

به نام خدا



# علم داده از تئوری تا عمل

مؤلفان

محمد جواد جعفری

کارشناسی ارشد مهندسی فناوری اطلاعات

پرمان محمد علیزاده

کارشناسی معماری کامپیوتر

هرگونه چاپ و تکثیر از محتویات این کتاب بدون اجازه کتبی ناشر ممنوع است. متخلفان به موجب قانون حمایت حقوق مؤلفان، مصنفان و هنرمندان تحت پیگرد قانونی قرار می‌گیرند.

## علم داده از تئوری تا عمل

مؤلفان: محمد جواد جعفری

پرمان محمد علیزاده

ناشر: مؤسسه فرهنگی هنری دیباگران تهران

حروفچینی و صفحه‌آرایی: شبنم هاشم زاده

طرح روی جلد: داریوش فرسای

چاپ: درج عقیق

نوبت چاپ: سوم

تاریخ نشر: ۱۳۹۸

تیراژ: ۱۰۰ جلد

قیمت: ۹۵۰۰۰۰ ریال

شابک: ۹۷۸-۶۰۰-۱۲۴-۵۸۰-۰

نشانی واحد فروش: تهران، میدان انقلاب،

خ کارگر جنوبی، روبروی پاساژ مهستان،

پلاک ۱۲۵۱

تلفن: ۲۲۰۸۵۱۱۱-۶۶۴۱۰۰۴۶

کد پستی: ۱۳۱۴۹۸۳۱۸۵

فروشگاههای اینترنتی:

[www.mftbook.ir](http://www.mftbook.ir)

[www.dibagaran-tehran.com](http://www.dibagaran-tehran.com)

نشانی تلگرام: @mftbook

نشانی اینستاگرام: Dibagaran\_publishing

سرشناسه: جعفری، محمد جواد، ۱۳۶۷-

عنوان و نام پدید آور: علم داده از تئوری تا عمل / مؤلفان  
محمد جواد جعفری، پرمان محمد علیزاده

مشخصات نشر: تهران - دیباگران تهران - ۱۳۹۶

مشخصات ظاهری: ۳۷۰ ص. مصور.

شابک: ۹۷۸-۶۰۰-۱۲۴-۵۸۰-۰

وضعیت فهرست نویسی: فیپا

موضوع: داده کاوی

موضوع: Data mining

شناسه افزوده: محمد علیزاده، پرمان، ۱۳۷۴-

رده بندی کنگره: ۱۳۹۶ ج ۶ ۷۶/۹/۵۲ QA

رده بندی دیویی: ۰۰۶/۳۱۲

شماره کتابشناسی ملی: ۴۶۶۳۸۷۷

## فهرست مطالب

21	بخش اول
21	شروع کار با علم داده
23	فصل اول
23	نگاهی اجمالی بر علم داده‌ها
24	بررسی راهکارهای پیشنهادی علم داده
24	ایجاد یک تیم درون سازمانی
25	برون‌سپاری کارها به مشاورین خصوصی علم داده
26	استفاده از راه‌حل‌های ابر محور
26	مزایای محسوس علم داده
27	فصل دوم
27	بررسی زیرساخت‌های مهندسی داده
28	تعریف کلان داده با 4 ویژگی اصلی آن
30	تنوع داده‌ها
31	منابع کلان داده کدامند؟
31	تفاوت بین مهندسی داده و علم داده
32	تعریف علم داده
32	تعریف مهندسی داده
33	مقایسه‌ی متخصصین داده و مهندسان داده
34	محدود کردن داده‌ها با استفاده از <i>mapreduce</i> و <i>hadoop</i>
34	نگاهی عمیق‌تر به <i>mapreduce</i>

35	نگاشت داده‌ها
35	کاهش داده
36	<i>Hadoop</i> چیست؟
39	معرفی چارچوب‌های پردازش بلادرنگ
40	معرفی پلتفرم‌های پردازش موازی عظیم <i>Mpp</i>
41	معرفی پایگاه داده‌های <i>NoSQL</i>
42	بررسی یک نمونه عملی از مهندسی داده
42	شناسایی چالش‌های مربوط به کسب‌وکار در این زمینه
43	حل مشکلات مربوط به کسب‌وکار با استفاده از مهندسی داده
43	چه مزایایی برای این شرکت حاصل شد؟
45	<b>فصل سوم</b>
45	<b>کاربرد علم داده در صنعت و کسب‌وکار</b>
46	ترکیب بینش‌های حاصل‌شده از داده‌ها در فرآیندهای کسب‌وکار
46	مزایای علم داده‌ی کسب‌وکار محور
47	شناسایی چالش‌های رایج در تجزیه و تحلیل
47	گردآوری داده‌های خام و تبدیل آن‌ها به بینش عملی
49	اقدام بر اساس بینش‌های حاصل‌شده
50	تفاوت هوش تجاری و علم داده
51	تعریف هوش تجاری
56	نگاهی به انواع داده‌ی مورد استفاده در هوش تجاری
56	بررسی تکنولوژی و مهارت‌هایی که در هوش تجاری مفید هستند
58	فرایند <i>ETL</i>
58	تعریف علم داده کسب‌وکار محور
59	بررسی انواع داده‌های مفید در علم داده‌ی کسب‌وکار محور

60	چه فناوری و مهارت‌هایی در علم داده کسب‌وکار محور مفید هستند؟
61	تفاوت‌های اصلی هوش تجاری و علم داده کسب‌وکار محور
63	علم داده در تجارت (داستان موفقیت یک تجارت مبتنی بر داده)
<b>65</b>	<b>بخش دوم</b>
<b>65</b>	<b>استفاده از علم داده در راستای استخراج بینش از داده‌ها</b>
<b>67</b>	<b>فصل چهارم</b>
<b>67</b>	<b>معرفی آمار و احتمالات</b>
68	معرفی مفاهیم اساسی احتمال
68	بررسی رابطه بین احتمال و آمار استنباطی
70	آشنایی با برخی توزیع‌های احتمال رایج
72	معرفی رگرسیون خطی
72	آشنایی با مدل‌های ساده رگرسیون خطی
74	آموزش ساخت خط رگرسیون با اندازه متناسب
75	روش‌هایی ساده جهت محاسبه رگرسیون حداقل مربعات معمولی شبیه‌سازی
79	چگونه این عدد تصادفی به رنگ چشم مرتبط است؟
81	استفاده از شبیه‌سازی برای ارزیابی خواص یک آماره آزمون
82	استفاده از شبیه‌سازی مونت‌کارلو برای ارزیابی خواص یک برآوردگر
85	معرفی تجزیه و تحلیل سری‌های زمانی
85	درک الگوهای سری زمانی
86	مدل‌سازی داده‌های سری زمانی تک متغیره
<b>89</b>	<b>فصل پنجم</b>
<b>89</b>	<b>خوشه‌بندی و طبقه‌بندی</b>
90	معرفی اصول خوشه‌بندی و طبقه‌بندی
91	درک الگوریتم‌های خوشه‌بندی

96	خوشه‌بندی با استفاده از الگوریتم <i>k-means</i> .....
97	روش <i>KDE</i> .....
98	خوشه‌بندی با الگوریتم‌های سلسله‌مراتبی و محلی .....
100	طبقه‌بندی داده‌ها با درخت تصمیم و الگوریتم‌های جنگل تصادفی .....
<b>103</b>	<b>فصل ششم .....</b>
<b>103</b>	<b>خوشه‌بندی و طبقه‌بندی .....</b>
103	با استفاده از الگوریتم نزدیک‌ترین همسایه .....
104	مفهوم بخشی به داده‌ها با استفاده از تجزیه و تحلیل نزدیک‌ترین همسایه .....
105	درک اهمیت خوشه‌بندی و طبقه‌بندی .....
106	طبقه‌بندی داده‌ها با استفاده از الگوریتم‌های میانگین نزدیک‌ترین همسایه .....
107	مقایسه شباهت‌های میانگین با تحلیلگر کسب‌وکار <i>Stu</i> .....
109	طبقه‌بندی با الگوریتم‌های <i>k</i> -آمین-نزدیک‌ترین همسایه .....
111	چه موقع باید از الگوریتم <i>KNN</i> استفاده کنیم؟ .....
112	کاربردهای رایج الگوریتم <i>KNN</i> .....
112	استفاده از نزدیک‌ترین فاصله برای استخراج مفاهیم الگوهای نقاط .....
113	حل مسائل دنیای واقعی به وسیله‌ی الگوریتم‌های نزدیک‌ترین همسایه .....
116	استفاده از روش‌های عددی در علم داده‌ها .....
117	بسط چندجمله‌ای تیلور .....
119	دوبخشی کردن توابع با استفاده از الگوریتم جستجوی دوبخشی .....
120	مدل‌سازی ریاضیاتی با زنجیره‌های مارکوف و روش‌های تصادفی .....
<b>123</b>	<b>فصل هفتم .....</b>
<b>123</b>	<b>مدل‌سازی داده‌های فضایی با آمار .....</b>
124	تولید سطح قابل پیش‌بینی از نقاط داده‌های فضایی .....
125	آشنایی با $(X, Y, Z)$ مدل داده‌های فضایی .....

126	..... معرفی <i>kriging</i>
127	..... کردن به منظور درون‌یابی <i>kriging</i> خودکار
127	..... انتخاب و استفاده از مدل‌های مناسب برای درون‌یابی <i>kriging</i> صریح و تعریف‌شده
127	..... میانگیری با مدل‌های <i>variogram</i> صریح و تعریف‌شده
128	..... تخمین <i>variogram</i>
129	..... نگاهی دقیق به روش <i>kriging rabbit hole</i>
129	..... بررسی روش‌های تخمین سطح <i>kriging</i> های معمولی
130	..... بررسی روش‌های تخمین سطح رگرسیون <i>kriging</i>
130	..... بررسی تخمین زن‌های سطح <i>kriging</i> بلوکی
132	..... بررسی تخمین زن‌های سطح <i>cokriging</i>
133	..... انتخاب بهترین روش تخمین در <i>kriging</i>
135	..... تحلیل باقی‌مانده‌ها برای تعیین بهترین مدل مناسب
138	..... شناخت گزینه‌های موجود در <i>kriging</i>
138	..... استفاده از تحلیل سطح روند بر داده‌های فضایی
<b>141</b>	<b>..... بخش سوم</b>
<b>141</b>	<b>..... ایجاد بصری‌سازی‌های معنادار از داده‌ها</b>
<b>143</b>	<b>..... فصل هشتم</b>
<b>143</b>	<b>..... به‌کارگیری اصول طراحی بصری‌سازی داده</b>
144	..... شناخت انواع بصری‌سازی
144	..... توصیف داستانی داده‌ها برای تصمیم‌گیرندگان سازمانی:
145	..... نمایش داده‌ها برای تحلیل‌گران
145	..... طراحی تابلوی نمایش داده برای فعالان این زمینه
145	..... تمرکز بر مخاطبان
148	..... انتخاب مناسب‌ترین حالت طراحی

148	..... استفاده از بصریسازی برای درک دقیق محاسباتی
149	..... استفاده از طراحی برای ایجاد شور و اشتیاق در مخاطب
151	..... استفاده از داده‌ها برای ایجاد متن
151	..... ایجاد متن با استفاده از حاشیه‌نویسی
153	..... ایجاد متن با استفاده از عناصر گرافیکی
153	..... چه زمانی باید از حالت متقاعدکننده استفاده کنیم؟
153	..... انتخاب مناسب‌ترین نوع داده گرافیکی
154	..... بررسی نمودار استاندارد گرافیکی
157	..... بررسی گرافیک مقایسه‌ای
161	..... بررسی طرح‌های آماری
163	..... بررسی ساختارهای توپولوژی
165	..... بررسی طرح‌ها و نقشه‌های فضایی
167	..... انتخاب گرافیک داده شما
168	..... تعیین دامنه سؤالات
168	..... در نظر گرفتن رسانه و کاربر
168	..... نگاهی مجدد به گام نهایی بیندازید
<b>169</b>	<b>..... فصل نهم</b>
<b>169</b>	<b>..... استفاده از <i>D3.js</i> برای بصریسازی داده</b>
170	..... معرفی کتابخانه <i>D3.js</i>
171	..... چه زمانی باید از <i>D3.js</i> استفاده کنیم؟
172	..... شروع به کار در <i>D3.js</i>
172	..... بررسی <i>HTML</i> و <i>DOM</i>
173	..... بررسی <i>JavaScript</i> و <i>SVG</i>
175	..... بررسی سرورهای وب و <i>PHP</i>



176	..... شناخت مفاهیم پیشرفته‌تر و شیوه عملکرد <i>D3.js</i>
180	..... بررسی ساختار زنجیره‌ای
181	..... بررسی مقیاس‌ها
182	..... مروری بر انتقال‌ها و تعاملات
<b>185</b>	<b>..... فصل دهم</b>
<b>185</b>	<b>..... اپلیکیشن مبتنی بر وب برای طراحی بصری سازی</b>
187	..... کار با تحلیلگر واتسون <i>IBM</i>
188	..... بصری سازی و کار تیمی با استفاده از <i>PLOTLY</i>
191	..... بصری سازی داده‌های فضایی با استفاده از ابزارهای جغرافیایی آنلاین
192	..... ایجاد نقشه‌های خیره‌کننده با <i>Open Heat Map</i>
193	..... ایجاد نقشه و تحلیل داده‌های فضایی با <i>CartoDB</i>
195	..... بصری سازی با ابزار متن‌باز: پلتفرم‌های بصری سازی داده مبتنی بر وب
195	..... ایجاد داده‌های گرافیکی زیبا با جدول‌های <i>Google fusion</i>
196	..... استفاده از <i>iCharts</i> برای بصری‌سازی داده مبتنی بر وب
197	..... استفاده از <i>RAW</i> برای بصری‌سازی داده مبتنی بر وب
199	..... چه زمانی باید از اینفوگرافیک استفاده کنیم؟
199	..... ساختن اینفوگرافیک‌های جذاب با استفاده از <i>Infogr.am</i>
201	..... مهاجرت درون ایالتی
201	..... ایجاد اینفوگرافیک‌های جذاب با استفاده از <i>PiktoChart</i>
<b>203</b>	<b>..... فصل یازدهم</b>
<b>203</b>	<b>..... بررسی بهترین روش‌ها در طراحی داشبورد</b>
204	..... تمرکز بر مخاطبین
205	..... طرح کلی
206	..... طراحی جزئیات

209.....	فصل دوازدهم.....
209.....	ساخت نقشه با استفاده از داده‌های فضایی.....
210 .....	آشنایی با اساس <i>GIS</i> .....
212 .....	فرمت‌های فایل‌های <i>GIS</i> .....
215 .....	درک مفهوم نقشه و سیستم‌های مختصات.....
218 .....	بافر و توابع مجاورت.....
219 .....	استفاده از تحلیل پوشش لایه‌ای.....
220 .....	طبقه‌بندی مجدد داده‌های فضایی.....
221 .....	عملیات ساده اعمال‌شده بر ویژگی‌های دارای هم‌پوشانی.....
221 .....	کار با <i>QGIS</i> منبع باز.....
221 .....	آشنایی با رابط <i>QGIS</i> .....
222 .....	اضافه کردن یک لایه‌برداری در <i>QGIS</i> .....
223 .....	نمایش داده در <i>QGIS</i> .....
229.....	بخش چهارم.....
229.....	محاسبات در علم داده.....
231 .....	فصل سیزدهم.....
231 .....	استفاده از <i>PYTHON</i> برای علم داده.....
232 .....	آشنایی با مفاهیم اساسی در <i>PYTHON</i> .....
233 .....	معرفی انواع داده <i>PYTHON</i> .....
234 .....	اعداد در <i>PYTHON</i> .....
234 .....	رشته‌ها در <i>PYTHON</i> .....
234 .....	لیست‌ها در <i>PYTHON</i> .....
235 .....	تاپل‌ها در <i>PYTHON</i> .....
235 .....	مجموعه‌ها در <i>PYTHON</i> .....

236	.....	واژه‌نامه‌ها در <i>PYTHON</i>
236	.....	استفاده حلقه‌ها در <i>PYTHON</i>
237	.....	آشنایی با توابع و کلاس‌ها
237	.....	معرفی توابع
239	.....	مزیت‌های استفاده از کلاس‌ها
241	.....	معرفی کتابخانه‌ی <i>Numpy</i>
243	.....	آشنایی بیشتر با کتابخانه <i>SciPy</i>
244	.....	استفاده از <i>Matplotlib</i> برای بصریسازی داده
245	.....	استفاده از <i>PYTHON</i> برای تحلیل داده – یک نمونه عملی
246	.....	نصب <i>PYTHON</i> بر روی سیستم‌عامل‌های <i>Mac</i> و <i>Windows</i>
247	.....	بارگذاری فایل‌های <i>CSV</i>
248	.....	محاسبه میانگین وزنی
250	.....	رسم خطوط روند
<b>253</b>	.....	<b>فصل چهاردهم</b>
<b>253</b>	.....	<b>استفاده از زبان <i>R</i> در علم داده</b>
254	.....	تسلط بر لغات مربوط به زبان <i>R</i>
256	.....	محاسبات مربوط به علم داده
257	.....	آشنایی بیشتر با توابع و عملگرها
260	.....	پیمایش در <i>R</i>
261	.....	آشنایی با عملکرد اشیا
264	.....	نگاهی بر پکیج‌های <i>R</i>
264	.....	نگاهی بر پکیج‌های مشهور تجزیه و تحلیل آماری
265	.....	بصری سازی، نگاشت و نمایش نموداری در <i>R</i>
265	.....	بصری سازی آماری <i>R</i> با استفاده از <i>ggplot2</i>

266.....	تجزیه و تحلیل شبکه‌ها با استفاده از <i>igraph</i> و <i>statnet</i>
267 .....	نقشه‌برداری و تجزیه و تحلیل الگوی نقطه‌ای فضایی با به‌کارگیری <i>spatstat</i>
<b>269 .....</b>	<b>فصل پانزدهم</b>
<b>269 .....</b>	<b>استفاده از <i>SQL</i> در علم داده</b>
270 .....	آغاز کار با <i>SQL</i>
270 .....	بررسی پایگاه داده‌های رابطه‌ای و <i>SQL</i>
274 .....	آشنایی با مفهوم طراحی پایگاه داده
274 .....	تعریف انواع داده
275 .....	توجه به قواعد جامعیت
275 .....	نرمال سازی پایگاه داده
277 .....	استفاده از <i>SQL</i> و توابع آن در علم داده
277 .....	یکپارچه‌سازی <i>SQL</i> ، <i>R</i> ، <i>PYTHON</i> و اکسل در استراتژی علم داده
277 .....	استفاده از توابع <i>SQL</i> در علم داده
282 .....	متن کاوی با استفاده از <i>SQL</i>
<b>283 .....</b>	<b>فصل شانزدهم</b>
<b>283 .....</b>	<b>نرم افزارهای متناسب با علم داده</b>
284 .....	مزیت‌های تسهیل‌کننده‌ی اکسل
285 .....	استفاده از اکسل برای درک سریع داده‌ها
286 .....	استفاده از فیلتر در اکسل
286 .....	فرمت بندی شرطی برای نشانه‌گذاری داده‌های خارج از محدوده و روندها
287 .....	رسم نمودار در اکسل برای مشاهده داده‌های خارج از محدوده و روندها
289 .....	فرمت بندی مجدد و خلاصه‌سازی با استفاده از جداول محوری
290 .....	خودکار سازی وظایف اکسل با استفاده از ماکرو (دستورات به کار رفته در <i>Excel</i> )
292 .....	استفاده از <i>KNIME</i> برای تحلیل پیشرفته داده‌ها

297	بخش پنجم .....
297	اعمال مهارت‌های حوزه‌های خاص .....
297	برای حل مسائل حقیقی با استفاده از علم داده .....
299	فصل هفدهم .....
299	بررسی تأثیر علم داده در خبرنگاری .....
301	چه کسی داده‌ها را تولید میکند؟ .....
303	نگاهی دقیق‌تر به اهمیت داستان خبری .....
303	چرا موضوعات شما برای مخاطبان اهمیت دارد؟ .....
305	بهترین زمان را برای انتشار مطالب انتخاب کنید .....
306	چگونه باید داستان‌های نهفته در داده‌های خود را دریابید؟ .....
307	چگونه باید داستان داده محور خود را ارائه دهید؟ .....
307	جمع‌آوری داده .....
308	استخراج داده‌ها برای داستان‌های خبری .....
308	بررسی روندهای غیر معمول و داده‌های خارج از محدوده .....
311	تأکید کردن بر داستان از طریق بصری‌سازی .....
315	فصل هجدهم .....
315	نگاهی به علم داده محیطی .....
316	مدل‌سازی تعامل انسان و محیط با استفاده از هوش محیطی .....
316	نگاهی به انواع مشکلات حل شده .....
317	تعریف هوش محیطی .....
318	شناسایی سازمان‌های بزرگ که در حوزه هوش محیطی کار میکنند .....
319	تأثیرگذاری از طریق هوش محیطی .....
321	مدل‌سازی منابع طبیعی خام .....
321	بررسی مدل‌سازی منابع طبیعی .....

322	مدل‌سازی منابع طبیعی برای حل مشکلات زیست‌محیطی
324	توصیف نقش علم داده در مدل‌سازی آمار فضایی
325	استفاده از آمار فضایی برای رسیدگی به مسائل زیست‌محیطی
<b>327</b>	<b>فصل نوزدهم</b>
<b>327</b>	<b>استفاده از علم داده برای رشد تجارت الکترونیکی</b>
330	ایجاد درکی شفاف از داده‌ها به منظور رشد تجارت الکترونیک
331	بهینه‌سازی سیستم‌های تجارت الکترونیک
332	بررسی جنبه‌های تحلیلی
333	ارزیابی اپلیکیشن‌های پرطرفدار تحلیل وب
334	دسترسی به تحلیل‌ها به منظور یافتن دستاورد
335	استفاده از تحلیل برای فعالیت
336	بررسی تحلیل‌ها برای حفظ کاربران
337	بررسی و آزمایش استراتژی
337	جمع‌بندی انواع روش‌های آزمایش در فرآیند رشد
339	تست دستیابی
339	آزمایش فعالیت
340	بخش‌بندی و هدف‌گذاری برای رسیدن به موفقیت
341	تقسیم‌بندی برای رشد سریع و راحت‌تر تجارت الکترونیک
342	پیدا کردن مخاطبان
342	بهینه‌سازی کانال‌های شبکه‌های اجتماعی
343	بخش‌بندی و هدف‌مندی برای حفظ کاربران
343	بخش‌بندی و هدف قرار دادن درآمدزایی
<b>345</b>	<b>فصل بیستم</b>
<b>345</b>	<b>استفاده از علم داده برای توصیف و پیش‌بینی اعمال مجرمانه</b>

346	تجزیه و تحلیل زمانی برای پیشگیری از جرائم
347	پیش‌بینی و نظارت بر جرائم فضایی
347	نقشه‌برداری جرائم به‌وسیله‌ی تکنولوژی <i>GIS</i>
348	گامی فراتر با مکان - تحلیل تخصیص
349	استفاده از آمار فضایی پیچیده برای درک بهتر جرائم
349	ریاضیات فضایی پیشرفته
350	آمار توصیفی
352	کاوش مشکلات با استفاده از علم داده در تحلیل جرائم
353	توجه به محدودیت‌های فنی
<b>355</b>	<b>بخش ششم</b>
<b>355</b>	<b>آشنایی با ابزار 10</b>
<b>357</b>	<b>فصل بیست و یکم</b>
<b>357</b>	<b>ده منبع پدیده برای داده‌های باز</b>
359	معرفی برنامه <i>Data.gov</i>
360	بررسی داده‌های باز کانادا
361	آشنایی با <i>data.gov.uk</i>
362	بررسی داده‌های دفتر سرشماری ایالات متحده
363	معرفی داده‌های <i>NASA</i>
363	گردآوری داده‌های بانک جهانی
365	آشنایی با داده <i>Knoema</i>
366	صف‌بندی با استفاده از داده‌های <i>Quandle</i>
368	بررسی داده‌های <i>Exversion</i>
369	نگاشت داده‌های فضایی <i>OpenStreetMap</i>
370	منابع

## خط مشی کیفیت انتشارات مؤسسه فرهنگی هنری دیباگران تهران در عرصه کتاب‌های است که بتواند خواسته‌های به روز جامعه فرهنگی و علمی کشور را تا حد امکان پوشش دهد.

حمد و سپاس ایزد منان را که با الطاف بی‌کران خود این توفیق را به ما ارزانی داشت تا بتوانیم در راه ارتقای دانش عمومی و فرهنگی این مرز و بوم در زمینه چاپ و نشر کتب علمی دانشگاهی، علوم پایه و به ویژه علوم کامپیوتر و انفورماتیک گام‌هایی هرچند کوچک برداشته و در انجام رسالتی که بر عهده داریم، مؤثر واقع شویم.

گسترده‌گی علوم و توسعه روزافزون آن، شرایطی را به وجود آورده که هر روز شاهد تحولات اساسی چشمگیری در سطح جهان هستیم. این گسترش و توسعه نیاز به منابع مختلف از جمله کتاب را به عنوان قدیمی‌ترین و راحت‌ترین راه دستیابی به اطلاعات و اطلاع‌رسانی، بیش از پیش روشن می‌نماید.

در این راستا، واحد انتشارات مؤسسه فرهنگی هنری دیباگران تهران با همکاری جمعی از اساتید، مؤلفان، مترجمان، متخصصان، پژوهشگران، محققان و نیز پرسنل ورزیده و ماهر در زمینه امور نشر درصدد هستند تا با تلاش‌های مستمر خود برای رفع کمبودها و نیازهای موجود، منابعی پُر بار، معتبر و با کیفیت مناسب در اختیار علاقمندان قرار دهند.

کتابی که در دست دارید با همت "آقایان محمد جواد جعفری - پرمان محمد علیزاده" و تلاش جمعی از همکاران انتشارات میسر گشته که شایسته است از یکایک این گرامیان تشکر و قدردانی کنیم.

### کارشناسی و نظارت بر محتوا: زهره قزلباش

در خاتمه ضمن سپاسگزاری از شما دانش‌پژوه گرامی درخواست می‌نماید با مراجعه به آدرس [dibagaran.mft.info](mailto:dibagaran.mft.info) (ارتباط با مشتری) فرم نظرسنجی را برای کتابی که در دست دارید تکمیل و ارسال نموده، انتشارات دیباگران تهران را که جلب رضایت و وفاداری مشتریان را هدف خود می‌داند، یاری فرمایید.

امیدواریم همواره بهتر از گذشته خدمات و محصولات خود را تقدیم حضورتان نماییم.

مدیر انتشارات

مؤسسه فرهنگی هنری دیباگران تهران  
[Publishing@mftmail.com](mailto:Publishing@mftmail.com)



## مقدمه مولف

مرور زندگی دانشمندان بزرگ چه در قرون اخیر و چه دورتر منبع الهام بخش مناسبی برای انسان امروز است. نکته ای که به وفور در این بررسی ها مشاهده می شود منابع اطلاعاتی و نوع دسترسی به آن ها توسط انسان است. تعدادی کتاب از گذشتگان یک علم خاص، مقداری مشاهدات محدود و در ادامه تحلیل و نتیجه گیری و ثبت نتایج برای استفاده آیندگان. چالش در این موضوع داشتن منابع اطلاعاتی بیشتر چه از طریق مشاهده و چه از طریق متن های نوشته شده بود. شاید شما هم داستان فردی را شنیده باشید که برای تهیه کتاب گرسنگی می کشید.

برگردیم به عصر حاضر، به قسمت تنظیمات حافظه کامپیوتر رومیزی، تلفن همراه هوشمند و یا سایر تجهیزات همراه خود نگاهی بیندازید. احتمالا تعداد زیادی کتاب با موضوعات مختلف در دسترس شما هست. اگر اهل مطالعه باشید این تعداد ممکن است به صدها عنوان هم برسد. می توانید سری به کتابخانه های اطراف خود بزنید یا یک سرچ ساده در گوگل در رابطه با عنوانی خاص داشته باشید.

دایره المعارف های بسیار بزرگ و کتابخانه های آنلاین را فراموش نکنید! شاید به دلایلی حوصله مطالعه کتاب های قطور را نداشته باشید. بیاید از فضای کتاب و کتاب خواندن خارج شویم. امروز هاوکینز جدیدترین نظریه خود را در صفحه شخصی اش در فضای مجازی منتشر کرد. نظر یکی از هم کلاسی های دوران دانشگاه را در مورد قسمت جدید فیلم جنگ ستارگان در صفحه مربوط به این فیلم خواندم.

خوب بهتر است کمی به کارهای شخصی خود برسیم. با دوستی که قرار است پس از سال ها به دیدنش بروم توسط سیستم آدرس دهی گوگل نقشه محل قرار را مشخص کردیم. همچنین سیستم پیشنهاد دهنده گوگل به ما یک رستوران خوب معرفی کرد. چند عکس یادگاری با گوشی هوشمند خود انداختیم و با سایر دوستان خودمان از طریق اینستاگرام به اشتراک گذاشتیم. البته آن روز همه چیز به خوبی و خوشی پیش نرفت و مجبور شدیم در نهایت به علت مسمومیت غذایی ناشی از حساسیت دوستم به بیمارستان با استفاده از نقشه گوگل برویم.

می توان ساعت ها از حجم بسیار عظیم منابع اطلاعاتی که ثانیه به ثانیه در حال ذخیره شدن و استفاده است صحبت کرد. اطلاعات زیادی از ما در حال ذخیره شدن و اطلاعات زیادی از بیرون در

معروض استفاده ما قرار گرفته اند. همه ی مواردی که گفته شد امروزه نقش کتاب ها و مشاهدات کمیاب گذشته را بازی می کنند که انسان باید برای بدست آوردنشان گرسنگی می کشید.

این مقدار از منابع اطلاعاتی و نیاز به مدیریت درست آن ها چالش جدیدی به نام استفاده از اطلاعات را بوجود آورد. در پس این چالش ها "علم داده" پا به عرصه زندگی گذاشت. سال ها پیش آقایان توماس دونپورت و دی جی پاتیل در سال 2012 در مقاله «علم داده: جذاب ترین شغل قرن بیست و یکم» متخصصین علم داده را این طور تعریف می کنند: کسانی که می دانند چگونه می توان از انبوه اطلاعات بدون ساختار پاسخ سوالهای کسب و کار را پیدا کرد. استنتون در سال 2013 علم داده را این طور تعریف می کند: علم داده رشته در حال ظهوری است که به جمع آوری، آماده سازی، تحلیل، بصری سازی، مدیریت و نگهداشت اطلاعات در حجم بالا می پردازد. در اسکول در سال 2014 علم داده را این طور تعریف می کند: علم داده مهندسی عمران داده هاست.

هدف کلی این کتاب آشناسازی خواننده با دنیای "علم داده" و موضوعات مرتبط با آن است. خواننده پس از مطالعه این کتاب شناخت کلی با زوایای مختلف این دنیای بزرگ خواهد داشت تا پس از مطالعه هر بخش بتواند مسیر خود را در راه تبدیل شدن به یک متخصص داده به درستی دنبال کند. برای افراد غیر مبتدی نیز بخش بندی کتاب بصورتی است که می توانند بخش های مرتبط با حوزه کاری خود را مطالعه کنند هر چند خواندن همه بخش ها جهت داشتن دید بهتر از علم داده مطمئناً خالی از لطف نخواهد بود. در این کتاب پس از بررسی زوایای مختلف این علم به سه بخش اصلی آن می پردازیم که شامل آماده سازی، تحلیل و تصویرسازی اطلاعات می باشد. علم آمار مطمئناً بخشی مهم و جدایی ناپذیر در علوم مرتبط با داده ها است که در این کتاب توجه ویژه ای به آن شده است. لابلا در کنار مطالب مطرح شده در هر ضمیمه ابزارها و زبان های برنامه نویسی کاربردی همراه با مختصری معرفی از نحوه استفاده آن ها عنوان شده است. در فصل های پایانی نگاهی به کاربرد علم داده در محیط زیست، کشف جرائم، خبرنگاری و تجارت الکترونیک انداخته ایم.

این کتاب مقدمه ایست جهت شروع یک راه طولانی که درباره جاده ها، ابزارهای مورد نیاز، مقصد و راه مواجهه با هر پستی و بلندی را به شما نشان می دهد که برای لذت بردن و داشتن انرژی و انگیزه کافی در هر سفر دانستن این موارد لازم است. پس از این کتاب در آینده ای نزدیک بصورت خاص سفر خود را با کتاب "کاربرد پایتون در علم داده" ادامه خواهیم داد.