

# Hiring Guide:

100%  
BS FREE



## *Machine Learning vs Data Scientist*

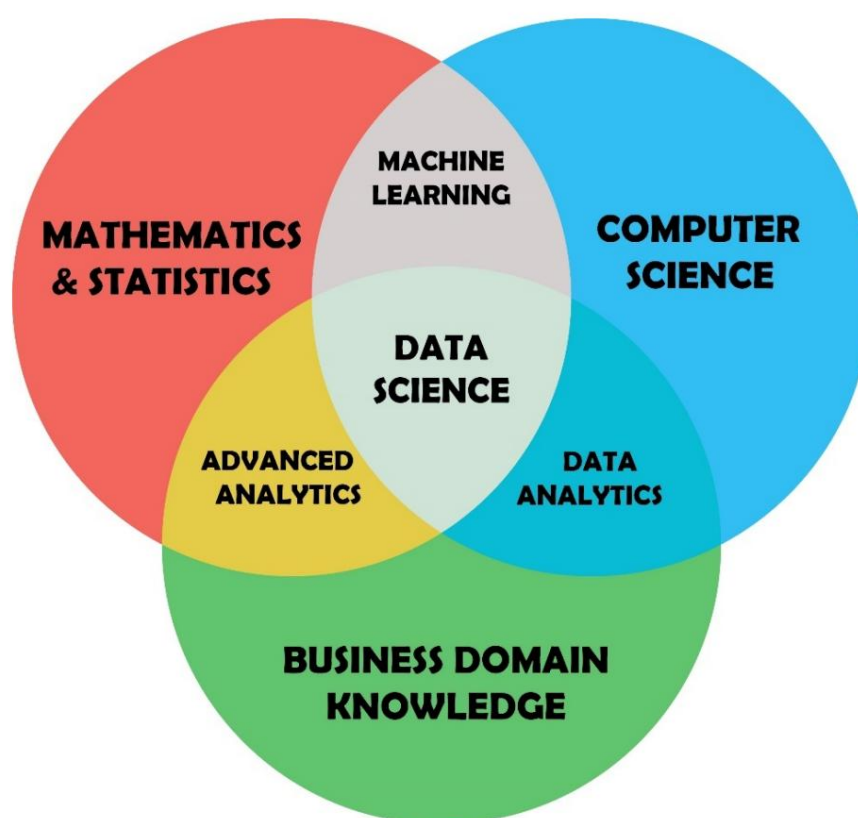
Por:

nabeel khan  
Monsterhub

## ¿Cómo contratar a un científico de datos y un ingeniero de aprendizaje automático?

Estoy bastante seguro de que aterrizaste aquí porque tú, de una forma u otra, te preguntaste, ¿cuál es el diferencia entre un ingeniero de aprendizaje automático y un científico de datos y probablemente buscado en Google "ingeniero de aprendizaje automático versus científico de datos"

Recientemente discutí con mi equipo, y después del debate y escuchando los argumentos de ambos partes, el Sr. Google.com y el equipo, pensé en producir un mapa detallado de dónde la cacofonía es más magnífica.



En términos prácticos, la diferencia entre un científico de datos y un ingeniero de aprendizaje automático difiere de una empresa a otra y su inversión en recursos humanos, y esto llevó a la máquina ingenieros de aprendizaje que enfrentan algoritmos genéricos y preguntas teóricas sobre Estadística Descriptiva y Inferencia bayesiana de probabilidad en entrevistas.

En el pasado, trabajábamos con los mejores investigadores (Ph.D.) sin conocimientos de SQL, dedicados a la desarrollo de una herramienta de BI y a merced de los ingenieros de datos para proporcionar los conjuntos de datos o incluso para realizar operaciones básicas en esos conjuntos de datos.

Estar en la industria durante los últimos 34 años y educar a las empresas sobre cómo aprovechar la tecnología en operaciones diarias, observamos que los reclutadores y el departamento de recursos humanos a veces usan datos Científicos, Analista de negocios, Analista de BI, Términos de aprendizaje automático indistintamente y no hay diferencia en una publicación de trabajo de un científico de datos y un ingeniero de aprendizaje automático.

El propósito de este artículo es compartir nuestra experiencia de contratación y ayudar a nuestros lectores a clasificar las desinformación desordenada y vea la ciencia de datos bajo una luz completamente nueva, más brillante y más clara.

La razón principal de la confusión que rodea la diferencia entre un aprendizaje automático ingeniero y un científico de datos es porque ambos son comparativamente nuevos en otras computadoras tecnologías

Otra causa de confusión que surge de la que acabo de mencionar proviene de recursos humanos, reclutadores, y gerentes de contratación, quienes razonablemente pueden quedar aturdidos con el bombardeo de nuevos términos y condiciones. palabras de moda dando vueltas.

Esto hace que etiqueten los puestos de trabajo de manera incorrecta, a menudo pareciendo que los están eligiendo en un capricho Un representante de recursos humanos puede llamar a un especialista en análisis de datos de puestos de trabajo cuando en realidad necesita un analista de datos, otro puede contratar a un científico de datos junior cuando requiere un negocio analista de inteligencia, por supuesto, hay muchas empresas que redactan sus ofertas de trabajo de manera brillante, pero esto no es estándar en todos los ámbitos, lo que puede causar aún más problemas ahora.

Puede buscar en Internet diferentes glosarios, pero ¿por qué perder el tiempo cuando hemos compilado toda la información para ti incluyendo la infografía.

En el documento a continuación, encontrará una estructura agregada concisa y precisa que contiene todos términos técnicos y comerciales que se utilizan con frecuencia en el campo de la ciencia de datos.

Comenzamos aclarando las similitudes y diferencias entre los términos de business analytics, data ciencia, inteligencia empresarial y aprendizaje automático.

Luego nos enfocamos en ayudarlo a digerir las definiciones que necesita saber de una manera efectiva para que puede sugerir títulos correctos, comprender la diferencia entre la descripción del trabajo de un científico de datos y un ingeniero de aprendizaje automático, sus calificaciones, habilidades técnicas y proporcionarle una plantilla en cómo calificar el nivel de habilidad haciendo las preguntas.

Vamos a sumergirnos directamente.



Léalo aquí.

Simplification es una comunidad para entusiastas del aprendizaje automático y la ciencia de datos, estudiantes, profesionales por igual. Una parte está dedicada a discutir los roles laborales y agradecemos las preguntas de usuarios no tecnológicos a unirse a nosotros y compartir su experiencia con nosotros.

Lo alentamos a unirse a nosotros en <https://discord.gg/uChNJt>



### Evolución de la ciencia de datos

De muchas, una causa de confusión entre un ingeniero de aprendizaje automático y un científico de datos es la constante evolución de la industria de la ciencia de datos, el rápido desarrollo de nuevas herramientas y plataformas, a su vez, complica más la situación. Por ejemplo, alguien que tenía el título de estadístico hace treinta años habría sido responsable de recopilar y limpiar conjuntos de datos y aplicar varios métodos estadísticos a los datos para obtener información significativa.



Del mismo modo, Minería de datos: avance rápido unos años más, el mismo estadístico debido a nuevos Los modelos matemáticos y estadísticos ahora podrían realizar pronósticos más precisos utilizando tecnología. y de nuevo otro término ha encontrado su camino en un glosario de negocios ya inflado llamado Analítica predictiva.

¿El estadístico cambió el trabajo en este punto?

No.

¿Son diferentes los objetivos?

No en realidad no.



Aunque con el crecimiento de los datos y la gran mejora en la tecnología, este estadístico

Ahora sería necesario extraer patrones de los datos. A partir de entonces, se acuñó una nueva palabra de moda.

Con suerte, es fácil ver ahora cómo estos roles se desarrollan y ramifican con el tiempo y cómo alguien

quien se calificaría como estadístico hace treinta años y se hubiera mantenido al día con las tecnologías modernas podría encajar en una multitud de categorías profesionales hoy en día.

La importancia de los datos en esta era: mantener un negocio saludable va de la mano con el trabajo

con datos y no se puede negar que los datos son la base de toda empresa exitosa.

Los empresarios que buscan más profundamente en los datos los harán sobresalir por encima de la competencia.

A medida que los datos crecían en tamaño, el término ciencia de datos se utilizó para describir el arte de extraer datos significativos.

información y conocimiento a partir de datos estructurados y no estructurados. Grandes gigantes tecnológicos como Facebook y

Google con los investigadores crea métodos científicos, procesos, software de algoritmos y herramientas para

administrar y procesar esos datos. A medida que el poder de cómputo creció durante un período de años —Artificial

la inteligencia se convirtió en aplicaciones prácticas de la vida real a partir de solo un tema novedoso en ciencias de la computación.

La mayoría de estas tareas requerían que se procesara una gran cantidad de datos, ya que los conocimientos se volvieron

claro a partir de los datos: el aprendizaje automático se deriva de la inteligencia artificial, donde los datos

Los algoritmos permiten que las aplicaciones se vuelvan altamente precisas y realicen cálculos pesados y

predecir y actuar sin necesidad de programación explícita.

Todos hemos experimentado el aprendizaje automático en acción de una forma u otra. Si usaste Netflix y

visto la recomendación de la película o comprado en Amazon y visto esos productos personalizados

recomendaciones, esas predicciones son el resultado del aprendizaje automático en acción.

Principalmente, el aprendizaje automático utiliza algoritmos y desarrolla modelos que pueden recibir datos de entrada y

prediga la salida, aproveche el poder de los modelos estadísticos y matemáticos mientras actualiza las salidas

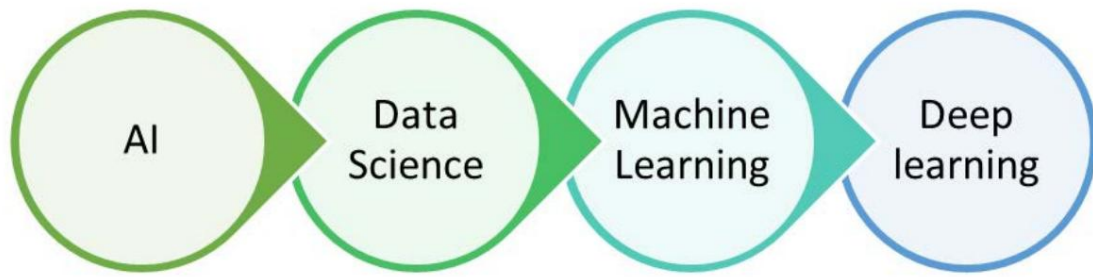
a medida que se disponga de nuevos datos. Los procesos involucrados tienen mucho en común con los procesos predictivos.

modelado y minería de datos. Esto se debe a que ambos enfoques exigen que uno busque a través de los datos.

para identificar patrones y ajustar el programa en consecuencia.







## Análisis vs Analítica

### Análisis

Datos pasados

Experiencia

análisis de los datos

análisis de negocios

### Analítica

predetección

Potencial

análisis de datos

Análisis de negocio

Digamos que queremos saber qué podemos hacer después de que el cliente ya haya comprado un producto de ¿a nosotros? Para resolver un problema empresarial, el ciclo suele comenzar con el equipo de datos. La base de datos los administradores harán una cantidad significativa de trabajo en los datos para que estén disponibles para más análisis, según disponibilidad. El equipo de inteligencia empresarial proporcionará información empresarial tablero, y el equipo de ciencia de datos es el principal responsable de traducir este problema comercial en un modelo matemático más técnico que, basado en una entrada de características en los retornos, proporcionará predicción precisa o tomará una acción basada en alguna predicción. Un ingeniero de aprendizaje automático probablemente tomaría ese modelo que desarrollaron los científicos de datos y lo integraría con el resto de la plataforma de la empresa, y eso podría implicar la creación, por ejemplo, de una API en torno a este modelo para que que pueda ser servido y consumido, y luego ser capaz de mantener la integridad y calidad de este modelo para que siga sirviendo predicciones realmente precisas.



## Diferencia entre un científico de datos y un ingeniero de aprendizaje automático

Piénsalo de esta manera. Un científico necesita entender completamente la *ciencia* detrás de su trabajo donde un ingeniero tiene la tarea de construir *algo*: una casa, una placa base o incluso un programa de computadora.

Los científicos de datos son expertos analíticos que utilizan sus habilidades tanto en tecnología como en ciencias sociales. para encontrar tendencias y gestionar datos. Utilizan el conocimiento de la industria, la comprensión contextual, el escepticismo. de los supuestos existentes, para descubrir soluciones a los desafíos comerciales.

*El campo de la ciencia de datos emplea disciplinas informáticas como matemáticas y estadísticas e incorpora técnicas como minería de datos, análisis de conglomerados, visualización y, sí, aprendizaje automático.*

Expliquémoslo con una analogía. ¿Cuál es la diferencia entre un científico farmacéutico y un ¿doctor en Medicina? Ambos conocen el uso de la medicina, pero uno de ellos es parte de crearlo y el segundo es prescriptor de ella. En el caso de un científico de datos y un ingeniero de aprendizaje automático, la diferencia es mucho elusivo.

## Científico de datos versus ingeniero de aprendizaje automático: qué hacen

Como se discutió anteriormente, existen similitudes cuando se trata de los roles de los científicos de datos y la máquina. ingenieros de aprendizaje. Ambos roles usan más o menos la misma pila de tecnología, donde el primero implementa lo que este último había diseñado.

Sin embargo, si observa ambos roles como miembros del mismo equipo, los científicos de datos son grandes datos wranglers, recopilando y analizando grandes conjuntos de datos estructurados y no estructurados. haciendo el análisis estadístico y matemático necesarios para determinar qué enfoque de aprendizaje automático usar o cómo combinar más de un enfoque, luego modelan el algoritmo y lo prototipo para pruebas. En ese momento, un ingeniero de aprendizaje automático toma el modelo prototipo y lo porta a el entorno de producción.

Volviendo a la división entre científico e ingeniero, no se espera necesariamente un ingeniero de aprendizaje automático para comprender los modelos predictivos y sus matemáticas subyacentes de la forma en que lo hace un científico de datos. Sin embargo, se espera que un ingeniero de aprendizaje automático domine las herramientas de software que hacen que estos modelos utilizables.





	Científico de datos	Ingeniero de aprendizaje automático
	Requisitos	
01.	Resolver problemas empresariales a través de investigación no dirigida y encuadre abierto preguntas terminadas de la industria	Resolviendo problemas de escalabilidad y confiabilidad a través de la investigación y el encuadre no dirigidos preguntas basadas en escenarios.
02.	Realizar análisis de datos exploratorios (EDA) para determinar cómo manejar los datos faltantes y para buscar tendencias y/u oportunidades.	Realizar análisis de datos exploratorios (EDA) para determinar cómo manejar la alimentación de datos y rendimiento de aprendizaje automático.
03.	Descubriendo nuevos algoritmos y desarrollando nuevos modelos.	Implementación de modelos ML y ajuste hiperparámetros del modelo.
04.	Identificar nuevas oportunidades, descubrir tendencias, correlaciones y patrones en complicados conjuntos de datos	Amplíe las bibliotecas ML existentes y Marcos para acomodar negocios requisitos
05.	Resuelve un problema empresarial utilizando estadísticas y algoritmos matemáticos	Implementa la solución en la producción. ambiente
06.	Desarrolla soluciones para solucionar el negocio problema.	Optimiza las soluciones para el rendimiento y escalabilidad
07.	Define la fuente de datos después de una prueba exhaustiva diferentes modelos de aprendizaje automático	Introduce datos en modelos y los amplía a la producción
	Educación	
	Doctorado preferido o al menos maestría en ingeniería, informática, o matemáticas y estadística	Maestría preferida al menos un licenciatura en informática, matemáticas o estadística.
	Científico de datos	Ingeniero de aprendizaje automático
	Habilidades	
08	R, Python, Java, C, C++, JavaScript, Scala, Julia, Octava	
09	Chispa, Hadoop, colmena, cerdo, SQL, Neo4J, MySQL,	
10	Theano, PyTorch, TensorFlow, Keras, Scikit-learn, Apache Spark	



	Algoritmos
11	Clasificador naïve bayesiano, agrupamiento de medios K, máquina de vectores de soporte, a priori, lineal Regresión, Regresión logística, Redes neuronales artificiales (ANN), Neural convolucional redes (CNN), redes neuronales recurrentes (RNN), bosques aleatorios, árboles de decisión y Vecinos más cercanos

## Conclusión

En la era actual, los científicos de datos y los ingenieros de ML comparten similitudes significativas en lo que respecta al conjunto de habilidades y, como hemos visto, no hay un conjunto estricto de instrucciones a seguir cuando se trata de contratar a cualquiera de ellos. Aquí está nuestra opinión sobre este tema.

- Un científico de datos con un Ph.D. será un experto con conocimientos avanzados en estadística modelado.
- Tanto los científicos de datos como los ingenieros de ML tienen un conocimiento profundo de los modelos de ML, Algoritmo y redes neuronales.
- Los científicos de datos y los ingenieros de ML son organizadores de big data, recopilan y analizan grandes conjuntos de datos estructurados y no estructurados. En promedio, el 70% de su tiempo dedicado a trabajar con datos: 20 % de modelado y 10 % operativo.
- Científicos de Datos e Ingeniero ML con maestría en Ciencias de la Computación y Las estadísticas tienen una comprensión firme de la teoría de algoritmos y temas como Gradient Descenso y Diferenciación Parcial.
- La experiencia puede ser un sustituto de un título.
- Tanto los científicos de datos como los ingenieros de ML deben tener la capacidad de analizar información objetivamente y emitir un juicio razonado. • Pensamiento crítico • Implica la evaluación de datos, hechos, fenómenos observables y hallazgos de investigación.
- Lingüística e Informática • Procesamiento de audio, video y lenguaje natural. Un El ingeniero de ML tendrá un conocimiento profundo de OpenCV, spaCy y NLTK.
- Los científicos de datos y los ingenieros de ML con experiencia en conjuntos deben ser preseleccionados primero.



• Siempre pídeles que expliquen los conceptos difíciles en términos simples.

Y recuerda

• Los proyectos de ML son un círculo vicioso de investigación y desarrollo.



## Escala

Hay muchas razones por las que podríamos querer una escala de experiencia, desde calificar a los empleados potenciales hasta crear un resumen de la experiencia dentro de una organización.

	Nombre del nivel	Descripción	Práctica
0	<b>ignorante</b>	Nunca escuché sobre eso.	ninguna
1	<b>interesado</b>	Escuché un poco sobre eso, pero no sé mucho.	1 hora
2	<b>perseguir</b>	He leído uno o dos artículos al respecto y entiendo el conceptos básicos de lo que es, pero nada en profundidad.	1 día (5 horas)
3	<b>principiante</b>	He leído un artículo detallado, un manual básico o un libro de instrucciones, y/o he jugado un poco con él.	1 semana (25 horas)
4	<b>aprendiz</b>	Lo he usado durante al menos unos meses y tengo completó con éxito un pequeño proyecto usándolo.	3 meses (250 horas)
5	<b>intermedio</b>	Lo he usado durante un año o más de forma diaria o regular, y me siento cómodo usándolo en moderadamente complejo proyectos	1 año (1.000 horas)
6	<b>avanzado</b>	Lo he estado usando durante muchos años, sé todo lo básico aspectos, y me siento cómodo usándolo como un elemento clave en proyectos complejos La gente de mi grupo viene a mí con sus preguntas	5 años (5.000 horas)
7	<b>cumplido</b>	Soy un experto local, con diez o más años de sólida experiencia. La gente de mi división viene a mí con sus preguntas.	10 años (10.000 horas)
8	<b>Maestro</b>	Soy un gurú de toda la empresa con veinte años o más de experiencia; gente de otras divisiones viene a mí con sus preguntas	20 años (20.000 horas)
9	<b>gran maestro</b>	Soy una autoridad internacional reconocida al respecto.	30 años (30.000 horas)
10	<b>genial</b> <b>gran maestro</b>	Yo lo creé, y soy el experto número 1 en el mundo.	50 años (50.000 horas)



## Clasificación de habilidades

Hemos creado una matriz de roles y habilidades, con el nivel de comprensión dividido en la tabla de escala uno.

	Negocio	BI	Datos	Datos	ML
	Analista	Analista	Analista	Científico	Ingeniero
<i>Correlación de datos</i>	✓	✓	✓	✓	✓
<i>Creando datos</i>	✓	✓	✓	✓	✓
<i>Creación de informes</i>	✓				
<i>Creando y ejecutando what-if simulaciones</i>	✓		✓	✓	✓
<i>Análisis de los datos</i>	✓	✓	✓	✓	✓
<i>Arquitectura de datos</i>			✓		
<i>Recopilación de datos</i>	✓	✓	✓	✓	✓
<i>Controles de datos</i>			✓		
<i>Desarrollo de datos</i>			✓		
<i>Gestión de datos</i>			✓		
<i>Modelado de datos</i>			✓		
<i>Procesamiento de datos</i>			✓	✓	
<i>Visualización de datos</i>			✓	✓	
<i>familiaridad con la base de datos</i>	✓	✓	✓	✓	✓
<i>Definición de métodos de acceso a datos</i>	✓	✓	✓	✓	✓
<i>Diseño de informes de nivel empresarial</i>	✓	✓			
<i>Desarrollo de consultas SQL complejas y informes</i>		✓	✓	✓	✓
<i>Encontrar tendencias/patrones</i>				✓	
<i>Identificar tendencias comerciales.</i>				✓	
<i>Conocimiento de la industria</i>				✓	
<i>MatLab</i>				✓	✓
<i>Modelado</i>			✓		
<i>Monitoreo de tendencias comerciales</i>	✓			✓	



<i>Supervisión de la calidad de los datos</i>		✓	✓	✓
<b>NO SQL</b>		✓	✓	✓
<b>Procesamiento analítico en línea (OLAP)</b>	✓			
<i>Pitón</i>			✓	✓
<b>Representar datos gráficamente</b>	✓		✓	
<b>Desarrollo de soluciones</b>			✓	✓
<b>programación SQL</b>	✓	✓	✓	✓
<b>análisis estadístico</b>	✓		✓	
<b>Conocimiento estadístico</b>			✓	
<b>Cuadro</b>			✓	✓
<b>Comprender las tendencias</b>	✓		✓	



# Machine Learning



**NEW WAY OF THINKING**

# Data Scientist



## Data Architect

designs the way data is generated, processed, retrieved and consumed



## Database Administrator

handles the control of data, optimize db retrieval and management tasks



## BI Consultant

external BI consultant



## Data Scientist

employs traditional statistical methods or unconventional ml techniques for making predictions



## ML learning

applies state of the art computational models

## Data Engineer

processes the data so that it can be used for further analysis



## BI Analyst

performs analysis and reporting on historic data



## BI developer

writing SQL and code to publish business intelligence dashboard

















































## Data Engineer

prepares, converts, designs flow of data for advance analysis





	<b>Data</b>		<b>Data Science</b>		
	TRADITIONAL	BIG	BUSINESS INTELLIGENCE	TRADITIONAL METHODS	MACHINE LEARNING
WHEN it is applied	At the beginning of your analysis		After the data has been gathered & organized	After BI reports have been created and discussed	
	PAST		NOW	FUTURE	
WHY you need it	data-driven decisions require well-organized and relevant raw data stored in a digital format		use data to create reports and dashboards to gain business insights	Predictive Analytics	
WHAT techniques are involved	<b>DATA COLLECTION</b> <b>PREPROCESSING</b> <ul style="list-style-type: none"> <li>class labeling (categorical vs numerical)</li> <li>data cleansing</li> <li>dealing with missing values</li> </ul> <b>CASE SPECIFIC</b> <ul style="list-style-type: none"> <li>e.g. balancing &amp; shuffling datasets</li> </ul>		<b>ANALYZE THE DATA</b> <b>EXTRACT INFO AND PRESENT IT IN THE FORM OF:</b> <ul style="list-style-type: none"> <li>metrics</li> <li>KPIs</li> <li>reports</li> <li>dashboards</li> </ul>	<b>REGRESSION</b>  <b>LOGISTIC REGRESSION</b> <b>CLUSTERING</b>  <b>FACTOR ANALYSIS</b> <b>TIME SERIES</b> 	<b>SUPERVISED LEARNING</b> <ul style="list-style-type: none"> <li>SVMs</li> <li>NNs</li> <li>deep learning</li> <li>random forests</li> <li>bayesian networks</li> </ul> <b>UNSUPERVISED LEARNING</b> <ul style="list-style-type: none"> <li>k-means</li> <li>deep learning</li> </ul> <b>REINFORCEMENT LEARNING</b> similar to supervised learning, but instead of minimizing the loss, one maximizes reward
	 				
WHERE	<b>BASIC CUSTOMER DATA</b> <b>HISTORICAL STOCK PRICE DATA</b>	<b>SOCIAL MEDIA</b> <b>FINANCIAL TRADING DATA</b>	<b>PRICE OPTIMIZATION</b> <b>INVENTORY MANAGEMENT</b>	<b>USER EXPERIENCE (UX)</b> <b>SALES FORECASTING</b>	<b>FRAUD DETECTION</b> <b>CLIENT RETENTION</b>
HOW using what tools	<b>PROGRAMMING LANGUAGES</b>     <b>SOFTWARE</b>  	<b>PROGRAMMING LANGUAGES</b>     <b>SOFTWARE</b>   	<b>PROGRAMMING LANGUAGES</b>     <b>SOFTWARE</b>     	<b>PROGRAMMING LANGUAGES</b>    <b>SOFTWARE</b>    	<b>PROGRAMMING LANGUAGES</b>       <b>SOFTWARE</b>    
WHO	<b>DATA ARCHITECT</b> <b>DATA ENGINEER</b> <b>DATABASE ADMINISTRATOR</b>	<b>BIG DATA ARCHITECT</b> <b>BIG DATA ENGINEER</b>	<b>BI ANALYST</b> <b>BI CONSULTANT</b> <b>BI DEVELOPER</b>	<b>DATA SCIENTIST</b> <b>DATA ANALYST</b>	<b>DATA SCIENTIST</b> <b>MACHINE LEARNING ENGINEER</b>
ARE YOU AWARE	200,000 lines of data is not necessarily big data. It is not just volume that defines a data set as 'big' - variety, variability, velocity, veracity, and other characteristics are determinative as well.		Qualitative analysis tools such as SWOT are not used for quantitative analysis. Hence, they are not part of business intelligence.	Software like Excel, SPSS, and Stata, can be successfully used by data science teams in many companies.	In deep learning, there is still a debate on WHY the algorithms used outperform all conventional methods.