



**گروه کامپیوتر – دانشکده مهندسی – دانشگاه فردوسی مشهد**

# مستندات گزارش کوتاه

**برای بخشی از فعالیت های کلاسی**

**درس روش پژوهش و شیوه ارائه مطالب علمی و فنی**

**موضوع:**

## علوم داده در کسب و کار

**Data science in Business**

**ارائه دهنده:**

**حمید رضا زه تاب**

**استاد درس:**

**دکتر رضا منصفی**

**زمستان 1400**

# چکیده

امروزه بیش از هرزمان دیگری داده داریم و با توجه به سریع تر شدن و قدرتمند شدن کامپیوتر ها و همچنین به لطف تکنولوژی ذخیره سازی اطلاعات می توانیم هزاران هزار ترابایت داده را مهار کنیم . به زبان ساده علم داده به ما کمک می کند تا این حجم گسترده از اطلاعات را برای خود قابل فهم کنیم . با کمک علم داده ، اطلاعات را طبقه بندی و دسته بندی میکنیم سپس آنها را مصور سازی می کنیم و سپس از داده ها نتیجه گیری می کنیم . نتیجه گیری نهایی می تواند در گستره بزرگی از شاخه ها من جمله کسب و کار به ما کمک کند . در این مقاله تلاش می کنیم تا ابتدا درک درستی از آنچه در علوم داده اتفاق می افتد ارایه کنیم و سپس به معرفی کاربرد های علوم داده در کسب و کار و نحوه به کار گیری آنها می پردازیم .

# فهرست مطالب


۱-مقدمه ..... ۱

۲- کاربرد های علوم داده ..... ۵

۲-۱-یادگیری ماشین (machine learning ) ..... ۵

۲-۲-اینترنت اشیاء( IOT ) ..... ۶

۲-۳-یادگیری عمیق ..... ۷

۳-ساخت تیم علوم داده ..... ۹

۳- ۱- اعضاء تیم ..... ۹

۳-۱-۱- مهندس داده ..... ۹

۳-۱-۲-تحلیلگر داده ..... ۱۰

۳-۱-۳-دانشمند یادگیری ماشین ..... ۱۱

۳-۱-۴-مرور ..... ۱۲

۶-جمع آوری داده های اضافی ..... ۱۳

۶-۱-حتی داده های بیشتر ..... ۱۳

۶-۱-۱-API های داده ..... ۱۳

6-1-2-سوابق عمومی ..... ۱۴

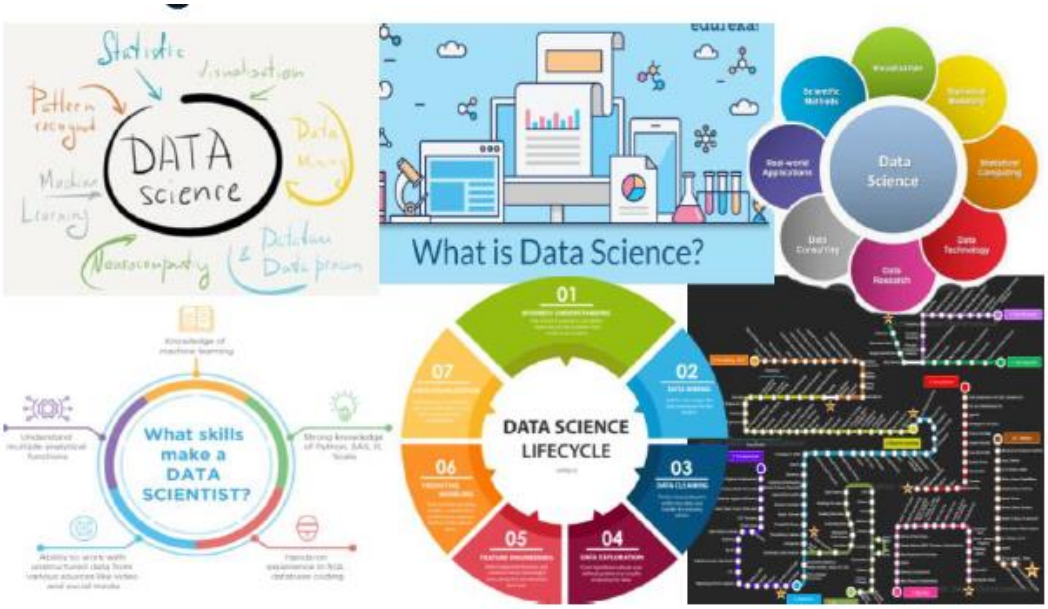
6-1-3-ترک مکانیکی ..... ۱۴

منابع ..... ۱۷

## ۱– مقدمه

علم داده یک رشته پویا است . ابزار هایی که ما استفاده می کنیم و توانایی های تیم هایمان هر روز در حال تغییر است . در ادامه خواهید دانست که علم داده چیست و چگونه می توانید از آن برای تقویت سازمان خود استفاده کنید .

اگر عبارت "علوم داده چیست ؟" را گوگل کنیم ، با حجم انبوهی از اطلاعات سردرگم کننده روبرو می شویم ، اما علم داده در حقیقت بسیار ساده است . علم داده مجموعه ای از روش ها برای گرفتن هزاران شکل از داده هایی است که امروزه در دسترس ما هستند و از آنها برای نتیجه گیری معنادار استفاده می کنیم . داده ها در همه جا هستند . هر لایک ، کلیک ، ایمیل ، تراکنش بانکی یا توییت یک داده جدید است که می تواند برای توصیف بهتر زمان حال یا پیش بینی بهتر آینده استفاده شود .




سوال اینجاست که داده ها چه کاری می توانند برای ما انجام دهند ؟ داده ها می توانند وضعیت فعلی ما را توصیف کنند . این را می توان با داشبوردها یا هشدارها انجام داد و فرآیندهای زمانبر گزارشدهی را با فناوری های جدید ساده کرد. علم داده می تواند به تشخیص رویدادهای غیر عادی کمک کند. اگر داده‌هایی در مورد آنچه قبلاً اتفاق افتاده است داشته باشیم، می‌توانیم با شناسایی خودکار یک رویداد جدید که غیرمنتظره است، کارایی را افزایش دهیم. داده ها همچنین می توانند علل رویدادها و رفتارهای مشاهده شده را تشخیص دهند. به جای تعیین همبستگی بین تعداد کمی از رویدادها، تکنیک‌های علم داده به ما کمک می‌کنند تا سیستم‌های پیچیده را با دلایل احتمالی متعدد درک کنیم. در نهایت، داده ها می توانند رویدادهای آینده را پیش بینی

کنند. ما می توانیم از تکنیک های جدید برای در نظر گرفتن علل مختلف و پیش بینی نتایج بالقوه استفاده کنیم. علاوه بر این، ما می توانیم احتمال پیش بینی خود را به صورت ریاضی ارزیابی کنیم تا سطح عدم قطعیت خود را روشن کنیم.

## 2– کاربرد های علوم داده

قبلاً با تعریف علم داده و مراحل یک گردش کار علم داده آشنا شدید. در ادامه ، شما یاد خواهید گرفت که چگونه علم داده را برای مشکلات واقعی کسب و کار به کار ببرید.

بیایید به سه حوزه مهیج علم داده نگاهی بیندازیم: یادگیری ماشین ، اینترنت اشیا و یادگیری عمیق.

### 2–1– یادگیری ماشین

فرض کنید در یک بانک بزرگ در اداره کشف کلاهبرداری و تقلب ( fraud detection )کار می کنید. می‌خواهید از داده‌ها برای تعیین احتمال جعلی بودن تراکنش استفاده کنید.

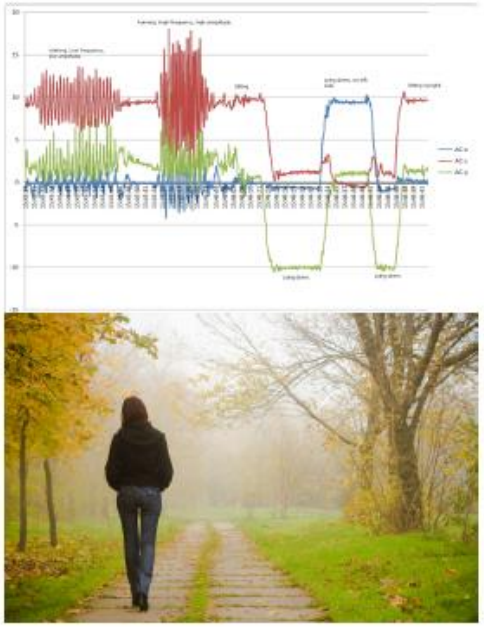
برای پاسخ به این سوال، ممکن است با جمع آوری اطلاعات در مورد هر خرید، مانند مبلغ، تاریخ، مکان، نوع خرید و آدرس دارنده کارت شروع کنید. شما به نمونه های زیادی از تراکنش ها، از جمله این اطلاعات، و همچنین برچسبی نیاز دارید که معتبر یا تقلبی بودن هر تراکنش را به شما بگوید . خوشبختانه، شما احتمالاً این اطلاعات را در یک پایگاه داده دارید. این رکوردها "داده های آموزشی( training data )" نامیده می شوند و برای ساخت یک الگوریتم استفاده می شوند. هر بار که یک تراکنش جدید رخ می دهد، اطلاعات الگوریتم خود مانند مبلغ و تاریخ را می دهید و به سوال اصلی پاسخ می دهد: احتمال تقلبی بودن( fraudulent ) این تراکنش چقدر است؟


| Amount | Date | Type | ... |
|---|---|---|---|
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |
| . | . | . | . |

قبل از اینکه بتوانیم به این سوال پاسخ دهیم، بیایید مثال خود را مرور کنیم و آنچه را که برای یادگیری ماشینی نیاز داریم تا جادوی خود را انجام دهد برجسته کنیم. اول، یک مسئله علم داده با یک سوال کاملاً تعریف شده شروع می شود. سوال ما این بود که "احتمال تقلبی بودن این تراکنش چقدر است؟" در مرحله بعد، ما به برخی داده ها برای تجزیه و تحلیل نیاز داریم. ماه‌ها تراکنش‌های قدیمی کارت اعتباری و ابرداده‌های مرتبط داشتیم که قبلاً تقلبی یا معتبر شناخته شده بودند. در نهایت، هر بار که می‌خواهیم پیش‌بینی و قضاوت جدیدی انجام دهیم، به داده‌های اضافی نیاز داریم. ما باید در مورد هر خرید جدید اطلاعات یکسانی داشته باشیم تا بتوانیم آن را به عنوان "تقلبی" یا "معتبر" بدانیم.

## 2–2– اینترنت اشیاء

حال، فرض کنید قصد دارید یک ساعت هوشمند برای نظارت بر فعالیت بدنی بسازید. شما می خواهید بتوانید فعالیت های مختلف مانند راه رفتن یا دویدن را به طور خودکار تشخیص دهید. ساعت هوشمند شما مجهز به حسگر ویژه ای به نام «شتاب سنج» است که حرکت را به صورت سه بعدی کنترل می کند. داده های تولید شده توسط این سنسور اساس یادگیری ماشین شما است. می توانید از چندین داوطلب بخواهید که ساعت شما را بپوشند و فعالیت خود را هنگام دویدن یا راه رفتن ضبط کنند. سپس می توانید الگوریتمی ایجاد کنید که داده های شتاب سنج را به عنوان نماینده یکی از این دو حالت تشخیص می دهد: راه رفتن یا دویدن.

ساعت هوشمند شما بخشی از یک زمینه در حال رشد به نام "اینترنت اشیا" است که به عنوان IOT نیز شناخته می شود، که اغلب با علم داده ترکیب می شود. اینترنت اشیا به ابزارهایی اطلاق می‌شود که کامپیوتر های استاندارد نیستند، اما همچنان توانایی انتقال داده‌ها را دارند. این شامل ساعت‌های هوشمند، سیستم‌های امنیتی خانگی متصل به اینترنت، سیستم‌های جمع‌آوری عوارض الکترونیکی، سیستم‌های مدیریت انرژی ساختمان و بسیاری موارد دیگر می‌شود . داده های تولید شده توسط اینترنت اشیا منبع عالی برای پروژه های علم داده است!

## 2–3– یادگیری عمیق

بیایید به مثال دیگری بپردازیم. یکی از وظایف کلیدی برای خودروهای خودران، شناسایی زمانی است که تصویری حاوی یک انسان است. مجموعه داده برای این مشکل چه خواهد بود؟

ما می توانیم تصویر را به صورت ماتریسی از اعداد بیان کنیم که در آن هر عدد نشان دهنده یک پیکسل است. با این حال، اگر ماتریس را به یک مدل یادگیری ماشین وارد کنیم، احتمالاً این رویکرد شکست خواهد خورد. به سادگی داده های ورودی بیش از حد وجود دارد!

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 1 | 3 | 1 | 1 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 1 | 3 | 1 | 1 | 2 |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 2 | 3 | 1 | 3 | 3 | 1 | 1 | 2 |
| 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 48 | 20 | 20 | 20 | 8 | 8 |
| 6 | 6 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 20 | 20 | 20 | 8 | 5 |
| 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 20 | 20 | 20 | 5 | 5 |
| 4 | 4 | 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 |
| 4 | 4 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |

ما به الگوریتم پیشرفته تری نیاز داریم که به عنوان یادگیری عمیق شناخته می شود. در یادگیری عمیق ، چندین لایه از مینی الگوریتم‌ها به نام « نورون‌ها » با هم کار می‌کنند تا نتیجه‌گیری‌های پیچیده‌ای را به دست آورند. یادگیری عمیق نسبت به یک مدل یادگیری ماشینی به داده های آموزشی بسیار بسیار بیشتری نیاز دارد، اما همچنین می تواند روابطی را بیاموزد که مدل های سنتی نمی توانند. یادگیری عمیق برای حل مشکلات پیچیده ای مانند طبقه بندی تصویر یا درک زبان استفاده می شود.

## 3– ساخت تیم علوم داده

در ادامه شما یاد خواهید گرفت که چگونه تیم داده خود را برای برآوردن نیازهای سازمان خود بسازید و ساختار دهید.

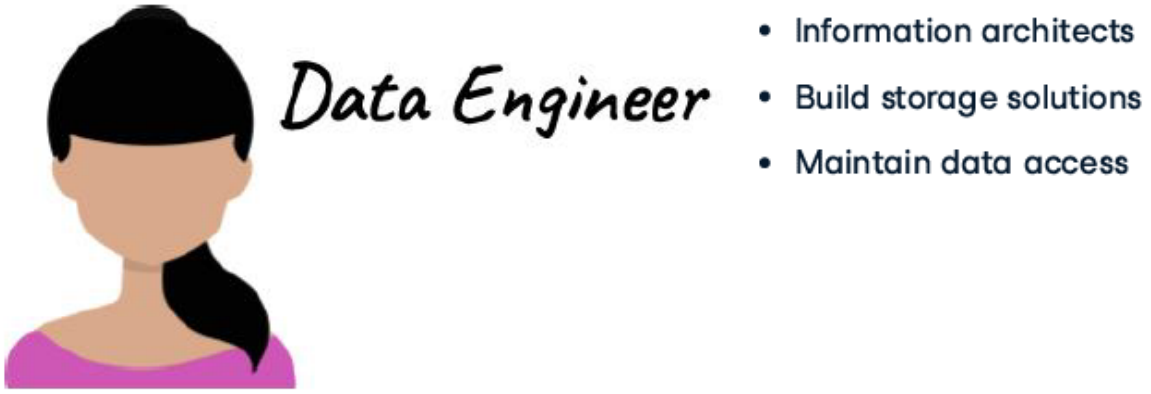
### 3–1– اعضاء تیم

شاید تعجب کنید که بدانید "علم داده" یک رشته واحد نیست. این در واقع سه شغل مختلف است: مهندس داده، تحلیلگر داده، و دانشمند یادگیری ماشین. بیایید هر یک را بررسی کنیم.




#### 3–1–1– مهندس داده

مهندسان داده جریان اطلاعات را کنترل می کنند: آنها سیستم های ذخیره سازی داده ها و زیرساخت های تخصصی ایجاد می کنند تا اطمینان حاصل کنند که داده ها به راحتی به دست می آیند و پردازش می شوند.



اکثر مهندسان داده با SQL آشنا هستند که از آن برای ذخیره و مدیریت کلان داده ها استفاده می کنند. آنها همچنین از یکی از زبان های برنامه نویسی  مانند جاوا، اسکالا یا پایتون برای پردازش داده ها و خودکارسازی وظایف مربوطه استفاده می کنند.


### 3–1–2– تحلیلگر داده

تحلیلگران داده ، حال را از طریق داده ها توصیف می کنند. آنها این کار را با داشبورد، آزمون فرضیه ها( hypothesis tests ) و تجسم( visualization ) انجام می دهند. آنها اغلب پیشینه ای در زمینه آمار یا علوم کامپیوتر دارند، اما تمایل دارند تجربه مهندسی کمتری نسبت به مهندسان داده و تجربه ریاضی کمتری نسبت به دانشمندان داده داشته باشند.


تحلیلگران داده از  spreadsheets  برای انجام تجزیه و تحلیل های ساده بر روی مقادیر کم داده استفاده می کنند. آنها ازSQL ، همان زبانی که توسط مهندسان داده استفاده می شود، برای تحلیل های بزرگتر استفاده می کنند. در حالی که مهندسان داده راه حل های ذخیره سازی SQL را می سازند و پیکربندی می کنند، تحلیلگران داده از پایگاه های داده موجود برای مصرف و خلاصه کردن داده ها استفاده می کنند. تحلیلگران همچنین از هوش تجاری یا  BI ، ابزارهایی مانند  Tableau ، Power BI  یا Looker  برای ایجاد داشبورد و به اشتراک گذاری تحلیل های خود استفاده می کنند.

Data Analyst

- Spreadsheets (Excel or Google Sheets)
  - Simple storage and analysis
- SQL
  - Large-scale analysis
- BI Tools (Tableau, Power BI, Looker)
  - Dashboarding and sharing information

### 3-1-3- دانشمند یادگیری ماشین

یادگیری ماشین شاید پرهیاهوترین بخش علم داده باشد. از آن برای تشخیص و حدس آنچه که احتمالاً درست است از آنچه قبلاً می دانیم استفاده می شود. این دانشمندان از داده های آموزشی برای طبقه بندی داده های بزرگتر و غیر طبقه بندی شده استفاده می کنند. یادگیری ماشینی می‌تواند به ما بگوید ارزش یک سهام در هفته آینده چقدر است، کدام تصاویر حاوی یک ماشین هستند یا چه احساساتی توسط مجموعه‌ای از توییت‌ها بیان می‌شوند.

Machine Learning Scientist

- Predictions and extrapolations
- Classification
- Stock price prediction
- Image processing
- Automated text analysis

دانشمندان یادگیری ماشین از پایتون یا R برای ایجاد مدل های پیش بینی خود استفاده می کنند. هر دو زبان برنامه نویسی عالی برای علم داده هستند و فردی که یک زبان را می داند احتمالاً می تواند کد را در زبان دیگر بخواند. به یاد داشته باشید، یادگیری زبان های برنامه نویسی به اندازه زبان های گفتاری دشوار نیست. اگر کسی بلد باشد فرانسوی صحبت کند، ممکن است سالها طول بکشد تا زبان آلمانی را یاد بگیرد. زبان های برنامه نویسی بیشتر شبیه ابزارها هستند. اگر می دانید چگونه از دریل برقی استفاده کنید، لزوماً نحوه استفاده از اره برقی را بلد نیستید، اما احتمالاً می توانید با کمی آموزش یاد بگیرید!

Machine Learning Scientist

- Python and R
  - Programming languages for creating predictive models

### 3–1–4– مرور

برای جمع‌بندی: مهندسان داده داده‌ها را ذخیره و نگهداری می‌کنند، تحلیلگران داده‌ها داده‌ها را تجسم( visualize ) و توصیف می‌کنند، و دانشمندان یادگیری ماشین با داده‌ها مدل‌سازی و پیش‌بینی می‌کنند. هر موقعیت از مجموعه کمی متفاوت از ابزارها برای رسیدن به اهداف خود استفاده می کند.

| Data Engineer | Data Analyst | Machine Learning Scientist |
|---|---|---|
| Store and maintain data | Visualize and describe data | Model and predict with data |
| SQL + Java/Scala/Python | SQL + BI Tools + Spreadsheets | Python/R |

## 6– جمع آوری داده های اضافه ( additional data )

در حالی که جمع‌آوری داده‌های داخلی برای برخی از پروژه‌های علم داده مفید است، اما تنها یک تکه از این پازل است. اغلب، شما نیاز به جمع آوری داده ها از منابع خارجی نیز دارید.


### 6–1– حتی داده های بیشتر

راه های زیادی وجود دارد که می توانید داده های اضافی را برای سازمان خود جمع آوری کنید. چند راه متداول شامل APIها، رکوردهای عمومی و Mechanical Turk است که در ادامه به همه آنها خواهیم پرداخت.

#### 6–1–1– API های داده

بیایید با API ها شروع کنیم .API مخفف Application Programming Interface است. این یک راه آسان برای درخواست داده از شخص ثالث از طریق اینترنت است. بسیاری از شرکت‌ها API دارند تا به تیم شما اجازه دسترسی به داده‌هایشان را بدهند. برخی ازAPI های قابل توجه عبارتند ازTwitter ، Wikipedia، Yahoo! ، و نقشه های گوگل، اما بسیاری از موارد دیگر وجود دارد. اگر با یک شریک کار می کنید و فکر می کنید که آنها ممکن است داده های مفیدی داشته باشند، یک جستجوی سریع در وب انجام دهید و ببینید آیا یک API وجود دارد یا خیر!


### 6–1–2– سوابق عمومی

سوابق عمومی یکی دیگر از راه های عالی برای جمع آوری داده های اضافی است. در ایالات متحده، data.gov داده های بهداشت، آموزش و تجارت را برای دانلود رایگان در دسترس دارد. در اتحادیه اروپا، data.europa.eu داده های مشابهی دارد. اینها می توانند منابع خوبی برای درک روندهای سطح جمعیت(population-level trends) یا جمع آوری داده های اقتصادی باشند.

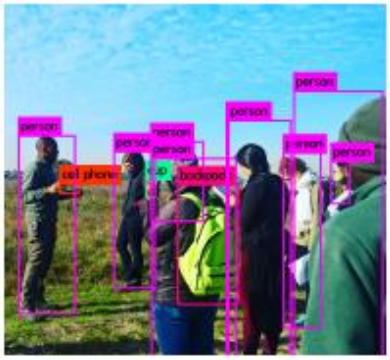


.

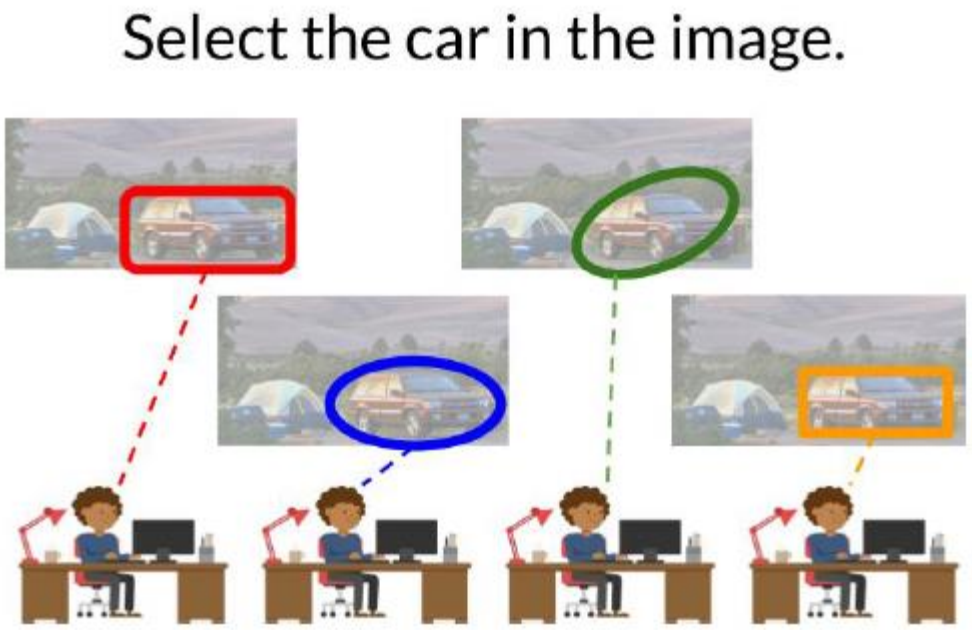
### 6–1–3– ترک مکانیکی

قبلا ، ما در مورد تشخیص تصویر به عنوان یک نوع مشکل علم داده بحث کردیم. برای ساختن یک الگوریتم تشخیص تصویر خوب، به مجموعه ای از تصاویر نیاز داریم که در آن تصاویر قبلاً برچسب گذاری شده باشند که به آن مجموعه آموزشی

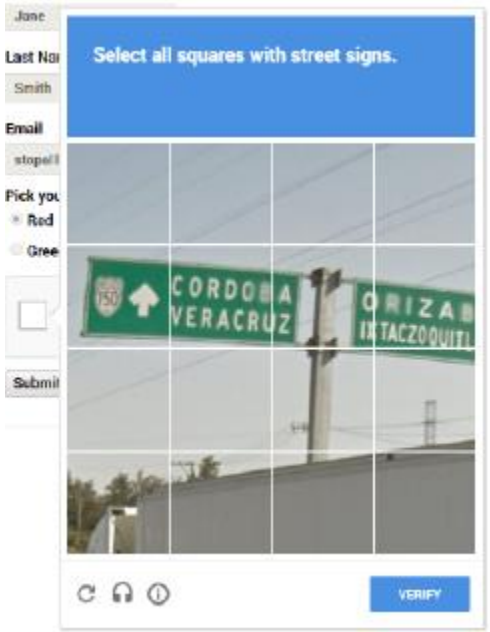
(training set) می گویند. اما ما فقط به یک یا دو عکس نیاز نداریم. ما به صدها یا هزاران عکس نیاز داریم. دریافت این تصاویر برچسب‌گذاری شده می‌تواند واقعاً سخت و زمان‌بر باشد و فقدان مجموعه آموزشی اغلب باعث می‌شود پروژه‌های علمی داده خوب تکمیل نشوند.


بسته به نوع مجموعه آموزشی مورد نیاز، Mechanical Turk که MTurk نیز نامیده می شود، می تواند یک گزینه عالی باشد. ترک مکانیکی به معنای درخواست از انسان برای تکمیل کاری است که ما در نهایت برای کامپیوتری کردن آن عمل برنامه ریزی می کنیم. در مثال قبلی ما، این به معنای برچسب زدن تعدادی عکس برای ایجاد یک مجموعه آموزشی برای تشخیص تصویر است. به جای اینکه از یک نفر بخواهیم هزاران تصویر را برچسب گذاری کند، هزاران نفر را استخدام می کنیم و به هر یک از آنها برای برچسب زدن چند تصویر پول می دهیم. برای اطمینان از کیفیت، ممکن است از دو یا سه نفر بخواهیم تصویر مشابه را مرور کنند و سپس رایج ترین پاسخ را در نظر می گیریم .




پلتفرم‌های زیادی برای کمک به Mechanical Turkو استخدام کمک‌کنندگان مانند AWS MTurk وجود دارد. ترک مکانیکی فقط برای تشخیص تصویر نیست. همچنین می توانید از آن برای برچسب زدن نظرات مشتریان به عنوان مثبت یا منفی، استخراج متن از فرم یا برجسته کردن کلمات کلیدی در یک جمله استفاده کنید. در مثال سمت راست، از کاربران خواسته می شود که مشخص کنند کدام بخش از تصویر دارای علامت خیابان است.



## منابع


1. Data Science for Business and Decision Making by Luiz Favero and Patrícia Belfiore
2. Harvard Business School at online.hbs.edu
3. Harvard Business Review at hbr.org
4. wikipedia.org
5. ibm.com/cloud/learn/data-science-introduction
6. coursera.org/browse/data-science
7. http://cloudscaling.com/blog/cloud-computing
8. https://www.redhat.com/en/technologies/management
9. https://github.com
10. https://www.darkreading.com/cloud