نقش علم داده ها و هوش مصنوعی در سیاست گذاری و حکمرانی منطقه ای

کورش پرند

استاد تمام گروه علوم داده ها و کامپیوتر دانشکده ریاضی

دانشگاه شهید بهشتی

آزمایشگاه هوش مصنوعی و محاسبات علمی

بهمن ۱۴۰۱

علم داده ها
و داده کاوی

داده و علم داده ها

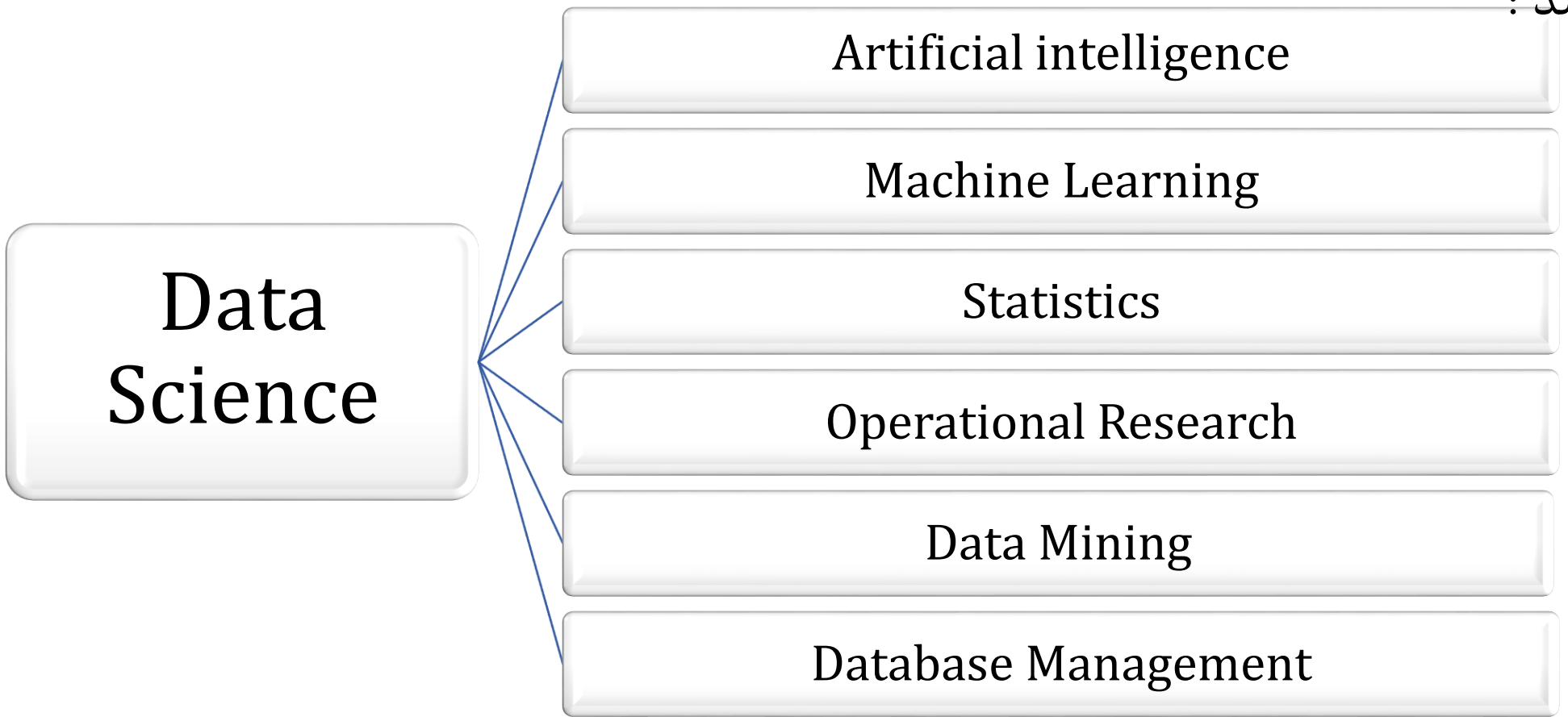
علم داده ها (Data Science) چیست ؟

علم داده با گسترش استفاده از فناوری اطلاعات و سیستم های اطلاعاتی مورد توجه سازمان ها قرار گرفته است.

علم داده با کاویدن و استخراج از منابع عظیم داده (Big Data) و پایگاه های داده (Data Base) ، اطلاعات گران بهایی که در حجم انبوهی از اطلاعات سطحی پنهان شده است را استخراج (Mining) می کند.

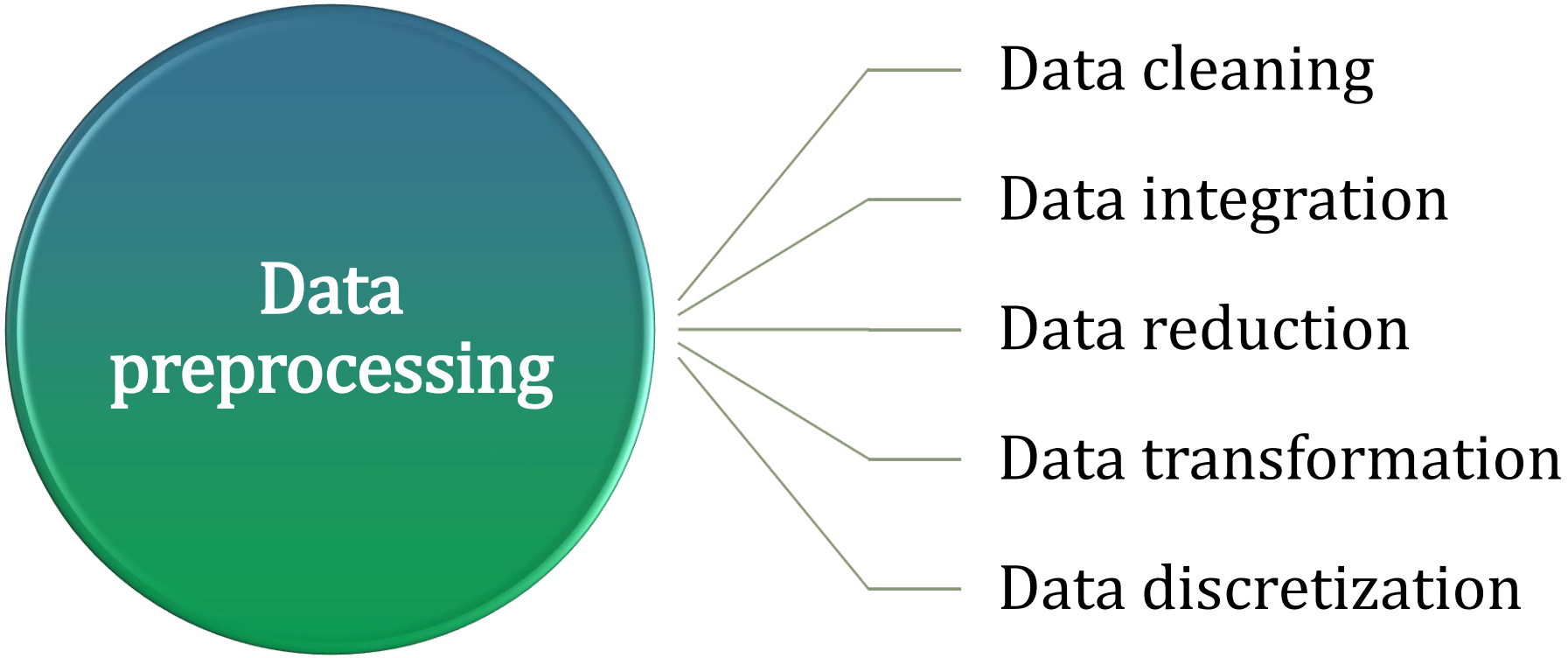
علم داده ها (Data Science) چیست ؟

علم داده ها از علوم زیر برای ساخت مدل‌ها و پاسخ به سوالات بهره می‌برد که هریک به تفصیل قابل بررسی هستند :



علم داده (Data Science) : پیش پردازش داده

اغلب پیش از کار با داده‌ها و به‌کار بردن مدل‌ها و روش‌ها، پیش‌پردازش روی داده‌ها انجام می‌شود.

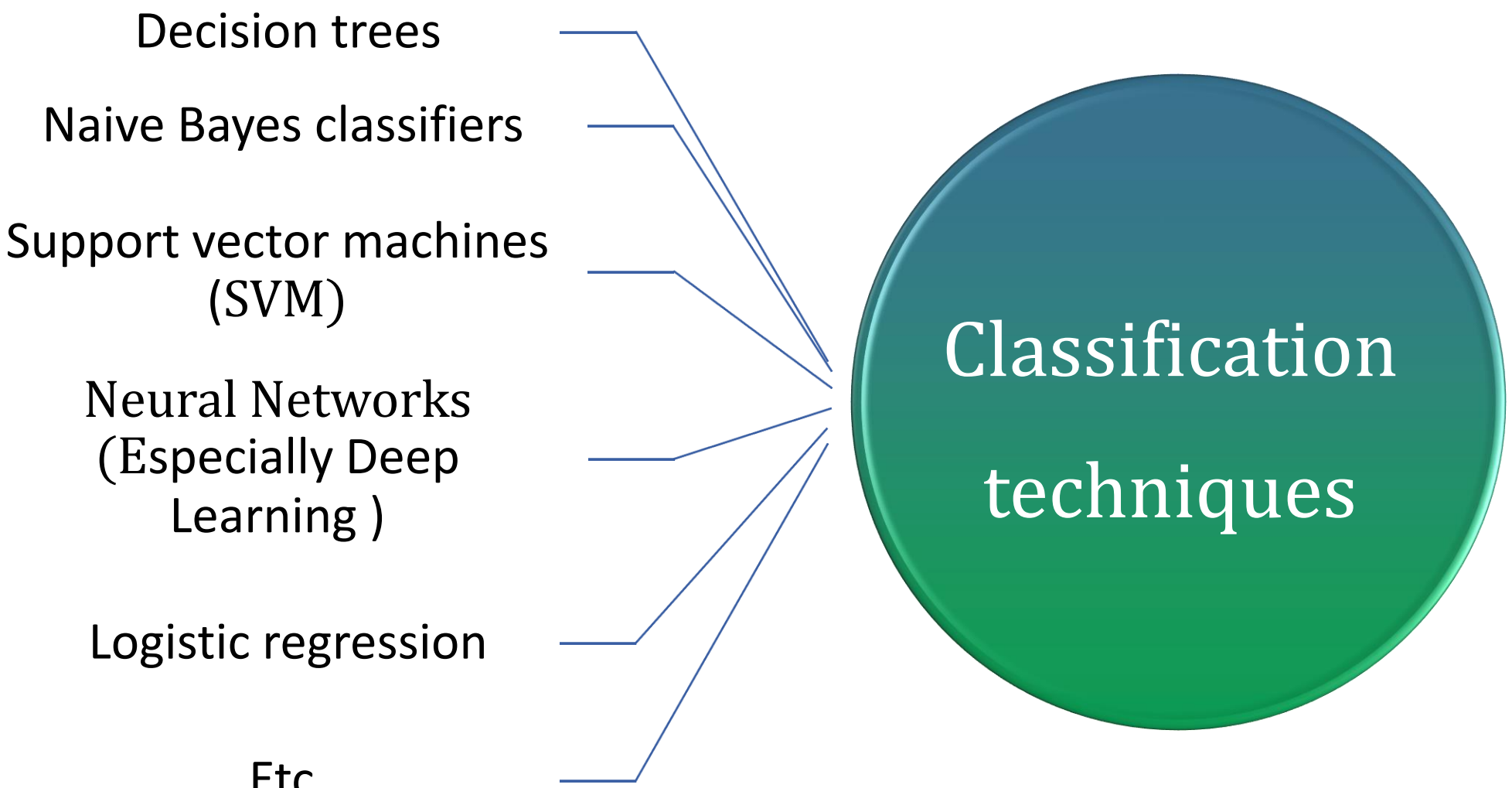


علم داده (Data Science) : ابزارها

درحوزه علم داده، مسائل مختلف با برخی ابزارهای خاص خود قابل حل هستند از جمله :

- خوشه‌بندی
- رگرسیون
- طبقه‌بندی

علم داده (Data Science) : ابزارهای طبقه بندی

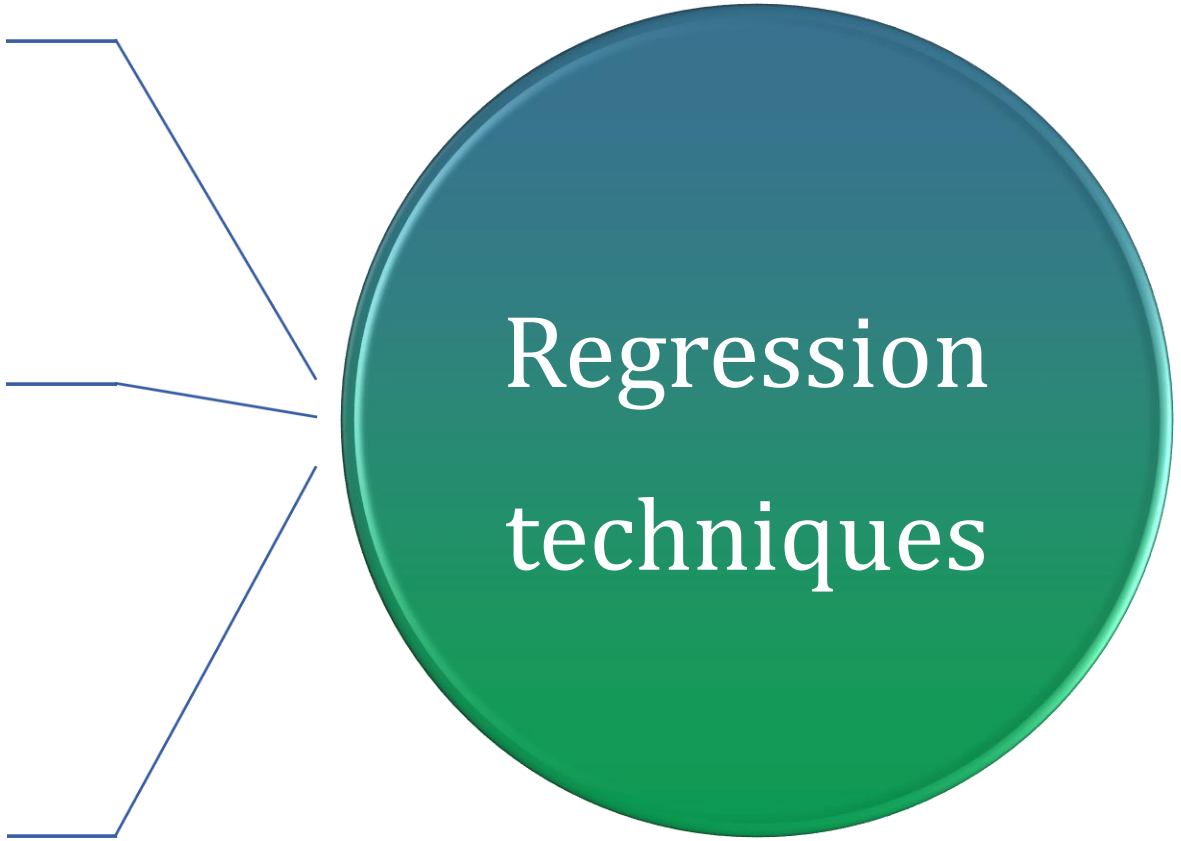


علم داده (Data Science) : ابزارهای رگرسیون

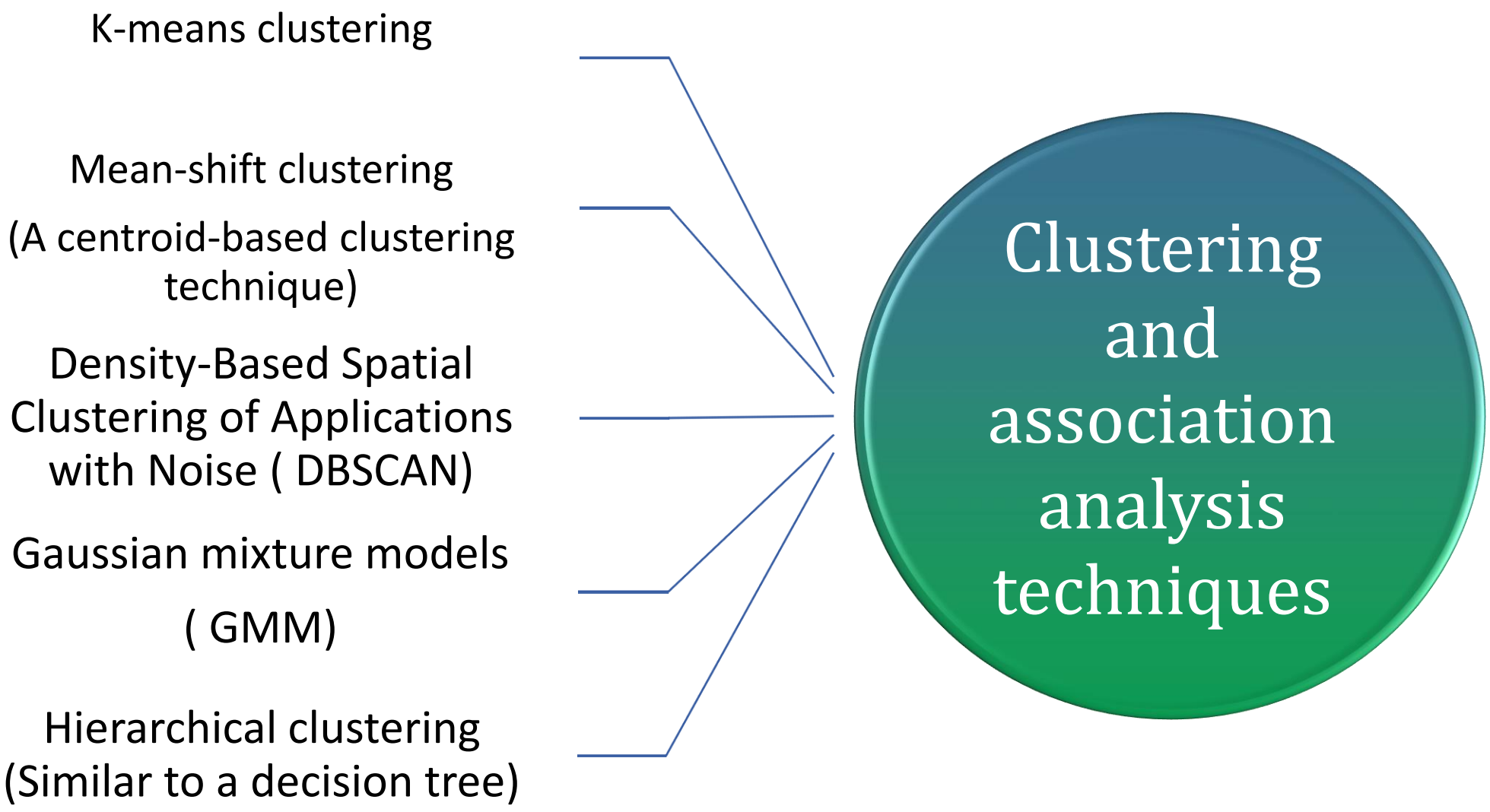
Linear regression

least absolute shrinkage and selection operator (LASSO regression)

Multivariate regression



علم داده (Data Science) : ابزارهای خوشه بندی



داده کاوی چیست ؟

داده کاوی فرآیند تجزیه و تحلیل کلان داده برای یافتن الگوها، کشف روندها و کسب بینش در مورد نحوه استفاده از آن داده ها است. سپس می توان از این یافته ها برای تصمیم گیری یا پیش بینی نتیجه استفاده کرد.

داده کاوی رشته ای است که آمار، یادگیری ماشین و هوش مصنوعی را در هم می آمیزد.

داده کاوی (Data Mining) چیست ؟

داده کاوی فنی است که از میان پایگاه داده سازمان ها، به دنبال الگوهای پنهان در میان داده ها، ارتباط میان آنها، روند و الگوی آنها می گردد.

داده کاوی با استفاده از توابع و الگوریتم های پیشرفته ریاضی ارتباط میان دو دسته از داده و امکان رخ دادن یک نتیجه را در آینده پیش بینی می کند.

داده کاوی (Data Mining) چیست ؟

- داده کاوی از **تکنیک‌های یادگیری ماشین** (و تا حدودی آمار) الهام می‌گیرد.

- داده کاوی با بکار گیری الگوریتم‌های یادگیری ماشین، الگوی مورد نیاز خود را از یک مجموعه داده دریافت می کند.
- همچنین ممکن است نواقص و اشکالاتی نیز درون دیتا وجود داشته باشد که در داده کاوی با تکنیک هایی خاص مرتفع می گردند.

- معمولا هدف از داده کاوی کشف و یا تولید روابط موجود میان دیتا و در ادامه پیش بینی مشاهدات و نتایج آینده به کمک الگوهای بدست آمده است.

اسامی دیگر داده کاوی

Knowledge discovery (mining) in databases (KDD)	• کشف دانش در پایگاه های داده
knowledge extraction	• استخراج دانش
data/pattern analysis	• تجزیه و تحلیل داده/الگو
data archaeology	• باستان شناسی داده ها
data dredging	• صید داده ها
Information harvesting	• برداشت داده
business intelligence	• هوش تجاری

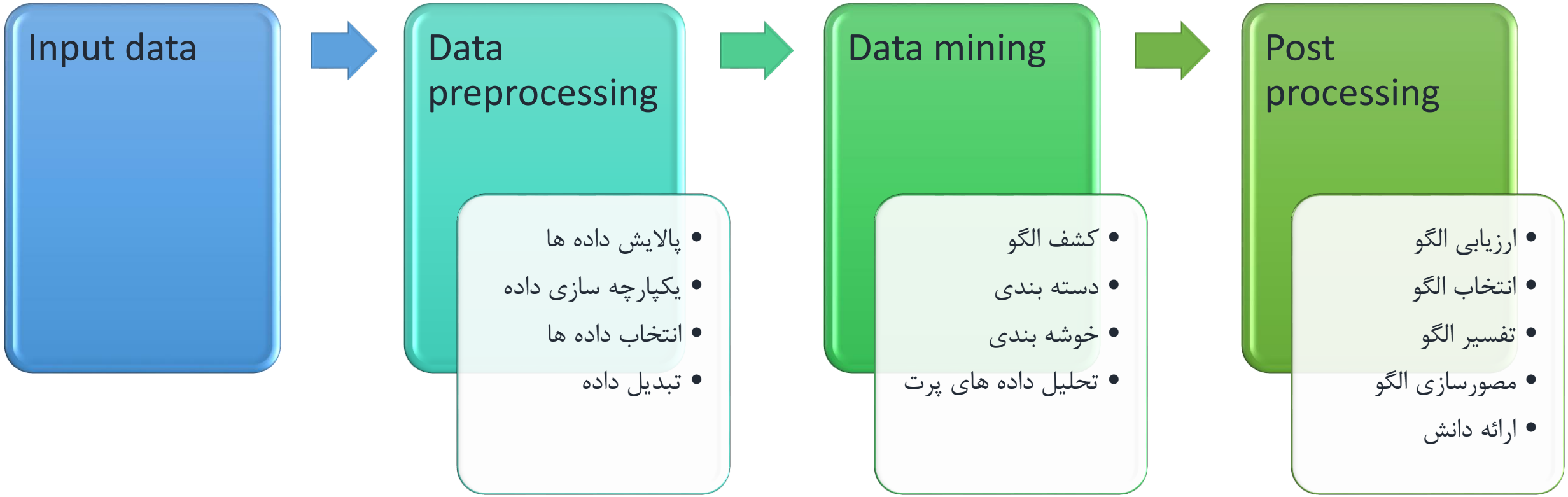
نمونه هایی از کاربرد داده کاوی

- تشخیص کلاه برداری از کارت های اعتباری
- پیش بینی ساعت اوج مصرف برق و گاز و آب
- پیش بینی های مالی و اقتصادی در بورس و رمز ارزها و غیره
- تشخیص های پزشکی
- ارزش گذاری املاک
- بهینه سازی تولید نیروگاه های آبی
- پیش بینی وضعیت آب و هوا
- حوزه تبلیغات

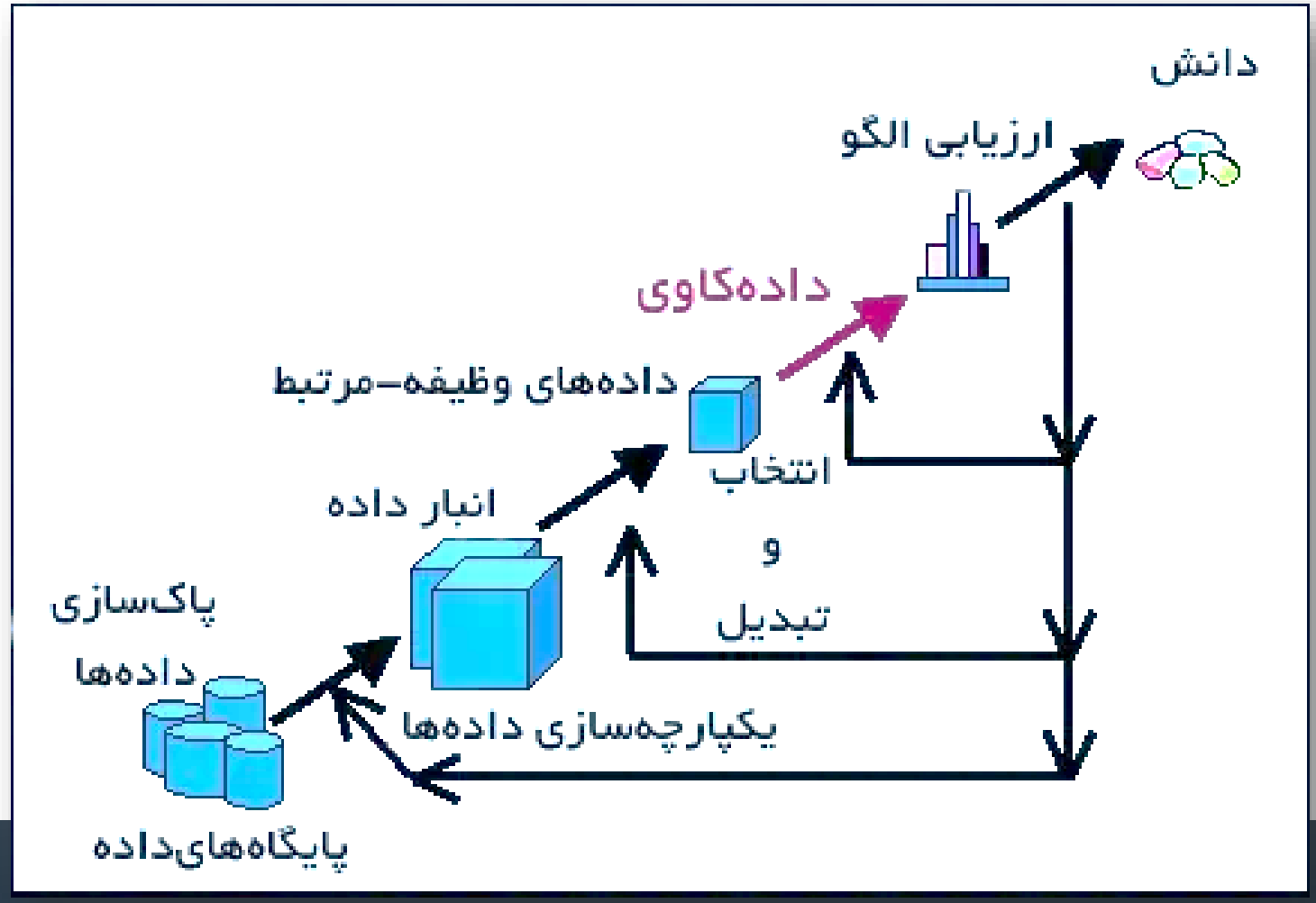
فرآیند کلی داده کاوی

فرآیند داده کاوی را می توان به شش مرحله اصلی تقسیم کرد:

فرآیند کلی داده کاوی



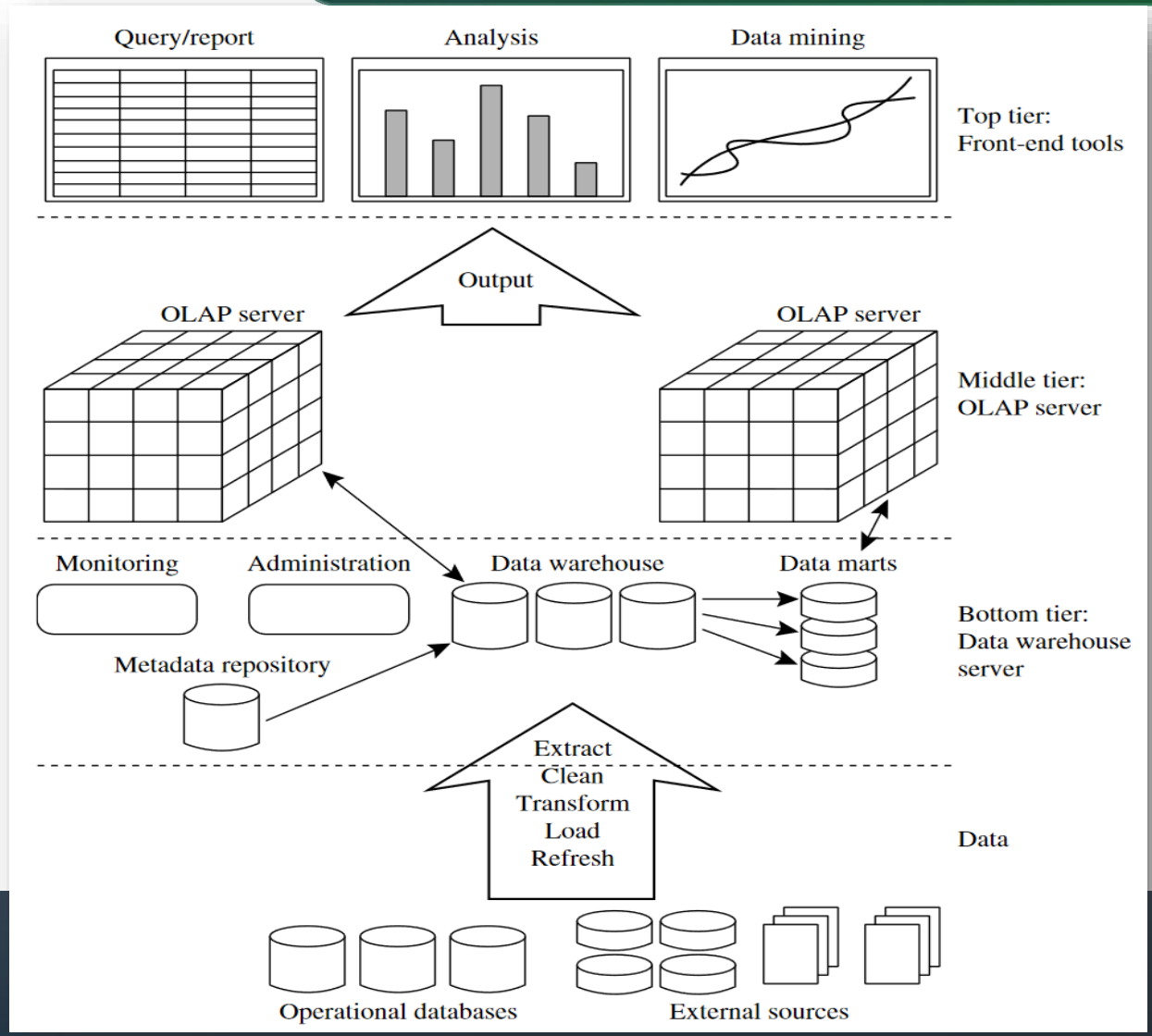
فرآیند کلی داده کاوی



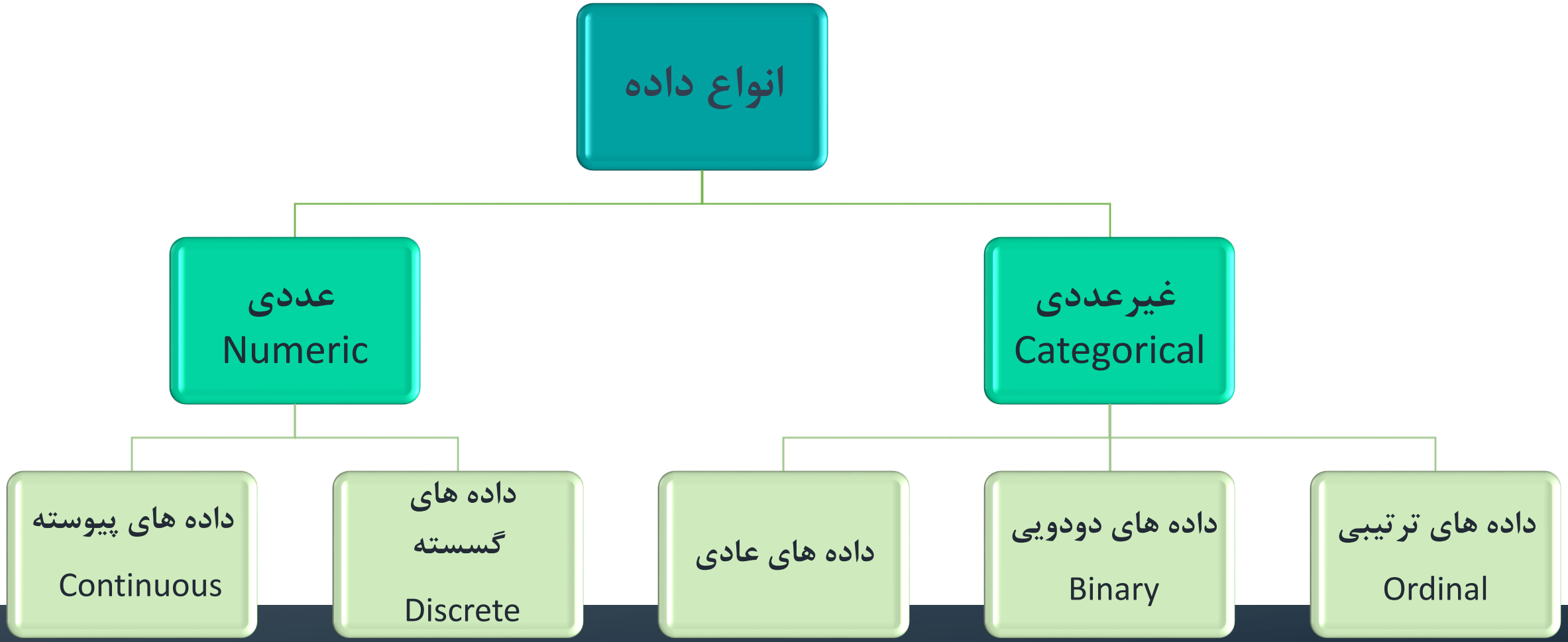
انبار داده ها Data Warehouse

انبار داده یا Data Warehouse پایگاه داده‌ای است که برای گزارش‌گیری و تحلیل داده به کار می‌رود و بعنوان هسته اصلی یک سیستم به شمار می‌آید. به عبارت دیگر انبار داده یک مخزن داده مرکزی از داده‌های تجمیع شده است که از سیستم‌ها و منابع مختلف سازمان جمع‌آوری شده است.

انبار داده ها : معماری چند لایه



انواع داده ها در داده کاوی



پیش پردازش داده ها

Data preprocessing

پیش پردازش داده

- پیش پردازش داده چیست؟
- اهمیت پیش پردازش داده ها
- مراحل پیش پردازش داده ها
- نمونه های پیش پردازش داده ها

پیش پردازش داده ها چیست؟

پیش پردازش داده گامی در فرآیند داده کاوی و تجزیه و تحلیل داده است که داده های خام را می گیرد و آن را به قالبی تبدیل می کند که توسط رایانه ها و یادگیری ماشینی قابل درک و تجزیه و تحلیل باشد.

داده‌های خام و واقعی در قالب متن، تصویر، ویدیو و غیره نامرتب هستند. نه تنها ممکن است حاوی خطاها و ناسازگاری باشد، بلکه اغلب ناقص است و طراحی منظم و یکنواختی ندارد.

پیش پردازش داده ها چیست؟

ماشین‌ها علاقه دارند اطلاعات خوب و مرتب را پردازش کنند زیرا داده‌ها را به صورت ۰ و ۱ می‌خوانند.

بنابراین محاسبه داده های ساختار یافته، مانند اعداد کامل و درصد، آسان است. با این حال، داده های بدون ساختار، به شکل متن و تصویر، ابتدا باید قبل از تجزیه و تحلیل، تمیز و قالب بندی شوند.

دلایل عدم صحت داده ها

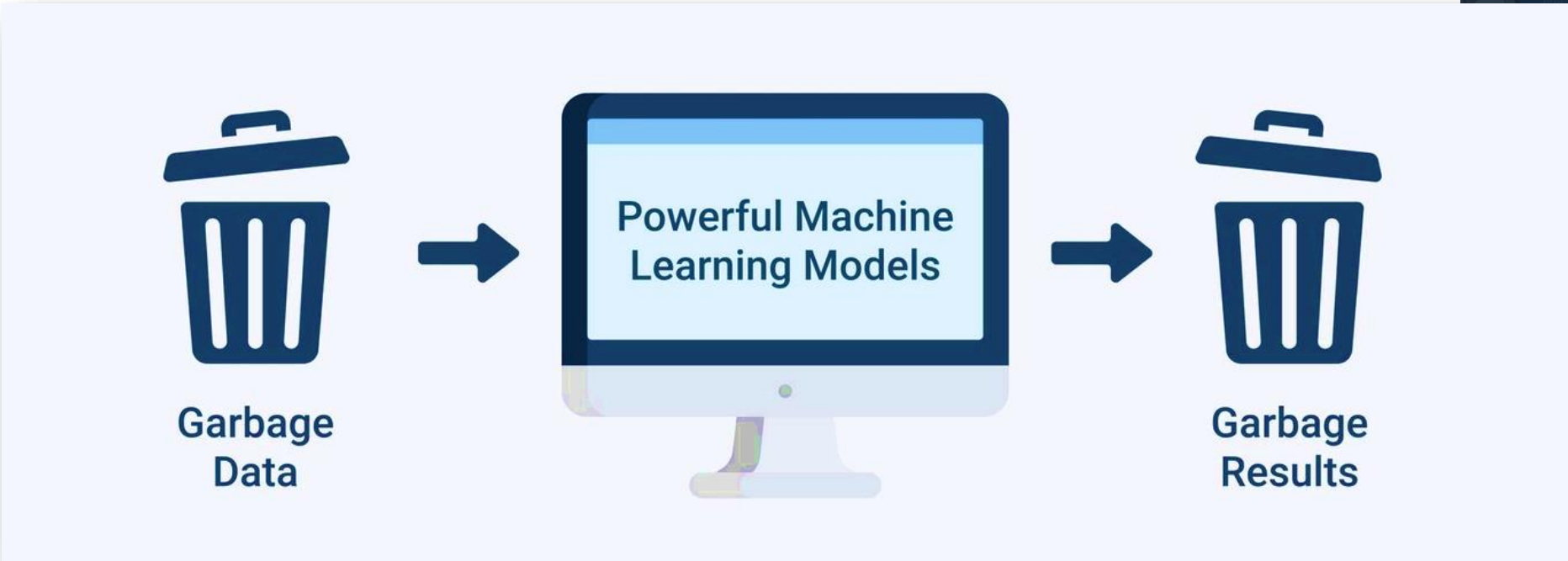
دلایل فراوان در عدم صحت و خطا داده ها:

- ابزار ها
- خطای انسانی
- خطای عمومی
- خطا در انتقال داده ها
- محدودیت فناوری
- احتمال تکراری بودن داده ها

اهمیت پیش پردازش داده ها

- اگر از داده‌های بد یا «کثیف» برای آموزش مدل خود استفاده کنید، در نهایت با یک نمونه بد مواجه خواهید شد.
- مدلی که به درستی آموزش دیده است ، در واقع به تحلیل شما مرتبط نخواهد بود.
- داده‌های خوب و از پیش پردازش‌شده حتی از قدرتمندترین الگوریتم‌ها نیز مهم‌تر هستند، تا جایی که مدل‌های یادگیری ماشینی که با داده‌های بد آموزش داده شده‌اند، می‌توانند در واقع برای تحلیلی که می‌خواهید انجام دهید مضر باشند و نتایج «بد» به شما بدهند.

اهمیت پیش پردازش داده ها



اگر از داده‌های بد یا «کثیف» برای آموزش مدل خود استفاده کنید، در نهایت با یک نمونه بد مواجه خواهید شد.

اهمیت پیش پردازش داده ها

بسته به روش‌ها و منابع جمع‌آوری داده‌ها، ممکن است در نهایت به داده‌هایی برسیم که خارج از محدوده هستند یا دارای ویژگی نادرستی هستند، مانند درآمد خانوار زیر صفر یا تصویری از مجموعه‌ای از «حیوانات باغ وحش» که در واقع یک درخت است.

مجموعه داده ممکن است مقادیر یا فیلدهای گمشده داشته باشد. یا داده های متنی، برای مثال، اغلب دارای کلمات غلط املایی و نمادهای نامربوط، URLها و غیره هستند.

اهمیت پیش پردازش داده ها

هنگامی که داده های خود را به درستی پیش پردازش و تمیز می کنید، خود را برای فرآیندهای پایین دستی بسیار دقیق تر آماده خواهید کرد.

ما اغلب در مورد اهمیت "تصمیم گیری مبتنی بر داده" می شنویم، اما اگر این تصمیمات توسط داده های بد هدایت شوند، آنها به سادگی تصمیمات بدی هستند.

مراحل پیش پردازش داده ها

Data Preprocessing Steps

مراحل پیش پردازش داده ها

۱. پاکسازی داده ها

Data Cleaning

۲. یکپارچه سازی داده ها

Data Integration

۳. کاهش داده ها

Data Reduction

۴. ارزیابی کیفیت داده ها

Data Quality Assessment

۵. تبدیل داده ها

Data Transformation

مراحل پیش پردازش داده ها

Data Preprocessing Steps

به داده های خود خوب نگاه کنید و از کیفیت کلی، ارتباط با پروژه خود و سازگاری آن ایده بگیرید.

تعدادی از ناهنجاری های داده و مشکلات ذاتی وجود دارد که تقریباً در هر مجموعه داده ای باید به آنها توجه کرد، به عنوان مثال:

داده های مفقود	داده های پرت	مقادیر داده های مختلط	انواع داده‌های نامتناسب

Data cleaning

پاکسازی داده ها

پاکسازی داده ها فرآیند افزودن داده های از دست رفته و تصحیح ، تعمیر یا حذف داده های نادرست یا نامربوط از یک مجموعه داده است.

پاکسازی داده مهمترین مرحله پیش پردازش است زیرا اطمینان حاصل می کند که داده های شما برای نیازهای پایین دست شما آماده است.

Data cleaning

پاکسازی داده ها

پاکسازی داده‌ها همه داده‌های متناقضی را که در ارزیابی کیفیت داده‌ها کشف کرده‌اید، تصحیح می‌کند.

بسته به نوع داده ای که با آن کار می کنید، تعدادی پاک کننده احتمالی وجود دارد که برای اجرای داده های خود به آنها نیاز دارید

- Missing Data
- Noisy Data

پاکسازی داده ها

داده های از دست رفته (Missing Data)

روش‌های مختلفی برای تصحیح داده‌های از دست رفته وجود دارد، اما دو روش رایج عبارتند از:

نادیده گرفتن تاپل ها

تاپل لیست یا دنباله ای از اعداد یا موجودیت های مرتب شده است. اگر چندین مقدار در تاپل ها وجود نداشته باشد، می توانید به سادگی تاپل هایی را که آن اطلاعات از دست رفته را دارند کنار بگذارید. این فقط برای مجموعه داده های بزرگ توصیه می شود، زمانی که چند تاپل نادیده گرفته شده به تجزیه و تحلیل بیشتر آسیب نمی رساند.

پر کردن دستی داده های از دست رفته

این می تواند خسته کننده باشد، اما قطعاً هنگام کار با مجموعه داده های کوچکتر ضروری است.

پاکسازی داده ها

داده های نویزی
(Noisy Data)

پاکسازی داده ها همچنین شامل اصلاح داده های "نویزدار" است. این داده‌هایی است که شامل نقاط داده غیرضروری، داده‌های نامربوط و داده‌هایی است که گروه‌بندی آنها با هم دشوارتر است.

نویز: یک خطای تصادفی یا اختلاف در بخش متغیر است، یا واریانس در یک متغیر اندازه گیری شده با مقادیر نادرست ویژگی به دلیل:

- نقص در ابزار جمع آوری داده ها
- مشکلات در ورود داده ها
- مشکلات در انتقال داده ها
- محدودیت فناوری
- تناقض در قرارداد های نامگذاری

پاکسازی داده ها

داده های نویزی (Noisy Data)

برخورد با داده های نویزی

۱. رگرسیون(Regression) : هموار کردن داده ها به کمک سازگار کردن آن ها با تابع رگرسیون

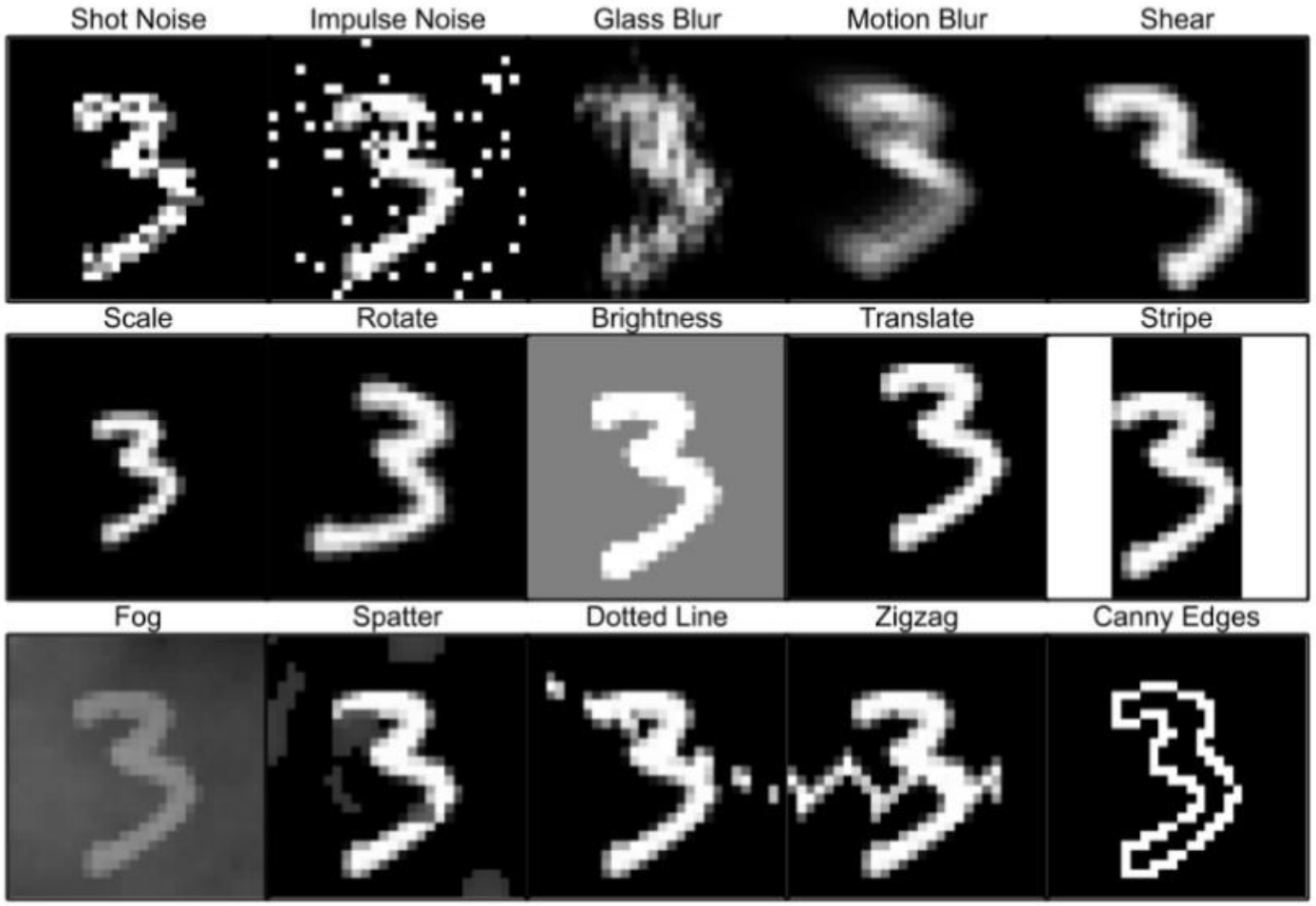
۲. خوشه بندی: شناسایی و حذف داده های پرت

۳. Binning

۴. ترکیب بازرسی کامپیوتر و انسان: تشخیص مقادیر مشکوک به صورت خودکار سپس بررسی آنها توسط عامل انسانی (به عنوان مثال تعامل با داده های دور افتاده)

پاکسازی داده ها

داده های نویزی
(Noisy Data)



پاکسازی داده ها

برخورد با داده های نویزی

خوشه بندی

الگوریتم‌های خوشه‌بندی برای گروه‌بندی مناسب داده‌ها استفاده می‌شوند تا بتوان آن‌ها را با داده‌های مشابه تجزیه و تحلیل کرد. آنها معمولاً در یادگیری بدون نظارت استفاده می شوند، زمانی که اطلاعات زیادی در مورد روابط درون داده های شما وجود ندارد.

رگرسیون

از رگرسیون برای تصمیم گیری اینکه کدام متغیرها واقعاً برای تحلیل شما اعمال می شوند استفاده می شود. تحلیل رگرسیون برای هموارسازی مقادیر زیادی از داده ها استفاده می شود. این به شما کمک می کند تا اطلاعات خود را کنترل کنید، بنابراین با داده های غیرضروری بیش از حد سنگین نشوید.

Binning

Binning داده های یک مجموعه داده گسترده را به گروه های کوچکتر از داده های مشابه تر مرتب می کند. اغلب هنگام تجزیه و تحلیل جمعیت شناسی استفاده می شود. به عنوان مثال، درآمد را می توان گروه بندی کرد: ۳۵,۰۰۰$-۵۰,۰۰۰$, ۵۰,۰۰۰$-۷۵,۰۰۰$ وغیره.

پاکسازی داده ها

برخورد با داده های نویزی

هموارسازی با کران bin

- Bin 1: 15, 4, 4
- Bin 2: 24, 21, 21
- Bin 3: 34, 25, 25

هموارسازی با میانگین bin

- Bin 1: 9, 9, 9
- Bin 2: 22, 22, 22
- Bin 3: 29, 29,29

افراز به bin ها (به تعداد برابر):

- Bin 1: 4, 8, 15
- Bin 2: 21, 21, 24
- Bin 3: 25, 28, 34

Binning

ابتدا داده ها مرتب شده و تقسیم به دسته های با فرکانس تکرار یکسان می شوند. مثال : داده ذخیره شده برای قیمت:

۴, ۸, ۱۵, ۲۱, ۲۱, ۲۴, ۲۵, ۲۸, ۳۴

پاکسازی داده ها

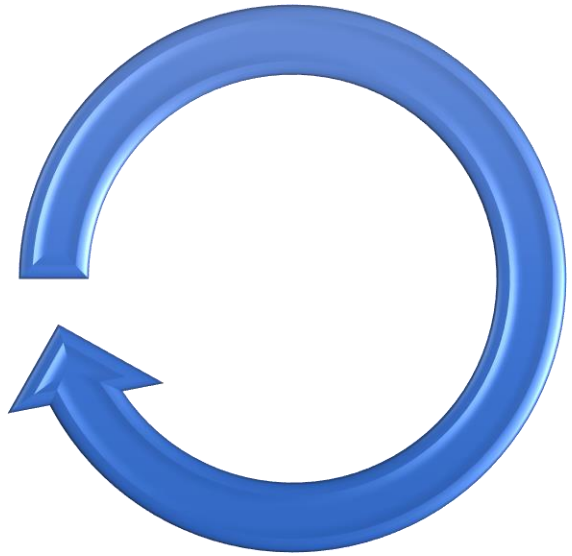
مثال

برای مثال، اگر با داده های متنی کار می کنید، برخی از مواردی که باید هنگام پاک کردن داده های خود در نظر بگیرید عبارتند از:

- نشانی‌های اینترنتی، نمادها، شکلک‌ها و غیره را که به تحلیل شما مرتبط نیستند، حذف کنید
- تمام متن را به زبانی که با آن کار می کنید ترجمه کنید
- برچسب های HTML را حذف کنید
- متن خالی غیر ضروری بین کلمات را حذف کنید
- داده های تکراری را حذف کنید

پاکسازی داده ها

مثال



- پس از پاکسازی داده ها، ممکن است متوجه شوید که داده های کافی برای کار در دست ندارید. در این مرحله همچنین می‌توانید برای افزودن مجموعه‌های داده جدید، غنی‌سازی داده‌ها یا data enrichment را انجام دهید و قبل از افزودن آن‌ها به داده‌های اصلی خود، آنها را از طریق ارزیابی کیفیت و تمیز کردن دوباره اجرا کنید.

یکپارچه سازی داده ها

Data Integration

یکپارچه سازی داده ها:

این مرحله شامل ترکیب داده ها از چندین منبع، مانند پایگاه های داده، صفحات گسترده و فایل های متنی است. هدف از یکپارچه سازی ایجاد یک نمای واحد و منسجم از داده ها است

Data Reduction

کاهش داده ها

هرچه با داده های بیشتری کار کنید، حتی پس از تمیز کردن و تبدیل آن ها، تجزیه و تحلیل آن سخت تر خواهد بود.

بسته به وظیفه ای که در دست دارید، ممکن است در واقع داده های بیشتری از آنچه نیاز دارید داشته باشید. به خصوص هنگام کار با تجزیه و تحلیل متن، بسیاری از گفتار منظم انسان اضافی یا بی ربط به نیازهای محقق است.

کاهش داده ها نه تنها تجزیه و تحلیل را آسان تر و دقیق تر می کند، بلکه ذخیره سازی داده ها را کاهش می دهد.همچنین به شناسایی مهم ترین ویژگی های فرآیند در دست اجرا کمک می کند.

Data Reduction

کاهش داده ها

از این مرحله برای انتخاب زیرمجموعه ای از داده ها استفاده می شود که مربوط به وظیفه داده کاوی است. این می تواند شامل انتخاب ویژگی (انتخاب زیر مجموعه ای از متغیرها) یا استخراج ویژگی (استخراج متغیرهای جدید از داده ها) باشد

انتخاب ویژگی Feature selection

- مانند بی اعتباری، انتخاب ویژگی می تواند داده های شما را در مجموعه های کوچکتر قرار دهد. این، اساسا، برچسب ها یا ویژگی ها را ترکیب می کند، به طوری که برچسب هایی مانند مرد/زن و استاد را می توان به استاد/استاد زن ترکیب کرد.

کاهش تعداد Numerosity reduction

- این به ذخیره سازی و انتقال داده ها کمک می کند. برای مثال می‌توانید از مدل رگرسیون استفاده کنید تا فقط از داده‌ها و متغیرهایی که به تحلیل شما مرتبط هستند استفاده کنید.

Data Reduction

کاهش داده ها

کاهش ابعاد Dimensionality Reduction

- این روش نیز مقدار داده های مورد استفاده برای کمک به تسهیل تجزیه و تحلیل و فرآیندهای پایین دستی را کاهش می دهد.
- الگوریتمK-Nearest Neighbor یا به اختصار KNN از تشخیص الگو برای ترکیب داده‌های مشابه و مدیریت پذیرتر استفاده می‌کنند.
- همچنین PCA یا Principal Component Analysis تحلیل مولفه های اصلی است که یکی از روشهای مهم کاهش ابعاد است.

Data transformation

تبدیل داده ها

با پاکسازی داده‌ها، ما قبلاً شروع به اصلاح داده‌های خود کرده‌ایم، اما Data transformation، فرآیند تبدیل داده‌ها به قالب(های) مناسبی که برای تجزیه و تحلیل و سایر فرآیندهای پایین دستی نیاز دارید، را آغاز می‌کند.
این معمولاً در یک یا چند مورد از موارد زیر رخ می دهد:

- رمزگذاری داده های طبقه بندی
  encoding categorical data
- ایجاد متغیرهای ساختگی
  creating dummy variables
- سلسله مراتب مفهومی
  Concept hierarchy generation
- بی اعتباری تولید
  Discreditization
- انتخاب ویژگی
  Feature selection
- نرمال سازی داده های عددی
  Normalization
- تجمع
  Aggregation

Data transformation

تبدیل داده ها

Data Normalization

نرمال سازی

نرمال سازی داده های شما را در یک محدوده منظم مقیاس بندی می کند تا بتوانید با دقت بیشتری آن ها را مقایسه کنید.

Data aggregation

تجمیع داده

تجمیع داده ها همه داده های شما را با هم در قالبی یکسان ترکیب می کند.

به عنوان مثال، اگر در حال مقایسه ضرر یا سود کارکنان در تعدادی از شرکت‌ها هستید (بعضی از آنها تنها با ۱۲ کارمند و برخی با بیش از ۲۰۰ کارمند)، باید آنها را در محدوده مشخصی مانند ۰.۱ - تا ۰.۱ یا ۰.۰ تا ۱ مقیاس کنید.

Data transformation

تبدیل داده ها

مهم است که به خاطر داشته باشید، که هر چه ویژگی‌های بیشتری را برای استفاده انتخاب کنید، روند آموزشی طولانی‌تر و گاهی اوقات نتایج شما دقیق‌تر می‌شود، زیرا برخی از ویژگی‌ها ممکن است همپوشانی داشته باشند یا کمتر در داده‌ها وجود داشته باشند.

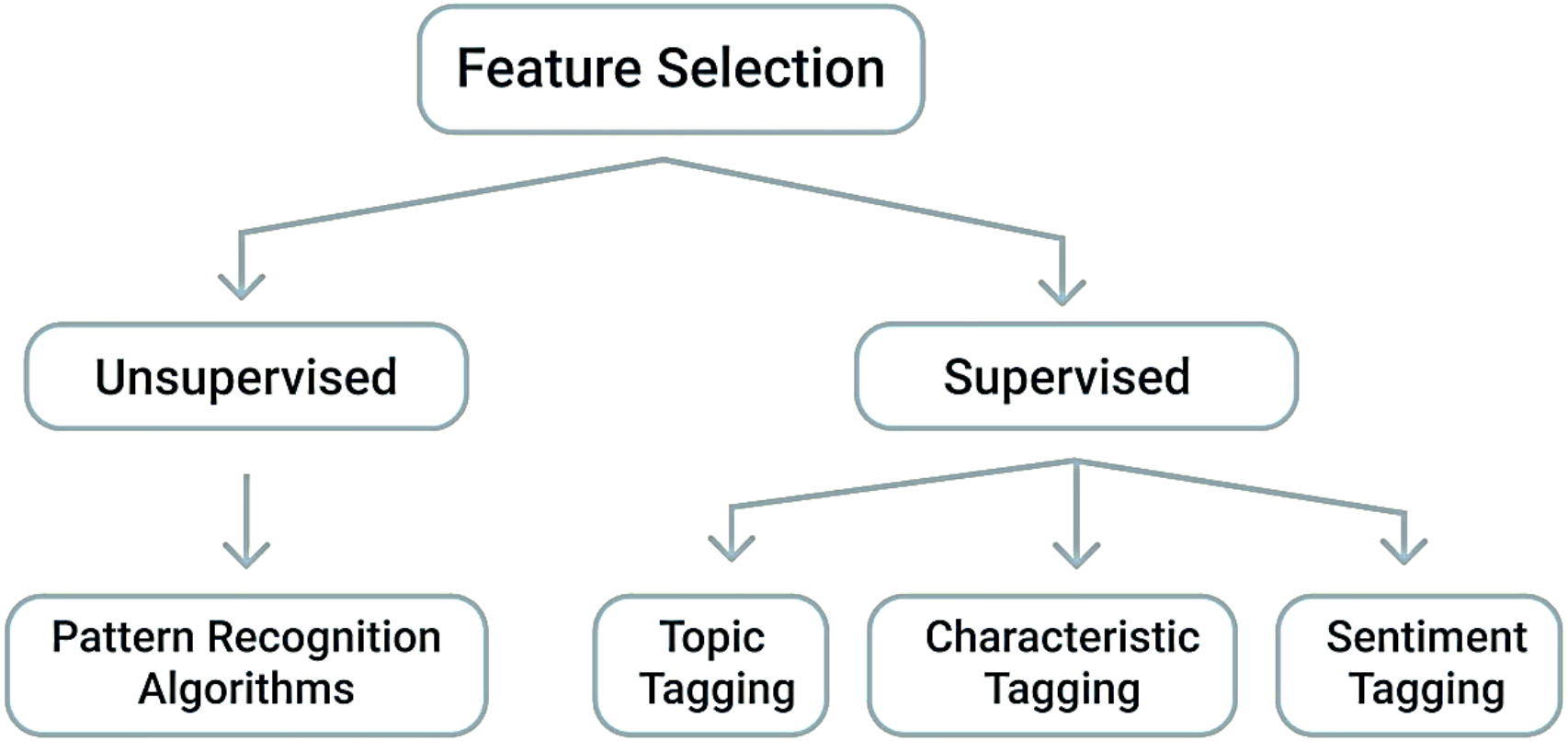
انتخاب ویژگی فرآیند تصمیم گیری است که کدام متغیرها (ویژگی ها، ویژگی ها، دسته ها و غیره) برای تجزیه و تحلیل شما مهم هستند. این ویژگی ها برای آموزش مدل های MLاستفاده خواهند شد.

انتخاب ویژگی

Feature selection

Data transformation

تبدیل داده ها



Data transformation

تبدیل داده ها

بی اعتباری

Discreditization

- بی اعتباری داده ها را در فواصل کوچکتر جمع می کند. این تا حدودی شبیه binningاست، اما معمولاً پس از پاک کردن داده ها اتفاق می افتد. به عنوان مثال، هنگام محاسبه میانگین تمرین روزانه، به جای استفاده از دقیقه ها و ثانیه های دقیق، می توانید داده ها را به یکدیگر متصل کنید تا به ۰-۱۵ دقیقه، ۱۵-۳۰ و غیره بیفتید.

ایجاد سلسله مراتب مفهومی

Concept hierarchy generation

- تولید سلسله مراتب مفهومی می تواند سلسله مراتبی را در داخل و بین ویژگی های شما اضافه کند که در داده های اصلی وجود نداشت. اگر تجزیه و تحلیل شما شامل گرگ ها و کایوت ها است، برای مثال، می توانید سلسله مراتب را برای جنس آنها اضافه کنید: canis.

انواع داده‌های نامتناسب: Mismatched data types

وقتی داده‌ها را از منابع مختلف جمع‌آوری می‌کنید، ممکن است در قالب‌های مختلف به دست شما برسد. در حالی که هدف نهایی کل این فرآیند، قالب‌بندی مجدد داده‌ها برای استفاده ماشین‌ها است، هنوز باید با داده‌هایی با فرمت مشابه شروع کنید.

به عنوان مثال، اگر بخشی از تحلیل شما شامل درآمد خانواده از چندین کشور باشد، باید هر مقدار درآمد را به یک ارز واحد تبدیل کنید.

ارزیابی کیفیت داده ها

Data Quality Assessment

داده های پرت : The Outliers

مقادیر داده پرت می توانند تأثیر زیادی بر نتایج تجزیه و تحلیل داده ها داشته باشند.

به عنوان مثال، اگر میانگین نمرات آزمون را برای یک کلاس در نظر می گیرید، و یک دانش آموز به هیچ یک از سؤالات پاسخ نداده است، ۰٪ آنها می تواند نتایج را تا حد زیادی منحرف کند.

Data Quality Assessment

ارزیابی کیفیت داده ها

مقادیر داده های مختلط :Mixed data values

شاید منابع مختلف از توصیفگرهای مختلفی برای ویژگی ها استفاده کنند و همه این توصیفگرهای ارزش باید یکنواخت شوند.

به عنوان مثال، man یا male .

ارزیابی کیفیت داده ها

Data Quality Assessment

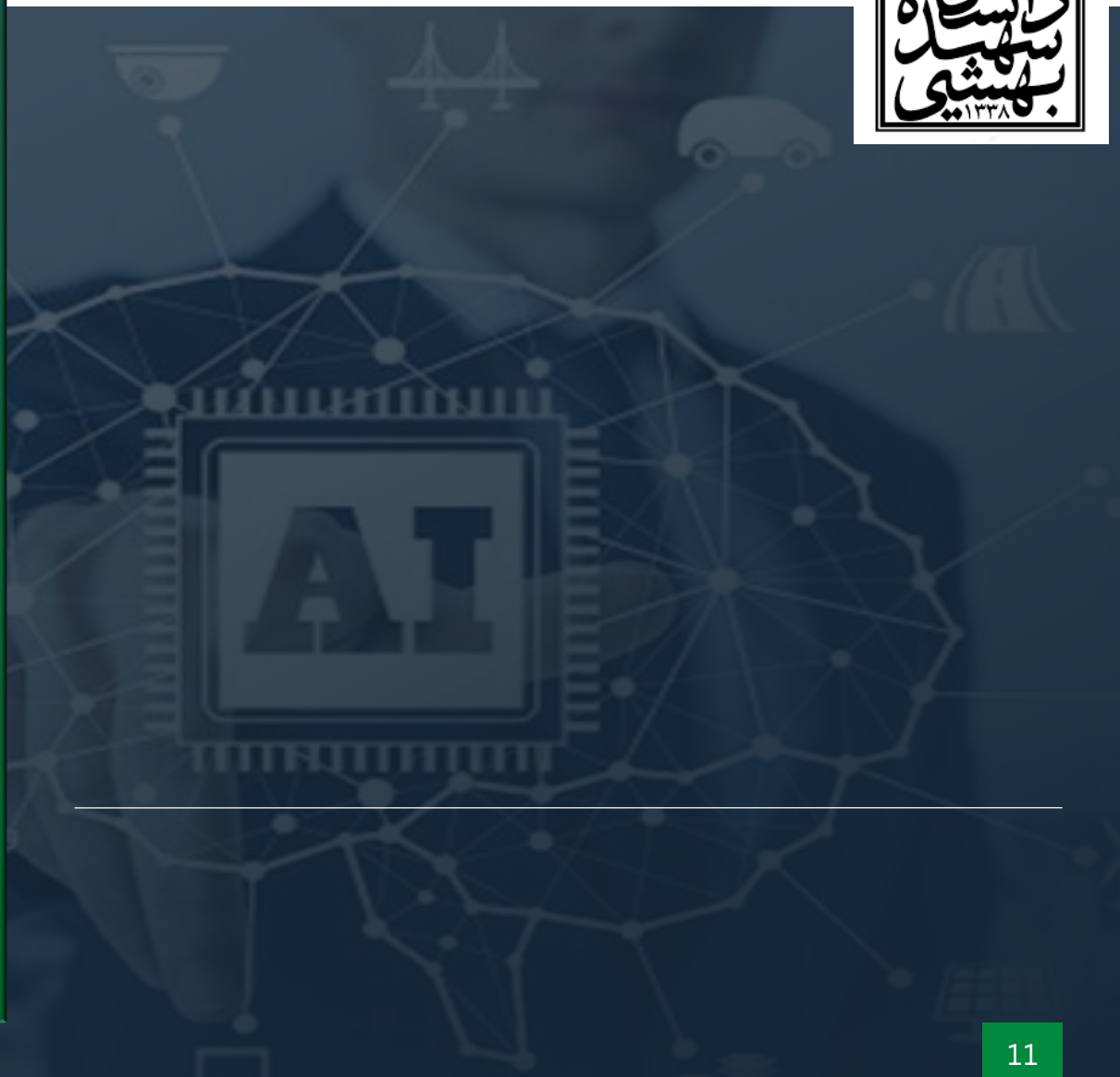
داده های مفقود :Missing data

برای جلوگیری از اثر بد داده های از دست رفته، باید پاکسازی داده ها را انجام دهید.

مثلا : به فیلدهای داده از دست رفته، فضاهای خالی در متن یا سؤالات بی پاسخ نظرسنجی نگاه کنید. این می تواند به دلیل خطای انسانی یا داده های ناقص باشد.

یادگیری عمیق

Deep Learning



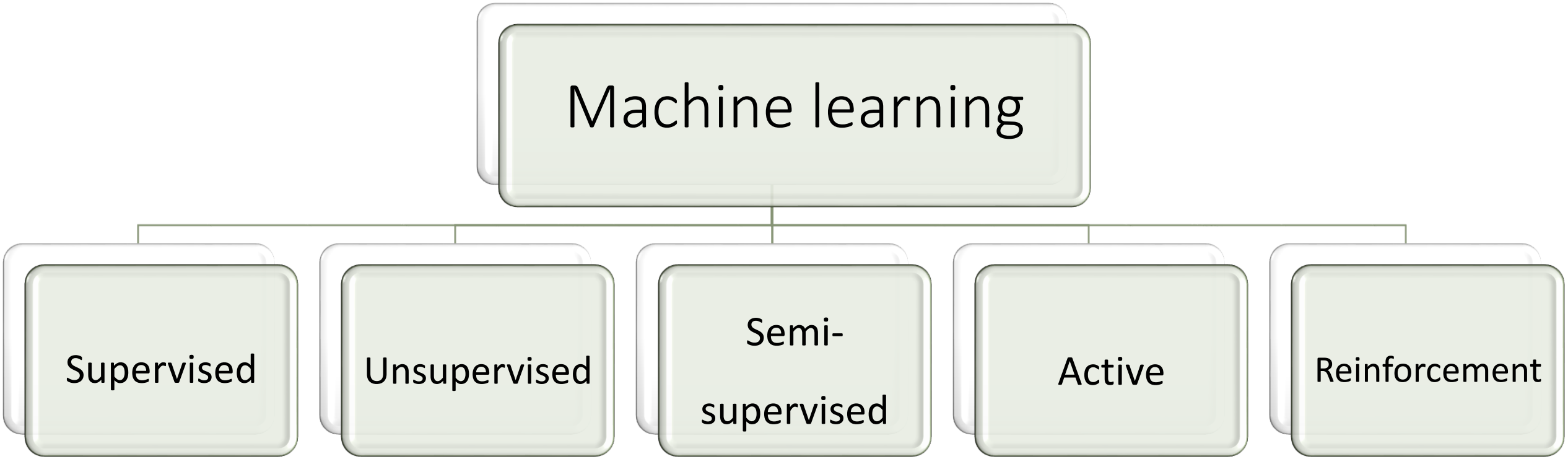
یادگیری ماشین (Machine Learning) چیست ؟

یادگیری ماشین زیر مجموعه ای از هوش مصنوعی بوده و تمرکز آن بر آموزش کامپیوتر است

غیر از برنامه نویسی و تنظیمات دقیق اولیه ، رایانه برای یادگیری داده ها و بهبود پیش بینی های خود به هیچگونه مداخله انسانی احتیاجی ندارد.

این سیستم از آنجایی که از داده های بیشتر و بیشتری تغذیه می کند، به خودی خود هوشمند خواهد شد

رویکردهای یادگیری ماشین (Machine Learning)



یادگیری ماشین (Machine Learning) چیست ؟

یادگیری نظارت شده (Supervised Learning)

در یادگیری ماشین نظارت شده ورودی‌هایی با برچسب خروجی دلخواه به ماشین داده می‌شود. هدف این الگوریتم این است که ماشین بتواند خروجی واقعی را با خروجی آموزش داده شده مقایسه کرده و مدل را متناسب با آن اصلاح کند؛ بنابراین یادگیری نظارت شده از الگوهایی برای پیش‌بینی ورودی‌های جدید استفاده می‌کند.

از الگوریتم یادگیری ماشین نظارت شده برای پیش‌بینی وقایع احتمالی آینده با کمک داده‌های تاریخی استفاده می‌شود. از دیگر کاربردهای یادگیری نظارت شده می‌توان به پیش‌بینی نوسانات آینده بازار بورس، فیلتر کردن ایمیل‌های هرزنامه و غیره اشاره کرد.

یادگیری ماشین (Machine Learning) چیست ؟

یادگیری بدون نظارت (Unsupervised Learning)

کشف الگوهای پنهان در یک مجموعه داده می‌تواند یکی از اهداف یادگیری نظارت نشده باشد. هدف دیگری که برای این الگوریتم می‌توان بیان کرد این است که یک ماشین محاسباتی بتواند داده‌های خام را به طور خودکار طبقه‌بندی کند.

در یادگیری بدون نظارت داده‌های ورودی برچسب ندارند؛ بنابراین الگوریتم یادگیری ماشین نظارت نشده تلاش می‌کند نقاط مشترک بین داده‌های ورودی را پیدا کند (در این نوع از یادگیری مدل ورودی‌ها بر اساس ویژگی ها، تفکیک و دسته بندی می‌شوند). از آنجایی که تعداد داده‌های بدون برچسب همواره بیشتر از داده‌های با برچسب است، فراگیری آموزش یادگیری ماشین نظارت نشده اهمیت ویژه‌ای دارد.

یادگیری ماشین (Machine Learning) چیست ؟

یادگیری تقویتی (Reinforcement Learning)

به طور کلی، یک عامل یادگیری تقویتی قادر است محیط خود را درک و تفسیر کند، اقداماتی انجام دهد و از طریق آزمون و خطا یاد بگیرد. این حوزه به خصوص در حوزه رباتیک کاربرد دارد.

یادگیری تقویتی یک روش آموزش یادگیری ماشینی است که بر اساس پاداش دادن به رفتارهای دلخواه و/یا تنبیه رفتارهای نامطلوب است.

یادگیری عمیق (Deep Learning)

یادگیری عمیق به عنوان یکی از زیرشاخه‌های حوزه یادگیری ماشین تلقی می‌شود. هدف یادگیری عمیق طراحی سیستم‌های کامپیوتری هوشمندی است که بتوانند مشابه انسان درباره موضوعی خاص، راه‌حل ارائه کنند و مفاهیم جدیدی را یاد بگیرند.

این حوزه از فناوری، شاخه‌ای مهم در «علم داده» است، زیرا اصلی‌ترین مباحث این شاخه، آمار و مدل‌سازی برای پیش‌بینی مسائل مختلف را شامل می‌شود.

یادگیری عمیق (Deep Learning)

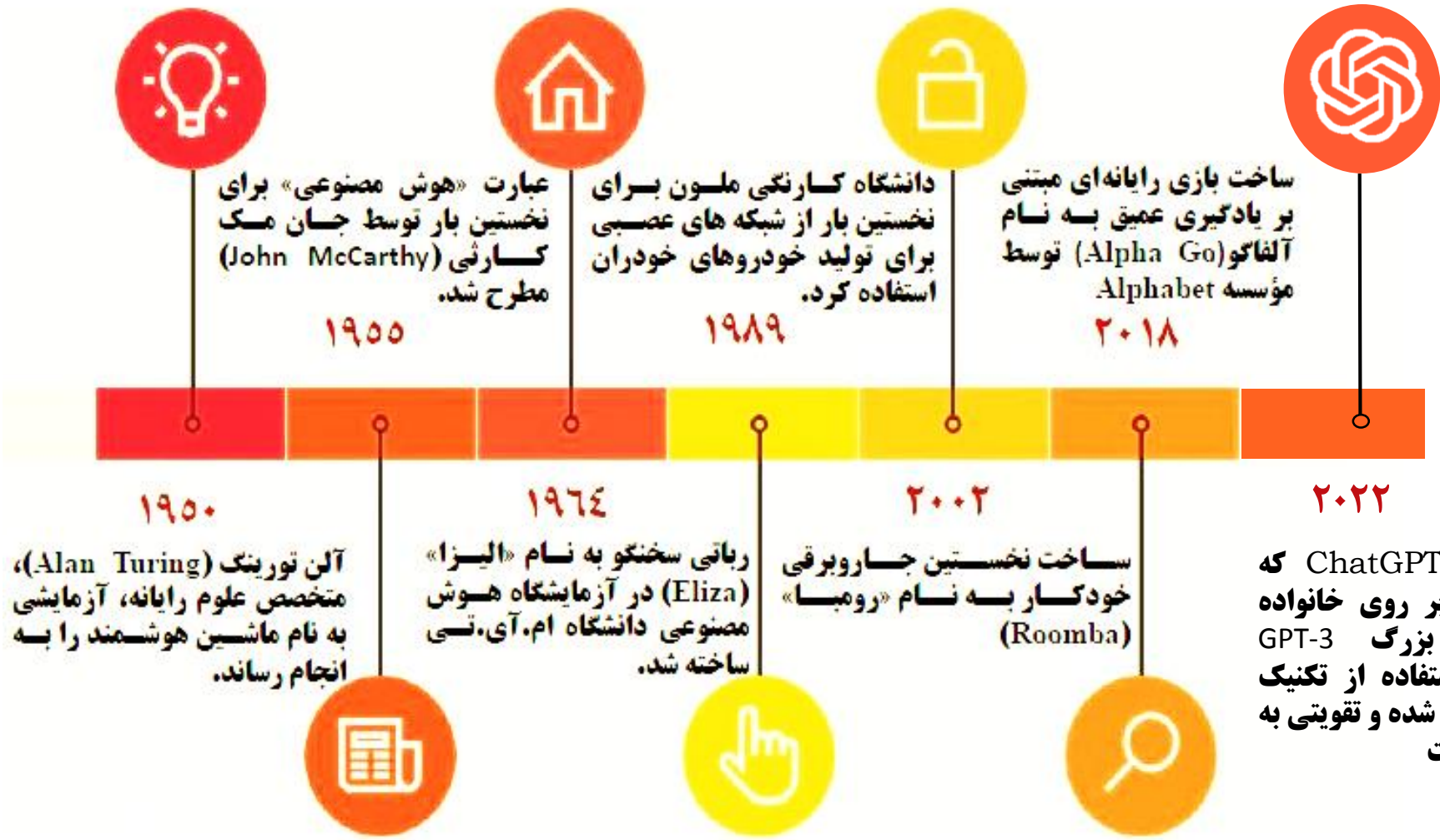
یادگیری عمیق (Deep Learning) بخشی از روش‌های یادگیری ماشین است که بر روش‌هایی تمرکز دارد که مبتنی بر شبکه‌های عصبی مصنوعی (Artificial neural network) هستند.

یادگیری عمیق به رایانه‌ها می‌آموزد آنچه را که به طور طبیعی برای انسان انجام می‌شود، انجام دهند

تفاوت یادگیری عمیق و یادگیری ماشین

الگوریتم‌های یادگیری ماشین به منظور یادگیری داده‌ها و پیش‌بینی مقداری در خروجی، به مفهومی با عنوان «مهندسی ویژگی» (Feature Engineering) یا «استخراج ویژگی» (Feature Extraction) و «انتخاب ویژگی» (Feature Selection) متکی بوده، درحالی که مدل‌های یادگیری عمیق به منظور یادگیری الگوهای داده‌ها، مبتنی بر مفاهیمی با نام‌های «لایه» (Layer) و «عمیق» (Deep) هستند.

تاریخچه هوش مصنوعی



ویژگی های کلیدی یک سیستم هوش مصنوعی

ویژگی های کلیدی یک سیستم هوش مصنوعی

یک سیستم هوش مصنوعی همان گونه که از اسمش پیداست باید فعالیت های ذهن انسان را به صورت مصنوعی انجام دهد . ذهن انسان قابلیت پردازش علومی همانند زبان شناسی، روان شناسی، فلسفه و ... را علاوه بر علوم ریاضی و کامپیوتری دارد، در نتیجه یک سیستم هوش مصنوعی باید توانایی هایی از قبیل پردازش، یادگیری، استدلال و حل مسئله را داشته باشد.

بنابراین هوش مصنوعی یک سیستم کامپیوتری است که رفتار انسان را طبق ۴ توانایی که اشاره کردیم، بروز می دهد؛ همچنین طبق تعریف بنیان گذار هوش مصنوعی آلن تورینگ: »علم ساخت ماشین های هوشمند و به ویژه برنامه های کامپیوتری است«.

ویژگی های کلیدی یک سیستم هوش مصنوعی

استدلال

توانایی استدلال و استنتاج بر اساس موقعیت

حل مسئله

قابلیت تجزیه و تحلیل و حل مشکلات پیچیده عمومی و تخصصی

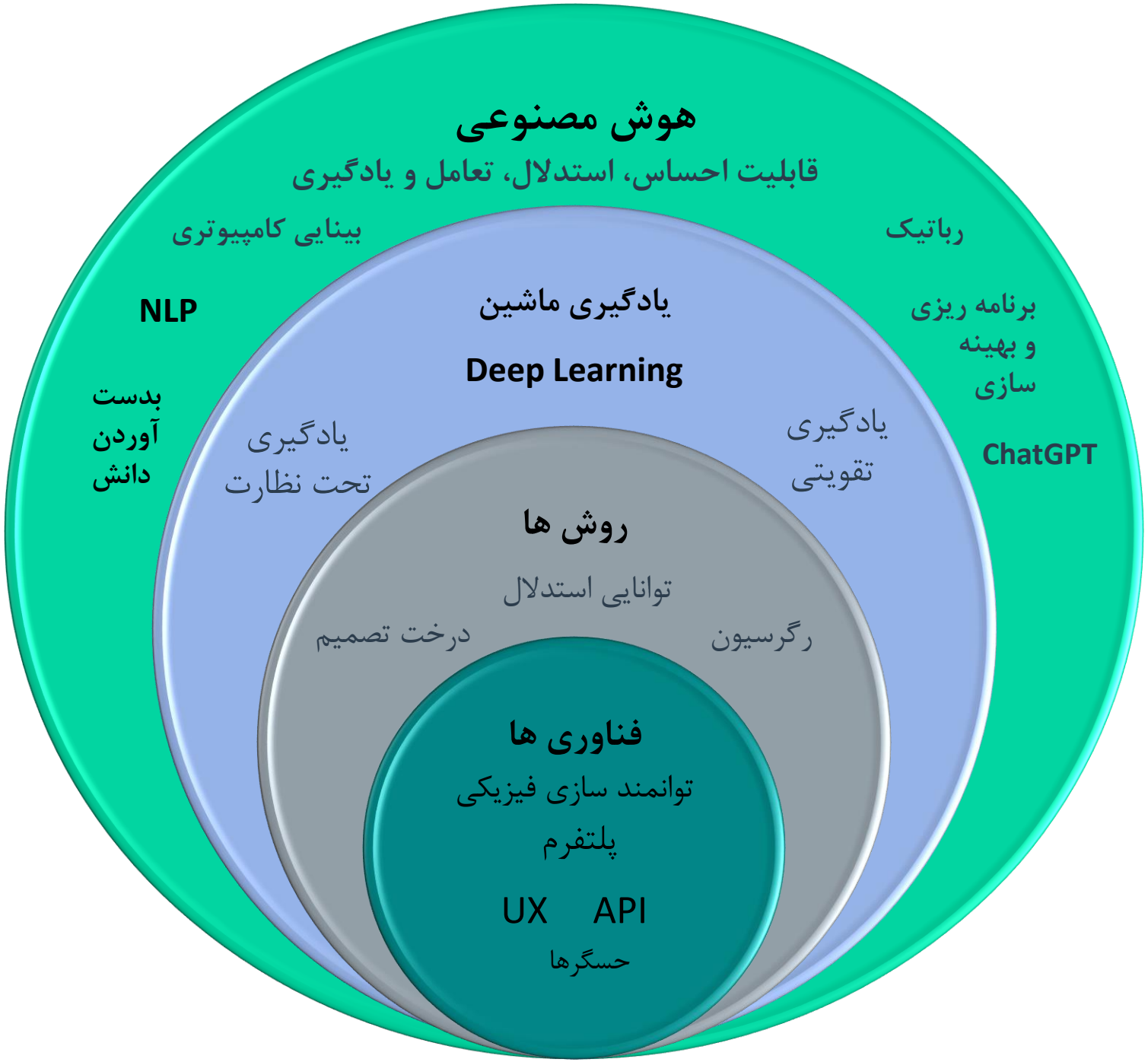
یادگیری

توانایی یادگیری بر اساس الگوهای تاریخی، ورودی های تخصصی و حلقه های بازخورد

پردازش اطلاعات

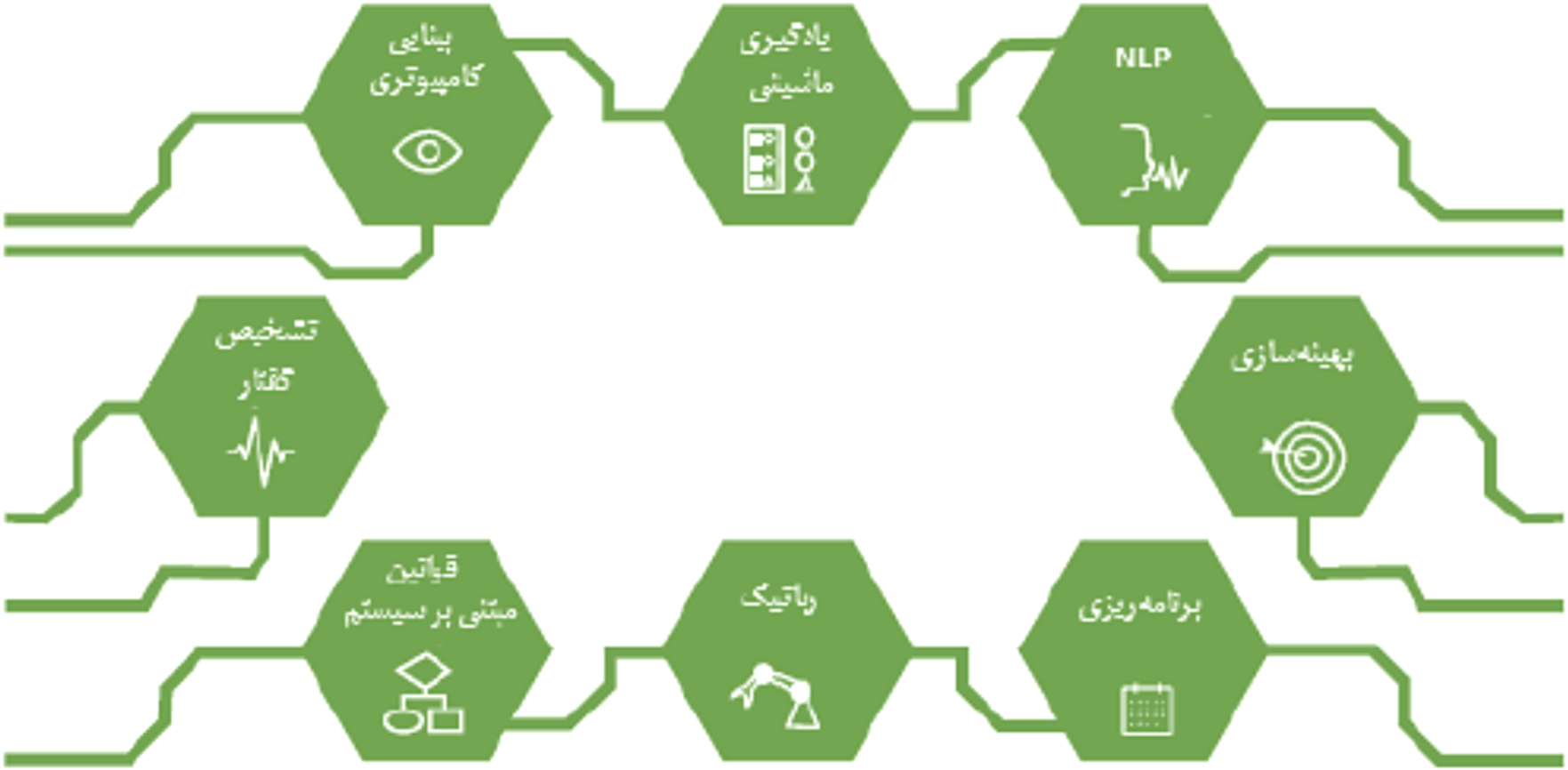
قادر به پردازش اطلاعات عظیم ساختار یافته یا بدون ساختار که دائما تغییر می کنند

ارتباط بین هوش مصنوعی ، یادگیری ماشین و روشهای پایه



یک سیستم هوش مصنوعی با ترکیب و استفاده از مهارت های یادگیری ماشین و سایر روش های تجزیه و تحلیل اطلاعات، به توانایی هایی نظیر رباتیک ، بهینه سازی و غیره دست می باید.

تاثیر هوش مصنوعی بر فناوری های شناختی



تاثیر هوش مصنوعی بر فناوری های شناختی

یادگیری ماشینی

به توانایی سیستم های کامپیوتری برای بهبود عملکرد خود با قرار گرفتن در معرض اطلاعات بدون نیاز به دستورالعمل های برنامه ریزی شده، یادگیری ماشینی گفته می شود. در این فناوری ماشین یکبار الگو را کشف می کند سپس از آن به صورت خودکار برای پیش بینی استفاده می کند.

NLP

پردازش زبان طبیعی به کامپیوتر ها توانی می دهد که مثل انسان با متن برخورد کنند. به عنوان مثال استخراج مفاهیم از متن یا تولید متنی که قابل خواند باشد و همچنین از نظر دستور زبان نیز صحیح باشد

رباتیک

رباتیک با ادغام دیگر فناوری های شناختی نظیر بینایی کامپیوتری، برنامه ریزی خودکار، حسگرهای با عملکرد بالا، موتورهای عملگر و سخت افزارهای هوشمند باعث ایجاد نسل جدیدی از ربات ها می شود که می توانند در کنار افراد کار کنند و با انعطاف پذیری بالا، وظایف مختلف در محیط های مختلف را به خوبی انجام دهند

چالش ها و فرصت های کلیدی حوزه علوم داده

- جمع آوری و تحلیل اطلاعات جمع آوری شده از فرآیندهای مختلف سازمان به منظور بهبود تصمیم گیری مدیران و پیش بینی آینده
- پردازش تصویر و ویدیو برای پیش بینی اتفاقات، کاهش اشتباهات و تحلیل عملکرد افراد بر اساس حرکات آن ها
- ایجاد دستیارهای مجازی برای کاهش هزینه، زمان و ارائه پیشنهاد به افراد و شرکت ها در جهت تسهیل انجام کارها
- ارتقا فضای امنیت سایبری و کاهش تقلب و کلاهبرداری با اتکا به توانمندی هوش مصنوعی و تحلیل رفتار کاربران

چالش ها و فرصت های کلیدی حوزه علوم داده

چالش/فرصت ۱: جمع آوری و تحلیل اطلاعات جمع آوری شده از فرآیندهای مختلف سازمان به منظور بهبود تصمیم گیری مدیران و پیش بینی آینده

راهکار:

۱- تحلیل اطلاعات مشتری برای شناسایی بهترین مدل بازاریابی و افزایش فروش

۲-تحلیل اطلاعات و سابقه بیماران به منظور دستیابی به الگوهای درمانی مشخص برای افراد

۳- تحلیل رفتار مخاطبین به منظور پیشنهاد محتوای شخصی شده برای آن ها و تولید محتوا با کمک یادگیری ماشینی توسط هوش مصنوعی

۴-تجمیع و تحلیل اطلاعات خط تولید به منظور بهبود تصمیم گیری و حتی گرفتن تصمیمات در حوزه های تکراری توسط هوش مصنوعی

۵-پیدا کردن الگوی مصرف خانه ها و کارخانه ها برای بهینه سازی مصرف انرژی

۶-تحلیل اطلاعات دوربین های کنترل ترافیک جهت کاهش ترافیک و بهبود حمل و نقل

چالش ها و فرصت های کلیدی حوزه علوم داده

چالش/فرصت ۲: پردازش تصویر و ویدیو برای پیش بینی اتفاقات، کاهش اشتباهات و تحلیل عملکرد افراد بر اساس حرکات آن ها

راهکار:

۱- خودکارکردن وسایل نقلیه با تحلیل‌حرکت دیگر وسایل و پردازش مسیر حرکت آن ها

۲-کنترل کیفیت هوشمند با تحلیل شکل قطعات تولید شده صورت بصری آن ها

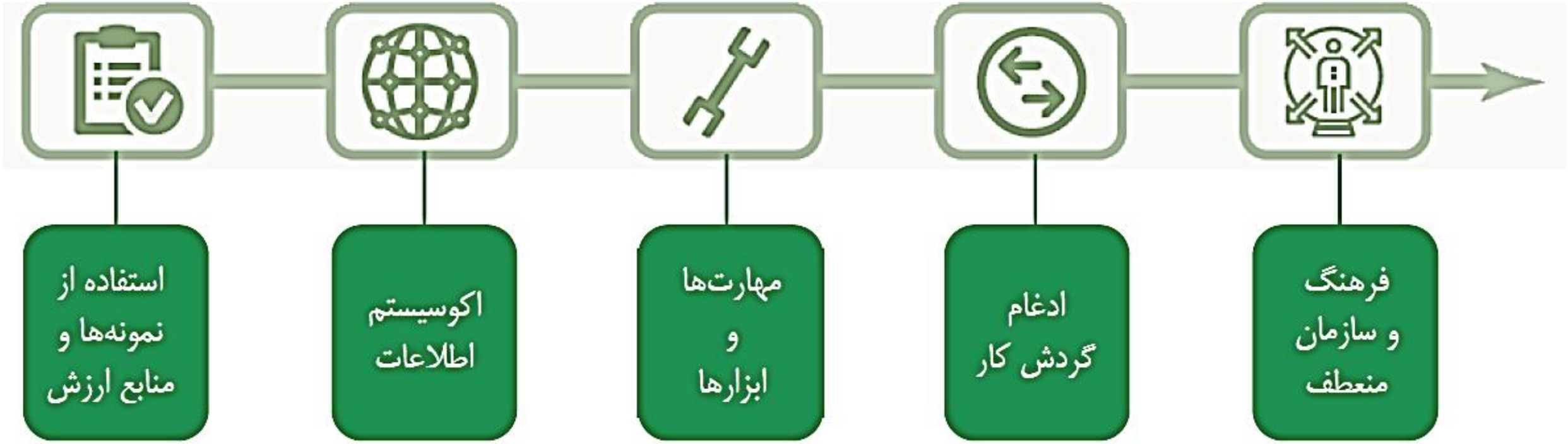
۳- تحلیل محل قرارگیری‌بازیکنان‌ورزش‌و آنالیزحرکت‌آن ها با کمک‌پردازش تصویر و هوش‌مصنوعی

۴- تحلیل و پردازش تصاویربه منظور ارتقا امنیت سیستم ها و تعریف سطوح دسترسی

۵- افزایش توان جسمی معلولان از طریق تحلیل‌ویدئو و جبران کردن ضعف بدنی آن ها

۶-رشد فکری کودکان توسط بازی‌های رایانه ای که به واسطه تحلیل تصاویر ایجاد شده

عناصر مهم در تحولات موفقیت آمیز هوش مصنوعی



فعالیت های حوزه هوش مصنوعی

تحلیل اطلاعات و بهینه سازی و پیش بینی مبتنی بر آن	• حل دغدغه هایی از قبیل کاهش ریسک، بهینه سازی و افزایش کارایی از طریق تجزیه و تحلیل کلان داده ها
هوشمند سازی خدمات	• شخصی سازی کسب وکارها با دیجیتال سازی خدمات، و افزایش بهره وری با ایجاد دستیار هوشمند در قسمت های مختلف آنها جهت انجام دقیق و سریع امور
پردازش صوت و تصویر	• پردازش زبان طبیعی و صوت، تحلیل عکس و ویدئو و تحلیل محتوای غیر متنی
هوشمند سازی تجهیزات ،ماشین آلات و کمک ابزارها	• هوش مصنوعی نیاز به اپراتور را در ربات کمتر کرده و باعث می شود آن ها به طور هوشمند و کاملا مستقل و خودکار عمل کنند.
ارتقاء امنیت	• بکارگیری فناوری هایی نظیر بلاک چین جهت جلوگیری از تقلب و ایجاد سازوکاری ایمن برای نهادها

کاربرد هوش مصنوعی در سیاست گذاری و حکمرانی منطقه ای

فناوری های نوین و توسعه

علم، فناوری و نوآوری را می توان یکی از مهمترین ارکان اقتدار ملی و محور توسعه و بالندگی در کشور دانست.

مهیا ساختن زمینه لازم برای **توسعه فناوری و نوآوری** و حرکت به سمت نفوذ آن در بخشهای مختلف اقتصادی، یکی از مهمترین موضوعاتی است که در سطوح مختلف برنامه ریزی و اجرا باید به آن توجه داشت.

فناوری های نوین و توسعه

تغییرات فناورانه ای که سالهای اخیر، صنایع و بخش های مختلف را تحت تأثیر خود قرار داده است، لزوم اندیشیدن به موارد زیر را بیش از پیش مشخص کرده است:

راهکارهای بهره گیری از این تغییرات در راستای حل مسائل و اولویتهای کشور و استفاده از ظرفیت آن در خلق ارزش اقتصادی

به کارگیری نیروی انسانی دانش آموخته و تحصیلکرده

اهمیت علم داده در سیاست گذاری

علم داده و هوش مصنوعی می توانند به استخراج اطلاعات و دانش مفید از کلان داده ها به منظور بهبود تصمیم گیری دولتی یا ارائه بینش برای تصمیم گیری های مبتنی بر داده، استفاده از تجزیه و تحلیل علّی پیش بینی کننده، تجزیه و تحلیل تجویزی، و یادگیری ماشین کمک کنند.

تعریف حکمرانی داده (Data governance)

تعاریف مختلفی از حکمرانی داده را می توان با تمرکز بر مسائل مختلف ارائه نمود:

موسسه MDM حاکمیت داده را اینگونه تعریف می کند:

"ترتیب رسمی افراد، فرآیندها و فناوری برای قادر ساختن سازمان به استفاده از داده ها به عنوان دارایی سازمانی"

- با تمرکز کلی بر افراد، فرآیندها و فناوری.

تعریف حکمرانی داده (Data governance)

تعاریف مختلفی از حکمرانی داده را می توان با تمرکز بر مسائل مختلف ارائه نمود:

موسسه Forrester حاکمیت داده را اینگونه تعریف می کند:

یک برنامه استراتژیک تجاری که داده‌های سود مالی سازمان‌ها را تعیین و اولویت‌بندی می‌کند و همچنین ریسک کسب‌وکار ناشی از شیوه‌ها و کیفیت داده‌های ضعیف را کاهش می‌دهد. در بطن این برنامه، مالکیت، مسئولیت پذیری، فرآیندها، برنامه ریزی و مدیریت عملکرد وجود دارد.

- با تمرکز بر مسئولیت امانتداری و برنامه ریزی سازمانی برای مدیریت داده ها.

تعریف حکمرانی داده (Data governance)

تعاریف مختلفی از حکمرانی داده را می توان با تمرکز بر مسائل مختلف ارائه نمود:

موسسه Data Governance Institute حاکمیت داده را اینگونه تعریف می کند:

«سیستمی از حقوق تصمیم‌گیری و مسئولیت‌پذیری برای فرآیندهای مرتبط با اطلاعات، اجرا شده بر اساس مدل‌های توافق‌شده، که توصیف می‌کند چه کسی می‌تواند با چه اطلاعاتی، چه زمانی، تحت چه شرایطی و با استفاده از چه روش‌هایی، چه اقداماتی را انجام دهد».

- با تمرکز بر مسئولیت های مدیران سطح متوسط و قوانین تعامل

تعریف حکمرانی داده (Data governance)

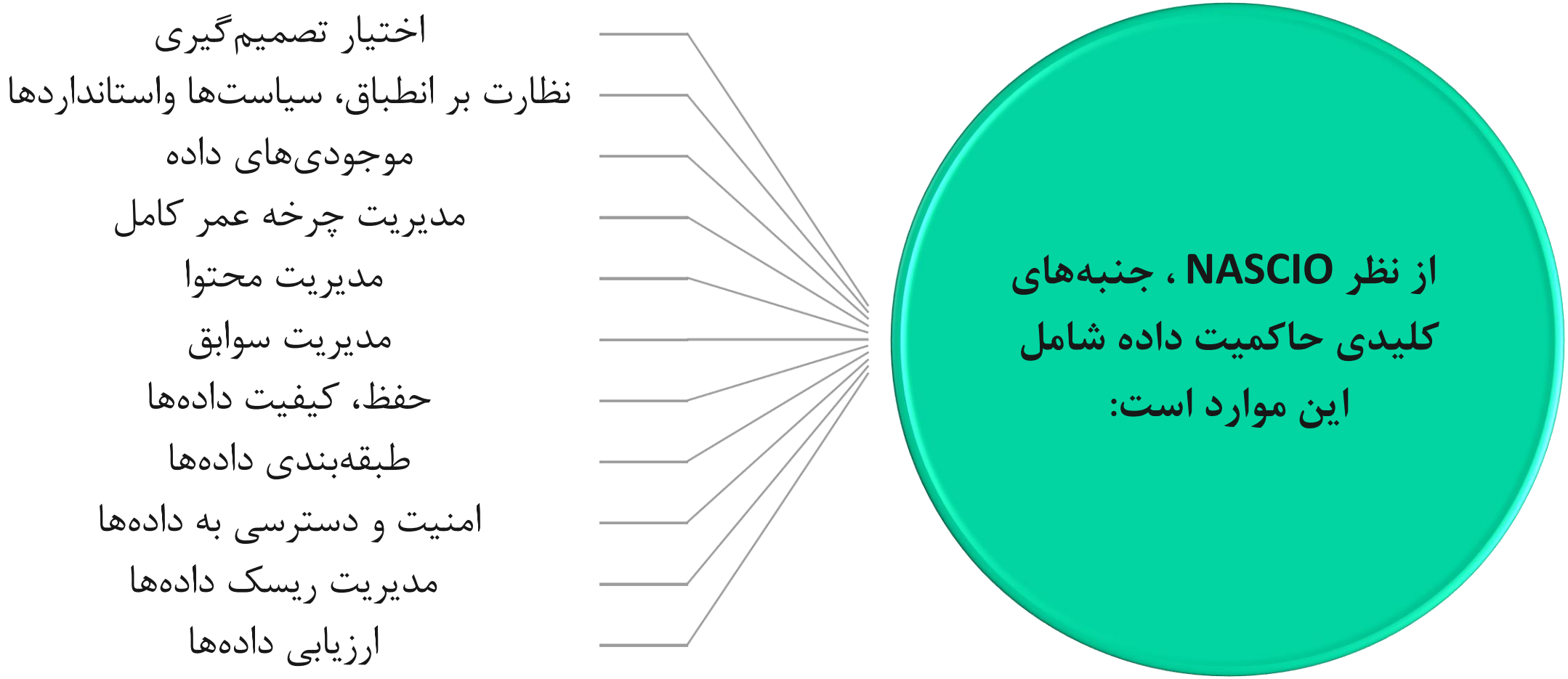
تعاریف مختلفی از حکمرانی داده را می توان با تمرکز بر مسائل مختلف ارائه نمود:

NASCIO حاکمیت داده را اینگونه تعریف می کند:

"رشته عملیاتی برای مدیریت داده ها و اطلاعات به عنوان یک دارایی کلیدی شرکت. این رشته عملیات شامل سازمان، فرآیندها و ابزارهایی برای ایجاد و اعمال حقوق تصمیم گیری در مورد ارزش گذاری و مدیریت داده ها است.

- با تمرکز بر قوانین تعامل و اصول عملیاتی برای مدیریت داده ها.

تعریف حکمرانی داده (Data governance)



تعریف حکمرانی داده (Data governance)

با توجه به تعاریف گفته شده، بطور کلی تعریف حکمرانی داده را می توان به صورت زیر خلاصه کرد:

قوانین تعامل برای نحوه مدیریت و حفظ داده ها توسط نهادها (افراد و سیاست ها) در سراسر شرکت و سازمان مربوطه، در طول چرخه عمر داده.

بنگاه اقتصادی در این زمینه شامل یک سازمان و ذینفعان داخلی و خارجی آن است. در ادامه این سند، این مضامین مشترک نشان دهنده حاکمیت داده است.

چارچوب های حاکمیت داده (DG Frameworks)

بطور کلی ، یک چارچوب حاکمیت داده توضیح می‌دهد که:

- چگونه تمام قطعاتی که حاکمیت داده را تشکیل می‌دهند با هم تناسب دارند.
- به توصیف مفاهیم اصلی و روابط متقابل آنها کمک می‌کنند.
- به سازماندهی پیچیدگی یک موضوع کمک می کنند.
- ارتباطات و بحث را تسهیل می کنند.

علاوه بر این، چارچوب‌های حاکمیت داده به نشان دادن چگونگی ارتباط حاکمیت داده با سایر جنبه‌های مدیریت داده، معماری داده و معماری سازمانی کمک می‌کنند.

چارچوب های حاکمیت داده (DG Frameworks)

Polices and procedures should cover four major areas:

Data Architecture – data models provide structure and requirements, single-source systems and metadata provide standardization

Data Quality – controls and monitoring to ensure accurate, complete, timely and consistent, checks throughout lifecycle for data movement, data use and data at rest

Data Security – begins with comprehensive inventory of databases and contents, data classifications drives requirements to ensure confidentiality, integrity, and availability

Data Usage – whether acquired or internally generated, reporting and risk modeling processes should reconcile and validate data, preference toward repeatable automated techniques to avoid manual errors

اسناد مورد نیاز برای ایجاد چارچوب حاکمیت داده های منطقه ای

- طرح کسب و کار داده که شامل:
  - چشم انداز و اهداف
  - چالش ها و نیازها
  - اولویت ها
  - فرآیند مدیریت
  - برنامه اقدام عملی
  - شیوه های تغییر مدیریت داده
- کاتالوگ مجموعه داده ها : لیست داده های موجود
  - داده های کسب و کار (تعیین فیچر های داده ها) برای مجموعه اولیه ها و زیر دامنه های تجاری
  - استانداردهای داده
  - مدل های داده
  - فرهنگ لغت داده
  - فرمت ها و روش های رابط فراداده: توضیحات و فایل‌هایی که داده‌ها را توصیف می‌کنند
- برنامه های بلند مدت مبتنی بر نظارت داده (توسط ذینفعان و حوزه تجاری) که شامل:
  - اولویت بندی مجموعه داده ها برای به روز رسانی و شرح مختصری از نحوه مدیریت اطلاعات (به عنوان مثال، تمدید، تبدیل، انتقال به فرمت قابل دسترس و غیره)
  - فرآیندهای جمع آوری، تبدیل و ترجمه داده ها
  - چرخه های به روز رسانی
  - فراداده و الزامات کیفیت رابط ها و روش های استاندارد و روش های دسترسی

توصیه هایی برای ایجاد چارچوب حاکمیت داده های منطقه ای

مرحله ۱: تعامل با ذینفعان

Stakeholder Engagement

- شناسایی ذینفعان
- توسعه رجیستری سهامداران
- برگزاری کارگاه طرح کسب و کار داده با ذینفعان کلیدی داده

مرحله ۲ : محدوده داده

Data Scope

- با هر یک از گروه های ذینفع، برای تعیین چالش های اصلی و کوتاه مدت، ملاقات کنید
- برنامه عملیاتی
- مجموعه داده های اولیه را بر اساس حوزه های کسب و کار یا ارزیابی در نظر بگیرید
- توسعه کاتالوگ داده (در ابتدا به عنوان Google Doc)

مرحله ۳ :استراتژی مباشرت

Steward Strategies

- برای هر موضوع مجموعه داده، اولویت ها و استانداردهای کیفیت و داده ها را توسعه دهید
- افزودن و انتشار کاتالوگ داده ها به صورت آنلاین برای ذینفعان اصلی

مرحله ۴ : ایجاد چارچوب حاکمیت داده

- منشور DGرا با ساختار سازمانی، نقش ها و مسئولیت ها توسعه دهید

ابزارهایی برای تجزیه و تحلیل داده ها در جهت افزایش ارائه خدمات

- به طور سنتی، دولت‌های محلی در مواردی، از «رویکرد ارتجاعی» هنگام تصمیم‌گیری در مورد زمان و مکان برآورده شدن نیازهای جامعه رنج می‌برند.

با این حال، شوراها در حال حاضر این توانایی را دارند که از داده‌های متنوع خود برای بهبود خدمات، بینشی به دست آورند.

- شهرهای هوشمند طبق تعریف، متخصصان کلان داده هستند. اگر بتوان اطلاعات مهم را استخراج و تجزیه و تحلیل کرد، کلان داده می تواند برای شهرهای هوشمند مفید باشد.

ابزارهایی برای تجزیه و تحلیل داده ها در جهت افزایش ارائه خدمات

داده ها نیازهای ما را برآورده می کنند و روشن می کنند که داده ها در همه جنبه های زندگی ما بومی هستند.

داده ها بر ما حکومت می کنند، قدرت را در اعداد جمع آوری می کنند و همیشه نقش اساسی در وجود ما داشته اند.

اینترنت اشیا یک «فناوری ضروری» است که بدون آن طرح‌های «شهر هوشمند» نمی‌توانند وجود داشته باشند. «چیزهای» اینترنت اشیا که شامل دستگاه‌ها، حسگرها و برنامه‌های کاربردی می‌شود، داده‌هایی را جمع‌آوری می‌کند که به راه‌حل‌های فناوری اجازه می‌دهد مؤثر واقع شوند.

کاربرد هوش مصنوعی در سیاست گذاری

هوش مصنوعی کاربردهای مختلف و فراوانی دارد بعنوان مثال، در حوزه‌های گوناگون زندگی اجتماعی، که به نوعی میدان عمل سیاست‌گذاری حساب می‌شوند، می توان از هوش مصنوعی استفاده نمود.

کاربرد هوش مصنوعی در سیاست گذاری و مدیریت دولت و تاثیراتش می‌تواند به صورت بنیادی بوده و پیش فرض‌ها و شاخص‌هایی که با استفاده از آن‌ها وضعیت و کیفیت یک حکومت ارزیابی می‌شود را تحت تاثیر قرار دهد. به طور مثال **انواع مدل‌های بهینه سازیِ توزیع یارانه‌ها، فهم شهروندان از عدالت و کارکردهای نهادهای نظام اقتصادی** را به شدت تحت تاثیر قرار خواهد داد .

در مثال دیگر می‌توان به **شاخص‌های مبارزه با فساد** که یکی از شاخص‌های اندازه‌گیری حکمرانی خوب به حساب می‌آید، اشاره کرد که به صورت جدی تحت تاثیر **استفاده از الگوریتم‌های جدید کشف فساد** قرار می‌گیرند. به طور کلی می‌توان این بخش را با عنوان جهش در سیستم‌های خبره در حوزه سیاست گذاری نیز نامگذاری کرد.

سیاست گذاری برای هوش مصنوعی

هوش مصنوعی در طبقه بندی انواع فناوری، در دسته فناوری‌های تحول آفرین قرار می‌گیرد و با توجه به شمول اثرات آن به صنایع و نهادها و طبقات مختلف جامعه، باید سیاست گذاری در ارتباط با آن کاملا جدی گرفته شود.

در صورتی که سیاست گذاری و به تبع آن برنامه ریزی و اجرای مناسبی در مورد هوش مصنوعی صورت نگیرد، قطعا جامعه و سیستم حاکمیتی را با مشکل روبرو خواهند شد.

حکمرانی توسط هوش مصنوعی

هوش مصنوعی بسیاری از معادلات و مفروضاتی که در حال حاضر دنیای سیاست گذاری و اداره امور عمومی را شکل داده است را به شدت تحت تاثیر قرار خواهد داد.

دموکراسی، بروکراسی، تخصص و بسیاری دیگر از کلیدواژه‌هایی که شکل دهنده نظام‌های اداری هستند، و با اهدافی مانند گسترش خیر عمومی، اداره بهتر امور ، شایسته سالاری و حفظ حاکمیت مردم مردم توسعه پیدا کرده‌اند، تحت تاثیر موضوع هوش مصنوعی باید به صورت جدی مورد باز تعریف قرار گیرند. علاوه بر این هیچ دور از ذهن نیست که در نتیجه رشد هوش مصنوعی هر کدام از این پایه‌های مدیریت در بخش عمومی به صورت کلی منسوخ شود.

پیشنهادات

- توجه به آموزش کارکنان و مدیران در خصوص علم داده و هوش مصنوعی
- گنجاندن احکامی در خصوص علم داده و حکمرانی داده در برنامه هفتم و ردیف بودجه ۱۴۰۲
- توجه ویژه به داده‌های ثبتی
- استفاده از قدرت داده‌ها در مدیریت استانها و حکمرانی منطقه‌ای
- ارتباط منسجم و یکپارچه بین معاونت‌های مختلف سازمان در خصوص استفاده از داده‌ها
- بکارگیری و استفاده از اکوسیستم فناوری (شرکت‌های دانش بنیان، استارت‌آپ‌ها، شتاب‌دهنده‌ها، صندوق‌ها، صندوق نوآوری و شکوفایی، ستادهای علم و فناوری و ...)

با تشکر

کورش پرند

استاد تمام گروه علوم داده ها و کامپیوتر دانشکده ریاضی

دانشگاه شهید بهشتی

آزمایشگاه هوش مصنوعی و محاسبات علمی

Contact info :

0912 489 3213

k_parand@sbu.ac.ir