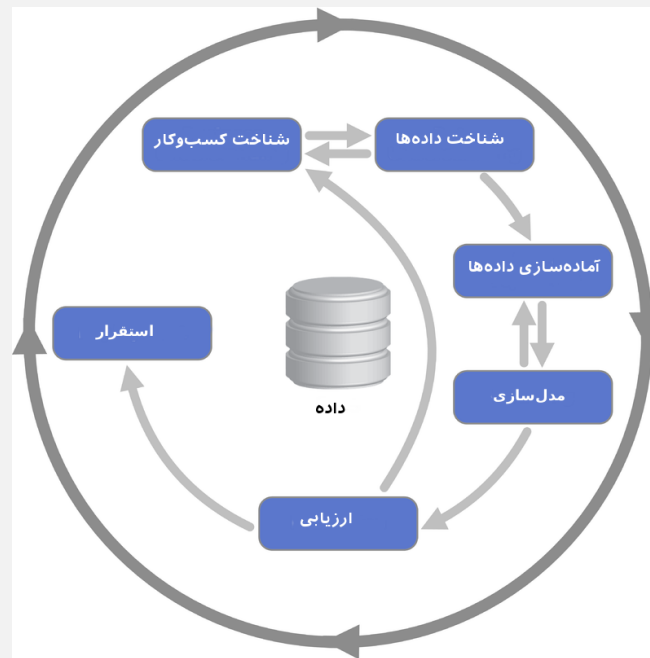


- پاکسازی داده (حذف نویز و داده‌های ناسازگار)
- یکپارچه‌سازی داده (ترکیب منابع داده چندگانه)
- انتخاب داده (انتخاب داده‌های مرتبط با تحلیل)
- نگاشت داده
- داده‌کاوی (استفاده از روش‌های هوشمندانه جهت کشف دانش)
- ارزیابی الگو (سنجش‌های جالب بودن)
- ارائه دانش

مدل فرآیند CRISP-DM

داده‌کاوی:

- *Cross Industry Standard Process for Data Mining*
- مدلی از رویکردهای مورد استفاده کارشناسان داده‌کاوی در حل مسائل



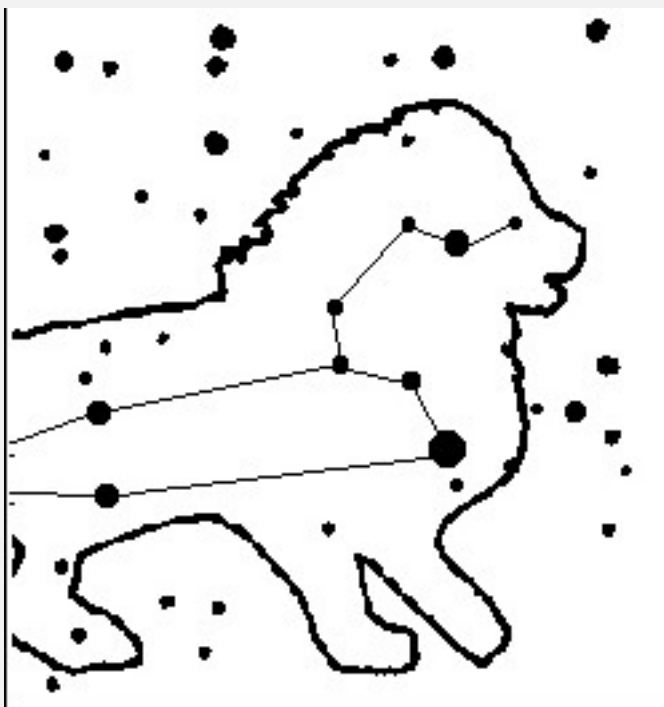
شکل ۷: مدل CRISP-DM.

انواع داده‌های قابل کاوش

- داده‌های پایگاه داده (رابطه‌ای)
- داده‌های تراکنشی (پایگاه داده تراکنشی)
- داده‌های انبار داده
- داده‌های جریان داده، داده‌های فضایی، زمانی
- داده‌های چندرسانه‌ای، متنی، ترتیبی
- گراف‌ها و درخت‌ها

انواع الگوهای قابل کشف

- توصیف کلاس/مفهوم (مشخصه‌سازی و تمایز)
- الگوهای مکرر، وابستگی و همبستگی
- طبقه‌بندی و رگرسیون
- خوشه‌بندی
- تحلیل دورافتادگی



شکل ۸: کشف الگو.

سنجه‌های جالبی الگوها

داده‌کاوی:

- نو و بدیع بودن
- قابل درک بودن برای انسان
- کارا بودن
- صحت داشتن برای داده‌های جدید یا آزمون با درجه‌ای از قطعیت
- (صحه گذاشتن بر یک پیش‌فرض منطقی)

یک الگوی جالب، دانش ارائه می‌کند.

زمینه‌های کاربرد

داده‌کاوی:

- همه‌جا، داده‌کاوی در هرکجا



شکل ۹: بدون شرح.

آشنایی با برخی از واژگان کلیدی

داده‌کاوی:

جدول ۲: برخی واژگان کلیدی در داده‌کاوی.

Instances	Attribute
Sample	Feature
Example	Dimension
Data Point	Column
Objects	Variable
Raw	Field
Record	
Tuple	

انواع خصیصه‌ها

جدول ۳: انواع خصیصه‌ها.

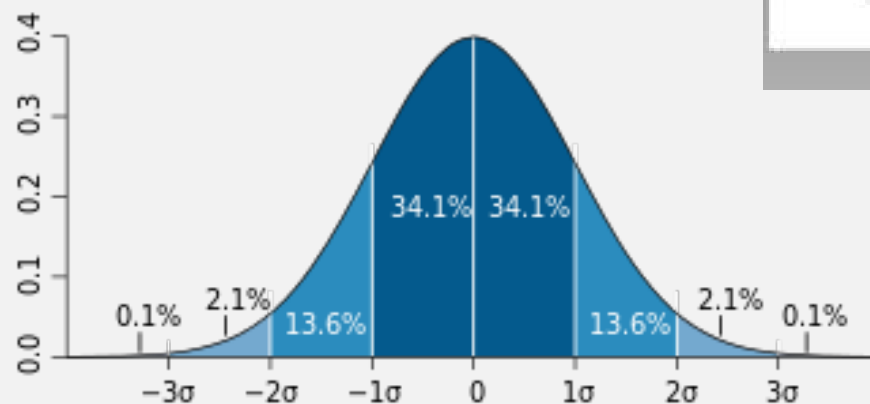
اسمی (دسته‌ای)	ترتیبی	عددی
- اسمی - دودویی (بولین)		- خصیصه‌های عددی فاصله‌ای - خصیصه‌های عددی نسبی
گسسته یا پیوسته		

مفاهیم آماری مقدماتی

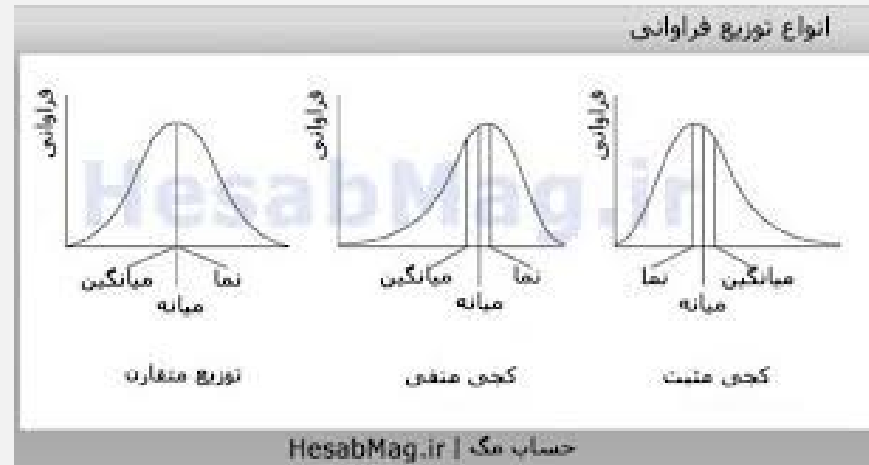
- گرایش مرکزی: میانگین، میانه، مد، رنج و وسط
- سنجش پراکندگی: رنج، چارک و چندک، رنج میان‌چارکی، واریانس و انحراف معیار
- نمودارهای آماری
- روش‌های اندازه‌گیری مشابهت و عدم مشابهت
- بصری‌سازی داده (مبتنی بر پیکسل، تصویرسازی هندسی، مبتنی بر نمایه، سلسله‌مراتبی، نمایش داده‌های پیچیده)

مفاهیم آماری مقدماتی

داده‌کاوی:



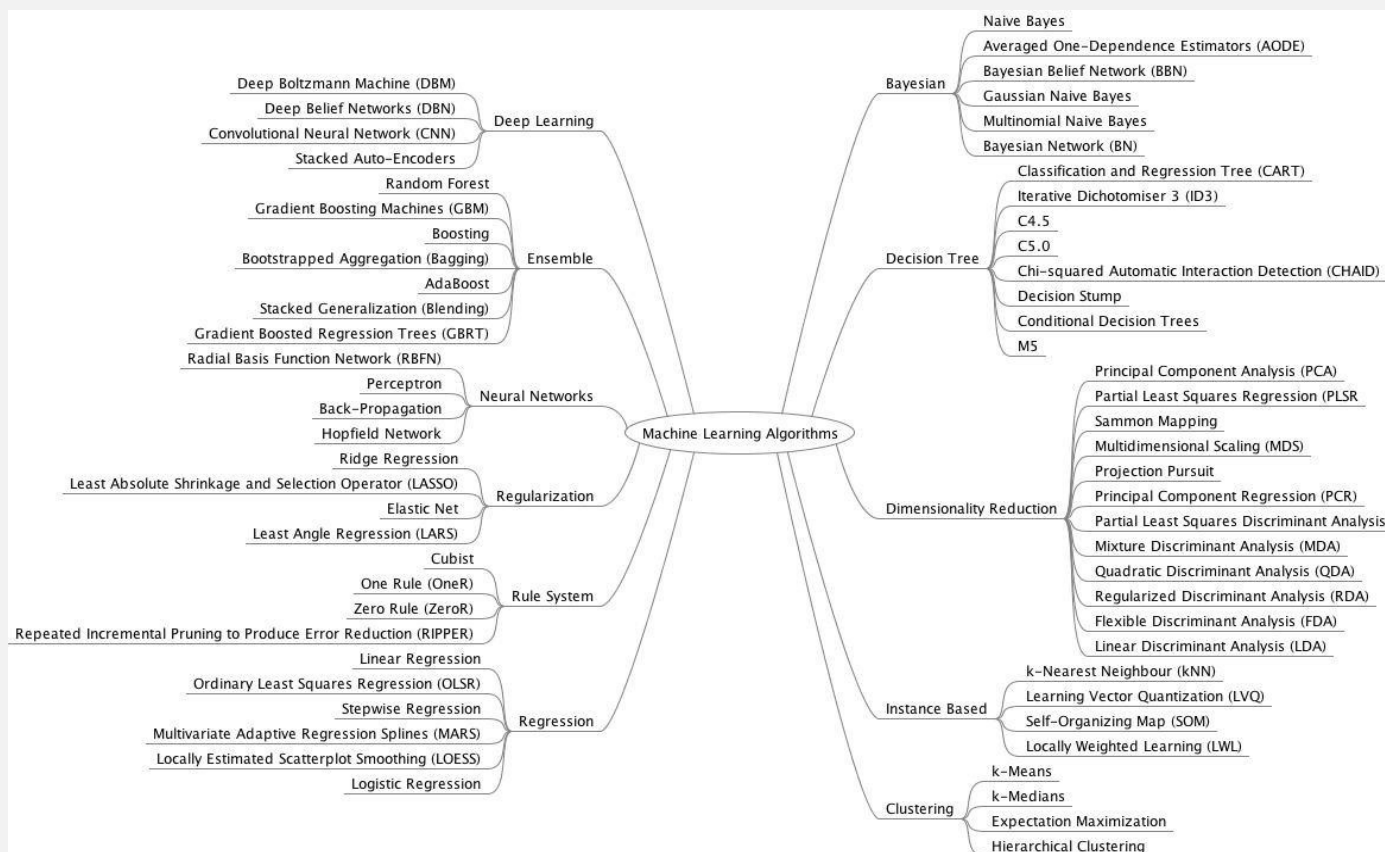
شکل ۱۱: چارک و چندک.



شکل ۱۰: گرایش‌های مرکزی.

روش‌های داده‌کاوی

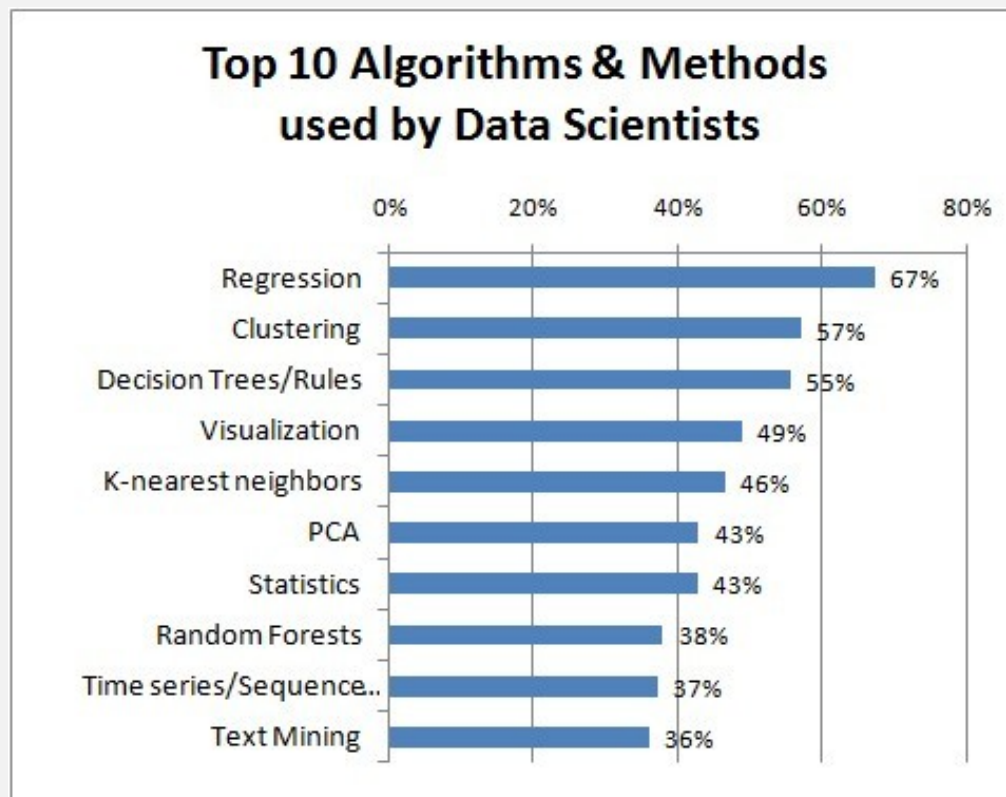
داده‌کاوی:



شکل ۱۲: دسته‌بندی الگوریتم‌ها.

مقایسه الگوریتم‌های داده‌کاوی

داده‌کاوی:



شکل ۱۳: ده الگوریتم و روش برتر استفاده شده توسط دانشمندان داده (kdnuggets.com).

مقایسه الگوریتم‌های داده‌کاوی

جدول ۴ : استفاده از الگوریتم‌ها بر اساس نوع کاربری (kdnuggets.com).

استفاده بر اساس نوع کاربری	رأی دهندگان	میانگین استفاده الگوریتم‌ها	نظارت شده	نظارت نشده	استفاده از متا	سایر روش‌ها
صنعت	۵۹٪	۸/۴	۹۴٪	۸۱٪	۵۵٪	۸۳٪
دولت/ناسودبر	۴/۱٪	۹/۵	۹۱٪	۸۹٪	۴۹٪	۸۹٪
دانش‌آموزی	۱۶٪	۸/۱	۹۴٪	۷۶٪	۴۷٪	۷۷٪
آکادمیک	۱۲٪	۷/۲	۹۵٪	۸۱٪	۴۴٪	۷۷٪
همه		۸/۳	۹۴٪	۸۲٪	۴۸٪	۸۱٪

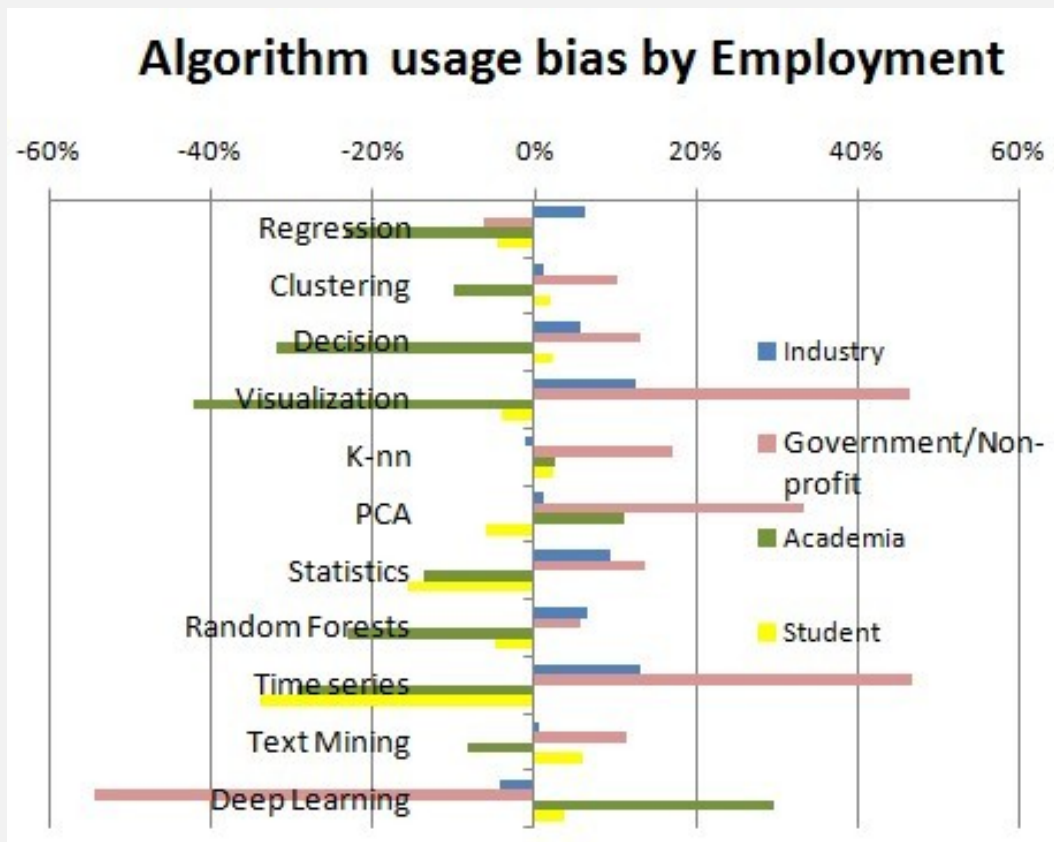
مقایسه الگوریتم‌های داده‌کاوی

جدول ۵ : ده الگوریتم برتر + یادگیری عمیق و نوع کاربری. (kdnuggets.com).

الگوریتم	صنعت	دولت/ناسودبر	آکادمیک	دانش‌آموزی	همه
رگرسیون	۷۱٪	۶۳٪	۵۱٪	۶۴٪	۶۷٪
خوشه‌بندی	۵۸٪	۶۳٪	۵۱٪	۵۸٪	۵۷٪
تصمیم	۵۹٪	۶۳٪	۳۸٪	۵۷٪	۵۵٪
بصری‌سازی	۵۵٪	۷۱٪	۲۸٪	۴۷٪	۴۹٪
KNN	۴۶٪	۵۴٪	۴۸٪	۴۷٪	۴۶٪
PCA	۴۳٪	۵۷٪	۴۸٪	۴۰٪	۴۳٪
آماري‌ها	۴۷٪	۴۹٪	۳۷٪	۳۶٪	۴۳٪
جنگل تصادفی	۴۰٪	۴۰٪	۲۹٪	۳۶٪	۳۸٪
سری‌های زمانی	۴۲٪	۵۴٪	۲۶٪	۲۴٪	۳۷٪
متن‌کاوی	۳۶٪	۴۰٪	۳۳٪	۳۸٪	۳۶٪
یادگیری عمیق	۱۸٪	۹٪	۲۴٪	۱۹٪	۱۹٪

مقایسه الگوریتم‌های داده‌کاوی

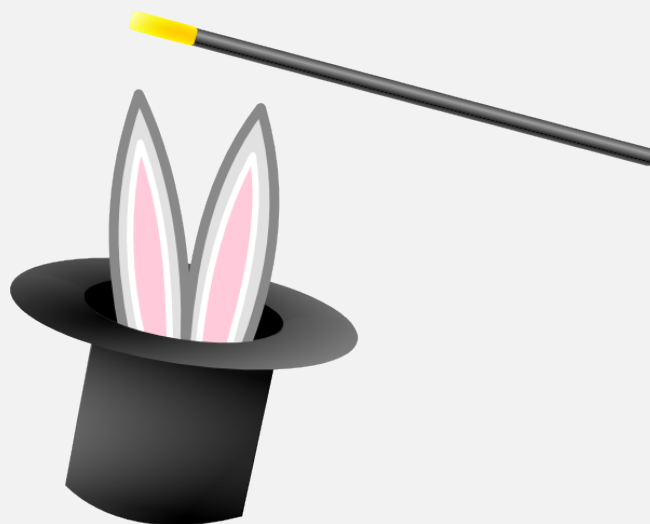
داده‌کاوی:



شکل ۱۴: سوگیری استفاده از الگوریتم‌ها بر اساس کاربرد (kdnuggets.com).

انتخاب روش داده‌کاوی

- روشی که به آن مسلط هستم را استفاده می‌کنم!
- دسته‌بندی الگوریتم‌ها بر اساس هدف مساله، نوع داده، حجم داده
- بررسی کارهای پیشین انجام شده در آن دامنه کاربرد



شکل ۱۵: بدون شرح.

ابزارهای داده‌کاوی

- زبان‌های برنامه‌نویسی (R، پایتون و سایر موارد)
- نرم‌افزارهای کاربردی (وکا، مطلب، IBM SPSS Modeler و سایر موارد)



شکل ۱۶: برخی ابزارهای آزاد داده‌کاوی.

مثال‌های جهان واقعی

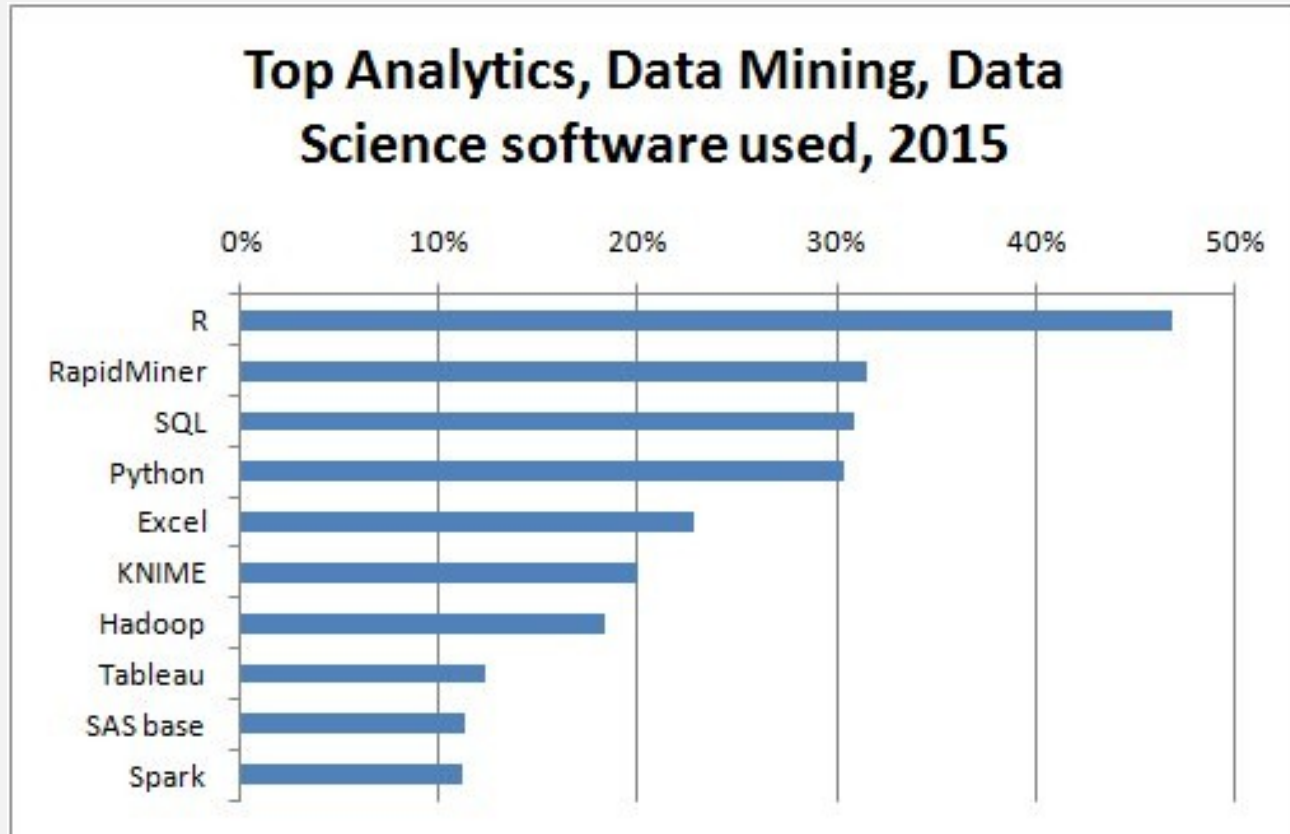
- گوگل
- پیش‌بینی سرطان
- تحلیل شبکه‌های اجتماعی در راستای تعیین نتایج انتخابات
- بخش‌بندی مشتریان (ایران - یک ارائه‌دهنده خدمات اینترنتی)
- شناسایی مشتریان خوش‌حساب و بدحساب (ایران-ایرانسل)
- و موارد بسیار دیگر (شامل داده‌کاوی پنهان)

زبان برنامه‌نویسی R

- زبان برنامه‌نویسی + محیط نرم‌افزاری برای محاسبات آماری و علم داده‌ها
- پیاده‌سازی شده براساس زبان‌های اس و اسکیم
- نرم‌افزار آزاد، تحت اجازه‌نامه عمومی همگانی گنو
- رایگان
- حاوی محدوده گسترده‌ای از تکنیک‌های آماری (مدل‌سازی خطی و غیرخطی، آزمون‌های کلاسیک آماری، تحلیل سری‌های زمانی، رده‌بندی، خوشه‌بندی و سایر موارد)
- قدرتمند در تولید اشکال گرافیکی و نمودارها
- در R، کدهای سی، سی++ و فورترن قابلیت اتصال و فراخوانی دارند

زبان برنامه‌نویسی R

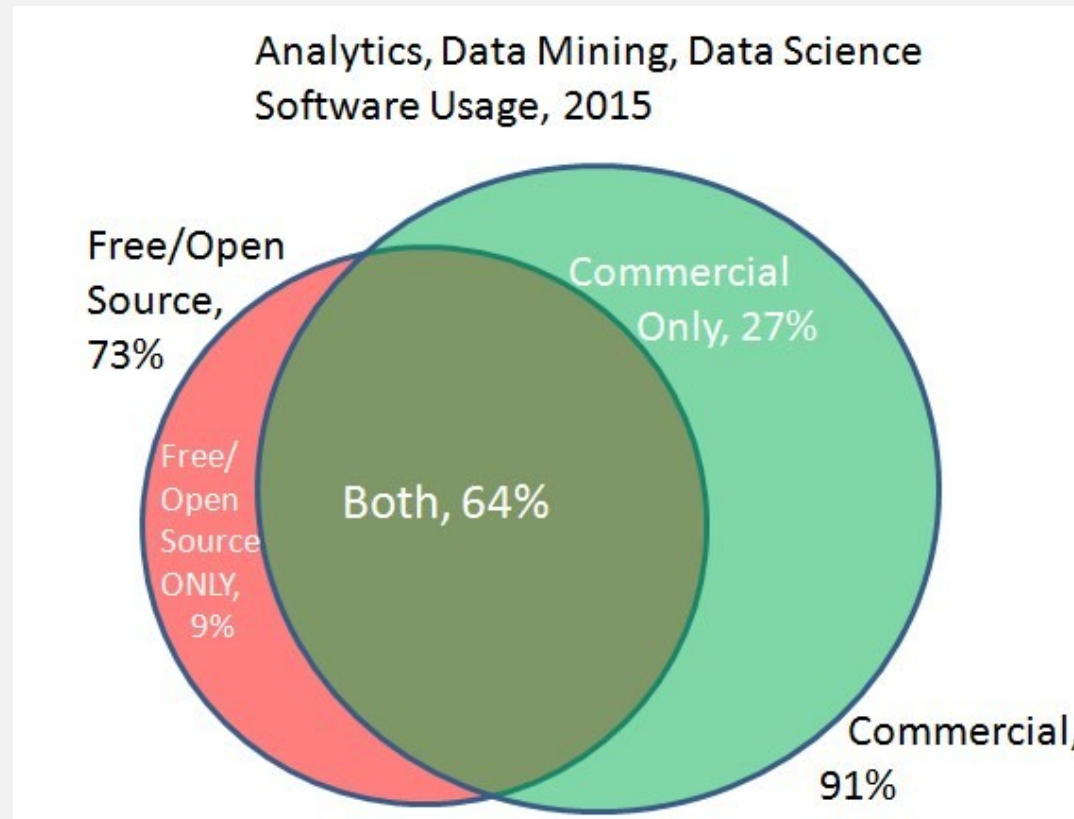
آر:



شکل ۱۷: ابزارهای داده‌کاوی برتر سال ۲۰۱۵ (kdnuggets.com).

زبان برنامه‌نویسی R

آر:



شکل ۱۸: استفاده از ابزارهای داده‌کاوی (kdnuggets.com).

محیط توسعه یکپارچه R

- یک محیط توسعه یکپارچه برای زبان برنامه نویسی R
- آزاد و تحت اجازه نامه نرم افزاری عمومی همگانی گنو آفرو
- دارای دو ویرایش رومیزی و سرور
- نوشته شده به زبان سی++
- استفاده از چارچوب کیوت جهت طراحی رابط کاربری گرافیکی

داده‌کاوی با R

- ابتدا R نصب کنید، سپس Rstudio
- نوع داده‌هایی که R پشتیبانی می‌کند: بردار، ماتریس، لیست، دیتافریم، کمیت‌های نرده‌ای
- دارای دو ویرایش رومیزی و سرور
- نوشته شده به زبان سی++
- استفاده از چارچوب کیوت جهت طراحی رابط کاربری گرافیکی

داده‌کاوی با R (ادامه در نرم‌افزار)

- نگاه مقدماتی به کدزندی در R
- ترسیم کلودتگ اشعار حافظ در R

منابع و مراجع

- [١] Bellinger, G., Castro, D., & Mills, A. (2004). Data, information, knowledge, and wisdom. URL: <http://www.systems-thinking.org/dikw/dikw.htm>.
- [٢] Mouthann N Effects of big data analytic on organization's value creation, Faculty of science and faculty of economics and business, University of Amsterdam
- [٣] Ras, P. (2012, May 4). Big data - wat was de vraag? Computable, 45.
- [٤] Leeuwen, R. V. (n.d.). De opvolger van de cloud: big data. Sync. Retrieved October 17, 2011, from <http://sync.nl/de-opvolger-van-de-cloud-big-data>.
- [٥] Hu H, Wen Y, Chua T.S, Li X "Toward sable systems for big data analytics: A technology tutorial", 2014, IEEE.
- [٦] Han J and Kamber M and Cerra d, "Data Mining: Concepts and Techniques", Morgan Kaufmann, Waltham, Massachusetts, USA, 2012.
- [٧] Kantardzic M, "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley & Sons, Inc, , Hoboken, NJ, USA, Second Edition, 2011.

منابع و مراجع

- [۸] Tamilselvi R and Kalaiselvi s, "An Overview of Data Mining Techniques and Applications", International Journal of Science and Research (IJSR), 2(2), 506-509, 2013.
- [۹] Gibert K, Sànchez-Marrè M and Codina V, "Choosing the Right Data Mining Technique: Classification of Methods and Intelligent Recommendation", International Congress on Environmental Modelling and Software Modelling for Environment's Sake, Ottawa, Canada, 2010.
- [۱۰] <http://www.kdnuggets.com/2016/09/poll-algorithms-used-data-scientists.html> آخرین (بازدید در تاریخ ۱۳۹۵/۰۷/۰۷).
- [۱۱] <http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html> آخرین (بازدید در تاریخ ۱۳۹۵/۰۷/۰۷).
- [۱۲] Zahao Y, "R and Data mining: Examples and Case Studies", junuary 2012.

