

Dive into

Data Science

شیرجه در علم داده

گروه دایکه . dayche.com





Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data. by Thomas H. Davenport and D.J. Patil

From the Magazine (October 2012)

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

گروه دایکه | dayche.com 



Data science growth over the years

650%

*Demand for data science jobs
has increased by 650% since
2012*

Source: LinkedIn

11.5M

*11.5 million new data jobs are
estimated to be created by
2026 in the USA alone*

Source: US Bureau of Labor Statistics

#1

*"Artificial Intelligence" skills
are skills with the fastest
growing roles since 2020*

Source: [LinkedIn](#)

تولید محتوا: محمد روزبه

daychegroup

daychegroup

گروه دایکه | dayche.com



How much does a Data Scientist make?

Updated Aug 13, 2023

Experience

All years of Experience



Industry

All industries



\$152,262 / yr

Total Pay

\$117,651 / yr

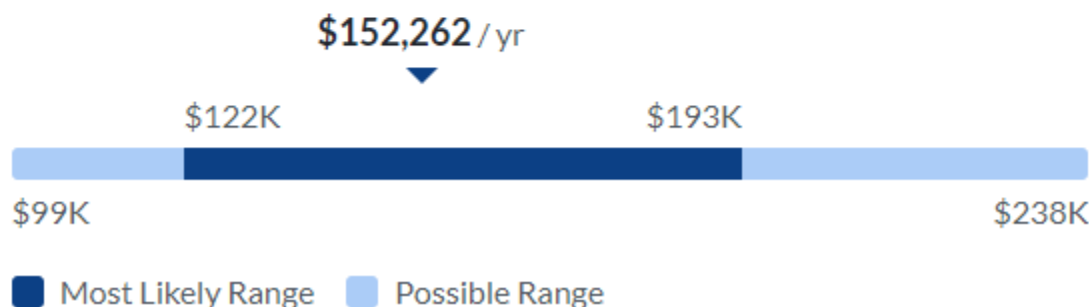
Base Pay

\$34,612 / yr

Additional Pay

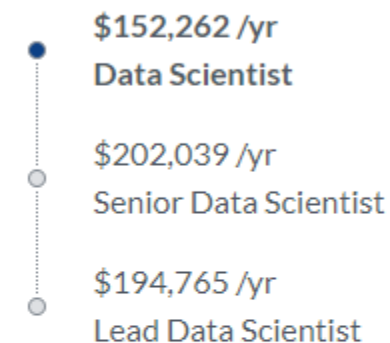


Confident



Total Pay Trajectory

For Data Scientist



[See Full Career Path](#)

[Download as data table](#)

تولید محتوا: محمد روزبه

daychegroup

daychegroup

گروه دایکه | dayche.com

علوم داده و هوش مصنوعی تهران

کارشناس

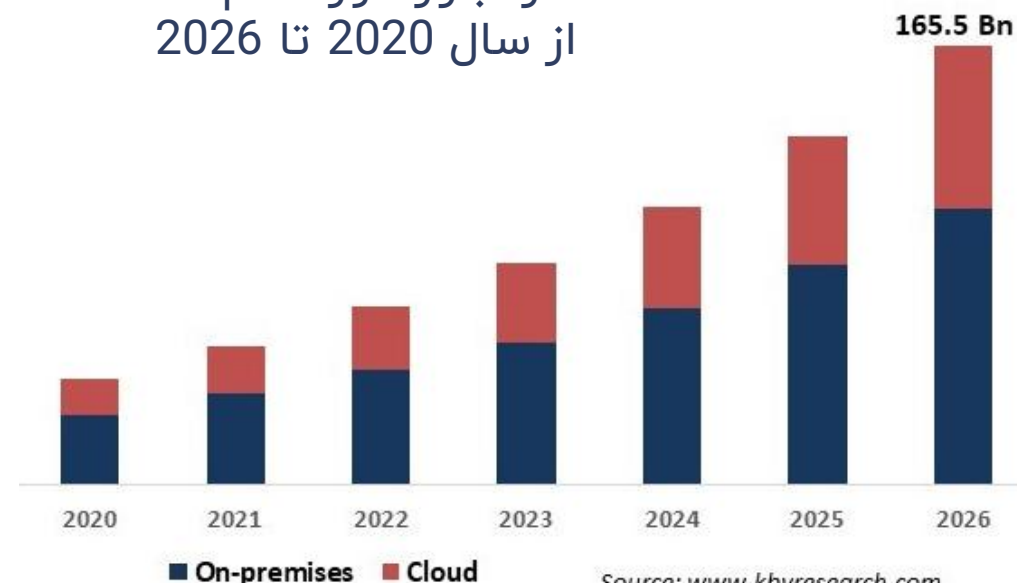


کارشناس ارشد



جایگزینی

نرخ رشد سالانه 27%
اندازه بازار حوزه علم داده
از سال 2020 تا 2026



Source: www.kbvresearch.com

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایچه



Why Data Science Important?



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

High Value Data Science Use Cases



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

□ نمونه 1: رانندگی خودکار

- مدل تشخیص اشیا
- مدل پیش بینی رفتار
- مدل طراحی مسیر/ قوانین رانندگی



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



□ نمونه 2: تحقیقات پزشکی و درمان

- مدل های تشخیص بیماری/ضایعه/تومور/...
- طراحی دارو و مدل اثربخشی درمان
- دستیار پزشکی هوشمند بر اساس شاخص های سلامتی بدن



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

□ نمونه 2: تحقیقات پزشکی و درمان

○ مثال: تشخیص کووید-19 از روی CT Scan ریه

در سامانه Daycher Medical



○ مدل شناسایی ضایعه ریوی و نوع آن

○ مدل تفکیک کووید-19 و آنفلوآنزا

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

نمونه 3: خرده فروشی و تجارت الکترونیک

(سرویس دهندگان B2C)

- بخش بندی و تحلیل رفتار خرید مشتریان
- مدل پیش بینی ریزش مشتریان
- سیستم های پیشنهاد دهنده به مشتریان
- بهینه سازی قیمت
- تحلیل رفتار و احساسات در شبکه های اجتماعی



تولید محتوا: محمد روزبه

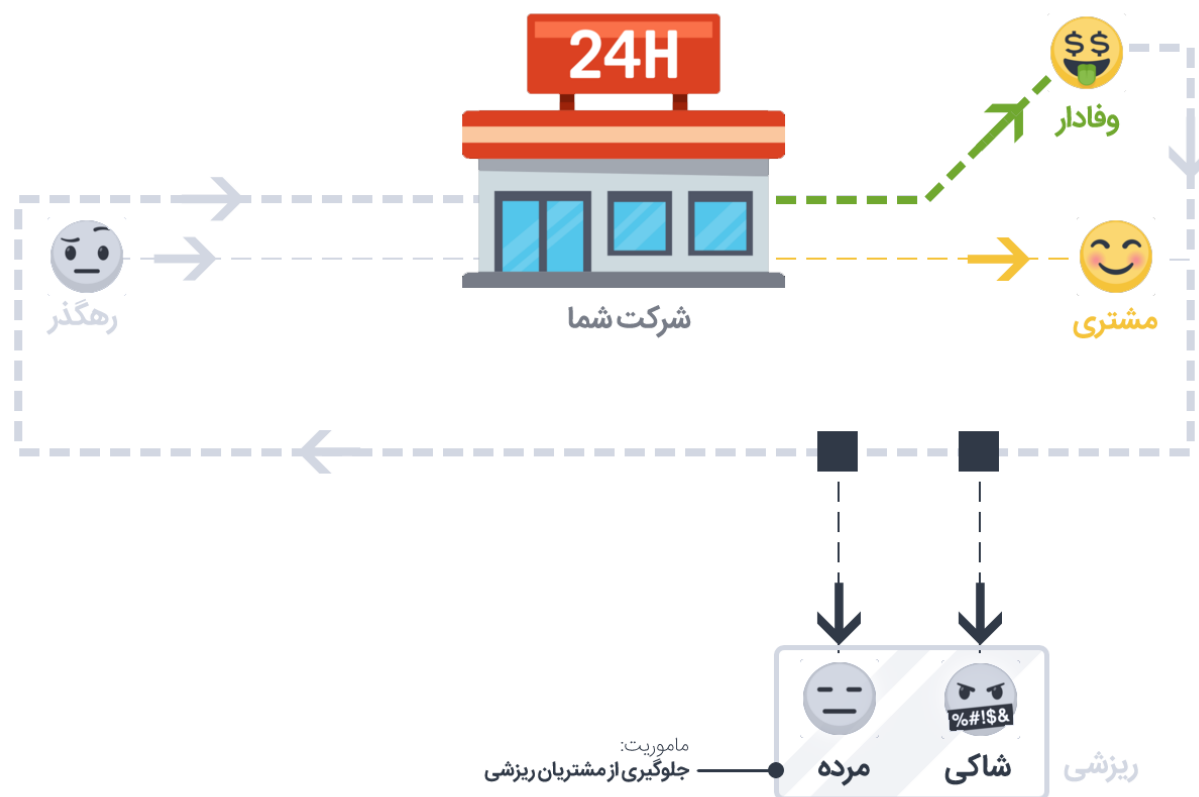
daychegroup

daychegroup

dayche.com | گروه دایکه

نمونه 3: خرده فروشی و تجارت الکترونیک

(سرویس دهندگان B2C)



○ مثال: تحلیل رفتار مشتریان در سامانه تحلیل رفتار Daycher CBM

تولید محتوا: محمد روزبه

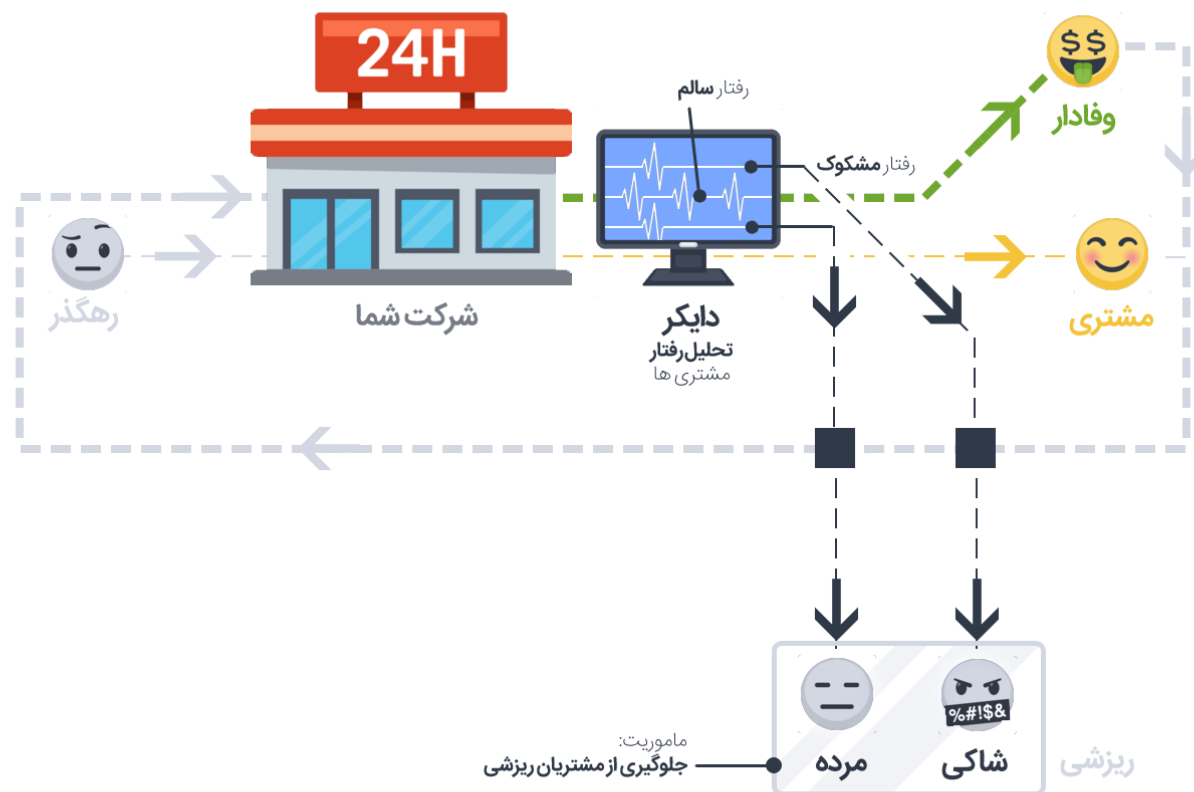
daychegroup

daychegroup

dayche.com | گروه دایکه

نمونه 3: خرده فروشی و تجارت الکترونیک

(سرویس دهندگان B2C)



مثال: تحلیل رفتار مشتریان در سامانه تحلیل رفتار Daycher CBM

تولید محتوا: محمد روزبه

daychegroup

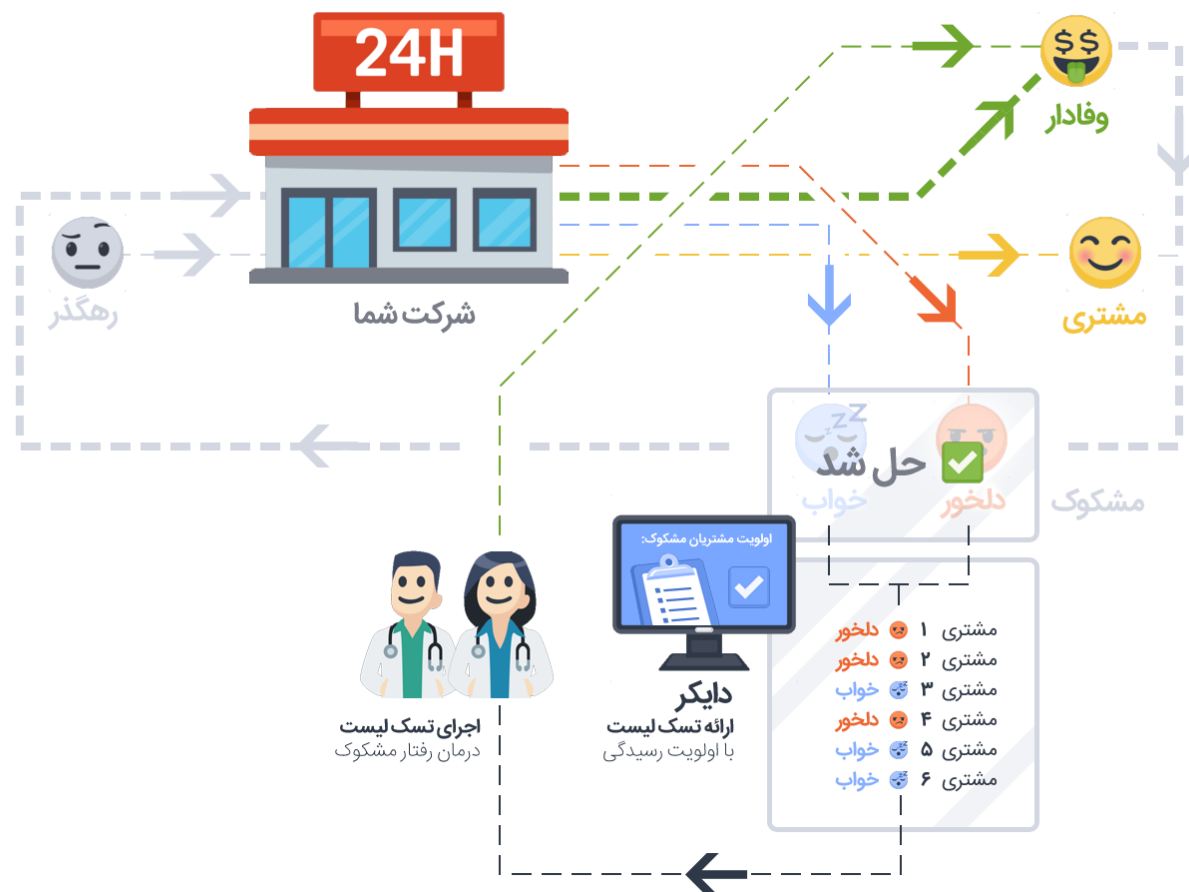
daychegroup

dayche.com | گروه دایکه

نمونه 3: خرده فروشی و تجارت الکترونیک

(سرویس دهندگان B2C)

مثال: تحلیل رفتار مشتریان در
سامانه تحلیل رفتار Daycher CBM



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

□ نمونه 3: خرده فروشی و تجارت الکترونیک (سرویس دهندگان B2C)

○ مثال: تحلیل رفتار مشتریان

سامانه تحلیل رفتار **Daycher CBM**

- شناسایی الگوهای رفتاری مشتریان
- مانیتورینگ تغییرات الگوی رفتاری
- پیش بینی تغییرات رفتاری مشتریان
- ارائه لیست های عملیاتی بر اساس اولویت



تولید محتوا: محمد روزبه

daychegroup

daychegroup

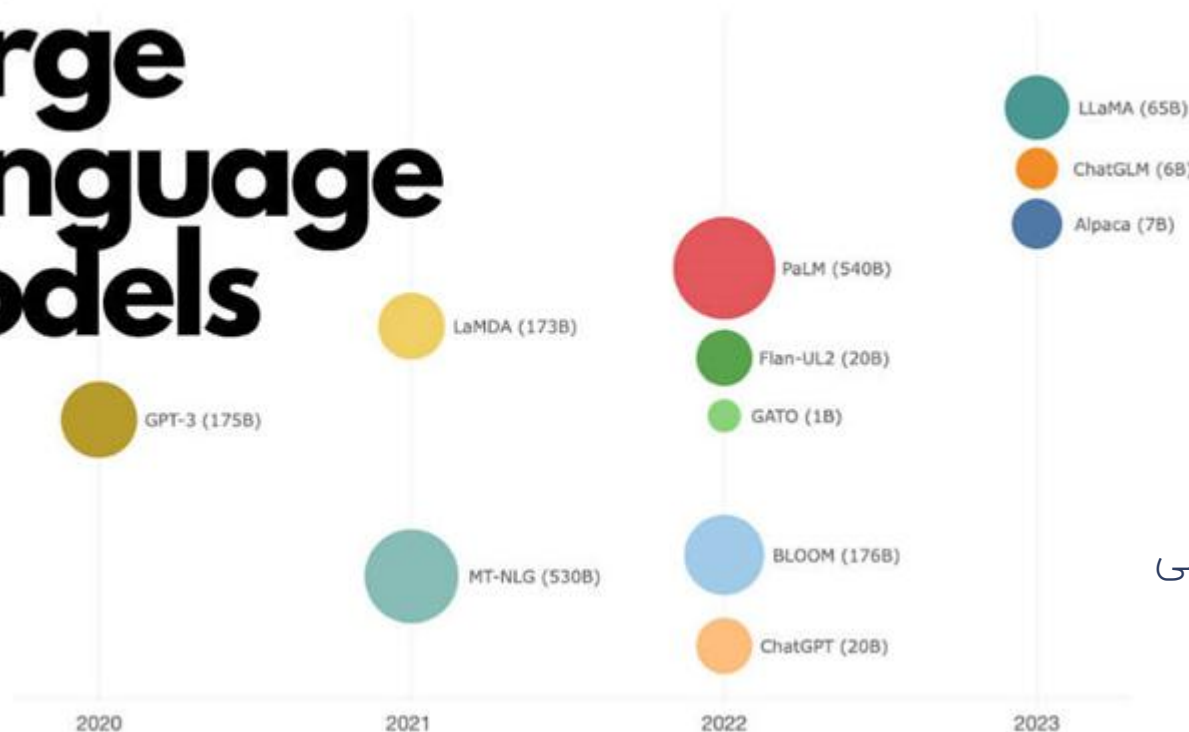
dayche.com | گروه دایکه

□ نمونه 4: مدل‌های زبانی بزرگ (LLMs)
(دستیاران هوشمند)

○ مثال: ChatGPT

- مکالمه هوشمند شبه انسانی
- توسعه چت بات‌ها
- تولید محتوا
- بهبود ملموس در کیفیت و سرعت کدنویسی
- و کاربردهای بسیار دیگر ...

Top Large Language Models



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

Data Science

مهارت های لازم

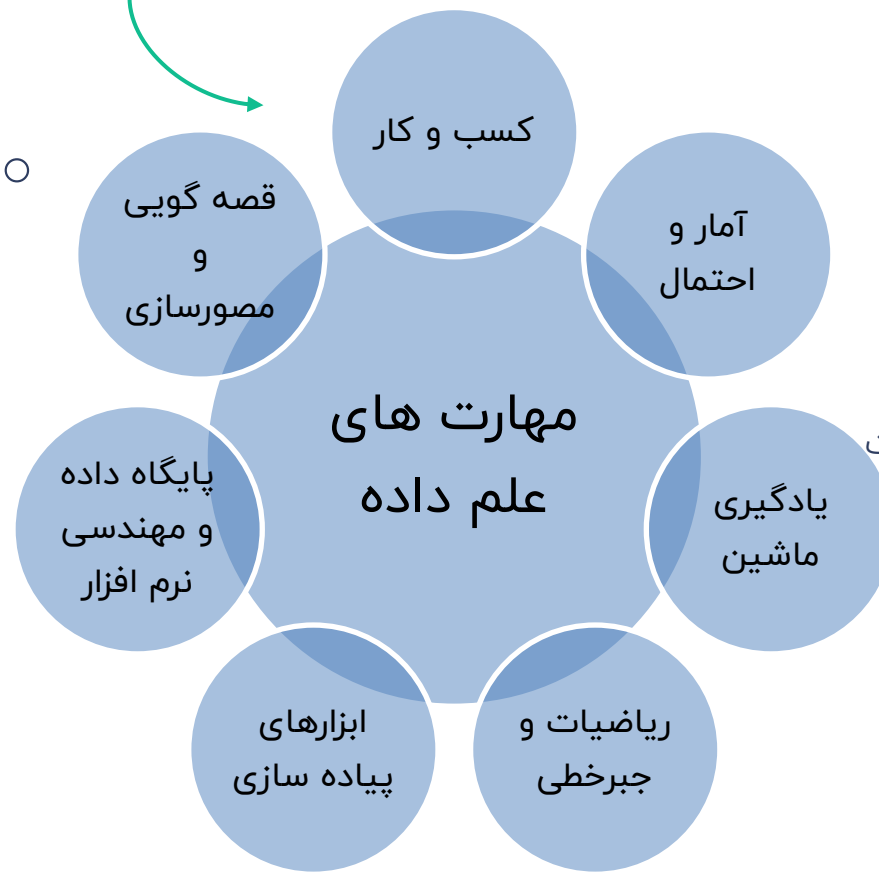
گروه دایکه . dayche.com



□ علم داده: موضوعی میان رشته ای

○ پایگاه داده و مهندسی نرم افزار:

- SQL / No-SQL
- API & Web Scraping
- Hadoop and Spark
- OS Concept and Git
- Software Deployment Concepts



○ ریاضیات و جبر خطی:

- احتمالات و ریاضیات عدم قطعیت
- حساب دیفرانسیل، مشتقات و روش های بهینه سازی
- جبر خطی، فضاها و برداری و تبدیلات

○ ابزارهای پیاده سازی:

- IBM SPSS, RapidMiner و ...
- برنامه نویسی Python / R

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



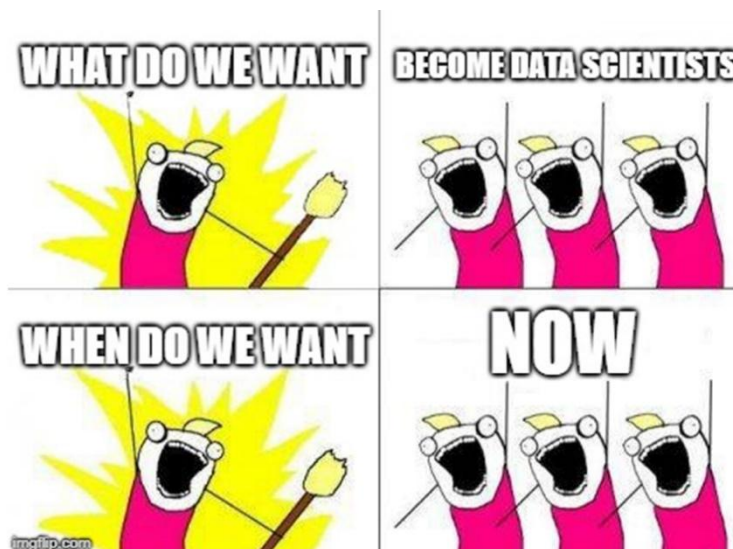
Josh Wills
@josh_wills

...

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

9:25 PM · May 3, 2012 · Twitter Web Client

1.8K Retweets 80 Quote Tweets 1.8K Likes



علم داده: موضوعی میان رشته ای □

دانشمند داده: شخصی که آمار را بهتر از هر مهندس کامپیوتری می داند و مهندسی کامپیوتر را بهتر از هر متخصص آماری می فهمد.

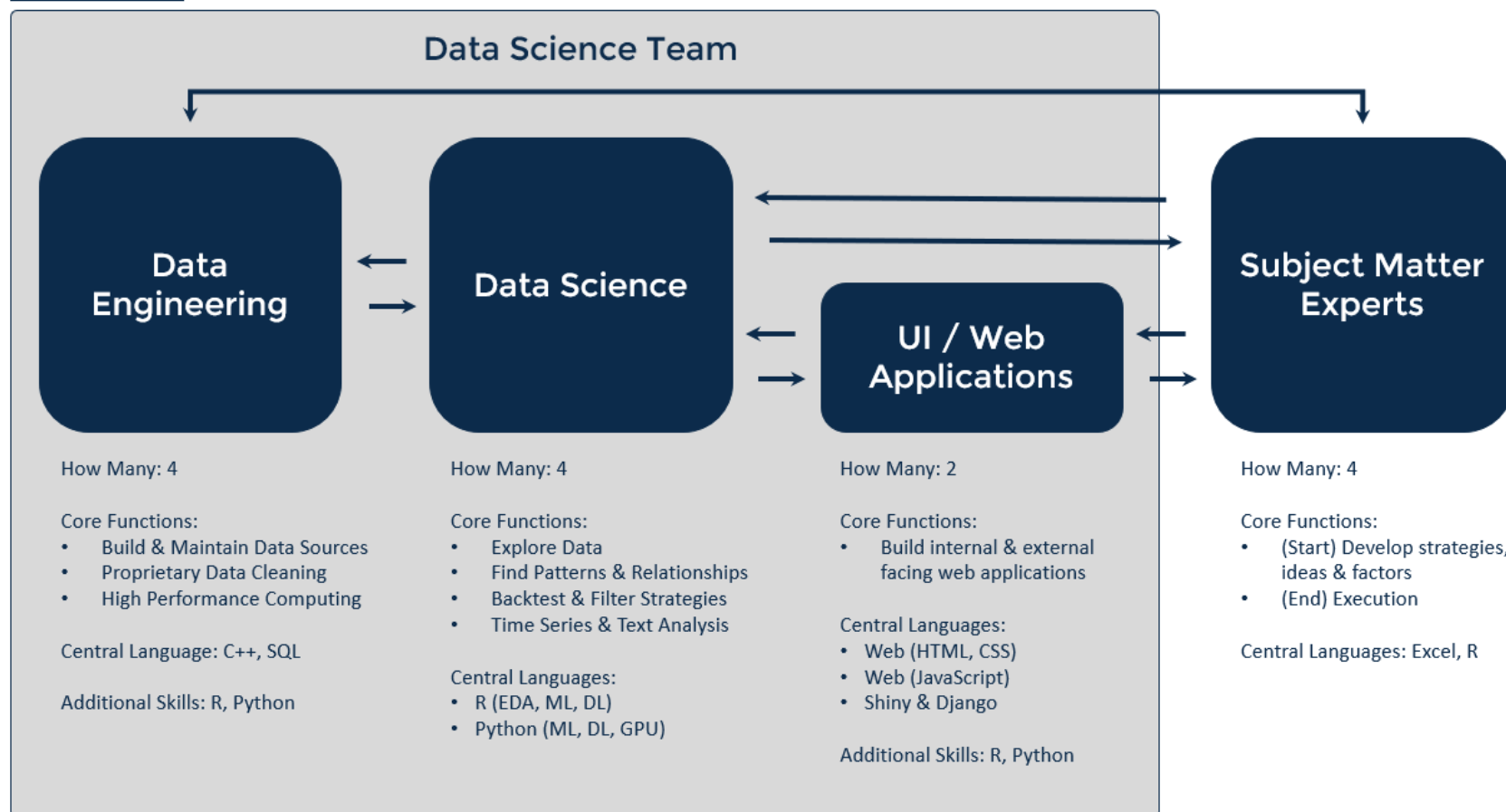
تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com گروه دایکه

High Performance Team Structure



□ مثال: ساختار تیم فنی

در حوزه دیتا

تولید محتوا: محمد روزبه

daychegroup

daychegroup

~ dayche.com | گروه دایکه

❑ موقعیت های شغلی در تیم علم داده

تحلیلگر داده

- تحلیل های توصیفی و مدل های پیش بینانه
- طرح مسئله، گزارش دهی و داشبوردهای مدیریتی

دانشمند داده

- ساخت و توسعه انواع مدل های تحلیلی پیشرفته
- تحلیل انواع داده ها ساخت یافته و غیر ساخت یافته

مهندس یادگیری ماشین

- اتوماتیک سازی مدل های یادگیری ماشین
- توسعه راهکارهای نرم افزاری

مهندس داده

- یکپارچه سازی زیرساخت داده ها
- ایجاد پایپ لاین داده

تولید محتوا: محمد روزبه

daychegroup 

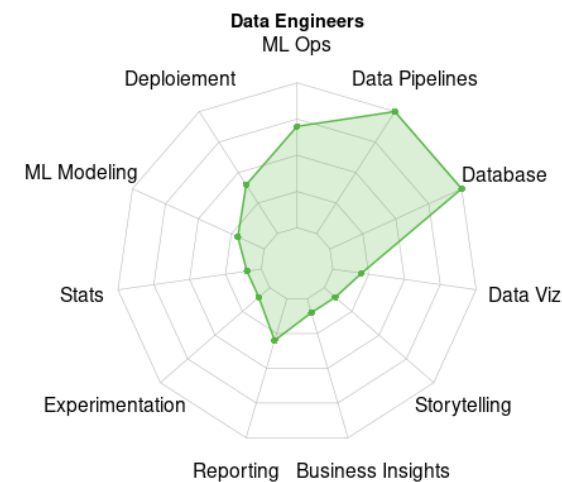
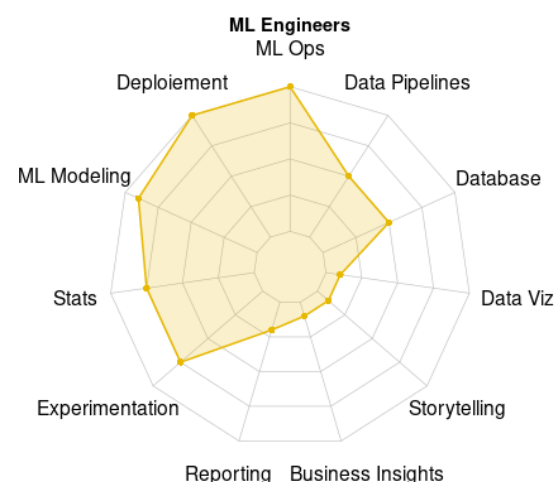
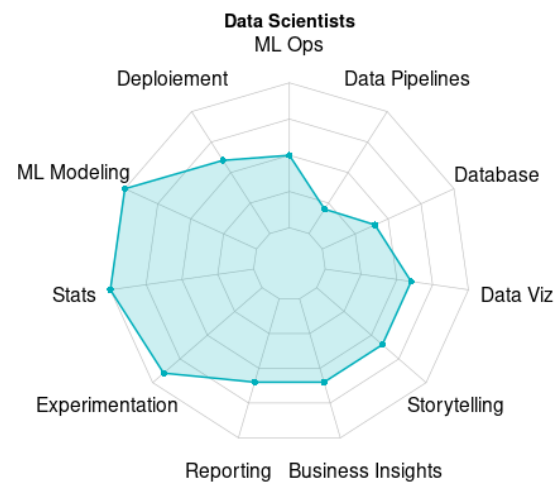
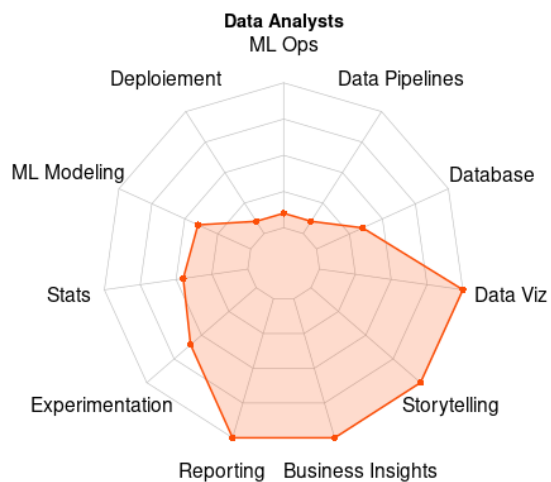
daychegroup 

dayche.com | گروه دایکه 

علم داده

مهارت های لازم

موقعیت های شغلی در تیم علم داده

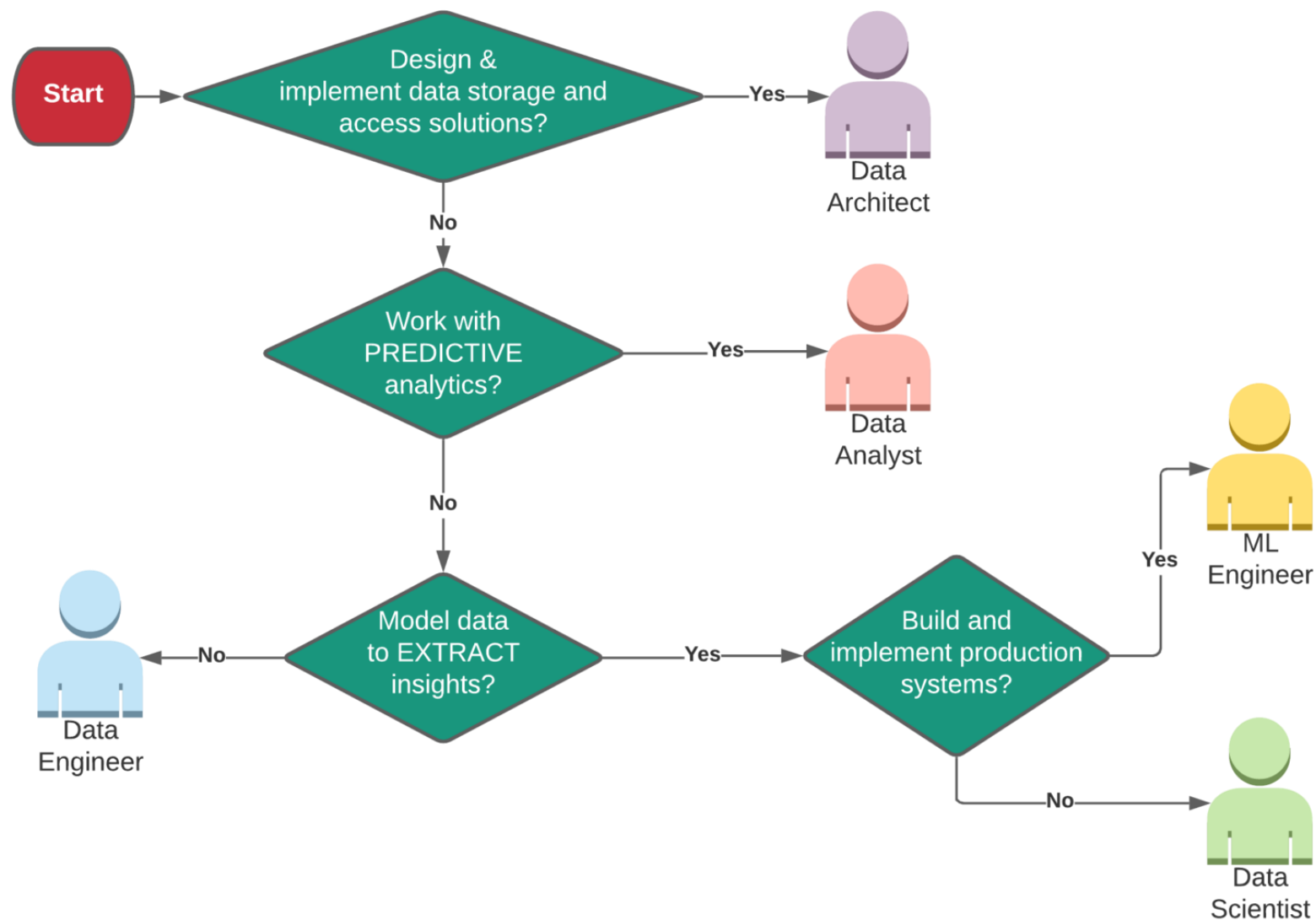


تولید محتوا: محمد روزبه

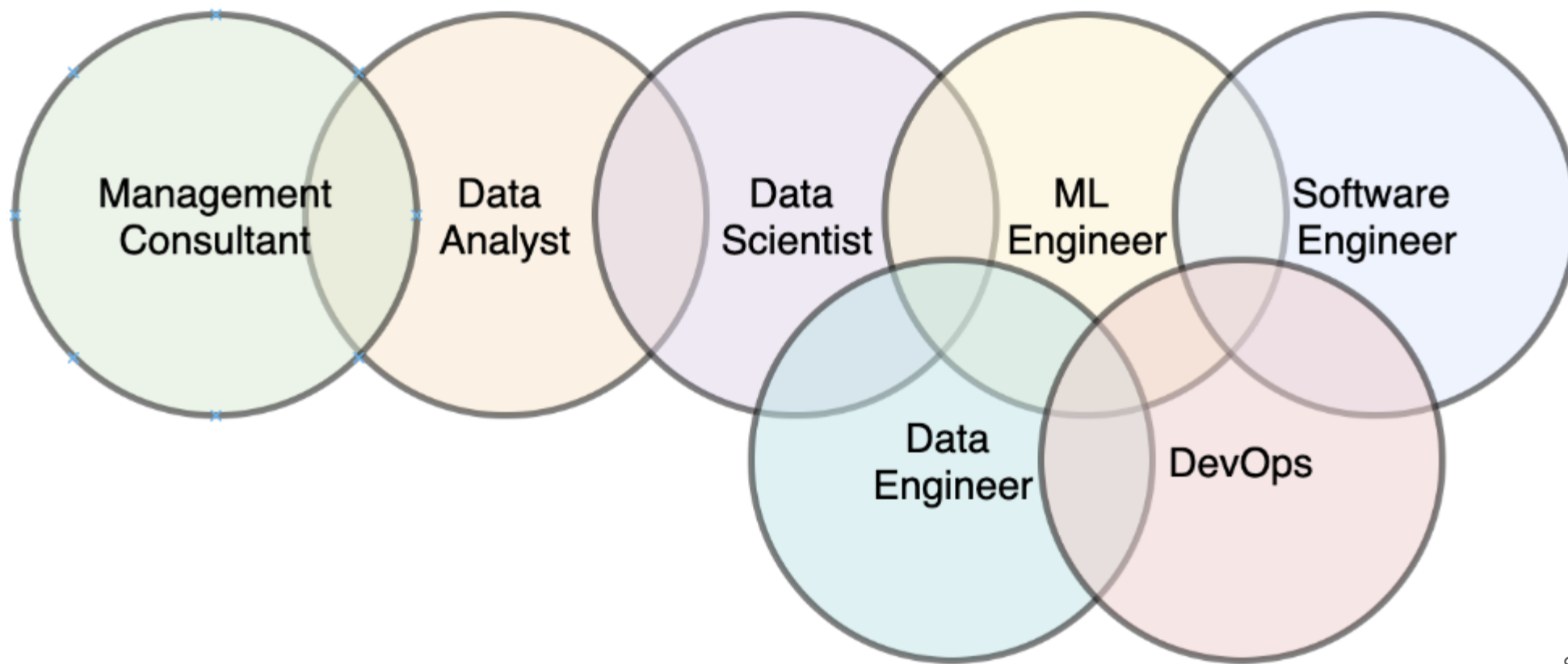
daychegroup

daychegroup

dayche.com | گروه دایکه



Business ← → IT

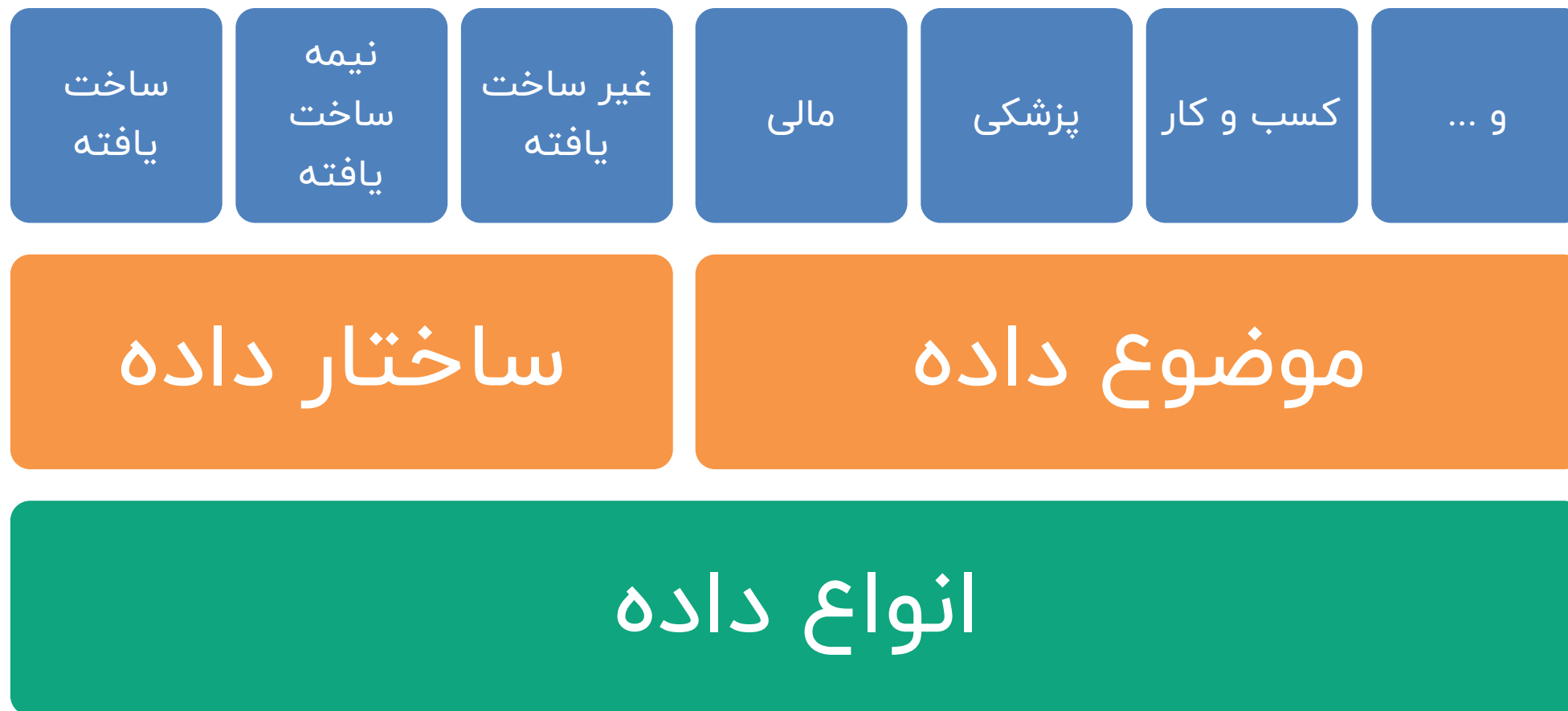


Data Science

انواع داده و رویکردهای تحلیلی

گروه دایکه . dayche.com





علم داده

انواع داده و رویکردهای تحلیلی

□ بر اساس موضوع داده

- هر کجا داده ای تولید شود، علم داده آنجا حضور دارد.
- کلیه مفاهیم، الگوریتم ها و روش های تحلیل داده در هر داده ای فارغ از موضوع محتوایی داده، بدون نیاز به تغییر یا با حداقل تغییرات قابل استفاده می باشد.



تولید محتوا: محمد روزبه

daychegroup 

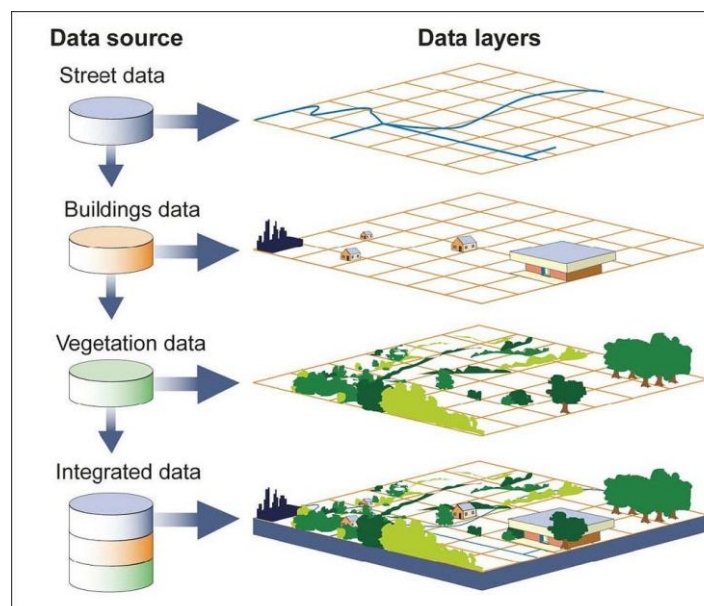
daychegroup 

dayche.com | گروه دایکه 

student_id	name	age
1	Akon	17
2	Bkon	18
3	Ckon	17
4	Dkon	18

subject_id	name	teacher
1	Java	Mr. J
2	C++	Miss C
3	C#	Mr. C Hash
4	Php	Mr. P H P

student_id	subject_id	marks
1	1	98
1	2	78
2	1	76
3	2	88



Source: GAO.

□ بر اساس ساختار داده

- داده های ساخت یافته
- جداول فلت مانند اکسل و فایل های CSV
- پایگاه داده های رابطه ای و ...
- پایگاه داده های جغرافیایی (GIS)

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

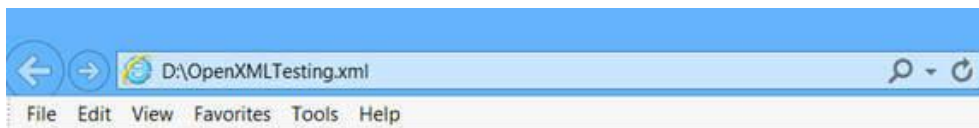
dayche.com | گروه دایکه 

□ بر اساس ساختار داده

○ داده های نیمه ساخت یافته

○ فایل های XML

○ فایل های JSON و ...



```
<?xml version="1.0"?>
- <ROOT>
  - <Customers>
    - <Customer CustomerName="Arshad Ali" CustomerID="C001">
      - <Orders>
        - <Order OrderDate="2012-07-04T00:00:00" OrderID="10248">
          <OrderDetail Quantity="5" ProductID="10"/>
          <OrderDetail Quantity="12" ProductID="11"/>
          <OrderDetail Quantity="10" ProductID="42"/>
        </Order>
      </Orders>
      <Address> Address line 1, 2, 3</Address>
    </Customer>
    - <Customer CustomerName="Paul Henriot" CustomerID="C002">
      - <Orders>
        - <Order OrderDate="2011-07-04T00:00:00" OrderID="10245">
          <OrderDetail Quantity="12" ProductID="11"/>
          <OrderDetail Quantity="10" ProductID="42"/>
        </Order>
      </Orders>
      <Address> Address line 5, 6, 7</Address>
    </Customer>
    - <Customer CustomerName="Carlos Gonzlez" CustomerID="C003">
      - <Orders>
        - <Order OrderDate="2012-08-16T00:00:00" OrderID="10283">
          <OrderDetail Quantity="3" ProductID="72"/>
        </Order>
      </Orders>
      <Address> Address line 1, 4, 5</Address>
    </Customer>
  </Customers>
</ROOT>
```

```
1 {
2   "users": [
3     {
4       "name": "John",
5       "age": 25
6     },
7     {
8       "name": "Mark",
9       "age": 29
10    },
11    {
12      "name": "Sarah",
13      "age": 22
14    }
15  ],
16  "dataTitle": "JSON Tutorial!",
17  "swiftVersion": 2.1
18 }
```

علم داده

انواع داده و رویکردهای تحلیلی

□ بر اساس ساختار داده

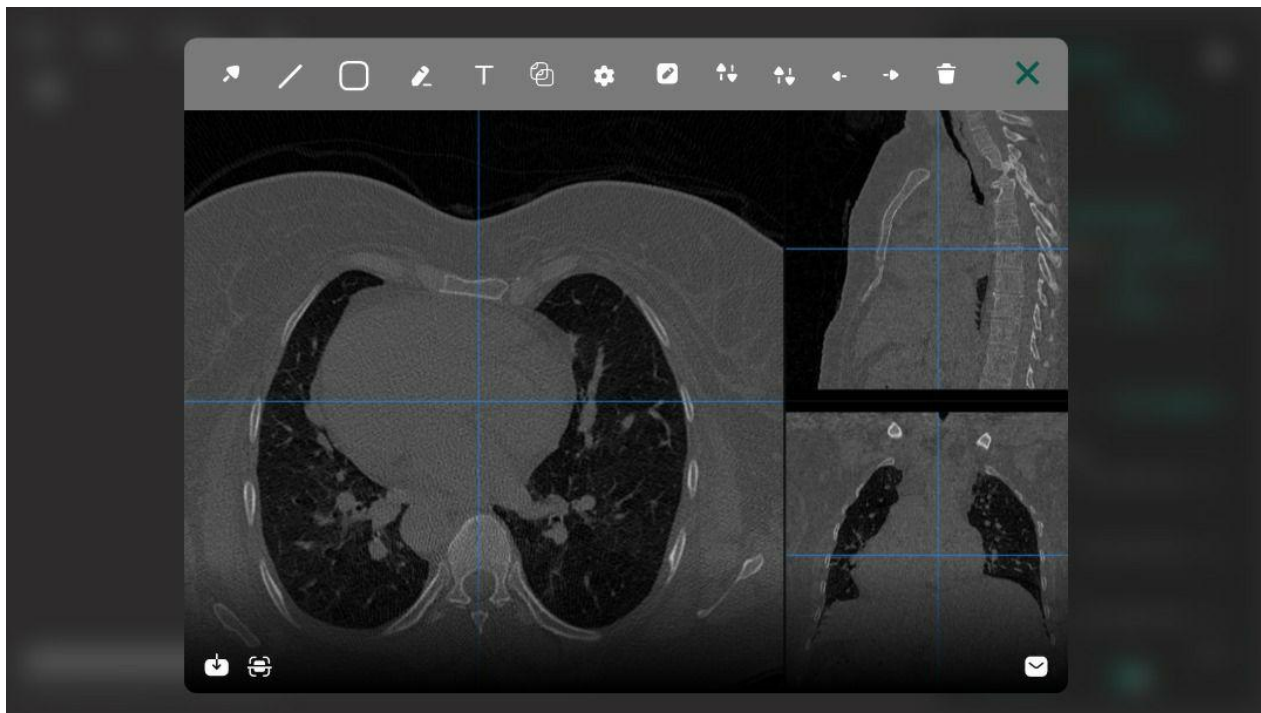
- داده های غیر ساخت یافته

- تصاویر

- متن

- سیگنال صوتی

- ویدئو و ...



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

□ بر اساس ساختار داده

- علم داده هیچ نوع محدودیتی در تحلیل داده با ساختارهای متنوع ندارد.

- کلیه مفاهیم، الگوریتم ها و روش های تحلیل داده، بنا به نوع ساختار ثبت داده قابل تعمیم و استفاده می باشد.

- مراحل آماده سازی، مدلسازی و ارزیابی در داده های غیر ساخت یافته دارای پیچیدگی های بیشتر بوده و نیازمند مهارت های جانبی افزون تری می باشد.

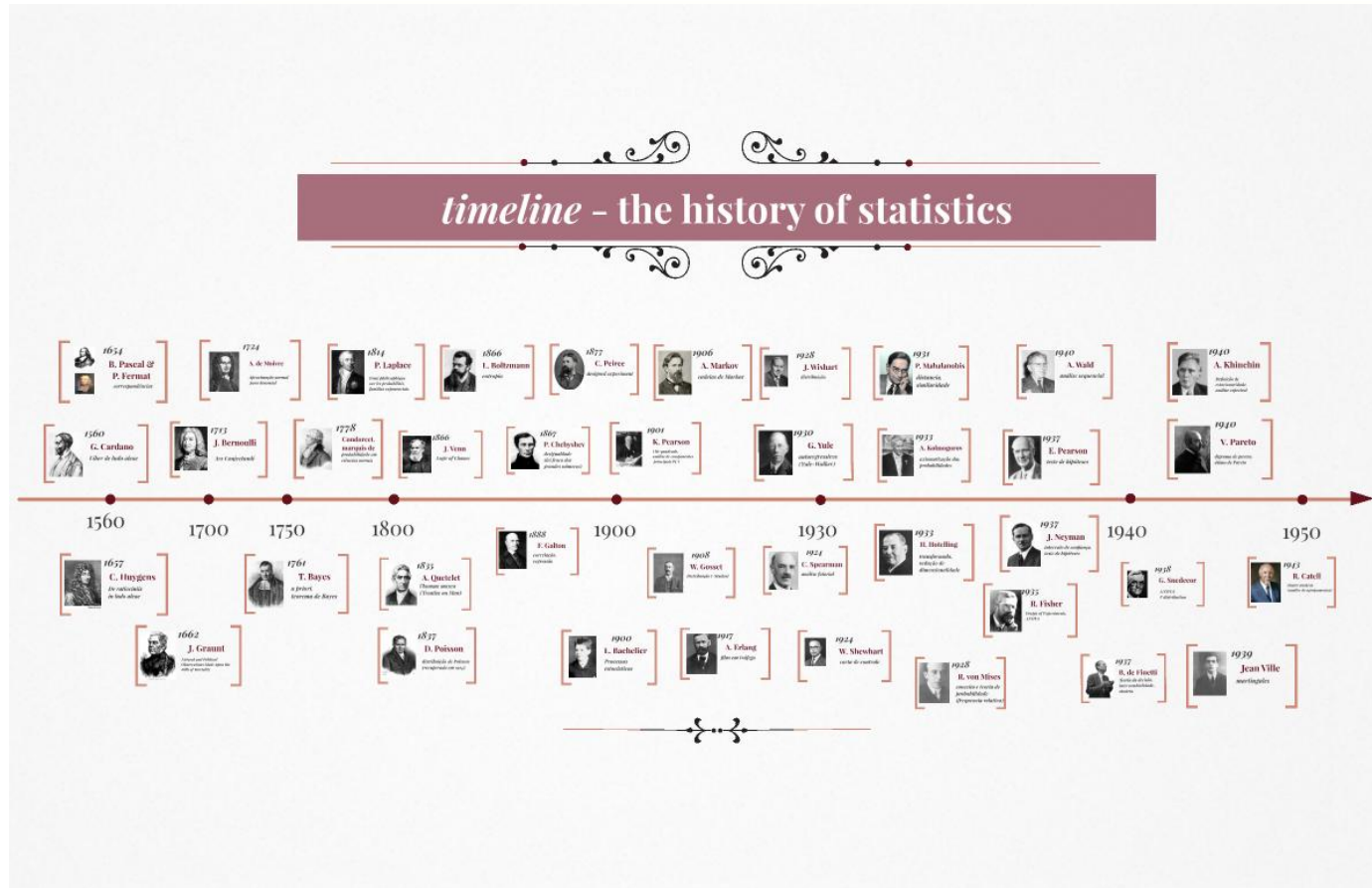
رویکردهای تحلیلی □

یادگیری
ماشین

آمار و
احتمال

آمار و احتمال

- احتمال، شاخه ای است از ریاضیات جهت مطالعه فضای عدم قطعیت.
- آمار، رشته ای است بر پایه نظریه احتمالات به منظور جمع آوری، تجزیه و تحلیل، تفسیر و ارائه اطلاعات مفید از داده ها.



آمار و احتمال

تئوری
احتمال

تعریف احتمال
احتمال شرطی
مفاهیم استقلال و سازگاری در
احتمالات
قانون جمع احتمال و بیز
تابع توزیع احتمال
قضیه حد مرکزی
و ...

متخصص
آمار
ریاضی

آمار

- ابزارهای تحلیل آماری:
- برآورد گرها (میانگین، واریانس، فاصله اطمینان و ...)
 - آماره ها (آماره t ، آماره کای دو، آماره پیرسون و ...)
 - انواع آزمون های فرض
 - مدل های آماری
 - و ...

تحلیلگران
داده

بیزینس:

- مالی
- پزشکی
- بازاریابی
- صنعت
- تکنولوژی
- و ...

01110001101
10001110111
00111011110
11011100001
10101111

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

کاردانو

گالیله

پیر فرما

بلز پاسکال



آمار و احتمال

- قرن 17: تئوری احتمالات
- علاقه مندی متفکران بزرگ به تحلیل شانس و تقلب در قمار
- انتشار کتاب بازی با تاس: از کاردانو و گالیله
- توسعه نظریه ترکیبات: نامه نگاری های پاسکال و فرما



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

آمار و احتمال

- قرن 18: آمار استنباطی
- معرفی قانون اعداد بزرگ: توسط برنولی
- معرفی قانون بیز: توسط تام بیز
- ایجاد پایه های مستحکم جهت تعمیم نتایج نمونه ها به جامعه نامعلوم بر اساس تئوری احتمالات

تولید محتوا: محمد روزبه

daychegroup 

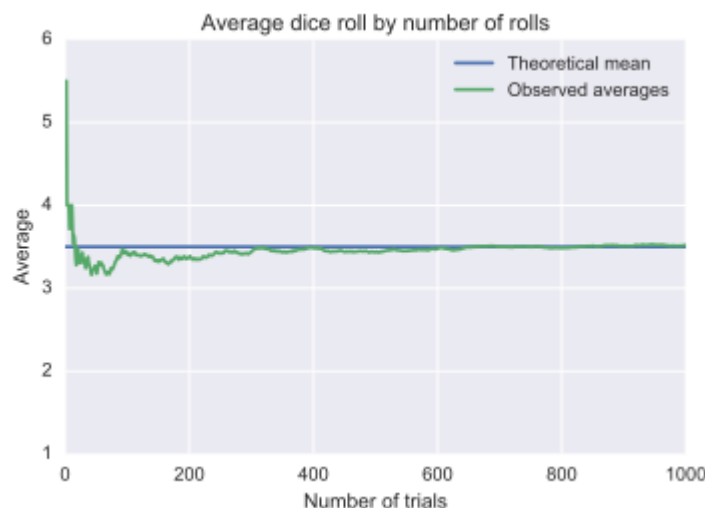
daychegroup 

dayche.com | گروه دایکه 

تام بیز



ژاکوب برنولی



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

□ آمار و احتمال

- قرن 19: گسترش کاربرد در حوزه های اجتماعی و علوم
 - توسعه قضایای مهم در احتمال توسط گوس و لاپلاس
 - بکارگیری نظریه احتمال در تحلیل علوم اجتماعی توسط کوئتل (پدر آمار نوین)
- بکارگیری آمار و احتمال در تحلیل پدیده های مقیاس بزرگ فیزیکی و بیولوژی
- معرفی ابزارهای مهم تحلیل آماری همچون همبستگی، رگرسیون، درستیابی و ...

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

گوس



لاپلاس



کوئتل



شانون

رونالد فیشر



□ آمار و احتمال

- قرن 20: مدل های پیش بینی و علوم کامپیوتر
- معرفی طراحی آزمایش، تحلیل واریانس و کوواریانس و نظریه مقدار حدی توسط رونالد فیشر
- معرفی و شروع نظریه اطلاعات توسط شانون (پدر نظریه اطلاعات)
- توسعه و بکارگیری ابزارهای تحلیل آماری متنوع همچون آزمون های فرض، روش های ناپارامتری و ...
- بکارگیری آمار در تحلیل داده های بزرگ و توسعه مدل های توصیفی و پیش بینی پیشرفته

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

□ آمار و احتمال

○ نقش آمار و احتمال در علم داده:

مدلسازی

ارتباط سنجی

خلاصه سازی

تئوری و مفاهیم احتمالات

□ آمار و احتمال

○ نقش آمار و احتمال در علم داده از دو جنبه مهم است:

روش های آماری

- شاخص های مرکزی و پراکندگی
- فاصله اطمینان
- آزمون های فرض
- تحلیل همبستگی و ...

نگرش آماری

- نهادینه شدن منطق آماری در تصمیم گیری
- تاکید بر تحلیل رفتار گروهی و ...

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

❑ محدودیت های آمار و احتمال

- صرفاً بر داده های کمی تاکید دارد.
- با فشردگی فکت ها روی رفتار گروهی از داده ها به تحلیل می پردازد نه روی رفتار یک داده منحصر بفرد.
- به علت استفاده از داده های کمی، امکان بررسی تمامی دلایل موثر را ندارد.
- به علت وجود روشهای متفاوت در محاسبه شاخص های آماری، نتایج یکسانی برای یک مسئله ثابت وجود ندارد.

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

هوش مصنوعی و یادگیری ماشین

- هوش مصنوعی: مجموعه تکنیک ها و روش هایی که سعی در ایجاد توانایی تقلید کامپیوترها از رفتار انسان دارد.
- سه مرحله در توسعه هوش مصنوعی:
 - هوش مصنوعی ضعیف
 - هوش مصنوعی قوی (جامع)
 - سوپر هوش مصنوعی



Narrow AI

Dedicated to assist with or take over specific tasks.



General AI

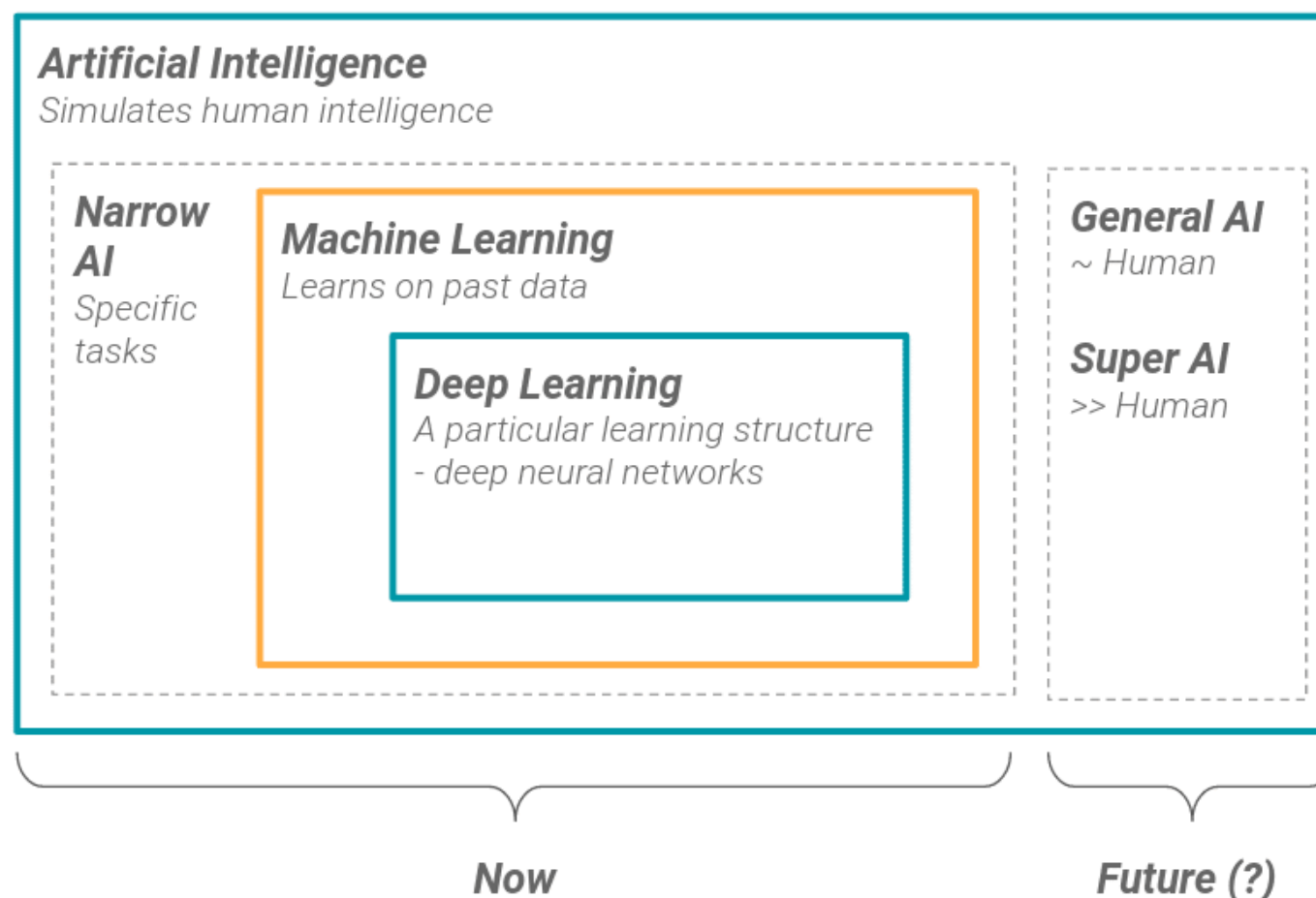
Takes knowledge from one domain, transfers to other domain.



Super AI

Machines that are an order of magnitude smarter than humans.

هوش مصنوعی و یادگیری ماشین



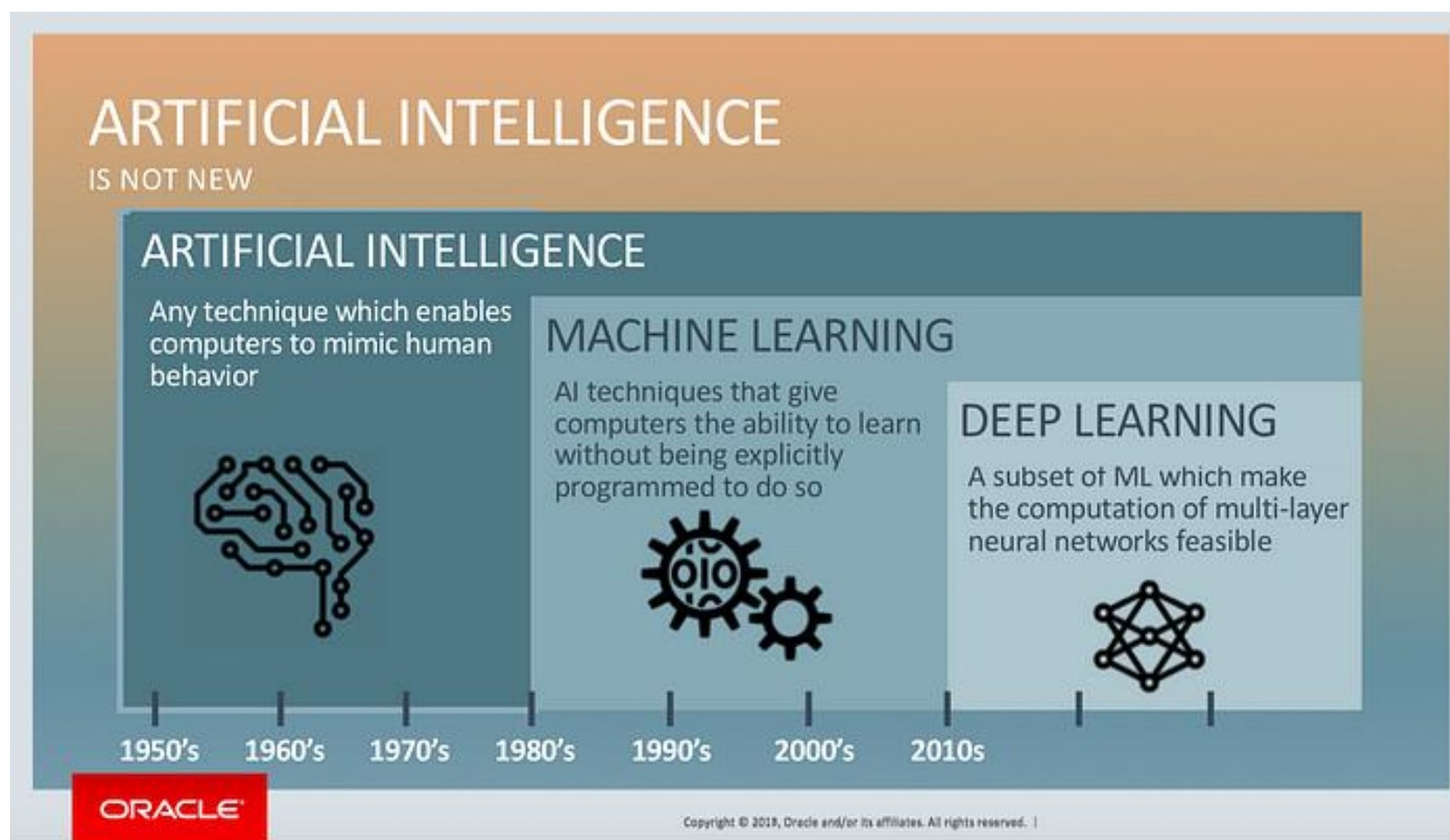
تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

هوش مصنوعی و یادگیری ماشین



دو رویکرد اساس در توسعه هوش مصنوعی:

○ رویکرد دانش - محور

○ رویکرد داده - محور

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

هوش مصنوعی و یادگیری ماشین

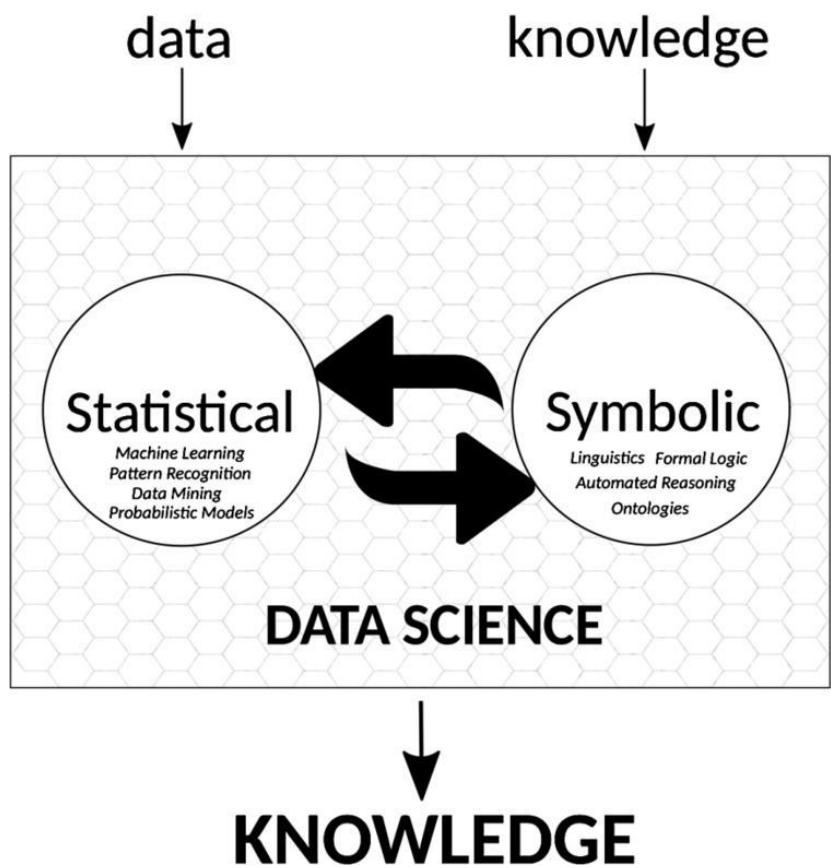
- 1956 الی 1973: هوش مصنوعی نمادین (عصر طلایی)
- حرکت به سوی هوش مصنوعی جامع: برای جایگزینی با هوش انسانی
- تاکید بر گزاره های منطقی و رویکردهای دانش محور و به حاشیه رفتن رویکردهای آماری داده محور
- معرفی اولین الگوریتم های شبکه عصبی (Perceptron) در زمینه ارتباط گرابی
- اختصاص بودجه های کلان توسط دارپا و دانشگاه هایی همچون MIT و استنفورد

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



هوش مصنوعی و یادگیری ماشین

1973 الی 1980: زمستان اول هوش مصنوعی

شکست پروژه ماشین ترجمه

شکست مدل های داده محور شبکه های عصبی

پرسپترون در زمینه ارتباط گرابی

شکست پروژه درک گفتار دانشگاه کارنگی ملون و دار

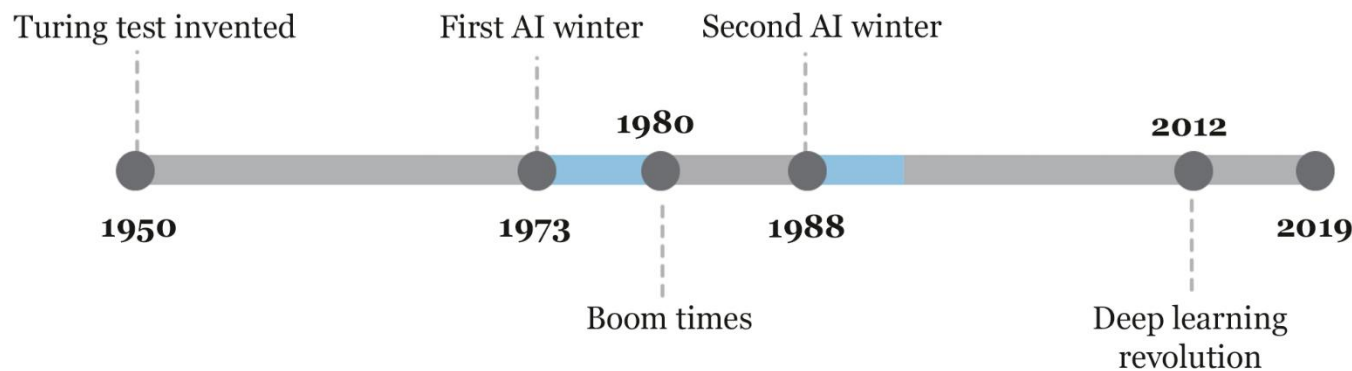
کاهش شدید و قطع بودجه پروژه های هوش مصنوعی

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



هوش مصنوعی و یادگیری ماشین

1980 الی 1988: دوره رونق دوباره

تاکید بر رویکردهای دانش محور "سیستم های خبره" در قالب گزاره های

اگر – آنگاه با نمادهای قابل درک توسط هوش انسانی

تخصیص بودجه کلان دولت ژاپن روی پروژه نسل پنجم کامپیوتر و

حرکت بسوی هوش مصنوعی

احیای ارتباط گرایی و امید به توسعه شبکه های عصبی نوین توسط

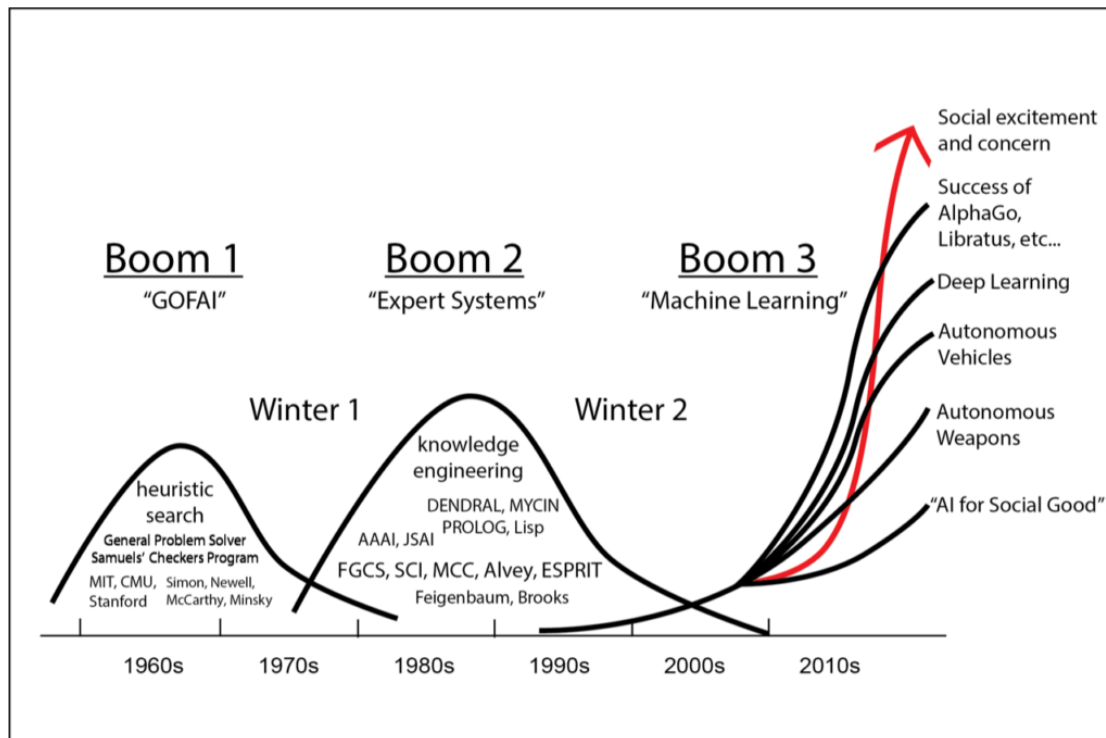
هاپفیلد و رومل هارت

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



هوش مصنوعی و یادگیری ماشین

○ 1988 الی 1993: زمستان دوم هوش مصنوعی

○ هزینه های گزاف نگهداری و بروزرسانی سیستم های خبره

○ صلاحیت سیستم های خبره در چند زمینه محدود

خاص بدون قابلیت یادگیری مداوم

○ شکست پروژه نسل پنجم کامپیوتر در ژاپن

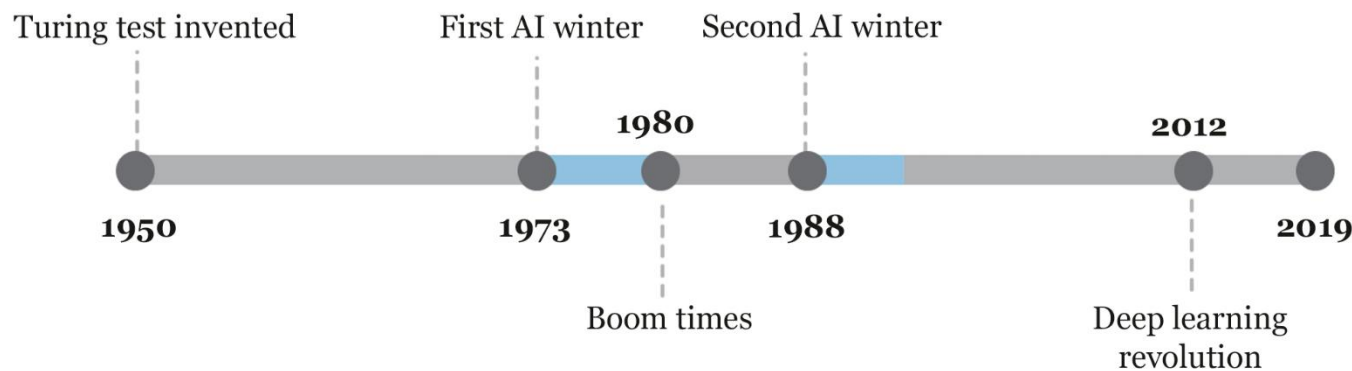
○ خروج سرمایه از بازار هوش مصنوعی

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



هوش مصنوعی و یادگیری ماشین

- 1993 الی 2011: هوش مصنوعی داده محور - یادگیری ماشین
- توسعه و بکارگیری مدل شبکه عصبی MLP و روش Backpropagation
- توسعه سایر مدل‌های داده محور همچون درخت تصمیم، SVM، مدل‌های جمعی
- ارتباط تنگاتنگ با آمار، جبر خطی و روش‌های بهینه‌سازی
- تغییر فضای هوش مصنوعی از رویکردهای دانش محور به رویکردهای داده محور
- تغییر نگرش از هوش مصنوعی جامع به هوش مصنوعی برای حل مسائل خاص

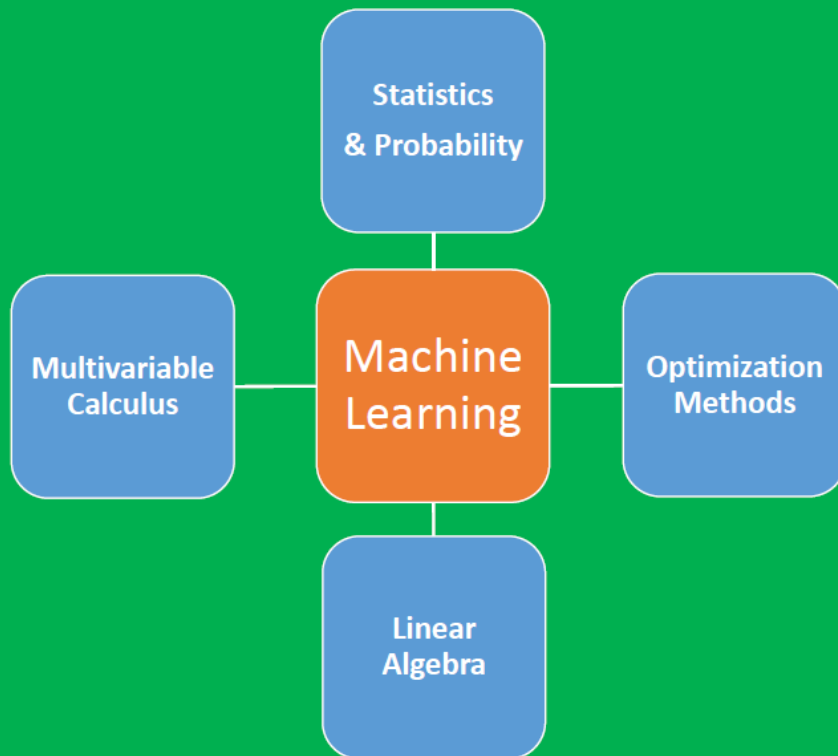
تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

4 Math Skills for Machine Learning



هوش مصنوعی و یادگیری ماشین

○ 1993 الی 2011: هوش مصنوعی داده محور - یادگیری ماشین

○ دستاورد های ویتربنی این دوره:

1993: اولین بازی کامپیوتری تخته نرد با استفاده از یادگیری شبکه های عصبی

1997: پیروزی هوش مصنوعی بر گری کاسپاروف در بازی شطرنج

2011: پیروزی ماشین هوش مصنوعی IBM Watson در بازی تلویزیونی Jeopardy

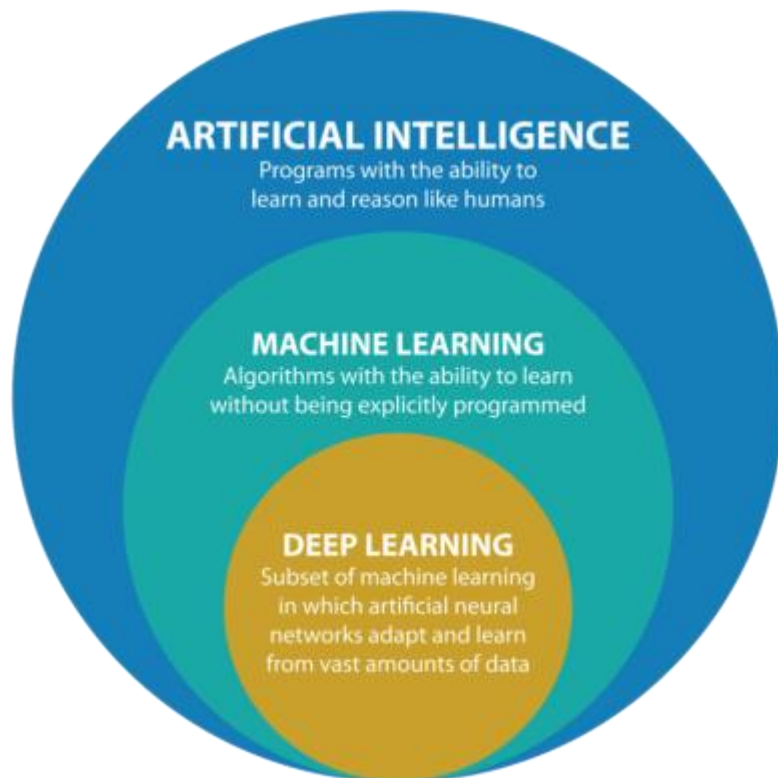
○ سوال بزرگ و بی پاسخ در این دوره:

یادگیری ماشین می تواند مانند انسان درکی از متن، تصویر، صدا و ... داشته باشد؟

Input	Label	Prediction
	CAT	
	NOT CAT	
	CAT	
		?

هوش مصنوعی و یادگیری ماشین

- 2011 تا الان: عصر جدید – یادگیری عمیق
- رشد اینترنت و شبکه های اجتماعی: مرجع گسترده از انواع داده های قابل دسترس
- رشد صنعت بازی: سرمایه گذاری و توسعه روی پردازنده های گرافیکی و ارزان شدن تکنولوژی
- امکان پیاده سازی ساختارهای پیچیده از شبکه های عصبی چند لایه تحت عنوان یادگیری عمیق به عنوان شاخه ای قدرتمند از یادگیری ماشین



هوش مصنوعی و یادگیری ماشین

- 2011 تا الان: عصر جدید – یادگیری عمیق
 - دستاوردهای ویتربنی این دوره:
- 2012: توسعه مدل شبکه عصبی عمیق برای شناسایی تصویر گربه در تصاویر بدون برچسب Youtube توسط تیم گوگل به سرپرستی Andrew Ng
- 2015: پیروزی 5 بر 0 هوش مصنوعی AlphaGo_Fan در مقابل قهرمان بازی GO با زیرساخت 176 پردازنده GPU بصورت توزیع شده
- 2016: پیروزی 4 بر 1 هوش مصنوعی AlphaGo_Lee در مقابل قهرمان بازی GO با زیرساخت 48 پردازنده TPU بصورت توزیع شده
- May, 2017: پیروزی 60 بر 0 هوش مصنوعی AlphaGo_Master در مقابل جمعی از قهرمانان بازی Go با زیرساخت 4 پردازنده TPU روی یک ماشین



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

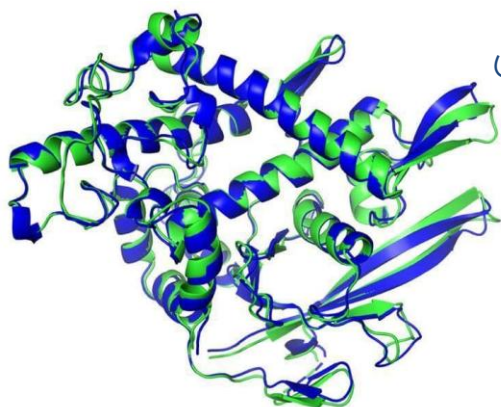
~ dayche.com | گروه دایکه

هوش مصنوعی و یادگیری ماشین

○ 2011 تا الان: عصر جدید – یادگیری عمیق

○ دستاوردهای ویتربنی این دوره:

- Oct, 2017: پیروزی 100 بر 0 هوش مصنوعی AlphaGo Zero در مقابل AlphaGo_Lee با زیرساخت 4 پردازنده TPU روی یک ماشین طی 3 روز
- Oct, 2017: پیروزی 89 بر 11 هوش مصنوعی AlphaGo Zero در مقابل AlphaGo_Master با زیرساخت 4 پردازنده TPU روی یک ماشین طی 21 روز
- Dec, 2017: معرفی الگوریتم AlphaZero که طی 24 ساعت می تواند به بالاترین سطح توانمندی قهرمانان جهان در بازی های Go، شطرنج و ... برسد.
- 2019: پیروزی اولین هوش مصنوعی در بازی پوکر توسط تیم تحقیقاتی دانشگاه کارنگی ملون و فیسبوک
- 2020: حل مسئله 50 ساله زیست شناسی با مدل پیش بینی و شبیه سازی ساختار 3 بعدی تاشدگی پروتئین
- 2022: معرفی ChatGPT و عمومیت یافتن سریع آن به عنوان یک دستاورد ملموس هوش مصنوعی



❑ محدودیت های یادگیری ماشین

- معمولا نیاز به حجم زیادی از داده های مناسب می باشد.
- هر مدل ساخته شده صرفا برای حل یک مسئله مشخص کارایی دارد و خارج از دامنه داده های آموزش داده شده قابل استفاده نیست.
- حریم خصوصی داده و محدودیت در دسترسی به داده های مناسب
- انتخاب نادرست نوع مسئله و الگوریتم مناسب توسط تحلیلگر

تولید محتوا: محمد روزبه

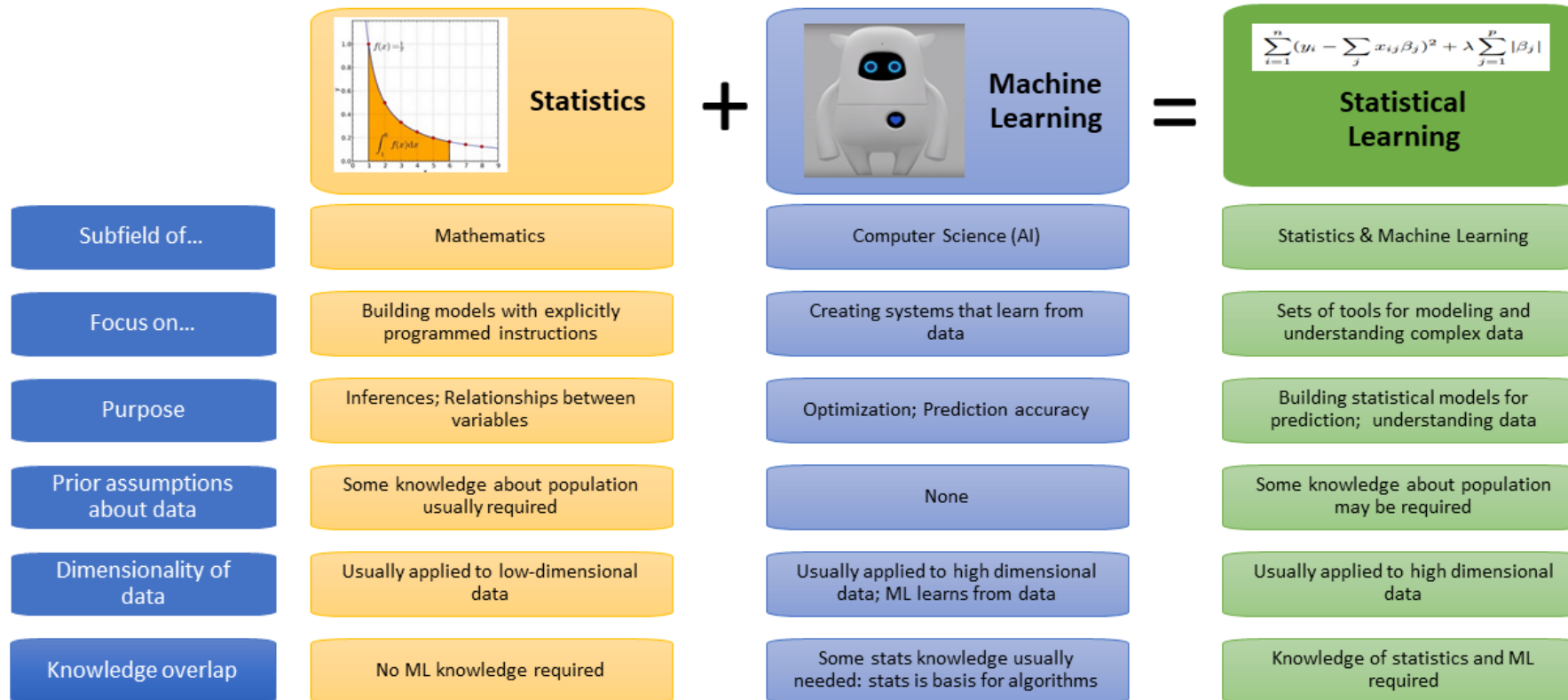
daychegroup 

daychegroup 

dayche.com | گروه دایکه 

آمار و یادگیری ماشین ← یادگیری آماری

○ یادگیری آماری: چارچوبی از یادگیری ماشین است که بر پایه آمار و آنالیز تابعی ریاضیات استوار است.



Musio image: Akawikipic [CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0>)]

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

~ dayche.com | گروه دایکه

Data Science

مسیر جامع علم داده دایکه

گروه دایکه . dayche.com



□ انواع منابع و روش های آموزشی

Youtube ○

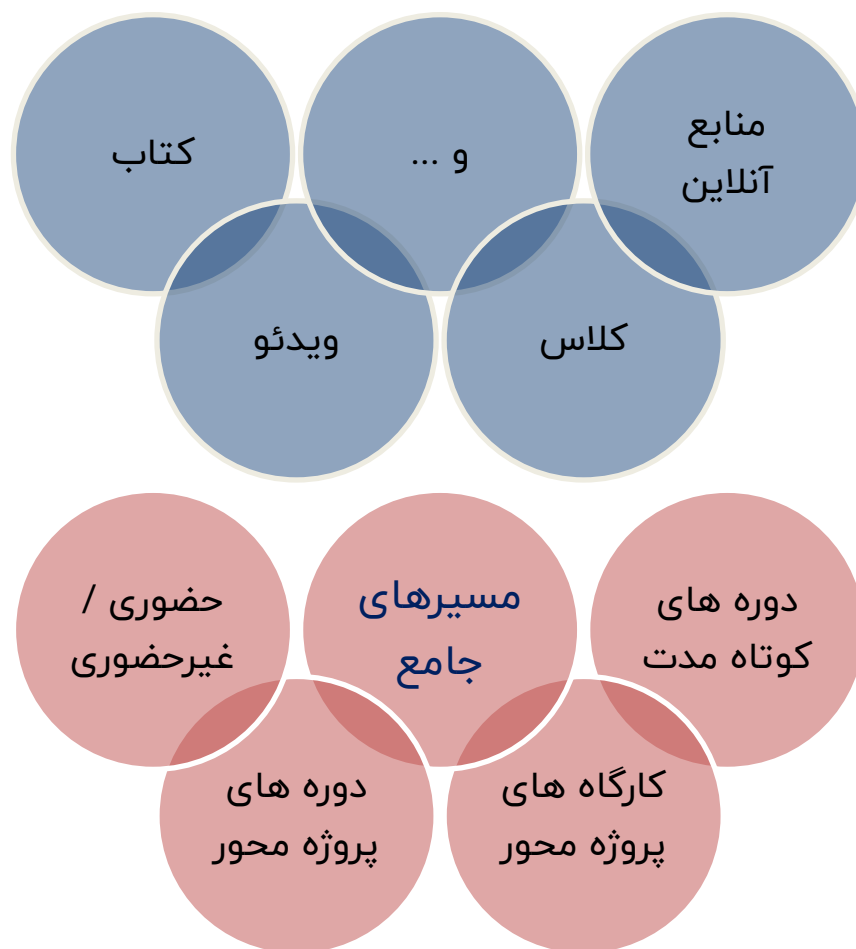
Corsera / Edx / ... ○

Wikipedia / Medium / Towards Data Science / ... ○

ChatGPT / Bard / ... ○

مکتبخونه / فرادرس / ... ○

انواع کلاس ها و کارگاه های آموزشی متنوع ○



Dayche Data Science Career
مسیر جامع علم داده دایکه

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

توسعه آموزش در دایکه



شروع فعالیت دایکه
همکاری با دکتر نخعی زاده
سال 1387

دوره های مهارتی
فقط حضوری
سال 1391

DA / DS

معرفی مسیرهای
آموزشی هدفمند

حضوری - غیرحضوری
سال 1397

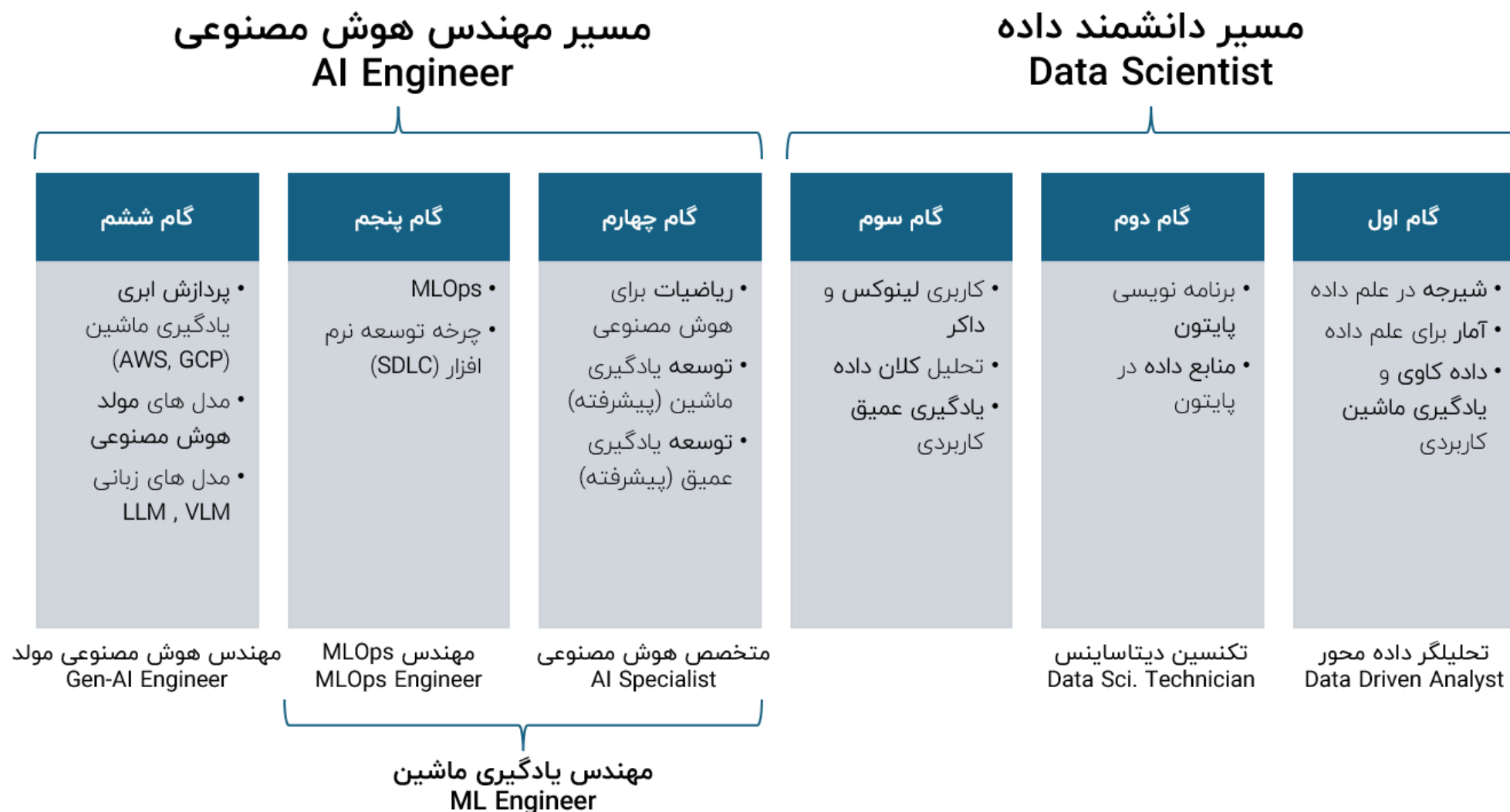
بازنگری و بروزرسانی در محتوا:
توسعه آموزش غیرحضوری
در بستر سرویس آنلاین
سال 1400

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مسیر دانشمند داده

- هنر حل مساله به صورت داده محور.
- توانایی تحلیل داده‌ها و کشف الگوهای پنهان.
- درک عمیق از یادگیری ماشین و هوش مصنوعی و بکارگیری آن.
- مهارت در برنامه‌نویسی و ابزارهای حرفه ای علم داده.
- توانایی ارتباط مؤثر برای ارائه نتایج به تیم‌های مختلف.

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

فرآیند داده کاوی

مدلسازی و ارزیابی

مدل های اکتشافی

مدل های پیش بینانه

روش های ارزیابی

قوانین انجمنی

انواع خوشه بندی

ترکیب مدل ها

داده های نامتوازن

روش های ارزیابی

انواع الگوریتم ها

شناخت و آماده سازی داده ها

یکپارچه سازی داده ها

کاهش ابعاد و انتخاب نمونه

تبدیل داده ها

توصیف و کاوش داده ها

نمونه گیری

استخراج ویژگی

انتخاب ویژگی

هموار سازی

تجمیع و فشرده سازی

گسسته سازی

شاخص سازی

نرمالسازی

گام 1: تحلیلگر داده محور

- تسلط بر نقشه راه پیشبرد پروژه های تحلیلی دیتاساینس و چالش های آن.
- آشنایی با تحلیل آماری داده، آماده سازی و ساخت مدل های یادگیری ماشین در چارچوب متدولوژی CRISP-DM.
- شیرجه در برنامه نویسی پایتون و بکارگیری کتابخانه های تخصصی علم داده در حل و پیاده سازی مسائل دیتاساینس.

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



ChatGPT + kaggle

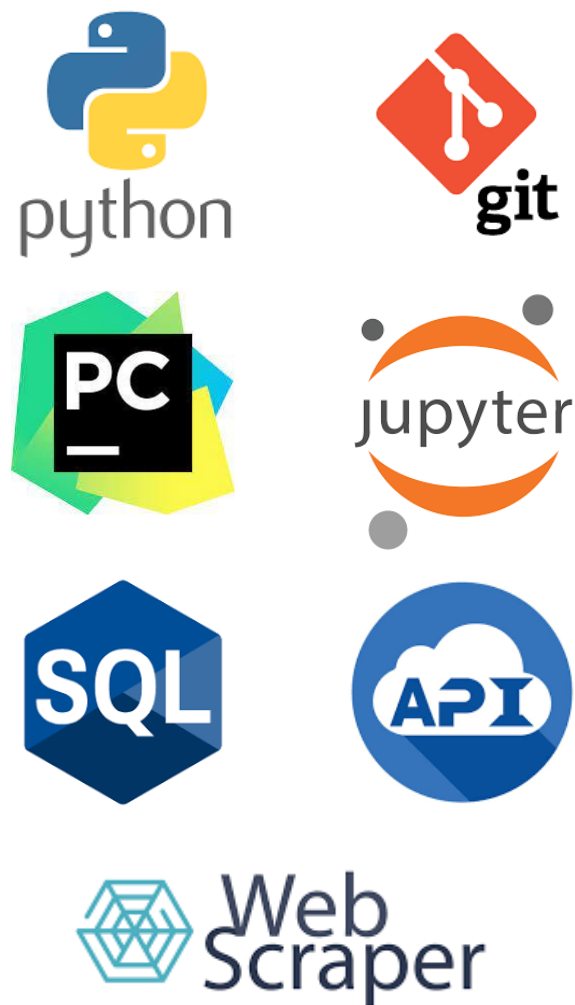
پروژه پایانی
A/B Test



Statistics
& Probability

پروژه تحلیل رفتار
مشتریان (سازمان
ایرانی)





□ گام 2: تکنسین دیتاساینس

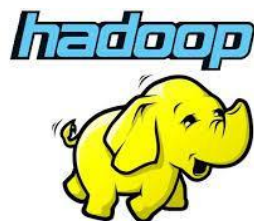
- مهارت برنامه نویسی مازولار و آشنایی با شی گزاری در پایتون.
- قابلیت استفاده از ابزارهای مدیریت ورژن کد Git و کار تیمی در Github.
- قابلیت استفاده از پایگاه داده مبتنی بر SQL، خواندن و نوشتن API و جمع آوری داده از وب Web Crawling.

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ گام ۳: ابزار و محیط حرفه ای

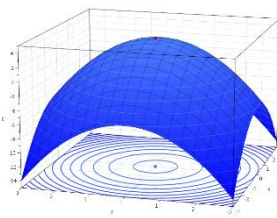
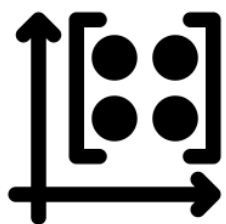
- آشنایی با سیستم عامل لینوکس و داکر برای محیط های داده محور.
- قابلیت استفاده از مهارت های تحلیل داده روی زیرساخت Big Data.
- حل مسایل پیچیده و تحلیل داده های متنی و تصویری با استفاده از الگوریتم های مدرن یادگیری عمیق.

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



DVC



□ گام 4: متخصص هوش مصنوعی (اختیاری)

- توانایی توسعه مدل های هوش مصنوعی با رویکرد عملیاتی.
- توانایی حرفه ای در استفاده از ابزارهای مدیریت ورژن کد و دیپتا Git & DVC.

تولید محتوا: محمد روزبه

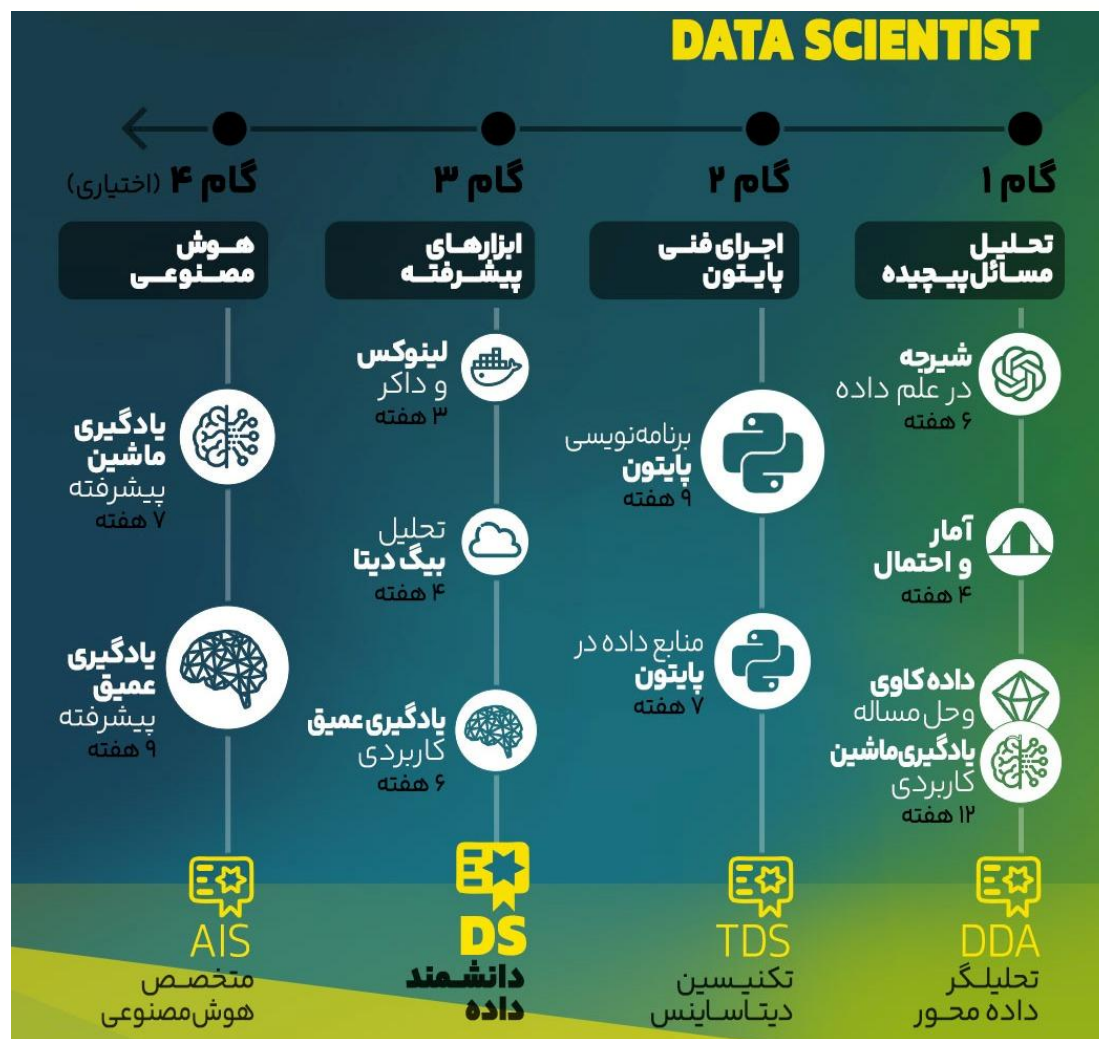
daychegroup

daychegroup

dayche.com | گروه دایکه

علم داده

مسیر جامع علم داده دایکه



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

□ چگونه مطالعه کنیم؟

فرآیند یادگیری موثر و سریع

Fast Learning Process

هدف از یادگیری، کاربرد در چه سطحی هست؟

چطور یک موضوع جدید را سریع یاد بگیریم؟

چطور تعیین اولویت و زمانبندی کنیم؟

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

چگونه مطالعه کنیم؟ □

سطح 1: آشنایی با **فرآیند** پیاده سازی علم داده

- تمرکز بر نقشه راه مسیر / اشراف کلی بر جانمایی ابزارها و مهارت ها در حل مسئله / تمرکز بر مفاهیم کلیدی و کاربردها بدون ورود به جزئیات
- مناسب برای: مدیران محصول، صاحبان کسب و کار و Citizen DS

سطح 2: یادگیری **چگونگی** پیاده سازی علم داده

- تمرکز بر اجرای عملی پروژه ها و تسلط بر ابزارهای پیاده سازی
- مناسب برای: Junior Data Scientist

سطح 3: فهم **عمیق** و **توسعه** الگوریتم ها

- یادگیری مفاهیم تئوری و ریاضیات الگوریتم ها و ابزارها
- مناسب برای: پژوهشگران داده، Senior Data Scientist

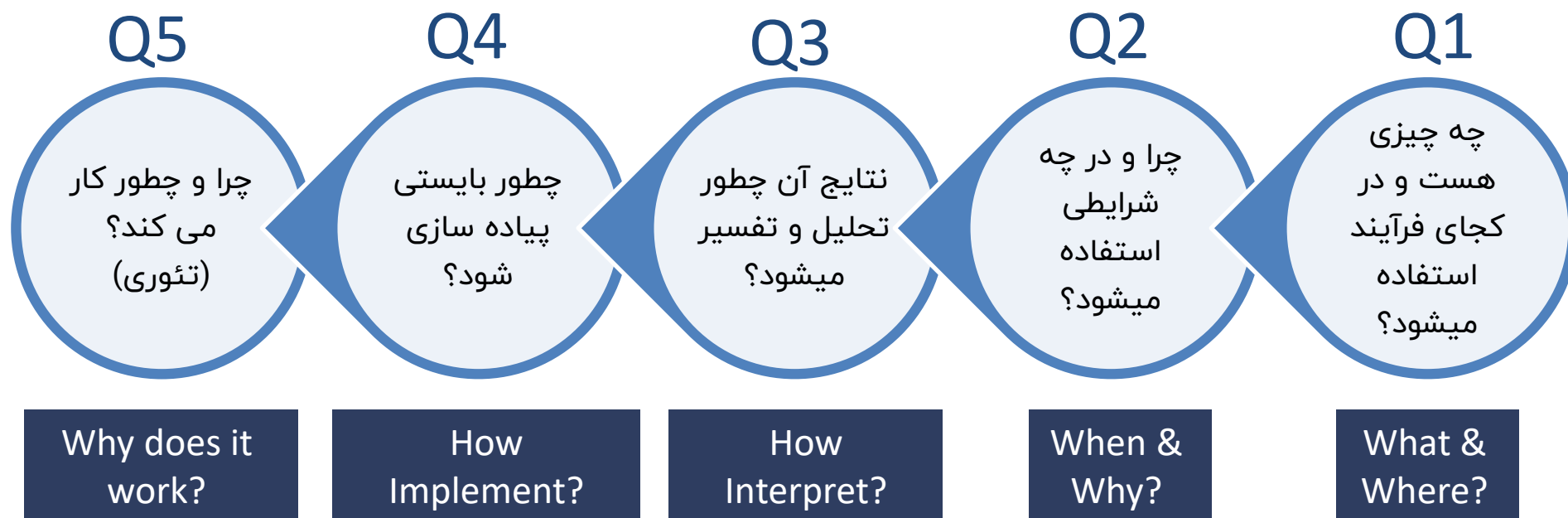
3
فهم
عمیق
جزئیات

2
یادگیری **چگونگی** پیاده
سازی علم داده

1
یادگیری **فرآیند** پیاده سازی علم داده

چطور یک موضوع جدید را سریع
یاد بگیریم؟

□ چگونه مطالعه کنیم؟



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

□ چگونه مطالعه کنیم؟

مثال: آزمون فرض t مستقل

- Q1 چه چیزی هست و در کجای فرآیند استفاده میشود؟
آزمونی آماری برای مقایسه میانگین دو گروه مستقل. در مرحله شناخت و تحلیل اولیه داده ها و ارزیابی مدلها مورد استفاده قرار می گیرد...
- Q2 چرا و در چه شرایطی استفاده میشود؟
جهت بررسی معنادار بودن اختلاف در دو گروه مستقل و تحت فرضیاتی مانند نرمال بودن، همگنی واریانس و ...
- Q3 نتایج آن چطور تحلیل و تفسیر میشود؟
محاسبه آماره t و همچنین محاسبه P -value. بزرگ بودن t و یا کوچکتر بودن P -value میتواند نشان از معناداری اختلاف باشد و
- Q4 چطور بایستی پیاده سازی شود؟
برای اجرای آزمون t مستقل در پایتون می توان از تابع `ttest_ind` در کتابخانه `SciPy` استفاده کرد و
- Q5 چرا و چطور کار می کند؟ (تئوری)
بر اساس توزیع احتمالی t -student تفاوت میانگین ها را نسبت به خطای استاندارد مقایسه میکند و

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

چطور تعیین اولویت و زمانبندی کنیم؟

□ چگونه مطالعه کنیم؟

سوال: محتوای هفتگی دایکه رو چطور زمانبندی و مطالعه کنیم؟

۱. زمانبندی محتوا

ویدئوهای هفتگی: ۳/۵ تا ۴ ساعت (متوسط)

کلاس های آنلاین: ۱/۵ تا ۲ ساعت + پروژه های هفتگی

زمان لازم برای مطالعه/پیشرفت: ۱۰ تا ۱۲ ساعت در هفته

۳. مدیریت زمان پیشنهادی

حداکثر ۳ تا ۴ ساعت در هفته:

مطالعه ویدئوهای تئوری

زمان باقیمانده (۶ تا ۸ ساعت):

مطالعه ویدئوهای عملی

انجام تمرینها و پروژههای هفتگی

۲. تقسیم محتوا (کورس های کلیدی آمار + داده کاوی)

۵۰٪ تئوری: مفاهیم پایه + جزئیات تئوریک

۵۰٪ عملی: پیاده سازی + تحلیل و تفسیر نتایج

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

Data Science

روش های یادگیری و انواع مسئله ها

گروه دایکه . dayche.com



❑ شناسایی الگوها – مثال (1)

- فرض کنید داده های خرید مشتریان شرکت X را داریم و به دنبال الگوهای خرید آنها هستیم.

CID	Date	Item	Price
1	99/01/07	z	100
2	99/01/07	a	50
3	99/02/15	a	50
1	99/02/25	d	200
2	99/03/05	s	20
3	99/03/17	z	100
2	99/03/22	d	200
2	99/03/28	s	20
2	99/04/07	b	150
3	99/04/07	s	20
.	.	.	.
.	.	.	.

تولید محتوا: محمد روزبه

daychegroup 

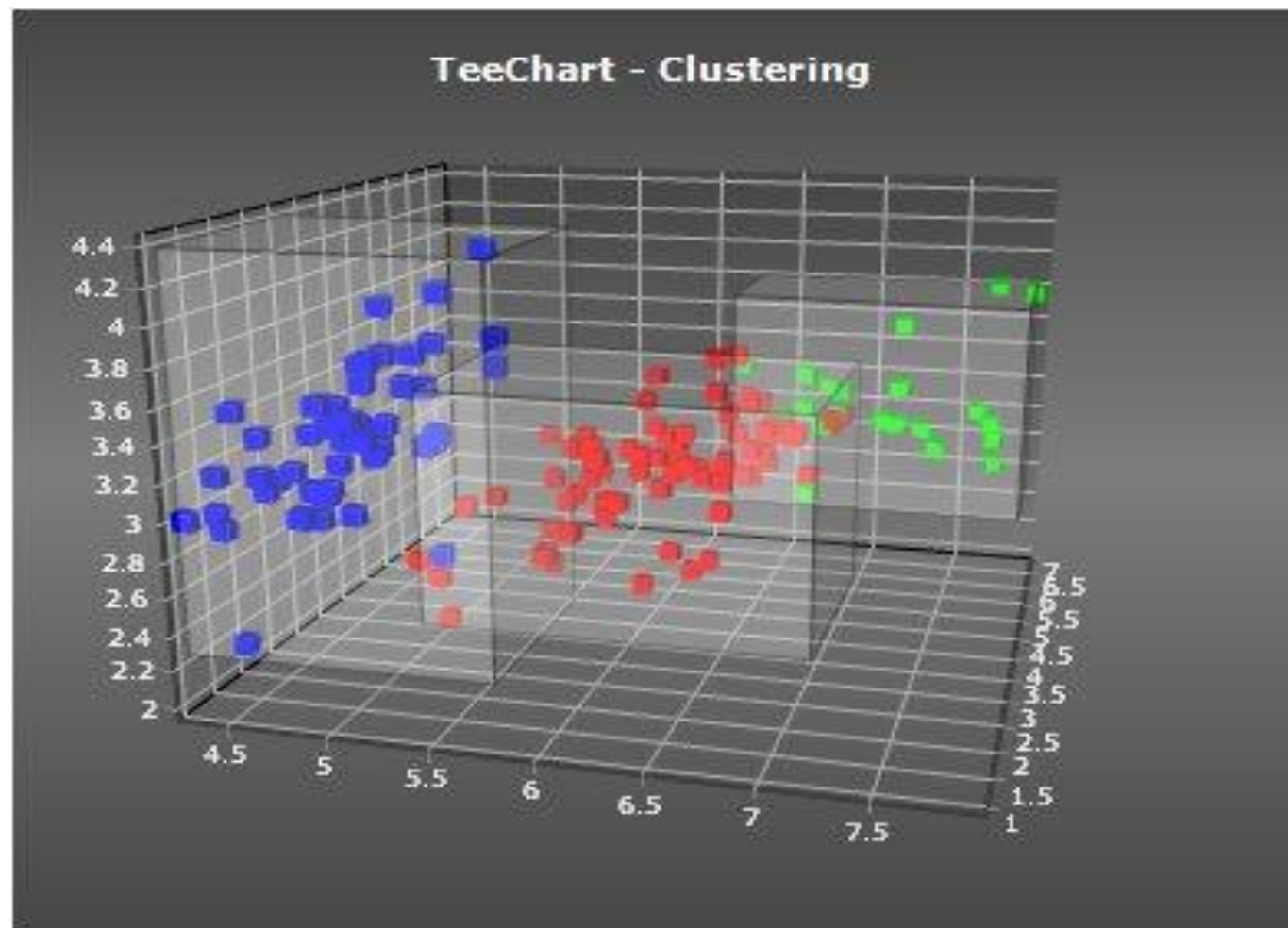
daychegroup 

dayche.com | گروه دایکه 

□ شناسایی الگوها – مثال (1)

○ ساخت شاخص های رفتاری

○ شناسایی الگو



CID	Recency	Frequency	Monetary
1	10	28	1850
2	3	35	750
3	35	3	170
4	12	18	400
5	5	1	20
6	45	1	1000
7	120	24	1600
8	17	22	580
9	155	18	1360
10	24	5	120
.	.	.	.
.	.	.	.

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

شناسایی الگوها – مثال (1)

- برچسب گذاری الگوی رفتاری هر مشتری (مشتریان با الگوی 9 را پیدا کنید!)

3 3 0 3 0 1 8 7 6 8 2 1 4 0 3 8 3 7 7 2 0 5 2 3 2 7 0 2 0
 7 1 4 6 0 2 1 3 2 7 6 0 2 5 6 3 2 5 7 6 3 3 0 2 0 3 0 7 2
 8 7 5 7 2 8 3 8 7 7 8 2 0 7 7 5 2 3 1 1 5 6 3 8 4 7 8 2 0
 0 5 0 5 1 6 1 7 5 6 8 0 4 4 6 7 4 7 1 4 0 0 8 4 4 3 0 3 2
 2 4 3 1 3 5 4 9 5 0 7 6 0 7 4 3 1 8 2 7 3 4 6 0 2 4 8 2 3
 8 6 2 2 6 5 4 6 7 0 7 6 0 0 3 9 0 2 4 7 1 7 2 3 3 5 8 7 0
 0 8 4 5 1 3 1 7 6 4 5 4 1 2 4 5 3 3 5 4 9 6 7 7 6 3 4 2 5
 4 7 7 0 2 2 0 1 1 7 7 7 0 2 6 6 4 7 5 8 6 1 4 3 7 8 5 4 6
 4 3 6 6 4 6 6 2 8 4 8 5 3 7 8 8 1 3 8 5 4 5 7 4 0 3 2 8 4
 5 5 0 3 5 3 5 3 8 3 2 3 8 2 3 1 6 2 7 2 4 6 3 6 4 4 3 2 5
 4 4 0 2 1 7 2 4 4 7 4 1 9 2 4 5 2 5 0 4 0 0 5 3 6 3 3 6 7
 7 4 6 6 8 7 5 7 9 2 0 2 8 8 8 8 3 2 4 2 6 4 0 4 6 3 7 2 1
 0 1 7 1 5 9 1 4 2 8 7 3 7 1 4 5 1 8 7 8 0 5 1 7 0 5 8 8 1
 2 8 5 2 1 2 8 7 7 6 2 5 6 2 6 4 1 5 1 6 1 2 1 1 0 5 6 4 0
 2 1 1 7 7 2 0 0 1 8 7 0 2 9 0 2 8 5 7 8 4 6 0 6 5 0 7 1 2
 0 5 2 4 1 5 3 3 1 5 5 1 4 0 1 6 4 3 3 9 8 8 3 4 6 8 4 8 6
 7 3 7 5 2 4 0 2 7 6 3 8 5 5 4 5 8 8 7 5 5 6 5 6 7 9 7 7 4
 0 3 2 8 1 4 4 6 0 8 2 3 0 1 3 4 6 2 0 5 7 7 3 6 1 8 7 3 5
 4 4 8 3 3 3 5 0 1 0 3 8 6 3 2 0 5 0 6 1 3 3 4 3 6 1 5 8 6
 1 0 2 2 7 6 3 3 0 8 8 0 3 1 8 8 1 2 1 7 5 2 9 3 5 8 3 2 5

CID	Recency	Frequency	Monitary	Segment
1	10	28	1850	S3
2	3	35	750	S2
3	35	2	250	S1
4	12	18	400	S2
5	5	1	20	S5
6	45	1	1000	S7
7	120	24	1600	S4
8	17	22	580	S2
9	155	18	1360	S4
10	24	5	120	S1
.	.	.	.	.
.	.	.	.	.

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

گروه دایکه | dayche.com 

شناسایی الگوها – مثال (1)

- ساخت داشبورد و گزارش های مدیریتی
- پیاده سازی سامانه های تعاملی خودکار



تولید محتوا: محمد روزبه

daychegroup

daychegroup

گروه دایکه | dayche.com

❑ شناسایی الگوها – مثال (2)

- چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

❑ شناسایی الگوها – مثال (2)

○ چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

If Income = High then Credit Rate = Good

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

شناسایی الگوها – مثال (2)

○ چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

If Income = Low then Credit Rate = Bad

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

❑ شناسایی الگوها – مثال (2)

○ چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

شناسایی الگوها – مثال (2)

چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

If *Income = Middle & Car = New*
then *Credit Rate = Good*

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

شناسایی الگوها – مثال (2)

چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

If *Income = Middle & Car = Old*
then *Credit Rate = Bad*

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

CID	Income	Car	Gender	Credit Rate
C-1	Low	New	F	Bad
C-2	Middle	Old	F	Bad
C-3	Middle	New	M	Good
C-4	Low	New	M	Bad
C-5	High	New	M	Good
C-6	High	New	F	Good
C-7	Middle	New	F	Good
C-8	High	Old	F	Good
C-9	Middle	Old	M	Bad
C-10	Low	Old	F	Bad

شناسایی الگوها – مثال (2)

چگونه می توان الگوی اعتباری مشتریان را شناسایی کرد؟

If Income = High then Credit Rate = Good

If Income = Low then Credit Rate = Bad

If Income = Middle & Car = New then Credit Rate = Good

If Income = Middle & Car = Old then Credit Rate = Bad

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

□ چالش های شناسایی الگوها

- انتخاب داده های مناسب
- ساخت شاخص های مناسب
- انجام تبدیلات مناسب
- انتخاب الگوریتم مناسب
- انتخاب طرح ارزیابی مناسب
- انتخاب الگوی مناسب و ...



داده کاوی
Data Mining

تعریف داده کاوی

- فرآیند کشف دانش از پایگاه های داده به منظور شناسایی الگوهای موجود در داده ها که شرایط زیر را داشته باشند:
- معتبر باشند
- کارا و قابل استفاده باشند
- بدیع باشند
- توجیه پذیر باشند.

(اسامه فیاض – 1996)

One of the most used definition (Fayyad et al 1996):

Knowledge Discovery in Databases (KDD) is a

process that aims at finding :

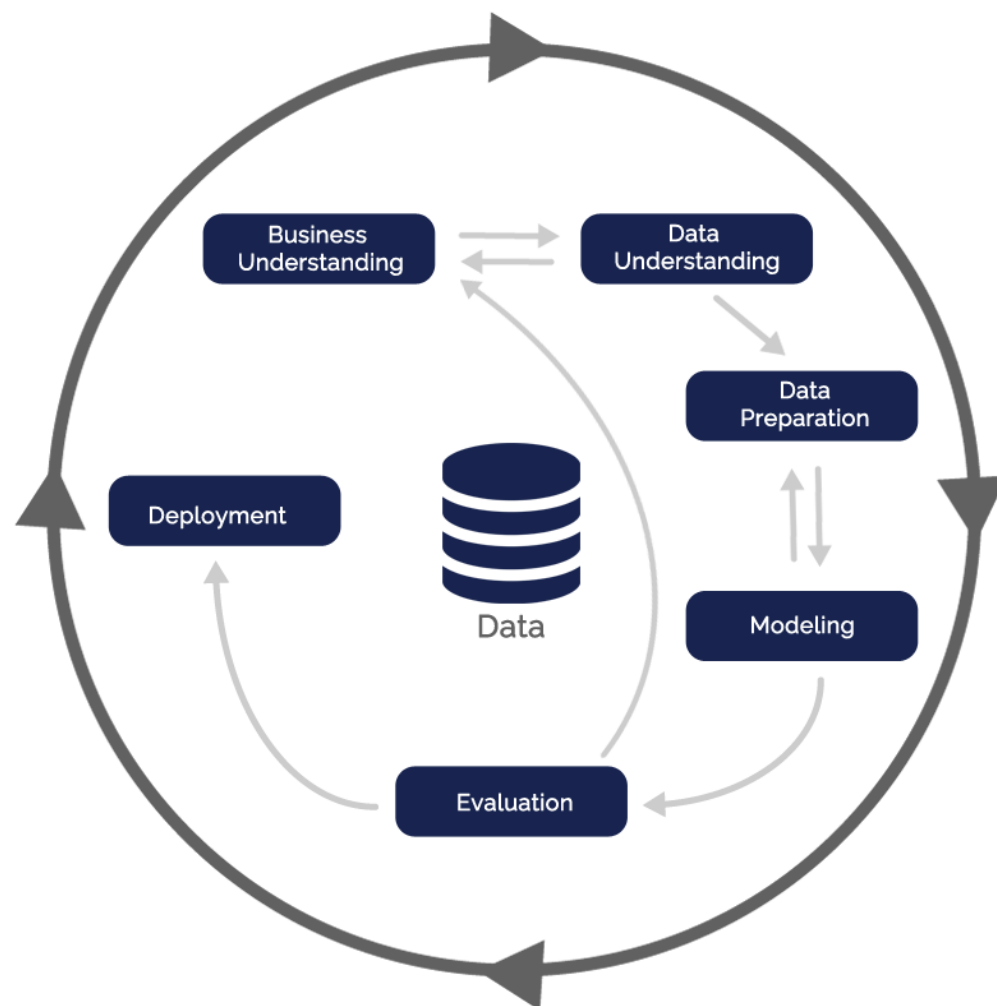
- valid
 - useful
 - novel
 - understandable
- patterns in data**

□ تعریف داده کاوی

- داده کاوی یک محصول نیست! داده کاوی یک نرم افزار نیست!
- داده کاوی یک الگوریتم نیست!
- داده کاوی فرآیند حل مسئله هست.

○ فازهای فرآیند داده کاوی

- شناسایی و درک مسئله
- شناسایی و درک داده ها
- آماده سازی داده ها
- مدلسازی
- ارزیابی
- گسترش و توسعه

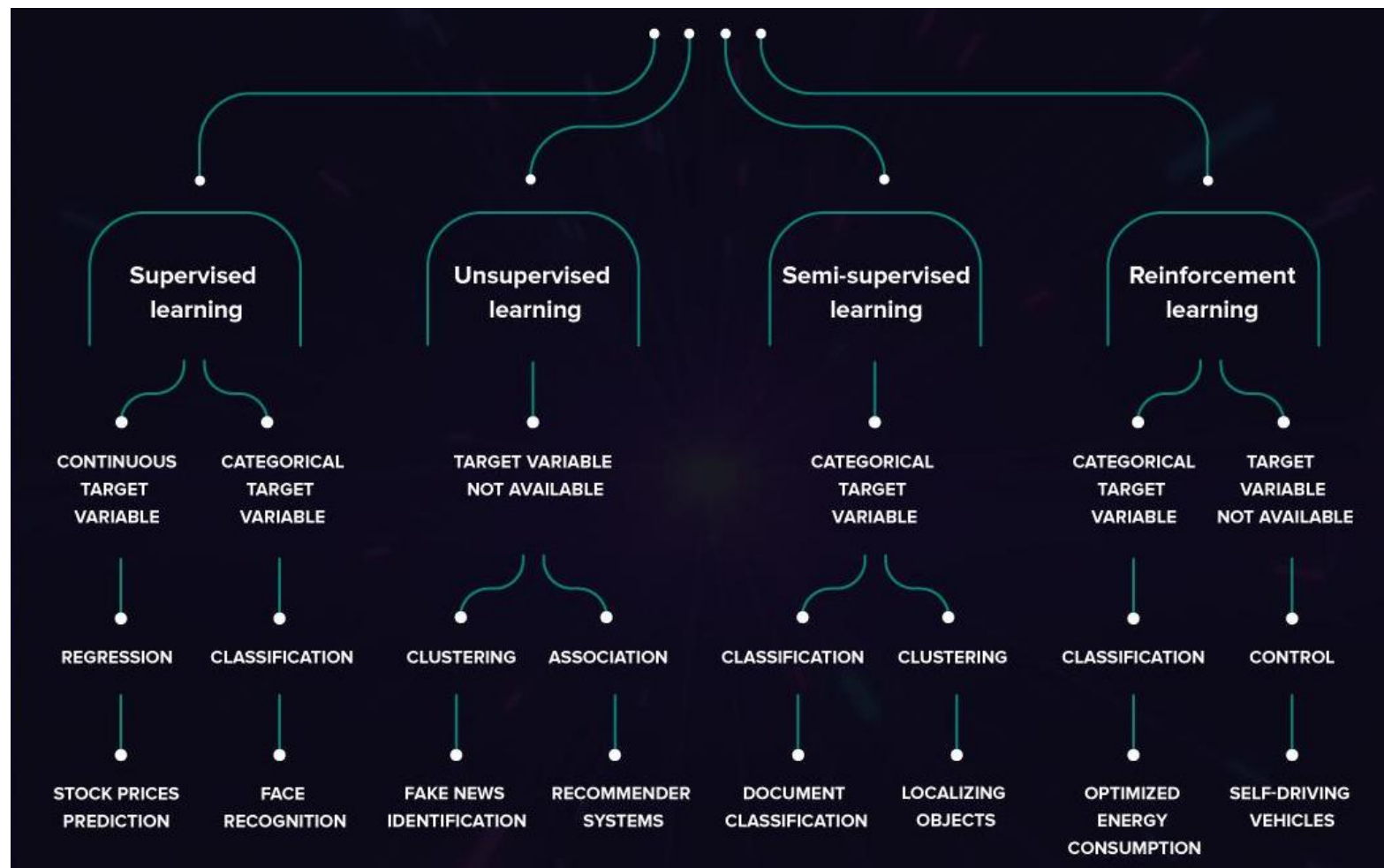


تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



روش های یادگیری

- یادگیری با نظارت
(Supervised Learning)
- یادگیری بدون نظارت
(Unsupervised Learning)
- یادگیری نیمه نظارتی
(Semi-supervised Learning)
- یادگیری تقویتی
(Reinforcement Learning)

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

روش های یادگیری □

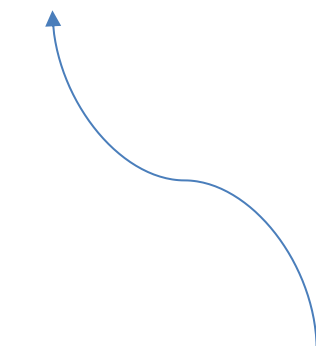
○ یادگیری با نظارت

(Supervised Learning)

○ رگرسیون (Regression): پیش بینی مقدار هدف با توزیع آماری پیوسته

Outcome Function Input

$$Y = F(X)$$

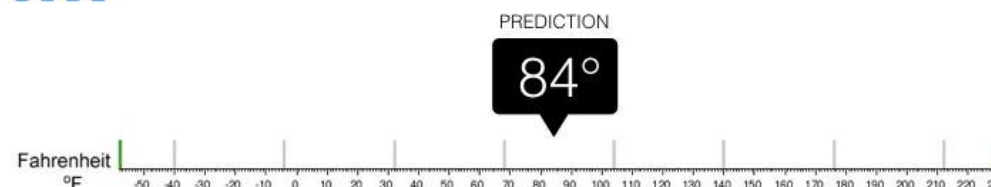


X1	X2	X3	X4	Target
.	.	.	.	54
.	.	.	.	32
.	.	.	.	65
.	.	.	.	74
.	.	.	.	45
.	.	.	.	23
.	.	.	.	18
.	.	.	.	77



Regression

What is the temperature going to be tomorrow?



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

روش های یادگیری

یادگیری با نظارت

(Supervised Learning)

طبقه بندی (Classification): پیش بینی برچسب هدف با توزیع آماری گسسته

$$Y = F(X)$$

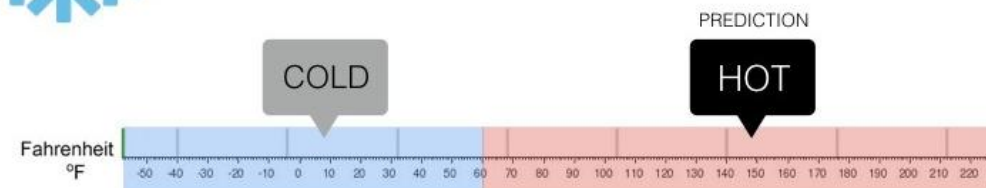
Outcome Function Input

X1	X2	X3	X4	Lable
.	.	.	.	COLD
.	.	.	.	COLD
.	.	.	.	HOT
.	.	.	.	HOT
.	.	.	.	COLD
.	.	.	.	COLD
.	.	.	.	COLD
.	.	.	.	HOT



Classification

Will it be Cold or Hot tomorrow?



تولید محتوا: محمد روزبه

daychegroup

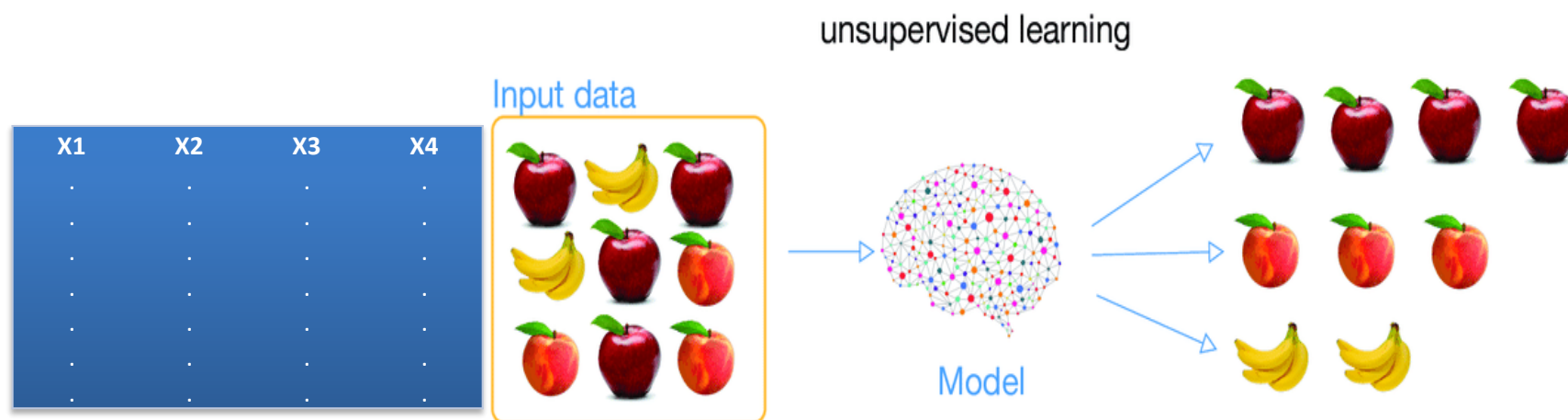
daychegroup

dayche.com | گروه دایکه

روش های یادگیری □

○ یادگیری بدون نظارت
(Unsupervised Learning)

○ خوشه بندی (Clustering): دسته بندی اتوماتیک مشاهدات بر مبنای معیار شباهت



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

روش های یادگیری

- یادگیری بدون نظارت
(Unsupervised Learning)

- قوانین انجمنی (Association Rules): تحلیل وابستگی بین وقوع رخدادها از نظر باهم آیی و تقدم-تاخر



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

روش های یادگیری و انواع مسئله ها

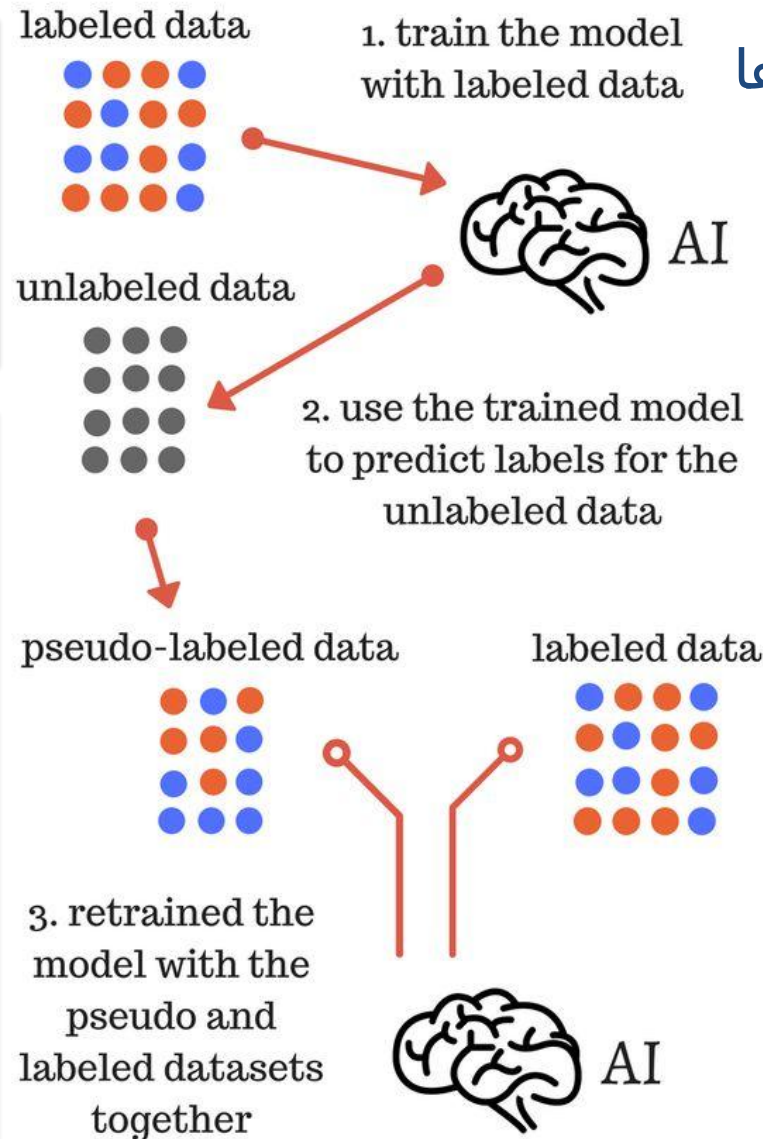
روش های یادگیری

○ یادگیری نیمه نظارتی

(Semi-supervised Learning)

- در این روش گروهی از داده ها دارای برچسب با توزیع احتمالی گسسته بوده و مابقی داده ها بدون برچسب هستند.

- رویکرد یادگیری هم بر اساس خوشه بندی و هم بر اساس طبقه بندی قابل توسعه می باشد.



X1	X2	X3	X4	Lable
.	.	.	.	Red
.	.	.	.	Red
.	.	.	.	Blue
.	.	.	.	Red
.	.	.	.	Blue
.	.	.	.	Blue
.	.	.	.	Red
.	.	.	.	Red

[illegible]

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

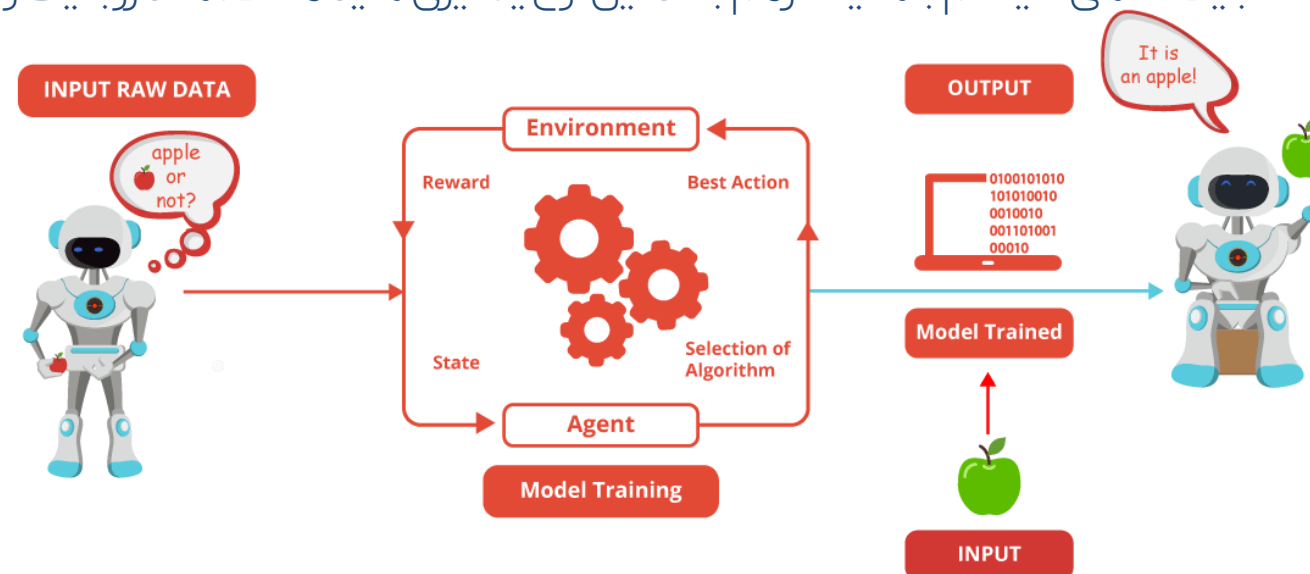
dayche.com | گروه دایکه

روش های یادگیری

یادگیری تقویتی

(Reinforcement Learning)

- در این نوع یادگیری زوج ورودی-خروجی وجود ندارد و سیستم بر اساس هر اکشن انجام شده از محیط پاداش می گیرد. بنابراین در نهایت سیستم تلاش می کند مجموع پاداش خود را بیشینه کند.
- معمولا در مواردی که قابلیت تعاملی سیستم با محیط فراهم باشد، این نوع یادگیری مفید هست: مانند رباتیک و ...



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

انتخاب الگوریتم؟

○ قدم اول:

آیا در مسئله مورد نظر به دنبال پیش بینی یک هدف هستیم (یادگیری با نظارت) یا به دنبال شناسایی ساختار درونی داده می گردیم؟ (یادگیری بدون نظارت)

○ قدم دوم:

در صورت داشتن هدف، آیا توزیع احتمالی آن پیوسته هست (رگرسیون) یا گسسته؟ (طبقه بندی)

○ قدم سوم:

آیا نوع الگویی که برای حل مسئله نیاز داریم محدودیت خاصی دارد:

مانند استخراج قوانین، مدل تابعی، مدل جعبه سیاه و ...؟

Unsupervised

Clustering &
Dimensionality Reduction

SVD

PCA

K-Means

Supervised

Continuous

Regression

Linear

Polynomial

Decision Trees

Random Forests

Nerual Networks

Categorical

Classification

KNN

Trees

Logistic Regression

Naive-Bayes

SVN

Nerual Networks

Association Analysis

Apriori

FP-Growth

Hidden Markov Model

Data Science

متدولوژی اجرای علم داده

گروه دایکه . dayche.com





استانداردهای قابل استفاده ☐

Methodology



☐ CRISP-DM

☐ SEMMA

☐ And ...

Deployment



☐ PMML

☐ PFA

☐ And ...

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

□ فاز توسعه و استقرار

Domain-Specific Languages

**R and
Scikit-Learn**

PMML

PFA

Java and C

General-Purpose Languages

○ استانداردهای قابل کاربرد در فاز توسعه و استقرار

توسعه داده شده توسط گروه DMG

(www.dmg.org)

○ PMML: خروجی مدل‌های پیش بینی به فرمت XML

Predictive Model Markup Language

○ PFA: خروجی مدل‌های پیش بینی به فرمت Json

Portable Format for Analytics

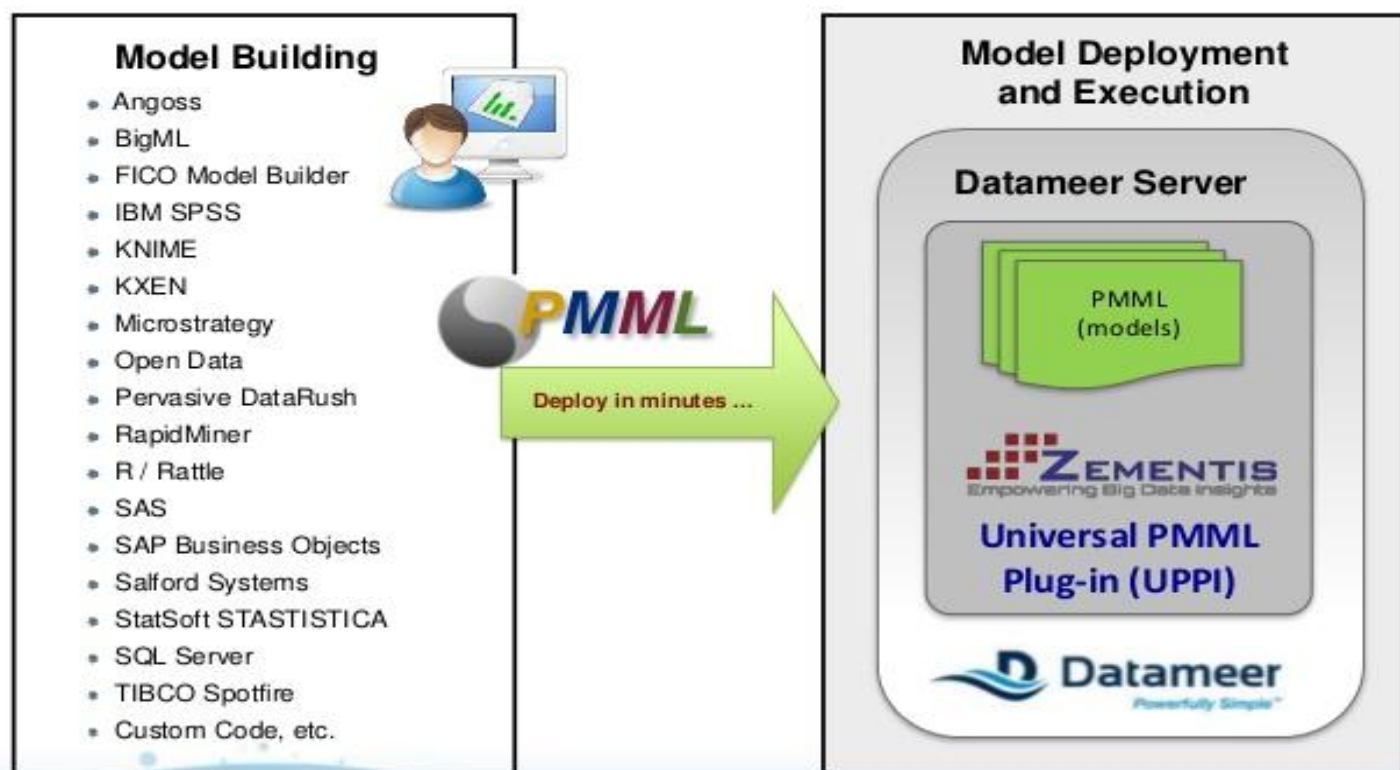
تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

From Model Building to Model Deployment (with PMML)



□ فاز توسعه و استقرار

- انتقال مدل‌های ایجاد شده به سایر محیط های نرم افزاری با محیط های عملیاتی

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

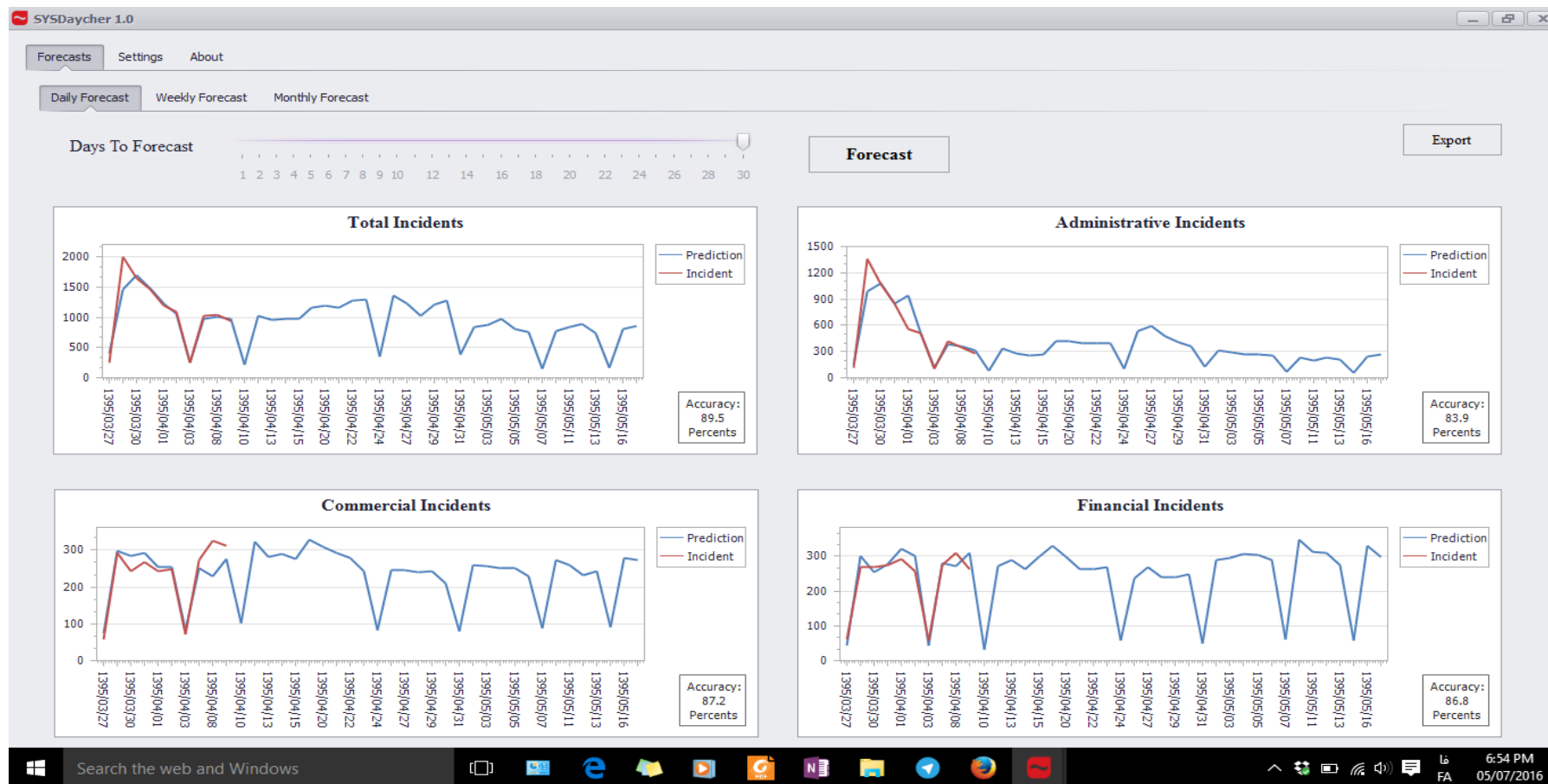
dayche.com | گروه دایکه 

علم داده

متدولوژی اجرایی

□ فاز توسعه و استقرار

○ مثال اجرایی از کاربرد PMML



تولید محتوا: محمد روزبه

daychegroup

daychegroup

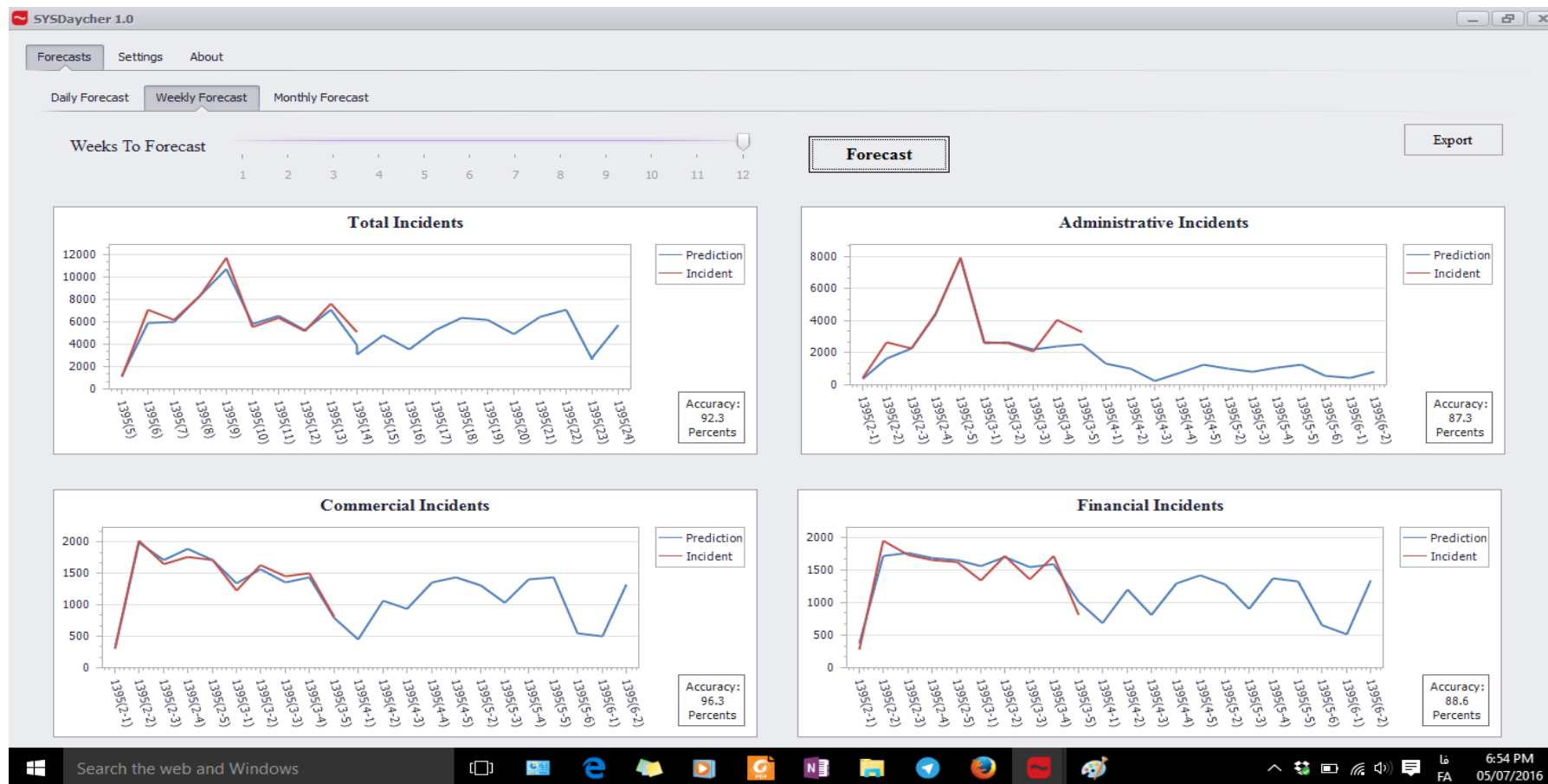
dayche.com | گروه دایکه

علم داده

متدولوژی اجرایی

□ فاز توسعه و استقرار

○ مثال اجرایی از کاربرد PMML



تولید محتوا: محمد روزبه

daychegroup

daychegroup

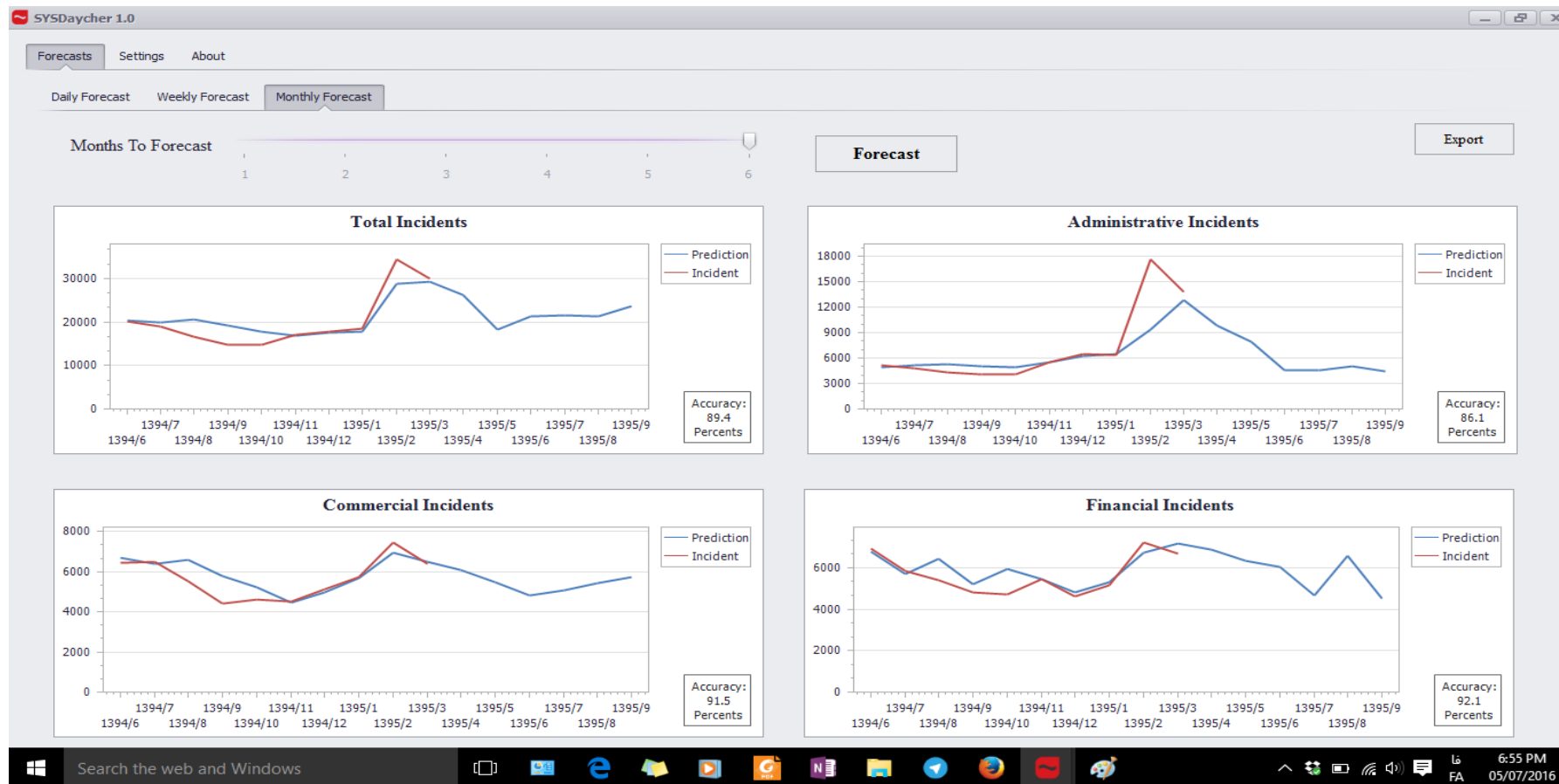
dayche.com | گروه دایکه

علم داده

متدولوژی اجرایی

❑ فاز توسعه و استقرار

○ مثال اجرایی از کاربرد PMML



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



فاز توسعه و استقرار



مثال اجرایی از کاربرد PMML



Forecasts	Settings	About
Model PMML		PMML
Daily Total Model Built Date	Thu Jun 30 22:23:16 2016	
Daily Administrative Model Built Date	Fri Jul 01 02:53:15 2016	
Daily Commercial Model Built Date	Fri Jul 01 03:08:12 2016	
Daily Financial Model Built Date	Fri Jul 01 03:11:44 2016	
Weekly Total Model Built Date	Fri Jul 01 03:17:03 2016	
Weekly Administrative Model Built Date	Fri Jul 01 03:21:28 2016	
Weekly Commercial Model Built Date	Fri Jul 01 03:25:25 2016	
Weekly Financial Model Built Date	Fri Jul 01 03:29:28 2016	
Monthly Model Built Date	Fri Jul 01 03:34:59 2016	
Monthly Administrative Model Built Date	Fri Jul 01 03:39:03 2016	
Monthly Commercial Model Built Date	Fri Jul 01 03:42:19 2016	
Monthly Financial Model Built Date	Fri Jul 01 03:45:36 2016	

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



❑ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

○ محصول کنسرسیومی متشکل از 3 شرکت (1996)

○ دایملر - کرایسلر

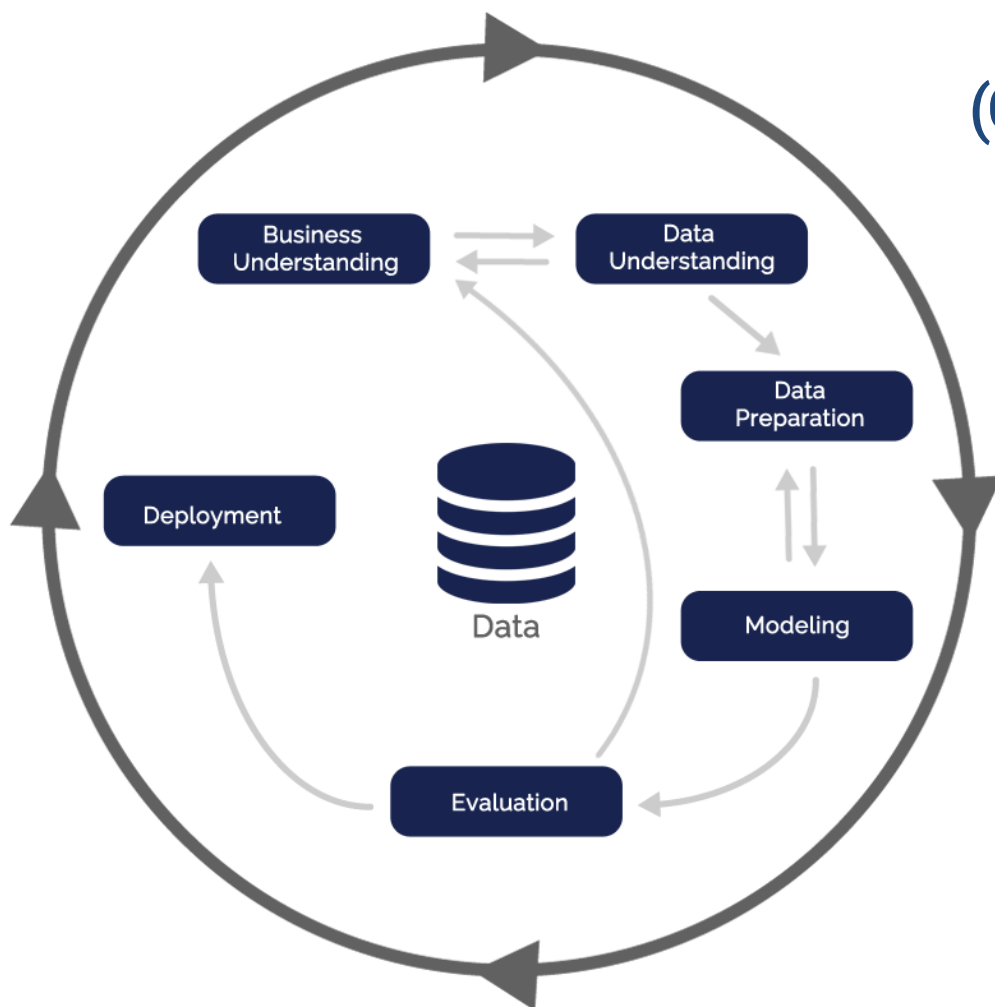
○ SPSS

○ NCR

○ با هدف کاهش ریسک اجرای پروژه های داده کاوی

○ ارائه متدولوژی جهت پیاده سازی گام به گام بر اساس تجربیات جهانی

○ ارائه چارچوبی برای مستندسازی پروژه های داده کاوی

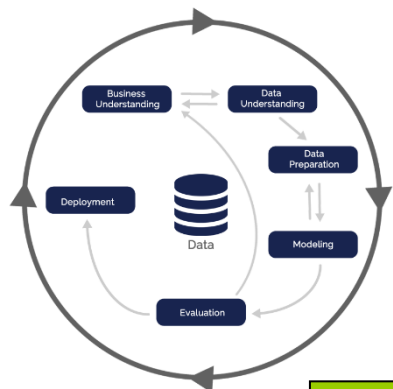


تولید محتوا: محمد روزبه

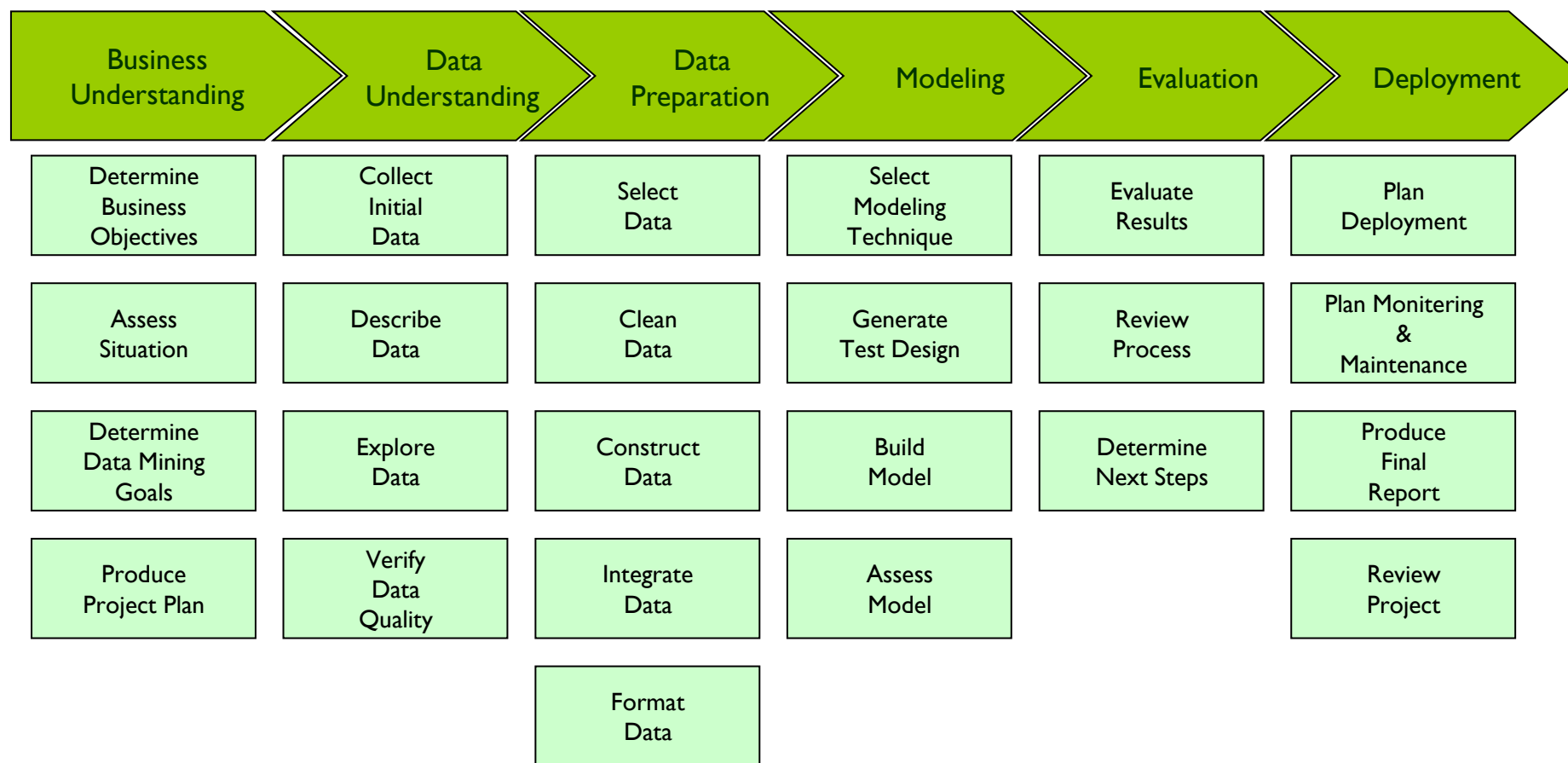
daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

تعریف دقیق مسئله

اهداف و ابعاد پروژه، معیارهای موفقیت به همراه شرح کاملی از وضعیت موجود و راهکارهای فعلی



فاز اول

شناسایی و درک
کسب و کار

ارزیابی وضعیت

شناسایی ذینفعان، افراد موثر در پیشبرد پروژه، منابع داده، محدودیت ها و ریسک ها



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

اهداف داده کاوی

ترجمه دقیق و طراحی سناریو داده کاوی بر اساس نوع یادگیری و انتخاب وظایف و الگوریتمهای احتمالی



ارائه طرح پروژه

مستند سازی نتایج و نقشه راه حل مسئله



فاز اول

شناسایی و درک
کسب و کار

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

گردآوری داده ها

در اغلب موارد داده های مورد نیاز در حل مسئله در جاهای مختلف و با فرمت های متنوع پراکنده هستند.



توصیف داده ها

استفاده از روش های آماری توصیفی و مستند سازی نحوه دسترسی به داده ها و چارچوب آنها تاثیر زیادی در شناخت اولیه و دید شهودی نسب به داده ها می گردد.



فاز دوم

شناسایی و درک داده ها

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

کیفیت داده ها

بررسی کیفیت داده های موجود، باعث می گردد تا با تعیین استراتژی های مناسب به پاکسازی داده پردازیم و با بهبود کیفیت داده ها اقدام به یادگیری مدل کنیم.



کاوش در داده ها

بررسی آماری دقیق از همبستگی های موجود و روابط معنادار بین آنها قدم اول در انتخاب، ایجاد و تغییر مجموعه داده ها است.



فاز دوم

شناسایی و درک
داده ها

تولید محتوا: محمد روزبه

daychegroup

daychegroup

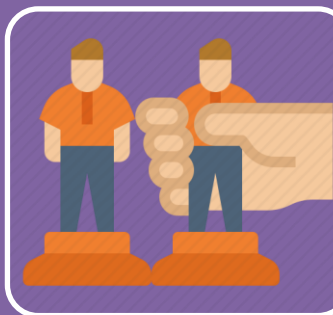
dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

انتخاب داده ها

انتخاب متغیرها: تکنیک های Feature Selection/Reduction
انتخاب مشاهدات: تکنیک های نمونه گیری (Sampling)



پاکسازی داده

برخورد مناسب با مقادیر گمشده
برخورد مناسب با مقادیر/رکوردهای پرت
برخورد مناسب با مقادیر غیر طبیعی



فاز سوم

آماده سازی داده ها

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

ساخت صفات مشتقه از داده ها

ساخت متغیرها: بر اساس شناخت آماری حاصل شده

ساخت مشاهدات: تکنیک های نمونه سازی مانند SMOTE



یکپارچه سازی داده ها و فرمت بندی داده

ایجاد فایل داده متمرکز حاوی کلیه داده های پراکنده

ایجاد الزامات مربوط به نرم افزار مورد استفاده



فاز سوم

آماده سازی داده ها

تولید محتوا: محمد روزبه

daychegroup

daychegroup

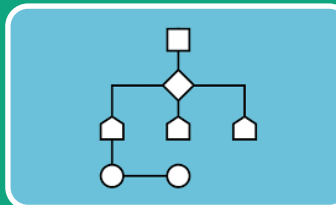
dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

انتخاب الگوریتم و طرح آزمون

مبتنی بر گام های اسلاید 78 الگوریتم های مناسب انتخاب و متناسب با حجم و توزیع داده ها طرح آزمون برای ارزیابی مدل ها تعیین می گردد.



آموزش الگوریتم و ساخت مدل

تنظیم پارامترها و توصیف مدل های ایجاد شده



ارزیابی مدل ها

بررسی نتایج مدل ها و انتخاب بهترین مدل



فاز چهارم

مدلسازی

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ فرآیند استاندارد فراصنعتی برای داده کاوی (CRISP-DM)

فاز پنجم

ارزیابی



ارزیابی کلیات راه حل
ارزیابی نتایج بدست آمده با طرح اولیه پروژه و معیارهای موفقیت
مروری بر فرآیند طی شده و مستندسازی

فاز ششم

توسعه و استقرار



طرح توسعه و مانیتورینگ فرآیند انجام شده
مستند سازی روال انجام شده و نتایج بدست آمده
بهره برداری از نتایج در قالب پرزنتیشن، داشبورد، سرویس یا نرم افزار

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

Dive into Data Science

گروه دایکه . dayche.com



پروژه 1: داده های بیماران یک کلینیک

گام اول: تعریف مسئله

	Age	Sex	BP	Cholesterol	Na	K	Drug
0	23	F	HIGH	HIGH	0.792535	0.031258	drugY
1	47	M	LOW	HIGH	0.739309	0.056468	drugC
2	47	M	LOW	HIGH	0.697269	0.068944	drugC
3	28	F	NORMAL	HIGH	0.563682	0.072289	drugX
4	61	F	LOW	HIGH	0.559294	0.030998	drugY

دیتاست روبرو از طرف یک کلینیک پزشکی به شما معرفی شد تا به صورت داده محور در روند تصمیم گیری و خدمات این کلینیک راهکار ارائه دهید.

فرض کنید، در این کلینیک 10 پزشک حضور دارند و برای هر بیمار پس از تکمیل پرونده و ویزیت توسط یکی از آنها، یک داروی مشخص تجویز می شود.

به نظر شما، چه نوع مسائلی در این شرایط فرضی و بر اساس داده های موجود می توان مطرح نمود تا برای مدیریت کلینیک جذاب و دارای ارزش افزوده باشد؟

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

پروژه 1: داده های بیماران یک کلینیک

	Age	Sex	BP	Cholesterol	Na	K	Drug
0	23	F	HIGH	HIGH	0.792535	0.031258	drugY
1	47	M	LOW	HIGH	0.739309	0.056468	drugC
2	47	M	LOW	HIGH	0.697269	0.068944	drugC
3	28	F	NORMAL	HIGH	0.563682	0.072289	drugX
4	61	F	LOW	HIGH	0.559294	0.030998	drugY

گام اول: تعریف مسئله

توصیف و کاوش داده ها Exploratory Data Analysis (EDA)

- چه داروهایی بطور عمده توسط پزشکان این کلینیک تجویز می شود؟
- چند درصد بیماران این کلینیک را مردان/زنان تشکیل می دهند؟
- توزیع سن بیماران این کلینیک به چه شکلی هست؟
- آیا رابطه ای بین فشار خون و میزان کلسترول در بین بیماران وجود دارد؟
- مقدار سدیم و پتاسیم اندازه گیری شده چه ارتباطی با هم دارند؟ و ...

خلاصه سازی + ارتباط سنجی آماری

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

پروژه 1: داده های بیماران یک کلینیک

گام اول: تعریف مسئله

	Age	Sex	BP	Cholesterol	Na	K	Drug
0	23	F	HIGH	HIGH	0.792535	0.031258	drugY
1	47	M	LOW	HIGH	0.739309	0.056468	drugC
2	47	M	LOW	HIGH	0.697269	0.068944	drugC
3	28	F	NORMAL	HIGH	0.563682	0.072289	drugX
4	61	F	LOW	HIGH	0.559294	0.030998	drugY

مدلسازی اکتشافی و شناسایی الگو

مدل های خوشه بندی

- آیا الگویی در بین پروفایل بیماران وجود دارد؟
- آیا می توان بین گروه های مشابه بیماران و نوع دارو تجویز شده ارتباطی برقرار کرد؟
- آیا امکان جایگزینی بین داروهای تجویز شده بر اساس پروفایل بیماران وجود دارد؟

پروژه 1: داده های بیماران یک کلینیک

گام اول: تعریف مسئله

	Age	Sex	BP	Cholesterol	Na	K	Drug
0	23	F	HIGH	HIGH	0.792535	0.031258	drugY
1	47	M	LOW	HIGH	0.739309	0.056468	drugC
2	47	M	LOW	HIGH	0.697269	0.068944	drugC
3	28	F	NORMAL	HIGH	0.563682	0.072289	drugX
4	61	F	LOW	HIGH	0.559294	0.030998	drugY

مدلسازی پیش بینانه

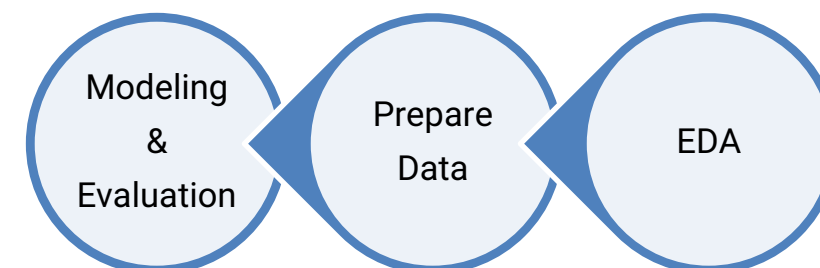
- آیا می توان بین پروفایل هر بیمار و نوع دارو ارتباطی پیدا کرد؟
- آیا امکان ساخت سامانه آنلایین و مشاوره تجویز دارو برای بیماران این کلینیک وجود دارد؟

مدل های رده بندی

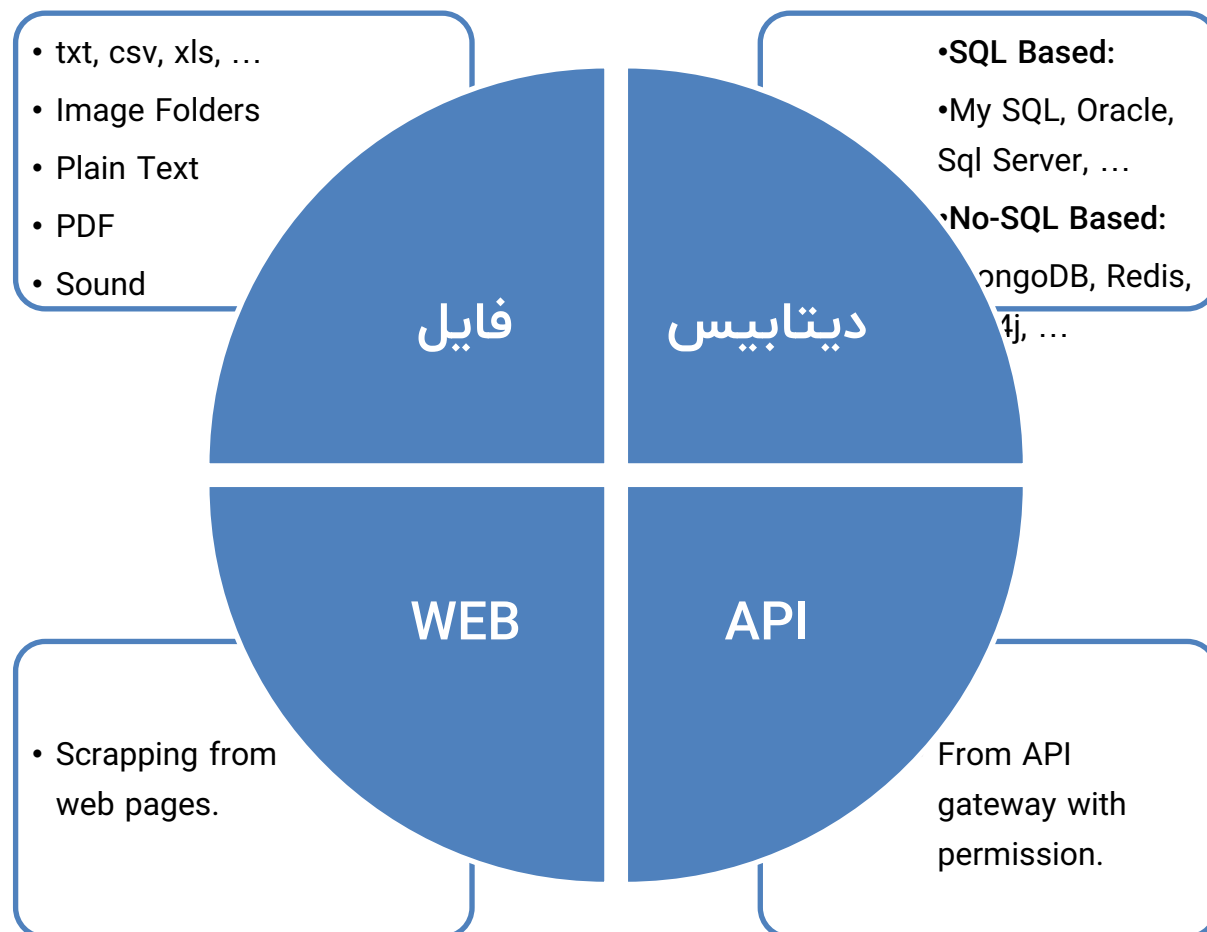
پروژه 1: داده های بیماران یک کلینیک

	Age	Sex	BP	Cholesterol	Na	K	Drug
0	23	F	HIGH	HIGH	0.792535	0.031258	drugY
1	47	M	LOW	HIGH	0.739309	0.056468	drugC
2	47	M	LOW	HIGH	0.697269	0.068944	drugC
3	28	F	NORMAL	HIGH	0.563682	0.072289	drugX
4	61	F	LOW	HIGH	0.559294	0.030998	drugY

گام دوم: پیاده سازی نرم افزاری



مروری بر فرآیند EDA و چالشهای آن



○ دسترسی به منابع داده

○ خلاصه سازی آماری

○ ارتباط سنجی آماری

○ نمودارهای بصری

○ کیفیت داده ها

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر فرآیند EDA و چالشهای آن

نقش آمار و احتمال در علم داده

مدلسازی

ارتباط سنجی

خلاصه سازی

تئوری و مفاهیم احتمالات

○ دسترسی به منابع داده

○ خلاصه سازی آماری

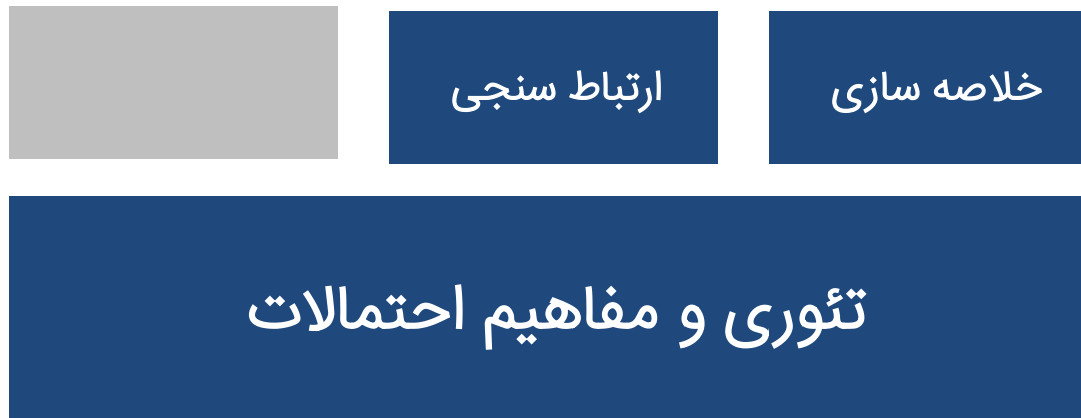
○ ارتباط سنجی آماری

○ نمودارهای بصری

○ کیفیت داده ها

مروری بر فرآیند EDA و چالشهای آن

نقش آمار و احتمال در فرآیند EDA



دسترسی به منابع داده ○

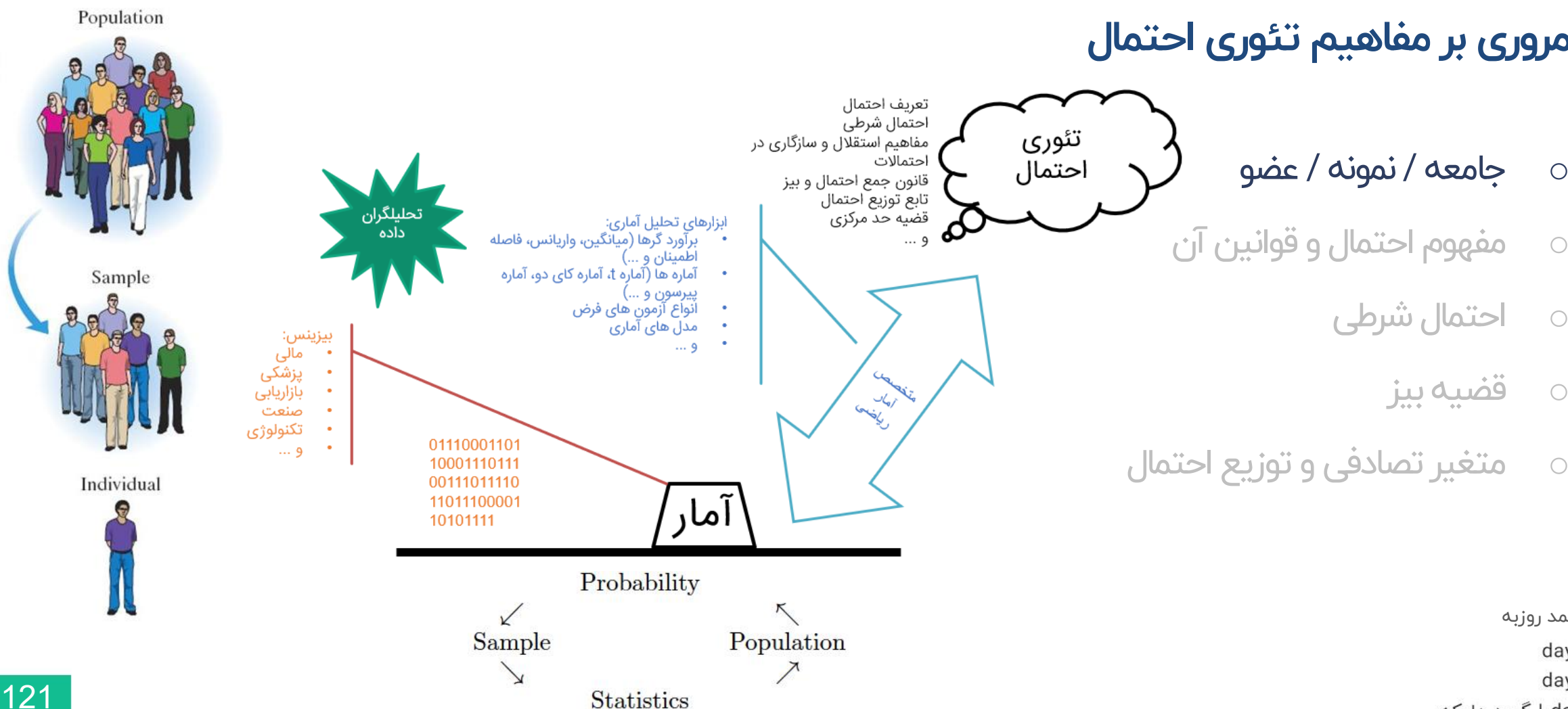
خلاصه سازی آماری ○

ارتباط سنجی آماری ○

نمودارهای بصری ○

کیفیت داده ها ○

مروری بر مفاهیم تئوری احتمال



جامعه / نمونه / عضو

مفهوم احتمال و قوانین آن

احتمال شرطی

قضیه بیز

متغیر تصادفی و توزیع احتمال

$$P(A) = \frac{n}{N} = \frac{\text{\# outcomes in } A}{\text{\# outcomes in Sample Space}}$$

مروری بر مفاهیم تئوری احتمال □

Drug:			$P(\text{Drug} = Y) = \frac{91}{200} = 0.455$
	Count	Percentage (%)	$P(\text{Drug} = X) = \frac{54}{200} = 0.27$
drugY	91	45.5	$P(\text{Drug} = A) = \frac{23}{200} = 0.115$
drugX	54	27.0	$P(\text{Drug} = C) = \frac{16}{200} = 0.08$
drugA	23	11.5	$P(\text{Drug} = B) = \frac{16}{200} = 0.08$
drugC	16	8.0	
drugB	16	8.0	

Sex:			$P(\text{Sex} = M) = \frac{104}{200} = 0.52$
	Count	Percentage (%)	$P(\text{Sex} = F) = \frac{96}{200} = 0.48$
M	104	52.0	
F	96	48.0	

جامعه / نمونه / عضو ○

مفهوم احتمال و قوانین آن ○

احتمال شرطی ○

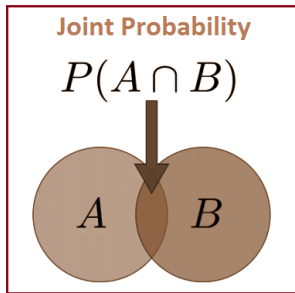
قضیه بیز ○

متغیر تصادفی و توزیع احتمال ○

$$P(\text{Drug} = Y) = \frac{91}{200} = 0.455$$

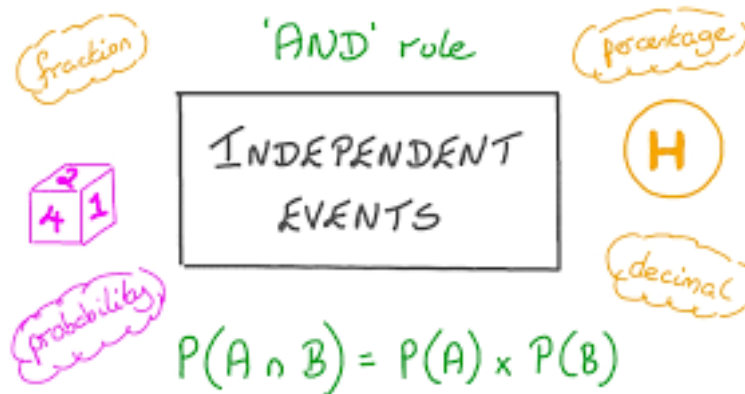
$$P(\text{Sex} = M) = \frac{104}{200} = 0.52$$

$$P(\text{Sex} = F) = \frac{96}{200} = 0.48$$



$$P(\text{Drug} = Y) \cap P(\text{Sex} = M) = \frac{44}{200} = 0.22$$

$$P(\text{Drug} = Y) \cap P(\text{Sex} = F) = \frac{47}{200} = 0.235$$



مروری بر مفاهیم تئوری احتمال

- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال

$$P(\text{Drug} = Y) = \frac{91}{200} = 0.455$$

$$P(\text{Sex} = M) = \frac{104}{200} = 0.52$$

$$P(\text{Sex} = F) = \frac{96}{200} = 0.48$$

$$P(\text{Drug} = Y | \text{Sex} = M) = \frac{44}{104} = 0.423$$

$$P(\text{Drug} = Y | \text{Sex} = F) = \frac{47}{96} = 0.49$$

Conditional Probability Formula

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Probability of A and B

Probability of A given B

Probability of B

مروری بر مفاهیم تئوری احتمال

- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

مروری بر مفاهیم تئوری احتمال

- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال

$$P(\text{Drug} = Y | \text{Sex} = M) = \frac{P(\text{Drug} = Y) \cap p(\text{Sex} = M)}{P(\text{Sex} = M)}$$

$$P(\text{Sex} = M | \text{Drug} = Y) = \frac{P(\text{Drug} = Y) \cap p(\text{Sex} = M)}{P(\text{Drug} = Y)}$$

$$P(\text{Drug} = Y | \text{Sex} = M) = \frac{P(\text{Sex} = M | \text{Drug} = Y) * P(\text{Drug} = Y)}{P(\text{Sex} = M)}$$

تابع احتمال پیشین (Prior)
 احتمال درستی فرضیه بدون در نظر گرفتن شواهد خاص
راست نمایی (Likelihood)
 احتمال مشاهده و وقوع شواهد موجود به شرط درستی فرضیه

$$P(y|x) = \frac{P(x|y) P(y)}{P(x)}$$

تابع احتمال پسین (Posterior)
 احتمال وقوع فرضیه با در نظر گرفتن شواهد موجود

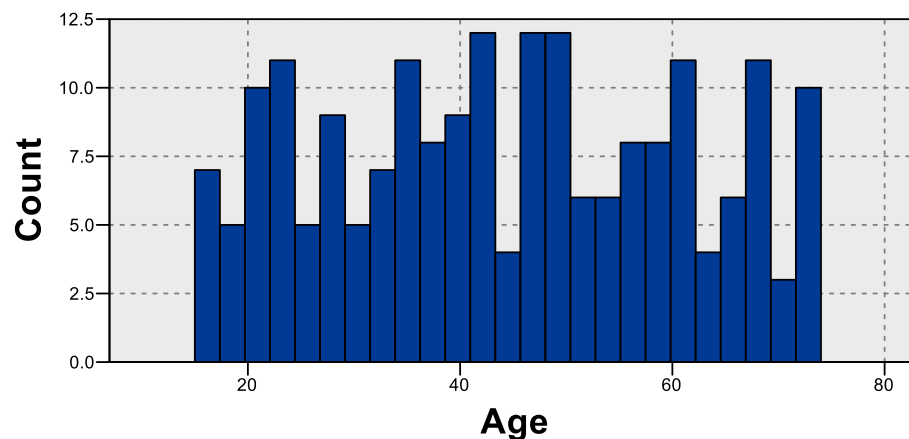
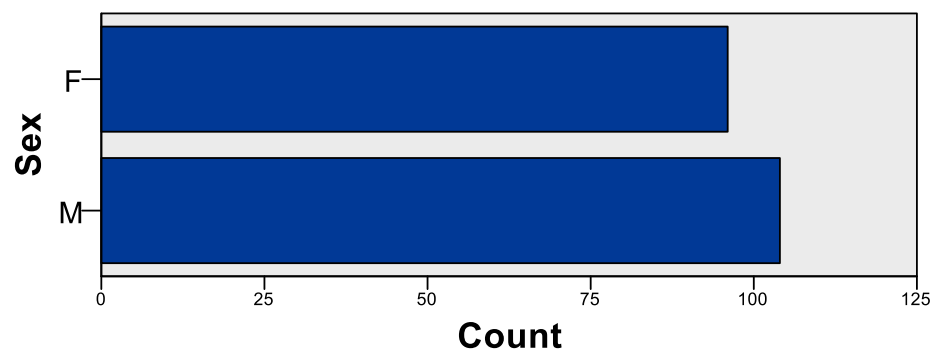
تابع احتمال مرزی یا حاشیه ای (Marginal)
 احتمال مشاهده و وقوع شواهد موجود با در نظر گرفتن همه فرضیات ممکن

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 



مروری بر مفاهیم تئوری احتمال

جامعه / نمونه / عضو

مفهوم احتمال و قوانین آن

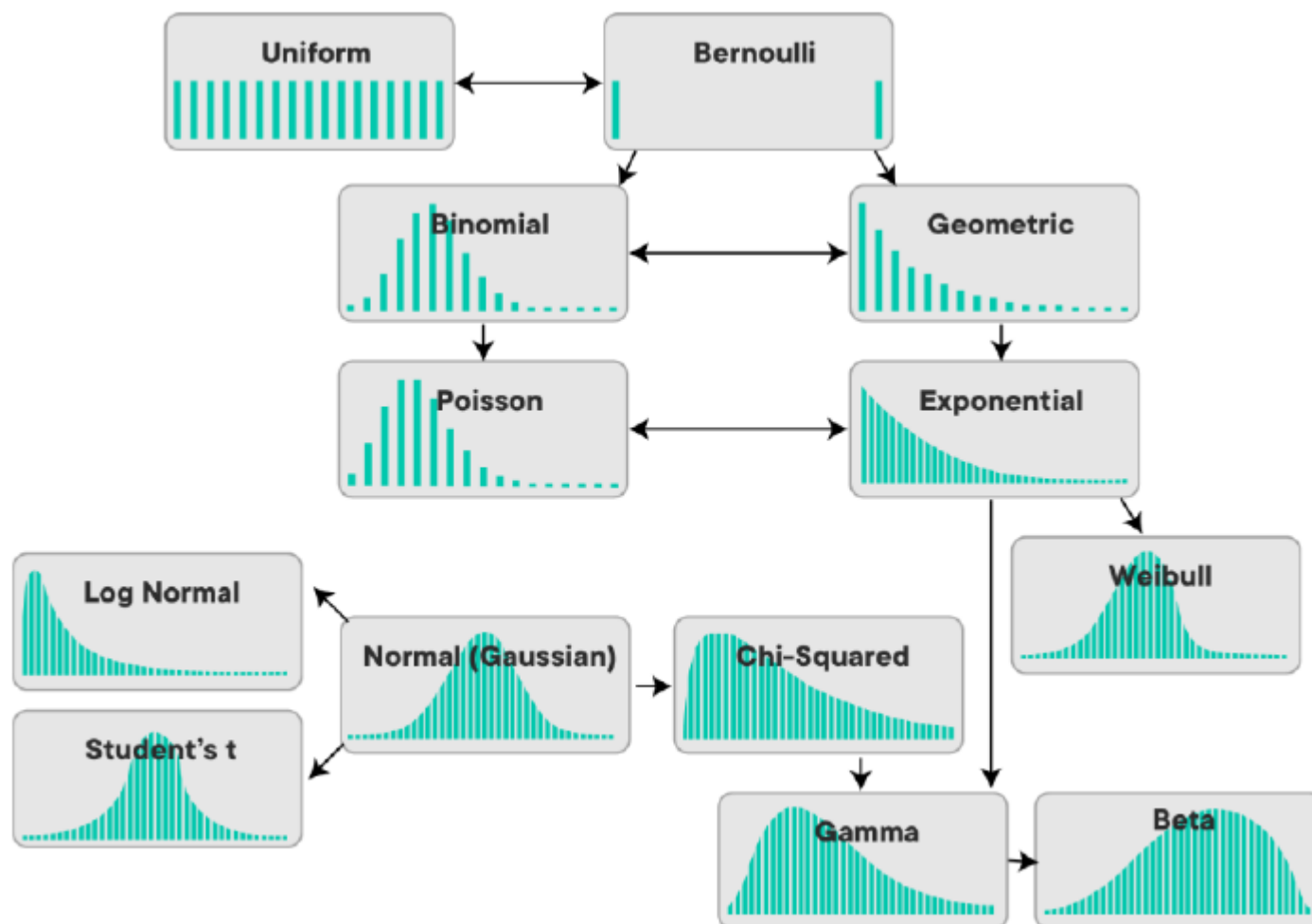
احتمال شرطی

قضیه بیز

متغیر تصادفی و توزیع احتمال

مروری بر مفاهیم تئوری احتمال

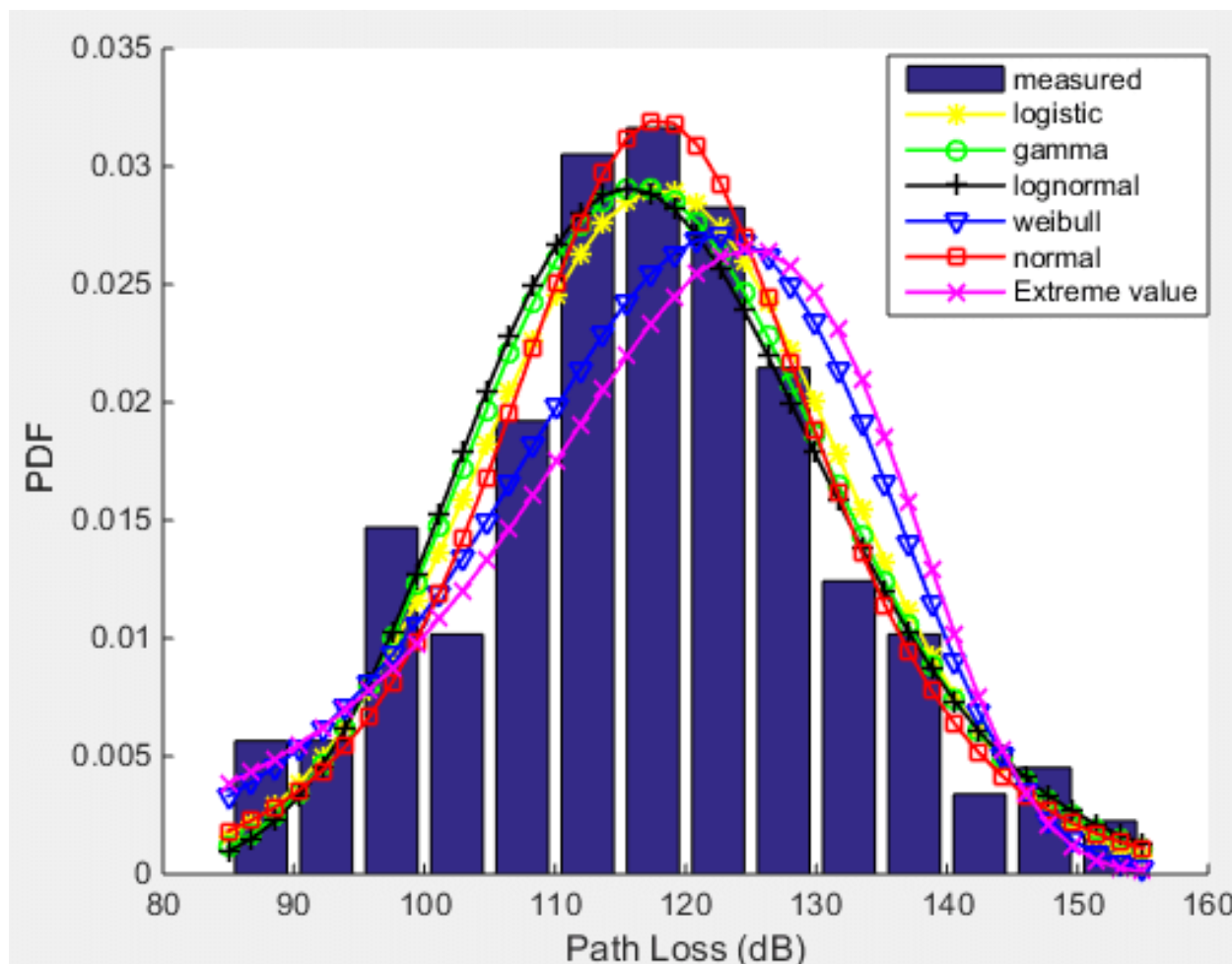
- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال





مروری بر مفاهیم تئوری احتمال

- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال

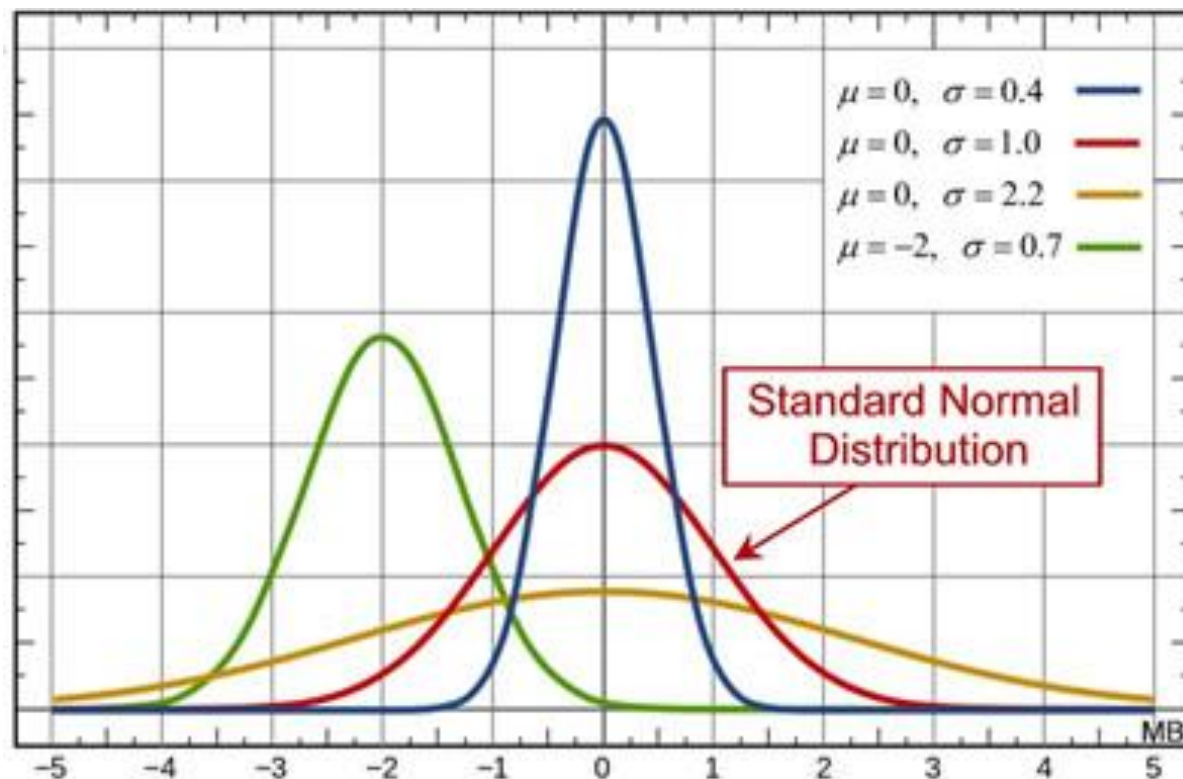


تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



□ مروری بر مفاهیم تئوری احتمال

- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ = mean of x

σ = standard deviation of x

Probability Density Function (PDF)

تولید محتوا: محمد روزبه

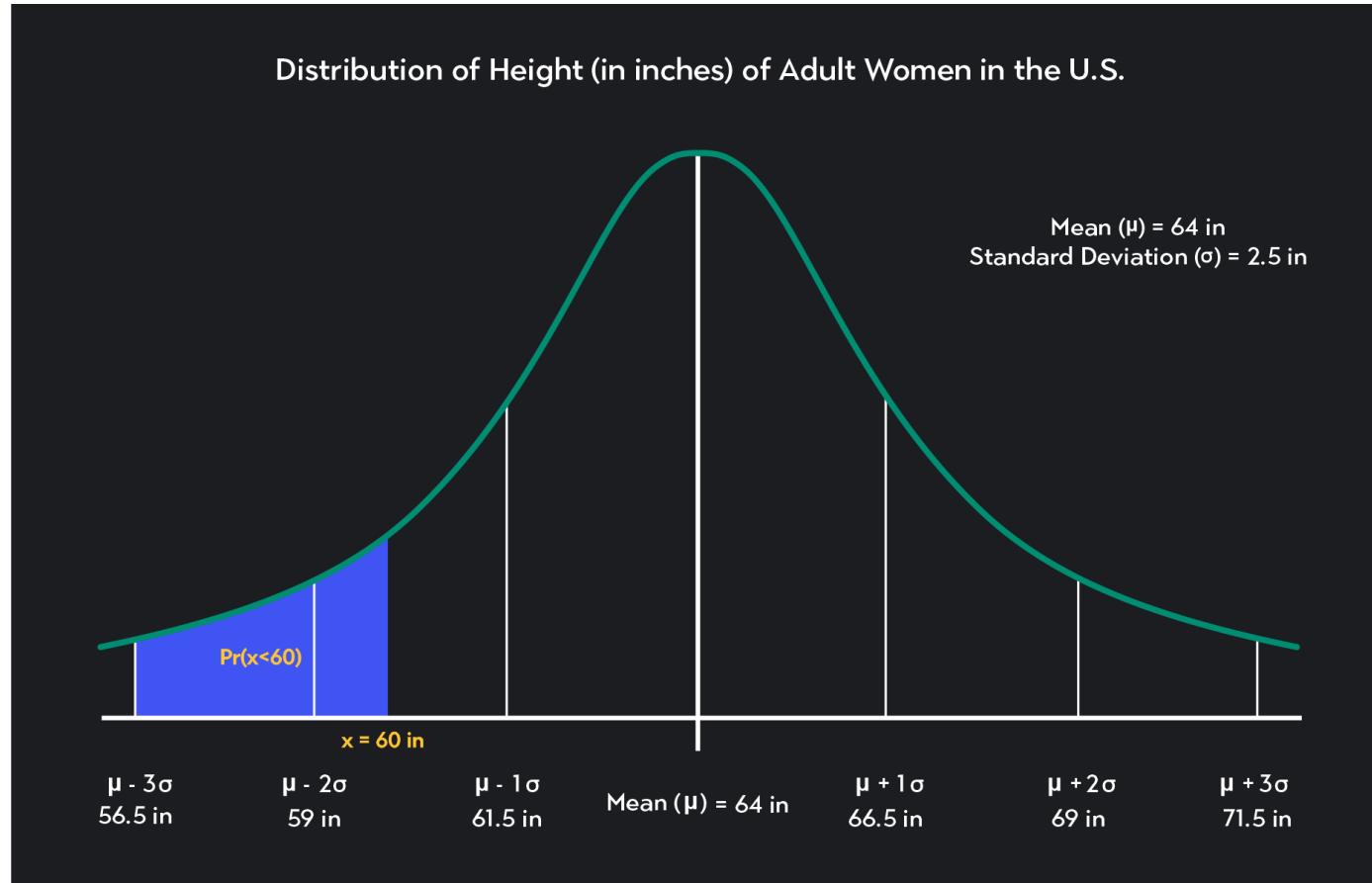
daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر مفاهیم تئوری احتمال

- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال



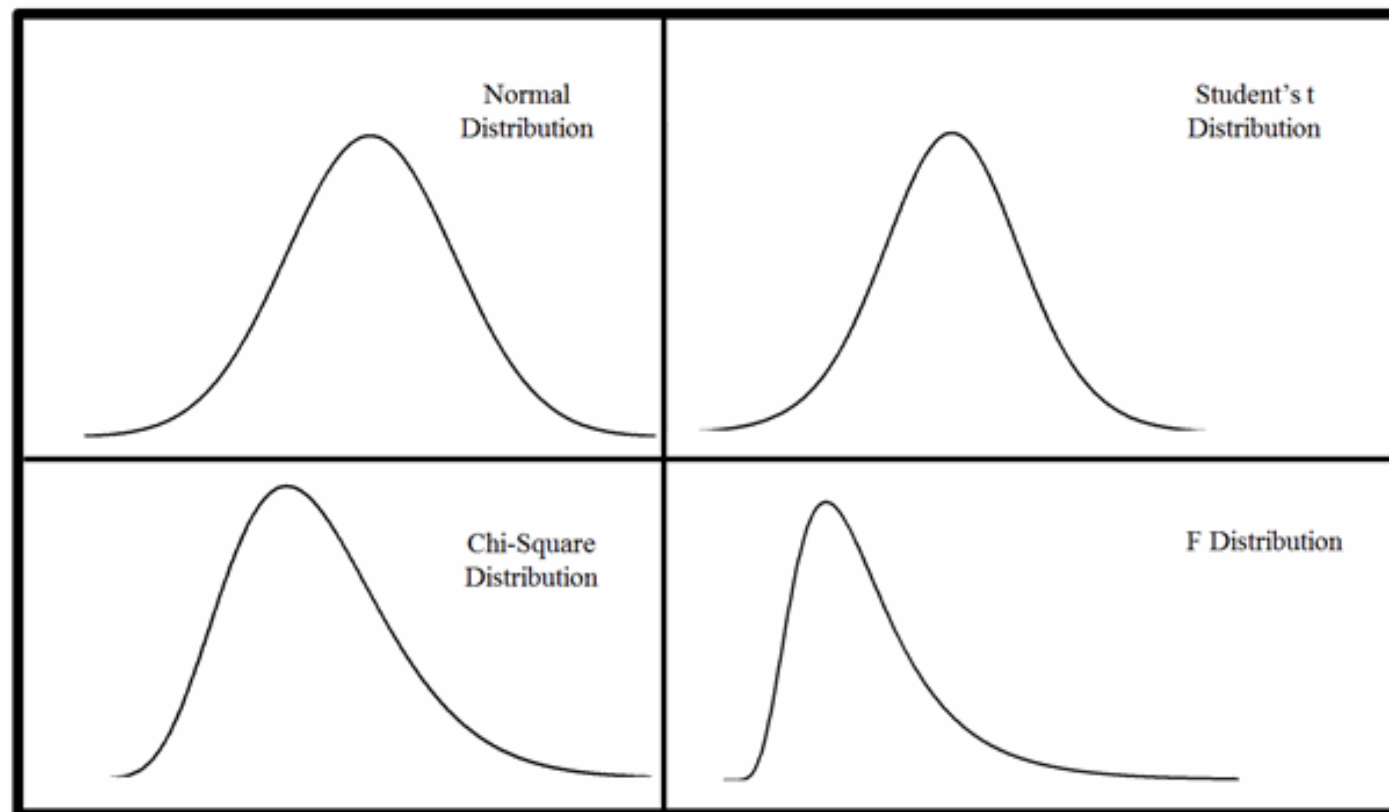
تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر مفاهیم تئوری احتمال



- جامعه / نمونه / عضو
- مفهوم احتمال و قوانین آن
- احتمال شرطی
- قضیه بیز
- متغیر تصادفی و توزیع احتمال

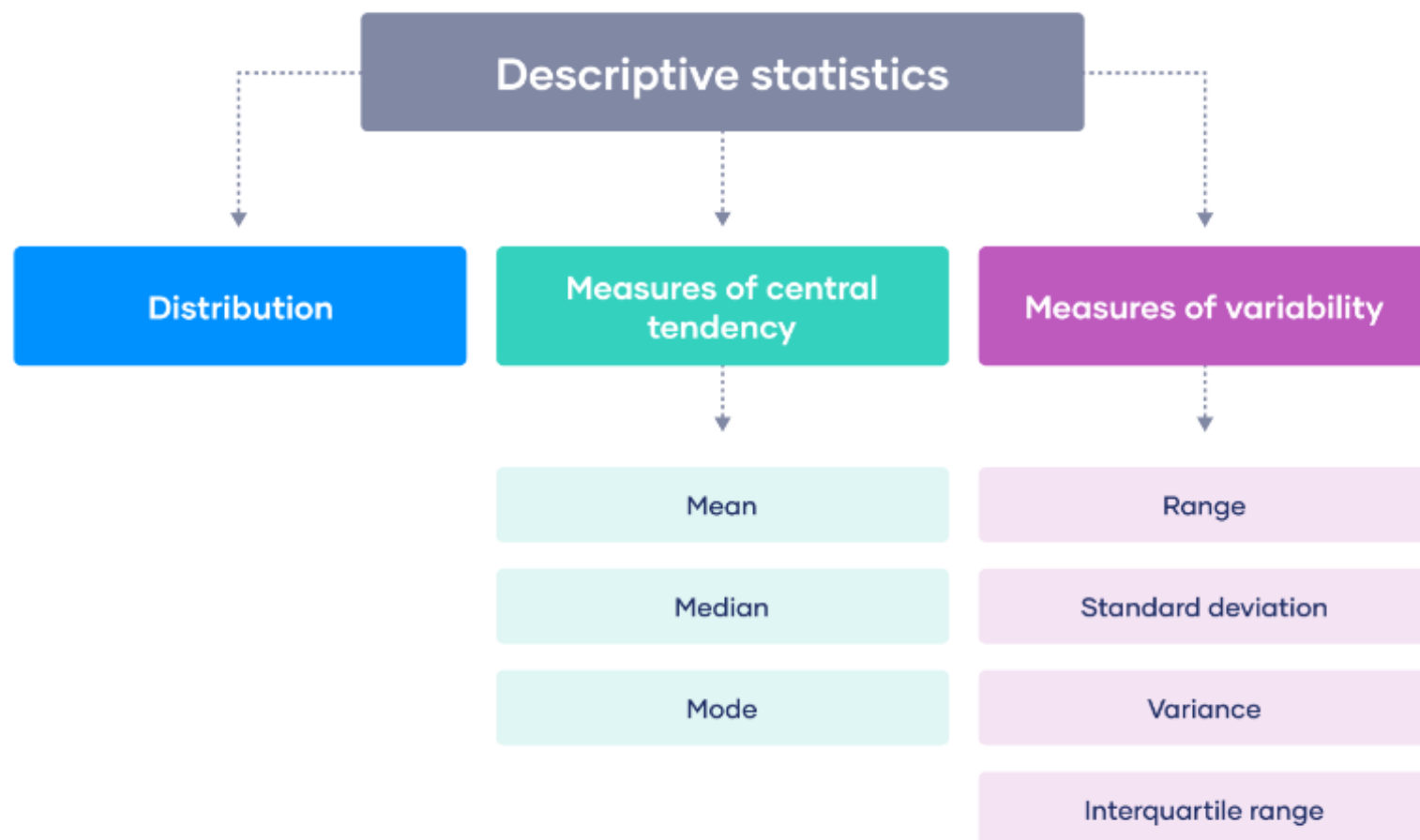
تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

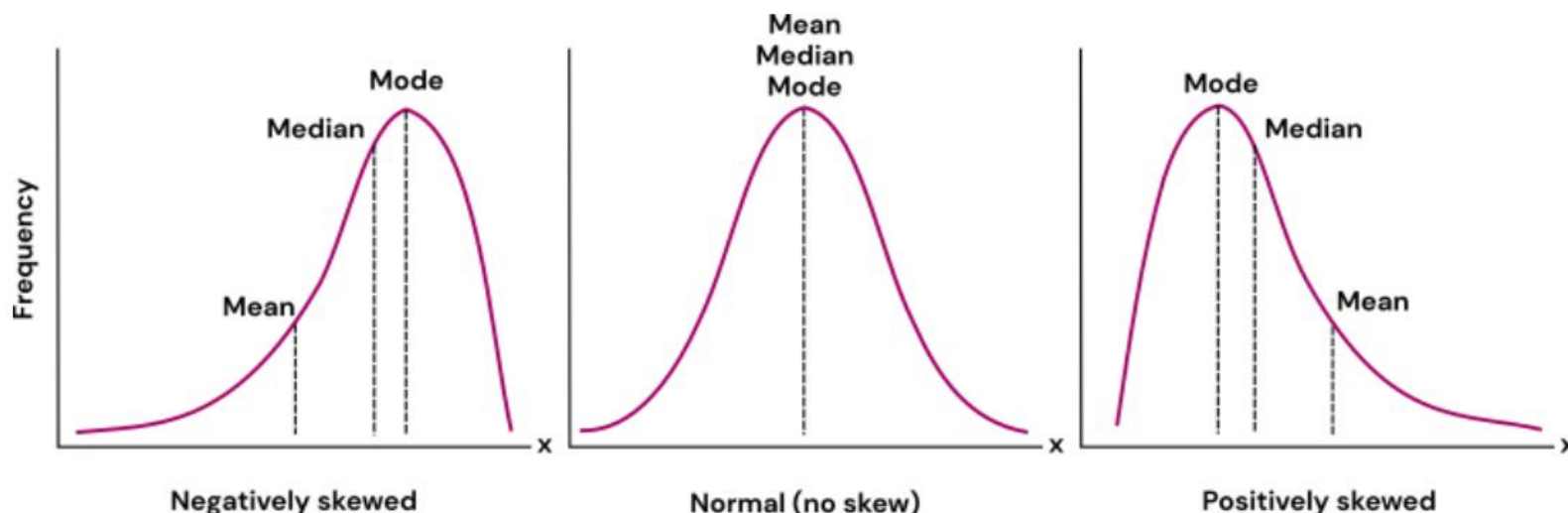
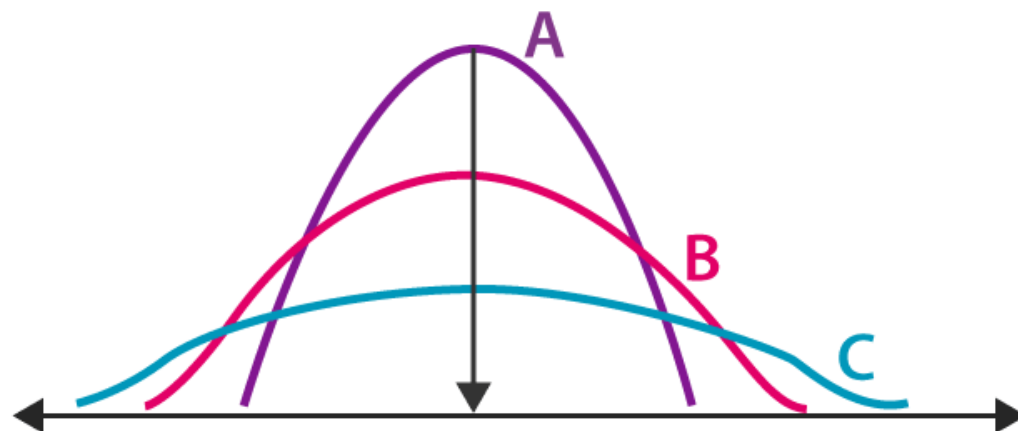
مروری بر خلاصه سازی آماری



- شاخص های مرکزی و پراکندگی
- نمودارهای بصری
- کیفیت داده ها

مروری بر خلاصه سازی آماری

- شاخص های مرکزی و پراکندگی
- نمودارهای بصری
- کیفیت داده ها



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

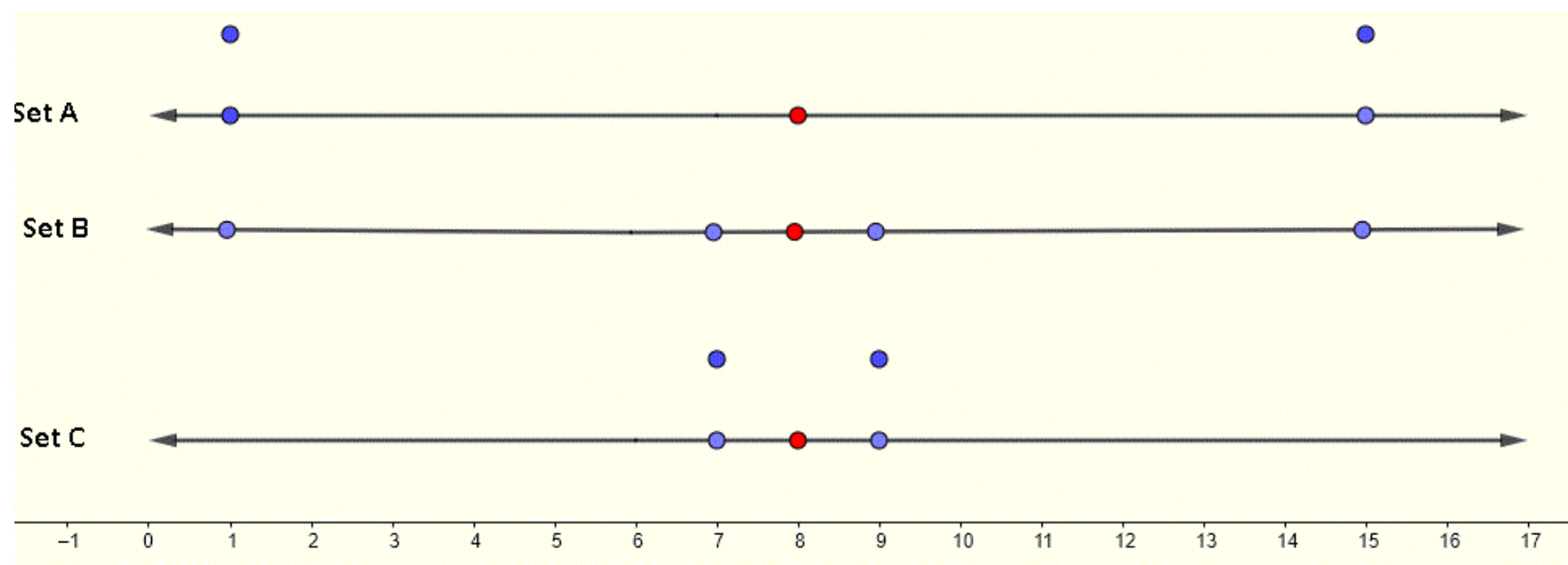
dayche.com | گروه دایکه 

مروری بر خلاصه سازی آماری

○ شاخص های مرکزی و پراکندگی

○ نمودارهای بصری

○ کیفیت داده ها



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

The formula to find the sample mean

$$\mu_x = \frac{\sum_{i=1}^n x_i}{n}$$

Formula to estimate sample standard deviation

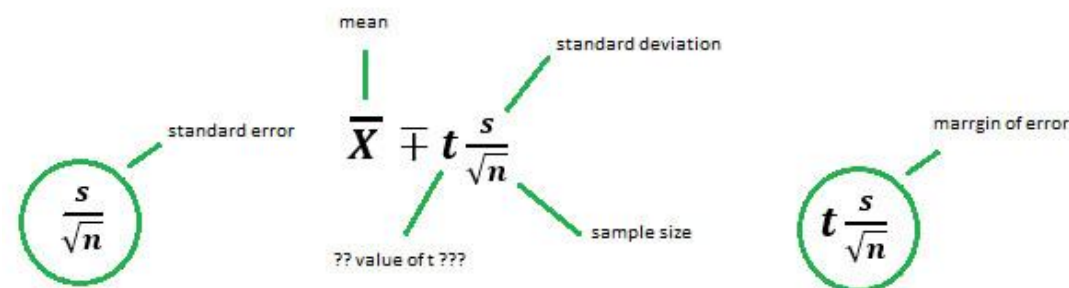
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \mu_x)^2}{n - 1}}$$

Formula to estimate **standard error (SE) of mean**

$$SE_{\mu_x} = \frac{s}{\sqrt{n}}$$

مروری بر خلاصه سازی آماری □

- شاخص های مرکزی و پراکندگی
- نمودارهای بصری
- کیفیت داده ها



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر خلاصه سازی آماری

○ شاخص های مرکزی و پراکندگی

○ نمودارهای بصری

○ کیفیت داده ها

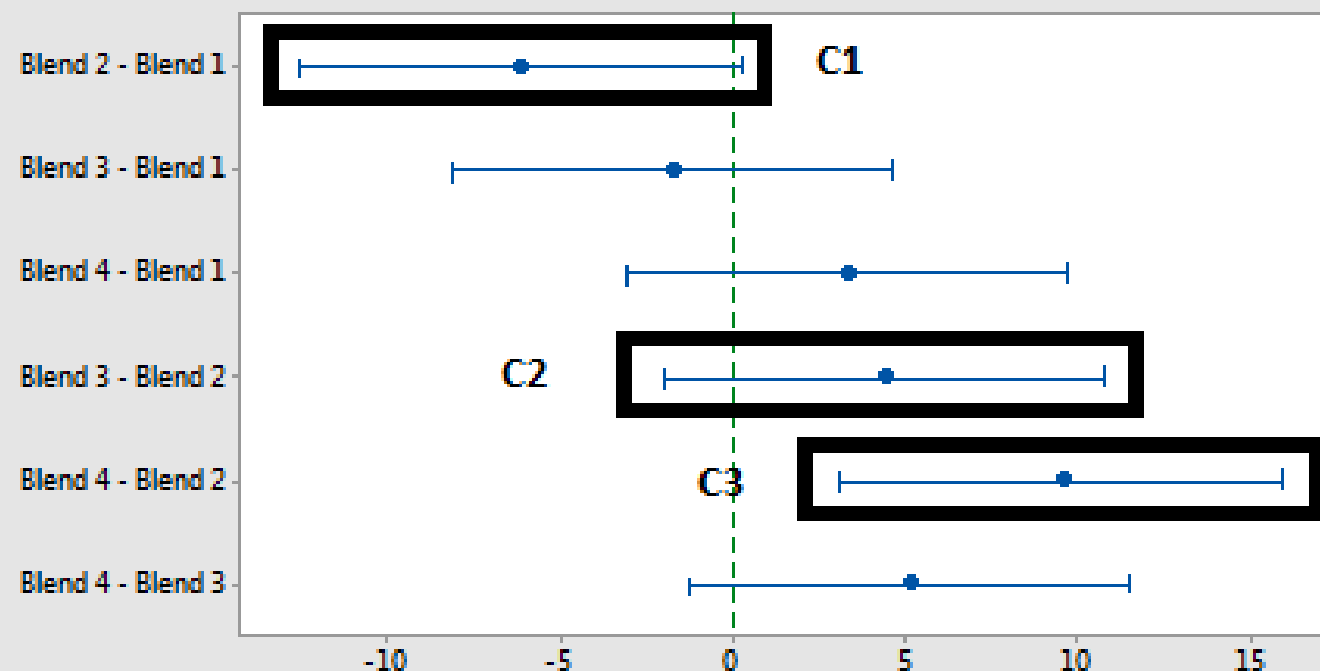
$C1 \approx C2$

$C2 \approx C3$

$C1 \not\approx C3$



Tukey Simultaneous 95% CIs
Differences of Means for Hardness



If an interval does not contain zero, the corresponding means are significantly different.

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

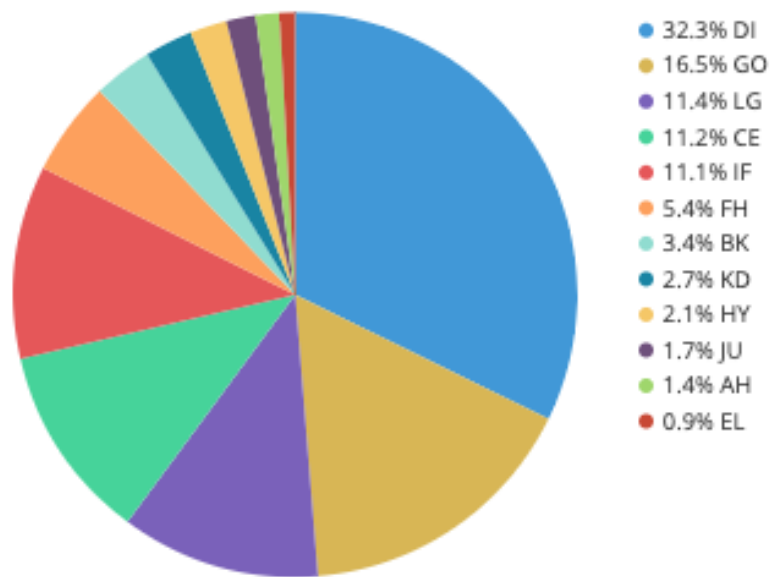
مروری بر خلاصه سازی آماری

شاخص های مرکزی و پراکندگی ○

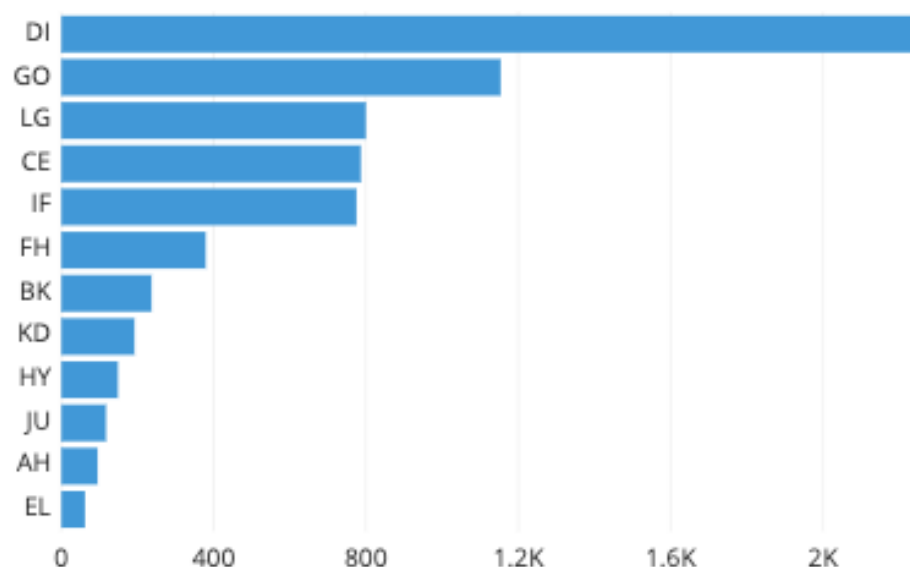
نمودارهای بصری ○

کیفیت داده ها ○

Production by district



Production by district



تولید محتوا: محمد روزبه

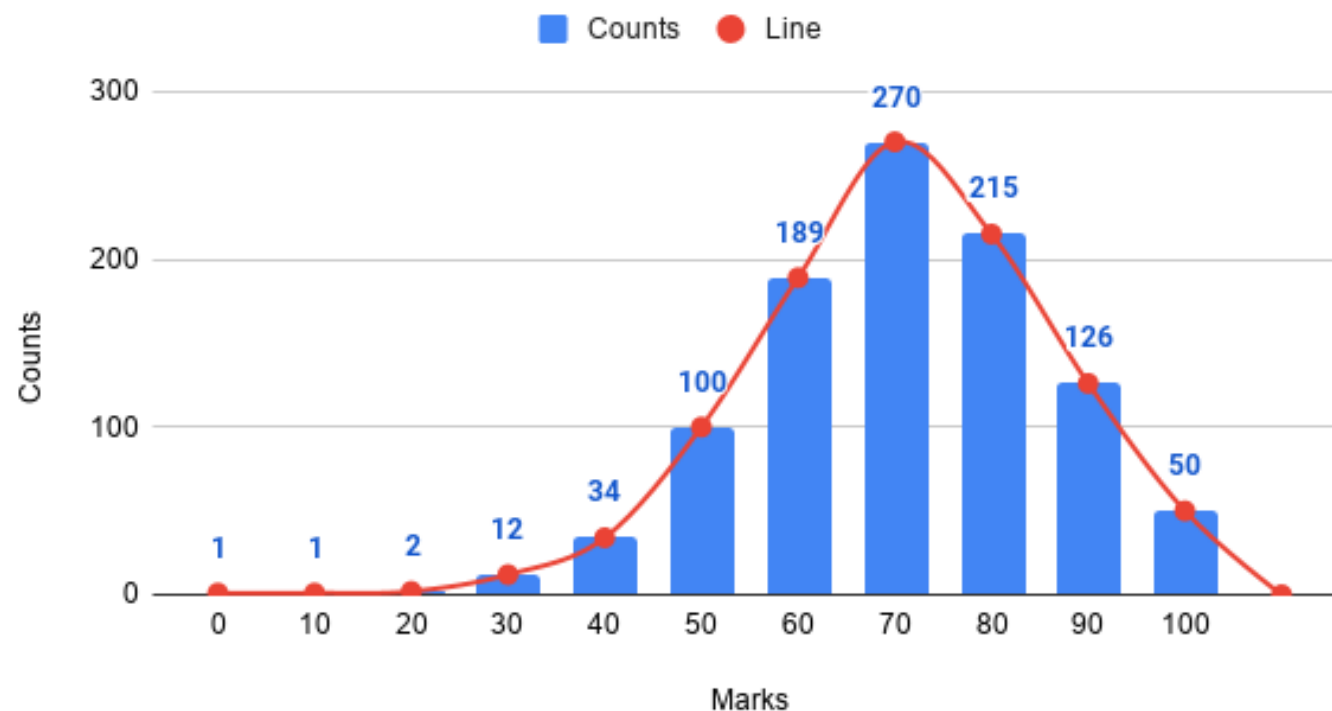
daychegroup

daychegroup

گروه دایکه | dayche.com

مروری بر خلاصه سازی آماری

Frequency chart for students marks



شاخص های مرکزی و پراکندگی ○

نمودارهای بصری ○

کیفیت داده ها ○

تولید محتوا: محمد روزبه

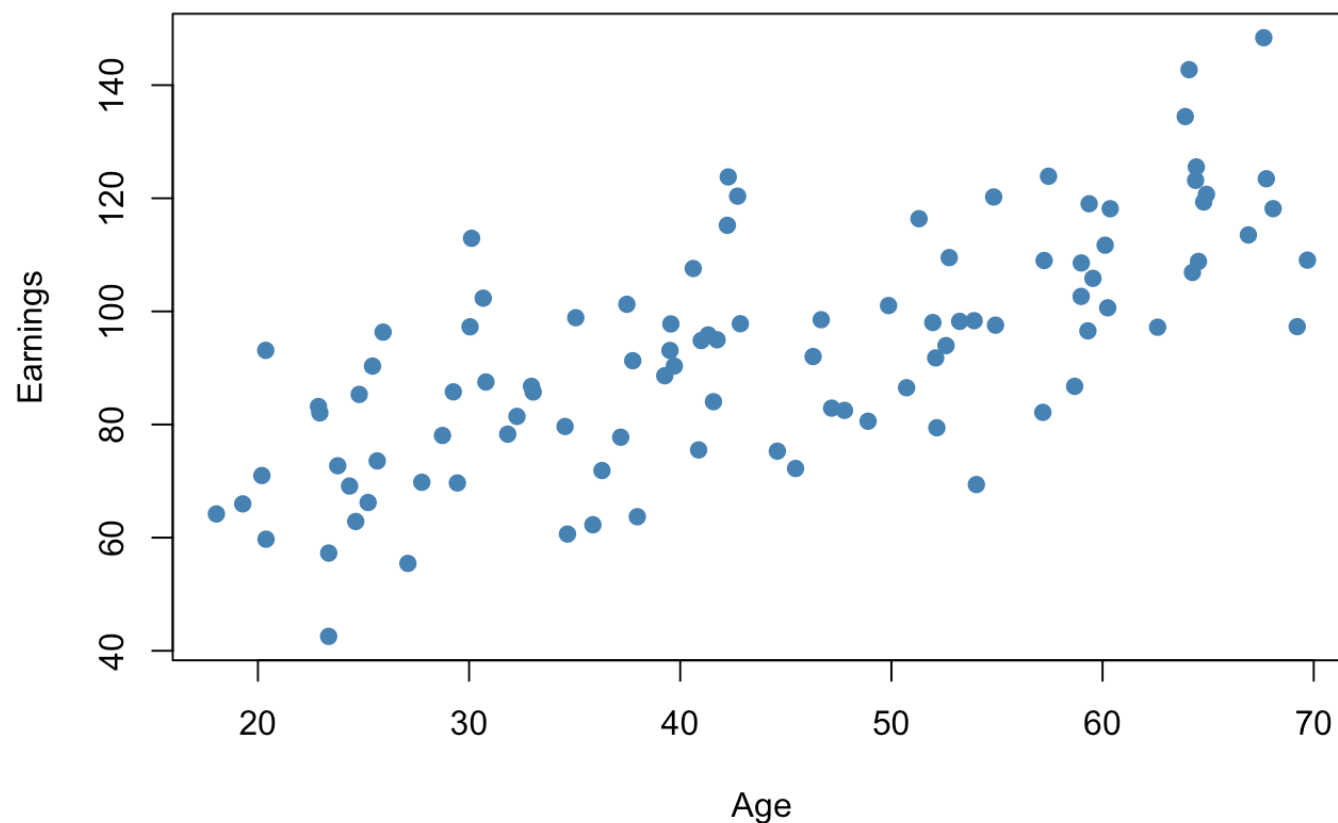
daychegroup 

daychegroup 

dayche.com | گروه دایکه 

مروری بر خلاصه سازی آماری

A Scatterplot of X and Y



شاخص های مرکزی و پراکندگی

نمودارهای بصری

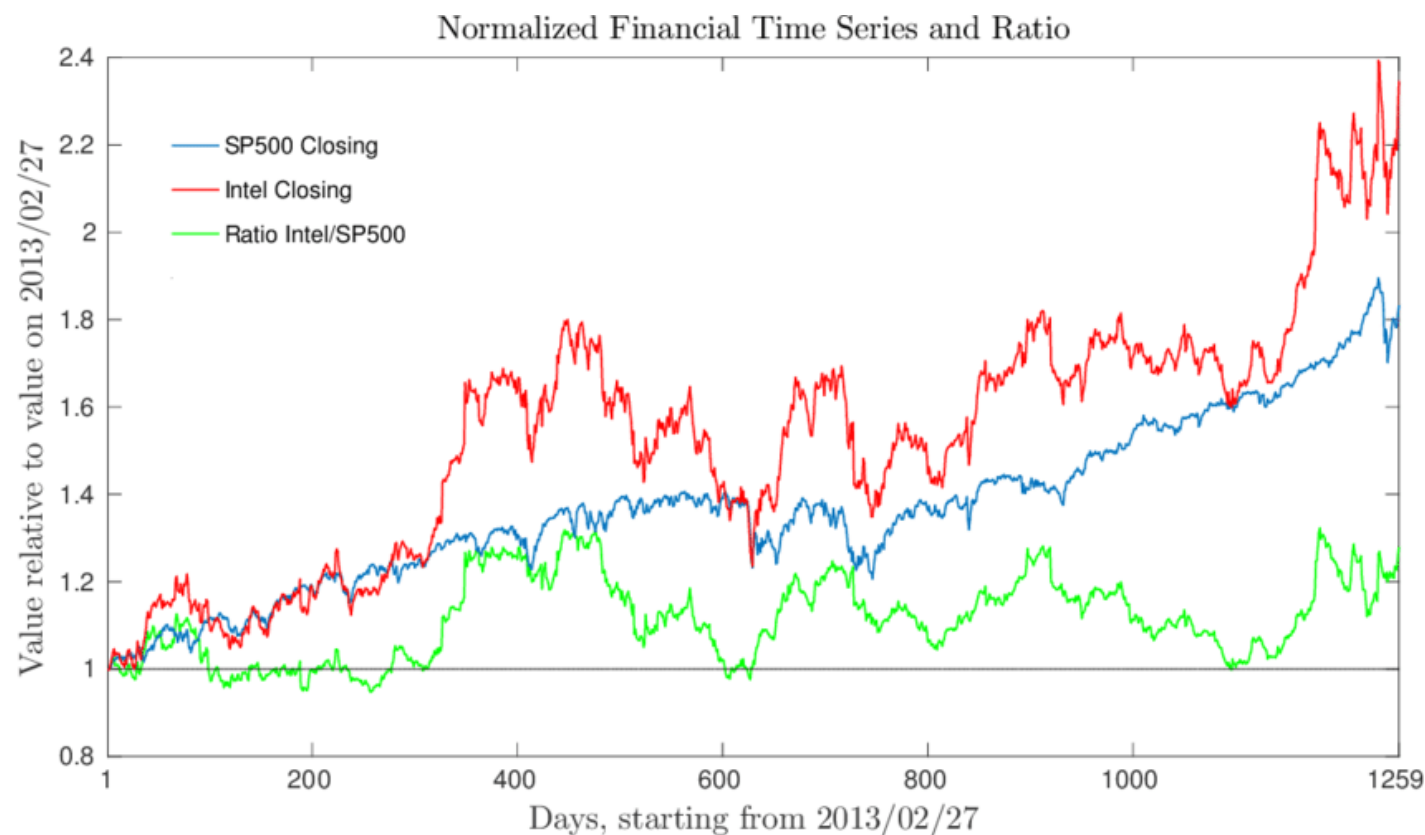
کیفیت داده ها

مروری بر خلاصه سازی آماری

○ شاخص های مرکزی و پراکندگی

○ نمودارهای بصری

○ کیفیت داده ها



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر خلاصه سازی آماری

○ شاخص های مرکزی و پراکندگی

○ نمودارهای بصری

○ کیفیت داده ها

کد مشتری	کد پستی	سن	جنسیت	درآمد (تومان)	شغل
c-25687	14546-32148	32	مرد	15.000.000	مدیر پروژه
c-62578	65489-32698	698	مرد	7.000.000	کارشناس
c-32689	32478-89543	37	زن	0	خانه دار
c-11457	2564898562		مرد	-14	آزاد
11111111	98742-25691	17			کارمند
c-19871	32659	56	زن	60.000.000	پزشک
c-36958	21698-52149	1368	زن	12.000.000	مدیرپروژه

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

مروری بر خلاصه سازی آماری

- شاخص های مرکزی و پراکندگی
- نمودارهای بصری
- کیفیت داده ها

کد مشتری	کد پستی	سن	جنسیت	درآمد (تومان)	شغل
c-25687	14546-32148	32	مرد	15.000.000	مدیر پروژه
c-62578	65489-32698	698	مرد	7.000.000	کارشناس
c-32689	32478-89543	37	زن	0	خانه دار
c-11457	2564898562		مرد	-14	آزاد
11111111	98742-25691	17			کارمند
c-19871	32659	56	زن	60.000.000	پزشک
c-36958	21698-52149	1368	زن	12.000.000	مدیر پروژه

داده های گمشده

Missing Data

داده های پرت

Outliers

ناسازگار

Inconsistent

خارج از بازه منطقی

Out of Logical Range

مروری بر خلاصه سازی آماری

○ شاخص های مرکزی و پراکندگی

○ نمودارهای بصری

○ کیفیت داده ها

- روش حذفی
- روش جانشینی

داده های گمشده

Missing Data

- نقاط پرت (تک متغیره)
- رکورد پرت (چند متغیره)

داده های پرت

Outliers

ناسازگار

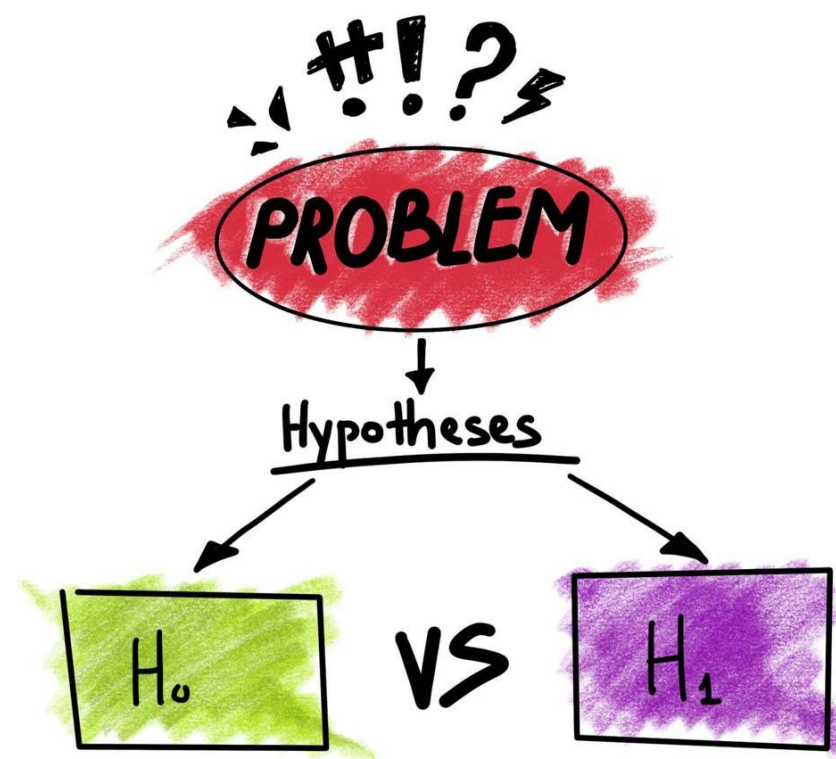
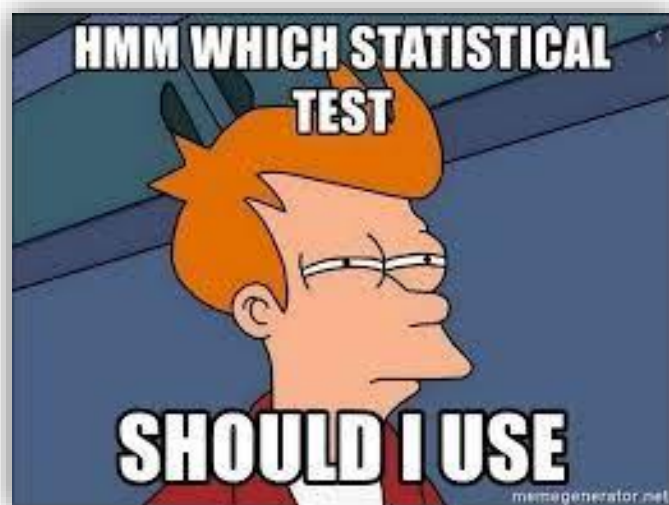
Inconsistent

خارج از بازه منطقی

Out of Logical Range

مروری بر ارتباط سنجی آماری

آزمون های فرض (Hypotheses Test)



تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com  گروه دایکه

مروری بر ارتباط سنجی آماری

آزمون های فرض (Hypotheses Test)

- فرض صفر حالت خنثی و عدم معناداری و عدم ارتباط فیلهاست. مانند استقلال دو فیلد یا برابری میانگین ها

در نظر گرفتن فرض صفر و فرض مقابل

- میزان آلفا معمولا 0/05 در نظر گرفته می شود. با افزایش آلفا تصمیم گیری سهل گیرانه تر و با کاهش آلفا تصمیم گیری سخت گیرانه تر خواهیم داشت.

تعیین میزان خطای آلفا

- آزمون مناسب با توجه به نوع فیلهای مورد بررسی انتخاب می شود.

انجام آزمون و بدست آوردن p-value

- اگر p-value کمتر از آلفا باشد فرض صفر رد می شود و رابطه معنادار است و اگر بیشتر از آلفا باشد فرض صفر رد نمی شود و رابطه معنادار نیست.

مقایسه p-value با خطای آلفا و نتیجه گیری

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر ارتباط سنجی آماری

آزمون های فرض (Hypotheses Test)

$$\begin{cases} H_0: \text{Sex and Drug are independent} \\ H_1: \text{Sex and Drug are NOT independent} \end{cases}$$

Contingency Table:

Sex	F	M
Drug		
drugA	9	14
drugB	6	10
drugC	7	9
drugX	27	27
drugY	47	44

Chi-Squared Statistic: 2.119

Degrees of Freedom: 4

p-value: 0.7138369773987128

$$\begin{cases} H_0: \text{Corr}(Na, K) = 0 \\ H_1: \text{Corr}(Na, K) \neq 0 \end{cases}$$

Correlation between Na and K:

Pearson's correlation coefficient: 0.0173 (p-value: 0.8075)

Spearman's rank correlation coefficient: -0.0052 (p-value: 0.9417)

Correlation between Na and Age:

Pearson's correlation coefficient: 0.1004 (p-value: 0.1573)

Spearman's rank correlation coefficient: 0.0931 (p-value: 0.1900)

$$\begin{cases} H_0: \text{Corr}(Na, Age) = 0 \\ H_1: \text{Corr}(Na, Age) \neq 0 \end{cases}$$

تولید محتوا: محمد روزبه

daychegroup 

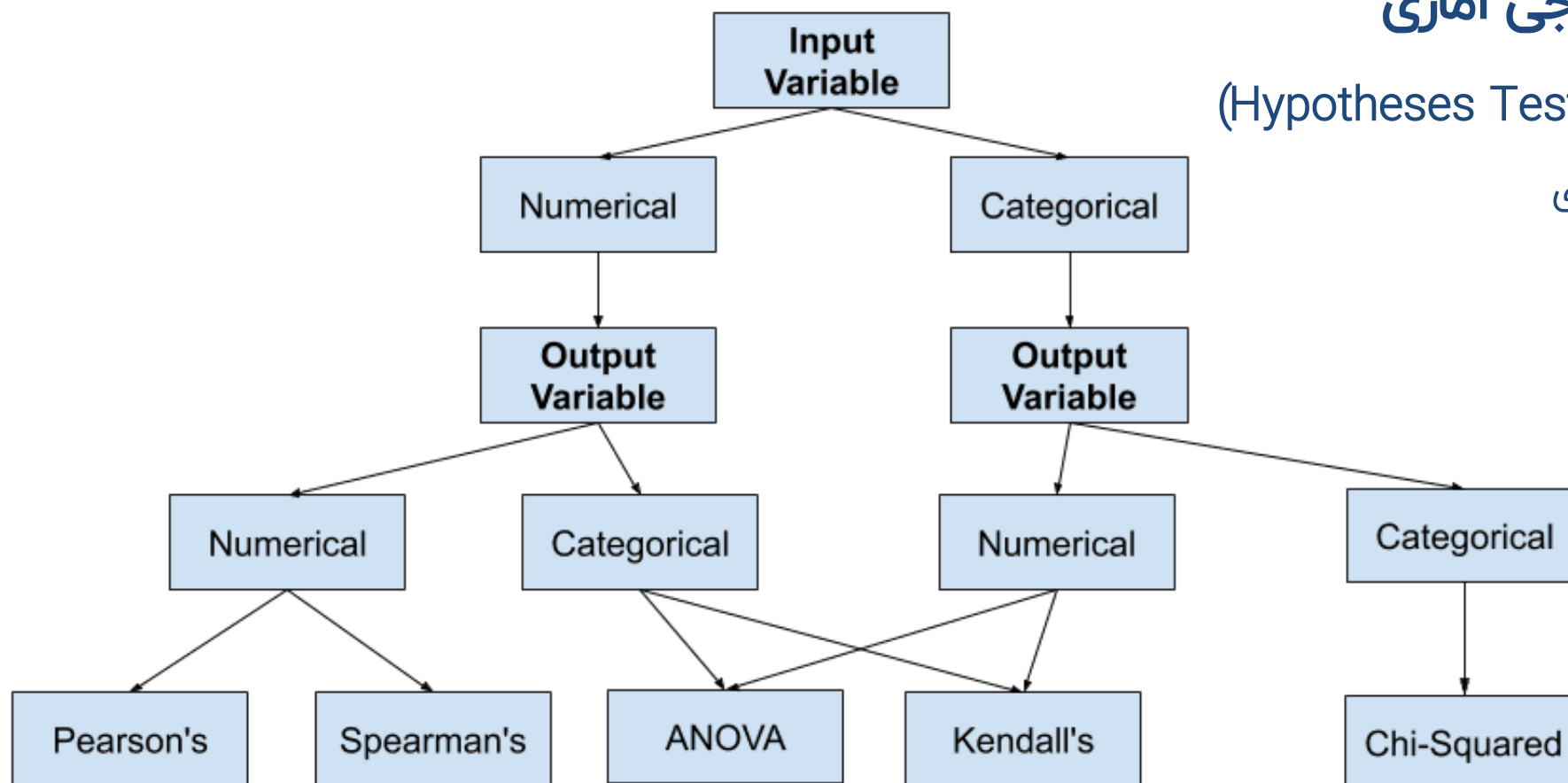
daychegroup 

dayche.com | گروه دایکه 

مروری بر ارتباط سنجی آماری

آزمون های فرض (Hypotheses Test)

انتخاب آزمون آماری



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه

مواردی که یاد گرفتیم..



گروه دایکه . dayche.com



□ مروری بر فرآیند پیاده سازی علم داده

فرآیند داده کاوی

مدلسازی و ارزیابی

شناخت و آماده سازی داده ها

□ مروری بر فرآیند پیاده سازی علم داده

فرآیند داده کاوی

مدلسازی و
ارزیابی

شناخت و آماده سازی داده ها

یکپارچه سازی
داده ها

کاهش ابعاد و
انتخاب نمونه

تبدیل داده ها

کیفیت داده ها

توصیف و کاوش
داده ها

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

مروری بر فرآیند پیاده سازی علم داده

شناخت و آماده سازی داده ها

تبدیل داده ها

یکپارچه سازی داده ها

کاهش ابعاد و انتخاب نمونه

کیفیت داده ها

توصیف و کاوش داده ها

هموار سازی

تجمیع و فشرده سازی

گسسته سازی

شاخص سازی

نرمالسازی

تولید محتوا: محمد روزبه

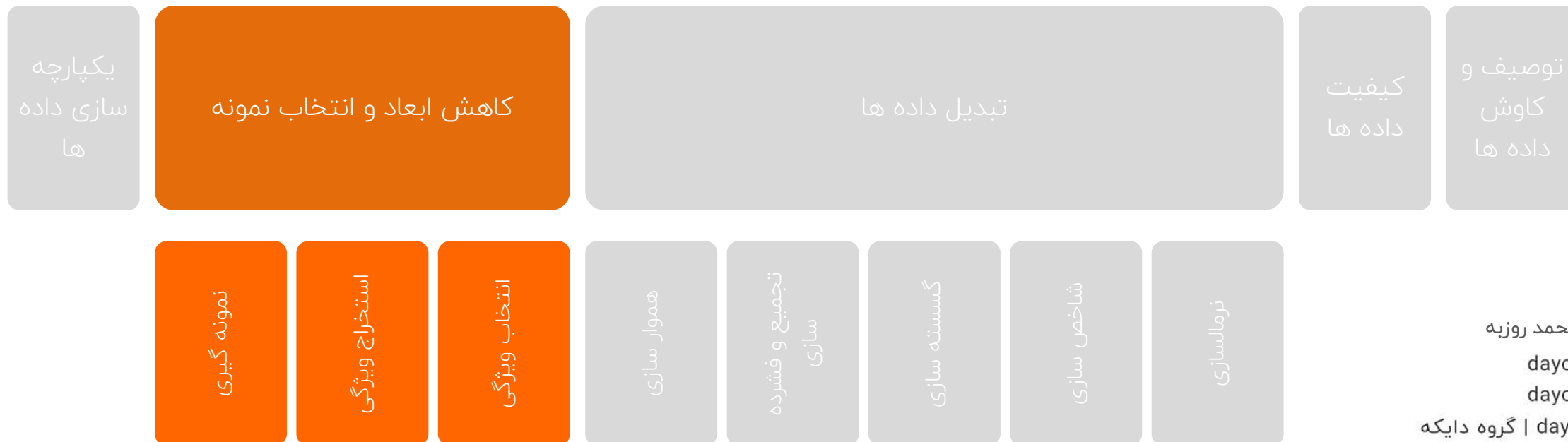
daychegroup

daychegroup

dayche.com | گروه دایکه

مروری بر فرآیند پیاده سازی علم داده

شناخت و آماده سازی داده ها



تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



❑ مروری بر فرآیند پیاده سازی علم داده

مدلسازی و ارزیابی

مدل های اکتشافی

مدل های پیش بینانه

ترکیب مدل ها

داده های نامتوازن

روش های ارزیابی

انواع الگوریتم ها

تولید محتوا: محمد روزبه

daychegroup 

daychegroup 

dayche.com | گروه دایکه 

□ مروری بر فرآیند پیاده سازی علم داده

مدلسازی و ارزیابی

مدل های اکتشافی

مدل های پیش بینانه

روش های ارزیابی

قوانین انجمنی

انواع خوشه بندی

ترکیب مدل ها

داده های نامتوازن

روش های ارزیابی

انواع الگوریتم ها

تولید محتوا: محمد روزبه

daychegroup

daychegroup

dayche.com | گروه دایکه



گروه داده کاوی دایکه | دای که
شرکت ندای کهن پرداز DAYCHE

تلفن: ۰۹۱۹۰۲۱۴۴۸۹

dayche.com info@dayche.com