علوم داده

مقدمه

علوم داده[1] حوزه ای بسیار گسترده و میان رشته‌ای است که از بخش های مختلفی مانند آمار و ریاضی، هوش مصنوعی و یادگیری ماشین، و مهندسی نرم‌افزار و برنامه‌نویسی بهره می‌برد. هدف علوم داده، استفاده از روش‌های آماری و الگوریتم‌های یادگیری ماشین جهت پردازش داده های گوناگون و معمولا غیرساختاریافته، مانند نوشته، صدا، عکس، و یا فیلم، در حجم های بسیار زیاد است تا بتوان از آن اطلاعات مفیدی استخراج کرد. به عنوان مثال، با استفاده از اطلاعات مختلفی که می‌توانیم از طریق سنسورهای موجود در طبیعت جمع‌آوری کنیم، احتمال وقوع زلزله را پیش‌بینی کنیم، و یا صدها هزار پست در شبکه های اجتماعی مثل توییتر را پردازش کرده و نظر یا طرز دید مردم را نسبت به یک موضوع خاص (اجتماعی، اقتصادی یا سیاسی) بسنجیم.

یک مهندس علوم داده، آماردانی است که از مهارت‌های برنامه نویسی لازم برای پردازش حجم‌های عظیم داده های دیجیتالی برخوردار بوده و با الگوریتم‌ها و روش‌های روز یادگیری ماشین آشنا باشد. به عبارت دیگر، او یک مهندس نرم‌افزار کامپیوتر است که از دانش آماری خوبی برخوردار بوده و توانایی بهره‌گیری از آنها در مسائل مختلف یادگیری ماشین را داشته باشد.

حوزه‌های دانشی مختلف که علوم داده از آن‌ها بهره می‌برد:

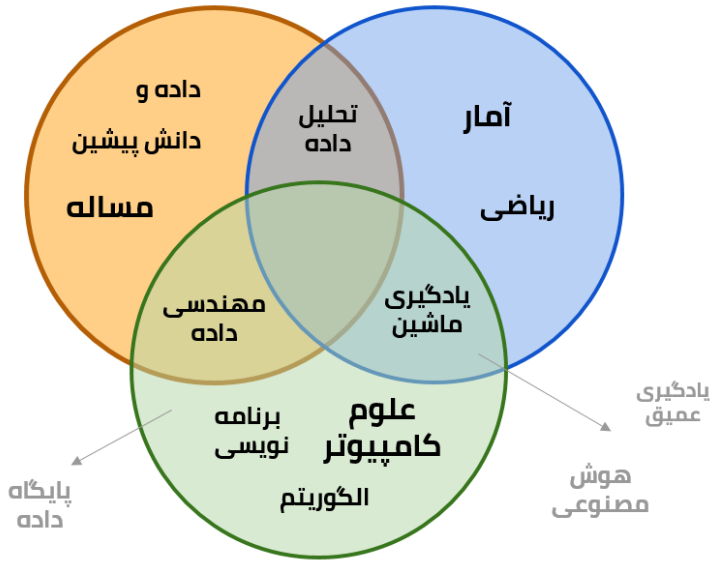
- **آمار:** پردازش داده تخصص های مختلفی میخواهد که یکی از مهمترین آنها آمار است. از آنجاییکه داده ها معمولا در جداول عریض و طویل ارائه میشوند، اولین قدم هنگام تحلیل داده‌ها نمایش تصویری مناسب آنها است. یک شکل خوب میتواند به اندازه هزاران کلمه اطلاعات منتقل کند. قدم اساسی بعدی یافتن مدل های ریاضی و آماری مناسب است. مثلا در مورد نحوه انتشار کرونا، اولین قدم یافتن مدل‌های دقیق ریاضی از نحوه انتشار ویروس است. این مدلها امکان پیشبینی آینده را به ما میدهند. نهایتا میزان اطمینان ما از دقت پیش بینی هم اهمیت دارد.

[1] Data Science

- **یادگیری ماشین:** یکی دیگر از ابزارهای اصلی برای تحلیل داده روش های یادگیری ماشین است. یادگیری ماشین یکی از زیرشاخه های اصلی هوش مصنوعی است: هدف هوش مصنوعی طراحی الگوریتم هایی ست که به کامپیوتر قابلیت انجام کارهای هوشمندانه بدهد (مثلا خواندن و درک محتوای متون، دیدن و فهمیدن عکس و فیلم، و...). در حال حاضر، یادگیری عمیق (روش های یادگیری مبتنی بر شبکه های عصبی) روش غالب در یادگیری ماشین است. به عنوان مثال، درمورد موضوع کرونا، در حال حاضر تحقیقاتی مبتنی بر یادگیری ماشین برای استفاده از صدا و سرفه افراد مبتلا جهت تشخیص از راه دور ابتلای آنها به کرونا در دست انجام است. ایده این است که صدای افراد همانند اثر انگشتشان حاوی اطلاعات یکتایی در مورد بیماری های آنها است که تشخیص دقیق آن تنها برای کامپیوتر ممکن است. برای کسب اطلاعات بیشتر می‌توانید به لینک‌های زیر مراجعه نمایید:
  https://www.covid-19-sounds.org/en/
  https://voca.ai/corona-virus/

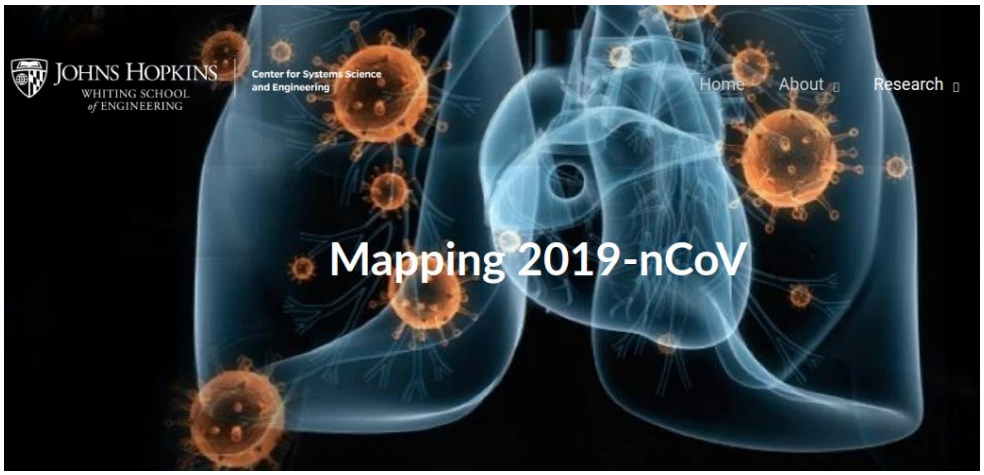
- **برنامه نویسی:** نهایتا یک دانشمند علم داده به طور روزمره برنامه های تحلیل داده می نویسد. برای این کار نیاز به فراگیری برنامه نویسی دارد. در برنامه نویسی، مهم ترین رکن مسلط بودن بر ساختمان داده و الگوریتم است. همچنین برای ذخیره سازی و بازیابی داده های حجیم نیاز به فراگیری دیتابیس است. با توجه به گسترش تکنولوژی های مبتنی بر شبکه، مثل موبایل، یک مهندس علوم داده در بسیاری از اوقات نیاز دارد تا کاربردهای مبتنی بر شبکه را مد نظر قرار دهد. در نهایت امر، تست کردن و درستی سنجی سیستم های تولیدی از اهمیت بالایی برخوردار است.

درنهایت می‌توان گفت تمامی حوزه‌های دانشی اشاره شده، بخشی از موضوعات مورد بررسی در ***رشته مهندسی کامپیوتر*** هستند.



کاربردهای علوم داده در جهان

علوم داده کاربردهای گسترده‌ای در دنیای امروز دارد. به عنوان مثال شیوع بیماری کرونا را در نظر بگیرید. درک ما از این بیماری و واکنش آن به داروهای مختلف و یا شرایط آب و هوایی مختلف هنوز ناقص است. در این راستا دانشگاه جانز هاپکینز داده‌های گوناگونی درباره این بیماری را جمع آوری کرده و بروز رسانی میکند. استفاده مناسب از این حجم عظیم داده در حیطه تخصصی علوم داده قرار می‌گیرد.



جایگاه رشته علوم داده در دنیا

پس از اتمام دوره تحصیلی و فراگیری مفاهیم بنیادی علوم داده، موقعیت‌های شغلی بسیاری چه در دانشگاه و چه در صنعت وجود دارد. در حال حاضر در دانشگاه‌ها محققین در حال کار بر روی مسائل بسیار زیادی با تنوع بالا بوده و فرصت‌های زیادی برای ادامه تحصیل در این رشته وجود دارد. در صنعت نیز یک متخصص علوم داده در سراسر دنیا بالاترین دستمزدها را دریافت می‌کند. به علاوه، امروزه اکثر دانشگاه‌های مطرح دنیا یک مرکز تحقیقاتی مربوط به تحقیقات علوم داده دارند که در حال کار کردن در مورد آخرین مسائل روز در زمینه‌های گوناگون علوم داده هستند.

علوم داده در تیاس

موسسه پژوهش‌های پیشرفته تهران (تیاس) در سال ۹۹ برای اولین بار اقدام به جذب دانشجو در رشته مهندسی کامپیوتر تحت گرایش علوم داده نموده است. درحال حاضر دوره های علوم داده موجود در ایران، در زیرگروه ریاضی کاربردی و آمار هستند. تیاس برای اولین بار در ایران گرایش علوم داده را به عنوان یک شاخه از مهندسی (کامپیوتر) ارائه نموده است.

برنامه های علوم داده درگروه علوم پایه به صورت محدود و در دانشکده‌های ریاضی و علوم کامپیوتر چند دانشگاه در تهران برگزار شده و عموما تاکید بر جنبه های ریاضی و ابعاد نظری علوم داده دارند. هرچند جنبه ی مهم دیگر در علوم داده، مهندسی سیستم‌های تحلیل و پردازش داده است که در دوره‌های گروه علوم پایه مورد تاکید قرار نمی‌گیرند. به عنوان مثال، درس پایگاه داده یکی از جنبه‌های سیستمی علوم داده است که ناظر بر چگونگی استخراج و بازیابی داده‌ها در عمل است که در جنبه ی تئوری خیلی مورد تاکید نیست. در ضمن، داده‌کاوی به معنای استخراج اطلاعات نهان یا الگوهای نهفته در حجم زیادی از داده است. داده‌کاوی اصطلاحی مربوط به دهه نود میلادی است که بیشتر در حوزه‌های مرتبط با پایگاه‌های داده به‌کار گرفته می‌شده است. در مقابل، علوم داده به مفاهیم وسیع‌تری از پردازش داده اشاره می‌کند که داده‌کاوی تنها یک زمینه از آن است. در علوم داده مفاهیمی نظیر جمع‌آوری، پالایش، تجرید، تبدیل و تحول داده مطرح می‌شوند که جزئی از رشته داده کاوی به معنای سنتی آن نیستند.
این در حالیست که گرایش ارائه شده برای رشته مهندسی کامپیوتر رویکردی عملیاتی و کاربردی داشته و بر جنبه‌های سیستمی و مهندسی نرم‌افزار تمرکز دارد. دانش‌آموختگان این گرایش مسلط بر ابعاد اجرایی و عملیاتی علوم داده بوده، توانایی پردازش و تحلیل سیستماتیک انواع داده در ابعاد بالا را داشته، و بر جنبه‌های تئوری آماری و مباحث امنیت اطلاعات و ... اشراف خواهند داشت.

ارتباط با گرایش هوش مصنوعی

تمرکز گرایش هوش مصنوعی بر بهبود الگوریتم‌ها و روش های یادگیری ماشینی بوده و در برگیرنده‌ی بخش عمده‌ای از جنبه‌های سیستمی و عملیاتی علوم داده نیست. به صورت خاص، دروس نرم‌افزاری مثل پایگاه داده، سیستم‌های توزیع شده و امنیت داده از مباحث این گرایش نبوده و بر دروس نظری مانند استنتاج آماری و تحلیل داده در ابعاد بالا تاکیدی ندارد. این در حالیست که گرایش علوم داده بر استفاده از روش‌های یادگیری ماشین برای پردازش حجم بالای داده در حوزه‌های مختلف به منظور دستیابی به اهداف خاص تمرکز کرده و رویکرد اجرایی آن به جنبه تئوری غالب است. علاوه بر این، تکیه‌ی این گرایش برای تحلیل و پردازش داده لزوما مبتنی بر روش‌های یادگیری ماشینی نبوده و از ابزارها و روش‌های استنتاج آماری و تحلیل نظری نیز در راستای این هدف بهره می برد. تمرکز بر مهندسی نرم‌افزار در نرم‌افزارهای داده محور نیازمند تخصصی خاص و بدیع است که هم اکنون در هیچ یک از زیرشاخه‌های مهندسی کامپیوتر، حتی شاخه‌ی مهندسی نرم‌افزار نیز به طور کامل و عمیق پوشش داده نمی‌شود. این تخصص نیاز به تاکید بر داده در تولید و آزمون نرم‌افزارهای تحلیل داده‌های حجیم دارد که فراتر از روش‌های سنتی تولید و آزمون نرم‌افزارهای تجاری و صنعتی است: تاکید بر نیازمندی‌های داده محور و پشتیبانی از حجم عظیم داده در کل مراحل تولید نرم‌افزار موجب به‌وجود آمدن روش‌های جدید مهندسی نرم‌افزار و همچنین ابزارهای سیستمی جدید برای نرم‌افزارهای داده محور شده است که در این زیر شاخه‌ی جدید مهندسی کامپیوتر پوشش داده خواهد شد.

حوزه‌های مختلف علوم داده توسط اساتید تیاس به شرح زیر پوشش داده می‌شود:

دکتر امین امین زاده گوهری: دکترا از دانشگاه برکلی، دارای مدال طلای المپیاد جهانی ریاضی

- زمینه تحقیقاتی: آمار و نظریه اطلاعات

دکتر حسین حجت: دکترا از EPFL، استاد همکار دانشگاه کرنل

- زمینه تحقیقاتی: مهندسی نرم افزار و روش های صوری

دکتر محمد طاهر پیله‌ور: دکترا از ساپینزای ایتالیا، استاد همکار دانشگاه کمبریج انگلستان

- زمینه تحقیقاتی: هوش مصنوعی و پردازش زبان طبیعی

درنهایت می‌توان گفت فضای حرفه‌ای و درعین حال صمیمانه موسسه تیاس، شرایط را برای انجام تحقیقات جدی با اساتید تمام وقت و پروژه های مشترک با محققین بین‌المللی تسهیل نموده است.