



Namatek
True Education



آشنایی با علم داده

فهرست مطالب

1. علم داده چیست؟ (Data Science)
2. ارکان تخصص علم داده
3. نمودارهای ون علم داده (The Data Scientist Venn Diagram)
4. اهداف و داده های علم داده
5. فرآیند Data Science

استخراج دانش و بینش از مجموعه ای از اطلاعات توسط علم داده امکان پذیر است که در بسیاری از سیستم های کاری مورد استفاده و کاربردی خواهد بود. Data Science یک علم نوظهور است که طرفداران بسیار زیادی را در مدتی اندک به خود جذب کرده است. اگر شما هم جزو علاقه مندان به آشنایی با این علم هستید این مقاله پاسخی برای سوال های شماست.

#1 علم داده چیست؟ (Data Science)

علم داده یک حوزه بین رشته ای محسوب می شود که از روش ها، فرآیندها، الگوریتم ها و سیستم های علمی برای استخراج دانش و بینش از بسیاری از داده های ساختاری و غیر ساختاری استفاده می کند.

در حقیقت علم داده علمی است که به داده کاوی، یادگیری ماشین و داده های بزرگ مربوط می شود. علم داده را می توان مفهومی برای متحد کردن آمار، تجزیه و تحلیل داده ها و روش های مربوط به آنها، به منظور درک و تجزیه و تحلیل پدیده های واقعی با استفاده از داده به شمار آورد.

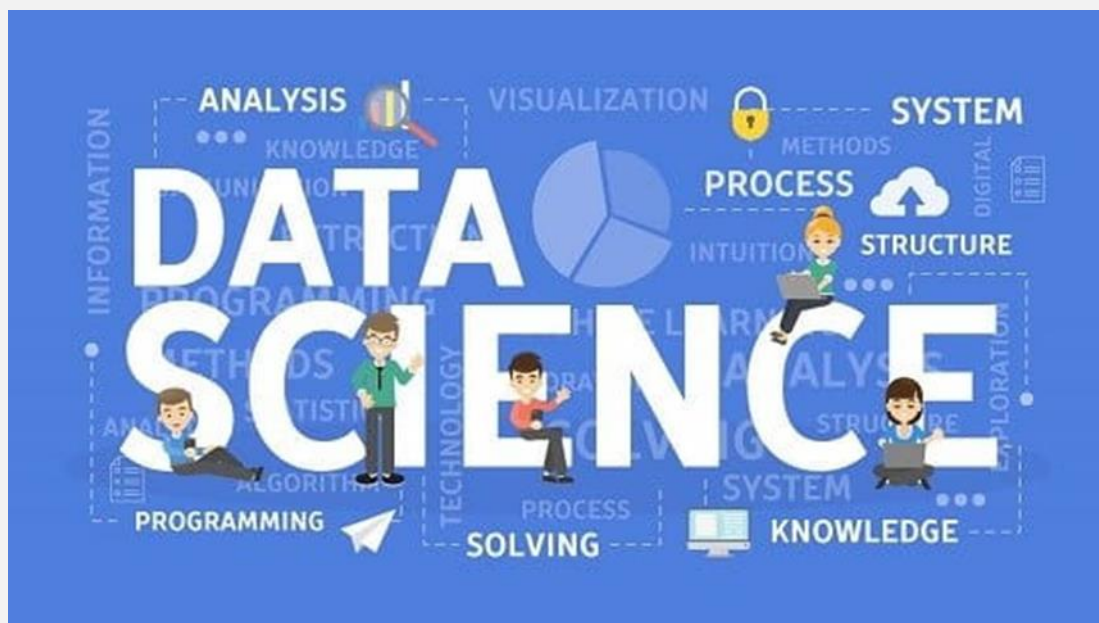
این حوزه از تکنیک ها و نظریه های گرفته شده از زمینه های مختلف در زمینه ریاضیات، آمار، علوم کامپیوتر، دانش دامنه و علوم اطلاعات استفاده می کند.

Data Science



جیم گری برنده جایزه تورینگ، علم داده را الگوی چهارم علم (تجربی، نظری، محاسباتی و اکنون مبتنی بر داده) دانسته و اظهار داشت که "همه چیز درباره علم به دلیل تأثیر فناوری اطلاعات در حال تغییر است" و با سیلی از داده ها مواجه هستیم.

#2 ارکان تخصص علم داده



در حالی که دانشمندان داده اغلب از زمینه های مختلف آموزشی و تجارب کاری استفاده می کنند، بیشتر آن ها باید در چهار زمینه اساسی قوی باشند یا در یک حالت ایده آل متخصص باشند. بدون هیچ ترتیب اولویت یا اهمیت خاصی، این موارد عبارتند از:

1. تجارت / دامنه
2. ریاضیات (شامل آمار و احتمال)
3. علوم کامپیوتر (به عنوان مثال، معماری و مهندسی نرم افزار / داده)
4. ارتباطات (کتبی و کلامی)

مهارت ها و تخصص های دیگری نیز وجود دارد که بسیار مطلوب هستند اما این چهار مهارت اصلی از اهمیت بیشتری برخوردار هستند. در ادامه

این مقاله از این عناصر به عنوان مهارت اصلی برای دانشمندان در حوزه علم داده استفاده می شود.

در حقیقت، مردم اغلب در یک یا دو موضوع قوی هستند اما معمولاً در هر چهار مورد به یک اندازه قوی نیستند. بر اساس این ارکان، دانشمند داده شخصی است که باید بتواند از منابع داده موجود استفاده کند و در صورت لزوم منابع جدیدی را برای استخراج اطلاعات معنی دار و بینش های عملی ایجاد کند.

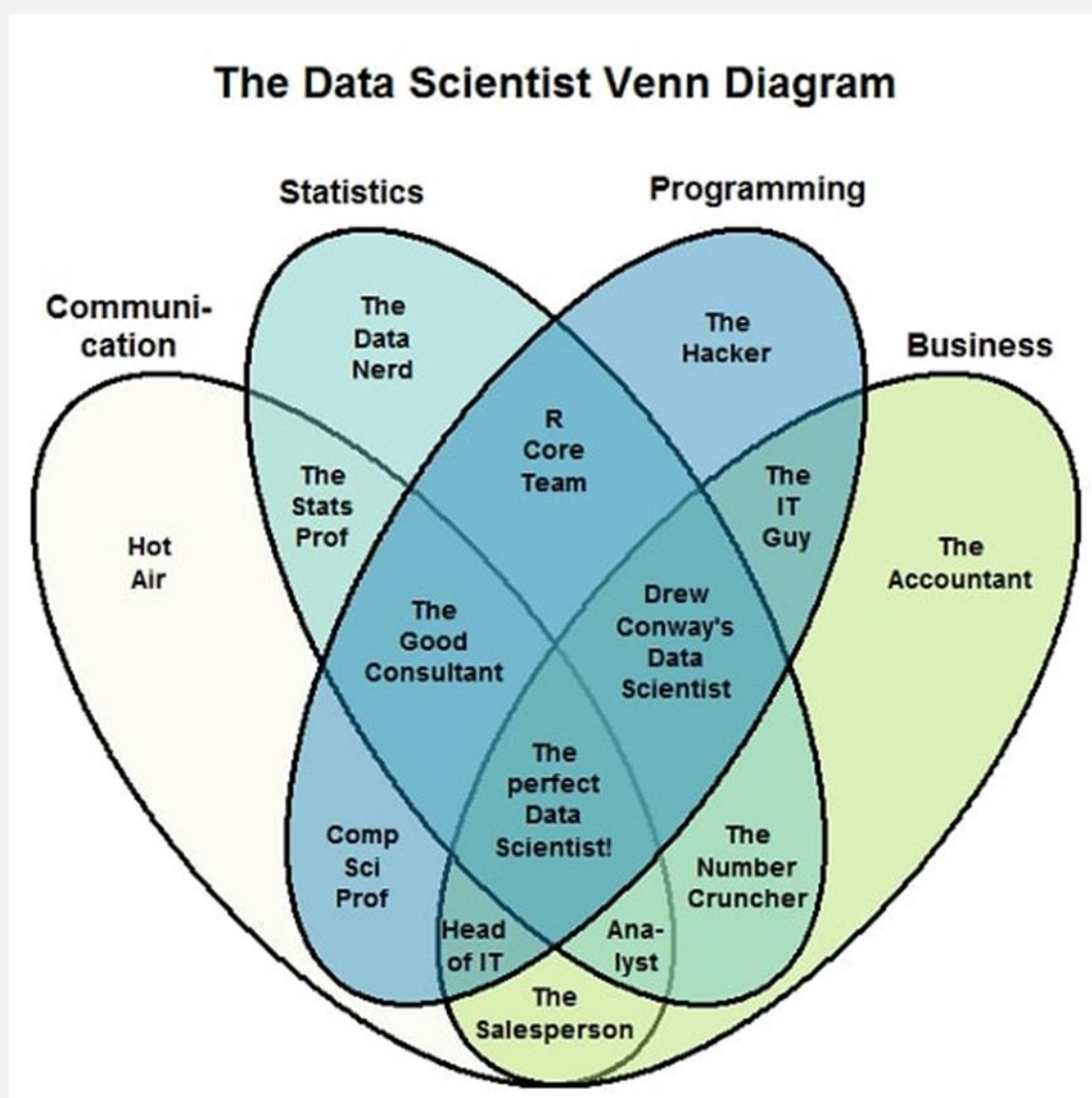
از این بینش ها می توان برای هدایت تصمیمات و تغییرات تجاری جهت دستیابی به اهداف تجاری استفاده کرد. این کار از طریق تخصص حوزه کسب و کار، ارتباطات موثر و تفسیر نتایج و استفاده از همه تکنیک های آماری مربوطه، زبان های برنامه نویسی، بسته های نرم افزاری و کتابخانه ها، زیرساخت های داده و غیره انجام می شود.

#3 نمودارهای ون علم داده (The Data Scientist Venn Diagram)

می توان نسخه های مختلفی از نمودار دانش داده را برای کمک به تجسم موارد این ارکان و روابط آن ها با یکدیگر پیدا کرد.

دیوید تیلور مقاله ای عالی در مورد این نمودارهای ون با عنوان، نبرد
 دیاگرام های ون در علم داده (Battle of the Data Science Venn
 Diagrams) نوشته است.

در تصویر زیر یکی از نمودارهای ون که توسط استفان کولاسا ایجاد شده
 است، مشاهده خواهید کرد که بیضی اولیه در نمودار بسیار شبیه ارکان
 ذکر شده در بالا است.



این نمودار و موارد دیگر مانند آن، سعی در تعیین برجسب ها و یا مشخص کردن شخص یا زمینه ای است که در هر یک از ارکان همپوشانی دارند.

همانطور که این نمودار نشان می دهد، استفان کولاسا برجسب "The Perfect Data Scientist" را به فردی اختصاص می دهد که در تجارت، برنامه نویسی، آمار و ارتباطات به یک اندازه قوی است.

#4 اهداف و داده های علم داده

برای درک اهمیت این ارکان، ابتدا باید اهداف معمولی و قابل بررسی مرتبط با ابتکارات علوم داده و همچنین فرآیند علم داده را درک کرد.

بهتر است ابتدا در مورد برخی از اهداف و داده های مشترک علم داده بحث کنیم. در اینجا لیستی کوتاه از مطالب قابل ارائه در زمینه علوم داده وجود دارد:

- پیش گویی (پیش بینی یک مقدار بر اساس ورودی)
- طبقه بندی (به عنوان مثال داده هرزنامه هست یا هرزنامه نیست)
- توصیه ها (به عنوان مثال، توصیه های آمازون و Netflix)
- شناسایی و گروه بندی الگوها (به عنوان مثال، طبقه بندی بدون کلاس های شناخته شده)
- تشخیص ناهنجاری (به عنوان مثال، تشخیص تقلب)

- تشخیص (تصویر، متن، صدا، فیلم، صورت و غیره)
- بینش عملی (از طریق گزارش، تجسم و غیره)
- فرآیندهای خودکار تصمیم گیری (به عنوان مثال، تأیید کارت اعتباری)
- امتیازدهی و رتبه بندی (به عنوان مثال، نمره FICO)
- تقسیم بندی (به عنوان مثال، بازاریابی مبتنی بر جمعیت)
- بهینه سازی (به عنوان مثال، مدیریت ریسک)
- پیش بینی ها (به عنوان مثال، فروش و درآمد)

هر یک از این ها برای دستیابی به یک هدف خاص و یا حل یک مشکل خاص در نظر گرفته شده است.



#4-1 هدف واقعی کدام است؟

سوال واقعی این است که کدام هدف است و هدف چیست؟

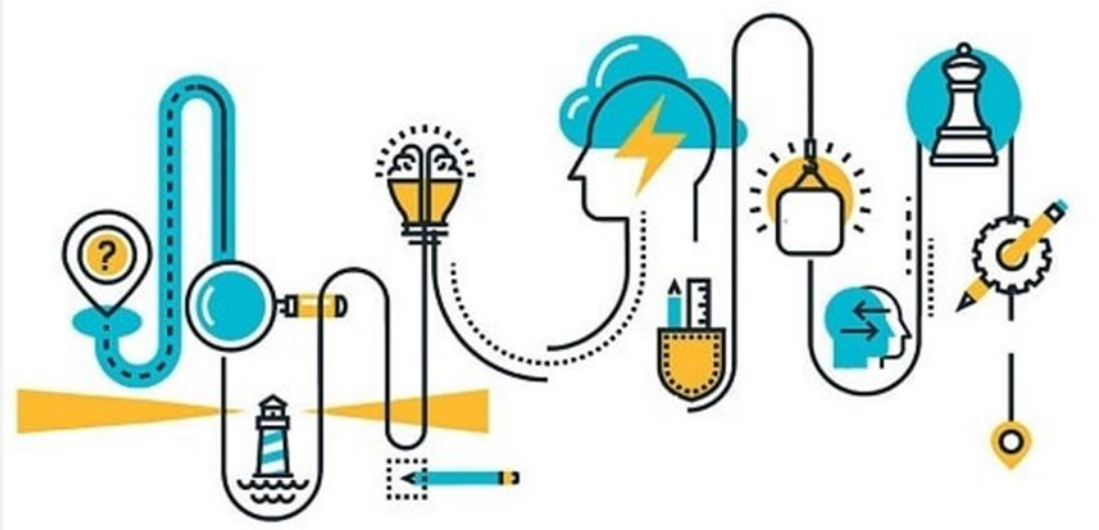
به عنوان مثال، یک دانشمند داده ممکن است فکر کند که هدف او ایجاد یک موتور پیش بینی با عملکرد بالا است. از طرف دیگر شغلی که قصد دارد از موتور پیش بینی استفاده کند، ممکن است هدفش افزایش درآمد باشد که می تواند با استفاده از این موتور پیش بینی حاصل شود.

گرچه به نظر می رسد این مسئله در نگاه اول مسئله مهمی نباشد اما در واقع وضعیت توصیف شده دلیل اصلی بودن رکن اول (تخصص حوزه کسب و کار) است. اغلب اعضای مدیریت عالی دارای سوابق تحصیلی کسب و کار محور مانند MBA هستند.

با وجود این که بسیاری از مدیران افراد فوق العاده باهوشی هستند اما ممکن است در تمام ابزارها، تکنیک ها و الگوریتم های در دسترس دانشمند داده (به عنوان مثال، تجزیه و تحلیل آماری، یادگیری ماشین، هوش مصنوعی و غیره) مهارت کافی نداشته باشند.

با توجه به این، آن ها ممکن است نتوانند به دانشمند داده نیازها و داده ها مانند منابع داده، ویژگی ها (متغیرها) و مسیر رسیدن به آنجا را پیشنهاد دهند. حتی اگر یک مدیر اجرایی بتواند تشخیص دهد که یک موتور توصیه دهنده خاص به افزایش درآمد کمک می کند، ممکن است متوجه نشود که احتمالاً روش های زیادی وجود دارد که می توان از داده های شرکت برای افزایش درآمد نیز استفاده کرد. بنابراین نمی توان به اندازه کافی تأکید کرد که دانشمند ایده آل داده نسبت به عملکرد کلی کسب و کارها و نحوه استفاده از داده های شرکت برای دستیابی به اهداف سطح بالا درک نسبتاً کاملی دارد.

#5 فرآیند Data Science



در زیر نمودار مدل فرآیند GABDO آورده شده است که در کتاب هوش مصنوعی برای افراد و مشاغل (AI for People and Business) ایجاد و معرفی شده است.

دانشمندان داده معمولاً فرآیندی شبیه به این را دنبال می کنند، خصوصاً هنگام ایجاد مدل با استفاده از یادگیری ماشین و تکنیک های مربوطه. مدل فرآیند GABDO از پنج مرحله تکراری تشکیل شده است:

1. اهداف (Goals)
2. دستیابی (Acquire)
3. ساخت (Build)
4. تحویل (Deliver)
5. بهینه سازی (Optimize)

از این رو با نام اختصاری GABDO نشان داده می شود. هر مرحله تکراری است زیرا هر مرحله می تواند به یک یا چند مرحله قبل برگردد.